Data115_2025Spring_PS2_Team7_Calderon_Hammad

February 20, 2025

DATA 115: Introduction to Data Analytics. Fall 2024

Problem Set 2

Team 7

Abdelaziz Hammad, Jeff Calderon

1. In your own words, provide a definition or a short description for each of the following terms (and explain how they are different for the pairs of items):

(a) Bar chart vs Histogram

(b) Mean vs Median

(c) Pearson correlation vs Spearman's rank correlation Bar chart vs Histogram

A Bar chart has two axes which can represent numerical and categorical data. In a column bar chart the horizontal axes will show categories and the vertical axes might show a value associated with each category. For instance you might be reporting exam scores from students in a class, so the names of each would be on the horizontal and their scores on the vertical axes. A histogram shows a distribution of a continuos numerical value. In keeping with the previous example, a histogram would show us the distribution of scores on the exam. We would set appropriate bin sizes and count the number of students which scored within each bin. The horizontal axes would show bin sizes, while the vertical height of the bin represents the number of students within that bin. The bins are also typically touching, unlike a bar graph which separates the bars. Since in the bar graph each bar represents a different category

Mean vs Median

Mean and median are both measures of central tendency. So they both inform us about what any observation is likely to be within a sample. They differ in important ways. The statistical mean is the result of averaging all the numbers. Simply add all the measured value of each observation and divide by the number of observations. The median is found by lining up all the values, sorted smallest to greatest, then determining the middle value. The mean is useful but is sensitive to outliers. In other words, infrequent but large magnitude observations might move the mean such that it is not representative of the central tendency. If most students received 99/100 but a few received 0/100 by not showing up on exam day, the mean will be affected downward and misrepresent the average score of test takers. On the other hand the mean will not be as affected.

Pearson Correlation VS Spearman's Rank Correlation

Pearson correlation is a way to determine the extent to which two continuous variables are related linearly. Two continuous variables would be a person weight (W) and Height (H). If we measure height vs weight then for any value H there is a value W for the same observation, so when compare how each value of H tends to vary with W using a pearson correlation, it will indicate if they are moving together in a line. If when H increases W also increases then they are positively correlated and the Pearson correlation would be positive. If W decreases as H increases (or vice-versa), then the correlation is negative. The Pearson Correlation is normalized so that the number is always in between -1 and 1. A zero Pearson correlation would indicate that there is no linear correlation.

Spearman's Rank Correlation is another way to test for a relationship between two variables but is more general and applicable in more circumstances. It can be used to check if one variable increases/decreases as the other increases/decrease in a way that measures how consistent those increases or decreases are. It does so based by ranking the two variables and examining how the differences in ranks vary in each data point. For instance if the heaviest person (rank 1) was the second tallest (rank 2) than this would have difference of -1. Summing up these differences and normalizing (among other things) is how we reach a value between -1 and 0.

2. Load the data in `GME_Stock` as a dataframe in Python.

(a) Plot the values in the closing prices column as a line plot.

(b) Make a scatterplot of opening prices vs. closing prices.

(c) Add a new column to your dataframe whose rows represent the change in price during each day and make a histogram of these values.

(d) Make a bar chart presenting the in-state tuition data presented in this table:

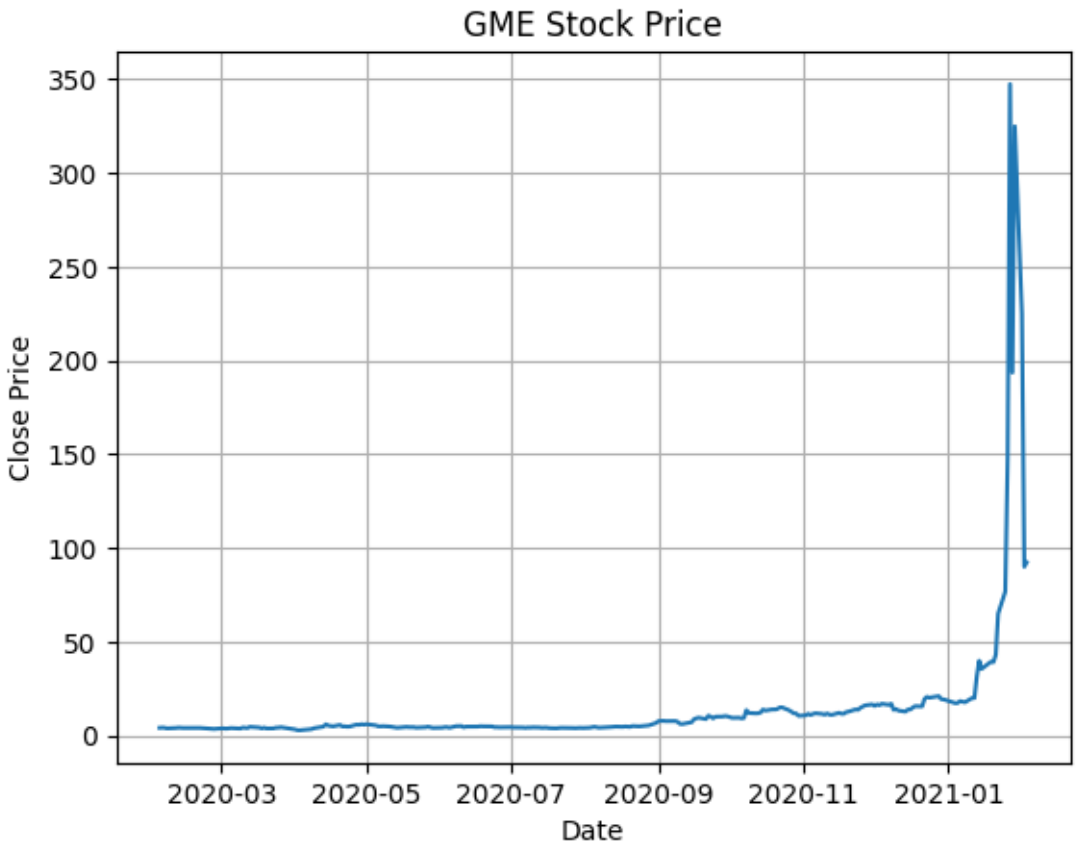
School	WSU	EWU	UW	UI	CWU
Tuition	11841	7526	11465	8304	8273

(e) Provide a brief justification for the order you placed the bars from left to right in (d).

