Assignment-based Subjective Questions

1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

- Conducted an analysis on categorical columns using boxplots and bar plots.
- Inferences:
    - Fall season attracted more bookings.
    - Booking count increased significantly in each season from 2018 to 2019.
    - Peak bookings occurred during May, June, July, August, September, and October.
    - Clear weather positively influenced bookings.
    - Thursday to Sunday had more bookings compared to the start of the week.
    - Non-holidays and working days showed lower booking counts.
    - Significant growth in bookings observed in 2019 compared to the previous year.

---

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

- Using `drop_first=True` is crucial to:
    - Reduce extra columns created during dummy variable creation.
    - Mitigate correlations among dummy variables.
    - Ensure k-1 dummies are obtained by removing the first level, avoiding multicollinearity issues.

---

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- After examining pair-plots among numerical variables, identified **'temp'** as having the highest correlation with the target variable.

---

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- Assumptions Checked:
    - Normality of error terms: Ensured normal distribution of error terms.
    - Multicollinearity: Verified insignificant multicollinearity among variables.
    - Linear relationship: Confirmed the presence of linearity among variables.
    - Homoscedasticity: Checked for no visible pattern in residual values.
    - Independence of residuals: Ensured no auto-correlation.

5. Top 3 Features in Final Model:

- **Identified the top three features significantly contributing to explaining the demand for shared bikes:**
  - **Temperature (temp): A unit increase in temperature increases bike hire numbers by 3695.86 units.**
  - **Year (yr): A unit increase in the year variable increases bike hire numbers by 2030.29 units.**
  - **Windspeed: A unit increase in windspeed decreases bike hire numbers by 1041.56 units.**

General Subjective Questions

1. Explain the Linear Regression Algorithm in Detail. (4 marks)

Answer:

Linear regression analyzes the linear relationship between a dependent variable and a set of independent variables.

It is a statistical method used in data science and machine learning for predictive analysis.

The mathematical representation is given by the equation (Y = mX + c), where:

- (Y) is the dependent variable.
- (X) is the independent variable.
- (m) is the slope representing the effect of (X) on (Y).
- (c) is the Y-intercept (value of (Y) when (X = 0)).

Types of Linear Regression:

- **Simple Linear Regression:** Involves one independent variable.
- **Multiple Linear Regression**: Involves multiple independent variables.

Assumptions:

1. **Multi-collinearity:** Assumes little or no dependency among independent variables.
2. **Auto-correlation:** Assumes minimal auto-correlation in the data.
3. **Linear Relationship:** Assumes a linear relationship between response and feature variables.
4. **Normality of Error Terms:** Assumes normal distribution of error terms.
5. **Homoscedasticity:** Assumes constant variance in residual values.

---

2. Explain the Anscombe’s Quartet in Detail. (3 marks)

Answer:

Anscombe’s Quartet, created by Francis Anscombe, comprises four datasets with identical summary statistics but distinct visual representations. The datasets share mean, variance, and correlation coefficient, yet their graphs tell different stories.

> Consider a scenario where a statistician is analyzing the relationship between the number of hours spent studying (variable X) and the exam scores achieved (variable Y) for four different groups of students. The statistician collects data from each group and calculates summary statistics, finding surprising similarities in the mean, variance, and correlation coefficient across all groups.

Dataset I: Linear Relationship

X (Study Hours)	Y (Exam Scores)
10	85
8	90
13	78
9	88
11	92

this dataset exhibits a clear linear relationship. If plotted, it would likely result in a well-fitted linear regression line.

Dataset II: Non-linear Relationship

X (Study Hours)	Y (Exam Scores)
10	85
8	90
13	55
9	78
11	95

Here, the relationship is non-linear. A linear regression model might not capture the underlying pattern effectively.

