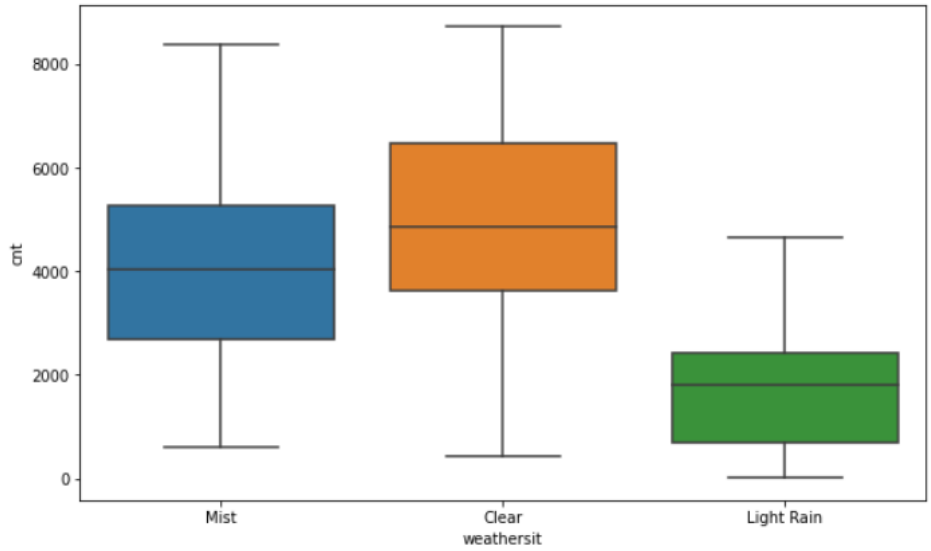
Assignment-based Subjective Questions

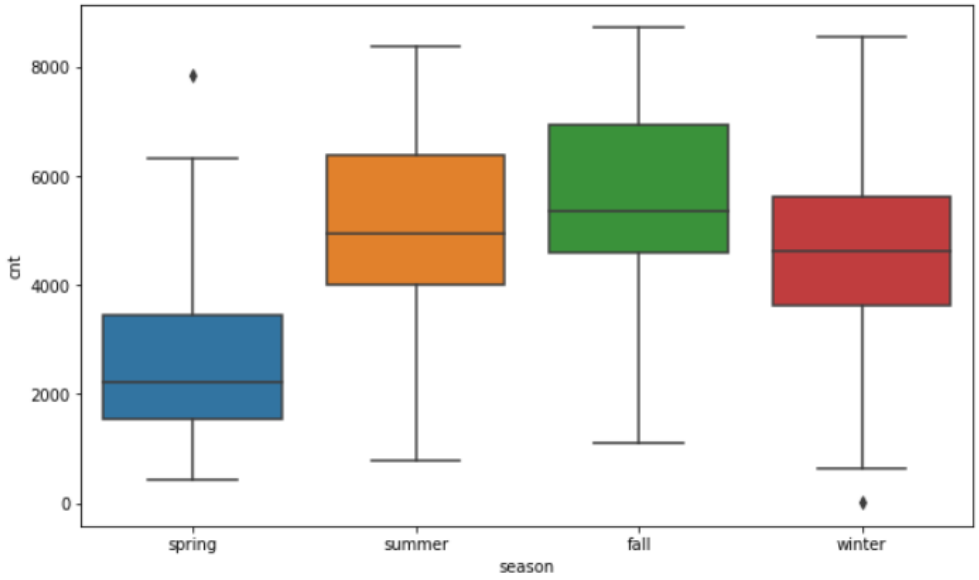
1. *From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?*

