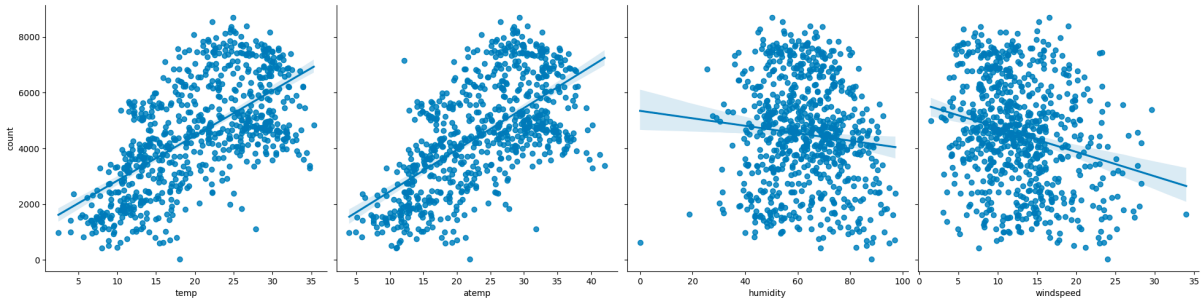
Assignment-based Subjective Questions/Answers

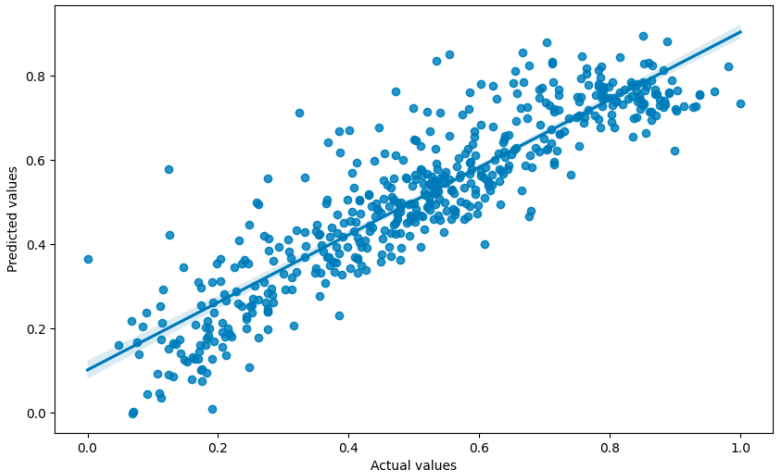
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   - Season
     - Demand for Bikes is highest in Fall, followed by Summer and Winter.
     - Strategic capacity planning and advertising can attract more business for the company.
   - Weather
     - Days with clear sky/sky with a few clouds are the most favorable for bike rides.
     - Data is not available for days with heavy rain/snow.
   - Year
     - Year 2019 has more bikes rented than the year 2018. That means company is having better business.
   - Month
     - Demand of bikes is higher during months May, June, July, Aug, Sep and Oct.
     - These months are of seasons Summer and Fall.
   - Holiday/WorkingDay
     - Higher demand of bikes observed during holidays/non-working days.
   - Weekday
     - Demand for bikes does not change a much for weekdays.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   - If the number of categories in a categorical variable is n, number of dummies required is n-1.
   - Hence when *drop_first=True* is used, an extra column is avoided during dummy variable creation.
   - And that helps avoiding multicollinearity being added in the model.

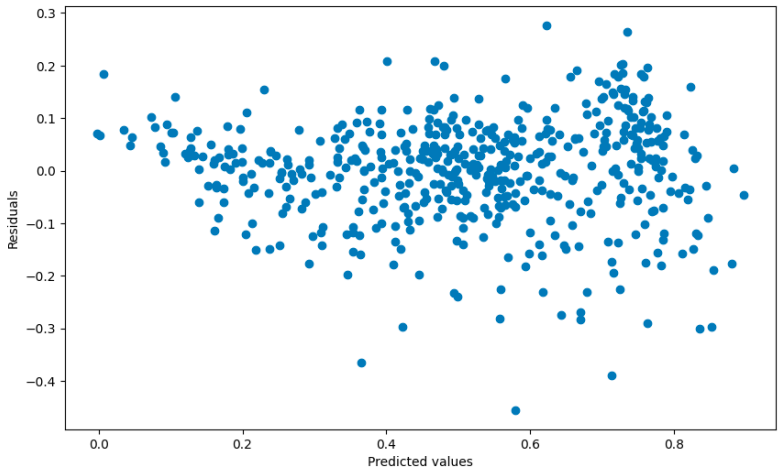
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   - 

