

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.

<u>ALL</u> 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 url¹ 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.

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

D Data Science & Engi▼				
① New	Job compute	Pools		
Workspace	ster policy 🗸 🗸			
ြန္နာ Repos	ster policy			
() Recents	me	\$	Policy	
 o□ Data	Location		Custer creation policy - Autoscaling	10.4
🔒 Compute	General		Cluster Creation Policy - Fixed	10.4
≚ Worknows			workers	

Figure 1: Databricks Cluster

Туре	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)

Table 1: Databricks Cluster Specification

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.

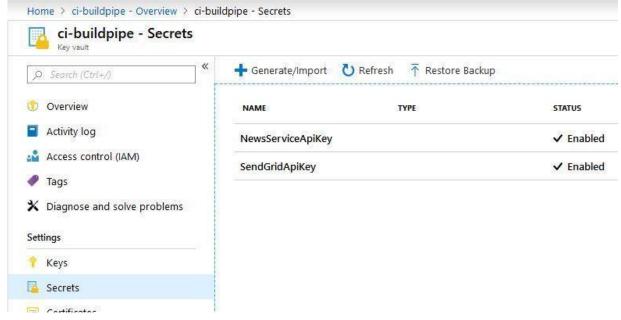


Figure 2: Azure key setup

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.

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

Micros	oft Azure databricks Q Search CTRL + P
•	User Settings
Ŧ	Access tokens Git integration Notebook settings Email preferences Language settings Preview
₽	
ន្រ	With co-versioned repo Databricks Repos allows you to clone a remote Git repo, which you can specify when you add a repo. Learn more
0	With individual notebooks
å	Although we recommended using co-versioned repo for Git integration, Databricks supports individual notebook version control integration with GitHub only).
£	Set your Git provider and credentials
<u>×=</u>	You can also set your Git provider credentials via API.Learn More
	Select your Git provider from the dropdown.
	Git provider
	Azure DevOps Services (personal access token)
	Change settings

Figure 3: GitHub Integration

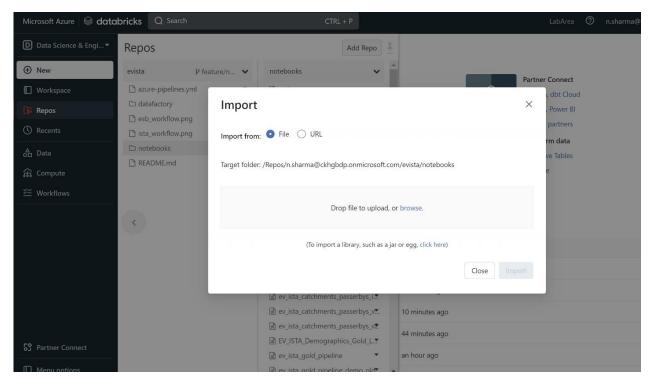


Figure 4: Import .ipynb in Databricks

Microsoft Azure 🗟 databri	icks Q Search	CTRL + P	LabArea 💿 n.sharma@ckhgbdp.onmicrosoft.c 🛩
D Data Science & Engi •	Repos	Add Repo	
• New	Add Repo		×
Repos	Location ①		
 Recents Data 	 /Repos/n.sharma@ckhgbdp.onmicrosoft.com Create repo by cloning a Git repository 		
🖧 Compute	Git repository URL ()	Git provider	
兰 Workflows	http://example.com/organization/project.git	Select a Git provider	~
	Advanced 🗸		
			Cancel Create Repo
		44 minutes ago	
52 Partner Connect		an hour ago	
Menu options			

Figure 5: Add GIT Repo into Databricks

Figure 6: Cli Installation

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

1 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.

1 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.

1 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

Storage Accounts dreistoredev	
✓	Attach to a resource
Slogs devstore ingest	Don't have permissions to access Azure subscriptions? You can connect to individual storage resources using various authentication methods.
> 🛋 File Shares	Tasks
> 🛄 Queues	Tables
> 📰 Tables	

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_ount—data.gov.uk* (2022)) 6. UK Postcode mapping.csv (*ukpostcode – NSPL* (2022)) 7. UK Qutputareas with population.geojson (*ukpostcode_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.

