Bayesian Bikeshare Demand Models: A Case Study of Seoul's Bikeshare system using Poisson and Negative Binomial Count models.

Amaia Rodriguez-Sierra, Peijin Chen, Maria Sckolnick

University of Vermont, Burlington, VT USA

Abstract.

As urban areas around the globe deal with the challenges of traffic congestion and pollution, bicycle sharing systems (BSS) have emerged as a sustainable and commercially viable alternative to traditional private or public transport. The popularity of the BSS in cities around the world has generated sizable data about their own usage and the underlying transportation dynamics of a city. In order to be more responsive to bike sharing demand, BSS companies have to anticipate demand, and this paper focuses on developing Bayesian regression and forecasting models for predicting bike rental demand prediction on an hourly basis. The goal of this paper is to do proof of concept study to assess whether bike rental demand in a large city can be probabilistically forecasted using a combination of meteorological conditions and prior usage data. This project uses publicly available data. The dataset utilized in this project consists of hourly bike rental counts in Seoul from December 2017 through November 2018. We experiment and compare a Poisson with a Negative Binomial model, often used for modeling count data, and see how these models perform on holdout data.

Introduction.

As urbanization increases around the world, especially in the developing world, urban transportation becomes an increasingly complex problem. The sheer level and scale of urban agglomeration creates problems of congestion, which leads to decreased economic efficiency, with some studies estimating that up to 1% of GDP is lost in areas of excessive congestion. Urban cycling is a sustainable, non-polluting form of transportation, but there are inconveniences that arise with urban bike ownership, from not having an infrastructure that is friendly to bike transport (buses that have bike racks, subways that allow bicycles to be transported, office space that accommodates bike parking) to the issues of bicycle theft and maintenance. The birth of BSS, in programs such as Vélib' in Paris or Citibike in New York City, was designed to alleviate such problems. A BSS will typically have a large number of rental stations positioned at various points in a city, where users can (either through membership plans or one-off payments), rent a bicycle. After paying and unlocking use of the bicycle, the user must return the bicycle, usually within a certain amount of time, to a station.

In large cities, BSS have become popular. There are sometimes not enough bikes to meet the demand under particular times and conditions (such as during morning or evening rush hours). Public bicycles must also be maintained and repaired, so to ensure that there are enough bicycles to meet demand, various mathematical models have been employed to forecast overall hourly bicycle demand.

In Seoul, South Korea, the bike sharing system Ddareungi (Korean: 따릉이) - Seoul Bike in English - was first introduced in Seoul in October 2015 in select areas of the right bank of the Han River. After a few months, the number of stations reached 150 and 1500 bikes were made available. A publicly available data set of aggregate hourly demand from December

2017-November 2018 (about one years' worth of time) is available and is the dataset we use for creating and testing our models.

Previous Work.

Smith et al. (2011) modeled bicycle usage (not necessarily bike sharing, but personal bikes) in Melbourne, Australia and found that there were strong seasonal trends as well as hourly trends (commute times, rush hours, etc). Gebhart et al. (2014) studied bike share counts in Washington, DC and found that cold temperatures, humidity, and rainfall affected the likelihood of bike shares being used. Sathishkumar et al. (2020) developed a rule based system for forecasting the bike share counts in Seoul (using the same dataset), and found that temperature and hour were the most important features across the several predictive models they developed. Leem et al. (2023) used a two-stage online learning based time series model to predict bike share demand (using the same dataset), and also found that hour and temperature were the most important features (though feature importance is measured differently, depending on the algorithm used). After that, they found that humidity to be the third most important feature in two out of their three test models, with humidity being the third most important feature in the remaining model. Similarly, El-Assi et al. (2017) found, in their study of Toronto, Canada's BSS, that hourly patterns varied depending on the season and temperature, as short-term or one-off usage dropped, but long-term members still tended to continue using the bikes, even as the weather got colder.

