

Unit 5:

Computing and Data Analysis



EXPLORING COMPUTER SCIENCE

Introduction

Managing and interpreting large amounts of data is part of the foundation of our information society and the economy. Where there are humans, there are data. Computing has enabled researchers to use data to explore questions related to large global issues such as climate change, animal habitats, and human behavior. With the vast amounts of data now available for everyone to use, effective members of society need to be able to use data to address their own questions and to corroborate or question “official statistics” when they are published in various formats. The ability to analyze, visualize and draw conclusions from large data sets is critical to computing.

This unit has been designed to allow students the opportunity to experience the process of data collection and analysis in real-world contexts. There are a variety of tools that can be used for data analysis. The choice of tool(s) to use is up to you and should be based on local needs and resources. While it may not be possible to utilize tools that are suitable for very large data sets and/or all of the analysis features discussed in this unit, every effort should be made to do multiple types of analysis and to discuss tradeoffs in the tools used. Students need to be able to use whatever tools are chosen in appropriate ways to manipulate the data, but the focus should be on conceptual understanding of data analysis--making appropriate inferences, using data to make a case or inform a discovery, and being able to justify conclusions.

The unit is divided into three main sections.

- Overview of the final project, data collection, and validating claims with data (Days 1-8)
- Data analysis techniques (Days 9-25)
- Final project (Days 26-30)

The goal is to prepare students to collect rich data, formulate queries that will inform whatever project is chosen and to use that information to either make a case or facilitate a discovery.

The data sets provided are examples. Teachers should choose data sets and examples that will be appropriate in local context. Attention should be paid to ensuring that the chosen data sets exercise the same data analysis techniques as those currently provided.

Specific topics for each instructional day are listed in the overview chart on the next page.

Daily Overview Chart	
Instructional Day	Topic
1-3	Review how data can be used for making a case/discovery. Explore pitfalls and challenges of putting together and managing large sets of data. Provide an overview of the final project.
4-5	Explore possible research questions for a selection of sample campaigns. Validate compelling stories with research data.
6-7	Assign groups. Discuss group roles and responsibilities. Choose campaigns and modes for data collection.
8	Data check-in—Discuss issues that arise (aggregating data, etc.).
9-12	Create maps using the latitude and longitude of a location and then create maps from a file of data.
13	Create maps with student data and related data set.
14-16	Discuss bar plots, categorical and continuous data, and mosaic plots as a vehicle for comparing categorical data, and looking at trends in data.
17	Create bar plots and mosaic plots with student data and related data set.
18-20	Review mean, median, minimum, maximum. Discuss various ways to subset data. Represent data with box plots and histograms.
21	Identify mean, median, minimum, maximum, create subsets, and create box plots and histograms with student data and related data set.
22-24	Use a variety of filters and queries to create subsets of text data. Create bar plots to graphically display the information.
25	Analyze text in student data and related data set.
26-27	Finalize data analysis for final project.
28-29	Develop website or Scratch program to present data analysis campaign.
30	Final project presentations

Daily Lesson Plans**Instructional Days:** 1-3

Topic Description: This lesson sets the stage for the unit. It provides a review of how data can be used for making a case or as a vehicle for discovery. Students explore the pitfalls and challenges of putting together and managing large sets of data. An overview of the final project is provided.

Objectives:

The student will be able to:

- Explain the possible themes for the final project
- Explain the difference between data used for making a case and data that informs discovery
- Identify and discuss the considerations that must be made in order for a large data set to be useful
- Consider how various types of data (numbers, text, dates, etc.) lend themselves to processing
- Collaborate with others to create, manage, and maintain a large set of data

Outline of the Lesson:

- Overview of final project (35 minutes)
- Review of data collection and making a case/discovery (5 minutes)
- Journal entry (15 minutes)
- Introduction to merging of diverse data sets about the same subject (10 minutes)
- Merge individual data sets together (45 minutes)
- Discussion of choices/decisions/compromises made during data merging (30 minutes)
- Discussion of creating an entire-class data set (25 minutes)

Student Activities:

- Participate in discussion of final project.
- Participate in discussion of data review.
- Complete journal entry.
- Participate in discussion about data formats.
- Groups work to merge individual data sets together.
- Participate in discussion of choices/decisions/compromises made during data merging.
- Participate in discussion about creating an entire class data set.

Teaching/Learning Strategies:

- Introduce students to some possible research topics for the final project by using the suggested videos or other high-interest and engaging resources. The goal is to spark their interest and desire to learn more about the topic, and to brainstorm how they can be citizen scientists on the topics. The topics below are arranged from less to more complex.
 - School safety and bullying

- Teen exercise and health
- Teens, social media, and online behavior
- Teens, video games, and civic engagement

Resources for the topics are included in Sample Final Project Topics. Additional resources related to these topics are included in later lessons.

- Journal Entry: In previous units we have discussed and considered many of the ways that data in the world are generated and collected. The term “Big Data” is often used to refer to the notion of very large data sets, and dealing with Big Data is something that many computer scientists concern themselves with. What does Big Data mean to you? In particular how big do you think a data set needs to be in order for it to qualify as Big Data? Do you think there is a formal definition? Why or why not?
 - Have students share their responses with their elbow partner.
 - Discuss student responses.
- Review the differences between data used for making a case and making a discovery.
- Data project—You can use the Room Data Project or create one of interest to your students that includes the following features and types of data:
 - Text
 - Spatial
 - Qualitative
 - Quantitative

and for which students can collect multiple entries by the next class period. You may wish to have them collect data related to one of the sample final project topics as an alternative. (The Room Data Project is adapted from a version by Baker Franke and Jeff Solin include in the *Taste of Computing* version of ECS.)

- Access to large data sets provides opportunities for in depth analysis, but in order to do this the data must be processed. In this project students will explore the challenges and pitfalls of putting together and managing a large set of data. Many choices can be made with no clear right or wrong answers. The goal is for the class to develop a single large set of data to use, formulate questions about the data, and answer them using an analysis tool. Each class’s will be different depending on the choices they make.
- Assign data collection for homework to be discussed the next class period.
- Break class into groups of 4 to merge their personal data sets into one.
 - Individuals in each group should share their school/room data spreadsheets with each other and discuss the choices they made.
 - How was Happiness recorded?
 - How were dates/times written?
 - How were locations recorded?
 - Did everyone record the same number of entries for each day?
 - If the goal is to merge the group’s data into one consistent set, how might the number of entries per student be affected?
 - It is likely that each person recorded their data differently.
 - The group’s task is to merge their four data sets into one consistent data set.
 - The group should agree on how to denote days, times, locations and happiness.

- It may be that some data needs to be compromised in order to conform to the group standard.
 - UNDER NO CIRCUMSTANCES should new data be “invented” after the fact to fit the mold.
 - The group should also agree on how many records (rows) of data should be allocated for each person.
 - It need not be exactly the same number per person, but it should be representative of the group
 - For example if one student has a minute-by-minute account of their day, while another just made one entry for each class they were in, how can they make that consistent?
 - Once the group agrees on a standard, they must produce one file that represents everyone’s data from the group
 - Students should use whatever collaboration technology is available to divide up the work of doing this - you shouldn’t have one student re-typing all the data. A “low-tech” suggestion is for each student to re-type/convert their personal data, save the file and send it to one person who can copy/paste the sheets together.
 - Note: this activity can also be done on poster paper without actual data entry, if desired.
 - By the end of class each group should have saved a well-formatted file containing the group’s data.
- For homework students should now collect data in the group’s agreed upon format.
- Groups share their merged data (e.g. printing, projector, online tool)
 - Have a member from each group explain some of the choices that they made, explain challenges they overcame.
 - Guiding questions:
 - Is it important or not that the resulting file be consistent?
 - What does consistency in the data mean?
 - Why might it be important or why might it not be important?
 - What about differences in the number of entries per student? How might that affect the data?
- Have each group develop an interesting question (or two) to ask of the data
 - Examples:
 - What is the happiest room in school? least happy?
 - What are the top 3 visited rooms by people in this class?
 - What time of day is everyone busiest?
 - What day of the week are people happiest? least happy?
- Introduction to thinking about big data.
 - Have groups share their questions and answers—foreshadow the need for a larger set of data.
 - These are questions that we can ask when we have a large set of data to work with.
- As an entire class discuss how to merge all groups’ data into one data set.
- Goal: For each of the four types of data in the Room Data Collection Project, the entire class must agree on an appropriate format that will make processing the data to answer questions of it possible.
- For homework students should now collect data in the class’ agreed upon format.

Resources:

- Room Data Collection Project
- Sample Final Project Topics

Room Data Collection Project

During the next several days, you will generate a collection of data about your movements within the school during the day and how you were feeling at that time. Ideally you will carry this paper with you and take notes over the course of your day. You may also transcribe it into your journal in an organized way. You should have at least one entry for each class you are in each day, but you can have more.

For the columns, *Day*, *Time* and *Location*, you should record the data in whatever way you think is most appropriate. For the last column labeled *Happy* record some measure of your happiness at that point in the day. How you record this is up to you.

Day	Time	Location/Room	Happy

Sample Final Project Topics

Topic 1: School Safety and Bullying

File: schoolSafety_dataset.csv

Introductory Resources:

- Bullying facts from the National Assoc. of School Psychologists: http://www.nasponline.org/resources/factsheets/bullying_fs.aspx
- More facts about bullying: <http://facts.randomhistory.com/facts-about-bullying.html>
- Youth Online: View state data. Select a state, click “GO”. Filter the data by selecting the Health Topic “Unintentional Injuries and Violence.” View the Table or Graph by selecting the appropriate tab: <http://nccd.cdc.gov/YouthOnline/App/Default.aspx>. Focus on the questions related to bullying.
- See the state-by-state grades for anti-bullying laws: <http://www.bullypolice.org/>
- YouTube video: You Are You <http://www.youtube.com/watch?v=l8ovOA7VhFo>
CAUTION: Preview before showing to evaluate suitability for your classroom.

Topic 2: Exercise and Health

File: exerciseHealth_dataset.

Introductory Resources:

- FitWebMD: <http://fit.webmd.com/teen/move/article/exercise-personality>
 - Play Bubble Rubble with Chicken Dawg
 - What’s your Exercise Personality?
 - Go From Couch to 5K in just 8 Weeks
- Physical activity recommendations: <http://www.cdc.gov/physicalactivity/everyone/guidelines/children.html>
- Youth Online: View state data. Select a state, click “GO”. Filter the data by selecting the Health Topic “Physical Activity.” View the Table or Graph by selecting the appropriate tab. <http://apps.nccd.cdc.gov/YouthOnline/App/Default.aspx>
- Rewards and benefits of exercise: http://kidshealth.org/teen/food_fitness/exercise/exercise_wise.html#cat20133

Topic 3: Teens, Social Media, and Online Behavior

File: Mobilize_MediaBehavior_2011.csv

Introductory Resources:

- Video Social Media Helps Teens See Other’s Needs: <http://www.webpronews.com/social-media-teens-others-needs-2012-02>
- Social Media, Social Life – What Do Teens Think About Twitter And Facebook? [INFOGRAPHIC] http://www.mediabistro.com/alltwitter/social-media-teens_b24749
- Video Tagged (cyber bullying) Teens and Social Media <http://planningwithkids.com/2011/09/30/teens-and-social-media/> Includes teaching resources: <http://www.cybersmart.gov.au/tagged/teenagers.htm>

- An Analytical Take on Youth, Social Networking, and Web 2.0: A Few Moments with Amanda Lenhart <http://dmlcentral.net/newsletter/05/2012/analytical-take-youth-social-networking-and-web-20-few-moments-amanda-lenhart>
- Teens, kindness and cruelty on social network sites: <http://www.pewinternet.org/2011/11/09/teens-kindness-and-cruelty-on-social-network-sites/>

Topic 4: Teens, Video Games, and Civic Engagement

File: Mobilize_GamesCivic_2008_xformed.csv

Introductory Resources:

- 22 Charts & Graphs on Video Games & Youth Violence <http://videogames.procon.org/view.resource.php?resourceID=003627>
- How much do you know about video games <http://www.esrb.org/about/video-game-industry-statistics.jsp>
- Video game stats <http://www.onlineeducation.net/videogame>
- Video: [Do video games cause violence in kids? Interesting facts and data](#)
- <http://www.youtube.com/watch?v=vVjUr2zXtb4>

Instructional Days: 4-5

Topic Description: Research questions for a selection of campaigns are explored. Validating claims with research data are discussed.

Objectives:

The student will be able to:

- Understand the complexities of collecting, processing, and analyzing large sets of data
- Describe the expectations of the final project
- Explore data sets related to sample final project topics
- Identify research questions that could be answered or stories that could be told from the data
- Select topic for the final project

Outline of the Lesson:

- Debrief of Room Data Collection Project as a Big Data problem (35 minutes)
- Description of final project expectations (20 minutes)
- Examine the available data sets on the sample final project topics (30 minutes)
- Discuss possible research questions that might be answered or stories that might be told with additional information (15 minutes)
- Select topic for the final project (10 minutes)

Student Activities:

- Participate in debrief of Room Data Collection Project.
- Participate in discussion of final project expectations.
- Analyze the available data sets for the sample final project topics.
- With an elbow partner discuss possible research questions that might be answered or stories that might be told with additional information.

