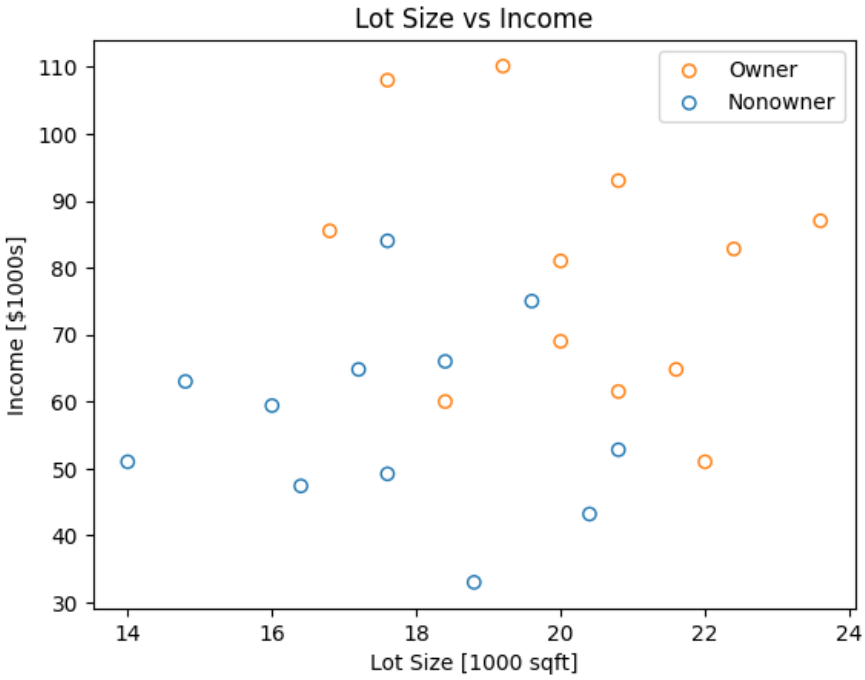
QM-7063 Data Mining
Professor: Dr. Abdulrashid
Learning Practice 1 – Noah L. Schrick

3.2 Sales of Riding Mowers: Scatter Plots.
A company that manufactures riding mowers wants to identify the best sales prospects for an intensive sales campaign. In particular, the manufacturer is interested in classifying households as prospective owners or nonowners on the basis of Income (in $1000s) and Lot Size (in 1000 ft2). The marketing expert looked at a random sample of 24 households, given in the file RidingMowers.csv.

a. Using Python, create a scatter plot of Lot Size vs. Income, color-coded by the outcome variable owner/nonowner. Make sure to obtain a well-formatted plot (create legible labels and a legend, etc.).

```
import pandas as pd
import matplotlib.pyplot as plt

mowers_df = pd.read_csv('RidingMowers.csv').squeeze("columns")

fig, ax = plt.subplots()
for val, color in ('Owner', 'C1'), ('Nonowner', 'C0'):
    subset_df = mowers_df[mowers_df.Ownership == val]
    ax.scatter(subset_df.Lot_Size, subset_df.Income, color='none', edgecolor=color)
ax.set_xlabel('Lot Size [1000 sqft]')
ax.set_ylabel('Income [$1000s]')
ax.legend(["Owner", "Nonowner"])
plt.title("Lot Size vs Income")
plt.show()
```



3.3 Laptop Sales at a London Computer Chain: Bar Charts and Boxplots.

The file LaptopSalesJanuary2008.csv contains data for all sales of laptops at a computer chain in London in January 2008. This is a subset of the full dataset that includes data for the entire year.

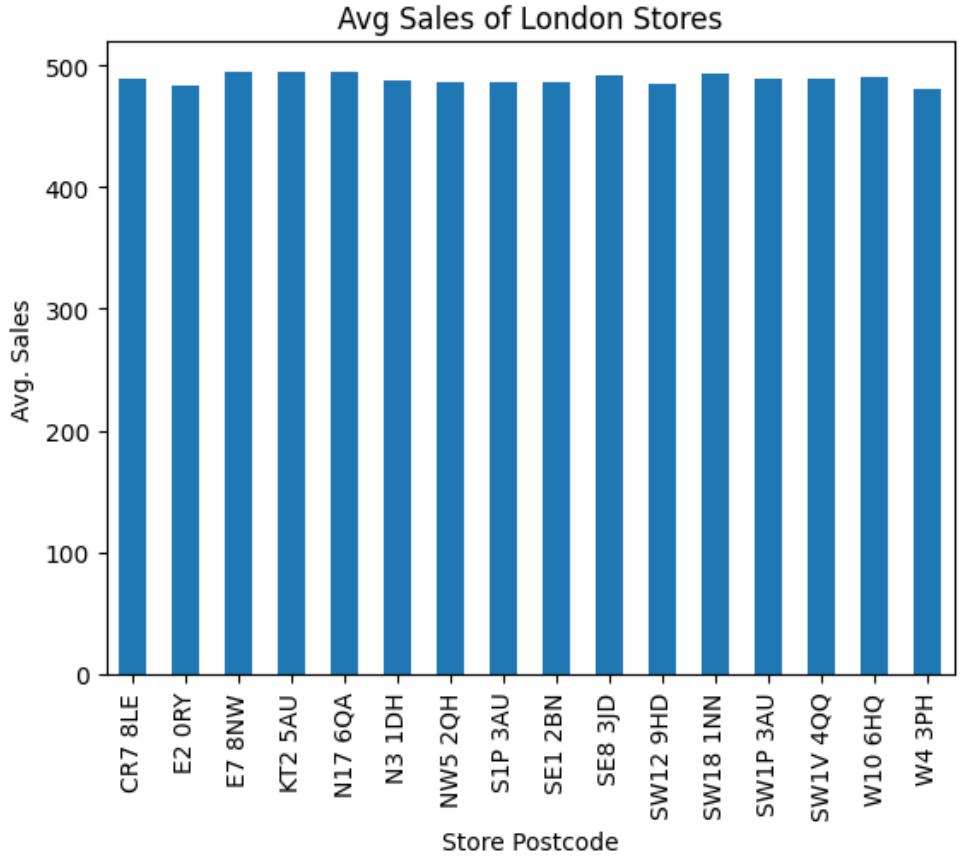
a. Create a bar chart, showing the average retail price by store. Which store has the highest average? Which has the lowest?

