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.

Solution: First, calculate the slope (b):

$$b = r\left(\frac{s_y}{s_x}\right) = 0.323\left(\frac{53.23}{61.26}\right) = 0.323(0.8689) \approx 0.2807$$

Next, calculate the y-intercept (a):

$$a = \bar{y} - b\bar{x} = 90.64 - (0.2807)(111.81) = 90.64 - 31.38 = 59.26$$

Rounding to two decimal places, the regression equation is:

$$\boxed{\hat{y} = 59.26 + 0.28x}$$

(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?

Solution: Slope: For every \$1 increase in the UWO price, the Amazon price is predicted to increase by only \$0.28.

Intercept: A book that is "free" (\$0) at UWO would theoretically cost \$59.26 on Amazon. This does not have meaningful interpretation in this context.

The intercept does suggest, however, that for cheaper books (i.e. close to \$0), Amazon might actually be the more expensive option. The slope being less than 1, indicates that the more a textbook costs in UWO's bookstore, the more likely Amazon is to be cheaper in comparison.

(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?

Solution: Predicted Amazon Price:

$$\hat{y} = 59.26 + 0.28(68.00) = 59.26 + 19.04 = \$78.30$$

Since the predicted Amazon price (\$78.30) is higher than the UWO price (\$68.00), the model suggests staying with UWO.

Actual Data Comparison: The actual Amazon price is \$126.76. Residual $= y - \hat{y} = 126.76 - 78.30 = 48.46$. The model underestimated how expensive Amazon was. In this case, UWO was definitely the better choice.

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

Solution: No, Amazon is not consistently cheaper. The correlation is weak ($r = 0.323$), meaning the relationship is inconsistent. While the mean price is lower at Amazon (\$90.64 vs \$111.81), the high intercept and low slope indicate that Amazon is preferable for expensive textbooks, but UWO is often cheaper for lower-priced textbooks.

Any reasonable explanation based on the regression results and data trends is acceptable here.

Coffee Shop Analytics

Latte Data

[30 marks]

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?

Solution: The total number of customers surveyed is:

$$30 + 70 + 55 + 25 + 50 + 45 + 35 + 30 + 40 + 10 + 5 + 5 = \boxed{400}$$

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

Solution: First, calculate column totals:

- Espresso: $30 + 50 + 40 = 120$
- Latte: $70 + 45 + 10 = 125$
- Cold Brew: $55 + 35 + 5 = 95$
- Tea: $25 + 30 + 5 = 60$

Marginal distribution (proportions):

Drink Type	Proportion
Espresso	$120/400 = 0.30$
Latte	$125/400 = 0.31$
Cold Brew	$95/400 = 0.24$
Tea	$60/400 = 0.15$

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

Solution: Row totals: 18–30: 180, 31–50: 160, 51+: 60

Age Group	Espresso (%)	Latte (%)	Cold Brew (%)	Tea (%)
18–30	16.67	38.89	30.56	13.89
31–50	31.25	28.13	21.88	18.75
51+	66.67	16.67	8.33	8.33

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

Solution:

$$\text{RR} = \frac{P(\text{Cold Brew}|18\text{–}30)}{P(\text{Cold Brew}|51+)} = \frac{55/180}{5/60} = \frac{55 \times 60}{180 \times 5} = \frac{3300}{900} = 3.67$$

Interpretation: Customers aged 18–30 are 3.67 times as likely to prefer Cold Brew compared to customers aged 51+.

(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.

Solution: Yes, there is evidence to suggest an association between age group and preferred drink type.

The conditional distributions differ substantially across age groups:

- Younger customers (18–30) show highest preference for Latte (38.89%) and Cold Brew (30.56%)
- Middle-aged customers (31–50) have more balanced preferences across all drink types
- Older customers (51+) strongly prefer Espresso (66.67%) and rarely choose Cold Brew (8.33%)

If age and drink preference were independent, we would expect similar conditional distributions across all age groups.

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.

Solution: The slope of -5.75 means that for each additional minute of wait time, the predicted customer satisfaction score decreases by approximately 5.75 points. This indicates a negative relationship: longer wait times are associated with lower customer satisfaction.

