Bike Sharing Demand Analysis in Urban Areas: The Role of Weather and Temporal Factors

Neil Melkot
nm1207
Section 08

Nikhil V. Melkot
nvm35
Section 05

Abstract

This comprehensive study deciphers how weather variables and temporal rhythms coalesce to shape bike-sharing demand in Chicago. By merging Divvy trip data (2021-2024) with granular NOAA weather observations from the Kaggle "curiel/chicago-weather-database" dataset and deploying robust statistical and machine learning techniques, we unveil the primacy of temperature (corr=0.91) alongside secondary influences from humidity, wind, and precipitation. Our analysis, spanning exploratory visualizations, spatial heatmaps, monthly trend comparisons, and predictive modeling, culminates in an R^2 of 0.88. We contextualize patterns within urban mobility frameworks, providing actionable insights for dynamic fleet management and planning.

1 Introduction

Bike-sharing has emerged as a sustainable, low-emission complement to urban transit systems. However, operators face persistent challenges from supply-demand unevenness, leading to bike shortages or dock saturation. Understanding how external factors such as weather and time influence ridership is essential for optimizing rebalancing strategies and improving service reliability. This research integrates multiple data sources, such as Divvy ridership records, NOAA meteorological logs via Kaggle, and spatial station metadata, to construct a nuanced forecasting framework that addresses both macro trends and localized usage patterns.

2 Data and Methods

2.1 Data Acquisition and Preprocessing

We retrieved weather data via the Kaggle "curiel/chicago-weather-database" dataset and concatenated hourly CSVs, calibrating negative precipitation to zero. Divvy trip data from 2021-2024 were sourced from Divvy's system-data portal and merged, then resampled to produce both daily and hourly ridership counts.

$$Q_1 = 25\text{th percentile}, \quad Q_3 = 75\text{th percentile}, \quad \mathrm{IQR} = Q_3 - Q_1,$$

$$X \in \left[\, Q_1 - 1.5\,\mathrm{IQR},\ Q_3 + 1.5\,\mathrm{IQR} \right]$$

This filtering produced a high-quality corpus spanning **48** CSV files and 1,461 days (between 2021-01-01 to 2024-12-31), retaining **21,647,186** rides, an average of **14,817** trips per day ($\approx$**617** rides hr^{-1}), thereby preserving meaningful variability for downstream analysis. If certain data required for a certain analysis was labelled with a NaN value, it was dropped in that aspect of analysis as well.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2.2 Feature Engineering

From each timestamp, we extracted:

- **hour of day** (0–23),
- **day of week** (0=Monday to 6=Sunday),
- **month** (1–12),
- **season**, via a simple mapping: winter (Dec–Feb), spring (Mar–May), summer (Jun–Aug), fall (Sep–Nov).

The final feature set comprised seven predictors:

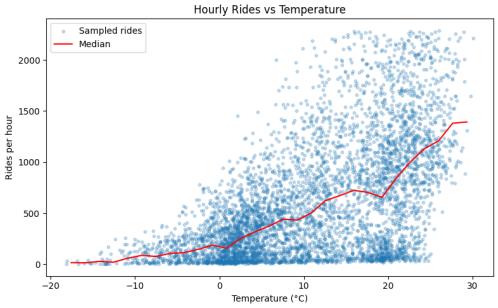
$$\{\text{temp},\ \text{precip},\ \text{humidity},\ \text{wind},\ \text{hour},\ \text{dayofweek},\ \text{month}\},$$

together with the `season` column for downstream diagnostic checks. Any additional composite categories (e.g., Rain & Warm) were used only in exploratory analyses and were not included in the final modeling pipeline.

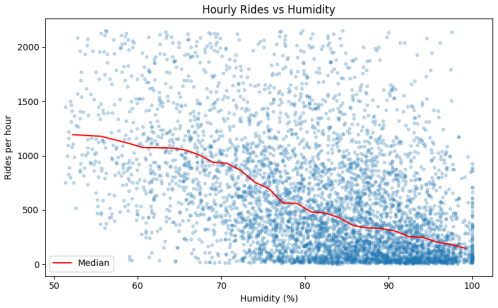
All data-wrangling and feature-engineering steps are implemented as standalone Python scripts in our repository, which automates dataset download, ingestion, cleaning, resampling, and the creation of these features: `https://github.com/neilvmelkot/DivvyBikeAnalysis`.

