OPTIMIZING PUBLIC TRANSIT OPERATIONS

An Exploratory Data Analysis of Public Transportation

JUSTINA AGBLO

MetroMove
TRANSIT SOLUTIONS

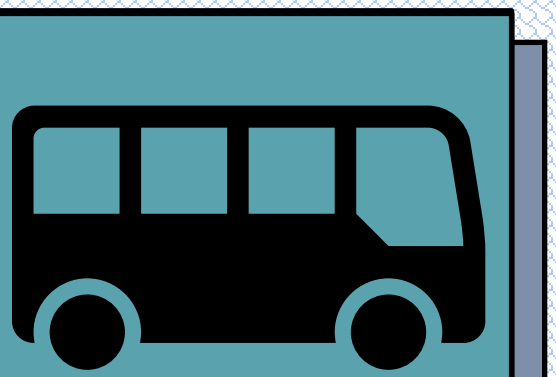
Business Introduction

- Public transportation provider operating in multiple cities
- Manages and analyzes trips by bus, train, ferry, tram
- Mission: efficient, affordable, timely public transport
- Wants to leverage data to improve passenger experience

Problem Statement

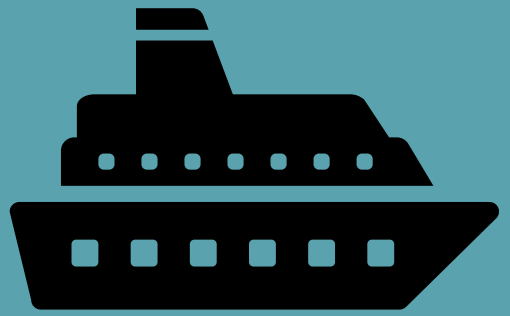
- Lots of trip data but limited insight
- Data is messy / inconsistent / incomplete
- They want to clean, explore, and summarize records
- Final aim: identify inefficiencies and patterns

Why This Analysis Matters

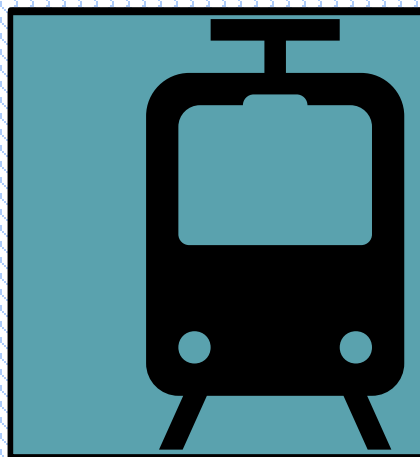


Understand passenger usage patterns

Evaluate performance of each transport mode



See how trip characteristics affect revenue



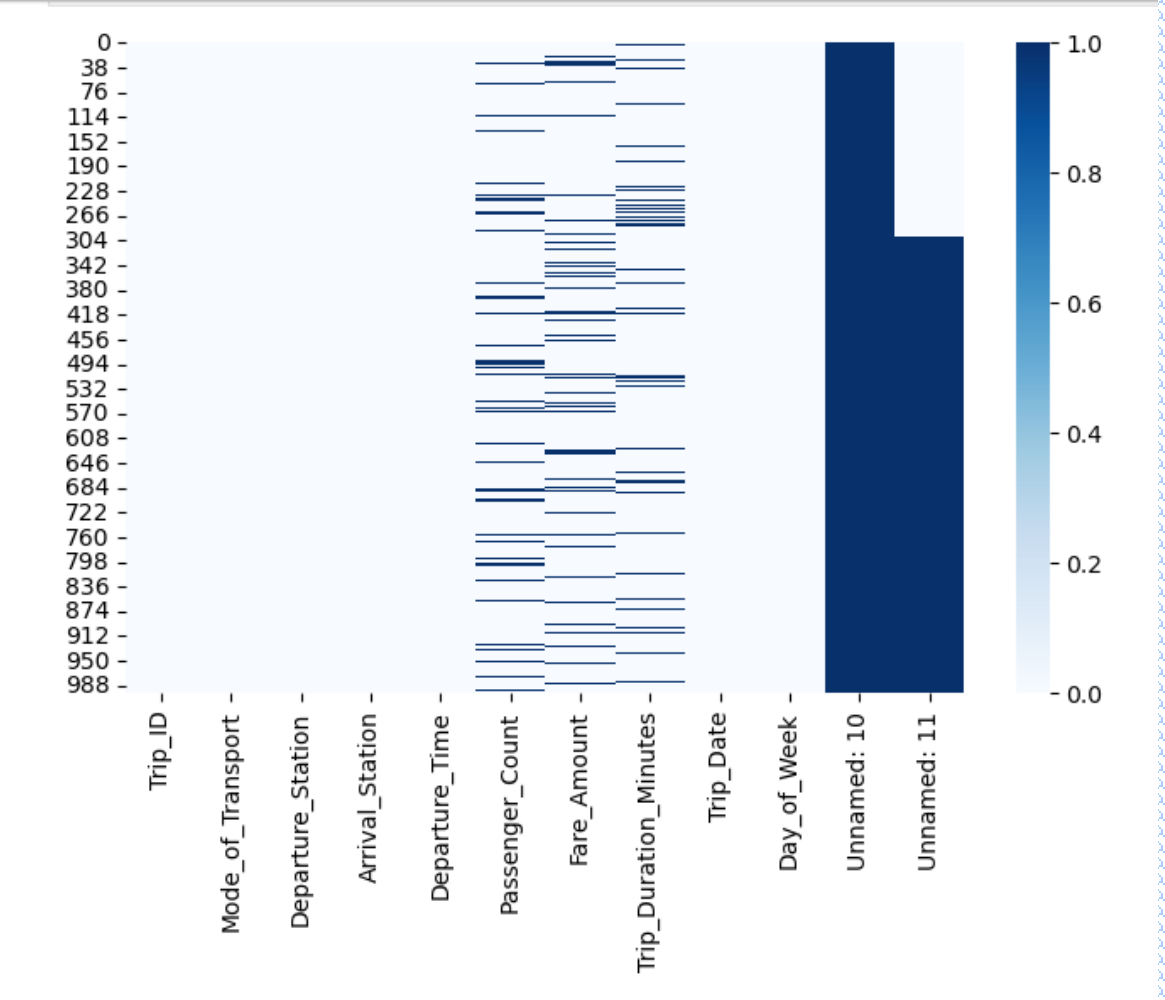
Practice real-world data cleaning + EDA + communication

