FINAL YEAR MAJOR PROJECT REPORT

"NEIGHBOURHOOD RECOMMENDER"

Submitted in Partial Fulfillment for the Award of Degree of Bachelor of Technology in Computer Science and Engineering from Rajasthan Technical University, Kota

COORDINATOR:

Mr. Ankit Kumar
(Dept. of Computer Science & Engineering)

SUBMITTED BY:

Harshvardhan Singh Khidiya (16ESKCS718)

MENTOR:

Mr. Ankit Kumar
(Dept. of Computer Science & Engineering)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SWAMI KESHWANAND INSTITUTE OF TECHNOLOGY,
MANAGEMENT & GRAMOTHAN
Ramnagaria (Jagatpura), Jaipur – 302017

SESSION2019-20

SWAMI KESHWANAND INSTITUTE OF TECHNOLOGY, MANAGEMENT &GRAMOTHAN
Ramnagaria (Jagatpura), Jaipur – 302017

CERTIFICATE

This is to certify that Final Year Major Project Report entitled "**Neighbourhood Recommender**" has been duly submitted by

Harshvardhan Singh Khidiya (16ESKCS718)

for partial fulfillment of the Degree of Bachelor of Technology of Rajasthan Technical University. It has been found satisfactory and hence approved for submission as Major Project during academic session 2019-2020.

Date: 5 October 2020

COORDINATOR:	**MENTOR:**	**HEAD OF DEPARTMENT:**
Mr. Ankit Kumar	**Mr. Ankit Kumar**	**Dr. Mukesh Kumar Gupta**
(Dept. of Computer Science & Engineering)	(Dept. of Computer Science & Engineering)	(Dept. of Computer Science & Engineering)

ABSTRACT

My project provides an interface for an Indian Restaurant Chain to help to decide where it should open its new Restaurant in the City of London, based upon the location of other Indian Restaurant and density of Indian Population living in various areas of London.

Our customer is ABC Restaurant, which is an International Indian Restaurant Brand and also a market leader.ABC Restaurant has recently planned to expand its business to the United Kingdom, and they want to start this journey from the heart of the UK itself. Given the extremely large size and the population of the city, our customer wants to identify the best neighbourhood area to open its first Indian Restaurant covering the majority of the population and facing least competition from other restaurants. The problem statement will be: Which neighbourhood has most Indian population and has lesser number of INDIAN RESTAUARNT's**?**

SWAMI KESHWANAND INSTITUTE OF TECHNOLOGY,
MANAGEMENT & GRAMOTHAN
Ramnagaria (Jagatpura), Jaipur – 302017

DECLARATION

We hereby declare that the report of the project entitled Neighbourhood Recommender is a record of an original work done by us at Swami Keshvanand Institute of Technology, Management &Gramothan, Jaipur under the mentorship of **Mr. Ankit Kumar** (Dept. of Computer Science & Technology) and coordination of **Mr. Ankit Kumar** (Dept. of Computer Science & Technology). This project report has been submitted as the proof of original work for the partial fulfillment of the requirement for the award of the degree of Bachelor of Technology (B.Tech) in the Department of Computer Science & Technology. It has not been submitted anywhere else, under any other program to the best of our knowledge and belief.

Team Members:	**Signatures:**
(16ESKCS718) **Harshvardhan Singh Khidiya**	

ACKNOWLEDGMENT

A project of such a vast coverage cannot be realized without help from numerous sources and people in the organization. We take this opportunity to express our gratitude to all those who have been helping us in making this project successful.

We are highly indebted to our faculty mentor **Mr. Ankit Kumar.** He has been a guide, motivator & source of inspiration for us to carry out the necessary proceedings for the project to be completed successfully. We also thank our project coordinator **Mr. Ankit Kumar** for his co-operation, encouragement, valuable suggestions and critical remarks that galvanized our efforts in the right direction.

We would also like to convey our sincere thanks to **Dr. Mukesh Kumar Gupta,** HOD, Department of Computer Science & Engineering, for facilitating, motivating and **s**upporting us during each phase of development of the project. Also, we pay our sincere gratitude to all the **Faculty Members** of Swami Keshvanand Institute of Technology, Management &Gramothan, Jaipur and all our Colleagues for their co-operation and support.

Last but not least we would like to thank all those who have directly or indirectly helped and cooperated in accomplishing this project.

Team Members:

(16ESKCS718) **Harshvardhan Singh Khidiya**

INDEX

INDEX OF FIGURES

Chapter 1

Introduction

1.1 OVERVIEW

My project provides an interface for an Indian Restaurant Chain to help to decide where it should open its new Restaurant in the City of London, based upon the location of other Indian Restaurant and density of Indian Population living in various areas of London. Our customer is ABC Restaurant, which is an International Indian Restaurant Brand and also a market leader.ABC Restaurant has recently planned to expand its business to the United Kingdom, and they want to start this journey from the heart of the UK itself. Given the extremely large size and the population of the city, our customer wants to identify the best neighbourhood area to open its first Indian Restaurant covering the majority of the population and facing least competition from other restaurants

1.2 MOTIVATION

The purpose of the document is to collect and analyze all assorted ideas that have come up to define the system, its requirements with respect to consumers. Also, we shall predict and sort out how we hope this product will be used in order to gain a better understanding of the project, outline concepts that may be developed later, and document ideas that are being considered, but may be discarded as the product develops.

In short, the purpose of this SRS document is to provide a detailed overview of our software product, its parameters and goals. This document describes the project's target audience and its user interface, hardware and software requirements. It defines how our client, team and audience see the product and its functionality.

1.2 OBJECTIVES OF PROJECT

- The Existing database is huge with less or no insights driven hence the strategies so formed to gain customer base & to increase profits are less effective.

- The neighborhood recommender project will help the company to improvise its online marketing strategies through email marketing, social media marketing, etc.

- Project will allow the company to highlight the major reasons for improper locations for new businesses hence reducing their cost to a considerable point.

- Pricing Elasticity model will help the company to increase its profit while making its customer presence more favorable.

- The data will then be integrated with the machine learning algorithms to make the system self independent in providing effective strategies.

Chapter 2

Introduction to Project

2.1 OVERVIEW OF PROJECT

This project provides an interface for an Indian Restaurant Chain to help to decide where it should open its new Restaurant in the City of London, based upon the location of other Indian Restaurant and density of Indian Population living in various areas of London.

Neighborhood Recommender collects the data from various sources such as Wikipedia, Geopy – Python Library for collecting Coordinate locations of an area based upon their Pin code area. Then the data will be further processed to obtain the precise locations that will be suitable for opening a new restaurant in the city of London.

Chapter 3

Description of Modules

3.1 Technologies Used

- **Python language**

Python is a language and environment for statistical computing and graphics. It is an Open Source Project which is similar to the Lisp, Java language and environment which was developed Guido van Rossum and colleagues. Python can be considered as a different implementation of Lisp. There are some important differences, but much code written for Python runs unaltered under Lisp.

Python provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The Python language is often the vehicle of choice for research in statistical methodology, and Lisp provides an Open Source route to participation in that activity.

One of Python's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

The Python Environment

Python is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hard copy, and

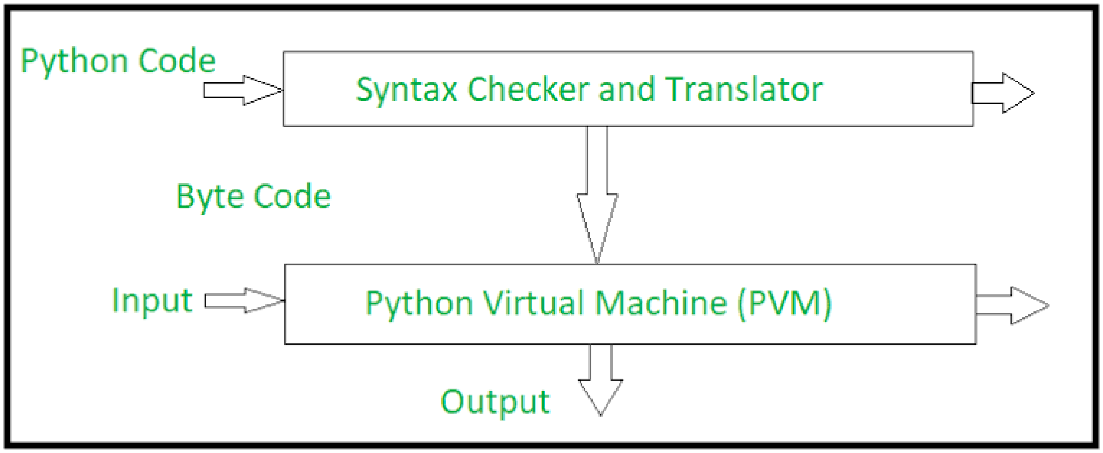


Fig 1.1 Basic Python Architecture

The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

Python, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the python dialect of Lisp, which makes it easy for users to follow the algorithmic choices made. For computationally-intensive tasks, C, C++ and FORTRAN code can be linked and called at run time. Advanced users can write C code to manipulate Python objects directly.

Many users think of python as a statistics system. We prefer to think of it of an environment within which statistical techniques are implemented. Python can be extended (easily) via packages. There are about eighty packages supplied with the Python distribution and many more are available through the py.org family of Internet sites covering a very wide range of modern statistics.

Python has its own Latex-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hard copy.

Anaconda

To increase the productivity of Python, a set of integrated tools known as Anaconda is used. Anaconda was created by Anaconda Inc, formerly known as Continuum Analytics Inc.

Its interface is organized so that the user can clearly view graphs, data tables, python code, and output all at the same time. It also offers an Import-Wizard-like feature that allows users to import CSV, Excel, SAS (*.sas7bdat), SPSS (*.sav), and Stata (*.dta) files into Python without having to write the code to do so.

Python Studio is partly written in the C++ programming language and uses the Qt framework for its graphical user interface. The bigger percentage of the code is written in Java, JavaScript is also amongst the languages used.

• PostgreSQL

PostgreSQL, often simply Postgres, is an object-relational database management system (ORDBMS) with an emphasis on extensibility and standards compliance. It can handle workloads ranging from small single-machine applications to large Internet-facing applications (or for data warehousing) with many concurrent users; on macOS Server, PostgreSQL is the default database; and it is also available for Microsoft Windows and Linux (supplied in most distributions).

