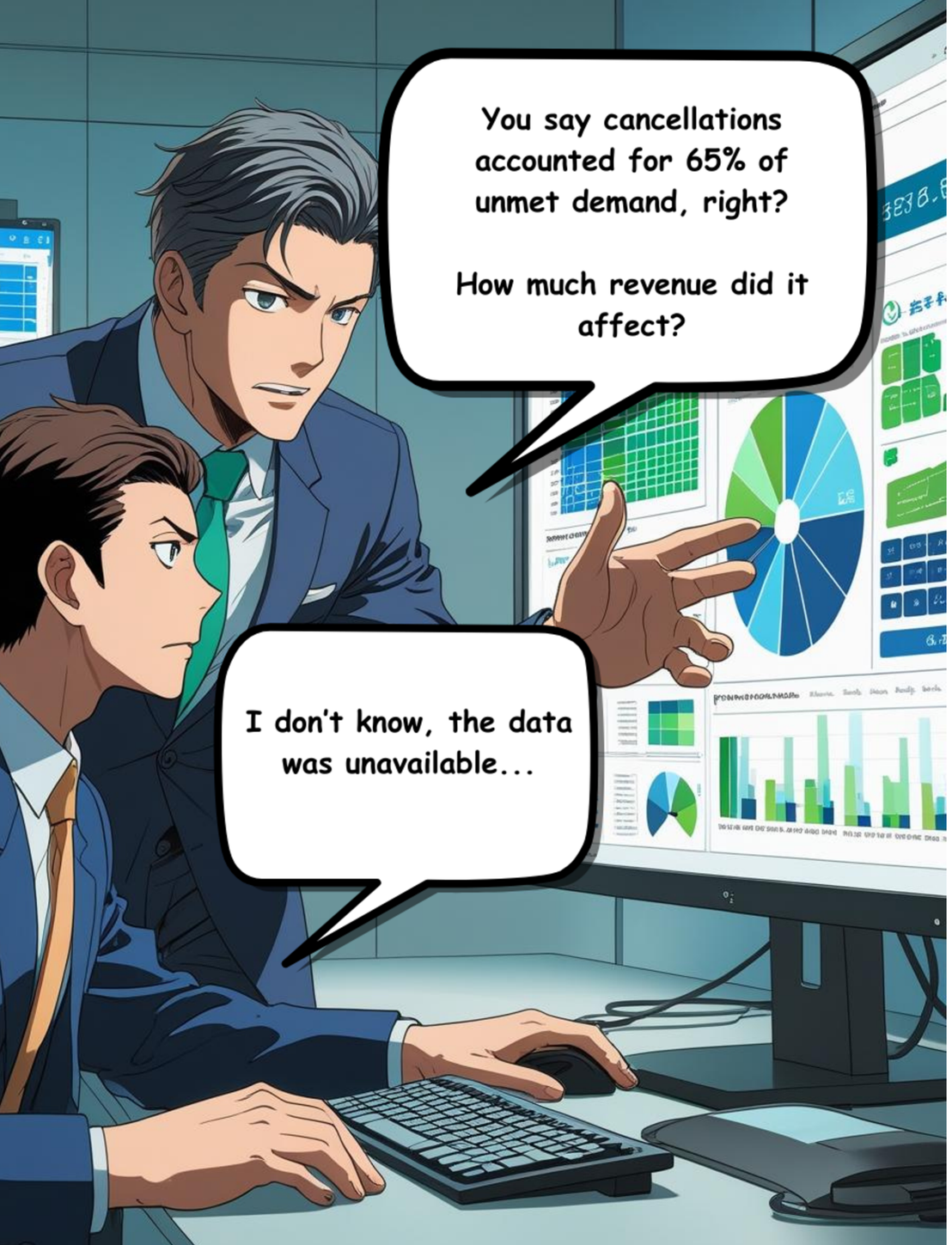


Missing data to Decision-making

- **In ride-hailing, every cancellation is lost revenue.**
- **Estimating that loss sounds simple: sum up cancelled booking values. But what if those values aren't recorded?**

The solution? Statistics. Enter Monte-Carlo Simulation!

Project Plan

Problem Statement Formulation:

Unmet demand in ride-hailing leads to significant revenue loss, but the challenge lies in understanding where this loss comes from and how much of it can realistically be recovered. The key questions were:

- How much revenue is lost due to unmet demand?
- Which segments contribute the most to this loss?
- Where should interventions be prioritized for maximum impact?

Resources

Data Source: Link

Tool and Techniques

Python: pandas, numpy, matplotlib, seaborn, plotly, SciPy

Statistics: Hypothesis testing (Chi-square analysis, t-test), Monte-Carlo Simulations (global and stratified)

Versioning: Git

Platforms: VS Code, Bricks (insights dashboard)

Approach

1. **Select Granularity –** Chi-Square tests to determine the level of granularity where variability is highest, ensuring deeper insights.
2. **Quantify Loss –** Estimate total revenue loss and break it down by segment.
3. **Identify Key Contributors –** Measure each segment’s share of loss and flag primary contributors (≥50% threshold).
4. **Drill Down into Hotspots –** Segment contributors further and identify high-impact hotspots for immediate intervention.
5. **Estimate Recovery Potential –** Calculate recoverable loss from interventions and project daily/hourly revenue uplift**.**

The Data

Date : Date of the booking

Time : Time of the booking

Booking ID: Unique identifier for each ride booking

Booking Status: Status of booking (Completed, Cancelled by Customer, Cancelled by Driver, etc.

