INDUSTRIAL TRAINING REPORT ON "DATA SCIENCE WITH PYTHON"

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology In Electronics & Communication Engineering

Submitted To: - Submitted By: - Nitish Kumar

Ms. Juhi Priyani

(Assistant Professor), ECE Enrollment No:- 00115607321



Department of Electronics & Communication Engineering
Dr. Akhilesh Das Gupta Institute of Technology & Management
Affiliated to Guru Gobind Singh Indraprastha University
Dwarka, New Delhi-110078.

CERTIFICATE





PROJECT COMPLETION CERTIFICATE

In recognition of the commitment to achieve professional excellence this is to certify that Ms./Mr.

NITISH KUMAR

has successfully completed an Industry-oriented project.

Project Name	IPL Data Analysis
Technologies Used	Data Science & Machine Learning - Jupyter Notebook
Reference No.	CE/2022/F/AUG/00290
Training Date	2 nd August, 2022 – 17 th September, 2022
Training Duration	90 Hours
Training Location	Dr. Akhilesh Das Gupta Institute of Technology & Management

Program Co-ordinator Industry/Academic Alliance



Training and Development Allsoft Solutions and Services

BIG DATA - ANALYTICS | IOT



ORACLE





CLOUD COMPUTING

ACKNOWLEDGEMENT

I would like to acknowledge the contributions of the following people, without whose help and

guidance this report would not have been completed.

I acknowledge the counsel and support of our training coordinator, Juhi Priyani, Assistant

Professor, ECE Department, with respect and gratitude, whose expertise, guidance, support,

encouragement, and enthusiasm has made this report possible. Their feedback vastly improved

the quality of this report and provided an enthralling experience. I am indeed proud and

fortunate to be supported by him/her.

I am also thankful to Prof. (Dr.) Niranjan Bhattacharya, H.O.D of Electronics and

Communication Engineering Department, Dr. Akhilesh Das Gupta Institute of

Technology & Management, New Delhi for his constant encouragement, valuable

suggestions and moral support and blessings.

Although it is not possible to name individually, I shall ever remain indebted to the faculty

members of Dr. Akhilesh Das Gupta Institute of Technology & Management, New Delhi

for their persistent support and cooperation extended during this work...

This acknowledgement will remain incomplete if I fail to express our deep sense of obligation

to my parents and God for their consistent blessings and encouragement.

Name of Student- Nitish Kumar

Enrollment No.- 00115607321

Signature (Assistant Professor)

ABSTRACT

In the Gujarat Titans Data Analysis Project the data is fetched and the desired information is been displayed over promptly. This project uses the python language with good enough easy to understand libraries. Gujarat Titan Data Analysis synchronysis the details of players to the selectors, which leads to the great enemuration in the tech field data analysis.

TABLE OF CONTENT

Certificate

Acknowledgement

Abstract

Table of content

CHAPTER 1: INTRODUCTION AND LITERATURE OVERVIEW

1.1 PYTHON

- History of python
- Ptython features
- Libraries

CHAPTER 2: TECHNOLOGY ADOPTED

2.1 DATA SCIENCE

2.2 MACHINE LEARNING

- How machine learning works?
- Advantages of machine learning

CHAPTER 3: DATA ANSALYSIS PROJECT

- Overview of Project
- Libraries
- Data Exploration

CHAPTER 4: CONCLUSION AND FUTIURE SCOPE

- Conclusion
- Future scope

REFRENCES

CHAPTER 1 – INTRODUCTION

1.1 PYTHON



Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

• Python is Interpreted:

Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

• Python is Interactive:

You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python is Object-Oriented:

Python supports Object-Oriented style or technique of programming that encapsulates code within objects.

• Python is a Beginner's Language:

Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

Python has a big list of good features:

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.

• IT supports automatic garbage collection.

It can be easily integrated with C, C++, COM, ActiveX, CORBA, and java.

1.1.1 HISTORY OF PYTHON

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands. Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, Small Talk, and Unix shell and other scripting languages. Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL). Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

1.1.2 PYTHON FEATURES

Python's features include:

• Easy-to-learn:

Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.

• Easy-to-read:

Python code is more clearly defined and visible to the eyes.

• Easy-to-maintain:

Python's source code is fairly easy-to-maintain.