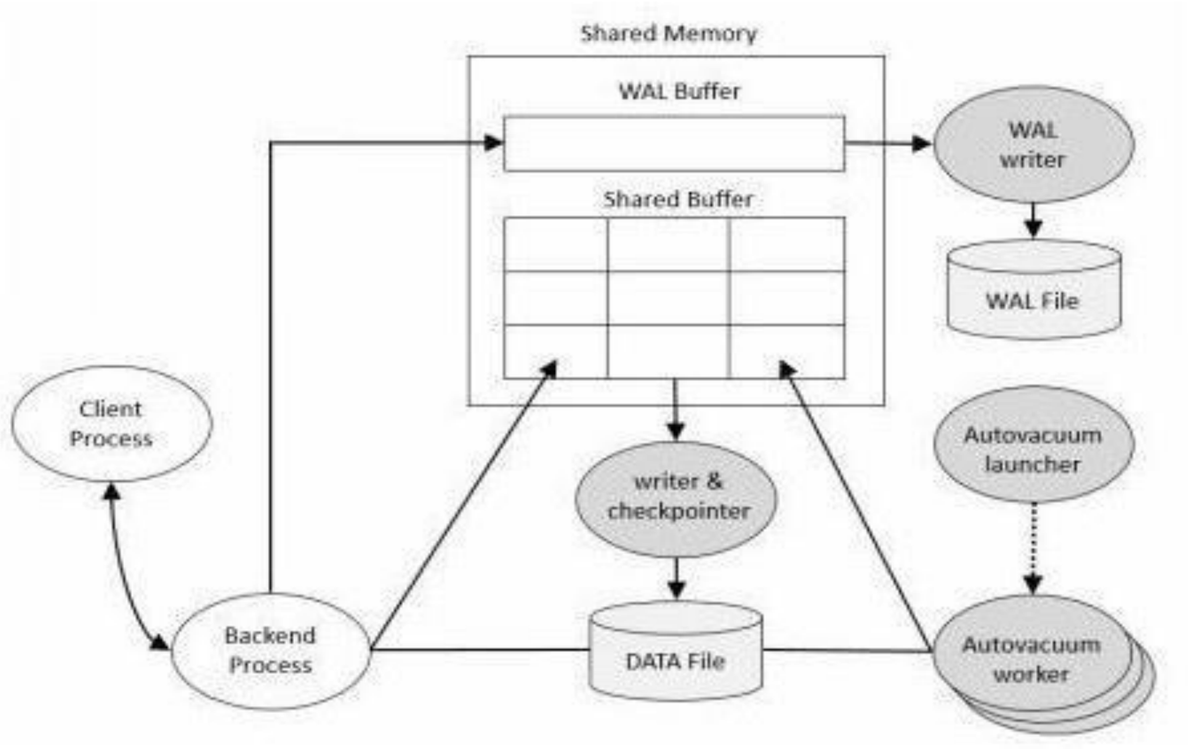


Fig 1.2 PostgreSQL Architecture

PostgreSQL is ACID-compliant and transactional. PostgreSQL has updatable views and materialized views, triggers, foreign keys; supports functions and stored procedures, and otherexpandability.

PostgreSQL is developed by the PostgreSQL Global Development Group, a diverse group of many companies and individual contributors. It is free and open-source, released under the terms of the PostgreSQL License, a permissive software license.

pgAdmin

pgAdmin is the most popular and feature rich Open Source administration and development platform for PostgreSQL, the most advanced Open Source database in the world. The application may be used on Linux, FreeBSD, Solaris, Mac OS X and Windows platforms to manage PostgreSQL 8.2 and above running on any platform, as well as commercial versions of PostgreSQL such as Mammoth PostgreSQL, EnterpriseDB Postgres Plus Advanced Server and Greenplum Database.

pgAdmin is designed to answer the needs of all users, from writing simple SQL queries to developing complex databases. The graphical interface supports all PostgreSQL features and makes administration easy. The application also includes a syntax highlighting SQL editor, a server-side code editor, an SQL/batch/shell job scheduling agent, support for the Slony-I replication engine and much more. Server connection may be made using TCP/IP or Unix Domain Sockets (on *nix platforms), and may be SSL encrypted for security. No additional drivers are required to communicate with the database server.

pgAdmin is developed by a community of PostgreSQL experts around the world and is available in more than a dozen languages. It is Free Software released under the PostgreSQL License.

• Google Analytics

Google Analytics is a freemium web analytics service offered by Google that tracks and reports website traffic, currently as a platform inside the Google Marketing Platform brand. Google launched the service in November 2005 after acquiring Urchin.

Google Analytics is now the most widely used web analytics service on the web. Google Analytics also provides an SDK that allows gathering usage data from iOS and Android Apps, known as Google Analytics for Mobile Apps.

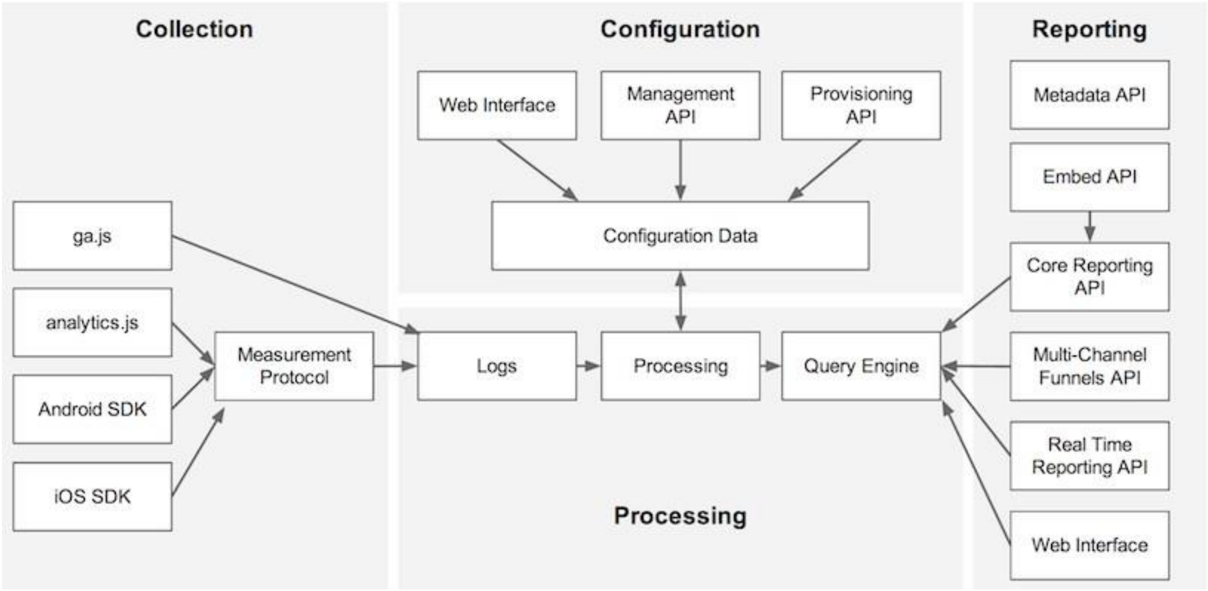


Fig 1.3 Google Analytics Architecture

Through Google Analytics, users can review online campaigns by tracking landing page quality and conversions (goals). Goals might include sales, lead generation, viewing a specific page, or downloading a particular file. Google Analytics' approach is to show high-level, dashboard-type data for the casual user, and more in-depth data further into the report set. Google Analytics analysis can identify poorly performing pages with techniques such as funnel visualization, where visitors came from (referrers), how long they stayed on the website and their geographical position. It also provides more advanced features, including custom visitor segmentation. Google Analytics e-commerce reporting can track sales activity and performance. The e-commerce report shows a site's transactions, revenue, and many other commerce-related metrics.

- ## Microsoft Excel

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. Excel forms part of Microsoft Office.

