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)

Ans –

a) The year 2019 witnessed a higher number of bookings compared to the previous year, indicating positive progress in terms of business.
b) The fall season has experienced a notable increase in bookings. Additionally, across all seasons, there has been a substantial rise in booking counts from 2018 to 2019.
c) On non-holidays, the booking count tends to be lower, which is reasonable as people may prefer spending time at home with family during holidays.
d) It's evident that clear weather conditions (labelled as Good in the notebook) played a significant role in attracting more bookings.
e) There appears to be a relatively equal distribution of bookings between working days and non-working days.
f) Bookings were more prevalent on Thursday, Friday, Saturday, and Sunday compared to the early days of the week.
g) Majority of bookings occurred in May, June, July, August, September, and October. The trend exhibited an increase from the beginning of the year until mid-year, followed by a decrease towards the year's end.

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

Ans- Using drop_first=True during dummy variable creation in Python helps prevent multicollinearity, which occurs when one variable can be linearly predicted from the others. In the context of dummy variables, it means avoiding redundant information. For example, if you have three categories (A, B, C), creating dummies would usually produce three columns, one for each category. However, if you know the values for two categories, you can infer the third. By using drop_first=True, one category is dropped (e.g., A), leaving two columns (B and C). This way, you avoid perfect multicollinearity, which can cause issues in regression models by making it difficult to determine the effect of each predictor accurately.

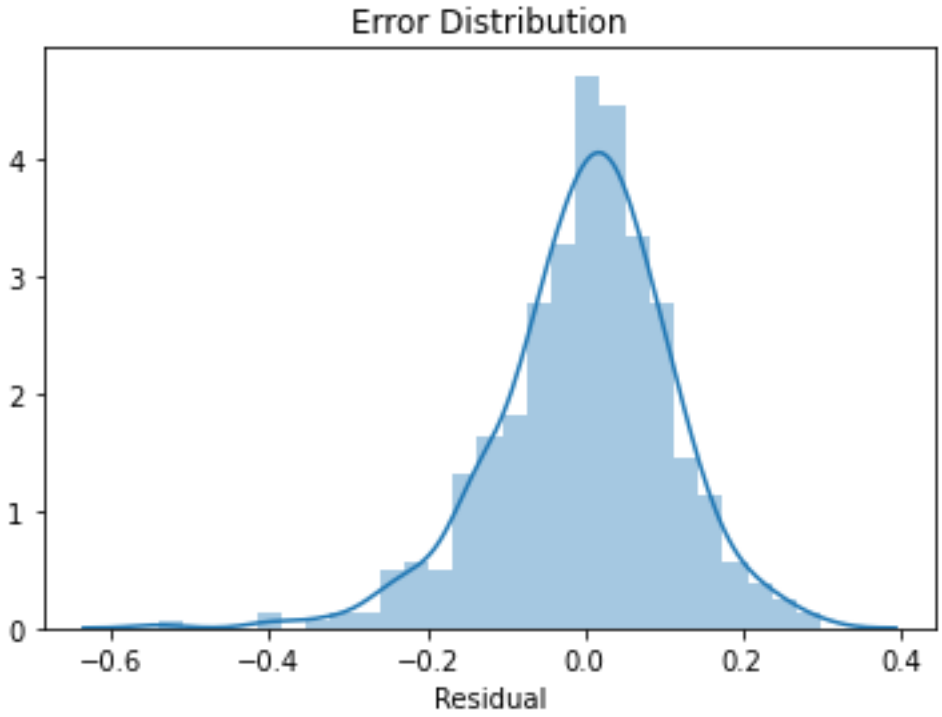
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans- The variable 'temp' exhibits the strongest correlation with the target variable ’cnt’
Given that 'atemp' and 'temp' are redundant variables, only one of them is selected during the determination of the best fit line.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans- To validate the assumptions of Linear Regression after building the model on the training set following steps to be followed:

I. Calculate the residual and see its distribution by plotting a distplot of residuals. It should give a normal distribution and should be centred around O.
II. Validation: Above shown image depicts a normal distribution and the residuals are distributed about mean zero.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans-Top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. Season_fall - A coefficient value of '0.31298' indicated that a unit increase in temp variable increases the bike hire count by 0.31298 units.

