UNIVERSITI MALAYA

Technical Report

WIX1002

Fundamental of Programming Assignment

Semester 1 2020/2021

Project 6: Donald Duck's Digital Data Duct

Donald Duck decides he's going to take up the coolest job of the decade, Data Scientist (debatable, chicken rice sellers are pretty much elite class in the UM hierarchy). But he is new to programming so he can't do it from scratch. He needs a library. And you'll be the ones to build it for him.You are required to develop a library to help Donald on his Data Science journey.

Tutorial & Lab Group: 1

Lecturer: Assoc. Prof. Dr. Ang Tan Fong

Prepared by:

Name	Matric No.
YAU DE MIN	U2005347/1
KONG JING YUAAN	U2005395/1
HONG ZHAO CHENG	U2005280/1
CHIEW ZHE WEI	U2005368/1
LOK JIA HUI	U2005425/1

TASK

We are assigned to write a program to develop a library. Our library is designed to help Donald on his Data Science journey. This library is designed with some additional features too.

TASK REQUIREMENTS

There are several methods which are needed in this library.

1. **DataFrame Object A DataFrame object to interface with the CSV file parsed.**

a. Method to save DataFrame to CSV file

b. Method to construct DataFrame from CSV file

2. **Manipulation methods**

a. Method to concatenate DataFrames.

- Column concatenation means stacking columns, whereas row concatenation means stacking rows.
- If the combining axis doesn't match, the program should provide an error to tell Donald he made a mistake.

b. Method to obtain a subset of DataFrame with range of row or column.
Rows are 0 indexed and the range is inclusive of the first element but exclusive of the last.

c. Method to sort the rows by a column in the DataFrame.

d. Method to remove duplicate rows based on the subset of columns.

- There should be a parameter to choose whether to keep the first, last, specific number or no occurrence.

e. Method to remove rows containing missing data in subset of columns

3. Statistics package and imputers

a. Method compute variance, standard deviation, min, max,mean, median, mode and range of a column

- Non numeric columns will only have mode

b. Method to fill in missing values of specified columns with a specified value.

4. **Scalers**

a. Method to perform Standard Scaling.

- Standard scaling is subtracting the mean from all values in the column and dividing by the standard deviation.

b. Method to perform Min Max Scaling

- Min max scaling is subtracting the min from all values in the column and dividing by the range

5. **K-Nearest Neighbors (k-nn)**

- K-nearest neighbors is a simple prediction algorithm that uses k-nearest known instances to try and predict unknown instances.
- For this question, k-nn is based on euclidean distance.
- Euclidean distance between 2 points is the root mean squared of differences between 2 points.

- Regressor: output the mean of the instances
- Classification: output the mode of the instances

a. Method to impute values for a column using k-nn regressor using subsets of other columns.

b. Method to impute values for a column using k-nn classifier using subset of other columns

6. **Error metric**

- Donald doesn't like errors so he isn't going to be particular about what we choose, he wants us to explore them on our own and implement what we think he should have.

APPROACH

We have designed a library for Donald so that he can store data in the library throughout his journey on Data Science. All the requirements are fulfilled and we have implemented several extra features in the library which are web scraper, bar chart generator, JavaScript Object Notation (JSON) output and data browser. The main reason we would implement these additional features in our system is to further enhance our system and ensure that Donald is able to use this library efficiently. Next, JSON output can also help in enhancing the efficiency of the system as it will parse the data easily and execute data in a faster way. As we all know, searching for valid data easily is the most important thing in a system, so we have decided to implement an additional feature, namely a data browser. Therefore, users can search the data they want just by entering a specific name and they will be able to get the row of the data of the name. As a result, the specific data will be stored directly to a new csv file. In short, this system is especially useful with the implementation of these additional features and these features are playing a crucial role in ensuring the system is always functioning effectively.

To make our system more user friendly and easy to use , our system will first prompt the users to enter one of the 8 steps according to their needs. There are 8 steps available which are adding new data (1), managing data (2), statistics (3), scalers (4), K-Nearest Neighbours (5), generate charts (6), convert file type to json (7) and web scraping (8). Users have the option which is entering ‘-1’ to exit the program.

In our system, if you have entered the wrong step, you can enter ‘-1’ to return to the previous step. This will prevent the users from wasting time to run the program again. Furthermore, the system will always prompt the users to enter ‘1’ which indicates “Confirm and will proceed to the next step” or ‘-1’ indicates ‘Return to previous step’ to make sure their decision is made accurately and precisely.

For adding new data, our system will allow users to either display data from a csv file or key in data. If you want to display data from a csv file, you are required to enter the name of the csv file. Apart from that, if you want to key in data, the system will prompt you to enter the number of rows and columns of data that you want to store and the name of the file you want to store the data.

Besides that, our system also enables the users to manage data in which you can concatenate data, obtain any specific data that you want, sort data in ascending or descending order, remove duplicated data in a row or remove data with null value and even browse data. Apart from that, for scaler, you can decide on what type of scaling you want to have. In our system, you can choose to have standard scaling or minimum and maximum scaling. Before deciding on what type of scaling you want to have, the system will prompt you to enter the file name, the column where the data located at and the values of the data. After that, the system will display the result of the values after scaling.

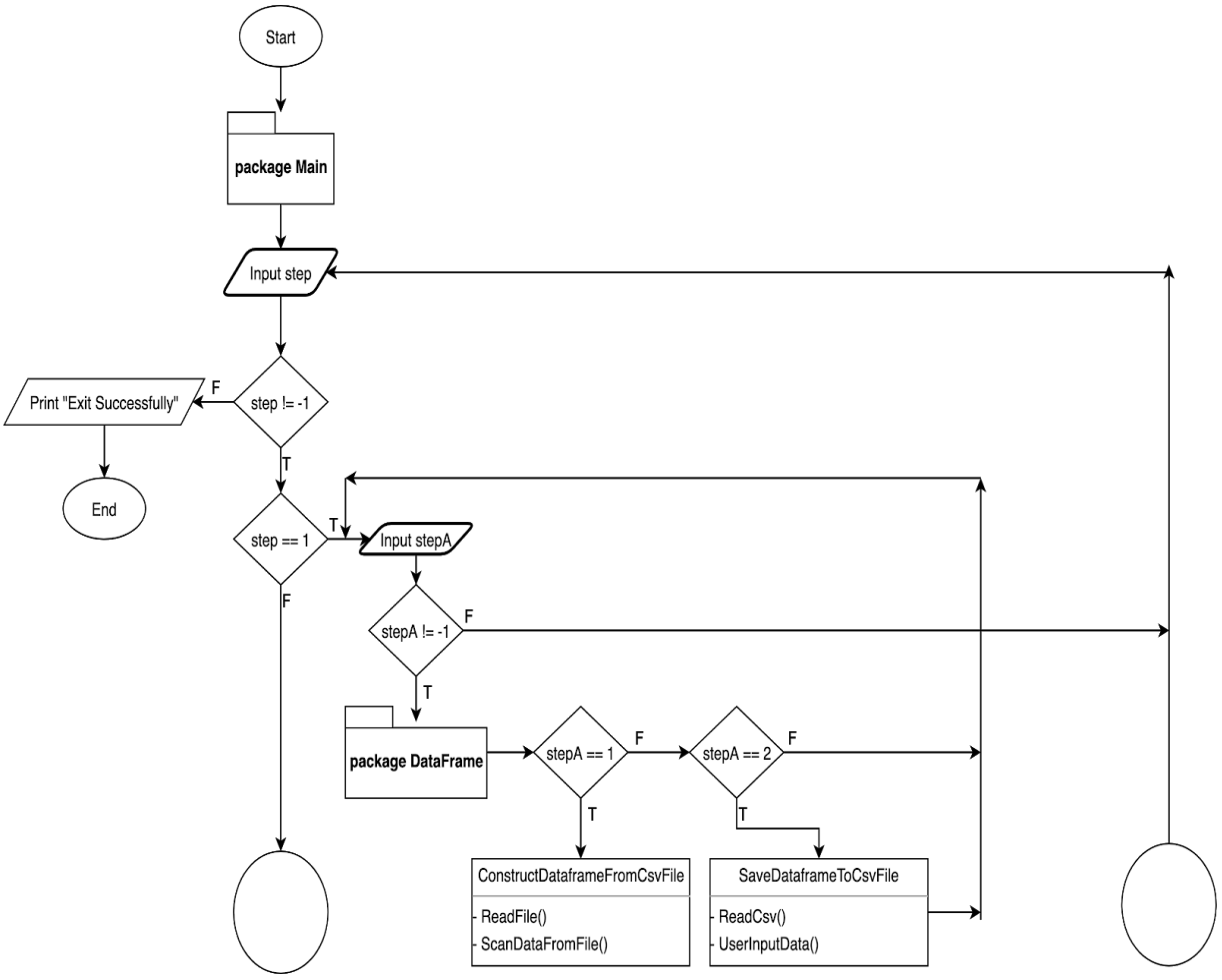
Next, users can also obtain the statistics easily in this system. You can choose to obtain overall statistics or fill missing values. In order to get overall statistics, you just need to enter the file name and the statistics of your data will be displayed. If you want to fill in missing values, you need to enter the file name so the system can detect whether there are any missing values in the file. To make our system more user friendly, it will show a line of words “There is no missing value” to let the users know there is no missing value. After that, the users can also choose to generate a JSON file if they want but this is not compulsory. As mentioned earlier, the JSON file will be generated simply because the execution of data using JSON file is faster.