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

Solution: The y-intercept of 101.5 represents the predicted satisfaction score when wait time is 0 minutes. While mathematically meaningful, a satisfaction score above 100 doesn't make practical sense since the maximum score is 100. This is an example of extrapolation beyond the observed data range, as the smallest observed wait time was 2 minutes.

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

Solution:

$$\hat{y} = 101.5 - 5.75(5) = 101.5 - 28.75 = \boxed{72.75}$$

(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.

Solution:

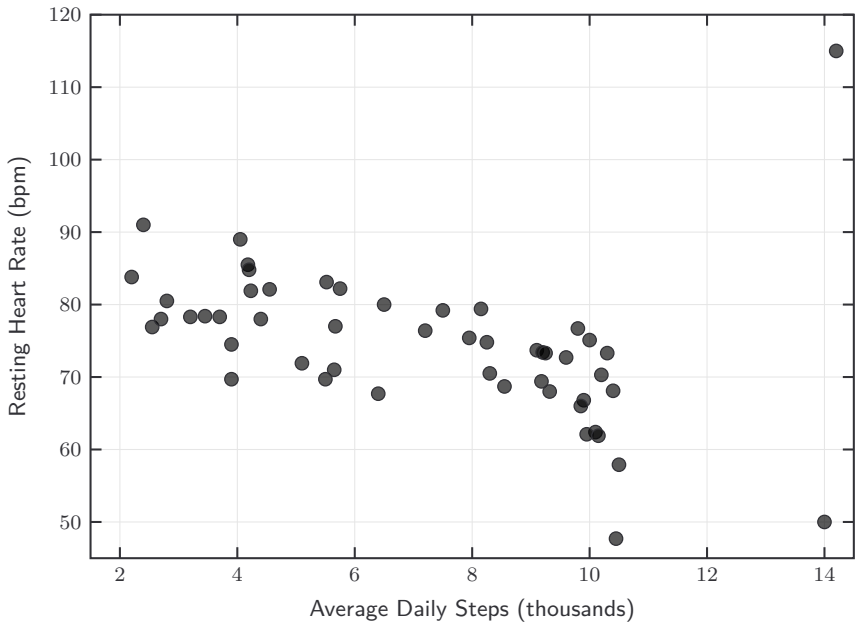
$$\text{Residual} = y - \hat{y} = 75 - 72.75 = \boxed{2.25}$$

The positive residual of 2.25 indicates that this coffee shop has a satisfaction score that is 2.25 points higher than predicted by the regression model. This shop is performing slightly better than expected given its wait time.

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.

Solution: Yes, the point with approximately 14.1k steps and resting heart rate of 110 is an influential observation. It is an outlier in terms of resting heart rate and has high leverage.

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

Solution: Direction: Negative (as daily steps increase, resting heart rate tends to decrease)

Form: Approximately linear

Strength: Moderate or strong (there is a clear downward trend, but considerable scatter around the pattern)

(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.

Solution:

$$R^2 = r^2 = (-0.64)^2 = 0.4096 \approx 0.41$$

Interpretation: Approximately 41% of the variation in resting heart rate can be explained by the linear relationship with daily step count.

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.

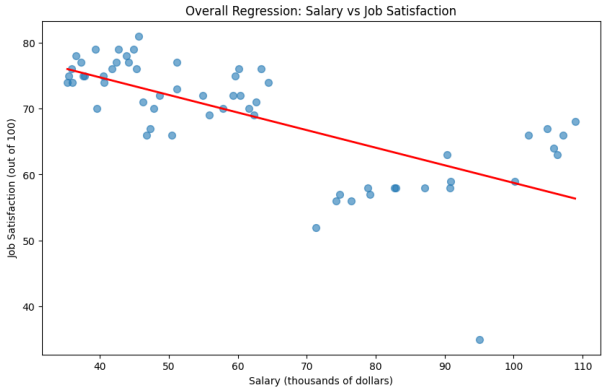
Solution:

Python Code

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import linregress
import numpy as np