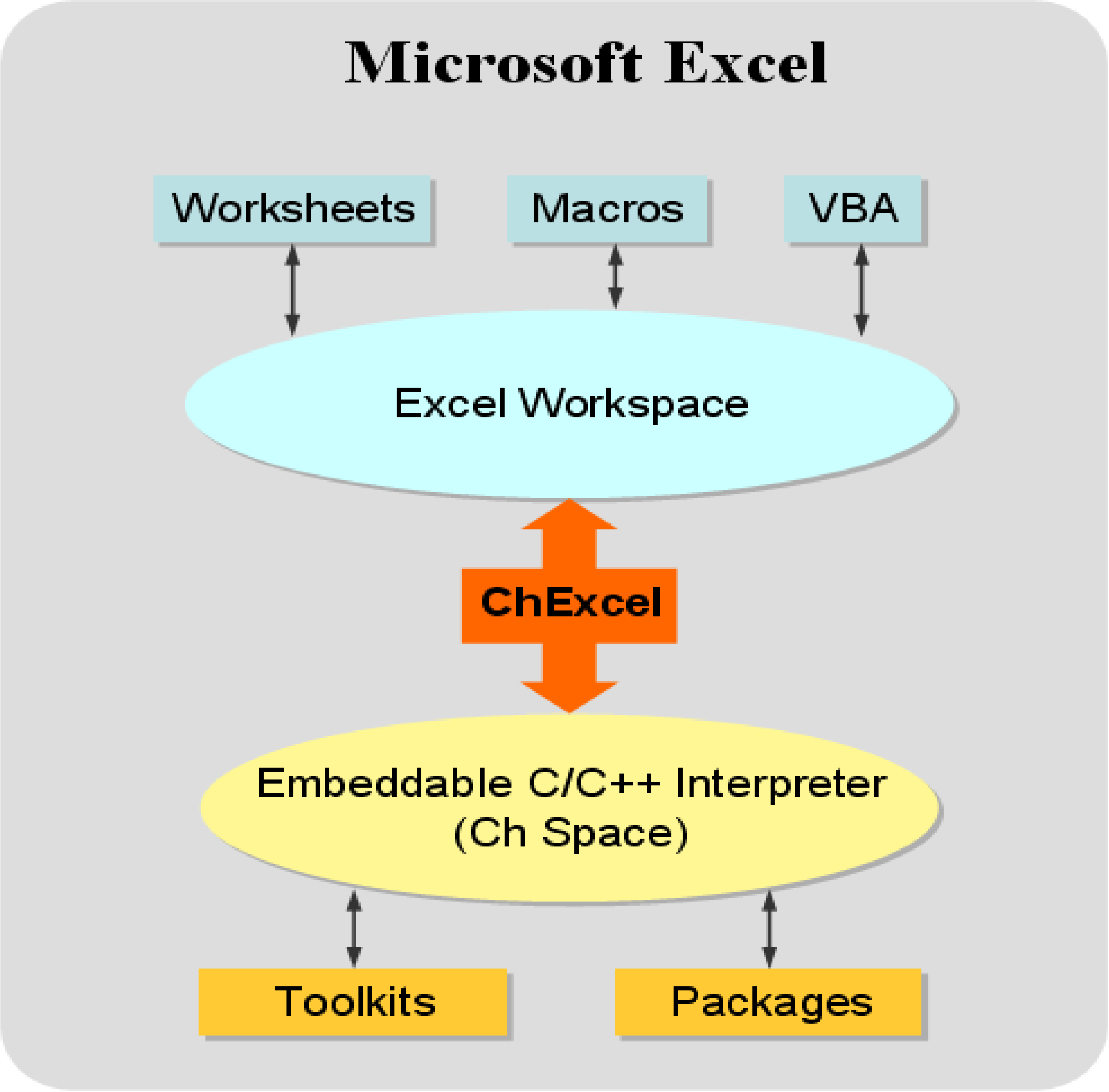


Fig 1.4 Microsoft Excel Architecture

Microsoft Excel has the basic features of all spreadsheets, using a grid of cells arranged in numbered rows and letter-named columns to organize data manipulations like arithmetic operations. It has a battery of supplied functions to answer statistical, engineering and financial needs. In addition, it can display data as line graphs, histograms and charts, and with a very limited three-dimensional graphical display. It allows sectioning of data to view its dependencies on various factors for different perspectives (using pivot tables and the scenario manager). It has a programming aspect, Visual Basic for Applications, allowing the user to employ a wide variety of numerical methods, for example, for solving differential equations of mathematical physics, and then reporting the results back to the spread sheet.

It also has a variety of interactive features allowing user interfaces that can completely hide the spreadsheet from the user, so the spreadsheet presents itself as a so-called application, or decision support system (DSS), via a custom-designed user interface, for example, a stock analyzer, or in general, as a design tool that asks the user questions and provides answers and reports. In a more elaborate realization, an Excel application can automatically poll external databases and measuring instruments using an update schedule, analyze the results, make a Word report or PowerPoint slide show, and e-mail these presentations on a regular basis to a list of participants.

3.2 Process of Data Analysis

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 are collected and analyzed to answer questions, test hypotheses or disprove theories.

Statistician John Tukey defined data analysis in 1961 as: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

There are several phases that can be distinguished, described below. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases.

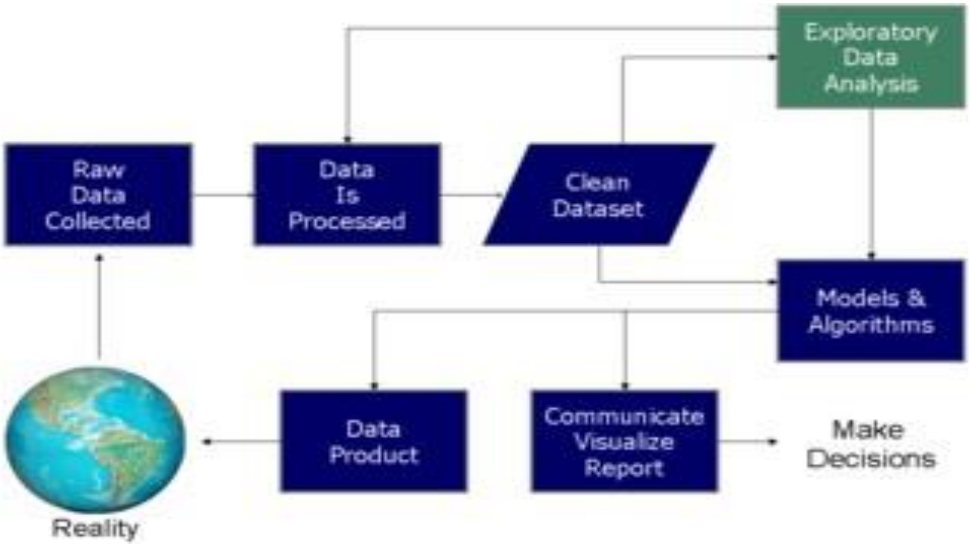


Fig. 1.5: Process of Data Analysis

3.2.1 Data Requirements:

The data are 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).

3.2.2 Data Collection:

Data are 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.

3.2.3 Data Processing:

Data initially obtained must be processed or organized 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.

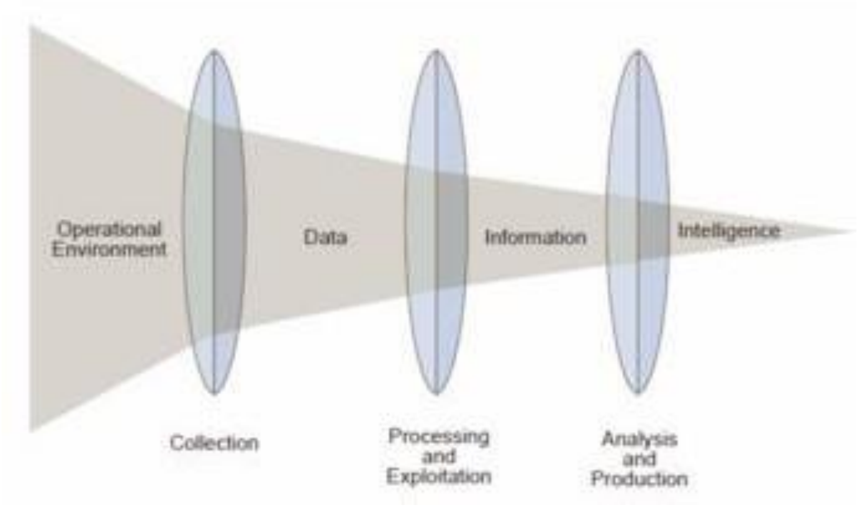


Fig.1.6: Relationship of Data, Information and Intelligence

3.2.4 Data Cleaning:

Once processed and organized, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that data are entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, and overall quality of existing data, de- duplication, and column segmentation. Such data problems can also be identified through a variety of analytical techniques. For example, with financial information, the totals for particular variables may be compared against separately published numbers believed to be reliable. Unusual amounts above or below pre-determined thresholds may also be reviewed. There are several types of data cleaning that depend on the type of data such as phone numbers, email addresses, employers etc. Quantitative data methods for outlier detection can be used to get rid of likely incorrectly entered data. Textual data spell checkers can be used to lessen the amount of mistyped words, but it is harder to tell if the words themselves are correct.

3.2.5 Exploratory Data Analysis:

Once the data are 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. The process of exploration may result in additional data cleaning or additional requests for data, so these activities may be iterative in nature. Descriptive statistics, such as the average or median, may be generated to help understand the data. Data visualization may also be used to examine the data in graphical format, to obtain additional insight regarding the messages within the data.

3.2.6 Modeling and Algorithms:

Mathematical formulas or models called algorithms may be applied to the data to identify relationships among the variables, such as correlation or causation. In general terms, models may be developed to evaluate a particular variable in the data based on other variable(s) in the data, with some residual error depending on model accuracy (i.e., Data = Model + Error).

Inferential statistics includes techniques to measure relationships between particular variables. For example, regression analysis may be used to model whether a change in advertising (independent variable X) explains the variation in sales (dependent variable Y). In mathematical terms, Y (sales) is a function of X (advertising). It may be described as Y = aX + b + error, where the model is designed such that a and b minimize the error when the model predicts Y for a given range of values of X. Analysts may attempt to build models that are descriptive of the data to simplify analysis and communicate results.

3.2.7 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.

3.2.8 Communication:

Once the data are 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.

When determining how to communicate the results, the analyst may consider data visualization techniques to help clearly and efficiently communicate the message to the audience. Data visualization uses information displays (such as tables and charts) to help communicate key messages contained in the data. Tables are helpful to a user who might lookup specific numbers, while charts may help explain the quantitative messages contained in the data.

Fig. 1.7: Data Visualization

3.3 Data Analytics Tools

3.3.1 R Programming:

R is the leading analytics tool in the industry and widely used for statistics and data modeling. It can easily manipulate data and present in different ways. It has exceeded SAS in many ways like capacity of data, performance and outcome. R compiles and runs on a wide variety of platforms viz -UNIX, Windows and MacOS. It has 11,556 packages and allows user to browse the packages by categories. R also provides tools to automatically install all packages as per user requirement, which can also be well assembled with Bigdata.

3.3.2 Tableau:

Tableau connects any data source be it corporate Data Warehouse, Microsoft Excel or web-based data, and creates data visualizations, maps, dashboards etc. with real-time updates presenting on web. They can also be shared through social media or with the client. It allows the access to download the file in different formats.

3.3.3 Python:

Python is an object-oriented scripting language which is easy to read, write, maintain and is a free open source tool. It was developed by Guido van Rossum in late 1980's which supports both functional and structured programming methods. Python is easy to learn as it is very similar to JavaScript, Ruby, and PHP. Also, Python has very good machine learning libraries viz. Scikitlearn, Theano, Tensorflow and Keras. Another important feature of Python is that it can be assembled on any platform like SQL server, a MongoDB database or JSON. Python can also handle text data very well.

3.3.4 SAS:

SAS is a programming environment and language for data manipulation and a leader in analytics, developed by the SAS Institute in 1966 and further developed in 1980's and 1990's. SAS is easily accessible, manageable and can analyze data from any sources. SAS introduced a large set of products in 2011 for customer intelligence and numerous SAS modules for web, social media and marketing analytics that is widely used for profiling customers and prospects. It can also predict their behaviors, manage, and optimize communications.