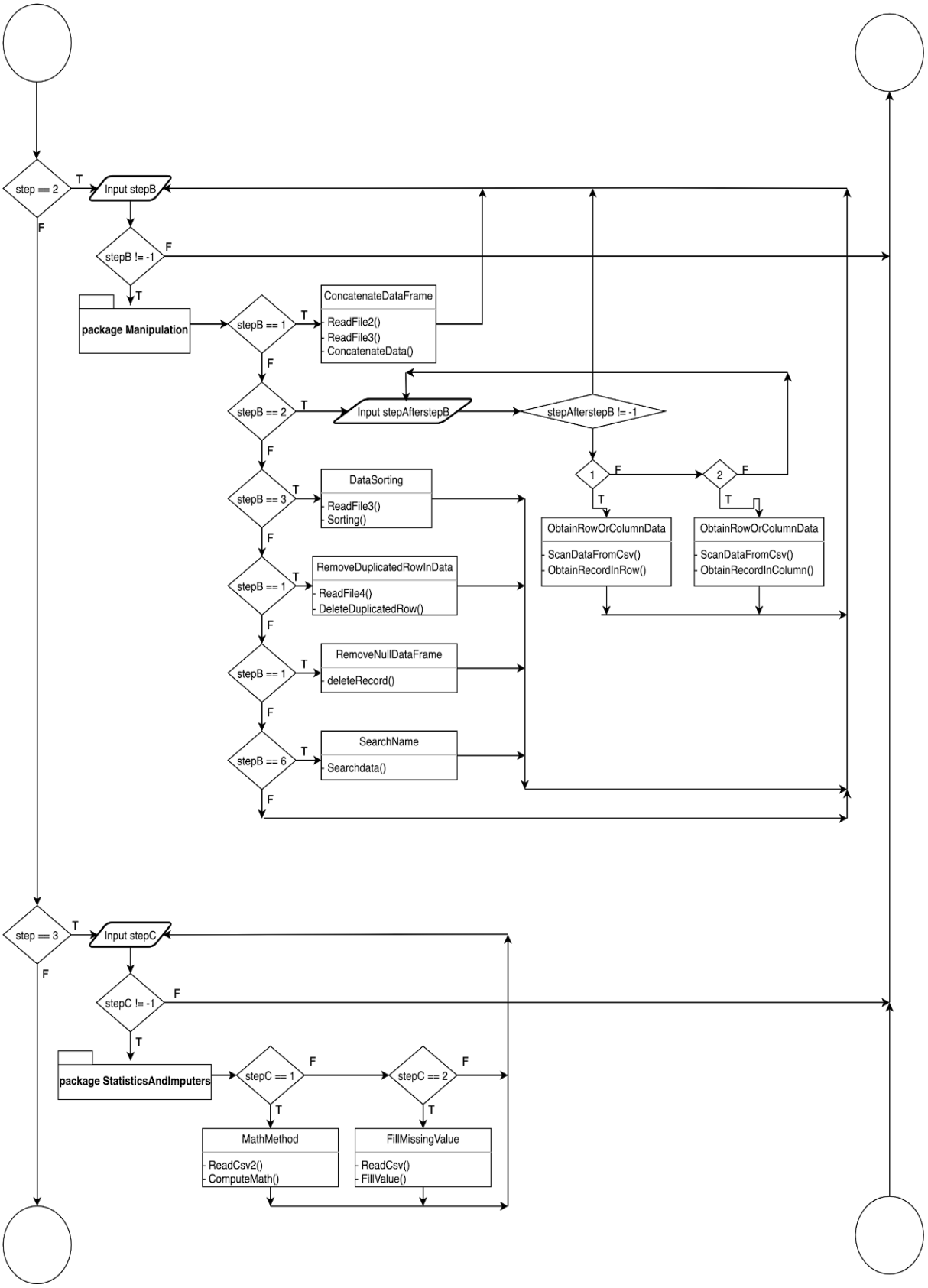
Moreover, users can also generate a bar chart using a bar chart generated which has been built in our system. If you want to generate a bar chart, you just need to enter the name of your csv file so that our system can access the data in the file and analyse the data in a short time of period and then display the bar chart to you. Furthermore, we have also implemented K-Nearest Neighbours (k-nn). This feature is basically to use a database in which the data points are separated into several classes to predict the classification of a new sample point. Besides that, KNN regression will approximate the association between the independent variables and the continuous outcomes by averaging the observations in the same neighbourhood.