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt

# Load the data
data = pd.read_csv('GME_Stock.csv')
data.head(2)
data.Date = pd.to_datetime(data.Date)
x_values = data.Date
y_values = data.Close

#line plot
print('A Line Plot')
fig, ax = plt.subplots()
```

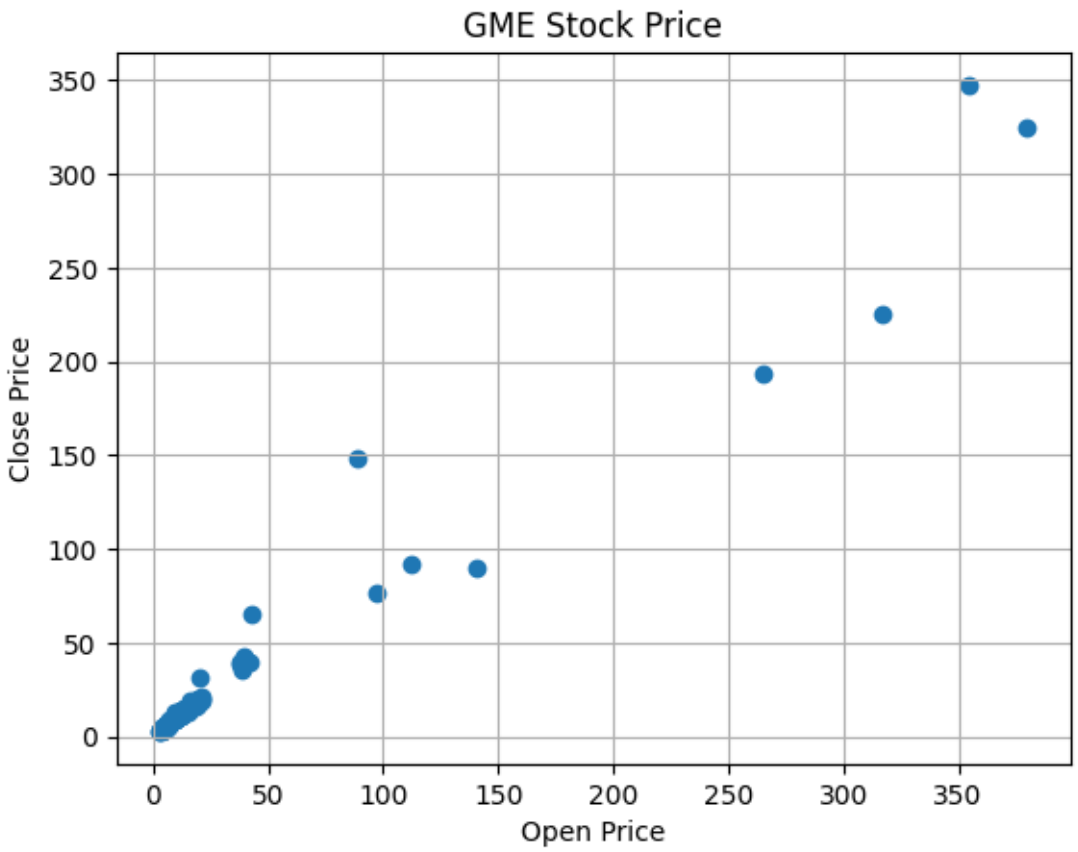
```
ax.plot(x_values, y_values)
ax.set(xlabel='Date', ylabel='Close Price', title='GME Stock Price')
ax.grid()
#plt.show()
```

```
A Line Plot
```



[2]:
```
data.columns
fig2, ax2 = plt.subplots()
ax2.scatter(data.Open, data.Close)
ax2.set(xlabel='Open Price', ylabel='Close Price', title='GME Stock Price')
ax2.grid()
print('(B) Scatter Plot')
plt.show()
```

```
(B) Scatter Plot
```

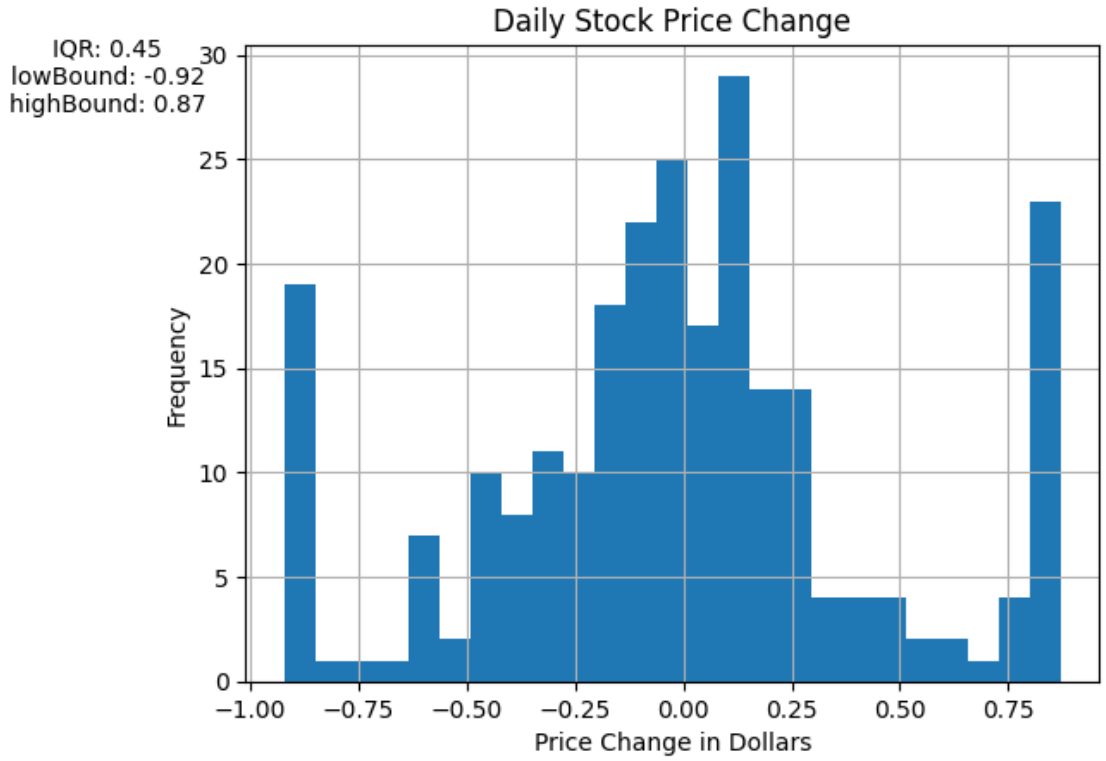


[3]:
```
# Add a new column
data['PriceChange'] = data.Close - data.Open

