1. **From your analysis of the categorical variables from the dataset, what could you infer about**

their effect on the dependent variable? (3 marks)

I have conducted analysis on the categorical columns using box plots and bar plots, and the visualizations reveal several noteworthy insights:

- The fall season appears to have garnered more bookings.
- Within each season, the booking count has experienced a significant increase from 2018 to 2019 with an increase of 63 %
- Clear weather/ Misty conditions appear to attract more bookings
- The majority of bookings occurred during the months of May, June, July, August, September, and October. There is a noticeable upward trend from the beginning of the year until the middle, followed by a gradual decline towards the year's end.
- On non-holidays, booking numbers tend to be lower, which is reasonable as people may prefer to spend time at home with family during holidays. However, there is no substantial difference
- Booking frequencies appear to be nearly equal on working days and non-working days.

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

When creating dummy variables from categorical variables, it's common to use the drop_first=True parameter to avoid the multiple variable for which can be explained by other variables as well. It is important to drop_first= True because

- Multicollinearity:

Incorporating all dummy variables without excluding one can result in multicollinearity, a scenario in which two or more variables in a multiple regression model exhibit high correlation.

- Redundancy:

When dealing with dummy variables that represent a categorical variable with two levels (0 and 1), incorporating both is redundant. Knowing the value of one dummy variable inherently provides information about the other.

- Interpretability:

Dropping the first dummy variable makes the interpretation of coefficients more straightforward.

- Computational Efficiency:

Including unnecessary dummy variables increases the computational burden. Dropping one dummy variable can make computations more efficient and save memory.

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

Temp variables (Temp/aTemp) have highest Corelation approx. 0.62

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

- Residual Analysis:

Done the plot between residuals and found that those were normally distributed

- Linearity:

Check the linearity assumption by plotting the observed values against the predicted values. The relationship is linear.

- Homoscedasticity (Constant Variance of Residuals):

Plot the residuals against the predicted values. The spread of residuals should be roughly constant across all levels

- Multicollinearity:

Checked for multicollinearity among independent variables by examining the variance inflation factor (VIF). High VIF values may indicate multicollinearity issues. And all the variables are in range

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

The importance features after testing the params are Temp, year, Winter and Sep (Winter/Sep are closely important and Winter is minorly important)

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Linear regression is a statistical technique employed to establish the connection between a dependent variable and one or more independent variables through the application of a linear equation to the provided data. In the context of simple linear regression, the scenario involves just one independent variable, whereas multiple linear regression incorporates more than one independent variable.

Simple Linear Regression:
In simple linear regression, the relationship between the independent variable X and the dependent variable Y is defined as a straight line:
$Y=\beta0+\beta1X+\varepsilon$

- Y - dependent variable.
- X - independent variable.
- $0\beta0$ - y-intercept
- $\beta1$ - slope of the line
- ε is the error term. Expressing the dispersion in Y that remains unaccounted for by the model. The objective is to determine the values of $0\beta0$ and $1\beta1$ that minimize the sum of squared differences between the observed and predicted values of Y. Typically, this is achieved through the application of the least squares method.

- **Multiple Linear Regression:**
  When there are multiple independent variables, the model becomes:
  $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta nXn+\varepsilon$
- Y - dependent variable.
- X 1,X2,---, Xn – Independent variables
- $\beta0$ - y-intercept
- $\beta1$, $\beta2$, $\beta3$, ---βn- slope of the line associated with the variables
- ε is the error term.

Assumptions of Linear Regression:

Linearity: The relationship between the independent and dependent variables is assumed to be linear.

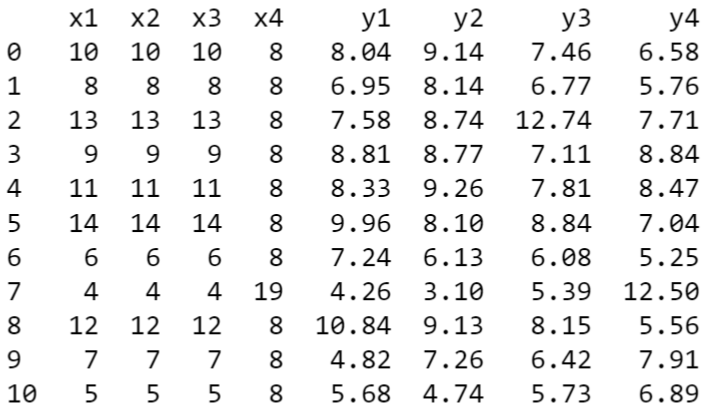
Independence: The residuals (the differences between observed and predicted values) should be independent of each other.

Homoscedasticity: Residuals should have constant variance across all levels of the independent variables.

Normality of Residuals: The residuals should be approximately normally distributed.

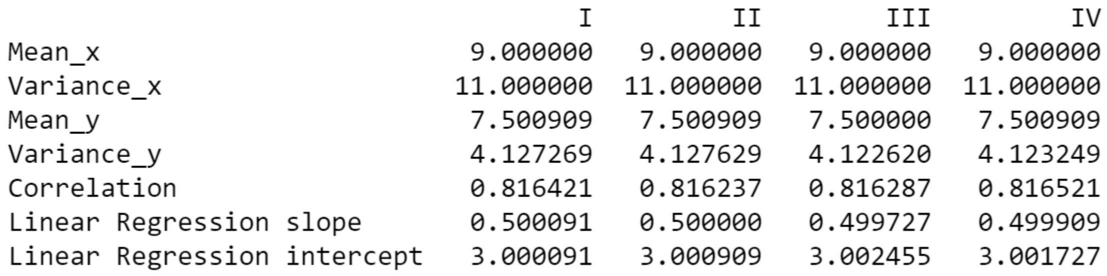
2. Explain the Anscombe’s quartet in detail. (3 marks)

Sample data set that is used:



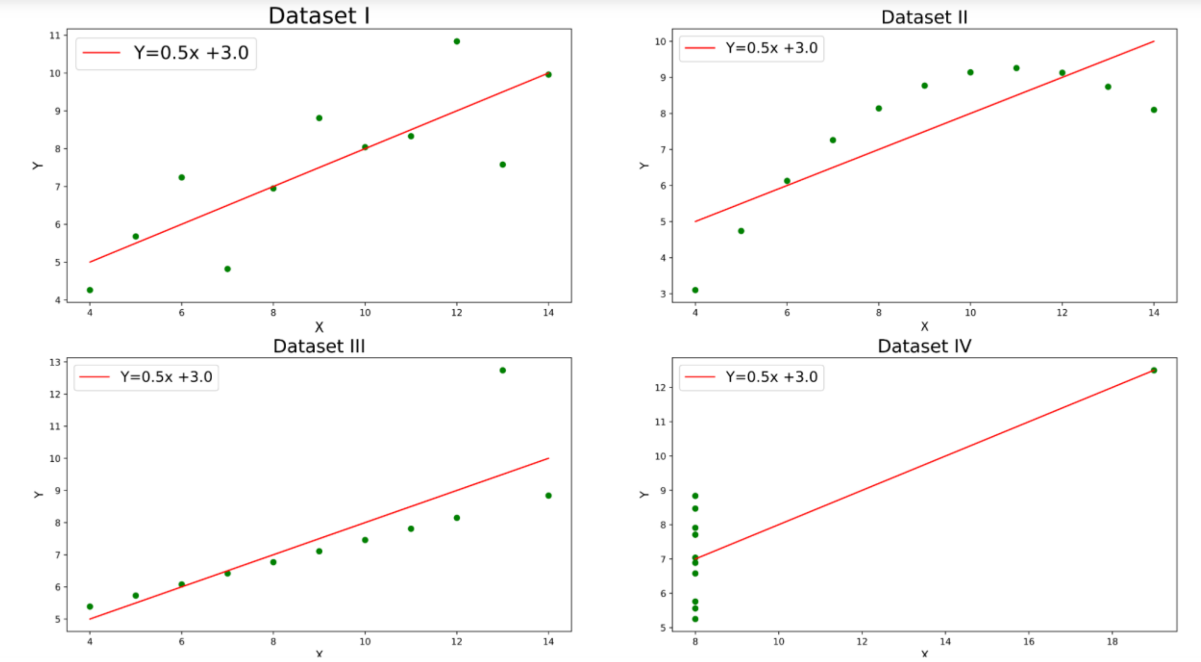
```
    x1  x2  x3  x4     y1    y2     y3     y4
0   10  10  10   8   8.04  9.14   7.46   6.58
1    8   8   8   8   6.95  8.14   6.77   5.76
2   13  13  13   8   7.58  8.74  12.74   7.71
3    9   9   9   8   8.81  8.77   7.11   8.84
4   11  11  11   8   8.33  9.26   7.81   8.47
5   14  14  14   8   9.96  8.10   8.84   7.04
6    6   6   6   8   7.24  6.13   6.08   5.25
7    4   4   4  19   4.26  3.10   5.39  12.50
8   12  12  12   8  10.84  9.13   8.15   5.56
9    7   7   7   8   4.82  7.26   6.42   7.91
10   5   5   5   8   5.68  4.74   5.73   6.89
```

