Maximizing Revenue for NYC Taxi Drivers

A Statistical Analysis of Payment Methods and Fare Pricing

PROJECT SCOPE:
Statistical Hypothesis Testing
Python / Pandas / Scipy
Method: Independent T-Test & EDA
BY: Deepanshu Gupta (Data Analyst)

“The Analytical Journey:
From Raw Data to Revenue Strategy

The Problem.
Driver revenue optimization.

The Data.
6.4M row NYC Taxi dataset.

The Cleanup.
Filtering noise & Feature Engineering

The Proof.
Hypothesis Testing (T-Test) & EDA.

The Strategy.
Actionable recommendations

DEEPANSHU GUPTA

In a low-margin industry, every trip counts

GOAL: Maximize revenue streams for drivers.

The Problem Statement

In the fast-paced taxi booking sector, long-term success requires data-driven insights. Drivers operate on thin margins and time efficiency.

We aim to determine if specific payment behaviors correlate with higher fare pricing.

Primary Variable: `Payment Type` vs `Fare Amount`

Two Key Questions Drive This Analysis

1. The Correlation

Is there a statistically significant relationship between the total fare amount and the payment type chosen by the passenger?

2. The Behavior

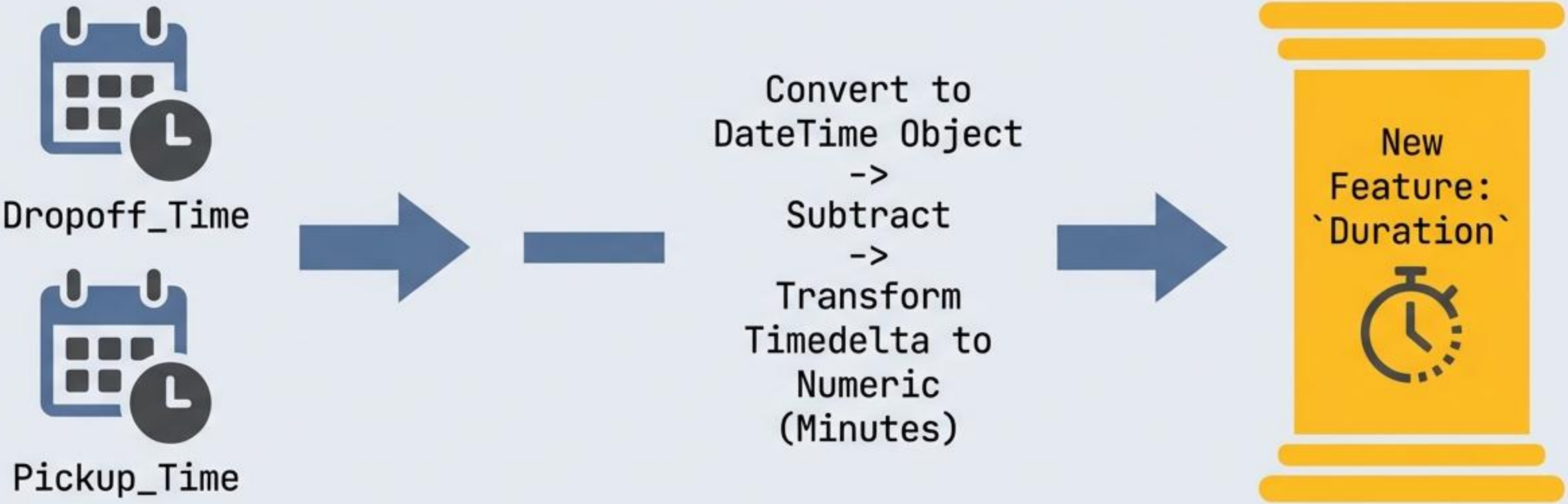
Can we nudge customers toward payment methods that generate higher revenue without negatively impacting their experience?

The Source Material: NYC Taxi Trip Records

VendorID	Pickup_DateTime	Dropoff_DateTime	Passenger_Count	Trip_Distance	Payment_Type	Fare_Amount
Creative Mobile	2023-01-01 12:00:00	2023-01-01 12:15:00	1	2.5 miles	Card	$15.50
VeriFone Inc	2023-01-01 12:30:00	2023-01-01 12:55:00	2	4.8 miles	Cash	$24.00
Creative Mobile	2023-01-01 13:10:00	2023-01-01 13:22:00	1	1.7 miles	Card	$11.50
VeriFone Inc	2023-01-01 13:45:00	2023-01-01 14:05:00	1	3.2 miles	Cash	$18.00

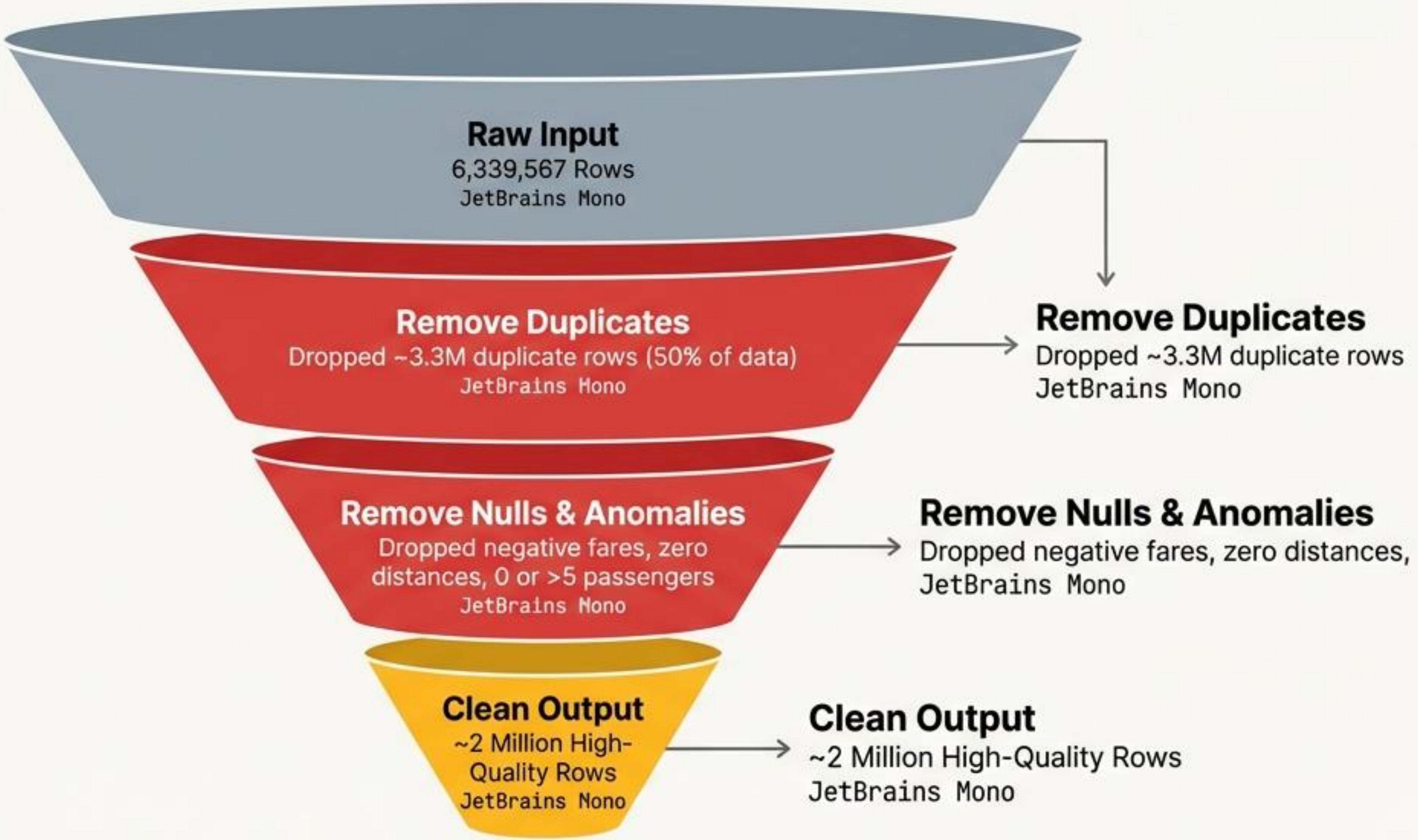
Initial Dataset Size:
~6.4 Million Rows

Engineering Value from Raw Timestamps



Insight: This derived feature allows us to analyze trip efficiency alongside distance and fare.

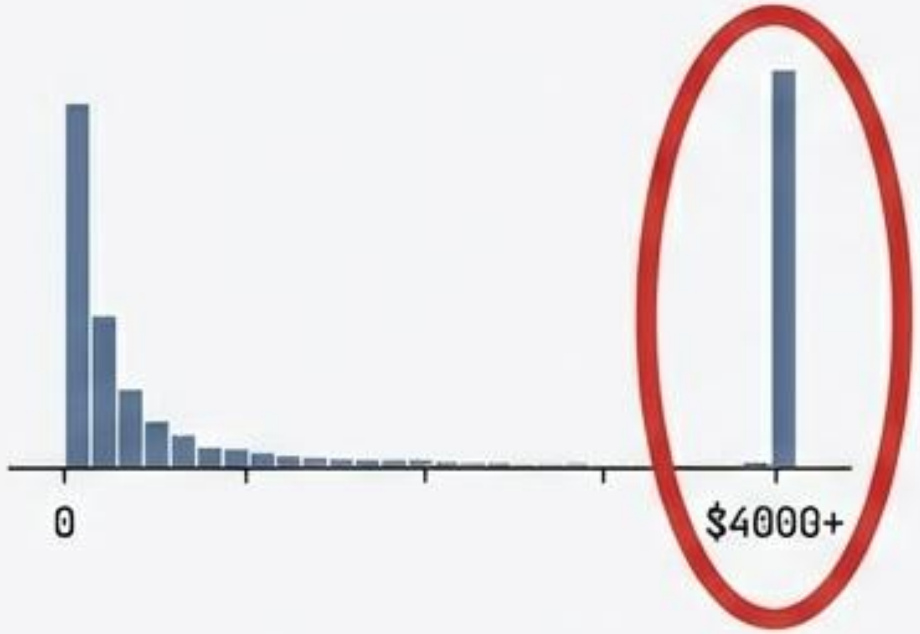
Filtering the Noise to Find the Signal



DEEPANSHU GUPTA

Statistical Cleaning: Handling Extreme Outliers

The Issue



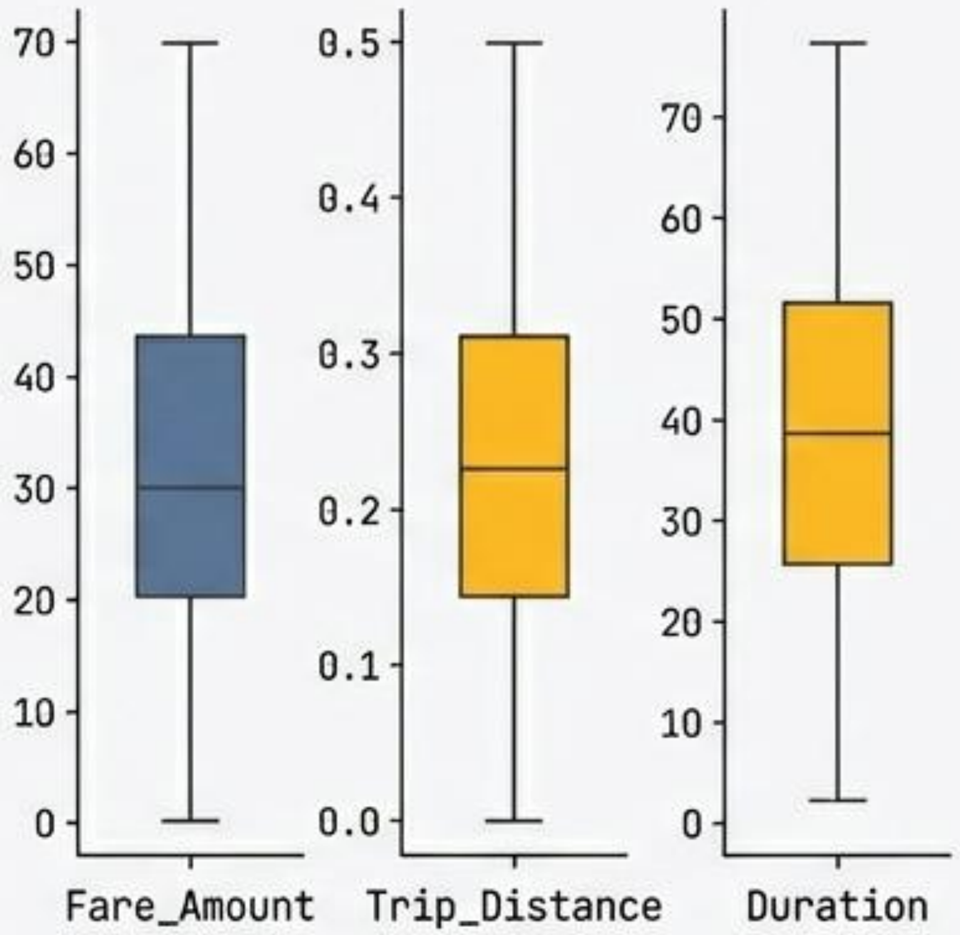
Extreme skew caused by data errors.

The Method

Lower = Q1 - 1.5 * IQR

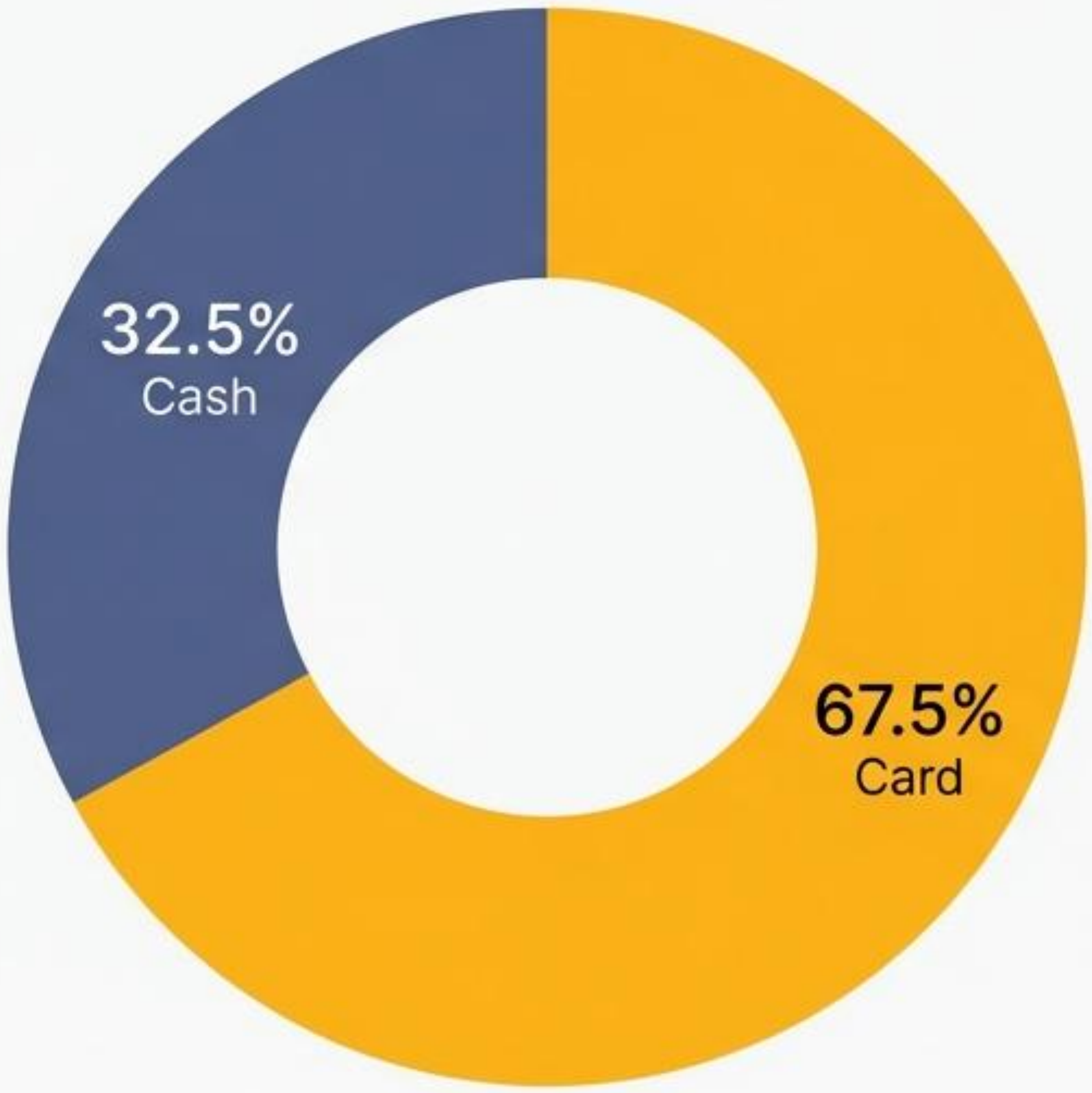
Upper = Q3 + 1.5 * IQR

The Result



DEEAPNSHU GUPTA

Card Payments Dominate the Market Share

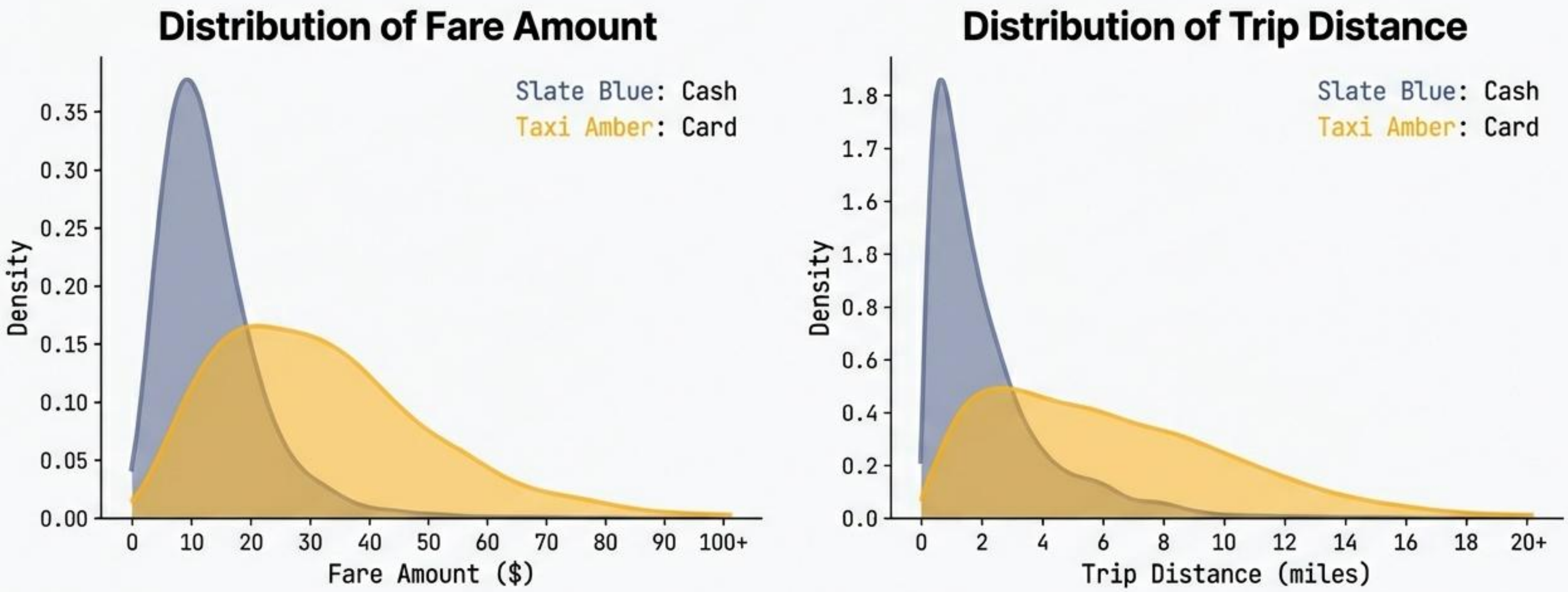


2/3rds of all transactions are already digital.

