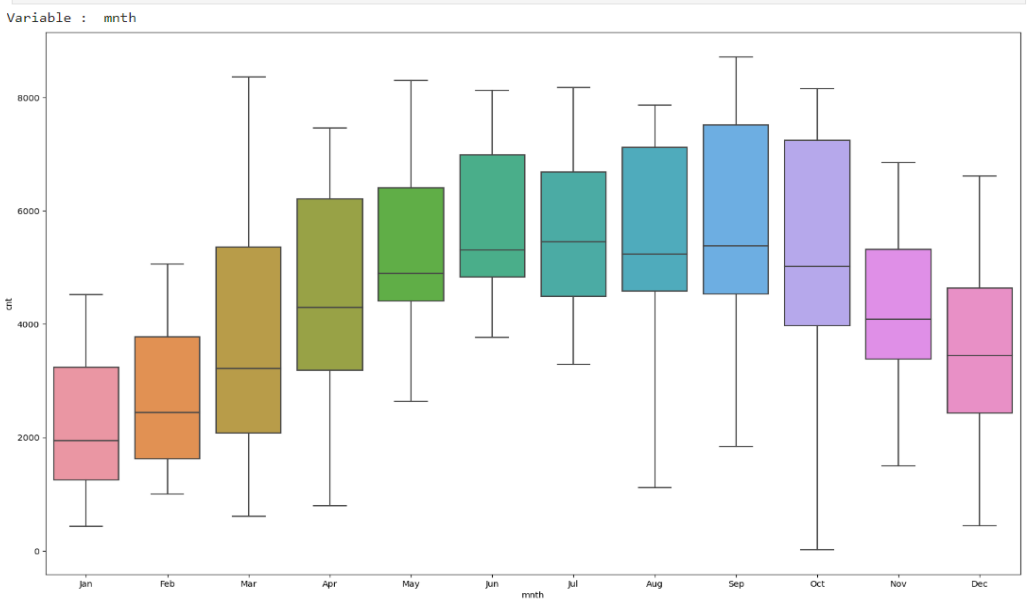
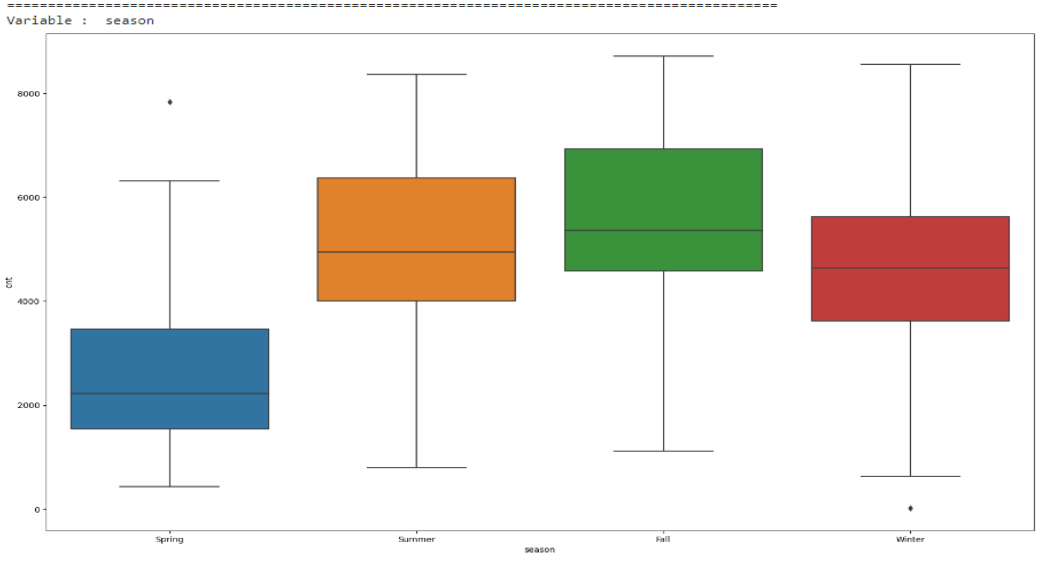
Assignment-based Subjective Questions

