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: Below are the few points we can infer from the visualization,

- The fall season has attracted more bookings. And Booking counts have increased in the year 2018, 2019.
- According to the graph June, July, September, and October has more booking.
- Clear weather has attracted more bookings.
- Thursday, Friday, Saturday, and Sunday have more bookings.
- The holiday season has more bookings than the non-holiday season.
- An equal number of bookings has been made on working and non-working days.
- In the year 2019, more bookings have been made.

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

Answer: drop_first: It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Answer: The 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:

- Error terms should be distributed normally
- There should be insignificant multicollinearity among variables.
- Linearity should be visible among variables
- There should be no visible pattern in residual values

- No auto-correlation

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

Answer: There are 3 features that contribute significantly towards explaining the demand for the shared bike:

- Temp
- Winter
- September

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

- Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares.
- The mathematical representation of linear regression is as follows:

Y = mX+c,

Where Y is an independent variable, we are trying to predict

X is an independent variable, we are using

m is the slope of the regression line which represents the effect X has on Y, and c is a constant

- There are two types of linear regression, which are as follows:
  - Simple Linear Regression
  - Multiple Linear Regression
- Assumptions:
  - Multicollinearity:
    - A linear relationship exists between two independent variables X and Y, with little or no multicollinearity between the features.

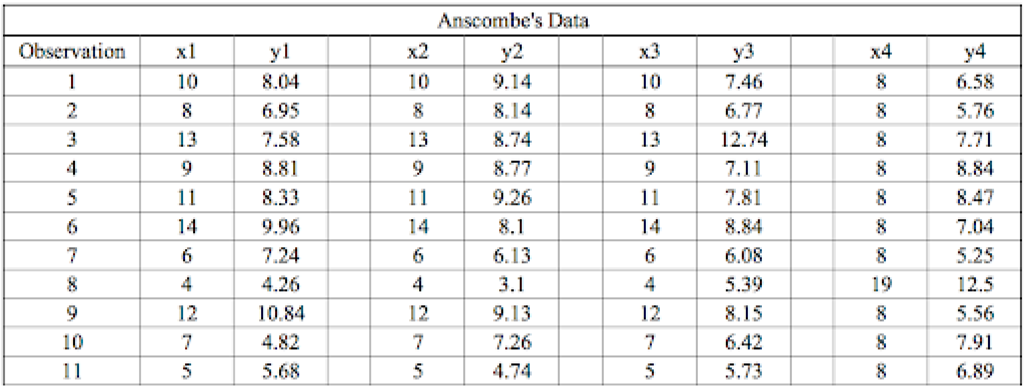
- Auto-Correlations:
  - The given dataset should not be auto-correlated. The phenomenon occurs when residual values or error terms are not independent

- Relationship between variables:
  - The relationship between the response and feature variable must be linear
- Normality of error terms:
  - Error terms should be normally distributed
- Homoscedasticity:
  - There should be no visible pattern in a residual variable.

2. Explain Anscombe's quartet in detail.

Answer:

- Anscombe's quartet was developed by Francis Anscombe. It consists of four datasets.
- It tells the importance of data visualization before applying various algorithms.
- The definition of four plots is as follows:



Anscombe's Data

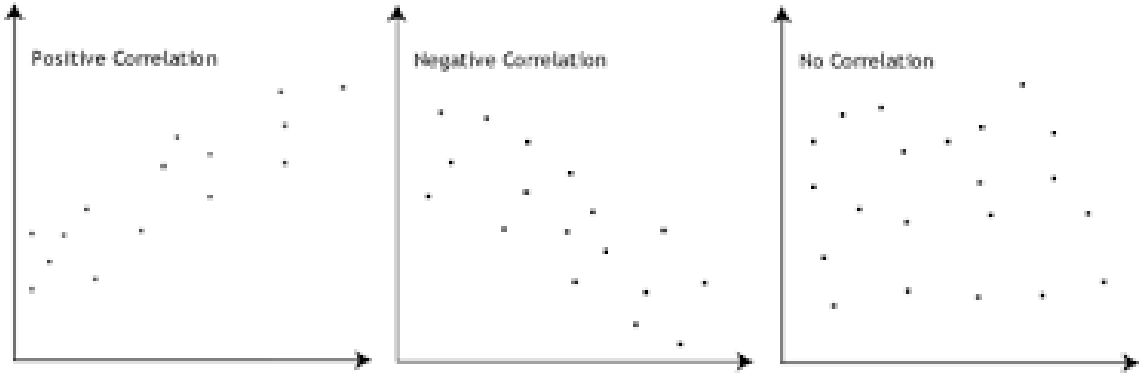
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

- The description of the four datasets is as follows:
  - Dataset 1: Fits the linear regression model
  - Dataset 2: Cannot fit the linear regression model because data is nonlinear.
  - Dataset 3: It shows the outliers involve in datasets, which cannot be handled by the linear regression model.
  - Dataset 4: It shows the outliers involve in datasets, which cannot be handled by the linear regression model.

3. What is Pearson’s R?

Answer:

- Pearson’s R is defined as the measurement of the strength of the relationship between the two variables and their association with each other.
- In simple words, Pearson’s correlation coefficient calculates the effects of change in one variable when another changes.
- It is the practice used for quantifying linear relationships through Pearson’s correlation coefficient.
- The strength and direction of the connection between two variables, take a value between -1 and +1.



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

Answer:

- Feature scaling is defined as a data preprocessing technique used to transform the values of a feature or variable in a dataset to a similar scale.
- It becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude.
- The reasons why scaling is performed are as follows:
  - Gradient Descent based algorithms:
    - Machine Learning algorithms like linear regression that use gradient descent as an optimization technique
  - Tree-based algorithms:

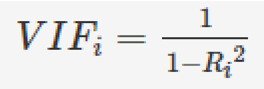
- Tree-based algorithms are insensitive to the scale of features. The decision tree splits a node on a feature that increases the homogeneity of the node.

Normalization	Standardization
Rescales values to a range between 0 and 1	Centres data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/standard deviation
Retains the shape of the original distribution	Changes the shape of the original distribution

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

Answer:

- The formula for calculating VIF is



$$VIF_i = \frac{1}{1-{R_i}^2}$$

  - Where i is the ith variable
- If R- Square value is equal to 1 then the denominator of the above formula become 0 and value becomes infinite.

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

Answer:

- The Q-Q plot is a graphical technique for determining if two datasets come from populations with a common distribution.
- The use of the Q-Q plot is as follows:
  - Q-Q plot is used to see if the points lie approximately on the line. If they don’t, it means our residuals are not Gaussian and hence, our errors are not Gaussian.
- The importance of the Q-Q Plot:
  - The sample size does not need to be equal
  - Many distributional aspects can be simultaneously tested.
  - The Q-Q plot can provide more insights into the nature of difference than analytical method.