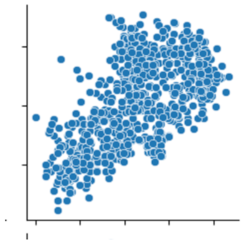
Bike Sharing Assignment

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?*
   Ans: After assessing the categorical variables we can assess that some of them are very highly correlated to the dependent variable. Below is the finding.
   Year: There is a significant increase in the number of Users in 2019 from 2018, Showing an upward trend.
   Month: Summers see increased usage as compared to winter months.
   Holiday: Usage is more on Working Days as compared to holidays, which shows that a significant number of people use the service for commuting to work
   Working Day: Although there is a slight increase in the usage on Working day the difference is not that significant.
   Weekday: We see that throughout the week the bike usage is almost similar with slightly higher usage on Monday. Overall the difference between the days is very low
   Weather Situation: As expected usage is high on clear days and lowest on snowy days. We don't have even a single record of bike rental on Rainy days.

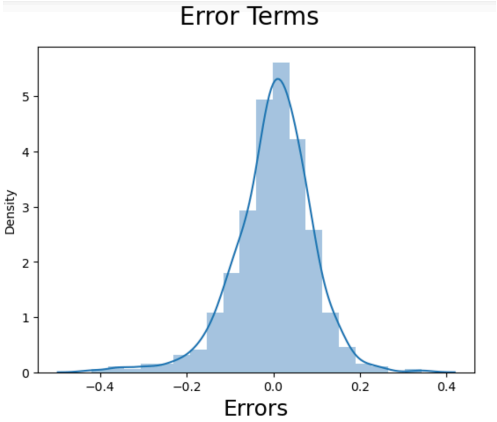
2. *Why is it important to use drop_first=True during dummy variable creation?*
   Ans: When we are dealing with categorical variables with values 'n' values, the idea of dummy variable creation is to create 'n-1' variables indicating the levels. Suppose we have a categorical variable 'Grade' with values A, B and C we can use 2 dummy variables to represent the values instead of 3. Such that if A is True it will take the value 1 and if B is true it will take the value 1 and if they both have the value 0 then it is clear that C is true and so we don't need third variable and that is why we drop one variable using 'drop first = True'

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



Looking at the pairplot the ***Temp*** and ***Atemp*** had the highest correlation with the target variable. It is understandable as it will be likely that people are willing to use bike more when the weather is pleasant and warm than when its

4. ***How did you validate the assumptions of Linear Regression after building the model on the training set?***



Ans: In order to validate the assumptions of Linear Regression we did a Residual Analysis of the Error terms to check and see if the error terms have a normal distribution. As per the histogram that we plotted on the error terms we did have a normal distribution centred at zero hance the model was good.

5. ***Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?***
Ans: The equation of our best-fitted line as per the model was as follows

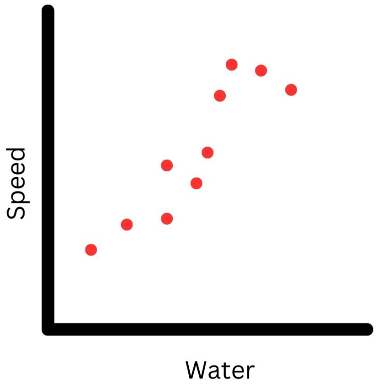
count = 0.2338 X yr + 0.055 X workingday + 0.4925 X temp - 0.150 X windspeed -0.0662 X spring +0.0469 X summer + 0.0851 X winter - 0.0489 X Jul + 0.0738 X Sep + 0.0666 X monday - 0.0830 X misty - 0.2892 X snow

As seen in the equation the 3 features contributing significantly are ‘temp’, ‘yr’ and ‘windspeed’

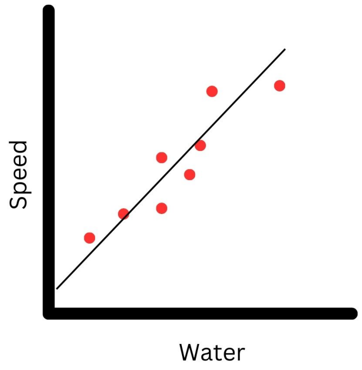
<u>General Subjective Questions</u>

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

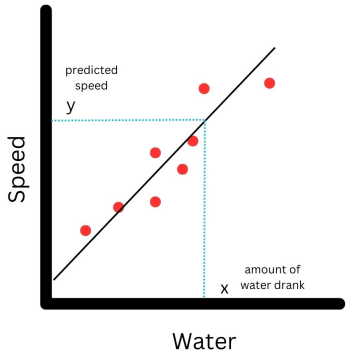
   To understand Linear Regression let us consider a situation where we measured how quickly someone can run 100 meters and how much water they drank. Now let us plot the results of 10 such people on a chart where the X-axis represents the amount of water they drank and the Y-axis represents the speed.



