Assignment 1

Predictive Modelling of Eating-Out Problem

Data Science Technology and Systems

11523 Semester 2

Introduction

This report represents an end to end data science workflow using the Zomato restaurant dataset. The study covers EDA, predictive modelling for regression and classification and reproducibility with Git, Git LFS, and DVC. Key findings highlight cuisine diversity and the ability to predict restaurant rating with high accuracy.

Exploratory Data Analysis (part A)

Data overview

The raw dataset contained 10,500 rows and 17 columns.

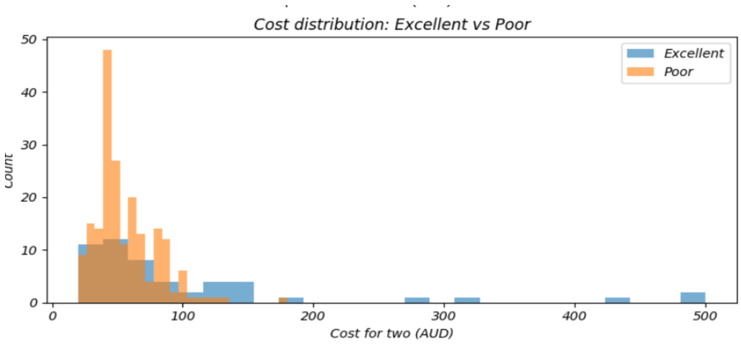
address	cost	cuisine	lat	link	lng	phone	rating_number	rating_text	subzone	title	type	vote
371A Pitt Street, CBD, Sydney	50.0	['Hot Pot', 'Korean BBQ', 'BBQ', 'Korean']	-33.876059	https://www.zomato.com/sydney/sydney-madang-cbd	151.207605	02 8318 0406	4.0	Very Good	CBD	Sydney Madang	['Casual Dining']	1311.
Shop 7A, 2 Huntley Street, Alexandria, Sydney	80.0	['Cafe', 'Coffee and Tea', 'Salad', 'Poké']	-33.910999	https://www.zomato.com/sydney/the-grounds-of-a...	151.193793	02 9699 2225	4.6	Excellent	The Grounds of Alexandria, Alexandria	The Grounds of Alexandria Cafe	['Café']	3236.
Level G, The Darling at the Star, 80 Pyrmont ...	120.0	['Japanese']	-33.867971	https://www.zomato.com/sydney/sokyo-pyrmont	151.195210	1800 700 700	4.9	Excellent	The Star, Pyrmont	Sokyo	['Fine Dining']	1227.
Sydney Opera House, Bennelong Point, Circular...	270.0	['Modern Australian']	-33.856784	https://www.zomato.com/sydney/bennelong-restau...	151.215297	02 9240 8000	4.9	Excellent	Circular Quay	Bennelong Restaurant	['Fine Dining', 'Bar']	278.

Summary statistic confirmed numeric ranges and identified missing values after cleaning (10,499, 17).

Cuisine Diversity

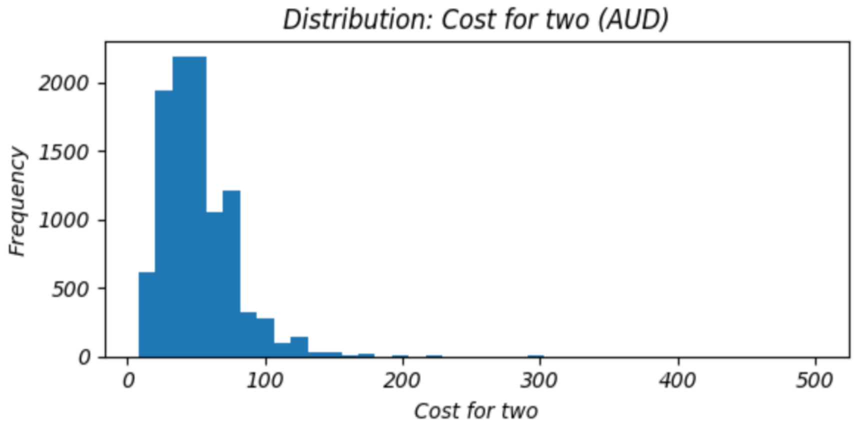
Unique cuisines served were 426, and the top 3 suburbs by counts:

```
subzone
CBD              476
Surry Hills      260
Parramatta       225
Name: count, dtype: int64
```

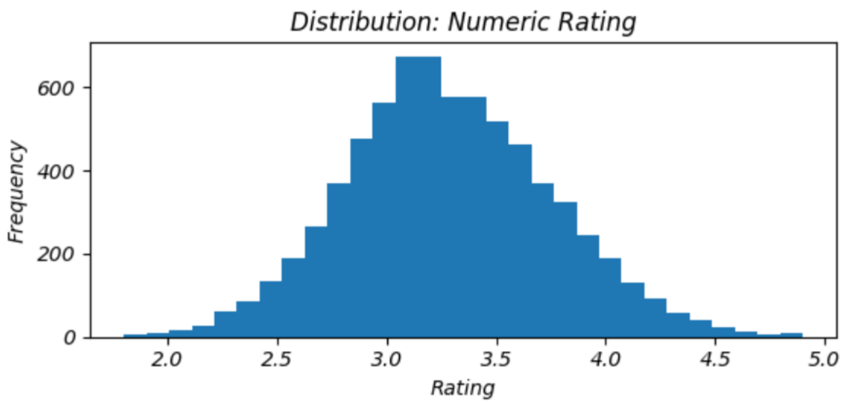


The histogram above shows the cost distribution for restaurants rated excellent vs poor. The median cost of excellent rated restaurants is AUD 60 and AUD 50 for poor rated restaurants. And the distribution shows that poor rated restaurants are more frequent in the lower cost range.

