**Project Report**

**Web-Based Data Analysis & Prediction**

## Submitted by

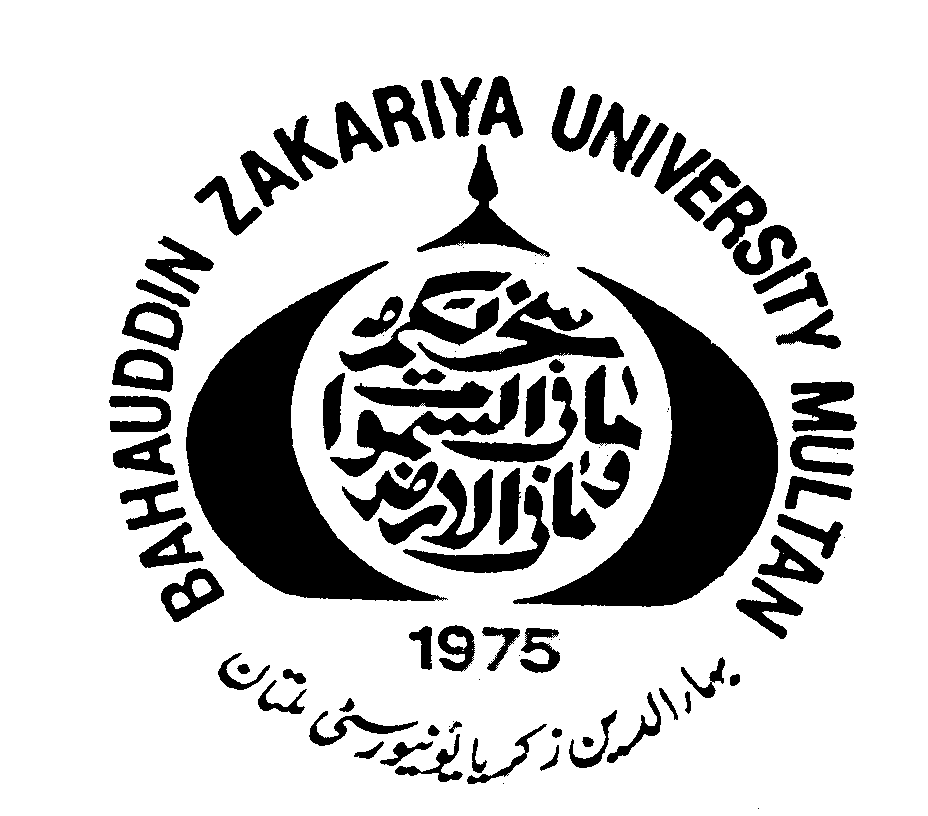
## BSCS-19-46

## Hazrat Bilal

## 2019 - 2023

## Supervised by

## Dr. Muhammad Asif Raza



## DEPARTMENT OF COMPUTER SCIENCE

**BAHAUDDIN ZAKARIYA UNIVERSITY MULTAN PAKISTAN**

**FINAL APPROVAL**

This is to certify that we have read this project report titled “Web-Based Data Analysis & Prediction” submitted by ***Hazrat Bilal*** *roll no* ***“BSCS-19-46”*** and it is our judgement that this report is of sufficient standard to warrant its acceptance by Bahauddin Zakariya University, Multan for the degree of BS(CS) (Bachelor of Science in Computer Science).

***Committee:***

**1. External Examiner \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Dr. Salman Qadri**

Associate Professor/chairman,

Department of Computer Science

Muhammad Nawaz Sharif University of Agriculture, Multan

**2. Supervisor \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Dr. Muhammad Asif Raza**

Assistant Professor,

Department Of Computer Science

Bahauddin Zakariya University, Multan

**3. Head of Department \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Dr. Minhaj Ahmed Khan**

Professor,

Department of Computer Science

Bahauddin Zakariya University, Multan

**DEDICATION**

***To my Extraordinary Parents and Remarkable Teacher, whose Unwavering Support and Guidance have shaped me into who I am today. To my Incredible Brother, who has been my Pillar of Strength and a Guiding Light, fulfilling the role of a Father. To my Loyal Friends, whose Unwavering Presence in both my Darkest Hours and Brightest Moments has been nothing short of a Blessing. And lastly, to the Dedicated Department and Exceptional Teachers, whose Wholehearted Efforts have nurtured and propelled me towards Success. This Dedication is a Testament to the Profound Impact these individuals have had on my life, and I am forever Grateful.***

**ACKNOWLEDGMENT**

In the completion of my project, I had to take the help and guidance of some respected persons, who deserve my greatest gratitude. The completion of the final project gave me much pleasure. I would like to show my gratitude to **Dr.** **Muhammad Asif Raza**, Assistant Professor in the Department of Computer Science, BZU Multanfor giving me a good guideline for the project through his valuable & fruitful Consultations. I would like to expand my deepest gratitude to all those who have directly and indirectly guided me in writing this project.

**Hazrat Bilal**

**BSCS-19-46**

**PROJECT BRIEF**

|  |  |
| --- | --- |
| PROJECT NAME | Web-Based Data Analysis & Predication |
| UNDERTAKEN BY | Hazrat Bilal |
| SUPERVISED BY | Dr. Muhammad Asif Raza |
| STARTING DATE | January 30, 2023 |
| COMPLETION DATE | June 7, 2023 |
| COMPUTER USED | Intel(R) Core (TM) i5-4210U CPU @ 1.70GHz 2.40 GHz Memory: 12,288MB RAM |
| OPERATING SYSTEM | Windows 10 |
| SOURCE LANGUAGE(S) | Python |
| DBMS USED | None |
| TOOLS/PACKAGES | Streamlit (Python Framework), Pandas, Numpy, Plotly, Matplotlib, Sciket-learn, AgGridPro |

**ABSTRACT**

Introducing the Web-Based Data Analysis & Prediction Application, an exceptional tool designed to empower users and organizations with comprehensive data analysis capabilities. With a user-friendly interface, this application enables deep analysis, efficient pre-processing, and data cleaning, ensuring accuracy and reliability. Say goodbye to coding and welcome a seamless experience that empowers users to make informed decisions and perform future predictions based on their data.

The project's core objective is to provide users with complete data security, eliminating concerns about unauthorized access. By offering a streamlined process for data cleaning and analysis, this application allows users to focus on gaining valuable insights and maximizing the potential of their data.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Contents** | **Page No.** | |
| *Chapter 01*  **Introduction** | | |
| 1.1 Project Introduction | | 11 |
| 1.1.1 Main Theme | | 11 |
| 1.1.2 Scope of the Project | | 12 |
| 1.1.3 Technical Details | | 12 |
| 1.1.4 Objectives of the Project | | 13 |
| 1.1.5 Features of the Project | | 14 |
| 1.1.5.1 User-Friendly Interface | | 14 |
| 1.1.5.2 Data Cleaning | | 14 |
| 1.1.5.3 Data Grouping and Pivoting | | 15 |
| 1.1.5.4 Data Analysis | | 15 |
| 1.1.5.5 Data Visualization | | 15 |
| 1.1.5.6 Deep Analysis | | 15 |
| 1.1.5.7 Data Predication | | 15 |
| 1.1.5.8 Data Storage | | 15 |
| *Chapter 02*  **System Analysis** | | |
| 2.1 Feasibility Study | | 16 |
| 2.1.2 Technical Feasibility | | 16 |
| 2.1.2 Economic Feasibility | | 16 |
| 2.1.3 Operational Feasibility | | 16 |
| 2.2 Linear Regression | | 17 |
| 2.2.1 Machine Learning Model Cycle | | 17 |
| *Chapter 03*  **System Design** | | |
| 3.1 Introduction to System Design | | 21 |
| 3.2 Proposed System and its Features | | 21 |
| 3.3 Features of the Project | | 21 |
| 3.3.1 User-Friendly Interface | | 20 |
| 3.3.2 Data Cleaning | | 22 |
| 3.3.3 Data Grouping and Pivoting | | 22 |
| 3.3.4 Data Analysis | | 22 |
| 3.3.5 Data Visualization | | 22 |
| 3.3.6 Deep Analysis | | 22 |
| 3.3.7 Data Predication | | 22 |
| 3.3.8 Data Storage | | 22 |
| 3.4 Dataset | | 23 |
| 3.4.1 Training Dataset | | 23 |
| 3.4.2 Analysis & Predication | | 23 |
| 3.5 System Design using UML | | 23 |
| 3.5.1 Sequence Diagram | | 23 |
| 3.5.2 Use Case Diagram | | 25 |
| *Chapter 04*  **System Implementation** | | |
| 4.1 Introduction | | 27 |
| 4.2 Tool/ Language/ Technology Selection Criteria | | 27 |
| 4.3 System Implementation | | 28 |
| 4.4 Front End (Streamlit: Python) | | 29 |
| 4.4.1 Streamlit | | 29 |
| 4.4.2 Python Language | | 29 |
| 4.4.3 Libraries | | 30 |
| 4.5 IDE (Visual Studio Code) | | 30 |
| 4.6 Operation system Selection | | 31 |
| 4.7 Streamlit | | 32 |
| 4.7.1 Advantages of Streamlit | | 32 |
| 4.8 Implementation | | 32 |
| *Chapter 05*  **User Guide** | | |
| 5.1 Main Page | | 35 |
| 5.2 Training | | 36 |
| 5.2.1 Load Dataset Button | | 37 |
| 5.2.2 Head & Tail | | 37 |
| 5.2.3 Statistics Button | | 38 |
| 5.2.4 Remove Duplicates Button | | 40 |
| 5.2.5 Handle the Null values | | 40 |
| 5.2.5.1 Ways to handle Null Values | | 41 |
| 5.2.6 See Changes Button | | 44 |
| 5.2.7 Data Operations | | 45 |
| 5.3 Analysis & Predication | | 52 |
| 5.3.1 Upload Dataset Button | | 52 |
| 5.3.2 Deep Drive into Dataset | | 53 |
| 5.3.3 Linear Regression | | 55 |
| References | | 58 |

**Chapter 01**

**INTORDUCTION**

* 1. **Project Introduction**

Many organizations and users have important data that they gain from questionnaires, sales, products, and observations. This data can provide valuable insights that can help inform decision-making. However, sharing sensitive data with others to gain these insights can be a concern for many due to privacy issues and the time-consuming process involved.

The aim of this project is to provide an opportunity for users or organizations to quickly and efficiently gain important insights without having to share their data with others. The project is designed to be user-friendly, allowing users to avoid writing any code and instead use buttons to perform operations on their data. This allows them to make informed decisions based on their own data, without having to rely on external parties.

The project contains two main modules:

1. **Training:** This module is designed to train the user for better understanding of their data, operations on data, and visualizations. It includes step-by-step instructions for performing operations on the data before the actual analysis takes place. This helps users become more familiar with their data and the tools available for analyzing it.
2. **Analysis & Prediction:** This module is designed to allow the user to add their own dataset for cleaning, grouping, pivoting, and deep analysis. It also includes the ability to perform predictions based on the dataset. The data will not be stored by the designer of the project, but the user is allowed to store their data, pivot tables, graphs, and charts without any restrictions. This gives users full control over their own data and allows them to perform in-depth analysis and make predictions based on it.
   * 1. **Main Theme**

Data can be analyzed, visualized, and manipulated in a number of ways using different tools and websites. However, these methods often require the user to share their data with others, which can raise concerns about privacy and security. Additionally, many of these tools require the user to have some knowledge of coding, which can be a barrier for some users.

The main theme of Web-Based Data Analysis & Prediction is to provide a secure and user-friendly way for users to analyze their own data without having to share it with others or write any code. The project is designed to provide a simple and intuitive interface that users can use without hesitation.

* + 1. **Scope of the Project**

This project aims to provide a comprehensive solution for users who want to analyze their own data without having to rely on external parties or learn how to code. The project includes features for cleaning, grouping, pivoting, and analyzing data, as well as the ability to perform predictions based on the dataset. This allows users to gain valuable insights from their own data in a secure and user-friendly way.

In addition, the project also allows users to save their cleaned data, pivot tables, grouped tables, graphs, and charts in various formats such as comma-separated values (CSV), Excel, text, Python, and pickle. This gives users even more control over their own data and allows them to easily share their findings with others.

The project has the potential to be widely adopted by organizations and individuals who want to analyze their own data without having to share it with others or learn how to code. It could be particularly useful for small businesses or individuals who want to gain insights from their data but don’t have the resources or expertise to do so using traditional methods.

* + 1. **Technical Details**

The Web-Based Data Analysis & Prediction project is built using the Streamlit framework, which is a popular Python library for building data-driven web applications. This allows the project to be easily accessed and used as a web application, as well as an application on a user’s PC or mobile phone.

Streamlit provides a simple and intuitive way to build interactive data applications, making it an ideal choice for this project. Users can easily interact with their data and perform operations using the buttons and widgets provided by the Streamlit interface.

* + 1. **Objectives of the Project**

The main objectives of the Web-Based Data Analysis & Prediction project are to provide a secure and user-friendly way for users to analyze their own data without having to share it with others or write any code. The project aims to provide a comprehensive solution for data analysis, including features for cleaning, grouping, pivoting, and analyzing data, as well as the ability to perform predictions based on the dataset.

The project is designed to provide a simple and intuitive interface that users can use without hesitation, allowing them to easily interact with their data and perform operations using buttons and widgets. Users will also be able to save their cleaned data, pivot tables, grouped tables, graphs, and charts in various formats such as CSV, Excel, Python, text, and pickle on their PC or mobile device for easy sharing and further analysis.

In addition, the project includes step-by-step instructions and training to help users become more familiar with their data and the tools available for analyzing it. The project will be easily accessible as a web application, as well as an application on a user’s PC or mobile phone.

* + 1. **Features of the Project**
       1. **User-Friendly Interface:** The project is designed to provide a simple and intuitive interface that users can use without hesitation. This allows them to easily interact with their data and perform operations using buttons and widgets, without having to write any code.
       2. **Data Cleaning:** The project includes features for cleaning data, allowing users to easily remove any errors or inconsistencies in their dataset. This helps ensure that the data is accurate and reliable before performing any analysis on it.
       3. **Data Grouping and Pivoting:** The project also includes features for grouping and pivoting data, allowing users to easily organize their data in a way that makes it easier to analyze. This can help reveal patterns and trends in the data that may not be immediately apparent.
       4. **Data Analysis:** The project includes powerful tools for analyzing data, allowing users to perform deep analysis and gain valuable insights from their own data. This includes the ability to create graphs and charts to visualize the data, as well as perform statistical analysis to identify trends and patterns.
       5. **Data Visualization:** In addition to analyzing data, the project also includes features for visualizing data in 1D (single column), 2D (two columns at a time), and 3D. This allows users to easily see patterns and trends in their data, making it easier to understand and analyze.
       6. **Deep Analysis:** The project also includes the ability to perform deep analysis using the Pandas Profiling library, which provides detailed reports on the characteristics of the dataset. This can help users gain a deeper understanding of their data and identify any potential issues or areas for further analysis.
       7. **Data Prediction:** In addition to analyzing data, the project also includes the ability to perform predictions based on the dataset. This allows users to make informed decisions about future events or trends based on their own data.
       8. **Data Storage:** The project allows users to save their cleaned data, pivot tables, grouped tables, graphs, and charts in various formats such as CSV, Excel, Python, text, and pickle on their PC or mobile device for easy sharing and further analysis.

**Chapter 02**

**SYSTEM ANALYSIS**

* 1. **Feasibility Study**

The purpose of this feasibility study is to assess the viability and potential success of developing a web-based data analysis and prediction tool. The tool will provide users with the ability to analyze their own data without having to share it with others or write any code. This study will evaluate the technical, economic, and operational feasibility of the project.

* + 1. **Technical Feasibility**

The technical feasibility of the web-based data analysis and prediction project is high. The required technologies and tools for cleaning, grouping, pivoting, and analyzing data, as well as performing predictions based on the dataset, are well-established and readily available. Python libraries such as Pandas and NumPy can be utilized to build the tool. Additionally, integrating the tool with existing systems, such as user databases and data storage solutions, is technically feasible.

* + 1. **Economic Feasibility**

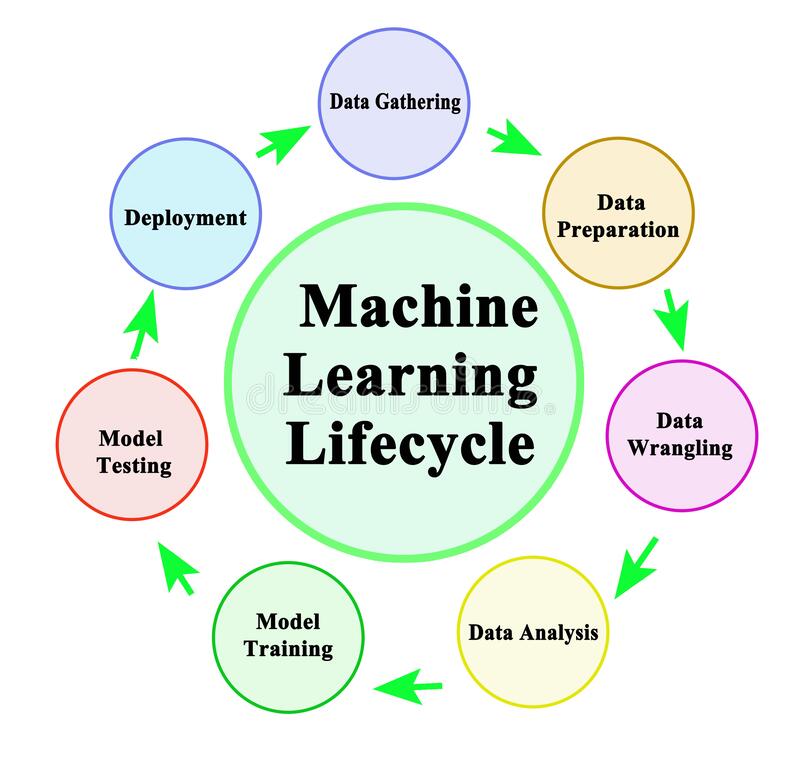
The economic feasibility of the web-based data analysis and prediction project is favorable. Implementing a tool that allows users to analyze their own data without having to share it with others or write any code can lead to cost savings by reducing the need for additional human resources or external parties to perform data analysis. The tool can improve operational efficiency and potentially increase user satisfaction by providing a secure and user-friendly way to analyze data. However, there will be upfront costs for development, integration, and ongoing maintenance of the tool.

