Configuration Manual

MSc Research Project
Data Analytics

Nayan Sharma
Student ID: X21167818

School of Computing
National College of Ireland

Supervisor: Prof. Paul Stynes

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Nayan Sharma
Student ID:	X21167818
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Prof. Paul Stynes
Submission Due Date:	15/12/2022
Project Title:	Configuration Manual
Word Count:	1391
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	15th December 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Nayan Sharma
ID: 21167818
MSCDAD JAN22 A

December 2022

1 Introduction

This paper has a detailed, step-by-step plan for how to carry out this research study. It will explain everything I did to finish my study and how those steps can be done again.

1.1 Prerequisites for conducting the research project

To conduct this research project, make sure you have an Azure and Datarbricks account set up and a small cost that would require you to use these services. There are very few prerequisites as the main aim of the research project was to make it simpler and automated following the best practices of security. Hence, we will do the most out of Azure services and integrate a few third-party libraries such as scikit-learn-1.1.3, Seaborn-0.12.1, and yellowbrick-1.5. TO integrate these libraries automated code is already written in .ipynb files.

- Using url [1] create Azure account
- Using url [2] create databricks free account where azure account credentails are required.
- Using url [3] create Github free account where code and dataset will be uploaded

2 Setup Environment

2.0.1 Setup Databricks environment

Setup the Databricks cluster as per Table 1 and create a cluster in data bricks as shown in Figure 1.

[1] https://azure.microsoft.com/en-gb/free/
[2] https://www.databricks.com/try-databricksaccount
[3] https://github.com/signup?source=login

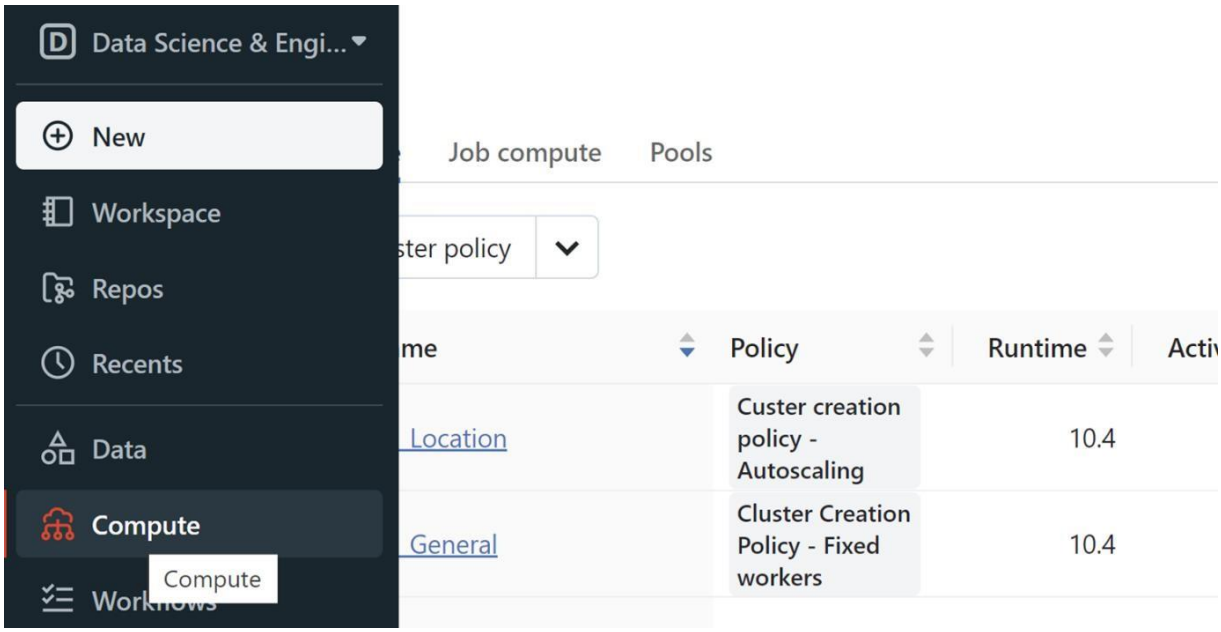

Figure 1: Databricks Cluster

Table 1: Databricks Cluster Specification

Type	Value
Databricks Runtime Version	9.1 LTS (includes Apache Spark 3.1.2, Scala 2.12)
Worker Type	Standard D8s v3 (32 GB, 8 Cores)
Driver Type	Standard D8s v3 (32 GB, 8 Cores)
Number of workers	Autoscale (1-4)

2.0.2 Setup Azure environment

Setup the Azure Key to access the Storage account and contains in Azure Data lake storage as shown in Figure 2.

