San José State University
Department of Applied Data Science

DATA 220
Mathematical Methods for Data Analysis

Spring 2021
Instructor: Ron Mak

Assignment #8

Assigned: Thursday, April 8
Due: Thursday, April 15 at 5:30 pm
Team assignment, 100 points max

Multiple regression analysis
In this assignment, you will perform multiple regression analysis on a dataset that you find and download. Since you will use a linear model, try to find data where the dependent variable is linearly dependent on at least two independent variables.

In a Jupyter notebook, you should at least perform the same analysis operations on your dataset as was performed in class on the California Housing dataset. See the example Jupyter notebook **`CaliforniaHousing.ipynb`**:

- Perform any necessary cleanup and transformation operations on the dataset.
  - What is the shape of the dataset after cleanup and transformation?
  - What are the independent variables (attributes) and the dependent variable?
- Create a scatterplot of each independent variable vs. the dependent variable.
  - Include a regression line within each graph.
  - If there are outliers that that cause the bulk of your plotted points to bunch up in a scatterplot, you can remove the outliers from the graph.
- Split your dataset for training and for testing.
  - What are the shapes of the training and testing data?
- Train your machine learning model on the training data.
  - Assume a linear model.
  - What are the regression coefficients?
- Which attributes have the most and the least influences on the dependent variable?
  - Which influences are positive, and which are negative?
  - What conclusions can you draw about the attributes?
- Test your model on the testing data.
  - How good are your model's predictions?
  - How strong is the correlation between the predicted values and the expected values of the dependent variable?
  - Draw a scatterplot of predicted vs. expected values.

Report
Incorporate your report about your data and your operations and answers to the above questions in the comments and markdowns of your notebook.

What to submit to Canvas
Submit your Jupyter notebook and your report (if separate) into Canvas:
Assignment #8.

Rubric
Your program will be graded according to these criteria:

Criteria	Max points
Chosen dataset • At least two independent variables. • Cleanup and transformation operations.	**20** • 10 • 10
Multiple regression analysis • Scatterplots of each independent variable vs. the dependent variable. • Split the dataset into training and testing data. • Train the data with a linear model to obtain the regression coefficients. • Test the model and evaluate its predictions.	**80** • 20 • 20 • 20 • 20