BHARAT DYNAMICS LTD.
(A Govt. of India Enterprise)

Internship Report

on

Data Science

At BDL Factory

Ministry of Defence, Government of India

By

Shivani Vadla

(16261A05H4)

B.Tech(Computer Science and Engineering)

MAHATMA GANDHI INSTITUTE OF TECHNOLOGY

(Affiliated To Jawaharlal Nehru Technological University, Hyderabad, A.P.) Chaitanya Bharathi P.O., Gandipet, Hyderabad – 500075

Under the Guidance of

K.Sunil Phani(MANAGER-ITD)

MAHATMA GANDHI INSTITUTE OF TECHNOLOGY

(Affiliated To Jawaharlal Nehru Technological University, Hyderabad, A.P.)

Chaitanya Bharathi P.O., Gandipet , Hyderabad – 500075

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

DATE:23-06-2018

This Is To Certify That the Project Work Entitled On Data Science and correlation, Is a Bonafide Work Carried Out By

Shivani Vadla(16261A05H4)

in partial fulfilment of the requirements for the degree of BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING by the Jawaharlal Nehru technological university, Hyderabad during the academic year 2018.

The results embodied in this report have not been submitted to any other university or institution for the award of any degree or diploma.

(Signature)

सुनील फणी/SUNIL PHANI K
प्रबंधक (आई-टी)/MANAGER (ITD)
भारत डायनामिक्स लि BHARAT DYNAMICS LTD.
भानूर संगारेड्डी BHANUR SANGAREDDY-502305

K.SUNIL PHANI(MANAGER-ITD)

ACKNOWLEDGEMENT

I express my deep sense of gratitude to our guide K.Sunil Phani(Manager-ITD), for his valuable guidance and encouragement in carrying out our project.

We are highly indebted to our faculty , who has given us all the necessary technical guidance in carrying out this internship.

We wish to express our sincere thanks our guide head of the IT department(CIM) ,BDL, for permitting us to pursue our project in STUDY OF DATA SCIENCE AND CORRELATION and encouraging us throughout the project.

Finally, we thank all the people who have directly or indirectly helps us through the course of our project.

SHIVANI VADLA

INDEX

1.Data Science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

The popularity of the term “data science” has exploded in business environments and academia, as indicated by jump in job openings.

Data science is a multidisciplinary blend of **data inference, algorithmm development, and technology** in order to solve analytically complex problems.

At the core is data. Troves of raw information, streaming in and stored in enterprise data warehouses. Much to learn by mining it.

Data science – discovery of data insight

This aspect of data science is all about uncovering findings from data. Diving in at a granular level to mine and understand complex behaviors, trends, and inferences. It's about surfacing hidden insight that can help enable companies to make smarter business decisions. For example:

- Netflix data mines movie viewing patterns to understand what drives user interest, and uses that to make decisions on which Netflix original series to produce.
- Target identifies what are major customer segments within it's base and the unique shopping behaviors within those segments, which helps to guide messaging to different market audiences.
- Proctor & Gamble utilizes time series models to more clearly understand future demand, which help plan for production levels more optimally.

How do data scientists mine out insights? It starts with data exploration. When given a challenging question, data scientists become detectives. They investigate leads and try to understand pattern or characteristics within the data. This requires a big dose of analytical creativity.

Then as needed, data scientists may apply quantitative technique in order to get a level deeper – e.g. inferential models, segmentation analysis, time series forecasting, synthetic control experiments, etc. The intent is to scientifically piece together a forensic view of what the data is really saying.

This data-driven insight is central to providing strategic guidance. In this sense, data scientists act as consultants, guiding business stakeholders on how to act on findings.

Data science – development of data product

A "data product" is a technical asset that:

(1) utilizes data as input, and

(2) processes that data to return algorithmically-generated results.

The classic example of a data product is a recommendation engine, which ingests user data, and makes personalized recommendations based on that data. Here are some examples of data products:

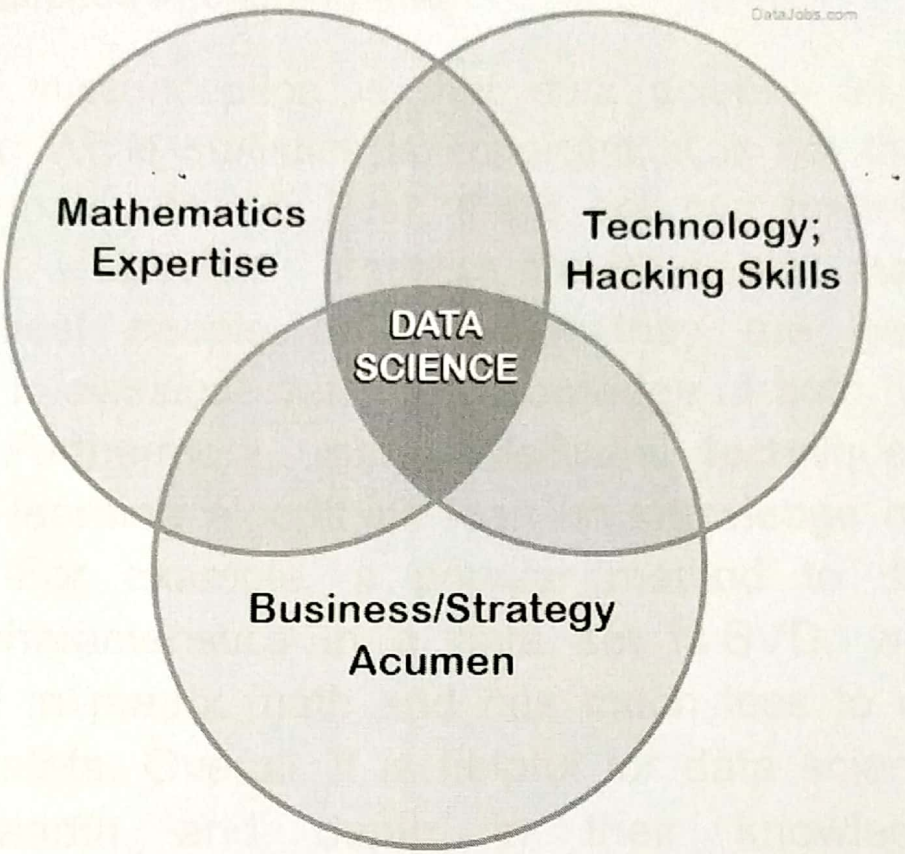
- Amazon's recommendation engines suggest items for you to buy, determined by their algorithms. Netflix recommends movies to you. Spotify recommends music to you.

- Gmail's spam filter is data product – an algorithm behind the scenes processes incoming mail and determines if a message is junk or not.
- Computer vision used for self-driving cars is also data product – machine learning algorithms are able to recognize traffic lights, other cars on the road, pedestrians, etc.

This is different from the "data insights" section above, where the outcome to that is to perhaps provide advice to an executive to make a smarter business decision. In contrast, a data product is technical functionality that encapsulates an algorithm, and is designed to integrate directly into core applications. Respective examples of applications that incorporate data product behind the scenes: Amazon's homepage, Gmail's inbox, and autonomous driving software.

Data scientists play a central role in developing data product. This involves building out algorithms, as well as testing, refinement, and technical deployment into production systems. In this sense, data scientists serve as technical developers, building assets that can be leveraged at wide scale.

What is data science – the requisite skill set
Data science is a blend of skills in three major areas:



Mathematics Expertise:

At the heart of mining data insight and building data product is the ability to view the data through a quantitative lens. There are textures, dimensions, and correlations in data that can be expressed mathematically. Finding solutions utilizing data becomes a brain teaser of heuristics and quantitative technique. Solutions to many business problems involve building analytic models

grounded in the hard math, where being able to understand the underlying mechanics of those models is key to success in building them.