* + 1. **Operational Feasibility**

The operational feasibility of the web-based data analysis and prediction project is high. The tool can provide users with a simple and intuitive interface for analyzing their own data without having to share it with others or write any code. The tool can handle tasks such as cleaning, grouping, pivoting, and analyzing data, as well as performing predictions based on the dataset. It can enhance the overall user experience and improve operational efficiency by automating routine tasks.

* 1. **Linear Regression**

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the line of best fit that can accurately predict the value of the dependent variable based on the values of the independent variables.

****

**Figure 2.1 Machine Learning Model Cycle [1]**

* + 1. **Machine Learning Model Cycle**

The cycle of Machine Learning Model typically involves the following steps:

1. **Data Gathering:** This involves collecting data on the dependent and independent variables. Data can be gathered from various sources such as surveys, experiments, or existing databases. It is important to ensure that the data is representative of the population being studied.
2. **Data preparation:** Once the data has been gathered, it must be cleaned and transformed into a format suitable for analysis. This involves removing any errors or inconsistencies in the data, as well as handling missing values and dealing with outliers. Data preparation may also involve normalizing or standardizing the data to ensure that all variables are on the same scale.
3. **Data wrangling:** This step involves manipulating the data to make it easier to work with. This may include merging multiple datasets, reshaping the data, or creating new variables. Data wrangling is an important step in ensuring that the data is in the right format for analysis.
4. **Data analysis:** Before fitting a model, it is important to perform exploratory data analysis to understand the relationships between the variables. This may involve visualizing the data using scatter plots, histograms, and box plots, as well as calculating summary statistics such as mean, median, and standard deviation.
5. **Model training:** Once the data has been prepared and analyzed, a model can be fit to the data. This involves estimating the coefficients of the regression equation that minimize the sum of squared errors between the observed values of the dependent variable and the predicted values. Model training may be done using various methods such as ordinary least squares, ridge regression, or lasso regression.
6. **Model testing:** After fitting the model, it must be evaluated to determine how well it fits the data. This involves calculating various measures of model fit, such as R-squared, adjusted R-squared, and root mean squared error, as well as performing statistical tests to assess the significance of the model coefficients. Model testing may also involve checking for violations of the assumptions of linear regression, such as linearity, independence, homoscedasticity, and normality.
7. **Deployment**: Once the model has been trained and tested, it can be used to make predictions on new data. This involves using the estimated coefficients to predict the value of the dependent variable based on the values of the independent variables. Deployment may also involve integrating the model into a larger system or application.

**Chapter 03**

**SYSTEM DESIGN**

* 1. **Introduction to System Design**

The purpose of system design in the context of a web-based data analysis and prediction tool is to create an efficient and reliable platform that allows users to analyze their own data without having to share it with others or write any code. The design will focus on software development activities to meet the requirements of the tool, including features for cleaning, grouping, pivoting, and analyzing data, as well as performing predictions based on the dataset.

The tool will be optimized for seamless integration with existing systems and will incorporate user-friendly interfaces and intuitive navigation to ensure a smooth and satisfying experience for users. The design will prioritize security, reliability, and efficiency to provide a comprehensive solution for data analysis.

During the system design phase, various activities such as system architecture design, database design, software component design, user interface design, and security design will be carried out. The goal of this phase is to create a software system that meets the requirements of the stakeholders and is scalable, maintainable, and efficient while ensuring the system’s security and reliability.

* 1. **Proposed System and its Features**

The Web-Based Data analysis & Prediction tool is a web-based application developed to provide users with the ability to analyze their own data without having to share it with others or write any code. The main purpose of the tool is to provide a secure and user-friendly way for users to gain important insights from their own data. The system provides features such as data cleaning, grouping, pivoting, analysis, and prediction.

* 1. **Features of the Project**

There are many features of the proposed system. Some of them are

* + 1. **User-Friendly Interface:** The project is designed to provide a simple and intuitive interface that users can use without hesitation. This allows them to easily interact with their data and perform operations using buttons and widgets, without having to write any code.
    2. **Data Cleaning:** The project includes features for cleaning data, allowing users to easily remove any errors or inconsistencies in their dataset. This helps ensure that the data is accurate and reliable before performing any analysis on it.
    3. **Data Grouping and Pivoting:** The project also includes features for grouping and pivoting data, allowing users to easily organize their data in a way that makes it easier to analyze. This can help reveal patterns and trends in the data that may not be immediately apparent.
    4. **Data Analysis:** The project includes powerful tools for analyzing data, allowing users to perform deep analysis and gain valuable insights from their own data. This includes the ability to create graphs and charts to visualize the data, as well as perform statistical analysis to identify trends and patterns.
    5. **Data Visualization:** In addition to analyzing data, the project also includes features for visualizing data in 1D (single column), 2D (two columns at a time), and 3D. This allows users to easily see patterns and trends in their data, making it easier to understand and analyze.
    6. **Deep Analysis:** The project also includes the ability to perform deep analysis using the Pandas Profiling library, which provides detailed reports on the characteristics of the dataset. This can help users gain a deeper understanding of their data and identify any potential issues or areas for further analysis.
    7. **Data Prediction:** In addition to analyzing data, the project also includes the ability to perform predictions based on the dataset. This allows users to make informed decisions about future events or trends based on their own data.
    8. **Data Storage:** The project allows users to save their cleaned data, pivot tables, grouped tables, graphs, and charts in various formats such as CSV, Excel, Python, text, and pickle on their PC or mobile device for easy sharing and further analysis.
  1. **Dataset**

The dataset is a crucial component of both the Training and Analysis & Prediction modules of the Web-Based Data analysis & Prediction tool.

* + 1. **Training Dataset**

For the training module, the famous “Titanic” dataset is used. This dataset includes a variety of data types, such as integers (int64), floating-point numbers (float64), strings (str), objects, and booleans (bool). The purpose of using the “Titanic” dataset is to introduce users to different data types and show them how to handle and work with them.

* + 1. **Analysis & Prediction**

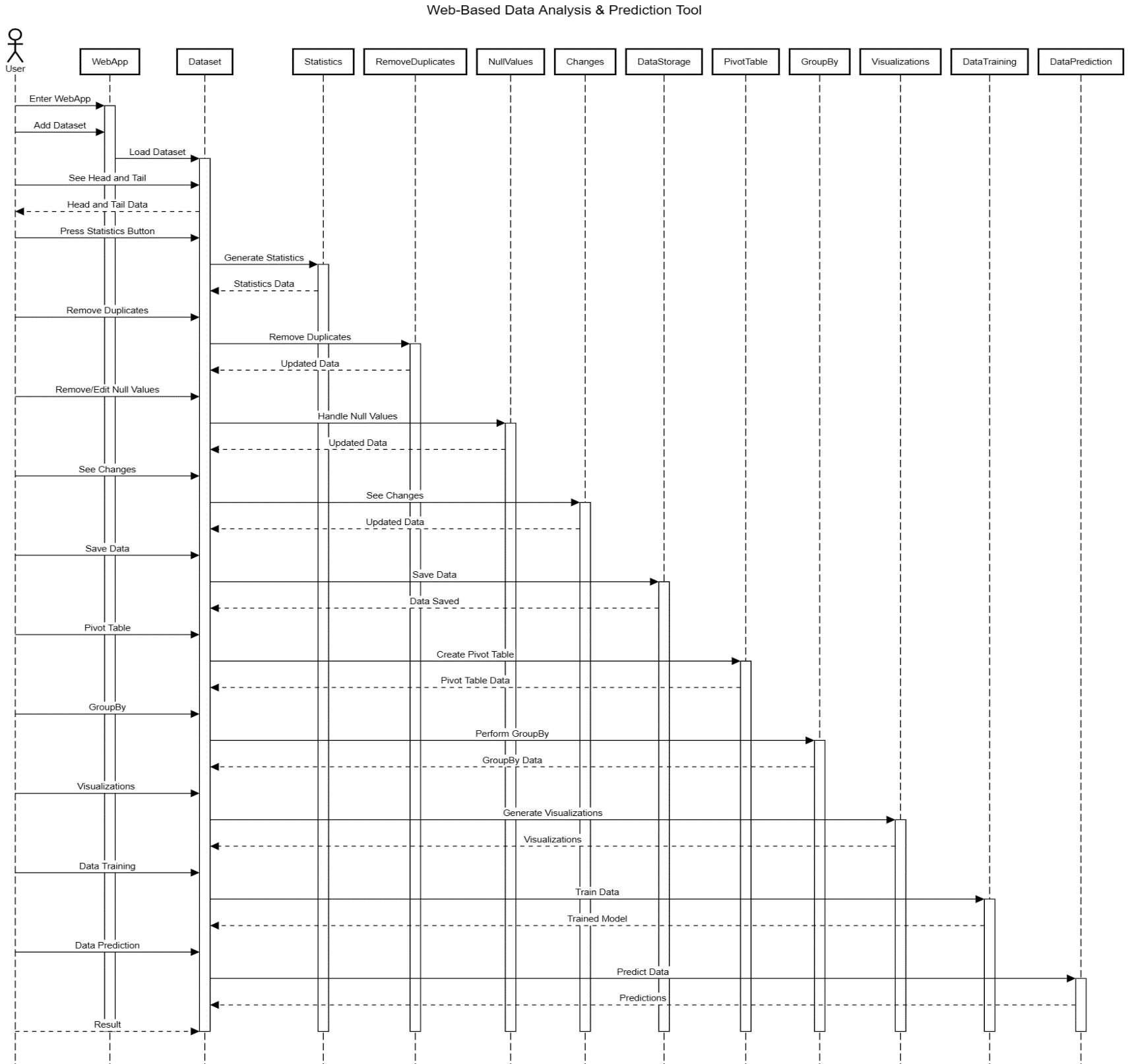
Analysis & Prediction: In the Analysis & Prediction module, the user is asked to provide their own dataset. The user can provide any dataset they wish to perform different operations on. This allows users to gain valuable insights from their own data in a secure and user-friendly way.

* 1. **System Design using UML**

In the context of my project, system design using UML (Unified Modeling Language) plays a crucial role in visualizing and communicating the structure, behavior, and interactions of the Web-Based Data Analysis & Prediction Tool. By utilizing UML diagrams, such as use case diagrams, I can effectively represent the functionalities and relationships between the User (actor) and the various components of the system. These diagrams serve as a common language for stakeholders, enabling clear communication and understanding of the system's design and functionality. The UML diagrams used for system design include:

* Sequence Diagram
* Use Case Diagram
  + 1. **Sequence Diagram**

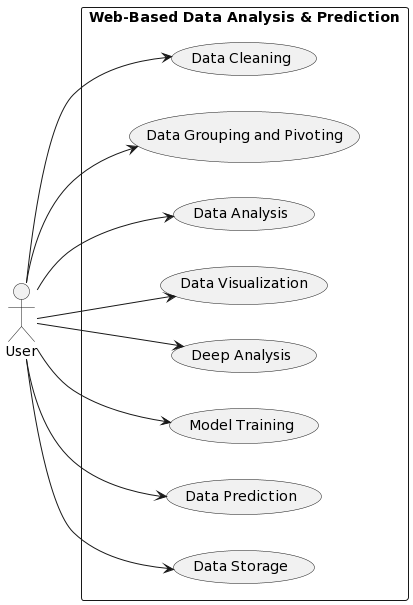
The sequence diagram illustrates the User's interaction with the "Web-Based Data Analysis & Prediction Tool." It shows the steps involved, such as data cleaning, grouping, analysis, visualization, deep analysis, prediction, and data storage. The diagram provides a visual representation of the flow of actions and information between the User and the WebApp.



**Figure 3.1 Sequence Diagram of Web-Based Data Analysis & Predication**

* + 1. **Use Case Diagram**

The use case diagram represents a user-friendly "Web-Based Data Analysis & Prediction Tool." It shows how a user interacts with different features of the system, such as cleaning data, grouping and pivoting data, analyzing it, visualizing it, performing deep analysis, making predictions, and storing data. This diagram helps visualize the main tasks performed by the system and how the user interacts with them.



**Figure 3.2 The use case diagram of Web-Based Data Analysis & Predication**

**Chapter 04**

**SYSTEM IMPLEMENTATION**

* 1. **Introduction**

Introduction Once the system has been proposed and the analysis and design phase has been completed, the designer moves towards the development phase of the software in accordance with the proposed system and design phase specifications. This involves transforming the theoretical concepts into a tangible and functional solution.

During the development phase, the designer will implement the various components and modules of the system, following the specifications outlined in the design phase. This can involve writing code to implement the various features and functionality of the system, as well as integrating it with existing systems and data sources.

* 1. **Tool / Language / Technology Selection Criteria**

The selection of appropriate tools is crucial for the successful development of the Web-Based Data Analysis & Prediction tool. In order to make an informed decision, several criteria have been considered to evaluate the suitability of the tools, including Python, Streamlit, Pandas, NumPy, Plotly, Seaborn, Matplotlib, scikit-learn, and Pandas Profiling.

The following criteria have been taken into account:

1. **Functionality:** The Streamlit framework provides a rich set of user interface controls and allows for the creation of intuitive and user-friendly interfaces. Pandas and NumPy provide powerful data manipulation capabilities, while Plotly, Seaborn, and Matplotlib offer a wide range of data visualization options. scikit-learn provides machine learning algorithms for performing predictions based on the dataset.
2. **Compatibility:** The tools should be compatible with each other and with the overall system architecture. Python is a widely used general-purpose language that is compatible with all the packages built for data analysis tasks.
3. **Scalability:** The chosen tools should support scalability to accommodate the growing needs of the Web-Based Data Analysis & Prediction tool. Python, Streamlit, Pandas, NumPy, Plotly, Seaborn, Matplotlib, scikit-learn, and Pandas Profiling provide high scalability that can boost the performance of the application.
4. **Reliability:** The tools should be reliable, ensuring the stability and consistent performance of the Web-Based Data Analysis & Prediction tool. Python, Streamlit, Pandas, NumPy, Plotly, Seaborn, Matplotlib, scikit-learn, and Pandas Profiling have proven track records and are widely used in the industry, indicating their reliability and stability. These tools have been extensively tested and refined over time, providing a solid foundation for the development of reliable and robust software systems.
5. **Community and Support:** Python, Streamlit, Pandas, NumPy, Plotly, Seaborn, Matplotlib, scikit-learn, and Pandas Profiling have active communities, extensive documentation, and readily available support channels. This facilitates the resolution of issues and promotes continuous learning for developers.
6. **Industry Adoption:** The tools should have a significant presence in the IT industry and be widely adopted. Python, Streamlit, Pandas, NumPy, Plotly, Seaborn, Matplotlib, scikit-learn, and Pandas Profiling are widely recognized and used in the IT industry, ensuring compatibility and future career prospects for developers.

By considering these criteria, we can confidently select Python, Streamlit, Pandas, NumPy, Plotly, Seaborn, Matplotlib, scikit-learn, and Pandas Profiling as the optimal tools for developing the Web-Based Data Analysis & Prediction tool. These tools offer the necessary functionality, compatibility, scalability, reliability, community support, and industry adoption to successfully realize the project goals.

* 1. **System Implementation**

During the system implementation phase, the designer will use the selected tools to develop the various components and modules of the Web-Based Data Analysis & Prediction tool. This can involve writing code to implement features such as data cleaning, grouping, pivoting, analysis, and prediction.

The designer will also need to integrate the tool with existing systems and data sources to ensure seamless operation. This can involve connecting to user databases or data storage solutions to access the data that will be analyzed by the tool.

* 1. **Front End (Streamlit: Python)**

The front end of the Smart chat and text analyzer leverages the following advantages of Streamlit 1.21.0, Python 3.10.2, HTML5 and CSS:

* + 1. **Streamlit**

The front end of the Web-Based Data Analysis & Prediction tool is developed using Streamlit, a popular Python library for building data-driven web applications. Streamlit provides a robust and comprehensive platform for web-based applications, offering a rich set of libraries and tools for efficient coding and enhanced productivity.

One of the key advantages of using Streamlit for the front end of the tool is its ability to create intuitive and user-friendly interfaces. Streamlit provides a wide range of user interface controls, such as buttons, sliders, and text inputs, that can be easily added to the application. This allows users to interact with their data and perform operations in a simple and intuitive way.

In addition to its user interface capabilities, Streamlit also ensures compatibility across different operating systems. This means that the Web-Based Data Analysis & Prediction tool can be accessed by users on a variety of devices, including desktop computers, laptops, and mobile phones.

* + 1. **Python Language**

In addition to Streamlit, the front end of the tool also leverages the power of the Python programming language. Python is known for its simplicity and readability, facilitating faster development and easier code maintenance.

One of the key advantages of using Python for the front end of the tool is its wide adoption and support by a vast developer community. This provides access to extensive resources and knowledge sharing, allowing developers to quickly resolve issues and learn new techniques.

Python also integrates seamlessly with various packages, allowing access to its extensive features and libraries. This includes powerful data manipulation capabilities provided by libraries such as Pandas and NumPy, as well as data visualization options offered by libraries such as Plotly, Seaborn, and Matplotlib.

* + 1. **Libraries**

The front end of the tool also makes use of various libraries to provide a user-friendly interface and powerful data analysis capabilities. These libraries offer a familiar development experience with Python, simplifying implementation.

One of the key advantages of using these libraries is their ability to provide simplified implementation in Visual Studio Code for rapid creation and customization of models. This allows developers to quickly prototype and test new features and functionality for the Web-Based Data Analysis & Prediction tool.

The libraries used in the front end of the tool also utilize an event-driven programming model, enabling easy handling of user interactions and system events. This means that when a user interacts with the application, such as by clicking a button or moving a slider, the appropriate code is automatically executed to handle that event.

In addition to their event-handling capabilities, these libraries also support robust data binding capabilities. This simplifies data display and synchronization with data sources, allowing users to easily view and interact with their data.

* 1. **IDE (Visual Studio Code)**

Advantages of Visual Studio Code for Web Application Development:

1. **Integrated Development Environment (IDE):** Visual Studio Code provides a comprehensive IDE with tools and features specifically designed for Python development, Model creation and other application development.
2. **Python Language Support:** Visual Studio Code offers robust support for the Python programming language, providing features like code completion, debugging, and IntelliSense for efficient and error-free coding.
3. **Python Packages Compatibility:** Visual Studio Code provides native support for third party python packages compatibility, allowing developers to manage data flow between these libraries and other modules.
4. **Debugging and Testing Tools:** The IDE includes powerful debugging and testing tools, enabling developers to quickly identify and fix issues in their Python code, and related other libraries.
5. **Enhanced Productivity Features:** Visual Studio Code offers numerous productivity-enhancing features such as code refactoring, code snippets, templates, and shortcuts, speeding up development tasks and improving overall efficiency.
6. **Collaboration and Version Control:** The IDE supports collaboration through features like Live Share, enabling real-time collaboration between team members working on Python Web Application, Python Desktop Application, and SQL Server projects. It also integrates with version control systems like Git for efficient code management.
7. **Publishing and Deployment Options:** Visual Studio Code provides flexible options for publishing and deploying applications developed with Python ensuring smooth deployment on various platforms.

These advantages collectively make Visual Studio Code an excellent choice for application development with Python by using different packages offering a comprehensive development environment, seamless integration, enhanced productivity, collaboration features, and efficient deployment options.

* 1. **Operating system Selection**

For this application, the operating system selection was not a challenging task. It was crucial to choose an operating system that supports the software requirements and is easily accessible to users. Windows 8 and Windows 10 were deemed suitable options, and **Windows 8.1 Pro** was selected as the operating system for the application.

* 1. **Streamlit**

Python is a modern, object-oriented programming language designed by Guido van Rossum to create desktop applications, Web applications, Android application. The graphical user interface for the application is developed using Streamlit which is a Framework developed for python. The advantages of using Streamlit for our application are:

* + 1. **Advantages of Streamlit**

1. **Easy to use:** Streamlit framework syntax is in python, making it easy for developers who have experience with these languages to pick it up quickly.

1. **Flexibility:** Python and Streamlit offer great flexibility during and after the initial project. This is important since the requirements of the project often change during its lifetime. With python and Streamlit, we can easily make changes to the application without losing valuable time.
2. **Platform Independent:** Our application will run smoothly on all browsers, and operating system providing a seamless experience to users.
   1. **Implementation**

The implementation of the “Web-Based Data Analysis & Prediction Tool” was carried out using the chosen programming language and framework for the development of the web-based application. The project was divided into two main modules: the Training module and the Analysis & Prediction module.

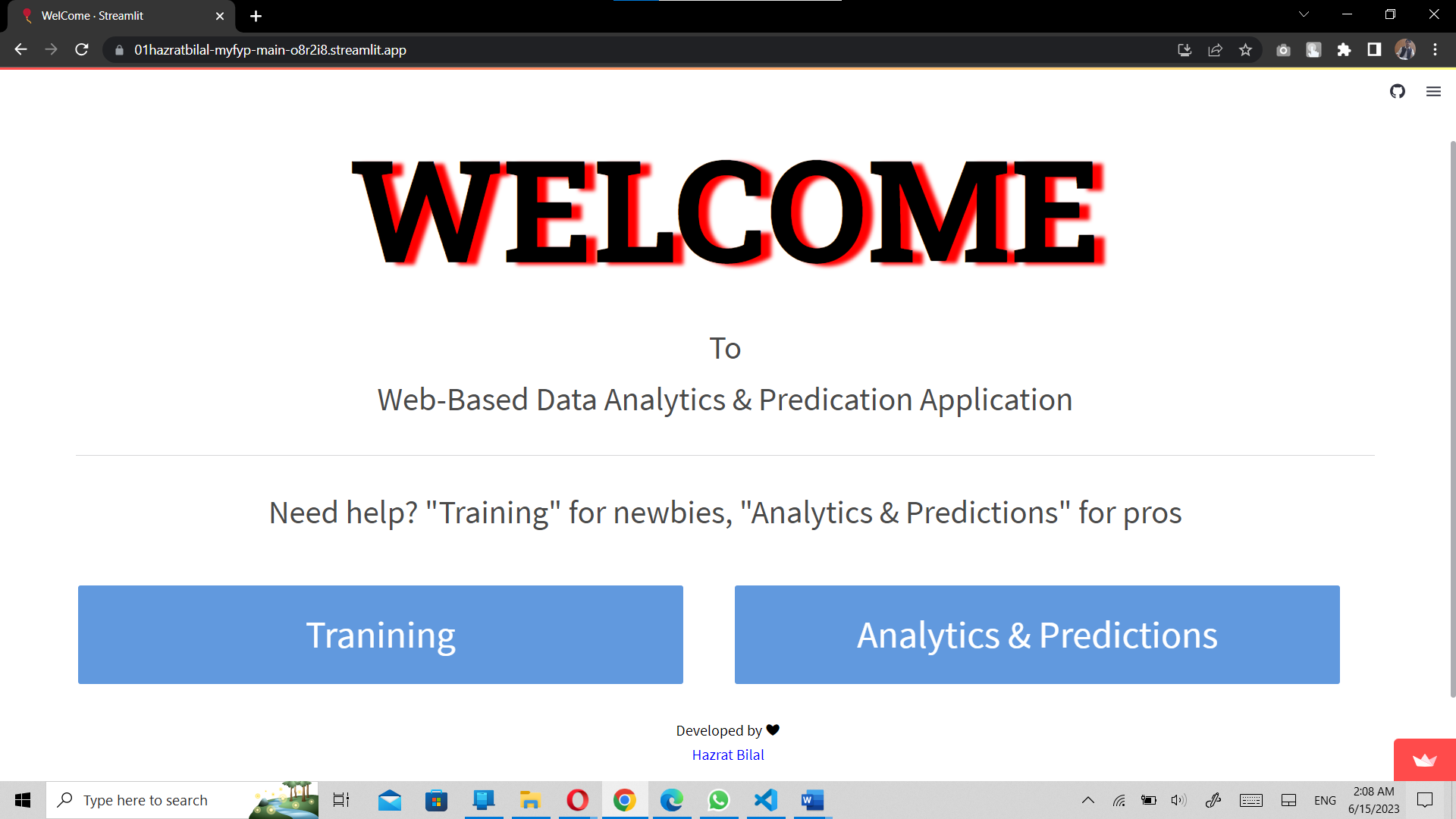
* The **Training** module uses the famous “Titanic” dataset to introduce users to different data types and show them how to handle and work with them. This module includes features such as data cleaning, grouping, pivoting, analysis, and visualization with instructions. Users can learn how to clean their data by removing any inconsistencies in their dataset, group and pivot their data to organize it in a way that makes it easier to analyze, and deep analysis to gain valuable insights from the data. The module also includes features for visualizing data in 1D, 2D, and 3D, allowing users to easily see patterns and trends in their data.
* The **Analysis & Prediction** module allows users to provide their own dataset for analysis. This module includes features such as data cleaning, grouping, pivoting, analysis, visualization, and prediction, allowing users to gain valuable insights from their own data in a secure and user-friendly way. Users can clean their data by removing any inconsistencies in their dataset, group and pivot their data to organize it in a way that makes it easier to analyze, and perform deep analysis to gain valuable insights from the data. The module also includes features for visualizing data in 1D, 2D, and 3D, allowing users to easily see patterns and trends in their data, as well as perform predictions based on the dataset.

**Chapter 05**

**USER GUIDE**

* 1. **Main Page**

Upon opening the Web-Based Data Analysis & Prediction Application by URL <https://01hazratbilal-myfyp-main-o8r2i8.streamlit.app/> or simply install the .exe file, the user is greeted with the Main Page, which warmly welcomes them to the application. The Main Page provides instructions for the user, directing them to either the Training page if they are unfamiliar with the usage of the application or to the Analysis & Prediction page if they already have information about the application and their data.

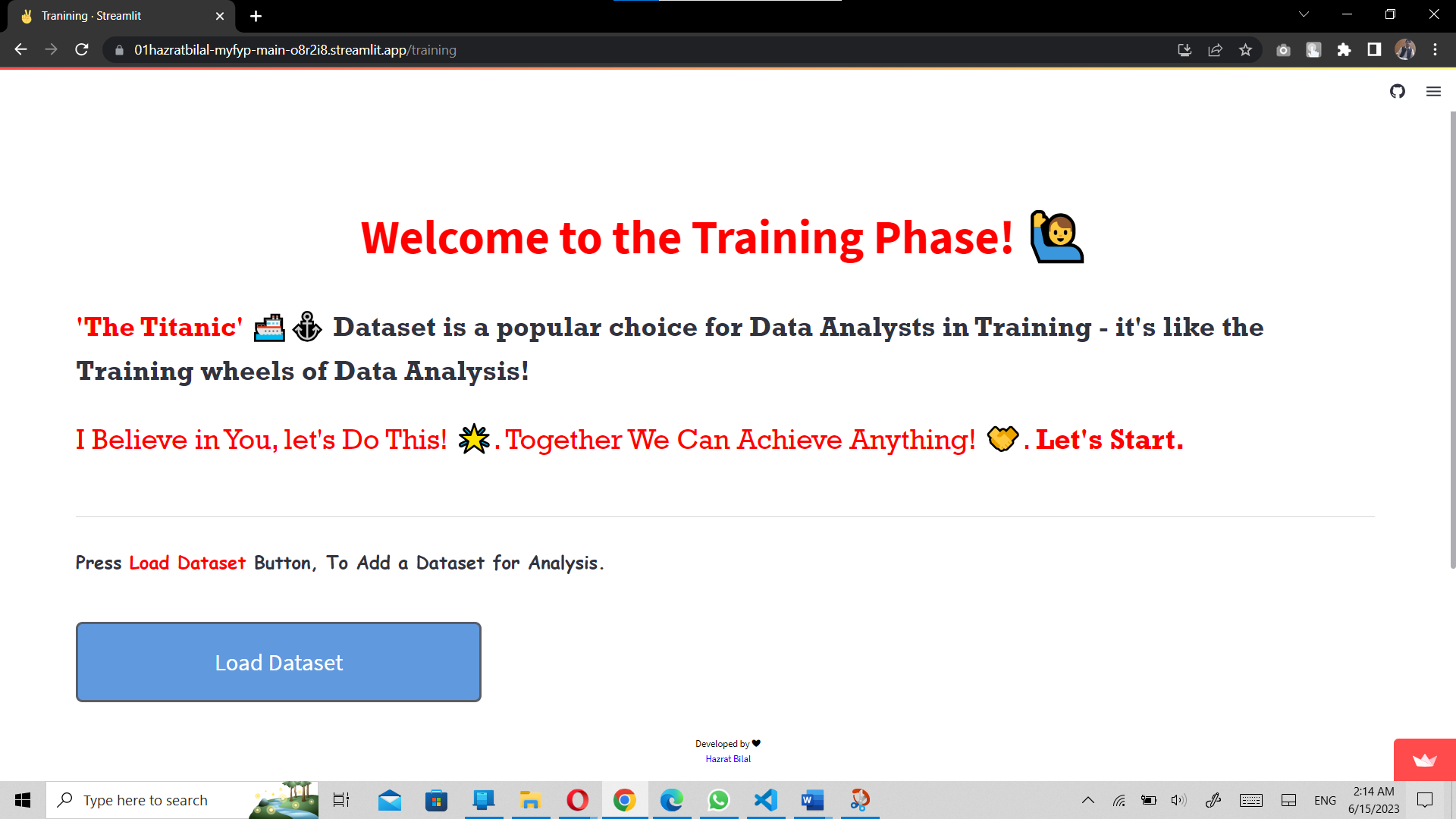
****

**Figure 5.1 Main Page**

* 1. **Training**

This module is designed to train the user for better understanding of their data, operations on data, and visualizations. It includes step-by-step instructions for performing operations on the data before the actual analysis takes place. This helps users become more familiar with their data and the tools available for analyzing it.

When the user clicks on the Training button, they are taken to a new interface. This page provides an introduction to the various features and functionalities of the Web-Based Data Analysis & Prediction Tool, using the famous “Titanic” dataset to demonstrate how to handle and work with different data types.



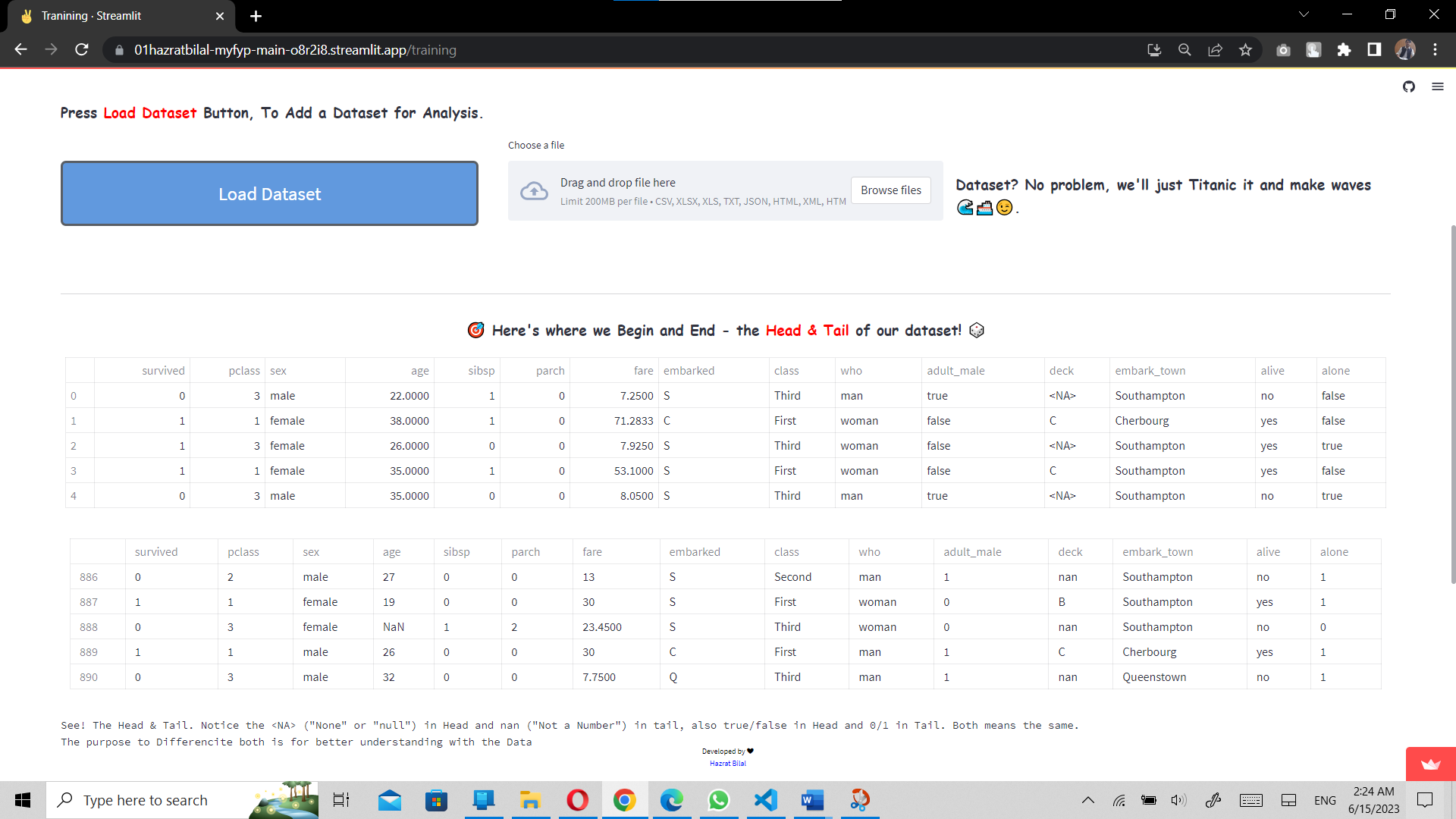
**Figure 5.2 Training Phase Beginning**

* + 1. **Load Datset Button**

The user is prompted to press the Load Dataset button to add a dataset for analysis, but for the Training module, the designer has provided the famous “Titanic” dataset for demonstration purposes.

* + 1. **Head & Tail**

After pressing the “Load Dataset” button, the user can see the head and the tail of the “Titanic” dataset and some bread information about the head and tail.



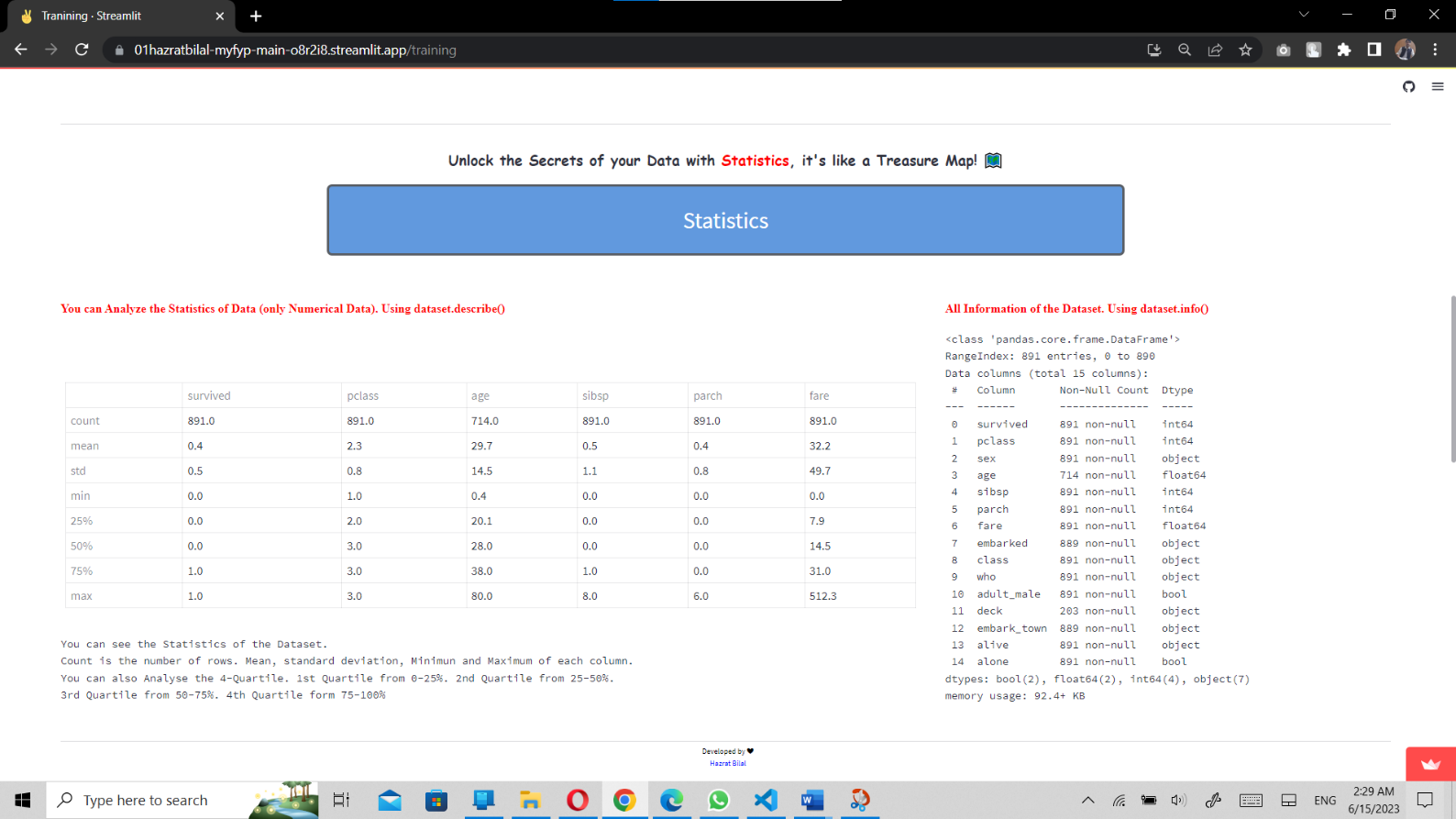
**Figure 5.3 Load Dataset Button**

* + 1. **Statistics Button**

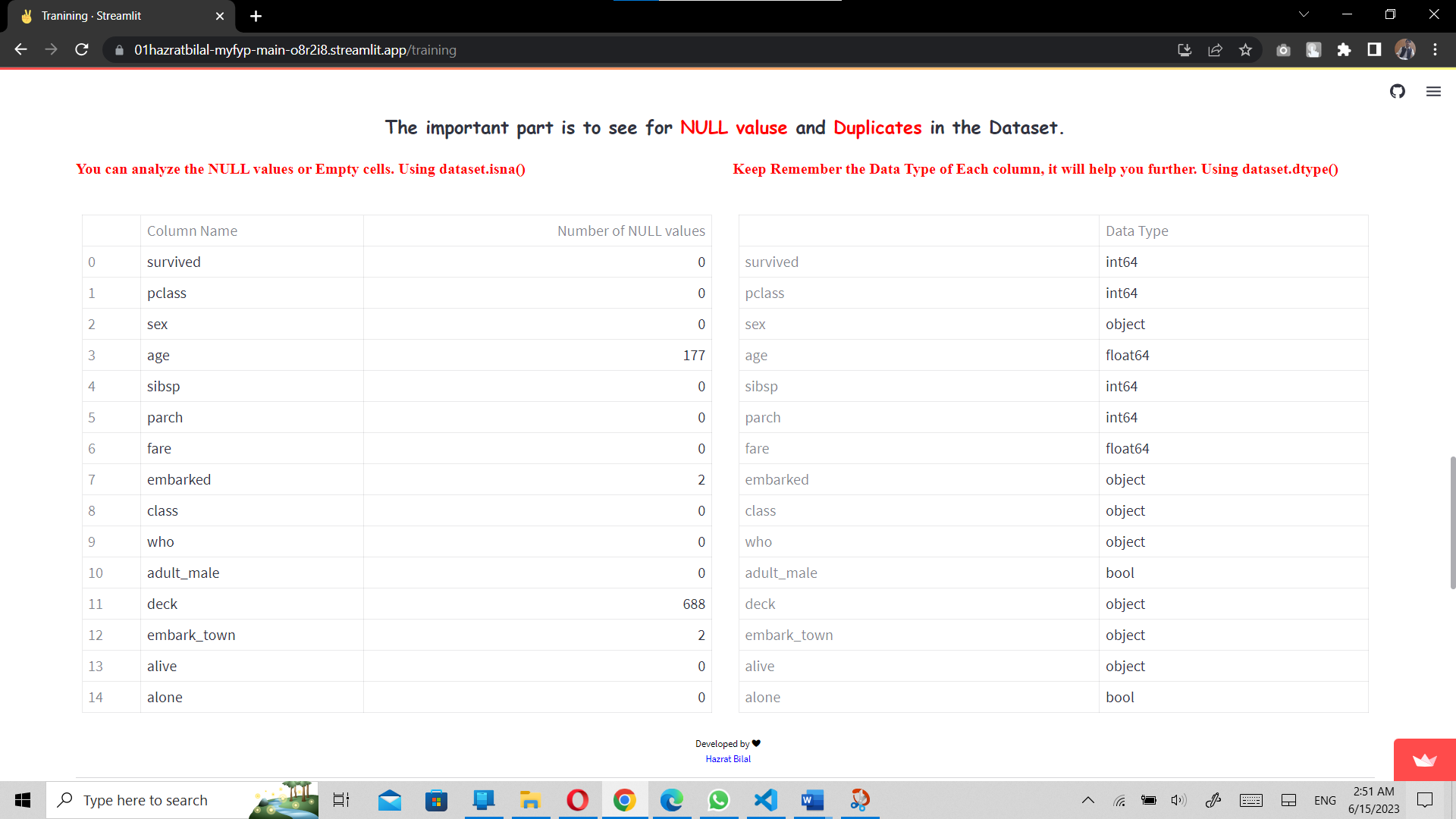
Just after the head & tail of the dataset, a button appears by the name of “Statistics”. The user is guided to know the statistics of the dataset.

Statistical analysis includes information about:

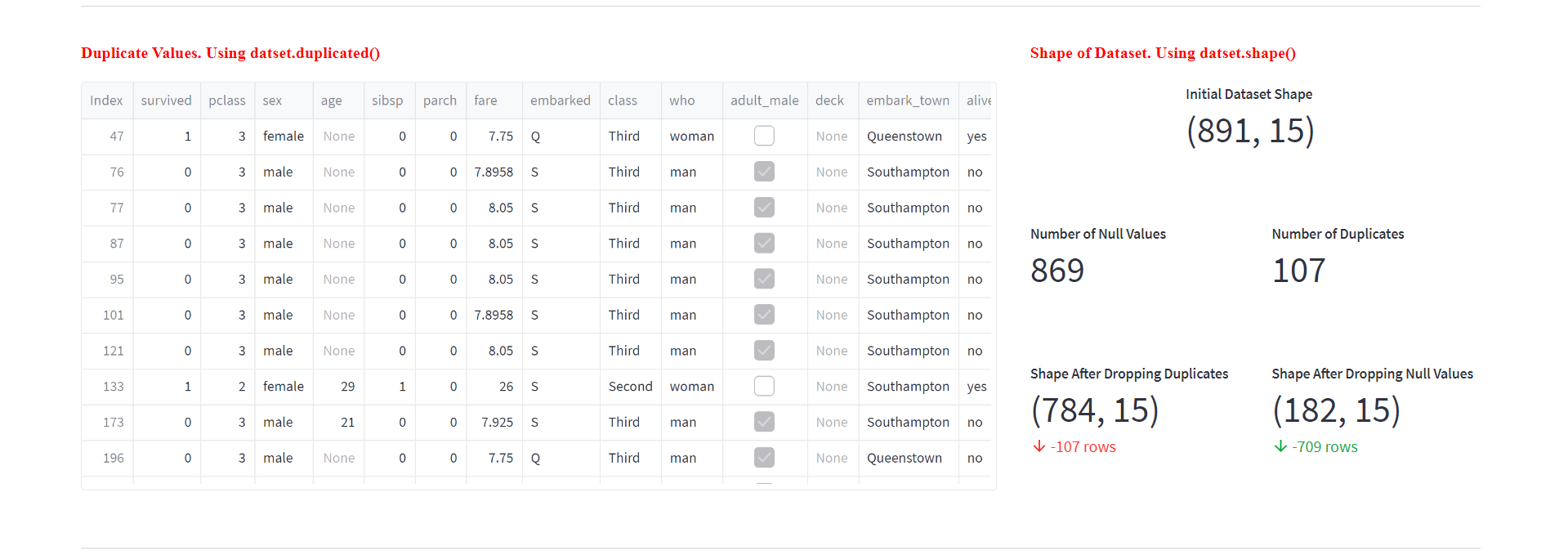
1. The statistics of the dataset (only numerical values) that includes the number of rows in each column, mean, standard deviation, minimum value, minimum value and the information about all 4 quartiles.
2. Deep information about the data set which tells the number of all the columns in the dataset, the number of non-null rows, the data type of each column.



**Figure 5.4 Information of dataset.**

1. ****The number of null values in each column.

**Figure 5.5 Information about the Null Values**

1. Shows the Duplicate values in the dataset.
2. ****A Metrix that informs about the shape of the dataset. The number of Duplicates and Null values in the dataset. This Metrix also inform about the number of rows when null values and duplicates are drop out of the dataset.

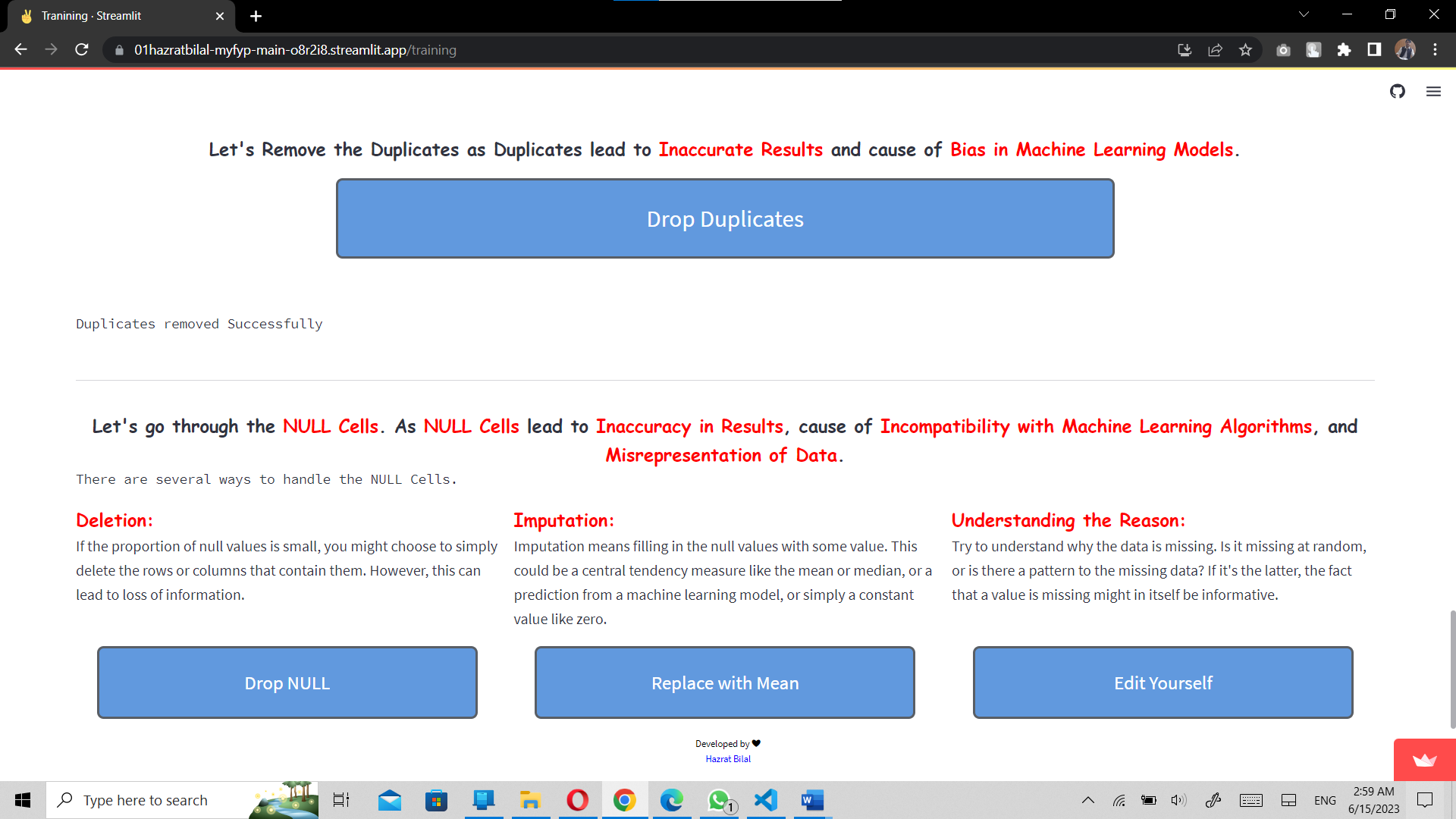
**Figure 5.5 Shape of the dataset and Duplicates in Dataset**

* + 1. **Remove Duplicates Button**

The user is informed about the duplicates in the dataset and their impact. The user is asked to press this button to remove the duplicates and go further.

The duplicates are removed from the dataset and other section is visible now to the user.

* + 1. **Handle the Null values**

****This session is designed to inform the user about the disadvantages of null values in the dataset.

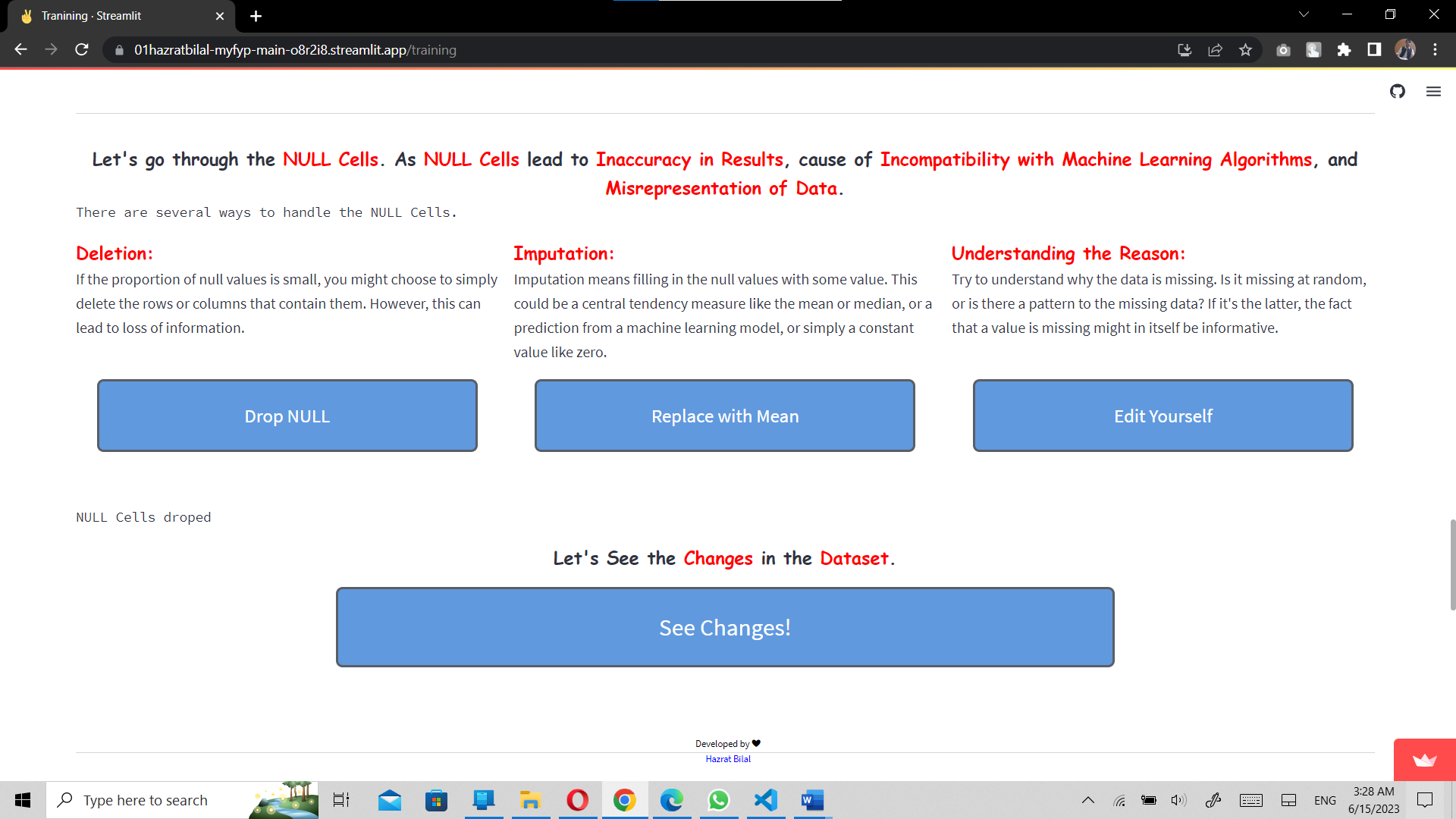
**Figure 5.6 Duplicates Dropped and the ways to handle the Null values**

* + - 1. **Ways to handle Null values**

The user is guided with three main ways to handle the noisy data.

1. **Delectation**

By using this method, all the null values will be removed from the dataset. But this method causes to reduce the number of rows in the dataset.

