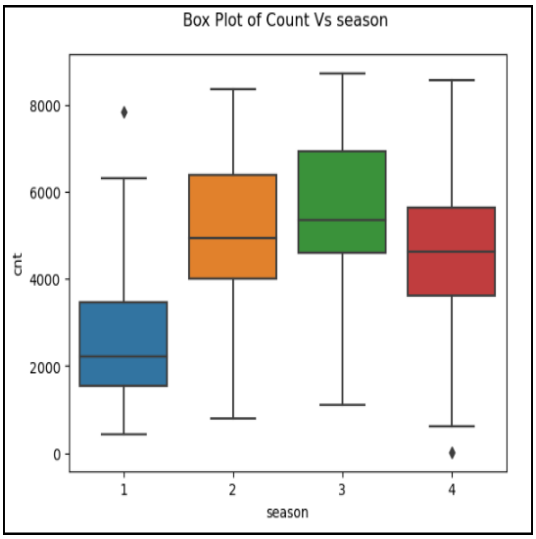
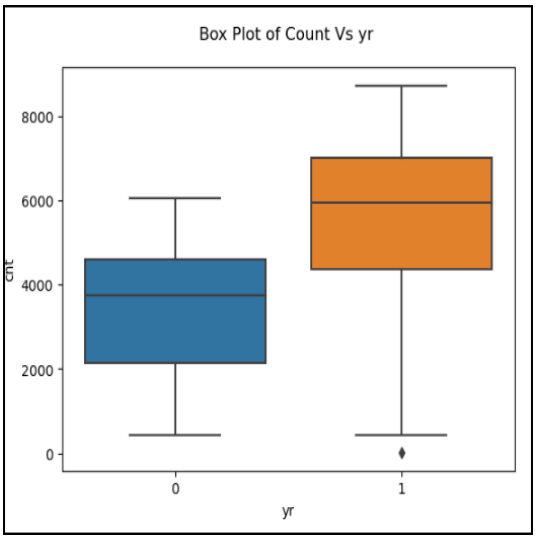
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

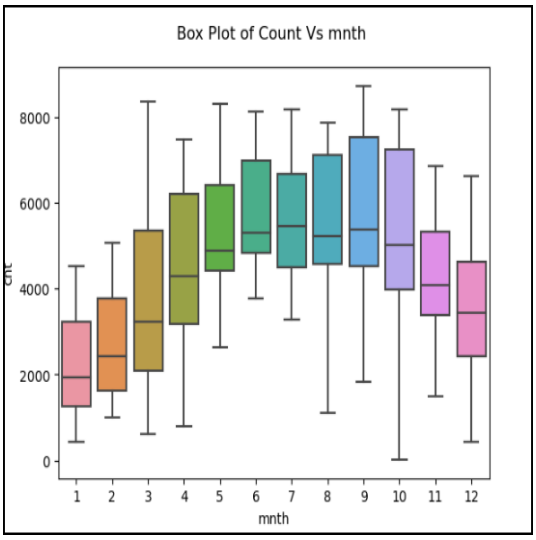
Ans:



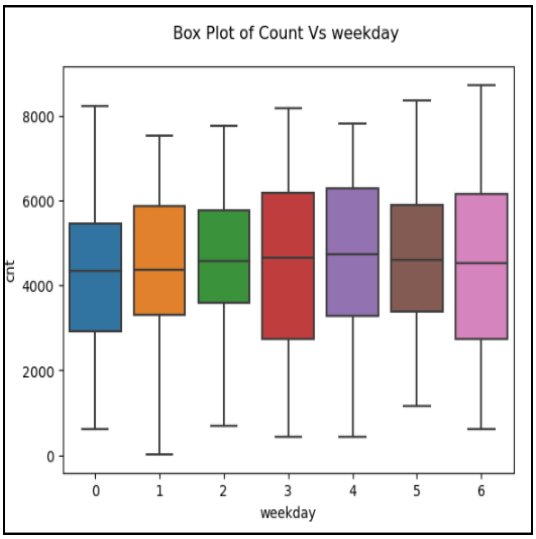
From the above graph, it can be observed that season is affecting the demand of the bikes. The demand is gradually increasing from season spring, summer, fall and it drops in winter.



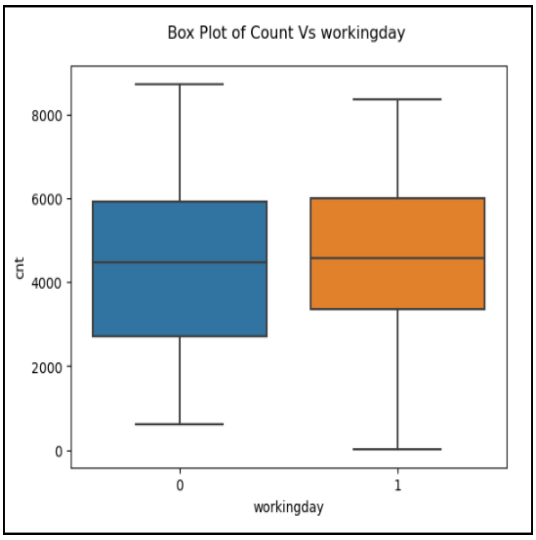
From the above graph it can be seen the demand of the bikes is increased in 2019.



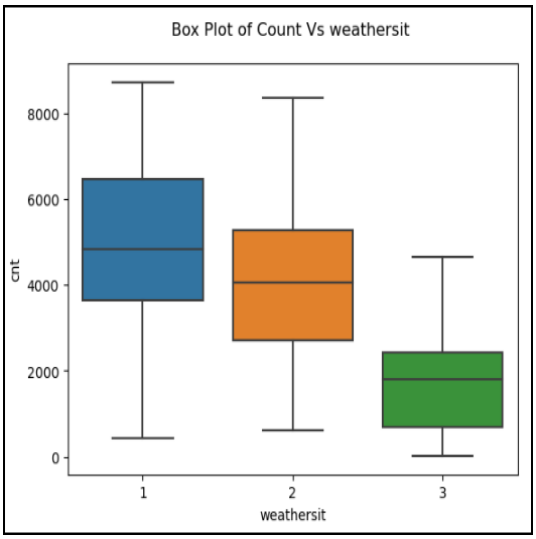
From above graph, it can be observed that the demand is varying with the months. In the months where winter starts the demand is less which is also true for season graph.



From above graph, it can be observed that there is not that much effect of weekdays on demands of bikes. Median of the demand is almost the same for all the weekdays. There is a variation in spread of the data.



From above graph, it can be observed that max demand in bike is more if there is a holiday, but overall there is not much effect on the spread of the bike demand on working day and non- working day. Median is almost same.



From above graph, it can be observed that the demand for the bike is decreasing with the following weather situation

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy

- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

Q2. Why is it important to use drop_first=True during dummy variable creation?

Without using 'drop_first=True' command it will create the columns for all the categories which is not necessary.

e.g. if three categorical variables are there and we create 3 dummy variables then the values will be