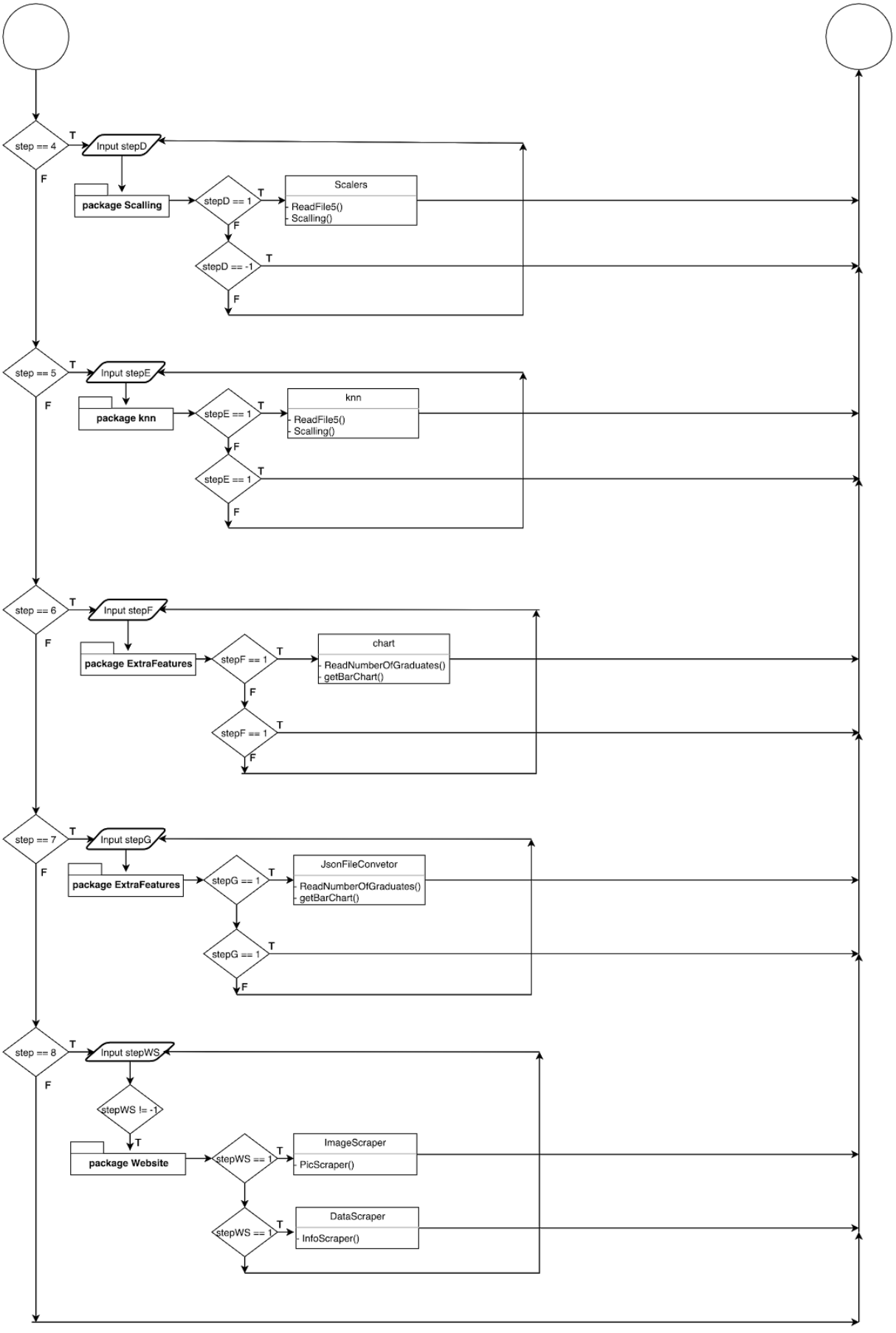
Lastly, we have implemented web scraping in our system. This feature is very important as there are users who may want to scrape the image or data from a website. If you want to get an image or information from a website, you need to enter the website URL. Then, our system will prompt you to enter a name of the file in which the data or image will be stored. The file type to store the data is usually a txt or csv file.

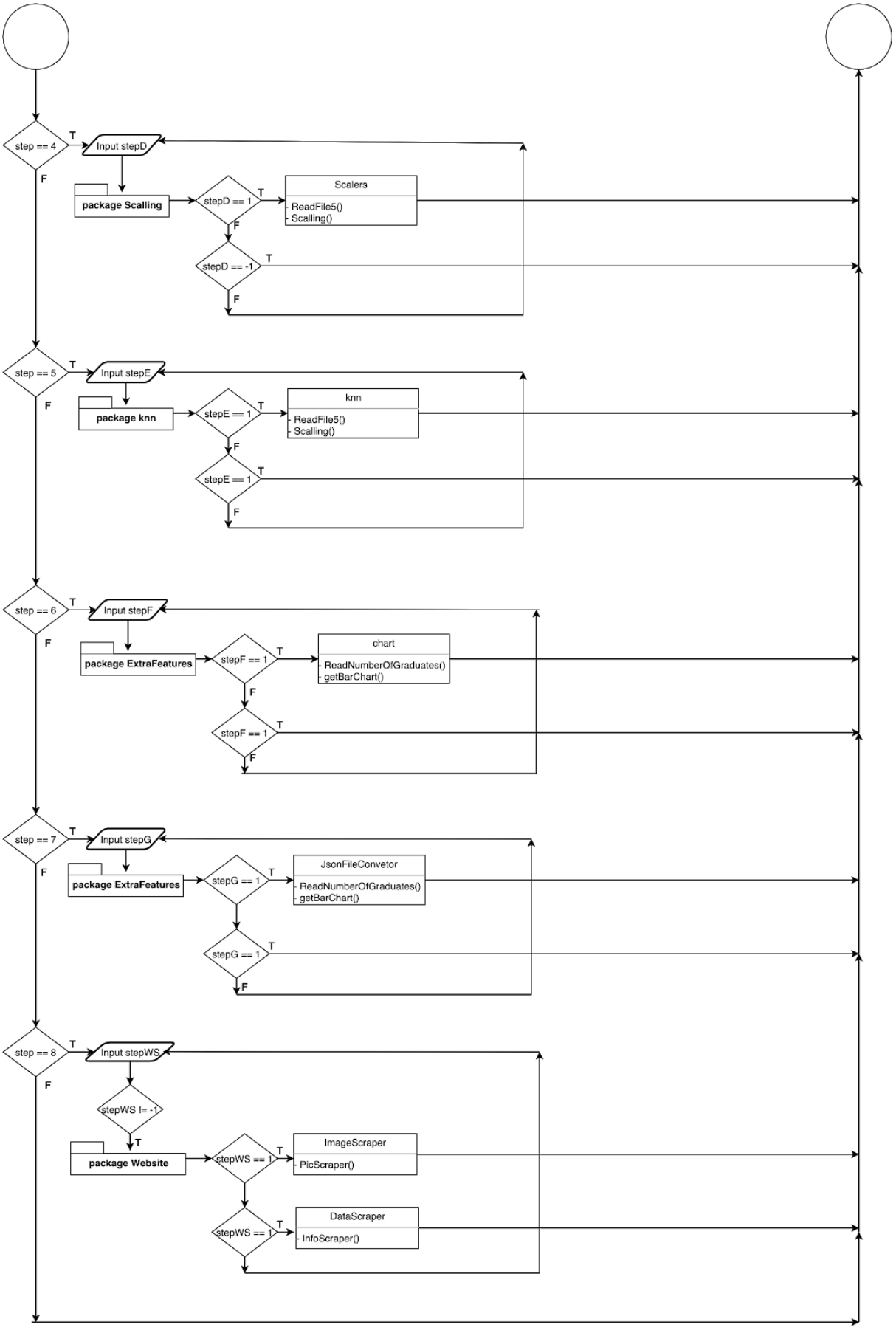
SOLUTION DESCRIPTION

Flowchart




step == 2
Input stepB
stepB != -1
package Manipulation
stepB == 1
ConcatenateDataFrame
ReadFile2()
ReadFile3()
ConcatenateData()
stepB == 2
Input stepAfterstepB
stepAfterstepB != -1
1
2
stepB == 3
DataSorting
ReadFile3()
Sorting()
ObtainRowOrColumnData
ScanDataFromCsv()
ObtainRecordInRow()
ObtainRowOrColumnData
ScanDataFromCsv()
ObtainRecordInColumn()
stepB == 1
RemoveDuplicatedRowInData
ReadFile4()
DeleteDuplicatedRow()
stepB == 1
RemoveNullDataFrame
deleteRecord()
stepB == 6
SearchName
Searchdata()
step == 3
Input stepC
stepC != -1
package StatisticsAndImputers
stepC == 1
stepC == 2
MathMethod
ReadCsv2()
ComputeMath()
FillMissingValue
ReadCsv()
FillValue()


step == 4
T
F
Input stepD
package Scalling
stepD == 1
Scalers
- ReadFile5()
- Scalling()
stepD == -1
step == 5
Input stepE
package knn
stepE == 1
knn
- ReadFile5()
- Scalling()
stepE == 1
step == 6
Input stepF
package ExtraFeatures
stepF == 1
chart
- ReadNumberOfGraduates()
- getBarChart()
stepF == 1
step == 7
Input stepG
package ExtraFeatures
stepG == 1
JsonFileConvetor
- ReadNumberOfGraduates()
- getBarChart()
stepG == 1
step == 8
Input stepWS
stepWS != -1
package Website
stepWS == 1
ImageScraper
- PicScraper()
stepWS == 1
DataScraper
- InfoScraper()


step == 4
T
F
Input stepD
package Scalling
stepD == 1
Scalers
- ReadFile5()
- Scalling()
stepD == -1
step == 5
Input stepE
package knn
stepE == 1
knn
- ReadFile5()
- Scalling()
stepE == 1
step == 6
Input stepF
package ExtraFeatures
stepF == 1
chart
- ReadNumberOfGraduates()
- getBarChart()
stepF == 1
step == 7
Input stepG
package ExtraFeatures
stepG == 1
JsonFileConvetor
- ReadNumberOfGraduates()
- getBarChart()
stepG == 1
step == 8
Input stepWS
stepWS != -1
package Website
stepWS == 1
ImageScraper
- PicScraper()
stepWS == 1
DataScraper
- InfoScraper()