Also, a misconception is that data science all about statistics. While statistics is important, it is not the only type of math utilized. First, there are two branches of statistics – classical statistics and Bayesian statistics. When most people refer to *stats* they are generally referring to *classical stats*, but knowledge of both types is helpful. Furthermore, many inferential techniques and machine learning algorithms lean on knowledge of linear algebra. For example, a popular method to discover hidden characteristics in a data set is SVD, which is grounded in matrix math and has much less to do with classical stats. Overall, it is helpful for data scientists to have breadth and depth in their knowledge of mathematics.

Technology and Hacking

First, let's clarify on that we are *not* talking about hacking as in breaking into computers. We're referring to the tech programmer subculture meaning of hacking – i.e., creativity and ingenuity in using technical skills to build things and find clever solutions to problems.

Why is hacking ability important? Because data scientists utilize *technology* in order to wrangle enormous data sets and work with complex algorithms, and it requires tools far more sophisticated than Excel. Data scientists need to be able to code — prototype quick solutions, as well as integrate with complex data systems. Core languages associated with data science include SQL, Python, R, and SAS. On the periphery are Java, Scala, Julia, and others. But it is not just knowing language fundamentals. A hacker is a technical ninja, able to creatively navigate their way through technical challenges in order to make their code work.

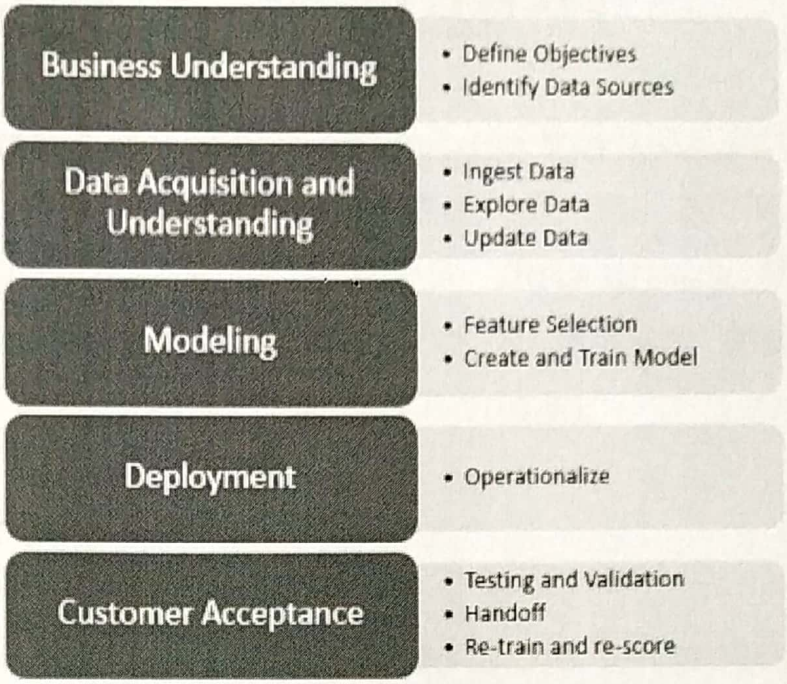
Along these lines, a data science hacker is a solid algorithmic thinker, having the ability to break down messy problems and recompose them in ways that are solvable. This is critical because data scientists operate within a lot of algorithmic complexity. They need to have a strong mental comprehension of high-dimensional data and tricky data control flows. Full clarity on how all the pieces come together to form a cohesive solution.

Strong Business Acumen

It is important for a data scientist to be a tactical business consultant. Working so closely with data, data scientists are positioned to learn from data in ways no one else can. That creates the responsibility to translate observations to shared knowledge, and contribute to strategy on how to solve core business problems. This means a core competency of data science is using data to cogently tell a story. No data-puking – rather, present a cohesive narrative of problem and solution, using data insights as supporting pillars, that lead to guidance.

Having this business acumen is just as important as having acumen for tech and algorithms. There needs to be clear alignment between data science projects and business goals. Ultimately, the value doesn't come from data, math, and tech itself. It comes from leveraging all of the above to build valuable capabilities and have strong business influence.

The Team Data Science Process



Before 20th century ,there was only a limited data. Henceforth data was stored in tables i.e in form of rows and colums. The main drawback of this is, it's mandatory that the data available should always be stored in rows and columns form, which leads to redundancy and always there should be a relationship existing between tables and having key constraints, imposing constraints in data. And also a huge man power is required to enter that data into tables.

And also accuracy of data is questionable. Spreadsheets , floppies and FLAT are used to store data. But after 20th century, large and large amount of data started coming from different IOT's(internet of things). By this data explosion occurred.

2.Data Explosion

The data **explosion** is the rapid increase in the amount of published information or **data** and the effects of this abundance. As the amount of available **data** grows, the problem of managing the information becomes more difficult, which can lead to information overload.

Growth patterns of data now-a-days are:

- The world's technological capacity to store information grew from 2.6 (optimally compressed) exabytes in 1986 to 15.8 in 1993, over 54.5 in 2000, and to 295 (optimally compressed) exabytes in 2007. This is equivalent to less than one 730-MB CD-ROM per person in 1986 (539 MB per person), roughly 4 CD-ROM per person of 1993, 12 CD-ROM per person in the year 2000, and almost 61 CD-ROM per person in 2007. Piling up the imagined 404 billion CD-ROM from 2007 would create a stack from the Earth to the Moon and a quarter of this distance beyond (with 1.2 mm thickness per CD).
- The world's technological capacity to receive information through one-way broadcast networks was 432 exabytes of (optimally compressed) information in 1986, 715 (optimally compressed) exabytes in 1993, 1,200 (optimally compressed) exabytes in 2000, and 1,900 in 2007.

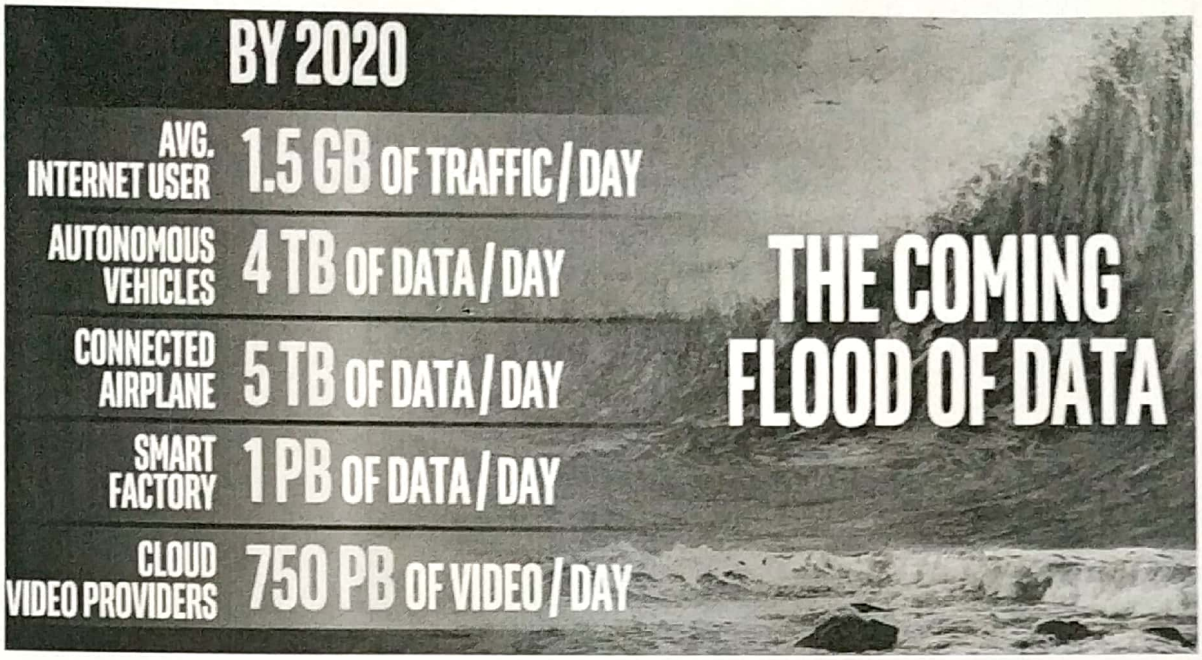
- The world's effective capacity to exchange information through two-way telecommunication networks was 0.281 exabytes of (optimally compressed) information in 1986, 0.471 in 1993, 2.2 in 2000, and 65 (optimally compressed) exabytes in 2007.[9]

According to Latanya Sweeney, there are three trends in data gathering today:

Type 1. Expansion of the number of fields being collected, known as the "collect more" trend.

