udemy

Project Report

On

“Result Analyzer”

Based on Python and Data Science

Udemy, Inc.

(“Python for Data Science and Machine Learning”)

Submitted By:

Mohit Bansal (05414802715)

B. Tech (Computer Science and Engineering)

Maharaja Agrasen Institute of Technology

Rohini, Delhi

Acknowledgement

It is a great pleasure for me to submit this report entitled “Result Analyzer” as my Summer Training Project for the partial fulfillment of under graduation.

I would like to present my utmost gratitude to Udemy, an online MOOC organization which made the power of knowledge available to all at a very cheap price or even for free. The availability of these world class resources on Python, which is one of the most popular and useful language these days along with concepts of Data Science. They provided me the golden opportunity to learn something I love and help me to kickstart in this field.

I would also like to thank my mentor **Mr. Jose Portilla**, a data scientist and mentor at Udemy’s course **“Python for Data Science and Machine Learning”** (Jose Portilla, n.d.) who taught me the skills of data Science and python which was the core part for my project.

Along with them, the community members of python must never be leaved. They deserve the thanks of all those using Python today, no matter for what. The descriptive documentation of Python and its libraries helped me to overcome the difficulties I faced during the development of the project.

There are various people whom I got to know during the development of this project, this vacation and it is almost impossible to thanks all of them. But few people cannot be bargained for some thanks, like Kushal Das and his community.

At last the support of my family to help me pursue what I want is invaluable. And finally, thanks to all of them who directly and indirectly motivated me to make a great project, naming few of them, Harshit Bisht, Divyank, Kishan, Sai, Pradhvan, Sayan and so on.

Thanking You
Mohit Bansal

Certificate

Declaration

I, Mohit Bansal, a third-year Undergraduate student enrolled at Maharaja Agrasen Institute of Technology, hereby declare that the project entitled “Result Analyzer” submitted as my “Summer Training Project” is an original work done by me under the guidance of my mentor “Jose Portilla”.

The information and the data given in the report is authentic to best of my knowledge and I take full responsibility of the consequences if any of the details provided deemed to be false.

Mohit Bansal Signature:

(05414802715) Date:

MAIT, Rohini, Delhi

Preface

This is the project report developed by “Mohit Bansal”, pre-final year student of Bachelor of Technology in Computer Science Engineering at Maharaja Agrasen Institute of Technology, Delhi while pursuing online Certificate Course on “Python for Data Science and Machine Learning” from a famous MOOC portal, Udemy.

Result Analyzer is a software which utilized the concept of data Science and machine learning made using python. This software named “Result Analyzer” takes the result data of students of an institute and then analyze various factors such correlation, mean and can be used to visualize their performance using different types of graph.

The aim of the project is to make the analysis of the result of students of different disciplines easily and effectively so that, teachers as well as students can improve their performance in the field they were lacking. This would help them to target their weaknesses.

Features:

1. Check the result
2. Analyze your result
3. Visualize your result
4. Compare results
5. Visualize the trends according to various factors

Organization Profile

Udemy.com is an online learning platform. It is aimed at professional adults. Unlike academic MOOC programs driven by traditional collegiate coursework, Udemy provides a platform for experts of any kind to create courses which can be offered to the public, either at no charge or for a tuition fee. Udemy provides tools which enable users to create a course, promote it and earn money from student tuition charges.

No Udemy courses are currently credentialed for college credit; students take courses largely as a means of improving job-related skills. Some courses generate credit toward technical certification. Udemy has made a special effort to attract corporate trainers seeking to create coursework for employees of their company.

Udemy, an online MOOC organization which made the power of knowledge available to all at a very cheap price or even for free.

CEO:	Dennis Yang
Founders:	Gagan Biyani
	Oktay Caglar
	Eren Bali
Founded:	2009

Figure 1 Udemy Logo

Abbreviations Used

PYM	Python you and me
MAIT	Maharaja Agrasen Institute of Technology
OO	Object Oriented
PDC	Project Development Cycle
GPL	General Public License
GNU	GNU Not Unix

List of Figures

List of Tables

Content

Thanks

Abstract

Data Science is an emerging field these days, and is very helpful to solve the dilemma of big data. The main intention of introducing data science is to reduce the manual work of handling this large chunk of data (over 1000ZB data). Data Science helps in filtering the useful information out of these raw data and analyze the various features/factors more statistically.