Teaching/Learning Strategies:

- Debrief of Room Data Collection Project
 - In their groups, have students review the latest set of data collected and respond to the following questions.
 - What do wish we'd thought about before starting the Room Data Collection Project?
 - Do you think it's possible to think of *everything* ahead of time? If not, what steps can you take to make sure that you can recover from missteps?
 - Any other data you wish we had collected from the outset?
 - Anything you wish we'd specified more clearly before collecting data?
 - When designing a method for collecting a large set of data using multiple sources, what kinds of things do you need to think about?

- Have each group share their responses and discuss as a class.
- Describe the final project expectations.
 - The class will select from one of the sample final project topics.
 - You may decide that each student group can choose a different topic, but this will mean that there is less student generated data for each.
 - You may have students generate a completely different list of topics that are of interest, but you will need to find appropriate data sets to use as the additional contextual data sets.
 - Each group will decide on the research questions to ask about the data and the story they hope to tell. Each group will pose their own research questions. Each group will contribute questions to the survey in order to gather the data necessary to answer their research questions.
 - Design the survey to get at the information they need to tell the story.
 - Gather data.
 - Analyze their collected data and data from other sources.
 - Draw conclusions, augment, reinforce, tell a story, verify truths, call into question, or further discover and describe the topic.
 - Report their findings through a website or Scratch project.
- Examine the available data set on each topic in a data analysis tool. (See Data Analysis and Research Questions.)
- Discuss possible research questions that might be answered or stories that might be told with additional information.
- **IMPORTANT CONSIDERATION:** The size of data sets for the four topics varies significantly in the number of variables. The complexity of the data sets adds to the complexity of strategies for analyzing the data and to the conceptual complexity for students to process the information, formulate research questions, and execute a successful project. The variables in the two larger data sets (Topics 3 & 4) have been rearranged to enable projects to be scaffolded more easily. It might be useful to guide students toward project topics or to subtopics within the four major topics that match their needs and resources. The numbering of topics (1 – 4) indicates increasing complexity. Each topic can be more narrowly focused to make the content manageable for most learners. Refer to the Possible Research Questions in Data Analysis and Research Questions for scaffolding suggestions.
- Guide students to select a topic for their project.
 - The entire class may use the same topic, and individual groups ask different research questions and include variables in the survey campaign to gather data unique to their questions.
 - An alternative strategy is to allow groups to select their own topic from the suggested four topics.
 - Other topics may be chosen, but students will need to have easy access to the contextual data files for their analysis.

Resources:

- Room Data Collection Project
- Data Analysis and Research Questions

Other Resources:

- The Human Face of Big Data. Rick Smolan, Jennifer Erwitte. November 20, 2012 | ISBN-10: 1454908270 | ISBN-13: 978-1454908272. AgainstAllOdds.com.
- Learning about Statistics videos <http://www.neok12.com/Statistics.htm>
- Basic Statistics and Probability <http://www.shmoop.com/basic-statistics-probability/designing-study.html>

Data Analysis and Research Questions

Topic 1: School Safety and Bullying

File: schoolSafety_dataset .csv

Introductory Resources:

- Bullying facts from the National Assoc. of School Psychologists:
http://www.nasponline.org/resources/factsheets/bullying_fs.aspx
- More facts about bullying: <http://facts.randomhistory.com/facts-about-bullying.html>
- Youth Online: View state data. Select a state, click “GO”. Filter the data by selecting the Health Topic “Unintentional Injuries and Violence.” View the Table or Graph by selecting the appropriate tab: <http://apps.nccd.cdc.gov/YouthOnline/App/Default.aspx> . Focus on the questions related to bullying.
- See the state-by-state grades for anti-bullying laws: <http://www.bullypolice.org/>
- YouTube video: You Are You <http://www.youtube.com/watch?v=l8ovOA7VhFo>
CAUTION: Preview before showing to evaluate suitability for your classroom.

Data analysis

See **DataSet_Quest-Ans_Codes.doc** for survey questions and answer codes.

1. What percentage of students has been bullied at school?
Electronically?
2. What is the range of days that students felt threatened at school?
What is the mean?
3. How many students who have felt unsafe at school have talked to a teacher about it?
4. Who (age, gender) is most likely to get help with bullying?

Possible research questions

1. How does experiencing bullying impact one’s willingness to help others being bullied?
2. What support strategies could be used to help victims of bullying?
3. How does intervening with bullies impact the bullies’ behavior?
4. Does the number of close friends a person have, impact the likelihood of being a victim?
5. What strategies defuse bullying?
6. What motivates bullies? Are there “trigger events?”

Topic 2: Exercise and Health

File: exerciseHealth_dataset.

Introductory Resources:

- FitWebMD: <http://fit.webmd.com/teen/move/article/exercise-personality>
 - Play Bubble Rubble with Chicken Dawg
 - What’s your Exercise Personality?
 - Go From Couch to 5K in just 8 Weeks
- Physical activity recommendations:
<http://www.cdc.gov/physicalactivity/everyone/guidelines/children.html>

- Youth Online: View state data. Select a state, click “GO”. Filter the data by selecting the Health Topic “Physical Activity.” View the Table or Graph by selecting the appropriate tab. <http://nccd.cdc.gov/YouthOnline/>
- Rewards and benefits of exercise: http://kidshealth.org/teen/food_fitness/exercise/exercise_wise.html#cat20133

Data analysis

See **DataSet_Quest-Ans_Codes.doc** for survey questions and answer codes.

1. How many students report being active for 60 minutes on 5 or more of the past 7 days?
2. Which students (age, gender) are most likely to get 8+ hours of sleep?
3. How is eating breakfast related to one’s description of their weight?
4. Rank the popularity/frequency of the various physical activities?
5. Compare the frequency of playing video games with being active 60 minutes or more on 5+ of the past 7 days.

Possible research questions

1. Does getting more sleep improve students’ activity levels?
2. What exercise activities do teenagers engage in that are likely to be continued for more than 1 year?
3. How does a family’s activity level impact the student’s activity level?
4. What role do friends have in maintaining consistent activity levels?
5. How does participation in a sports team impact diet?
6. Are active students healthier than inactive students? Are they happier?

Topic 3: Teens, Social Media, and Online Behavior

File: Mobilize_MediaBehavior_2011.csv

Introductory Resources:

- Video Social Media Helps Teens See Other’s Needs: <http://www.webpronews.com/social-media-teens-others-needs-2012-02>
- Social Media, Social Life – What Do Teens Think About Twitter And Facebook? [INFOGRAPHIC] http://www.mediabistro.com/alltwitter/social-media-teens_b24749
- Video Tagged (cyber bullying) Teens and Social Media <http://planningwithkids.com/2011/09/30/teens-and-social-media/> Includes teaching resources: <http://www.cybersmart.gov.au/tagged/teenagers.htm>
- An Analytical Take on Youth, Social Networking, and Web 2.0: A Few Moments with Amanda Lenhart <http://dmlcentral.net/newsletter/05/2012/analytical-take-youth-social-networking-and-web-20-few-moments-amanda-lenhart>
- [Teens, kindness and cruelty on social network sites:](http://www.pewinternet.org/2011/11/09/teens-kindness-and-cruelty-on-social-network-sites/) <http://www.pewinternet.org/2011/11/09/teens-kindness-and-cruelty-on-social-network-sites/>

<p>Data analysis</p> <p>See DataSet_Quest-Ans_Codes.doc for survey questions and answer codes.</p> <ol style="list-style-type: none"> 1. Which privacy setting is most common? 2. What percentage of students has posted something they thought might reflect badly in the future? 3. What is the average number of text messages students send and receive? 4. Which gender has the most positive view of the kindness of others online? 5. How many students have been treated cruelly online? 	<p>Possible research questions</p> <ol style="list-style-type: none"> 1. What is the relationship between students who are most careful with their online presence and their kindness factor? 2. Are students who spend lots of time online more or less likely to join other organizations? 3. What is the relationship between the time spent online in social settings and time spent socializing offline? 4. What is the relationship between time spent on social media and academic success? 5. Which is more enduring, online relationships or physical relationships? 6. How important is honesty in online relationships? Are some online relationships more prone to honesty than others?
<p>Topic 4: Teens, Video Games, and Civic Engagement</p> <p>File: Mobilize_GamesCivic_2008_xformed.csv</p>	
<p>Introductory Resources:</p> <ul style="list-style-type: none"> • 22 Charts & Graphs on Video Games & Youth Violence http://videogames.procon.org/view.resource.php?resourceID=003627 • How much do you know about video games http://www.esrb.org/about/video-game-industry-statistics.jsp • Video game stats http://www.onlineeducation.net/videogame • Video: Do video games cause violence in kids? Interesting facts and data • http://www.youtube.com/watch?v=vVjUr2zXtb4 	
<p>Data analysis</p> <p>See DataSet_Quest-Ans_Codes.doc for survey questions and answer codes.</p> <ol style="list-style-type: none"> 1 How many engage in civic activities? 2. What percentage of students believe it is important to be involved in improving their community? 3. What is the most popular game type (fighting, puzzles, races, etc.)? 4. Create a word cloud of favorite games. Which names appear most 	<p>Possible research questions</p> <ol style="list-style-type: none"> 1. Are social-media-using students more or less socially aware of the needs of others? 2. Which social media venues foster the greatest tendency to get involved in charitable activities? 3. What types of games seem to promote more civic engagement? 4. How does parental involvement in the students' game playing influence the choice of games students play? 5. What is the relationship between accepting people from different backgrounds and playing video games? 6. What is the relationship between playing video games and crime or other anti-social behavior at school or in the

often? 5. What is the relationship between the frequency someone plays games and how often he/she help or guide others?	community?
--	------------

Instructional Days: 6-7

Topic Description: Groups for the final project are created and roles and responsibilities for group members are assigned. Students develop a list of possible questions and method for data collection to use for the final project.

Objectives:

The student will be able to:

- Draw conclusions about the data set selected for the class-topic
- Identify issues related to the topic or aspects of the topic of particular interest to the groups
- Identify research questions or describe the stories or discoveries they plan to uncover through the analysis of gathered data
- Identify the specific variables needed to analyze the data
- Develop a method for data collection
- Collect data
- Evaluate initial survey questions to determine if they were easily understood and elicited the data they had planned on gathering
- Predict how useful the data would be in answering the research question if the survey items were scaled to the whole class
- Revise the survey questions to better fit the goal of answering the research questions

Outline of the Lesson:

- Create groups and assign roles (10 minutes)
- Identify research questions or describe stories or discoveries to uncover through the analysis of data gathered on the final project topic (15 minutes)
- Identify the specific variables needed to analyze the data (15 minutes)
- Data collection method for the final project (15 minutes)
- Collect data for homework
- Evaluate the effectiveness of the initial survey questions (15 minutes)
- Predict how useful the data will be in answering the research questions if the survey items were scaled to the whole class (10 minutes)
- Revise the survey questions (30 minutes)

Student Activities:

- Groups discuss roles and responsibilities.
- Groups identify research questions or describe the stories or discoveries that can be discovered through gathered data.
- Groups list the specific variables needed to analyze the data.
- Groups develop data collection survey questions.
- Students collect data for homework.
- Groups evaluate the effectiveness of the initial survey questions.

- Predict how useful the data would be in answering the research questions if the survey items were scaled to the whole class.
- Groups revise the survey questions to better fit the goal of answering the research questions.

Teaching/Learning Strategies:

- Group roles and responsibilities
 - Explain to students that everyone in the group is accountable for the work on the various stages and that they need to alternate roles. They also need to come to consensus on decisions related to data collection, questions for analysis, etc.
 - Provide students with dates for the data check-ins/analysis with their own data. (These may need to be revised if sections take longer than anticipated.) Students should plan to collect some data prior to Day 7 and then before day 8 data check-in to use as a base for discussion.
- Identify research questions or describe the stories or discoveries that can be made through analysis of data. It is important that every group member contributes a few ideas before the group refines the ideas and selects, by consensus, one or two research questions.
- Groups identify the specific variables needed to analyze the data in order to answer their research questions.
- Review the traits of “good” questions for effective surveys that will reveal the information needed to answer research questions. Use content from <http://www.shmoop.com/basic-statistics-probability/designing-study.html>.
- Guide each group to write 4–5 survey questions with the variables that they believe will help them answer their research questions. Refer to their work on Day 5.
- For homework, have students collect data based on their survey questions.
- Evaluate the survey questions to determine if they were easily understood and if they elicited the data they had planned on gathering.
 - In their groups, have students review the latest set of data collected and respond to the following questions. (Note: These are the same questions that were considered during the debrief of the Room Data Collection Project.)
 - What do wish we’d thought about before starting to collect the data?
 - Do you think it’s possible to think of *everything* ahead of time? If not, what steps can you take to make sure that you can recover from missteps?
 - Any other data you wish we had collected from the outset?
 - Anything you wish we’d specified more clearly before collecting data?
 - When designing a method for collecting a large set of data using multiple sources, what kinds of things do you need to think about?
- Have each group share their responses and discuss as a class.
 - Prompt them with questions such these:
 - Why do you think this question worked well to elicit the response you were after?
 - What made you think that this question (that wasn’t effective) would elicit the response you were after? How can you modify it to be more effective?
- Predict how useful the data would be in answering the research questions if the survey items were scaled to the whole class.

