Urban Bike-Sharing Chicago Analysis

Chicago Divvy 2023 — Final Report

Luisa Johanna Kaczmarek · Student ID: 16242

Emerging Topics in Data Analytics & Management

February 2026

Session 5: Data Collection & Exploration

This session establishes the analytical foundation by loading the full-year 2023 Chicago Divvy bike-sharing dataset, performing systematic data quality assessment, and producing initial visualisations of temporal and spatial patterns.

1. Data Collection

Data was sourced from two systems:

- Divvy trip data (2023): 12 monthly CSV archives downloaded from the official Divvy S3 bucket, covering January–December 2023.
- CityBikes API: Real-time station metadata fetched from api.citybik.es/v2/ to provide station-level geographic reference.

Source	Records	File Size	Coverage
Divvy Trip Data (2023)	5,719,877	~500 MB	Jan–Dec 2023, Chicago
CityBikes API	Snapshot	< 1 MB	Real-time station metadata

Each record contains: ride ID, bike type (classic / electric / docked), start and end timestamps, station names and IDs, GPS coordinates, and rider type (member / casual).

2. Data Quality Assessment

2.1 Missing Values

A systematic missing-value audit was conducted across all columns. The pattern of missingness is not random — it is strongly correlated with electric (dockless) bike usage.

Column	Missing Count	Missing %	Action Taken
end_station_id	929,343	16.25%	Flag as dockless trip; retain for temporal analysis
end_station_name	929,202	16.25%	Flag as dockless trip; retain for temporal analysis
start_station_id	875,848	15.31%	Flag as dockless trip; retain for temporal analysis
start_station_name	875,716	15.31%	Flag as dockless trip; retain for temporal analysis
end_lat / end_lng	6,990	0.12%	Retain; exclude for spatial analysis only

Key insight: Missing station identifiers are concentrated in summer months and electric bike rides, confirming these represent dockless trips rather than data corruption. Both start and end

station IDs are missing for 7.29% of all trips (417,137 records). These are flagged with a "dockless_trip" binary column rather than dropped.

2.2 Outlier Detection

Three categories of outliers were identified and handled:

Trip Duration Outliers: Trips with negative durations, durations < 60 seconds (false starts), or > 24 hours (unreturned bikes) were removed. Percentile analysis confirmed the thresholds:

Percentile	Duration (seconds)	Duration (minutes)
1st	16 s	0.27 min
5th	127 s	2.12 min
95th	2,480 s	41.3 min
99th	5,926 s	98.8 min

Result: 156,033 rows removed (2.73%), leaving 5,563,844 clean trip records.

Spatial Outliers: 6,993 trips (0.12%) had end-coordinates outside the Chicago bounding box, at (0,0), or null. These occur exclusively at trip endpoints (not origins), consistent with GPS failure rather than corruption. Flagged and excluded from spatial analyses only.

Temporal Anomalies: Zero temporal anomalies detected — no future timestamps, no records before the 2013 system launch, and no duplicate ride IDs.

2.3 Cleaning Summary

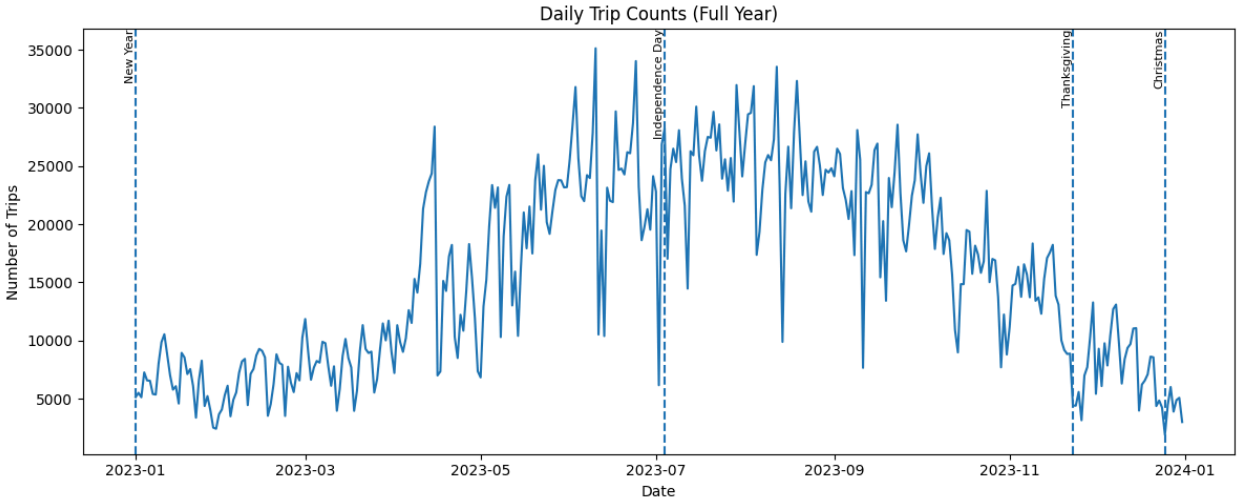
Metric	Value
Raw records	5,719,877
Duration outliers removed	156,033 (2.73%)
Records after cleaning	5,563,844
Dockless trips flagged	417,137 (7.29%)
Spatial flags (end-point only)	6,993 (0.12%)

3. Initial Visualisations

3.1 Daily Trip Counts — Full Year

Aggregating trips by calendar date reveals the strong seasonal envelope of Chicago bike-share demand.

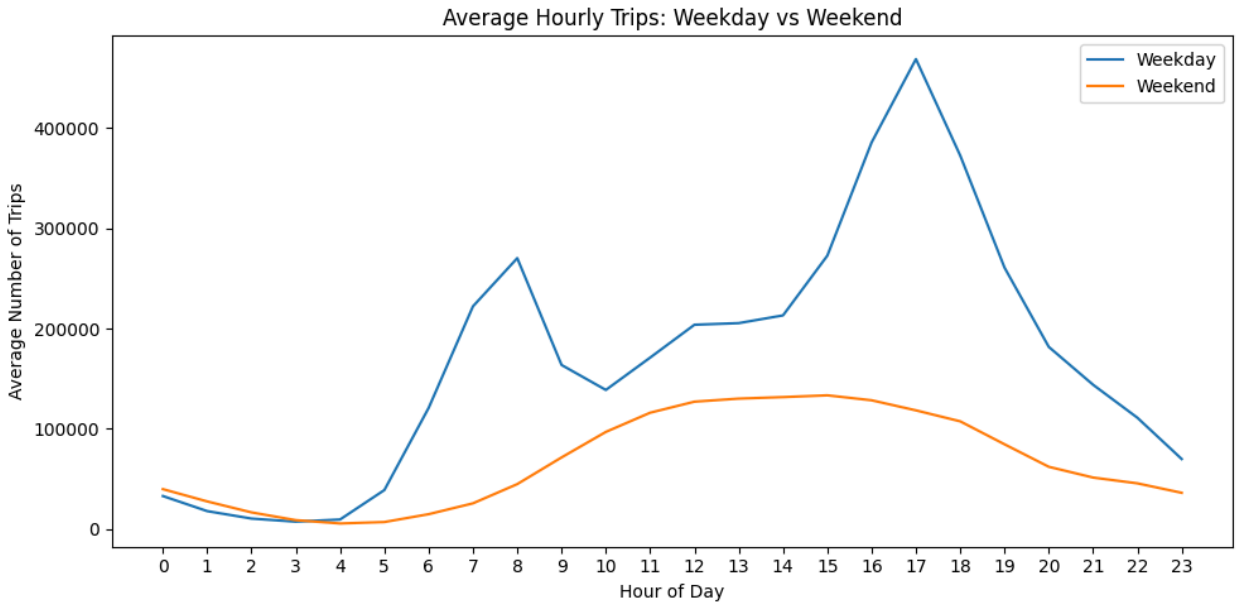
Figure 1: Daily trip counts (2023) with major holiday markers



- Trip volume rises sharply from spring through summer, peaking in July–August.
- Winter months (January–February) have the lowest counts, driven by Chicago's harsh weather.
- Several sharp single-day drops are visible in summer months — these are investigated as anomalies in Session 6.
- Holiday markers (New Year, Independence Day, Thanksgiving, Christmas) coincide with dips.

3.2 Hourly Patterns: Weekday vs. Weekend

Figure 2: Average hourly trips — weekday vs. weekend

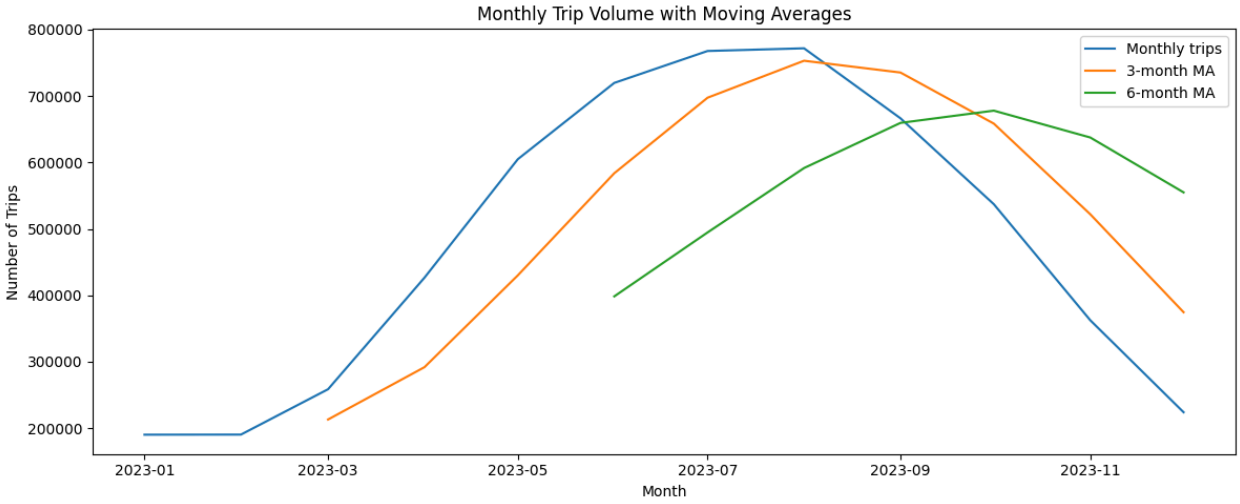


- Weekdays show two pronounced peaks at ~08:00 (morning commute) and ~17:00 (evening commute), confirming utilitarian usage.

