

|  |
| --- |
| Sport Participation webserver & data analysis Interim report  Hdip Computer Science 2021 |
|  |
| February 4  SETU  Authored by: Sheila Kirwan |

# Abstract

Write the abstract at the end once the project is complete

# Acknowledgements

I would like to sincerely thank all the lecturing staff on the Hdip in Computer Science at South East Technological university for your guidance and enthusiasm throughout the course.

I would like to express my thanks to Brenda Mullaly, my project supervisor who provided great support throughout this project.

I would like to thank my husband Oliver and children, Robyn, and Oliver for providing a supportive environment for me to continue learning and growing.

# Preface

Write the preface at the end once the project is complete

# Table of Contents

Contents

[Abstract 2](#_Toc126499506)

[Acknowledgements 3](#_Toc126499507)

[Preface 4](#_Toc126499508)

[Table of Contents 5](#_Toc126499509)

[Table of Figures 6](#_Toc126499510)

[Table of Tables 7](#_Toc126499511)

[List of Abbreviations 7](#_Toc126499512)

[Appendices 7](#_Toc126499513)

[Project Introduction 9](#_Toc126499514)

[Methodology 10](#_Toc126499515)

[Research and Analysis of prospective tools 15](#_Toc126499516)

[Website Research and Methodology 15](#_Toc126499517)

[Data Analysis Research and Methodology 21](#_Toc126499518)

[Final Methodology decided on based on research and learnings 24](#_Toc126499519)

[Design 28](#_Toc126499520)

[Customer requirements and input 29](#_Toc126499521)

[Screen Designs 29](#_Toc126499522)

[Walkthrough 30](#_Toc126499523)

[Data model 30](#_Toc126499524)

[CLASS DIAGRAM 30](#_Toc126499525)

[Bibliography 31](#_Toc126499526)

[Project Tools 32](#_Toc126499527)

[Appendices 34](#_Toc126499528)

# Table of Figures

[Figure 1 13](#_Toc126499642)

[Figure 2 13](#_Toc126499643)

[Figure 3 14](#_Toc126499644)

[Figure 4 15](#_Toc126499645)

[Figure 5 17](#_Toc126499646)

[Figure 6 19](#_Toc126499647)

[Figure 7 19](#_Toc126499648)

[Figure 8 20](#_Toc126499649)

# Table of Tables

# List of Abbreviations

# Appendices

[Appendix A 36](#_Toc126417718)

Researched how a data analyst can provide a portal for a customer to upload and view visualization

Decided on which tool to use. Streamlit with Python

Researched Data Analysis Tool, see list

Decided on which tools to use python using the pandas module, matplotlib, plotly, numpy, Jupyter notebook.

Ramp up phase began

Watched tutorial on python.

What were my main learnings

Watched tutorials on Streamlit

What were my main learnings.

Share all links in bibliography

Website Methodology (software engineering lifecycle)

Decided on website model.

iteration plans – show how the website evolved from one to multipage

Show Models here.

Screen designs

UML, Architectural Model, Process Flow Charts, Data Models (E-R), UI wireframes etc.

Data Analysis Methodology

Show model diagram here.

Walkthrough

Data model

reflection

# Project Introduction

The following interim report outlines my plan for the design and implementation of my project. My project will encompass both web design and data analysis with a high emphasis on data analysis and data visualization.

The customer for this project has a mass participation sports event company. This company organizes many races which can have many participants. Each Participant registers for an event online on their website.

The product of this project is a website that the customer can log into. This website includes user authentication. The website allows the customer to upload a .csv file is a predefined format. The website has many pages each of which provide customer insights and data visualizations using tables and graphs. The data insights provided were specified by the customer as the first steps in this project.

This project is python focused. The website is built using Streamlit which is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science.

The tool and modules that I have selected and learned in order to complete the data analysis are as follows:

* Data Generator website
* Streamlit
* Python
* Pandas
* Matplotlib
* Plotly Express

Planning organizing and diagraming

* Penzu
* Trello
* Figma

IDE

* VSCode
* Jupiter Notebooks

From the project perspective, the main emphasis was on learning how to programmatically analyze data and programmatically create data visualization based on that data through the use of the python tools mentioned. This was of higher importance than the creation of the website created to host this however the creation of the website and selecting the most relevant tool to do so was also a task that proved interesting and useful.

# Methodology

This project is broken up into to two separate areas, including website creation and data analysis and visualization.

The overall project methodology applied to this project is the Agile Methodology.

The Agile Manifesto is as follows:

*We are uncovering better ways of developing  
software by doing it and helping others do it.  
Through this work we have come to value:*

*Individuals and interactions over processes and tools  
Working software over comprehensive documentation  
Customer collaboration over contract negotiation  
Responding to change over following a plan*

*That is, while there is value in the items on  
the right, we value the items on the left more.*

(Beck et al., 2019)

An agile methodology works well when the outcome of the project is unknown. As I have not previously studied any of the module’s, platforms and tools used for this project apart from python language which was part of the Hdip course content, this methodology was a good fit.

The following was the main benefit during this project of working with an agile philosophy

* Ability to manage changing priorities
  + At the beginning of this project, in order decide on a course of action to complete the project proposal website with data visualization, it was necessary for me to ramp up on all of the tools available to me. I began the process with a totally different plan to the plan that I later settled on. A lot of research of these areas and liaising with my supervisor who has experience in data analysis altered my trajectory to the current one.

Within the Agile Methodology, I choose to implement the Scrum Methodology. This methodology was well suited to my chosen project it allows a project to be completed in smaller increments at a time, with continuous experimentation and feedback loops to allow me to improve my project as I progress.

I incorporated scrum events such as Sprints to my planning process. I assessed the timeline available for completion of the project and set up 5 two-week sprint periods.