# Check that it worked
print(data.head(2))

# More informative without the outliers
Q1 = data['PriceChange'].quantile(0.25)
Q3 = data['PriceChange'].quantile(0.75)
IQR = Q3 - Q1
lowBound = round(Q1 - 1.5 * IQR, 2)
highBound = round(Q3 + 1.5 * IQR, 2)
IQR = round(IQR, 2)
#remove outliers
clipped_data = data['PriceChange'].clip(lower=lowBound, upper=highBound)

#Now we can make a histogram
fig3, ax3 = plt.subplots()
print('\n\n(C) Histogram')
```

```
ax3.hist(clipped_data, bins=25)
ax3.set(xlabel='Price Change in Dollars', ylabel='Frequency', title='Daily␣
 ↪Stock Price Change')
plt.gcf().text(0, 0.8, f'IQR: {IQR}\nlowBound: {lowBound}\nhighBound:␣
 ↪{highBound}', ha = 'center', fontsize = 10)

ax3.grid()
plt.show()
```

	Date	Open	High	Low	Close	Adj Close	Volume	PriceChange
0	2020-02-04	4.03	4.25	3.97	4.07	4.07	3563100	0.04
1	2020-02-05	4.15	4.41	4.14	4.18	4.18	2641700	0.03

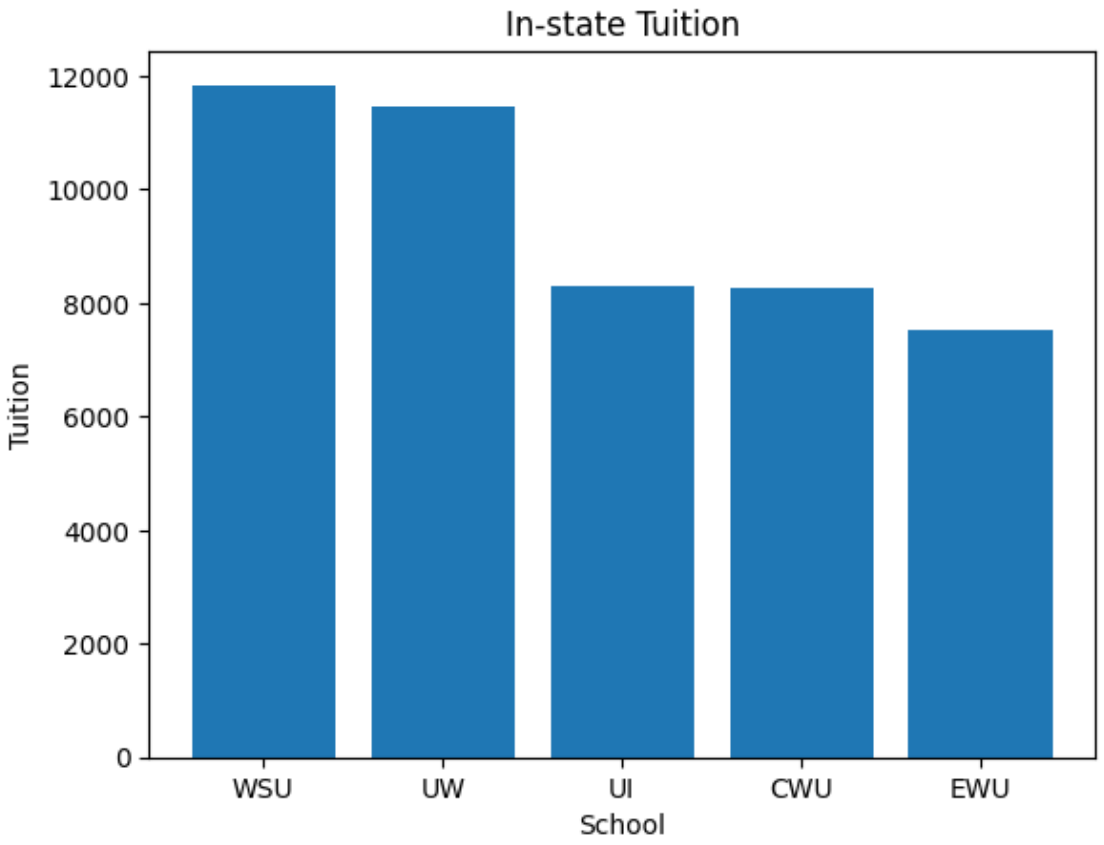
(C) Histogram



[4]:
```
# Create data for a bar chart
School_data = {'School': ['WSU', 'EWU', 'UW', 'UI', 'CWU'],}
Tuition = {'Tuition': [11841, 7526, 11465, 8304, 8273],}
df = pd.DataFrame(School_data)
df['Tuition'] = Tuition['Tuition']
df.sort_values(by='Tuition', ascending=False, inplace=True)
```