While cards are dominant, cash remains a significant 32% of the market—representing a major conversion opportunity.

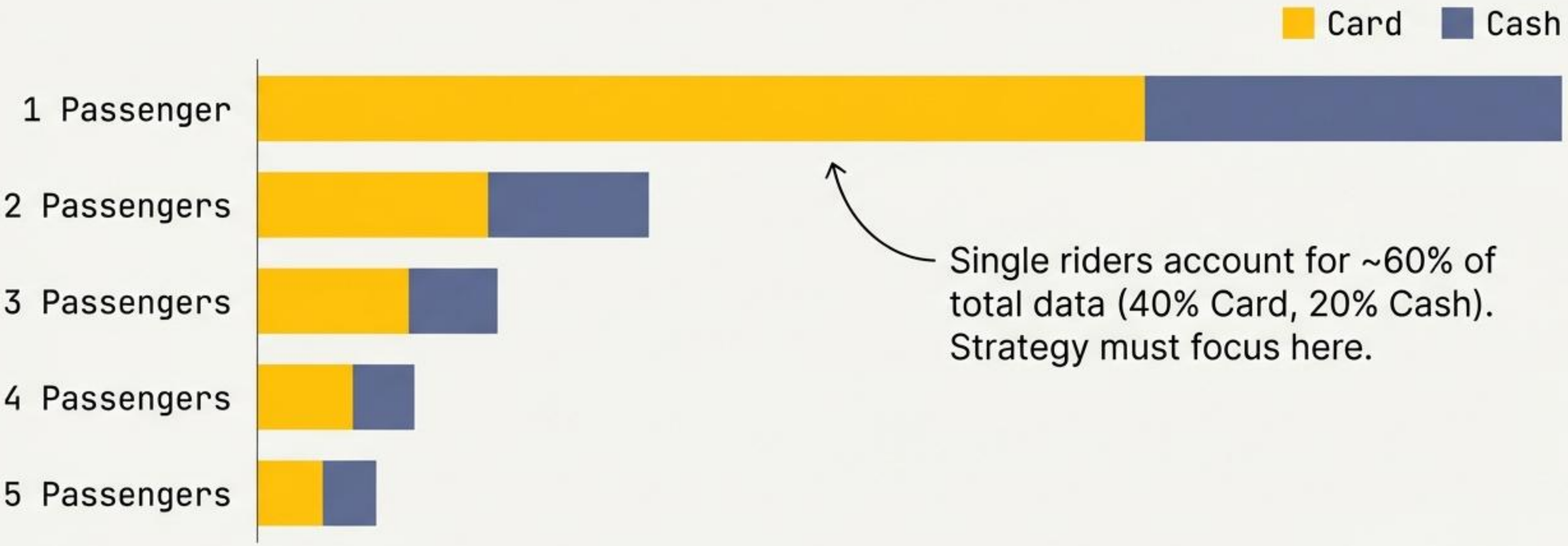
DEEPANSHU GUPTA

Card Users Take Longer Trips and Pay Higher Fares



Visual evidence suggests a correlation between higher value trips and card usage.

Single Passengers Drive the Majority of Revenue



DEEPANSHU GUPTA

The Raw Numbers: A Clear Value Gap

Payment Type	Mean Fare ($)	Mean Distance (Miles)	Standard Deviation
Card	$ 14.50	6.8 Miles	High SD
Cash	$ 11.00	3.2 Miles	Low SD

The Revenue Gap

DEEPANSHU GUPTA

Putting the Observation to the Test

Null Hypothesis (H_0)

There is NO difference in average fare between Card and Cash users.

