Pedal and Predict: Capital Bikeshare's Data-Driven Rebalancing Strategies

Code

Xiaxin Tang and Regy Septian

April 23, 2024

1 Introduction

Capital Bikeshare, initiated in September 2010, is the primary bike-sharing system across the Washington, DC metro region. This service is a partnership among multiple jurisdictions, including Arlington County, Alexandria, Montgomery County, Fairfax County, Prince George's County, and the City of Falls Church. With more than 4,700 bikes and over 575 docking stations, Capital Bikeshare offers convenient, economical transportation options for residents and visitors, enhancing mobility and accessibility throughout the region.

The demand for such a widespread service necessitates an effective re-balancing strategy to ensure that bikes are available when and where they are needed. To address this, the DDOT's re-balancing efforts could incorporate the use of dedicated fleets for moving bikes between high-demand areas and incentivizing users through programs that reduce rental costs when bikes are moved to specified locations that are in need of more bikes. This strategy would ensure operational efficiency and improve service reliability for users of the bike-share system [1][2][3].

2 Data Set Up

2.1 Import Bike Share Data

For this project, we used Capital Bikeshare data from March 2022[4], noting an increase in ridership from the previous month[5]. To prepare the data for analysis, we analyzed the entries in time intervals of 15 and 60 minutes. This organization helps maintain a consistent format for DateTime values, marked by year, month, day, hour, minute, and second. Additionally, we extracted the week of the observation (ranging from 1-52 throughout the year) and labeled each day of the week to facilitate daily trend analysis.

Show

2.2 Import Census Info

Census information is essential for providing spatial context to the demographic profile of areas surrounding the bike-share stations. For this project, we incorporated census data from the same year as our bike-share data, 2022. This approach ensures consistency and relevance in our analysis. The census data primarily helped us map the starting point of bike trips to their respective census tracts, simplifying the interpretation of our findings. Furthermore, this demographic data will later be crucial for testing the adaptability and robustness of our predictive models, ensuring they can be applied interchangeably across different urban contexts.

Show

2.3 Import Weather Data

To further understand the impact of environmental factors on bike share usage, we incorporated weather data from Ronald Reagan Washington National Airport (DCA) for March 2022. This period marks the beginning of spring in Washington, DC, which typically experiences varied weather conditions. We specifically analyzed hourly temperature, wind speed, and precipitation levels to observe how they might influence ridership patterns. Notably, Figure 2.3.1 shows that this month saw precipitation levels reaching up to 2 inches, wind speeds up to 20 mph, and temperatures ranging from 20 to 80°F. These factors are crucial for assessing the variability in bike share demand influenced by weather conditions.

Show

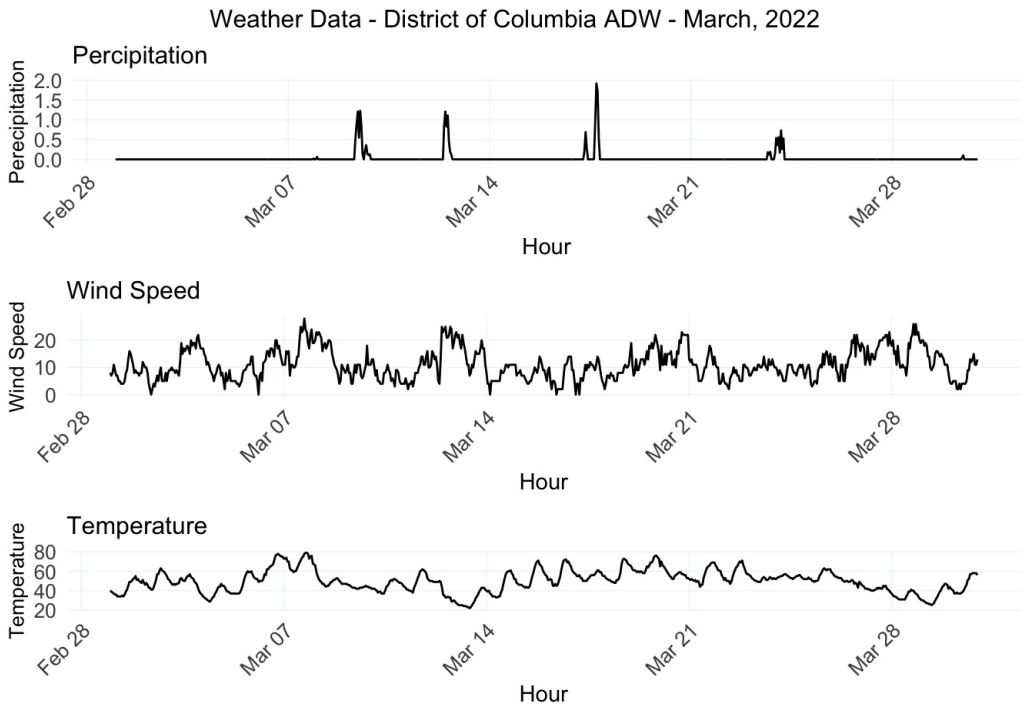

Figure 2.3.1

3 Exploratory Data Analysis

3.1 Bike Share Trip accross the Time

In this section of the analysis, we explored how bike share usage varied over time, focusing on hourly trends throughout the month. By examining the data, we observed a daily periodicity in bike usage, with distinct patterns emerging based on the time of day. Additionally, usage typically dipped during the weekends. Notably, the data from March 13-14 shows an anomaly, suggesting an unusual event or external factor that caused a deviation from normal usage patterns during these dates. See Figure 3.1.1.

