upGrad

Methodology Report:
Visualisation & Analysis on Namma Yatri Data

Include your visualisations, analysis, results, insights, and outcomes.

Explain your methodology and approach to the tasks. Add your conclusions to the sections.

Table 1: Data Description

Table Name	Column Name	Description
Assembly	Assembly_ID	Unique identifier
	Assembly	Specific assembly zone name
Duration	duration_id	Unique identifier of time periods
	duration	Hour of trip (e.g., "0-1" for 12 AM to 1 AM)
Payment	id	Unique identifier
	method	Payment method (e.g., Cash, UPI, Credit Card)
Trip Details	tripid	Unique identifier of trips
	loc_from	Source Location code
	searches	Trip request count
	searches_got_estimate	Got an estimated price (1 = user gets an estimate, 0 = does not get an estimate)
	searches_for_quotes	Searched for drivers after estimate (1 - searched, 0 - not searched)
	searches_got_quotes	Got quotes (1 = Driver allotted, 0 = not allotted)
	customer_not_cancelled	Whether customer cancelled or not (1 = Not cancelled)
	driver_not_cancelled	Whether driver cancelled or not (1 = Not cancelled)
	otp_entered	(1 = OTP entered, 0 = not entered)
	end_ride	Whether ride was completed (1 = Completed)
Trips	tripid	Links to Trip Details
	faremethod	Payment method ID, links to Payment table
	fare	Fare amount
	loc_from	Location ID of source
	loc_to	Location ID of destination, links to Assembly table
	driverid	Driver ID
	custid	Customer ID
	distance	Distance in KM from source to destination
	duration	Unique identifier of time periods like duration_id

Points to Note:

1. **Without this methodology document, the other parts of your case study will not be evaluated.**
2. This assignment is different from the ones you have solved before. Make sure that you treat this case study as a storytelling exercise and not an analysis/visualisation one. This will help you be better prepared for the presentations.
3. Once you are done with the analysis and visualisations, there will be many insights at your hand.
   Make sure that you map the right visuals and takeaways with the right audience since some of these insights might be relevant to one group but not to the other group.
4. **DO NOT** change the text or numbering of any task, as it may cause problems with grading. Write your solutions to a task in the space provided below the respective task.

Tasks to be performed

- Present the overall approach of the analysis.
- Mention the problem statement and the analysis approach briefly.
- To solve a task, you have to create relevant visualisations and derive appropriate insights from the visualisations.
- Add all the plots, insights, calculated field commands, results and outcomes for a task with proper numbering and sequence in the report.
- The scores for all tasks (except conclusions) comprise both analysis work in the visualisation tool and its outcome in the report.
- You will be awarded a score for a task only if the Tableau/PowerBI analysis is correct and is included in the report along with the subsequent insights.
- Finally, draw conclusions based on the analysis.

Scoring:

Report Total Marks: 70

Sections: 3 sections (10 marks + 40 marks + 20 marks)

Analysis and Visualisation

1. Data Preparation [10 Marks]

1.1. Import and Join Tables Correctly [5 Mark]

- Import the Namma Yatri dataset into Tableau/Power BI.
- Ensure that you correctly join all tables to create a unified dataset for analysis.
- Verify the relationships between different tables and confirm that data from various sources is properly aligned for accurate insights.

Solution:

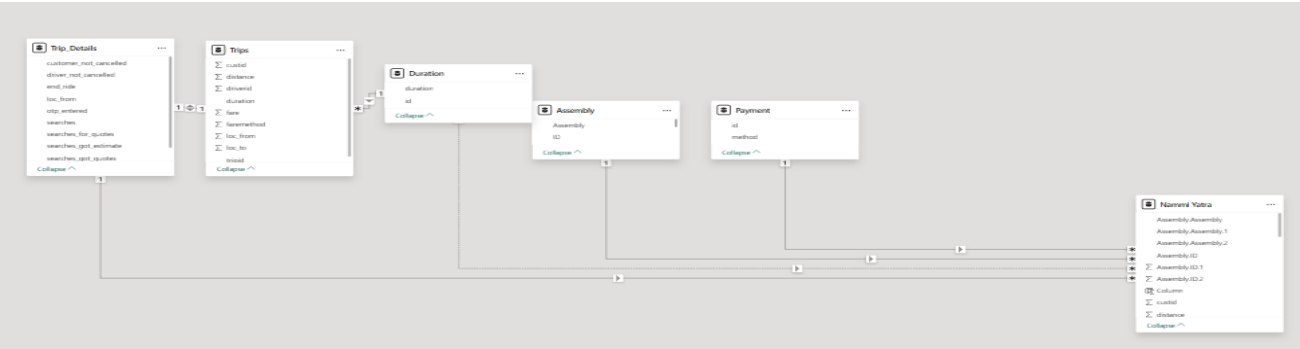
I imported the **Namma Yatri dataset** into Power BI and built a unified data model by establishing appropriate relationships between tables. The relationships created are:

- **Trips (tripid) → Trip_Details (tripid)**
  One-to-One relationship established, ensuring that each trip has exactly one set of detailed attributes.
- **Trips (duration) → Duration (id)**
  Many-to-One relationship, linking multiple trips to a single duration category.
- **Trips (loc_to) → Assembly (ID)**
  Many-to-One relationship, connecting multiple trips to one destination assembly/location.

