Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Here is the analysis of categorical variables on dependent variable:

- Majority of the bike bookings were happened in Seasjon3: "fall". This was followed by summer and winter seasons
- Majority of the bike booking were happened in the months of May, Jun, Jul, Aug & Sep with a median of over 4000 booking per month.
- Weekday variable shows almost similar trend for all the days having their independent medians between 4000 to 5000 bookings.
- Majority the bike bookings were happened during 'Clear' with a median of close to 5000 bookings. This was followed by Misty with 30% of total booking.
- Majority of the bike booking were happened when it is not a holiday based on the median of bookings
- Approximately 70% of the bike booking were happened on 'workingday' with a median of close to 5000 booking
- Year 2018 has average bookings of approx 3800 whereas Year 2019 has average bookings of approx 6000

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

1) The condition **drop_first=True** place vital role in dummy variable creation as it helps in creating an additional column.
2) If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.
3) Hence if we have categorical variable n levels then we need to use n-1 columns to represent the dummy variables

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

'temp' has highest correlation with target variable

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

I’ve validated the assumptions of Linear Regression after building the model on training dataset with below specified checks:

1) Normality of Error Terms: Error terms are normally distributed based on histogram drawn
2) Homoscedasticity: There is no visible pattern on residual values from the graph drawn
3) Multicollinearity: From the final model VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 5
4) Auto Correlation: Durbin-Watson value of final model lr6 is 2.027, which signifies there is no autocorrelation.
5) Linear Relationship: There is clear linear relationship among the variables

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Based on the final model, top 3 features contributing significantly towards explaining the demand of the shared bikes are :

1) Temp
2) Year
3) Windspeed

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear Regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal is to find the best-fitting line (or hyperplane in the case of multiple predictors) that predicts the dependent variable based on the values of the independent variables.

The equation of a simple linear regression line is:

$$Y=mX+b$$

Where:

- Y = Dependent variable
- X = Independent variable (the predictor)
- m = Slope of the line
- b = Y-intercept (the value of Y when X=0)

Types:

- **Simple Linear Regression**: One independent variable. ($Y = \beta_0 + \beta_1 X$)
- **Multiple Linear Regression**: Multiple independent variables. ($y=\beta 0+\beta 1x1+\beta 2x2+...+\beta pxp+ \epsilon$)
- **Polynomial Regression**: A form of linear regression where predictors are transformed into polynomial terms.

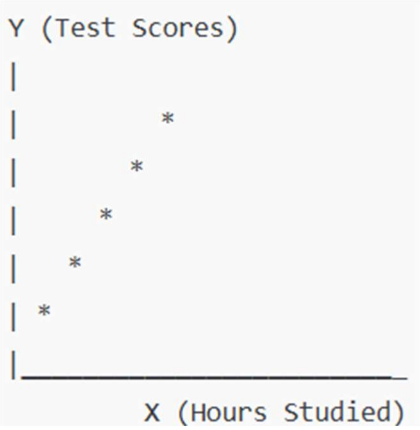
Steps Involved in Simple Linear Regression:

1. Data Collection
2. Data Quality Checks
3. Handling Outliers
4. Data Preparation
5. Splitting the data into train & test datasets
6. Rescaling the Features
7. Building the Linear Model
8. Residual Analysis
9. Model Predictions
10. Model Evaluation

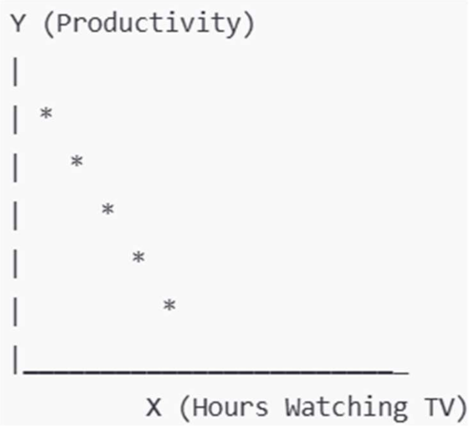
Assumptions:

1. **Linearity**: Relationship between variables is linear.
2. **Independence**: Residuals must be independent.
3. **Homoscedasticity**: Constant variance of residuals.
4. **Normality of Residuals**: Errors are normally distributed.
5. **No Multicollinearity**: Independent variables shouldn't be highly correlated with each other.

Positive & Negative Linear Regressions Example:



```
Y (Test Scores)
|
|         *
|       *
|     *
|   *
| *
|__________________________
        X (Hours Studied)
```



```
Y (Productivity)
|
| *
|   *
|     *
|       *
|         *
|_________________________
         X (Hours Watching TV)
```

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four different data sets that have the same basic statistical properties, like mean, variance, and correlation, but look very different when plotted on a graph. It was created by the philosopher and statistician Francis Anscombe in 1973 to show the importance of visualizing data before analyzing it using just numbers.

