**CUMULATIVE CALL DETAIL**

**ANALYSIS REPORT**

**Bachelor of Technology**

**In**

**Computer Science & Engineering**

**Under the supervision of**

***Manoj Kumar Tripathy***

**By**

*NEHA HOODA* (*60015002714*)

**To**



**Maharaja Surajmal Institute of Technology**

Affiliated to Guru Gobind Singh Indraprastha University C-4,

Janakpuri, New Delhi-58

i

**CERTIFICATE**

**This is to certify that the project work done on *“Cumulative Call Detail Analysis Report*” submitted to Maharaja Surajmal Institute of Technology, Janakpuri Delhi by “NEHA HOODA” In partial fulfillment of the requirement for the award of degree of Bachelor of Technology, is a bonafide work carried out by him/her under my supervision and guidance. This project work is the original one and has not submitted anywhere else for any other degree.**

*Manoj Kumar Tripathy Koyel Datta Gupta*

**Delivery Head-IMS (HOD, CSE MSIT)**

ii

**ACKNOWLEDGEMENT**

Team effort together with precious words of encouragement and guidance makes daunting tasks achievable. It is a pleasure to acknowledge the direct and implied help I received at various stages in the task of developing the report. It would not have been possible to develop such a report without the furtherance on part of numerous individuals. I find it impossible to express my thanks to each one of them in words, for it seems too trivial when compare to the profound encouragement that they extended to me.

I am grateful to Mr. Manoj Kumar Tripathy, Delivery Head, IMS, NIIT Technologies, and New Delhi for having given me opportunity to do this report, which was of great interest to me.

It gives me immense pleasure to express my heartfelt thanks to Mr. **Vinay Sachdeva**, Senior System Engineer and **Mr. Virendra Singh**, Senior Associate - Data Center Services for guiding me during the report.

My sincere thanks to **Ms. Tanya Pradhan,** PMO - IMS, NIIT Technologies, New Delhi **& Mr. Tapan Biswas**,System Analyst, NIIT Consultant for believing in me and providing motivation all through. Without their guidance this project would not be such a success.

At last I thank the almighty, who had given the strength to complete this report on time.

Finally I would like to thank my parents, all friends, and well-wishers for their valuable help and encouragement throughout the project and **Mr. Vivek Aggarwal**, Program Manager, MS and **Mr. DP Goswami**, Project Lead IMS, NIIT Technologies, New Delhi for their constant support during the project.

*Neha Hooda (Signature)*

*60015002714*

iii

**ABSTRACT**

Programmable Infrastructure and Continuous Deployment is the way forward. With businesses wanting fail-fast capabilities and the business environment becoming ever more unpredictable, the IT build function needs to be more agile and nimble than ever. This is not merely limited to the software wing of build, but also includes the infrastructure arm. Businesses today need to change rapidly in response to the changing market conditions.

At NIIT Technologies, we deliver end-to-end Infrastructure Management Services (IMS) aligned to our service vision of “*Continuous Delivery”* for today’s volatile business environment. We deploy our proprietary delivery framework, CARE, which empowers enterprises to Change-and-Run with newer technologies while we optimize their IT infrastructure.

Our IMS offerings include a mix of traditional and digital services across key verticals such as Travel and Transportation, Insurance, Banking and Financial Services, Media, and Manufacturing. Our capabilities, especially in DevOps and Cloud retrofitting solutions, enable Agile Infrastructure for the customers.

Our service lines are as follows:

* Data Center as a Service
* Workplace as a Service
* Infrastructure Operations Management Services
* DevOps and Automation Services
* Landscape Management Services
* Cloud Services

We have an ecosystem of partners that enables us to develop new capabilities and allow us to invest in our customers to support their Change-and-Run organizations. NIIT Technologies’ partner ecosystem includes the likes of AppDynamics, Dynatrace, AWS, Azure, CA Technologies, Micro Focus-Suse, VistaraIT, and BMC Remedy.

iv

**TABLE OF CONTENTS**

Title i

Certificate ii

Acknowledgement iii

Abstract iv

Chapter 1: Introduction 1

Chapter 2: Importance of Data Analysis 2

Chapter 3: Types of Data Analysis

3. A

3. B

3. C 4

Chapter 4: Process of Data Analysis 9

Chapter 5: Techniques used 39

5. A: Bar Chart 41

5. B:

5. C: 46

5. D:

5. E

Chapter 6: Final Report 51

Chapter 7: Conclusion 53

Appendix A 55

Appendix B 56

References 57

Glossary 58

**CUMULATIVE CALL DETAIL ANALYSIS REPORT**



**CHAPTER 1: INTRODUCTION**

Every business counts on collected sales, sales, customer and retail data to understand its stand in the present scenario. But too much of raw data is difficult to manage and harder to analyze properly. Thus, it’s always smarter to come up with a data analysis report so that all the data can get a structured form that further helps in convenient understanding of the situation.

At NIIT Technologies, they have been providing Infrastructure maintenance and support services to some very large clients. They are providing a wide range of services through a dedicated and specialized team of IT consultants , system administrators and support specialists ensuring that they not only meet but exceed the agreed upon SLA terms.

1

**CHAPTER 2: IMPORTANCE OF DATA ANALYSIS**

The purpose of analysing data is to obtain usable and useful information. The analysis, irrespective of whether the data is qualitative or quantitative, may:

• describe and summarise the data

• identify relationships between variables

• compare variables

• identify the difference between variables

• forecast outcomes

SCALES OF MEASUREMENT

Many people are confused about what type of analysis to use on a set of data and the relevant forms of pictorial presentation or data display. The decision is based on the scale of measurement of the data. These scales are nominal, ordinal and numerical.

Nominal scale

A nominal scale is where: the data can be classified into a non-numerical or named categories, and the order in which these categories can be written or asked is arbitrary.

Ordinal scale

An ordinal scale is where: the data can be classified into non-numerical or named categories an inherent order exists among the response categories. Ordinal scales are seen in questions that call for ratings of quality (for example, very good, good, fair, poor, and very poor) and agreement (for example, strongly agree, agree, disagree, and strongly disagree).

Numerical scale

A numerical scale is: where numbers represent the possible response categories there is a natural ranking of the categories zero on the scale has meaning there is a quantifiable difference within categories and between consecutive categories.

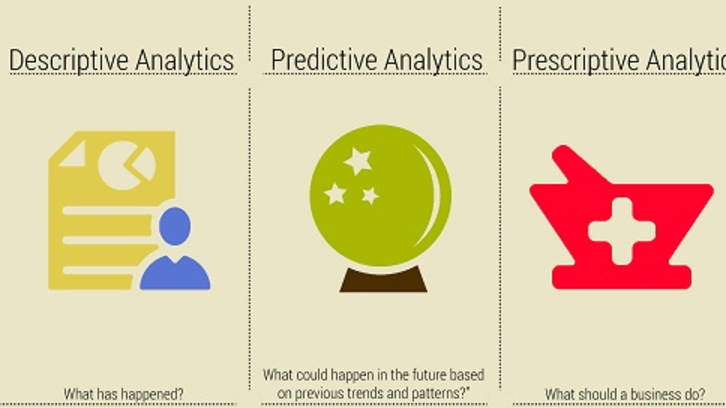
2

There are many benefits of data analysis however; the most important ones are as follows: - data analysis helps in structuring the findings from different sources of data collection like survey research. It is again very helpful in breaking a macro problem into micro parts. Data analysis acts like a filter when it comes to acquiring meaningful insights out of huge data-set. Every researcher has sort out huge pile of data that he/she has collected, before reaching to a conclusion of the research question. Mere data collection is of no use to the researcher. Data analysis proves to be crucial in this process. It provides a meaningful base to critical decisions. It helps to create a complete dissertation proposal.

One of the most important uses of data analysis is that it helps in keeping human bias away from research conclusion with the help of proper statistical treatment. With the help of data analysis a researcher can filter both qualitative and quantitative data for an assignment writing projects. Thus, it can be said that data analysis is of utmost importance for both the research and the researcher. Or to put it in another words data analysis is as important to a researcher as it is important for a doctor to diagnose the problem of the patient before giving him any treatment.

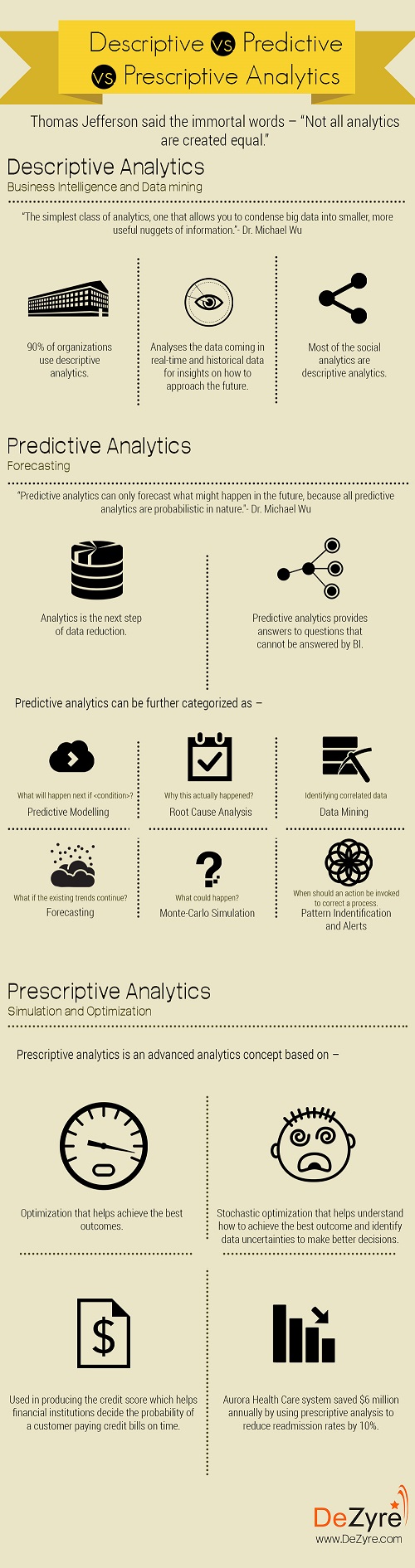
3

**CHAPTER 3: TYPES OF DATA ANALYSIS**

**** 4

In NIIT Technologies, one get a highly skilled and experienced team that can provide comprehensive, reliable and cost effective services. So one can concentrate on their business, while the qualified experts help you run your application systems and infrastructure.

Data analytics helps a business understand the requirements and preferences of a customer, so that businesses can increase their customer base and retain the existing ones with personalized and relevant offerings of their products or services. According to IDC, the big data and analytics industry is anticipated to grow at a CAGR of 26.4% reaching a value of $41.5 billion by end of 2018. The big data industry is growing at a rapid pace due to various applications like smart power grid management, sentiment analysis, fraud detection, personalized offerings, traffic management, etc. across myriad industries. After the organizations collect big data, the next important step is to get started with analytics. Many organizations do not know where to begin, what kind of analytics can nurture business growth and what these different types of analytics mean.



What is Descriptive Analytics?

90% of organizations today use descriptive analytics which is the most basic form of analytics. The simplest way to define descriptive analytics is that, it answers the question “What has happened?” This type of analytics, analyses the data coming in real-time and historical data for insights on how to approach the future. The main objective of descriptive analytics is to find out the reasons behind precious success or failure in the past. The ‘Past’ here, refers to any particular time in which an event had occurred and this could be a month ago or even just a minute ago. The vast majority of big data analytics used by organizations falls into the category of descriptive analytics.

A business learns from past behaviors to understand how they will impact future outcomes. Descriptive analytics is leveraged when a business needs to understand the overall performance of the company at an aggregate level and describe the various aspects.

Dr. Michael Wu, chief scientist of San Francisco-based Lithium Technologies describes descriptive analytics as -“The simplest class of analytics, one that allows you to condense big data into smaller, more useful nuggets of information.”

Descriptive analytics are based on standard aggregate functions in databases, which just require knowledge of basic school math. Most of the social analytics are descriptive analytics. They summarize certain groupings based on simple counts of some events. The number of followers, likes, posts, fans are mere event counters. These metrics are used for social analytics like average response time, average number of replies per post, %index, number of page views, etc. that are the outcome of basic arithmetic operations.

The best example to explain descriptive analytics are the results that a business gets from the web server through Google Analytics tools. The outcomes help understand what actually happened in the past and validate if a promotional campaign was successful or not based on basic parameters like page views.

What is Predictive Analytics?

The subsequent step in data reduction is predictive analytics. Analyzing past data patterns and trends can accurately inform a business about what could happen in the future. This helps in setting realistic goals for the business, effective planning and restraining expectations. Predictive analytics is used by businesses to study the data and ogle into the crystal ball to find answers to the question “What could happen in the future based on previous trends and patterns?”

Dr. Michael Wu, chief scientist of San Francisco-based Lithium Technologies said -"The purpose of predictive analytics is NOT to tell you what will happen in the future. It cannot do that. In fact, no analytics can do that. Predictive analytics can only forecast what might happen in the future, because all predictive analytics are probabilistic in nature."

Organizations collect contextual data and relate it with other customer user behavior datasets and web server data to get real insights through predictive analytics. Companies can predict business growth in future if they keep things as they are. Predictive analytics provides better recommendations and more future looking answers to questions that cannot be answered by BI.

Predictive analytics helps predict the likelihood of a future outcome by using various statistical and machine learning algorithms but the accuracy of predictions is not 100%, as it is based on probabilities. To make predictions, algorithms take data and fill in the missing data with best possible guesses. This data is pooled with historical data present in the CRM systems, POS Systems, ERP and HR systems to look for data patterns and identify relationships among various variables in the dataset. Organizations should capitalize on hiring a group of data scientists in 2016 who can develop statistical and machine learning algorithms to leverage predictive analytics and design an effective business strategy.

Predictive analytics can be further categorized as –

Predictive Modelling –What will happen next, if?

Root Cause Analysis-Why this actually happened?

Data Mining[2]- Identifying correlated data.

Forecasting- What if the existing trends continue?

Pattern Identification and Alerts –When should an action be invoked to correct a process.

Sentiment analysis is the most common kind of predictive analytics. The learning model takes input in the form of plain text and the output of the model is a sentiment score that helps determine whether the sentiment is positive, negative or neutral.

Organizations like Walmart, Amazon and other retailers leverage predictive analytics to identify trends in sales based on purchase patterns of customers, forecasting customer behavior, forecasting inventory levels, predicting what products customers are likely to purchase together so that they can offer personalized recommendations, predicting the amount of sales at the end of the quarter or year. The best example where predictive analytics find great application is in producing the credit score. Credit score helps financial institutions decide the probability of a customer paying credit bills on time.

What is Prescriptive Analytics?

Big data might not be a reliable crystal ball for predicting the exact winning lottery numbers but it definitely can highlight the problems and help a business understand why those problems occurred. Businesses can use the data-backed and data-found factors to create prescriptions for the business problems that lead to realizations and observations.

Prescriptive analytics is the next step of predictive analytics that adds the spice of manipulating the future. Prescriptive analytics advises on possible outcomes and results in actions that are likely to maximize key business metrics. It basically uses simulation and optimization to ask “What should a business do?”

Prescriptive analytics is an advanced analytics concept based on –

Optimization that helps achieve the best outcomes.

Stochastic optimization that helps understand how to achieve the best outcome and identify data uncertainties to make better decisions.

Simulating the future, under various set of assumptions, allows scenario analysis - which when combined with different optimization techniques, allows prescriptive analysis to be performed. Prescriptive analysis explores several possible actions and suggests actions depending on the results of descriptive and predictive analytics of a given dataset.

Prescriptive analytics is a combination of data, mathematical models and various business rules. The data for prescriptive analytics can be both internal (within the organization) and external (like social media data).Business rules are preferences, best practices, boundaries and other constraints. Mathematical models include natural language processing, machine learning, statistics, operations research, etc.

Prescriptive analytics are comparatively complex in nature and many companies are not yet using them in day-to-day business activities, as it becomes difficult to manage. Prescriptive analytics if implemented properly can have a major impact on business growth. Large scale organizations use prescriptive analytics for scheduling the inventory in the supply chain, optimizing production, etc. to optimize customer experience.

Aurora Health Care system saved $6 million annually by using prescriptive analytics to reduce re-admission rates by 10%. Prescriptive analytics can be used in healthcare to enhance drug development, finding the right patients for clinical trials, etc.

As increasing number of organizations realize that big data is a competitive advantage and they should ensure that they choose the right kind of data analytics solutions to increase ROI, reduce operational costs and enhance service quality.

**CHAPTER 5:**

#### 4 DESCRIPTIVE ANALYTICS:

Called the “simplest class of analytics”, descriptive analytics allows you to condense big data into smaller, more useful bits of information or a summary of what happened.

It has been estimated that more than 80% of business analytics (e.g. social analytics) are descriptive. Some social data could include the number of posts, fans, followers, page views, check-ins, pins, etc. It would appear to be an endless list if we tried to list them all.

Data Analysis is a process of collecting, transforming, cleaning, and modelling data with the goal of discovering the required information. The results so obtained are communicated, suggesting conclusions, and supporting decision-making. Data visualization is at times used to portray the data for the ease of discovering the useful patterns in the data. The terms Data Modelling and Data Analysis mean the same.

Data Analysis Process consists of the following phases that are iterative in nature −

* Data Requirements Specification
* Data Collection
* Data Processing
* Data Cleaning
* Data Analysis
* Communication
  1. Data Requirements Specification

The data required for analysis is based on a question or an experiment. Based on the requirements of those directing the analysis, the data necessary as inputs to the analysis is identified (e.g., Population of people). Specific variables regarding a population (e.g., Age and Income) may be specified and obtained. Data may be numerical or categorical.

* 1. Data Collection

Data Collection is the process of gathering information on targeted variables identified as data requirements. The emphasis is on ensuring accurate and honest collection of data. Data Collection ensures that data gathered is accurate such that the related decisions are valid. Data Collection provides both a baseline to measure and a target to improve.

Data is collected from various sources ranging from organizational databases to the information in web pages. The data thus obtained, may not be structured and may contain irrelevant information. Hence, the collected data is required to be subjected to Data Processing and Data Cleaning.

* 1. Data Processing

The data that is collected must be processed or organized for analysis. This includes structuring the data as required for the relevant Analysis Tools. For example, the data might have to be placed into rows and columns in a table within a Spreadsheet or Statistical Application. A Data Model might have to be created.

* 1. Data Cleaning

The processed and organized data may be incomplete, contain duplicates, or contain errors. Data Cleaning is the process of preventing and correcting these errors. There are several types of Data Cleaning that depend on the type of data. For example, while cleaning the financial data, certain totals might be compared against reliable published numbers or defined thresholds. Likewise, quantitative data methods can be used for outlier detection that would be subsequently excluded in analysis.

* 1. Data Analysis

Data that is processed, organized and cleaned would be ready for the analysis. Various data analysis techniques are available to understand, interpret, and derive conclusions based on the requirements. Data Visualization may also be used to examine the data in graphical format, to obtain additional insight regarding the messages within the data.

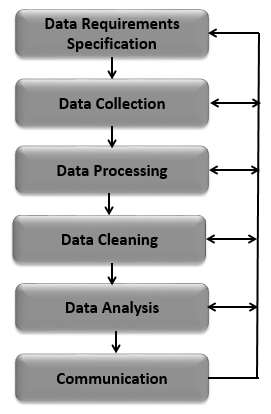
Statistical Data Models such as Correlation, Regression Analysis [1] can be used to identify the relations among the data variables. These models that are descriptive of the data are helpful in simplifying analysis and communicate results.

The process might require additional Data Cleaning or additional Data Collection, and hence these activities are iterative in nature.

* 1. Communication

The results of the data analysis are to be reported in a format as required by the users to support their decisions and further action. The feedback from the users might result in additional analysis.

The data analysts can choose data visualization techniques, such as tables and charts, which help in communicating the message clearly and efficiently to the users. The analysis tools provide facility to highlight the required information with color codes and formatting in tables and charts.



**CHAPTER 5**

Some of the tools used in Descriptive Analysis are as follows:

**TECHNIQUES:**

# 5.1 **BOX PLOT**

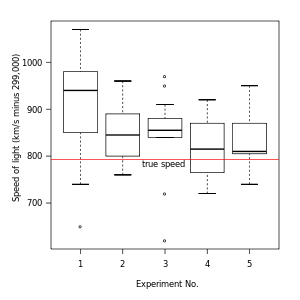
[](https://en.wikipedia.org/wiki/File:Michelsonmorley-boxplot.svg)

Figure 1. Box plot of data from the Michelson–Morley experiment

A simple way of representing statistical data on a plot in which a rectangle is drawn to represent the second and third quartiles, usually with a vertical line inside to indicate the median value. The lower and upper quartiles are shown as horizontal lines either side of the rectangle.

In [descriptive statistics](https://en.wikipedia.org/wiki/Descriptive_statistics), a box plot or boxplot is a method for graphically depicting groups of numerical data through their [quartiles](https://en.wikipedia.org/wiki/Quartile). Box plots may also have lines extending vertically from the boxes (*whiskers*) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. [Outliers](https://en.wikipedia.org/wiki/Outlier) may be plotted as individual points. Box plots are [non-parametric](https://en.wikipedia.org/wiki/Non-parametric): they display variation in samples of a [statistical population](https://en.wikipedia.org/wiki/Statistical_population) without making any assumptions of the underlying [statistical distribution](https://en.wikipedia.org/wiki/Probability_distribution). The spacing between the different parts of the box indicate the degree of [dispersion](https://en.wikipedia.org/wiki/Statistical_dispersion) (spread) and [skewness](https://en.wikipedia.org/wiki/Skewness) in the data, and show [outliers](https://en.wikipedia.org/wiki/Outlier). In addition to the points themselves, they allow one to visually estimate various [L-estimators](https://en.wikipedia.org/wiki/L-estimator), notably the [interquartile range](https://en.wikipedia.org/wiki/Interquartile_range), [midline](https://en.wikipedia.org/wiki/Midhinge), [range](https://en.wikipedia.org/wiki/Range_(statistics)), [mid-range](https://en.wikipedia.org/wiki/Mid-range), and [tri mean](https://en.wikipedia.org/wiki/Trimean). Box plots can be drawn either horizontally or vertically. Box plots received their name from the box in the middle.

The box plot is a quick way of examining one or more sets of data graphically. Box plots may seem more primitive than a [histogram](https://en.wikipedia.org/wiki/Histogram) or [kernel density estimate](https://en.wikipedia.org/wiki/Kernel_density_estimation) but they do have some advantages. They take up less space and are therefore particularly useful for comparing distributions between several groups or sets of data. Choice of [number and width of bins](https://en.wikipedia.org/wiki/Histogram#Number_of_bins_and_width) techniques can heavily influence the appearance of a histogram, and choice of bandwidth can heavily influence the appearance of a kernel density estimate.

#### **Box Plot in python**

import matplotlib.pyplot as plt

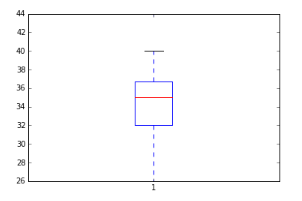
import pandas as pd

fig=plt.figure()

ax = fig.add\_subplot(1,1,1)

ax.boxplot(df['Age'])

plt.show()

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/05/Box_Plot_Violin_Plot1.png)

# 5.2 HISTOGRAM

A histogram is an accurate graphical representation of the [distribution](https://en.wikipedia.org/wiki/Frequency_distribution) of numerical data. It is an estimate of the [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution) of a [continuous variable](https://en.wikipedia.org/wiki/Continuous_variable) (quantitative variable) and was first introduced by [Karl Pearson](https://en.wikipedia.org/wiki/Karl_Pearson).[[1]](https://en.wikipedia.org/wiki/Histogram#cite_note-pearson-1) It is a kind of bar graph. To construct a histogram, the first step is to "[bin](https://en.wikipedia.org/wiki/Data_binning)" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping [intervals](https://en.wikipedia.org/wiki/Interval_(mathematics)) of a variable. The bins (intervals) must be adjacent, and are often (but are not required to be) of equal size.[[2]](https://en.wikipedia.org/wiki/Histogram#cite_note-2)

If the bins are of equal size, a rectangle is erected over the bin with height proportional to the [frequency](https://en.wikipedia.org/wiki/Frequency_(statistics)) — the number of cases in each bin. A histogram may also be [normalized](https://en.wikipedia.org/wiki/Normalization_(statistics)) to display "relative" frequencies. It then shows the proportion of cases that fall into each of several [categories](https://en.wikipedia.org/wiki/Categorization), with the sum of the heights equaling 1.

However, bins need not be of equal width; in that case, the erected rectangle is defined to have its *area* proportional to the frequency of cases in the bin.[[3]](https://en.wikipedia.org/wiki/Histogram#cite_note-3) The vertical axis is then not the frequency but *frequency density* — the number of cases per unit of the variable on the horizontal axis. Examples of variable bin width are displayed on Census bureau data below.

As the adjacent bins leave no gaps, the rectangles of a histogram touch each other to indicate that the original variable is continuous.

Histograms give a rough sense of the density of the underlying distribution of the data, and often for [density estimation](https://en.wikipedia.org/wiki/Density_estimation): estimating the [probability density function](https://en.wikipedia.org/wiki/Probability_density_function) of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the *x*-axis are all 1, then a histogram is identical to a [relative frequency](https://en.wikipedia.org/wiki/Relative_frequency) plot.

A histogram can be thought of as a simplistic [kernel density estimation](https://en.wikipedia.org/wiki/Kernel_density_estimation), which uses a [kernel](https://en.wikipedia.org/wiki/Kernel_(statistics)) to smooth frequencies over the bins. This yields a [smoother](https://en.wikipedia.org/wiki/Smooth_function) probability density function, which will in general more accurately reflect distribution of the underlying variable. The density estimate could be plotted as an alternative to the histogram, and is usually drawn as a curve rather than a set of boxes.

The histogram is one of the [seven basic tools of quality control](https://en.wikipedia.org/wiki/Seven_Basic_Tools_of_Quality).

