ML Model assistant

Deepak Kumar. C. R 1833009

Tilak Vijayaraghavan. S 1833055

# Introduction:

Human-computer interaction (HCI) is a multidisciplinary field of study focusing on the design of computer technology and, in particular, the interaction between humans (the users) and computers. While initially concerned with computers, HCI has since expanded to cover almost all forms of information technology design.

Exploratory Data Analysis (EDA) is an approach to analyse the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. Exploratory data analysis is a necessity to do any data science project or to gain insights on the collected data. Facilitating ease of EDA will save valuable time.

This project aims to provide the necessary information to start of a data science project with a blast. Pairing EDA with a ML model selector which will give you an estimate of which ML model might suit you the best.

# Problem Statement:

To create an application using HCI principles incorporated in the design.

# Tools Used:

Python was the main programming language used for the creation of the application, we’ve also used some important libraries like,

1. Pandas
2. Numpy
3. Plotly

For data manipulation and visualisation to facilitate EDA and gain insights. We pair this with a series of machine learning models like

1. Random Forest Regressor / Classifier
2. SVM
3. KNN / K-means
4. Different types of regressions

The application helps select the necessary variables and fits the model to find the most probably best fit model for the selected dataset.

We’ve mainly used streamlit and its dependencies to create the UI / UX designs to facilitate our applications. Streamlit is a python API which facilitates development and deployment of various application.

