

Taxi Price Analysis Project (Simple Explanation)

Project Idea (In Simple Words)

This project looks at taxi trip data to understand **how taxi prices are calculated** and **what really affects the final fare**.

The goal is to explain the results in a way that **anyone can understand**, even without a technical background.

The project answers questions like: - What makes a taxi trip expensive or cheap? - Does traffic increase the price? - Are prices higher during peak hours?

Problem Definition

Taxi pricing is often unclear to passengers and decision makers. Many people assume that traffic congestion or peak hours automatically increase fares, but this is not always true.

The problem addressed in this project is: - Identifying the real factors that influence taxi trip prices - Verifying whether peak hours lead to higher fares - Explaining pricing behavior using real data instead of assumptions

Literature Review (Simple Overview)

Previous studies on transportation pricing show that: - Distance and trip duration are usually the strongest factors affecting taxi fares - Traffic congestion increases travel time more than direct pricing - Many taxi systems use fixed pricing rules rather than surge pricing

Research in data analysis and urban mobility suggests that statistical analysis and visualization are effective tools for understanding pricing behavior and detecting unfair pricing patterns.

This project follows similar approaches by using real trip data, exploratory analysis, and hypothesis testing to reach clear conclusions.

Tools Used (Very Simply Explained)

- **Python:** Used to analyze numbers and data.
- **Charts & Graphs:** Used to visually explain patterns.
- **Statistics:** Used to make sure conclusions are correct and not random.
- **Public Dataset:** Real taxi trip data.

Phase 1: Getting the Data

What Happened

- A real taxi dataset was downloaded.
- The data was loaded so it could be analyzed.

Why This Matters

You cannot analyze anything without data. This step makes sure we start with real information.

Phase 2: Understanding How Taxi Prices Work

What Happened

We assumed taxi prices depend on: - Distance traveled - Time spent in the trip - Fixed base fare

Why This Matters

Knowing how prices *should* work helps us check if the data makes sense.

Phase 3: Exploring the Data

What Happened

- Checked what information exists in the data.
- Looked at averages, minimums, and maximums.
- Used charts to see how values are distributed.

Why This Matters

This helps us understand the data before making conclusions.

Phase 4: Cleaning the Data

What Happened

- Removed duplicate records.
- Filled missing values.
- Removed unrealistic values (very large or very small numbers).
- Simplified prices using a mathematical transformation.

Why This Matters

Clean data gives more accurate and trustworthy results.

Phase 5: Creating Helpful Information (Feature Engineering)

What Happened

New useful information was created, such as: - Trip duration in hours - Average speed - Whether the trip happened in peak hours - Whether the trip was on a weekend - Traffic level converted into numbers

Why This Matters

This helps explain pricing behavior more clearly.

Phase 6: Looking at Relationships

What Happened

- Compared distance, time, traffic, and price together.
- Used charts to see how multiple factors interact.

Why This Matters

Taxi prices depend on more than one thing at the same time.

Phase 7: Finding What Matters Most

What Happened

Different methods were used to find the most important factors affecting price.

Final Important Factors

- Distance
- Trip duration
- Price per kilometer
- Price per minute
- Base fare
- Traffic conditions

Why This Matters

This shows what really controls taxi pricing.

Phase 8: Peak vs Off-Peak Analysis

Question

Are taxi fares higher during peak hours?

What Happened

- Compared prices during peak and non-peak times.
- Used statistical tests to confirm results.

Final Answer

There is **no clear difference** between peak and off-peak fares.

Phase 9: Understanding Trip Duration

What Happened

- Studied how long trips usually take.
- Checked which statistical shape fits trip durations best.

Result

Trip durations follow a realistic and expected pattern.

Phase 10: Simplifying the Big Picture (PCA)

What Happened

- Combined many factors into two main patterns.
- Used visuals to understand traffic and distance effects.