TECHNICAL DIFFICULTIES

Difficulties	Ways to overcome
Big amount of data at output in java - All methods involve displaying the whole csv file, running multiple methods subsequently will causes the output to be cramped	Only display full table outputs in java for those methods that alter and update the csv file. For other methods, the output can be obtained from the csv file in the project folder.
InputMismatchException - When user's input format is wrong, (Ex: int to String) the program will encounter error and start over	Use try and catch to catch InputMismatchException, use while loop to loop until the user enters a valid input.
Scanner bug - After scanning an integer, the next scanner for String will be skipped.	Use two scanners, one for scanning integer values, one for scanning String values to save us the hassle of capturing the enter key.
Some methods require specific conditions of csv file - Concatenate row and column method will only run when 2 file have same number of columns and rows	Prepare the required csv files beforehand to ensure the method runs smoothly. Make sure the csv we are using is appropriate for the method. For example, knn methods require a big amount of data for the prediction to be precise.
Large amount of coding from various group members. - Vast amount of codes makes it impractical to use copy paste in compiling the whole project.	Use Github application for cooperation between members, each member can do their coding in the same project repository.

SAMPLE INPUT OUTPUT

2. Manipulation Method:

a) Method to concatenate DataFrames

Main File Before Row Concatenation

	A	B	C	D	E	F
1	Name	Department	Current CGPA	Expected Graduation Salary	Actual graduation salary	
2	Meow	pharmacy	4.89	1000000	4500	
3	Woof	Software Engineering	3	4200	6788	
4	LWY	medicine	4	1000	1000000	
5	kong	datascience	2	4500	1000	
6						

SubFile to Concatenate Row

	A	B	C	D	E	F
1	Name	Department	Current CGPA	Expected Graduation Salary	Actual Graduation Salary	
2	Chiew Zhe Wei	Artificial Intelligence	4	5000	2101	
3	Tan	Data Sceince	2.89	3000	2010	
4						

Main File After Row Concatenation

	A	B	C	D	E	F
1	Name	Department	Current CGPA	Expected Graduation Salary	Actual graduation salary	
2	Meow	pharmacy	4.89	1000000	4500	
3	Woof	Software Engineering	3	4200	6788	
4	LWY	medicine	4	1000	1000000	
5	kong	datascience	2	4500	1000	
6	Chiew Zhe Wei	Artificial Intelligence	4	5000	2101	
7	Tan	Data Sceince	2.89	3000	2010	
8						

Main File Before Column Concatenation

	A	B	C	D	E	F
1	Name	Department	Current CGPA	Expected Graduation Salary	Actual graduation salary	
2	Meow	pharmacy	4.89	1000000	4500	
3	Woof	Software Engineering	3	4200	6788	
4	LWY	medicine	4	1000	1000000	
5	kong	datascience	2	4500	1000	
6	Chiew Zhe Wei	Artificial Intelligence	4	5000	2101	
7	Tan	Data Sceince	2.89	3000	2010	
8						

SubFile to Concatenate Column

	A	B	C
1	Married	Age	
2	yes	18	
3	no	18	
4	yes	18	
5	no	20	
6	yes	20	
7	yes	18	
8			

Main File After Column Concatenation

	A	B	C	D	E	F	G	H
1	Name	Department	Current CGPA	Expected Graduation Salary	Actual graduation salary	Married	Age	
2	Meow	pharmacy	4.89	1000000	4500	yes	18	
3	Woof	Software Engineering	3	4200	6788	no	18	
4	LWY	medicine	4	1000	1000000	yes	18	
5	kong	datascience	2	4500	1000	no	20	
6	Chiew Zhe Wei	Artificial Intelligence	4	5000	2101	yes	20	
7	Tan	Data Sceince	2.89	3000	2010	yes	18	
8								

b) Method to obtain a subset of DataFrame with range of row or column

Data to Obtain Row and Column

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	Meow	Artificial Intelligence	3.7	1000000	1000	
3	Woof	Software Engineering	3	4200	4200	
4	LWY	Information system	4.3	1000	1000000	
5	kong	datascience	2	10000	1000	
6						

Obtained Row

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	Meow	Artificial Intelligence	3.7	1000000	1000	
3	LWY	Information system	4.3	1000	1000000	
4	kong	datascience	2	10000	1000	
5						

Obtained Column

	A	B	C
1	Department	Actual Graduation Salary	
2	Artificial Intelligence	1000	
3	Software Engineering	4200	
4	Information system	1000000	
5	datascience	1000	
6			

c) Method to sort the rows by a column in the DataFrame

Data to be Sorted

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	Meow	Artificial Intelligence	3.7	1000000	1000	
3	Woof	Software Engineering	3	4200	4200	
4	LWY	Information system	4.3	1000	1000000	
5	kong	datascience	2	10000	1000	
6						

Sorted Data (e.g. ascending order of name)

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	kong	datascience	2	10000	1000	
3	LWY	Information system	4.3	1000	1000000	
4	Meow	Artificial Intelligence	3.7	1000000	1000	
5	Woof	Software Engineering	3	4200	4200	
6						

d) Method to remove duplicate rows based on a subset of columns

Duplicated Name Data

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	kong	datascience	2	10000	1000	
3	Xavier Tan	datascience	2.5	4000	4200	
4	LWY	Information system	4.3	1000	1000000	
5	Xavier Tan	Multimedia	2.5	4000	5000	
6	Meow	Artificial Intelligence	3.7	1000000	1000	
7	Woof	Software Engineering	3	4200	4200	
8	Xavier Tan	Software Engineering	3	4000	1000000	
9						

Data After Remove Duplicated Row (e.g. keeping Row 4 - Xavier Tan)

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	kong	datascience	2	10000	1000	
3	LWY	Information system	4.3	1000	1000000	
4	Xavier Tan	Multimedia	2.5	4000	5000	
5	Meow	Artificial Intelligence	3.7	1000000	1000	
6	Woof	Software Engineering	3	4200	4200	
7						

e) Method to remove row containing missing data in subset of columns

Null Value Data

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	kong	datascience	2	10000	1000	
3	LWY	Information system	4.3	1000	1000000	
4	Xavier Tan	Multimedia	2.5		5000	
5	Meow	Artificial Intelligence	3.7	1000000	1000	
6	Woof	Software Engineering	3	4200	4200	
7						

Data After Remove Null Value

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	kong	datascience	2	10000	1000	
3	LWY	Information system	4.3	1000	1000000	
4	Meow	Artificial Intelligence	3.7	1000000	1000	
5	Woof	Software Engineering	3	4200	4200	
6						

3. Statistics Package and Imputers:

a) Method compute variance, standard deviation, min, max, mean, median, mode and range of a column

Data to be Converted to Statistics

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	kong	datascience	2	10000	1000	
3	LWY	Information system	4.3	1000	1000000	
4	Meow	Artificial Intelligence	3.7	1000000	1000	
5	Woof	Software Engineering	3	4200	4200	
6						

