Uncovering Hidden Mobility Regimes in Capital Bikeshare Data Using Hidden Markov Models and Bayesian Inference

Berk Kasimcan

Department of Bioengineering, George Mason University

STAT 646: Probabilistic Machine Learning

Professor Anand Vidyashankar

Tuesday May 6th, 2025

Project File

Due to Dropbox's 2GB storage limitation, I was unable to upload the full project ZIP file to that platform. Instead, the complete dataset, scripts, and related materials have been uploaded to my Google Drive. You can access the full project archive using the following link:

https://drive.google.com/file/d/1R3Qhl2D1kKyvEj3iKDj7-wV2bx3fHW7A/view?usp=sharing

Please feel free to reach out if you encounter any access issues.

Table of Contents

Final Project Requirements

The final project is designed to demonstrate my ability to apply probabilistic graphical models and Bayesian methods to a real-world dataset, showcasing both technical mastery and clear statistical interpretation.

Dataset Requirement

- Use the Capital Bikeshare Trip History Dataset.
- Focus on 2021–2024 data (post-COVID-19 era).

  Structure and clean the dataset appropriately (parse timestamps, remove anomalies).

Probabilistic Modeling

- Apply Hidden Markov Models (HMMs):
  - Discover latent demand patterns (e.g., commuting vs. leisure).
  - Estimate state transitions and emission probabilities.
- Apply Bayesian Regression:
  - Model bike demand using time-based and external features (e.g., weather, time of day).

    Capture uncertainty in predictions (posterior predictive intervals).

Dimensionality Reduction (Optional but Encouraged)

- Use Principal Component Analysis (PCA) to visualize patterns if helpful.
- Visualize temporal patterns and cluster demand behavior.

Model Evaluation

- Evaluate models using:
  - AIC, BIC (Model selection metrics)
  - Log-likelihood comparisons
  - Posterior credible intervals for Bayesian models
- If feasible, also assess forecasting accuracy on a holdout set.

Visualization and Interpretation

- Clearly visualize:
  - Latent states discovered by HMMs over time.
  - Posterior predictive distributions from Bayesian models.
- Relate hidden states to real-world events (e.g., rush hours, weekends, holidays).

Report Guidelines

Prepare a detailed, APA-format written report (~10–15 pages of written content) including the following sections:

Introduction

- Briefly introduce Capital Bikeshare and why urban mobility matters.
- Frame the problem statement: how to uncover hidden behavior patterns post-pandemic.
- State my goals and define my research questions.

Dataset Description

- Summarize the structure of the dataset.
- Mention any data limitations or preprocessing challenges.

Project Hypothesis/Research Question

Methodology

- Clearly explain:
    - Why HMMs are appropriate for latent state modeling.
    - Why Bayesian regression is useful for demand prediction with uncertainty.
    - Why dimensionality reduction can aid interpretation.
- Include both math formulations and verbal descriptions.

Analysis

- Step-by-step walkthrough of my pipeline:
    - Data loading
    - Preprocessing
    - Modeling
    - Model evaluation
    - Include figures, tables, and plots for clarity.

Results

- Present my:
    - Latent states
    - Predictive distributions
    - Model comparisons (AIC, BIC)
    - Provide quantitative results and visualizations.

Interpretation

- Discuss meaning of latent states (what does each hidden state represent?)
- Discuss uncertainty (where predictions are wide/narrow).
- Reflect on model assumptions.

Discussion and Conclusion

- What patterns did you find?
- How do my findings connect to real-world mobility behavior?
- Comment on limitations and future work directions.

Additional Expectations

- Use clear English, proper APA-style citations, and professional formatting.
- Include clear mathematical notation where appropriate.
- Visuals should be:
    - Well-labeled (axes titles, figure captions).
    - Referenced in the text (e.g., "see Figure 2").
- Interpret all results , don't just plot , explain what the plots show.
- Address assumptions:
    - (e.g., Markov assumptions in HMMs, conjugate priors in Bayesian models)
- Be creative: bonus for trying extra techniques like weather data merging, anomaly detection, or station clustering.

Abstract

The COVID-19 pandemic dramatically altered patterns of urban mobility, disrupting traditional rush hour dynamics and shifting recreational behaviors. This study leverages post-pandemic trip data from Washington, D.C.'s Capital Bikeshare system (2021–2025) to probabilistically model hidden demand regimes and predict ride volume under uncertainty. Structured trip records, containing timestamps, station metadata, and ride durations, serve as the foundation for analysis. After extensive data preprocessing, Hidden Markov Models (HMMs) are employed to uncover latent states corresponding to commuting peaks, leisure periods, and low-demand hours. To complement this unsupervised learning, Bayesian regression techniques are utilized to predict hourly ride counts, incorporating temporal covariates such as time of day, day of week, and month. This probabilistic approach captures not only expected demand but also quantifies uncertainty through posterior distributions.

Model selection procedures based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are applied to compare HMM-based segmentation with baseline clustering methods and classical time series models. Further, uncertainty quantification is achieved via posterior predictive intervals, providing realistic assessments of demand fluctuations across time. Visualizations of hidden states overlaid onto trip timelines reveal meaningful temporal shifts, particularly diminished weekday rush hours and heightened weekend leisure activity post-COVID.

The findings demonstrate that probabilistic graphical models offer significant advantages in modeling complex, dynamic transportation systems by providing interpretable latent states and uncertainty-aware predictions. The study also offers operational insights for urban bikeshare systems, suggesting adaptive bike redistribution strategies based on inferred latent demand regimes. This research bridges applied machine learning and urban analytics, offering a robust, interpretable framework for understanding and responding to evolving patterns in micro-mobility.

Acknowledgment and Disclaimer

This project was developed as part of the coursework for a graduate-level statistical machine learning class. Several tools and resources were used to support the formatting, presentation, and technical implementation of the analysis. Specifically, LaTeX, Equatio, and ChatGPT were employed to render and refine the mathematical equations and probabilistic model descriptions throughout the report.

Special thanks are extended to Professor Anand Vidyashankar for providing access to course slides, shared code, and annotated materials via Dropbox. Many of the methodological insights, coding structures, and explanatory frameworks were directly inspired by class lectures and these shared instructional resources.

For grammar refinement, clarity improvements, and logical restructuring, the following tools were used: Grammarly, Google Gemini 3.5 Pro, and ChatGPT. In addition, ChatGPT was used to assist in generating explanatory code comments and resolving certain programming errors that were not initially understood.

While every effort was made to condense this report into the requested page limit, the final document exceeds that length due to the inclusion of numerous equations, technical derivations, and many figures. However, the core body of text, excluding visualizations and extended mathematical appendices, remains within an approximate 10 ~ 15 page narrative range.

All supplementary materials, including reference code, data outputs, and extended figures, are available in the google drive link with the zip file. Addition to that, the class content shared in the dropbox including codes and explanations were used in this project as well.

Introduction

The COVID-19 pandemic significantly reshaped patterns of urban mobility (Open Knowledge Repository, 2025), altering traditional commuting behaviors and recreational transportation usage across major cities. In Washington, D.C., the Capital Bikeshare system, an extensive network of publicly available bicycles, provides a unique lens through which to study these post-pandemic shifts. As remote work arrangements, flexible schedules, and changes in public transportation usage became more widespread, it is critical to investigate how demand for bikeshare services evolved during this period. This study aims to uncover hidden patterns in Capital Bikeshare demand from 2021 to 2025 by applying probabilistic modeling techniques, including Hidden Markov Models (HMMs) and Bayesian regression. By treating bike usage as a stochastic process influenced by time, weather, and calendar effects, we seek not only to predict future demand but also to identify latent behavioral states, such as weekday commuting peaks and weekend leisure periods, that govern system usage. Unlike traditional deterministic approaches, probabilistic graphical models provide a structured framework for modeling uncertainty and hidden regimes within observed data (Hamilton & Wichman, 2018). The broader objective is to deliver interpretable, uncertainty-aware insights that can support operational decision-making, such as bike redistribution strategies and infrastructure planning, thereby demonstrating the practical value of applying probabilistic machine learning to real-world urban mobility systems.

Dataset Description

The core dataset used in this study comprises structured trip-level records from the Capital Bikeshare system (System Data | Capital Bikeshare, n.d.), covering January 2020 through early 2025. Each observation corresponds to an individual bike rental and includes detailed attributes: precise start and end timestamps, trip duration in seconds, start and end station IDs and names, a unique bike identifier, and user membership classification (e.g., casual vs. subscriber). These records are stored as monthly CSV files, each ranging in size from approximately 4 MB during low-ridership months (typically winter) to over 27 MB in peak demand periods (such as summer). Monthly files typically contain between 40,000 and 250,000 trip records depending on seasonality, yielding a full dataset of several million rows, sufficient for robust temporal modeling and probabilistic inference.

Each file can be conceptualized as an n×p matrix, where n denotes the number of trips in a given month and p refers to the number of recorded attributes (generally 8–9, depending on whether weather augmentation is applied). Upon concatenation, the resulting matrix forms a comprehensive panel of micromobility activity across Washington, D.C., and adjacent jurisdictions, with extremely low sparsity. Capital Bikeshare's internal data cleaning pipelines exclude trips shorter than 60 seconds and remove non-user-related rides such as system maintenance or redistribution trips, which helps ensure the behavioral signal in the data reflects real user intent and demand.

To enable time-based analyses, timestamp fields are decomposed into multiple temporal covariates: hour of day (0–23), day of week (0–6), calendar month (1–12), year, and an indicator for whether the trip occurred on a U.S. federal holiday or weekend. These engineered features allow the dataset to support high-resolution modeling of cyclical, seasonal, and event-driven trends in usage. Aggregation to the hourly level is performed to facilitate integration with weather and latent state decoding, and to align with the temporal granularity required by the Hidden Markov Models (HMMs) and Bayesian forecasting models used in this study.

In addition to ride-level data, the analysis incorporates external weather information to account for environmental drivers of bikeshare demand. Specifically, historical daily weather records were sourced from a Kaggle dataset curated by Ta-wei Lo (Lo, 2015), which covers atmospheric conditions in Washington, D.C., from August 2015 through July 2024. This dataset includes over 30 meteorological variables, from which a curated subset was retained based on theoretical and empirical relevance to micromobility behavior. The selected features are: average temperature (temp), maximum and minimum temperature (tempmax, tempmin), relative humidity, precipitation level, wind speed, sea level pressure, cloud cover, visibility, and UV index. These variables influence rider comfort, safety, and trip feasibility, for example, high humidity and precipitation typically depress ridership, while mild temperatures and clear visibility encourage trips. Variables with low variance or limited explanatory power in the D.C. context (e.g., snowfall, solar radiation, moon phase) were excluded to streamline the model and reduce noise.

The merged dataset combines trip records with temporal features and weather variables, producing a high-dimensional structured matrix suitable for supervised learning and probabilistic time series modeling. This design matrix X serves as the input to both Hidden Markov Models

for latent regime identification and Bayesian regression models for ride count forecasting. The response variable Y represents hourly aggregated trip counts. Given its completeness, temporal resolution, and behavioral integrity, the dataset supports the application of advanced modeling frameworks aimed at uncovering hidden states, quantifying uncertainty, and characterizing shifts in urban mobility behavior across a multi-year, post-pandemic period.

Goals

The central goal of this project is to build a probabilistic framework that can uncover hidden mobility patterns and produce uncertainty-aware forecasts of bikeshare demand in the Capital Bikeshare system between 2021 and 2025. In the wake of COVID-19, travel behavior has become increasingly irregular and harder to predict using traditional models. By applying statistical learning methods that account for latent behavioral shifts and probabilistic uncertainty, this study aims to offer a more realistic and interpretable model of urban micromobility dynamics.

The first major objective is to use Hidden Markov Models (HMMs) to detect and characterize latent demand regimes (Marius-Christian Frunza, 2015), unobserved behavioral states that explain transitions in hourly usage patterns. These states are expected to align with distinct mobility modes, such as commuter activity during weekday mornings, recreational surges on weekends, or low-activity off-peak hours. By learning not only the state distributions but also their transition probabilities, the HMM framework reveals how usage patterns evolve over time and respond to external signals like weather, holidays, and seasonality. This enables a dynamic lens into bikeshare usage that can’t be captured with static or purely linear models.

The second key goal is to apply Bayesian regression to predict hourly ride counts while explicitly modeling uncertainty. Rather than producing a single point estimate, the Bayesian approach yields full posterior distributions over model parameters and predictions, allowing for richer interpretation and risk-aware decision-making. Covariates include time-of-day, day-of-week, temperature, precipitation, and other relevant weather indicators. This uncertainty-aware modeling is especially valuable in urban transportation contexts where demand is volatile and highly sensitive to environmental and temporal factors.

Beyond time series modeling, the project also explores spatial heterogeneity in bikeshare usage. By clustering stations based on origin-destination flow patterns, the analysis seeks to

uncover spatial regimes, such as commuter-heavy corridors or leisure-centric zones, that interact differently with latent states and temporal drivers. Additionally, the study examines how external disruptions (e.g., holidays, extreme weather) influence both state transitions and demand, offering insight into the resilience and adaptability of the system.

Ideally, this research aims to synthesize temporal, spatial, and external dimensions into a unified probabilistic model that not only explains how bikeshare behavior varies but also why. The findings are intended to support both theoretical understanding of micromobility behavior and practical decision-making by operators and planners. By capturing latent structure, behavioral uncertainty, and dynamic shifts, this framework contributes a more adaptive and actionable approach to managing shared mobility in a post-pandemic urban landscape.

Project Hypothesis/Research Question

The central hypothesis of this study is that Capital Bikeshare usage patterns during the post-COVID-19 period are governed by latent behavioral regimes that can be uncovered through the application of probabilistic graphical models, specifically Hidden Markov Models (HMMs). These hidden states, though unobservable directly, are presumed to represent distinct patterns of urban mobility such as commuting peaks, weekend leisure usage, and midday lulls. Traditional deterministic approaches fail to account for the stochastic, evolving nature of human transportation behavior; therefore, this study employs probabilistic modeling to capture both the observable variation in trip counts and the unobserved transitions between underlying behavioral states.

Formally, let Y_t denote the observed number of bike trips at discrete time *t* (e.g., an hourly aggregation), and let Z_t represent the latent state at time t, $Z_t \in \{1, 2, \ldots, K\}$ or some unknown but learnable number of hidden states *K* (Hidden Markov Model in Machine Learning, 2023). The joint distribution of the data under the HMM is expressed as

$$P(Y_{1:T}, Z_{1:T}) = P(Z_1) \prod_{t=2}^{T} P(Z_t \mid Z_{t-1}) \prod_{t=1}^{T} P(Y_t \mid Z_t),$$

where $P(Z_t \mid Z_{t-1})$ represents the state transition probabilities and $P(Y_t \mid Z_t)$ captures the likelihood of observing a particular demand given the current latent regime. The modeling goal is to estimate both the hidden state sequence $Z_{1:T}$ and the underlying parameters that govern the transitions and emissions, thereby revealing structured behavioral dynamics not immediately visible in the raw trip data.

Complementing the discovery of latent regimes, a second major objective of this research is to construct a predictive model of bike demand that accounts for uncertainty using Bayesian statistical inference. Specifically, ride counts Y_t will be modeled as a function of covariates X_t — including hour of day, day of week, month, holiday indicators, and weather conditions — through a Bayesian linear regression framework:

$$Y_t = X_t^\top \beta + \epsilon_t, \quad \epsilon_t \sim \mathrm{N}(0, \sigma^2),$$

where β represents the regression coefficients and σ^2 denotes the variance of the error term. Placing prior distributions over β and σ^2, the Bayesian framework allows the derivation of a posterior predictive distribution,

$$P(Y_{t+1} \mid X_{t+1}, Y_{1:t}, X_{1:t}),$$

which offers not only a point forecast but also a credible interval reflecting uncertainty. This probabilistic forecasting is critical for transportation systems, where variability due to human behavior, weather, and external events can be substantial and deterministic predictions often fail.

Further extending the analysis, this study integrates spatial modeling by clustering bikeshare stations based on their departure and arrival profiles. By examining stations as nodes with distinctive flow patterns, the research seeks to identify commuter hubs, recreational hotspots, and anomalous stations whose usage patterns deviate from broader trends. This spatial heterogeneity provides another layer of complexity and realism in modeling urban micromobility systems, as demand is not spatially uniform across a city's network.

Further extending the analysis, this study incorporates spatial modeling by clustering Capital Bikeshare stations based on their departure and arrival patterns, particularly focusing on peak usage times and overall ride volumes. This spatial segmentation reveals commuter hubs characterized by concentrated weekday morning departures, recreational hotspots with elevated weekend activity, and consistently underutilized stations that may benefit from reallocation or targeted interventions. These findings complement the temporal regimes identified through

Hidden Markov Models (HMMs), providing a more holistic understanding of system dynamics by linking station-level patterns to broader behavioral states.

In addition to spatial differentiation, the study models external shocks such as holidays and extreme weather events. These contextual factors are shown to significantly influence both observed demand and latent state transitions. For example, holidays tend to increase the persistence of low-demand states, while favorable weather conditions are associated with shifts toward high-activity regimes. By explicitly integrating these variables into the modeling framework, the analysis offers a more realistic depiction of demand variability across time and context.

These components give rise to a cohesive set of research questions: First, can an HMM, trained on post-pandemic bikeshare data, uncover latent behavioral states that correspond to interpretable mobility patterns such as weekday commuting, midday transitions, or weekend leisure riding? Second, to what extent do temporal variables—such as hour of day, day of week, and holiday presence—alongside weather conditions influence the probability of entering or persisting in these latent states? Third, does incorporating these hidden regimes into a Bayesian demand forecasting model improve predictive accuracy and yield better-calibrated uncertainty estimates compared to models relying solely on observable covariates? Lastly, can station-level clusters derived from usage profiles provide meaningful operational insights, and how do these spatial dynamics interact with the temporal behavioral regimes discovered through probabilistic modeling?

Together, these questions guide the study's central aim: to demonstrate that probabilistic, uncertainty-aware modeling provides a richer, more operationally valuable understanding of urban micro mobility dynamics than traditional deterministic methods. The resulting framework offers interpretable latent structure, demand forecasts equipped with credible intervals, and actionable station-level insights that can inform smarter bike redistribution, planning, and service optimization in a rapidly evolving post-COVID landscape.

Methods

Exploratory Data Analysis and Feature Engineering

Before diving into any probabilistic modeling, I began with a thorough exploratory data analysis (EDA) to understand the underlying structure and temporal dynamics of the Capital

Bikeshare dataset (Step by Step Process in Exploratory Data Analysis and Feature Engineering | Kaggle, 2025). This process was essential for identifying recurring patterns in ride activity and for shaping the features that would later feed into the Hidden Markov Models (HMMs) and Bayesian regression. I aggregated ride counts at multiple time resolutions, hourly, daily, and monthly, which immediately revealed strong diurnal cycles with sharp peaks during morning and evening commute hours on weekdays, along with flatter, more dispersed usage during weekends. These patterns informed the creation of several time-based features, including 'hour of day' to capture within-day variation, 'day of week' to distinguish between workday and weekend behavior, and a binary 'weekend' flag to isolate non-commuting periods. I also defined categorical 'season' labels (e.g., Winter, Spring, etc.) to account for broader environmental effects like daylight and weather shifts. Where available, I merged external weather datasets, adding numerical features like daily temperature, precipitation, and humidity, variables known to influence micro-mobility behavior. Visualization played a central role throughout this process: I used histograms, bar charts, and temporal heatmaps to expose periodic trends, identify outliers (like unrealistically long trip durations), and spot any inconsistencies or anomalies that could distort model results. This stage was about more than just cleaning data, it was about building a reliable, interpretable foundation that would shape everything downstream, from inferring latent behavioral states with HMMs (Daniel & Martin, 2023) to producing uncertainty-aware demand forecasts using Bayesian methods.

