DATA 602 Group Project

Data Analysis with Hypothesis Testing and Linear Regression

This project aims to test your knowledge in applying some of the main concepts we learn through this course and to test your skills in communication and project development.

DATA 602 instructors will be creating groups within the first couple of weeks. Once your group information is shared, you can start communicating with your members to initiate your group project work.

Project Details:

The best learning happens when you develop your own data analysis! You are required to apply explanatory data analysis (data visualization and numerical summaries), confidence intervals, hypothesis testing and regression modeling for a data set of your choice. Please invest in finding the right data set based on the hypothesis of interest and the regression model you are planning to work with. You can investigate multilinear regression as well, but it is not mandatory.

TASK I: Research and find the right data set for your study.

Few suggested sites to investigate an appropriate data set:

https://www.kaggle.com/datasets

https://www.tableau.com/learn/articles/free-public-data-sets

TASK II: Identify and apply appropriate EDA techniques that you learnt in class and summaries your findings.

TASK II: Formulate the research question such as the hypothesis (or multiple hypotheses) that you plan to investigate with the choice of your data. Conduct a detailed test of hypothesis to reach your conclusion based on the formulated hypothesis.

TASK III: Choose an appropriate target variable and an explanatory variable in your dataset which you believe a linear relationship exists. In this part, please use data visualization in justifying why you have decided to go with the corresponding variables. Conduct regression analysis to develop the model and evaluate your model based on residual analysis.

Your report/presentation must consist of the following components:

- Introduction (Provide details of the objective of this project and explain the data set you decided to choose
- EDA and Data visualization (please provide any data visualizations that you have used to elaborate on your answer. This part does not necessarily have to be a separate section. Data visualization can be included within introduction, hypothesis tests and regression analysis)
- Hypothesis test results (please do not just copy paste outputs from R. In fact, make sure to add detailed information of all results)
- Regression analysis (please do not just copy paste outputs from R. In fact, make sure to add detailed information of all results)
- Conclusions and future steps/recommendations
- References

Submission of your reports:

- For the presentation, preferred media would be PPT or pdf.
- For the report submission, please submit the pdf of your R(rmd) file along with your presentation media (PPT or pdf).
- Your grade will be based on the quality of your analysis, the clarity and organization of your presentation, and the completeness of your report submission.