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

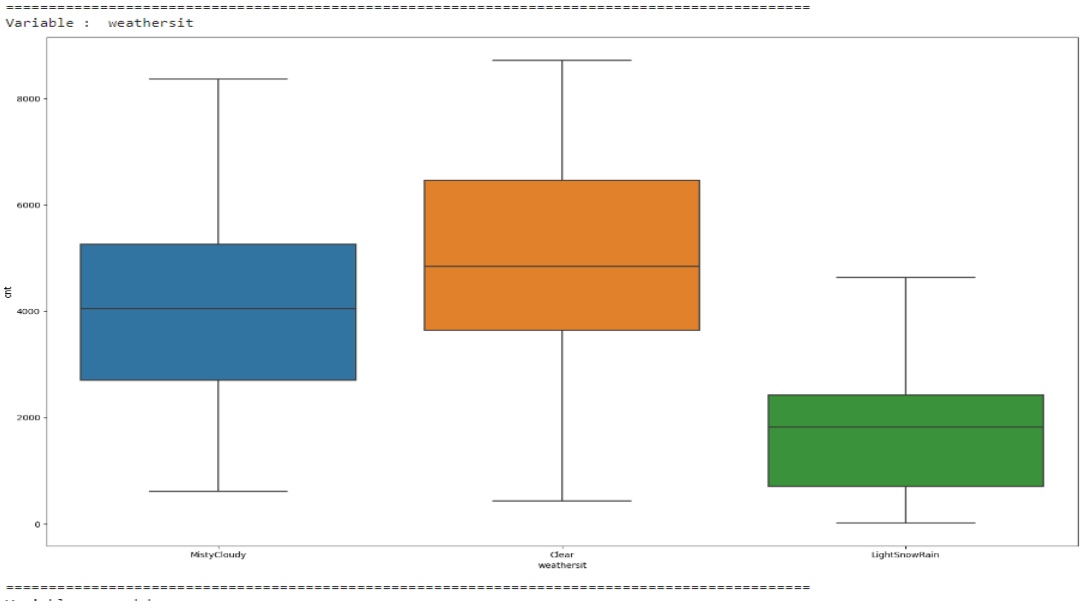
- **mnth** : Usage of rental bikes is very high From May month to September and demand decreases from November to April.



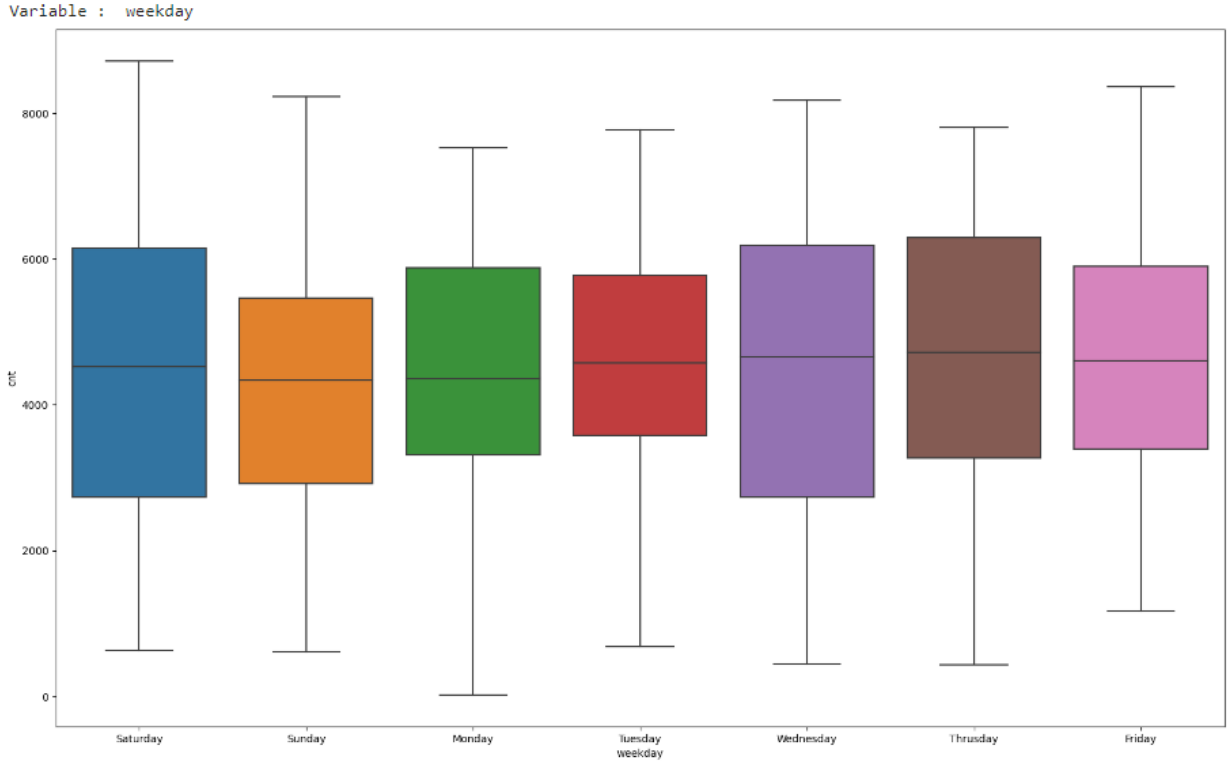
- **Season** : Rental bike usage is very high in summer and fall season where as very less in spring season.