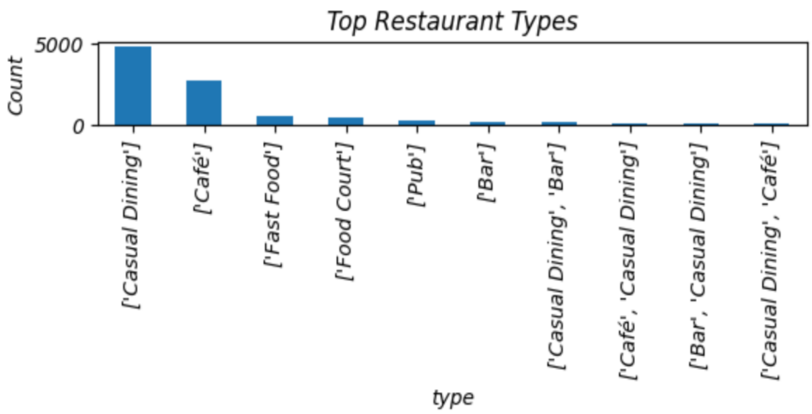
Distribution of Key Variables



Right-skewed distribution; most restaurant fall below 100 AUD.

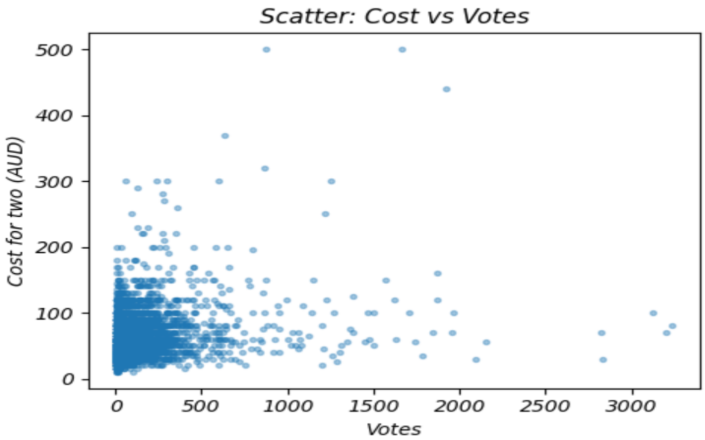


Bell-shaped distribution centered around 3.0 and 3.5.



Dominated by Casual Dining and Café with smaller counts for fast food, pubs and bars.

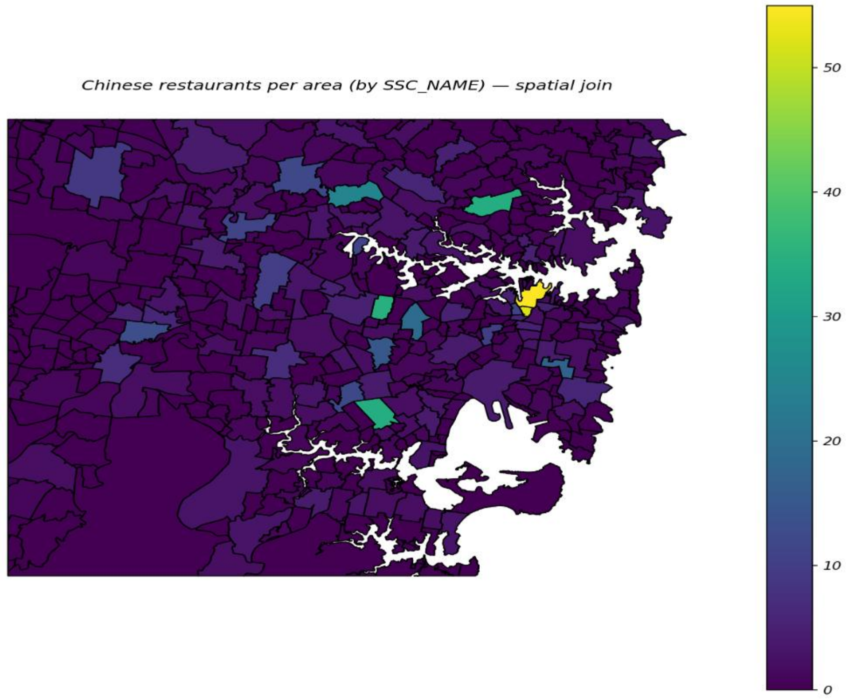
```
Correlation (cost ~ votes): 0.270
```



The scatterplot of cost vs votes shows a weak positive correlation (r:0.27) higher cost restaurants generally attract more votes.

