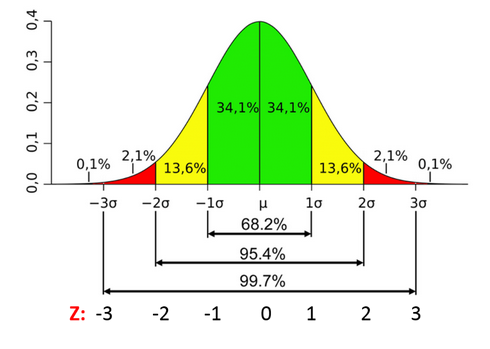
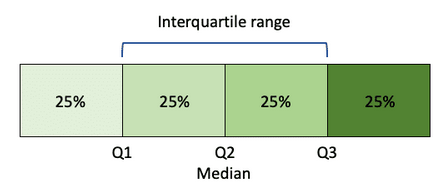
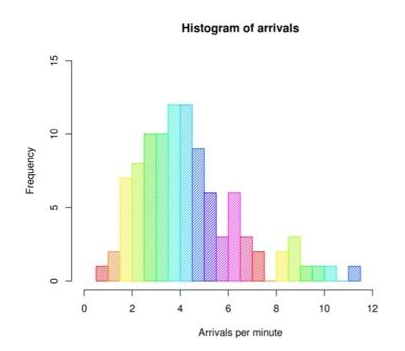
1. Univariate Analysis (Numerical Variable: (also called quantitative variable) is a quantifiable characteristic whose values are numbers. Numeric variables may be either continuous or discrete.)
   1. Mean: is the average of a data set.
   2. Median: is the most common number in a data set.
   3. Mode: is the middle of the set of numbers.
   4. Count: is how many items or observations you have in the data sets.
   5. Max: is the largest value in the data set.
   6. Min: is the smallest value in the data set.
   7. Quantiles(분위수): defines a particular part of a data set, i.e. a quantile determines how many values in a distribution are above or below a certain limit. Special quantiles are the quartile (quarter), the quintile (fifth) and percentiles (hundredth).
   8. Variance: The variance is a measure of variability. It is calculated by taking the average of squared deviations from the mean.
   9. Standard Deviation: The average distance between each value in your data set and the mean.



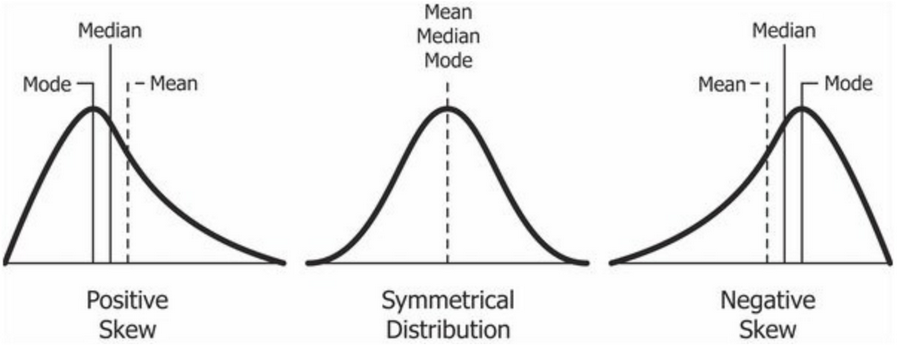
* 1. Variance vs Standard Deviation: Since the units of variance are much larger than those of a typical value of a data set, it’s harder to interpret the variance number intuitively. That’s why standard deviation is often preferred as a main measure of variability.
     1. Standard deviation is expressed in the same units as the original values (e.g., meters).
     2. Variance is expressed in much larger units (e.g., meters squared)
  2. Interquartile Range: In descriptive statistics, the interquartile range tells you the spread of the middle half of your distribution.



* 1. Histogram: is a bar graph-like representation of data that buckets a range of outcomes into columns along the x-axis.