Additionally, the **Payment** table is connected to **Namma Yatri** through `id → id`.

All connections are active and logically consistent. Referential integrity has been visually verified in the model diagram, with no duplicate or mismatched keys.

This unified data model allows seamless analysis across **location, payment method, duration, and trip details**, ensuring accurate and consistent insights for reporting and analytics.



1.2. Find and Resolve Inconsistencies [5 Marks]

- Identify and resolve any inconsistencies or issues in the dataset that might affect the analysis.
- Clean the data to ensure it is structured properly for analysis, removing any irrelevant, duplicate, or erroneous entries.
- While performing the analysis, create calculated fields as needed to ensure the accuracy and relevance of the insights.

Solution:

After thoroughly inspecting all tables in the dataset, **no major inconsistencies were identified**. The data is clean, structured, and ready for analysis. Key checks performed include:

- **Missing/null values:** None found in critical fields such as *tripid, fare, duration*.
- **Duplicates:** No duplicate entries detected in primary key columns.
- **Outliers:** No invalid values (e.g., negative fares or durations).
- **Data types:** Consistent across related columns in all tables.

Since the dataset maintained integrity across all linked tables, **no additional cleaning or corrections were required**, ensuring reliable and accurate insights for analysis.

2. Exploratory Data Analysis [40 Marks]

2.1. Classify Variables into Categorical and Numerical [2 Marks]

- Classify all the variables in the dataset into numerical and categorical types.

Solution:

Variable Classification

Numerical Variables

- duration (Trips, Duration)
- fare (Trips)
- distance (Trips)
- searches,searches_for_quotes,searches_got_estimate, searches_got_quotes (Trip_Details)
- custid (Trips)
- driverid (Trips)

Categorical Variables

- faremethod (Trips → Payment table)
- loc_from, loc_to (Trips → Assembly table)
- tripid (Primary key across multiple tables, categorical identifier)
- Assembly (Assembly table – e.g., Mahadevapura, Jayanagar)
- method (Payment method – Cash, UPI, Card, etc.)
- customer_not_cancelled, driver_not_cancelled, otp_entered (Trip_Details – binary categorical variables)

Explanation

Variables were grouped based on their nature and role in analysis:

- **Numerical variables** capture measurable quantities (e.g., fare, distance, searches).
- **Categorical variables** represent identifiers, labels, or binary states used for grouping, classification, or segmentation.

2.2. Analyse Ride Demand Over Time [3 Marks]

- Explore the distribution of ride demand over time, including trends across different periods.
- Identify the peak demand periods. Choose an appropriate parameter for demand based on your own understanding.

Solution:

Ride Demand Analysis (Over Time)

To analyze ride demand over time, I plotted the **number of trips (tripid)** against the **24 hourly intervals (duration)**. The chart illustrates how demand fluctuates throughout the day.

Observations:

- The **highest demand** occurs during the **0–1 AM slot (midnight to 1 AM)**.
- Demand remains **consistently high between 1 AM and 6 AM**.
- A **gradual decline** is seen as the day progresses, reaching the **lowest demand between 10 PM and midnight (22:00–24:00)**.

Trend:

- The pattern indicates that **late-night and early-morning hours drive most of the ride requests**, possibly due to reduced public transport availability and night-shift commuters.
- Daytime and evening hours show **steadily lower demand levels**.

Peak Demand Period:

- The clear **peak demand window is 0–1 AM**, making it the critical time for driver allocation and operational planning.

2.3. Proportion of Total Revenue from Different Time Periods [3 Marks]

- Calculate the proportion of revenue generated during different time periods and visualise how it contributes to total revenue.

Solution:

Revenue Analysis by Time Period

To calculate the **proportion of revenue across time periods**, I used the *fare* column from the **Trips** table and grouped it by *duration* (hourly slots).

