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:

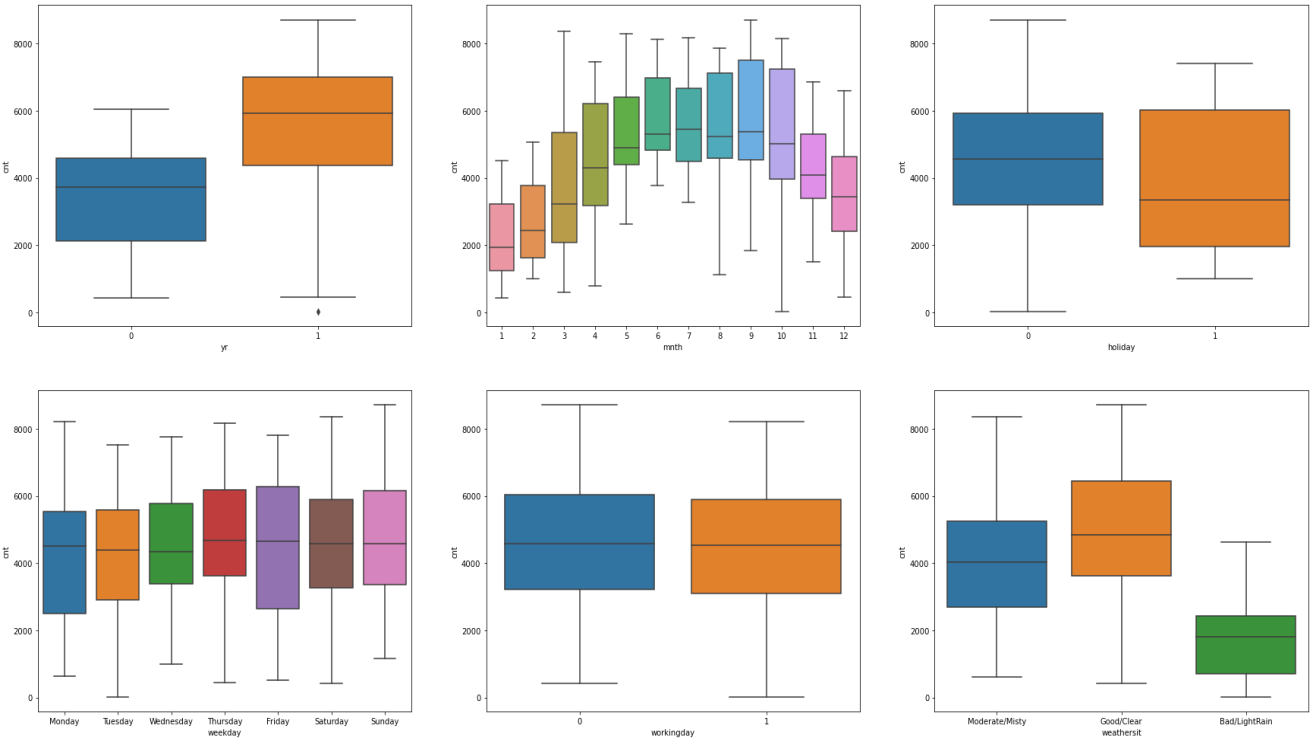
As per analysis we can conclude effect on the dependent variable is:'cnt'

Demand and hiring of bike is more in the month of winter season as compared to other season.

In the year of 2019 the bike rental demand is more as compared to year 2018 as we see the higher the median.

Clear weather is most desire for renting bike because the humidity and temperature is less.

As we see on working day the renting bikes is more as compare to Non working day.



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

ANS:

It is important to use because it helps to reducing the extra columns which created during the dummy variable creation.

drop_first=True it is very important in reducing the dimensionality

It also reduces the correlation among the dummy variables.

