Version 4.0

Unit 5:

Computing and Data Analysis

A High-School Curriculum Exploring What Computer Science Is and What It Can Do

ECS

EXPLORING
COMPUTER
SCIENCE

Introduction

Managing and interpreting large amounts of data is part of the foundation of our information society and the economy. 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. Although it is highlighted throughout the unit, it is important to make special note of the fact that R and the graphical data analysis system, Deducer, are just tools for the data analysis that the students will be doing throughout the unit. Students need to be able to use the commands in appropriate ways to manipulate the data, but the focus should be on conceptual understanding. This is especially important since R is only one of many tools that can be used for data analysis.

The unit is divided into three main sections.

- Overview of the final project and the data collection process (Days 1-6)
- Data analysis techniques (Days 7-23)
- Final project (Days 24-30)

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

Due to the possibility of technical difficulties at any point in the unit, extra time has been built in to “catch up” should these occur. By the time students reach the final project, much of the data analysis will be complete; therefore, fewer days can be allocated the project, if that seems reasonable. This extra time also provides flexibility to spend additional time on lessons where discussions are rich rather than cutting those short. For example, additional time could be allotted for students to get to know the members in their group and plan their campaign.

Teachers who do not have mobile technology available can still teach this unit. Some possible options for the days on which students use their own data include: Pose more questions related to the data sets provided in the curriculum or use the contextual data sets included in the Canonical Campaign Supplement to create additional activities for each analysis type. A possible option for the final project is to extend the project from Unit 2 to include more detailed analysis and then proceed with the design of a web page or Scratch program.

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

Note: Information on Deducer can be found in the Deducer Quick Start Guide. Information specific to the canonical phone applications can be found in the Canonical Campaigns Supplement. Both of these documents are included with the Supplemental Materials.

Daily Overview Chart

Instructional Day	Topic
1	Review how data can be used for making a case/discovery and provide an overview of the final project.
2	Discuss photo ethics and student safety related to android phone use.
3-5	Distribute phones. Create groups. Discuss group roles and responsibilities. Navigate the android application. Navigate the online system.
6	Data check-in—Discuss issues that arise (aggregating data, etc.).
7-10	Introduce R/Deducer. Create maps using the latitude and longitude of a location and then create maps from a file of data.
11	Create maps with student data and related data set.
12-14	Discuss bar plots, categorical and continuous data, and mosaic plots as a vehicle for comparing categorical data, and looking at trends in data.
15	Create bar plots and mosaic plots with student data and related data set.
16-18	Review mean, median, minimum, maximum. Discuss various ways to subset data. Represent data with box plots and histograms.
19	Identify mean, median, minimum, maximum, create subsets, and create box plots and histograms with student data and related data set.
20-22	Use a variety of filters and queries to create subsets of text data. Create bar plots to graphically display the information.
23	Analyze text in student data and related data set.
24-26	Finalize data analysis for final project.
27-29	Develop website or Scratch program to present data analysis campaign.
30	Final project presentations

Daily Lesson Plans

Instructional Day: 1

Topic Description: This lesson sets the stage for the unit. It provides a review of the data collection and analysis that will be needed in order to complete the final project.

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 a discovery.

Outline of the Lesson:

- Collect permission slips (5 minutes)
- Review of data collection and making a case/discovery (30 minutes)
- Overview of final project (20 minutes)

Student Activities:

- Participate in discussion of data review.
- Participate in discussion of final project.

Teaching/Learning Strategies:

- Collect parent permission slips.
- Overview of the Final Project
  - If you have access to videos of sample projects, show a few of them.
  - Use the appropriate parts of the Canonical Campaigns Supplement as a guide for this discussion. The overview will help students understand why they are receiving phones, what they will be doing with the phones, what issues they should think about in data collection, and what questions they might choose to ask as part of their analysis. You may choose to provide students with a printed copy of this overview, but the most important thing is to encourage the discussion. Highlight each of the following in detail.
    - Context and examples
    - Data collection with the android application
    - Analysis phases
    - Making a case v. discovery

Resources:

- Canonical Campaigns Supplement

Instructional Day: 2

Topic Description: In this lesson, photo ethics and student safety related to android phone use are discussed.

Objectives:

The student will be able to:

- Explain why they are not allowed to take photos of people during their data collection.
- Describe situations in which they should not take any type of photos.

Outline of the Lesson:

- Journal Entry (10 minutes)
- Photo ethics (20 minutes)
- Safety of the photographer (10 minutes)
- How to represent people in photos (15 minutes)

Student Activities:

- Complete journal entry.
- Participate in photo ethics discussion.
- Participate in photographer safety discussion.
- Participate in how to represent people in photos discussion.