This Software can be used to help those students who want to prepare for exams by analyzing their performance in previous papers. Other uses and scope of this software are as follows:

Uses:

1. Check the result
2. Analyze your result
3. Visualize your result
4. Compare results
5. Visualize the trends according to various factors

Scope:

1. In Educational institutes to help students improve their performance
2. In Survey, to see the trend of education among students.

Methodology:

General Methodology of any software development project is being followed during the development of the project "Result Analyzer":

1. Requirement Phase: Features required by the users to be integrated within the software are elicited.
2. Design Phase: The Structure of software and its modules is decided.
3. Development Phase: The project is developed as whole.
4. Coding Phase: The product is started to be coded and distributed.
5. Testing Phase: The product before distribution is tested whether all the requirements are being satisfied.

As you may have observed, coding skills only take 20% of the work of whole software product.

1. Introduction

1.1 Overview of Data Science

The project "Result Analyzer" utilizes the concepts of Data Science and Machine Learning, and uses the power of highly popular and powerful language Python3.

Data Science is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured. It is a concept to unify statistics, data analysis and their related methods in order to understand and analyze actual phenomena with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information Science, and various other fields. (Wikipedia)

Data Scientist is someone who use their data and analytically able to find and interpret rich data sources, manage large amounts of data despite hardware, software, bandwidth constraints, merge data sources, ensure consistency of datasets, create visualizations to aid in understanding data.

Data Science is not only about technology and math, effective data scientists require a combination of technical skills and soft skills to turn data into actionable insight.

It is often said "Data Scientist is a Unicorn that never exists", that means the diverse range of skills required by a Data Scientist is almost impossible to learn and hence it never exists but a team of experts of diverse skills, as a unit can be considered as Data Scientist.

1.2 Need for Data Science: Big Data

In this era, where computers promise High computational power to the world and it leads to a very big problem of Big Data. This in effect lead us to the dilemma, how to manage, store, analyze these large chunks of data efficiently. Here comes Data Science to the rescue. Using data science, companies have become intelligent enough to push & sell products as per customers purchasing power & interest. Here's how they are ruling our hearts and minds.

Data Science can be thought as the basis for empirical research, for data is used to inform our hypothesis and provide observations. In many cases, this data is used either by businesses or by scientists to inform their understanding of a phenomenon. Because there are often large troves of data which we can mine for insights, we often call this big data. Insight is a term we use to refer to the data product of data science. It is extracted from a diverse amount of data through a combination of exploratory data analysis and modelling. The question we ask are sometimes quite specific, but sometimes it takes looking at the data and patterns in it to come up with a specific question. Another important point to recognize is that data science is not a static, one-time analysis. It involves a process where the models we generate lead to insights and those insights are then improved by gathering further empirical evidence, or simply, data. For example, a book retailer like amazon.com can constantly improve the model of a customer's book preferences using the customer demographics, his or her previous purchases and prior book reviews by the customer. Their models also likely consider the similarity of customers to detect common interests. The book retailer can also use this information to predict which customers are likely to like a new book and act to market the book to those customers. This is where we see insights being turned into action. A we have seen in the book marketing example, using data science and analysis of the past and current information, data science generates actions. This is not just an analysis of the past, but rather generation of an actionable information for the future.

