DS 1000B – Assignment 2

Total Marks: 100

Due date: Monday, February 2nd, 2026 at 8:00 PM

Submission Platform: All assignments must be submitted via Gradescope.

File Format: Submit a **single** PDF file containing all of your work. Please note that Gradescope only displays your most recent submission and this is the version that will be graded. **You will receive a grade of zero in each case where**

- Submission is not in PDF format.
- **Questions have no pages assigned to them on Gradescope** (i.e. did not submit anything for that question).
- Submission is illegible (e.g. blurry, too small to read comfortably without zooming in).

You must submit the following as a single PDF file:

- Part 1 – Written responses (these may be handwritten or typed). Multiple choice questions do not require work to be shown.

  Unless otherwise specified, please show your work.
- Part 2 – Python coding exercises with some written responses. *Acceptable submission formats:*
  - Screenshot showing both the code cell and the corresponding output (from Google Colab, a local Jupyter notebook, or another Python environment).
  - Copy-pasted code with the output clearly labeled below it in a screenshot or text (where applicable).

Individual Work: Each student must submit their own original work. You may discuss questions with your classmates, but you must write up your solutions independently. Provided your submission adheres to the requirements outlined above, the method you use to generate the document is at your discretion.
Rounding requirements: Please round answers to 2 decimal places.

Part 1 – Written Responses

Data Science in Finance

[26 marks]

Making Cents of Subscriptions

Rocket Money is a leading personal finance application that helps users manage subscriptions and monitor spending habits. As a Senior Data Scientist (Decisions), your team has the results of the following survey to assess subscription usage among University of Western Ontario students.

Q1. The following is a preview of the data collected:

student_id	gender	num_subscriptions	monthly_cost_usd	monthly_income_usd	monthly_discretionary_usd
582334538	Female	5	36.51	900	180
398704996	Female	5	38.00	0	0
219540831	Male	0	0.00	850	160
553035110	Male	2	22.42	0	0
890779946	Non-binary	1	18.11	920	200

(a) (4 marks) Who are the individuals in this dataset? What are the variables?

Solution: The individuals in this dataset are University of Western Ontario students who participated in the survey. The variables are `student_id`, `gender`, `num_subscriptions`, `monthly_cost_usd`, `monthly_income_usd`, and `monthly_discretionary_usd`.

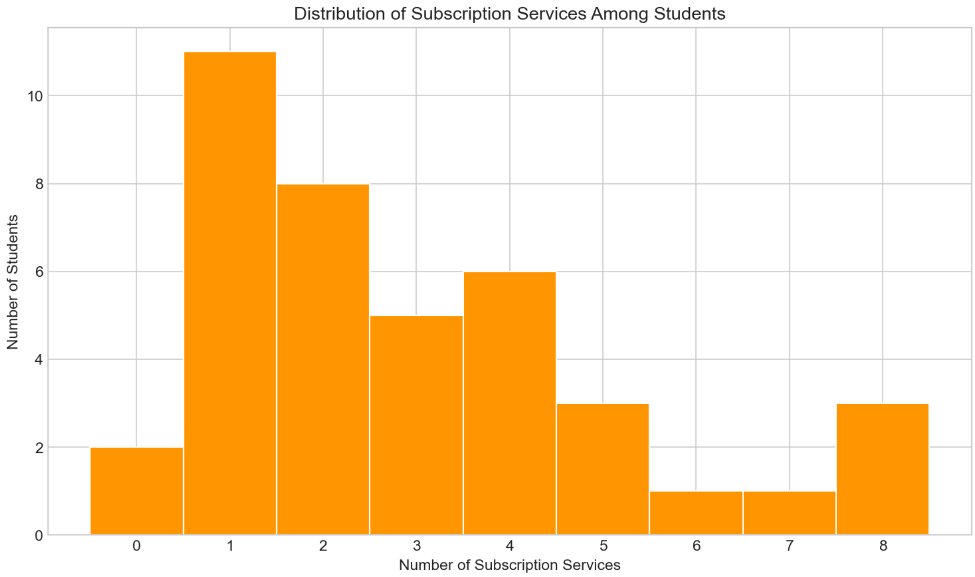
(b) (2 marks) Which variables should be considered categorical, and which should be considered numerical?

Solution: The categorical variables are `gender` (which describes the gender of the student) and `student_id`. The numerical variables are `num_subscriptions` (which counts the number of subscriptions), `monthly_cost_usd`, `monthly_income_usd`, and `monthly_discretionary_usd`.

(c) (2 marks) List all appropriate visualizations from Chapter 1 to display the distribution of gender in this dataset.

Solution: A bar chart or pie chart would be appropriate since gender is a categorical variable.

Q2. As part of your analysis, you produced the following histogram of the distribution of number of subscriptions:



(a) (2 marks) According to this histogram, how many individuals were surveyed?

Solution: Adding up the heights of all the bars in the histogram, we find that the total number of individuals surveyed is $2 + 11 + 8 + 5 + 6 + 3 + 1 + 1 + 3 = 40$.

