

PROFESSIONAL TRAINING REPORT

at

**Sathyabama Institute of Science and Technology
(Deemed to be University)**

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Technology Degree in

Information Technology

By

**VIJAY VINAYAK R
(Reg.No.37120084)**



**DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING**

**SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119**

AUGUST 2020



SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

(Established under Section 3 of UGC Act, 1956)
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai -600119
www.sathyabama.ac.in



DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **R.VIJAY VINAYAK** (Reg.no.37120084) who carried out the project entitled “**PUBG DATA SCIENCE**” under our supervision from DEC 2020 to JAN 2021.

Internal Guide

E.BRUMANCIA M.E,(Ph.D.,)

Head of the Department

Dr.R SUBHASHINI M.E,Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I ,R.VIJAY VINAYAK (Reg.No.37120084) hereby declare that the Project Report entitled “**PUBG DATA SCIENCE**” done by me under the guidance of **E.BRUMANCIA M.E(Ph.D).**, is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology Degree in Information Technology .

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D, Dean**, School of Computing , **Dr.SUBHASHINI M.E., Ph.D.** , Head of the Department of Information and Technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **E.Brumania M.E,(Ph.D.,)** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.



TO WHOMSOEVER IT MAY CONCERN

This is to certify that Vijay Vinayakhas successfully completed his Internship with Perficient India Pvt, Ltd, in Web application using MongoDB, Express, Angular, Node JS and Agile SCRUM Principles from December 11,2019 to January 06, 2020.

During the training his conduct was GOOD.

We wish him good luck in his future endeavors.

For Perficient India Pvt Ltd

**Parvez Fakir Maideen
Director**

PERFICIENT INDIA PRIVATE LIMITED

Perficient Tower, Plot No. A17 & A18, Thiru Vi Ka Industrial Estate, Guindy, Chennai - 600 032.
Phone : 95516 50494 / 95516 50496 URL : www.perficient.com CIN : U72200TN2002PTC049576 GST : 33AABCE3223F1ZB



07/12/2020

Vijay Vinayak

has successfully completed

Blockchain Basics

an online non-credit course authorized by University at Buffalo and The State University of New York and offered through Coursera

Bina Ramamurthy
Teaching Professor
Computer Science and Engineering

**COURSE
CERTIFICATE**



Verify at coursera.org/verify/QCB9SYXVTXD7
Coursera has confirmed the identity of this individual and
their participation in the course.

ABSTRACT

The project titled PlayerUnknown's Battlegrounds (PUBG) Data Science. It takes PUBG Statistics of players to build a Regression model using the Random Forest Algorithm to determine the Placement Percentile more than 4 Million players. This project will use the Exploratory Data Analysis, with many important libraries such as TensorFlow and Scikit Learn. The method to follow Exploratory Data Analysis on PUBG Data Feature Engineering of new attributes Building a Regression Model using Random Forest. The data provided with a large number of anonymous PUBG game stats, formatted so that each row contains one player's post-game stats. The data comes from matches of all types: solos, duos, squads & custom; there is no guarantee of there being 100 players per match, nor at most four players per group. It create a model which predicts players' finishing placement based on their final stats, on a scale from 1 (1st place) to 0 (last place). I've created a data dictionary for more information regarding the data & to make our lives simpler.

TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE.NO
1	ABSTRACT LIST OF ABBREVIATIONS LIST OF FIGURES INTRODUCTION	
	1.1 INTRODUCTION TO THE PROJECT	1
	1.2 AIM AND SCOPE	1
	1.3 FEASABILITY STUDY	1
	1.4 SYSTEMS	3
2	ALGORITHMS AND METHODS	
	2.1 SDLC METHODOLOGY	4
	2.2 ACTIVITY DIAGRAM	6
3	SYSTEM DESIGN AND TESTING PLAN	
	3.1 MODULES	7
	3.2 INPUT DESIGN	8
	3.3 DATA FLOW DIAGRAM	10
	3.4 UML DIAGRAM	11
	3.5 TEST PLAN	15
	3.6 SYSTEM REQUIREMENTS	19
	3.7 NON-FUNCTIONAL REQUIREMENTS	19
4.	RESULTS AND DISCUSSIONS	
	4.1 SUMMARY	22
	4.2 SCREEN SHOTS	23
5	CONCLUSION AND FUTURE ENHANCEMENTS	
	5.1 CONCLUSION	36
	5.2 FUTURE ENHANCEMENT	37
	5.3 REFERENCES	37
6	APPENDIX	
	6.1 SOURCE CODE	38

LIST OF FIGURES

FIGURE NO.	TITLE OF THE FIGURE	PAGE NO.
2.1	Activity Diagram	6
3.1.2	System Architecture	7
3.3	Data Flow Diagram	9
3.4.1	Use Case Diagram	10
3.4.1	Class Diagram	11
3.4.1	Sequence Diagram	12
3.4.1	Component Diagram	13

LIST OF ABBREVIATIONS

ACRONYM

DFD

UML

EPL

JDT

IDE

GPS

EXPANSION

Data Flow Diagram

Unified Modeling Language

Eclipse Public License

Java Development Tools

Integrated Development
Environment

Global Positioning System

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO THE PROJECT

PUBG needs no introduction. It is one of the popular and the most played games right now. Players fight to death until one remains, so it is a survival game. There are pure statistical models to analyze the survival times. Using PUBG data, we will try to use one such survival models to understand how different strategies can improve the player's survival rates.

1.2. AIM AND SCOPE:

1.2.1. AIM:

The main aim of the project is "To Analyze the PUBG with help of data science".

1.2.2. SCOPE:

PUBG analysis using data science' is an end-to-end case study to understand the different stages in the Data Science life cycle. This will deal with 'data manipulation' with pandas, Numpy, and 'data visualization' with Matplotlib with the PUBG dataset.

1.3. Feasibility Study

1.3.1. Technical Feasibility

Evaluating the technical feasibility is the trickiest part of a feasibility study. This is because, .at this point in time, not too many detailed design of the system, making it difficult to access issues like performance, costs on (on account of the kind of technology to be deployed) etc. A number of issues have to be considered while doing a technical analysis.

Understand the different technologies involved in the proposed system before commencing the project we have to be very clear about what are the technologies that are to be required for the development of the new system. Find out whether the

organization currently possesses the required technologies. Is the required technology available with the organization?

1.3.2. Operational Feasibility

Proposed project is beneficial only if it can be turned into information systems that will meet the organizations operating requirements. Simply stated, this test of feasibility asks if the system will work when it is developed and installed. Are there major barriers to Implementation? Here are questions that will help test the operational feasibility of a project:

Does killing more people increases the chance of winning the game?

Hint: Use the correlation between the match winning percentage and number of kills to determine the relationship

How do we catch the fraudsters in the game?

Hint: Use various logical conditions based on game knowledge to determine fraudsters in the game

Can we predict the finishing position of a player in the game?

Hint: Regression Problem: Train and test a model using regression algorithm to predict the final position of the player at the end of the game. Create a model which predicts players' finishing placement based on their final stats, on a scale from 1 (first place) to 0 (last place).

1.3.3.Economic Feasibility

Economic feasibility attempts to weigh the costs of developing and implementing a new system, against the benefits that would accrue from having the new system in place. This feasibility study gives the top management the economic justification for the new system.

A simple economic analysis which gives the actual comparison of costs and benefits are much more meaningful in this case. In addition, this proves to be a useful point of reference to compare actual costs as the project progresses. There could be various

types of intangible benefits on account of automation. These could include increased customer satisfaction, improvement in product quality better decision making timeliness of information, expediting activities, improved accuracy of operations, better documentation and record keeping, faster retrieval of information, better employee morale.

1.4 SYSTEMS:

1.4.1 EXISTING SYSTEM

The existing system consists of entering the details in the Microsoft Excel Sheets for the storing of the data. When a manager needs information of the employee he searches for the specified file in the file system. He opens the file and takes the information. Report Generation done manually by copying the content of the different files into another file. The Manually generated report was then printed.

DISADVANTAGES OF EXISTING SYSTEM :

- Information retrieval is a very big process.
- Lack of organization of the files may lead to information loss due to accidental deletion of files.
- No security because the files are visible to the users.
- Report generation will be a big task.

1.4.2 PROPOSED SYSTEM

The Proposed system is a browser which is completely related to online system, which provides the centralized database. It stores Defects data and description of the particular Defect data. It can also create reports and documents based on the information in its database.

ADVANTAGES OF PROPOSED SYSTEM :

- The performance is increased due to well designed database.
- Security is increased
- Time saving in report generation
- Easy to update the details

CHAPTER 2

ALGORITHM AND METHODS

2.1. SDLC METHODOLOGY:

This document play a vital role in the development of life cycle (SDLC) as it describes the complete requirement of the system. It means for use by developers and will be the basic during testing phase. Any changes made to the requirements in the future will have to go through formal change approval process.

SPIRAL MODEL was defined by Barry Boehm in his 1988 article, "A spiral Model of Software Development and Enhancement. This model was not the first model to discuss iterative development, but it was the first model to explain why the iteration models.

As originally envisioned, the iterations were typically 6 months to 2 years long. Each phase starts with a design goal and ends with a client reviewing the progress thus far. Analysis and engineering efforts are applied at each phase of the project, with an eye toward the end goal of the project.

The steps for Spiral Model can be generalized as follows:

- The new system requirements are defined in as much details as possible. This usually involves interviewing a number of users representing all the external or internal users and other aspects of the existing system.
- A preliminary design is created for the new system.
- A first prototype of the new system is constructed from the preliminary design. This is usually a scaled-down system, and represents an approximation of the characteristics of the final product.
- A second prototype is evolved by a fourfold procedure:
 1. Evaluating the first prototype in terms of its strengths, weakness, and risks.
 2. Defining the requirements of the second prototype.
 3. Planning an designing the second prototype.
 4. Constructing and testing the second prototype.

- At the customer option, the entire project can be aborted if the risk is deemed too great. Risk factors might involved development cost overruns, operating-cost miscalculation, or any other factor that could, in the customer's judgment, result in a less-than-satisfactory final product.
- The existing prototype is evaluated in the same manner as was the previous prototype, and if necessary, another prototype is developed from it according to the fourfold procedure outlined above.
- The preceding steps are iterated until the customer is satisfied that the refined prototype represents the final product desired.
- The final system is constructed, based on the refined prototype.
- The final system is thoroughly evaluated and tested. Routine maintenance is carried on a continuing basis to prevent large scale failures and to minimize down time.

