A Project Report On
"An Ontology-based Approach to Develop Data Cleaning Extension - 0Data"

Submitted By
Shouman Barua Shuvo
181-35-2321
Department of Software Engineering
Daffodil International University

Supervised by
Md. Shohel Arman
Senior Lecturer
Department Of Software Engineering
Daffodil International University

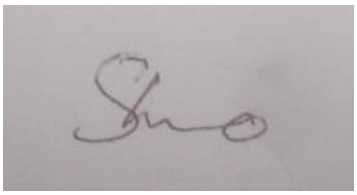
This project report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering.

© All right reserved by Daffodil International University

APPROVAL OF PROJECT

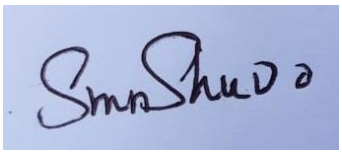
This project titled “0Data : A Basic Data Cleaning Extension for VSCode”, submitted by Shouman Barua Shuvo, ID: 181-35-2321 to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Software Engineering and approved as to its style and contents

Supervised by



Md. Shohel Arman
Senior Lecturer
Department Of Software Engineering
Faculty Of Science & Information Technology
Daffodil International University

Submitted by



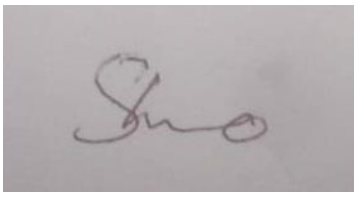
Shouman Barua Shuvo
181-35-2321
Department of Software Engineering
Faculty Of Science & Information Technology
Daffodil International University

© Daffodil International University

DECLARATION

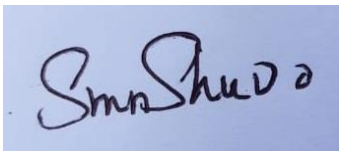
I hereby declare that I have done this project under the supervision of Md. Shohel Arman, Senior Lecturer, Department of Software Engineering, Daffodil International University. I also declare that this project or any part of this is unique and has not been submitted elsewhere for the award of any degree.

Supervised by:



Md. Shohel Arman
Senior Lecturer
Department Of Software Engineering
Faculty Of Science & Information Technology
Daffodil International University

Submitted by



Shouman Barua Shuvo
181-35-2321
Department of Software Engineering
Faculty Of Science & Information Technology
Daffodil International University

© Daffodil International University

ACKNOWLEDGEMENT

At first, I am blessed that I successfully moved towards the last semester. I am pleased with my almighty. First, at the beginning of university life, I have learned a lot about software development as well as computer science-related knowledge from my university's knowledgeable teachers and helpful course mates. Teachers teach us ethics, morality, and politeness as well as software knowledge and related knowledge. I must be thankful to my parents and my family to give me the opportunity and always be to myself. My family always supports me. I am highly indebted to Md. Shohel Arman for his guidance and constant supervision as well as for providing necessary information regarding the project & for his support in completing the project. My supervisor supports me to make this project "SDLC Manager" a successful end. My thanks and appreciations go to my course mates in developing the project and people who have willingly helped me out with their abilities.

© Daffodil International University

ABSTRACT

“0Data : A Basic Data Cleaning Extension for VSCode” is an extension based project to help the individual developer or developer group to have some basic data cleaning, where currently they are having to use third party websites or tools that uses online sync to do the basic cleaning.

The main feature here is the extension is totally offline. I started the project, doesn’t mean any other developer cannot extend it. Being open-source developers have the option to extend it till the end they need.

Having clean data will enhance overall productivity and help you to make decisions based on the best quality information available. If you have to do buy a software and/or to use a website you have to share a data that shouldn’t be fair in any circumstance. This is why this extension came in handy. With the touch of the extension the data cleaning will be just a matter of click. I think the extension currently includes the basic data cleaning features that a data science students or developer might need.

© Daffodil International University

List of Contents

APPROVAL OF PROJECT ii
DECLARATION iii
ACKNOWLEDGEMENT iv
ABSTRACT v
CHAPTER 1: INTRODUCTION 1
1.1 Project Overview 1
1.2. Project Purpose 1
1.2.1 Background 2
1.2.2 Benefits & Beneficiaries 2
1.3 Stakeholders 2
1.3.1 Owner 2
1.3.2 Member 3
1.4 Modules of 0Data 3
1.5 Project Schedule 4
1.5.1 Gantt Chart 5
1.5.2 Release Plan or Milestone 6
1.6 Glossary 7
CHAPTER 2: REQUIREMENTS ENGINEERING 9
2.1 Functional Requirements (FR): 9
2.1.1 9
2.1.2 9
2.1.3 10
2.1.4 10
2.1.5 10
2.1.6 11
2.1.7 11
2.1.8 11
2.2 Non-Functional Requirements 12
2.2.1 12
2.2.2 12
2.2.3 12

© Daffodil International University

CHAPTER 3: SYSTEM ANALYSIS, DESIGN & SPECIFICATION 14
3.1 Development Model.....14
3.2 Use Case Diagram 15
3.2.1 Show Contents 16
3.2.2 Clean Empty Row 17
3.2.2 Fill missing values with Zero 18
3.2.2 Fill missing values with Average 19
3.2.2 Fix Structure 20
3.2.2 Clean Empty Row 21
3.3 Activity Diagram.....22
3.3.1 Dashboard Activity 23
3.3.2 Databoard Activity.....24
3.4 Sequence Diagram.....25
3.4.1 Visualization.....25
3.4.2 Data Cleaning 26
CHAPTER 4: SYSTEM TESTING.....28
4.1 Feature Testing 28
4.1.1 Features to be tested.....28
4.2 Testing strategies 29
4.2.1 Test Approach.....29
4.2.2 Pass/Fail Criteria.....29
4.2.3 Testing Schedule 29
4.3 Testing Environment.....30
4.4 Test Cases 30
4.4.1 Data Visualization 31
4.4.2 Deduplication 32
4.4.3 Cleaning Empty Rows.....33
4.4.4 Fill Missing Values with Zero.....34
4.4.5 Fill Missing Values with Average 35
CHAPTER 5: USER MANUAL 36
5.1 Context Menu 36
5.2 Visualizing Data.....37
5.3 Removing Empty Rows 38

© Daffodil International University

5.4 Deduplication.....40
5.5 Fill Missing Values with Zero/Average 42
Chapter 6: Conclusion.....45
6.1 Project Summary.....45
6.2 Limitations 45
6.3 Obstacles and Achievements 45
6.4 Future Scopes.....45
Chapter 6: Reference.....47

© Daffodil International University

CHAPTER 1: INTRODUCTION

1.1 Project Overview

"0Data : A Basic Data Cleaning Extension for VSCode" is developed keeping in mind of minimizing the hassles that a developer find before starting a project. It's every time s/he has to open Microsoft Excel and do some manual tasks like formatting. Also, s/he has to clean the empty rows. Then check for duplicate entries and so on. This is fine with a small amount of data. But when the amount of data is too large, it becomes almost impossible to do it manually. Then they have to use some online tool, where there's a chance of hijacking data.

1.2. Project Purpose

The purpose of this project is significant to the developer. Developers have been looking for a solution where they can be sure of the privacy of their data. It's not that there is no tool that can do the thing. But the tools are not open source. Also, most of them first uploads the data to their server and then process the data there and then, you get the refined data. Though they say they don't store your data, but who knows if they do!

We can assure these things about the project:

- You can trust 0Data because it is open source.
- 0Data cannot steal your data as it works completely offline and if you still have doubts you can go through the source code and compile the project yourself.

© Daffodil International University

1.2.1 Background

I tried to find an extension to follow before starting the project. There are many extensions on the market place but unfortunately there's none which does the data cleaning for us. So, I had to traverse many of the websites and do research what are the basic functions a data scientist may do on the web to clean their data. I went up following the line and tried to add the very basic functionalities in my extension. I have started the project hoping that it will be extended in future by the developers.

The reason behind choosing the IDE Visual Studio Code was on purpose. There are many IDEs that are used nowadays. I personally do prefer PyCharm by InteliIJ. But if I look of the performance and resource consumption PyChram takes a lot of RAM and processing power itself. Being light weight and less resource hungry I have seen VSCode is more accepted in this community than others.

1.2.2 Benefits & Beneficiaries

This system will be helpful for developers and help them to be sure about the privacy of their data. I also think that it will help the developers to lessen their efforts.

1.3 Stakeholders

A person who is actively involved in this system and is not a developer. According to project management, project stakeholders "a person, group or organization will be influenced or affected by in a decision, activity or outcome of the project". In my system the developer is the stakeholder.

1.3.1 Owner

As my system is user and/or team based the user/team who is using this extension is the owner.

© Daffodil International University

1.3.2 Member

Being a single user base system, the owner is the only member who will be using the system.

