DS 1000B – Assignment 3

Total Marks: 100

Due date: Friday, February 13th, 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.

Part 1 – Written Responses

Textbook Price Comparison

[20 marks]

Price is Write

A student group is investigating textbook affordability. They want to predict Amazon prices based on the campus store (UWO) prices to determine if buying online is consistently cheaper. They collected data on 12 randomly selected textbooks.

Q1. The following table lists the prices (in CAD) for the selected books at both retailers.

Course	**Book Title**	$x =$ **UWO Price**	$y =$ **Amazon Price**
CAL2503B	Multivariable Calculus	97.00	95.22
MAT1229B	Elementary Linear Algebra	75.00	39.90
MAT3152B	A Walk Through Combinatorics	190.70	39.97
STA2503B	Differential Equations	68.00	126.76
STA2864B	Statistical Programming with R	52.00	53.95
STA4861B	Time Series and Forecasting	130.00	137.40
ASC2427B	Actuarial Mathematics	57.00	62.95
FMO3613B	Quantitative Finance	181.55	27.97
NMM1411B	Elementary Linear Algebra	62.00	65.99
NMM2270B	Advanced Engineering Mathematics	221.40	207.34
GEO2162B	Planning Canada	142.80	135.00
PHI2020	An Introduction to Logic	64.25	95.22

Summary Statistics:

Statistic	Value
Mean UWO Price ($\bar{x}$)	\$111.81
Standard Deviation UWO (s_x)	\$61.26
Mean Amazon Price ($\bar{y}$)	\$90.64
Standard Deviation Amazon (s_y)	\$53.23
Correlation Coefficient (r)	0.323

(a) (7 marks) Calculate the least squares regression line equation for predicting Amazon prices from UWO prices.

(b) (5 marks) Interpret the slope and y-intercept of your regression line. What do these values imply about the relationship between the two stores?

(c) (3 marks) A student needs to buy *Differential Equations* (UWO Price: \$68.00). Using your model, determine if they should buy it from Amazon or UWO. How does this compare to the actual Amazon price?

(d) (5 marks) Based on your analysis, is Amazon consistently the "preferable" (cheaper) option? Explain why or why not.

Coffee Shop Analytics

[30 marks]

Latte Data

Q2. A coffee chain is interested in understanding the relationship between customer age group and preferred drink type. The following contingency table shows the results of a survey they conducted.

Age Group	Preferred Drink Type			
	Espresso	Latte	Cold Brew	Tea
18–30	30	70	55	25
31–50	50	45	35	30
51+	40	10	5	5

(a) (2 marks) How many customers were surveyed in total?

(b) (4 marks) Calculate the marginal distribution for preferred drink type.

(c) (8 marks) Calculate the conditional distribution of preferred drink type for each age group.

(d) (2 marks) Calculate the relative risk of preferring Cold Brew for customers aged 18–30 compared to customers aged 51+. Interpret your result.

(e) (4 marks) Based on your calculations, is there evidence of an association between age group and preferred drink type? Justify your answer. Note that you may use visualizations (either drawn or computer-generated) to support your response.

Q3. A researcher collected data from 8 coffee shops to study the relationship between average customer wait time (in minutes) and customer satisfaction score (out of 100).

wait_time	satisfaction	wait_time	satisfaction
2	92	6	68
4	84	8	55
3	88	5	75
7	60	10	42

The least squares regression line for predicting satisfaction from wait time is:

$$\hat{y} = 101.5 - 5.75x$$

(a) (2 marks) Interpret the slope of the regression line in the context of the problem.

(b) (2 marks) Interpret the y-intercept of this regression line. Is this interpretation meaningful in this context? Explain.

