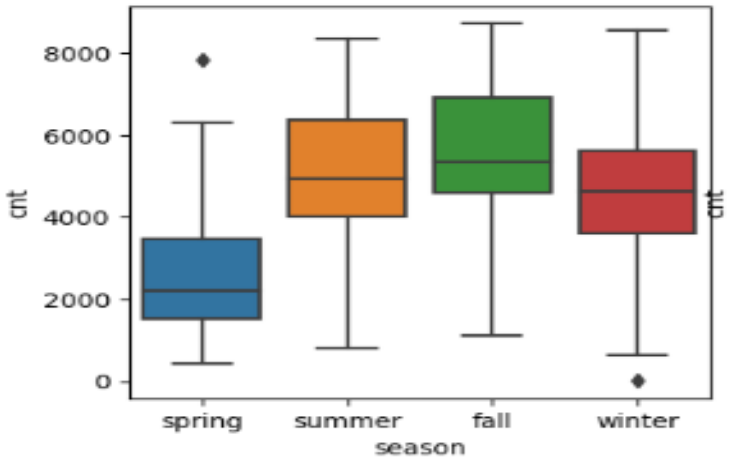
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

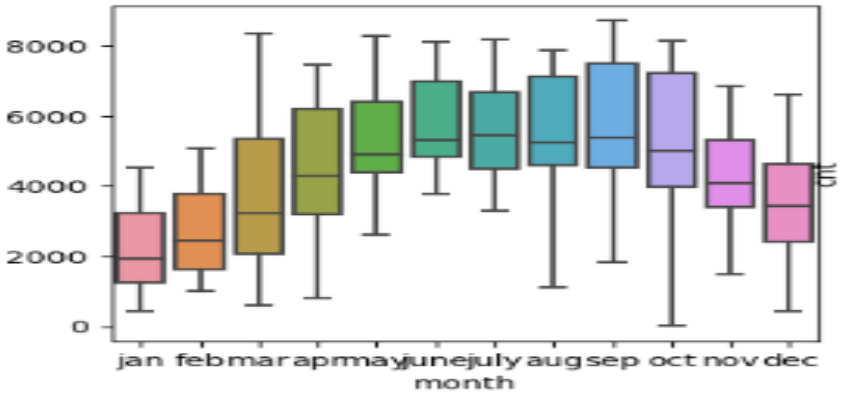
Answer:

Season:



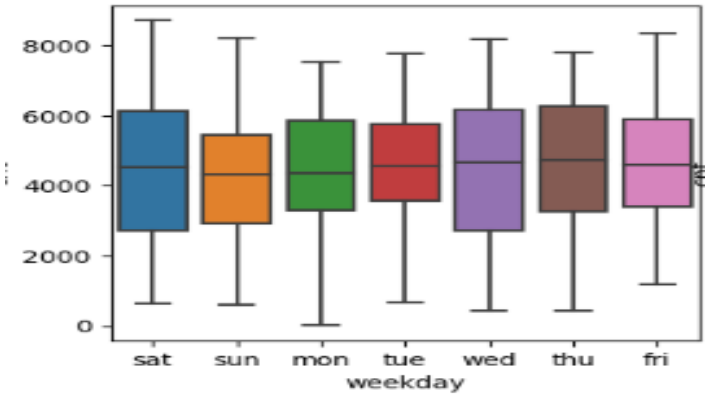
- The Bikes count changes according seasons and Spring season has the least business and fall has the highest business.
- According to the data we have we can Increase our business in Summer and Fall we can have good online marketing and we can increase our stations according to the season and rent some places with extreme weathers as we can Lease it out when it is required and leave it if not required.

Month:



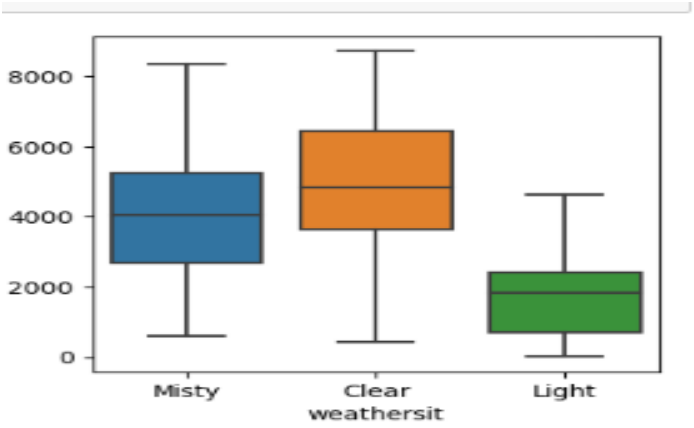
- As we can see the count of bikes is fluctuating according to the weather conditions. We cannot ride a bike in extreme weathers and according to the data we have we can conclude that.