Data Description

- 1,000 trip records
- 3 key numeric fields: Passenger_Count, Fare_Amount, Trip_Duration_Minutes,
- 4 key categorical fields: Mode_of_Transport, Day_of_Week, Departure_Station, Arrival_Station
- 1 column was empty

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Trip_ID                1000 non-null   object
 1   Mode_of_Transport      1000 non-null   object
 2   Departure_Station      1000 non-null   object
 3   Arrival_Station        1000 non-null   object
 4   Departure_Time         1000 non-null   datetime64[ns]
 5   Passenger_Count        900 non-null    float64
 6   Fare_Amount            900 non-null    float64
 7   Trip_Duration_Minutes  900 non-null    float64
 8   Trip_Date              1000 non-null   datetime64[ns]
 9   Day_of_Week            1000 non-null   object
 10  Unnamed: 10            0 non-null      float64
 11  Unnamed: 11            299 non-null    object
dtypes: datetime64[ns](2), float64(4), object(6)
memory usage: 93.9+ KB
```

Data Quality – Missing Values



- Passenger_Count, Fare_Amount, Trip_Duration_Minutes → 100 missing each (10%)
- Unnamed: 10 and Unnamed: 11 → entirely missing
- Time fields have no missing values
- empty → to be dropped

Data Cleaning Steps

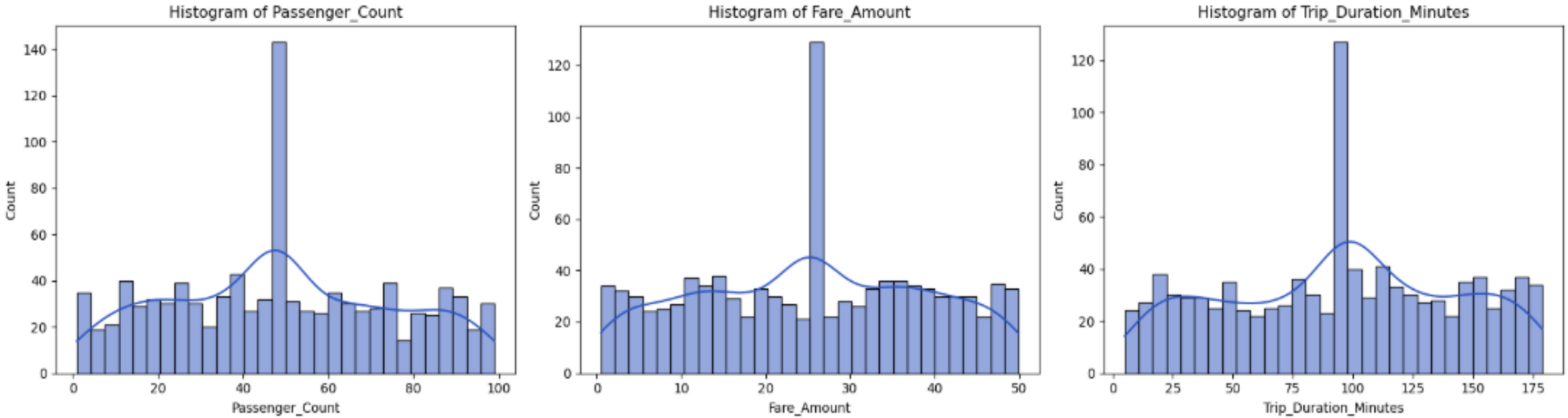
- Dropped: Trip_ID, Unnamed: 10, Unnamed: 11
- Filled numeric columns with median
- Extracted Hour and Day_of_Week
- Created Time_of_Day
- Created Revenue = Passenger_Count × Fare_Amount
