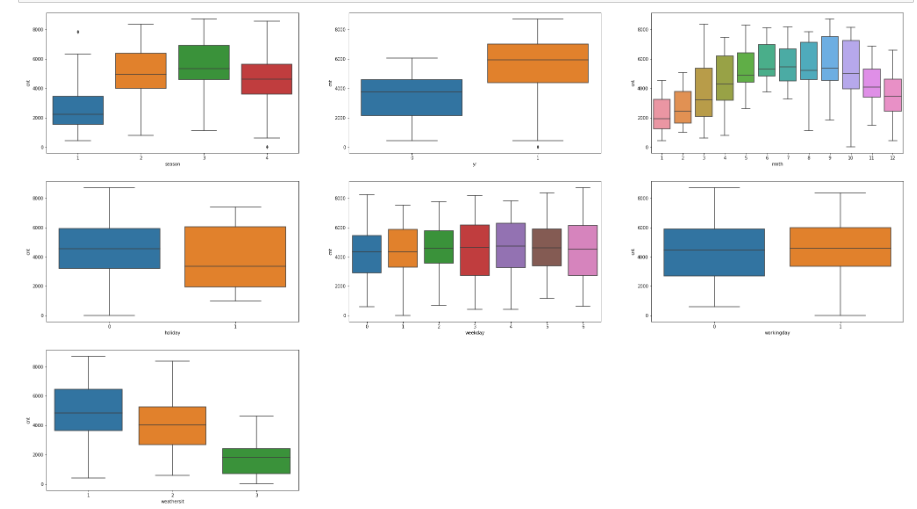
**Assignment-based Subjective Questions**

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:



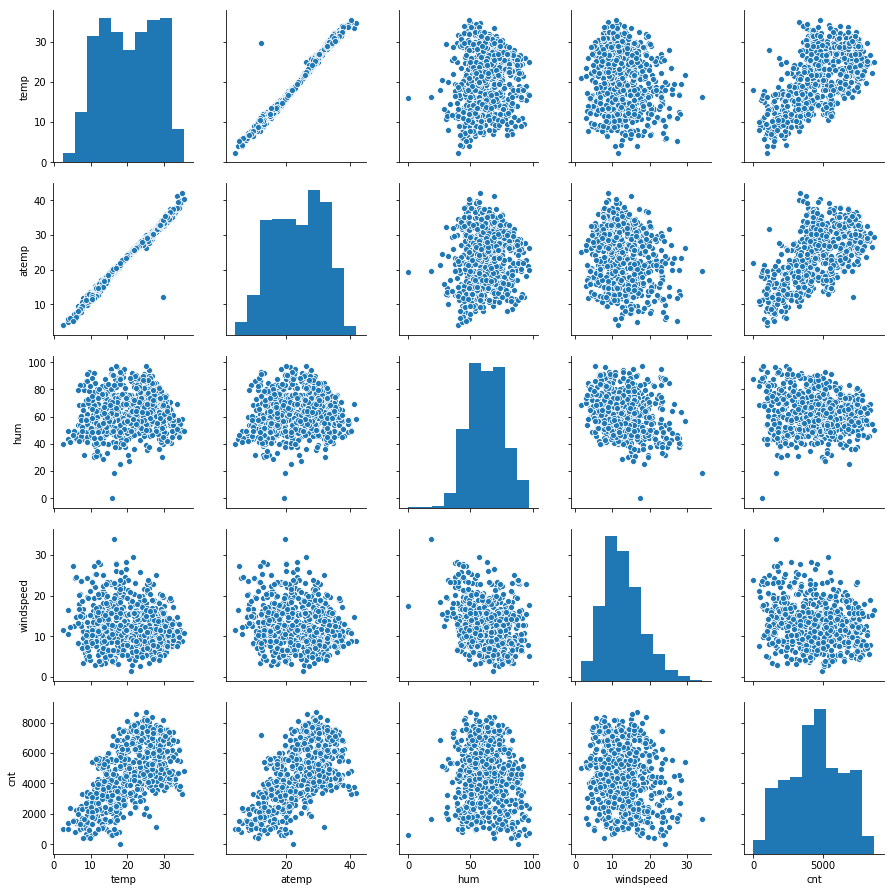
1. Season Vs Cnt (Usage) : From the EDA analysis it can observed that as the season change from Winter to Spring there is a great increase this slightly increases in summer and drops down slightly in Fall. In general, this suggests seasons does have an effect in usage of bikes.
2. Year Vs Cnt (Usage) : As the year moves from 2018 to 2019 this increased usage of the bikes, this can be due to increased popularity.
3. Month Vs Cnt(Usage) : This reflects the information presented through Season vs cnt as this suggest as the month moves from March there is a steady increase until September then it falls down until the next March.
4. Holiday Vs Cnt(Usage) : Median for non-holiday is higher with closer upper and lower quartiles suggesting there might be slight indication that people prefer to use the bikes during non-holiday peiod.
5. Weekday Vs Cnt(Usage) : As the weekday goes into working days Wednesday and Thursday there is a slight increase in median , showing a slight indication that people prefer to use it during mid week more than weekends
6. Workingday Vs Cnt(Usage) : this correlates with holiday and weekday outcome where slight indication that people prefer to use bikes during working days, but not enough to conclude as a critical trend.
7. Weathersit Vs Cnt(Usage) : when the weather is Clear, Few clouds, Partly cloudy, Partly cloudy the usage is at its highest, when the weather worsens to Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds the usage drops same as it worsens to Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds. There is no usage as it turns to Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog. This shows the usage is dependent on weathersit.
8. Why is it important to use drop\_first=True during dummy variable creation?