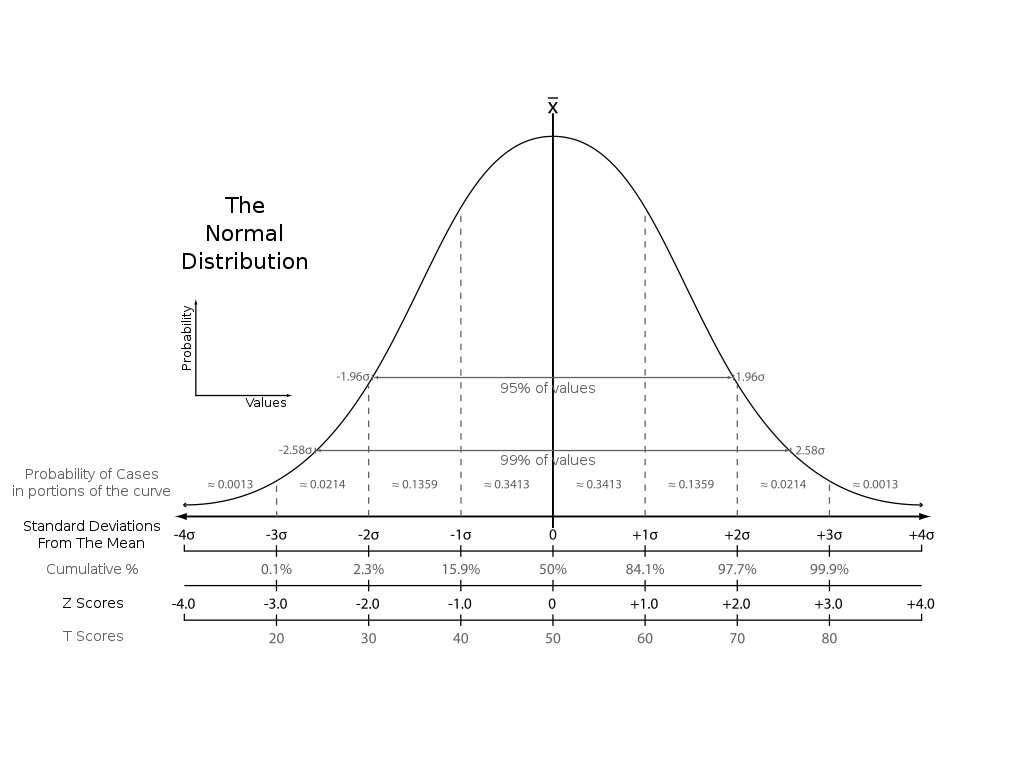
* 1. Skewness: in statistics represents an imbalance and asymmetry from the mean of a data distribution.



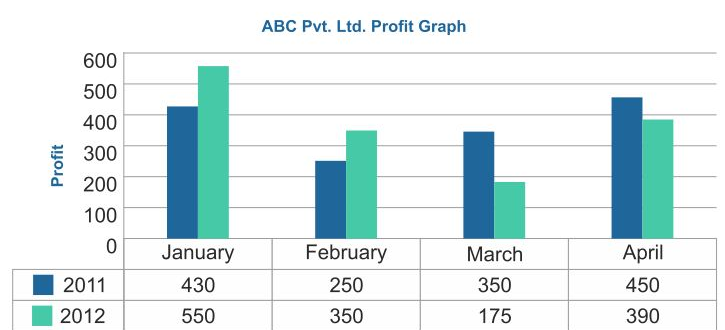
* 1. Data Standardization: is the process of putting different variables on the same scale. This process allows you to compare scores between different types of variables.
  2. Z-Score(Standard Score): is the number of standard deviations away from the mean a certain data point.

It is useful to standardized the values (raw scores) of a normal distribution by converting them into z-scores because:

* + 1. it allows researchers to calculate the probability of a score occurring within a standard normal distribution;
    2. and enables us to compare two scores that are from different samples (which may have different means and standard deviations).

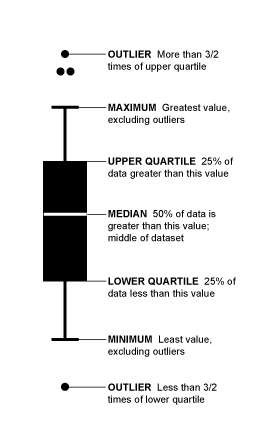
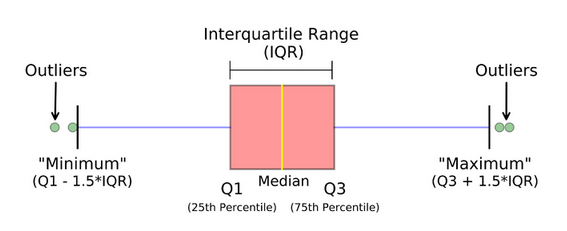


1. Univariate Analysis (Categorical Variable: (also called qualitative variable) take on values that are names or labels. The color of a ball (e.g., red, green, blue).)
   1. Count: is how many items or observations you have in the data sets.
   2. Frequency: is how many items or observations you have in the data sets.
   3. Percentage: is a representation that displays the number of observations within a given interval.
   4. Bar Chart: show the frequency counts of values for the different levels of a categorical or nominal variable.



