**Machine Learning as a Platform (MLaaS)**

**Phase 1 - Requirement Document**



Table of Content

[Objective 4](#_Toc61996951)

[Scope 4](#_Toc61996952)

[Login 4](#_Toc61996953)

[User Actions 4](#_Toc61996954)

[Validations 4](#_Toc61996955)

[Navigation Panel 5](#_Toc61996956)

[Activities Timeline 6](#_Toc61996957)

[Data Ingestion Module 7](#_Toc61996958)

[Create Dataset Page 7](#_Toc61996959)

[Create Project Page 9](#_Toc61996960)

[All Projects Page 10](#_Toc61996961)

[Data clean-up and Visualization Module 12](#_Toc61996962)

[Data Detail Page 12](#_Toc61996963)

[Schema Mapping Page 12](#_Toc61996964)

[Data Exploration Page 14](#_Toc61996965)

[Data Visualization Page 15](#_Toc61996966)

[Data Clean-up Page 17](#_Toc61996967)

[Formulas/Logic used 19](#_Toc61996968)

[1. Distribution Graph 19](#_Toc61996969)

[2. Datatype Determination 19](#_Toc61996970)

## 

**Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Revised by** | **Date** | **Reason for Changes** | **Version** |
| Swati Srivastava | 12/11/2020 | First Draft | 1.0 |
| Swati Srivastava | 12/29/2020 | 1. Added requirements for **Schema Mapping** and **Data Exploration** page  2. Updated the validation for **Create Dataset** page | 2.0 |
| Swati Srivastava | 01/04/2020 | Added requirements for **Login** Page and **Activities Timeline** | 3.0 |
| Swati Srivastava | 01/19/2020 | 1. Added requirements for **Data Exploration**, **Data Visualization** and **Data clean-up** page  2. Updated validation for **Schema Mapping** page | 4.0 |

## Objective

Objective of this POC is to develop a platform for Machine Learning as a Service. This platform will offer services for data clean-up, data visualization, and predictive analytics with machine learning algorithm.

## Scope

Scope of this POC is to develop an end-to-end pipeline to let users-

1. Upload the dataset in CSV format
2. Clean-up the data
3. Perform data exploration
4. Define data splitter parameter
5. Predict the data for target column

## Login

### User Actions

1. This page will let user login to the application by entering registered username and password

2. This page also provides feature to regenerate the password if the user forgets password.

### Validations

* Application should show error message when user enters invalid username or password

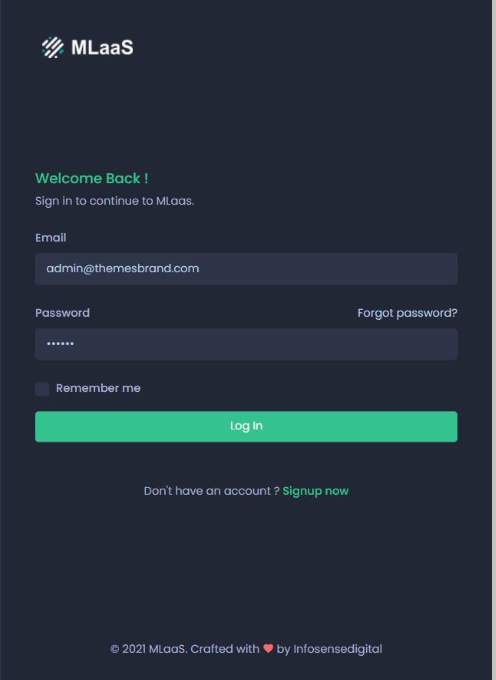


Figure : Login Page Mock-up

## Navigation Panel

* A navigation panel is provided at the left side of the application
* The panel will show the links of all the pages. User can navigate to the desired page by clicking the link
* The navigation panel will highlight the current page link
* The pages where user is not eligible to navigate at any given point of time will be shown as inactive link