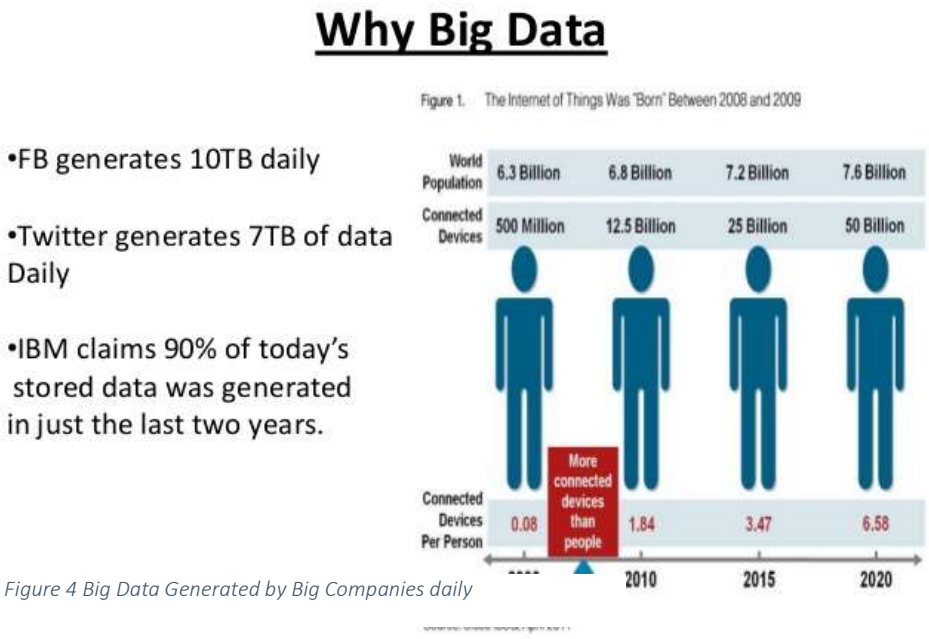

Figure 4 Big Data Generated by Big Companies daily

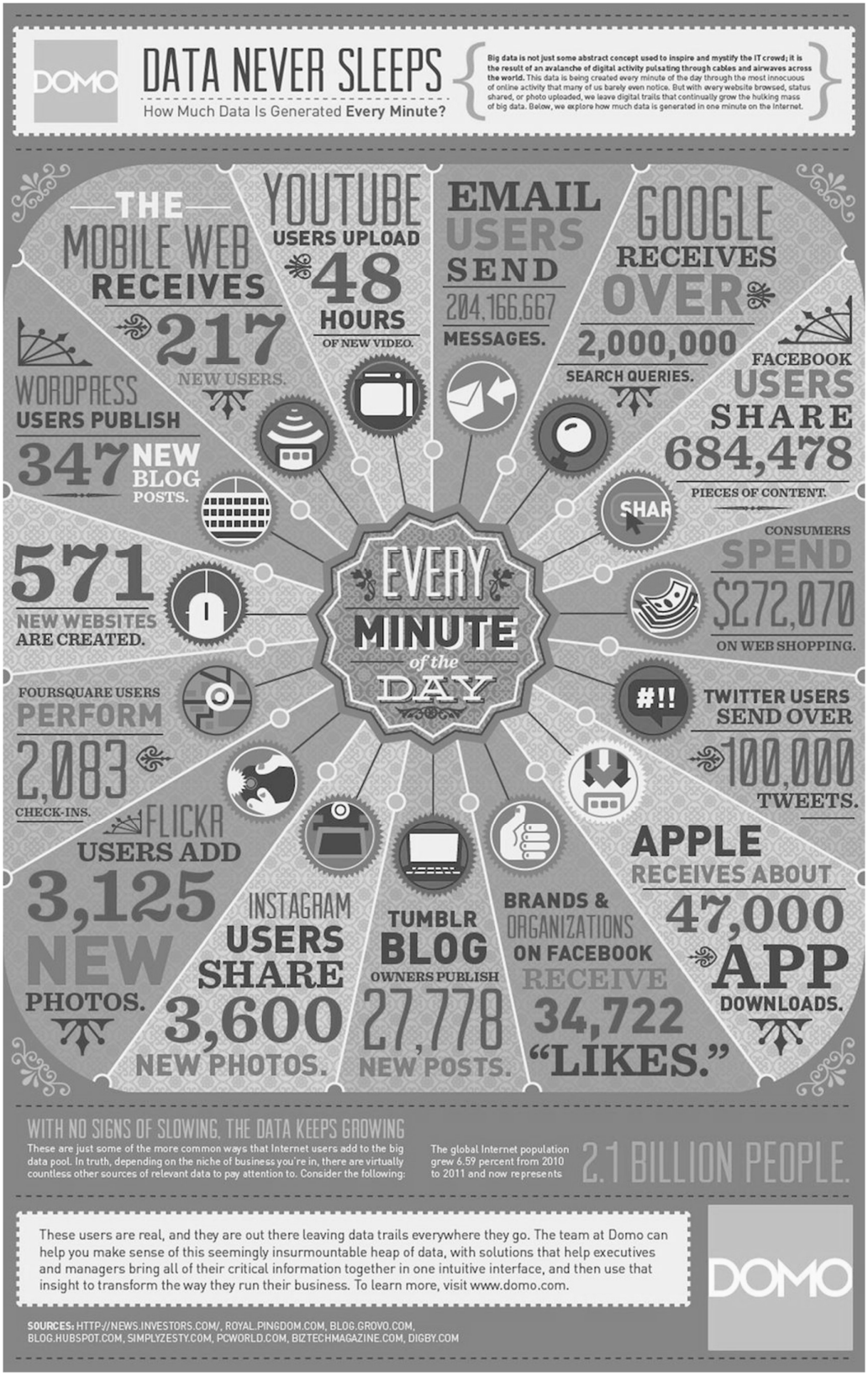

Figure 5 Big Data

1.3 Applications of Data Science

Data Science has vivid number of applications to manage and to visualize these large chunks of data. (analyticsvidhya, 2015)

1. **Internet Search**: When we speak of search, we think 'Google'. Right? But there are many other search engines like Yahoo, Bing, Ask, AOL, Duckduckgo etc. All these search engines (including Google) make use of data science algorithms to deliver the best result for our searched query in fraction of seconds. Since, Google processes more than 20 petabytes of data every day. Had there been no data science, Google wouldn't have been the 'Google' we know today.

2. **Targeted Advertising and re-targeting**: If you thought Search would have been the biggest application of data science and machine learning, here is a challenger – the entire digital marketing spectrum. Starting from the display banners on various websites to the digital bill boards at the airports – almost all of them are decided by using data science algorithms.

