R PROJECT

You can work in groups of three. Please turn in one assignment with all names at the top, in bold.

Final write-up is due in class Sunday November 29th. Your final should be written in an R Markdown file that contains all the R code in `blocks` (“‘{r} CODE_HERE “‘). You will need to `knit` a pdf/html including your written answers, R code and R output where asked to do so.

Often data comes unstructured. There is a bit of work to do before you can readily apply the concepts learned in this course. To receive credit for a project you will complete the following.

1. Load a data set into R. This can be any data set that has at least two quantitative variables and one qualitative variable and can be used to complete the rest of this assignment.
    - If you have an interest (chess, avocados, climate change), this is a great opportunity to explore that topic. Or just browse some data sets like this one. I listed potential data sources at the end of this pdf. Look for data that is “.csv” file type.
    - If you do not have an interest in mind, you can use the default data set from the 2016 American Community Survey from Colorado. It is located in the example final zip file with the name “ACS_2016_CO.csv” with an accompanying codebook describing the variable values. NOTE: the default data set is nice to have, but a lot of the variables are top coded as 9999* because the information is missing. This isn’t valuable data when calculating statistics or creating plots and should not be used! (the command subset(-) in R is helpful for dealing with this)
2. Describe the data set with words. How many variables are there? How many observations? What is the unit of observation for the data set (person-year, state, month)? Is this a cross-section, or multiple observations overtime? Do we have repeated observations for the same subject? Describe a few key variables that you will use in your data set including their units (feet, miles, $).
3. Summarize one of the quantitative variables for the full sample using sample statistics. Then, summarize the same quantitative variable for a subset the observations that meet a specific condition. (e.g. report the average, and standard deviation, monthly price of avocados in 20 major cities in the US from 2010 to 2015, then summarize the avocado price for all 20 cities only in the month of February). Try to choose the subsample in a way that is meaningful. How do the summary statistics compare and what do you learn from that? Include R code and output here, along with your interpretation.
4. Create a histogram of the variable that you summarize in part 3 with properly labeled axis and title. (Bonus points if you can create two histograms of the same variable, split by some other variable, that are strikingly different). Include R output here, along with your interpretation.
5. Calculate a confidence interval. You can choose to calculate a confidence interval of one variable of a

difference of means confidence interval. Pick something interesting to you and interpret your findings. Include your R code here and output here, along with your interpretation.

6. Formalize a hypothesis you wish to test with these data (e.g. is the average salary from men the same as the average salary for women?). You might not have all the knowledge to test the exact hypothesis you are interested in. You will mostly be interested in doing a difference of means test. Or if the mean of a variable is equal to a specific value.

7. Conduct the hypothesis at the α=0.05 level of significance and interpret your results in a meaningful way. Include your R code and output here, along with your interpretation.

8. Create a scatterplot from the data set, with an appropriate title, axis labels, and legend. The goal is for this image to tell a story that is clear to the reader. Include R code and output here, along with your interpretation.

9. Finally, think and write about who would be a good "consumer" of this information. Who would be interested in the facts you present here, and how you could improve the analysis in the future by incorporating new data or using the existing data to answer a more interesting question.

In terms of grading, I will be going through the following rubric:

Section	Points
Formatting	10
Dataset description	10
Summary statistics	10
Histogram	10
Confidence interval	15
Formalize hypothesis	5
Conduct hypothesis	15
Visualization	15
Write up	10

POTENTIAL DATA SOURCES (IF YOU DO NOT CARE TO USE THE DEFAULT ACS DATA)

Note: if you'd like to use your own dataset, you need to get it approved by me beforehand.

