UNIVERSITY OF WESTMINSTER

INFORMATICS INSTITUTE OF TECHNOLOGY

6DATA007C – Final Project Report

Comparative Need State Analysis for Ride-Hailing Platforms: A Case Study on PickMe and Uber in Colombo, Sri Lanka

Student: Dinith Perera (w1838861 | 20200912)
Supervisor: Miss Abarnah Kirupananda
Co-Supervisor: Mr. Fouzal Hassan

This report is submitted in partial fulfillment of the requirements for the BSc (Hons) Data Science and Analytics at the University of Westminster

School of Computer Science & Engineering
University of Westminster

Date: 2025/05/15

Document Scope

The report conducts a comparative need state analysis of ride-hailing customers for PickMe and Uber users in Colombo, Sri Lanka. The scope includes data collection via survey, preprocessing the data, clustering the user segments using K-Modes clustering, testing and evaluation of the model, dashboard development, interpretation of the key findings, and providing actionable insights to PickMe and Uber. The project discusses the similar literature related to the project scope, ethical and sustainability concerns of the project, and the tools and technical methods implemented.

Declaration

This report has been prepared based on my work. Where other published and unpublished source materials have been used, these have been acknowledged in references.

Word Count: 10235
Student Name: Vidanalage Dinith Mario Perera
Date of Submission: 2025/05/15

Use of Generative AI

In this assessment, I used ChatGPT to brainstorm the analytics approach and Grammarly to refine the document with more formality. Everything else is derived from my own research, experience, and academic learning.

Abstract

This study analyzes Uber and PickMe users in Colombo, Sri Lanka, in a comparative need state analysis. The project segments users into groups driven by their needs, motivations, and preferences when using the ride-hailing services. The overall work of this study is twofold: (a) to make better business build recommendations for PickMe and Uber, using these insights to achieve better user engagement, loyalty, and service optimization while driving the use of actionable insights, and (b) to demonstrate the application of a powerful, but often underutilized, clustering technique, K-Modes, in a business context. It is well-suited to the data collected (fully categorical) and because K-Modes clustering solutions yield personas that are understandable, actionable, and relatable to Uber and PickMe user experiences.

Acknowledgements

I would like to express my appreciation to my supervisors, Miss Abarnah Kirupananda and Mr. Fouzul Hassan, for their valuable guidance and support throughout the completion of this project.

I would like to thank the University of Westminster and the Informatics Institute of Technology for creating the academic atmosphere, resources, and opportunities that made it possible for me to finalize this research.

I would also like to thank my parents for their constant support, understanding, and encouragement, which have largely helped me pursue my studies and the completion of this project successfully.

Table of Contents

List of Figures

List of Tables

1. Introduction

Chapter Overview

This chapter outlines the initial business problem and its criticality. It then discusses the blueprint of the solution implemented, the aims and objectives that guide it, the tools that were selected for the project, and the project scope.

1.1 Problem Statement

Ride-hailing giants in Sri Lanka, PickMe & Uber, have made a significant contribution to urban transportation with various service options that make life easier for the community. PickMe holds around 70% of the Sri Lankan ride-hailing market share, and Uber holds around 30% of the ride-hailing market share. They have become a necessity for daily commuting, travelling, and logistics. However, **with the expansion of the market, evolving customer expectations and needs, and competition, both companies face the challenge of understanding their customers sustaining their loyalty while providing optimized services, which impacts the revenue growth.** The growing ride-hailing market in Sri Lanka behaves significantly differently from the more mature markets. Therefore, competing not just on price but also on technological innovation, convenience, and user experience is also critical for the data-driven strategies.

Several studies emphasize the importance of understanding customer needs for improving service delivery and increasing user loyalty. Rafq and McNally (2022) highlight that recognizing alternative travel behaviours and preferences helps companies adjust their offerings to retain users. Similarly, Tirachini (2019) underscores the importance of analyzing travel behaviour to enhance ride-hailing services and encourage frequent use. A critical gap exists in the Sri Lankan ride-hailing domain in understanding user needs, the underlying motivations, preferences, and situational factors that affect their decision-making. While the companies collect customer data, they are mostly incomplete or unreliable and don’t explicitly extract the user's needs and motivations. By grouping customers by their needs and wants, the companies can approach them with personalized services, targeted marketing strategies, and new pricing models.

Business decision makers, being primary stakeholders of this project, will be able to use this project to identify potential opportunities in the emerging Sri Lanka market for service optimization & user retention. Ride-hailing customers will benefit from more personalized service offerings. Furthermore, the methodological approach, which uses the clustering technique

K-Mode with data preprocessing and interactive dashboarding, may serve as a blueprint for similar projects for academics and data analysts in the relevant industries.

1.2 Aims & Objectives

This project will conduct a comparative need state analysis of the PickMe and Uber users of Colombo, Sri Lanka. The project aims to **develop a data-driven unsupervised customer segmentation model that analyses the need states of PickMe & Uber customers in Colombo, Sri Lanka, creates user personas that show distinct customer needs, preferences, and behavioural patterns, which provide insights into marketing and operational, and other business strategies that increase user engagement, retention, and finally, revenue generation.**

The project accomplishes the following objectives;

1. Survey analysis to collect data from the user perspective.
2. Selection & the justification of the most appropriate clustering algorithm based on the nature of the collected dataset.
3. Development and implementation of the clustering model.
4. Evaluation of the model.
5. Design an interactive dashboard that visualizes the key insights that support decision-making.
6. Documentation of the project methodology, results, and providing business recommendations.

The final solution, including the clustering model and the dashboard, accomplishes the following objectives;

1. Customer segmentation into user personas provides actionable insights.
2. Design and develop an interactive dashboard that can be used fr decision making.
3. Provide actionable recommendations to PickMe and Uber for each cluster based on their features.

1.3 Project Scope

The project targets the collection and analysis of ride-hailing user data from PickMe and Uber customers in the Colombo district, Sri Lanka. It will concentrate on identifying customer

segments and presenting those insights through an interactive dashboard. The project won't extend to predictive modeling of customer behavior or real-time application deployment.

1.4 Research Gap and Novelty

Even though ride-hailing services in Sri Lanka are becoming popular, there is very little academic research on this market. In particular, there were no existing studies that have used need state analysis in the context of Sri Lankan ride-hailing services, nor any research focused on user behavior or segmentation specifically for PickMe, which is currently the country's leading ride-hailing platform was found. Additionally, while some consumer behavior and transportation studies have used clustering techniques, no studies have applied the K-Modes algorithm to categorize ride-hailing customers based on categorical behavioral data. Consequently, this gap in the existing research represents an opportunity not only for methodologically innovative work to be done but also for research with a direct relatable to practical consumer insights.

Chapter Summary

This chapter has provided the project's background, pinpointed the problem space, and laid out the aims and objectives that will steer the research and development work. The project initiates from a survey analysis collecting user perspective data on their needs, wants, and behaviour when using ride-hailing platforms, PickMe and Uber, so that they will be clustered into segments using an unsupervised machine learning algorithm, which will be helpful to the ride-hailing companies to create targeted strategies based on the characteristics of the user segments to increase the user engagement and by extension, revenue of the companies.

2. Background

Chapter Overview

This section examines the relevant literature on the need state analysis in the ride-hailing context and customer segmentation approaches that are valuable to the project. It also evaluates the tools and libraries used for the cluster analysis and the visualization. This section will act as a foundation for the justified methodological approach, decisions taken along the project, and the novelty of the project.

2.1 Literature Review

2.1.1 Ride-Hailing Needs State Analysis and Customer Segmentation

Targeted marketing starts from customer segmentation, which is used to understand user behaviour. Rafiq and McNally (2022) state that the emergence of technology-enabled (app-based) on-demand ride services expands the set of travel alternatives and substantially increases flexibility in activity scheduling and travel choices, thus affecting travel behavior in multiple ways.

Companies in their early stages can demographically and statistically segment their customers, but once the market starts to become dynamic, behavioural segmentation becomes necessary. It presents deeper information about situational and motivational factors for user decision making (Gomes and Meisen, 2023). Customer segmentation is also used for customizing product/ service offerings to diverse groups of users.

Need state analysis is an extension of customer segmentation that grasps the needs and wants of users for using a product or service. However, it remains suboptimal in the Sri Lankan ride-hailing market. Studies like Clewlow and Mishra, Gouri S (2017), and Sikder (2019) have analysed the demographics and the behaviours in ride-hailing but lack the depth of the need state segmentation.

Gonaldson and Sunitiyoso (2024) have used hybrid methods, including K-Means and NBClust, for a customer segmentation study in Indonesia that cohorts customers based on their ride-hailing preferences, showcasing the ability to create customer profiles using clustering. Juma James Masele and Shayo (2025) analysed the determinants of customer satisfaction in Tanzania without applying advanced clustering techniques of segmentation. Lee et al., (2021) and Shah and Kubota (2022) highlight factors such as convenience, cost, and quality that influence ride-hailing

adoption and satisfaction in developing countries, which can be easily embodied in need state frameworks. Rafiq and McNally (2022) and Young and Farber (2019) examined the United States' ride-hailing usage patterns, recognizing temporal and spatial factors that affect user choices. However, these studies also used descriptive methods rather than a segmentation-focused approach.

Piñón Rodríguez (2024) applied clustering techniques to public transportation, showing the potential of such segmentation in the mobility industry, but did not focus on ride-hailing or categorical clustering suitable for this project. He also talks about the importance of user personas, stating that they reveal user needs and motivations of customer segments, which transportation planners can use to improve ride-hailing infrastructure, develop targeted interventions, and improve the overall effectiveness of management strategies.

Study	**Method(s)**	**Dataset(s)**	**Key insights**
Clewlow and Mishra, Gouri S (2017)	Descriptive analysis	US ride-hailing users' survey data	Trends are identified, but no segmentation
Gomes and Meisen, (2023)	Literature Review	e-commerce	The importance of behavioural segmentation is recognized
Gonaldson and Sunitiyoso (2024)	Factor analysis, Cluster analysis, Regression analysis	Indonesian ride-hailing users' survey data	Segmentation using K-Means clustering. Personas were created using segmentation.
Juma James Masele and Shayo (2025)	Regression Analysis	Tanzanian ride-hailing users' survey data	Identified customer satisfaction factors
Lee et al., (2021)	Factor analysis	US & Chinese millennials' survey data	Need states influencing adoption
Shah and Kubota (2022)	Factor analysis	Lahore ride-hailing users' survey data	Service quality determinants
Young and Farber (2019)	Statistical analysis	A cross-sectional household travel survey conducted by Toronto Transit	Usage patterns analysis

		Services.	
Rafiq and McNally (2022)	Latent class analysis	2017 National Household Travel Survey - District of Columbia, USA	Need states for ride-hailing
Piñón Rodríguez (2024)	Hierarchical clustering and K-Means	HSL Travel Survey of 2018 - Helsinki region	Demonstrated the value of clustering in mobility studies. Behavioural variables were identified.

Table 1: Key literature in customer segmentation and need state analysis

2.2 Review of Methods and Applications

2.2.1 Customer Segmentation Techniques

Clustering algorithms are used to uncover hidden segments within large datasets. One such method is K-Means clustering, which is known for its simplicity and efficiency but is suitable for numerical data only (Gomes and Meisen, 2023). Another method is DBSCAN, which can identify irregular clusters but is limited in handling categorical data. Hierarchical clustering presents a visual hierarchy of clusters created through an iterative process, but struggles with scalability and is sensitive to outliers (Piñón Rodríguez, 2024; Cendana and Kuo, 2024).

K-Mode clustering is the most appropriate option for handling categorical data. By replacing the mean with modes for cluster centroids, K-Modes uses a simple matching dissimilarity measure, unlike K-Means clustering, which uses the Euclidean distance. It makes K-Mode ideal for datasets with categorical variables (Goyal, 2017; Sharma and Gaud, 2015). Saxena (2023) clustered travel perceptions during the COVID-19 pandemic, using K-Modes clustering. This proves that K-Mode is already a viable option for clustering ride-hailing users based on their needs and wants.

K-Prototype clustering is an extension of the K-Means and K-Modes algorithms built to handle mixed datasets with both numerical and categorical data. It has a distance function that mixes the Euclidean distance for numerical attributes with a dissimilarity measure for categorical ones. Studies like Dake et al. (2023) and Mohd et al. (2024) used K-Prototype clustering with this distance function. Delali Kwasi Dake, Gyimah, and Buabeng-Andoh (2023) modeled the behaviors of university students using mixed attribute profiles. Mohd, Lay Eng Teoh, and Hooi

Ling Khoo (2024) clustered shared-taxi ride requests using spatial and behavioral data. However, K-Prototype's hybrid design poses certain limitations when working with entirely categorical datasets, such as the one used in this study. When there are no numerical variables, the algorithm is just K-Modes and doesn't give any additional benefits. This further supports using K-Modes as the viable method for this project.

A comprehensive taxonomy of categorical data clustering methods is provided by Cendana and Kuo (2024), confirming the k-modes capability of building interpretable and meaningful clusters. V Kondur (2018) utilized k-mode to create user personas supporting its application in behavioural segmentation.