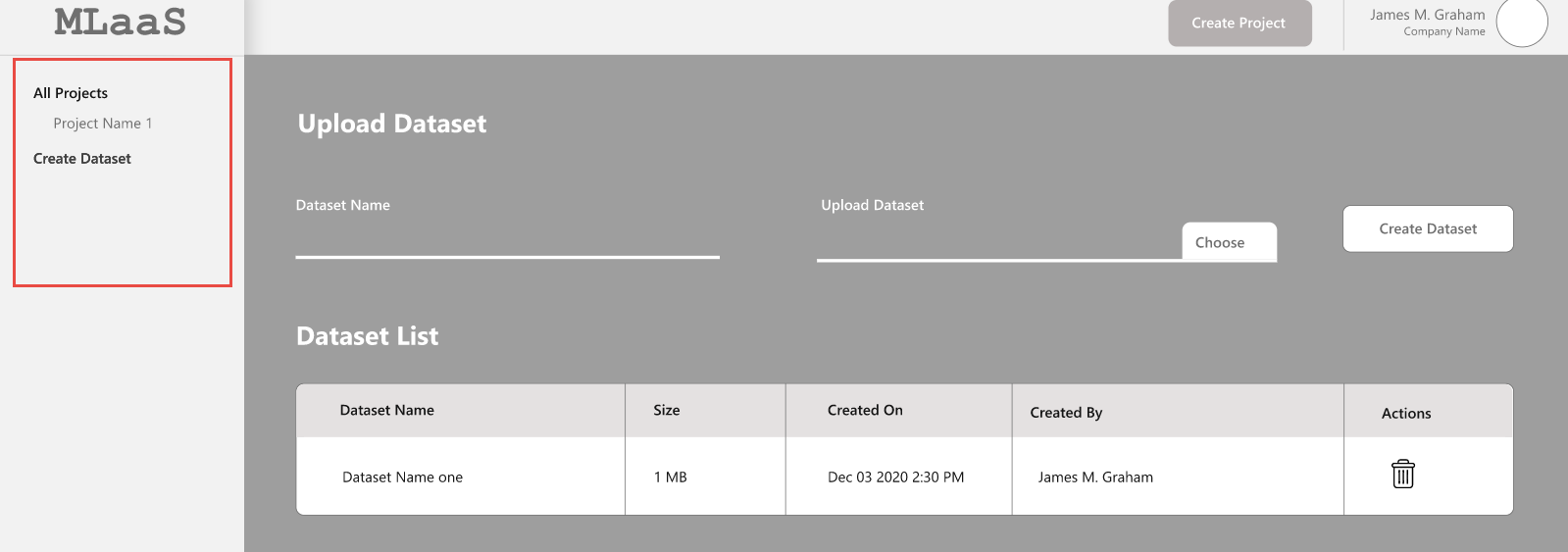


Figure 2: Navigation Panel Mock-up Screen

## Activities Timeline

Timeline panel of the application shows the history of recent activities performed by user.

Following are the activities and description to be shown for Data ingestion module-

|  |  |  |
| --- | --- | --- |
| **Sr. no** | **Activities** | **Description** |
| 1 | Created dataset | You have created dataset “ABC” |
| 2 | Deleted dataset | You have deleted dataset “ABC” |
| 3 | Created project | You have created project “ABC” |
| 4 | Deleted project | You have deleted project “ABC” |

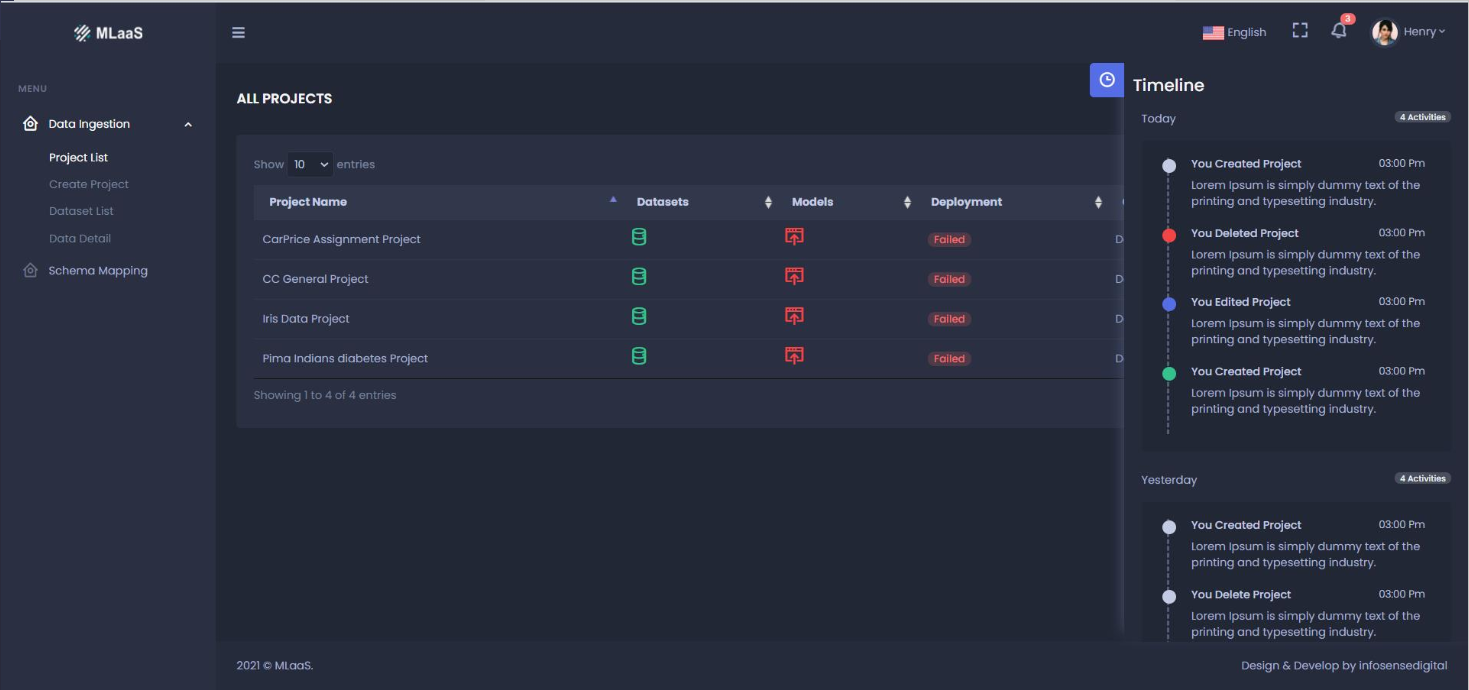


Figure : Activities Timeline Mock-up

## Data Ingestion Module

### Create Dataset Page

Create Dataset page will allow users to upload dataset in MLaaS page. These datasets can be used for multiple projects

#### User Actions

##### 1. Upload Dataset

To upload a dataset, user need to-

* Enter Dataset name
* Click **Choose file** to select the dataset from local system
* Click **Create Dataset** button to upload the dataset