   This is the reason why digital ads have been able to get a lot higher CTR than traditional advertisements. They can be targeted based on user's past behavior. This is the reason why I see ads of analytics trainings while my friend sees ad of apparels in the same place at the same time.

Figure 6 Targeted Marketing by google

3. **Recommender Systems**: Who can forget the suggestions about similar products on Amazon? They not only help you find relevant products from billions of products available with them, but also adds a lot to the user experience.

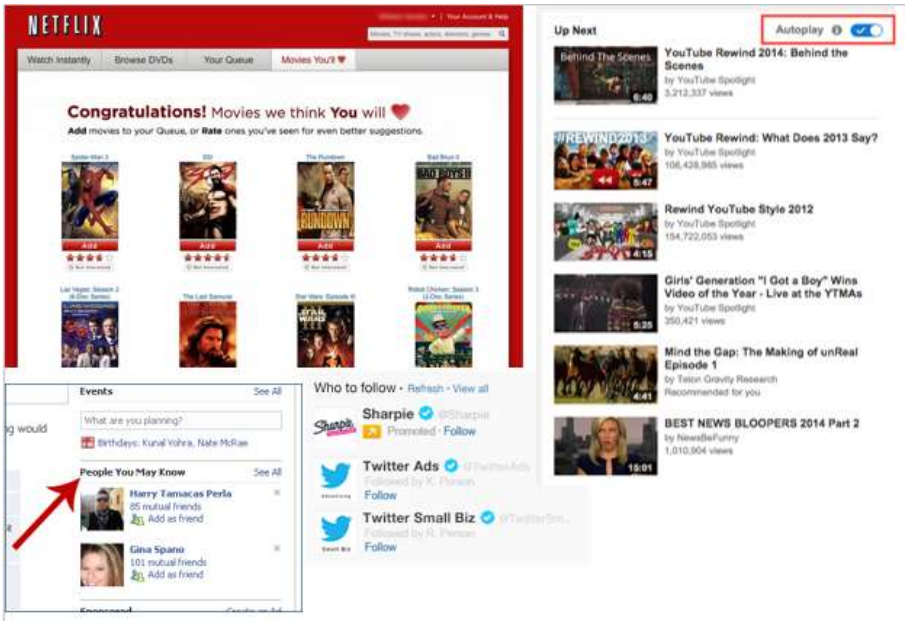

Figure 7 Recommender System of Netflix

4. **Image Recognition**: You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm. Similarly, while using WhatsApp web, you scan a barcode in your web browser using your mobile phone. In addition, Google provides you the option to search for images by uploading them. It uses image recognition and provides related search results. To know more about image recognition, check out this amazing (1:31) mins video:

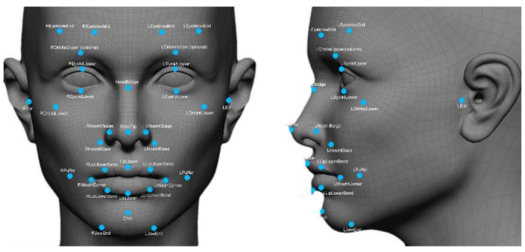

Figure 8 Face Recognition used in iPhone 8

5. **Speech recognition**: Some of the best example of speech recognition products are Google Voice, Siri, Cortana etc. Using speech recognition feature, even if you can't type a message, your life wouldn't stop. Simply speak out the message and it will be converted to text. However, at times, you would realize, speech recognition doesn't perform accurately. Just for laugh, check out this hilarious video (1:30 mins) and the conversation between Cortana & Satya Nadella (CEO, Microsoft)

Figure 9 Schema of Speech Recognition

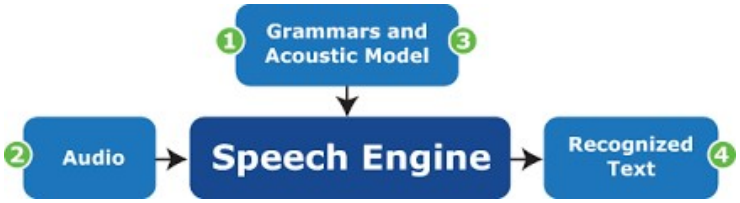

Figure 10 Process of Audio to Text Transcription

6. **Data Analysis and Data Visualization**: Data Analysis and its visualization is a sub domain of Data Science and is extensively used to analyses big data generated by a corporation to improve their productivity.
   The applications of Data Science are infinite and will never end.

