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)
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

General Subjective Questions

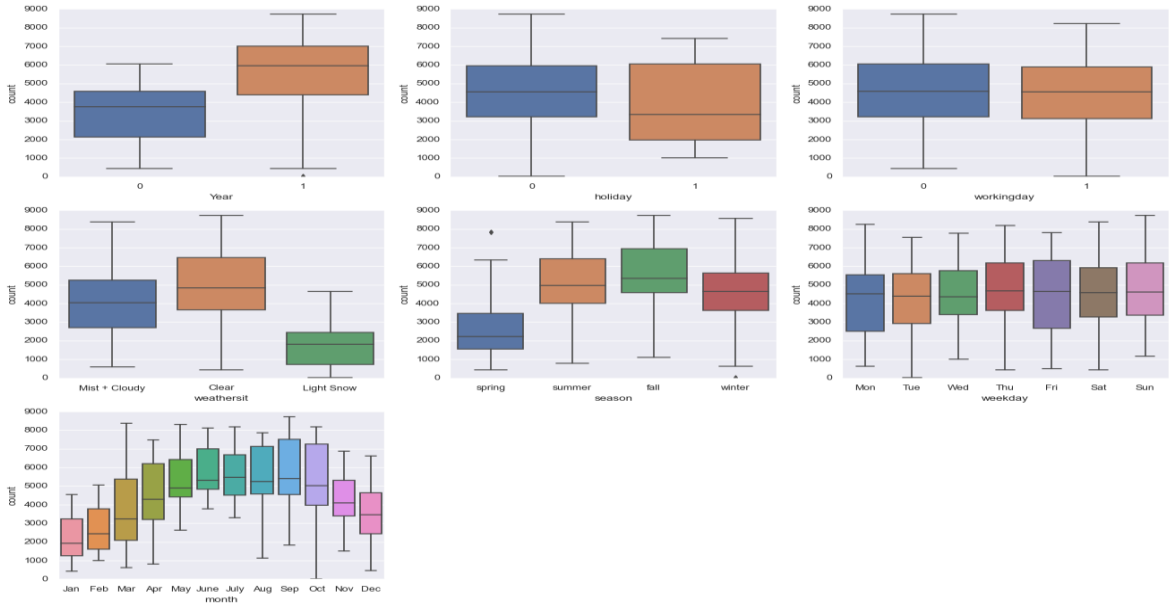
1. Explain the linear regression algorithm in detail. (4 marks)
2. Explain the Anscombe's quartet in detail. (3 marks)
3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have 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)

Answers

- **Assignment-based Subjective Questions**

ANS 1.

In the dataset, we have seven categorical variables: Season, Weather, Holiday, Working-day, Month, Weekday and Year. These variables were visualized using a Box-plot. These variables influenced our dependent variable as follows:



❖ ***Season VS Count:***

- *The majority of the Bike rental count is in the fall season, followed by winter season and summer.*

❖ ***Weather VS Count:***

- *We see that the Bike rental count is high when the day is clear when compared to others.*
- *On days with light snow the Bike rental count is very less.*

❖ ***Holiday VS Count:***

- *We observe that the Bike rental count is significantly more when there's a holiday as compared to the not holiday.*

❖ ***Working-day VS Count:***

- *The Bike rental count is comparatively high on non-workingdays when compared to that of the workingday.*

❖ ***Month VS Count:***

- *We observe a high peak in the Bike rental count for the month of September followed by March and May.*

❖ ***Weekday VS Count:***

- *We observe more Bike rental count on Saturdays and Sundays.*
- *While on the rest of the days it's almost similar*

