Instruction Manual
Web Scraping App: YouScrape

Table of Contents

I. Introduction

1. **What is the app**

   YouScrape is a web scraping application that allows you to extract data from your preferred websites. Beside scraping the information of the websites, YouScrape also provides another function where you can visualise and customise the gathered data using different types of graph.

2. **Who is it for**

   This application is for anyone and everyone, it is designed to be able to be used by anyone wanting to do some web scraping without having any actual coding knowledge. Simply someone who wants to get and/or analyse larger amounts of data from the internet than is reasonably expectable of a user.

3. **Why we made it**

   Before making YouScrape, we had some ideas on what application we would like to make and web scraping is one of them. Web scraping is something that both of us have never tried or learned before. We wanted to try and learn new things as we made this application and combine it with skills we had already had throughout years of studying. After searching more and more about web scraping, we tried some available web scraping applications such as ParseHub and OctoParse to see how they work. After trying and testing the applications, we were inspired to make a similar application with different functions. Two URLs and data visualisation are some functions that ParseHub and OctoParse do not have.

4. **Ethical web scraping**

   There are many questions about web scraping, if taking someone else data from their website and using it for yourself is a moral thing to do. As with most things in life, the answer is not that simple. But we have found a few guiding principles that we believe can help with ensuring no ethical boundaries are crossed when web scraping.

   These are the principles:

   - Good intention, web scraping can be used to overburden and attack websites; it should be done with consideration and restraint
   - Do not copy any information that wasn't already fully publicly available and not behind any password authentication barrier.
   - Literal facts, make sure that the information being copied is factual and does not overstep anyone else's rights
   - The data copied should be used for personal use or in a transformative way, it should not be used to steal business from competitors by undercutting or directly stealing customers.

   These principles are not all encompassing, they are a good starting point to consider when web scraping. While there is nothing literally illegal about web scraping we advise

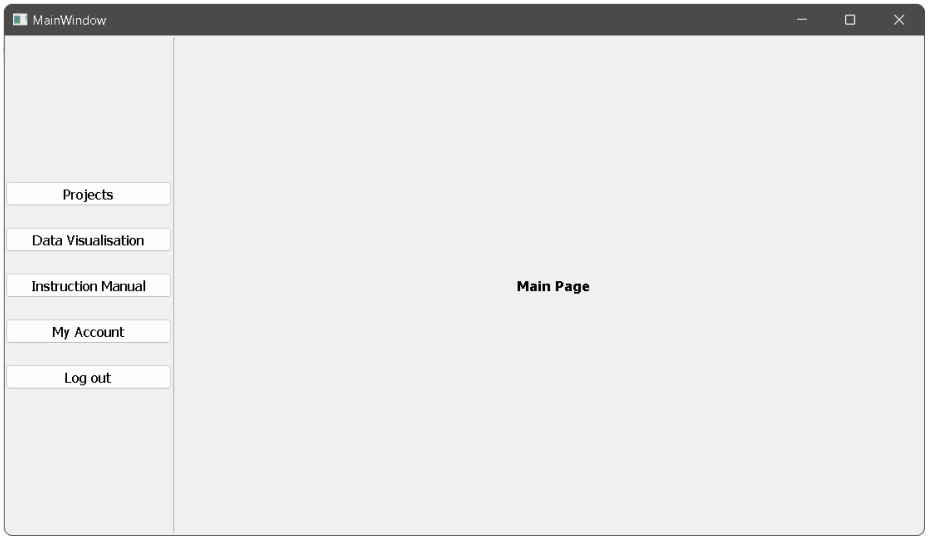
to please use wisdom and common sense when operating YouScrape. Thank you and happy scraping!

II. Projects

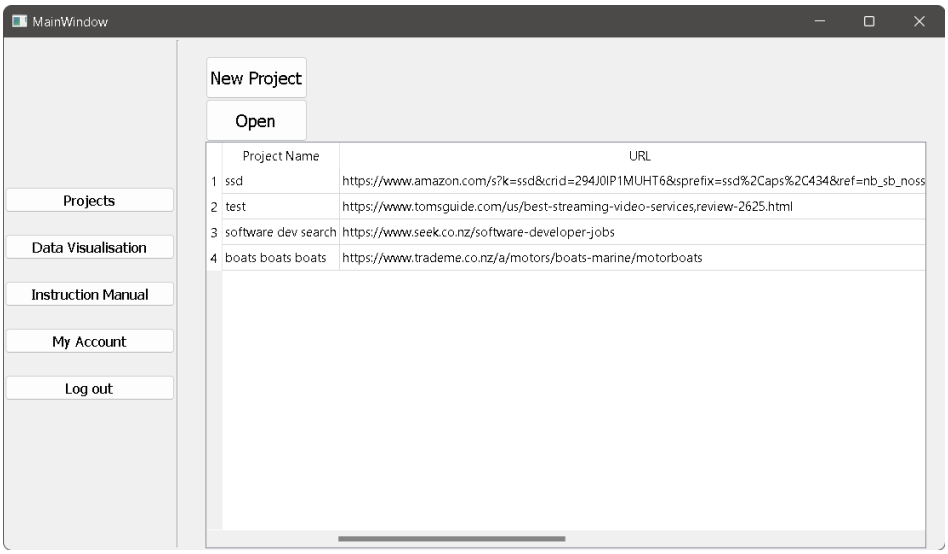
Creating a project is the main function that YouScrape provides. Now what is making a Project? What is a Project? A project for the purposes of this application is where you load whatever webpage you are interested in getting information from and you select the different elements or data that you want to scrape. That is if you wanna scrape financial data you might select the stock names and their values. Or if your interested in comparing prices of laptops in order to get a new one, then you select the brand, specifications, prices or whatever you wish to consider for your purchase.