Show

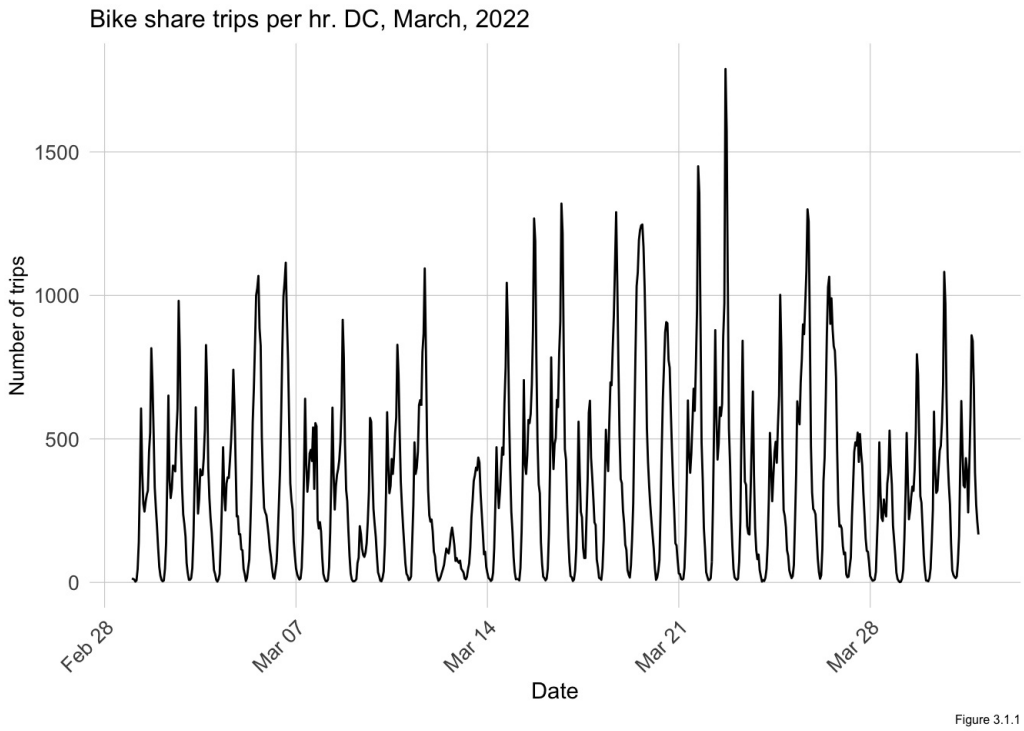

Figure 3.1.1

3.2 Hourly Trips per Station

In this analysis, we investigated the distribution of bike share trips across different times of the day and at various stations. Initially, we observed a distinct pattern in trip volume, with significant fluctuations between high-volume periods and times with lower usage. These variations suggested the need for different modeling approaches, such as Poisson regression for count data or linear models to better capture the fluctuations during peak times. By examining daily trends, we were able to further refine our understanding of how different times (morning rush, mid-day, evening rush, midnight) and types of days (weekends vs. weekdays)—see Figures 3.2.1-3.2.4—impact usage. Additionally, Figure 3.2.5 maps the distribution of trips per hour by stations across DC. This comprehensive temporal analysis helps in identifying specific periods that require attention for service improvement or targeted marketing strategies.

