Lab 07 (Module 08)

This lab will be done in Python. For the lab to be complete, all questions must be answered.

Section 1: Data Preparation

We'll be using the train.csv data set for this lab. The data set covers the characteristics and prices for used cars sold in India.

1) Import data and examine a few rows.
2) Look at the data types of the variables using `dtypes`.
3) To use them in a model, we need Engine and Power to be numeric. To make them numeric, you will need to strip out the text characters and convert the data type to numeric. Use the following code:
    a. `df1["Engine"] = df1["Engine"].str.rstrip(" CC")`
    b. `df1["Power"] = df1["Power"].str.rstrip(" bhp")`
    c. `df1["Power"] = df1["Power"].replace(regex="null", value = np.nan)`
    d. `df1["Power"] = df1["Power"].astype("float")`
    e. `df1["Engine"] = df1["Engine"].astype("float")`
4) Compute some summary statistics of the variables using `describe()`. Notice that the object variables do not appear in the `describe()` output. What do you notice when you examine the summary statistics? Hint: Look at the counts and compare means with medians.
5) Because `describe()` will not provide information about the object variables, you will need to check whether there is missing data in those using a different method. Which variable had the most missing values?
6) Use `pairplot()` to examine the distributions and scatterplots of these variables. Do you see potential problems such as non-normal or potentially correlated data?
7) Transform Price so that it looks more normal, then create a new plot of the transformed variable.

Section 2: Predicting Used Car Prices

We are interested in predicting the price of a car given Power, Engine, Kilometers Driven, and Year. We will attempt to build a linear regression model of Price.

1) Build a model of transformed price based on the other 4 variables. Be sure to examine both the output and the notes. You will get an error if any of the covariates have missing values. Hint: Only drop rows which have missing values in the variables you are using for the model.
    a. How much variance is explained?
    b. How many observations were used to fit the model?
2) Based on your plots and the note about condition indices in the output, multicollinearity may be a problem. Compute the VIF of each variable. Which variables does it show are correlated?
3) Dimensionality reduction through PCA is one way to manage collinear variables. Use PCA to create principal components and create a scree plot. For this example, which criteria would you use to determine the number of components to use: proportion of variance, eigenvalues > 1, or elbow on the plot?

4) Create / select 2 components and rerun the regression with transformed price. What is the amount of variance explained?
    a. When merging the transformed price onto the principal components dataframe, you will likely need to reset the index.
    b. `principalDf = pd.concat([principalDf, df1['lnPrice'].reset_index(drop=True)], axis = 1)`
5) What is the VIF of the components?