- What conclusions, if any, can be drawn from the survey data?
 - What do you predict you will learn when this project is scaled to the entire class? What additional insights might you learn if it was scaled to the entire school? All students in the state? In the US? In the world?
 - What actions might result from what is learned?
- Revise the questions to better fit the goal of answering the research questions. The revised questions will be combined on Day 8 with the questions of all of the classroom teams for the final survey project.

Resources

- Data Analysis and Research Questions
- <http://www.shmoop.com/basic-statistics-probability/designing-study.htm>

Instructional Day: 8

Topic Description: Data collection check-in.

Objectives:

The students will be able to:

- Identify issues related to the data collection process
- Explain aggregation of data

Outline of the Lesson:

- Data check-in (10 minutes)
- Journal Entry (5 minutes)
- Data collection issues and aggregation of data (40 minutes)

Student Activities:

- Groups discuss the data collected to date.
- Complete journal entry.
- Participate in discussion about data collection issues and aggregation of data.

Teaching/Learning Strategies:

- Data check-in
 - Have students upload their individual data and share with the rest of their group.
- Journal Entry: Consider the data that your group collected. What issues did you have with collecting the data?
- Discuss issues that have arisen with data collection.
 - Have each group describe the data they have collected to date.
 - Clarify any misconceptions about what they should be collecting.
 - Did they understand the various prompts and possible responses?
 - How many entries did they collect? How does that compare with other groups?
 - Discuss aggregation of their data.
 - Why is it important for each member of the group to collect data?
 - Why will we want to pool the data from all of the groups at the end of the unit even if each group is working on a different set of research questions?
 - NOTE—modify based on whether there is one or more projects/class
- Build consensus to select, condense, and revise the questions into the class survey question set, if all students are working with the same overall topic.
 - There are several strategies that can be used to complete this task as efficiently as possible through consensus. The goal is to have a set of questions that can be combined into a single survey that will enable all teams to analyze the variables they need to answer their research questions. One such strategy is

- Each group writes each of their prompts/questions, possible responses for multiple or single choice prompts, and the variable name on 8.5 X 11 inch sheets of paper (one prompt and variable per sheet).
 - Group the members of 2 or 3 groups together so that they can spread the pieces of paper such that each group member can see all of the prompts.
 - A spokesperson for each group should share the team's research questions and state why each of the variables/prompts is important to their research.
 - The groups then look for duplication in questions, rewriting duplicates into a final question that all members of the groups can agree upon. The teacher may need to intervene in disagreements and guide consensus building.
 - Repeat the process by combining these refined questions from all groups. The spokesperson of each group will be the active participant in working to build consensus among all teams. The prompts in round two can be displayed on the wall. Only the spokesperson for each group can talk.
 - If the class is smaller or there are several campaign topics, the process can be compressed into just one round of consensus building.
- Each group convenes to verify that their prompts and variables are represented in the list of final prompts.
 - Remind students of check-in dates and periodically remind students of their survey and sharing responsibilities.

Resources:

- No additional resources needed

Instructional Days: 9-12

Topic Description: The basic features of loading and saving files, sorting and creating subsets are explored. Maps are created by using the latitude and longitude of a location and then maps of points and bubble charts are created from a file of data.

Objectives:

The students will be able to:

- Translate addresses into latitude/longitude
- Sort files of data
- Create subsets of data
- Read location data from a file and plot points on maps
- Create bubble plots

Outline of the Lesson:

- Journal Entry (5 minutes)
- Describing location (30 minutes)
- Exploring LA Bike Data (30 minutes)
- LA Bike Activity (45 minutes)
- LA Bus Stops Activity (45 minutes)
- Bubble charts (20 minutes)
- Bubble Charts Activity (45 minutes)

Student Activities:

- Complete journal entry.
- Participate in discussion of location and LA Bike Data.
- Complete LA Bike Activity.
- Complete LA Bus Stops Activity.
- Participate in discussion of bubble charts.
- Complete Bubble Chart Activity.

Teaching/Learning Strategies:

- Journal Entry: Consider the data that you have been collecting. How might seeing the data on a map help you analyze it?
- Install data analysis tool
 - You may want to do this installation yourself before the class to save time.
- Describing location
 - Use Walking and Biking in LA as an introduction to the LA bike data and describing location with latitude and longitude. (You may want to share a version of this resource with students.)

- Take the opportunity to point out that this is a campaign similar to what they will be doing for the final project—people like them concerned about a topic, collecting data to inform their cause.
- Load the Google Map with the locations of the 56 intersections for the survey.
- Ask questions about these locations: Are any of these locations in your neighborhood? Near the school? Near your home? Other questions that may be of interest.
- Go to <http://www.latlong.net> and demonstrate how to translate a place on the map to latitude and longitude.
 - Type in an address. Use the example in Walking and Biking in LA Teacher Resource or another that you choose
 - The appropriate numbers are under the map inside the parenthesis next to WKT: POINT(-118.2796304, 34.0916803). That means the location is 34.0916803 north of the equator and 118.2796304 west of the Prime Meridian. (If the example provided is used.) Note: Latitude is often referred to first, but the coordinates of the point are (longitude, latitude).
 - Have students try finding locations for their house and/or school.
 - Point out that the system of longitude and latitude will allow them to draw spatial objects such as points (house, school). They can think of longitude as the x-direction (it runs along the equator from east to west) and latitude as the y-direction (it runs from the North Pole to the South Pole along the Prime Meridian) when making a plot from these coordinates.
- Exploring LA Bike data
 - Note: Throughout this unit, use a variety of strategies as you introduce new features of analysis. Many of the same types of strategies that were used to introduce new features of Scratch can be effective. Such strategies include, but are not limited to, student demonstrations, prediction of behavior, driver/navigator pairs, jigsawing of several new features.
 - Load the labike data file.
 - Point out that this is what they will be doing with the data they collect.
 - Point out the following features as you discuss the layout of the file (You can use Exploring LACBC as a reference.):
 - Header, number of rows, categories
 - Note what you see in the file (many of rows of data).
 - The first line is the header and describes the names of each variable or column.
 - Each **row** refers to a different intersection, and so there are 38 intersections represented in the data set. Each **column** refers to the various data that were collected about the intersection.
 - Navigate through the survey to show the variables in the data
 - set:
 - "**name**" is the location of the intersection,
 - "**longitude**" is the longitude of the location
 - "**latitude**" is the latitude of the location

- "type" is the type of bike transportation available at the intersection (bike lane, bike path, bike route, none)
- "bike_count_pm" is the evening count of bikes
- "ped_count_pm" is the evening count of pedestrians
- Demo how to obtain a table of frequencies for type.
 - Note that 20 of the intersections have nothing.
 - There will be more discussion of frequencies later in the unit, but note that this file is small enough that the counting could be done by hand (as with the data collected in Unit 2), but that later data sets will be much larger.
- Demo how to sort by bike count and pedestrian count.
 - Which intersections have the most bike traffic/ pedestrian traffic? Are they the same?
- Demo how to create a subset of locations where the bike count is greater than or equal to the pedestrian count.
 - Ask students what other subsets might be interesting to create (e.g., locations with bike routes).
 - Have students create a few subsets of their own and list the questions they might want to ask about those subsets.
- Demo how to plot the intersections on a map. Include a title, axes and background. Also demo the various sizes and shapes of points and how to zoom.
 - Before each plot feature is added, ask students questions that will guide them to the need for the feature.
 - Ask students questions about the plot such as: Are there any outliers? Are there clusters of points? Does the plot match the table?
- Throughout this section and the other analysis sections, continue to pose questions to prompt students to reflect on their analysis and justify their conclusions.
- LA Bike Activity
 - Have students complete the LA Bike Activity on their own. Circulate the room and answer questions.
 - Allow sufficient time at the end of this part to have students share their responses with their group members and lead a discussion to ensure that they understand each of the features of the data analysis tool discussed so far.
- LA Bus Stops Activity
 - Have students complete the LA Bus Stops Activity on their own. Circulate the room and answer questions.
 - Allow sufficient time at the end of this part to have students share their responses with their group members and lead a discussion about the data analysis and the patterns they have described. In particular, have groups share and discuss answers to questions 10-13 as these require a higher level of analysis.
- Bubble charts with LA Bike Count Data

- Describe bubble charts. (You can use Bubble Charts as a resource.)
- Ask questions such as: What is being described when longitude and latitude is plotted on a map? Is there a way to distinguish counts of pedestrians and bicyclists?
- Demo how to create a bubble chart with the pedestrian counts in the labike file.
- Demo how to change the size and color of the bubbles.
- Bubble Chart Activity
 - Have students complete the Bubble Charts Activity on their own. Circulate the room and answer questions.
 - Allow sufficient time at the end of this part to ask students for their responses and lead a discussion to ensure that they understand each of the features of data analysis discussed so far. Challenge them to ask additional questions. Identify and clarify misconceptions.

Resources:

- Walking and Biking in LA
- Exploring LACBC
- LA Bike Activity
- LA Bus Stops Activity
- Bubble Charts
- Bubble Charts Activity

Walking and Biking in LA

**Survey Description

The data that will be considered first in this lesson were collected in September of 2009 by the Los Angeles County Bicycle Coalition (LACBC, <http://la-bike.org>), a non-profit organization that works to "make the entire L.A. region a safe and enjoyable place to ride." For two days in late September, the LACBC recruited volunteers to count the number of bicyclists and pedestrians that pass 56 different intersections within Los Angeles County. Some of the survey locations were chosen because they are known to be popular with cyclists and pedestrians, others because they are near locations where a traffic-related change is about to take place, and still others because they are the site of a large number of bike accidents each year.

LACBC volunteers surveyed each location in the morning (7:00-9:30 am) and evening (4:00-6:30 pm) on Tuesday September 22 and Wednesday September 23 of 2009. Data were also collected on Saturday the 26th, but will not be considered here. The volunteers produced a report summarizing their findings (<http://lacbc.files.wordpress.com/2010/06/labikecountreport.pdf>).

** Describing location

When someone asks you to describe your current location, you might respond informally by saying you are in class or at school. Friends and family will know where that is, but if a relative were visiting from out of town and were unfamiliar with the area, the street address for school and maybe a nearby intersection would be necessary. These descriptions are excellent for looking up a location on a map or for walking, biking, or driving somewhere new. Roads and intersections and street addresses create a network that we regularly navigate. This network may change as old roads and buildings are replaced by others. Also, some of the important places in our lives do not have a street address (like the peak of a mountain or a hiking trail in the Santa Monica Mountains). Finally, in order to draw a map of LA, it would be helpful to be able to specify positions in a more consistent way, tracking a road as it turns a corner or veers to the left. For all of these reasons, there is a need to associate positions with a fixed set of "coordinates" on the earth.

One of the most popular such coordinate systems involves specifying a point's latitude and longitude. These are two numbers that represent angles (in degrees) from the center of the earth to a point on its surface. Latitudes are angles from north to south—in this case the North Pole is assigned a value of 180 degrees, the equator is at 0 degrees and the South Pole is at -180 degrees. Longitudes are angles from east to west with 0 occurring at the Prime Meridian, a line running from the North Pole to the South Pole and crossing Greenwich, England. (Greenwich is also used in defining Greenwich Mean Time or GMT.) A description of longitude and latitude can be found at the following url.

http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Georeferencing_and_coordinate_systems

These two ways of describing a location (an address versus latitude-longitude) can be compared. The LACBC produced a Google Map with the locations of the 56 intersections for their survey—this was actually a recruitment tool and there are references to shifts that are needed for people to lend a hand.

<http://tinyurl.com/LABikeECS>

To translate the location on the map for the intersection of Sunset and Hyperion into latitude and longitude, a service like Get Lat Lon (<http://www.getlatlon.com/>) can be used.

In the box at the top of the page enter

Sunset & Hyperion, Los Angeles, CA

and the map should center on the right intersection. The latitude and longitude of the point will appear at the bottom of the page. In this case the latitude and longitude of the point is 34.0916803, -118.2796304—that is, 34.0916803 north of the equator and 118.2796304 west of the Prime Meridian.

The system of longitude and latitude allows people to draw spatial objects like points (house, school) or lines (a street) or shapes (the grounds of a high school, a park). Longitude can be thought of as the x-direction (it runs along the equator from east to west) and latitude as the y-direction (it runs from the North Pole to the South Pole along the Prime Meridian) when making a plot from these coordinates.

Exploring LACBC

Load the labike data file.