The following diagram shows how a spiral model acts like:

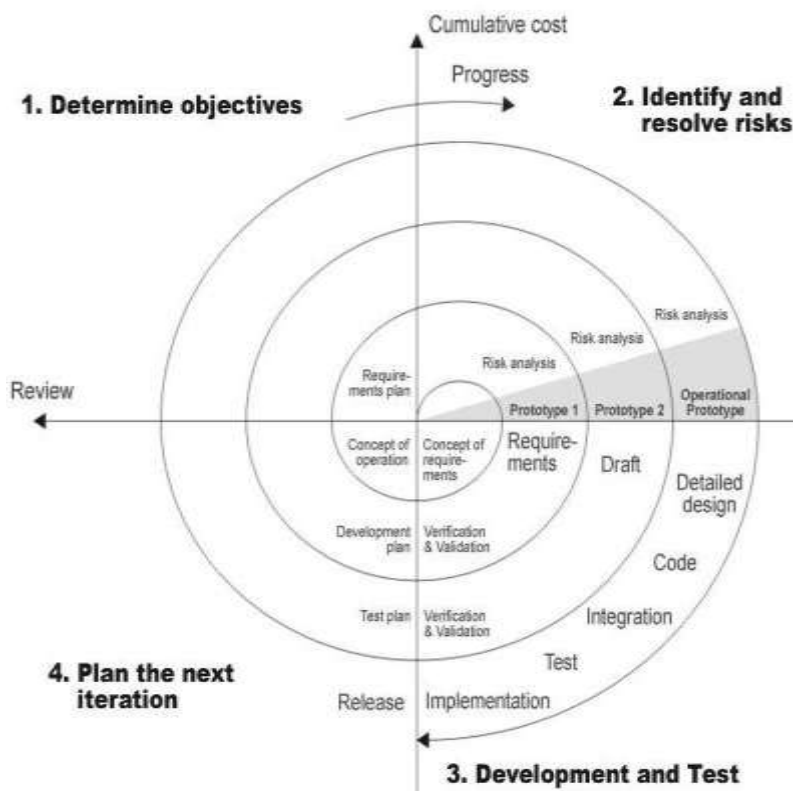


Fig 1.0-Spiral Model

ADVANTAGES

- Estimates(i.e. budget, schedule etc .) become more realistic as work progresses, because important issues discovered earlier.
- It is more able to cope with the changes that are software development generally entails.
- Software engineers can get their hands in and start working on the core of a project earlier.

2.2 ACTIVITY DIAGRAM:

ACTIVITY DIAGRAM

Admin:

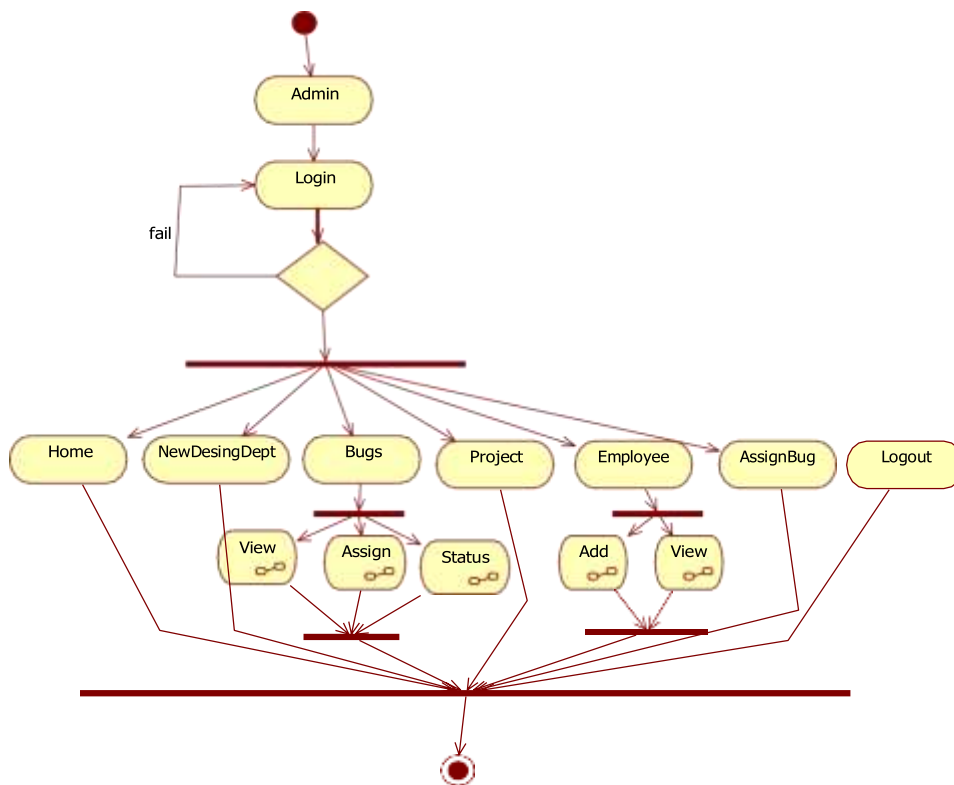


Fig 2.1. Activity diagram

CHAPTER 3

SYSTEM DESIGN AND TESTING PLAN

3.1.1 MODULES:

ADMIN: This module has the entire access to all other modules, admin creates the project and assigning the projects to the created manager, adding members to the managers, assigning Defects based on the priority. Can update the manager, members and access to the particular project data. Generating reports based on the managers report submission.

Manager: Manager has the full access to the particular project assigned by the admin and controls the team members access to the Defects assigned. Has the permission to generate the reports and update the information of team members and adding members to the project.

Developer: Can access the task or Defect assigned by the manager, view assigned projects and resolving the assigned Defect. Developer can view the Defects list assigned by the manager.

Tester: Tester can access to the projects , can view the assigned projects and can add a new Defect to the list and send the bug back to the manager. Tester can login to the system and access the assigned projects list.

Reports: Both Admin and Manager can access this module and generate the reports based on the requirements.

3.1.2 SYSTEM ARCHITECTURE

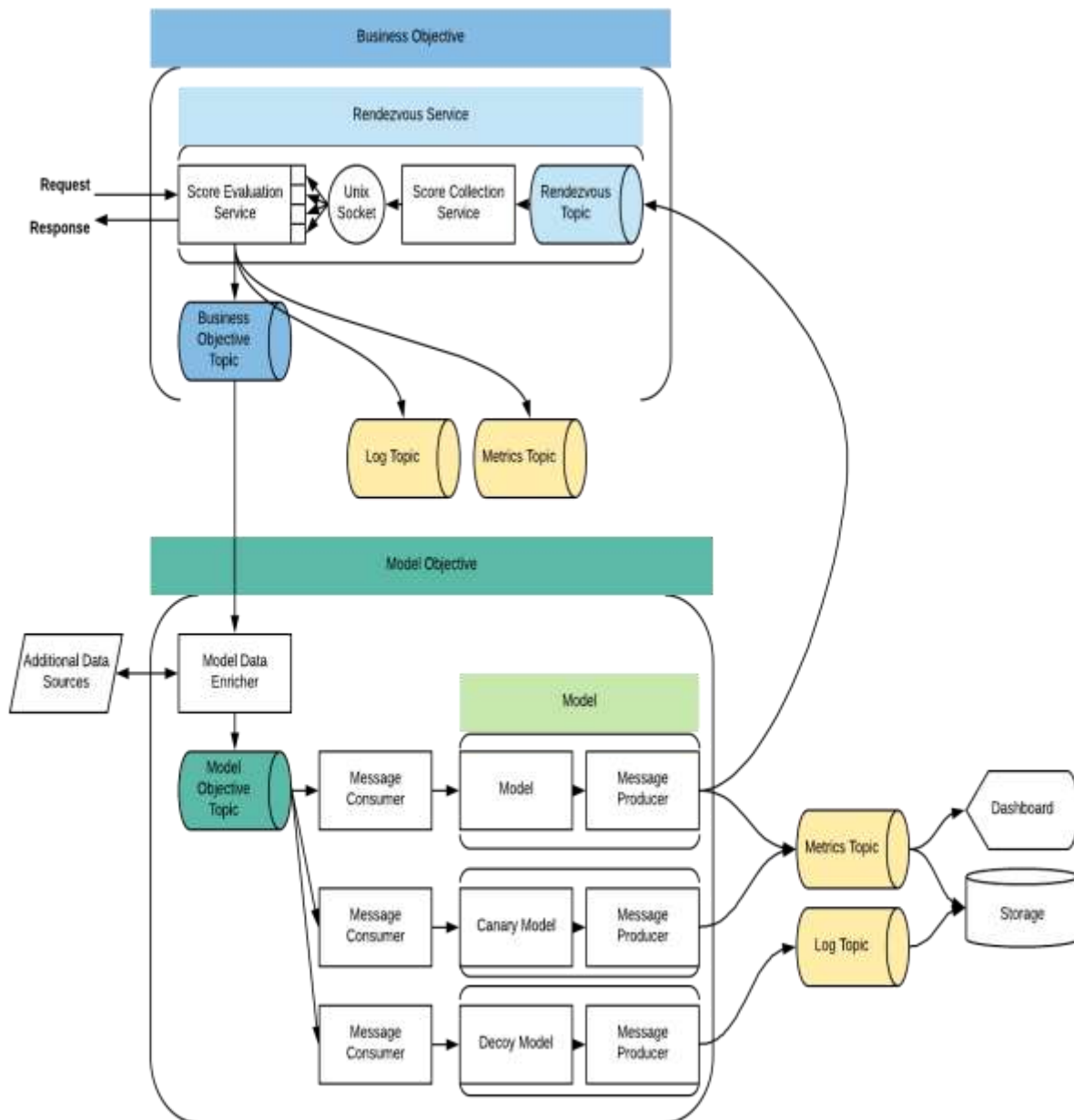


Fig .3.1.2 shows system architecture

3.2 INPUT DESIGN:

3.2.1. THE PRESENTATION LAYER

It Also called as the client layer comprises of components that are dedicated to presenting the data to the user. For example: Windows/Web Forms and buttons, edit boxes, Text boxes, labels, grids, etc.

3.2.2. THE BUSINESS RULES LAYER

This layer encapsulates the Business rules or the business logic of the encapsulations. To have a separate layer for business logic is of a great advantage. This is because any changes in Business Rules can be easily handled in this layer. As long as the interface between the layers remains the same, any changes to the functionality/processing logic in this layer can be made without impacting the others. A lot of client-server apps failed to implement successfully as changing the business logic was a painful process

3.2.3. THE DATA ACCESS LAYER

This layer comprises of components that help in accessing the Database. If used in the right way, this layer provides a level of abstraction for the database structures. Simply put changes made to the database, tables, etc do not affect the rest of the application because of the Data Access layer. The different application layers send the data requests to this layer and receive the response from this layer.

3.2.4. THE DATABASE LAYER

This layer comprises of the Database Components such as DB Files, Tables, Views, etc. The Actual database could be created using SQL Server, Oracle, Flatfiles,etc.

In an n-tier application, the entire application can be implemented in such a way that it is independent of the actual Database. For instance, you could change the Database Location with minimal changes to Data Access Layer. The rest of the Application should remain unaffected.

3.3. DATA FLOW DIAGRAM (DFD)

A data flow diagram is a way of representing a flow of data of a process or a system. The DFD also provides the information about the input and output of each entity and the process itself.

