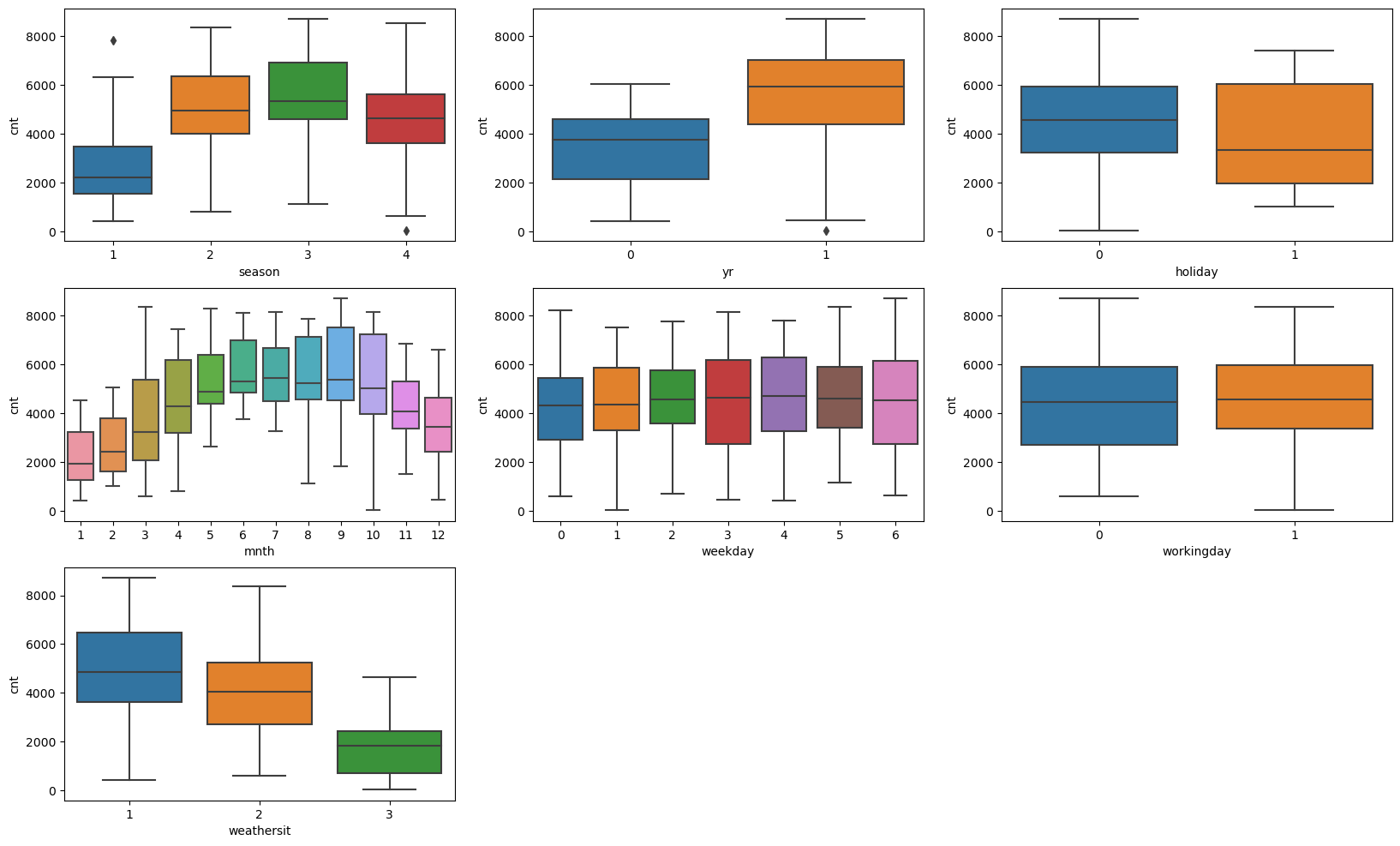
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Categorical variables present in the dataset are as follow: -

1. Season (season)
2. Year (yr)
3. Holiday (holiday)
4. Month (mnth)
5. Weekdays (weekday)
6. Workingday (workingday)
7. Weather Situation (weathersit)



From the initial analysis (EDA-Box plot following inferences could be observed)

-The average shared bike demand was more in Summer and Rainfall

-The average shared bike demand was more in the year 2019 compared to 2018

- The average shared bike demand was more in the holidays and Weekends compared to Week days

-The average shared bike demand was max in the month of June

-The average shared bike demand was max on Thursdays

-The average shared bike demand was more on clear climates and least during precipitation days

At the end of feature selection Wednesday, March, April, Jan and Dec turns out to be highly significant with less p value and VIF in the considerable range.

