RIDERSHIP PATTERN ANALYSIS USING EXPLORATORY DATA ANALYSIS IN R FOR A U.S. METROPOLITAN TRANSIT AUTHORITY

1. Background

A city transit authority in the Pacific Northwest was reviewing its post-pandemic ridership performance. While raw boarding counts were available, there was no consolidated analysis to identify when, where, and among whom ridership had shifted.

We were brought in to perform an exploratory data analysis (EDA) using R to help reveal patterns in rider behavior by route, daypart, and rider type. These insights supported route optimization, schedule redesign, and fare policy adjustments.

2. Objective

- To identify high- and low-traffic bus and rail routes based on timestamped boarding data
- To detect time-of-day usage trends and demographic ridership shifts
- To generate recommendations for service optimization and targeted transit improvements

3. Data Used

Source: Automated fare collection + smart card tap-in records (Jan–Dec 2023)

Dataset Details:

- 4.8 million ride records
- Key fields:
  - Rider_ID, Route_ID, Mode (Bus/Light Rail), Station_ID, Tap_Time, Fare_Type, Age_Bracket, Zip_Code, Ride_Duration_Min

Preprocessing in R:

- Extracted Hour, Day_of_Week, and Peak_Offpeak_Flag using lubridate
- Cleaned anomalies (e.g., negative durations, missing tap-outs)
- Created weekday vs. weekend segments, and grouped demographic fields

statssy

4. EDA Methodology

4.1 Temporal Analysis

- Line plots of daily and hourly ridership with ggplot2
- Peak load periods visualized using heatmaps: Hour × Day_of_Week
- Seasonal analysis by quarter with facet_wrap

4.2 Route and Station-Level Analysis

- Ranked routes by total riders, peak riders, and weekend usage
- Mapped tap-ins geographically using ggmap and sf (for station locations)

4.3 Demographic Segment Analysis

- Cross-tabulated Fare_Type × Age_Bracket × Mode
- Boxplots of Ride_Duration_Min by Age_Bracket
- Identified underutilized services by ZIP code

5. Key Findings

Area	**Insight**
Peak Time Load	6:45–9:00 AM and 4:30–6:00 PM remain core peaks, but with 14% lower volume than 2019
Route-Specific Trends	Route 5 and Route 12 exceeded pre-pandemic volume by 18% and 22% respectively
Underused Services	Weekend late-night rail under 30% utilization across 3 corridors
Demographic Shifts	Riders aged 18–25 increased use of tap-and-go cards by 35% in Q3
Location Gaps	3 ZIP codes with population growth >10% showed no added frequency or stops

6. Interpretation and Recommendations

- **Realign schedules** for low-use late-night services and reinvest into growing corridors
- Increase frequency on **Route 12** to reduce crowding; piloted headway reduction on weekdays

Statssy

- Target outreach and discounts to young riders using **tap-and-go** cards
- Propose micro-transit or on-demand pilots in 3 underserved ZIP codes
- Provide quarterly data snapshots to planners via **R Markdown automation**

7. Reporting Output

- **R Markdown Report (PDF, 31 pages)**:
  - Route-level visual summaries
  - Time heatmaps and rider demographics
  - Interactive HTML version for internal dashboards
- **Excel Summary Files**:
  - Route-level KPIs
  - ZIP code level rider statistics
  - Pivot tables for internal planning workshops
- **Shiny App (Prototype)**:
  - Filter by route, day, time, and age group
  - Displays top/bottom usage hours with chart exports

8. Institutional Impact

- Informed **2024 transit plan** with schedule changes on 9 routes
- Identified **$1.1M in low-yield services** reallocated to growing demand corridors
- Demographic findings fed into **city-wide student fare pilot**
- Transit agency adopted the R dashboard format for **quarterly operations review**

statssy