1. Create a project

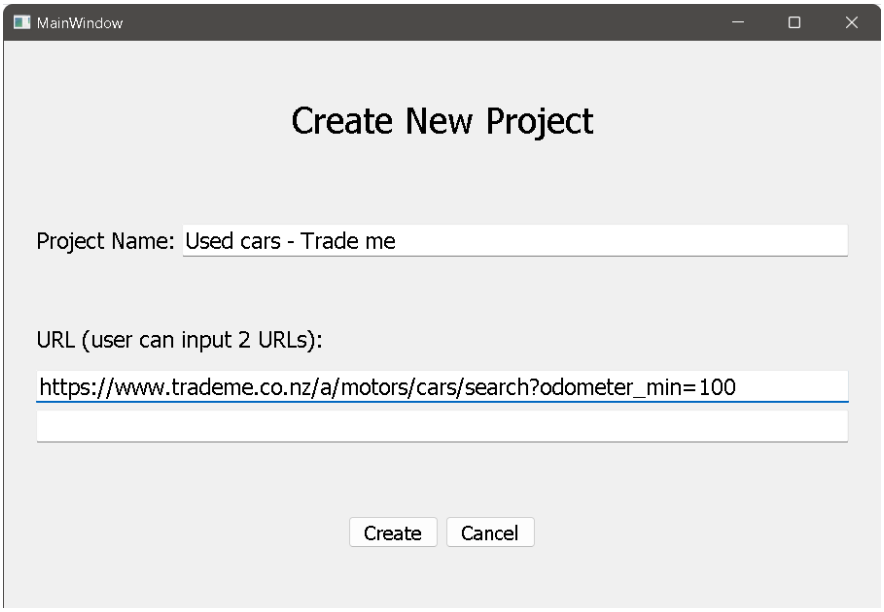
First of all you must know how to create a project. After logging into your account. You will be led to the main page. On the left side you will find the main menu. Click on the first tab labelled “Projects”:



Once you are on the Projects tab then click on “New Project”:

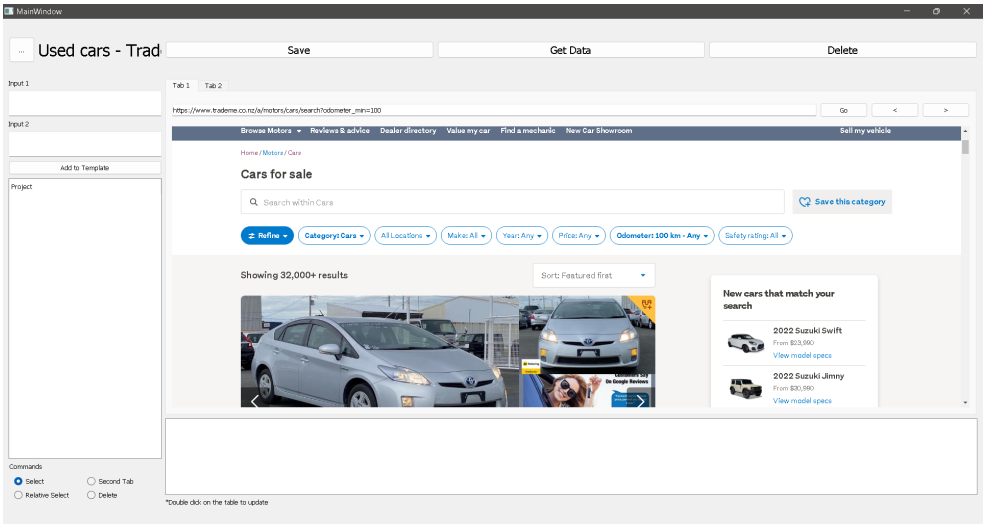


A new window will appear where you name your project and input up to 2 URLs. Try to make your Project name relevant, that way if you chrome back to it on a later date it will be easy to know what each project is. Once everything is ready click “create”:



*Note, while the url can be changed later on in the project, it is recommended that you enter in a precise url. For example see above, instead of just inputting tradem.co.nz the user used the url where the data they wants is on, in this case they are searching for cars.

Your project is now created:



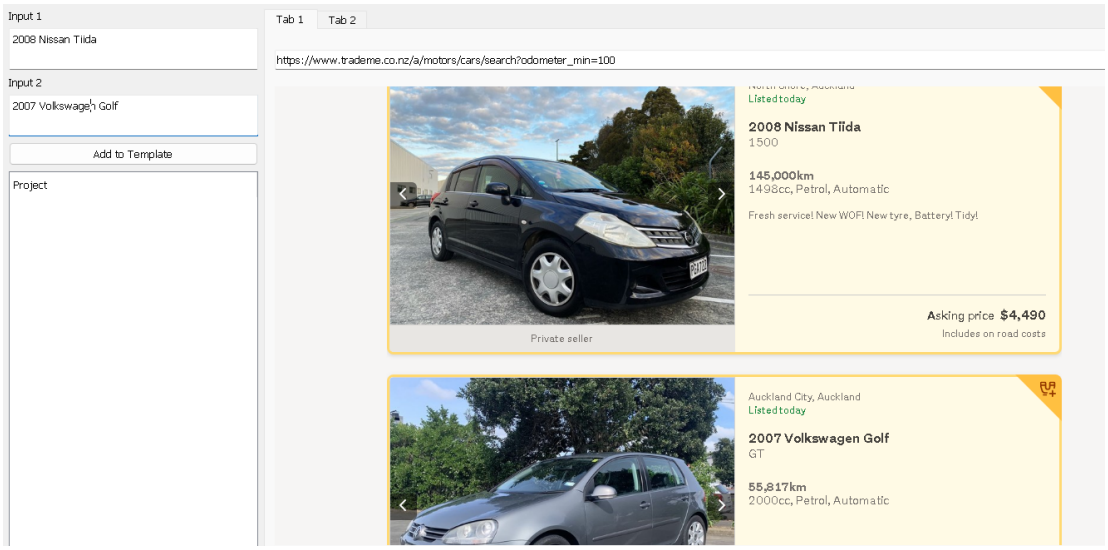
2. Edit a project

Now that you have created a Project it is important to know what you can do and how to do it.

<u>Add inputs</u>

When a project is created or when it is loaded the website(s) will automatically load on the browser section of the screen. Here you can look at your page and think about what you want to scrape.
Once you have identified what you want either copy paste the test from the webpage into the input fields or write them in yourself.