##### 2. Dataset Access-

User will get an option to either keep the dataset as **Public** or **Private**

##### 3. Dataset List

Once the dataset is uploaded, following details will be shown in Dataset List-

* Name of the dataset
* No. of rows in the dataset
* Size of the dataset
* Date and time the dataset was uploaded on
* Name of the user who has uploaded the dataset
* User will get an option to remove the Dataset

#### Validations

* Datasets which are already being used by projects cannot be deleted.
* Application should show error message when user tries to delete a dataset being used by project
* User can only delete the dataset uploaded by him/her
* Confirmation window should appear when user deletes a dataset
* Users are only allowed to upload CSV dataset
* Application should not let user upload the CSV files with junk data in Name of Rows and Columns
* System should not let user create datasets with duplicate name (Same as existing datasets)

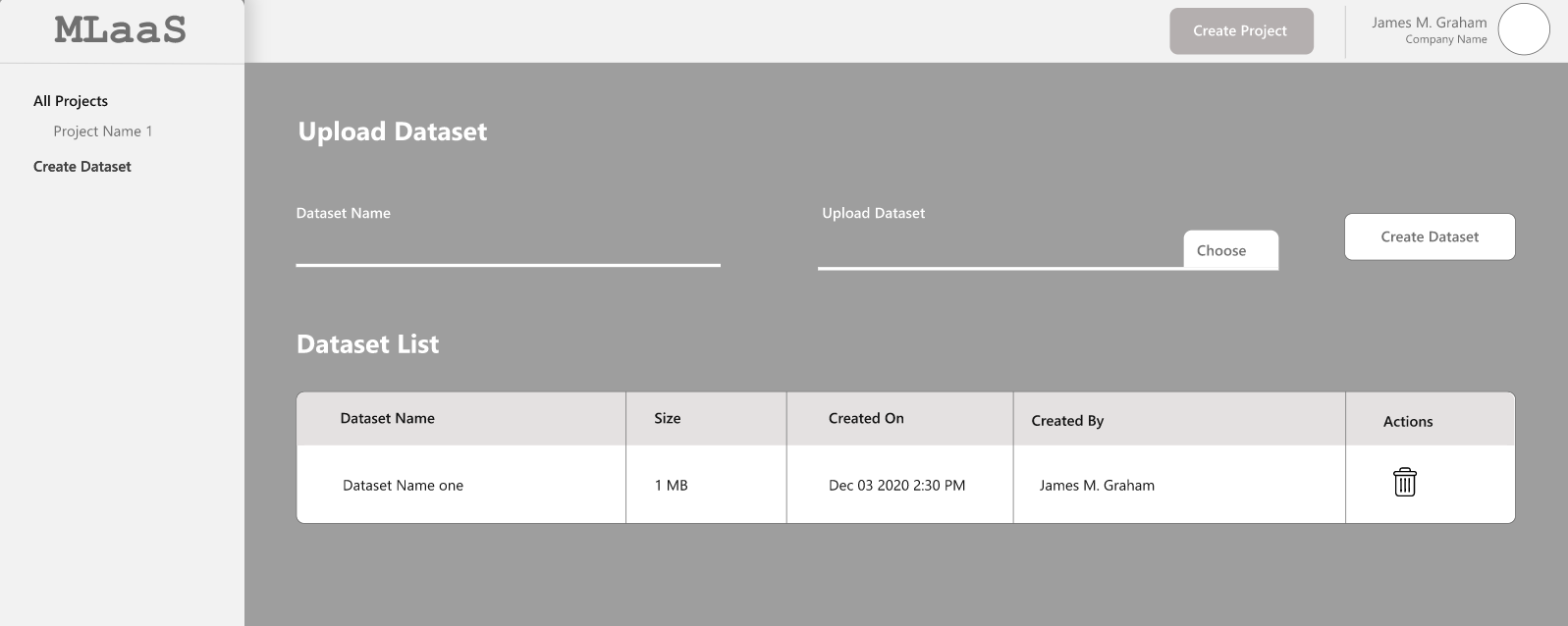


Figure 4: Create Dataset Mock-up Screen

### Create Project Page

This page let users create new projects and add dataset to the project.

#### User Actions

##### 1. Create Project

To create a project, user need to-

* Enter Project Name
* Provide a brief description of the project
* Add dataset to the project. A dataset can be added by-
  + Uploading a new dataset by clicking **Choose File,** defining the access- Private or Public**,** and provide the name of dataset
  + Selecting an existing dataset from the **Select Dataset** dropdown list
* Once the dataset is uploaded, click **Create Project**

#### Validations

* Select Dataset Dropdown list should display all the public dataset listed on the **Create Dataset** page and datasets private to the user
* On clicking **Create Project**, user should be directed to [**All Projects**](#AllProjectPage) page with the new project added to the list
* System should not let user create projects with duplicate name (Same as existing projects)