A. The categorical variables in the dataset are weathersit, season, mnth and weekday.

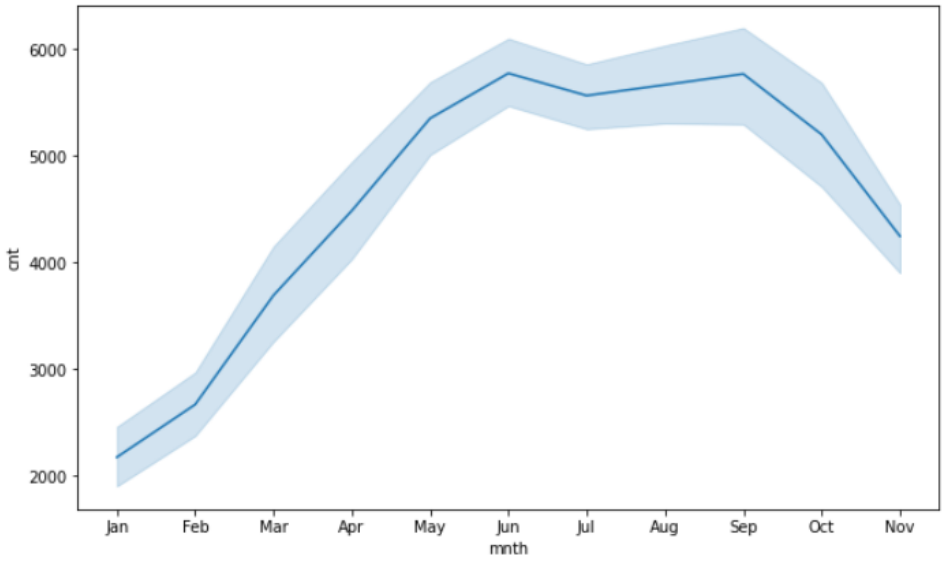
Weathersit: Count of rentals (cnt) are lower on Rainy day and highest on clear sky.



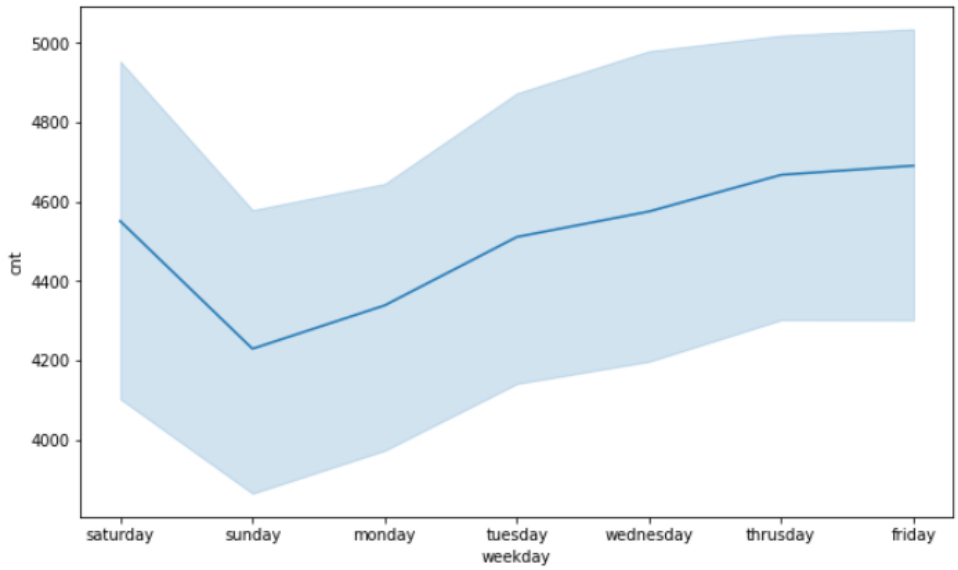
Season: Count of rentals (cnt) are more in fall and less in spring season.



Mnth: Count of rentals are more in months of June and September and least in December due to heavy snow.



Weekday: Count of rentals(cnt) more on Fridays and least on Sundays.