DFD- LEVEL

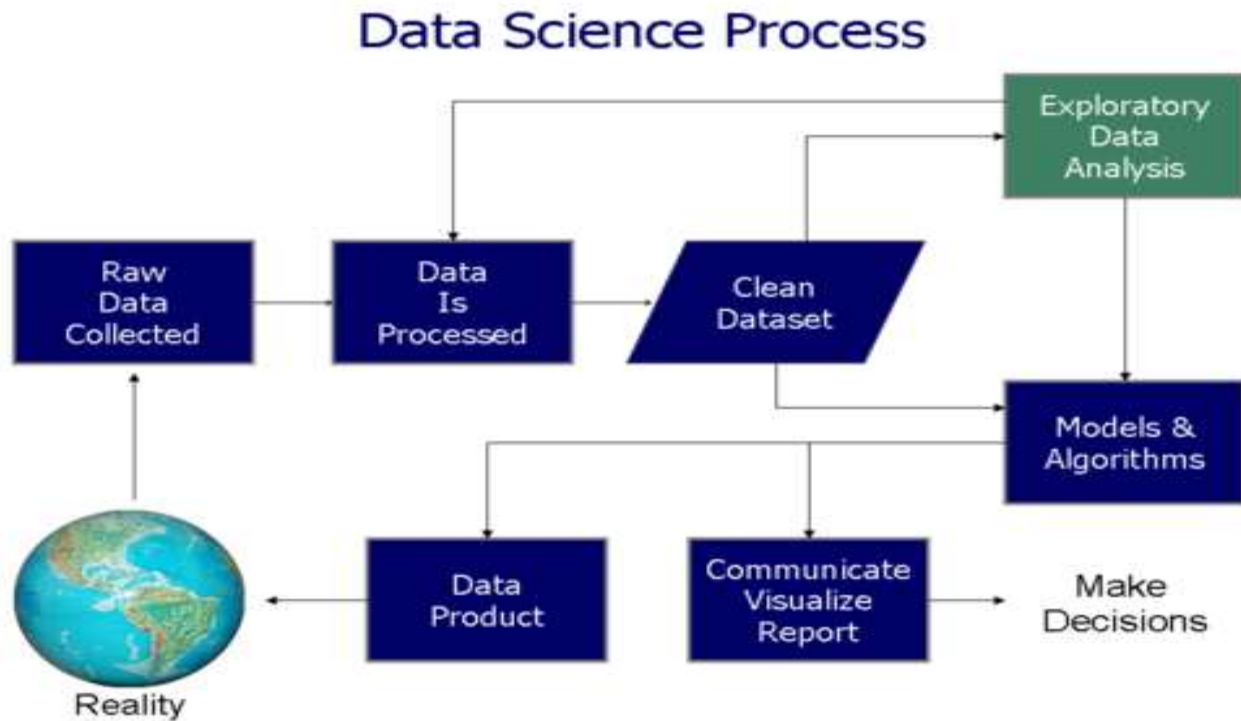


Fig 3.3. shows data flow diagram

3.4. UML DIAGRAMS

Unified Modelling language includes a set of graphic notations techniques to create a visual models of object oriented software intensive systems

3.4.1 TYPES OF UML DIAGRAMS

- Use Case Diagram
- Class diagram
- Sequential diagram
- Component diagram

Use case diagram :

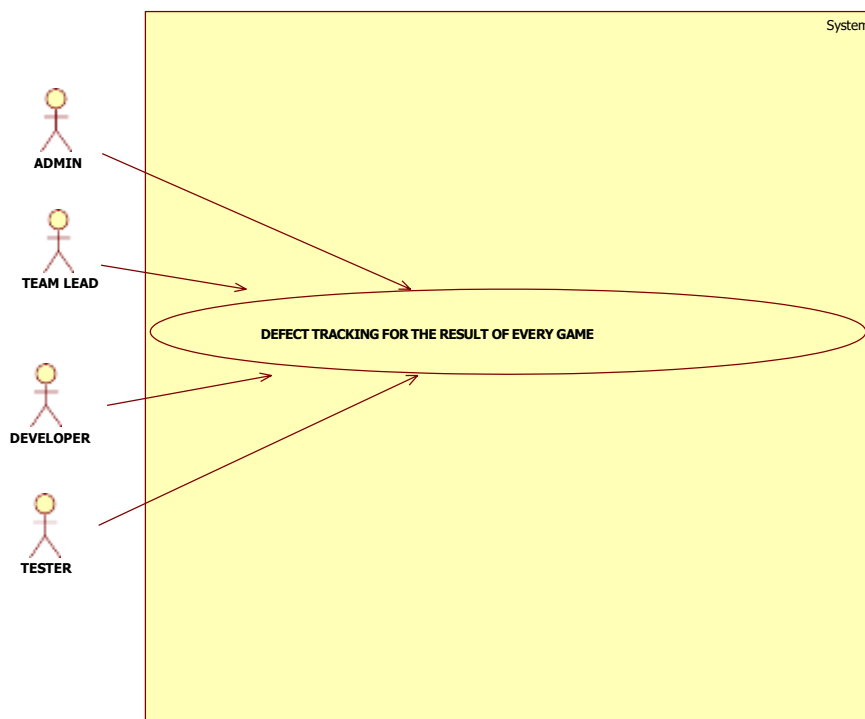


Fig. 3.4.1 Shows use case diagram

CLASS DIAGRAM :

In software Engineering , a class diagram in the unified modeling language is a type of structure diagram that describes the structure of a system by showing the systems ,classes ,their structures,operations and their relationships among objects.

In the diagram , classes are represented with bases which contain three parts.

- The top part contains the name of the class.
- The middle point contains the attributes of the class.
- The bottom part gives the methods or operations in the class can take or undertake .

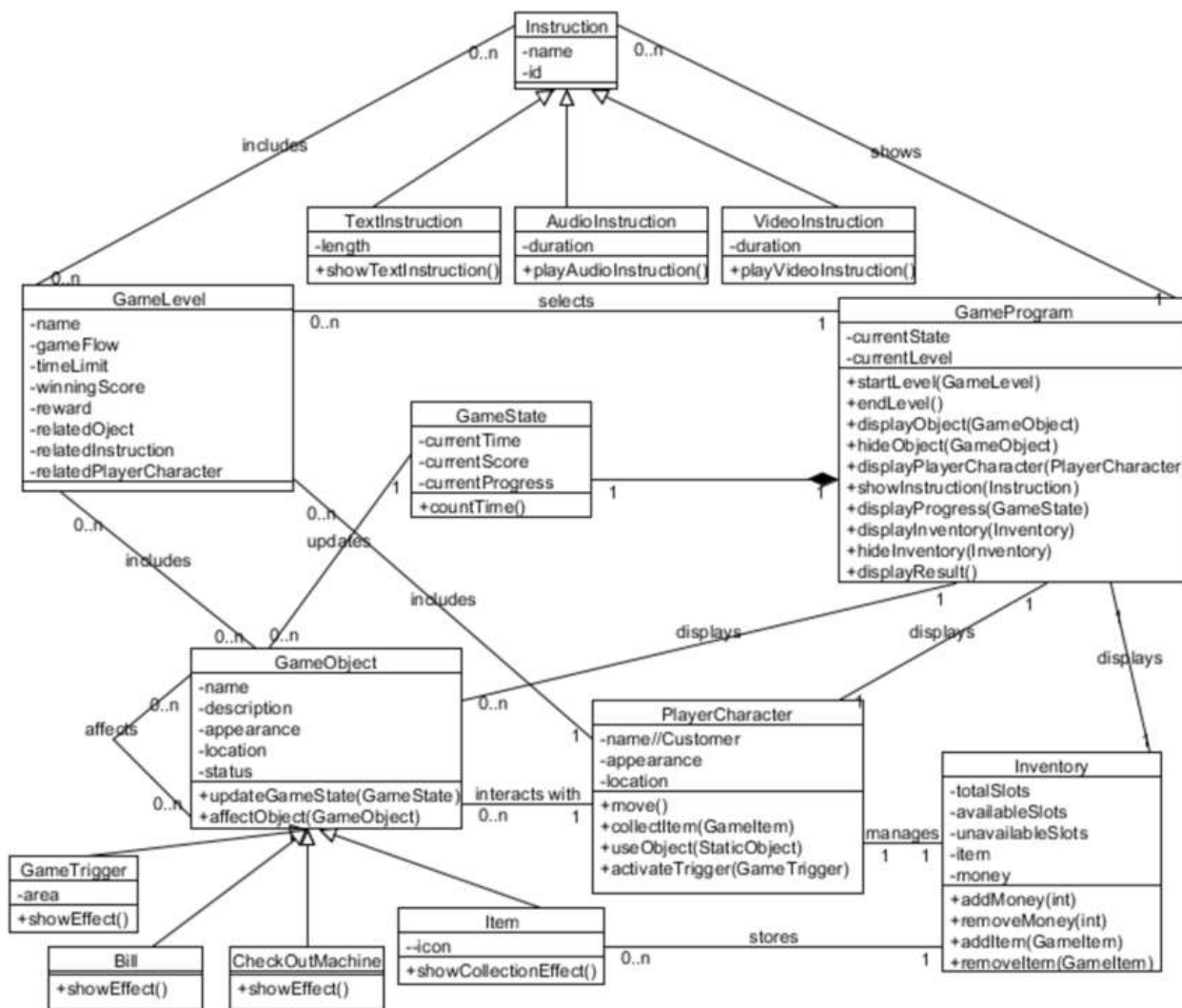


Fig.3.4.1 Shows a class diagram

SEQUENCE DIAGRAM :

- The Sequence diagram shows the ordering of the messages .
- It shows the order of the messages that the admin should perform.
- The interaction that takes place in a collaboration that either realizes a use case or an operation.
- High-level interactions between user of the system and the system,between the system and other systems,or between subsystems.

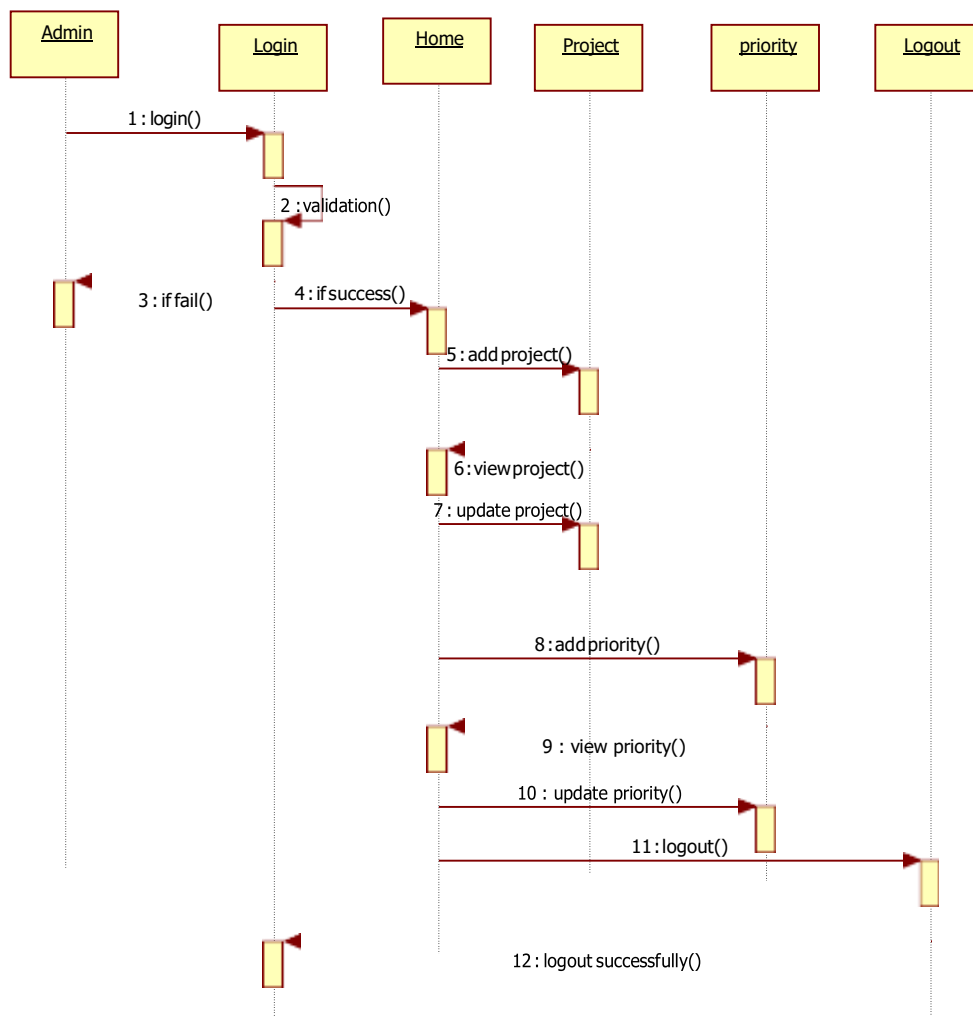


Fig. 3.4.1 Shows sequential diagram

COMPONENT DIAGRAM :

- A component diagram also known as a UML component diagram .
- It describes the organization and wiring of the physical components in a system.
- Component diagrams are often drawn to help model implementation details and double check that every aspect of the systems required.

