

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.

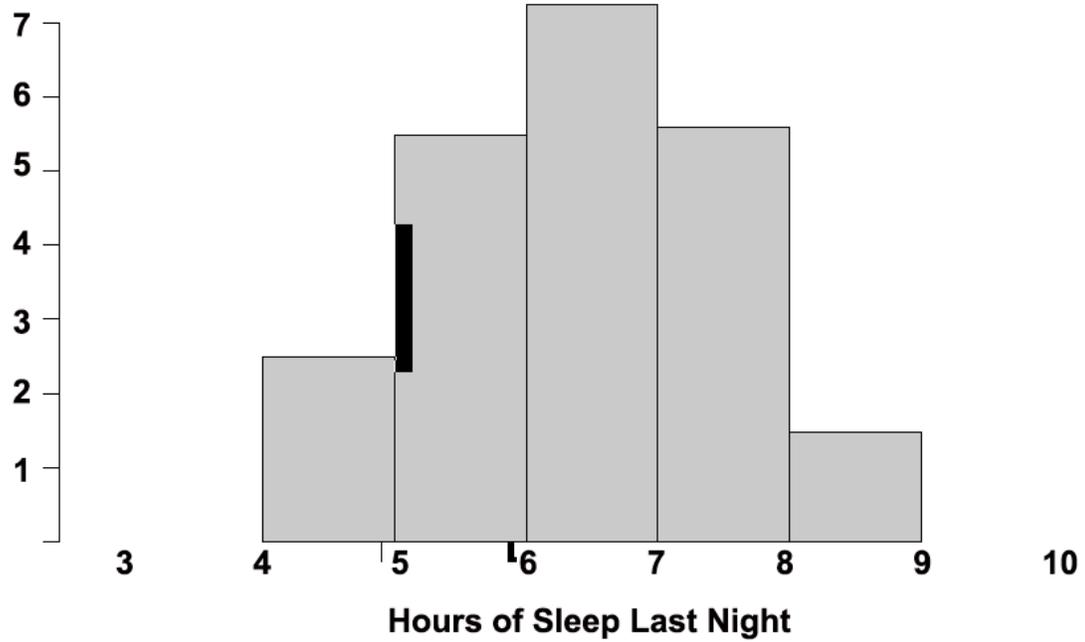
- For the most part, even if someone says they slept 8 hours that is an estimate. They really slept close to that amount (7.9 hours, etc).
- The histogram creates groupings of the data.
- To reinforce this idea, students cannot answer that they slept exactly 7 hours. They will have to choose either the 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