# Module Design:

The project consists of two distinct modules,

1. The Exploratory Data Analysis Module
2. The Machine Learning Module

The EDA module consists of 6 parts described with an image below,

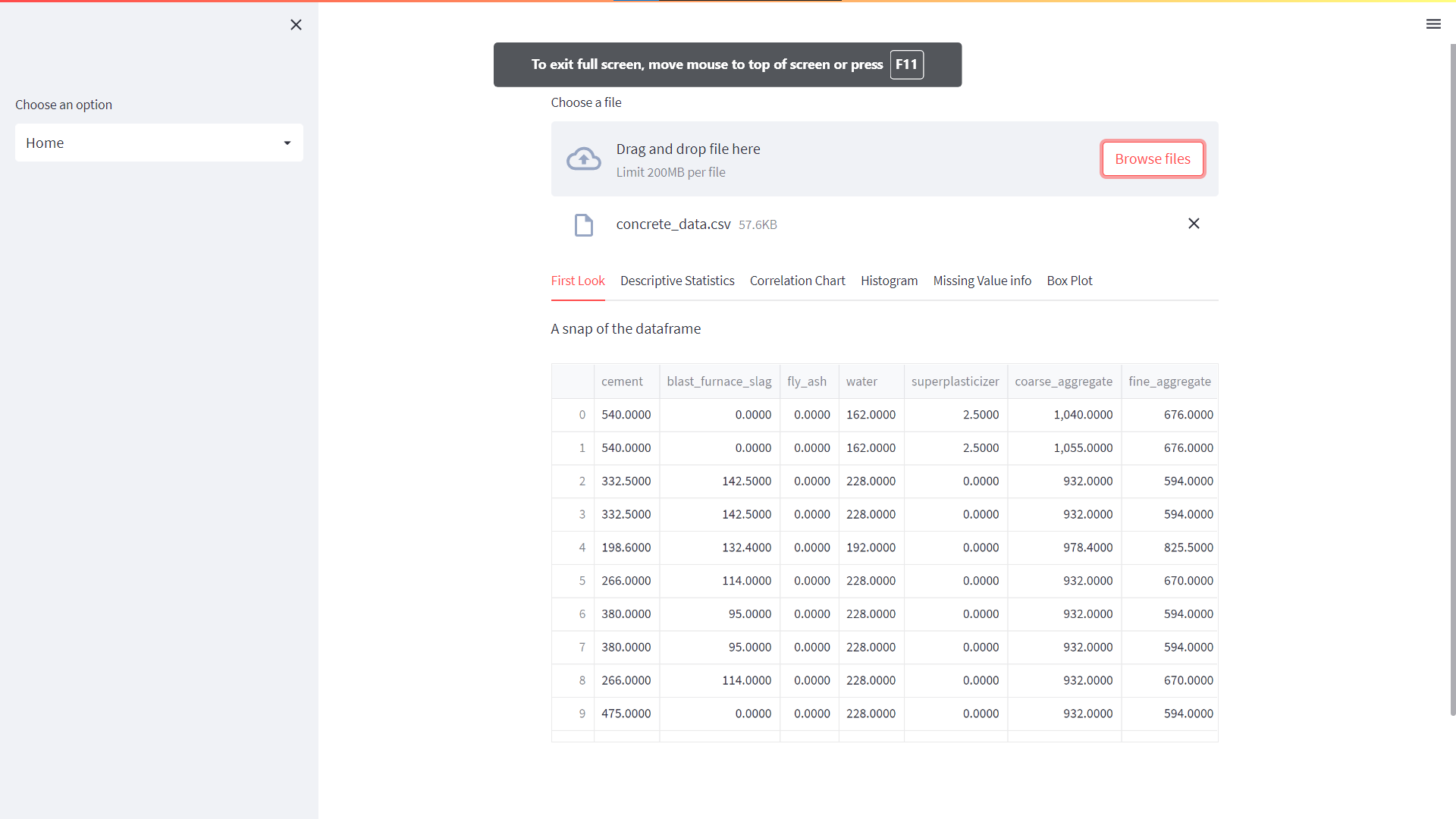
Figure 1

Figure 1 is the first output obtained which shows the data loaded into the application with options to explore the dataset uploaded into the application. This gives the user the oppurtunity to explore the uploaded dataset once again.