Teaching/Learning Strategies:

- Journal Entry: Have you ever felt uncomfortable when taking a photo of someone? Have you ever felt uncomfortable when someone took a photo of you? Why?
- Discuss photo ethics. (See Photoethics.ppt in the Supplemental Materials for the photographs to display.)
    - Project Dorothea Lange photograph of "Migrant Mother".
        - Ask students if they recognize the photograph.
        - Ask questions such as: What does the image depict?
            - Probe as necessary—ask them other questions around the image itself. For example: What is this woman doing in the photograph? Who do you think she is? Where do you think she might be from? What do you think she does for a living and how can you tell? Who do you think the children are? Why do you think they are turned away from the camera?
        - Ask a question such as: How does this photo make you feel?
            - Most likely, students will say things such as "sad", "depressed", "hungry", or maybe "nothing." If students say that they don't get any reaction from the image, ask them what they think the *woman* is feeling in the image.

        - Ask questions such as: What in the picture makes you feel that way? (You might want to get them thinking about things like the lack of color, the lighting, the children's faces turned away, the skinny woman with a worried look, and the wrinkles in her face, etc.)
        - Finally, explain the history of the photo: From 1935-42, the US Farm Security Administration created to address agricultural problems and rural poverty during the Great Depression sent out photographers to help document what life was like in rural and poor areas of the country. This photograph is by Dorothea Lange and became world-famous for illustrating rural poverty in America.
    - Next, show the picture from the BBC News website
        - Ask a question such as: What is happening in this photograph?
            - Probe as necessary—ask them other questions such as: Where do you think this photo was taken? Who do you think this woman is? What is she doing in the picture? Who else is in the photograph? What is the soldier doing? Why do you think he is laughing? How do you think the woman is feeling at this moment? Why would she be picking up rice?
            - It is possible that students won't recognize anything wrong at first since they are probably used to seeing these types of images on television or in magazines/newspapers.
        - Then ask questions such as: How does this photograph make you feel? Why?
        - Next, show both photos (flipping back and forth between slides 1 and 2) and ask the students: Does anything seem wrong about these photos?
        - Try to get students thinking about the ethics of the photographer's role. You could ask questions such as: Who took these pictures? Where were the photographers when these pictures were taken? What do you think happened to these people after the photographer took the picture? What did the photographer win out of these photos compared to the subjects of these photos? How do you think the people in these photos felt when the photographer took their pictures? Happy to be photographed? Embarrassed? Unsure of themselves? Uncomfortable?
        - Finally, explain the story of the BBC photograph: This picture was published on the BBC News Website. It is a photo taken in the Middle East of a woman who is picking up spilt rice grains to feed her family. The soldier in the background may be laughing at her or may be smiling at the camera—it isn't really clear. Regardless, you can tell that the soldier doesn't really care about the woman's situation and that the woman is so starved and desperate that she is taking the time to pick up tiny grains of rice on the ground. Students may provide insights such as: There are many homeless people and people picking through the trash in the United States and they may have gotten so used to seeing these people that they don't even realize they are there; this might be the situation for the soldier. As human beings we have the power to help people in need around us and we shouldn't ignore others or take advantage of them by gaining a photograph out of their misery.

        - Explain to students that, as photographers, they should be conscious of the lives, privacy, and experiences of people around them; that these photographs may feel disturbing or uncomfortable—and if they feel uncomfortable looking at the images, this is a good thing because they are beginning to recognize how a photograph can really take advantage of people and be unethical. Discuss their journal entries and ask them to think of times when they might have felt uncomfortable when someone they didn't trust took their photo or when somebody they didn't know very well posted their photo on MySpace or Facebook.
        - Remind students that photographers have an important responsibility to respect the privacy of the things/people they take photos of (even if people aren't in the photograph) because people surrounding a person with a camera might feel uncomfortable too without the photographer even realizing it.
        - When students receive the cell phones and start taking pictures of things, they should be aware of how people near the object of the photo might feel about the camera/phone being pointed at them.

- Safety of the Photographer
    - Next, show the war photograph.
    - Ask questions such as: What is happening in this photograph? Where do you think this is taking place? Where is the photographer? What feelings does the image evoke?
    - Discuss how the photographer is at risk in this picture—can be easily shot down just like the soldiers.
    - Explain that as they collect data for this unit they should **NEVER** put themselves at risk. If taking the photograph is going to hurt them in any way—because somebody doesn't like that they're taking a photo, because someone sees their phone and wants to steal it from them, because they're standing in the middle of a busy street and might get hit by a car, etc.—then they should NOT take the photo at that moment, but should wait and take a picture of something that REPRESENTS what they wanted to take a picture of.
- Photo ethics and safety related to cell phone use
    - Explain to students: When they receive the cell phones, they will not be allowed to take any pictures of people.
    - This will prevent people from feeling uncomfortable if they take pictures around them
    - This will prevent the students being put at risk if they take pictures in dangerous areas of their neighborhoods, etc.
- Summary:
    - Don't make people feel uncomfortable by taking pictures of things with the cell phones.
    - Don't put yourself in an awkward or dangerous position by taking a photo.
- How do you represent people in the photos?
    - Brainstorm photography alternatives with the students by doing the following:
        - Ask students to imagine their best friend. Ask students to make a list of things that remind them of that friend in their journal. These could be objects related to memories they share with that friend, experiences they have had with that friend, inside jokes, etc.