In this example we are looking at cars so we just have to take the names of two of the listing and use those:



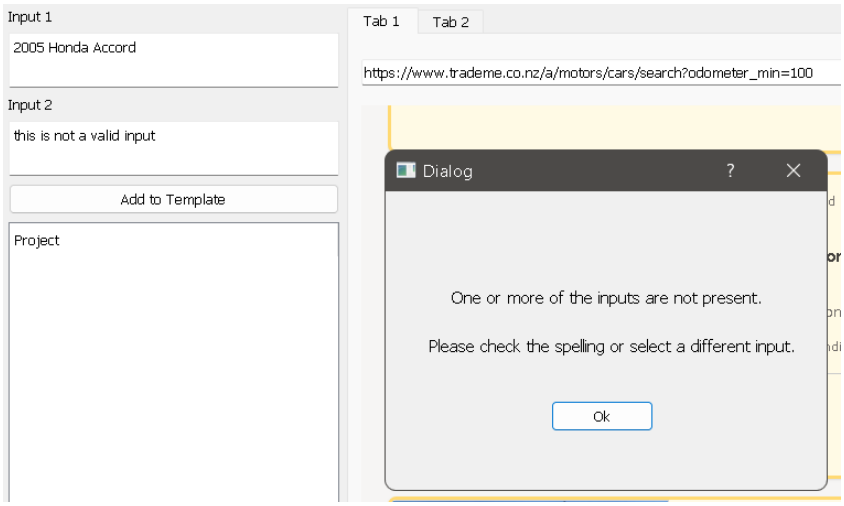
Selecting your inputs is arguably the most important part, based on these the program will find all others like them and put it into your final document. Therefore there are a few things to take note of, or just general recommendations:

<u>Input recommendations:</u>

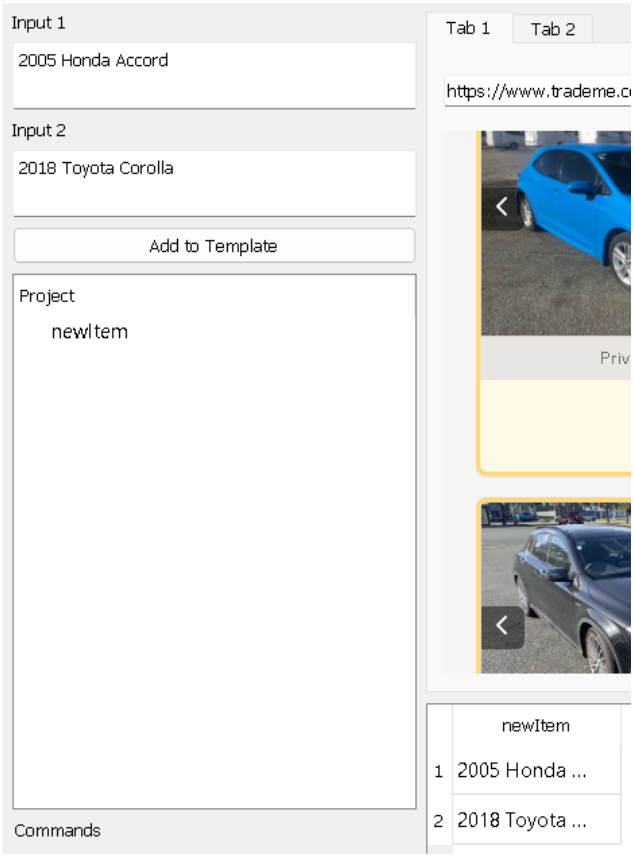
- Be Precise: make sure that you have the correct spelling to match what you see on the page. That includes capital letters and special symbols.
- The more unique the better: the program will search the page for anywhere it can find the text for the inputs, so to ensure its finding the correct elements, try and select things that do not appear anywhere else on the page.
- Avoid featured items: some pages will push certain listings to the top with special code, this code can sometimes be different from the main results and can mess up with the data scraping.
- If at first you don't succeed: in case the program can't seem to find your input just try a different one, maybe its spelt or formatted in a way that can make it difficult to identify so another input might solve your issue.
- Dynamic results: web pages today, especially ones that are selling products and services, are constantly adding and taking away options, So even if you input a valid option one day, if you save it and come back another time it might no longer be present. Just delete the old inputs and add new ones based on the changes.

After choosing your inputs click “Add to template”, the program will check the inputs to see if they are present on the webpage (this might take a minute). If they cant be found a new window will pop up and tell you the inputs are not present, just check spelling or select new ones. Once it finds them you will see a new item in the template and a new column in the preview table.

Invalid input



Valid inputs



The template

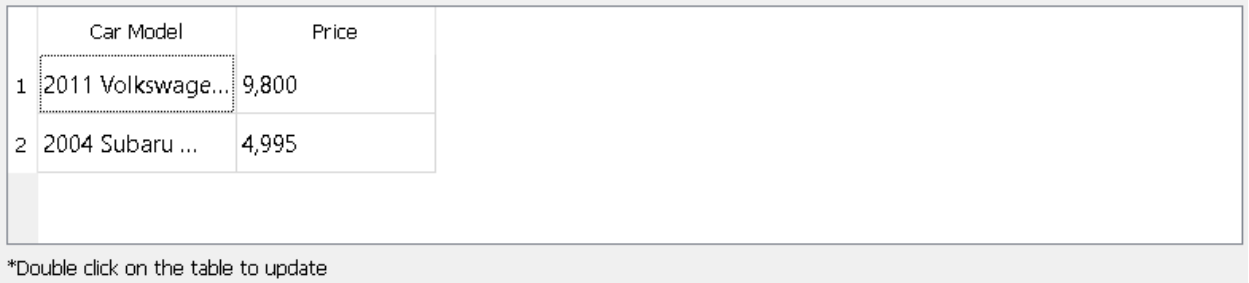
The template is the structure of your data. Here you will select how the different things you want to scrape relate to each other.
The main commands of the template can be found just underneath it:

- The *Select* command is the default command, when it is checked and you add inputs, a new item will be added to the template. It will appear with the name “newItem”, make sure you double click on it and CHANGE THE NAME. Try and make sure that every item on the template is unique.
- The next one is *Relative Select,* this command allows you to add new items to the template relative to an already existing template item. In order to use it make sure you (1) select an already existing item on the template so its highlighted and make sure (2)*relative select* is checked. Then just click (3) *add to the template* like with regular *select.* Once it is added you will see an (4)arrow next to the selected item click on it and you will see “newSubItem”, (5)edit the name to what you want.
- *Delete* is a straight forward command, check its circle and then click on the item of the template you want deleted. BE CAREFUL, make sure you click on the right item.
- Finally, *Second Tab* is only applicable in projects with two URLs. Its objective is to make sure that for every selection in the first Website, there is a corresponding selection in the second one. It works the same way as the *relative select: make* sure it's checked and the template item it corresponds to is highlighted, then *add to template*. This wont make any visible changes on the template itself but if you look at the *preview table* the inputs will be visible there.