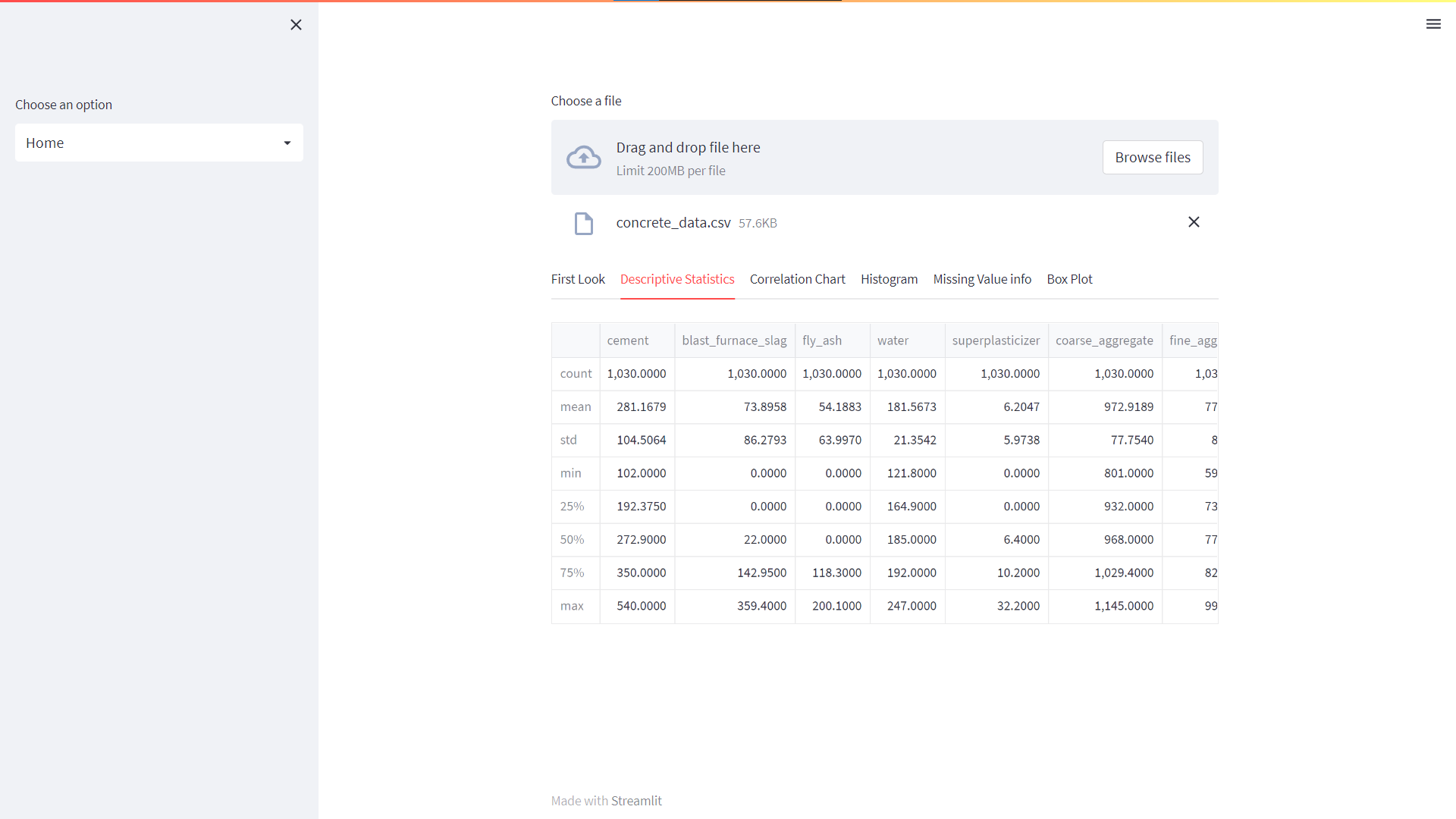
Figure 2

Figure 2 showcases the necessary descriptive statistics obtained from the dataset this includes the regular

1. Mean
2. Median
3. Max
4. Min
5. Quartiles

For all the columns listed in the dataset. This is useful when you are trying to understand the distribution the data might follow which helps with coming up with an assumption for the model to be fit. Once, again, the user can gain more insights on the collected data by understanding what the data says statistically.