Method	**Advantages**	**Disadvantages**
K-Means	Widely used. Simple and efficient. Works well with large datasets	Requires numerical data and is not suitable for categorical variables
Hierarchical clustering	Does not require a predetermined number of clusters	Difficult to handle large datasets. Sensitive to outliers Not ideal for categorical data
DBSCAN	Can identify irregular segments. Handles outliers and noise well	Not ideal for categorical data. Not compatible with high-dimensional data
K-Protoype	Handles both numerical and categorical data in a single model.	Required feature weighting. Model faces complexities when only categorical data is present
K-Modes	Specifically designed for categorical data. Easy to implement and efficient with large datasets. Relatively fast. Robust to noise	Sensitive to initialization. Requires a predefined number of clusters. Unbalanced cluster size (wasn't a problem in this project)

Table 2: Comparison of clustering techniques for the analysis

2.2.2 Application in Ride-Hailing

Clustering techniques are not as widespread in ride-hailing as in other domains. This is largely because most studies have employed basic statistical or exploratory methods to derive their

findings. For example, the foundational research of Gonaldson and Sunitiyoso (2024) and Rafiq and McNally (2022) work at a largely superficial level, failing to penetrate the rich structure of the ride-hailing data that one can easily cluster. This project thereby fills a significant gap in the existing ride-hailing literature. Moreover, the undertaking includes the creation of an interactive dashboard, a feature commonly found in business intelligence but infrequently combined with clustering results in academic studies of ride-hailing services. By adding the dashboard to the project, I aim to increase the utility and real-world applicability of the analytical work.

2.3 Review of Tools

2.3.1 Data Analysis and Modeling Tools

The Data preprocessing and clustering model development use Python. Its strong libraries, such as Pandas, Numpy, and the kmodes package, give it the flexibility and efficiency to work with categorical data (Cendana and Kuo, 2024). Jupyter Notebook is used for coding due to its modular format.

Microsoft Excel is appropriate for the initial phases of data cleaning and exploratory analysis. Its nearly universal accessibility, along with its ease of use, make it a suitable tool for early-stage data manipulation. Excel lacks the advanced statistical modeling capabilities needed for this project.

2.3.2 Visualization and Dashboarding Tools

Power BI was chosen to develop the dashboard because it is easy to use, flexible, and works with a lot of different data sources. It lets us create professional, interactive dashboards that are great for letting stakeholders dig into the different customer segments and for letting them explore all the different behavioral patterns associated with those segments.

Although celebrated individually for their capabilities in data science and business intelligence, both Python and Power BI are often used together. This is because they complement one another in a way that balances flexibility, scalability, and, most importantly, usability for the business end user.

Chapter Summary

This chapter has provided a complete and thorough review of the literature relevant to the project. This body of work has two parts: the first part is customer segmentation; the second part is tool and method review, which directly leads into the project's choice of segmentation method and dashboard tool. The next chapter will outline the legal, ethical, sustainable, and other considerations.

3. Legal, Ethical, Sustainability, and Other Considerations

Chapter Overview

This chapter discusses the legal, ethical, social, professional, and sustainability concerns associated with the project. Given the nature of the project, which includes data collection via survey, following the relevant guidelines was a must to protect the integrity of the project.

3.1 Legal Considerations

To collect the data required for the project, a survey was designed and distributed among the ride-hailing users in Colombo, Sri Lanka. Starting from the design up to the data storage Personal Data Protection Act, No. 9 of 2022 of Sri Lanka was followed. The participants were well informed about the project before requesting assistance for the data collection, and the participants had the freedom and right to discontinue the survey at their convenience. Since the Data Protection Act in Sri Lanka is relatively new, the General Data Protection Regulation (GDPR) introduced by the European Union was also taken into account and adhered to.

The collected data didn't include any personally identifiable information (PII data). Not even an email address or a respondent's name was collected. All the data were anonymised. Furthermore, the survey didn't include any questions regarding sensitive information such as religion or health. Additional secure data storage steps were taken, adhering to the University of Westminster's data management policies, which align with the GDPR as well as the UK data protection standards.

3.2 Ethical Considerations

Before the data collection, an ethics form was filled out and sent to the University Research Ethics Committee at the University of Westminster (UoW) as part of the adherence to the ethical guidelines. During the data collection stage, the participants were provided with information about the project and how their data would be used. No sensitive information was asked from the participants, nor was their data collected. Furthermore, the literature used in the project is correctly cited to acknowledge their studies while allowing readers to verify the information.

3.4 Social and Professional Considerations

The project is conducted for the users in Colombo to identify their needs and wants and personalize the service offerings. During each stage of the project a special attention was given to avoid being biased and to extract accurate information. The project helps enhance urban mobility. Transparency was maintained during the project by detailed documentation of the project and the results. The results were verified thoroughly with clear acknowledgments of limitations and assumptions to maintain the accountability of the project. Furthermore, the findings were presented truthfully without any kind of manipulation.

3.5 Sustainability Considerations

The project embodies several sustainability practices. By identifying the needs and wants of users and customizing their services, the companies can indirectly support sustainable ride-hailing, potentially reducing vehicle emissions. A novel data science and analytics project was built, and it can be implemented in other industry sectors that are looking for sustainable technological and data initiatives. Additionally, the survey was conducted online, removing the need for paper-based surveys and physical travel.

Chapter Summary

This chapter has shown the steps taken to maintain the integrity, transparency, and accountability of the project by considering ethical, social, and legal aspects. The steps taken ensure that the project works within a framework of ethical regularities, securing participants' data and promoting technological innovations.

4. Methodology

Chapter Overview

This chapter gives an overview of the project cycle, the frameworks & algorithms integrated, and the nature of the collected data. The chapter also discusses the key tasks conducted, the results of the exploratory data analysis (EDA), and the analytical methods applied throughout the project.

4.1 Methodological Frameworks and Key Tasks

The project adopted the Cross-Industry Standard Process for Data Mining (CRISP—DM) framework, which was used to break down the project into phases. CRISP-DM is a widely recognized framework in data science projects as it includes both technical and business aspects. The project also followed agile principles, especially in the iterative model and dashboard building, enabling refinements to the project. These adoptions were responsible for the flexible response to stakeholder inputs and the quality of the project.

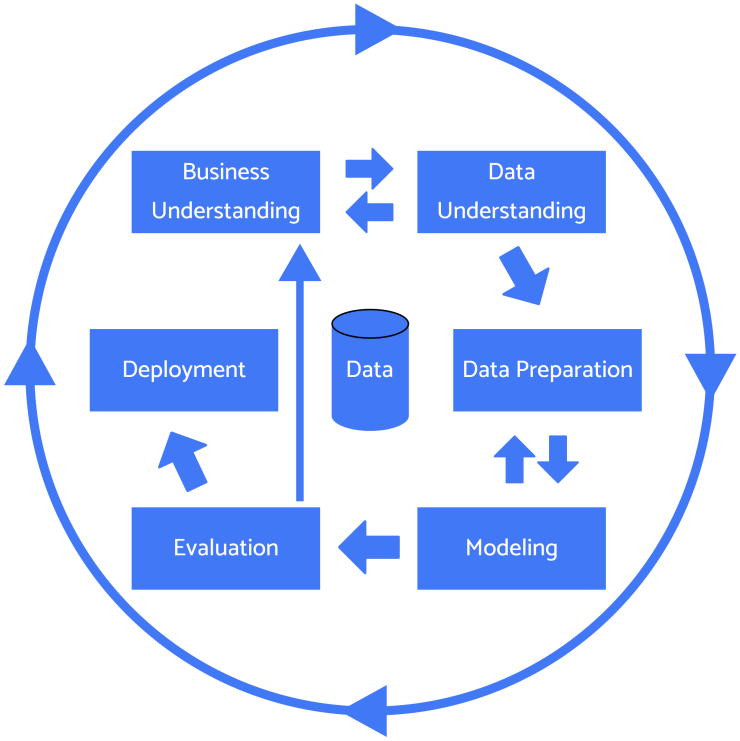

Figure 1: CRISP-DM framework

Key Tasks

1. **Problem Definition**
   To increase revenue by increasing engagement and customizing the service offerings, companies need to understand their customers. Therefore, the main aim is to build a clustering model and create user personas of ride-hailing customers.
2. **Data Collection & Sample Selection**
   A survey was designed and distributed in both English and Sinhala to collect user perspective data of the ride-hailing users in Colombo.
3. **Data Preprocessing**
   Microsoft Excel and Python were used to clean the data, preparing the data for clustering.
4. **Model Development**
   To handle the 100% categorical dataset, the K-Modes algorithm was introduced using the kmodes library in Python.
5. **Dashboard Development**
   Microsoft Power BI was used to build an interactive dashboard with multiple tabs to visualize the key insights found by clustering the customers.
6. **Evaluation and Validation**
   The results obtained by clustering and visualising were thoroughly evaluated for accuracy and alignment with the business requirements.
7. **Business Recommendations**
   Based on the clusters obtained and the key findings extracted from each cluster, recommendations were presented to each company separately and combined that can be executed within business strategies.

4.2 Data Collection

A survey was conducted to collect the required data for the project. This approach was taken because

1. The relevant data were not publicly available from the respective companies.
2. User perspective data provides in-depth insights into the project.

The data was collected in January 2025 by distributing two Google Forms, one in English and the other in Sinhala, due to the different language proficiency levels of the ride-hailing users in Colombo. The survey collected information on demographics, ride frequencies, purposes for using PickMe/ Uber or both, and user preferences. The research done by Shah and Kubota (2022) and V Kondur, N. (2018) was referred to when designing the survey.

Sample Size

The sample size was determined using **Cochran's formula:**

$$n_0 = \frac{Z^2 \cdot p \cdot (1-p)}{e^2}$$

Where:

- Z = 1.96 (95% confidence level)
- p = 0.5 (maximum variability assumption)
- e = 0.05 (5% margin error)

$$n_0 = \frac{(1.96)^2 \cdot 0.5 \cdot 0.5}{(0.05)^2} = 384.16 \approx 384$$

Cochran's formula is widely used in survey research and marketing studies when a large population is involved and needs to use a margin of error and calculate the sample size. A total of **535** responses were collected, and **384** responses from Colombo were filtered as the sample for the analysis.

Data Summary

The survey questions were replaced with column names according to a custom naming convention to easily identify the data.

	Survey Question	**Column Name**
1	What is your age group?	age
2	What is your gender?	gender
3	Please select the district where you spend most of your time.	district
4	What is your approximate monthly household income?	monthly_income
5	Which ride-hailing app(s) do you use? (Your answer will direct you to relevant questions.)	ridehailing_app
6	Why do you choose PickMe as your preferred ride-hailing app?	p_pickmepreference

7	For what purposes do you most often use PickMe's ride-hailing services?	p_pickmepurpose
8	How often do you use PickMe rides?	p_pickmefrequency
9	Which type of vehicle(s) do you use most often?	p_pickmevehicle
10	When do you most frequently use ride-hailing services?	p_pickmetime
11	What's your preferred payment method?	p_pickmepayment
12	Are you a “PickMe Pass" subscriber?	p_pickmepass
13	Why do you choose Uber as your preferred ride-hailing app?	u_uberpreference
14	For what purposes do you most often use Uber's ride-hailing services?	u_uberpurpose
15	How often do you use Uber rides?	u_uberfrequency
16	Which type of vehicle(s) do you use most often?	u_ubervehicle
17	When do you most frequently use ride-hailing services?	u_ubertime
18	What's your preferred payment method	u_uberpayment
19	Are you an “Uber One" subscriber?	u_uberone
20	Why do you use both PickMe and Uber?	b_preference
21	How often do you switch between apps for better deals or availability?	b_switchingfrequency
22	For what purposes do you most often use ride-hailing services?	b_purpose
23	How often do you use ride-hailing services?	b_frequency
24	When do you most frequently use ride-hailing services?	b_time
25	Which type of PickMe vehicle(s) do you use most often?	b_pickmevehicle
26	Which type of Uber vehicle(s) do you use most often?	b_ubervehicle
27	What's your preferred payment method	b_paymentmethod
28	Are you a “PickMe Pass" subscriber?	b_pickmepass
29	Are you an “Uber One" subscriber?	b_uberone

