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)

From the analysis of the categorical variables I have found that –

- The fall season attracts the highest number of customers, while the spring season attracts the fewest.
- In the month of June, July, August and September most bookings are done.
- Booking activity is fairly consistent across weekdays. This consistency suggests a steady demand pattern throughout the week.
- On Clear weather high amount of bookings are done and in case of heavy rain no booking were done.
- In 2019 we have seen more number of customer then previous year which indicates progress in business.

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)

Using drop_first=True when creating dummy variables is important because it helps avoid multicollinearity among variables.
While creating dummy variables for all the categories they add up to represent the same information. Dropping one dummy variable reduces the number of features the model has to learn, making it faster to train and helping to avoid overfitting

---

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 and atemp variables has the highest correlation with the target variable.

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)

If relationship between the predictors and the target variable assumption is likely satisfied.
Error terms should be normally distributed.
There should be very little multicollinearity between variables.
Residual values doesn't have strong pattern.
The spread of errors (residuals) should be the same for all predicted values.

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
Year
Light Rain Weather

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 basic yet powerful algorithm for predicting continuous outcomes by finding a linear relationship between independent (predictor) variables and a dependent (target) variable. Linear regression aims to model the relationship between one or more independent variables and a continuous dependent variable by fitting a line (or plane, in the case of multiple variables) that best represents this relationship.

Equation –
In its simplest form (simple linear regression with one predictor variable), linear regression is represented by the equation:

$$y = b_0 + b_1 x$$

y is the predicted output (dependent variable).
b0 is the intercept (value of y when x=0).
b1 is the slope or coefficient for x, showing how much y changes for a one-unit change in x.

Linear regression uses the Least Squares method to find the best-fitting line by minimizing the sum of the squared errors (residuals), represented by:

$$\text{Error} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Linear regression is a foundational algorithm, ideal for simple, interpretable models when there is a clear linear relationship. It's widely used for trend analysis and forecasting in various domains, from finance to healthcare, and remains a valuable tool for understanding relationships in data.

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)

Anscombe's quartet is a set of four datasets created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before interpreting statistical results. The quartet contains four different datasets with nearly identical statistical properties, such as mean, variance, correlation, and linear regression line, yet they look very different when graphed. Anscombe created these datasets to illustrate that summary statistics alone can be misleading and do not capture the full story of the data's structure.

1. Overview of Anscombe's Quartet

The quartet consists of four datasets, each with:

- The same mean for the x-values and y-values
- The same variance for both x and y
- The same linear regression line with a similar slope and intercept

2. The Four Datasets

Dataset I

- This dataset resembles a classic linear relationship.
- When plotted, it shows a fairly strong linear correlation between x and y, with points scattered closely along the regression line.

Dataset II

- This dataset shows a nonlinear relationship, but a linear regression line is still fitted.
- The xxx values are identical across observations, with an outlier near the end.
- The linear regression line does not accurately capture the relationship in this dataset because the data follows a curved, parabolic shape.

Dataset III

- Dataset III has most points clustered on a vertical line, with an extreme outlier affecting the summary statistics.
- Here, the relationship is weakly linear with one influential point that lies on the regression line.

Dataset IV

- In Dataset IV, most points are nearly identical along the y-axis, with a single influential point at the far right.
- This influential point aligns closely with the regression line, which otherwise poorly represents the data.

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)

Pearson's r, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. This coefficient is commonly used in statistics to understand the degree to which two variables are linearly related.

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

To correctly interpret Pearson's r, several assumptions should be met:

Linearity: Pearson's r only captures linear relationships, so it may not accurately reflect nonlinear relationships.
Continuous Data: Both variables should be continuous (interval or ratio scale).
Normality: It's ideal if both variables are normally distributed, although Pearson's r is robust to slight deviations.
No Outliers: Outliers can heavily influence the correlation, making the coefficient misleading.

