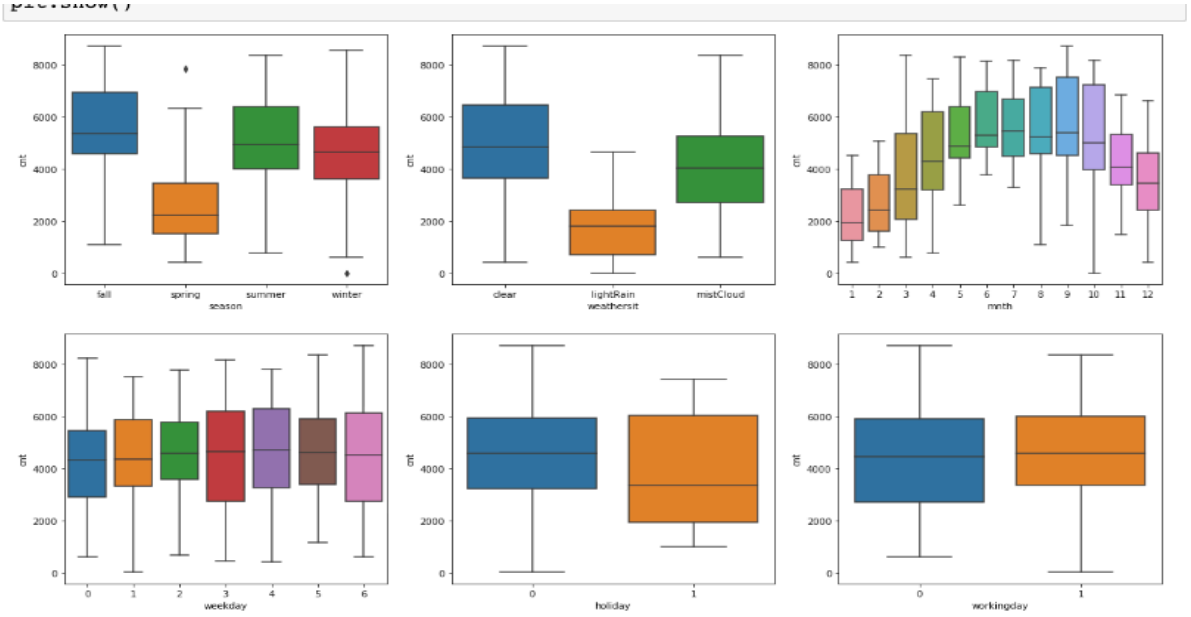
Assignment-based Subjective Questions

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

Ans: Season, month, weekday, weather situation, holiday , working days are categorical variables. These categorical variables influence the dependent variable in following ways: -



i)Season: Spring season had lowest value of cnt value and fall season is highest value of cnt value. Summer and winter have median almost to 5k-6k.

ii)Weather situations: Bikes sharing (cnt) is more when weather clear and mild cloud .

iii)Month: September is having more rental while January is less .

iv)Weekday : Median value of Bike sharing (cnt) is almost same in weekdays. weekday is very less influence towards dependent variable

v)Holiday: Rental found lower during holiday.

vi)Working day: Rental more on working day

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

Ans: Creating dummy variables from categorical data, the parameter **drop_first=True** is used to prevent multicollinearity in regression analysis and to improve the interpretability of the model. Let's break down why this parameter is important and how it addresses these issues:

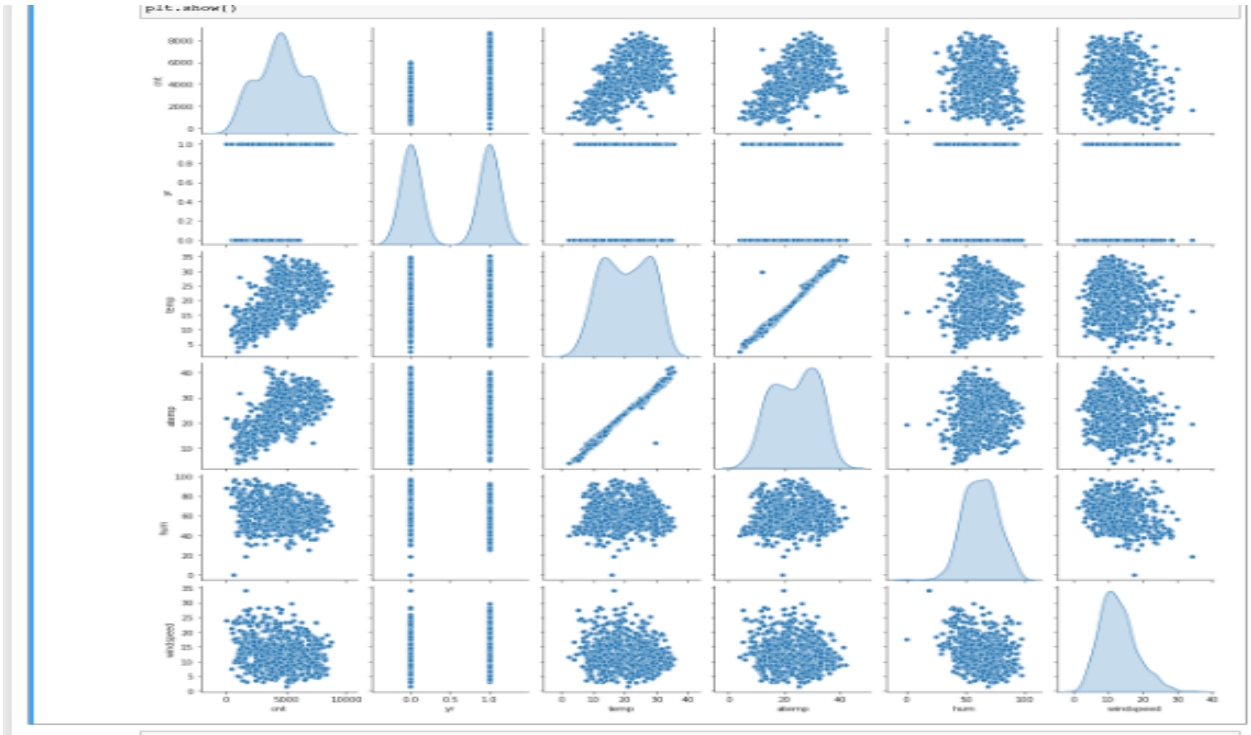
i)Multicollinearity:

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated.

ii)Interpretability:

Including all dummy variables in a regression model can make it difficult to interpret the effects of each category. By dropping the first category, you establish a reference point against which the effects of the other categories can be compared. This makes it easier to understand the impact of each category on the response variable.

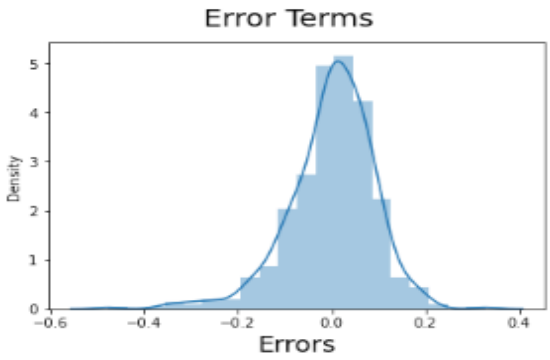
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



temp and atemp are two parameter which is highly correlated with cnt.

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

Ans: The distribution of residual should be normal and centered around 0. The residual are centered around 0 as its clear from below figure.




```
In [ ]: ##From the above histogram, we could see that the Residuals are normally distribute
```

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

Ans:- Below are top 3 predicator variables that influencing the bike variables .

Temperature(temp): with coefficient 0.374182. with unit increase in temperature bike share wll increase by 0.374182 times.

Weather situation(Weathersit) : with coefficient -0.327160.if weather situations are of type (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) with u nit
increase in weathersit variable reduces bike sharing by -0.327160.

Year: with coefficient 0.235293. with unit increase in year bike sharing increasing by 0.235293.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