Customer ID: Unique identifier for customers

Vehicle Type: Type of vehicle (Go Mini, Go Sedan, Auto, eBike/Bike, UberXL, Premier Sedan)

Pickup Location: Starting location of the ride

Drop Location: Destination location of the ride

Avg VTAT: Average Vehicle Time at Arrival

Avg CTAT: Average Customer Time at Arrival

Booking Value: Total fare amount for the ride

Ride Distance: Distance covered during the ride (in km)

Challenge: Imputing missing Revenue data

Problem

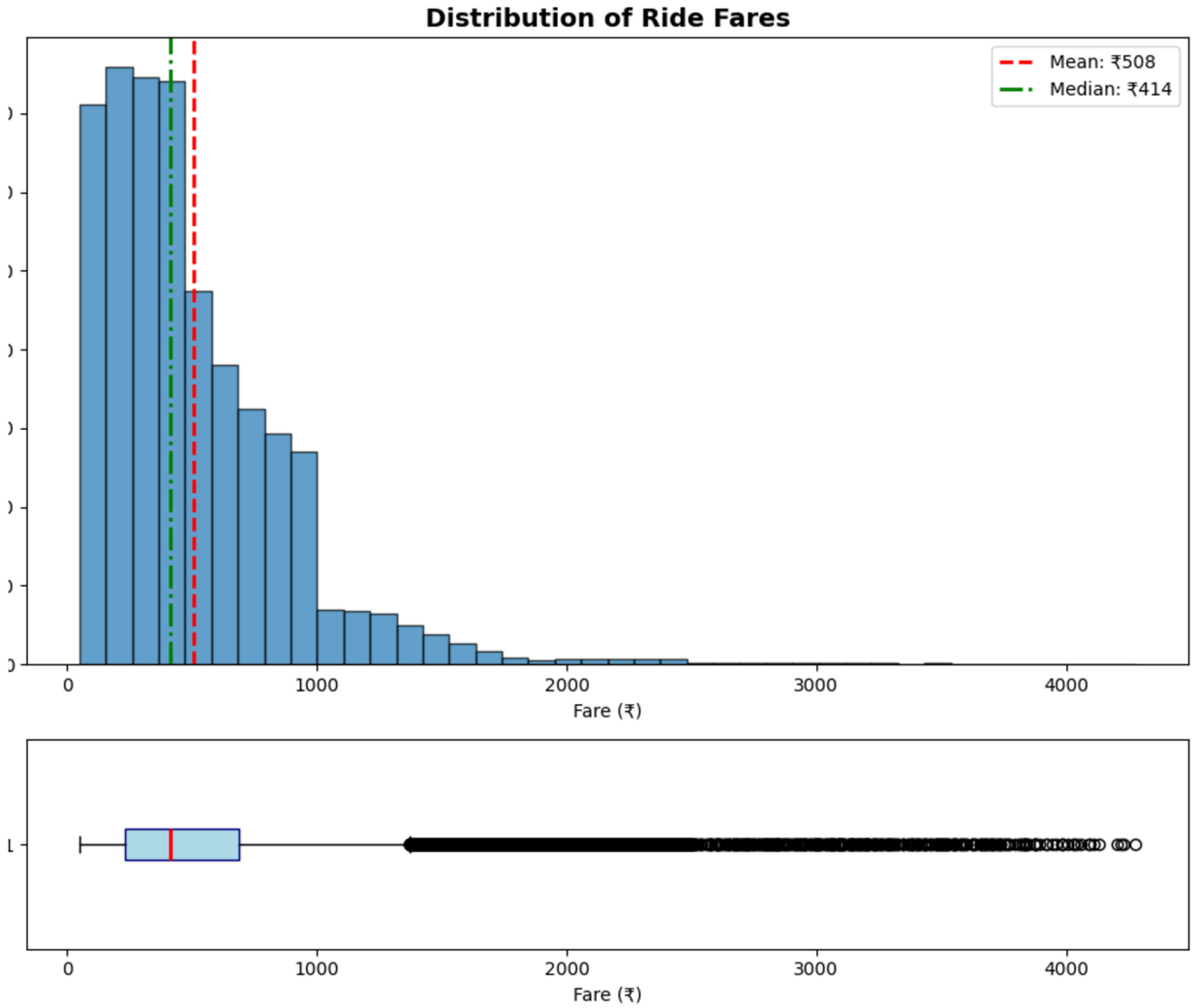
- Dataset lacked ride values for cancellations & driver unavailability (null).
- Imputation required to estimate revenue loss.

Challenge

- Ride fares are right-skewed.
- **Mean imputation** → overestimates loss.
- M**edian imputation** → underestimates loss.

Approach

- Standard imputation not suitable.
- Required method to account for skewed distribution.
- Solution space: **Statistical Estimation & Predictive Modeling**.



Predictive modeling via Regression won’t work.

The Pricing Challenge

- Ride pricing is dynamic → depends on demand, supply, traffic, weather, events, distance, etc.
- Dataset lacked these features → replicating pricing model not feasible.

