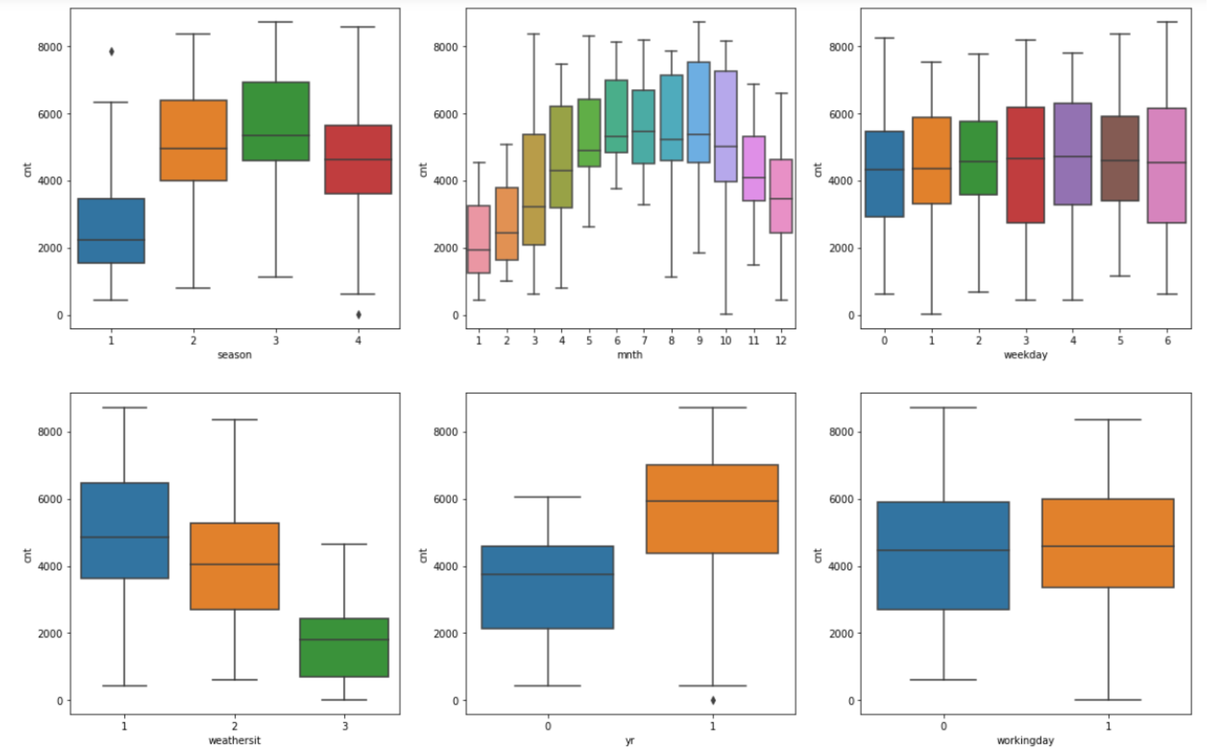
Assignment-based Subjective Questions

1. ***From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?***

Ans: We are having below categorical variables:

1) season
2) mnth
3) weekday
4) weathersit
5) yr
6) workingday



weathersit :

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

We see Light snow has least median count close to 2000 while other categories are close to 4000.

weekday : day of the week :: All the categories count are same approximately close to 4500

mnth : month (1 to 12) :: All the months median count ranges from 2000 to close to 5000
season : season (1:spring, 2:summer, 3:fall, 4:winter) :: We see that spring season has median count close to 2000 while other seasons have close to 5000 count.

yr : year (0: 2018, 1:2019) :: We can see yr=1 (cnt = 6000) has median count much higher than yr=0 (cnt =4000)

workingday : if day is neither weekend nor holiday is 1, otherwise is 0 :: We infer that median for both the working days as approximately same.

Effect of categorical variables on the final model:

season(summer)
season(winter)
mnth(sep)
weathersit(Light Snow)
weathersit(misty)

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

   Ans: It is important to use drop_first = True because it helps in deleting the extra column created during dummy variable encoding. Hence it reduces the correlations created among dummy variables. i.e., let's say we have 3 types of values in Categorical column and we want to create dummy variable create dummy variable for that column. If one variable is not summer, not spring and not fall, then it is obvious winte . So, we do not need 4th variable to identify the winter.

