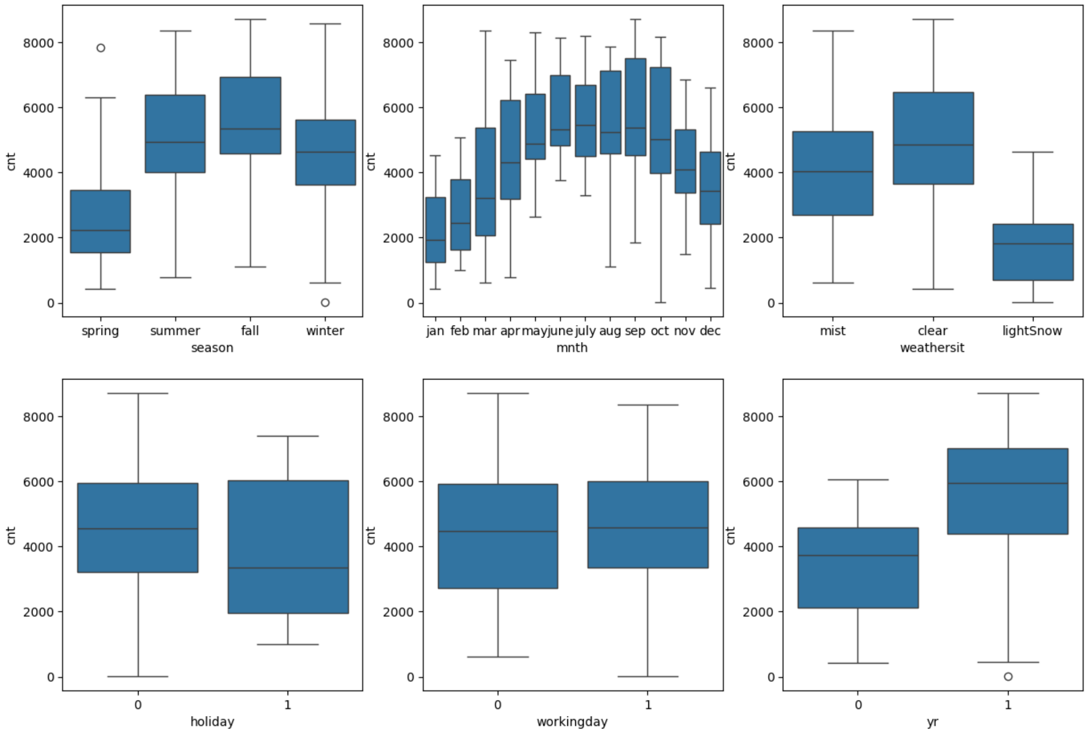
Assignment-based Subjective Questions

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

   Answer:
   I analysed the categorical variables using boxplots, and the following insights were observed:

   - Bike bookings increase during summer and winter seasons.
   - From May to October i.e Q2 and Q3, there is a noticeable rise in bookings, while the first and fourth quarters of the year show a drop.
   - During clear weather, the number of rentals is higher compared to other weather conditions.
   - Holidays show higher bookings than non-holiday days.
   - The year 2019 recorded more bookings compared to the previous year, 2018.
   - We can also observe two outliers in the spring and winter seasons.



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

Answer:
drop_first=True removes one of the dummy columns when creating dummy variables which helps avoid creating extra columns that can cause high correlation among the dummy variables.

drop_first=True keeps only k–1 dummy columns instead of k.

Code

```
season_dummies = pd.get_dummies(dfDay['season'], drop_first=True, dtype=int)
```

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

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

   Answer:
   The dataset has been split into two datasets one is training and other is validation/testing.
   The training and test data split was done 70:30 ratio.
   The sklearn train_test_split function is used to split the data.

   Code: train_test_split(dfDay, train_size = 0.7, random_state = 100)

   Linear Regression model is then trained with training dataset with multiple iterations then performing VIF checks to ensure no multicollarnarity.