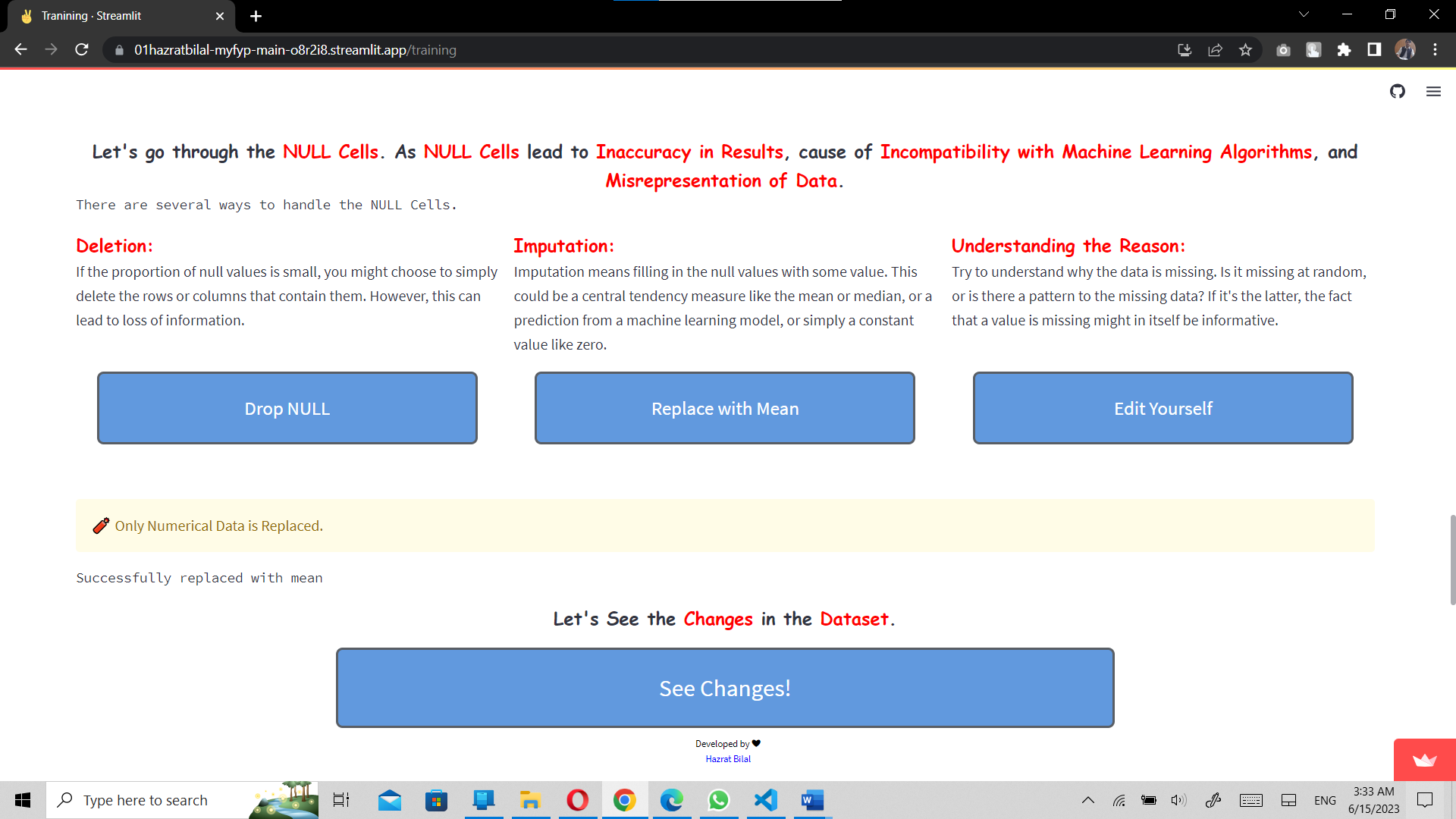
“Drop NULL” button is provided to the user to drop all the null values. A message “NULL Cells droped” is shown to the user.

**Figure 5.7 Dropping Null values**

1. **Imputation**

This method involves replacing null cells with the mean of the column they belong to. This approach is suitable for datasets with only numerical values, as it only fills null values in numerical columns with the mean of that column.

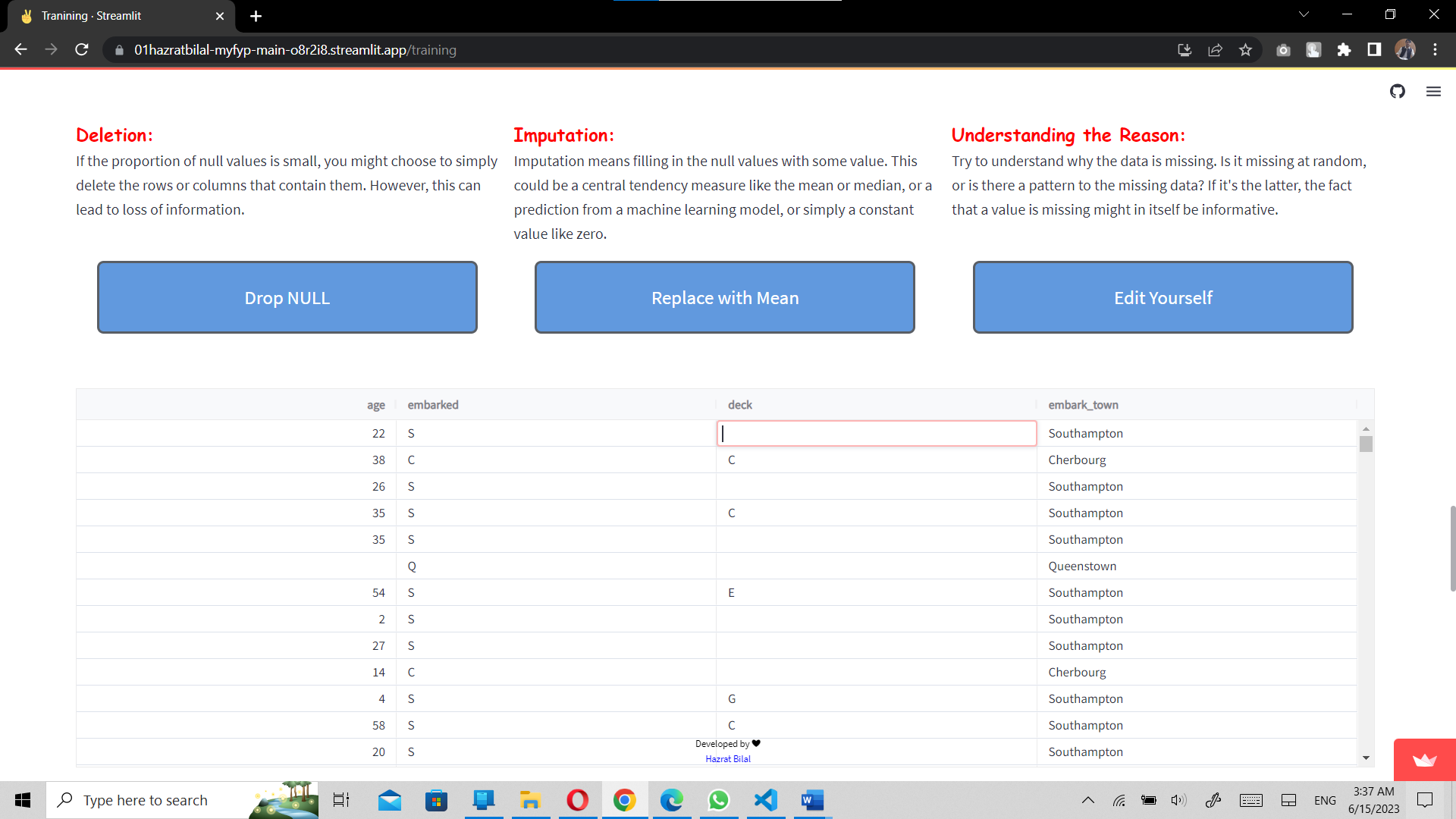
The use is allowing to replace the null values with mean by pressing the “Replace with Mean” button. A message “🧨 Only Numerical Data is Replaced.” and “Successfully replaced with mean”.



**Figure 5.8 Replace Null values with Mean**

1. **Understanding the Reason**

This method allows the user to understand the dataset and full the null values. This is one of the best approaches but it takes time to fill if the number of null cells is large.

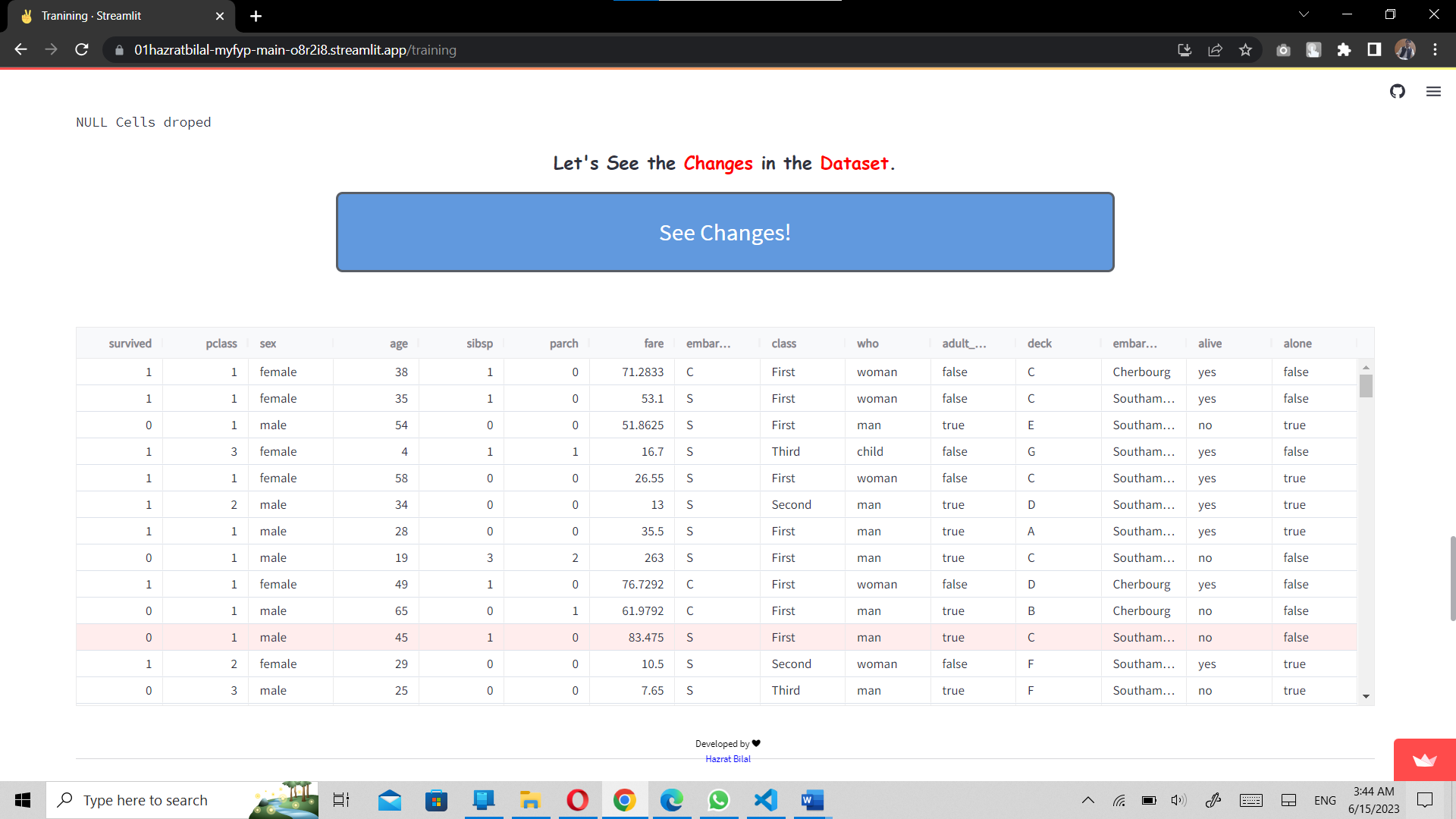
****After pressing “Edit Yourself” button an editable table is shown to the user to full the null values after understand the dataset.

**Figure 5.9 Filling Null values by the user**

* + 1. **See Changes Button**

When user Drop the nulls, replace it with mean or edit yourself. This button shows the changes made by the user. The user is also allowed to go back and change the decision of choosing approach to handle the null values.

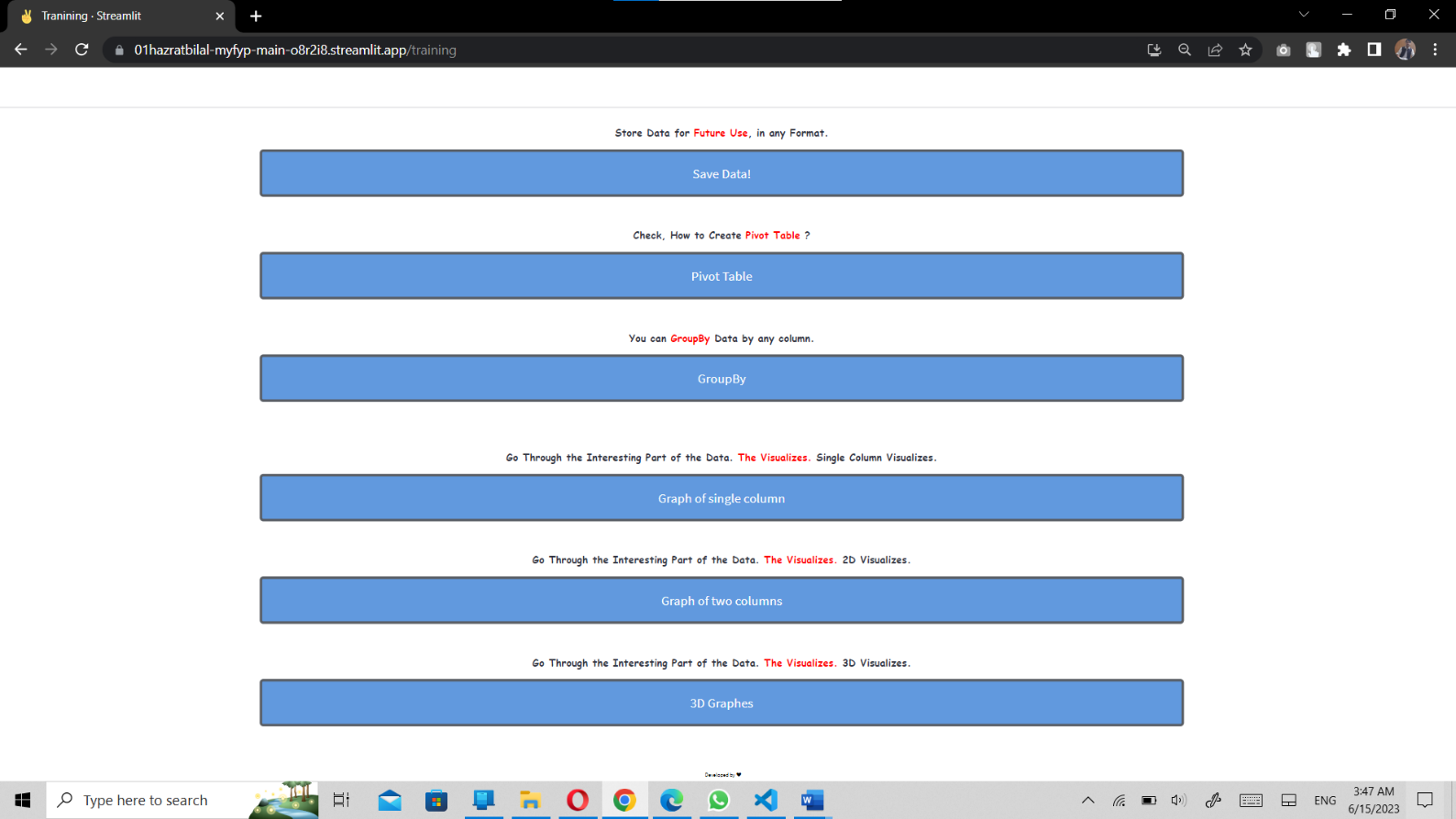
The user once again can analyze the whole dataset here in this section.



**Figure 5.10 Analyzing the dataset after handling with Null values**

* + 1. **Data Operations**

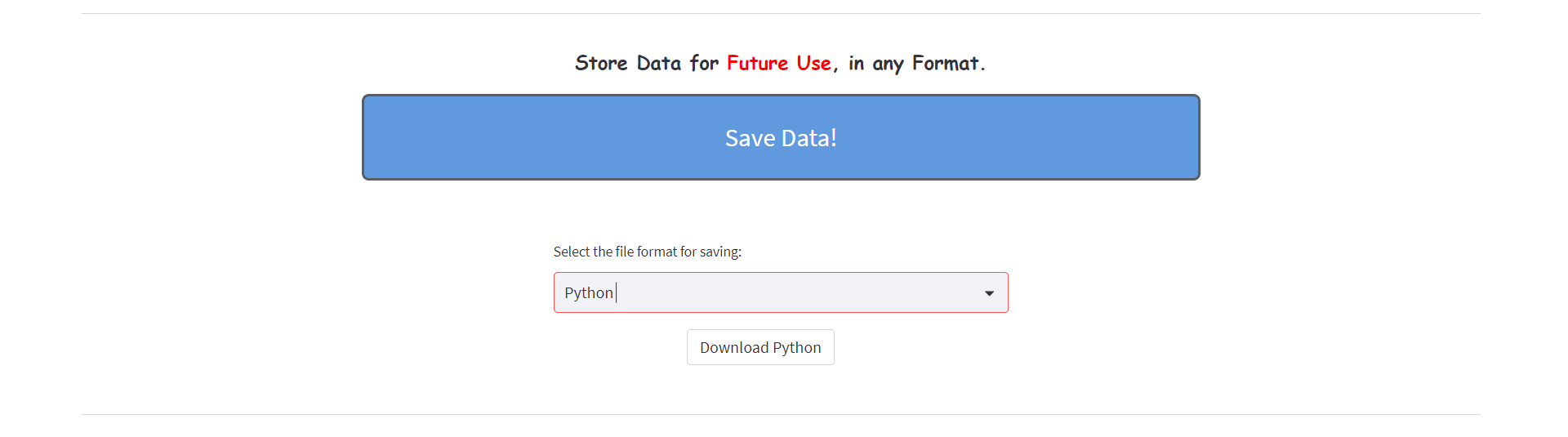
After cleaning the dataset or pre-processing, the user is presented with a range of optional operations that can be performed on the data. These include saving the data in different formats, creating pivot tables and saving them in different formats, grouping the data and saving it, and visualizing the data in 1D (single column), 2D (two columns at a time), and 3D (three columns at a time). These operations allow the user to further manipulate and analyze their data to gain valuable insights.



**Figure 5.11 Other Optional buttons for the Training**

1. **Save Button**

This button allows the user to save the cleaned dataset in comma separated values (csv), excel, text, python and pickle format.