3.3.5 Excel:

Excel is a basic, popular and widely used analytical tool almost in all industries. Excel becomes important when there is a requirement of analytics on the client's internal data. It analyzes the complex task that summarizes the data with a preview of pivot tables that helps in filtering the data as per client requirement. Excel has the advance business analytics option which helps in modeling capabilities which have prebuilt options like automatic relationship detection, a creation of DAX measures and time grouping.

3.3.6 RapidMiner:

RapidMiner is a powerful integrated data science platform developed by the same company that performs predictive analysis and other advanced analytics like data mining, text analytics, machine learning and visual analytics without any programming. RapidMiner can incorporate with any data source types, including Access, Excel, Microsoft SQL, Tera data, Oracle, Sybase, IBM DB2, Ingres, MySQL, IBM SPSS, Dbase etc. The tool is very powerful that can generate analytics based on real-life data transformation settings, i.e. the formats and data sets for predictive analysis can be controlled.

3.3.7 QlikView:

QlikView has many unique features like patented technology and has in-memory data processing, which executes the result very fast to the end users and stores the data in the report itself. Data association in QlikView is automatically maintained and can be compressed to almost 10% from its original size. Data relationship is visualized using colors – a specific color is given to related data and another color for non-related data.

3.3.8 Power BI:

Power BI is a business analytics service provided by Microsoft. It provides interactive visualizations with self-service business intelligence capabilities, where end users can create reports and dashboards by themselves, without having to depend on information technology staff or database administrators.

3.4 Real Time Applications

6.1 Education Industry:

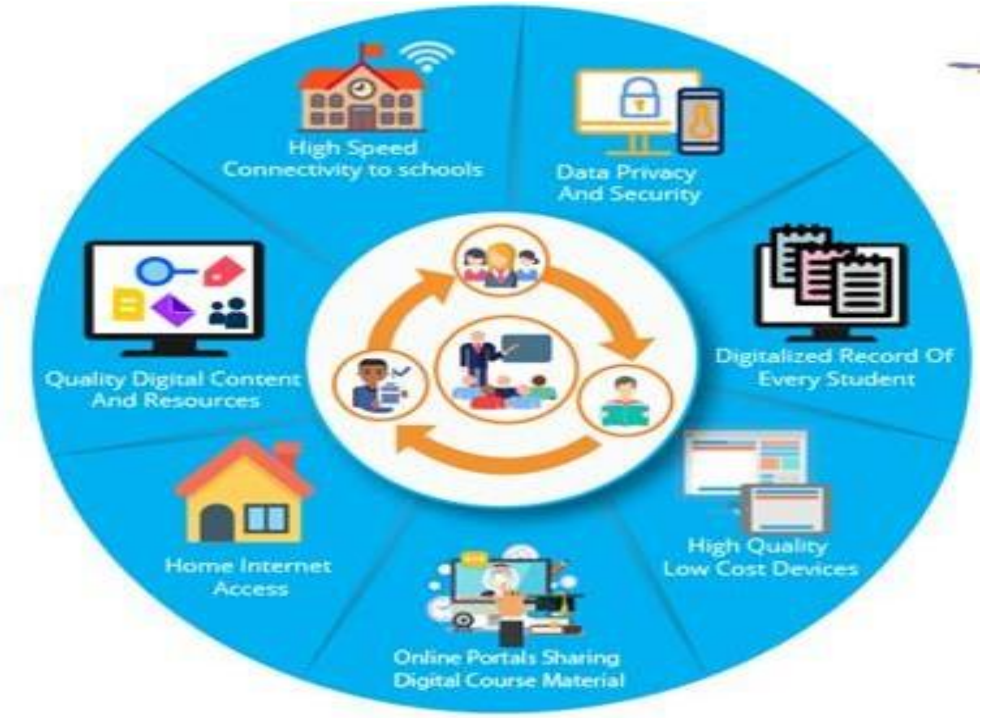

High Speed Connectivity to schools
Data Privacy And Security
Digitalized Record Of Every Student
High Quality Low Cost Devices
Online Portals Sharing Digital Course Material
Home Internet Access
Quality Digital Content And Resources

Fig.1.8: Analytics in Education Industry

Education Industry is flooding with a huge amount of data related to students, faculties, courses, results and what not. The proper study and analysis of this data provides insights that can be used to improve the operational effectiveness and working of educational institutes.

Following are some of the fields in education industry that have been transformed by data motivated changes.

- **Customized and dynamic learning programs:**

  Customized programs and schemes for each individual can be created using the data collected on the bases of a student's learning history to benefit all students. This improves the overall student results.

- **Reframing course material:**

Reframing the course material according to the data that is collected on the basis of what student learns and to what extent by real time monitoring of what components of a course are easier to understand.

- **Grading Systems:**

  New advancements in grading systems have been introduced as a result of proper analysis of student data.

- **Career prediction:**

  Proper analysis and study of every student's records will help in understanding the student's progress, strengths, weaknesses, interests and more. It will help in determining which career would be most appropriate for the student in the future.

Example:

The University of Alabama has more than 38000 students and an ocean of data. In the past when there were no real solutions to analyze that much data, some of that data seemed useless. Now administrators are able to use analytics and data visualizations for this data to draw out patters with students revolutionizing the university's operations, recruitment and retention efforts.

Tracking students' performance across cohort, departments and courses and creating clusters based on different characteristics enables targeted strategies for specific segments of students. Such as Students pursuing a particular course and performing exceptionally well on average or below average students finding the course very tough. For the below average cluster, the University administration can initiate structures intervention and provide them some special training to ensure retention and improved performance.

Analyzing the attendance data and focusing on students who miss the assigned course credit can help identify likely dropouts. Specific actions or retention programs for such students can have a significant impact on dropout rates.

6.2 Healthcare Industry:

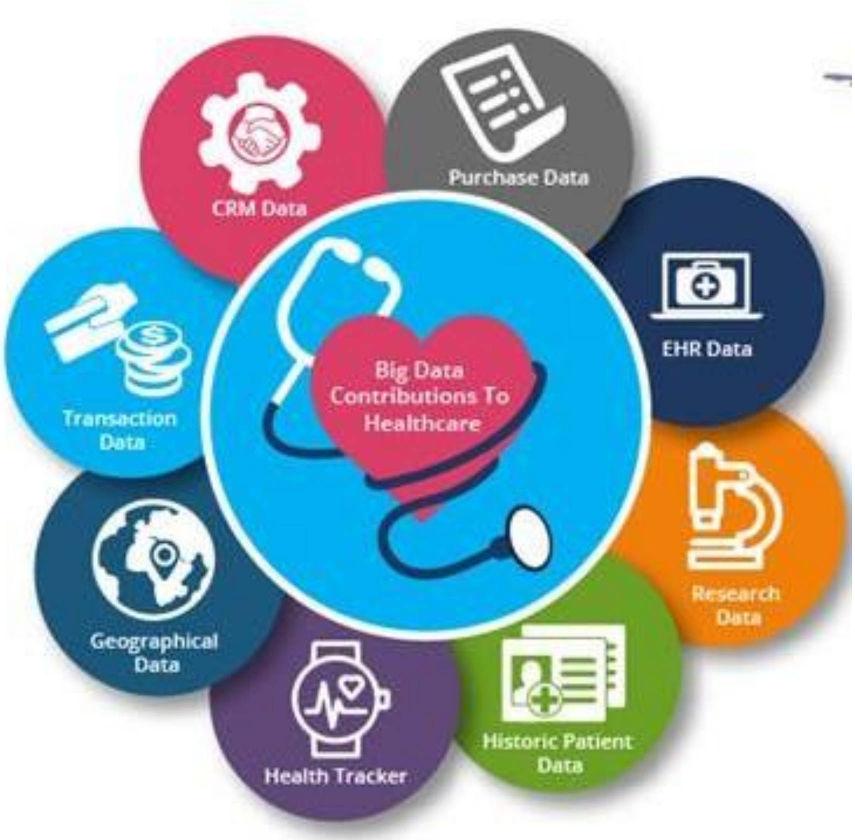


Fig. 1.9: Analytics in Healthcare Industry

Healthcare is yet another industry which is bound to generate a huge amount of data. Following are some of the ways in which big data has contributed to healthcare.

- Big data reduces costs of treatment since there is less chances of having to perform unnecessary diagnosis.
- It helps in predicting outbreaks of epidemics and also helps in deciding what preventive measures could be taken to minimize the effects of the same.

- It helps avoid preventable diseases by detecting diseases in early stages and prevents it from getting any worse which in turn makes the treatment easy and effective.
- Patients can be provided with the evidence based medicine which is identified and prescribed after doing the research of past medical results.

Example:

Wearable devices and sensors have been introduced in healthcare industry which can provide real time feed to the electronic health record of a Patient. One such technology is from Apple.

Apple has come up with what they call Apple Health Kit, Care Kit and Research Kit. The main goal is to empower the iPhone users to store and access their real time health records on their phones.