1st variable	2nd variable	3rd variable
1	0	0
0	1	0
0	0	1

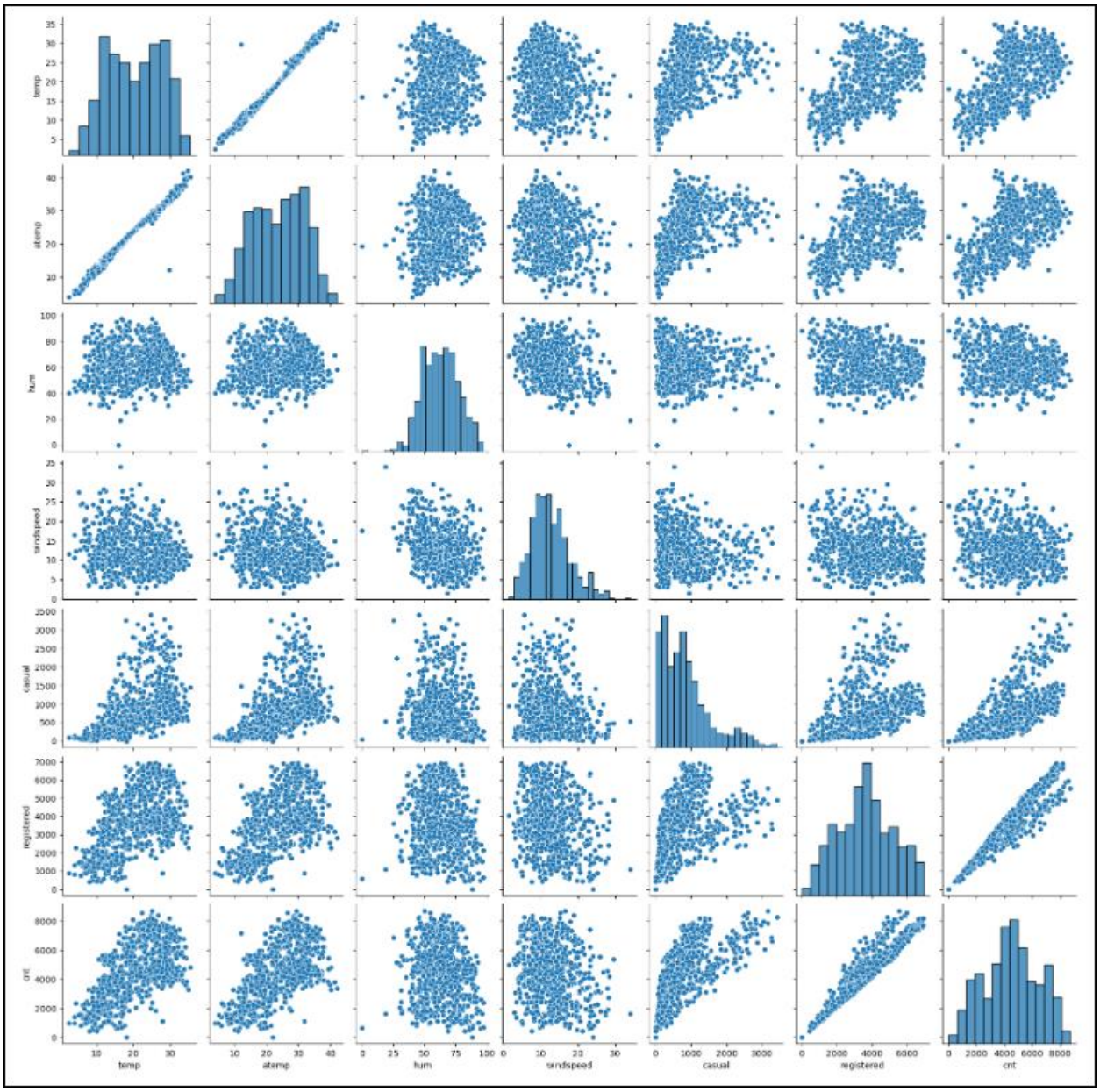
Now from the above table it can be seen that we can explain the first variable by 0 for 2nd and 0 for 3rd , so we can drop the first variable and values will be as below

2nd variable	3rd variable	
0	0	Represents first variable
1	0	Represents second variable
0	1	Represents third variable

By this we can save time and make the model less complex.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

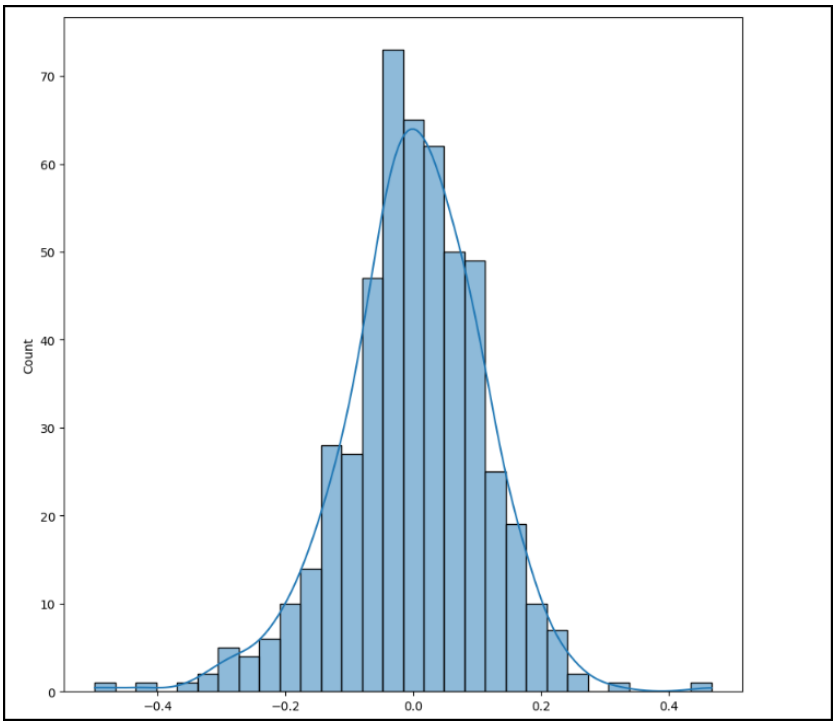
Ans:



It can be seen from the above pair plot, the highest correlation is between atemp and temp, the data is almost matching which is obvious as it is feeling temperature and real temperature. Feeling temperature is derived from the real temperature.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

By plotting the histogram of error terms= Y actual-Ypred and it can be seen from the graph that the error terms are normally distributed and mean of error terms is zero. Below is the plot of the same.



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The equation for the Multilinear regression model is as below,

Y_pred=0.2306+0.2476*Year+0.0622*weekday-0.1366*windspeed-0.0835WS2-0.3065*WS3+0.2664*summer+0.3369*fall+0.2604*winter

looking at the equation above top three features are as below

1. Fall season (+ve impact on demand)
2. Weathersituation3 i.e. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-ve impact on demand)
3. Summer season (+ve impact on demand)

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear regression is a technique to predict the value of dependent variable with respect to one or more independent variable based on available data set.

Basically, there are two types of linear regression.

1. Simple linear regression: in this y values are predicted only on one independent variable.
2. Multiple linear Regression: In this y values are predicted on more than one independent variable.

The main aim of the linear regression is to best fit the line within the given spread of dependent and independent variable, so that the square of error terms = y prediction- y actual is minimum.

The equation of the linear regression is

$Y=\beta_0+\beta_1*X_1+ \beta_2*X_2+……..+ \beta_n*X_n$

where,

β_0 = intercept

$\beta_1, \beta_2…. B_n$ =Slope

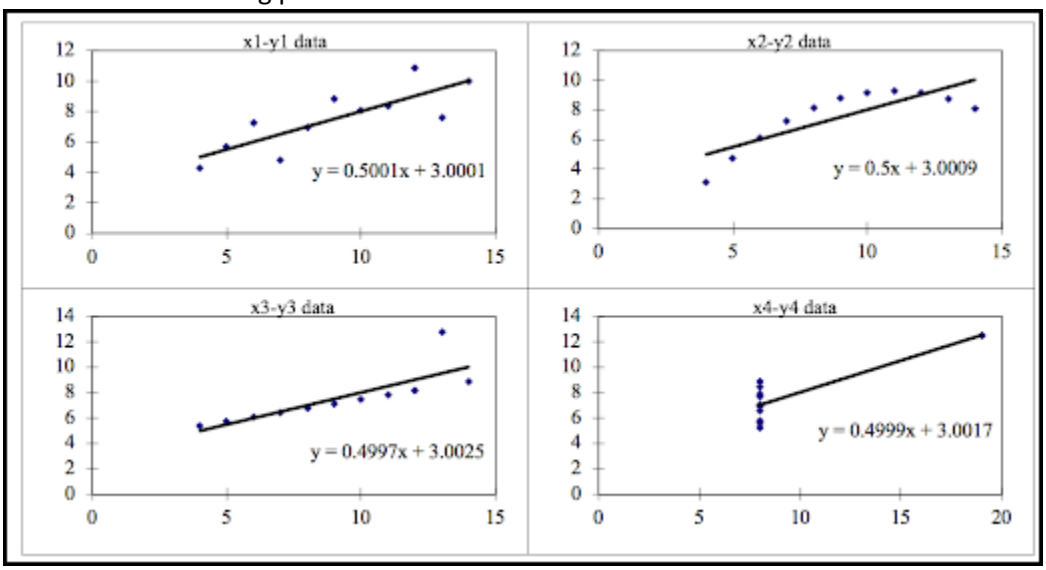
Below are the steps which are to be followed to build a linear regression model