Histogram of Hourly Trips per Station

Show

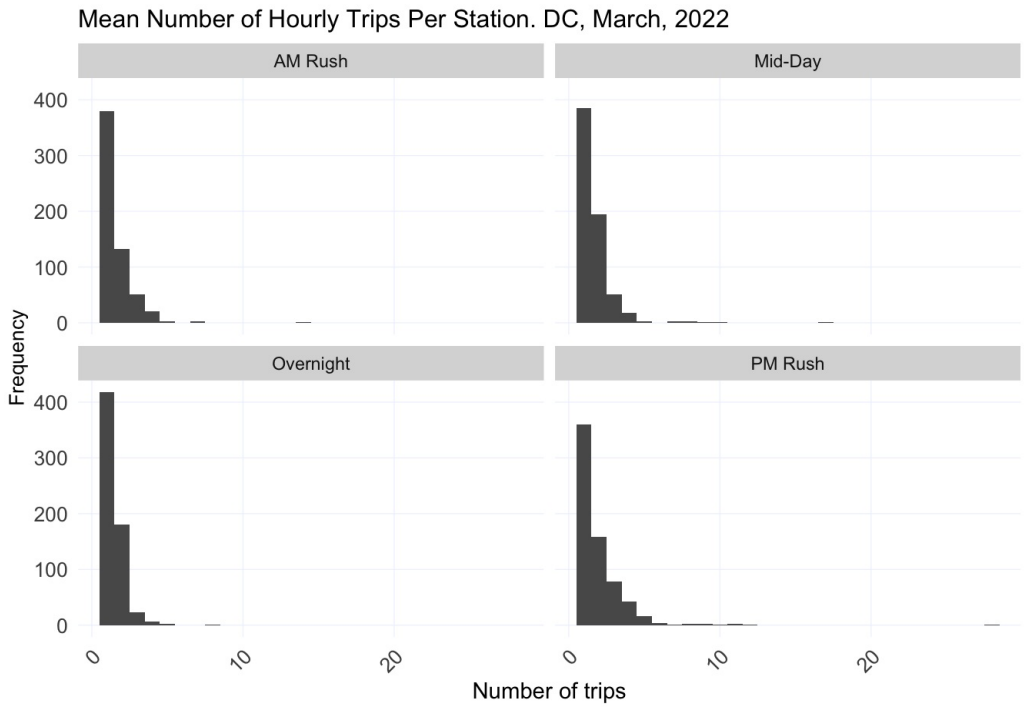

Figure 3.2.1

Visualization of Daily Trends

Show

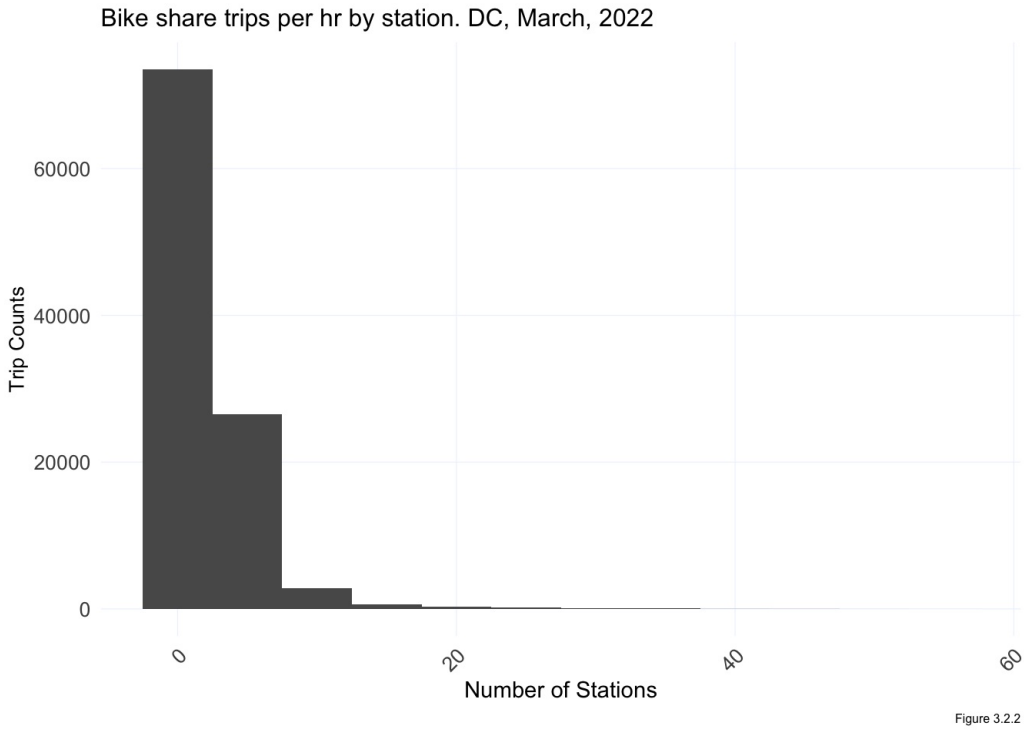

Figure 3.2.2

Weekday vs. Weekend Usage Patterns

Show

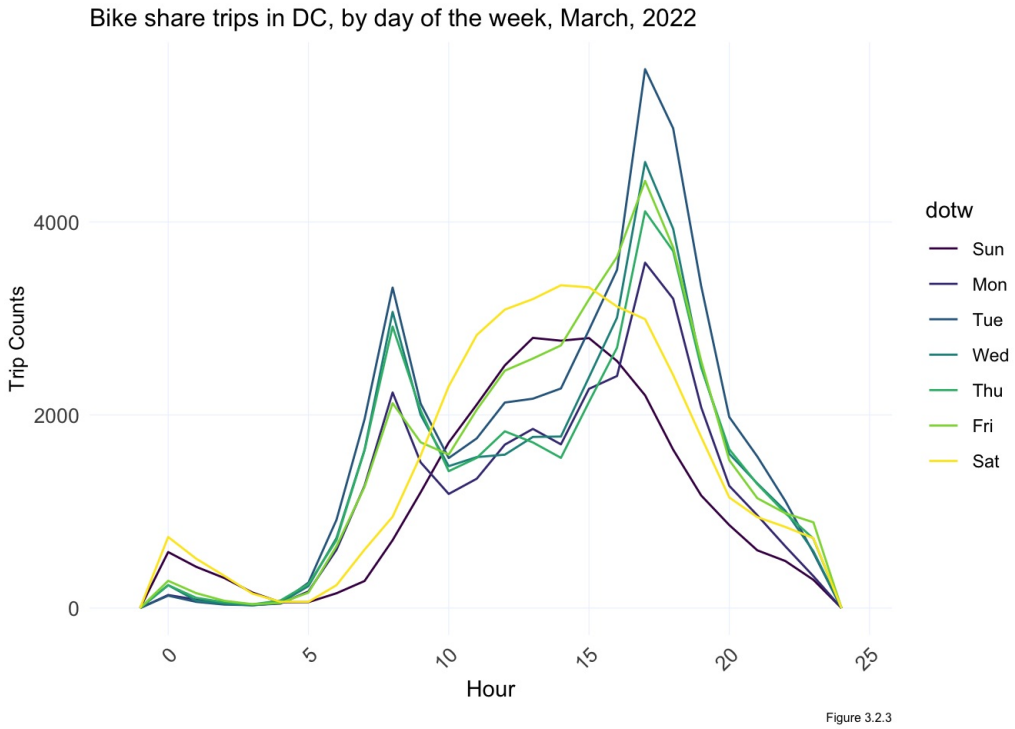

Figure 3.2.3

Show

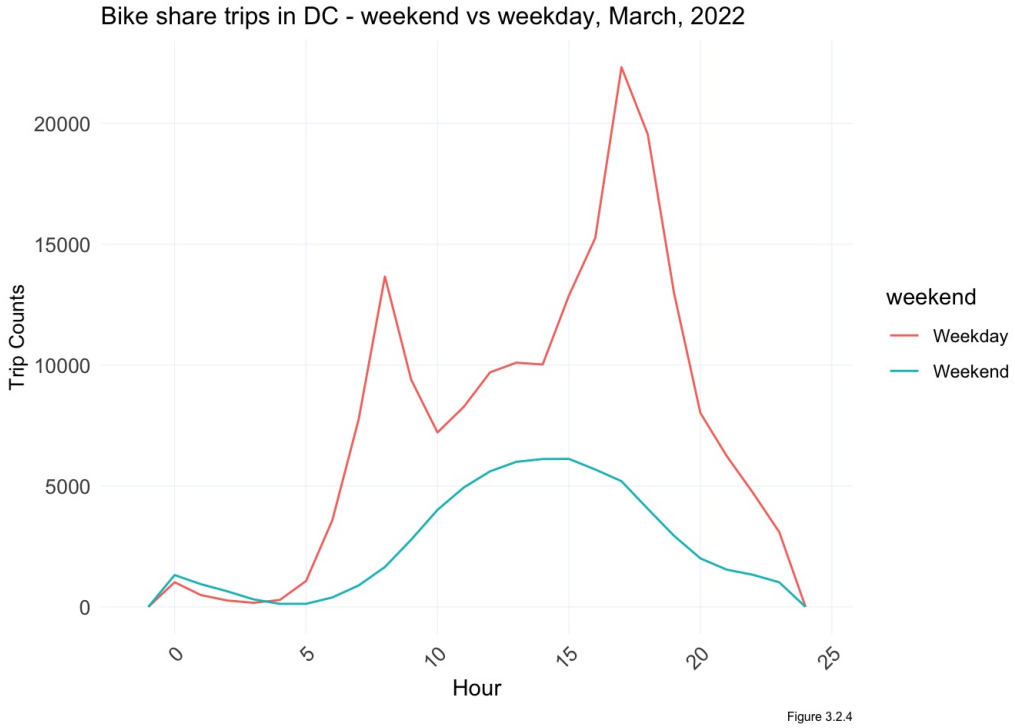

Figure 3.2.4

Map of Bike Share Trips

Show

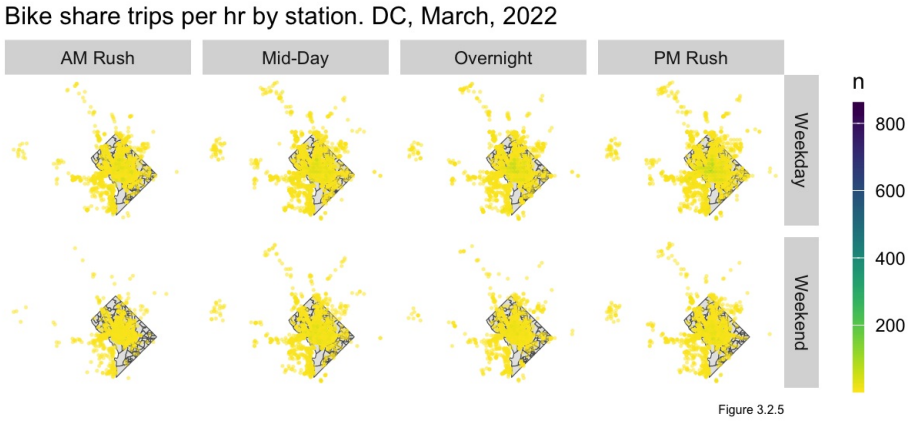

Figure 3.2.5

3.3 Create Space-Time Panel

To effectively analyze the temporal and spatial distribution of bike share usage, we ensure that our dataset represents every possible combination of station and time (hour/day). This method involves creating a 'panel' data structure, where each row corresponds to a unique station and time combination, regardless of whether any trips were recorded at that time. This comprehensive approach allows us to fill gaps with zeroes, ensuring that our analysis considers periods of inactivity as well as peak usage times.

Show

```
## [1] 504497
```

Show

```
## [1] 504497
```

We begin by identifying all possible combinations of times and stations. After that, we create an initial panel by pairing each unique time with every station. This process involves creating a comprehensive grid that combines these elements, ensuring no potential data point is overlooked.

Furthermore, we enhance this panel with relevant data such as station names and geographic coordinates, extracting only unique entries for accuracy. Additionally, we align this dataset with external data sources like weather and census information to enrich our analysis, allowing us to examine the influence of external factors on bike share usage patterns.

Show

3.4 Create time lags

To capture the dynamics of bike share demand over time, we introduce time lag variables into our dataset. These lags help us understand how past usage patterns influence future demand. Specifically, we calculate lags for various intervals—from hourly up to one day back—to analyze the immediate and delayed effects of past trip counts on current demand.

Additionally, we consider the impact of unusual events/holidays, such as the "Bans off Our Bodies" rally that occurred on March 14[6], by introducing dummy variables that indicate the proximity to these special dates. This analysis helps identify deviations in normal usage patterns during holidays and allows us to adjust operational strategies accordingly.