   Code: [variance_inflation_factor(df.values, i) for i in range(df.shape[1])]

   Later is train dataset is validated with test dataset finally R2 Score is used for evaluation.
   r2_test = r2_score(y_test, y_test_pred)
   r2_train = r2_score(y_train, y_train_pred)

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

   Answer:
   - temp,
   - year,
   - winter season

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Answer:
   Linear regression is a method used to understand the relationship between input variables (X) and an output or target variable (Y). We usually visualize this relationship by drawing a straight line through the data points on a graph. This line shows the trend in the data and is called the best-fit line**.**

   The equation of this line is: Y = m*X* + *b*

   Y → the value we want to predict (dependent variable)

   X → the input used for prediction (independent variable)

   m → the slope, which tells us how much Y changes when X changes b → the intercept; it tells us the value of Y when X = 0

If we use more than one input variable, the method is called multiple linear regression. In that case, the equation becomes:

Equation: y(x) = w0 + w1x1 + w2x2 + … + w(n)x(n),

Here, each w represents a weight or coefficient that shows how strongly each variable affects the output.

The goal of the linear regression algorithm is to get the best values for w0 and w1 to find the best fit line and the best fit line should have the least error.

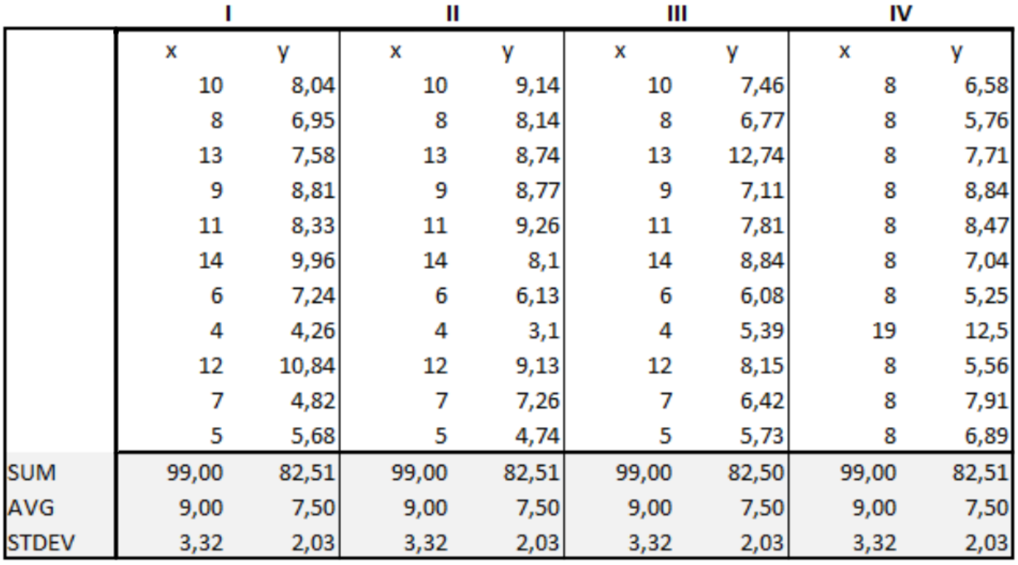
In Linear Regression, RFE or Mean Squared Error or cost function is used, which helps to figure out the best possible values for w0 and w1, which provides the best fit line for the data points.

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

Answer:
Anscombe's quartet was created by the statistician Francis Anscombe. It consists of four datasets, each containing eleven (x, y) pairs. They share identical descriptive statistics same averages, same variances, and even the same correlation.