# Load data
df = pd.read_csv(DATA_DIR / "job_satisfaction.csv")

# Fit the linear regression model
slope, intercept, r_value, p_value, std_err = linregress(
    df['salary_thousands'],
    df['satisfaction']
)

print(f"Slope: {slope:.4f}")
print(f"Intercept: {intercept:.4f}")

plt.figure(figsize=(10, 6))
sns.regplot(
    data=df,
    x='salary_thousands',
    y='satisfaction',
    scatter_kws={'alpha': 0.6, 's': 50},
    line_kws={'color': 'red', 'linewidth': 2},
    ci=None
)
plt.xlabel('Salary (thousands of dollars)')
plt.ylabel('Job Satisfaction (out of 100)')
plt.title('Overall Regression: Salary vs Job Satisfaction')
plt.savefig(FIGS_DIR / "q6a_regression.png", bbox_inches='tight')
```

Output

```
Slope:  -0.2668
Intercept:  85.4135
```



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

Solution:

Python Code

```
industries = df['industry'].unique()
colors = {'Non-profit': 'blue', 'Education': 'green', 'Corporate': 'red'}

plt.figure(figsize=(10, 6))

# Plot scatter points
for industry in industries:
    industry_data = df[df['industry'] == industry]
    plt.scatter(industry_data['salary_thousands'],
                industry_data['satisfaction'],
                alpha=0.6, s=50, color=colors[industry], label=industry)

# Fit and plot regression lines
for industry in industries:
    industry_data = df[df['industry'] == industry]

    slope_ind, intercept_ind, _, _, _ = linregress(
        industry_data['salary_thousands'], industry_data['satisfaction'])

    print(f"{industry} - Slope: {slope_ind:.4f}")
```

Output

```
Non-profit - Slope:  0.3361
Education - Slope:  0.2101
Corporate - Slope:  0.2707
```

(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?

Solution: The overall regression line from part (a) shows a negative slope (-0.27), indicating that as salary increases, job satisfaction tends to decrease. However, when we look at the industry-specific lines from part (b), we see that all three industries (Non-profit, Education, and Corporate) have positive slopes (approximately 0.21 to 0.34). This suggests that within each industry, higher salaries are associated with higher job satisfaction.

This is an example of Simpson's Paradox. When we look at all the data together, it looks like higher

salaries mean lower job satisfaction. But within each industry, higher salaries actually mean higher job satisfaction. The overall negative trend happens because the industries have different average salaries and satisfaction levels. When we ignore industry, this difference creates a misleading result.

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.

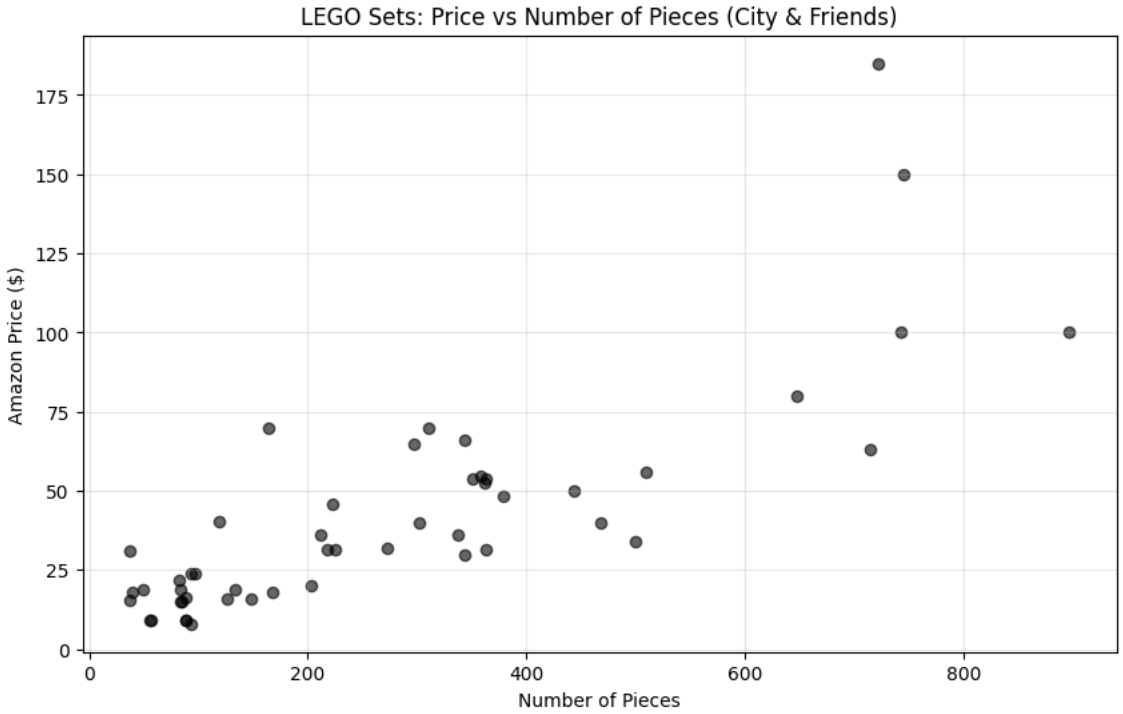
Q6. First, we'll explore if there is a relationship between the Amazon price and the number of pieces per set.

(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.*

Solution:

Python Code