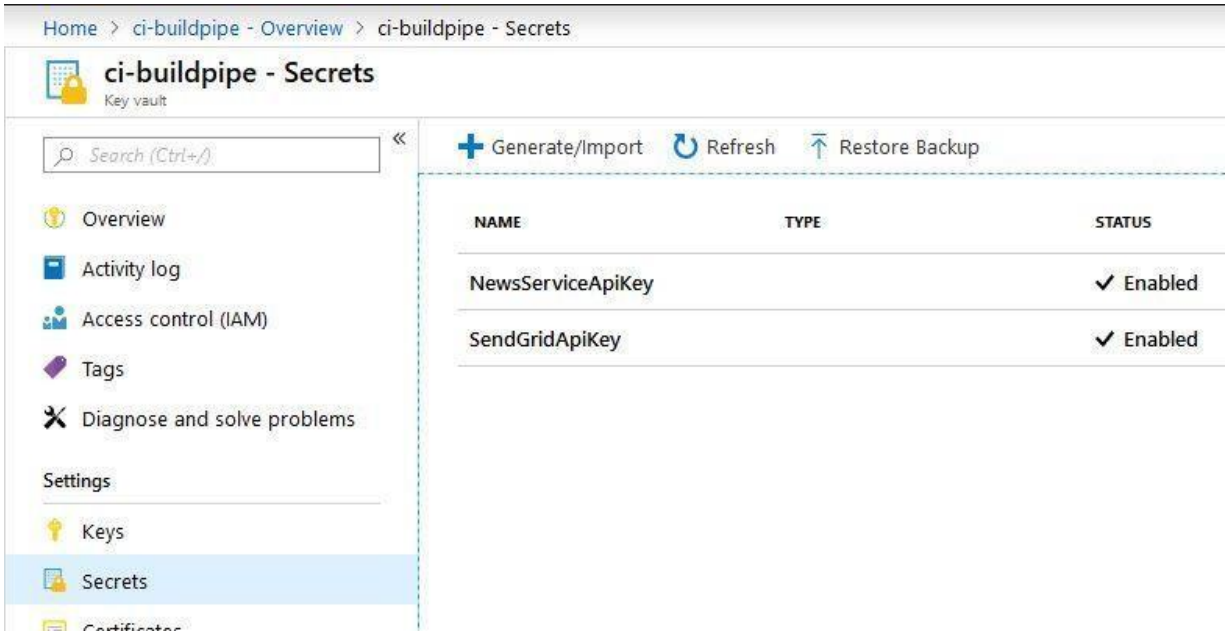

Figure 2: Azure key setup

2.0.3 Setup GitHub

Integrate the GitHub account with Azure data bricks by going into the data bricks user setting as shown in Figure 3. Before that, we need to create a repository for the code and generate a token/key from GitHub to authenticate the user from Databricks. We can follow standard documentation from url [4]. Either you can upload all .ipynb files into databricks as shown in Figure 4 or upload file in GIT and add repo into Databricks as shown in Figure 5

2.0.4 Setup databricks cli

- Install latest version of python from [5] into personal laptop.
- Install data bricks cli by using the command into PowerShell or cmd terminal is shown in Figure 6.

[4]https://docs.databricks.com/repos/index.html
[5]https://www.python.org/downloads/