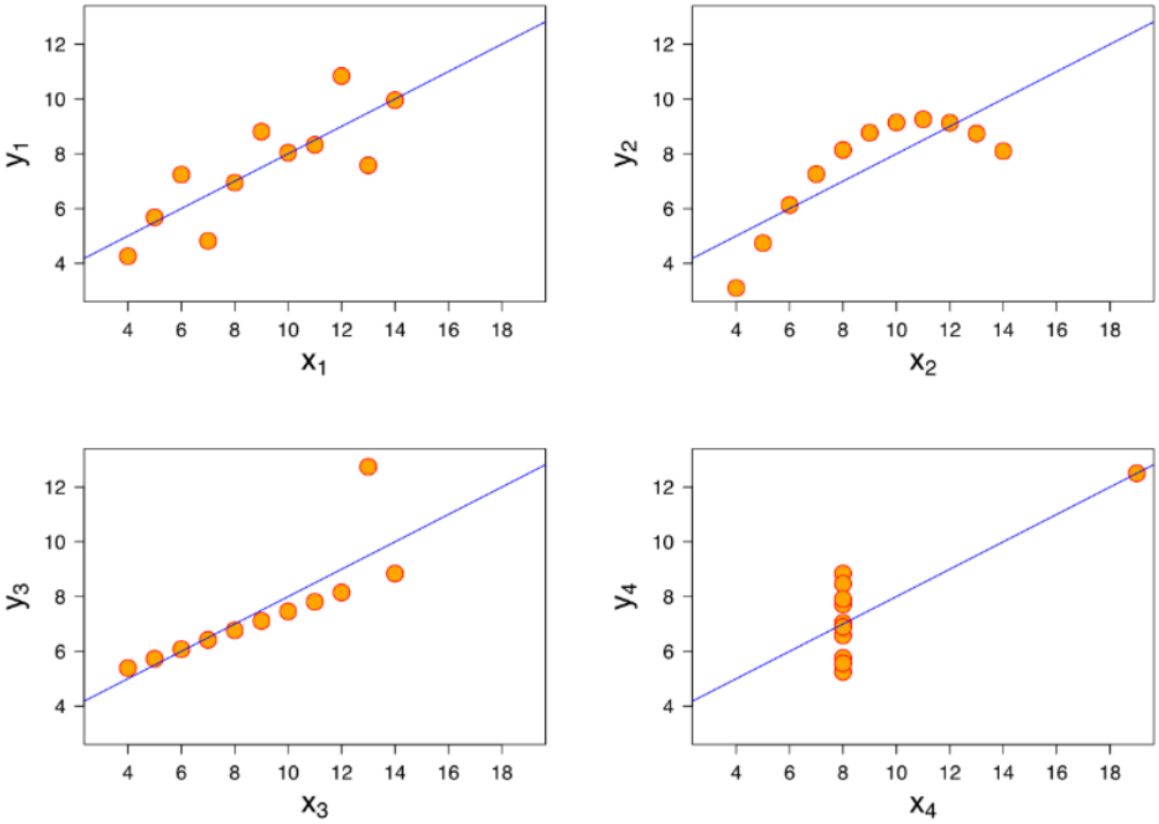
But when you graph them each dataset tells a very different story despite having the same summary statistics.

Example: If we look at the table and the SUM,AVG and STDEV looks similar



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

However, once plotted, the four graphs look totally different:



- Data I fits a clean, straight linear model
- Data II forms a curve and is clearly not linear
- Data III looks linear, but one outlier distorts the regression
- Data IV shows how a single extreme value can create a high correlation even when most points don't follow any pattern

  Summary statistics itself can hide important patterns or problems. When we look at the actual graphs, we get a true picture of the structure and meaning behind the data. So Anscombe's Quartet tells why data visualization is essential in data analysis.

3. What is Pearson's R?

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

Formula:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \ \sum(y-\bar{y})^2}}$$

Types of Correlation:

- Positive Correlation (r > 0): As one variable increases, the other also increases, r is between 0 and +1
- Negative Correlation (r < 0): As one variable increases, the other decreases, r is between 0 and –1
- Zero Correlation (r = 0): No linear relationship between the variables. r is 0 or close to zero

Where the Pearson correlation coefficient, r, can take a range of values from +1 to -1.