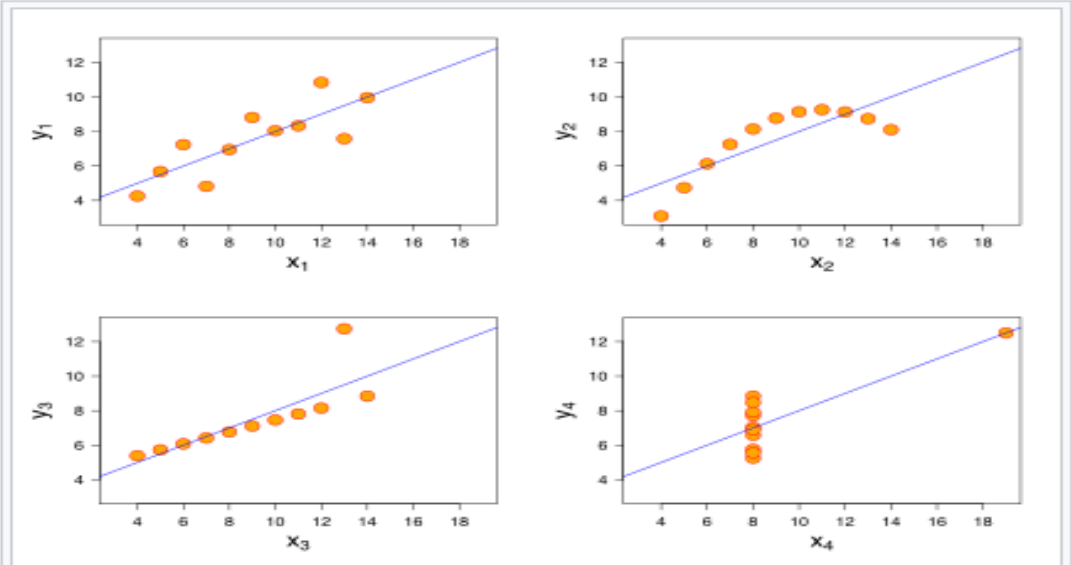
Ans: - Linear Regression is a type of supervised Machine learning algorithm that is used for prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly use predictive analysis model. Linear regression is based on the popular equation "y=mx+c".
It is assumed that there is a linear relationship between the dependent variable(y) and the independent variable(x). in regression, we calculate the best fit line which describes the relationship between independent and dependent variable. Regression is performed when dependent is of continuous datatype and predictors or independent variable could be of any data type like continuous, categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with last 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 2 types: -

1. Simple liner 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.

   The equation for MLR will be: -
   Y=B0+B1X1+B2X2+B3X3+…...+BnXn.
   B1=coefficient for x1 variable.
   B2=coefficient for x2 variable.
   B3=coefficient for x3 variable.
   B0 is the intercept (constant term).

2.Explain the Anscombe's quartet in detail.

Ans:- Anscombe's Quartet was developed by statistician Francis Anscombe.it includes four data dets that have almost identical statistical features, but they have a very different distribution and looks 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 influencing observations on it.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Statistical properties: -
1)The first scatter plot appear to be simple and linear relationship.
2) The second graph is not distributed normally. There is a relationship between them but it's not linear.
3)Third graph the distribution is linear but there is outlier in that.
4)Fourth graph: All the Y values related to same x.

3)What is Pearson's R?

Ans: - Pearson's r is numerical representation of the strength of the linear relationship between the variables. Its value ranges from -1 to. +1. It depicts the linear relationship of two sets of data. In simple term it asks if we can draw a line graph to represent the data.
r=1 means the data is perfectly linear with positive slope.
r=-1 means the data is perfectly linear with negative slope.
r=0 means there is no linear association.

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

Ans: Feature scaling is a method 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.

a) 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-Neighbors.
b) Standardization on the other hand can be helpful in case where the data follows a gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have an bounding range. So even if you have outliers in your data they will not be affected by standardization.

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

Ans: - The VIF indicates how much collinearity has increased the variance of the coefficient estimate.(VIF) is equal to 1/(1-Ri2). VIF = infinity if there is perfect correlation. Where R-1 denotes the R-square value of the independent variable for which we want to see how well it is explained by other independent variable. if an independent variable can be completely described by other independent variable, it has perfect correlation and has a R-square value of 1. As a result VIF=1/(1-1) which is infinity .

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

Ans: - The Quantiles of the first data set are plotted against the Quantiles of the second data set in a q-q graphic. It's a tool for comparing the shapes of different distributions. A scatterplot generated by plotting two sets of quantiles against each other is known as Q-Q plot.

Because both sets of quantiles came from the same distribution, the points should form a line. The q-q plot is used to answer the following questions: -

i) Do two data sets come from populations with a common distribution?
ii) Do two data sets have common location and scale?
iii) Do two data sets have similar distribution shapes?
iv) Do two data sets have similar tail behavior?