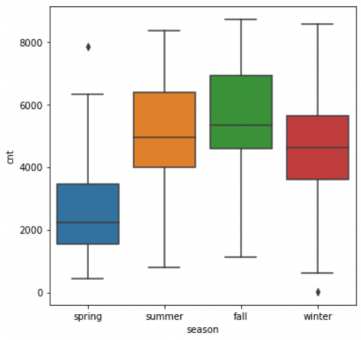
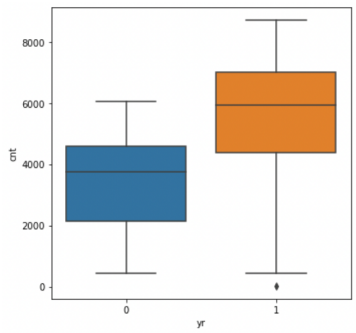
Assignment-based Subjective Questions

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

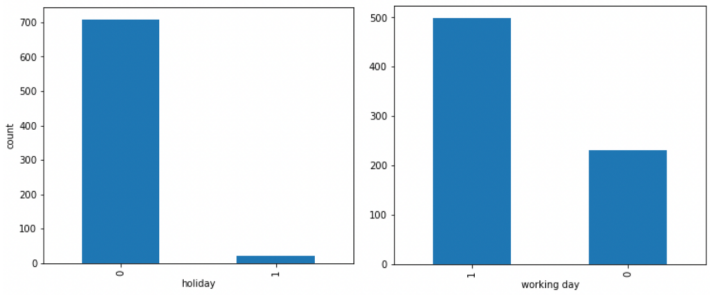
    - Based on the seasons the no of users are more during summer and fall and least during spring.



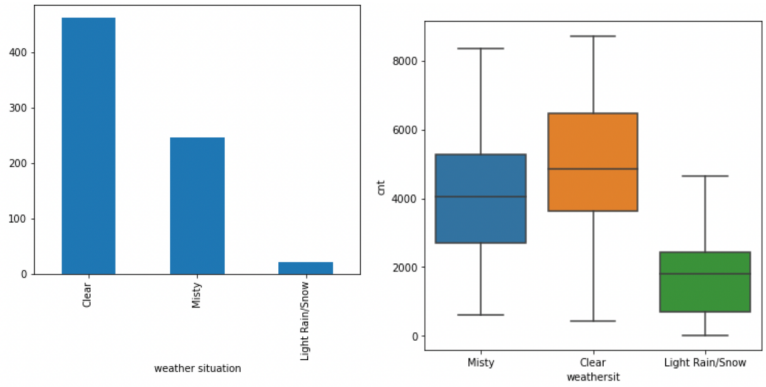
    - The users in 2019 is more that 2018 which tells us that the demand increases with time.



    - In general the no of users are more during working days and less during holiday.



    - The no of users are more when the weather situation is clear



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

During model creation the categorical variables cannot be used as it is, since it cannot be quantified. Hence this needs to be converted to numerical.

This is done by creating dummy variables. The number of dummy variables required for a categorical variable depends on the number of possible values possible for the later. The dummy variable represent each of these possible values of categorical variable. The value of this dummy variable for each instance is set to 1 if the corresponding instant's value for categorical variable is the same as the dummy variable, if not it will be set to 0.

If the categorical variable can have 3 different values then the number of variables required is 2. The reason why we removed one here is because when the value of both the dummy variables is 0, it means that the 3rd value is 1 and this need not to be represented. This will then reduce the correlation between the dummy variables.

For ex:

Consider the weather situation. As per our analysis on the data set, it could have values Clear, Misty, Light Rain/Snow. The dummy variables created can be Misty & Light Rain/Snow. Now the data would look like:

Index	Weather Situation
1	Clear
2	Misty
3	Light Rain/Snow
4	Clear

Index	Misty	Light Rain/Snow
1	0	0
2	1	0
3	0	1
4	0	0

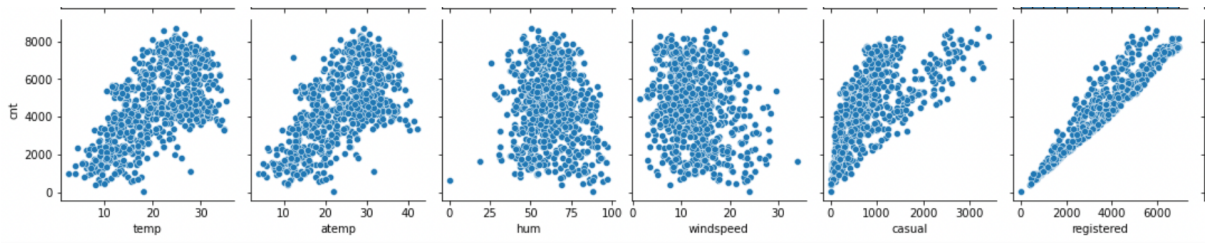
In the above table the index 1 and 4 have weather situation clear and on the table on the RHS with the dummy variables, the value for Misty and Light Rain/Snow is represented as 0 which indicates that the weather situation is Clear.

Hence while creating dummy variable we perform drop_first=True, if not done it will create one extra dummy variable which is not required and it may cause correlation between dummy variables.

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

The target value chosen is count(cnt), it has highest correlation with variable registered.

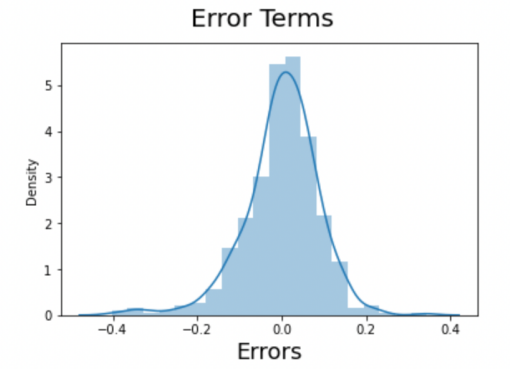
Next comes the variable temperature(temp) or the feeling temperature (atemp) with high correlation with the target variable count(cnt).



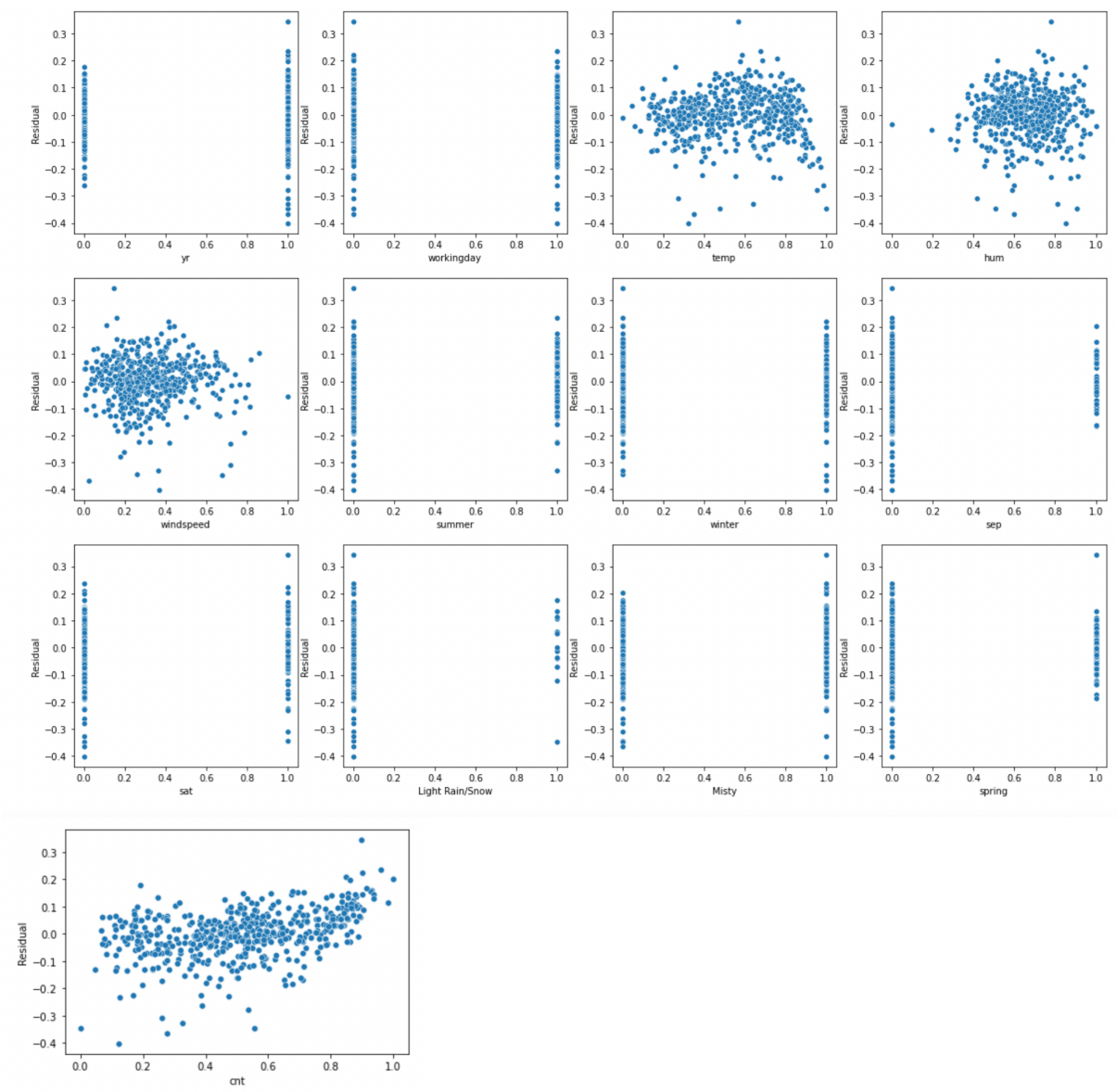
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The following steps are done to validate the assumptions:

1. A histogram on error terms (y_train – y_train_pred) is ploted, and this should be normally distributed.



2. There should not be any visible pattern b/w error term and the variables i.e error terms should be independent. This was validated by plotting scatter plots b/w error terms and other variables.



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

    1. Temperation – temp (+ve impact)
    2. Light Rain/Snow – Weather situation (-ve impact)
    3. Year – yr (+ve impact)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithm is a supervised learning method, where we provide a set of historical data on which the model will be trained to create a linear equation.

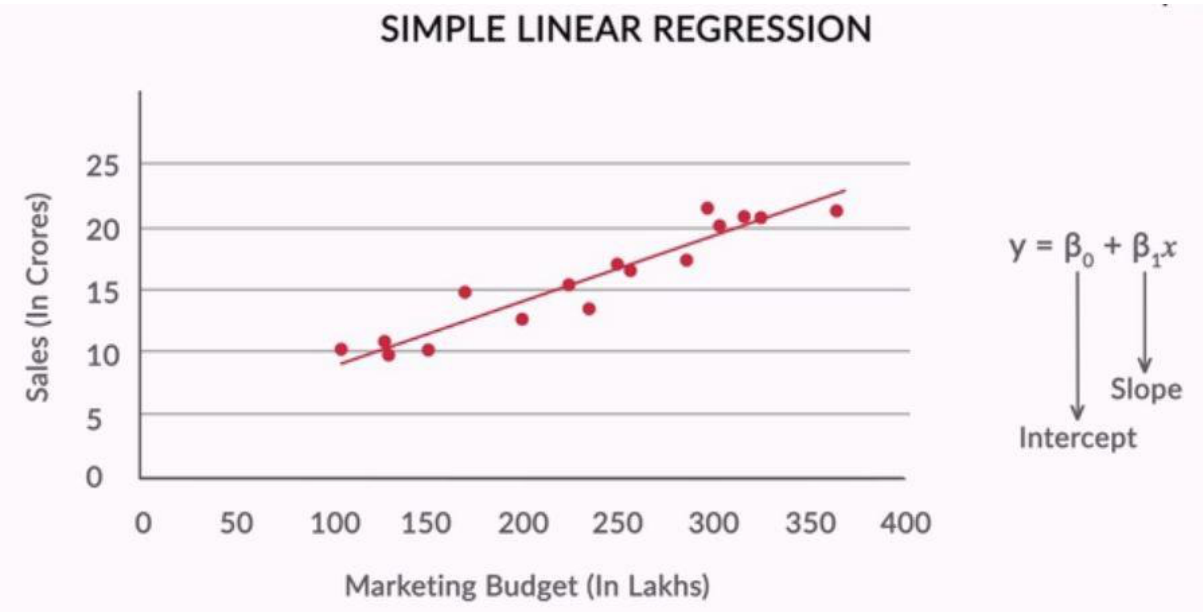
The linear equation comprises of dependent variable which is the target variable (the value we want to predict) and one or more independent variable using which the dependent variable is predicted.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

To understand linear regression in more detail, it can be classified into two:

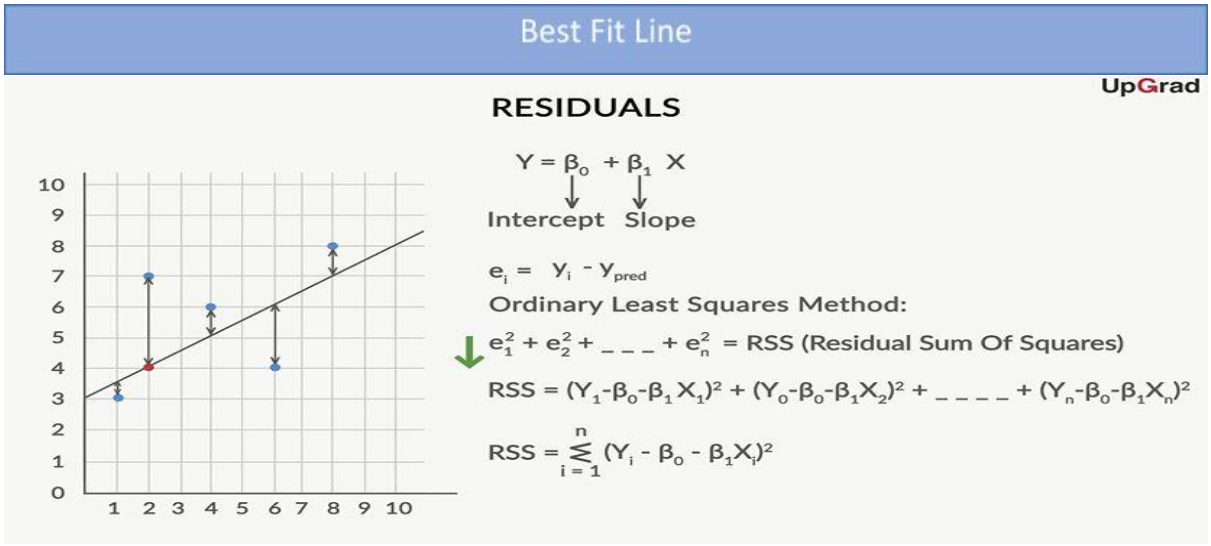
- Simple Linear regression
- Multiple Linear regression

Simple Linear regression is the most elementary type of regression model. It 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 line $y = \beta_0 + \beta_1 X_1$ here would be used by model to predict the values for y based on X_1. In order to have good accuracy the value of β_0 & β_1 should be such that the line should be a best fit, so that this can be used to predict the unseen data as well.

To find the best fit we use the metric **RSS, Residual Sum of Squares**. Residual is the perpendicular distance b/w actual y value and the predicted y-value represented by the line. RSS is the sum of the squares of these residual values.
The value of β_0 & β_1 should be such that the RSS value should be minimum.



Apart from RSS there is one more metric call TSS, Total Sum of Squares. It is the sum of residuals of the datapoints from mean of response variable (if we draw a line passing through the mean of all the data points). This would be the max error that a model could possibly have.

Now the ratio **RSS/TSS** will give us the total variation in the outcome that the model couldn't explain (as the ratio represent the error). Hence in order to find the strength of the model we could do **1 - RSS/TSS**. This is called the $\mathbf{R^2}$.

Multiple Linear Regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

The metrics used to evaluate the strength is same as that of simple linear regression.

Assumptions of Linear Regression

1. Linear relationship between $X_{1..n}$ and Y
2. Error terms are normally distributed
3. Error terms are independent to each other
4. Error terms have constant variance(homoscedasticity)

One thing to note when handling multiple independent variable is the multicollinearity, i.e the dependency of X's (independent variables) with each other. This should be minimised as this could cause issues during interpretation of the model.