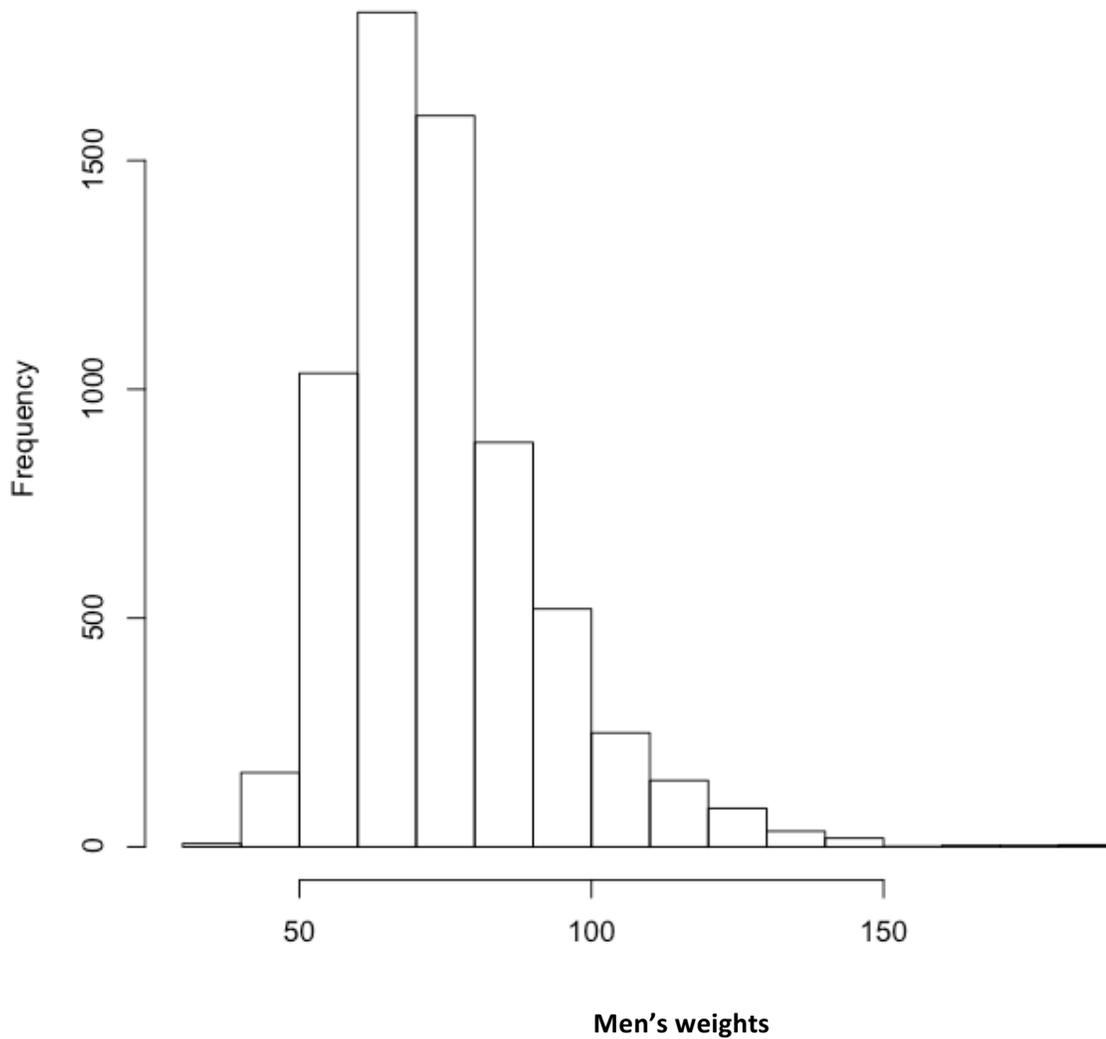
2007 National YRBS Data Users Manual

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.