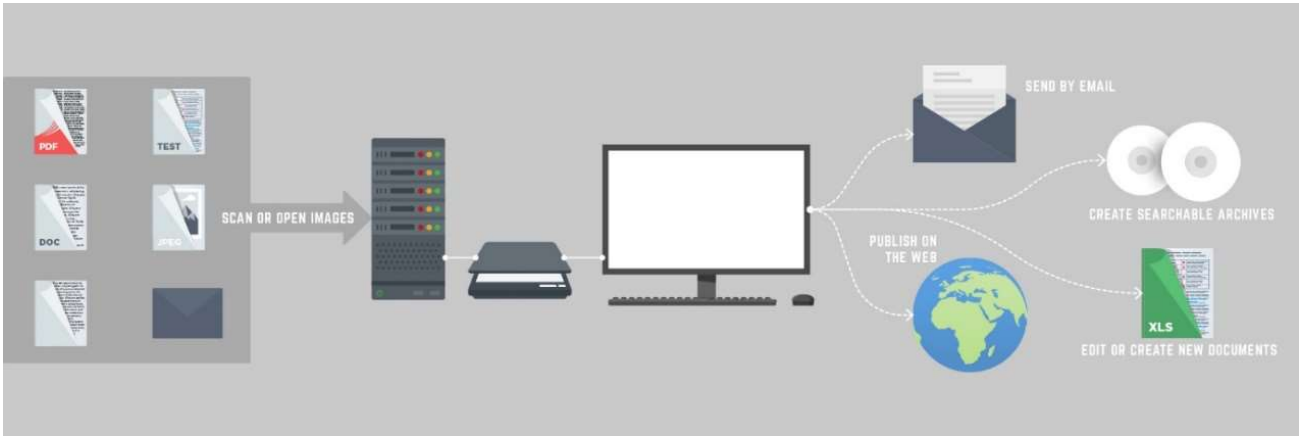

Figure 11 Data Analysis

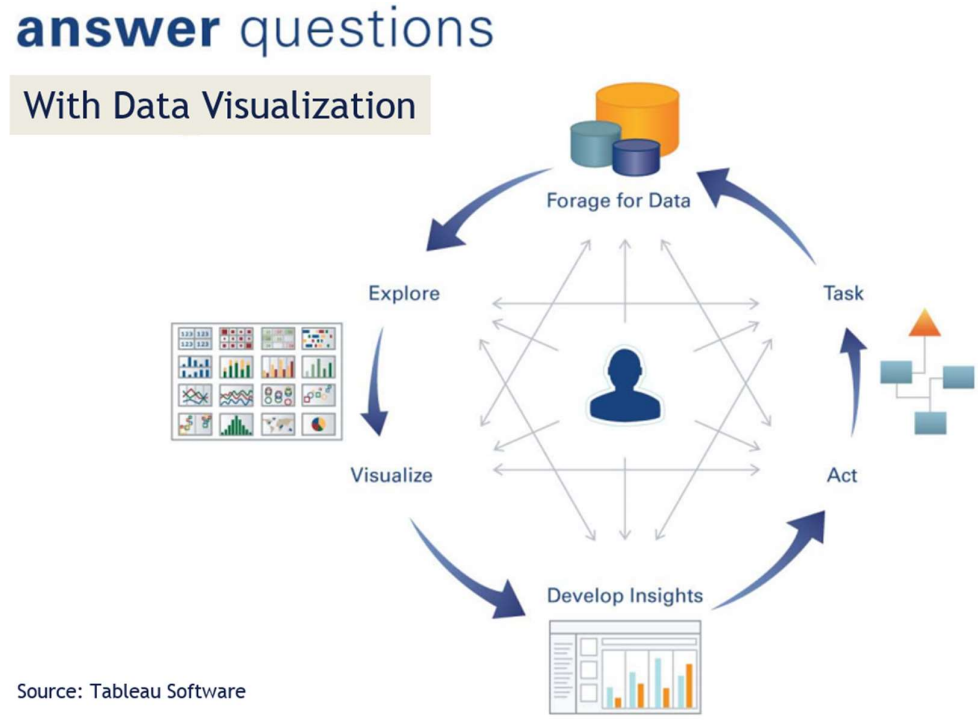

Figure 12 Process of Data Visualization

2. Technologies Used

2.1 Python 3.6

Figure 13 Python Logo

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL). (Tutorialspoint) An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such as C++ or Java. The language provides constructs intended to enable writing clear programs on both a small and large scale.

Python features a dynamic type system and automatic memory management and supports multiple programming paradigms, including object-oriented, imperative, functional programming, and procedural styles. It has a large and comprehensive standard library.

Python interpreters are available for many operating systems, allowing Python code to run on a wide variety of systems. (Wikipedia)

Why Python for Project?

There is not any specific reason to why Python. Like every other programming language, each language offers some advantages as well disadvantages.

1. Python is among those languages, which offers advantage to data science and machine learning.

2. Python has very large number libraries to get started.
3. Python is an open language with a vibrant community to help you.
4. Above of all the reasons, Python is very easy to learn if you are familiar with basic programming knowledge.
5. Python is interpreted language, which is beneficial for machine learning and data science programs.

Features of Python

- **High Level Language**: Such languages cannot be directly being understood by computer machines. Hence programs written in these languages are first converted into assembly and later into machine-level languages (which are lower level languages) for the computer to understand what exactly to do.

- **Object Oriented (OO) Language**: Object-oriented languages are those in which it is possible to imitate real-life objects as programming objects for more efficient code. A programmer can define objects in programs, their properties, how they relate to each other and the logic regarding how to manipulate them.

- **Simple Syntax Formats**: Python has very easy syntax for writing codes with very little formatting and therefore it is easy to learn. Hence it is one of the most preferred entry-level language.

- **Interpreted Language:** Python is an interactive language and codes are usually shorter in length. Shorter codes are generally easier to debug and understand.

Figure 14 Working of an Interpreted Language like Python

2.2 Data Analysis

1. Numpy

Numpy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. (Numpy)

Figure 15 Numpy Logo

2. Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. (Wikipedia)

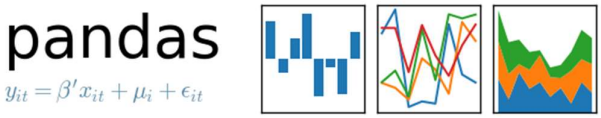

Figure 16 Pandas Logo

2.3 Data Visualization

1. Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension Numpy. (wikipedia)

Figure 17 Matplotlib Logo

2. Project Details (Result Analyzer)

Result Analyzer is a utility tool to analyze and visualize the result data of students to help students to improve their performance by targeting their weaknesses.

