Assignment – Individual Project
Due: Nov 02

For this assignment you will be conducting a statistical analysis of a dataset, producing a set of results that will be placed in a 4 minute slide presentation, and that presentation will be delivered before the class.

Working independently you will look for a dataset, or collect one of your own. The dataset should have data relevant to one of these topics:

- Health care
- Tourism
- Food and culinary trends (including prices)
- Transportation
- Finance (including crypto, currency etc)
- Cars (automotive industry)
- Residential prices/trends/predictions

The dataset should contain at least 100 data points. Some recommendations of where you can search for datasets are:

- (Google's dedicate data search tool) https://datasetsearch.research.google.com/
- (Kaggle) https://www.kaggle.com/
- (R datasets) https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html
- (US government data) https://data.gov/
- (Open Data on StackExchange) https://opendata.stackexchange.com/

You will investigate the associations contained in the dataset using methodologies covered in the class. Specifically your approach should utilize at least 2 of the following methods:

- Linear regression
- LOESS
- Decision Trees
- Random Forests
- XGBoost

(**other methodologies not covered in class can also be used but must be described*)

The presentation should have 8-16 slides in total. The structure of the presentation should follow this format:

- Title slide containing the title of your project, and your name
- A single slide containing the brief answers to these 3 questions 1) what the study aims to investigate and why it is important 2) which dataset is used to investigate this question 3) which methodologies are used in this investigation.
- Data section 1-3 slides stating where the data was obtained, a summary of the variables used in the data, the size of the dataset and any extra information you deem to be useful to describe. You can optionally include a slide displaying 'descriptive statistics' such as histograms/boxplots etc.
- Methodology/Model definition section 1-3 slides covering the methodologies used and the models defined. Eg. state what are the dependent and independent variables the models are fitting for your study.
- Results section of 2-5 slides where the results of the model inferences are presented. Results can include results of hypothesis tests, model fits, comparison of different model fitting to look at R^2 for instance or RMSE etc, predictive accuracy in training testing, confusion matrix, or some other quantifiable comparison based upon the investigation.
- Conclusion 1 slide summarizing the results in the context of the motivational question of slide 2.

(**screen shots are fine but should be clear to the audience*)

Presentations will take place during class times in front of the peers and myself. The presentation will be on the classroom projector. Not everyone will be able to present on the same day. At the end of the presentation you will be asked a question from the audience and a question from the instructor.

Points you will be graded on:

- The choice of the dataset on its originality and potential insight
- Basic analysis of the dataset
- Motivation for why these results are important
- Overall quality of slides and structure of text / images
- Description of the methodology
- Inference and Results presented