(b) (2 marks) What is the mean of the distribution?

Solution: The mean number of subscriptions is given by

$$\bar{x} = \frac{0(2) + 1(11) + 2(8) + 3(5) + 4(6) + 5(3) + 6(1) + 7(1) + 8(3)}{40} = \frac{118}{40} = 2.95$$

(c) (1 mark) What is the shape of the distribution?

Solution: This distribution is right-skewed.

(d) (2 marks) Would you use the mean or the median to describe the typical number of subscriptions? Justify.

Solution: Given the right-skewed distribution, the median would be a better measure of central tendency than the mean. The mean is pulled toward the right tail by the larger values, making it less representative of a typical observation. The median is resistant to outliers and skewness.

Q3. The following table is a sample of the monthly subscription costs (in USD) for 12 participants from the study.

70.30	78.40	13.62	65.68
23.48	9.99	30.00	12.99
155.00	30.99	40.99	29.99

Table 1: Monthly subscription costs (in USD) for 12 participants.

(a) (5 marks) Write down the five-number summary for this distribution. *No justification necessary but incorrect answers without justification (where appropriate) will receive no partial marks.*

Solution: Sorting the data in ascending order, we obtain:

$$9.99,\ 12.99,\ 13.62,\ 23.48,\ 29.99,\ 30.00,\ 30.99,\ 40.99,\ 65.68,\ 70.30,\ 78.40,\ 155.00$$

With $n = 12$ observations:

- Minimum: 9.99
- Q_1: Average of 3rd and 4th values: $Q_1 = \dfrac{13.62 + 23.48}{2} = 18.55$
- Median (Q_2): Average of 6th and 7th values: $Q_2 = \dfrac{30.00 + 30.99}{2} = 30.50$ (rounded to 2 decimal places)
- Q_3: Average of 9th and 10th values: $Q_3 = \dfrac{65.68 + 70.30}{2} = 67.99$
- Maximum: 155.00

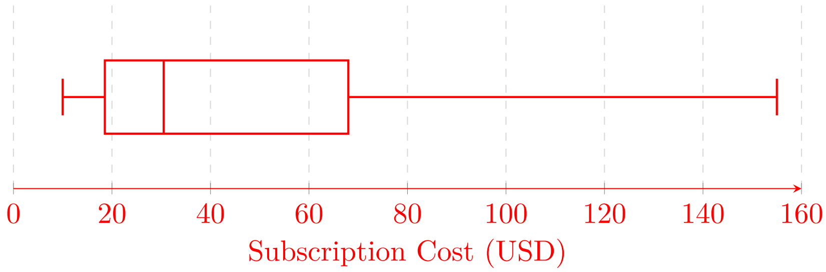
Five-number summary: (9.99, 18.55, 30.50, 67.99, 155.00)

Comment to Graders:

If a student computes Q_1 or Q_3 as 19.05, they are likely using my solutions from last term. Please deduct marks appropriately (e.g. 0.5 or 1 marks off for that error), **do not comment on it in the marking** and notify me via Piazza.

(b) (2 marks) Draw a boxplot for the data from Table 1.

Solution:



(c) (1 mark) Use the IQR rule to check for potential outliers. If there are any, what are they? If there are none, write “NONE".

Solution: First, calculate the IQR:

$$\text{IQR} = Q_3 - Q_1 = 67.99 - 18.55 = 49.44$$

Lower fence: $Q_1 - 1.5 \times \text{IQR} = 18.55 - 1.5(49.44) = 18.55 - 74.16 = -55.61$

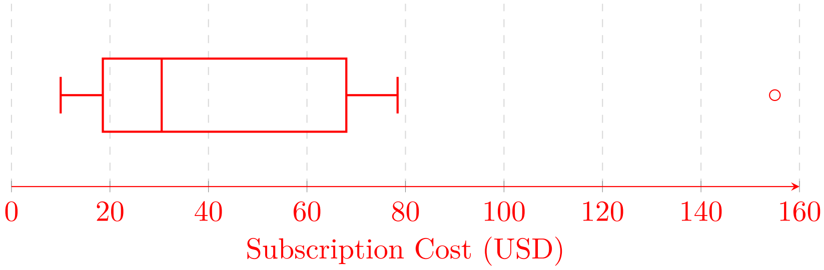
Upper fence: $Q_3 + 1.5 \times \text{IQR} = 67.99 + 1.5(49.44) = 67.99 + 74.16 = 142.15$

Any values below -55.61 or above 142.15 are potential outliers.

Since $155.00 > 142.15$, the value \$155.00 is a potential outlier. No values are below the lower fence.

(d) (3 marks) Draw a modified boxplot for the data from Table 1.

Solution:



Environmental Data Science [13 marks]

As a Machine Learning Scientist working at the Scripps Institution of Oceanography, your mission is to analyze the Mauna Loa CO_2 data, which reports monthly average concentrations from 1958 to 2023, in order to detect long-term trends and assess potential climate risks.

Q4. The following are the CO_2 concentrations (in parts per million) from the first 10 rows of the dataset.