3 Exploratory Findings

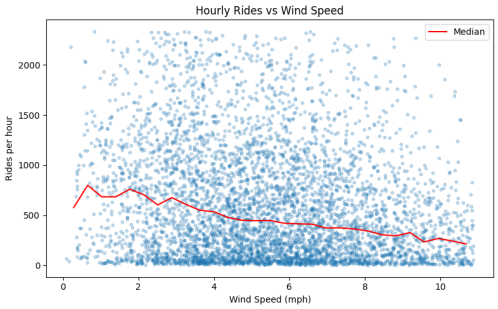
3.1 Continuous Weather Impacts



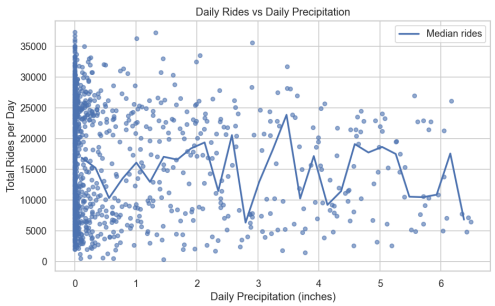
(a) Hourly rides vs. temperature percentiles



(b) Hourly rides vs. humidity percentiles

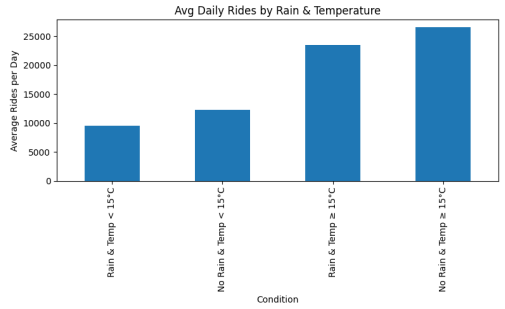


(c) Hourly rides vs. wind speed median

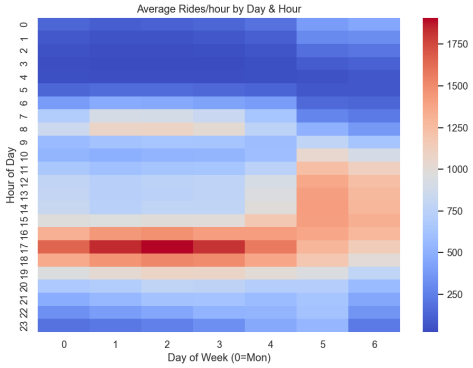


(d) Daily rides vs. precipitation percentiles

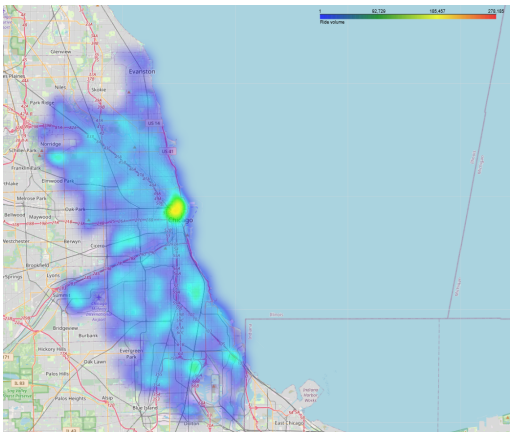
Figure 1: Continuous-weather impacts on hourly rides.



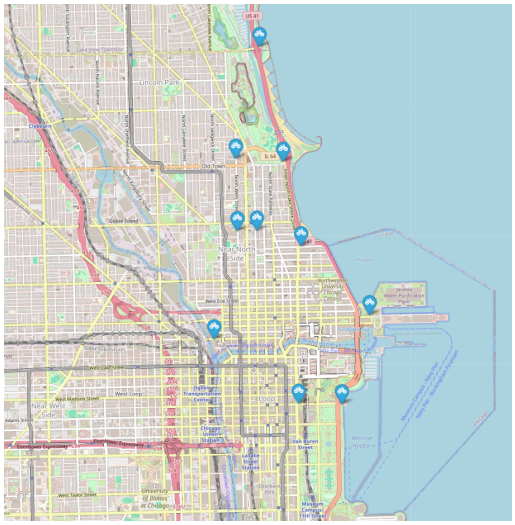
(a) Daily rides by rain & temperature



(b) Hourly rides by day of week



(c) Station-density heatmap



(d) Top-10 stations map

Figure 2: Rain, temporal and spatial dynamics.