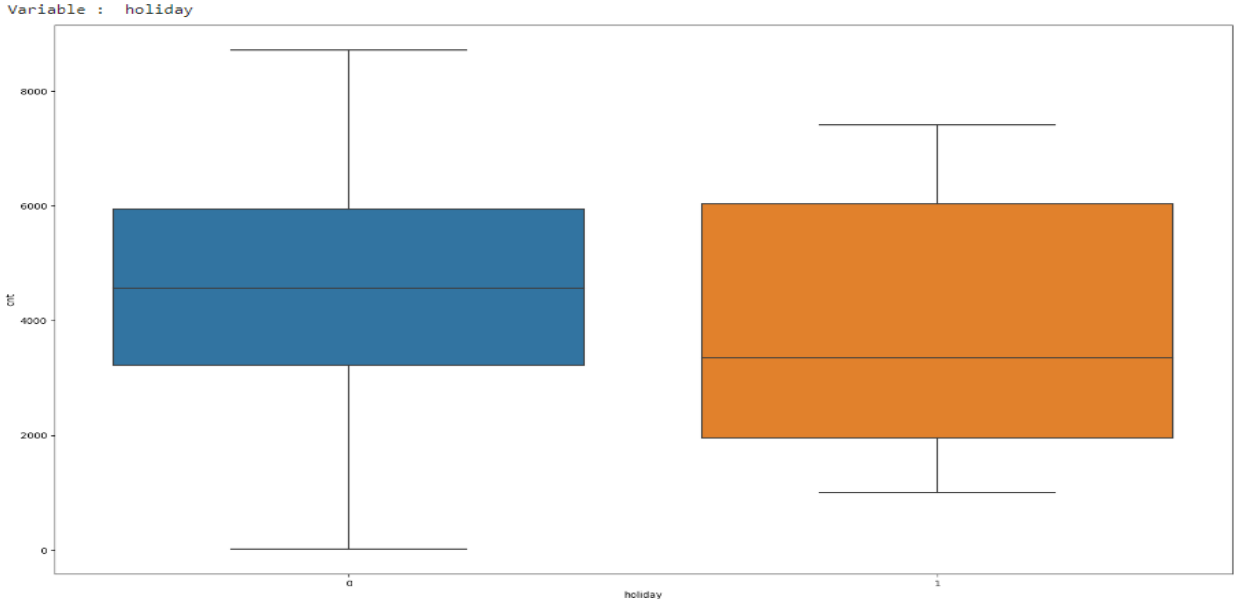
- **Weathersit**: Usage of rental bikes is higher in clear weather conditions like clear sky/partially cloudy where as the usage is very less during adverse conditions like Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.