315.7	316.4	316.5	315.9	314.9
313.2	313.3	314.7	315.6	320.5

Table 2: CO_2 concentrations (in parts per million) from the first 10 rows of the dataset.

(a) (2 marks) Draw a stemplot for the data from Table 2.

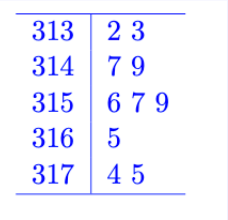
Solution:

313	2 3
314	7 9
315	6 7 9
316	4 5
317	
318	
319	
320	5

Key: 315 | 7 = 315.7 ppm

Comment to Graders:

If you see a student give the following stemplot:



313	2 3
314	7 9
315	6 7 9
316	5
317	4 5

If a student gives this stemplot, please deduct marks appropriately (e.g., 1 or 2 marks off for that error) and notify me via Piazza.

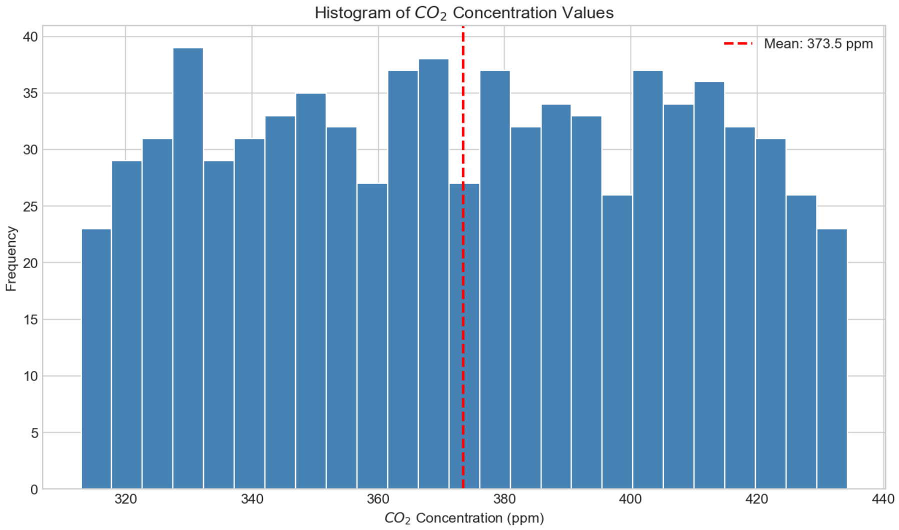
(b) (2 marks) Suppose the average of the above datapoints is 315.67.

Based on the stemplot you produced, are there any outliers? What is the shape of the distribution?

Solution: Based on the stemplot, the value 320.5 appears to be a potential outlier as it is noticeably separated from the main cluster of data (there are gaps at stems 317, 318, and 319). The distribution is right-skewed due to this high value pulling the right tail.

Alternatively, one could describe the main body of data (313–316) as approximately symmetric (after removing the outlier, in line with Marieke's approach).

Q5. The following histogram depicts monthly average CO_2 concentrations from 1958 to 2023 from the Mauna Loa Observatory.



(a) (1 mark) Describe the shape of the distribution.

Solution: This distribution appears to be very slightly right-skewed as the right tail is longer than the left tail. One can also argue that it is approximately symmetric.

(b) (2 marks) In discussions with climate scientists, you learn that elevated CO_2 levels are a primary driver of global warming. A widely recognized benchmark is 450 ppm (parts per million), since some climate models associate this concentration with approximately +2°C of global temperature increase.

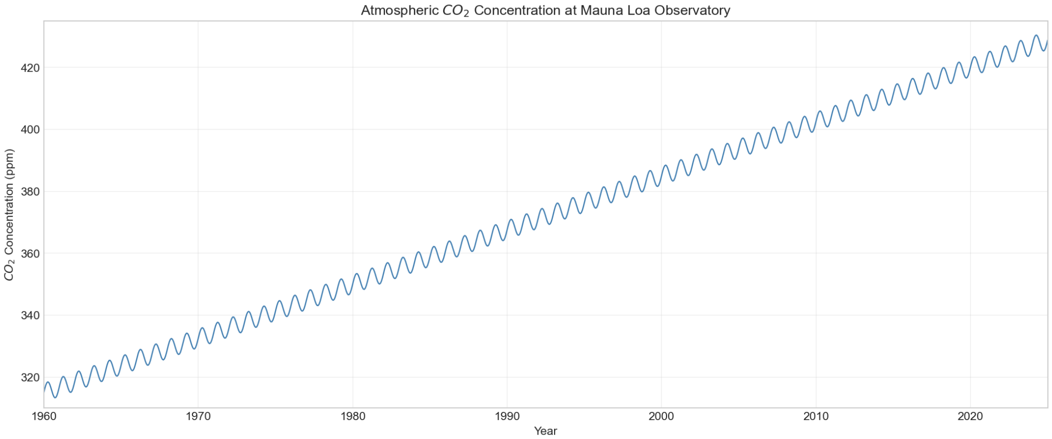
Based on the histogram and the mean CO_2 concentration, should we be concerned about the 450 ppm threshold? Explain your reasoning.