6.3 Government Industry:

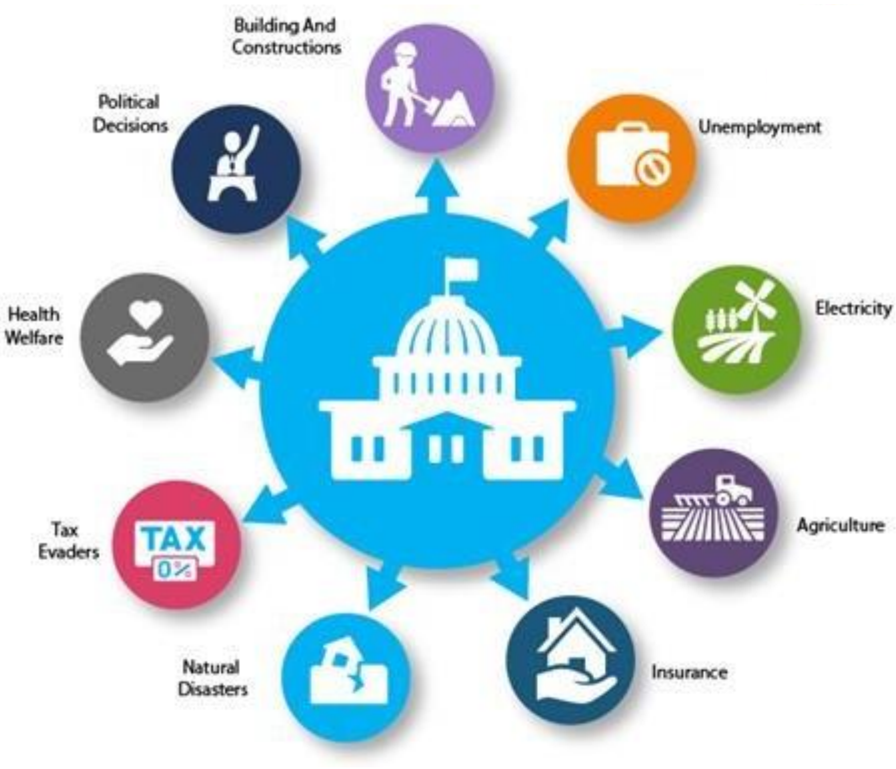


Fig. 1.10: Analytics in Government Industry

Governments, be it of any country, come face to face with a very huge amount of data on almost daily basis. Reason being, they have to keep track of various records and databases regarding the citizens, their growth, energy resources, geographical surveys and many more. All of this data contributes to big data. The proper study and analysis of this data helps the Governments in endless ways. Few of them are:

Welfare schemes:

- In making faster and informed decisions regarding various political programs.
- To identify the areas that is in immediate need of attention.
- To stay up-to-date in the field of agriculture by keeping track of all the land and livestock that exists.
- To overcome national challenges such as unemployment, terrorism, energy resource exploration and more.

Cyber security:

- Big Data is hugely used for deceit recognition
- Governments are also finding the use of big data in catching tax evaders.

Example:

The Food and Drug Administration (FDA) which runs under the jurisdiction of the Federal Government of US leverages from the analysis of big data to discover patters and associations in order to identify and examine the expected or unexpected occurrences of food based infections.

6.4 Media and Entertainment Industry:

With people having access to various digital gadgets the generation of large amount of data is inevitable and this is main cause of rise in big data in media and entertainment industry.

Other than this, social media platforms are also another way in which huge amount of data is being generated. Although business in media and entertainment industry have realized the importance of this data and they have been able to leverage from it to help their businesses grow.

Some of the benefits extracted from the big data in media and entertainment industry:

- Predicting the interests of audiences.
- Optimized or on-demand scheduling of media streams in digital media distribution platforms.
- Getting Insights into customer’s reviews and pin pointing their animo sities.
- Effective targeting of the advertisements for media.

Example:

Spotify, which is an on-demand music providing platform, uses big data analytics and collects data from all of the users around the globe and then uses the analyzed data to give informed music recommendations and suggestions to every individual user.

Amazon Prime, that offers, videos, music and Kindle books in a one-stop shop is also big on using big data.

6.5 Transportation Industry:

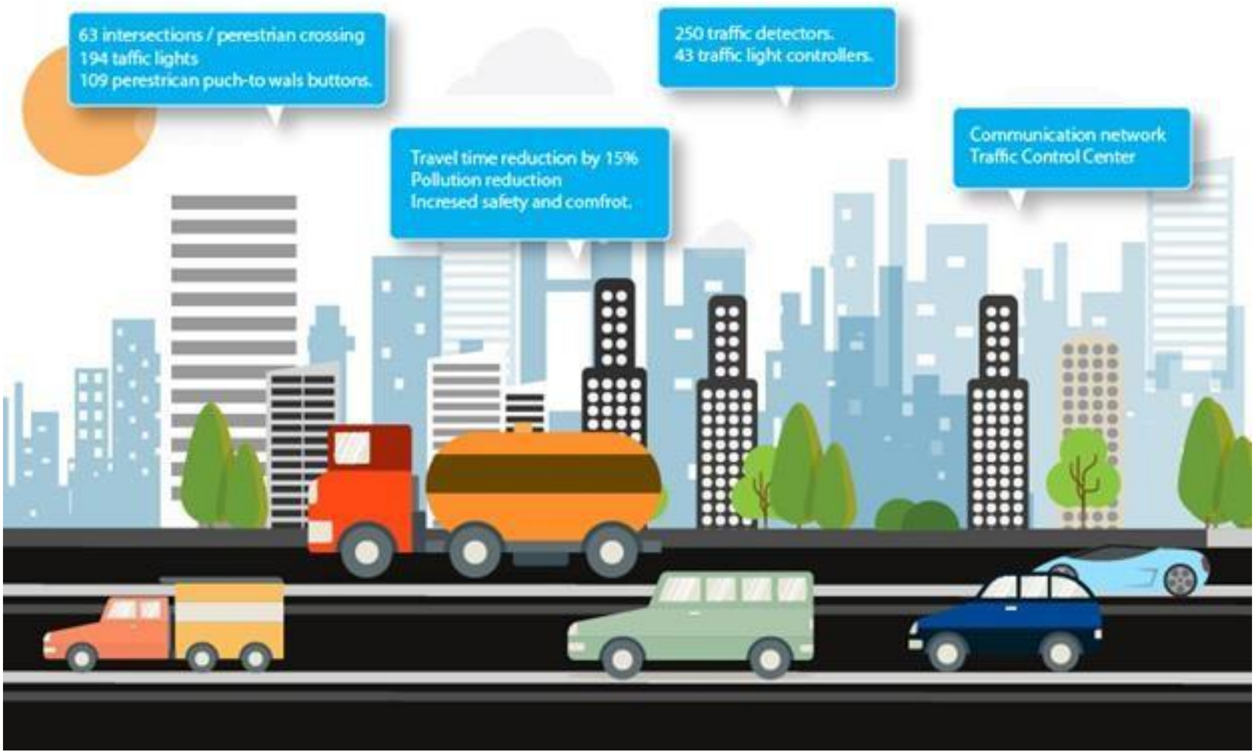

63 intersections / perestrian crossing
194 taffic lights
109 perestrican puch-to wals buttons.
250 traffic detectors.
43 traffic light controllers.
Travel time reduction by 15%
Pollution reduction
Incresed safety and comfrot.
Communication network
Traffic Control Center

Fig. 1.11: Analytics in Transportation Industry

Since the rise of big data, it has been used in various ways to make transportation more efficient and easy. Following are some of the areas where big data contributed to transportation.

- **Route planning:** Big data can be used to understand and estimate the user’s needs on different routes and on multiple modes of transportation and then utilizing route planning to reduce the users wait times.
- **Congestion management and traffic control:** Using big data, real time estimation of congestion and traffic patterns is now possible. For examples, people using Google Maps to locate the least traffic prone routes.
- **Safety level of traffic:** Using the real time processing of big data and predictive analysis to identify the traffic accidents prone areas can help reduce accidents and increase the safety level of traffic.

Example:

User generates and uses a huge amount of data regarding drivers, their vehicles, locations, every trip from every vehicle etc. All of this data is analyzed and then used to predict the supply, demand, location of the drivers and the fares that will be set for every trip.

6.6 Banking Sector:

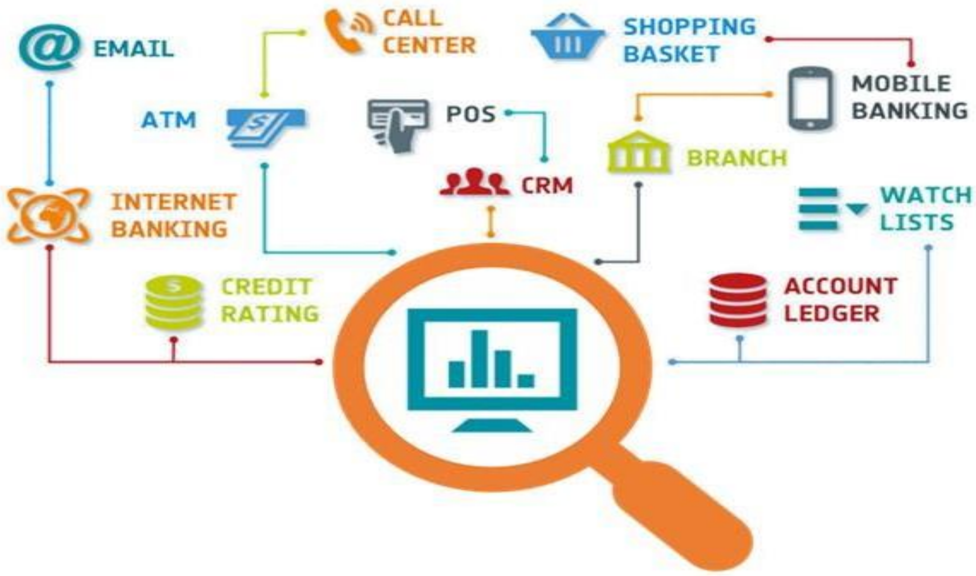


Fig. 1.12: Analytics in Banking Sector

The amount of data in banking sectors is skyrocketing every second. According to GDC prognosis, this data is estimated to grow 700% by 2020. Proper study and analysis of this data can help detect any and all the illegal activities that are being carried out such as,

- The misuse of credit cards
- Misuse of debit cards
- Venture credit hazard treatment
- Business clarity
- Customer statistics alteration
- Money laundering
- Risk Mitigation

Example:

Anti-money laundering software such as SAS AML and Actimize are deployed by various financial enterprises for the main purpose of detecting suspicious transactions and analyzing customer data. One such financial enterprise is the Bank of America who have been a SAS AML customer for more than 25 years.