It should appear similar to the table below.

name	longitude	latitude	type	bike_count_pm	ped_count_pm
1st & Alameda	-118.238125	34.049175	none	62	241
4th & Wilton	-118.313441	34.06713	bike route	48	87
7th & Figueroa	-118.259883	34.049388	none	216	1979
8th & La Brea	-118.344641	34.060446	none	72	272
9th & Pacific	-118.287306	33.735118	none	58	160

The first line of the file is known as a header and describes the names of each variable or column (e.g., "name", "longitude", and "latitude"). Each row refers to a different location at which volunteers counted the bike and pedestrian counts during the evening rush hour. ("Objects" in this table are positions in Los Angeles and the "variables" measured for each include the name, the longitude and latitude, and the counts of pedestrians and cyclists.)

The data file "labike" has 38 rows, each referring to a different location. The first column is the name of the locations, similar to the list in the LACBC Google Map. (The data here reproduce Table 14 from LACBC's report and there are only 38 of the 56 locations included.) The next two columns give the positions' longitude and latitude. These coordinates can be used to place the locations on a map. The fourth column describes the type of bike transportation available at the intersection (a bike lane, a bike path, a bike route or nothing) and the last two columns represent the evening counts of bikes (column 5) and pedestrians (column 6) crossing the intersection.

A variety of operations can be performed on the data. (e.g., create a table that shows the frequency of type, sort by bike_count_pm, sort by ped_count_pm, create a subset of locations where the bike count is greater than the pedestrian count ($\text{bike_count_pm} \geq \text{ped_count_pm}$) .

There are several kinds of spatial data. Their structure is best described by their look—that is, positions or points on the map (a house, the Sandstone Peak in the Santa Monica Mountains); points that are connected to form paths or lines (the route of a walk to school or the driving route to a friend's house); and areas or regions (the footprint of a school's buildings or the area covered by Los Angeles County). Points, lines and regions are basic spatial structures that are used for computation.

In the case of the LA Bike Count data, there are intersections where survey takers stood (points). The transportation system in Los Angeles can also be consulted for bus routes (lines), and the U.S. Census Bureau can provide demographic data about people living in different Census blocks (small geographic areas or regions).

Since the data set includes the longitude and latitude in columns 2 and 3, the intersections can be plotted on a map.

LA Bike Activity

Load the labike data.

1. Create a subset of the locations with no special routes for bikes.
 - How many locations are in the subset?
 - Sort by `bike_count_pm`. Which intersection has the greatest count? Which has the least count?
 - Plot this subset on a map. Include a title, axes and a background.
2. Create a subset of the locations with special routes for bikes.
 - Use a different color and shape to plot this subset on the same map.
 - Describe any patterns you see.

LA Bus Stops Activity

Load the bus_stops data file.

1. What are the variables in this survey?
2. Form a frequency table to see the number of stops along each street. Which street has the most stops? What might be a reason for this?
3. How many total stops are there?
4. Look at the data for the 6000th row—a bus stop on Sunset Boulevard at Anita Avenue. Go to <http://getlatlon.com> and type in Sunset & Anita, Los Angeles, CA and check that the longitude and latitude listed in the data file are the same as from Get Lat Lon. What do you notice? Why might this be the case?
5. Create a subset of the bus stops that are along Sunset or Vermont. How many stops are there?
6. Create a plot of bus stops that are along Sunset or Vermont. Include a title, axes, and a background. Describe what you see in the plot.
7. Create a plot of bus stops that are along Myrtle or Mulholland. Include a title, axes, and a background. Describe what you see in the plot. How does this compare to the plot of Sunset or Vermont.
8. Create a plot of bus stops that are along Gayley or Hilgard. Include a title, axes, and a background. Describe what you see in the plot. How does this compare to the previous plots?
9. Create a plot of bus stops that are along a few streets in your neighborhood. Include a title, axes, and a background. Describe what you see in the plot. Why might this data be useful for someone to have?
10. What is an advantage to plotting the data on a map instead of just looking at the latitude/longitude numbers?
11. What would happen to the map if you had less data in the file? More data in the file? How would that affect your interpretation of the map?
12. If you were trying to make a case that you needed more bus stops in your neighborhood would it be enough to show that the count of bus stops is less than those along Sunset? Explain your answer.
13. How could you use what you learned about plotting points on a map with the data you have collected?

Bubble Charts

The points at which LA Bike volunteers stood and the position of bus stops exhibit the geometry of these things. However, spatial objects can have other data associated with them. The LA Bike Counts are associated with counts of pedestrians and bicyclists. When the intersections at which the LA Bike volunteers stood were plotted there was nothing that could be determined about the number of bikers or pedestrians.

A bubble chart uses numerical values to scale the diameter of circles located at a given spatial location. Consider the pedestrian totals from the LA Bike data.

In a bubble plot of the pedestrian counts each intersection where volunteers collected data is the center of a circle—the larger the circle, the greater the number of pedestrians counted there. If this plot were drawn by hand, a number of choices would need to be made. First, the size of the circles relative to each other is fixed by the data. If a volunteer at one intersection saw twice as many pedestrians as another volunteer saw at a different intersection, the first circle should be twice as big as the second. The relationship between the circles and the map, over which they are plotted, however, is not fixed and can be changed (again, assuming the relative sizes of the circles remains the same).

Bubble Chart Activity

Load the labike data file.

1. Create a bubble chart of the pedestrian counts.
2. Create a bubble chart of bike counts and add it to the pedestrian counts. What happens to the chart?
3. Change the color for bike counts. Describe what you see now.
4. What else might you change to get an even clearer visual picture of bike and pedestrian counts? Try these ideas. Explain how this changes the chart.
5. Based on your graph, what questions might you ask?
6. Try zooming in on a part of the graph. Describe what you see.
7. Does your graph make sense based on the counts in the table? Explain why or why not.
8. Create a subset of all locations that have a special route for bikes. Create a bubble chart with the counts for the subset. Describe what you see.
9. Create a subset of all locations that have no special route for bikes. Add a bubble chart of the counts for the subset to the previous chart in a different color. Describe what you see now. What conclusions might you draw? Justify your answer.
10. Create another pair of subsets that are of interest to you. Create a bubble chart. Describe the story you see.

Instructional Day: 13

Topic Description: In this lesson students use the data they have collected and additional contextual data sets to do spatial analysis for use in the final project.

Objectives:

The students will be able to:

- Analyze the data they have collected using spatial analysis techniques

Outline of the Lesson:

- Create spatial plots with student generated and contextual data sets (55 minutes)

Student Activities:

- Groups create spatial plots and use spatial analysis techniques with the data they have collected and additional contextual data sets.

Teaching/Learning Strategies:

- Students work in their groups to analyze the spatial aspects of the data they have collected pulling in the additional contextual data sets as appropriate.

Resources:

- Student data
- Additional contextual data sets

Instructional Days: 14-16

Topic Description: Bar plots and the differences between categorical and continuous data are explored. Mosaic plots are introduced as a vehicle for comparing categorical data and looking for trends in data.

Objectives:

The students will be able to:

- Read and interpret a bar plot
- Create bar plots
- Differentiate between categorical and continuous data
- Compare two categorical sources with mosaic plots
- Look for trends by analyzing various plots

Outline of the Lesson:

- Journal Entry (5 minutes)
- Birth Month Bar Plot (15 minutes)
- Experiment with bar plot commands (30 minutes)
- Public Agenda Bar Plot Activity (45 minutes)
- Journal Entry (5 minutes)
- Public Agenda data and mosaic plots (60 minutes)
- Wrap up Question (5 minutes)

Student Activities:

- Complete journal entry.
- Participate in Birth Month Bar Plot discussion.
- Experiment with bar plot commands.
- Complete Public Agenda Bar Plot Activity.
- Complete journal entry.
- Respond to questions during guided discussion.
- Complete questions in Public Agenda Data and Mosaic Plots Activity.
- Provide responses to the wrap up question and participate in discussion.

Teaching/Learning Strategies:

- Journal Entry: If everyone were going to be put in a different group based on the MONTH in which they were born, how many groups would there be? Which group do you think would have the most people?
- Birth Month Bar Plot
 - Tell students that you are going to create a bar plot (also called a bar graph or bar chart) of everyone's birth month to answer the journal question.
 - Have students help you create the skeleton of a bar plot like Sample Birth Month Bar Plot. You should end up with a similar chart, but without any counts (bars).

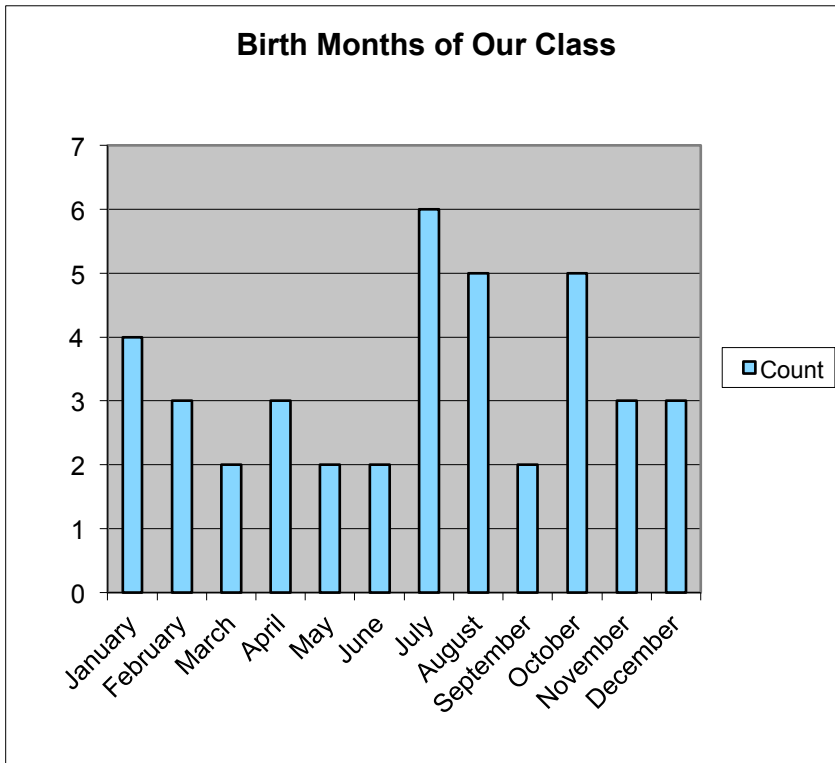
- Ask each student what their birth month is. Increase the height of the corresponding bar by one until the entire class has responded.
- The bar plot should be used for categorical data only.
- Categorical data is expressed in terms of specific category values or labels (e.g., days of the week, answers to a multiple choice survey).
- Explain to the students that if we tried to do a bar plot of every student's exact height (example of 68.901 inches), we would most likely end up with a bar plot with a bar for every student each with a height of one. This type of data is quantitative (e.g., decimal numbers).
- Public Agenda Data and Bar Plots
 - Explain that the survey data file holds data collected by a private research group called Public Agenda (www.publicagenda.org). It is a survey of high school students and their parents designed to see if both groups have the same view of what's working (or not) with our schools. The people in the survey were identified by random selection from a list of all high school students in the United States. Respondents were asked over 100 questions—the file that will be used is a small subset.
 - Have students load the survey data file. Ask questions such as: How many different students are represented? (1293) How many different questions were asked of a student. (*Survey contains four (4) of the over 100 questions.*)
 - Ask: What are the variables?
 - "year" is their year in school
 - "effort" describes how hard they are working at doing well in school
 - "homework" describes their view of the amount of homework they are getting, and
 - "grades" records how well they said they are doing in school
 - Highlight the type of each variable and possible values that are assigned to each variable.
 - Demonstrate how to create a bar plot. Point out that a bar plot is a graphical representation of the table and each bar should correspond to the count in the table.
 - Have students complete Public Agenda Bar Plot Activity individually.
 - Lead a discussion of the answers to Public Agenda Bar Plot Activity.
 - Each of the responses should generate a discussion beyond the simple solutions. Ask questions that encourage students to provide justifications for their conclusions and highlight possibilities for additional research.
- Journal Entry: Do you think there is a relationship between grades and effort? If so, what type of relationship do you think grades and effort might have?
- Public Agenda Data and Mosaic Plots
 - Reload the survey data file, if necessary.
 - Demo looking at two variables at once with mosaic plots and guide a discussion with students.
 - Note that in the previous section bar plots about grades and effort were looked at separately.
 - A good question to ask is “are the two related?” If so, how?

- Discuss the journal entry.
- Create a contingency table with data to show the relationship between the answers to the two questions.
 - Ask students to explain what the items in the table mean. For example, there are 311 students that earned an A and are trying their best to do well in school. To represent this graphically, we can use a mosaic plot.
- Demo how to create a mosaic plot to graphically compare the 2 categorical variables grade and effort.
- How to interpret the mosaic plot:
 - The wider the columns, the more responses there are in that category.
 - Point out that the labels may not line up correctly.
 - Allow students time to respond individually to questions such as the following before discussing them as a group.
 - What grade is the most common?
 - What grade is the least common?
 - Does that reflect the numbers in the table?
 - Within each column, the taller the row, the more responses there are in that category.
 - Allow students time to respond individually to questions such as the following before discussing them as a group.
 - Within those students with A's, are most of them trying their best or could they try harder?
 - Within those students with B's, are most of them trying their best or could they try harder?
 - All the sizes are proportional to the numbers in the tables. So if twice as many respond a certain way, then the height would be twice as tall in the mosaic plot.
 - Looking at the mosaic plot as a whole, is there a trend? What story does it tell?
- Have students complete the Public Agenda Data and Mosaic Plots Activity individually.
- Discuss student responses and ask probing questions that will lead to discussion of the data.
- Wrap up question: Which items in the data you are collecting are categorical?
 - Ask students to provide a response. Discuss their responses to make sure they understand the difference between categorical and quantitative data.

