## 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:

- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of May, June, July, Aug, sep and Oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year. Number of booking for each month seems to have increased from 2018 to 2019.
- Clear weather attracted more booking which seems obvious. And in comparison to previous year, i.e. 2018, booking increased for each weather situation in 2019.
- Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
- Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
- Booking seemed to be almost equal either on working day or non-working day. But, the count increased from 2018 to 2019.
- 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

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

Answer: It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Answer: ‘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?**

Answer:

- Error terms should be normally distributed.
- There should be insignificant multi-collinearity among variables.
- Linearity should be visible among variables.
- There should be no visible pattern in residual values.
- 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?**

Answer: In the analysis I found that, Demand of bikes depend on year, holiday, temp, wind speed, sep, Light_snowrain, Misty, spring, summer and winter.

Top 3: 1. temp, 2. winter, 3.sep.

## General Subjective Questions:

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

Answer: Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable (target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables.

Mathematically the relationship can be represented with the help of following equation −
y = mx + c
Here, y is the dependent variable we are trying to predict.
x: is the independent variable we are using to make predictions.
m: is the slope of the regression line which represents the effect X has on Y
c: is a constant, known as the Y-intercept.

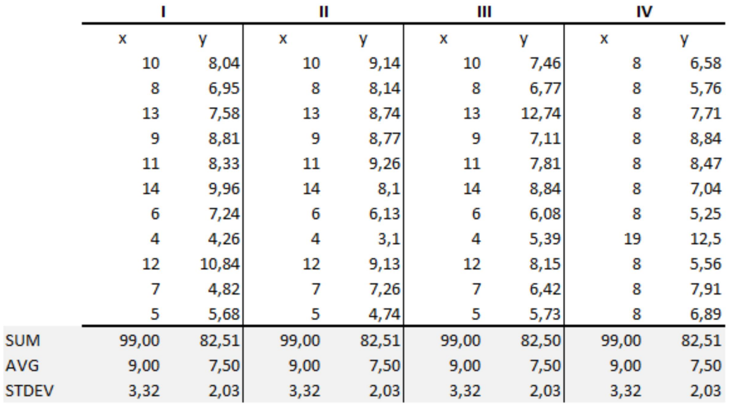
If x = 0, Y would be equal to c.

- Linear relationship can be positive or negative in nature.
- Linear regression is of the following two types −

a). Simple Linear Regression, b). Multiple Linear Regression.

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

Answer: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

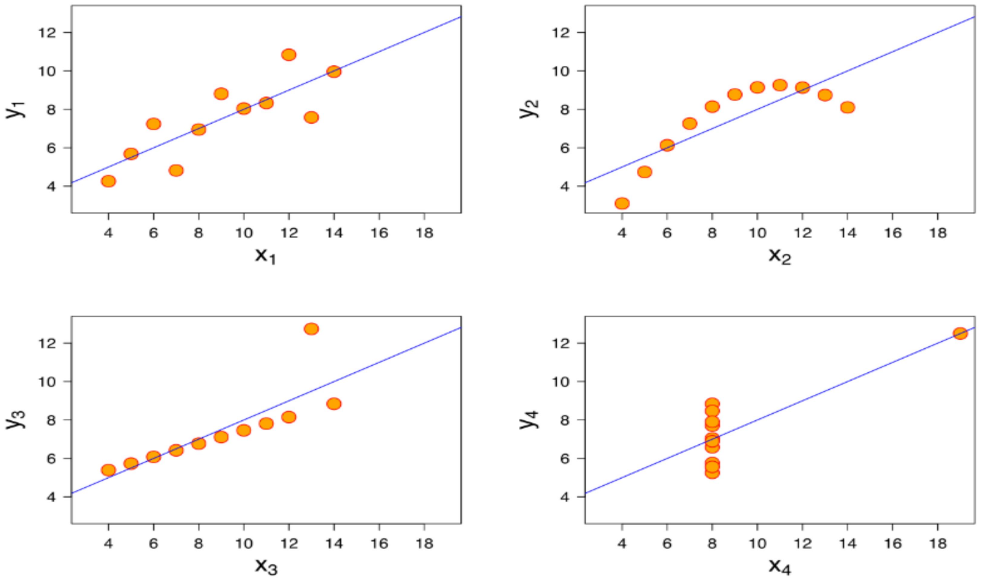


| | I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|---|
| | x | y | x | y | x | y | x | y |
| | 10 | 8,04 | 10 | 9,14 | 10 | 7,46 | 8 | 6,58 |
| | 8 | 6,95 | 8 | 8,14 | 8 | 6,77 | 8 | 5,76 |
| | 13 | 7,58 | 13 | 8,74 | 13 | 12,74 | 8 | 7,71 |
| | 9 | 8,81 | 9 | 8,77 | 9 | 7,11 | 8 | 8,84 |
| | 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 |
| SUM | 99,00 | 82,51 | 99,00 | 82,51 | 99,00 | 82,50 | 99,00 | 82,51 |
| AVG | 9,00 | 7,50 | 9,00 | 7,50 | 9,00 | 7,50 | 9,00 | 7,50 |
| STDEV | 3,32 | 2,03 | 3,32 | 2,03 | 3,32 | 2,03 | 3,32 | 2,03 |

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.
- This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

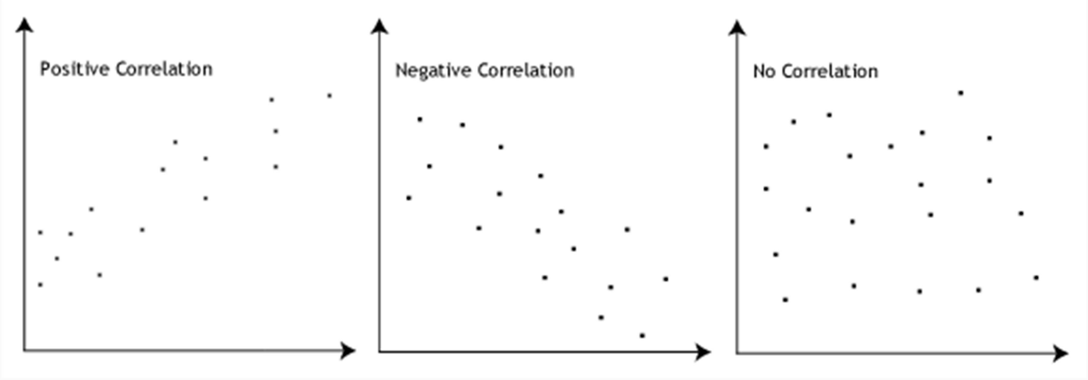
### 3. What is Pearson's R?

Answer: In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope ( i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope ( i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association

- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



**Pearson r Formula**

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

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

Answer: A lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods.

You can scale the features using two very popular method:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

2. Min-Max Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

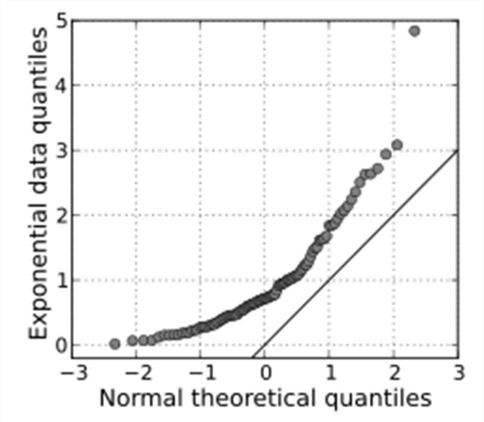
**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

Answer: 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. 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? Explain the use and importance of a Q-Q plot in linear regression.**

Answer: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot 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.