Assignment-based Subjective Questions

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

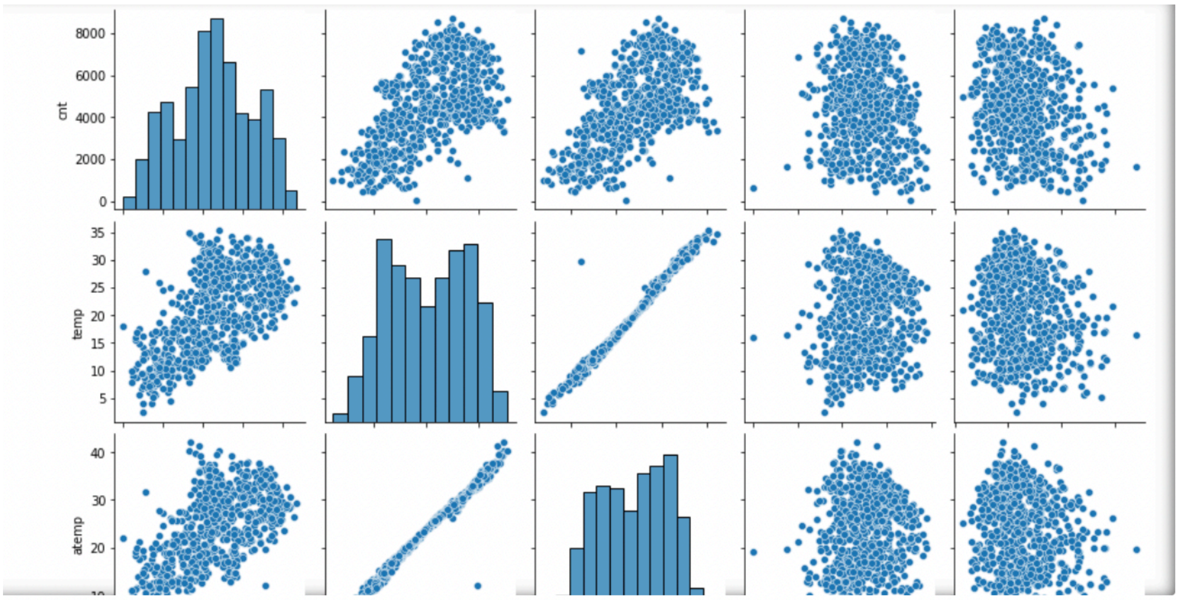
Answer 1 :
The categorical variable in the dataset were season,weathersit,holiday,mnth,yr and weekday.These were visualized using a boxplot .These variables had the following effect on our dependant variable:-
1. Season - in spring season cnt decreased but in fall had maximum value of cnt. Summer and winter had intermediate value of cnt.
2. Weathersit - There are no users when there is heavy rain/ snow . weathersit ’ Clear, Partly Cloudy’, cut increases.
3. Holiday - rentals decreases in holidays.
4. Mnth - In September no of rentals was more but in December less.
5. Yr - The number of rentals in 2019 was more than 2018

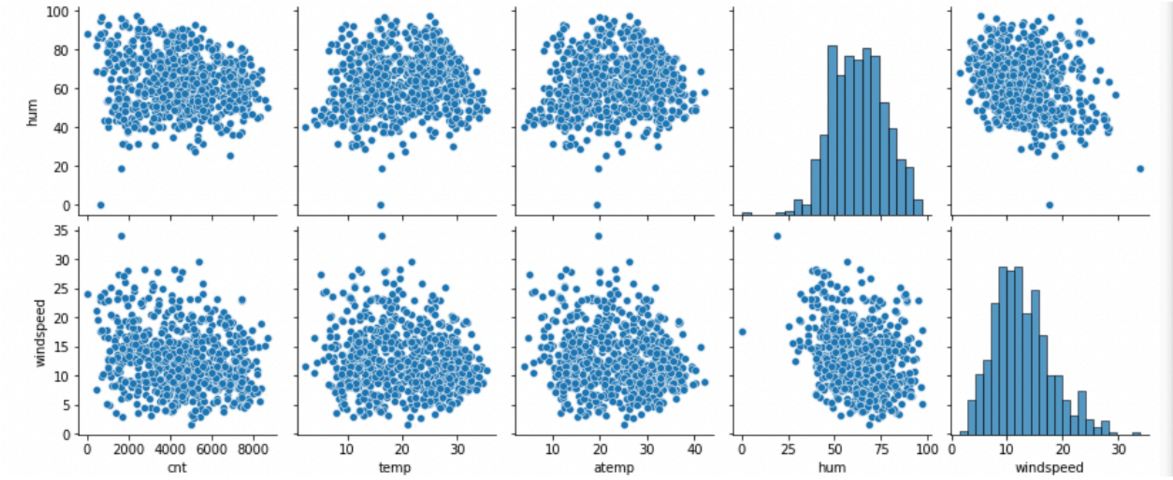
Question 2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer 2:
Since no of dummy variables to be created is N-1 where N levels are there in a categorical variable. After the creation of dummy variables, N-1 columns are created and now the actual categorical column becomes insignificant, so needs to be dropped as this might increase the correlation with all the other dummy variables and thus will have high p value and VIF.

