Exploratory Data Analysis
G2M Case Study – XYZ Cab Industry Analysis

Magdalena Ivanova | 18-September-2025

Agenda

Data Glacier
Your Deep Learning Partner
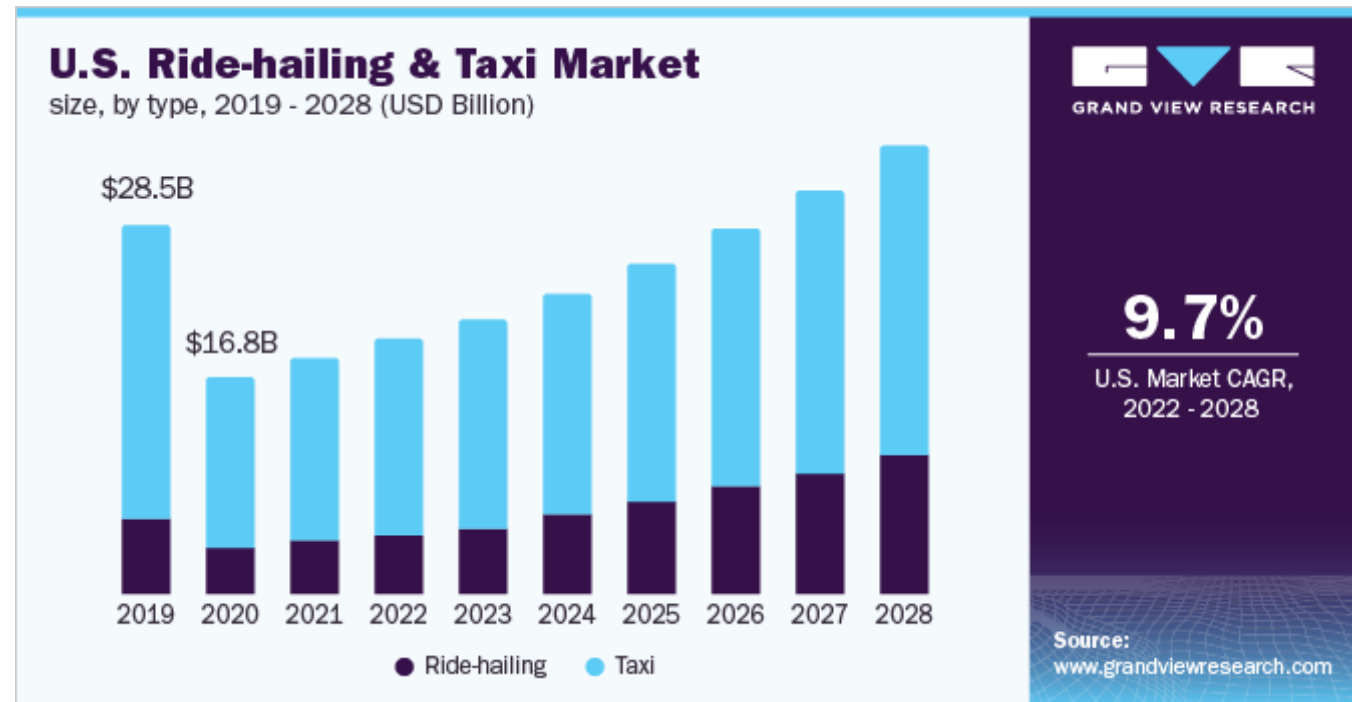
Background – Case study

- XYZ, a private firm, is evaluating an investment in the U.S. cab market. Two companies were analyzed: Yellow Cab and Pink Cab.

- This presentation uses cab industry data from 2016–2018 to test key hypotheses on market share, profitability, customer segments, payment modes, city penetration, and forecasting.

- Objective: Recommend the stronger investment opportunity for XYZ based on data-driven insights.

The Client & Problem Statement

- XYZ, a private firm, seeks entry into the U.S. cab market, which has grown rapidly but remains fragmented.

- **The decision:** Which company, Yellow Cab or Pink Cab, is the stronger long-term investment?

- **Key factors assessed:** profitability, customer retention, demographics, geographic penetration, and growth outlook.

U.S. Ride-hailing & Taxi Market
size, by type, 2019 - 2028 (USD Billion)

GRAND VIEW RESEARCH

$28.5B

$16.8B

2019 2020 2021 2022 2023 2024 2025 2026 2027 2028

● Ride-hailing ● Taxi

9.7%
U.S. Market CAGR, 2022 - 2028

Source:
www.grandviewresearch.com

Dataset Overview

- Period: Jan 2016 – Dec 2018
- 4 datasets merged into Master (~355K records)
- Enables analysis across: transactions, demographics, geography, profitability

Dataset	Description	Records	Key Fields
Cab_Data	Transaction details: fare charged, cost of trip	~359,000	'Transaction ID', 'Price Charged', 'Cost of Trip', 'City'
Customer_ID	Customer demographics (age, gender, income)	~49,000	'Customer ID', Age, 'Income'
Transaction_ID	Maps transactions to customers & payment modes	~440,000	'Transaction ID', 'Customer ID', 'Payment Mode'
City	U.S. cities with population & cab users	20	'City', 'Users'

Data Intake & Preparation

The datasets were cleaned and prepared before analysis:

- Removed duplicate entries across transactions.
- Checked for missing values and handled them appropriately.
- Derived new fields: Revenue = Price Charged, Profit = Revenue – Cost, Margin% = Profit ÷ Revenue.
- Merged files using Transaction_ID, Customer_ID, and City as keys.

**Final outcome**: A structured master dataset suitable for analysis.

| Raw Data | → | Cleaning | → | Master Dataset |
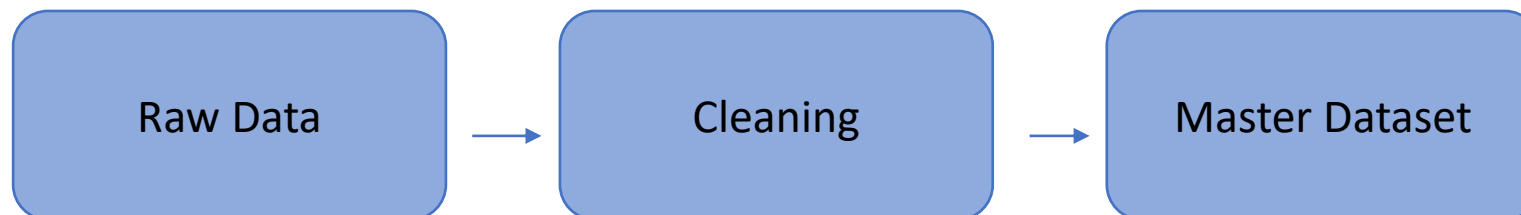
Data Relationships

The data sources are linked as follows:

- Cab_Data provides transaction-level details.
- Transaction_ID maps each trip to a customer.
- Customer_ID provides demographics for each customer.
- City.csv provides city-level context (population, cab users).

