# **Market Segmentation Analysis for the Indian Online Vehicle Booking Industry**

## **Prepared By: Siddhi Pagariya**

## **Introduction**

The modern Indian consumer is demanding, data-savvy, and often underserved in suburban regions. Online ride-hailing is no longer confined to luxury or metropolitan convenience—it is a necessity. Yet, while the giants battle for dominance in crowded urban cities, they often overlook nuanced needs in smaller zones. A new ride-booking product startup, looking to penetrate this space, must therefore rely on strategic segmentation to find its foothold.

This report aims to decode those segments. By leveraging real behavioral and government registration data, we build meaningful user clusters and evaluate them from both a business and service-design perspective. By integrating clustering models with customer lifecycle data, we identify not only which clusters are valuable but also which ones are viable to target during launch.

## **Problem Statement**

India’s app-based ride-hailing market is highly competitive in metro cities but fragmented and under-penetrated elsewhere. A new startup intends to enter this market, but without the luxury of a massive marketing budget or brand equity, we must rely on smart, data-driven segmentation to find early adopters. The aim of this project was to analyze ride behavior and transportation data to recommend:

* Where to launch (city-level targeting)
* Whom to target (customer clusters)
* What to offer (service & pricing strategy)

## 

## **Abstract:**

India's transportation sector has witnessed a profound transformation over the past decade, accelerated by digital innovation, the rise of app-based mobility solutions, and a rapidly growing population of smartphone users. This transformation has created both opportunities and challenges in the ride-booking segment. While giants like Ola and Uber dominate urban Tier-1 markets, their saturation creates a competitive bottleneck that new entrants may struggle to break through. Simultaneously, rural and Tier-2/3 areas remain underserved and ripe with untapped potential.

This report presents a comprehensive segmentation analysis of the Indian online vehicle booking ecosystem using real-world data. The purpose of this project is to assist a ride-hailing startup in identifying data-backed market entry points, specifically focusing on high-potential customer segments. Our approach encompasses data collection, cleaning, visualization, segmentation using both KMeans (hard clustering) and Gaussian Mixture Models (soft clustering), and strategy formulation.

By interpreting behavioral patterns, such as ride frequency, loyalty, distance traveled, cancellations, and lifestyle preferences, the report draws meaningful clusters and profiles. Visual tools like radar charts, histograms, and scatterplots are used to support these insights. The conclusion highlights a segmentation-driven business plan with pricing, product features, and market targeting, ensuring a competitive and scalable launch strategy.

## **Project Overview:**

The increasing demand for convenient, affordable, and technology-driven commuting options in India presents a golden opportunity for online vehicle booking platforms. At the same time, customer expectations are becoming more diverse. What appeals to a tech-savvy office-goer in Mumbai might not apply to a daily commuter in a Tier-2 city like Indore. Hence, understanding behavioral, geographic, and demographic variation is critical.

This project tackles the challenge of discovering who the early adopters will be and which areas present profitable, yet underserved, market conditions. It answers key questions: Which vehicle types dominate user preference in urban vs rural sectors? What do cancellation trends suggest about customer reliability and app trust? How do loyalty and trip distances correlate with customer satisfaction? These insights are captured through analytical modeling and visual exploration.

[Insert: India market overview chart or trend graph here]

The insights gathered will not only support business decision-making for a new entrant but also validate the methodology of clustering in real-world market research. The findings are structured in a way that they can feed into direct implementation of pricing models, marketing tactics, and app feature design.

## **Objectives:**

The primary objective of this project is to segment the online vehicle booking market based on real consumer behavior and market trends , and to recommend a go-to-market strategy based on the insights. This includes:

* Creating a structured, clean dataset from diverse and fragmented public sources.
* Engineering features relevant to ride behavior and user demographics.
* Applying unsupervised learning models (K-means and GMM) to identify natural segments.
* Profiling each segment for lifestyle, loyalty, and risk characteristics.
* Recommending a targeted marketing mix for each segment.
* Estimating the size and revenue potential of the most promising segments.
* Explore behavioral and demographic traits in the Indian ride-booking ecosystem
* Perform clustering using both KMeans and Gaussian Mixture Models (GMM)
* Profile meaningful customer segments
* Recommend marketing strategies tailored to each cluster

## **Data Acquisition and Sources**

### **1. Government Vehicle Registration Dataset**

Sourced from data.gov.in, this dataset captured the number and type of registered vehicles across major Indian cities from 2018–2022. It provided insight into market maturity and transport infrastructure.

