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 following observations are come out:

- If we see month wise bookings it has increasing trend from January and its on peak in September and starts decreasing again. Maximum bookings were made in June to October month.
- We had 2018 booking data we can see that initially it started low but increased in 2019. Maximum bookings were made in 2019 than 2018 which is good actually for business it should increase by year.
- Fall season attracted maximum bookings than others. Also, we can see that compare to 2018 each season has increased booking in 2019.
- Clear weather made more peoples to book bikes as they may preferred clear weather than others. Peoples would have refused to go out in rainy weather. And in comparison, to the previous year, i.e. 2018, bookings increased for each weather situation in 2019.
- Within working day and non-working day there is not much difference as booking quite similar in both days.
- Mon, Tue, Fri, and Sun booking were maximum rather than other days.
- On holiday booking were maximum than no holiday may be people want to go out maximum or spent day with family at outside.

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 −1dummyvariables 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' and ‘atemp' variable has highest correlation with the target i.e. ‘cnt’ 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 shared bikes.

- **Temperature(temp)**- As temp as coefficient of 0.4 and 0 p value.
- **Year(yr_2019) –** It has coefficient of 0.2354 and 0 p value.
- **Weather condition (Light Rain) -** As temp as coefficient of -0.2906 and 0 p value. It negatively related to target variable.

General Subjective Questions

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

Answer:

Linear regression may be defined as the statistical model that analyses 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 which is used to make predictions.
- 'm' is the slope of the regression line which represents the effect of X on y
- 'c' is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

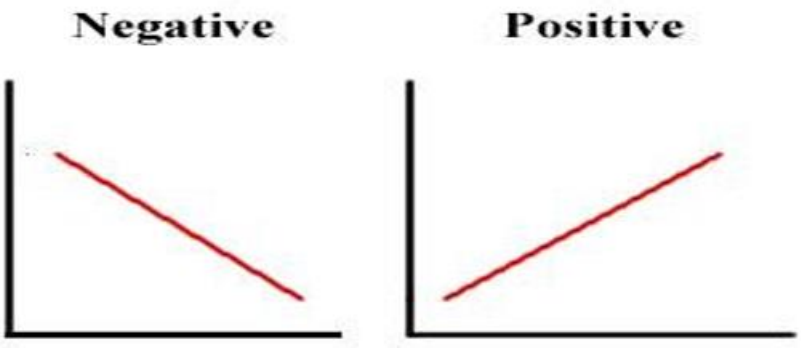
Linear regression can be extended to Multiple linear Regression as well as shown below:

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

here, $x1$, $x2$, …, xn are independent variables and β1, β2, …, βn 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.



Following are some assumptions about the dataset 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

Key Points of Anscombe's Quartet:

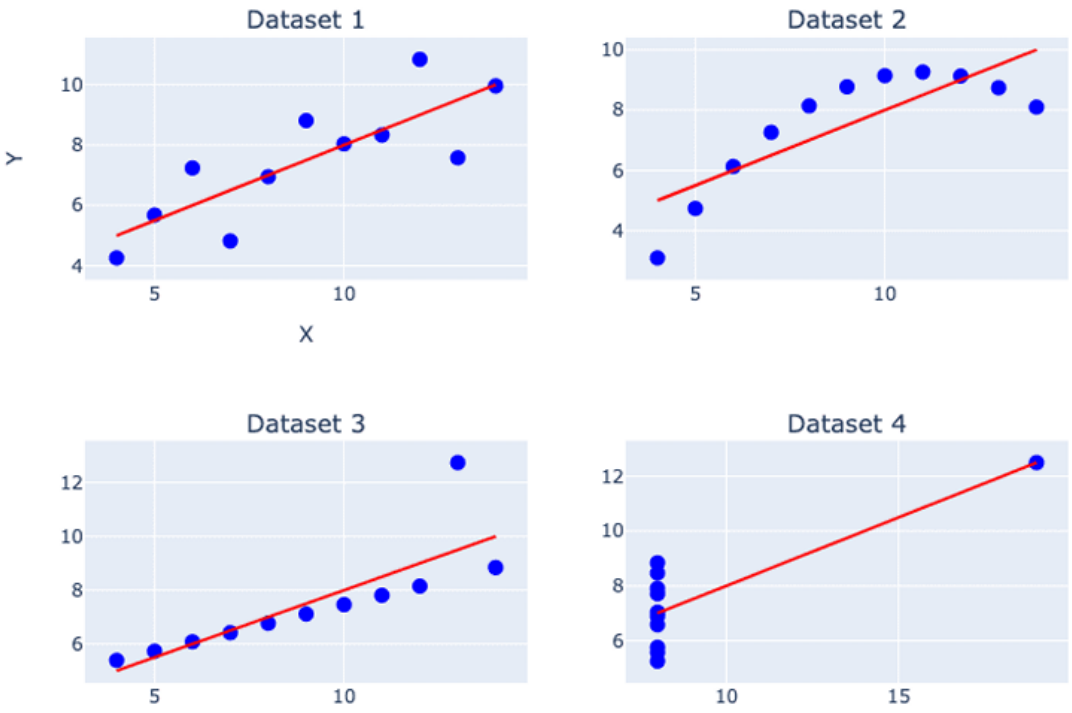
1. **Identical Statistical Properties:**

- All four datasets have the same mean for the x and y values.
- They have the same variance for x and y.
- Each dataset has the same correlation coefficient between x and y.
- The linear regression line (y = mx + c) is nearly the same for all four datasets.

2. **Visual Differences:**

- **Dataset 1:** Shows a typical linear relationship between x and y, which would be expected based on the regression analysis.
- **Dataset 2:** The data is more curvilinear, indicating a non-linear relationship despite the linear regression line.
- **Dataset 3:** Contains an outlier, which heavily influences the regression line, leading to misleading interpretations if only the statistics are considered.
- **Dataset 4:** All x-values are the same except for one, creating a vertical line. The single differing point (an outlier) forces the regression line to fit in a misleading way.

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 appear 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.

Importance of Anscombe's Quartet:

- **Visual Exploration:** The quartet illustrates the crucial role of visualizing data before jumping to conclusions based on summary statistics. By plotting the data, one can identify patterns, outliers, or structures that simple statistics might miss.
- **Misleading Conclusions:** If one relies solely on statistical summaries without visual inspection, one might draw incorrect or oversimplified conclusions about the data.
- **Teaching Tool:** Anscombe's Quartet is widely used in statistics education to teach the importance of graphical analysis and to caution against the over-reliance on summary statistics.

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

Answer:

Pearson's correlation coefficient (often denoted as Pearson's r) is a statistical measure that r 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 change when the other changes, assuming a linear relationship between them.

Key points about Pearson's correlation coefficient:

1. **Range:** The value of 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.
2. **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).
3. **Magnitude:** The closer r is to 1 or-1, the stronger the linear relationship. A value closer to 0 implies a weaker linear relationship.

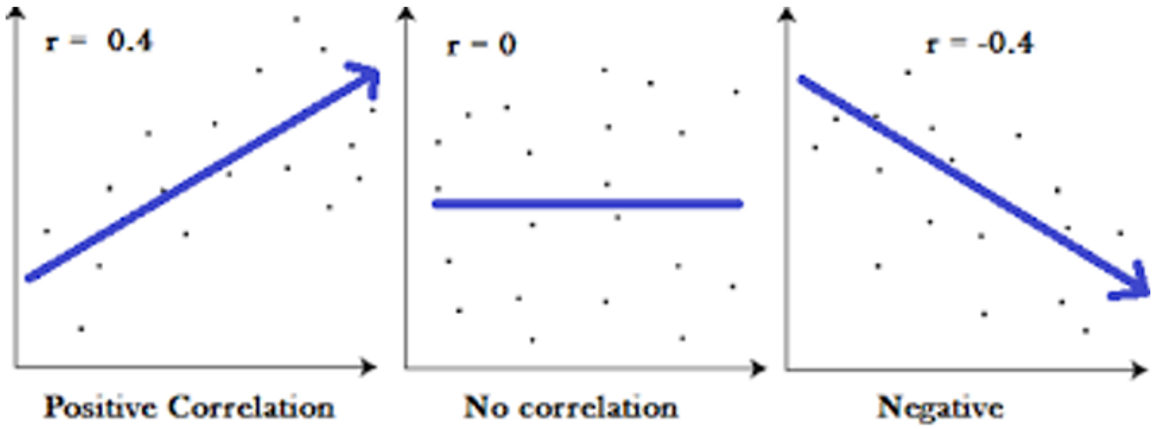
Assumptions: Pearson's r assumes a linear relationship between variables and is sensitive to outliers and non-linear patterns.

