WEEK 3

Exploring Categorical and Quantitative Data

3.1 Week 3 Reading Guide: Introduction to R, Categorical Variables, and a Single Quantitative Variable

Chapter 3 (Applications: Data)

Videos

- Starting_with_R

Notes

R is case sensitive, meaning it reads `data` differently from `Data`. If you get an error message, check that your capitalization is correct.

R does not like spaces or special characters. This means the column and row headers in the data set should not have spaces, periods, commas, etc. Instead of titling the variable `column header`, use `column_header` or `ColumnHeader`.

Tidy data: Data frames should have

1 row per ________________________,

1 column per ___________________.

We highly recommend completing the R/RStudio tutorials in section 3 to help understand R better.

We will not expect you to be able to write full code independently for this course. For Stat 216, you will need to understand types (categorical or quantitative) and roles (explanatory or response) of variables, as well as the structure of data, in order to fill in a few blanks in provided code to graph or analyze data.

Functions

State what these introductory functions do in R:

`glimpse(data_set_name)`

`head(data_set_name)`

`data_set_name$variable_name`

`<-`

`%>%`

Chapter 4 (Exploring categorical data)

Videos

- 4.1
- 4.2
- 4.4

Vocabulary

Frequency table:

Relative frequency table:

Contingency or two-way table:

Association (between two variables):

Unconditional proportion:

Conditional proportion:

Row proportions:

Column proportions:

Statistic:

Sample proportion:

Notation:

Parameter:

Population proportion:

Notation:

Bar plot:

Segmented bar plot:

Mosaic plot:

Simpson's Paradox:

Notes

In a contingency table, which variable (explanatory or response) generally will make the columns of the table? Which variable will make the rows of the table?

In a segmented bar plot, the bars represent the levels of which variable? The segments represent the levels of which variable?

What type of plot(s) are appropriate to display a single categorical variable?

What type of plot(s) are appropriate to display two categorical variables?

What is the difference between a standardized segmented bar plot and a mosaic plot?

True or false: Pie charts are generally highly recommended ways to graphically display categorical data.

True or false: Two categorical variables are associated if the conditional proportions of a particular outcome (typically of the response variable) differ across levels of the other variable (typically the explanatory variable).

True or false: When a segmented bar plot has segments that sum to 1 (or 100%), the segment heights correspond to the proportions conditioned on the **segment**.

Review of Simpson's Paradox

Based on the segmented bar plot in Figure 4.6, which race of defendant was more likely to have the death penalty invoked?

Based on the segmented bar plot in Figure 4.7 and Table 4.9, which race of defendant was more likely to have the death penalty invoked when the victim was Caucasian?

Based on the segmented bar plot in Figure 4.7 and Table 4.9, which race of defendant was more likely to have the death penalty invoked when the victim was African American?

The direction of the relationship between the ____________________ and ____________________ variables is **reversed** when accounting for a ____________________ variable.

Chapter 5 (Exploring quantitative data)

Videos

- 5.2to5.4
- 5.5to5.6
- 5.7

Type of Plots

Scatterplot:

Dot plot:

Histogram:

Density plot:

Box plot:

Vocabulary

Four characteristics of a scatterplot:

Form:

Strength:

Direction:

Unusual observations or outliers:

Distribution (of a variable):

Four characteristics of the distribution of one quantitative variable:

Center:

Variability:

Shape:

Outliers:

Point estimate:

Histogram:

Data density:

Tail:

Skew:

Symmetric:

Modality:

Density plot:

Deviation:

Variance:

Standard deviation:

Boxplot:

Five number summary:

Median:

X^{th} percentile:

Interquartile range (IQR):

Robust statistics:

Notes

What type of plot(s) are appropriate for displaying one quantitative variable?

What type of plot(s) are appropriate for displaying two quantitative variables?

What type of plot(s) are appropriate for displaying one quantitative variable and one categorical variable?

What are the two ways to measure the 'center' of a distribution? Which one is considered robust to skew/outliers?

What are the three ways to measure the 'variability' of a distribution? Which one is considered robust to skew/outliers?

How are variance and standard deviation related?

Fill in the following table with the appropriate notation.

Summary Measure	Parameter	Statistic
Mean		
Variance		
Standard deviation		

How are outliers denoted on a box plot? How can you mathematically determine if a data set has outliers?

3.1.1 Summarizing Chapters 4 and 5

Look at the table of vocabulary terms in the final section of each chapter. If there are any you do not know, be sure to review the appropriate section of your text.

Notes

Statistics summarize ____________________ .
Parameters summarize ____________________.

Fill in the following table with the appropriate notation for each summary measure.