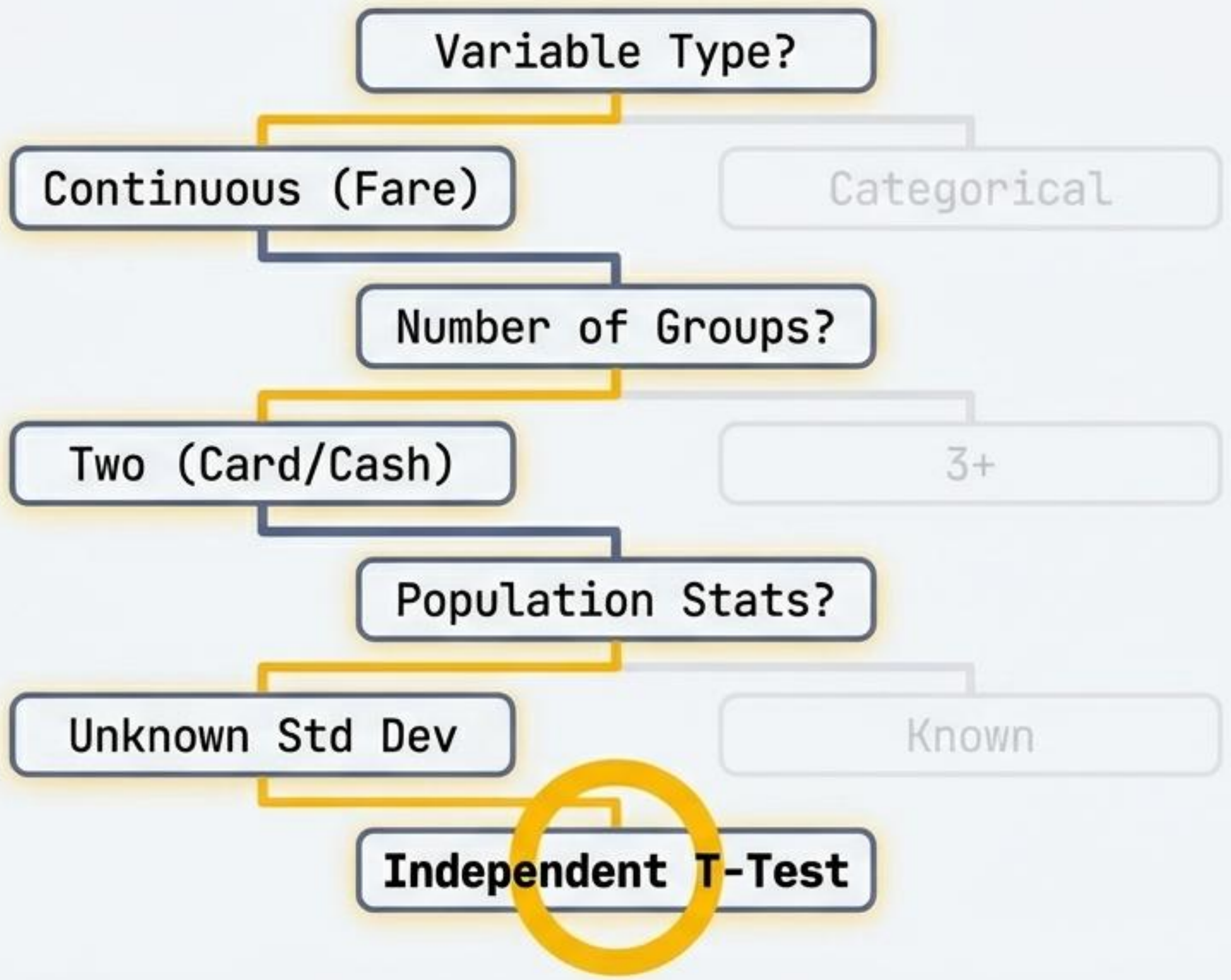
Alternate Hypothesis (H_1)

There IS a **significant difference** in average fare.

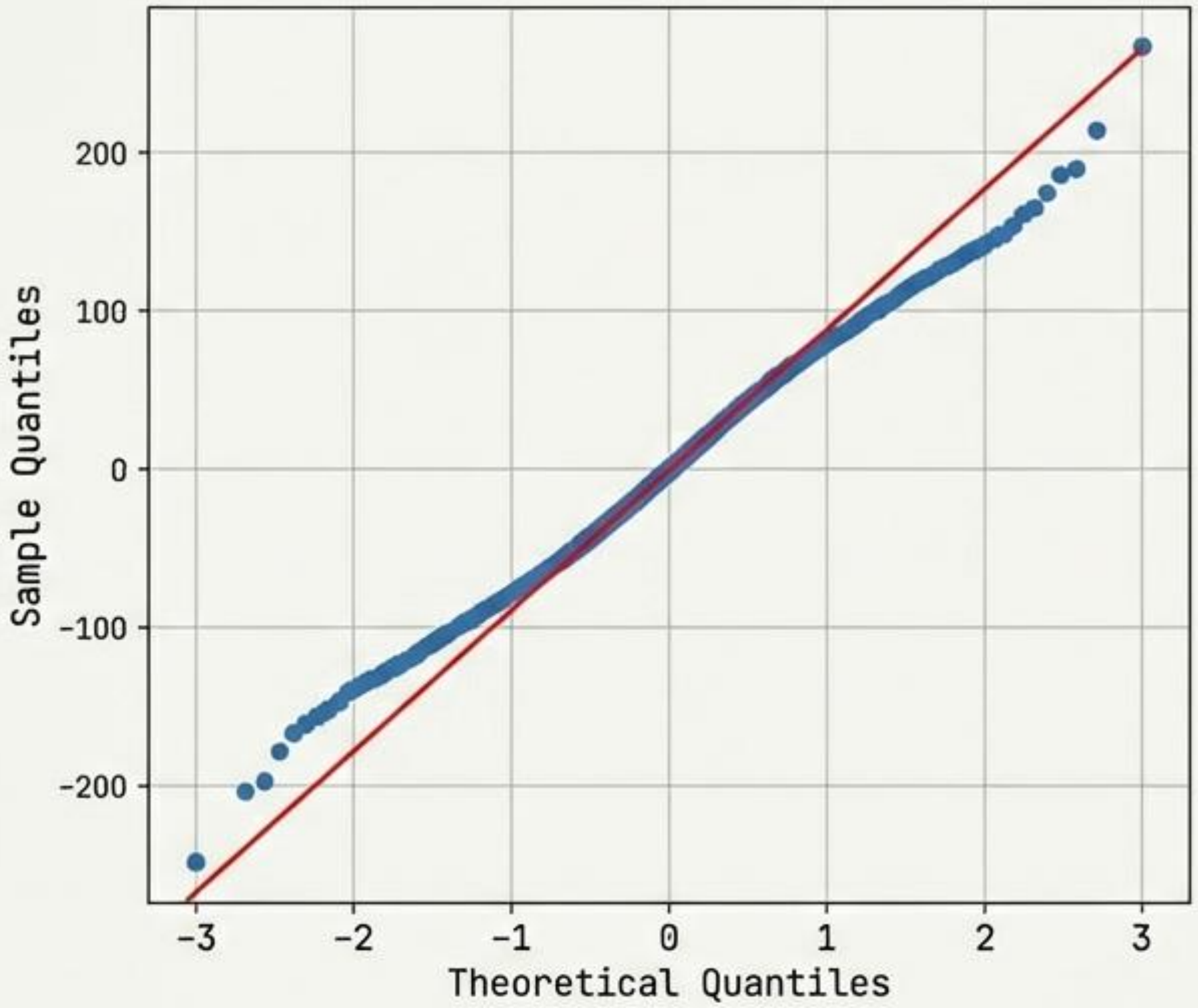
`Methodology: Comparing two independent samples (Card Sample vs. Cash Sample).`