Formula: The formula for Pearson's correlation coefficient between two variables and with 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 neighbours, 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 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	Min max scaling used for Normalization	Standard scale is used for standardization
6	It helps when we don't know the distribution	Used when distribution is normal or gaussian distributed
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:

Variance Inflation Factor (VIF) is used to detect multicollinearity in regression models. A high VIF indicates that a predictor variable is highly correlated with other predictor variables, which can make the coefficient estimates unstable and inflate their variances.

When the value of VIF is infinite, it typically points to a specific issue with multicollinearity:

1. Perfect Multicollinearity

- **Definition:** Perfect multicollinearity occurs when one predictor variable is an exact linear combination of other predictor variables. In other words, there is a perfect correlation ($r = \pm 1$) between one predictor and a linear combination of other predictors.
- **Implication:** When perfect multicollinearity is present, the matrix used to calculate the regression coefficients becomes singular or non-invertible. As a result, the calculation of VIF, which involves the inverse of this matrix, can yield an infinite value.

2. Implications for the Model

- **Model Issues:** An infinite VIF indicates a severe problem with multicollinearity. In practical terms, this means the model cannot provide unique estimates for the coefficients of the involved variables due to their perfect linear dependence.

- **Solution:** To address perfect multicollinearity, you need to:
    - **Remove Variables:** Drop one of the perfectly collinear variables from the model.
    - **Combine Variables:** If variables are conceptually related, consider combining them into a single predictor or using dimensionality reduction techniques (e.g., Principal Component Analysis).
    - **Reevaluate Data:** Check the data collection and preprocessing steps to ensure that no redundant or duplicate predictors have been included.

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

Answer:

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a theoretical distribution, commonly the normal distribution. In the context of linear regression, Q-Q plots are particularly useful for evaluating whether the residuals (errors) of the model meet the assumption of normality.

Use and Importance of a Q-Q Plot in Linear Regression

1. **Checking Normality of Residuals:**
    - **Assumption of Normality:** In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. This is important for the validity of hypothesis tests and confidence intervals for the regression coefficients.
    - **Q-Q Plot Usage:** By plotting the quantiles of the residuals against the quantiles of a normal distribution, you can visually assess if the residuals follow a normal distribution. If the plot shows a roughly straight line, this supports the normality assumption.
2. **Detecting Deviations from Normality:**
    - **Systematic Deviations:** If the residuals deviate significantly from the straight line, this may indicate non-normality. Common patterns include:
        - **Heavy Tails:** If points deviate from the line at the extremes, it may suggest heavy tails (leptokurtosis).

        - **Skewness:** A systematic curve (e.g., S-shaped) might indicate skewness in the residuals.
    - **Model Diagnostics:** Identifying these deviations can help you recognize if there are underlying issues with the model or if transformation of the dependent variable or residuals is needed.

3. **Evaluating Goodness of Fit:**
    - **Model Adequacy:** Although the Q-Q plot is primarily used for assessing normality, deviations from normality can hint at inadequacies in the model. For example, it might suggest that some predictors are missing or that a non-linear model might be more appropriate.

4. **Complementing Other Diagnostics:**
    - **Holistic View:** While the Q-Q plot is valuable, it should be used alongside other diagnostic tools (like residual plots, histograms of residuals, and statistical tests) to get a comprehensive understanding of the residuals and the model fit.