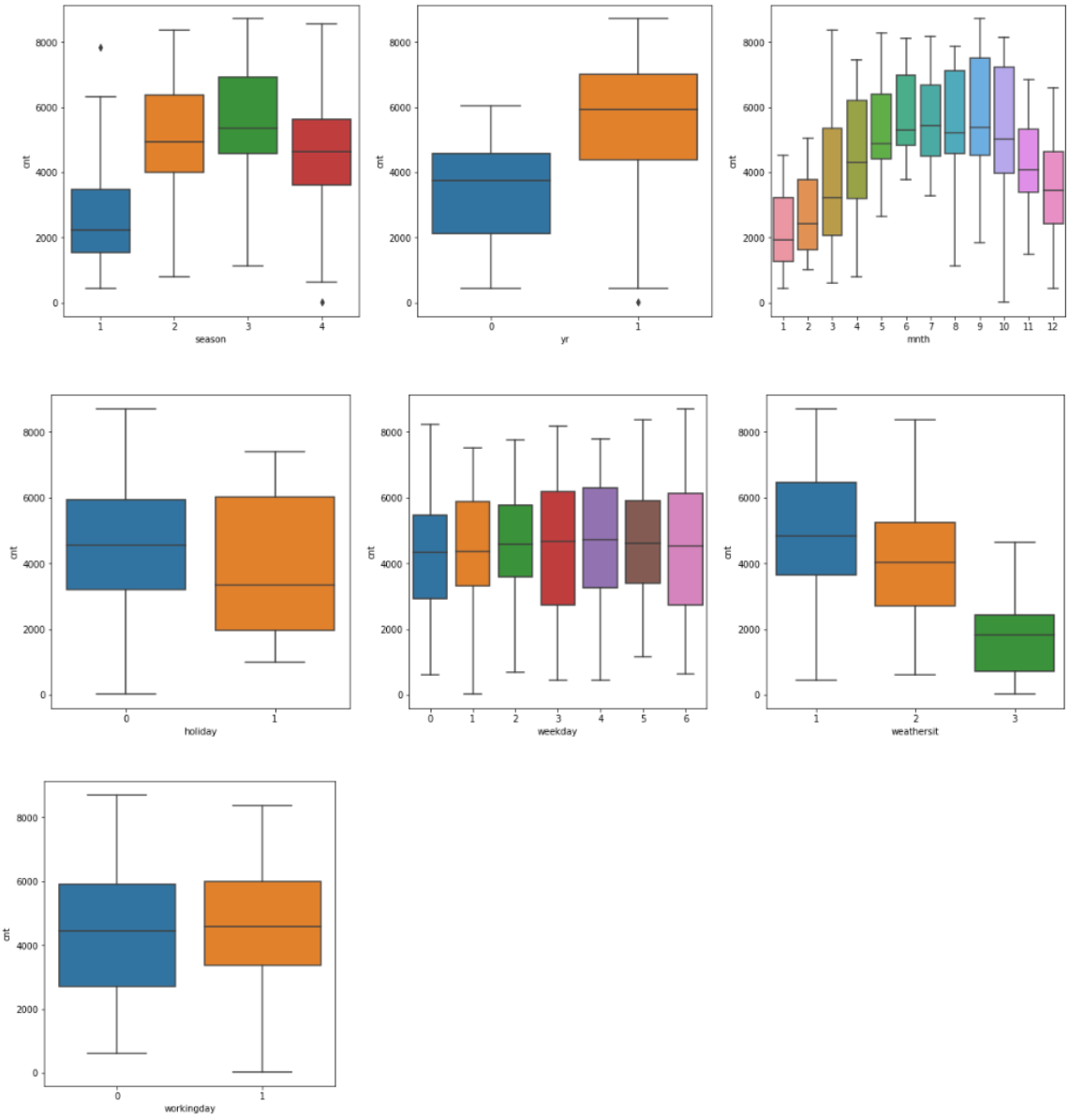
Assignment-based Subjective Questions:

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

ANS.



From the above box plots derived from the code in the uploaded python file, below can be concluded

- As we see in the box plots, there are enough up's and down's in the median of variables like 'season', 'yr', 'month' , 'weathersit' which indicates a significant influence over target variable 'cnt'.
- And also if we observe the median of variables like 'holiday', 'weekday', 'workingday' doesn't

varies much, indicating that they have lesser influence over target variable.

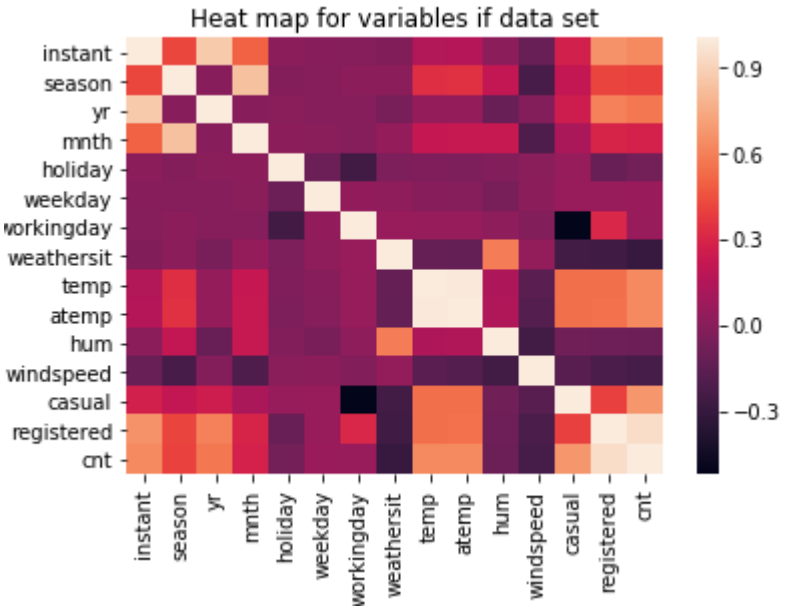
2. Why is it important to use drop_first=True during dummy variable creation?