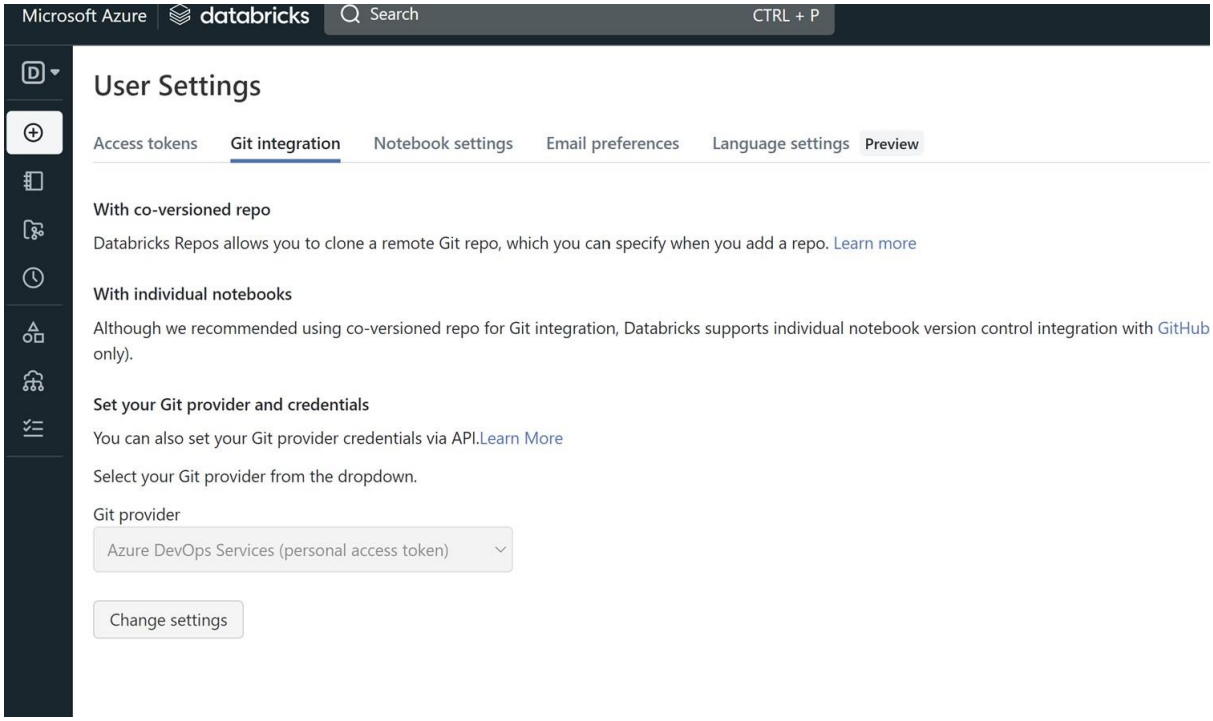

Figure 3: GitHub Integration

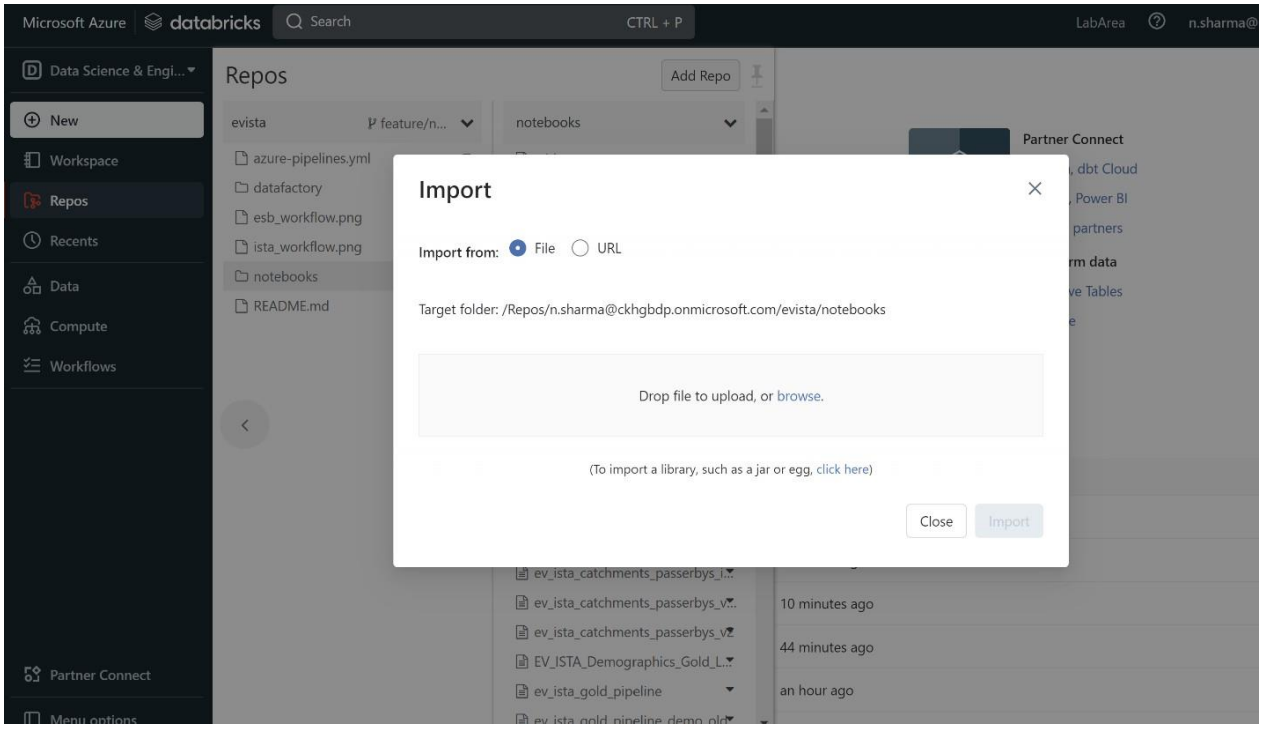

Figure 4: Import .ipynb in Databricks

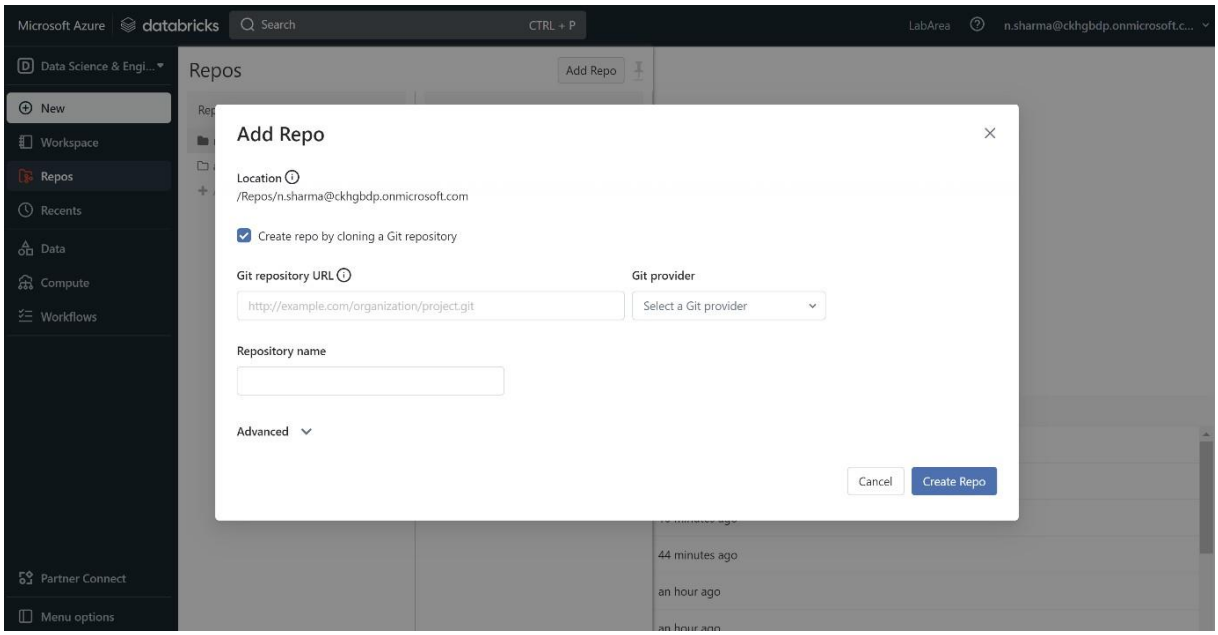

Figure 5: Add GIT Repo into Databricks

```
pip install databricks-cli
```

Figure 6: Cli Installation

- Setup secret scope in data bricks to access azure with the personal token is shown in Figure 7.

```
databricks secrets create-scope --scope <your_name_of_the_scope>
```

Figure 7: Scope in Databricks

- Check the list of scopes already created is shown in Figure 8.

```
databricks secrets list-acls --scope <your_name_of_the_scope>
```

Figure 8: List Scope

- Put the principle id and key created in Azure as scope in databricks is shown in Figure 9.

```
databricks secrets put-acl --scope <your_name_of_the_scope> --principal <microsoftusername>@
```

Figure 9: Scope creation

2.0.5 Create Storage account and containers in Azure

Create a storage account and a container in Azure to store datasets and intermediate and output external tables data to feed into the Spatial model as shown in Figure 10

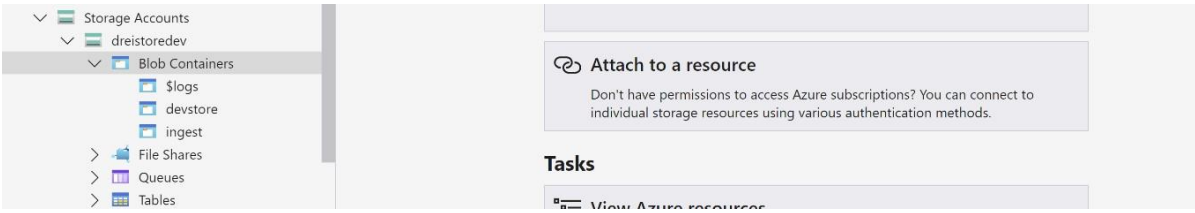

Figure 10: Azure Container creation

3 Collecting data-sets

For this research, we are using seven datasets 1. UK postcode dataset.csv (*uk-postcode-electricity-consumption* 2022) 2. UK Car ownership dataset.csv (Transport 2022) 3. Mobility Footfall dataset.csv(*Uk Mobility Footfall* (2022)) 4. Mobility passerby dataset.csv(*Uk Mobility passerby* (2022)) 5. UK traffic count.csv (*Traffic count—data.gov.uk* (2022)) 6. UK Postcode mapping.csv (*uk postcode – NSPL* (2022)) 7. UK Outputareas with population.geojson (*uk postcode geojson* (2022)). There is a requirement to upload all these datasets to Azure containers as shown in Figure 11 so that Databricks notebook can consume data for processing and feed it into the model.