Month and Weekdays are the categorical variables which could significantly tell about the shared bike demand

1. Why is it important to use drop\_first=True during dummy variable creation?

It is necessary to use drop function while using dummy variable creation because to avoid the extra columns and correlation and hence the multicollinearity

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

If the relation ship between the numerical variable and the target variable has a linear relationship and can be explained by a straight line has the highest correlation. It can be either negative or positive based on the direction of straight line that could be fitted in the data points

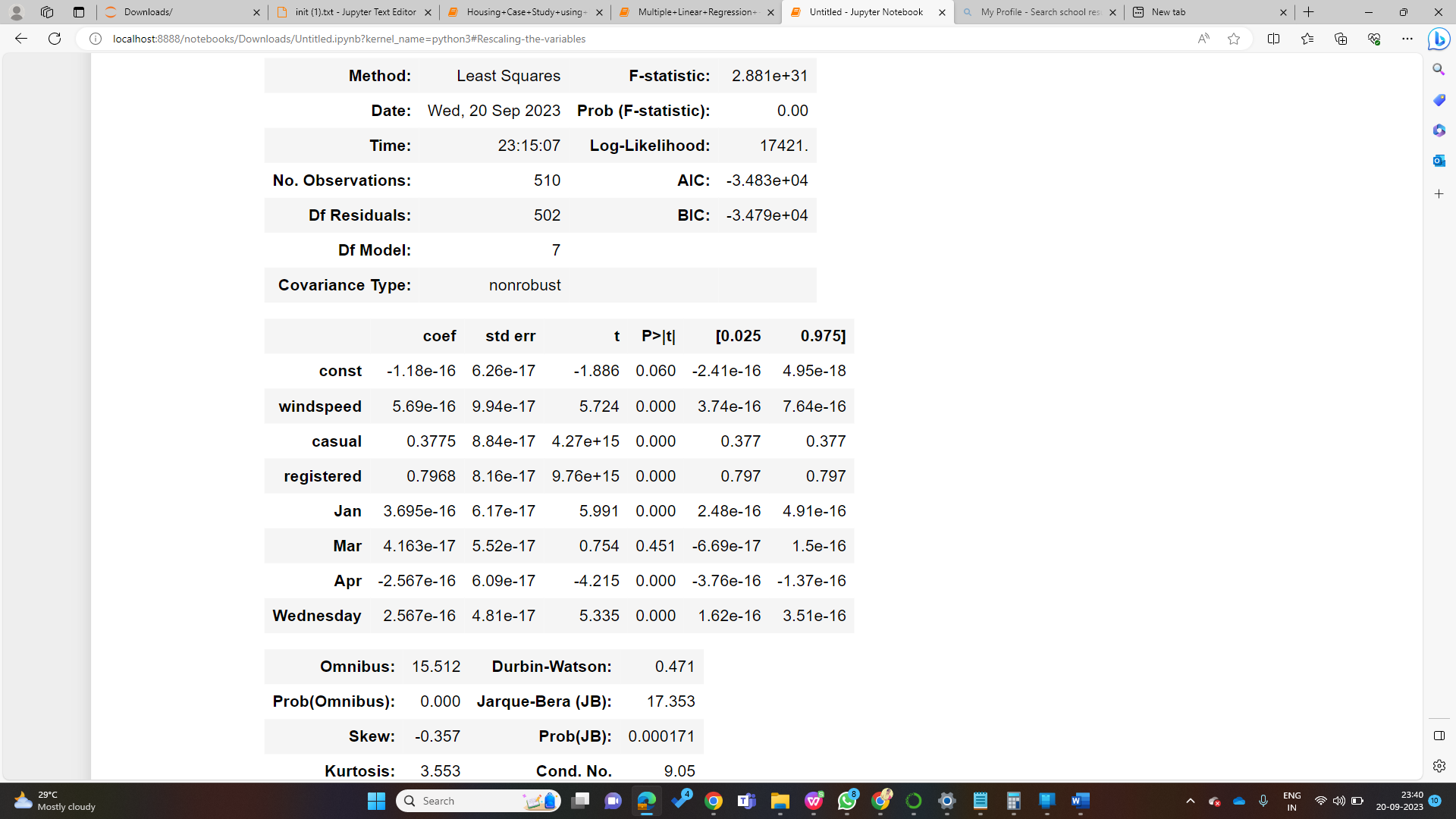
1. How did you validate the assumptions of Linear Regression after building the model on the training set?