ANS:

- Usually when we create a dummy variables extra columns will get generate which increases the correlation among variables.
- Therefore to reducing the extra column created during dummy variable creation we have to make **drop_first=True** which heps in reducing 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 above heat map we see that 'registered' variable has the highest correlation with the target variable 'cnt'.

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

ANS: For validating we basically use r2_score of the model. calculating the r2_score involves below steps

- Using the obtained linear regression model for predicting the target variable for the given test data.
- Once we obtain the predicted values of target variables, we use its original values and predicted values to calculate the r2_score of the model where we expect it to be as high as possible.

- For the present model we see the r2_score is **0.7815289564483857**

Note: Once we obtain the predicted values using our model, We perform residual analysis to calculate error terms or residuals where we expect their distplot should be a normal distribution centered around zero

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

ANS:

- Significance of the variable in the model building is usually calculated using p value where ideally a signficant variable will have its p value less than 0.05.
- So in our final model, based on the p values we see below are the top 3 features contributing significantly towards explaining the demand of the shared bikes

a. weekday

b.yr

c.weathersit

General Subjective Questions:

1. Explain the linear regression algorithm in detail ?

ANS:

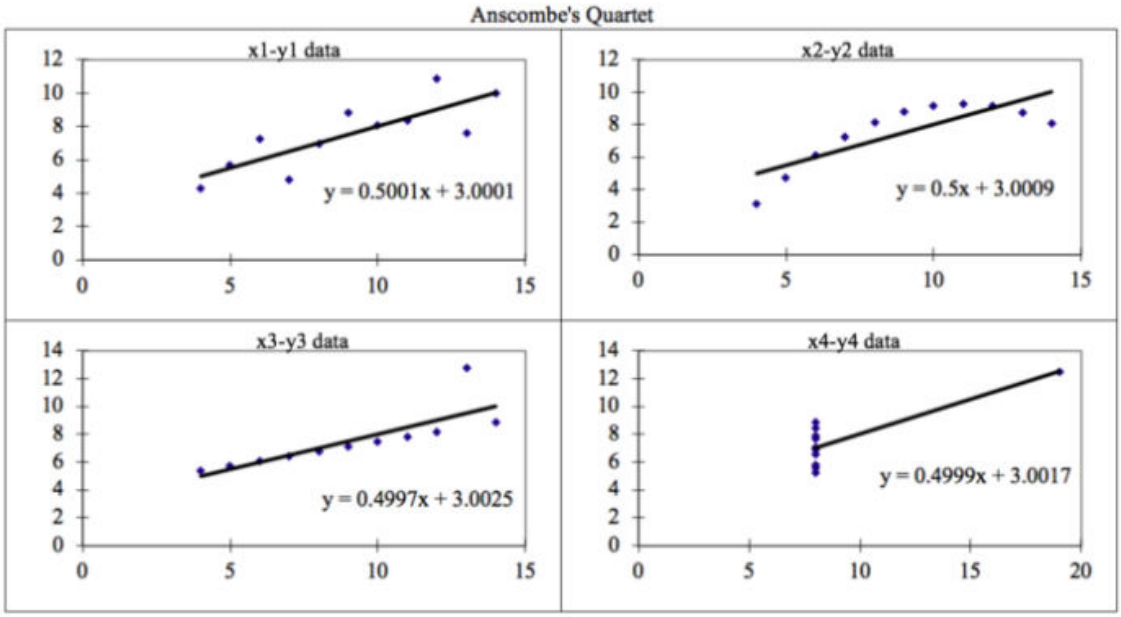
- Linear regression is a ML algorithm where we predict a target variable value (y) based on a given variables (x).
- So, Basically this is an approach of finding out a linear relationship between x and y. Hence, the name is Linear Regression.
- y = b1x + b0
- Based on number of variables we have 2 types of linear regression. One is simple linear regression and another is multiple linear regression.
- Our target should be finding the best b1 and b2 values, with which we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

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 like mean, variance etc, but there are some peculiarities in the dataset that fools the regression model if built.

- When plotted on scatter plots we find them very different as seen in below image.
- It helps us in understanding the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.
- Below are the four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



3. What is Pearson's R?

ANS:

- Pearson's r is nothing but a numerical summary of the strength of the linear association between the variables.
- If the variables tend to go up and down together, the correlation coefficient will be positive.
- If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
- The Pearson's correlation coefficient varies between -1 and +1 where:
- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association

- r > 8 means there is a strong association

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

ANS:

- It is a process where we normalize the data within a particular range.
- As the data sets usually contains variables with highly varying in values, 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.
- It helps in speeding up the calculations in an algorithm.
- "Normalized scaling": Brings all of the data in the range of 0 and 1.
- "Standardized scaling":Replaces the values by their Z scores thereby bringing all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

ANS:

- VIF = infinity happens If there is a perfect correlation. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
- To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

ANS:

- Q-Q Plots (Quantile-Quantile plots) are nothing but plots of two quantiles
- If two sets of data come from the same distribution then it can be verified using these Q-Q plots .
- A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution.