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?

   - The categorical variables in the dataset are yr., season, month, holiday, weekday, working day and weathers it. The above variables were visualised using a Boxplot and following are my inference about their effect on dependent variable "cnt".

   1.Yr – 2019 showed clear increase in cnt compared to 2018.
   2.Season - season 2 & 3 (Summer & Fall) showed high usage, while season 1 (spring) had the shows lowest footfall.
   3.Mnth – Overall from Mar to Oct the business was reasonable, with highest rental count happening in September.
   4.Holiday– In general Holidays had relatively less rentals
   5.Weekday– There is marginal variation observed in the the mean rental count with the days of week. Monday starts with relatively low count and gradually increases all the way until Friday, with a dip on Saturday.
   6.Working day– When the data is split by years, we don't see significant variation of mean between a working and a nonworking day
   7.Weathersit– The business model shows significant dependency on Weather. While Clear or partly cloudy weather seems preferred, the foot fall decreases with increase in snow / rain and no users when there is heavy rain/ snow.

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

- When we have a categorical variable with, say, 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels, as we will still be able to explain all

the levels with n-1 variables. Any additional variable will be redundant and can lead to multicollinearity within the dummy variables.

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

- Both 'temp' and 'attempt' had highest positive correlation with the target variable

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

- Residuals (Error term) distribution should follow normal distribution, centred around Mean = 0 and should be independent with constant variance. I validated these assumptions by plotting a distplot of residuals and it followed a normal distribution and cantered around 0. Also, I plotted the error term to see if there are any patterns, and the plot showed completed random spread around zero.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- Based the linear regression model I built, following are the top 3 significant features explaining the demand:

1.Temp: 0.437655 (positive correlation)

2.Yr: 0.234287 (positive correlation)

3.Light Snow & Rain: - 0.292892 (negative correlation)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

- Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

Y = mX + b

Here, Y is the dependent variable we are trying to predict

X is the dependent variable we are using to make predictions.

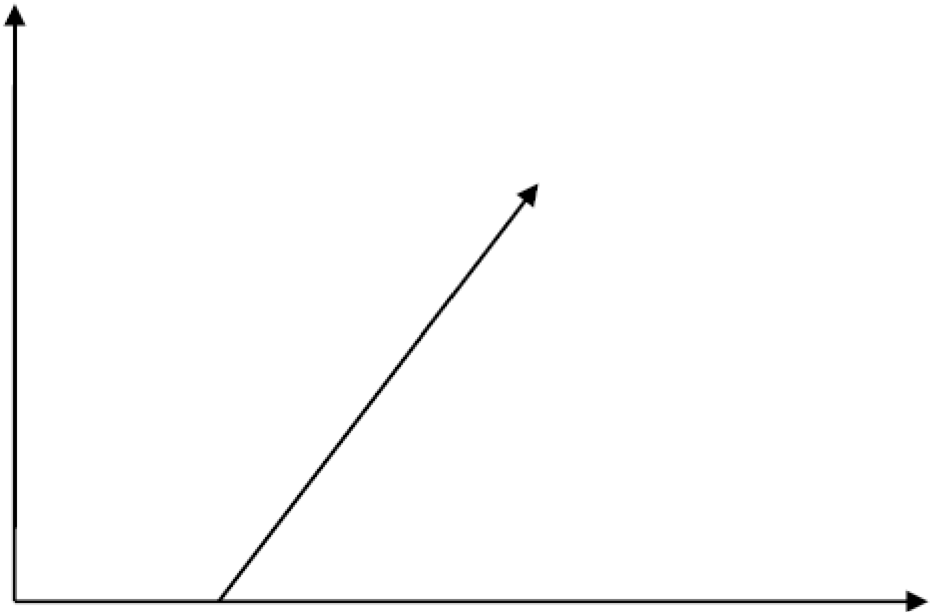
m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept. If X = 0,Y would be equal to b.

Furthermore, the linear relationship can be positive or negative in nature as explained below −

Positive Linear Relationship

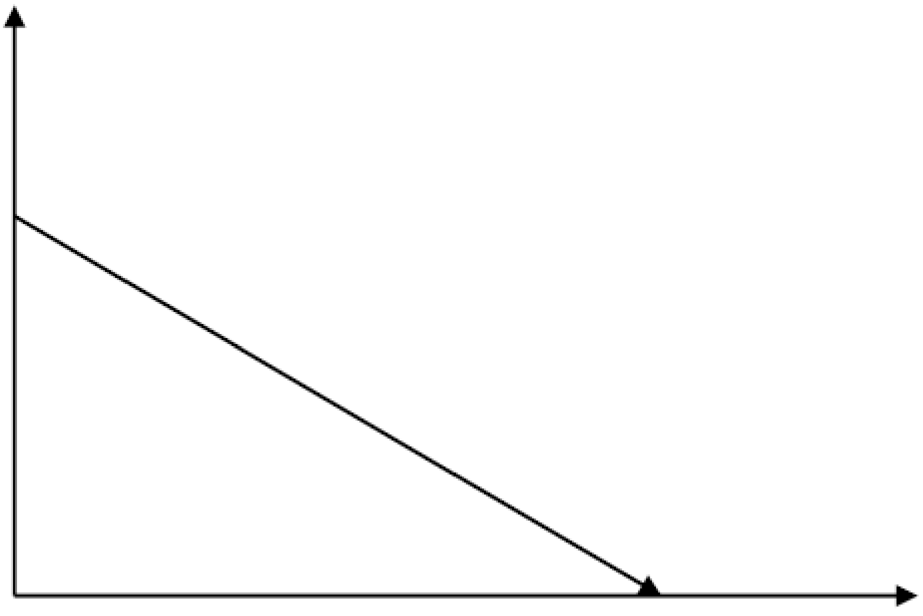
A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