1.4 Modules of 0Data

Cleaning Empty Rows	This will clean the rows where there is no data to handle null values.
Deduplication	Sometimes there are more than one rows with exactly same data. Here comes the deduplication tool to remove the duplicate rows.
Fill Missing Values with Average	If there are missing values in a column, this module will help you to fill them with the average of the rest.
Fill Missing Values with Zero	This is the same as filling with average unlike it fills the empty column values with zero.
Fix Structure	If there is Structural error in the document it will try to fix the error.
Remove Duplicate Entry	This is same as deduplication but it goes through columns rather than rows.

© Daffodil International University

Show Contents	This module is my custom developed module which does the job of popular excel viewer. It doesn't support lots of features but it does show a simple excel file just fine.

1.5 Project Schedule

Having a complex thinking I had to invest most of my time in the R&D, having a mind that I do at least have to make the base so that it gets the extendibility and acceptability. I had an option to develop it either for *PyCharm, InteliIJ* or *Visual Studio Code*. Being light weight and open source VSCode has a great user base and most of the developer prefer it nowadays. I have prepared a schedule keeping my target in mind and deadline in the head. The management also refers the tasks that need to be done within a short time and the priority of their works.

© Daffodil International University

1.5.1 Gantt Chart

Activities		Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14
Planning	Ideas														
	Problem Identification														
	Proposal														
Requirements	Requirements Specification														
	Requirement Analysis														
QA-1	Quality Assurance														
System Design	Sketching														
	Design Specification														
	Model Design														
R&D	Testing APIs														
Implementation-1	Write Review														
QA-2	Test Cases														
Implementation-2	Case Imposing and Finding Defects														
Testing	Unit Testing														
	Black Box Testing														
Delivery	Software Release														

Time	
Allocated Time	

Figure 1.5.1f

© Daffodil International University

1.5.2 Release Plan or Milestone

Activities	Week Selection	Total Week Duration
Brainstorming	W1, W2	2
Problem Identification	W1, W2	2
Requirement Specification	W2	1
Requirement Analysis	W2, W3	2
Feasibility Testing	W3	1
Context Designing	W4, W5	2
API Testing	W6	1
Review Writing	W5, W6, W7, W8	4
Unit Testing	W4, W5, W7, W11, W12	5
Test Case Designing	W11	1
Blackbox Testing	W11, W12, W13	3
Software Release	W14	1

© Daffodil International University

1.6 Glossary

Term	Definition
Deduplication	Removing the duplicate values
SRS(Software Requirement Specification)	A document that completely describes all of the functions of a proposed system and the constraints under which it must operate. For example, this document.
Stakeholder	Any person with an interest in the project who is not a developer.
End-User	Intended persons for whom the software is built.
API	Application Programming Interface
R&D	Research and Development
SDLC	Software Development Life Cycle

1.7 Objectives

The main objective of this project is developing an extension which will help other data scientists and/or developer on developing in a hassle free environment.

© Daffodil International University

© Daffodil International University

CHAPTER 2: REQUIREMENTS ENGINEERING

2.1 Functional Requirements (FR):

Functional requirements referred to as the mandatory functions, a software must have. Functional requirements capture the intended behavior of the system. This behavior can be written as functions, services, tasks, or which system is required to perform. In my extension almost every feature is a functional requirement.

2.1.1

FR 1	Cleaning Empty Rows
Description	This will clean the rows where there is no data to handle null values.
Stakeholders	End-users

2.1.2

FR 2	Deduplication
Description	Sometimes there are more than one rows with exactly same data. Here comes the deduplication tool to remove the duplicate rows.
Stakeholders	End-users

© Daffodil International University

2.1.3

FR 3	Fill Missing Values with Average
Description	If there are missing values in a column, this module will help you to fill them with the average of the rest.
Stakeholders	End-users

2.1.4

FR 4	Fill Missing Values with Zero
Description	This is the same as filling with average unlike it fills the empty column values with zero.
Stakeholders	End-users

2.1.5

FR 5	Fill Missing Values with Zero
Description	This is the same as filling with average unlike it fills the empty column values with zero.
Stakeholders	End-users

© Daffodil International University

2.1.6

FR 6	Fix Structure
Description	If there is Structural error in the document it will try to fix the error.
Stakeholders	End-users

2.1.7

FR 7	Remove Duplicate Entry
Description	This is same as deduplication but it goes through columns rather than rows.
Stakeholders	End-users

2.1.8

FR 8	Show Contents
Description	This module is my custom developed module which does the job of popular excel viewer. It doesn’t support lots of features but it does show a simple excel file just fine.
Stakeholders	End-users

© Daffodil International University

2.2 Non-Functional Requirements

Nonfunctional requirements are used to judge the operation of an application and/or system. In my case I have kept some of the non-functional requirements' priority high keeping my aspect in mind.

2.2.1

NFR1	Security
Description	Keeping in mind that data needs privacy I have not embedded any library that uses online services to process the data. All off the library and even the app is offline based as well as open source.
Stakeholders	End-users

2.2.2

NFR2	Scalability
Description	Thinking of scalability, the project is kept open source with proper documentation.
Stakeholders	End-users

2.2.3

NFR3	Usability
Description	The application could be a console-based application, which runs via command line interface. Keeping the target of reaching users of all level the app is GUI based and also a part of an open-source extension.
Stakeholders	End-users

The other non-functional requirements such as performance was maintained on the application. Being priority not high in this application those requirements were not mentioned dedicatedly.

© Daffodil International University

© Daffodil International University

CHAPTER 3: SYSTEM ANALYSIS, DESIGN & SPECIFICATION

3.1 Development Model

My project had lots of confusion, so I had to test and make sure every time that the implementation will work, I had to run through different APIs and, check again and again. Prototype model was suiting my development life cycle just fine. So, I have chosen prototype model for the SDLC period of my extension.

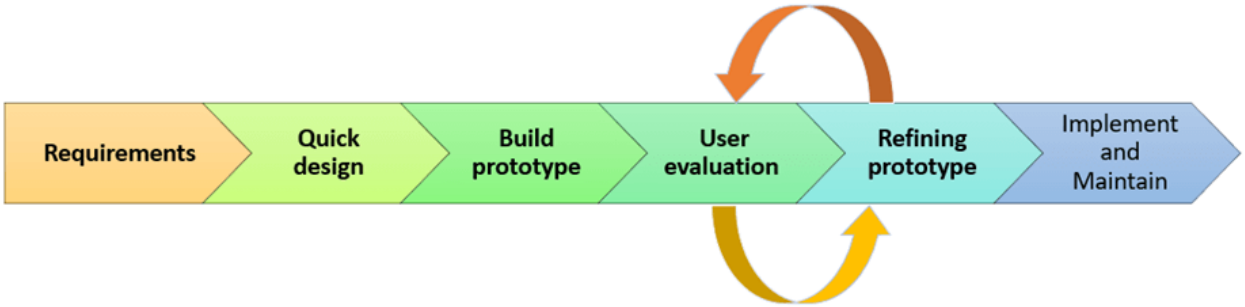
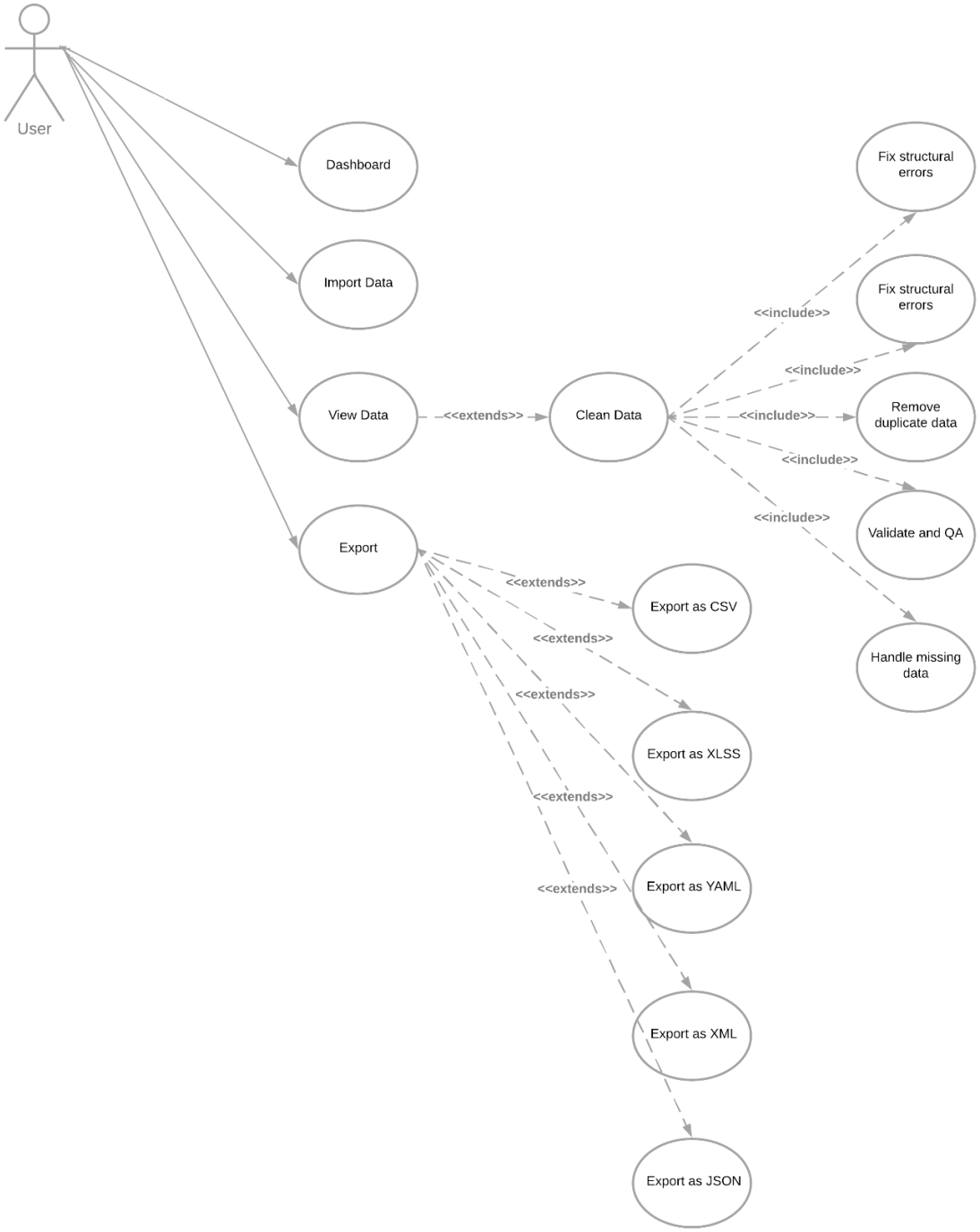


