**Uber data Analysis**

A

Project Report

Submitted In Partial Fulfillment of the Requirements For the Award Of

**Bachelor of Technology**

***Under Guidance Of***

**RAHUL SHARMA**

**MICROSOFT & HPE CERTIFIED TECHNICAL TRAINER**

Project Carried Out At



**Ardent Computech Pvt. Ltd.**

**(An ISO 9001:2015 Certified)**

##### SDF Building, Module No: 132, Ground Floor Sector V,

GP Block, Kolkata- 700091

**Submitted By-**

Srilopa Ghosh

**ACKNOWLEDGEMENT**

Success of any project depends largely on the encouragement and guidelines of many others. We take this sincere opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project work.

Our heartful thanks to **Avick Kr. Dey** for providing us the opportunity to develop the project at **Ardent Computech Pvt. Ltd.**

### We would like to show our greatest appreciation to **Rahul Sharma, Technical Trainer at Ardent,** Durgapur. We always feel motivated and encouraged every time by his valuable advice and constant inspiration; without his encouragement and guidance this project would not have materialized. We also want to thank them for sharing their pearls of wisdom with us during the course of this project.

We are also immensely grateful to **Rahul Sir** and **Avick Sir** for their comments on earlier versions of the manuscripts, although any errors are our own and should not tarnish the reputations of these esteemed professionals.

Words are inadequate in offering our thanks to the other trainees**, project assistants** and other members at **Ardent Computech Pvt. Ltd.** for their encouragement and cooperation in carrying out this project work. The guidance and support received from all the members and who are contributing to this project, was vital for the success of this project.

1. Title of the Project : **Uber Data Analysis**
2. Name of the Guide : **Mr. Rahul Sharma**

#### Ph.D.(Reg) M-Tech MCA M.Sc. MBA B-Tech

##### Educational Qualification of the Guide(s)

Y

1. Working / Teaching experience of the Guide**: 4 Years**
2. Software used in the Project
   * Python 3.7
   * Anaconda 3.5
   * Ms- Office
   * Draw.io

Signature of the Guide

**For Office Use Only** …………………………………

#### Signature, Designation Stamp of the Project Proposal Evaluator

Not Approved

Approved

**Date:….……………………….**

Self Certificate

This is to certify that the dissertation/project proposal entitled **“Uber Data Analysis”** is done by Srilopa Ghosh is an authentic work carried out for the partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** under the guidance of **Mr. Rahul Sharma**. The matter embodied in this project work has not been submitted earlier for award of any degree to the best of my knowledge and belief.

### Name of the Students:

1. Srilopa Ghosh :

# Certificate by Guide

This is to certify that this project entitled **“Uber Data Analysis”** submitted in partial fulfillment of the degree of **Bachelor of Technology (B.Tech)** by **Maulana Abul Kalam Azad University of Technology** through: **Ramkrishna Mahato Government Engineering College** done by Srilopa Ghosh

### Is an authentic work carried out under my guidance & best of our knowledge and belief.

Signature of Student Signature of the Guide

Date: Date:

***Certificate of Approval***

This is to certify that this documentation of **Summer Vacation Training Program 2020**, entitled **“Uber Data Analysis”** is a record of bona-fide work, carried out by Srilopa Ghosh under my supervision and guidance.

In my opinion, the report in its present form is in fulfillment of all the requirements, as specified by the **Maulana Abul Kalam Azad University of Technology** and as per regulations of the **Ardent Computech Pvt. Ltd.** In fact, it has attained the standard, necessary for submission. To the best of my knowledge, the results embodied in this report, are original in nature and worthy of incorporation in the present version of the report for **Bachelor of Technology**.

**Mr. Rahul Sharma Microsoft & HP Certified Technical Trainer Ardent Computech Pvt Ltd**

### (An ISO 9001:2015 Certified)

##### (Approved by NCVT & Ministry of HRD, Government of India)

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.NO.** | **Name of the Topic** | **Page No.** |
| **1.** | **Company Profile** | **1** |
|  |  |  |
| **2.** | **Abstract** | **2** |
|  |  |  |
| **3.** | **Introduction** | **3** |
| **4.** | **Algorithms** | 4 |
|  |  |  |
|  |  |  |
|  |  |  |
| **4.** | **Problem Statement** | **5** |
| **6.** | **Background Dataset** | **5** |
|  |  |  |
| **7.** | **What is Machine Learning** | **6** |
| **8.** | **System Analysis** | **7** |
| 8.1 | Identification of Need | 7 |
| 8.2 | Feasibility Study | 7 |
| 8.3 | Work Flow | 8 |
| 8.4 | Hardware and Software Requirements | 9 |
|  |  |  |
| **9.** | **System Design** | **10-14** |
| 9.1 | Gnatt Chart | 10 |
| 9.2 | Pert Chart | 11 |
| 9.3 | Sequence Diagram | 12-13 |
| 9.4 | Activity Diagram | 14-15 |
| 9.5 | Class Diagram | 16-17 |
| 9.6 | Use Case Diagram | 18 |
| **10.** | **User Inreface Design** | **19-21** |
|  |  |  |
| **11.** | **Implementation and Testing** | **21-25** |
| 8.1 | Introduction | 21 |
| 8.2 | Objective of Testing | 22 |
| 8.3 | Process Overview | 22-23 |
| 8.4 | Test Cases | 24 |
| 8.5 | Testing Step | 24 |
| 8.6 | Validation | 24-26 |
| 8.7 | White Box Testing | 27 |
| 8.8 | Black Box Testing | 27 |
| 8.9 | System Testing | 27 |
| 8.10 | Output Testing | 27 |
| 8.11 | User Acceptance Testing | 28 |

|  |  |  |
| --- | --- | --- |
| 8.12 | Integration Testing | 29 |
| 8.13 | Functional Testing | 29 |
|  |  |  |
| **9.** | **System Security Measures** | **30** |
| 9.1 | Database Security | 30 |
| 9.2 | System Security | 30 |
| 9.3 | Limitations | 30 |
|  |  |  |
| **11.** | **Cost Estimation** | **31-34** |
|  |  |  |
| **12.** | **Conclusion** | **35** |
|  |  |  |
| **13.** | **Future Scope & Further Enhancements** | **35** |
|  |  |  |
| **14.** | **Bibliography/Reference** | **35** |

1

1. **ARDENT COMPUTECH PVT.LTD.**

Ardent Computech Private Limited is an ISO 9001-2008 certified Software Development Company in India. It has been operating independently since 2003. It was recently merged with ARDENT TECHNOLOGIES.

Ardent Technologies

ARDENT TECHNOLOGIES is a Company successfully providing its services currently in UK, USA, Canada and India. The core line of activity at ARDENT TECHNOLOGIES is to develop customized application software covering the entire responsibility of performing the initial system study, design, development, implementation and training. It also deals with consultancy services and Electronic Security systems. Its primary clientele includes educational institutes, entertainment industries, resorts, theme parks, service industry, telecom operators, media and other business houses working in various capacities.

Ardent Collaborations

ARDENT COLLABORATIONS, the Research Training and Development Department of ARDENT COMPUTECH PVT LTD is a professional training Company offering IT enabled services & industrial trainings for B-Tech, MCA, BCA, MSc and MBA fresher’s and experienced developers/programmers in various platforms. Summer Training / Winter Training / Industrial training will be provided for the students of B.TECH, M.TECH, MBA, MCA and BCA only. Deserving candidates may be awarded stipends, scholarships and other benefits, depending on their performance and recommendations of the mentors.

Associations

Ardent is an ISO 9001:2008 company.

It is affiliated to National Council of Vocational Training (NCVT), Directorate General of Employment & Training (DGET), Ministry of Labor & Employment, and Government of India.