- Standardized Mode_of_Transport to BUS / FERRY / TRAIN / TRAM

Numerical Summary

- Passenger_Count: median 48, max 99
- Fare_Amount: median $25, max $50
- Trip_Duration: median 97.5 mins, max 179 mins
- Revenue: median $984, max $4,826
- Max ≫ 75th percentile → a few high trips → possible outliers

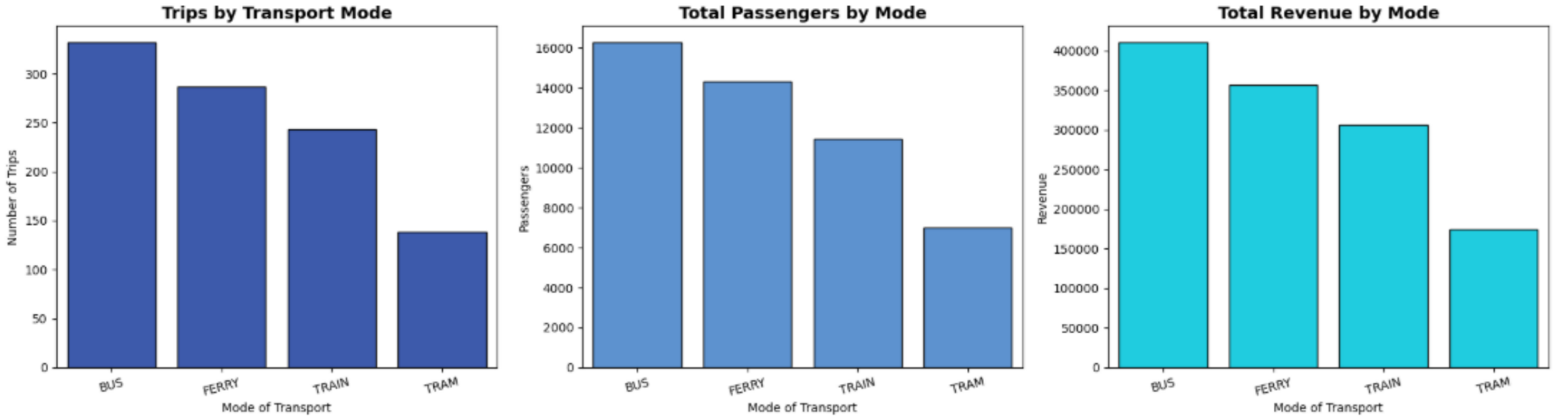
	Passenger_Count	Fare_Amount	Trip_Duration_Minutes	Revenue
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	49.039000	25.365053	94.593000	1248.668177
std	26.277709	13.721526	48.043673	1027.743269
min	1.000000	0.500576	5.000000	4.191420
25%	28.000000	13.917364	55.000000	393.404987
50%	48.000000	25.403856	97.500000	983.888969
75%	70.000000	36.580122	132.250000	1914.469950
max	99.000000	49.945184	179.000000	4826.548112