```
df = pd.read_csv(DATA_DIR / "lego_sample.csv")

# Filter for City and Friends only
df_subset = df[df['theme'].isin(['City', 'Friends'])].copy()

# Plot
plt.figure(figsize=(10, 6))
plt.scatter(df_subset['pieces'], df_subset['amazon_price'],
            alpha=0.6, color='black')
plt.xlabel('Number of Pieces')
plt.ylabel('Amazon Price ($)')
plt.title('LEGO Sets: Price vs Number of Pieces (City & Friends)')
plt.grid(True, alpha=0.3)
plt.savefig(FIGS_DIR / "q7a_lego_scatter.png", bbox_inches='tight')
```



(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.

Solution:

Python Code

```
slope, intercept, r_value, p_value, std_err = linregress(
    df_subset['pieces'], df_subset['amazon_price']
)
print(f"Equation: Price = {intercept:.2f} + {slope:.4f} * Pieces")
print(f"R-squared: {r_value**2:.4f}")
```

Output

```
Equation:  Price = 6.42 + 0.1281 * Pieces
R-squared:  0.6616
```

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

Solution: The slope of 0.1281 means that for each additional piece in a LEGO set, the predicted Amazon price increases by approximately \$0.13 (about 13 cents per piece). This represents the average "cost per piece" for City and Friends sets combined.

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:

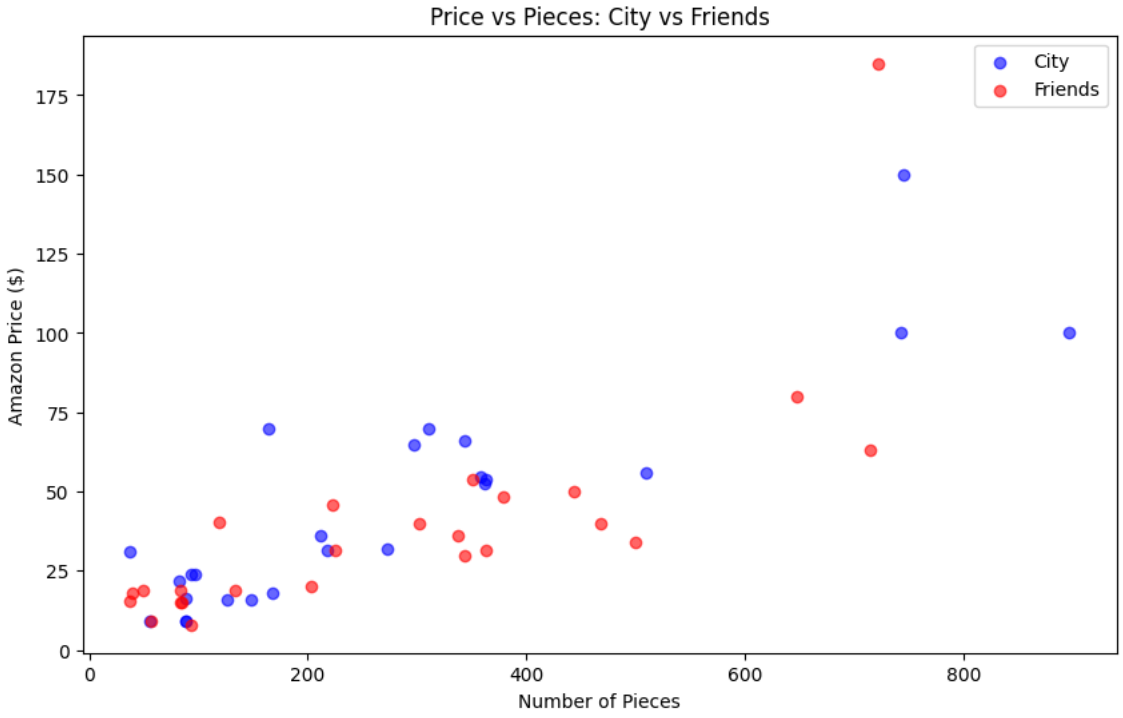
Does the relationship between price and number of pieces differ between LEGO City and LEGO Friends sets?

(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.

Solution:

Python Code