As you work on your template you can reorganise the items into whichever order you want even moving relatively selected sub items from one item to another.

<u>The preview table</u>

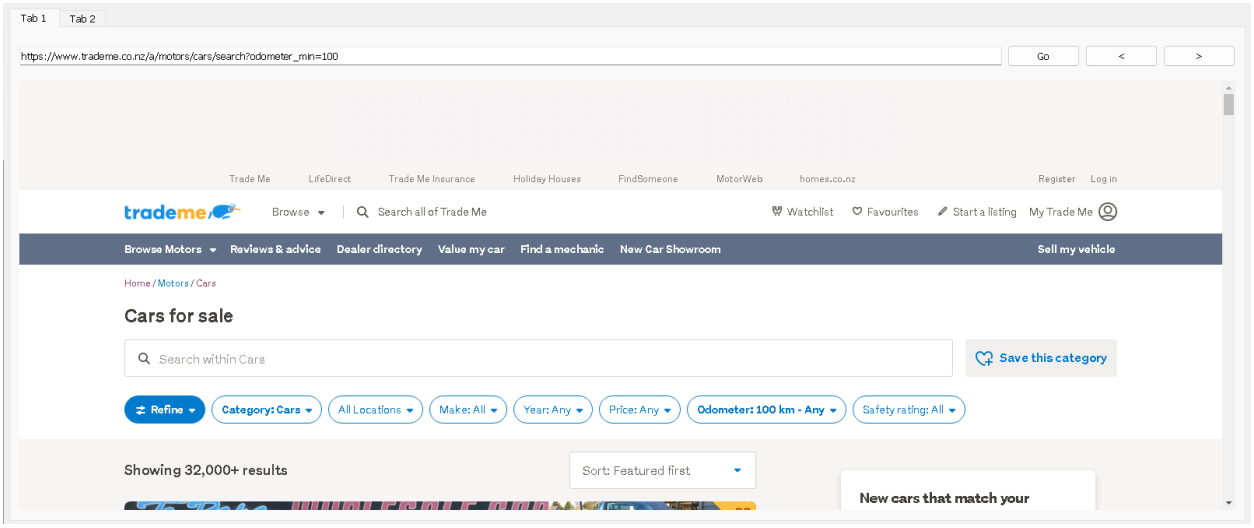
The *Preview Table* is there for your assistance and clarification. Its directly tied to the inputs and the template. Whenever you add something to the template it will appear on the table. However, if you make edits like changing the name of an item or the order that they appear just double click on the table and it will update itself.



It will show the name of the template item and the inputs you chose for it. The reason its called the preview table is because it simulates what your data will look like in a Microsoft Excel file if you choose to download the data as a csv; not in a 100% accurate capacity, just with the inputs and the item name.

Browser view

The *Browser* is where you can see the webpage(s) selected and interact with it.



It has the basic functions of any web browser: displaying the page, a navigation bar and *Go* button for going to a new page, and backwards and forwards buttons to go between pages you have navigated. If you input two URLs then you can find the second one on *Tab 2*.

If the page you input when creating the project is not working or not the one you desired then you can enter a new one and, as long as you click on the save button, the next time you open the project the new URL will still be there. Keep this in mind however if you don't want your URL to change.

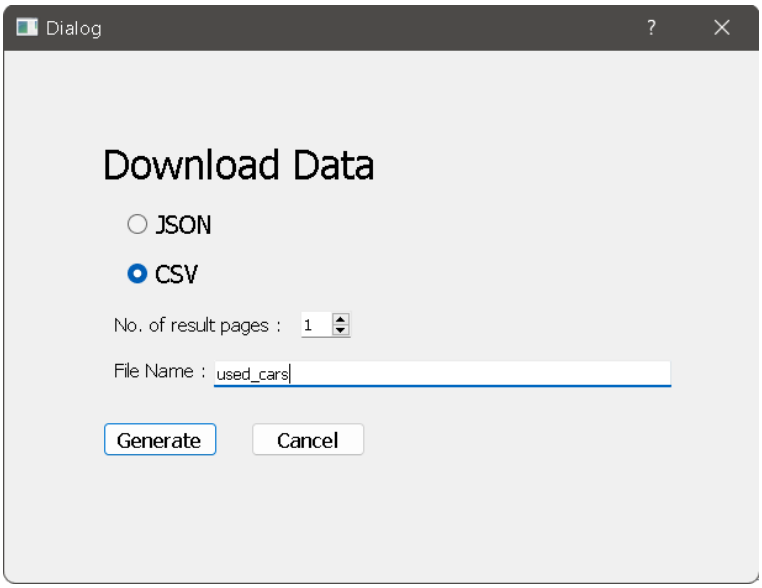
Home/Save/Delete/Get data

The buttons on the top of the project screen are all basic functions.

On the far left, next to the project name, the button with 3 dots will take you back to the projects screen, so make sure you save before clicking it.

Save and Delete refer to the project itself. Saving will save the progress made with the template and whatever URLs are in the browser view. Deleting will get rid of the entire project so make sure you are certain before deleting.