Histograms are sometimes confused with bar charts. A histogram is used for [continuous data](https://en.wikipedia.org/wiki/Continuous_data), where the bins represent ranges of data, while a [bar chart](https://en.wikipedia.org/wiki/Bar_chart) is a plot of categorical variables. Some authors recommend that bar charts have gaps between the rectangles to clarify the distinction.

#### **Histogram in python**

fig=plt.figure() #Plots in matplotlib reside within a figure object, use plt.figure to create new figure

#Create one or more subplots using add\_subplot, because you can't create blank figure

ax = fig.add\_subplot(1,1,1)

#Variable

ax.hist(df['Age'],bins = 7) # Here you can play with number of bins

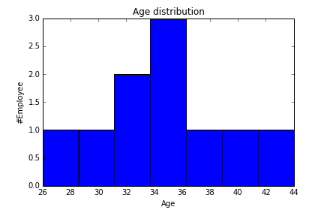
Labels and Tit

plt.title('Age distribution')

plt.xlabel('Age')

plt.ylabel('#Employee')

plt.show()



# 5.3 SCATTER PLOT

A scatter plot (also called a scatter graph, scatter chart, scatter gram, or scatter diagram) is a type of [plot](https://en.wikipedia.org/wiki/Plot_(graphics)) or [mathematical diagram](https://en.wikipedia.org/wiki/Mathematical_diagram) using [Cartesian coordinates](https://en.wikipedia.org/wiki/Cartesian_coordinate_system) to display values for typically two [variables](https://en.wikipedia.org/wiki/Variable_(mathematics)) for a set of data. If the points are color-coded, one additional variable can be displayed. The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the [vertical axis](https://en.wikipedia.org/wiki/Vertical_axis).

A scatter plot can be used either when one continuous variable that is under the control of the experimenter and the other depends on it or when both continuous variables are independent. If a [parameter](https://en.wikipedia.org/wiki/Parameter) exists that is systematically incremented and/or decremented by the other, it is called the *control parameter* or [independent variable](https://en.wikipedia.org/wiki/Independent_variable) and is customarily plotted along the horizontal axis. The measured or [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) is customarily plotted along the vertical axis. If no dependent variable exists, either type of variable can be plotted on either axis or a scatter plot will illustrate only the degree of [correlation](https://en.wikipedia.org/wiki/Correlation)  between two variables.

A scatter plot can suggest various kinds of correlations between variables with a certain [confidence interval](https://en.wikipedia.org/wiki/Confidence_interval). For example, weight and height, weight would be on y axis and height would be on the x axis. Correlations may be positive (rising), negative (falling), or null (uncorrelated). If the pattern of dots slopes from lower left to upper right, it indicates a positive [correlation](https://en.wikipedia.org/wiki/Correlation) between the variables being studied. If the pattern of dots slopes from upper left to lower right, it indicates a negative correlation. A line of [best fit](https://en.wikipedia.org/wiki/Curve_fitting) (alternatively called 'trendline') can be drawn in order to study the relationship between the variables. An equation for the correlation between the variables can be determined by established best-fit procedures. For a linear correlation, the best-fit procedure is known as [linear regression](https://en.wikipedia.org/wiki/Linear_regression) and is guaranteed to generate a correct solution in a finite time. No universal best-fit procedure is guaranteed to generate a correct solution for arbitrary relationships. A scatter plot is also very useful when we wish to see how two comparable data sets agree to show nonlinear relationships between variables. The ability to do this can be enhanced by adding a smooth line such as [LOESS](https://en.wikipedia.org/wiki/Local_regression). Furthermore, if the data are represented by a mixture model of simple relationships, these relationships will be visually evident as superimposed patterns.

The scatter diagram is one of the [seven basic tools](https://en.wikipedia.org/wiki/Seven_Basic_Tools_of_Quality) of [quality control](https://en.wikipedia.org/wiki/Quality_control).

Scatter charts can be built in the form of [bubble](https://en.wikipedia.org/wiki/Bubble_chart), marker, or/and [line charts](https://en.wikipedia.org/wiki/Line_chart).

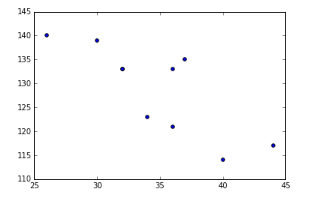
#### **Scatter Plot**

fig = plt.figure()

ax = fig.add\_subplot(1,1,1)

ax.scatter(df['Age'],df['Sales']) #You can also add more variables here to represent color and size.

plt.show()



**5.4 VIOLIN PLOT**

A **violin plot** is a method of plotting numeric data. It is similar to [box plot](https://en.wikipedia.org/wiki/Box_plot) with a rotated [kernel density plot](https://en.wikipedia.org/wiki/Kernel_density_estimation) on each side.

The violin plot is similar to [box plots](https://en.wikipedia.org/wiki/Box_plot), except that they also show the [probability density](https://en.wikipedia.org/wiki/Probability_density_function) of the data at different values (in the simplest case this could be a [histogram](https://en.wikipedia.org/wiki/Histogram)). Typically violin plots will include a marker for the median of the data and a box indicating the interquartile range, as in standard box plots. Overlaid on this box plot is a [kernel density estimation](https://en.wikipedia.org/wiki/Kernel_density_estimation). Like box plots, violin plots are used to represent comparison of a variable distribution (or sample distribution) across different "categories". For example temperature distribution compared between day and night or distribution of car prices compared across different car makers.

A violin plot is more informative than a plain box plot. In fact while a box plot only shows summary statistics such as mean/median and interquartile ranges, the violin plot shows the full distribution of the data. The difference is particularly useful when the data distribution is multimodal (more than one peak). In this case a violin plot clearly shows the presence of different peaks, their position and relative amplitude. This information could not be represented with a simple box plot which only reports summary statistics. The inner part of a violin plot usually shows the mean (or median) and the interquartile range. In other cases, when the number of samples is not too high, the inner part can show all sample points (with a dot or a line for each sample).

Although more informative than box plots, a disadvantage of violin plots is that they are less popular. Because of their unpopularity, their meaning can be harder to grasp for many readers not familiar with the violin plot representation. In this case, a more accessible alternative can be plotting a series of stacked histograms or [Kernel density distributions](https://en.wikipedia.org/wiki/Kernel_density_estimation).

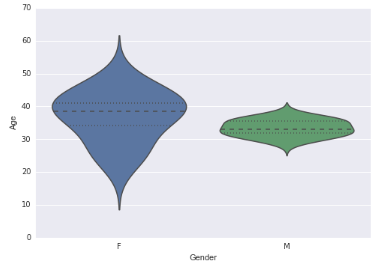
Violin plots are available as extensions to a number of software packages, the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) libraries [matplotlib](https://en.wikipedia.org/wiki/Matplotlib" \o "Matplotlib)[[3]](https://en.wikipedia.org/wiki/Violin_plot#cite_note-3) and Seaborn.[[4]](https://en.wikipedia.org/wiki/Violin_plot#cite_note-4)

#### **Violin Plot in Python**

import seaborn as sns

sns.violinplot(df['Age'], df['Gender']) #Variable Plot

sns.despine()



**5.6 BUBBLE PLOT**

A bubble chart is a type of [chart](https://en.wikipedia.org/wiki/Chart) that displays three dimensions of data. Each entity with its triplet (*v*1, *v*2, *v*3) of associated data is plotted as a disk that expresses two of the *vi* values through the disk's *xy* location and the third through its size. Bubble charts can facilitate the understanding of social, economic, medical, and other scientific relationships.

Bubble charts can be considered a variation of the [scatter plot](https://en.wikipedia.org/wiki/Scatter_plot), in which the data points are replaced with bubbles. As the documentation for [Microsoft Office](https://en.wikipedia.org/wiki/Microsoft_Office) explains, "You can use a bubble chart instead of a scatter chart if your data has three data series that each contain a set of values. The sizes of the bubbles are determined by the values in the third data series.".[[1]](https://en.wikipedia.org/wiki/Bubble_chart" \l "cite_note-1)

## Choosing bubble sizes correctly

The [human visual system](https://en.wikipedia.org/wiki/Human_visual_system) naturally experiences a disk's size in terms of its [area](https://en.wikipedia.org/wiki/Area). And the area of a disk—unlike its [diameter](https://en.wikipedia.org/wiki/Diameter) or [circumference](https://en.wikipedia.org/wiki/Circumference)—is not proportional to its [radius](https://en.wikipedia.org/wiki/Radius), but to the square of the radius. So if one chooses to scale the disks' radii to the third data values directly, then the apparent size differences among the disks will be non-linear and misleading. To get a properly weighted scale, one must scale each disk's radius to the [square root](https://en.wikipedia.org/wiki/Square_root) of the corresponding data value *v*3 This scaling issue can lead to extreme misinterpretations, especially where the range of the data has a large spread. And because many people are unfamiliar with—or do not stop to consider—the issue and its impact on perception, those who are aware of it often have to hesitate in interpreting a bubble chart because they cannot assume that the scaling correction was indeed made. So it is important that bubble charts not only be scaled in this way, but also be clearly labeled to document that it is area, rather than radius or diameter, that conveys the data.[[2]](https://en.wikipedia.org/wiki/Bubble_chart#cite_note-2)