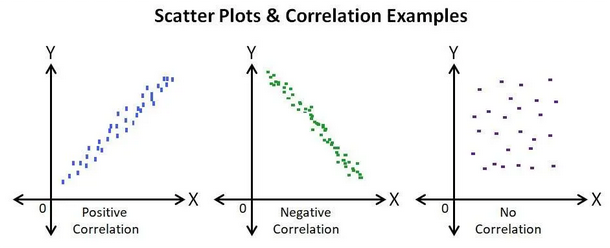
* 1. Pie Chart: is a way of summarizing a set of nominal data or displaying the different values of a given variable

1. Data Clearning (Missing Values) by Pandas
   1. Removing N/A (Counting N/A)
   2. Replacing N/A with 0
   3. Replacing N/A with Mean
2. Data Clearning (Outliners) by Pandas, Scipy
   1. Filtering Data: (same as Excel Advanced Filter)
   2. Detection by Boxplot:

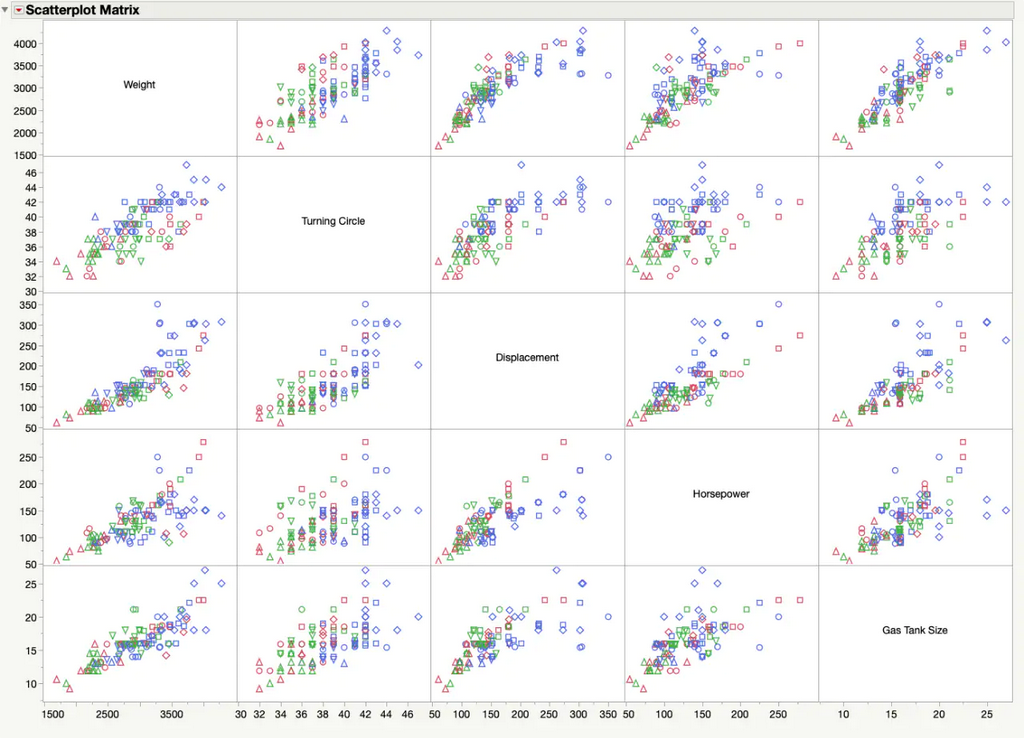
 

* 1. Detection by Z-Score: All the X values are transformed into the Z values. Means become 0 and Standard Deviation becomes 1. Any values which goes beyond 3 Z-Score are termed as an outlier.
  2. Capping: excluding every value beyond x(e.g 99) percentile.
  3. Flooring: excluding every value behind x(e.g 1) percentile.

