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: -

I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –

- The fall season seems to have attracted more bookings. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of May, June, July, Aug, Sep and Oct. Trend increased starting the year till mid of the year and then decreased as we approached the end of the year.
- Clear weather attracted more bookings which seems obvious.
- Thu, Fir, Sat and Wed have a greater number of bookings as compared to the start of the week.
- When it's not a holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy themselves with family.
- Booking seemed to be almost equal either on the working day or non-working day in the year 2018 but in 2019 bookings is high on working days.
- 2019 attracted more number of booking than the previous year, which shows good progress in terms of business.

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

drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. This is done to avoid multicollinearity in the regression model and to improve the interpretability of the coefficients

Syntax -

drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

Let's say we have 3 types of values in Categorical column and we want to create dummy

variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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:

I have validated the assumption of Linear Regression Model based on below 2 assumptions Normality of error terms

- Error terms should be normally distributed

Multicollinearity check

- There should be insignificant multicollinearity among variables.

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:

Below are the top 3 features contributing significantly towards explaining the demand of the
shared bikes –

1. temp
2. year
3. season_winter

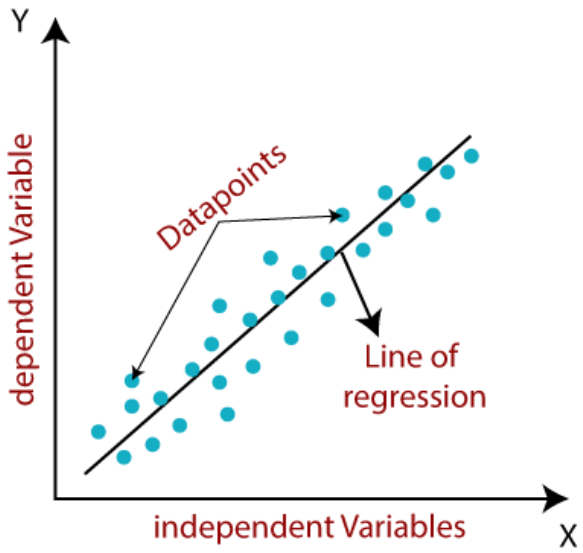
General Subjective Questions

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

Linear regression, a widely used and straightforward machine learning algorithm, offers valuable insights through statistical analysis. It enables predictive modeling for continuous or numeric variables like sales, salary, age, and product price.

By establishing a linear relationship between a dependent variable (y) and one or more independent variables (x), linear regression captures the essence of this connection. It examines how changes in the independent variables influence the dependent variable, providing meaningful interpretations.

The Standard equation of the regression line is given by the following expression:



$Y = mX + C$

Here:

- Y = Dependent Variable (Target Variable)
- X = Independent Variable (predictor Variable)
- C = intercept of the line (Gives an additional degree of freedom)
- m = Linear regression coefficient (scale factor to each input value).

Now there can be positive or negative Linear Relationship.

- **Positive linear relationship: -** When the both independent and dependent variables increases then we can say that the relation is positive linear relationship.

- **Negative linear relationship:** - This type of relationship is formed when independent variables and dependent variable decreases.

Types of Linear Regression:-

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

There are some assumptions of linear regression. There are the formal checks while building a linear regression are as follows: -

1. Multi-collinearity: -

   Linear regression model assumes that there is very low or no multi-collinearity in the data. The problem multi-collinearity occurs when the independent variables or features have dependency in them.

2. Auto-correlation: -

   Another assumption linear regression model assumes is that there is very low or no auto correlation in the data.

3. Relationship between variables: -

   Linear regression model assumes that the relationship between response and feature variables must be linear.

4. Normality of error terms: -

   Error terms should be normally distributed.