## ABSTRACT

This paper presents an evaluation of the critical factors influencing choice of destinations this study focuses on the internal factors that motivate tourists to choose their preference of destination. This study can assist relevant authorities and travel agencies to plan and promote the places of attraction with effective marketing strategies besides assisting tourists to decide where to go to main attractions. an attempt has been made to do comparative study of  taxi aggregators that have radically changed the way "the great Indian middle class" commutes daily- UBER. Currently, UBER cabs are following the strategy of expanding their operations and building customer base in key metropolitan cities across India. The motive is to increase market share and achieve economies of scale and at the same time providing customer satisfaction. This article seeks to understand the dynamics of India s taxi market by studying various factors like the pricing, market share, revenue models, etc. The Project is qualitative in nature and based on secondary data collected from different sources.

.

## INTRODUCTION

Uber has a massive database of drivers, so as soon as you request a car, Uber’s algorithm goes right to work – in 15 seconds or less, it matches you with the driver closest to you. In the background Uber is storing data for every trip taken — even when the driver has no passengers. All of this data is stored and leveraged to predict supply and demand, as well as [setting fares](https://cbpp.georgetown.edu/sites/cbpp.georgetown.edu/files/ConsumersurplusatUber_PR.PDF). Uber also looks at how transportation is handled across cities and tries to adjust for bottlenecks and other common issues.

Uber also gathers data on its drivers. In addition to collecting non-identifiable information about their vehicle and their location, [Uber also monitors their speed and acceleration](http://www.chicagotribune.com/news/local/breaking/ct-uber-telematics-getting-around-20161218-column.html" \t "_blank), and checks to see if they are working for a competing company as well (such as Lyft).

If you’re reading this wondering if it’s a gross invasion of privacy – you aren’t the first. But Uber is very clear about how it uses the data gathered on its platform. A section of its privacy policy for U.S. customers and drivers reads: Uber uses your personal data in an anonymised and aggregated form to closely monitor which features of the Service are used most, to analyze usage patterns and to determine where we should offer or focus our Service. We may share this information with third parties for industry analysis and statistics.While the ever-present spectre of data misuse is nipping at Uber’s heels, there’s no doubt that the anonymous, aggregated data they collect insights from is nothing short of amazing.All of this data is collected, crunched, analyzed and used to predict everything from the customer’s wait time, to recommending where drivers should place themselves via heatmap in order to take advantage of the best fares and most passengers. All of these items are implemented in real-time for both drivers and passengers alike.

## 4. ALGORITHMS

For our project ,we have applied a K-Means clustering algorithm whose main goal is to group similar elements or data points into a cluster. **“K” in K-means represents the number of clusters.**

**What is Clustering: “**Clustering is the process of dividing the datasets into groups, consisting of similar data-points”. Clustering is a type of unsupervised machine learning, which is used when you have unlabeled data.

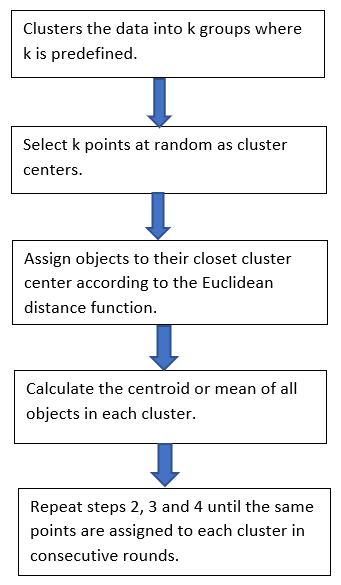
Let’s understand in the real scenario, Group of diners sitting in a restaurant. Let’s say two tables in the restaurant called T1 and T2. People in Table T1 might be related to each other, or maybe a group of family members or colleagues or anything. In a similar way, people in table T2 might be related to each other or maybe a group of colleagues or anything. But, when comparing people sitting in table T1 and people sitting in table T2, they are entirely different and not at all related to each other.

Clustering also works in the same way. Data points in one cluster are entirely different from data points in another cluster. All the points in the same cluster are either the same or related to each other.

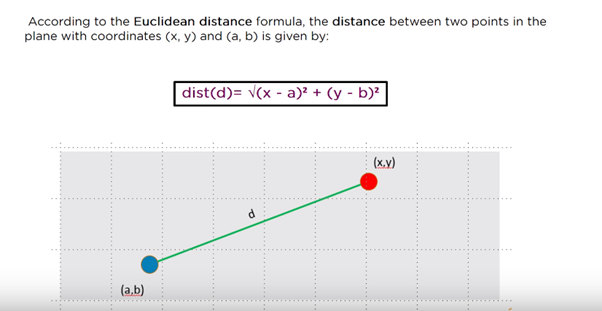
**What is K-means Clustering?**

K-Means is a clustering algorithm whose main goal is to group similar elements or data points into a cluster. **“K” in K-means represents the number of clusters.**

**K-means clustering steps:**

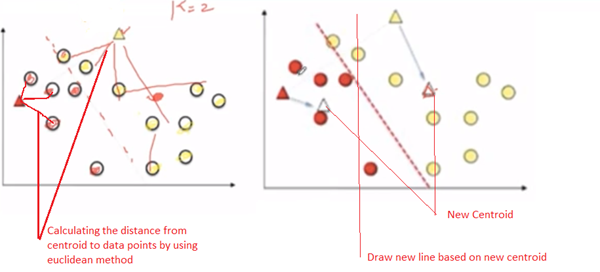


Distance measure will determine the similarity between two elements and it will influence the shape of the clusters. Normally, Euclidean distance will be used in K-Means Clustering . The Euclidean distance is the “ordinary” straight line. It is the distance between two points in Euclidean space.



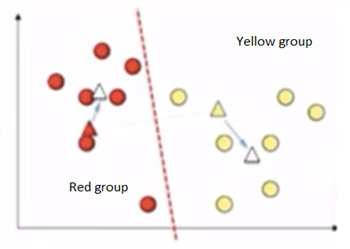
**How does K-Means Algorithm work?**

Let’s take two data points. Assume K =2. Then, it will take two random centroids anywhere in the data and based on that it will draw a line in the middle. One centroid represents in red and another one in yellow. Then all the data points towards the yellow centroid classified into yellow. All the data points towards the red centroid classified as a red. This is the first iteration step.

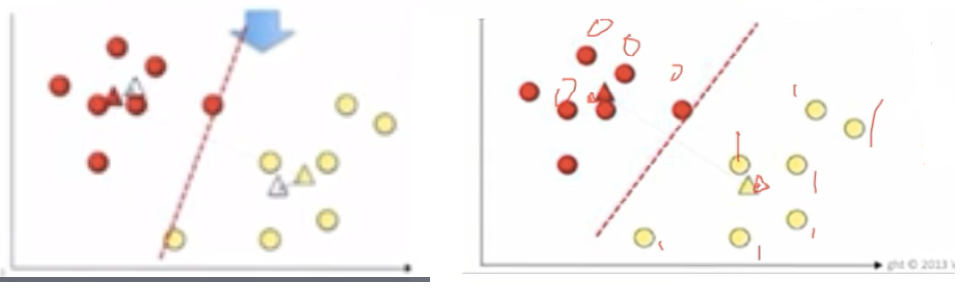


Next step, it calculates the distance from the centroid to data points by using the Euclidean method. And both red and yellow centroid points moves into new points by calculating the mean of all the points.

Again, it draws a new line between new centroid points. Automatically, all the points falling towards the red centroid will be recognized as a red group and yellow centroid will be recognized as a yellow group.



Again the same step will be repeated with new points and it will calculate new centroids. We can clearly see that data points has been moved to different groups.



It keeps on going until centroid movements become almost negligible. Then, it becomes cluster 1 and cluster 2. Here, the output Y label will give 0’s and 1’s. 0’s represent cluster 1 and 1 represent cluster 2. If it is three clusters, then it gives 0, 1 and 2.