- Ask students to share what they have written.
- Explain to students that this is how they can take pictures of people—by taking photographs of things that represent the person instead of taking pictures of the person him/herself.
- Whenever they want to take a picture of a specific person, they should wait and take pictures of an object or scene or something else that represents that person.
- Ask students to write in their journals about what they would take photographs of to represent an important person in their life, for example their mother or grandmother.

Resources:

- Photoethics.ppt (Supplemental Materials)

Instructional Days: 3-5

Topic Description: Phones are distributed. Groups for the final project are created and roles and responsibilities for group members are assigned. Students navigate the chosen android application, navigate the online application, and develop a list of possible questions to use for the final project.

Objectives:

The student will be able to:

- Explain the rules for sharing the phones and why those rules are in place.
- Login and navigate through the basic features of the phone application.
- Login and navigate through the basic features of the online system.
- Develop a method for data collection.

Outline of the Lesson:

- Phone distribution (25 minutes)
- Online system (25 minutes)
- Phone application (25 minutes)
- Group roles and responsibilities (35 minutes)
- Data collection method for the final project (55 minutes)

Student Activities:

- Groups complete phone paperwork.
- Experiment with the online application.
- Experiment with the phone application.
- Groups discuss group roles and responsibilities.
- Groups discuss potential questions regarding advocacy/discovery and data collection.

Teaching/Learning Strategies:

- Distribute phones
    - Have groups complete all phone paperwork.
    - Sharing the phone—Remind students that they need to share the phones with their group. Tell them to coordinate with their group how they will be sharing the phone. Each phase will be long enough that everyone should have a few days of data collection on their own, even though they are sharing the phone.
    - Remind students to follow school-wide rules when using phones.
    - Explain limitations on phone functionality (data collection device only).
- Determine who in each group will be responsible for the phone on the first night.
- Navigate the chosen application both online and on the phone. (You can use the Canonical Campaigns Supplement as a resource.)

  - Demo the online system.
  - Explain how the phone application works.
  - Give students time to explore both applications to become familiar with the features. Make sure that each student has an opportunity to use the phone application and understands how to enter data.
- Group roles and responsibilities
  - Explain to students that everyone in the group is accountable for the work on the in the various stages and that they need to alternate roles depending on who is the data collector for a given night, etc. 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 6 data check-in to use as a base for discussion.
  - Review the examples that were provided on day 1 of the unit.
- Discuss campaigns
  - Have student groups consider the research questions in the Canonical Campaigns Supplement as they develop a method for data collection.
  - Answer questions and guide them as appropriate.

Resources:

- Phone distribution forms
- Canonical Campaigns Supplement

Instructional Day: 6

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 the person in charge of the first phase of data collection upload their 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 be a data collector for some period of time?
        - 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?

Resources:

- No additional resources needed

Instructional Days: 7-10

Topic Description: This lesson introduces R/Deducer as a data analysis tool. 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 and Deducer (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, LA Bike Data and Deducer.
- 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 with the phone. How might seeing the data on a map help you analyze it?
- Install Deducer package. (See Deducer Quick Start Guide.)
    - 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.getlatlon.com/ 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 and Deducer
    - Load the labike data file.
    - Point out that this is what they will be doing with the data they collect with the phones.
    - Point out the following features as you discuss the layout of the table shown in the data viewer (You can use Exploring LACBC and Deducer as a reference.):
        - Header, number of rows, categories
        - Note what you see in the data viewer (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.
            - The table appears in the Console window.
            - 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.
        - Point out that the system provides a default title for the subset, but it is better to create a new title.
        - 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?
  - 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 ask students for their responses and lead a discussion to ensure that they understand each of the features of Deducer 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 ask students for their responses and lead a discussion to ensure that they understand each of the features of Deducer discussed so far.
  - 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 Deducer discussed so far.

Resources:

- Deducer Quick Start Guide
- Walking and Biking in LA
- Exploring LACBC and Deducer
- 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 and Deducer

Load the labike data file into the Deducer Data Viewer.

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 frame "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 (bike_count_pm ≥ 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 file into the Deducer Data Viewer.

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 into the Deducer Data Viewer.

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 collected on the phone?

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 into the Deducer Data Viewer.

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: 11

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: 12-14

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.