Figure 3.1f: Prototype Model

© Daffodil International University

3.2 Use Case Diagram



© Daffodil International University

Figure 3.2f: Use Case Diagram

Use Case Description

3.2.1 Show Contents

Use Case Name	Show Contents
Trigger	On click on the context menu
Pre Condition	File format have to be excel
Basic Path	1. User hover mouse over the excel file 2. User clicks the right button to open context menu 3. User clicks on the show content action
Alternative Path	Mouse events can be done with keyboard shortcuts
Post Condition	Contents are displayed
Exception Path	Images and charts cannot be displayed
Other	

© Daffodil International University

3.2.2 Clean Empty Row

Use Case Name	Clean Empty Row
Trigger	On click on the context menu
Pre Condition	Excel has to contain rows with no values
Basic Path	1. User hover mouse over the excel file 2. User clicks the right button to open context menu 3. User clicks on the clean empty rows action
Alternative Path	Mouse events can be done with keyboard shortcuts
Post Condition	New excel file is exported with non-empty rows.
Exception Path	The file stays as if.
Other	

© Daffodil International University

3.2.2 Fill missing values with Zero

Use Case Name	Fill missing values with Zero
Trigger	On click on the context menu
Pre Condition	Selected column must contain numeric values
Basic Path	4. User hover mouse over the excel file 5. User clicks the right button to open context menu 6. User clicks on the clean empty rows action
Alternative Path	Mouse events can be done with keyboard shortcuts
Post Condition	New excel file is exported with non-empty rows.
Exception Path	The file stays as if.
Other	

© Daffodil International University

3.2.2 Fill missing values with Average

Use Case Name	Fill missing values with Average
Trigger	On click on the context menu
Pre Condition	Selected column must contain numeric values
Basic Path	7. User hover mouse over the excel file 8. User clicks the right button to open context menu 9. User clicks on the clean empty rows action
Alternative Path	Mouse events can be done with keyboard shortcuts
Post Condition	New excel file is exported with non-empty rows.
Exception Path	The file stays as if.
Other	

© Daffodil International University

3.2.2 Fix Structure

Use Case Name	Fix Structure
Trigger	On click on the context menu
Pre Condition	The excel value may not contain dynamic data.
Basic Path	10. User hover mouse over the excel file 11. User clicks the right button to open context menu 12. User clicks on the clean empty rows action
Alternative Path	Mouse events can be done with keyboard shortcuts
Post Condition	New excel file is exported with non-empty rows.
Exception Path	The file stays as if.
Other	

© Daffodil International University

3.2.2 Clean Empty Row

Use Case Name	Remove Duplicate Entry
Trigger	On click on the context menu
Pre Condition	The column must contain more than one exact same values.
Basic Path	13. User hover mouse over the excel file 14. User clicks the right button to open context menu 15. User clicks on the clean empty rows action
Alternative Path	Mouse events can be done with keyboard shortcuts
Post Condition	New excel file is exported with non-empty rows.
Exception Path	The file stays as if.
Other	

© Daffodil International University

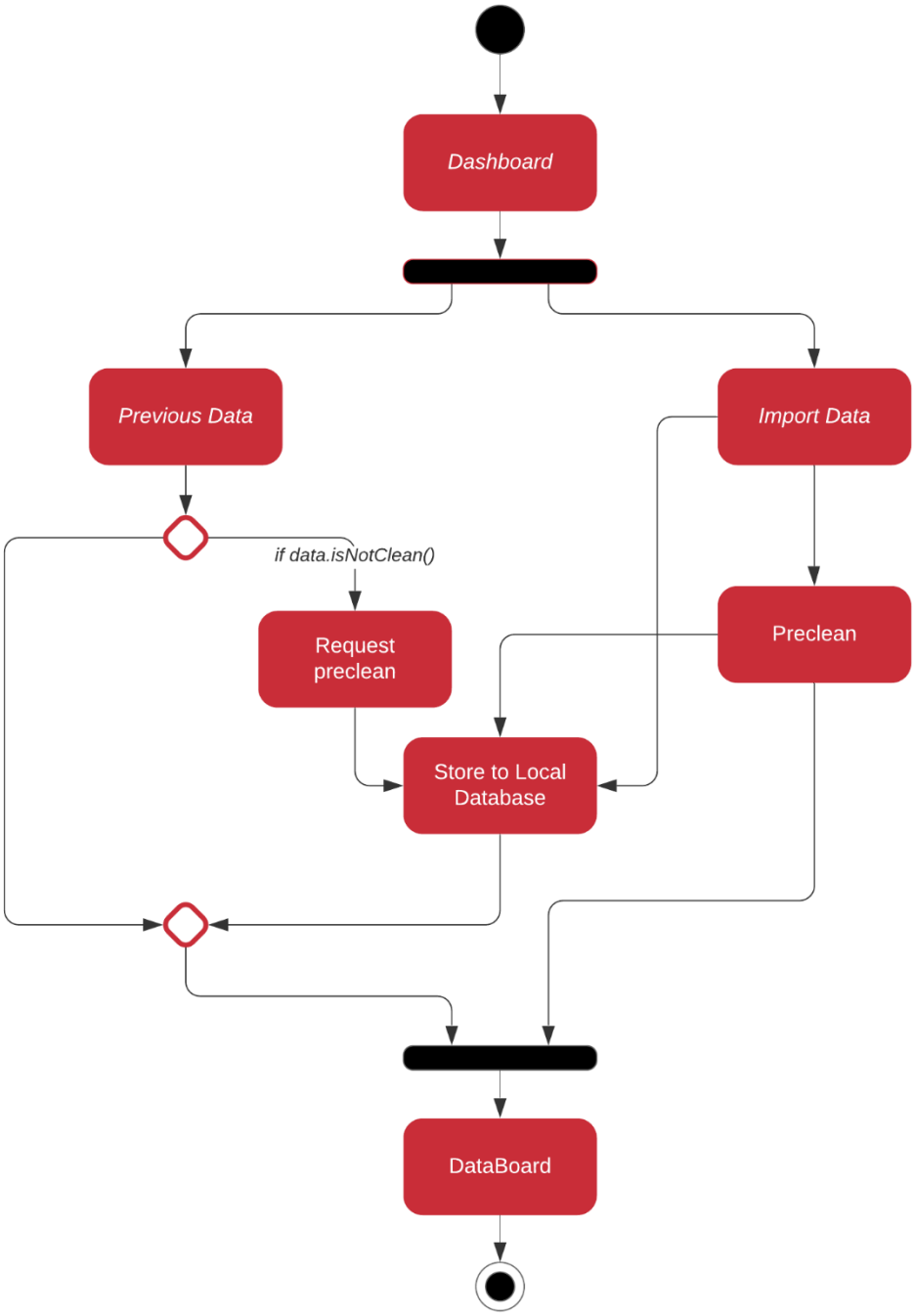
3.3 Activity Diagram

I have prepared an activity diagram for my project. Though my project doesn’t have much activity flow to be displayed. By the time I was planning the functions I didn’t had a plan for any specific library or API.