Table 1: Variable definitions

Variable	name	Type	Definition
Hour	Hour	datetime	year-month-day hour:minute:second
Rented Bike count	BikeCount	numeric	Count of bikes rented at each hour
Temperature	Temperature	numeric	Temperature in Celsius
Humidity	Humidity	numeric	% humidity
Windspeed	WindSpeed	numeric	meters/second
Visibility	Visibility	numeric	in 10m
Dew point temperature	Dewpoint	numeric	Celsius
Solar radiation	SolarRadtion	numeric	MJ/m2
Rainfall	Rainfall	numeric	mm
Snowfall	Snowfall	numeric	cm
Seasons	Seasons	categorical	Winter, Spring, Summer, Autumn
Holiday	Holiday	categorical	Holiday/No holiday
Functional Day	FunctionalDay	categorical	NoFunc(Non Functional Hours), Fun(Functional hours)
Workday	Workday	categorical	A workday is a weekday that is not a holiday.

To our knowledge, there is not much work on modeling this dataset using Bayesian methods, such as Bayesian General Linear Models. We hope that this probabilistic forecasting model can not only help estimate the number of bikes that will be rented, but will help understand the variance and uncertainty inherent in such estimates.

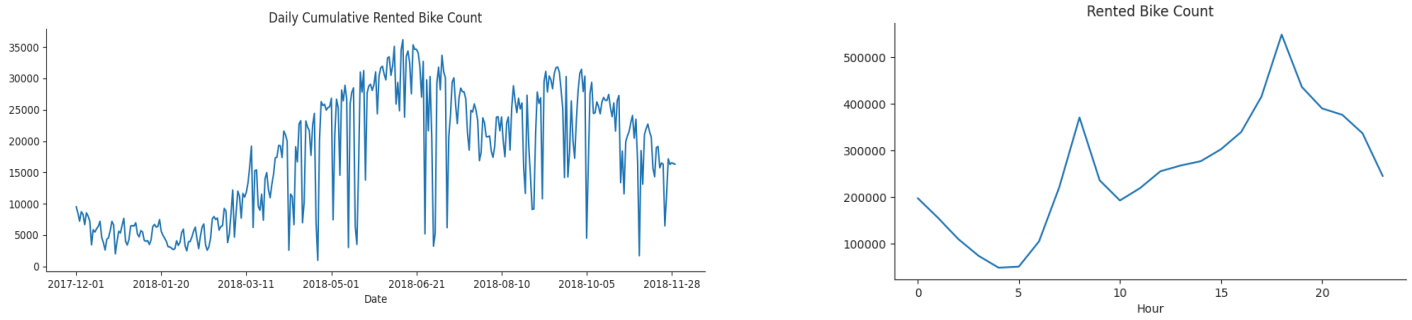

Figure 1. Rental Bike Counts over a period of months (Left) and Mean Bike Rental Counts by Hour (Right)

Data (Exploratory Data Analysis)

The data is publicly available at the UCI Machine Learning Repository. Table 1 lists the variables that are included in the dataset. After removing data from days when the bike rental places were not open for business (coded as functioning day = 0 in the data), bike counts over time follow the trend pictured in Figure 1. We can see that there are clear seasonal trends, with the late spring and summer being the times of year that experience the highest level of demand while bike counts are low during the winter months. Two things stand out when examining the trends in the data set: (1) There is a dip in counts around August 2018, that is inconsistent with the overall trend. (2) There are several spikes in the data corresponding to days with counts much lower compared to the average bike counts around those events. These low count dates are not associated with a specific weekday or holidays. Furthermore, we were not able to find an association of the August dip or the dates with the abnormally low bike counts with any other variables in the data set or any other reported events in Seoul that might explain these data features. We expect that these features will unfortunately affect our signal-to-noise ratio negatively. Figure 1 (Right) shows cumulative bike counts by time of day. Bike demand throughout the day is characterized by clear intra-day trends, with spikes during the morning and evening commute hours.