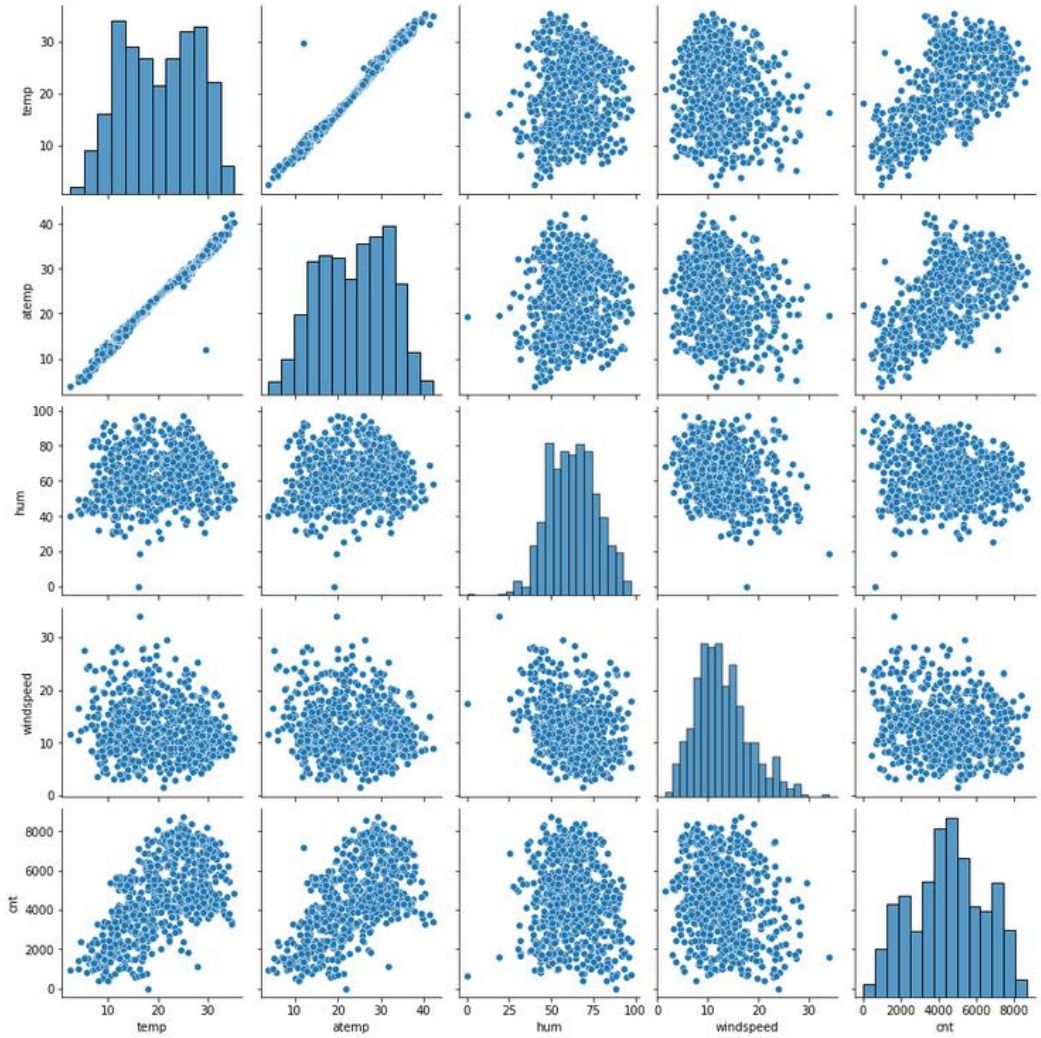


2. Why is it important to use drop_first=True during dummy variable creation?

A. For each categorical variable with levels m we take m-1 dummy variables because we only need m-1 variables to fully explain all the information related to that categorical variable. If we don't drop the first variable then the data become redundant.

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

A. The numerical variables are Temp, atemp, humidity, windspeed and target variable is cnt.

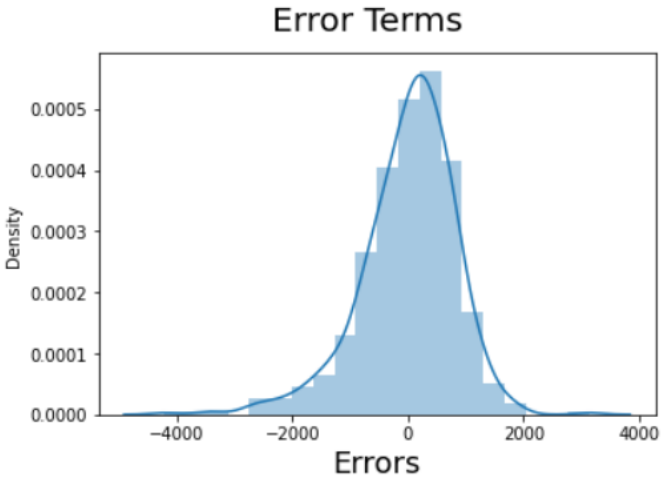


temp and atemp has strong positive correlation with count of rentals(cnt).