## PROBLEM STATEMENT

This data set is a masked data set which is similar to what data analysts at Uber handle.You may have some experience of travelling to and from the airport. Have you ever used Uber or any other cab service for this travel? Did you at any time face the problem of cancellation by the driver or non-availability of cars?Well, if these are the problems faced by customers, these very issues also impact the business of Uber. If drivers cancel the request of riders or if cars are unavailable, Uber loses out on its revenue. Let’s hear more about such problems that Uber faces during its operations. We also find the number of trips for each category BUSINESS OR PERSONAL and calculate duration of each trip in minutes and number of trips per hour. Then find the purpose of travelling in a bar representation followed by mean purpose of various travels.Representation in a bar form of miles month and duration minutes .Finally, get the scatter plot in hours vs speed in KM.

## BACKGROUND DATASET

* UBER DATA Analysis is a data set that provides1156 entries of PERSONAL OR BUSINESS travels to various starts and stops including distance travelled over the trips in miles and also their purpose
* There were a total of 12204.7miles travelled
* Users of UBER were selected randomly.
* All users Travelled either business or for personal reasons to destinations

## WHAT IS ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING ?

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

The word Artificial Intelligence comprises of two words “Artificial” and “Intelligence”. Artificial refers to something which is made by human or non natural thing and Intelligence means ability to understand or think. There is a misconception that Artificial Intelligence is a system, but it is not a system .AI is implemented in the system. There can be so many definition of AI, one definition can be “It is the study of how to train the computers so that computers can do things which at present human can do better.”Therefore It is a intelligence where we want to add all the capabilities to machine that human contain.

## SYSTEM ANALYSIS

#### IDENTIFICATION OF NEED

System analysis is a process of gathering and interpreting facts, diagnosing problems and the information to recommend improvements on the system. It is a problem solving activity that requires intensive communication between the system users and system developers. System analysis or study is an important phase of any system development process. The system is studies to the minutest detail and analyzed. The system analyst plays the role of the interrogator and dwells deep into the working of the present system. The System is viewed as a whole and the input to the system are identified. The outputs from the organization are traced to the various processes. System analysis is concerned with becoming aware of the problem ,identifying the relevant and Decisional variables, analysis and synthesizing the various factors and determining an optimal or at least a satisfactory solution or program of action.

A detailed study of the process must be made by various techniques like interviews, questionnaires etc. The data collected by these sources must be scrutinized to arrive to a conclusion. The conclusion is an understanding of how the system functions. This system is called the existing system. Now the existing system is subjected to close study and problem area are identified. The designer now function as a problem solver and tries to sort out the difficulties that the enterprise faces. The solutions are given as proposals. The proposal is then weighed with the existing system analytically and the best one is selected. The proposal is presented to the user for an endorsement by the user. The proposal is reviewed on user request and suitable changes are made. This is loop that ends as soon as the user is satisfied with proposal.

#### FEASIBILITY STUDY

Feasibility study is made to see if the project on completion will serve the purpose the organization for the amount of work

Effort and time spent on it : Feasibility study lets the developer foresee the future of the project and the usefulness. A feasibility study of a system proposal is according to its workability, which is the impact on the organization, ability to meet their user needs and effective use of resources. Thus when a new application is proposed it normally goes through a feasibility study before it is approved for development.

The document provides the feasibility of the project that is being designed and lists various area that were considered very carefully during the feasibility study of this project such as Technical, Economic and operational feasibilities.

**Technical Feasibility:** This project is technically feasible as all it has got to do to extract tweets, is to get the proper credentials from the Developer’s console provided by Twitter. After the credentials are obtained, i.e. the Access Token Key, Access Token Secret Key, Consumer Key and Consumer Secret key,Twitter gives us

access to its tweets. Hence we get a sufficiently large dataset to conduct sentiment analysis. Also the range of tweets obtained is limited to 300 tweets per page,which ensures that the results do not go out of bounds. Thus, this is technically feasible.

**Economic Feasibility :** This project work is economically feasible as it does not take into account any additional costs. Whatever data is extracted, it is done without any charges. Twitter provides free use of this data that is non-encrypted and publicly available for analysis purpose.Hence, this work is economically feasible as well.

**Operational Feasibility :** This is operationally feasible as well. As already mentioned, it takes in 300 tweets per page as that is the limit set by Twitter. Therefore it is operationally feasible as well. The system won’t hang when getting the results.

#### WORK FLOW

* First Responsible for scheduling work flows and tasks. The scheduler takes into account various factors such as schedule interval, task dependencies, trigger rules and retries, and uses this information to calculate the next set of tasks to run.
* Next create a graph representing total number of trips for business or personal category
* Now, we will calculate the duration of each trips in minutes
* Then we find the duration of trips per hour and find a representation through frequency
* Finally, we find the bar graph representation of values of various purposes involved in the trips and mean of their purposes and scatter graph showing duration in hours vs speed in KM.

#### HARDWARE AND SOFTWARE REQUIREMENTS

* + 1. **Hardware Requirements**

Standard computer with at least i3 processor Standard computer with 2GB of RAM Standard computer with 100GB of free space

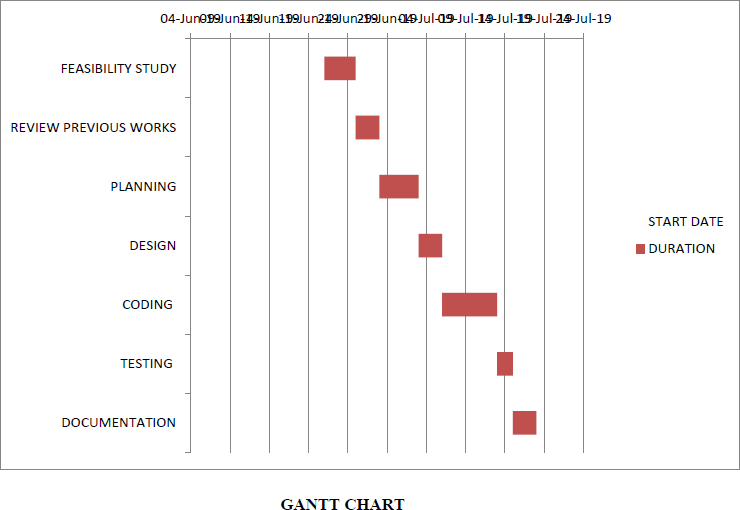
#### Software Requirements

python 3.7 Anaconda-3.5 Ms Office Draw.io

## SYSTEM DESIGN

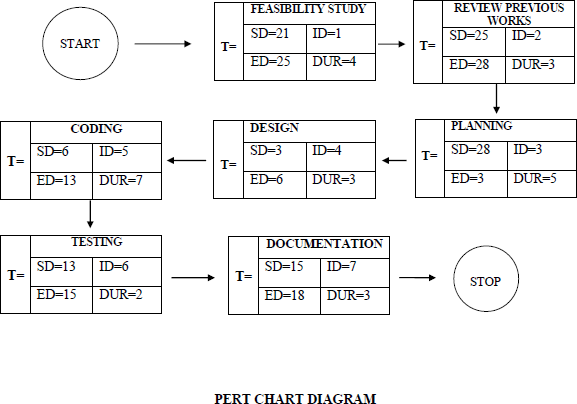
#### GANTT CHART

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | **TASK** | **START DATE** | **END DATE** | **DURATION** |
| 2 | FEASIBILITY STUDY | 21-Jun-19 | 25-Jun-19 | 4 |
| 3 | REVIEW PREVIOUS WORKS | 25-Jun-19 | 28-Jun | 3 |
| 4 | PLANNING | 28-Jun-19 | 03-Jul-19 | 5 |
| 5 | DESIGN | 03-Jul-19 | 06-Jul-19 | 3 |
| 6 | CODING | 06-Jul-19 | 13-Jul-19 | 7 |
| 7 | TESTING | 13-Jul-19 | 15-Jul-19 | 2 |
| 8 | DOCUMENTATION | 15-Jul-19 | 18-Jul-19 | 3 |