Solution: Based solely on the histogram, it does not appear to be a concern since 450 ppm does not even appear in the histogram: all recorded values fall below approximately 430 ppm. The mean of 365 ppm is well below the 450 ppm threshold. However, this assessment may be misleading because the histogram does not show the temporal trend in the data.

Comment to Graders:

Accept any answer that demonstrates an understanding of both the histogram and the real-world context, even if the explanation differs in wording or emphasis.

Q6. The following time series plot is based on monthly average CO_2 concentrations from 1958 to 2023 from the Mauna Loa Observatory.



(a) (3 marks) Describe the features you observe in the time series plot.

Solution: The time series shows a strong upward trend in CO_2 concentrations over time, rising from about 315 ppm in 1960 to over 420 ppm in recent years. There is also a clear seasonal pattern: CO_2 levels decrease each summer and increase each winter, reflecting the annual cycle of plant growth and decay in the Northern Hemisphere.

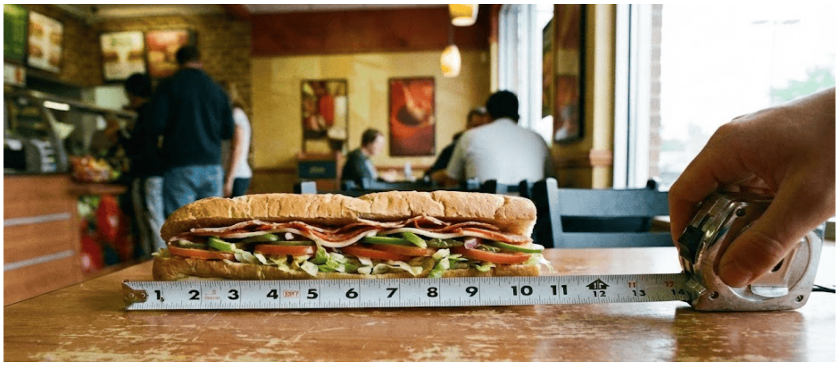
(b) (3 marks) Based on the time series plot, should we be concerned about the 450 ppm threshold? Does your answer differ from your assessment based on the histogram from the previous question?

Solution: Yes, the time series plot shows a clear upward trend toward 450 ppm, so we should be concerned. Unlike the histogram, which hides the increase over time, the time series reveals that CO_2 levels are rising and may reach the threshold soon. This highlights the importance of using time series plots to assess trends.

Data Science in Consumer Protection [18 marks]

Let's get this bread

In 2013, Subway was the subject of a class action lawsuit claiming that its footlong sandwiches were often less than 12 inches long. As part of the investigation, you have been asked to collect data from multiple Subway locations to analyze whether there is a relationship between sandwich length and profit margin, hence determining if there is a financial incentive for making sandwiches shorter.



Q7. As part of your investigation, you collected data from 8 randomly selected Subway franchise locations. For each location, you recorded:

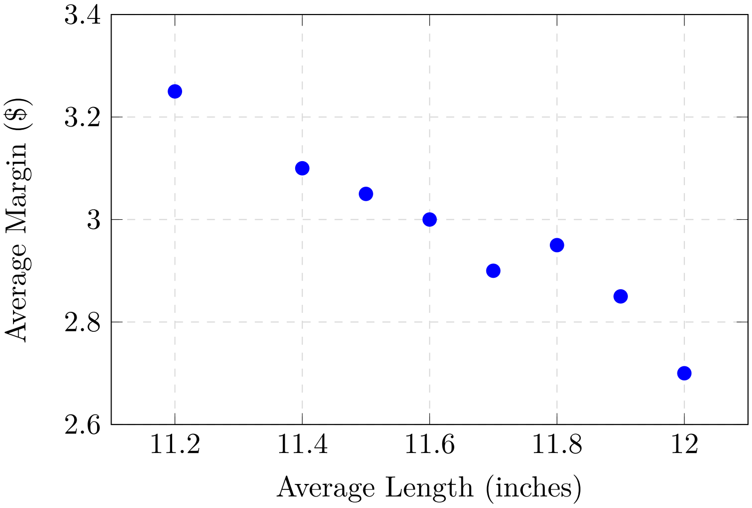
- `avg_length`: The average length of footlong sandwiches (in inches)
- `avg_margin`: The average profit margin per sandwich (in dollars)

The data are shown in the table below:

avg_length	avg_margin	avg_length	avg_margin
11.4	3.10	11.6	3.00
11.8	2.95	11.9	2.85
11.2	3.25	11.5	3.05
12.0	2.70	11.7	2.90

(a) (3 marks) Plot the data on a scatterplot with average sandwich length on the x-axis and average profit margin on the y-axis. Make sure to label your axes appropriately.

Solution:



- x-axis: Average Length (inches), ranging from approximately 11.0 to 12.2
- y-axis: Average Margin ($), ranging from approximately 2.4 to 3.5
- 8 data points showing a downward trend

(b) (3 marks) Based on your scatterplot, describe the direction, form, and strength of the relationship between average sandwich length and average profit margin.