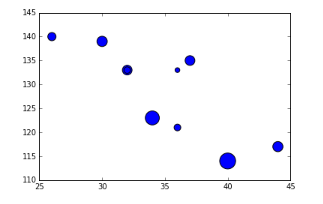
#### **Bubble Plot in python**

fig = plt.figure()

ax = fig.add\_subplot(1,1,1)

ax.scatter(df['Age'],df['Sales'], s=df['Income']) # Added third variable income as size of the bubble

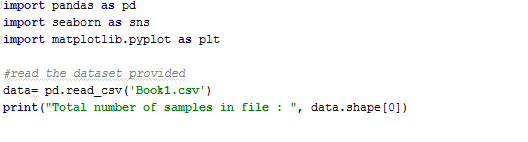
plt.show()



50

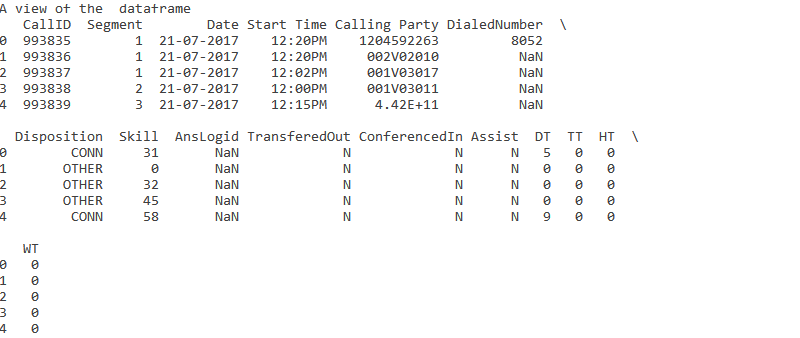
**CHAPTER 6**

**FINAL REPORT**

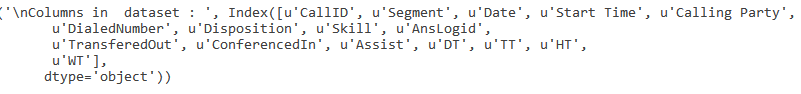




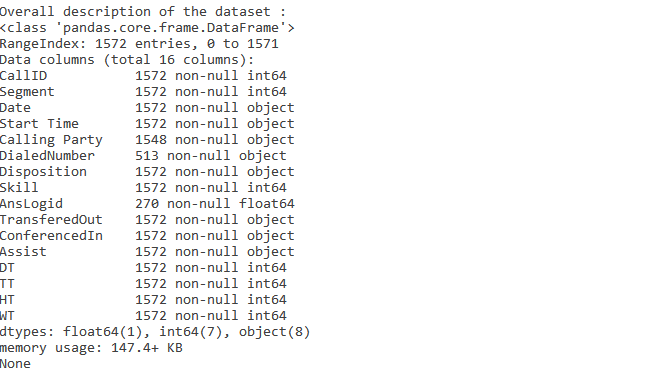




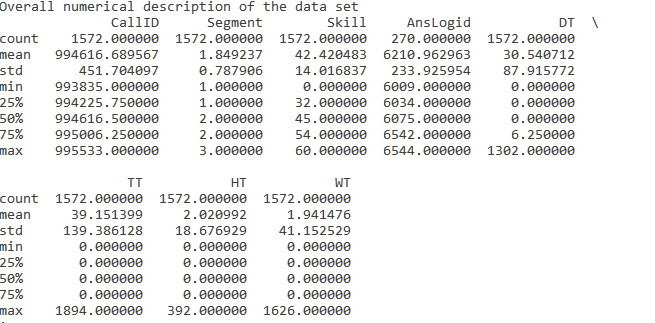


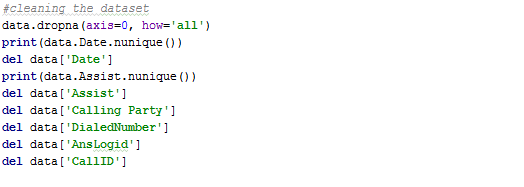




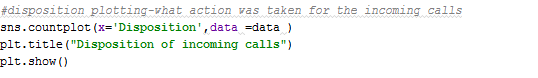


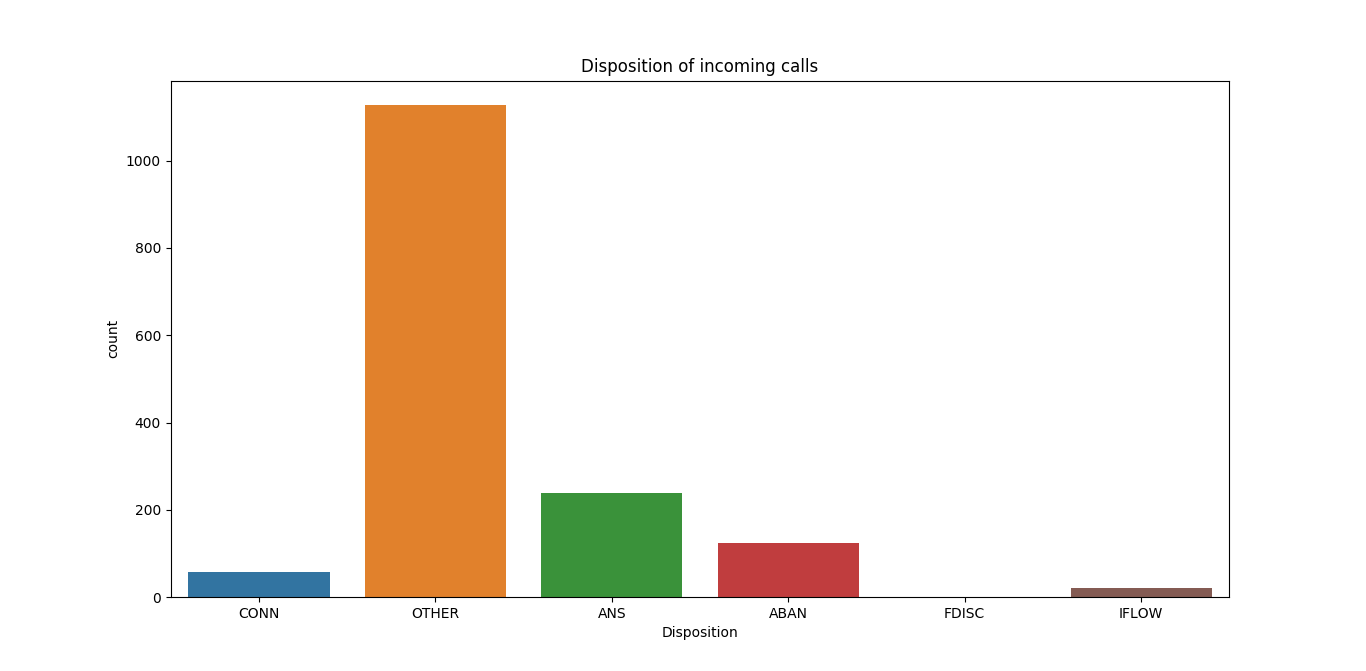


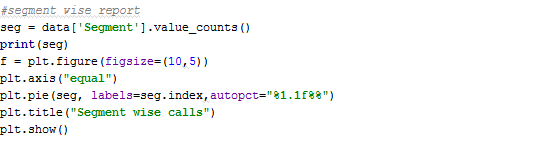


