

**SIX WEEK SUMMER TRAINING REPORT**

on

**DATA SCIENCE & MACHINE LEARNING**

(June – July 2022)

Submitted by

**RANJITH SK**

**Registration Number: 12007594**

Under the Guidance of

**PRANAY SHARMA**

(Instructor for the course/project at IBM(All Soft Solutions))

**School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

**DECLARATION**

I hereby declare that I have completed my six weeks summer training at  **All Soft Solutions(IBM), Punjab** through online classesfrom **June 6** to **July 8** under the guidance of **Pranay Sharma.** I declare that I have worked with full dedication during this training and the completion of this course/project partially fulfil the requirements for the award of degree of **B.Tech Computer Science Engineering (Hons)**.

*Specialization****:* Data Science**

**RANJITH SK**

Registration Number: **12007594**

Date:15/07/2022

Expert from IBM/All Soft Solutions(Teacher): **Pranay Sharma**

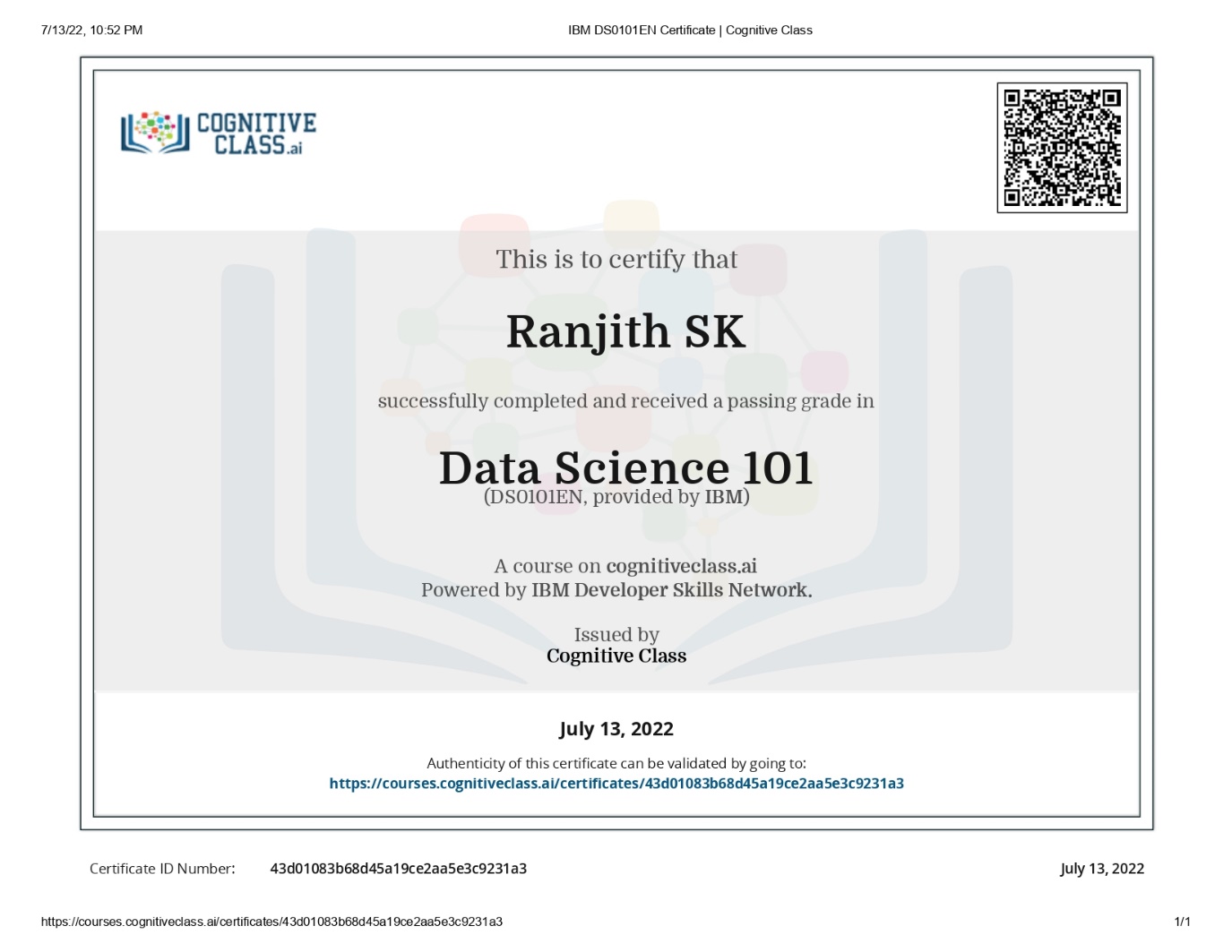
**ACKNOWLEDGEMENT**

I would like to express my respect and gratitude to the teachers at **Lovely Professional University** and the training coordinators for providing this opportunity to do this program and project regarding **Data Science & Machine Learning.**