Solution: Direction: Negative (as average length increases, average margin decreases)

Form: Approximately linear

Strength: Strong (the points closely follow a linear pattern with minimal scatter)

(c) (7 marks) Calculate the correlation coefficient between average sandwich length and average profit margin. *No justification necessary but incorrect answers without justification (where appropriate) will receive no partial marks.*

Solution: First, calculate the means and standard deviations for both variables ($n = 8$):

$$\bar{x} = \frac{93.1}{8} = 11.6375 \quad \text{and} \quad \bar{y} = \frac{23.8}{8} = 2.975$$

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{0.49875}{7}} \approx 0.2669$$

$$s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{0.195}{7}} \approx 0.1669$$

Let $z_x = \frac{x-\bar{x}}{s_x}$ and $z_y = \frac{y-\bar{y}}{s_y}$, then

x	y	z_x	z_y	$z_x z_y$
11.4	3.10	-0.89	0.75	-0.67
11.8	2.95	0.61	-0.15	-0.09
11.2	3.25	-1.64	1.65	-2.71
12.0	2.70	1.36	-1.65	-2.24
11.6	3.00	-0.14	0.15	-0.02
11.9	2.85	0.98	-0.75	-0.74
11.5	3.05	-0.52	0.45	-0.23
11.7	2.90	0.23	-0.45	-0.10
Sum				$\sum z_x z_y \approx -6.80$

$$r = \frac{1}{n-1} \sum z_x z_y = \frac{1}{7} - 6.80 \approx -0.97$$

(d) (2 marks) What does the correlation coefficient tell you about the relationship between average sandwich length and average profit margin? Would you describe this as a strong, moderate, or weak correlation?

Solution: The correlation coefficient of approximately $r \approx -0.97$ indicates a very strong negative linear relationship. This means that locations with longer average sandwich lengths tend to have lower profit margins, and this relationship is very consistent. This makes business sense: longer sandwiches require more ingredients, which reduces profit margins.

(e) (3 marks) Suppose you converted all profit margins from dollars to cents (1 dollar = 100 cents). What would be the new correlation coefficient? Be sure to justify.

Solution: The correlation coefficient would remain unchanged at $r \approx -0.97$. Correlation is unitless and invariant under linear transformations. Multiplying all profit margins by 100 (converting to cents) is a linear transformation that does not affect the strength or direction of the linear relationship.

Part 2 – Python

Movies

[11 marks]

Q8. The Internet Movie Database (IMDb) is a popular online database for information on the movie and television industry. As a data analyst for a film production company, you are examining box office performance patterns of horror movie directors. The data set `IMDB.csv` contains data scraped from IMDb in 2021, including the gross box office revenues (in US dollars), of movies (released between 1980–2016) directed by John Carpenter, David Cronenberg, George A. Romero and Wes Craven.

(a) (1 mark) Using Python, print the first six rows of the dataset.

Solution:

Python Code

```python
import pandas as pd

# Load dataset
df = pd.read_csv("data/IMDB.csv")

# Print first six rows
print(df.head(6).to_string())
```

```
                   name   genre  year  score          director      budget  \
0               The Fog  Horror  1980    6.8    John Carpenter   1000000.0
1  Escape from New York  Action  1981    7.2    John Carpenter   6000000.0
2              Scanners  Horror  1981    6.8  David Cronenberg         NaN
3       Deadly Blessing  Horror  1981    5.6        Wes Craven   2500000.0
4             The Thing  Horror  1982    8.1    John Carpenter  15000000.0
5             Creepshow  Comedy  1982    6.9  George A. Romero   8000000.0

        gross
0  21448782.0
1  25244626.0
2  14225876.0
3   8279042.0
4  19632053.0
5  21028755.0
```

(b) (3 marks) Using Python, print the minimum, maximum, mean, median, and standard deviation of the gross box office revenues.

Solution:

Python Code

```python
print("Summary Statistics for Gross Box Office Revenues:")
print(f"Minimum: ${df['gross'].min():,.2f}")
print(f"Maximum: ${df['gross'].max():,.2f}")
print(f"Mean: ${df['gross'].mean():,.2f}")
print(f"Median: ${df['gross'].median():,.2f}")
print(f"Std Dev: ${df['gross'].std():,.2f}")
```

```
Summary Statistics for Gross Box Office Revenues:
Minimum: $386,078.00
Maximum: $173,046,663.00
Mean:     $29,600,208.33
```

```
Median:  $18,074,865.00
Std Dev: $40,553,011.88
```