Table 3: Survey questions and their replaced column names

Exploratory Data Analysis

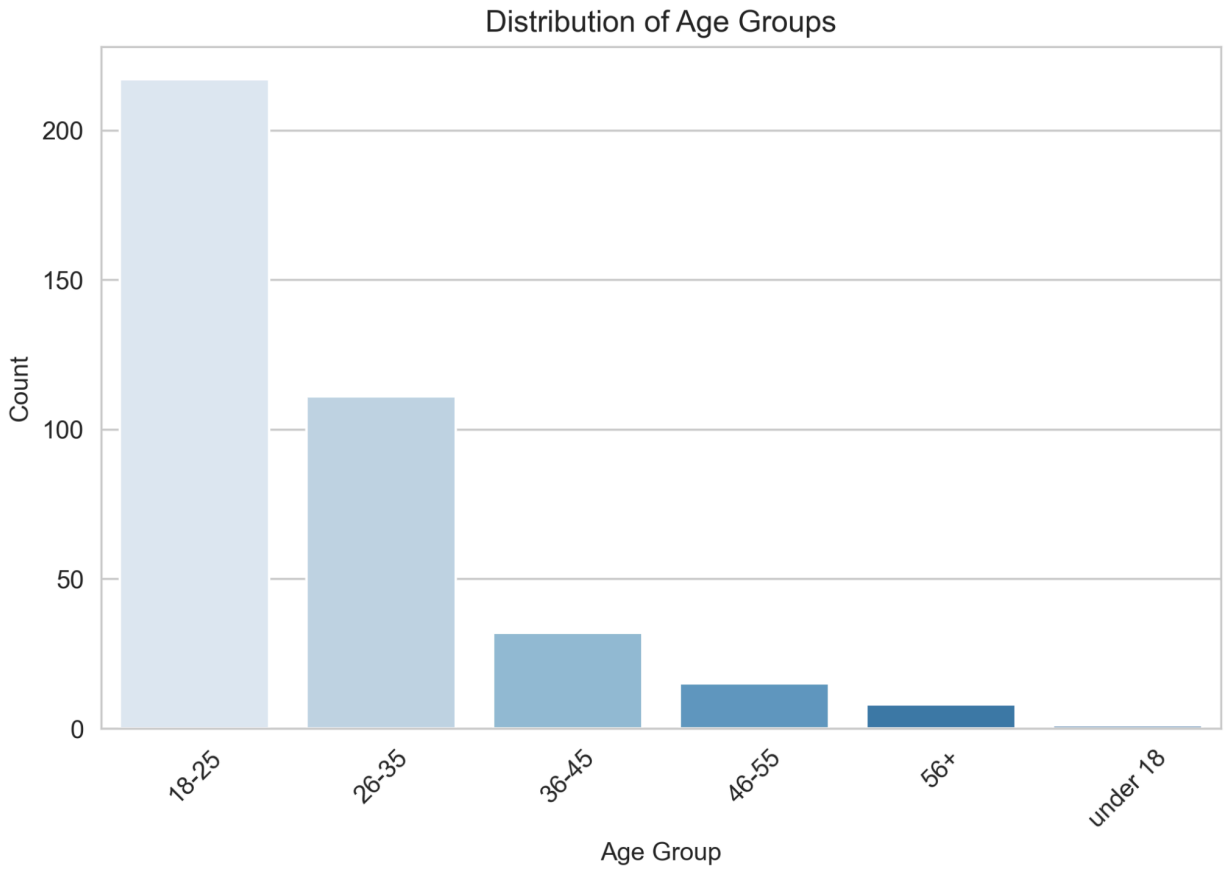

Figure 2: Distribution of age groups

As per Figure 2, more than half of the population falls under the age group 18-25 age group. The second highest group is age 26-35, making the customer base primarily Gen-Z and Millennials. A low turnout of people above age 45 suggests that the young generation is still economically stabilizing, and most of them are without private vehicles. The situation is the opposite for people who are well above the age of 45. Another interpretation is that the tech-savviness of the young generation is more than the older generations. We can even see a small percentage of underage customers who use the ride-hailing apps, further confirming this. A similar pattern of age distribution is observed in the studies of Juma James Masele and Shayo (2025), Clewlow and Mishra, Gouri S (2017), and Shah and Kubota (2022). These studies provide a similar interpretation for this kind of age distribution.

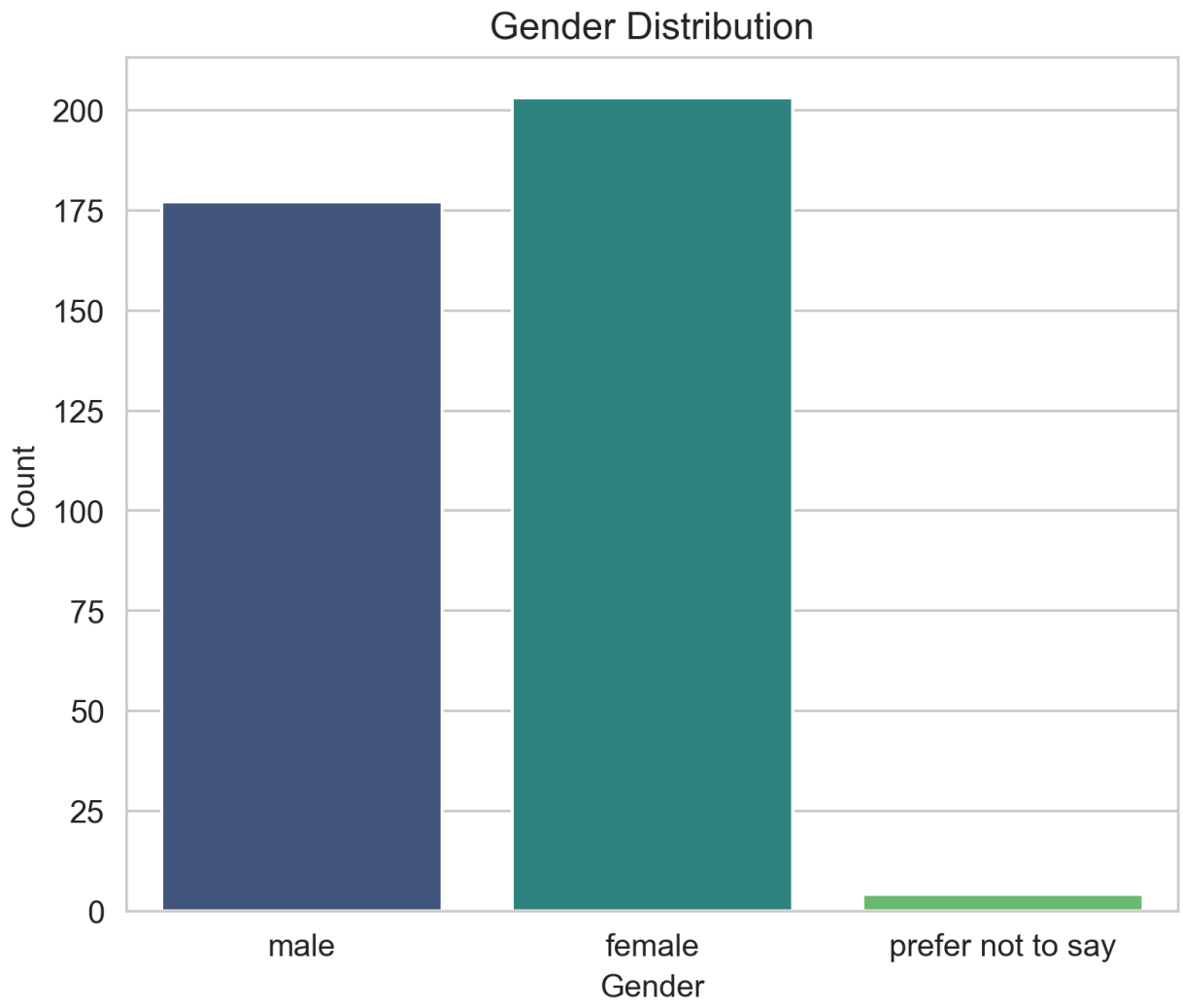

Figure 3: Gender distribution

A gender balanced distribution can be observed among the respondents, with a slightly higher proportion of female riders. This suggests that ride-hailing apps are adopted across all genders, indicating trust in the services provided. Around 11% of the customers have not revealed their gender.

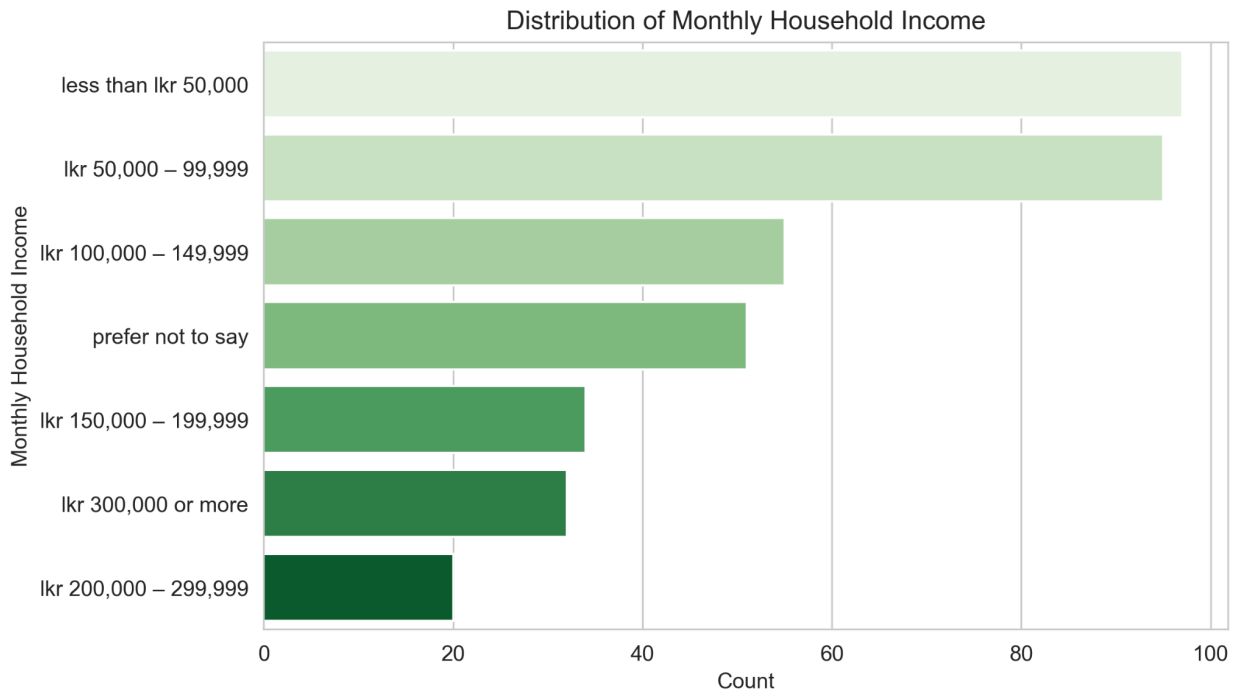

Figure 4: Distribution of monthly household income

As per Figure 4, the highest number of customers is from low household income levels of below Rs.100,000. Most of them being young and still becoming economically stable could be a reason for such a higher number.

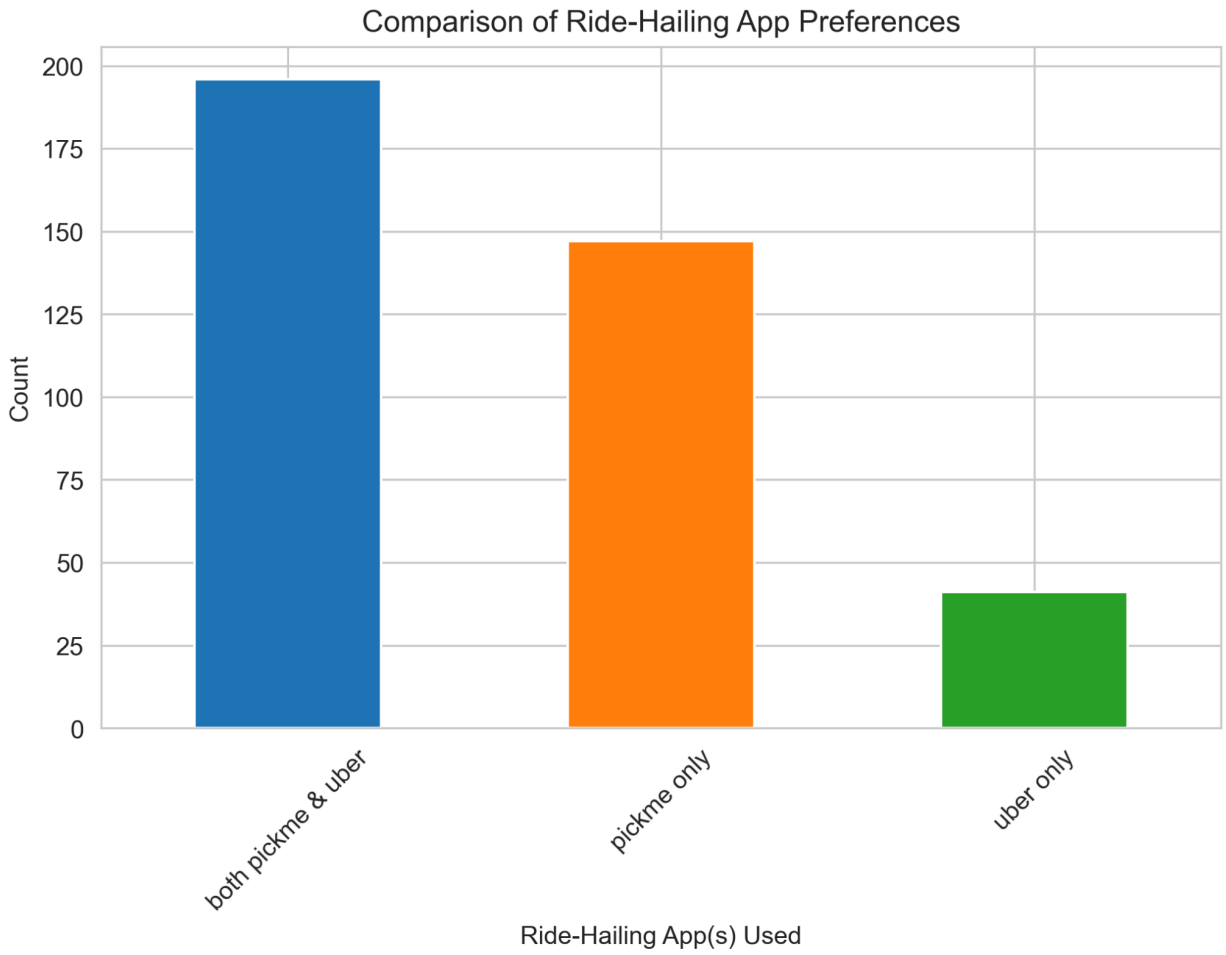

Figure 5: Comparison of ride-hailing app preferences

According to Figure 5, around 50% of respondents use both PickMe and Uber, while around 40% use only PickMe and 11% use only Uber. High numbers of dual app users suggest that both apps provide a competitive level of customer service, making it hard for customers to be loyal to a single app. Since PickMe is the leading ride-hailing service in the country with a much larger driver fleet, more people use it because of the better driver availability with shorter waiting times.