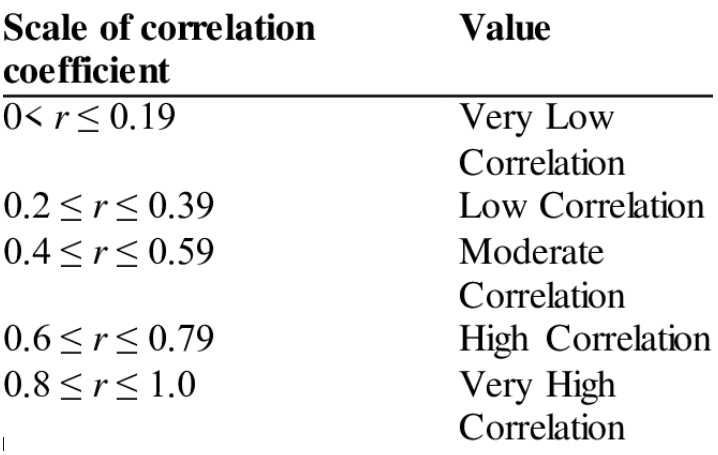
1. Data Clearning (Outliners) by Pandas, Scipy
   1. Scatter Plot: is a graphic tool used to display the relationship between two quantitative variables.

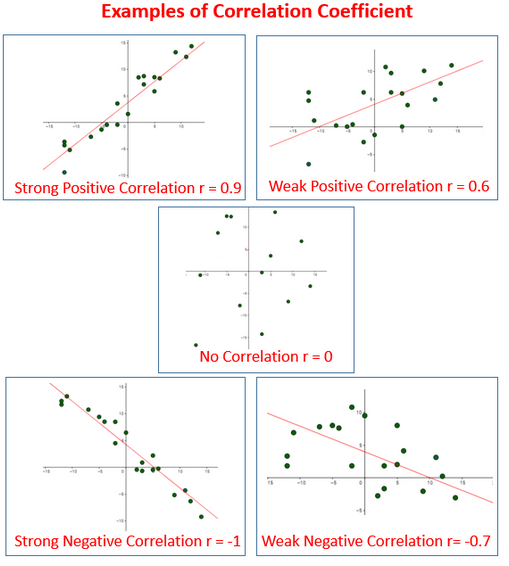


* 1. Scatter Matrix: can show how multiple variables are related. After plotting all the two-way combinations of the variables, the matrix can show relationships between variables to highlight which relationships are likely to be important. The matrix can also identify outliers in multiple scatter plots.



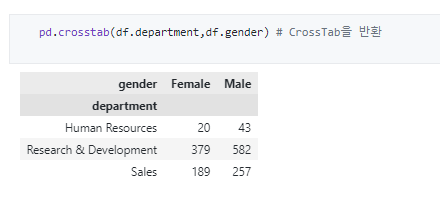
* 1. Covariance(공분산: 공분산은 서로 다른 변수들 사이에 얼마나 의존하는지를 수치적으로 표현하며, 그것의 직관적 의미는 어떤 변수(X)가 평균으로부터 증가 또는 감소라는 경향을 보일 때, 이러한 경향을 다른 변수(Y 또는 Z 등등)가 따라 하는 정도를 수치화 한 것이다.): measures the directional relationship between the returns on two assets. A positive covariance means that asset returns move together while a negative covariance means they move inversely.
  2. Correlation Coefficient: The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movement of the two variables.



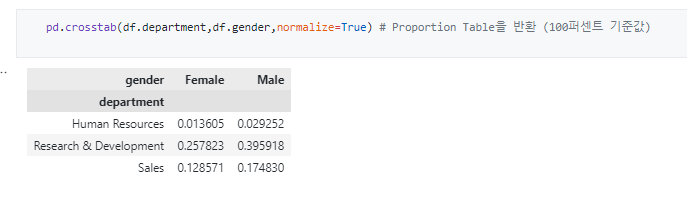


* 1. Correlation Coefficient: The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movement of the two variables.