Dataset III: Linear Relationship with Outlier

X (Study Hours)	Y (Exam Scores)
10	85
8	90
13	78
9	88
20	50 (Outlier)

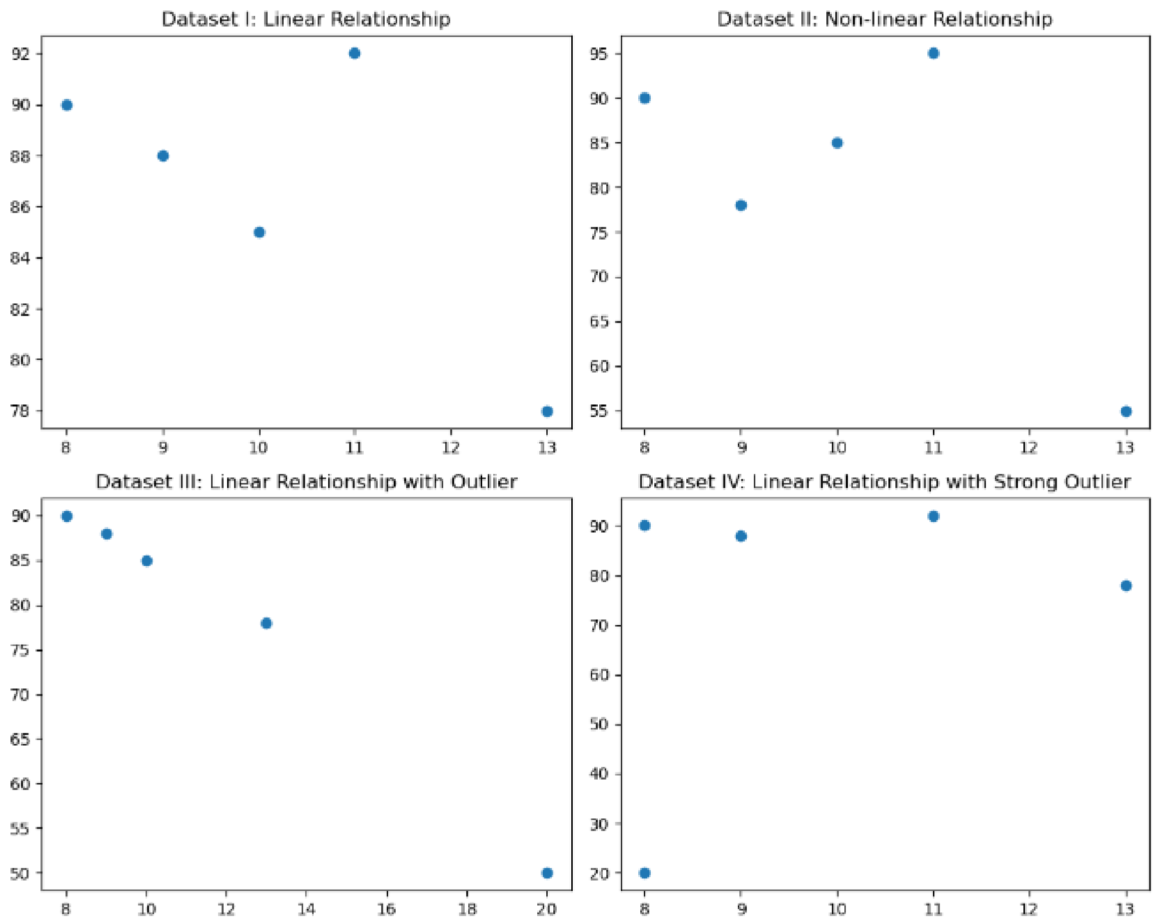
This dataset has a linear relationship but is heavily influenced by an outlier. The presence of the outlier could significantly impact regression analysis.

Dataset IV: Linear Relationship with Strong Outlier

X (Study Hours)	Y (Exam Scores)
8	20 (Outlier)
8	90
13	78
9	88
11	92

In this case, a single extreme outlier dominates the dataset, potentially leading to a misleadingly high correlation coefficient.