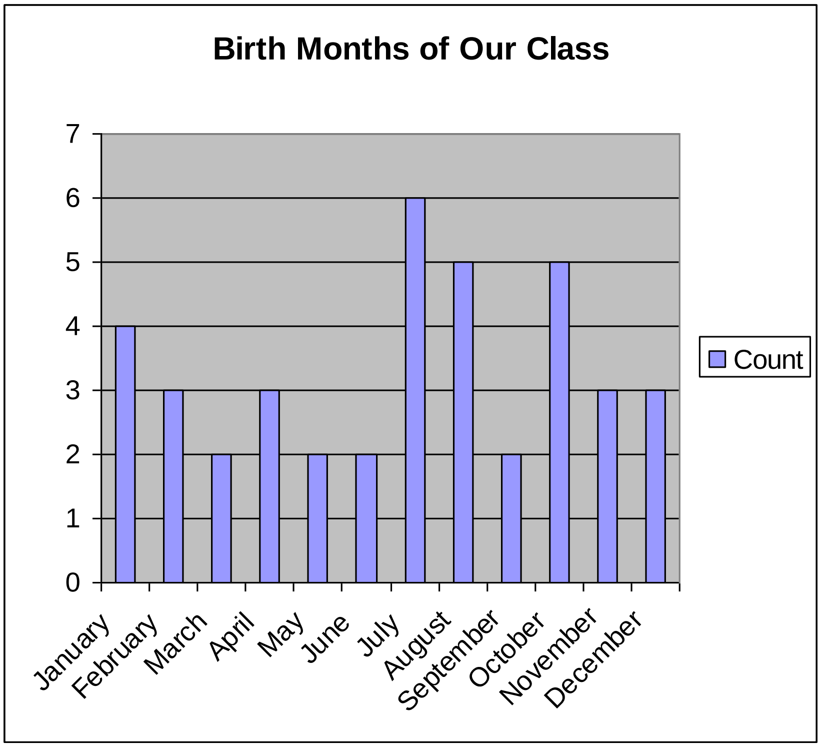
  - Have students help you create the skeleton of a bar plot like Sample Birth Month Bar Plot. You should end up with a similar chart, but without any counts (bars).
  - Ask each student what their birth month is. Increase the height of the corresponding bar by one until the entire class has responded.
  - 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
  - Look at the Variable View in the data viewer. Here you see the type of each variable and possible values that are assigned to each variable. Factor is the type for categorical variables.
  - Demonstrate how to use the plot tool 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.
- 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?"
            - Discuss the journal entry.
        - Create a contingency table with data to show the relationship between the answers to the two questions.
            - The table will appear in the Console window. 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 used when tagging events with phones 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
- Deducer Quick Start Guide

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: 15

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: 16-18

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 descriptives, 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 descriptive 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

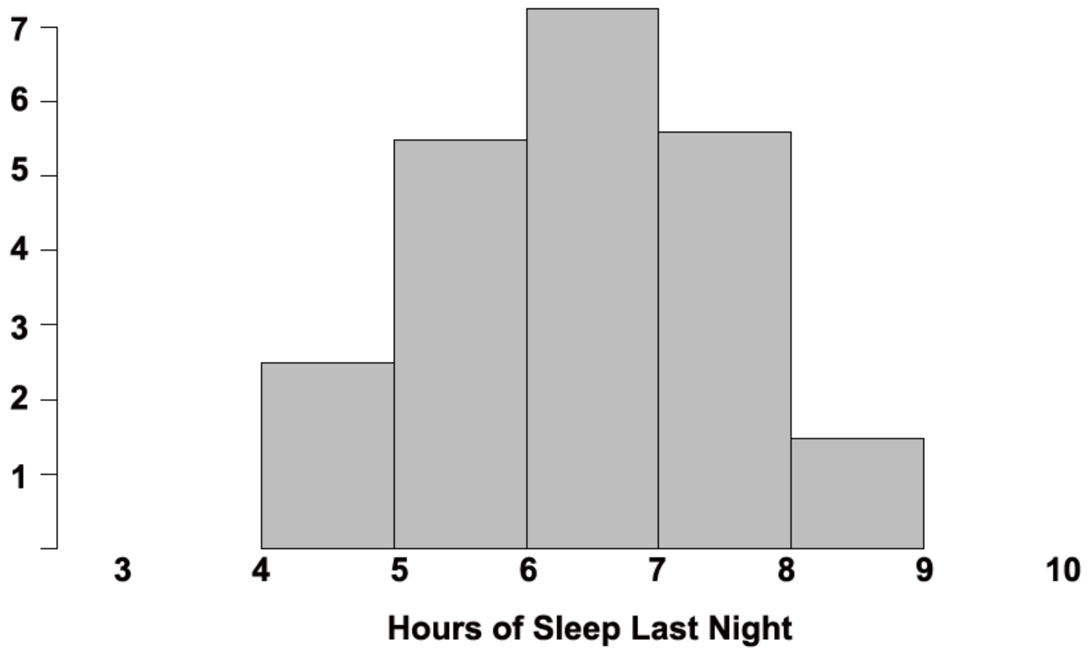
                them when one measure might be better than another. For example, the mean is more sensitive to outlying data. Students may ask what 1st Qu. and 3rd Qu. refer to. Explain that it will be easier to discuss that when you start doing graphical representations.

        - 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 items used when tagging events with the phones 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
- Deducer Quick Start Guide

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

Reminder: R/Deducer describes categorical data with the word “factor”and a decimal value with the word “double”.