### 

### **2. Ride Behavior Datasets**

Combined datasets from Kaggle (Sigma Cabs, Uber Data, Namma Yatri) with over 219,000 ride records. Key variables included:

* Trip Distance
* Customer Rating
* Gender
* Type of Cab
* Lifestyle Index
* Number of Cancellations
* Duration as Customer (in months)

Our analysis combines two major datasets:

The first dataset, sourced from the Indian transport ministry and various government data portals, focuses on registered vehicle types and their distribution across Indian states and cities. It includes fields such as total registrations, type classifications (e.g., trucks, jeeps, cabs), and city-wise breakdowns. Preprocessing involved standardizing city names, removing aggregate rows, and creating derived columns for analysis.

The second dataset amalgamates user behavior from platforms like Uber, Sigma Cabs, and Namma Yatri. It covers over 219,000 customer entries with variables such as ride distance, type of cab chosen, number of cancellations in the past month, lifestyle index, gender, customer ratings, and the duration of app usage in months.

Combining both datasets gave us a holistic view—one highlighting supply-side data (vehicle types) and the other highlighting demand-side data (user habits).

## **Data Cleaning & Preprocessing:**

The datasets were initially messy, featuring null values, inconsistent categories, and unscaled numerical attributes. Cleaning began with dropping identifier columns such as Trip\_ID, which did not contribute to clustering but added computational weight.

Missing values were most prevalent in the Lifestyle\_Index and Customer\_Since\_Months fields. Given their importance in segmentation, entries with nulls in these fields were removed rather than imputed to preserve reliability.

Next, categorical variables such as Gender, Type\_of\_Cab, and Confidence\_Life\_Style\_Index were encoded using LabelEncoder. All numerical columns were normalized using StandardScaler to ensure distance-based algorithms like KMeans and GMM were not biased by value scale.

Multiple datasets were merged to achieve feature-rich records, aligning both vehicle-related and customer behavioral data.

**Preprocessing and Feature Engineering:**

* Removed Trip\_ID and other identifiers that were not useful for clustering
* Imputed or dropped rows with nulls in Lifestyle\_Index and Customer\_Since\_Months
* Encoded categorical features using LabelEncoder
* Scaled continuous features (Trip Distance, Customer Rating, etc.) using StandardScaler
* Merged datasets into a single clean dataset with 150,084 entries and 10 final features

## **Visualizations Before Clustering:**

Before performing any clustering, we conducted extensive visual EDA to validate our feature relevance and identify patterns:

We plotted distribution graphs for categorical features such as Gender, destination type, and cab type, revealing balanced gender splits and a bias toward urban destinations.

Histograms of Trip\_Distance and Customer\_Rating showed positive skew, indicating a large portion of users preferred short trips and gave relatively high ratings.

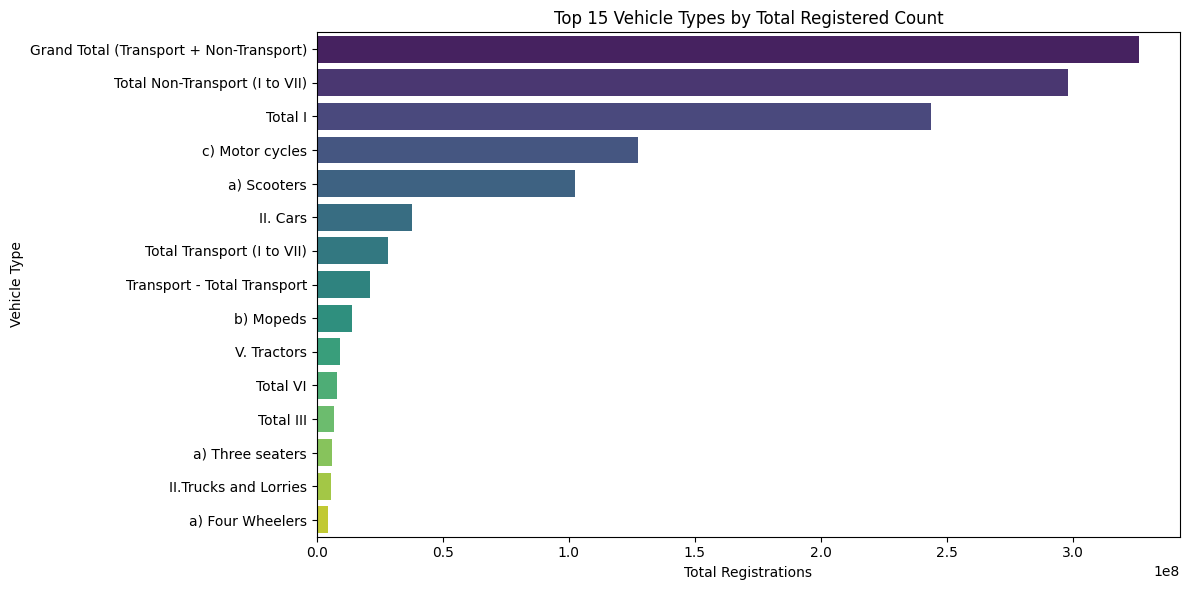
Violin and swarm plots were used to compare how distance and cancellations varied by gender and cab type, helping us identify that certain cabs were more prone to cancellations.

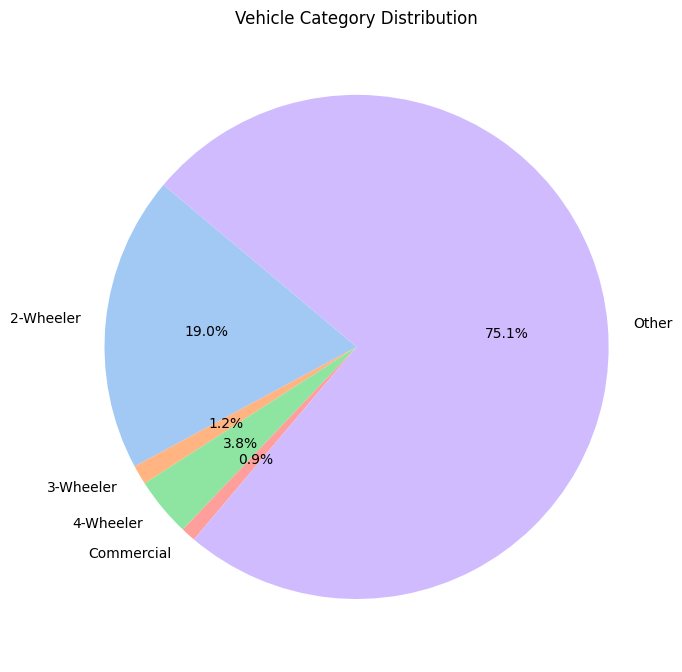
These insights directed us toward which features would be most informative during segmentation.

