Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

As per my observation, the below are the categorical variables:

- 'mnth' with values Jan,Feb,Mar,Apr,May,Jun,Jul,Aug,Sep,Oct,Nov,Dec
- 'season' (1:spring, 2:summer, 3:fall, 4:winter)
- 'Weathersit'
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- 'Yr' : year (0: 2018, 1:2019)

Here is my analysis:

- spring, Mist and Cloudy affects the Business negatively as there is very less demand.
- Demand of Bike Rent has been significantly increased in the 2019 than 2018
- Especially during May till October , the demand is high with peak demand during June till September
- The Demand of Bikes is more in the Winter and Summer season, mostly user don't like to travel using Bikes in Rainy Day or Rainy Season.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

The drop_first = True is used while creating dummy variables to drop the base/reference category. The reason for this is to avoid the multi-collinearity getting added into the model if all dummy variables are included. The reference category can be easily deduced where 0 is present in a single row for all the other dummy variables of a particular category.

Another reason is to reduce the dimension/size of the data.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

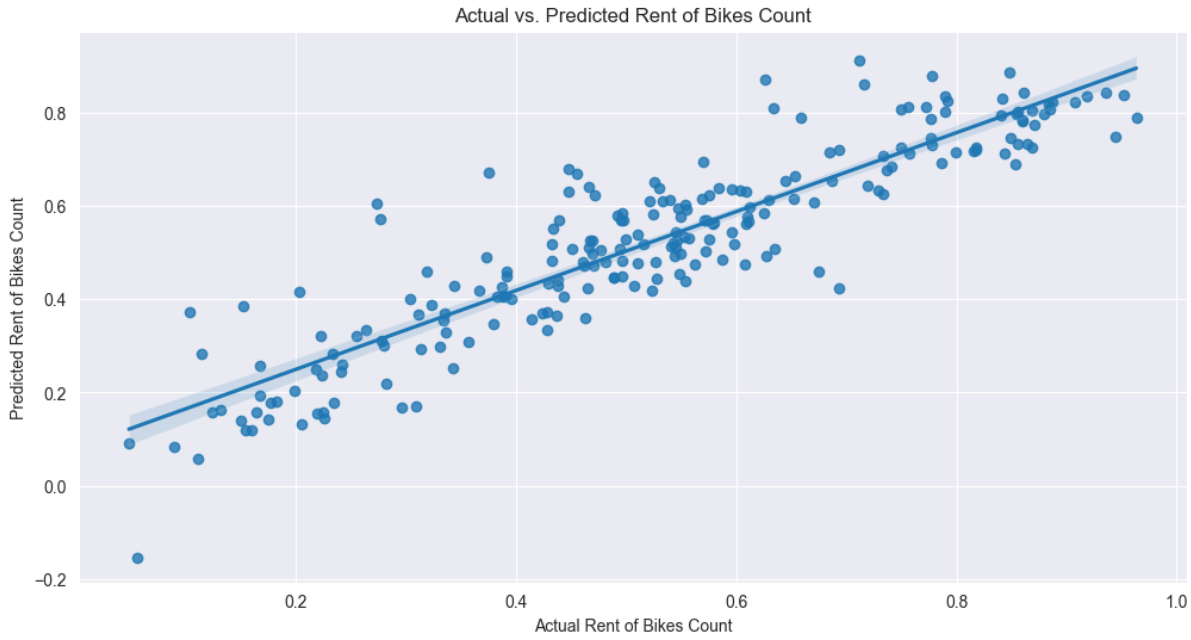
- "temp" is the variable which has the highest correlation with target variable i.e. 0.63.
- The casual and registered variables are actually part of the target variable as values of these columns sum up to get the target variable, hence ignoring the correlation of these 2 variables.
- "atemp" is the derived parameter from temp, humidity and windspeed, hence not considering it as it is eliminated in the model preparation.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