Resources:

- Sample Birth Month Bar Plot
- Public Agenda Bar Plot Activity
- Public Agenda Data and Mosaic Plots Activity

Sample Birth Month Bar Plot



Public Agenda Bar Plot Activity

1. Create a bar plot for effort.
 - Copy the plot to a document.
 - How does the effort of the students that responded compare?
2. Create a bar plot for homework.
 - Copy the plot to a document.
 - How much homework did most students respond that they have?
 - How do you think that compares with students at your school?
 - If you think responses about homework are different from those at your school, why do you think students in this survey might have responded as they did? How could you test your assumption?
3. Create a bar plot for grades.
 - Copy the plot to a document.
 - What grade is most common?
 - How do you think that compares with students at your school?
 - If you think grades are different from those at your school, why do you think students in this survey might have responded as they did? How could you test your assumption?

Public Agenda Data and Mosaic Plots Activity

1. Create a contingency table with effort as the row and grade as the column.
 - How does this table compare to the one with grade as the row and effort as the column?
2. Create a mosaic plot with (effort, grades)
 - What do you see in this plot?
 - Compare your plot to the one done previously. Does it tell a different story? Justify your answer with features of the plot.
3. Try making mosaic plots with three different combinations of the available variables: year, effort, homework, grades. Choose one of these other plots, describe what you see, and explain what story it tells.

Instructional Day: 17

Topic Description: In this lesson students use the data they have collected and additional contextual data sets to create and analyze bar and mosaic plots for use in the final project.

Objectives:

The students will be able to:

- Analyze the data they have collected using bar and mosaic plots

Outline of the Lesson:

- Bar and mosaic plots with student generated and contextual data sets (55 minutes)

Student Activities:

- Groups create bar and mosaic plots with the data they have collected and additional contextual data sets.

Teaching/Learning Strategies:

- Students work in their groups to analyze the data they have collected with bar and mosaic plots pulling in the additional contextual data sets as appropriate.

Resources:

- Student data
- Additional contextual data sets

Instructional Days: 18-20

Topic Description: In this lesson, the statistical measures of mean, median, minimum, and maximum are reviewed. Various ways to subset data are discussed and data is represented using box plots and histograms.

Objectives:

The students will be able to:

- Read and interpret a histogram
- Create a histogram
- Read and interpret a box plot
- Create box plots
- Explain mean, median, minimum, maximum
- Create and query subsets of a data set

Outline of the Lesson:

- Journal Entry (5 minutes)
- Sleep Histogram Activity (15 minutes)
- Quantitative Data and the CDC Survey Activity, Parts I-III (140 minutes)
- Wrap up Question (5 minutes)

Student Activities:

- Complete journal entry.
- Participate in Sleep Histogram Activity.
- Complete Parts I-III of the Quantitative Data and the CDC Survey Activity and participate in the discussions associated with the activity.
- Provide responses to the wrap up question and participate in discussion.

Teaching/Learning Strategies:

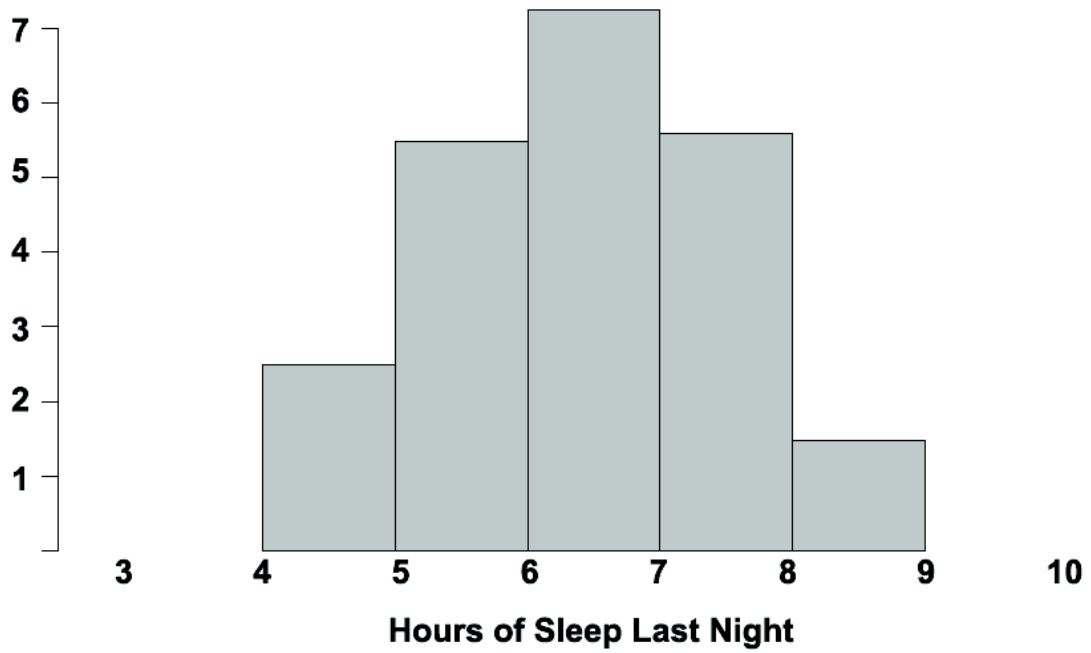
- Journal Entry: Thinking back to the bar and mosaic plots, why are we making graphical representations of our data instead of just using the numbers? Are their advantages to the different representations of data?
- Quantitative vs. Categorical revisited
 - Review the differences between these two types of data before starting on the Hours of Sleep Histogram.
- Sleep Histogram Activity
 - Create a histogram based on the amount of sleep each of the students got last night. (See Sample Hours of Sleep Histogram.)
 - Construct the bottom part of the histogram on the board, chart paper, or other display.
 - The lines should be right at the number of hours.
 - Ask each student how much sleep they got last night.

- For the most part, even if someone says they slept 8 hours that is an estimate. They really slept close to that amount (7.9 hours, etc).
- The histogram creates groupings of the data.
- To reinforce this idea, students cannot answer that they slept exactly 7 hours. They will have to choose either the 6-7 group or the 7-8 group.
 - As each student picks a group, increase the size of the corresponding column.
- Emphasize:
 - Histogram is for quantitative data and not categorical.
 - Histogram is similar to the bar plot, but notice that there is no space between columns.
 - Histogram is similar to a bar plot, but counts are within a pre-constructed range of numbers.
 - In the CDC data they are about to explore, the hours of sleep question was asked in a multiple-choice format so the resulting data is categorical rather than quantitative. Therefore, they will need to use a bar plot rather than a histogram.
- CDC Data
 - Explain to students that for the next few days they will be working with a survey of students by the Centers for Disease Control and Prevention (CDC). The CDC is part of the Department of Health and Human Services and addresses health issues facing students in the United States. (For more information: <http://www.cdc.gov/healthyyouth/yrbs/>). CDC Data Subset Description provides the survey questions that explain the columns in the data set.
 - Part I—Getting familiar with the survey
 - Students should complete this on their own as teacher circulates.
 - Before going on to Part II, discuss Part I, demo how to find frequencies and descriptive statistics, and conduct the discussion as indicated below.
 - Remind students that frequencies provides the number of responses for each option of a categorical variable. Show the frequencies for gender.
 - How many responses are there for each gender?
 - Is the sum of those the same as the total?
 - Point out N/A's.
 - Descriptives provide a numerical summary for a quantitative variable. You can include the minimum value (the shortest student's height), the maximum value (the tallest student's height), two measures of the "center" of the distribution, the mean (exact average) and median (if everyone was standing in order by height, the person in the middle) and the number of NA's or missing values. You can also include the 25th and 75th percentiles. Show the descriptives for height.
 - Explain that the heights are currently in meters, but that they will be converted in a later section. Discuss mean, median, maximum, and minimum. Make sure students understand what those terms mean in a general sense (They don't need to do the calculations.). In particular, ask them when one measure might be better than another. For example, the mean is more sensitive to outlying data.

- Part II—Subsets
 - Remind students how to create a subset and how to use more complicated conditions to subset by creating a subset of students that are “Female” **AND** “16 years old”.
 - Students should complete Part II of the activity on their own as teacher circulates.
 - Before going on to Part III, discuss Part II.
- Part III—Graphical Representations
 - Demo how to create a histogram of heights.
 - Discuss the plot.
 - Demo how to create a box plot of men’s heights.
 - Explain each piece of the box plot (median, middle half of people, maximum, minimum). Discuss each statistic and have students label their graph.
 - Box Plots can be placed side by side broken up by answers to a category. Demo a box plot of only the female’s weight on the left and only the male’s weights on the right. This shows a relationship between a quantitative variable (weight) and a categorical variable (gender).
 - Ask questions such as: According to the box plots, on average are females or males heavier?
 - Demo how to transform data. Show the transformation from meters to inches for height.
 - Note that the transformed variable appears at the end of the columns.
 - Students should complete Part III of the activity on their own as teacher circulates.
 - Distribute Different Types of Plots as a reference after completing all three parts so students can see the types of plots they have learned up to this point.
- Wrap up question: Which data values that you are collecting are quantitative?
 - Ask students to provide a response.

Resources:

- Sample Hours of Sleep Histogram
- CDC Data Subset Description
- Different Types of Plots
- Quantitative Data and the CDC Survey Activity

Sample Hours of Sleep Histogram

CDC Data Subset Description

Data Names	Data Type	Question
Age	categorical	12 years old or younger; 13 years old; 14 years old; 15 years old; 16 years old; 17 years old; 18 years old or older
Gender	categorical	Male, Female
Grade	categorical	
Hisp_latino	categorical	Are you Hispanic or Latino?
Race	categorical	What is your race?
Height	double	Height in meters
Weight	double	Weight in kilograms
Helmet	categorical	When you rode a bicycle in the past 12 months, how often did you wear a helmet?
Seat_belt	categorical	How often do you wear a seat belt when riding in a car driven by someone else?
Fights	categorical	During the past 12 months, how many times were you in a physical fight?
Depressed	categorical	During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities?
Days_Smoking	categorical	During the past 30 days, on how many days did you smoke cigarettes?
Describe_Weight	categorical	How do you describe your weight? (Very underweight, slightly underweight, about the right weight, slightly overweight, very overweight)
Eat_Fruit	categorical	During the past 7 days, how many times did you eat fruit? (Do not count fruit juice.)
Eat_Salad	categorical	During the past 7 days, how many times did you eat green salad?
Drink_Soda	categorical	During the past 7 days, how many times did you drink a can, bottle, or glass of soda or pop, such as Coke, Pepsi, or Sprite? (Do not include diet soda or diet pop.)
Drink_Milk	categorical	During the past 7 days, how many glasses of milk did you drink? (Include milk you drank in a glass or cup, from a carton, or with cereal. Count the half pint of milk served at school as equal to one glass.)
Days_Exercise_60	categorical	During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day? (Add up all the time you spend in any kind of physical activity that increases your heart rate and makes you breathe hard some of the time.)
Hours_TV	categorical	On an average school day, how many hours do you watch TV?
Hours_Videogame	categorical	On an average school day, how many hours do you play video or computer games or use a computer for something that is not school work? (Include activities such as Nintendo, Game Boy, PlayStation, Xbox, computer games and the Internet.)
Number_Teams	categorical	During the past 12 months, on how many sports teams did you play? (Include any teams run by your school or community groups.)
Asthma	categorical	Has a doctor or nurse ever told you that you have asthma?

Days_Exercise_30	categorical	On how many of the past 7 days did you participate in physical activity for at least 30 minute that did not make you sweat or breathe hard, such as fast walking, slow bicycling, skating, pushing a lawnmower, or mopping floors?
Days_Exercise_20	categorical	On how many of the past 7 days did you exercise or participate in physical activity for at least 20 minutes that made you sweat and breathe hard, such as basketball, soccer, running, swimming laps, fast bicycling, fast dancing, or similar aerobic activities?
Sunscreen	categorical	When you are outside for more than an hour on a sunny day, how often do you wear sunscreen with an SPF of 15 or higher?
Hours_Sleep	categorical	On an average school night, how many hours of sleep do you get?
General_Health	categorical	How do you describe your health in general?