5. Homoscedasticity: -

   There should be no visible pattern in residual values

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

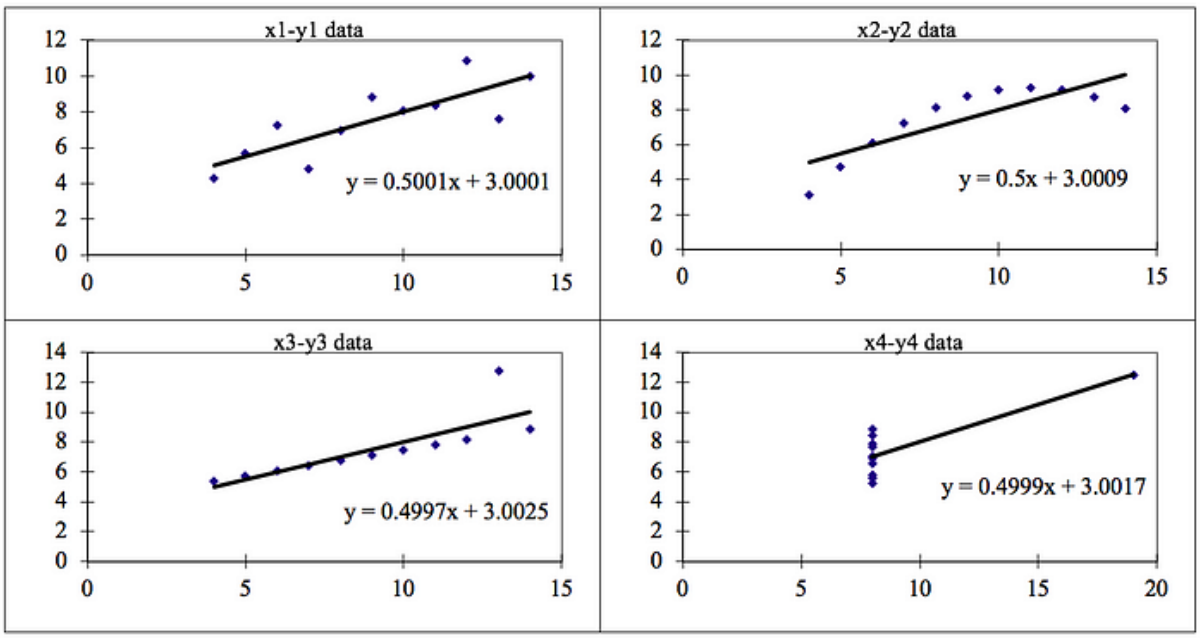
Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

There are four datasets, each consisting of eleven (x, y) pairs. It's important to highlight that these datasets have identical descriptive statistics. However, when visualized on a graph, the narrative changes dramatically—completely, I must emphasize. Each graph presents a distinct story, disregarding their comparable summary statistics.

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
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

The summary statistics show that the means and the variances were identical for x and y across the groups.

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- The variance of x is 11 and variance of y is 4.13 for each dataset.
- The correlation coefficient between x and y is 0.82 for each dataset



The four datasets can be described as:

1. **Dataset 1:** this **fits** the linear regression model pretty well.
2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

3. What is Pearson’s R? (3 marks)
Answer:

The Pearson correlation coefficient (r) is widely used to quantify linear correlations. Ranging from -1 to 1, it indicates 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 scaling and standardized scaling? (3 marks)

Answer:

Feature scaling is a critical data preprocessing step in machine learning. It addresses biases in distance-based algorithms caused by varying feature scales. Scaling is particularly beneficial for improving training and convergence speed in machine learning and deep learning algorithms.

There are two ways to scale the feature

- **Standardizing:-** The Vanables are scaled in such a way that their mean is zero and standard deviation is one
- **MinMax Scaling:-** The variables are scaled in such a way that all the values lle between zero and one using the maximum and the minimum values in the data.

Normalization	Standardization
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

In situations where there is an absolute perfect correlation between variables, the VIF (variance inflation factor) will indeed be infinite. A large VIF value indicates the presence of correlation between variables. If the VIF is -5, it signifies that the variance of the model coefficient is inflated by a factor that warrants investigation. On the other hand, if the VIF is greater than 10, it is advisable to remove the variable because its inclusion leads to high multicollinearity.

When the VIF approaches infinity, it reveals the presence of perfect correlation among independent variables, resulting in an R-squared value of 1. To address this issue, it is necessary to drop one of the variables from the dataset that is causing this perfect multicollinearity.

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

The Quantile-Quantile (Q-Q) plot is a useful graphical tool used to assess whether a given dataset follows a specific theoretical distribution, such as the Normal, exponential, or Uniform distribution. It is also employed to determine if two datasets originate from populations with a shared distribution.

In the context of linear regression, the Q-Q plot can be utilized when we have separate training and test datasets. By examining the Q-Q plot, we can verify whether both datasets are derived from populations with the same underlying distribution. This is particularly valuable in ensuring the validity and reliability of the linear regression model, as it confirms the assumption of consistent distributions between the training and test datasets.
The use of Q-Q plot can be described as follows:

In scenarios where we have two datasets, such as dataset 1 and dataset 2, with different sample sizes, we can employ the Q-Q plot to compare the distributions of specific variables, such as the age variable in this case. This comparison allows us to determine if the distributions of the variables in the two datasets are similar or different.