- **Weekday**: comparatively bike usage is slightly less on Sundays ,



- **Holiday :** Clearly the deman for the bike is less on holidays.



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

Answer:

The **drop_first=True** argument is used to avoid the multicollinearity in the data which is nothing but two independent variables having high correlation between them and one variable can be predicted from the other.

If you include all dummy variables, they can perfectly predict the constant term. For example, lets consider season variable which has 4 categories(Spring,Summer,Fall,Winter) if we know that season=Spring or season =Summer or season=Fall, if its not then season is winter ,No need to specify all the categorical values otherwise The perfect relationship among the dummy variables leads to multicollinearity, making the regression coefficients unstable and difficult to interpret.

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

Answer:

Feeling temperature (0.63) and Temperature in Celsius (0.63) has highest correlation with target variable.

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

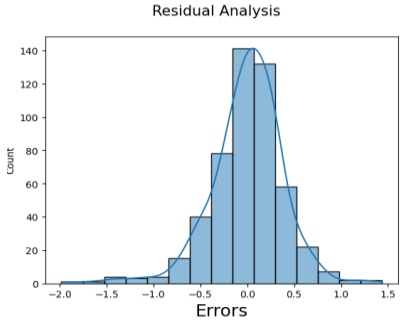
Following are the assumptions to build linear regression model.

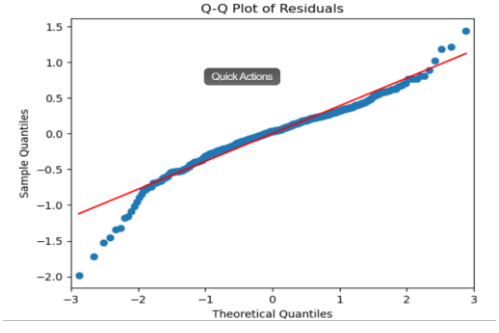
- Linear relationship between X and Y .

  From the pair plot ,it is clear that all the variables are in linear relationship with dependent variable ‘cnt’. Hence the assumption holds.

- Error terms are normally distributed (not X, Y)

  I ploted the error distribution graph and Q-Q plot for the residuals . below graphs clearly indicates that the distribution is normal and the mean error is 0. Hence the assumption holds.



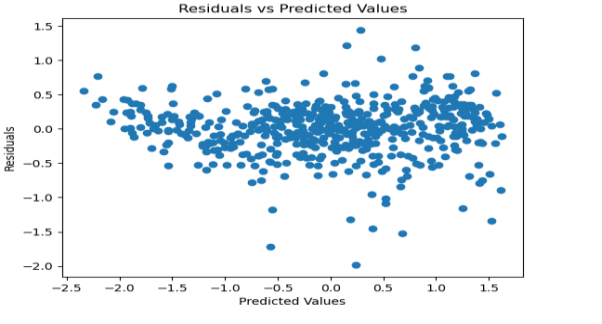


- Error terms are independent of each other .

  Durbin-Watson statistic result is 2.02 which indicates no auto correlation and the values are significantly different from each other

- Error terms have constant variance (homoscedasticity).

  Plotted Residuals vs predicted value scatter plot, I did not observe any patterns and hence the assumption holds.



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

- weathersit_LightSnowRain with co-efficient -1.13.
- yr variable with co-efficient - 1.04.
- atemp variable with co-efficient 0.40(positive)

General Subjective Questions.

Question-1 : Explain the linear regression algorithm in detail. ?

Linear regression is a supervised learning algorithm used for predicting a continuous target variable Y based on one or more independent variables (X OR X1,X2,X3….Xn).
Following are the assumptions to build linear regression model.

- Linear relationship between X and Y .
- Error terms are normally distributed (not X, Y) .
- Error terms are independent of each other .
- Error terms have constant variance (homoscedasticity).

Equation of linear regression

$$Y = \beta 0 + \beta 1 \cdot X1 + \beta 2 \cdot X2 + \ldots + \beta n \cdot Xn + \epsilon$$

- y: Dependent variable (target variable) we want to predict.
- X1,X2…Xn : Independent variables (features) that influence Y.
- β0: Intercept (constant term), where the regression line crosses the y-axis.
- β1,β2,…,βn: Coefficients (parameters) representing the relationship between each independent variable and dependent variable.
- ϵ Error term (residuals), representing the deviation of the actual values from the predicted values.