Geospatial Analysis

Cuisine density varies by suburb, with central suburbs showing higher concentrations.

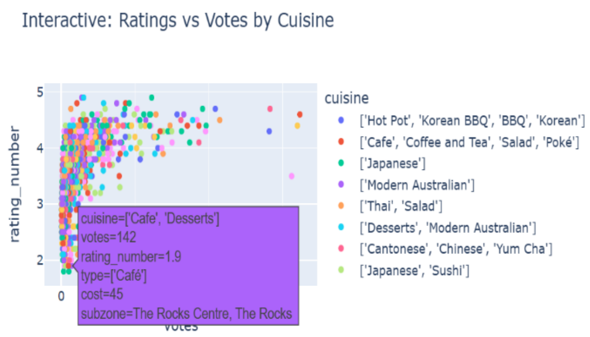


- Exact matches: 116 suburbs contained at least one Chinese restaurant.
- After fuzzy matching: 133 suburbs matched to GeoJSON regions.
- Hotspots are concentrated in the CBD and inner-west suburbs, shown in yellow/green shades.

Interactive Visualisation

To overcome the limit of static plots, I have built an interactive ploty scartterplot of rantings vs votes, coloured by cuisine.

This provides a richer insights than static charts, especially for identifying outliers and comparing cuisines.



Predictive Modelling (Part B)

Feature Engineering

Three new features were engineered to enhance the dataset. Cuisine_count captures how many cuisines a restaurant serves, reflecting diversity of offerings. Cost_bin groups restaurants into low, medium, and high-cost categories, making affordability easier to compare. Is_chain flags whether a restaurant is part of a chain based on repeated names, since chains often have consistent quality and pricing. These features add more structure and predictive value for the modelling stage.

Regression Models

Two regression approaches were applied to predit rating number:

- Linear Regression (Scikit-Learn) achieved an extremely low MSE of 0.000487, showing a very strong fit.
- Manual Gradient Descent Regression reached an MSE of 0.02864, higher dueto its iterative approximation.

[35]:

	Model	MSE
0	LinearRegression (sklearn)	0.000487
1	Manual Gradient Descent	0.028640

This demonstrates that Scikit-Learn's optimized solver provides superior accuracy compared to manual implementation.

Classification Models

The target rating_text was simplified into two classes:

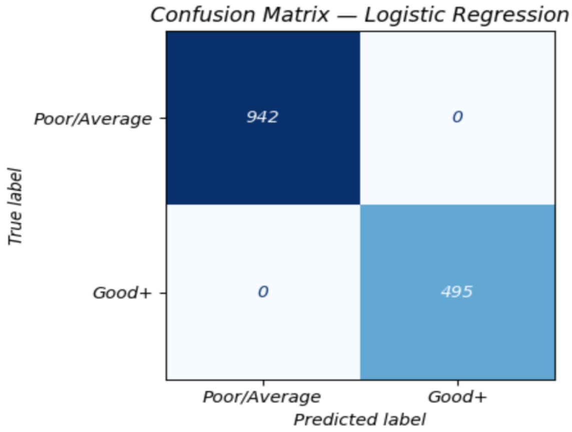
- Class 0: Poor + Average
- Class 1: Good + Very Good + Excellent

Four classifiers were trained and evaluated (80/20).

[31]:

	Model	Precision	Recall	F1
0	LogisticRegression	1.00000	1.00000	1.000000
1	SGDClassifier	1.00000	0.99798	0.998989
2	KNN	0.99596	0.99596	0.995960
3	DecisionTree	1.00000	1.00000	1.000000

Logistic Regression achieved perfect precision recall. And the confusion matrix confirms flawless classification with all 942 poor/average and 495 good/very good/excellent.



Workflow Management

To ensure reproducibility, Git, Git LFS and DVC were used:

- Git init – initialise repository.
- Git lfs install – handle large CSV dataset.
- Dvc init – initialise pipeline.
- Dvc add data/raw/Zomato_df_final_data.csv – track dataset
- Dvc repro - reproduce pipeline
- Dvc push – push artifacts to remote.

This setup allowed automatic tracking of dataset version, transformations and results.

PySpark vs Scikit-Learn

Scikit-Learn was easy to implement, efficient for small/medium datasets and produced consistent and accurate results.

PySpark was difficult to implement as when attempted for regression and classification, faced repeated worker crashes and gateway errors, for this dataset PySpark was overkill.

Scikit-Learn was the more practical choice, while PySpark is better suited for distributed computing on very large datasets.

Reference

Prodregosa, F. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. *vol*, *12*, 2825-2830.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., ... & Stoica, I. (2012). Resilient distributed datasets: A {Fault-Tolerant} abstraction for {In-Memory} cluster computing. In *9th USENIX symposium on networked systems design and implementation (NSDI 12)* (pp. 15-28).