(Move the next page)

© Daffodil International University

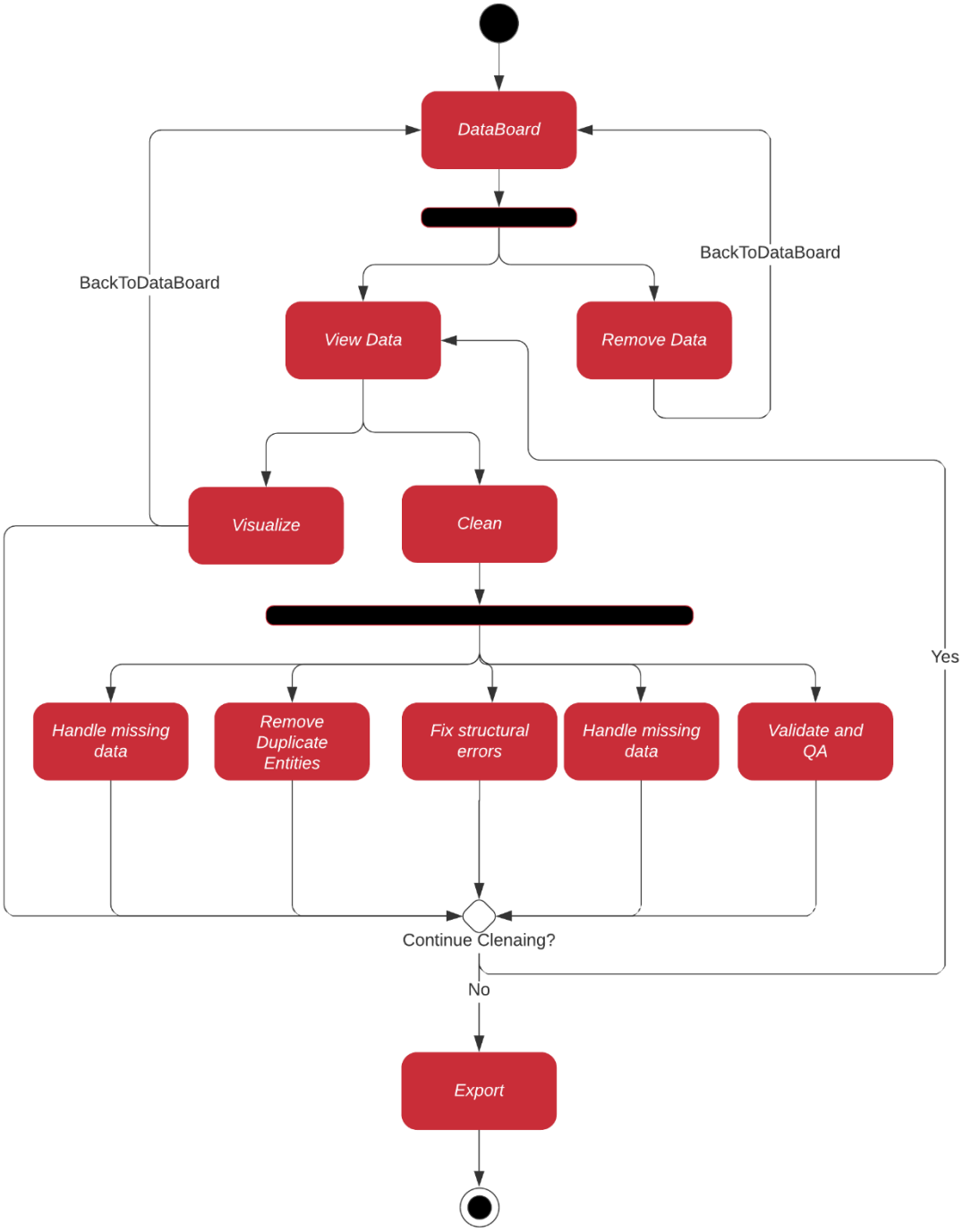
3.3.1 Dashboard Activity



© Daffodil International University

Figure 3.3.1f: Dashboard Activity

3.3.2 Databoard Activity



© Daffodil International University

Figure 3.3.2f: Databoard Activity

3.4 Sequence Diagram

3.4.1 Visualization

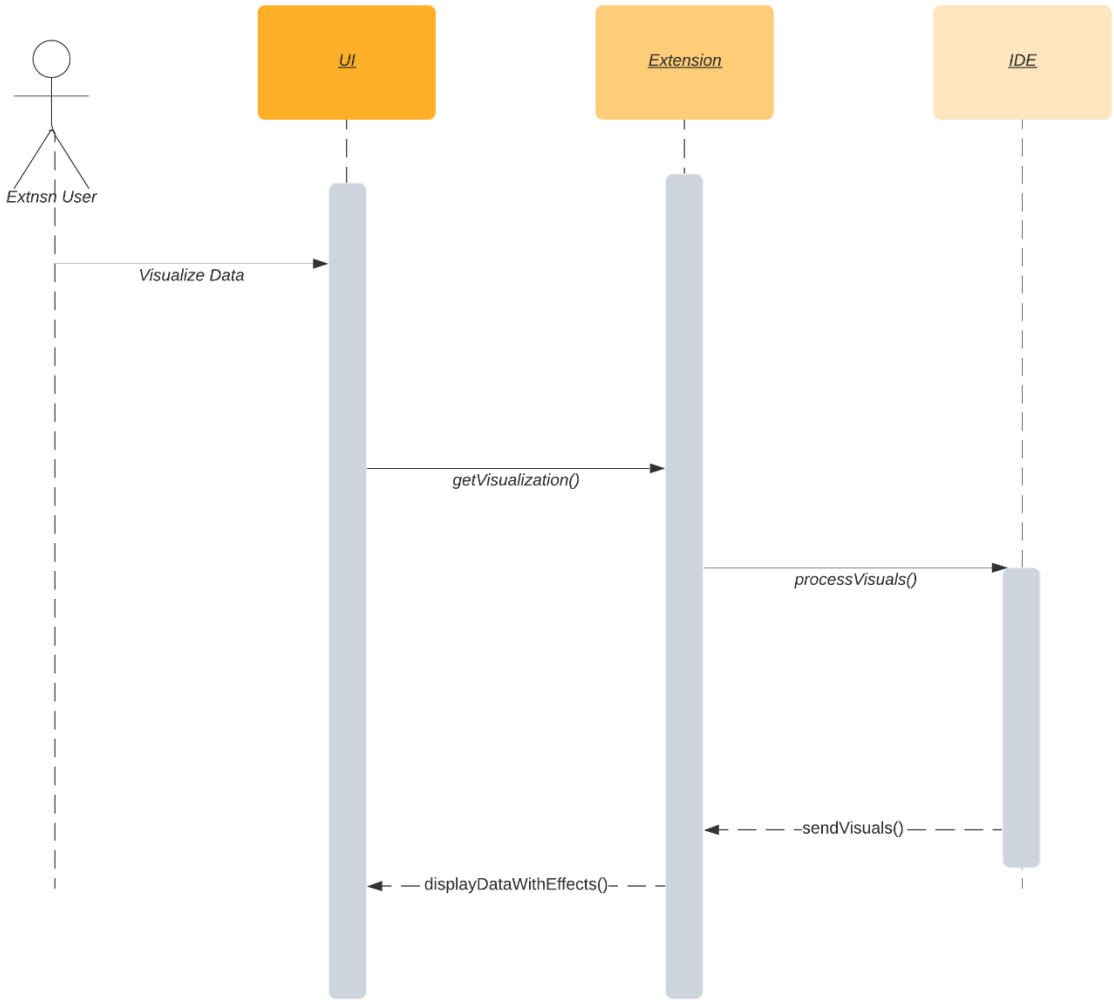


Figure 3.4.1f Visualization

© Daffodil International University

3.4.2 Data Cleaning

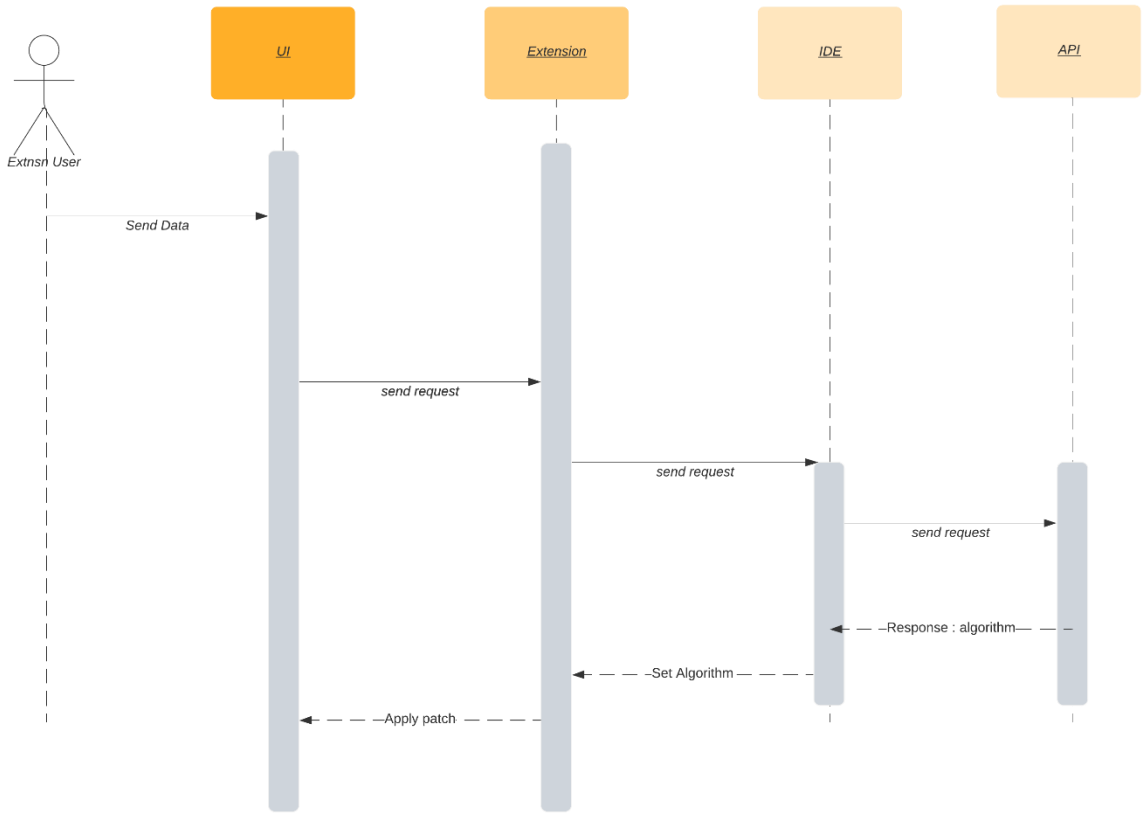


Figure 3.4.1f Data Cleaning

© Daffodil International University

© Daffodil International University

CHAPTER 4: SYSTEM TESTING

4.1 Feature Testing

Feature testing is considered to add or modify the new functionality to the existing system. Every feature and functionality have different characteristics. In my case I have listed the following feature to be tested.

4.1.1 Features to be tested

Features	Priority	Description
Data Visualization	1	The feature that is responsible for visualizing the excel data.
Validating Excel File	1	To validate if working with the correct file format.
Exporting	1	Exporting the file without damaging the data.
Deduplication	2	It removes the duplicate rows, one of the core functions that my extension supports.
Filling empty value with average/zero	2	It is another core feature which is responsible for handling empty fields.
Here, 1 = High Priority, 2 = Medium Priority, 3 = Low Priority		

© Daffodil International University

4.2 Testing strategies

4.2.1 Test Approach

I have two different types of testing to ensure the quality of my system. These two testing systems include functional and structural testing.

- Black Box Testing was used in my apps to track the pace between the expected output and desired output by changing the input values.
- White Box Testing was used to validate if the internal mechanism of my extension was working perfectly fine.

4.2.2 Pass/Fail Criteria

I have created a few input files and kept the expected output in a separate file. After processing I have compared both the expected output and actual output. The tests that got accuracy over 98% was considered as pass, any value lower than this was considered failed.

4.2.3 Testing Schedule

Test Phase	Time
Testing plan creation	1 week
Unit Testing	During development time
Component test	During development time
Integration testing	8 days
UI testing	4 days
Load testing	2 week
Performance testing	1 week

© Daffodil International University

Accessibility testing	1 week

4.3 Testing Environment

Testing environment means to prepare the environment with hardware and software so that testers can be able to execute test cases as required. The following was used to do the testing

- Test data
- Manually processed data
- API (Microsoft)
- Solid State Drive (To test with the faster transfer rate)
- Pen drive (To test the performance with slower transfer rate)
- Third Party Tools
- Visual Studio Code
- System and application

4.4 Test Cases

Test cases are those by which a tester can determine whether a system can be able to perform better under test cases properly.

© Daffodil International University

4.4.1 Data Visualization

Test case # 1			Test case name: Data Visualization			
Designed By: Shouman Barua Shuvo			Designed Date:			
Executed by:			Executed Date:			
Short description: The extension will display the excel file in its context						
