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, we can have the following inferences.

- More bookings are seen in the season 'Fall' followed by 'summer, winter, spring'. 2019 has more booking than 2018 for all seasons.
- Booking is more during the months June, July, August, September. The booking is gradually increasing from January till September and then decreases.
- Clear weather shows more booking and booking is more in 2019 than 2018.
- While analysing the booking, we can see the bookings are comparable. Out of these Saturday, Sunday and Monday shows high bookings.
- Booking is less on holidays.
- 2019 attracted a greater number of bookings from the previous year, which shows good progress in terms of 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 is important to use during dummy variable creation, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

By using drop_first=True, we drop one of the dummy variables (typically the first category), leaving n-1 dummy variables. This ensures that the remaining dummy variables do not add up to 1 and hence avoids the multicollinearity problem, making the model more interpretable and reliable.

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' variable has the highest correlation with the target variable 'cnt'.

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)

Verified the assumptions as follows:

- Normality of error terms: By plotting the histogram of the residuals and verified if it approximates a normal distribution (bell-shaped curve).
- Multicollinearity: Verified all the independent variables are not correlated
- Linearity: Verified the relationship between the independent variables and the dependent variable is linear.
- Homoscedasticity: Verified by plotting the residuals against the predicted values
- Independence of residuals

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)
Top 3 features that has significant impact towards explaining the demand of the shared bikes are:

- temp
- winter
- sep

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 type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data. When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.

Linear regression fits a linear relationship between the dependent variable Y and one or more independent variables X. The simplest form is simple linear regression, where there is one independent variable. When there are multiple independent variables, it's called multiple linear regression.

Simple linear regression involves predicting a single dependent variable y from one independent variable x using the equation of a straight line:

$y=\beta_0+\beta_1 x+\epsilon$

where:
y is the dependent variable (target).
x is the independent variable (predictor).
β_0 is the intercept (the value of y when x=0).
β_1 is the slope or coefficient of the independent variable x (shows the change in y for a one-unit

change in x).

ϵ is the error term (captures the difference between the predicted and actual values of y).

Multiple linear regression, where there are multiple independent variables $x_1, x_2, \ldots, x_p$ the equation becomes:

$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\epsilon$

The goal of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the actual values of the dependent variable and the predicted values. This is achieved by minimizing the sum of squared errors (SSE) or residual sum of squares (RSS). To minimize the sum of squared errors and find the optimal parameters $\beta_0, \beta_1, \ldots, \beta_p$, we typically use a method called Ordinary Least Squares (OLS).

Once the coefficients are computed, we can make predictions for new data points by plugging the values of the independent variables into the regression equation. For a new observation with input features $x_1, x_2, \ldots, x_p$, the predicted value of the dependent variable $y\hat{}$ is:

$\hat{y}=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$, and this predicted value $\hat{y}$ is the output of the linear regression model.

Once the model is trained and predictions are made, we evaluate the model's performance using several metrics like, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2).

Linear regression makes certain assumptions about the data and the relationship between variables:

Linearity: The relationship between the dependent and independent variables is linear.

Independence: The residuals (errors) are independent.

Homoscedasticity: The variance of the residuals is constant across all levels of the independent variable(s).

Normality: The residuals should be normally distributed.

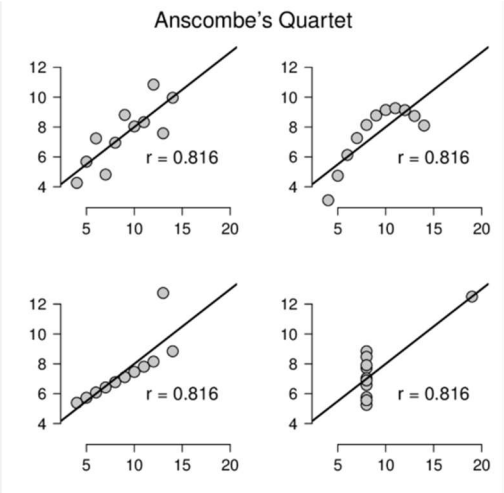
No multicollinearity: The independent variables should not be highly correlated with each other (for multivariate regression).

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 famous dataset created by the statistician Francis Anscombe in 1973. It consists of four datasets that have nearly identical descriptive statistics (mean, variance, correlation, etc.) but different underlying distributions and relationships. Anscombe created this dataset to emphasize the importance of visualizing data, as it illustrates how summary statistics alone can be misleading, and how important it is to consider the context and patterns in the data.

