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:

From the analysis of the categorical features of the given dataset i observed below.

- Most of the bookings were made during the months of May, June, July, Aug, Sept and Oct. Trend increased from the start of the year till the middle of the year and then it started decreasing as we approached the end of year. The number of bookings for each month seems to have increased from 2018 to 2019.
- Fall season seems to have attracted more bookings. And, in each season the booking count has increased drastically from 2018 to 2019.
- Booking seemed to be almost equal either on working day or non-working day. But the count increased from 2018 to 2019.
- Thu, Fir, Sat and Sun 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 with family.
- 2019 attracted a greater number of bookings from the previous year, which shows good progress in terms of business.
- Clear weather attracted more bookings, which seems obvious. And in comparison, to the previous year, i.e. 2018, bookings increased for each weather situation in 2019.

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

When creating dummy variables from categorical data, drop_first=True is used to prevent multicollinearity, where one predictor variable in a regression model can be predicted from the others. This parameter ensures that one level of each categorical variable is dropped, creating n − 1 dummy variables for n categories.

By dropping to the first level, you avoid perfect multicollinearity, where one dummy variable becomes a linear combination of the others. This helps in the interpretation of the coefficients in regression models. It also prevents issues like the "dummy variable trap," where having all levels of a categorical variable as predictors could lead to issues with the model's performance and interpretation.

Ultimately, using drop_first=True helps in improving the performance and interpretation of regression models by eliminating redundant information caused by perfect 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:

For validating the assumptions I’ve done Residual Analysis, In which I basically tested 5 assumptions on the model as shown below,

- **Normality of Error Terms:**
  - This assumption states that the error terms (residuals) should follow a normal distribution. It implies that most of the residuals should be clustered around zero
- **Multicollinearity:**
  - Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. So assumption states that there should not be any significant multicollinearity amongst the features.
- **Linearity:**
  - Linearity assumes that the relationship between the independent variables and the dependent variable is linear. Residuals should be randomly scattered around zero when plotted against predicted values.
- **Homoscedasticity:**
  - Homoscedasticity refers to the assumption that the variance of the residuals should remain constant across all levels of the predictor variables. In simpler terms, there should not be any visible patterns in the residual plot.
- **Independence of Variables:**
  - Independence of variables assumes that the predictor variables used in the regression model are not correlated with each other, meaning there should not be significant auto-correlation.

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,

- temp
- Light_snowrain
- year

General Subjective Questions

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

Answer:

Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with a given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

$$y = mx + c$$

Here,

- Y is the dependent variable we are trying to predict.
- X is the independent variable we are using to make predictions.
- m is the slope of the regression line which represents the effect X has on Y
- c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

And Linear regression can be extended to **Multi linear Regression** as well as shown below,

$$y = \beta 0 + \beta 1 \cdot x1 + \beta 2 \cdot x2 + \ldots + \beta n \cdot xn$$

here, $x1, x2, \ldots, xn$ are independent variables and $\beta 1, \beta 2, \ldots, \beta n$ are their respective coefficients.

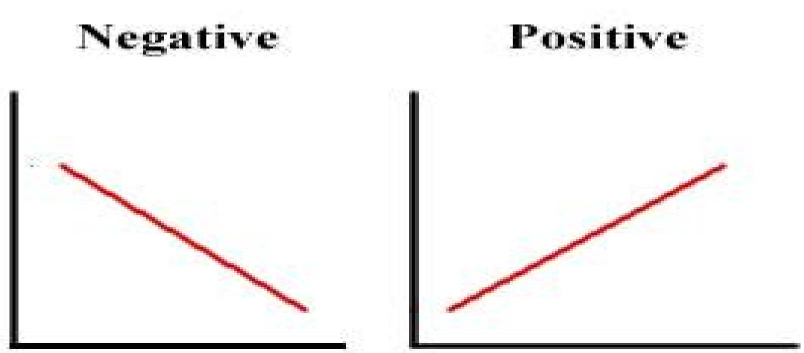
Furthermore, the linear relationship can be positive or negative in nature as explained below−

Positive Linear Relationship:

A linear relationship will be called positive if both independent and dependent variables increase. It can be understood with the help of following graph

Negative Linear Relationship:

A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph



The following are some assumptions about dataset that is made by Linear Regression model −

