UNIVERSITÀ DEGLI STUDI DI BERGAMO
Dipartimento di Ingegneria Gestionale, dell'Informazione e della Produzione

Urban Bike Sharing in Washington, DC: A Spatio-Temporal and Statistical Analysis

"Optimizing bike-sharing through data-driven approaches"

Academic Year: 2024 - 25

Master Degree in COMPUTER ENGINEERING

Data Science and Data Engineering Curriculum

Course:
Statistics for High Dimensional Data and Compstat Lab

Professor
Prof. Francesco Finazzi

Student
Arash Abedi

Introduction

I. Key Question:

- Can bike-sharing systems in urban areas be optimized using spatio-temporal models?

II. Main Approach:

- Analyze 2023 bike-sharing data from Washington, DC using statistical models.

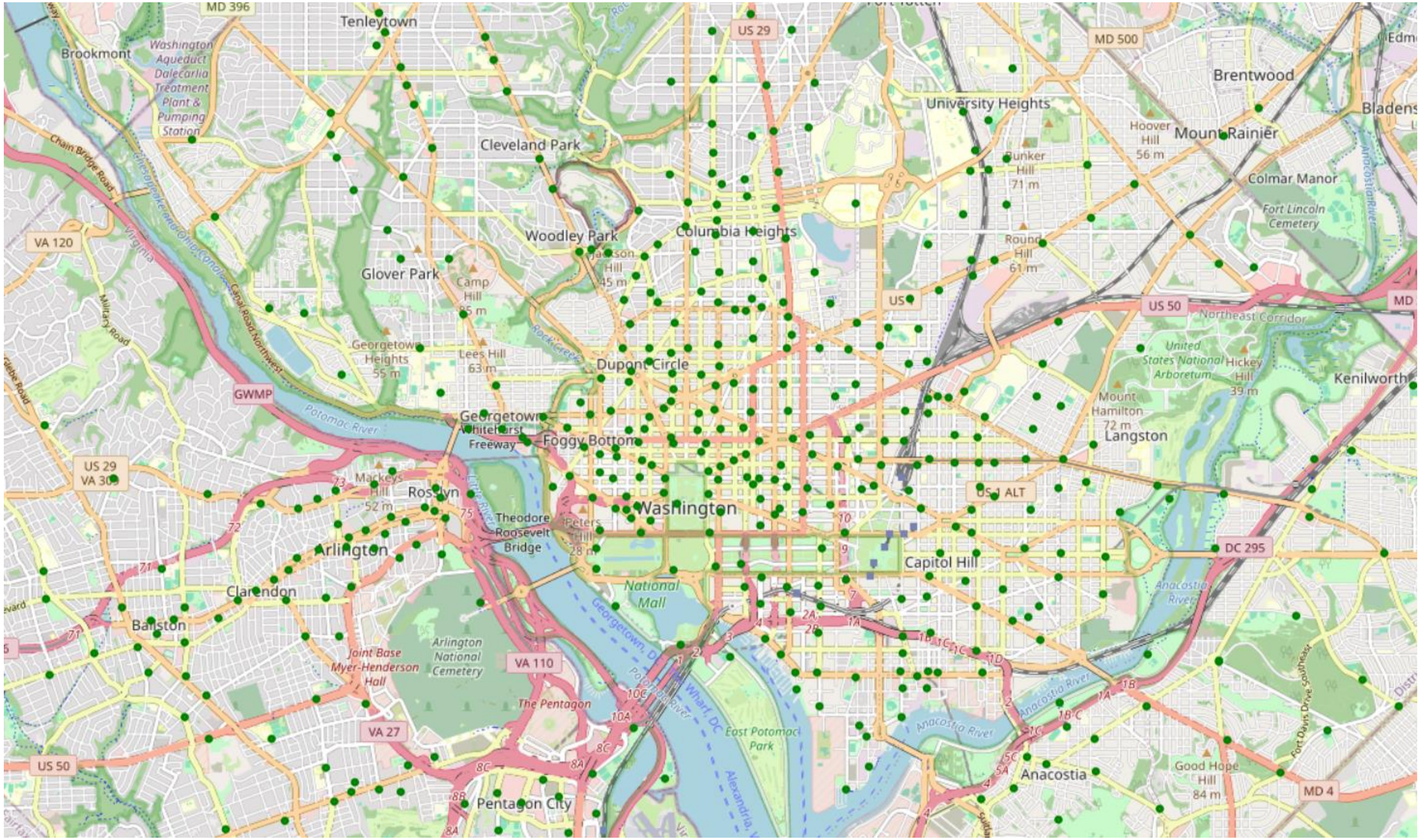
III. Goal:

- Improve resource allocation and operational efficiency.

Dataset Description

- **Data Sources:**
  - Capital Bikeshare: 700+ stations, 6,000+ bikes
  - Weather Data: Historical weather records (temperature, precipitation, etc.)
- **Key Variables:**
  - Bike pickups, trip duration, station locations
  - Weather conditions (temperature, humidity, precipitation, etc.)

Map of Washington, DC: Bike Stations Highlighted



Exploratory Data Analysis

- **Bike Usage Patterns:**
  - Daily and hourly aggregation of rides
  - Weekends/holidays vs. working days

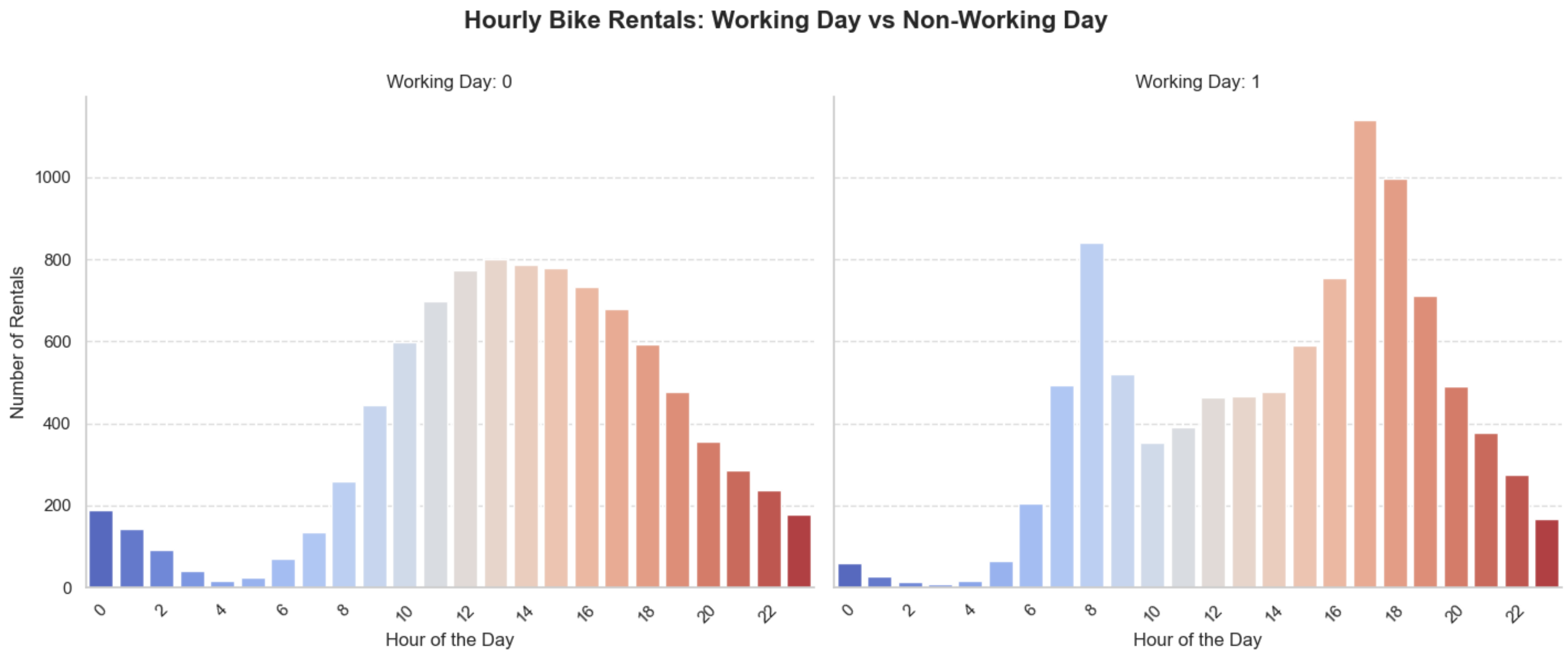
- **Key Statistics:**

Variable	Unit	Min	Max	Mean	Median	Std	Skew	Kurt
Mean pickups	-	1.67	34.23	13.24	12.99	5.85	1.08	2.21
Mean trip duration	min	0.48	837.54	18.10	11.42	33.82	14.73	411.14

Key descriptive statistics for bike-sharing variables.

Hourly Rental Patterns

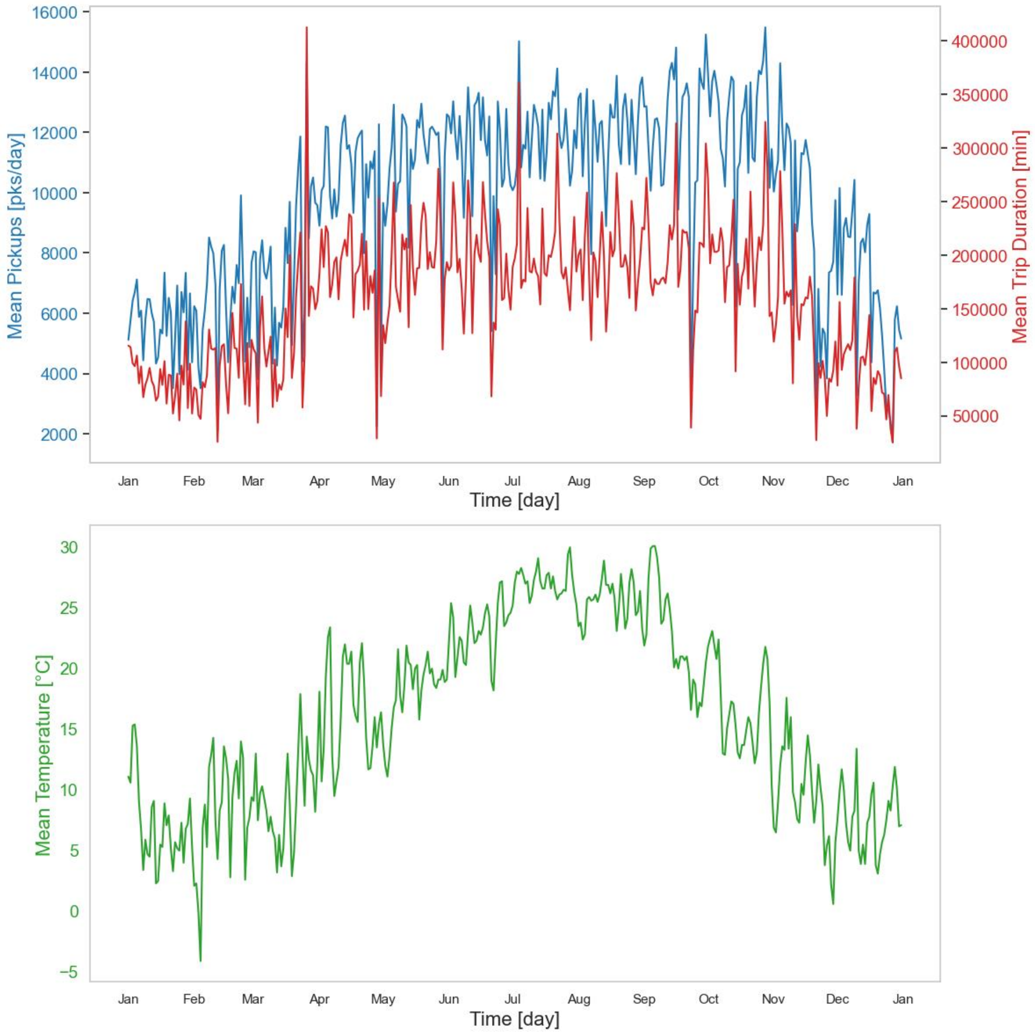
- **Observation:**
  - Peak hours on working days: Morning and evening (*commuting times*).
  - Peak hours on weekends: Midday and afternoon (*recreational use*)



