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?

Ans: Of the categorical variable, year seems to have maximum influence on the user count. The balance of the categorical variable were data regarding date, season and weather situation. These data seemed to be corelated to each other too.

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

Ans 2. When dummy variables are created using pd. get_dummies(), it assigns 1 to indicate that the corresponding particular value of variable is represented in that particular column. If the gender column has male, female and others, then the representation will be as follows

Male	Female	Others	Gender Represented
1	0	0	Male
0	1	0	Female
0	0	1	Others

From the above it is clear that the male representation can be assigned even when only female and others column is represented. Thus removing the male column reduces the multicollinearity between eh gender columns as well as reduce the overall number of variables. Hence its important to use drop_first.

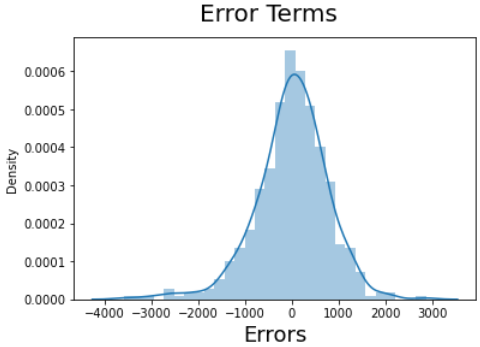
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark).

Ans 3. Temperature

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

Ans 4. After the model was built, the error terms were plotted to see two aspects.

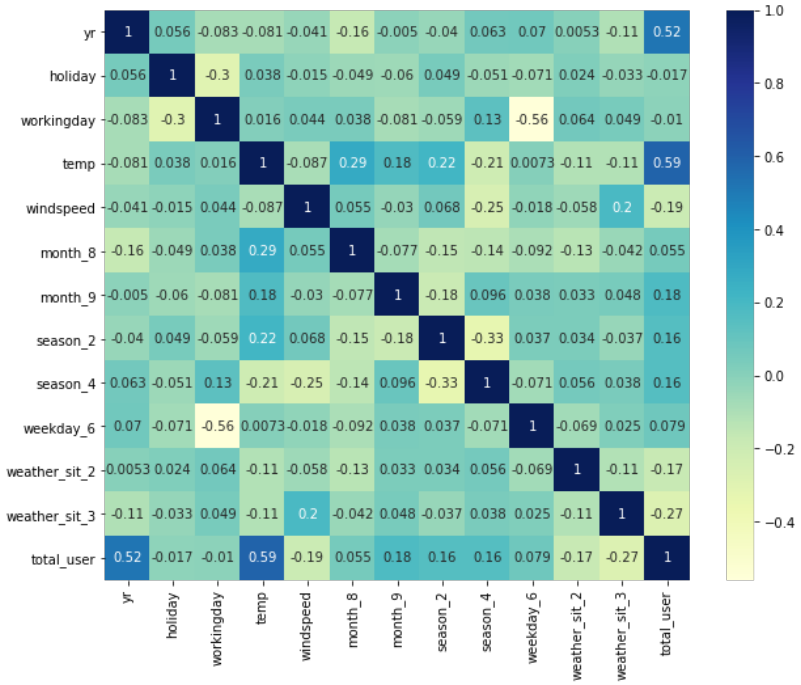
First: The mean is close to zero
Second: That the values lie between 1 standard Deviation.



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

Ans 5. The top three contribution is done by the following features

1. Temperature : Positive Correlation of 0.59
2. Year: Positive correlation of 0.52
3. Weather Sit 3, (Light snow): Negative correlation of 0.27



General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans1: Linear regression algorithm is the equation for line expressed as

$y = b_0 + b_i X_1 + b_2 X_2 \ldots\ldots b_n Xn$

here $\mathbf{b_0}$ stands for the constant and $\mathbf{b_{1\text{-}n}}$ stands for the coefficient of each variable independent variable represented as X and n is the number of variables. Bo is the intercept and b1 is the slope.

2. Explain the Anscombe's quartet in detail.

Ans3: Anscombe's quartet was constructed by Francis Anscombe in the year 1973. He used four nearly identical datasets but with different distributions that appears very different when graphically represented. This demonstrates the effect of outliers and influential observations.

3. What is Pearson's R?

Ans 3. Pearson's R or Pearson's Coefficient is the measure of linear correlation between two sets of data. It is the normalised measurement of the covariance, that the measurement falls always between +1 and -1.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Ans 4. Scaling is the process of standardising numerical values in a dataset. Scaling is performed to negate the impact of the measuring units of the values. For eg, the impact of 10 years of age and 1000 dollars in wages will be different on the target variable. Hence there is a need to scale the value such that impact is calculated correctly. The two types of scaling are

Min Max Scaling : This is also called normalised scaling. In normalised scaling the value will always lie between 0 and 1. This will also take care of any outliers in the dataset.

Standard Scaling: This is when the values are centred around mean with unit standard deviation. In this case the values are not restricted to a scale like in normalisation. Outliers wont be affected by standardization.

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

Ans 5. VIF if infinite when the correlation between two variable is perfect. This means that there exist multicollinearity between independent variables if VIF is infinite. This also means that the corresponding variable can be expressed as a combination of other variables.

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

Ans6: Q-Q plot is a plot used to compare two probabilistic distributions. This is done by plotting their quantile against each other. It can be used to compare the shape of the distribution. This is more powerful than using histograms for comparison. This graphically represents "goodness of fit" of the line.