1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    a. Company should focus on expanding business during Spring
    b. Company should focus on expanding business during September.
    c. Based on previous data it is expected to have a boom in number of users once situation comes back to normal, compared to 2019.
    d. There would be less bookings during Light Snow or Rain,
    e. Significant variables:-
        i. holiday
        ii. temp
        iii. hum
        iv. windspeed
        v. Season
        vi. months(January, July, September, November, December)
        vii. Year (2019)
        viii. weathersit

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

    drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
    Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level. Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. Error terms should be normally distributed
    b. There should be insignificant multicollinearity among variables.
    c. Linearity should be visible among variables
    d. Homoscedasticity
    e. No auto-correlation
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. year
    b. temp

c. sep

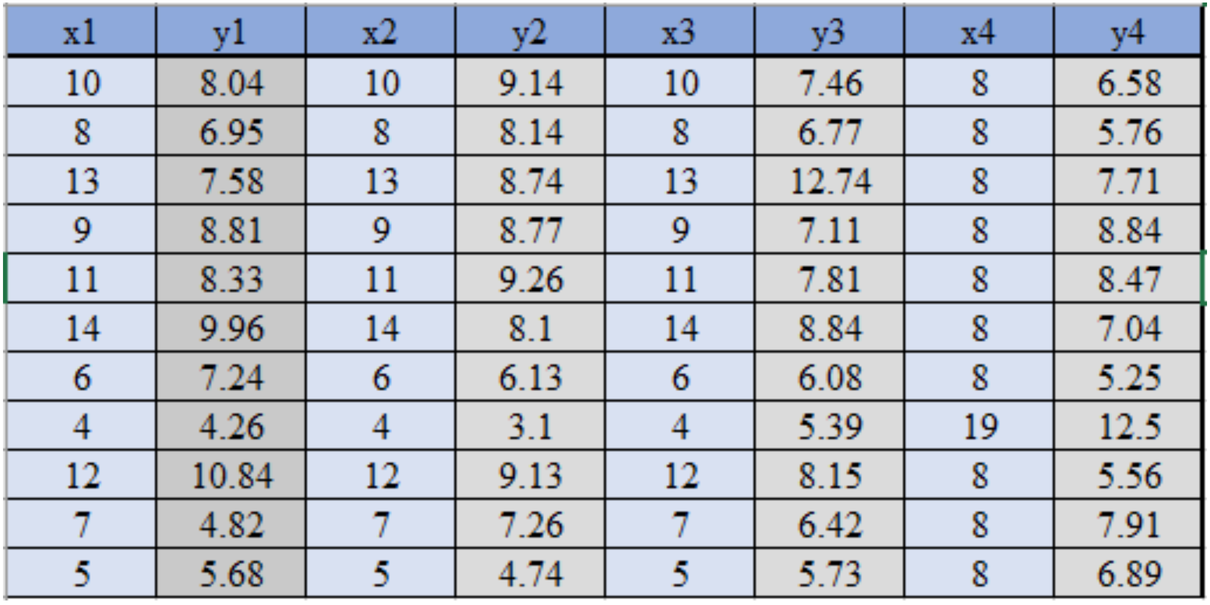
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

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 analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze 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 behavior irrespective of statistical analysis.



x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Four Data-sets

Apply the statistical formula on the above data-set,

Average Value of x = 9

Average Value of y = 7.50

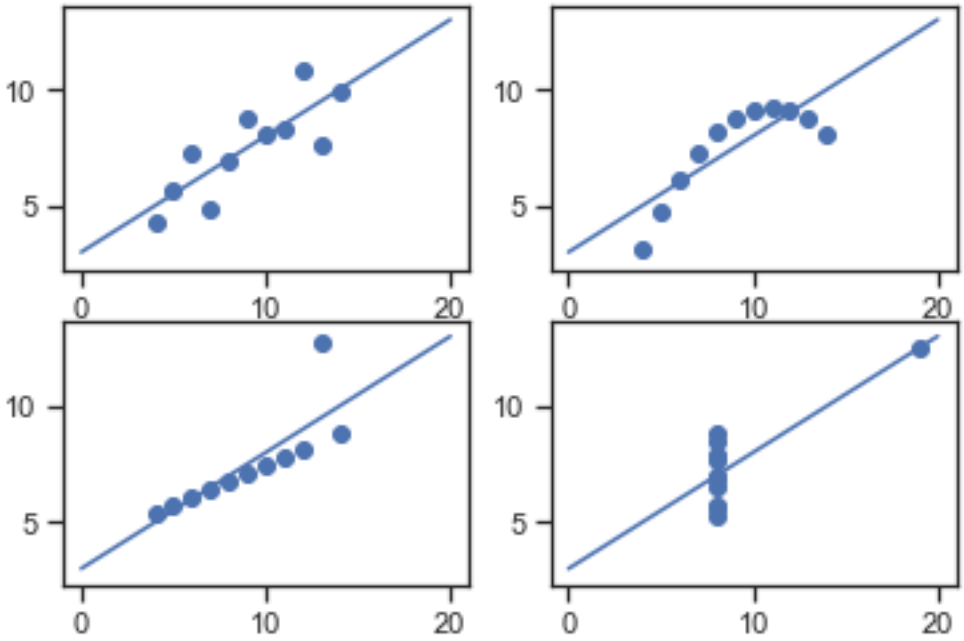
Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are 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.

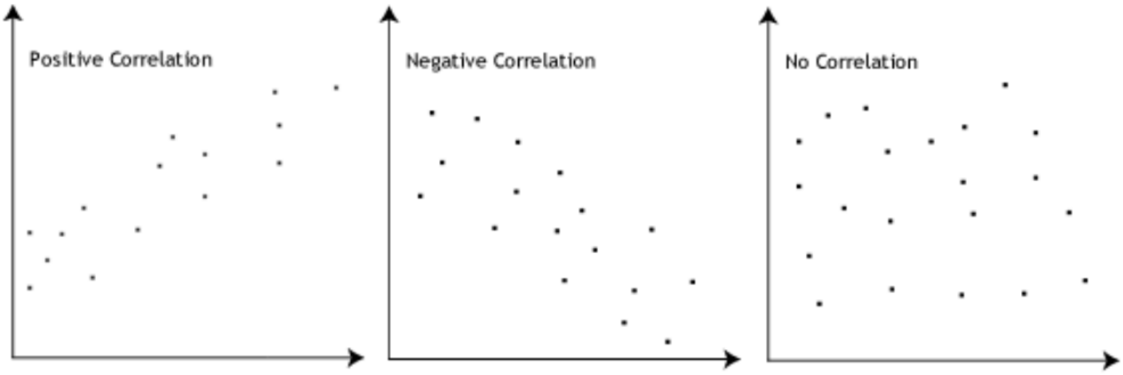


Graphical Representation of Anscombe’s Quartet

- 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 (might be quadratic?).
- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
- Data-set IV — looks like the value of x remains constant, except for one outlier as well.

3. What is Pearson’s R?
    a. Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive.
    b. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
    c. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables.

d. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   a. Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. I
   b. t is performed during the data pre-processing to handle highly varying magnitudes or values or units.
   c. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
   d. We use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

S.NO.	Normalization	Standardization
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   a. If there is perfect correlation, then VIF = infinity.
   b. A large value of VIF indicates that there is a correlation between the variables.
   c. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity.
   d. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
6. What is a Q-Q plot?
   a. The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
   b. Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.
   c. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.
   d. Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.