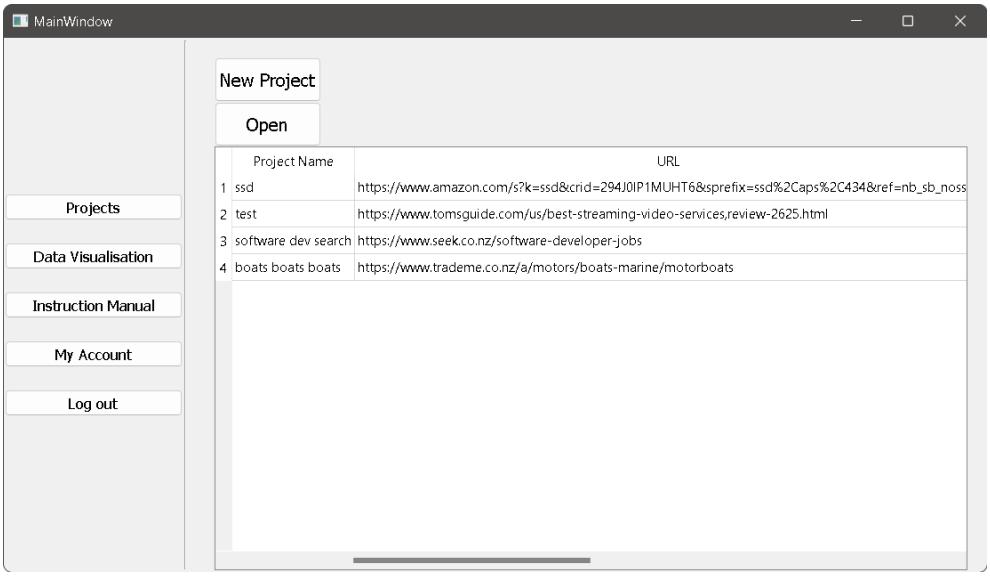
Finally, *Get Data* is where you go to download the data you want at the end of the project. It will open up a new window:



Here you select whether you want JSON or CSV type file and you give the file a name. There is also an option for number of result pages, this refers to the number of result pages from the website you want it to scrape. It has a maximum of 99 pages it can iterate through.
Once you are satisfied with all the fields click on generate and depending on how much information you are looking to scrape it will take a few minutes to complete and return the file to you.

3. **Open Project**

If you saved a project and want to return to it then you can access it from the projects tab on the main window. Just select the project you want on the table and click on Open:

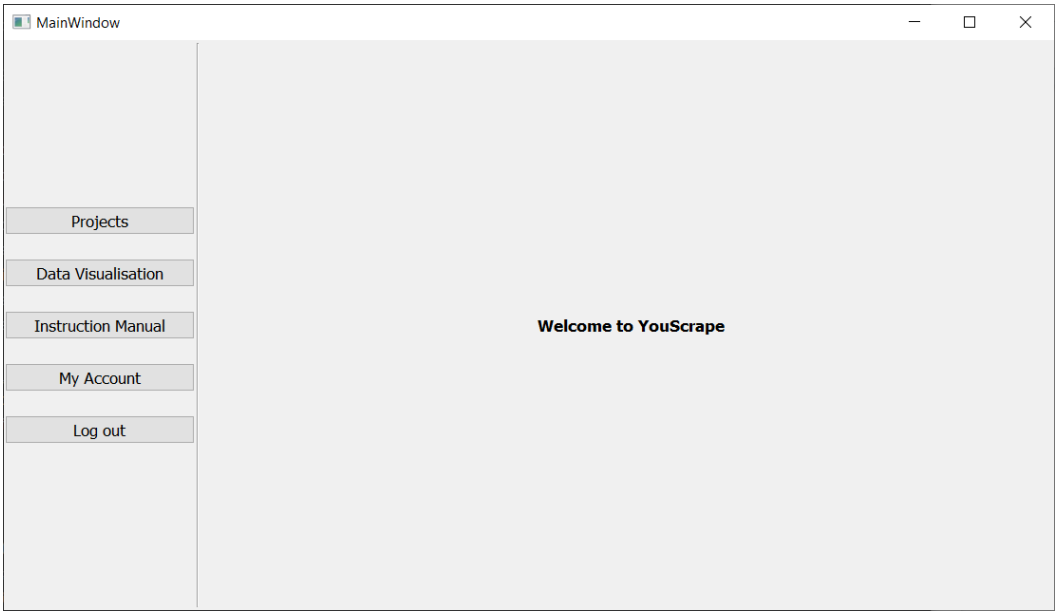


Also, if you wish to delete a project you must first open it and then delete it from the project page by clicking on the delete button you will find there.

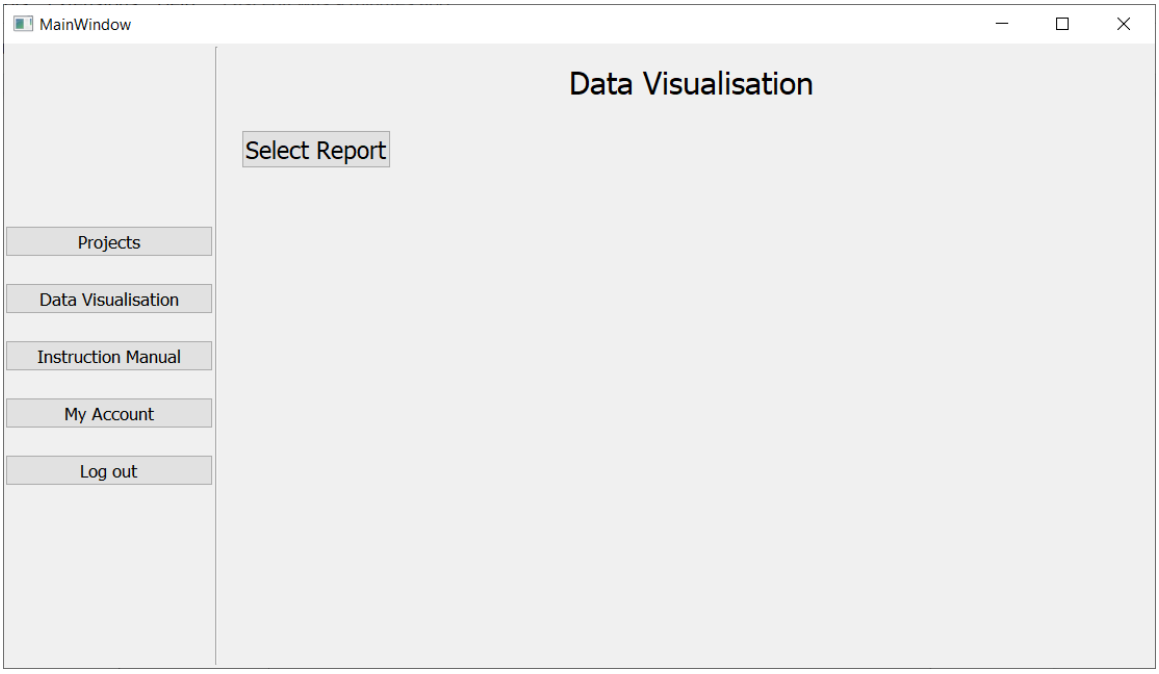
III. Reports

Another function that YouScrape provides is you can create a visualised report using generated data or any CSV or JSON file you want to visualise. This makes it easier to spot the differences between the data you choose by using a suitable graph.