Methods

Data were modeled using Bayesian methods for estimating the outcome as a General Linear Model (GLM). The most parsimonious way to model count data is to assume a Poisson distribution of the outcome. A poisson distribution assumes that the variance of the data is equal to the mean. To include information from the predictors into the model we assume that the mean varies as a function of the predictors in the following way:

$$Y_i = Poisson\,(\lambda_i))$$
$$\lambda_i = exp(X_{ij}\beta_j + \alpha)$$
$$\beta_j \sim Normal(0, 0.2)$$
$$\alpha \sim Normal(6, 0.2)$$

If the assumption that mean is equal to variance does not hold, i.e. if the data are either overdispersed or underdispersed a negative binomial distribution will provide a better fit, allowing the variance to be modeled separately through the inclusion of an additional scaling parameter, as detailed below.

$$Y_i = Negative\,Binomial\,(\lambda_i), \varphi)$$

$$\varphi \sim Lognormal(0, 0.2)$$

$$\lambda_i = exp(X_{ij}\beta_j + \alpha)$$

$$\beta_j \sim Normal(0, 0.2)$$

$$\alpha \sim Normal(6, 0.2)$$

Data preprocessing included removal of data from dates when "Functioning Day" was equal to 0. Those dates showed zero bike rental counts and we assumed that those represented days where the bike rental service either did not operate or counting was omitted due to some system error. Weekday was extracted from date and encoded into a binary indicator for weekend. Precipitation values (snowfall and rainfall) were sparse and model convergence issues prompted us to include these data as a binary variable indicating whether there was any precipitation (snow or rain) at a given time. The hour variable was grouped into night (hours 0-5), morning (hours 6-11), afternoon (hours 12-17) and evening (hours 18 - 23) and encoded one-hot with afternoon as the baseline. Season was also encoded one-hot with Winter as the baseline. All other categorical variables were binary already and included as provided in the data set. Our final data set that we included in the model contained 15 predictor variables.

To ensure reasonable modeling time, 2000 randomly selected data points were chosen as the training set. An additional 200 data points were randomly chosen and used as the testing data set. Models were estimated in Stan with 4 chains and 8000 warmups/iterations. Model convergence was assessed by checking trajectories for divergences. Rhat values were all close to 1.0.

Results

Posterior Predictive Checks

We performed posterior predictive checks. Figure 2 compares simulations of outcomes using our prior distributions to the observed outcome distribution. Our prior choices give reasonable predictions for both the Poisson as well as the negative binomial model.

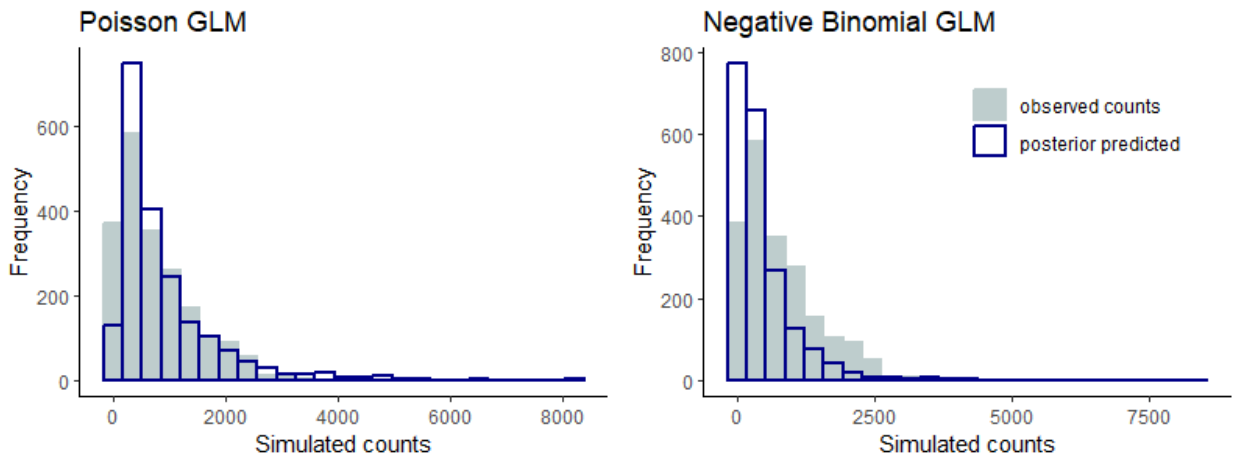

Figure 2: Posterior predictive checks: Histograms of the observed bike counts compared to a histogram of one simulation of outcome with our choice of prior distributions.