Summary measure	Statistic	Parameter
Sample size		
Proportion (used to summarize one categorical variable)		
Mean (used to summarize one quantitative variable)		
Correlation (used to summarize two quantitative variables)		
Regression line slope (used to summarize two quantitative variables)		

Data visualization summary

Fill in the following table to help associate type of plot for each of several scenarios.

	Appropriate plot(s)
One categorical variable (categorical response, no explanatory)	
One quantitative variable (quantitative response, no explanatory)	
Two categorical variables (categorical response, categorical explanatory)	
One of each (quantitative response, categorical explanatory)	
Two quantitative variables (quantitative response, quantitative explanatory)	

3.2 Activity 3A: Graphing Categorical Variables

3.2.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question involving categorical variables.
- Plots for a single categorical variable: bar plot.
- Plots for association between two categorical variables: segmented bar plot, mosaic plot.

3.2.2 Terminology review

In today's activity, we will review summary measures and plots for categorical variables. Some terms covered in this activity are:

- Proportions
- Bar plots
- Segmented bar plots
- Mosaic plots

To review these concepts, see Chapter 4 in the textbook.

3.2.3 Graphing categorical variables

For today's activity we will begin to use the statistical package R to analyze data through the IDE (integrated development environment) RStudio. For almost all activities and labs it will be necessary to upload the provided R script file from D2L for that day. Follow these steps to upload the necessary R script file for today's activity:

- Download the Myopia Activity R script file from D2L.
- Click "Upload" in the "Files" tab in the bottom right window of RStudio. In the pop-up window, click "Choose File", and navigate to the folder where the Myopia Activity R script file is saved (most likely in your downloads folder). Click "Open"; then click "Ok".
- You should see the uploaded file appear in the list of files in the bottom right window. Click on the file name to open the file in the Editor window (upper left window).

Notice that the first three lines of code contain a prompt called, `library`. Packages needed to run functions in R are stored in directories called libraries. When using the MSU RStudio server, all the packages needed for the class are already installed. We simply must tell R which packages we need for each R script file. We use the prompt `library` to load each **package** (or library) needed for each activity. Note, these `library` lines MUST be run each time you open a R script file in order for the functions in R to work. Before class today you should have worked through an R tutorial to prepare for class and to make sure you can login to the RStudio server. This tutorial will be a great resource as you begin to use R.

Highlight and run lines 1–3 to load the packages needed for today's activity. Notice the use of the # symbol in the R script file. The # sign is not part of the R code. It is used by these authors to add comments to the R code and explain what each call is telling the program to do. R will ignore everything after a # sign when executing the code. Refer to the instructions following the # sign to understand what you need to enter in the code.

Nightlight use and myopia

In a study reported in Nature (Quinn et al. 1999), a survey of 479 children found that those who had slept with a nightlight or in a fully lit room before the age of two had a higher incidence of nearsightedness (myopia) later in childhood.

In this study, there are two variables studied: `Light`: level of light in room at night (no light, nightlight, full light) and `Sight`: level of myopia developed later in childhood (high myopia, myopia, no myopia).

1. Which variable is the explanatory variable? Which is the response variable?

An important part of understanding data is to create visual pictures of what the data represent. In this activity, we will create graphical representations of categorical data.

R code

Throughout these activities, we will often include the R code you would use in order to produce output or plots. These "code chunks" appear in gray. In the code chunk below, we demonstrate how to read the data set into R using the `read.csv()` function. The line of code shown below (line 6 in the R script file) reads in the data set and names the data set `myopia`. Highlight and run line 6 in the R script file to load the data from the Stat 216 webpage.

```r
# This will read in the data set
myopia <- read.csv("https://math.montana.edu/courses/s216/data/ChildrenLightSight.csv")
```

2. Click on the data set name (`myopia`) in the Environment tab (upper right window). This will open the data set in a 2nd tab in the Editor window (upper left window). R is case sensitive, which means that you must always type the name of a variable EXACTLY as it is written in the data set including upper and lower case letters and without misspellings! Write down the name of each variable (column names) as it is written in the data set.

Displaying a single categorical variable

If we wanted to know how many children in our data set were in each level of myopia, we could create a frequency bar plot of the variable `Sight`. In the R script file, enter the variable name, `Sight` (*note the capital* S), for `variable` into the `ggplot` code at line 10. Highlight and run lines 9–15 to create the plot. Note: this is a **frequency** bar plot plotting counts (the number of children in each level of sight is displayed on the y-axis).