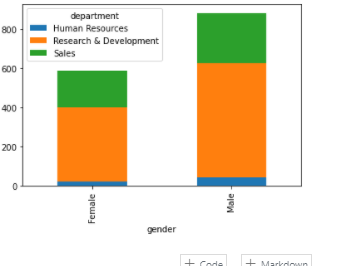
1. Bi-variate Analysis (Categorical and Categorical)
   1. Cross-Tab: is a method to quantitatively analyze the relationship between multiple variables.



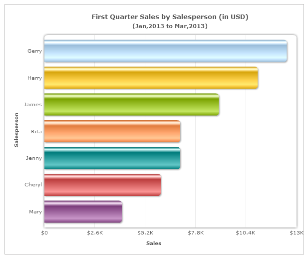
* 1. Proportion Table:



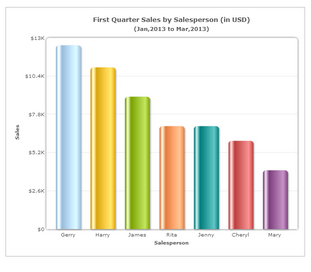
* 1. Slacked Chart: also known as a stacked bar graph, is a graph that is used to break down and compare parts of a whole. Each bar in the chart represents a whole, and segments in the bar represent different parts or categories of that whole. Different colors are used to illustrate the different categories in the bar.



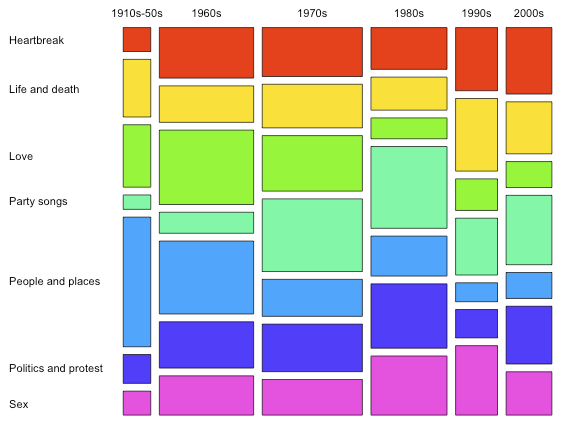
* 1. Bar Chart:



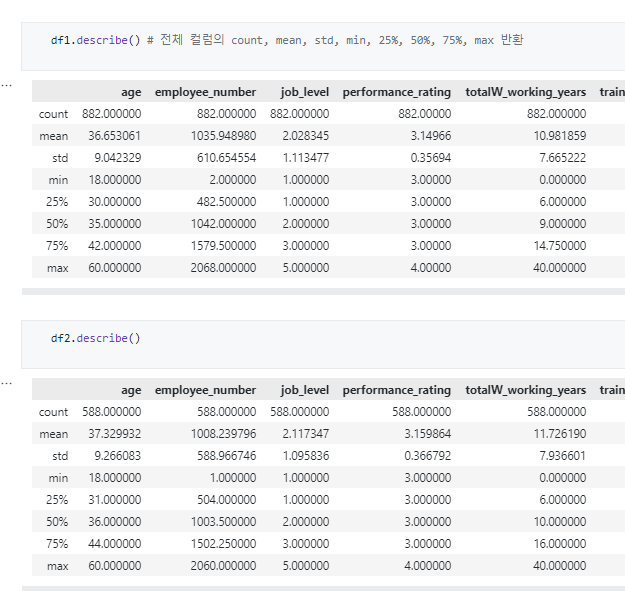
* 1. Column Chart:



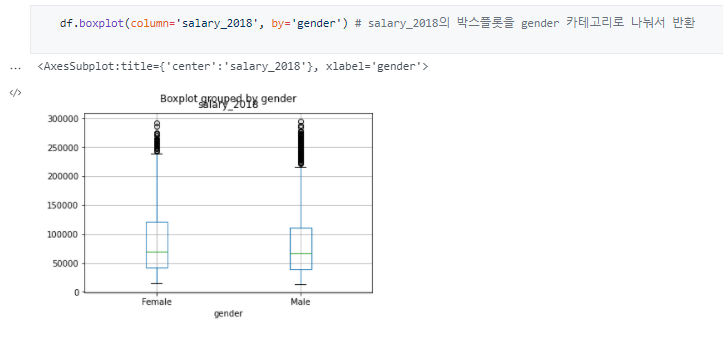
* 1. Mosaic Plot: plot is a special type of stacked bar chart that shows percentages of data in groups.