Model results

For the Poisson Model, approximately 60% of the replications were underestimates of the corresponding observed value in the training set. The Mean Absolute Error (MAE) of the replications (simulated values) on the training set was approximately 526. By way of contrast, a Random Forest Regression Model (RF), achieved a MAE of 206 on the training set and MAE of 494 on the testing set.

PM and NBM results were qualitatively very similar. Graphs below will only present results for the NBM which was a better fit. Numerical results in the text are given as NBM (and PM in parentheses).

For the Negative Binomial model, Figure 3 shows simulations against observed values. In the train data some of the peaks and troughs are captured in the estimations but there are also many instances of overestimation. Estimates for the train data set do not appear to reflect the true values very accurately. These results suggest that the model does not predict bike count variability very accurately and additional predictors will have to be considered to make the model more useful.

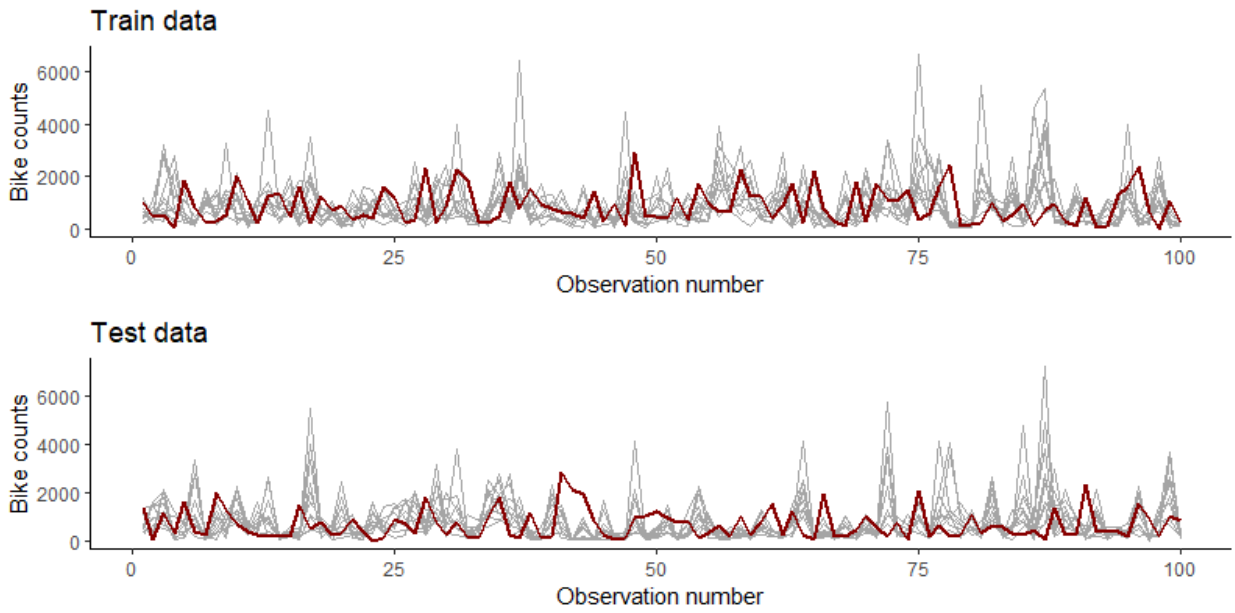

Figure 3: 10 estimates (grey) and corresponding observed values (red) of bike counts for 100 data points each from the training and testing data set.

Posterior distributions of parameter estimates.

Figure 4 shows histograms of a subset of posterior distributions of a subset of all parameters to illustrate the shapes of their distributions. Odds ratios and plausible intervals for all parameters are listed in Table 3 for the PM and in Table 4 for the NBM.