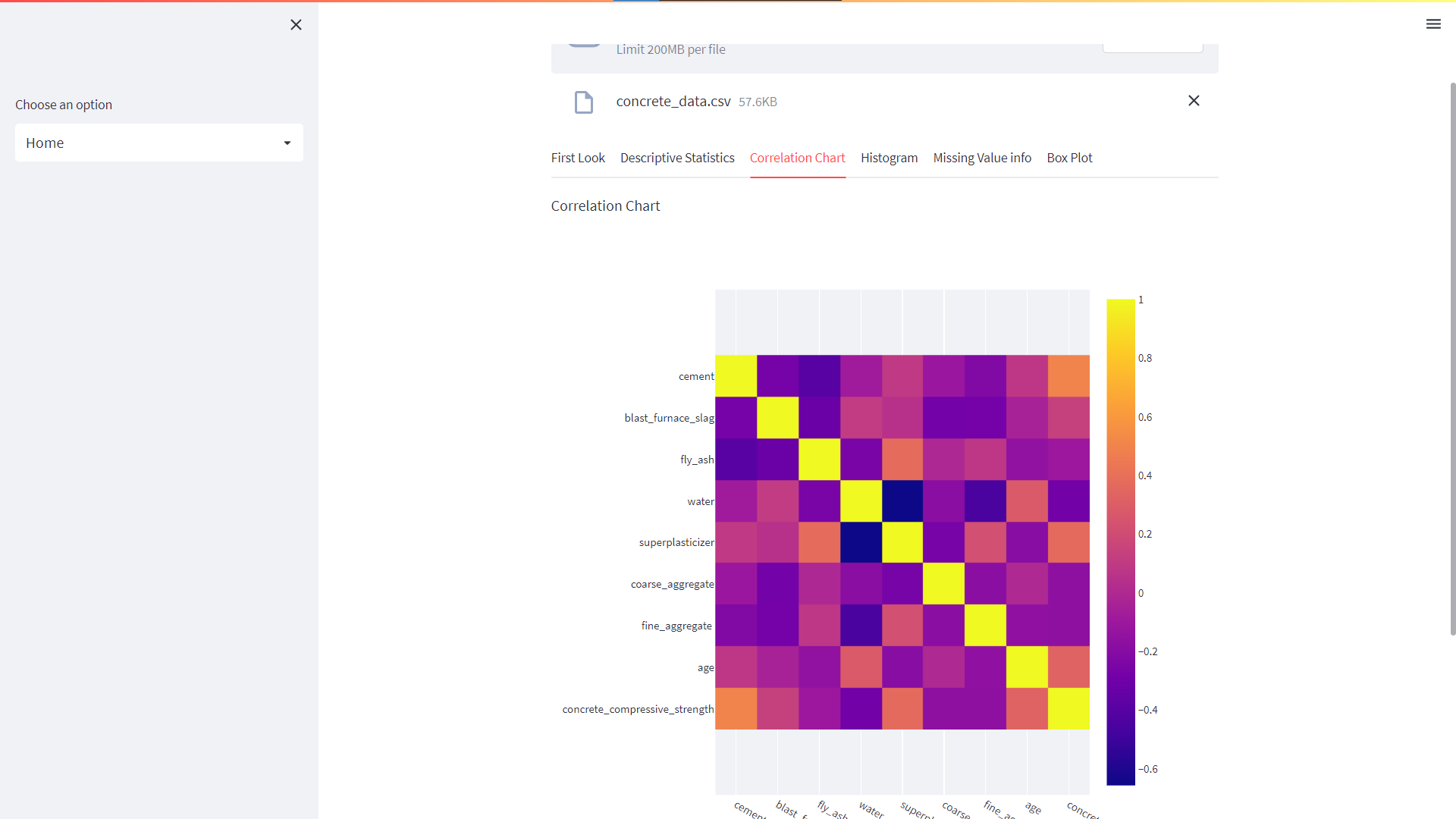
Figure 3

Figure 3 shows the correlation that exists between the variables in the dataset. This helps in identifying the possible variables to avoid during the construction of the model. This is because multi-collinearity could severely affect the performance of the model that is being built. This helps weed out the unnecessary variables seen in the data.