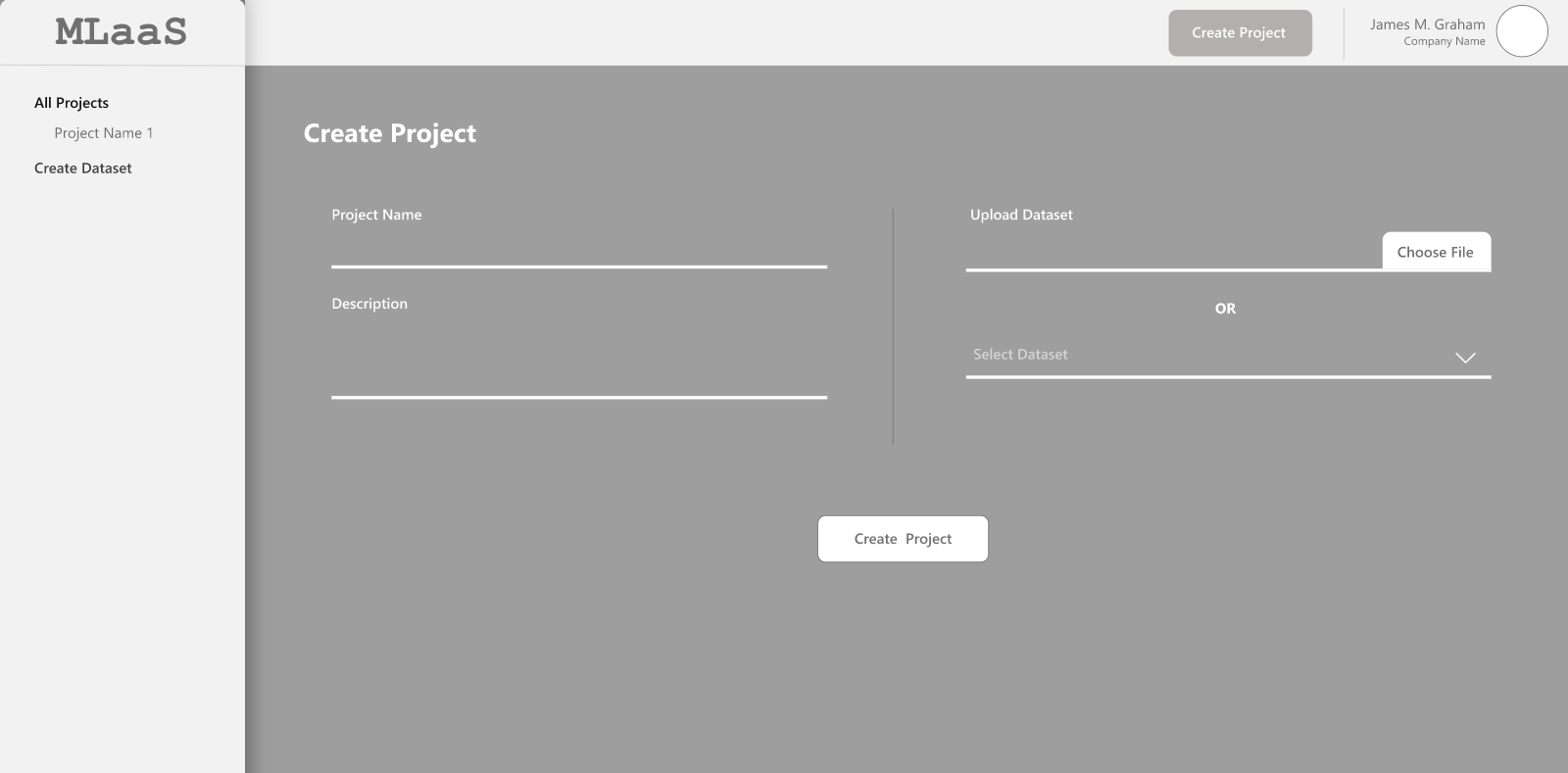


Figure 5: Create Project Mock-up Screen

### All Projects Page

This page shows all the existing projects in the system and its status. Following details are shown in this page-

* **Project Name** – Name of the Project
* **Dataset** - Shows the status of Dataset-
  + **Green** denotes correctly mapped datasets
  + **Red** stands for an error related to the dataset.
* **Model**- Shows the training status of the model
  + **Green** denotes a successfully trained model
  + **Yellow** denotes training in progress
  + **Red** stands for "Training Failed" status for a model.
* **Deployment** – Shows thedeployment status of the project-
  + "Active" deployment
  + "Suspended" deployment
* **Created** – Project creation Date and Time
* **Actions** – Provide option to see Dataset details and delete the project

#### User Actions

##### 1. See Data Details

User can see the details of the dataset uploaded for the project by clicking **See Detail** icon under Actions Tab.

Clicking this icon will direct the users to [**Data Detail**](#DataDetailPage) page.

##### 2. Delete Project

To delete a project, user need to click the **Delete** icon provided under the Actions tab

#### Validations

* Confirmation window should appear when user deletes a project
* User cannot delete a project if it’s deployment is going on. In such scenario, if user tries to delete a project, system should provide an error message.
* User cannot delete a project if the model training is in progress. System should display error message in such scenario
* When user deletes a project, its dataset will not get deleted