****

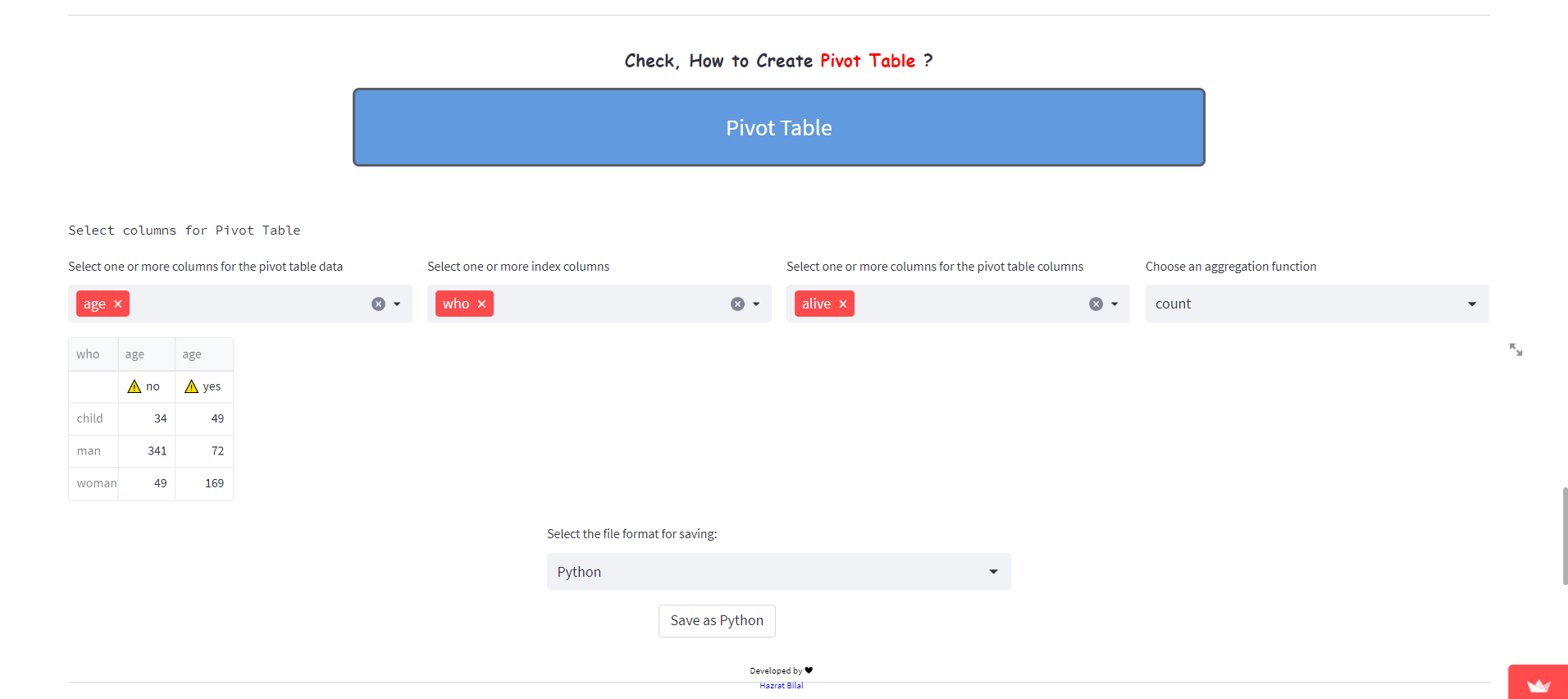
**Figure 5.12 Save Data button View**

1. **Pivot Table Button**

A pivot table is a tool that summarizes and reorganizes data from a larger table, such as a spreadsheet or a database, by grouping values and applying statistics, such as sums, averages, or ranges. A pivot table does not change the original data, but allows you to view it from different perspectives and explore trends. Pivot tables are useful for analyzing large amounts of data quickly and easily.

Pivot tables can be used for a variety of purposes, including calculating sums or averages, showing totals as a percentage of a whole, generating a list of unique values, creating a summary of a complex report, identifying the maximum and minimum values of a dataset, and querying information directly from an online analytical processing (OLAP) server.

The user is able to create a pivot table of the dataset just by clicking the pivot table button and provide the columns according to his/her will. Beside this, the user can also select pivot table by mean, count, and sum as aggregate function. The user is also able to save the pivot table in his own storage in different formats.

****

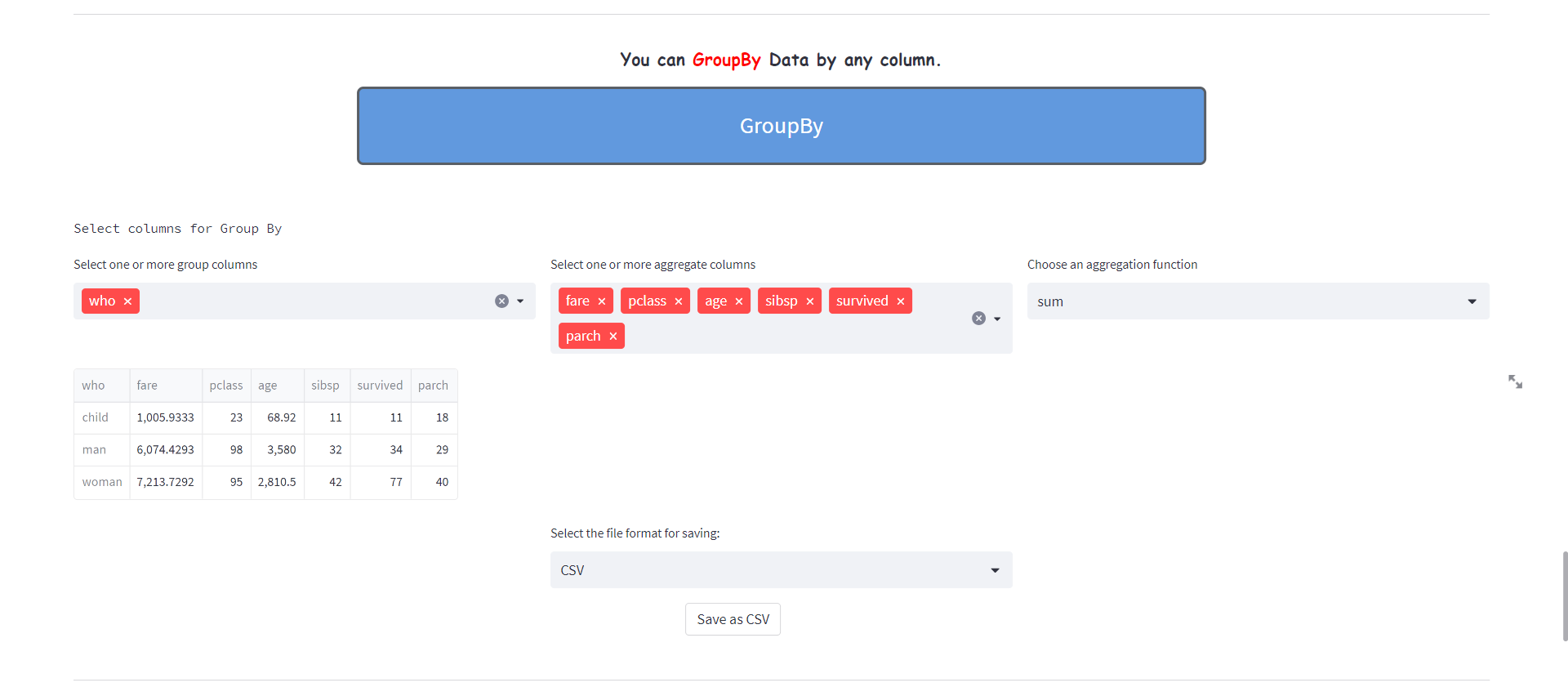
**Figure 5.13 Session of Pivot Table and to save that table**

1. **GroupBy Button**

GroupBy is a process that involves splitting data into groups based on some criteria, applying a function to each group independently, and then combining the results into a data structure. This can be used to group large amounts of data and compute operations on these groups. In Pandas (python), the groupby() function can be used to split data into groups based on some criteria, apply a function to each group independently, and then combine the results.

GroupBy can be used for a variety of purposes, including calculating sums or averages, identifying maximum and minimum values of a dataset, and generating a list of unique values.

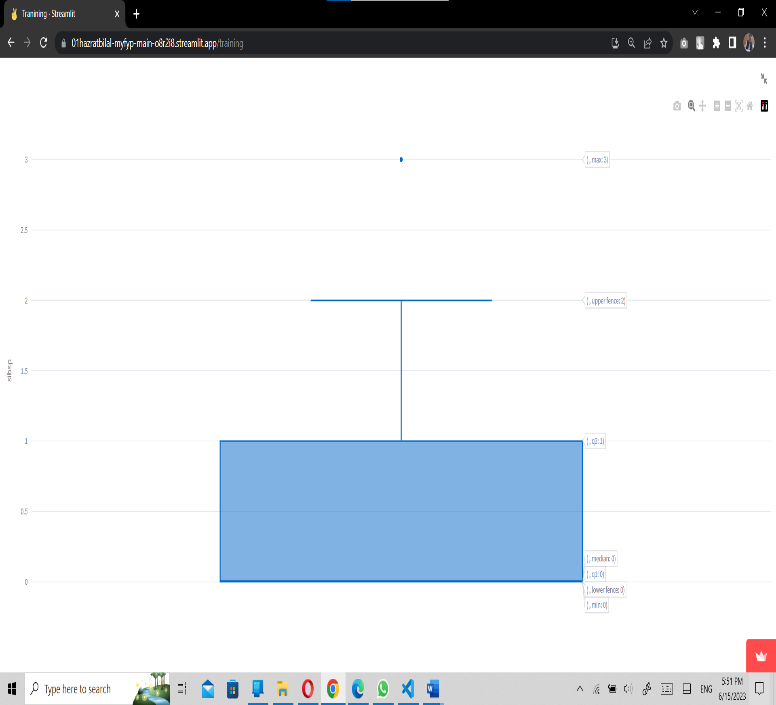
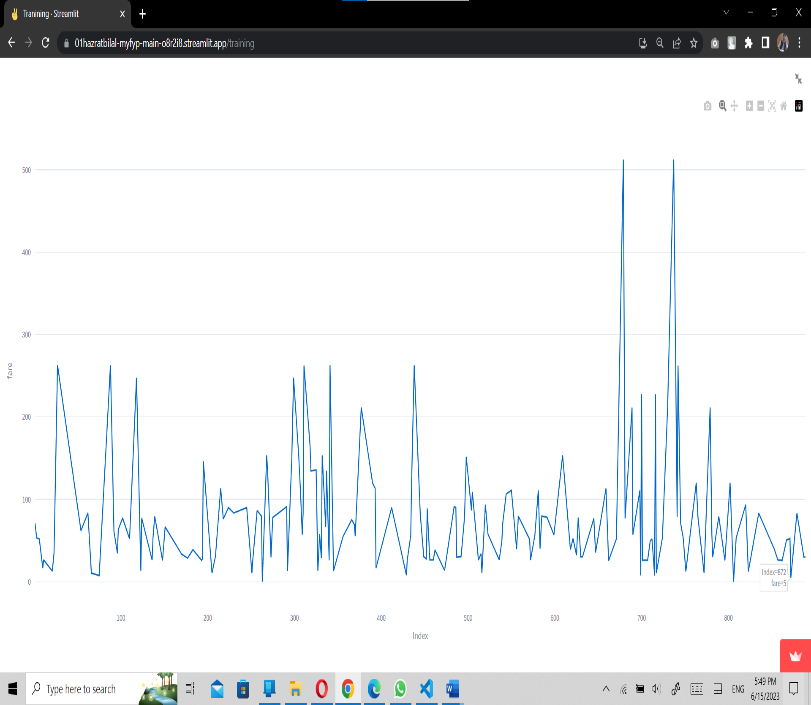
By pressing this button, the user can group his data by a single column or multiple columns based on the count, minimum, minimum, sun, and mean aggregate function. The user is all able to save the groupby table in his/her secondary storage.



**Figure 5.14 Group By the data and options to save the table**

1. **Graph of Single Column Button**

This section is able to draw an interactive graph of single column of the dataset. These graphs include Histogram Plot, Line Graph, Pie Chart, and box plot depending on the data type of the column.

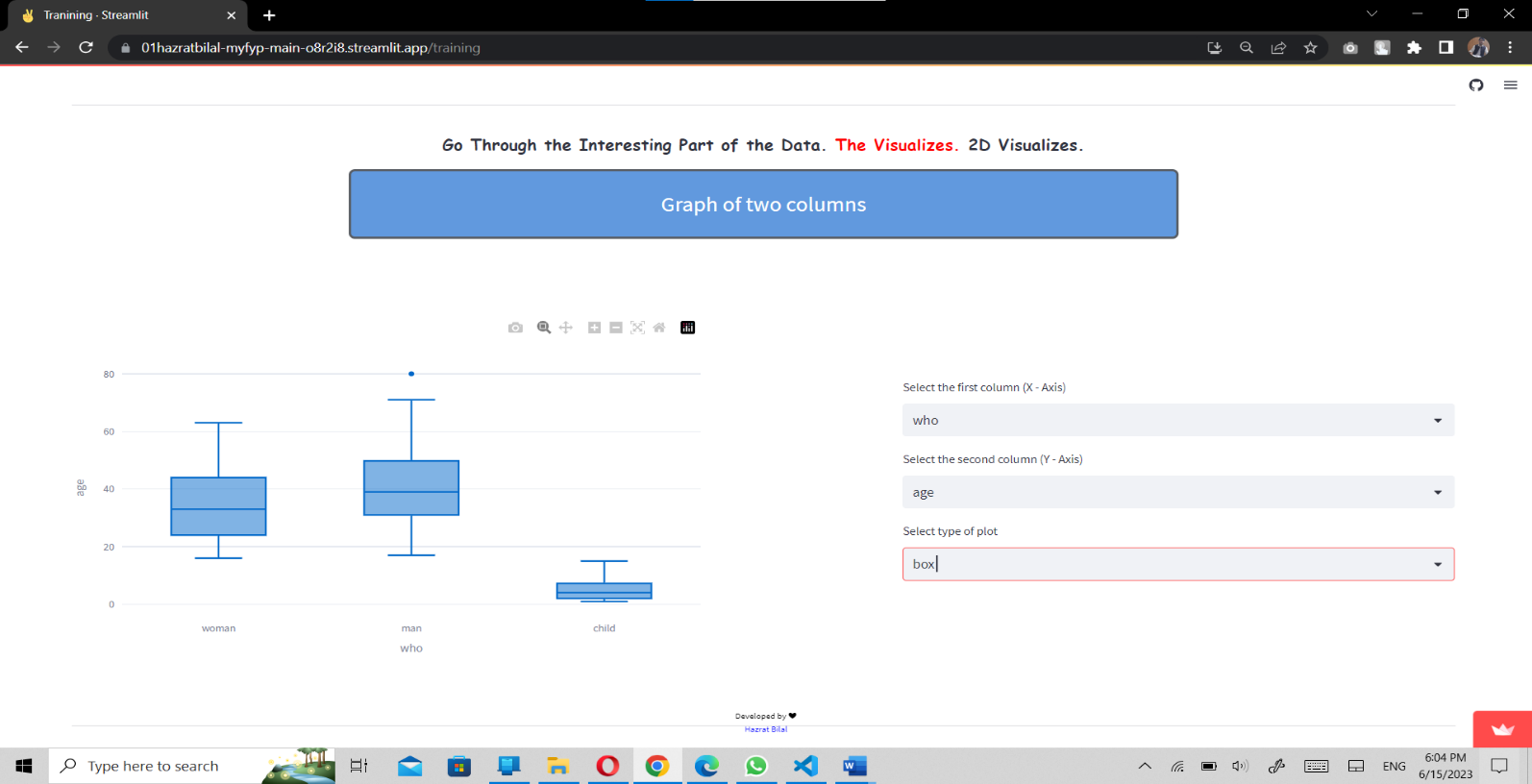


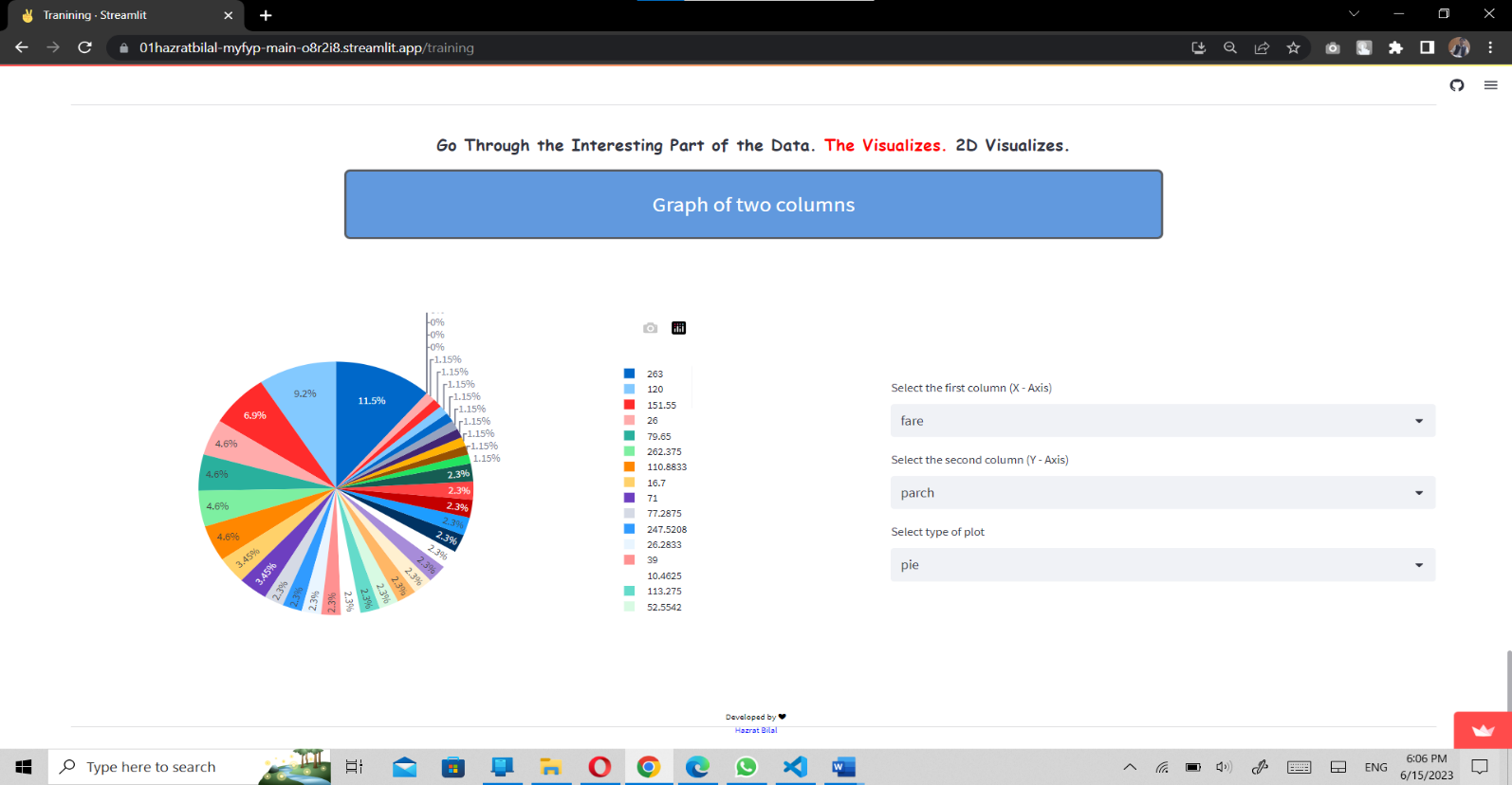
**Figure 5.15 Single Colum Charts and Graphs including Histogram, Pie chart, Box plot, and Line chart**

The Graphs and Chats are full interactive. The user only needs to hover to see the information.