I would also like to extend my gratitude to **Pranay Sharma**(Teacher from IBM/All Soft Solutions) who helped us by giving guidance and support all through the program and also in completing the project.

The success and final outcome of this project required a lot of guidance and assistance from my instructor and my friends, and I am extremely fortunate to have got this along the completion of my project work. Whatever I have done is only due to such guidance and assistance and I would not forget to thank them.

**SUMMER TRAINING CERTIFICATE**



Certificate issued by **Cognitive class** after completion of the course and test on **DATA SCIENCE 101 (IBM)**



Certificate issued by **All Soft Solutions(IBM)** after the completion of the course and project **WhatsApp Chat and Chat Sentiment Analysis**

|  |  |  |
| --- | --- | --- |
| **S.No** | **TABLE OF CONTENTS** | **PAGE** |
| **1.** | Introduction | 1 |
| **2.** | Technology Learnt | 2 |
| **3.** | Project Objective | 4 |
| **4.** | Idea of the Project | 4 |
| **5.** | Project Architecture | 4 |
| **6.** | System Implementation | 6 |
| **7.** | Advantages of this Project | 8 |
| **8.** | Project Implementation | 9 |
| **9.** | Reason for choosing this Project | 17 |
| **10.** | Reason for choosing this Course | 17 |
| **11.** | Learning Outcome | 18 |

**INTRODUCTION**

Data Science is the art of uncovering the insights and trends that are hiding behind data. Data Science is the journey toward achieving three significant kinds of results: discovery, insights, and innovation through data. As such I have interest in the field and chosen Data Science course by IBM-All soft Solutions, where I have been taught in detail about Data Science and lots more.

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization’s data. It has incalculable benefits in business, research, and our everyday lives. Your route to work, your most recent search engine query for the nearest coffee shop, your Instagram post about what you ate, and even the health data from your fitness tracker are all important to different data scientists in different ways. Sifting through massive data lakes, and looking for connections and patterns, data science is responsible for bringing us new products, delivering breakthroughs, insights and making our lives more convenient. These insights can be used to guide decision-making and strategic planning.

The project I have completed is based on the **Analysis of WhatsApp chat,** I have tried my best to make the complicated process of chat analysis as simple as possible using the necessary requirements. I have tried to design the project in such a way that user may not have any difficulty in using this package & further expansion is possible without much effort. Even though I cannot claim that this work to be entirely exhaustive, the main purpose of our work is Chat analysis which is the investigation and modelling of a specific WhatsApp chat’s thorough worktime and analytics, so that we can save our time and manage our worktime and activities.

**TECHNOLOGY LEARNT**

**Data Science**

**Module 1 - Defining Data Science**

* + What is data science?
  + There are many paths to data science
  + Any advice for a new data scientist?
  + What is the cloud?

**Module 2 - What do data science people do?**

* + A day in the life of a data science person
  + Data science tools and technology

**Module 3 - Data Science in Business**

* + How should companies get started in data science?
  + Recruiting for data science

**Module 4 - Use Cases for Data Science**

* + Applications for data science

**Module 5 - Fun with Data Science**

* + Things data science people say

**Python for Data Science**

**Module 1 - Python Basics**

* Types
* Expressions and Variables
* String Operations

**Module 2 - Python Data Structures**

* Lists and Tuples
* Sets
* Dictionaries

**Module 3 - Python Programming Fundamentals**

* Conditions and Branching
* Loops
* Functions
* Objects and Classes

**Module 4 - Working with Data in Python**

* Reading files with open
* Writing files with open
* Loading data with Pandas
* Working with and Saving data with Pandas

**Module 5 - Working with NumPy Arrays**

* NumPy 1D Arrays
* NumPy 2D Arrays

**PROJECT OBJECTIVE**

In this decade the upcoming technologies are mainly dependent on **data**. This data can only be obtained if there is some research applied on the context of the requirements of the project/tool. Since a lot of machine learning enthusiasts develop models which helps solve multiple problems the requirements of appropriate data are very large scale this project aims to provide a better understanding towards various types of chats. This analysis proves to be better input to machine learning models which essentially explore the chat data. These models require proper learning instances which provides better accuracy for these models. Our project ensures to provide an in-depth exploratory data analysis on various types of WhatsApp chats.