Statistics

	A	B	C	D	E	F
1	<Current CGPA>	<Expected Graduation Salary>	<Actual Graduation Salary>	<Name>	<Department>	
2	Min: 3.0	Min: 1000.0	Min: 1000.0	Mode: No mode (0)	Mode: No mode (0)	
3	Max: 4.3	Max: 1000000.0	Max: 1000000.0			
4	Mode: No mode (0)	Mode: No mode (0)	Mode: 1000.0 (2)			
5	Median: 3.35	Median: 7100.0	Median: 2600.0			
6	Mean: 3.25	Mean: 253800.0	Mean: 251550.0			
7	Range: 1.2999999999999998	Range: 999000.0	Range: 999000.0			
8	Variance: 0.9766666666666666	Variance: 2.4748696E11	Variance: 2.4897001E11			
9	Standard Deviation: 0.9882644720249062	Standard Deviation: 497480.61268757	Standard Deviation: 498968.9469295659			
10						

b) Method to fill in missing values of specified columns with a specified value

Data to Be Filled Up

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	kong		2	10000	1000	
3	LWY	Information system	4.3	1000	1000000	
4	Meow	Artificial Intelligence	3.7	1000000	1000	
5	Woof	Software Engineering	3		4200	
6						

Data After Filling Up Null Value

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	kong	datascience	2	10000	1000	
3	LWY	Information system	4.3	1000	1000000	
4	Meow	Artificial Intelligence	3.7	1000000	1000	
5	Woof	Software Engineering	3	4200	4200	
6						

4. Scalers:

a) Method to perform Standard Scaling

Data to Perform Scaling

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	kong	Data Science	2	10000	1000	
3	LWY	Information system	4.3	1000	1000000	
4	Meow	Artificial Intelligence	3.7	1000000	1000	
5	Woof	Software Engineering	3	4200	4200	
6						

Standard Scaling (e.g. on Actual Graduation Salary)

	A	B	C	D	E
1	Standard scaling				
2	Actual Graduation Salary:				
3					
4	-0.502135457	1.499993145	-0.502135457	-0.495722232	
5					

b) Method to perform Min Max Scaling

Data to Perform Scaling

	A	B	C	D	E	F
1	Name	Department	CurrentCGPA	Expected Graduation Salary	Actual Graduation Salary	
2	kong	Data Science	2	10000	1000	
3	LWY	Information system	4.3	1000	1000000	
4	Meow	Artificial Intelligence	3.7	1000000	1000	
5	Woof	Software Engineering	3	4200	4200	
6						

Min Max Scaling (e.g. on Expected Graduation Salary)

	A	B	C	D	E
1	MinMax scaling				
2	Expected Graduation Salary:				
3					
4	0.009009009	0	1	0.003203203	
5					

5. K-Nearest Neighbors (k-nn):

Training data used (DATA.csv)

	A	B	C	D	E	F
1	Name	Department	CurrentCGP	Expected Gr	Actual Graduation	Salary
2	Ma Ning	Information	4	9000	10000	
3	Wen Li	Artificial Int	4	9000	10000	
4	Zhao Cheng	Multimedia	3.5	7000	8000	
5	Zou Fang	Information	3.5	7000	8000	
6	Yap Jia Yi	Software Er	3.5	7000	8000	
7	Zhu Huan	Artificial Int	2.5	7000	2000	
8	Ma Ning	Artificial Int	2.5	1000	2000	
9	Xia Shu Xu	Computer S	2.5	1000	2000	
10	Zou Fang	Artificial Int	3	5000	6000	
11	Lim Jia Yi	Artificial Int	2.75	3000	4000	
12	Lim Mei Ling	Multimedia	2.5	1000	2000	
13	Lee Jing Yua	Multimedia	4	9000	10000	
14	Lee Jian Hui	Data Scienc	4	9000	10000	
15	Loo Ying Yin	Software Er	3	5000	6000	
16	Lisa Tan	Artificial Int	3.5	7000	8000	
17	Longan Mur	Artificial Int	2.75	3000	4000	
18	Muhamad H	Artificial Int	4	9000	10000	
19	Lim Jia Jia	Computer S	4	9000	10000	
20	Muhammed	Data Scienc	3.5	7000	8000	
21	Murugan Ali	Artificial Int	4	9000	10000	
22	Misa Lim	Software Er	2.75	3000	4000	
23	Mia Ng	Information	2.75	3000	4000	
24	Nigel Ng	Multimedia	2.5	1000	2000	
25	Nithira Ng	Information	3	5000	6000	
26	Nick Lim	Data Scienc	3.5	7000	8000	
27	Lim Mei Me	Software Er	4	9000	10000	
28	Nicholas Ng	Data Scienc	4	9000	10000	
29	Lim Mei Hui	Computer S	4	9000	10000	
30	Muhammed	Data Scienc	3.5	7000	8000	

a. Method to impute values for a column using **k-nn regressor** using subset of other columns

	A	B
1	distance	dependant value
2	0.2000001	7000
3	0.2000001	7000
4	0.2000001	7000
5	0.2000001	7000
6	0.2000001	7000
7	0.2000001	7000
8	0.2000001	7000
9	0.2000001	7000
10	0.2000001	7000
11	0.3	5000
12	0.3	5000
13	0.3	5000
14	0.3	5000
15	0.3	5000
16	0.3	5000
17	0.55	3000
18	0.55	3000
19	0.55	3000
20	0.55	3000
21	0.55	3000
22	0.55	3000
23	0.55	3000
24	0.55	3000
25	0.55	3000
26	0.7000001	9000
27	0.7000001	9000
28	0.7000001	9000
29	0.7000001	9000
30	0.7000001	9000

b. Method to impute values for a column using **k-nn classifier** using subset of other columns

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	class	frequency within k neighbors											
2	9000	0											
3	7000	9											
4	1000	0											
5	5000	6											
6	3000	0											
7	Predicted classification of Expected Graduation Salary based on 15 nearest neighbor for the given CurrentCGPA which is 3.25 is class: 7000.0 with the frequency of 9												

6. Error Metric:

For k-NN regression: MAE(Mean Average Error) and MAPE(Mean Average Percentage Error)

k	Mean Average Error
1	1102.04
2	551.02
3	367.35
4	275.51
5	220.41
6	183.67
7	192.42
8	198.98
9	204.08
10	355.10
11	378.48
12	336.73
13	414.44
14	428.57
15	440.82
16	420.92
17	484.99
18	541.95
19	534.91
20	602.04
21	697.76
22	751.39
23	800.35
24	784.01
25	857.14

k	Mean Average Percentage Error(%)
1	110.20
2	55.10
3	36.73
4	27.55
5	22.04
6	18.37
7	16.44
8	15.00
9	13.88
10	19.63
11	20.99
12	21.25
13	24.17
14	24.91
15	26.50
16	26.39
17	28.17
18	29.75
19	31.25
20	33.87
21	37.40
22	39.49
23	41.83
24	42.52
25	46.68
26	50.64

For k-NN classification: Confusion Matrix and F1 Score

k	Accuracy from Confusion Matrix(%)
1	62.75
2	50.98
3	54.90
4	50.98
5	50.98
6	50.98
7	50.98
8	50.98
9	50.98
10	47.06
11	50.98
12	50.98
13	50.98
14	50.98
15	50.98
16	50.98
17	50.98
18	50.98
19	50.98
20	50.98
21	50.98
22	50.98
23	50.98
24	50.98
25	50.98
26	50.98
27	50.98
28	50.98
29	50.98
30	50.98

k	F1 score
1	0.70
2	0.32
3	0.26
4	0.24
5	0.44
6	0.19
7	0.00
8	0.00
9	0.19
10	0.31
11	0.00
12	0.00
13	0.00
14	0.00
15	0.00
16	0.00
17	0.00
18	0.00
19	0.00
20	0.00
21	0.00
22	0.00
23	0.00
24	0.00
25	0.00
26	0.00
27	0.00
28	0.00
29	0.00
30	0.00