Show

Through correlation analysis, we measure the strength of the relationship between these lags and current trip counts. The one-hour lag shows a very strong correlation of 0.89, suggesting that recent past demand is an excellent predictor of immediate future demand. Other lags, such as two and three hours, also show significant positive correlations (0.71 and 0.52, respectively), although their predictive power decreases as the time gap increases. Notably, the negative correlation at twelve hours (-0.46) likely reflects daily demand cycles, with peaks occurring at different times of the day. The strong positive correlation for one-day lags (0.73) indicates that the demand pattern from the previous day can reliably predict the current day's demand at the same time. These insights are pivotal for planning short-term bike repositioning and informing long-term strategic decisions, ensuring operational efficiency and enhanced service reliability.

Show

```
## # A tibble: 6 × 2
##   Variable   correlation
##   <fct>            <dbl>
## 1 lagHour           0.89
## 2 lag2Hours         0.71
## 3 lag3Hours         0.52
## 4 lag4Hours         0.36
## 5 lag12Hours       -0.46
## 6 lag1day           0.73
```

4 Modeling

4.1 Run Models

Splitting the data into training and test sets based on time (e.g., weeks) rather than randomly is crucial for time-series analysis. This approach ensures the model is tested on completely unseen data, simulating a real-world scenario where future data points are predicted based on past observations. It helps to validate the model's effectiveness over time and ensures that temporal trends and seasonal effects are captured and appropriately understood.

Using data from the end of the month to predict the beginning of the following month can strategically capture monthly cyclic trends and changes in user behavior that are not immediately apparent within the same month. This method is applied to better understand how end-of-month activities, such as payday spending trends or monthly subscription renewals, might influence bike share usage.

Show

After the splitting process, we created five models by incorporating fixed effects, such as station name and day of the week. This approach helps control for unobserved heterogeneity, allowing the model to focus on the influence of other variables like weather conditions or lag effects. It provides a clearer picture of what factors significantly affect bike share usage.

Show

This sequence of models allows for a systematic examination of different factors influencing bike share usage. Each model builds on the previous one by adding new variables, enabling an assessment of how each additional factor contributes to the model's predictive power. Testing these models on separate data ensures that the findings are robust and that the models can generalize well to new, unseen data scenarios.

4.2 Predict for test data

Structuring the test data into nested data frames by week allows for a more organized and segmented approach to modeling. Nesting organizes the data so that each week's data is contained within its own subset, similar to having a folder for each week. This setup is crucial for applying different predictive models to each subset independently yet systematically.

Show

Using the `purrr::map()` function, we applied a custom function, `model_pred`, to each nested data frame. This function leverages various regression models—ranging from basic models with temporal controls to more complex ones incorporating weather conditions, time lags, and holiday effects. Each model was tasked with predicting bike share trip counts based on the features it was designed to analyze.

After the predictions were made, we reshaped the data from wide to long format using the `gather()` function. This transformation pooled all model predictions into a single column, simplifying subsequent analyses and comparisons of the models' effectiveness.

Finally, we calculated error metrics to assess each model's accuracy. We computed absolute errors between observed and predicted values, along with the Mean Absolute Error (MAE) and the standard deviation of these errors, using `map2()` and `map_dbl()`. These metrics are crucial for evaluating the precision and reliability of our predictive models across different weeks.

This structured approach not only enhances our understanding of each model's predictive capabilities but also aligns with sophisticated data science practices that effectively accommodate complex and large-scale data sets.

Show

5 Result

5.1 Examine Error Metrics for Accuracy

The most effective models, particularly those including time lags, show an accuracy of less than one ride difference per hour on average. This level of precision is quite satisfactory, indicating that these models can reliably predict bike share demand.

However, Figure 5.1.1 indicates that the holiday time lags may appear to have minimal impact due to their infrequency. Since holidays do not occur often, the models might not have enough data to effectively learn from these occasions. Additionally, the overall effect of holidays could be overshadowed by more dominant daily or seasonal trends affecting bike usage (see Figure 5.1.2).

