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? (3 marks)

Answer: There are few points that were inferred from the categorical variables of the dataset. These are,

i. Use of bike was increased in the 2nd year.
ii. The median value is higher in holidays.
iii. Mostly, the medians are same throughout the week.
iv. Month 1,2,11 and 12 are the lower side of bike usage.
v. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered always been the reasons of less bike rents.
vi. Summer and fall are the most preferable season for cycling.

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

Answer: drop_first=True, it drops the first column of the newly created dummy variable. The data complexity is reduced by dropping a column and also reduces the correlation created among the dummy variables.
Lastly, if 'n' number of variable can be defined with 'n-1' dummy variable then dropping one column is logical.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: The target variable is 'cnt'. Therefore the highest correlation is 0.95 with 'registered'.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: There are two steps were performed for the assumptions of the linear regression.

a. The plot of error terms and visually analyzed with the normal distribution from the residual analysis.
b. Keep check and update the variables (for thir p-value and VIF) to avoid multicollinearity.

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

Answer:
Top 3 contributed features:

a. atemp
b. yr
c. season_4 or winter.

General Subjective Questions

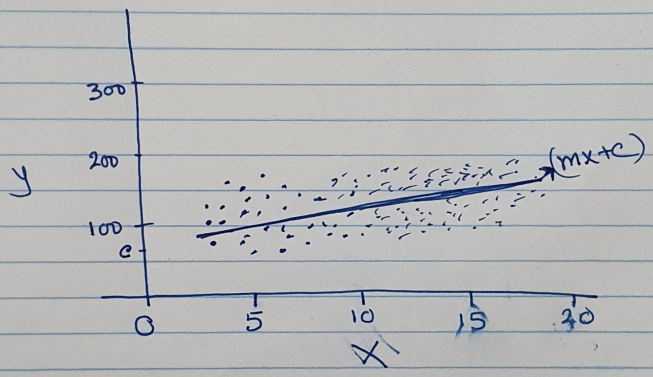
1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

It is a machine learning algorithm, based on Supervised learning model. The linear regression model performs the regression task to model a prediction target based on the variables. The main Job of the model is to find a relationship between different variable and the predictive values.

Suppose we have a dependent variable 'y' based on the independent variable 'x'. So the regression model try to find a linear relationship 'y = mx + c' in the data set, in the figure below.

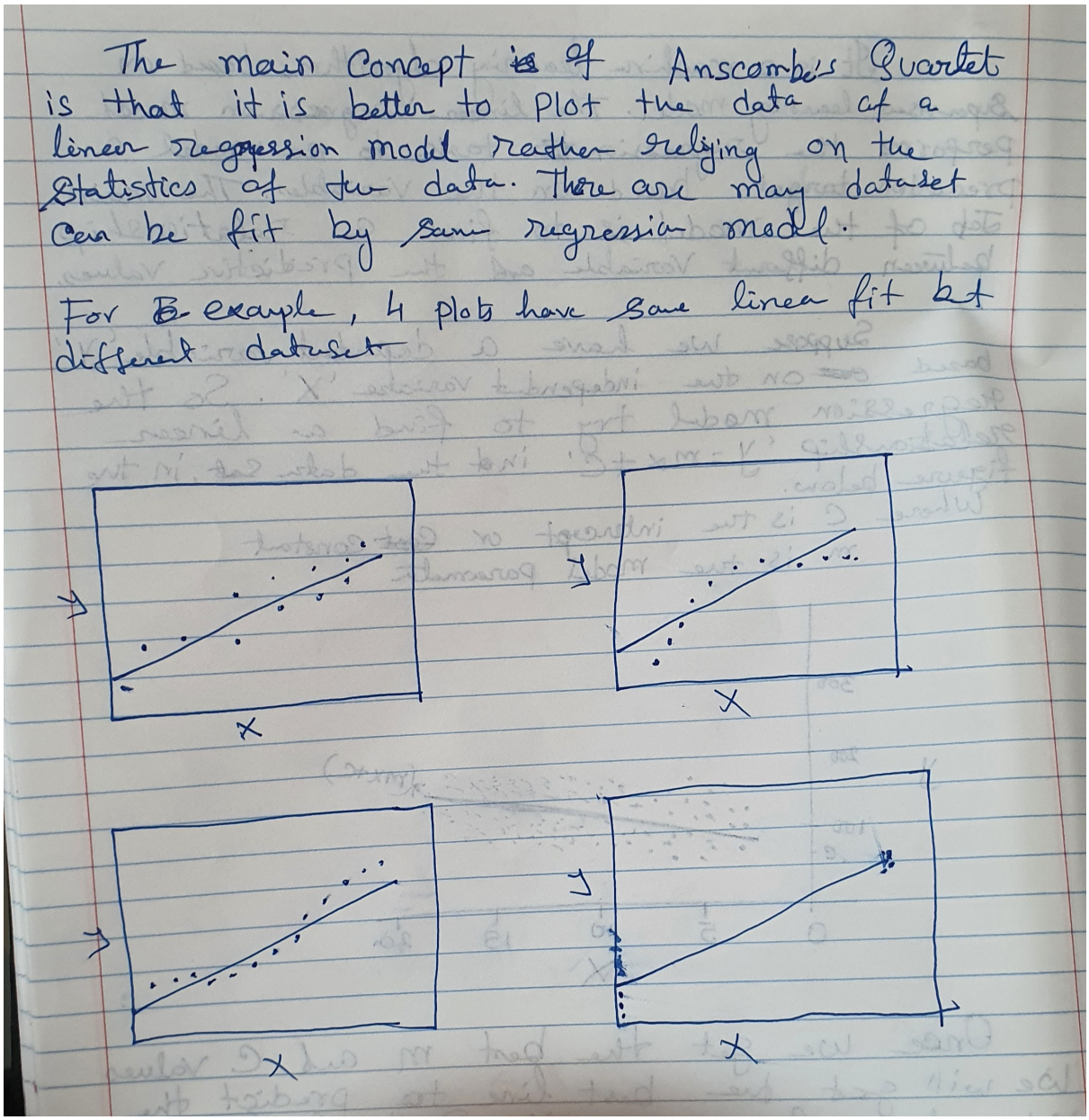
Where c is the intercept or Constant
m is the model parameter