Experiment

- Tried regression (LightGBM + Bayesian tuning).
- Result: $R^2 = 0.05$ → dataset variables explained almost no variance in pricing (expected).

Workaround

- Adopted Monte Carlo Simulation.
- Idea: use the observed distribution of Completed rides as proxy.

Plan

1. Run 100,000 simulations sampling from completed ride fare distribution.
2. Validate against Incomplete rides (where booking value was available).
3. Once validated, extend to Cancelled rides to estimate potential revenue loss.

Monte Carlo Simulation

- **What it is**
  - A statistical method that uses random sampling to estimate outcomes when exact values are unknown.

- **How it works**
  a. Start with a known distribution (from data or assumptions).
  b. Generate a large number of random samples (e.g., 100,000).
  c. Aggregate results to approximate the range and probability of outcomes.

- **Why it’s useful here**
  - Handles uncertainty and missing data.
  - Captures not just averages, but the entire spread of possible values.
  - Widely used in finance, risk analysis, and demand forecasting.

Result validation after Simulation runs

Total potential revenue= INR 76M
Revenue lost due to unsuccesful rides= INR 29M (based on global fare distribution)

Results of MC simulations:

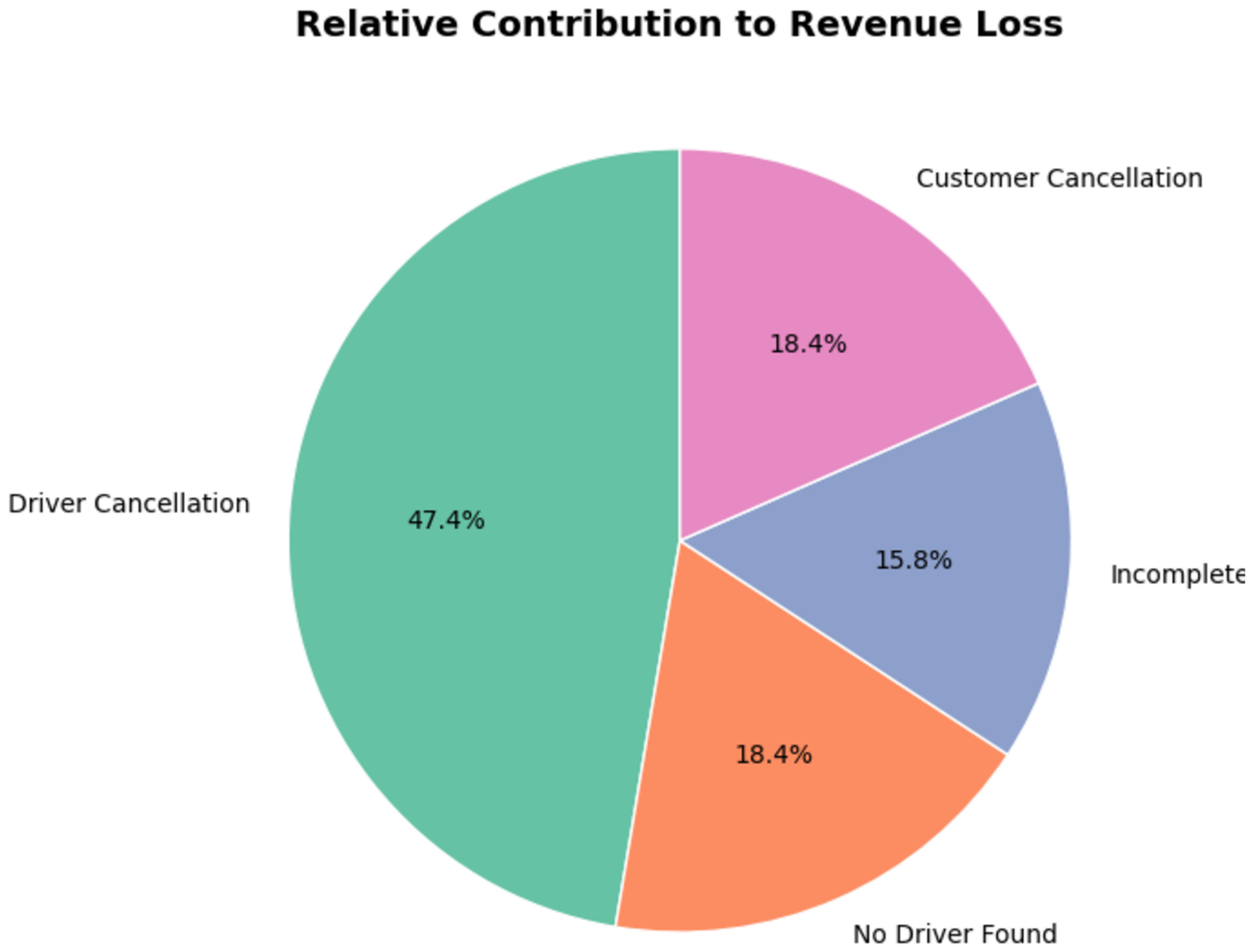
Driver of Loss	Unmet Rides	Mean Loss (₹)	CI 2.5% (₹)	CI 97.5% (₹)	Share of Total (%)
Driver Cancellation	27,000	13,720,880	13,593,100	13,848,640	47.37
Customer Cancellation	10,500	5,335,776	5,255,616	5,414,611	18.42
No Driver Found	10,500	5,335,749	5,256,699	5,415,756	18.42
Incomplete	9,000	4,573,706	4,500,255	4,648,150	15.79
TOTAL	57,000	28,966,110	28,784,100	29,153,580	100.00

Revenue lost due to ‘Incomplete’ rides = INR 4,585, 609 (Ground Truth)

Comparing to the ground truth value for ‘incomplete rides’, it is evident that the true mean falls within the 95% Confidence Interval.