1. Jan 16 – Jan 29 – Develop website
2. Jan 30 – Feb 12 – Develop website
3. Feb 13 – Feb 26 – Data Analysis and Data Visualization
4. Feb 27 – Mar 12 – Data Analysis and Data Visualization
5. Mar 13 – Mar 26 – Fixes

The following table is my original project timeline plan:

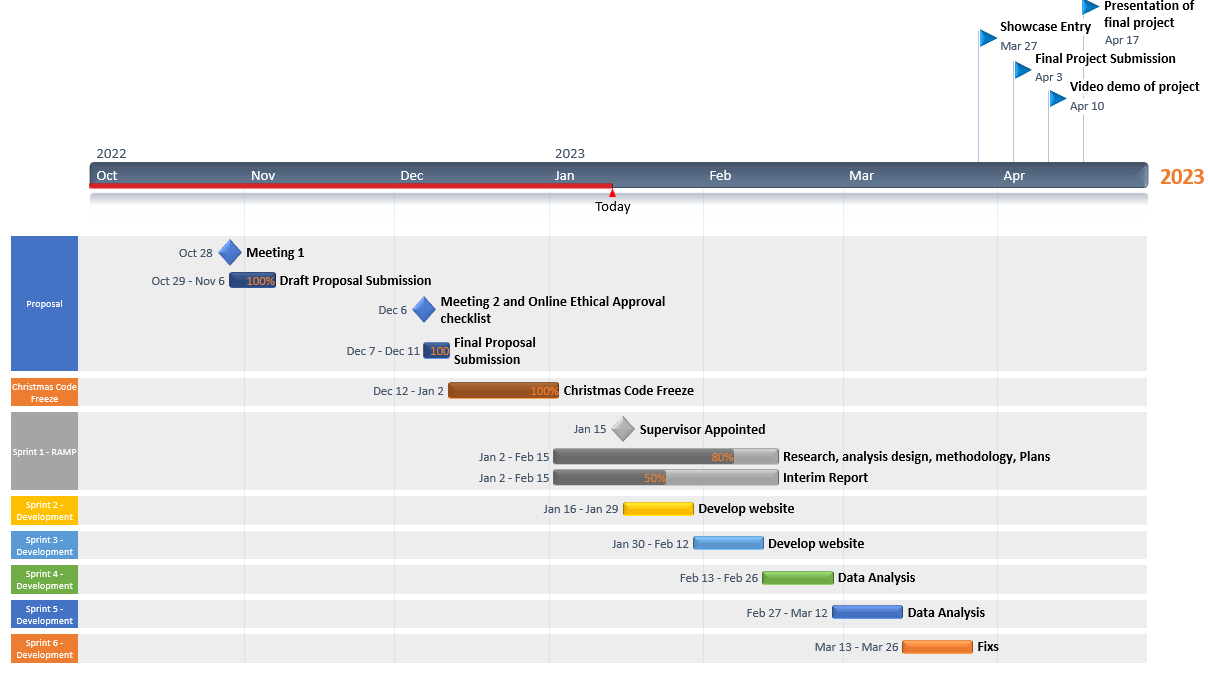


Figure 1

As part of the product backlog required in scrum methodology. I used the Trello as an agile tool to track tasks.

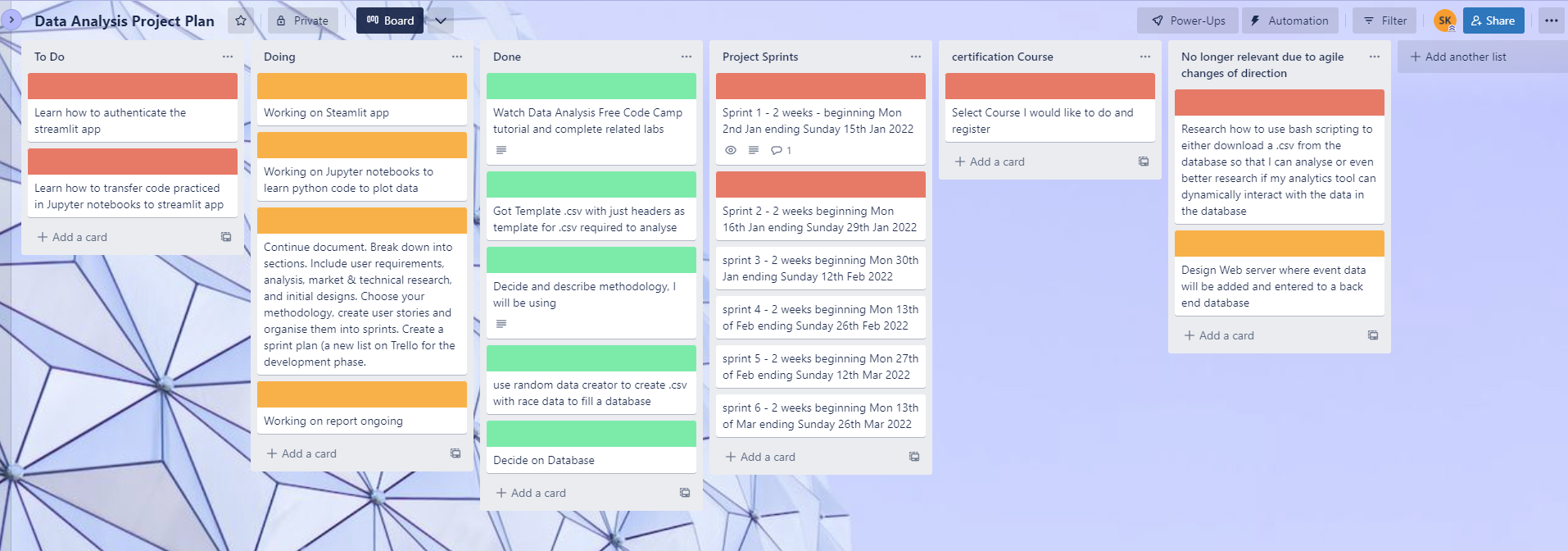


Figure 2

Finally, another tool that I employed throughout the project was the simple Penzu online Journal to track my work and provide a record of my decision making processes throughout the project timeline.