• A broad standard library:

Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.

• Interactive Mode:

Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.

• Portable:

Python can run on a wide variety of hardware platforms and has the same interface on all platforms.

• Extendable:

You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.

• Databases:

Python provides interfaces to all major commercial databases.

• GUI Programming:

Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the XWindow system of Unix.

• Scalable:

Python provides a better structure and support for large programs than shell scripting.

1.1.3 LIBRARIES

• NUMPY:

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- 1. a powerful N-dimensional array object
- 2. sophisticated (broadcasting) functions
- 3. tools for integrating C/C++ and Fortran code
- 4. useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multidimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

NumPy is licensed under the BSD license, enabling reuse with few restrictions. The core functionality of NumPy is its "ND array", for n-dimensional array, data structure. These arrays are stride views on memory. In contrast to Python's built-in list data structure (which, despite the name, is a dynamic array), these arrays are homogeneously typed: all elements of a single array must be of the same type. NumPy has built-in support for memory mapped arrays

- 1. zeros (shape [, dtype, order]) Return a new array of given shape and type, filled with zeros.
- 2. array (object [, dtype, copy, order, lubok, ndim]) Create an array
- 3. as array (a [, dtype, order]) Convert the input to an array.
- 4. As an array (a [, dtype, order]) Convert the input to an ND array, but pass ND array subclasses through.
- 5. Arange ([start,] stop [, step,] [, dtype]) Return evenly spaced values within a given interval.

6. linspace (start, stop [, num, endpoint, ...]) - Return evenly spaced numbers over a specified interval. etc. there many functions which are used to perform specified operation on the given input values

Here is some function that are defined in this NumPy Library.



PANDAS:

Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-touse data structures and data analysis tools for the Python programming language. Python with

Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc. In this tutorial, we will learn the

various features of Python Pandas and how to use them in practice.

The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data.

In 2008, developer Wes McKinney started developing pandas when in need of high performance, flexible tool for analysis of data.

Prior to Pandas, Python was majorly used for data munging and preparation. It had very little

contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin

of data — load, prepare, manipulate, model, and analyse.

Python with Pandas is used in a wide range of fields including academic and commercial

domains including finance, economics, Statistics, analytics, etc.



Key Features of Pandas

- Fast and efficient Data Frame object with default and customized indexing.
- Tools for loading data into in-memory data objects from different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of date sets.
- Label-based slicing, indexing and sub setting of large data sets.
- Columns from a data structure can be deleted or inserted.
- Group by data for aggregation and transformations.
- High performance merging and joining of data.
- Time Series functionality.

• MATPLOTIB

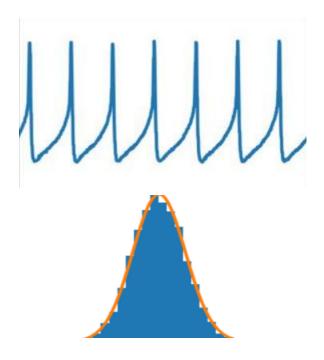


Matplotlib is a Python 2D plotting library which produces publication quality figures in a

variety of hardcopy formats and interactive environments across platforms. Matplotlib can be

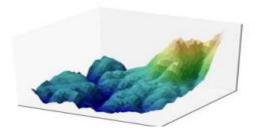
used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application

servers, and four graphical user interface toolkits.



Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object-oriented interface or via a set of functions familiar to MATLAB users.



0

CHAPTER 2 – TECHNOLOGY ADOPTED

2.1 DATA SCIENCE



"Data science" is just about as broad of a term as they come. It may be easiest to describe what it is by listing its more concrete components:

- Data exploration & analysis: -.
 Included here: Pandas; NumPy; SciPy; a helping hand from Python's Standard Library.
- 2) Data visualization: A pretty self-explanatory name. Taking data and turning it into something colourful.
 - Included here: Matplotlib; Seaborn; Data shader; others.
- 3) Classical machine learning: Conceptually, we could define this as any supervised or unsupervised learning task that is not deep learning (see below). Scikit-learn is farand-away the go-to tool for implementing classification, regression, clustering, and dimensionality reduction, while Stats Models is less actively developed but still has a number of useful features.
 - Included here: Scikit-Learn, Stats Models.
- 4) Deep learning: This is a subset of machine learning that is seeing a renaissance, and is commonly implemented with Keras, among other libraries. It has seen monumental improvements over the last ~5 years, such as Alex Net in 2012, which was the first design to incorporate consecutive convolutional layers.
 - Included here: Keras, TensorFlow, and a whole host of others.
- 5) Data storage and big data frameworks: Big data is best defined as data that is either literally too large to reside on a single machine, or can't be processed in the absence of a distributed environment. The Python bindings to Apache technologies play heavily here.
 - Apache Spark; Apache Hadoop; HDFS; Dask; h5py/pytables.
- 6) Odds and ends. Includes subtopics such as natural language processing, and image

