Project

Description: This project will be done in pairs. The goal will be to use the techniques learned in class to analyze real data. You should find a partner with whom you will work by Feb 12. If you can't find a partner by that date, email me so I can find a match. As a group, you will find a robust data set with at least 100 observations. There should be one continuous response variable, at least 8 viable numerical predictors and at least 3 viable categorical predictors in the data set. You must see me and get the data set approved by me on or before 3:30 PM on February 20. The project will done in two stages.

Stage 1: Simple Linear Regression.

a) Find the numerical predictor with the highest correlation to the response variable.

b) Use this variable to perform a simple linear regression analysis.

c) Write a paper on your analysis using the following format:

* Introduction: The introduction should include a brief description of the data you are analyzing, the important finding from an exploratory analysis of the data including scatter plots of the data, the correlation you ran to decide on the predictor, summary data of the predictor and response. Not only should you perform a simple linear regression, you develop 2- 3 interesting research questions your model can answer. You can also provide a summary of other existing research results/literatures on similar topic if there is any.
* Methods and Analysis: Describe the whole analysis you have done. You should do a complete simple linear regression analysis for the research questions you have listed in the Introduction section. If you find any seriously violations in the assumptions, you should use appropriate remedial methods to adjust your model to an appropriate model.
* Results: Include graphs or tables in the results section that assist in answering your research questions. Make sure they are properly labeled and titled. In this part, you should include the statistical conclusions based on the graphs/tables you get. Interpret your results assuming your audience are people who have some background in linear regression analysis;
* Conclusion and Discussion: A conclusion paragraph that summarizes your findings from the project. Also include some questions you have and any idea for potential follow-up work;
* References: Choose a particular citation format (for example APA, Chicago, MLA, etc, styles). You should cite any resources you used. They could be journal articles, website, etc. Remember to cite the source of the dataset as well.
* One paper per group must be uploaded to SAKAI by midnight on Friday, March 13.

Stage 2: Multiple Linear Regression.

a) Perform a multiple linear regression analysis on your whole data set to find 2 -4 candidate models for the best model. Using techniques learned in class to determine the best model

b) Write a paper on your analysis using the following format:

* Introduction: The introduction should include a brief description of the data you are analyzing, further findings from the exploratory analysis of the data that was not included in stage 1. Develop 2- 3 interesting research questions your model can answer. You can also provide a summary of other existing research results/literatures on similar topic if there is any.
* Methods and Analysis: Describe the whole analysis you have done. You should do a complete multiple linear regression analysis for the research questions you have listed in the Introduction section. If you find any seriously violations in the assumptions, you should use appropriate remedial methods to adjust your model to an appropriate model.
* Results: Include graphs or tables in the results section that assist in answering your research questions. Make sure they are properly labeled and titled. In this part, you should include the statistical conclusions based on the graphs/tables you get. Interpret your results assuming your audience are people who have some background in linear regression analysis;
* Conclusion and Discussion: A conclusion paragraph that summarizes your findings from the project. Also include some questions you have and any idea for potential follow-up work;
* References: Choose a particular citation format (for example APA, Chicago, MLA, etc, styles). You should cite any resources you used. They could be journal articles, website, etc. Remember to cite the source of the dataset as well.
* One paper per group must be uploaded to SAKAI by midnight on Friday, April 24.

IMPORTANT DATES:

2/12 - Name of your partner

2/20- Data set being analyzed

3/13 - Simple Linear Model paper due

4/24 - Multiple Regression Model paper due.

Where to find data:

1) Kaggle.com

2) https://archive.ics.uci.edu/ml/datasets.html

3) <https://data.cityofchicago.org> or search any other city <city name> data portal.

4) <https://data.neonscience.org/data-products/explore>

5) http://catalog.data.gov/dataset

6) <https://dasl.datadescription.com/>