Type 2. Replace an existing aggregate data collection with a person-specific one, known as the "collect specifically" trend.

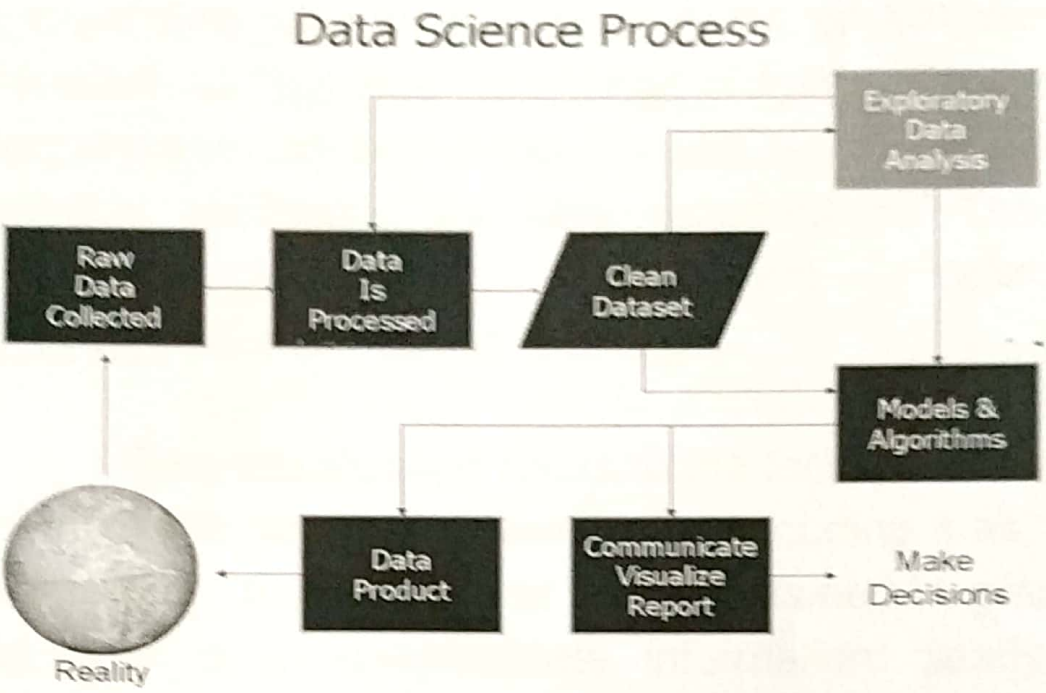
Type 3. Gather information by starting a new person-specific data collection, known as the "collect it if you can" trend.



So to adjust and support the large amount of data and something which is more efficient then DBMS. Now the concept of data science comes into action. Data science possess mainly two characteristics namely:

- Data Visualization
- Data analysis

The below figure depicts the practical usage data visualization and data analysis in practical life .



Data Visualization

Data visualization or **data visualisation** is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data.

To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools.

Data visualization is both an art and a science.[3] It is viewed as a branch of descriptive statistics by some, but also as a grounded theory development tool by others. Increased amounts of data created by Internet activity and

an expanding number of sensors in the environment are referred to as "big data" or Internet of things. Processing, analyzing and communicating this data present ethical and analytical challenges for data visualization.[4] The field of data science and practitioners called data scientists help address this challenge.

Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics. The goal is to communicate information clearly and efficiently to users. It is one of the steps in data analysis or data science.

Graphical displays should (i.e) main features of data visualization are:

- Show the data.
- Induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else.
- Avoid distorting what the data has to say.
- Present many numbers in a small space.
- Make large data sets coherent.
- Encourage the eye to compare different pieces of data.

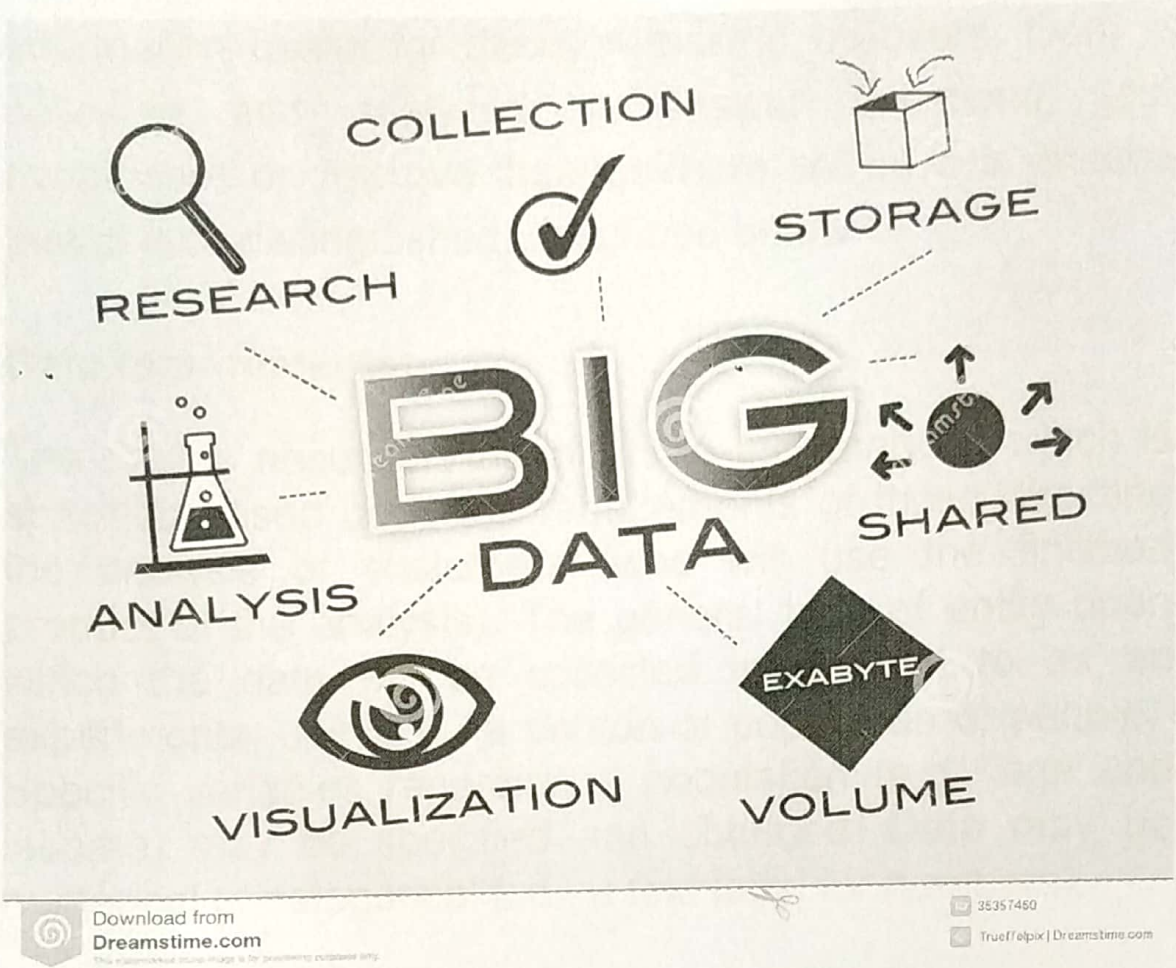
- Reveal the data at several levels of detail, from a broad overview to the fine structure.
- Serve a reasonably clear purpose: description, exploration, tabulation or decoration.
- Be closely integrated with the statistical and verbal descriptions of a data set.