2. Season Summer - A coefficient value of '0.271339' indicated that a unit increase in yr variable increases the bike hire count by 0.271339 units.

3. Year (weathersit_3) - A coefficient value of '0.244155' indicated that a unit increase in weathersit_3 variable increased the bike hire count by 0.244155 units.

General Subjective Questions

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

Ans- Linear Regression is a fundamental statistical method used for predictive analysis. It models the relationship between a dependent variable (target) and one or more independent variables (predictors) by fitting a linear equation to observed data. The main types are Simple Linear Regression (one predictor) and Multiple Linear Regression (multiple predictors).

Simple Linear Regression:

It models the relationship between two variables using a straight line:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- y: Dependent variable (response).
- x: Independent variable (predictor).
- β_0: Intercept (value of y when x is 0).
- β_1: Slope (change in y for a unit change in x).
- ϵ: Error term (captures the deviation from the line).

Multiple Linear Regression:

It extends the simple model to include multiple predictors:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where $x_1, x_2, \ldots, x_n$ are independent variables.

Example:

Suppose we want to predict house prices based on size and location. Using multiple linear regression, we could model it as:

$$\text{Price} = \beta_0 + \beta_1(\text{Size}) + \beta_2(\text{Location}) + \epsilon$$

Here, β_1 indicates how much the price changes with each square foot increase, and β_2 shows how location affects the price. The model estimates $\beta_0, \beta_1, \beta_2$ from data, allowing predictions for new input values.

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

Ans- Anscombe's Quartet is a set of four distinct datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression lines. However, when these datasets are visualized, they reveal vastly different distributions and relationships between the variables. The quartet was constructed by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show how different datasets with similar statistical properties can have different structures.

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.

Importance of Anscombe's Quartet:

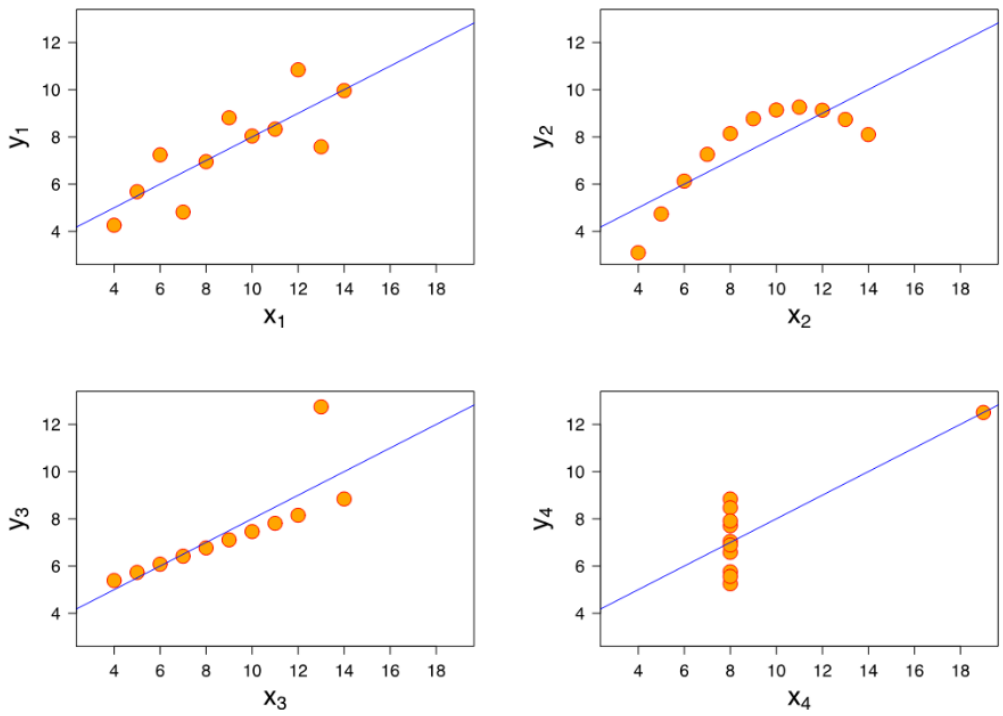
- Visual Exploration: The quartet illustrates the crucial role of visualizing data before jumping to conclusions based on summary statistics. By plotting the data, one can identify patterns, outliers, or structures that simple statistics might miss.

