Assignment-based Subjective Questions

Question 1**. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)**

Total Marks**: 3 marks (Do not edit)**

Answer: **<Your answer for Question 1 goes below this line> (Do not edit)**

1. Fall season have more bookings

2.Most of the bookings are done during the month of may, june, july, aug, sep, oct.

3. Number of booking are more for 2019.

4.Clear have more booking.Thu, Fir, Sat and Sun have more number of bookings.

5. NON holiday day booking are less in number

6. Clear sky lead to more bookings ,

7. Bookings on working and non working days were same .

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: **<Your answer for Question 2 goes below this line> (Do not edit)**

For dummy variable creation its used and to assure extra column is dropped. Ie (K-1 dummies)

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: **<Your answer for Question 3 goes below this line> (Do not edit)**

Temp have highest correlation with cat.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: **<Your answer for Question 4 goes below this line> (Do not edit)**

Error graph to check normality .

No auto correlation

No multicollinearity

No pattern in residual

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: **<Your answer for Question 5 goes below this line> (Do not edit)**

Temp

Sep

Winter

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)

Total Marks: 4 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear Regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to fit a straight line (or hyperplane in multiple

dimensions) to the data that minimizes the difference between the predicted and actual values. This is done by finding the best-fitting line using a technique called **Ordinary Least Squares (OLS)**, which minimizes the **sum of squared errors** (the vertical distance between the data points and the line).

Equation of line :

y = beta_0 + beta_1 x

Where beta0 is the intercept and

beta1 is the slope (coefficient). Linear regression assumes a linear relationship between the variables, and it is widely used for prediction, forecasting, and identifying trends in data.

The model assumes a linear relationship and requires assumptions like homoscedasticity, no multicollinearity, and normally distributed residuals for accurate predictions. Linear regression is widely used in finance, economics, and natural sciences for trend analysis, forecasting, and understanding the relationships between variables.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.)

Each dataset consists of 11 data points, and although the statistical summary (such as mean, variance, and correlation) is the same across all four, their scatterplots reveal distinct characteristics. One dataset shows a linear relationship, another shows a perfect curve, a third exhibits a clear outlier, and the fourth is highly concentrated with one unusual point. Anscombe's Quartet underscores the point that relying solely on summary statistics (like mean and variance) can be misleading, and it's crucial to examine the data visually for a deeper understanding of its underlying patterns and potential anomalies.

Question 8. What is Pearson's R? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R, is a measure of the **linear relationship** or **correlation** between two variables. It quantifies the strength and direction of the linear relationship on a scale from **-1 to +1**.

- **+1**: A perfect positive linear relationship (as one variable increases, the other also increases in a perfectly linear manner).
- **-1**: A perfect negative linear relationship (as one variable increases, the other decreases in a perfectly linear manner).
- **0**: No linear relationship (the variables do not have any linear correlation)

Key Points:

- **Strength**: The closer r is to 1 to -1 , stronger is linear relation
- **Direction**: Positive values indicate a positive correlation, and negative values indicate a negative correlation.
- **Limitations**: Pearson's R only measures **linear** relationships, meaning it will not capture non-linear relationships like exponential or quadratic correlations. Additionally, it can be influenced by outliers.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is the process of adjusting the range and distribution of data features so that they are comparable and on the same scale. It is an essential preprocessing step in machine learning, particularly when different features have different units or magnitudes. The goal of scaling is to transform features to a common range so that machine learning algorithms, especially those sensitive to the scale of data, can perform optimally.

Normalization (Min-Max Scaling) is useful when you want to scale your data into a specific range, typically for algorithms that rely on bounded values like neural networks.

Standardization (Z-score Scaling) is more appropriate when your data is not normally distributed and you want to remove the mean and scale the data based on variance, making it suitable for most machine learning algorithms, especially those based on distance or linear models.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

(VIF) measures the extent to which the variance of a regression coefficient is inflated due to collinearity (correlation) among the predictor variables in a multiple regression model. A **high VIF** indicates that a predictor is highly correlated with other predictors, which can cause problems like multicollinearity and unreliable parameter estimates.

A **VIF** becomes infinite when there is **perfect multicollinearity** or **perfect correlation** between a predictor variable and one or more other predictor variables. This means that one of the independent variables is **linearly dependent** on another variable (or a combination of others), leading to an issue called **perfect multicollinearity**.

Vif = 1/ (1-R square)

Infinite VIF occurs when:

R Square = 1 , Xi can be perfectly predicted from the other variables in the model.

In this case, the denominator becomes zero, leading to an infinite VIF.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A Q-Q **plot** is a graphical tool used to assess if a dataset follows a particular theoretical distribution, typically the **normal distribution**. It compares the quantiles of the observed data with the quantiles of a theoretical distribution. If the data follows the specified distribution, the points on the Q-Q plot should lie approximately along a straight line. Deviations from this line indicate that the data does not follow the assumed distribution.

In linear regression, one of the **key assumptions** is that the **residuals** (errors) of the model should be normally distributed. The Q-Q plot is often used to check this assumption:

1 **Checking Normality of Residuals**:

- In linear regression, it's important that the residuals (the differences between the observed and predicted values) follow a normal distribution. This is because many statistical tests, such as hypothesis tests for regression coefficients (e.g., t-tests), rely on the assumption that the residuals are normally distributed.
- A **Q-Q plot of residuals** allows you to visually assess if the residuals deviate from normality. If the residuals are normally distributed, the Q-Q plot should show the points following a straight line.

2 **Identifying Problems**:

- **Non-normal residuals** can indicate model problems such as:
  - **Outliers**: If the residuals contain extreme values, the Q-Q plot will show deviations at the ends of the plot.
  - **Skewness**: If the residuals are skewed, the Q-Q plot will show a noticeable curve.
  - **Heteroscedasticity**: Non-constant variance of residuals may sometimes also be detected through deviations in the Q-Q plot.

3 **Implications for Model Accuracy**:

- If the residuals are not normally distributed, it can impact the reliability of the regression results, including confidence intervals, significance tests, and predictions.
- While **non-normality of residuals** is less of a problem for large sample sizes (thanks to the **Central Limit Theorem**), for smaller datasets, non-normal residuals may lead to **biased or inefficient parameter estimates**.