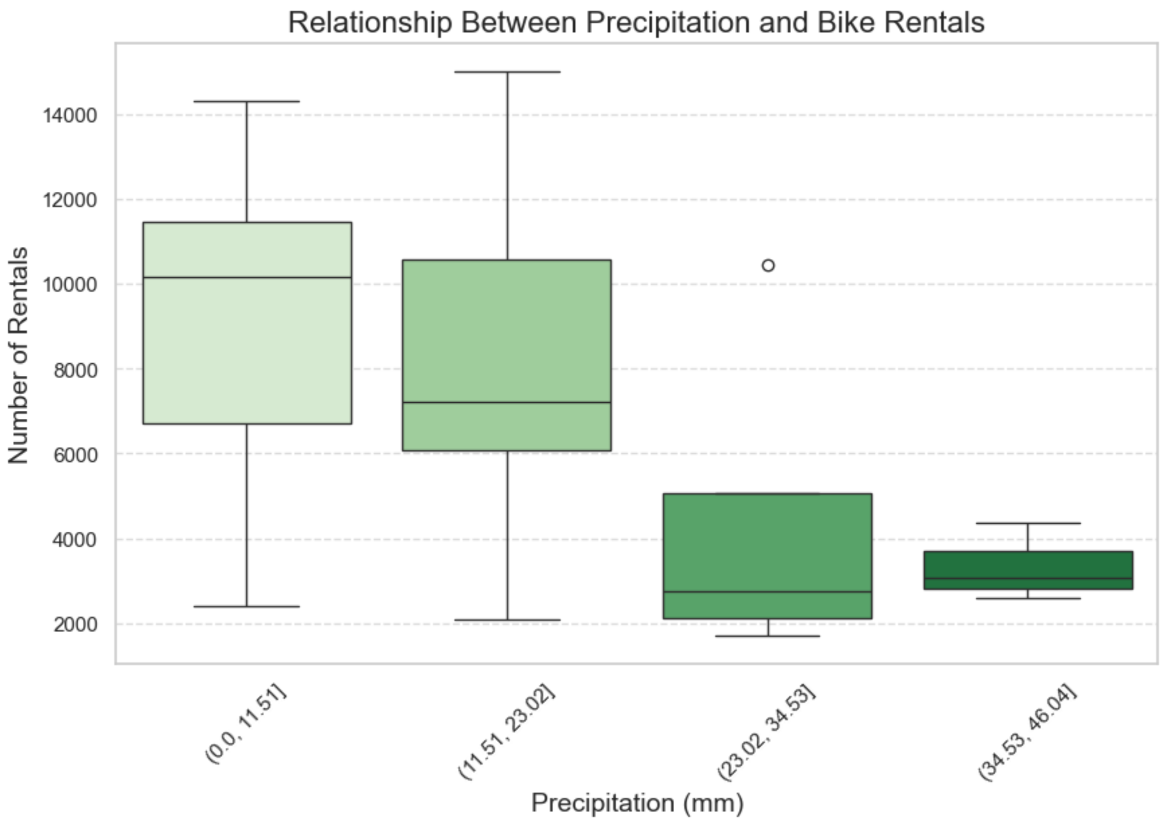
Impact of Weather on Bike Usage

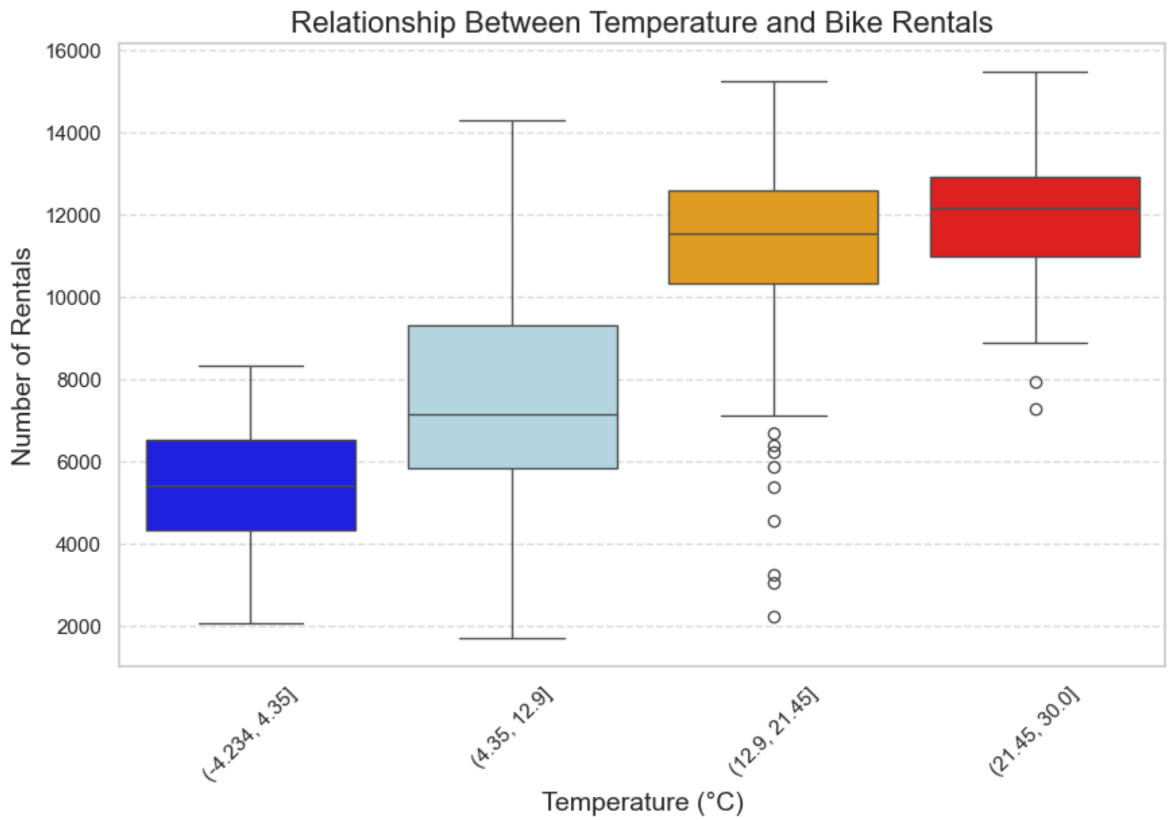
- **Hypothesis:**
  - Weather significantly influences bike rental demand.

- **Findings:**
  - Higher temperature → More rentals & longer trips
  - High precipitation → Lower rentals