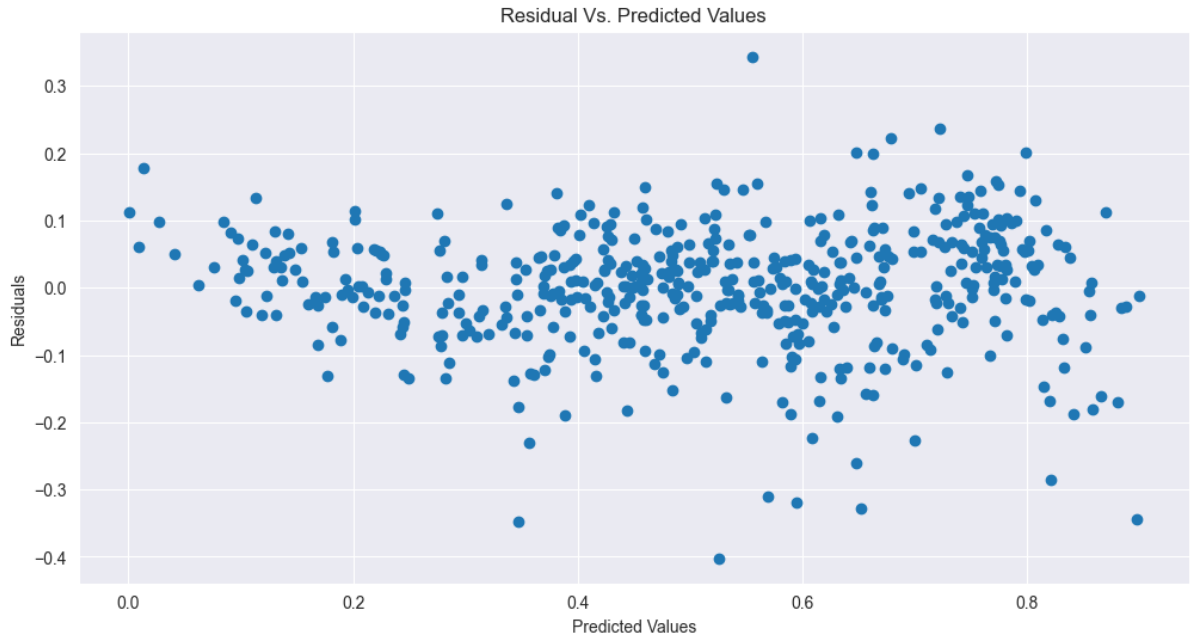
Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

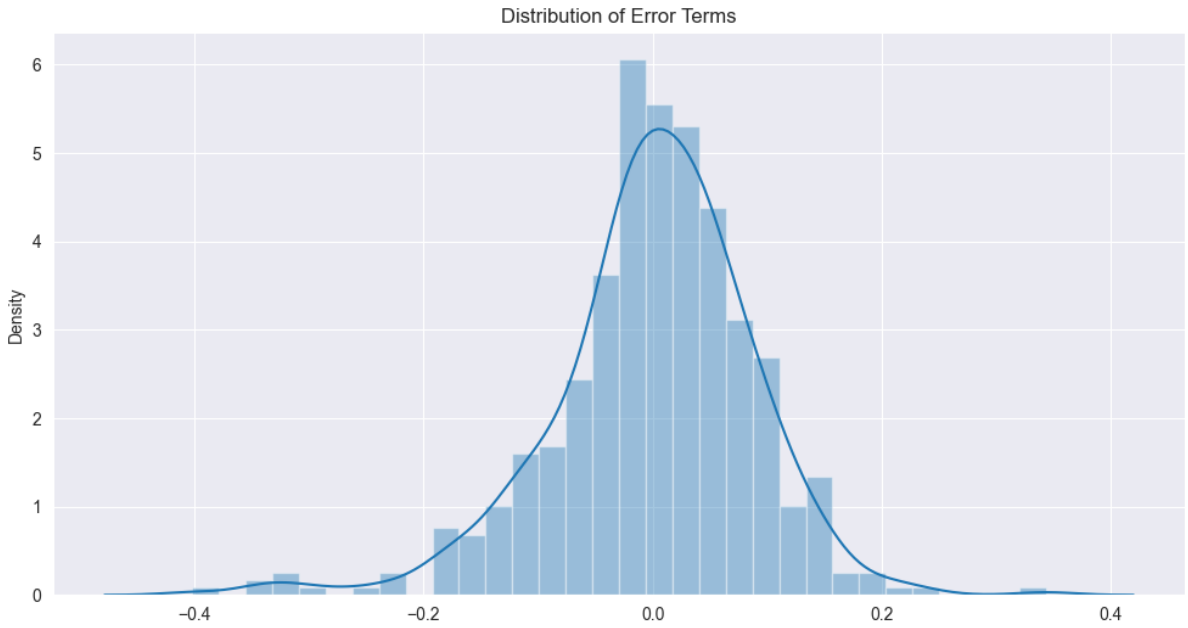
1. **Linear relationship between independent and dependent variables** – The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the below figure.