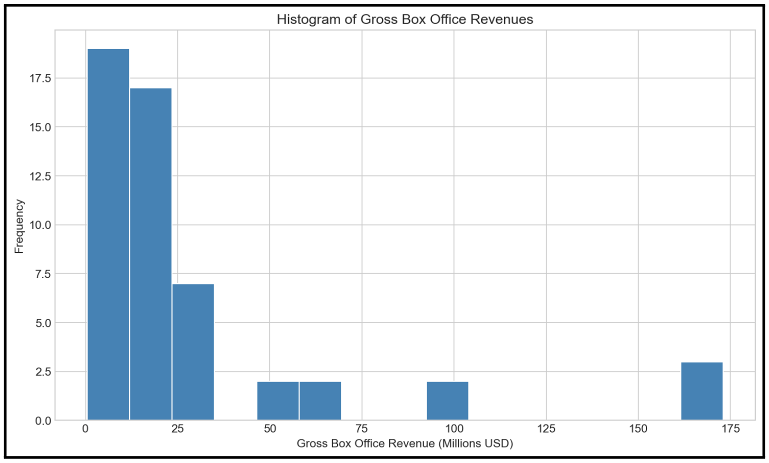
(c) (4 marks) Using Python, create a histogram of the gross box office revenues (found in the `gross` column of the dataset). Set the number of bins to 15 in the plot.

Solution:

Python Code

```python
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.histplot(df['gross'], bins=15)
plt.xlabel('Gross Box Office Revenue (USD)')
plt.ylabel('Frequency')
plt.title('Histogram of Gross Box Office Revenues')
plt.show()
```



(d) (3 marks) Using Python, print the number of movies that had a gross box office of over $100,000,000. Write down the names of these movies.

Solution:

Python Code

```python
high_gross_movies = df[df['gross'] > 100_000_000]
num_high_gross = len(high_gross_movies)
print(f"Number of movies with gross > $100,000,000: {num_high_gross}")
print("\nNames of movies:")
print(high_gross_movies['name'].to_string(index=False))
```

```
Number of movies with gross > $100,000,000: 3

Names of movies:
  Scream
Scream 2
Scream 3
```

Comment to Graders:

Please accept student submissions where they write out the relevant movies by hand. However, they need to present how they found this.

If students just write it down by hand without showing how they obtained it through code, please take off very minor marks for not following the instructions (at most 1 mark). If they indicate that they obtained it by inspecting the dataset, leave a comment indicating that this is not feasible for large datasets.

Rock, Paper, Statistics [8 marks]

Q9. In an in-class survey, students were asked to choose a move in a hypothetical game of Rock, Paper, Scissors. The table below summarizes the percentage of respondents who selected each move:

Move	% of respondents
Paper	16
Rock	37
Scissors	47

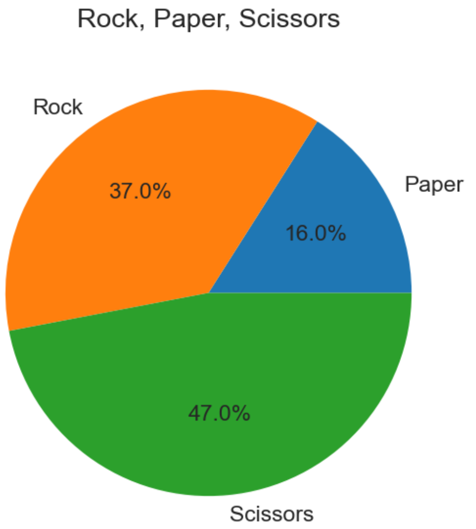
(a) (3 marks) Using Python, create a pie chart to visualize the percentage of respondents who selected each move.

Solution:

Python Code

```python
import matplotlib.pyplot as plt
labels = ['Paper', 'Rock', 'Scissors']
percentages = [16, 37, 47]

plt.figure(figsize=(6, 4))
plt.pie(percentages, labels=labels,
  autopct='%1.1f%%')
plt.title('Rock, Paper, Scissors')
plt.show()
```



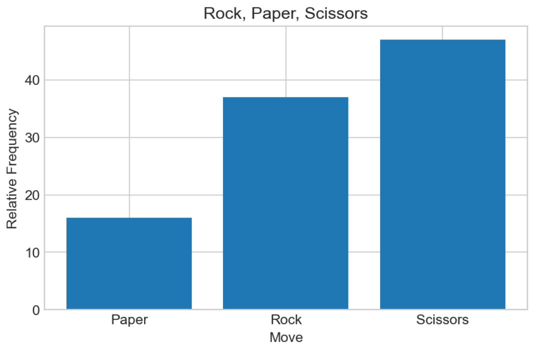
(b) (3 marks) Using Python, create a bar chart that shows the percentage of respondents who chose each move.

Solution:

Python Code

```python
import matplotlib.pyplot as plt
labels = ['Paper', 'Rock', 'Scissors']
percentages = [16, 37, 47]

plt.figure(figsize=(6, 4))
plt.bar(labels, percentages)
plt.xlabel('Move')
plt.ylabel('Relative Frequency')
plt.title('Rock, Paper, Scissors')
plt.show()
```



(c) (2 marks) Is a stemplot an appropriate visualization for this data? Explain your reasoning.

Solution: A stemplot is not an appropriate visualization for this data. A stemplot is designed for quantitative (numerical) variables where you can meaningfully split digits into stems and leaves.

