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?

Across categorical variables, the median of the variable **cnt** is changing across categories, that means the levels in each category has an impact on the variable **cnt**

As and when the weather situation worsens that is from clear sky to snow, the likeliness of people using boom bikes reduces

Working Day does not have an adverse effect on the **cnt** variable, and hence shall be modelled carefully and can be looked at a potential variable to remove

The variation in **cnt** variable is not very significant to visualize on the graph shall be better explored during model building

Like Weather Season has a impact on the **cnt** variable, users use it to more in the fall and summer compared to other seasons

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

Its important because now that we have transformed the original variable into a form that can be fed to the model, we no more need the original variable

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

It's the variable temp with a correlation of 0.63

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

Initially using correlation plots observed variables such as temp being linearly correlated to **cnt** variable and hence LR is possibe, plotted Distribution of error terms that is y_train-y_train_pred to check if it's a normal distribution, plotted errors terms to check homoskedasticity

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

Temperature, Weather Situation being Snowy and Year

Year over Year, the business is increasing, temperature has a direct correlation with the cnt variable and has a co efficient of 0.43, increase in 1 unit scale of temp can increase demand by 0.43 units

If the weather situation is snowy sorts, the demand is expected to decrease by 0.28 units

General Subjective Questions

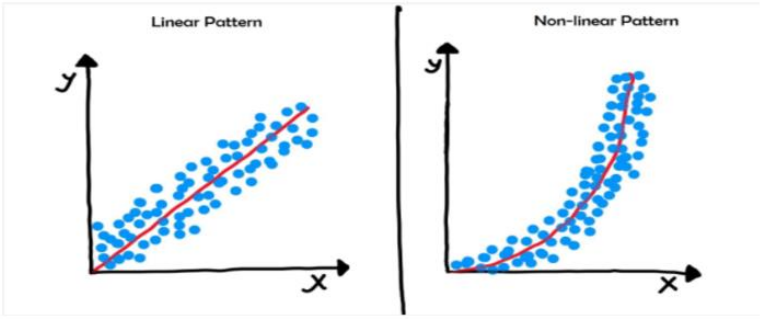
1. Explain the linear regression algorithm in detail.

Linear Regression is a ML technique that falls into Supervised section of models and is aimed at being able to predict a continuous variable

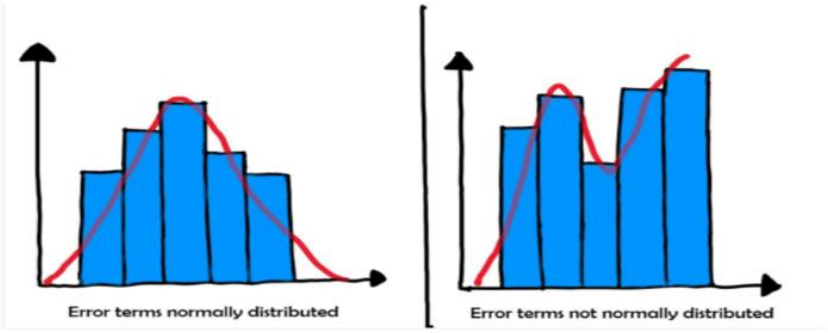
We always have one target variable which needs to be predicted and one independent variable in Simple Linear Regression and multiple independent variables in Multiple Linear Regression

There are few assumptions which should hold good before applying this model:

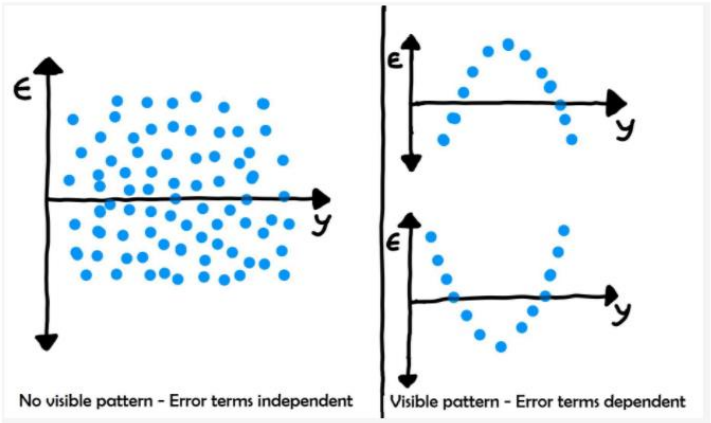
There is a linear relationship between independent and dependent variable