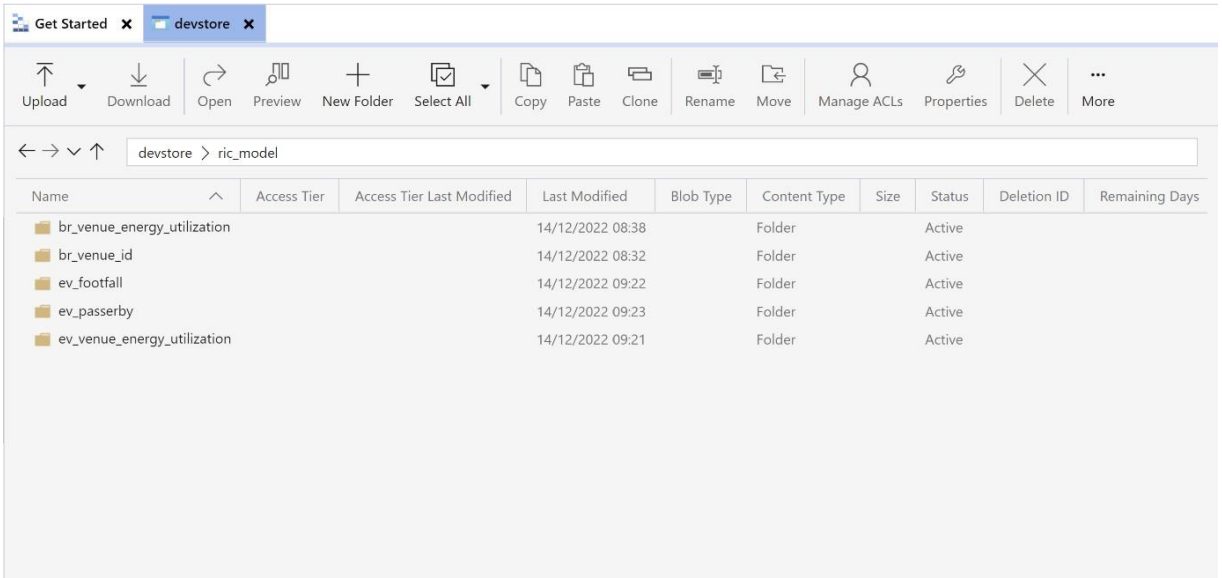

Figure 11: Files at Azure

4 Notebooks

There are eight Notebooks involved in this research which are described below:

- ev_static_csv_file_load.ipynb
- ev_carownership_excel_ingestion.ipynb
- ev_outputareas_carownership_ingestion.ipynb
- ev_model_dataset_creation.ipynb
- ev_create_dataset_traffic_carownershp.ipynb
- ev_unsupervised_model_monthly.ipynb
- ev_regression_model_monthly_data.ipynb
- ev_spatial_model.ipynb

4.1 ev_static_csv_file_load.ipynb

This is the first notebook that should run. The main purpose of this notebook is to create a database in data bricks with name ric_model use all the necessary datasets and create a final table that acts as input to ev_model_dataset_creation.ipynb. You have to run this notebook multiple times with the table name and path of the files uploaded on Azure as shown in Table 2. Figure 12 shows the notebook where we need to mention the path and table name.

Table 2: Table Creation

File Path	Table Name
/Raw/ric/ev passerby.csv	ric model.ev passerby
/Raw/ric/ev footfall.csv	ric model.ev footfall
/Raw/ric/ev passerby.csv	ric model.ev passerby
/Raw/ric/energy utilization.csv	ric model.ev venue energy utilization
/Raw/ric/post code uk.csv	ric model.br postcodes uk
/Raw/ric/traffic count.csv	ric model.br traffic counts uk

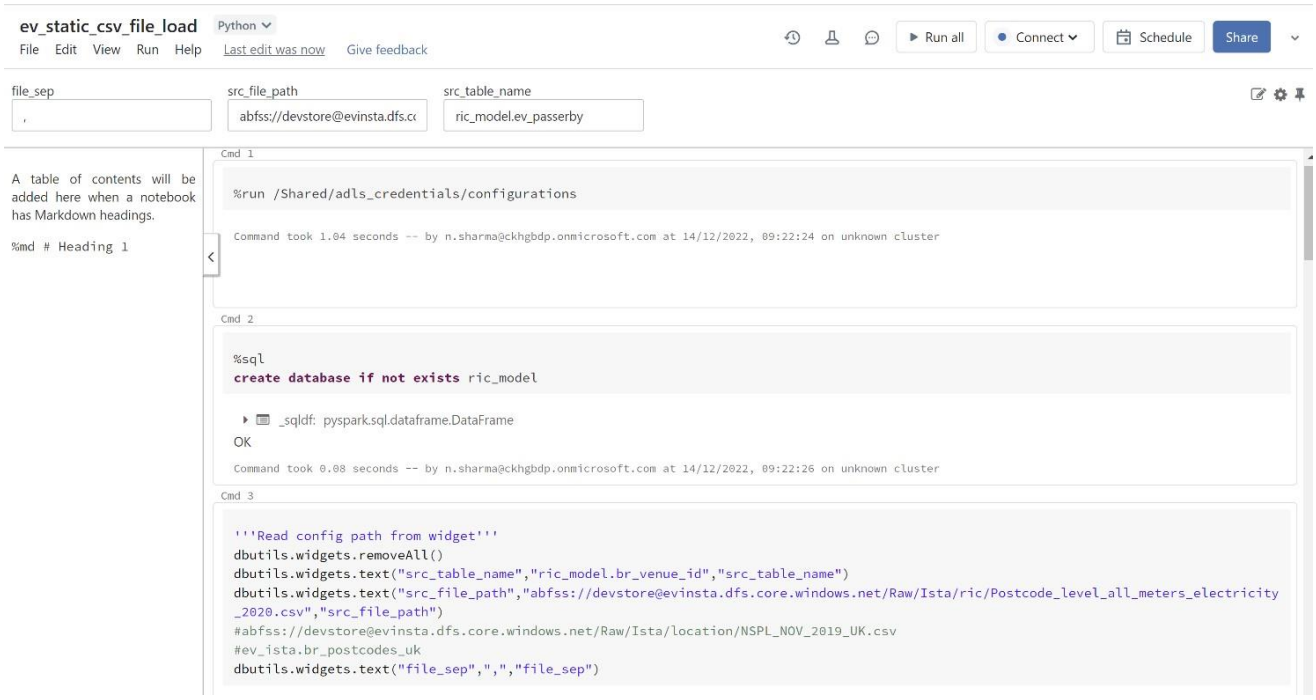

Figure 12: ev_static_csv_file_load

4.2 ev_carownership_excel_ingestion.ipynb

The main purpose of this notebook is to download car ownership datasets from the [6] and load it into stage table ric_model.br_carownership_stat. The notebook as shown in Figure 13.