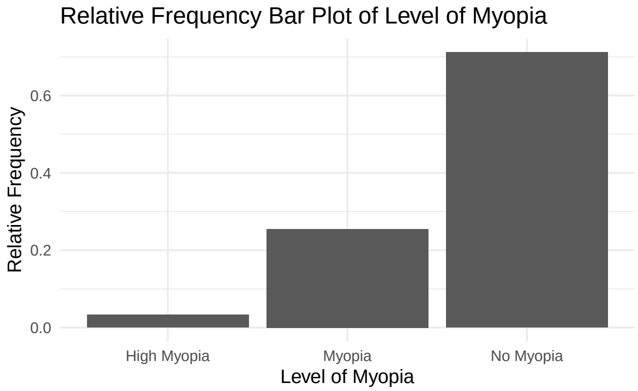
```r
myopia %>% # Data set piped into...
ggplot(aes(x = variable)) +   # This specifies the variable
  geom_bar(stat = "count") +  # Tell it to make a bar plot
  labs(title = "Frequency Bar Plot of Level of Myopia",  # Give your plot a title
       x = "Level of Myopia",   # Label the x axis
       y = "Frequency")  # Label the y axis
```

3. Sketch the bar chart created below. Be sure to label the axes.

4. Using the bar chart created, estimate how many children have some level of myopia.

We could also choose to display the data as a proportion in a **relative frequency** bar plot. To find the relative frequency, divide the count in each level of myopia by the sample size. These are sample proportions. Notice that in this code we told R to create a bar plot with proportions.

```
myopia %>% # Data set piped into...
ggplot(aes(x = Sight)) +   # This specifies the variable
  geom_bar(aes(y = ..prop.., group = 1)) +  # Tell it to make a bar plot with proportions
  labs(title = "Relative Frequency Bar Plot of Level of Myopia",  # Give your plot a title
       x = "Level of Myopia",   # Label the x axis
       y = "Relative Frequency")  # Label the y axis
```



5. Which features in the relative frequency bar plot are the same as the frequency bar plot? Which are different?

Displaying two categorical variables

Is there an association between the level of light in a room and the development of myopia? To examine the differences in level of myopia for the level of light, we would create a segmented bar plot of `Light` segmented by `Sight`. To create the segmented bar plot enter the variable name, `Light` for `explanatory` and the variable name, `Sight` for `response` in the R script file in line 26. Highlight and run lines 25–31.

```
myopia %>% # Data set piped into...
ggplot(aes(x = explanatory, fill = response)) +   # This specifies the variables
  geom_bar(stat = "count", position = "fill") +  # Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Night Light Use by Level of Myopia",
       # Make sure to title your plot
       x = "Level of Light",   # Label the x axis
       y = "")  # Remove y axis label
```

6. Sketch the segmented bar plot created here. Be sure to label the axes.

7. From the segmented bar plot, estimate the proportion of no myopia for those that used a nightlight.

8. Which level of light has the highest proportion of `No Myopia`?

We could also plot the data using a mosaic plot. Fill in the variable name, `Light` for `explanatory` and the variable name, `Sight` for `response` in line 36 in the R script file. Highlight and run lines 34–40.

```
myopia %>% # Data set piped into...
  ggplot() +   # This specifies the variables
  geom_mosaic(aes(x=product(explanatory), fill = response)) +  # Tell it to make a mosaic plot
  labs(title = "Mosaic Plot of Night Light Use by Level of Myopia",
       # Make sure to title your plot
       x = "Level of Light",   # Label the x axis
       y = "")  # Remove y axis label
```

9. What is similar and what is different between the segmented bar chart and the mosaic bar chart?

10. Explain why the bar for `Nightlight` is the widest in the mosaic plot.

Fill in the name of the explanatory variable and the response variable in line 43 in the R script file, highlight and run line 43 to get the counts for each combination of levels of variables.

```
myopia %>% group_by(explanatory) %>% count(response)
```

11. Fill in the following table with the values from the R output.

	Light Level			
Myopia Level	Full Light	Nightlight	No Light	Total
High Myopia				
Myopia				
No Myopia				
Total				

In the following questions, use the table to calculate the described proportions. Notation is important for each calculation. Since this is sample data, it is appropriate to use statistic notation for the proportion, $\hat{p}$. When calculating a proportion dependent on a single level of a variable, subscripts are needed when reporting the notation.

12. Calculate the proportion of children with high myopia. Use appropriate notation.

13. Calculate the proportion of children with high myopia among those that slept with full light. Use appropriate notation.

14. Calculate the proportion of children with high myopia among those that slept with no light. Use appropriate notation.

15. Calculate the difference in proportion of children with high myopia for those that slept with full light minus those who slept with no light. Give the appropriate notation. Use full light minus no light as the order of subtraction.