```
print(df.head())
# Create a bar chart
fig4, ax4 = plt.subplots()
ax4.bar(df.School, df.Tuition)
ax4.set(xlabel='School', ylabel='Tuition', title='In-state Tuition')
print('(D) Bar Chart')
plt.show()
```

```
  School  Tuition
0    WSU    11841
2     UW    11465
3     UI     8304
4    CWU     8273
1    EWU     7526
(D) Bar Chart
```



(e) Order of Bars
I ordered the bars from largest tuition to smallest. This is because the data is comparing the schools by the in-state cost. Presumably the reader would be concerned about the overall cost of attendance and want to compare the cost of each school to other schools.

3. Load the data in `COL` as a dataframe in Python.

(a) Decide which rows are outliers in this data and describe and justify how you determined their outlier status.

(b) For each row you identified, if you were performing EDA on this dataset, would you include its values in your analysis and plots? Why or why not?

(c) Make a scatterplot matrix of the numeric columns.

(d) Write a brief (no more than three sentences) summary of what you observe in the plot in (c).

(e) Choose a single subplot that seems most interesting to you and make a separate scatterplot of just those two columns with the points colored by the salary value.

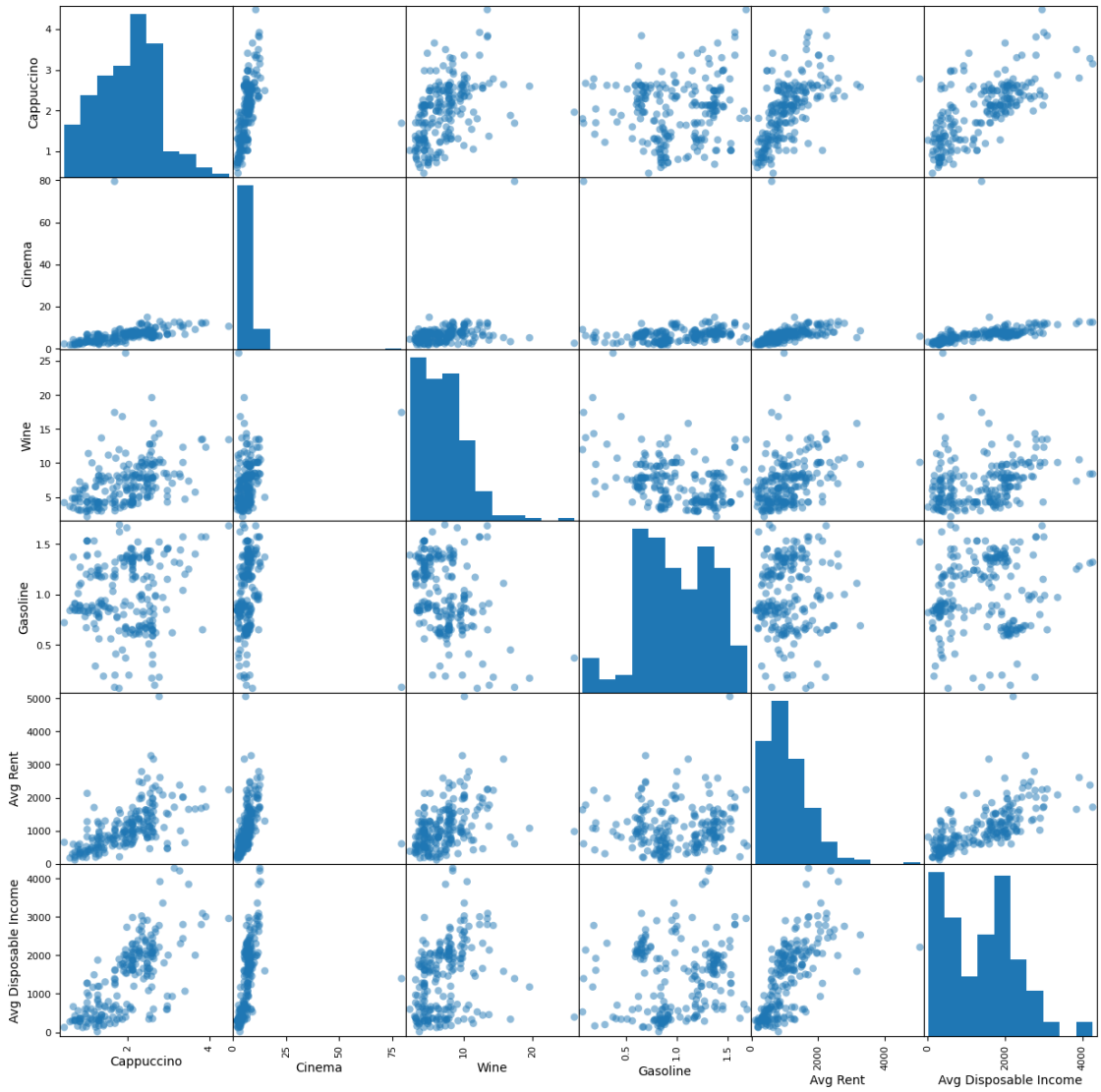
(f) Write a brief (no more than two sentences) summary of what you observe in the plot in (e).

[5]:
```
df2 = pd.read_csv('COL.csv')
#check format
print(df2.head())
# drop categorical data
df2_mod = df2.drop(columns=['City'], axis=1)
# find outliers
for col in df2_mod.columns:

    Q1 = df2[col].quantile(0.25)
    Q3 = df2[col].quantile(0.75)
    IQR = Q3 - Q1
    lowBound = round(Q1 - 1.5 * IQR, 2)
    highBound = round(Q3 + 1.5 * IQR, 2)
    IQR = round(IQR, 2)
    clipped_data = df2_mod[col].clip(lower=lowBound, upper=highBound)
print('(A) outlier rows where removed based on +-1.5*IQR\n')
print('This is in keeping with the standard practice of removing outliers\n')
print('(B) I would not include the value on the plots because they will change␣
 ↪the scale of the histograms and skew other metrics that we can use for␣
 ↪descriptive statistics\n')
print(' Rent and Disposable Income seem to have a correlation. while it is not␣
 ↪as strong as some others is something to note\n')