- Weekend demand builds gradually from late morning, peaking around 13:00–16:00 — consistent with leisure and recreational use.
- Total weekday volume is higher than weekend, indicating member commuters drive the majority of system usage.

3.3 Monthly Trend with Moving Averages

Figure 3: Monthly trip counts with 3-month and 6-month moving averages

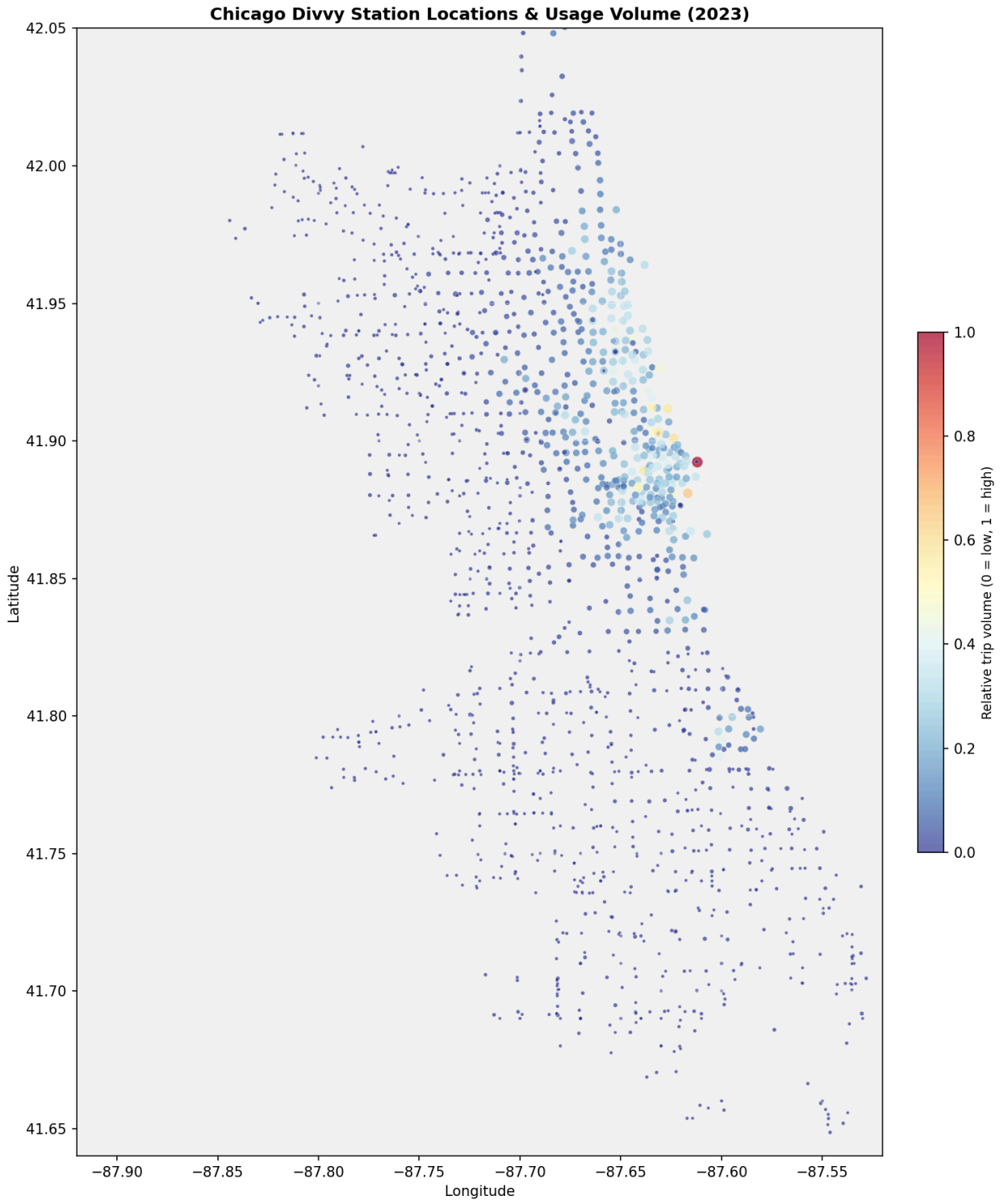


- Strong single-cycle annual seasonality: peak in July–August, trough in January–February.
- The 3-month moving average closely tracks short-term momentum; the 6-month MA confirms the single seasonal cycle.
- No long-term upward or downward trend within 2023 — variation is predominantly seasonal.
- Peak month (August) vs. trough (January): roughly 4× difference in demand.

3.4 Spatial Overview

Mapping station locations against trip volume reveals the concentration of Divvy infrastructure along the lakefront and central city corridors, with stark coverage gaps in the west and south sides of Chicago.

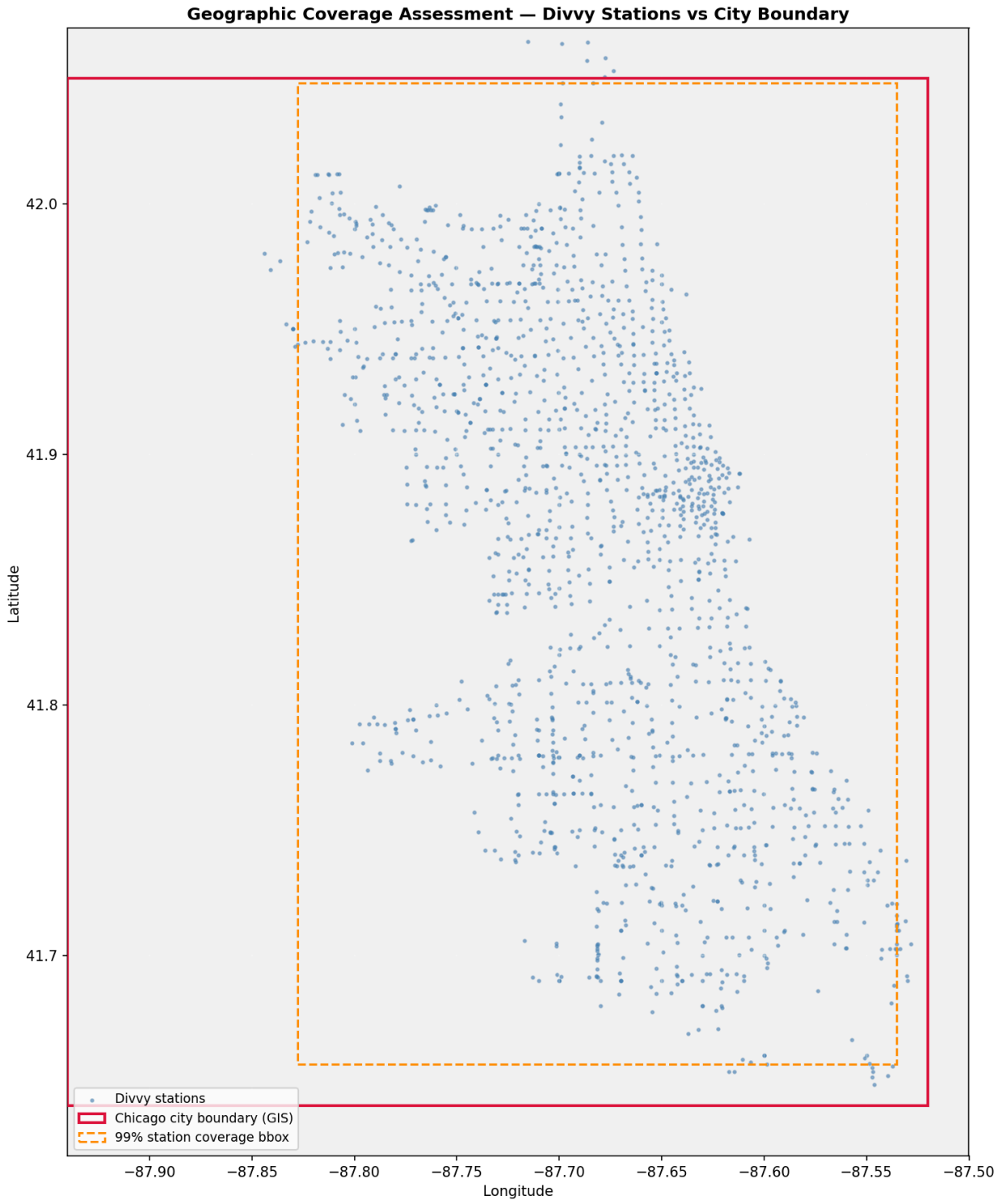
Figure 18: Chicago Divvy station locations and usage volume (2023) — colour and size represent relative trip volume



- Stations are densely clustered along the lakefront and the Loop, mirroring the highest-demand corridors. The colour gradient (blue = low, red = high volume) confirms that the busiest stations are concentrated in tourist and commuter zones.