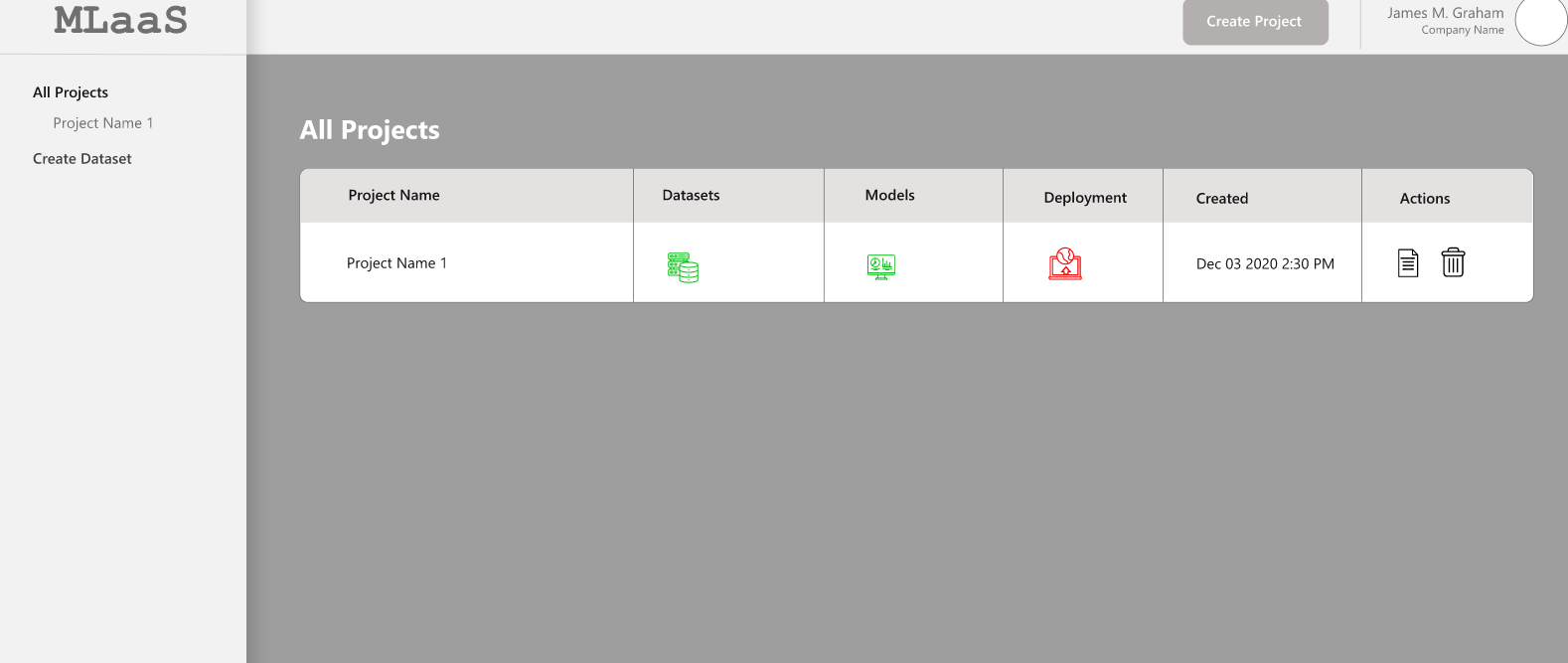


Figure 6: All Projects Mock-up Screen

## Data clean-up and Visualization Module

This Module is comprised of the processes of fixing/removing incorrect, corrupted, duplicate, or incomplete data within a dataset and provide the graphical representation of information.

### Data Detail Page

This page shows the raw data of the dataset uploaded for the project

User can navigate through the rows using page navigation and columns using scroll bar.

