Cars Python Project Code

# Code to read in csv file

import pandas as pd

df = pd.read\_csv(r'C:\Users\G01317496\Downloads\Cars\_Data.csv')

df.head(20)

#importing other libraries

import numpy as np

import seaborn as sns

import matplotlib

import matplotlib.pyplot as plt

plt.style.use('ggplot')

from matplotlib.pyplot import figure

%matplotlib inline

matplotlib.rcParams['figure.figsize'] = (12, 8) # This adjusts the configuration of the plots I'm going to create

# Visualisation to look at correlation between engine size and breakhorse power

# This graph shows us that the correlation between break horse power ansd engine size is very high but what would the value be?

sns.regplot(x = 'EngineSize', y = 'Horsepower', data = df)

# This will show us correlation of all numeric fields, but we will only look at the fields enginge size and break horse power

# We see that the correlation coeefficient is 0.787435 which shows there is a high level of correlationb

df.corr()

# Removing first character from MSRP and Invoice

df['MSRP'] = df['MSRP'].str[1:]

df['Invoice'] = df['Invoice'].str[1:]

df.head()

# Replacing comma in columns to stop it being a string

df['MSRP'] = df['MSRP'].str.replace(',', '')

df['Invoice'] = df['Invoice'].str.replace(',', '')

df.head()

# Converting MSRP and Invoice to float

df['MSRP'] = df['MSRP'].astype(dtype='float')

df['Invoice'] = df['Invoice'].astype(dtype='float')

df.head(10)

# Replacing null values in dataset with the mean average

df.isnull().sum()

df['Cylinders'].fillna(df['Cylinders'].mean(), inplace = True)

# The different makes of cars and the sum of eaxch occurance of make

df['Make'].value\_counts()

# All cars where the origin is Asia or Europe

df[(df['Origin'] == 'Asia') | (df['Origin'] == 'Europe')]

df[df['Origin'].isin(['Asia', 'Europe'])]

# Coode to remove all records where Weight > 4000

df[df['Weight'] > 4000] # First to check how many records (103 rows)

df[~(df['Weight'] > 4000)]

# increasing all MPG\_City by 3

df['MPG\_City'] = df['MPG\_City'].apply(lambda x:x +3)

df