2. **Error terms are independent of each other** – We can see there is no specific Pattern observed in the Error Terms with respect to Prediction, hence we can say Error terms are independent of each other



3. **Error terms are normally distributed:** Histogram and distribution plot helps to understand the normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The top 3 variables are:

Temperature is the Most Significant Feature which affects the Business positively, Whereas the other Environmental condition such as Humidity, Windspeed and Cloudy affects the Business negatively.

'Yr':

The growth year on year seems organic given the geological attributes.

'season':

Winter season is playing the crucial role in the demand of shared bikes.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

- Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.

- The algorithm uses the best fitting line to map the association between independent variables with dependent variable.

- There are 2 types of linear regression algorithms

- ✓ Simple Linear Regression – Single independent variable is used.
  - $Y = \beta_0 + \beta_1 X_1$ is the line equation for SLR.

- ✓ Multiple Linear Regression – Multiple independent variables are used.
  - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$ is the line equation for MLR.
  - $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
  - $\beta 1, \beta 2, \ldots, \beta p = Slope\ or\ the\ gradient.$

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

- Statistics like variance and standard deviation are usually considered good enough parameters to understand the variation of some data without actually looking at every data point. The statistics are great to for describing the general trends and aspects of the data.
- Francis Anscombe realized in 1973 that only statistical measures are not good enough to depict the data sets. He created several data sets all with several identical statistical properties to illustrate the fact.
- Anscombe's Quartet signifies that multiple data sets with many similar statistical properties could still be different from one another when plotted.
- The dangers of outliers in data sets are warned by the quartet. Check the bottom 2 graphs. If those outliers would have not been there the descriptive stats would have been completely different in that case.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearon's R returns values between -1 and 1. The interpretation of the coefficients are:

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

What - The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.

Why – Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results in to the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance. The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.

- Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

$$\boldsymbol{MinMax\ Scaling}: \boldsymbol{x} = \frac{\boldsymbol{x} - \boldsymbol{min}(\boldsymbol{x})}{\boldsymbol{max}(\boldsymbol{x}) - \boldsymbol{min}(\boldsymbol{x})}$$

- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

$$\boldsymbol{Standardization}: \boldsymbol{x} = \frac{\boldsymbol{x} - \boldsymbol{mean}(\boldsymbol{x})}{\boldsymbol{sd}(\boldsymbol{x})}$$

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

The VIF formula clearly signifies when the VIF will be infinite. If the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below

Interpretations:

- Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
- Y values < X values: If y-values quantiles are lower than x-values quantiles.
- X values < Y values: If x-values quantiles are lower than y-values quantiles.
- Different distributions – If all the data points are lying away from the straight line.

Advantages:

- Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
- The plot has a provision to mention the sample size as well.