Each dataset in the quartet has:

- The same average value for the x and y variables.
- The same correlation between x and y.
- The same variance for both x and y.

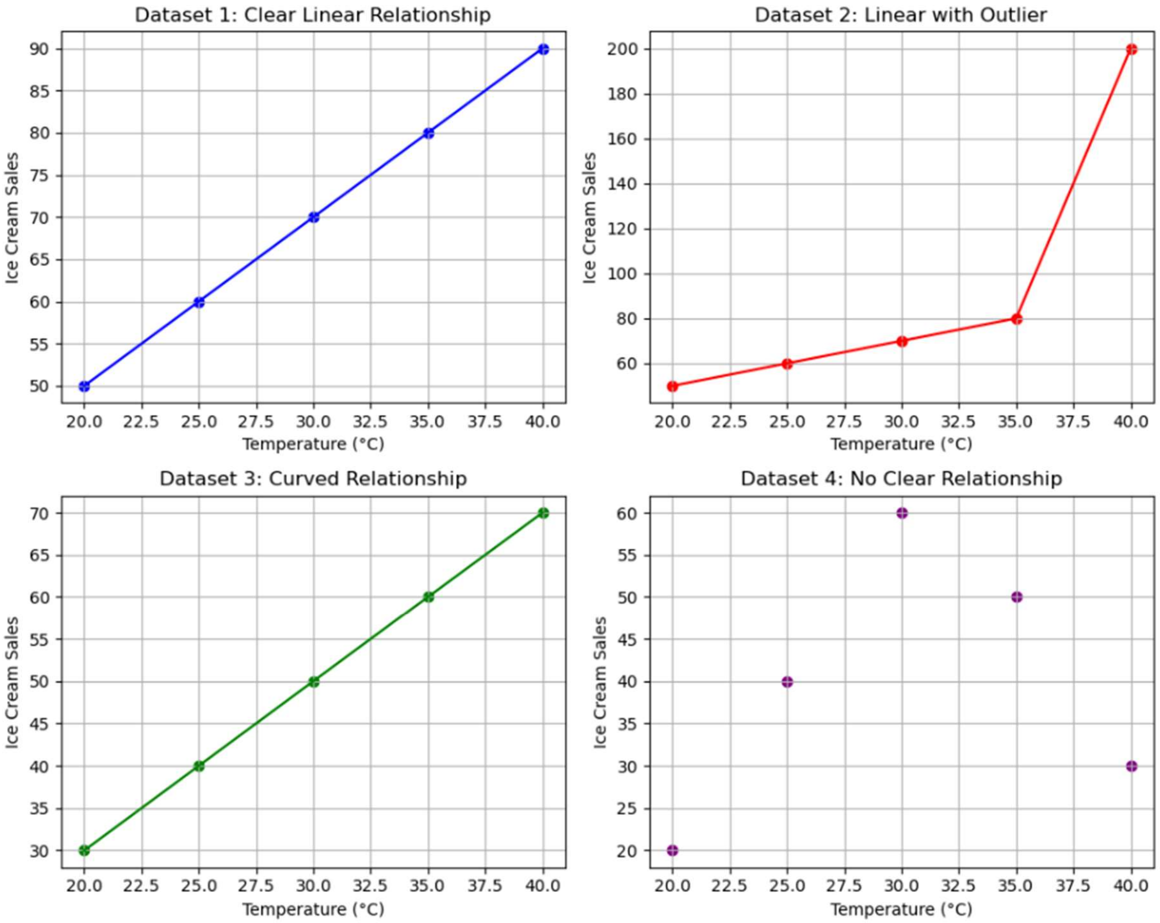
Despite having these identical statistics, when we plot each dataset, they look totally different. Some appear as straight lines, while others form curves or have outliers. This highlights how relying solely on summary statistics (like the mean or correlation) can be misleading, and it shows why it’s important to visualize data before drawing conclusions.