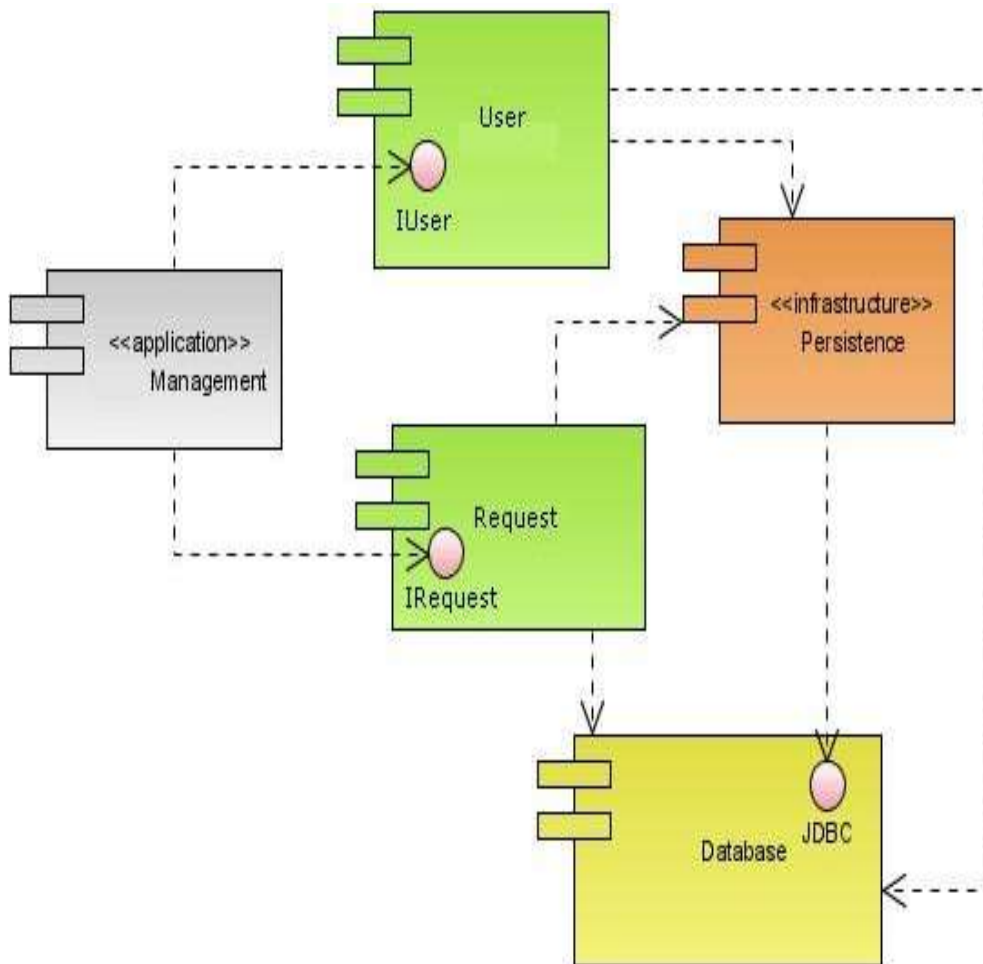


Fig .3.4.1 Shows a component diagram

3.5. TEST PLAN:

- The process of executing a system with the intent of finding an error.
- Testing is defined as the process in which defects are identified, isolated, subjected for rectification and ensured that product is defect free in order to produce the quality product and hence customer satisfaction.
- Quality is defined as justification of the requirements
- Defect is nothing but deviation from the requirements
- Defect is nothing but bug.
- Testing --- The presence of bugs
- Testing can demonstrate the presence of bugs, but not their absence
- Debugging and Testing are not the same thing!
- Testing is a systematic attempt to break a program or the AUT
- Debugging is the art or method of uncovering why the script /program did execute properly.

Testing Methodologies:

- **Black box Testing:** is the testing process in which tester can perform testing on an application without having any internal structural knowledge of application.

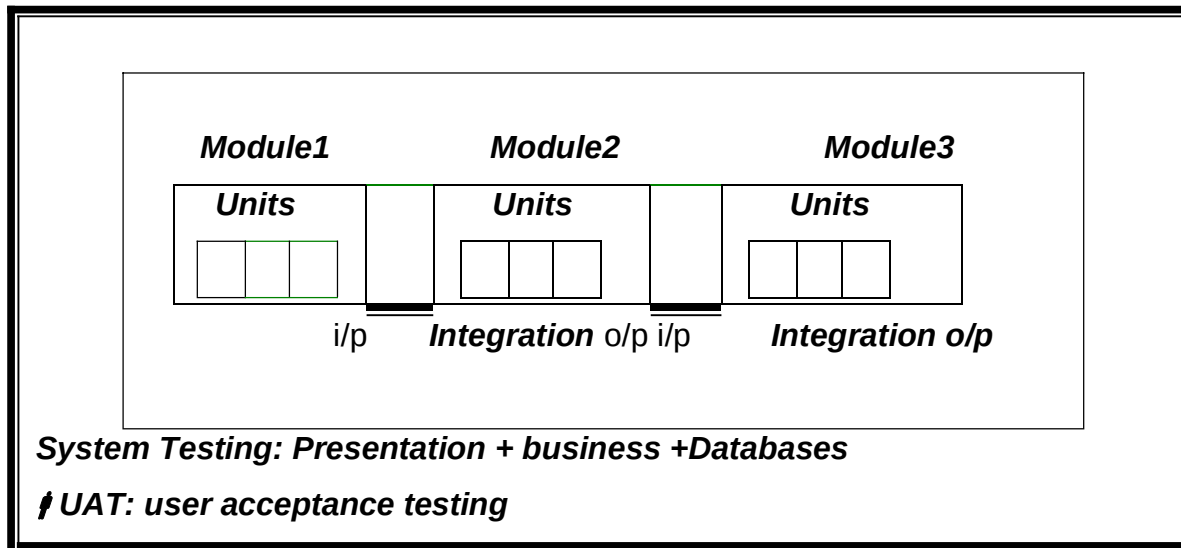
Usually Test Engineers are involved in the black box testing.

- **White box Testing:** is the testing process in which tester can perform testing on an application with having internal structural knowledge.

Usually The Developers are involved in white box testing.

- **Gray Box Testing:** is the process in which the combination of black box and white box tonics' are used.

Levels of Testing:



STLC (SOFTWARE TESTING LIFE CYCLE)

Test Planning:

1. Test Plan is defined as a strategic document which describes the procedure how to perform various testing on the total application in the most efficient way.
 2. This document involves the scope of testing,
 3. Objective of testing,
 4. Areas that need to be tested,
 5. Areas that should not be tested,
 6. Scheduling Resource Planning,
 7. Areas to be automated, various testing tools
- Used....

Test Development:

1. Test case Development (check list)
2. Test Procedure preparation. (Description of the Test cases).
3. Implementation of test cases. Observing the result.

Result Analysis:

1. Expected value: is nothing but expected behavior Of application.

2. Actual value: is nothing but actual behavior of application

Bug Tracing: Collect all the failed cases, prepare documents.

Reporting: Prepare document (status of the application)

Types Of Testing:

🌱 > **Smoke Testing:** is the process of initial testing in which tester looks for the availability of all the functionality of the application in order to perform detailed testing on them. (Main check is for available forms)

🌱 > **Sanity Testing:** is a type of testing that is conducted on an application initially to check for the proper behavior of an application that is to check all the functionality are available before the detailed testing is conducted by on them.

🌱 > **Regression Testing:** is one of the best and important testing. Regression testing is the process in which the functionality, which is already tested before, is once again tested whenever some new change is added in order to check whether the existing functionality remains same.

🌱 > **Re-Testing:** is the process in which testing is performed on some functionality which is already tested before to make sure that the defects are reproducible and to rule out the environments issues if at all any defects are there.

🌱 **Static Testing:** is the testing, which is performed on an application when it is not been executed.ex: GUI, Document Testing

🌱 **Dynamic Testing:** is the testing which is performed on an application when it is being executed.ex: Functional testing.

🌱 **Alpha Testing:** it is a type of user acceptance testing, which is conducted on an application when it is just before released to the customer.

🌱 **Beta-Testing:** it is a type of UAT that is conducted on an application when it is released to the customer, when deployed in to the real time environment and being accessed by the real time users.

🌱 **Monkey Testing:** is the process in which abnormal operations, beyond capacity operations are done on the application to check the stability of it in spite of the users abnormal behavior.

🌱 **Compatibility testing:** it is the testing process in which usually the products are

tested on the environments with different combinations of databases (application servers, browsers...etc) In order to check how far the product is compatible with all these environments platform combination.

🌱 **Installation Testing:** it is the process of testing in which the tester try to install or try to deploy the module into the corresponding environment by following the guidelines produced in the deployment document and check whether the installation is successful or not.

🌱 **Adhoc Testing:** Adhoc Testing is the process of testing in which unlike the formal testing where in test case document is used, with out that test case document testing can be done of an application, to cover that testing of the future which are not covered in that test case document. Also it is intended to perform GUI testing which may involve the cosmetic issues.

TCD (Test Case Document:

Test Case Document Contains

- **Test Scope (or) Test objective**
- **Test Scenario**
- **Test Procedure**
- **Test case**

This is the sample test case document for the Academic details of student project:

Test scope:

- Test coverage is provided for the screen “ Academic status entry” form of a student module of university management system application
- Areas of the application to be tested

Test Scenario:

- When the office personals use this screen for the marks entry, calculate the status details, saving the information on student's basis and quit the form.

