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?

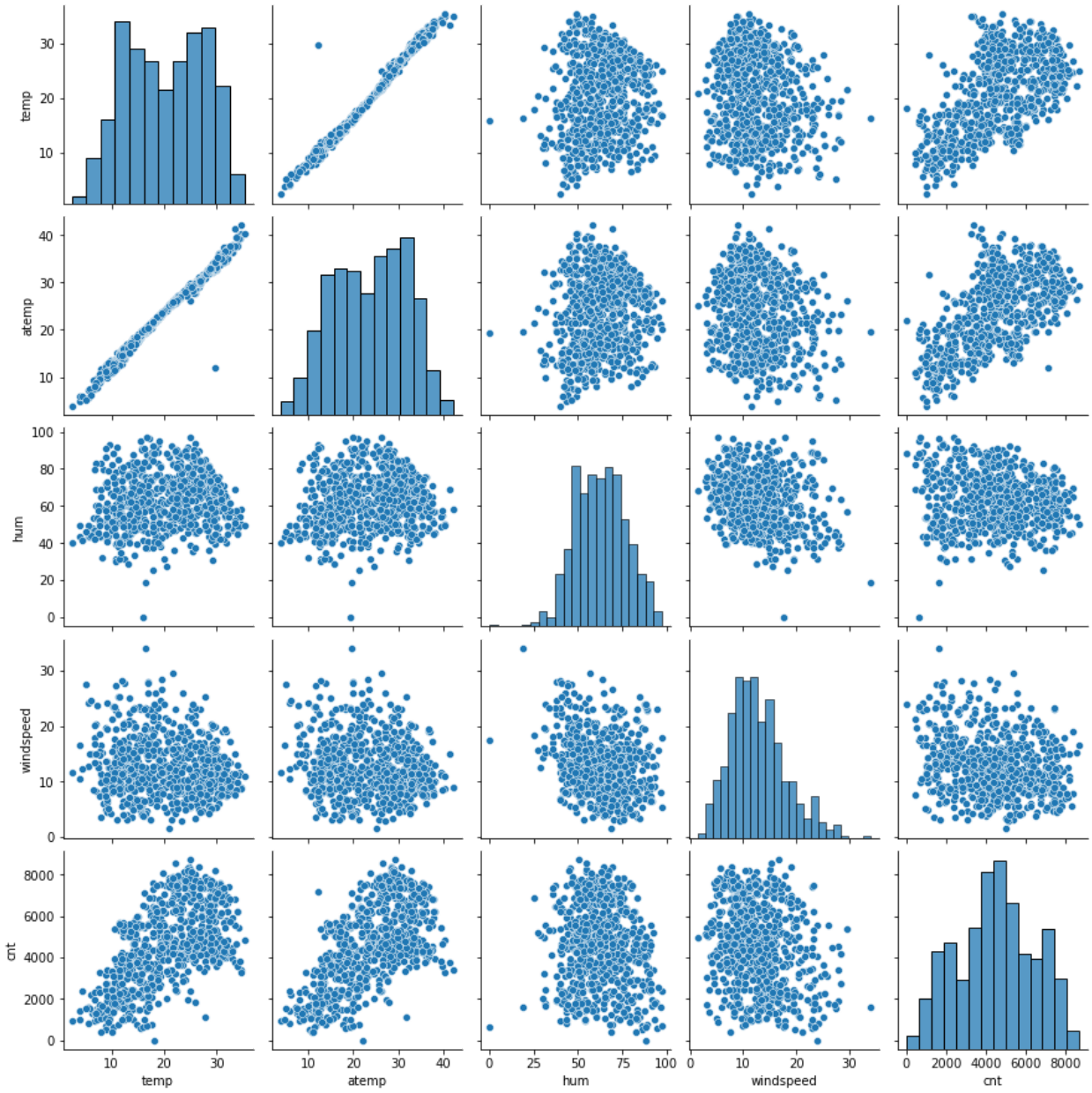
In the given data set we have 'season','mnth','weekday','weathersit','yr and ‘workingday’ as categorical Variables. We have visualised it using boxplot and have reached the following conclusions:

- ‘Season’: From the box plot it is clear that spring has the least count and fall has the highest count on the bike rentals
- ‘Mnth’: Month september has the highest count of bike rentals according to the box plot
- ‘Yr’:2019 has more bike rentals compared to 2018
- ‘weathersit’:From the box plot it is clear that there won’t be any bike rentals in extreme climate (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)
- ‘Workingday’: There is a slight decrease in the rentals on holidays

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

When we create dummy variables , the variables created are highly correlated and it becomes difficult to interpret the predicted coefficients of variables from the regression models. So inorder to avoid that we use drop_first .