This relational structure enables analysis across customer behavior, city coverage, and profitability.
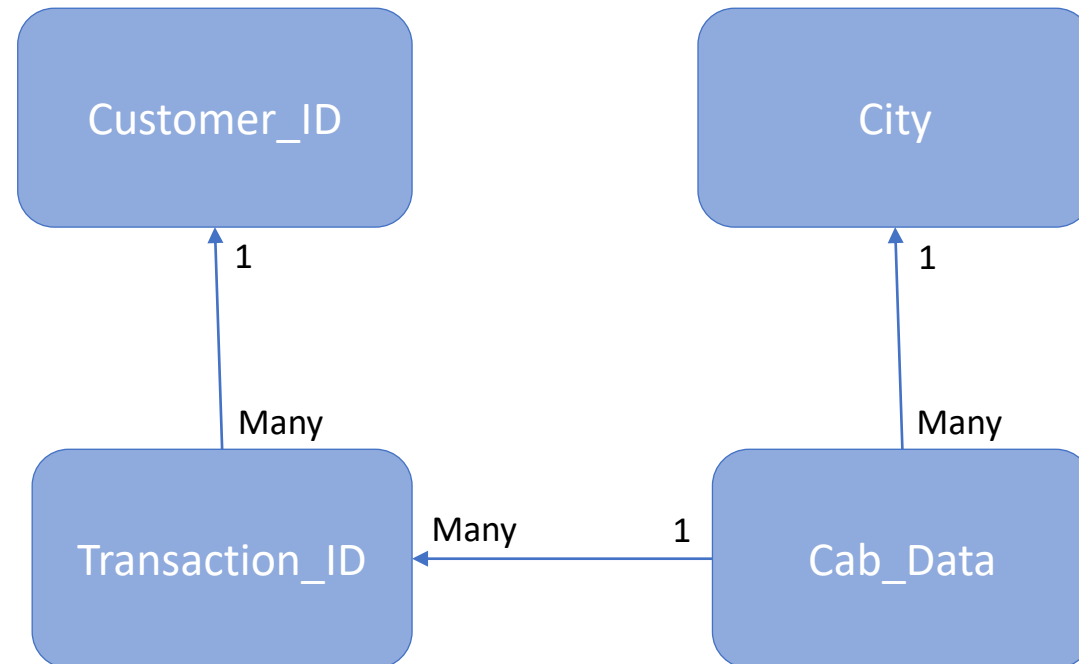
Hypotheses to Test

I investigated several hypotheses to understand the market:

- **H1:** Yellow Cab holds larger market share across time and cities.
- **H2:** Margin rises with customer count (monthly correlation).
- **H3:** There is monthly seasonality (peaks mid-year, troughs in Feb).
- **H4:** Higher-income and 25–44 age segments contribute more revenue.
- **H5:** Card payments dominate and are associated with slightly higher average revenue per trip.
- **H6:** Penetration (Users/Population) relates to revenue concentration (NY, CHI, DC, LA, BOS).

Company Performance

- Yellow Cab consistently generates higher revenue and profit across the 2016–2018 period.
- This indicates stronger operational efficiency and broader market share compared to Pink Cab.
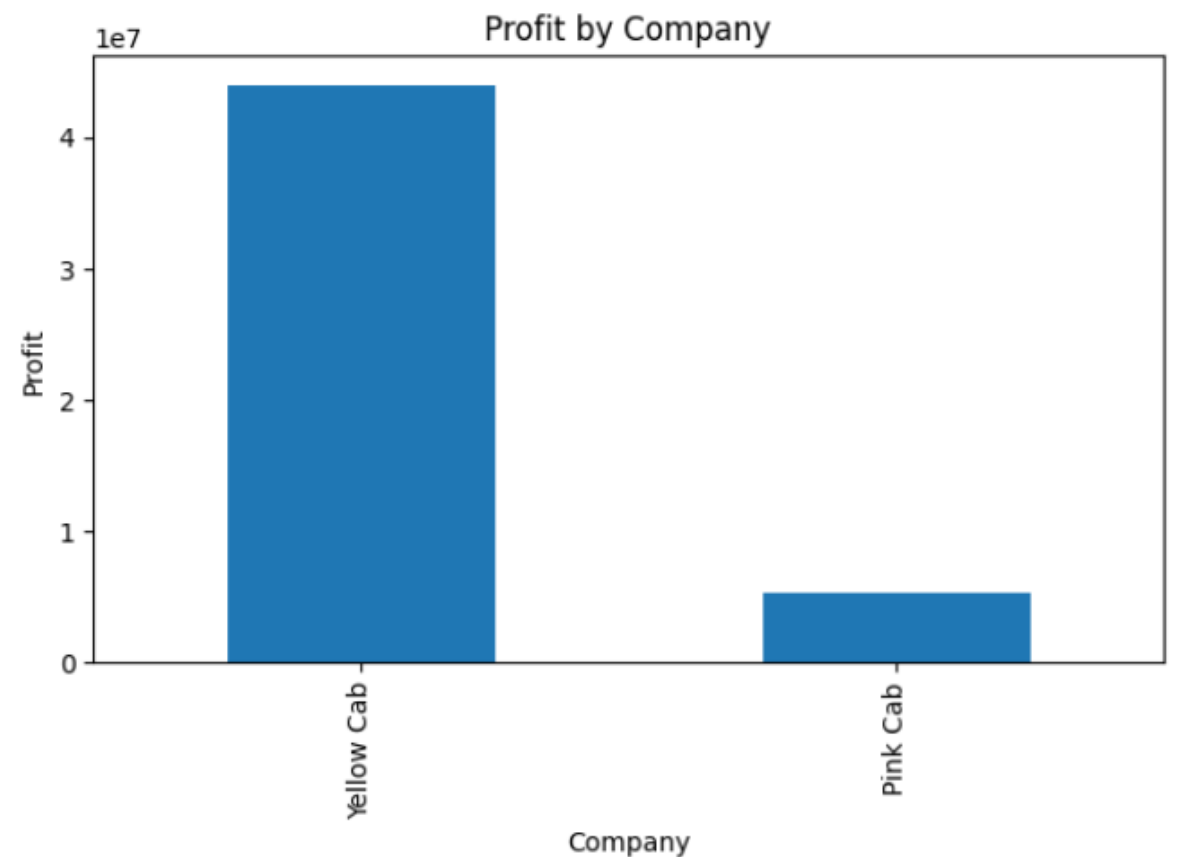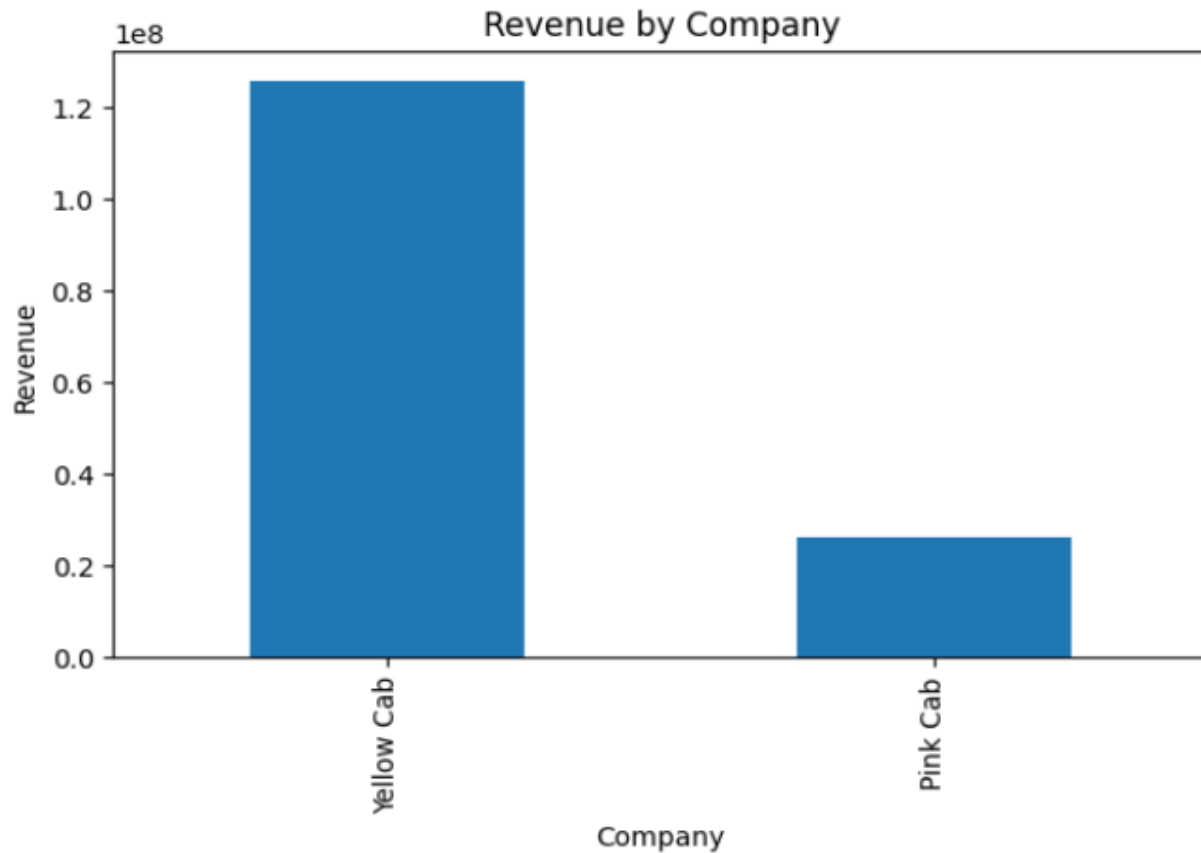
Profit Margin vs Customer Count