**IDEA OF THE PROJECT**

WhatsApp chat analysis is a **statistical analysis** project for WhatsApp chats. Working on the chat files that can be exported from WhatsApp it generates various plots showing, for example, which another participant a user responds to the most. We propose to employ dataset manipulation techniques to have a better understanding of WhatsApp chat present in our phones.

**PROJECT ARCHITECTURE**

The project architecture shows the general architecture of this project. This includes, data collection stage, the data input state, data transformation, data exploration and data visualization. The data collection stage is shown further down while the data input, data transformation, data exploration and data visualization are handled by Python and its libraries.

Diagram

Description automatically generated with medium confidence

* **Data Collection Stage**

This stage involves the point where the WhatsApp data is collected. This was done by visiting the chat group to be analyzed, to export the WhatsApp data file that was used. The procedures involved, visiting the WhatsApp group page, clicking on the settings, select export data and then select either add media or without media. This simply means to know whether you intend to export the file with the media or not. Note that exporting with the media, will lead to use of larger volume of data and waste of time for data collection. The below figure will be a brief illustration of the steps involved in the data collection.

Diagram

Description automatically generated

**SYSTEM IMPLEMENTATION**

**Python:** It is an interpreted, high-level general-purpose programming language. Created by Guido Van Rossum and first released in 1991. Its language constructs and objects-oriented approach aim to help programmer with clear, logical code for small and large-scale tools. Python is used for web development (server-side), software development, mathematics, it can be used alongside software to create workflows, it can connect to database systems, it can also read and modify files, it can be used to handle big data and perform complex mathematics and can be used for rapid prototyping, or for production-ready software development.

**Pandas:** This is an open-source Python libraries which is mainly used in Data Science and machine learning subjects. This library provides analysis tool for data manipulation, using its data structures this are used for analyzing data for manipulating time series analysis and numerical data.

**NumPy:** NumPy can be name come from Numeric Python, it is a data analysis library for Python that contains various numerical functions and methods for numerical analysis and having multi-dimensional array objects and to process these arrays contains collection of routines.

**Matplotlib**: Matplotlib is easy to use and an amazing visualizing library in Python. It is built on NumPy arrays, and it work with the broader SciPy stack and consists of several plots like pie, line, bar, graph, scatter, histogram, etc. In this project, Matplotlib is used for various visualizations for analysis of WhatsApp chats. Visualizations like bar charts, line charts, pie charts are used.

**Seaborn**: Seaborn is a library mostly used for statistical plotting in Python. To make statistical plots more attractive it provides beautiful color palettes and default styles. In this project, Seaborn is used for heatmap visualization for showing 24 hours with 7 days with different scale of color for getting hour with max to min messages.

**NLP:** In this project, Features of NLP are used like parsing text, eliminating stop words and analyzing text. Parsing text is used for splitting messages into words for analysis like total words and mostly used words. A file is used that contains all stop words which is given to the python program to show meaningful words only by eliminating all stop words. Text analysis is used to identify how many media are shared; how many links are shared.

**ADVANTAGES OF WHATSAPP CHAT ANALYSIS PROJECT**

• Works on all devices.

• Shows data based on WhatsApp chat file.

• Shows different visualizations.

• Total Messages.

• Media shared.

• Link shared.

• Monthly timeline.

• Most busy day.

• Most busy month.

• Weekly activity.

• Most busy users.

**IMPLEMENTATION**

**WHATSAPP GROUP CHAT ANALYSIS**

**### Importing libraries**

Graphical user interface, text, application

Description automatically generated

**### Extracting date from chat file**

Logo, company name

Description automatically generated

**### Regex pattern to extract username of Author**

A picture containing shape

Description automatically generated

**### Extracting Date, Time, Author, and message from the chat file**

Text

Description automatically generated

**### Finally creating a data frame and storing all data inside that data frame**

A picture containing graphical user interface

Description automatically generated

Text

Description automatically generated

**### Checking head part of Dataset**

Graphical user interface, application

Description automatically generated

**### Checking shape of Dataset**

Graphical user interface, text, application

Description automatically generated

**### Checking information of Dataset**

Text

Description automatically generated

**### Dropping Nan values**

Graphical user interface, text, application

Description automatically generated

**### Checking no. of authors of group**

Text

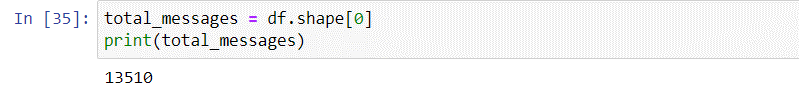
Description automatically generated with low confidence