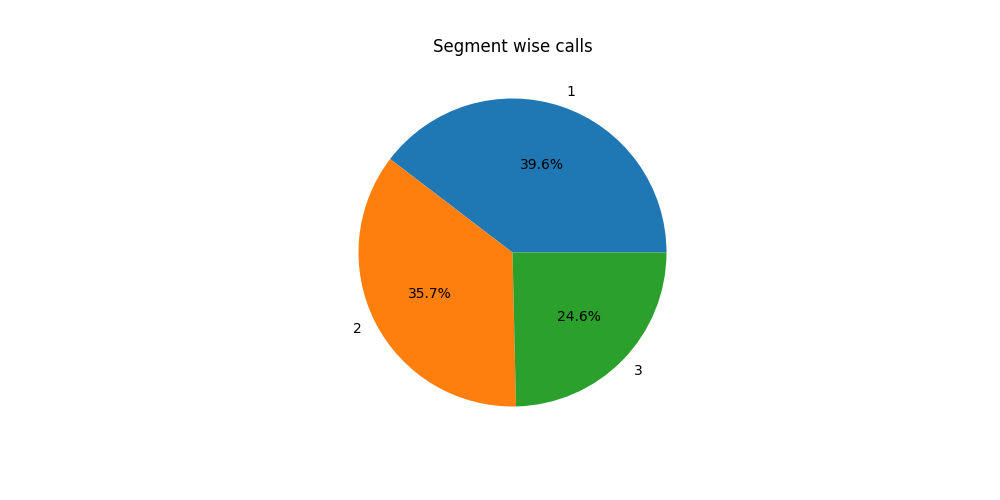


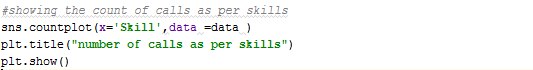


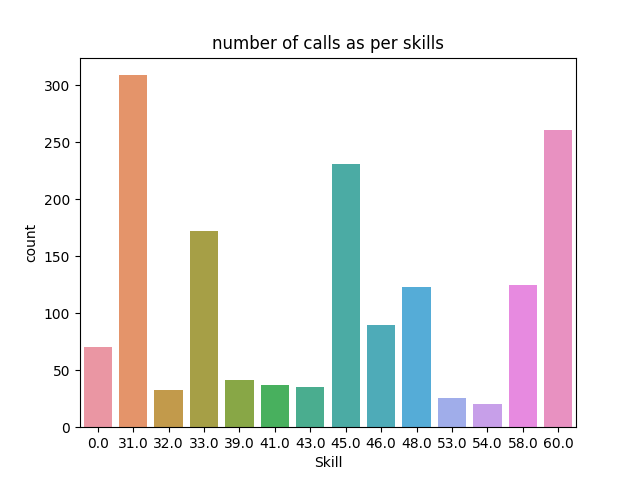


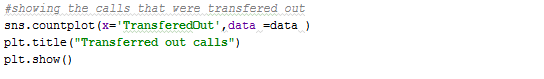


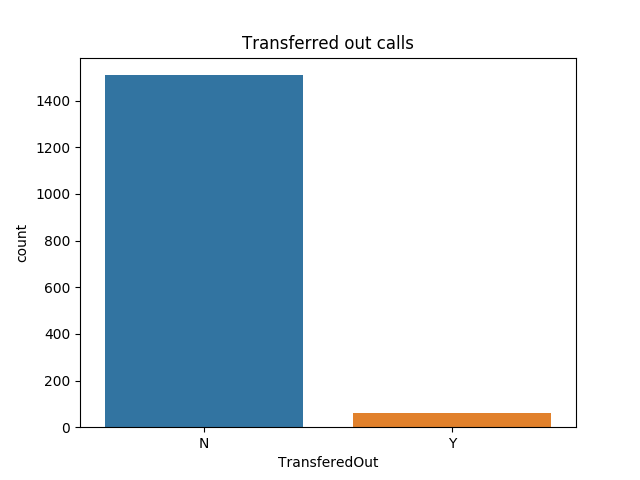


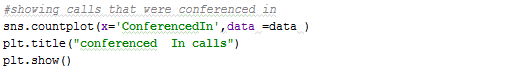


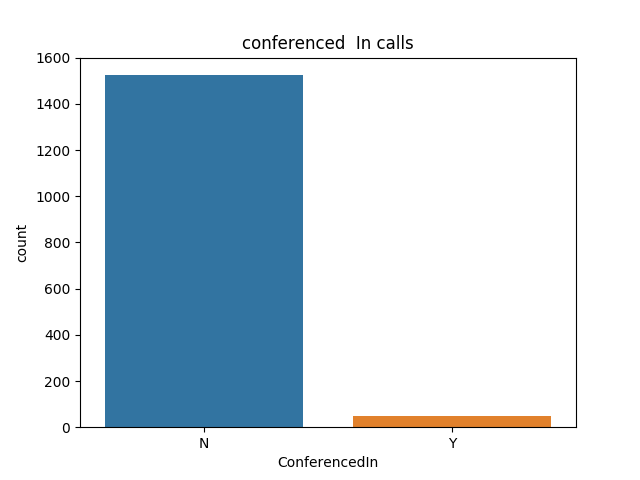


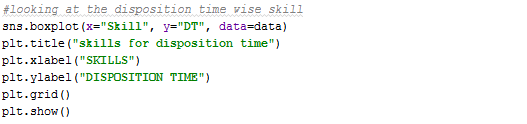


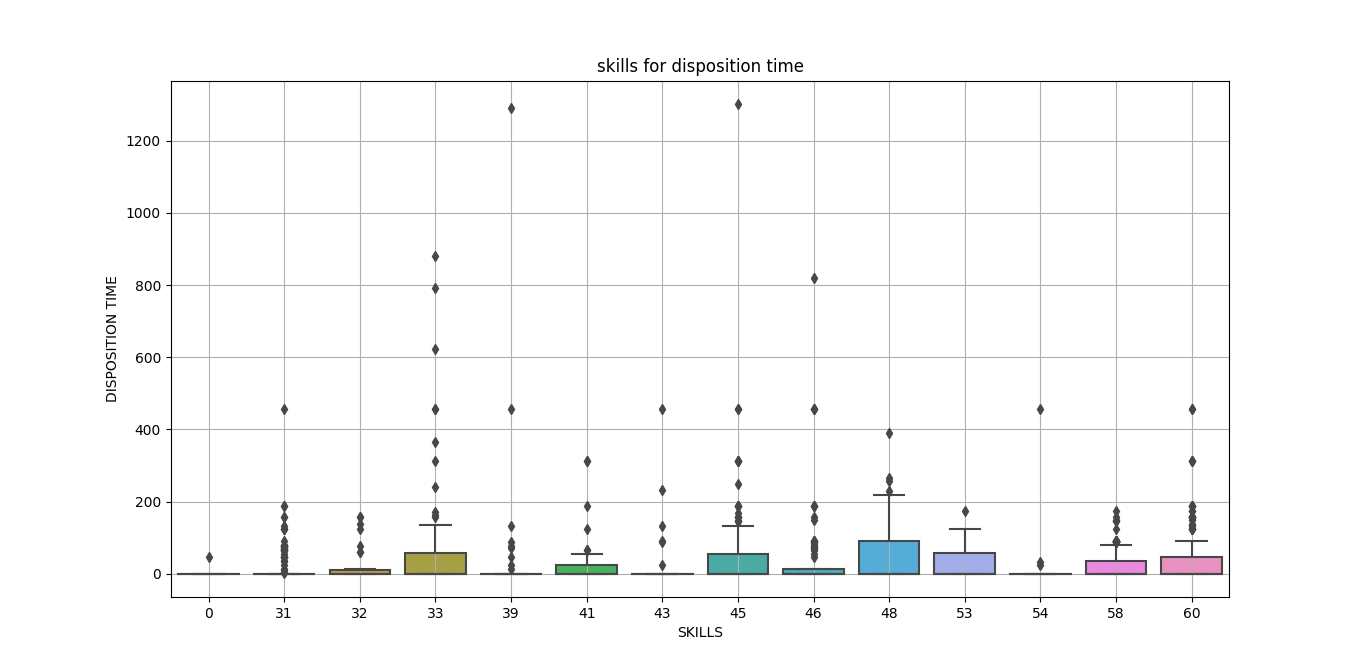


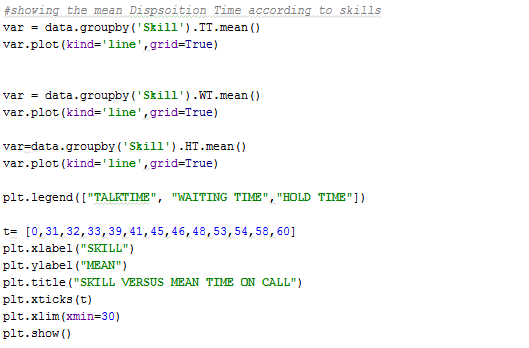


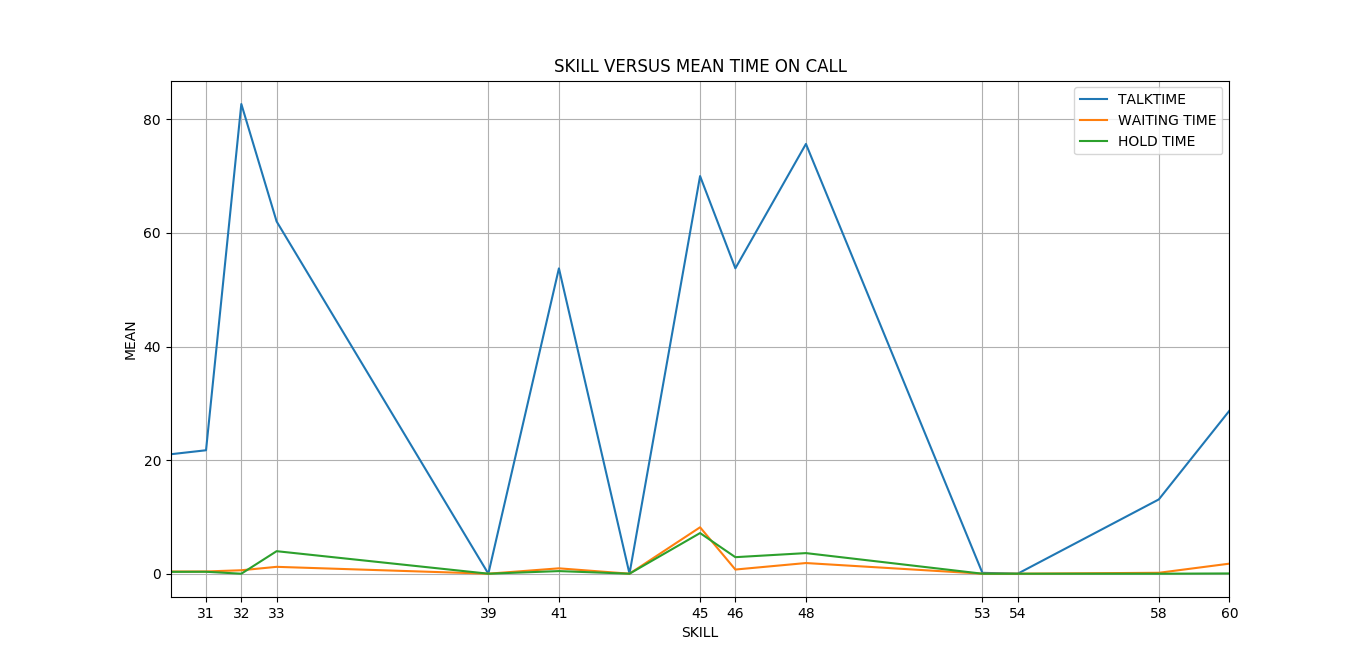




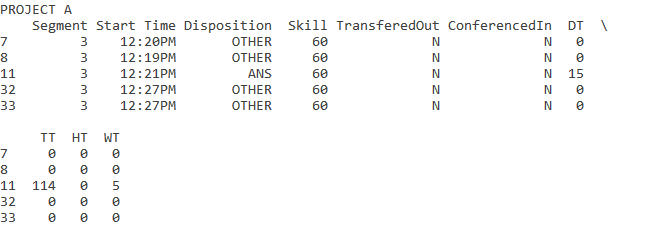


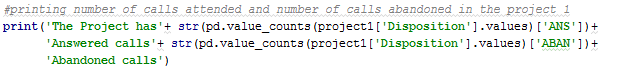




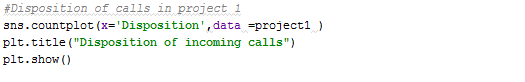


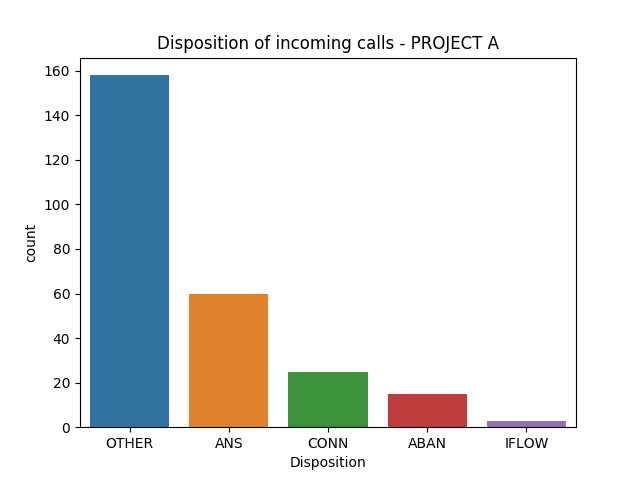