Univariate Analysis – Trip Metrics



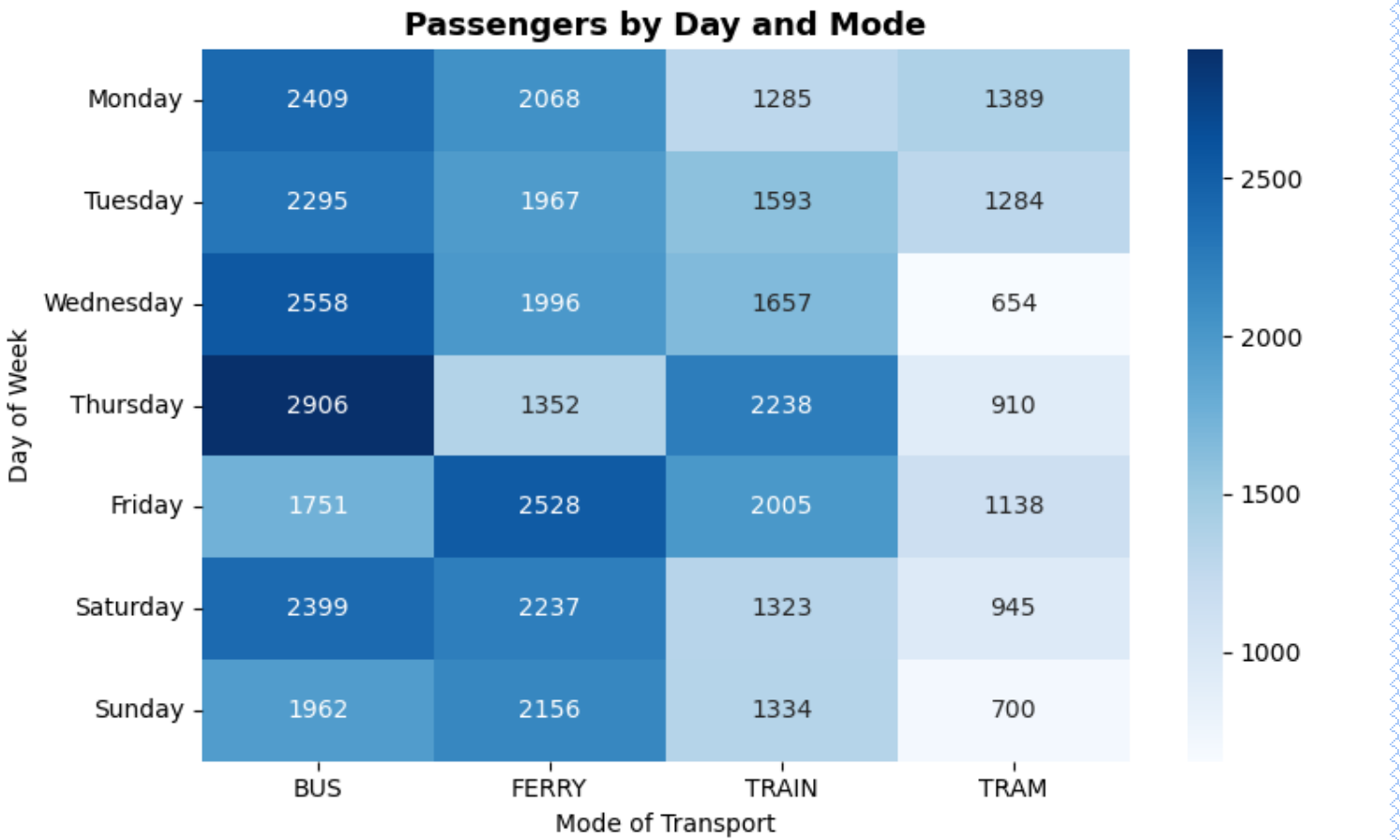
- All three metrics do not have any strong left or right skewness
- The tall spikes suggest that most trips cluster around the median
- A few very large trips stretch the distribution

Which Transport Mode Leads?



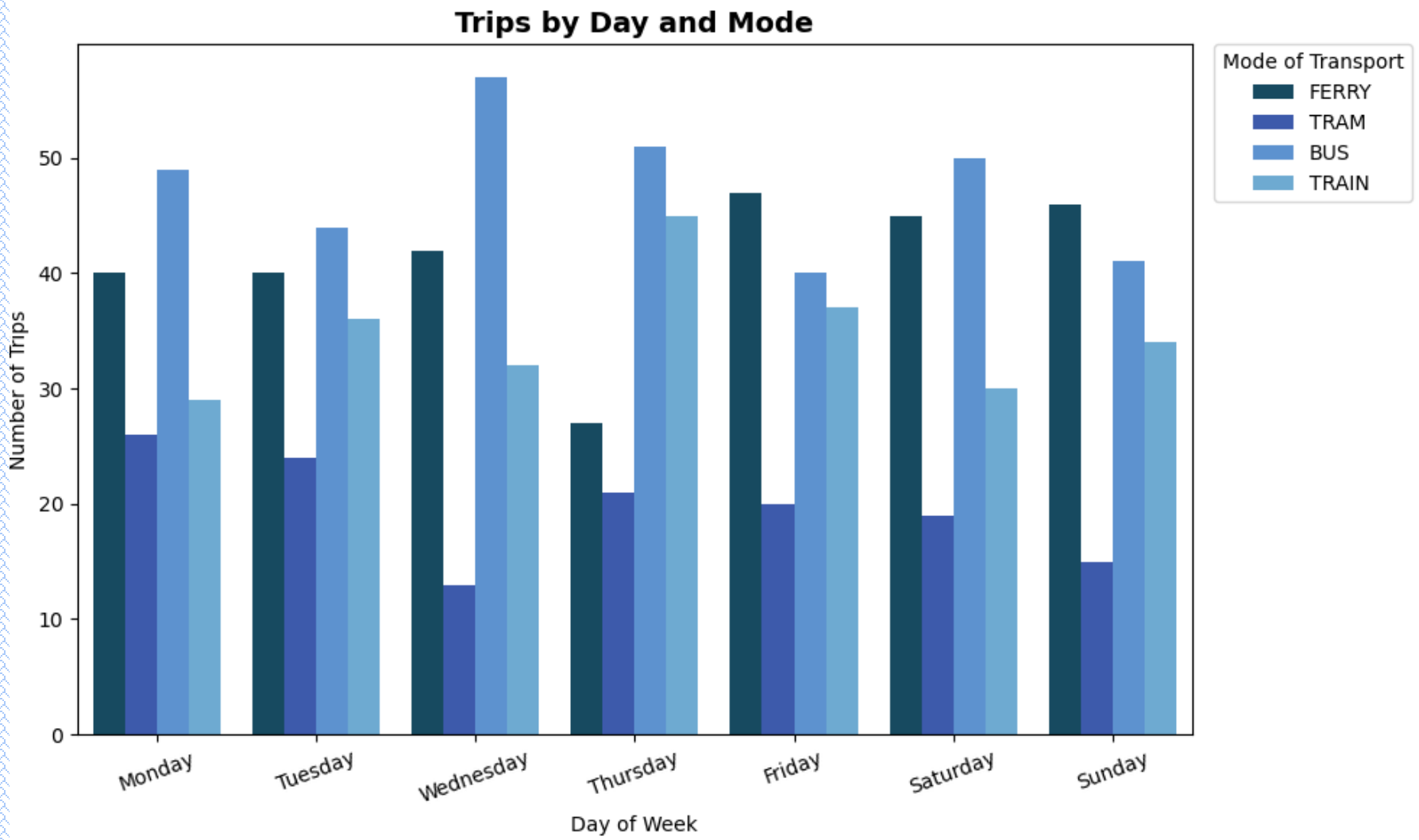
- BUS is #1 for trips, passengers and revenue
- FERRY and TRAIN are mid-tier
- TRAM is consistently lowest.

When Are People Travelling?



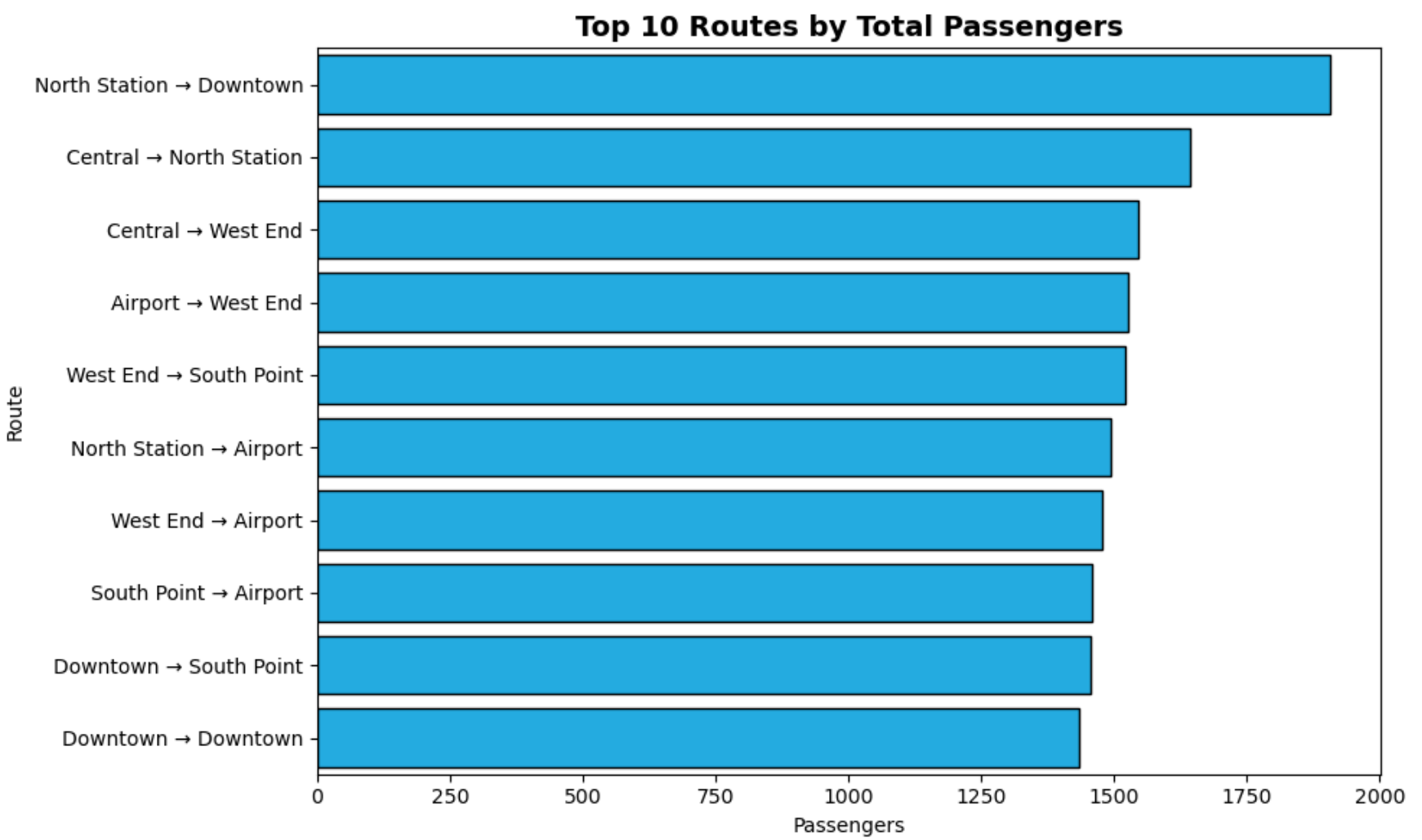
- Thursday and Wednesday are heavy days
- Bus dominates most days of the week
- Tram is light across the week

Trips by Day & Mode



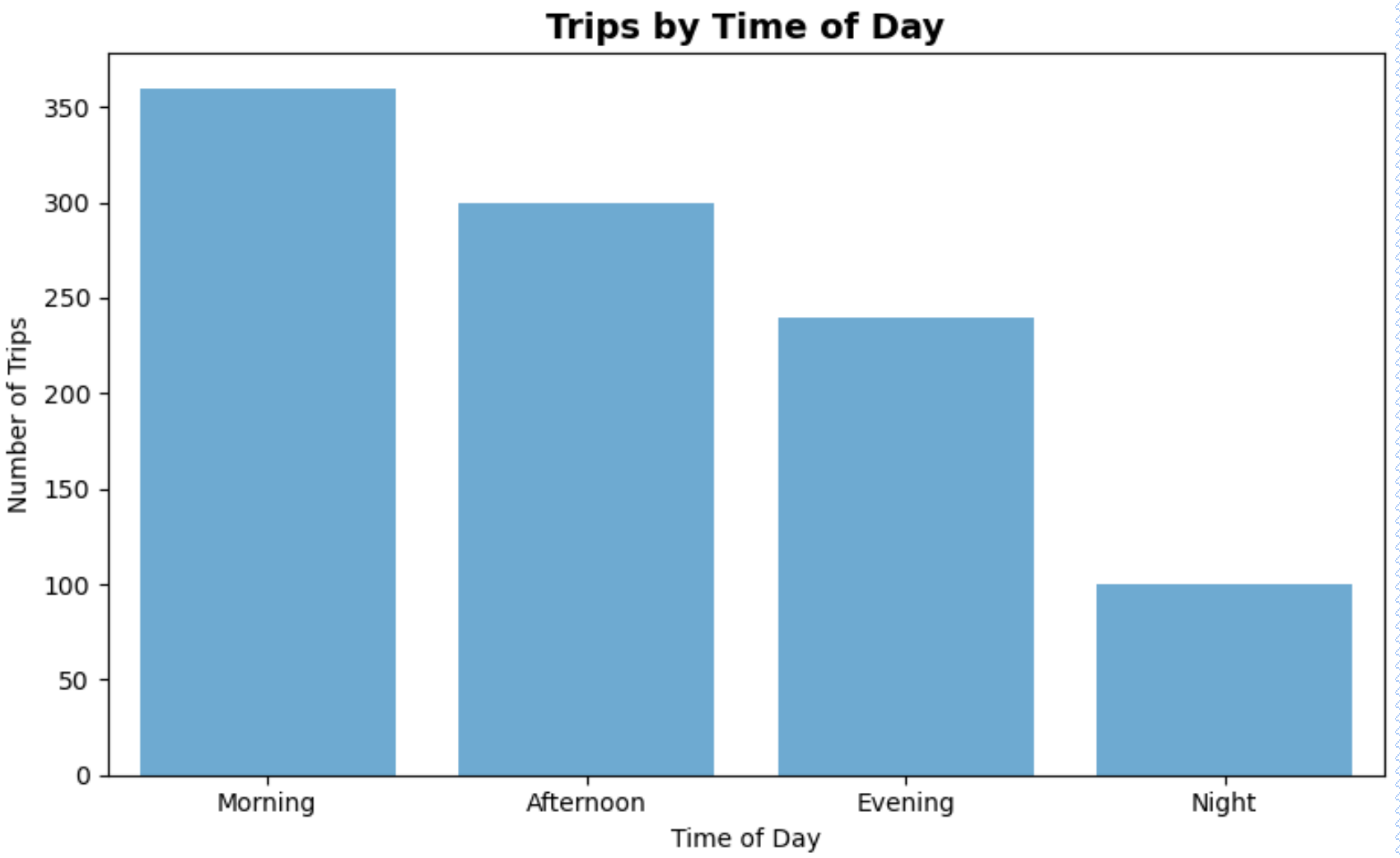
- Ferry & Bus stay relatively strong across the week