4.3 Methods

Data Preprocessing

Initial data preprocessing included the Translation of Sinhala survey data into English. The process was initially executed within the Python environment and failed. Therefore, a manual translation took place by manually entering the Sinhala data into the English Google Form, which automatically updated the respective Google Sheet. Next, the Colombo district data were

filtered. Since the dataset is categorical, data validation steps were conducted. Data was standardized by converting it to lowercase and removing extra spaces. This method ensured that the data didn't have redundant values due to typographical errors. Encoding methods were discarded as K-Modes doesn't require them. The survey question "When do you most frequently use ride-hailing services? " (b_time) in the Sinhala survey was mistakenly omitted when publishing the survey. Therefore, that column was empty. To fill the b_time column, a separate dataframe was created with the relevant data, and Imputation methods such as Exact Match Imputation and Partial Match Imputation were used by referring to the values of the b_time column in the English survey. Finally, all the data was merged into a single dataset and was prepared for clustering.

K-Modes Clustering Algorithm

K-Modes algorithm sorts the data into a user-defined number of K clusters, where each cluster is represented by a centroid, which is the mode of the categorical data. K-Modes algorithm tries to minimize the cost function, which is the sum of the dissimilarity between each data point and the centroid of its assigned cluster (Saxena 2023).

Dashboard Development

An interactive dashboard was designed to present the key findings and business insights using Microsoft Power BI. This approach allows business stakeholders to understand the clustered data and take business decisions based on the insights. The dashboard was designed following agile principles to make sure it meets the analytical and business objectives of the project. The dashboard includes core metrics, cluster information, and comparisons along with dynamic filtering options, making it more user-friendly and insightful.

Chapter Summary

This chapter has presented how the project was executed according to a framework, following its principles and applying methods to obtain the best results, aligning with professional standards. The project implements various methods to tackle each stage of the project and uses multiple tools and algorithms for that. The following chapter will elaborate on the Tools used and skills acquired.

5. Tools and Skills

Chapter Overview

This chapter discusses the tools and environments that were used for the project. Each of these tools was selected based on its effectiveness in fulfilling certain tasks and supporting project objectives.

5.1 Python Environment

Python, a programming language that is widely used in data science and machine learning, was used as the primary work environment of the project. Python was used for important tasks such as data preprocessing, model building, and clustering users into meaningful segments. Python's user friendliness and vast range of library ecosystem make it the best choice for the project. Jupyter Notebook was used for code writing and inline visualization.

Inside Python, the following libraries were utilized;
pandas and numpy - These libraries were used for data manipulation and pre-processing tasks. They handled the survey dataset-related computations and analyses.
matplotlib - This library was used to design Python visualizations depicting data distributions and exploratory data analysis (EDA) related charts.
Kmodes - This is the specialized library that can handle categorical data in clustering. It perfectly aligns with the project requirements and the nature of the dataset.

5.2 Google Sheets and Google Forms

Google Forms was used to design the conditional survey and to gather data. Google Sheets was utilized as a data entry tool for the collected data from the Google Forms. Google Sheets was used for tasks like initial data preprocessing before importing to Python and then to Power BI. Because both of these products can be synced so that Sheets get updated in real-time with the response, both of them were selected as the best choice for data collection.

5.3 Microsoft Power BI

Microsoft Power BI was used to present the findings as actionable insights in an understandable way to non-technical stakeholders. Once the clusters were formed, the cleaned and updated dataset was imported into Power BI to seamlessly. Power BI is highly user-friendly and interactive. Power BI acts as a decision support tool. Therefore, it was chosen as the ideal tool for presentation purposes.

5.4 Skills Development

The project required a new set of skills to be learned to complete the project. Categorical clustering is the major new skill that was developed during the project. It included understanding the K-Modes algorithm, how to evaluate clusters, and how to evaluate the model. Skills like exact match imputation and partial match imputation were learnt as part of handling missing values during the data pre-processing stage. For data collection, conditional survey design skills were needed as data should be gathered efficiently and responsibly. Ethical skills were developed as well. Skills such as creating measures, calculated tables, and DAX were practiced and developed during the dashboard development phase.

Chapter Summary

The combination of these tools, languages, and libraries supports the completion of a successful project. Google Forms was used to collect data, and Google Sheets was used to store data. The Python environment helped preprocess the data and build the model to cluster the data. Finally, the findings were presented in Power BI, supporting stakeholders to make business decisions. The perks of each tool made them ideal options for the relevant tasks.

6. Model Development

Chapter Overview

This chapter presents a detailed explanation of the model-building process, including the preparations done up to the introduction of the model, implementation of the model, testing and evaluations conducted, and the key results. This chapter also discusses the technical and analytical aspects of the model-building process, highlighting the relevance to the business objectives, originality, practicality, and complexity of the codes. The objective of model building is to segment the ride-hailing customers in Colombo into need state clusters using categorical data analysis, and then create user personas based on the cluster features to finally provide business insights and recommendations.

6.1 Model Development Process

6.1.1 Data Preparation code

The data preparation code is an essential part of the model development process as it directly affects the determination of the number of clusters, the quality of the clusters, and the overall accuracy of the model. Several steps were executed in code to prepare the data.

At first, 2 datasets were loaded into the Python environment, main_df and dual_platform_df. main_df contains all responses from both the English and translated Sinhala surveys, excluding Sinhala entries where respondents selected "Both PickMe and Uber" and skipped the "b_time" question (due to its absence in that version of the survey). Dual_platform_df contains the translated responses for Sinhala participants who selected "Both PickMe and Uber" but originally missed the "b_time" question used for targeted missing value imputation.

Initial Cleaning

```
# Remove duplicate rows, if any
main_df.drop_duplicates(inplace=True)

# Drop rows where all elements are missing
main_df.dropna(how='all', inplace=True)

# Filter rows for Colombo district
```

```
colombo_df = main_df[main_df['District'] == 'Colombo']

# Convert all column names to lowercase
colombo_df.columns = colombo_df.columns.str.lower()

# Standardize column names (remove leading/trailing spaces and replace multiple spaces with a single space)
colombo_df.columns = colombo_df.columns.str.strip().str.replace(r'\s+', ' ', regex=True)

# Standardize string values in all object-type columns
colombo_df = colombo_df.apply(lambda x: x.str.strip() if x.dtype == "object" else x)

# Convert all object-type column values to lowercase
colombo_df = colombo_df.apply(lambda x: x.str.lower() if x.dtype == "object" else x)
```

Explanation:

The above code is responsible for removing the duplicates, blank rows, and filtering only the Colombo district data, which represents the scope of the project. Since the dataset comprises categorical data, the column names and the values were all standardized by converting every letter to lowercase, removing extra spaces, and converting the datatype of all the string values into objects. This way, the model won't make any mistakes when reading the values. Next, any response in the ridehailing_app column other than "PickMe only", "Uber only", or "Both PickMe & Uber" was discarded because they are out of the scope of this project. These steps were repeated for both datasets.

Imputation of missing b_time values

As mentioned in 4.3 Methods, the related question to the b_time column was accidentally omitted in the Sinhala Survey. It affected 31 missing values among users who use "Both PickMe & Uber." Therefore, 3 stages of imputation methods were applied to fill these values as they are valuable when clustering.

Exact Match Imputation

```
# Exact Matching
exact_matches = []
exact_filled = 0
for idx, row in colombo_df2[colombo_df2["b_time"].isna()].iterrows():
    matches = filtered_both_df[
        (filtered_both_df["b_frequency"] == row["b_frequency"]) &
        (filtered_both_df["b_purpose"] == row["b_purpose"])
    ].copy()
    if not matches.empty:
        b_time_mode = matches["b_time"].mode()
```

```
        if not b_time_mode.empty:
            selected_time = np.random.choice(b_time_mode) if len(b_time_mode) > 1 else b_time_mode[0]
            exact_matches.append((idx, selected_time))
            exact_filled += 1

print(f"Exact matching filled {exact_filled} rows.")
```

The first method used was Exact Match Imputation. The approach fills exact b_time values by recognizing users with identical ride frequency (b_frequency) and ride purpose (b_purpose). When such users are found, the method will impute the mode value among the matched users. If multiple mode values exist, the method will select a random mode value to avoid bias. 14 missing values were filled by the exact match imputation method.

Partial Match Imputation

```
# Partial Matching
partial_matches = []
partial_filled = 0
for idx, row in colombo_df2[colombo_df2["b_time"].isna()].iterrows():
    if idx not in [x[0] for x in exact_matches]:
        matches = filtered_both_df[
            (filtered_both_df["b_frequency"] == row["b_frequency"]) &
            filtered_both_df["b_purpose_list"].apply(
                lambda x: sum(p in row["b_purpose_list"] for p in x) >= 2 if x and row["b_purpose_list"] else
False
            )
        ].copy()
        if not matches.empty:
            b_time_mode = matches["b_time"].mode()
            if not b_time_mode.empty:
                selected_time = np.random.choice(b_time_mode) if len(b_time_mode) > 1 else b_time_mode[0]
                partial_matches.append((idx, selected_time))
                partial_filled += 1

print(f"Partial matching filled {partial_filled} rows.")
```

The values with no exact match were filled by partial matches by identifying if the respondent shared at least two purposes with others who had the same b_frequency, and the mode values from the matching rows were imputed. Another 14 rows were filled using partial match imputation.

Mode Fallback

```
# Mode Fallback
mode_fallback = []
mode_filled = 0
overall_mode = filtered_both_df["b_time"].mode()[0] if not filtered_both_df["b_time"].mode().empty else
"don't have a specific time"
for idx, row in colombo_df2[colombo_df2["b_time"].isna()].iterrows():
    if idx not in [x[0] for x in exact_matches + partial_matches]:
        mode_fallback.append((idx, overall_mode))
        mode_filled += 1

print(f"Mode fallback filled {mode_filled} rows.")
```

For the remaining three values in the b_time column, the overall mode value was imputed.

6.1.2 Implementation of K-Modes

```
# Import the nessasary libraries for clustering
!pip install kmodes
from kmodes.kmodes import KModes
```

First, the kmodes package was installed and imported.

```
# Replace nulls with "Not Applicable" due to conditional survey design
df_for_clustering = preprocessed_data.fillna("Not Applicable")
print(df_for_clustering.isnull().sum())  # Should show 0 for all columns
```

Next, the remaining null values, due to the nature of the survey dataset, were replaced with “Not Applicable”. K-Modes doesn’t handle null values effectively. That’s the reason this step was taken.

```
# Make sure all columns are treated as categorical (convert to string)
df_for_clustering = df_for_clustering.astype(str)
```

Since K-Mode requires categorical data, the data type of each column was converted to string values.

Determining the Optimal Number of Clusters

Three techniques were followed when determining the optimal number of clusters for this project.

1. **Elbow - Method (Cost Function)**
2. **Dunn Index**
3. **Categorical Silhouette Score**

1. Elbow Method

```
# Determine the optimal number of clusters using the cost function
costs = []
K_range = range(2, 10)  # Testing clusters from 2 to 10

for K in K_range:
    km = KModes(n_clusters=K, init='Huang', n_init=5, verbose=0)
    km.fit(df_for_clustering)
    costs.append(km.cost_)
```

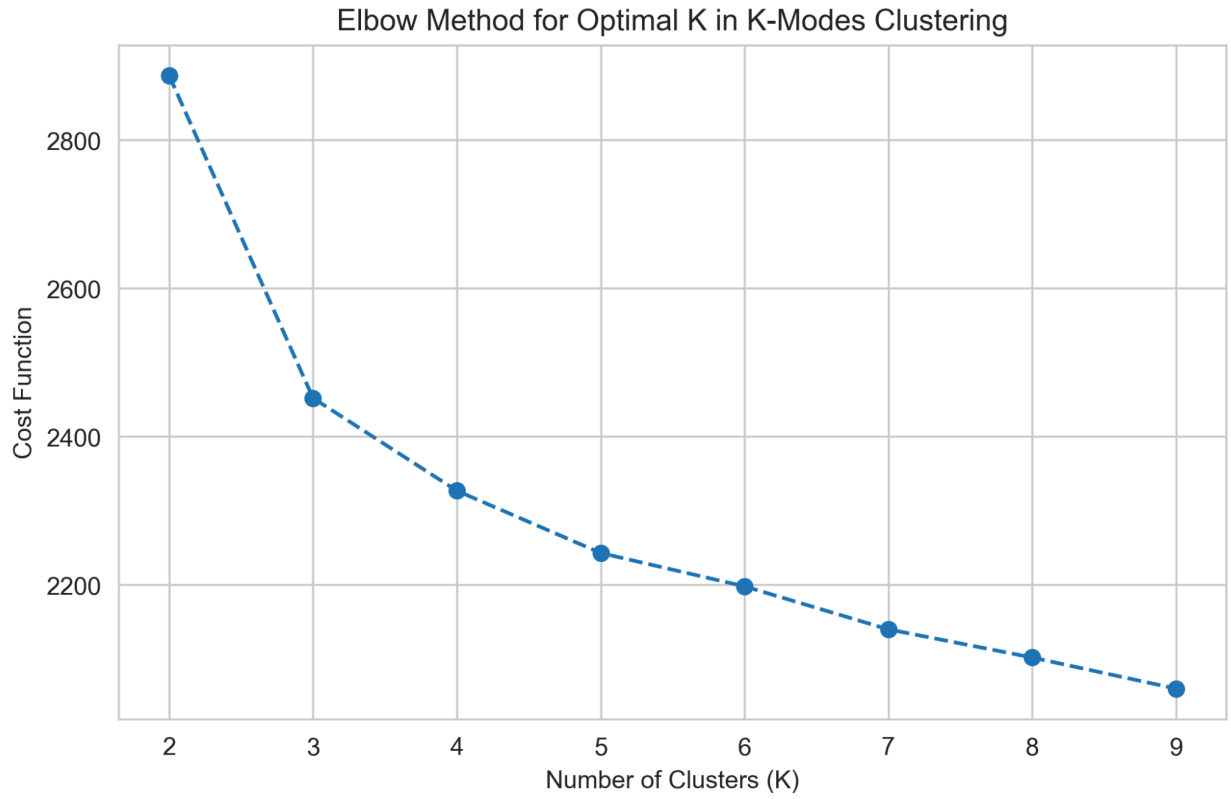

Figure 6: Elbow method for optimal K in K-Modes Clustering

The elbow method finds the optimal number of clusters by plotting the cost function against the number of clusters. The cost function measures the overall dissimilarity within each cluster. As per Figure 6, the optimal number of clusters using the elbow method is 4.