DEEPANSHU GUPTA

Selecting the Right Statistical Tool



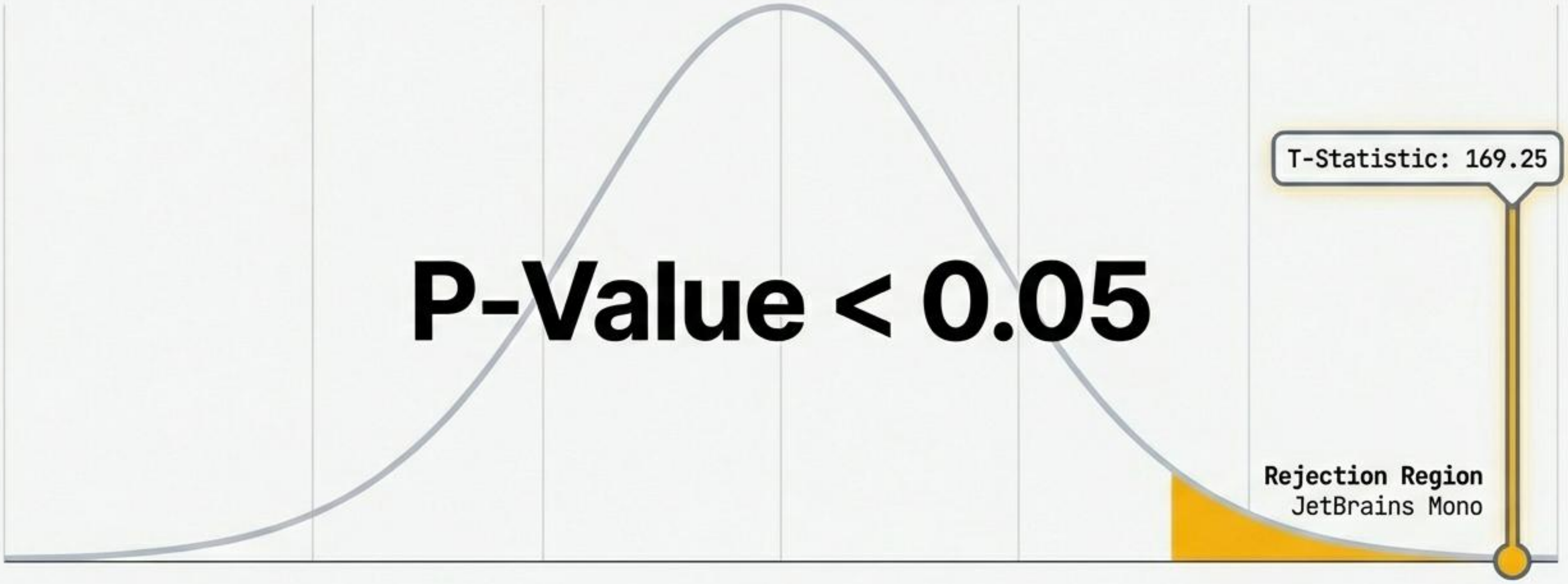
Validating Assumptions: The Normality Check



- **Observation:** Data deviates from the diagonal line -> **Not Normal**.
- **Justification:** Despite non-normality, the massive sample size (>2M rows) satisfies the Central Limit Theorem, making the **T-Test robust**.

DEEPANSHU GUPTA

The Verdict: The Difference is Statistically Significant



We **REJECT the Null Hypothesis**. The revenue difference is not due to chance.

Synthesis: The ‘Card Premium’ is Real

Higher Revenue. Card users generate higher average fares per trip.

Longer Trips. Card users are sigfficantly more likely to take longer journeys.