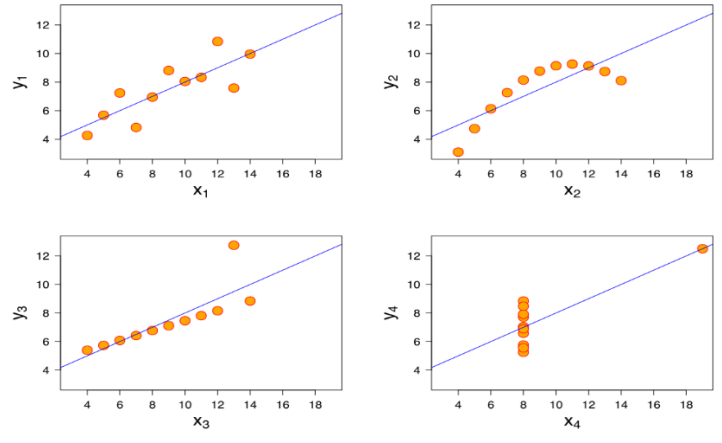
And The objective of linear regression is to find the coefficients of independent variables that minimize the sum of squared differences between the actual values and the predicted values.

Question 2 : Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is the evidence to prove the importance of graphical visualization of data instead of relying only on descriptive statistics.

It consists of 4 datasets which has similar statistical properties like mean, variance, correlation coefficients. But totally looks different when we visualize these four datasets(refer below fig) .



Question-3 : What is Pearson's R?

Answer:

- Pearson's R is the statistical method to identify the relationship between two variables.
- Pearson's R value ranges between -1 to +1.
- Formula r = Cov(X,Y) / SD(X).SD(Y)
- Interpretation of Pearson's R.

  a. if r > 0 then Y increases with the increase in X.

  b. If r<0 then Y decreases with the increase in X and vice versa.

  c. r=0 No correlation between X & Y.
- We can apply correlation only if X & Y are in linear relationship.

Question-4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

- What is scaling? .

  Scaling is a type of data preprocessing where the data is transformed into specific range .There are two types of scaling

  a. Normalized scaling.

  b. Standardized scaling.

- Why scaling ?.
    a. Many machine learning models perform better if all numeric variables are on a similar scale. especially in distance-based algorithms produces biased results if the variables scale different is huge.
- Difference between Normalized Scaling and Standardized Scaling.

Normalized Scaling	Standardized Scaling
Rescales to the range (0-1)	Rescales to the range (-3,+3) with mean 0 and S.D 1
It is useful ,if the data distribution is not normal	It is useful if the data distribution is normal
It is more sensitive to outliers compared to standardized scaling	It is less sensitive to outliers compared to normalized scaling
Formula , X= (X-mean(X))/S.D(X)	Formula, X= (X-min(X))/(Max(X)-Min(X)))

Question-5 : You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer:

VIF will indicate the multi-collinearity in the data. If it is infinite then it means that one of the independent variable is having perfert correlation with other variable/variables i.e .In such cases the variable can be perfectly predicted by other variables

- Why does it happen ?

  VIF = 1/(1-Ri^2) , where Ri^2 is the R^2 value from the regression of Xi against Xi-1
  In the perfect linear correlation, R2 will be 1 So
  VIF = 1/(1-1)
  VIF = 1/0 = Infinity

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

- What is Q-Q Plot ?

  It is also called as Quantile-Quantile plot, it is a method used to verify whether the dataset follows a expected distribution or not. It compares the quantile of the given data against the quantile of theoretical distribution.
- Use and importance of Q-Q Plot in linear regression.

  We assume that residuals of linear regression model is normally distributed with mean 0 and S.D =1 ,but we need statistical technique to prove the normal distribution. In such situations Q-Q Plot becomes handly, it compares the quantile of residual with the quantile of theoretical distributions and concludes whether assumption (“Error Term is normally distributed or not”) holds or not .