Lab 06

Predicting used car prices

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. We are interested in predicting the price of a car given some characteristics. For today we will focus on Price, Power, Engine, Kilometers Driven, and Year, and we will attempt to build a linear regression model of Price.

1. Visualize the distribution of price
2. Transform Price so that it looks more normal
3. Build a model of transformed price based on the other 4 variables, how much variance is explained?
4. Compute the VIF of each variable, which 2 are potential problems?
5. Use PCA to create principal components and create a scree plot
6. Create / select 2 components and rerun the regression with transformed price, what is the amount of variance explained?
7. What is the VIF of the components?

Note:

1. Here are some transformations for the variables:

```
df1["Mileage"] = df1["Mileage"].str.rstrip(" kmpl")
df1["Mileage"] = df1["Mileage"].str.rstrip(" km/g")
df1["Engine"] = df1["Engine"].str.rstrip(" CC")
df1["Power"] = df1["Power"].str.rstrip(" bhp")
df1["Power"]= df1["Power"].replace(regex="null", value = np.nan)
df1["Fuel_Type"]=df1["Fuel_Type"].astype("category")
df1["Transmission"]=df1["Transmission"].astype("category")
df1["Owner_Type"]=df1["Owner_Type"].astype("category")
df1["Mileage"]=df1["Mileage"].astype("float")
df1["Power"]=df1["Power"].astype("float")
df1["Engine"]=df1["Engine"].astype("float")
df1["Company"]=df1["Name"].str.split(" ").str[0]
df1["Model"]=df1["Name"].str.split(" ").str[1]+df1["Name"].str.split(" ").str[2]
```

2. When merging the transformed price onto the principal components dataframe you will likely need to reset the index, here is a sample

```
pd.concat([principalDf, df1['lnPrice'].reset_index(drop=True)], axis = 1)
```