# **Final Project – Instructions guidelines and deadlines**

The final project for our class will be completed in groups. The group should work on selecting data of their interest, explore them/review them, analyze them, and answer one or more claims/theses of their choice.

Here is an overview of the Final Project:

For this assignment, the goal is to explore online data repositories to find an interesting and suitable data set to complete a basic exploratory data analysis, and run statistical modeling of your choice with the goal to answer a hypothesis that you will set ahead of time.

Take some time to look through the data repositories below. Once you have looked through the data sets, choose one that interests you whether due to the topic or the contents.

Once you have chosen your data, you should complete the steps of statistical analysis as we have already review them in class.

**Framework**

1. Set your hypothesis, have a specific goal.
2. Review your data to ensure that they are appropriate and complete and can help you prove or disprove your hypothesis.
3. Complete literature review on the subject/hypothesis and determine if there is any relevant research/study has been already completed. Study the literature and include citations in your final report.
4. Once you ensure that your data are sufficient and you can initially rely on them to run your modeling, clean them up.

**Exploratory Analysis**

1. Run Exploratory analysis on at least 5 total variables where 2 of them are quantitative.
   1. Check for missing data
   2. Check for outliers, IQR, and summarize the statistics.
   3. Disect your variables in a way that will help you with your analysis.
   4. Determine the distribution ( if any that your data follow, experimentally and theoretically)
   5. Show your analysis in both tables/charts and visually ( histograms, qqnorm plots, boxplots etc…

**Statistical Modeling**

1. Determine correlations, do correlation comparisons ( technically and visually ) use both plots or pairs for your graphical representations. Split your graphs in ways that will help you to conclude and infer based on your model.
2. Your model and main hypothesis should be answered either using ANOVA or Regression Analysis, or both. – This may mean, depending on your data, that you may need to use a categorical variable to dissect your data, and that you may need to have data with many more than just 2 quantitative variables.

**Report**

The report should be written in R Markdown and then transformed to HTML.  Your report needs to have the following sections:

* Introduction – In this section, explain briefly the purpose of your analysis. Identify your hypothesis, and in a single sentence refer to the results of your work.
* Data - a section describing the data set and how you loaded and transformed it in R.  Include R code blocks within your comments and explain what the code is doing.
* Analysis - walk through the analysis that you performed. Include R code blocks within your comments and explain what the code is doing.
* Issues – Refer to any issues you had with collecting your data, cleaning your data, or implementing your model.
* Results - any plots, tables, or other results which gave you the answers to the questions that you were asked. Include R code blocks for any plots and explain what the code is doing.
* Discussion - a brief (one or two paragraphs) discussion of the results and how they validate or not your initial claims.

**Presentation**

You should prepare a brief (5-8 minutes) presentation on your project

You should create a Powerpoint presentation on your report.

Your Presentation should be narrated, or you should create a recorded video, presenting the Powerpoint.

The presentation should include:

* Introduction - quick reference to the problem you are analyzing and your data
* Hypothesis – main claim or claims you worked on
* Methodology - what statistical tools you used to complete your analysis and why?
* Issues - any issues you had while you run your analysis and how did you resolve them
* Tables, Charts and Graphs and measures with explanation for your EDA
* Results of Correlation, Regression, and ANOVA analysis
* Conclusions
* Future questions and further analysis not addressed in this project
* References

**DataBases / DataSets and Data Repositories:**

There are a number of websites to find data online:

* + Governments - more cities and countries are becoming part of the "open data" movement and publishing data about their operations to the web.  For example, Chicago, IL (home of the SPS Loop campus!) publishes quite a bit of data for anyone to download and analyze on its data portal : [https://data.cityofchicago.org](https://data.cityofchicago.org/)
  + Worldwide organizations - Orgs like the United Nations ([data explorer.](http://data.un.org/Explorer.aspx)) and the World Bank ([open data catalog.](http://datacatalog.worldbank.org/)) offer data sets collected from countries around the world covering crime, demographics, financial data, environmental impact, etc.
  + The most famous is Kaggle which bills itself as the "The Home of Data Science". Check out [https://www.kaggle.com/datasets](https://www.kaggle.com/datasets?utm_medium=paid&utm_source=google.com+search&utm_campaign=datasets&gclid=Cj0KCQjw9fntBRCGARIsAGjFq5HPuFci4hJUEAFIsZyYg0V0Na1Xja_hsg8X01O6T_Et3CXBrVLKzdcaAtJGEALw_wcB)
  + KDNuggets are websites and communities which surround all things data.  All of the data sets listed in this article are public use and most are available from a sharing site called GitHub (another great place to poke around for data): [Awesome Public Datasets on GitHub](http://www.kdnuggets.com/2015/04/awesome-public-datasets-github.html).
  + University Repositories: [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/)
  + Bureau of labor statistics: <https://www.bls.gov/data/>
  + Unicef data: <https://data.unicef.org/dv_index/>
  + Census public data: <https://www.census.gov/programs-surveys/ces/data/public-use-data.html>
  + Education Department data: <https://www2.ed.gov/rschstat/catalog/colleges-universities.html>