The strongest predictors for large bike counts were temperature and dew point, increasing the probability of observing a bike rental by 51% (50%) and 43% (32%) respectively. The strongest predictors in descending order for observing reduced bike counts were night (NBM 53%, PM 63% reduction), precipitation(NBM 48%, PM 62% reduction). Summer (NBM 31%, PM 27% reduction), humidity (NBM 26%, PM 23% reduction) and holiday (NBM 25%, PM 22% reduction) had only a moderate to small effect on bike counts.

The effect of summer as having a negative impact on bike counts is particularly interesting since data trends overall suggest that bike counts peak in summer. Posterior estimates of outcome by season and time of day confirm this as shown in Figure 5 for both our training and our testing data set. Likely, this general effect of bike counts being high in summer is due to confounding effects from other variables. Summer tends to coincide with pleasant temperatures and less precipitation. Since the model accounts for these other measures directly the residual effect of summer on bike count will capture only those effects that are not weather related. This residual summer effect appears to be negative in our model.

	Poisson Model		
Regression Variable	Odds Ratio for Mean	95% plausible interval	
		Lower bound	Upper bound
Temperature	1.50	1.48	1.52
DewPoint	1.32	1.30	1.35
Evening	1.04	1.03	1.04
Visibility	1.04	1.03	1.04
Spring	1.01	1.01	1.02
Wind Speed	1.01	1.01	1.01
Fall	1.00	0.37	2.65
Weekend	0.89	0.89	0.89
Solar Radiation	0.86	0.85	0.86
Morning	0.84	0.83	0.84
Holiday	0.78	0.78	0.79
Humidity	0.77	0.77	0.78
Summer	0.73	0.73	0.73
Precipitation	0.38	0.37	0.38
Night	0.37	0.36	0.37

Table 3: Odds ratio estimates of bike counts based on the regression parameters from the PM and 95% plausible interval of the odds ratio for each parameter.

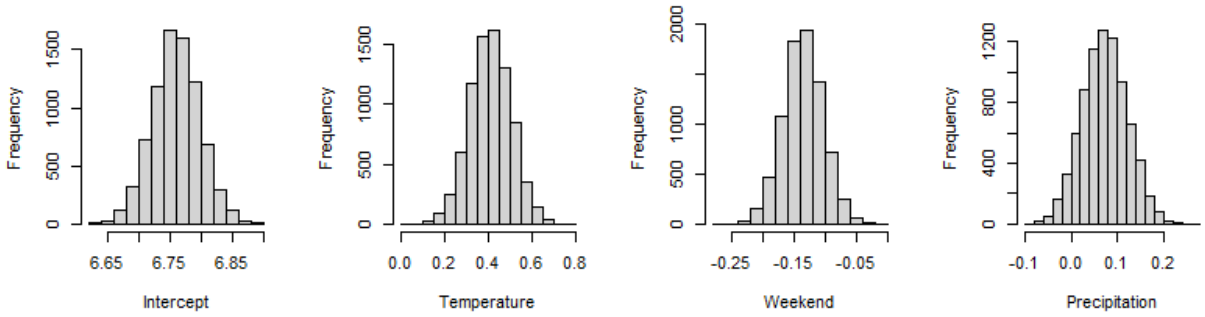

Figure 4: Posterior distributions of a subset of four parameter estimates.