manipulation with libraries such as OpenCV.

Included here: nltk; Spacy; OpenCV/cv2; scikit-image; Python

2.2 MACHINE LEARNING



- Machine learning is a subset of artificial intelligence in the field of computer science
 that often uses statistical techniques to give computers the ability to "learn" (i.e.,
 progressively improve performance on a specific task) with data, without being
 explicitly programmed.
- Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field.
- Machine learning (ML) is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed.
 The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

2.2.1 How Machine Learning works?

- Machine learning algorithms are often categorized as supervised or unsupervised. Supervised algorithms require a data scientist or data analyst with machine learning skills to provide both input and desired output, in addition to furnishing feedback about the accuracy of predictions during algorithm training. Data scientists determine which variables, or features, the model should analyse and use to develop predictions. Once training is complete, the algorithm will apply what was learned to new data.
- Unsupervised algorithms do not need to be trained with desired outcome data. Instead, they use an iterative approach called deep learning to review data and arrive at conclusions. Unsupervised learning algorithms -- also called neural networks -- are used

for more complex processing tasks than supervised learning systems, including image recognition, speech-to-text and natural language generation. These neural networks work by combing through millions of examples of training data and automatically identifying often subtle correlations between many variables. Once trained, the algorithm can use its bank of associations to interpret new data. These algorithms have only become feasible in the age of big data, as they require massive amounts of training data.

2.2.2 Advantages of Machine Learning

1. Trends and Patterns Are Identified With Ease

Machine Learning is adept at reviewing large volumes of data and identifying patterns and trends that might not be apparent to a human. For instance, a machine learning program may successfully pinpoint a causal relationship between two events. This makes the technology highly effective at data mining, particularly on a continual, ongoing basis, as would be required for an algorithm.

2. Machine Learning Improves Over Time

Machine Learning technology typically improves efficiency and accuracy over time thanks to the ever-increasing amounts of data that are processed. This gives the algorithm or program more "experience," which can, in turn, be used to make better decisions or predictions.

A great example of this improvement over time involves weather prediction models. Predictions are made by looking at past weather patterns and events; this data is then used to determine what's most likely to occur in a particular scenario. The more data you have in your data set, the greater the accuracy of a given forecast. The same basic concept is also true for algorithms that are used to make decisions or recommendations.

3. Automation

Machine Learning is a key component in technologies such as predictive analytics and artificial intelligence. The automated nature of Data Science means it can save time and money, as developers and analysts are freed up to perform high-level tasks that a computer simply cannot handle.

On the flip side, you have a computer running the show and that's something that is certain to make any developer squirm with discomfort. For now, technology is imperfect. Still, there are workarounds. For instance, if you're employing Data Science technology in order to develop an algorithm, you might program the Data Science interface so it just suggests improvements or changes that must be implemented by a human.

This workaround adds a human gatekeeper to the equation, thereby eliminating the potential for problems that can arise when a computer is in charge. After all, an algorithm update that looks good on paper may not work effectively when it's put practice.

CHAPTER 3 – PROJECT

PROJECT NAME – Gujarat Titans Data Analysis

3.1INTRODUCTION

Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

The Dataset

A Data Set is a container that holds the data you upload to Analytics. Data Sets control how uploaded data gets joined with existing data. You configure Data Sets at the Property level. Data Sets must be associated with at least one View, and can be associated with multiple Views.

Prerequisites

Before starting with this Python project with source code, you should be familiar with the computer vision library of Python that is **OpenCV** and **Pandas**.