Pre-conditions: ● The file must have to an excel file ● The file must contain plain data only						
Serial	File Type	Contains Plain Data		Expected result	Pass/Fail	Remarks
1	word document	NA		Show error message	Pass	
2	Excel File with images	No		Show error message	Pass	
3	Excel File with plain texts	Yes		Visualize the file	Pass	
Post-conditions: User will be able successfully open the file						

© Daffodil International University

4.4.2 Deduplication

Test case # 1			Test case name: Deduplication			
Designed By: Shouman Barua Shuvo			Designed Date:			
Executed by:			Executed Date:			
Short description: The extension will remove the duplicate rows from the extension						
Pre-conditions: ● The file must have to an excel file ● The file must contain plain data only ● Not mandatory but having duplicate rows is a plus point						
Serial	File Type	Contains Duplicates		Expected result	Pass/Fail	Remarks
1	word document	NA		Show error message	Pass	
2	Excel File	No		Processes the file	Pass	
3	Excel File	Yes		Processes the file	Pass	
Post-conditions: User will get a new excel file with no duplicate values						

© Daffodil International University

4.4.3 Cleaning Empty Rows

Test case # 3	Test case name: Cleaning Empty Rows
Designed By: Shouman Barua Shuvo	Designed Date:
Executed by:	Executed Date:
Short description: The extension will remove the empty rows from the excel file	
Pre-conditions: ● The file must have to an excel file ● The file must contain plain data only ● Not mandatory but having duplicate empty is a plus point	

Serial	File Type	Contains Empty Rows		Expected result	Pass/Fail	Remarks
1	word document	NA		Show error message	Pass	
2	Excel File	No		Processes the file	Pass	
3	Excel File	Yes		Processes the file	Pass	

Post-conditions: User will get a new excel file with no duplicate values

© Daffodil International University

4.4.4 Fill Missing Values with Zero

Test case # 4			Test case name: Fill Missing Values with Zero			
Designed By: Shouman Barua Shuvo			Designed Date:			
Executed by:			Executed Date:			
Short description: The extension will fill the column values that are empty with a value of zero						
Pre-conditions: ● The file must have to an excel file ● The file must contain plain data only ● The column should contain numeric data only						
Serial	File Type	Contains Numeric Data		Expected result	Pass/Fail	Remarks
1	word document	NA		Show error message	Pass	
2	Excel File	No		Processes the file	Pass	
3	Excel File	Yes		Processes the file	Pass	
Post-conditions: User will get a new excel file with empty value filled with zero						

© Daffodil International University

4.4.5 Fill Missing Values with Average

Test case # 5	Test case name: Fill Missing Values with Average
Designed By: Shouman Barua Shuvo	Designed Date:
Executed by:	Executed Date:
Short description: The extension will fill the column values that are empty with a value of the average of the column.	
Pre-conditions: ● The file must have to an excel file ● The file must contain plain data only ● The column should contain numeric data only	

Serial	File Type	Contains Numeric Data	Selected row doesn't exist	Expected result	Pass/Fail	Remarks
1	word document	NA		Show error message	Pass	
2	Excel File	No	True	Processes the file	Pass	
3	Excel File	Yes	False	Processes the file	Pass	
4	Excel File	NA	True	Shows error message	Pass	

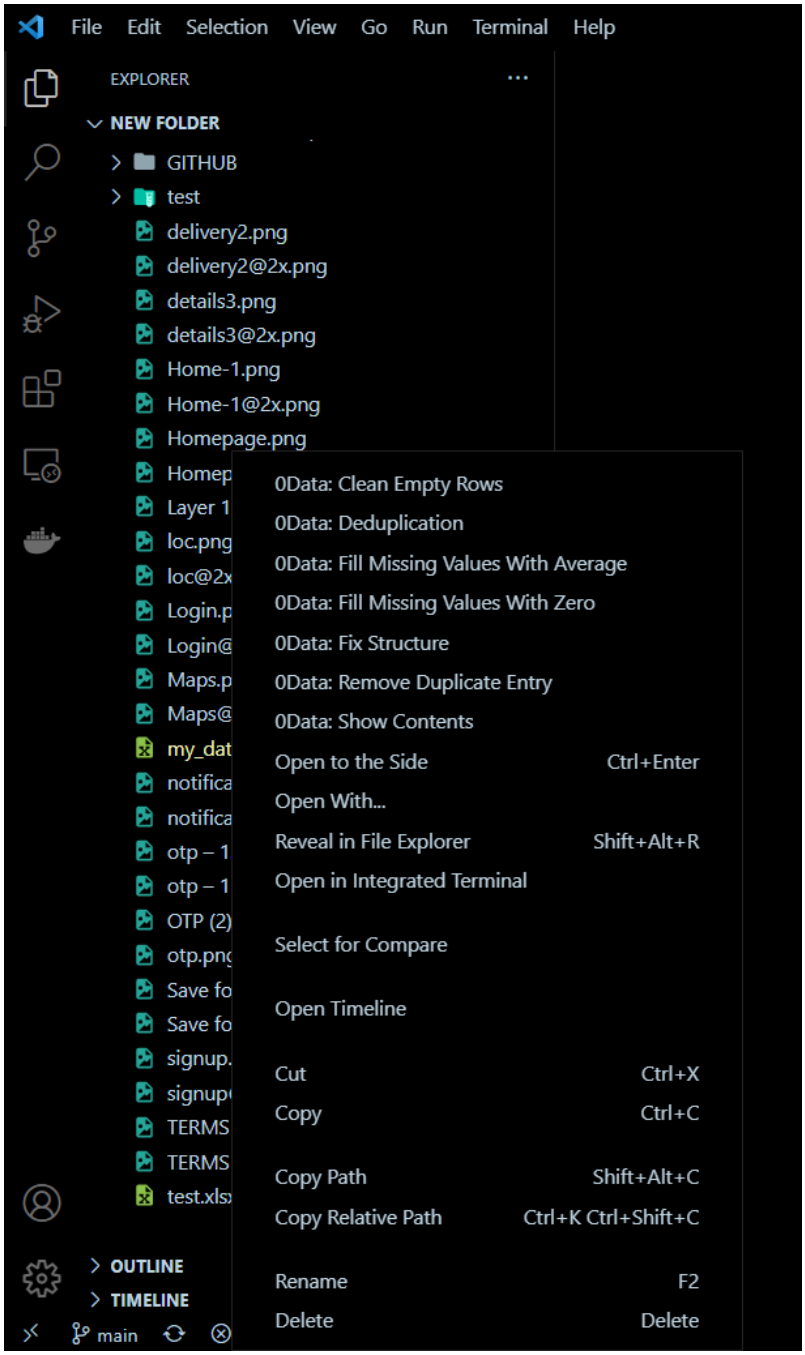
Post-conditions: User will get a new excel file with empty value filled with the average value

