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)

Ans - Based on the analysis of the categorical variables from the dataset, here are some key insights about their effect on the number of bookings:

- Seasonal Trends: The fall season had the highest number of bookings. There was a noticeable increase in bookings from 2018 to 2019 in each season.
- Monthly Trends: Bookings peaked during the months of May, June, July, August, September, and October. The trend showed an increase in bookings from the beginning of the year, peaking in mid-year, and then declining towards the end of the year.
- Weather Conditions: Clear weather conditions saw more bookings, which makes sense as people prefer to go out when the weather is good.
- Day of the Week: Thursdays, Fridays, Saturdays, and Sundays had more bookings compared to the start of the week (Monday to Wednesday).
- Holidays: There were fewer bookings on holidays, likely because people prefer to stay home and spend time with family on these days.
- Working Days: The number of bookings was almost the same on working days and non-working days.
- Yearly Trends: There was a significant increase in bookings in 2019 compared to 2018, indicating positive business growth.

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

Ans - Using drop_first=True during dummy variable creation is important because it helps in reducing the number of columns created and avoids multicollinearity issues among the dummy variables.

When drop_first=True is used, it removes the first category level from the categorical variable when creating dummy variables. This is based on the assumption that the information about the first category can be inferred from the absence of the other categories. This reduction in the number of columns helps to simplify the model and avoid multicollinearity, where the dummy variables are highly correlated with each other. Multicollinearity can cause issues in statistical models, leading to unstable estimates and inflated standard errors. Therefore, drop_first=True is a practical approach to handle categorical variables with multiple levels by ensuring that the model remains interpretable and the results are more reliable.

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

Ans- Temp variable has the highest corelation with the target variable

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

Ans –

- Normality of error terms: The error terms (residuals) should follow a normal distribution. This assumption ensures that the statistical tests and confidence intervals derived from the model are valid.
- Multicollinearity check: There should be no significant multicollinearity among the independent variables. Multicollinearity occurs when independent variables are highly correlated, which can lead to unstable estimates of regression coefficients.
- Linear relationship validation: There should be a linear relationship between the independent variables and the dependent variable. This assumption ensures that changes in the independent variables result in proportional changes in the dependent variable.

- Independence of residuals: The residuals should be independent of each other. There should be no autocorrelation or pattern in the residuals when plotted against time or other variables. Independence of residuals ensures that the errors are not influenced by the errors from previous observations.

These assumptions collectively help ensure that the linear regression model is appropriate for the data and that the estimates of the regression coefficients are unbiased and efficient. Therefore, it is essential to validate these assumptions before interpreting the results of a linear regression analysis.

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 - Demand of bikes depend on year, holiday, temp, windspeed, sept, Light_snowrain, Misty, spring, summer and winter.

But top 3 are – temp, winter and sept variables that contribute significantly on demand of shared bikes.

General Subjective Questions

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

Ans - Linear regression is a statistical method used to model the relationship between a dependent variable Y and one or more independent variables X. The goal is to fit a linear equation to the observed data that can predict the value of the dependent variable based on the independent variables.

Types of Linear Regression -

- Simple Linear Regression: Involves a single independent variable.
- Multiple Linear Regression: Involves two or more independent variables.

A linear regression line equation is written in the form of:

$Y = a + bX$

Linear regression assumes that the relationship between X and y is linear, residuals (errors) are normally distributed, residuals have constant variance (homoscedasticity), and there is little or no multicollinearity among predictors.

To evaluate this, we need to find-

- R-squared: Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Higher values indicate a better fit.
- Adjusted R-squared: Adjusts R-squared for the number of predictors in the model.
- Residual Analysis: Checking residuals helps assess model goodness-of-fit and adherence to assumptions.

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

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.), yet they exhibit markedly different properties when graphically analysed. Anscombe's quartet consists of four datasets, each containing 11 paired observations (x, y). Despite the datasets having identical summary statistics (mean, variance, correlation, regression line parameters), they are fundamentally different in their distributions and relationships.

Anscombe's quartet highlights the importance of visualizing data. It demonstrates that datasets with identical summary statistics can exhibit vastly different structures and relationships. Visual inspection of data through scatter plots, histograms, and other graphical methods can provide deeper insights that summary statistics alone may obscure. Despite their graphical differences, all datasets meet the assumptions of standard linear regression analysis when considering summary statistics. However, this overlooks the practical significance of outliers and the shape of the data. Anscombe's quartet is widely used in statistics education to emphasize. Anscombe's quartet challenges the assumption that datasets with similar statistical properties are necessarily similar in all respects. It underscores the importance of data visualization as a crucial step in understanding and interpreting data effectively.

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

Ans- Pearson's r, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables, typically denoted as X and Y.

Pearson's r quantifies the strength and direction of the linear relationship between two continuous variables.

Range: It ranges from -1 to +1. Pearson's r is widely used in statistics and data analysis to:

- Assess the strength and direction of relationships between variables.
- Determine whether there is a linear association between variables.
- Aid in understanding patterns in data and informing predictive models.

It is calculated by this formula – where X bar and Y bar are the means of X and Y, respectively, and n is the number of data points.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

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

Ans- Scaling in the context of data preprocessing refers to the process of transforming data into a specific range or distribution.

Purpose of Scaling:

- Normalization: Scaling is performed to bring all features of the data into a similar scale or range. This ensures that no single feature dominates due to its larger magnitude, which can lead to biased or inaccurate results in algorithms that rely on distances or weights
- Standardization: Scaling is also done to transform data into a standard normal distribution (mean = 0, standard deviation = 1). This can be useful for algorithms that assume normally distributed data, such as linear regression, logistic regression, and methods that use regularization.

Difference Between Normalized and Standardized Scaling:

- Normalization adjusts the values of numeric data to a common scale without distorting differences in the ranges of values or the original distribution shape. It is useful when the algorithm does not assume a normal distribution of the data.
- Standardization transforms the data to have a mean of 0 and a standard deviation of 1, making it suitable for algorithms that assume normally distributed data or benefit from the zero-mean, unit-variance properties (like many linear models).

6. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (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.

an infinite VIF occurs when there is perfect multicollinearity among predictor variables, where one variable can be perfectly predicted from others. This situation leads to R2 values of 1 in the VIF calculation, resulting in division by zero and hence an infinite VIF value. It highlights the severe issues multicollinearity can cause in regression analysis, making interpretation and inference challenging or impossible for affected variables.

7. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans - A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a set of data follows a certain theoretical distribution, such as the normal distribution. A Q-Q plot compares the quantiles of the dataset against the quantiles of a theoretical distribution (usually the normal distribution).

By plotting the observed quantiles against the expected quantiles from the theoretical distribution, the Q-Q plot helps visualize how closely the data approximate the theoretical distribution. Departures from the diagonal line in the Q-Q plot indicate deviations from the assumed distribution.

Importance in Linear Regression:

- Assumption Checking: In linear regression, it is often assumed that the residuals (errors) of the model are normally distributed. The Q-Q plot of residuals is crucial for verifying this assumption.
- Validity of Inference: If the residuals are normally distributed, it ensures the validity of statistical inference, such as hypothesis testing (e.g., for regression coefficients) and confidence interval estimation.
- Model Performance: Deviations from normality in residuals could indicate issues with the model, such as misspecification or the presence of outliers or influential points.

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 data sets 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 and Kolmogorov-Smirnov 2-sample tests.