2. Dunn Index

```
def dunn_index(X, labels):
    unique_labels = np.unique(labels)
    intra_distances = []
    inter_distances = []
    for label in unique_labels:
        cluster_points = X[labels == label]
        if len(cluster_points) > 1:
            intra_distances.append(np.mean([np.sum(cluster_points != row) for row in cluster_points]))
    for i, label_i in enumerate(unique_labels):
        cluster_i = X[labels == label_i]
        for j, label_j in enumerate(unique_labels):
            if i != j:
                cluster_j = X[labels == label_j]
                inter_distances.append(np.mean([np.sum(cluster_i != row) for row in cluster_j]))
    return min(inter_distances) / max(intra_distances) if intra_distances and inter_distances else -1

# Convert categorical values to numeric codes (only for evaluation)
X_numeric = df_for_clustering.apply(lambda col: col.astype("category").cat.codes).to_numpy()
# Initialize evaluation metric containers
dunn_indices = []

# Evaluate clustering performance for different values of K
for K in K_range:
    km = KModes(n_clusters=K, init='Huang', n_init=5, verbose=0)
    labels = km.fit_predict(df_for_clustering)

    # Append evaluation metrics
    dunn_indices.append(dunn_index(X_numeric, labels))
```

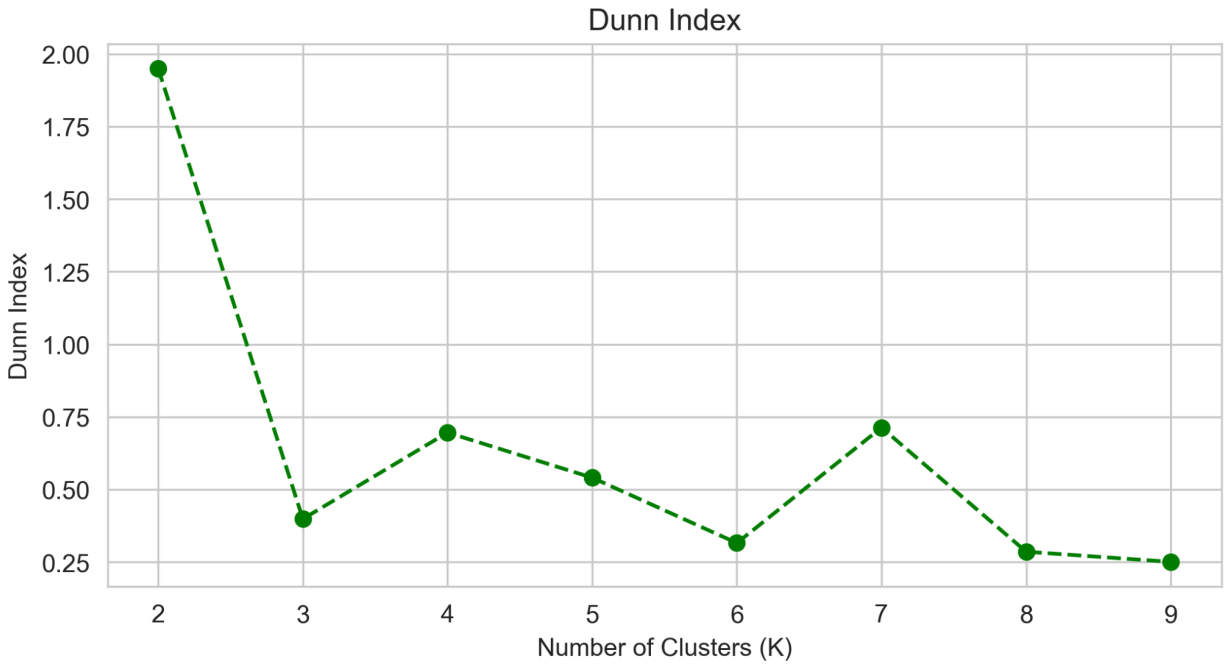

Figure 7: Dunn index for optimal K in K-Modes Clustering

The Dunn Index evaluates the quality of clusters by measuring the compactness of the clusters and the separation of the clusters from each other. As per Figure 7, the Dunn Index peaks at k=4 and k=7. Even though k=7 is higher, it is impractical to be taken as the optimal cluster count as it would produce fewer actionable micro clusters. Therefore, k=4 can be taken as the optimal number of clusters.

3. Categorical Silhouette Score

```
def categorical_silhouette_score(X, labels):
    unique_labels = np.unique(labels)
    silhouette_scores = []
    for i, label in enumerate(labels):
        same_cluster = X[labels == label]
        other_clusters = X[labels != label]
        if len(same_cluster) > 1:
            a = np.mean([np.sum(same_cluster != row) for row in same_cluster])
            b = np.min([np.mean(np.sum(other_clusters != row, axis=1)) for row in same_cluster])
            silhouette_scores.append((b - a) / max(a, b))
    return np.mean(silhouette_scores) if silhouette_scores else -1

# Initialize evaluation metric containers
silhouette_scores = []
```

```
# Evaluate clustering performance for different values of K
for K in K_range:
    km = KModes(n_clusters=K, init='Huang', n_init=5, verbose=0)
    labels = km.fit_predict(df_for_clustering)

    # Append evaluation metrics
    silhouette_scores.append(categorical_silhouette_score(X_numeric, labels))
```

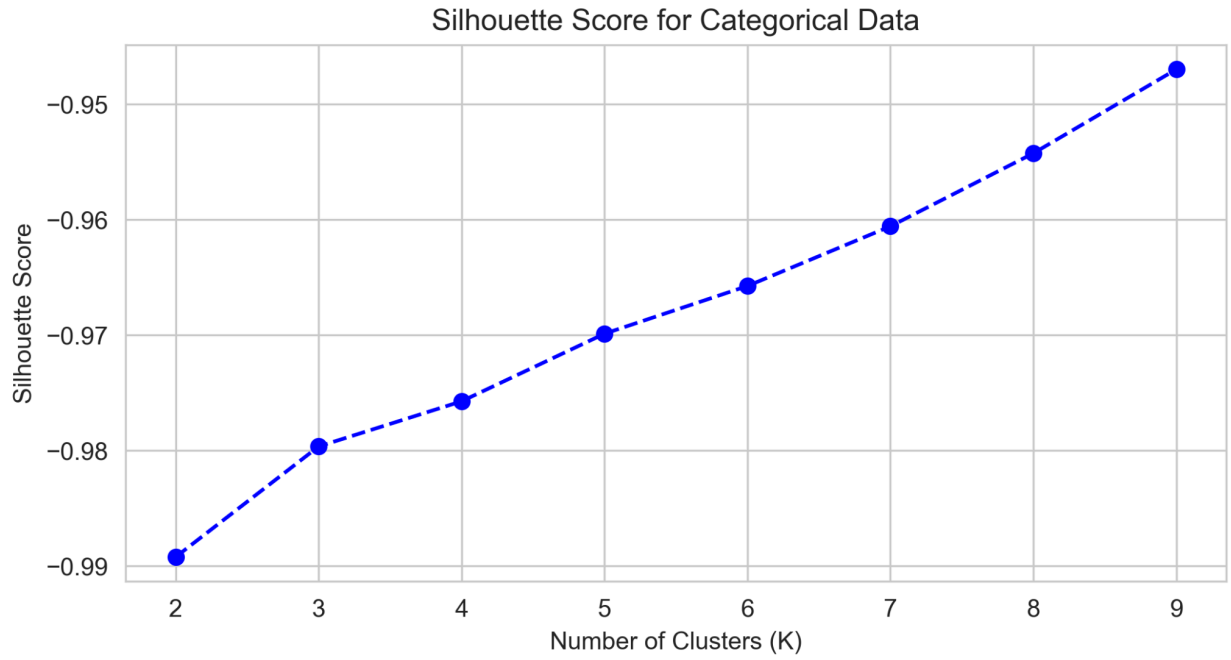

Figure 8: Silhouette score for categorical data

A custom function was implemented to evaluate silhouette scores for categorical data using numerical codes as proxies. As per Figure 8, k=4 is a strong midpoint since further improvements beyond k=4 are small. Therefore, k=4 can be considered as the optimal number of clusters according to the Silhouette score.

Execution of the Model

Based on the above 3 metrics, the model was applied to the dataset with the cluster count of 4. The "cao" initialization method was used to improve the cluster stability. The "n_init = 15" ensured that the model ran multiple times and selected the best output to ensure the robustness of the clustering.

```
# Optimal clusters (k=4) selected using elbow method, Dunn index, and categorical silhouette score.
# Apply K-Modes clustering to df_for_clustering
k = 4
kmodes_model = KModes(n_clusters=k, init="Cao", n_init=15, verbose=1)
cluster_labels = kmodes_model.fit_predict(df_for_clustering)
```

[62]

```
Initialization method and algorithm are deterministic. Setting n_init to 1.
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 1, iteration: 1/100, moves: 22, cost: 2340.0
Run 1, iteration: 2/100, moves: 10, cost: 2331.0
Run 1, iteration: 3/100, moves: 16, cost: 2320.0
Run 1, iteration: 4/100, moves: 1, cost: 2320.0
```

Figure 9: Model execution and clustering

6.1.3 Cluster Assignment

Once the clustering model was run and the clusters were assigned to each respondent, the next step was to summarize the results.

```
# Generate cluster summaries

cluster_summary = preprocessed_data.groupby("Cluster")["ridehailing_app"].value_counts(normalize=True).unstack().fillna(0) * 100

# Display as percentage distribution in each cluster
print(cluster_summary.round(2))
```

```
ridehailing_app  both pickme & uber  pickme only  uber only
Cluster
0                             100.0          0.0        0.0
1                               0.0        100.0        0.0
2                               0.0          0.0      100.0
3                             100.0          0.0        0.0
```

Figure 10: Cluster summaries

The purpose of the code in Figure 10 was to find out how the ride-hailing apps have been distributed among the clusters. This helps to identify what makes each cluster distinct. As per the result, cluster 0 and 1 have the customers who use both PickMe and Uber, cluster 1 has only customers who use PickMe, and cluster 2 only has the customers who use Uber. These insights provide logic that actionable segmentation has taken place in the model.

```python
text_profiles = []

for cluster in sorted(preprocessed_data["Cluster"].unique()):
    desc = f"🔹 **Cluster {cluster}**:\n"
    for col in profile_columns:
        cluster_data = preprocessed_data[preprocessed_data["Cluster"] == cluster][col].dropna()

        if not cluster_data.empty:
            try:
                mode_value = cluster_data.mode()[0]
                mode_pct = cluster_data.value_counts(normalize=True).loc[mode_value] * 100
                desc += f"- Most users are '{mode_value}' for **{col}** ({mode_pct:.1f}%)\n"
            except:
                pass # Silently skip if something goes wrong
    print(desc)
    print("-" * 60)
```

```
🔹 **Cluster 0**:
- Most users are '18-25' for **age** (62.1%)
- Most users are 'male' for **gender** (55.2%)
- Most users are 'colombo' for **district** (100.0%)
- Most users are 'lkr 50,000 – 99,999' for **monthly_income** (24.8%)
- Most users are 'both pickme & uber' for **ridehailing_app** (100.0%)
- Most users are 'to compare prices' for **b_preference** (23.4%)
- Most users are 'often' for **b_switchingfrequency** (53.1%)
- Most users are 'comfort and convenience' for **b_purpose** (18.6%)
- Most users are 'weekly' for **b_frequency** (48.3%)
- Most users are 'don't have a specific time' for **b_time** (49.7%)
- Most users are 'tuk' for **b_pickmevehicle** (36.6%)
- Most users are 'tuk' for **b_ubervehicle** (58.6%)
```

Figure 11: Text-based cluster profiles

The code block in Figure 11 provides text-based summaries of each cluster using the columns in the dataset. The value with the highest percentage in each column is presented under the summary of each cluster. This simplifies the complex dataset into an understandable interpretation of the characteristics of the user personas.

```
# Define a mapping from cluster number to descriptive label
cluster_label_map = {
    0: "Flexible Weekly Dual-App Users",
    1: "PickMe-Focused, Low-Frequency Cash Users",
    2: "Occasional Price Sensitive Uber Users",
    3: "Loyal Female Dominant Daily Riders with Subscriptions"
}

# Create the new column using the mapping
preprocessed_data["cluster_label"] = preprocessed_data["Cluster"].map(cluster_label_map)
```

Figure 12: Labelling the clusters

Next, based on the text-based profiles and the business knowledge, the clusters were assigned labels that represent their characteristics so that it is easier for the business stakeholders to have an understanding of each cluster.

Cluster 0: "Flexible Weekly Dual-App Users"
Cluster 1: "PickMe-Focused, Low-Frequency Cash Users"
Cluster 2: "Occasional Price-Sensitive Uber Users"
Cluster 3: "Loyal Female Dominant Daily Riders with Subscriptions"

These labels bring clarity to the analysis.

```
preprocessed_data.to_csv("clustered_data.csv", index=False)
print("Clustered data saved to 'clustered_data.csv'")
```

```
Clustered data saved to 'clustered_data.csv'
```

Figure 13: Clustered dataset export

Finally, the dataset was updated with the cluster numbers and the cluster labels, which were then exported to be used in the dashboard development.