- 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.

Practical Takeaway:

Even though two datasets might have identical statistical properties, their underlying distributions and relationships can be entirely different. Always visualize your data to understand its true nature before relying solely on statistical summaries or models.



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

Ans- Pearson's R in Linear Regression is a measure of the linear correlation between two variables, typically the independent variable (predictor) and the dependent variable (outcome). It quantifies the strength and direction of the linear relationship between these two variables.

- The Pearson's R is calculated as the ratio of the covariance of the two variables to the product of their standard deviations. Mathematically, it is expressed as:

  [R = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}]

where:

- (\text{Cov}(X, Y)) is the covariance of the variables (X) (independent) and (Y) (dependent),
- (\sigma_X) and (\sigma_Y) are the standard deviations of (X) and (Y), respectively.

In the context of linear regression, Pearson's R helps to understand how well the independent variable predicts the dependent variable. A high absolute value of R indicates that the independent variable is a good predictor of the dependent variable, which supports the linear regression model. However, it's important to remember that correlation does not imply causation; a strong correlation doesn't necessarily mean that one variable causes change in the other.

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

Ans- Scaling refers to the process of adjusting the range of features in your dataset so that they can be compared on a common scale. This is particularly important in algorithms where the distance between data points or the magnitudes of features matter, such as in gradient descent optimization, support vector machines, or k-nearest neighbours.

Scaling is performed to:

- **Improve Model Performance:** Algorithms like gradient descent converge faster when features are scaled because the optimization process is more efficient when the features are within the same range.
- **Ensure Equal Contribution:** Scaling ensures that all features contribute equally to the model. Without scaling, features with larger ranges could disproportionately influence the model's predictions.
- **Enhance Interpretability:** It makes it easier to interpret the results of the model, particularly in distance-based algorithms.

- Normalized Scaling:
    - Definition: Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1]. It adjusts the values to be within a specific range without affecting the relative differences between data points.
    - Formula:

      [X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}]
    - Use Case: Normalization is useful when you need to bound the values within a certain range, for instance, in algorithms that need bounded inputs like neural networks.
- Standardized Scaling:
    - Definition: Standardization scales the data based on the mean and standard deviation, resulting in features with a mean of 0 and a standard deviation of 1.
    - Formula:

      [X_{\text{std}} = \frac{X - \mu}{\sigma}]

- where (\mu) is the mean and (\sigma) is the standard deviation of the feature.
    - Use Case: Standardization is preferred when features have different distributions or when the model assumes that the data is normally distributed (e.g., in linear regression or logistic regression).

In summary, normalization constrains data within a specific range, making it useful for bounded inputs, while standardization adjusts data based on statistical properties, making it more suitable when the data's distribution is important.

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

Ans- The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to the presence of multicollinearity among the independent variables. A VIF value becomes infinite when perfect multicollinearity is present, meaning that one independent variable is an exact linear combination of one or more other independent variables.

Cause of Infinite VIF:

An infinite VIF occurs when the correlation between one independent variable and a combination of the other independent variables is perfect (correlation coefficient of 1 or -1). In this case, the model cannot distinguish between the perfectly correlated variables, and the variance of the affected variable's coefficient is infinitely inflated. Mathematically, this happens when the determinant of the matrix (X'X) used to compute VIF is zero, leading to a division by zero.

Implication and Solution :

- **Implication:** An infinite VIF indicates severe multicollinearity, which can destabilize the regression coefficients, making them highly sensitive to changes in the model. This undermines the reliability of the model and can lead to incorrect interpretations.
- **Solution:** To address infinite VIF, you may need to:
    - **Remove one of the perfectly correlated variables** from the model.

- **Combine the correlated variables** into a single feature through techniques like Principal Component Analysis (PCA).
- **Rethink the model design** to avoid including redundant predictors.

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

(3 marks)

Ans- Quantile-Quantile (Q-Q) plot, is a graphical technique to help assess if a set of data came from some theoretical distribution such as Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution

Interpretation of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

• **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis.

• **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.

• **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.

• **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis.

Uses of Q-Q plot:

• can be used with sample sizes also.

• Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

Q-Q plot in linear regression is important because:

when training and test data sets are received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.