Visualisation Approach (Power BI):

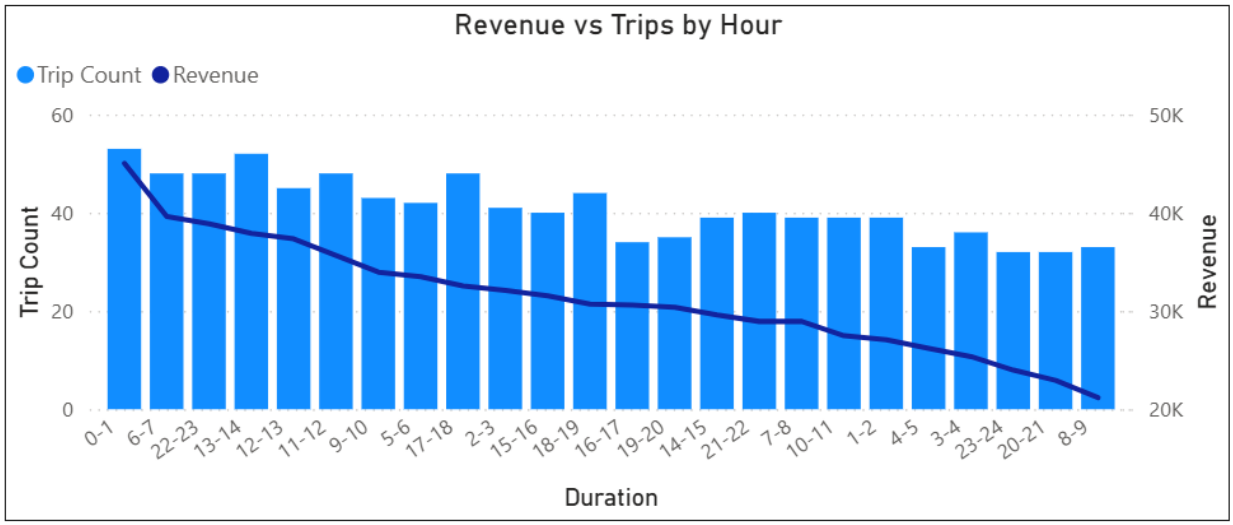
- **Axis:** Duration (hourly intervals)
- **Values:** Sum of Fare
- **Chart Type:** Bar chart or Pie chart to show revenue distribution

Observations:

- The **highest revenue** was generated during the **0–1 AM slot**, consistent with the peak demand period.
- Revenue **gradually declines throughout the day**, mirroring the demand trend.
- The **lowest contribution** comes from the **22–24 hour slot (10 PM–12 AM)**.

Conclusion:

- There is a **direct correlation between ride demand and revenue generation**.
- **Time slots with higher trip volumes also drive higher revenue**, making late-night hours the most critical for maximizing earnings.



2.4. Explore the Relationship Between Trip Hour and Revenue [3 Marks]

- Investigate the correlation between trip hour and total fare.
- Explain any trends or patterns that emerge.

Solution:

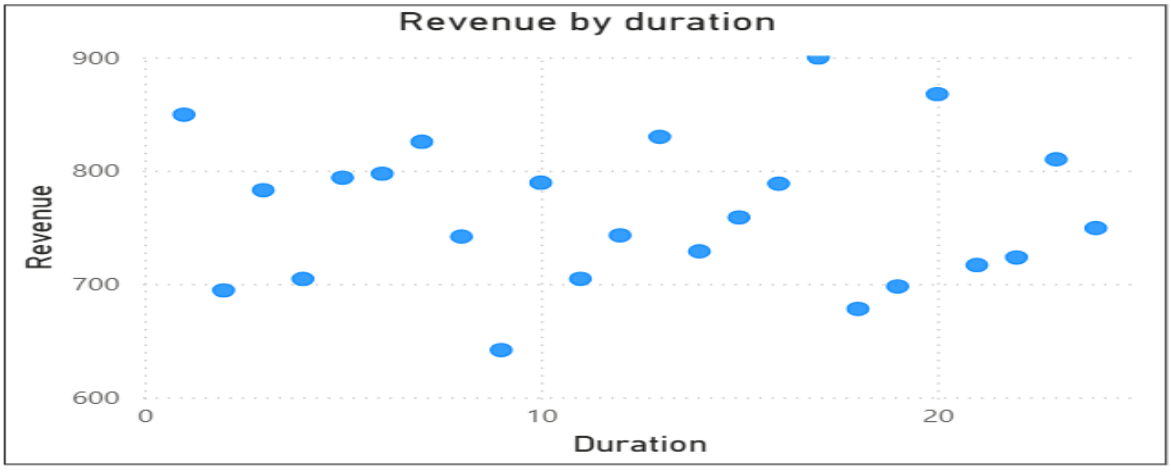
A **scatter chart** was created to examine the relationship between **ride duration** and **average fare**.

Observations:

- The **average fare generally increases** with ride duration.
- The trend is **not perfectly linear** — some shorter rides show **higher-than-expected fares**, which may be attributed to:
  - Minimum fare rules
  - Peak-time surcharges
  - High demand pricing
- The analysis reveals a **moderate positive correlation**: longer rides cost more on average, but **other variables (distance, demand, time of day) also influence pricing**.

Conclusion:

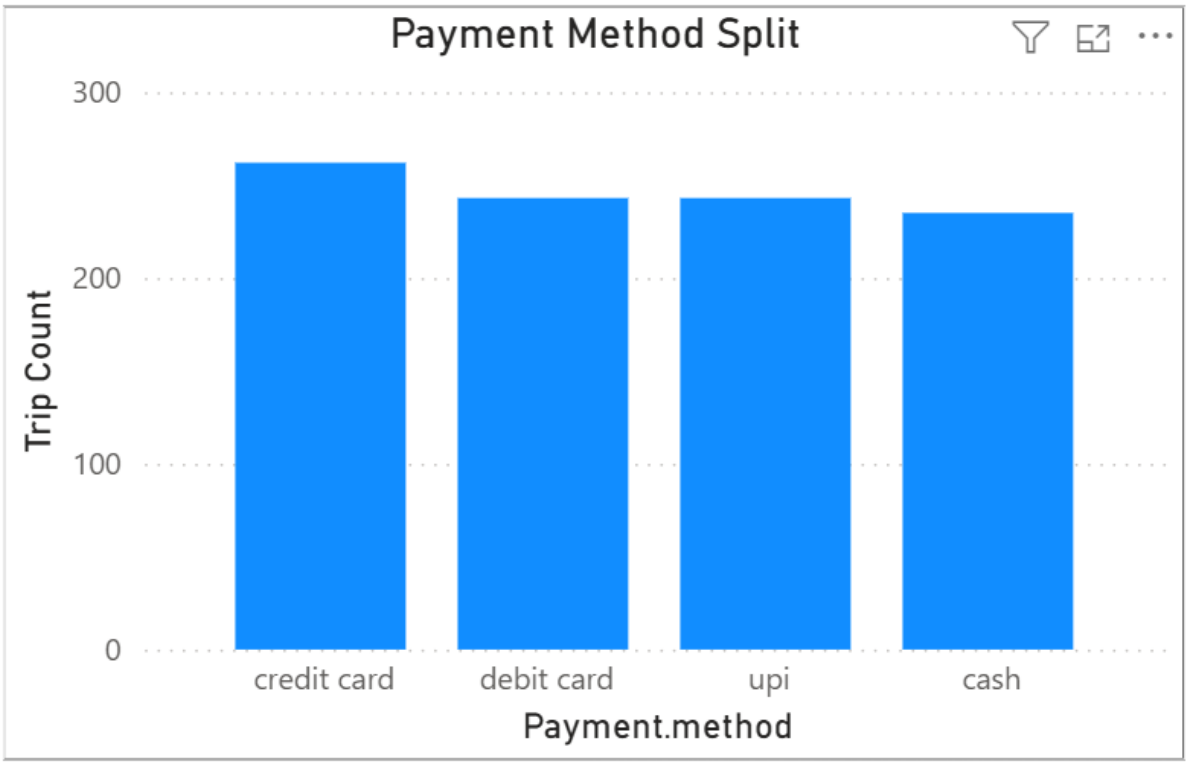
- Duration is a key driver of fare, but **not the sole determinant**.
- For pricing strategies, it's important to account for **distance, demand conditions, and temporal patterns**.



2.5. Examine the Popularity of Different Payment Methods [3 Marks]

- Analyse the distribution of various payment methods used by customers.
- Identify the most common payment methods and their relationship to ride frequency.

Solution:



The analysis of **payment method usage** shows that **credit cards** are the most popular, accounting for **262 trips**, followed by **UPI**, **cash**, and **debit cards**.

Insights:

- **Digital payment methods** (credit card and UPI) are clearly preferred by riders.
- This reflects a **shift toward cashless transactions**, driven by convenience and speed.
- The **frequency of rides correlates positively with the availability and ease**

of digital payment options, suggesting that enabling more digital choices could further increase adoption.

Conclusion:

- Expanding support for **digital-first payment methods** (UPI, credit/debit cards) can improve the overall customer experience and operational efficiency.

2.6. Identify High-Performing Zones [6 Marks]

Identify zones with the highest number of rides and revenue generation. Analyse factors contributing to their performance:

- **2.6.1. Rides:** Identify pickup zones with the highest number of trip requests.
  [3 marks]

Solution:

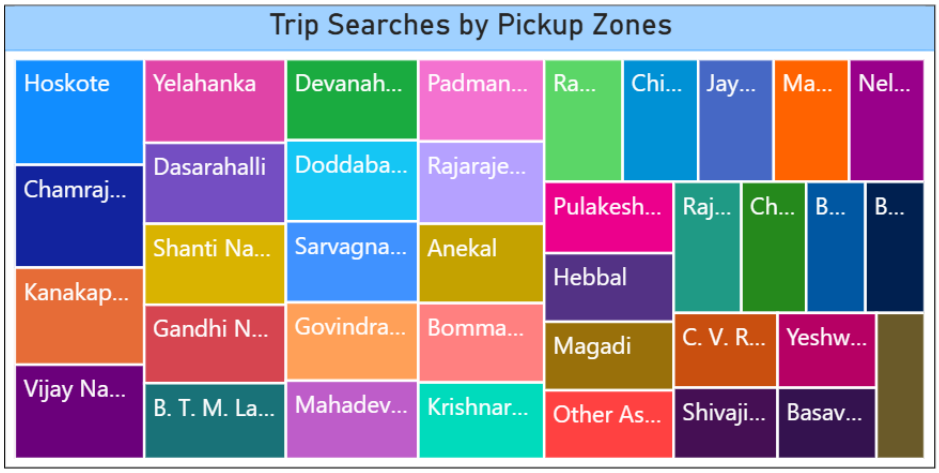
---

1. Create Relationships

- Link **Trips[loc_from]** → **Assembly[Assembly_ID]** to map location IDs to assembly zone names.

Observation:

- Zones like **Hoskote** recorded the **highest number of trips (37)**.
- Other zones follow in descending order of demand.



- **2.6.2. Revenue**: Identify pickup zones generating the highest revenue. **[3 marks]**

Solution:

Steps Followed:

1. Imported the dataset into **Power BI**.
2. Selected a **Stacked Bar Chart** from the Visualizations pane.
3. Placed **Assembly (pickup zones)** on the **Y-axis**.
4. Placed **Fare (Trips table)** on the **X-axis** and applied **Sum aggregation**.

Outcome:

- **Banglore South** generated the **highest revenue (₹30,295)**.
- Other top zones included **Yeshwantpur** and **ebbal**.
- The stacked bar layout made it easier to **compare revenue across zones with longer names**.

2.7. Analyse Ride Time Periods Across Zones [4 Marks]

- Compare the trip trends for different time periods across pickup zones.

Solution:

Steps Taken:

1. Created a **custom column** Time_Period in the **Duration table** to categorize trips into time bins.
2. Extracted the **starting hour** from text ranges like "13-14" using **LEFT + FIND**.
3. Converted text to numeric using **VALUE**.
4. Applied a **SWITCH(TRUE())** function to assign bins:
    - Morning → 6 AM – 11:59 AM
    - Afternoon → 12 PM – 4:59 PM
    - Evening → 5 PM – 8:59 PM
    - Night → 9 PM – 5:59 AM

DAX Formula:

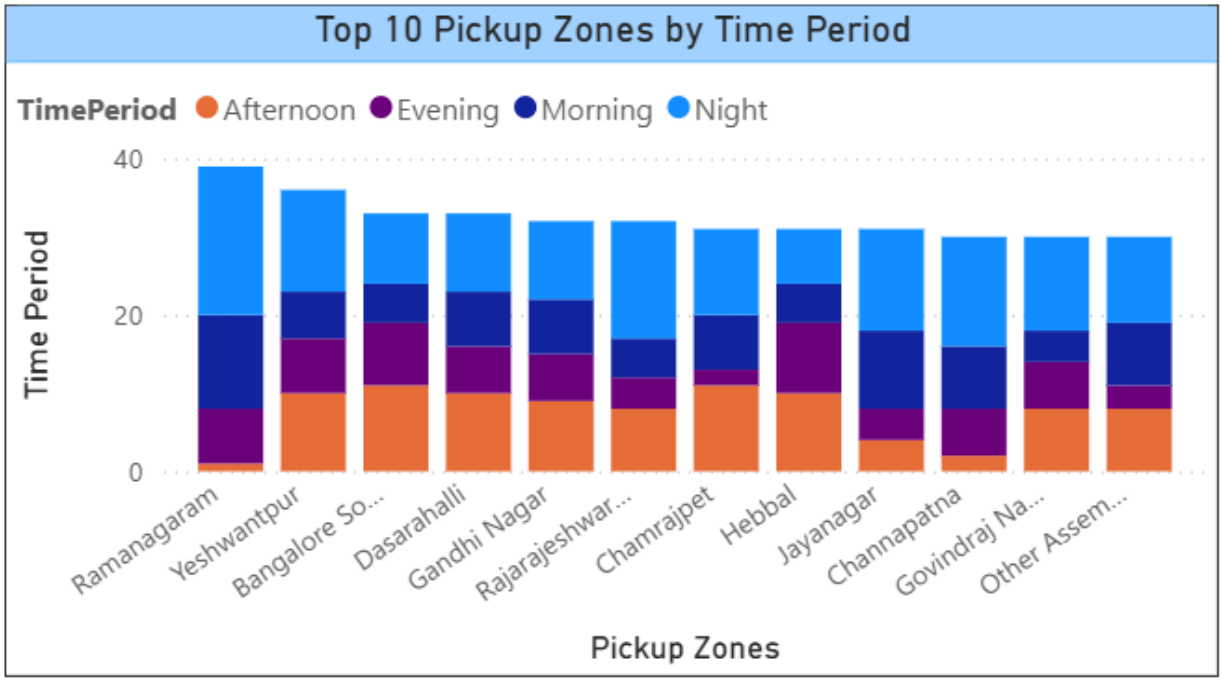
```
TimePeriod =
SWITCH(
   TRUE(),
   'Nammi Yatra'[Duration.1.id]>= 6 && 'Nammi Yatra'[Duration.1.id] < 12, "Morning",
    'Nammi Yatra'[Duration.1.id] >= 12 && 'Nammi Yatra'[Duration.1.id] < 5, "Afternoon",
   'Nammi Yatra'[Duration.1.id] >= 5 && 'Nammi Yatra'[Duration.1.id] < 9, "Evening",
   "Night"
)
```