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

Answer:

Scaling is the process of transforming data so that its values fit within a specific range or distribution. This helps ensure that different features or variables can be compared fairly and do not dominate a model simply because of their larger values.

Why its performed:

- Scaling helps different features have similar impact, instead of just the largest numbers taking over.
- It helps in faster learning and better accuracy for algorithms in machine learning and statistics.

Difference:

S.No	Normalized Scaling	Standardized Scaling
1	Uses minimum and maximum values of the feature for scaling.	Uses mean and standard deviation of the feature for scaling.
2	Scales values to a fixed range, usually (0,1) or (-1,1).	Values are not bounded to any specific range.
3	Highly affected by outliers (because min/max shift).	Less affected by outliers.
4	Also called Min–Max Scaling or Normalization.	Also called Z-score Normalization or Standardization.

Code:

```
scaler = MinMaxScaler() or StandardScaler()

numeric_vars = ['temp','hum','windspeed','count']

scaler.fit(df_train[numeric_vars])
```

```
df_train[numeric_vars] = scaler.transform(df_train[numeric_vars])
```

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

   Answer:

   Variance Inflation Factor (VIF) in order to determine if we have a multicollinearity problem. VIF becomes infinite when two features are perfectly correlated with each other. This means one column can be predicted exactly from another column in the dataset. When this happens, the regression matrix cannot be inverted, and in the VIF formula:

   $$VIF = \frac{1}{1 - R^2}$$

   if the correlation is perfect, then become infinite as R2 is equal to 1

   $$R^2 = 1$$

   $$VIF = \frac{1}{1-1} = \frac{1}{0} = Infinite$$

   This makes the denominator zero, causing the VIF value to shoot to infinity**.**

   Therefore the range of VIF is between 1 and infinity.

   An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model.

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

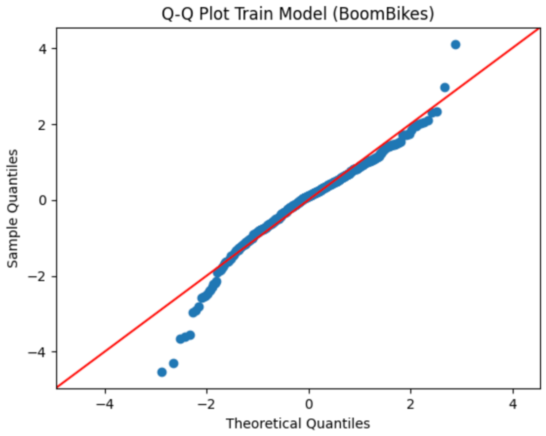
Answer:

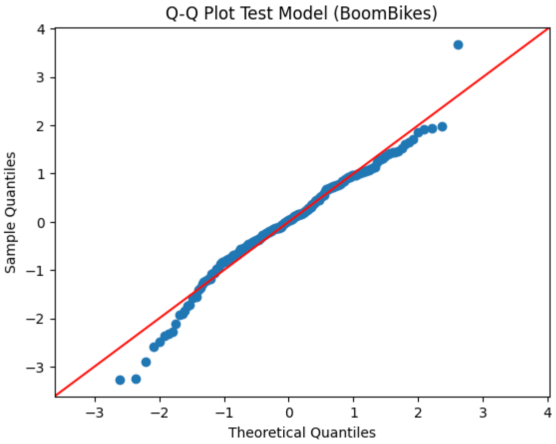
Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Q-Q plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the Q-Q plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression :
In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Q-Q plot based on Boom Bikes test and train dataset gives this distribution this tells the trained and tested models are more linear.


Q-Q Plot Train Model (BoomBikes)
Sample Quantiles
Theoretical Quantiles


Q-Q Plot Test Model (BoomBikes)
Sample Quantiles
Theoretical Quantiles