Lab #5 - Data Visualization and Machine Learning

Re-submit Assignment

Due Nov 15, 2019 by 5pm **Points** 4 **Submitting** a file upload
Available after Nov 14, 2019 at 4:55pm

Lab #5 - Data Visualization and Machine Learning

Complete these exercises and submit a single Jupyter Notebook file (in .html format, not .ipynb) that contains your responses by **5PM on Friday November 15th**. Late assignments will be penalized up to 2 points per day (20%), unless prior arrangements have been made to submit the assignment after the deadline. *Please put your names at the top of the lab and note if any assigned group members are absent.*

The notebook should be well organized. Each section should be **clearly labeled with the exercise (and part) that it addresses** (e.g., Exercise #1a, #1b, #2) in a Markdown cell block. Use (clear and concise) comments as needed to help describe each step of your process. All notebook cells that contain essential steps should be executed and the output should be visible, so as to demonstrate your successful completion of the exercise. If you cannot complete an exercise in its entirety, you should make an effort to demonstrate your intermediate progress in order to maximize partial credit, and move forward as best as possible. You may submit any written answers to the exercises in the notebook as text cells.

Academic Integrity: Each group is expected to submit its own original work. You may consult with the instructor and the TAs for help, but collaboration with other groups should be kept to a minimum. Submissions that contain significant similarities will be reported directly to the Office of Student Conduct.

Instructions/Background

Diamonds are the crown jewel of the jewelry industry. As a luxury item, there is a broad range of quality-related characteristics and prices, designed to accommodate many customer segments. The attached data file contains sample data on 6,000 diamonds. Each row contains a summary of characteristics for each diamond, including the 4 C's (carat weight, color, clarity, cut), polish, symmetry, certifying agency (report), and price.

Step 0 (Ungraded): Data Preparation

Read in the data. Look at the descriptive statistics (continuous variables) or value counts (categorical variables) for each column.

Exercise #1: Data Visualization

Use appropriate techniques for data analysis and visualization, along with relevant features in matplotlib, pandas, Seaborn, or plotly to create one high-quality data visualization. Write 4-6 sentences describing the findings of your visualization. You will be graded on the quality of the visualizations and your explanations of the insights, as well as the richness of your story. Consider filtering the data in multiple ways to look for the most interesting patterns.

Exercise #2: Prediction Using OLS Regression

In exercise 2, we will analyze the relationship between the diamond characteristics and the price of the diamond. You will then try to predict the price from the characteristics for a subset of the data.

Step 2.1: Prepare the data for regression using stats models:

- In the regression, we will use the following columns for the independent (exogenous) variables: ['Carat Weight', 'Cut', 'Color', 'Clarity', 'Polish', 'Symmetry']. Which of these columns are continuous? Which are categorical? Create dummy variables for the categorical columns and add all of the columns to the X dataframe (see lecture).
- Add the price divided by 100 to the y column. What do you think is the advantage of dividing the price by 100? If you are not sure, check the results of the regression using the price without this change and compare the results.

Step 2.2: Run the regression on the full sample using statsmodels and view the results. Recall that in regression, you can interpret the coefficients as follows: "On average, a 1 unit increase in __x_variable__ is associated with a __coefficient_value__ increase in __y_variable__." If the coefficient is negative, that indicates a decrease in the y variable.

- Rewrite a version of the sentence above to construct a written description of the coefficient on carat weight. Make sure you mention the units of the y variable.
- When looking at dummy variables, we need to add to the interpretation that the coefficient is "compared to __left_out_dummy_category__." Rewrite the sentence above to interpret the coefficient on color "E" - what is the left out category that we are comparing the "E" color to?

Step 2.3: Use scikit learn to create predicted prices for a subset of the data.

- Split the sample into 30% test data and 70% training data and rerun the model. What is the out of sample R^2?
- Predict the price for the test data. What is the correlation between the predicted and actual values? Is this a high correlation? (Note: Unlike the example from class, we do not have to exponentiate the predicted/actual values.)

Lab #5 Rubric