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 -

I have done the visualization on the categorical columns using boxplots and bar plots. Here's a summary of the key points inferred from the visualizations:

- The fall season appears to have attracted more booking and within each season, the booking count has shown a significant increase from 2018 to 2019.
- The majority of booking occurred during the months of May, June, July, August, September and October. The trend increased from the beginning of the year until the middle of the year and then started decreasing towards the end of the year.
- Clear weather is associated with more bookings which is expected.
- Thursday, Friday, Saturday and Sunday have a higher number of bookings compared to the start of the week.
- When it's not a holiday, the number of the bookings seems to be less, which is reasonable as people may prefer to spend time at home and enjoy with the family during holiday.
- In 2019, there was an increase in number of bookings as compared to the previous year, indicating positive progress in terms of business.

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

Answer -

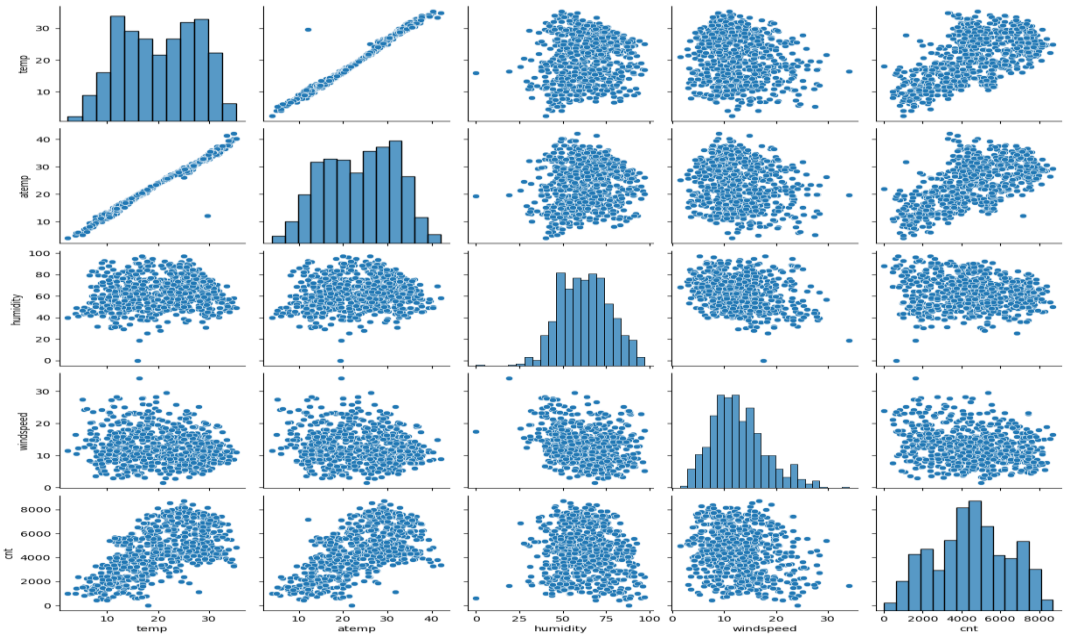
drop_first=True helps avoid multi-collinearity issues, simplifies the interpretation of coefficients, and improves the efficiency of the regression model.

Using drop_first=True is crucial as it minimizes the creation of an additional column during dummy variable creation, thereby mitigating correlations among dummy variables.

The syntax for drop_first is a boolean, defaulting to False, indicating whether to generate k-1 dummies out of k categorical levels by excluding the first level.

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

Answer -



The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.

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

ANSWER-

I have validated the assumption of Linear Regression Model based on Normality of error terms, Multicollinearity check, Linear relationship validation, Homoscedasticity and Independence of residuals.

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

ANSWER -

The top 3 features contributing significantly towards explaining the demand of the shared bikes as 'temp', 'winter', 'sep'.