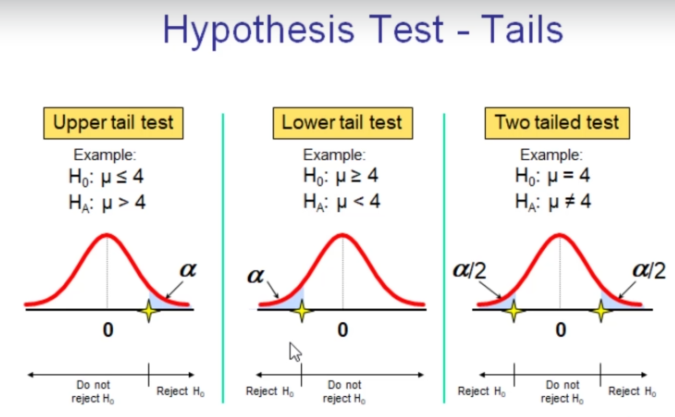
1. Bi-variate Analysis (Numerical and Categorical)
   1. Descriptive Statistics: is a set of brief descriptive coefficients that summarize a given data set representative of an entire or sample population.



* 1. Box Plot



1. Hypothesis Test
   1. One Sample T-Test(1 group, n < 30): is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.



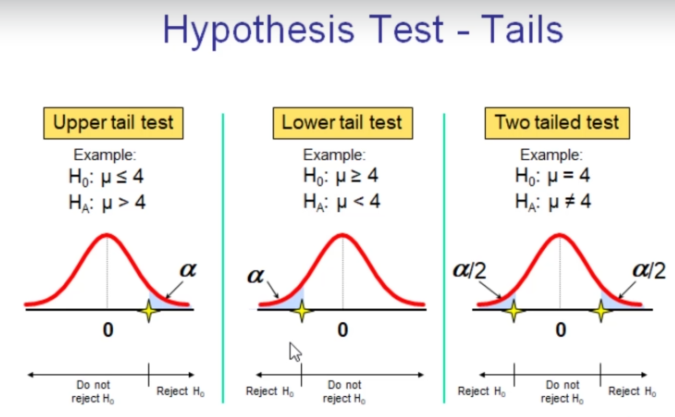
* 1. Paired T-Test (2 groups, paired): A paired t-test is used to compare two population means where you have two samples in which observations in one sample can be paired with observations in the other sample. ... Before-and-after observations on the same subjects
  2. Chi-Square: Chi-Square test of independence uses the following null and alternative hypotheses:

H0: (null hypothesis) The two variables are independent.

H1: (alternative hypothesis) The two variables are not independent. (i.e. they are associated)

* 1. ANOVA (more than 2 groups): ANOVA is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups.

1. Hypothesis Test
   1. One Sample T-Test(1 group, n < 30): is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.



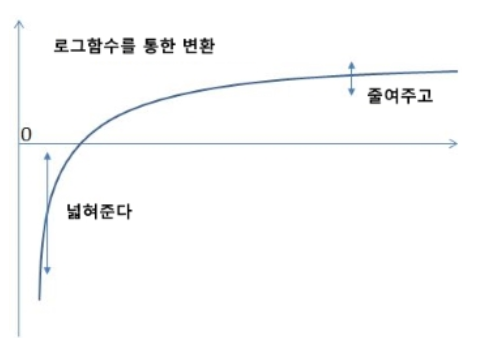
* 1. Paired T-Test (2 groups, paired): A paired t-test is used to compare two population means where you have two samples in which observations in one sample can be paired with observations in the other sample. ... Before-and-after observations on the same subjects
  2. Chi-Square: Chi-Square test of independence uses the following null and alternative hypotheses:

H0: (null hypothesis) The two variables are independent.