Weekday:



- Saturday and Wednesday are busiest days for the business.
- So what we can do is if bike bike requires servicing and according to data we have the bikes can be delivered to servicing and minor repairs on Tuesday so that we don't loose any business.

Weather Sit:



- The Visualization of data tells us that if heavy snow is there is no one that rents a bike and according to the weather the renting of bikes change.

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

Answer :

- **drop_first = True** This syntax is used to help us not create any extra columns during the dummy variable creation.
- For the data if the command is not given it is creating 34 features and when given only 30 features removing the dummy columns.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable? (1 mark)

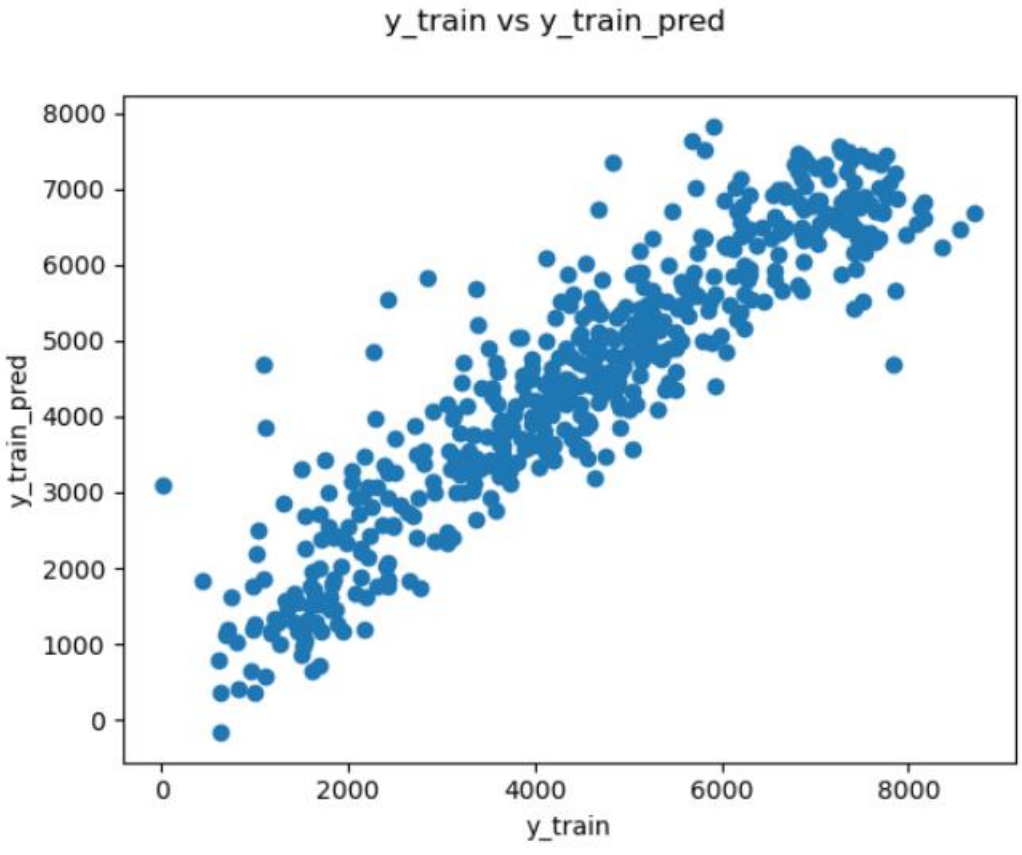
Answer :