As a result, we can safely use the results of the simulation as proxy to estimate the Revenue Loss due to cancellations and unsuccessful rides.

As evident, ~65% of the estimated revenue loss is due to cancellations.



Where to take immediate action ?

We first quantified the contribution of each segment to total revenue loss. Naturally, **not all issues can be addressed simultaneously, and businesses often prioritize interventions based either on ease of implementation or on the magnitude of financial impact.**

While cancellations emerged as the single largest driver of loss, it is equally possible that, from a strategic perspective, the company may consider addressing supply-side gaps (e.g., “No Driver Found”) by increasing fleet size if that is the more immediately viable option.

However, since such strategic decisions lie outside the scope of this analysis, the focus of this project was narrowed to **Cancellations**, enabling a deeper exploration of their patterns and potential recovery opportunities.

From Revenue Leakage to Targeted Analysis

Having identified **cancellations** as the primary channel of revenue leakage, the next step was to design an analysis funnel to systematically localize and quantify loss drivers.

Segmentation Strategy

- Each cancellation type (driver-led, customer-led) was segmented along two key dimensions:
  a. Hour of Day – to capture temporal patterns in cancellations.
  b. Vehicle Type – to capture operational differences across fleet categories.

- For each segment, three metrics were calculated:
  - Unmet Count (volume of cancellations)
  - Estimated Revenue Loss (stratified Monte Carlo Simulation)
  - Loss Share (relative contribution to total loss)

Pareto-Based Narrowing

- Within each segmentation, cancellation hotspots were identified using a **70% contribution threshold.**
- This ensured analytical focus on the relatively small subset of conditions that accounted for the majority of revenue loss.

Cross-Segmentation (Hour × Vehicle Type)

- Hotspots were then combined at the intersection of hour of day and vehicle type, **yielding granular "cells"** where interventions could be targeted.
- This enabled a two-dimensional prioritization:
  - When cancellations were most costly.
  - Where (in terms of vehicle type) the impact was greatest.

Target Setting

- Although mathematically, full recovery from these hotspots was possible, such an assumption would be unrealistic operationally.
- A **more pragmatic target was set: 40% recovery of losses from these high-impact cells.**
- Even this level of recovery was considered ambitious yet achievable, balancing statistical optimism with business feasibility.

How well do *per segment cancellation estimates* align with the global cancellation estimate?

Validation of Segment-Level Estimates

Loss Driver	Global Estimate (₹)	Reaggregated by Vehicle (₹)	Reaggregated by Hour (₹)
Driver Cancellation	13,720,880	13,720,454	13,720,387
Customer Cancellation	5,335,776	5,335,753	5,335,001

The differences are tiny (<0.01%), proving that segment-level Monte Carlo simulations are consistent with the global estimates.

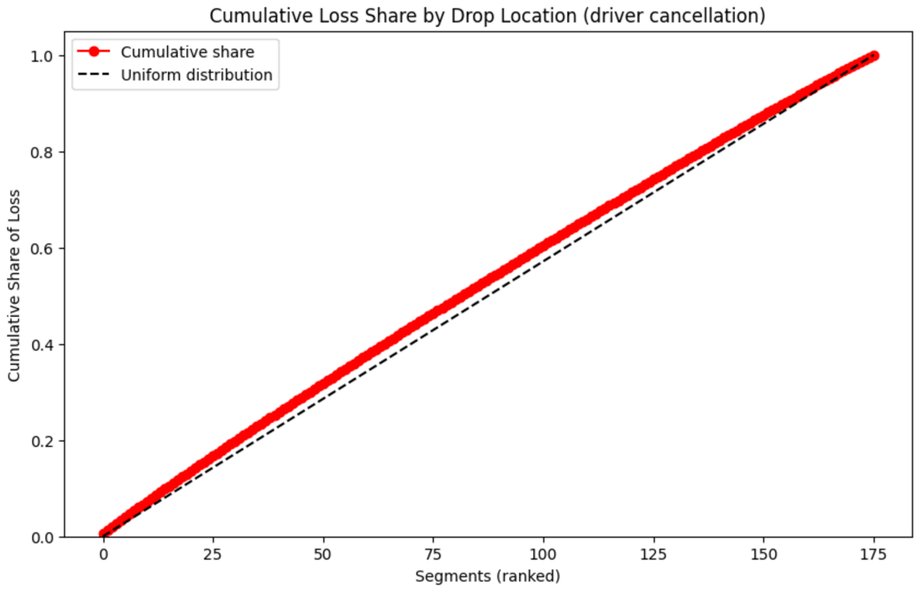
Location-based prioritization is not viable due to lack of variability in the dataset

- Observations on Location-Level Analysis
- Minimal variability across locations in booking volumes.
- Distribution of met vs. unmet demand remained nearly identical across locations.
- Prioritizing “top” locations would not yield meaningful insights, since the remaining locations still represent significant portions of demand.
- This uniformity does not reflect typical real-world patterns, suggesting the dataset may have been synthetic or selectively sampled.