There are many good resources to find data online:

- Kaggle Datsets: https://www.kaggle.com/datasets
- Google's dataset search: https://toolbox.google.com/datasetsearch

- Data is plural structured archive, list of interesting data: https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit
- Bureau of Labor Statistics, prices, unemployment: https://www.bls.gov/data/
- Five Thirty Eight project data: https://github.com/fivethirtyeight/data
- Energy Information Agency: https://www.eia.gov/
- Census data at IPUMS: https://www.ipums.org/
- Economics data at the Federal Reserve: https://fred.stlouisfed.org/
- Economic history data: http://eh.net/databases/
- Bureau of Economic Analysis: https://www.bea.gov/data
- Agricultural data at USDA: https://www.ers.usda.gov/data-products/

R MARKDOWN

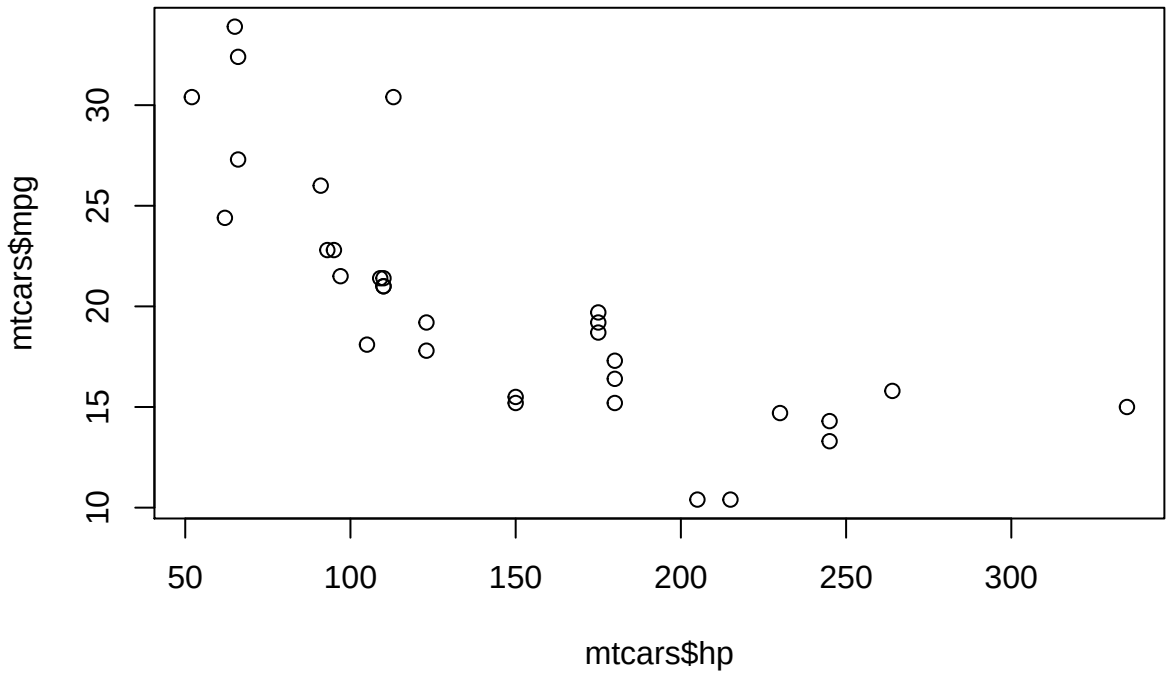
For more details on using R Markdown see http://rmarkdown.rstudio.com.

One thing you haven't learned how to do in Rmd from the homework, is to not include code you don't want your reader to see. For example, you don't want a reader to have to scroll through all the code of loading your data and cleaning it.

At the beginning of code blocks "‘‘‘{r}", you can add a bunch of options to change what gets displayed.

1. You can add the option "‘‘‘{r, include= FALSE}" which will not display the code nor any output produced, i.e. the reader will never know that the code was run.
2. You can add the option "‘‘‘{r, echo= FALSE}" which will not display the code, but will display any output produced by the code (`summary()` or `plot()`/`hist()`).

The below graph was produced with `echo= FALSE` option



The below graph was produced without the `echo= FALSE` option

```
plot(mtcars$hp, mtcars$mpg)
```