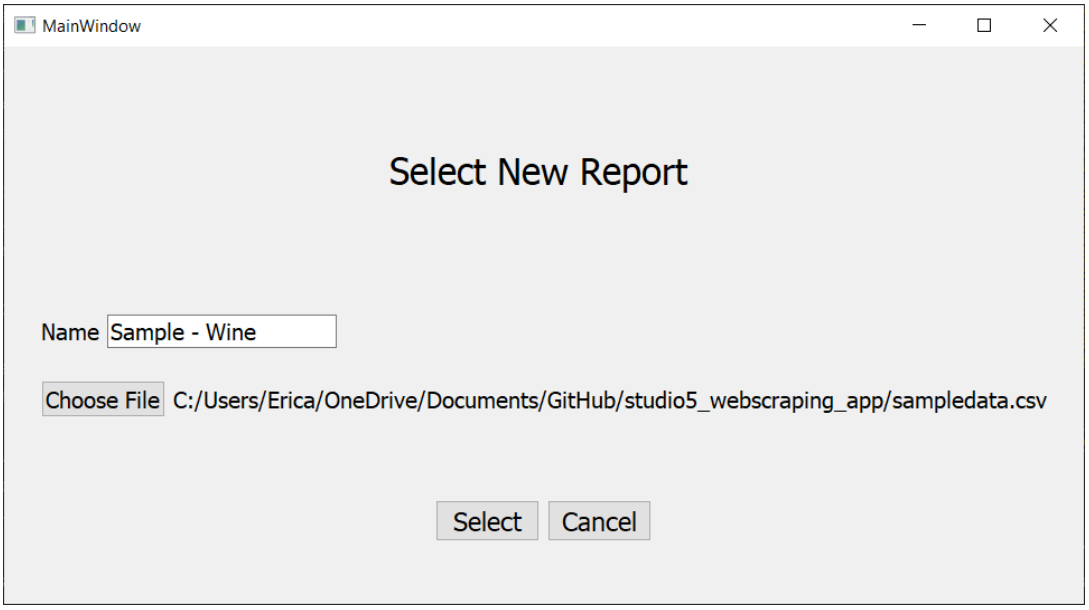
When you are in the Main Menu, you can see that we have a “Data Visualisation” tab on the left side. You can select the tab to select a new report.



1. **Create Report**

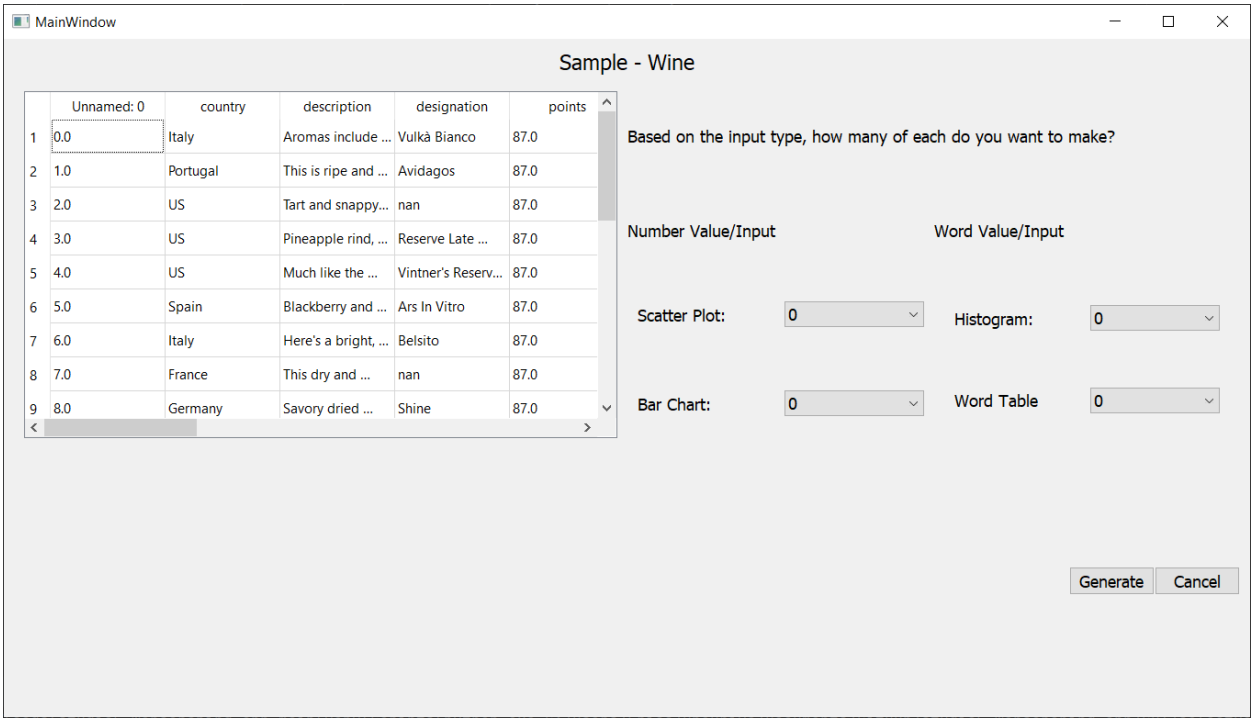


After clicking the “Select Report” button, you **have to input** the report name and select one CSV or JSON file.