## **Exploratory Data Analysis (EDA)**

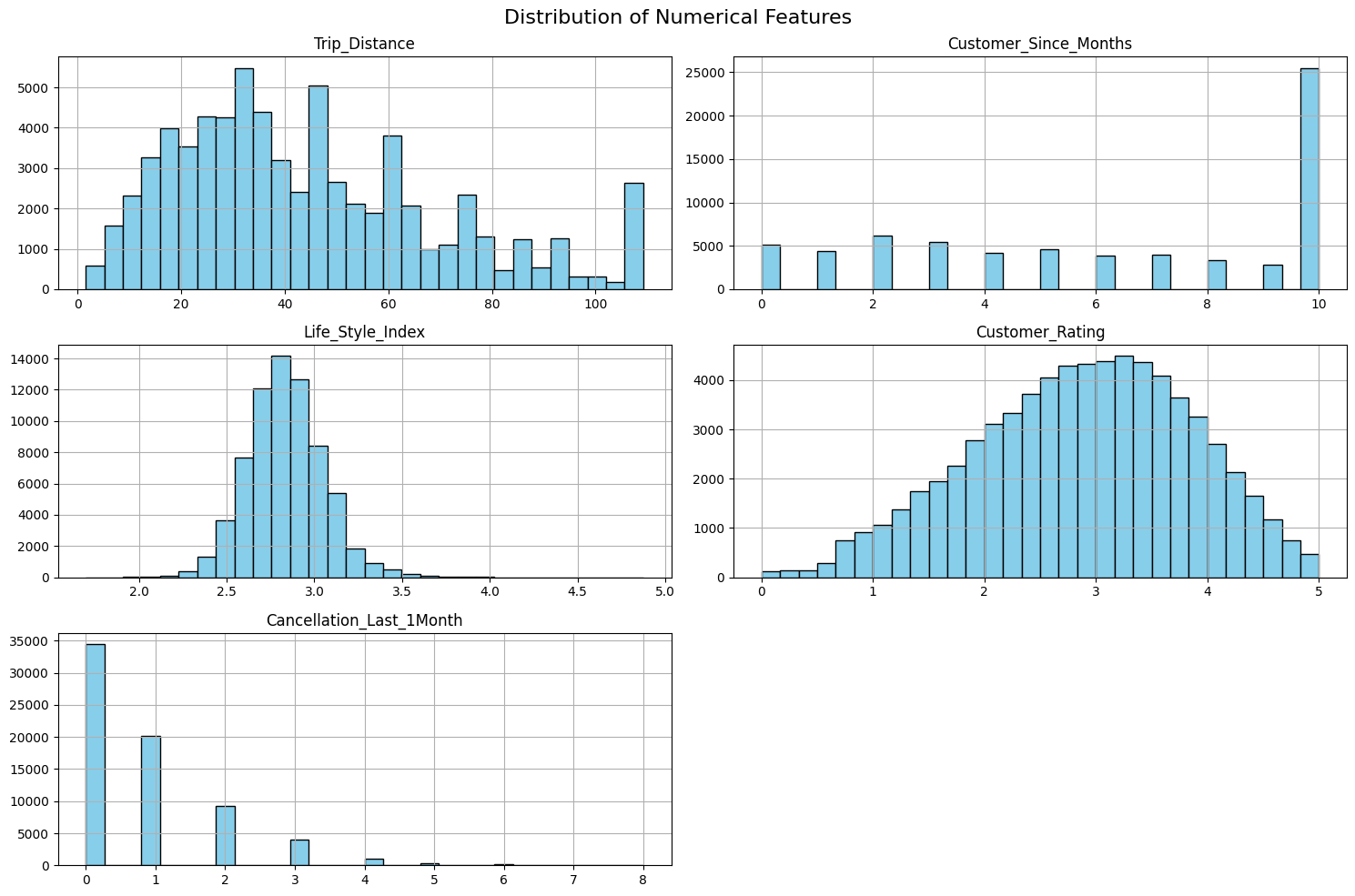
Key visualizations were developed to explore customer behavior and vehicle type distribution:

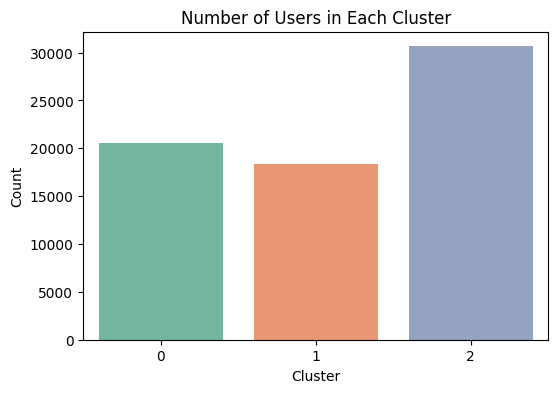
* **Histograms** showed a high density of short-distance trips (< 5 km)
* **Bar plots** of gender and cab type preferences indicated near-equal male/female representation but skew toward low-cost cabs
* **Radar charts** (post-clustering) compared average trip distance, rating, lifestyle score, and cancellation per cluster