Data Analysis

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

The main aspect of data analysis is data mining.Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes, while business intelligence covers data analysis that relies heavily on aggregation, focusing mainly on business information.[1]

In statistical applications, data analysis can be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data while CDA focuses on confirming or falsifying existing hypotheses. Predictive analytics focuses on application of statistical models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All of the above are varieties of data analysis.



\

Another important concept of data analysis is data integration. Data integration is a precursor to data analysis, and data analysis is closely linked to data visualization and data dissemination. The term *data analysis* is sometimes used as a synonym for data modeling.

Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into

information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theoriesThere are several phases that can be distinguished, described below.

Data requirements:

The data is necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis or customers (who will use the finished product of the analysis). The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people). Specific variables regarding a population (e.g., age and income) may be specified and obtained. Data may be numerical or categorical (i.e., a text label for numbers).

Data collection:

Data is collected from a variety of sources. The requirements may be communicated by analysts to custodians of the data, such as information technology personnel within an organization. The data may also be collected from sensors in the environment, such as traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloads from online sources, or reading documentation.

Data processing:

The phases of the intelligence cycle used to convert raw information into actionable intelligence or knowledge are conceptually similar to the phases in data analysis.

Data initially obtained must be processed or organised for analysis. For instance, these may involve placing data into rows and columns in a table format (i.e., structured data) for further analysis, such as within a spreadsheet or statistical software.

Data cleaning:

Once processed and organised, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that data is entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, overall quality of existing data,[5] deduplication, and column segmentation.

Exploratory data analysis:

Once the data is cleaned, it can be analyzed. Analysts may apply a variety of techniques referred to as exploratory data analysis to begin understanding the messages contained in the data.

Data product:

A data product is a computer application that takes data inputs and generates outputs, feeding them back into the environment. It may be based on a model or algorithm. An example is an application that analyzes data about customer purchasing history and recommends other purchases the customer might enjoy.

Communication:

Once the data is analyzed, it may be reported in many formats to the users of the analysis to support their requirements. The users may have feedback, which results in additional analysis. As such, much of the analytical cycle is iterative.

3.Correlation

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

Correlation works for quantifiable data in which numbers are meaningful, usually quantities of some sort. It cannot be used for purely categorical data, such as gender, brands purchased, or favorite color.

Correlation is a measure of association between two variables. The variables are not designated as dependent or independent. The two most popular correlation coefficients are: Spearman's correlation coefficient rho and Pearson's product-moment correlation coefficient.

Types of Correlation

The value of a correlation coefficient can vary from minus one to plus one.

A minus one indicates a perfect negative correlation, while a plus one indicates a perfect positive correlation.

A correlation of zero means there is no relationship between the two variables.

- When there is a negative correlation between two variables, as the value of one variable increases, the value of the other variable decreases, and vise versa. In other words, for a negative correlation, the variables work opposite each other.

For example(using python program):

What happens to our correlation figure if we invert the correlation such that an increase in x results in a decrease in y?

In [3]:

```
# 1000 random integers between 0 and 50
x = np.random.randint(0, 50, 1000)


# Negative Correlation with some noise
y = 100 - x + np.random.normal(0, 5, 1000)


np.corrcoef(x, y)
```

Out[3]:

```
array([[ 1.        , -0.94957116],
       [-0.94957116,  1.        ]])
```

Our correlation is now negative and close to 1.

- When there is a positive correlation between two variables, as the value of one variable increases, the value of the other variable also increases. The variables move together.

For example(using python program):

Let's take a look at a positive correlation. Numpy implements a `corrcoef()` function that returns a matrix of correlations of x with x, x with y, y with x and y with y. We're interested in the values of correlation of x with y (so position (1, 0) or (0, 1)).

In [1]:

```
import numpy as np


np.random.seed(1)


# 1000 random integers between 0 and 50
x = np.random.randint(0, 50, 1000)


# Positive Correlation with some noise
y = x + np.random.normal(0, 10, 1000)
```

```
np.corrcoef(x, y)
```

Out[1]:

```
array([[ 1.        ,  0.81543901],
       [ 0.81543901,  1.        ]])
```

This correlation is 0.815, a strong positive correlation.

The standard error of a correlation coefficient is used to determine the confidence intervals around a true correlation of zero. If your correlation coefficient falls outside of this range, then it is significantly different than zero. The standard error can be calculated for interval or ratio-type data (i.e., only for Pearson's product-moment correlation).

The significance (probability) of the correlation coefficient is determined from the t-statistic. The probability of the t-statistic indicates whether the observed correlation coefficient occurred by chance if the true correlation is zero. In other words, it asks if the correlation is significantly different than zero. When the t-statistic is calculated for Spearman's rank-difference correlation coefficient, there must be at least 30 cases before the t-distribution can be used to determine the probability. If there are fewer than 30 cases, you must refer to a special table to find the probability of the correlation coefficient.

Limitations of Correlation Analysis

The correlation analysis has certain limitations:

- Two variables can have a strong non-linear relation and still have a very low correlation.
- Recall that correlation is a measure of the linear relationship between two variables. The correlation can be unreliable when outliers are present.
- The correlation may be spurious. Spurious correlation refers to the following situations:
- The correlation between two variables that reflects chance relationships in a particular data set. The correlation induced by a calculation that mixes each of two variables with a third variable

Uses of Correlation Analysis

The uses of correlation analysis are highlighted through six examples in the curriculum. Instead of reproducing the examples, the specific scenarios where they are used are listed below:

- Evaluating economic forecasts: Inflation is often predicted using the change in the consumer price index (CPI). By plotting actual vs predicted inflation,

analysts can determine the accuracy of their inflation forecasts

- Style analysis correlation: Correlation analysis is used in determining the appropriate benchmark to evaluate a portfolio manager's performance. For example, assume the portfolio managed consists of 200 small value stocks. The Russell 2000 Value Index and the Russell 2000 Growth Index are commonly used as benchmarks to measure the smallcap value and small-cap growth equity segments, respectively. If there is a high correlation between the returns to the two indexes, then it may be difficult to distinguish between small-cap growth and small-cap value as different styles.
- Exchange rate correlations: Correlation analysis is also used to understand the correlations among many asset returns. This helps in asset allocation, hedging strategy and diversification of the portfolio to reduce risk. Historical correlations are used to set expectations of future correlation. For example, suppose an investor who has an exposure to foreign currencies. He needs to ascertain whether to increase his exposure to the Canadian dollar or to Japanese Yen. By analyzing the historical correlations between

USD returns to holding the Canadian dollar and USD returns to holding the Japanese yen, he will be able to come to a conclusion. If they are not correlated, then holding both the assets helps in reducing risk.