```
plt.figure(figsize=(10, 6))
colors = {'City': 'blue', 'Friends': 'red'}

for theme in ['City', 'Friends']:
    theme_data = df_subset[df_subset['theme'] == theme]
    plt.scatter(theme_data['pieces'], theme_data['amazon_price'],
                color=colors[theme], label=theme, alpha=0.6)

plt.xlabel('Number of Pieces')
plt.ylabel('Amazon Price ($)')
plt.title('Price vs Pieces: City vs Friends')
plt.legend()
plt.savefig(FIGS_DIR / "q7b_lego_comparison.png", bbox_inches='tight')
```



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

Solution:

Python Code

```
for theme in ['City', 'Friends']:
    theme_data = df_subset[df_subset['theme'] == theme]
    slope_t, intercept_t, _, _, _ = linregress(
        theme_data['pieces'], theme_data['amazon_price'])
    print(f"{theme}")
    print(f" Slope: {slope_t:.4f}, Intercept: {intercept_t:.4f}")
```

Output

```
City
Slope:  0.1305, Intercept:  9.4375
Friends
Slope:  0.1258, Intercept:  3.4029
```

(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.

Solution: Yes, the results support this interpretation.

- Slopes: The slopes are nearly identical (0.1305 for City vs 0.1258 for Friends), indicating that for both themes, each additional piece adds approximately 13 cents to the price.
- Intercepts: The intercepts differ significantly (9.44 for City vs 3.40 for Friends).

This suggests an "additive" relationship: LEGO City sets have a base price that is approximately \$6.00 higher than Friends sets ($9.44 - 3.40$), but the rate at which price increases with size is consistent across both themes.

(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?

Solution: Yes, since the slopes are roughly equal, the difference in price is consistent across different set sizes. The premium is determined by the difference in intercepts:

$$\text{Premium} = \text{Intercept}_{\text{City}} - \text{Intercept}_{\text{Friends}} = 9.44 - 3.40 = \$6.04$$

There is a flat premium of approximately \$6.00 for City sets compared to Friends sets of the same size.

Solution: Alternative solution (using `price` instead of `amazon_price`):

Note: The following results arise if the `'price'` column (List Price/MSRP) is used for the analysis instead of `'amazon_price'`.

Part (a): Overall Regression

Python Code

```
slope, intercept, r_value, p_value, std_err = linregress(
    df_subset['pieces'], df_subset['price']
)
print(f"Equation: Price = {intercept:.2f} + {slope:.4f} * Pieces")
print(f"R-squared: {r_value**2:.4f}")
```

Output

```
Equation:  Price = 1.97 + 0.1182 * Pieces
R-squared:  0.8046
```

Part (b): Comparison by Theme

Python Code

```
for theme in ['City', 'Friends']:
    theme_data = df_subset[df_subset['theme'] == theme]
    slope_t, intercept_t, _, _, _ = linregress(
        theme_data['pieces'], theme_data['price']
    )
    print(f"{theme} -> Slope: {slope_t:.4f}, Intercept: {intercept_t:.4f}")
```

Output

```
City -> Slope:  0.1362, Intercept:  2.4698
Friends -> Slope:  0.0977, Intercept:  2.3159
```

Interpretation Differences for Part (c): Unlike the Amazon price analysis (which showed similar slopes but different intercepts), the List Price analysis shows different slopes (0.1362 for City vs 0.0977 for Friends) and similar intercepts (≈ 2.40).

This suggests that based on List Price, LEGO City sets are more expensive because they have a higher *cost per piece*, not because of a fixed "base" premium.