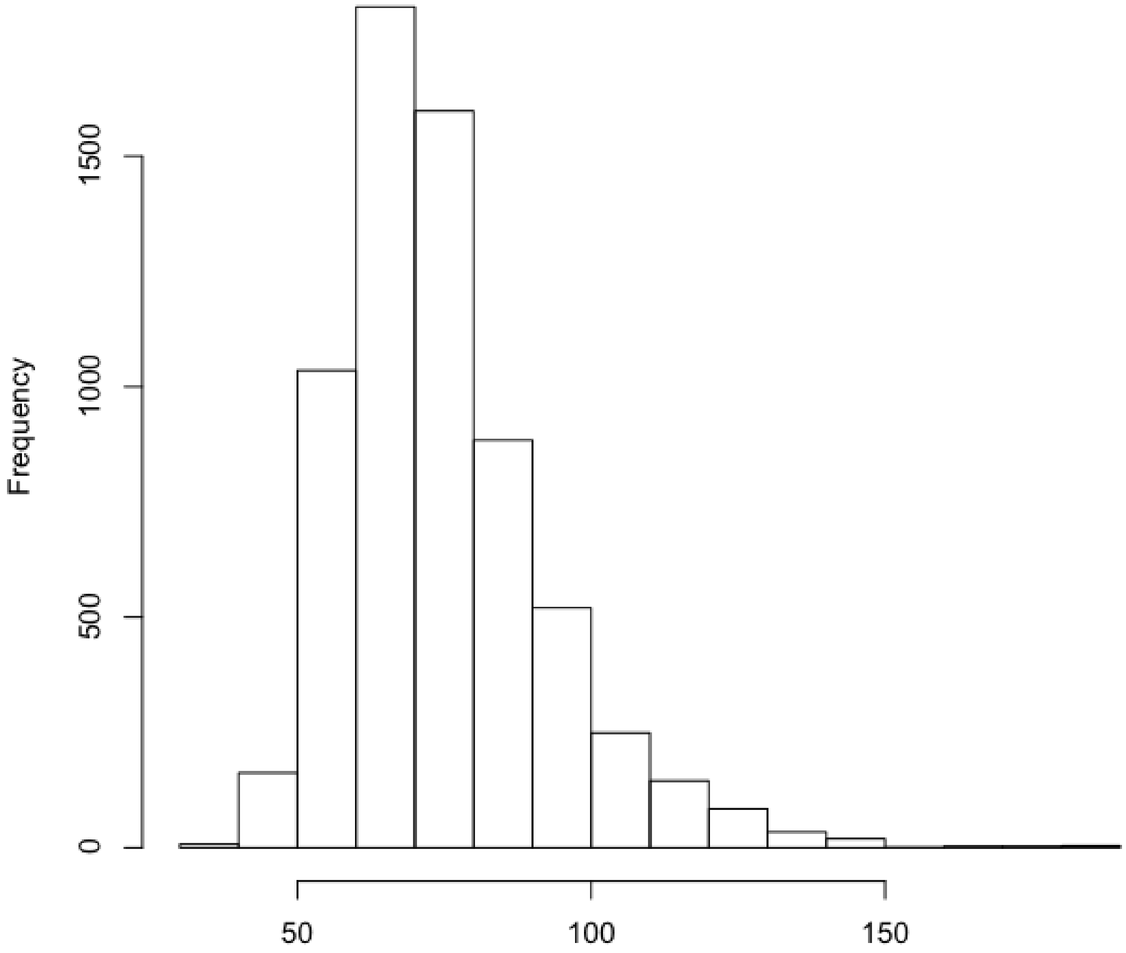
Source:
http://www.cdc.gov/HealthyYouth/yrbs/data/index.htm
2007_National_YRBS_Data_Users_Manual.pdf
2007 National YRBS Data Users Manual