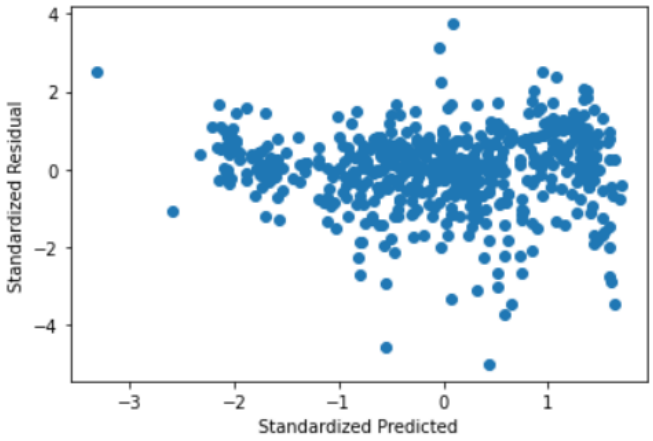
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A. The assumptions of Linear Regression are

1. Residuals must be normally distributed and centred around 0. We check this by plotting distplot on residuals.



2. The variables must be independent of each other and should not have multicollinearity. We can test this using variable Inflation factor(VIF) which needs to be below 5.
3. Homoscedasticity : there should not a any pattern on the plot between standardized residuals and standardized predicted values.



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

A. Top three features are :

1.temp
2.yr
3. weathersit_Light Rain

General Subjective Questions

1.Explain the linear regression algorithm in detail.

A. Linear regression is supervised machine learning algorithm that can be applied on continuous numerical or discrete numerical variables which is widely used in Predictive analytics and forecasting .It works on principle of line equation **y=mx+c.** where c is the intercept of line at origin , m is the slope and x is and independent variable and y is the dependant variable.
Linear Regression Types:

1. Simple Linear regression (1 Independent variable)
2. Multiple Linear Regression (more than one independent variable)
3. Polynomial linear regression
4. Bayesian Linear regression etc.

Equation of Multi-Linear Regression

$$Y_n=B_0+B_1X_1+B_2X_2+\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots+B_{n-1}X_{n-1}+B_nX_n$$

B_1 = Coefficient of variable X_1
B_2= Coefficient of variable X_2
$B_{n=}$ Coefficient of variable X_n
B_0= Constant

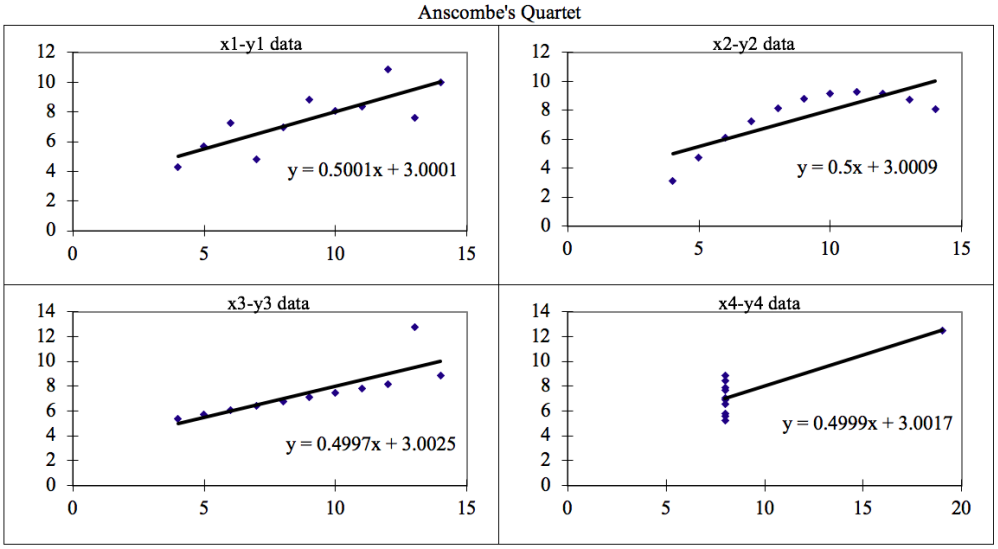
Assumptions:

1. Linear Relationships.
2. No Multicollinearity
3. Residuals must be normally distributed.
4. No auto correlation
5. Homoscedasticity

2. Explain the Anscombe's quartet in detail.

A. 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?

A. 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 ofdata. 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 sloper = 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?

A. 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 thinks that greater values have more importance than smaller values. Normalization typically means **rescales the values into** a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

A. 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 tocheck 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"

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

A. 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. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.