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)

In our dataset, the dependent variable was cnt indicating booking count and categorical variables were season, year, weather situation, holiday, month, working day, and weekday. A boxplot was used to analyse the effect of these categorical variables on the dependent variable.

a. Season: Fall season seems to have attracted more booking count while spring season had the lowest value of cnt. And, in each season the booking count has increased drastically from 2018 to 2019.
b. Year: Booking increased significantly from year 2018 to 2019
c. Month: Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year. Number of booking for each month seems to have increased from 2018 to 2019.
d. Weather situation: Clear weather attracted more booking which seems obvious. And in comparison, to previous year, i.e 2018, booking increased for each weather situation in 2019.
e. Weekends and weekdays: Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
f. When its holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
g. Working day: Booking seemed to be almost equal either on working day or non-working day, signifying that there is not much effect on the dependent variable.

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

Ans 2. It is important to use **drop_first=True** during dummy variable creation to avoid the issue of multicollinearity. When dealing with categorical variables in regression or machine learning models, it's common practice to convert them into binary dummy variables. However, including dummy variables for all categories can lead to multicollinearity, where independent variables become highly correlated, affecting coefficient estimates and model interpretability.

For instance, consider a categorical variable "color" with options: red, green, and blue. Creating dummy variables for each color results in three columns. Using all three columns introduces redundancy, hindering model performance.

To mitigate multicollinearity, we drop one dummy variable, often the first category. This leaves us with n−1 columns, maintaining necessary information while avoiding collinearity issues. This practice enhances model stability and facilitates clearer interpretation of coefficients.

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

Ans 3. 'temp' variable has the highest correlation with the target variable 'cnt'

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

Ans 4. By validating the assumptions of linear regression after building the model on the training set:

- Normality: Using Q-Q plot or histogram, checking whether the residuals are normally distributed.
- Homoscedasticity: by plotting scatterplot between the dependent variable 'cnt' and residuals. No visible pattern observed from above plot for residuals.
- Independence of residuals: Verifying that the residuals are independent of each other and not exhibiting autocorrelation. This was checked using Durbin-Watson statistics.
- Linearity: Assessing whether the relationship between the independent variables and the dependent variable is linear. This was examined visually using scatterplots of the independent variables against the dependent variable.
- Multicollinearity – evaluated using VIFs and correlation matrices.

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 5. Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are: temp, year and Lightsnowrain.

General Subjective Questions

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

Ans 1. Linear regression is a foundational and widely used supervised learning algorithm for modeling the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the independent variables and the dependent variable. The goal of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the predicted and actual values of the dependent variable.

Assumptions:

- Linear relationship: It assumes that there is a linear relationship between the independent variables and the dependent variable.
- Independence: The observations (data points) are independent of each other.
- Homoscedasticity: The variance of the residuals (the differences between predicted and actual values) is constant across all levels of the independent variables.
- Normality: The residuals follow a normal distribution.
- No multicollinearity: The independent variables are not highly correlated with each other.

A linear regression line has an equation of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

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

Ans 2. Anscombe's quartet is a set of four datasets created by the statistician Francis Anscombe in 1973. Each dataset consists of eleven (x, y) points and has nearly identical simple descriptive statistics, such as mean, variance, correlation coefficient, and linear regression line. However, when plotted, these datasets reveal different patterns and relationships between the variables.

The purpose of Anscombe's quartet is to emphasize the importance of visualizing data and not relying solely on summary statistics. Despite having similar statistical properties, the datasets exhibit distinct patterns that can only be fully understood through graphical exploration.

The quartet highlights the limitations of relying solely on summary statistics and underscores the need for data visualization in exploratory data analysis. It serves as a cautionary example against making assumptions about data without visual inspection, as well as the importance of considering the context and underlying patterns in data analysis.

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

Ans 3. Pearson's correlation coefficient (often denoted as Pearson's r or simply r) is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the linear association between the variables. Pearson's r ranges from -1 to 1, where:

r = 1 indicates a perfect positive linear relationship: As one variable increases, the other variable also increases in a perfectly linear fashion.

r = -1 indicates a perfect negative linear relationship: As one variable increases, the other variable decreases in a perfectly linear fashion.

r = 0 indicates no linear relationship: There is no systematic linear relationship between the variables.

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

Ans 4. Scaling is a data pre-processing step which is applied to independent variables to normalize the data within a particular range. Scaling is performed to ensure that features are on a similar scale and have comparable magnitudes.

It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Difference between normalized and standardized scaling:

Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1.

MinMax Scaling: x = x-min(x)/max(x)-min(x)

Standardized Scaling: Standardized scaling replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

z = x-mean(x)/sd(x)

One disadvantage of normalized scaling over standardized scaling is that it loses some information in the data, especially about outliers.

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

Ans 5. When the VIF for a particular predictor variable is infinite, it typically means that there is perfect multicollinearity between that predictor variable and the other variables in the model. Perfect multicollinearity occurs when one or more of the predictor variables in the model are perfectly linearly related to each other, meaning that one predictor variable can be exactly predicted from the others.

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

Ans 6. A Q-Q plot, also referred as quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the dataset's empirical distribution to the quantiles of a theoretical distribution, typically the normal distribution. This comparison allows us to visually determine whether the data deviates from the assumed distribution.

Use: Q-Q plots are commonly used to check the assumption of normality in linear regression models. Linear regression models assume that the residuals (the differences between observed and predicted values) are normally distributed.

Importance: By plotting the residuals on a Q-Q plot against the quantiles of the normal distribution, we can visually assess whether the residuals follow a normal distribution.