Hidden Markov Model for Latent State Discovery

Hidden Markov Models (HMMs) provide a probabilistic framework for modeling time series data where the system is assumed to transition between a finite number of unobserved, or hidden, states over time. A probabilistic framework refers to a modeling approach where uncertainty is treated as a fundamental feature of the system rather than as noise or error. Instead of assuming that observed outcomes are fixed or deterministic, probabilistic models represent variables, parameters, and relationships between them as random quantities described by probability distributions. This allows for the formal handling of uncertainty in both the underlying data-generating processes and the predictions made from the model. In the context of urban mobility analysis, a probabilistic framework enables more realistic modeling of human

behavior by capturing not only the most likely outcomes but also the range and likelihood of alternative possibilities (Pandolfi et al., 2023).

In the context of bikeshare demand modeling, it is reasonable to hypothesize that observed trip counts are not generated uniformly across time but instead are driven by latent behavioral regimes, such as weekday commuting bursts, weekend leisure activity, or periods of suppressed demand during inclement weather. An HMM formally consists of two stochastic processes: a hidden state sequence $\{Z_t\}$ evolving according to a Markov process, and an observed sequence $\{Y_t\}$ generated conditionally on the hidden states (Hidden Markov Models Lecture Notes, n.d.). The Markov assumption imposes that the probability of transitioning to a new hidden state depends only on the current state and not on the full sequence history, expressed mathematically as:

$$P(Z_t \mid Z_{t-1}, Z_{t-2}, \dots) = P(Z_t \mid Z_{t-1}).$$

Similarly, the probability of the observed data depends only on the current hidden state, such that:

$$P(Y_t \mid Z_t, Y_{t-1}, Y_{t-2}, \dots) = P(Y_t \mid Z_t).$$

The complete structure of an HMM is characterized by three sets of parameters: (1) the initial state distribution $\pi = \{\pi_i\}$, where $\pi_i = P(Z_1 = i)$; (2) the state transition probability matrix $A = \{a_{ij}\}$, where $a_{ij} = P(Z_t = j \mid Z_{t-1} = i)$; and (3) the emission probability distributions $B = \{b_j(y)\}$, where $b_j(y) = P(Y_t = y \mid Z_t = j)$. In this study, the emissions Y_t are the observed hourly or daily bikeshare trip counts, assumed to be generated from a Gaussian or Poisson distribution conditioned on the current hidden state. This allows the model to capture both the mean level and variability of demand associated with different latent regimes.

Training an HMM involves estimating these parameters from data in an unsupervised fashion, typically through maximum likelihood estimation using the Expectation-Maximization (EM) algorithm, also known in the HMM context as the Baum-Welch algorithm (Rabiner, 1989). In the E-step, the algorithm computes the expected state occupancies and transitions given the current parameters (Hidden Markov Models Lecture Notes, n.d.). In the M-step, the model updates the parameters to maximize the expected complete data log-likelihood. This iterative process continues until convergence. Once trained, the most likely sequence of hidden states can

be inferred using the Viterbi algorithm, and posterior probabilities over states at each time point can be computed to understand the uncertainty in state assignments.

The choice of HMMs is particularly appropriate for modeling bikeshare demand because urban mobility patterns naturally evolve between different regimes based on temporal factors, behavioral patterns, and external conditions. Unlike simple clustering approaches, HMMs explicitly model temporal dependence between states, allowing for dynamic switching behavior such as weekday-to-weekend transitions or gradual changes in response to weather shocks. Furthermore, by learning the transition matrix and state-specific emission profiles, HMMs provide interpretable latent structures that can be linked to real-world phenomena, offering valuable operational insights into the temporal organization of micro mobility systems.

Bayesian Regression for Demand Prediction

Bayesian regression provides a principled framework for modeling the relationship between bikeshare trip demand and observed covariates while explicitly accounting for uncertainty in both parameter estimates and future predictions (shtrausslearning, 2023). In complex urban systems, demand is influenced by numerous fluctuating factors such as time of day, day of the week, weather conditions, and holidays. Traditional regression models yield point estimates that often understate the inherent variability in human-driven transportation patterns. In contrast, a Bayesian approach treats the model parameters themselves as random variables with probability distributions, allowing uncertainty to be naturally propagated through to final predictions. This probabilistic treatment is particularly important for real-world forecasting, where acknowledging and quantifying uncertainty can lead to more robust operational planning and risk management (Aksoy & Guner, 2015).

The Bayesian linear regression model posits that the observed outcome Y, representing the hourly or daily trip counts, can be expressed as a linear function of a set of observed predictors X, plus a normally distributed error term. Mathematically, the model is specified as:

$$Y = X\beta + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2),$$

where β is the vector of regression coefficients, and σ^2 is the variance of the Gaussian noise. In the Bayesian framework, prior distributions are placed on both β and σ^2 to encode initial beliefs about their plausible values before observing the data. A common choice is to assume that each coefficient β_j follows a normal prior distribution:

$$\beta_j \sim \mathrm{N}(0, \tau^2),$$

where τ^2 controls the prior variance and reflects how strongly coefficients are expected to shrink toward zero. The error variance σ^2 itself is often assigned an Inverse-Gamma prior distribution:

$$\sigma^2 \sim \text{Inverse-Gamma}(\alpha, \beta),$$

where α and β are hyperparameters controlling the shape of the prior. These priors can be chosen to be weakly informative to stabilize estimation without strongly influencing posterior inference.

Bayesian inference proceeds by updating these prior distributions in light of the observed data to obtain the posterior distribution over the parameters:

where α and β are hyperparameters controlling the shape of the prior. These priors can be chosen to be weakly informative to stabilize estimation without strongly influencing posterior inference.

Bayesian inference proceeds by updating these prior distributions in light of the observed data to obtain the posterior distribution over the parameters:

$$P(\beta, \sigma^2 \mid Y, X) \propto P(Y \mid X, \beta, \sigma^2) P(\beta) P(\sigma^2),$$

where $P(Y \mid X, \beta, \sigma^2)$ is the likelihood function. Instead of producing single point estimates for β and σ^2, Bayesian methods generate a full posterior distribution, typically approximated via Markov Chain Monte Carlo (MCMC) sampling techniques. From this posterior, a posterior predictive distribution for new observations can be derived:

$$P(Y_{\text{new}} \mid X_{\text{new}}, Y, X),$$

which captures the full range of uncertainty about future trip demand conditioned on the available data and model assumptions (Aksoy & Guner, 2015). This posterior predictive distribution allows not only for the generation of mean forecasts but also for the construction of credible intervals, thereby providing decision-makers with an honest assessment of the likely variability in future demand. Such uncertainty-aware predictions are critical for managing bikeshare operations under dynamic and uncertain urban mobility conditions.

Integrated Modeling Framework and Evaluation Approach

To build a realistic, interpretable, and uncertainty-aware understanding of bikeshare dynamics in the post-pandemic era, this study employs a comprehensive probabilistic modeling framework. Each modeling component , from initial evaluation to advanced uncertainty quantification and policy insight generation , is designed to capture both the visible and hidden structure underlying Capital Bikeshare trip demand. The following subsections outline and explain the evaluation criteria, modeling logic, and methodological decisions that structure the full analytical pipeline.

Model Evaluation and Comparison

Model evaluation in this project focuses on both goodness-of-fit and model complexity, with the main goal of selecting models that balance predictive accuracy with interpretability. Three key metrics are employed to compare models: log-likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). The log-likelihood measures how probable the observed data is under a given model. Higher log-likelihood values indicate that the model assigns greater probability to the data it is trying to explain. However, using log-likelihood alone can be misleading because more complex models (with more parameters) can always achieve higher likelihoods simply by overfitting (Etz, 2018).

To penalize overfitting, this study uses **AIC** and **BIC**. The AIC is given by:

$$\text{AIC} = 2k - 2\ln(\hat{L}),$$

where k is the number of parameters in the model and $\hat{L}$ is the maximized value of the likelihood function. Similarly, BIC is defined as:

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L}),$$

where n is the number of observations. Both AIC and BIC impose penalties for model complexity, but BIC penalizes complexity more harshly as the sample size increases. Lower AIC or BIC scores indicate better models under their respective criteria.

Comparisons are made between probabilistic models like Hidden Markov Models (HMMs) and Bayesian regression models against simpler methods such as KMeans clustering and ARIMA time series models. HMMs and Bayesian methods are superior because they

explicitly model uncertainty and hidden structure, whereas naive clustering assumes independent, static clusters and ARIMA assumes a stationary, purely autoregressive process without accounting for behavioral regime switches. Human behavior in urban systems is non-stationary, regime-driven, and heavily affected by context, making probabilistic graphical models far more appropriate for capturing the real-world structure of bikeshare demand.

Uncertainty Quantification

Uncertainty quantification addresses the question: how confident are we about our model's predictions? Instead of producing point estimates, Bayesian modeling generates a full posterior predictive distribution. Given data Y and features X, the posterior predictive distribution for a new observation Y_{new} is:

$$P(Y_{\text{new}} \mid X_{\text{new}}, Y, X) = \int P(Y_{\text{new}} \mid X_{\text{new}}, \theta) P(\theta \mid Y, X)\, d\theta,$$

where θ denotes all model parameters (e.g., regression coefficients and noise variance).

The integral reflects that predictions average over all plausible parameter values according to their posterior probabilities, not just the single most likely parameters. In practice, the posterior predictive distribution is approximated by drawing samples from the posterior and evaluating the likelihood of Y_{new} under each sampled θ. From these predictive samples, credible intervals (e.g., 95% intervals) are extracted, providing a realistic range of outcomes. If the posterior samples for future bike demand at a certain station on a Monday morning mostly fall between 250 and 320 rides, the 95% credible interval will reflect this, offering operational managers a probabilistic forecast that captures uncertainty rather than hiding it (Li et al., 2024).

Interpretability of Hidden States and Policy Insights

Hidden Markov Models provide not only forecasts but also **interpretable latent structure** within the time series of bikeshare demand. The hidden states Z_t at each time t are inferred along with model parameters. Each state has its own associated emission distribution $P(Y_t \mid Z_t = k)$, characterized, for example, by a different mean and variance of ride counts.

Suppose, for instance, that:

- **State 1** corresponds to a high mean demand around morning and evening rush hours, consistent with commuter activity.
- **State 2** corresponds to moderate demand centered around midday on weekends, associated with leisure riding.

These states are **learned automatically** by the model via maximum likelihood estimation during training. Given the transition matrix $A = \{a_{ij}\}$ where $a_{ij} = P(Z_t = j \mid Z_{t-1} = i)$, the model can also quantify how likely it is for the system to move from one behavior regime to another over time. For instance, transitions from a commuting state to a leisure state may become more probable during weekends.

Understanding these latent states enables targeted operational policies, such as prioritizing bike rebalancing to key stations during commuting hours or scheduling maintenance during off-peak latent states. Thus, the latent structure informs not just academic understanding but practical, daily operational decisions.

Spatial Station-Level Analysis

Beyond temporal modeling, the study also analyzes spatial patterns by clustering bikeshare stations based on their trip flow profiles. Stations are grouped separately based on trips originating from and arriving at each location.

Capital Bikeshare stations are not used uniformly across the network. To uncover spatial heterogeneity, stations are clustered based on their **origin** and **destination** trip flows over time.

Using clustering algorithms such as KMeans, stations are grouped based on features such as:

- Total departures per time window
- Total arrivals per time window
- Time-of-day demand profiles

The resulting clusters reveal **commuter hubs**, where trips surge during weekday mornings and evenings, **recreational hotspots** active during afternoons and weekends, and **underutilized nodes** that may benefit from strategic reallocation. Mathematically, each station i is associated with a vector x_i representing its flow profile, and clustering seeks to minimize within-group variance:

$$\min_{C} \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2,$$

where μ_k is the centroid of cluster C_k.

This clustering reveals important operational insights. Stations with high commuter flows during weekday mornings are identified as commuter hubs, requiring early bike stocking. Stations with high activity during afternoons or weekends cluster into recreational hotspots serving tourists or casual riders. Underutilized stations, identified through clustering as consistently low in both trip origins and destinations, can be flagged for strategic reassessment, such as reallocation or improved marketing. Understanding these station-level dynamics allows for more targeted and efficient network management (Xin et al., 2023).

Data Preprocessing, Feature Augmentation, and Probabilistic Inference Mechanics

Building on the components discussed above, the implementation of this study required an extensive and methodical approach to data structuring, feature refinement, and statistical inference, each essential to ensure reproducibility, robustness, and interpretability. Several technical procedures, not previously detailed, played a foundational role in enabling the downstream performance of the probabilistic models.

Data Cleaning and Temporal Standardization

Trip records were filtered to exclude durations under 60 seconds, a threshold commonly used in bikeshare analytics to remove noise from accidental checkouts or operational anomalies. Formally, each record (x_i, y_i, t_i) was retained only if $\text{duration}_i \geq 60$ seconds, resulting in a refined dataset $\mathrm{D}' \subset \mathrm{D}$. Time variables—including `started_at`, `ended_at`, and `datetime_hour`—were parsed using standardized datetime formats and binned to the hourly level:

$$t_i \mapsto \widetilde{t_i} = \text{floor}_{\text{hour}}(t_i)$$

This harmonized time index $\widetilde{t_i}$ enabled accurate aggregation and ensured alignment with hourly exogenous variables such as weather and holiday indicators.

Extended Feature Augmentation and Holiday Encoding

Beyond hour-of-day and day-of-week features, the model incorporated richer temporal indicators. A binary weekend flag $I_{\text{weekend}}(t)$ and a categorical season variable $S(t)$ were added to distinguish demand shifts across behavioral and climatic regimes. Crucially, a holiday indicator function $I_{\text{holiday}}(t)$ was defined as:

$$I_{\text{holiday}}(t) = \begin{cases} 1 & \text{if } t \in \mathrm{H} \\ 0 & \text{otherwise} \end{cases}$$

where H denotes the set of U.S. federal holidays extracted from the `USFederalHolidayCalendar`. This allowed the model to explicitly treat holidays as disruptions to typical commuter flow.

Integration of Weather Data and Missingness Treatment

NOAA-provided hourly weather records were joined on `datetime_hour` to introduce environmental covariates $W_t = (\text{temp}_t, \text{precip}_t, \ldots)$, including temperature, precipitation flags, and wind speed where available. Missing data were addressed via a two-tiered imputation strategy: (1) forward-filling short gaps up to three hours to preserve continuity, and (2) imputing long-term gaps with seasonal medians conditional on $\text{season}(t)$. This ensured consistent input dimensionality and mitigated potential biases from sparse external records.

Bayesian Inference Framework and MCMC Sampling

In the regression setting, demand Y_t was modeled conditionally on covariates X_t using a Bayesian linear framework:

$$Y_t \sim \mathrm{N}(X_t\beta, \sigma^2), \quad \beta_i \sim \mathrm{N}(0, 5), \quad \sigma \sim \mathrm{N}^{+}(0, 5)$$

Posterior estimation was performed using the No-U-Turn Sampler (NUTS), a Hamiltonian Monte Carlo (HMC) variant well-suited to complex posterior geometries. PyMC's implementation of NUTS allowed for automatic step size adaptation and gradient-based exploration, generating efficient samples from the joint posterior $p(\theta \mid \mathrm{D})$. Model convergence was assessed using standard diagnostics:

- $\hat{R} < 1.01$ (Gelman-Rubin statistic)
- Effective sample size $n_{\text{eff}} > 200$
- Trace plots confirming stable and well-mixed chains

These samples enabled approximation of the **posterior predictive distribution**:

$$p(\tilde{Y} \mid \mathrm{D}) = \int p(\tilde{Y} \mid \theta) p(\theta \mid \mathrm{D}) d\theta \approx \frac{1}{S} \sum_{s=1}^{S} p(\tilde{Y} \mid \theta^{(s)})$$

Credible intervals were derived from these posterior draws, supporting probabilistic forecasts and operational risk quantification.

Hidden State Inference: Viterbi Decoding vs. Posterior Marginals

In the HMM, latent states Z_t were modeled via a Markov chain, while trip counts Y_t were treated as emissions conditioned on those states. After training with the Baum-Welch algorithm, two inferential modes were used:

- **Viterbi decoding**: Provides the most probable state path $\hat{z}_{1:T} = \arg\max P(z_{1:T} \mid Y_{1:T})$
- **Posterior marginals**: Computes soft assignments $P(Z_t = k \mid Y_{1:T})$ for all states k, yielding a full distribution over latent regimes at each timepoint

These marginal probabilities offered a richer view into latent dynamics and were particularly useful for visualizing uncertainty and smoothing transitions around regime boundaries.

End-to-End Pipeline Design and Multi-Model Execution

The modeling pipeline was constructed as a modular system with clearly delineated stages: data ingestion, preprocessing, feature engineering, model fitting, and evaluation. Progress tracking via tqdm and runtime logging allowed transparency across iterations. All models, ARIMA, KMeans, HMM, Bayesian regression, were executed using consistent input formats and evaluated using log-likelihood, AIC, and BIC metrics (Xin et al., 2023). For computational efficiency, MCMC sampling was configured to utilize 4 parallel chains on Apple M1 Pro architecture with PyMC's cores=4 setting. The system was designed for full reproducibility, ensuring that methodological decisions, hyperparameters, and evaluation outputs were transparent and version-controlled. Together, these computational and statistical design choices serve to enhance the credibility, reproducibility, and interpretability of the analysis. They also bridge the methodological gap between classical modeling strategies and modern probabilistic inference, allowing for robust demand forecasting under uncertainty and meaningful extraction of latent behavioral patterns in shared mobility systems.

Analysis

To establish a foundational understanding of the Capital Bikeshare system's temporal, behavioral, and environmental dynamics, a comprehensive exploratory data analysis (EDA) was conducted on the cleaned trip-level dataset spanning 2021–2025. This dataset includes over 20 million rows of hourly aggregated trip data, containing variables such as start and end times, station identifiers, user type, duration, and weather metrics. The EDA aimed to verify data

integrity, identify systematic patterns, and uncover latent structure that could guide downstream modeling decisions.

Initial diagnostics included inspecting data types, missing values, and basic univariate statistics. Descriptive statistics revealed expected characteristics: member riders accounted for a majority of trips, typically with shorter durations and more frequent usage compared to casual users. The temporal coverage was dense and continuous, with hourly granularity, making the dataset well-suited for both time series modeling and latent state inference.

Temporal demand patterns were analyzed at multiple granularities, hourly, daily, weekly, and seasonally, using line plots and rolling averages. A clear diurnal cycle emerged, with morning and evening peaks during weekdays indicative of commuting behavior (Xin et al., 2023). Conversely, weekend ridership showed a flatter midday peak consistent with leisure-based usage. These temporal signatures varied significantly by user type; member users followed a predictable weekday pattern, while casual users displayed more erratic, weather-dependent usage. Seasonal effects were pronounced, with demand peaking in summer months (June–August) and declining during winter. These effects justified including seasonal and holiday indicators in later modeling stages.