Why This Matters

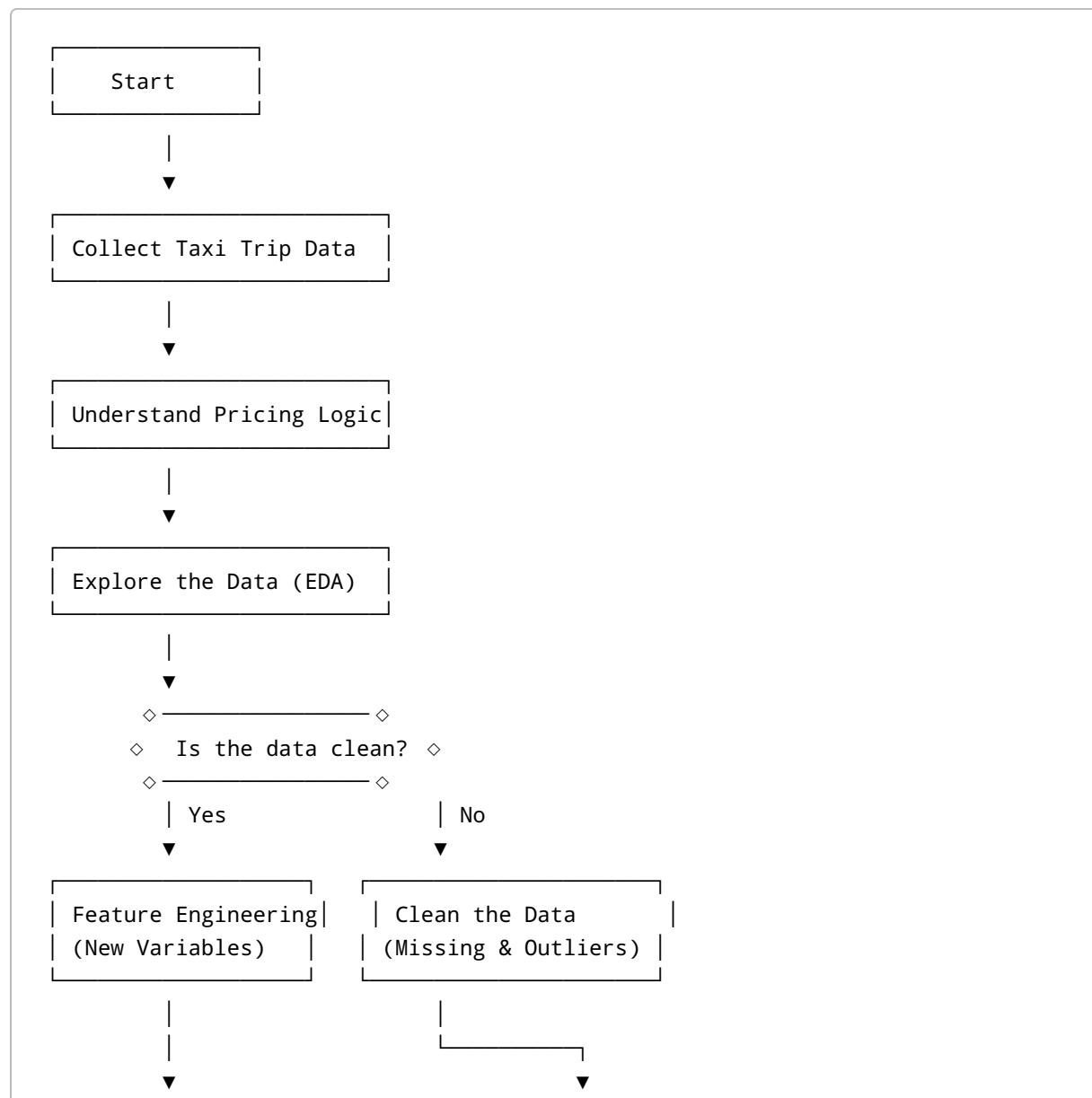
It makes complex data easier to understand.

