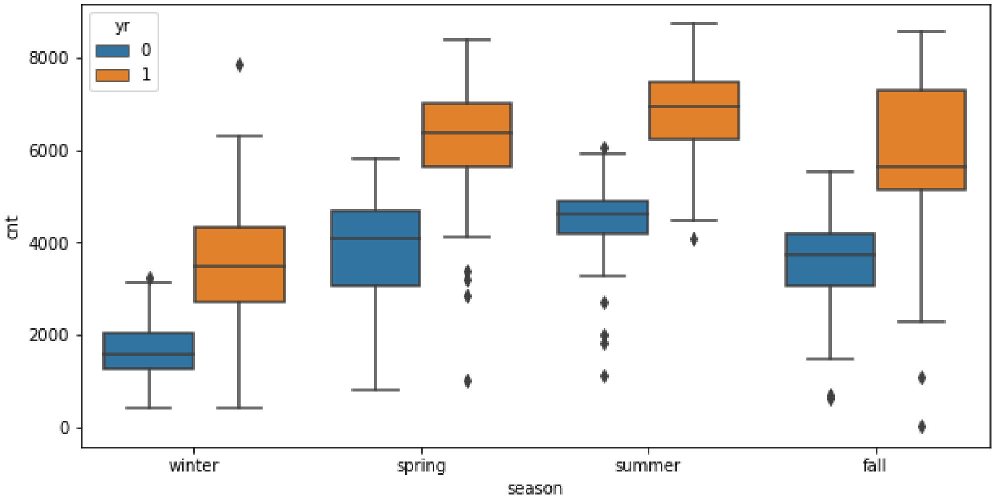
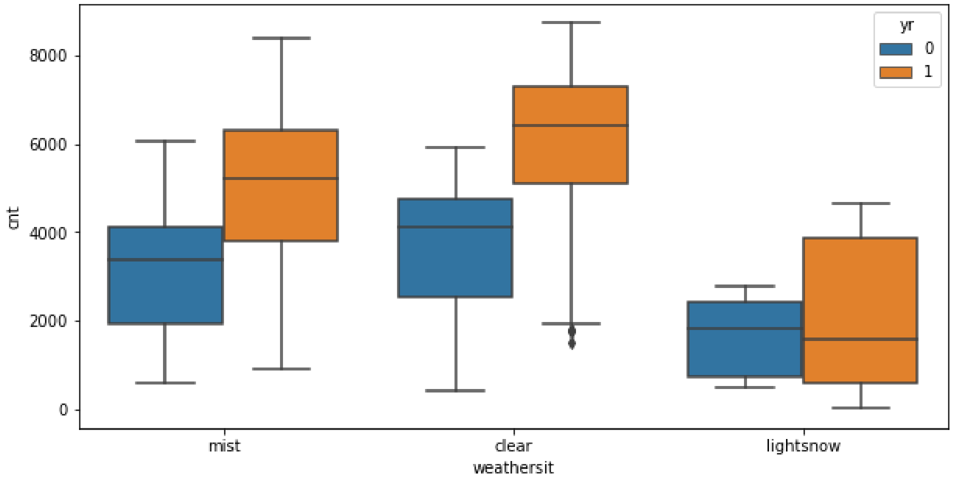
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

- As could understand from below picture that except Winter, Spring, Summer and Fall have higher cycle rentals and it is significantly increased in the year 2018 from year 2017.



- Below picture clearly explains that clearly explains that clear weather boosts the rentals whereas light snow weather conditions have a smaller number of rentals.



$$cnt=0.247\times yr+0.03\times spring+0.09\times summer-0.177\times windspeed-0.227\times winter-0.288\times lightsnow-0.086\times mist$$

- From the above best fitted line equation spring, summer are positively correlated with Rentals count whereas winter, lightsnow and misty weather conditions are negatively correlated with rentals count.

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

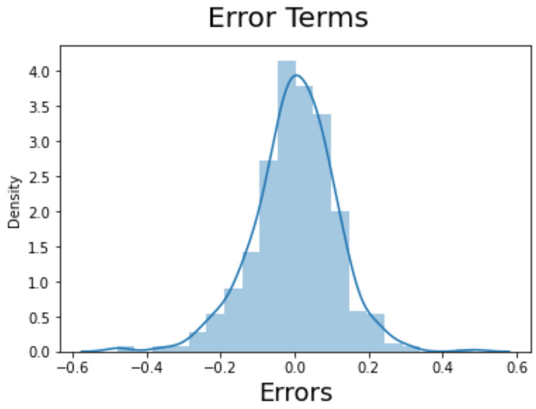
- When there is a category of N level it can have N-1 of dummy variable. So, it is required to drop the first column. Since we can identify the first column with the values in other dummy variables.

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

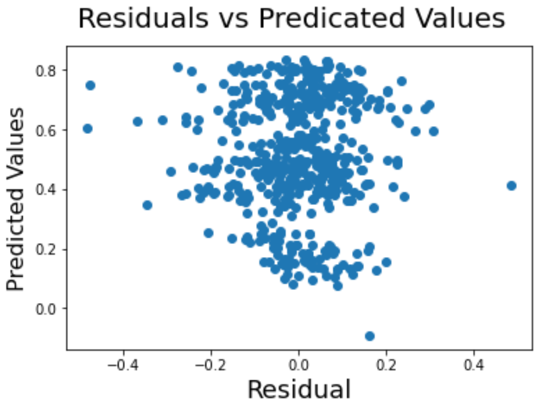
- atemp is the variable which is highly correlated with target variable cnt.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

- Error Range (Residuals) are normally distributed is validated by plotting the graph.



- No Multicollinearity is validated against VIF (Variance Inflation Factor) and RFE technique.
- Homoscedastic can be validated by plotting the graph between Residuals and Predicted values. There is no clear pattern.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

- Yr, summer and lightsnow, winter are the top 3 features.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

- It is one of the basics of machine learning which falls under supervised learning where we allow the machine to train on the certain portion of the data and to predict the data for the remaining test data.
- Its equation is y=mx+c, where c is the intercept and m are the coefficient.
- It means the two variables in x and y axis should be linearly correlated.
- It also some features on Residuals like it should be normally distributed, independent to each other and homoscedastic.
- It is the method to find best fitting line.

2. **Explain the Anscombe's quartet in detail. (3 marks)**

- It comprises of four datasets that have identical statistical properties but appear different when graphed
- It is useful to analyze relationship graphically before actual analysis.

3. **What is Pearson's R? (3 marks)**

- It is the measure of correlation coefficient between two variables in linear relationship. It helps in finding how well they are related. In simple, it is the numerical summary of the strength of the linear relationship of the two variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

- Scaling is the technique to standardize the independent features present in the data in a fixed range.
- It is done to avoid ML algorithms to misinterpret the values like 3000 Meters > 5 KMS which is not actually.
- Normalization is the feature to rescale value in the range 0 and 1.
- Standardized is the feature to re-scale the distribution with mean 0 and standard deviation of 1.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

- VIF is infinite sometime, if there is a perfect or exact correlation is expressed by the variables.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

- It is a graphical tool to assess if the data comes from some theoretical distribution like normal, exponential, and uniform. Also, it helps to find if two data set comes from same population.
- It is important in linear regression to find if train and test data are from same population in case if they are received separately.