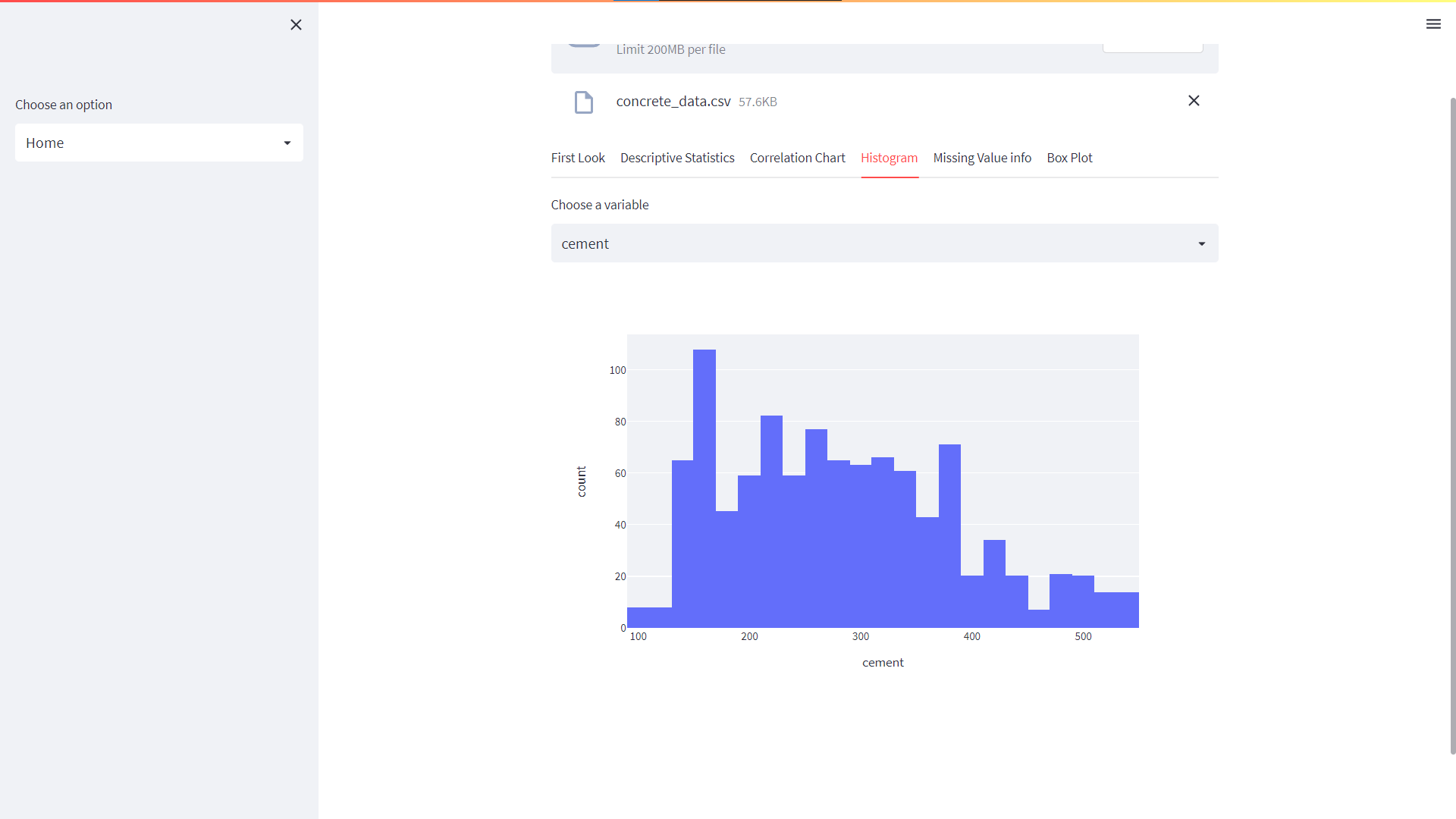
Figure 4

Figure 4 is a histogram plot which shows the data distribution of data to help assume the nature of the data present in each column. Once again, providing additional information to facilitate data exploration help with developing a better hypothesis and a better model.