Once we get the best m and c values we will get the best line to predict the value of y for a given X. (for Best Fit)

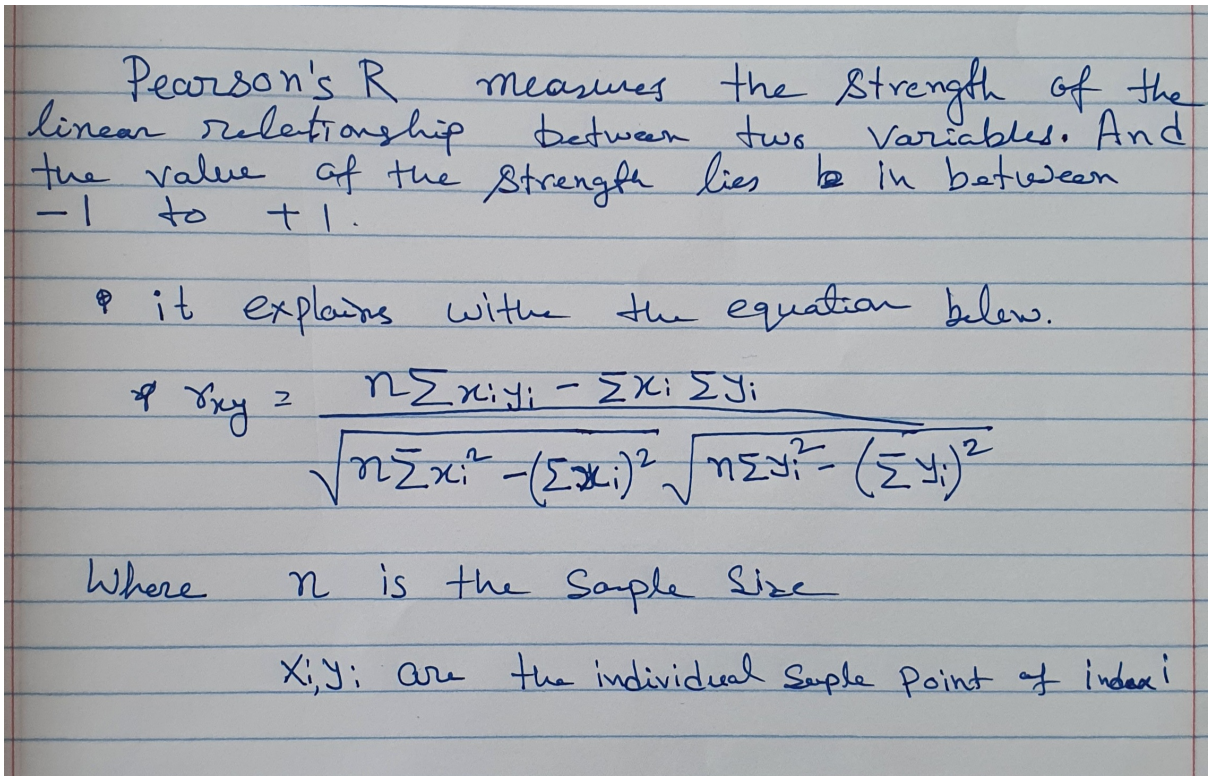
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:



3. What is Pearson's R? (3 marks)

Answer:



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

Answer: The scaling is a technique to standardize the independent feature into a fixed range. In a dataset, the different features can be in different values range. All the features can be transformed into a fixed range to make a balance of large and small value coefficients.

When the values are centered around the mean with unit standard deviation is called standardized scaling. In normalized scaling, values are transformed into 0 to 1 range.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: VIF shows the correlations among the variables. VIF=1.0 stands for orthogonal. And if it perfectly correlated then VIF= infinity.

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

Answer: The Q-Q plot is graphical probability quantile plot for comparing two probability distributions.
It is important to define whether the dataset came from normal, uniform or exponential distribution.