1. **Graphes of Two Columns Button**

The user can analyze two columns at a time and specify them in to X and Y axis to have better understand with the patterns inside the data to gain informed Decisions from the data.

This part includes important charts and plots such as Scatter plot, Bar chart, Box plot, area plot, and Pie chart (not applicable on all data types). User can select any one based on his/her requirements.

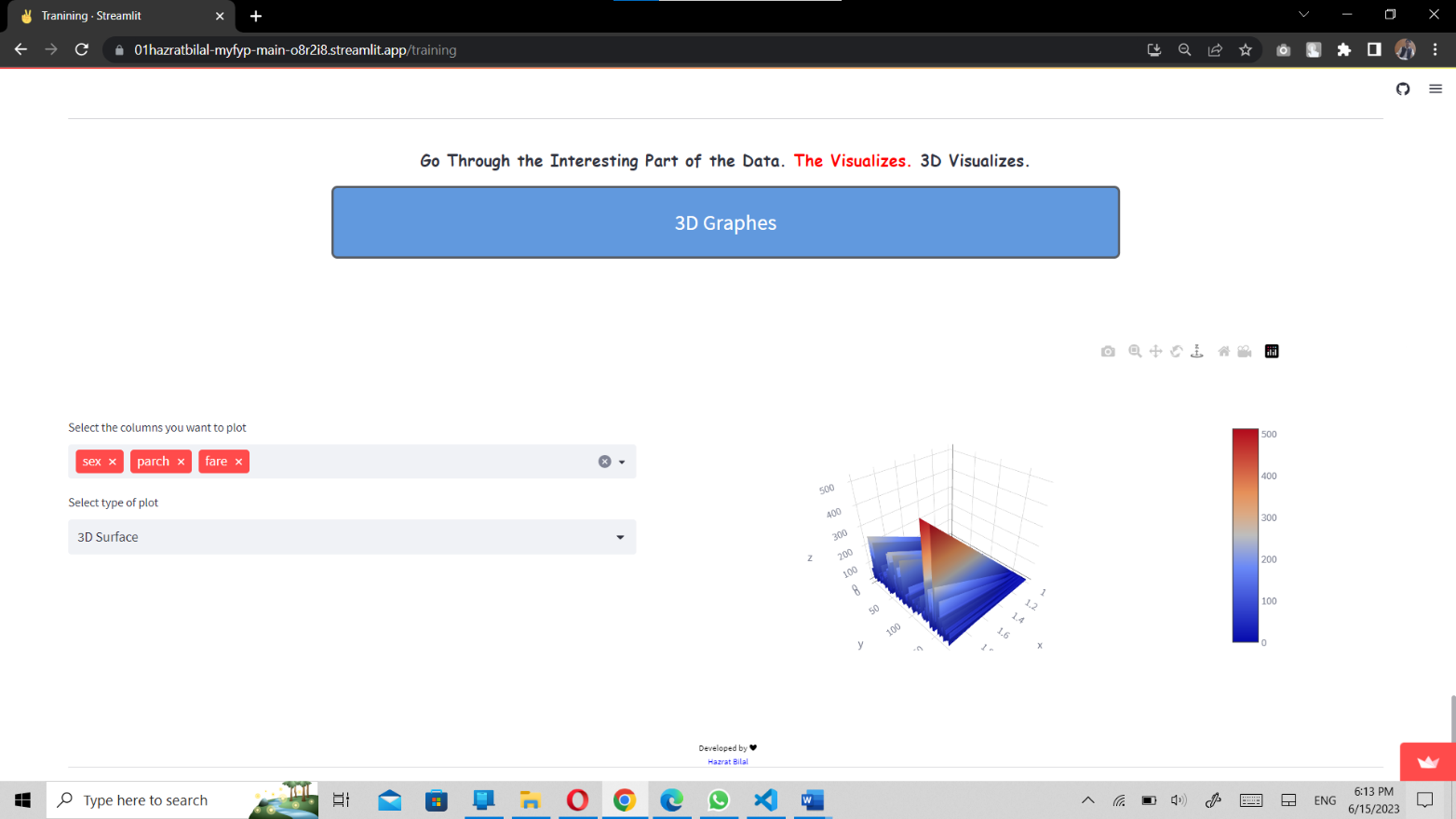
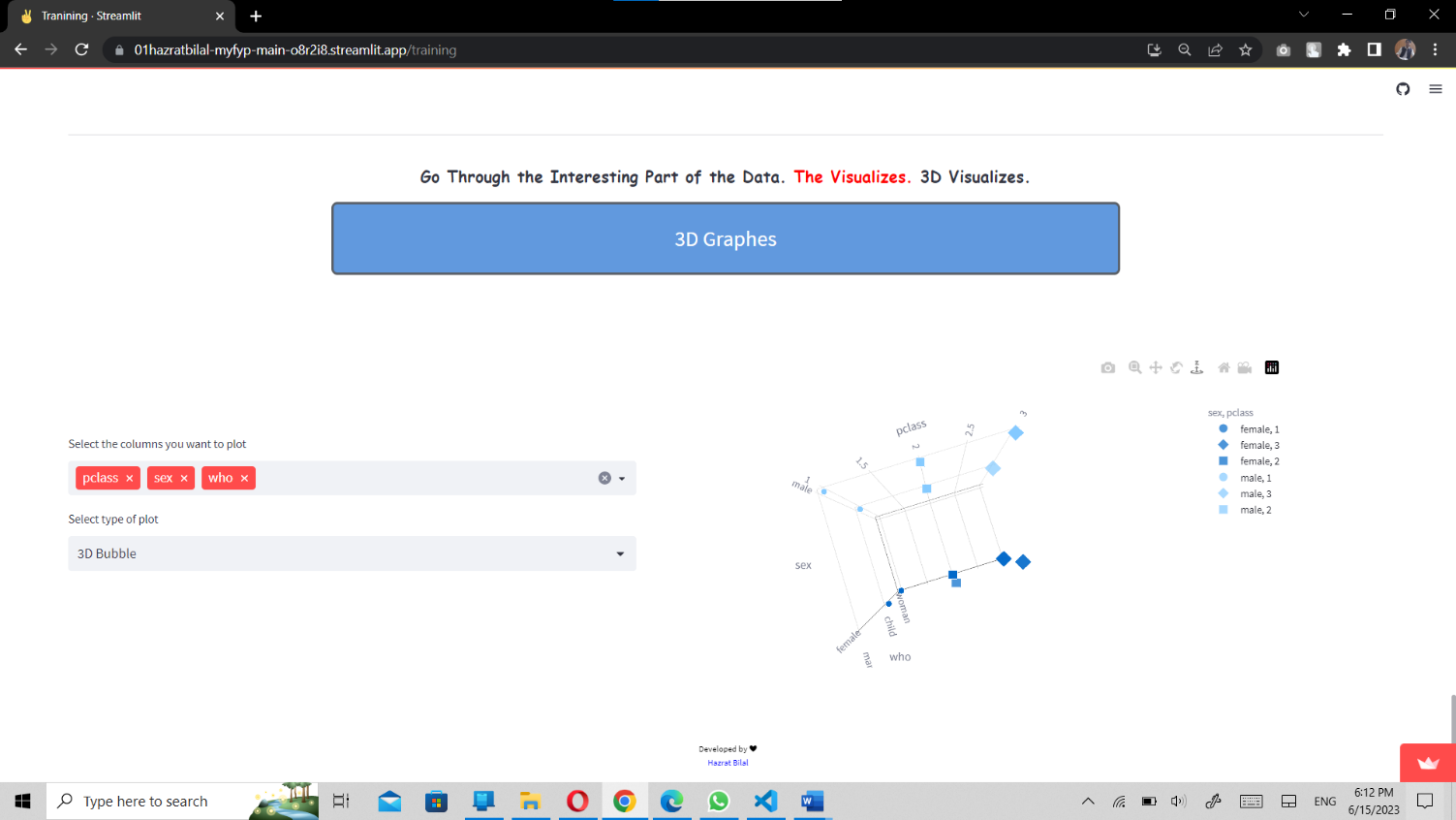


**Figure 5.15 Two columns Visuals**

1. **3D Graphes Button**

Here are the most advanced graphes that provide 3D visualizations of three columns at a time. They are fully interactive, can be zoom in, zoom out, can see the other sides and points deeply.

These graphes in include 3D Line, 3D Surface, 3D Scatter, 3D Bubble, and 3D Cone charts. These charts need good knowledge and experience to be understood.



**Figure 5.16 3D Visualization**

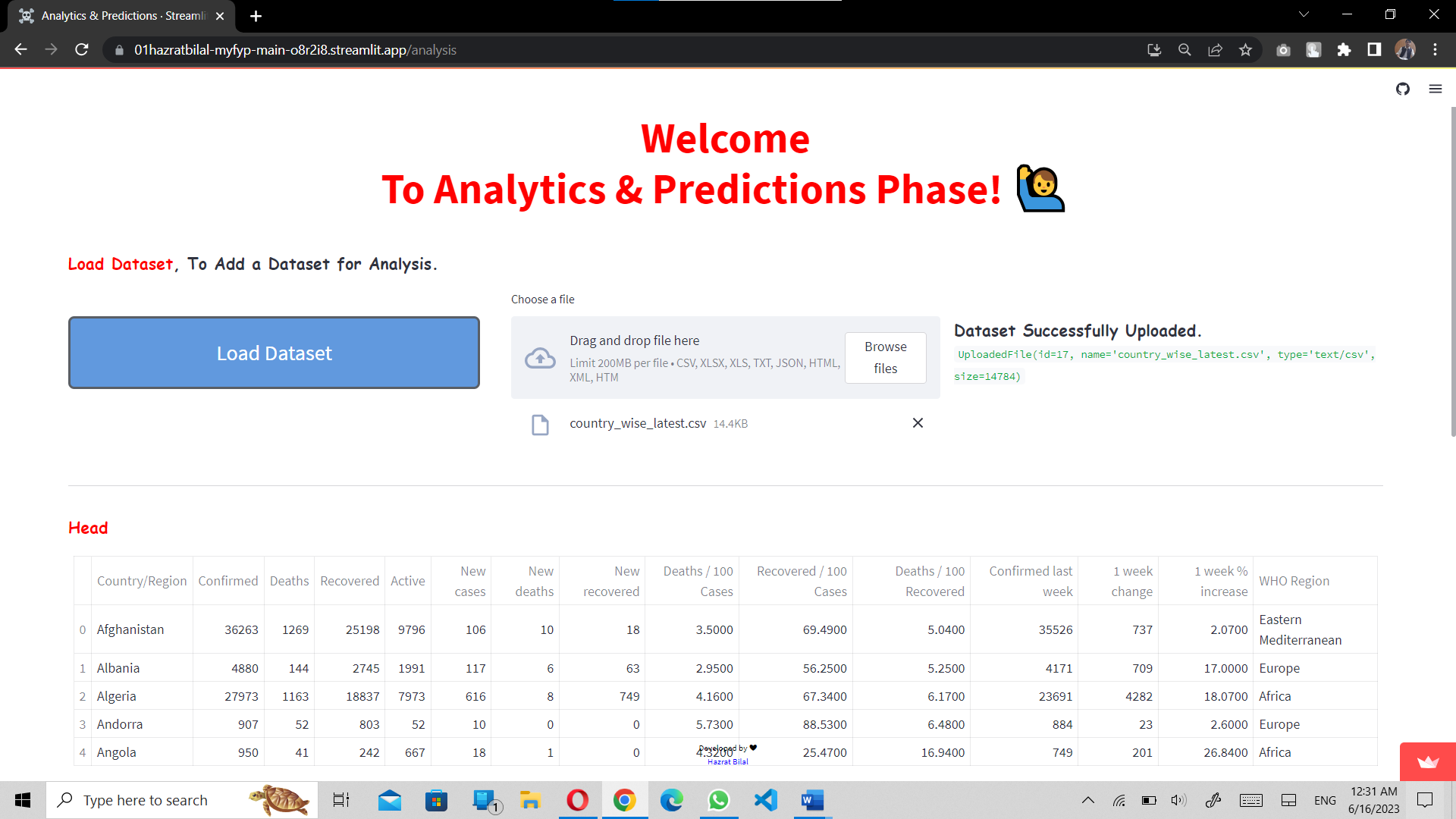
* 1. **Analysis & Prediction**

The actual analysis and the main page of the project is Analysis & Predication. If the user already about data and the basics of the Data Analysis, he/she can directly land upon this page by pressing **Analysis & Predication** button. This allows any user to upload his/her own dataset and perform operations and analysis of the data set. Most of the buttons are same as the Training but everything will be different with another dataset.

Let’s upload a dataset:

* + 1. **Upload Dataset Button**

This is the first button to make the journey to successful analysis. After pressing this button, the upload section can be seen. Where user can upload a Dataset with different extension. If data is in the form of excel, csv, text, html, and json under 200 MB which can be quite large file.

After uploading the file, the user will have some brief information about the dataset. Like the size, type and id of the dataset.

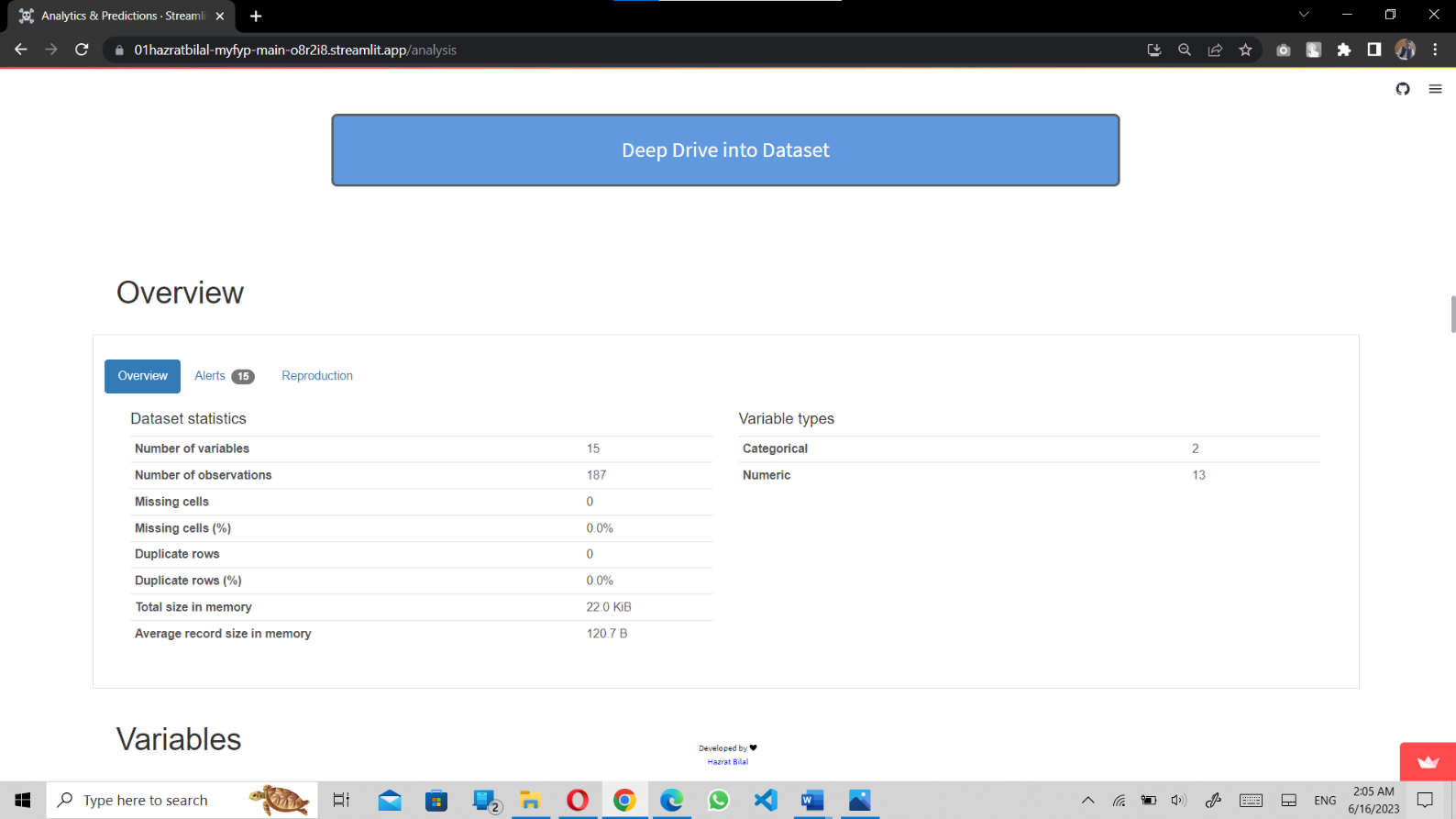
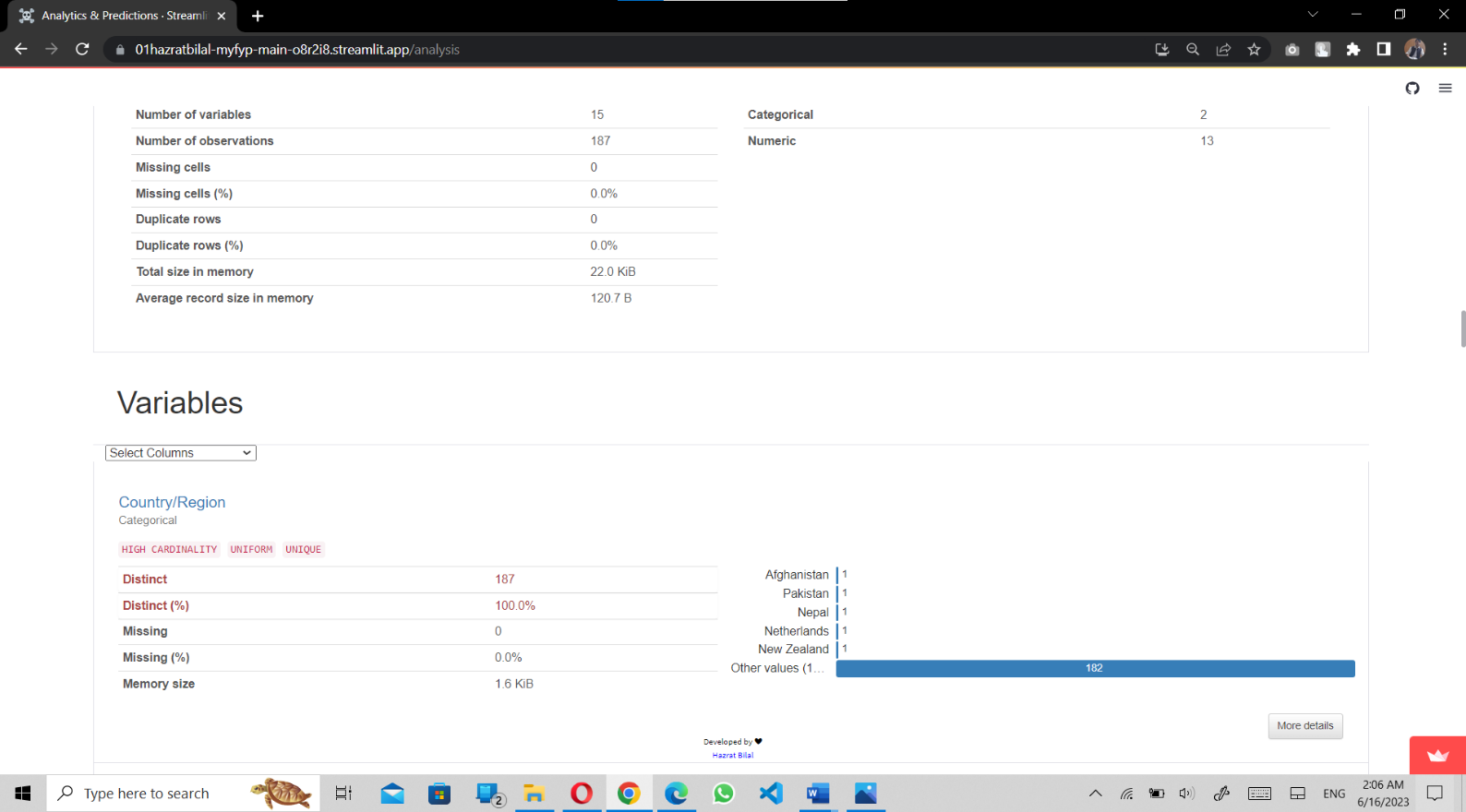
**Figure 5.16 Analysis and Prediction Page**