The quartet consists of four different data sets, each with two variables (X and Y), where the descriptive statistics are nearly identical across the four datasets. However, when we visualize the data (using scatter plots), we will see very different relationships between X and Y in each case.



- Dataset I: Linear relationship with some random noise.
- Dataset II: Non-linear relationship, resembling a curve.
- Dataset III: Linear relationship but with an outlier that significantly influences the fit.
- Dataset IV: Vertical line due to an outlier, which affects the correlation and regression analysis.

Even though the summary statistics (mean, variance, correlation) are identical across the datasets, the patterns in the data are very different. Visualizing the data via scatter plots reveals the underlying differences that the summary statistics miss. Anscombe's Quartet is a perfect example of the importance of data visualization and contextual analysis.

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, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It indicates how strongly two variables are related to each other and whether the relationship is positive or negative. Pearson's R is widely used in statistics to assess the degree to which two variables are linearly associated.

The value of Pearson's R ranges from -1 to 1. +1 indicates a perfect positive linear relationship. -1 indicates a perfect negative linear relationship. 0 indicates no linear relationship.

Pearson's R measures only linear relationships. If the relationship between two variables is non-linear, Pearson's R may not adequately reflect the strength of the relationship. Pearson's R is sensitive to outliers, if there is a single extreme value can dramatically affect the correlation coefficient. Pearson's R is also scale-invariant, so it doesn't matter whether the variables are measured in different units. It compares the relative movements of the two variables, not their absolute values.

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 transforming features so that they have a specific range or distribution. Scaling is an essential preprocessing step, particularly for algorithms that are sensitive to the magnitude of the features. It ensures that each feature contributes equally to the analysis, helping to improve the performance of machine learning models and ensuring faster convergence during training. Scaling is a crucial step, especially when the features of the dataset have different units or magnitudes.

For optimization algorithms, scaling can lead to faster convergence. Algorithms like gradient descent benefit from features having similar scales, which ensures the gradients are appropriately balanced. Scaling ensures that no single feature disproportionately influences the model due to its magnitude. This leads to more balanced learning.

There are mainly two types of scaling techniques:
Normalized scaling (Min-Max Scaling)
Standardized scaling (Z-score Scaling)

Normalized scaling	Standardized scaling
Scales the data to a fixed range (typically [0, 1]).	Data is centered around 0 with a standard deviation of 1.
Sensitive to outliers, as they affect the minimum and maximum values.	Less sensitive to outliers since scaling is based on mean and standard deviation.
Assumes that the data should lie in a specific range (e.g., [0, 1]).	Assumes that the data follows a Gaussian distribution.
Often used in neural networks or algorithms that rely on a bounded input range.	Typically used for algorithms that assume normality, such as linear regression, logistic regression.
Values are scaled to a fixed range (e.g., [0, 1]).	Values are centered around 0 and have unit variance.

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>

The Variance Inflation Factor (VIF) is a measure used to assess the level of multicollinearity in a set of features (independent variables) in a linear regression model. A high VIF indicates that a predictor is highly collinear with other predictors, which can cause issues in model interpretation and stability.

The formula for VIF for a given predictor variable is $1/(1-R_i^2)$. A VIF becomes infinite when the R-squared value (R_i^2) is 1, which means that the i^{th} predictor is perfectly collinear with the other predictors in the model. This happens when one predictor variable is a linear combination of one or more other predictor variables in the model. If two or more predictors are highly collinear, we can remove one of them from the model. This will reduce the multicollinearity and eliminate the 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 (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, most commonly the normal distribution. It is a scatter plot that compares the quantiles of the dataset with the quantiles of a reference probability distribution (e.g., normal distribution).

In a Q-Q plot, the x-axis represents the theoretical quantiles (e.g., from the normal distribution), while the y-axis represents the quantiles of the data being tested. If the data follows the reference distribution, the points will lie approximately along a straight line, typically a 45-degree line.

In the context of linear regression, a Q-Q plot is commonly used to check the assumption of normality of the residuals. Ensuring that the residuals are approximately normally distributed is crucial for making valid inferences about the regression model.

Use and importance of Q-Q Plot in Linear Regression:

A key assumption in linear regression is that the residuals are normally distributed. This is important for the validity of inferential statistics, such as t-tests for coefficients, confidence intervals, and hypothesis testing. The Q-Q plot helps to visually check whether the residuals of the regression model approximately follow a normal distribution. A non-linear pattern in the Q-Q plot can indicate that the residuals deviate from normality. Outliers are extreme values that do not follow the pattern of the rest of the data. These can be detected in a Q-Q plot as points that deviate significantly from the straight line, especially in the tails of the plot.