   In the above figure, each red dot is a person and corresponding to the point in the graph we can know how much water they drank and how fast they ran. Given this pretend data we can see that the more water they drank the faster they ran and completed the 100 meters. We can fit a line through the data to show the trend.



   We can also use the black line for predictions. Suppose someone said that they drank x amount of water we can use the black line to predict how fast that person can run.



The black line is the type of Machine Learning that can be used for making predictions. This whole process of fitting the best line after analysing and understanding the data is called Linear Regression.

The same principle of linear regression can be used for making predictions in real-life scenarios like what is the probability of a person repaying his loan. We can take some variables like his/her income, age, and liabilities and use them to predict the probability of him/her repaying the loan.

In the above example we used the amount of water drank to predict the speed. Here, the amount os water drank is called Independent Variable and Speed is called the dependent variable. In Linear Regression we can have more than one Independent Variable to predict one Dependent Variable.

In order to find calculate and figure out the best line that fits our model and which can be used to best predict the outcomes we use the following equation

$y = mx + c$

You might have figured out that this is the equation for a straight line. There are 2 parameters in this equation m and c. m denotes the slope of the line and c denotes the point where the line intercepts the Y-axis. Basically c is the value when x is 0, m signifies how strong is the relationship between y and x, if you increase x by some units how much y will increase will depend on the value of m which is the slope.

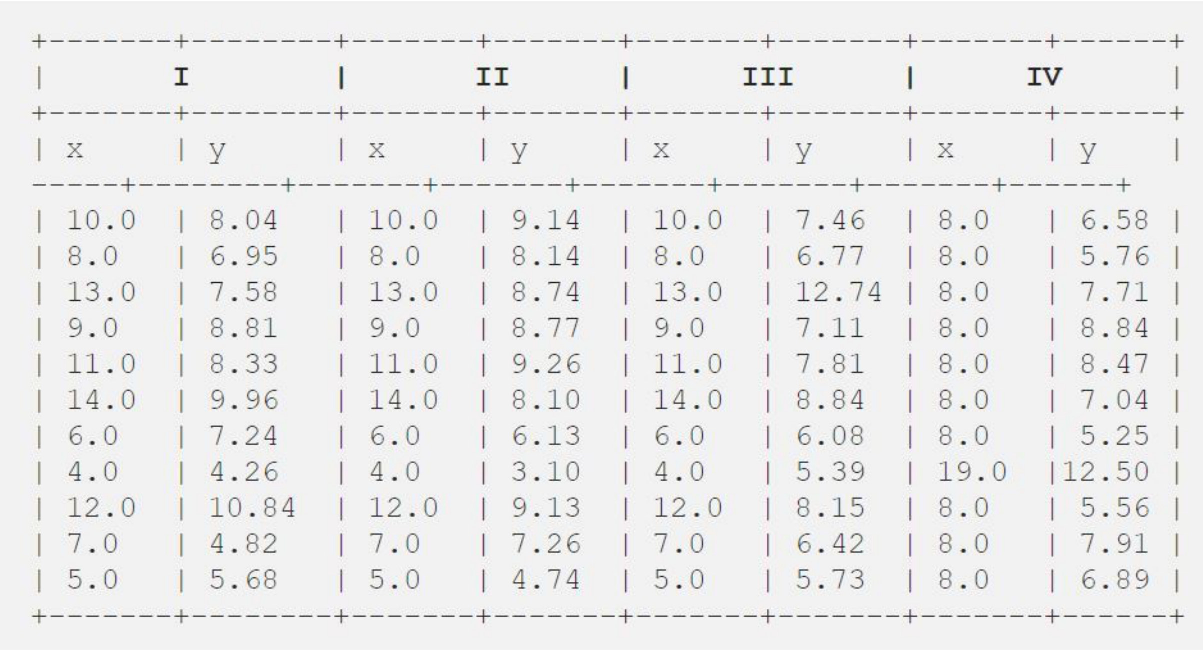
In Linear Regression, we write the equation as below

$y = B0 + B1x$

This is the standard format used for Linear Regression. Where B1 represents the slope and B0 is the intercept. Where B0 tells us what will the value of y when x is 0.