OpenCV, Pandas, and numpy are the Python packages that are necessary for this project in Python. To install them, simply run this pip command in your terminal:

1. pip install opency-python numpy pandas

Then we save our file with **.py**(python) extension and a sample image for color detection with **.csv**(A CSV (comma-separated values) file is a text file that has a specific format which allows data to be saved in a table structured format.) extension.

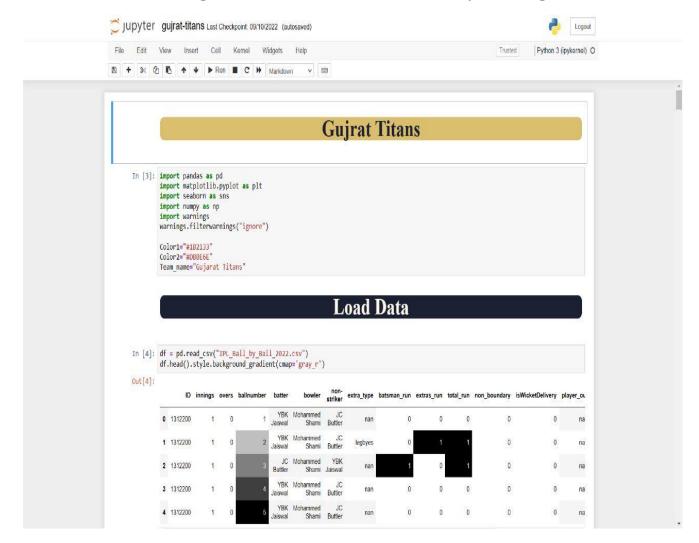
3.2 PROCEDURE

Step1: Import the data & read the CSV file with pandas

We are using argparse library to create an argument parser. We can directly give an data path from the command prompt:

We are using argparse library to create an argument parser. We can directly give an image path from the command prompt.

The pandas library is very useful when we need to perform various operations on data files like CSV. **pd.read_csv()** reads the CSV file and loads it into the pandas DataFrame. We have assigned each column with a name for easy accessing.

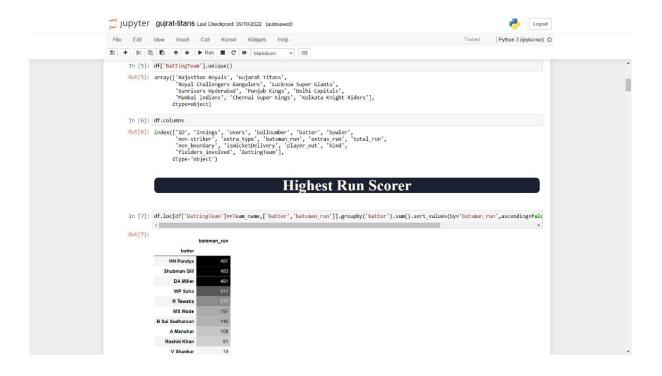


Step2: Now we will analyse the data and find highest run scorer

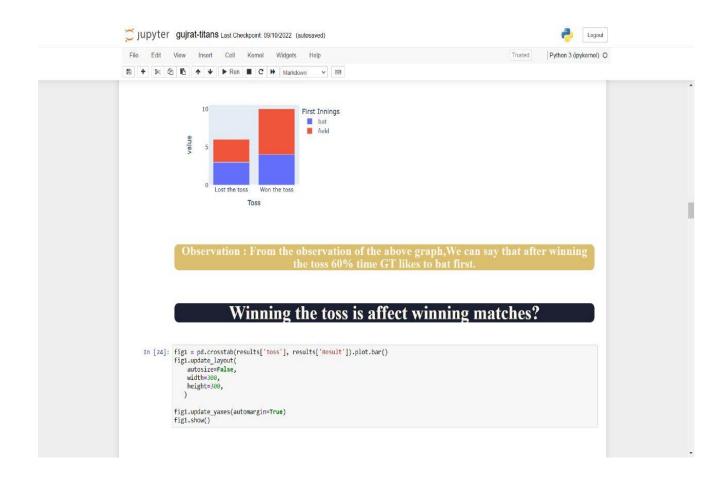
It will calculate the values of the battsman which we double click. Ih will show the brief details of the battsman.