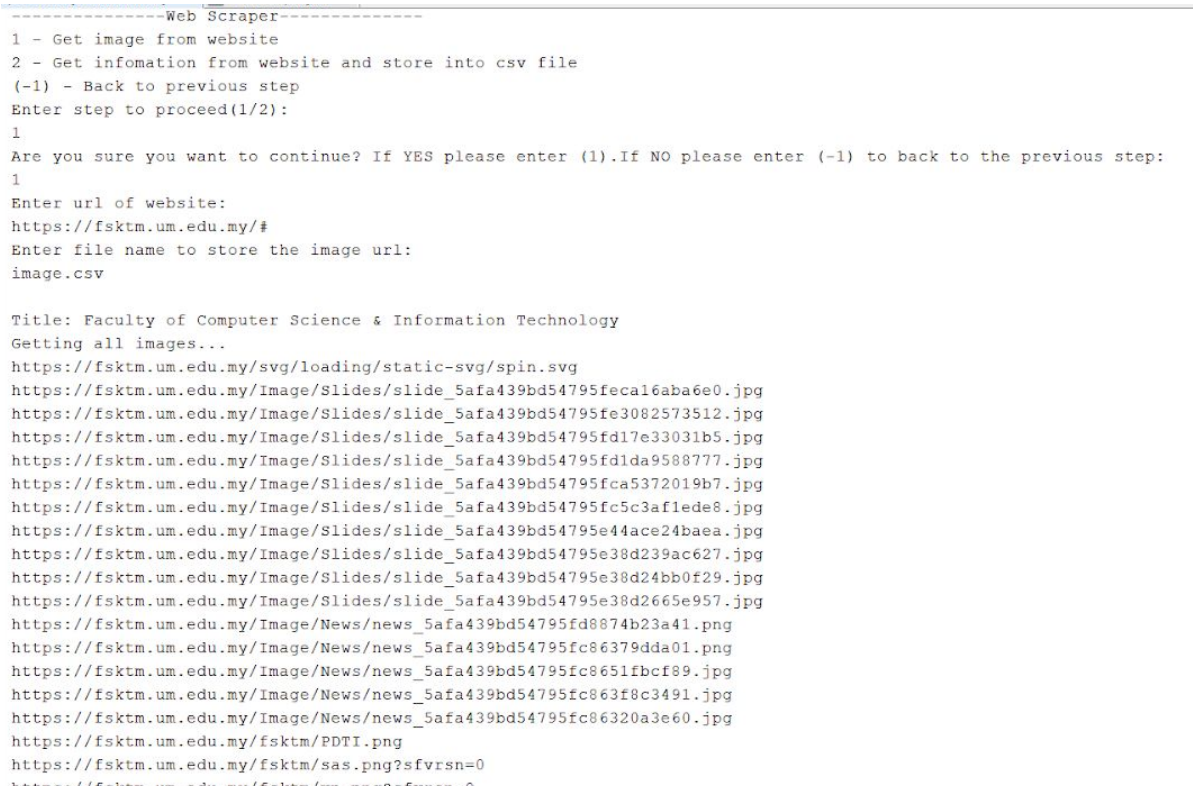
EXTRA FEATURES

These are a few of our major extra features:

Web Scraper

We added a web scraper which allows users to scrape the data from the website by inputting website URL. After scraping the data from the website, the data from the website will automatically convert to a txt or csv file. In web scraper, users can input choice 1 or 2 to proceed.

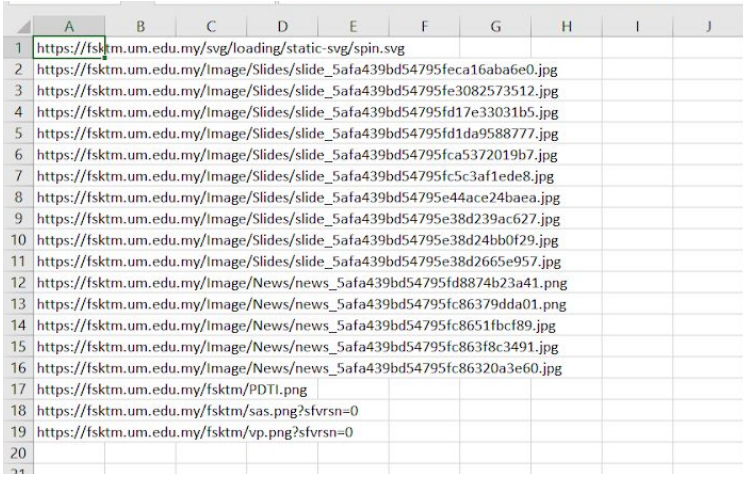
For choice 1, the web scraper will scrape all the images in the specific website.



```
--------------Web Scraper--------------
1 - Get image from website
2 - Get infomation from website and store into csv file
(-1) - Back to previous step
Enter step to proceed(1/2):
1
Are you sure you want to continue? If YES please enter (1).If NO please enter (-1) to back to the previous step:
1
Enter url of website:
https://fsktm.um.edu.my/#
Enter file name to store the image url:
image.csv

Title: Faculty of Computer Science & Information Technology
Getting all images...
https://fsktm.um.edu.my/svg/loading/static-svg/spin.svg
https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795feca16aba6e0.jpg
https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795fe3082573512.jpg
https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795fd17e33031b5.jpg
https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795fd1da9588777.jpg
https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795fca5372019b7.jpg
https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795fc5c3af1ede8.jpg
https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795e44ace24baea.jpg
https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795e38d239ac627.jpg
https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795e38d24bb0f29.jpg
https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795e38d2665e957.jpg
https://fsktm.um.edu.my/Image/News/news_5afa439bd54795fd8874b23a41.png
https://fsktm.um.edu.my/Image/News/news_5afa439bd54795fc86379dda01.png
https://fsktm.um.edu.my/Image/News/news_5afa439bd54795fc8651fbcf89.jpg
https://fsktm.um.edu.my/Image/News/news_5afa439bd54795fc863f8c3491.jpg
https://fsktm.um.edu.my/Image/News/news_5afa439bd54795fc86320a3e60.jpg
https://fsktm.um.edu.my/fsktm/PDTI.png
https://fsktm.um.edu.my/fsktm/sas.png?sfvrsn=0
```

Sample output in csv file:



	A
1	https://fsktm.um.edu.my/svg/loading/static-svg/spin.svg
2	https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795feca16aba6e0.jpg
3	https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795fe3082573512.jpg
4	https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795fd17e33031b5.jpg
5	https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795fd1da9588777.jpg
6	https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795fca5372019b7.jpg
7	https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795fc5c3af1ede8.jpg
8	https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795e44ace24baea.jpg
9	https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795e38d239ac627.jpg
10	https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795e38d24bb0f29.jpg
11	https://fsktm.um.edu.my/Image/Slides/slide_5afa439bd54795e38d2665e957.jpg
12	https://fsktm.um.edu.my/Image/News/news_5afa439bd54795fd8874b23a41.png
13	https://fsktm.um.edu.my/Image/News/news_5afa439bd54795fc86379dda01.png
14	https://fsktm.um.edu.my/Image/News/news_5afa439bd54795fc8651fbcf89.jpg
15	https://fsktm.um.edu.my/Image/News/news_5afa439bd54795fc863f8c3491.jpg
16	https://fsktm.um.edu.my/Image/News/news_5afa439bd54795fc86320a3e60.jpg
17	https://fsktm.um.edu.my/fsktm/PDTI.png
18	https://fsktm.um.edu.my/fsktm/sas.png?sfvrsn=0
19	https://fsktm.um.edu.my/fsktm/vp.png?sfvrsn=0
20	

For choice 2, the web scraper will scrape all the information from the table in the website.

```
---------------Web Scraper---------------
1 - Get image from website
2 - Get infomation from website and store into csv file
(-1) - Back to previous step
Enter step to proceed(1/2):
2
Are you sure you want to continue? If YES please enter (1).If NO please enter (-1) to back to the previous step:
1
Enter url:
 https://computersciencestudentsalary.blogspot.com/2020/12/2020-worlds-richest-people-ranking.html

Enter file name to store the data from website:
website.csv
Extracting data from URl....

Rank Name Total Net Worth Country Industry
1 Jeff Bezos $182.5 B United States Technology
2 Elon R Musk $147.0 B United States Technology
3 Bill Gates $146.1 B France Technology
4 Bernard Arnault $120.1 B United States Consumer
5 Mark Zuckerberg $98.3 B United States Technology
```

