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)**

   Season – The numbers of users are less during the spring season as compared to others. The bike users are highest in the fall season followed by summer and winter.

   Year - The number of users has certainly increased from 2018 to 2019.

   Month - The trend shows that in the first quarter the numbers of bikes rented is quite low and slowly start rising in the next quarter. It peaks around the September in the autumn season and then gradually slows in the later part of year.

   Holiday – The median numbers of rented bikes are higher on non-holidays rather than the holidays. It looks like, holidays is quite insignificant to numbers.

   Working Day – The weekdend and weekdays are almost following the same trend with a similar median number.

   Weathers – The boxplot chart for the spring season shows that the numbers are highest for the bikes rented. The numbers are very low in the rainy season which is quite natural as less people go out for travelling.

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

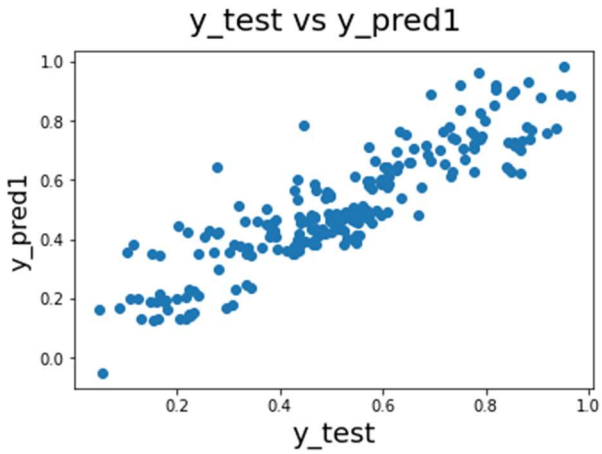
   It is used because it helps in reducing the number of the columns, thus reducing the dummy variable correlations.

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

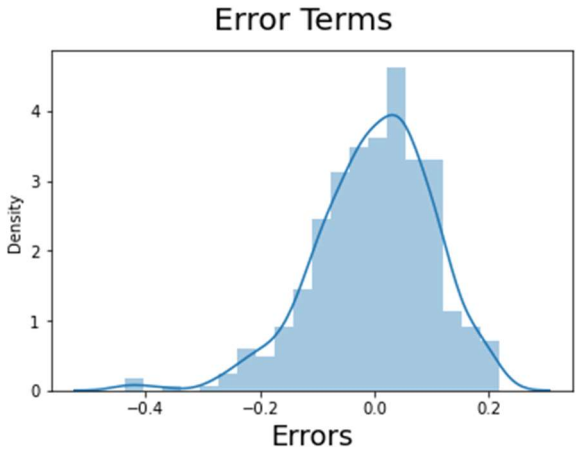
   The count of the registered users is the most correlated to the target variable.

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

   - **Homoscedasticity** – The plot between the residual and fitted values shows the a linear relationship



- **VIF** for the independent variables shows no multicollinearity.
- The error terms graph follow a normal distribution curve with a mean of zero.



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

   3 major features contributing significantly towards demand of shared bikes are

   Year, casual users of rental bikes are contributing positively & Spring is affecting negatively.

   cnt = 35.431 + 22.477 * yr + 13.622 * casual - 17.435 * Spring - 6.360 * Cloudy - 8.474 *Rain

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Linear regression model is the supervised learning method. Regression is used for the predicting the value of target variable on basis of the independent variables. It helps in finding out the relationship between the target variable and independent variable.

   The linear regression helps in determining the which are independent variables which are significant in predicting the target variable.

   The general formula for linear regression is as below: -

   $y = c + ax_1 + bx_2 + \ldots + gx_n$

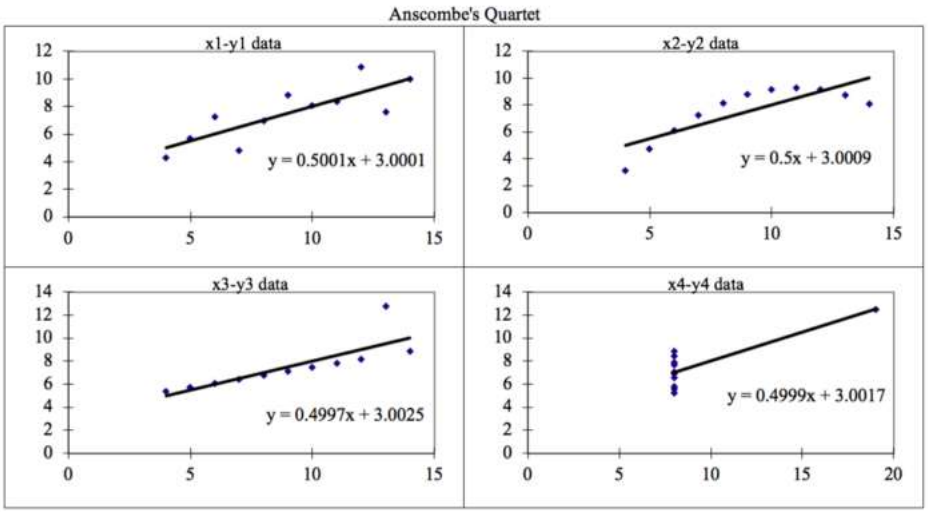
   y is the dependent or the target variable which we want to predict
   c is the intercept
   x_1, x_2, $\ldots x_n$ are the independent variable or the features impacting the target variables

   Linear regression used the least-squares method to fit a line to the data. It is a method used to predict the best fitted line explaining the data.

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's quartet is phenomena which was observed by a statistician Francis Anscombe, in 1973. This emphasises on the importance of visualizing the data before analysing the model building, and the effect of other observation on other statistical properties.

   It a group of four data, which are identical from descriptive statistics point of view, however when plotted has different distributions and appear differently when plotted as graph.



3. **What is Pearson's R?**

   Pearson R is the correlation coefficient which indicates the how two variables are changing with respect to each other. The Pearson coefficient value varies between -1 to 1.

   A value of 1 indicate the perfect correlation and two variables following a positive linear relationship.

   A value of -1 indicates negative linear relationship.

   A value to 0 indicate no relationship between two variables.

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

   Scaling is method used in the data preparation phase. In the pre-processing of the date, we come across set of independent variables which are highly varying in nature due to different magnitudes, units and range.

   If the scaling of the data is not done the algorithm might consider only the magnitudes into account leading to incorrect modelling. In order to overcome such issues, we scale the data.

   There are two types of scaling that is done.

   1. Normalized or Min- Max Scaling – In the process, it brings all the variables in a range on 0 and 1.

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

   2. Standardization: It replaces the values by their Z scores. It brings the data into standard normal distribution which has mean zero and standard deviation 1.

      Standardization: x = x – mean(x)/sd(x)

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

   The VIF infinite means that there are some independent variables which are perfectly correlated.

   The formula for VIF is $1/1\text{-}R^2$. In case of perfect correlation, we get $R^2 = 1$, which will make the denominator 0, making VIF infinity.

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

   QQ plots helps in determining if the two datasets come from population with a common distribution.

   It is a scatter plot which is created by plotting two sets of quantiles against one another. If the two data sets come from same distribution, it will form a straight line.