1. Reading and understanding the data
    - Reading the data
    - Cleaning the data
    - Understanding the data by using various commands such as describe info etc.
    - Visualizing the data by univariate and bivariate analysis and get insights into the data.
2. Data preparation
    - Dividing the data into train and test data set
    - Creating dummy variables for categorical data
    - Scaling the data
3. Building a linear model
    - For Simple linear model fitting the line between dependent and independent variable
    - For multiple linear model we can fit the line by finding which variables are significant for the model by observing P values, also dropping the variables which show the multicollinearity among themselves i.e., which are having VIF value>5. This can be done manual, automated or by combination of both methods.
4. Residual analysis
    - In this we verify the underlying assumptions in linear regression are true so that we can verify our model can be explained by linear regression
5. Model Evaluation
    - Model can be evaluated by calculating R-squared value
    - Testing the model on test data set checking, the R-squared value

- If there is not significant difference in both R square values, then we can say that our model working well

Q2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet is four datasets which are identical in statistical properties but visually are different. This tells us the importance of visualizing the data. In Anscombe's quartet, for four different data sets the mean, variance and correlation coefficient is exactly same for all the models. But after plotting the variables the following patterns has been observed



- In the first graph, it can be observed that the data is scattered, and which is explained greatly by Linear model.
- In the second graph, it can be observed that the data is nonlinear and linear regression line is not a good choice here.
- In third graph there is outlier which can not be explained by linear model
- In the fourth graph x value is same and y values are different and one outlier is there, which can not be explained by linear model.

By above observation it can be concluded that data visualization is very important before building the model.

Q3. What is Pearson's R?

Pearson's R is nothing but the correlation coefficient. It explains the linear correlation between two variables. The value of Pearson's R lies between -1 to +1. -ve Value indicated negative correlation between variables and +ve value indicated positive correlation between the variables. Lower the absolute value of R, lower is the linear relation between the variables.

Pearson's r is calculated by dividing the covariance of the two variables by the product of their standard deviations. The formula for Pearson's r is:

$r = (\Sigma (xi - x) * (yi - y)) / ((n-1) * sx * sy)$
where xi and yi are the values of the two variables, x and y are their means, sx and sy are their standard deviations, and n is the sample size.

Pearson's R only measures the linearity between the data, and it does not explain non linearity. It is also sensitive to outliers and can be affected by distribution of the data.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a term in which all the variables are brought to one scale so that they can be easily understood and explained.

Consider an example in which two variables have different units e.g., Area in feet and area in meter, now fitting the linear model with these two variables will have different values of slope coefficient as the unit is different. Now consider this for more than two variables which are very different from each other. So, to explain and understand the model correctly it is wise to do the scaling and bring it to one scale.

Below are the scaling techniques which are normally used

1. Normalized scaling
2. Standardized scaling

Difference in them is as follows,

Normalized Scaling	Standardized Scaling
Normalization or Min-Max Scaling is used to transform features to be on a similar scale	Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation.
X_new = (X - X_min)/(X_max - X_min)	X_new = (X - mean)/Std

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF is a measure of multicollinearity between the variables.

The formula for calculating the VIF is,

$VIF=1/(1-Ri^2)$.

By looking at the formula we can say that when denominator will be zero the value will be infinite which is possible when Ri^2 term is 1 or very near to 1 which is possible in following conditions

1. When the variables are very highly corelated
2. When there is error in the data e.g., duplicated values are there in variables.

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

Q-Q plot is a graph used to compare two probability distributions by plotting their quantiles against each other. It is used to assess the normality of a data distribution.

In linear regression, a Q-Q plot is used to check the normality of the residuals. Residuals is the difference between observed values and predicted values. If the residuals are normally distributed, then the points in Q-Q plot is near to straight line. Q-Q plot can also be used to check heteroscedasticity.