The head and the tail, the statistics are the same as the training. But you will see a congrats message if the user has no Duplicates and Null values.

* + 1. **Deep Drive into Dataset**

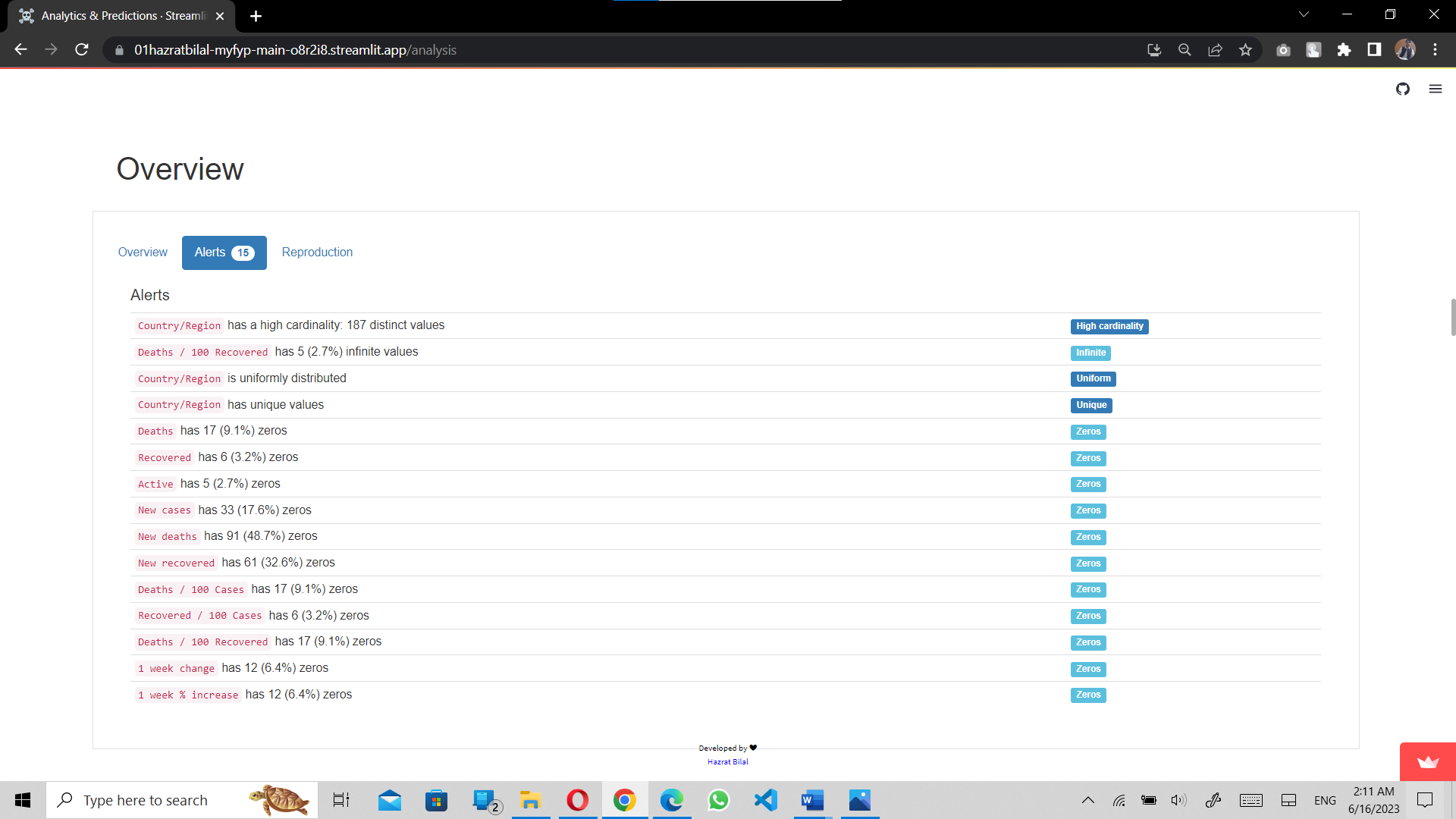
Introducing the new feature of the Analysis & Prediction module. This button hides deep analysis of the dataset and consists of three main features.

1. **Overview**

As the name suggests, this section describes the dataset 360 degree. It includes the statistics, variables and their information with graph, Interactions, Missing values, and the Sample of the dataset.

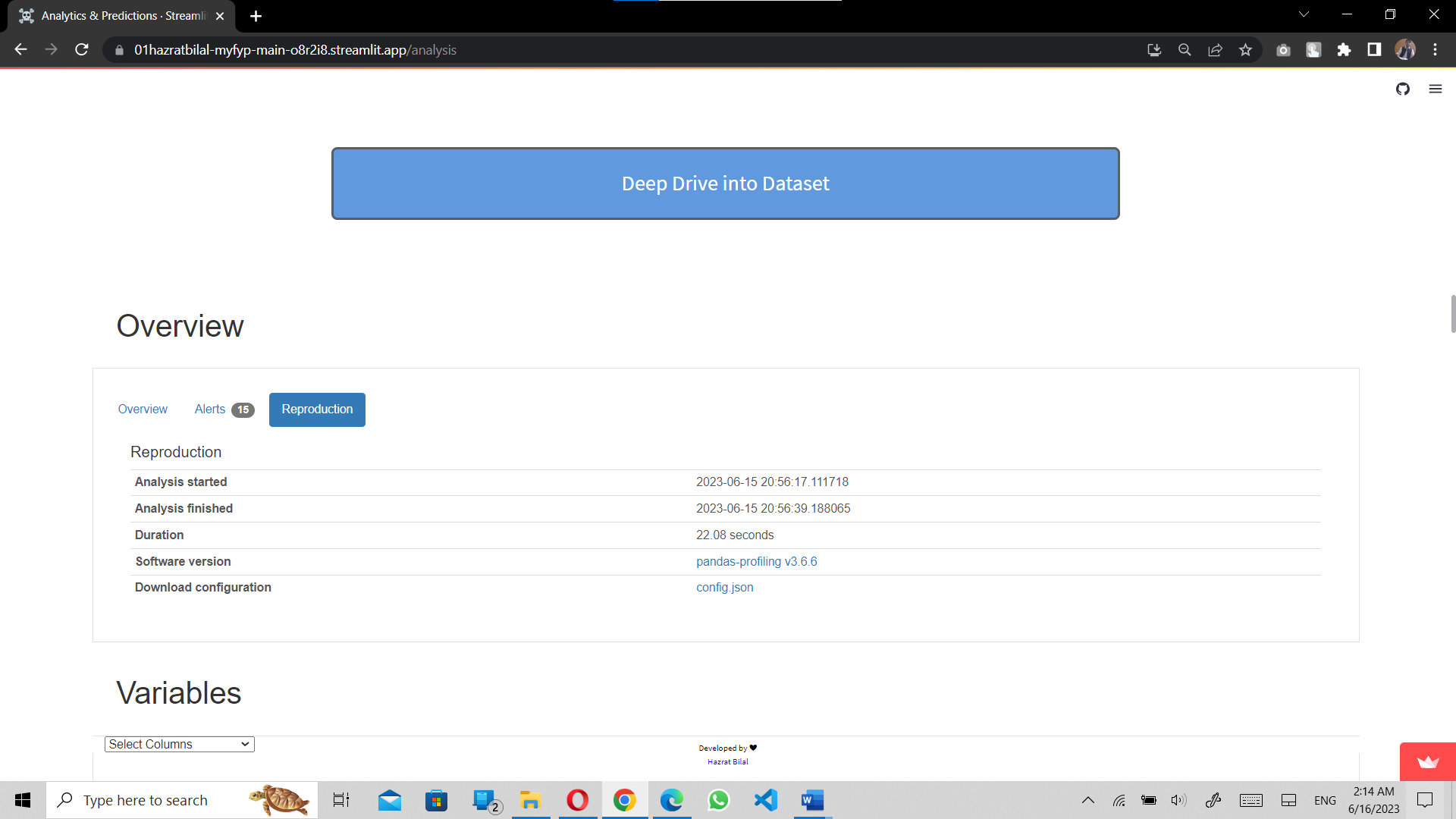
**Figure 5.17 Deep Drive session Overview**

1. **Alerts**

****This session provides the information about the correlation to other columns.

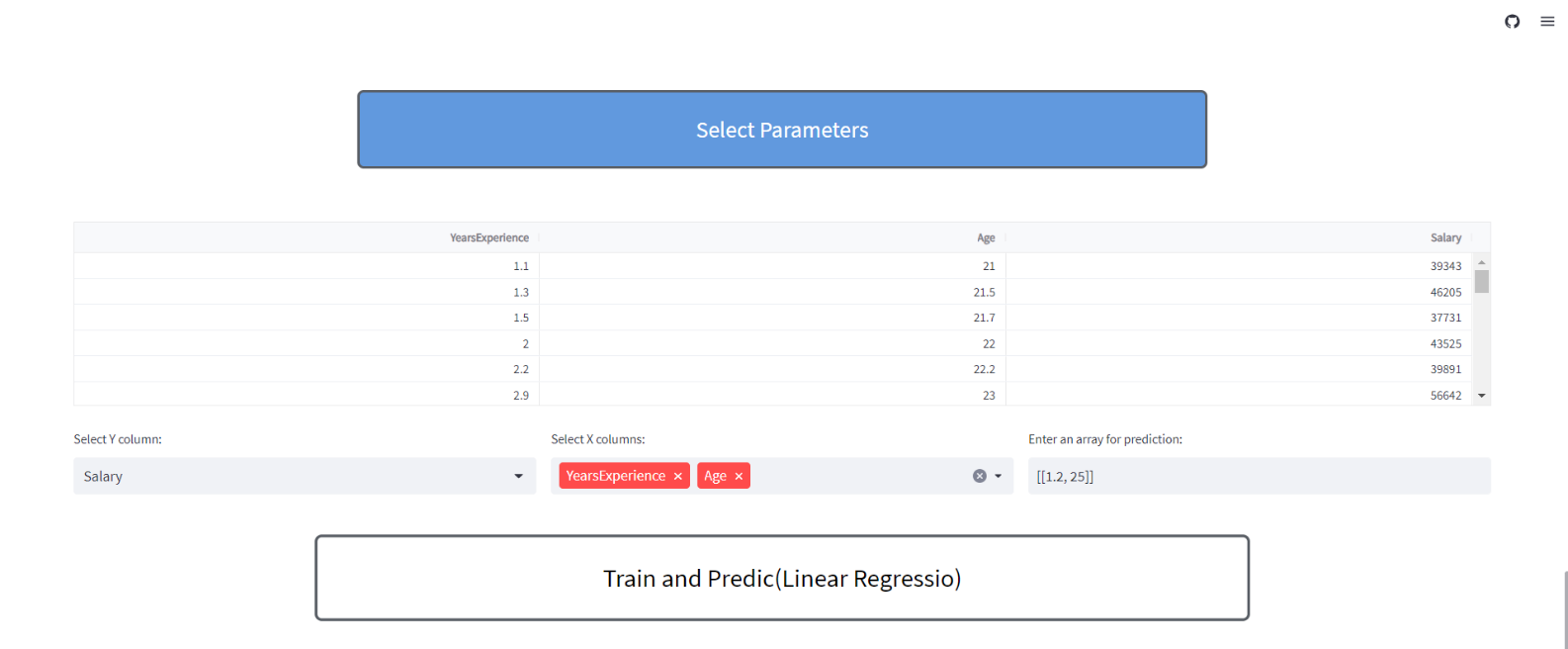
**Figure 5.18 Deep Drive session Alerts**

1. **Reproduction**

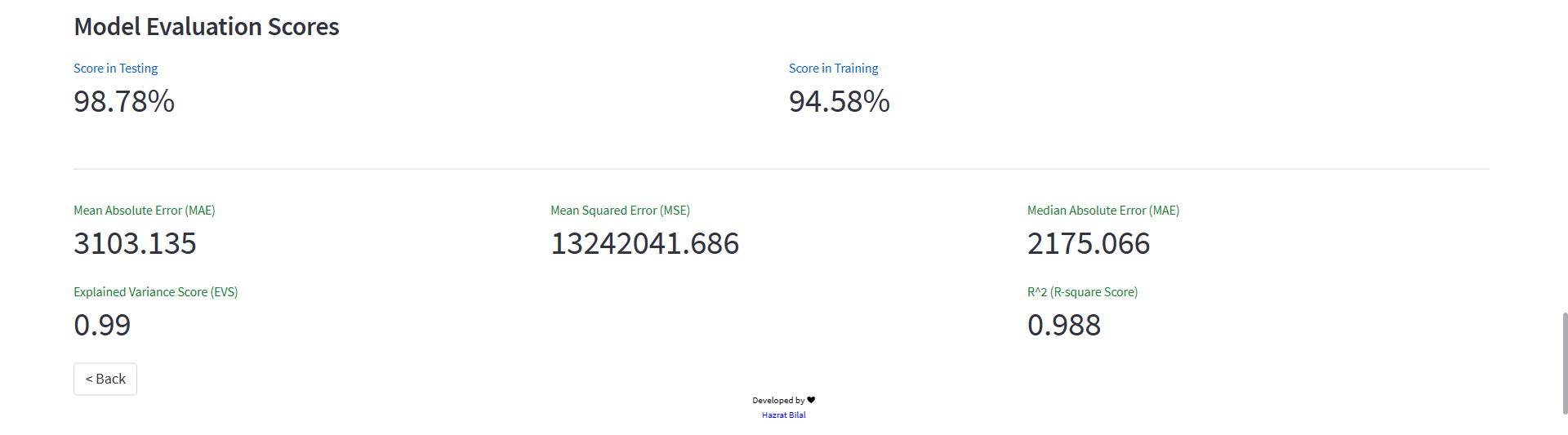
****This session provides about the analysis of dataset such as start data of analysis, finished data of analysis and duration.

**Figure 5.19 Deep Drive session Reproduction**

* + 1. **Linear Regression Button**

This button offers a supervised machine learning algorithm that is use to predict or get a very informed decision based on the data set. Linear Regression can be trained only by numerical data only. This session will also inform the user about the accuracy of training and the testing plus inform about the errors such as Mean Absolute Error, Mean Squared Error, Median Absolute Error, Explained Variance Score, R-square Error.

**Figure 5.20 Linear Regression and Prediction**

****

**Figure 5.21 Score of the Model and Different Errors**

**Chapter 06**

**CONCLUSION**

* 1. **Conclusion**

The Web-Based Data Analysis & Prediction revolutionizes the way users interact with their data. With its intuitive interface and step-by-step training module, it empowers users to explore, analyze, and predict with ease. From data cleaning to advanced statistical analysis, the project offers a comprehensive suite of features that cater to users of all levels.

It provides valuable insights into the dataset, allowing users to handle duplicates, address null values, and perform statistical analysis effortlessly. The application goes beyond basic functionalities, enabling users to create pivot tables, group data, and generate interactive graphs. With the Analysis & Prediction module, users can seamlessly upload their own datasets and conduct customized analysis, including machine learning algorithms.

The Web-Based Data Analysis & Prediction democratizes data analysis by offering a user-friendly platform for both beginners and experts. It streamlines complex tasks, facilitates informed decision-making, and unlocks the true potential of data. With its versatility, compatibility with different data formats, and advanced functionalities, it sets a new standard in data analysis, making it accessible and empowering for users across various domains.

References:

[1] dreastime.com