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

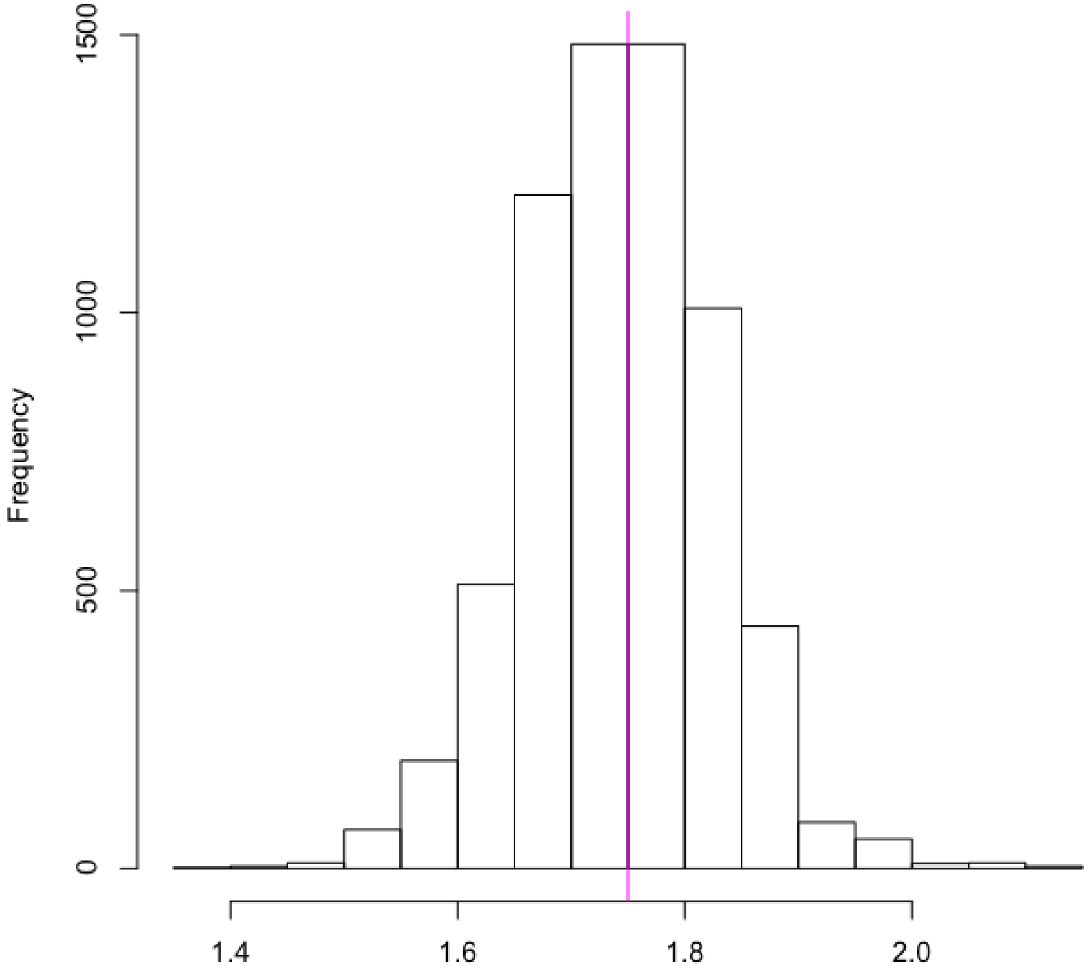


Men's weights

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



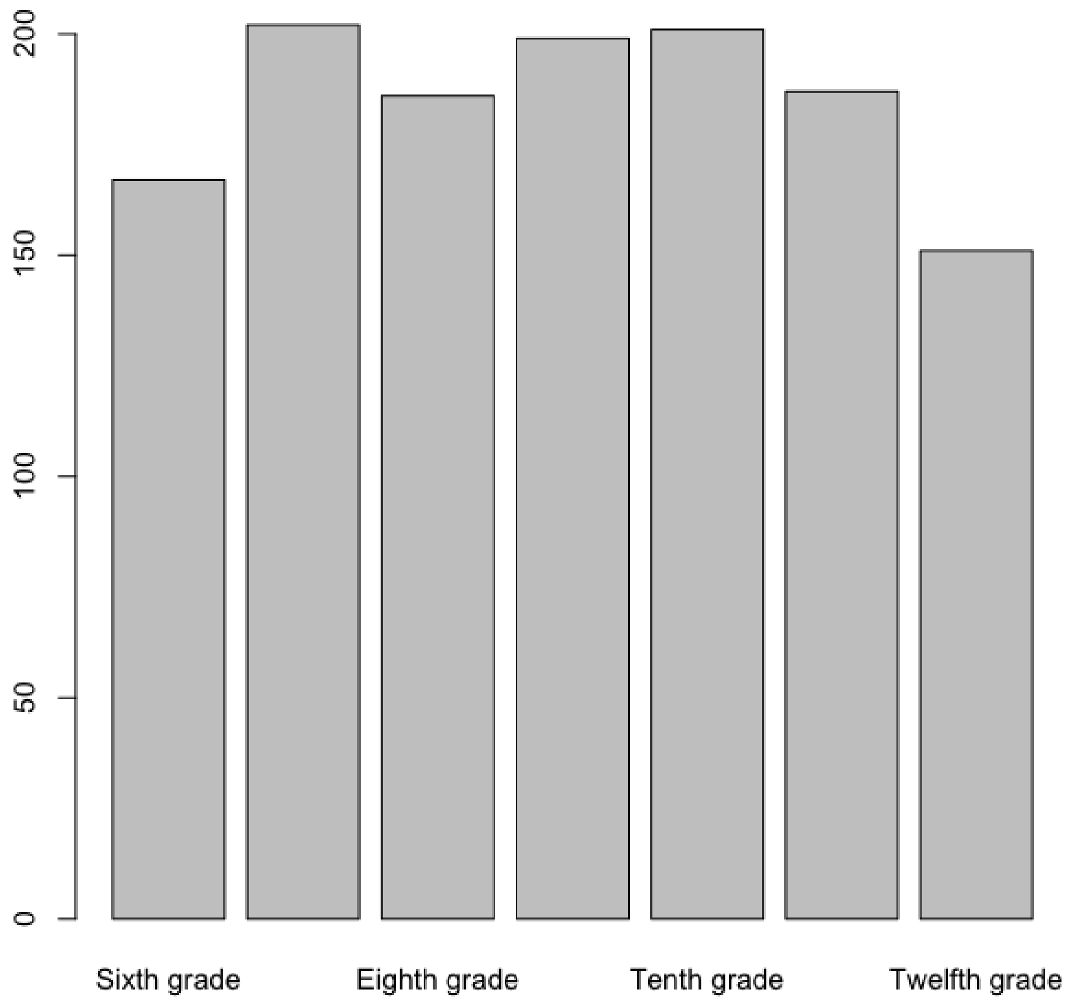
Men's heights

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.