© Daffodil International University

CHAPTER 5: USER MANUAL

5.1 Context Menu

The list of operation that can be done with the current version of my extension.



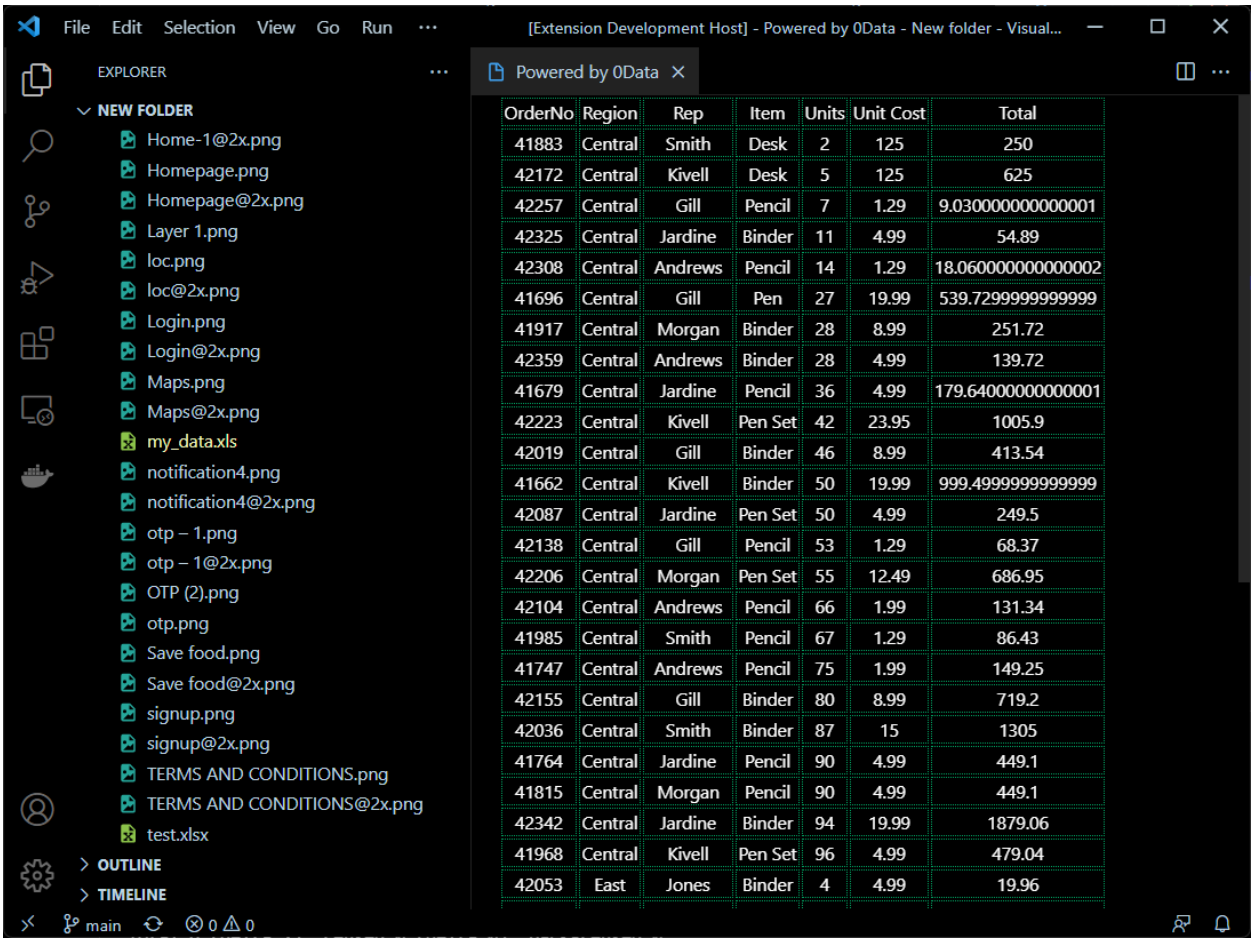
© Daffodil International University

Figure 5.1f: Context Menu

This the set of actions that will be available in the menu
that appears when an user clicks a file.

5.2 Visualizing Data

No third-party extension needed. You can use 0Data's default viewer to visualize the excel sheet without reopening MS Excel (or similar app) again and again. 0Data can handle simple excel files.



OrderNo	Region	Rep	Item	Units	Unit Cost	Total
41883	Central	Smith	Desk	2	125	250
42172	Central	Kivell	Desk	5	125	625
42257	Central	Gill	Pencil	7	1.29	9.030000000000001
42325	Central	Jardine	Binder	11	4.99	54.89
42308	Central	Andrews	Pencil	14	1.29	18.060000000000002
41696	Central	Gill	Pen	27	19.99	539.7299999999999
41917	Central	Morgan	Binder	28	8.99	251.72
42359	Central	Andrews	Binder	28	4.99	139.72
41679	Central	Jardine	Pencil	36	4.99	179.64000000000001
42223	Central	Kivell	Pen Set	42	23.95	1005.9
42019	Central	Gill	Binder	46	8.99	413.54
41662	Central	Kivell	Binder	50	19.99	999.4999999999999
42087	Central	Jardine	Pen Set	50	4.99	249.5
42138	Central	Gill	Pencil	53	1.29	68.37
42206	Central	Morgan	Pen Set	55	12.49	686.95
42104	Central	Andrews	Pencil	66	1.99	131.34
41985	Central	Smith	Pencil	67	1.29	86.43
41747	Central	Andrews	Pencil	75	1.99	149.25
42155	Central	Gill	Binder	80	8.99	719.2
42036	Central	Smith	Binder	87	15	1305
41764	Central	Jardine	Pencil	90	4.99	449.1
41815	Central	Morgan	Pencil	90	4.99	449.1
42342	Central	Jardine	Binder	94	19.99	1879.06
41968	Central	Kivell	Pen Set	96	4.99	479.04
42053	East	Jones	Binder	4	4.99	19.96

Figure 5.2f: Visualizing Data

The action shows the excel file in a basic row and column so that the
user don't need any external excel viewer for basic excel formats.