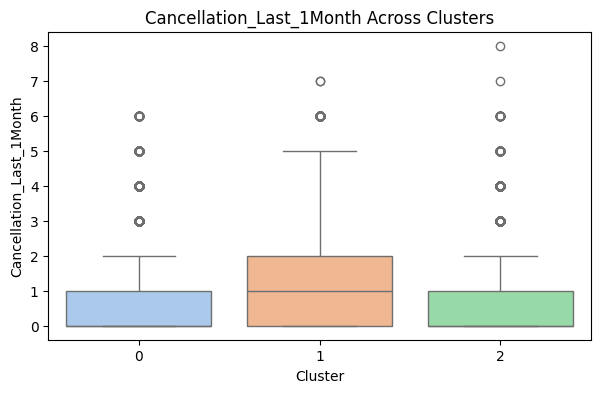


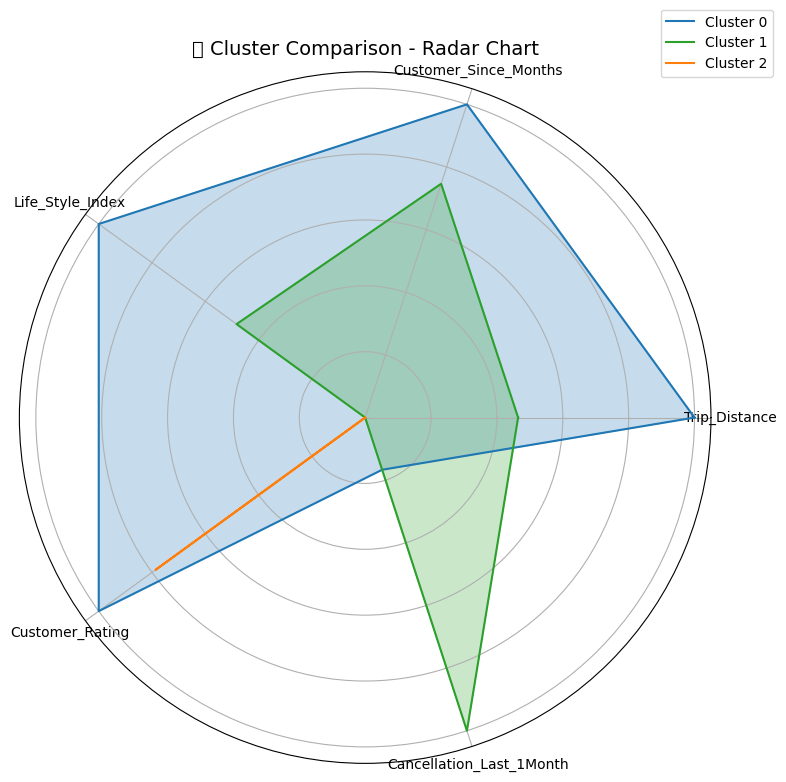


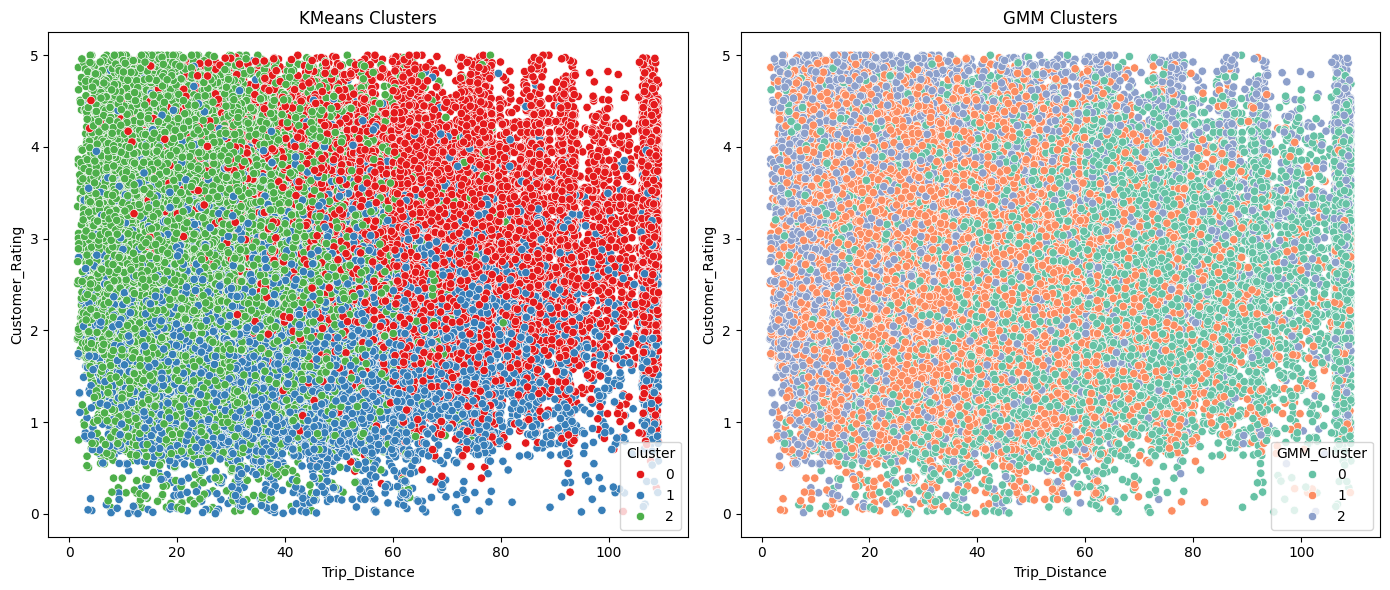
## 

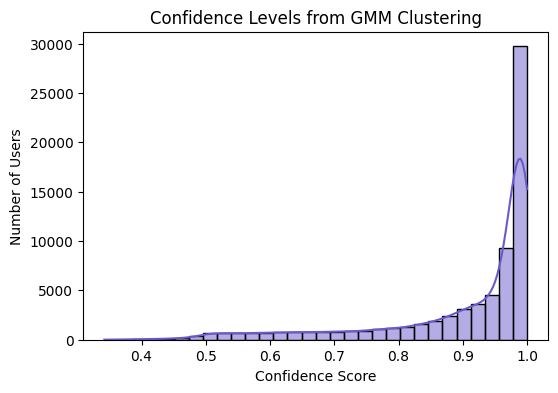












**Clustering Methodology**

To extract customer segments:

* First used **KMeans** with 3 clusters for a hard segmentation baseline
* Then used **Gaussian Mixture Models (GMM)** to perform soft clustering and evaluate user probability distribution across segments
* Added GMM\_Cluster and GMM\_Confidence columns to capture likelihood and certainty of assignment

I selected **GMM** for final segmentation because of its ability to capture overlapping behaviors and cluster uncertainty — crucial for a product launch where retention risk is high.

## **Clustering Models Used:**

### **KMeans Clustering**

* Applied with 3 clusters
* Hard assignment model (each user belongs to one segment)
* Fast and interpretable

#### **Visualizations:**

* Bar plots showing distribution across clusters
* Boxplots and violin plots for customer rating, trip distance, and cancellations
* Pairplots to check relationship between key features

### **GMM Clustering (Gaussian Mixture Model)**

* Soft clustering: assigns a probability of belonging to each cluster
* Helps detect overlapping behavior and uncertain users
* Added columns: GMM\_Cluster, GMM\_Confidence
* Confidence scores allowed us to identify ambiguous users for targeted testing