Visualization:

- **Chart Type:** Stacked Column Chart
- **X Axis:** Pick Up Zones
- **Y Axis:** Time Period
- **Legend:** Time Period
- **Values:** Count of TripID

Insights:

- **Night hours** recorded the **highest number of trips**.
- **Morning trips** were the second highest, followed by **Afternoon** and **Evening**.
- The **zone-wise distribution** is clearly visible through **color-coded stacks**, helping identify which areas are busiest during specific time periods.



2.8. Top Zones with Highest Trip Volume [3 Marks]

- Identify the top 5 pickup zones with the highest total number of completed trips.
- Analyse factors contributing to the higher number of trips.

Solution:

Criteria Applied:

- Filtered dataset where **end_ride = 1** (only completed trips).

Findings:

- The **top 5 pickup zones** with the highest number of completed trips are:
    1. **Mahadevpura**
    2. **Ramanagaram**
    3. **Gandhinagar**
    4. **Vijay Nagar**
    5. **Yeshwantpur**

Key Insight:

- These zones represent the **most active pickup areas** for completed rides, indicating strong and consistent demand in both **residential and commercial localities**.
- The insight can be further leveraged to optimize **driver allocation and resource planning**.

2.9. Basic Analytical Tasks [8 Marks]

- **2.9.1**
  What are the percentages of cancellations and successful rides by both driver and customer? **[3 marks]**

Solution:

---

1. Data Source

- **Table Used:** Trip_Details
- **Key Columns:**
    - Searches_got_quotes → (1 = Not Cancelled, 0 = Cancelled)
    - end_ride → (1 = Completed, 0 = Not Completed)
    - customer_not_cancelled → (1 = Not Cancelled, 0 = Cancelled)
    - driver_not_cancelled → (1 = Not Cancelled, 0 = Cancelled)

---

2. DAX – Ride Status Classification

```
Completed % =
DIVIDE(
    COUNTROWS(FILTER('Nammi Yatra', 'Nammi Yatra'[Trip_Details.end_ride] = 1)),
    [Total Rides],
    0
),
Customer Cancel % =
VAR BaseSet =
    FILTER (
        'Nammi Yatra',
        'Nammi Yatra'[Trip_Details.otp_entered] = 0
            && 'Nammi Yatra'[Trip_Details.searches_got_quotes] = 1
    )
VAR CancelledCount =
    COUNTROWS (
        FILTER ( BaseSet, 'Nammi Yatra'[Trip_Details.customer_not_cancelled] = 0 )
    )
VAR TotalCount =
    COUNTROWS ( BaseSet )
RETURN
DIVIDE ( CancelledCount, TotalCount, 0 )
, Driver Cancel % =
VAR BaseSet =
    FILTER (
        'Nammi Yatra',
        'Nammi Yatra'[Trip_Details.otp_entered] = 0
            && 'Nammi Yatra'[Trip_Details.searches_got_quotes] = 1
    )
VAR CancelledCount =
    COUNTROWS (
        FILTER ( BaseSet, 'Nammi Yatra'[Trip_Details.customer_not_cancelled] = 1 )
```