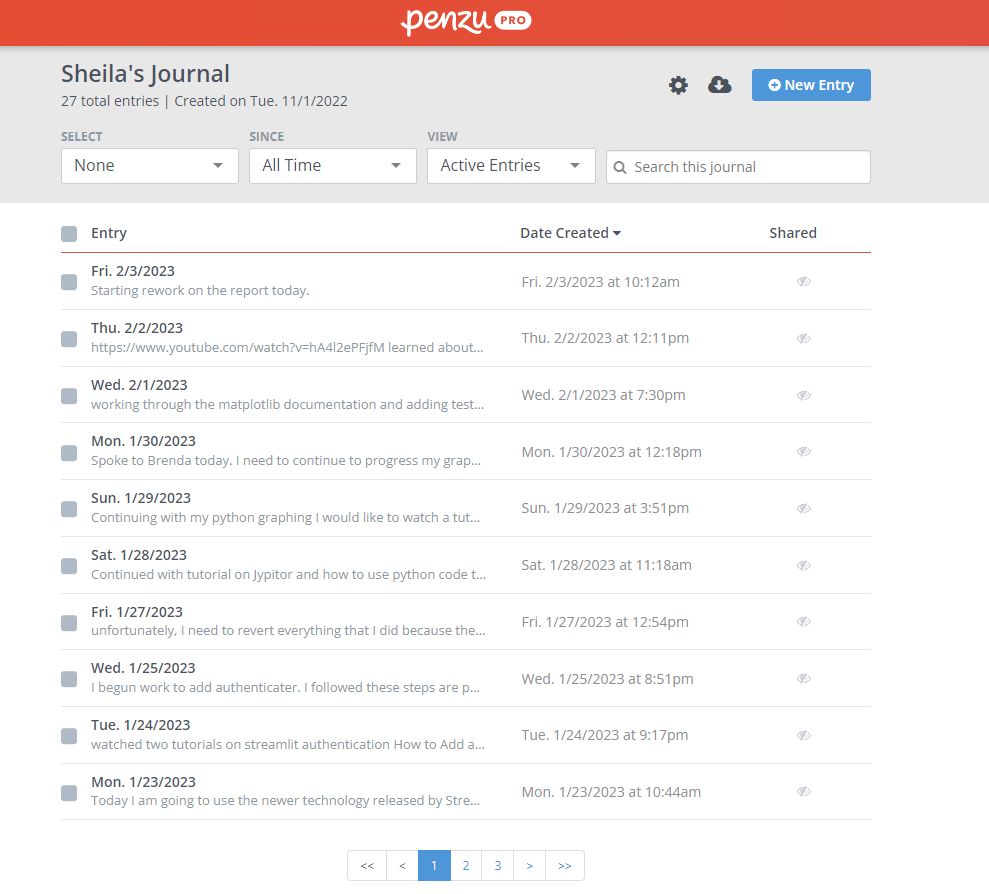


Figure 3

A very important first step as part of the development of this website with Data Analysis was ramping up on the possible portals, modules and processes which could be used to implement this project.

## Research and Analysis of prospective tools

### Website Research and Methodology

Initially, I planned on building a website using the following tools; Node Framework, Hapi Framework and following the Model View Controller (MVC) architectural pattern. The original website was to employ Joi validation, Bulma, font Awesome and be deployed to Heroku. As part of my agile methodology, I decided that the use of these technologies for my data analysis project were not suitable for the following reasons.

* To build this website within the timeline would mean that my project would be a mainly a web development project rather than centered on Data Analysis.
* These tools were not specifically targeted at Data Analysts

This was my original website model.

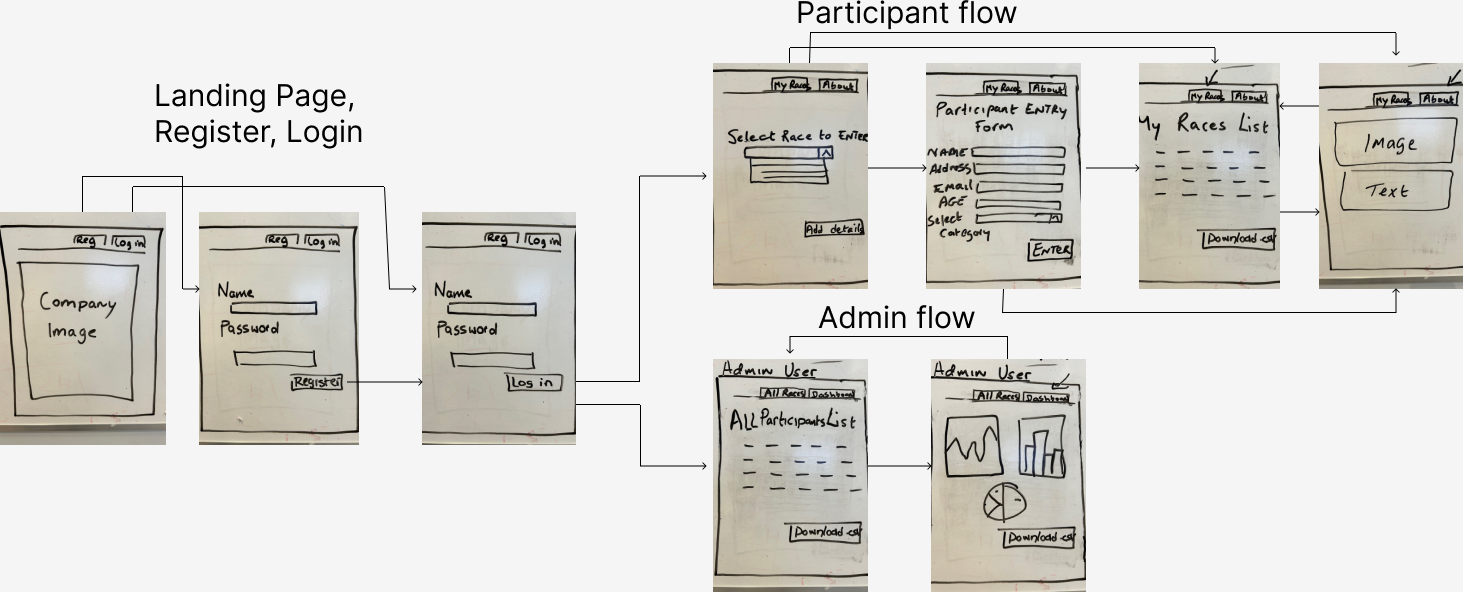


Figure 4

Instead after much research, I decided to use Streamlit to build my data analysis website.

Why did I choose Streamlit?

Streamlit is an open-source python based dashboarding tool. It facilitates the creation and sharing of custom web apps for machine learning and data science. (docs.streamlit.io, n.d.)

Moving data science models from a Proof-Of-Concept (POC) in some notebooks (Jupiter notebook) to a Minimal Viable Product (MVP) that provides business value can be a tough transition. (Willemsen, 2022)

This can have a multitude of reasons:

* The amount of engineering work required to get your model integrated into existing systems.
* The slow (or nonexistent) feedback cycle with your users, resulting in a very slow cycle time for the last few remaining tweaks that need to be done.
* Lack of trust by your users in your model, resulting in your amazing model remaining unused.

Streamlit is tool which can be incorporated to reduce these impacts

In this case, the trade off between integrating the data analysis into a node based web app which would involve a lot of front end work and time or the into a Streamlit dashboard which is more light weight and data analysis targeted was the right choice for this project.

The benefits of using Streamlit for this project.

* Streamlit is not studied as part of the Hdip and hence was new learnings.
* The Streamlit tool is specifically used for data analysis type websites
* Developing the web app using the python based Streamlit app is quicker than traditional web development tools
* Streamlit is Python based which means I can apply the programming knowledge that I have learned from the Hdip in Computer Science course (SETU) to code in Streamlit.

Why not Flask or Django?

Django and Flask are both well-known python-based web framework.

* The most significant benefit of using Streamlit as alternative to Flask is the ability to include HTML code inside the framework Python file. Streamlit does not involve different templates and CSS formatting for the front-end UI.
* The speed of development with Streamlit is faster than with Django.