$\rightarrow \checkmark \uparrow$	devstore > ric_	model								
Name	^	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size	Status	Deletion ID	Remaining Day
💼 br_venue_er	nergy_utilization			14/12/2022 08:38		Folder		Active		
💼 br_venue_id				14/12/2022 08:32		Folder		Active		
💼 ev_footfall				14/12/2022 09:22		Folder		Active		
💼 ev_passerby	,			14/12/2022 09:23		Folder		Active		
💼 ev_venue_er	nergy_utilization			14/12/2022 09:21		Folder		Active		

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

File Edit View Run Help	Last edit was now Give feedback			DA	ŝ	Run all	 Connect ¥ 	Schedule	Share	~
file_sep	src_file_path abfss://devstore@evinsta.dfs.cc	src_table_name ric_model.ev_passerby							20	F #
A table of contents will be added here when a notebook has Markdown headings. %md # Heading 1	Cmd 1 %run /Shared/adls_credent: Command took 1.04 seconds b Cmd 2 %sql create database if not exi >	<pre>y n.sharma@ckhgbdp.onmicrosoft ists ric_model me.DataFrame y n.sharma@ckhgbdp.onmicrosoft vidget''' () table_name","ric_model.br file_path","abfss://devst ') .dfs.core.windows.net/Raw/</pre>	.com at 14/12/2022, 09:2 venue_id","src_tabl core@evinsta.dfs.core	2:26 on e_name .windo	unknown ") ws.net/	cluster Raw/Ista/r	ic/Postcode_le	vel_all_meters_u	electricity	

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.

Note: This notebook must	Command complete							
All vehicles veh0122.ods	Cad 37							
Low emission vehicles ve	xsql (sql)							
	CREATE TABLE IF NOT EXISTS ric_model.br_carownership_stat							
	(postcode_dist string,cars_count int,car_type string,year int,quarter string,country string,date string,load_datetime timestamp);							
	MERGE INTO ric_model.br_carownership_stat a							
	USING carownership							
	ON a.postcode_dist = carownership.postcode_dist and							
	a.car_type = carownership.car_type and							
	a.year = carownership.year and							
	a.quarter = carownership.quarter and							
	a.country = carownership.country							
	WHEN MATCHED THEN							
	UPDATE SET a.cars_count = carownership.cars_count,							
	a.load_datetime = carownership.load_datetime							
	WHEN NOT MATCHED THEN							
	INSERT (postcode_dist, car_type,year,quarter,country,cars_count,date,load_datetime) VALUES (carownership.postcode_dist,							
	carownership.car_type,carownership.year,carownership.quarter,carownership.country,carownership.cars_count,carownership.date,carowner							
	hip.load_datetime);							
	Command complete							
	Cnd 38							
	%sql							

Figure 13: ev car ownership excel ingestion

⁶https://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.

_outputareas_carowr Edit View Run Help									
able of contents will be	Command complete								
d here when a notebook Aarkdown headings.	Cnd 19								
d # Heading 1	*sql <								
	CREATE TABLE IF NOT EXISTS ric_model.br_outputareas_carownership_stats (outputareacode string,percentage_district_overlap double,postcode_dist string,cars_count int,car_type string,year int,quarter string,country string,cars_count_overlap double,load_datetime timestamp);								
	MERGE INTO ric_model.br_outputareas_carownership_stats a								
	USING merged_geog_intersection b								
	ON								
	a.outputareacode = b.outputareacode and								
	a.postcode_dist = b.postcode_dist and								
	a.car_type = b.car_type and								
	a.year = b.year and								
	a.quarter = b.quarter and								
	a.country = b.country								
	WHEN MATCHED THEN								
	UPDATE SET a.cars_count = b.cars_count,								
	a.percentage_district_overlap=b.percentage_district_overlap,								
	a.load_datetime = b.load_datetime,								
	a.cars_count_overlap =b.cars_count_overlap								
	WHEN NOT MATCHED THEN								
	INSERT (outputareacode,percentage_district_overlap,postcode_dist,								
	car_type,year,quarter,country,cars_count,cars_count_overlap,load_datetime) VALUES								
	(b.outputareacode,b.percentage_district_overlap,b.postcode_dist,								
	b.car_type,b.year,b.quarter,b.country,b.cars_count,b.cars_count_overlap,b.load_datetime);								

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.