The variable in this case (Move: Paper/Rock/Scissors) is categorical: it represents discrete categories, not numerical measurements. For categorical data, bar charts or pie charts are appropriate visualizations.

Air Quality in Canadian Cities [12 marks]

Q10. Environment Canada monitors air quality across the country using the Air Quality Health Index. As an environmental data scientist, you are comparing air quality patterns in three major Canadian cities.

The dataset `airquality.csv` contains monthly Air Quality Index (AQI) readings for Toronto, Vancouver, and Calgary from 2018–2023. The AQI scale ranges from 0 (excellent) to 500 (hazardous), with values above 100 considered "unhealthy for sensitive groups."

(a) (4 marks) Create side-by-side boxplots comparing the distribution of AQI values across the three cities.

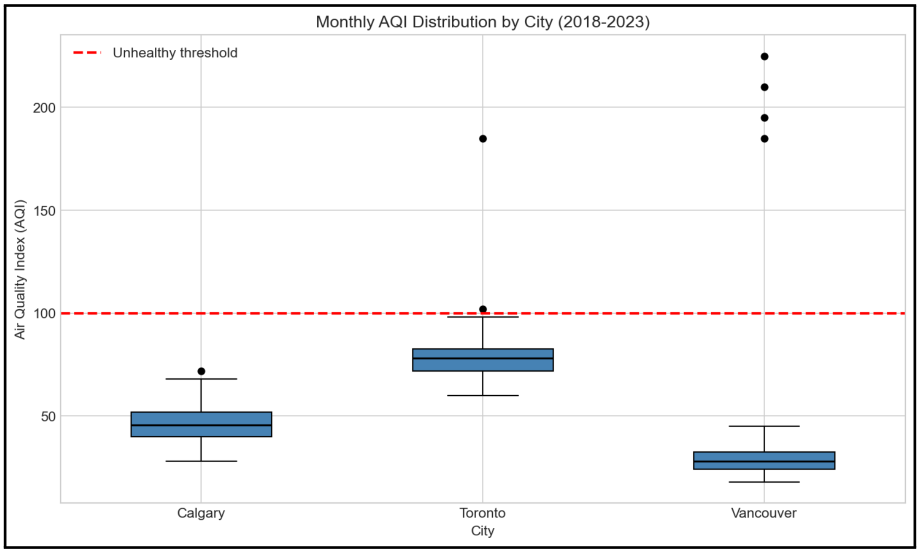
Solution:

Python Code

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

df = pd.read_csv("data/airquality.csv")

plt.figure(figsize=(10, 6))
sns.boxplot(x='city', y='aqi', data=df)
plt.xlabel('City')
plt.ylabel('Air Quality Index (AQI)')
plt.title('Monthly AQI Distribution by City (2018-2023)')
plt.axhline(y=100, color='red',
  linestyle='--', label='Unhealthy threshold')
plt.legend()
plt.show()
```



(b) (3 marks) Provide numerical summaries (count, mean, median, standard deviation, min, max) for the AQI distribution in each city.

Hint: You may use the following code snippet would obtain the AQI data entries for Toronto.

Python Code

```
toronto_aqi = df[df['city'] == 'Toronto']['aqi']
```

Solution:

Comment to Graders:

Please note that students do not have to produce the std, 1st or 3rd quartiles in their output. Please notify me if anyone does or their solution looks identical to the following one (but do not Deduct marks if they included all of the relevant numerical summaries.)

While filtering by city and calculating statistics individually is possible, a more efficient approach is to use the `groupby` method in pandas to compute summary statistics for all cities at once

Python Code

```python
summary = df.groupby('city')['aqi'].describe()
print("Summary Statistics by City:")
print(summary.to_string())

print("\nMedian AQI by City:")
print(df.groupby('city')['aqi'].median())
```

```
Summary Statistics by City:
           count       mean        std   min   25%   50%    75%    max
city
Calgary     72.0  46.444444   9.670470  28.0  40.0  45.5  52.00   72.0
Toronto     72.0  79.583333  15.383135  60.0  72.0  78.0  82.75  185.0
Vancouver   72.0  38.166667  41.029018  18.0  24.0  28.0  32.50  225.0