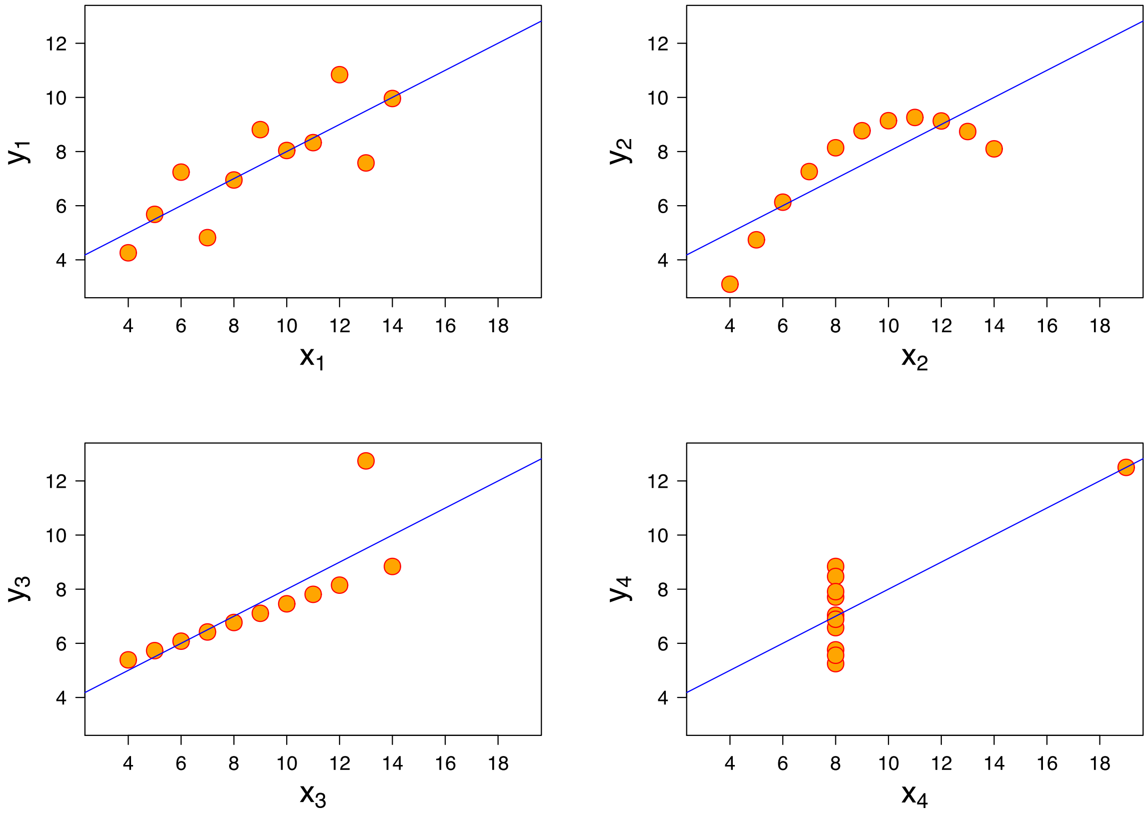
2. *Explain the Anscombe's quartet in detail.*

How could sets of data with more or less the same statistical properties such as average value or mean, variance or how much is the data spread out from each other look totally different when presented in a graph. In 1973 an English statistician Francis Anscombe show exactly how by giving a striking example. He presented a paper with 4 graphs which later came to be known as Anscombe's Quartet. The graphs contained eleven data points each with identical statistical properties of Mean, Standard Deviation and Correlation.



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

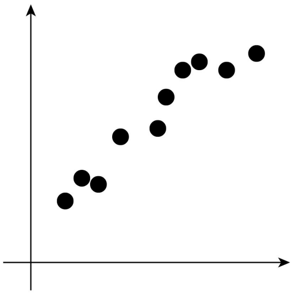
Looking at the data alone would have concluded that when plotted on a graph with x-axis and y-axis the graphs would look identical. It was common belief at the time that once the numbers have been calculated graphing was irrelevant. Anscombe showed how wrong that common belief was. The four datasets which were very identical in terms of statistical properties when plotted on a graph looked completely different.



It clearly showed that statistics can easily fool us if used incorrectly or if we fail to take in the full picture of what's going on.