6.2 Testing and Validation

6.2.1 Data Quality Validation

In the data cleaning stage, thorough validations were done to maintain the quality of the dataset by handling missing values, duplicates, and filtering.

```
# Verify missing
remaining_missing = colombo_df2["b_time"].isna().sum()
print(f"\nRemaining missing b_time values after imputation: {remaining_missing}")
```

```
Remaining missing b_time values after imputation: 0
```

Figure 14: Imputation verification

After the 3 levels of imputation methods were applied, the b_time column was tested for any missing values left, and the output of 0 confirms that the imputations were successful and all 31 rows are now filled.

6.2.2 Cluster Evaluation Metrics

K-Modes is an unsupervised clustering algorithm for categorical data, which means the traditional evaluation metrics cannot be applied here. Instead, the following metrics were used for the algorithm evaluation: they evaluate the cluster compactness, separation, and algorithm strength.

As discussed above, three metrics, the Elbow Method, the Dunn Index, and the Silhouette Score, were tested to validate the algorithm's optimal number of clusters. As per Figure 6, the total cost drops by around 200 from k=4 to k=7 compared to the 500 drop from k=2 to k=3. This shows no significance of adding more clusters beyond k=4. As per Figure 7, the Dunn Index also supports k=4 as the optimal number of clusters. Even though k=7 is mathematically accurate, k=4 brings more practicality to the model. As per Figure 8, the Silhouette Score increases gradually, although the values are negative, which is expected in categorical data. Using multiple metrics for this validated that the applied optimal number of clusters, which is k=4, is accurate.

```
Initialization method and algorithm are deterministic. Setting n_init to 1.
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 1, iteration: 1/100, moves: 22, cost: 2340.0
Run 1, iteration: 2/100, moves: 10, cost: 2331.0
Run 1, iteration: 3/100, moves: 16, cost: 2320.0
Run 1, iteration: 4/100, moves: 1, cost: 2320.0
```

Figure 15: Algorithm iterations

As per Figure 15, it is shown that the cost value converged just after four iterations from 2340 to a stable 2320. This indicates that the algorithm efficiently reached a stable solution, supporting that the model is logically sound.

The cluster sizes were checked after the model execution. Cluster 1: 147 (~38.3%), Cluster 0: 145 (~37.8%), Cluster 3: 51 (~13.3%), Cluster 2: 41 (~10.7%). This shows that the clusters are well balanced without any cluster being over 50%. Two large clusters can be seen, which is not a significant imbalance, and in segmentations like this, it is preferable.

6.3 Suggestions for Further Development

The success of the model lays the foundation for future improvements. While the current project is limited to the Colombo district, future projects can expand the geographical area, adding regional insights to the project. The project can also broaden the amount of data collected to have a more insightful cluster analysis, uncovering more user segments that were not revealed from this project. Furthermore, if Uber and PickMe provide ride activity data of the users, it will provide more accurate results for the behavioural patterns of the customers. Additionally, future work could develop and compare other models with the K-Mode algorithm to find out the optimal algorithm, as well as to find out the strengths and weaknesses of the algorithms. Future work could integrate multiple models and methods, and expand the scope of the project, possibly towards behavioural predictions or churn predictions.

Chapter Summary

The use of K-Modes for categorical data clustering was completely successful. The development was done by aligning the technical methodology and the real-world application. The multi-level testing techniques conducted demonstrated that this model wasn’t just developed and deployed; every critical step was thoroughly validated. Tests conducted in data cleaning stages made sure that there is a clean dataset with no structural errors, and the multi-run comparison of the algorithm verified stable cluster structures.

7. Results Analysis and Discussion

Chapter Overview

This chapter discusses the summary of the behavioural patterns, the demographic characteristics, and other key features of each cluster identified by the algorithm. The chapter also discusses the comparative analysis of the clusters conducted via dashboard visualizations, which is also readily available as an insights tool to the stakeholders to base their decision-making on. Additionally, actionable insights and recommendations are provided for the stakeholders that are based on the key findings and results of this project. Finally, potential improvements for the analysis are discussed for future developments in this chapter.

7.1 Cluster Summaries

The K-Modes clustering presented 4 distinct clusters by segmenting 384 customers who use ride-hailing services, PickMe and/or Uber in Colombo, Sri Lanka. These clusters were labeled as personas based on their key characteristics. Analyzing and comparing the key characteristics of these clusters provides a deep understanding of the needs and wants of their customers, which acts as a foundation for the respective companies to personalize their engagement strategies towards their customers.

7.1.1 Cluster 0: Flexible Weekly Dual-App Users (FWDA Users)

Around 38% of the customers belong to this cluster, making it the second-largest customer segment. Customers use both PickMe and Uber, often switching between the apps to compare the prices, making them price-sensitive customers. This segment is mostly comprised of young male customers with average income around Rs. 50000 to Rs. 100000. Closer to half of the customers use the apps weekly, however, not during a specific time of the day. Although the purpose of comfort and convenience drives these customers to use the ride-hailing apps, they are not motivated enough to use any subscription-based payment options.

```
**Cluster 0**:
- Most users are '18-25' for **age** (62.1%)
- Most users are 'male' for **gender** (55.2%)
- Most users are 'colombo' for **district** (100.0%)
- Most users are 'lkr 50,000 - 99,999' for **monthly_income** (24.8%)
- Most users are 'both pickme & uber' for **ridehailing_app** (100.0%)
- Most users are 'to compare prices' for **b_preference** (23.4%)
- Most users are 'often' for **b_switchingfrequency** (53.1%)
- Most users are 'comfort and convenience' for **b_purpose** (18.6%)
- Most users are 'weekly' for **b_frequency** (48.3%)
- Most users are 'don't have a specific time' for **b_time** (49.7%)
- Most users are 'tuk' for **b_pickmevehicle** (36.6%)
- Most users are 'tuk' for **b_ubervehicle** (58.6%)
- Most users are 'cash' for **b_paymentmethod** (57.9%)
- Most users are 'no' for **b_pickmepass** (82.1%)
- Most users are 'no' for **b_uberone** (76.6%)
```

Figure 16: Cluster 0 text-based summary

7.1.2 Cluster 1: PickMe-Focused, Low-Frequency Cash Users (PFLFC Users)

This is the largest segment of the model, with around 57% of them being female users. All the customers in this segment use only PickMe. Closer to 50% of them use PickMe occasionally, making them inactive customers of PickMe. Their low monthly income level could be a reason for their low engagement. The preference for using PickMe is driven by the better driver availability in the region and the familiarity of the app. Around 70% of the customers use cash as their preferred payment method. The purpose of using ride-hailing services diversifies in this segment, but comfort and convenience can be identified as priorities.

```
**Cluster 1**:
- Most users are '18-25' for **age** (49.0%)
- Most users are 'female' for **gender** (57.1%)
- Most users are 'colombo' for **district** (100.0%)
- Most users are 'less than lkr 50,000' for **monthly_income** (33.3%)
- Most users are 'pickme only' for **ridehailing_app** (100.0%)
- Most users are 'better availability in my area' for **p_pickmepreference** (22.4%)
- Most users are 'comfort and convenience' for **p_pickmepurpose** (19.0%)
- Most users are 'occasionally' for **p_pickmefrequency** (45.6%)
- Most users are 'tuk' for **p_pickmevehicle** (49.7%)
- Most users are 'don't have a specific time' for **p_pickmetime** (42.9%)
- Most users are 'cash' for **p_pickmepayment** (70.7%)
- Most users are 'no' for **p_pickmepass** (65.3%)
```

Figure 17: Cluster 1 text-based summary

7.1.3 Cluster 2: Occasional Price-Sensitive Uber Users (OPSU Users)

Being the smallest and the only cluster with male dominance, cluster 2 comprises customers who use only Uber as their ride-hailing app. A high percentage of the users in this category are between 18 to 24 years of age (70%). The users in this cluster are highly price sensitive. Affordable fares and better availability are the common choices among the users in this cluster.

Even though they use Uber, they are not motivated enough to purchase the subscription plans offered by loyalty. Around 73% of the users are without the Uber One subscription plan. Being occasional ride-hailing users could be the reason behind that. Around 76% of the users prefer cash as their primary payment method. Primary purposes for using Uber diversify among the users, comfort and convenience, and school and education are becoming the dominant purposes. The group appears to be students or budget travellers.

```
**Cluster 2**:
- Most users are '18-25' for **age** (70.7%)
- Most users are 'male' for **gender** (56.1%)
- Most users are 'colombo' for **district** (100.0%)
- Most users are 'less than lkr 50,000' for **monthly_income** (31.7%)
- Most users are 'uber only' for **ridehailing_app** (100.0%)
- Most users are 'affordable fares' for **u_uberpreference** (9.8%)
- Most users are 'comfort and convenience' for **u_uberpurpose** (17.1%)
- Most users are 'occasionally' for **u_uberfrequency** (34.1%)
- Most users are 'tuk' for **u_ubervehicle** (56.1%)
- Most users are 'don't have a specific time' for **u_ubertime** (34.1%)
- Most users are 'cash' for **u_uberpayment** (75.6%)
- Most users are 'no' for **u_uberone** (73.2%)
```

Figure 18: Cluster 2 text-based summary

7.1.4 Cluster 3: Loyal Female Dominant Daily Riders with Subscriptions (LFDRWS)

This cluster is dominated by female (72%) daily riders (43%) with a modest income level (Rs.50000 - Rs.100000). All the users in this segment are dual-app users. Most of the customers are subscribed to subscription services offered by the ride-hailing companies. Customers in this cluster don’t prefer a certain payment method, they use both card and cash (70%). Given their high engagement level, this group appears to be daily commuters.

```
**Cluster 3**:
- Most users are '18-25' for **age** (51.0%)
- Most users are 'female' for **gender** (72.5%)
- Most users are 'colombo' for **district** (100.0%)
- Most users are 'lkr 50,000 – 99,999' for **monthly_income** (29.4%)
- Most users are 'both pickme & uber' for **ridehailing_app** (100.0%)
- Most users are 'urgency or convenience' for **b_preference** (19.6%)
- Most users are 'always' for **b_switchingfrequency** (51.0%)
- Most users are 'comfort and convenience' for **b_purpose** (9.8%)
- Most users are 'daily' for **b_frequency** (43.1%)
- Most users are 'don't have a specific time' for **b_time** (41.2%)
- Most users are 'tuk, car' for **b_pickmevehicle** (47.1%)
- Most users are 'tuk, zip' for **b_ubervehicle** (49.0%)
- Most users are 'both cash & card' for **b_paymentmethod** (70.6%)
- Most users are 'yes' for **b_pickmepass** (54.9%)
- Most users are 'yes' for **b_uberone** (74.5%)
```

Figure 19: Cluster 4 text-based summary

7.2 The Dashboard Presentation and Insights

A dashboard was developed to present the key findings of the analysis. The visualizations created reveal trends and insights among clusters as well as opportunities to target customers strategically for business objectives. The dashboard has 4 main sections: Overview, Demographics by Profile, Behavioural Insights, and PickMe vs Uber Comparison.

7.2.1 Overview tab

This tab provides an outline of the dataset used for the model building. A total of 384 customers were clustered into 4 segments. The most common primary ride purpose was "comfort and convenience," followed by "emergencies" and "lack of public transport". A donut chart shows app usage: over 51% use both PickMe and Uber, 38% use PickMe only, and only 11% use Uber exclusively. The bar chart shows the app usage by cluster. Cluster 1 has only PickMe users, Cluster 2 has only Uber users. Clusters 0 and 3 have dual-app users.

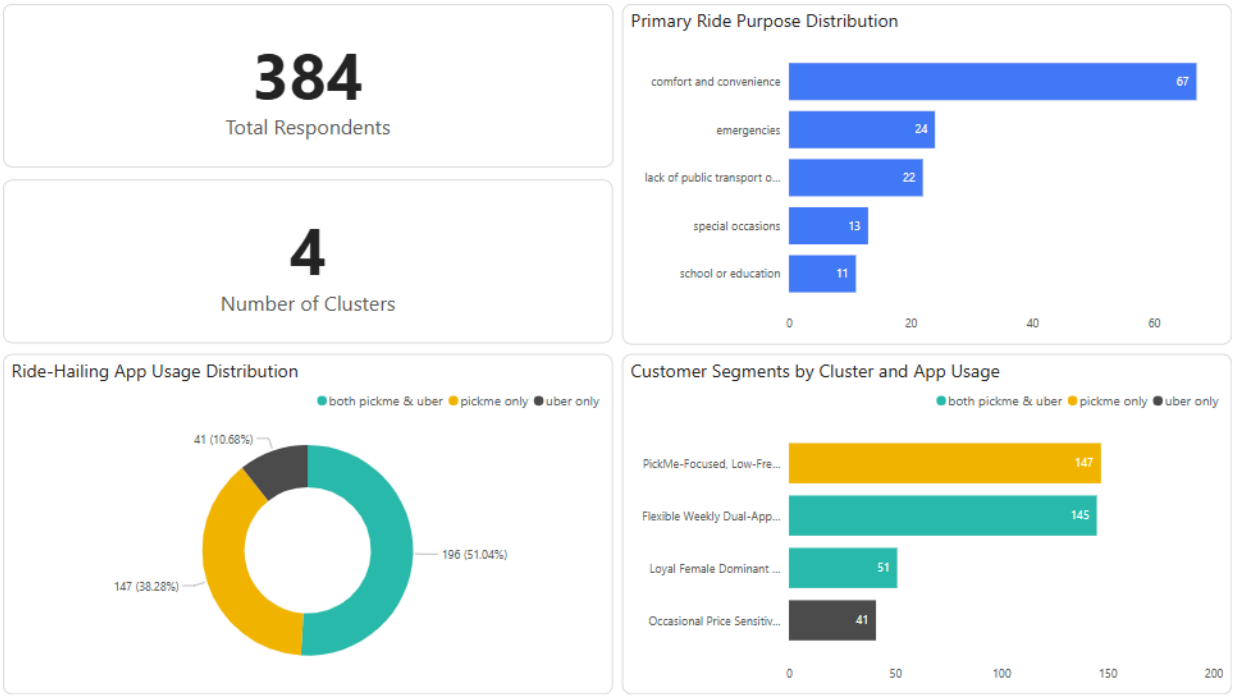

Figure 20: Overview tab

7.2.2 Demographics by Profile tab

This tab shows the demographics insights of the clusters. 4 100% stacked bar charts with clusters as legends used to show the distributions of clusters against demographic data.

