Assignment1_DM (1) (1)

March 30, 2020

[53]:
```
# PROBLEM 1
import pandas as pd
from sklearn.preprocessing import MinMaxScaler, StandardScaler, RobustScaler
data = pd.DataFrame({"Age": [25,56,65,32,41,49],"Income":
 ↪[49000,156000,99000,92000,39000,57000]})

# Using standard scalar
scaler = StandardScaler()
standardized_data = norm_df = pd.DataFrame(scaler.fit_transform(data),␣
 ↪index=data.index, columns=data.columns)
print("StandardScaler: \n", standardized_data, "\n")

# Using minmax
scaler = MinMaxScaler()
standardized_data2 = pd.DataFrame(scaler.fit_transform(data), index=data.index,␣
 ↪columns=data.columns)
print("MinMaxScaler: \n",standardized_data2, "\n")

# using robust scalar
transformer = RobustScaler().fit(data)
normalized_data = transformer.transform(data)
print("Using robust scalar: \n",normalized_data, "\n")

# Normalizing a data frame using pandas
norm_df = (data - data.mean()) / data.std()
print("Normalized data using Panda : \n", norm_df, "\n" )

# Rescaling using pandas
recaled_df = (data - data.min()) / (data.max() - data.min())
print("Rescaled data using Panda : \n", recaled_df, "\n" )
```

```
StandardScaler:
         Age    Income
0 -1.438597 -0.832491
1  0.829022  1.866799
2  1.487363  0.428859
3 -0.926554  0.252270
```

```
4 -0.268213 -1.084762
5  0.316979 -0.630675

MinMaxScaler:
      Age    Income
0  0.000  0.085470
1  0.775  1.000000
2  1.000  0.512821
3  0.175  0.452991
4  0.400  0.000000
5  0.600  0.153846

Using robust scalar:
 [[-1.          -0.55135135]
 [ 0.55         1.76216216]
 [ 1.           0.52972973]
 [-0.65         0.37837838]
 [-0.2         -0.76756757]
 [ 0.2         -0.37837838]]

Normalized data using Panda :
         Age    Income
0 -1.313253 -0.759957
1  0.756790  1.704147
2  1.357770  0.391493
3 -0.845824  0.230290
4 -0.244844 -0.990247
5  0.289361 -0.575725

Normalized data using Panda :
      Age    Income
0  0.000  0.085470
1  0.775  1.000000
2  1.000  0.512821
3  0.175  0.452991
4  0.400  0.000000
5  0.600  0.153846
```