- Wednesday → spike for Bus
- Weekend volume drops but doesn't disappear

Top 10 Routes by Passengers



- North Station → Downtown is the busiest route
- Routes involving Central and Airport are repeatedly in top 10
- These are the priority corridors for optimization

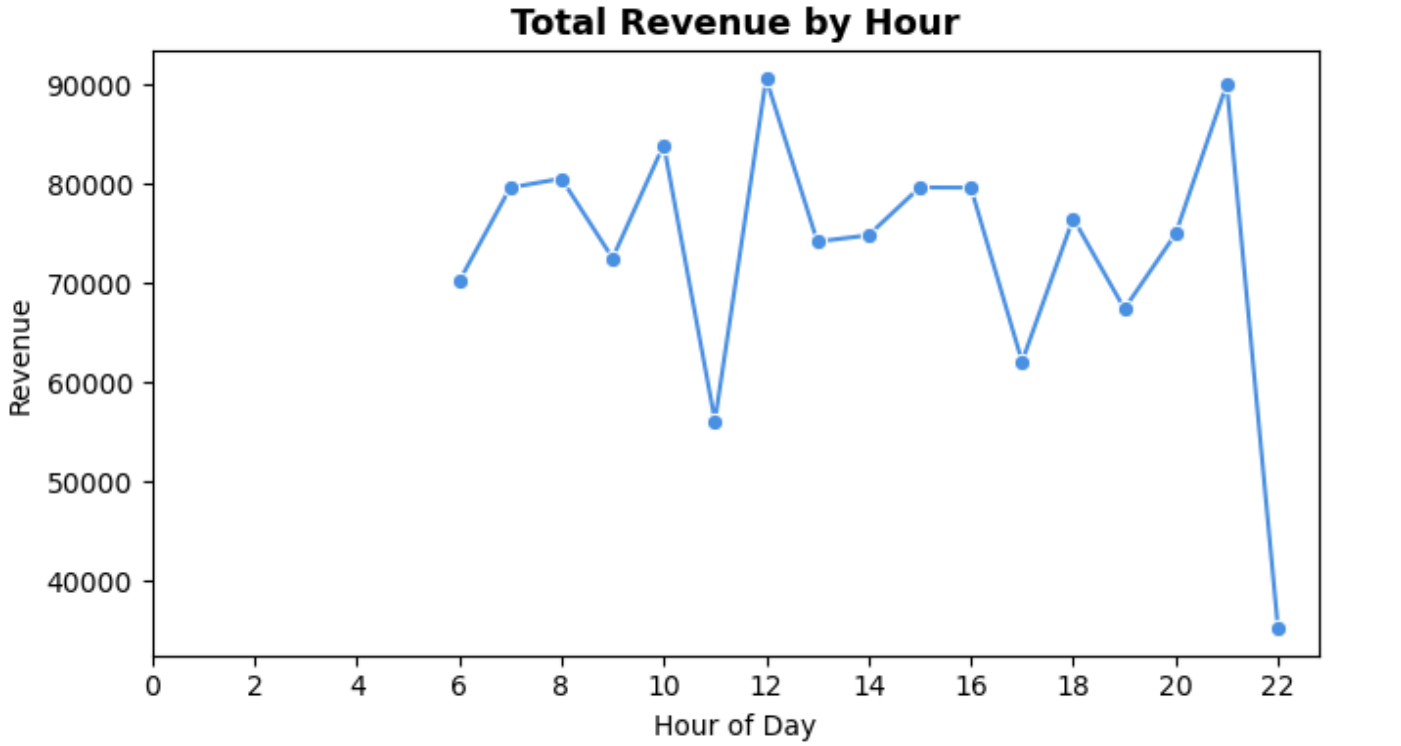
Trips by Time of Day



- Morning (360) and Afternoon (300) dominate
- Evening is moderate
- Night is low (100)

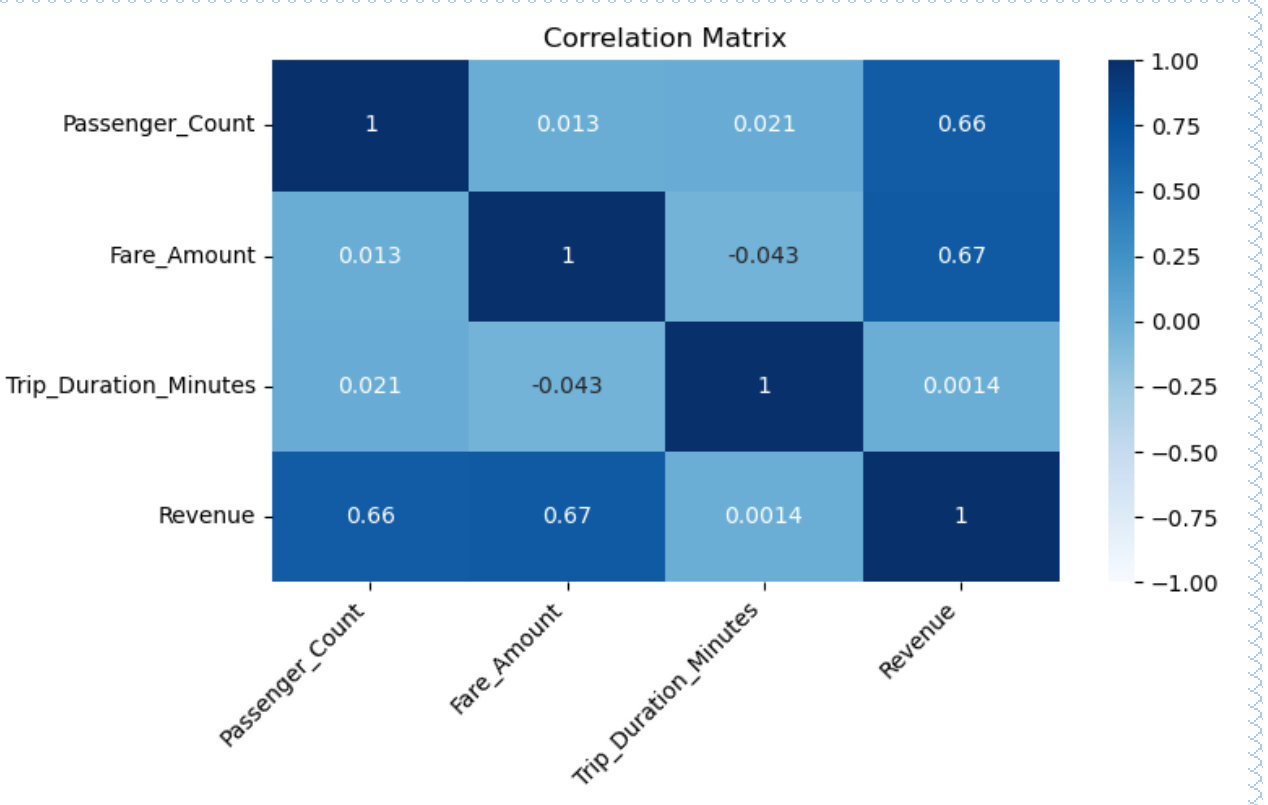
Revenue Drivers

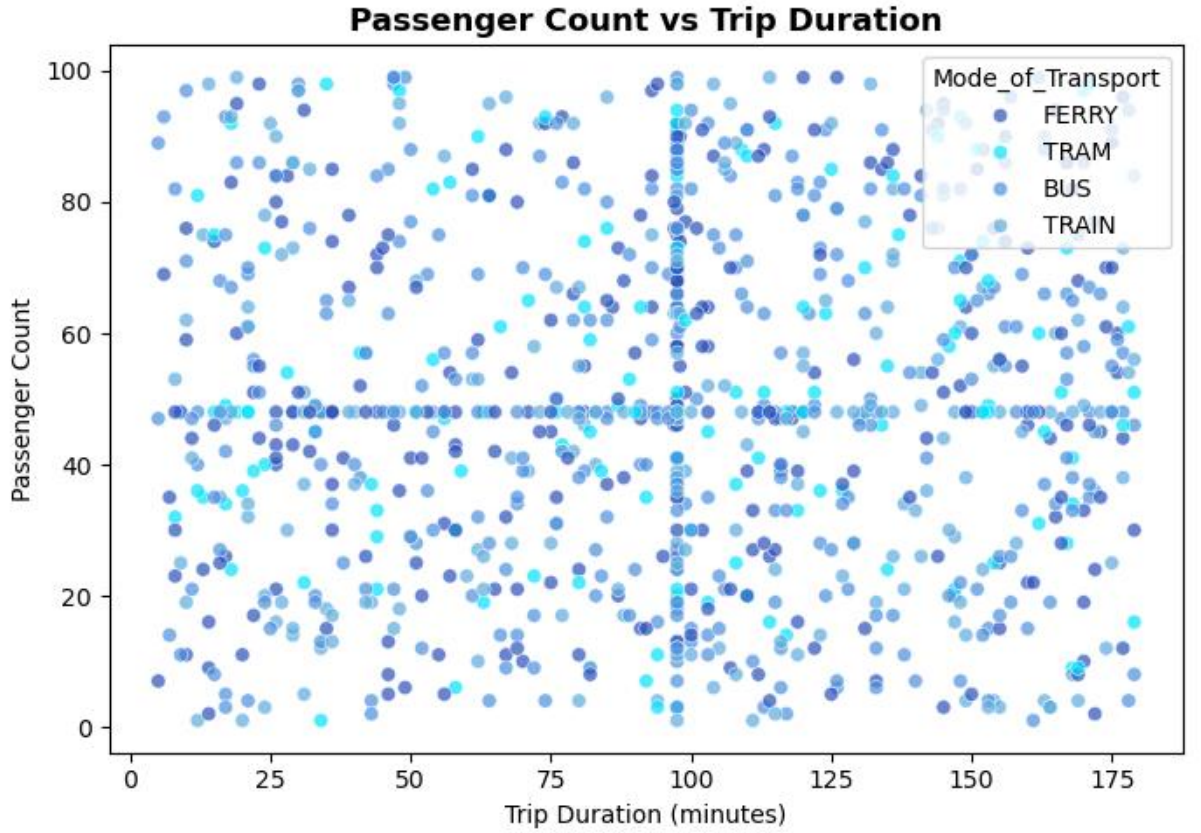
	Mode_of_Transport	Revenue
0	BUS	410929.647242
1	FERRY	356903.111660
2	TRAIN	306669.390974
3	TRAM	174166.026843



- Bus is the top revenue generator
- Revenue is not flat over the day → mid-day/late-day spikes