   - As per the pair-plot among the numerical variables, it is clear that temp and atemp have the highest correlation with the target variable.
   - atemp is a derived variable from temp, humidity and windspeed. Hence it also has high correlation with temp. It is dropped during model creation to reduce multicollinearity.

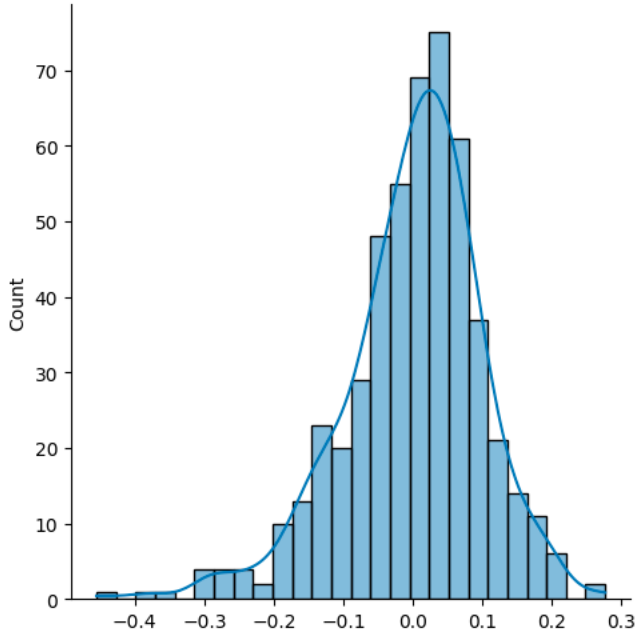
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   - The relationship between the dependent and independent variables is linear.
     - This is validated by plotting graph of values predicted by the model and the actual values of the target variable.



   - The variance of error is constant.
     - This is validated with a plot of residual vs predicted values.



   - The independent variables are not highly correlated with each other.
     - This is validated with VIF results of variables participating in the model.
   - Error Terms are normally distributed.
     - This is validated by creating histogram of distribution of error terms.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - Temperature
     - The temperature has the highest coefficient value 0.57. That means with an increase of one unit of temperature value the count value will be increased by 0.57 units.
   - Year
     - Year has the second highest coefficient value 0.23. It is evident that the bike rent number has gone up from year 2018 to 2019.
   - Winter
     - Coefficient value for winter is 0.13. This suggests that with an increase of one unit in value of winter will increase the bike rent number by 0.13 units.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   - Linear regression is a foundational algorithm in data science. It is a supervised learning algorithm that finds the best linear relationship between a dependent variable and one or more independent variables.
   - There are two types of linear regression algorithms.
     i. Simple Linear Regression – attempts to explain the relationship between one dependent variable and one independent variable.
        1. Line Equation: $Y = \beta_o + \beta_1 X$
     ii. Multiple Linear Regression – attempts to explain the relationship between one dependent variable and multiple independent variables.
        1. Line Equation: $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$
   - Line represented by above equations represents the relationship between the dependent and independent variable(s).

- Slope of the line ($\beta_0, \beta_1, \beta_2 \ldots \beta_n$) indicates how much the dependent variable changes for a unit change in the dependent variable(s).
- Cost function helps to identify the best values for the slope of the line.
- While finding the best fit line, we encounter that there are errors while mapping the actual values to the line. These errors are called residuals.
- Linear regressions can be used in business to evaluate trends and make estimates or forecasts. It is used in many domains such as finance, economics, healthcare to understand and predict the behavior of a variable.