[16]:
```
# PROBLEM 2 Sub-problem-1:
# Creating binary dummy variables for three categorical variables
import pandas as pd
cars_data = pd.read_csv('/Users/nivethida/Downloads/ToyotaCorolla___(1).csv')
print(cars_data.head(5))
print(pd.get_dummies(cars_data.Fuel_Type))
print(pd.get_dummies(cars_data.Color))
print(pd.get_dummies(cars_data.Model))
```

```
   Id                                        Model    Price  Age_08_04  \
0   1  TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors  13500.0         23
1   2  TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors  13750.0         23
2   3  TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors  13950.0         24
3   4  TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors  14950.0         26
4   5    TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors  13750.0         30

   Mfg_Month  Mfg_Year     KM Fuel_Type  HP  Met_Color  … Powered_Windows  \
0         10      2002  46986    Diesel  90          1  …                1
1         10      2002  72937    Diesel  90          1  …                0
2          9      2002  41711    Diesel  90          1  …                0
3          7      2002  48000    Diesel  90          0  …                0
4          3      2002  38500    Diesel  90          0  …                1

   Power_Steering  Radio  Mistlamps  Sport_Model  Backseat_Divider  \
0               1      0          0            0                 1
1               1      0          0            0                 1
2               1      0          0            0                 1
3               1      0          0            0                 1
4               1      0          1            0                 1

   Metallic_Rim  Radio_cassette  Parking_Assistant  Tow_Bar
0             0               0                  0        0
1             0               0                  0        0
2             0               0                  0        0
3             0               0                  0        0
4             0               0                  0        0

[5 rows x 39 columns]
      CNG  Diesel  Petrol  Petrole
0       0       1       0        0
1       0       1       0        0
2       0       1       0        0
3       0       1       0        0
4       0       1       0        0
…   …     …      …       …
1431    0       0       1        0
1432    0       0       1        0
1433    0       0       1        0
1434    0       0       1        0
1435    0       0       1        0

[1436 rows x 4 columns]
      Beige  Black  Blue  Green  Greenish  Grey  Red  Redish  Silver  Violet  \
0         0      0     1      0         0     0    0       0       0       0
1         0      0     0      0         0     0    0       0       1       0
2         0      0     1      0         0     0    0       0       0       0
3         0      1     0      0         0     0    0       0       0       0
```

```
4          0      1     0      0          0     0    0       0       0       0
…      …     …    …     …         …    …   …      …      …      …
1431       0      0     1      0          0     0    0       0       0       0
1432       0      0     0      0          0     1    0       0       0       0
1433       0      0     1      0          0     0    0       0       0       0
1434       0      0     0      0          0     1    0       0       0       0
1435       0      0     0      1          0     0    0       0       0       0

      White  Yellow  Yellowish
0         0       0          0
1         0       0          0
2         0       0          0
3         0       0          0
4         0       0          0
…     …      …          …
1431      0       0          0
1432      0       0          0
1433      0       0          0
1434      0       0          0
1435      0       0          0

[1436 rows x 13 columns]
      TOYOTA Corolla   \
0                   0
1                   0
2                   0
3                   0
4                   0
…               …
1431                0
1432                0
1433                0
1434                0
1435                0

      TOYOTA Corolla ! 1.6-16v vvt-i sol airco sedan 4/5-Doors  \
0                                                           0
1                                                           0
2                                                           0
3                                                           0
4                                                           0
…                                                       …
1431                                                        0
1432                                                        0
1433                                                        0
1434                                                        0
1435                                                        0
```

```
      TOYOTA Corolla 1 6-16v VVT-i Linea Terra Comfort Airco 5drs 4/5-Doors  \
0                                                     0
1                                                     0
2                                                     0
3                                                     0
4                                                     0
…                                                   …
1431                                                  0
1432                                                  0
1433                                                  0
1434                                                  0
1435                                                  0

      TOYOTA Corolla 1.3 16V 2/3-Doors  \
0                                    0
1                                    0
2                                    0
3                                    0
4                                    0
…                                  …
1431                                 0
1432                                 0
1433                                 0
1434                                 0
1435                                 0

      TOYOTA Corolla 1.3 16V HATCHB 2/3-Doors  \
0                                           0
1                                           0
2                                           0
3                                           0
4                                           0
…                                         …
1431                                        0
1432                                        0
1433                                        0
1434                                        0
1435                                        0

      TOYOTA Corolla 1.3 16V HATCHB G6 2/3-Doors  \
0                                              0
1                                              0
2                                              0
3                                              0
4                                              0
…                                            …
1431                                           1
1432                                           0
```

```
1433                                          0
1434                                          0
1435                                          0

      TOYOTA Corolla 1.3 16V HATCHB LINEA LUNA 2/3-Doors  \
0                                                     0
1                                                     0
2                                                     0
3                                                     0
4                                                     0
…                                                   …
1431                                                  0
1432                                                  0
1433                                                  0
1434                                                  0
1435                                                  0

      TOYOTA Corolla 1.3 16V HATCHB LINEA LUNA AUT3 2/3-Doors  \
0                                                     0
1                                                     0
2                                                     0
3                                                     0
4                                                     0
…                                                   …
1431                                                  0
1432                                                  0
1433                                                  0
1434                                                  0
1435                                                  0

      TOYOTA Corolla 1.3 16V HATCHB LINEA TERRA 2/3-Doors  \
0                                                     0
1                                                     0
2                                                     0
3                                                     0
4                                                     0
…                                                   …
1431                                                  0
1432                                                  1
1433                                                  1
1434                                                  1
1435                                                  0

      TOYOTA Corolla 1.3 16V HATCHB LINEA TERRA AUT3 2/3-Doors  …  \
0                                                     0         …
1                                                     0         …
2                                                     0         …
3                                                     0         …
```

```
4                                                    0          …
…                                                  …          …
1431                                                 0          …
1432                                                 0          …
1433                                                 0          …
1434                                                 0          …
1435                                                 0          …

      TOYOTA Corolla LIFTBACK 1.6-16v Linea Luna 2/3-Doors  \
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0
…                                                  …
1431                                                 0
1432                                                 0
1433                                                 0
1434                                                 0
1435                                                 0

      TOYOTA Corolla LIFTBACK 1.6-16v Linea Luna 4/5-Doors  \
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0
…                                                  …
1431                                                 0
1432                                                 0
1433                                                 0
1434                                                 0
1435                                                 0

      TOYOTA Corolla LIFTBACK 1.9 D Linea Terra 4/5-Doors  \
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0
…                                                  …
1431                                                 0
1432                                                 0
1433                                                 0
1434                                                 0
1435                                                 0

      TOYOTA Corolla Liftback 1.6 VVTI 4/5-Doors  \
```

```
0                                                 0
1                                                 0
2                                                 0
3                                                 0
4                                                 0
…                                                 …
1431                                              0
1432                                              0
1433                                              0
1434                                              0
1435                                              0

      TOYOTA Corolla Linea Luna 1.6i 16V 2/3-Doors  \
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   0
…                                                   …
1431                                                0
1432                                                0
1433                                                0
1434                                                0
1435                                                0

      TOYOTA Corolla Luna 5drs D4D116 PK 4/5-Doors  \
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   0
…                                                   …
1431                                                0
1432                                                0
1433                                                0
1434                                                0
1435                                                0

      TOYOTA Corolla Sw 20D Linea Terra Stationwagen  \
0                                                     0
1                                                     0
2                                                     0
3                                                     0
4                                                     0
…                                                     …
1431                                                  0
1432                                                  0
1433                                                  0
```

```
1434                                                 0
1435                                                 0

      TOYOTA Corolla VERSO 2.0 D4D LINEA SOL MPV  \
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   0
…                                                   …
1431                                                0
1432                                                0
1433                                                0
1434                                                0
1435                                                0

      TOYOTA Corolla VERSO 2.0 D4D SOL (7) BNS MPV  \
0                                                     0
1                                                     0
2                                                     0
3                                                     0
4                                                     0
…                                                     …
1431                                                  0
1432                                                  0
1433                                                  0
1434                                                  0
1435                                                  0

      TOYOTA Corolla VERSO 2.0 D4D SOL (7) MPV
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   0
…                                                   …
1431                                                0
1432                                                0
1433                                                0
1434                                                0
1435                                                0

[1436 rows x 319 columns]
```

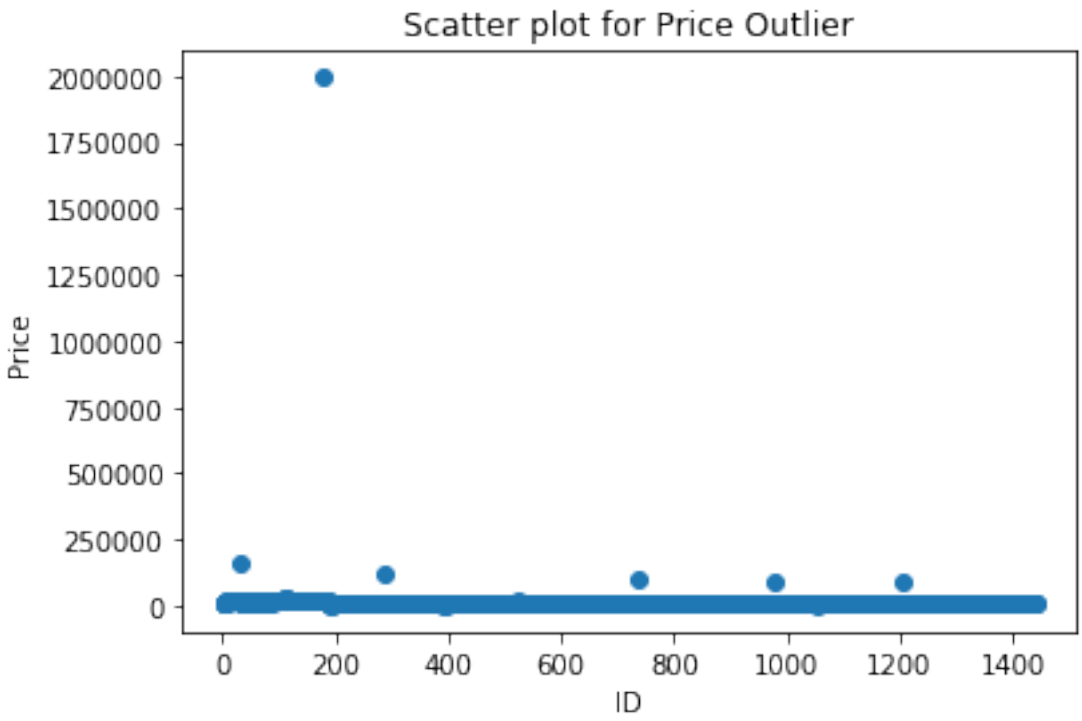
[17]:
```
# PROBLEM 2 Sub-problem-2:
# Partitioning the dataset
from sklearn.model_selection import train_test_split
```