Dataset 1: A Clear Linear Relationship
Temperature (x): 20, 25, 30, 35, 40
Ice Cream Sales (y): 50, 60, 70, 80, 90

Dataset 2: A Linear Relationship, But with an Outlier
Temperature (x): 20, 25, 30, 35, 40
Ice Cream Sales (y): 50, 60, 70, 80, 200

Dataset 3: A Curved Relationship
Temperature (x): 20, 25, 30, 35, 40
Ice Cream Sales (y): 30, 40, 50, 60, 70

Dataset 4: No Clear Relationship
Temperature (x): 20, 25, 30, 35, 40
Ice Cream Sales (y): 20, 40, 60, 50, 30



Anscombe's Quartet shows us a powerful insight about data analysis:

- Summary statistics (like means, variances, and correlation) can tell you a lot about a dataset, but they don’t tell the whole story.
- The context and the visual representation of data can reveal important patterns, anomalies, and relationships that you might miss if you only focus on numbers.
- Outliers and non-linear relationships can distort the interpretation of the data. A good graph will help uncover these issues.

The Key Takeaways:

In the ice cream example, even though all four datasets have the same statistical properties (mean, variance, and correlation), they look very different when plotted. Some show clear trends, while others don’t. This is why it's crucial to visualize data before making conclusions based on statistics alone.

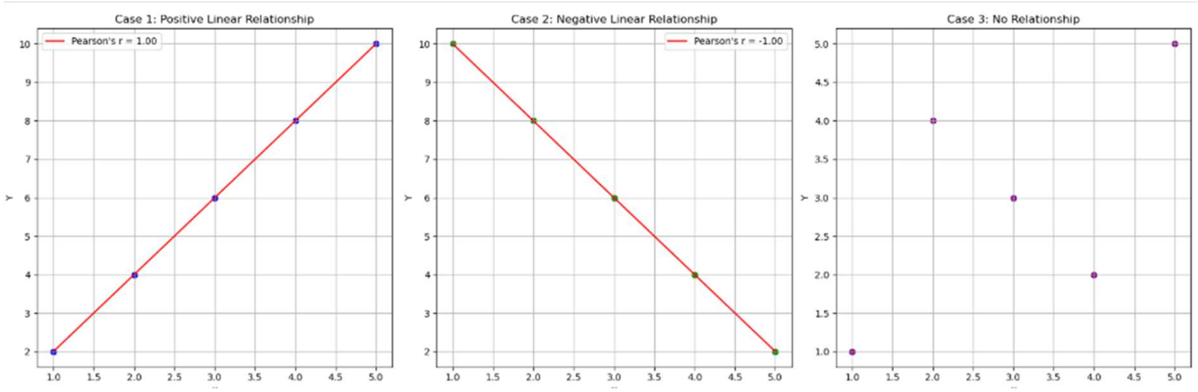
Question 8. What is Pearson’s R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

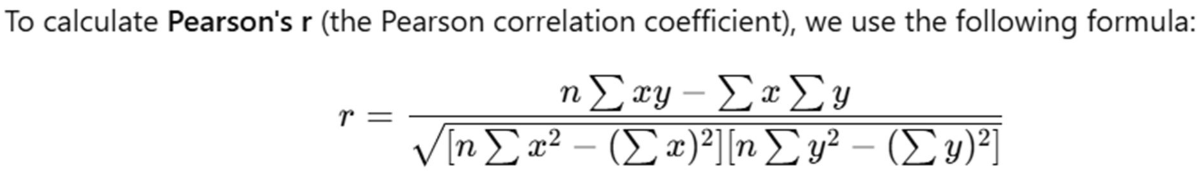
<Your answer for Question 8 goes here>

Pearson's R, also known as the **Pearson correlation coefficient**, is a statistic that measures the strength and direction of a **linear relationship** between two variables. It’s a number that tells you how closely the data from two variables are related to each other in a straight-line fashion.

Here's how to think about it:

- **Values of Pearson's r** range from **-1 to 1**.
  - **r = 1** means there's a perfect positive linear relationship (as one variable increases, the other increases in a perfectly straight line).
  - **r = -1** means there's a perfect negative linear relationship (as one variable increases, the other decreases in a perfectly straight line).
  - **r = 0** means there is no linear relationship between the variables—changes in one variable don't predict changes in the other in a straight-line fashion.





To calculate **Pearson's r** (the Pearson correlation coefficient), we use the following formula:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Feature Scaling is a technique used to standardize or normalize the range of independent variables (features) in a dataset. It is an essential step in the data pre-processing phase of machine learning. This is typically done to ensure that each feature contributes equally to the model or analysis.

When the values of different features (variables) vary widely, scaling helps in bringing them to a comparable scale, improving the efficiency and performance of machine learning models.