# Create a matrix plot
print('(C) Matrix Plot')
pd.plotting.scatter_matrix(df2_mod, figsize=(15,15), grid=True, marker='0')
plt.show()
print('(D) Summary of plots\n')
```

```
print('Cinema and Cappuccino seem to have some correlations with the other␣
 ↪variables, but the other variables do not seem to have any correlation with␣
 ↪each other\n')
print('This is discernible by looking at the shape of the scatter plots and␣
 ↪imagining a trend line. If the points are all balled up, then there is no␣
 ↪probably no correlation\n')

print('(E) Rent and Cappuccino')
fig5, ax5 = plt.subplots()
scat_plot = ax5.scatter(df2_mod['Avg Rent'], df2_mod.Cappuccino, c=df2_mod['Avg␣
 ↪Disposable Income'], cmap='viridis', edgecolors='k')
cbar = plt.colorbar(scat_plot)
cbar.set_label('Average Disposable Income')
ax5.set(xlabel='Rent', ylabel='Cappuccino', title='Rent vs Cost of Cappuccino')
ax5.grid()
plt.show()
print('\n\n (F) Cappuccino and Rent prices are positively correlated.')
print('There is also a correlation with high disposable income since the colors␣
 ↪lighten as the rent and cappuccino prices rise.\n')
print('On the other hand, some low rent places with cheap cappuccino can be␣
 ↪seen.')
print('These light color points are interesting outliers. Viewers may note␣
 ↪desirable places with low cost of living yet high incomes')
```

```
       City  Cappuccino  Cinema   Wine  Gasoline  Avg Rent  \
0  Lausanne        3.15   12.59   8.40      1.32   1714.00
1    Zurich        3.28   12.59   8.40      1.31   2378.61
2    Geneva        2.80   12.94  10.49      1.28   2607.95
3     Basel        3.50   11.89   7.35      1.25   1649.29
4     Perth        2.87   11.43  10.08      0.97   2083.14

   Avg Disposable Income
0                4266.11
1                4197.55
2                3917.72
3                3847.76
4                3358.55
(A) outlier rows where removed based on +-1.5*IQR

This is in keeping with the standard practice of removing outliers

(B) I would not include the value on the plots because they will change the
scale of the histograms and skew other metrics that we can use for descriptive
statistics