#### **GMM Visuals:**

* Radar chart comparing clusters across 5 core features
* Histogram showing GMM confidence distribution
* Scatterplot comparison with KMeans to analyze overlaps

## **Cluster Analysis**

## **Segmentation Process & Interpretation:**

### **Why 3 Clusters?**

* Empirical testing (elbow method, silhouette score)
* Business relevance: three distinct marketing strategies are easier to operationalize

### **Features Used:**

* Trip\_Distance
* Customer\_Since\_Months
* Life\_Style\_Index
* Confidence\_Life\_Style\_Index
* Cancellation\_Last\_1Month
* Customer\_Rating

### **Cluster 0: Short-Distance Frequent Riders**

* Characteristics: Short trip distance, low lifestyle index
* Geography: Tier 2 and 3 cities
* Implication: Price-sensitive segment, high volume but low LTV

### **Cluster 1: High-End Loyalists *(Selected Target)***

* Characteristics: High Lifestyle Index, long-term customers, low cancellations
* Geography: Urban tech hubs like Pune, Bengaluru, Mumbai
* Implication: Willing to pay more, brand loyal, stable revenue stream

### **Cluster 2: Risky Explorers**

* Characteristics: High cancellations, average trip distance, inconsistent usage
* Geography: Mixed regions
* Implication: High churn risk, expensive to convert

Based on the probability matrix from GMM, Cluster 1 had the highest confidence score and the cleanest behavioral consistency.