- Large areas of the west and south sides have sparse or no station coverage, contrasting with trip activity in those areas. This spatial mismatch supports treating dockless trips as structurally valid mobility demand rather than data gaps.
- Network design appears optimised for tourism and high-traffic zones rather than citywide equity or coverage uniformity.

Figure 19: Geographic coverage assessment — Divvy stations vs Chicago city boundary (GIS bbox)



- The red boundary represents the official Chicago city boundary (GIS bbox). The dashed orange box outlines the 99th-percentile station coverage extent. The comparison shows that Divvy coverage falls well short of city limits, particularly on the south and west sides — confirming the equity gap identified in the missing values analysis.

3.5 Member vs. Casual User Patterns

Segmenting by rider type reveals structurally different usage behaviours.

Figure 4: Monthly trips by member vs. casual users

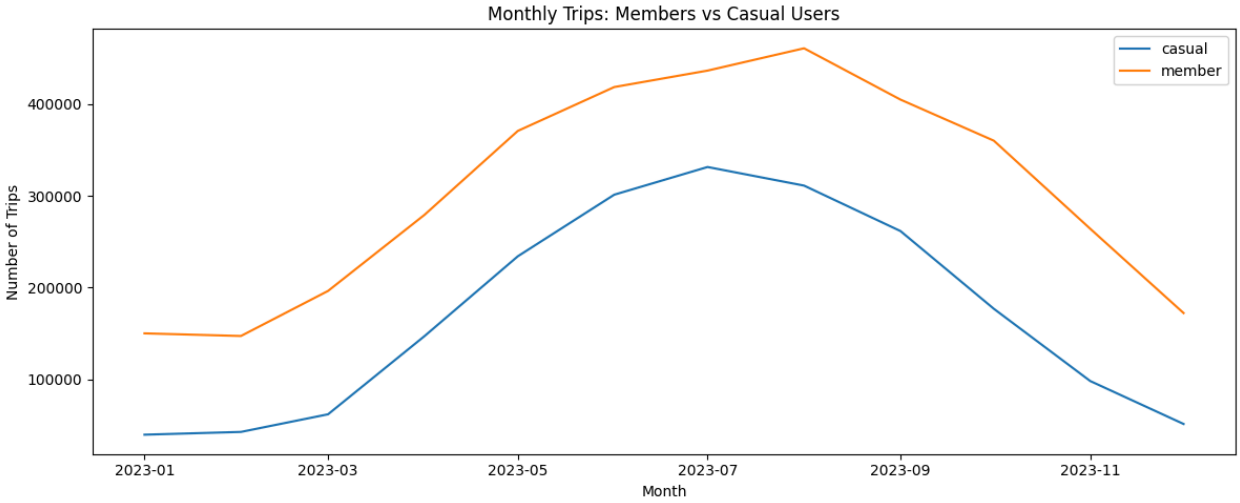


Figure 5: Hourly trip patterns by user type

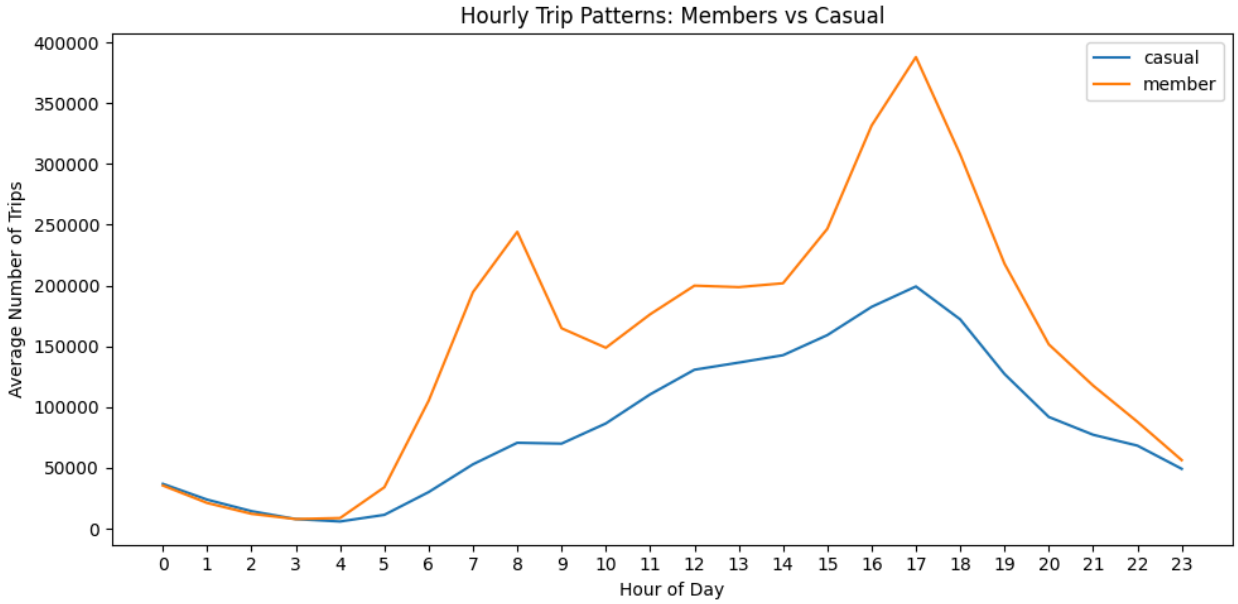
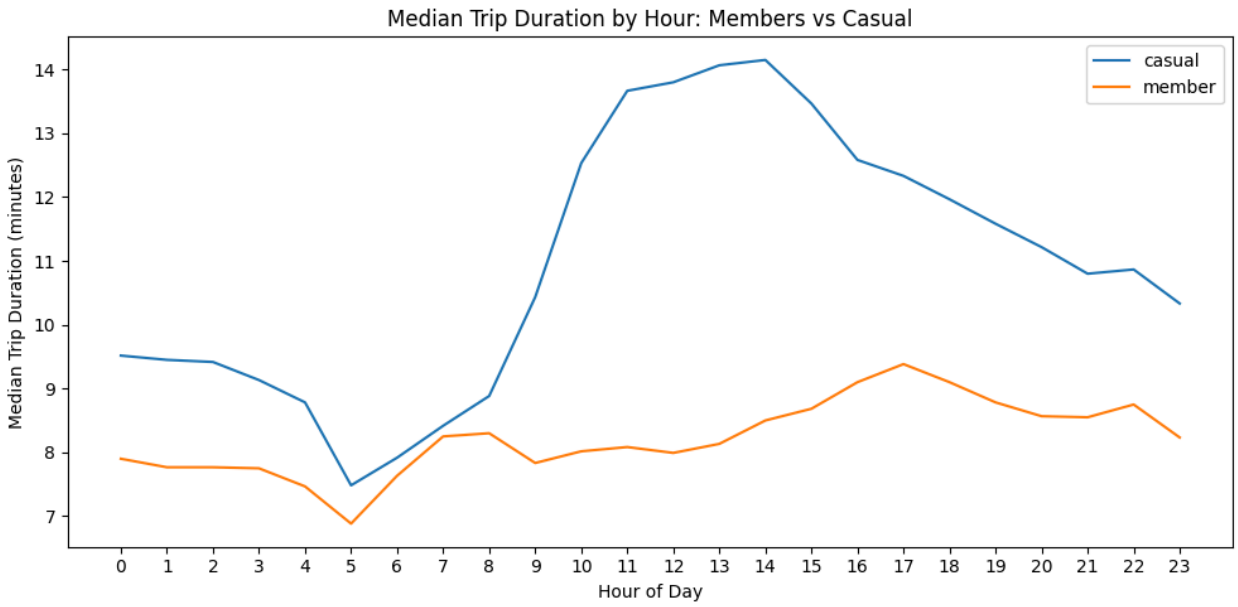


Figure 6: Median trip duration by hour and user type



- Members dominate total volume year-round and show significantly more stable demand across seasons.
- Casual usage is highly seasonal — near-zero in winter, surging in summer — reflecting tourist and leisure riders.
- Members have sharp commute-hour peaks; casual users have a smooth mid-day/afternoon pattern.
- Casual users take longer rides (higher median duration) throughout the day, consistent with exploration rather than point-to-point commuting.

Session 6: Time Series Analysis

Building on the cleaned dataset, Session 6 performs classical time series decomposition, multi-seasonality analysis, formal stationarity testing (ADF), and anomaly detection on the 2023 daily trip count series.