Note: only two data types are “numerical”—height, weight

Source:

<http://www.cdc.gov/HealthyYouth/yrbs/data/index.htm>

2007_National_YRBS_Data_Users_Manual.pdf

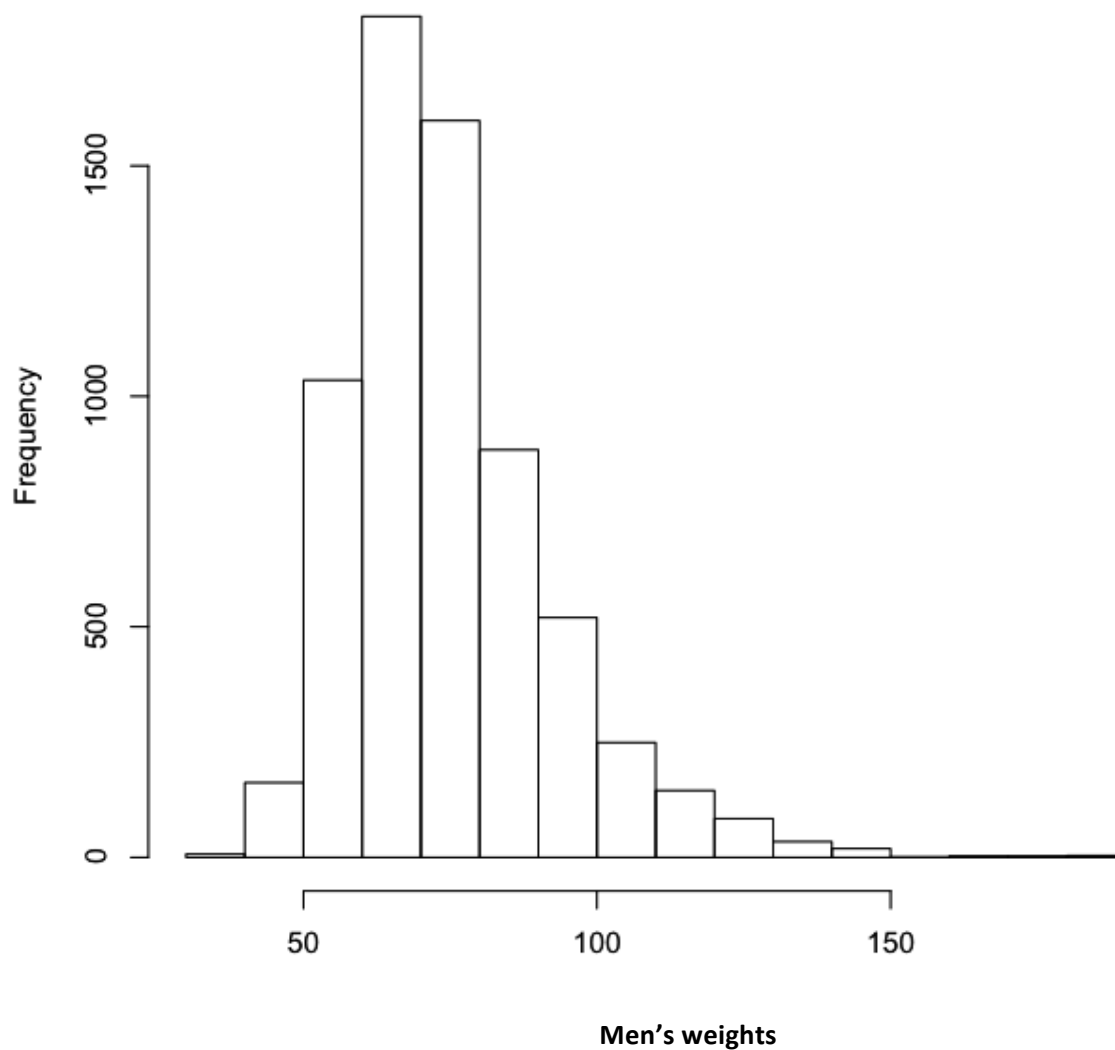
2007 National YRBS Data Users Manual

Different Types of Plots

Histogram (quantitative data)

What plot shows: How often a group of numbers occurs in a dataset. Each segment also represents its percentage of the entire data set.

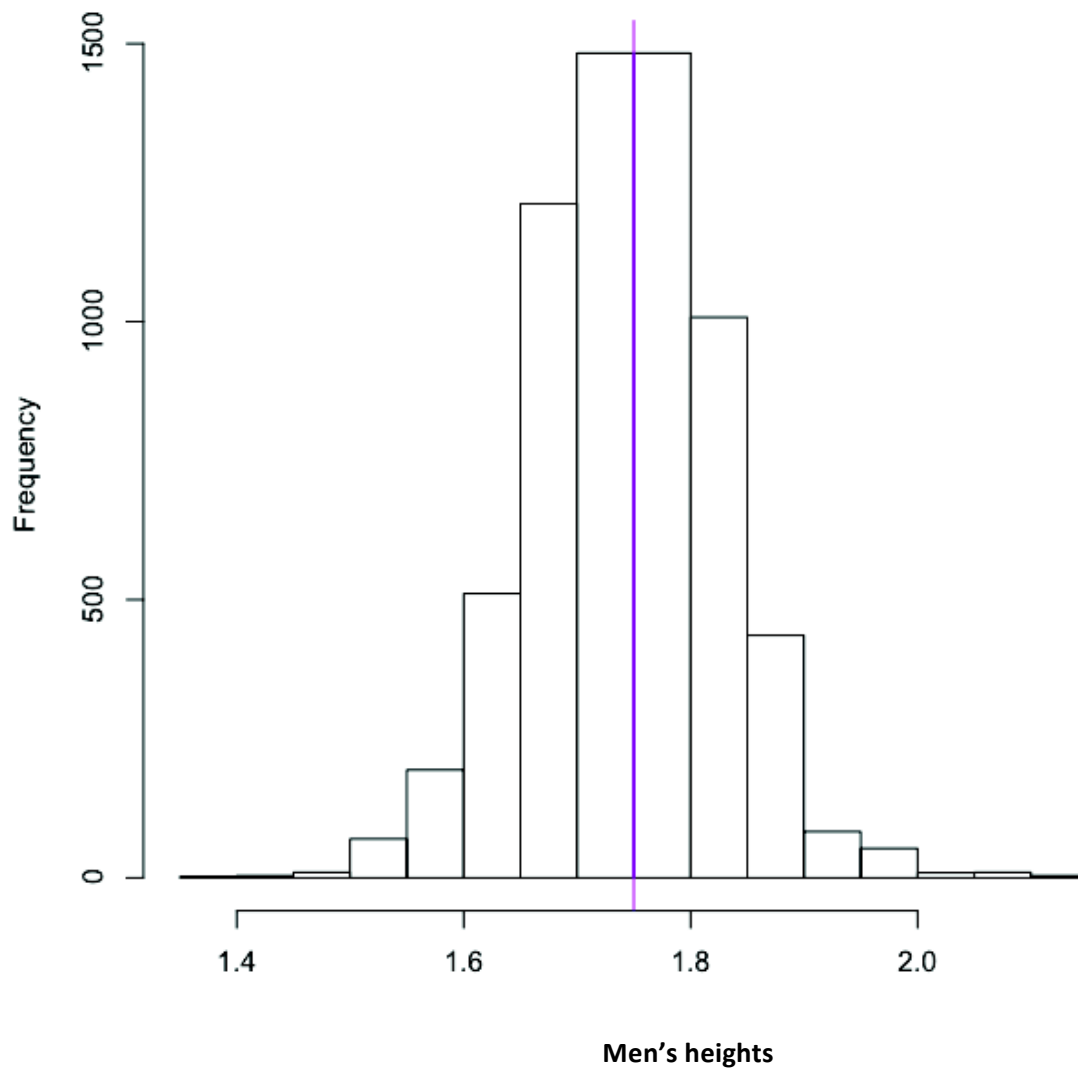
Example: Histogram of men's weights



Median (quantitative data)

What the plot shows: Median is the value that is literally in the middle—the point where half the data have larger values and half the data have smaller values.

Example: Median indicated on histogram of men's heights



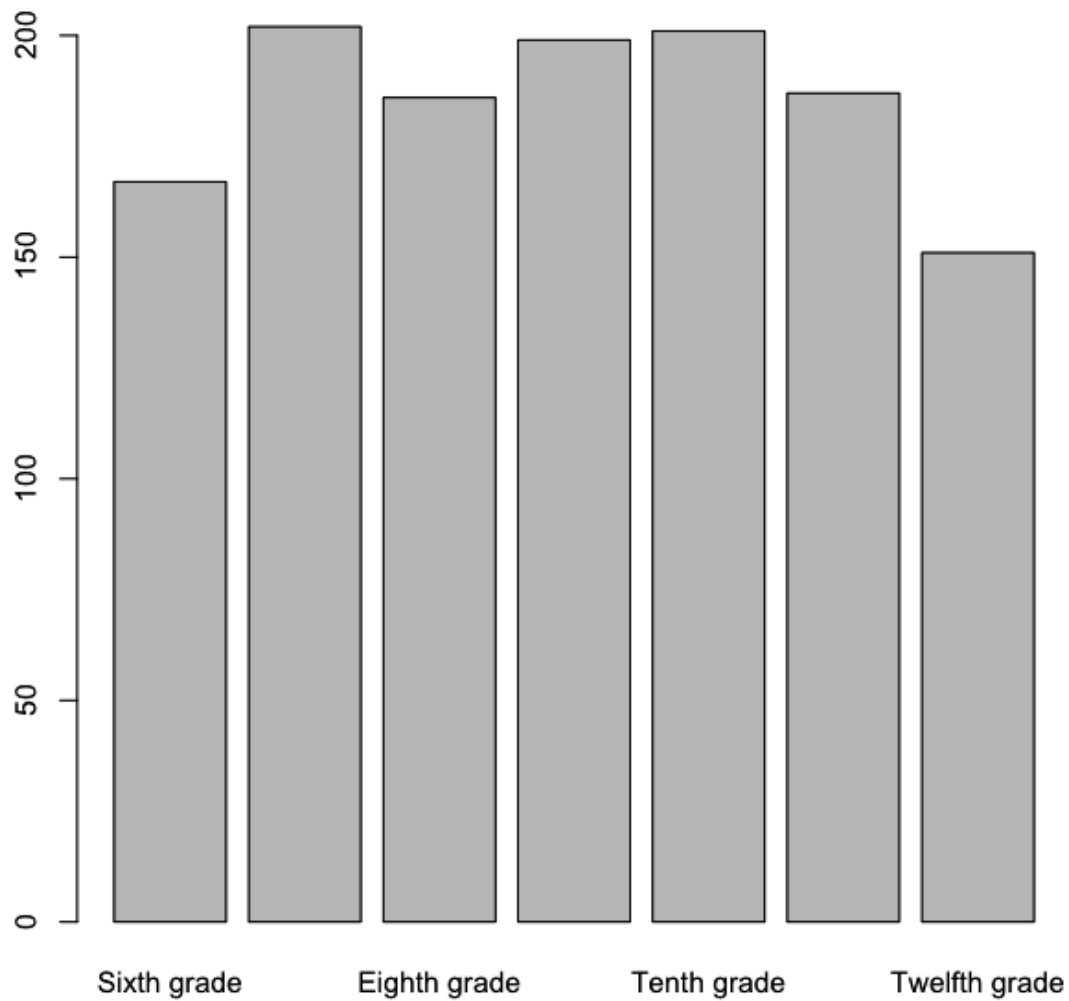
Bar Plot

What the plot shows: The number of occurrences within a given category

The data are quantitative (number of students) and categorical (grade level)

Ideal when categories are roughly the same size.

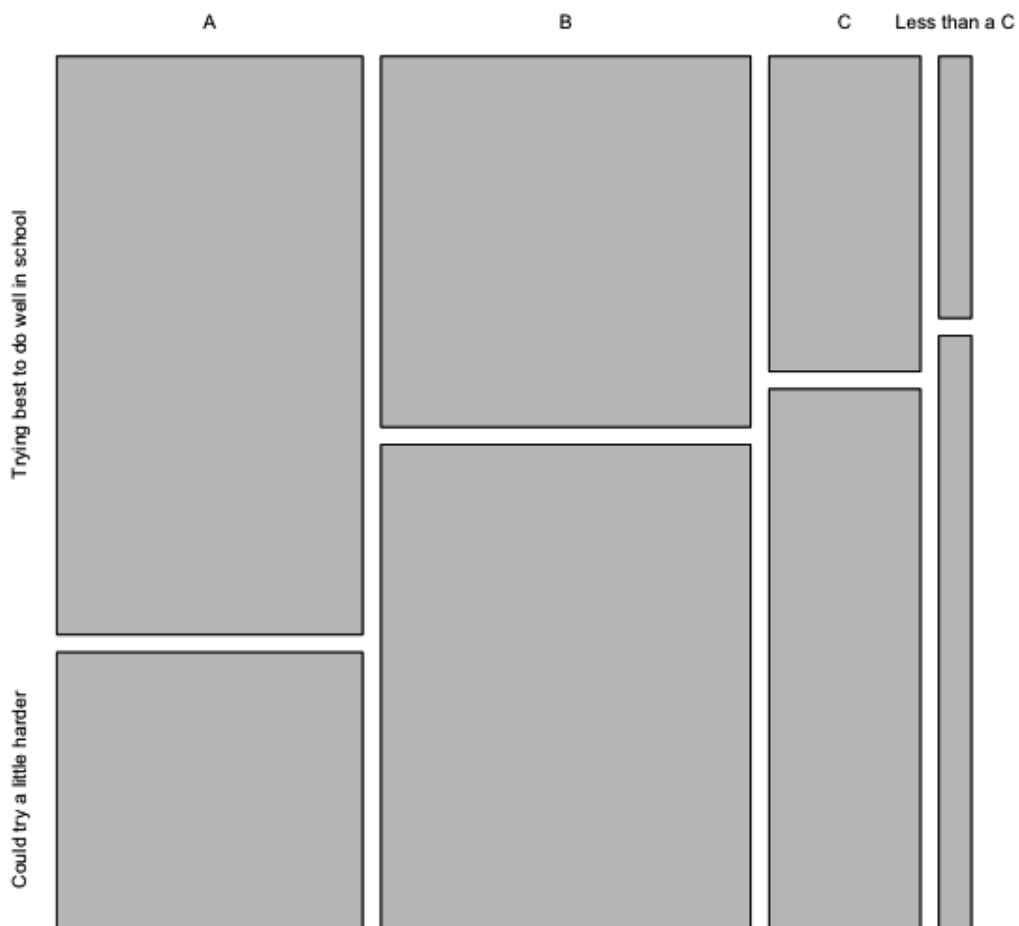
Example: Bar plot of years



Mosaic Plot – Sample 1 (categorical data)

What the plot shows: The possible relationships among categorical data.