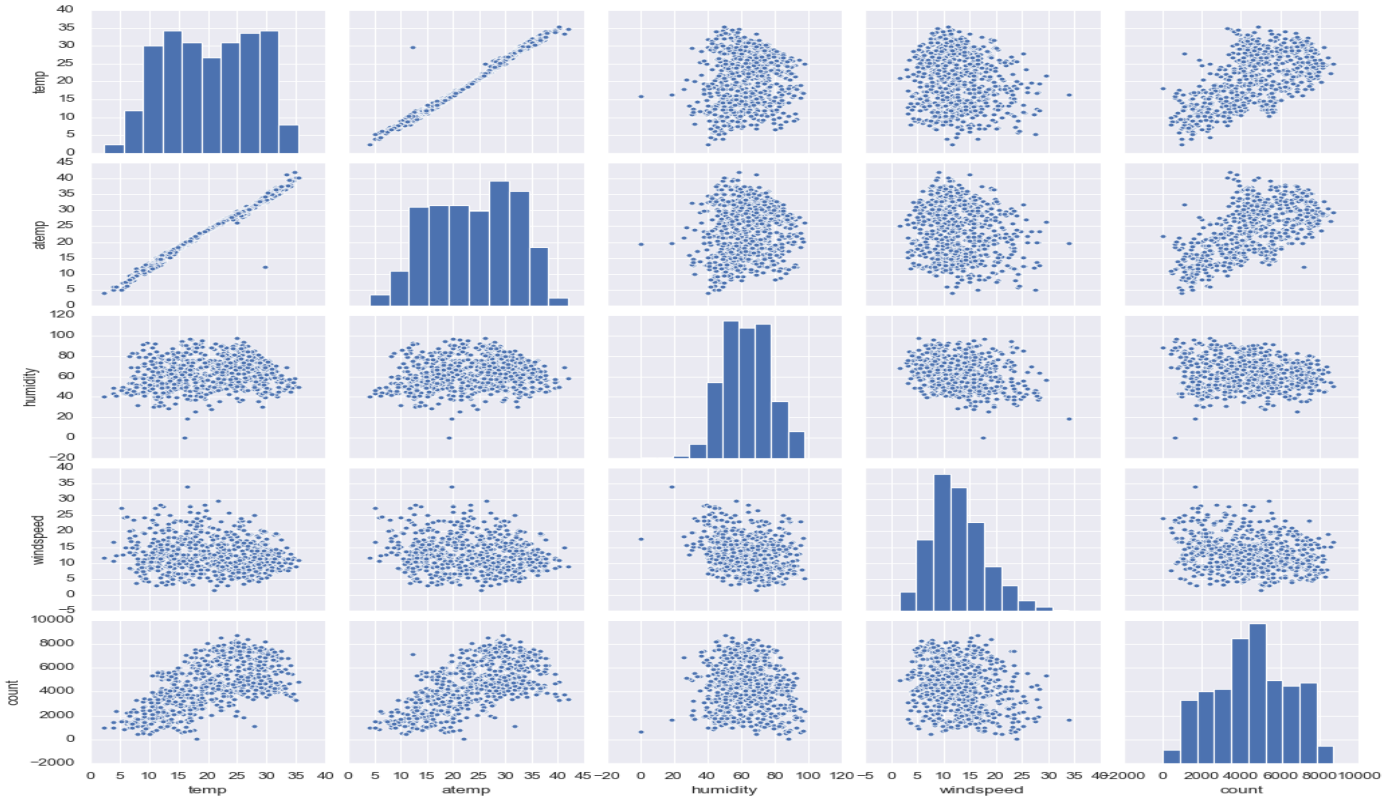
❖ ***Year VS Count:***

- *We observe a significant increase in Bike rental count for the year of 2019 when compared with the year 2018.*
- *Highest Bike rental count for the year of 2018 is between 6000 and 7000 while the Bike rental count for the year 2019 is between 8000 and 9000.*

ANS 2.