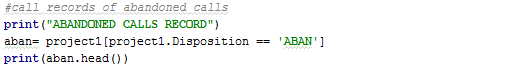


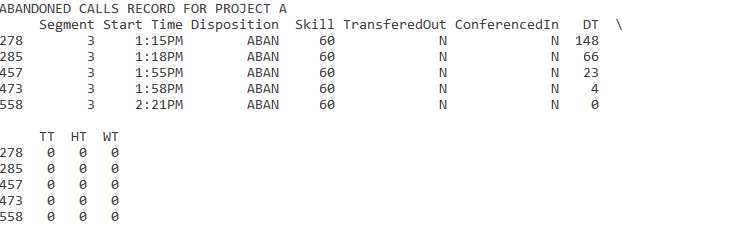


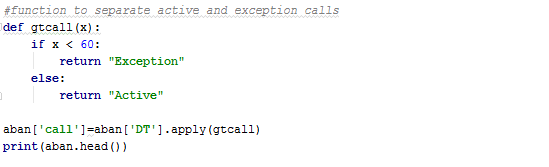


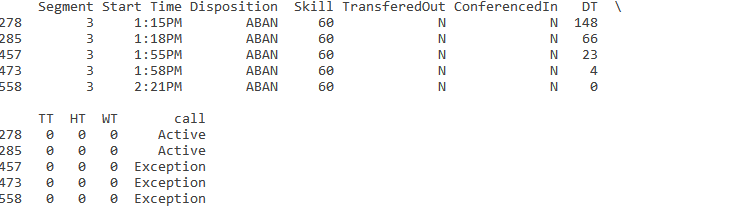


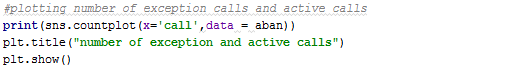


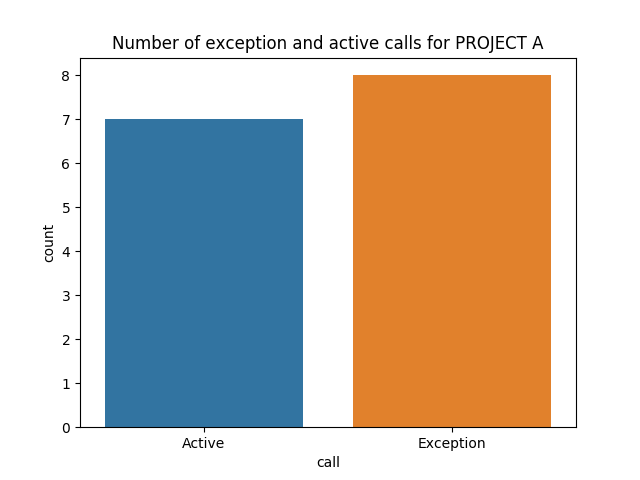


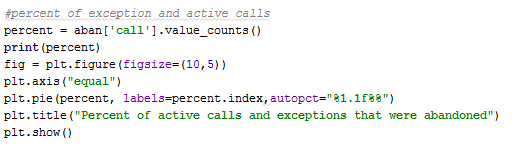


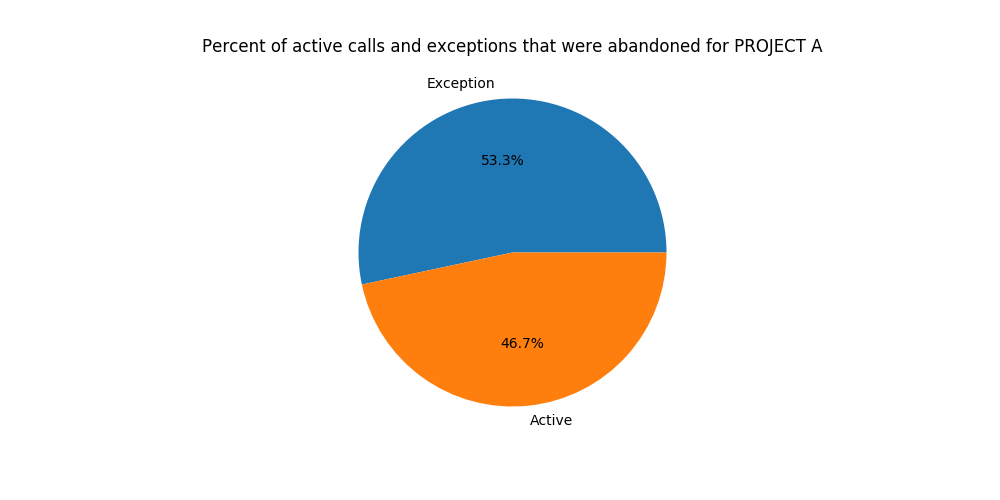












52

**CHAPTER 7: CONCLUSION**

The pace of business is accelerating around the globe, as leading organizations in computer-intensive industries, such as financial, trading, and manufacturing and retail industries move toward a real-time business model in which transactions and information sharing are near-instantaneous.