FWDA users are the only segment with underage customers. They also have the highest percentage of youngest customers (42% of the age 18-25). Every other age group is dominated by the PFLFC users, with above 30% of the customers in each age group. They are also slightly female dominant (41%), while FWDA users are slightly male dominant (45%). LFDRWS ride daily, while other clusters ride mostly weekly and occasionally. PFLFC users and OPSU users skew toward lower-income brackets.

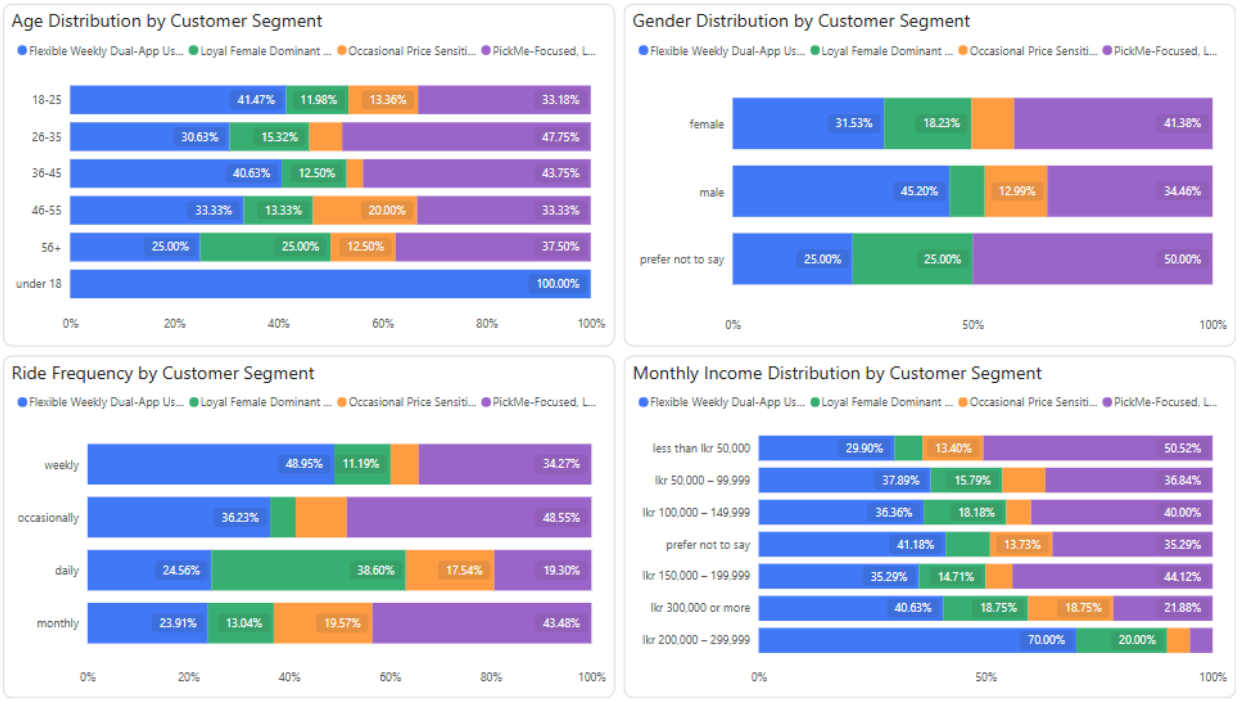

Figure 21: demographics by profile tab

7.2.3 Behavioural Insights tab

This tab has a slicer option to select the cluster and look into its behavioural data. Across all clusters, most users don't have a specific period of the day to use ride-hailing apps, and that is not the only common behaviour shown across all clusters. Most users tend to use ride-hailing services during office hours and evening, suggesting that the congestion in public transport could be a prominent reason for that. FWDA users and PFLFC users rely heavily on cash. LFDRWS relies on both cash and cards. The reason for FWDA users to choose both apps is to compare the prices at the time of the booking. Another reason is urgent situations, so that they can select the app that provides a driver in a lesser amount of time. OPSU users say that the low prices afforded by Uber are their reason for choosing Uber over PickMe. PFLFC users prefer better availability and shorter waiting times. They prioritize that over factors like pricing. Every cohort of customers uses the ride-hailing apps primarily because of comfort and convenience, and emergencies. This can be backed by the poor public transport conditions of the country and the high prices of private taxis.

Figure 22: Behavioural insights tab

7.2.4 PickMe vs Uber Comparison tabs

These two tabs compare user preferences with what the platform does.
PickMe users have a clear preference for tuk-tuks, with 317 responses indicating this. Uber users also prefer tuk-tuks with 216 responses. This shows that most people engage in long-distance rides.
User subscriptions: 105 users on PickMe Pass; 83 on Uber One.
App Preference Factors:
Availability, shorter wait times, and app interface are the attributes that PickMe users value.
Affordability, ease of use, and bad past experiences with PickMe are the reasons why Uber users praise the service.

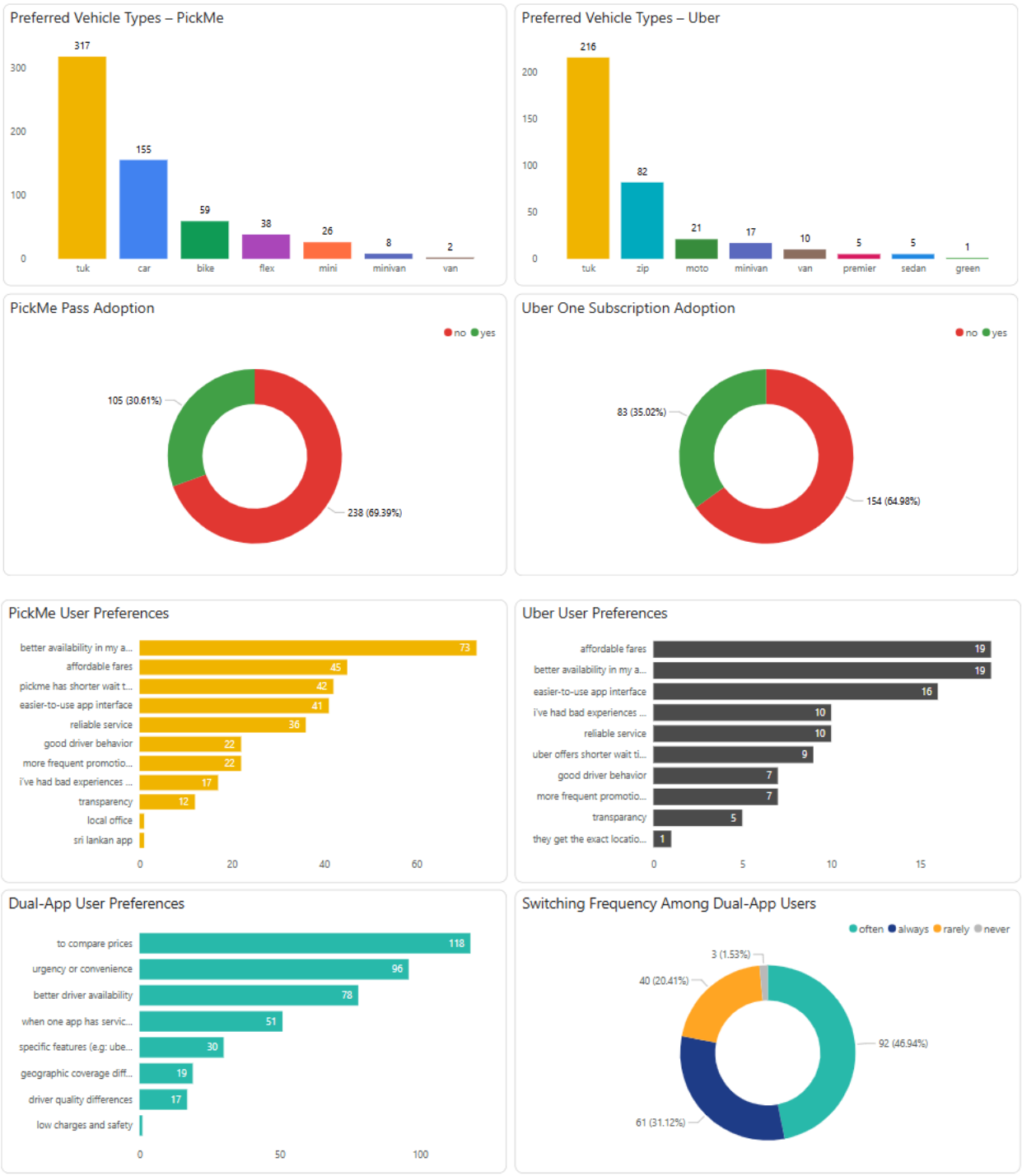

Figure 23: PickMe vs Uber comparison tabs

7.3 Business Recommendations

This section provides actionable recommendations for PickMe and Uber based on the insights from the cluster analysis and the dashboard findings. Businesses can strategize targeted marketing campaigns, operational campaigns to increase engagement, loyalty, and finally the revenue by approaching users based on their cluster-defined personas.

7.3.1 Recommendations for PickMe

Cluster 0: Flexible Weekly Dual-App Users (FWDA Users)
Since these customers are price-sensitive dual app users, PickMe can offer dynamic pricing with transparent fare breakdowns to have a competitive edge over Uber. These customers aren't motivated to subscribe to PickMe Pass, potentially due to their income level. Therefore, instead of focusing on subscriptions, PickMe can focus on a feathery reward scheme to increase their engagement with PickMe that will nudge to motivate them to go for the PickMe Pass. It will also help customer retention.

Cluster 1: PickMe-Focused, Low-Frequency Cash Users (PFLFC Users)
Even though this is the largest segment, they are not actively engaged with PickMe. What PickMe can focus on is reengagement strategies. PickMe seems to be already pushing SMS on ride discounts for the next few rides to their customers. This can be directed at customers who belong in this cluster to increase their engagement. For the customers who use PicMe for long-distance trips, PickMe could explore flexible payment options and partner with a BNPL provider like MintPay or Koko to split ride payments over multiple weeks. This will motivate users to shift from cash to card payments. It will also motivate the customers with financial difficulties to be loyal customers to PickMe.

Cluster 2: Occasional Price-Sensitive Uber Users (OPSU Users)
This cluster includes PickMe's competition. PickMe can launch awareness campaigns highlighting factors such as better availability of drivers, low prices, and other services provided by PickMe, other than ride-hailing. PickMe can also focus on improving customer support, which will be a crucial factor to motivate a customer to leave the service for a new one. Referrals via current users and new customer discounts can also be used to penetrate this cluster.

Cluster 3: Loyal Female Dominant Daily Riders with Subscriptions (LFDRWS)
PickMe could work on improving its subscription model by gamifying it with discounts as rewards for this segment. PickMe could focus on improving the quality of the service by improving the driver discipline, vehicle conditions, and providing options to have a preferred driver. PickMe could use the recommendations provided above for FWDA because both of the clusters are dual app users.

7.3.2 Reccomendations for Uber

Cluster 0: Flexible Weekly Dual-App Users (FWDA Users)
Uber could introduce mini ride packages with a lower purchase price instead of Uber One to acquire more engagement from the users, moving them away from PickMe. It could be designed in a manner that motivates customers to go for Uber One. Just like recommended to PickMe above, Uber could provide a transparent fare breakdown to customers, as they are a price-sensitive cohort.

Cluster 1: PickMe-Focused, Low-Frequency Cash Users (PFLFC Users)
Uber could use their international reputation to leverage the acquisition of current PickMe users by launching awareness and marketing campaigns showcasing use cases on how Uber has impacted other countries and even within Sri Lanka. As these customers are already less active with PickMe, they could be incentivized to join Uber with app download discounts, newcomer benefits, and providing discounted pricing for the first selected number of rides.

Cluster 2: Occasional Price-Sensitive Uber Users (OPSU Users)
This cohort is dominated by young users who are most likely students. A student package could be designed and tried out to improve the engagement of the young users. Also, marketing campaigns can be launched targeting young users to attract more young users to Uber, as well as to motivate the current young users. Additionally, promotions could be introduced in collaboration with universities and educational institutions to further increase the user count and user engagement.

Cluster 3: Loyal Female Dominant Daily Riders with Subscriptions (LFDRWS)
This is a high-value cluster for Uber. The recommendations provided above for PickMe can be adopted by Uber as well. Additionally, providing features such as female driver preference options exclusively for Uber One subscribers could be applied.