File Edit View Run Help	p	ast edit was 9 hours ago Give feedbac	ick	9	8		Run all	● Connect ∨	🗄 Schedule	Share 👻
granularity										Z ¢ ∓
monthly										
Notebook to create datas	A	Cmd 1								
Engergy utilization Footfall		Notebook to create	e dataset to be used fo	r EV regr	essio	n m	nodels.			
PasserBy		Cmd 2								
Venue Overall agg	<u> </u>	%run /Shared/adls_credential	ls/configurations							
		Cmd 3 from pyspark.sql import func from pyspark.sql import Wind from pyspark.sql.types impor import numpy as np	dow						Python D	
		Command took 0.01 seconds by n	n.sharma@ckhgbdp.onmicrosoft.com at 14	/12/2022, 09:57:	58 on un	kaown	cluster			
		Cmd 4								
		dbutils.widgets.dropdown("gr granularity=dbutils.widgets. print('Running pipeline for		monthly'])						
		Running pipeline for : mont	thly							

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.

Combining carownership venue table	select	t a.*,b.total	table ric_mod motor,b.car_d	counts			· Anna · Santa an			
Adding Mobility Data	<pre>from ric_model.ev_model_dataset_monthly a left outer join ric_model.br_venue_trawhere b.year=2021 </pre>						Notebook detached cluster not in usable state			
	Image: Command Comm	eturned no resul d took 11.88 se t * from ric_ Spark Jobs	ts conds by n.sh model.br_ev_ma	arma@ckhgbdp.onn odel_dataset	ected_rows: long, num_ins microsoft.com at 14/12/ d: string, year: integer 74	2022, 15:32:11 on	DE_General			
	• ==									
		e 🗸 +								
		e v + venue_id	year 🍝	month 🔺	energy_utilization 🔺	total_footfall 🔺	age_band_18_24 🔺	age_band_25_34 🔺	age_band_35_4	
			year 2022	month A	energy_utilization A	total_footfall A	age_band_18_24	age_band_25_34	age_band_35_4 0.14377741606.	
	Table	venue_id d	,			-	-	-		
	Table 1	venue_id ABD18 3ST	2022	9	229.33	144988	0.10329130686677518	0.18471873534361463	0.14377741606.	
	Table 1 2	venue_id BD18 3ST BD18 3ST	2022 2022	9 12	229.33 148.49	144988 null	0.10329130686677518 null	0.18471873534361463 null	0.14377741606. null	
	Table 1 2 3	venue_id BD18 3ST BD18 3ST BD18 3ST	2022 2022 2022	9 12 8	229.33 148.49 291.28	144988 null 164001	0.10329130686677518 null 0.10604813385284236	0.18471873534361463 null 0.1867854464302047	0.14377741606. null 0.14424302290;	
	Table 1 2 3 4	venue_id BD18 3ST BD18 3ST BD18 3ST BD18 3ST	2022 2022 2022 2022 2022	9 12 8 10	229.33 148.49 291.28 197.14	144988 null 164001 148782	0.10329130686677518 null 0.10604813385284236 0.09980373969969486	0.18471873534361463 null 0.1867854464302047 0.1866153163689156	0.14377741606. null 0.14424302290; 0.14281969593	

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.

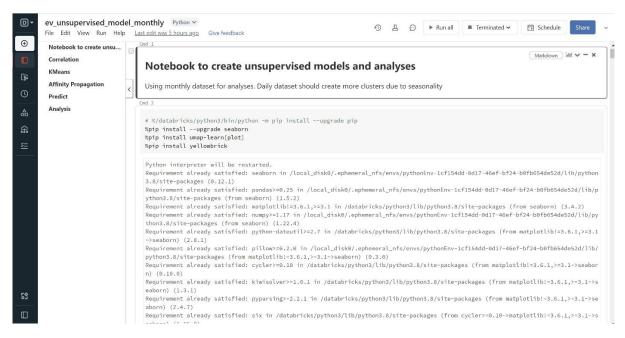


Figure 17: ev_unsupervised_model_monthly

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.