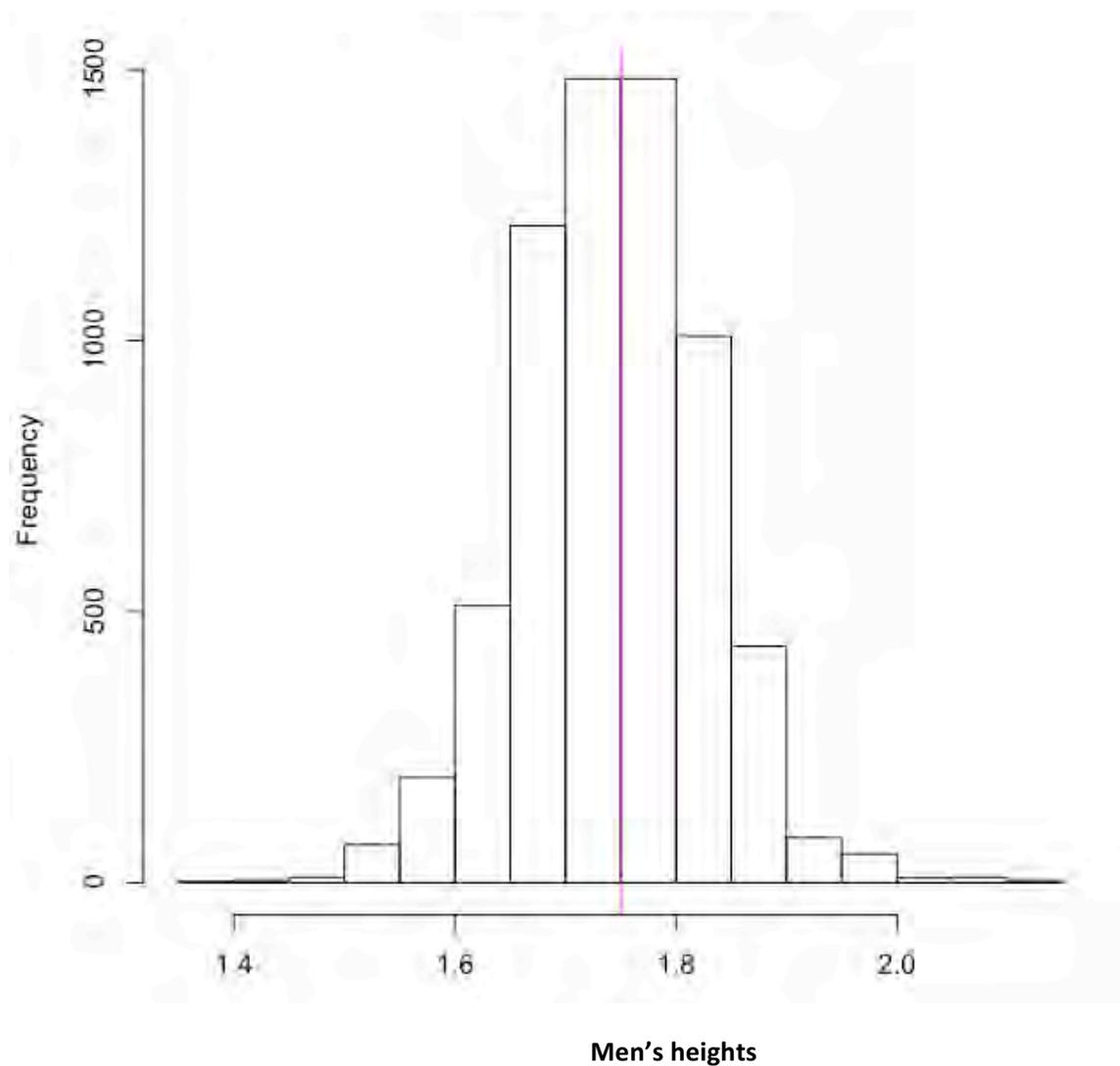
Example: Histogram of men's weights



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



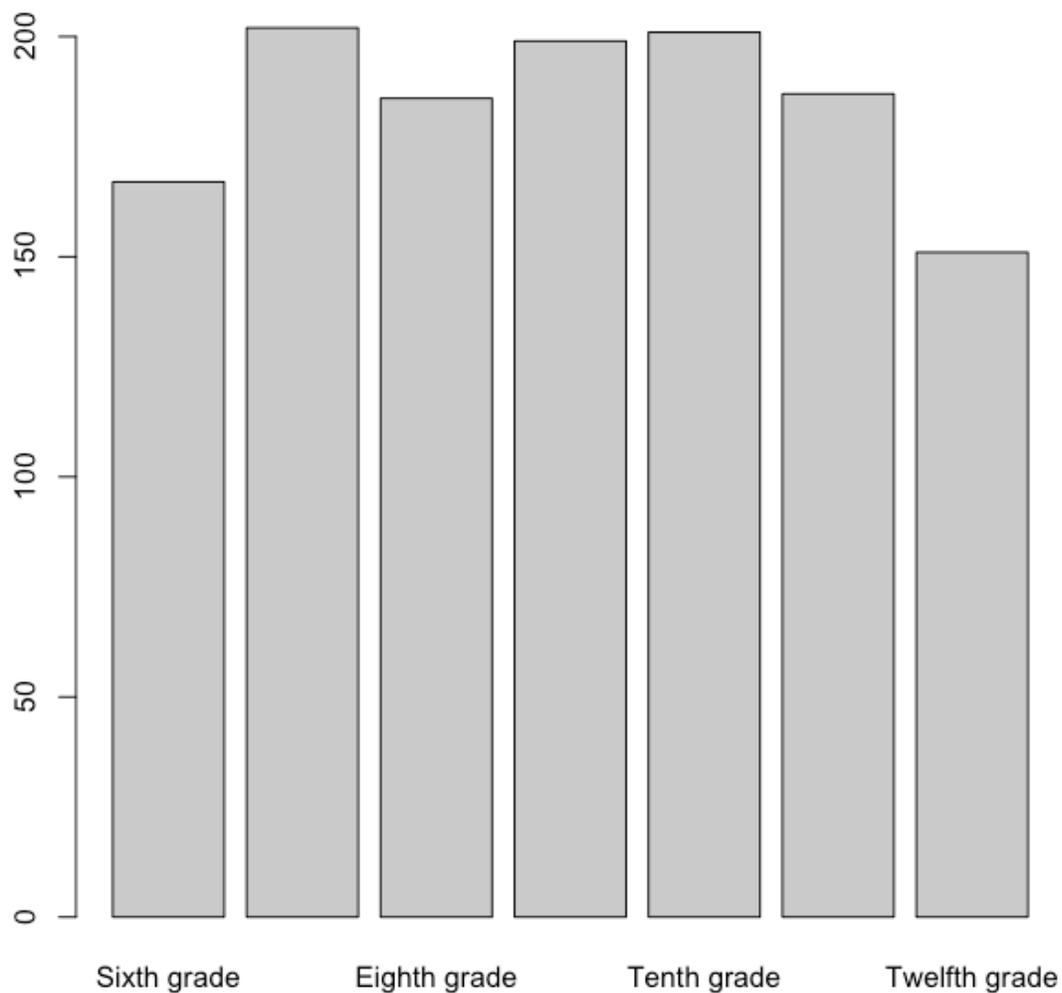