- Correlations among stock return series: Analyzing the correlations among the stock market indexes such as large-cap, small-cap and mid-cap helps in asset allocation and diversifying risk. For instance, if there is a high correlation between the returns to the large-cap index and the small-cap index, then their combined allocation may be reduced to diversify risk.
- Correlations of debt and equity returns: Similarly, the correlation among different asset classes, such as equity and debt, is used in portfolio diversification and asset allocation. For example, high-yield corporate bonds may have a high correlation to equity returns, whereas long-term government bonds may have a low correlation to equity returns
- Correlations among net income, cash flow from operations, and free cash flow to the firm: Correlation analysis shows if an analyst's decision to value a firm based only on NI and ignore CFO and FCFF is correct. FCFF is the cash flow available to debt holders and shareholders after all operating expenses

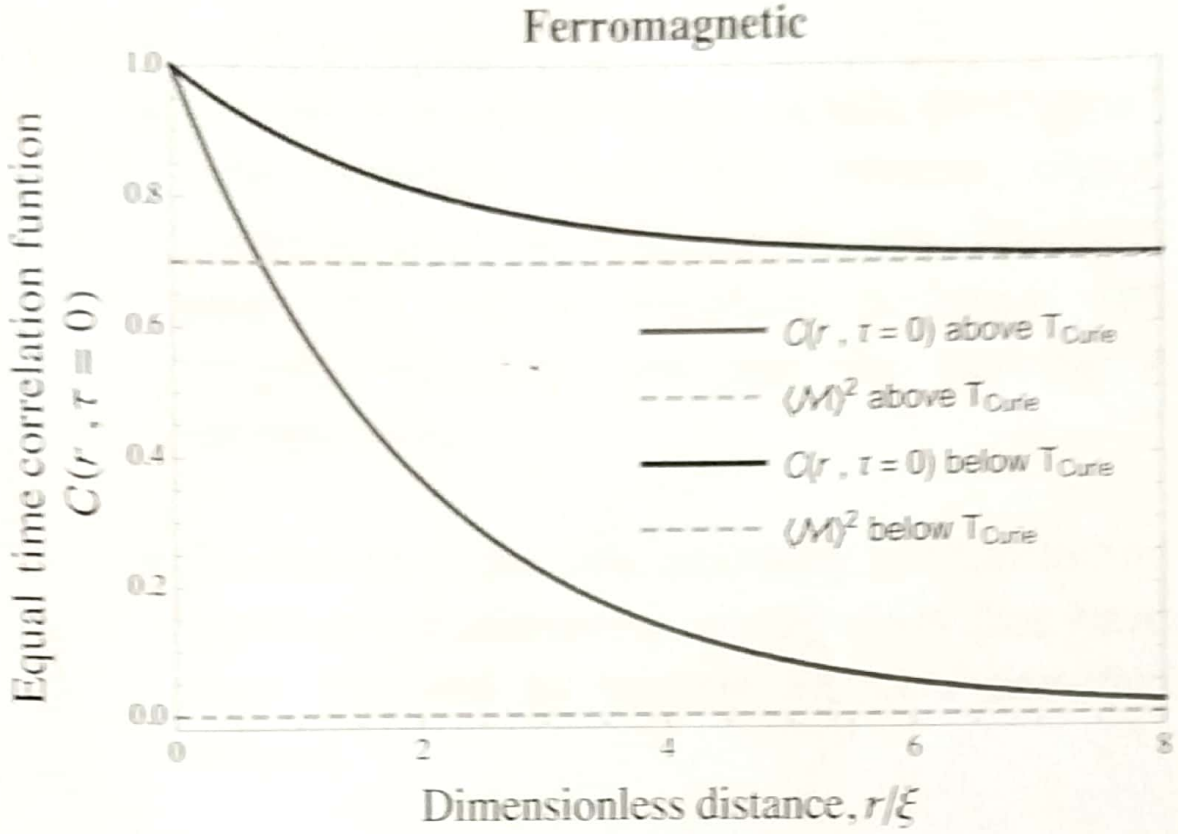
have been paid and investments in working and fixed capital have been made. If there is a low correlation between NI and FCFF, then the analyst's decision to use NI instead of FCFF/CFO to value a company is questionable.

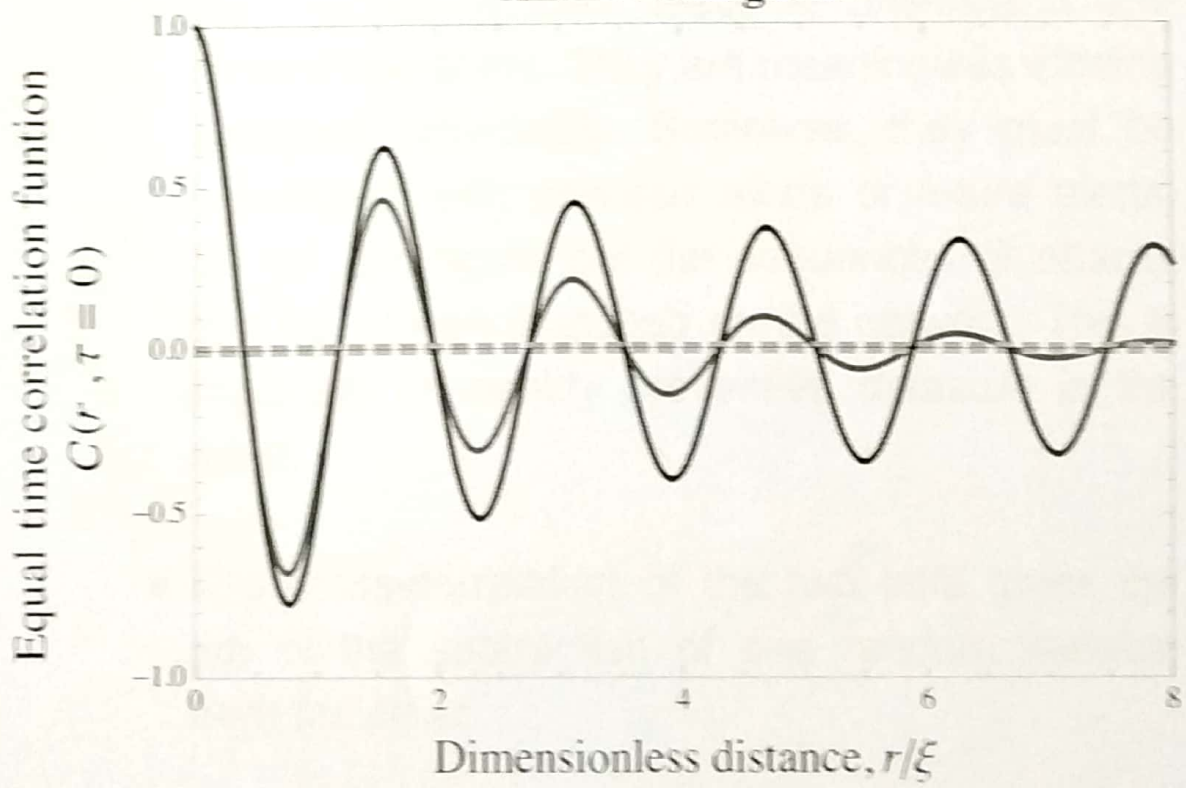
Correlation in Data Science:

Correlation measure how two observed variables are related to each other . It has been used in many different ways in data science.

1. Correlation is used im univariate analysis to identify which feature is more predictive for classification of regression task.
2. To identify multicollinearity in the feature set . Multicollinearity reduse the accuracy of model.
3. Identify casual relationship between variables.
4. Their are many other extension like cca (canonical correlation analysis).CCA is used to identify and measure the associations among two sets of variables.

The below figure shows us the correlation function in two modes (i.e) ferro and anti-ferro.





Application of Correlation:

- Correlation is used to extract second (and higher) order statistics from any random signal, (de)convolution is inherent in any operation where the Fourier transform is taken with irregular sampling: one real life example = interferometry

- Correlation is the only promising method to find significant relationships among alerts that have been triggered by multiple intrusion detection sensors.

- Security Admin (SA) needs to understand and study these alerts. They are meaningless if being analyzed individually. Somehow, they must be 'connected' with previous alerts or future alerts. So SA can figure out the sequences of attacks that have been launched on the network. This is important to identify preventive measure in the future.

