R Project

Imagine that you are tasked with the job to create a report using a dataset that will be turned in to your work supervisor. To that end, you will choose a dataset you find interesting to analyze. Although you will have discretion in choosing the dataset and what variables you report on, you should focus on producing content that is interesting and meaningful (don't just pick a variable at random and plot the distribution; try to learn something). The analysis should be written in full sentences in a professional manner.

You can work in groups of three. Please turn in one assignment with all names at the top, in bold. Each individual should submit an `.Rmd` file and the knitted `.html` file by Sunday, November 28th. Unlike in previous assignments, you should use the chunk option `echo = FALSE` to hide all the code from the final output. This way the final document will just contain text and the plots you create.

To receive credit for a project you will complete the following:

1. Load a data set into R. See the section on data below for requirements on the data.
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 over time? 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, $). *Note* This should be highly personalized to the data at hand.
3. Summarize one of the quantitative variables for the full sample using sample statistics. Then, summarize the same quantitative variable for a subset of 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 for the full sample and the subsample? What *practical* information do you learn from that?
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).
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.
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?). This can either be a one-mean or a difference-in-means test. Write why this test is important for the stakeholder to learn about.
7. Conduct the hypothesis at the $\alpha = 0.05$ level of significance and interpret your results in a *meaningful* way.
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. Add to the plot a line of best fit to try and describe the relationship in the data. Then in the text below the figure, explain information gleaned from the plot.

Potential data sources

Note: You/Your Group need to get your dataset approved by me beforehand. You can just send an email that describes the dataset, i.e. what's a unit of observation, and a link so I can view the data.

The most important data requirement is that observations should be at the individual level.

Good Examples:

- Sports with individual players
- Movies with review
- Survey data

Bad Examples

- Country averages
- Aggregated data (e.g. mean responses to survey)

If you have an interest (chess, sports, movies, climate change, crypto, etc.), this is a great opportunity to explore that topic. If nothing in particular comes to mind, I listed potential data sources at the end of this pdf. Look for data that is .csv file type. If you can't find a data set you want to use, come talk to me and I will help you find one.

There are many good resources to find data online:

- Kaggle Datsets
- Google’s dataset search
- Sports Reference
- Data is plural structured archive
- Bureau of Labor Statistics, prices, unemployment
- Five Thirty Eight project data
- Energy Information Agency
- Census data at IPUMS
- Economics data at the Federal Reserve
- Economic history data
- Bureau of Economic Analysis
- Agricultural data at USDA

Grading

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

	Points		Points
Professional Write up	20%	Confidence interval	15%
Dataset description	10%	Formalize hypothesis	50%
Summary statistics	10%	Conduct hypothesis	15%
Histogram	10%	Visualization	15%

A few comments:

- A full 20% of your grade will be determined by how professional the final report looks and reads.
- For the statistical requirements (testing, plotting, summarizing, etc.), it is not enough to just do it. I want to see you make choices that are interesting. For example, don't just test if the average age of your sample is significantly different from zero, because of course they are!!