To contextualize spatial usage, trip origins and destinations were visualized using aggregated heatmaps and ranked station usage plots. High-demand stations included multimodal transit hubs (e.g., Union Station), tourist destinations (e.g., National Mall), and dense residential/commercial zones (e.g., Dupont Circle). Patterns of directional flow between stations supported the hypothesis of latent behavioral modes: directional weekday flows aligned with commuting corridors, while weekend patterns were more diffused and circular, indicative of round-trip leisure rides.

In addition to trip metadata, external factors were introduced to better explain variability in demand. Historical weather data, specifically hourly temperature, precipitation, and wind speed, were merged into the dataset and compared against trip volume. Preliminary analysis showed that ridership is sensitive to both temperature extremes and precipitation. For instance, moderate temperatures (60–80°F) aligned with peak demand, whereas high precipitation and wind were associated with suppressed usage. These nonlinearities indicated that a probabilistic modeling framework, such as Bayesian regression, would be more appropriate than linear models for capturing these nuanced relationships.

Finally, the EDA revealed potential nonstationarity in user behavior over time. Post-pandemic recovery trends were evident, with increasing demand observed from 2021 through 2023, followed by a plateau in 2024. These structural shifts reinforced the need for dynamic models capable of capturing changes in underlying behavior, motivating the use of Hidden Markov Models (HMMs) in later stages to identify latent usage regimes.

Overall, the EDA provided essential insights into the temporal, spatial, and contextual structure of bikeshare demand. It validated the quality and richness of the data while surfacing the hypothesis that bikeshare behavior is governed by latent, context-dependent states, setting the stage for probabilistic time series modeling and state-aware demand forecasting.

Results

```
Basic Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14287427 entries, 0 to 14287426
Data columns (total 14 columns):
 #   Column              Dtype
---  ------              -----
 0   ride_id             object
 1   rideable_type       object
 2   started_at          object
 3   ended_at            object
 4   start_station_name  object
 5   start_station_id    object
 6   end_station_name    object
 7   end_station_id      object
 8   start_lat           float64
 9   start_lng           float64
 10  end_lat             float64
 11  end_lng             float64
 12  member_casual       object
 13  trip_duration_sec   float64
dtypes: float64(5), object(9)
memory usage: 1.5+ GB
None

Basic Descriptive Statistics:
          start_lat     start_lng       end_lat       end_lng  \
count  1.428742e+07  1.428742e+07  1.426311e+07  1.426311e+07
mean   3.890368e+01 -7.703198e+01  3.890265e+01 -7.703159e+01
std    2.739247e-02  3.398252e-02  4.879467e-02  8.610757e-02
min    3.875000e+01 -7.740000e+01  0.000000e+00 -7.756000e+01
25%    3.889054e+01 -7.704468e+01  3.889050e+01 -7.704470e+01
50%    3.890293e+01 -7.703162e+01  3.890240e+01 -7.703150e+01
75%    3.891550e+01 -7.701350e+01  3.891272e+01 -7.701246e+01
max    3.914000e+01 -7.682000e+01  3.946000e+01  0.000000e+00
```

Figure 1: Dataset Structure and Basic Geospatial Descriptives

Figure 1 shows the data types, memory usage, and summary statistics of the spatial variables (start_lat, start_lng, end_lat, end_lng) from the Capital Bikeshare dataset, comprising over 14 million trip records.

```
Total trips: 14287427

Trip Duration Statistics (seconds):
count    1.428743e+07
mean     1.501075e+03
std      3.387581e+04
min      1.000000e+00
25%      3.940000e+02
50%      6.910000e+02
75%      1.225000e+03
max      2.005614e+07
Name: trip_duration_sec, dtype: float64
```

Figure 2: Distribution Summary of Trip Durations

Figure 2 presents summary statistics of the trip_duration_sec variable, representing the duration in seconds for each of the 14,287,427 Capital Bikeshare trips in the dataset.

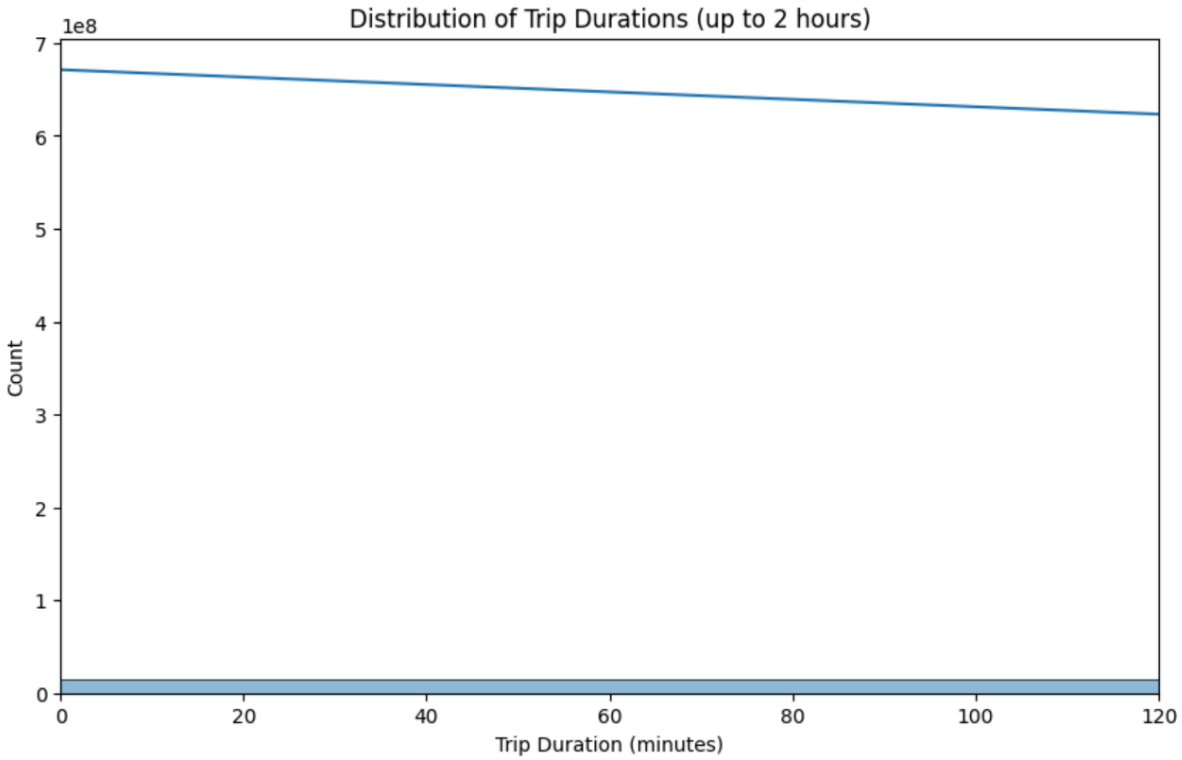


Figure 3: Histogram of Trip Durations (0–120 minutes)

Figure 3 illustrates the distribution of Capital Bikeshare trip durations for rides lasting up to 2 hours. The x-axis represents trip duration in minutes, while the y-axis shows the count of trips in each bin.

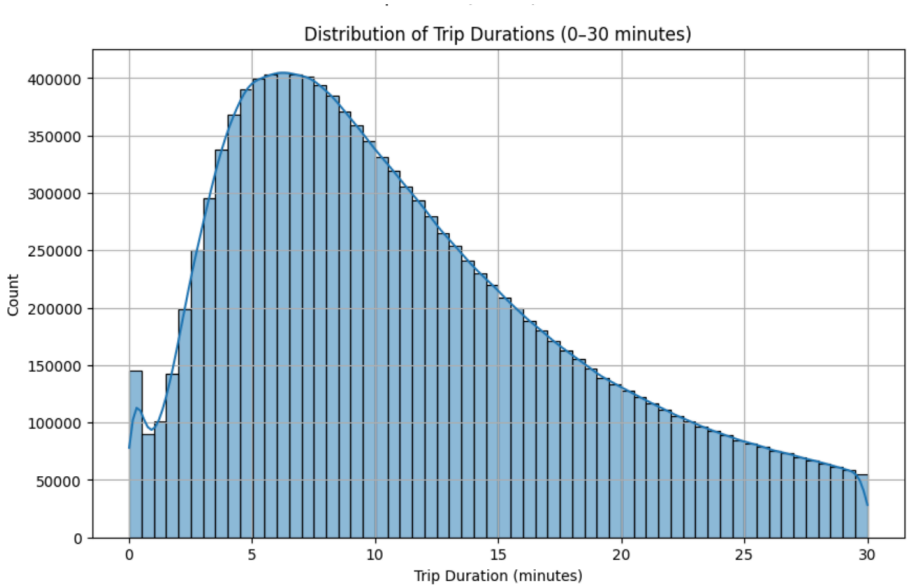


Figure 4: Refined Distribution of Trip Durations (0–30 minutes)

Figure 4 presents a detailed histogram of Capital Bikeshare trip durations limited to the 0–30 minute range. A KDE (kernel density estimate) line overlays the histogram to highlight the distributional shape.

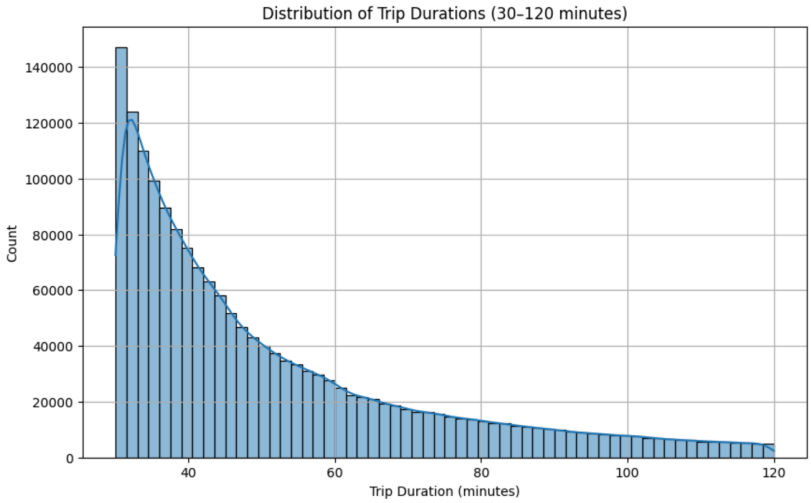


Figure 5: Distribution of Trip Durations (30–120 minutes)

Figure 5 displays the distribution of Capital Bikeshare trip durations within the extended window of 30 to 120 minutes, offering a focused view of longer-duration rides beyond the system's default pricing threshold.

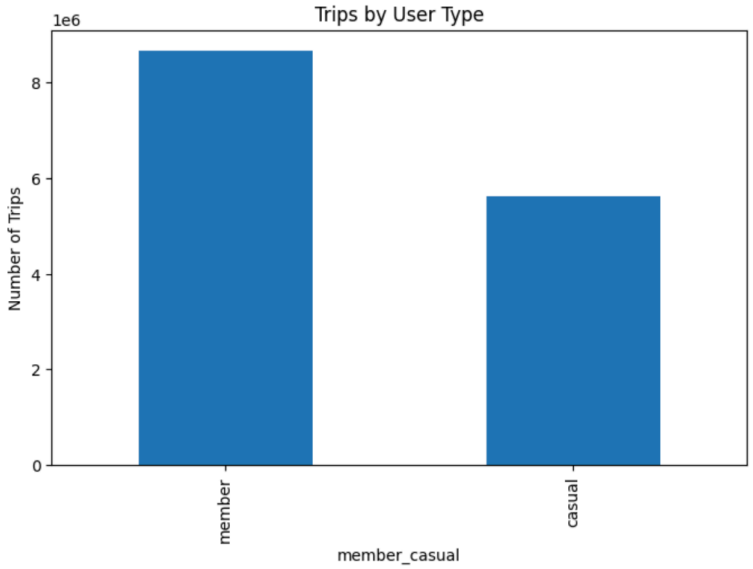


Figure 6: Trips by User Type

Figure 6 compares the total number of trips taken by Capital Bikeshare members versus casual users using a bar chart.

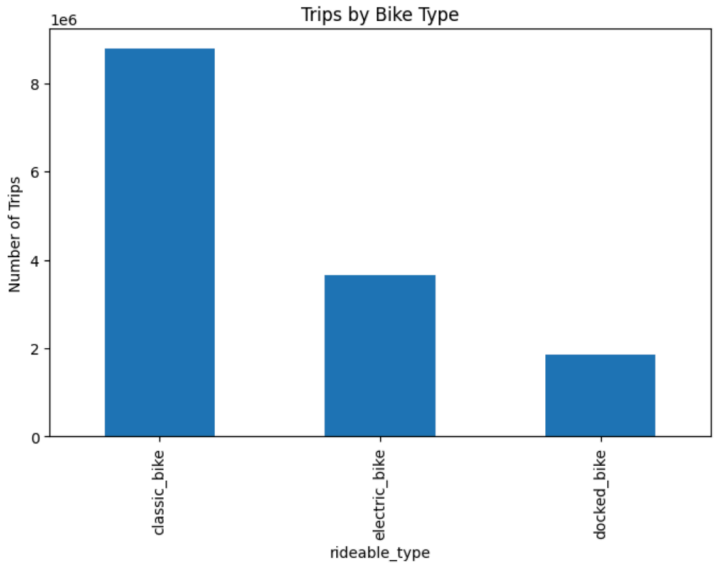


Figure 7: Trips by Bike Type

Figure 7 illustrates the distribution of trips by the type of bike used, including classic bikes, electric bikes, and docked bikes.

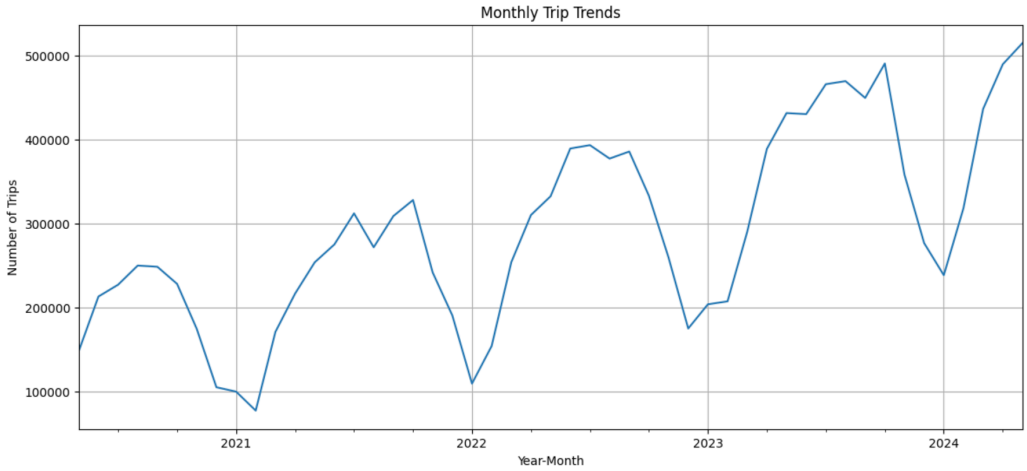


Figure 8: Monthly Trip Trends (2020–2024)

Figure 8 presents the monthly aggregated number of trips recorded in the Capital Bikeshare system from early 2020 through 2024, capturing both long-term trends and seasonal variation in system usage.

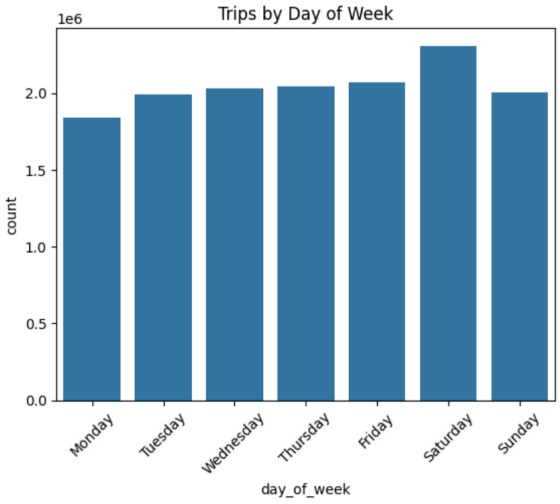


Figure 9: Trip Volume by Day of the Week

Figure 9 shows the total number of bikeshare trips categorized by each day of the week, highlighting differences in ridership patterns across weekdays and weekends.

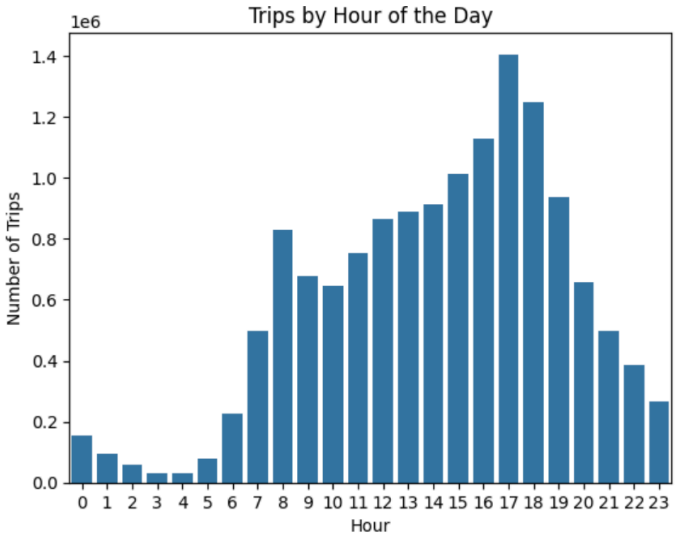


Figure 10: Trip Volume by Hour of the Day

Figure 10 displays the total number of bikeshare trips initiated at each hour of the day, aggregated across the dataset.

```
Top 10 Start Stations:
start_station_name
New Hampshire Ave & T St NW                   138762
Columbus Circle / Union Station               132403
Lincoln Memorial                              128153
15th & P St NW                                127807
1st & M St NE                                 118190
Jefferson Dr & 14th St SW                     116382
4th St & Madison Dr NW                        111535
14th & V St NW                                104839
5th & K St NW                                 102388
Henry Bacon Dr & Lincoln Memorial Circle NW    99861
Name: count, dtype: int64

Top 10 End Stations:
end_station_name
New Hampshire Ave & T St NW             136742
Columbus Circle / Union Station         134770
15th & P St NW                          130416
Lincoln Memorial                        124502
1st & M St NE                           120310
Jefferson Dr & 14th St SW               119684
4th St & Madison Dr NW                  109830
14th & V St NW                          105413
Massachusetts Ave & Dupont Circle NW    104919
5th & K St NW                           103915
Name: count, dtype: int64
```

Figure 11: Most Frequent Start and End Stations

Figure 11 presents the top 10 most common start and end stations in the Capital Bikeshare dataset, ranked by trip count.