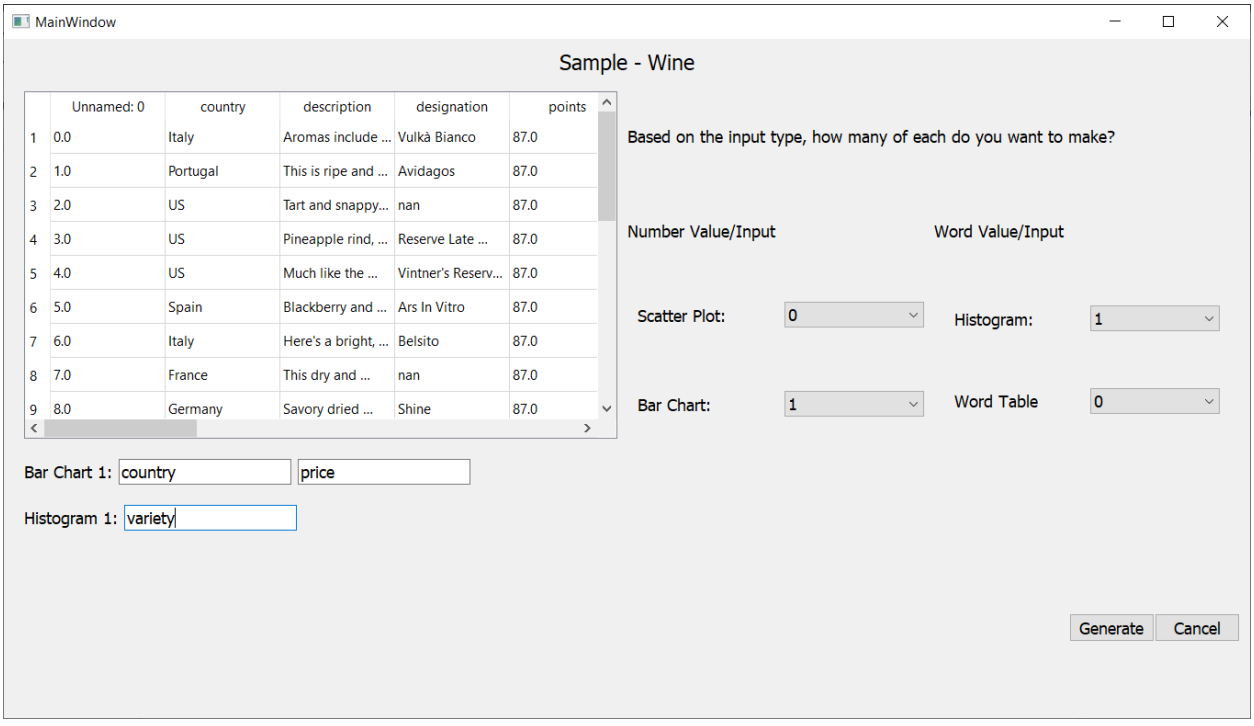


2. **Select visualisations**

You can see what is inside the CSV/JSON file in the table (top left).



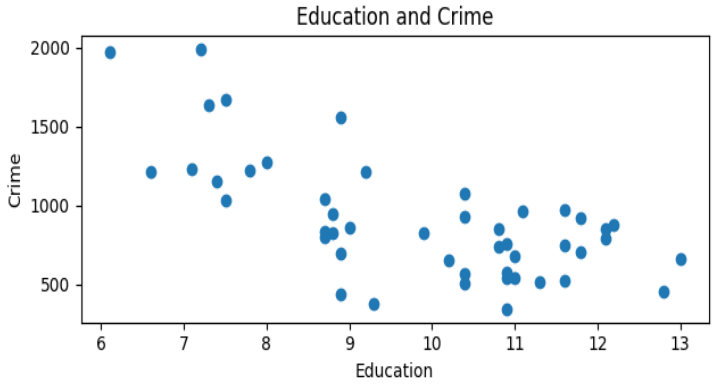
In this window, you can choose any graph that is suitable for your data with their quantity (maximum two). The graph is divided by two types, number and word input or value.



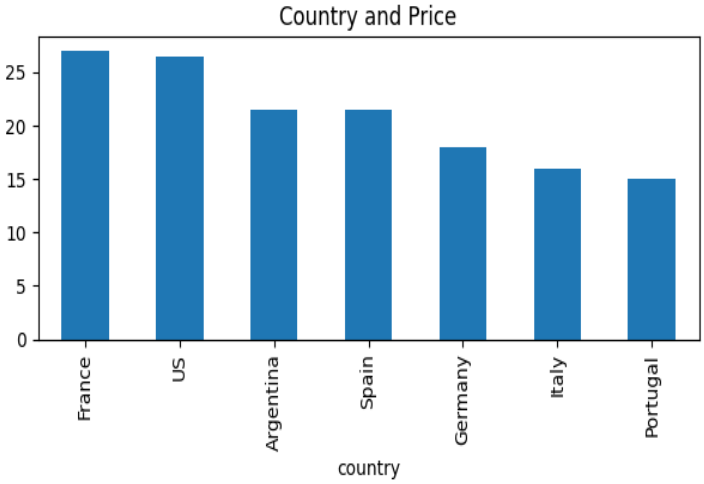
You have to input the values based on the column's name (name can be seen on the top left table). **Make sure** to at least select one of the graphs to generate the report.

Graph Types

- Number Value/Input
    1. Scatter plot, user will have to input two numeric values. You can use Scatter plot to show the correlations between the values. For example, you have data about crime and education. You would like to see whether higher levels of education affect the crime rate. In that case, scatter plot is the best type of graph to present your data.



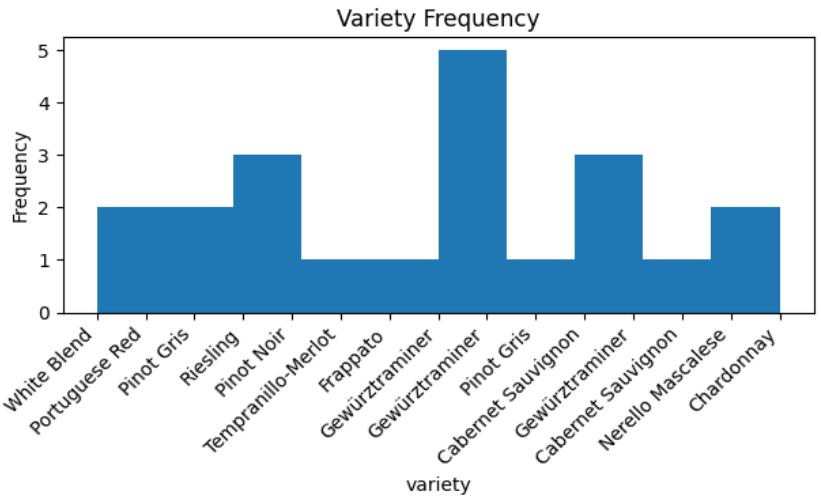
    2. Bar chart, user can input two numeric values or one numeric and one word/string value. Bar chart can be used to show two different values and categories