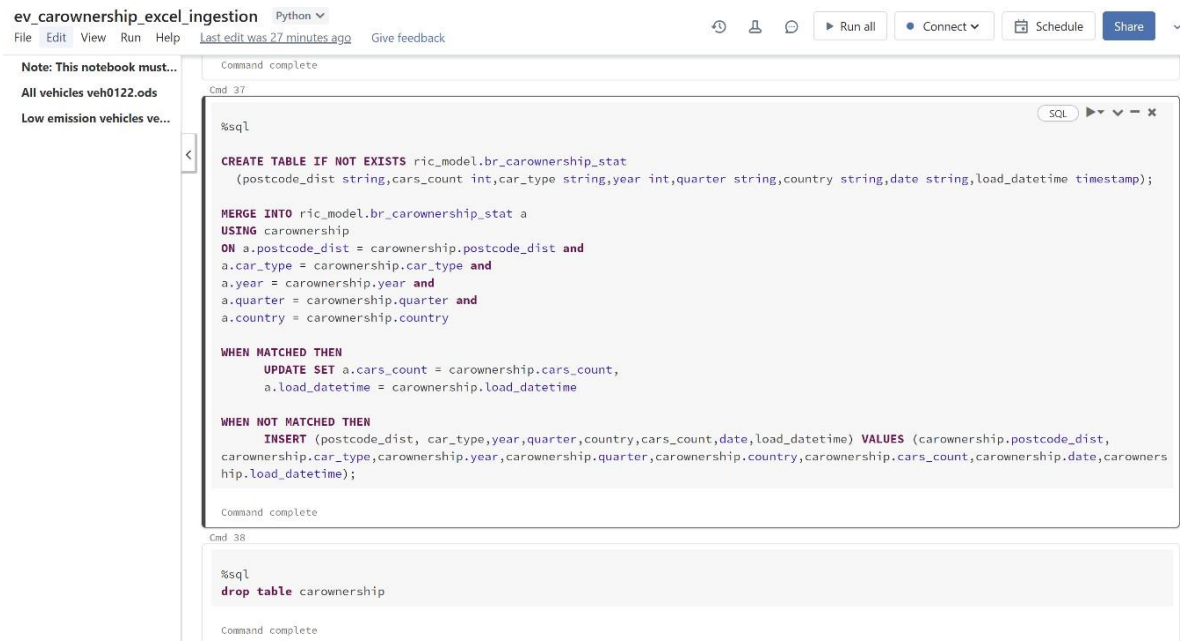

Figure 13: ev car ownership excel ingestion

[6]https://www.gov.uk/government/collections/vehicles-statistics

4.3 ev_outputareas_carownership_ingestion.ipynb

The main purpose of this notebook is to use ric_model.br_carownership_stat table and load it into the final table ric_model.br_outputareas_carownership_stats after mapping it with the postcode. The notebook as shown in Figure 14.

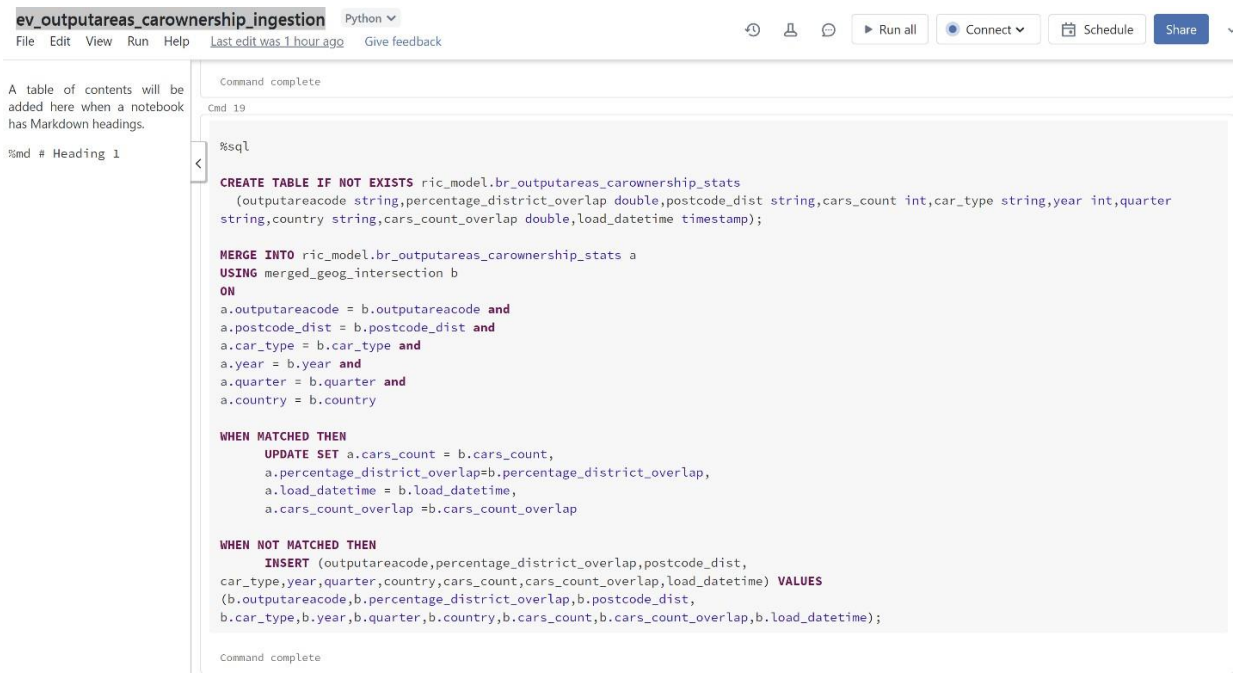

Figure 14: ev_outputareas_carownership_ingestion

4.4 ev_model_dataset_creation.ipynb

This is the second notebook that should run. The main purpose of this notebook is to use energy consumption, passerby, and footfall, and aggregate them all and create a single output table with the name ric_model.ev_model_dataset_monthly that acts as input to various machine-learning models. The notebook is shown in Figure 15.