Visualizing Anscombe's Quartet:



Now, if we were to graphically represent these datasets, we would observe the following:

- **Dataset I:** Clear linear trend.
- **Dataset II:** Non-linear pattern.
- **Dataset III:** Linear, but influenced by an outlier.
- **Dataset IV:** Linear, but strongly affected by a single outlier.

Despite the diverse patterns, all datasets share similar summary statistics. This showcases the limitation of relying solely on numerical summaries and emphasizes the need for visual exploration to truly understand the nature of the data.

Key Takeaways:

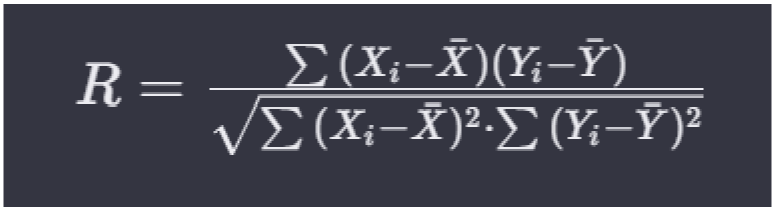
- **Lesson 1:** Statistical summaries may not capture the full story.
- **Lesson 2:** Visual inspection is crucial for detecting nuances.
- **Lesson 3:** Real-world datasets can have unique characteristics that demand a comprehensive analysis.

3. What is Pearson's R? (3 marks)

Answer:

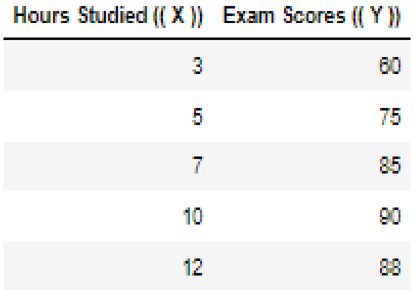
Pearson's correlation coefficient, denoted as R, is a statistical measure quantifying the strength and direction of a linear relationship between two continuous variables. Here are six key points about Pearson's R:

1. **Range**: The coefficient ranges from -1 to 1, indicating the strength and direction of the correlation.
   - (R = 1): Perfect positive linear correlation
   - (R = -1): Perfect negative linear correlation
   - (R = 0): No linear correlation
2. **Formula:**



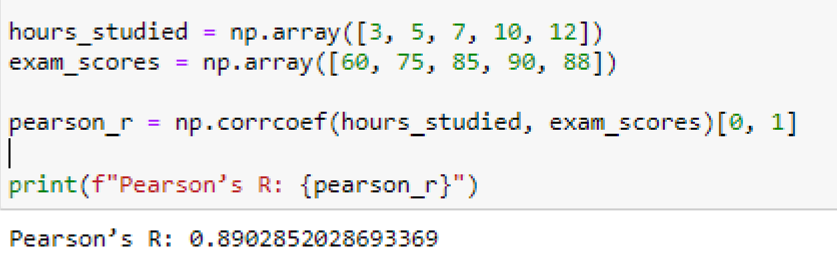
$$R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \cdot \sum (Y_i - \bar{Y})^2}}$$

3. Interpretation:
   - Positive (R): Direct proportional relationship
   - Negative (R): Inverse proportional relationship
   - (R) close to 0: Weak or no linear relationship
4. Strength of Relationship:
   - (|R|) near 1: Strong linear relationship
   - (|R|) near 0: Weak or no linear relationship
5. Example: Consider the following dataset representing hours studied ((X)) and corresponding exam scores ((Y))



Hours Studied ((X))	Exam Scores ((Y))
3	60
5	75
7	85
10	90
12	88

6. --



```
hours_studied = np.array([3, 5, 7, 10, 12])
exam_scores = np.array([60, 75, 85, 90, 88])

pearson_r = np.corrcoef(hours_studied, exam_scores)[0, 1]

print(f"Pearson's R: {pearson_r}")
```

```
Pearson's R: 0.8902852028693369
```