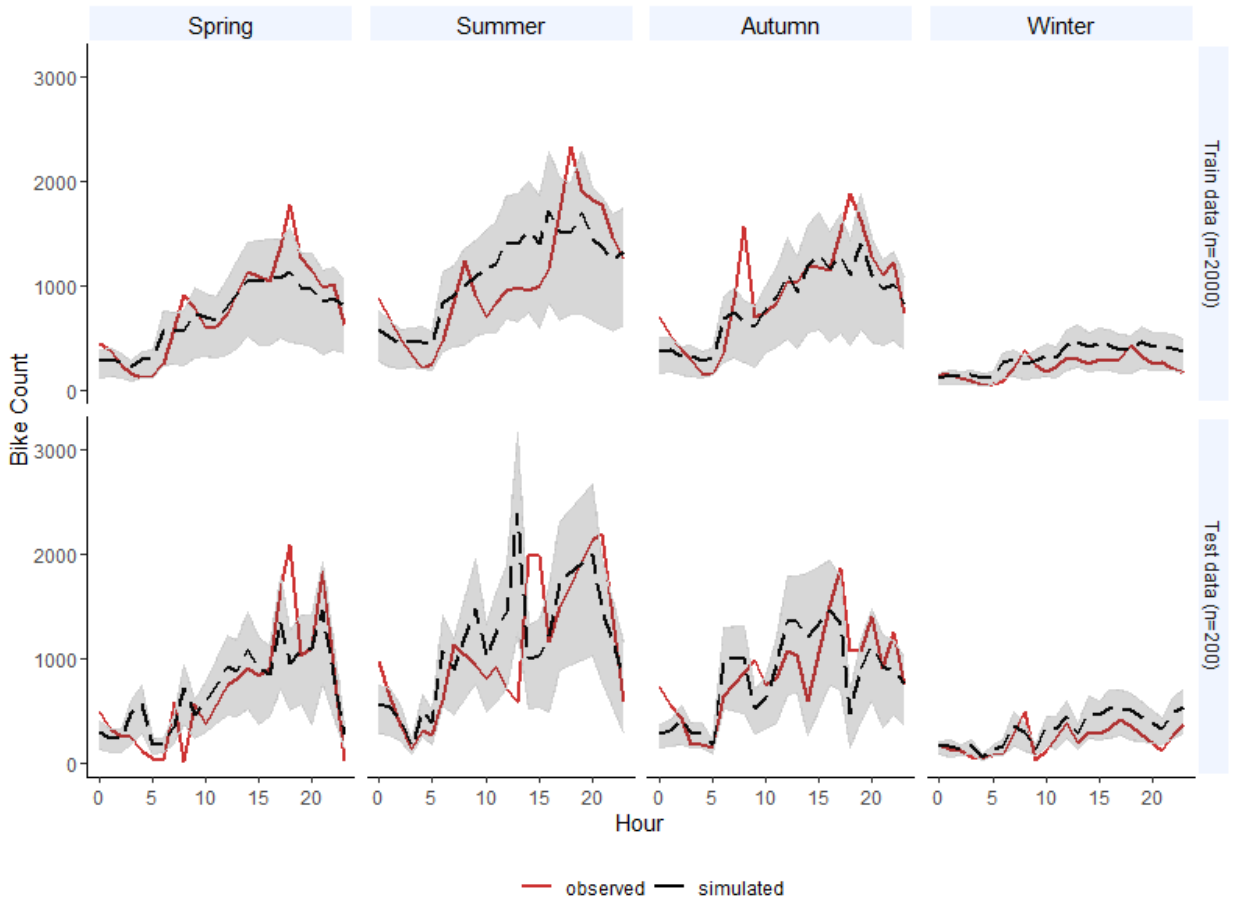

Figure 5: Comparison of observed and simulated bike counts at different times of day across seasons for test and train data. The grey areas are 95% plausible intervals around the predicted mean counts.

	Negative Binomial model		
Regression Variable	Odds Ratio for Mean	95% plausible interval	
		Lower bound	Upper bound
Dew Point	1.43	1.15	1.77
Temperature	1.51	1.25	1.81
Fall	1.00	0.68	1.48
Evening	1.08	0.98	1.18
Visibility	1.03	1.00	1.07
Wind Speed	0.99	0.96	1.03
Morning	0.96	0.88	1.05
Spring	0.92	0.86	0.99
Solar Radiation	0.91	0.87	0.95
Weekend	0.87	0.82	0.93
Holiday	0.75	0.66	0.86
Summer	0.69	0.63	0.77
Humidity	0.74	0.67	0.82
Precipitation	0.52	0.47	0.58
Night	0.43	0.39	0.48

Table 4: Odds ratio estimates of bike counts based on the regression parameters from the NBMand 95% plausible interval of the odds ratio for each parameter.