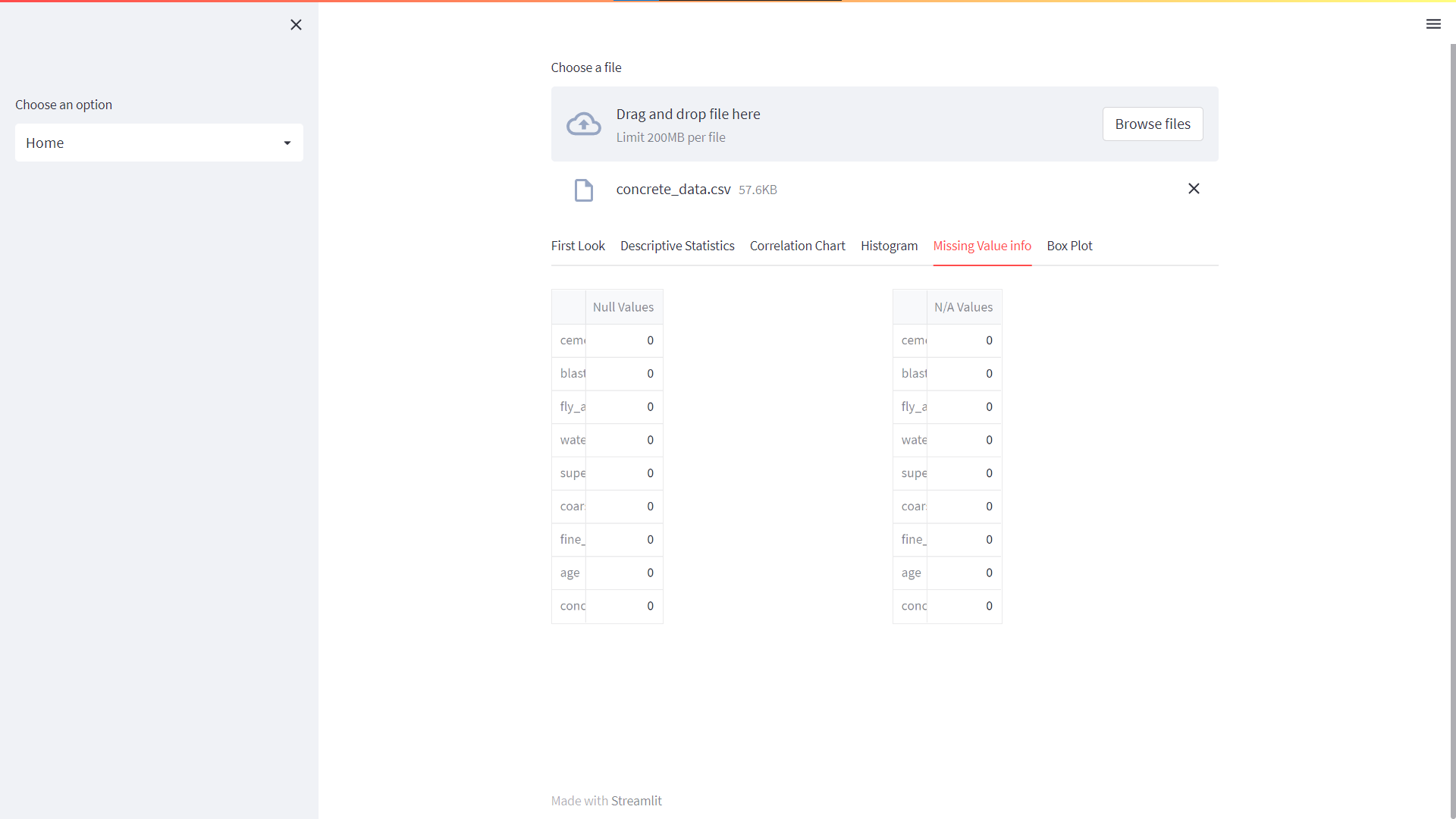
Figure 5

Figure 5 help with understanding any anomalies within a column. This is seen by the number of gaps left in-between the rows in the dataset. This is also a valuable information as this helps with deciding if a model needs interpolation of data or not. This could be crucial if the data needs fine tuning.