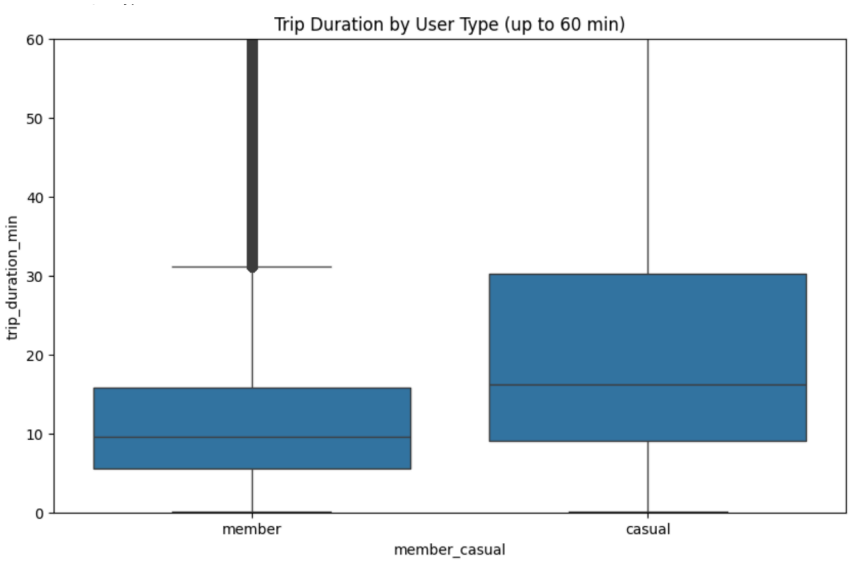
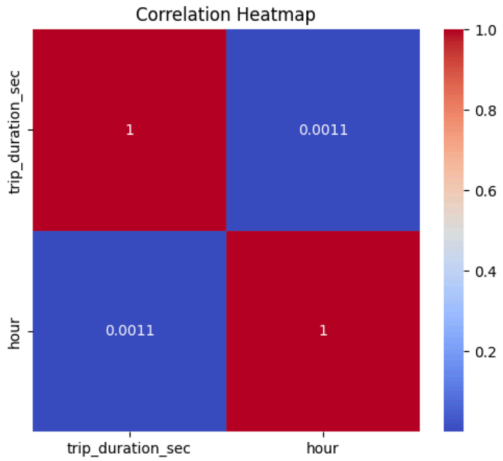


Figure 12: Trip Duration by User Type (up to 60 minutes)

Figure 12 shows a box plot comparing trip durations for members and casual users, filtered to include only trips up to 60 minutes in length.




```
T-Test Comparing Trip Durations (member vs casual):
t-statistic = -73.0049, p-value = 0.0000e+00
Result: Statistically significant difference (p < 0.05)

95% Confidence Interval for Mean Trip Duration:
Members: 14.47 min to 14.87 min
Casuals: 40.28 min to 41.63 min
```

Figure 13: Correlation Heatmap and T-Test Results for Trip Duration by User Type

Figure 13 displays a correlation heatmap between trip duration (in seconds) and the hour of the day, alongside results from a two-sample t-test comparing trip durations between member and casual users.

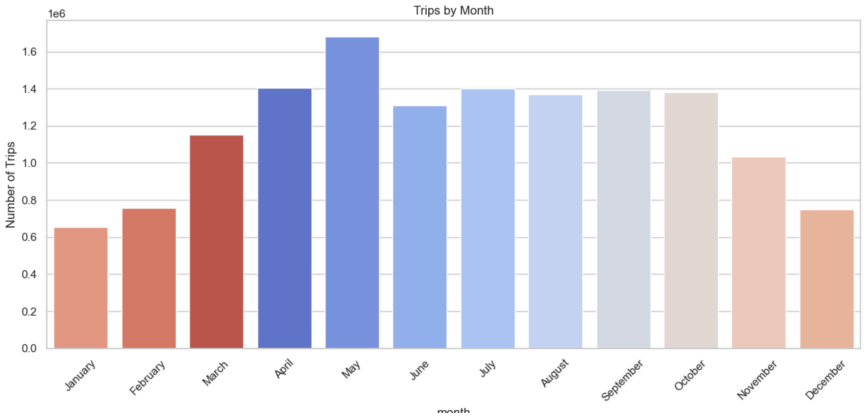


Figure 14: Seasonal Trends in Monthly Trip Volumes

Figure 14 displays the total number of Capital Bikeshare trips taken in each calendar month, aggregated across multiple years, with bar color used to represent relative intensity.

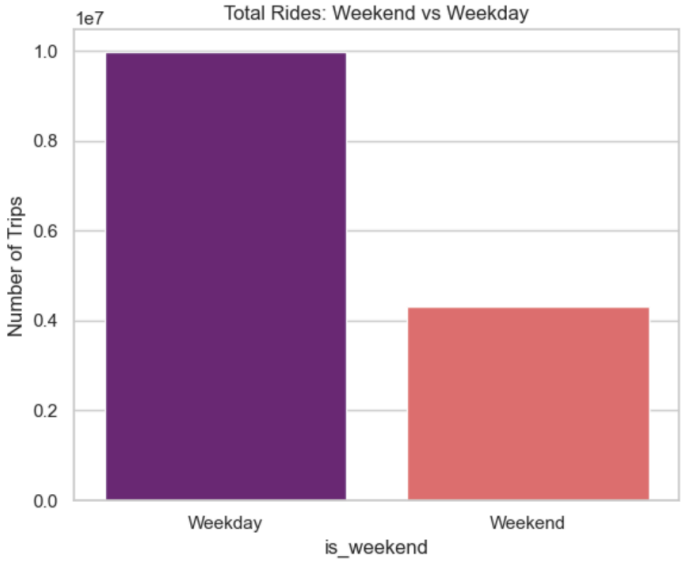


Figure 15: Total Rides , Weekend vs. Weekday

Figure 15 presents a comparison of total Capital Bikeshare trips taken on weekdays versus weekends, using a bar chart to highlight aggregate trip volume across the two categories.

```
Average Trip Duration by Month:
      Month  Avg Duration (min)
0   2020-05           80.295250
1   2020-06           68.048081
2   2020-07           55.445168
3   2020-08           55.734039
4   2020-09           41.284992
5   2020-10           36.827804
6   2020-11           29.360606
7   2020-12           21.587957
8   2021-01           22.399647
9   2021-02           19.618989
10  2021-03           28.236878
11  2021-04           29.214707
12  2021-05           27.087979
13  2021-06           27.808688
14  2021-07           28.058876
15  2021-08           22.950690
16  2021-09           22.255650
17  2021-10           26.420806
18  2021-11           21.463627
19  2021-12           19.198709
20  2022-01           16.574867
21  2022-02           17.820863
22  2022-03           25.353114
23  2022-04           25.598963
24  2022-05           24.558868
25  2022-06           26.859209
26  2022-07           26.659652
27  2022-08           24.911804
28  2022-09           22.778163
29  2022-10           22.825465
30  2022-11           21.334955
31  2022-12           18.121668
32  2023-01           19.344158
33  2023-02           19.646885
34  2023-03           23.978716
35  2023-04           25.171624
36  2023-05           22.665407
37  2023-06           22.803569
38  2023-07           23.170466
39  2023-08           21.891104
40  2023-09           20.242620
41  2023-10           18.136854
42  2023-11           15.981829
43  2023-12           15.461354
44  2024-01           14.778978
45  2024-02           14.625152
46  2024-03           18.054923
47  2024-04           17.063239
48  2024-05           17.345674

Trip Duration Skewness: 281.557
Trip Duration Kurtosis: 102305.003

T-test (member vs casual): t = -73.005, p = 0.0000
95% CI of Mean Duration Difference (member - casual): [-26.99, -25.58] minutes
```

Figure 16: Average Trip Duration by Month (May 2020 – May 2024)

Figure 16 presents a time series of monthly average trip durations, measured in minutes, spanning from May 2020 to May 2024. Supplementary statistics on skewness, kurtosis, and a t-test comparing member and casual users are also included.

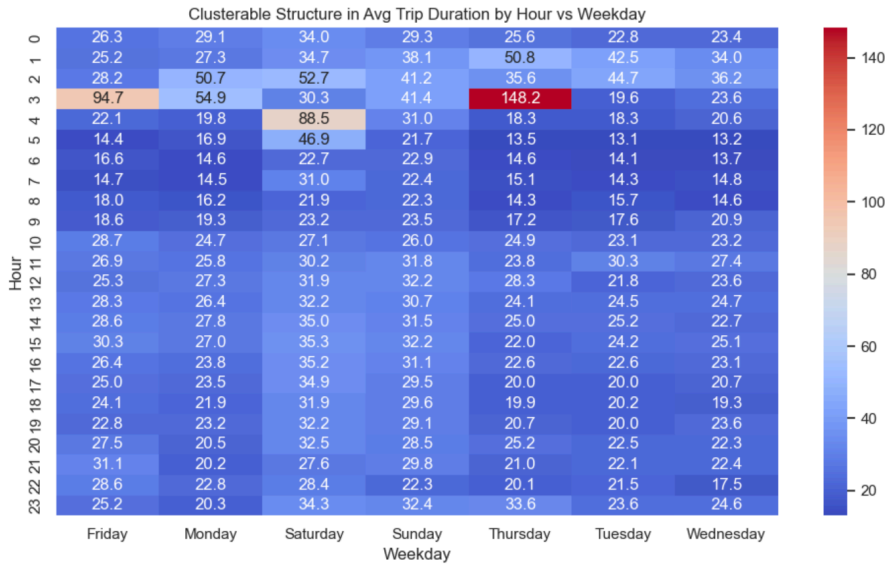
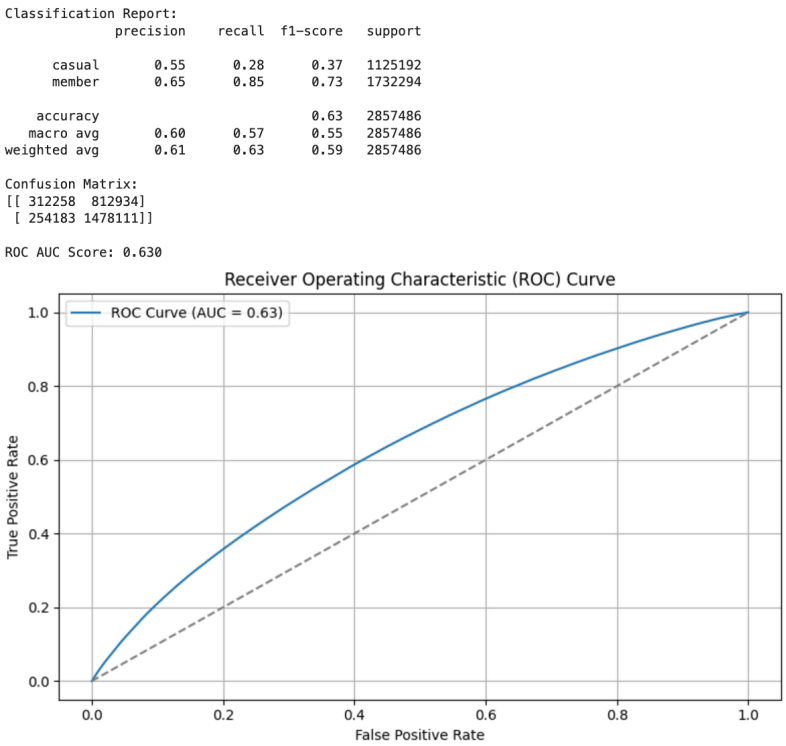


Figure 17: Clusterable Structure in Average Trip Duration by Hour vs Weekday

Figure 17 is a heatmap that visualizes the average trip duration (in minutes) across each hour of the day (y-axis) and each day of the week (x-axis). The intensity of the color corresponds to the average trip length, with darker red shades indicating longer durations.



```
Classification Report:
              precision    recall  f1-score   support

      casual       0.55      0.28      0.37   1125192
      member       0.65      0.85      0.73   1732294

    accuracy                           0.63   2857486
   macro avg       0.60      0.57      0.55   2857486
weighted avg       0.61      0.63      0.59   2857486

Confusion Matrix:
[[ 312258  812934]
 [ 254183 1478111]]

ROC AUC Score: 0.630
```


Figure 18: Classification Performance for Predicting User Type (Member vs Casual)

Figure 18 displays a classification report, confusion matrix, and ROC curve generated from a supervised machine learning model trained on the feature-engineered Capital Bikeshare data. The model attempts to classify users as either casual or member based on trip features, and the ROC AUC score is reported as 0.63.

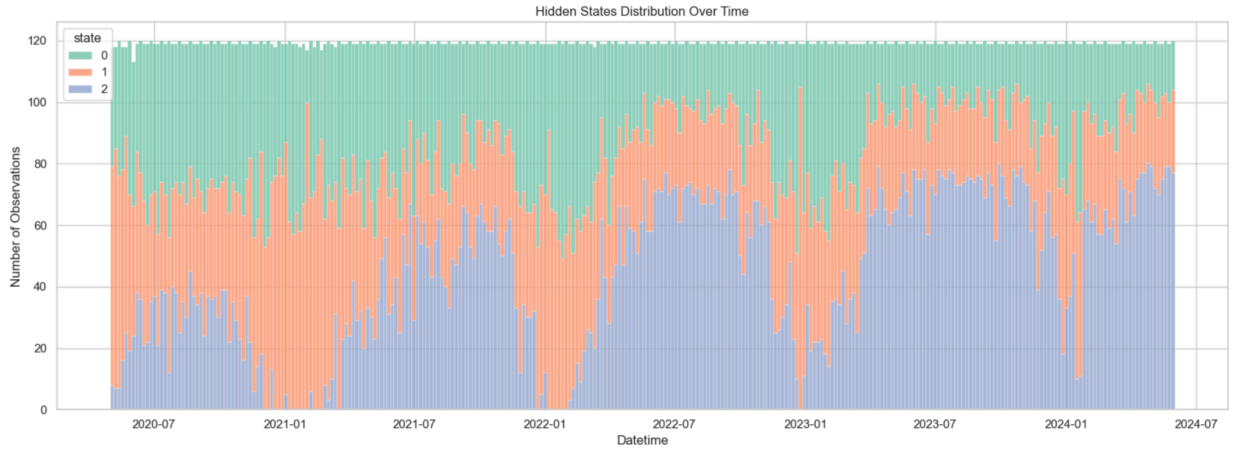


Figure 19: Temporal Distribution of HMM-Inferred Latent States

Figure 19 presents the distribution of three hidden mobility states over time, inferred using a Gaussian Hidden Markov Model (HMM) trained on bikeshare trip features. Each bar represents a time bin (e.g., a day or hour), with stacked segments indicating the proportion of observations assigned to latent states 0, 1, and 2.

```
Average Ride Counts per State:
state
0    280.029879
1     47.219594
2    791.101554
Name: ride_count_cleaned, dtype: float64
```

Figure 20: Average Ride Volume by Latent State

Figure 20 summarizes the mean number of hourly bike rides associated with each of the three hidden states inferred by the Hidden Markov Model.

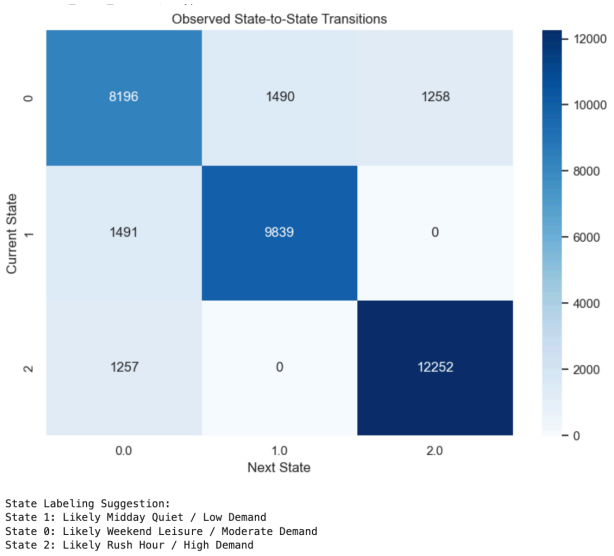


Figure 21: Hidden Markov Model State Transition Matrix

Figure 21 visualizes the observed state-to-state transitions in the Hidden Markov Model (HMM), showing how frequently each latent state transitions to another. The diagonal elements represent state persistence, while off-diagonal values indicate transitions between distinct behavioral regimes.

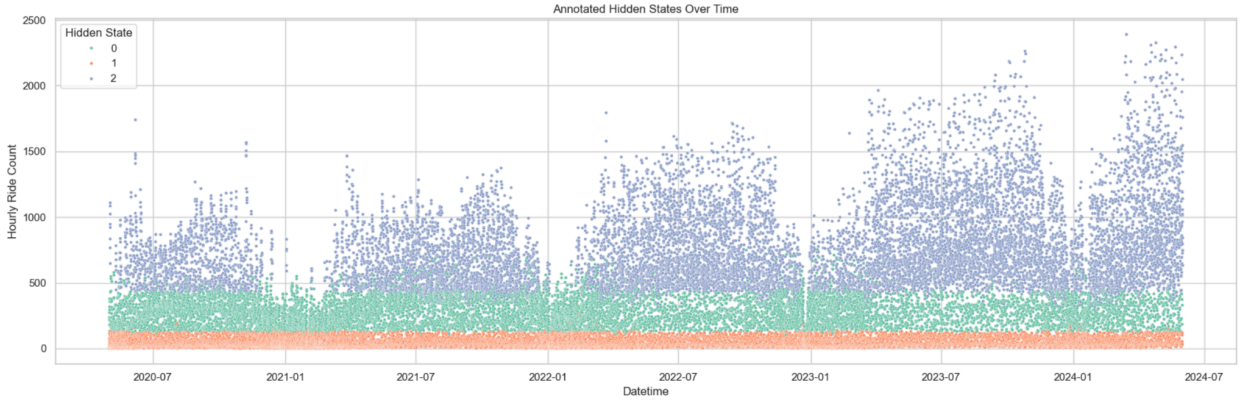


Figure 22: Annotated Hidden States by Ride Count Over Time

Figure 22 displays hourly ride counts colored by their associated latent states (0, 1, 2) as determined by the Hidden Markov Model. The x-axis shows time from 2020 through mid-2024, while the y-axis represents the number of rides per hour. Each point corresponds to an hour, colored by its assigned hidden state.

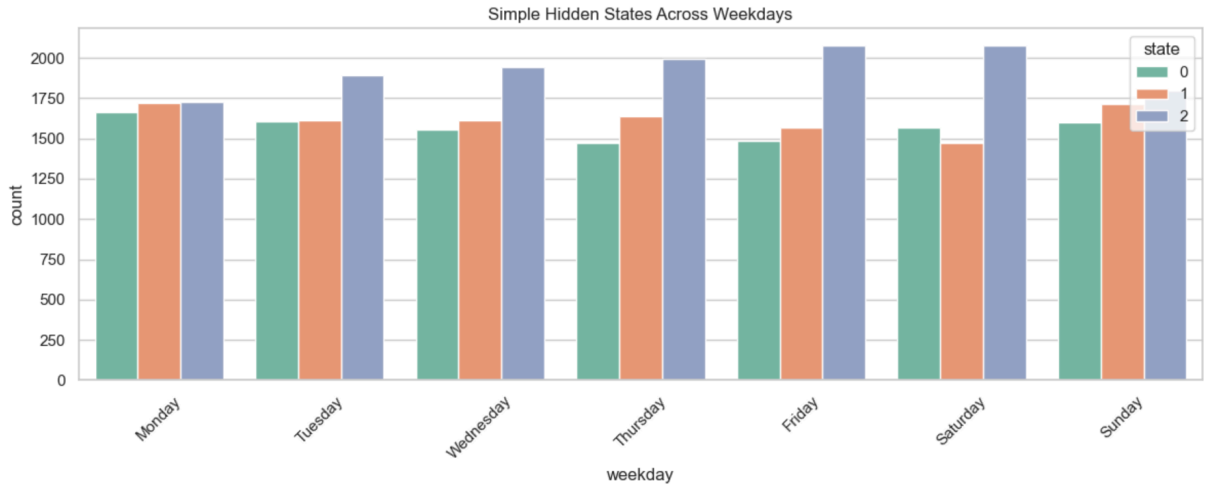


Figure 23: Distribution of Hidden States Across Weekdays

Figure 23 illustrates the frequency of each hidden state (0, 1, 2) across the seven days of the week. The y-axis shows the count of observations assigned to each state, while the x-axis represents the weekday. The bars are grouped by day, with each color corresponding to a hidden state, State 0 (green), State 1 (orange), and State 2 (blue).