3.2.4 Take-home messages

1. Bar charts can be used to graphically display a single categorical variable either as counts or proportions. Segmented bar charts and mosaic plots are used to display two categorical variables.
2. Segmented bar charts always have a scale from 0 - 100%. The bars represent the outcomes of the explanatory variable. Each bar is segmented by the response variable. If the heights of each segment are the same for each bar there is no association between variables.
3. Mosaic plots are similar to segmented bar charts but the widths of the bars also show the number of observations within each outcome.

3.2.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

3.3 Activity 3B: IMDb Movie Reviews — Displaying Quantitative Variables

3.3.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question for quantitative data.
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.

3.3.2 Terminology review

In today's activity, we will review summary measures and plots for quantitative variables. Some terms covered in this activity are:

- Two measures of center: mean, median
- Two measures of spread (variability): standard deviation, interquartile range (IQR)
- Types of graphs: box plots, dot plots, histograms
- Identify and create appropriate summary statistics and plots given a data set or research question for a single categorical and a single quantitative variable.
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.
- Given a plot or set of plots, describe and compare the distribution(s) of a single quantitative variable (center, variability, shape, outliers).

To review these concepts, see Section 2.3 in the textbook.

3.3.3 Movies released in 2016

A data set was collected on movies released in 2016 ("IMDb Movies Extensive Dataset" 2016). Here is a list of some of the variables collected on the observational units, movies released in 2016.

Variable	Description
`budget_mil`	Amount of money (in US $ millions) budgeted for the production of the movie
`revenue_mil`	Amount of money (in US $ millions) the movie made after release
`duration`	Length of the movie (in minutes)
`content_rating`	Rating of the movie (`G`, `PG`, `PG-13`, `R`, `Not Rated`)
`imdb_score`	IMDb user rating score from 1 to 10
`genres`	Categories the movie falls into (e.g., Action, Drama, etc.)
`facebook_likes`	Number of likes a movie receives on Facebook

Summarizing a single quantitative variable

The `favstats()` function from the `mosaic` package gives the summary statistics for a quantitative variable. Here we have the summary statistics for the variable `imdb_score`. The summary statistics give the two measures of center and two measures of spread for IMDb score. Highlight and run lines 1 – 8 in the provided `R` script file to load the data set. Check that the summary statistics match that printed in the coursepack.

```
# Read in data set
movies <- read.csv("https://math.montana.edu/courses/s216/data/Movies2016.csv")
movies %>% # Data set piped into...
  summarise(favstats(imdb_score)) # Apply favstats function to imdb_score
```

```
#>   min   Q1 median  Q3 max     mean       sd  n missing
#> 1 3.4 5.65    6.4 7.1 8.2 6.309783 1.086689 92       0
```

1. Give the values for the two measures of center (mean and median).

2. Report the value for quartile 1 and interpret this value in context of the problem.

3. Calculate the interquartile range (IQR = Q3 - Q1).

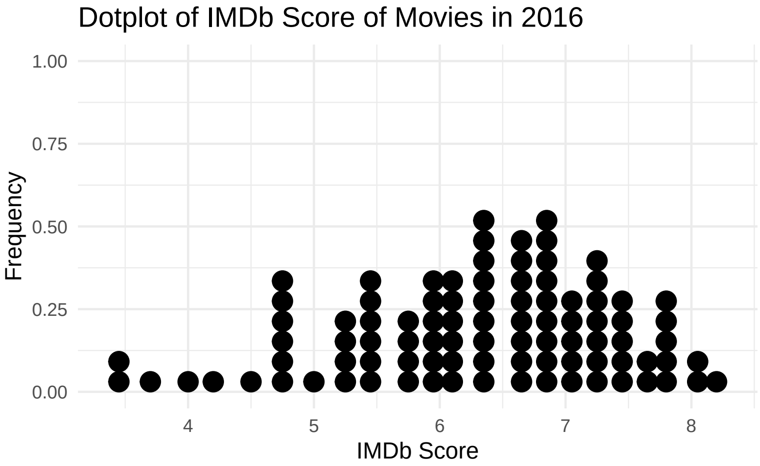
4. Report the value of the standard deviation and interpret this value in context of the problem.

Displaying a single quantitative variable

5. What are the three types of plots used to plot a single quantitative variable?

A dotplot will plot a dot for each value in the data set. The following code will create a dotplot of IMDb scores. Notice that we put in the variable name `imdb_score` for `x =` in the ggplot function.

```
movies %>% # Data set piped into...
ggplot(aes(x = imdb_score)) +   # Name variable to plot
  geom_dotplot() +  # Create dotplot
  labs(title = "Dotplot of IMDb Score of Movies in 2016", # Title for plot
       x = "IMDb Score", # Label for x axis
       y = "Frequency") # Label for y axis
```



6. What is the shape of the distribution of IMDb scores?

To create a histogram of the IMDb scores, enter the variable name, `imdb_score` in the provided `R` script file for `variable` at line 20, highlight and run lines 19–24. Visually, this shows us the range of IMDb scores for Movies released in 2016.

