BIKE SHARE REGRESSION MODEL QUESTIONS

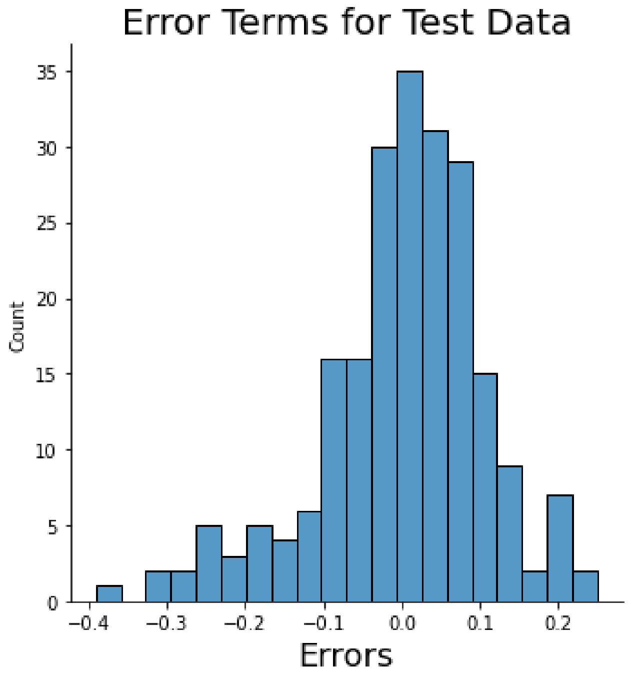
Assignment Based Subjective Questions:

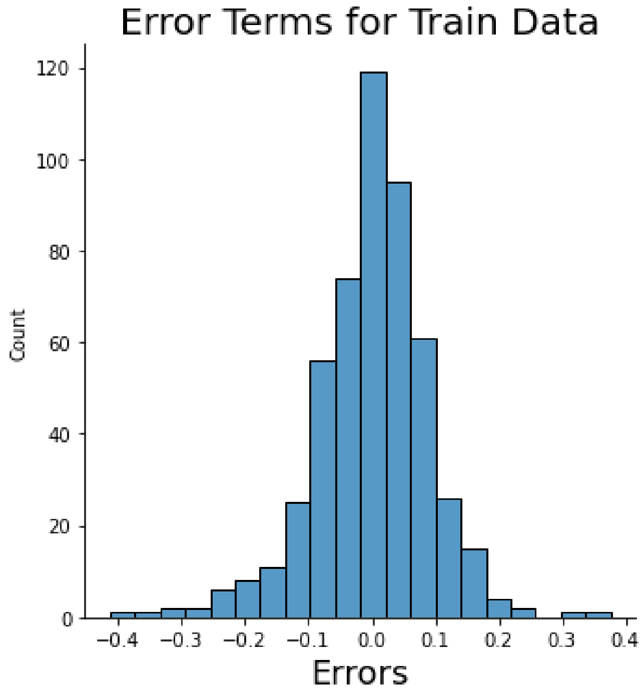
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   - Converting the numerical values of categorical variables into categorical values using the given dictionary would provide us with more insights about what day, month (dependent on September only), season (dependent on Spring only) or weather would have different effects on the dependent variable, which is the count of bike rentals. Thus, the assignment of the categorical values for these categories will give us a deeper analysis of what factors really impact the bike rentals and how can the company focus on generating revenue based on those factors.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   - In order to avoid Multicollinearity (i.e., correlation between different variables), we must use drop_first=True during dummy variable creation and also because the number of dummy variables allowed = (Total number of Variables)-1

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   - The numerical variables ‘temp’ for actual temperature and ‘atemp’ for feels like temperature have the highest correlation with the target variable ‘cnt’ if we are looking at the pair plots.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   - By Doing the Residual Analysis of the train data and test data, A Normal Distribution of the Error terms was observed with the mean 0 and the Error terms were independent of each other. The Graphs below demonstrate the assumptions.





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

- The Top 3 features are Temperature, Year and the Month of September that contribute significantly towards the demand of shared bikes.

General Subjective Questions:

1. **Explain the linear regression algorithm in detail.**
   - Linear Regression is one of the types of the machine learning models where the output variables to be predicted is Continuous in nature, e.g., the price of the car rental. Linear Regression models the relationship between dependent variable and an independent variable by fitting a linear equation using a Straight Line. Linear Regression also fall under the category of Supervised Learning Methods where the already given data (past data) is used for building the model.

2. **Explain the Anscombe's quartet in detail.**
   - Anscombe's quartet is a group of Four data sets that have Identical Simple Descriptive Statistics, yet have different distributions when they are shown Graphically. Additionally, each of the four data sets contains 11 points in (x,y) coordinate form.

3. **What is Pearson's R?**
   - Pearson's R, also, known as the correlation coefficient is the measure of the linear correlation between two data points where its values range from -1 to 1. When the value of R is -1, then there is a Negative Correlation between two data variables and when its 0, then there is No Correlation between two data variables. When the value of R is 1, then there is a Positive Correlation between two data points.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   - (i). Scaling is a technique to normalize the range of independent variables or the features of data so that the process of the model building is smoother. (ii). Scaling is done during the pre-processing

step of the model building in order to handle highly varying magnitudes of some features. And, the reason scaling is performed is because we want the predictor values to have a Mean 0. (iii). The difference is that the normalised scaling converts all the data in the range of 0 and 1 and the standardized scaling converts all the data into a standard normal distribution with the mean 0 and standard deviation 1.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   - The value of VIF (Variance Inflation Factor) is infinite only in an ideal situation where there is a Perfect correlation between two independent variables. In case of perfect correlation, the value of R-Squared=1 and hence, 1/ (1- R-Squared) is infinite, which is the value of VIF. It also means one variable could be expressed as the linear combination of the other variable.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   - Q-Q plot is a graphical plot to check if a set of data has a reasonable distribution such as normal distribution, uniform distribution etc. and also used to check if two data sets have a similar distribution. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.