© Daffodil International University

5.3 Removing Empty Rows

Clicking the action will remove any rows that doesn't contain any value

Before

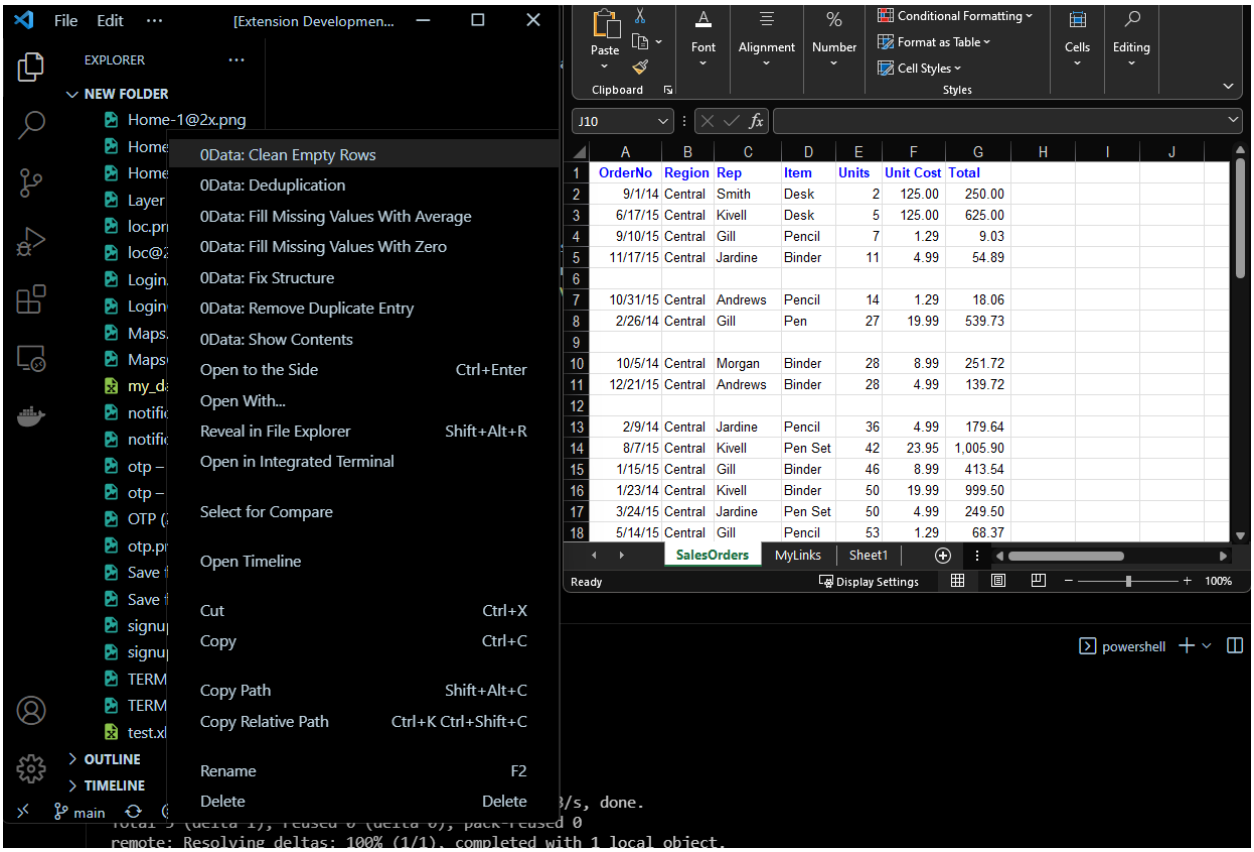


Figure 5.3.1f: Removing Empty Rows

The row number 6,9 and 12 is empty here which will be removed in this action.

© Daffodil International University

After

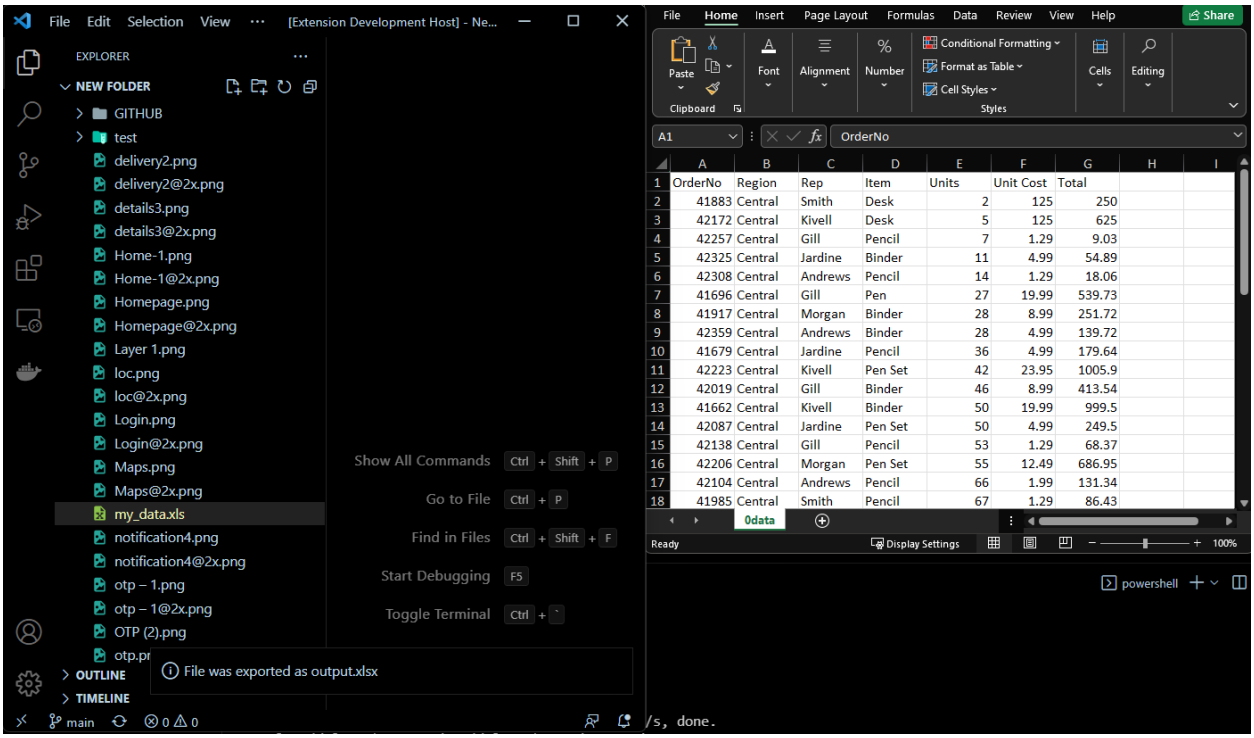


Figure 5.3.2f: Removing Empty Rows (After)

After removing have a look at the rows that were highlighted in the previous figure. This excel file doesn't contain any empty row now.

© Daffodil International University

5.4 Deduplication

Deduplication removes the duplicate rows that are exactly same

Before

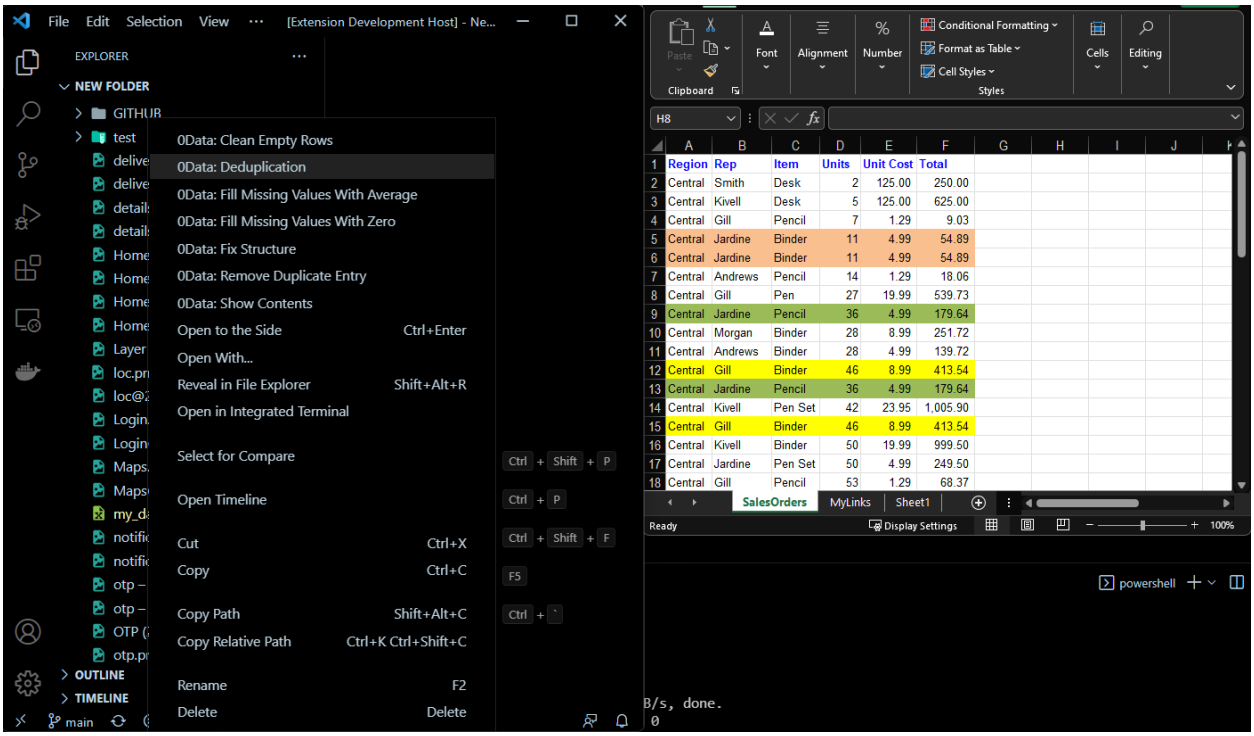


Figure 5.4.1f: Deduplication

There are some rows that contain exactly same values
These rows are highlighted here.