Notice that the **bin width** is 0.5. For example the first bin consists of the number of movies in the data set with an IMDb score of 3.25 to 3.75. It is important to note that a movie with a IMDb score on the boundary of a bin will fall into the bin above it; for example, 4.75 would be counted in the bin 4.75–5.25.

```
movies %>% # Data set piped into...
ggplot(aes(x = variable)) +   # Name variable to plot
  geom_histogram(binwidth = 0.5) +  # Create histogram with specified binwidth
  labs(title = "Histogram of IMDb Score of Movies in 2016", # Title for plot
       x = "IMDb Score", # Label for x axis
       y = "Frequency") # Label for y axis
```

7. Sketch the histogram created here.

8. Which range of IMDb scores have the highest frequency?

To create a boxplot of the IMDb scores, enter the variable name, `imdb_score` in the provided `R` script file for `variable` at line 28, highlight and run lines 27–32.

```
movies %>% # Data set piped into...
ggplot(aes(x = variable)) +    # Name variable to plot
  geom_boxplot() +  # Create boxplot
  labs(title = "Boxplot of IMDb Score of Movies in 2016", # Title for plot
       x = "IMDb Score", # Label for x axis
       y = "Frequency") # Label for y axis
```

9. Sketch the boxplot created and identify the values of the 5-number summary (minimum value, first quartile (Q_1), median, third quartile (Q_3), maximum value) on the plot. Use the following formulas to find the invisible fence on both ends of the distribution. Draw a dotted line at the invisible fence to show how the outliers were found.

$$\text{Lower Fence: values} \leq Q_1 - 1.5 \times IQR$$

$$\text{Upper Fence: values} \geq Q_3 + 1.5 \times IQR$$

10. Compare the three graphs of IMDb scores created above.

    Which graph is best used to show the shape of the distribution?

    Which graph is best used to show the outliers of the distribution?

Summary statistics for a single categorical and single quantitative Variable

Is there an association between content rating and budget for movies in 2016? To use the `favstats()` function in the mosaic package with two variables, we will enter the variables as a formula, response~explanatory. This function will give the summary statistics for budget for each content rating. Highlight and run lines 35–37 in the provided `R` script file and check that the summary statistics match those provided in the coursepack.

```
movies %>% # Data set piped into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  summarise(favstats(budget_mil~content_rating)) # Find the summary measures for each content rating
```

```
#>   content_rating min    Q1 median      Q3 max     mean       sd  n missing
#> 1             PG 0.5 11.00   74.0 151.250 175 86.54167 71.52795 12       0
#> 2          PG-13 0.0 17.25   33.5 138.750 250 74.17500 74.15190 46       0
#> 3              R 0.0  7.75   19.5  29.625  60 21.09375 16.99926 32       0
```

11. Which content rating has the largest IQR?

12. Report the mean budget amount for the PG rating. Use appropriate notation.

13. Report the mean budget amount for the R rating. Use appropriate notation.

14. Calculate the difference in mean budget amount for movies in 2016 with a PG rating minus those with a R rating. Use appropriate notation with informative subscripts.

Displaying a single categorical and single quantitative variable

The boxplot of movie budgets (in millions) by content rating is plotted using the code below. Enter the variable `budget_mil` for `response` and the variable `content_rating` for explanatory at line 42, highlight and run code lines 40–46. This plot compares the budget for different levels of content rating.

```
movies %>%  # Data set piped into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  ggplot(aes(y = response, x = explanatory))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Side by side box plot of budget by content rating",  # Title
       x = "Content Rating",    # x-axis label
       y = "Budget (in Millions)")  # y-axis label
```

15. Sketch the box plots created using the `R` code.

16. Answer the following questions about the box plots created.

    a. Which content rating has the highest center?

    b. Which content rating has the largest spread?

    c. Which content rating has the most skewed distribution?

    d. Fifty percent of movies in 2016 with a PG-13 content rating fall below what value? What is the name of this value?

17. Which variable is the explanatory variable? Response variable?

3.3.4 Take-home messages

1. Histograms, box plots, and dot plots can all be used to graphically display a single quantitative variable.

2. The box plot is created using the five number summary: minimum value, quartile 1, median, quartile 3, and maximum value. Values in the data set that are less than $Q_1 - 1.5 \times IQR$ and greater than $Q_3 + 1.5 \times IQR$ are considered outliers and are graphically represented by a dot outside of the whiskers on the box plot.