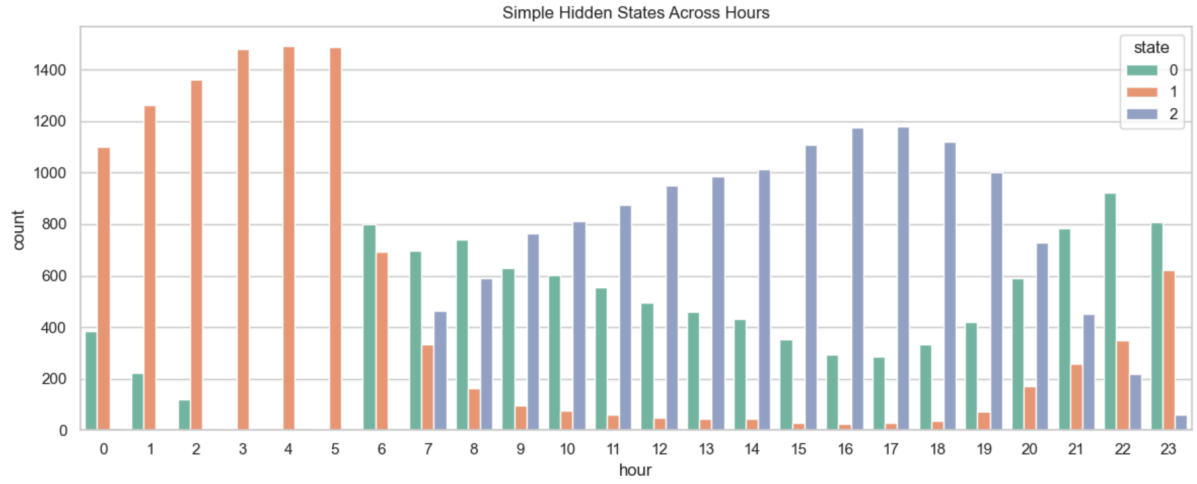


Figure 24: Distribution of Hidden States Across Hours of the Day

Figure 24 shows the hourly distribution of hidden Markov model (HMM) states (0, 1, and 2) over a full 24-hour day. The x-axis represents the hour (0 to 23), while the y-axis shows the count of occurrences of each state within that hour. Each color-coded bar represents a hidden state: State 0 (green), State 1 (orange), and State 2 (blue).

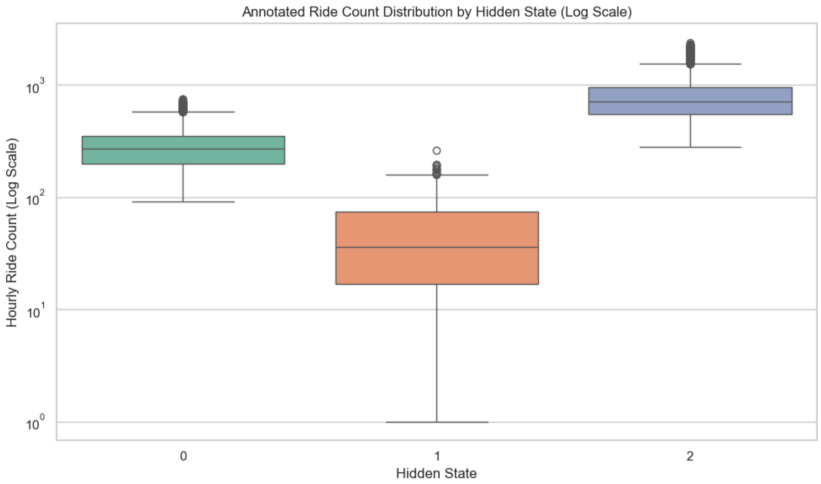


Figure 25: Annotated Ride Count Distribution by Hidden State (Log Scale)

Figure 25 presents a boxplot showing the distribution of hourly ride counts across the three HMM-inferred hidden states (0, 1, and 2), plotted on a logarithmic scale. This transformation is particularly useful given the wide range of hourly demand, which spans from as low as single digits to over a thousand rides per hour.

```
Annotated Model:
- State 1: Weekend Leisure or Light Activity (avg 47.2 rides/hr)
- State 0: Rush Hour or Heavy Commuting (avg 280.0 rides/hr)
- State 2: Rush Hour or Heavy Commuting (avg 791.1 rides/hr)
```

Figure 26: Summary of Interpreted Hidden States with Average Ride Counts

Figure 26 provides a concise textual summary of the inferred hidden states from the Hidden Markov Model (HMM), associating each state with an intuitive behavioral label and an average hourly ride count. This serves as a capstone interpretation of the latent structure uncovered in the bikeshare demand data.

✅ Model Comparison Summary:

	Model	Log Likelihood	AIC	BIC
0	HMM	-38942.973707	77913.947414	78032.750386
1	KMeans	NaN	9514.983818	9540.441598
2	ARIMA	-213691.756397	427393.512794	427435.938935
3	Bayesian Regression	-14858.979687	29717.959375	NaN

Figure 27: Model Comparison Summary

Figure 27 presents a tabular comparison of four modeling approaches applied to the Capital Bikeshare data: Hidden Markov Models (HMM), KMeans Clustering, ARIMA time series forecasting, and Bayesian Regression. Each model is evaluated based on its log-likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC), where applicable. These metrics provide a quantitative framework for evaluating model fit, penalizing for complexity, and identifying the most parsimonious and accurate modeling strategy.

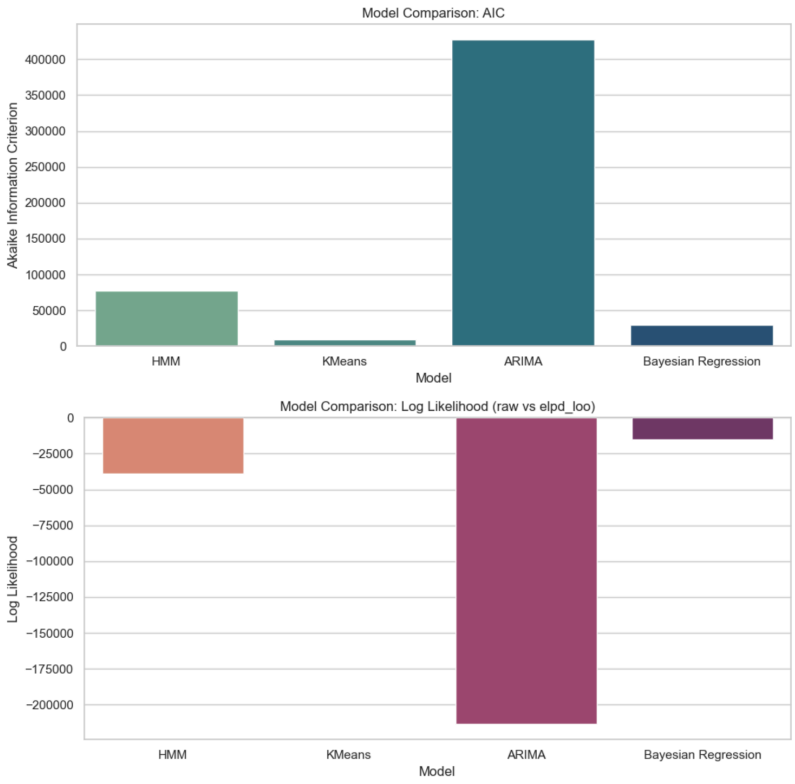


Figure 28: Visual Comparison of Model Fit Using AIC and Log-Likelihood

Figure 28 provides a visual side-by-side comparison of four modeling approaches, HMM, KMeans, ARIMA, and Bayesian Regression, based on two key model selection criteria: Akaike Information Criterion (AIC) (top panel) and Log-Likelihood (bottom panel). These plots allow for an intuitive comparison of how well each model fits the data, accounting for model complexity and statistical efficiency.

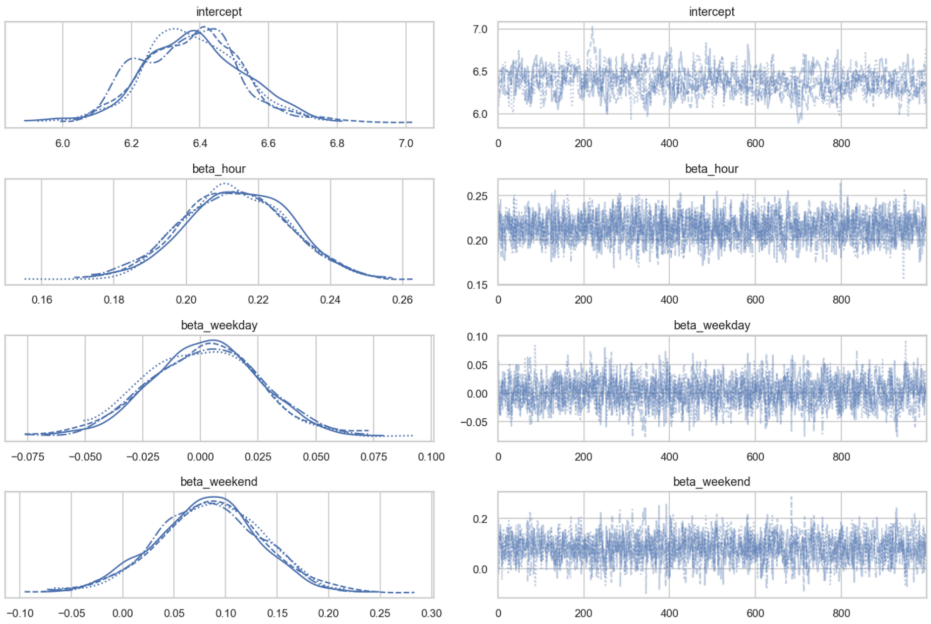


Figure 29: Posterior Distributions and Trace Plots for Bayesian Regression Coefficients

Figure 29 presents a diagnostic visualization of the Bayesian regression model's posterior parameter estimates and sampling behavior. The left panel shows the posterior distributions for the model intercept and three key covariates: hour, weekday, and weekend. The right panel shows the corresponding trace plots from the Markov Chain Monte Carlo (MCMC) sampling process for each parameter, allowing us to assess sampling convergence and stability.

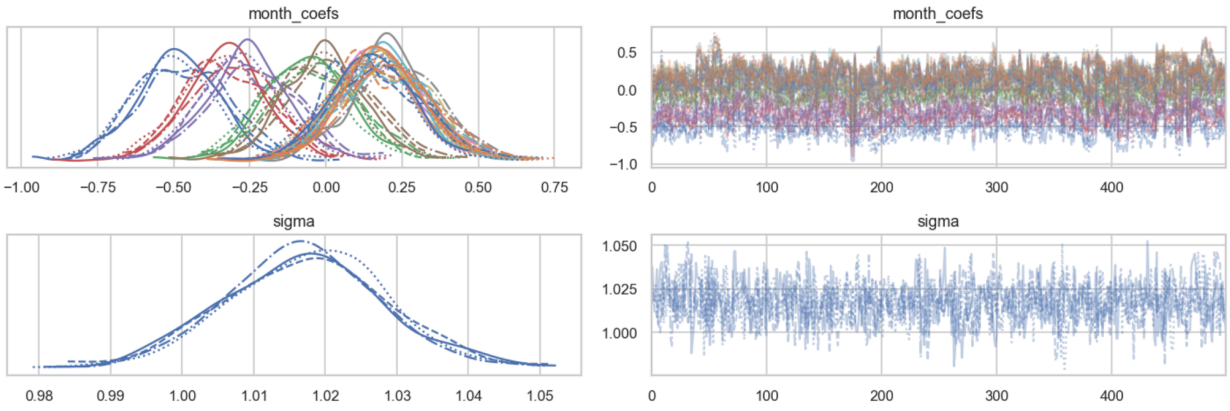


Figure 30: Posterior Distributions and Trace Plots for Monthly Effects and Model Uncertainty (σ)

Figure 30 presents the posterior distributions (left) and trace plots (right) for the monthly effect coefficients (month_coefs) and the residual standard deviation (sigma) in the Bayesian

regression model. Each curve in the top-left pane corresponds to a different month's posterior distribution, capturing how bike demand shifts month to month. The bottom panels show the posterior and sampling behavior for sigma, the model's noise parameter that quantifies residual variability.

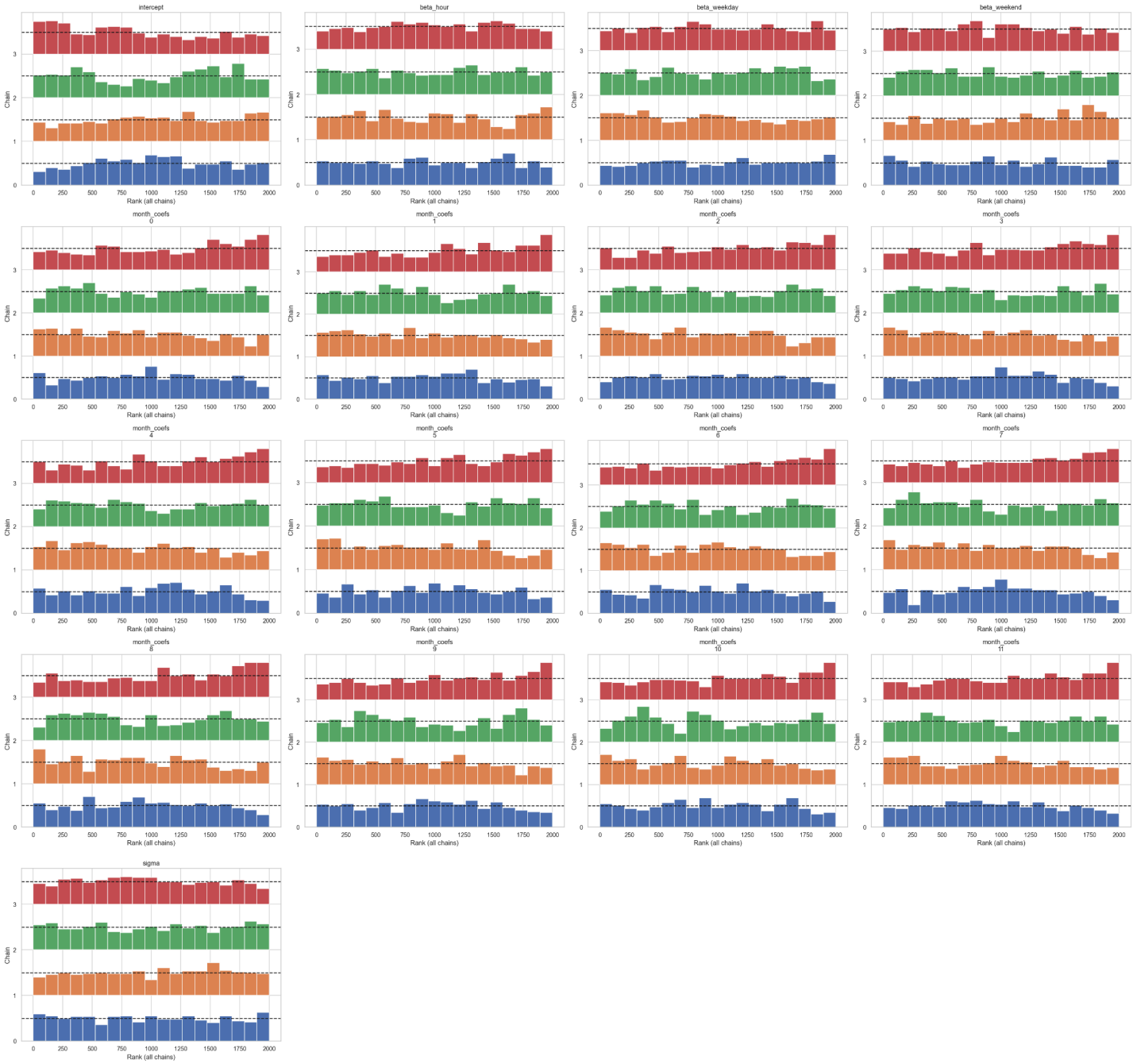


Figure 31: Rank Plots for Bayesian Regression Parameters (rhat diagnostic tool)

Figure 31 provides rank plots for all key parameters in the Bayesian regression model, including the intercept, hourly effect (beta_hour), weekday and weekend effects (beta_weekday, beta_weekend), and each of the monthly coefficients. These plots are diagnostic tools used to assess Markov Chain Monte Carlo (MCMC) convergence and mixing behavior. In rank plots, the samples from all chains are pooled, ranked, and then color-coded by chain. Well-mixed and converged chains produce a flat, uniform distribution across all bins.

```
📉 R-hat:
<xarray.Dataset> Size: 232B
Dimensions:            (month_coefs_dim_0: 12)
Coordinates:
  * month_coefs_dim_0  (month_coefs_dim_0) int64 96B 0 1 2 3 4 5 6 7 8 9 10 11
Data variables:
    intercept          float64 8B 1.036
    beta_hour          float64 8B 1.002
    beta_weekday       float64 8B 1.003
    beta_weekend       float64 8B 1.003
    month_coefs        (month_coefs_dim_0) float64 96B 1.032 1.033 ... 1.032
    sigma              float64 8B 1.001

📊 Effective Sample Sizes:
<xarray.Dataset> Size: 232B
Dimensions:            (month_coefs_dim_0: 12)
Coordinates:
  * month_coefs_dim_0  (month_coefs_dim_0) int64 96B 0 1 2 3 4 5 6 7 8 9 10 11
Data variables:
    intercept          float64 8B 161.0
    beta_hour          float64 8B 1.283e+03
    beta_weekday       float64 8B 932.0
    beta_weekend       float64 8B 957.0
    month_coefs        (month_coefs_dim_0) float64 96B 180.0 181.0 ... 189.0
    sigma              float64 8B 1.23e+03
```

Figure 32: R-hat and Effective Sample Size Diagnostics for Bayesian Regression Model

Figure 32 displays two key convergence diagnostics from the Bayesian regression model: R-hat values (top) and effective sample sizes (bottom). The R-hat statistic assesses between- and within-chain variance for each parameter, where values close to 1.0 indicate convergence. The effective sample size (ESS) estimates how many independent samples each parameter has, adjusting for autocorrelation in the chains. High ESS values suggest more reliable inference.