- Yellow Cab shows higher margins (30–40%) with more trips, proving strong economies of scale.
- Pink Cab stays at lower margins (10–25%), even with higher volumes.
- Growth in customers translates into sustained profitability for Yellow Cab.
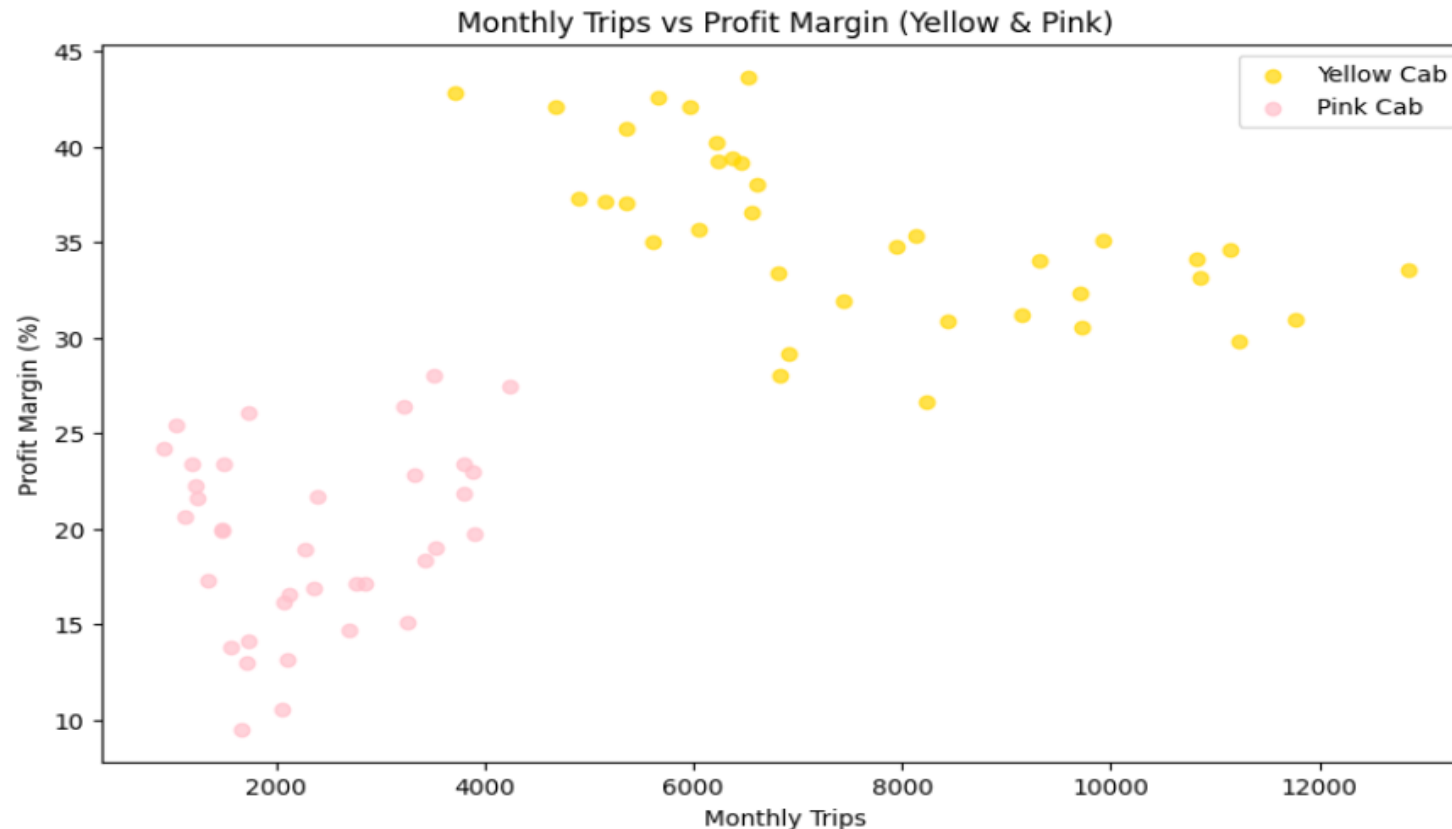
Seasonality trends- Monthly Trips by Company

- Yellow Cab consistently achieves **3× the trips** of Pink Cab, confirming a much larger market share.
- Both companies show **clear seasonal patterns**: demand peaks mid-year and early year, with sharp dips around February.
- Yellow Cab's growth trajectory is steeper, showing rising demand over the 2016–2018 period.


Monthly Trips by Company
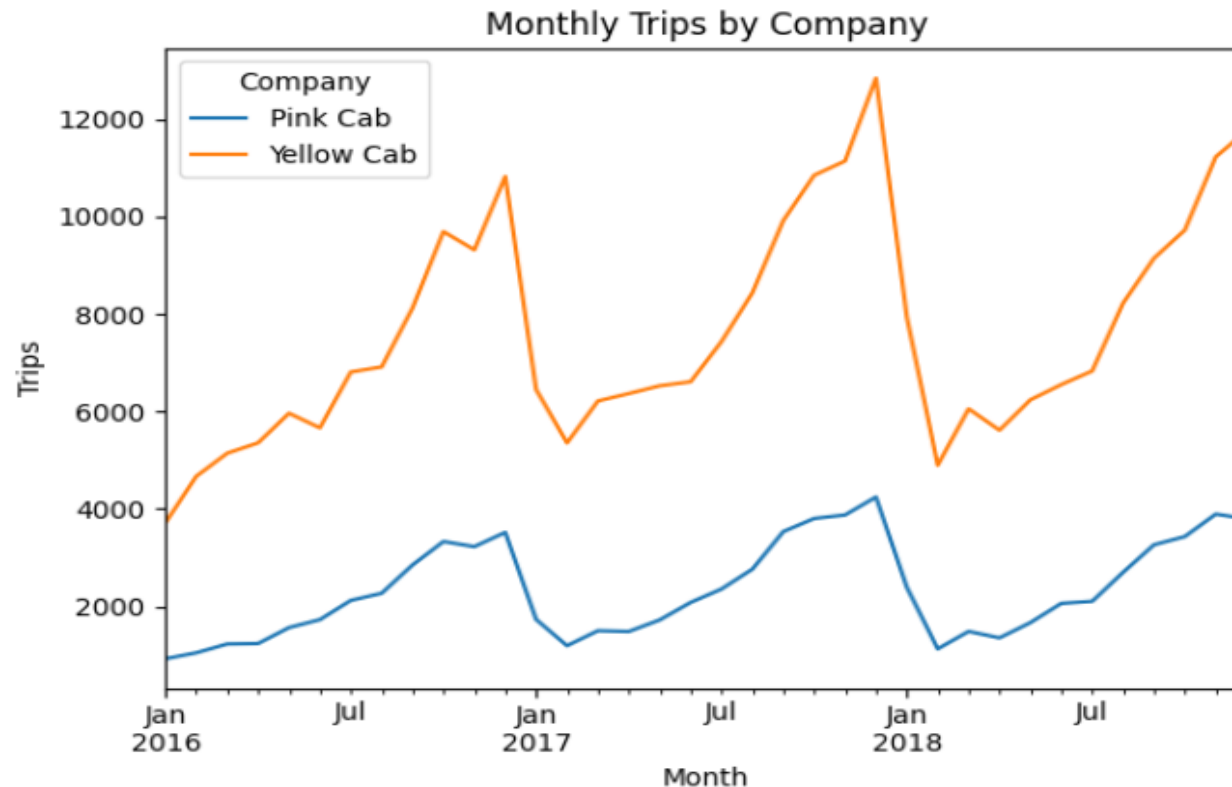
Seasonality trends - Monthly Revenue by Company