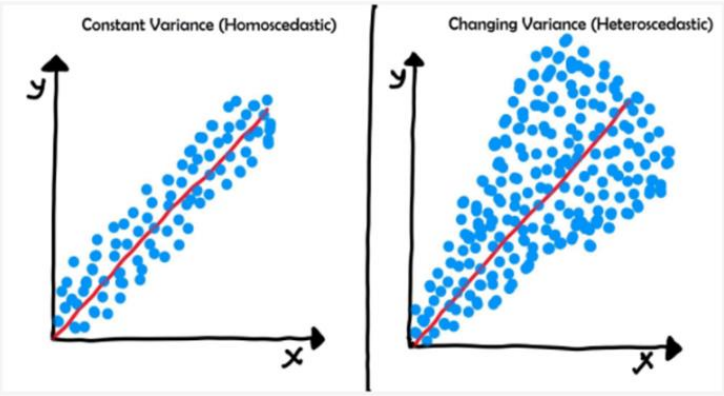
Error terms are normally distributed



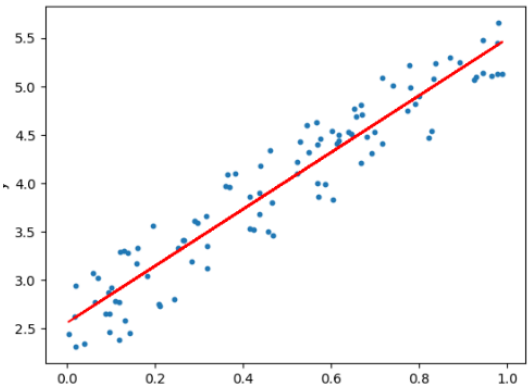
Error terms are independent of each other



Error terms have a constant variance (Homoskedasticity)



To visualize how an LR Model is built on a graph, imagine a scatter plot between X and Y and there is only one line which possibly covers all these scatter dots in the best possible way,



There are different methods to achieve this best fit line,

The equation for the regression line is given by below for n independent variables,

$$Y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$$

Mostly the cost function in this model is the sum of squared residuals, its just the approach to minimize this is different across different methods

Ordinary Least Squares or OLS method:

This procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seeks to minimize.

Gradient Descent:

This works by starting with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

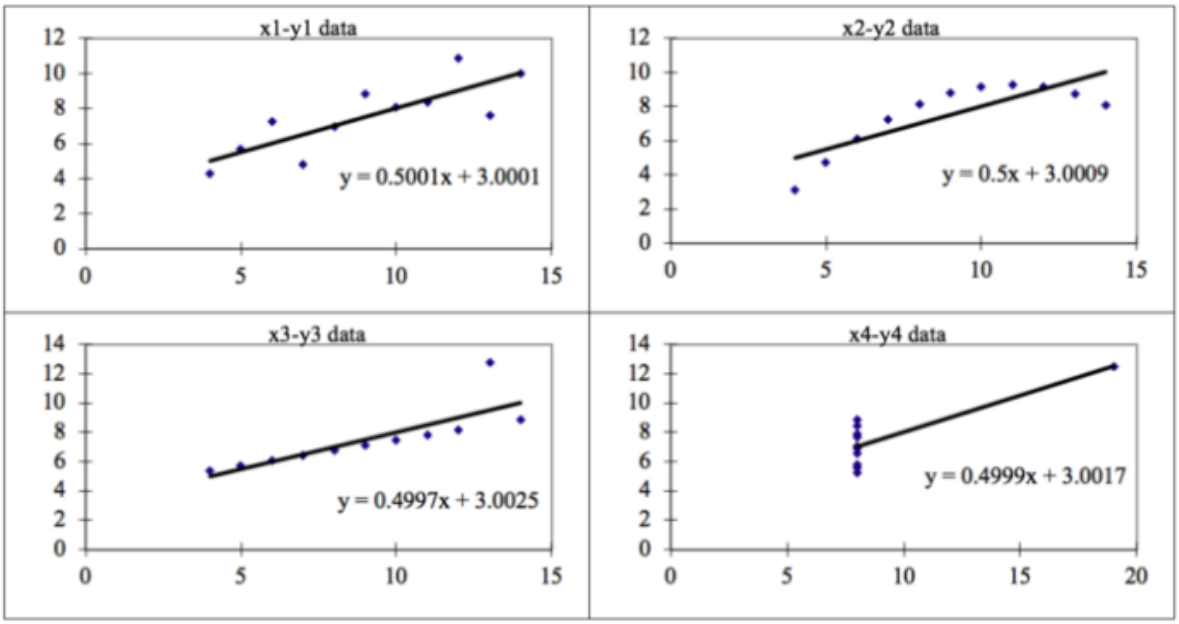
There are certain hypotheses on the co-efficients of the equation for a model's coefficients to be true to consume, the null hypothesis is that the coefficient of a variable is 0, we reject the null hypothesis in this context to approve of the model's co-efficients