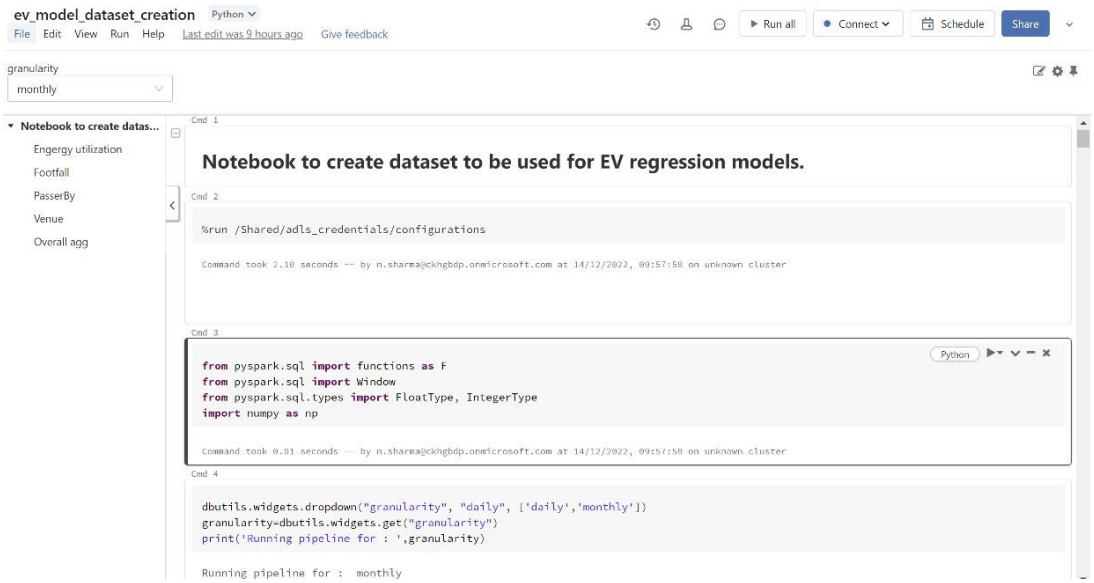

Figure 15: ev_model_dataset_creation

4.5 ev_create_dataset_traffic_carownershp.ipynb

This is the third notebook that should run. The main purpose of this notebook is to combine traffic count, car ownership, and output table from the second notebook i.e ric_model.ev_model_dataset_monthly. Output is ric_model.br_ev_model_dataset which contains the total motor and total car count. The notebook is shown in Figure 16.

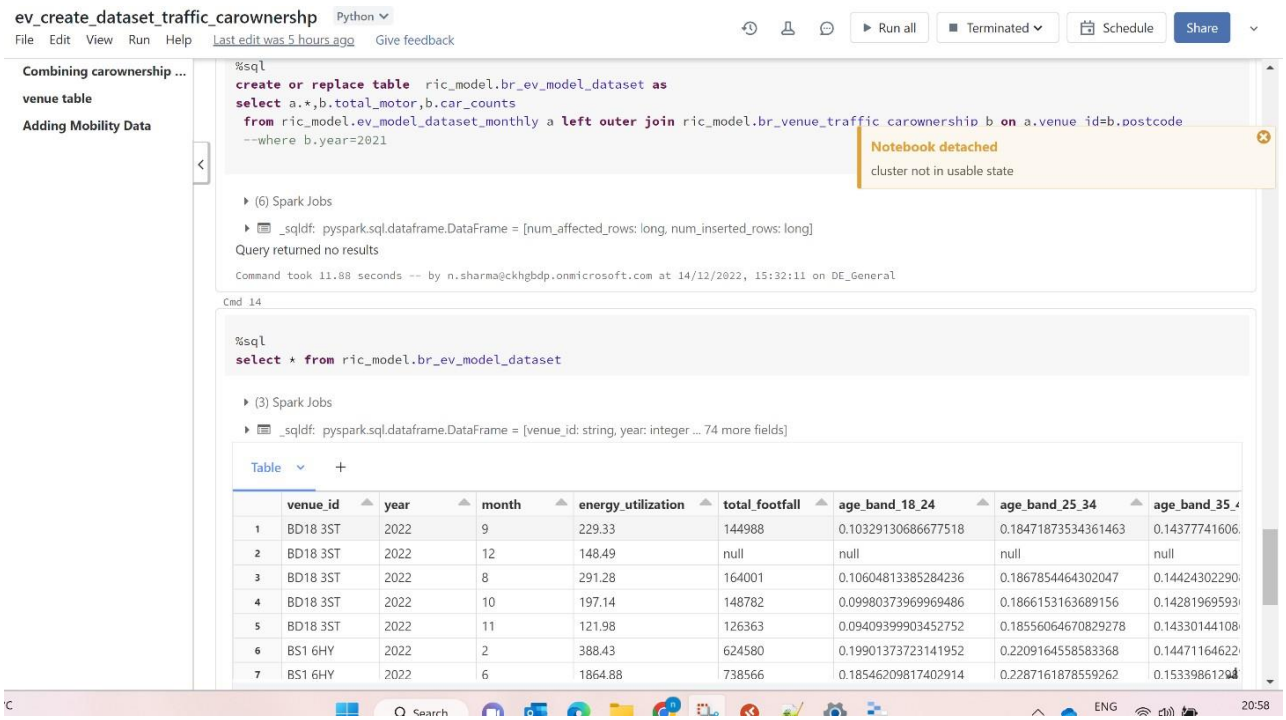

Figure 16: ev_create_dataset_traffic_carownershp

4.6 ev_unsupervised_model_monthly.ipynb

The main purpose of this notebook is to take the ric_model.br_ev_model_dataset table created and feed into the KMean clustering model. Before feeding it into the Machine learning model EDA and min max scaler have been performed. The notebook is shown in Figure 17.

