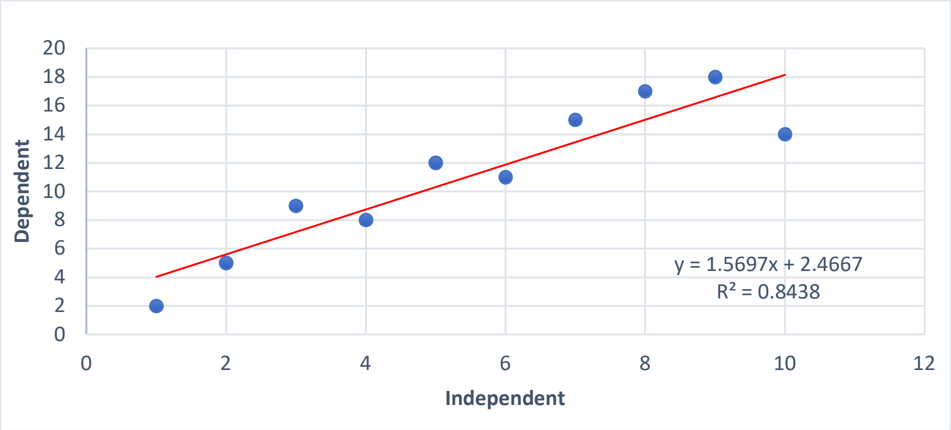
Assignment-based Subjective Questions:

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   - There is a difference in the number of bikes shared during different season. for season 1(Spring) the number is very less than that of other seasons. We can infer that fewer bikes will be shared during spring season.
   - Number of bikes shared increased in the year 2019. So, we can infer that the business has a scope of developing after CoVID.
   - Most number of bikes are shared in the Clear weather. Then somewhat less in Misty and fewest in Rainy/Snow. There is no data for heavy rain/snow. We can infer that a greater number of bikes will be shared when the weather is clear.
   - Slightly higher number of bikes are shared on Monday. We can infer that if it is Monday, chances are that a bit greater number of bikes will be shared.
2. **Why is it important to use drop_first=True during dummy variable creation?**
   - It is important to use drop_first=True during dummy variable creation for the sake of model performance.
   - We do not want to deal with a very large number of data if we can represent the same data with fewer variables.
   - This will make our model to consume less time and resources.
   - For e.g., if we want to represent the values of 3 columns: A, B, or C in binary, we can represent it as: 10 -> A, 01 -> B and 00 -> C.
   - This way, we don’t need a separate column for C, and the model has to deal with 1 less variable.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   - **temp** or **atemp** can have highest correlation, but according to correlation table, atemp has the highest correlation with target variable.
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   - **linearity and additivity** of the relationship between dependent and independent variables:
     - Plotted a scatter plot between predicted and original values of the test data and it was a distributed along diagonal line.
   - **statistical independence** of the errors:
     - Plotted a scatter plot - residuals vs. index. No pattern found in the graph.
   - **homoscedasticity** (constant variance) of the errors:
     - Verified the constant variance in the previous - residuals vs. index graph.
   - **normality** of the error distribution:
     - Plotted a histogram to verify the normality of error terms.
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - temp, year, weather (Light Rain/Snow), windspeed

General Subjective Questions:

1. **Explain the linear regression algorithm in detail.**

- Linear Regression is quite simple and statistical method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.



- The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.
- To calculate best-fit line linear regression uses a traditional slope-intercept form: **$y = a_0 + a_1x$.** Here,
  - y = Dependent variable
  - x = Independent variable
  - a_0 = Intercept of the line
  - a_1 = Slope of the line / Linear Regression coefficient
- A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.
  - If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.
  - If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.

- The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.
- **Cost Function:**
  - The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.
  - The best fit line is obtained by minimizing the cost function.
  - In Linear regression, the cost function used is Mean Squared Error (MSE), which is the average of squared error that occurred between the predicted value and the actual value.