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

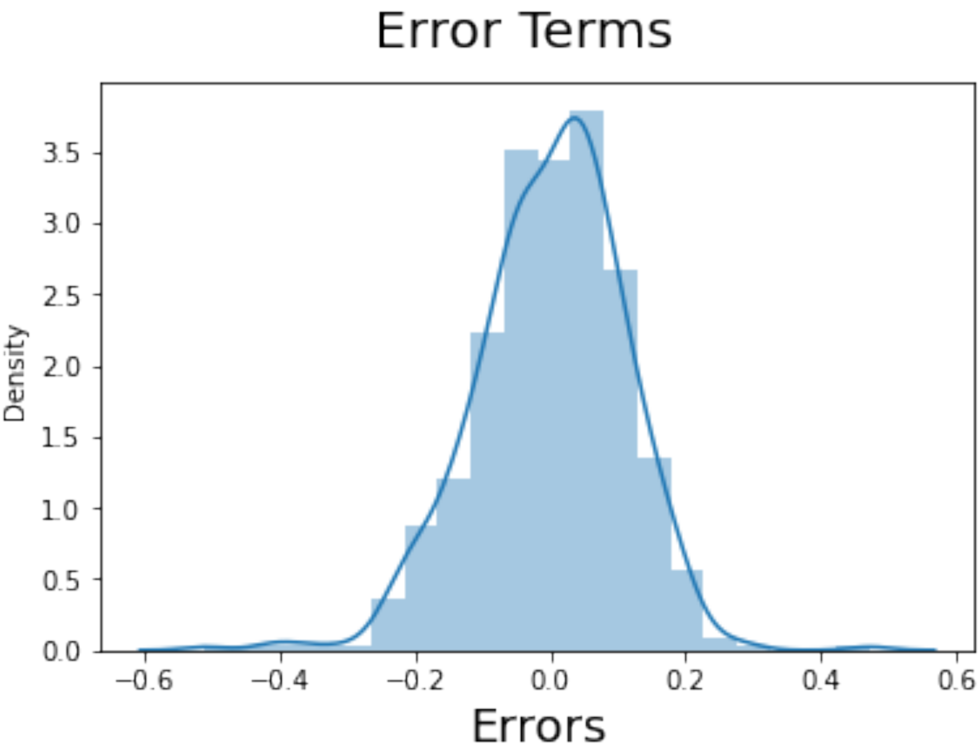
Answer 3:





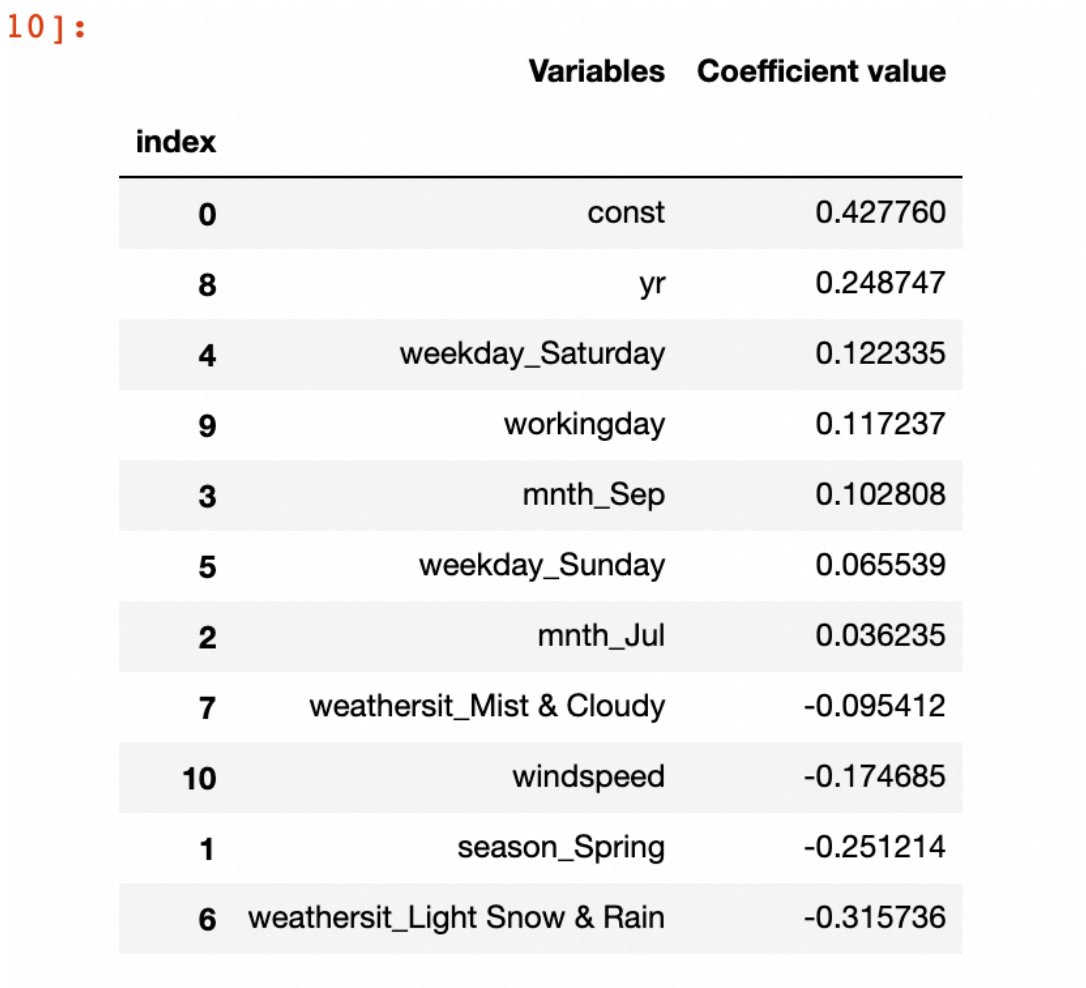
Temp and temp
We validate it by by plotting a graph distplot - residuals then check if residuals are having normal distribution or not.
We can see in the graph that the residuals are distributed around mean = 0.

