Bike-Rental-Regression-Assignment – Manoj Kumar.

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: EDA on Categorical Variable:

   a. SUMMER:
      I. Summer and fall season have a higher average rental bike per day than overall average rental.
      II. Summer and fall season have a higher average rental bike per day than Winter and Spring.

   b. YR:
      I. 2019 year have a higher average rental bike per day than overall average rental.
      II. 2019 year have a higher average rental bike per day than 2018 year.

   c. MNTH:
      I. Bike Rentals are more prominent in the month of September and October.

   d. HOLIDAY:
      I. Holidays have a higher average rental bike per day than overall average rental.
      II. Holidays have a higher average rental bike per day than Non-Holidays.

   e. WEEKDAY :
      I. Weekday 2,3,4,5,6 have a higher average rental bike per day than overall average rental
      II. Weekday 2,3,4,5,6 have a higher average rental bike per day than Weekday 1,2

   f. WORKINGDAY :
      I. Working day have a higher average rental bike per day than overall average rental
      II. Working day have a higher average rental bike per day than Non-Working Day

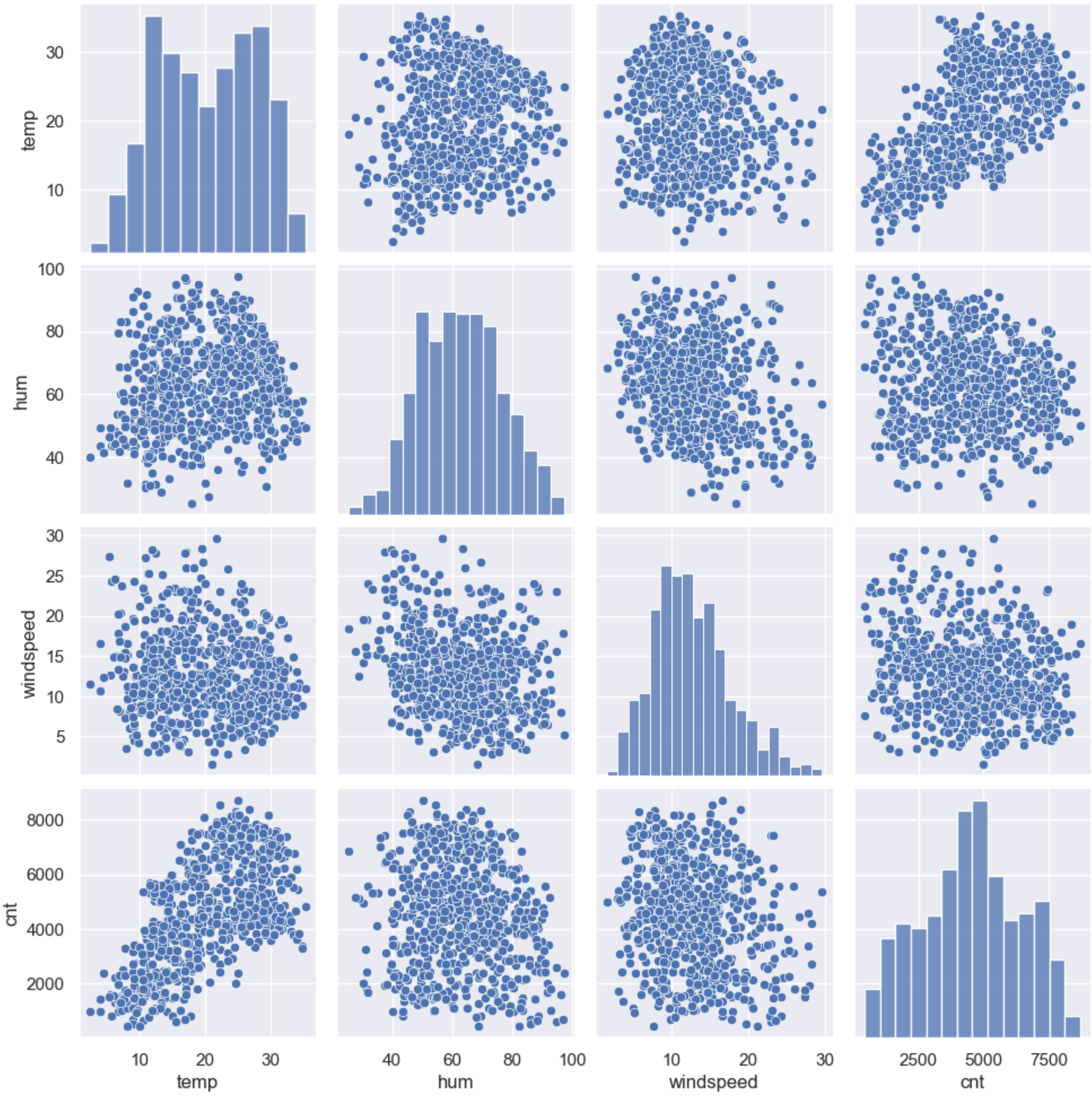
   g. WEATHERSIT :
      I. Clear, Few clouds, Partly cloudy, Partly cloudy Weather have a higher average rental bike per day than overall average rental.
      II. Clear, Few clouds, Partly cloudy, Partly cloudy Weather have a higher average rental bike per day than Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist, Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