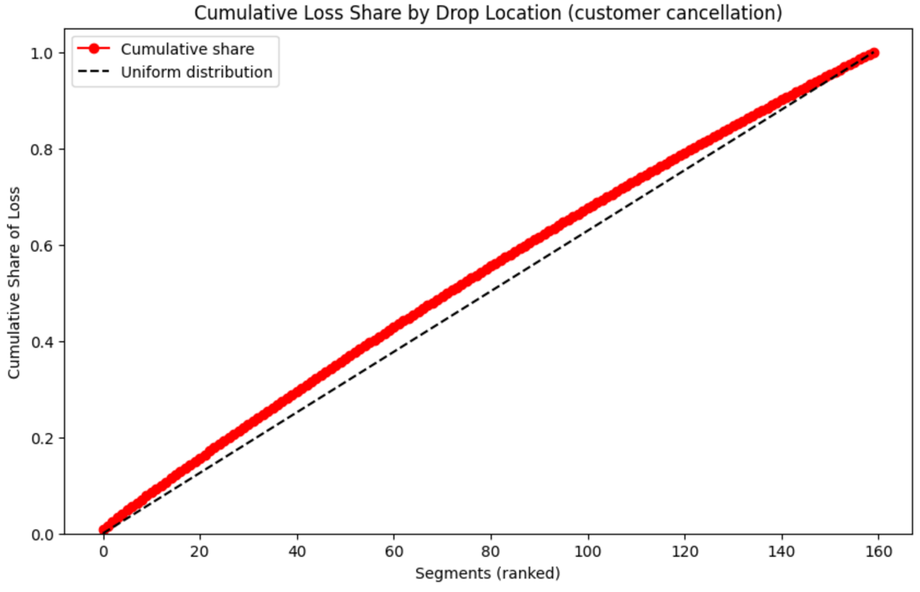
Testing the uniformity hypothesis

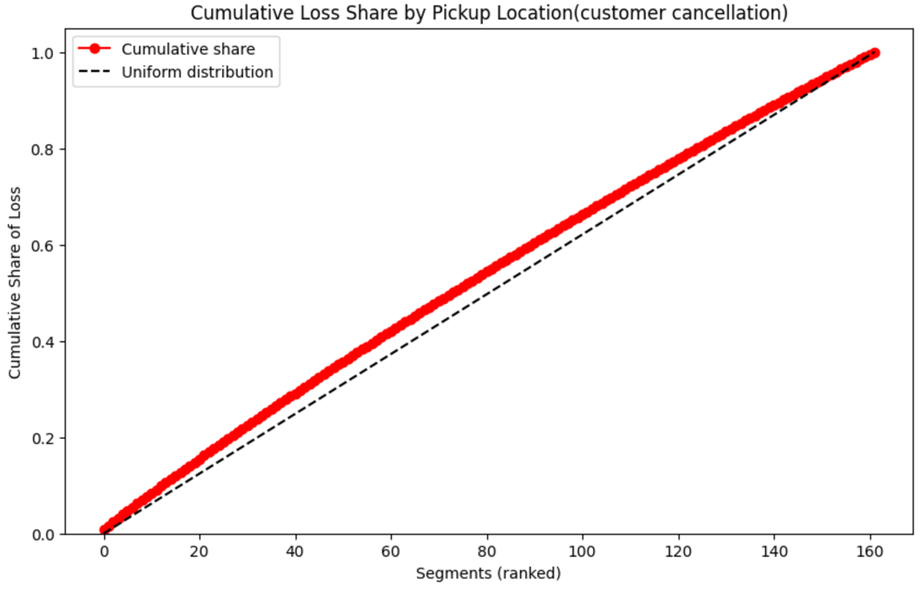
Lorenz-Style Plot of Loss Distribution

- **What it is:**
- A Lorenz-style plot shows the cumulative share of loss contributed by segments, ranked from highest to lowest.
  - X-axis → Ranked segments (e.g., pickup or drop locations)
  - Y-axis → Cumulative share of total estimated loss
- **Interpretation:**
  - If the curve rises steeply at first → a small number of segments contribute disproportionately (Pareto effect / hotspots).
  - If the curve is close to a straight diagonal → losses are spread almost uniformly across all segments.
- **Insight from this data:**
  - The plot lies very close to the diagonal, showing no dominant locations.
  - This suggests cancellations are distributed evenly across pickup and drop points, so isolating a few “top” locations would not meaningfully help target overall loss.