6.7 Weather Patterns:

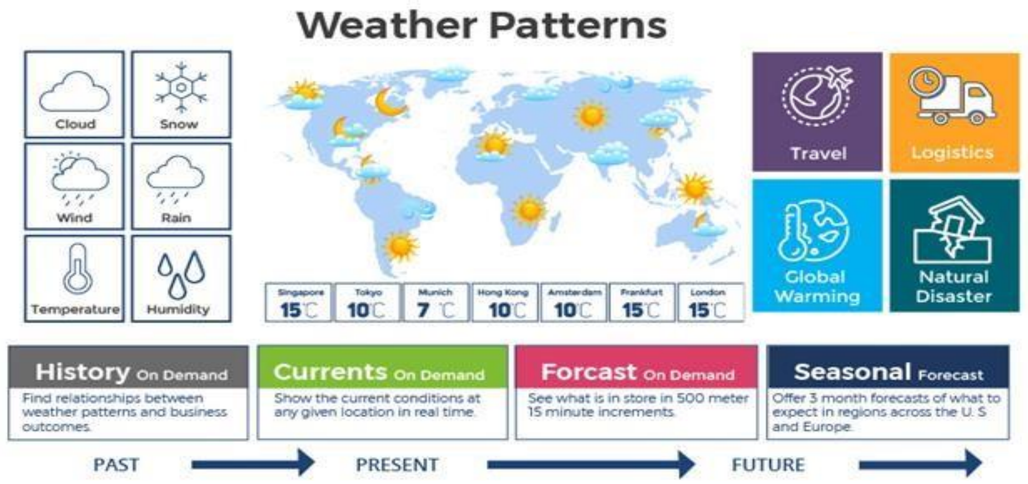

Weather Patterns
Cloud
Snow
Wind
Rain
Temperature
Humidity
Singapore 15℃
Tokyo 10℃
Munich 7 ℃
Hong Kong 10℃
Amsterdam 10℃
Frankfurt 15℃
London 15℃
Travel
Logistics
Global Warming
Natural Disaster
History On Demand
Find relationships between weather patterns and business outcomes.
Currents On Demand
Show the current conditions at any given location in real time.
Forcast On Demand
See what is in store in 500 meter 15 minute increments.
Seasonal Forecast
Offer 3 month forecasts of what to expect in regions across the U. S and Europe.
PAST
PRESENT
FUTURE

Fig. 1.13: Analytics in Weather Patterns

There are weather sensors and satellites deployed all around the globe. A huge amount of data is collected from them and then this data is used to monitor the weather and environmental conditions.

All of the data collected from these sensors and satellites contributes to big data and can be used in different ways such as:

- In weather forecast
- To study global warming
- Understanding the patterns of natural disasters
- To make necessary preparations in case of crisis
- To predict the availability of usable water around the world.

Example:

IBM deep thunder which is a research project by IBM provides weather forecasting through high performance computing of big data. IBM is also assisting Tokyo with the improved weather forecasting for natural disasters or probability of damaged power lines in order to plan successful 2020 Olympics.

3.5 Specific Requirements Description

3.5.1 Overall Description

This section and its subsections contain the description of the project components such as interfaces, performance requirements, design constraints, assumptions and dependencies etc.

3.5.2 Product Perspective

The application will be a Windows / Linux based product.

3.5.3 System Interfaces

List each system interface and identify the functionality of the system (hardware and software both) to accomplish the system requirement and interface description to match the system.

3.5.4 Hardware Interfaces

- Screen resolution of at least 800 x 600 pixels is required for proper and complete viewing of screens. Higher resolutions in wide-screen mode will be better for a better view.
- Support for printer (dot-matrix / desk jet / inkjet / laser) is required. This implies that appropriate drivers should be installed and printer device should be connected for printing of reports and mark sheets.
- A network connection (internet / intranet) is required to make the web service accessible on other systems connected over the network.
- Other hardware interface specifications are as follows

HARDWARE INTERFACES (Minimum)

HARDWARE	RAM	DISK SPACE
Intel Core i5 / i7 2.27 GHz and higher Or AMD 4XXX and higher	2048 MB	20 GB

Table 3.5.1 – Minimum Hardware Interfaces

HARDWARE INTERFACES (Recommended)

HARDWARE	RAM	DISK SPACE
Intel Xeon higher Or AMD equivalent	4096 MB	40 GB

Table 3.5.2 – Recommended Hardware Interfaces

3.5.5 Software Interfaces

- Any Microsoft Windows 7 and higher (Windows 7 / 8 / 8.1 / 10) or equivalent Linux based operating system with minimum kernel support3.X.
- Crystal Reports 8 for generation and viewing of reports.
- Anaconda Environment, pgAdmin, Google Analytics & Microsoft Excel for coding and developing of the application.

SOFTWARE INTERFACES (Minimum)

Software Tool	Version	Purpose of Use
Operating system	Windows 7 and higher or Linux with Kernel 3.x and higher	Installation and operational platform
Web Browser	Internet Explorer 6 and other higher compatible	Access to the web application
Application	Anaconda	Development of algorithms & Visualization of insights
Database	Google Analytics, PostreSQL & pgAdmin	Fetching required database & linking it with R

Table 3.5.3 – Minimum Software Interfaces

SOFTWARE INTERFACES (Recommended)

Software Tool	Version	Purpose of Use
Operating system	Windows 8 &higher or Linux with Kernel 4.x & higher	Installation and operational platform
Web Browser	Internet Explorer 11 and other higher compatible	Access to the web application
Application	Anaconda	Development of algorithms & Visualization of insights
Database	Google Analytics, PostreSQL & pgAdmin	Fetching required database & linking it with R

Table 3.5.4– Recommended Software Interfaces

3.5.6 Memory Constraints

- At least 2048 MB of RAM and 20 GB of space on hard disk will be required for running the application.

3.5.7 Operations

- The DBA will be assumed responsible for manually deleting or achieving obsolete or non-required data from the database as per client’s requirements.

- This will include database backup and recovery options also.

- The algorithms developed accept dynamic data so with every change in data the insights received will be changed henceforth bringing changes in visualizations.

- Proper constraints on data fetched by Google Analytics have to be put before analysis.

3.5.8 Site Adaptations

The computing terminals connected to network (internet / intranet) will be required to support the hardware and software interfaces specified in above sections.

3.5.9 Project Functions

Only authorized users can fetch data from the dump for analysis and deriving insights. Depending upon the user's role, he / she will be able to access only specific part of the database.

A summary of the major functions that the developed algorithms in the project will perform:

1. The location data will be clustered on the basis of distance, regions, preferred population type, purchasing behavior, etc. so as to perform location clustering.
2. The clustered data will be used in framing online marketing strategies so as to target the high potential location.
3. K-Means Clustering will be used for grouping the locations into clusters based upon the common characteristic.

3.5.10 User Characteristics

- *Educational Level:* User should be at least graduate and comfortable with English.
- *Experience:* User should be well versed / informed about the structure of the program. Data entry and modification can be done only by the user authorized for this job.
- *Technical Expertise:* User should be comfortable using general purpose applications on a computer.

3.5.11 Constraints

Since the DBMS being used in this project is PostgreSQL, and Anaconda which are free open source tools, the technologies are out of any guarantees, unless specifically purchased for enterprise environment.

3.5.12 System Product Features

- **Security**

  The whole database fetching, analysis & visualization will be password protected. Users will have to enter correct username, password and role in order to access the database modules allowed to their privilege.

- **Maintainability**

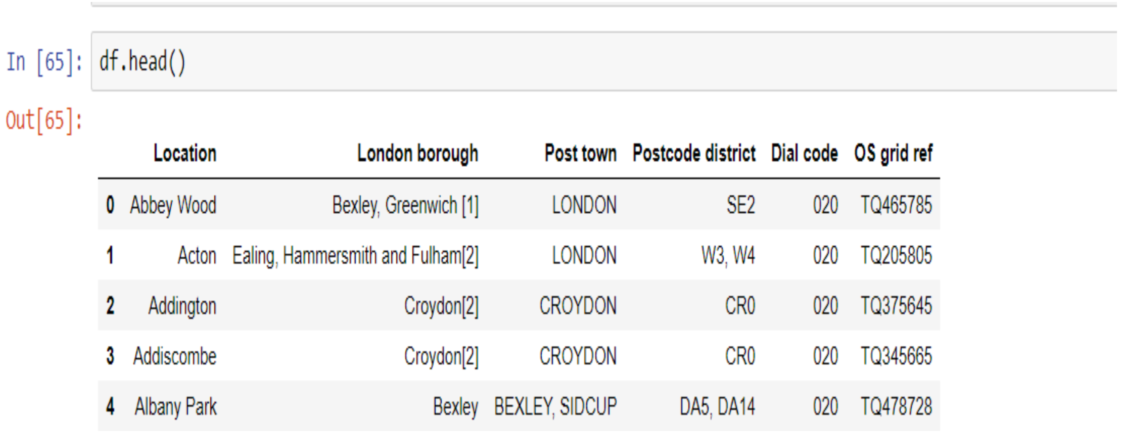
  The algorithm can accept changes in data hence maintaining its credibility.

- **Portability**

  The algorithms will be easily portable among any windows or linux based systems that have Anaconda and pgAdmin installed.