#### Installation of Streamlit

Step 1. I created a folder on my windows 11 computer and installed Python

Step 2. Next, I used the command *pip install Streamlit* to install the Streamlit library.

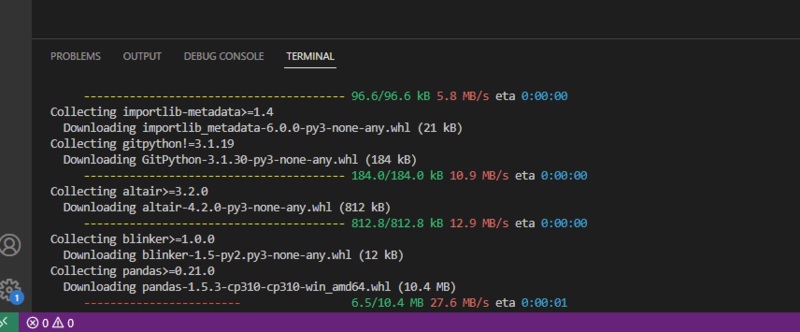


Figure 5

Step 3. I followed instructions in the Streamlit documentation for installation prerequisites (docs.streamlit.io, n.d.)

<https://docs.streamlit.io/library/get-started/installation#prerequisites>

Step 4. I set up virtual environment to run the following commands. I did this by running a PowerShell script (firstly setting the environment to where I will be creating my project folder)

Step 5. I ran the command *pip install virtualenv* to create a virtual environment

Step 6. I then created my project folder at this location and created a new folder within it called Streamlit.

Step 7. I then moved to this folder *cd streamlit* where I then ran the command virtualenv steamlitenv

Step 8. From that project path run the following command to activate the streamlitenv

C:\Users\User\PythonFinalProject\streamlit project>\Scripts\activate

This activated the new environment live with this response on the terminal

(streamlitenv) PS C:\Users\User\PythonFinalProject\streamlit project>

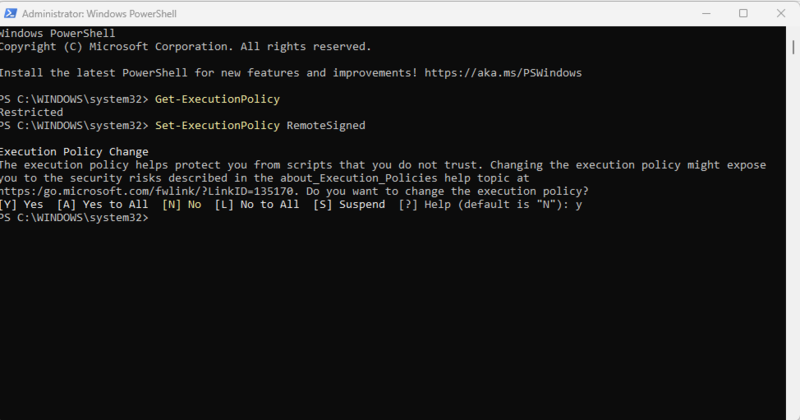
Step 8. I then installed the Streamlit within this environment.

(streamlitenv) PS C:\Users\User\PythonFinalProject\streamlit project> pip install streamlit

This took a few moments to install.

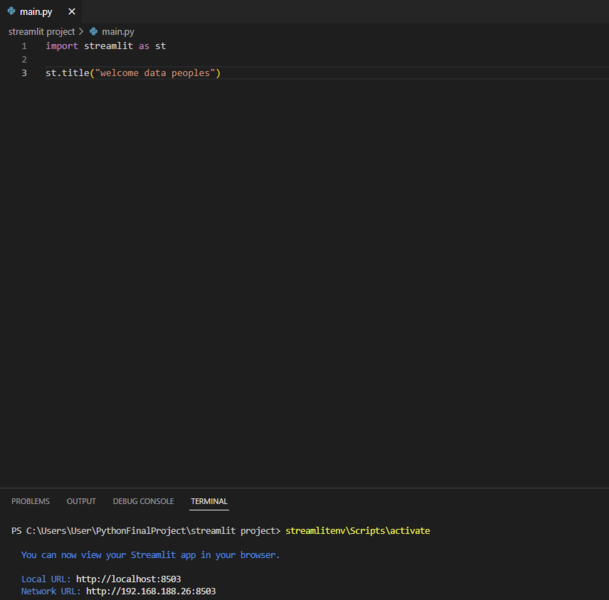
Step 9. To test, type the command *streamlit hello.* This will automatically open streamlit page in the chrome browser

As an aside, I ran into an issue running scripts on Windows and followed the following steps to resolve it by enabling the execution on windows to allow me to run scripts. This is something that I will revert for security reasons.



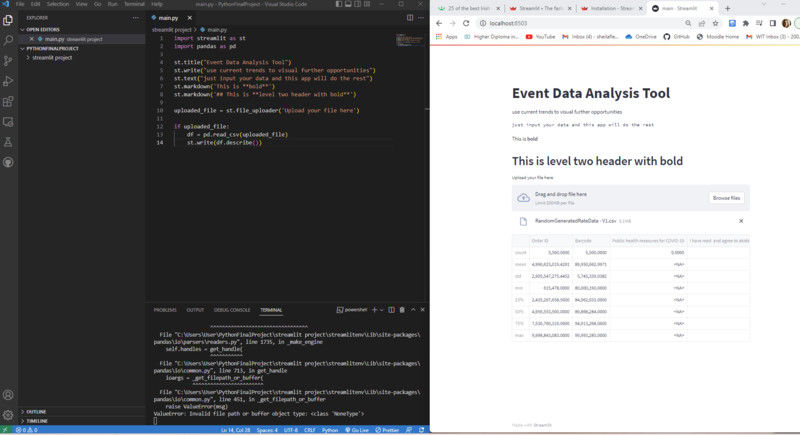
Figure

This is an example of how Streamlit appears in VSCode



Figure

The following is an example of what the Streamlit tool can accomplish with not a lot of code



Figure

It was necessary for me to install Pandas, Matplotlib and Plotly express to my Streamlit folder location via the terminal on vs code.

Authentication and the database used.

To do

Authentication, either through Streamlit or google authentication.

Find out how long a Streamlit app will remember a file that is uploaded?

### Data Analysis Research and Methodology

I completed the following analysis of all data analysis tools in 2022/2023 to find which would fit my requirement most. See this analysis at [Appendix 1](#_Appendices)

Based on the research and analysis undertaken, I selected to use Python. The reasons are as follows:

1. Requires Programming Knowledge and Python was a programming language that I studied as part of the Hdip in Computer Science.