****

* 1. **PERT CHART:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | **TASK** | **START DATE** | **END DATE** | **DURATION** |
| 2 | FEASIBILITY STUDY | 21-Jul-20 | 25-Jul-20 | 4 |
| 3 | REVIEW PREVIOUS WORKS | 25-Jul-20 | 28-Jul-20 | 3 |
| 4 | PLANNING | 28-Jul-20 | 03 -Aug -20 | 5 |
| 5 | DESIGN | 03-Aug-20 | 06-J Aug -20 | 3 |
| 6 | CODING | 06- Aug -20 | 7-J Aug -20 | 2 |
| 7 | TESTING | 7- Aug 20 | 8- Aug -20 | 1 |
| 8 | DOCUMENTATION | 8- Aug -20 | 9- Aug -20 | 1 |

****

* 1. **SEQUENCE DIAGRAM**

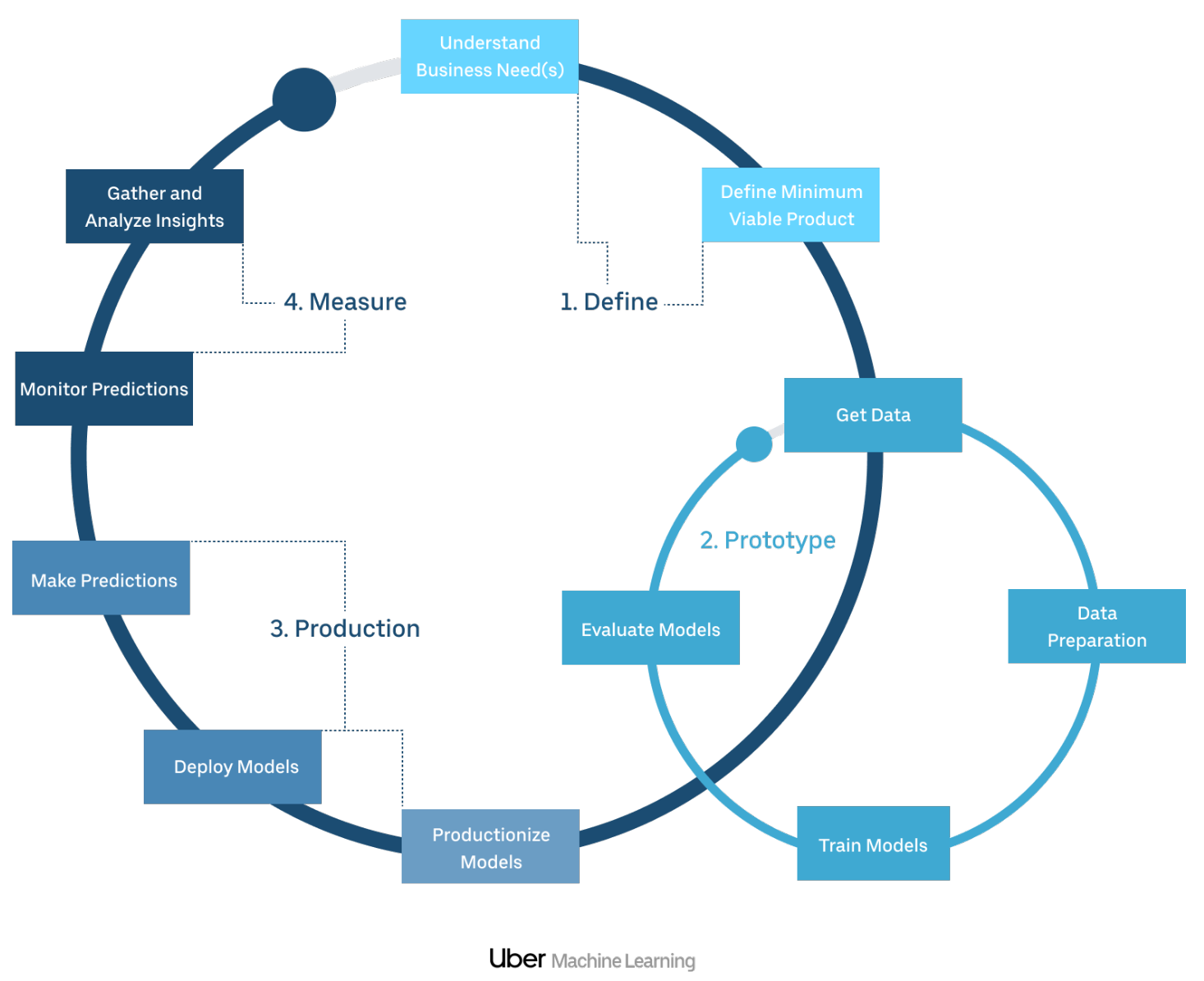
##### Sequence diagrams can be useful reference diagrams for businesses and other organizations. Try drawing a sequence diagram to:

* + - Represent the details of a UML use case.
    - Model the logic of a sophisticated procedure, function, or operation.
    - See how tasks are moved between objects or components of a process.
    - Plan and understand the detailed functionality of an existing or future scenario.

**9.3.1 Popular Sequence Diagram Uses:**

**Usage Scenario** – A usage scenario is a diagram of how your system could potentially be used. It’s a great way to make sure that you have worked through the logic of every usage scenario for the system. **Method Logic** - Just as you might use a UML sequence diagram to explore the logic of a use case, you can use it to Usage Scenario - A usage scenario is a diagram of how your system could potentially be used. It's a great explore the logic of any function, procedure, or complex process.

**Service Logic -** If you consider a service to be a high-level method used by different clients, a sequence diagram is an ideal way to map that out.



## Fig:Sequence Diagram

#### ACTIVITY DIAGRAM

Activity diagram is another important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another. This flow can be sequential, branched, or concurrent. Activity diagrams deal with all type of flow control by using different elements such as fork, join, etc.

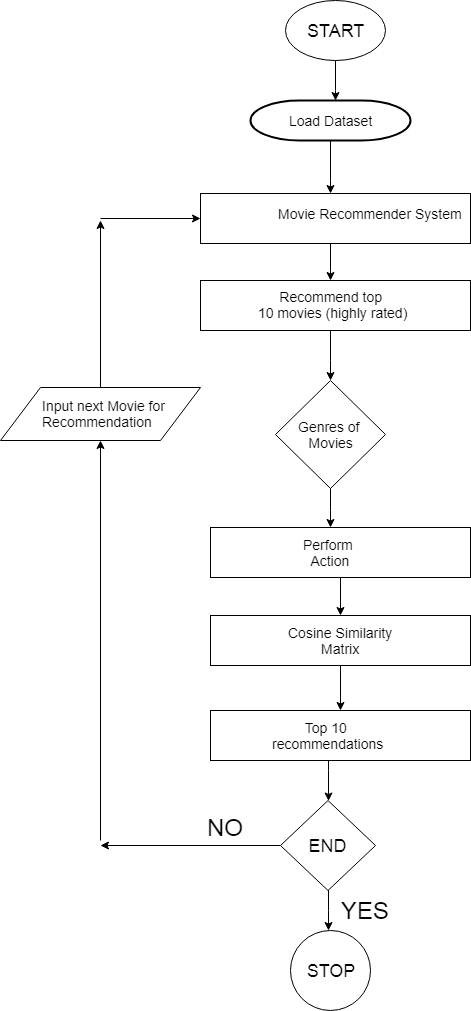
#### Purpose of Activity Diagrams

##### The basic purposes of activity diagrams is similar to other four diagrams. It captures the dynamic behavior of the system. Other four diagrams are used to show the message flow from one object to another but activity diagram is used to show message flow from one activity to another.

**Where to Use Activity Diagrams?**

The basic usage of activity diagram is similar to other four UML diagrams. The specific usage is to model the control flow from one activity to another. This control flow does not include messages.