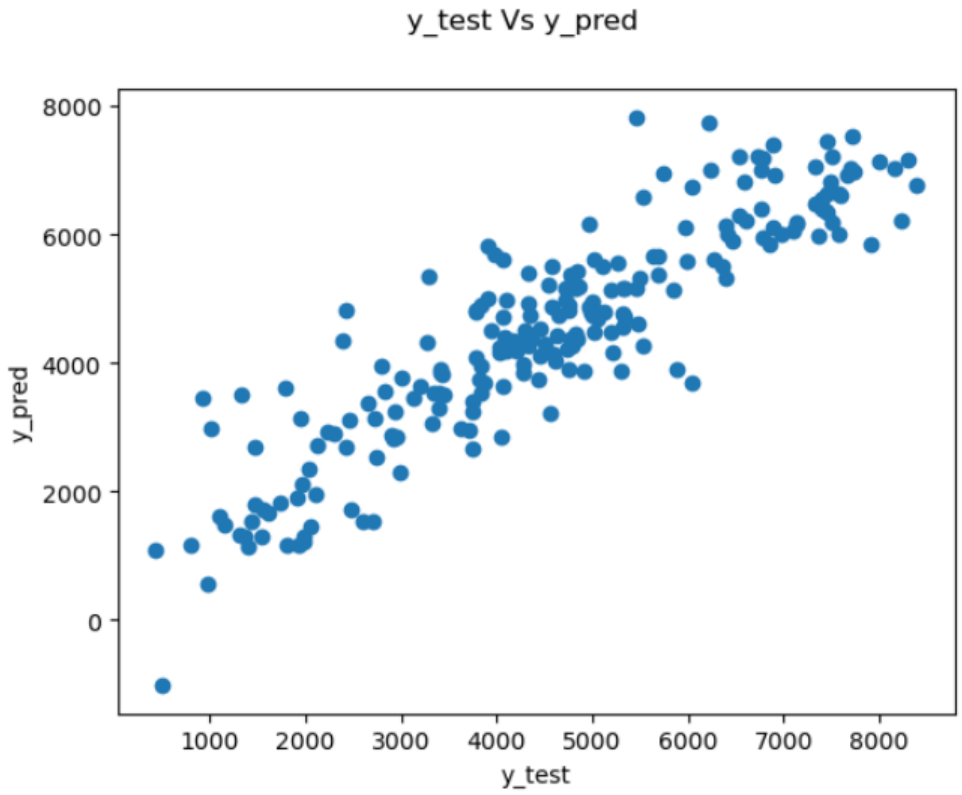
- The Variable is **'temp'** with 0.63 correlation.

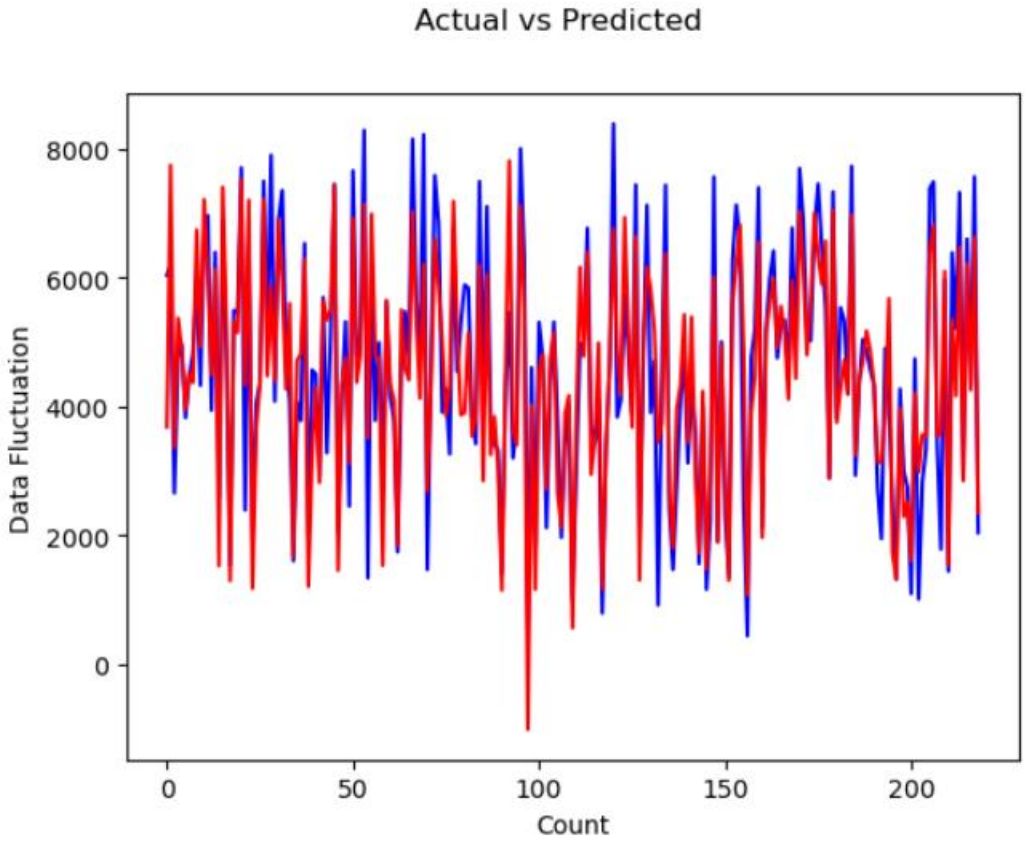
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