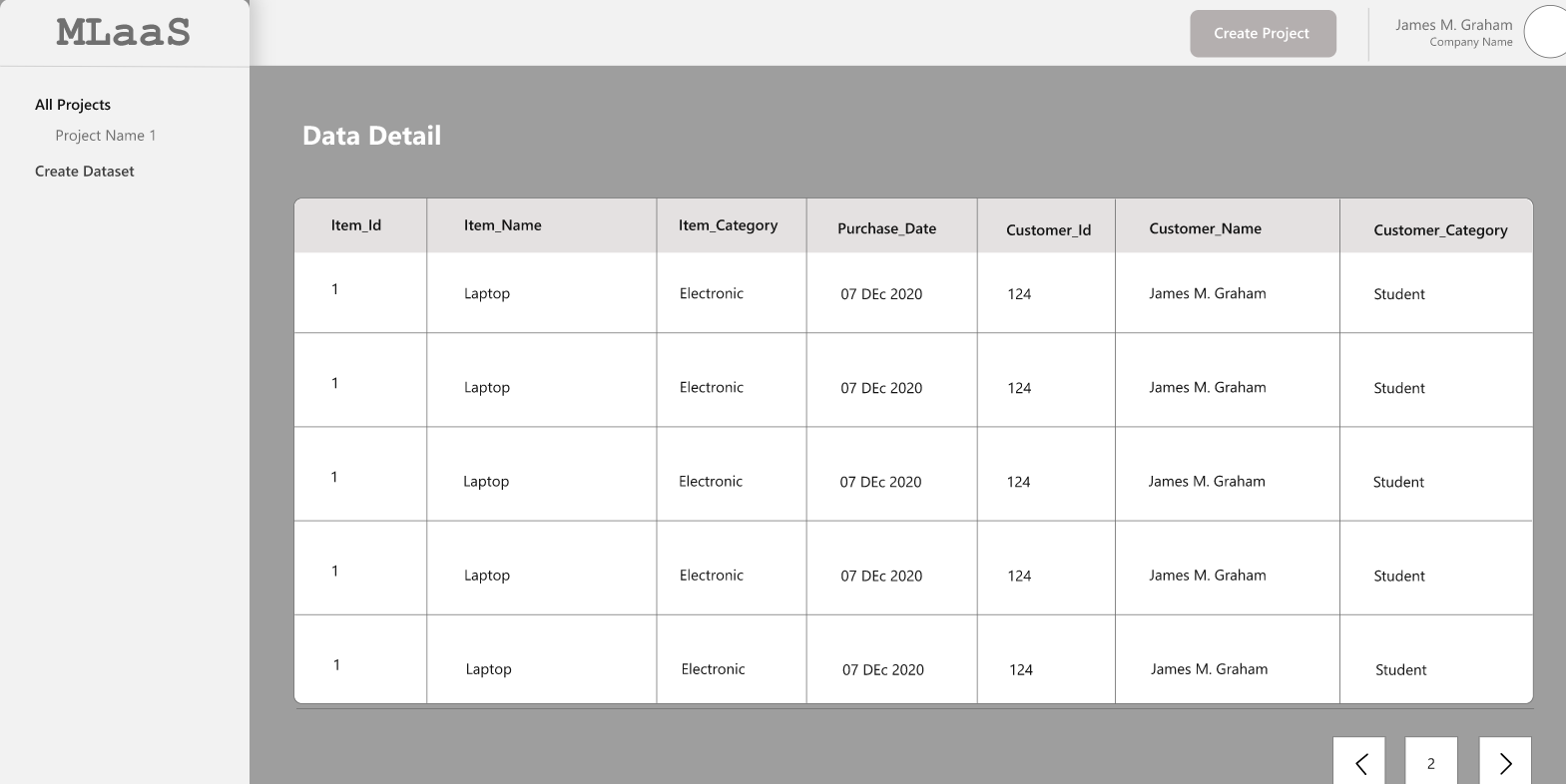


Figure : Data Detail Page Mock-up Screen

### Schema Mapping Page

#### User Actions

##### 1. [Update Datatype](#_2._Datatype_Determination)

Schema Mapping page shows the datatype of each column. Datatype can be-

* Categorical
* Categorical List
* Numerical
* Text
* Timestamp

##### 2. Select column attribute

User can select attributes for each column.

* **Target Column** -user can select a column as Target column on which the prediction is to be done
* **Ignore**- User can select Ignore for the columns to be not considered for the experiment

##### 3. Change column Name

User can choose to update the name of the column from this page.

#### Validations

* Multiple columns can be selected as Target Column
* **Assumption** – User cannot change the datatype
* Column Name and Changed Columsn Name cannot be same
* The target columns should get highlighted
* On clicking Reset, application should initialize all user-input values
* It is NOT mandatory to select the Target column.