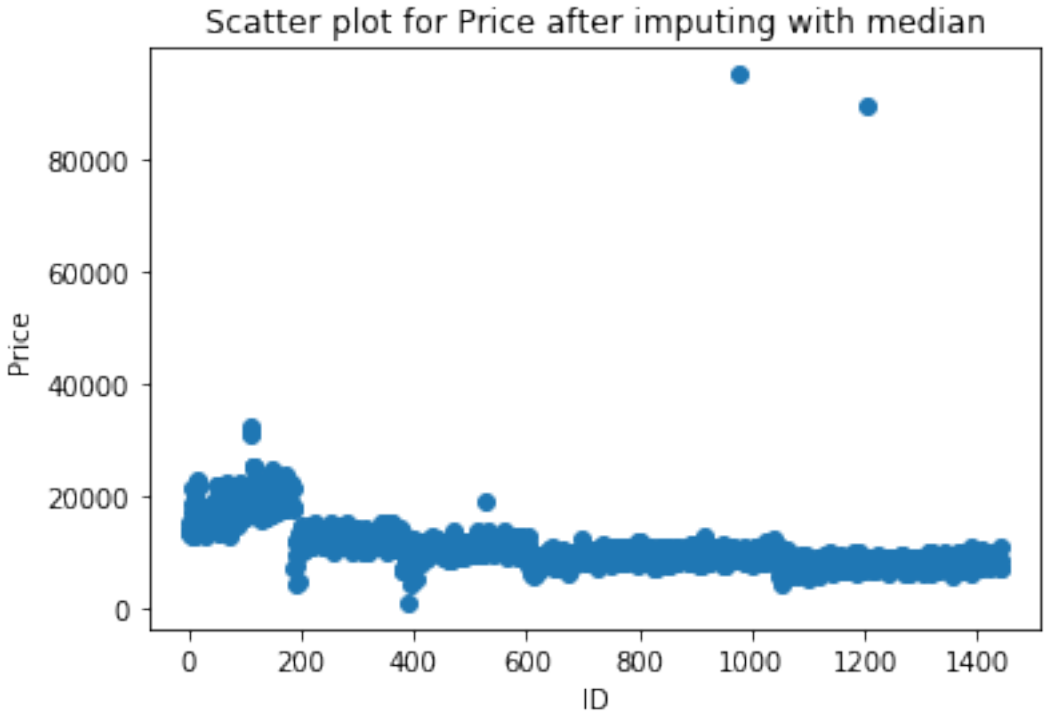
```
cars_data = pd.read_csv('/Users/nivethida/Downloads/ToyotaCorolla___(1).csv')
trainData, temp = train_test_split(cars_data, test_size=0.5, random_state=1)
validData, testData  = train_test_split(cars_data, test_size=0.3,␣
 ↪random_state=1)
print('Training Data: ', trainData.shape)
print('Validation Data: ', validData.shape)
print('Test Data: ', testData.shape)
```

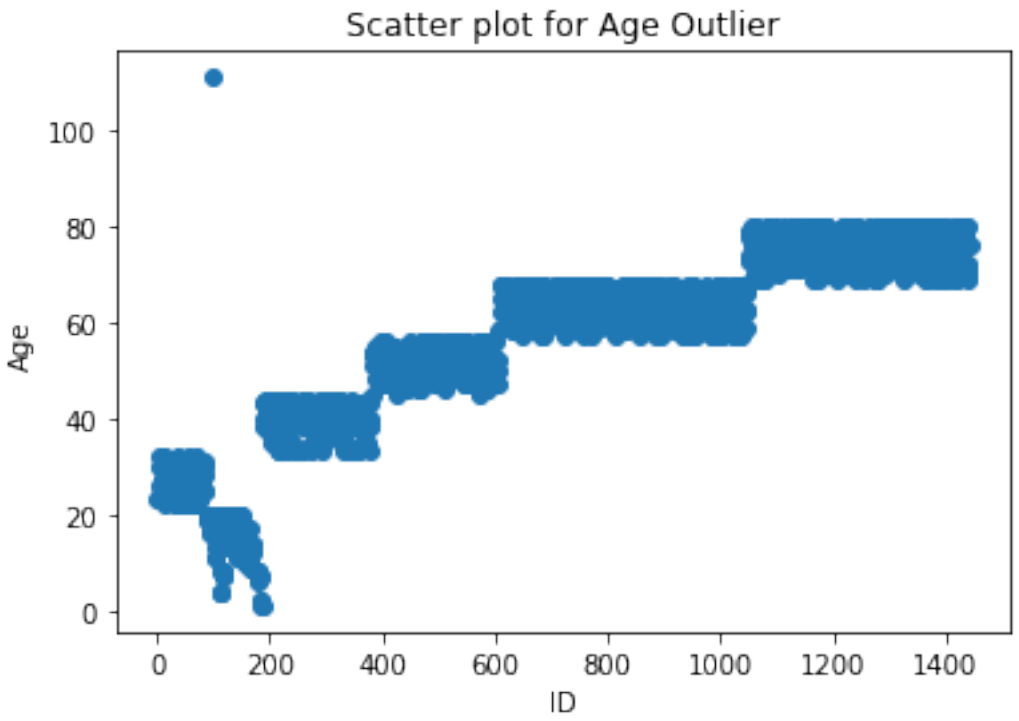
```
Training Data:  (718, 39)
Validation Data:  (1005, 39)
Test Data:  (431, 39)
```