**### Checking authors of group**

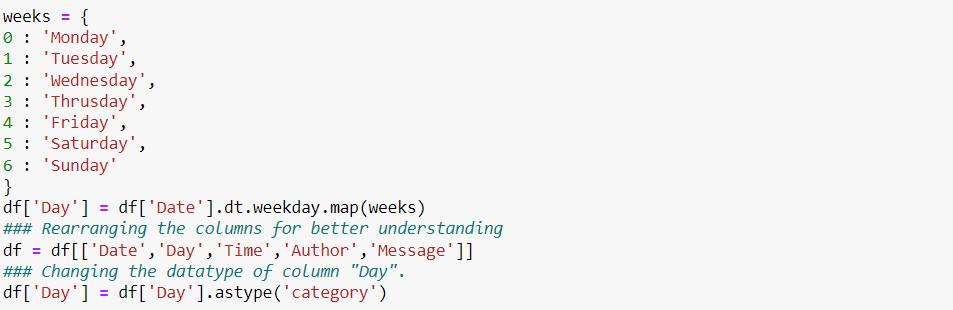
Graphical user interface, text, application

Description automatically generated

**### Total Number of Messages**



**### Adding one more column of “Day” for better analysis, here we use datetime library which help us to do this task easily**



**### Updated Dataset**

Graphical user interface, application

Description automatically generated

**### Function to count number of links in dataset, it will add extra column and store information in it.**

Graphical user interface, application

Description automatically generated

**### Group Chatting Stats**

Graphical user interface, text, application

Description automatically generated

**### Creates a list of unique Authors**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**### Most Active Author in the Group**

Text

Description automatically generated

Chart, bar chart

Description automatically generated

**### Most Active day in the Group**

Text

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

**REASON FOR CHOOSING THIS PROJECT**

* I chose ‘WhatsApp Chat Analysis’ as my project because, in our day-to-day life we greatly indulge in the usage of social media such as WhatsApp, Instagram, Twitter and so.
* As such by choosing this as my topic while I complete my project with regards to the course program, I wanted to learn how WhatsApp affect our daily life.
* By learning the time period spent(chat analysis) in WhatsApp using the extracted data, we can greatly adjust our time management with other respective work and activities in our daily.
* Hence, this project helps us both in developing our knowledge in Data Science and Python but also helps us in our daily life with regards to time management.

**REASON FOR CHOOSING THIS COURSE**

* According to *The Economist*, data has supplanted oil as the most valuable resource in the world. This could easily rank as the number one reason to pursue a career or education in data science.
* The core of the data science field, centres around precise and often minutia-driven analysis, building strong decision capabilities, and can at times be a quiet and solitary field.
* There is data science in business, accounting, education, science, engineering, healthcare, technology, energy sector, government, and so on. So having the ability to work with data is an essential skill.
* Whether it is the ubiquity of this technology, its fast-evolving nature, its applications in solving real-world problems, its scope to integrate previous knowledge with new ideas, its capacity to tame unstructured data, or build predictive capabilities.
* This course has given me an opportunity to expand my knowledge about data science and it was worth spending my summer holidays by acquiring more knowledge through this course.

**LEARNING OUTCOME FROM THE TRAINING**

* In this summer training, I completed a course on “**Data Science**”. The reason I chose this course is because I have interest in the field of Data Science. My specialization in the coming years is Data Science. So, to have completed this introductory course is a good start. Though I have prior experience in Python it was clearer to revise them with the basics and learn other intermediate parts while learning.
* It was my first-time learning, Machine Learning and it was fun learning and knowing how every field I have interest in like, Data Science, Machine Learning, Programming are correlated. And it was also helpful to me in many ways, like showing us how data science and machine learning were related to real life problems.
* I can clearly say that I am satisfied with the training as I have covered what I wanted to know about data science and machine learning before going through the specialization class next semester.