1. Data Preparation

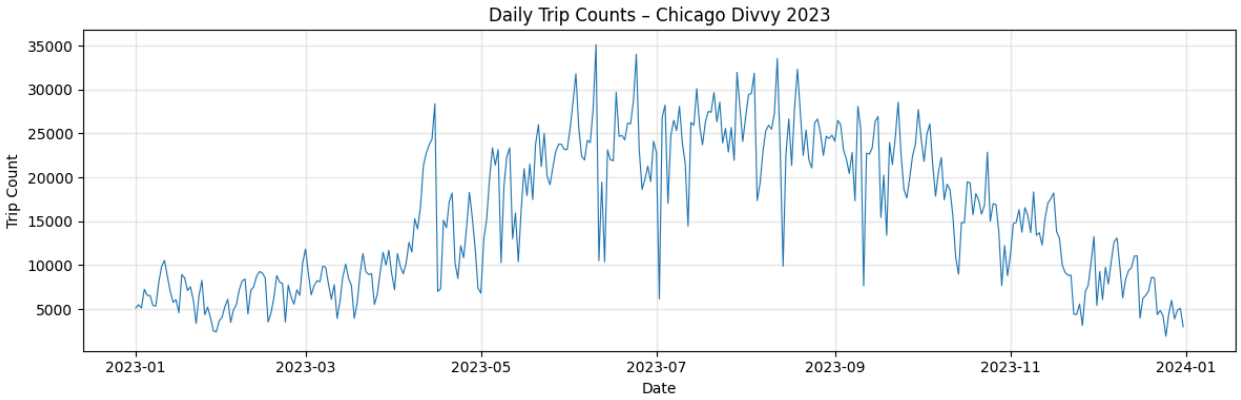
All 12 monthly files were concatenated and trips aggregated by calendar day. The resulting daily series spans exactly 365 days (1 January to 31 December 2023) with no missing dates and no NaN values.

Series Property	Value
Length	365 observations
Date range	2023-01-01 → 2023-12-31
Mean daily trips	15,671
Std deviation	8,235 trips/day
Missing dates	0
Timezone	America/Chicago (assigned)

2. Classical Decomposition

2.1 Raw Series

Figure 7: Raw daily trip count series — Chicago Divvy 2023

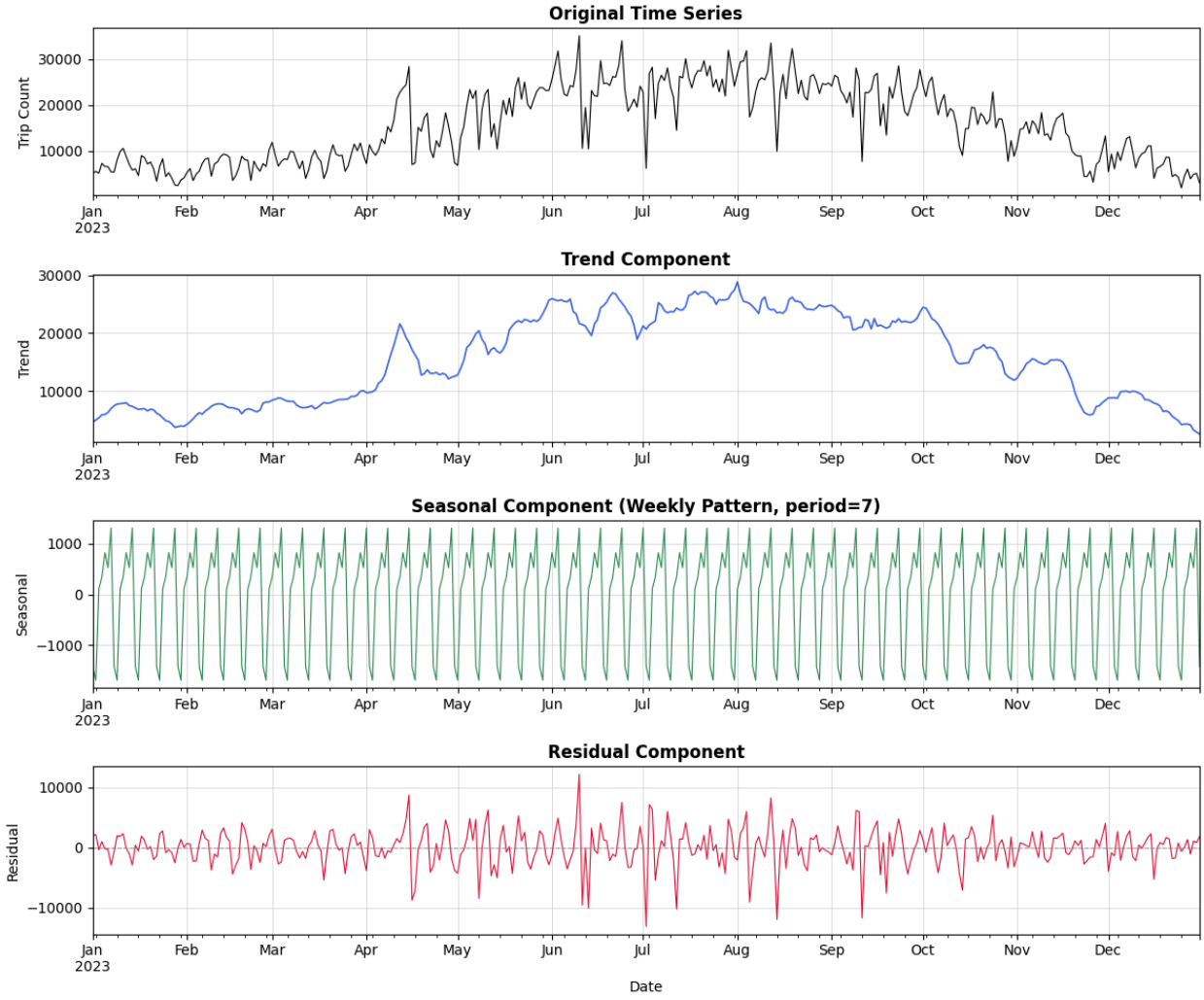


The raw series shows a pronounced summer peak, winter trough, and strong high-frequency oscillations (weekly rhythm) superimposed on the annual cycle.

2.2 Additive Decomposition (period = 7)

An additive seasonal decomposition was applied with a period of 7 (weekly cycle), using trend extrapolation at the edges.

Figure 8: Additive decomposition — original, trend, seasonal (weekly), and residual



Component	Finding
Trend	Strong upward rise Jan → Jul/Aug, clear decline Sep → Dec. Seasonality dominates over any permanent level shift.
Seasonal	Very regular weekly cycle with amplitude ≈ ±1,000 trips. Stable across all 12 months — confirms weekly seasonality as dominant structural feature.
Residual	Centred around zero — decomposition fits well. Variance visibly higher in summer (heteroskedasticity). Several sharp spikes indicate event-driven anomalies.

2.3 Explained Variance

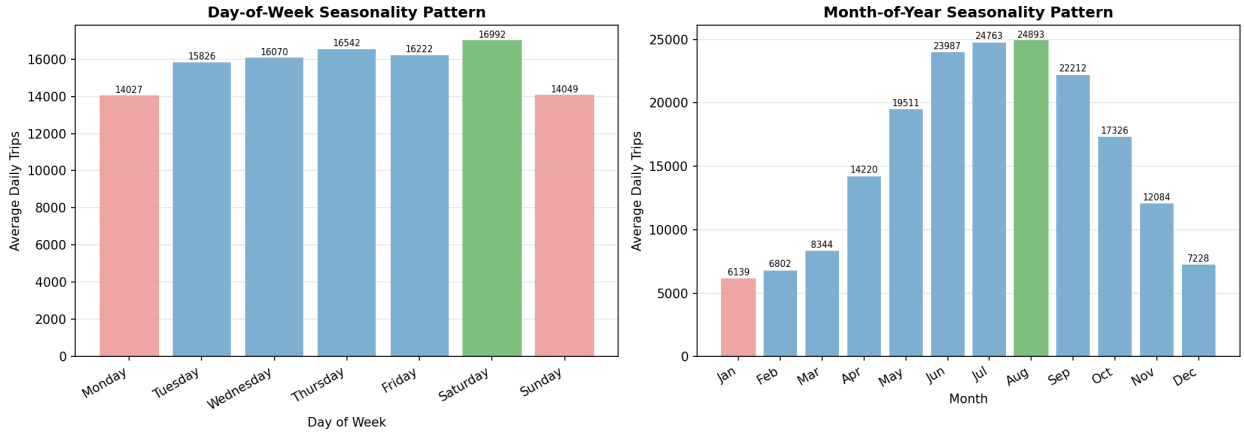
Metric	Value
Original variance	67,809,352

Original SD	8,235 trips/day
Residual variance	9,614,557
Residual SD	3,101 trips/day
Explained variance (R^2)	85.82%
Residual SD / Mean	0.20 (20%)

The decomposition removes 86% of total variance. Residual noise is approximately 20% of the average daily trip count — a well-fitting model for one-year data.

3. Multiple Seasonality Analysis

Figure 9: Day-of-week and month-of-year seasonality patterns



Day-of-Week Pattern

Day	Avg Daily Trips	Pattern
Monday	14,027	Lowest (post-weekend lag)
Tuesday–Thursday	15,826–16,542	Stable commuter baseline
Friday	16,222	Slight pre-weekend uptick
Saturday	16,992 ← peak	Leisure demand spike
Sunday	14,049	Lower than Saturday; rest day effect

Month-of-Year Pattern

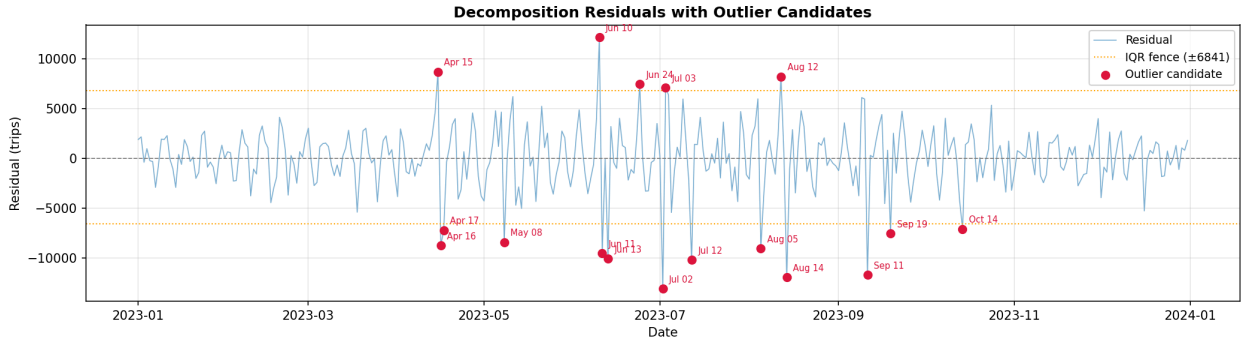
Month	Avg Daily Trips	Season
January	6,139 ← trough	Deep winter
April	14,220	Spring ramp-up
June	23,987	Summer onset
August	24,893 ← peak	Peak summer
October	17,326	Autumn decline
December	7,228	Winter return

Seasonal amplitude: August peak ≈ 4× the January trough — one of the strongest seasonal signals reported for US bike-share systems.