 Rent and Disposable Income seem to have a correlation. while it is not as
strong as some others is something to note
```

(C) Matrix Plot

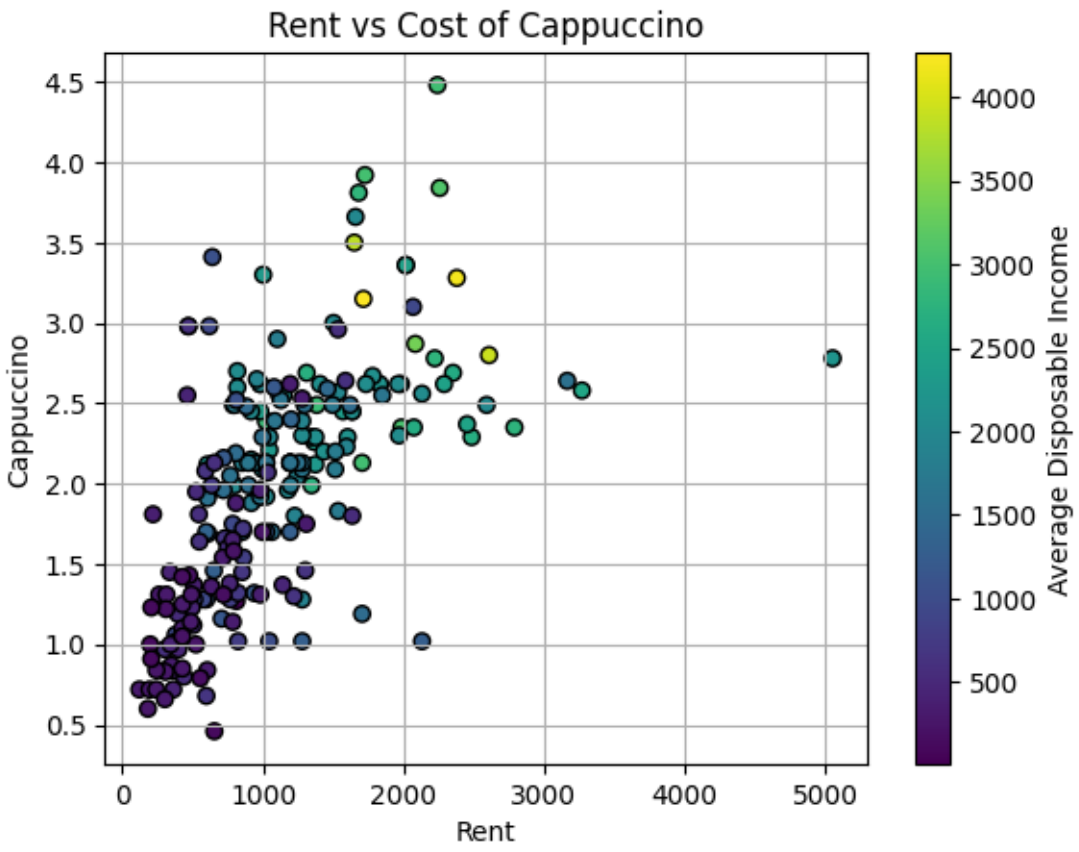


(D) Summary of plots

Cinema and Cappuccino seem to have some correlations with the other variables, but the other variables do not seem to have any correlation with each other

This is discernible by looking at the shape of the scatter plots and imagining a trend line. If the points are all balled up, then there is no probably no correlation

(E) Rent and Cappuccino



(F) Cappuccino and Rent prices are positively correlated.
There is also a correlation with high disposable income since the colors lighten as the rent and cappuccino prices rise.

On the other hand, some low rent places with cheap cappuccino can be seen. These light color points are interesting outliers. Viewers may note desirable places with low cost of living yet high incomes

4. Load the data in `Height_Weight_Age_Sex` as a dataframe in Python.

(a) Create boxplots for the height and weight columns separately. Comment on the symmetry and skewness, if any, for their distributions using these plots.

(b) Create histograms for the height and weight columns separately. Comment on the symmetry and skewness, if any, for their distributions using these plots. Are your conclusions based on the boxplots consistent with those based on densities?

(c) Create separate boxplots for the weight data separated by the male variable. What do you observe about the two distributions?

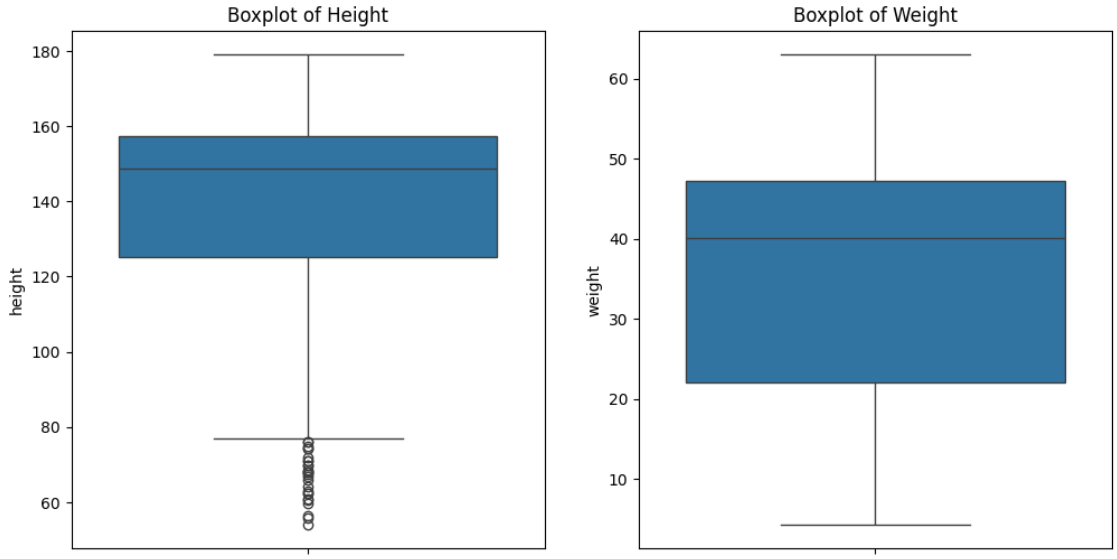
(d) Define variable `BMI`, applying the formula for metric system: $BMI = Weight/(\frac{Height}{100})^2$. Create a binary indicator variable `Underweight` that takes value "1" for individuals with BMI < 18.5 and value "0" otherwise. Add both of these variables as a columns to the dataframe.

(e) Create separate histograms for the BMI column separated by the male variable. What do you observe about the two distributions?

(f) Make a scatterplot of height vs. weight for the full dataset that distinguishes both the male variable and the underweight variable. What do you observe?

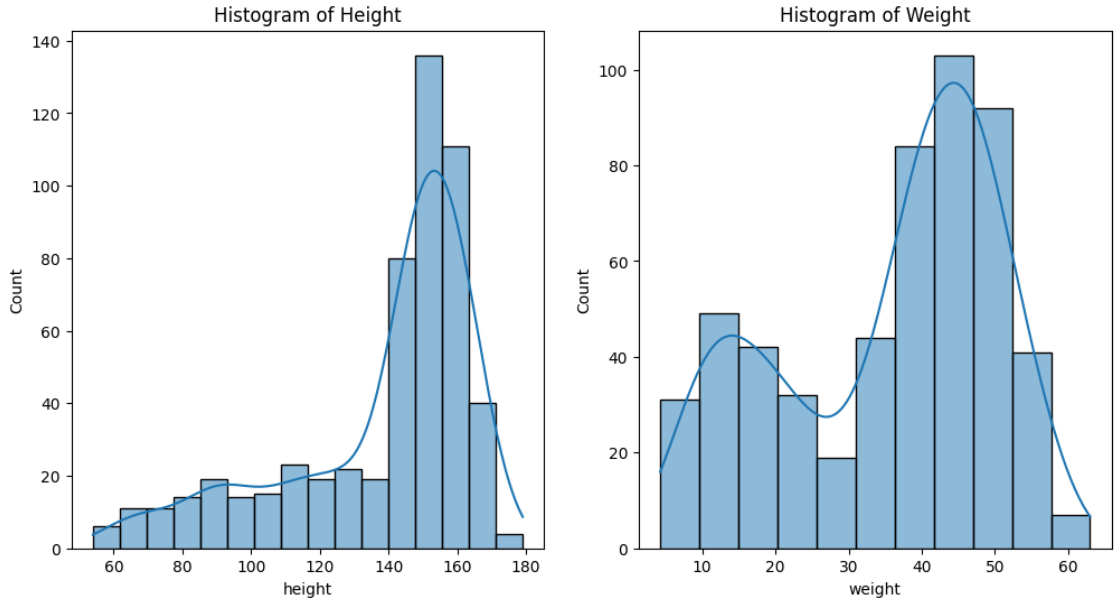
(a) since the boxplot for height shows that the median is closer to the top of the box and the upper whisker is shorter, the distribution of height is left skewed compared to weight boxplot with no noticeable skewness.

[6]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('Height_Weight_Age_Sex.csv')
plt.figure(figsize=(12, 6))

plt.subplot(1, 2, 1)
sns.boxplot(y=df['height'])  # Use 'height' instead of 'Height'
plt.title('Boxplot of Height')

plt.subplot(1, 2, 2)
sns.boxplot(y=df['weight'])  # Use 'weight' instead of 'Weight'
plt.title('Boxplot of Weight')

plt.show()
```