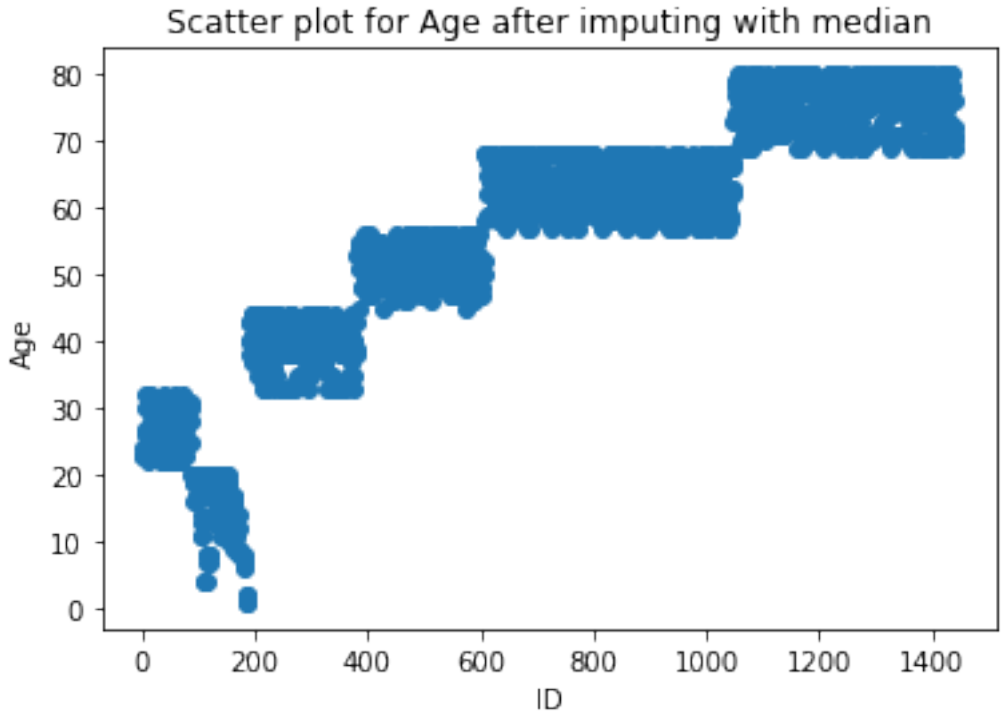
[30]:
```
# PROBLEM 2 Sub-problem-3:
# Exploring the data for the outliers
import matplotlib.pyplot as plt
import numpy as np
np.warnings.filterwarnings('ignore')
cars_data = pd.read_csv('/Users/nivethida/Downloads/ToyotaCorolla___(1).csv')

# Finding outlier in Price
plt.scatter(x=cars_data['Id'], y=cars_data['Price'])
plt.title('Scatter plot for Price Outlier')
plt.xlabel('ID')
plt.ylabel('Price')
plt.show()

# Creating Indicator for outlier
cars_data.loc[cars_data['Price'] > 100000, 'Price_outlier'] = 1

# Considering all values greater than 100000 as outliers and filling it with␣
 ↪null values
cars_data.loc[cars_data['Price'] > 100000, 'Price'] = np.nan

# Judging the outlier as data error and replacing it with median value
median_Price = cars_data['Price'].median()
cars_data['Price'].fillna(value=median_Price)
plt.scatter(x=cars_data['Id'], y=cars_data['Price'])
plt.title('Scatter plot for Price after imputing with median')
plt.xlabel('ID')
plt.ylabel('Price')
plt.show()

# Finding outlier in age
plt.scatter(x=cars_data['Id'], y=cars_data['Age_08_04'])
plt.title('Scatter plot for Age Outlier')
plt.xlabel('ID')
plt.ylabel('Age')
```

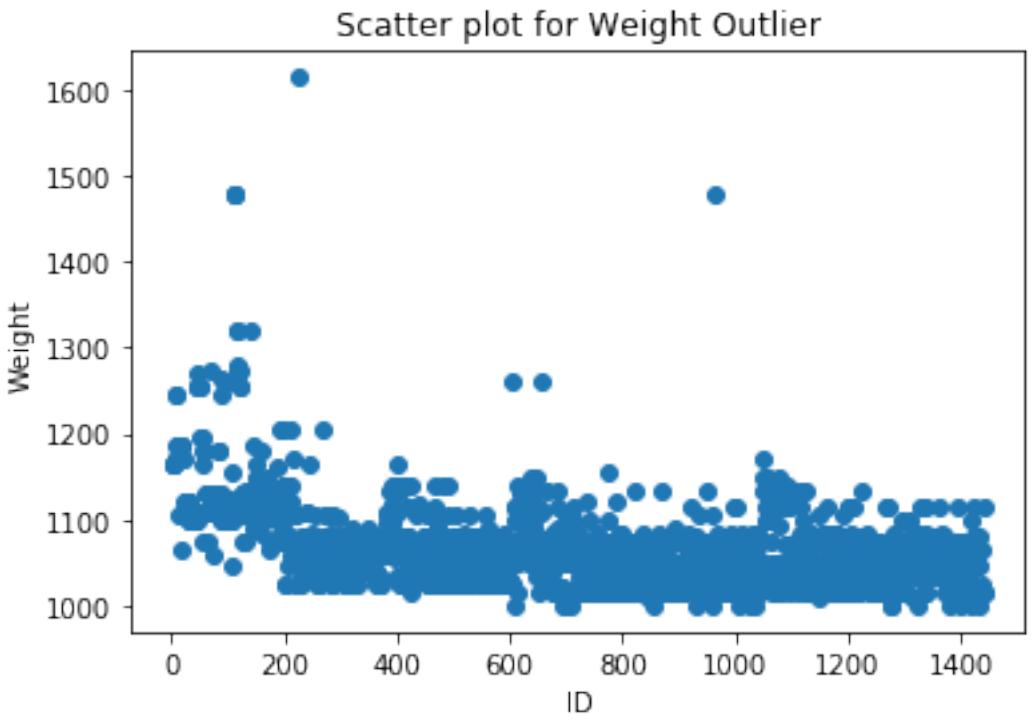
```
plt.show()

# Creating Indicator for outlier
cars_data.loc[cars_data['Age_08_04'] > 100, 'Age_outlier'] = 1

# Considering all values greater than 100 as outliers and filling it with null␣
 ↪values
cars_data.loc[cars_data['Age_08_04'] > 100, 'Age_08_04'] = np.nan

# Judging the outlier as data error and replacing it with median value
median_Age = cars_data['Age_08_04'].median()
cars_data['Age_08_04'].fillna(value=median_Age)
plt.scatter(x=cars_data['Id'], y=cars_data['Age_08_04'])
plt.title('Scatter plot for Age after imputing with median')
plt.xlabel('ID')
plt.ylabel('Age')
plt.show()

# Finding outlier in Weight
plt.scatter(x=cars_data['Id'], y=cars_data['Weight'])
plt.title('Scatter plot for Weight Outlier')
plt.xlabel('ID')
plt.ylabel('Weight')
plt.show()
max_Weight = cars_data['Weight'].max()

# Creating Indicator for outlier
cars_data.loc[cars_data['Weight'] > 1600, 'Weight'] = 1
cars_data.loc[cars_data['Weight'] < 200, 'Weight'] = 1

# Considering all values greater than 1600 or less than 200 as outliers and␣
 ↪filling it with null values
median_val = cars_data['Weight'].median()
cars_data.loc[cars_data['Weight'] > 1600, 'Weight'] = np.nan
cars_data.loc[cars_data['Weight'] < 200, 'Weight'] = np.nan

# Judging the outlier as data error and replacing it with median value
cars_data['Weight'].replace(max_Weight, np.nan, inplace=True)
median_Weight = cars_data['Weight'].median()
cars_data['Weight'].fillna(value=median_Weight)
plt.scatter(x=cars_data['Id'], y=cars_data['Weight'])
plt.title('Scatter plot for Weight after imputing with median')
plt.xlabel('ID')
plt.ylabel('Weight')
plt.show()
```