- Yellow Cab generates significantly higher revenue, consistently outpacing Pink Cab each month.
- Seasonal peaks in revenue align with trip spikes, reflecting strong, recurring demand cycles.
- Revenue growth over time is much stronger for Yellow Cab, while Pink Cab remains stagnant at lower levels.
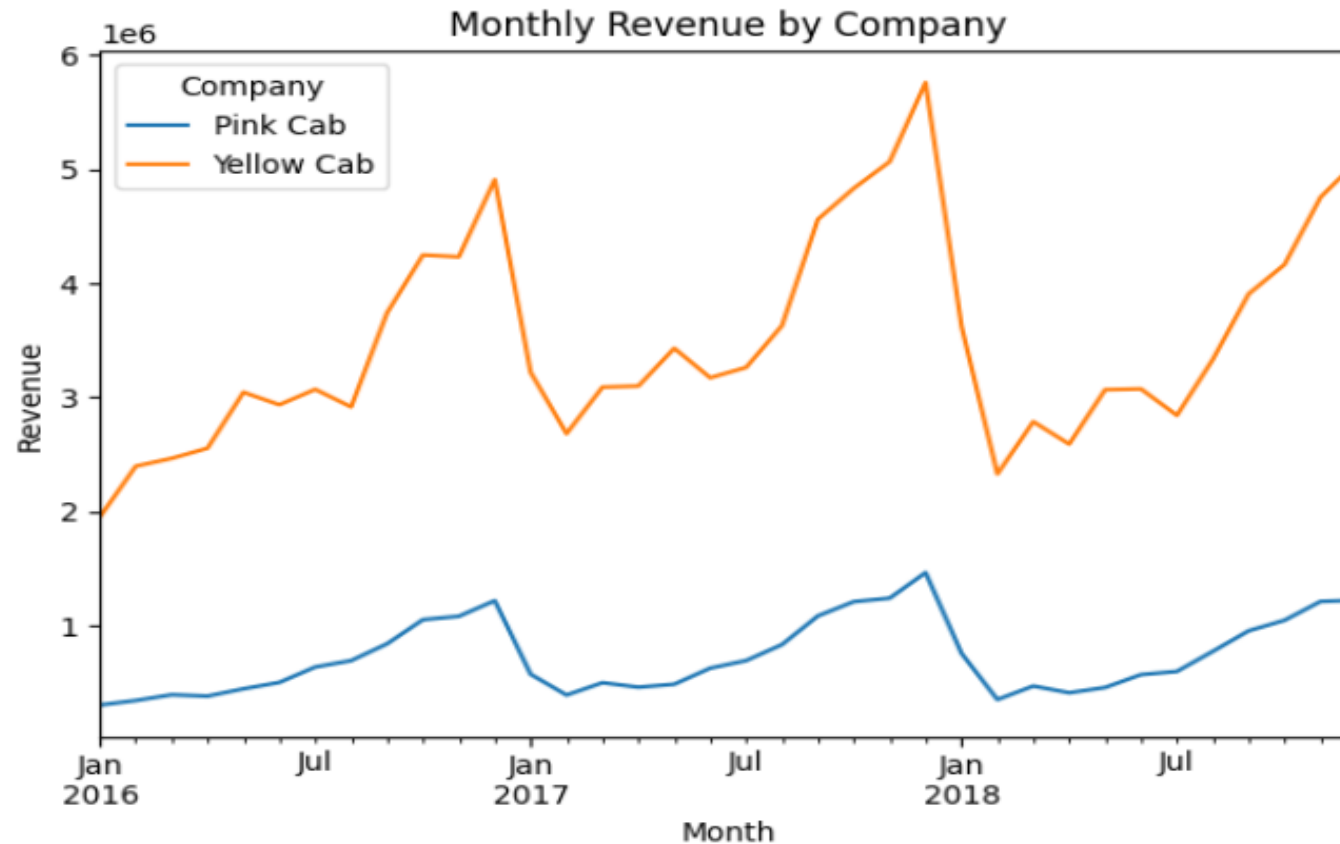
Customer Segmentation: Age

- The 25–34 age group generates the highest revenue, followed by 35–44 and ≤24.
- Revenue contribution drops steadily after age 45, with very limited demand among seniors (65+).
- Younger and middle-aged adults are therefore the core customer base for cab services.


Revenue by Age Group
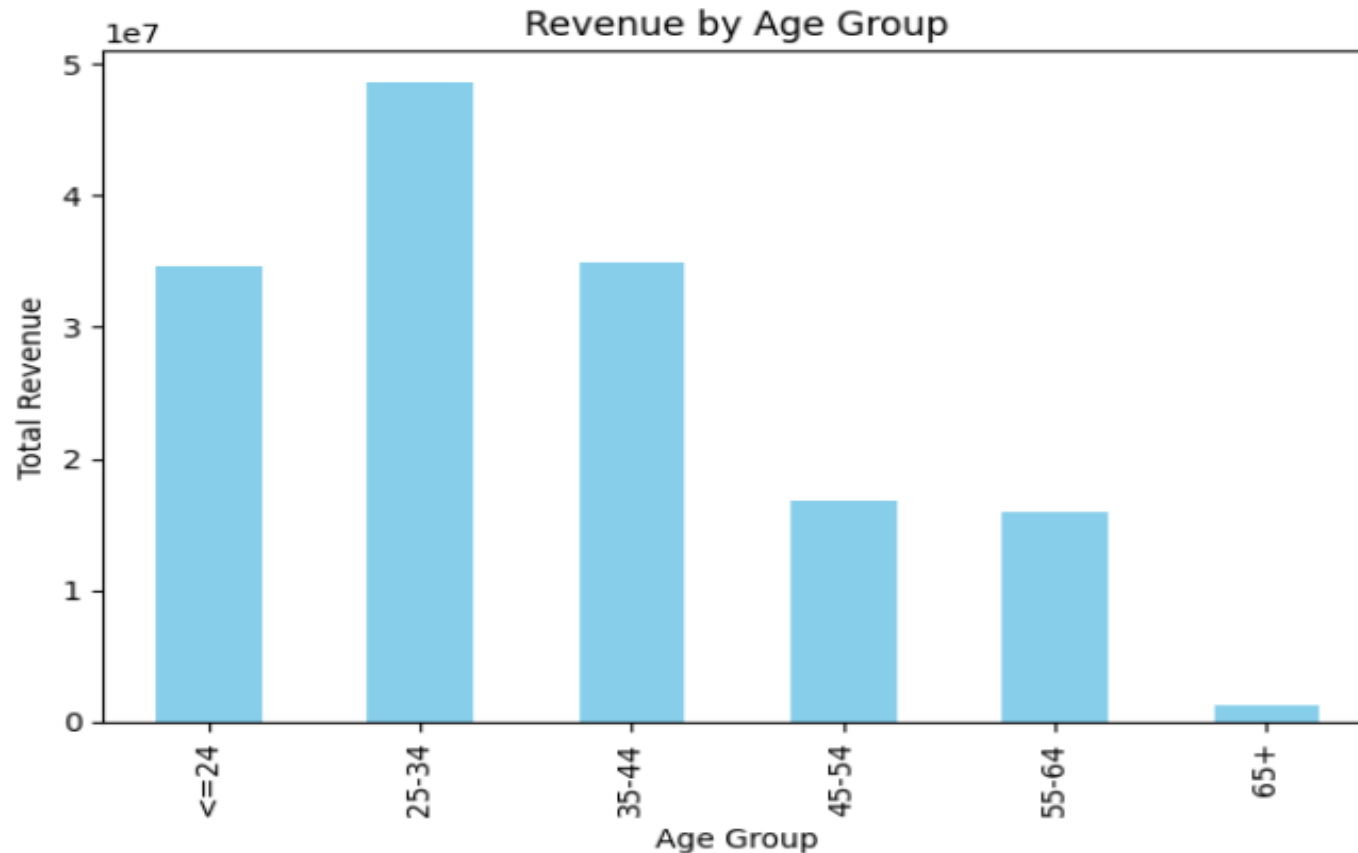
Customer Segmentation: Income