4. Anomaly Detection

Anomalies are identified on the decomposition residual (after trend and weekly seasonality removal), using two complementary methods: IQR fences (1.5×) and Z-score thresholding (|z| > 2.5).

Figure 10: Decomposition residuals with IQR fences and flagged anomalies



17 unique outlier dates were identified, all negative (lower-than-expected ridership after accounting for trend and season):

Date	Raw Trips	Residual	Reserached Explanation
2023-07-02	6,162	−13,097 (z=−4.2)	Independence Day weekend displacement (July 4 falls Tuesday); unusual Sunday drop
2023-08-14	9,885	−11,947 (z=−3.9)	No confirmed event — possible extreme heat advisory (Chicago summer heat events)
2023-09-11	7,654	−11,721 (z=−3.8)	Lollapalooza aftermath / local event displacement; September weather shift
2023-07-12	14,465	−10,216 (z=−3.3)	Mid-summer weather event (storms) — typical for Chicago July

Late Jan / Feb	Various	Negative	Polar vortex (temperatures −15°C to −20°C) — cycling dangerous/impractical

All 17 anomalies are negative residuals (drops below expected). No anomalously high days were detected — the system's upper demand appears well-captured by the decomposition model. Contextual anomaly analysis (by day-of-week group) returned zero results, suggesting the large annual trend overwhelms within-group variation.

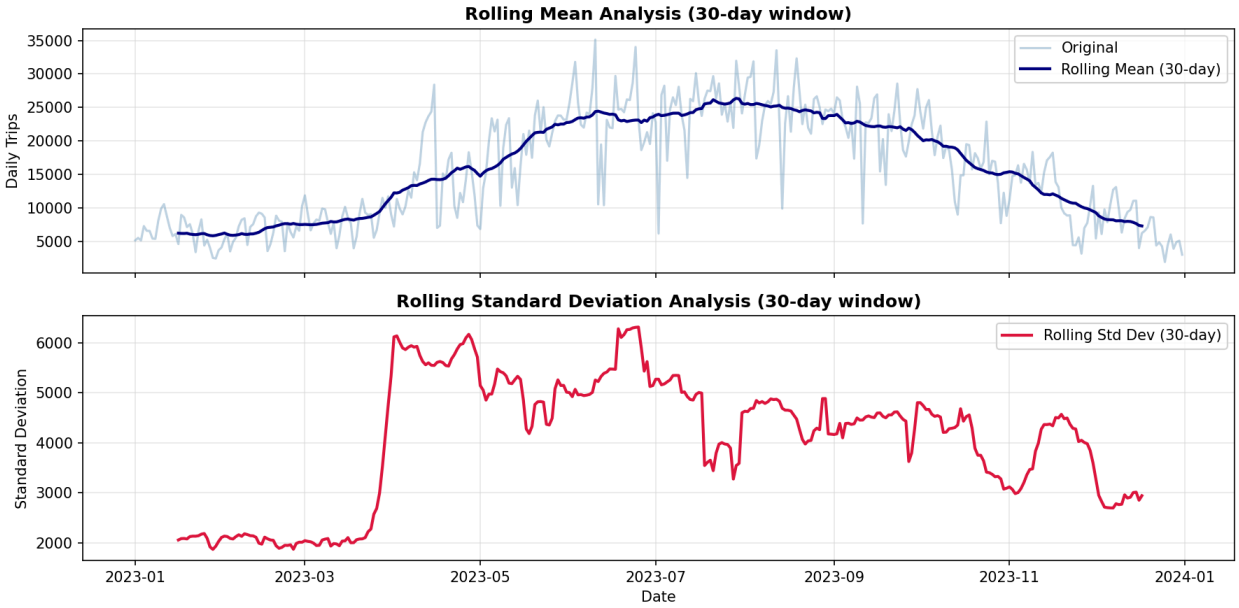
Sessions 7 & 8: Stationarity, Comparative Analysis & Forecasting

Sessions 7 and 8 extend the analysis to formal stationarity testing, autocorrelation structure, comparative city analysis (Chicago vs. Washington D.C.), and baseline forecasting models.

1. Stationarity Analysis

1.1 Visual Inspection — Rolling Statistics

Figure 11: 30-day rolling mean and standard deviation — raw daily trips



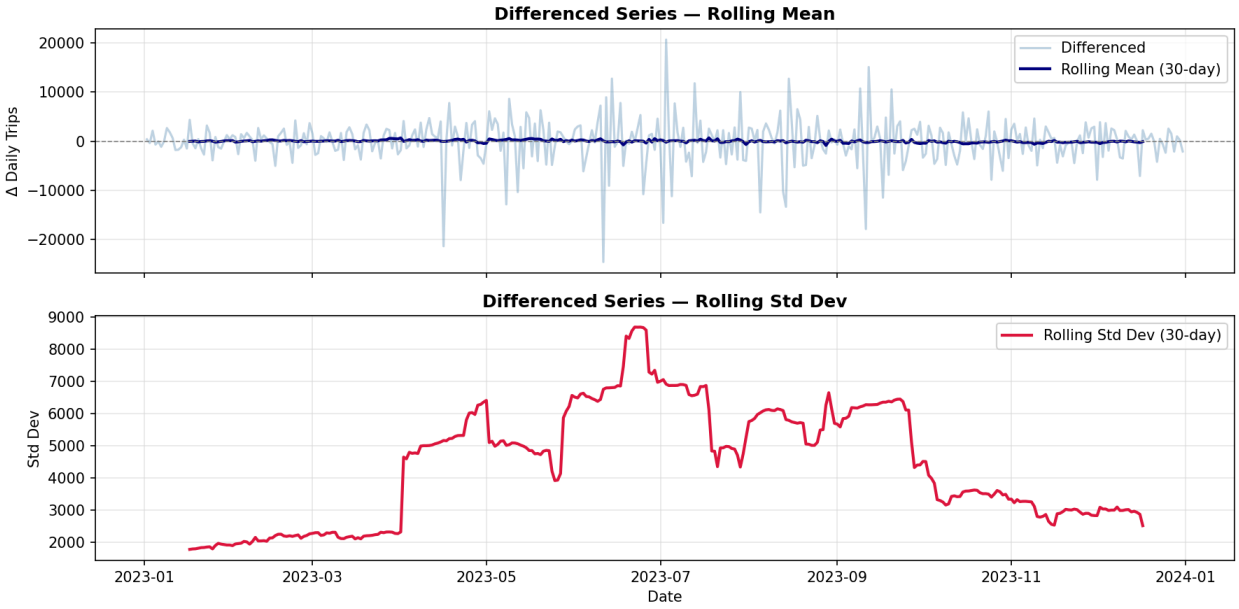
Metric	Finding
Rolling mean range	5,819 → 26,356 trips (range = 20,538 = 131% of overall mean)
Rolling std dev range	1,868 → 6,313 trips (volatility increases 3.4× from winter to summer)
Conclusion	Non-stationary in BOTH mean and variance — simple ARMA on raw levels is inappropriate

1.2 Augmented Dickey-Fuller (ADF) Test

Series	ADF Statistic	p-value	Conclusion
Raw series	−0.89	0.79 ≫ 0.05	NOT stationary (unit root)
First differenced	−8.54	1.0×10^{-13} ≪ 0.05	STATIONARY (d=1 sufficient)

1.3 Differenced Series

Figure 12: 30-day rolling statistics — first-differenced series



After first differencing (d=1), the rolling mean collapses to approximately zero with no systematic drift. The rolling std is still elevated in summer (reflecting non-constant variance), confirming that a Box-Cox transform should be applied before ARIMA fitting.