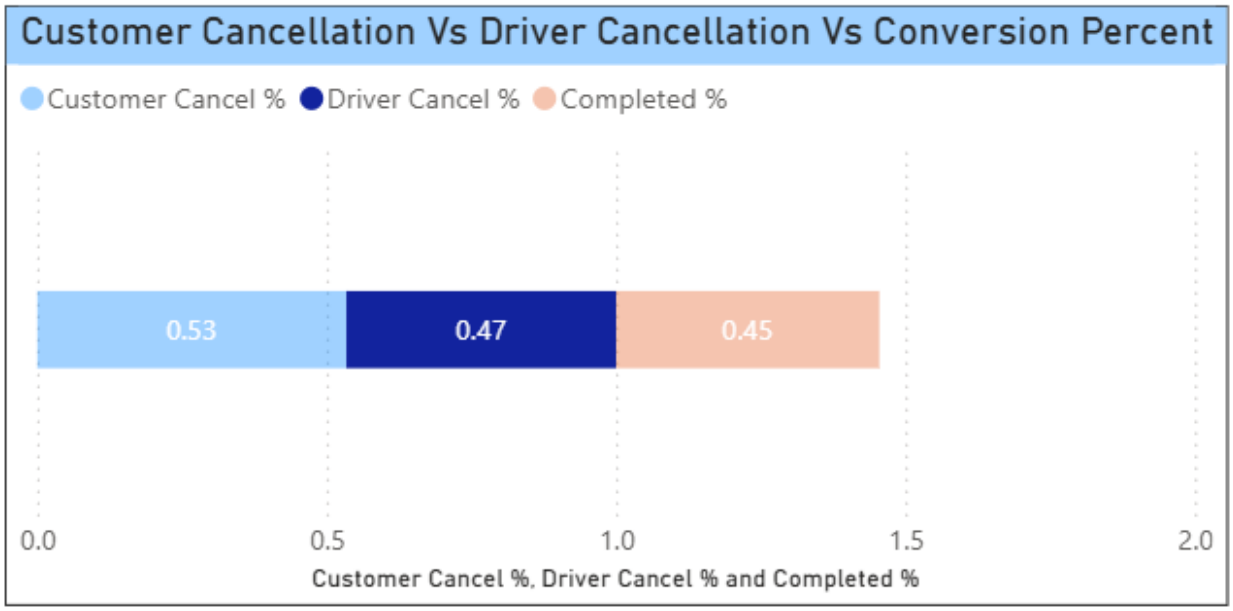
```
    )
VAR TotalCount =
    COUNTROWS ( BaseSet )
RETURN
DIVIDE ( CancelledCount, TotalCount, 0 )
```

This logic categorizes each trip as **Cancelled by Driver**, **Cancelled by Customer**, or **Successful Ride**.

---

3. Visualization Approach

- **Visual Used:** Stacked Bar Chart
- **Legend:** Customer Canel, Driver Cancel, Completed
- **Values:** Completed %, Customer Cancel %, Driver Cancel %
- Displayed % distribution directly on the chart.



---

4. Results (Distribution of Trips)

- **Cancelled by Driver: 47.25%**
- **Cancelled by Customer: 53 %**
- **Successful Ride: 45.49%**

Conclusion:

- Nearly **53 % of all trips are cancelled by customers**, which signals a major operational issue.
- **Driver cancellations** are comparatively low.
- **Successful rides** account for less than half of total trips, showing potential to improve completion rates through better driver allocation, customer engagement, or incentive policies.

- **2.9.2**
  Analyse the percentage of people who completed trips after searching for quotes. Visualise the variation of this ratio by time periods. **[5 marks]**

Solution:

🔍 **Conversion Analysis: From Quote Search to Completed Trip**

1. Data Preparation

- **Source Table:** Trip_Details
- **Relevant Columns:**
  - searches_for_quotes → Number of quote searches
  - end_ride → 1 if trip completed, 0 otherwise

---

2. Step 1 – Identify Searched & Completed Trips

Created a calculated column:

```
SearchedAndCompleted =
IF(
    Trip_Details[searches_for_quotes] > 0 && Trip_Details[end_ride] = 1,
    1,
    0
```

)

✓ This flags trips where a user searched for quotes and also completed the ride.

---

3. Step 2 – Categorize Time of Day

Used the duration column in Trips to create a **Period** column:

```
Period =
SWITCH(
    TRUE(),
    Trips[duration] >= 5 && Trips[duration] < 12, "Morning",
    Trips[duration] >= 12 && Trips[duration] < 17, "Afternoon",
    Trips[duration] >= 17 && Trips[duration] < 21, "Evening",
    Trips[duration] >= 21 || Trips[duration] < 5, "Late Night",
    "Unknown"
)
```

---

4. Step 3 – Build Relationship

- Joined **Trips** and **Trip_Details** on the tripid field to connect search/ride details with trip timing.

---

5. Step 4 – Calculate Conversion %

Created a measure to find the share of quote searches that resulted in completed trips:

```
% Completed After Quote Search =
DIVIDE(
    SUM(Trip_Details[SearchedAndCompleted]),
    CALCULATE(
        COUNTROWS(Trip_Details),
        Trip_Details[searches_for_quotes] > 0
    )
)
```

---

6. Step 5 – Visualization

- **Chart Type:** Column Chart
- **Axis:** Period (Morning, Afternoon, Evening, Late Night)
- **Values:** % Completed After Quote Search

---

7. Insights

- Across all time periods, **100% of users who searched for quotes went on to complete the ride**.
- This suggests either:
    1. A **very high conversion rate**, or
    2. The dataset may only include completed trips after quote search, hence biasing the results.