- Customers earning USD 15k+ per month contribute the majority of total revenue.
- Lower and middle income groups (<15k) are relatively balanced but far behind in total revenue share.
- Indicates that affluent customers drive cab profitability, though broad participation exists across all income levels.
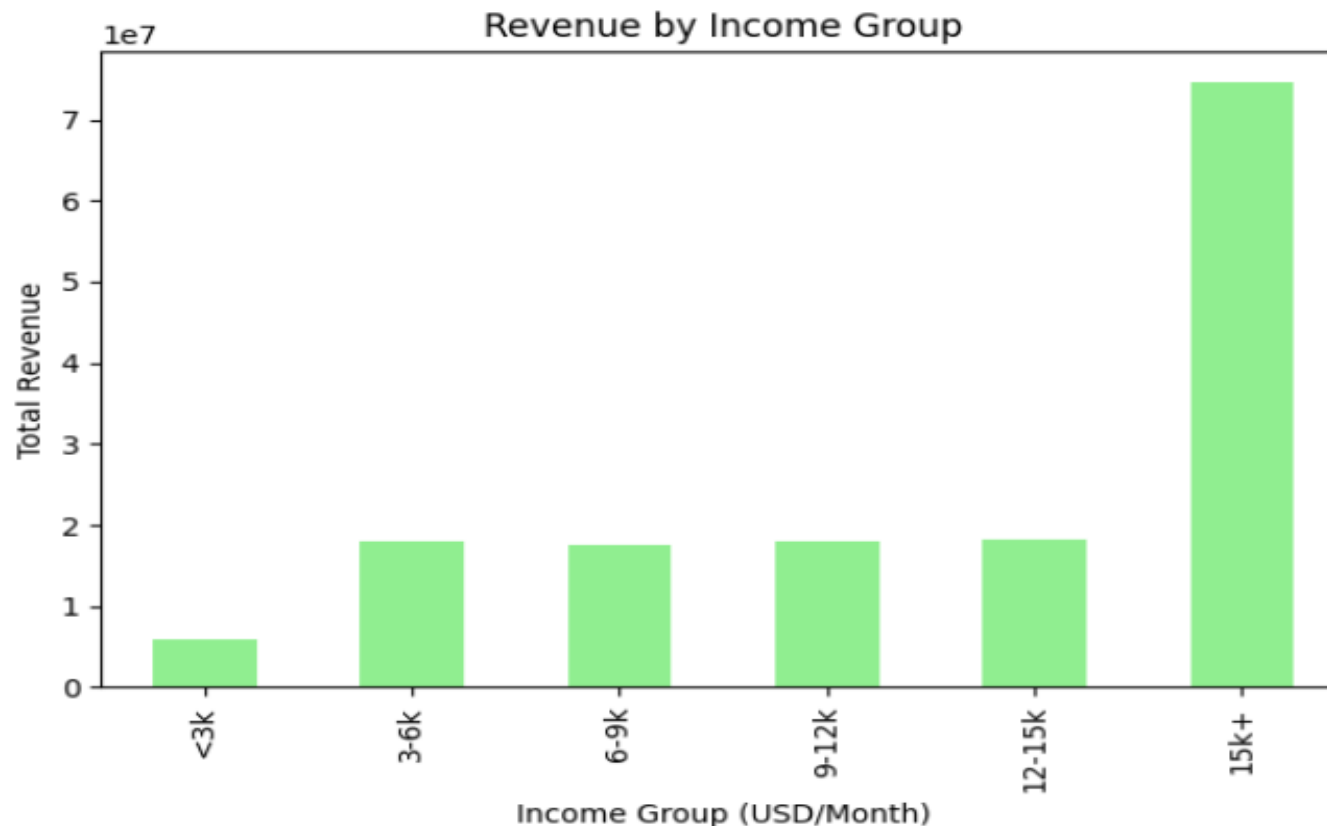
Payment Mode Preferences (Overall)

- Card payments dominate at 60% of total rides.
- Card trips generate slightly higher average revenue per trip compared to cash.
- Indicates a strong shift toward digital payments across the market.

Payment Mode Share

Revenue per Trip by Payment Mode
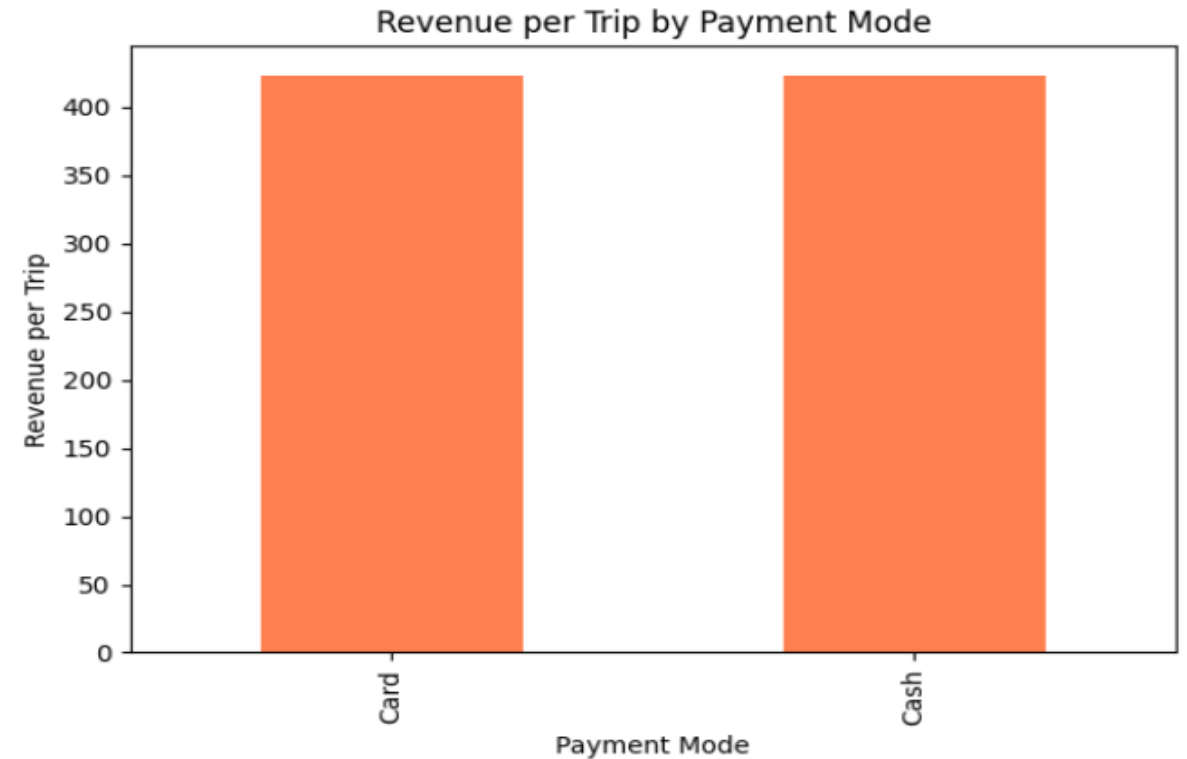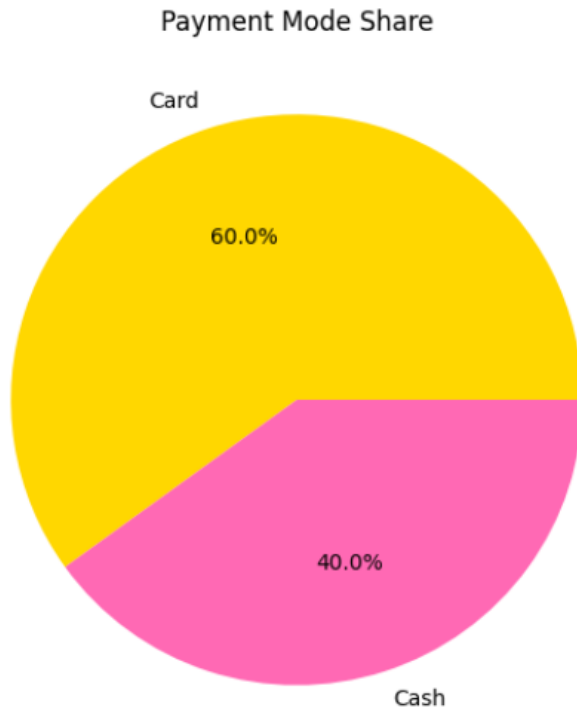
Payment Mode by Company