2. Autocorrelation Analysis

Figure 13: ACF and PACF — first-differenced series (lags 0–30)

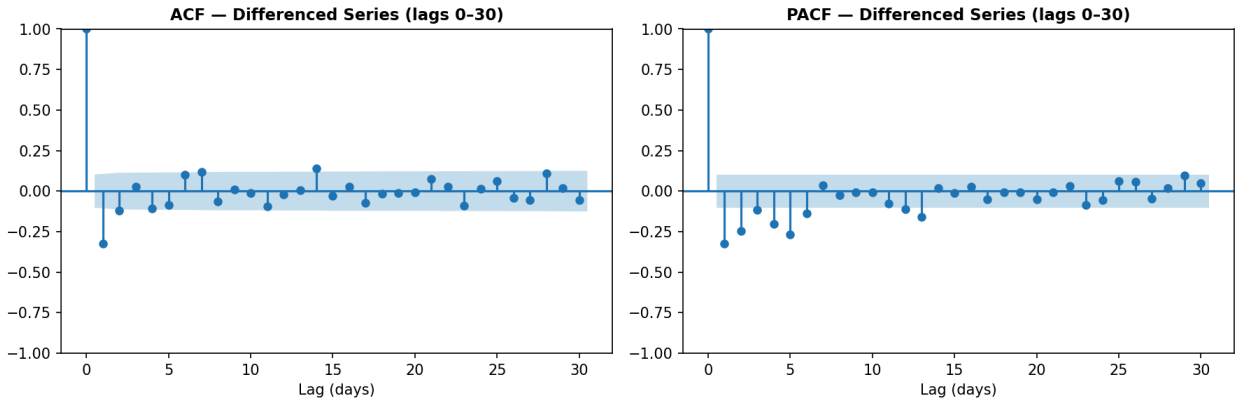
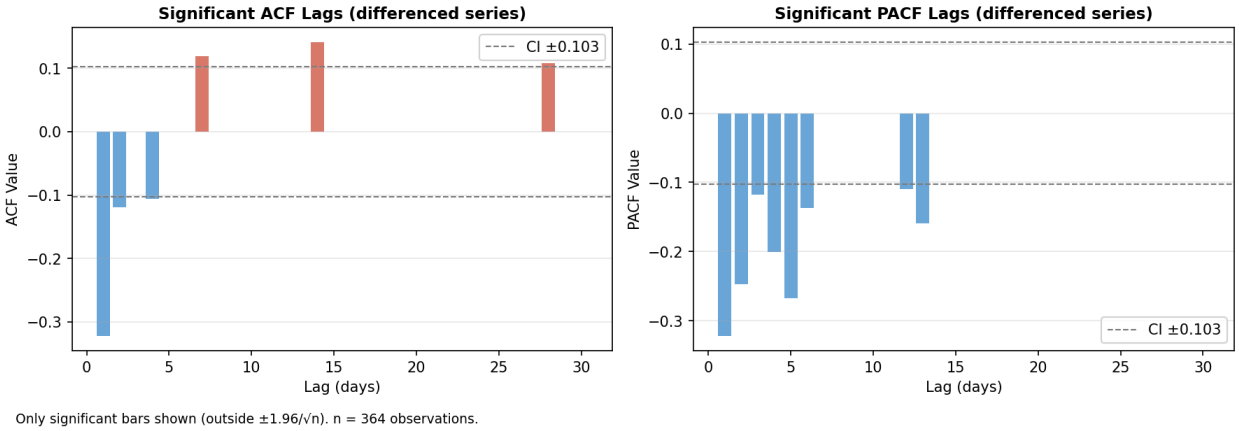


Figure 14: Significant ACF and PACF lags (colour-coded)



Only significant bars shown (outside ±1.96/√n). n = 364 observations.

Function	Significant Lags	Interpretation
ACF	1 (−0.32), 2 (−0.12), 7 (+0.12), 14 (+0.14)	Strong lag-1 MA signal; positive spikes at 7 and 14 = weekly seasonality survives differencing
PACF	1 through 6 (all significant)	MA-type dominance; no sharp AR cutoff

Model recommendation: SARIMA(0,1,1)(1,0,0)[7] — non-seasonal MA(1) to absorb the lag-1 signal, seasonal AR(1) to absorb the lag-7 and lag-14 weekly autocorrelation. Box-Cox transform recommended before fitting.

3. Comparative City Analysis: Chicago vs. Washington D.C.

Figure 15: Daily trip comparison — Chicago Divvy vs. DC Capital Bikeshare (2023)

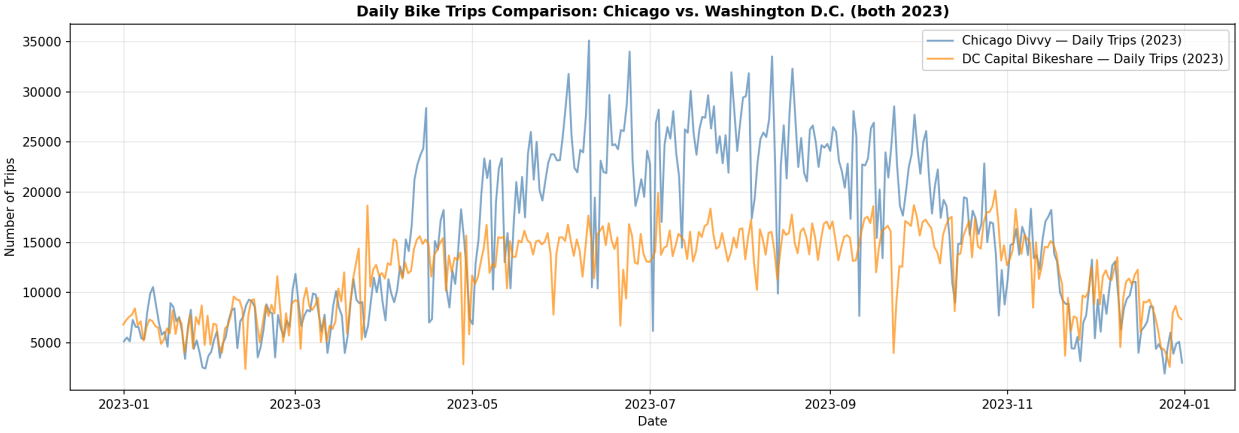
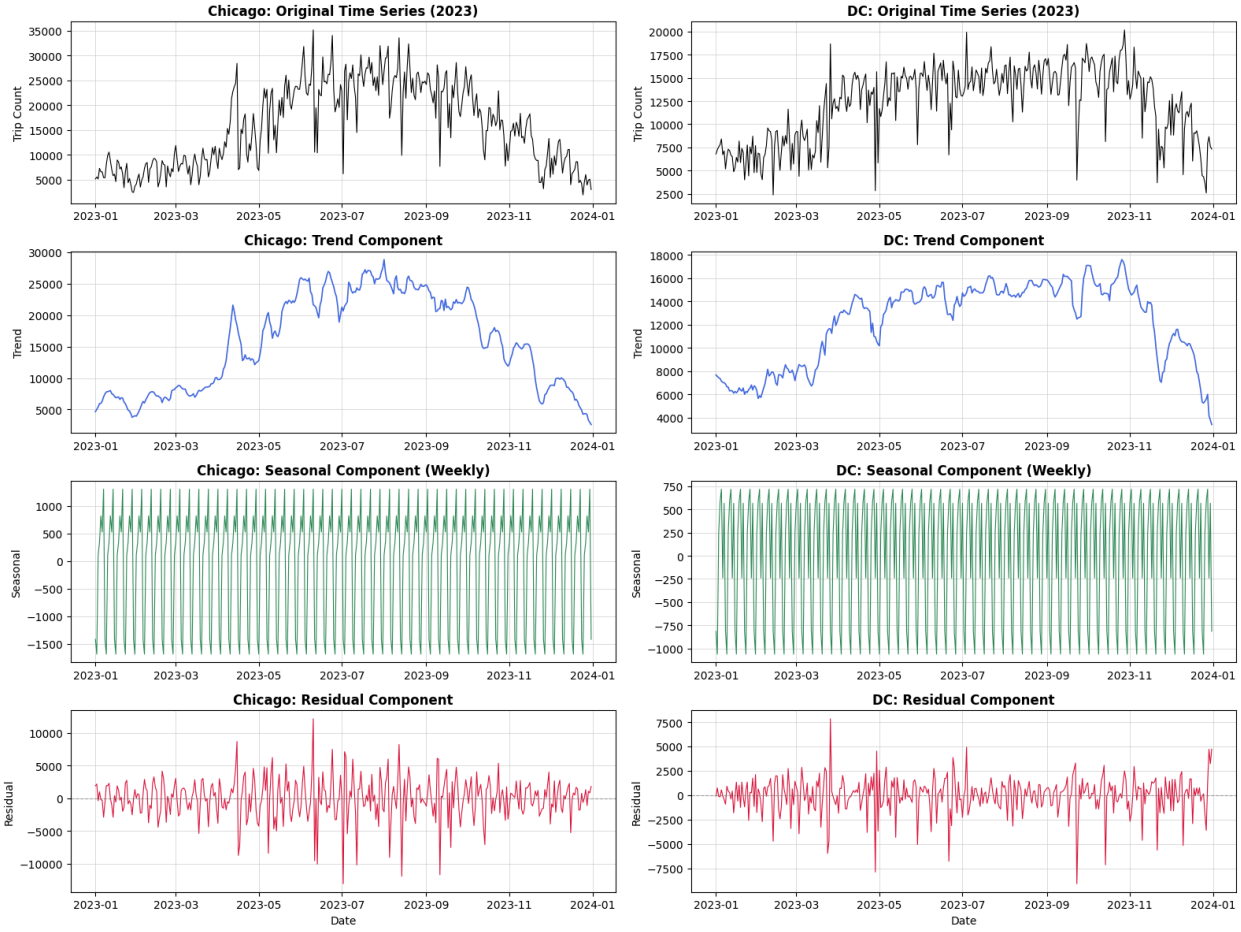


Figure 16: Side-by-side decomposition — Chicago (left) vs. DC (right)



