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 analysis is performed using the boxplot and bar plot, below are the observations:

- We can say that there are more bookings in Fall season
- The bookings have increased from 2018 to 2019
- Majority of bookings are high from May to October and then starts to dip by end of the year
- Most of the bookings are done when the weather is clear with drastic difference in the Light snow
- Overall bookings are pretty much high in week from Tuesday and increasing till Saturday. The bookings dip on Sunday and is slightly more on Monday.
- It is also evident that on holidays people usually do less bookings than working day

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

Answer:

- If there is an additional column created during dummy variable creation, we can remove it
- Also using drop_first=True ensures that we prevent multicollinearity

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

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? (3 marks)

Answer:

In order to validate the Linear Regression model, we had following assumptions

- Multicollinearity validation
- Significance of the variable
- Verifying the VIF values
- Normally distribution of the errors

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:

Based on the final model, 3 top contributing features are temp, winter, sep

General Subjective Questions

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

Answer:

Linear Regression is a supervised Machine Learning algorithm which predicts a target variable based on the independent variables.

The linear regression equation can be represented as:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn + \varepsilon$

Below are the steps performed In Linear Regression Algorithm:

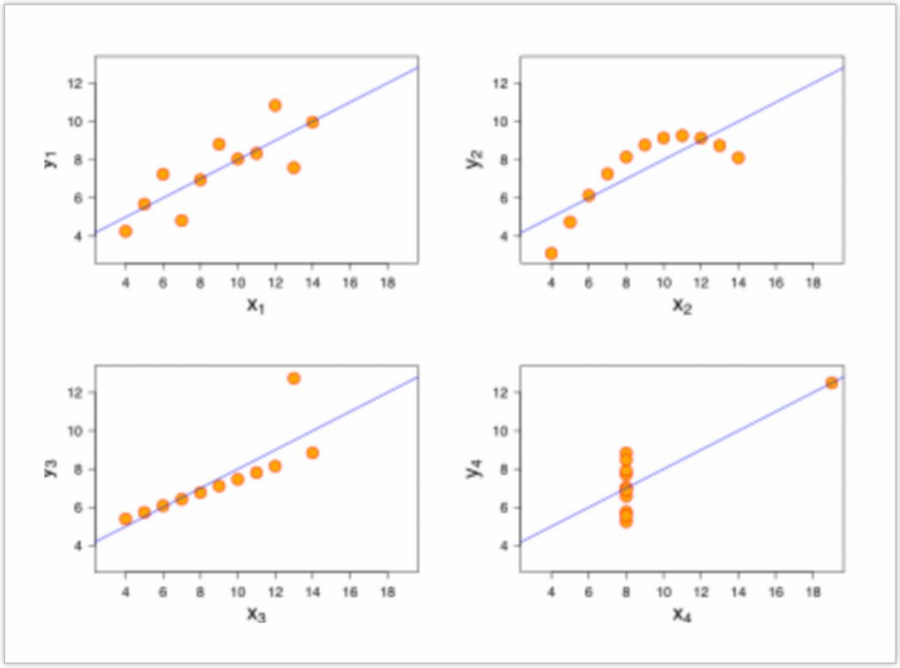
- Data Preparation
- Features Engineering
- Model Training
- Model Evaluation
- Model Tuning
- Model Deployment

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

Answer:

Anscombe's quartet are collection of 4 data sets which have quite identical summary stats (mean, variance and correlation), but have different distributions when plotted graphically. This helps to illustrate the importance of Exploratory Data Analysis and drawbacks of depending only on one variable. Below are the 4 datasets of Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



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

Answer:

Pearson's R is the corelation coefficient which measures the linear co relation between 2 data sets. It is a number between -1 & 1 which measures the strength and direction of relationship between 2 variables.

Pearson correlation coefficient is calculated using the formula r=cov(X,Y)√var(X)√var(Y).

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

Answer:

Scaling is the process of transforming the features of a dataset so they fall in a similar range. They are of 2 types:

- Normalization (Min-Max Scaling) : In this type, the values are transformed within the 0 & 1 range. The formula for Min-Max scaling is given by:
  $X\text{normalized} = \frac{X - X\text{min}}{X\text{max} - X\text{min}}$
  Where:
  X =Original value
  X min = minimum value of feature
  X max = maximum value of feature

- Standardization (Z-score normalization): This scaling process transforms the features to have the mean of 0 and standard deviation of 1. The formula for this scaling is given by:
  $X\text{standardized} = \frac{X - \mu}{\sigma}$
  Where:
  X =Original value
  μ = mean of feature
  σ = standard deviation of feature

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

Answer:

If there is perfect corelation, then it means that there is perfect linear combination which further can be observed by the infinite VIF value. This may happen when we have duplicate variables in dataset or if the datapoints are less compared to the predictors.

We may take following precautions to avoid infinite VIFs:

- Check for duplicates in dataset
- Enough datapoints to support the number of predictors
- Remove the high corelated variables
- Removed perfectly predicted variables

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

Answer:

Q-Q plot also known as Quantile-Quantile (Q-Q) plot, is used to assess if the dataset follows a theoretical distribution. The Q-Q plot compares the quantiles of the observed data to the quantiles of a theoretical distribution. The data is sorted in ascending order.

Q-Q plots are primarily used to assess the normality of residuals. They may also help to predict the outliers. There are 3 types of plots:

- Straight Line Pattern
- Deviation from Line
- S-Shape or Curvature