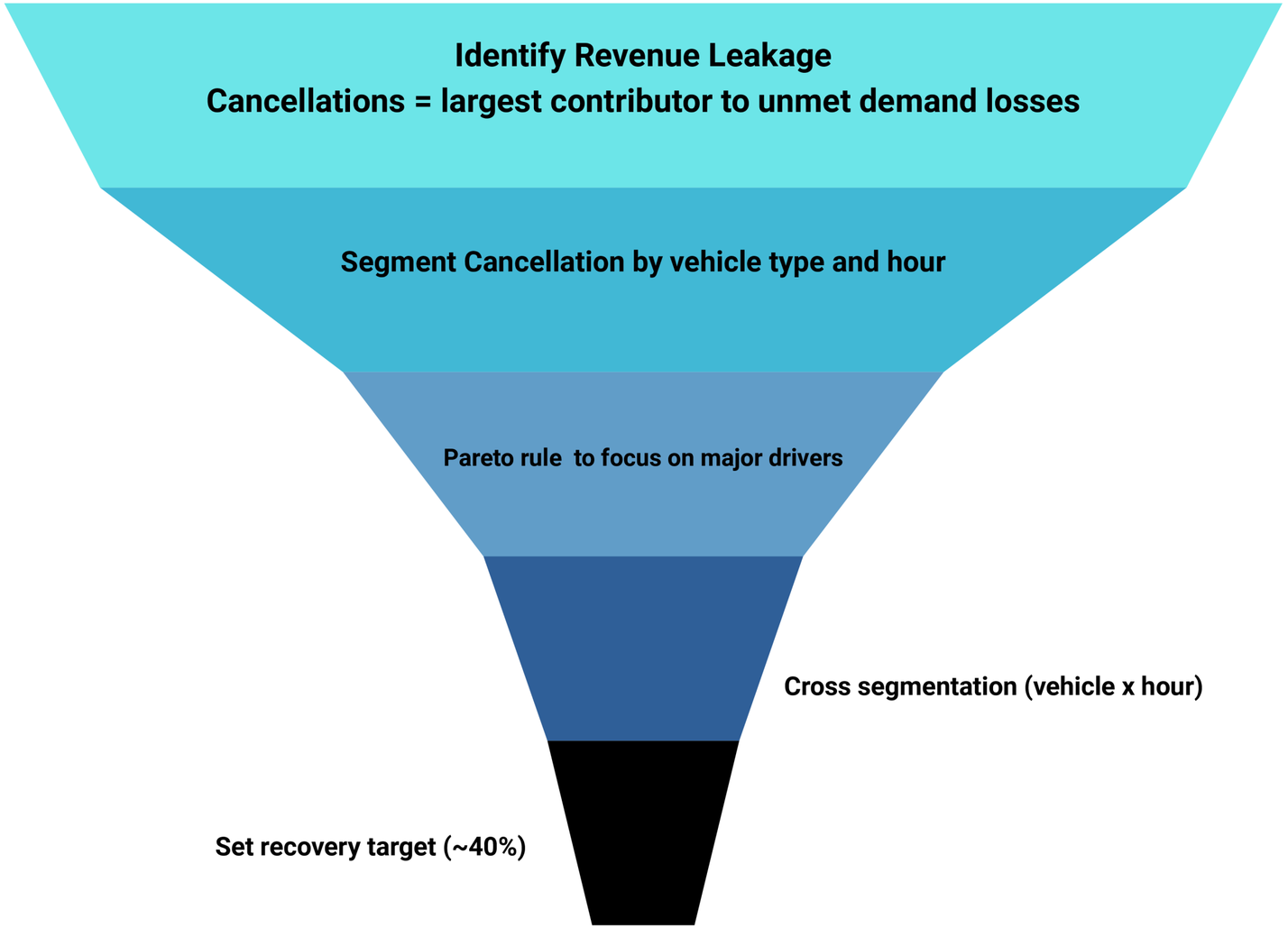

Cumulative Loss Share by Drop Location (driver cancellation)
Cumulative share
Uniform distribution
Cumulative Share of Loss
Segments (ranked)
0.0
0.2
0.4
0.6
0.8
1.0
0
25
50
75
100
125
150
175