7.3.3 Combined recommendations for both platforms

Both companies are recommended to use their internal data sources and conduct need state analysis every six months to understand how the need states of customers change with the external factors such as economic and financial shifts. Both apps could improve the user interface, which is appealing to younger generations as well as simple for middle-aged and older generations. Companies should halt general promotions which doesn't cause much impact and tailor them to each cluster going forward. Both companies can continue to invest in driver training and customer care services, which are key factors for customer churn if negatively performed. While ride-hailing is the primary service of the two companies, that's not the only reason why customers purchase subscriptions, PickMe Pass, or Uber One. That's because they

provide benefits in food delivery as well. Therefore, improving those services will also act as a motivations for people to purchase the subscriptions as well.

Chapter Summary

This chapter analysed the cluster results and extracted actionable insights. The chapter discusses the dashboard visualizations as well, which provide further information about the clusters. Based on the findings, recommendations were provided to PickMe and Uber, respectively, enabling them to approach the clusters effectively. Additionally, combined recommendations were provided to both companies.

8. Conclusions and Reflections

Reflecting on the Project

The project successfully delivered the core objective of the project, comparative need state analysis for ride-hailing customers who use PickMe and Uber in Colombo, Sri Lanka, using K-Modes clustering. The project involved the design and publishing survey to collect data, preprocessing the data, Implementation of the clustering algorithm, validation of the results, visualizing the key findings as a dashboard, and providing business recommendations to PickMe and Uber.

K-Modes clustering was taken as the best algorithm to segment the customers, given the nature of the dataset. While no other research has used K-Modes in the ride-hailing context, the results from this project show clearly that the model is excellent at providing interpretable clusters from categorical data. The clustering process was led by filtering Colombo district ride-hailing customers and imputations to handle missing values, along with an exploratory data analysis. A total of 384 customers were analysed. Evaluation metrics such as the Elbow method, Dunn Index, and Categorical Silhouette Score were used to test the quality of the algorithm and to decide the optimal number of clusters. The balanced cluster distribution further proved the robustness of the algorithm.

Four user personas were created by analysing the cluster characteristics, namely, Flexible Weekly Dual-App Users, PickMe-Focused, Low-Frequency Cash Users, Occasional Price-Sensitive Uber Users, and Loyal Female Dominant Daily Riders with Subscriptions.

The clusters were summarised and analysed to create user personas to help stakeholders understand the types of customer segments there are who use ride-hailing in Colombo. A dashboard was developed using Power BI to further identify characteristics and patterns of the clusters. Demographic, behavioural, and app comparison data were visualized to assist in decision-making. Finally, the insights gained from clustering, analysing, and visualizing were transformed into business recommendations for both companies.

Strengths of the Project

The primary strength of the project lies in being the first academic application to apply K-Modes clustering in the Sri Lankan ride-hailing context. The project follows end-to-end execution from

survey design to presentation of an interactive dashboard, the full life cycle of a data science project. The insights also align directly with what PickMe and Uber would consider for engagement strategies. The evaluation results further strengthen the integrity of the project.

Limitations of the Project

As for the limitations, this project reveals only the insights of Colombo district users. It doesn't apply to other areas of the country as their needs, wants, and behaviours will be different from the users in Colombo. Also, since the data collection was done using surveys, it could lead to selection bias and other respondent mistakes. Furthermore, the project is static. It only captures the current trends and patterns. Therefore, this analysis has to be repeated from time to time to capture the evolving needs of the customers.

Skills Development

The project improved technical skills such as conditional survey design, Python, clustering, and Power BI. It also helped to use frameworks like CRISP-DM along with agile principles. The project also improved critical thinking, especially when applying analytical results to business recommendations. Data preprocessing skills were also developed during instances like missing value imputations.

Future Work

This project can be further developed by integrating real-time data from PickMe and Uber to acquire app usage data, which will reveal more insights into customer segments. The analysis can be branched out to a predictive model of customer churn using clusters as the features. Furthermore, this analysis can be implemented to other services provided by the apps, such as food delivery and parcel delivery, and extract insights about those customers as well. The project is not limited to ride-hailing. It can be used in other industries such as retail and telecommunications as well.

This project has bridged the gap between user behaviour and business strategy. It contributes to practical decision-making.

9. References

Cendana, M. and Kuo, R.-J. (2024) ‘Categorical Data Clustering: a Bibliometric Analysis and Taxonomy’, Machine Learning and Knowledge Extraction, 6(2), pp. 1009–1054. Available at: https://doi.org/10.3390/make6020047.

Clewlow, R.R. and Mishra, Gouri S (2017) Disruptive Transportation: the Adoption, Utilization, and Impacts of Ride-Hailing in the United States, Escholarship.org. Available at: https://escholarship.org/uc/item/82w2z91j.

Delali Kwasi Dake, Gyimah, E. and Buabeng-Andoh, C. (2023) ‘University Students Behaviour Modelling Using the K-Prototype Clustering Algorithm’, Mathematical Problems in Engineering, 2023, pp. 1–13. Available at: https://doi.org/10.1155/2023/5507814.

Gomes, M.A. and Meisen, T. (2023) ‘A Review on Customer Segmentation Methods for Personalized Customer Targeting in e-commerce Use Cases’, Information Systems and e-Business Management, 21(21), pp. 527–570. Available at: https://doi.org/10.1007/s10257-023-00640-4.

Gonaldson, V. and Sunitiyoso, Y. (2024) ‘Ride-Hailing Landscape in Indonesia: a Segmentation Analysis Perspective’, International Research Journal of Economics and Management Studies, 3(5), pp. 220–233. Available at: 10.56472/25835238/IRJEMS-V3I5P127.

Goyal, M. (2017) ‘A Review on K-Mode Clustering Algorithm’, International Journal of Advanced Research in Computer Science, 8, pp. 725–729. Available at: https://doi.org/10.26483/ijarcs.v8i7.4301.

Juma James Masele and Shayo, E.E. (2025) ‘Determinants of Customer Satisfaction for Ride-hailing Services in Tanzania’, Tanzania Journal of Development Studies, 22(2), pp. 1–25. Available at: https://doi.org/10.56279/njiy8787/tjds.v22i2.1.

Lee, S., Lee, W., Vogt, C.A. and Zhang, Y. (2021) ‘A Comparative Analysis of Factors Influencing Millennial Travellers’ Intentions to Use ride-hailing’, Information Technology & Tourism [Preprint]. Available at: https://doi.org/10.1007/s40558-021-00194-6.

Mohd, A., Lay Eng Teoh and Hooi Ling Khoo (2024) ‘Passengers’ Requests Clustering with k-prototype Algorithm for the first-mile and last-mile (FMLM) shared-ride Taxi Service’,

Multimodal Transportation, 3(2), pp. 100132–100132. Available at: https://doi.org/10.1016/j.multra.2024.100132.

Piñón Rodríguez, A. (2024) Enhancing Public Transport Utilisation through User Behaviour Clustering Case Study from Helsinki Region.

Rafiq, R. and McNally, M.G. (2022) 'An Exploratory Analysis of Alternative Travel Behaviors of ride-hailing Users', Transportation [Preprint]. Available at: https://doi.org/10.1007/s11116-021-10254-9.

Saxena, P. (2023) Clustering COVID-19 Travellers' Perceptions and Behaviours for Tourism Recovery: Using K-Modes.

Shah, S.A.H. and Kubota, H. (2022) 'Passenger's Satisfaction with Service Quality of app-based Ride Hailing Services in Developing countries: Case of Lahore, Pakistan', Asian Transport Studies, 8, p. 100076. Available at: https://doi.org/10.1016/j.eastsj.2022.100076.

Sharma, N. and Gaud, N. (2015) 'K-modes Clustering Algorithm for Categorical Data', International Journal of Computer Applications, 127(17), pp. 1–6. Available at: https://doi.org/10.5120/ijca2015906708.

Sikder, S. (2019) 'Who Uses Ride-Hailing Services in the United States?', Transportation Research Record: Journal of the Transportation Research Board, 2673(12), pp. 40–54. Available at: https://doi.org/10.1177/0361198119859302.

Tirachini, A. (2019) 'Ride-hailing, Travel Behaviour and Sustainable mobility: an International Review', Transportation, 47. Available at: https://doi.org/10.1007/s11116-019-10070-2.

V Kondur, N. (2018) Using K-Mode Clustering to Identify Personas for Technology on the Trail. MSc Thesis.

Young, M. and Farber, S. (2019) 'The who, why, and when of Uber and Other ride-hailing trips: an Examination of a Large Sample Household Travel Survey', Transportation Research Part A: Policy and Practice, 119, pp. 383–392. Available at: https://doi.org/10.1016/j.tra.2018.11.018.

Zhou, Y., Yuan, Q., Yang, C. and Wang, Y. (2021) 'Who You Are Determines How You travel: Clustering Human Activity Patterns with a Markov-chain-based Mixture Model', Travel Behaviour and Society, 24, pp. 102–112. Available at: https://doi.org/10.1016/j.tbs.2021.03.005.

10. Bibliography

Acheampong, R.A., Siiba, A., Okyere, D.K. and Tuffour, J.P. (2020) 'Mobility-on-demand: an Empirical Study of internet-based ride-hailing Adoption factors, Travel Characteristics and Mode Substitution Effects', Transportation Research Part C: Emerging Technologies, 115, p. 102638. Available at: https://doi.org/10.1016/j.trc.2020.102638.

Ali, N., Javid, M.A., Campisi, T., Chaiyasarn, K. and Saingam, P. (2022) 'Measuring Customers' Satisfaction and Preferences for Ride-Hailing Services in a Developing Country', Sustainability, 14(22), p. 15484. Available at: https://doi.org/10.3390/su142215484.

Elnadi, M. and Gheith, M.H. (2022) 'What Makes Consumers Reuse ride-hailing services? An Investigation of Egyptian Consumers' Attitudes Towards ride-hailing Apps', Travel Behaviour and Society, 29, pp. 78–94. Available at: https://doi.org/10.1016/j.tbs.2022.06.002.

Lee, C.K.H. and Wong, A.O.M. (2021) 'Antecedents of Consumer Loyalty in ride-hailing', Transportation Research Part F: Traffic Psychology and Behaviour, 80, pp. 14–33. Available at: https://doi.org/10.1016/j.trf.2021.03.016.

Muchlis Muchlisin, Soza-Parra, J., Susilo, Y.O. and Ettema, D. (2024) 'Unraveling the travel patterns of ride-hailing users: A latent class cluster analysis across income groups in Yogyakarta, Indonesia', Travel Behaviour and Society, 37, pp. 100836–100836. Available at: https://doi.org/10.1016/j.tbs.2024.100836.

Nguyen, D.G. and Ha, M.-T. (2022) 'What Makes Users Continue to Want to Use the Digital Platform? Evidence from the Ride-Hailing Service Platform in Vietnam', SAGE Open, 12(1), p. 215824402110691. Available at: https://doi.org/10.1177/21582440211069146.

Nguyen-Phuoc, D.Q., Su, D.N., Tran, P.T.K., Le, D.-T.T. and Johnson, L.W. (2020) 'Factors Influencing customer's Loyalty Towards ride-hailing Taxi Services – a Case Study of Vietnam', Transportation Research Part A: Policy and Practice, 134(1), pp. 96–112. Available at: https://doi.org/10.1016/j.tra.2020.02.008.

11. Appendix

Appendix A: Links to Supplementary Materials

1. GitHub Repository (Source Code & Documentation):
   Click here to access GitHub

2. Interactive Power BI Dashboard:
   Click here to access the dashboard

3. Video Presentation (Google Drive Link):
   Click here to access the demo video

Appendix B: Additional Screenshots

Name of the Supervisee	**Dinith Perera**
Student ID	**20200912**
Project Title	**The Need State Analysis**

Meeting #	Date	Time	Discussion/Guidance given	Tasks to complete before next meeting
1	30-Aug	4:30 PM	Presented the idea of doing a need state analysis for PickMe customer base to identify their wants and neeeds in order to cater them through marketing and operation strategies.	1. Fill the Ethics form so that IIT can give a confirmation letter 2. Consult PickMe whether they would give the releant data and whats the process of obtaining them.
2	7-Oct	10:30 AM	PickMe told they will have to take another week to give a reply about the data request. I was asked to create a survey to gather data just in case PickMe doen't provide the data and to do a literature survey.	1. Create a sample survey 2. Literature Survey
3	22-Oct	14:30 PM	PickMe declined the provide the data for the analysis. Therefore I created and presented the survey questions flow and received feeback. Will complete the survey and send it to Mrs. Abarnah for feeback within the week.	1. Send the complete survey to Mrs.Abarnah
4	25-Nov	10:00 AM	Discussed about the technical approach that can be taken to cluster the data points. Currently going with a k-mode analysis.. Presented the improved flow of the survey.	1. Fix the issues related to the flow of the survey 2. Make the nessasary changes to the survey according to the feedback
5	17-Dec	10.30 AM	Discussed the current issues in the survey and issues related to the timeline of the project. Decided to remove some questions and add another questions to the survey.	1. Send the finalized survey to both supervisors. 2. Publish the survey and collect data.
6	10-Jan	2:00 PM	Discussed progress of the data collection. Was advised to start writing the report. while working on the code.	1. Data collection 2. Report upto Literature review
7	29-Jan	4:00 PM	Discussed about the Interim Progress submission. Was advised to finish the code upto model building.	1. Finish the model
8	21-Feb	3:00 PM	Presented the progress. Discussed the issues of the code. Advises were given on potential remedies.	1. Finish the dashboard 2. Finish the report
9	25-Apr	3:30 PM	Discussed about the final report. Was advised to finish the report early for feedback. Presented the progress of the report.	1. Finish the dashboard 2. Finish the report
10	2-May	2:30 PM	Discussed about the content of the report and the progress.	

Figure 24: Meetings log