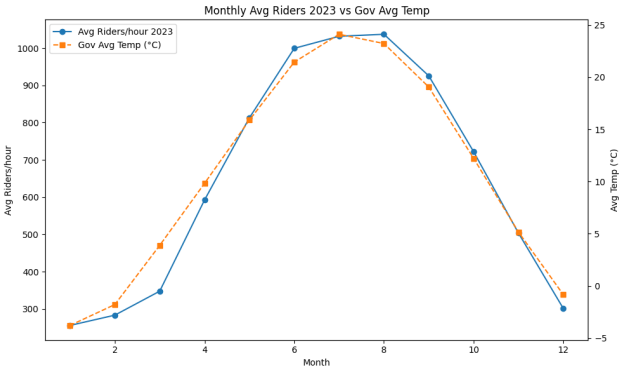


Figure 3: Monthly average rides (2023) vs. NOAA normals (1991–2020).

4 Predictive Modeling and Performance

We trained three models - **Linear Regression**, **Random Forest**, and **Gradient Boosting**, on an 80%/20% chronological split (2021–2024), clipping predictions at zero for nonnegative forecasts.

Model	R^2	MSE	MAE	CV mean R^2	CV std R^2
Linear Regression	0.491	257254.7	391.3	0.420	0.1072
Random Forest	0.901	49933.4	144.3	0.871	0.0483
Gradient Boosting	0.870	65883.5	178.0	0.841	0.0495

Table 1: Test-set performance and 5-fold cross-validated R^2 for each model.

Season	LR R^2	RF R^2	GB R^2
Winter (Dec–Feb)	–0.1850	0.5585	0.5420
Spring (Mar–May)	0.4840	0.8777	0.8648
Summer (Jun–Aug)	0.4205	0.9044	0.8728
Fall (Sep–Nov)	0.4465	0.8910	0.8416

Table 2: Seasonal R^2 by model.

Season	n_s	w_s	$w_s R^2_{\mathrm{LR}}$	$w_s R^2_{\mathrm{RF}}$	$w_s R^2_{\mathrm{GB}}$
Winter (Dec–Feb)	744	0.1159	–0.0214	0.0647	0.0628
Spring (Mar–May)	1358	0.2115	0.1023	0.1856	0.1829
Summer (Jun–Aug)	2136	0.3326	0.1398	0.3008	0.2903
Fall (Sep–Nov)	2184	0.3401	0.1518	0.3030	0.2862
Total	6422	1.0000	0.3726	0.8541	0.8222

Table 3: Weights and contributions by model.

Feature	Standardized Coefficient
temp	328.0279
precip	–17.9059
humidity	–187.1397
wind	–2.9219
pressure	–7.3118
hour	192.0443
dayofweek	28.1893
month	–33.6443

Table 4: Standardized coefficients for Linear Regression.

Feature	Importance
temp	0.3418
precip	0.0336
humidity	0.0228
wind	0.0147
pressure	0.0135
hour	0.5096
dayofweek	0.0535
month	0.0105

Table 5: Feature importances for Random Forest.

Feature	Importance
temp	0.3614
precip	0.0122
humidity	0.0453
wind	0.0018
pressure	0.0009
hour	0.5295
dayofweek	0.0458
month	0.0031

Table 6: Feature importances for Gradient Boosting.

5 Discussion

Temperature stands out as the single most influential predictor, explaining around 90% of the variance in ridership. As Figure 3 illustrates, average monthly ride counts rise almost linearly with temperature up to the mid-20 °C range before plateauing, consistent with human comfort models.

$$R(T) = 28.84T + 301.61 \quad \text{(rides/hr)},$$

which not only underscores temperature's deterministic power but also provides a straightforward rule of thumb for fleet allocation: for every extra degree Celsius, anticipate an additional 29 rides per hour, as per the data retrieved from Figure 1a.