Assumptions of Linear regression model are as below: -

* The error terms are independent -by plotting (res Vs Y)
* The error terms have constant variance-using scatter plot and check whether divergence or convergence is there
* The error terms are normally distributed – using Distplot
* The X and Y must have a linear relationship-using scatter plot

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

Top Features are as follow:-



1. Windspeed, Casual, Registered, Jan, Apr, Wednesday
2. Mar
3. Explain the linear regression algorithm in detail.

Linear Regression Algorithm is a type of Supervised learning wherein the linear relationship between dependent and independent variable is used to build a model and predict the data.

SLR- (Simple Linear regression) -(y=Bo+B1\*x)

When there is one independent variable it is called SLR

MLR- (Multiple Linear regression) -(y=Bo+B1\*x1…...+Bi\*xi)

When there is more than one independent variable

Linear Regression model fits the straight line using Least Mean Square Method

Minimising the RSE Value

1. Differentiation
2. Gradient Descent method

The model is checked using R2 Parameter

Assumption for Linear Regression Model

Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.

Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.

Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.

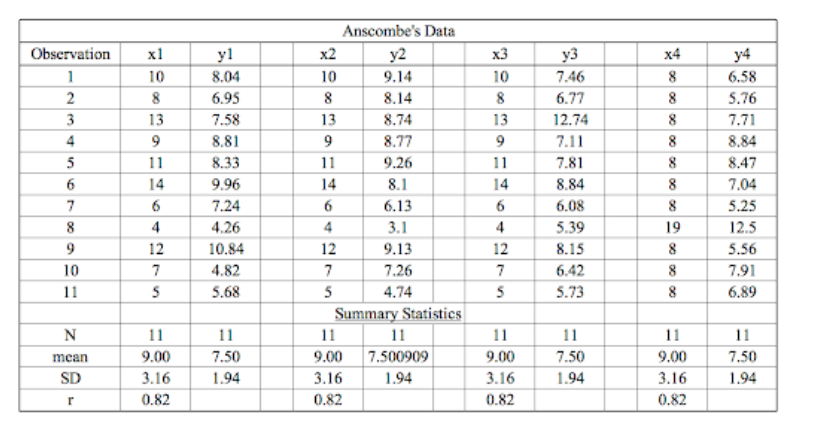
Normality: The errors in the model are normally distributed.

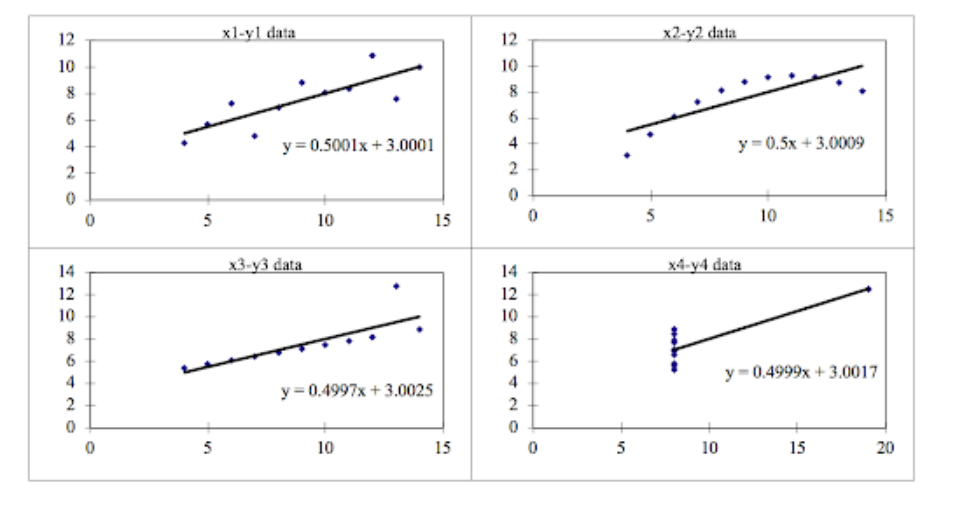
No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics but are entirely different when plotted in scatter plot.

For e.g. There are 4 Datasets with same statistical values but when plotted the it generates a different kind of plot and cannot be represented by regression model





3. What is Pearson’s R?

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. It is used in Bivariate Analysis.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude other wise it will affect the coefficients

Normalised Scaling -Scaling

It brings all of the data in the range of 0 and 1.

Standardised Scaling

It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

VIF =1/(1-R2)

When R2 is equal to one VIF=infinity which means perfect correlation between variables

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

The purpose of the quantile-quantile (QQ) plot is to show if two data sets come from the same distribution. Plotting the first data set’s quantiles along the x-axis and plotting the second data set’s quantiles along the y-axis is how the plot is constructed.

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