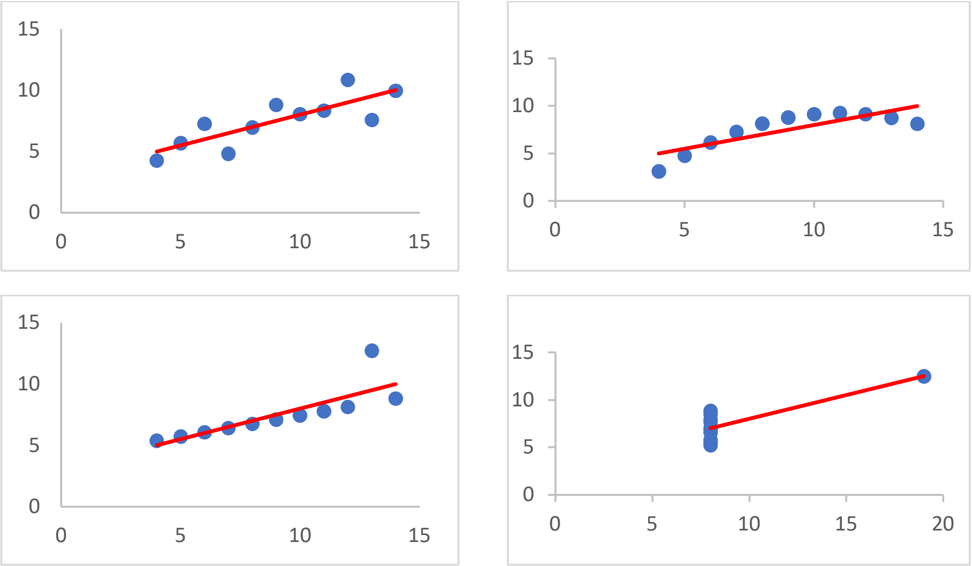
2. Explain the Anscombe's quartet in detail.

- Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analysing it with statistical properties.
- It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyse about these data-sets is that they all share the same descriptive statistics (mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis.

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

- Here, after applying statistical formula on above datasets,
  - Avg. value of x = 9
  - Avg. value of y = 7.50
  - Variance of x = 11
  - Variance of y = 4.12
  - Correlation coefficient = 0.816
  - Linear regression equation: y = 0.5x + 3
- However, the statistical analysis of these four data-sets is pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we

get the following results & each pictorial view represent the different behaviour.



- Data-set I — consists of a set of (x, y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship.
- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.

3. What is Pearson's R?

- The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
- The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Pearson correlation coefficient (r)	Correlation type	Interpretation
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.
0	No correlation	There is no relationship between the variables.

Between 0 and -1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
    - Scaling is a data pre-processing technique that involves transforming the values of features or variables in a dataset to a similar scale. This is done to ensure that all features contribute equally to the model and to prevent features with larger values from dominating the model. Feature scaling is essential when working with datasets where the features have different ranges, units of measurement, or orders of magnitude. Common feature scaling techniques include standardization, normalization. By applying feature scaling, the data can be transformed to a more consistent scale, making it easier to build accurate and effective machine learning models.
    - **Normalization:**
        - Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
        - The formula for normalization: $X' = (X - X_{min}) / (X_{max} - X_{min})$
        - Here, X_{max} and X_{min} are the maximum and the minimum values of the feature, respectively.
            - When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0
            - On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator, and thus the value of X’ is 1
            - If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1
        - Normalization is sensitive to outliers.
    - **Standardization:**
        - Standardization is another scaling method where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.
        - The formula for standardization: $X' = (X - \mu) / \sigma$
        - Here, μ = mean of the feature values and σ is the standard deviation of the feature values.
        - Standardization is sensitive to outliers.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
    - If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
    - An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
    - Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.
    - Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s)
    - Whenever we are interpreting a Q-Q plot, we shall concentrate on the 'y = x' line. We also call it the 45-degree line in statistics. It entails that each of our distributions has the same quantiles. In case if we witness a deviation from this line, one of the distributions could be skewed when compared to the other.