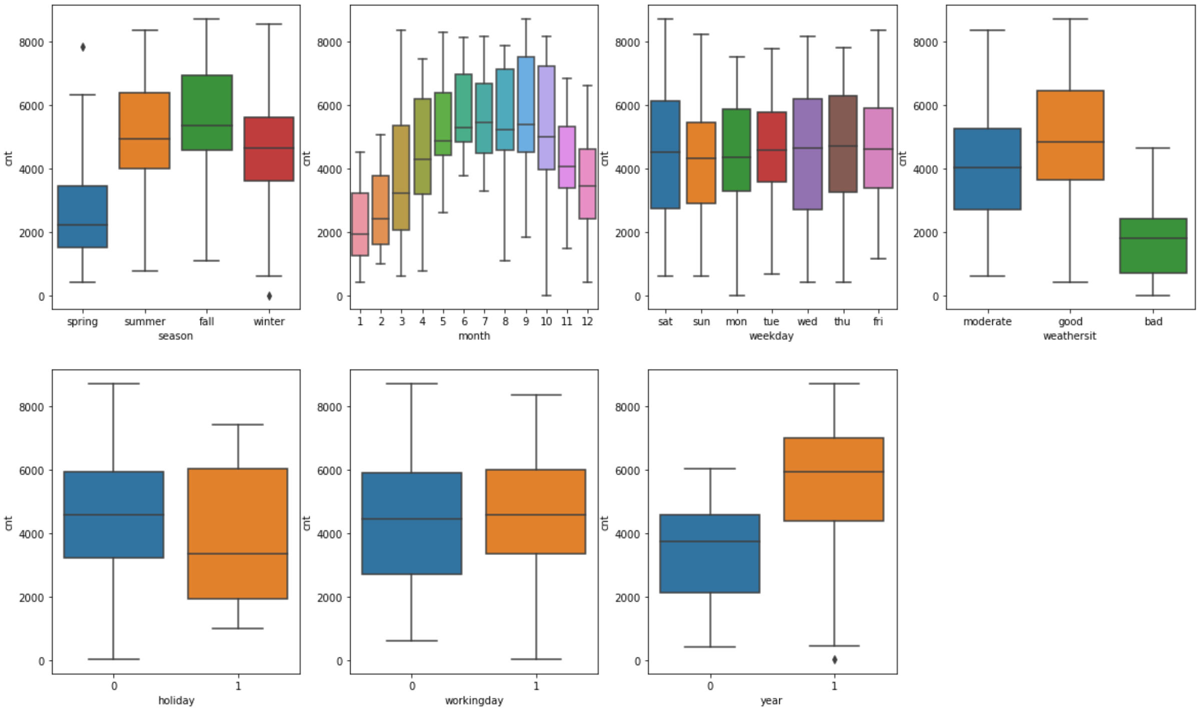
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   Ans- As observed from plotting the categorical variables in jupyter notebook, it is been seen that, there are increase in number of demands for shared number of bikes varies with different season, different months, different weekdays and different weather components also demand in year 2019 was much higher compared to 2018.

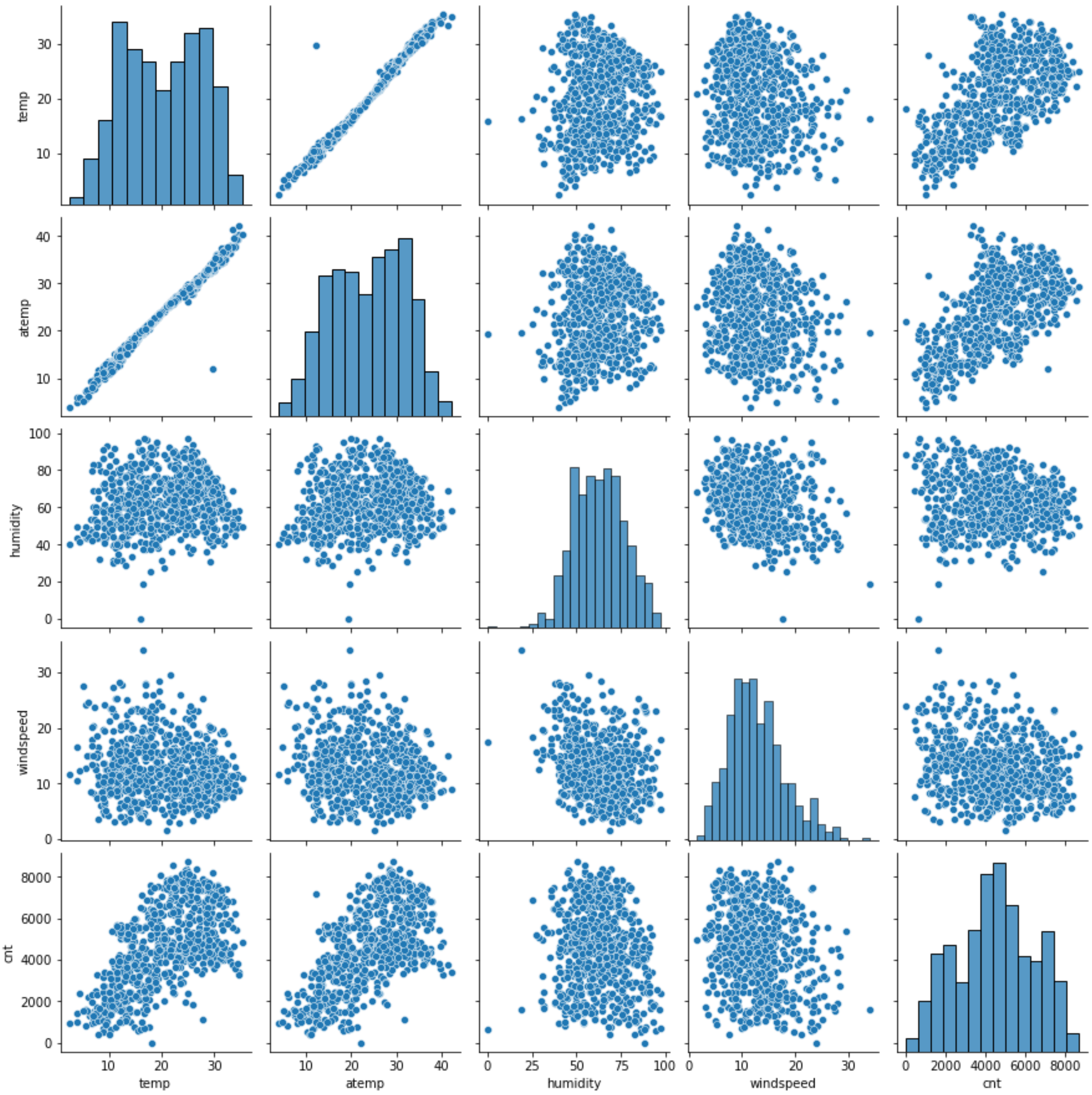


2. Why is it important to use drop first=True during dummy variable creation?

Ans- drop first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Ans- From the pairplot below it is clear that temp and atemp are highly correlated with the target variable.



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

Ans – Linear regression models are validated based on Linearity, No auto correlation, Normality of errors, Homoscedasity.

4. 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 to the demand are windspeed, year and working day.

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Ans- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

2, Explain the Anscombe’s quartet in detail.

Ans- Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3. What is Pearson’s R?

Ans- In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

5. 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. It also helps in speeding up the calculations in an algorithm and Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

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

Ans- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

6. 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.