ZURICH BUS TRANSPORT ANALYSIS

Data Science for Efficient Inner-City Bus Transport

- Goal: Operational Research (Route Clustering)

- Stakeholders: Transport authority,

- Role: Senior Data Scientist

For DATAPAO

By Saransh Bhargava

PROBLEM STATEMENT

Business Problem:
Analyze ZT Bus Data

Data Science
Objectives:

Operational Efficiency
Optimization

Performance-based
Route Segmentation

Service Planning

Success Metrics:

Clustering Quality
Metrics

Operational
Performance Metrics

DATA SOURCES

- Operational & Bus Movement Data
- DATAPAO

- Bonus – GitHub Project Link
- Please note: The GitHub project is a preliminary code, which shall be later converted into a project.

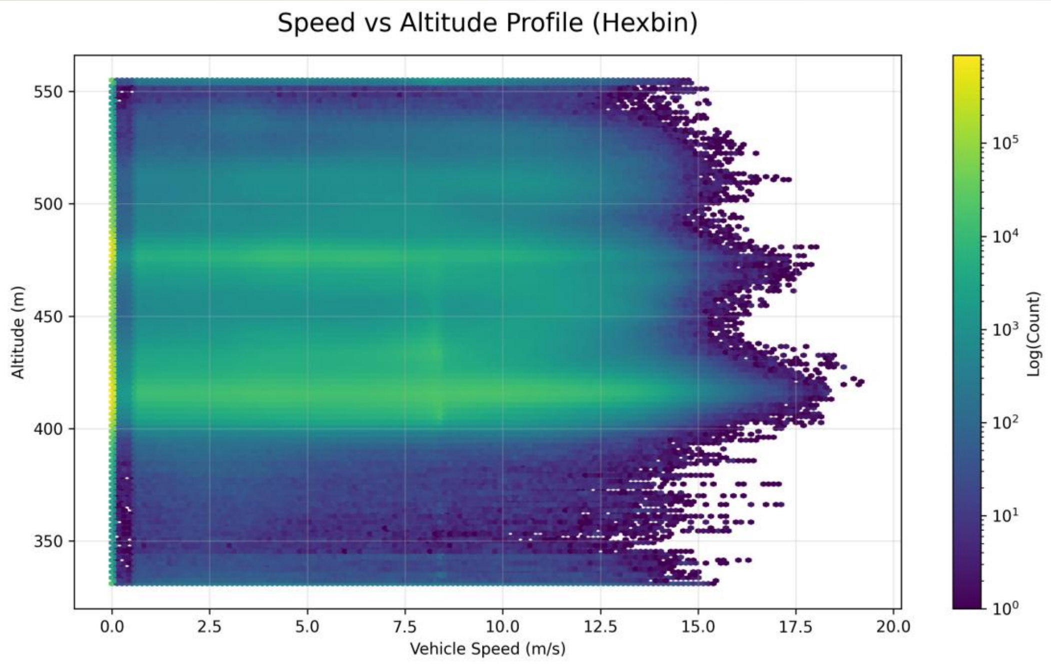
DATA CHALLENGES

- Missing data points in GNSS sensors (2.7%).
- Aggregated bus routes in metadata (unknown level of accuracy from actual bus routes).
- Sensor Data Inaccuracies (negative speed, missing stops, negative electric power demand, extreme altitude data points).

METHODOLOGY

- Handle all missing and erroneous data points.
- Understand levels of correlation and handle features.
- Generate behavioral features from final features.
- Route Classification through kMeans Clustering
- Silhouette and Inertia Scoring for optimal Clustering