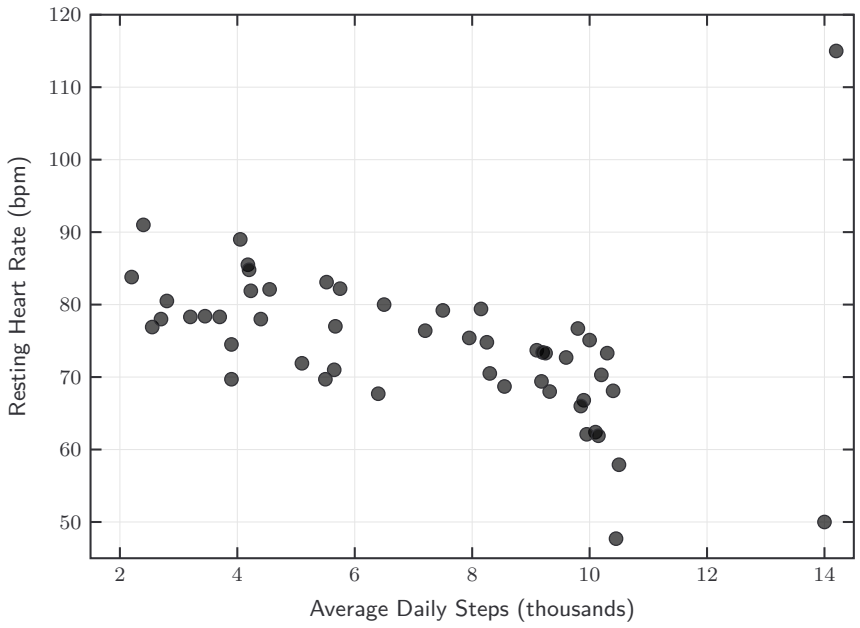
(c) (2 marks) Use the regression equation to predict the satisfaction score for a coffee shop with an average wait time of 5 minutes.

(d) (4 marks) The coffee shop with a wait time of 5 minutes has an observed satisfaction score of 75. Calculate the residual and explain what it indicates about this coffee shop's performance relative to the model's prediction.

Fitness Tracker Study [10 marks]

A health researcher is studying the relationship between daily step count and various health metrics using data from fitness trackers. The dataset contains information from 50 participants, including their average daily steps and average resting heart rate.

Q4. The following scatterplot shows the relationship between average daily steps (in thousands) and average resting heart rate (beats per minute) for the 50 participants.



(a) (3 marks) Are there any observations that appear influential? If so, list them.

(b) (3 marks) Based on the scatterplot, describe the relationship between daily steps and resting heart rate in terms of direction, form, and strength.

(c) (4 marks) The correlation coefficient for this data is $r = -0.64$. Calculate the coefficient of determination (R^2) and interpret it in context.

Part 2 – Python

Job Satisfaction Analysis [17 marks]

Q5. The file `job_satisfaction.csv` includes a variable indicating the employee's industry sector (Non-profit, Education, or Corporate). We will explore how the relationship between salary and job satisfaction changes when we account for industry.

(a) (4 marks) Fit a linear regression model predicting job satisfaction from salary, ignoring industry for now. Print the slope and intercept of the regression line.

(b) (8 marks) Fit a separate linear regression model for each industry (Non-profit, Education, and Corporate). Print the slope for each industry.

(c) (5 marks) Compare the overall regression from part (a) with the industry-specific regressions from part (b). What do you notice, and what might explain these differences?

LEGO Set Analysis [23 marks]

The `lego_sample.csv` contains data on 75 LEGO sets, including the number of pieces, Amazon price, and theme. In this analysis, we will explore the relationship between the number of pieces and price, focusing on two popular themes: City and Friends.

Q6. First, we'll explore if there is a relationship between the Amazon price and the number of pieces per set.

(a) (3 marks) Load the dataset and create a filtered DataFrame containing only "City" and "Friends" themes. Then, create a scatterplot of the number of pieces (x) versus Amazon price (y) for these sets. *Hint: The pandas `.isin()` method may be useful for filtering.*

(b) (3 marks) Fit a simple linear regression model predicting price from the number of pieces for this subset. Report the regression equation and the R^2 value. You do not need to plot the regression line.

(c) (2 marks) Interpret the slope in the context of this study.

Q7. Online discussions suggest that "City" bricks are among the most expensive per piece by theme. We'll compare City to Friends sets to investigate:

Does the relationship between price and number of pieces differ between LEGO City and LEGO Friends sets?

(a) (4 marks) Create a scatterplot of pieces (x) vs. price (y) for the filtered data. Use different colours for each theme (e.g., blue for City, red for Friends) and include a legend.

(b) (5 marks) Fit separate regression lines for the City sets and the Friends sets. Print the slope and intercept for each.

(c) (3 marks) Compare the slopes and intercepts from part (b). Do the results suggest that the cost per piece is similar between themes, but one theme has a higher base price? Explain.

(d) (3 marks) Based on your results in part (c), does one theme have a consistent price premium over the other, holding the number of pieces fixed? If so, what is the estimated value of this premium according to the models?