- Both companies show higher usage of card payments.
- Yellow Cab has a stronger card preference (≈60%+ card vs 40% cash).
- Pink Cab also skews toward cards, but with a smaller share.
- Digital adoption is a consistent trend, with Yellow Cab leading.
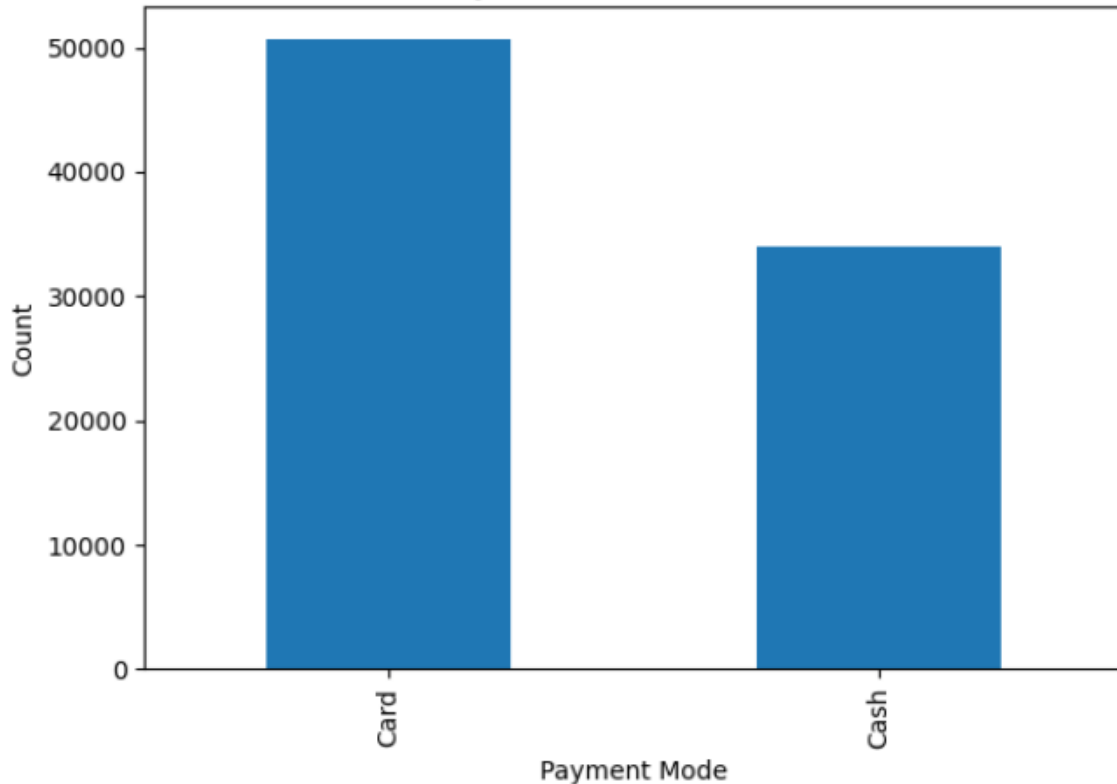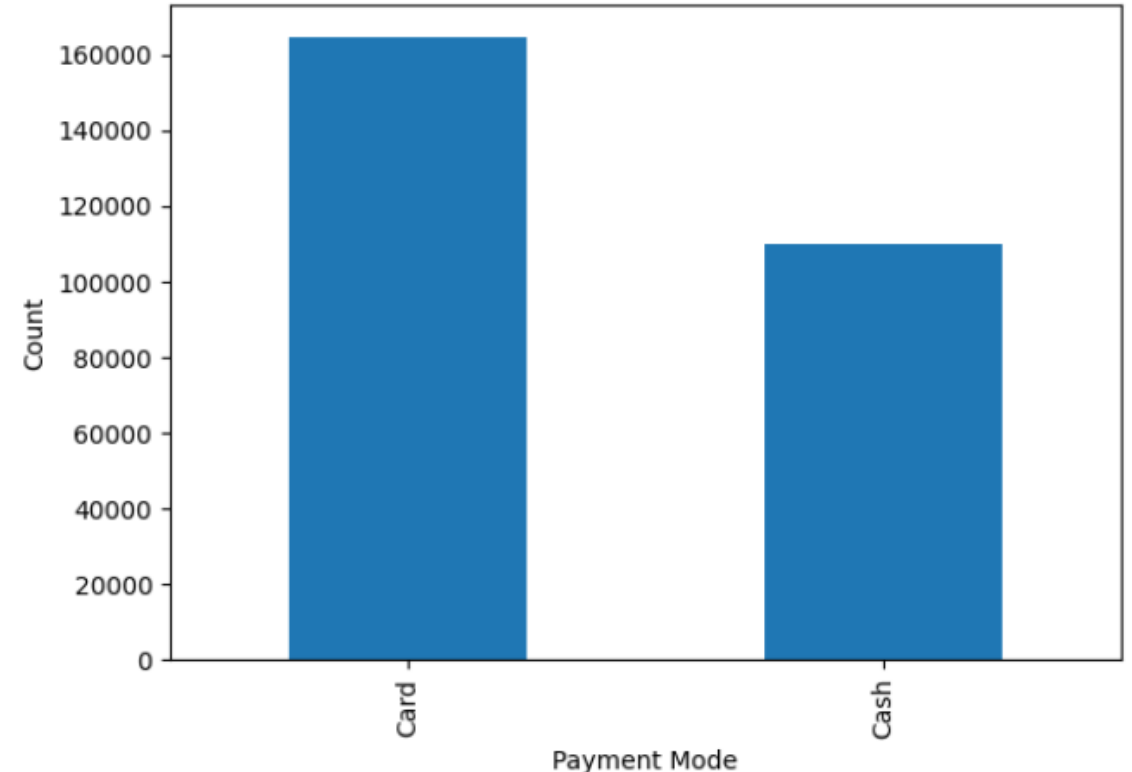
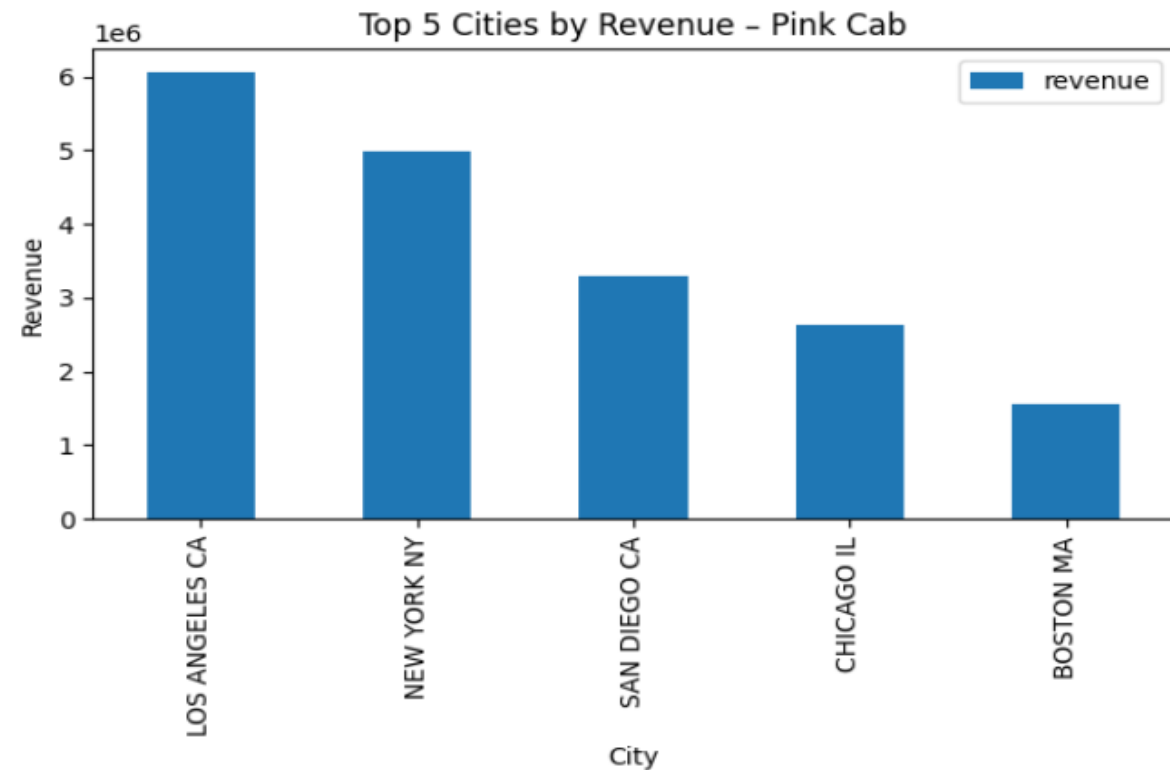# Top Cities by Revenue (Pink vs Yellow Cab)