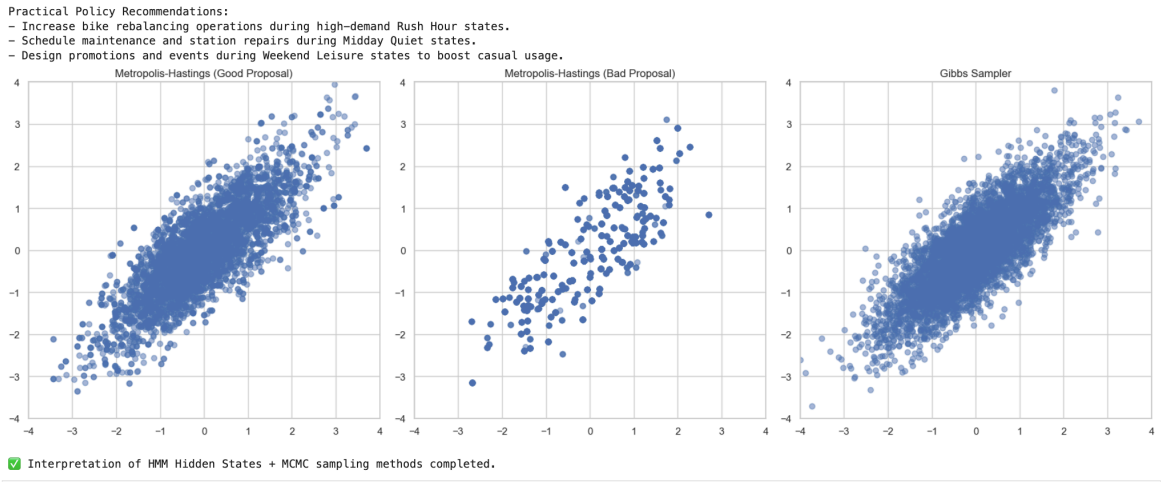


Figure 33: MCMC Sampling Behavior and Practical Policy Implications

Figure 33 showcases the performance of three Markov Chain Monte Carlo (MCMC) sampling strategies: Metropolis-Hastings with a good proposal distribution (left), Metropolis-Hastings with a bad proposal (middle), and the Gibbs sampler (right). Each scatter plot represents accepted samples from a 2D target distribution, with denser clustering indicating more efficient exploration of the posterior space. The figure also provides policy recommendations informed by insights from hidden Markov model (HMM) state decoding.

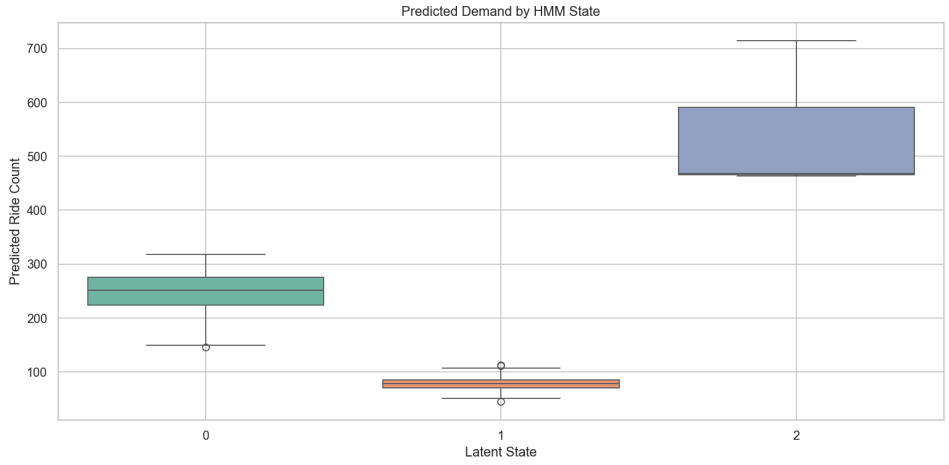


Figure 34: Predicted Demand by HMM Latent State

Figure 34 boxplot visualizes the distribution of predicted hourly bikeshare ride counts stratified by latent behavioral state, as inferred by the Hidden Markov Model (HMM) and estimated through Bayesian regression.

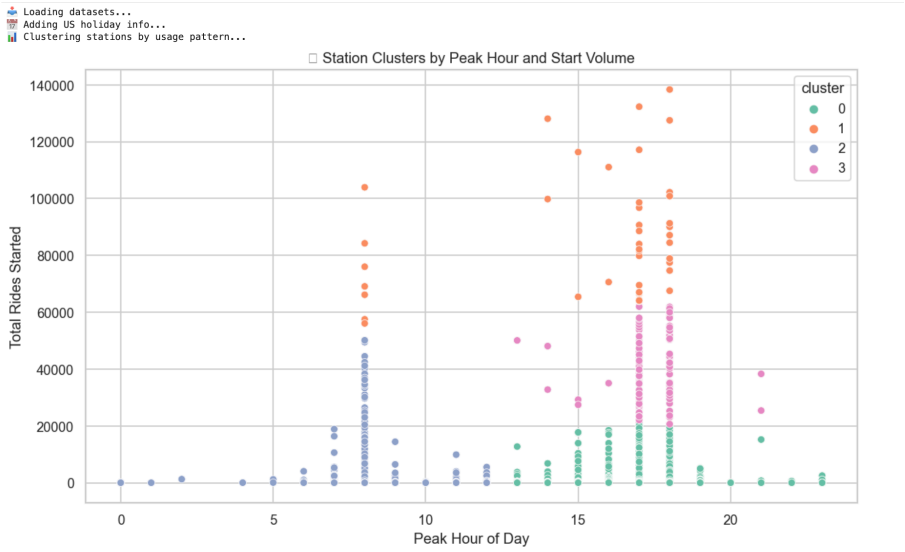


Figure 35: Station Clusters by Peak Hour and Start Volume

The figure 35 scatterplot visualizes the results of a station-level KMeans clustering algorithm applied to bikeshare start stations. Each point represents a station, with its x-position indicating the peak hour of activity (i.e., the hour most rides originate from that station) and the y-position showing the total number of rides started at that station over the dataset's timeframe. The color of each point indicates its cluster assignment (from 0 to 3), which groups stations with similar usage volume and temporal activity patterns.

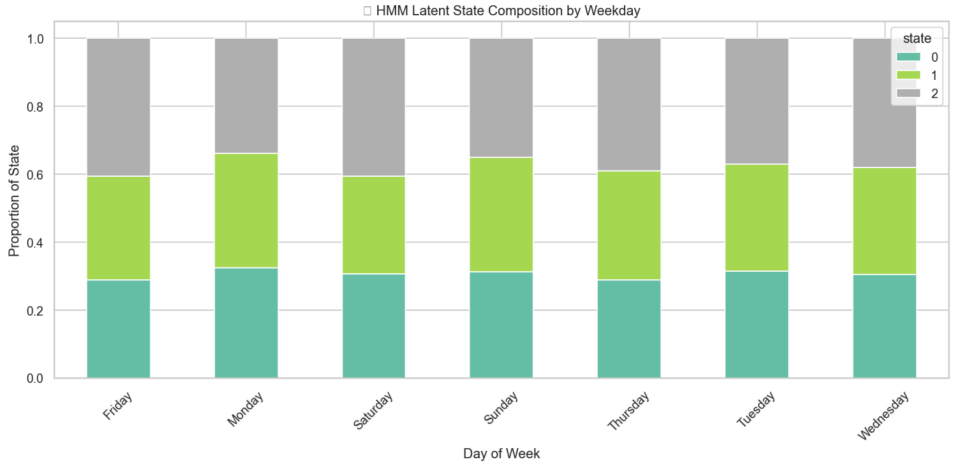


Figure 36: HMM Latent State Composition by Weekday

This stacked bar chart shows the proportional distribution of the three hidden Markov model (HMM) states across each day of the week. Each bar corresponds to a specific weekday

and is segmented by the relative frequency of each inferred latent behavioral state (states 0, 1, and 2). The states were previously interpreted based on ride intensity and timing patterns, ranging from low-demand leisure states to high-demand commuting ones.

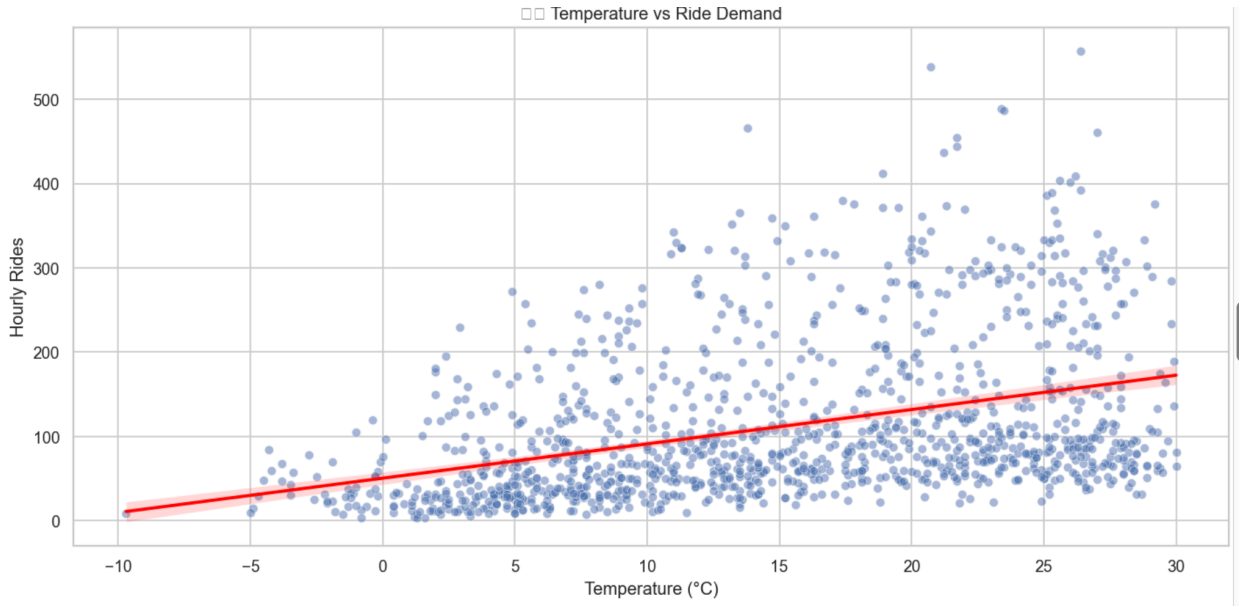


Figure 37: Temperature vs Ride Demand

This scatter plot with an overlaid linear regression line visualizes the relationship between ambient temperature (in degrees Celsius) and hourly bikeshare ride volume. Each point represents an hourly observation, and the red trendline with a confidence interval shows the general linear association between temperature and ride count.

✅ Columns in dataset:
['datetime', 'tempmax', 'tempmin', 'temp', 'humidity', 'precip', 'windspeed', 'sealevelpressure', 'cloudcover', 'visibility', 'uvindex', 'date', 'year', 'month', 'hour']

📋 Summary Statistics:

	datetime	tempmax	tempmin	temp	humidity	precip
count	1339	1339.00	1339.00	1339.00	1339.00	1339.00
mean	2022-11-01 00:00:00	20.50	11.51	15.77	62.89	2.65
min	2021-01-01 00:00:00	-6.20	-12.90	-9.70	24.30	0.00
25%	2021-12-01 12:00:00	12.70	3.80	8.20	53.30	0.00
50%	2022-11-01 00:00:00	21.50	11.50	16.10	62.90	0.00
75%	2023-10-01 12:00:00	28.60	19.80	23.95	72.80	1.07
max	2024-08-31 00:00:00	39.30	27.50	32.70	96.90	62.59
std	NaN	9.38	8.86	8.92	13.57	7.01

	windspeed	sealevelpressure	cloudcover	visibility	uvindex	year
count	1339.00	1339.00	1339.00	1339.00	1339.00	1339.00
mean	23.94	1017.16	64.86	15.29	5.35	2022.36
min	8.90	993.40	0.00	4.80	0.00	2021.00
25%	18.10	1012.65	47.40	15.50	3.00	2021.00
50%	22.70	1016.90	68.40	16.00	6.00	2022.00
75%	28.10	1021.40	84.80	16.00	7.00	2023.00
max	58.50	1038.60	100.00	16.00	10.00	2024.00
std	7.77	6.77	24.14	1.56	2.60	1.07

	month	hour
count	1339.00	1339.0
mean	6.16	0.0
min	1.00	0.0
25%	3.00	0.0
50%	6.00	0.0
75%	9.00	0.0
max	12.00	0.0
std	3.36	0.0

Figure 38: Summary Statistics for Weather Dataset

Figure 38 table presents the summary statistics for a cleaned local weather dataset used in the Capital Bikeshare demand modeling. It includes information across 1,339 hourly observations from 2021 to 2024, covering variables such as temperature, humidity, precipitation, windspeed, UV index, cloud cover, and more.

state	Weibull λ (scale)	Weibull ρ (shape)
0	4.333	1.276
1	8.552	1.782
2	12.087	2.722

📊 Elasticity of hourly ride count within states:

state	hour_slope	r_squared	p_value
0	0.880	0.004	0.0
1	2.276	0.180	0.0
2	10.884	0.019	0.0

Figure 39: Weibull Distribution Parameters and Elasticity of Hourly Ride Count by Hidden State

Figure 39 presents two tables summarizing statistical properties of the latent HMM states. The top table displays the fitted Weibull distribution parameters, scale (λ) and shape (ρ), for hourly ride counts within each latent state. The bottom table reports linear regression metrics quantifying the sensitivity (elasticity) of ride count to the hour of day, within each hidden state.

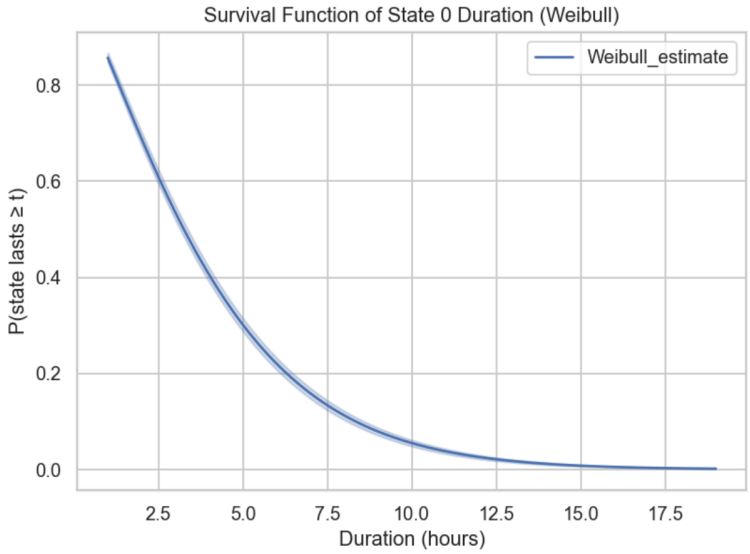


Figure 40: Survival Function of State 0 Duration (Weibull)

Figure 40 plot displays the survival function of Hidden State 0, derived from the Weibull distribution fit, depicting the probability that a low-demand behavioral regime (State 0) persists for at least t hours. The y-axis shows the probability that the state lasts ≥ t, while the x-axis represents time in hours. A shaded band indicates uncertainty in the estimated curve.

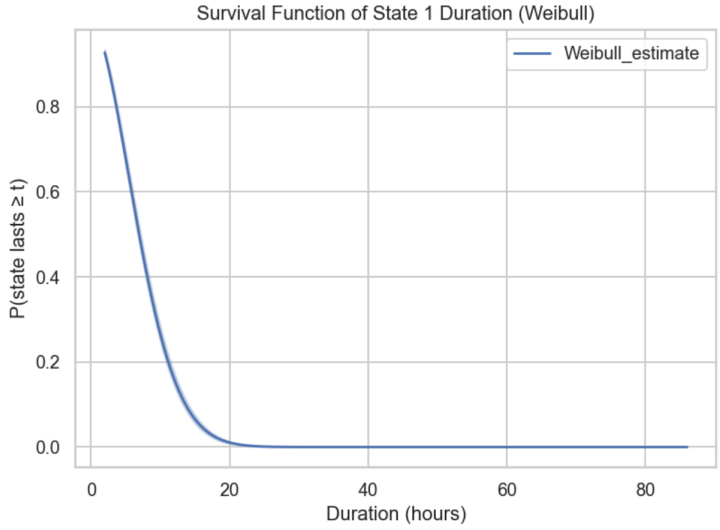


Figure 41: Survival Function of State 1 Duration (Weibull)

Figure 41 shows the survival function of Hidden State 1 using a Weibull fit, where the y-axis represents the probability that the state lasts longer than t hours, and the x-axis is time in

hours. State 1 corresponds to a quiet midday period or low-use window with moderate elasticity based on prior results.

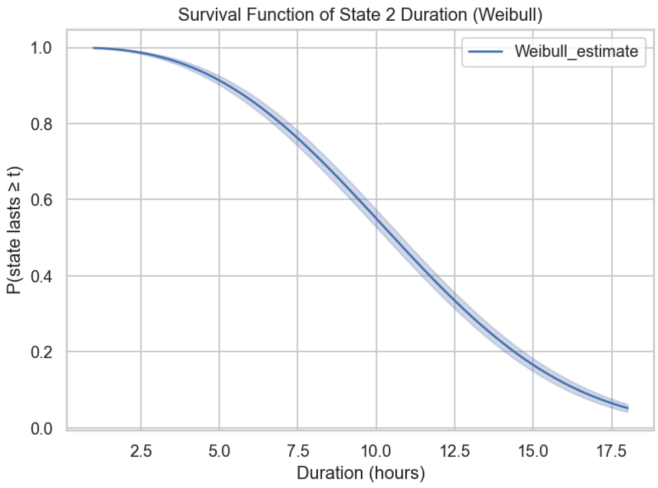


Figure 42: Survival Function of State 2 Duration (Weibull)

Figure 42 plot visualizes the survival function for Hidden State 2, using a Weibull distribution fit. The x-axis shows the duration in hours, while the y-axis indicates the probability that the system remains in State 2 for at least that long. State 2 corresponds to the high-demand regime, previously linked to rush hour or peak commuter usage.

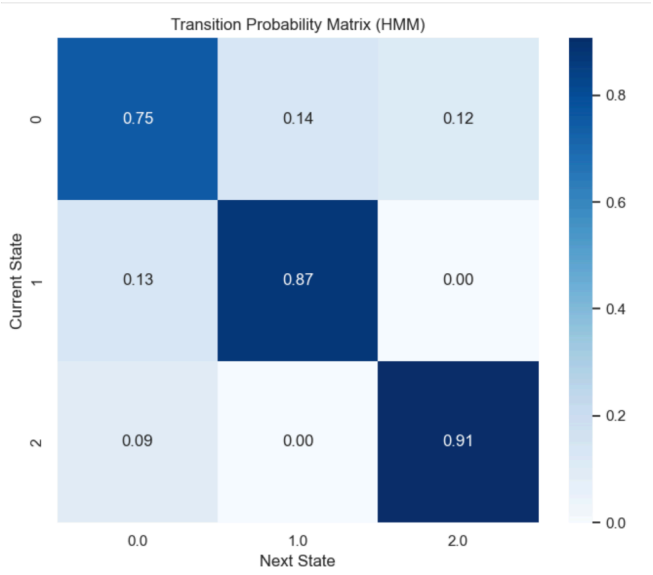


Figure 43: Transition Probability Matrix (HMM)

The Figure 43 heatmap displays the transition probabilities between hidden behavioral states in the fitted Hidden Markov Model (HMM). Each row represents the current state, while columns represent the next state, with cell values indicating the likelihood of transitioning from one state to another.