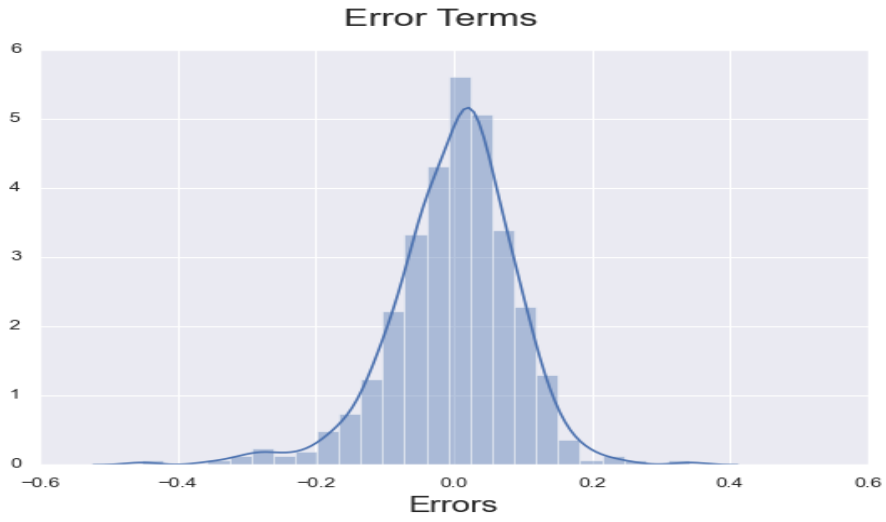
If we don't drop the first column, then our dummy variables will be correlated (redundant). Using "drop_first = True" helps reduce the number of extra columns created when creating a dummy variable. As a result, it will reduce the correlations created among dummy variables. For instance, we wish to create a dummy variable for Categorical columns for three types of values: Furnished, Semi-Furnished, and Unfurnished. Unfurnished is evident if one variable is neither furnished nor semi-furnished. Hence, we do not need a third variable to identify unfurnished housing.

ANS 3.



Temperature has the highest correlation with count, we can observe as the temperature rises the Bike rental count increases.

ANS 4.



The residuals distribution should follow a normal distribution and should be centred around 0(mean = 0). By plotting a distplot of residuals, we can verify if the residuals follow a normal distribution or not. The graph demonstrates that the residuals are distributed around mean 0.

ANS 5.

Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- *Temperature [co-efficient = 0.4391]*
- *Year [co-efficient = 0.2349]*
- *Weathersit_LightSnow [co-efficient = - 0.2949]*

- **General Subjective Questions**

ANS 1.

Linear regression is a type of supervised machine learning algorithm which is used in the prediction of numerical values. Linear regression is the most fundamental form of regression. Regression is the most widely used for predictive analytical model.

Linear regression is based on the popular equation "y = mx + c".

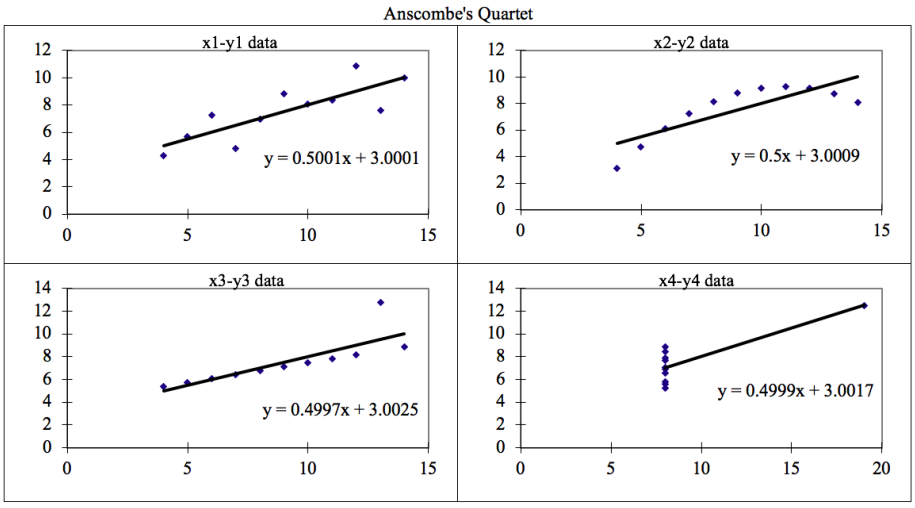
It assumes that a linear relation exists between the dependent variable (y) and the predictor(s)/independent variable(x). In the regression, we calculate the best fit line that describes the relation between the independent variable and the dependent variable. The regression method tries to find the best fit line that shows the relationship between the dependent variable and predictors having the least error.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. ***Simple Linear Regression: SLR** is used when the dependent variable is predicted using only one independent variable.*
2. ***Multiple Linear Regression: MLR** is used when the dependent variable is predicted using multiple independent variables.*

ANS 2.