RESULTS I



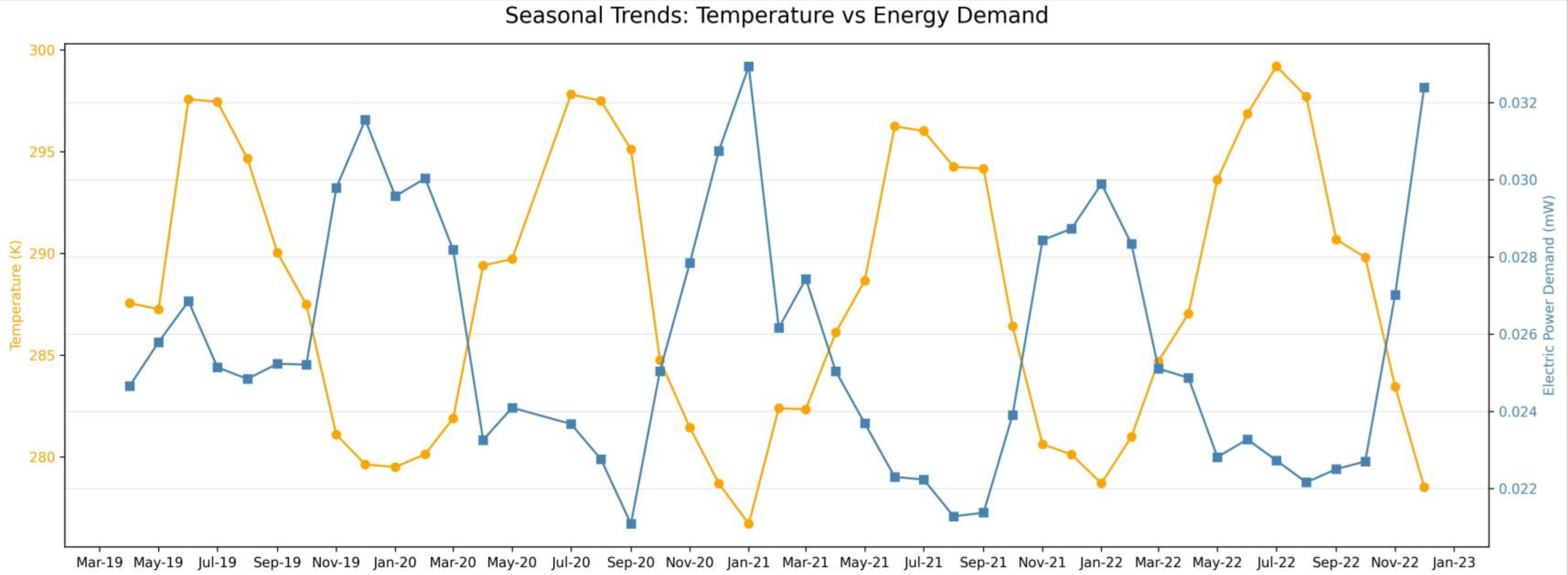
- Average bus speed and altitude speed to not have a close relationship.
- From information ahead, we shall see how clustered buses seem to behave.
- Higher altitude with a yellow-ish hue indicates high concentration of bus movements, i.e., 7.5 m/s ~ 27 km/hr.
- Buses travel slower speeds and that holds true in the clustering as well.