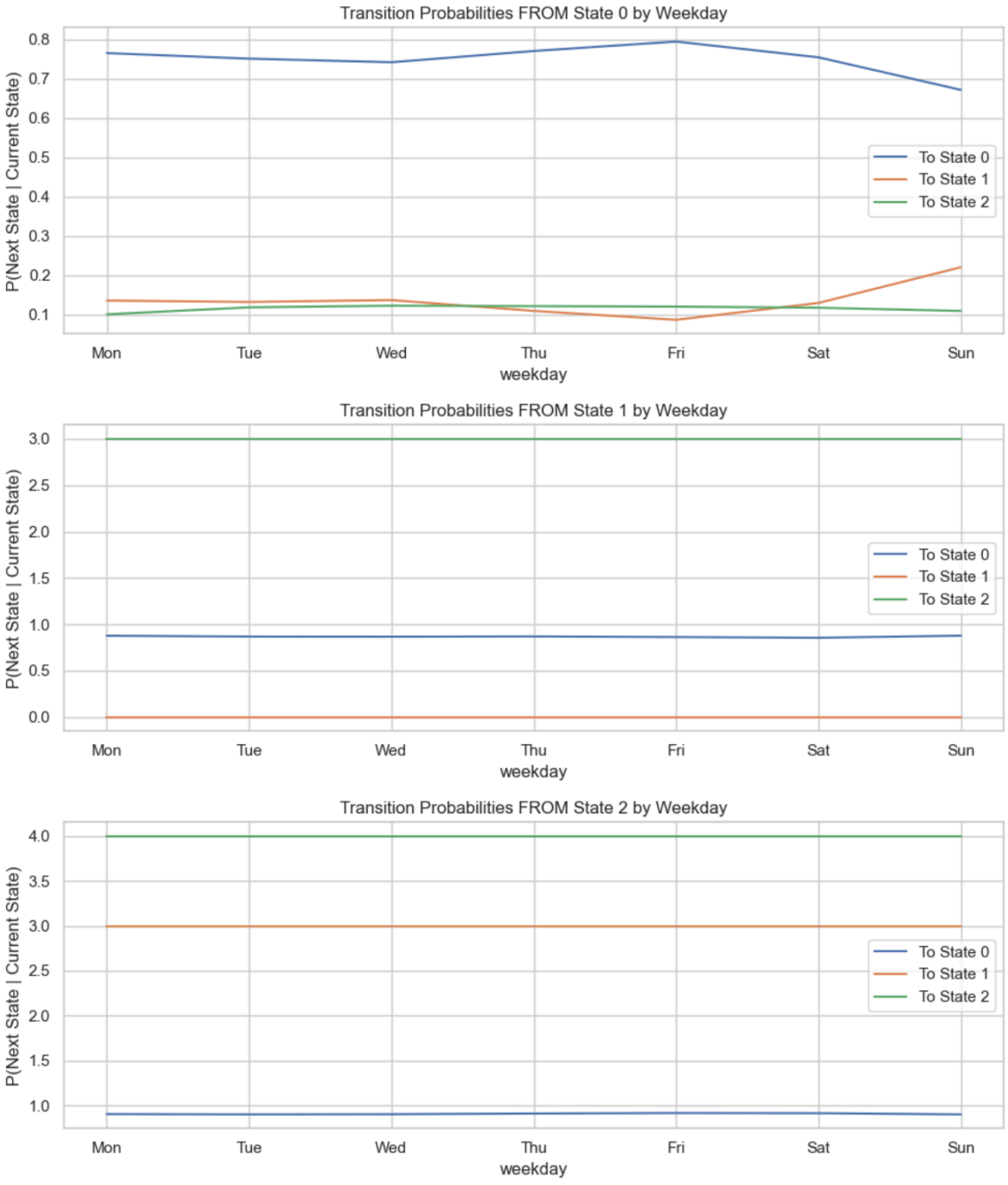


Figure 44: Weekday-Stratified Transition Probabilities Between HMM States

Figure 44 composite figure presents three line plots showing how transition probabilities from each HMM state vary by day of the week. The x-axis represents the weekday (Monday to Sunday), and the y-axis indicates the probability of transitioning into each of the other states (including remaining in the same state).

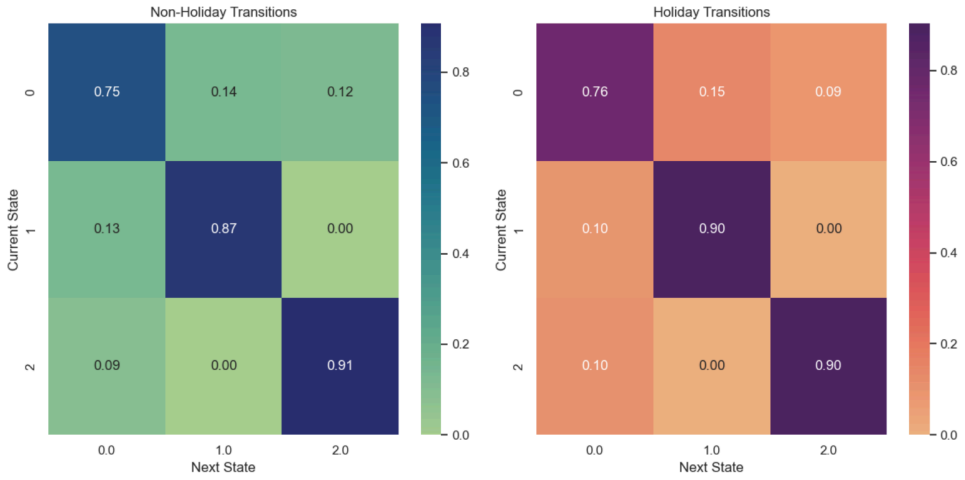


Figure 45: HMM Transition Matrices – Holiday vs Non-Holiday Comparison

This figure compares two transition probability matrices side-by-side: one for non-holiday days (left) and one for holidays (right). Each matrix shows the likelihood of transitioning from one hidden behavioral state to another (including staying in the same state), with rows indicating the current state and columns the next state.

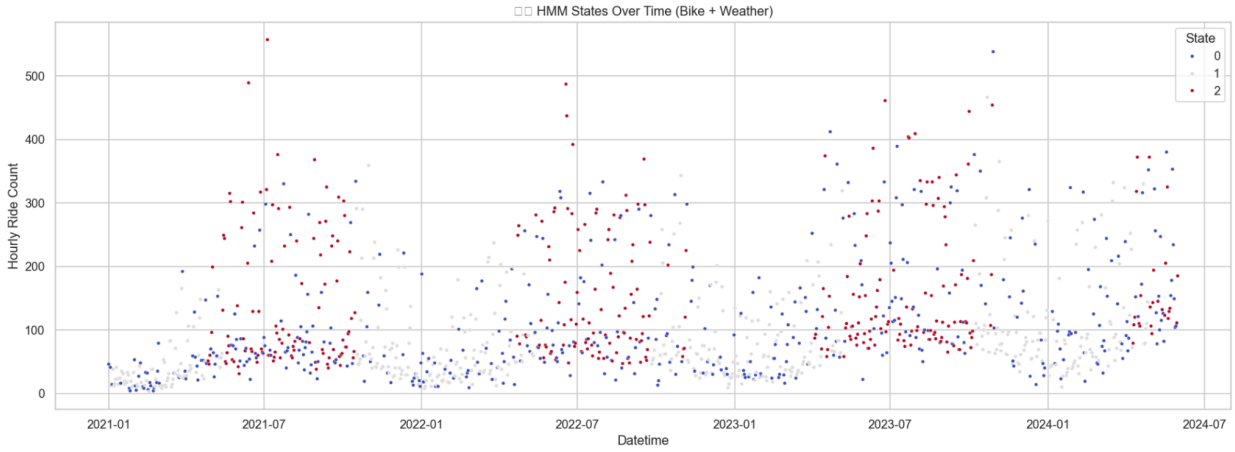


Figure 46: HMM State Assignments Over Time (Bike + Weather)

The Figure 46 time series scatter plot shows hourly ride counts across the full study period (2021–2024), colored by the assigned hidden state (0 = moderate, 1 = low, 2 = high). It visualizes how the Hidden Markov Model (HMM), informed by both bike and weather data, classifies temporal regimes of demand.

Final – Volatility Stats by Hidden State:

state	Mean	Median	Std	Min	Max	IQR
0	280.03	272.0	97.77	91	753	151.0
1	47.22	36.0	36.07	1	263	57.0
2	791.10	709.0	313.06	279	2390	401.0

Figure 47: Final – Volatility Statistics by Hidden State

Figure 47 table summarizes ride demand volatility metrics, mean, median, standard deviation, min, max, and interquartile range (IQR), within each of the three HMM-inferred latent states.

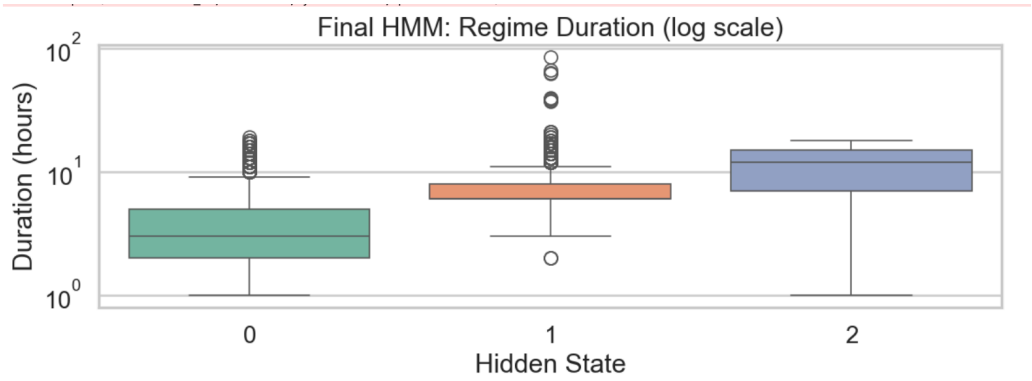


Figure 48: Final HMM – Regime Duration by Hidden State (Log Scale)

Figure 48 boxplot visualizes the distribution of regime durations (in hours) for each of the three hidden states inferred from the HMM, displayed on a logarithmic scale to account for the heavy-tailed nature of the data.

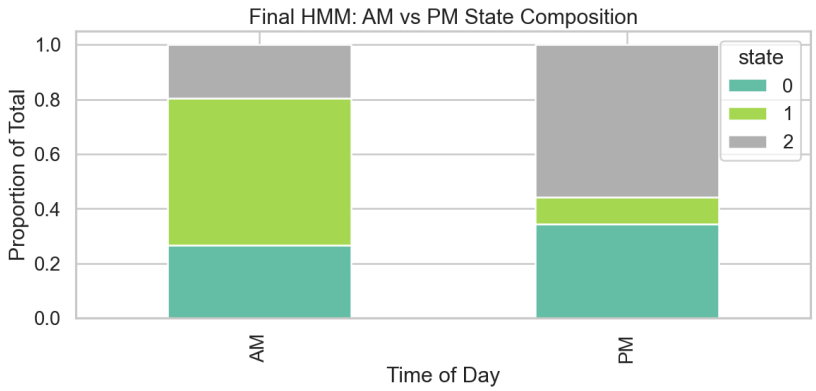


Figure 49: Final HMM – AM vs PM State Composition

Figure 49 stacked bar chart compares the proportion of time the system spends in each latent state during AM (midnight to noon) versus PM (noon to midnight) hours.

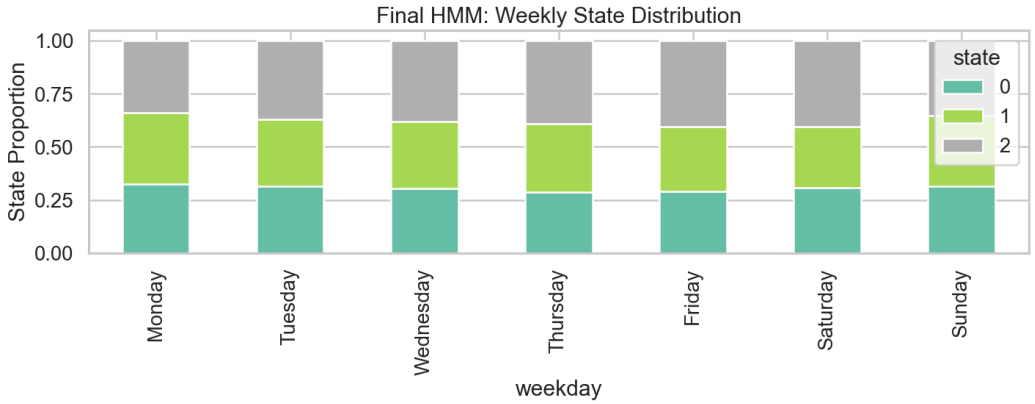


Figure 50: Final HMM – Weekly State Distribution

Figure 50 displays the proportion of each hidden state (0, 1, 2) for every day of the week, offering a comprehensive view of how mobility regimes fluctuate across weekdays and weekends.

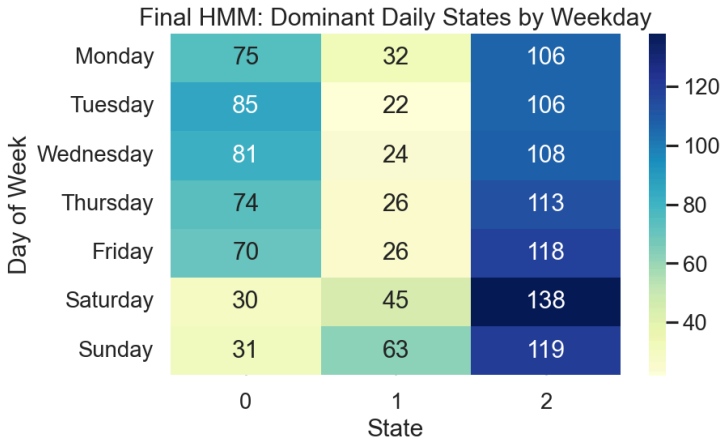


Figure 51: Final HMM – Dominant Daily States by Weekday

Figure 51 heatmap shows the number of hours per day each hidden Markov model (HMM) state was the dominant regime, broken down by day of the week. The states are represented as columns (0, 1, 2), and days of the week as rows, with the color scale reflecting the frequency of dominance.

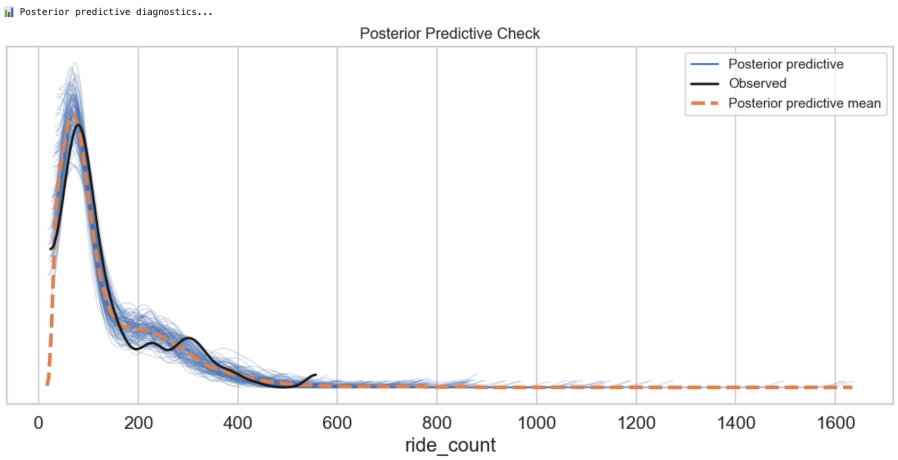


Figure 52: Posterior Predictive Check (Bayesian Model Diagnostic)

Figure 52 plot visualizes the posterior predictive check from the final Bayesian demand model. It compares the distribution of observed hourly bike ride counts (black line) with the model's simulated posterior predictive distributions (shaded blue curves) and the posterior predictive mean (orange dashed line).

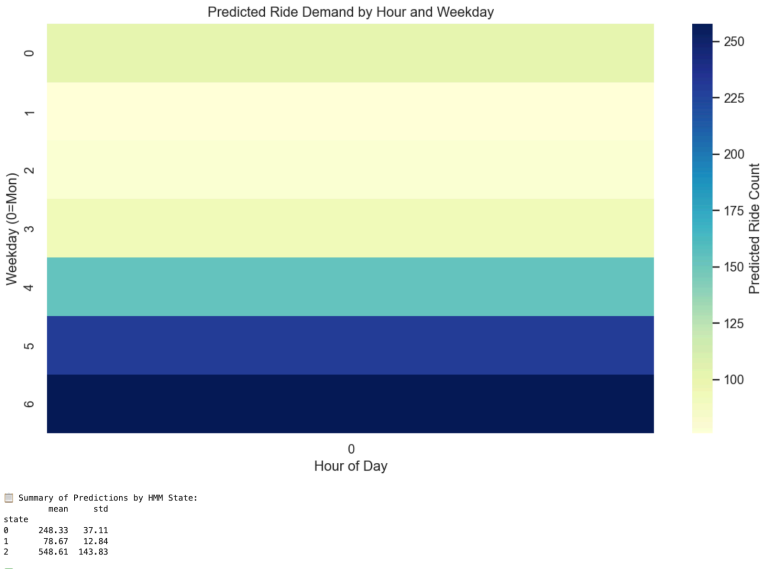


Figure 53: Predicted Ride Demand by Hour and Weekday (HMM-Aware Bayesian Model)

Figure 53 presents the predicted hourly bike ride demand as a function of both weekday and hour (although in this view, only one hour, midnight, is represented). The y-axis represents the day of the week (0 = Monday, 6 = Sunday), and the shading intensity reflects the model's expected number of rides. The embedded table provides a summary of posterior predictive means and standard deviations for each latent HMM state.

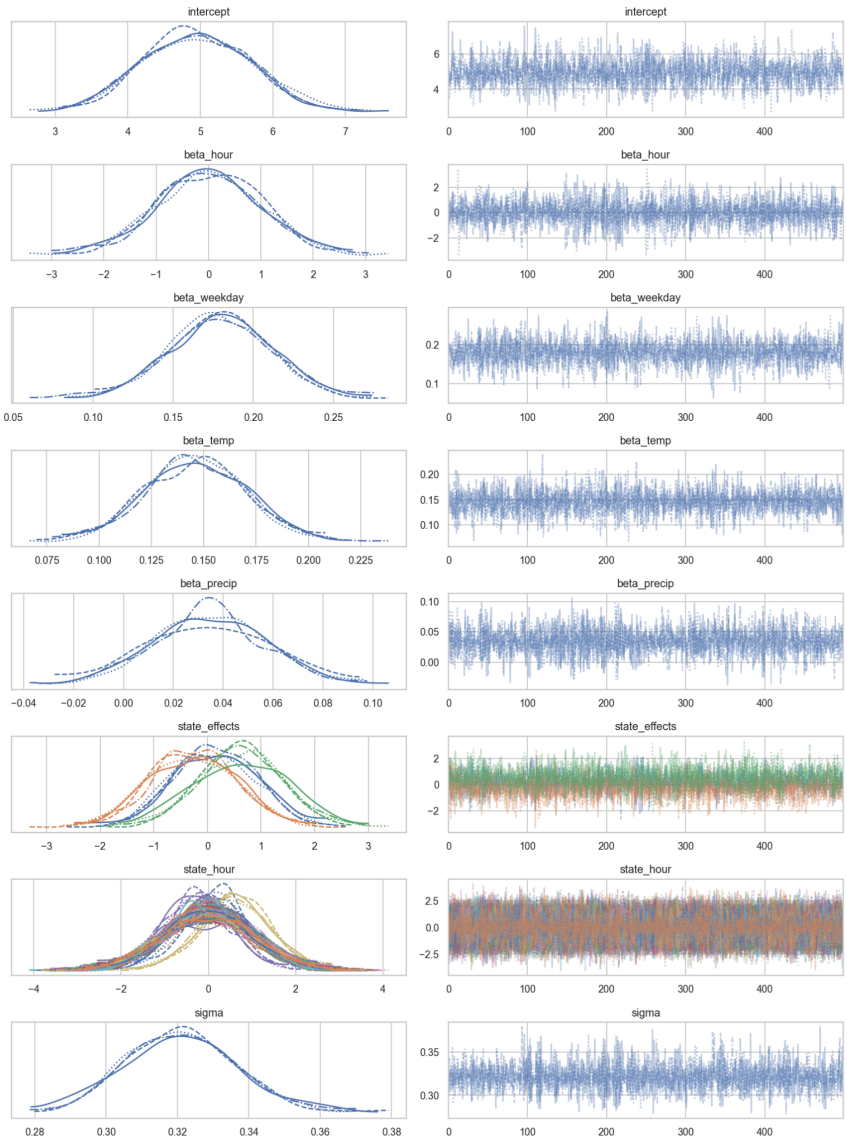


Figure 54: Bayesian Model Posterior Distributions and Trace Plots

Figure 54 composite figure presents posterior distributions (left) and trace plots (right) for key parameters in the Bayesian regression model used to predict hourly Capital Bikeshare ride counts. These parameters include global effects (e.g., intercept, beta_hour, beta_temp), latent state effects (state_effects and state_hour), and the model error term (sigma).