Like every Data Analysis project, "Result Analyzer" started with:

1. **Data Collection**

   Sample Data of student results are elicited from the result portal for various field as in computer Science, Electrical Engineering, Information Technology etc.

2. **Describing Data Set**

   The gathered raw data is processed to convert them into useful information i.e. unnecessary fields and attributes were removed to focus on the specific factors we want to work with like, Photos of the students is not required, so is trimmed of the raw data.

3. **Statistical Analyses of Dataset**

   Once the data is filtered and processed into useful information, required analysis is performed on that useful data such as:

   - Performance over the year of a student
   - Comparison between performance of two students, and so on.

4. **Interpretation of Analyzed Data**

   After the analysis, the data we get is used to visualize the trend to better get the understanding of their performance which saves our time as well because the need to look for each entry is eliminated.

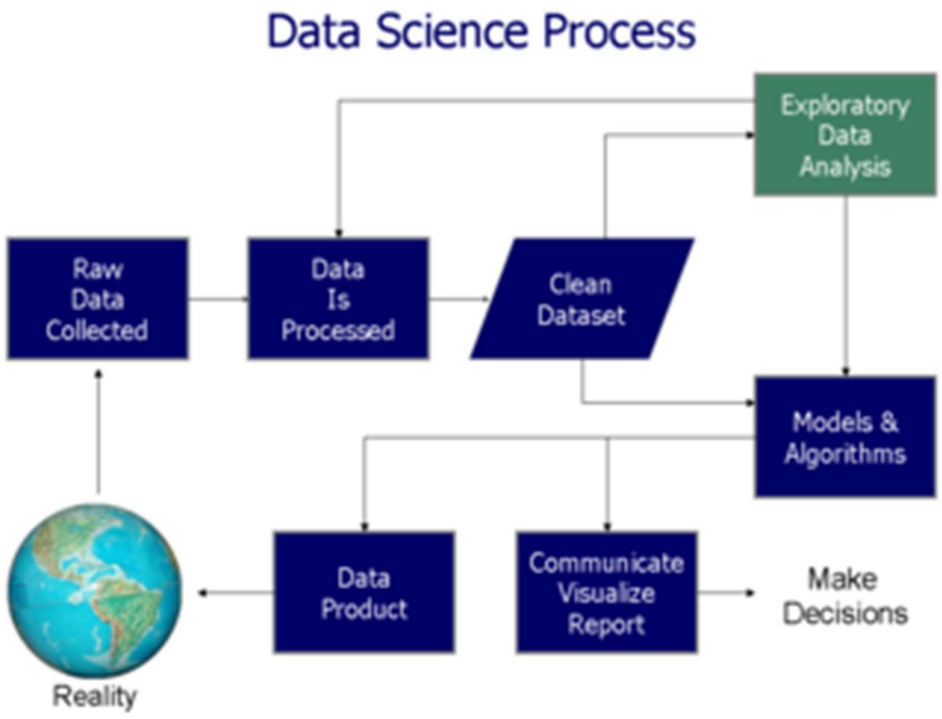

Figure 18 Data Science Process

5. Screenshots

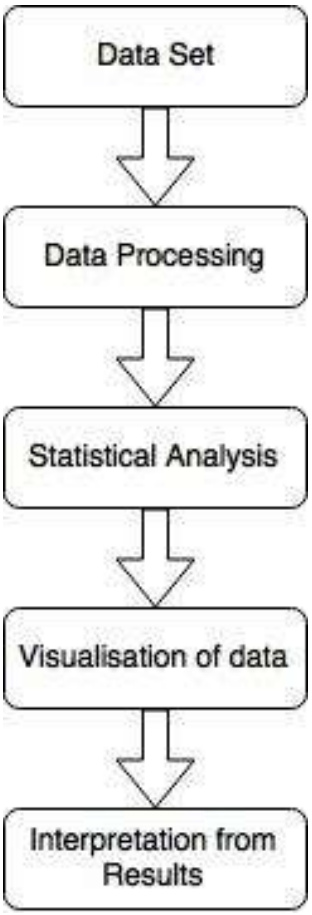

Figure 19 Project Development Cycle

6. Project Tables

Table 1 Result of Students

Enrollment No.	Subject1	Subject2	Subject3	Subject4	Subject5
00114802715	89	75	68	79	82
02114802715	56	89	71	52	50