Test Procedure:

- The procedure for testing this screen is planned in such a way that the data entry, status calculation functionality, saving and quitting operations are tested in terms of Gui testing, Positive testing, Negative testing using the corresponding

Gui test

cases, Positive test cases, Negative test cases respectively

3.6. SYSTEM REQUIREMENTS

3.6.1. Software Requirements

Operating System	:	Windows XP/2003 or Linux/Solaris
User Interface	:	HTML, CSS
Client-side Scripting	:	Python
Programming Language	:	Python
Web Applications	:	JDBC, JNDI, Servlets, JSP
IDE/Workbench	:	Anaconda
Database	:	SQLite
Server Deployment	:	Tomcat6.0

3.6.2. Hardware Requirements

Processor	:	Pentium IV
Hard Disk	:	40GB
RAM	:	256MB

3.7 NON-FUNCTIONAL REQUIREMENTS:

Performance Requirements

Some Performance requirements identified is listed below:

- The database shall be able to accommodate a minimum of 10,000 records of students.
- The software shall support use of multiple users at a time. There are no other specific performance requirements that will affect development.

Security Requirements

Some of the factors that are identified to protect the software from accidental or malicious access, use, modification, destruction, or disclosure are described below. Specific requirements in this area could include the need to:

- Utilize certain cryptographic techniques
- Keep specific log or history data sets
- Assign certain functions to different modules
- Restrict communications between some areas of the program
- Check data integrity for critical variables
- Later version of the software will incorporate encryption techniques in the user/license authentication process.
- The software will include an error tracking log that will help the user understand what error occurred when the application crashed along with suggestions on how to prevent the error from occurring again.
- Communication needs to be restricted when the application is validating the user or license.(i.e., using https)

Portability Requirements

Some of the attributes of software that relate to the ease of porting the software to other host machines and/or operating systems.

This may include:

Java is used to develop the product. So it is easiest to port the software in any environment.

Maintainability

The user will be able to reset all options and all stored user variables to default settings.

Reliability

Some of the attributes identified for the reliability is listed below:

- All data storage for user variables will be committed to the database at the time of entry.
- Data corruption is prevented by applying the possible backup procedures and techniques.

Usability requirements

Some of the usability requirements identified for this system are listed below:

- A logical interface is essential to an easy to use system, speeding up common tasks.
- Error prevention is integral to the system and is provided in a number of formats from sanity hacks to limiting free-text input.

Availability

- All cached data will be rebuilt during every startup. There is no recovery of user data if it is lost. Default values of system data will be assigned when necessary

Software System Attributes

- There are a number of attributes of software that can serve as requirements. It is important that required attributes be specified so that their achievement can be objectively verified. The following items provide a partial list of examples.
- The input system will allow for inputting numbers, operands, special symbols and letters of the alphabet.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 SUMMARY

PUBG analysis using data science' is an end-to-end case study to understand the different stages in the Data Science life cycle. This will deal with 'data manipulation' with pandas, Numpy, and 'data visualization' with Matplotlib with the PUBG dataset. After Data manipulation and Data visualization, an ML model will be built on the PUBG dataset to get predictions. I am using sci-kit learn library to implement the machine learning algorithm.

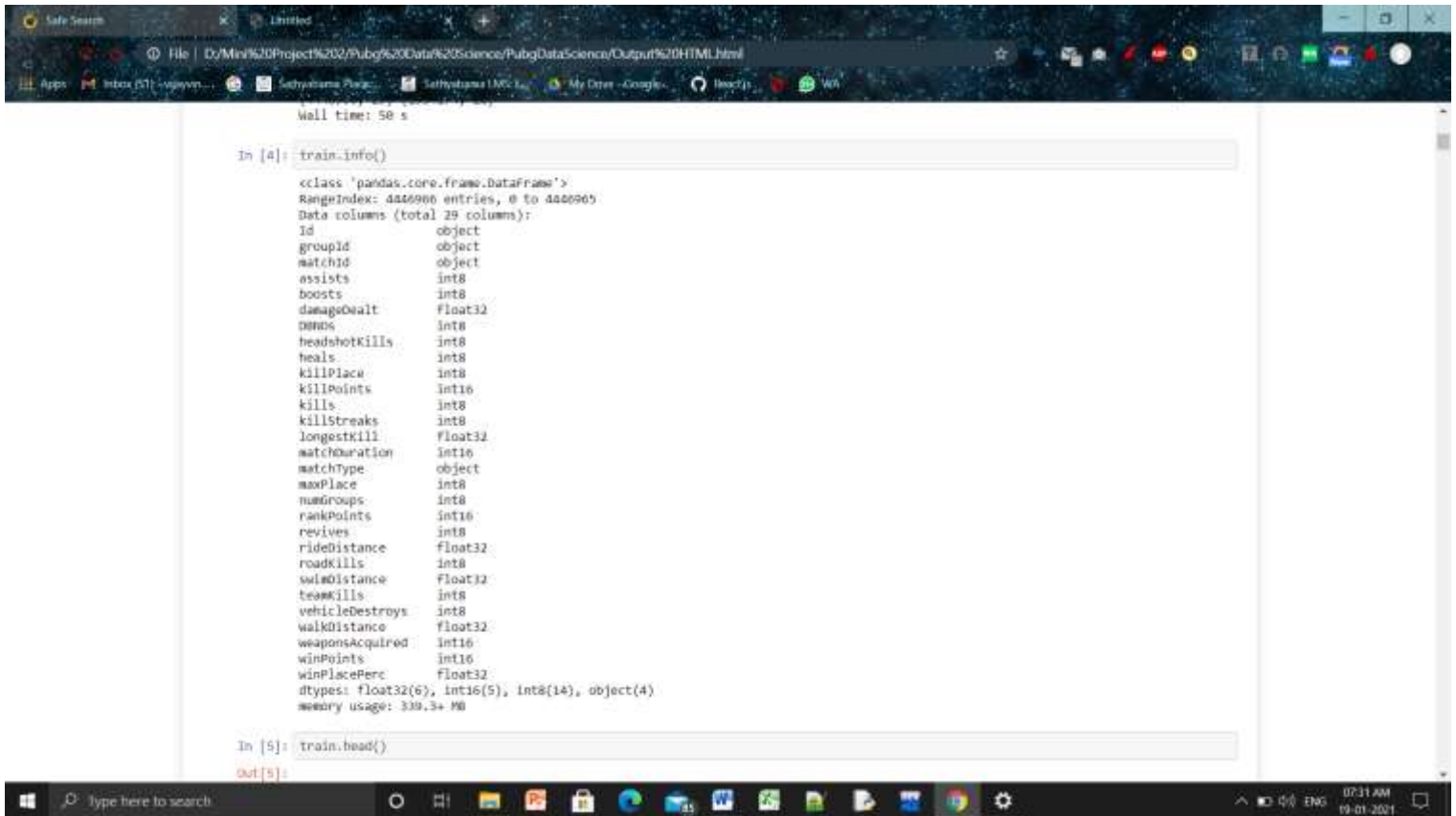
In a PUBG game, up to 100 players start in each match (matchId). Players (Id) can be on teams (groupId) which get ranked at the end of the game (winPlacePerc) based on how many other teams are still alive when they are eliminated. During the game, players can pick up different amunitions, revive downed-but-not-dead (knocked) teammates, drive vehicles, swim, run, shoot, and experience all of the consequences — such as falling too far or running themselves over and eliminating themselves.

The team at PUBG has made official game data available for the public to explore and scavenge outside of "The Blue Circle." This is not an official or affiliated PUBG site. Its based on the data collected by Kaggle and made available through the PUBG Developer API.

You are provided with a large number of anonymized PUBG game stats, formatted so that each row contains one player's post-game stats. The data comes from matches of all types: solos, duos, squads, and custom; there is no guarantee of there being 100 players per match, nor at most 4 player per group.