- The cross-correlation of the two pdfs gives the pdf of the subtraction of one random variable from the other.

- Correlation is how you would obtain a new velocity distribution from a new position distribution and an old position distribution.

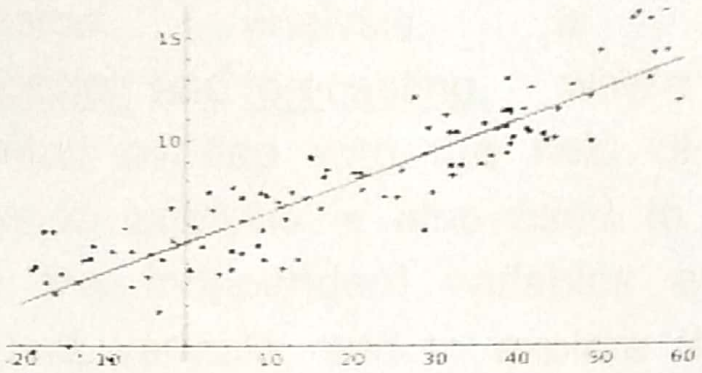
- The correlation is used to appreciate the similarity between a signal and its traslated variants,

- Correlation is applicable whenever you are interested in finding statistical relationships between quantitative variebles and later predict one of them from another.
- Correlation of pseudorandom binary codes is what makes GPS work, lots of radar systems, and lots CDMA (code division multiple access) systems. It is why GPS is power hungry, it takes a lot of processing to find the correct correlation, with correct satellite (correct code) and the correct delay.

4.REGRESSION

The basic idea of Regression is predicting the past based on the selected future.(i.e) for example :one first sets a goal and then decides what should he do to achieve.

In statistical modeling, **regression analysis** is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.



Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, a function of the independent variables called the **regression function** is to be estimated. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the prediction of the regression function using a probability distribution. A related but distinct approach is Necessary Condition Analysis[1] (NCA), which estimates the maximum (rather than average) value of the dependent variable for a given value of the independent variable (ceiling line rather than central line) in order to identify what value of the independent variable is necessary but not sufficient for a given value of the dependent variable.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable;[2] for example, correlation does not prove causation.

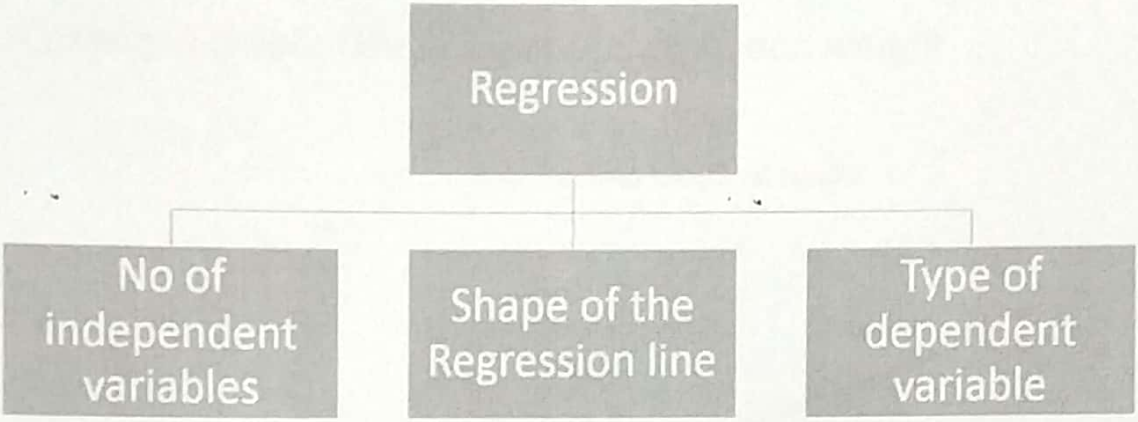
There are multiple benefits of using regression analysis.

They are as follows:

1. It indicates the **significant relationships** between dependent variable and independent variable.

2. It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Types of Regerssion



There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line).

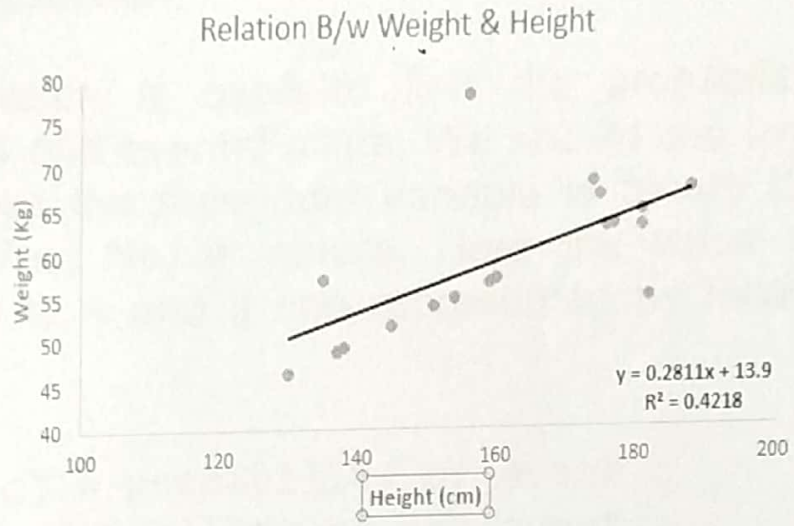
1. Linear Regression

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

It is represented by an equation **Y=a+b*X + e**, where a is intercept, b is slope of the line and e is error term.

For example graph: Graph between height and weight



Important Points:

- There must be **linear relationship** between independent and dependent variables
- Multiple regression suffers from **multicollinearity, autocorrelation, heteroskedasticity**.
- Linear Regression is very sensitive to **Outliers**. It can terribly affect the regression line and eventually the forecasted values.
- Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable

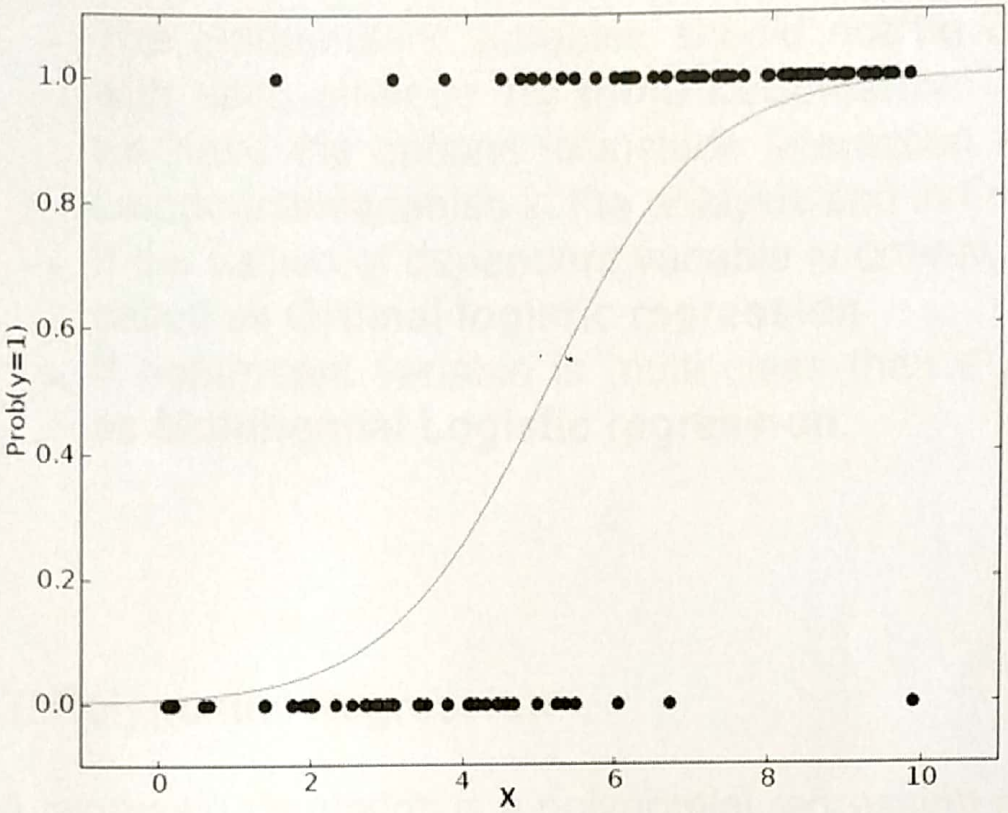
- In case of multiple independent variables, we can go with **forward selection**, **backward elimination** and **step wise approach** for selection of most significant independent variables.

2. Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

```
odds= p/ (1-p) = probability of event
occurrence / probability of not event
occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) =
b0+b1X1+b2X2+b3X3....+bkXk
```

Above, p is the probability of presence of the characteristic of interest.



Important Points:

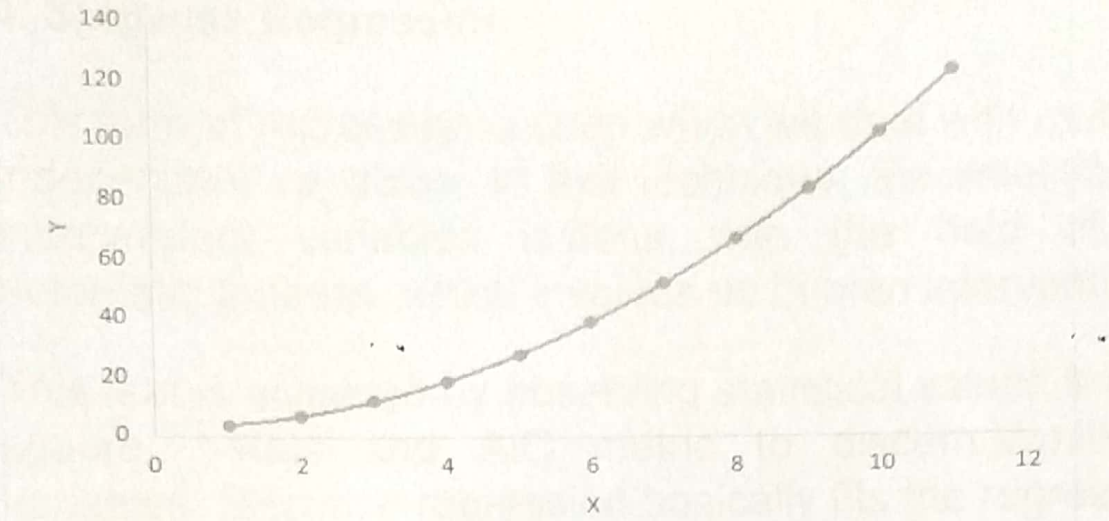
- It is widely used for **classification problems**
- Logistic regression doesn't require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio
- To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression
- It requires **large sample sizes** because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square

- The independent variables should not be correlated with each other i.e. **no multi collinearity**. However, we have the options to include interaction effects of categorical variables in the analysis and in the model.
- If the values of dependent variable is ordinal, then it is called as **Ordinal logistic regression**
- If dependent variable is multi class then it is known as **Multinomial Logistic regression**.

3. Polynomial Regression

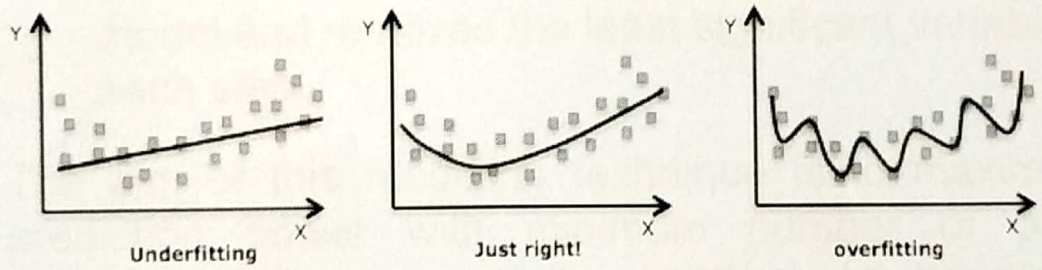
A regression equation is a polynomial regression equation if the power of independent variable is more than 1.

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.



Important Points:

- While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in over-fitting. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem. Here is an example of how plotting can help:



- Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher polynomials can end up producing wierd results on extrapolation.

4. Stepwise Regression

This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves *no* human intervention.

This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by adding/dropping co-variates one at a time based on a specified criterion. Some of the most commonly used Stepwise regression methods are listed below:

- Standard stepwise regression does two things. It adds and removes predictors as needed for each step.
- Forward selection starts with most significant predictor in the model and adds variable for each step.
- Backward elimination starts with all predictors in the model and removes the least significant variable for each step.

The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the method to handle higher dimensionality of data set.

5. Ridge Regression

Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Important Points:

- The assumptions of this regression is same as least squared regression except normality is not to be assumed
- It shrinks the value of coefficients but doesn't reaches zero, which suggests no feature selection feature
- This is a regularization method and uses l2 regularization.

6. Lasso Regression

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the

accuracy of linear regression models. Look at the

$$= \underset{\beta \in R^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

equation below:

Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. This results to variable selection out of given n variables.

Important Points:

- The assumptions of this regression is same as least squared regression except normality is not to be assumed
- It shrinks coefficients to zero (exactly zero), which certainly helps in feature selection
- This is a regularization method and uses l1 regularization
- If group of predictors are highly correlated, lasso picks only one of them and shrinks the others to zero.

7. ElasticNet Regression

ElasticNet is hybrid of Lasso and Ridge Regression techniques. It is trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

A practical advantage of trading-off between Lasso and Ridge is that, it allows Elastic-Net to inherit some of Ridge's stability under rotation.

Important Points:

- It encourages group effect in case of highly correlated variables
- There are no limitations on the number of selected variables
- It can suffer with double shrinkage.

How to select the right regression model?

Within multiple types of regression models, it is important to choose the best suited technique based on type of independent and dependent variables, dimensionality in the data and other essential characteristics of the data. Below are the key factors that you should practice to select the right regression model:

1. Data exploration is an inevitable part of building predictive model. It should be you first step before selecting the right model like identify the relationship and impact of variables
2. To compare the goodness of fit for different models, we can analyse different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term. Another one is the Mallow's Cp criterion. This essentially checks for possible bias in your model, by comparing the model with all possible submodels (or a careful selection of them).
3. Cross-validation is the best way to evaluate models used for prediction. Here you divide your data set into two group (train and validate). A simple mean squared difference between the observed and predicted values give you a measure for the prediction accuracy.
4. If your data set has multiple confounding variables, you should not choose automatic model selection method because you do not want to put these in a model at the same time.
5. It'll also depend on your objective. It can occur that a less powerful model is easy to implement as compared to a highly statistically significant model.
6. Regression regularization methods(Lasso, Ridge and ElasticNet) works well in case of high dimensionality and multicollinearity among the variables in the data set.