Show

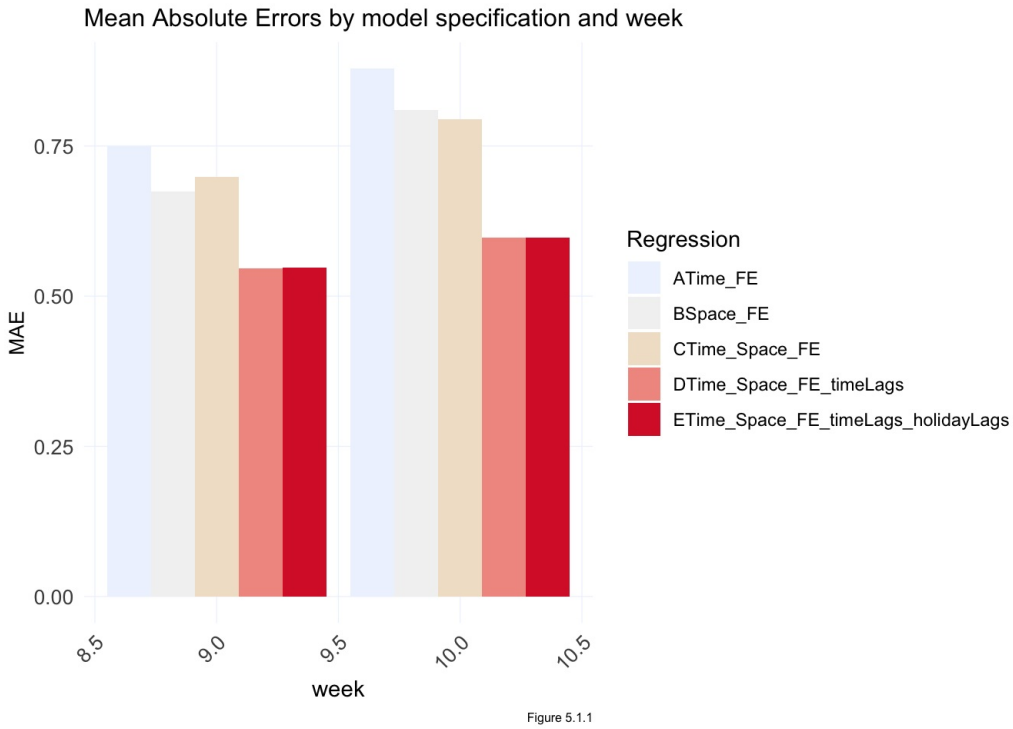

Figure 5.1.1

Show

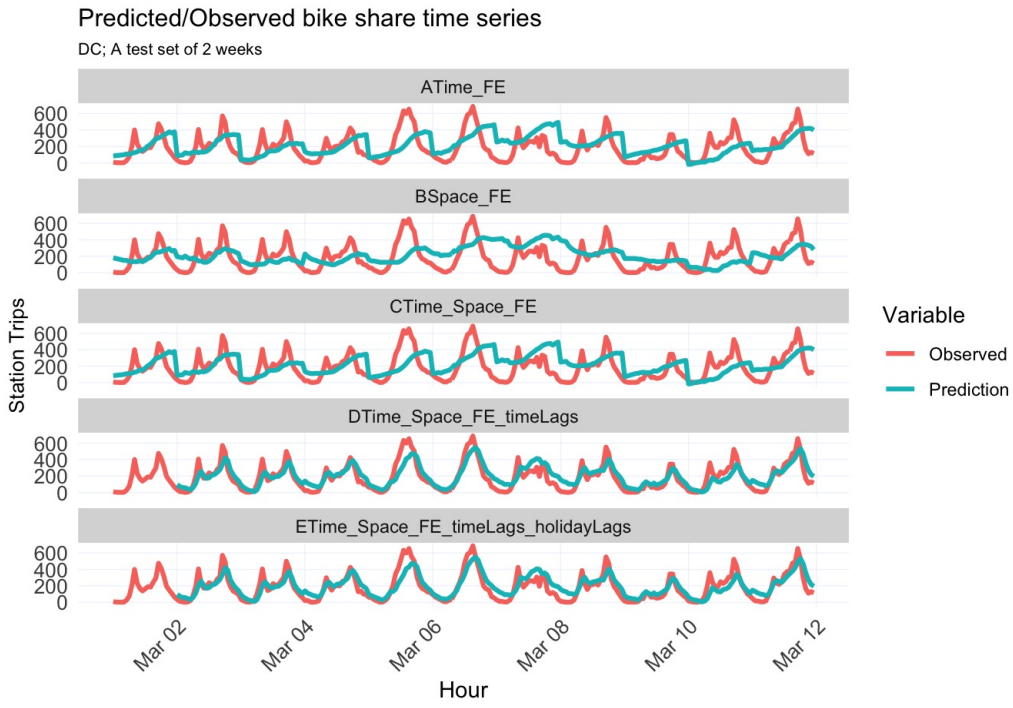

Figure 5.1.2

The errors displayed across different stations suggest that certain areas consistently experience overpredictions or underpredictions. This could be influenced by unique local factors at each station, such as its proximity to popular destinations or its accessibility, which may not be fully accounted for in the models.

Visualizing errors geographically (see Figure 5.1.3) helps pinpoint specific stations where predictions are less accurate. This approach is particularly useful for identifying potential improvements in the bike share network, such as adjusting bike availability or updating station facilities to better meet actual user demand.

Show

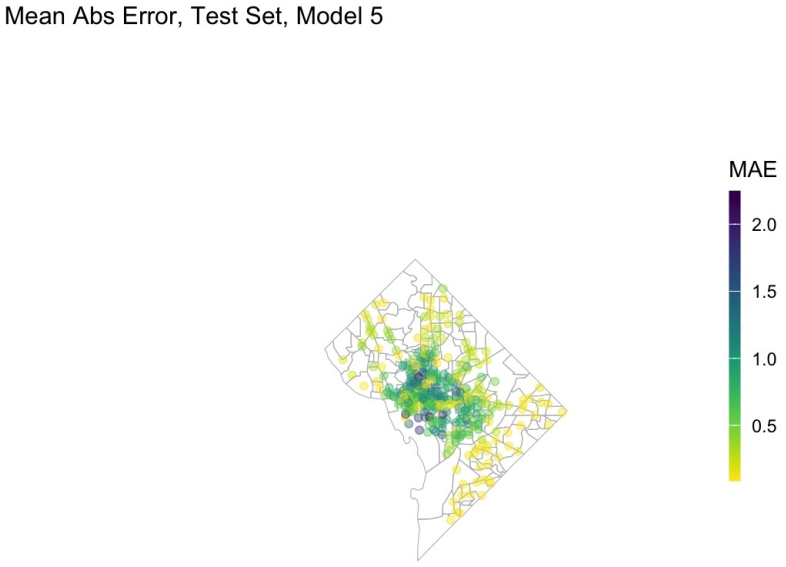

Figure 5.1.3

5.2 Space-Time Error Evaluation

In general, Figure 5.2.1 shows that our model tends to underpredict the reality of bike share trips in Washington, DC. While this is still acceptable, we have also noticed from the plot of observed vs. predicted trips that there are differences which might be influenced by factors that cannot be explained by our model, such as:

1. Time of Day Variability: By plotting observed versus predicted trips for various times of the day, you likely notice that prediction errors vary significantly based on the time. For instance, higher errors during peak rush hours might suggest that the model struggles with sudden spikes in usage.
2. Weekend and Weekday Differences: The model may also exhibit different error patterns on weekends compared to weekdays. During weekends, especially in recreational areas like along water bodies or in central business districts during weekdays, the prediction errors might increase. This could indicate that the model underestimates leisure-related trips or overestimates normal weekday commuting patterns.
3. Unexplained Variability by the Model: The patterns you mentioned—specifically underprediction—suggest that there are additional factors influencing bike share usage that the model hasn't captured. This could be unrecorded events, sudden weather changes, or other non-modeled variables like local traffic conditions or temporary road closures.

Show

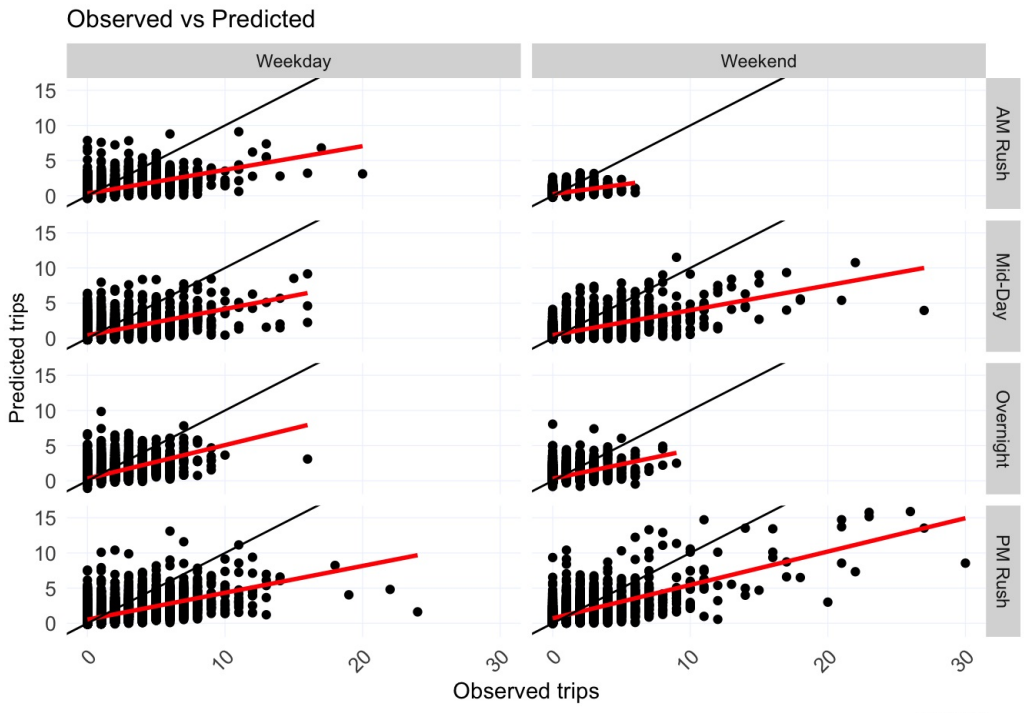

Figure 5.2.1

Visualizing the errors across time of the day and weekday/weekend across geography, as shown by Figure 5.2.2, helps us to understand where our models struggle the most. The observation that errors cluster along the downtown area happen almost every time of the day during the weekday (except overnight time) and strikingly intensed during the mid-day and evening over the weekend, points towards:

1. Recreational vs. commuter trips: differentiating model strategies between commuter and recreational usage can help reduce errors. Recreational trips might require different modeling considerations, like seasonal effects.
2. Geographical focus for model improvements: areas with consistently high errors (e.g., downtown areas) should be the focus for additional data collection and model refinement, perhaps by incorporating more localized or dynamic factors to the model.

Thus, to deal with this issue, strategies like re balancing this error is significant (particularly as they are concentrated in high-volume locations), implying that:

1. High volume locations: errors in high traffic areas can lead to significant impacts on bike availability. If the model consistently under predicts demand in these areas, stations might run out of bikes, leading to customer dissatisfaction.
2. Rebalancing strategies: understanding where and when these errors occur can help in planning more active rebalancing actions, such as scheduling more frequent transfers of bikes to busy stations during predicted peak times, or ensuring that extra capacity is available during weekends in recreational areas.

Show

Figure 5.2.2

Aside from the environmental factors, socio-economic factors are also important to be considered for our bike share rebalancing strategies. Figure 5.2.2 shows how different how different socio-economic factors such as median income, percent of people taking public transportation, and percent of white population might affect the bike share trips in Washington, DC. The following plots show how these socio-economic factors correlate to morning rush hour trips, particularly in downtown areas with high accessibility to public transportations:

1. Median income: there does not seem to be a strong link between income levels and the errors our model is making. This suggests that whether a station is in a high or low income area does not consistently change how well the model predicts bike usage
2. Public transit usage: interestingly, a model seems to stumble a bit more in areas where fewer people are using public transit. Perhaps the model is less equipped to handle bike share demand in places where residents have more transportation options, possibly opting for bikes over buses or trains less predictability.
3. Percentage of White population: as the percentage of White residents goes up, so do the errors. This could mean that our model might not be capturing all the factors that affect bike share use in these communities.