   For representing **k** level of variable data, we need **k-1** columns to represent the data. Because 2 columns are enough to represent 3 variables i.e. **00, 01, 10**

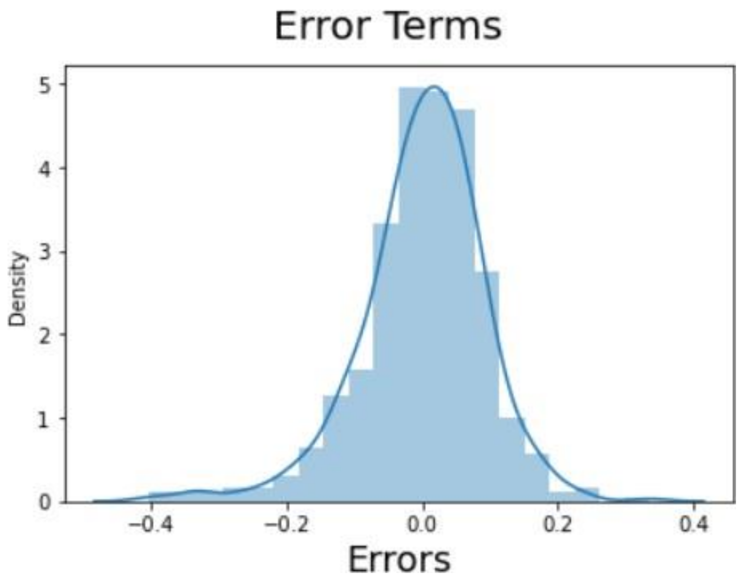
3. ***Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?***

   Ans: 'temp' variable is having the highest correlation with the target variable 'cnt' and is **0.64**

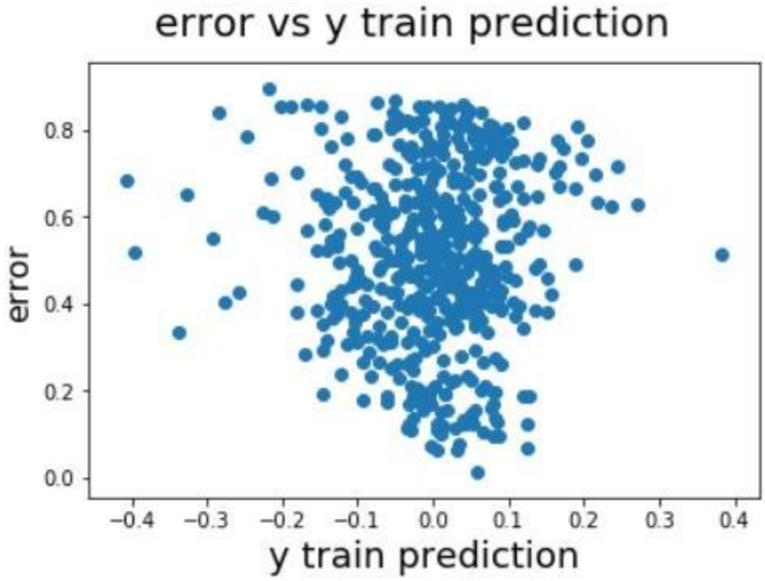
4. ***How did you validate the assumptions of Linear Regression after building the model on the training set?***

   Ans: We have various assumptions in Linear Regression, which we validate after model building.

   1- Error terms are normally distributed with zero mean



2- Error terms have constant variance (homoscedasticity)



3- Linear relationship between X and Y

4- Error terms are independent of each other

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

Ans: Below are the top three features of the final model:

1- temp

2- yr

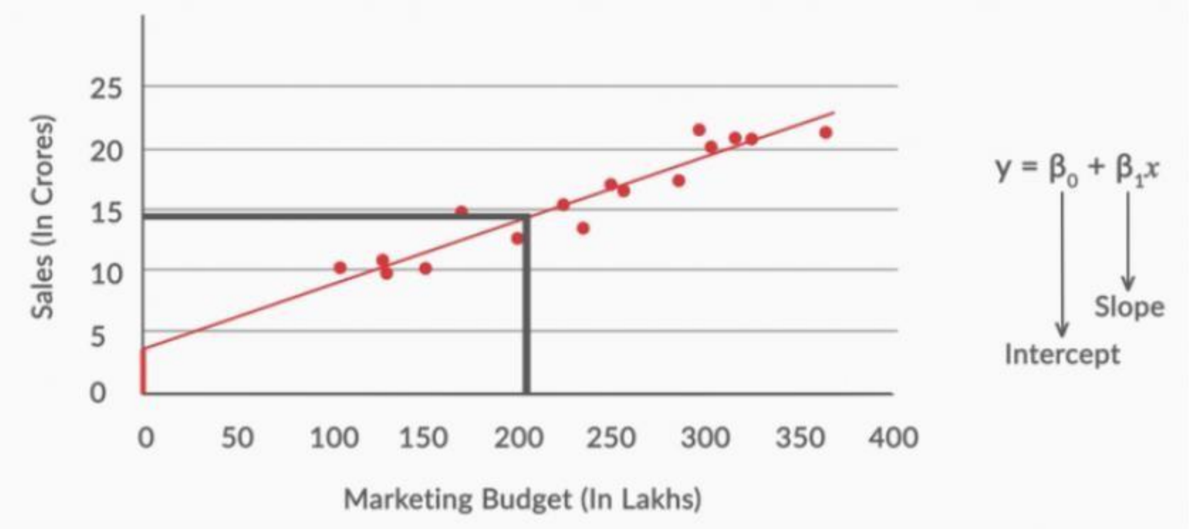
3- weathersit_Light Snow

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: The most basic and widely used model in the industry for model building is the Linear Regression. Simple Linear regression is basically a Linear plot of an independent variable to a dependent variable. In simple maths we can say **y = mx + c**. So change in x independent variable makes impact on dependant variable y by m times, which we basically call coefficient or slope.

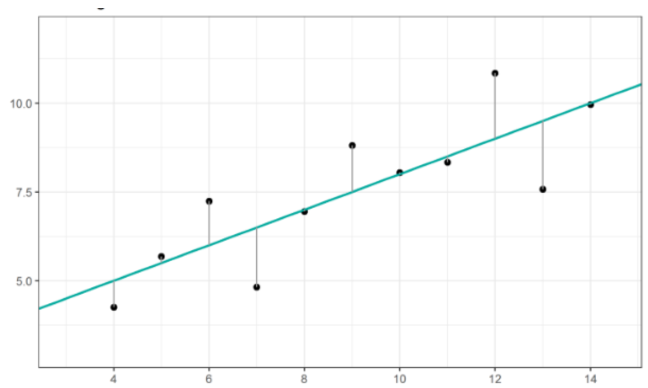
The standard equation for Simple linear regression is given by **$Y = \beta_0 + \beta_1 X$**



Now, there will be so many data points which will be there when we make a **scatter plot** between independent and dependant variable. So we try to find the best fit line which passes form the maximum data points and try to fit maximim possible values.

The best-fit line is found by minimising expression of **Residual Sum of Squares (RSS)** which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.

RSS 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:



Formulae:

$$RSS = \sum_{i=1}^{n}(Yi - \beta 0 - \beta 1Xi)^2$$

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

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

The strength of the linear regression model can be assessed using 2 metrics:

1 - **R²** or Coefficient of Determination

2 - **Residual Standard Error** (RSE) R² or Coefficient of Determination

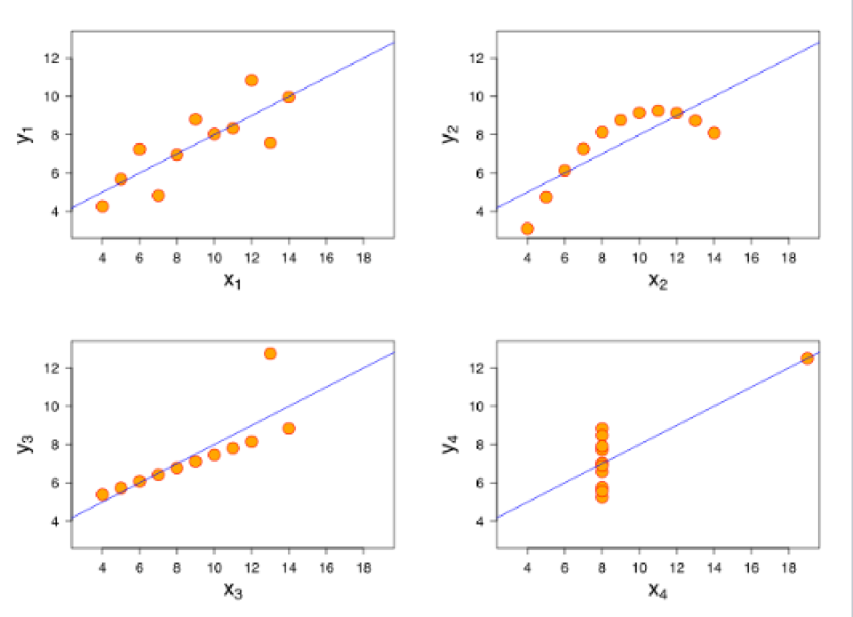
You also learnt an alternative way of checking the accuracy of your model, which is R2 statistics. R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. Overall, the higher the R-squared, the better the model fits your data. It is represented as: **R² = 1 - (RSS / TSS)**