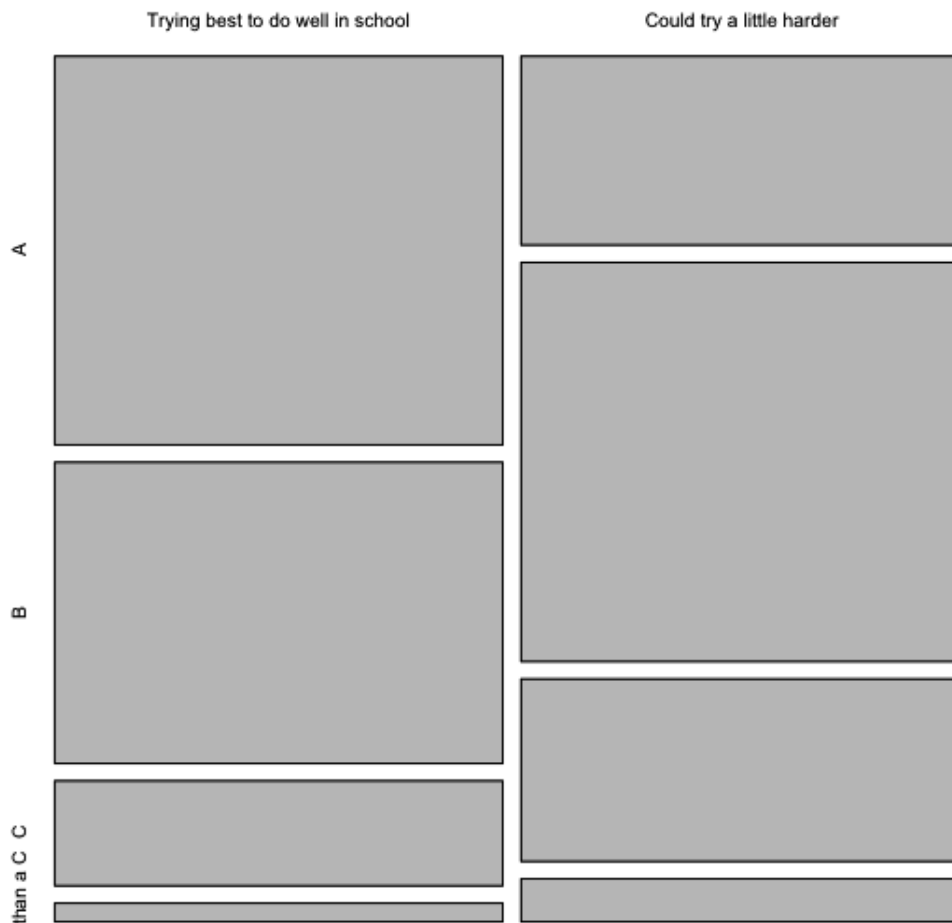
Example: Mosaic plot of grades compared to effort



Mosaic Plot – Sample 2 (categorical data)

Sometimes interchanging the relationship of the values helps clarify or present different points of view.

Example: Mosaic plot with effort compared with grades

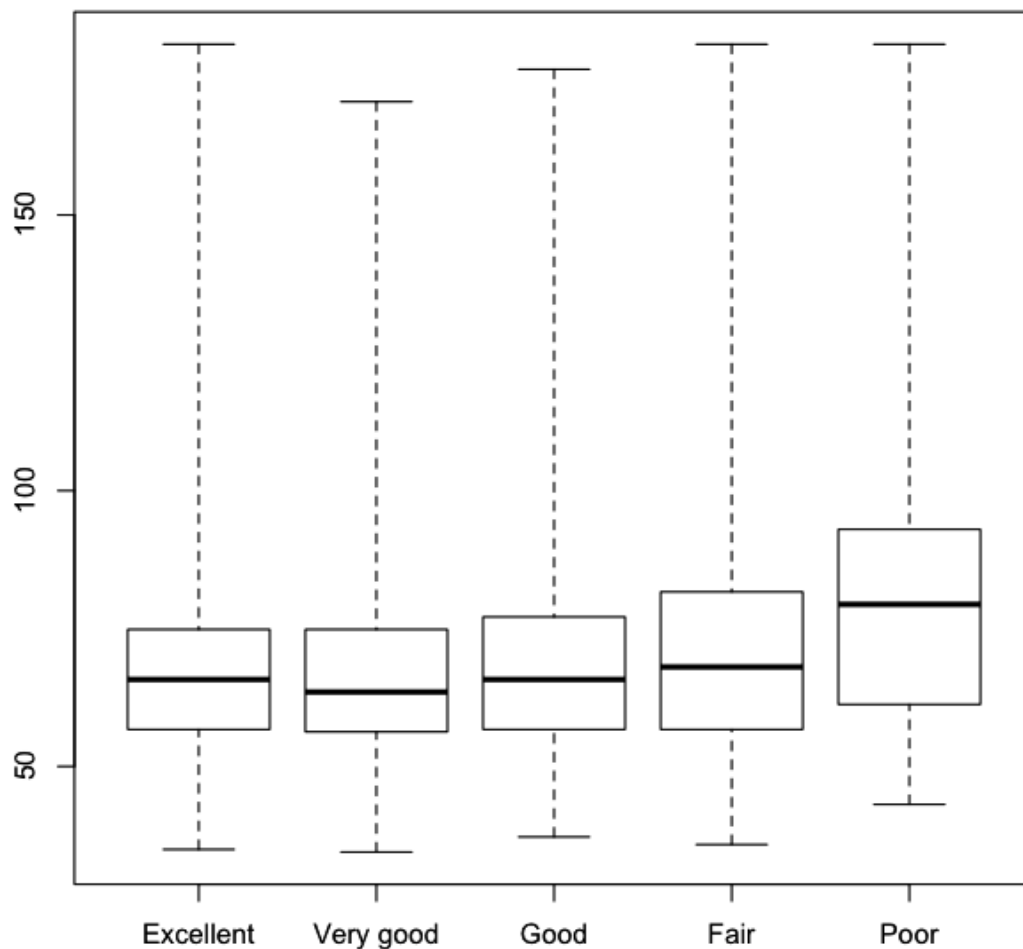


Box Plot

What the plot shows: A visualization of a range of values: the median (the thick black line), the upper and lower bounds of the values (the ends of the dotted lines), the 25th quartile (the lower part of the box), and the 75th quartile (the upper part of the box)

Data are categorical (general_health) and numerical (weight)

Example: Box plot of weight compared to general health



Quantitative Data and the CDC Survey Activity

Part I—Getting familiar with the survey

1. Load the CDC data file.
2. How many students responded to the survey?
3. What type of variable is gender?
4. What are the possible responses for gender?
5. What kind of data do you see for height?
6. What type of variable is height?
7. What type of variable is weight?

Part II—Creating subsets

Reminder that subset allows you to make a copy of just the data that meets certain requirements. For instance, you could separate out the responses for female students or students of a specific height.

1. Look at the age, gender, height and weight of everyone that has a height of 1.27
2. How many students are 1.27 m tall?
3. Save the subset of all Females as women.
4. Save the subset of all Males as men.
5. How many men are there?
6. How would you subset out the males that are 17 years old? How many men are in this subset?
7. What are the possible answers to general_health?
8. Create a subset of students that have a general_health of “Fair” **OR** “Poor”
9. How would you subset out students with excellent or very good health?
10. Why might subsetting be useful for your final project?

Part III—Graphical Representations

1. Make a histogram of men's weight. Paste the resulting plot into a document.
2. Add a line at the mean in magenta and a line at the median in cyan.
3. Why would you want to identify both the mean and the median on the same graph?
4. Create side by side box plots to compare weight and general_health? Paste the resulting box plots in a document.
5. What story do you see?

**** Transforming data**

1. Create a plot of men's height in inches.
2. What is the new mean height?
3. Transform weight from kg to pounds (1kg=2.2 lbs).
4. What is the mean weight now?
5. Write three questions that are interesting to you related to this data. Create plots of each and paste them into a document. At least one of them should be a Mosaic Plot comparing 2 categorical variables. For each plot write a description of the story it tells and how the data support the story.

Instructional Day: 21

Topic Description: In this lesson students use the data they have collected and contextual data sets to do statistical analysis with mean, median, maximum, minimum, and display information with a variety of plots for use in the final project.

Objectives:

The students will be able to:

- Analyze the data they have collected using statistical analysis and a variety of plots

Outline of the Lesson:

- Statistical analysis with mean, median, maximum, and minimum with student generated data and contextual data sets (55 minutes)

Student Activities:

- Groups do statistical analysis with mean, median, maximum, and minimum using the data they have collected and additional contextual data sets.

Teaching/Learning Strategies:

- Students work in their groups to analyze the data they have collected pulling in the additional contextual data sets as appropriate.

Resources:

- Student data
- Additional contextual data sets

Instructional Days: 22-24

Topic Description: In this lesson, computation with text is explored. A variety of filters and queries are used to create subsets of text data. Bar charts are used for graphical display.

Objectives:

The students will be able to:

- Read in a file containing text as data
- Filter a text data set (remove punctuation, remove case, remove stop words, stemming)
- Create a bar chart as one method of analyzing text
- Create and query subsets of a text data set

Outline of the Lesson:

- Introduction to text data (25 minutes)
- Introduction to Text Activity (30 minutes)
- Basic Analytics (20 minutes)
- Computing with Text Activity Part I (30 minutes)
- Journal Entry (5 minutes)
- Focusing on the words (20 minutes)
- Computing with Text Activity Part II (30 minutes)
- Wrap up Question (5 minutes)

Student Activities:

- Participate in discussion of text data.
- Complete Introduction to Text Data Activity.
- Complete journal entry.
- Complete Parts I and II of the Computing with Text Activity and participate in the discussions associated with the activity.
- Provide responses to the wrap up question and participate in discussion.

Teaching/Learning Strategies:

- Introduction to text data (You can use Introduction to Text Data as a resource.)
 - Explain secondary uses of data.
 - Provide Twitter background.
 - Ask students what words or phrases people may use to describe the first warm day after winter.
 - Look at Jillamore pdf.
 - Have students plot her location on a map.
 - Look at Weather Underground site.
 - Have students search for Jillamore’s location.

- Load the weather data file.
- Navigate through the file.
 - Ask students to provide an explanation of what the variables are, etc.
- Load the twitter data file. Navigate through the file.
 - Ask students to explain the variables, etc. Ask questions such as: What terms were searched for? Which had the most tweets?
 - Point out that this original file is very large, so to improve performance of the computers and make it easier for them to work with the file they will be using a subset. This is a good opportunity to remind students that computing is a powerful tool and allows working with large data sets, but is limited by things like processor speed and memory. (This can be linked back to the lessons in Unit 1.)
 - Demo how to create a subset of the data for a specific region of the country by picking latitude and longitude boundaries.
- Load the twitterwithdate data file.
 - Point out the created date variable.
- Have students complete the Introduction to Text Activity.
 - Discuss results.
 - Note that students may have some difficulty creating the subset expressions.
- Text Analysis
 - Part I—Basic Analytics
 - Text can be analyzed many different ways. Research areas like "stylometrics" attempt to say something quantitative about an author's work; e.g., by computing the average number of words per sentence or the average number of letters per word written by an author. Analytics can also be used to find patterns in other types of text. (You can use Computing with Text Background as a resource.) In the first part, the basics of counting words in a file and creating bar charts based on those counts will be addressed by working with the California subset of the twitter data file. Demonstrate how to do the following:
 - Load the CATwitter data file.
 - Look at the tweets.
 - Change the size of the column in order to view the entire tweet.
 - Scroll and look at the variety of tweets.
 - Text mining—analyzing word counts. Demonstrate how to do the following:
 - This is a good opportunity to explain that the tweets are stored in an array (or vector) where the numbers in front indicate the place the tweet is in the vector. Arrays are an important concept in computer science. Storing items in an array allows us to access particular elements, search and sort.
 - Demo how to view the vector and point out that each of the array elements of the corpus matches the corresponding tweet in the data file.
 - Create a frequency table that separates out each word and counts how many times it appears in all the tweets.
 - Ask questions such as: What is the word that appears least frequently? What is the word that appears most frequently?