Linear Regression with all	
Predict	(Pythion) 🕨 🗸 🗸
Fredict	# %/databricks/python3/bin/python -m pip installupgrade pip
Linear Regression with to	%pip installupgrade seaborn
Regression model with tr	%pip install umap-learn[plot]
Kegression model with this	%pip install yellowbrick
	Python interpreter will be restarted.
	Requirement already satisfied: seaborn in /local_disk0/.ephemeral_nfs/envs/pythonEnv-3f4488e5-cf6e-4a18-8f54-937ead82fcf0/lib/pytho
	3.8/site-packages (0.12.1)
	Requirement already satisfied: pandas>=0.25 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-3f4488e5-cf6e-4a18-8f54-937ead82fcf0/lib,
	ython3.8/site-packages (from seaborn) (1.5.2)
	Requirement already satisfied: matplotlib:=3.6.1,>=3.1 in /databricks/python3/lib/python3.8/site-packages (from seaborn) (3.4.2)
	Requirement already satisfied: numpy>=1.17 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-3f4488e5-cf6e-4a18-8f54-937ead82fcf0/lib/
	thon3.8/site-packages (from seaborn) (1.22.4)
	Requirement already satisfied: python-dateutil>=2.7 in /databricks/python3/lib/python3.8/site-packages (from matplotlib!=3.6.1,>=3.
	->seaborn) (2.8.1)
	Requirement already satisfied: pillow>=6.2.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-3f4488e5-cf6e-4a18-8f54-937ead82fcf0/lib
	python3.8/site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (9.3.0)
	Requirement already satisfied: cycler>=0.10 in /databricks/python3/lib/python3.8/site-packages (from matplotlib!=3.6.1,>=3.1->seabc
	n) (0.10.0)
	Requirement already satisfied: kiwisolver>=1.0.1 in /databricks/python3/lib/python3.8/site-packages (from matplotlibl=3.6.1,>=3.1-> eaborn) (1.3.1)
	<pre>easo(n) (1:3:1) Requirement already satisfied: pyparsing>=2.2.1 in /databricks/python3/lib/python3.8/site-packages (from matplotlib!=3.6.1,>=3.1-></pre>
	aborn (2.4.7)
	Requirement already satisfied: six in /databricks/python3/lib/python3.8/site-packages (from cycler>=0.10->matplotlib!=3.6.1,>=3.1-3
	eaborn) (1.15.0)
	Requirement already satisfied: pytz>=2020.1 in /databricks/python3/lib/python3.8/site-packages (from pandas>=0.25->seaborn) (2020.5
	Command took 10,55 seconds by n.sharmaekkgbdp.ommicrosoft.com at 14/12/2022, 15:34:03 on DE_General
	Cod 2

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.

_

_

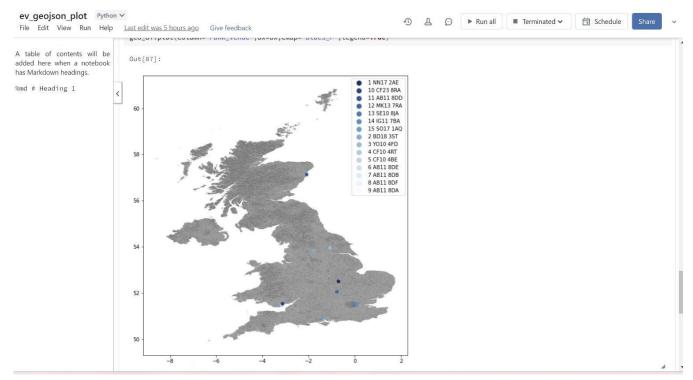


Figure 19: ev spatial model

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

Microsoft Azure 😂 databricks 🛛	Q Search CTRL + P	LabArea ⑦ n.sha
Workflows → Jobs → EV_MODEL_P EV_MODEL_PIPELINE		
Runs Tasks	EV_MODEL_PIPELINE_RIC A formy halmmed inkeptog reminism. & Sensed_int_uterm ev_dataset_creation - Advance.tutermate biologicg contents & Senset_uter_uter.tuter	Job details Job ID 216504091535 Creator @ n.sharma@ Run as ⑦ @ n.sharma@ Tags ⑦ + Tag Git
33	ev_traffic_car_ownership D- Aneros hatesets-dougledge ownerses # Sheerig ph_share ev_unsupervised_learning A Sheerig sh_stare ev_regr_ev_regr_everses # Sheerig sh_stare	Not configured Add Git settings Schedule None Add schedule + Compute Shared_job_cluster Driver: Standard_DS3_v2 · Workers: Sta

Figure 20: Databricks Worflow

6 Resources URL

Artefact:

https://studentncirlmy.sharepoint.com/:f:/g/personal/x21167818 student ncirl ie/EqSYdxrs 8rBDjj07278KkT8BSAsbnFseK5UclUafd4LTUA

References

*Traffic*count-data.gov.uk(2022), https://www.data.gov.uk/dataset/208c0e7b - 353f - 4e2d - 8b 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/evf* **ootfall.json?key = 9eb8e7e0**

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_postcode₃eojson(2022), https://geoportal.statistics.gov.uk/.[Accessed24-Nov - 2022].

ukpostcode-NSPL(2022), https://geoportal.statistics.gov.uk/datasets/national - statistics - Nov - 2022].