3. Data should be summarized numerically and displayed graphically to give us information about the study.

4. When comparing distributions of quantitative variables we look at the shape, center, spread, and for outliers. There are two measures of center: mean and the median and two measures of spread: standard deviation and the interquartile range, $IQR = Q_3 - Q_1$.

3.3.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

3.4 Week 3 Lab: IPEDs

3.4.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question for quantitative data.
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.

3.4.2 Terminology review

In today's lab, we will review summary measures and plots for quantitative variables. Some terms covered in this activity are:

- Two measures of center: mean, median
- Two measures of spread (variability): standard deviation, interquartile range (IQR)
- Types of graphs: box plots, dot plots, histograms
- Identify and create appropriate summary statistics and plots given a data set or research question for a single categorical and a single quantitative variable.
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.
- Given a plot or set of plots, describe and compare the distribution(s) of a single quantitative variable (center, variability, shape, outliers).

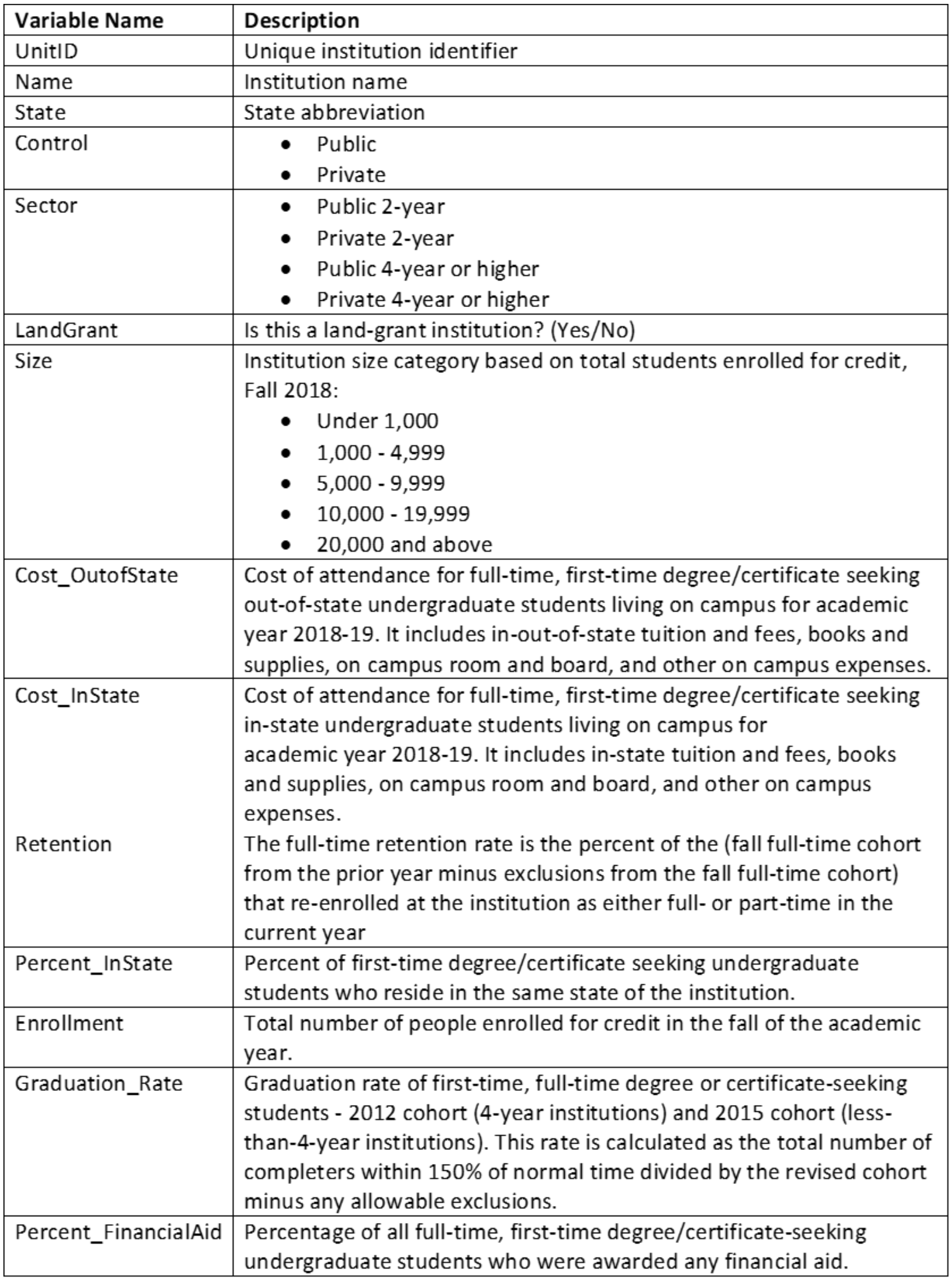
To review these concepts, see Chapter 5 in the textbook.