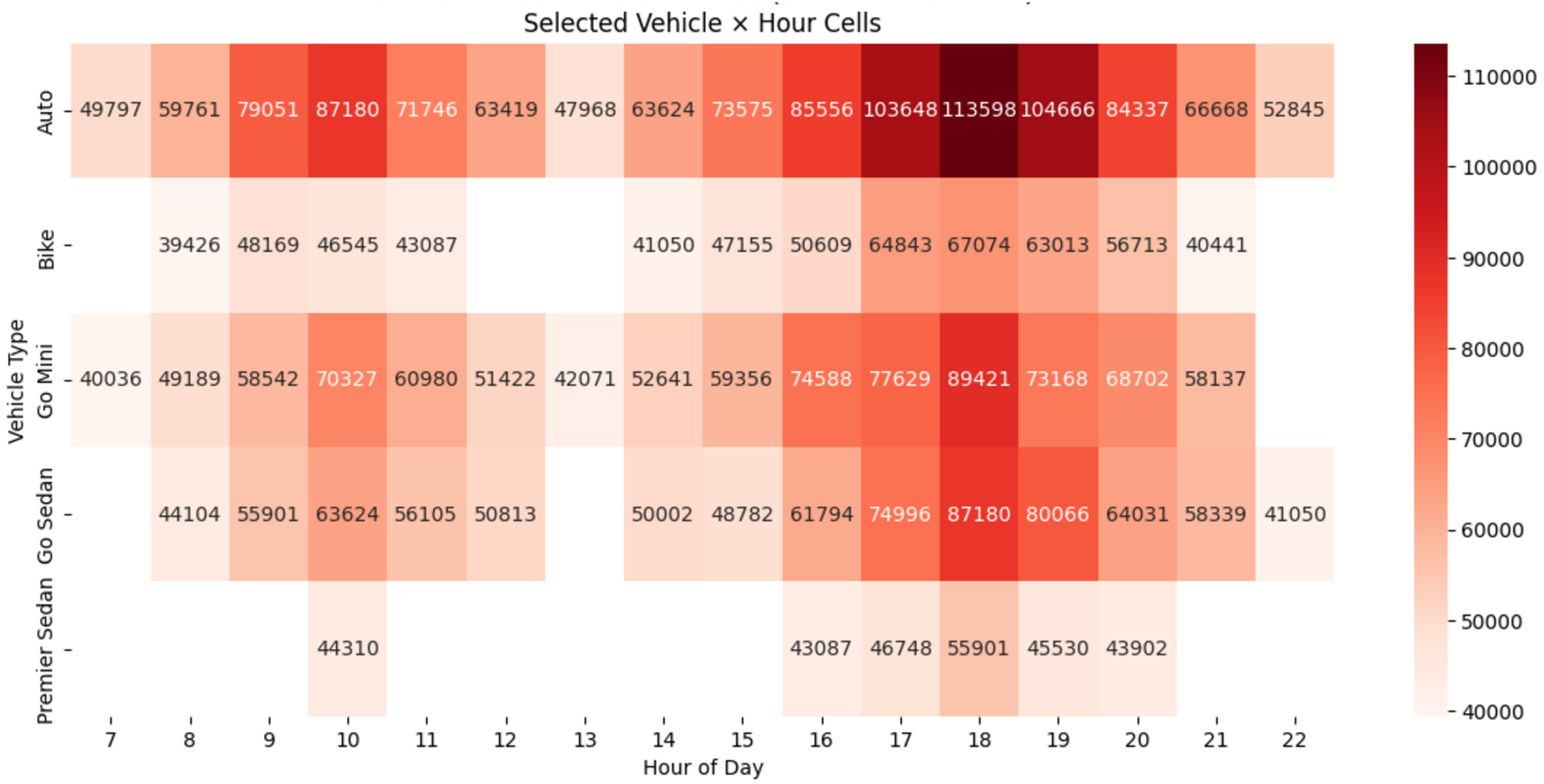

Cumulative Loss Share by Pickup Location (driver cancellation)
Cumulative share
Uniform distribution
Cumulative Share of Loss
Segments (ranked)
0.0
0.2
0.4
0.6
0.8
1.0
0
25
50
75
100
125
150
175

