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)**

   Ans.

   -Bike demand is highest in fall.

   -Bike demand takes a dip in Spring

   -Bike demand in year 2019 is higher as compared to 2018.

   -Bike demand is high in the months from May to October

   -Bike demand is high if weather is clear or with misty clouds, while the demand is low when where there is light rain or snow.

   -Bike demand is almost similar throughout the weekdays.

   -Bike demand does not change whether the day is working day or not.

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

   Ans.

   - It is important to achieve k-1 dummy variables as it can be used to delete extra columns while creating dummy variables.
   - Also, using **drop_first=True,** significantly helps in reducing multicollinearity between dummy variables.

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

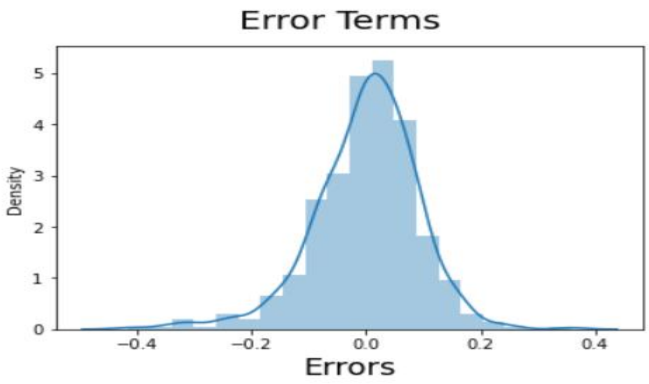
   Ans.

   - atemp and temp both have same correlation with target variable of 0.63 which is indeed the highest among all numerical values.

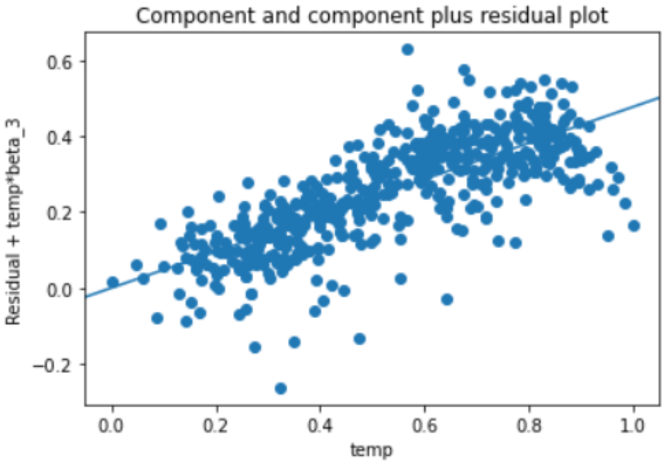
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

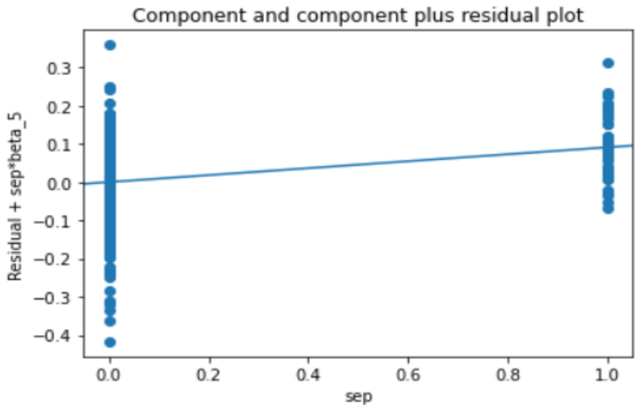
   Ans.

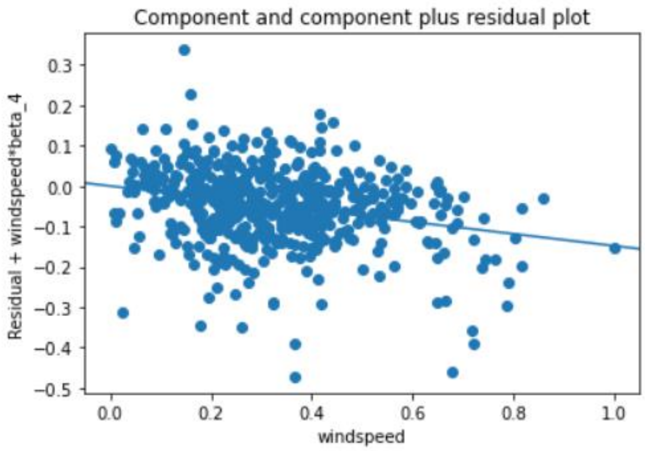
   - We validate the assumptions of Linear Regression firstly by plotting a distplot of the residuals and analyzing them to see if it is a normal distribution or not and if it has mean = 0. From our analysis we can see that the error terms have a normal distribution with mean =0.