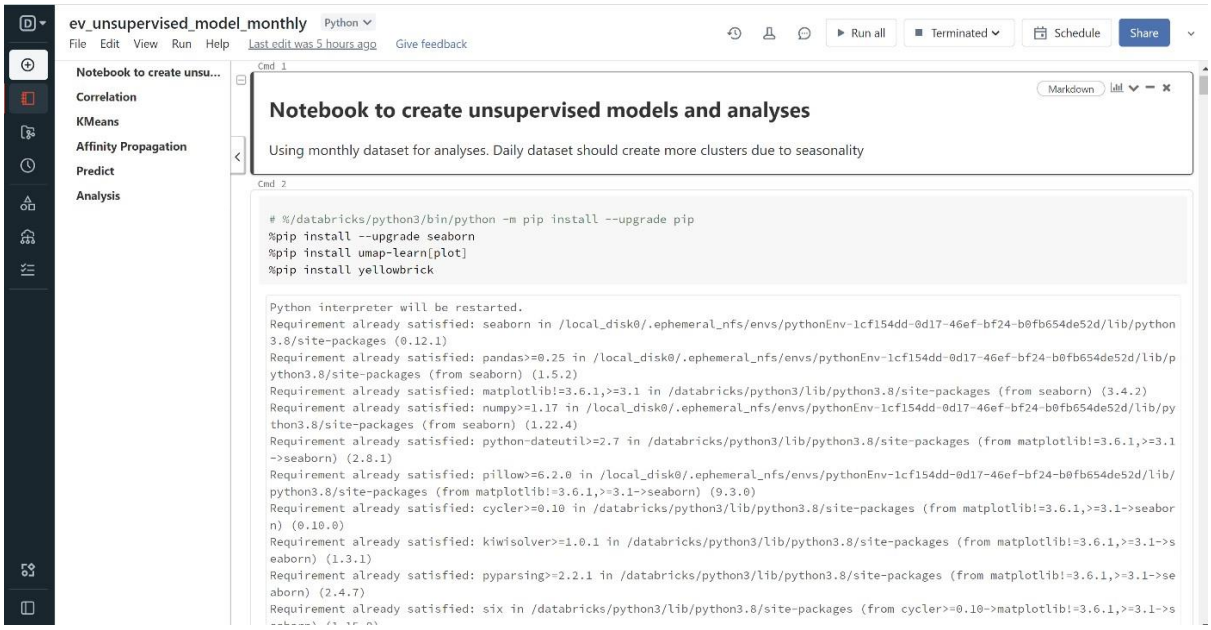

Figure 17: ev_unsupervised_model_monthly

4.7 ev_regression_model_monthly_data.ipynb

The main purpose of this notebook is to take the ric_model.br_ev_model_dataset table created in the first step as input, feed it into six machine-learning models, and save the output into a table that acts as input for the spatial model table ric_model.regression_model_output. The notebook is shown in Figure 18.

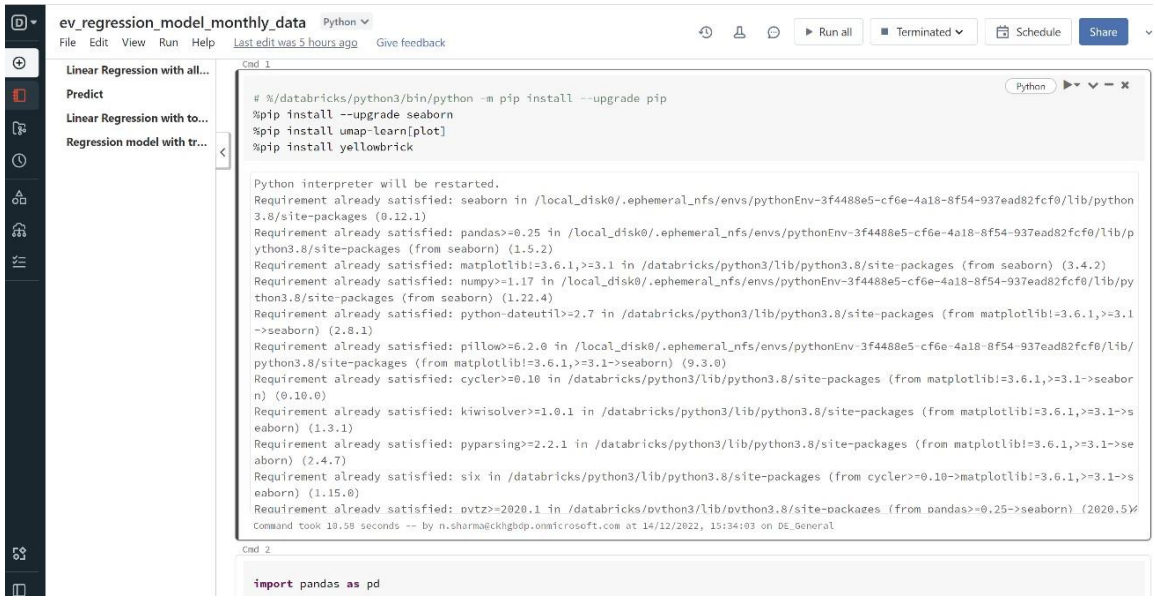

Figure 18: ev_regression_model_monthly_data

4.8 ev_spatial_model.ipynb

The main purpose of this notebook is to take the final output of the model and run the spatial model. The output of this notebook is. a JPEG file that contains the locations where to place Electric Vehicle charging stations are as per the energy utilization ranking. The notebook is shown in Figure 19.