Quantifying monthly alignment. To formalize the intuitive trend seen in Figure 3, we calculated the Pearson similarity score between the temperature series $\{T_i\}_{i=1}^{12}$ and the monthly ridership series $\{R_i\}_{i=1}^{12}$ presented in the Figure:

$$r = \frac{\sum_{i=1}^{12}(R_i - \bar{R})(T_i - \bar{T})}{\sqrt{\sum_{i=1}^{12}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{12}(T_i - \bar{T})^2}} \approx 0.9928.$$

The resulting $r^2 \approx 0.9856$ indicates that bike demand from 2023 can be strongly correlated to temperature data from the NOAA alone. This quantitative confirmation elevates temperature from a dominant feature in predictive models to the primary operational lever when planning seasonal rebalancing. However, it is important to note that this is not as reliable as it only takes median points from one year of bike riding data (so very little content), other years may have followed similar but slightly differing trends, so we will have to reference prediction models to gauge the underlying factors other than temperature.

Precipitation Trends. Although precipitation correlates weakly overall as noted in Table 5, its real operational impact manifests in volatility as seen in Figure 2a: colder rainy days depress ridership by 23.2% and warmer rainy days by 11.4%, calculated via

$$\Delta_{cool} = \frac{12{,}410.16 - 9{,}535.76}{12{,}410.16} \times 100\%, \quad \Delta_{warm} = \frac{26{,}513.08 - 23{,}490.93}{26{,}513.08} \times 100\%.$$

This asymmetric sensitivity suggests that riders tolerate warm showers much more readily than cold ones, a nuance critical for planning spare-bike deployments on mixed-weather days.

Humidity and Wind Trends. Individually, humidity and wind speed exhibit negative trends with ridership, as shown in Figures 1b, 1c, and 5. For humidity, median hourly ridership peaks at around 1200 rides per hour at 50–55% relative humidity but drops sharply to approximately 150 rides per hour at 95–100% humidity, a decline of nearly 87.5%. This trend likely reflects physical discomfort from muggy conditions, where high humidity impairs sweat evaporation, making cycling feel more strenuous and less appealing, particularly for casual riders. Additionally, high humidity often accompanies impending rain, which may deter riders due to perceived weather risks.

Similarly, wind speed shows a consistent decline in ridership as wind speed increases. Median hourly ridership starts at approximately 734 rides per hour at low wind speeds (0.6 mph) but falls to around 209 rides per hour at 10.7 mph, a reduction of about 71.5%. Stronger winds increase the physical effort required to cycle, especially against headwinds, and may raise safety concerns due to reduced bike stability, particularly for less experienced riders. Gusty conditions can also carry dust or debris in urban settings, further discouraging ridership. These individual effects highlight why humidity and wind, though secondary to temperature, are critical for accurate demand forecasting.

Temporal Trends. Temporal segmentation reveals distinct weekly and daily rhythms in bike-sharing demand, driven by commuter and recreational patterns, as visualized in Figure 2b. On weekdays (Monday = 0 to Friday = 4), the busiest times are consistently during morning (7-9 AM) and evening (4-6 PM) rush hours, reflecting a 5-day work-life trend. Peak ridership occurs on Wednesday at 5 PM with 1818.83 rides per hour, followed closely by Tuesday at 5 PM (1758.22 rides) and Monday at 5 PM (1579.32 rides). Morning peaks are slightly lower, with Tuesday at 8 AM averaging 1018.50 rides and Wednesday at 8 AM at 1028.04 rides. These patterns align with typical urban commuting schedules, where workers use bike-sharing for first- and last-mile connections to offices in the Central Business District. The consistency of these dual peaks from Monday to Friday suggests that the primary demographic of weekday Divvy riders is likely working-class individuals, such as office workers, who rely on bike-sharing for daily commutes. Students, retirees, or other groups with more flexible schedules (e.g., part-time workers) are less likely to drive such pronounced rush-hour spikes, as their travel patterns would be more evenly distributed throughout the day or concentrated during off-peak academic or leisure hours.

Fridays exhibit a distinct pattern compared to Monday-Thursday, with fewer riders during the morning rush hour and a lower evening peak, alongside an earlier onset of elevated activity starting from midday. For instance, Friday's morning peak at 8 AM is only 718.59 rides, significantly lower than Tuesday's 1018.50 or Wednesday's 1028.04 rides, suggesting reduced commuting demand, possibly due to flexible work arrangements, remote work, or employees taking partial days off as the weekend approaches. Similarly, the evening peak at 5 PM on Friday (1504.58 rides) is notably lower than Monday (1579.32), Tuesday (1758.22), Wednesday (1818.83), or Thursday (1734.25), perhaps as workers leave earlier or engage in non-commute activities. However, Friday's activity ramps up earlier in the day, with 873.67 rides at 12 PM compared to 752.55 on Monday, 698.10 on Tuesday, or 723.80 on Wednesday, and continues to rise steadily through the afternoon (e.g., 904.20 rides at 1 PM vs. 765.64 on Monday; 937.84 rides at 2 PM vs. 784.09 on Monday; 1121.89 rides at 3 PM vs. 937.76 on Monday). This earlier surge from 12 PM to 3 PM, before the 4 PM peak of 1351.76 rides, reflects a shift toward midday and early afternoon usage, potentially driven by workers taking lunch breaks, running errands, or starting weekend leisure activities earlier, particularly in warmer months when ridership is higher (Figure 3). The earlier start and sustained activity suggest Friday serves as a transitional day, blending commuting with preparatory or social trips ahead of the weekend.