Discussion

The models, particularly the Negative Binomial (NBM), exhibited limitations in accurately predicting bike counts. The poor fit might be because the way we grouped time of day did not fully capture the detailed trends throughout the day. Future models should consider refining this categorization to better reflect the actual demand patterns throughout the day.

The model's performance was affected by outlier spikes and irregularities in the data, significantly impacting its accuracy. Understanding and addressing these anomalies could substantially enhance the model's predictive capabilities. A deeper exploration of these spikes and data noise could refine the predictive power of the models.

Additionally, expanding the dataset to include a broader time frame might significantly improve model performance. The current dataset spans only a single year, and the difference between the data of November 2017 and November 2018 might indicate ongoing program developments. Collecting data across multiple years could offer a more comprehensive understanding of the system's evolution and demand patterns.

Comparative analyses with data from other cities could provide valuable insights. Assessing bike rental demand across various urban settings could uncover common trends or unique factors influencing demand, thereby refining the models and offering broader applicability.

In conclusion, addressing categorization refinement, outlier analysis, dataset expansion, and cross-city comparisons can substantially enhance the accuracy and reliability of predictive models for bike rental demand.

Conclusion

In this study, Bayesian regression models, specifically the Poisson General Linear Model (GLM) and Negative Binomial Model (NBM), were explored to predict bike rental demand in Seoul's bike.sharing system. Utilizing meteorological conditions and temporal features, our analysis aimed to provide insights into demand forecasting. However, while these models showed promise, several challenges and opportunities for improvement surfaced.

Key Findings

- Clear seasonal trends were evident in bike rental demand, notably peaking during late spring and summer while diminishing in winter.
- Anomalies such as irregular drops in counts and sporadic low counts posed challenges for accurate demand prediction.
- Significant predictors like temperature and dew point exerted varying impacts on bike rental coins, although certain variables presented residual effects, contradicting overall trends.

Potential Pitfalls

- Outlier spikes and irregularities in the data affected model accuracy, potentially undermining precise predictions.
- Insufficient capturing of detailed intra-day trends restricted the models' ability to reflect detailed time-of-day variations accurately.
- Certain variables, like summer, exhibited unexpected influences, indicating underlying factors not entirely encompassed by the models.

Future Directions

- Investigate outlier spikes and data noise to refine model predictions and enhance accuracy.
- Refine time of day categorization to better capture described trends and intra-day variations.
- Explore more comprehensive dataset spanning multiple years and compare data across cities for broader insights and context.

By addressing these identified limitations and pursuing these future research directions, we aim to refine the Bayesian modeling framework for predicting bike rental demand. This endeavor will contribute to more accurate predictions and deeper insights into the dynamics of urban bike-sharing systems, enabling more effective planning and management strategies.

References

Etienne Come, Latifa Oukhellou. Model-based count series clustering for Bike Sharing System usage mining, a case study with the Velib system of Paris. ACM Transactions on Intelligent Systems and Technology, 2014, 27p. ffhal-01052621

Leem, S., Oh, J., Moon, J. et al. Enhancing multistep-ahead bike-sharing demand prediction with a two-stage online learning-based time-series model: insight from Seoul. J Supercomput (2023). https://doi.org/10.1007/s11227-023-05593-6

Gebhart, K., Noland, R.B. The impact of weather conditions on bikeshare trips in Washington, DC. Transportation 41, 1205–1225 (2014). https://doi.org/10.1007/s11116-014-9540-7

VE, Sathishkumar, and Yongyun Cho. "A rule-based model for Seoul Bike sharing demand prediction using weather data." European Journal of Remote Sensing 53, no. sup1 (2020): 166-183.

El-Assi, Wafic, Mohamed Salah Mahmoud, and Khandker Nurul Habib. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. Transportation 44 (2017): 589-613.