Positive Linear Relationship

Negative Linear relationship

A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph −



Negative Linear Relationship

Types of Linear Regression

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Simple Linear Regression (SLR)

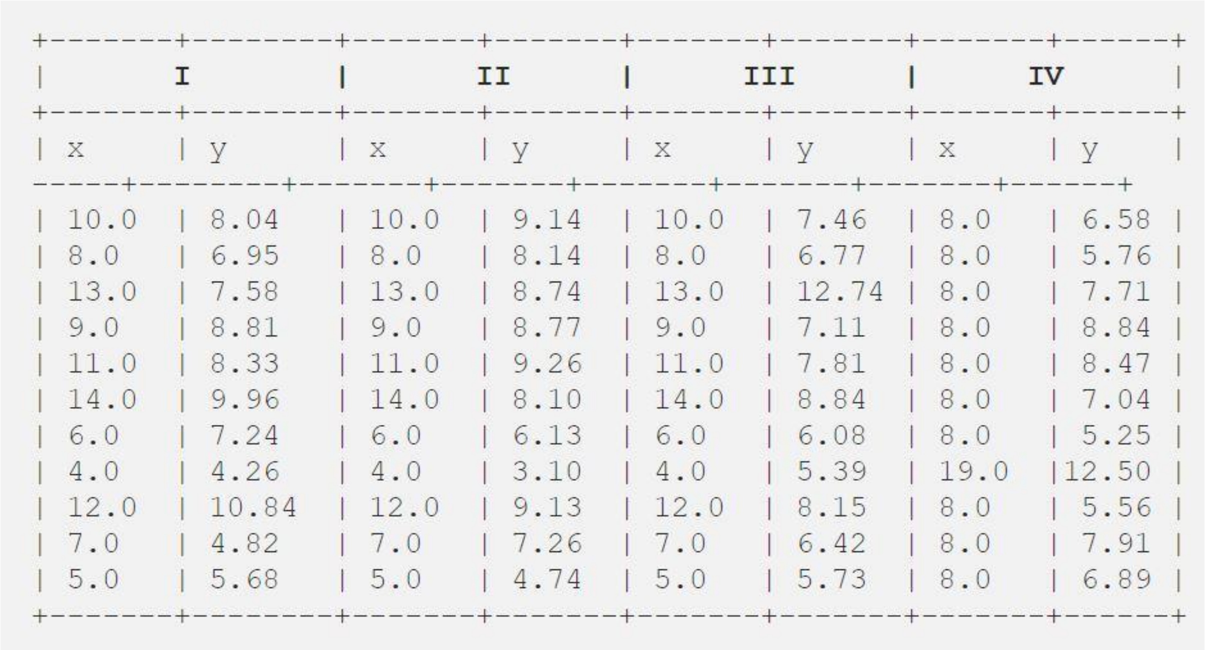
It is the most basic version of linear regression which predicts a response using a single feature. The assumption in SLR is that the two variables are linearly related.

2. Explain the Anscombe's quartet in detail. (3 marks)

- 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 analyzing it and the effect of outliers on statistical properties.

Simple understanding:

Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.



```
+-------+--------+-------+-------+-------+-------+-------+------+
|       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 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

3. What is Pearson’s R? (3 marks)

- Pearson correlation coefficient or **Pearson’s** correlation coefficient or **Pearson’s r** is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

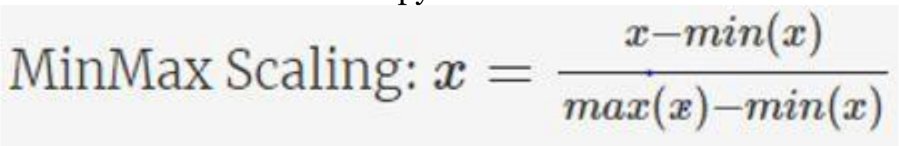
- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

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/Min-Max Scaling:

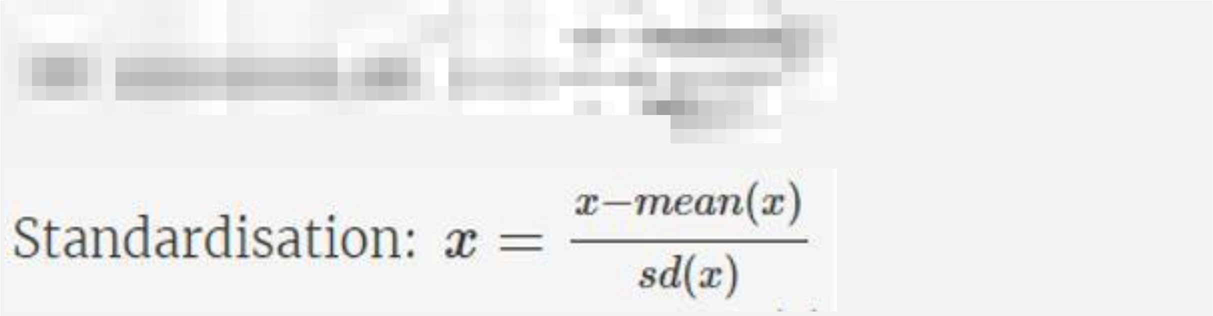
- It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.



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

- **sklearn.preprocessing.scale** helps to implement standardization in python.

- 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? (3 marks)

- If there is **perfect correlation**, then VIF = infinity. 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.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

- 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.