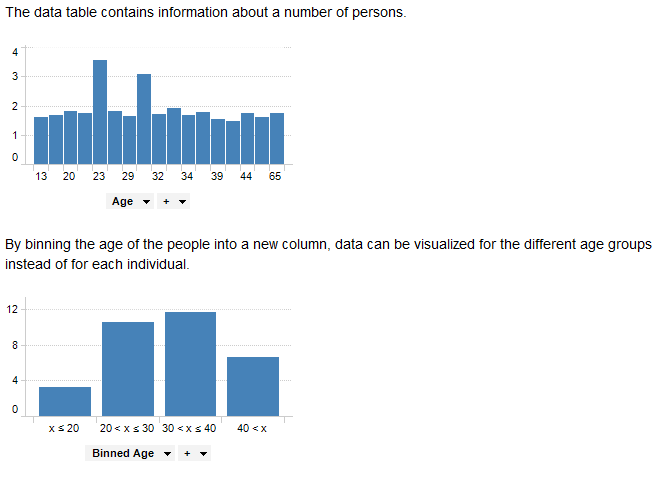
H1: (alternative hypothesis) The two variables are not independent. (i.e. they are associated)

* 1. ANOVA (more than 2 groups): ANOVA is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups.

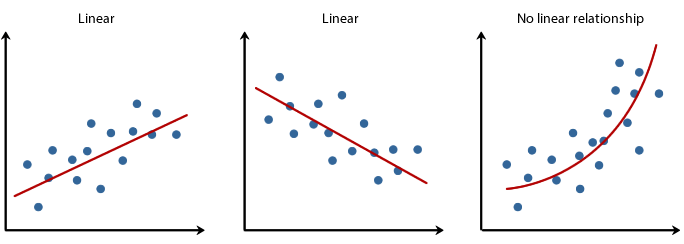
1. Variation Creation
   1. Tenure: Years from the Original Hire Date
   2. Compa Ratio: Salary / Median of the Salary
   3. Dummy Variable: Converting Category into 1 or 0
2. Variable Transformation
   1. Normalization: is a scaling technique method in which data points are shifted and rescaled so that they end up in a range of 0 to 1. It is also known as min-max scaling.
      1. x ′ = ( x − x m i n ) / ( x m a x − x m i n )
   2. Log Transformation: Log를 취해주게 되면 큰 숫자를 같은 비율의 작은 숫자로 만들어주며, 첨도와 왜도가 줄어들면서 정규성이 높아진다. 이는 분석의 정확도를 더 높일 수 있는 결과를 가져온다.



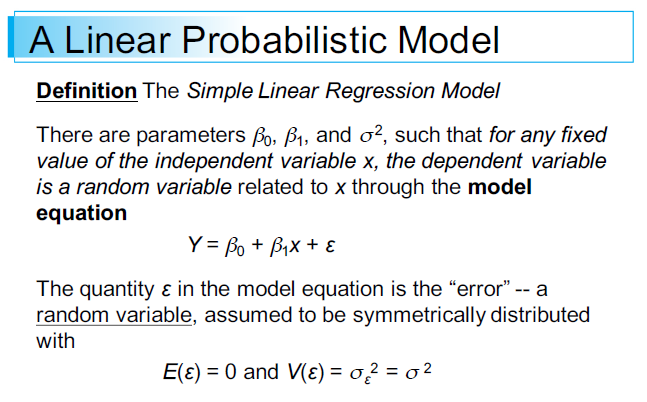
* 1. Square Root Transformation
  2. Cube Root Transformation
  3. Shapiro -Wilk Test: is a test of normality in frequentist statistics
  4. Binning: is a way to group a number of more or less continuous values into a smaller number of "bins". For example, if you have data about a group of people, you might want to arrange their ages into a smaller nmber of age intervals.



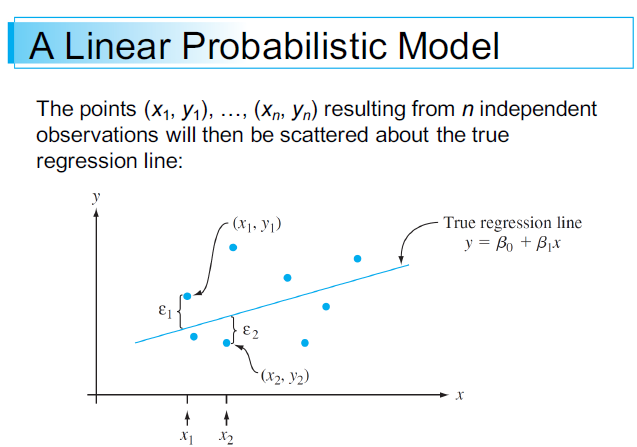
1. Linear Regression
   1. Regression Analysis: is used to predict the value of a dependent variable based on the value of at least one independent variable. And also explains the impact of changes in an independent variable on the dependent variable.
   2. Linear Relationship: is a statistical term used to describe a straight-line relationship between two variables.



