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)

I have done categorical variables analysis using boxplot and following are the findings.

- Fall season has attracted the highest number of bookings.
- July, August, September, October has been the peak season for bookings. Bookings have started increasing since January and then peak reached in October then started decreasing.
- Maximum bookings happened when weather conditions were ‘clear’.
- During 2019, received more bookings in comparison to previous year.

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)

It helps in reducing the extra column created during the dummy variable creation.
Consider an example where there is a categorical variable with ‘k’ features, then we will use the **drop_first=True** argument and it will only create k-1 dummy variables, it avoids multicollinearity and still captures all the necessary information.

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 the highest correlation coefficient with the 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)

- Error terms should be Normally distributed with mean zero
- There should be insignificant Multicollinerality among variables
- The relationship between predictors and response is linear.
- No pattern in residual values
- Variance of residuals should be constant across all levels of predicted values.

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)
List of top 3 features explaining the bikesharing demand -:
'temp'
'sept'
'winter'

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)

Linear regression is a supervised machine learning algorithm which is used to model the linear relationship between dependent variable and one or more independent variables. As the value of independent variables increases or decreases, the value of the dependent variable varies as well. Goal is to predict the best fitted linear equation that predicts the dependent variable(y) based on the independent variables (x1,x2…..xn)
- Equation of Simple linear regression is defined as y = a + bx + c
- Equation of Multiple linear regression is defined as y = a + b1x1 b2x2 + b3x3 + c
y - Dependent/Target variable
xi - Independent variables
a - Intercept of line equation where x=0
c - Error

Linear Regression Steps:: Regression algorithm takes the following assumptions.
- Linearity: Relationship between predictors and response is linear.
- **Independence**: Observations are independent.
- **Normality of Errors**: Residuals are normally distributed.
- **No Multicollinearity**: Predictors are not highly correlated.
- **Homoscedasticity**: There is no pattern visible in residual values.

To find the best fitted line, the algorithm minimizes the difference between the actual and predicted values via minimizing the **Sum of Squared Errors (SSE)**:

SSE = Sum of Squares of all values (y_predicted - y_actual)

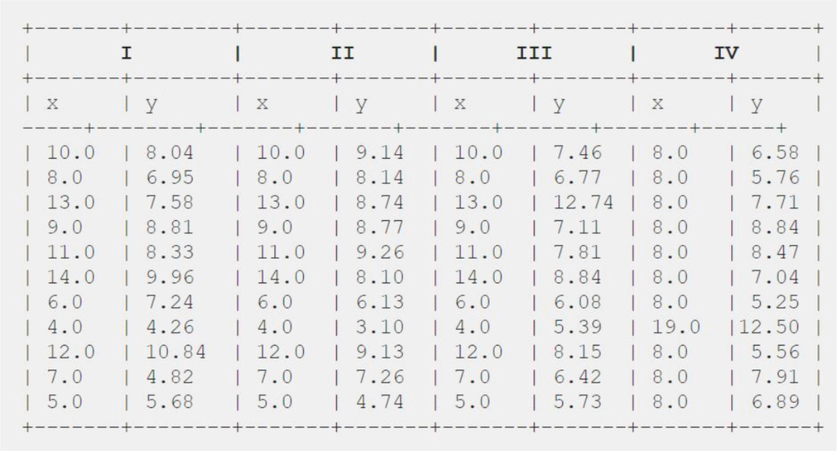
To optimize the Error between predicted and actual values there various algorithms which model can leverage to achieve results e.g. Ordinary Least Squares(OLS), Gradient Descent etc…

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)

Anscombe's Quartet is a set of four datasets that have nearly identical statistical properties, such as mean, variance, correlation coefficient, and linear regression line, yet appear very different when visualized.