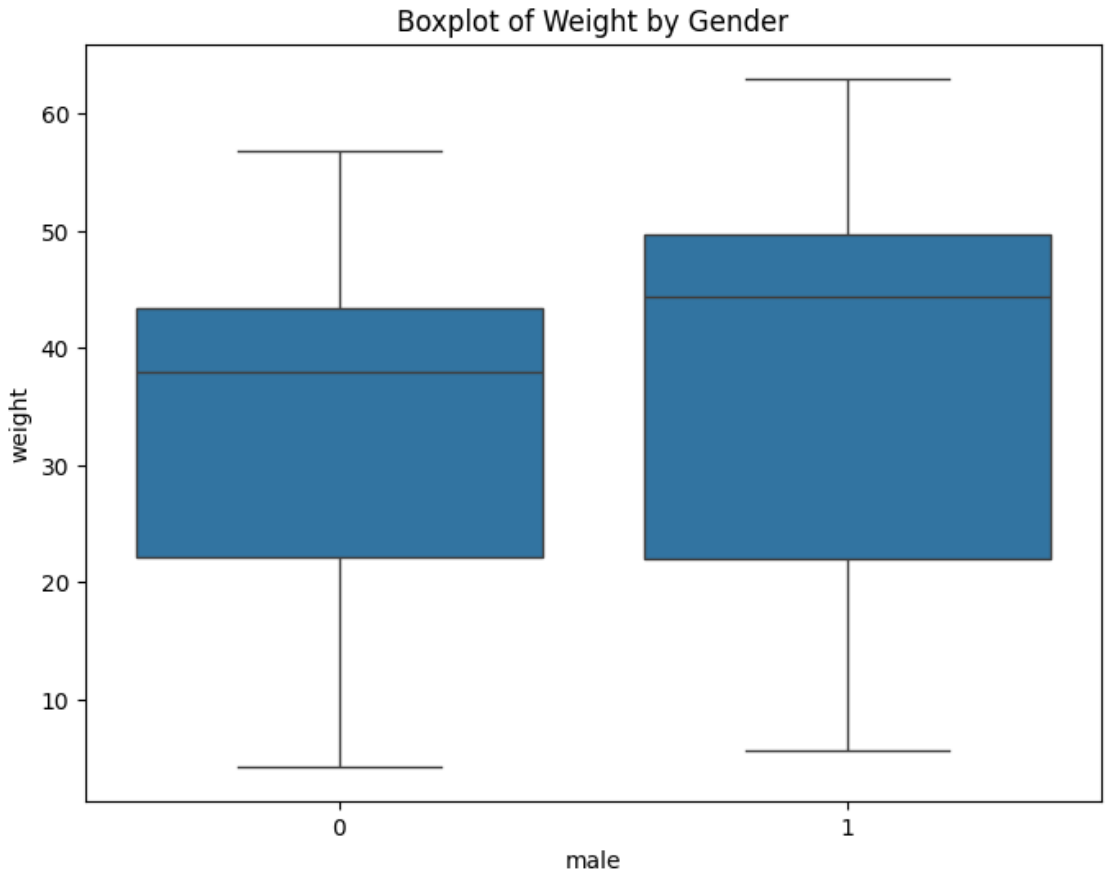
(b) both histograms for height and weight are skewed to the left. There are some very short people causing this, which pulls the average left (down) even though most samples are bigger larger. Similarly we see the left skew of the weight plot, due to a some very light people pulling the average down. So, both of these means will be lower than the median.

[7]:
```
df = pd.read_csv('Height_Weight_Age_Sex.csv')
plt.figure(figsize=(12, 6))

plt.subplot(1, 2, 1)
sns.histplot(df['height'], kde=True)
plt.title('Histogram of Height')

plt.subplot(1, 2, 2)
sns.histplot(df['weight'], kde=True)
plt.title('Histogram of Weight')

plt.show()
```



(c) the boxplot is left skewed since the median is closer to the top and the tail is shorter

[8]:
```
df = pd.read_csv('Height_Weight_Age_Sex.csv')
plt.figure(figsize=(8, 6))
sns.boxplot(x='male', y='weight', data=df)
plt.title('Boxplot of Weight by Gender')
plt.show()
```



(d)

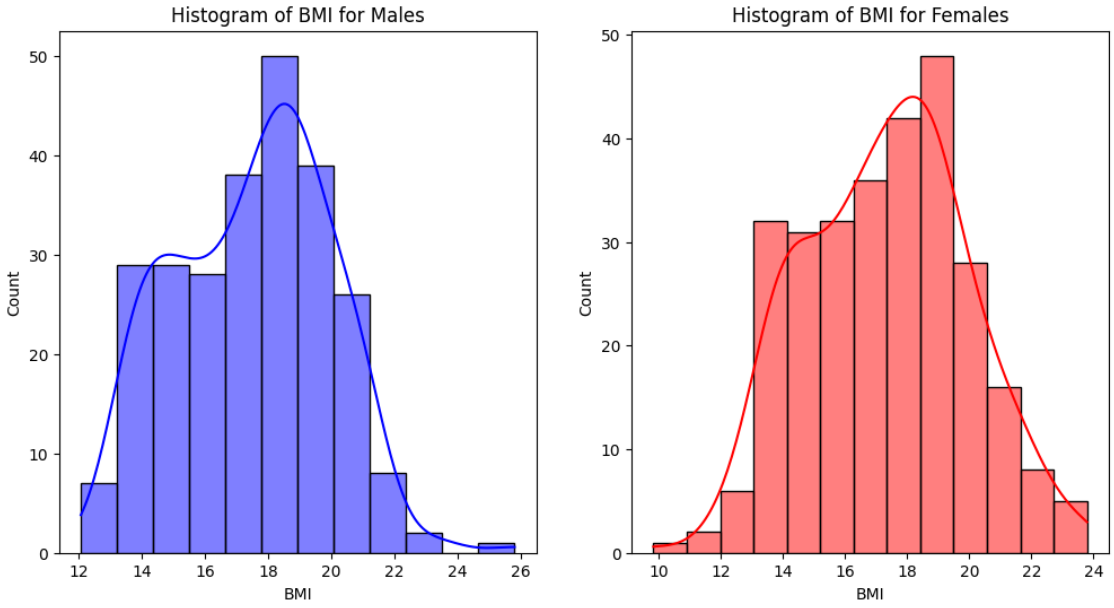
[9]:
```
df = pd.read_csv('Height_Weight_Age_Sex.csv')
df['BMI'] = df['weight'] / (df['height'] / 100) ** 2
df['Underweight'] = (df['BMI'] < 18.5).astype(int)
print(df.head())
```

```
    height     weight   age  male        BMI  Underweight
0  151.765  47.825606  63.0     1  20.764297            0
1  139.700  36.485807  63.0     0  18.695244            0
2  136.525  31.864838  65.0     0  17.095718            1
3  156.845  53.041914  41.0     1  21.561444            0
4  145.415  41.276872  51.0     0  19.520384            0
```

