## 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?**

**Answer**: From the analysis done by me of the categorical columns the infer we could derive are the followings-

a) Except winter season, spring, summer & fall are in increasing manner. And the Highest number of bookings in Fall season, so we can expect High bookings on Fall season.
b) As we can see that year 2019 has more bookings compared to 2018, it means we can expect more bookings on next years.
c) There is increasing manner from the month of Jan to Sep, the maximum numbers of bookings in the month of May, Jun, Jul, Aug & Sep.
d) From the Holyday feature we could strongly say that the bookings count is more when it's non-holyday.
e) Count of booking is more on Thu, Fri, Sat & Sun as compared to Mon, Tue & Wed.
f) We can see, there is not much difference in Working day.
g) From the weathersit we can strongly say clear weather has more numbers of bookings followed by mist.

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

**Answer**: **drop_first = True** is important because it helps to drop the 1st column while creating dummy variables. Like, if we have a column for gender that contains 4 variables- Male, Female, Other, Unknown. So, a person can be Male, Female or Other, if not in these three then it will consider as unknown.
The main purpose is to reduce the number of columns or we can say extra column which is not necessary.

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

**Answer**: From the Pair-Plot among the numerical variables, temp & atemp has the highest correlation with the target variable cnt.

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

**Answer**: I have validated the assumptions of Linear Regression after building the model on the training set are the followings-

1) Residual Analysis – Error terms should be normally distributed
2) Linearity - There should be a linear relationship between dependent variable and independent variable.
3) There shouldn't be any correlation among the residual (error) terms.
4) Multicollinearity - The independent variables should not be correlated.
5) Homoscedasticity-The error terms must have constant variance.

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

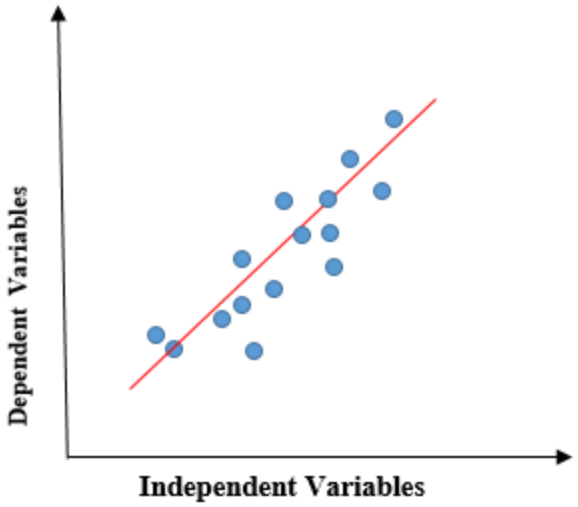
**Answer**: Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are the followings-

i. Temp
ii. Year
iii. Light Rain

## General Subjective Questions

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

**Answer**: Linear regression is one of the Machine Learning Algorithm where we train the model to predict the behavior of the data based on some variables. Linear regression shows the relationship between Dependent Variable & Independent Variable(s). In Linear Regression If there is one input variable it's known as Simple Linear Regression and if there is more than one input variables it's known as Multiple Linear Regression.




The above graph we can understand 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 the best fit straight line. Based on the given data points.

**Equation** -> Y = MX + C

Y = Dependent Variable

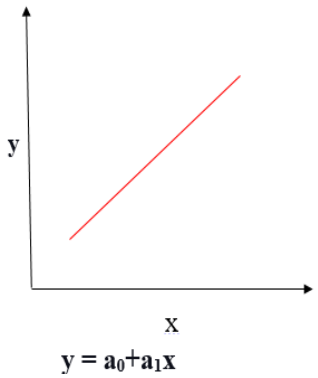
M = Slope of the Line

X = Independent Variable

C = Intercept of the line

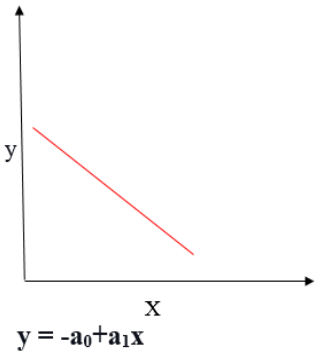
Linear Relationship can be two type-

1. **Positive Linear Relationship** – When the values of dependent variable on Y axis increases and the independent variable on X axis increases, it's known as Positive Linear Relationship.