- Word Value/Input

  Both Histogram and Word table only need one input/value

    1. Histogram, can be used if you want to see the frequency of each word

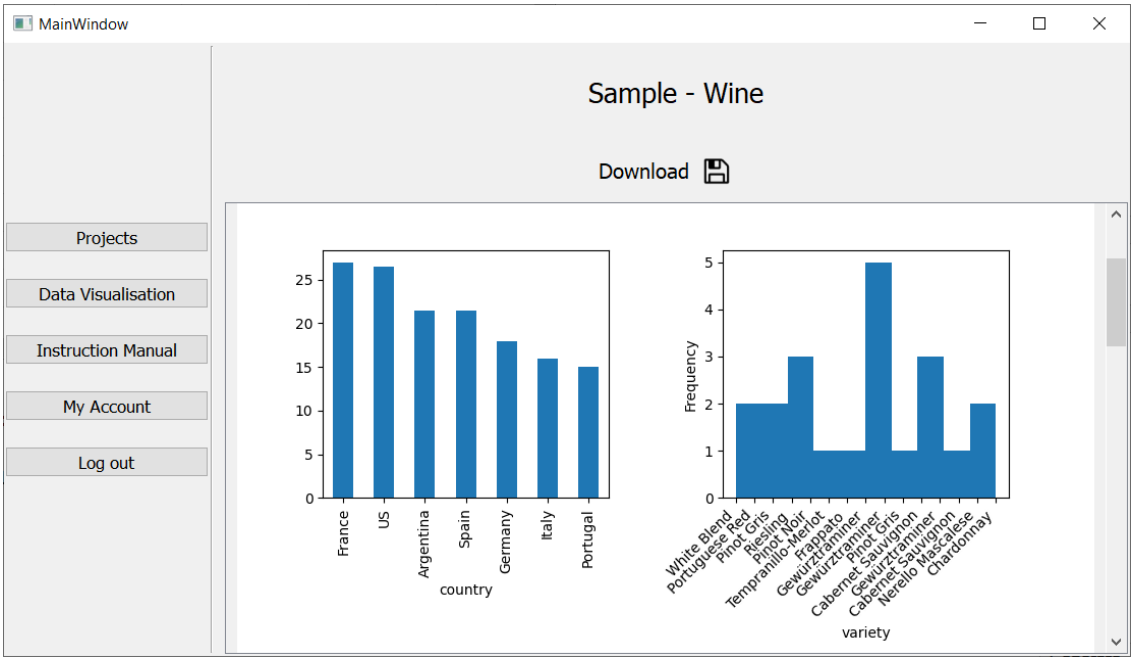


2. Word table, will split and count how many of the same words

Job title	Frequency
developer	33
software	24
engineer	5
full	5
end	5
stack	5
net	4
web	3
senior	3
intermediate	3
front	3
cnet	2
jobs	2
development	2
back	2
lead	2
tech	1
company	1
team	1
typescript	1
systems	1
intermediatesenior	1
contract	1
react	1
native	1
awardwinning	1
principal	1
assistant	1
developer2	1
operations	1
nz	1
juniorgraduate	1
golang	1
embedded	1
roles	1
amp	1
fullstack	1
remote	1
mobile	1

3. **Save/edit**

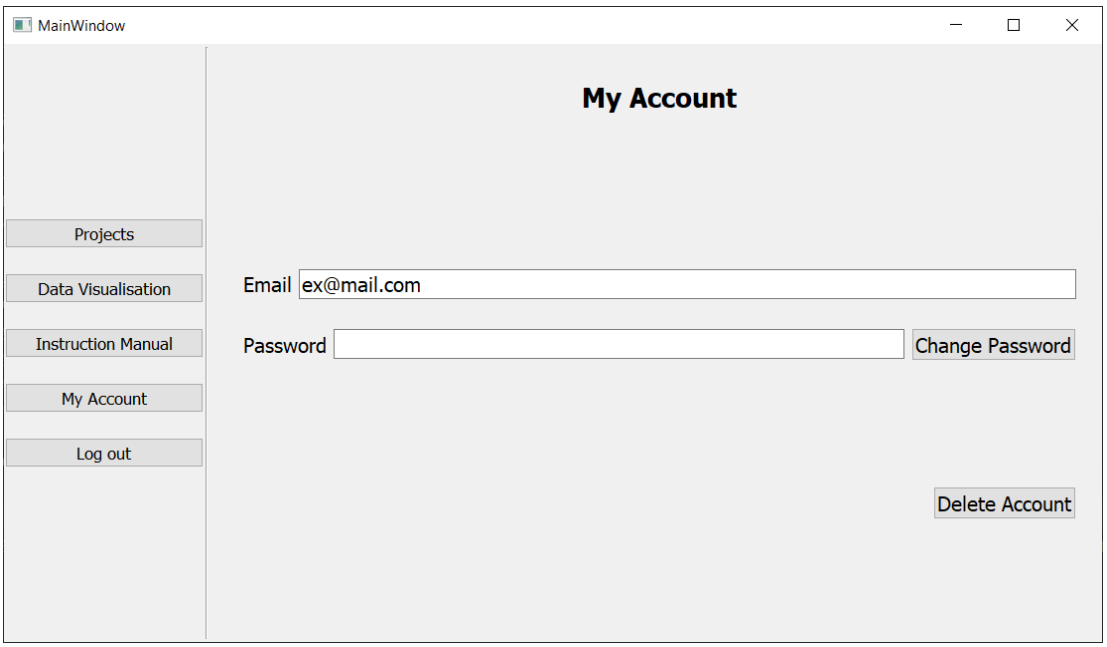
After choosing and inputting values based on the input type, you can generate the report and will be directed to the report page where you can see their chosen graph(s).



The generated report will not be saved in your database. If you want to keep the report, you can download it as Image (PNG or JPG). You can also go back to the previous page if you want to delete or add a certain graph.

IV. Account

Each user has their email address as their username and password to log in. If user does not have an account, they can sign up and log in with the new account.



1. **Change Password**
   Passwords can be changed by user while the email address cannot be changed.
2. **Delete Account**
   If you would like to delete your account, you have to be aware that all of your projects that have been created will be deleted.