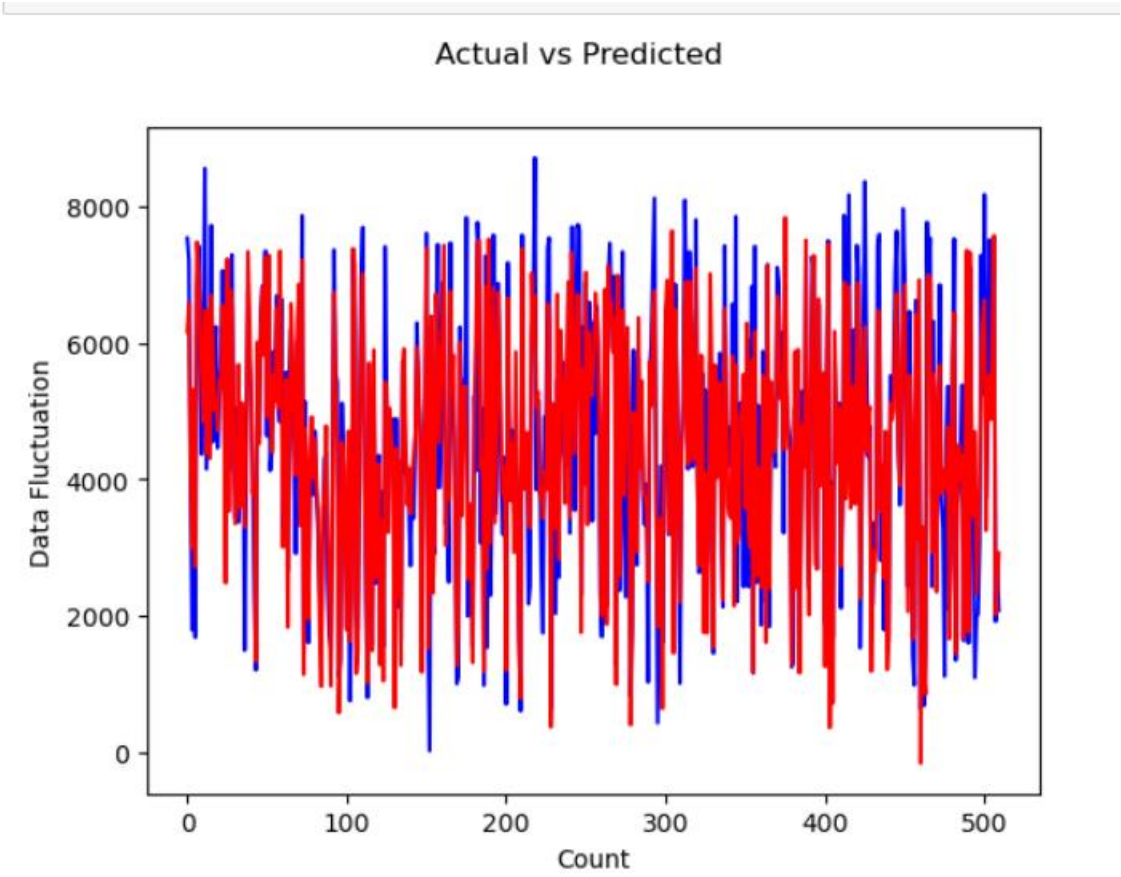
Answer :

- Error terms
- Multi collinearity
- Linear relationship between the features
- Residual Analysis
- Homoscedasticity




y_test Vs y_pred
8000
6000
4000
2000
0
y_pred
1000
2000
3000
4000
5000
6000
7000
8000
y_test


Actual vs Predicted
8000
6000
4000
2000
0
Data Fluctuation
0
50
100
150
200
Count



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer :

The top three features are:

- Windspeed (windspeed)
- Working day (workingday)
- Temperature (temp)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer :

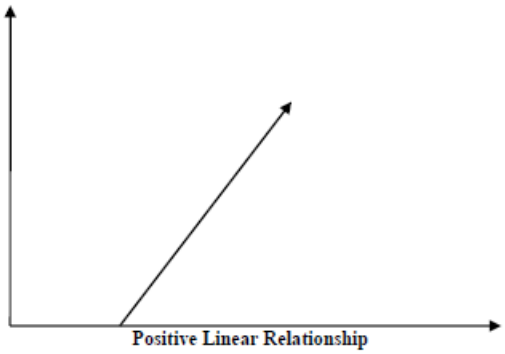
Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more

independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
Mathematically the relationship can be represented with the help of following equation −