Bar Plot

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

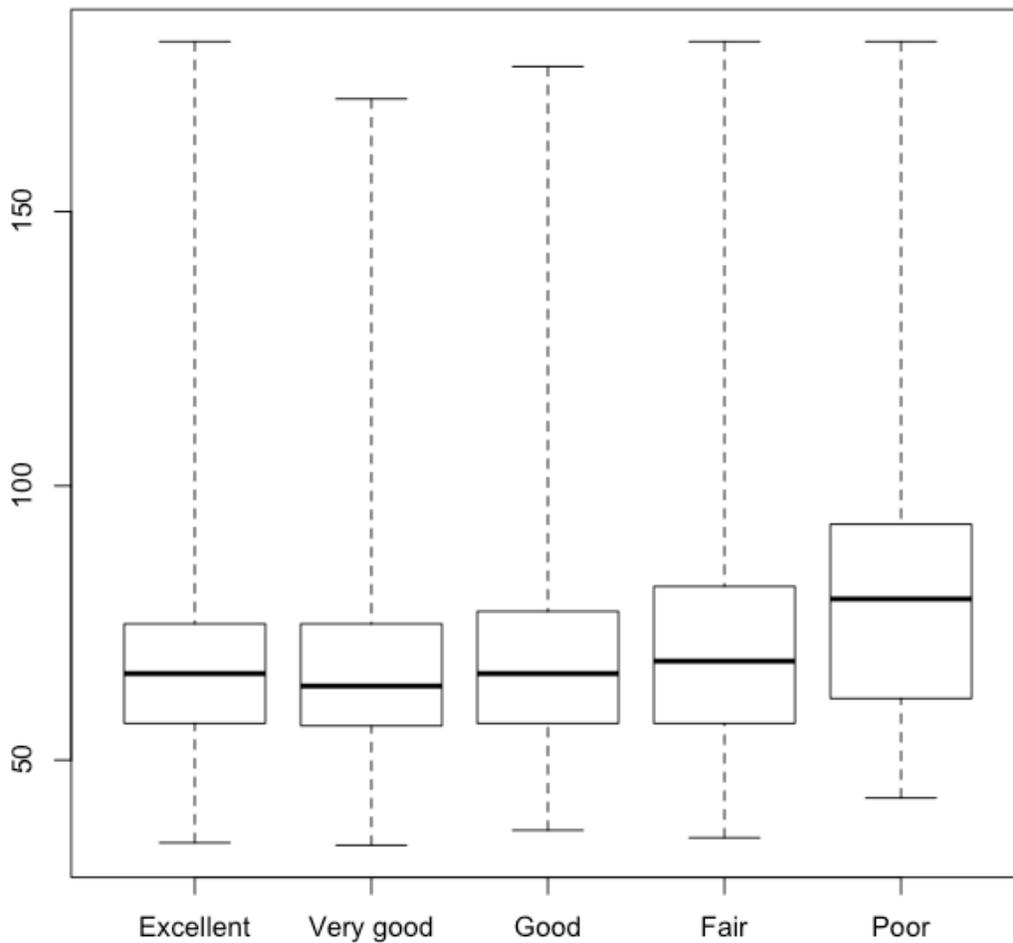


Box Plot

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.