2. Python Library 'Pandas' make data manipulation easier

4. Its Free and Open-source platform

5. Data Analysis is easier due to its simple syntax and built-in functions and libraries

6. Data Visualizations provided by several excellent graphing libraries loaded with a wide range of capabilities



Figure

Image Credit (Karczewski, 2021)

Because, I have never approached data analysis previously, it was necessary to me to learn the basics as well as get a visual introduction to how data analysis with Python can be implemented. In order to build up my knowledge of the area and the steps that I would need to take, I selected to the following free course online.

Data Analysis with Python – Full Course for Beginners (NumPy, Pandas, Matplotlib, Seaborn) (Basulto, 2020)

This is a tutorial over 4 hours with the following course contents:

Part 1: Introduction

What is Data Analysis, why Python? what other options are there? what is the cycle of a Data Analysis project? What is the difference between Data Analysis and Data Science?

Part 2: Real Life Example of a Python/Pandas Data Analysis project

A demonstration of a real-life data analysis project using Python, Pandas, SQL, and Seaborn.

Part 3: Jupyter Notebooks Tutorial

A step-by-step tutorial to learn how to use Jupyter Notebooks

Part 4: Intro to NumPy

Learn why NumPy was such an important library for the data-processing world in Python. Learn about low level details of computations and memory storage, and why tools like Excel will always be limited when processing large volumes of data.

Part 5: Intro to Pandas

Pandas is arguably the most important library for Data Processing in the Python world. Learn how it works and how its main data structure, the Data Frame, compares to other tools like spreadsheets or DFs used for Big Data

Part 6: Data Cleaning

Learn the different types of issues that we will face with our data: null values, invalid values, statistical outliers, etc., and how to clean them.

Part 7: Reading Data from other sources

Part 8: Python Recap

This section is a recap of Python main features and control flow structures.

### Final Methodology decided on based on research and learnings

* This Data Analysis will be completed with Python and the entire Py data stack to perform this data analysis. However, the main modules that I will use are pandas, NumPy, matplotlib and Plotly express from the Py stack.
* I will be using Jupyter notebooks as my python text editor.
* I will be reading data from csv files.
* I will clean and transform the .csv data using statistical functions and create useful visualizations

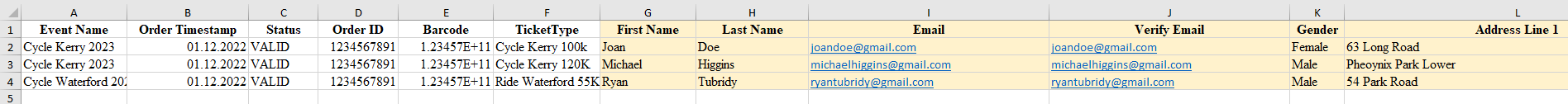
# Design

[Data](https://www.simplilearn.com/what-is-data-article) is perhaps one of the most valuables assets that a business can have today. Data defines the market intelligence that businesses large and small can gather about their customers and the market they are operating in. In other words, it can make or break a company. (Van Loon, 2022)

## Customer requirements and input

In order to understand the customers’ requirements, I requested an excel spreadsheet with the specific column headers as per the .csv file which is the output a of system used by the customer.

The following is a sample of this file with random generated data.