RESULTS II

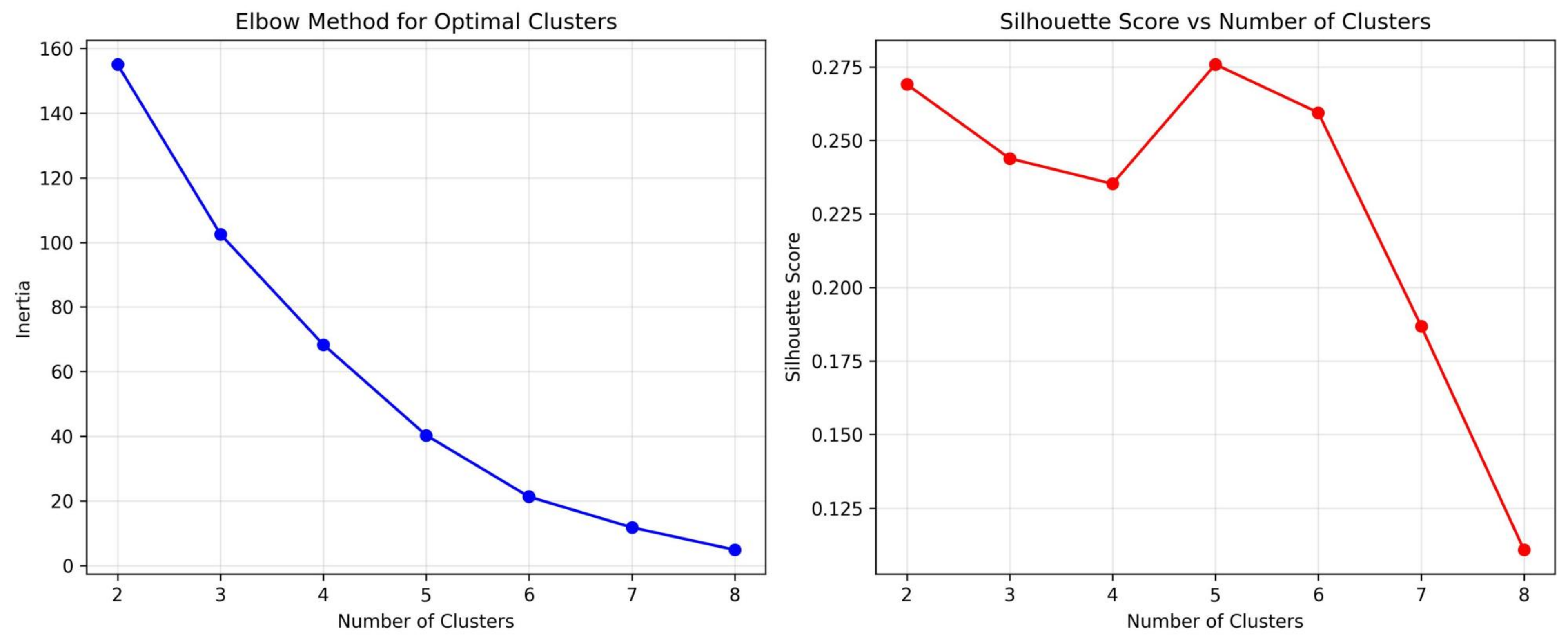
- The key results we receive are:
- Cluster 2: (bus routes N4, N2, N1) are potentially intercity routes with either fewer stops or lower frequency, whereas the other buses are regular multi-stop buses.
- Power demand is almost directly proportional to average speed, with exceptions.
- Average passenger load of intercity routes is low, yet power consumption and potentially maintenance costs are high.

RESULTS III

- As temperature seems to go lower, energy demand of buses goes up, which suggests heating costs rise.
- To reduce heating costs, bus routing and scheduling can be optimized.