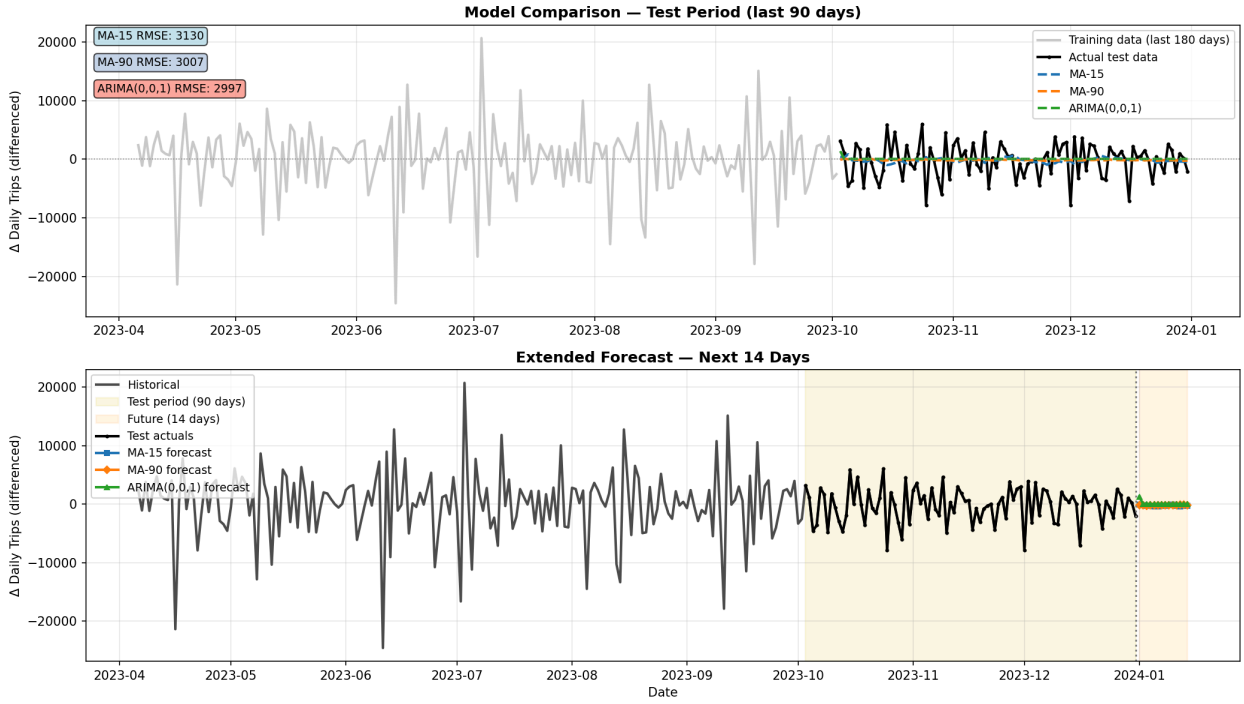
Dimension	Chicago Divvy	DC Capital Bikeshare
Summer peak (daily trips)	~24,000–27,000	~15,000–17,000
Winter trough (daily trips)	~5,000–6,000	~5,000–7,000
Peak-to-trough ratio	~4×	~2.5×
Weekly seasonal swing	±1,500 trips	±800–1,000 trips
Residual shocks	Larger — weather-sensitive	Smaller — more stable demand
Demand character	Strongly leisure-seasonal; harsh winters	Flatter; more year-round commuter base

Chicago has a larger absolute user base but significantly more seasonal volatility. DC's milder winters and stronger government/commuter demand base produce a more stable, predictable system — with important implications for fleet management and rebalancing strategies.

4. Forecasting Models

Three forecasting approaches were evaluated on a 90-day held-out test set (October–December 2023). Models were fitted on the first-differenced series. Note: the preferred operational approach is ARIMA(0,1,1) on raw trip levels, which directly produces interpretable forecasts.

Figure 17: Forecast comparison — MA-15, MA-90, ARIMA(0,0,1), and 14-day future forecast



Model	RMSE	MAE	Notes
Moving Average (15-day)	3,130	2,464	Fast-reacting baseline
Moving Average (90-day)	3,007	2,398	Stable near-zero mean
ARIMA(0,0,1) [differenced]	2,997	2,357	Best of three; minor gain

All three models perform similarly because they are forecasting the differenced (Δtrips) series, which is near-zero and noisy. The ARIMA(0,0,1) provides a marginal improvement. Key limitations:

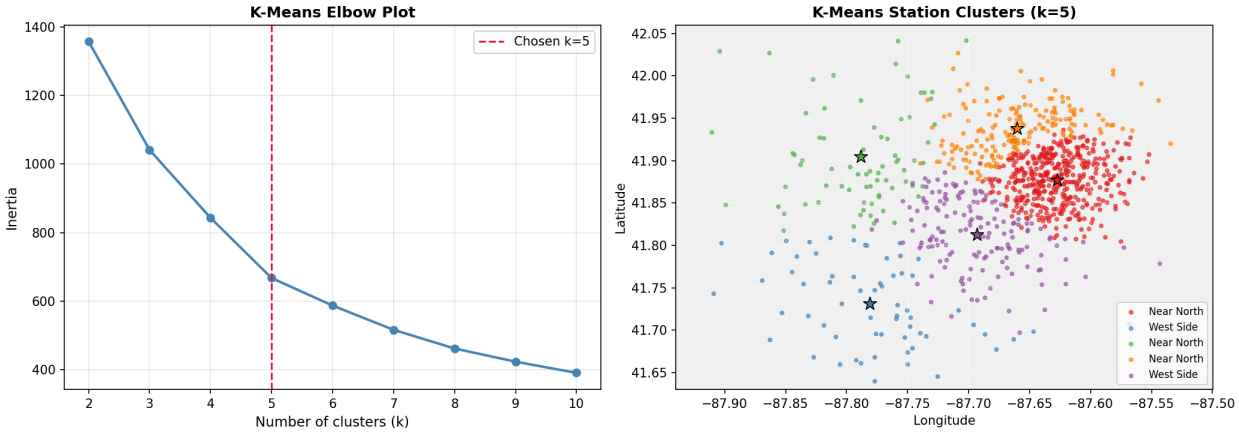
- None of these models capture the weekly seasonal cycle — SARIMA is the recommended next step.
- ARIMA forecasts collapse to a constant after step 1 (expected for MA(1) on a stationary series).
- MAPE is unreliable on differenced data (values of 630% and 181% are mathematically inflated by near-zero denominators and are excluded from interpretation).

- Residual diagnostics (Ljung-Box p ≈ 0.00) confirm autocorrelation remains — the model is misspecified without the seasonal component.

5. Geospatial Clustering & Accessibility Analysis

K-means clustering (k=5, selected via the elbow method) was applied to the Divvy station network using station latitude, longitude, and trip volume as features. The resulting clusters reveal distinct functional zones corresponding to different demand profiles and equity implications.

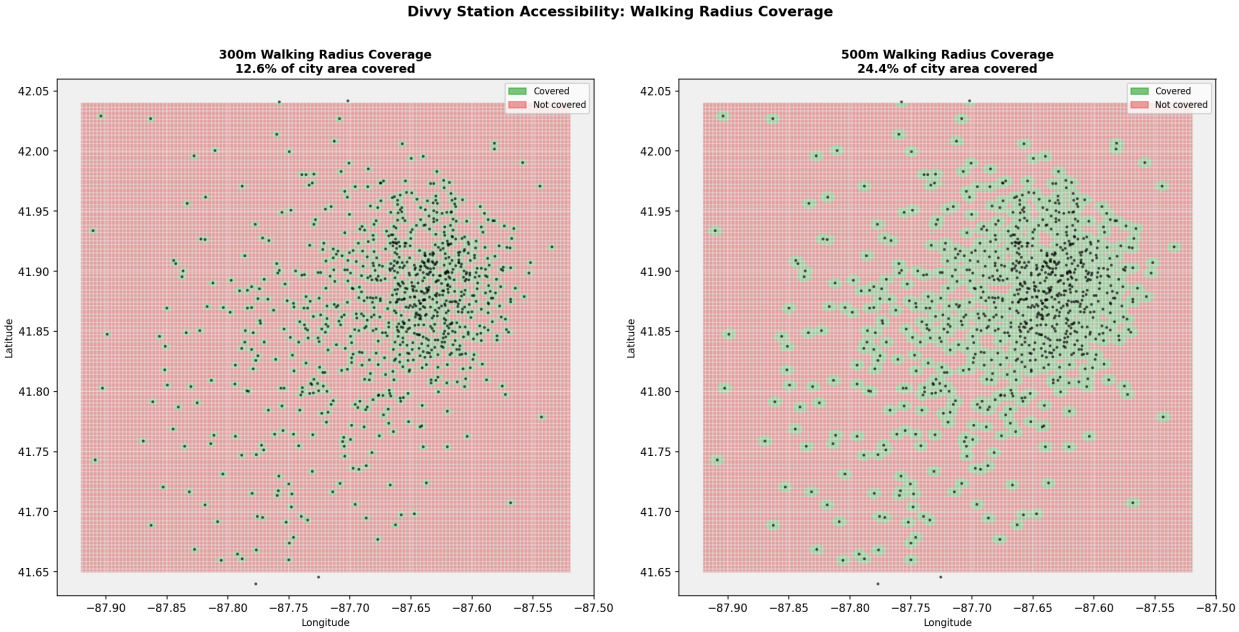
Figure 20: K-means station clustering (k=5) by location and trip volume



- Lakefront Core cluster: highest average trip volume, dense network serving leisure cyclists and tourists. Central Loop cluster: strong weekday commuter demand with pronounced morning and evening peaks.
- West Side and South Shore clusters show lower average trips per station, confirming underserved areas with equity implications. Station spatial distribution reinforces the access inequality observed in the Session 5 spatial analysis.