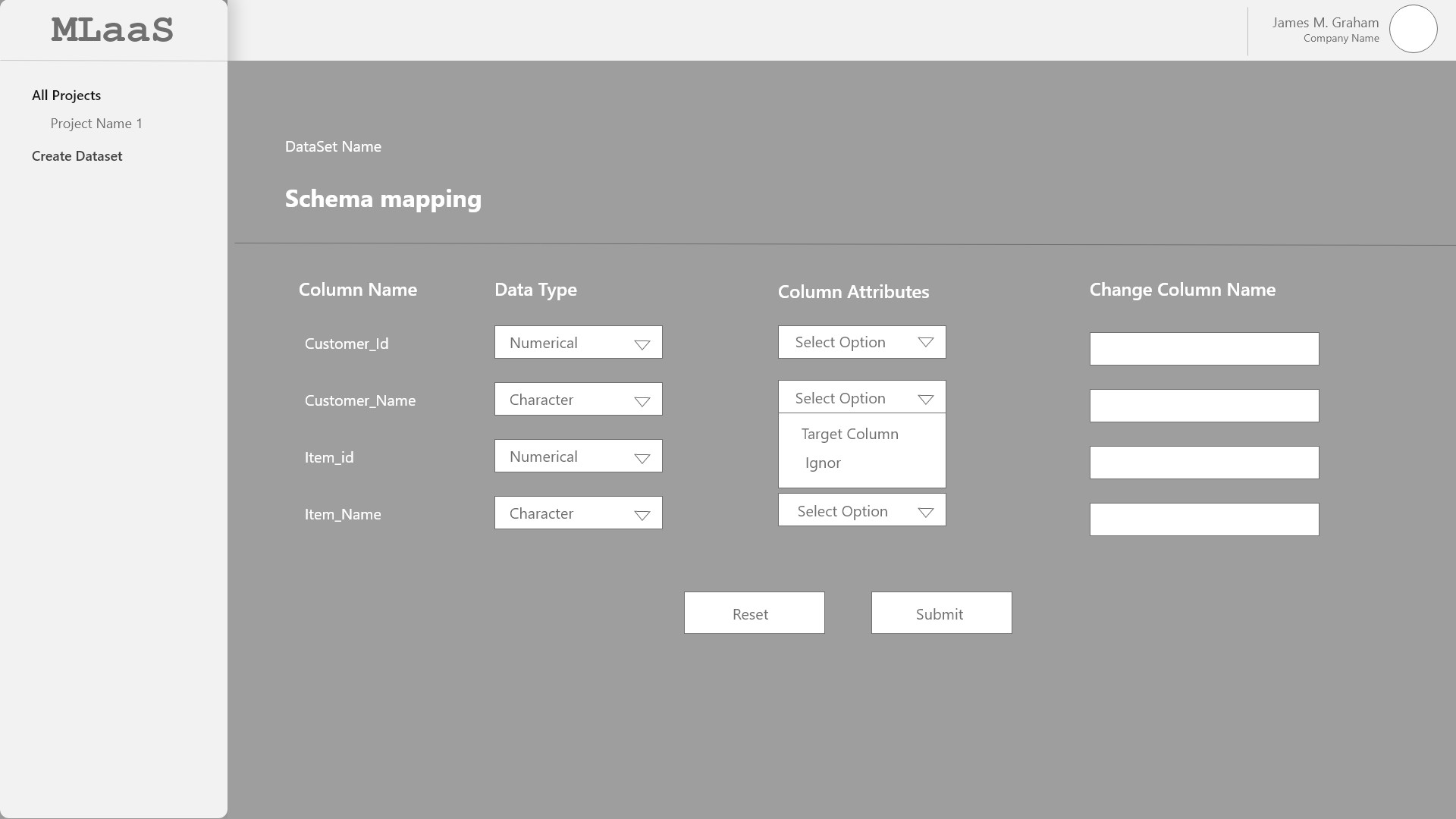


Figure : Schema Mapping Mock-up Screen

### Data Exploration Page

This page shows all the columns selected from the data Schema page and its statistics

#### User Action

##### 1. [Distribution graph](#_1._Distribution_Graph)

User can see the distribution graph for each of the column on this page. These distributions show the spread (dispersion, variability, scatter) of the data.

* Histogram is shown for columns with Numerical, Timestamp datatype
* Count plot is shown for columns with Categorical, Categorical list, Text datatype

##### 2. Data Statistics

User can see the data statistics based on the type of column-

For Continuous column-

* Data Type
* Data Count
* Missing data
* Mean value
* Min Value
* Max Value
* Std Deviation
* Unique Data

For Categorical column-

* Data Type
* Data count
* Missing Data
* Most Occurring
* Least occurring
* Unique data

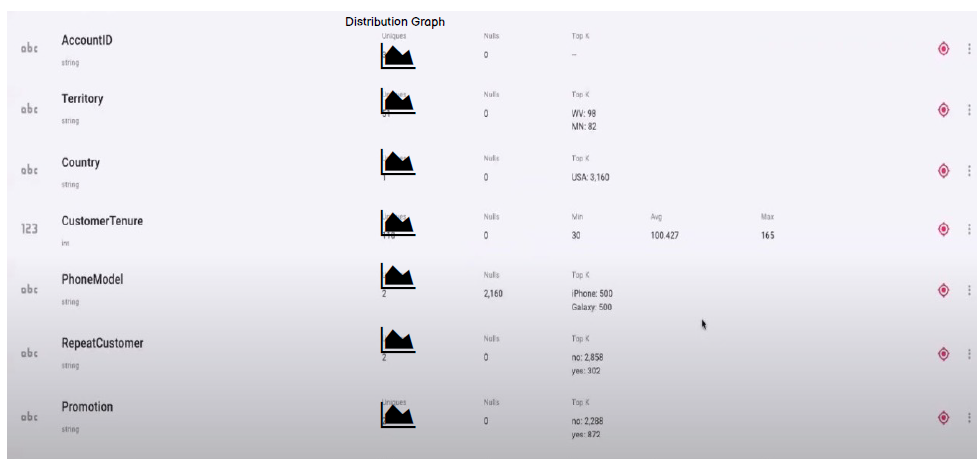


Figure 9: Data Exploration Mock-up

### Data Visualization Page

This page let user select the columns from dataset and explore the details with advanced visualization.

#### User Actions

##### 1. Select plot type

User can choose form one of the following plot type-

* Histogram
* Box Plot
* Scatter Plot
* Distribution Plot
* Count Plot
* Heat Map

##### 2. Select column

User can select the single or multiple columns to plot the graph based on the type of plot selected.

##### 3. Change range of axis

User can change the range of X and Y axis of the plot

##### 4. Legend

Plots should have Legend to reflect the data displayed in graph

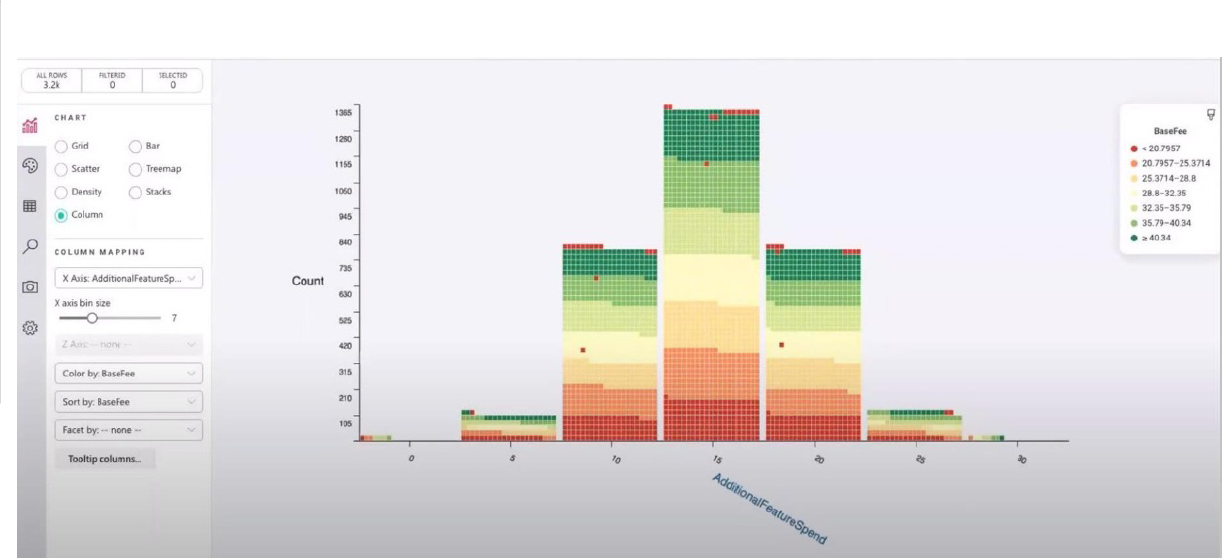


Figure 10: Data Visualization Page Mock-up

### Data Clean-up Page

#### User Actions

##### 1. Select column

User can select multiple columns of the dataset to perform the clean-up activities