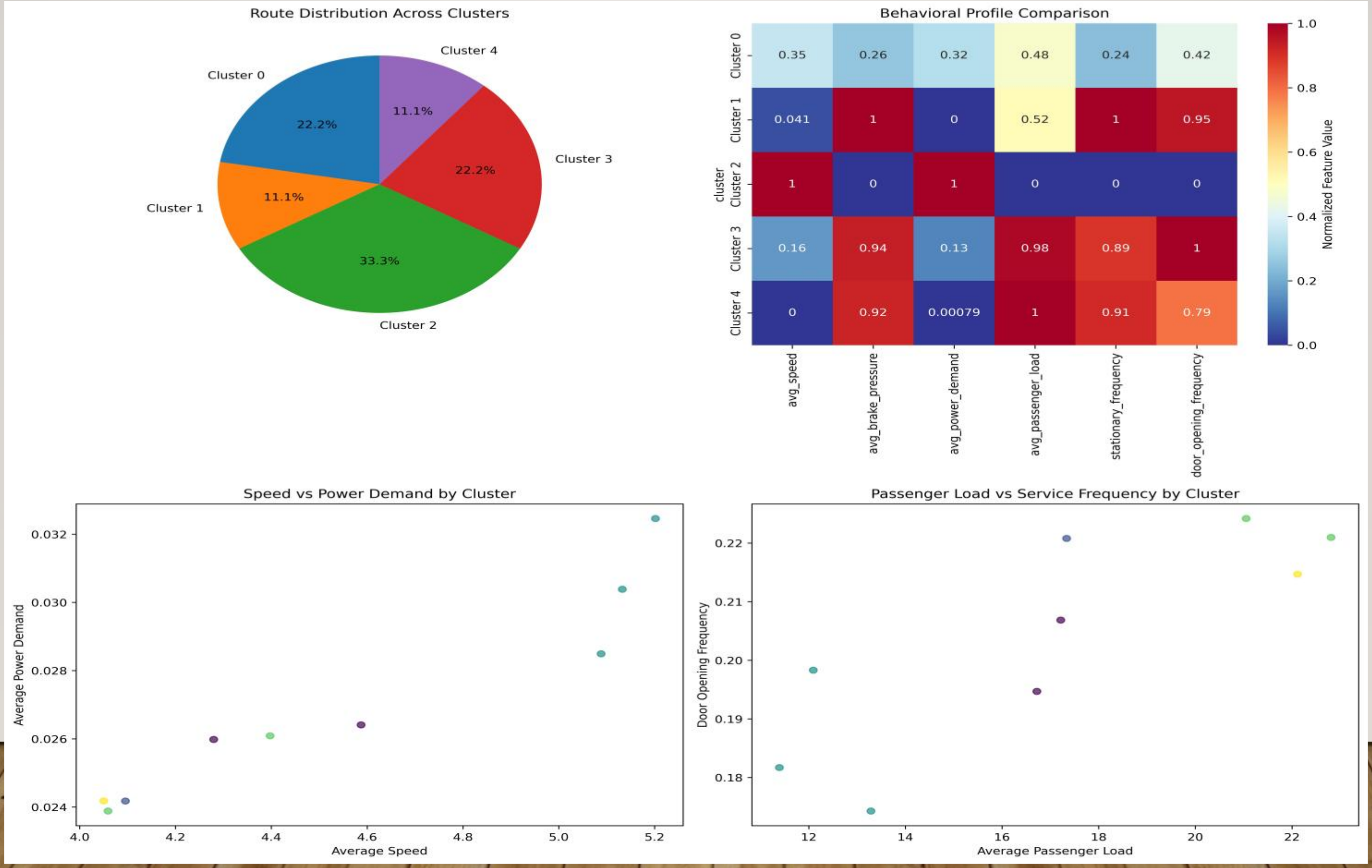


RESULTS IV


Elbow Method for Optimal Clusters
Inertia
0
20
40
60
80
100
120
140
160
2
3
4
5
6
7
8
Number of Clusters
Silhouette Score vs Number of Clusters
Silhouette Score
0.125
0.150
0.175
0.200
0.225
0.250
0.275
2
3
4
5
6
7
8
Number of Clusters

RESULTS V

Cluster	Routes	Reasoning
0	33, 46	Sub-Urban (low frequency)
1	83	Urban (external)
2	N4, N2, N1	Sub-Urban
3	72, 32	Urban (city-centre)
4	31	Semi-Urban



CHALLENGES & TRADE-OFFS

- Due to size of dataset, feature engineering and modelling is a massive undertaking.
- This can be circumvented by utilizing better data structures like Dask or Parquet filing systems.
- Additional information about current zonal districts would be useful to identify accurate bus usage statistics.

NEXT STEPS

- Convert the jupyter notebook-based project into a package-based project.
- Implement further effective structure in feature engineering.
- Perform further classification and try to gather fare information to predict appropriate fare system based on usage statistics.

CONCLUSION

- There are various other avenues of analysis to be explored within this dataset. Visit the GitHub link at a future date as there shall be updates to the project.
- Some examples are:
- Time Series Segmentation(peak vs off-peak), Urban vs Sub-Urban level Classification(N4/N2 vs 31/33), Operational & Scheduling Optimization (time period vs usage).