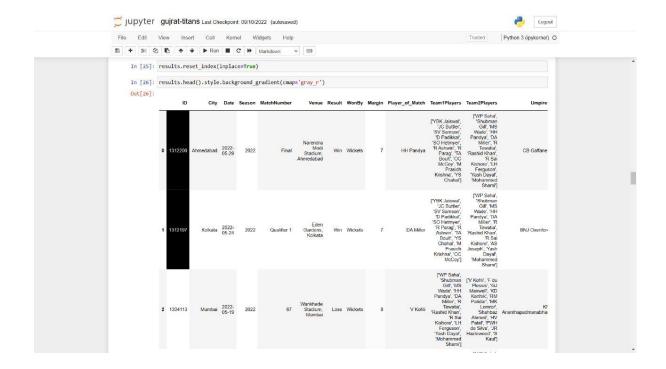
It will show a output like name of the battsman with the scores that every player had.

And the code is also manipulated the outcome very well.

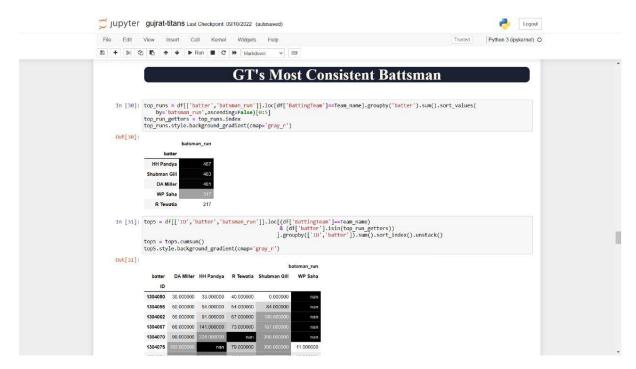


Step3: Calculating the winning toss has any affect on winning matches





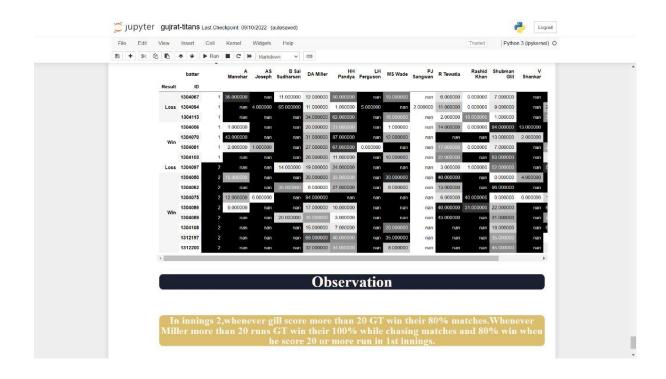
Step4: Gujarat Titans most consistent battsman

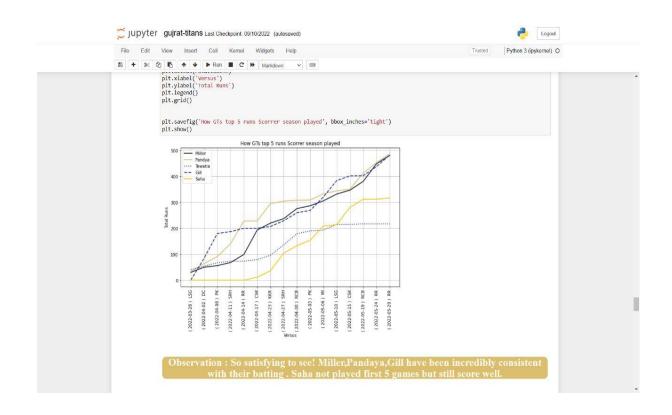


Step5: Run Python File

The beginner Python project is now complete, you can run the Python file. Our desired will be shown as like in result image.

RESULT IMAGES





CHAPTER 4

***** CONCLUSION

In this project we have studied the surveys of the team Gujarat titans and received the very clearful results. We have manipulated the code with outgoing performances of the Gujarat Titans in overall aspects.

This project was manily acquired of the certain desired aspects to give brief overview of the team to other person with minimizing thie hurdles.

❖ FUTURE WORK

In the future this project leaves a great impact to users by having deep and accurate data overview of any team,Battsman,Bowler,etc .This project can be made more interesting by using the real time APIs.It would be more convenient to fetch real time data.

.

REFRENCES

1.Github.com2.kaggel.com