Show

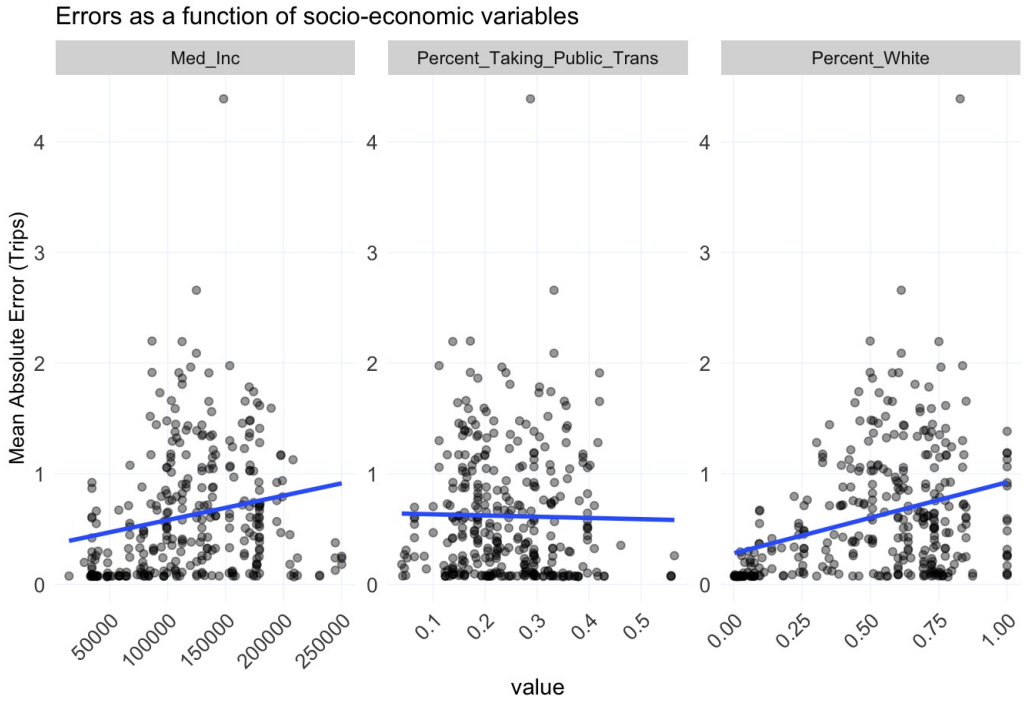

Figure 5.2.2

Through our space-time and socio-economic errors evaluation, we see that the implications for rebalancing are quite direct. In high-demand areas, especially during busy commuting hours, being off by a small number bikes can have a big impact. If the errors cluster around key locations–like bustling downtown areas or leisure areas on the weekend–it could disrupt service and user experience. This is even more critical if these locations are already seeing a lot of bike share traffic.

For a more fine-grained analysis, if we zoom in on the weekday morning rush hour, we might notice that stations serving more affluent, less diverse areas with fewer public transit users are tougher for the model to predict accurately. This insight could be crucial for improving the model, suggesting we might need to look closer at how different demographic choose to travel, especially when they are heading into dense urban centers for work.

6 Conclusion

In conclusion, our predictive analysis, tailored specifically to the Capital Bikeshare network in Washington, D.C., demonstrates a substantial capability to forecast regular demand patterns, which is instrumental for the operational objective of effective rebalancing. Yet, the models fall short during critical peak demand periods, essential for Capital Bikeshare's rebalancing efforts. These high-demand peaks exhibit specific spatial and demographic trends within the D.C. metro region that our current models have yet to fully encapsulate.

The challenge of underpredicting at times of high demand can lead to insufficient bikes in the network when users most need them. This situation could undermine the system's reliability, resulting in a decline in ridership and an increase in the expense of rebalancing operations.

Refining our model to better account for the unique attributes of different stations within the bike share system, such as local demographic shifts or specific station traits, is crucial for increasing predictive precision. A deeper dive into the characteristics of stations that consistently yield prediction errors may reveal additional predictive variables or point towards new, focused rebalancing approaches.

The foundation laid by our algorithm is solid, yet it requires ongoing enhancements to serve as the backbone of an efficient rebalancing strategy for Capital Bikeshare. By zeroing in on the specific conditions and locations where the algorithm underperforms, we can employ these insights practically, ensuring that the Capital Bikeshare network is reliably stocked to navigate the ebb and flow of bustling urban life in D.C.

---

1. DDOT DC (https://ddot.dc.gov/page/capital-bikeshare)↩
2. DDOT DC Expands Access to $5 Annual Memberships (https://ddot.dc.gov/release/capital-bikeshare-expands-access-5-annual-memberships)↩
3. DDOT Capital Projects (https://projects.ddot.dc.gov/datasets/capital-bikeshare)↩
4. Capital Bike Share Trip History Data (https://s3.amazonaws.com/capitalbikeshare-data/index.html)↩
5. Statista (https://www.statista.com/statistics/1457901/monthly-bike-share-trips-capital-wheels-washington-dc-us/)↩
6. Washington Post (https://www.washingtonpost.com/dc-md-va/2022/05/14/dc-bans-off-our-bodies-protest/)↩