If first variable not dropped then this can lead to dummy variable trap. This happens when the independent variable are multicollinear. When new dummy variable is created, each categorical variable will be represented using 0 or 1 using the one hot encoding method. 0 represent the absence of it and 1 represent the present of it. A simple explanation is when two or more variables are highly correlated; one variable can be predicted by the outcome of the other variables. For example take Season from this data set, one hot encoding will turn this into

|  |  |  |  |
| --- | --- | --- | --- |
| **Spring** | **Summer** | **Fall** | **Winter** |
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |

By removing the Winter we can still predict the outcome, absence of all three spring, summer and fall would suggest that its winter. If not dropped, it will make it harder to predict individual effect on prediction model due to multicollinearity.

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



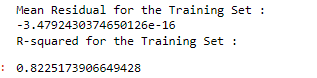
It can be observed that there is a high correlation between ‘temp’ and ‘atemp’ against target variable ‘cnt’. Also a linear relationship between ‘atemp’ and ‘temp’ can be observed as well. ‘Windspeed’ and ‘hum’ does not have any visible relationship with any of the other numerical variables through pariplot.

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

Assumption 1: **Linearity**: Relationship between the dependant variable and independent variable to be linear

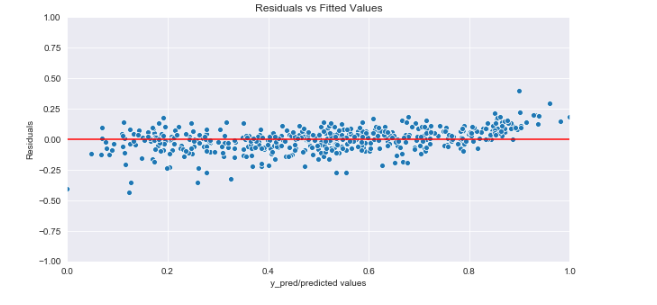
Pair plot shows the linearity between dependant and independent variable model to be linear. Then it was confirmed using r2, which was 0.822517 in this model. R2 explains the difference between variance explained by the model divided by the total variance, the value was closer to 1 hence there a clear indication of linearity.

Assumption 2: **Mean of Residuals**: Residual is the difference between the actual value and predicted value, the assumption is the value should be 0. In the designed model value is close to 0, hence proving the model.



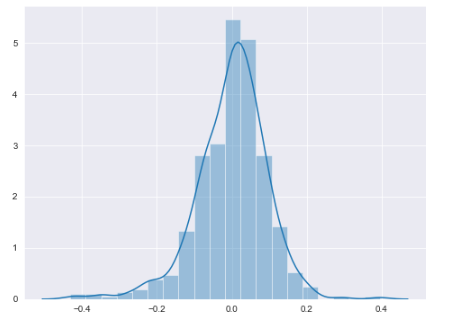
Assumption 3: **Homoscedasticity**: Homoscedasticity means the residuals have a equal or almost equal variance across the regression line.

By plotting, the residual it is observed as shown in the image below the data is equally distributed across both side of the 0 point.

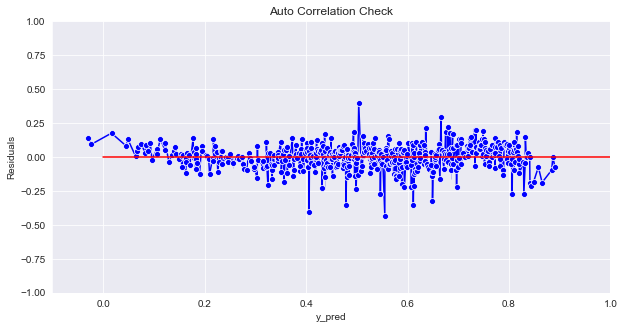


Assumption 4 : **Normality of error terms/residuals** : Residual terms are normally distributed and meets the central limit theorem, it will also be centred around 0.

This can be observed in the image below from the model and the curve is narrower.



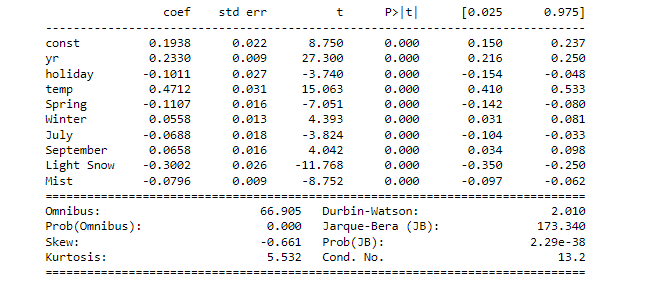
Assumption 5 : **No Residual Autocorrelation** : When there is unexplained pattern in the Y variable this can show up in Y variable. Which means the current value is dependent on the previous one. From the image below it can be seen that it did not form any pattern.



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

Top three features from the model would be as follows:

* Temp : Temperature
* Light Snow : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
* Yr : year



**General Subjective Questions**

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