- **Normality of Error Terms:**
  - This assumption states that the error terms (residuals) should follow a normal distribution. It implies that most of the residuals should be clustered around zero
- **Multicollinearity:**
  - Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. So assumption states that there should not be any significant multicollinearity amongst the features.
- **Linearity:**
  - Linearity assumes that the relationship between the independent variables and the dependent variable is linear. Residuals should be randomly scattered around zero when plotted against predicted values.
- **Homoscedasticity:**
  - Homoscedasticity refers to the assumption that the variance of the residuals should remain constant across all levels of the predictor variables. In simpler terms, there should not be any visible patterns in the residual plot.
- **Independence of Variables:**
  - Independence of variables assumes that the predictor variables used in the regression model are not correlated with each other, meaning there should not be significant auto-correlation.

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

Answer:

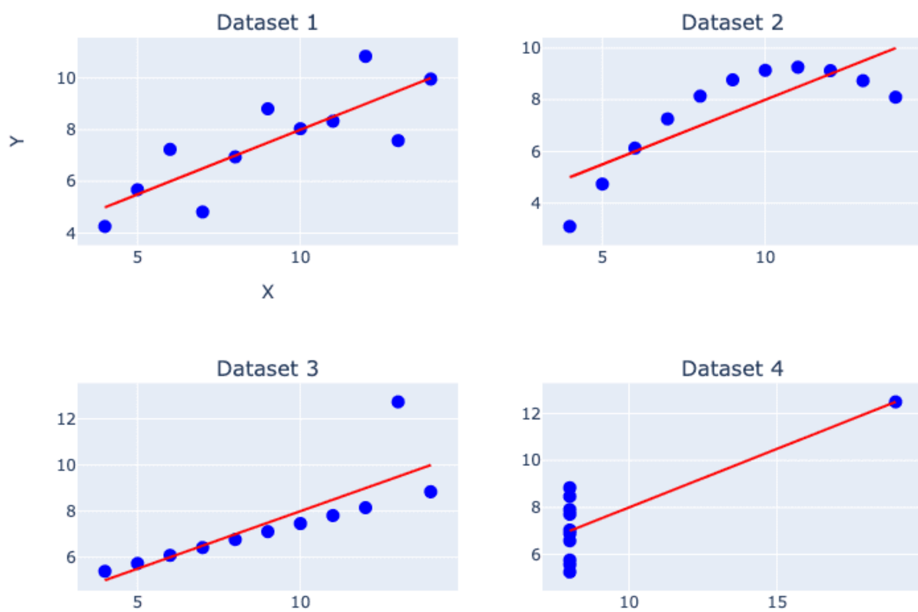
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics

x1	y1	x2	y2	x3	y3	x4	y4
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	8.1	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	3.1	4.0	5.39	19.0	12.5
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

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

	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Mean X	9.0	9.0	9.0	9.0
Mean Y	7.500909	7.500909	7.5	7.500909
Variance X	10.0	10.0	10.0	10.0
Variance Y	3.752063	3.75239	3.747836	3.748408
Correlation	0.816421	0.816237	0.816287	0.816521

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

Answer:

Pearson's correlation coefficient (often denoted as Pearson's r) is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It assesses the linear association between two variables, indicating how much one variable changes when the other changes, assuming a linear relationship between them.

Key points about Pearson's correlation coefficient:

Range: The value of r ranges between -1 and 1.

- $r = 1$ implies a perfect positive linear relationship.
- $r = -1$ implies a perfect negative linear relationship.
- $r = 0$ means no linear relationship exists between the variables.

Direction:

- Positive r values indicate a positive linear relationship (both variables increase or decrease together).
- Negative r values indicate a negative linear relationship (one variable increases while the other decreases).

xMagnitude: The closer r is to 1 or -1, the stronger the linear relationship. A value closer to 0 implies a weaker linear relationship.