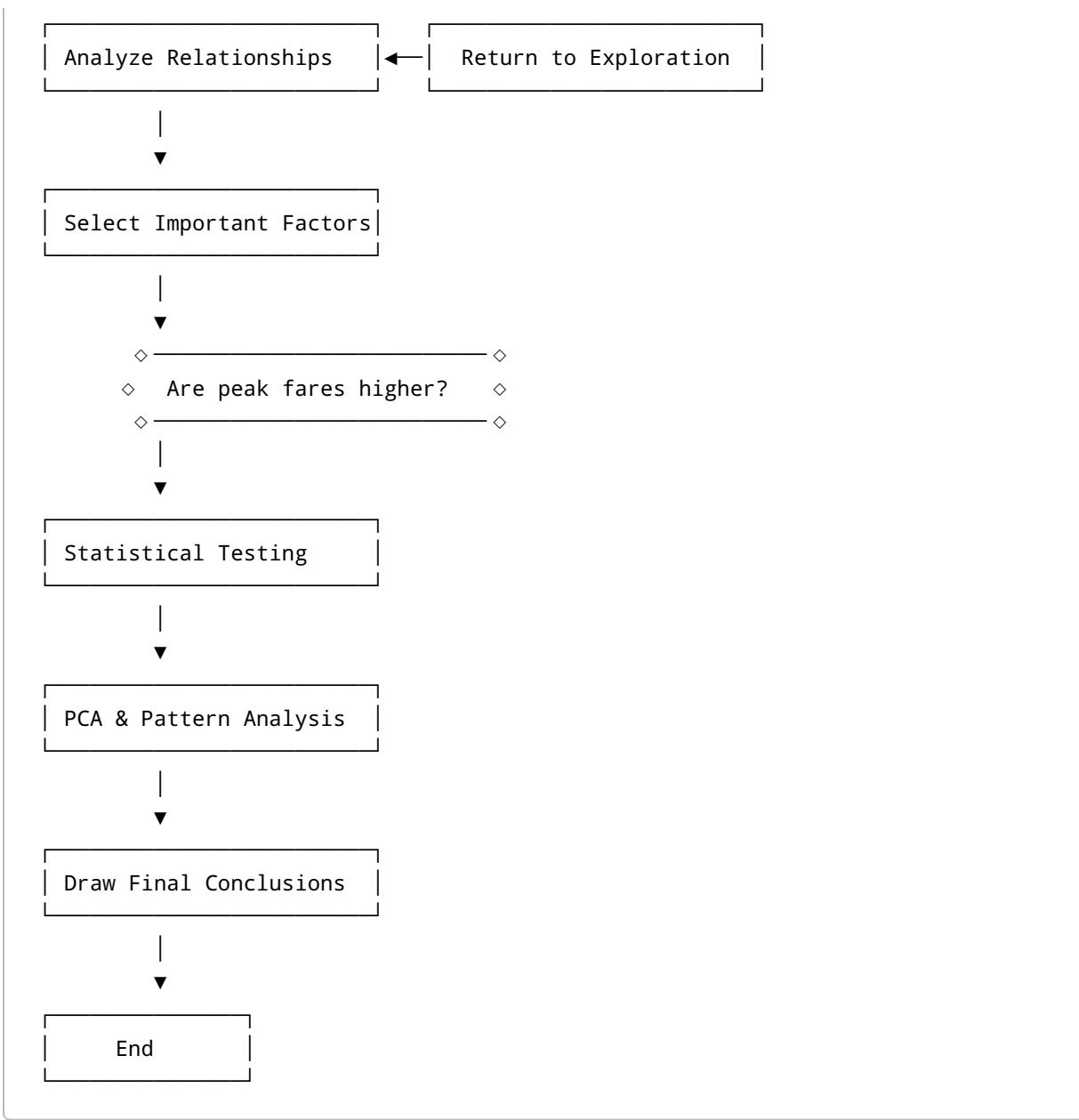
Final Conclusions (In Plain English)

- Taxi prices mainly depend on **distance and time**.
- Traffic affects how long trips take, not the price directly.
- Peak hours do **not** significantly increase fares.
- Pricing appears fair and consistent.

Project Flowchart (Step-by-Step)

Below is a simple flowchart using **basic shapes**: - **Rectangle** → Process / Action - **Rhombus (Diamond)** → Decision





References

1. Kaggle Taxi Price Prediction Dataset
2. Scikit-learn Documentation – Feature Selection & PCA
3. SciPy Documentation – Statistical Tests
4. Pandas Documentation – Data Analysis
5. Research articles on transportation pricing and urban mobility