2. Explain the Anscombe's quartet in detail.

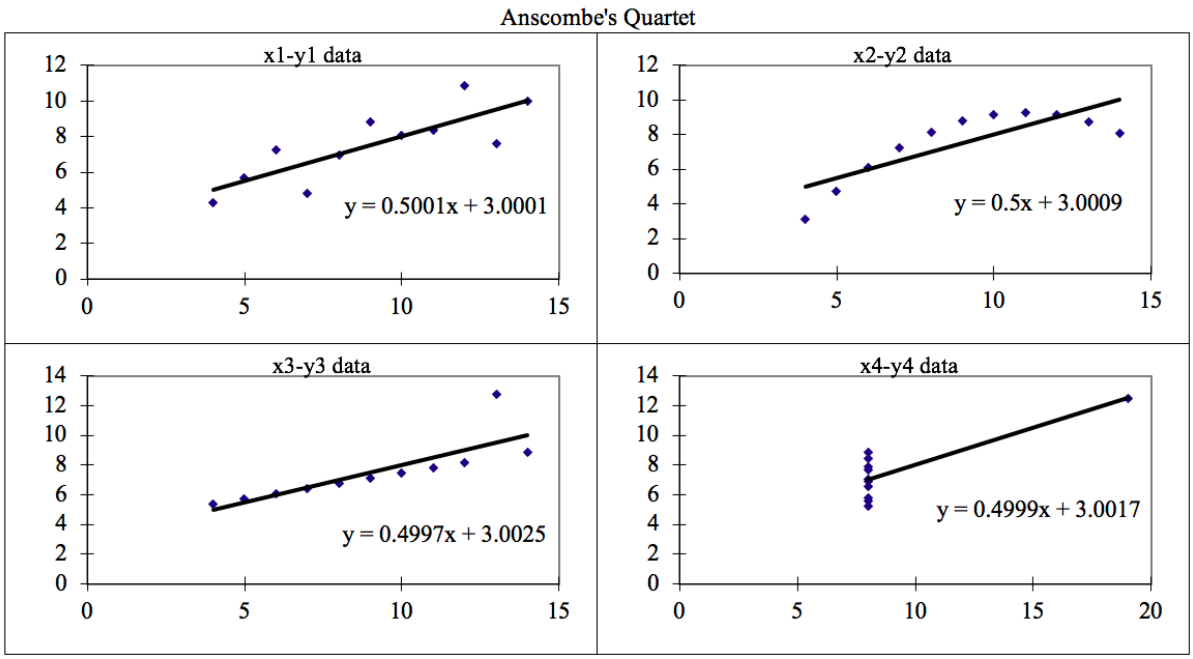
Anscombe's quartet is a group of four data sets which are nearly identical in simple descriptive statistics, but they have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. This tells us the importance of visualising the data before applying various algorithms to build models out of them. This helps to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, etc.

These four plots can be defined as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

In the summary statistics section we can see that the statistical information for all these four datasets are approximately similar.
When these plots are plotted in the scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm.



In the above figure we can see that the linear equation that satisfies all the four datasets are similar. These have nearly same statistical observations, which provides same statistical information that involves variance**,** and mean of all x,y points in all four datasets.

The four datasets can be described as:

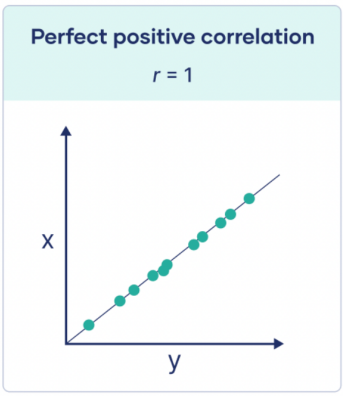
- Dataset 1**:** this fits the linear regression model pretty well.
- Dataset 2**:** this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
- Dataset 4**:** shows the outliers involved in the dataset which cannot be handled by linear regression model

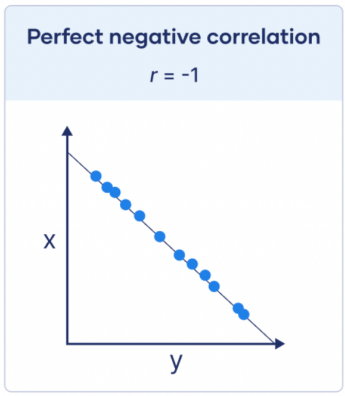
3. What is Pearson's R?

Pearson's R is the way of measuring a linear correlation. It is the number between -1 and 1 which measures the strength and direction of the relationship between two variables.

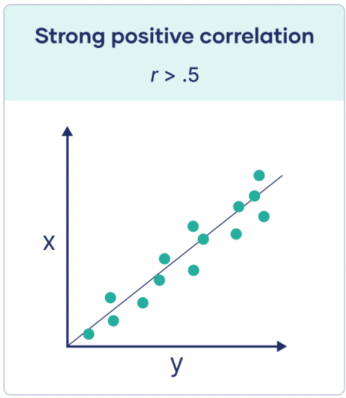
It is a measure of how close the observations are to the line of best fit. It also tells whether the slope is positive or negative. When the slope is negative the Pearson's R is negative, when the slope is positive then the Pearson's R is positive.

