# 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.

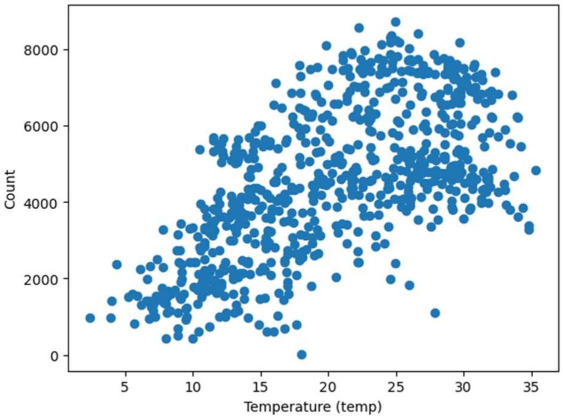
- Weather – Weather has a very good impact on the bike booking count. Most of the bookings happened in clear weather followed by misty weather and least booking happened in light snow/ rainy weather.
- Season – Most of the bookings happened in Fall followed by summer. Winter stands third and the least number of bookings happened in Spring.
- Months - Most of the bookings happened from May to September.
- Holiday – More bookings happened on a non-holiday.
- Year – There has been a significant rise in booking from 2018 to 2019.
- Weekday – The bookings are almost evenly distributed through out the week.

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

Ans. Dropping the first column in dummy variable creation helps to avoid the multicollinearity in regression models. This provides more stable and interpretable coefficients. If we do not drop the first variable, it becomes perfectly predictable from others. For example, in gender (male/female) encoding, dropping the first dummy (e.g., male) prevents perfect correlation with the second (female).

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

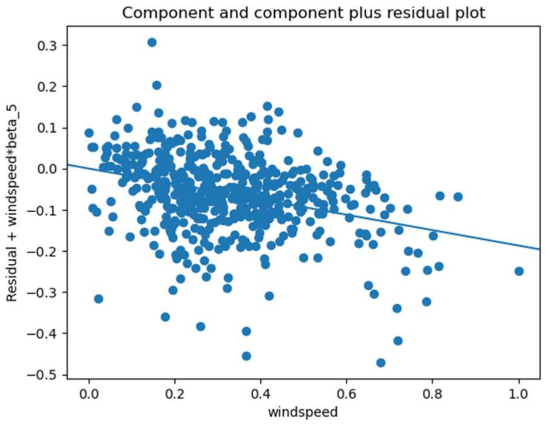
Ans. As per my observation, the variable with highest co-relation with count is “temp”. “Atemp” also has a very good correlation as it is a derived variable from temp.

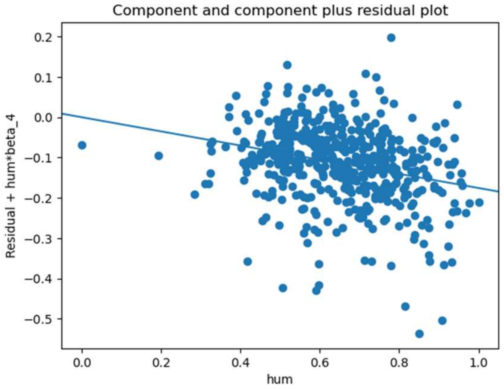


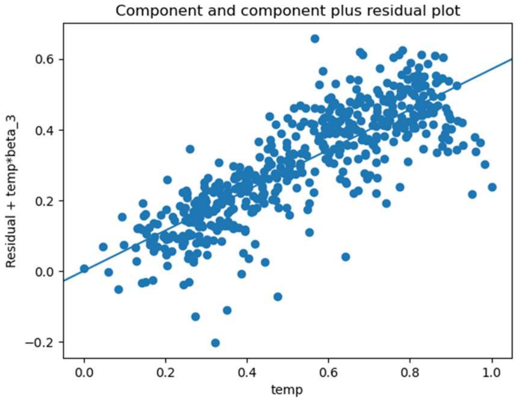
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. Validating the assumption of Linear Regression Model :

- **Linear Relationship**

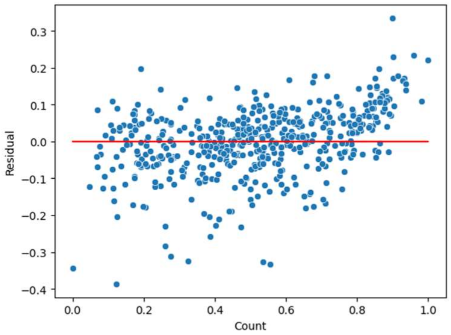






The above plots depict the relationship between the model and the different predictor variables. We can clearly see that the linearity is well preserved.

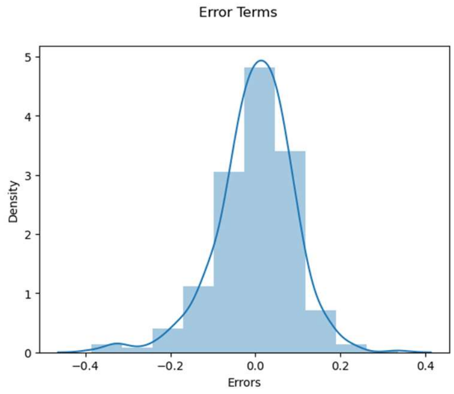
- **Homoscedasticity**



No visible pattern is observed in residual values. Homoscedacity is well preserved.

- **Normality of Errors**

The error is normally distributed.



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

Ans. The top 3 features as per my model are Temperature, Weather (clear weather) and year.

# General Subjective Questions

1. Explain the linear regression algorithm in detail

Ans. In linear regression, the target variable is predicted based on input features through finding a best-fit line. The algorithm attempts to draw a straight line with as few vertical differences (residuals) between predicted and actual values. It seeks the smallest sum of square, and finds such a line with minimum possible sum of squares. The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by minimising the cost function (RSS in this case, using the Ordinary Least Squares method) which is done using the following two methods: Differentiation and Gradient descent method.

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) but vastly differ in their distributions. It highlights the importance of visualizing data to understand its underlying patterns, cautioning against relying solely on summary statistics when interpreting datasets.

3. What is Pearson's R?

Ans. The Pearson correlation coefficient aka Pearsons R, is an indicator of linear association between two variables. It signifies the size and direction of their relationship, which runs from -1 (a perfect negative correlation) to 1 (a perfect positive correlation), with zero indicating no linear connection. It is frequently used in statistics to measure the collinearity between variables.

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

Ans. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. Sometimes the data that we have can be of different magnitudes or values. This may result variations in the algorithms. Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean ($\mu$) zero and standard deviation one ($\sigma$). Normalized scaling brings all of the data in the range of 0 and 1.

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

Ans. A VIF (Variance Inflation Factor) becomes infinite when perfect multicollinearity exists among variables. This means one or more variables can be perfectly predicted from others. In such cases, the VIF cannot be computed, as it involves dividing by zero, indicating an issue in the regression model due to redundant predictors.

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

Ans. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distribution.