Median AQI by City:
city
Calgary      45.5
Toronto      78.0
Vancouver    28.0
Name: aqi, dtype: float64
```

(c) (2 marks) Do the boxplots indicate the presence of outliers in any of the three distributions? If so, identify the cities and roughly what the corresponding AQI values are.

Solution: Yes, the boxplots indicate outliers in two cities:

- Vancouver shows extreme outliers with AQI values reaching approximately 185–225.
- Toronto also has one major outlier around 185.

Calgary shows one mild outlier around 72 but nothing extreme.

These extreme AQI values correspond to wildfire smoke events: Vancouver's outliers occur during BC wildfire seasons (summers of 2018, 2020, 2021, 2023) and Toronto's outlier occurred in June 2023 when Quebec wildfire smoke drifted south.

(d) (3 marks) Compare the distributions of AQI for Toronto and Vancouver. Discuss differences in both center and spread.

Solution: Center (Location):

- Toronto has a much higher typical AQI: median of 78 vs. Vancouver's median of 28
- Toronto's air quality is consistently worse on a day-to-day basis

Spread (Variability):

- Vancouver has much higher variability: standard deviation of 41 vs. Toronto's 15

- Vancouver's IQR is small (most readings 24–32), but extreme outliers stretch the distribution
- Toronto's distribution is more tightly clustered around its median

In summary: Toronto has worse typical air quality but more consistent readings, while Vancouver has better typical air quality but more extreme fluctuations due to wildfire smoke events.

Comment to Graders:

Students do not need to comment on the standard deviation. However, if they only comment on standard deviation, please apply a minor penalty (0.5 marks) and remind them that the relevant measure of spread used for distributions having outliers is the IQR.

Properties of Variance and Standard Deviation [12 marks]

I ran out of puns

In this section, you will use Python to investigate three fundamental properties of spread: non-negativity, translation invariance, and scaling sensitivity.

Q11. Variance and Standard Deviation measure spread. Since distance cannot be negative, these measures must be non-negative.

(a) (3 marks) Using Python and `numpy`, calculate and print the **variance** for the following three lists.

- `data_spread = [1, 2, 3, 4, 5]`
- `data_identical = [5, 5, 5, 5, 5, 5, 5]`
- `data_negative = [-1, -2, -3, -4, -5]`

Ensure you use `ddof=1` for your calculations if using `numpy`.

Solution:

Python Code

```python
import numpy as np

data_spread = [1, 2, 3, 4, 5]
data_identical = [5, 5, 5, 5, 5,5, 5]
data_negative = [-1,-2, -3,-4, -5]

print(f"Variance (spread):    {np.var(data_spread, ddof=1)}")
print(f"Variance (identical): {np.var(data_identical, ddof=1)}")
print(f"Variance (negative):  {np.var(data_negative, ddof=1)}")
```

```
Variance (spread):    2.5
Variance (identical): 0.0
Variance (negative):  2.5
```

(b) (1 mark) Based on your output, what are the variance and standard deviation when all values in a dataset are identical? Why does this make sense?

Solution: The variance and standard deviation are both 0.0. This is because these measures quantify the spread of data points from the mean; if all data points are identical, the distance of every point from the mean is zero.

Q12. Translation Invariance: If you add a constant to every value in a dataset, the location shifts, but the distance between points should remain constant.

(a) (3 marks) Using Python, perform the following:

i. Create a `numpy` array with values `[10, 20, 30]`.

ii. Calculate the **variance** and **standard deviation** (using `ddof=1`).

iii. Create a new array by adding `1000` to the original values.

iv. Calculate the **variance** and **standard deviation** of this new array.

Print all four values.

Solution:

Python Code

```python
import numpy as np

data = np.array([10, 20, 30])
var_original = np.var(data, ddof=1)
sd_original = np.std(data, ddof=1)

# Shift data by adding 1000
data_shifted = data + 1000
var_shifted = np.var(data_shifted, ddof=1)
sd_shifted = np.std(data_shifted, ddof=1)

print(f"Original Var: {var_original}, Original SD: {sd_original}")
print(f"Shifted Var:  {var_shifted}, Shifted SD:  {sd_shifted}")
```

```
Original Var: 100.0, Original SD: 10.0
Shifted Var:  100.0, Shifted SD:  10.0
```

(b) (1 mark) Does adding a constant to the data change the variance or standard deviation? Explain why this makes sense.

Solution: No, neither the variance nor the standard deviation changes. Both measures quantify the spread (distance between points). Shifting all points by the same amount preserves the relative distances between them, so the spread remains identical.

Q13. Scaling Sensitivity: Unlike addition, multiplication affects the scale of the data.

(a) (3 marks) Using Python, perform the following:

i. Create a `numpy` array with values `[2, 4, 6]`.

ii. Calculate the **variance** AND **standard deviation** (`ddof=1`).

iii. Create a new array by multiplying the original values by `10`.

iv. Calculate the **variance** AND **standard deviation** of this new array.

Print all four values.

Solution:

Python Code

```python
data = np.array([2, 4, 6])
var_original = np.var(data, ddof=1)
sd_original = np.std(data, ddof=1)

# Scale data by multiplying by 10
data_scaled = data * 10
var_scaled = np.var(data_scaled, ddof=1)
sd_scaled = np.std(data_scaled, ddof=1)

print(f"Original Var: {var_original}, Original SD: {sd_original}")
print(f"Scaled Var:   {var_scaled}, Scaled SD:   {sd_scaled}")
```

```
Original Var: 4.0, Original SD: 2.0
Scaled Var:   400.0, Scaled SD:   20.0
```

(b) (1 mark) How does multiplying the data by 10 affect the variance and standard deviation differently?

Solution: The standard deviation is multiplied by 10 (scaling linearly). However, the variance is multiplied by $10^2 = 100$. This is because variance is calculated using squared distances, so the scaling factor is squared.