3. What is Pearson's R?

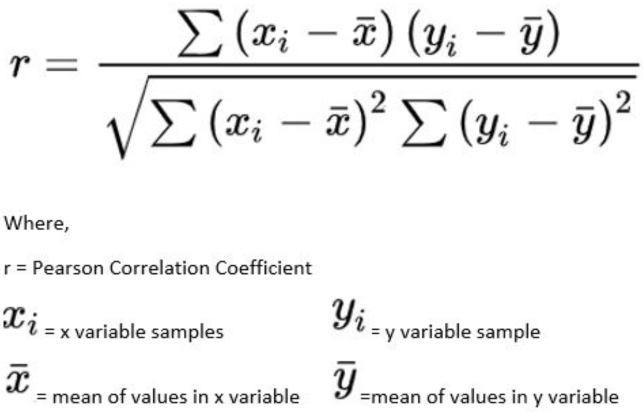
Pearson's R or Pearson's correlation analyses the relationship between two variables. For example, is there a relationship between a person's salary and age



In this scatter plot every single point is a person. If the relationship is confirmed between the salary and age we can predict the salary of a person by age using a regression model. But it is not as easy as it sounds. There must be a clear causal relationship for this. Just because there is a correlation we can tell which way the relationship is going. But with the help of Pearson's R or Pearson's Correlation we can measure the linear relationship between two variables.

Pearson's R helps us identify how strong the correlation is and if it is a positive or negative correlation. The value of Pearson's R is between -1 and 1. Where -1 represents a perfect negative correlation and 1 represents a perfect positive correlation and 0 shows that there no relationship between the two variables.

To calculate the Pearson's R we use the following formula



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

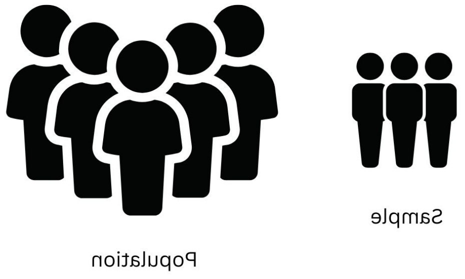
x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

In the equation we can see that the respective mean value is first substracted from both variables. So in our example, we calculate the mean values of age age and salary. We then substract the mean values from each person's age and salary. We then multiply both values. Then we sum up the individual results of the multiplication. The expression in the denominator ensures that the correlation coefficient is scaled between -1 and 1.



The correlation coefficient is usually calculated with data taken from a sample. However we often want to test a hypothesis about the population. In the case of correlation analysis, we then want to know if there is a correlation in the population.For this we check, whether the correlation coefficient in the sample is statistically significantly different from zero.

The null hypothesis in the pearson correlation is that the correlation coefficient does not differ significantly from zero which means there is no linear relationship and the alternative hypothesis is that the correlation coefficient differs from zero which means we can reject the null hypothesis and state that there is a linear relationship between the variables.

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

Scaling is the process of converting different types of numerical variable values so that their weightage can be assessed and understood easily for building a machine learning algorithm or model. The datasets that we use for training models in machine learning can sometimes have values that might vary from each other on a broad scale. Numerical values might have big differences amongst themselves, especially when they represent different scales, and this would make it difficult to compare them, for example : kg, miles, centimeters, feet, pixels, etc.
Scaling is introduced to solve this problem. It adjusts the numbers to make it easy to compare the values. This helps increase the accuracy of the models.

There are two techniques of scaling:

1. Normalized Scaling:
   Normalization is a data preprocessing technique used to adjust the values of features in a dataset to a common scale. This is done to facilitate data analysis and modeling, and to reduce the impact of different scales on the accuracy of machine learning models.
   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.

1. Standardized Scaling :
   Standardization is another scaling method where the values are centered 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.

It might be difficult to choose between the normalized scaling or standardized scaling. However, it will depend on the problem and the machine learning algorithm we are using. There is no hard and fast rule to tell us when to which technique.

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

   VIF or is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

   A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients.

   If all the independent variables are orthogonal to each other, then VIF = 1.0. **If there is perfect correlation, then VIF = infinity.**

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

   The purpose of the Q-Q plot or the quantile-quantile plot is to show if two data sets come from the same distribution. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed. In practice, many data sets are compared to the normal distribution. The normal distribution is the base distribution and its quantiles are plotted along the x-axis as the "Theoretical Quantiles" while the sample quantiles are plotted along the y-axis as the "Sample Quantiles".