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



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

Answer 5:



10]:

index	Variables	Coefficient value
0	const	0.427760
8	yr	0.248747
4	weekday_Saturday	0.122335
9	workingday	0.117237
3	mnth_Sep	0.102808
5	weekday_Sunday	0.065539
2	mnth_Jul	0.036235
7	weathersit_Mist & Cloudy	-0.095412
10	windspeed	-0.174685
1	season_Spring	-0.251214
6	weathersit_Light Snow & Rain	-0.315736

Based on my model building these are: yr, weekday_Saturday and workingday

General Subjective Questions:

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

Answer 1:

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values.Linear Regression is the most basic form of regression analysis.Regression is the most commonly used predictive analysis model. Linear regression is based on the popular equation "y = mx + c". It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term. Regression is broadly divided into simple linear regression and multiple linear regression. 1. Simple Linear Regression : SLR is used when the dependent variable is predicted using only one independent variable. 2. Multiple Linear Regression :MLR is used when the dependent variable is predicted using multiple independent variables.

Question 2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer 2:
Anscombe’s Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph.It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties The first scatter plot (top left) appears to be a simple linear relationship. The second graph (top right) is not distributed normally; while there is a relation between them,it’s not linear. • In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

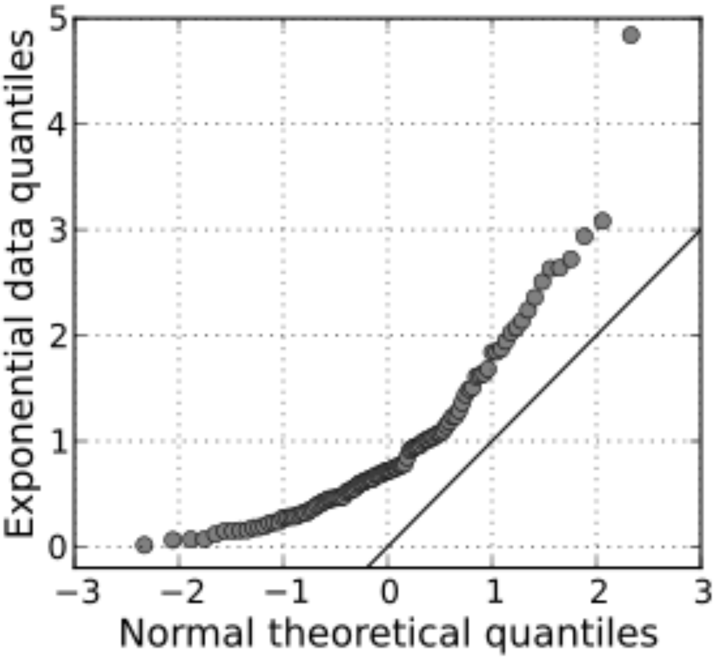
3. What is Pearson’s R? (3 marks)
Answer 3:
Pearson's r is a numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data? r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association
It s a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

Question 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer 4:
Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values. Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks. Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also,

unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

Question 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Answer 5:
VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.(VIF) =1/(1-R_1^2). If there is perfect correlation, then VIF = infinity.Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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



In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.[1] First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.