- Demo how to produce frequency tables that show only the most frequently appearing words and the different sorting options.
- Demo how to produce a bar chart of frequently occurring words.
- Journal Entry: What do you think would happen if you did all of these same things on the NJ subset?
 - Have students complete Part I of Computing with Text Activity.
 - Discuss results before going on to Part II.
- Part II—Focusing on the Words
 - Demonstrate how to do the following:
 - Remove case. Make “Spring” and “spring” be the same thing by making everything lowercase. (Note: each new corpus that is created should be assigned a new name.)
 - Removing "stop" words. Some words like “a” and “the” are probably always going to appear frequently because they are common parts of speech. Those words can be removed to emphasize the other less common words. Demo the method for removing stop words.
 - Deleting punctuation. Notice that many of the captions include symbols other than numbers and letters. Demo the method for deleting punctuation.
 - Stemming. It might be useful to ignore the ending of words such as “s”, “ing”, etc. In other words, change words like “boats” and “boating” to just “boat”. This is called stemming. Demo stemming.
 - Have students complete Part II of Computing with Text Activity.
 - Discuss results.
- Wrap up question: What is the source of the words that will be analyzed for your final project?
 - Ask students to provide a response. Make sure they understand that the answer is any of the text they enter that is “free text”.

Resources:

- Introduction to Text Data
- Jillamore.pdf
- Introduction to Text Activity
- Computing with Text Background
- Computing with Text Activity

Introduction to Text Data

****Secondary uses of data**

Data that are publicly available on the web are subject to a host of secondary uses. As the consumer of these data, there are a few questions to ask: Who collected the data and why were they collected? When was the data collected and how old is the data? What was their original purpose? What are the strengths and limitations of these data for your problem? How are the data organized? How do you access them? Is there someone who you can ask for help if you have questions?

****Twitter and the Jillamore file**

Twitter is a micro-blogging site that handled 4 billion messages or tweets in the first three months of 2010 alone. As a social network, Twitter culls activities from millions of people and there have been several studies of what people are posting to Twitter.

It is possible to look to Twitter and its users for signs of spring. Somewhere in the daily observations of millions of people it should be possible to find comments about the changing season.

The easiest place to start is with "Spring is here". On April 5, 2010 at about 1 pm a search for the phrase "spring is here" was submitted. Jillamore.pdf is a screen shot of the search results.

The last tweet was from the user "jillamore" who comments that it is a beautiful day in her part of the country, with temperatures in the upper 70s. She declares "Spring is here!" A bit more about this person can be learned from the last page in the Jillamore.pdf file. She lives at 40.360171 latitude-74.079609 longitude. The point can be plotted on a map. (Enter latitude and longitude into Google Maps, for example.)

This point is Red Bank, New Jersey. To get a sense of what the weather has been like in her part of the world, a service like the Weather Underground can be examined.

<http://www.wunderground.com>

It allows searches for weather anywhere in the country. This site is interesting to us both because it is possible to see what the weather has been like for jillamore and also because of where the data to make this judgment comes from. The Weather Underground culls data from about 10,000 officially run weather stations (e.g., the National Oceanic and Atmospheric Administration or NOAA) and 8,500 that are privately run but subject to strict data quality controls. The idea that citizens would install sensors and volunteer their data is very much in the spirit of the phone applications being used.

Search the Weather Underground for Red Bank, NJ. It shows historical weather data for this city. A file starting from January 1, 2010 to April 5, 2010, when the tweet about spring was posted was created from this information.

Load the weather data file and look at it in the Data Viewer.

Each row represents a different day and the names of the different variables recorded for each day appear in the first row.

**** Some historical data**

A group of researchers made hourly requests from the Twitter API using several different phrases in addition to "spring is here"—some relating to things turning green or trees beginning to bud. The researchers were also interested in the beginning of fall, so they also collected data on phrases like "fall is here" and comments about leaves turning colors.

Load the twitter data file and look at it in the Data Viewer.

Each row represents a single tweet. The variables include "created" which is a timestamp; Twitter "username" from the person who wrote the tweet; the "longitude" and "latitude" of their location (either their home or, if they are using a smart phone, the place where they typed in their tweet); which of the researcher's "search_term"s the tweet matched; and then the "message" itself. The variable "search_term" is a factor (categorical).

Load the twitter with date data file and look at it in the Data Viewer.

This file was created from the original twitter file and in addition to being a small subset; it includes a variable that indicates the created date in date format. (Note: to sort by date, the numerical created date needs to be used.) This file can be subsetted further by choosing latitude and longitude boundaries. For example, the New Jersey area would be approximately bounded by latitude between 38.5 and 41.5 and longitude between -75.5 and -73.5.

Introduction to Text Activity

Load the twitterwithdate data file and the weather data file.

1. Create a subset of the twitter data that includes only the tweets that contain "Spring is here".
 - How many tweets contain "Spring is here"?
 - What other search terms could you include that might indicate spring?
 - Create a subset that includes "Spring is here" and at least one other search term. How many tweets were added?
2. Create a subset of all tweets from approximately the New Jersey area.
 - Look at locations near where Jillamore lives. How many tweets are from that area? Do they match Jillamore's description? What other ways could you use to verify this?
 - Plot the New Jersey subset on a map. Experiment with different point sizes and zoom levels. What inferences can you make from the plot?
 - Sort by created. Do the dates and search terms correspond correctly?
 - Who has the most tweets? How does that impact the total number of tweets from New Jersey? How does that impact the number of tweets that include "Spring is here"?
3. Create a subset of California.
 - How many tweets are from that area?
 - Plot the California subset on a map.
4. What reasons can you think of to explain the difference in the number of tweets between New Jersey and California? How might you test your reasoning?

Computing with Text Background

A book usually has a fairly predictable structure. There are chapters which are made up of paragraphs which are made up of sentences which are made up of words. Research areas with names like "stylometrics" attempt to say something quantitative about an author's work. It is possible to compute the average number of words per sentence or the average number of letters per word written by an author. Some authors write in short, choppy sentences, while others craft sentences that are over a page long, adding phrase after phrase. Some authors choose simple vocabulary, while others prefer long, complex words. Statistics of this kind can not only point out interesting ways to think about the differences between authors, but they can even be used to help us figure out who wrote texts if their author is unknown or uncertain. One of the earliest analyses of this kind was of the famed Federalist Papers, a collection of documents describing the philosophy motivation behind our system of government. The papers are thought to be written by Alexander Hamilton, James Madison and/or John Jay. In the mid 1960s, a group of statisticians considered a number of novel statistics to differentiate the writing styles of the three men.

The counts of the different words in a document have also been used to characterize something about the document's subject.

The idea that the frequency with which words appear in a document might reflect something of its content has real-world applications. For example, the spam filter that intercepts junk e-mail is working on the frequency of words in each message. If a message makes too many references to "sales" or "won" or "Visa", there is a strong suspicion that the e-mail is spam.

The goal of this section is not to develop any of the topics above in any great depth. Instead, it will provide some basic tools for simple analysis on text.

Computing with Text Activity****Part I—Basic Analytics**

Load the NJTwitter data file.

1. Create a vector from NJTwitter.
2. View the vector. Do the elements of the corpus match the messages in the NJTwitter data file?
3. Create a frequency table that counts how many times each word appears in all the tweets. What is the word that appears least frequently? What is the word that appears most frequently?
4. Create a frequency table that shows only the top 1% of frequently appearing words sorted by ascending frequency. What are the most frequently occurring words?
5. Create a bar chart of the most frequently occurring words. Does this match your frequency table?
6. Experiment with different percentages to determine the greatest percentage that allows you to read all of the words on the chart. Describe your process and the reasoning for your final answer.
7. Notice that there is a “spring” and a “spring!” (with an exclamation point). Do you think we should include those counts together or keep them separate? Why?

****Part II—Focusing on the Words**

Another data set was collected by using the API (Application Programming Interface) to conduct a search on Flickr for images that were tagged with the word "chill". The first 3,000 image captions in the list of search results were downloaded. Those captions containing words that were inappropriate were removed. In order to make it easier to work with the file, a subset was created.

Load the smallcaptions data file. Enlarge the column so that the entire caption can be viewed.

1. Change the list into a vector. View several of the elements of the vector. Do they match the file?
2. Create a frequency table for the entire list. What is the most frequently occurring word? Scroll to look through the entire file.
3. Run the frequency for only the top 10%. What does this do? Describe anything that you notice.
4. Run the frequency for the top 1%. Make a list of the words and their counts.
5. Create a bar chart for the top 1%. Describe what you see. Save the chart to a document.
6. Experiment with a few more % choices. Which gives the most information? Explain your choice.
7. Create a frequency table of the top 1% of words after making them all lower case and without punctuation. Make a list of the words and their counts.
8. Create a bar chart for the top 1%. Describe what you see. Save the chart to a document.
9. Experiment with a few more % choices. Which gives the most information? Explain your choice.
10. Create a frequency table of the previous corpus after deleting stop words. How did the file change?
11. Create a bar chart of the file without stop words. What are some of the words that disappeared from your bar chart? Why might it be useful to delete these stop words?
12. Create a frequency table of the previous corpus after deleting stems. How did the file change?
13. Create a bar chart of the file without stems. What are some of the words that disappeared from your bar chart? Why might it be useful to delete these stems?

Instructional Day: 25

Topic Description: In this lesson students use the data they have collected and additional contextual data sets to do text analysis for use on the final project.

Objectives:

The students will be able to:

- Analyze the data they have collected using text analysis techniques

Outline of the Lesson:

- Analyze text in student generated and contextual data sets (55 minutes)

Student Activities:

- Groups do text analysis with the data they have collected and the additional contextual data sets.

Teaching/Learning Strategies:

- Students work in their groups to analyze the data they have collected with text analysis techniques pulling in the additional contextual data set as appropriate.

Resources:

- Student data
- Additional data set

Instructional Days: 26-27

Topic Description: Students complete final projects.

Objectives:

The students will be able to:

- Incorporate all objectives of the unit into the final project

Outline of the Lesson:

- Review of final project expectations (20 minutes)
- Overview of rubric (15 minutes)
- Final project (~3 days)

Student Activities:

- Teams complete final projects.

Teaching/Learning Strategies:

- Review of expectations and overview of rubric
 - Discuss the rubric and answer questions.
- Final project
 - Teams work on final projects.
 - Help student teams with projects as necessary.

Resources:

- Final Project
- Final Project Sample Rubric

Instructional Days: 28-29

Instructional Days: 28-29

Topic Description: Students complete Scratch projects or websites to use with the presentation of their final projects.

Objectives:

The students will be able to:

- Incorporate all objectives of the unit into the final project

Outline of the Lesson:

- Review of final project expectations (10 minutes)
- Overview of rubric (10 minutes)
- Final project presentation development (~2 days)

Student Activities:

- Teams complete final project presentation.

Teaching/Learning Strategies:

- Review of expectations and overview of rubric
 - Remind students of project expectations.
 - Discuss the rubric and answer questions.
- Final project presentation development
 - Teams work on final projects.
 - Help student teams with projects as necessary.

Resources:

- Final Project
- Final Project Sample Rubric

Instructional Day: 30

Topic Description: Students present final projects.

Objectives:

The students will be able to:

- Incorporate all objectives of the unit into the final project

Outline of the Lesson:

- Final project presentations (55 minutes)

Student Activities:

- Teams present final projects.

Teaching/Learning Strategies:

- Final project presentations
 - Student teams present their findings to the class.
 - Other teams ask questions and participate in the discussion.

Resources:

- Final Project
- Final Project Sample Rubric

Final Project

Analyzing Your Data

In the course of the past few weeks, your group has collected data using the phone application. Now it is your turn to tell an interesting story based on the data. You will present your story to the class (it can be a series of web pages or a Scratch program). You must include plots/graphics that support the story.

You may include data from any of the other data sets you've seen in the lessons. However, these data cannot be the primary source of your story.

You will have access to data from your classmates as well as students in other classes that have also collected data. Keep in mind that you have already done some analysis on your data. This is your opportunity to pull it all together, modify as necessary and tell a compelling story that makes a case or highlights a discovery.

Final Project Sample Rubric

	Points Possible	Yes/No	Points Earned
Does your web page or Scratch program:			
1. Have a title with your group members' names?	5		
2. Tell a story based on your data? (Why does this data support your story?)	10		
3. Have 2 or more descriptive plots?	10		
4. Have other types of visuals?	10		
5. Have a description of why you chose the visuals that you did?	10		
6. Address how you can use this data to make a difference?	10		
8. Bring in data from an outside source that supports your story?	10		
Does your presentation include:			
How many items are in your data sets? Based on that how valid is your story? An explanation of this?	10		
An explanation of what you learned in this unit (analysis techniques, etc)?	10		
Web Page or Scratch program uses appropriate features for the medium	15		
Total	100		