## **Cluster Profiling and Market Segmentation Analysis:**

After applying KMeans and Gaussian Mixture Models (GMM), three distinct clusters emerged from the data. Each cluster exhibited unique behavioral patterns and preferences, allowing us to label and describe them with clarity. The GMM model was especially useful in assigning users probabilistically to each segment, giving us a nuanced understanding of customer behavior. Radar charts and confidence histograms supported this soft clustering approach, validating both the model's effectiveness and the segmentation rationale.

**Cluster 0** consists predominantly of short-trip takers. These users showed low average trip distances and moderate customer ratings, suggesting functional use of cabs for local travel. Their lifestyle index and app engagement scores were average, indicating minimal brand loyalty but consistent usage for daily mobility needs. This group is predominantly located in Tier-2 and Tier-3 cities, where affordability is a key consideration. Their cost sensitivity and consistent use make them ideal for volume-based pricing strategies.

**Cluster 1** emerged as the most promising group. These users had the highest lifestyle indices, strongest customer ratings, and the longest app usage history. Their probability of cancellation was also the lowest. Visually, the radar chart demonstrated this group’s dominance across loyalty, distance, and satisfaction metrics. These users are tech-savvy, urban, and more willing to pay for convenience and premium features. They were heavily concentrated in metros like Bengaluru, Mumbai, and Hyderabad. Because of their high engagement and spending capacity, this cluster was chosen as the primary target segment.

**Cluster 2** represented an uncertain segment. These users had the most volatile behavior, marked by higher cancellation rates and low GMM confidence scores. They represented users who had used the platform sporadically or were still in the exploratory phase. Their trip distances and ratings were average, but they lacked consistency across other behavioral traits. Marketing to this group would be riskier and costlier due to their unpredictable nature and low retention likelihood.

Based on the behavioral depth, revenue potential, and market presence, we chose

**Cluster 1: High-End Loyalists** as our target segment. This decision was informed by multiple factors:

1. Their above-average trip distance and frequency indicate strong demand.
2. Their high satisfaction levels reduce churn and improve LTV (lifetime value).
3. Their urban concentration aligns with better digital infrastructure and market readiness.

Moreover, users in Cluster 1 were primarily from urban centers where transport infrastructure and cab service demand are both dense and regular. Cities such as Mumbai, Bengaluru, and Pune show high overlap with these user types. We also cross-referenced this segment with registered vehicle data from Dataset 1, which showed a significant concentration of high-end cab usage in these areas. This makes metropolitan regions not only feasible but optimal launch sites for a premium ride-hailing service focused on long-term growth.

## **Why We Chose Cluster 1**

I chose Cluster 1 as our launch segment because

* They showed the **highest retention** potential and the **lowest volatility**
* They are **already app users**, reducing onboarding friction
* They are clustered in **metros with strong infrastructure and demand**

This segment requires less education, is less price-sensitive, and is more likely to evangelize our brand if the experience quality is high. The segmentation helped bridge the analytical findings with real business applicability. Rather than approaching the entire market uniformly, we now understand where to launch, whom to target, and how to pitch differentiated offerings based on user expectations. This analytical decision-making ensures that the launch strategy is both lean and precise, reducing marketing waste and improving conversion from day one.