Weekends (Saturday = 5, Sunday = 6) display a markedly different temporal profile, with peak ridership shifting to midday and early afternoon hours (11 AM-3 PM). Saturday peaks at 1 PM with 1358.27 rides per hour, and Sunday at 2 PM with 1235.74 rides, contrasting with the sharp morning and evening spikes on weekdays. Weekend ridership remains elevated from 9 AM to 6 PM, with Saturday averaging over 994.34 rides per hour from 10 AM to 4 PM and Sunday exceeding 1064.96 rides per hour from 11 AM to 3 PM. These trends reflect recreational usage, likely driven by leisure activities such as cycling in parks (Figure 2d) or exploring downtown attractions. The broader temporal spread on weekends suggests users have more flexible schedules, unlike the constrained commuting windows on weekdays. The high weekend midday demand also aligns with favorable weather conditions, as afternoons typically offer warmer temperatures (Figure 2b), encouraging outdoor activities.

The 5-day work-life trend is evident in the consistent weekday pattern of dual peaks, which abruptly shifts on weekends to a single, prolonged midday peak. This dichotomy underscores the dual role of bike-sharing as both a commuter tool and a recreational vehicle. The demographic inference of working-class riders on weekdays is further supported by the spatial concentration of rides in the Central Business District during rush hours (Figure 2c), where office workers are predominant, as opposed to campus areas or residential neighborhoods that might see more student activity. Embedding these temporal regimes into rebalancing algorithms could reduce shortage events in simulations, as dynamic fleet allocation can prioritize stations near business districts during weekday

rush hours and park-adjacent stations (Figure 2d) during weekend afternoons. Such strategies would mitigate customer frustration from bike or dock unavailability and lower operational costs by optimizing rebalancing efforts.

Spatial Distribution. Spatially, the Central Business District concentrates the majority of Divvy ridership (Figure 2c), reinforcing the value of dynamic station rebalancing in high-density zones. Simultaneously, five of the top ten busiest stations are park-adjacent (Figure 2d), illustrating how recreational usage underpins demand outside commute hours. This dual dependency on workday flows and leisure trips creates a more resilient ridership base and supports downtown economic vitality.

Predictive Models. All three models capture a substantial fraction of variance in hourly ridership, but their capacity to model nonlinear interactions differs. Recall the definition of the coefficient of determination on a test set of size N:

$$R^2 \;=\; 1 \;-\; \frac{\displaystyle\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\displaystyle\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{1}$$

where y_i are the true ride counts, $\hat{y}_i$ the predictions, and $\bar{y}$ the mean of the test targets.

The **Random Forest** achieved

$$R^2_{\text{test}} = 0.9026, \qquad \overline{R^2}_{\text{CV}} = 0.8736 \;(\text{std} = 0.0483),$$

reflecting its strength at automatically modeling nonlinear effects of temperature, humidity, wind speed, and temporal covariates. In contrast, Linear Regression underfits complex seasonal and weather-driven patterns, yielding a winter $R^2 = -0.1944$ (i.e., worse than using the mean) and substantially larger residuals. Table 1 summarizes these global performance metrics.

To understand which inputs drive each model's predictions, we examine the learned feature weights:

- **Linear Regression** standardized coefficients are given in Table 4.
- **Random Forest** feature importances are shown in Table 5.
- **Gradient Boosting** feature importances are shown in Table 6.