---

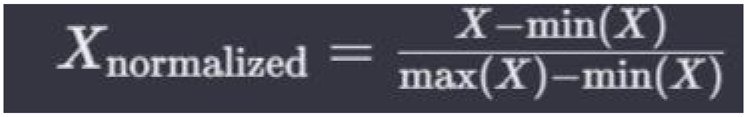
4. What is Scaling? Why is Scaling Performed? What is the Difference Between Normalized Scaling and Standardized Scaling? (3 marks)

Answer:

Scaling:

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

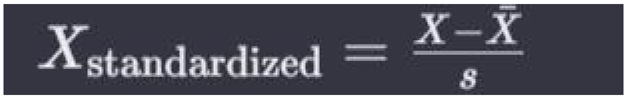
Normalized Scaling:



$$X_{\text{normalized}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- Range: 0 to 1.
- Use Case: Suitable for algorithms sensitive to magnitude (e.g., K-Nearest Neighbors).

Standardized Scaling:



$$X_{\text{standardized}} = \frac{X - \bar{X}}{s}$$

- Properties: Mean (bar{X}) becomes 0, Standard Deviation ((s)) becomes 1.
- Use Case: Suitable for algorithms where magnitude matters less (e.g., linear regression).

Aspect	Normalized Scaling	Standardized Scaling
Range of Transformed Values	**Typically between 0 and 1**	**No fixed range; values can be positive or negative**
Sensitivity to Outliers	**More sensitive, as it uses the range of the data**	**Less sensitive, as it is based on mean and standard deviation**
Interpretability	**Easier to interpret as values are within a specific range (e.g., 0 to 1)**	**Values are centered around 0 with a standard deviation of 1, making interpretation less intuitive**
Computation	**Requires knowledge of the minimum and maximum values of each feature**	**Requires computation of mean and standard deviation for each feature**
Algorithm Sensitivity	**Suitable for algorithms sensitive to feature scales, like k-NN**	**Suitable for algorithms that assume a standard normal distribution of features, such as PCA**

---

5. VIF and Infinite Values: Why Does this Happen? (3 marks)

Answer:

Variance Inflation Factor (VIF) quantifies multicollinearity in regression models. Infinite VIF occurs with perfect multicollinearity:

- **Perfect Multicollinearity:**
    - One predictor perfectly predicts another.
    - Results in numerical instability.
- **Consequences:**
    - **Infinite VIF arises due to perfect prediction.**
    - **Solution: Drop one variable causing perfect multicollinearity.**

When the variance inflation factor (VIF) becomes infinite, it indicates a perfect linear relationship between the predictor variables. The VIF measures the extent to which a predictor variable can be expressed as a linear combination of other predictor variables in the model. If this relationship is perfect (i.e., one predictor can be precisely predicted by a linear combination of others), the VIF becomes infinite.

This perfect multicollinearity can happen when one or more predictor variables in the model are a linear combination of others, leading to an exact relationship. In such cases, it becomes impossible to estimate the coefficients accurately, and the model's stability is compromised. To address this issue, it's common to identify and remove one of the highly correlated variables or re-evaluate the model structure.

---

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution. It compares the quantiles of the observed data against the quantiles of a specified theoretical distribution, typically the normal distribution.

Purpose and Use in Linear Regression

Use of Q-Q Plot:

- Visualizing Distribution: Q-Q plots help visualize how well the observed data aligns with the expected distribution.
- Normality Check: In linear regression, Q-Q plots are often used to check the assumption of normality in the residuals.

Importance in Linear Regression:

- Normality Assumption: One of the key assumptions in linear regression is that the residuals (the differences between observed and predicted values) should be normally distributed. Q-Q plots provide a visual way to assess this assumption.
- Residual Analysis: If the points in the Q-Q plot deviate significantly from the theoretical line, it suggests deviations from normality in the residuals, which can impact the reliability of statistical inferences.

Interpreting Q-Q Plot:

- Points on the Line: Good alignment with the reference line indicates that the residuals are approximately normally distributed.
- Deviation from Line: Deviations or patterns in the plot may indicate departures from normality.