- We also plotted a heatmap and see that there is no multicollinearity observed.
- We also validated the linear relationship using the Component and Component Plus Residual Plot(CCPR).







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

Ans.

Following are the top three features contributing significantly towards explaining the demand of shared bikes:

- **Temp** (Temperature) A coefficient value of “0.4777” indicates that a unit increase in the temp variable, increases the bike hire by 0.4777.
- **Year** - A coefficient value of “0.2341” indicates that a unit increase of this variable, will increase the bike hire by 0.2342 units.
- **Light_snowrain(weathersit)** – A coefficient value of “-0.2850” indicates that, a unit increase of this variable will decrease the bike hire by -0.2850 units.

General Subjective Questions

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

Ans.

- Linear regression is a statistical method that is used for predictive analysis. Linear regression basically makes predictions for continuous/real or numeric variables such as salary, age, etc.

- Linear regression algorithm shows a linear relationship between a dependent variable (y) and one or more independent variables(x), hence called as linear regression.

- Mathematically, linear regression can be represented as follows:

- **Y = B1 + B0X**

Where, Y = Dependent Variable

X = Independent Variable

B0 = Intercept of the line

B1 = Linear regression Coefficient

- **Types of Linear regression**:
- **Simple Linear Regression:**

  If a single independent variable is used to predict the value of a numerical dependent variable, then such linear regression is called Simple Linear Regression.

- **Multiple Linear Regression:**
- If more than one dependent variable is used to predict the value of a numerical dependent variable, then such linear regression is called Multiple Linear Regression.
- **Best Fit Line:**

  When working with linear regression, our main goal is to find the best fit line which means that the errors between the predicted values and actual values should be minimized.

  -**Linear regression has five key assumptions:**

  -Linear Relationship: There exists a linear relationship between the independent variable x and the dependent variable y

  -Independence: The residuals are independent.

  - Homoscedasticity: The residual at constant variance at every level of x.

  -Normality: The residuals of model are normally distributed.

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

Ans.

- Anscombe's quartet is used to illustrate the importance of Exploratory data analysis and the drawbacks of depending upon only the summary statistics.
- These datasets were created by Francis Anscombe in 1973 to demonstrate the importance of visualizing data and show that summary statistics alone can be misleading.
- It also signifies the importance of data visualization to spot trends, outliers and other crucial details that might not be obvious from the summary statistics alone.
- Anscombe's quartet comprises a set of four dataset, having identical descriptive properties in terms of means, variance, R-squared, correlations and linear regression lines but having different representations when we scatter plot on graph.
- The 4 datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths.
- Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient and linear regression line.

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

Ans.

- The Pearson's r is a numerical summary of the strength of the linear association between variables.
- If the variables tend to go up and down together, the correlation coefficient will be positive.
- If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, then the correlation coefficient will be negative.
- The Pearson's correlation coefficient lies between -1 to +1 where:

  r=1 , means data is perfectly linear with a positive slope
  r=-1, means data is perfectly linear with a negative slope
  r=0, means there is no linear association
  r>0<5, means there is a weak association
  r>5<8, means there is a moderate association
  r>8, means there is a strong association

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

Ans.

- Scaling is a step of data pre-processing which is applied to the independent variables to normalize the data within a particular range.
- It also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitude, units, and range.
- **If scaling is not done**, the algorithm takes only magnitude in account and not units hence incorrect modelling.
- **To solve this issue**, we must do scaling to bring all the variables to the same level of magnitude.
- **Normalized scaling** brings all the data in the range of 0 & 1.
- Whereas **standardized scaling** replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean 0 and standard deviation 1.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans.

- An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model.
- As, VIF is given by the below equation:

  VIF = 1/(1-R^2)

- We can say that when the value of R^2 approaches to 1, VIF values become infinity.

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

Ans.

- Quantile-Quantile plot is a graphical tool to help us access if a set of data plausibly cake from some theoretical distribution such as a Normal, Exponential or Normal Distribution.
- Also, it helps to determine if two data sets come from populations within a common distribution.
- This helps in a **scenario of linear regression** where we have training data and testing data set received separately and then we can confirm using the Q-Q plot, that both the data sets are from populations with same distributions.
- In linear regression, it is used to check the following scenarios:
- If two data sets:
- Come from population with a common distribution
- Have common location and scale
- Have similar distributional shapes
- Have similar tail behavior