4.2 SCREEN SHOTS

Page 1:



```
Wall time: 58 s

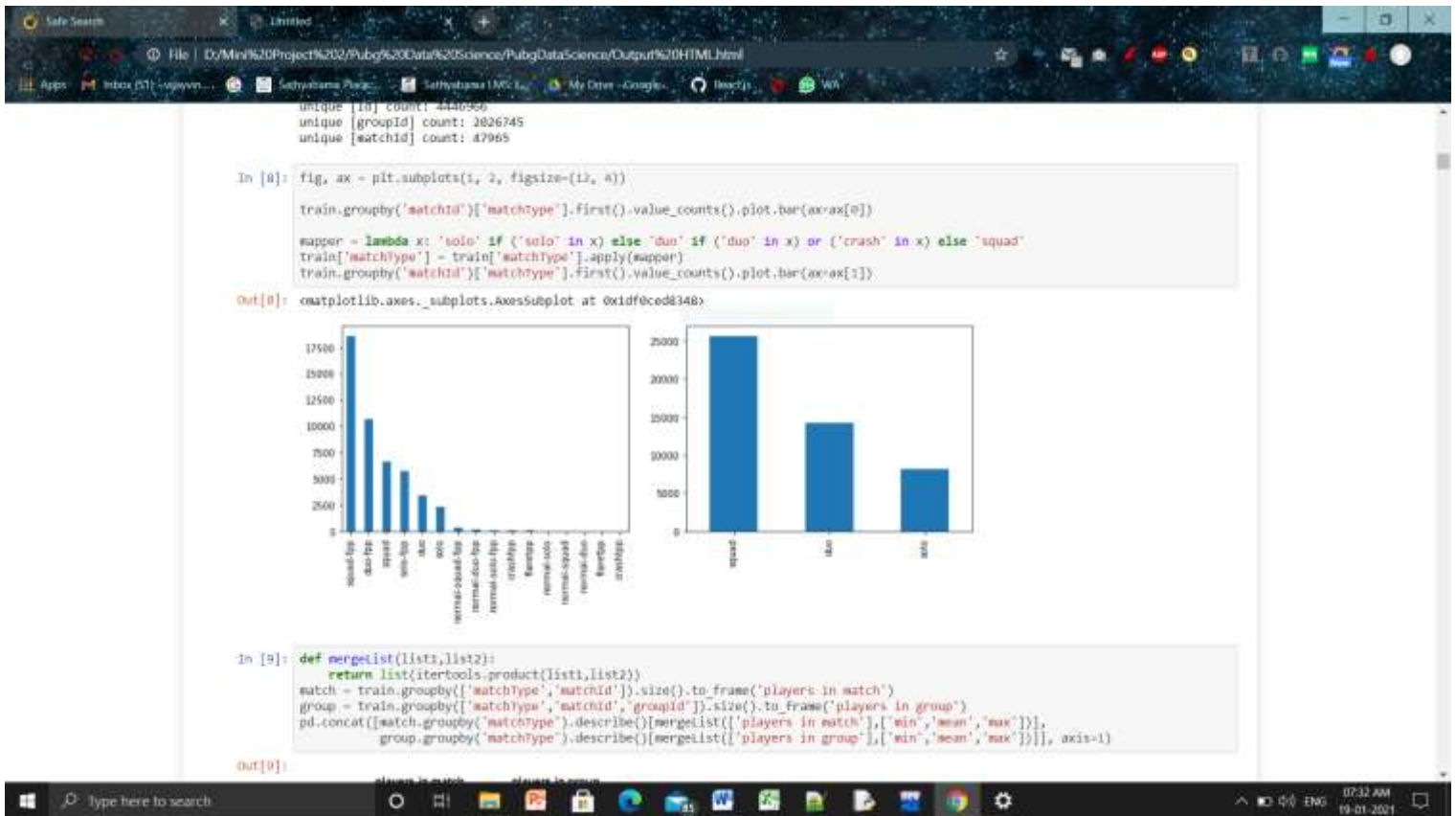
In [4]: train.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4446966 entries, 0 to 4446965
Data columns (total 29 columns):
Id                object
groupId           object
matchId           object
assists           int8
boosts            int8
damageDealt       float32
deaths            int8
headshotKills     int8
heals             int8
killPlace        int8
killPoints        int16
kills             int8
killStreaks       int8
longestKill       float32
matchDuration     int16
matchType         object
mapPlace          int8
numGroups         int8
rankPoints        int16
revives           int8
rideDistance      float32
roadKills         int8
swimDistance      float32
teamKills         int8
vehicleDestroys   int8
walkDistance      float32
weaponsAcquired   int16
winPoints         int16
winPlacePerc      float32
dtypes: float32(6), int16(5), int8(14), object(4)
memory usage: 339.3+ MB

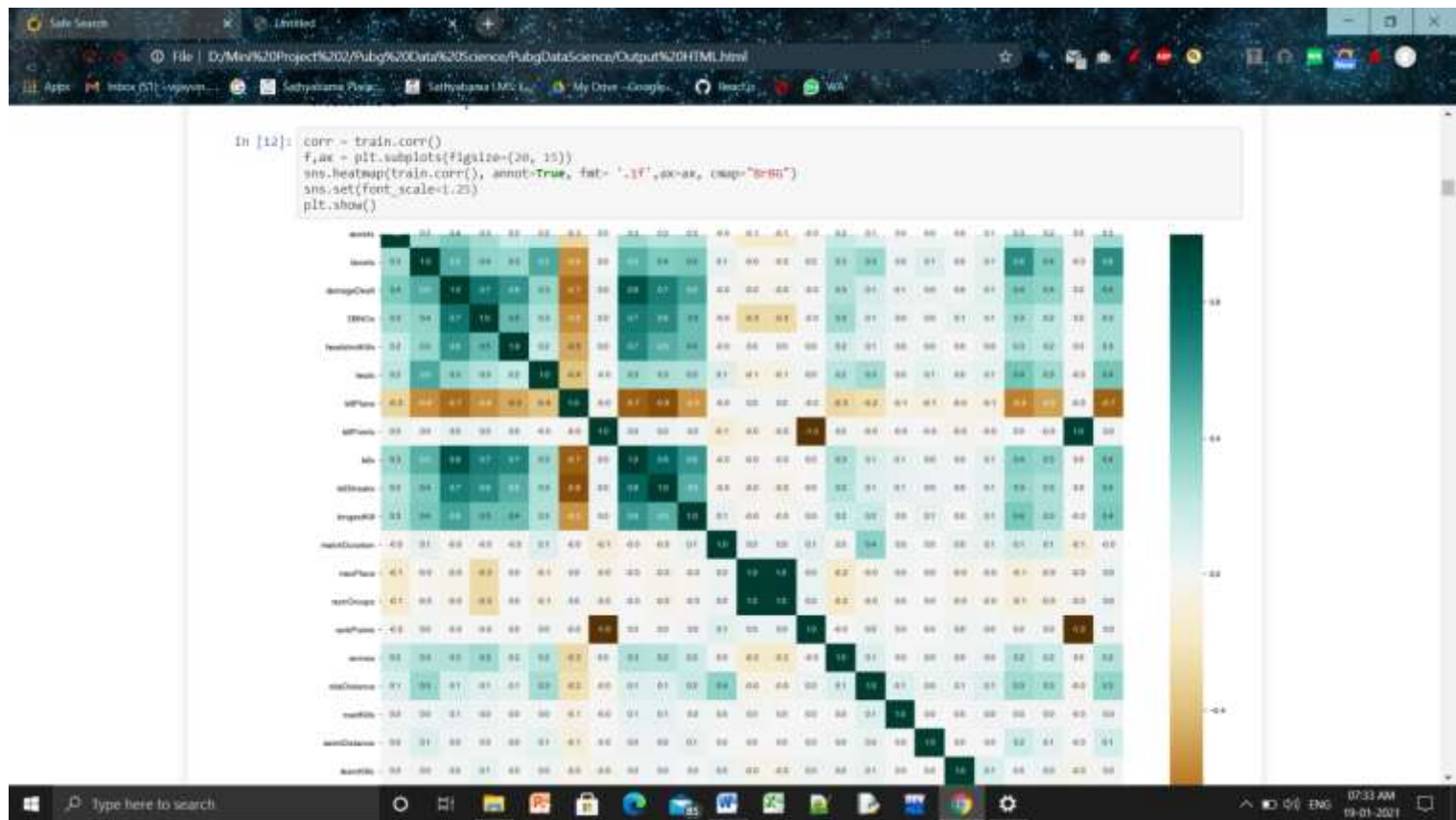
In [5]: train.head()

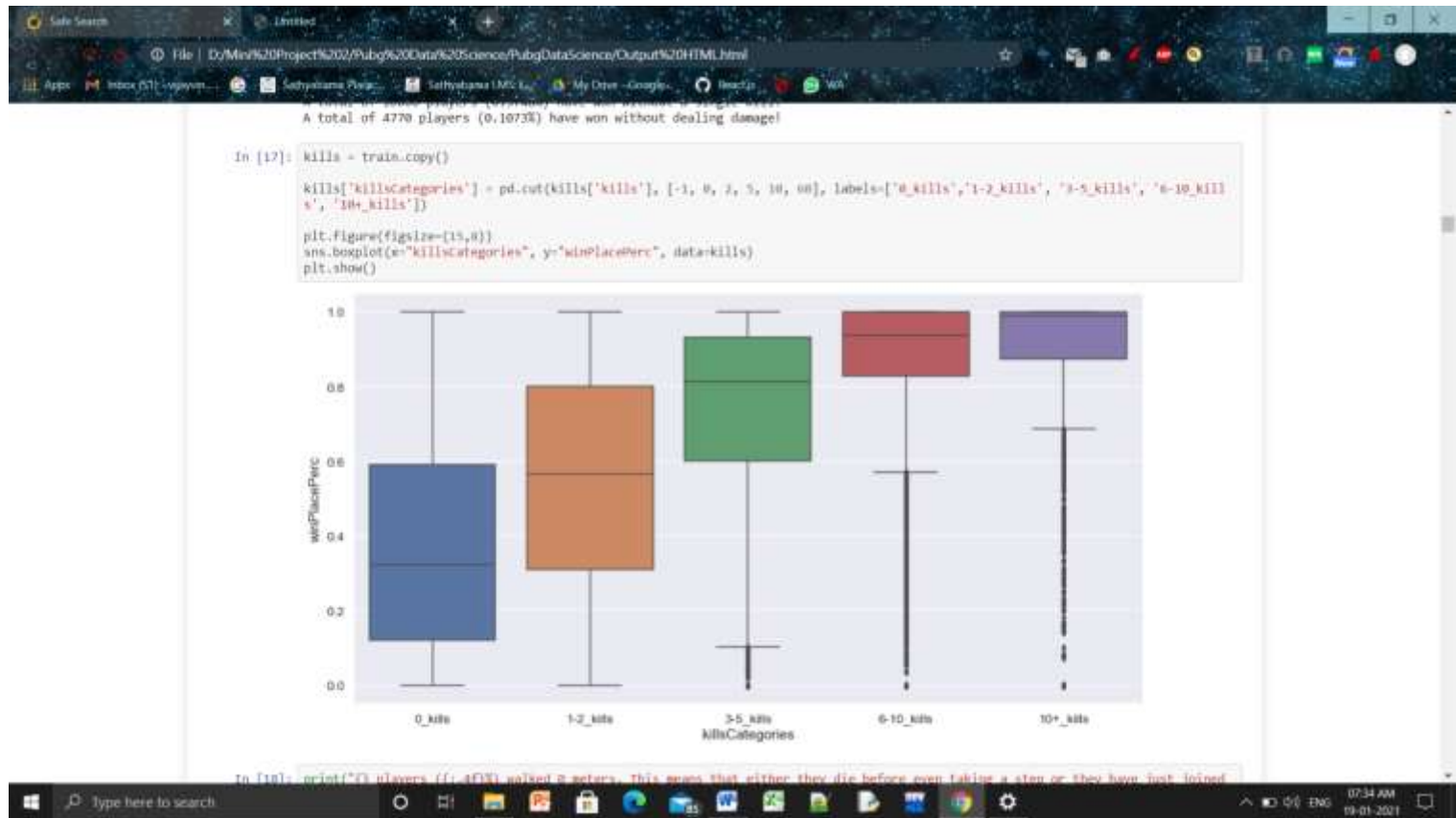
Out[5]:
```

Page 2:



Page 3:

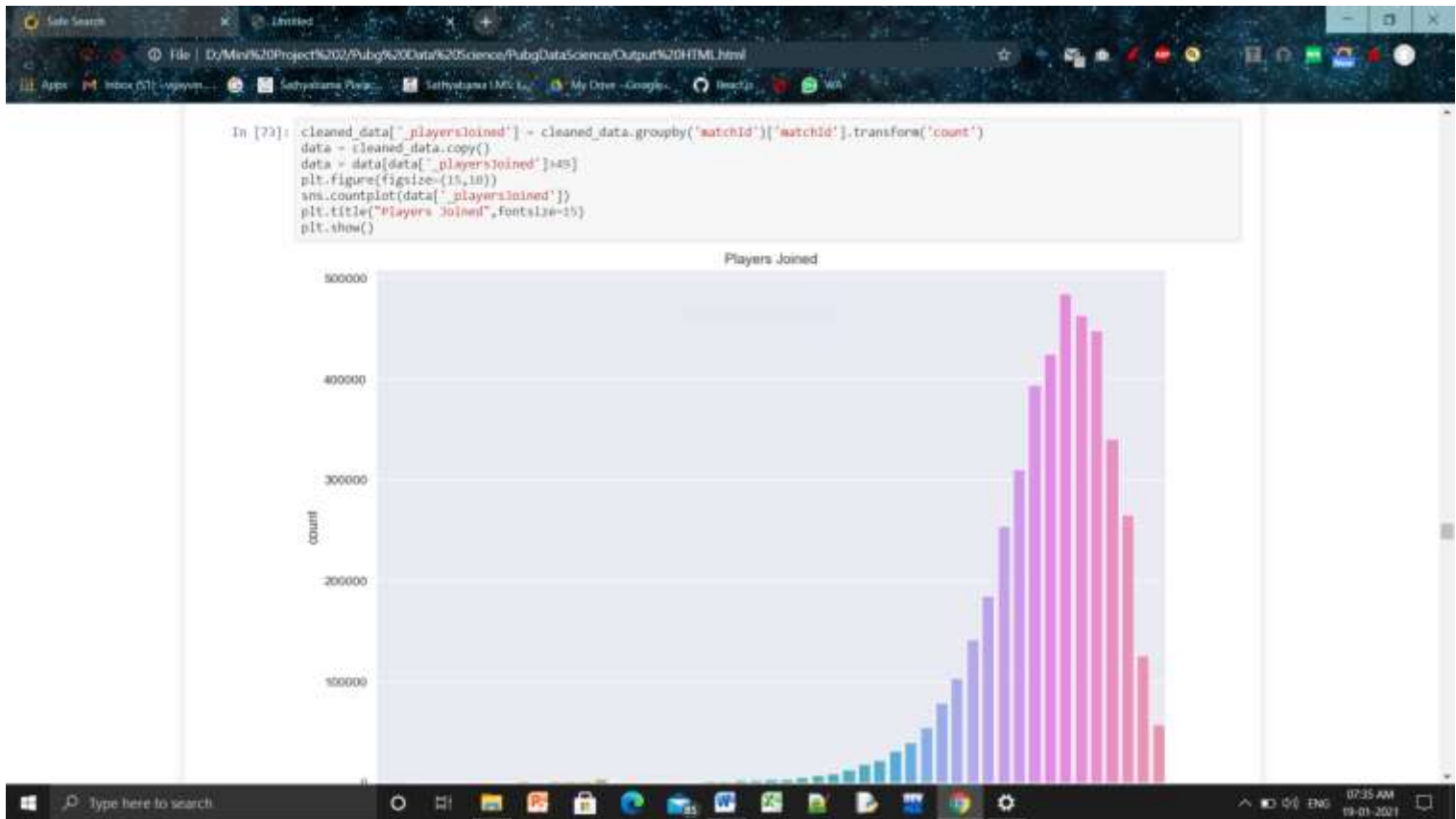




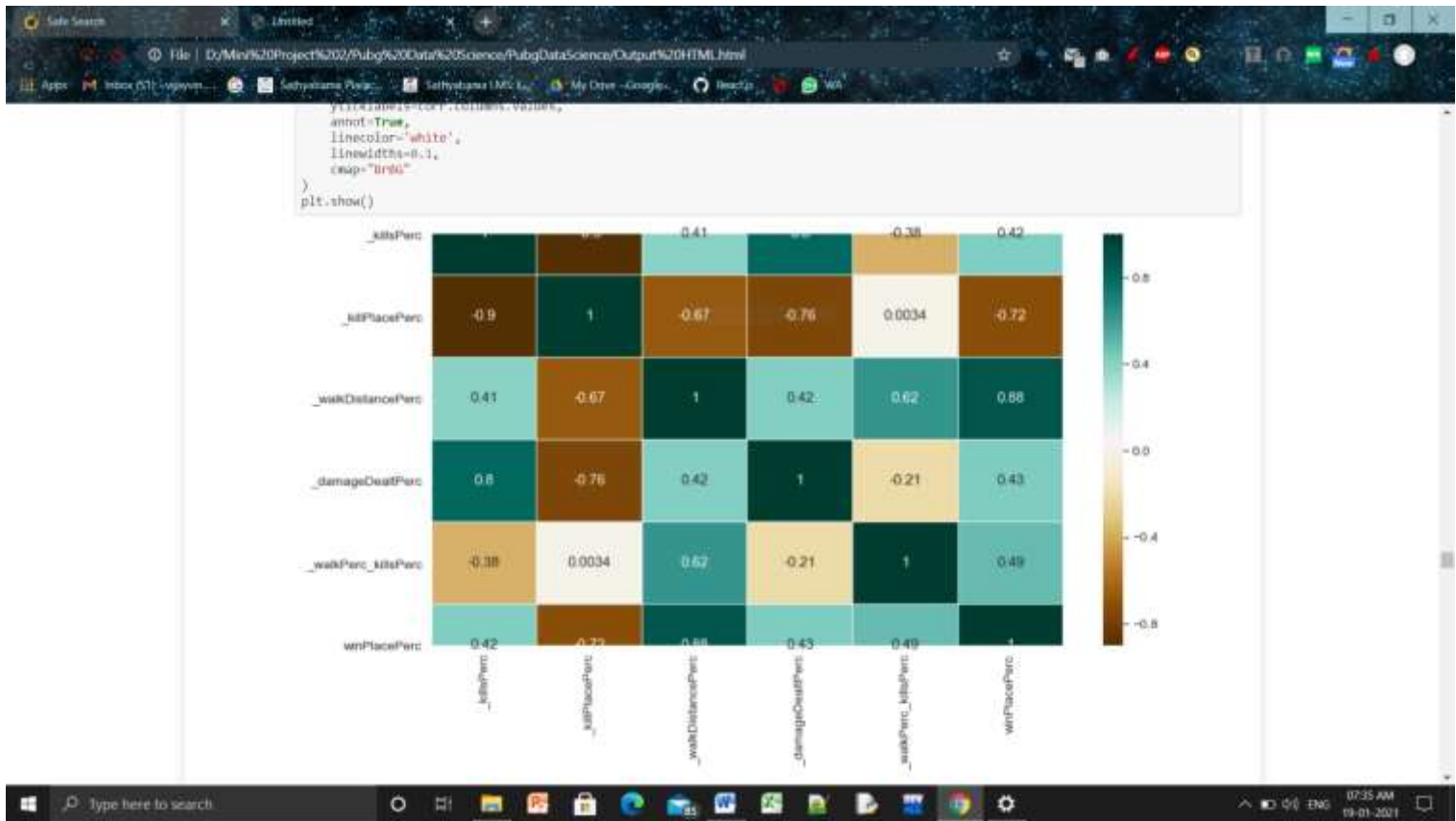
Page 5:



Page 6:



Page 7:



CHAPTER 5

CONCLUSION AND FUTURE ENHANCEMENT

5.1. CONCLUSION

Data Science makes use of several statistical procedures. These procedures range from data transformations, data modeling, statistical operations (descriptive and inferential statistics) and machine learning modeling. Statistics is the primary asset of every Data Scientist. In order to gain predictive responses from the models, it is an essential requirement to understand the underlying patterns of the data model. Furthermore, optimization techniques can be utilized to meet the business requirements of the user.

Using various statistical tools, a Data Scientist has to develop models. With the help of these models, they help their clients in the decision-making process. Furthermore, these models support demand generation initiatives.

Analytic objectives and approaches are planned and defined by the Data Scientists who collaborate with the internal consulting team. Data Scientists also formulate work plans to provide support – programming as well as analytical to internal consulting. There is also a provision of statistical procedures that utilize Microsoft Office and SAS suite.

5.2. FUTURE ENHANCEMENT

Data Science is a colossal pool of multiple data operations. These data operations also involve machine learning and statistics. Machine Learning algorithms are very much dependent on data. This data is fed to our model in the form of training set and test set which is eventually used for fine-tuning our model with various algorithmic parameters. By all means, advancement in Machine Learning is the key contributor towards the future of data science.

In particular, Data Science also covers:

- a. Data Integration.
- b. Distributed Architecture.

- c. Automating Machine learning.
- d. Data Visualisation.
- e. Dashboards and BI.
- f. Data Engineering.
- g. Deployment in production mode
- h. Automated, data-driven decisions.

5.3. REFERENCES :

- N. Schuessler and K. Axhausen, "Processing raw data from global positioning systems without additional information," *Transp. Res. Rec., J. Transp. Res. Board*, no. 2105, pp. 28–36, 2009.
- J. Wolf, "Using GPS data loggers to replace travel diaries in the collection of travel data," Georgia Inst. Technol., Citeseer, Tech. Rep., 2000.
- J. Ogle, R. Guensler, W. Bachman, M. Koutsak, and J. Wolf, "Accuracy of global positioning system for determining driver performance parameters," *Transp. Res. Rec., J. Transp. Res. Board*, no. 1818, pp. 12–24, 2002.

APPENDIX

6.1 SOURCE CODE

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import itertools
import warnings
warnings.filterwarnings("ignore")
# Memory saving function credit to https://www.kaggle.com/gemartin/load-data-reduce-memory-usage
def reduce_mem_usage(df):
    """ iterate through all the columns of a dataframe and modify the data type
        to reduce memory usage.
    """
    start_mem = df.memory_usage().sum() / 1024**2

    for col in df.columns:
        col_type = df[col].dtype

        if col_type != object:
            c_min = df[col].min()
            c_max = df[col].max()
            if str(col_type)[:3] == 'int':
                if c_min > np.iinfo(np.int8).min and c_max < np.iinfo(np.int8).max:
```

```

        df[col] = df[col].astype(np.int8)
    elif c_min > np.iinfo(np.int16).min and c_max < np.iinfo(np.int16).max:
        df[col] = df[col].astype(np.int16)
    elif c_min > np.iinfo(np.int32).min and c_max < np.iinfo(np.int32).max:
        df[col] = df[col].astype(np.int32)
    elif c_min > np.iinfo(np.int64).min and c_max < np.iinfo(np.int64).max:
        df[col] = df[col].astype(np.int64)
    else:
        #if c_min > np.finfo(np.float16).min and c_max < np.finfo(np.float16).max:
        #    df[col] = df[col].astype(np.float16)
        #el
        if c_min > np.finfo(np.float32).min and c_max < np.finfo(np.float32).max:
            df[col] = df[col].astype(np.float32)
        else:
            df[col] = df[col].astype(np.float64)
    #else:
    #df[col] = df[col].astype('category')

end_mem = df.memory_usage().sum() / 1024**2
print('Memory usage of dataframe is {:.2f} MB --> {:.2f} MB (Decreased by
{:.1f}% )'.format(
    start_mem, end_mem, 100 * (start_mem - end_mem) / start_mem))
return df

%%time
train = pd.read_csv('train_V2.csv')
train = reduce_mem_usage(train)
test = pd.read_csv('test_V2.csv')
test = reduce_mem_usage(test)

```

```
print(train.shape, test.shape)
```

```
    print(f'unique [{i}] count:', train[i].nunique())
```

```
unique [Id] count: 4446966
```

```
unique [groupId] count: 2026745
```

```
unique [matchId] count: 47965
```

```
fig, ax = plt.subplots(1, 2, figsize=(12, 4))
```

```
train.groupby('matchId')['matchType'].first().value_counts().plot.bar(ax=ax[0])
```

```
mapper = lambda x: 'solo' if ('solo' in x) else 'duo' if ('duo' in x) or ('crash' in x) else  
'squad'
```

```
train['matchType'] = train['matchType'].apply(mapper)
```

```
train.groupby('matchId')['matchType'].first().value_counts().plot.bar(ax=ax[1])
```

```
def mergeList(list1,list2):
```

```
    return list(itertools.product(list1,list2))
```

```
match = train.groupby(['matchType','matchId']).size().to_frame('players in match')
```

```
group = train.groupby(['matchType','matchId','groupId']).size().to_frame('players in  
group')
```

```
pd.concat([match.groupby('matchType').describe()[mergeList(['players in  
match'], ['min','mean','max'])],
```

```
          group.groupby('matchType').describe()[mergeList(['players in  
group'], ['min','mean','max'])]], axis=1)
```

```
'''
```

```
subset = train[train['matchId']=='3e029737889ce9']
```

```
sub_grp = subset[subset['groupId']=='b8275198faa03b']
```

```

print('matchId ==\`3e029737889ce9\` & groupId ==\`b8275198faa03b\`)
print('-'*50)
print('players:',len(subset))
print('groups:',subset['groupId'].nunique())
print('numGroups:',subset['numGroups'].unique())
print('maxPlace:',subset['maxPlace'].unique())
print('-'*50)
print('max-group players:',len(sub_grp))
print('max-group winPlacePerc:',sub_grp['winPlacePerc'].unique())
print('-'*50)
print('winPlacePerc:',subset['winPlacePerc'].sort_values().unique())

```

```

corr = train.corr()
f,ax = plt.subplots(figsize=(20, 15))
sns.heatmap(train.corr(), annot=True, fmt= '.1f',ax=ax, cmap="BrBG")
sns.set(font_scale=1.25)
plt.show()

```

```

print("On an average a person kills {:.4f} players, 99% of people have {} kills or less,
while the maximum kills ever recorded is
{}.".format(train['kills'].mean(),train['kills'].quantile(0.99), train['kills'].max()))

```

On an average a person kills 0.9248 players, 99% of people have 7.0 kills or less, while the maximum kills ever recorded is 72.