Figure

## Screen Designs

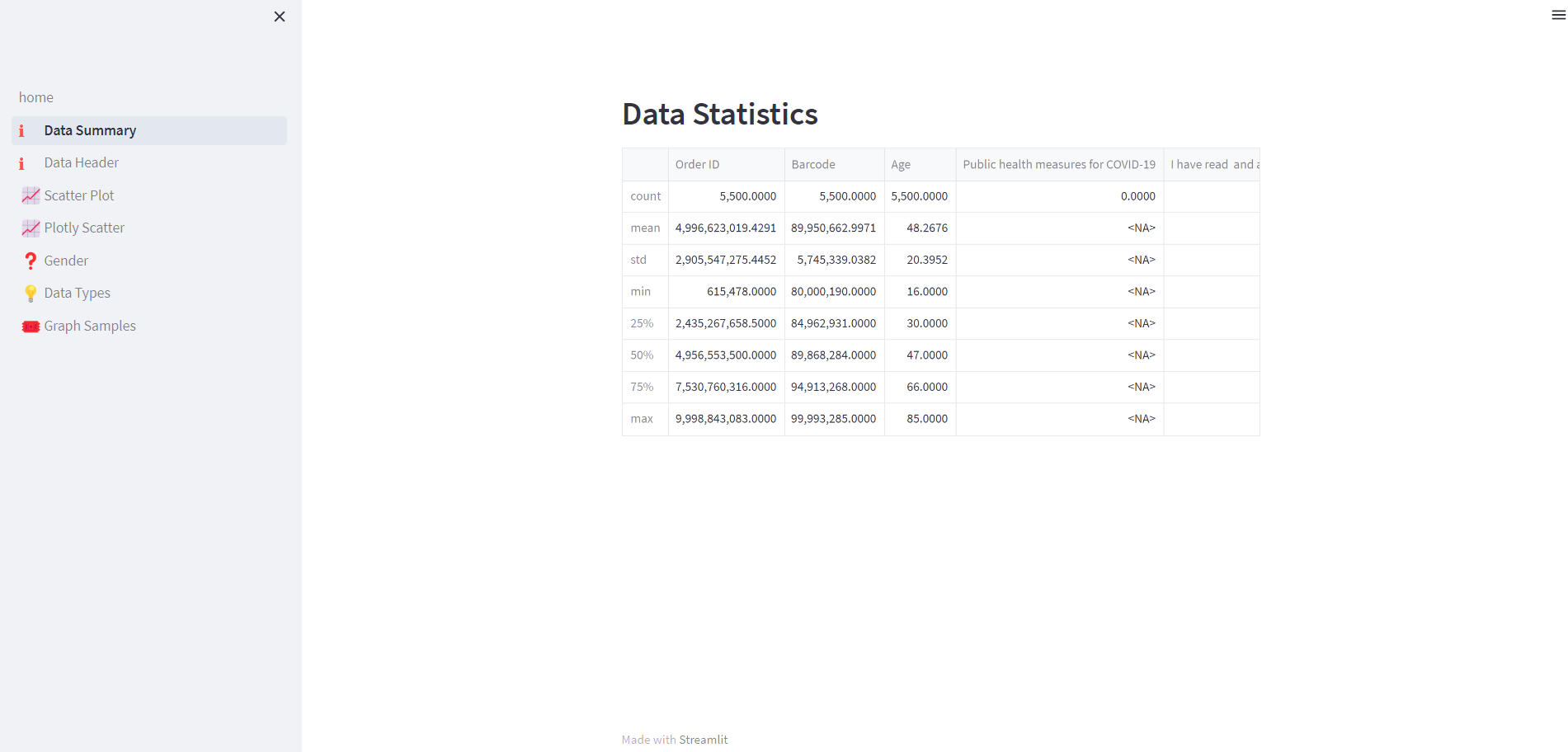
### Initial Screen Designs



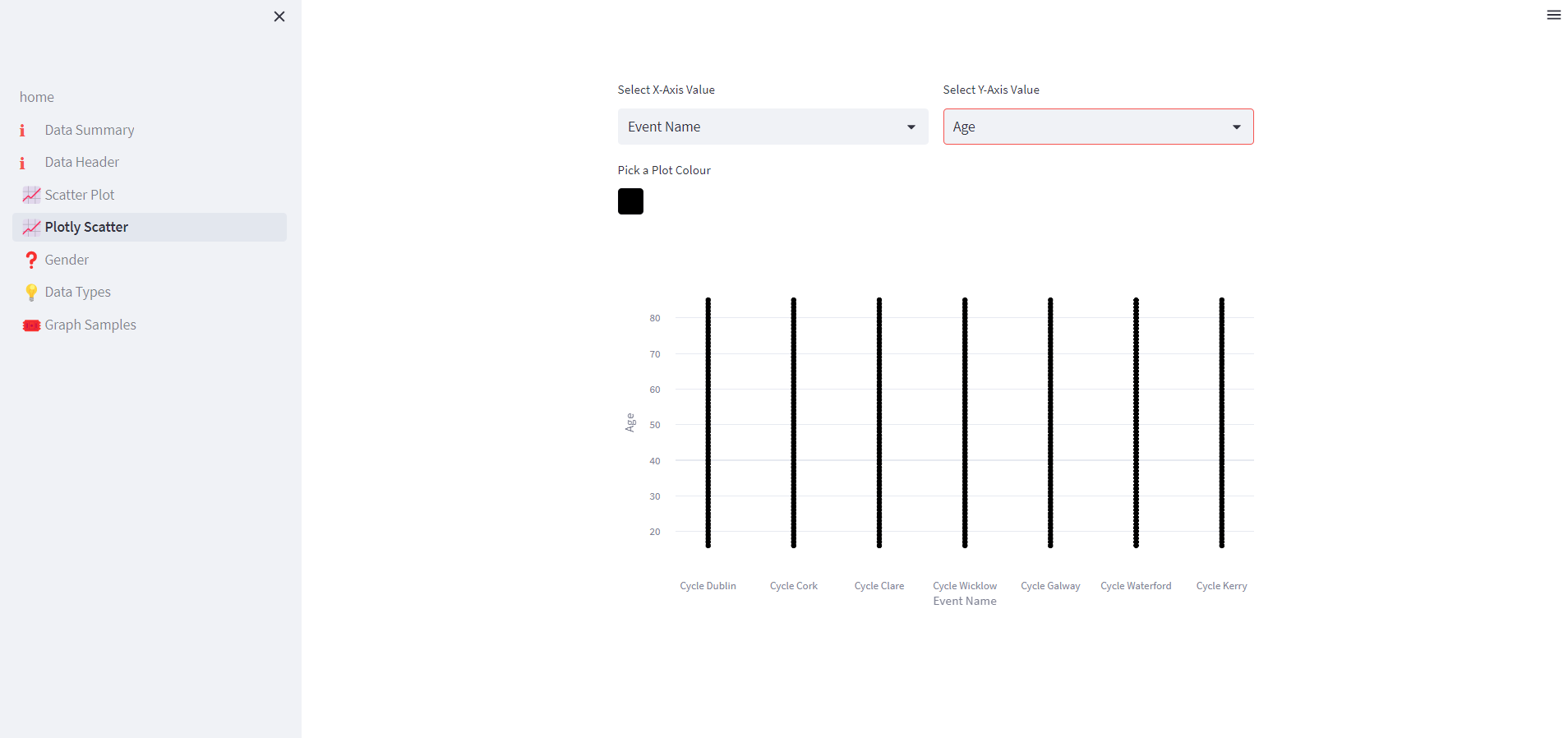
Drag and drop file



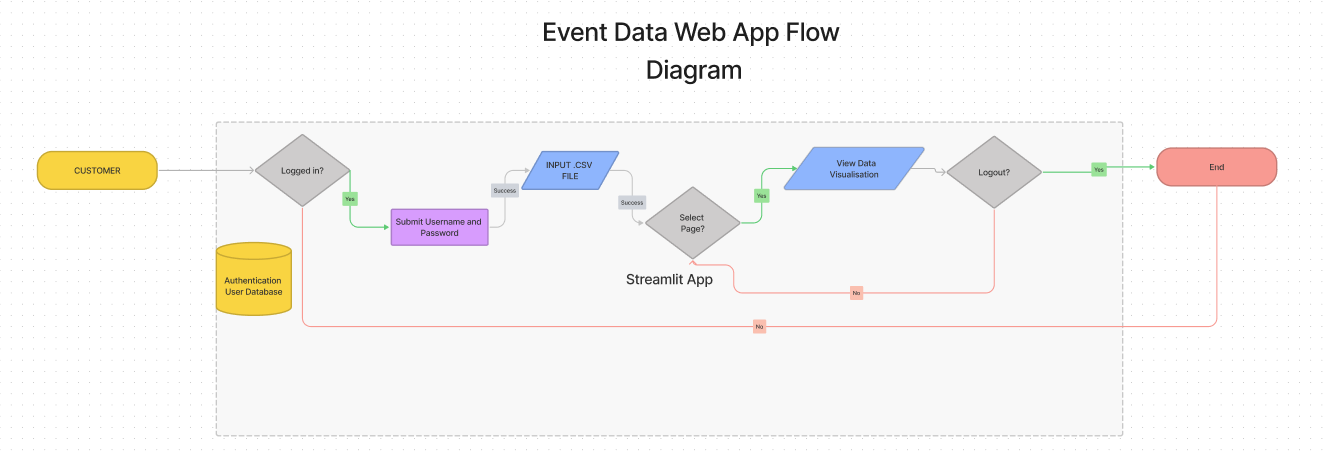
Select Page called Data Summary



Select Plotly Scatter Page



Final Screen Designs



Figure

Please see detailed Diagram below in Appendix B

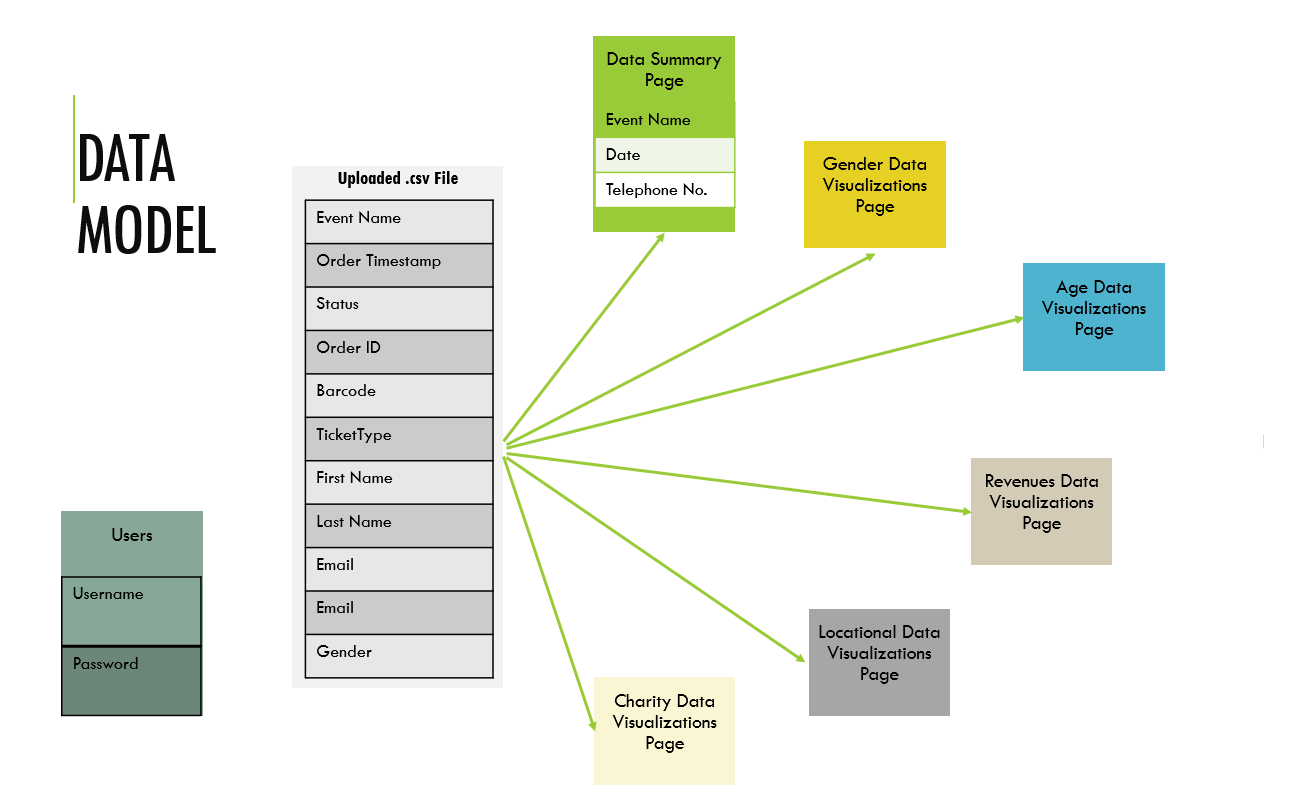
* Log in (authentication)
* Upload file
* Select Page with Data insights and visualizations required
* Log out



Figure

## Walkthrough

## Data model



I have also included a larger version of this in the Appendix section C.

[UNFINISHED]

## Class Diagram

# Bibliography

There are no sources in the current document.

# Project Tools

[Creation of timeline plan - https://www.officetimeline.com/office-timeline/14-days-trial#download-office-timeline1](Creation%20of%20timeline%20plan%20-%20https://www.officetimeline.com/office-timeline/14-days-trial#download-office-timeline1)