- Pink Cab revenue is concentrated in Los Angeles, New York, San Diego, Chicago, and Boston, but total levels remain modest.
- Yellow Cab dominates in New York, followed by Chicago, Washington DC, LA, and Boston.
- Yellow Cab captures significantly larger absolute revenue in top cities, confirming stronger market dominance.
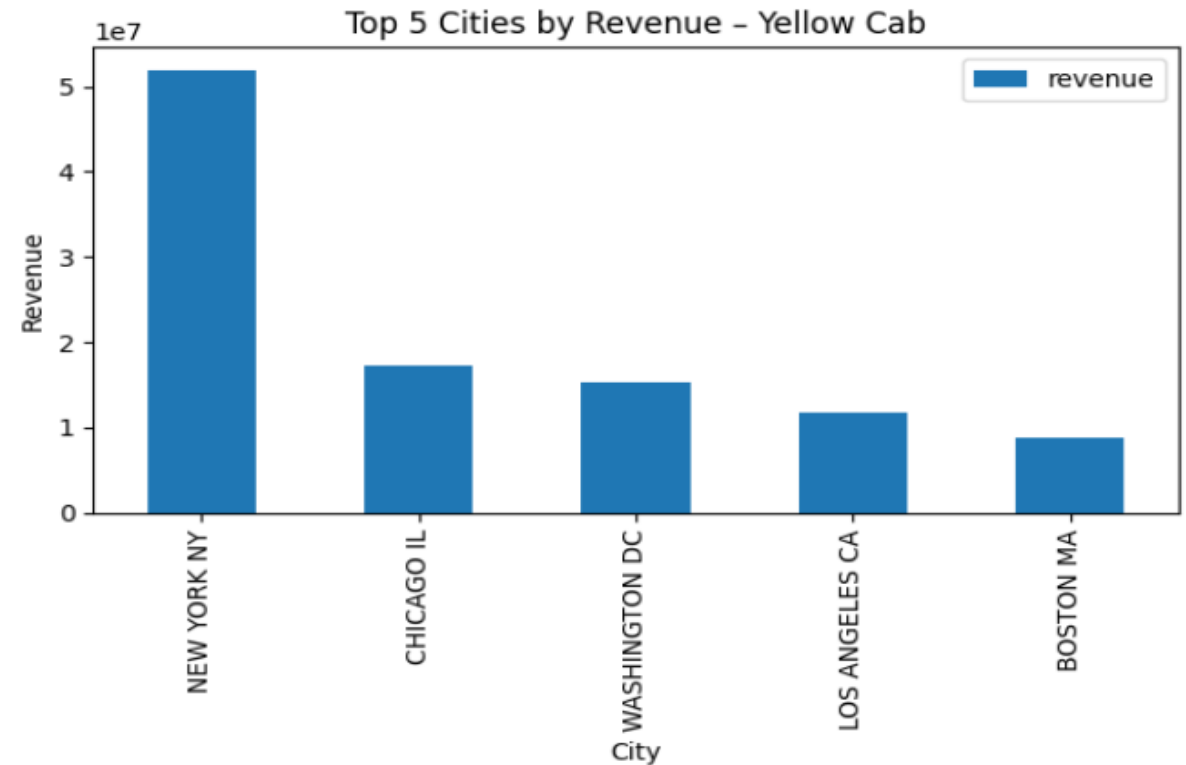
Penetration vs Revenue (City-Level Comparison)

- Revenue is strongly linked to penetration (users ÷ population).
- High-revenue cities like New York, LA, DC, and Boston also show higher penetration.
- Yellow Cab achieves better penetration in high-value cities, reinforcing its leadership in both scale and profitability.
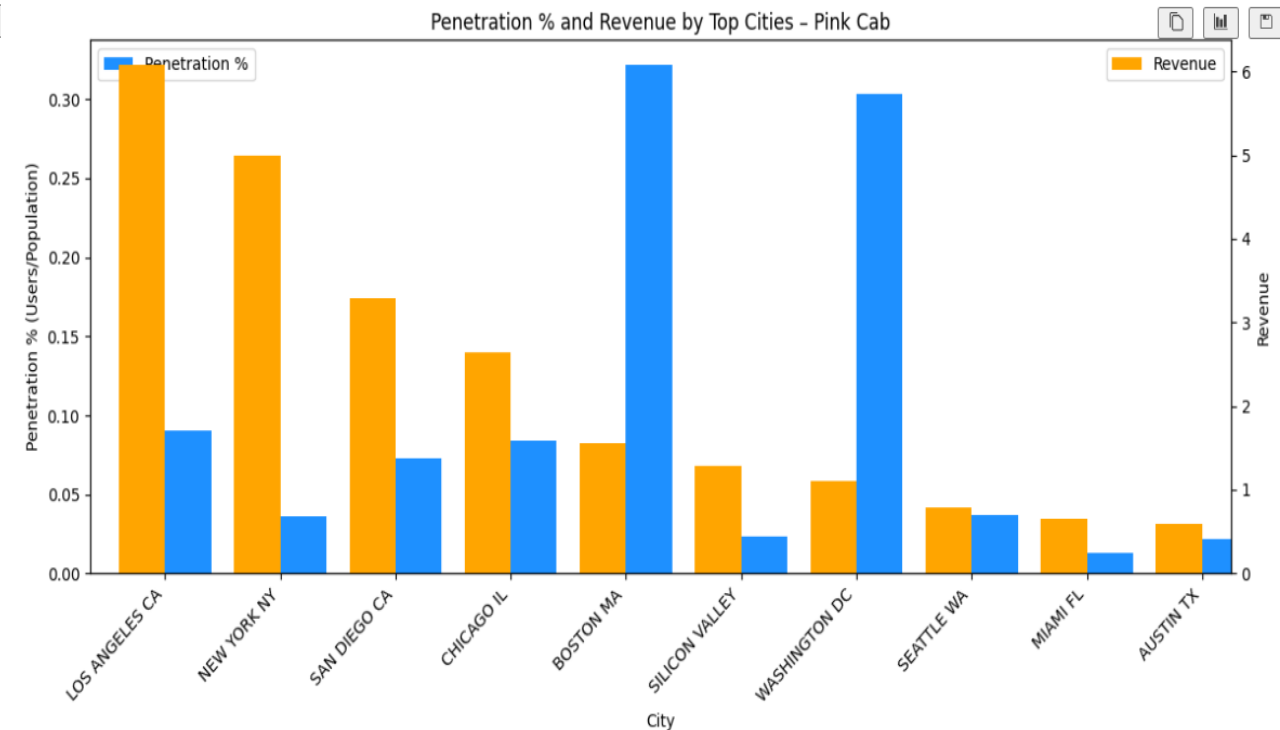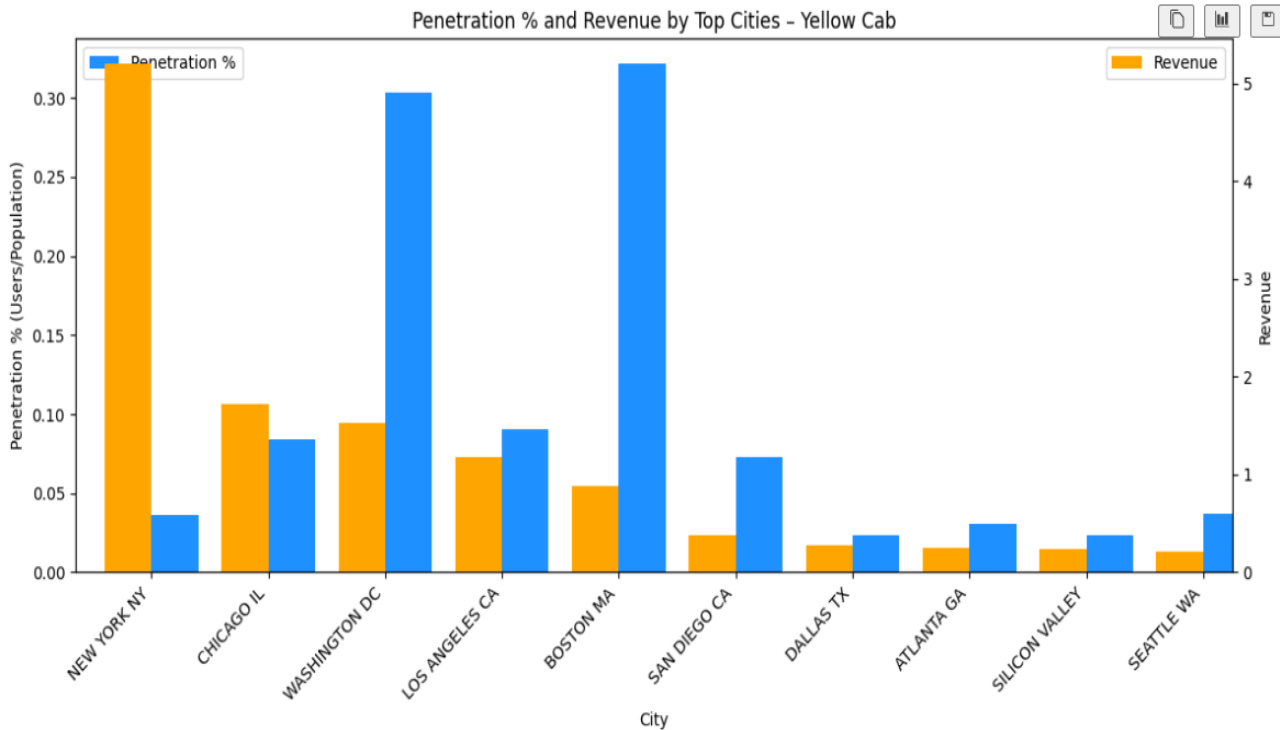
Forecasting (2019)