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

   Ans: It is important to use drop first as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. By dropping one of the one-hot encoded columns from each categorical feature, we make sure that there are no reference columns and the remaining columns become linearly independent.

   If there are n categorical values then n-1 dummy columns have to be created.
   In the BoomingBikes Sharing Assignment the following are the categorical columns for which we have to create the dummy variables. **SEASON,MNTH,WEEKDAY,WEATHERSIT**

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



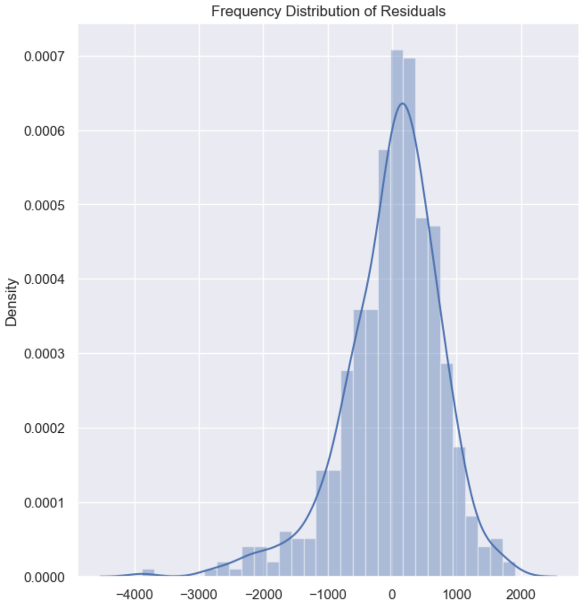
Ans: Looking at the above pair plot we can see the “temp” and “atemp” have the linear correlation with the target variable “cnt”. Looking at the couple of points on the edges of the scatter plot we can conclude that the “temp” has the highest correlation with the target variable.

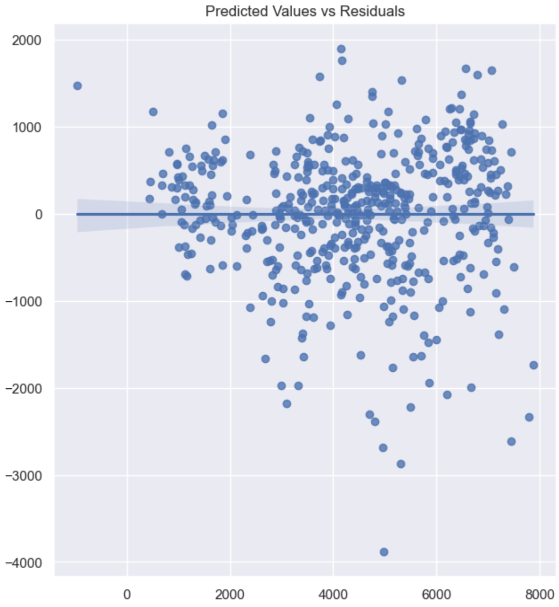
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans:

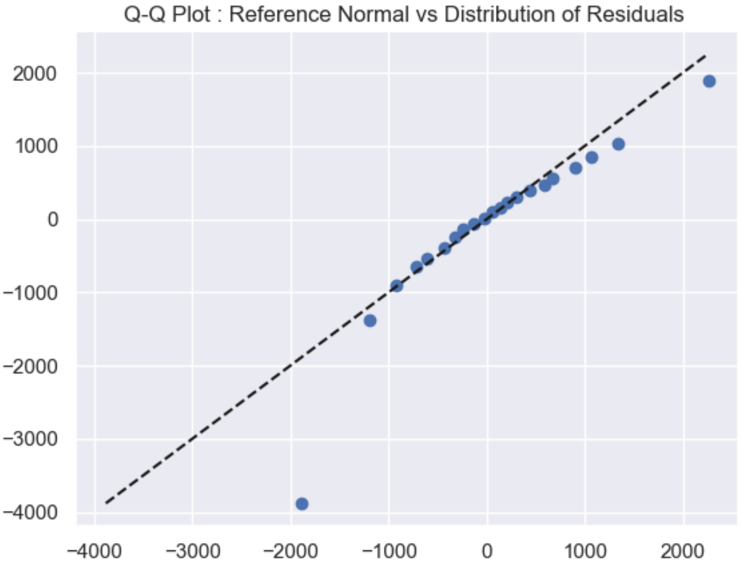
a. Assumption of Normally Distributed Error Terms Validated by plotting adistribution of the residuals.

b. Assumption of Error Terms Being Independent
Validated by plotting a regression plot of the residual’s vs predicted values(0-1).