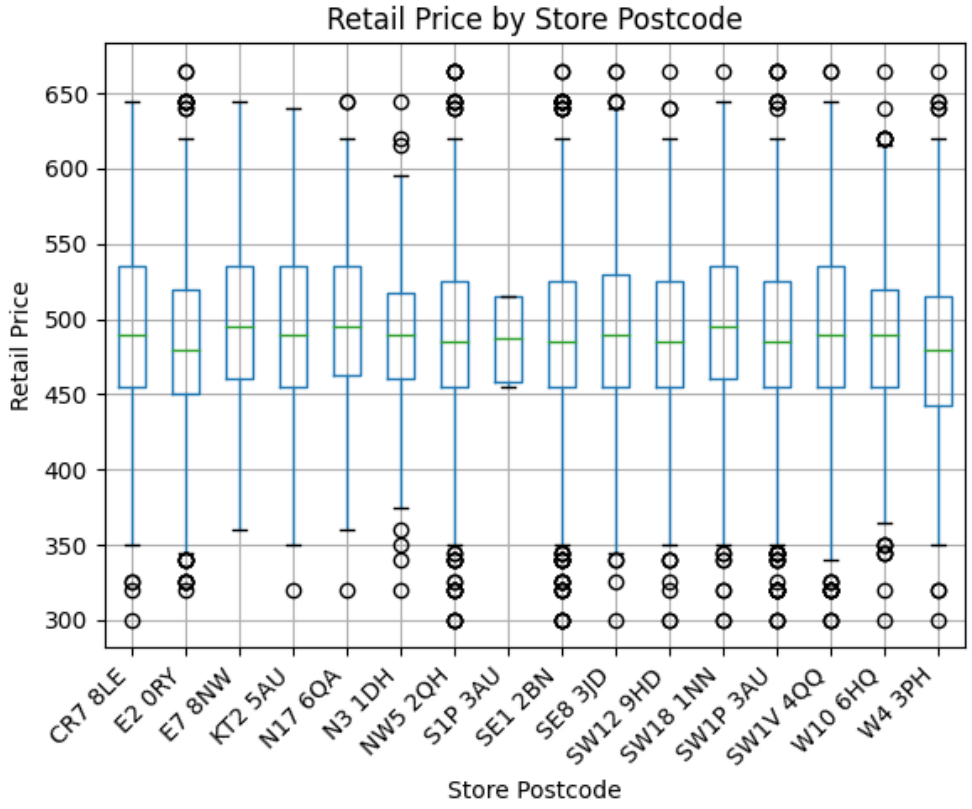
b. To better compare retail prices across stores, create side-by-side boxplots of retail price by store. Now compare the prices in the two stores from (a). Does there seem to be a difference between their price distributions?

```
laptops_df = pd.read_csv('LaptopSalesJanuary2008.csv')

## Part A:
# bar chart
laptops_retail_df = laptops_df.groupby('Store Postcode').mean(numeric_only=True)['Retail Price']
laptop_avg_sales = laptops_retail_df.plot(kind='bar')
laptop_avg_sales.set_ylabel('Avg. Sales')
laptop_avg_sales.set_title('Avg Sales of London Stores')

# highest avg
print("Max Avg. Sales:", round(laptops_retail_df.max(), 2), "from Store:", laptops_retail_df.idxmax())

# lowest avg
print("Min Avg. Sales:", round(laptops_retail_df.min(), 2), "from Store:", laptops_retail_df.idxmin())

## Part B:
# box plots
ax = laptops_df.boxplot(column='Retail Price', by='Store Postcode')
ax.set_ylabel('Retail Price')
plt.xticks(rotation=45, ha='right')
plt.suptitle('')  # Suppress the given title
plt.title('Retail Price by Store Postcode')
```

Max Avg. Sales: 494.63 from Store: N17 6QA
Min Avg. Sales: 481.01 from Store: W4 3PH

The median of each store's sales are all relatively similar, all within 25 pounds. The minimum and maximum of each store varies slightly more, with the exception of store S1P 3AU, which has only a few sales that are all around the same Retail Price. Some stores, such as SE1 2BN, have noticeably more outliers when compared to stores such as E7 8NW.


Avg Sales of London Stores
Avg. Sales
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


Retail Price by Store Postcode
Retail Price
300
350
400
450
500
550
600
650
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