2. Explain the Anscombe's quartet in detail.

Ans: Ansombes quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

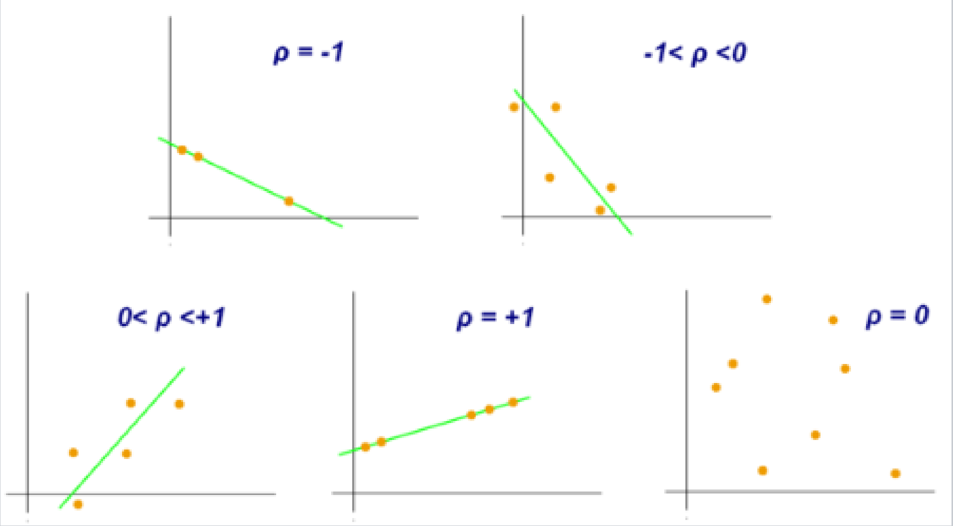
They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

3. *What is Pearson's R?*

Ans: Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables. Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1



Examples of scatter diagrams with different values of correlation coefficient (ρ)

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decrease in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn’t a positive or negative increase. The two just aren’t related.

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-0.75| = 0.75, which has a stronger relationship than 0.65

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

Ans: Scaling is a method used to normalize the range of independent variables or features of data. When you have 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.

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

Standardization technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

2- **MinMax 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.

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

MinMax scaling is a technique in which values are shifted and rescaled so that they end up ranging between 0 and 1.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1 r = +1 data lie on a perfect straight line with a positive slope
- If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1

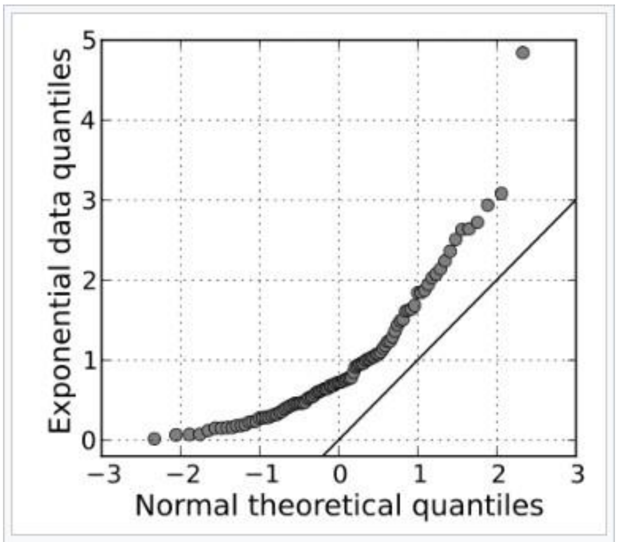
5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

Ans: Infinite Variance Inflation Factor indicates a perfect correlation between all the independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To overcome this problem, 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.**

Ans: Quantile-Quantile plots (Q Q 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 45degree 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.

Quantile-Quantile plots are important in linear regression as it let you check that the data meet the assumption of normality. They compare the distribution of your data to a normal distribution by plotting the quartiles of your data against the quartiles of a normal distribution. (If your data are normally distributed then they should form an approximately straight line)


Exponential data quantiles
Normal theoretical quantiles