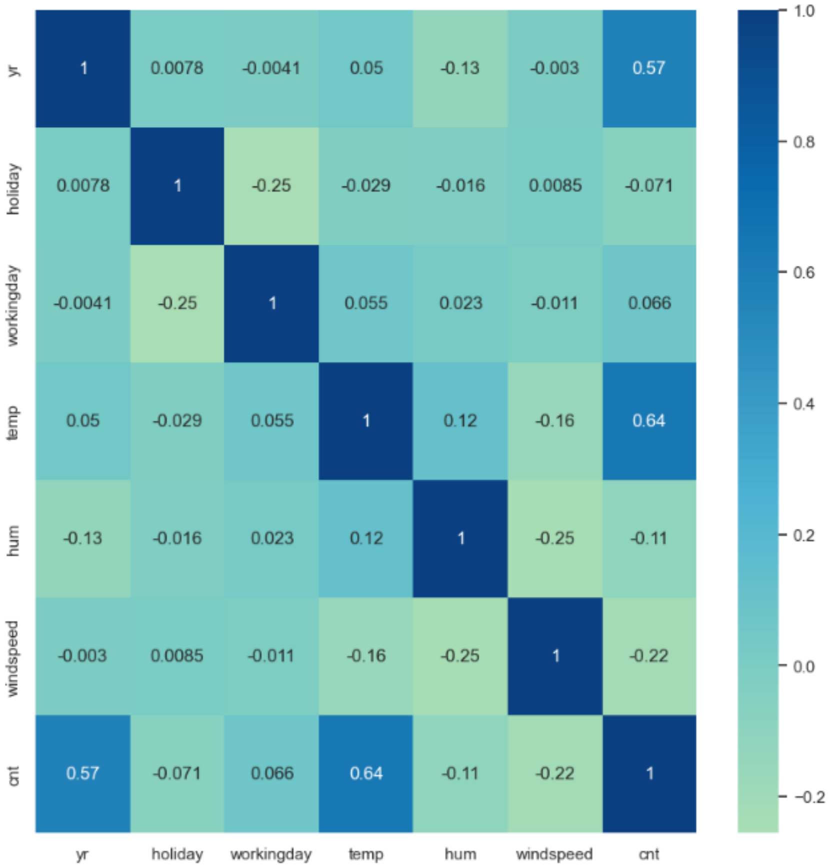




c. This plot further shows that the residual distribution is approximately normal for all test data with values within range of training data. Extrapolated points show significant predict inaccuracy.



d. Assumption of No or little multi-correlation
Validated by plotting a heat map for the correlation matrix and Variance inflation factor (VIF) of the independent variables



	yr	holiday	workingday	temp	hum	windspeed	cnt
yr	1	0.0078	-0.0041	0.05	-0.13	-0.003	0.57
holiday	0.0078	1	-0.25	-0.029	-0.016	0.0085	-0.071
workingday	-0.0041	-0.25	1	0.055	0.023	-0.011	0.066
temp	0.05	-0.029	0.055	1	0.12	-0.16	0.64
hum	-0.13	-0.016	0.023	0.12	1	-0.25	-0.11
windspeed	-0.003	0.0085	-0.011	-0.16	-0.25	1	-0.22
cnt	0.57	-0.071	0.066	0.64	-0.11	-0.22	1

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 contributing significantly towards the demand of the shared bikes are the **temperature**, the **year** and the **holiday variables**.

Bike-Rental-Regression-Assignment – Manoj Kumar.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans: Linear Regression is an ML algorithm used for supervised learning. It helps in predicting a dependent variable(target) based on the given independent variable(s). The regression technique tends to establish a linear relationship between a dependent variable and the other given independent variables. There are two types of linear regression- simple linear regression and multiple linear regression. Simple linear regression is used when a single independent variable is used to predict the value of the target variable. Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable. A linear line showing the relationship between the dependent and independent variables is called a regression line. A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

2. **Explain the Anscombe's quartet in detail.**

Ans: Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give an accurate representation of two datasets being compared.

3. **What is Pearson's R?**

Ans: Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.

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

Ans: Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range.

The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.

The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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

Ans: The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

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

Ans: The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.