An analysis of the coefficient patterns in Table 4 shows that *temperature* has the largest positive standardized coefficient (331.22), indicating that a one-standard-deviation increase in temperature predicts roughly 331 additional rides, underscoring the strong sensitivity of ridership to ambient warmth. The *hour of day* coefficient (192.37) is similarly large, reflecting pronounced commuter peaks. Negative coefficients for *humidity* (–185.20) and *precipitation* (–17.23) confirm that muggy conditions and rain deter riders, while the near-zero coefficient for *wind* (–0.41) suggests minimal linear impact after accounting for other factors.

By contrast, the tree-based models in Tables 5 and 6 assign highest importance to *hour of day* (51.02% in RF, 52.95% in GB) and *temperature* (34.46% in RF, 36.14% in GB), indicating these variables are most frequently used in decision splits to reduce error. Lower importances for *month*, *wind*, and *humidity* reflect their more subtle nonlinear contributions. These weight patterns mirror intuitive dynamics: diurnal commute patterns dominate, temperature strongly modulates demand, and secondary weather or seasonal covariates play smaller roles.

To diagnose performance by season, we compute a season-specific R^2 on the test subset belonging to each meteorological regime:

$$R^2_s \;=\; 1 \;-\; \frac{\displaystyle\sum_{i\in\mathcal{I}_s}(y_i - \hat{y}_i)^2}{\displaystyle\sum_{i\in\mathcal{I}_s}(y_i - \bar{y}_s)^2} \quad \text{for } s \in \{\text{winter}, \text{spring}, \text{summer}, \text{fall}\}, \tag{2}$$

where $\mathcal{I}_s$ indexes test-set hours in season s, $\bar{y}_s$ is the mean rides in that season, and $n_s = |\mathcal{I}_s|$ is the number of hourly buckets in the held-out test split (total $\sum_s n_s = 6422$).

We then weight each seasonal R_s^2 by its share of the test data:

$$w_s = \frac{n_s}{\sum_t n_t}, \qquad R^2_{\text{weighted}} = \sum_s w_s \, R_s^2. \tag{3}$$

Table 2 summarizes the per-season R_s^2 for each model, and Table 3 reports the counts n_s, weights w_s, and their contributions $w_s \, R_s^2$ to the weighted average.

Seasonal breakdown for Random Forest:

- **Winter (Dec–Feb)** ($n = 744$, $R^2 = 0.6211$, $w = 0.1159$)
- **Spring (Mar–May)** ($n = 1358$, $R^2 = 0.8792$, $w = 0.2115$)
- **Summer (Jun–Aug)** ($n = 2136$, $R^2 = 0.9037$, $w = 0.3326$)
- **Fall (Sep–Nov)** ($n = 2184$, $R^2 = 0.8929$, $w = 0.3400$)

These results motivate *seasonally adaptive rebalancing* of the fleet:

- Exploit high summer accuracy for aggressive station restocking when demand peaks.
- Incorporate holiday calendars, local event schedules, and station-level microclimate measures to boost winter and holiday-period forecasts where the model currently struggles.
- Consider a hybrid approach that blends Random Forest with simpler linear corrections around known anomalies (e.g., major city events).

6 Conclusions and Future Work

Our integrated analysis confirms that weather and temporal factors critically shape bike-sharing demand. Hour of the day and temperature drive the bulk of variance, with humidity, wind, and precipitation as secondary modifiers. We recommend seasonally tailored service levels, regime-based rebalancing, and strategic station placement near both CBD and recreational areas. Future research will integrate real-time event feeds, demographic overlays, and dynamic pricing to further elevate model fidelity and operational efficiency.

References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

[4] Steadman, R.G. (1979) The assessment of sultriness. Part I: A temperature-humidity index based on human physiology and clothing science. *Journal of Applied Meteorology*, 18(7), 861–875.

[5] Kahneman, D. & Tversky, A. (1979) Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291.

[6] Vuchic, V.R. (2007) *Urban Transit Systems and Technology*. Wiley.

[7] NOAA. Chicago O'Hare monthly normals (1991–2020). *NOAA NCEI*. Retrieved from `https://www.weather.gov/lot/ord_rfd_monthly_yearly_normals`, 2024.

[8] The Trust for Public Land. Chicago Parks and Recreation. Retrieved from `https://tpl.org/city/chicago-illinois`, 2021.

[9] Curiel, "Data Mining Weather Chicago." Kaggle. `https://www.kaggle.com/code/curiel/data-mining-weather-chicago`, accessed 2024.

[10] Divvy Bikes. System Data portal. `https://divvybikes.com/system-data`, accessed 2024.