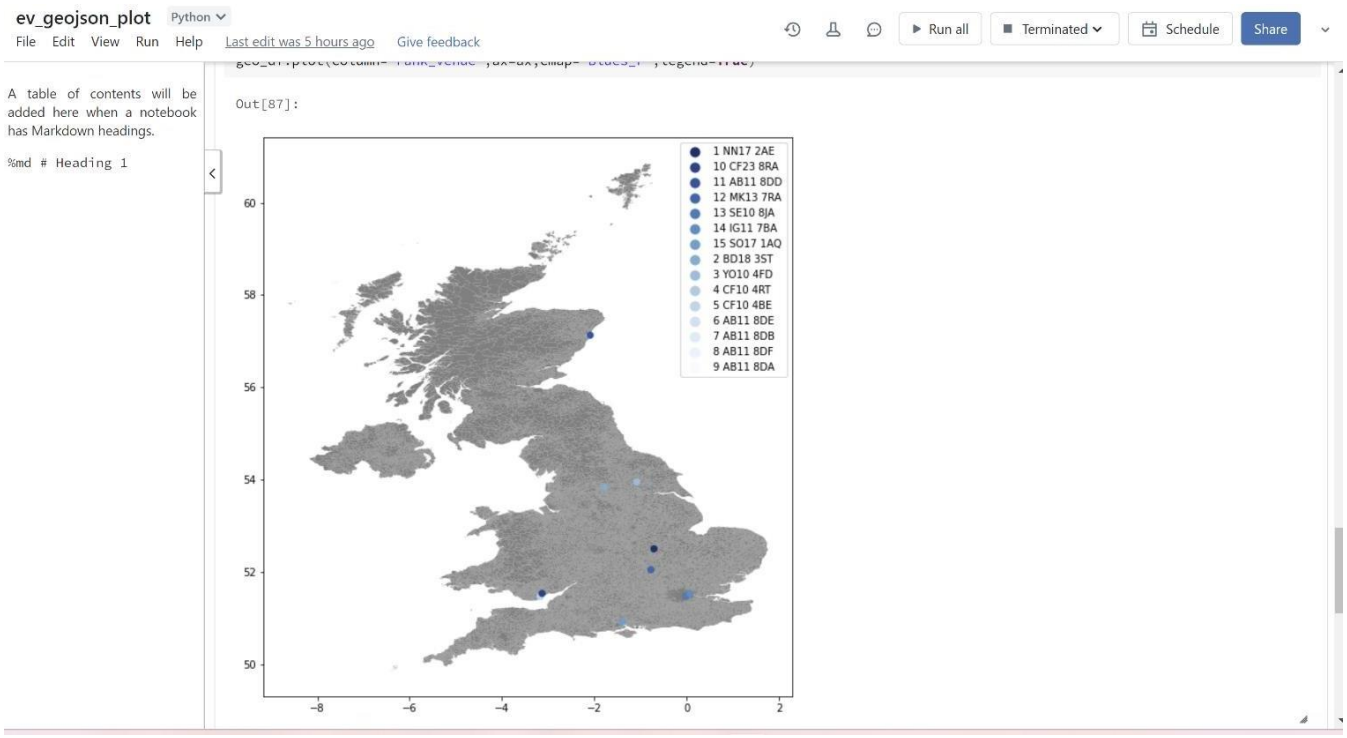

Figure 19: ev spatial model

5 Pipeline Setup

The pipeline of notebooks is set up in the Databricks workflow manager as shown in Figure 20. The pipeline contains all the notebooks in the below order:

- ev_static_csv_file_load.ipynb
- ev_model_dataset_creation.ipynb
- ev_create_dataset_traffic_carownershp.ipynb
- ev_unsupervised_model_monthly.ipynb
- ev_regression_model_monthly_data.ipynb
- ev_spatial_model.ipynb

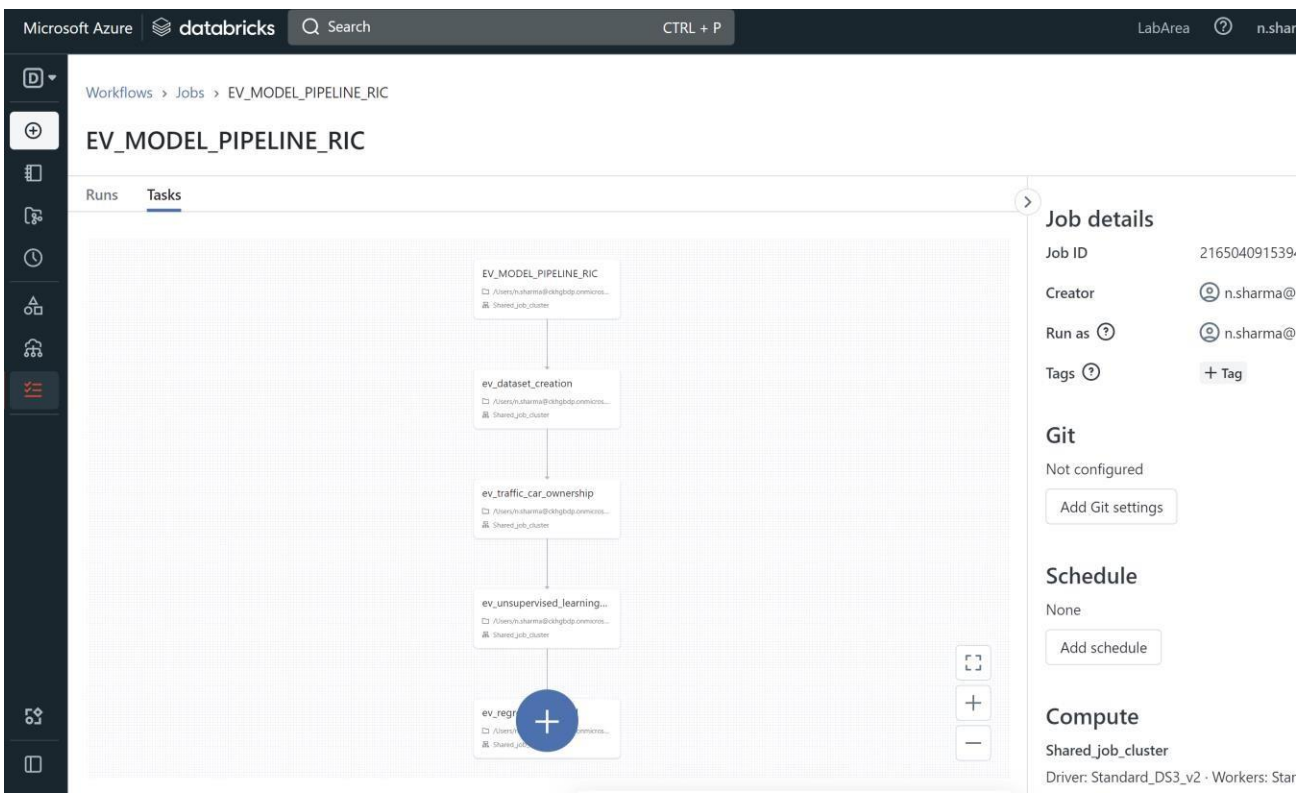

Figure 20: Databricks Worflow

6 Resources URL

Artefact:

https://studentncirl-my.sharepoint.com/:f:/g/personal/x21167818_student_ncirl_ie/EqSYdxrs8rBDjjQ7278KkT8BSAsbnFseK5UclUafd4LTUA

References

Traffic$_c$ount−data.gov.uk(2022), *https* : *//www.data.gov.uk/dataset/*208*c*0*e*7*b* − 353*f* − 4*e*2*d* − 8*b* *Nov* − 2022].

Transport, D. f. (2022), 'Vehicles statistics'.
URL: *https://www.gov.uk/government/collections/vehicles-statistics*

Uk Mobility Footfall (2022).
URL: *https://my.api.mockaroo.com/ev$_f$ ootfall.json*?*key* = 9*eb*8*e*7*e*0

Uk Mobility passerby (2022).
URL: *https://api.mockaroo.com/api/67042b30?count=1000key=9eb8e7e0*

uk-postcode-electricity-consumption (2022), https://www.data.gov.uk/dataset/e7d4c1cf-45a0-4070-878f-24ad9641f655/domestic-electricity-and-gas-estimates-by-postcode-in-great-britain. [Accessed 24-Nov-2022].

uk$_p$ostcode$_g$eojson(2022), *https* : *//geoportal.statistics.gov.uk/.*[*Accessed*24− *Nov* − 2022].

uk$_p$ostcode−*NSPL*(2022), *https* : *//geoportal.statistics.gov.uk/datasets/national* − *statistics* − *Nov* − 2022].