Scatter plot for Price Outlier
2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
Price
0
200
400
600
800
1000
1200
1400
ID

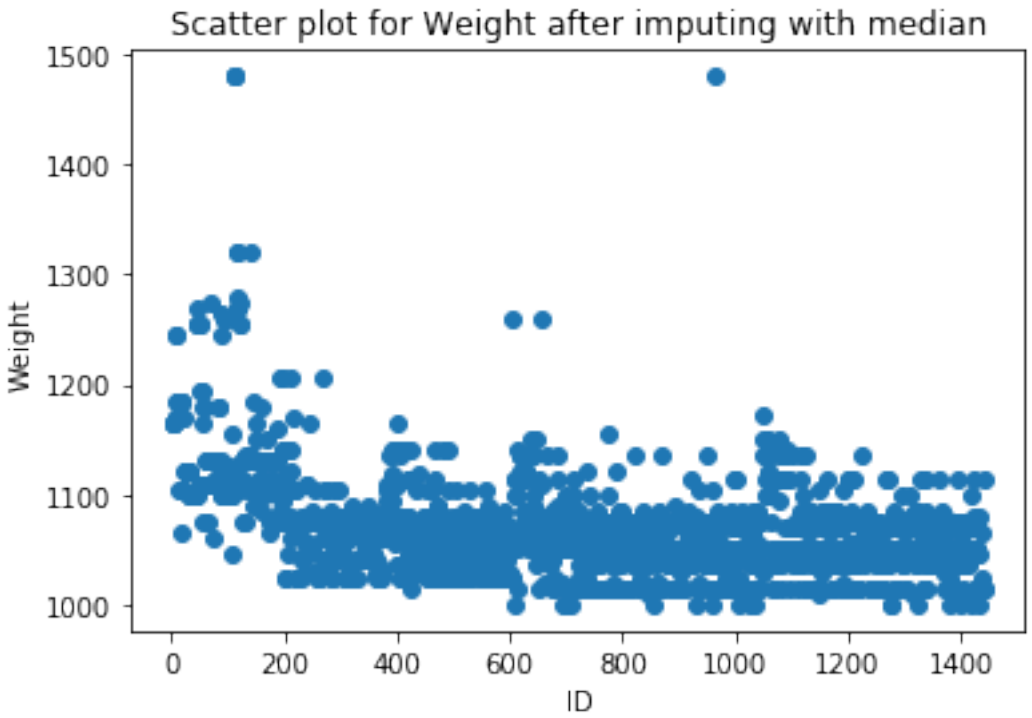

Scatter plot for Price after imputing with median
80000
60000
40000
20000
0
Price
0
200
400
600
800
1000
1200
1400
ID


Scatter plot for Age Outlier
Age
ID
0
20
40
60
80
100
0
200
400
600
800
1000
1200
1400


Scatter plot for Age after imputing with median
Age
ID
0
10
20
30
40
50
60
70
80
0
200
400
600
800
1000
1200
1400