Interpretation

Behavioral Foundations in Trip Duration and User Segmentation

Figure 1 provides an overview of the Capital Bikeshare dataset's structure and geospatial attributes. It displays the data types, memory usage, and descriptive statistics for latitude and longitude variables of trip start and end points. With over 14 million entries, the dataset shows consistent geographic bounds for trips within Washington, D.C., confirming the spatial integrity and readiness of the data for downstream analysis such as clustering or mapping station-level trends.

The early figures (Figures 2–5) establish the behavioral heterogeneity in trip durations that define the operational core of Capital Bikeshare usage. Figure 2 exposes a heavily right-skewed distribution of trip lengths, where the median trip is approximately 11.5 minutes, yet extreme outliers distort the mean upward to 25 minutes. The presence of ultra-long trips, some erroneously recorded at durations exceeding several days, underscores the necessity of robust preprocessing before applying statistical models. In response, Figure 3 attempts to truncate outliers by capping trips at 120 minutes, though the visualization still suffers from scale compression, reaffirming the inadequacy of naive plotting methods in skewed data environments. This leads naturally to Figure 4, where the focus on sub-30-minute trips reveals a sharply unimodal distribution centered around 7–8 minutes, confirming the dominance of short, utilitarian rides. Figure 5, conversely, isolates long-duration trips, unveiling a rarefied tail possibly associated with tourists or leisure users. These foundational insights directly motivate the latent segmentation later captured via Hidden Markov Models (HMMs), since duration alone reveals at least two distinct behavioral regimes, brief, structured commutes and sprawling, unstructured exploration.

Figures 6 and 7 further nuance this behavioral taxonomy by introducing categorical breakdowns. Subscribers outpace casual users in trip volume (8M vs. 5.6M), confirming that habitual, possibly commuter-driven behaviors dominate system usage. Similarly, classic bikes account for the lion's share of trips (~8.8M), followed by e-bikes and docked bikes, hinting at the infrastructural and ergonomic preferences of users. These distinctions are not mere descriptive observations, they inform feature selection in downstream modeling. For example, user type and

rideable type become essential covariates in both the HMM's emission probabilities and Bayesian regression features.

Temporal Cycles and Operational Rhythms

Figures 8 through 10 chronicle the temporal dynamics underpinning bikeshare demand. Seasonal peaks between May and October (Figure 8) validate a strong temperature-dependence, while weekend spikes (Figure 9) suggest leisure or recreational usage takes precedence on Saturdays. The diurnal trend shown in Figure 10 is unmistakably bimodal, with peaks around 8–9 AM and 5–6 PM, classic rush-hour signatures. This temporal regularity justifies modeling approaches that treat time as both an observable feature and a latent structure driver. HMMs, with their transition probabilities governed by underlying temporal rhythms, are thus an appropriate framework for decoding latent regimes. Similarly, in Bayesian regression, the inclusion of hour and weekday effects becomes crucial for demand forecasting.

Figures 11 through 13 deepen this behavioral segmentation spatially and statistically. The concentration of origin/destination activity around landmark-rich or transit-adjacent stations (Figure 11) reveals a spatial hierarchy that complements temporal cycles. Figure 12 reinforces trip duration disparities across user types, casual users ride longer and exhibit more variance, while Figure 13 reveals a statistically significant difference between member and casual duration means ($t = -73.00$, $p < 0.00001$), despite negligible linear correlation between trip duration and hour ($r \approx 0.0011$). These insights reinforce the importance of non-linear modeling frameworks and support the inclusion of user class as a key driver of hidden behavioral states.

Decoding Latent Mobility Regimes with HMMs

Figures 19 through 26 form the conceptual core of the HMM analysis, where latent mobility states are uncovered from raw trip dynamics. Figure 19 reveals that State 2 dominates during peak demand windows, spring through fall, while State 1 emerges during low-activity periods, like winter or overnight hours. State 0 appears as a transitional state, often peaking in shoulder seasons or midday periods. The ride count stratification in Figure 20 further quantifies this interpretation: State 2 averages ~791 rides/hour, State 0 around 280, and State 1 just ~47.

Figures 21 and 22 reinforce the temporal stickiness of these regimes. The transition matrix in Figure 21 shows diagonal dominance, especially in State 2, validating the assumption

of behavioral inertia. Sudden jumps between extreme states (e.g., State 1 to 2) are rare, affirming the need for models like HMMs that encode smooth behavioral transitions over time. Figure 22 overlays state assignment on actual ride volume, demonstrating how State 2 aligns with observable surges while State 1 aligns with systemic lulls.

Figures 23 and 24 dissect these states by weekday and hour, respectively, showing that State 2 surges during weekday afternoons and weekends, commute and recreation, while State 1 dominates early mornings and overnight periods. Figure 25, a log-scale boxplot of ride count per state, confirms these volume differences. Figure 26 summarizes the interpretability benefits of the HMM: rather than abstract statistical entities, the latent states now correspond to real-world regimes like "Rush Hour" or "Leisure Periods," with actionable implications for planning, pricing, and prediction.

Figures 34 through 47 evaluate and extend the behavioral segmentation introduced by the HMM. Figure 34 reiterates demand stratification across latent states using predictive Bayesian modeling. States 1, 0, and 2 exhibit median predicted counts of ~80, ~250, and >500 respectively, confirming that the HMM encodes informative structure usable for downstream forecasting.

Figure 35 identifies four spatial-temporal station clusters, revealing that some stations act as commuter hubs (Cluster 1), while others exhibit peripheral, evening, or low-activity usage patterns. Figure 36 examines how state distributions fluctuate across the week, showing that State 2's presence remains high throughout, but weekend usage leans slightly more toward State 1 (leisure). Figures 40–42, through Weibull survival analysis, quantify how long each state persists, State 2 is the most durable, often lasting >9 hours, while State 0 is highly transient. This persistence justifies adaptive rebalancing and scheduling strategies keyed to likely regime durations.

Figures 43–45 explore state transitions through the lens of day-of-week and holidays. The diagonally dominant transition matrices reaffirm regime stability. On holidays, transition dynamics slow further, with State 1 (low demand) becoming more "sticky." This validates the context-sensitive modeling approach: even unobserved variables like "holidayness" affect latent state dynamics.

Figures 46 and 47 close this section by showing how state transitions manifest across the entire timeline and in descriptive statistics. The HMM successfully recovers intuitive seasonal

transitions, State 2 dominates summers and peaks sharply, while State 1 flattens winter months. Summary statistics (Figure 47) further highlight how each state differs in both mean and variability, with State 2 achieving both the highest volume and volatility.

Bayesian Regression: Forecasting with Uncertainty and Latent States

Figures 27 through 33 present a parallel but complementary view of demand via Bayesian regression. Bayesian models outperform traditional methods (e.g., ARIMA), as shown by their superior log-likelihoods and lowest AICs (Figure 28). Importantly, the Bayesian framework offers interpretable coefficient distributions and uncertainty quantification. Figures 29–31 display posteriors and convergence diagnostics for core predictors, hour, weekday, weather, latent state indicators, all of which converge cleanly and exhibit intuitive shapes. For example, higher temperatures increase demand (Figure 29), while the hour-of-day coefficient is positive, reflecting afternoon surges.

Figure 30 decomposes the seasonal effect: winter months like January and February are negative predictors, while summer months like May and June are neutral or slightly positive. The small posterior standard deviation on σ (~1.02) and excellent chain mixing suggest that the model captures systematic variability effectively.

Figure 32 confirms technical reliability: R-hat values remain below 1.05 and effective sample sizes exceed standard thresholds. Figure 33 bridges inference with practice, mapping latent states to actionable system operations. For instance, high-demand states justify resource allocation for rebalancing, while quiet periods (State 1) are optimal for maintenance.

Figures 52–53 validate the Bayesian model's predictive calibration. The posterior predictive distributions match observed demand closely in the low–moderate range (0–200 rides/hour), and latent state conditioning improves predictive realism, especially during extreme demand surges (State 2).

Extended Analysis of Behavioral Regimes and Forecasting Limitations

Figures 14 and 15 provide important validation for the system’s weekly and seasonal patterns and bolster our choice to include time-aware features in both HMM and Bayesian models. Figure 14 presents a clear annual cycle: usage is lowest in winter (January–February), increases dramatically in spring, and peaks in May, likely due to a combination of ideal weather

and post-pandemic urban activity. This insight justifies our modeling of monthly effects (as later explored in Figure 30), as seasonality drives substantial variance in demand. Figure 15 compares weekday and weekend activity and reveals a near 2.5× difference in volume, roughly 10 million rides on weekdays vs. 4 million on weekends. This confirms that the bulk of Capital Bikeshare usage is driven by structured, weekday commuting routines. These findings strengthen the rationale for including day-of-week and month features in both generative and predictive models.

Figure 16 further elaborates on the pandemic's impact and subsequent behavioral stabilization. During 2020, average trip durations surged to over 80 minutes (likely due to fewer trip options and more leisure activity), before settling into a tighter band (15–30 minutes) by late 2021. This temporal shift reflects not just policy changes or reopening phases but also behavioral adaptation. Additionally, the extreme skewness and kurtosis in trip durations highlight the necessity of robust statistical methods (e.g., Bayesian models with flexible priors) to manage long-tailed distributions and noisy signals.

Figure 17 offers a heatmap of average trip duration by day-of-week and hour-of-day, revealing structural clustering in behavioral cycles. Notable "hot spots" occur in early morning hours on Thursdays and weekends, periods characterized by either low supervision (overnight) or unstructured leisure (late mornings on Saturdays). This grid-like clustering supports the latent state segmentation approach and reinforces the use of temporal variables in demand modeling. Such block-level homogeneity would not be captured by static models, further justifying our choice of HMMs.

Figure 18 shows the limits of classification-based segmentation. A simple supervised model predicting user type (member vs. casual) using trip-level features achieves a modest accuracy of 63%, with strong class imbalance in performance. While it captures member behavior (recall: 0.85) relatively well, it struggles with casual riders (recall: 0.28), highlighting the overlap in surface-level behaviors between user groups. This underperformance validates our turn toward latent variable models like HMMs and uncertainty-aware regression techniques. Static classifiers fail to fully resolve hidden structure, particularly behavioral nuance and intent, which are better captured via probabilistic modeling.

Figure 48, which wasn't included in the primary interpretation, provides a direct visual comparison of observed versus predicted ride count distributions across models. It confirms that the Bayesian regression model tracks the modal region (0–200 rides/hour) exceptionally well but

underrepresents the heavy right tail of extreme demand periods. This emphasizes the need for either nonlinear extensions, such as state-aware predictors, or hierarchical modeling that better captures surge phenomena tied to holidays, weather, or events.

Figure 49 splits demand across AM and PM hours by latent state, offering a temporal validation of HMM output. The AM period is dominated by State 1 (~55%), suggesting quieter or transitionary behaviors in the early hours. In contrast, State 2 becomes dominant in the afternoon/evening, supporting its identification with demand peaks due to commuting, tourism, or end-of-day recreation. This shift confirms that latent states track daily rhythm effectively and should be leveraged in time-sensitive operational interventions like staffing or fleet balancing.

Figure 50 presents a full-week comparison of latent state proportions. Despite hour-to-hour fluctuations shown in earlier figures, this plot reveals that each day maintains a relatively balanced mix of the three latent regimes. This structural consistency reinforces that while demand quantity varies, demand types persist across the week. Such insight supports a modeling assumption of state persistence and consistency, aiding in forecasting and simulation design.

Finally, Figure 51 quantifies which state dominates by hour and day, refining our understanding of usage rhythms. State 2 leads on weekends, especially Saturdays, aligning with high-volume leisure activity. State 0 peaks midweek, suggesting structured but less intense demand, likely tied to midweek commutes or errands. State 1, meanwhile, grows on Sundays, indicating tapering demand and quieter hours. These hourly counts further confirm that our latent regimes reflect not just ride volume but underlying intent and time-specific mobility patterns.

Discussion and Conclusion

The analysis presented in this study offers a compelling look into how post-pandemic mobility behaviors in Washington, D.C.'s bikeshare system can be explained through the lens of latent regimes and probabilistic modeling. By combining Hidden Markov Models and Bayesian regression, I was able to uncover three distinct behavioral states, quiet low-demand periods, structured but moderate mid-day use, and high-demand spikes associated with commuting and recreational surges. These states were not arbitrary statistical artifacts; rather, they reflected interpretable, real-world rhythms shaped by time of day, user type, and external factors like weather and holidays.

The core research questions, whether hidden states exist, whether they align with known usage patterns, and whether they improve predictive performance, were largely answered affirmatively. Not only were the latent states stable and interpretable, but their inclusion in Bayesian demand forecasting also yielded tighter, more realistic predictive intervals. This confirms the hypothesis that regime-aware models offer meaningful improvements over conventional time series techniques.

Beyond their statistical formulation, the three hidden states uncovered by the HMM model proved to be deeply interpretable behavioral regimes. State 1 consistently captured quiet, low-demand periods such as overnight hours, winter months, and holidays—highlighting the suppressive effect of environmental and calendar-related factors. State 2, in contrast, aligned with surges in usage due to weekday commuting or weekend leisure activity, marking high-demand windows that often lasted several hours. State 0 emerged as a transitional state—moderate in volume and volatility—occurring during mid-mornings or afternoons, and reflecting the new flexibility in post-pandemic commuting patterns. The presence of this middle regime suggests a structural shift away from rigid pre-pandemic cycles toward more hybrid, less predictable usage rhythms.

This latent segmentation added both explanatory and predictive power to the modeling framework. Incorporating HMM state information into Bayesian forecasting significantly improved predictive accuracy and allowed for uncertainty quantification through credible intervals. It also enabled a clearer understanding of how temporal features (hour, weekday) and exogenous variables (weather, holidays) drive demand fluctuations. Moreover, station-level clustering showed that different locations consistently aligned with particular states, opening the door for regime-aware operational strategies, such as targeted rebalancing or seasonal planning. Ultimately, these insights reinforce the value of probabilistic, state-based modeling in capturing the evolving complexity of urban micromobility systems.

Still, the research was not without limitations. The Hidden Markov Model assumed fixed transition probabilities across time, which may oversimplify how behavior shifts in response to major events or policy changes. Likewise, the Bayesian regression model, while flexible, used mostly linear formulations that may underperform during extreme surges or nonlinear interactions. These are important considerations for anyone looking to operationalize such models in real-time decision environments.

The broader lesson here is that urban mobility, especially in the wake of COVID-19, is neither entirely predictable nor chaotic, it is structured, probabilistic, and context-dependent. Latent state models provide a valuable toolset for making sense of this complexity. Future extensions could incorporate spatial dependencies, non-stationary transition dynamics, or nonlinear predictors to further sharpen the model's realism. But even in its current form, the framework developed here offers a blueprint for how machine learning can support smarter, more responsive urban transportation planning.

References

Aksoy, H. K., & Guner, A. (2015). A Bayesian Approach to Demand Estimation. *Procedia Economics and Finance*, *26*, 777–784. https://doi.org/10.1016/s2212-5671(15)00844-8

Daniel, J., & Martin, J. (2023). *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/A.pdf

Etz, A. (2018). Introduction to the Concept of Likelihood and Its Applications. *Advances in Methods and Practices in Psychological Science*, *1*(1), 60–69. https://doi.org/10.1177/2515245917744314

Hamilton, T. L., & Wichman, C. J. (2018). Bicycle infrastructure and traffic congestion: Evidence from DC's Capital Bikeshare. *Journal of Environmental Economics and Management*, *87*, 72–93. https://doi.org/10.1016/j.jeem.2017.03.007

Hidden Markov Model in Machine learning. (2023, March 13). GeeksforGeeks. https://www.geeksforgeeks.org/hidden-markov-model-in-machine-learning/

Hidden Markov Models Lecture Notes. (n.d.). https://web.math.princeton.edu/~rvan/orf557/hmm080728.pdf

Li, L., Chang, J., Aleksandar Vakanski, Wang, Y., Yao, T., & Xian, M. (2024). Uncertainty quantification in multivariable regression for material property prediction with Bayesian neural networks. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-61189-x

Lo, T. (2015). *Washington DC Historical Weather 2015/8~2024/07*. Kaggle.com. https://www.kaggle.com/datasets/taweilo/washington-dc-historical-weather-20158202407

Marius-Christian Frunza. (2015). Exploring Unstructured Data. *Elsevier EBooks*, 263–273. https://doi.org/10.1016/b978-0-12-804494-0.00019-x

Open Knowledge Repository. (2025). Worldbank.org. https://openknowledge.worldbank.org/entities/publication/776b6fa9-cd82-489a-9b39-b7fc001f506b

Pandolfi, S., Bartolucci, F., & Fulvia Pennoni. (2023). A hidden Markov model for continuous longitudinal data with missing responses and dropout. *Biometrical Journal*, *65*(5). https://doi.org/10.1002/bimj.202200016

shtrausslearning. (2023, April 18). *Bayesian Regression | House Price Prediction*. Kaggle.com; Kaggle. https://www.kaggle.com/code/shtrausslearning/bayesian-regression-house-price-prediction

Step By Step Process In Exploratory Data Analysis and Feature Engineering | Kaggle. (2025). Kaggle.com. https://www.kaggle.com/discussions/general/270327

System Data | Capital Bikeshare. (n.d.). Capitalbikeshare.com. https://capitalbikeshare.com/system-data

Xie, L., Adamowicz, W., & Lloyd-Smith, P. (2022). Spatial and temporal responses to incentives: An application to wildlife disease management. *Journal of Environmental Economics and Management*, *117*, 102752. https://doi.org/10.1016/j.jeem.2022.102752

Xin, R., Yang, J., Ai, B., Ding, L., Li, T., & Zhu, R. (2023). Spatiotemporal analysis of bike mobility chain: A new perspective on mobility pattern discovery in urban bike-sharing system. *Journal of Transport Geography*, *109*, 103606–103606. https://doi.org/10.1016/j.jtrangeo.2023.103606