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)

   On analysing categorical variable against target variable, I have got following findings:

   - For all categorical variable one observation is common, 2019 has more number of booking compared to previous year. This implies that there is good growth in business
   - Categorical variable named season have four values like spring, summer, fall, winter. Among this season fall has more bookings.
   - On analysing categorical variable "mnth" against dependent variable Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Booking increased starting of the year and then it started decreasing during the end of year.
   - Weather have 3 option like misty, clear and rain. During clear weather the booking is more. When the weather is rainy, the booking dropped to great extent. Misty weather condition got better booking than rainy but less than clear weather.
   - Comparing to weekdays, weekend have more bookings. Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
   - Booking seemed to be almost equal either on working day or non-working day.

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

   - By setting drop_first=True when creating dummy variables, you omit one of the dummy variables.
   - This approach avoids the multicollinearity problem by ensuring that only n-1 dummy variables are used, which are enough to capture the information about the categorical variable.
   - The coefficients of the remaining dummy variables represent the effect relative to the omitted category, making interpretation straightforward.

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

   - Temp has the highest correlation with target variable.
   - correlation between temp and cnt is 0.63

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

After building a model I have validated the assumption of linear regression using below criteria

- **Linear relationship** between X and y: X and Y should always display some sort of a linear relationship; otherwise, there will not be any use of fitting a linear model between them.
- **Normal distribution of error terms:** It represents the assumption of normality. Which exhibits that error terms generally follow a normal distribution with mean equal to zero in most cases.

- **Multicollinearity** denotes when independent variables in a linear regression equation are correlated. Multicollinear variables can negatively affect model predictions
- **Homoscedasticity** means the error is constant across the values of the dependent variable. The easiest way to check homoscedasticity is to make a scatterplot with the residuals against the dependent variable. There should be no visible pattern in residual values.
- **Independence of residuals:** The residuals (errors) of the observations are independent of each other. The value of the residual for one observation should not be influenced by the value of the residual for another observation. 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)

- Temp
- season_winter
- mnth_sept

General Subjective Questions

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

- Linear regression is a fundamental statistical and machine learning technique used to model the relationship between a dependent variable (or target) and one or more independent variables (or predictors).
- The goal is to predict the dependent variable based on the values of the independent variables.

Types of linear regression

- Simple Linear Regression: This involves a single independent variable. The model assumes a linear relationship between the dependent variable and the independent variable.
- Multiple Linear Regression: This involves two or more independent variables. The model extends the simple linear regression to capture relationships between the dependent variable and multiple predictors.

Mathematical formula

- **Simple Linear Regression:** In simple linear regression, the relationship between the dependent variable y and the independent variable x is expressed as:

$$y=\beta0+\beta1x+\epsilon$$

1. β0 is the intercept of the regression line (the value of y
2. β1 is the slope of the line (the change in y for a one-unit change in x).
3. ϵ is the error term (the difference between the observed value and the value predicted by the model).

- **Multiple linear regression:** In multiple linear regression, the relationship is extended to multiple predictors

y=β0+β1x1+β2x2+⋯+βnxn+ϵ

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

- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

Dataset I:

- **Description:** This dataset represents a classic linear relationship. When plotted, it shows a clear linear trend with a positive slope.

Dataset II:

- **Description:** This dataset also represents a linear relationship, but with a slight curve and an outlier.The data generally follows a linear trend, but there's an obvious outlier that affects the relationship.

Dataset III:

- **Description:** This dataset shows a perfect quadratic relationship. The data forms a perfect parabola, demonstrating a non-linear relationship.

Dataset IV:

- **Description:** This dataset has a linear relationship but includes an outlier with a high leverage point.The data shows a linear trend, but a single extreme outlier has a significant impact on the slope.

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

- Pearson's R is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by 'r'.
- it has a value between +1 and −1, where
- 1 is total positive linear correlation,
- 0 is no linear correlation,
- −1 is total negative linear correlation.

Mathematically, Pearson's correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations.

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

Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.

The two most discussed scaling methods are Normalization and Standardization.

- Normalization typically scales the values into a range of [0,1].
- Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Formula of Normalized scaling:

$x= (x-\min(x)) / (\max(x)- \min(x))$

Formula of Standardized scaling:

$x= (x-\mathrm{mean}(x)) / (\mathrm{sd} (x)$

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

The value of VIF is calculated by the below formula:

$Vifi= 1/(1- Ri^2)$

Where, 'i' refers to the ith variable.

- If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.
- 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.

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

- This q-q or quantile-quantile is a graphical technique which helps us validate the assumption of normal distribution in a data set.
- Using this plot we can infer if the data comes from a normal distribution.
- If yes, the plot would show fairly straight line. The straight lines shows the absence of normality in the errors.

Uses of Q-Q Plots:

1. **Normality Testing**: To visually assess if a dataset is approximately normally distributed. Points should roughly follow a straight line if the data is normally distributed.
2. **Checking Other Distributions**: Q-Q plots can be used to assess other distributions (e.g., exponential, log-normal) by comparing the quantiles of the data to the quantiles of those distributions.
3. **Identifying Outliers**: Q-Q plots can help detect outliers and deviations from the expected distribution.

In a Q-Q plot comparing data to a normal distribution:

- **X-Axis**: Quantiles from the theoretical normal distribution.

- **Y-Axis**: Quantiles from your data.

If the data is normally distributed, the plot will show points that approximately lie on a 45-degree line.

If the data is not normally distributed, the points will deviate from this line in a manner that can reveal the nature of the deviation (e.g., heavy tails, skewness).