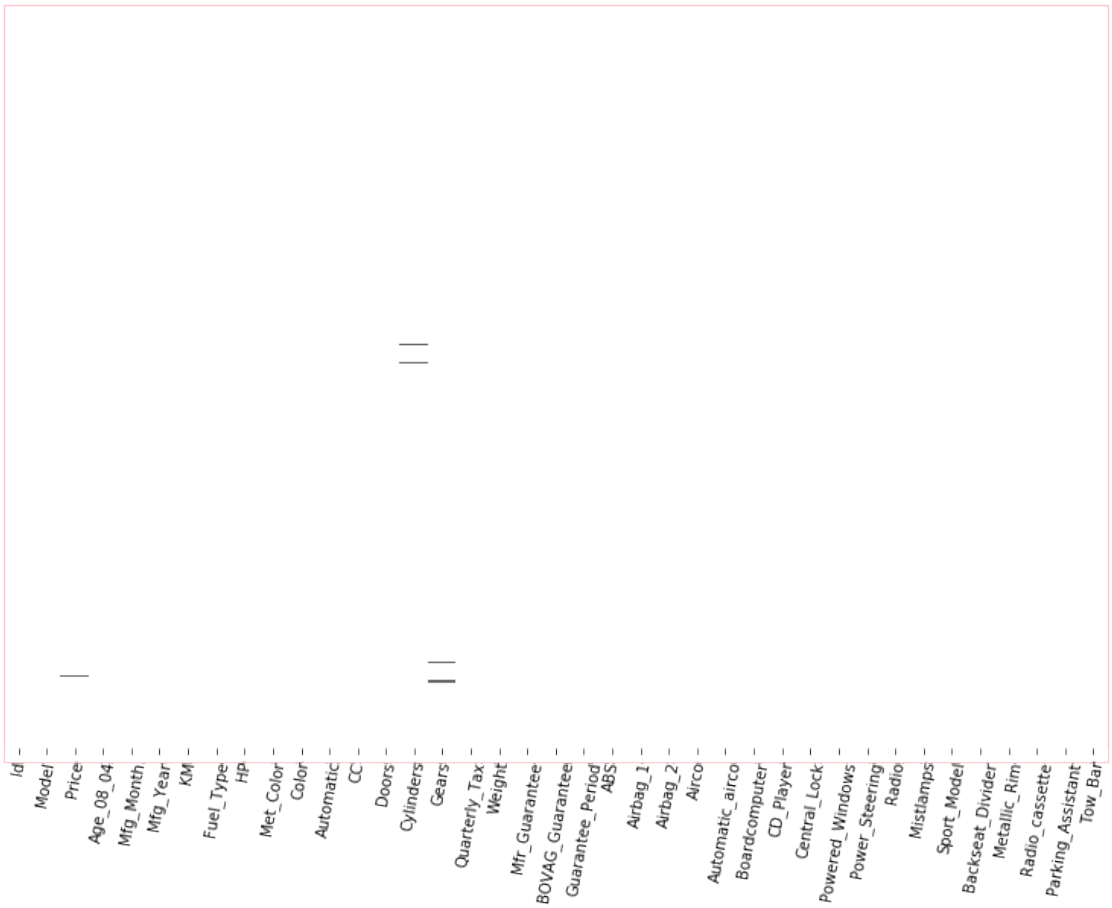

Scatter plot for Weight Outlier
Weight
ID
1600
1500
1400
1300
1200
1100
1000
0
200
400
600
800
1000
1200
1400

1615


Scatter plot for Weight after imputing with median
Weight
ID
1500
1400
1300
1200
1100
1000
0
200
400
600
800
1000
1200
1400

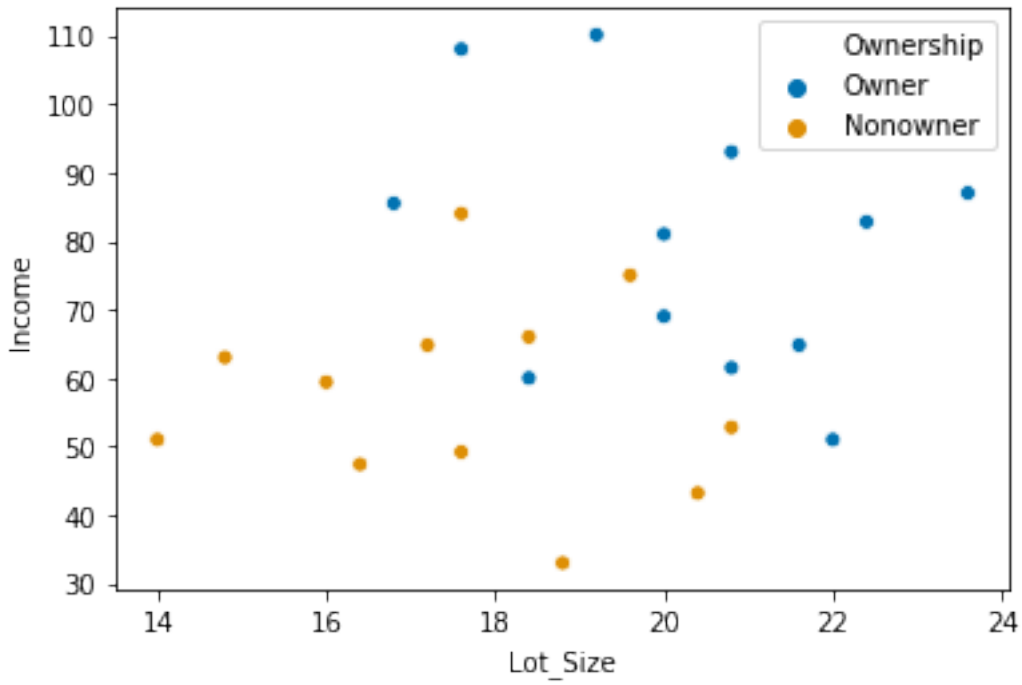
[39]:
```
# PROBLEM 2
# Sub-problem-4:
# Exploring the data for missing values
import numpy as np
import seaborn as sns
car_df = pd.read_csv('/Users/nivethida/Downloads/ToyotaCorolla___(1).csv').
 ↪sort_values(['Mfg_Year'])
naInfo = np.zeros(car_df.shape)

# Filling the missing values with 1
naInfo[car_df.isna().values] = 1
naInfo = pd.DataFrame(naInfo, columns=car_df.columns)
fig, ax = plt.subplots()
fig.set_size_inches(13, 9)

# Plotting the heat map
ax = sns.heatmap(naInfo, vmin=0, vmax=1, cmap=['white', '#666666'], cbar=False,␣
 ↪ax=ax)
ax.set_yticks([])
rect = plt.Rectangle((0, 0), naInfo.shape[1], naInfo.shape[0], linewidth=1,␣
 ↪edgecolor='pink', facecolor='none')
rect = ax.add_patch(rect)
rect.set_clip_on(False)
plt.xticks(rotation=80)
```

[39]:
```
(array([ 0.5,  1.5,  2.5,  3.5,  4.5,  5.5,  6.5,  7.5,  8.5,  9.5, 10.5,
        11.5, 12.5, 13.5, 14.5, 15.5, 16.5, 17.5, 18.5, 19.5, 20.5, 21.5,
        22.5, 23.5, 24.5, 25.5, 26.5, 27.5, 28.5, 29.5, 30.5, 31.5, 32.5,
        33.5, 34.5, 35.5, 36.5, 37.5, 38.5]),
 <a list of 39 Text xticklabel objects>)
```



[2]:
```
# Problem 3
# Creating a scatter plot of Lot Size vs. Income, color-coded by the outcome␣
 ↪variable owner/nonowner
import seaborn as sns
import pandas as pd

# Read using pandas
df = pd.read_csv('/Users/nivethida/Downloads/RidingMowers.csv')

# Display with seaborn scatterplot using ownership in hue
sns.scatterplot(x="Lot_Size", y="Income", hue="Ownership",
                data=df, palette='colorblind', legend='full')
```

