## 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) Analysis of categorical variables from the dataset on the dependent variables
bike rides are increase Year al 2019th has a high median then 2018 is median. the fact that bike is getting popular and the peoples are becoming more awareness about the environment
Spend in the month not reflected of the season plot has fall month have high medium
Working and non-working days has almost the same median who's weak for the non-working day have a plant and does not want to rent bike.
Daily Trend: Registered users demand more bike on weekdays as compared to weekend or holiday.
Rain: The demand of bikes will be lower on a rainy day as compared to a sunny day. Similarly, higher humidity will cause to lower the demand and vice versa.
Time: Total demand should have higher contribution of registered user as compared to casual because registered user base would increase over time.
which is expected as weather condition are most optimal to ride bick followed by summer
People rent more on non holiday compared to the holiday so reason maybe they perfect spend the time with a family and use personal vehicle instead of bike rental.

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

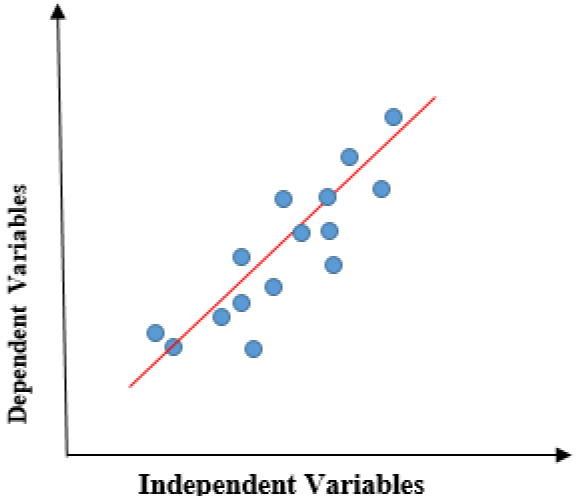
ans) Drop_first=True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have types of values in Categorical column and we want to create dummy variable for that column. If one variable is not season and summer, then It is obvious winter. So we do n ot need 3rd variable to identify the winter

## General Subjective Questions and answer

1. Explain the linear regression algorithm in detail.

Ans) Simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points



The standard equation of the regression line is
$Y = \beta_0 + \beta_1 X$
$\beta_0$=intercept
$\beta_1$=slop
x = Independent variable from dataset
y = Dependent variable from dataset

**The best-fit line**

The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

The strength of the linear regression model can be assessed using 2 metrics:
1. $R^2$ or Coefficient of Determination
2. Residual Standard Error (RSE)

**R2 = 1 – (RSS/TSS)**

Where,
$R^2$ = Coefficient of Determination
RSS = Residuals sum of squares
TSS = Total sum of squares

**TSS (Total sum of squares):** It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

**RSS (Residual Sum of Squares):** In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.
Linear regression assumes the linear relationship between the dependent and independent variables.
Used to find how much change can be expected in a dependent variable with change in an independent variable.

2. Explain the Anscombe's quartet in detail.
Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, Each dataset consists of eleven (x,y) points.
Analyze the 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.
when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.

3. What is Pearson's R?

Ans) 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 normalized 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

Formula

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

r=correlation coefficient
$x_i$=values of the x-variable in a sample
$x^{-}$=mean of the values of the x-variable
$y_i$=values of the y-variable in a sample
$y^{-}$=mean of the values of the y-variable

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

Ans) It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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-squared**, etc.

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

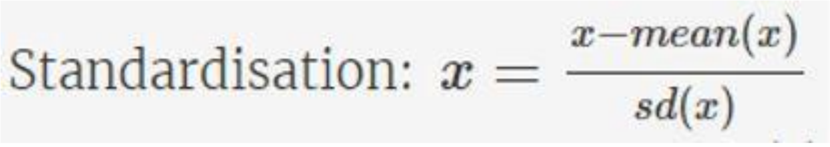
Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1 **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).



Standardisation: $x = \frac{x - mean(x)}{sd(x)}$

- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

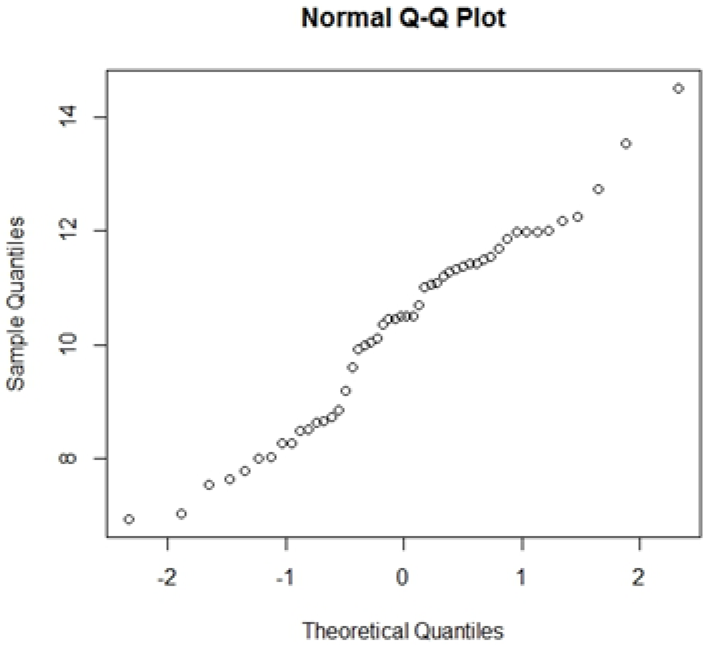
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

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?

Ans) The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.