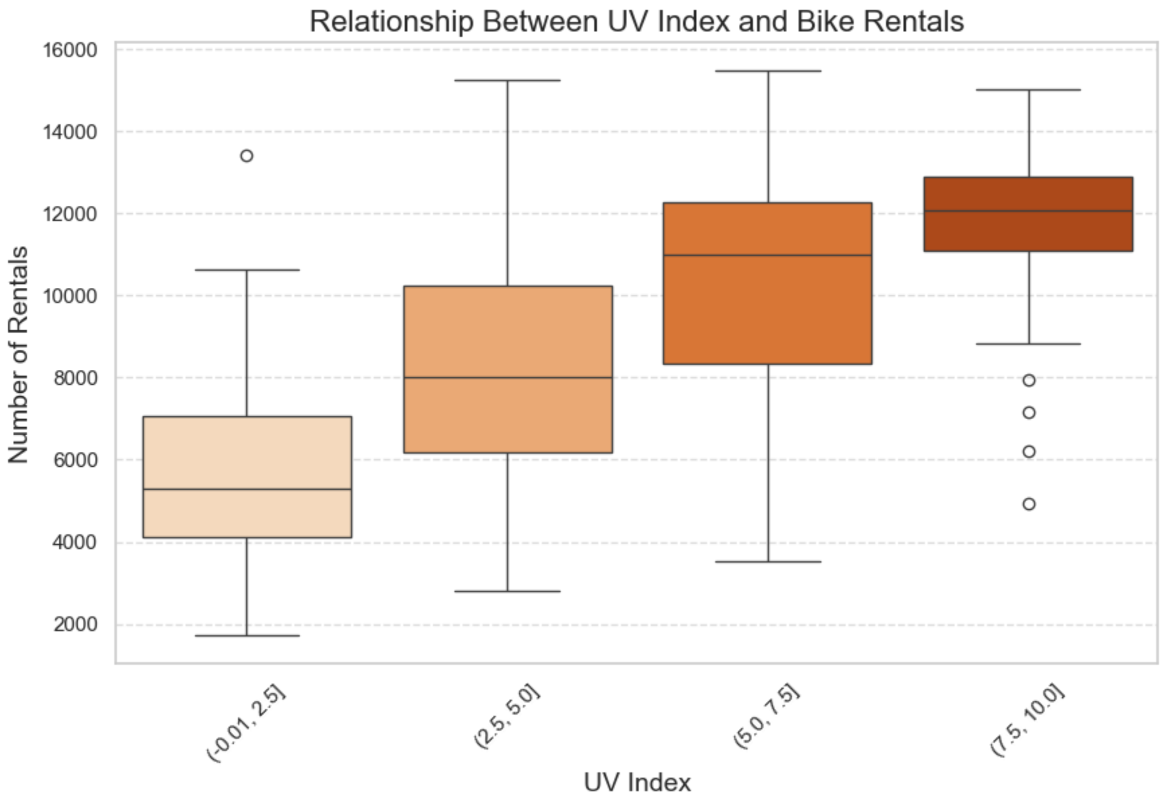


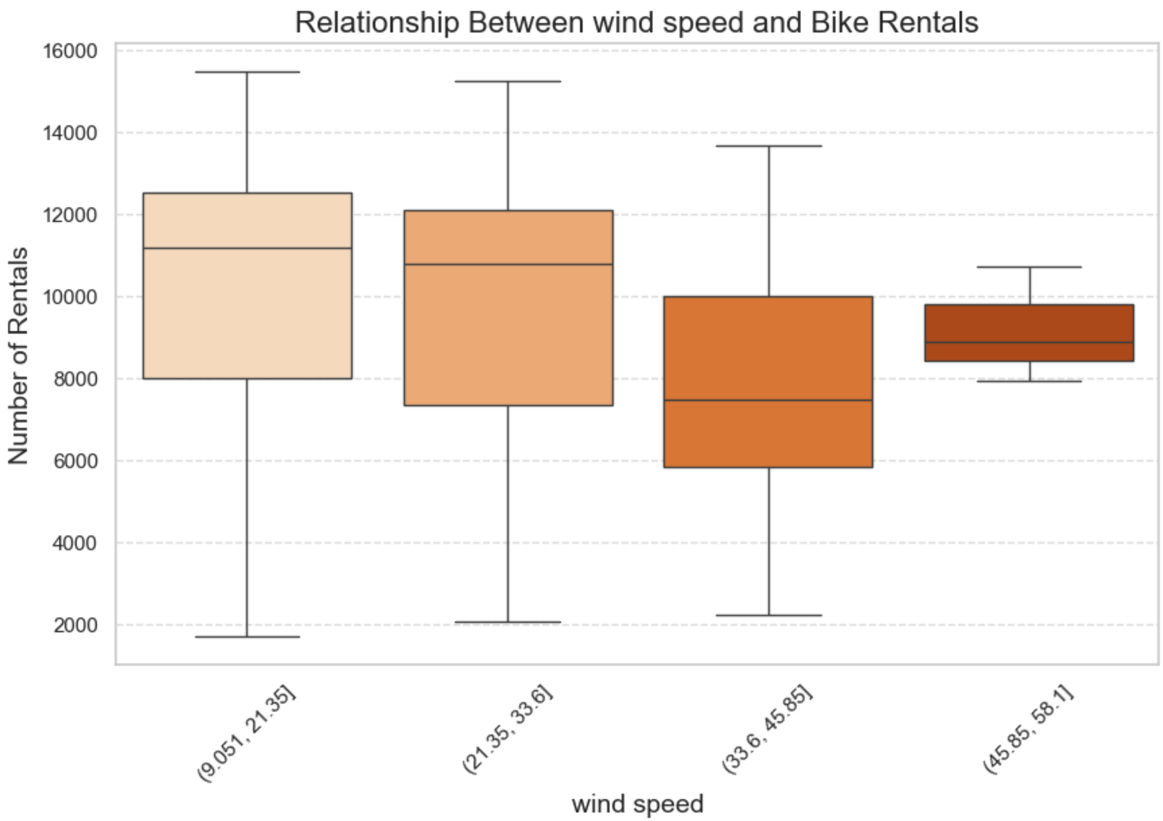
Impact of Weather on Bike Usage





Impact of Weather on Bike Usage





Research Objectives

- **Main Questions:**
  - Which external factors (weather, events) impact bike-sharing demand?
  - Which spatio-temporal models best predict bike rentals and trip duration?
- **Outcome:**
  - Develop a framework to improve urban bike-sharing efficiency.

Spatio-Temporal Models Used

- **Models Implemented:**
  - Dynamic Coregionalization Model (DCM)
  - Hidden Dynamic Geostatistical Model (HDGM)
- **Key Differences:**
  - *DCM*: Accounts for multiple latent processes
  - *HDGM:* More flexible and computationally efficient

Mathematical Foundations of DCM

- **Dynamic Coregionalization Model (DCM)**

$$y(s,t) = x_\beta(s,t)'\beta + x_z(s)'z(t) + \alpha w(s,t) + \varepsilon(s,t)$$

$$z(t) = Gz(t-1) + \eta(t)$$

- **Key Terms:**
  - $x_\beta(s,t)$: *Time-varying covariates (e.g., weather).*
  - $z(t)$: *Latent temporal process with Markov dynamics.*
  - $w(s,t)$: *Spatially correlated random effect.*

Mathematical Foundations of HDGM

- **Univariate Hidden Dynamic Geostatistical Model (HDGM)**

$$y(s,t) = x_{\beta}(s,t)'\beta + \alpha z(s,t) + \varepsilon(s,t)$$

$$z(s,t) = gz(s,t-1) + \eta(s,t)$$

- **Key Terms:**
  - $z(s,t)$*: Unified spatio-temporal latent process.*
  - *Simpler structure than DCM (single latent component).*

Model Selection & Validation

- **Methodology:**
  - Covariate selection based on statistical significance
  - Training (70%) & testing (30%) split
- **Performance Metrics:**
  - RMSE (Root Mean Squared Error)
  - MSE (Mean Squared Error)

Results – DCM Model

- **Key Findings:**
  - Temperature, UV index, and precipitation are the strongest predictors.
  - Model struggles slightly with overfitting.

- **Performance Metrics:**
  - RMSEs for pickups: 0.6045
  - RMSEs for trip duration: 0.8403

Results – HDGM Model

- **Key Findings:**
  - Outperforms DCM in predictive accuracy.
  - More stable due to its simplified latent process.