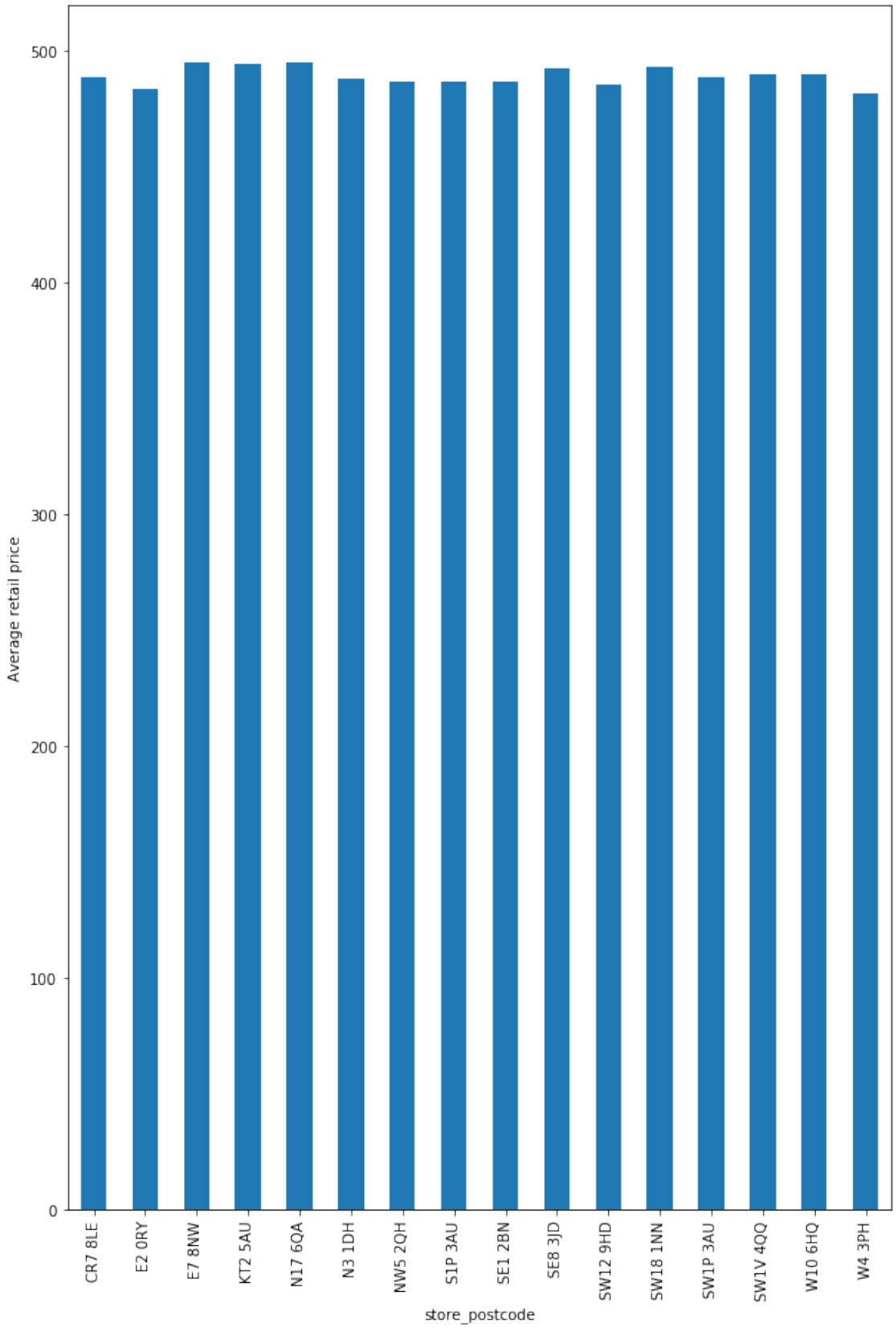
[2]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a226dda90>
```



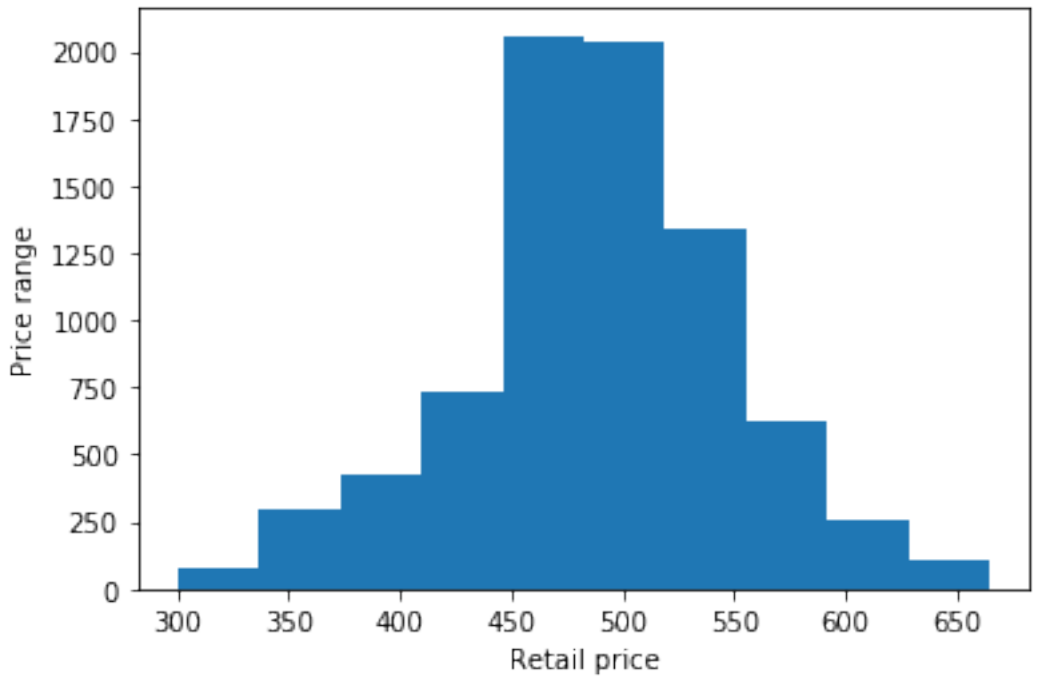
[59]:
```
# Problem 4
import matplotlib.pyplot as plt
laptop_df =  pd.read_csv('/Users/nivethida/Downloads/LaptopSalesJanuary2008.
 ↪csv')