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet is a set of four datasets that have nearly identical summary statistics but completely different patterns when plotted on graph.
- They were constructed in 1973 by statistician Francis Anscombe to demonstrate both the importance of graphing data while analyzing it and the effect of outliers.
- Anscombe's quartet illustrates the importance of exploratory data analysis and drawbacks of depending only on summary statistics.
- Important points:
    i. Understanding the importance of visualizing data can help us make better decisions and avoid making incorrect assumptions.
    ii. It can also help us identify errors or inconsistencies in our data.

3. What is Person's R?

- The Person correlation coefficient (r), also known as Pearson's R is the measure of the strength of a linear regression between two variables.
- Value of r can range between -1 and +1.
- Value 0 indicates that there is no correlation between the variables.
- Positive r value indicates the positive correlation that means if one variable increases the other one increases too.
- Negative r value indicates the negative correlation which means increase in one variable decreases the other variable and vice versa.
- r = 1 indicates the perfect positive relationship between the two variables.
- r = -1 indicates the perfect negative relationship between the two variables.

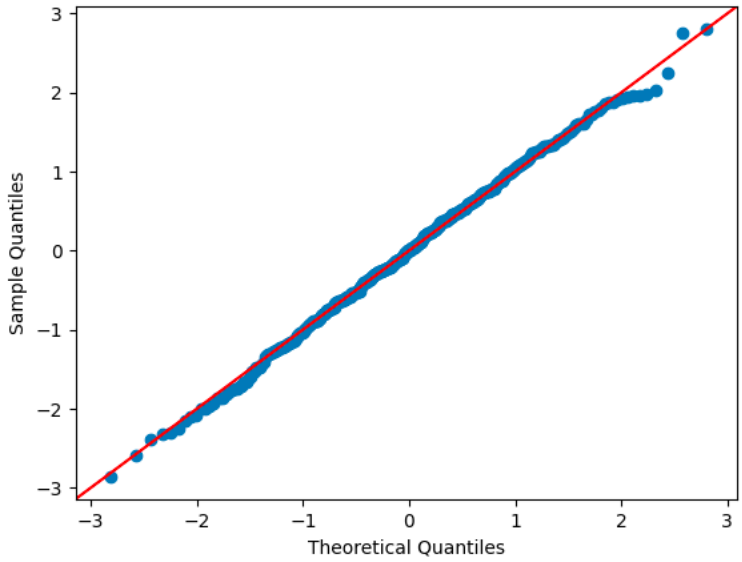
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling vs standardized scaling?

- Scaling is the process of transforming values of independent features in the dataset such that they are withing specific range.
- Scaling is necessary when features have different ranges, units of measurement or order of magnitude. Without scaling, coefficients are impaired.
- Normalized scaling/Min-Max scaling normalizes the data withing the range of 0 and 1. It also helps to normalize the outliers.
    i. $x = \frac{x - \min(x)}{\max(x) - \min(x)}$
- Standardized scaling brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation (σ) 1.

i. $x = \frac{x-\mu}{\sigma}$

5. **You might have observed that sometimes value of VIF is infinite. Why does this happen?**
   - Formula of VIF is $VIF = \frac{1}{1-R^2}$
   - It is clear from the formula that when R = 1, the VIF ends up being infinite.

6. **What is a Q-Q plot? Explain the use and importance of Q-Q plot in linear regression.**
   - Q-Q Plot is also known as Quantile-Quantile plot. It is a plot of quantiles of two distributions against each other. It helps in determining if a dataset follows any particular type of probability distribution such as normal, uniform, exponential.
   - It is also useful to determine
     i. If two populations are of the same distribution
     ii. If residuals follow a normal distribution
     iii. Skewness of distribution.
   - If the datasets being compared of the same type of distribution type, the Q-Q plot will result in a straight line.



   -
   - The Q-Q plots are in linear regression to identify if train dataset and test dataset are from the populations with the same distributions.