For example(using python program):

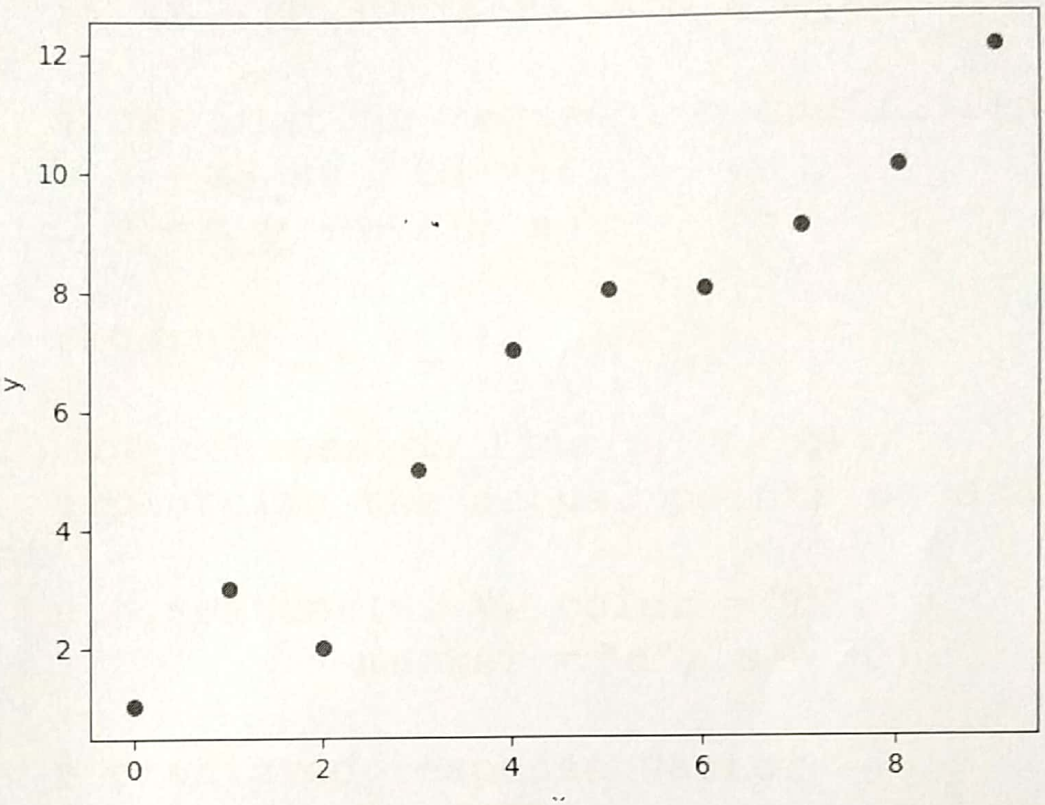
Let us consider a dataset where we have a value of response y for every feature x:

x	0	1	2	3	4	5	6	7	8	9
y	1	3	2	5	7	8	8	9	10	12

For generality, we define:

x as **feature vector**, i.e x = [x_1, x_2, ..., x_n],
y as **response vector**, i.e y = [y_1, y_2, ..., y_n]

The graphical representation of the above problem is:



Program:

```
import numpy as np
import matplotlib.pyplot as plt

def estimate_coef(x, y):
    # number of observations/points
    n = np.size(x)

    # mean of x and y vector
    m_x, m_y = np.mean(x), np.mean(y)

    # calculating cross-deviation and
```

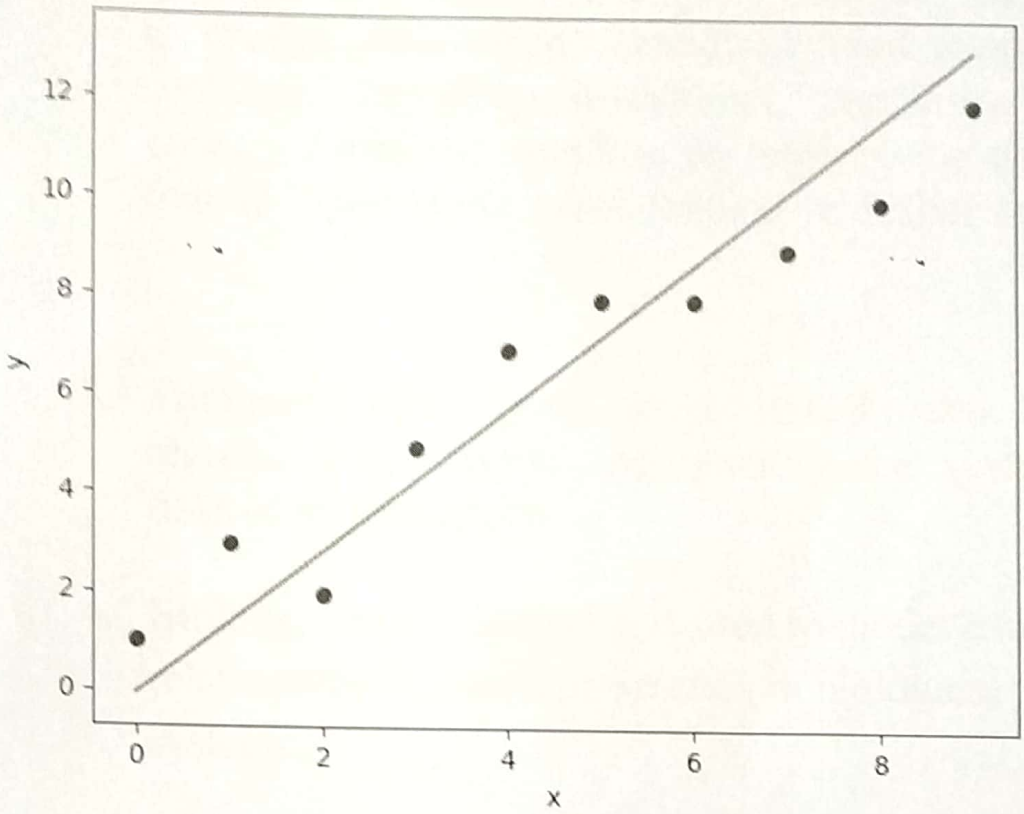
```
deviation about x
    SS_xy = np.sum(y*x - n*m_y*m_x)
    SS_xx = np.sum(x*x - n*m_x*m_x)

    # calculating regression coefficients
    b_1 = SS_xy / SS_xx
    b_0 = m_y - b_1*m_x

    return(b_0, b_1)

def plot_regression_line(x, y, b):
    # plotting the actual points as scatter
plot
    plt.scatter(x, y, color = "m",
                marker = "o", s = 30)

    # predicted response vector
    y_pred = b[0] + b[1]*x

    # plotting the regression line
    plt.plot(x, y_pred, color = "g")

    # putting labels
    plt.xlabel('x')
    plt.ylabel('y')

    # function to show plot
    plt.show()

def main():
    # observations
```

```
    x = np.array([0, 1, 2, 3, 4, 5, 6, 7, 8,
9])
    y = np.array([1, 3, 2, 5, 7, 8, 8, 9, 10,
12])

    # estimating coefficients
    b = estimate_coef(x, y)
    print("Estimated coefficients:\nb_0 = {}  \
          \nb_1 = {}".format(b[0], b[1]))

    # plotting regression line
    plot_regression_line(x, y, b)

if __name__ == "__main__":
    main()
```

Run on IDE
Output of above piece of code is:

```
Estimated coefficients:
b_0 = -0.0586206896552
b_1 = 1.45747126437
```

And graph obtained looks like this:



Applications of regression (mainly linear regression):

1. **Trend lines:** A trend line represents the variation in some quantitative data with passage of time (like GDP, oil prices, etc.). These trends usually follow a linear relationship. Hence, linear regression can be applied to predict future values. However, this method suffers from a lack of scientific validity in cases where other potential changes can affect the data.

2. **Economics:** Linear regression is the predominant empirical tool in economics. For example, it is used to predict consumption spending, fixed investment spending, inventory investment, purchases of a country's exports, spending on imports, the demand to hold liquid assets, labor demand, and labor supply.

3. **Finance:** Capital price asset model uses linear regression to analyze and quantify the systematic risks of an investment.

4. **Biology:** Linear regression is used to model causal relationships between parameters in biological systems.