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)

Answer: Based on the observations from your analysis using box plots and bar plots on categorical columns, here are the key points derived:

Weather: Bookings are most frequent when the weather is clear. Similar to the month, there is an increasing trend in bookings from January to October, followed by a decrease. The number of bookings also significantly increased for each weather condition from 2018 to 2019.

Day of the Week: Thursdays, Fridays, and Saturdays attract the highest number of bookings.

Holidays: The number of bookings is generally lower during holidays, which aligns with the expectation that people prefer to spend time at home with family during these periods.

Working Day vs. Non-Working Day: The number of bookings appears to be approximately the same on working days and non-working days. However, there is an overall increase in the number of bookings from 2018 to 2019.

Business Progress: From a business perspective, there is progress evident from 2018 to 2019, with an increase in the overall number of bookings, indicating positive growth for the business.

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

Ans: drop_first = True, is import to use as it helps in reducing the extra column created during dummy varirable creation. Hence using drop_first=True during dummy variable creation is important to avoid multicollinearity and enhance the interpretability of regression coefficients. It reduces correlation between variables and sets a reference category for comparison, leading to more meaningful and efficient results.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: The 'temp' variable has the highest correlation with the target varible.

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

Ans: I have validated the assumption of Linear Regression Model based on below 5 assumptions -

- Normality of error terms: Error terms should be normally distributed

- Multicollinearity check : There should be insignificant multicollinearity among variables.
- Linear relationship validation: Linearity should be visible among variables
- Homoscedasticity: There should be no visible pattern in residual values.
- ndependence of residuals: No auto-correlation

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

- temp
- winter
- sep

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression models the relationship between independent variables and a dependent variable using a linear equation.

- The model aims to find the best-fitting line (or hyperplane) that minimizes the difference between predicted and actual values.
- Simple linear regression has one independent variable, while multiple linear regression involves multiple independent variables.
- Coefficients in the equation represent the slope and intercept, indicating the impact of independent variables on the dependent variable.
- The model is trained by estimating the coefficients using methods like ordinary least squares or gradient descent.
- Model performance is evaluated using metrics like mean squared error, root mean squared error, and R-squared.
- Predictions are made by plugging in new independent variable values into the trained model.
- Assumptions of linearity, normality of residuals, independence, and homoscedasticity should be

validated.

- Extensions and improvements include polynomial regression, regularization techniques (ridge, lasso), and handling multicollinearity.

2**. Explain the Anscombe's quartet in detail. (3 marks)**

Ans: Anscombe's quartet is a collection of four datasets that have nearly identical statistical properties, but exhibit vastly different patterns when visualized. The quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visual exploration and analysis in understanding data. It serves as a reminder that summary statistics alone may not provide a complete understanding of the underlying relationships in the data.

Here is a detailed explanation of Anscombe's quartet:

- Dataset Overview: Anscombe's quartet consists of four datasets, each containing 11 (x, y) paired values. The datasets are labeled I, II, III, and IV.
- Statistical Properties: Despite having different patterns, all four datasets share nearly identical summary statistics, including means, variances, correlation coefficients, and regression coefficients.
- Dataset I: Dataset I appears to have a linear relationship between x and y. It follows a simple linear regression model and has relatively low variance.
- Dataset II: Dataset II is non-linear, forming a curved relationship. It demonstrates that even with a clear non-linear pattern, the summary statistics can be similar to Dataset I.
- Dataset III: Dataset III exhibits a strong linear relationship with an outlier point, significantly influencing the regression line and correlation coefficient. This highlights the impact of outliers on summary statistics.
- Dataset IV: Dataset IV consists of three distinct clusters of points. Despite having no clear linear relationship, the summary statistics remain similar to the other datasets.
- Visual Analysis: Visualizing the datasets through scatter plots, regression lines, or other graphical methods is crucial for gaining insight into the patterns and relationships present.
- Implications: Anscombe's quartet serves as a cautionary example that relying solely on summary statistics can be misleading. It highlights the importance of exploring and visually analyzing data to uncover patterns, outliers, and other essential characteristics.

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

Ans: Pearson's correlation coefficient, often referred to as Pearson's R, is a statistical measure that

quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1, with a value of 0 indicating no linear correlation.

Summary of Pearson's R:

- Definition: Pearson's R measures the degree of association or correlation between two variables. It assesses how closely the data points align to a straight line relationship.
- Range: The correlation coefficient ranges from -1 to +1. A value of -1 indicates a perfect negative correlation, +1 indicates a perfect positive correlation, and 0 indicates no linear correlation.
- Strength: The absolute value of Pearson's R represents the strength of the correlation. Values closer to -1 or +1 indicate a stronger relationship, while values closer to 0 indicate a weaker or no correlation.
- Direction: The sign of Pearson's R indicates the direction of the relationship. A positive value indicates a positive correlation, meaning that as one variable increases, the other tends to increase. A negative value indicates a negative correlation, where as one variable increases, the other tends to decrease.
- Interpretation: The magnitude and direction of Pearson's R can be interpreted to understand the relationship between variables. A value close to -1 or +1 suggests a strong and consistent relationship, while values closer to 0 suggest a weaker or non-linear relationship.
- Assumptions: Pearson's R assumes that the relationship between the variables is linear, the variables are normally distributed, and there is homoscedasticity (constant variance) in the data.
- Limitations: Pearson's R measures only linear relationships and may not capture non-linear associations. It is also sensitive to outliers and can be influenced by extreme observations.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

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

(3 marks)

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value.That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along thisreference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both datasets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square andKolmogorov-Smirnov 2-sample tests.