$y = a_0 + a_1 x$

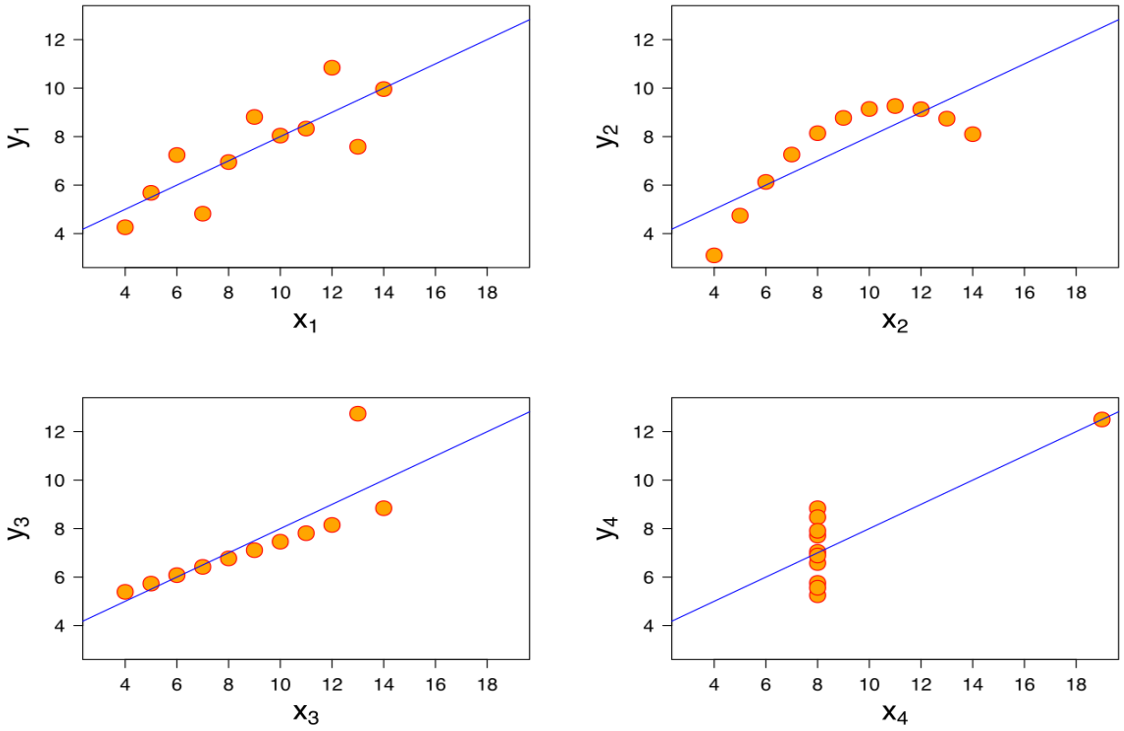
2. **Negative Linear Relationship** – When the values of dependent variable on Y axis decreases and the independent variable on X axis increases, it's known as Negative Linear Relationship.




3.

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

**Answer**: Anscombe's Quartet has designed by the statistician Francis Anscombe consists of four datasets with eleven (x, y) data-points. He found this dataset in his dream & requested the council that his last wish to plot those data-points.
When the council analyzed the Data-point using descriptive statistics and found the mean, standard deviation and correlation between x, y pretty same, yet appear very different when graphed.




## Explanation:

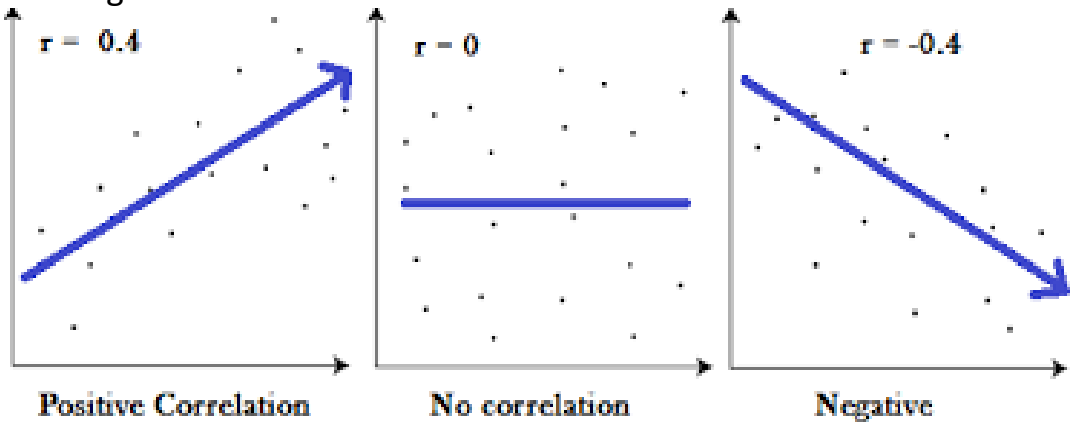
- If we look at the first one (Top Left) it clearly represents a linear relationship between x & y.
- If we look at the second one (Top Right) indicates a non-linear relationship between x & y.

