Submission 2 - Linear Regression Subjective Questions – Akash Komarla Ashok Kumar

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?

A - The following insights were derived from the visual analysis of categorical variables:

- Seasonal Trends: The fall season recorded the highest number of bookings. Additionally, there was a notable increase in bookings from 2018 to 2019.
- Monthly Distribution: Bookings peaked during the months of May through October, with a steady rise from the beginning of the year, reaching a maximum mid-year, and declining towards year-end.
- Weather Conditions: Clear weather conditions were associated with a higher volume of bookings, which aligns with general expectations.
- Day of the Week: Bookings were more frequent on Thursdays through Sundays, compared to the earlier part of the week.
- Holiday Impact: Booking volumes were lower on non-holidays, suggesting a preference among individuals to spend holidays at home with family.
- Workday Comparison: There was no significant difference in booking volumes between working and non-working days.
- Year-over-Year Growth: The year 2019 experienced a higher number of bookings compared to 2018, indicating positive business growth.

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

A - Using drop_first=True during dummy variable creation is essential to avoid the dummy variable trap, which occurs when multicollinearity is introduced due to redundant information among the dummy variables. By dropping the first category, we reduce the number of dummy variables from *k* to *k−1*, where *k* is the number of unique categories in the original categorical variable.

This approach ensures that the model remains interpretable and avoids issues related to perfect multicollinearity, which can negatively impact the performance of regression models.

Syntax: drop_first: bool, default=False

This parameter determines whether to drop the first level of categorical variables, thereby generating *k−1* dummy variables instead of *k*.

Example: If a categorical column contains three categories: A, B, and C, creating dummy variables with drop_first=True will result in only two variables (e.g., B and C). If both are 0, it is implicitly understood that the category is A.

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

A - Upon examining the pair plot of the numerical features, the variable temp demonstrates the strongest positive correlation with the target variable. This indicates that as the temperature increases, the target variable tends to increase as well, suggesting a meaningful linear relationship between the two.

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

A - To ensure the reliability and validity of the Linear Regression model, I evaluated the following five key assumptions:

- Normality of Error Terms
    - Verified that the residuals (error terms) are approximately normally distributed, which is essential for accurate hypothesis testing and confidence intervals.
- Multicollinearity Check
    - Assessed the presence of multicollinearity among independent variables using metrics such as the Variance Inflation Factor (VIF). The goal was to ensure that multicollinearity is minimal or insignificant.
- Linearity
    - Confirmed that a linear relationship exists between the independent variables and the dependent variable, which is a fundamental assumption of linear regression.
- Homoscedasticity
    - Checked for constant variance of residuals across all levels of the independent variables. The absence of any discernible pattern in the residual plots indicates homoscedasticity.
- Independence of Residuals
    - Ensured that residuals are independent and not autocorrelated, which is crucial for the validity of regression estimates.

5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A - The final regression model identified the following three features as the most significant contributors to predicting the demand for shared bikes:

- Temperature (temp) – A key continuous variable indicating that higher temperatures are positively associated with increased bike usage.
- Winter Season (winter) – A seasonal indicator suggesting that demand patterns vary significantly during the winter months.
- September (sep) – A monthly indicator reflecting a notable increase in demand during the month of September.

These features collectively capture both environmental and temporal influences on bike-sharing demand, enhancing the model's predictive accuracy.

General Subjective Questions

1) Explain the linear regression algorithm in detail.
A - Definition: Linear Regression is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X) by fitting a linear equation to observed data.

Mathematical Representation: For Simple Linear Regression (one independent variable), the model is:

Y=mX+c$Y=mX+c$

Where:

- Y = Dependent variable (target)
- X = Independent variable (predictor)
- m = Slope of the line (effect of X on Y)
- c = Y-intercept (value of Y when X = 0)

Types of Linear Regression:

- Simple Linear Regression – One independent variable.
- Multiple Linear Regression – More than one independent variable.

Nature of Relationship:

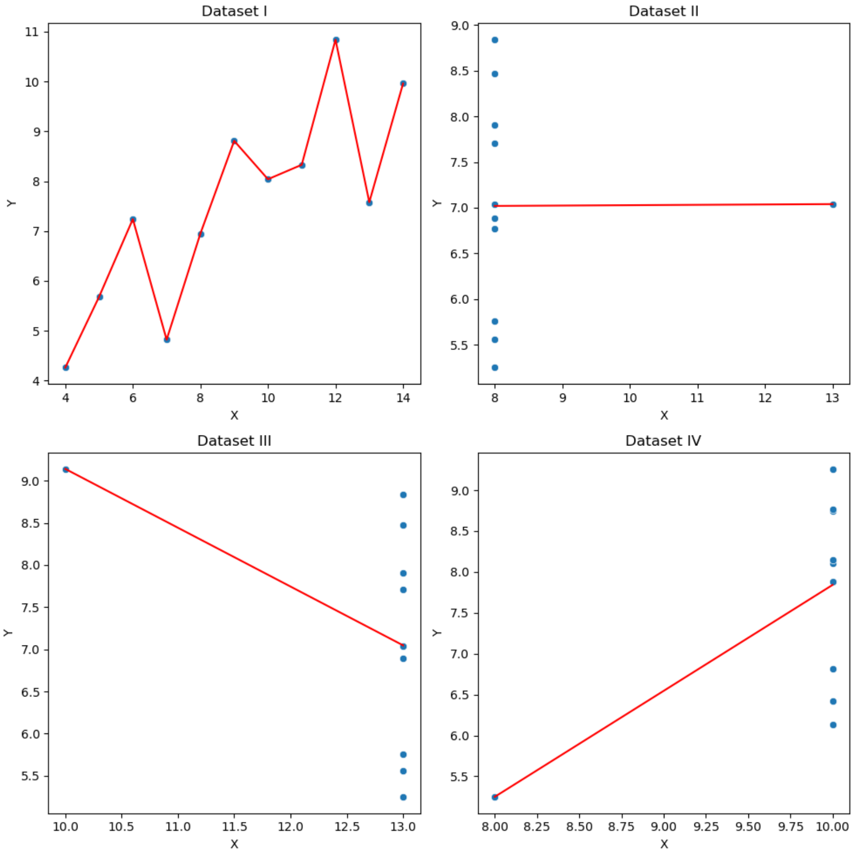
- Positive Linear Relationship: As X increases, Y increases.
- Negative Linear Relationship: As X increases, Y decreases.

Assumptions of Linear Regression:

- Linearity: The relationship between X and Y is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: Constant variance of residuals.
- Normality: Residuals (errors) are normally distributed.
- No Multicollinearity: Independent variables are not highly correlated.
- No Autocorrelation: Residuals are not correlated with each other.

2) Explain the Anscombe's quartet in detail.

A - Here is the visual representation of Anscombe's Quartet, showing how four datasets with nearly identical statistical properties can look dramatically different when plotted:



Key Observations from the Graphs:

Dataset I: Shows a clear linear relationship — a good fit for linear regression.

Dataset II: Curved pattern — not suitable for linear modeling.

Dataset III: Appears linear, but one outlier distorts the regression line.

Dataset IV: Most points are constant, with one extreme outlier creating a misleading correlation.

This visualization powerfully demonstrates why data visualization is essential in data analysis — summary statistics alone can hide critical insights.

3) What is Pearson's R?

A - Pearson's R is a statistical measure that calculates the strength and direction of the linear relationship between two continuous variables. It is also known as the Pearson correlation coefficient.

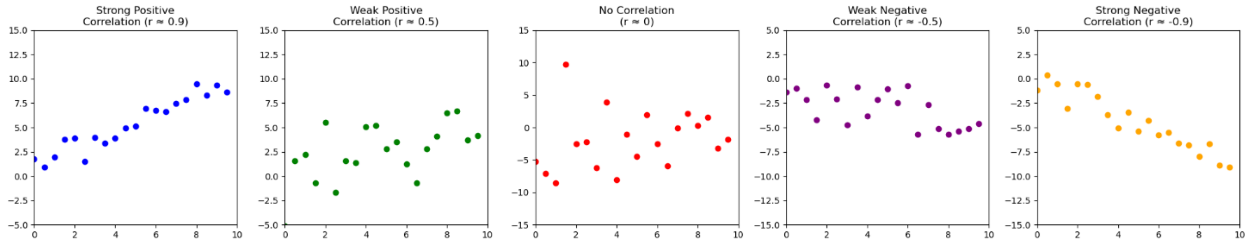
Formula:

r=∑(xi−x¯)(yi−y¯)∑(xi−x¯)2∑(yi−y¯)2r=∑(xi−x¯)2∑(yi−y¯)2∑(xi−x¯)(yi−y¯)

Where:

- xi,yi*xi*,*yi* are individual sample points
- x¯,y¯*x¯*,*y¯* are the means of X and Y

Here is a visual representation of Pearson's correlation coefficient (r) across different scenarios:



Interpretation of Each Plot:

- Strong Positive Correlation (r ≈ +0.9)
  As X increases, Y increases in a nearly linear fashion.
- Weak Positive Correlation (r ≈ +0.5)
  A general upward trend exists, but with more scatter.
- No Correlation (r ≈ 0)
  No discernible pattern between X and Y.
- Weak Negative Correlation (r ≈ -0.5)
  A general downward trend, but not tightly clustered.
- Strong Negative Correlation (r ≈ -0.9)
  As X increases, Y decreases in a nearly linear fashion.