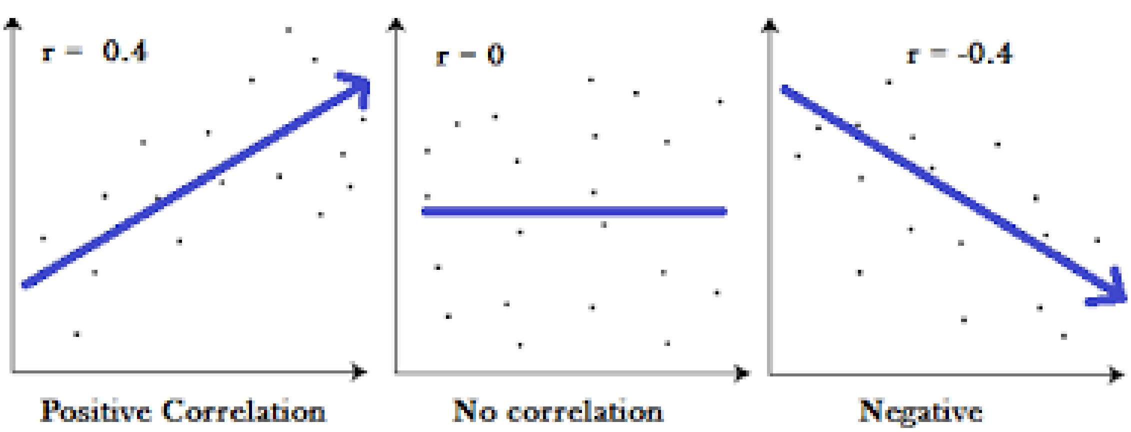
Assumptions: Pearson's r assumes a linear relationship between variables and is sensitive to outliers and non-linear patterns.

The formula for Pearson's correlation coefficient between two variables X and Y with n points is:

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2\Sigma(Y-\overline{Y})^2}}$$

Here, $\overline{X}$ and $\overline{Y}$ represent the means of variables X and Y respectively.

Pearson's r is widely used in various fields like statistics, social sciences, finance, and more to measure the strength and direction of linear relationships between variables.



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

Answer:

Scaling in data analysis refers to adjusting the range or distribution of data to ensure that different features or variables have comparable scales. It's done for several reasons:

Consistent Comparison: Scaling helps in comparing features that have different units or scales. For instance, if one feature ranges from 0 to 1000 and another from 0 to 1, the one with larger values might dominate the analysis. Scaling brings them to a common scale for fair comparison.

Algorithm Performance: Many machine learning algorithms, like SVM, K-nearest neighbors, and neural networks, are sensitive to the scale of input features. Scaling helps these algorithms converge faster and prevents features with larger scales from disproportionately influencing the model.

Normalization and standardization are two common scaling techniques:

SN	Normalization	Standardization
1	We use min and max values of feature for scaling.	Mean and Standard deviation is used for scaling.
2	It Scales values between [0,1] or [-1,1]	There are no such range bounds in standardization.
3	we use this when values are of different scales.	We use this to ensure zero mean and unit standard deviation.
4	Highly effected by outliers	Less/Rarely affected by outliers.
5	MinMaxSclaer() is used in Normalization	StandardScaler is used for Standardization
6	Helps when we don't know the distribution	Used when feature is normal or guassian disributed
7	Called as Scaling normalization	Called as Z-Score normalization

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, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. 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 leads to $\frac{1}{1-r^2}$ infinity. To solve this we need to drop one of the variables from the dataset which 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:

A Q-Q (quantile-quantile) plot is a graphical technique used to assess whether a dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution.

Here's how it works:

1. Quantiles: Quantiles are points taken at regular intervals from a cumulative distribution function (CDF). For example, the median is a quantile that divides the data into two equal parts.

2. Constructing a Q-Q Plot: In a Q-Q plot, the quantiles of the dataset are plotted against the quantiles of a theoretical distribution. If the data matches the theoretical distribution, the points in the Q-Q plot should roughly fall along a straight line.

The use and importance of a Q-Q plot in linear regression are as follows:

1. Assumption Checking: In linear regression, several assumptions are made about the data, including the assumption of normality of residuals (errors). The Q-Q plot helps in checking if the residuals follow a normal distribution. If the residuals are normally distributed, the points in the Q-Q plot will roughly form a straight line.

2. Detecting Departures from Normality: If the Q-Q plot shows a departure from a straight line (e.g., if the points deviate significantly from the line), it suggests that the residuals might not follow a normal distribution. This departure could indicate issues like outliers, skewness, or heavy-tailed distributions in the residuals.

3. Model Validity: Assessing the normality of residuals is crucial in linear regression because violating the assumption of normally distributed errors might affect the validity of statistical inferences and predictions made by the model.

In summary, a Q-Q plot is a valuable tool in linear regression analysis as it helps to visually examine whether the residuals adhere to the assumption of normality. This assessment is important for understanding the reliability of the regression model and the accuracy of its predictions.