© Daffodil International University

After

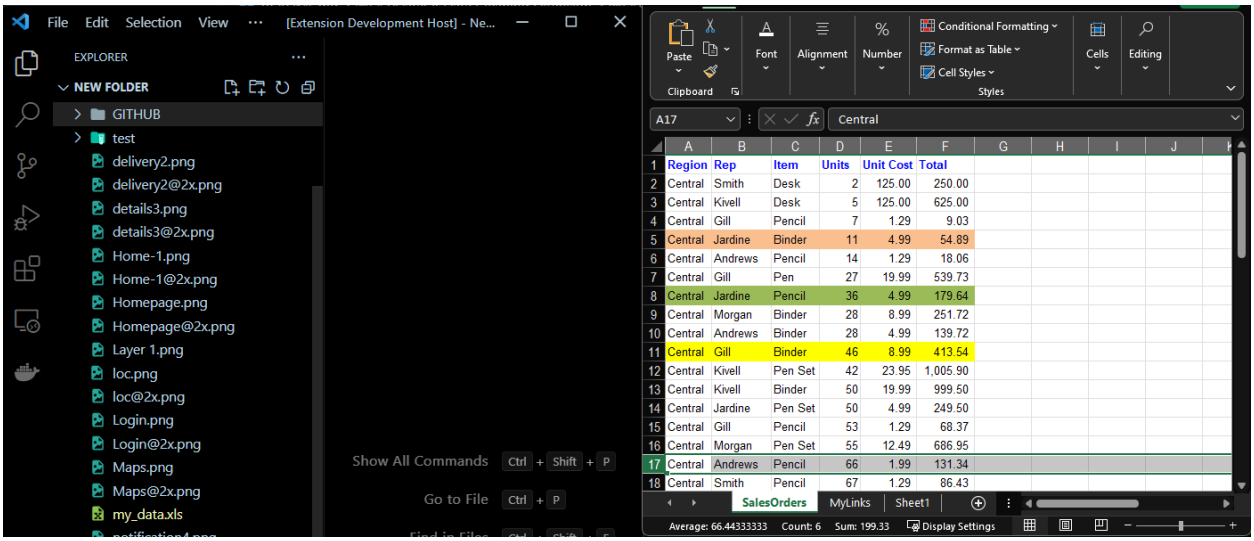


Figure 5.4.2f: Deduplication (After)

The highlighted rows are now removed in this excel file.

© Daffodil International University

5.5 Fill Missing Values with Zero/Average

In case there are columns that has empty fields this field comes handy.

Before

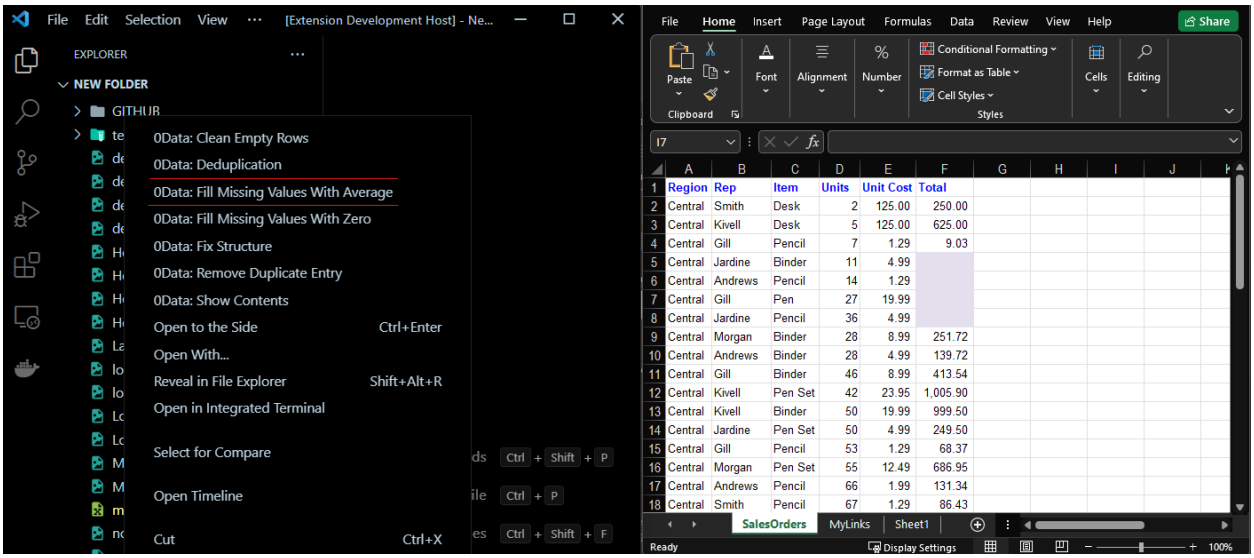


Figure 5.5.1f: Fill Missing Values with Zero/Average

The highlighted fields don't contain any value
which may create exception when dealing with this data.

© Daffodil International University

After

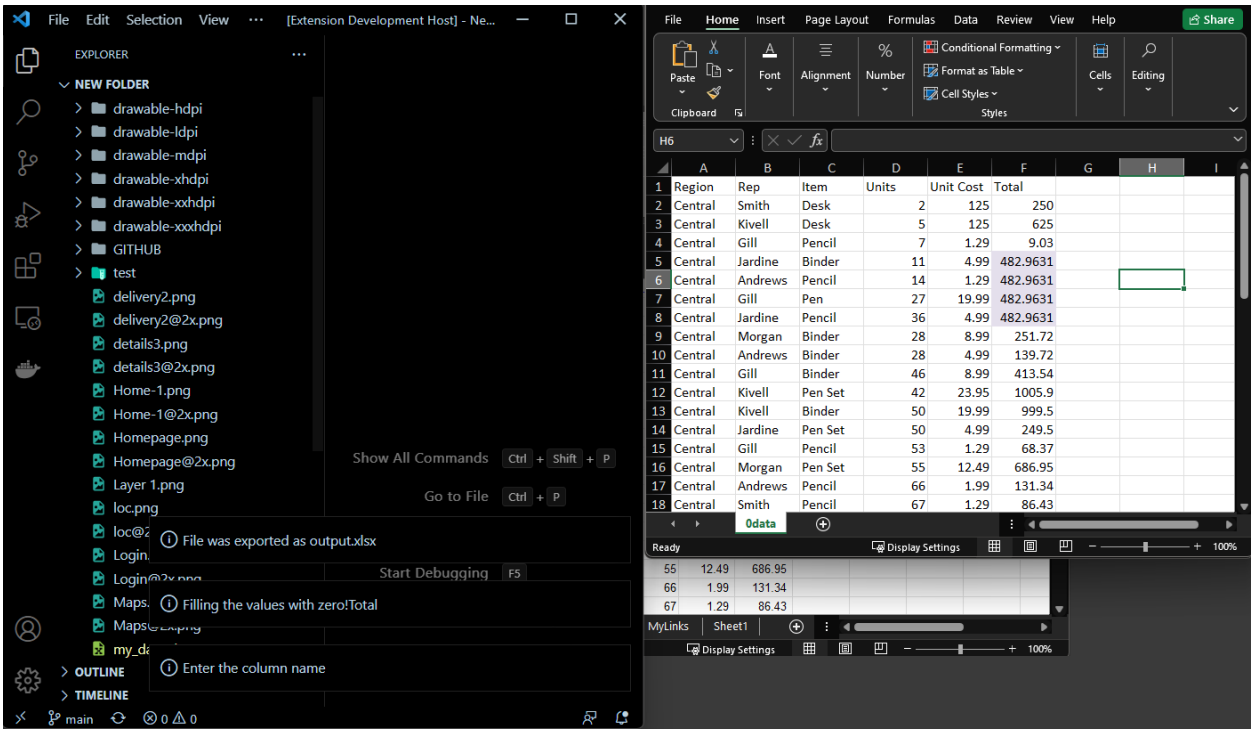


Figure 5.5.2f: Fill Missing Values with Zero/Average (After)

After the operation the fields are now filled with their average or zero.

© Daffodil International University

© Daffodil International University

Chapter 6: Conclusion

6.1 Project Summary

The project was an IDE based project which can be a helping hand to the data scientist if extended properly. I had the idea in mind and designing the base of this platform was the goal. I was able to achieve my goal in this project. The project was made offline as well as offline keeping the privacy of data in mind as data can be sensitive. The project is open for contributing in the github and if you look at the code, you will see that the code was documented well. Each method, their parameter and return type is explained properly so that it becomes easy for the next developer to modify / extend the project.

6.2 Limitations

As this is pretty basic version of the extension, I could not include all the rich features that currently premium applications are providing. Which may include:

- Being limited to excel files only
- Not being able to handle complex scenario

6.3 Obstacles and Achievements

Having no such extension like this before make itself an achievement for me. I will say being the very first data cleaning extension deserves appreciations. I had a lot of obstacles like not getting the proper documentation, selecting open source libraries and so on. The difficulties were not easy to solve but with the time being I was able to solve them.

6.4 Future Scopes

If this gets acceptability in the developer community it has lots of scopes in the future. One of its features was extendibility.

© Daffodil International University

The extension can more options in the context menu and I do believe will get extended with the time being.

© Daffodil International University

Chapter 6: Reference

1. Visual Studio Code API Documentation
   Reference url: https://code.visualstudio.com/api/references/vscode-api
2. WebView API for Visual Studio Code
   Reference url: https://code.visualstudio.com/api/extension-guides/webview
3. Visual Studio Code, Extension Development Documentation
   Reference url: https://code.visualstudio.com/api/get-started/your-first-extension

© Daffodil International University