Table 2 Course Details

Course Id	Subject1	Subject2	Subject3	Subject4	Subject5
148	FCS	Math	STLD	DS	OOP
147	Signal Systems	Wireless Systems	STLD	Communication System	Math

Table 3 Student Details

S no.	Enrollment No.	Name	Institute Name	Course Id
1.	05451802715	Mukul	MAIT	518
2.	02114802715	Shiva	MAIT	148

(Note: The data and tables presented above are not real and tend to change)

7. Conclusion

This is the era of high computation and high-speed Internet where every device produces massive of data every second of every minute, and taking into the effect of tens of devices used by billions of people produces a large amount of data. To solve the problem of high computation, lead us to another problem better known as "Big Data". From the images illustrated in this report, it can be concluded that to manage this large amount of data is rather difficult but is very necessary.

Data Science is an emerging field of computer science to help us solve this problem of big data by statistically filtering out the irrelevant data and analyze useful information to determine the useful factors and features out of that big data which is like ***"Finding Needle in haystack".***

Data Science is rather a tough field to get a grasp on and a data scientist is needed to learn all the relevant disciples to analyze data which is almost impossible and so it is often said:

"Data Scientist is an Unicorn"

Result Analyzer is one of the attempt to handle this big data for student result which can help students to improve their performance by targeting their weaknesses.

Further Developments

- Since the current project is in development phase, and thus still in console/terminal based utility tool. To add the graphical Feature to help non-technical users.
- Can be Integrated with an existing webapp or android app.
- It can be used to analyze result in colleges to help students improve their performance.

Limitations

- The project is still in console/terminal phase
- To analyze the data, the structure of the dataset is necessary to know, to perform analytical operations
- To incorporate different datasets for different organization, program needs to be changed a little.

Reference/Bibliography

References

- analyticsvidhya. (2015, 9 21). *Blog.* Retrieved from AnalyticsVidhya: https://www.analyticsvidhya.com/blog/2015/09/applications-data-science/
- Jose Portilla. (n.d.). *Udemy*. Retrieved from https://www.udemy.com/python-for-data-science-and-machine-learning-bootcamp/
- Numpy. (n.d.). *Wikipedia.* Retrieved from https://en.wikipedia.org/wiki/NumPy
- Tutorialspoint. (n.d.). *Tutorialspoint.* Retrieved from https://www.tutorialspoint.com/python/
- Wikipedia. (n.d.). *en.wikipedia.org/wiki/Data_science.* Retrieved from Wikipedia.org: https://en.wikipedia.org/wiki/Data_science
- wikipedia. (n.d.). *Wikipedia.* Retrieved from https://en.wikipedia.org/wiki/Matplotlib
- Wikipedia. (n.d.). *Wikipedia.* Retrieved from https://en.wikipedia.org/wiki/Pandas_(software)
- Wikipedia. (n.d.). *WIkipedia.* Retrieved from https://en.wikipedia.org/wiki/Python_(programming_language)

Bibliography

- ✓ Kushal das. *document.* Retrieved from http://pymbook.readthedocs.io/en/latest/
- ✓ DataCamp, course. Retrieved from https://courses.edx.org/courses/course-v1:Microsoft+DAT208x+2T2017/course
- ✓ Stackoverflow. Retrieved from https://stackoverflow.com/
- ✓ Numpy. Retrieved from http://www.numpy.org/
- ✓ Pandas. Retrieved from https:// http://pandas.pydata.org/
- ✓ Matplotlib. Retrieved from https://matplotlib.org/
- ✓ Python. Retrieved from https://docs.python.org/3/
- ✓ Emht: Retrieved from https://github.com/emht/ResultAnalyzer

Thank You for Time!