##### 2. Missing values handling

User can select the column and choose one of the following options to handle the missing values-

* Replace Missing values with Median
* Replace missing values with Mean
* Drop the affected rows

##### 3. Rescaling

User can select one of the following methods for rescaling the data for selected columns-

* Min-Max
* Std. Scaler
* Robust
* Normalization

##### 4. Remove Noise

User will can choose to remove the noise form selected column by opting one of the following options-

* Regression
* Binning

##### 5. Remove duplicate data

User can select this checkbox to remove the duplicate data from the selected columns

##### 6. Select Encode type

User can select one of the following encoding types-

* One hot encoding
* Label encoding

**Apply –** On clicking apply, the selected options should be get applied to for the selected columns. All the options must be visible under the **Handling** section for respective columns.

**Reset-** On clicking Reset, application should initialize all user-input values.

**Save and Proceed –** On clicking **Save and Proceed**, all the selected options will get save and user will be directed to next phase for Data Modelling.

#### Validations

* Multiple columns can be selected to apply Data clean-up options
* Multiple options CANNOT be selected for Missing values handling, Rescaling, Remove Noise by, and Select Encode type
* Once the cleanup criteria are selected and user has clicked **Apply**, user should be able to reselect the columns and change the options

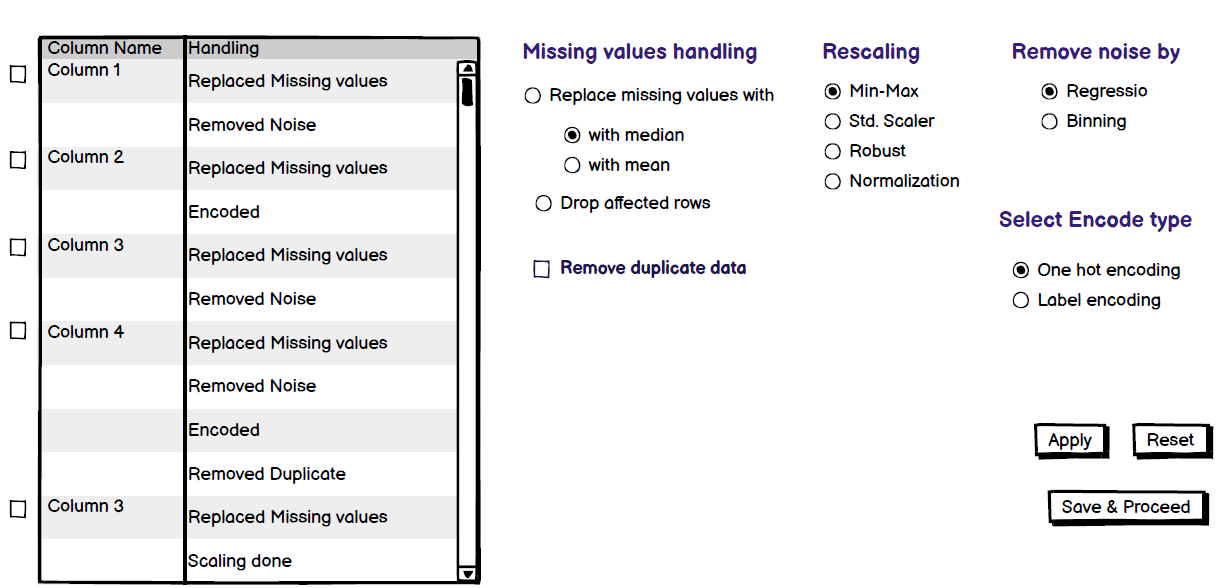
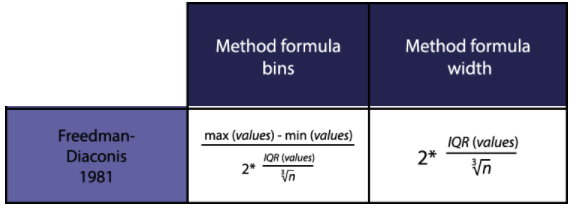


Figure 11: Data Clean-up Page Mock-up

## Formulas/Logic used

### 1. Distribution Graph

To plot the graph, bins are decided on based on **Freedman – Diaconis formula**

* Histogram graphs will be plotted with the bins in the range of 2-20
* Count graph will be plotted for all the unique values of the column

### 2. Datatype Determination

Application is reading the CSV file (dataset) using pandas library and based on that datatype will be defined-

|  |  |
| --- | --- |
| **Datatype based on pandas library** | **Datatype to be assigned on Schema Mapping** |
| float64, float32, int32, int64 | Numerical |
| object (String) | Text |
| datetime64[ns] | Timestamp |
| string of comma separated values | Categorical list |
| (unique column values/total rows) < 20% | Categorical |