- If we look at the third one (Bottom Left) indicates the linear relationship between x & y except one which could be an outlier, far away from the line.
- If we look at the fourth one (Bottom Right) provides an example of high-leverage point is enough to produce a high correlation coefficient.

This Quartet conclude the importance of looking at the dataset graphically before starting to analyze according to a particular type of relationship.

**3. What is Pearson's R?**

**Answer**: In statistics, Pearson correlation coefficient also referred to Pearson's R. It is nothing but a normalized measurement of two variable's Covariance. The value of the correlation coefficient varies between +1 and -1. A value of ± 1 indicates a perfect degree of association between the two variables.
A value of 0 indicates that there is no relationship between the two variables. A value greater than 0 indicates a positive relationship; that is, as the value of one variable increases, the value of the other variable increases. 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. We can check from the below diagram.




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

**Answer**: Scaling is a technique to standardize the independent featured present in the data to a fixed range. It is used in during the data pre-processing to control highly varying values. When data has different values, and even different measurement units, it can be difficult to compare them.
Suppose, if we don't perform scaling method than it can consider the value 3000 miligrams to be greater than 5 kilograms, we know that's not true actually. In this case the algorithm will output wrong prediction.
Techniques of Scaling –

1. **Min-Max Scaling** – This technique rescales the values between 0 & 1

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

2. **Standardization** – This technique rescales the values so that it has distribution with 0 mean value & variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

| S.NO | Normalized Scaling | Standardized Scaling |
|---|---|---|
| 1. | Minimum and Maximum value 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 0 mean and unit standard deviation |
| 3. | It scales values between (0,1) or (-1,1) | It is not fixed 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 |

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

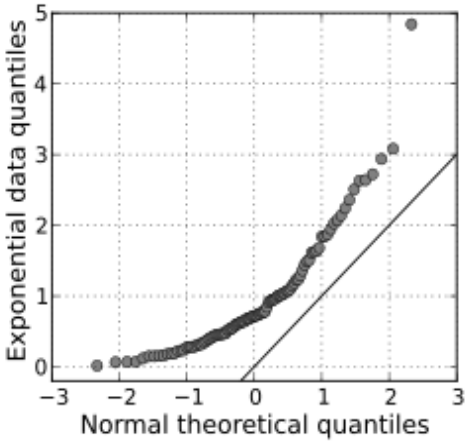
**Answer**: VIF = infinite means there is perfect correlation between two independent variables. In the case of perfect correlation, the value of R2 = 1, which mean 1/(1-R2) infinite. We need to drop that variable which causing the multicollinearity to solve this problem.
An infinite VIF value indicates that the variable concerned can be expressed exactly by a linear combination of the other variables, which also show an infinite VIF.
The value of VIF calculated by the below formula

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where 'I' means the ith variable.
If R2 value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite, it denotes perfect correlation between variable.

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

**Answer**: Q-Q plot (Quantiles-Quantiles plots) are plots of two Quantiles against each other. Q-Q plot is a graphical technique for determining if two datasets come from the population with same distributions.
A Quantiles is a fraction where certain values fall below that quantile.
Like, 25% Quantile is the point where 25% of the data fall below and 75% fall above that value.
A 45 Degree angle is plotted on the Q-Q plot, if the two datasets come from the same distribution, the points will fall on that reference line.
A Q-Q plot showing the 45 Degree reference line,




If the two Distribution being compared are similar, the points on the Q-Q plot will approximately lie on the line y = x, if the distribution are linearly related, the points on the Q-Q plot will approximately lie on a line, but not necessarily on the line y = x.
Use of Q-Q plot-

1. No need of equal size for the Sample
2. It is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale and skewness are similar or different in the two distributions.
3. The Q-Q plot can provide more insight into the nature of the difference than analytical methods.
4. Many distributional aspects can be simultaneously tested.