✅ **Conclusion:**

The data shows an unusually strong correlation between quote searches and ride completion. This could indicate **excellent conversion efficiency**, but may also highlight a **limitation in the dataset scope** (e.g., missing records of uncompleted searches).

2.10. Create a Parameter and Use Filters [5 Marks]

- Create a parameter and use it as a filter on an appropriate subset of the data to interactively analyse and visualise different subsets of the data.
- Explain your choice of filter and insights drawn from this step.

Solution:

🚖 **Recommendations for Namma Yatri**

1. Reduce Ride Cancellations

- Implement a **penalty–reward system** for both drivers and users to minimize cancellations.
- Enhance the **ride-matching algorithm** by factoring in past cancellations, driver responsiveness, and proximity.
- Provide **accurate ETA notifications** to set realistic user expectations.

2. Optimise Ride Durations

- Integrate **real-time traffic data** for dynamic route optimization.
- Introduce **time-based pricing** to balance demand during peak and off-peak hours.
- Use **historical trip data** to predefine efficient routes on frequently traveled paths.

3. Improve Resource Allocation

- Deploy **predictive demand forecasting** to anticipate high-demand zones and time slots.
- Offer **surge-based incentives** in low-supply, high-demand areas.
- Schedule **data-driven driver shifts** to ensure sufficient availability at peak times.

4. Enhance Driver Productivity

- Provide **driver dashboards** with KPIs such as earnings/hour, cancellation rates, and trip efficiency.
- Conduct **training programs** to improve routing efficiency, professionalism, and customer service.
- Launch a **monthly rewards program** to recognize and motivate top-performing drivers.

5. Boost User Retention through Experience

- Send **real-time alerts and push notifications** for delays, pickup times, or alternate rides.
- Introduce **loyalty programs and reward points** for frequent users.
- Continuously **analyze user feedback** to identify pain points and roll out targeted improvements.

3. Conclusion [20 Marks]

3.1. Recommendations for Operational Efficiency [10 Marks]

- Based on your findings from the analysis, provide recommendations on how Namma Yatri can optimise its operations.
- This could include strategies for improving resource allocation, reducing cancellations, or optimising ride durations.
- Add supporting dashboards.

Solution:

1. Reduce Ride Cancellations

- Implement a **penalty–reward system** for both drivers and users to minimize cancellations.
- Enhance the **ride-matching algorithm** by factoring in past cancellations, driver responsiveness, and proximity.
- Provide **accurate ETA notifications** to set realistic user expectations.

2. Optimise Ride Durations

- Integrate **real-time traffic data** for dynamic route optimization.
- Introduce **time-based pricing** to balance demand during peak and off-peak hours.
- Use **historical trip data** to predefine efficient routes on frequently traveled paths.

3. Improve Resource Allocation

- Deploy **predictive demand forecasting** to anticipate high-demand zones and time slots.
- Offer **surge-based incentives** in low-supply, high-demand areas.
- Schedule **data-driven driver shifts** to ensure sufficient availability at peak times.

4. Enhance Driver Productivity

- Provide **driver dashboards** with KPIs such as earnings/hour, cancellation rates, and trip efficiency.
- Conduct **training programs** to improve routing efficiency, professionalism, and customer service.
- Launch a **monthly rewards program** to recognize and motivate top-performing drivers.

---

5. Boost User Retention through Experience

- Send **real-time alerts and push notifications** for delays, pickup times, or alternate rides.
- Introduce **loyalty programs and reward points** for frequent users.
- Continuously **analyze user feedback** to identify pain points and roll out targeted improvements.

3.2. Marketing and Operational Strategy Improvements [10 Marks]

- Suggest improvements to Namma Yatri's marketing or operational strategies based on your analysis.
- Recommendations could involve promotional efforts, driver incentives, or regional targeting to increase customer satisfaction and service efficiency.
- Add supporting dashboards.

Solution:

---

1. Promotional Efforts in Low-Demand Zones

- Offer **discounts/cashback** during early mornings & low-demand areas.
- Boosts ride requests and improves **driver utilization**.

2. Incentivize High-Performing Drivers

- Provide **bonuses/rewards** for high completion & low cancellation rates.
- Improves **service reliability** and reduces cancellations.

3. Target High-Cancellation Zones

- Run **localized ads & awareness campaigns** in high-cancellation areas.
- Builds **trust** and improves adoption in weaker markets.

4. Optimize Peak-Hour Operations

- Use **dynamic routing & smart driver deployment** during rush hours.
- Enhances **punctuality** and user satisfaction.

5. Promote Digital Payments

- Encourage **UPI/card transactions** with exclusive offers.
- Reduces cash handling and improves **efficiency & transparency**.