Types of Scaling:

There are two common types of scaling methods:

1. **Normalized Scaling**
2. **Standardized Scaling**

Normalized Scaling (Min-Max Scaling)

Normalization (also called Min-Max scaling) refers to transforming the data such that it fits within a predefined range, typically between 0 and 1, or -1 and 1. It is sensitive to outliers. Easier to interpret. Rescales data proportionally between minimum and maximum values

Standardized Scaling (Z-Score Scaling)

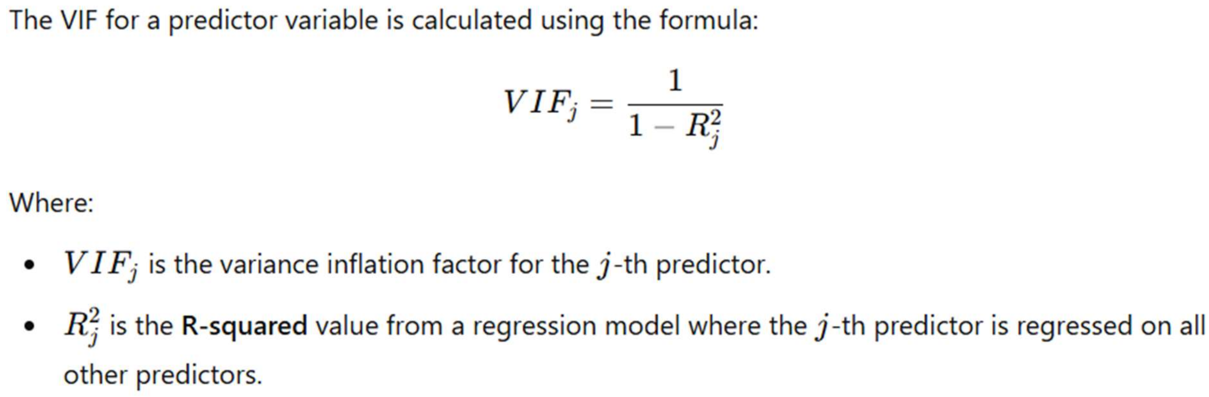
Standardization (or Z-score scaling) refers to transforming the data so that it has a **mean of 0** and a **standard deviation of 1**. This is done by subtracting the mean of the feature and then dividing by the standard deviation. It is less sensitive to outliers because is relies on mean and SD. Centres data around zero with a scale based on standard deviations

---

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The **Variance Inflation Factor (VIF)** is used to measure how much the variance of a regression coefficient is inflated due to multicollinearity, or correlation, among the predictor variables in a regression model. A high VIF value indicates that a predictor variable is highly correlated with other predictors, leading to instability in the regression estimates.



The VIF for a predictor variable is calculated using the formula:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where:

- VIF_j is the variance inflation factor for the j-th predictor.
- R_j^2 is the **R-squared** value from a regression model where the j-th predictor is regressed on all other predictors.

VIF can become infinite when the R-squared value R_j^2 for a predictor variable is equal to 1, or very close to 1. This happens in cases where one predictor is perfectly (or nearly perfectly) correlated with one or more other predictors in the model.

- If $\mathbf{R_j^2}$ =1, it means that the predictor variable X_j can be perfectly predicted using the other predictor variables in the model.
- In this case, the denominator of the VIF formula becomes 0 (1− $\mathbf{R_j^2}$ =0), leading to an infinite VIF.

Why Does This Happen?

1. **Perfect Multicollinearity**: Infinite VIF typically occurs due to **perfect multicollinearity**, where one predictor is an exact linear function of one or more other predictors. .
2. **Redundant Predictors**: Sometimes, redundant or duplicate features in the dataset, such as including both a variable and its transformation can create perfect correlation.
3. **Data Issues**: Certain data issues, such as errors or incorrect feature engineering, can lead to artificially high correlations between variables, which in turn causes the VIF to be infinite.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the sample data to the quantiles of a theoretical distribution. In the plot:

- The **x-axis** represents the quantiles of the theoretical distribution (e.g., normal distribution).
- The **y-axis** represents the quantiles of the observed data.

In a **normal Q-Q plot**, for example, if the data points closely follow a straight line, it suggests that the data follows a normal distribution. If the points deviate significantly from the line, it indicates that the data does not follow the assumed distribution.

To construct a Q-Q plot, you:

1. **Sort** the observed data in ascending order.
2. **Calculate** the quantiles of the observed data and the theoretical distribution.
3. **Plot** the quantiles of the observed data against the quantiles of the theoretical distribution

Q-Q Plot in Linear Regression

In linear regression, a **Q-Q plot** is often used to check the residuals (the differences between observed and predicted values) for normality. The assumptions in linear regression include that:

1. The errors (residuals) should be normally distributed.
2. The residuals should have constant variance (homoscedasticity).

Importance of Q-Q Plot in Linear Regression:

1) Checking Residual Normality:
2) Model Diagnostics
3) Identifying Outliers
4) Evaluating Model Assumptions