$Y = mX + c$

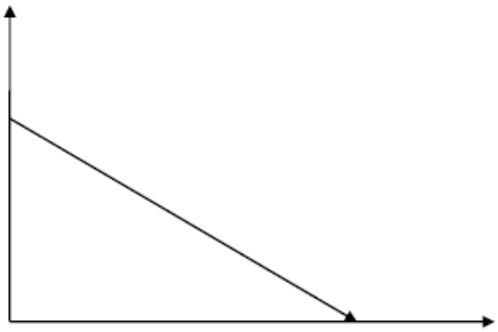
Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.
Furthermore, the linear relationship can be positive or negative in nature as explained below−

1. Positive Linear Relationship : When the X-axis increases Y-axis also increases.



Positive Linear Relationship

2. Negative Linear Relationship: When X-axis increases then Y-axis decreases i.e X and Y axis are inversely proportional.



Negative Linear Relationship

Linear regression is of the following two types −

- ✓ Simple Linear Regression
- ✓ Multiple Linear Regression

Assumptions :

- Multi-collinearity
- Auto-correlation

- Relation between the features
- Error terms
- Homoscedasticity

2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer :

Anscombe's quartet is a set of four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

Dataset 1:

```
x, y
10, 8.04
8, 6.95
13, 7.58
9, 8.81
11, 8.33
14, 9.96
6, 7.24
4, 4.26
12, 10.83
7, 4.82
5, 5.48
```

Dataset 2:

```
x, y
10, 9.14
8, 8.14
13, 8.74
9, 8.77
11, 9.26
14, 8.1
6, 6.13
4, 3.1
12, 9.13
7, 7.26
5, 4.74
```

Dataset 3:

```
x, y
10, 7.46
8, 6.77
13, 7.2
9, 6.11
11, 7.59
14, 8.86
6, 4.26
4, 3.75
12, 7.31
7, 5.73
5, 5.32
```

Dataset 4:

```
x, y
10, 9.14
8, 8.14
13, 8.74
9, 8.77
11, 9.26
14, 8.1
6, 6.13
4, 3.1
12, 9.13
7, 7.26
5, 4.74
```

If we look at the summary statistics for each dataset, we see that they are all very similar. The mean and variance of x and y are the same for all four datasets, and the correlation coefficient between x and y is also the same. However, when we plot the datasets, we see that they look very different from each other.

The first dataset shows a clear linear relationship between x and y. The second dataset also shows a linear relationship, but it is not as strong as the first dataset. The third dataset shows a linear relationship, but there is also an outlier that is far away from the other points. The fourth dataset does not show any clear relationship between x and y.

Anscombe's quartet is a reminder that we should always visualize our data before we start to analyze it. Summary statistics can be misleading, and they can hide important information about the distribution of the data. By visualizing our data, we can identify outliers, patterns, and other important features that we might not have noticed otherwise.

Anscombe's quartet is also a reminder that we should be careful about making assumptions about our data. Just because two datasets have the same summary statistics does not mean that they are the same. We should always look at the data itself to see how it is distributed and what kind of relationship there is between the variables.

Anscombe's quartet is a valuable tool for teaching and learning about data visualization and statistical analysis. It is a reminder that we should always be critical of the data we use and the methods we use to analyze it.

3. What is Pearson's R? (3 marks)

Answer :

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.

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

4. What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling? (3 marks)

Answer :

Scaling is a data preprocessing technique that involves transforming the values of features or variables in a dataset to a similar scale. This can be done for a number of reasons, including:

- To make features more comparable. For example, if you are training a machine learning model to predict house prices, you might want to scale the features "square footage" and "number of bedrooms" to a similar scale so that the model does not give undue weight to either feature.
- To improve model performance. Some machine learning algorithms, such as gradient descent, are sensitive to the scale of the features. Scaling the features can help these algorithms to converge more quickly and to find better solutions.
- To prevent overfitting. Overfitting occurs when a machine learning model learns the training data too well and is unable to generalize to new data. Scaling the features can help to reduce overfitting by making the model less sensitive to the specific values of the features in the training data.

Normalized scaling and **Standardized scaling** are two common techniques for scaling features.

Normalized scaling transforms the values of a feature to a range between 0 and 1.
Standardized scaling transforms the values of a feature to have a mean of 0 and a standard deviation of 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

VIF can be infinite if there is perfect multicollinearity between two or more independent variables. This means that one independent variable can be perfectly predicted from the other variables in the model. In this case, the regression coefficients for the collinear variables are indeterminate, and the VIF is infinite.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer :

A Q-Q plot, or quantile-quantile plot, is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Q-Q plot can also be used to test distribution amongst 2 different datasets. For example, if dataset 1, the age variable has 200 records and dataset 2, the age variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same.