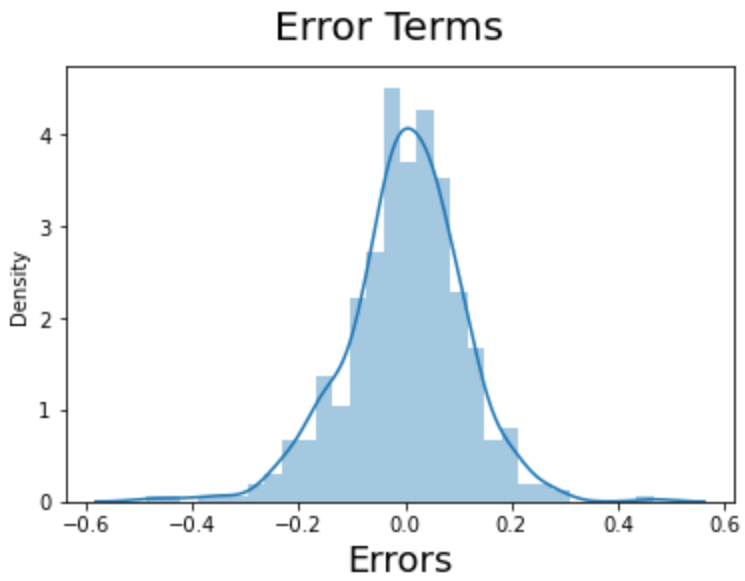
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



From the above figure it is clear that the temp and atemp are the highly correlated variables

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

According to the assumptions of Linear Regression the errors should be normally distributed. We validated it by plotting the errors using the distplot and verified whether it is following a normal distribution or not



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

Based on the final model the top 3 features are:
Yr . coefficient 247068
weathersit_lightSnow/lightRain&thunderstrom : coefficient -0.197202
season_spring : coefficient -0.268305

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a machine learning algorithm for predicting the target variable(y) from independent variables(X) . It is based on the equation “y=mx+c”. Linear Regression is based on the assumption that there is a linear relationship between the dependent variable and the independent variables.It tries to find the best fit line which explains the relationship between x and y with minimum errors.

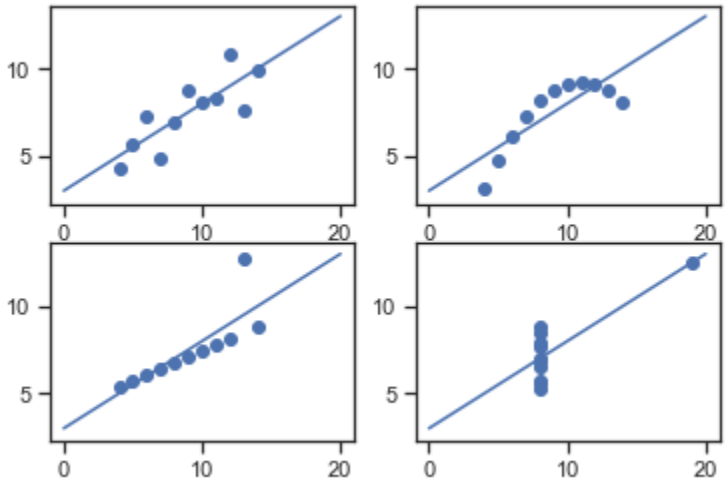
Linear Regression can be classified into simple linear regression and multiple linear regression

In simple Linear Regression there will be only one independent variable to predict the dependent variable. Multiple linear regression comes in to picture when there are more than one independent variables to predict the dependent variables

The equation of multiple linear regression is
$Y=\beta 0+\beta 1X1+\beta 2X2+.....+\beta pXp+\epsilon$

2. Explain the Anscombe's quartet in detail

Anscombe's quartet helps us to understand the importance of data visualization along with statistical results .It consists of four data sets and each data-set consists of eleven (x,y) points**.** These data sets have almost the same descriptive statistics but have a very different distribution when plotted in graphs.



Plot 1: It shows a linear relationship between x and y
Plot 2: It shows a normal distribution between x and y
Plot 3:It shows a linear relationship between x and y except for an outlier
Plot 4: It shows that values of x is a constant except for one outlier

3. What is Pearson's R

Pearson's correlation coefficient denoted by r is the numerical summary of the statistical relationship or association between two continuous variables. It varies from -1 to 1.Magnitude of association or correlation and direction of relationship can be inferred from it.

r = 1 indicates the data is perfectly linear with a positive slope
r = -1 indicates the data is perfectly linear with a negative slope
r = 0 indicates there is no linear association.

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

Scaling is the process of standardizing or normalising the independent variables present in the data set.It should be done for correct prediction and result . If scaling is not performed on the data set variables with high magnitude tend to weigh more than variables with low values even if that variable has more weightage in predicting the output. And also some of algorithms will take less execution time when scaled.

Normalized scaling scales the value between -0 to 1 or -1 to 1.It uses the min and max values of scaling and is affected by outliers.We choose normalized scaling when we don't know about the distribution.

In Standardized scaling, mean and standard deviation are used for scaling and thus it is not affected by outliers.It doesn't have a bounded range . We choose standardized scaling when we know that the distribution is gaussian or normal

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

VIF, Variance inflation factor is the measure of amount of multicollinearity in multiple linear regression.VIF=1/1-R^2 In the case of perfect correlation, we get R^2 =1, which lead to 1/(1-R^2) infinity.That is if there is perfect correlation, then VIF = infinity.

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.It is mainly used to identify if two data set come from population with same distribution,,have similar distributional shapes and have similar tail behaviour.The q-q plot can provide more understanding of the nature of the difference of two data set than analytical methods