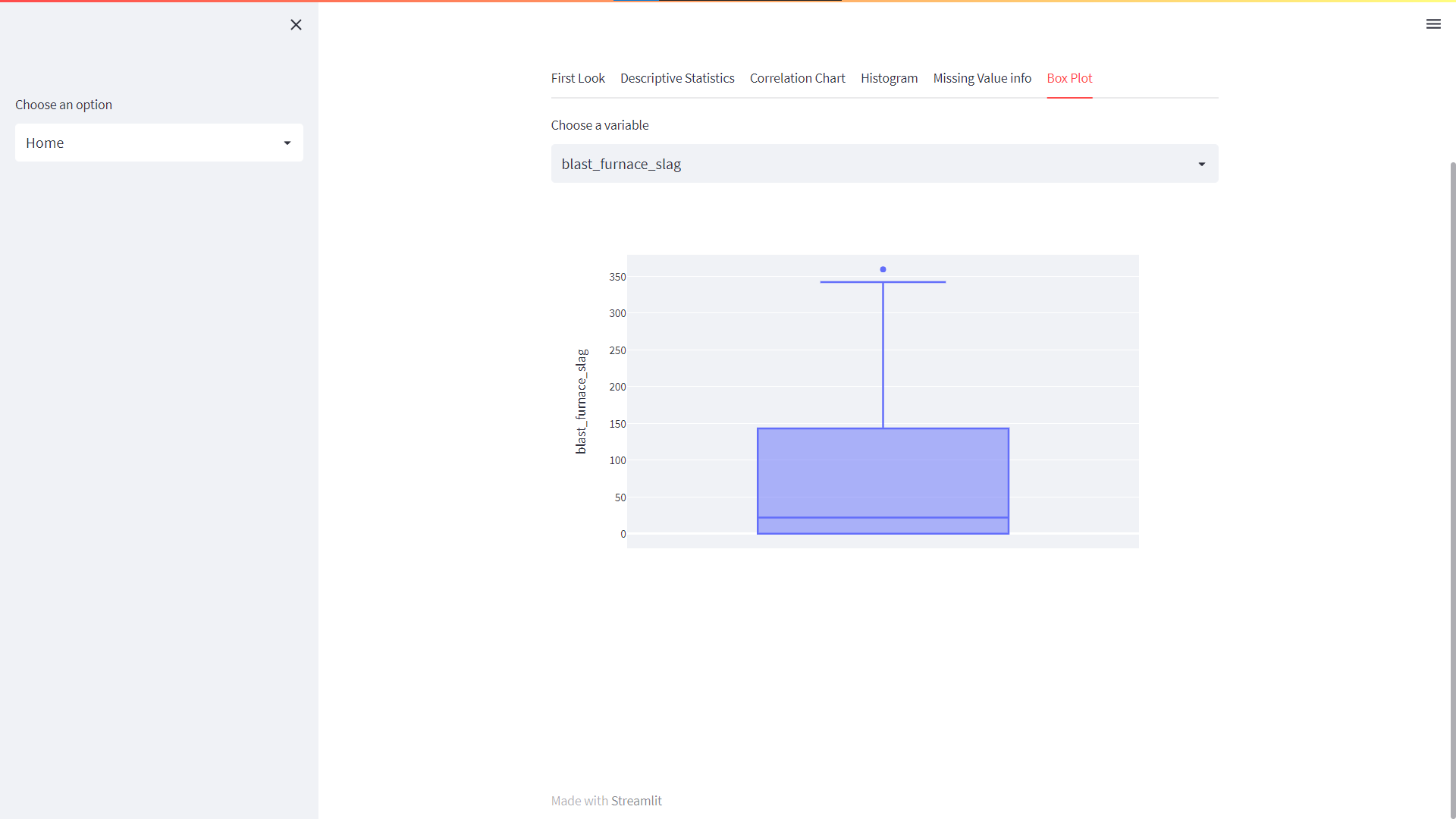
Figure 6

Figure 6 shows the distribution of data in terms of its quartiles and shows the distribution based on its spread from the median. This helps us identify the outlier present in the data and whether or not we need to remove them.

The second module in the application helps with choosing a machine learning model for the selected dataset. This is achieved by providing the necessary inputs like the independent and dependent variables. Once the inputs are given and the nature of the model is specified, different models are ran to determine their possible scores when built individually. This will give the necessary head start required for the project.