Activity diagram is suitable for modeling the activity flow of the system. An application can have multiple systems. Activity diagram also captures these systems and describes the flow from one system to another. This specific usage is not available in other diagrams. These systems can be database, external queues, or any other system.



**Fig:Activity Diagram**

**9.5. CLASS DIAGRAM**

Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application.

Class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modeling of object oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages.

**Purpose of Class Diagrams**

The purpose of class diagram is to model the static view of an application. Class diagrams are the only diagrams which can be directly mapped with object-oriented languages and thus widely used at the time of construction.

UML diagrams like activity diagram, sequence diagram can only give the sequence flow of the application, however class diagram is a bit different. It is the most popular UML diagram in the coder community.

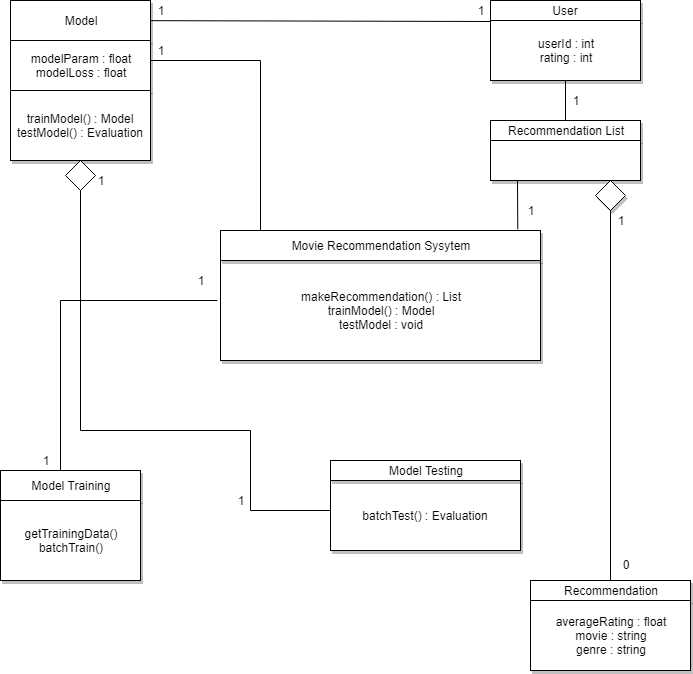
**Where to Use Class Diagrams?**

Class diagram is a static diagram and it is used to model the static view of a. The static view describes the vocabulary of the system.

Class diagram is also considered as the foundation for component and deployment diagrams. Class diagrams are not only used to visualize the system static view of the system but they are also used to construct the executable code for forward and reverse engineering of any system.

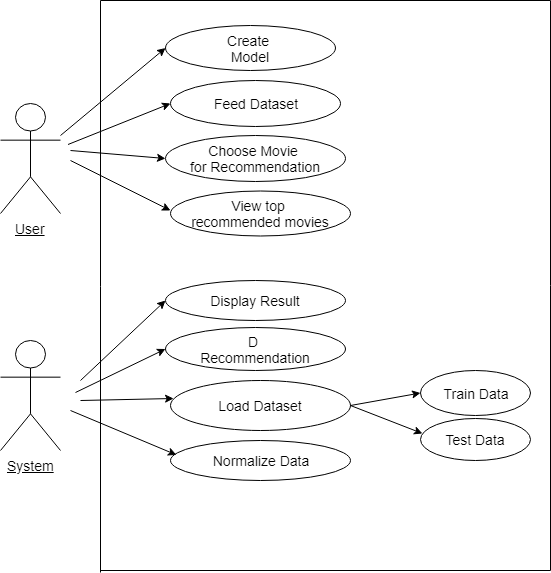
Generally, UML diagrams are not directly mapped with any object-oriented programming languages but the class diagram is an exception.

Class diagram clearly shows the mapping with object-oriented languages such as Java, C++, etc. From practical experience, class diagram is generally used for construction purpose.