General Subjective Questions

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

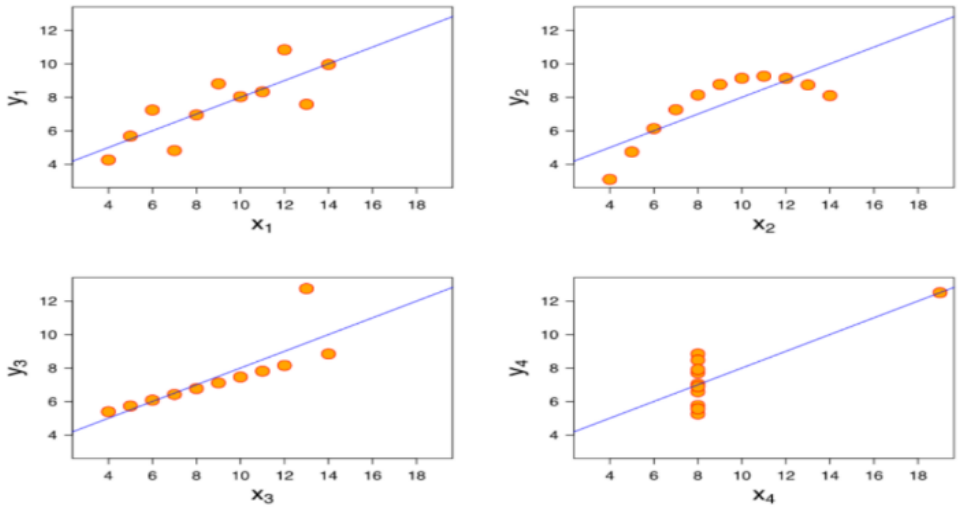
ANSWER -

Linear regression is a predictive modeling technique that reveals the relationship between the dependent (target variable) and independent variables (predictors). It establishes a linear relationship, indicating how the value of the dependent variable changes in accordance with the value of the independent variable. When there is a single input variable (x), it is termed simple linear regression. In cases with more than one input variable, it is referred to as multiple linear regression. The linear regression model produces a sloped straight line depicting the relationship between the variables, which can be either a positive or negative linear relationship. The objective of the linear regression algorithm is to determine the optimal values for a0 and a1, aiming to achieve the best-fit line with minimal error. RFE, Mean Squared Error (MSE), or cost function are employed in linear regression to determine the most suitable values for a0 and a1, resulting in the best-fit line for the data points..

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

ANSWER-

Anscombe's Quartet consists of four datasets that exhibit nearly identical simple descriptive statistics, including mean, variance, and other statistical measures. Despite their similar statistical characteristics, these datasets differ significantly in their distributions and appearances when visualized on scatter plots. The quartet was intentionally created to underscore the importance of graphically exploring data before conducting analysis or building models. It emphasizes how relying solely on summary statistics can be misleading and how graphical exploration is essential to understanding the true nature of the data. Each of the four datasets within Anscombe's Quartet provides the same statistical information, such as mean and variance, for both x and y variables, yet their visual representations reveal distinct patterns and relationships.



- ✓ 1 st data set fits linear regression model as it seems to be linear relationship between X and y
- ✓ 2 nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.

- ✓ 3 rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
- ✓ 4 th data set has a high leverage point means it produces a high correlation coeff.

Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model.

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

ANSWER-

Pearson's correlation coefficient, also known as Pearson's R, is a measure that tells us how two things are related. It gives a number between -1 and 1:

- If it's close to 1, it means they go up and down together.
- If it's close to -1, it means as one goes up, the other goes down.
- If it's around 0, there's not much of a relationship.
- It helps us see if there's a pattern between two sets of numbers. Keep in mind that it assumes the relationship is a straight line and doesn't work well if things are more complicated.

The formula for Pearsons correlation coefficient (r) between variables X and Y is given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

ANSWER-

Scaling is a way of adjusting your data so that it fits into a specific range. It's a step in preparing data for analysis, making calculations faster for algorithms. When your data has features that vary in size and range, scaling is important. If you don't scale, the algorithm might give too much importance to large values and ignore other factors, leading to inaccurate results.

Here are some differences between Normalizing Scaling and Standardize Scaling:

- ➢ In normalized scaling, we use the minimum and maximum values of features however in standardize scaling, we use the mean and standard deviation for scaling.

- Normalized scaling is suitable when features have different scales but Standardized scaling is used to make sure the data has a mean of zero and a standard deviation of one.

- Normalized scaling scales values between (0,1) or (-1,1) however Standardized scaling doesn't have a specific range.

- Normalized scaling is influenced by outliers but Standardized scaling is not affected by outliers.

- Normalized scaling is used when we're not sure about the distribution of the data moreover Standardized scaling is used when the distribution is normal.

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

ANSWER -

When there's a perfect correlation between two variables, the Variance Inflation Factor (VIF) becomes infinity. VIF measures how much the variance of a model coefficient is inflated due to multicollinearity. A high VIF suggests correlation between variables. For example, if the VIF is 4, it means the variance is inflated by a factor of 4 because of multicollinearity.

If the VIF is infinite, it indicates a perfect correlation between two independent variables. In cases of perfect correlation, the R-squared (R2) value becomes 1, leading to a division by zero issue (1/(1-R2) equals infinity). To address this, it's necessary to remove one of the variables causing this perfect multicollinearity from the dataset. Dropping one of the correlated variables resolves the issue.

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

A Q-Q plot, or quantile-quantile plot, is a graphical tool used in statistics to check if a dataset follows a certain theoretical distribution, like a normal distribution. It compares the quantiles of the dataset with the expected quantiles from the theoretical distribution.

In simple terms, a Q-Q plot helps us see if our data is close to a normal distribution. In linear regression, it's important because assuming normality is a key assumption. If the points in the Q-Q plot roughly form a straight line, it suggests that our data is approximately normally distributed, which is good for making accurate predictions with linear regression. If the points deviate from the line, it may indicate a departure from normality, prompting further investigation.