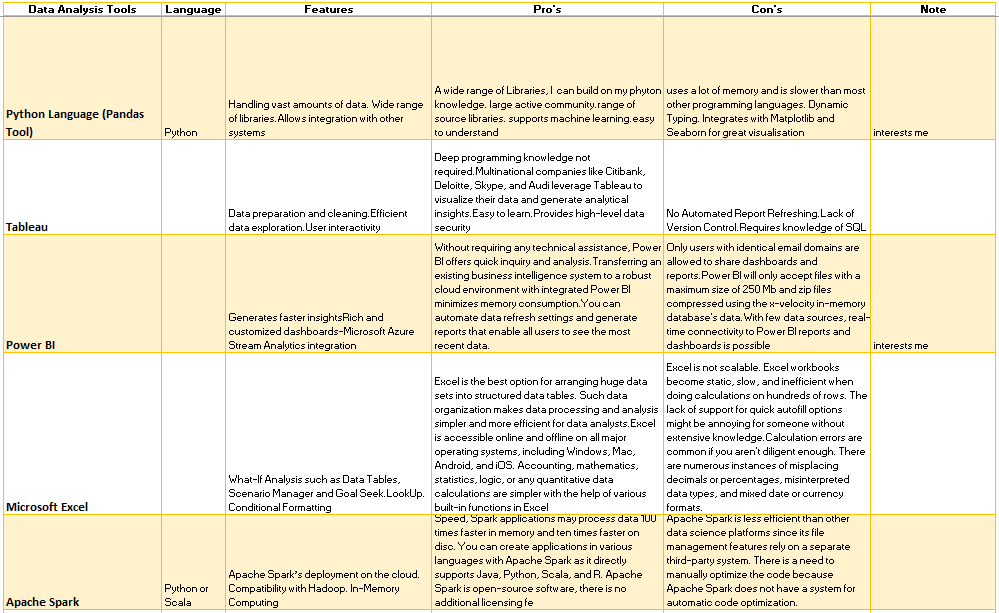
DECLARATION OF AUTHENTICITY

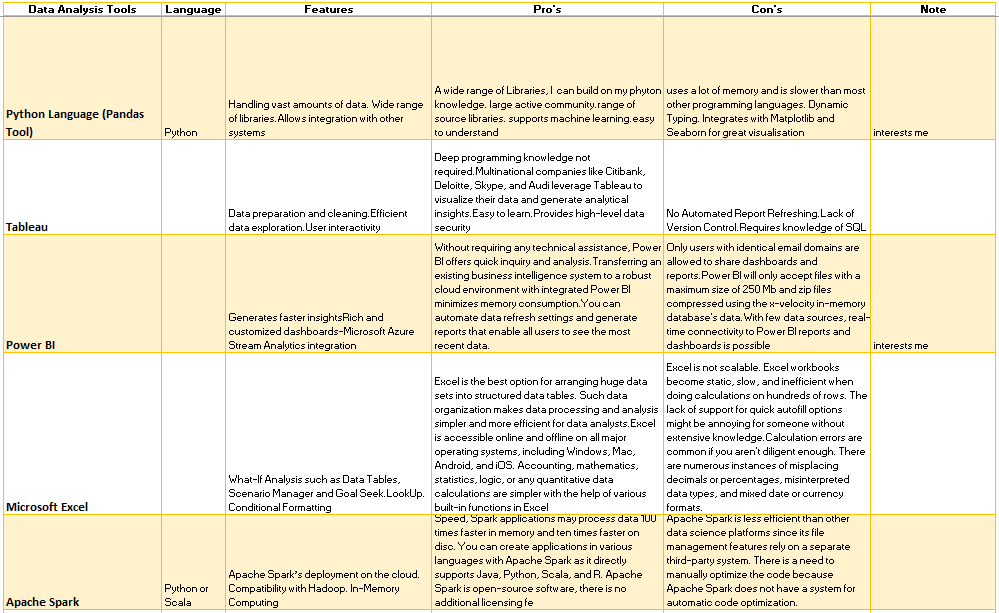
I declare that all material presented to South East Technological University (SETU) is my own work, or fully and specifically acknowledged wherever adapted from other sources. I understand that if at any time it is shown that I have significantly misrepresent material presented to South East Technological University (SETU), and degree or credits awarded to me based on that material may be revoked.

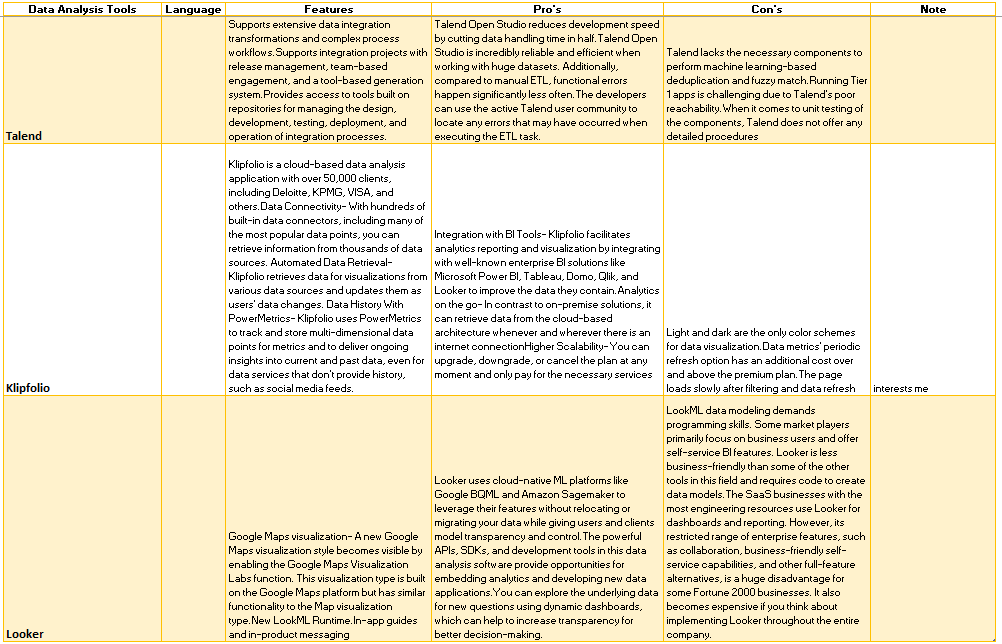
|  |  |  |  |
| --- | --- | --- | --- |
| Student’s Signature |  | Date |  |
| Supervisor’s Signature |  | Date |  |

# Appendices

Appendix A

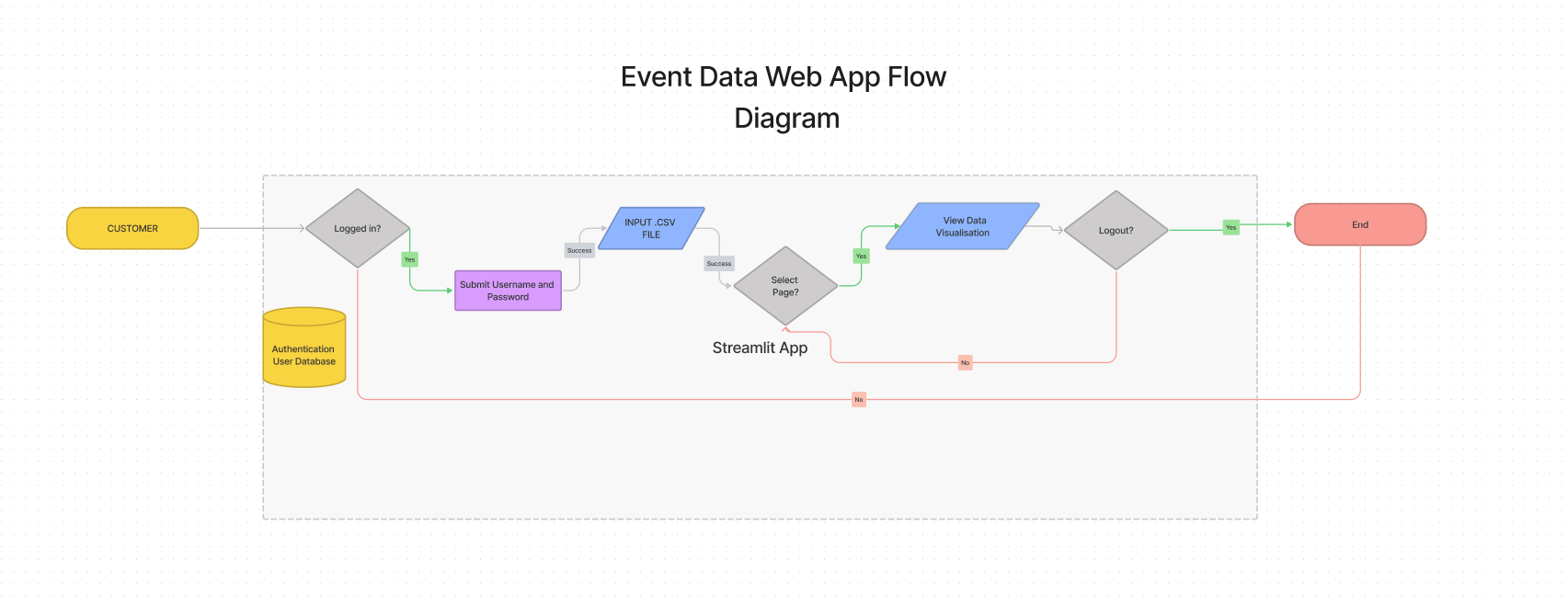






(ProjectPro, 2023)

Appendix B



Appendix

Appendix C

Appendix