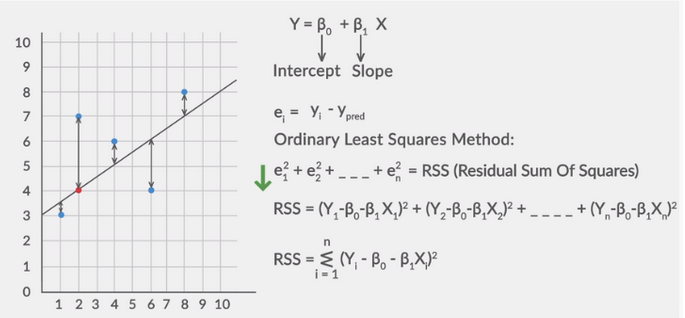
* 1. Simple linear regression: has only one x and one y variable. For instance, when we predict rent based on square feet alone that is simple linear regression.



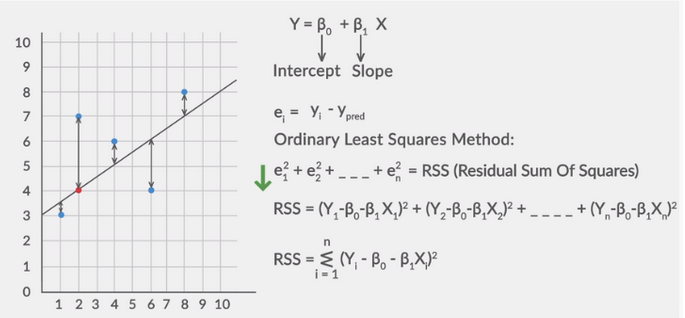
B1 = Slope



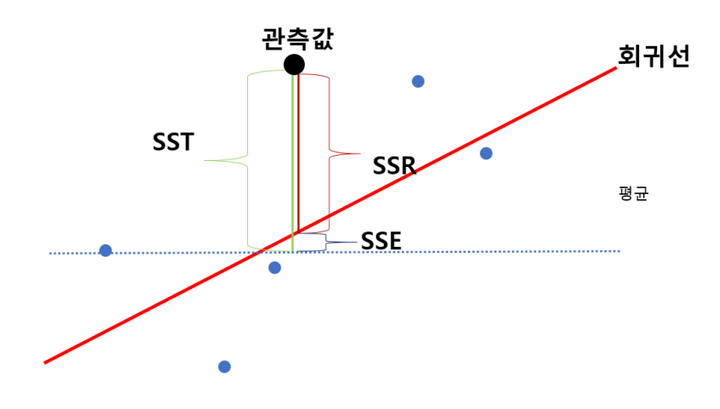
* 1. Multiple linear regression: has one y and two or more x variables. When we predict rent based on square feet and age of the building that is an example of multiple linear regression.
  2. Least-squares Method:

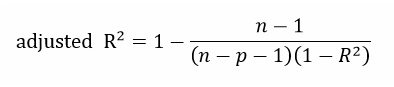


* 1. Least-squares Method:



* 1. R-Square(결정계수 = 피어슨 상관계수의 제곱): is a statistical measure of how close the data are to the fitted regression line. If this value is 45%, independent variable X can explain 45% of dependent variable Y.
  2. Adjusted R-Square: 결정계수는 독립변수 개수가 많아질수록 그 값이 커지게 됩니다. 따라서 종속변수의 변동을 별로 설명해 주지 못하는 변수가 모형에 추가된다고 하더라도 결정계수값이 커질 수 있습니다. 이러한 문제를 보정한 것이 수정된 결정계수(adjusted coefficient of determination)입니다. 표본의 크기와 독립변수의 수를 고려하여 계산하게 되는데 그 식은 아래와 같습니다





* 1. F-Test: is for Overall Significance of the Model and shows if there is a linear relationship between all of the X variables considered together and Y.
     1. H0: B1 = B2 = B3 = … = Bk = 0 (no linear relationship)
     2. Ha: At least one Bi is not equal = 0 (at least one independent variable affects y)
     3. If F-value is less than 0.05, reject Null(H0) and Accept Alternative(Ha).