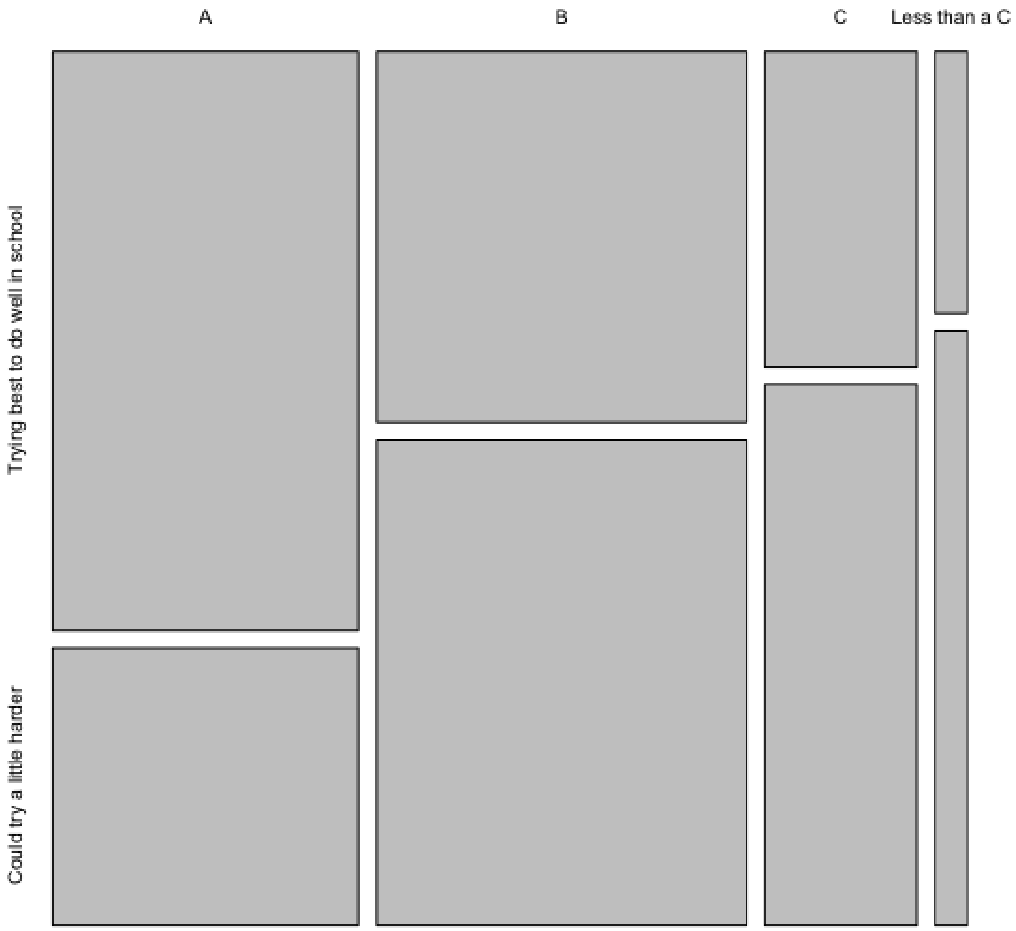
Example: Bar plot of years



Mosaic Plot – Sample 1 (categorical data)

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

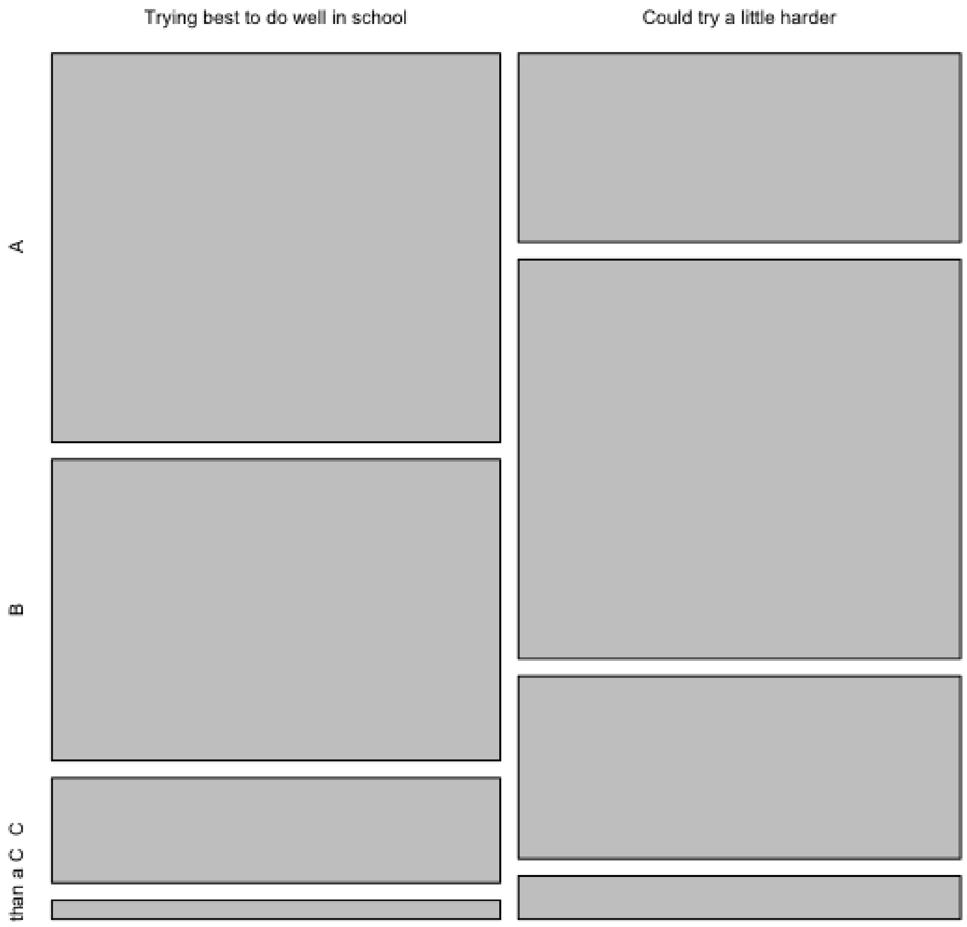
Example: Mosaic plot of grades compared to effort