- Both companies face a decline in 2019.
- Yellow Cab: Trips stabilize at 6k–12k/month, decline ≈ –1.8%.
- Pink Cab: Stays much lower at 2k–4k/month, decline ≈ –3.1%.
- The forecast reinforces Yellow Cab's larger scale and greater resilience, even under slowing market conditions.

Actual vs Forecasted Monthly Trips (Yellow & Pink Cab)
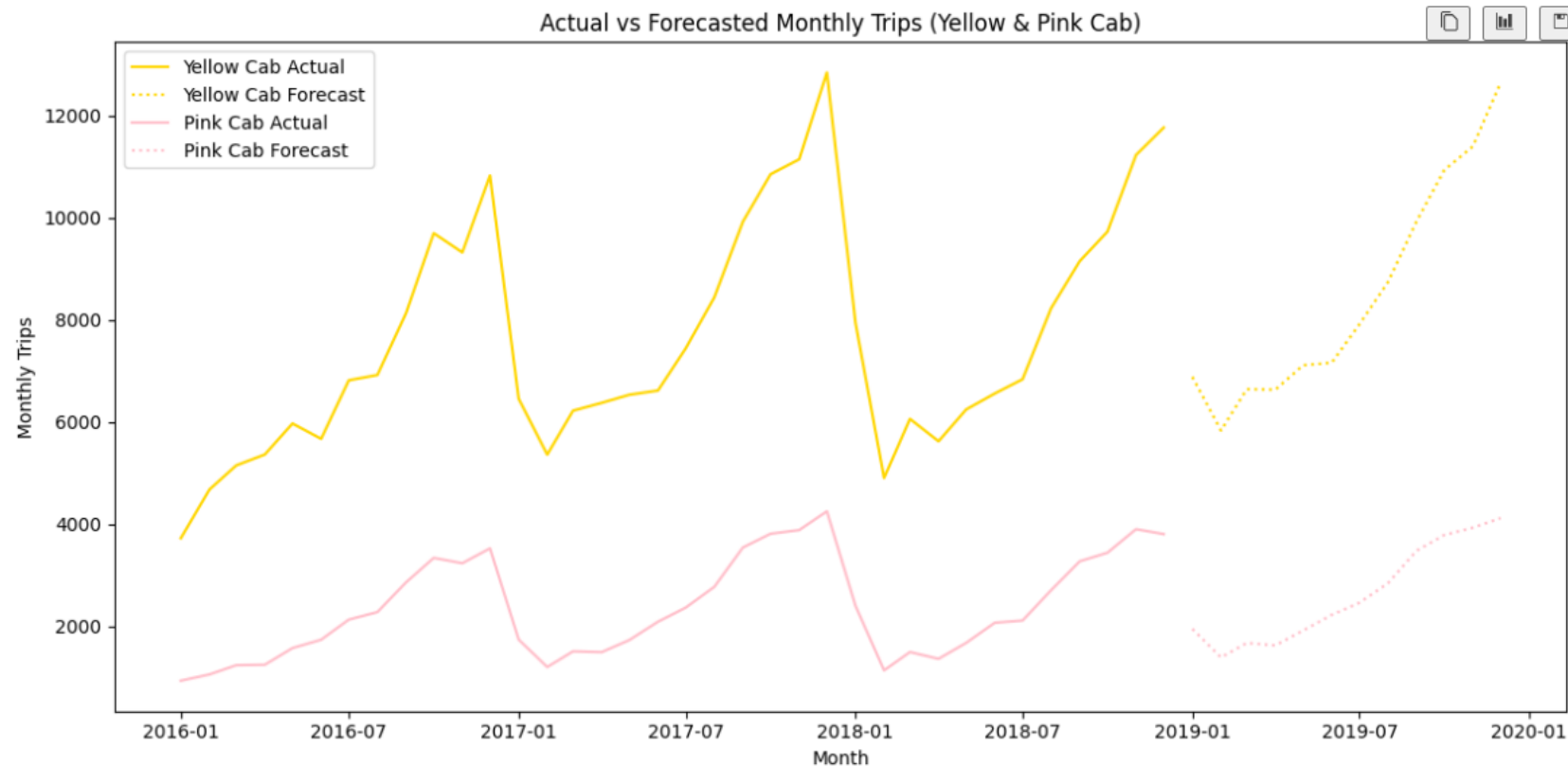
Key Insights & Recommendation

Strengths (Why Yellow Cab)

- ~3× more trips & ~5× revenue vs. Pink (2016–2018)
- Higher profit margins, scales better with demand
- Stronger city presence (NYC, Chicago, DC, LA, Boston)
- Broader customer base (age, income, payment mix)

Risks (What to Monitor)

- Margin variability across cities
- Seasonal dips (February troughs)
- Payment mix differences (cash-heavy cities)

Opportunities (Growth Levers)

- Optimize pricing in high-demand cities
- Deepen loyalty in 25–44 age group
- Expand digital adoption (card payments)
- Target senior customers where Yellow already has reach

Recommendation: Invest in Yellow Cab – stronger scale, higher profitability, broader reach, and greater resilience compared to Pink Cab.

Thank You