Anscombe's Quartet is a collection of four data sets that are almost identical in terms of simple descriptive statistics, but have some irregularities in the dataset that trick the regression model if built. When plotted on scatter plots, they have extremely different distributions and appear differently. Francis Anscombe, a statistician, built it in 1973 to demonstrate the significance of plotting graphs before analysing and modelling, as well as the impact of other data on statistical features. There are four data set plots with virtually identical statistical observations and statistical information, including variance and mean of all x, y points in all four datasets.



The four datasets are as follows:

<u>Dataset 1</u>: this is a good fit for the linear regression model.

<u>Dataset 2</u>: due to the non-linear nature of the data, a linear regression model could not be fit adequately.

<u>Dataset 3</u>: depicts the outliers in the dataset that the linear regression model cannot manage.

<u>Dataset 4</u>: depicts the outliers in the dataset that the linear regression model cannot manage.

ANS 3.

The Pearson's r is also known as Pearson's Correlation Coefficient, or bivariate correlation, in statistics. It's a metric for determining the linear relationship between two variables. It has a numerical value between -1.0 and +1.0, just like all correlations. Pearson's correlation coefficient is commonly used when discussing correlation in statistics. It, on the other hand, is incapable of capturing nonlinear interactions between two variables or distinguishing between dependent and independent variables.

Pearson's correlation coefficient is the product of the two variables' standard deviations divided by their covariance. The name comes from the fact that the definition includes a "product moment," or the mean (the first moment near the origin) of the product of the mean-adjusted random variables.

1. r = 1 means the data is perfectly linear with a positive slope.
2. r = -1 means the data is perfectly linear with a negative slope.
3. r = 0 means there is no linear association.

ANS 4.

It is a data pre-processing technique that is applied to independent variables in order to normalise the data within a particular range. It also aids in the acceleration of algorithmic calculations.

The majority of the time, the acquired data set contains features with a wide range of magnitudes, units, and ranges. If scaling is not done, the algorithm will only consider magnitude rather than units, resulting in erroneous modelling. To solve this problem, we must scale all of the variables to the same magnitude level. Scaling only changes the coefficients and not the other parameters such as the t-statistic, F-statistic, p-values, R-squared, and so on.

Normalization is the process of rescaling values into a range of [0, 1]. Typically, standardisation entails rescaling data to a mean of 0 and a standard deviation of 1. (Unit variance). Normalization has the disadvantage of losing some data information, particularly about outliers, when compared to standardisation.

ANS 5.

The variance inflation factor (VIF) measures the degree of correlation between one predictor and the rest of the model's predictors. It is used to detect collinearity and multicollinearity. Higher values indicate that assessing the contribution of predictors to a model properly is difficult to impossible.

If there is perfect correlation, then VIF = infinity. This demonstrates that two independent variables have a perfect correlation. We get R2 =1 in the event of perfect correlation, which leads to 1/(1-R2) infinite. We need to remove one of the factors from the dataset that is creating this perfect multicollinearity to address this problem.

ANS 6.

Two quantiles are shown against each other in Q-Q plots (Quantile-Quantile charts). A quantile is a percentage of the population in which specific values fall below it. The median, for example, is a quantile where 50% of the data falls below it and 50% of the data falls above it. Q-Q plots are used to determine whether two sets of data are from the same distribution. On the Q-Q plot, a 45 degree angle is drawn; if the two data sets are from the same distribution, the dots will fall on that reference line.

A few advantages:

- *a) It may be used with sample sizes of any size;*
- *b) It can identify several distributional features such as shifts in location, shifts in scale, changes in symmetry, and the existence of outliers.*

If two data sets are compared, it is used to examine the following scenarios:

- *i. Originate from populations having a similar distribution.*
- *ii. Have a same scale and location*
- *iii. Have distributional forms that are comparable*
- *iv. Having a tail that behaves similarly*