3.6 Coding and Project Data

- The coding for complete Neighborhood Recommendation was done through PYTHON programming language.
- The real time data was extracted from the WIKIPEDIA. It contained various features of location like Areas of interest, distance from popular places etc.
- Data visualization was used to predict and design meaningful strategies to target more and more customer and to increase the company's profit.
- A snapshot of the data we have is presented below.



```
In [65]: df.head()
Out[65]:
```

	Location	London borough	Post town	Postcode district	Dial code	OS grid ref
0	Abbey Wood	Bexley, Greenwich [1]	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham[2]	LONDON	W3, W4	020	TQ205805
2	Addington	Croydon[2]	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon[2]	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728

Fig. **Snapshot of data obtained from Wikipedia**

Chapter 4

Results

4.1 Results for NEIGHBOURHOOD RECOMMENDER

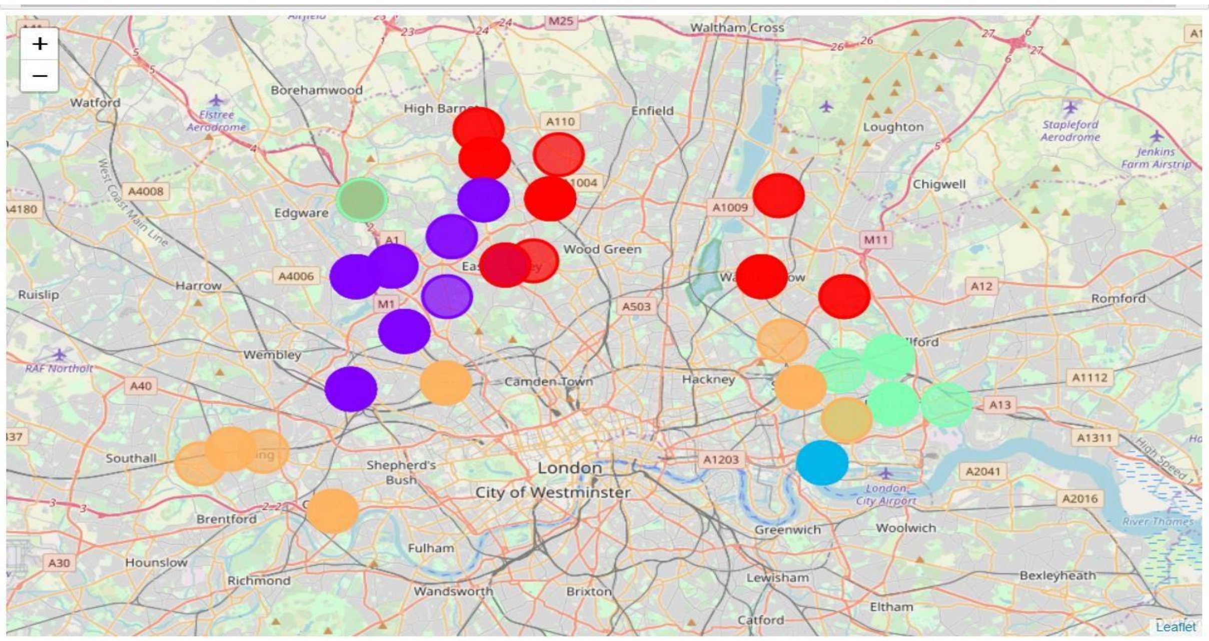
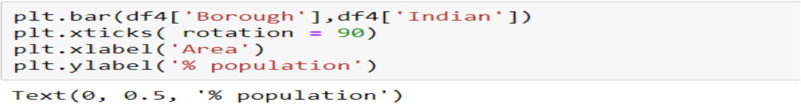


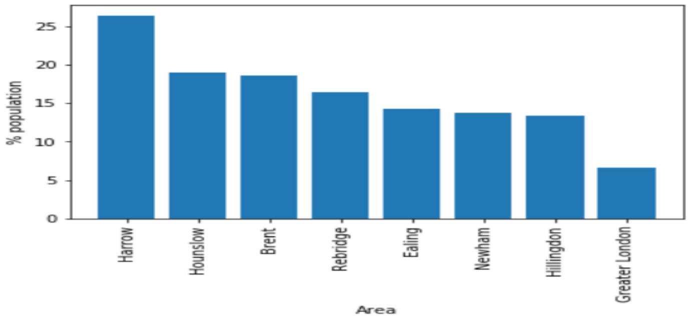
Fig 4.1: segmented locations clusters

- There are 5 observable clusters in the above figure.
- The main targets are the places with high Indian population but low restaurant density.
- It was identified that locations which were falling under the blue clusters are most suitable for opening a new Indian restaurant. This deduction was made on the basis of correlation coefficient between population density and restaurant density.



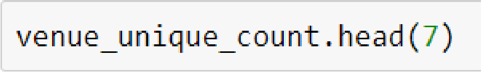
```
plt.bar(df4['Borough'],df4['Indian'])
plt.xticks( rotation = 90)
plt.xlabel('Area')
plt.ylabel('% population')
```

```
Text(0, 0.5, '% population')
```

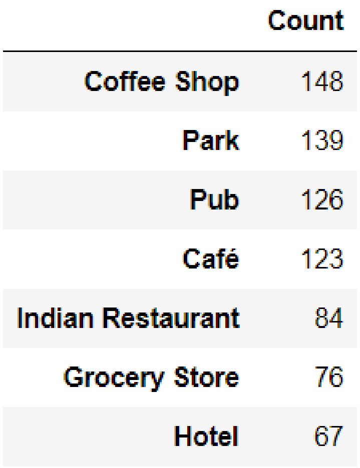


The above figure shows about the percentages of Indian population living in London area.

These are the various no. of venues popular in our interested pin code locations in London.



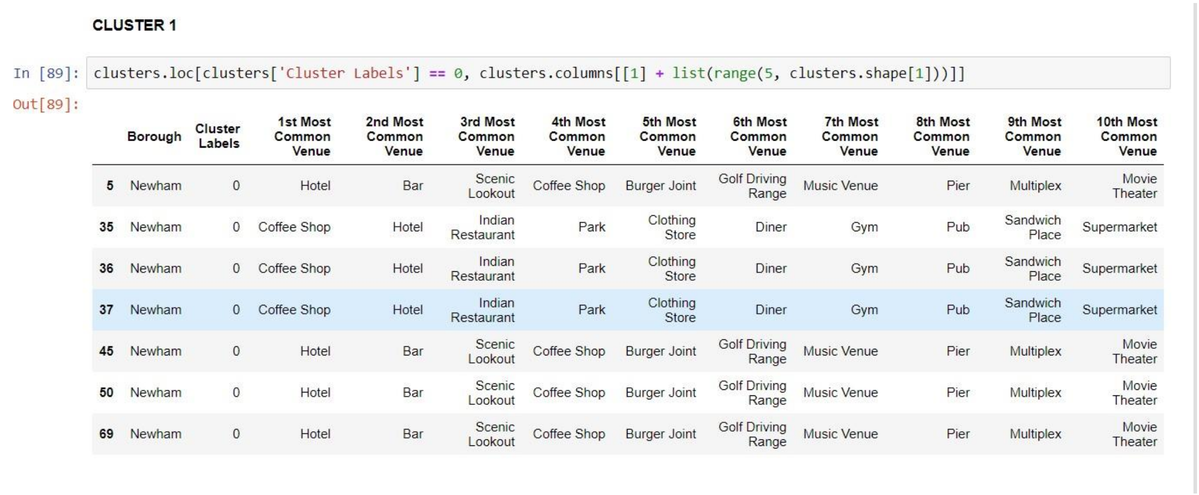
```
venue_unique_count.head(7)
```



	Count
Coffee Shop	148
Park	139
Pub	126
Café	123
Indian Restaurant	84
Grocery Store	76
Hotel	67

4.2 Results for clusters

In the first cluster we can see that the 3rd most preference or common venue in Indian restaurant so this is not suitable for us to open our outlet in this area.



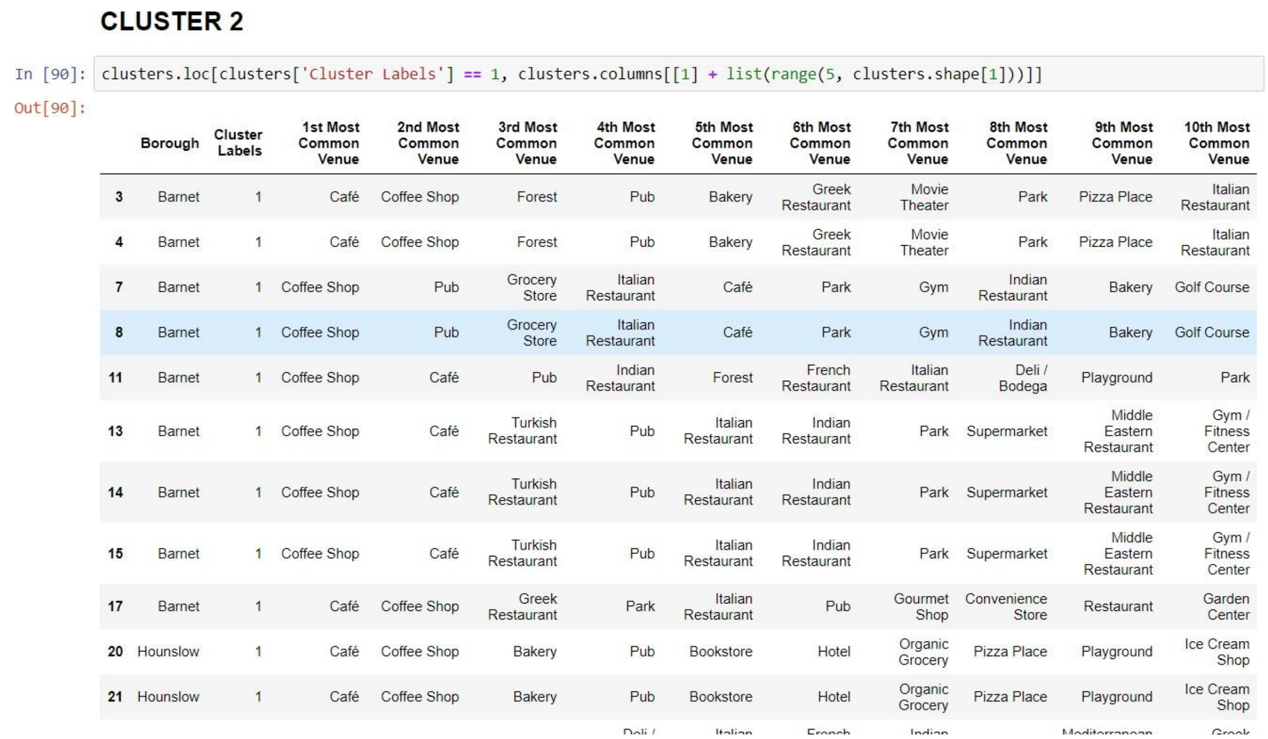
CLUSTER 1

```
In [89]: clusters.loc[clusters['Cluster Labels'] == 0, clusters.columns[[1] + list(range(5, clusters.shape[1]))]]
Out[89]:
```

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	Newham	0	Hotel	Bar	Scenic Lookout	Coffee Shop	Burger Joint	Golf Driving Range	Music Venue	Pier	Multiplex	Movie Theater
35	Newham	0	Coffee Shop	Hotel	Indian Restaurant	Park	Clothing Store	Diner	Gym	Pub	Sandwich Place	Supermarket
36	Newham	0	Coffee Shop	Hotel	Indian Restaurant	Park	Clothing Store	Diner	Gym	Pub	Sandwich Place	Supermarket
37	Newham	0	Coffee Shop	Hotel	Indian Restaurant	Park	Clothing Store	Diner	Gym	Pub	Sandwich Place	Supermarket
45	Newham	0	Hotel	Bar	Scenic Lookout	Coffee Shop	Burger Joint	Golf Driving Range	Music Venue	Pier	Multiplex	Movie Theater
50	Newham	0	Hotel	Bar	Scenic Lookout	Coffee Shop	Burger Joint	Golf Driving Range	Music Venue	Pier	Multiplex	Movie Theater
69	Newham	0	Hotel	Bar	Scenic Lookout	Coffee Shop	Burger Joint	Golf Driving Range	Music Venue	Pier	Multiplex	Movie Theater

Fig 4.2: Cluster 1

In the second cluster we can see that the 4th and the 8th most common venue are Indian restaurant so it is also not suitable for us to open our outlets near these areas.