## Fig: Class Diagram

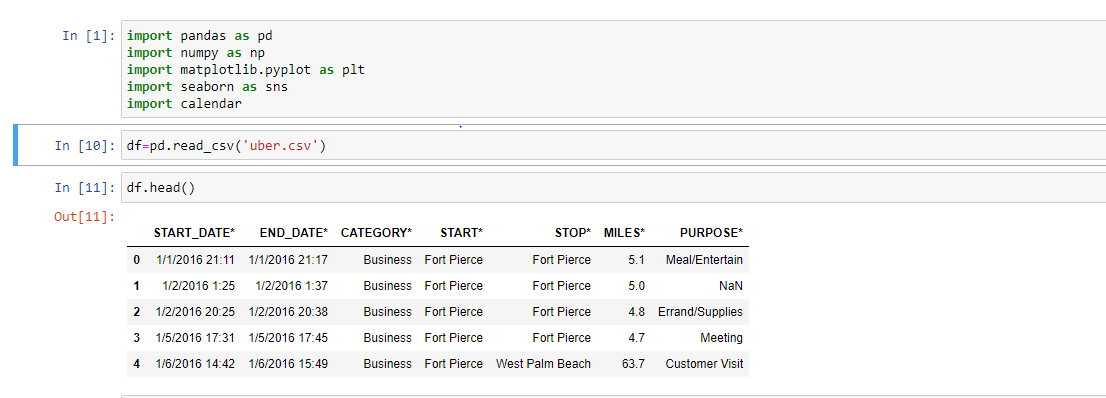
#### 9.6 USE CASE DIAGRAM

****

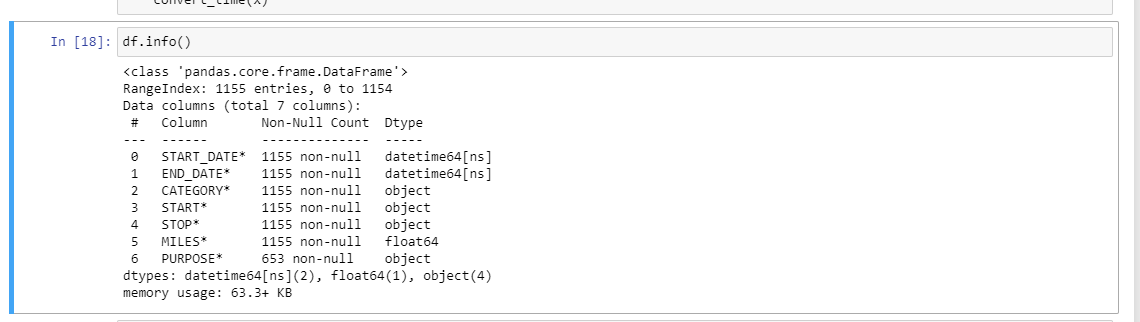
**Fig: Use Case Diagram**

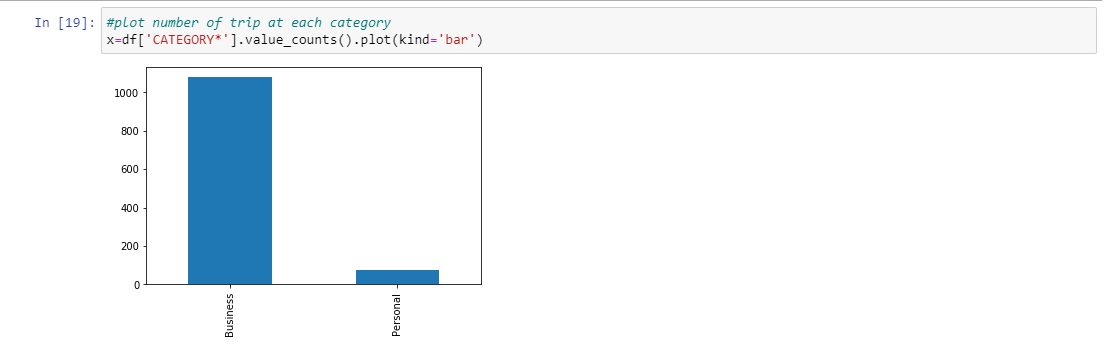
1. **USER INTERFACE DESIGN**
2. The [design](http://en.wikipedia.org/wiki/Design) of [user interfaces](http://en.wikipedia.org/wiki/User_interface) for [machines](http://en.wikipedia.org/wiki/Machine) and [software,](http://en.wikipedia.org/wiki/Software) such as computers, [home appliances,](http://en.wikipedia.org/wiki/Home_appliance) [mobile devices,](http://en.wikipedia.org/wiki/Mobile_device) and other [electronic devices,](http://en.wikipedia.org/wiki/Electronics) with the focus on maximizing the [user experience.](http://en.wikipedia.org/wiki/User_experience) The goal of user interface design is to make the user's interaction as simple and efficient as possible, in terms of accomplishing user goals ([user-centered design](http://en.wikipedia.org/wiki/User-centered_design)).
3. Good user interface design facilitates finishing the task at hand without drawing unnecessary attention to it. [Graphic design](http://en.wikipedia.org/wiki/Graphic_design) and typography are utilized to support its [usability,](http://en.wikipedia.org/wiki/Usability) influencing how the user performs certain interactions and improving the aesthetic appeal of the design; design aesthetics may enhance or detract from the ability of users to use the functions of the interface. The design process must balance technical functionality and visual elements (e.g., [mental model](http://en.wikipedia.org/wiki/Mental_model)) to create a system that is not only operational but also usable and adaptable to changing user needs.