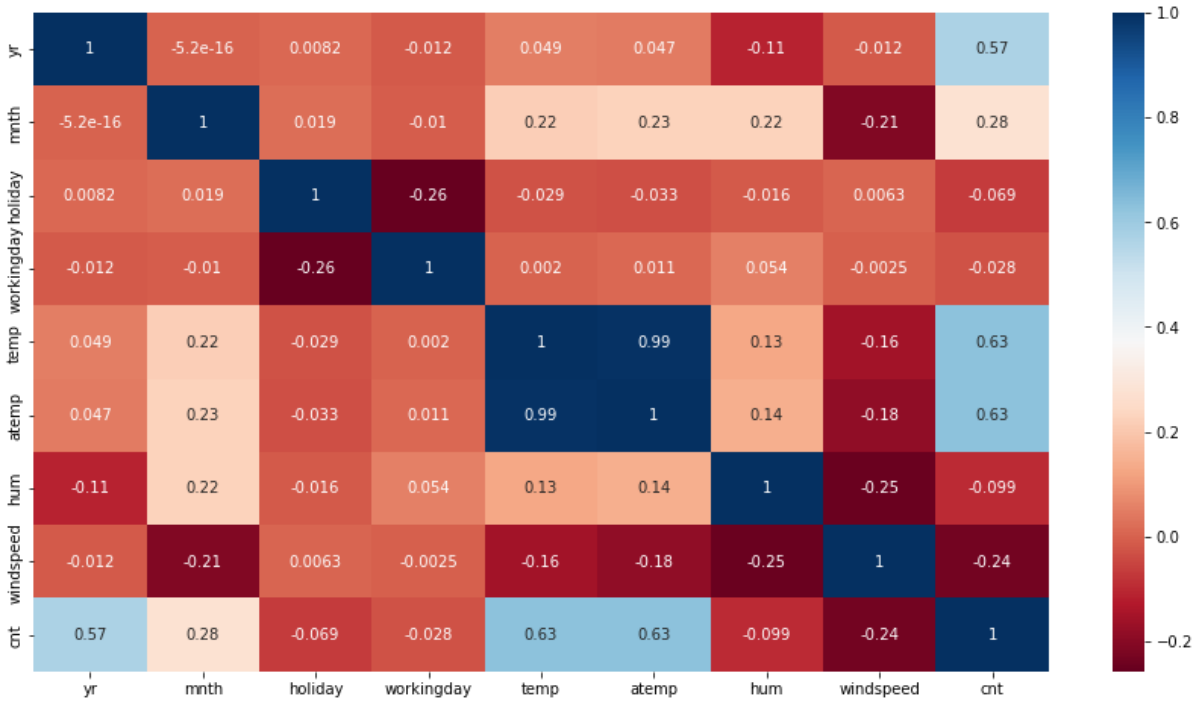
Drop_first= Ture drops the first column during dummy variable creation. For ex: consider you have a column for gender that contains 4 variables- Male, Female, Trans, Unknown. So a person is either Male or Feamle, or Trans. If they are not either from 3, their gender is Unknown.

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

ANS:

Both temp and atemp have same correlation with target variable of 0.63 which isthe highest

Among all numerical variables. As u can see in the correlation dia:



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

ANS:

Predictors (x) Are Independent and Observed with Negligible Error.

Residual Errors Have a Mean Value of Zero.

Residual Errors Have Constant Variance. If the data points on the graph near to the best fit line form straight line that means it met with the assumption.

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

ANS:

season_summer and season_winter feature plays major role in demand and hiring the rental bikes because if the weather is good and clear people will step out and rent a bike as much as they can .

weathersit_Good/Clear is also contributing features as the weather clear and good so less humidity and less temperature not raining and thunderstorm then it increased the demand and hiring the rental bikes.

Bike rents are increasing year on as year 2019 has higher median than 2018, it might be due to the fact that bike rentals are getting popular day by day and people are becoming more aware about environment.

General Subjective Questions

1.Explain the Linear Regression algorithm in detail.

Ans:

Linear regression is a machine learning algorithm classified in to supervised learning which use the labelled datasets to train algorithms that to classify data or predict outcomes accurately.

Regression models target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Linear regression is a machine learning technique in which model predicts the output as a continuous numerical variable.

Linear regression is commonly used for predictive analysis and modelling.

For example:

It can be used to quantify the relative impacts of age, gender, and diet on height.

The goal of linear regression is to build an artificial system that can learn the mapping between the input and the output, and can predict the output of the system given new inputs.

2.Explain the Anscombe’s quarter in detail.

ANS:

Anscombe's Quartet can be defined as a bunch of four applied data sets which are identical in informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population

Anscombe’s Quartet it plays very important role in highlighting data points and confirms the validity of fit model.

Anscombe’s demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

3.What is pearson’s R?

ANS:

Pearson's R is a numerical summary of the strength of the linear association between the variables.

Pearson’s R is the test statistics that measures the statistical relationship between two continuous variable . It is the best method to cal relationship between two variable because it is based on method of covariance .

Pearson’s R is the most common way of measuring a linear correlation it is between -1 and 1 which measures strength and direction.

The values of R ranges from -1 to 1. If the R=0 then it means there no linear relationship between two continuous variable.

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

ANS:

Scaling is the process which is transformed the original data into scalable manner so that data fit in moderate or mention scale .

Scaling is used for the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.

Normalized scaling means rescales the values into a range of 0 to 1. It is also called as min max scaling.

Normalized scaling is useful when there are no outliers as it cannot cope up with them.

Standarddized scaling is another scaling technique where the values are centered around the mean with a unit standard deviation.

It the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

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. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

6.What is 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

In Q-Q plot the sample sizes do not need to be equal. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.\

Whenever we are interpreting a Q-Q plot, we shall concentrate on the 'y = x' line.

We also call it the 45-degree line in statistics. It displays that each of our distributions has the same quantiles. In case if we witness a deviation from this line, one of the distributions could be skewed when compared to the other.