CLUSTER 2

```
In [90]: clusters.loc[clusters['Cluster Labels'] == 1, clusters.columns[[1] + list(range(5, clusters.shape[1]))]]
Out[90]:
```

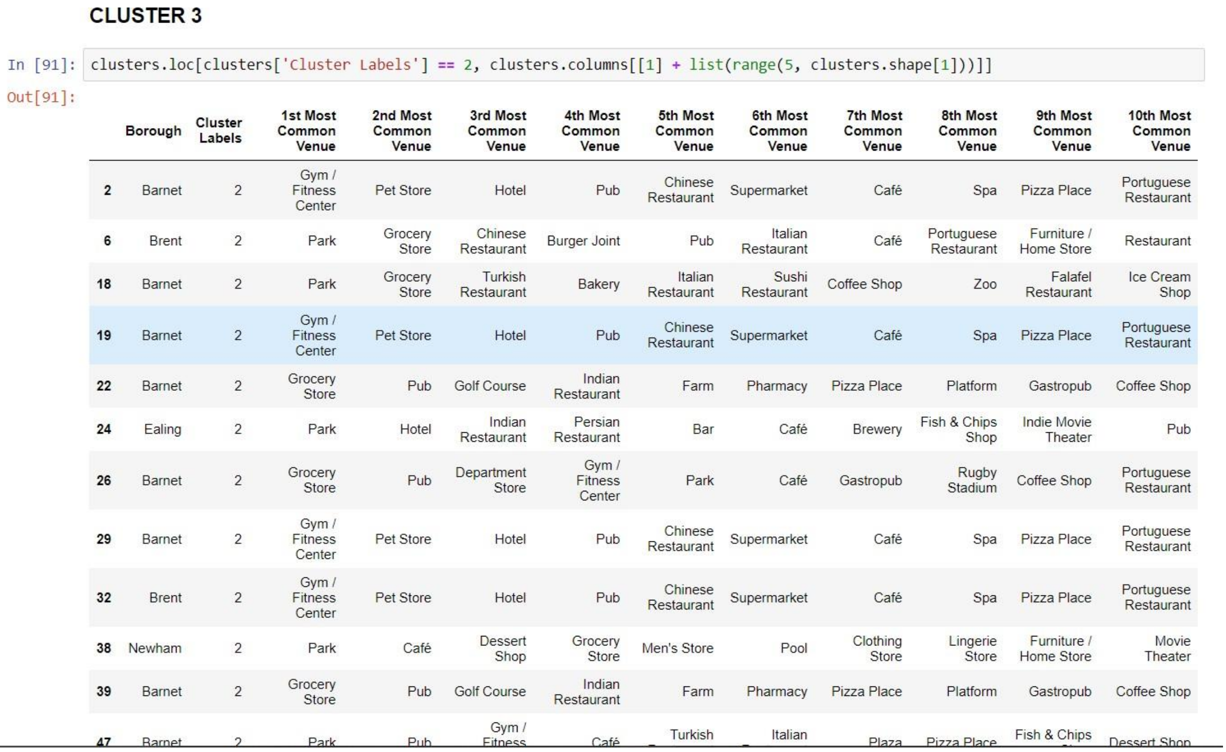
	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	Barnet	1	Café	Coffee Shop	Forest	Pub	Bakery	Greek Restaurant	Movie Theater	Park	Pizza Place	Italian Restaurant
4	Barnet	1	Café	Coffee Shop	Forest	Pub	Bakery	Greek Restaurant	Movie Theater	Park	Pizza Place	Italian Restaurant
7	Barnet	1	Coffee Shop	Pub	Grocery Store	Italian Restaurant	Café	Park	Gym	Indian Restaurant	Bakery	Golf Course
8	Barnet	1	Coffee Shop	Pub	Grocery Store	Italian Restaurant	Café	Park	Gym	Indian Restaurant	Bakery	Golf Course
11	Barnet	1	Coffee Shop	Café	Pub	Indian Restaurant	Forest	French Restaurant	Italian Restaurant	Deli / Bodega	Playground	Park
13	Barnet	1	Coffee Shop	Café	Turkish Restaurant	Pub	Italian Restaurant	Indian Restaurant	Park	Supermarket	Middle Eastern Restaurant	Gym / Fitness Center
14	Barnet	1	Coffee Shop	Café	Turkish Restaurant	Pub	Italian Restaurant	Indian Restaurant	Park	Supermarket	Middle Eastern Restaurant	Gym / Fitness Center
15	Barnet	1	Coffee Shop	Café	Turkish Restaurant	Pub	Italian Restaurant	Indian Restaurant	Park	Supermarket	Middle Eastern Restaurant	Gym / Fitness Center
17	Barnet	1	Café	Coffee Shop	Greek Restaurant	Park	Italian Restaurant	Pub	Gourmet Shop	Convenience Store	Restaurant	Garden Center
20	Hounslow	1	Café	Coffee Shop	Bakery	Pub	Bookstore	Hotel	Organic Grocery	Pizza Place	Playground	Ice Cream Shop
21	Hounslow	1	Café	Coffee Shop	Bakery	Pub	Bookstore	Hotel	Organic Grocery	Pizza Place	Playground	Ice Cream Shop
						Deli /	Italian	French	Indian		Mediterranean	Greek

Fig 4.3: Cluster 2

- Based on above interpretations offers for suitable places were made to the company.

- The offers were made by a product recommendation system which was implemented using machine learning algorithm.

In the 3rd cluster we can observe that in the top 10 common venue there are very less number of Indian restaurant so it is suitable for us to open our outlet here so that it is more profitable and successful.



CLUSTER 3

```
In [91]: clusters.loc[clusters['Cluster Labels'] == 2, clusters.columns[[1] + list(range(5, clusters.shape[1]))]]
Out[91]:
```

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Barnet	2	Gym / Fitness Center	Pet Store	Hotel	Pub	Chinese Restaurant	Supermarket	Café	Spa	Pizza Place	Portuguese Restaurant
6	Brent	2	Park	Grocery Store	Chinese Restaurant	Burger Joint	Pub	Italian Restaurant	Café	Portuguese Restaurant	Furniture / Home Store	Restaurant
18	Barnet	2	Park	Grocery Store	Turkish Restaurant	Bakery	Italian Restaurant	Sushi Restaurant	Coffee Shop	Zoo	Falafel Restaurant	Ice Cream Shop
19	Barnet	2	Gym / Fitness Center	Pet Store	Hotel	Pub	Chinese Restaurant	Supermarket	Café	Spa	Pizza Place	Portuguese Restaurant
22	Barnet	2	Grocery Store	Pub	Golf Course	Indian Restaurant	Farm	Pharmacy	Pizza Place	Platform	Gastropub	Coffee Shop
24	Ealing	2	Park	Hotel	Indian Restaurant	Persian Restaurant	Bar	Café	Brewery	Fish & Chips Shop	Indie Movie Theater	Pub
26	Barnet	2	Grocery Store	Pub	Department Store	Gym / Fitness Center	Park	Café	Gastropub	Rugby Stadium	Coffee Shop	Portuguese Restaurant
29	Barnet	2	Gym / Fitness Center	Pet Store	Hotel	Pub	Chinese Restaurant	Supermarket	Café	Spa	Pizza Place	Portuguese Restaurant
32	Brent	2	Gym / Fitness Center	Pet Store	Hotel	Pub	Chinese Restaurant	Supermarket	Café	Spa	Pizza Place	Portuguese Restaurant
38	Newham	2	Park	Café	Dessert Shop	Grocery Store	Men's Store	Pool	Clothing Store	Lingerie Store	Furniture / Home Store	Movie Theater
39	Barnet	2	Grocery Store	Pub	Golf Course	Indian Restaurant	Farm	Pharmacy	Pizza Place	Platform	Gastropub	Coffee Shop
47	Barnet	2	Park	Pub	Gym / Fitness	Café	Turkish	Italian	Plaza	Pizza Place	Fish & Chips	Dessert Shop

Fig. Cluster 3

Chapter 5

Conclusion

5.1 Existing System

The Existing database is huge with less or no insights driven hence the strategies so formed to gain customer base & to increase profits are less effective.

5.2 Proposed System

The NEIGHBOURHOOD RECOMMENDER project will help the new businesses to decide its new locations for opening another branch.

Given the extremely large size and the population of the city, our customer wants to identify the best neighbourhood area to open its first Indian Restaurant covering the majority of the population and facing least competition from other restaurants.

5.3 Future Scope

The data will then be integrated with the machine learning algorithms to make the system self-independent in providing effective strategies.

References

[1] Shaun Bangay, “Visview: A system for the visualization of Multi-dimensional data”, in “Visual Data Exploration and Analysis V”. (page number670-672)

[2] DavidS.Ebert,RandallM.Rohrer,ChristopherD.Shaw,PradyutPanda,JamesM.Kukla,
D. Aaron Roberts, “Procedural Shape Generation for Multi-dimensional Data Visualization”, in “Data Visualization ‘99”. (page umber 222-230)

[3] Daniel Keim, “Visual Support for Query Specification and Data Mining”. (page number 930-945)

[4] Kamran Parsaye, Mark Chignell, “Intelligent Database Tools & Applications”. (page number 121-122)