Metrices are:



```
                                    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
```

The Graphs when plotted looks like :



Explanation of this output:

- In the first one Dataset One the scatter plot seems to be a linear relationship between x and y.

- In the second one Dataset Two is a non-linear relationship between x and y.
- In the third one Dataset Three there is a perfect linear
- Finally, the fourth one Dataset Four shows that when one high-leverage point is enough to produce a high correlation coefficient.

The quartet continues to be frequently employed to emphasize the significance of visually inspecting a dataset before engaging in a specific type of analysis. It underscores the limitations of relying solely on fundamental statistical properties to describe complex datasets realistically.

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

Pearson's correlation coefficient, commonly represented as *r* or Pearson's *r*, serves as a statistical metric that measures the strength and direction of a linear association between two continuous variables. Its values fall within the range of -1 to +1, with specific implications for the nature of the relationship.

- r=1 indicates a perfect positive linear relationship,
- r=−1 indicates a perfect negative linear relationship,
- r=0 indicates no linear relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling? (3 marks)

Scaling is the method used to adjust numerical variables within a dataset to a standardized range or scale. The main purpose of scaling is to ensure that all variables share a common scale, facilitating equitable comparisons and preventing variables with larger magnitudes from exerting undue influence on the analysis
Normalized Scaling: All values are transformed to a scale between 0 and 1.

Standardized Scaling (Z-Score Normalization): All values are centered around 0, and the spread is adjusted to have a standard deviation of 1.

Scaling plays a vital role as an essential preprocessing step in both data analysis and machine learning. Two prevalent techniques, normalized scaling and standardized scaling, serve distinct purposes. Normalized scaling focuses on restricting variables within a predefined range, whereas standardized scaling centers variables around 0, ensuring a standard deviation of 1. The selection between these techniques hinges on the specific needs of the analysis and the inherent characteristics of the dataset.

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

(3 marks)

The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a regression analysis. The occurrence of an infinite VIF typically arises when perfect multicollinearity is detected among predictor variables. Perfect multicollinearity manifests when one predictor variable within a regression model can be precisely predicted using a linear combination of the other predictor variables. Essentially, this indicates the existence of a flawless linear relationship among the predictors.

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

(3 marks)

A Quantile-Quantile (Q-Q) plot serves as a visual aid in statistics for examining the adherence of a dataset to a specified theoretical distribution. Specifically in linear regression, Q-Q plots are frequently utilized to appraise the normality of residuals, which represent the variations between observed and predicted values within a regression model. It is used in Normality Assessment:, Identify outliers and skewness and is used in decision making