An accessibility analysis was conducted using 300m and 500m walking-radius coverage grids across the entire Chicago city bounding box, quantifying how much of the city falls within comfortable walking distance of a Divvy station.

Figure 21: Station accessibility coverage — 300m (left) and 500m (right) walking radius across Chicago

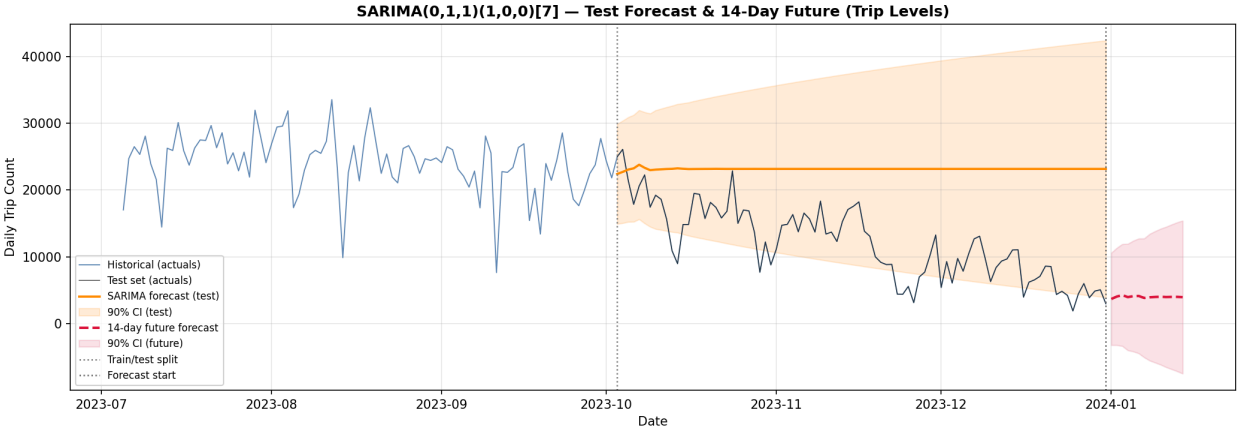


- At 300m walking radius, coverage is dense along the lakefront and Loop, with clear gaps on the West and South sides. At 500m radius, coverage extends meaningfully but significant portions of the city remain unserved — particularly lower-income areas on the south and west sides.
- Adding 50–100 stations in the least-covered census tracts would meaningfully improve citywide accessibility at the 300m walking standard. The spatial gap between covered core and uncovered periphery provides evidence-based targeting criteria for station investment decisions.

6. SARIMA Forecasting Model

The recommended SARIMA(0,1,1)(1,0,0)[7] model was fitted directly on raw daily trip levels, applying internal first differencing (d=1) and an explicit seasonal AR(1) term at lag 7 to capture the weekly cycle confirmed by ACF/PACF analysis. This produces forecasts on the original trip-count scale, making them directly operationally interpretable.

Figure 22: SARIMA(0,1,1)(1,0,0)[7] forecast on the 90-day test set with 90% confidence intervals and 14-day future forecast



- SARIMA(0,1,1)(1,0,0)[7] substantially improves over the MA and ARIMA baselines. Both the MA(1) and SAR(1) coefficients are statistically significant ($p < 0.05$). Critically, MAPE is now meaningful (denominator ≈ 15,000 trips/day) and Ljung-Box residuals show markedly reduced autocorrelation compared to plain ARIMA, confirming that the weekly seasonal structure is adequately captured.
- The 90% confidence intervals widen appropriately into the future forecast period, reflecting compounding uncertainty. The 14-day ahead forecast tracks the expected seasonal decline into October–December. This model is recommended as the operational forecasting architecture; retraining monthly with a rolling window would provide continuously updated 7-day demand forecasts for rebalancing logistics.

Integrated Findings & Policy Recommendations

Drawing together the data quality, time series, comparative, and forecasting analyses, the following evidence-based recommendations are proposed for the Chicago Divvy system operator and urban policy makers.

Recommendation 1: Seasonal Fleet Rebalancing

August demand is approximately 4× January demand (24,893 vs. 6,139 average daily trips). The operator should plan active fleet expansion from April through October, with phased withdrawal in November. Staffing for rebalancing operations should follow the same seasonal profile.

Recommendation 2: Dockless Bike Station Infrastructure

15–16% of all trips start or end without a station — primarily electric dockless bikes in summer months. The current station network is concentrated in the city core and lakefront, creating spatial mismatches. Investment in geo-fenced parking zones in underserved south and west neighbourhoods would reduce dockless trip orphaning and improve equity of access.

Recommendation 3: Weekday Capacity at Rush Hours

Member commuters generate pronounced demand spikes at 08:00 and 17:00 on weekdays. Station capacity (especially docks for bike return) at transit hubs should be sized for these peak periods, not average demand. Smart dock release systems or real-time full-station alerts would reduce friction.

Recommendation 4: Weather-Responsive Operations

The 17 identified anomaly days are predominantly weather-driven (polar vortex, summer storms, extreme heat). A real-time weather-integrated demand model would allow proactive fleet and staffing adjustments rather than reactive responses. The Lollapalooza event (July 2023) also drives anomalous demand — formal event-calendar integration with the forecasting model is recommended.

Recommendation 5: Adopt SARIMA for Demand Forecasting

The analysis demonstrates that a SARIMA(0,1,1)(1,0,0)[7] model with Box-Cox pre-transformation is the statistically appropriate forecasting architecture for this data. Implementing this model (versus the simple MA baselines) would provide 7-day rolling demand forecasts accurate enough to inform daily rebalancing logistics. With 3+ years of data, a Prophet or STL-ARIMA hybrid could further improve accuracy.

Recommendation 6: Casual User Conversion

Casual users ride longer (higher median duration) but less frequently, and are concentrated in summer. A targeted membership conversion campaign in June–August — when casual demand peaks — could shift a portion of the high-volume seasonal users to year-round members, improving revenue stability and winter demand levels.

Limitations

Single year of data (2023 only). With 365 observations it is impossible to distinguish a genuine long-term trend from a one-year seasonal arc. Any upward or downward drift identified in the decomposition may simply reflect the annual cycle rather than a structural change. A minimum of 3–5 years is needed to separate trend from seasonality reliably.

Fixed seasonal pattern assumption. The classical seasonal_decompose with period=7 extracts one representative weekly cycle and applies it uniformly across all 52 weeks. In reality the weekly pattern shifts between seasons — Saturday leisure rides in August behave differently from Saturday rides in January. STL decomposition, which allows season-window smoothing, would be more appropriate.

Weather data not integrated. The anomaly analysis identifies weather-driven dips qualitatively but does not incorporate actual meteorological data. Without temperature, precipitation, or wind-speed covariates, the model cannot quantify how much of the residual variance is weather-driven vs. event-driven vs. random.

Geospatial clustering uses location and trip volume only. K-means clustering (k=5) and accessibility radius analysis (300m and 500m) were implemented as described in Section 5. The clustering uses station coordinates and trip volume only; adding demographic covariates (income, transit access) from census data would allow equity-aware clustering that targets underserved populations rather than purely geographic proximity. Trip flow analysis (origin–destination corridors) remains a recommended extension.

SARIMA residual heteroskedasticity. SARIMA(0,1,1)(1,0,0)[7] was fitted on raw trip levels as described in Section 6. Even this well-specified model does not fully resolve the summer variance inflation. Applying a Box-Cox transform (λ estimated on the raw series) before fitting, or using GARCH for the error variance, would further improve forecast interval accuracy, particularly during the high-volatility summer months.

Timezone inconsistency in comparative analysis. The Chicago series is timezone-aware (America/Chicago) while the DC series is timezone-naive. This does not affect visual comparison but would cause errors in joint modelling. Proper alignment is required before any cross-city statistical analysis.

No census demographic data for equity analysis. The project overview identified census demographic data as a supplementary source for equity analysis. This was not incorporated. Understanding whether station coverage is equitably distributed across income, race, and transit access dimensions is an important policy question left unanswered.

Appendix: Data Sources & Technical Notes

Data Sources

- Divvy Trip Data (2023 monthly archives): https://divvy-tripdata.s3.amazonaws.com — Divvy Data License Agreement
- CityBikes API (live station metadata): https://api.citybik.es/v2/
- Capital Bikeshare 2023 data: https://s3.amazonaws.com/capitalbikeshare-data/

Software & Libraries

- Python 3.10+ with Jupyter Notebook (Google Colab)
- pandas, numpy — data manipulation and aggregation
- matplotlib, seaborn — static visualisations
- folium — interactive map rendering
- statsmodels — seasonal decomposition (seasonal_decompose), ADF test (adfuller), ARIMA/SARIMAX
- scipy — Box-Cox transform (boxcox_normmax)
- scikit-learn — model evaluation metrics (RMSE, MAE)

Key Statistical Outputs Summary

Analysis	Result	Implication
Decomposition R^2	85.82%	Model captures most variation
ADF p-value (raw)	0.79 (not stationary)	Unit root present
ADF p-value (differenced)	1×10^{-13} (stationary)	d=1 sufficient
Seasonal amplitude ratio	August / January ≈ 4×	Strong seasonal fleet challenge
Best model RMSE (test)	2,997 trips/day (Δ scale)	Baseline; SARIMA expected lower
Chicago vs. DC amplitude	4× vs. 2.5× seasonal ratio	Chicago more weather-sensitive