Sample output in csv file:

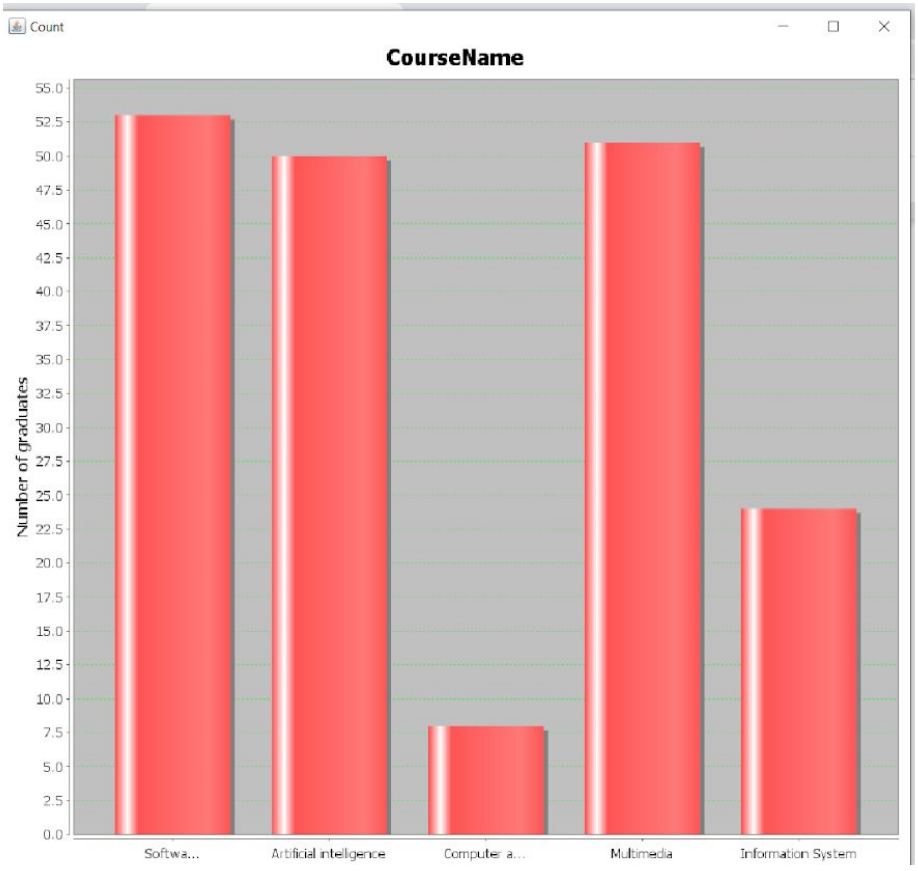
	A	B	C	D	E
1	Rank	Name	Total Net Worth	Country	Industry
2	1	Jeff Bezos	$182.5 B	United States	Technology
3	2	Elon R Musk	$147.0 B	United States	Technology
4	3	Bill Gates	$146.1 B	France	Technology
5	4	Bernard Arnault	$120.1 B	United States	Consumer
6	5	Mark Zuckerberg	$98.3 B	United States	Technology
7					
8					
9					

Chart

We also added a bar chart generator in our project. By calculating the data in the csv file, a bar chart will be generated based on the data in the csv file.

```
--------------Main Page--------------
1 - Adding new data
2 - Managing data
3 - Statistic
4 - Scalers
5 - K-Nearest Neighbors(k-nn)
6 - Generate charts
7 - Convert file type to (.json)
8 - Web scraping
-1 - Exit
Enter step(1-8) to be proceed:
6
Are u sure u want to proceed? If yes please enter (1). If NO please enter (-1) to back to the previous step:
1
Enter file name:
OverallData.csv
Total number graduates for software engineering: 53
Total number graduates for artificial intelligence: 50
Total number graduates for computer and system network: 8
Total number graduates for multimedia: 51
Total number graduates for information system: 24
Total number graduates for data science: 40
```

Sample output:



JSON(JavaScript Object Notation) output

To further enhance the system, we added an extra feature in which the files can be saved into a JSON file. We have chosen to do this because it is a platform-independent format. Any language that can parse a string can handle JSON. It is also supported by many databases such as PostgreSQL and MySQL, NoSQL databases, MongoDB were built upon this format and use JSON documents to store records.

Why we choose JSON:

1. JSON is Faster:

Easy parsing of data and faster data execution. Since its syntax is very small and light weighted that's the reason that it executes the response in the faster way.

2. Schema Support:

It has a wide range of supported browser compatibility with the operating systems. The applications made with the coding of JSON doesn't require much effort to make it all browser compatible.

3. Server Parsing:

If the parsing will be fast on the server side then only the users can get the fast response. JSON server-side parsing is the strong point that indicates us to use the JSON on the server side.

4. Tool for sharing data:

JSON is the best tool for sharing data of any size, even audio, video. This is because JSON stores the data in the arrays so it makes data transfer easier.