Dominant Volume. Cards already account for 67% of transactions.

Recommendation 1: The Digital Nudge

Strategy: Default the payment terminal to **'Card' for trips over 5 miles.**

Rationale: Reduce friction. Leverage the correlation between distance and card usage by making the preferred behavior the default option.

Recommendation 2: Incentivize the Switch

Strategy:
Offer micro-discounts or loyalty points for card payments on high-value fares.

Rationale: Target “swing” users—those who have cards but pay cash out of habit. Decouple the “pain of paying” by gamifying the card experience.

DEEPANSHU GUPTA

Recommendation 3: Tech & Trust Infrastructure

'Security Verification'

'Speed/NFC'

Strategy:

Deploy visible 'Secure Payment' badges and contactless NFC terminals.

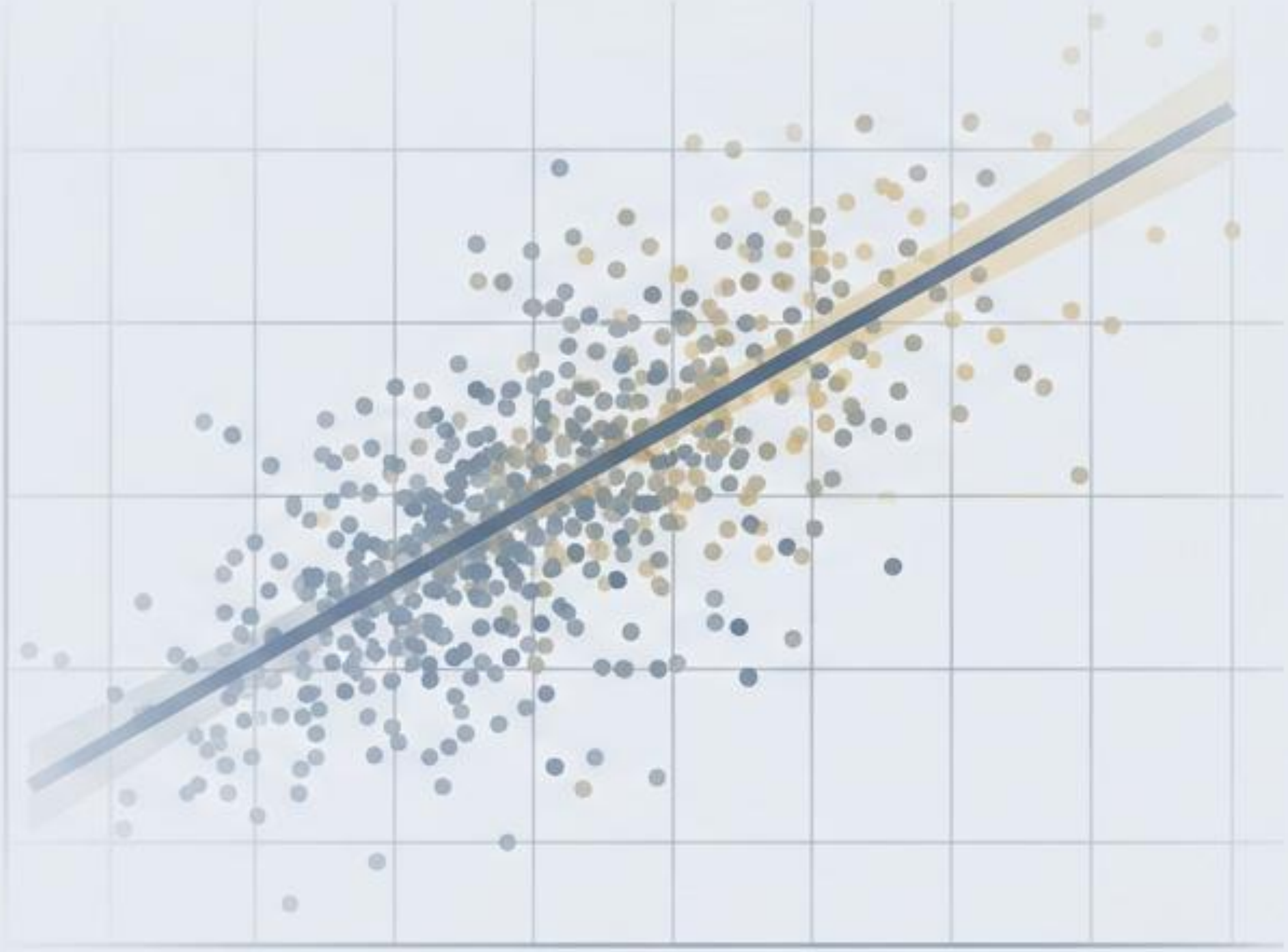
Rationale:

Remove the barrier of trust. Many users stick to cash due to security fears. **Visible verification encourages higher-value digital tannes-value digital transactions.**

Future Roadmap: Beyond Descriptive Stats

Current State:
Correlation Proven.

Next Step:
Regression Analysis.
Using the engineered `Duration` feature to predict exact fare amounts.



Data + Statistics = Maximized Profitability.

DEEPANSHU GUPTA