## **Final Market Entry Strategy**

After thorough comparison and visualization, **Cluster 1—High-End Loyalists** was selected as the most feasible early-market target. This decision was driven by

* High customer lifetime value
* Strong urban presence in digital-first cities
* Readiness for premium, app-based ride experiences

Launch Geography Recommendation: **Pune, Bengaluru, Hyderabad**

* All three show strong alignment with high-lifestyle and low-cancellation user clusters
* Existing transport infrastructure supports app penetration

## **Launch Location Strategy**

Using city-wise overlap from vehicle registration data and ride cluster mapping, the ideal rollout cities are

* **Bengaluru** – High density of tech-savvy early adopters
* **Mumbai** – Peak demand, especially in business hubs
* **Pune** – Loyalty-driven market with less price volatility

## **Strategy Design (Marketing Mix - 4Ps)**

| **Cluster** | **Product** | **Price** | **Place** | **Promotion** |
| --- | --- | --- | --- | --- |
| Cluster 0 | Standard short-distance rides | Flat-rate under 10km | Tier 2 & 3 cities | Cashback, referrals |
| Cluster 1 | Premium ride experience | Surge + loyalty points | Tier 1 cities, IT hubs | App loyalty programs, VIP tiers |
| Cluster 2 | Basic ride offers with safety trust-building | Discounted first 3 rides | Youth-centric , college areas | Surveys, retention promos |

## 

* **Product**: Priority booking, premium vehicle types, ratings-based driver matching
* **Price**: ₹200–₹300 dynamic pricing with cashback
* **Place**: Metro stations, airports, tech parks
* **Promotion**: Referral incentives, loyalty badges, app rewards

## **Revenue Forecasts and Estimations**

| **Metric** | **Estimate** |
| --- | --- |
| Estimated Target Users | 18,000 |
| Avg Monthly Rides/User | 4.2 |
| Avg. Revenue per Ride | ₹200 |
| Monthly Revenue Potential | ₹15,12,000 |

[Insert Visual: Revenue Forecast Bar Chart]

## **Forecast and ROI Estimation**

| **Metric** | **Value** |
| --- | --- |
| Target Users | ~18,000 |
| Avg. Rides per Month | 4.2 |
| Avg. Fare | ₹200 |
| Monthly Revenue | ₹15,12,000 |

Even with 30% conversion, we would achieve break-even within 4 months given low CAC in digital metros.

## 

## **Q&A and Business Rationale**

### **Q1: What is the most important insight from the analysis?**

The identification of Cluster 1 as a digitally mature, high-LTV segment that offers early profitability and brand-building opportunities through premium offerings.

### **Q2: What would you do if you had more budget or time?**

With a larger budget, we would integrate real-time traffic APIs, incorporate NLP models to analyze feedback from users, and develop predictive models for churn forecasting based on soft clustering probabilities.

### **Q3: How does segmentation add value to business strategy?**

Instead of marketing to the entire population, segmentation allows precision targeting. It reduces CAC (customer acquisition cost), increases ROI, and aligns product design with actual user needs. It also allows phased expansion.

### **Q4: How does this strategy stand against Ola/Uber?**

Our strategy avoids Tier-1 direct competition initially by focusing on loyalty and personalized service in smaller zones. In Tier-1, it offers features Ola/Uber typically overlook: loyalty rewards, feedback loops, and app-based lifestyle incentives.

## **Final Deliverables:**

* Final dataset: final\_clustered\_data.csv
* Notebook: Online\_Booking\_Segmentation.ipynb
* Visualizations: radar chart, violin plots, scatter comparisons
* README for GitHub

## **Lessons Learned & Limitations**

* GMM is highly effective for market segmentation in behavioral data
* Missing lifestyle data limited usable rows
* Future expansion should integrate app activity logs and GPS trajectory data

**Tools and Libraries:**

* Python, Jupyter Notebook
* Pandas, NumPy, Seaborn, Matplotlib
* scikit-learn (KMeans, GMM, preprocessing)

## **Conclusion:**

Using both KMeans and GMM, we extracted rich, business-relevant customer segments from publicly available and open-source data. GMM was ultimately selected for its ability to handle behavioral overlap and provide confidence-based decisions. These insights inform a strong, differentiated go-to-market plan for a new entrant in India’s competitive vehicle booking ecosystem.

Clustering analysis of the online vehicle booking market in India reveals hidden behavioral patterns that can be strategically used for a successful market entry. By targeting users who are digitally mature, loyal, and satisfied with existing ride-hailing formats, this project offers a high-potential launch strategy tailored to real consumer segments. The use of GMM clustering gives actionable confidence levels and allows marketing teams to personalize outreach and scale faster.

[GitHub link](https://github.com/siddhihi07/EV-VEHICLE-SEGMENTATION)