- **Performance Metrics:**
  - RMSEs for pickups: 0.3609
  - RMSEs for trip duration: 0.7024

Model Results Summary

- **Most Significant Covariates:**
  - **Temperature:** Strong positive effect on pickups and trip duration.
  - **Precipitation:** Negative effect on pickups.
  - **UV Index:** Positive effect (likely linked to sunny weather).
  - **Weekends/Holidays:** Increased trip duration (likely due to leisure rides).

- **Performance Comparison:**
  - **HDGM outperformed DCM** in predicting pickups and trip duration, likely due to its simpler structure.
  - For trip duration, both models performed similarly, but HDGM was slightly better.

Comparison of Spatio-Temporal Models

Feature	DCM	HDGM
Model Complexity	Higher	Lower
Latent Processes	Two	One
Computational Cost	Higher due to complex structure	Lower, more efficient
Best for Predicting	Trip duration	Bike pickups
Key Covariates	Temperature, UV index, precipitation, wind speed, weekends/holidays	Temperature, UV index, precipitation, weekends/holidays
RMSE (Pickups)	**0.6045** (full) / **0.6023** (selected)	**0.3673** (full) / **0.3650** (selected)
Log-Likelihood (Best Model)	-339.12 (lower, worse fit)	-248.72 (higher, better fit)
Overall Performance	Good, but slightly overfits	More stable, better generalization

Comparative Performance Analysis

- **Observation:** HDGM outperforms DCM in most cases.
- **Why?**
  - HDGM's single latent component avoids overfitting.
  - Captures spatial dependencies effectively.

Index	*Variable*	*Model*	*Min.*	*Max.*	*Mean*	*Median*	*Std*
RMSE$_t$	Pickups [pks/day]	DCM	0.56	54.50	13.20	11.30	10.10
		HDGM	0.11	38.20	11.80	10.90	6.50
	Duration [min]	DCM	5.50	2300	45	13	180
		HDGM	5.20	2280	43	12	175
RMSE$_s$	Pickups [pks/day]	DCM	5.20	49.50	12.80	8.10	11
		HDGM	3.80	32.00	11	7.90	7.50
	Duration [min]	DCM	10	630	110	48	155
		HDGM	6	625	108	47	153

Limitations & Discussion

- **Data Limitations:**
  - Only one year (2023) analyzed → No long-term trends.
  - Only Washington, DC → Findings may not generalize to all cities.
- **Model Limitations:**
  - Assumes stationarity, may not capture behavioral shifts.
  - High computational cost.

Key Takeaways

- **Weather Drives Demand:**
  - **Temperature** (+), **UV Index** (+), and **Precipitation** (−) are the strongest predictors of bike rentals.
  - Warmer, sunnier days increase rentals by ~20-30%; rain reduces demand.
- **Temporal Patterns Matter:**
  - **Peak Hours:** Morning/evening (workdays) vs. midday (weekends).
  - **Weekends/Holidays**: Longer trips (leisure use).
- **HDGM Outperforms DCM:**
  - **Lower RMSE (0.36** vs. 0.60 for pickups) → More accurate predictions.
  - **Simpler Structure**: Single latent process avoids overfitting.

Conclusion

- **Key Findings:**
  - Temperature, precipitation, and UV index are the main predictors.
  - HDGM is the best-performing model for predictive analysis.
- **Impact:**
  - Helps optimize bike allocation.
  - Supports data-driven urban mobility planning.

References

- **Spatio-Temporal Modeling Framework**
  - Finazzi, F., Wang, Y., & Fassò, A. (2021). D-STEM v2: A software for modeling functional spatiotemporal data. Journal of Statistical Software, 99(10), 1-29. DOI: 10.18637/jss.v099.i10
- **Data Sources**
  - Bike-sharing data: Capital Bikeshare System Data (2023).
  - Weather data: Visual Crossing Weather API.
- **Statistical & Modeling Tools**
  - MATLAB – Used for implementing D-STEM and statistical analysis
  - D-STEM v2 – Toolbox for spatio-temporal data modeling
  - Python – Used for data preprocessing, exploratory data analysis, and visualization