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:

The Observations based on the analysis of categorical variables by plotting a bar plot are as below

- Highest Booking was in the fall season
- Booking are found to increase till June and is almost constant till September and then

  Drops till December.
- Booking increase all the way from Sunday till Saturday with lowest on Sunday
- Higher Bookings on clear weather
- lower bookings are seen on holidays and doesn't seem to be affected by working days or not
- Higher bookings were seen in 2019

---

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

Answer:

If drop_first = True is not used , then n dummy variables will be created which means a dummy variable is created for every variable and will be corelated to each other causing multicollinearity. Setting drop_first = True drops the first category and ensures inly n-1 dummy variables are created

---

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

Answer:

Temp and atemp have the highest correlation

---

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

Answer:

Assumptions:

- The error terms are normally distributed
- No Auto Correlation
- Multicollinearity

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

6. Atemp
7. Windspeed
8. Yr

General Subjective Questions

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

Answer:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Mathematically, linear regression equation as:

$y = a + bx$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

---

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

Answer:

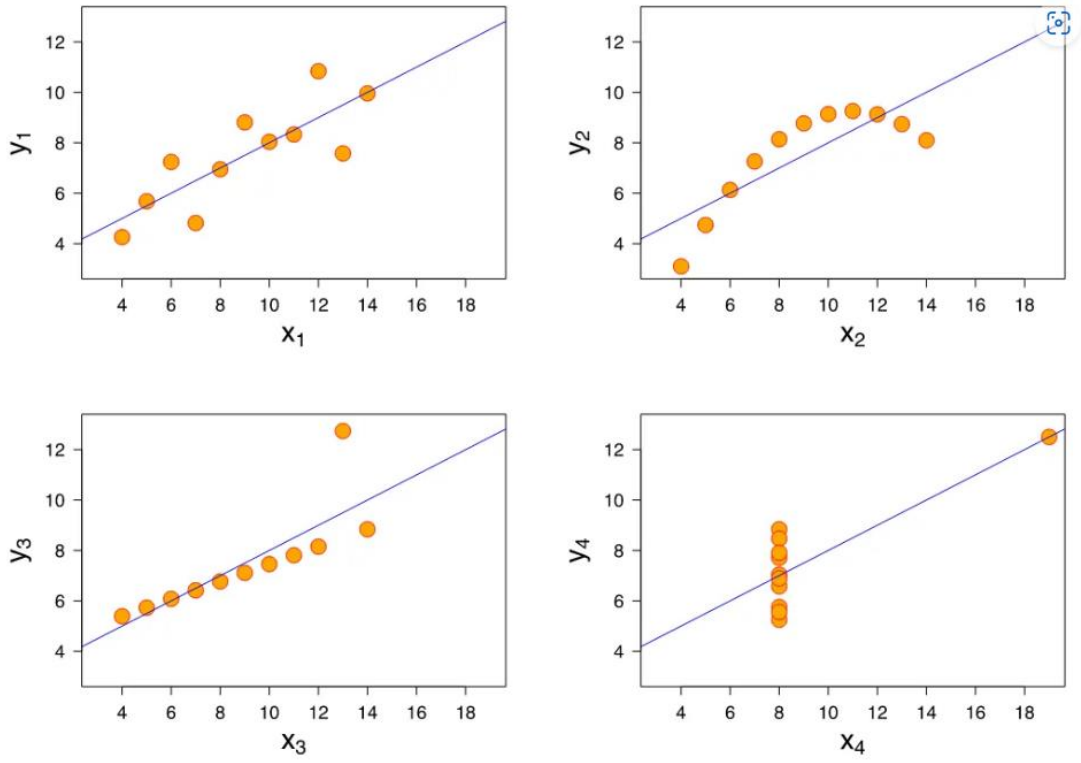
It's a group of four datasets that appear to be similar when using typical summary statistics, are totally different when graphed.Consider below dataset

	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

The statistics when computed for the dataset is similar.Please find the observations

- The average *x* value is 9 for each dataset
- The average *y* value is 7.50 for each dataset
- The variance for *x* is 11 and the variance for *y* is 4.12
- The correlation between *x* and *y* is 0.816 for each dataset

But when plotting these four data sets on an x/y coordinate plane, we get the following results:



Hence it's important to visualize the data to get a clear picture of what's going on.

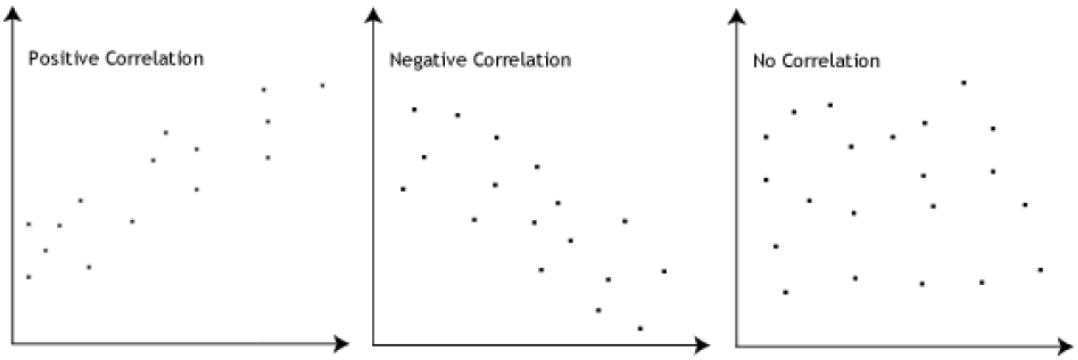
---

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

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.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



Pearson correlation coefficient formula:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

---

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

Answer:

When there are a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very different coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

You can scale the features using two very popular method:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

---

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

Answer:

If there is perfect correlation between two independent variables , then VIF = infinity.
When the value of VIF is infinite it shows a perfect correlation between two independent variables.
In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity.

This can be solved by dropping one of the variables from the dataset which is responsible for this multicollinearity.

---

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

Answer:

Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset.
Quantile, refers to the fraction (or percent) of points below the given value. i.e, 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
.
Importance of Q-Q plot:
A Q–Q plot is used to compare the shapes of distributions, if the dataset follows any particular type Of probability distribution like normal, uniform, exponential.