A study and its results can be assessed in multiple ways. Without a valid design, valid scientific conclusions cannot be drawn. Internal validity is an inductive estimate of the degree to which conclusions about causal relationships can be made (e.g., cause and effect), based on the measures used, the research setting, and the whole research design. External validity concerns the extent to which the (internally valid) results of a study can be held to be true for other cases, such as to different people, places, or times. In other words, it is about whether findings can be validly generalized. Learning about and applying statistics (as well as knowing their limitations) can help you better understand sociological research and studies. Knowledge of statistics helps you makes sense of the numbers in terms of relationships, and it allows you to ask relevant questions about sociological phenomena.

Further, today’s IT infrastructure has become pervasive in these organizations. It encompasses front, back and middle offices, covers customers, suppliers, employees and partners, and permeates every type of operations like strategizing, planning, manufacturing, servicing, etc. In the process, mission critical business processes heavily depend on the IT infrastructure. Yet another development is that this IT infrastructure is getting increasingly complex and specialized, compelling the institutions to develop expertise and specialties in these areas.

This transition is putting increasing demands on the performance, capacity, availability, and agility of underlying IT infrastructure. As process timelines are compressed from weeks or days, to hours, minutes or even seconds, the cost of downtime skyrockets. From supplier and customer transactions, to employee communications and financial reporting, business-critical functions must be up and running at all times. Business availability and continuity is critically poised on and directly correlated to and depends on the capacity, availability and reliability of the IT Infrastructure.

Therefore, the management of such a massive and complex infrastructure has become the current focus of discussions in many forums.

Business units are asking more of IT organizations. The rate at which new demands are being placed on the IT infrastructure is outpacing the capacity of IT organizations to effectively manage and support them. The complexity of what to manage and how to manage is compounded by the sheer number of new managed objects, as well as the legacy objects that must be maintained.

53

Businesses. Improving network efficiency is highly critical.

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. Data analysis is a process, within which several phases can be distinguished.

One way in which analysis can vary is by the nature of the data. Quantitative data is often analyzed using regressions. Regression analyses measure relationships between dependent and independent variables, taking the existence of unknown parameters into account.

Qualitative data can be coded--that is, key concepts and variables are assigned a shorthand, and the data gathered are broken down into those concepts or variables. Coding allows sociologists to perform a more rigorous scientific analysis of the data.

Sociological data analysis is designed to produce patterns. It is important to remember, however, that correlation does not imply causation; in other words, just because variables change at a proportional rate, it does not follow that one variable influences the other.

Without a valid design, valid scientific conclusions cannot be drawn. Internal validity concerns the degree to which conclusions about causality can be made. External validity concerns the extent to which the results of a study are generalizable.

**APPENDIX A**

Regression Analysis

In statistics, regression analysis 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. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

The type of data analysis employed can vary. One way in which analysis often varies is by the quantitative or qualitative nature of the data.

Quantitative data can be analyzed in a variety of ways, regression analysis being among the most popular. Regression analyses measure relationships between dependent and independent variables, taking the existence of unknown parameters into account. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. A large body of techniques for carrying out regression analysis has been developed. In practice, the performance of regression analysis methods depends on the form of the data generating process and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a large amount of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of causality based on observational data, regression methods give misleading results.

55

**APPENDIX B**

## **What is 'Data Mining'**

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. Data mining depends on effective data collection and [warehousing](http://www.investopedia.com/terms/w/warehousing.asp) as well as computer processing.

## **BREAKING DOWN 'Data Mining'**

Grocery stores are well-known users of data mining techniques. Many supermarkets offer free loyalty cards to customers that give them access to reduced prices not available to non-members. The cards make it easy for stores to track who is buying what, when they are buying it and at what price. The stores can then use this data, after analyzing it, for multiple purposes, such as offering customers [coupons](http://www.investopedia.com/terms/c/coupon.asp) targeted to their buying habits and deciding when to put items on sale or when to sell them at full price. Data mining can be a cause for concern when only selected information, which is not representative of the overall sample group, is used to prove a certain hypothesis.

## **Data Warehousing**

When companies centralize their data into one database or program, it is called data warehousing. With a data warehouse, an organization may spin off segments of the data for specific users to analyze and utilize. However, in other cases, analysts may start with the type of data they want and create a data warehouse based on those specs. Regardless of how businesses and other entities organize their data, they use it to support management's decision-making processes.

## **Data Mining Software**

Data mining programs analyze relationships and patterns in data based on what users request. For example, data mining software can be used to create classes of information. To illustrate, imagine a restaurant wants to use data mining to determine when they should offer certain specials. It looks at the information it has collected and creates classes based on when customers visit and what they order.

In other cases, data miners find clusters of information based on logical relationships, or they look at associations and sequential patterns to draw conclusions about trends in consumer behavior.

## **Data Mining Process**

The data mining process breaks down into five steps. First, organizations collect data and load it into their data warehouses. Next, they store and manage the data, either on in-house servers or the cloud. Business analysts, management teams and information technology professionals access the data and determine how they want to organize it. Then, application software sorts the data based on the user's results, and finally, the end user presents the data in an easy-to-share format, such as a graph or table.

56

**REFERENCES**

# [1] Big Data Analytics: Descriptive vs. Predictive vs. Prescriptive [online] Available at: [http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279 [Accessed](%20http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279%20%5bAccessed) 8 July 2017]

[2] dissertation-help-uk.blogspot. Solutions. [Online] Available at: http://dissertation-help-uk.blogspot.in/2011/12/importance-of-data-analysis-in-research.html [Accessed 5 July 2017]

[3] Big Data Analytics: Descriptive, Predictive and Prescriptive. [Online] Available at: https://www.charterglobal.com/big-data-analytics-descriptive-predictive-and-prescriptive/ [Accessed 7July 2017]

# [4]9 popular ways to perform Data Visualization in Python [online] Available at: https://www.analyticsvidhya.com/blog/2015/05/data-visualization-python/ [Accessed 18 July 2017]

[5] [Monte Carlo Simulation: What Is It and How Does It Work? - Palisade](http://www.palisade.com/risk/monte_carlo_simulation.asp). [Online] Available at: http://www.palisade.com/risk/monte\_carlo\_simulation.asp 25July2017]

[6] Matplotlib (2017). [Online] Available at: https://en.wikipedia.org/wiki/Matplotlib [Accessed 6 July 2017]

57

**GLOSSARY**

The following table identifies the terms referred to in this project report:

|  |  |  |
| --- | --- | --- |
| 1. | correlation | It is a statistic that measures the degree to which two [securities](http://www.investopedia.com/terms/s/security.asp) move in relation to each other. Correlations are used in advanced [portfolio management](http://www.investopedia.com/terms/p/portfoliomanagement.asp). Correlation is computed into what is known as the [correlation coefficient](http://www.investopedia.com/terms/c/correlationcoefficient.asp), which has value that must fall between -1 and 1. |
| 2. | Big Data | Extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions. |
| 3. | Business Intelligence | It comprises the set of strategies, processes, [applications](https://en.wikipedia.org/wiki/Application_software), [data](https://en.wikipedia.org/wiki/Data), technologies and technical architectures which are used by enterprises to support the collection, [data analysis](https://en.wikipedia.org/wiki/Data_analysis), presentation and dissemination of [business information](https://en.wikipedia.org/wiki/Business_information). BI technologies provide historical, current and predictive views of [business operations](https://en.wikipedia.org/wiki/Business_operations). |
| 4. | Stochastic | Having a random probability distribution or pattern that may be analysed statistically but may not be predicted precisely. |
| 5. | Natural Language Processing | It is a field of [computer science](https://en.wikipedia.org/wiki/Computer_science), [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) and [computational linguistics](https://en.wikipedia.org/wiki/Computational_linguistics) concerned with the interactions between [computers](https://en.wikipedia.org/wiki/Computer) and [human (natural) languages](https://en.wikipedia.org/wiki/Natural_language), and, in particular, concerned with programming computers to fruitfully process large [natural language corpora](https://en.wikipedia.org/wiki/Corpus_linguistics). |
| 6. | Machine Learning | It is the subfield of [computer science](https://en.wikipedia.org/wiki/Computer_science) that, gives "computers the ability to learn without being explicitly programmed." Evolved from the study of [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition) and [computational learning theory](https://en.wikipedia.org/wiki/Computational_learning_theory) in [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence), machine learning explores the study and construction of [algorithms](https://en.wikipedia.org/wiki/Algorithm) that can learn from and make predictions on [data](https://en.wikipedia.org/wiki/Data)– such algorithms overcome following strictly static [program instructions](https://en.wikipedia.org/wiki/Computer_program) by making data-driven predictions or decisions |
| 7. | Monte Carlo Simulation | It is a computerized mathematical technique that allows people to account for risk in quantitative analysis and decision making. The technique is used by professionals in such widely disparate fields as finance, project management, manufacturing, engineering, research and development, insurance, oil & gas, transportation, and the environment. |
| 8. | NumPy | It is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects, and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more. |
| 9. | Matplotlib | Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. |
| 10 | Seaborn | Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics. |

59