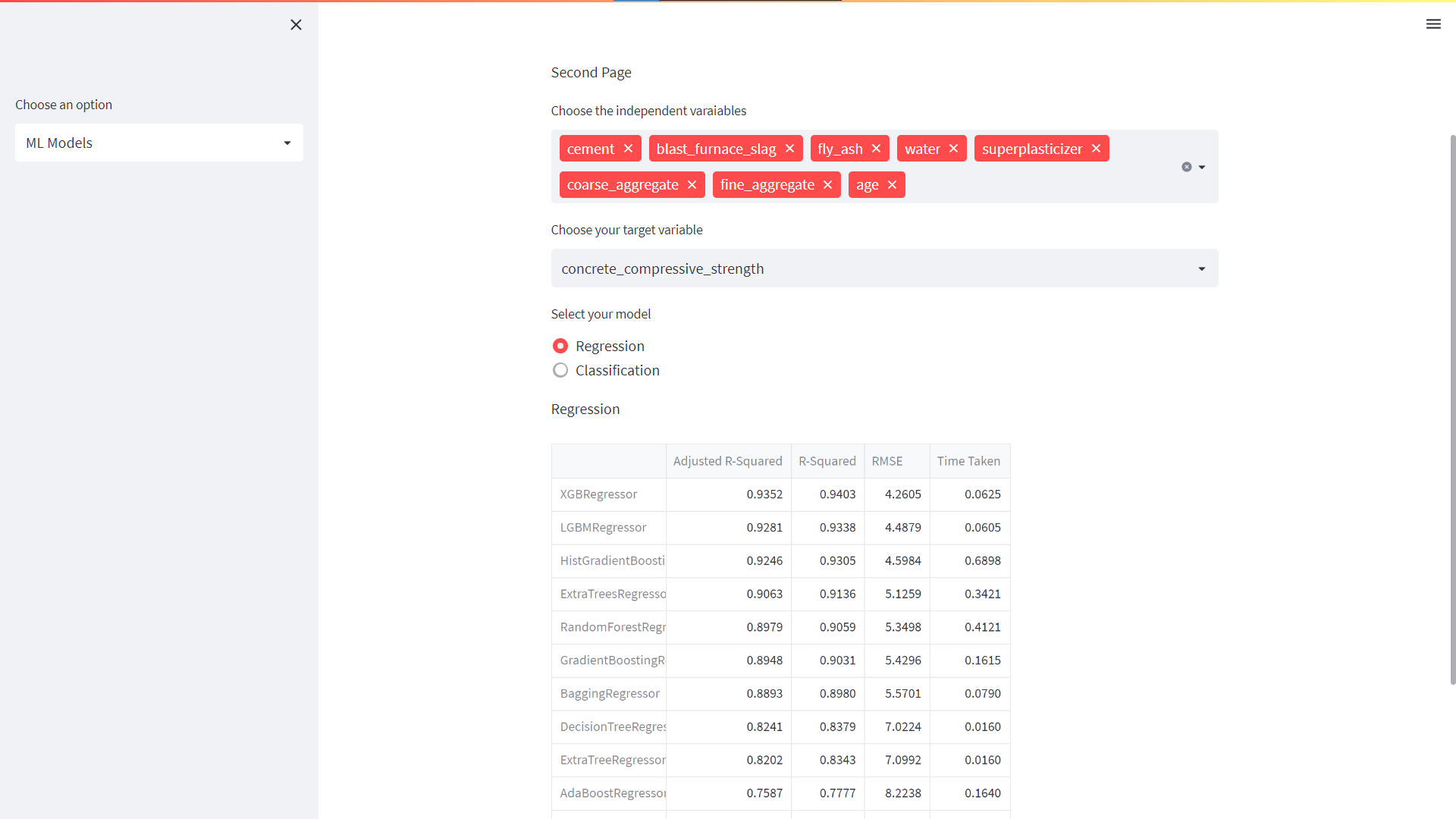
Figure 7

Figure 7 shows the snapshot of the second module involved in the project. This page takes in the necessary variables as inputs and gives a suggestive output.

# Results and Discussion:

Once, the application was ready, we went ahead and tried to use a dataset to try out the model’s performance. The dataset chose for testing was dataset on the strength of cement obtained from the different items that makeup its constituents.

The dataset has 8 explanatory and 1 explained variable. We’ve used the dataset in the images to show a sample of the application above. We’ve used variables like (all the variables represent their quantity)

1. Cement
2. Fly ash
3. water
4. superplasticizer
5. coarse aggregate
6. fine aggregate
7. age
8. blast furnace slag

By giving them as input, we found that a random forest regressor is the best model for the project and we built the model with the following metrics,

# Conclusion:

We conclude the project by successfully creating an application which can assist with data science project and built with HCI principles.