Exercise 04

Section 1: Predicting used car prices

We'll be using the cars.csv data set for this section of the exercise. 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. We will attempt to build a linear regression model of Price. We are going to work on filling in the missing data that we previously dropped.

Here is some handy code to help formatting the data

```
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]
```

1. Transform Price so that it looks more normal, produce histograms of the variable before and after transformation
2. How many values are missing for Power and Engine?
3. Which column has the most missing values and what should we do about it?
4. Build a model of transformed price based on Power, Mileage, Kilometers Driven, and Year, how much variance is explained?
5. How many rows were used to train the model?
6. Fill the missing values in Power and Mileage with their respective means and rebuild the model. Now how much variance is explained?
7. How many rows were used to train the model?
8. Impute the missing data using MICE and rebuild the model

MICE documentation:
https://www.statsmodels.org/dev/generated/statsmodels.imputation.mice.MICE.html

9. How have the parameter estimates changed from step 4?
10. Plot the distribution of Power with and without MICE
11. Plot the distribution of Engine with and without MICE

Section 2: Predicting customer spending

We’ll be using the lab06_customers.csv data set for this lab. The data set covers the demographic characteristics of some customers and the amount they spent over the past year at an online retailer. For this exercise it is recommended to use the sklearn packages for linear regression, ridge, and lasso. Sklearn documentation linked below.

Linear regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Ridge: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

Lasso: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

In order to interact the categorical variables you will need to dummy code them and manually multiply, an example is given below.

```
customerDf = pd.get_dummies(data=customerDf, columns=['sex', 'race'], prefix=['sex','race'])
customerDf["Hispanic_Male"] = np.multiply(customerDf["race_hispanic"],customerDf["sex_male"])
```

1. Build a linear regression with all the dependent variables and two way interactions between sex and race, consider the other category for race and sex to be the reference category and treat it appropriately
2. Build ridge models with various values for alpha. Create a chart showing how the coefficients change with alpha values
3. Build lasso models with various values for alpha. Create a chart showing how the coefficients change with alpha values
4. Compare the coefficients from linear regression, ridge, and lasso (select an alpha value using your chart)
5. Compare the R2 from lr, ridge, and lasso
6. Which model would you choose, and why?
7. Which variables are dropped from the chosen model that were not dropped in linear regression?