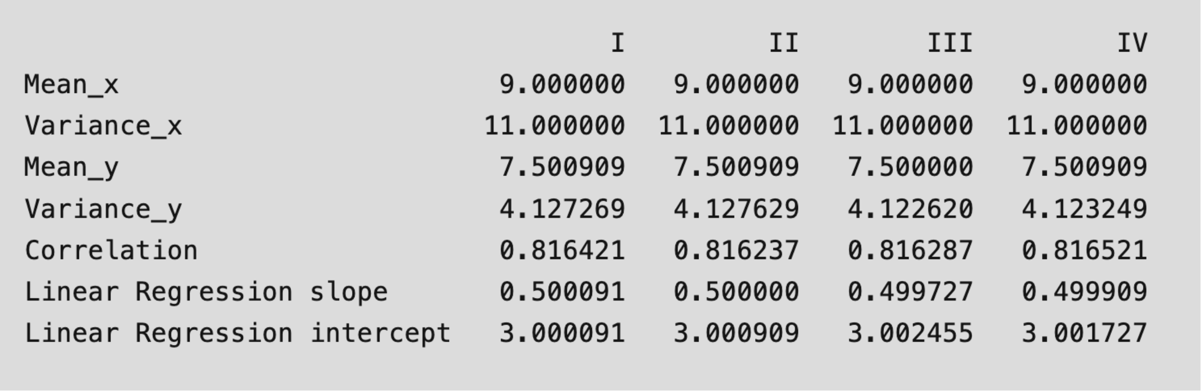
Purpose of Anscombe's Quartet:
1: Summary statistics such as mean, variance, correlation, and linear regression coefficients do not always capture the true nature of the data.
2: Graphical analysis reveals patterns, relationships, or anomalies (e.g., outliers, non-linearity) that are not apparent from numerical summaries alone.
3: Anscombe's quartet demonstrates the importance of combining numerical and visual analysis to understand datasets better.

Consider the following 4 dataset:

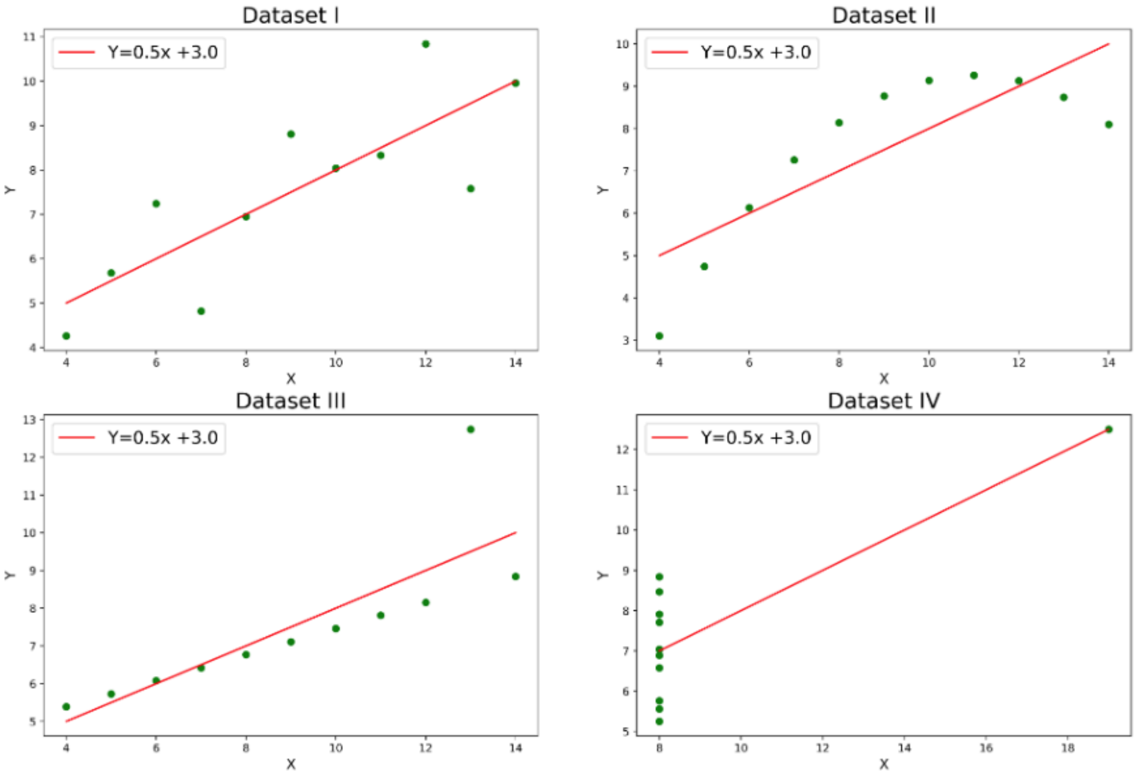


```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 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  | 8.10  | 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   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 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 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

Following the calculated variables of each dataset:



```
                                    I          II         III          IV
Mean_x                       9.000000    9.000000    9.000000    9.000000
Variance_x                  11.000000   11.000000   11.000000   11.000000
Mean_y                       7.500909    7.500909    7.500000    7.500909
Variance_y                   4.127269    4.127629    4.122620    4.123249
Correlation                  0.816421    0.816237    0.816287    0.816521
Linear Regression slope      0.500091    0.500000    0.499727    0.499909
Linear Regression intercept  3.000091    3.000909    3.002455    3.001727
```



Dataset I

Dataset II

Dataset III

Dataset IV

DataSet1:: There seems to be linear relationship between between x and y
DataSet2:: There is non-linear relationship between between x and y
DataSet3:: There is perfect linear relationship between between x and y
DataSet4:: All the data points distributed to the bottom.

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)

Pearson's R, also known as the **Pearson correlation coefficient**, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is represented by the symbol r and ranges from −1 to +1

Interpretation of Pearson's R

Value of r	**Interpretation**
r = +1	Perfect positive correlation
r = -1	perfect negative correlation
r = 0	No linear correlation
0 < r < 1	Positive linear correlation
-1 < r < 0	Negative linear correlation

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)

Scaling is the process of adjusting the range and distribution of feature values in a dataset. It ensures that all features contribute equally to a model's learning process by bringing them to the same scale.

Scaling is performed because of following reasons:
1- To Improve Model Performance:: Without scaling some features with larger ranges dominate the learning process, leading to biased results.
2- To Speed up Training:: In many optimization-based algorithms like Gradient Descent, scaling ensures faster convergence by reducing the distortion caused by features with different magnitudes.
3- Prevent Numerical Instability: Models may face numerical instability (e.g., overflow or underflow) when handling large or small feature values.

Difference between normalized scaling and standardized scaling

#	Normalized	Standardized
1	Adjust data to specific range e.g. [1, -1]	Scaled data is not bound within a range
2	Minimum and Maximum values of features are used for scaling. e.g. (X - X_min)/ (X_max - X_min)	Mean and Standard deviation is used for scaling. e.g. X - mean/sigma
3	It is highly sensitive to Outliers	It is less sensitive to

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)

Infinite values for VIF due to the perfect multicollinearity. This happens when two or more independent variables in a model are perfectly linearly dependent. That is, one independent variable in the model can be entirely predicted by another independent variable.

VIF for Xi = 1/1-R-Squarei

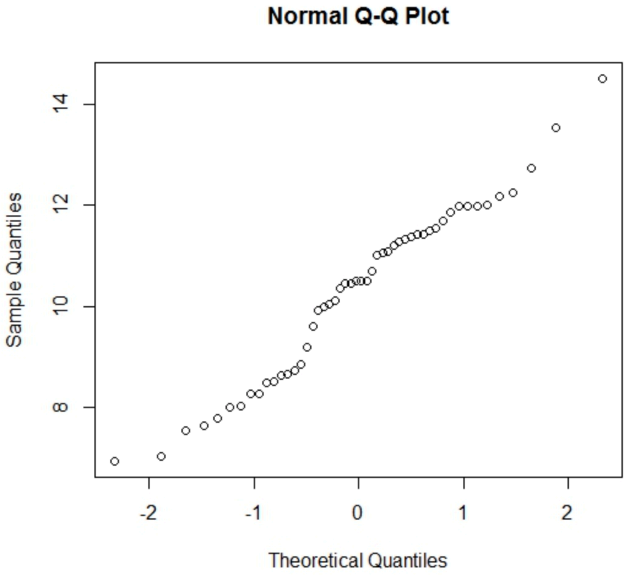
If R-Squarei = 1, it means that Xi can be perfectly predicted by a linear combination of the other predictors indicating perfect multicollinearity.

To Address infinite VIF, need to perform the following::
1- Remove the perfect multicollinearity
2- Drop one dummy variable
3- Remove or combine collinear predictors

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)

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset against a theoretical distribution (commonly the normal distribution). It plots the quantiles of the data against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points will lie approximately on a straight 45-degree line.

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



Importance of Q-Q plot in Linear Regression:

Model Diagnostics:
Ensures that the model satisfies the assumption of normality for residuals.
Detects deviations that may require remedial actions

Detecting Outliers:
Extreme deviations from the 45-degree line indicate potential outliers.

Improving Model Fit:
Identifying non-normality may suggest transformations (e.g., log, square root) or alternative models to better fit the data.