3.4.3 The Integrated Postsecondary Education Data System (IPEDS)

Upload and open the provided R script file for the week 3 lab to answer the following questions. **Remember bolded questions will be answered on Gradescope for your group.**

These data are on a subset of institutions that met the following selection criteria (Education Statistics 2018):

- Degree granting
- United States only
- Title IV participating
- Not for profit
- 2-year or 4-year or above
- Has full-time first-time undergraduates
- Note that several variables have missing values for some institutions (denoted by "NA").



Variable Name	Description
UnitID	Unique institution identifier
Name	Institution name
State	State abbreviation
Control	• Public • Private
Sector	• Public 2-year • Private 2-year • Public 4-year or higher • Private 4-year or higher
LandGrant	Is this a land-grant institution? (Yes/No)
Size	Institution size category based on total students enrolled for credit, Fall 2018: • Under 1,000 • 1,000 - 4,999 • 5,000 - 9,999 • 10,000 - 19,999 • 20,000 and above
Cost_OutofState	Cost of attendance for full-time, first-time degree/certificate seeking out-of-state undergraduate students living on campus for academic year 2018-19. It includes in-out-of-state tuition and fees, books and supplies, on campus room and board, and other on campus expenses.
Cost_InState	Cost of attendance for full-time, first-time degree/certificate seeking in-state undergraduate students living on campus for academic year 2018-19. It includes in-state tuition and fees, books and supplies, on campus room and board, and other on campus expenses.
Retention	The full-time retention rate is the percent of the (fall full-time cohort from the prior year minus exclusions from the fall full-time cohort) that re-enrolled at the institution as either full- or part-time in the current year
Percent_InState	Percent of first-time degree/certificate seeking undergraduate students who reside in the same state of the institution.
Enrollment	Total number of people enrolled for credit in the fall of the academic year.
Graduation_Rate	Graduation rate of first-time, full-time degree or certificate-seeking students - 2012 cohort (4-year institutions) and 2015 cohort (less-than-4-year institutions). This rate is calculated as the total number of completers within 150% of normal time divided by the revised cohort minus any allowable exclusions.
Percent_FinancialAid	Percentage of all full-time, first-time degree/certificate-seeking undergraduate students who were awarded any financial aid.

Summary statistics for a single quantitative variable

Look through the provided chart above showing the description of variables measured. The UnitID and Name are identifiers for each observational unit, *US degree granting institutions in 2018.*

1. Identify in the chart above which variables collected on the US institutions are categorical (C) and which variables are quantitative (Q).

In Wednesday's activity, the code was provided to import the data set needed directly from the Stat 216 website. Follow these steps to upload and import the data set for today's lab.

- Download the provided data set `IPEDS_Data_2018` from D2L
- Upload the data set `IPEDS_Data_2018` to the RStudio server using the same steps to upload the R script file.
- Click on "Import Dataset" in the Environment tab in the upper right hand corner.
- Choose "From Text(base)" in the drop-down menu and select the correct csv file.
- Be sure that "Yes" is selected next to "Heading" in the pop-up screen. Click "Import".
- To view the data set, click on the data set name (`IPEDS_Data_2018`). Verify that that column names match the first column in the chart on the previous page. If the columns are named V1, V2, V3…etc, you did not select "Yes" for "Heading".

Enter the name of the data set (see the environment tab) for `datasetname` in the R script file in line 6. We will look at the retention rates for the 4-year institutions only. Enter the variable name `Retention` for `variable` in line 12. Highlight and run lines 1 – 12. **Note that the two lines of code (lines 8 and 10) are filtering to remove the 2-year institutions so we are only assessing Public 4-year and Private 4-year institutions.** The `favstats()` function from the `mosaic` package gives the summary statistics for a quantitative variable. The summary statistics give the two measures of center and two measures of spread for retention rate.

```r
IPEDS <- datasetname #Creates the object IPEDS
IPEDS <- IPEDS %>%
  filter(Sector != "Public 2-year") #Filters the data set to remove Public 2-year
IPEDS <- IPEDS %>%
  filter(Sector != "Private 2-year") #Filters the data set to remove Private 2-year
IPEDS %>%
  summarise(favstats(variable)) #Gives the summary statistics
```

2. **Report the value for quartile 3 and interpret this value in context of the study.**

3. Calculate the interquartile range ($IQR = Q_3 - Q_1$) for this study.

4. How many missing values are there? What does this indicate?

Displaying a single quantitative variable

We will create both a histogram and a boxplot of the variable `Retention`. Enter the name of the variable in both line 16 and line 23 for `variable` in the R script file. **Give each plot a descriptive title.** Highlight and run lines 15 – 27 to give the histogram and boxplot. Notice that the **bin width** for the histogram is 10. For example, the first bin consists of the number of 4-year institutions in the data set with a retention rate of 0 to 10%. It is important to note that a 4-year institution with a retention rate on the boundary of a bin will fall into the bin above it; for example, 10 would be counted in the bin 10–20.