This diagram clearly shows how the value of Pearson's r reflects the strength and direction of a linear relationship between two variables.

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A - Scaling is a data preprocessing technique used to adjust the range or distribution of features (independent variables) in a dataset so they can be compared on a common scale.

Scaling is important because:

- Many machine learning algorithms (e.g., KNN, SVM, gradient descent-based models) are sensitive to the magnitude of feature values.
- Features with larger ranges can dominate those with smaller ranges, leading to biased models.
- It helps speed up convergence in optimization algorithms.

Difference Between Normalized and Standardized Scaling:

Feature	Normalized Scaling	Standardized Scaling
Also known as	Min-Max Scaling	Z-score Normalization
Formula	*X'=fracX−XminXmax−Xmin*	*X'=fracX−musigma*

Output Range	[0, 1] (or any custom range)	Mean = 0, Standard Deviation = 1
Use Case	When data is not normally distributed	When data is normally distributed
Sensitive to	Outliers	Less sensitive to outliers

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   - VIF (Variance Inflation Factor) measures how much the variance of a regression coefficient is inflated due to multicollinearity among the independent variables. Hence, becomes infinite.
   - When VIF Becomes Infinite:
     VIF is calculated as: VIF*i*=1/*Ri*^2
     Where:
   - Ri^2*Ri*^2 is the coefficient of determination when the i^*th* independent variable is regressed on all other independent variables.

   If:

   - Ri^2=1, then:

   VIFi=1/1−1=∞

   This happens when:

   One independent variable is a perfect linear combination of one or more other independent variables.

   In other words, there is perfect multicollinearity.

   Conclusion:

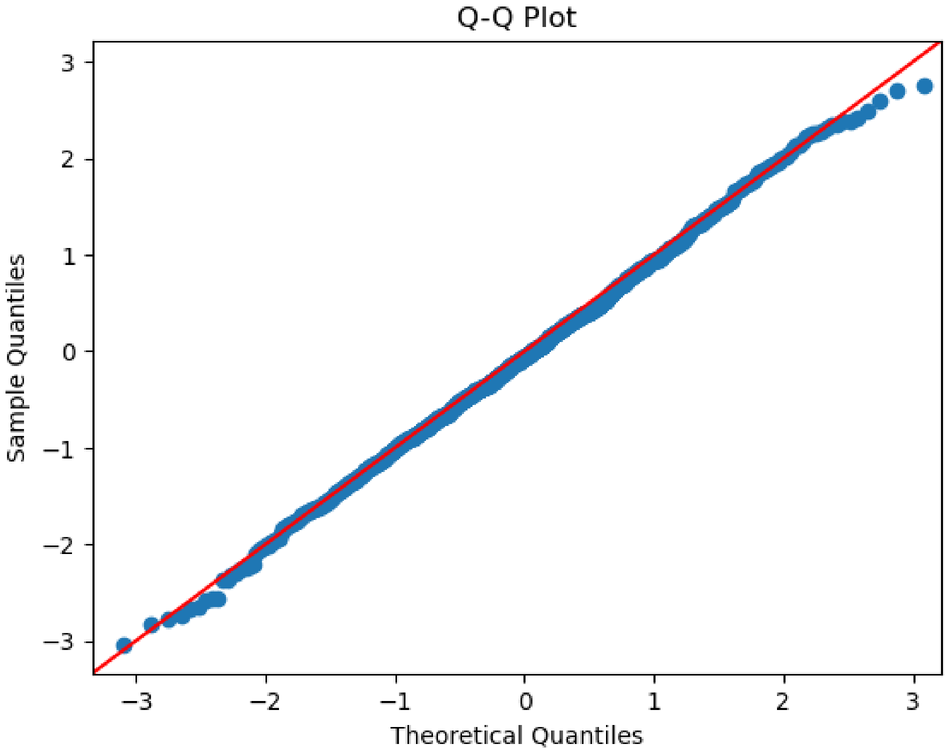
   An infinite VIF indicates perfect multicollinearity, which makes it impossible to estimate the regression coefficients uniquely. This is a serious issue in regression modeling and must be addressed by removing or combining correlated variables.

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

- A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset with a theoretical distribution—typically the normal distribution. It plots the quantiles of the sample data against the quantiles of the theoretical distribution.
- In linear regression, a Q-Q plot is primarily used to:
  - Check the normality of residuals (errors).
  - Validate one of the key assumptions of linear regression:
    1. Residuals should be normally distributed.

- If the residuals lie along the 45° line, they are approximately normally distributed.
- Deviations from the line indicate non-normality, which may affect the validity of statistical tests and confidence intervals.

- Here is a Q-Q plot of normally distributed data:



This plot shows that the data points closely follow the 45° line, indicating that the residuals are normally distributed.