**SNAPSHOTS**

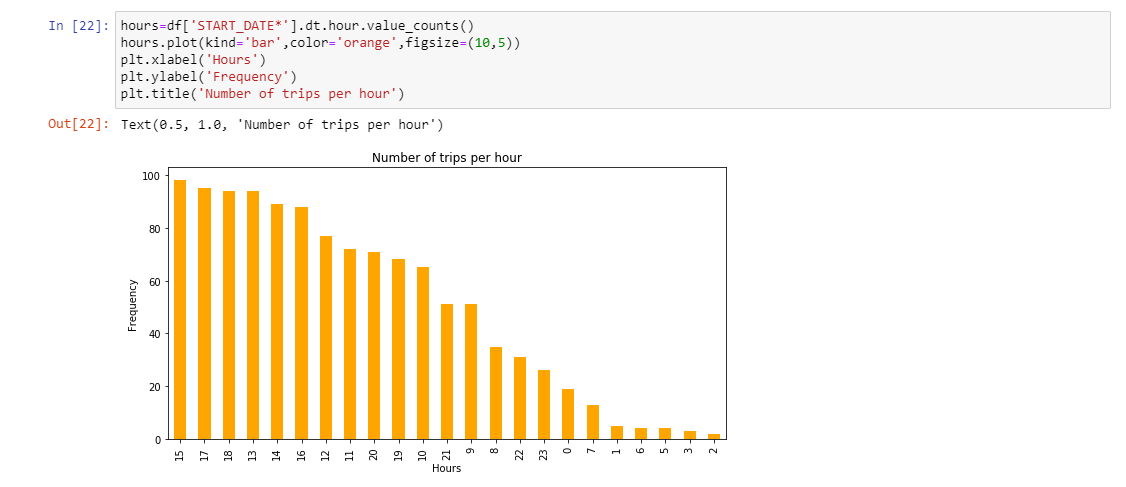
****

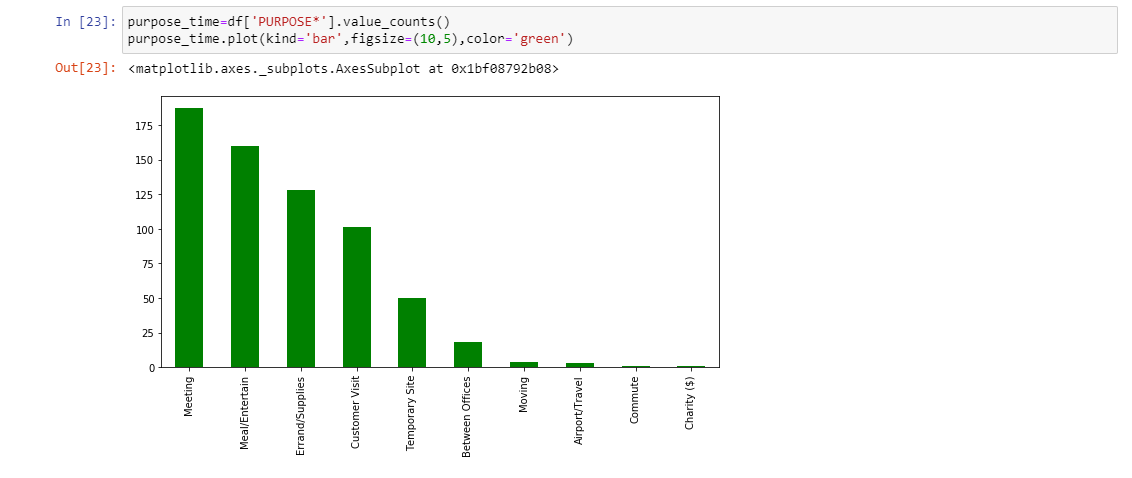
****

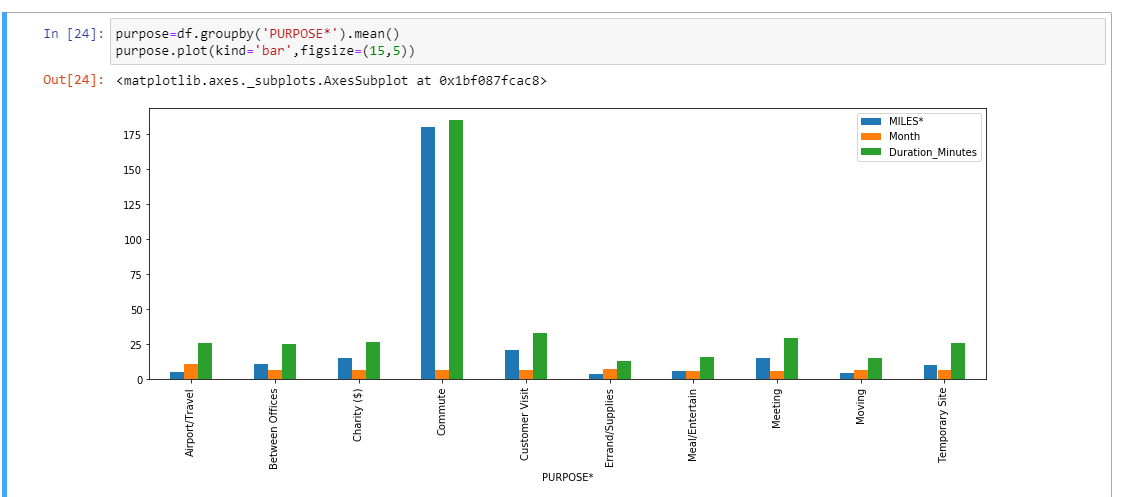


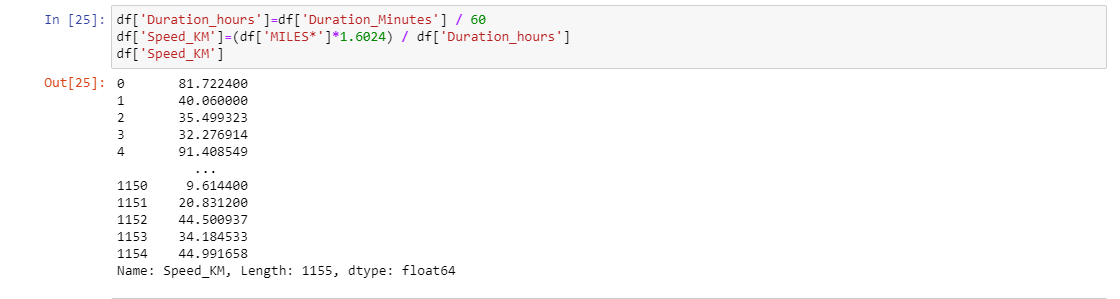
****

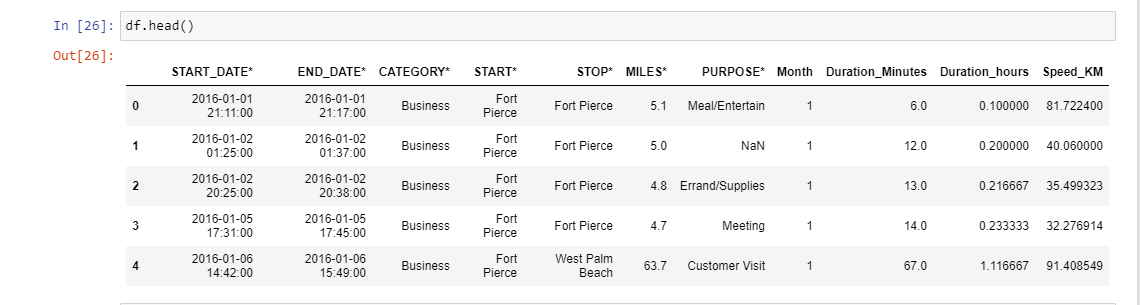
****

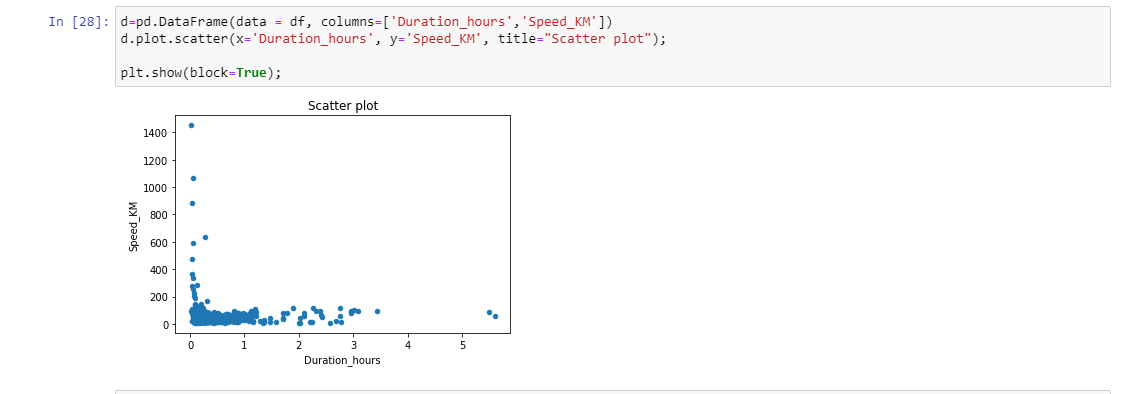
****

****

****

****

****

****

1. **IMPLEMENTATION AND TESTING**

A software system test plan is a document that describes the objectives, scope, approach and focus of software testing effort. The process of preparing a test plan is a usual way to think the effort

needed to validate the acceptability of a software product. The complete document will help people

outside the test group understand the "WHY" and "HOW" product validation. It should be through

enough to be useful but not so through that no one outside the test group will read it.

* 1. **Introduction**

Testing is the process of running a system with the intention of finding errors. Testing enhances the integrity of a system by detecting deviations in design and errors in the system. Testing aims at detecting error-prone areas. This helps in the prevention of errors in a system. Testing also adds value to the product by conforming to the userrequirements.

The main purpose of testing is to detect errors and error prone areas in a system.Testing must be through well planned.A partially tested system is to detect errors and error prone areas in a system.Testing must be through well planned.A partially tested system is as bad as an untested system.And the price of an untested and under tested system is high.

* 1. **Objectives Of Testing**

The objective our test plan is to find and report as many bugs as possible to improve the integrity of our program. Although exhaustive testing is not possible, we will exercise a broad range of tests to achieve our goal. Our user interface to utilize these functions is designed to be user-friendly and provide easy manipulation of the tree. The application will only be used as a demonstration tool, but we would like to ensure that it could be run from a variety of platforms with little impact on performance or usability.

* 1. **Process Overview**

The following represents the overall flow of the testing process:

* Identify the requirements to be tested. All test cases shall be derived using the current Program Specification.
* Identify which particular test(s) will be used to test each module.
* Review the test data and test cases to ensure that the unit has been thoroughly verified and that the test data and test cases are adequate to verify proper operation of the unit.

#### Test Cases

##### A test case is a document that describe an input, action, or event and expected response, to determine if a feature of an application is working correctly. A test case should contain particular such as test case identifier, test condition, input data.

Requirement expected results. The process of developing test cases can help find problems in the requirements or design of an application since it requires completely thinking through the operations of the application.

* 1. **Testing Steps**
* **Unit Testing**

Unit testing focuses efforts on the smallest unit of software design. This is known as module testing. The modules are tested separately. The test is carried out during programming stage itself. In this step, each module is found to be working satisfactory as regards to the expected output from the module.

* **Integration Testing**

Data can be lost across an interface. One module can have an adverse effect on another, sub functions, when combined, may not be linked in desired manner in major functions. Integration testing is a systematic approach for constructing the program structure, while at the same time conducting test to uncover errors associated within the interface.

* 1. **Validation**

At the culmination of the integration testing, Software is completely assembled as a package. Interfacing errors have been uncovered and corrected and a final series of software test begin in validation testing. Validation testing can be defined in many ways, but a simple definition is that the validation succeeds when the software functions in a manner that is expected by the customer. After validation test has been conducted, one of the three possible conditions exists.

a)The function or performance characteristics confirm to specification and are accepted. b)A deviation from specification is uncovered and a deficiency lists is created.

c)Proposed system under consideration has been tested by using validation test and found to be working satisfactory.

|  |  |
| --- | --- |
| **Tested By:** | Srilopa Ghosh |
| **Test Type** | White box Testing |
| **TestCaseName** | Uber data analysis |
| **Test Case Description** | Analysis the data of 2016 .of Uber   1. Number of trip in business and personal purpose 2. Frequency of trip throughout the day with respect to time 3. In which purpose user uses the car 4. Comparative analysis of amount of trip in miles , duration . |

* 1. **White Box Testing**

In white box testing, the UI is bypassed. Inputs and outputs are tested directly at the code level and the results are compared against specifications. This form of testing ignores the function of the program under test and will focus only on its code and the structure of that code. Test case designers shall generate cases that not only cause each condition to take on all possible values at least once, but that cause each such condition to be executed at least once. To ensure this happens, we will be applying Branch Testing. Because the functionality of the program is relatively simple, this method will be feasible to apply.

* 1. **Black box testing**

Black box testing typically involves running through every possible input to verify that it results in the right outputs using the software as an end-user would. We have decided to perform Equivalence Partitioning and Boundary Value Analysis testing on our application.

* 1. **System Testing**

The goals of system testing are to detect faults that can only be exposed by testing the entire integrated system or some major part of it. Generally, system testing is mainly concerned with areas such as performance, security, validation, load/stress, and configuration sensitivity.

* 1. **Output Testing**

After performing the validation testing, the next step is output testing of the proposed system, since no system could be useful if it does not produce the required output in a specific format. The output format on the screen is found to be correct. The format was designed in the system design time according to the user needs. For the hard copy also; the output comes as per the specified requirements by the user. Hence output testing did not result in any correction for the system.

* 1. **User Acceptance Testing**

User acceptance of a system is the key factor for the success of any system. The system under consideration is tested for the user acceptance by constantly keeping in touch with the prospective system users at the time of developing and making changes whenever required.