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

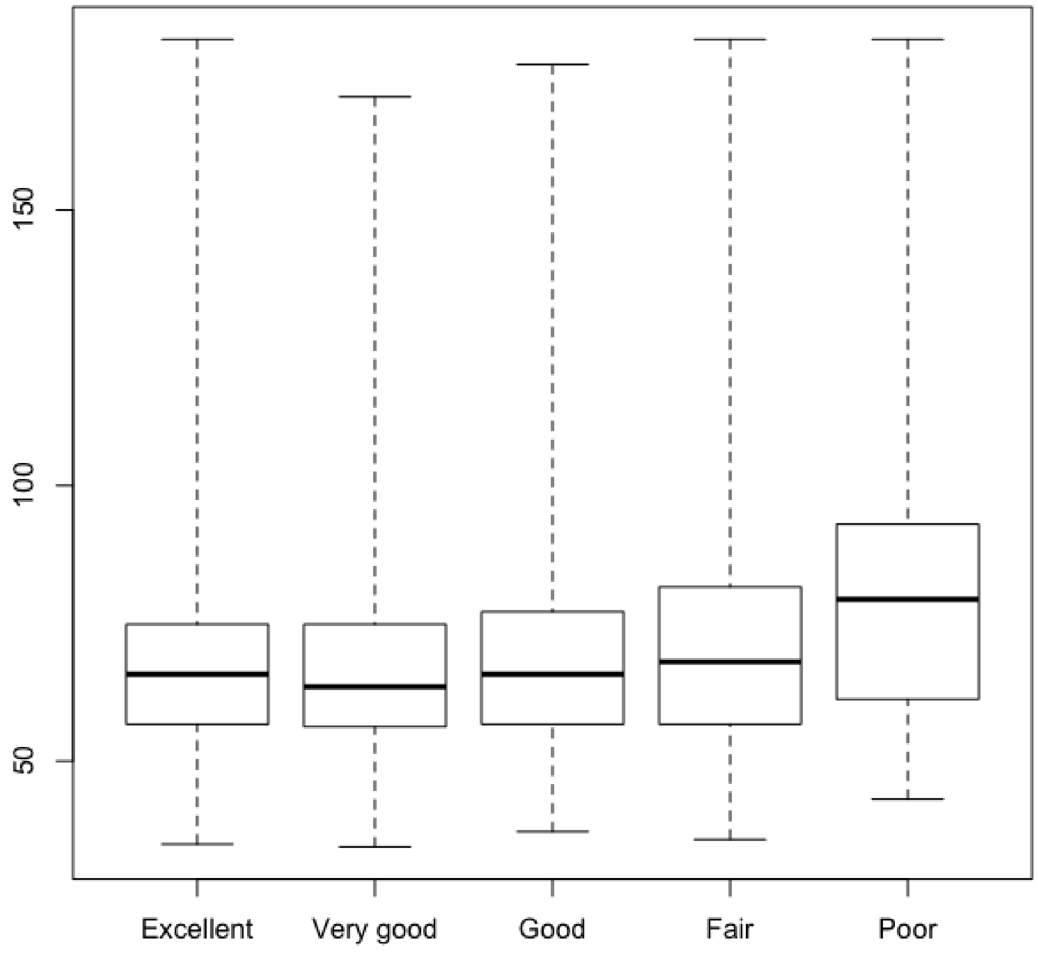


Box Plot

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. What command did you use? 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: 19

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: 20-22

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:
            - Turn the vector of tweets into a "corpus". A corpus is the term used to describe a collection of writings. This is necessary in order to do some more sophisticated analysis.
            - Demo creating the corpus.
            - 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 corpus 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
- Deducer Quick Start Guide

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 into the Deducer Data Viewer.

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 into theDeducer Data Viewer.

1. Create a corpus from NJTwitter.
2. View the corpus. 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 into the Deducer Data Viewer. Enlarge the column so that the entire caption can be viewed.

1. Change the list into a corpus. View several of the elements of the corpus. 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: 23

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: 24-26

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: 27-29

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 (~3 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 at other schools 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		