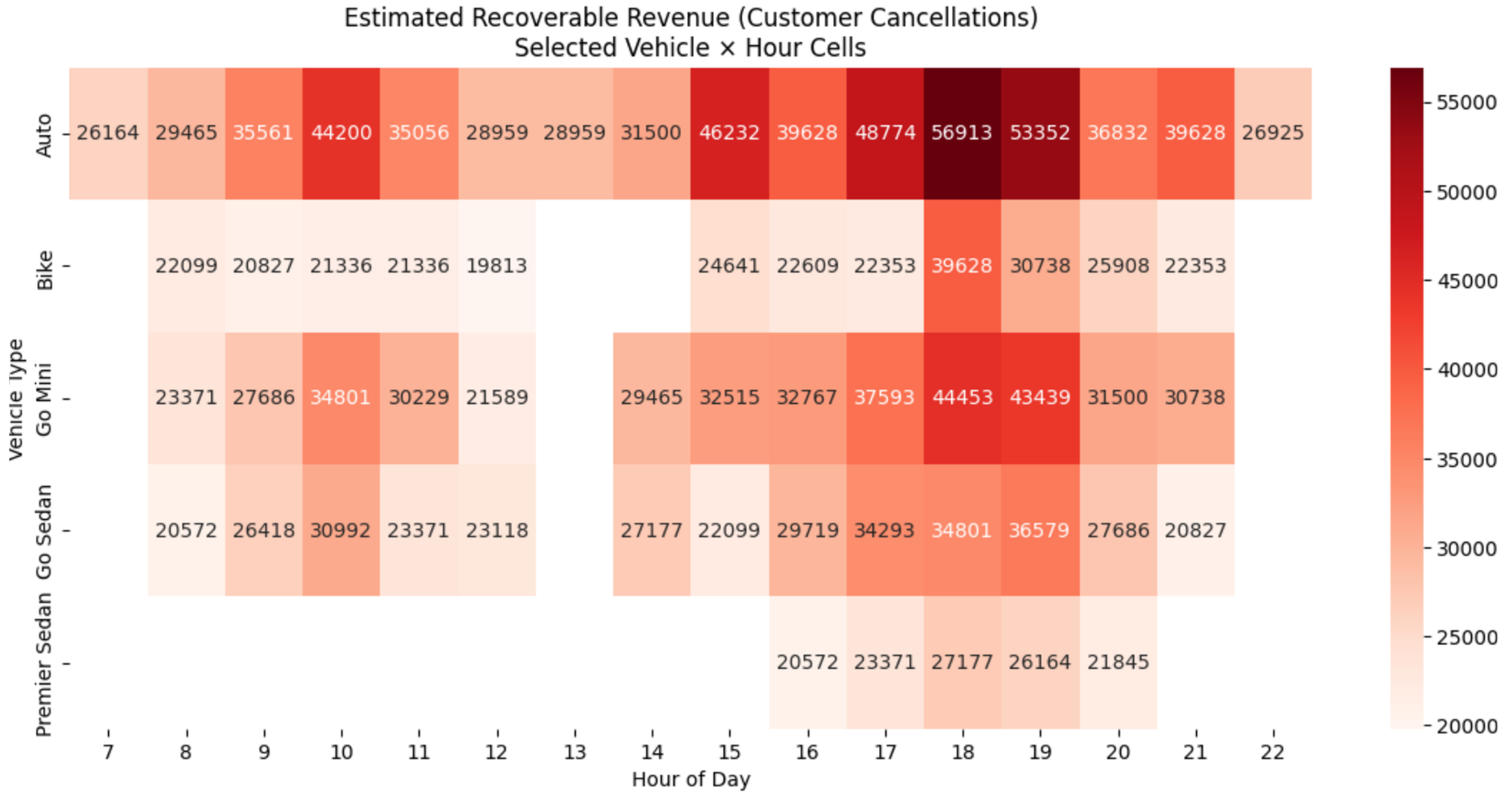

Cumulative Loss Share by Drop Location (customer cancellation)
Cumulative share
Uniform distribution
Cumulative Share of Loss
Segments (ranked)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
140
160


Cumulative Loss Share by Pickup Location(customer cancellation)
Cumulative share
Uniform distribution
Cumulative Share of Loss
Segments (ranked)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
140
160

The Analysis Funnel




Selected Vehicle × Hour Cells
Vehicle Type
Auto
Bike
Go Mini
Go Sedan
Premier Sedan
49797 59761 79051 87180 71746 63419 47968 63624 73575 85556 103648 113598 104666 84337 66668 52845
39426 48169 46545 43087 41050 47155 50609 64843 67074 63013 56713 40441
40036 49189 58542 70327 60980 51422 42071 52641 59356 74588 77629 89421 73168 68702 58137
44104 55901 63624 56105 50813 50002 48782 61794 74996 87180 80066 64031 58339 41050
44310 43087 46748 55901 45530 43902
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Hour of Day
110000
100000
90000
80000
70000
60000
50000
40000


Estimated Recoverable Revenue (Customer Cancellations)
Selected Vehicle × Hour Cells
Vehicle Type
Auto
Bike
Go Mini
Go Sedan
Premier Sedan
26164 29465 35561 44200 35056 28959 28959 31500 46232 39628 48774 56913 53352 36832 39628 26925
22099 20827 21336 21336 19813 24641 22609 22353 39628 30738 25908 22353
23371 27686 34801 30229 21589 29465 32515 32767 37593 44453 43439 31500 30738
20572 26418 30992 23371 23118 27177 22099 29719 34293 34801 36579 27686 20827
20572 23371 27177 26164 21845
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Hour of Day
55000
50000
45000
40000
35000
30000
25000
20000

Final narrowing and Impact Optimisation

1. **Hotspot Prioritization**
    - From the cells contributing to 70% of total loss, the Top 10 cells were selected.
    - These represent the most critical combinations of hour × vehicle type.

2.Recovery Potential

- Estimated recoverable amount by assuming 40% recovery from these Top 10 hotspots.
- This conservative assumption balances feasibility with meaningful impact.

3. Benchmarking Against Ground Truth

- Compared potential recovery to total revenue from completed rides.
- Provided a business context for the magnitude of recoverable loss.

4. Revenue Uplift Estimation

- Calculated daily and hourly average uplift, under the assumption that intervention achieves full success in recovering 40% from the targeted hotspots

Hotspots for Driver Cancellations

Vehicle Type	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Auto										x	x	x					x	x	x	x	x			
Go Mini																		x						
Go Sedan																		x	x					

Hotspots for Customer Cancellations

Vehicle Type	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Auto											x	x		x	x	x	x	x			x	x		
Bike																		x						
Go Mini																		x	x					

Targeted intervention at Top 10 hotspots (Hour × Vehicle Type) to recover 40% of cancellation-related losses.

Metric	Value
2024 Completed Ride Revenue	₹47,260,574
Potential Recoverable Loss (40%)	₹1,370,949
Post-Intervention Revenue	₹48,631,522 *(+3% uplift)*
Average Daily Revenue (Pre → Post)	₹129,481 → ~₹133,400 *(+3%)*
Average Hourly Revenue (Pre → Post)	₹5,395 → ~₹5,560 *(+3%)*

From missing data to decision-making

- The project began with a data gap: no booking values for cancellations and unmet rides.
- Traditional imputation methods were unsuitable due to the right-skewed nature of ride fares.
- By applying statistical estimation and Monte Carlo Simulation, we reconstructed a realistic picture of the missing values.
- Validation against incomplete rides confirmed the reliability of the simulation, turning an uncertain dataset into actionable insights.
- This approach enabled evidence-based decision making: pinpointing cancellation hotspots, setting pragmatic recovery targets, and quantifying revenue uplift.
- Ultimately, it demonstrates how statistics can transform absence of data into a foundation for strategic action.

Thank you