Export and upload both plots to Gradescope for your group.

- To export the graphs: in the bottom right corner in the Plots tab, click on `Export`.
- Then choose `Save as Image`. Save the image as a png. This will save your graph to the server.
- In the Files tab, click on the box next to your saved image file, click `More` and choose `Export`. This will save your file to your downloads folder on your computer.

```
IPEDS %>% # Data set piped into...
ggplot(aes(x = variable)) +   # Name variable to plot
  geom_histogram(binwidth = 10) +  # Create histogram with specified binwidth
  labs(title = "Title", # Title for plot
       x = "Rentention Rate", # Label for x axis
       y = "Frequency") # Label for y axis
```

```
IPEDS %>% # Data set piped into...
ggplot(aes(x = variable)) +   # Name variable to plot
  geom_boxplot() +  # Create boxplot
  labs(title = "Title", # Title for plot
       x = "Retention Rates", # Label for x axis
       y = "Frequency") # Label for y axis
```

5. What is the shape of the distribution of retention rates?

6. Identify any outliers in the data set.

Robust Statistics

Let's examine how the presence of outliers affect the values of center and spread.

7. Report the two measures of center (mean and median) for retention rates given in the R output.

8. Report the two measures of spread (standard deviation and IQR) for retention rates given in the R output.

To show the effect of outliers on the measures of center and spread, the smallest values of retention rate in the data set were increased by 30%. Highlight and run lines 30–38.

```
IPEDS %>% # Data set piped into...
  summarise(favstats(Retention_Inc))
```

```
IPEDS %>% # Data set piped into...
  ggplot(aes(x = Retention_Inc)) +   # Name variable to plot
  geom_boxplot() +  # Create histogram with specified binwidth
  labs(title = "Boxplot of Retention Rates for US Higher Education Institutions", # Title for plot
       x = "Retention Rate", # Label for x axis
       y = "Frequency") # Label for y axis
```

9. Report the two measures of center for this new data set.

10. Report the two measures of spread for this new data set.

11. **Which measure of center is robust to (not affected by) outliers? Explain your answer.**

12. Which measure of spread is robust to outliers? Explain your answer.

Summarizing a single categorical and single quantitative variable

Is there a difference in retention rates for public and private 4-year institutions? In the next part of the activity we will compare retention rates for public and private 4-year institutions. Note that this variable (public or private) is labelled `Control` in the data set.

13. **Based on the research question, which variable will we treat as the explanatory variable? Response variable?**

Enter the name of the explanatory variable and the name of the response variable in lines 42 and 45 of the R script file. Remember that the variable name must be typed in EXACTLY as it is written in the data set. Highlight and run lines 41 – 49 to find the summary statistics and create side by side boxplots of the data.

```
IPEDS %>%  # Data set piped into...
  summarise(favstats(response~explanatory)) # Summary statistics for retention rates by sector
```

```
IPEDS %>%  # Data set piped into...
  ggplot(aes(y = response, x = explanatory))+  # Identify variables
  geom_boxplot()+  # Create box plot
  labs(title = "Side by side box plot of retention rates by control",  # Title
       x = "Control",    # x-axis label
       y = "Retention Rates")  # y-axis label
```

14. **Compare the two boxplots.**

    Which type of university has the highest center?

    Largest spread?

    What is the shape of each distribution?

    Does either distribution have outliers?

15. Report the difference in mean retention rates for private and public universities. Use private minus public as the order of subtraction. Use the appropriate notation.

16. Does there appear to be an association between retention rates and type of university? Explain your answer.

Summarizing two categorical variables

Are private 4-year institutions smaller than public one? The following set of code will create a segmented bar plot of size of the institution by sector. Enter the variable `Sector` for explanatory and `Size` for response in line 53. Highlight and run lines 52 - 57 in the R script file.

```
IPEDS %>%
  ggplot(aes(x=explanatory, fill = response)) + # Enter the explanatory and response variables
  geom_bar(stat = "count", position = "fill") + # Create a segmented bar plot
  labs(title = "Segmented Bar Plot of Sector by Size", # Title
       x = "Sector", # x-axis label
       y = "") # remove y-axis label
```

17. Does there appear to be an association between sector and size of 4-year institutions? Explain your answer using the plot.