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:

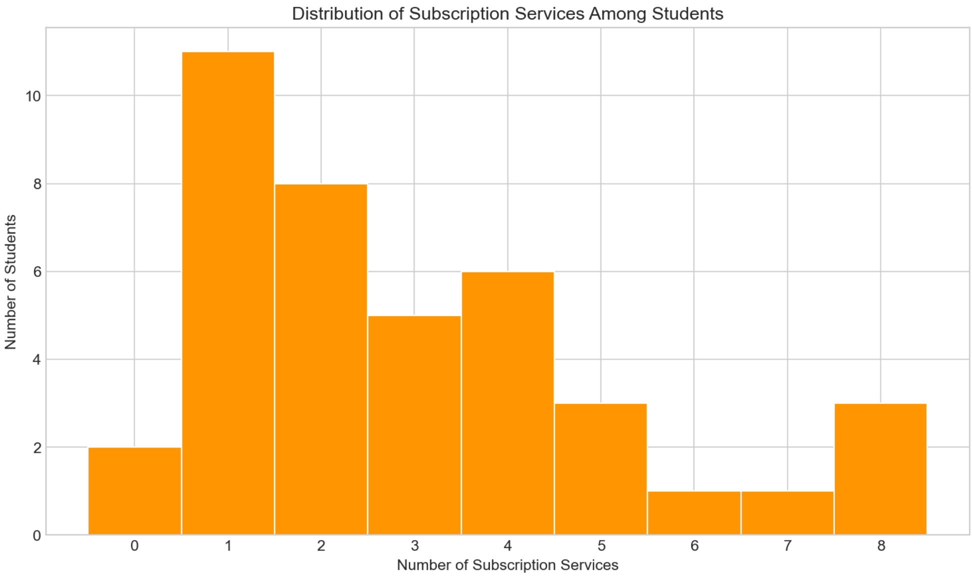
student_id	gender	num_subscriptions	monthly_cost_usd	monthly_income_usd	monthly_discretionary_usd
582334538	Female	5	36.51	900	180
398704996	Female	5	38.00	0	0
219540831	Male	0	0.00	850	160
553035110	Male	2	22.42	0	0
890779946	Non-binary	1	18.11	920	200

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

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

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

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?

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

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

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

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.*

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

(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".

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

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

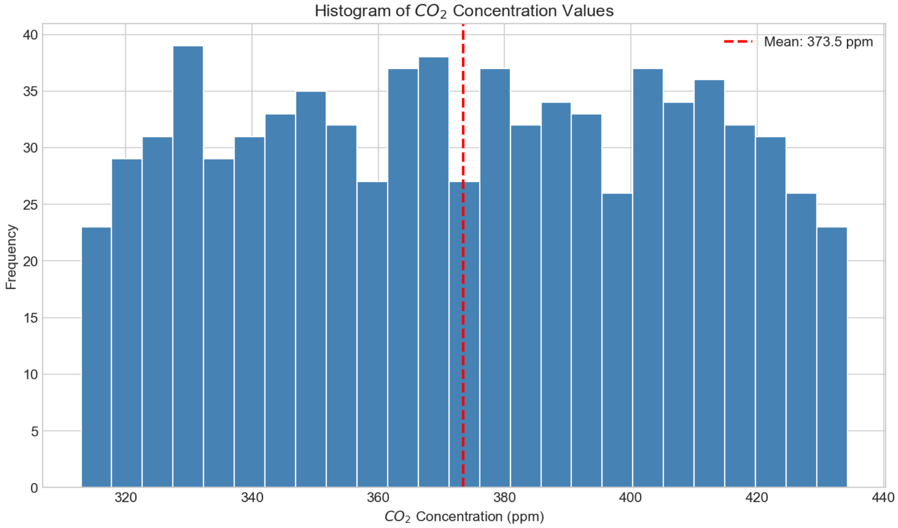
Table 2: CO_2 concentrations (in parts per million) from the first 10 rows of the dataset.

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

(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?

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

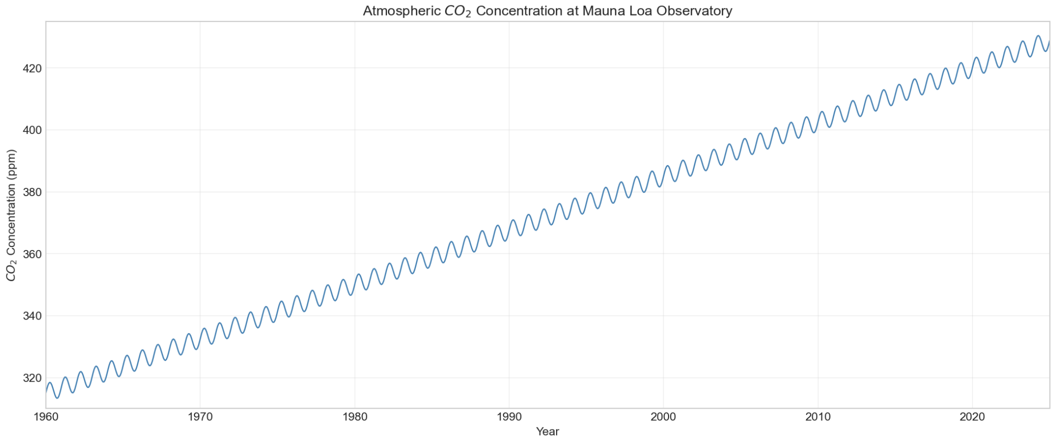


(a) (1 mark) Describe the shape of the distribution.

(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.

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



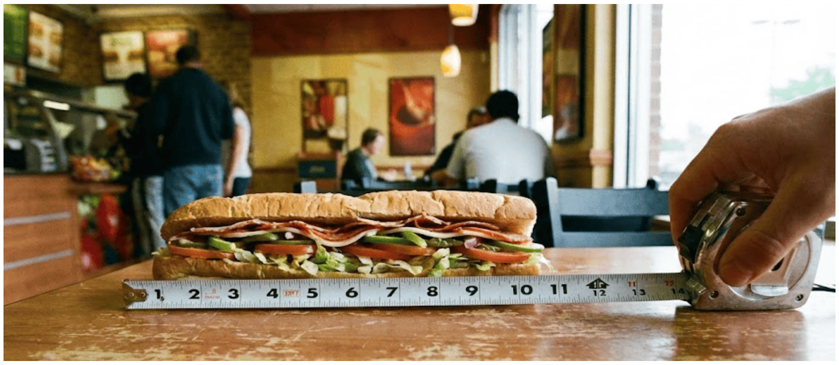
(a) (3 marks) Describe the features you observe in the time series plot.

(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?

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.

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

(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.*

(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?

(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.

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.

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

(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.

(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.

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.

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

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

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.

(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']
```

(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.

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

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`.

(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?

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.

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

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.

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