# Sub-problem-1: Clean up column headers
laptop_df.columns = laptop_df.columns.str.strip().str.lower().str.replace(' ',␣
 ↪'_').str.replace('(', '').str.replace(')', '').str.replace("?","")
print("Cleaned up colums \n",laptop_df.head())

# Sub-problem-2: Bar chart with the average retail price by store
data_plot = laptop_df.groupby('store_postcode').mean()['retail_price']
axis = data_plot.plot(kind='bar', figsize=[10, 15])
axis.set_ylabel('Average retail price')
print("Average values: \n", data_plot)
print("Maximum value : ", max(data_plot)) #Store E7 8NW has highest average
print("Minimum value : ", min(data_plot)) #Store W4 3PH has lowest average
```

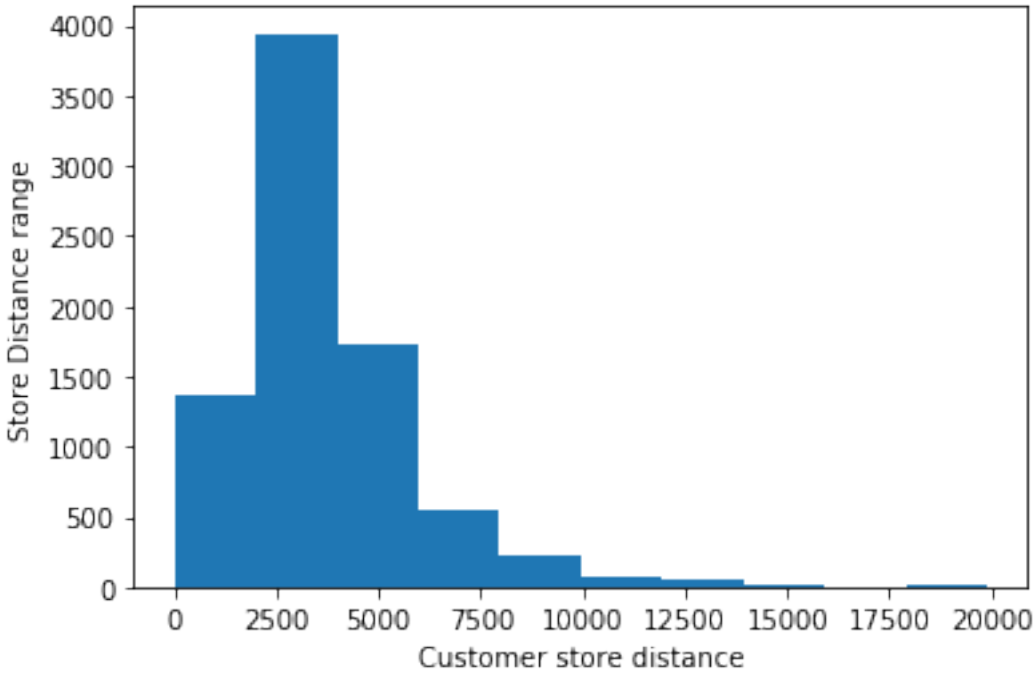
```
Cleaned up colums
             date  configuration customer_postcode store_postcode  \
0  1/1/2008 0:01            163          EC4V 5BH        SE1 2BN
1  1/1/2008 0:02            320           SW4 0JL       SW12 9HD
2  1/1/2008 0:04             23          EC3V 1LR         E2 0RY
3  1/1/2008 0:04            169          SW1P 3AU        SE1 2BN
4  1/1/2008 0:06            365          EC4V 4EG       SW1V 4QQ
```

```
   retail_price  screen_size_inches  battery_life_hours  ram_gb  \
0           455                  15                   5       1
1           545                  15                   6       1
2           515                  15                   4       1
3           395                  15                   5       1
4           585                  15                   6       2

   processor_speeds_ghz integrated_wireless  hd_size_gb bundled_applications  \
0                   2.0                 Yes          80                  Yes
1                   2.0                  No         300                   No
2                   2.0                 Yes         300                  Yes
3                   2.0                  No          40                  Yes
4                   2.0                  No         120                  Yes

   os_x_customer  os_y_customer  os_x_store  os_y_store  customerstoredistance
0         532041         180995    534057.0    179682.0            2405.873022
1         529240         175537    528739.0    173080.0            2507.558574
2         533095         181047    535652.0    182961.0            3194.001409
3         529902         179641    534057.0    179682.0            4155.202281
4         531684         180948    528924.0    178440.0            3729.298057
Average values:
 store_postcode
CR7 8LE     488.619048
E2 0RY      483.171729
E7 8NW      494.381443
KT2 5AU     493.904762
N17 6QA     494.634146
N3 1DH      487.368421
NW5 2QH     486.580460
S1P 3AU     486.250000
SE1 2BN     486.680195
SE8 3JD     492.177778
SW12 9HD    485.295699
SW18 1NN    493.038922
SW1P 3AU    488.506858
SW1V 4QQ    489.344978
W10 6HQ     489.866667
W4 3PH      481.006289
Name: retail_price, dtype: float64
Maximum value :    494.6341463414634
Minimum value :    481.0062893081761
```


Average retail price
0
100
200
300
400
500
CR7 8LE
E2 0RY
E7 8NW
KT2 5AU
N17 6QA
N3 1DH
NW5 2QH
S1P 3AU
SE1 2BN
SE8 3JD
SW12 9HD
SW18 1NN
SW1P 3AU
SW1V 4QQ
W10 6HQ
W4 3PH
store_postcode

[5]:
```
# Problem 4
# Sub-problem-3: Histograms for Retail Price, CustomerStoreDistance
from matplotlib import pyplot as plt
import numpy as np
laptop_df =  pd.read_csv('/Users/nivethida/Downloads/LaptopSalesJanuary2008.
 ↪csv')
laptop_df.head(5)

# Create histograms for Retail Price
fig, ax = plt.subplots()
ax.hist(laptop_df['Retail Price'])
ax.set_xlabel('Retail price')
ax.set_ylabel('Price range')
plt.show()

# Create histograms for CustomerStoreDistance
fig, ax = plt.subplots()
ax.hist(laptop_df['CustomerStoreDistance'])
ax.set_xlabel('Customer store distance')
ax.set_ylabel('Store Distance range')
plt.show()
```




Store Distance range
4000
3500
3000
2500
2000
1500
1000
500
0
0
2500
5000
7500
10000
12500
15000
17500
20000
Customer store distance