Sample output:


```
{"Department":["Information System","Artificial Intelligence","Multimedia","Information System","Software
Engineering","Artificial Intelligence","Artificial Intelligence","Computer System and Network","Artificial
Intelligence","Information System","Multimedia","Software Engineering","Artificial Intelligence","Artificial
Intelligence","Artificial Intelligence","Multimedia","Multimedia","Multimedia","Computer System and Network","Artificial
Intelligence","Software Engineering","Multimedia","Information System","Computer And System Network","Data
Science","Software Engineering","Multimedia","Information System","Computer And System Network","Artificial
Intelligence","Software Engineering","Multimedia","Information System","Computer And System Network","Artificial
Intelligence","Data Science","Multimedia","Information System","Computer System and Network","Information
System","Information System","Data Science","Artificial Intelligence","Computer System and Network","Multimedia","Computer
System and Network","Information System","Software Engineering","Software Engineering","Computer System and Network","Data
Science","Software Engineering","Multimedia","Multimedia","Software Engineering","Information System","Information
System","Artificial Intelligence","Artificial Intelligence","Computer System and Network","Data Science","Artificial
Intelligence","Multimedia","Multimedia","Data Science","Software Engineering","Information System","Information
System","Multimedia","Artificial Intelligence","Multimedia","Multimedia","Data Science","Software Engineering","Artificial
Intelligence","Artificial Intelligence","Artificial Intelligence","Data Science","Artificial Intelligence","Information
System","Data Science","Software Engineering","Artificial Intelligence","Information System","Information
System","Multimedia","Multimedia","Information System","Data Science","Computer System and Network","Data
Science","Artificial Intelligence","Artificial Intelligence","Artificial Intelligence","Artificial Intelligence","Computer
System and Network","Data Science","Artificial Intelligence","Multimedia","Information System","Computer And System
Network","Artificial Intelligence","Data Science","Multimedia","Information System","Computer And System
Network","Artificial Intelligence","Software Engineering","Data Science","Information System","Computer And System
Network","Artificial Intelligence","Software Engineering","Multimedia","Information System","Computer And System
Network","Data Science","Software Engineering","Data Science","Multimedia","Information System","Data Science","Software
Engineering","Multimedia","Artificial Intelligence","Software Engineering","Artificial Intelligence","Artificial
Intelligence","Software Engineering","Artificial Intelligence","Data Science","Artificial Intelligence","Software
Engineering","Data Science","Information System","Computer And System Network","Artificial Intelligence","Software
Engineering","Computer System and Network","Data Science","Artificial Intelligence","Software Engineering","Information
Systems","Multimedia","Information Systems","Data Science","Software Engineering","Data Science","Computer System and
Network","Data Science","Data Science","Artificial Intelligence","Software Engineering","Software Engineering","Software
Engineering","Multimedia","Information Systems","Multimedia","Multimedia","Multimedia","Artificial
Intelligence","Multimedia","Multimedia","Data Science","Artificial Intelligence","Software Engineering","Data
Science","Data Science","Information Systems","Software Engineering","Artificial Intelligence","Software
Engineering","Data Science","Computer System and Network","Data Science","Artificial Intelligence","Artificial
Intelligence","Software Engineering","Data Science","Software Engineering","Artificial Intelligence","Information
Systems","Artificial Intelligence","Computer System and Network","Computer System and Network","Software
Engineering","Artificial Intelligence","Computer System and Network","Data Science","Artificial
Intelligence","Multimedia","Multimedia","Data Science","Software Engineering","Artificial Intelligence","Information
Systems","Computer System and Network","Information Systems","Data Science","Software Engineering","Software
```

```
Engineering","Multimedia","Information Technology","Aritificial Intelligence"],"Current CGPA":
["3.5","4","3.5","3.5","3.5","2.5","2.5","2.5","3","3.5","2.75","2.75","2.5","2.5","2.75","2.5","2.5","4","3.5","2.5","3",
"4","3.5","2.75","2.5","3","4","3.5","2.75","2.5","3","4","2.5","3","4","2.5","3","4","3.5","2.75","3","3","2.5","2.75","3
","2.75","2.5","3.5","2.75","4","3.5","2.75","3.5","3","2.75","2.5","2.5","3.5","3.5","2.5","2.5","2.75","2.5","4","4","3"
,"3","2.75","3","2.75","2.5","4","4","3","3.5","2.75","4","2.5","2.5","4","2.75","3","2.75","3","2.75","3","2.75","2.75","
3","2.75","3.5","3","2.75","2.75","4","4","3.5","4","3","4","2.5","3","4","2.5","3","4","3.5","2.75","2.5","3","2.5","3","
4","3.5","2.75","2.75","2.5","3","3.5","2.5","3","4","3.5","2.5","3","4","2.5","3","4","3.5","2.75","2.5","3","4","3.5","2
.75","2.5","2.5","4","3.5","4","2.75","2.75","2.5","3","3.5","4","4","4","3.5","4","2.75","2.75","2.75","3.5","3","3","4",
"2.75","3","3","3.5","2.5","2.5","3.5","3.5","3.5","3","2.5","4","3","3.5","2.5","3","2.75","3.5","3","2.75","3","2.75","3
","3","4","3.5","2.5","2.5","2.75","3.5","2.5","2.5","2.75","2.5","4","4","3","2.5","2.5","2.75","2.75","3","2.5","2.75","
2.75","2.5","3","3.5","4","4","4","3.5","4","2.75","2.75","2.75","3.5","3","3","4","2.75","3","3","3.5","2.5","2.5","3.5",
"3.5","3.5","3","2.5","4","3","3.5","2.5","2.75","4","2.75","3","2.5","3","2.67","4","3","3.33","3","2","2.33","2.67","3.6
7","3.33","3.33","4","4","2.33","2.33","2","1.67","4","3.67","3.67","1.33","2.33","2.75","1.33","4","4","3","3","3.67","2.
67","1.67","3.67","2.67","3.33","2.67","3.33","3","3","2.33","2.75","3","3.67","3.33","1.33","4","2.33","2.67","2.67","3.3
3","3.33","1.67","1.33","3.67","4","3.67","1.67","2.33","3","3.33","4"],"Actual Graduation Salary":
["10000","5000","5000","5000","3000","10000","5000","10000","4000","3000","5000","8000","8000","4000","8000","5000","10000
","5000","4000","4000","3000","5000","8000","10000","4000","3000","5000","8000","10000","4000","3000","4000","3000","5000"
,"8000","4000","4000","3000","5000","8000","4000","5000","5000","5000","4000","10000","8000","8000","4000","3000","10000",
"10000","3000","4000","5000","4000","8000","5000","8000","10000","5000","3000","10000","10000","4000","8000","3000","5000"
,"8000","3000","10000","10000","4000","8000","10000","10000","10000","10000","8000","3000","8000","4000","5000","3000","50
00","8000","10000","8000","5000","10000","8000","5000","10000","10000","10000","5000","4000","5000","10000","5000","4000",
"3000","5000","8000","10000","5000","8000","10000","4000","3000","5000","8000","10000","8000","10000","4000","3000","5000"
,"5000","8000","10000","4000","3000","5000","8000","10000","8000","10000","4000","3000","5000","8000","10000","10000","400
0","3000","5000","8000","5000","4000","5000","3000","4000","5000","8000","4000","10000","4000","10000","4000","4000","4000
","4000","4000","8000","4000","3000","8000","8000","4000","10000","3000","8000","5000","8000","10000","4000","8000","8000"
,"10000","8000","3000","10000","5000","10000","8000","5000","10000","10000","5000","10000","8000","8000","3000","8000","10
000","3000","8000","10000","5000","3000","10000","10000","4000","8000","8000","4000","4000","8000","5000","4000","3000","4
000","5000","8000","4000","10000","4000","10000","4000","4000","4000","4000","8000","4000","3000","8000","8000","4000
00","10000","3000","8000","5000","8000","10000","4000","8000","8000","10000","8000","3000","10000","3000","8000","10000","
```

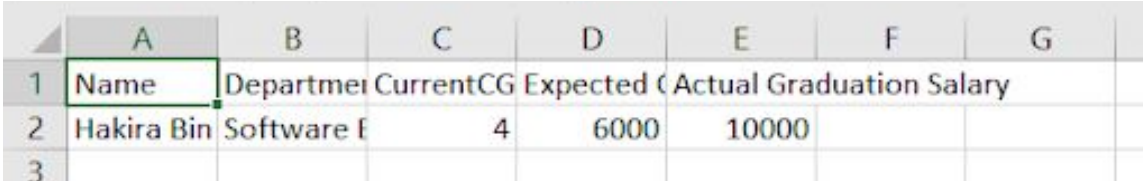
Search Name

In order to search the valid data easily, we implemented a function called data browser. For example, by entering a specific parameter of the entity (name), we can easily get the row of the data (entity) with its attributes included. The specific data will directly store to a new csv file.

Sample input:

```
--------------Managing data--------------
1 - Concatenate data
2 - Obtain specific data
3 - Data sorting
4 - Remove duplicated data in row
5 - Remove data with null value
6 - Data Browser
(-1) - Back to previous step
Enter step to proceed(1-6):
6
Are you sure you want to continue? If YES please enter (1).If NO please enter (-1) to back to the previous step:
1
Enter file name:
DATA.csv
Enter the name of person you wish to search:
Hakira Bin Hadi
Hakira Bin Hadi found at row 40
Name                    Department              CurrentCGPA             Expected Graduation Salary    Actual Graduation Salary
Hakira Bin Hadi         Software Engineering    4                       6000                          10000
```

Sample output:



	A	B	C	D	E	F	G
1	Name	Departmer	CurrentCG	Expected (	Actual Graduation Salary		
2	Hakira Bin	Software E	4	6000	10000		
3							

LIMITATIONS

Limitations	Explanations
Row and column concatenation method limitation	When numbers of row and column are the same for both csv files(2x2/3x3), there will be semantic error in output because the program could not differentiate row from column.
RemoveDuplicatedRow Method limitation	Program can only detect duplication of one type of name. More than one type of name being duplicated will mess up the system.
Knn classifier limitation	F1 score error metric can only be used when there are only 2 classes(true/false).
Knn regressor limitation	For k-nn regression, the output will be inaccurate when the instance value is out of the range of training data or near to the edge of training data
Error metric computation limitation	To compute MAPE error metric for k-nn regressor, there are limitations towards the training data where data points with the actual value zero need to be excluded to avoid a division by zero error.
Web scraper (Info Scraper)	Only the data from the website in table form can be obtained. Other than that, the data will be very messy and unorganized.