(e) the bmi distribution of males on the histogram is approximately symetrical, but females histogram looks left skewed and a longer tail towards lower values

[10]:
```
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
sns.histplot(df[df['male'] == 1]['BMI'], kde=True, color='blue', label='Male')
plt.title('Histogram of BMI for Males')
```

```
plt.subplot(1, 2, 2)
sns.histplot(df[df['male'] == 0]['BMI'], kde=True, color='red', label='Female')
plt.title('Histogram of BMI for Females')
plt.show()
```



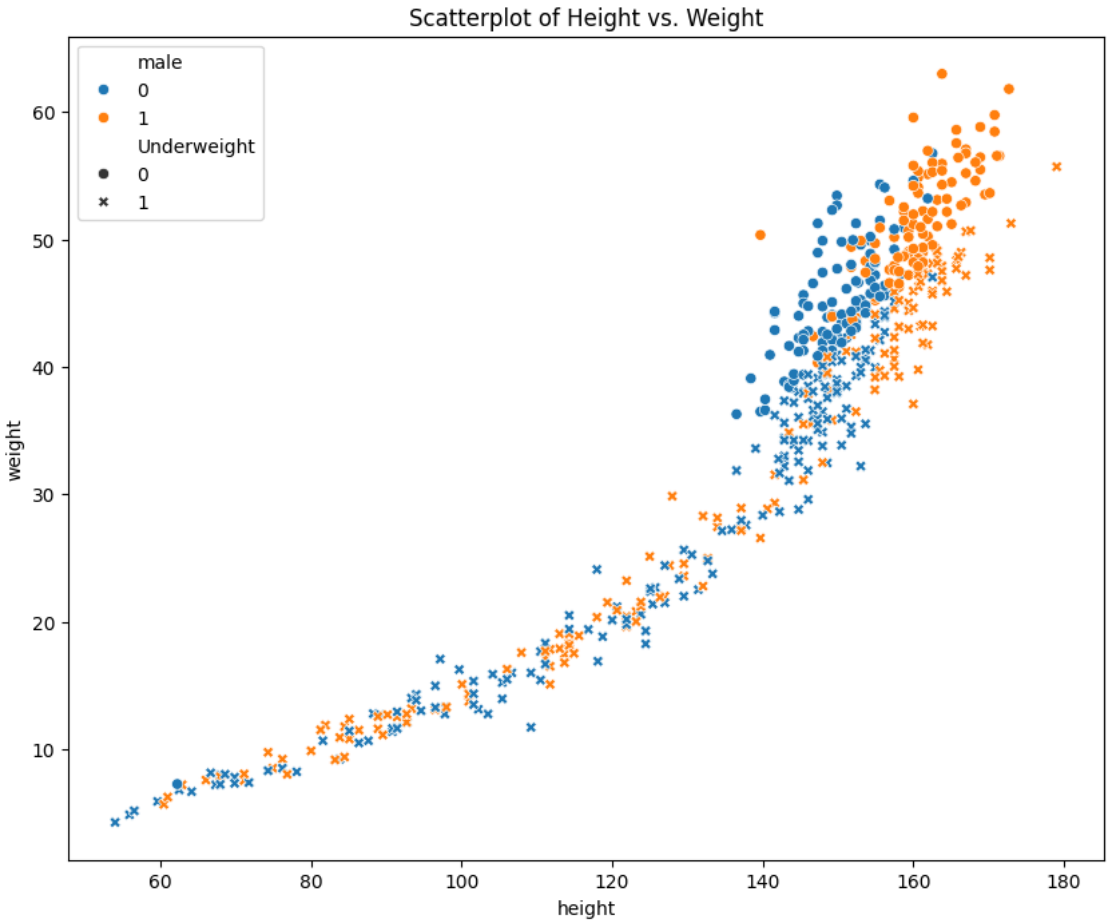
(f) the scatterplot shows us the correlation between weight and gender and height

1.Males tend to have a higher weight than females

2. as height increases the weight increases aswell

3. underweight individuals are scattered throughout different heights but they tend to be more at lower heights

[11]:
```
plt.figure(figsize=(10, 8))
sns.scatterplot(x='height', y='weight', hue='male', style='Underweight',␣
↪data=df)
plt.title('Scatterplot of Height vs. Weight')
plt.show()
```



5. Read the following examples about Simpson's Paradox: How to lie with statistics?. Fill in the following table with ratios of hits to attempts so that player A has a higher batting average in both season 1 and season 2 but player B has a higher overall batting average for the two seasons combined.

hits/attempts	Season 1	Season 2
Player A	25 / 100	35 / 120
Player B	15 / 100	50 / 200