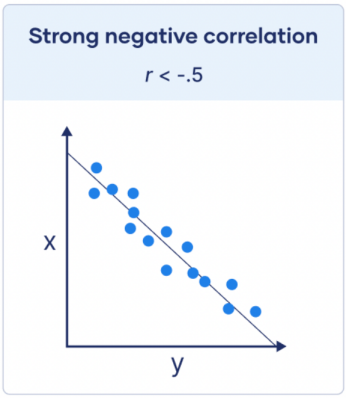
When Pearson's R is 1 or -1 the observations fall exactly on the best fit line.



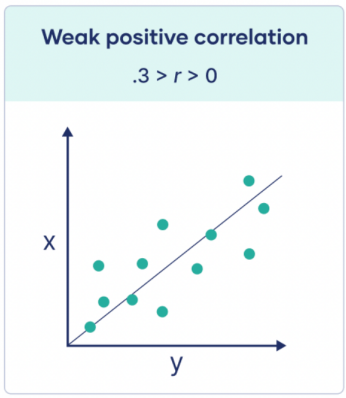


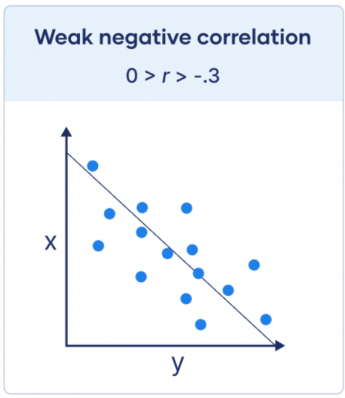
When Pearson's R is greater than .5 or less than –.5, the points are close to the line of best fit:



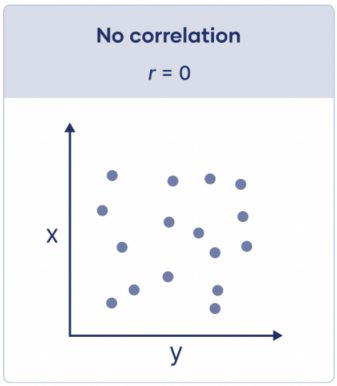


When Pearson's R is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:





When Pearson's R is 0, a line of best fit is not helpful in describing the relationship between the variables:



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

Scaling is the mechanism to standardize the independent features within a fixed range. It is done in the pre-processing step.

Scaling is done so that all the features are in comparable scales. If scaling is not performed then the coefficients for some of the features obtained by the model would be very large or very small compared to other features. This makes it difficult to interpret the model and for model evaluation.

Normalized scaling and standardized scaling are the two common methods of scaling the features.

1. Normalized Scaling:

   This is also called Min-Max Scaling.

   Formula:

   X` = (X - X_min)/(X_max - X_min)

   Where,

   X` is the new value,
   X is the current value
   X_min is the min value of X in the data set
   X_max is the max value of X in the data set

   This will scale the values b/w the range [0,1].

2. Standardized Scaling:

   It is the transformation of features by subtracting from mean and dividing by standard deviation.

Formula:

X` = (X - mean)/Std

Where,

X` is the new value,
X is the current value
mean is the avg/mean value of the data set
std is the standard deviation of the data set.

Here the result is not bound to any range. This is usually used when the feature distribution is normal. It is also called as Z-Score Normalization.

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

VIF- Variance Inflation Factor indicates how dependent a feature is with all the other features(multicollinearity). This is one of the key metric used to determine how significant a feature in a model creation. Higher the value indicates more the dependency. As a general thumb rule anything above 5 means it has good dependency and could be ignored from the model.

Formula:

$VIF = 1/(1-R^2)$

Where, R-squared is the coefficient of determination in linear regression. Its value lies between 0 and 1.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. From the formula we can see that, VIF will be infinite if R^2 is 1, this will happen if there is a perfect correlation between the independent variables.

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

Q-Q Plot is called Quantile-Quantile plot. It is obtained when the quantiles of two variables are plotted 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. This plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations.

If two samples have similar distribution all points on the plot lies on or close to straight line with an inclination of 45 deg.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.