```

data = train.copy()
data.loc[data['kills'] > data['kills'].quantile(0.99)] = '8+'
plt.figure(figsize=(15,8))
sns.countplot(data['kills'].astype('str').sort_values())

```

```
plt.title("Kill Count",fontsize=14)
```

```
plt.show()
```

```
print("A total of {} players {:.4f}% have won without a single  
kill!".format(len(data[data['winPlacePerc']==1]),  
100*len(data[data['winPlacePerc']==1])/len(train)))
```

```
data1 = train[train['damageDealt'] == 0].copy()
```

```
print("A total of {} players {:.4f}% have won without dealing  
damage!".format(len(data1[data1['winPlacePerc']==1]),  
100*len(data1[data1['winPlacePerc']==1])/len(train)))
```

A total of 16666 players (0.3748%) have won without a single kill!

A total of 4770 players (0.1073%) have won without dealing damage!

```
kills = train.copy()
```

```
kills['killsCategories'] = pd.cut(kills['kills'], [-1, 0, 2, 5, 10, 60], labels=['0_kills','1-2_kills',  
'3-5_kills', '6-10_kills', '10+_kills'])
```

```
plt.figure(figsize=(15,8))
```

```
sns.boxplot(x="killsCategories", y="winPlacePerc", data=kills)
```

```
plt.show()
```

```
print("{} players {:.4f}% walked 0 meters. This means that either they die before even  
taking a step or they have just joined the game but are away from keyboard (more  
possible)".format(len(data[data['walkDistance'] == 0]),  
100*len(data1[data1['walkDistance']==0])/len(train)))
```

98054 players (2.0329%) walked 0 meters. This means that either they die before even taking a step or they have just joined the game but are away from keyboard (more possible).

```
data = train.copy()
```

```
data = data[data['walkDistance'] < train['walkDistance'].quantile(0.99)]
```

```
plt.figure(figsize=(15,10))
```

```
plt.title("Walking Distance Distribution",fontsize=15)
sns.distplot(data['walkDistance'])
plt.show()
```

```
f,ax1 = plt.subplots(figsize =(15,8))
sns.pointplot(x='vehicleDestroys',y='winPlacePerc',data=data,color='lime',alpha=0.5)
plt.xlabel('Count of Vehicle Destroys',fontsize = 16,color='blue')
plt.ylabel('Win Percentage',fontsize = 16,color='blue')
plt.title('Vehicle Destroyed/ Win Ratio',fontsize = 20,color='blue')
plt.grid()
plt.show()
```

```
print("In the game on an average a person uses {:.1f} heal items, 99% of people use {}
or less, while the doctor used {}".format(train['heals'].mean(),
train['heals'].quantile(0.99), train['heals'].max()))
```

```
print("In the game on an average a person uses {:.1f} boost items, 99% of people use
{} or less, while the doctor used {}".format(train['boosts'].mean(),
train['boosts'].quantile(0.99), train['boosts'].max()))
```

In the game on an average a person uses 1.4 heal items, 99% of people use 12.0 or less, while the doctor used 80.

In the game on an average a person uses 1.1 boost items, 99% of people use 7.0 or less, while the doctor used 33.

```
data = train.copy()
data = data[data['heals'] < data['heals'].quantile(0.99)]
data = data[data['boosts'] < data['boosts'].quantile(0.99)]
```

```
f,ax1 = plt.subplots(figsize =(15,8))
sns.pointplot(x='heals',y='winPlacePerc',data=data,color='lime',alpha=0.8)
sns.pointplot(x='boosts',y='winPlacePerc',data=data,color='blue',alpha=0.8)
```

```

plt.text(4,0.6,'Heals',color='lime',fontsize = 16,style = 'italic')
plt.text(4,0.55,'Boosts',color='blue',fontsize = 16,style = 'italic')
plt.xlabel('Number of heal/boost items',fontsize = 16,color='blue')
plt.ylabel('Win Percentage',fontsize = 16,color='blue')
plt.title('Heals vs Boosts',fontsize = 20,color='blue')
plt.grid()
plt.show()

```

```

solos = train[train['numGroups']>50] #Solo game played by more than 50 different
groups
duos = train[(train['numGroups']>25) & (train['numGroups']<=50)] #Duo game can
consist of a maximum of 50 groups(50x2=100)
squads = train[train['numGroups']<=25] # Squad game can consist a maximum of 25
groups(25x4=100)
print("{} ({:.2f}%) of games are solo, {} ({:.2f}%) are duo and {} ({:.2f}%) are squad
games.".format(len(solos), 100*len(solos)/len(train), len(duos),
100*len(duos)/len(train), len(squads), 100*len(squads)/len(train),))
709111 (15.95%) of games are solo, 3295326 (74.10%) are duo and 442529 (9.95%)
are squad games.
f,ax1 = plt.subplots(figsize =(20,10))
sns.pointplot(x='kills',y='winPlacePerc',data=solos,color='black',alpha=0.6)
sns.pointplot(x='kills',y='winPlacePerc',data=duos,color='red',alpha=0.6)
sns.pointplot(x='kills',y='winPlacePerc',data=squads,color='blue',alpha=0.6)
plt.text(14,0.5,'Solos',color='black',fontsize = 16)
plt.text(14,0.45,'Duos',color='red',fontsize = 16)
plt.text(14,0.4,'Squads',color='blue',fontsize = 16)
plt.xlabel('Number of kills',fontsize = 15,color='blue')
plt.ylabel('Win Percentage',fontsize = 15,color='blue')

```

```
plt.title('Solo vs Duo vs Squad Kills',fontsize = 20,color='blue')
plt.grid()
plt.show()
```

```
train['_totalDistance'] = train['rideDistance'] + train['walkDistance'] +
train['swimDistance']
```

```
# Engineer _headshot_rate feature --- headshots made per kill
```

```
train['_headshot_rate'] = train['headshotKills'] / train['kills']
```

```
train['_headshot_rate'] = train['_headshot_rate'].fillna(0)
```

```
#Defining some functions for plotting graphs, we will be needing a lot of countplot and
distplot
```

```
def show_countplot(column):
```

```
    plt.figure(figsize=(15,8))
```

```
    sns.countplot(data=train, x=column).set_title(column)
```

```
    plt.show()
```

```
def show_distplot(column):
```

```
    plt.figure(figsize=(15, 8))
```

```
    sns.distplot(train[column], bins=50)
```

```
    plt.show()
```

```
show_countplot('kills')
```

```
cleaned_data['_killsNorm'] = cleaned_data['kills']*((100-
cleaned_data['_playersJoined'])/100 + 1)
```

```
cleaned_data['_damageDealtNorm'] = cleaned_data['damageDealt']*((100-
cleaned_data['_playersJoined'])/100 + 1)
```

```
cleaned_data['_maxPlaceNorm'] = cleaned_data['maxPlace']*((100-
cleaned_data['_playersJoined'])/100 + 1)
```

```
cleaned_data['_matchDurationNorm'] = cleaned_data['matchDuration']*((100-
cleaned_data['_playersJoined'])/100 + 1)
```

```

# Compare standard features and normalized features
to_show = ['Id', 'kills', '_killsNorm', 'damageDealt', '_damageDealtNorm', 'maxPlace',
'_maxPlaceNorm', 'matchDuration', '_matchDurationNorm']

cleaned_data[to_show][0:11]

agg = cleaned_data.groupby(['groupId']).size().to_frame('players_in_team')
cleaned_data = cleaned_data.merge(agg, how='left', on=['groupId'])
cleaned_data['_healthItems'] = cleaned_data['heals'] + cleaned_data['boosts']
cleaned_data['_headshotKillRate'] = cleaned_data['headshotKills'] / cleaned_data['kills']
cleaned_data['_killPlaceOverMaxPlace'] = cleaned_data['killPlace'] /
cleaned_data['maxPlace']
cleaned_data['_killsOverWalkDistance'] = cleaned_data['kills'] /
cleaned_data['walkDistance']
cleaned_data['_killsOverDistance'] = cleaned_data['kills'] /
cleaned_data['_totalDistance']
cleaned_data['_walkDistancePerSec'] = cleaned_data['walkDistance'] /
cleaned_data['matchDuration']
cleaned_data.head()

```

Drop Id column, because it probably won't be useful for our Machine Learning algorithm,

```

# because the test set contains different Id's
highly_corr.drop(columns = ['Id'], inplace=True)
# Take sample for debugging and exploration
sample = 500000
df_sample = highly_corr.sample(sample)
# Split sample into training data and target variable
df = df_sample.drop(columns = ['winPlacePerc']) #all columns except target
y = df_sample['winPlacePerc'] # Only target variable
# Function for splitting training and validation data

```

```

def split_vals(a, n : int):
    return a[:n].copy(), a[n:].copy()
val_perc = 0.12 # % to use for validation set
n_valid = int(val_perc * sample)
n_trn = len(df)-n_valid
# Split data
raw_train, raw_valid = split_vals(df_sample, n_trn)
X_train, X_valid = split_vals(df, n_trn)
y_train, y_valid = split_vals(y, n_trn)

# Check dimensions of samples
print('Sample train shape: ', X_train.shape,
      'Sample target shape: ', y_train.shape,
      'Sample validation shape: ', X_valid.shape)
Sample train shape: (440000, 34) Sample target shape: (440000,) Sample validation
shape: (60000, 34)
# Metric used for the PUBG competition (Mean Absolute Error (MAE))
from sklearn.metrics import mean_absolute_error
from sklearn.ensemble import RandomForestRegressor

# Function to print the MAE (Mean Absolute Error) score

def print_score(m : RandomForestRegressor):
    res = ['mae train: ', mean_absolute_error(m.predict(X_train), y_train),
          'mae val: ', mean_absolute_error(m.predict(X_valid), y_valid)]
    #Score of the training dataset obtained using an out-of-bag estimate.
    if hasattr(m, 'oob_score_'): res.append(m.oob_score_)
    print(res)

```

```

# Train basic model
m1 = RandomForestRegressor(n_estimators=40, min_samples_leaf=3,
max_features='sqrt', n_jobs=-1)
m1.fit(X_train, y_train)
print_score(m1)
['mae train: ', 0.030842484940609604, 'mae val: ', 0.051213798273597545]
def rf_feat_importance(m, df):
    return pd.DataFrame({'cols':df.columns,
'imp':m.feature_importances_}).sort_values('imp', ascending=False)
# What are the most predictive features according to our basic random forest model
fi = rf_feat_importance(m1, df); fi[:15]

```