- Combine this with time-of-day to plan pricing / frequency

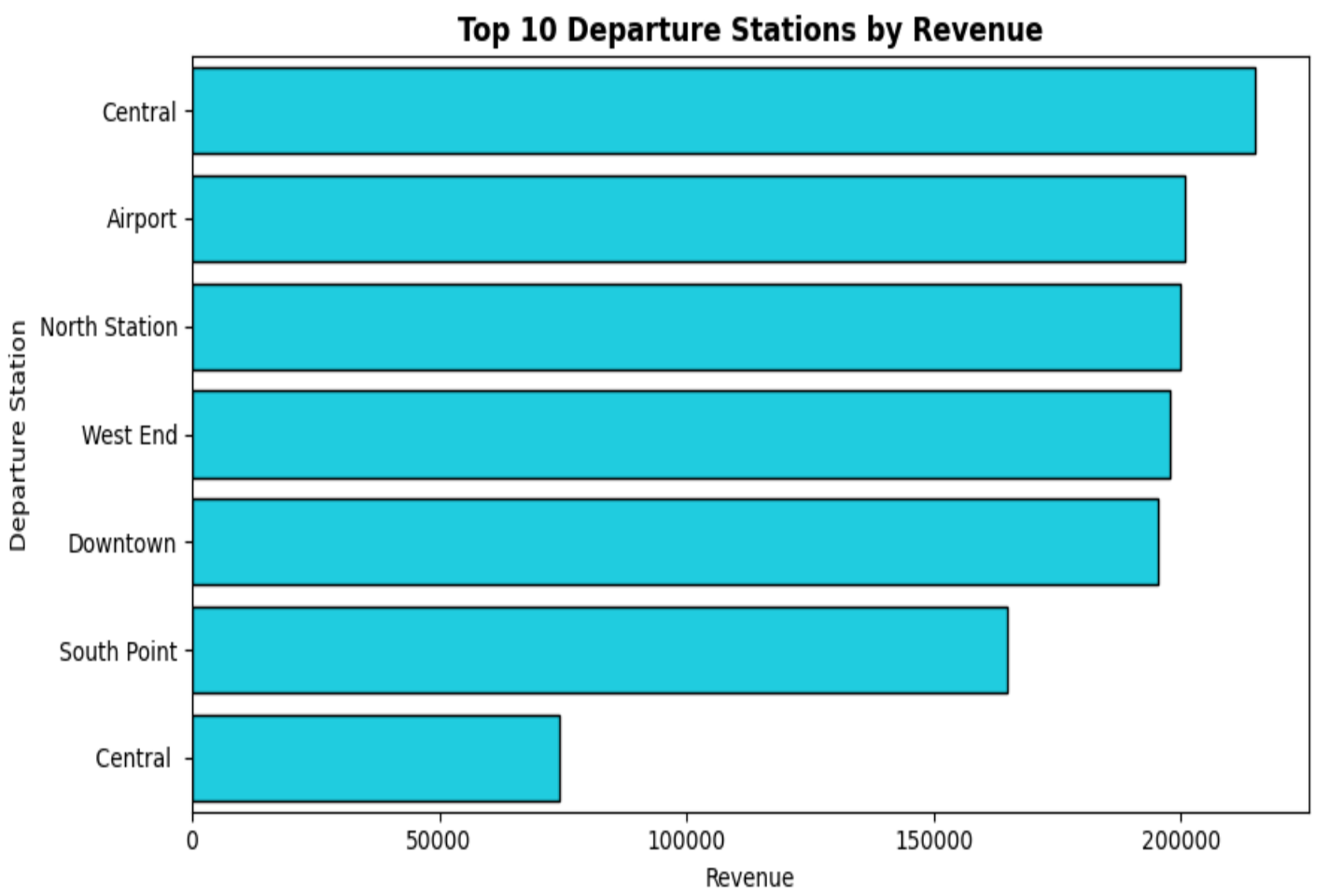
Relationships in the Data





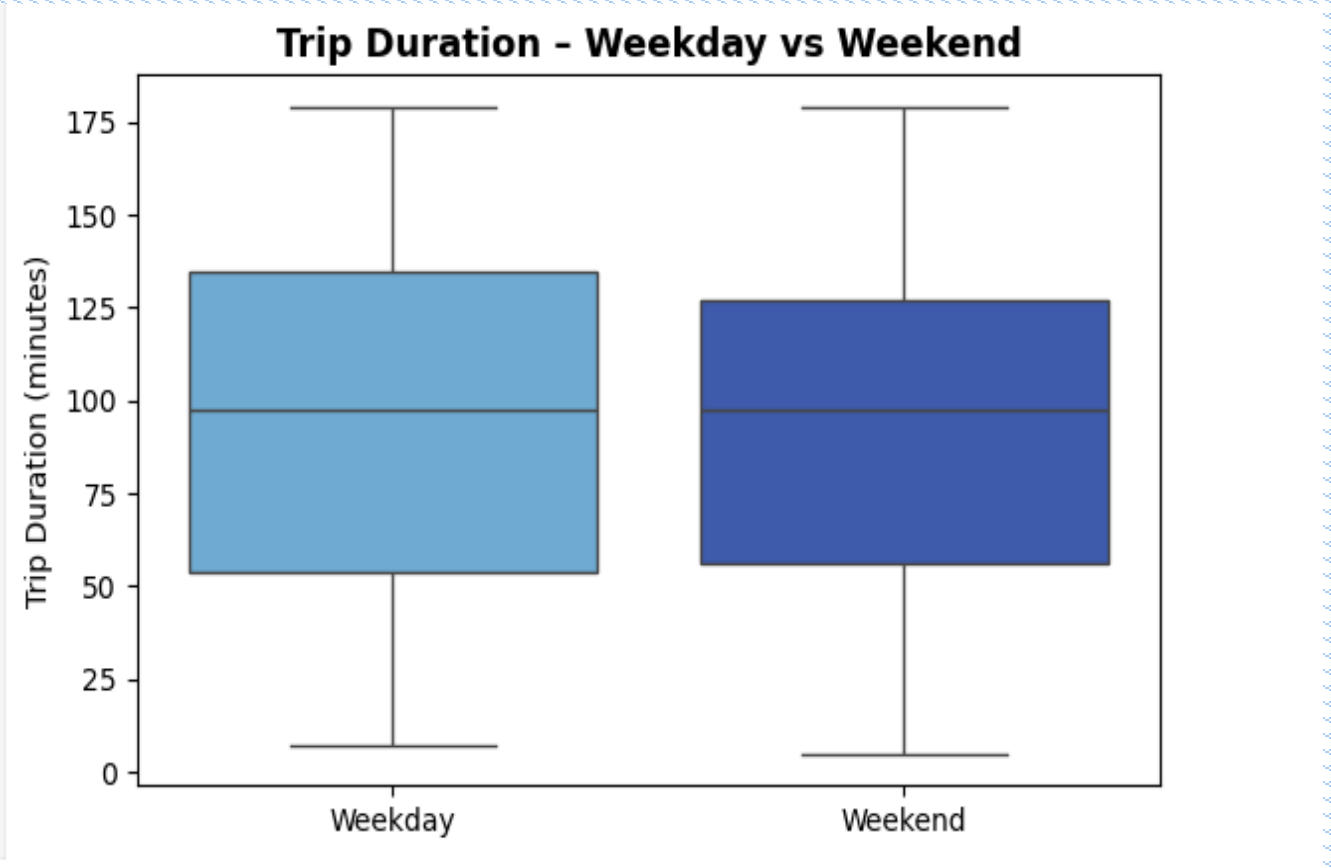
- Revenue correlates strongly with Passenger_Count and Fare_Amount (0.66–0.67)
- Trip duration has weak relationship with revenue
- Passenger counts stay fairly stable across trip lengths (no strong correlation), and this pattern is consistent for all transport modes.

Top Departure Stations by Revenue



- Central is the most valuable departure station
- Airport and North Station follow closely
- These stations justify better service or priority routes

Trip Duration – Weekday vs Weekend



- Weekends show slightly wider spread
- But median durations are similar → service time is consistent
- No outliers

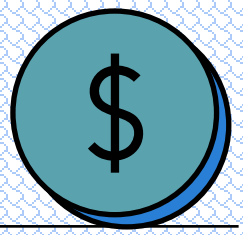
Insights

Bus is the backbone → highest trips, passengers, revenue

Demand peaks on weekday mornings/afternoons

North Station – Downtown and Central/Airport routes are the busiest

A few high-revenue / long trips exist → outliers, monitor separately



Revenue is driven more by volume + fare than by duration

Recommendations

Prioritize capacity planning for buses on busy weekdays

Create outlier reports for very long / very high-revenue trips

Focus revenue optimization on boosting passenger volume and fare strategies

Monitor and optimize top 10 routes + top departure stations

Consider time-of-day scheduling (reduced night, stronger morning)

Thank you !!!