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)

Scaling is a crucial preprocessing step in data analysis and machine learning that involves transforming the features of a dataset to a common scale. This process helps improve the performance of algorithms that are sensitive to the magnitude of the input features.

Scaling is performed for several reasons:

Improves Convergence: For optimization algorithms (e.g., gradient descent), scaling can help them converge faster by ensuring that all features are on a similar scale.

Enhances Performance: Many machine learning algorithms, particularly those based on distance metrics (like KNN and support vector machines), perform better when features are scaled. Otherwise, features with larger ranges can dominate the distance calculations.

Equal Weighting: Scaling ensures that all features contribute equally to the distance calculations, preventing features with larger ranges from disproportionately influencing the model.

Stability: Scaling can enhance the numerical stability of algorithms, particularly those that involve matrix inversions or other mathematical operations sensitive to the scale of input data.

Differences Between Normalized Scaling and Standardized Scaling

Aspect	Normalized Scaling	Standardized Scaling
Purpose	Rescales data to a specific range (e.g., [0, 1])	Centers data around the mean with a standard deviation of 1
Formula	$X' = \frac{X - \min(X)}{\max(X) - \min(X)}$	$X' = \frac{X - \mu}{\sigma}$
Output Range	[0, 1] or [-1, 1] depending on the implementation	No fixed range; can be any value, positive or negative
Data Distribution	Useful for non-Gaussian distributions	Assumes data is normally distributed
Sensitivity to Outliers	Highly sensitive; outliers can affect min/max values	Less sensitive; outliers can still affect mean and standard deviation but typically less drastically

---

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)

A VIF value can become infinite (or extremely large) under specific conditions, primarily related to the linear relationships among the predictor variables. Here are the key reasons:

Perfect Multicollinearity:
Perfect multicollinearity occurs when one independent variable is an exact linear combination of one or more other independent variables.

Degenerate Cases:

This can also occur in cases where the design matrix (the matrix of independent variables) is not full rank. If the columns of the matrix are linearly dependent (meaning one or more columns can be expressed as a linear combination of others), the inverse of the matrix cannot be computed, leading to a scenario where the VIF becomes infinite.

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)

A Q-Q plot, or quantile-quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (often the normal distribution) or to compare the distributions of two datasets. It plots the quantiles of one dataset against the quantiles of another dataset.

Axes: In a Q-Q plot, the x-axis typically represents the quantiles of a theoretical distribution (like the normal distribution), while the y-axis represents the quantiles of the sample data.
Data Points: Each point on the plot corresponds to a pair of quantiles: one from the theoretical distribution and one from the sample data.
Line of Identity: A 45-degree reference line (line of identity) is usually included in the plot. If the sample data follows the theoretical distribution, the points will closely align with this line.

Use of Q-Q Plots

- **Normality Check**: One of the primary uses of a Q-Q plot in the context of linear regression is to assess whether the residuals of the regression model are normally distributed. This is an important assumption of linear regression.
- **Distribution Comparison**: Q-Q plots can also be used to compare the distribution of residuals to other theoretical distributions (e.g., exponential, uniform) to see if a different distribution might be a better fit.
- **Identify Outliers**: Q-Q plots can help identify outliers or extreme values in the dataset. Points that deviate significantly from the line of identity may indicate the presence of outliers.

Importance of Q-Q Plots in Linear Regression

1. **Assumption Verification**:
   - **Normality of Residuals**: Linear regression assumes that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot helps verify this assumption. If the residuals are not normally distributed, it can affect hypothesis tests and confidence intervals related to the regression coefficients.
2. **Model Diagnostics**:
   - **Assessment of Fit**: By examining the distribution of residuals, analysts can assess how well the regression model fits the data. Deviations from normality might

suggest that the model is not appropriately capturing the underlying data structure.

- **Transformation Decisions**: If a Q-Q plot indicates non-normality, it may prompt the analyst to consider data transformations (e.g., log, square root) to achieve normality.