This is done in regard to the following point:

1. Input Screen Design.
2. Output Screen Design.
3. Format of reports and other outputs

#### Integration Testing

##### Software testing is always used in association with verification and validation. In the testing phase of this project our aim is to find the answer to following two questions.

* Whether the software matches with the specification (i.e.process base) to verify the product.
* Whether this software in one client what wants (i.e. product base) to validate the product.
* Unit testing and integration testing has been carried out to find the answer to above questions. In unit testing each individual module was test to find any unexpected behaviour if exists. Later all the module was integrated and flat file was generated.

#### Functional Testing

##### These are the points concerned during the stress test:

* Nominal input: character is in putted in the place of digits and the system has to flash the message "Data error"
* Boundary value analysis: exhaustive test cases have designed to create an output report that produces the maximum (and minimum) allowable number of table entries.

## SYSTEM SECURITY MEASURES

#### Database Security

##### System security measure is meant to be provided to make your system reliable and secured from unauthorized user may create threats to the system. So you should follow some security measures. We have used security levels in database level at system level.

* 1. **System Security**

If we talk about the system security in our proposed system we have implemented with the help of maintain the session throughout the system’s use. Once a user has logged out than he/she will not be able to perform any task before signing back again.

A high level of authentic login is given to the system so this is a very tedious task to enter without authorization and authentication.

1. **COST ESTIMATION**

The **Constructive Cost Model** (COCOMO) is a procedural **software cost estimation model** developed by **Barry W. Boehm**. Intermediate COCOMO takes these Cost Drivers into account and Detailed COCOMO additionally accounts for the influence of individual project phases.

**Types of COCOMO**

1. **Basic COCOMO**

Basic COCOMO computes software development effort (and cost) as a function of program size. Program size is expressed in estimated thousands of source lines of code ([SLOC,](https://en.wikipedia.org/wiki/Source_lines_of_code) [KLOC](https://en.wikipedia.org/wiki/Source_lines_of_code#Related_terms)).

COCOMO applies to three classes of software projects:

* + Organic projects - small teams with good experience working with less than rigid requirements.
  + Semi-detached projects - medium teams with mixed experience working with a mix of rigid and less than rigid requirements.
  + Embedded projects - developed within a set of tight constraints. It is also combination of organic and semi-detached projects. (Hardware, software, operational, ...)

The basic COCOMO equations take the form: Effort Applied (E) = a\*(KLOC)b[man-months ] Development Time (D) = c\*(Effort Applied)d [months]

People required (P) = Effort Applied / Development Time [count] Where,

KLOC is the estimated number of delivered lines (expressed in thousands) of code for project. The coefficients a, b, c and d are given in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Software project** | **a** | **b** | **c** | **d** |
| Organic | 2.4 | 1.05 | 2.5 | 0.38 |
| Semi-detached | 3.0 | 1.12 | 2.5 | 0.35 |
| Embedded | 3.6 | 1.20 | 2.5 | 0.32 |

Basic COCOMO is good for quick estimate of software costs. However, it does not account for differences in hardware constraints, personnel quality and experience, use of modern tools and techniques, and so on.

1. **Intermediate COCOMO**

Intermediate COCOMO computes software development effort as function of program size and a set of cost drivers that include subjective assessment of product, hardware, personnel and project attributes. This extension considers a set of four cost drivers, each with a number of subsidiary attributes: -

Product attributes:

1. Required software reliability
2. Size of application database
3. Complexity of the product Hardware attributes:
4. Run-time performance constraints
5. Memory constraints
6. Volatility of the virtual machine environment
7. Required turnabout time Personnel attributes
8. Analyst capability
9. Software engineering capability
10. Applications experience
11. Virtual machine experience
12. Programming language experience Project attributes
13. Use of software tools
14. Application of software engineering methods
15. Required development schedule

Each of the 15 attributes receives a rating on a six-point scale that ranges from very low to extra high (in importance or value). An effort multiplier from the table below applies to the rating. The product of all effort multipliers results in an effort adjustment factor (EAF). Typical values for EAF range from 0.9 to 1.4. The Intermediate COCOMO formula now takes the form:

E = a\*(KLOC)b(EAF)

Where,

E is the effort applied in person-months, KLOC is the estimated number of thousands of delivered lines of code for the project, and EAF is the factor calculated above. The coefficient a and the exponent b is given in the table:

|  |  |  |
| --- | --- | --- |
| **Software project** | **a** | **B** |
| Organic | 3.2 | 1.05 |
| Semi-detached | 3.0 | 1.12 |
| Embedded | 2.8 | 1.20 |

The Development time D calculation uses P in the same way as in the Basic COCOMO.

1. **Detailed COCOMO**

Detailed COCOMO incorporates all characteristics of the intermediate version with an assessment of the cost driver's impact on each step (analysis, design, etc.) of the software engineering process. The detailed model uses different effort multipliers for each cost driver attribute. These Phase Sensitive effort multipliers are each to determine the amount of effort required to complete each phase. In detailed COCOMO, the whole software is divided in different modules and then we apply COCOMO in different modules to estimate effort and then sum the effort.

In detailed COCOMO, the effort is calculated as function of program size and a set of cost drivers given according to each phase of software life cycle. A Detailed project schedule is never static. The Six phases of detailed COCOMO are: -

* + Plan and requirement.
  + System design.
  + Detailed design.
  + Module code and test.
  + Integration and test.
  + Cost Constructive model

**Detailed cost estimation of Movie Recommender System**

So by considering all the facts, if we calculate the cost, it will be like;

**The OES is having 2.1 Kilo Lines of Code**.According to the COCOMO model the comparison is:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mode** | **Project Size** | **Innovation** | **Deadline** | **Development Environment** |
| **Organic** | Typically 2-50 KLOC | Little | Not tight | Familiar and in house |
| **Semi- detached** | Typically 50-300 KLOC | Medium | Medium | Medium |
| **Embedded** | Typically Over 300 KLOC | Significant | Tight | Complex hardware/Customer interface is required |

The project has 2.1 KLOC, so it’s under **organic** category. Putting the facts in the formulas,

**Effort Applied (E)** = a\*(KLOC)b [ [person-months](https://en.wikipedia.org/wiki/Man-month) ]

= 2.4\*(2.1)1.05 PM

= **5 PM**

**Development Time (D)** = c\*(Effort Applied)d [months]

= 2.5\*(5)0.38 M

= **4.6 M**

## FUTURE SCOPE AND FURTHER ENHANCEMENTS

The scope of Machine Learning is expanding across all fields such as banking and finance, information technology, media & entertainment, gaming, and the automotive industry. As the Machine Learning scope is very high, there are some of the areas where researchers are working toward revolutionizing the world for the future. Let us discuss them in detail.

## CONCLUSION

Thus the project teaches Data visualization makes it easier to understand the core values of the databases.Data science is very interesting and this is one of the projects which prove it. This project is easily implemented and very useful for a number of apps. Not only Uber but there is a lot more application which will need to extract information from their huge databases. This project can help in that situation..

## BIBLIOGRAPHY

1. Takeo Kanade. volume 47. Birkh¨auser Basel, 1977.
2. Lawrence Sirovich and Michael Kirby. Low-dimensional procedure for the characterization of human faces. Josa a, 4(3):519–524, 1987.
3. M. Turk and A. Pentland. Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1):71– 86, Jan 1991.
4. Dong chen He and Li Wang. Texture unit, texture spectrum, and texture analysis. IEEE Transactions on Geoscience and Remote Sensing, 28(4):509–512, Jul 1990.

5] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In 2009 IEEE 12th International Conference on Computer Vision, pages 32– 39, Sept 2009.

1. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. ﬁsherfaces: recognition using class speciﬁc linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):711–720, Jul 1997.
2. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features.