Handling multi collinearity with scores such as VIF and scaling all the independent variables are very crucial steps to building this model

2. Explain the Anscombe's quartet in detail.

As the word Quartet suggests 4, it consists of 4 datasets that were developed by Anscombe himself in the year 1973 to demonstrate the importance of visualizing the data through graphs and also to learn about the effect of outliers and other influential observations on different statistical properties, he wanted to counter the impression that most of them have which is graphs are rough visualizations vs numbers being accurate

These 4 datasets that were developed by him had identical or similar descriptive statistics but their data distributions appeared very different when visualized via graph



The 4 datasets can be described as below,

Dataset 1 fits the Linear Regression model very well

Dataset 2 could not fit that well, as the data depicts a non-linear relationship

Dataset 3 shows that the outliers in a dataset cannot be handled by a LR model

Dataset 4 shows that even one data is enough to produce a correlation co-efficient, even though the other variables not at all linear

These datasets were created to show importance of visualization and how someone can be fooled by a simple LR algorithm, hence all-important features shall be visualized before implementing a good model and interpreting the model

3. What is Pearson's R?

Pearson's R is also known as the correlation co-efficient, which helps us with the quantification of the association between variables. The formula is like below between 2 sets of variables X and Y,

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

The r value is positive if the association or relation between 2 variables is positive i.e., X increases as Y increases

Similarly, it can be negative if the relation is negative, X decreases as Y increases

If the value is too small or equal to 0, it depicts no relation between the variables at all

The value of r lies between -1 and 1, the value here indicates the direction and also the value

It is better to consume and interpret the Pearson's R values when:

Variables are approx. normally distributed

Outliers are removed

Scale of measurement being interval or ratio

Hoping the actual relationship to be linear

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

Scaling is a method that is applied on independent variables to scale numeric features in the same scale or range

Scaling features is the last step in the data pre-processing and before model building, especially in Multiple Linear Regression when we have a lot of independent variables and all these variables are on different scales, when we build a model using these variables without scaling them, the co-efficients that come out of the model will be weird and will be hard to interpret

As per ML methods, we fit the feature scaling basis the train data and transform it both on train and test data, we should never fit the scaling on test data

Scaling is essential because of 2 main reasons:

a. Ease of interpretation
b. Faster convergence for Gradient Descent Methods

When scaling happens, it is affects only the co-efficients and has no impact on the model accuracy or various model statistic values like F statistic, t statistic, p-value, R-Square etc.

There are 2 very popular methods in Feature Scaling:

a. Standardized Scaling: Formula goes as below, here variables are scaled in such a way that the Mean is 0 and Standard Deviation is 1

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

b. Normalized Scaling: Formula goes as below, here variables are scaled in such a way that, all values lie between 0 and 1 using the max and min values in the data

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

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

Judging this from a formula POV,

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

VIF can here be infinite only if the denominator is 0 and that can happen only when R-Square is 1

R Square can only be if the fit is perfect and the model is able to explain complete variance, such models are still not good, because they don't generalise well against new data like test data, this will also mean that a given variable can be exactly explained by other variables in a linear fashion and should ideally be removed from the analysis

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

Q-Q Plot is one type of a plot that does help us understand if a set of data points possibly come from theoretical distributions like Normal, Exponential or Uniform

It also helps us in determining if 2 datasets come from populations with a common distribution

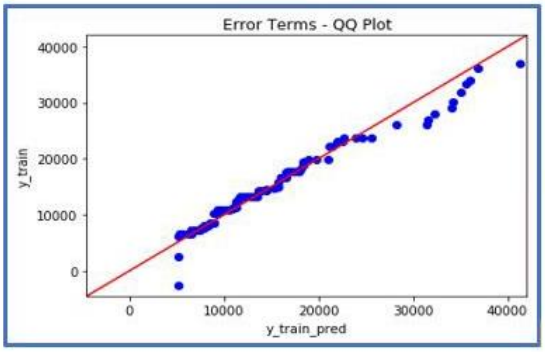
This is also a plot that plots a set of quantiles from first dataset against the second

It is important in linear regression to sometimes check if the provided train and test datasets come from the same set of population, or also if the predictions and actuals come from the same distribution of the data, if they do come from a similar distribution that means their behavior is similar and we are on the verge of creating a model that predicts realistic outcomes

We can have the below interpretations from the Q-Q plot:

Similar Distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

Y Values < X Values or X Values < Y Values: If one set of quantiles are lower than the other (maybe might help us understand if our model is underpredicting or over predicting)



Different Distribution: All point of quantiles lie away from the straight line which is at a 45 degree line from X-Axis