Washington D.C. Capital Bikesharing Case Study Report

Name: Marcel Gumulak

Universität Bielefeld

314008 Statistisches Praktikum (Pr) (WiSe 2024/2025)

Modul 31-SW-StiP Statistik in der Praxis

Modulverantwortlicher: Prof. Dr. Dietmar Bauer

November 26, 2024

Contents

1 Executive Summary

This report evaluates the factors influencing the rental counts of bikes in Washington D.C.'s Capital Bikeshare system, with a specific focus on weather-related variables and station-specific dynamics. More specifically, this study examines the extent to which the number of bikes rented is linked to the daily weather conditions and which of these factors contribute the most to bike rental. Data from 100 different bike stations were collected over 334 days in 2022 to investigate this relationship. The purpose of the analysis is to identify actionable insights that can be employed to optimise operational processes, whose insights are obtained through predictive modelling and the identification of hot zones within Washington D.C.. The key findings of the analysis are summarised in the following points:

Weather Impact The mean temperature was identified as the primary factor influencing bike rentals. The data indicated a clear and strong relationship between an increase in bike rental counts and mean temperatures, which exhibited a general U-shaped relationship. It is essential to note that this has significant implications for the operational procedures during target periods, as there is a discernible downwards trend in rental counts when mean temperatures exceed a threshold value of 76°F, with the effect intensifying as mean temperatures rise. This phenomenon occurs consistently throughout the year and is not influenced by the underlying station. Conversely, precipitation and wind speeds demonstrated a significant negative effect, although this effect was secondary to that of the primary factor. To illustrate, the predicted mean bike rental count on an average May day at the 10th & K St NW station and in the absence of any precipitation amounts to 39 rentals, based on the model presented here. In contrast, the predicted mean count in the event of heavy rain only reduces to 34 predicted mean rentals. Therefore, their impact is only moderate, contributing more as supplementary information to temperatures. Furthermore, the occurrence of snowfall and resulting snow cover also have a significant negative effect on bike rental counts, particularly during extreme conditions.

Predictive Model A log-normal panel data model was employed to quantify the relationship between weather variables and bike rentals. This approach offers greater flexibility in modelling and provides access to a wide range of statistical test statistics that have been specifically developed for panel data. The results of the finalised model reiterate the preceding assertion, underscoring the pivotal role of weather conditions in influencing bike rental activity. The model exhibits a high degree of explanatory power, with a total of 59.2% of the variance in the dependent variable explained by its explanatory variables and the fixed effects structure. Moreover, the model allows for the projection of future mean bike rental counts. While 59.2% is notably high for the limited number of explanatory variables employed in the model, it is important to recognise that the remaining 41.8% of model uncertainty will have an impact on future predictions. To provide a more detailed explanation, the model uncertainty is related to the unexplained variance in the observed bike rental counts by our model, and this increases the width of the interval considerably when calculating a 95% out-of-sample prediction interval. This indicates that the predicted mean bike rental count for an average August day at the 10th & K St NW station is 45 bikes, with the interval width encompassing a range of 61 units (between -22 and +39 units around the predicted mean bike count of 45). Thus, it is evident that the model's accuracy could be enhanced by reducing the associated uncertainty, which could be achieved by incorporating additional sources of information that explain the remaining unexplained variance. However, at this stage, it is unclear whether and by how much this additional information will assist with the prognosis, in contrast to the cost

associated with obtaining this greater quantity of information.

Hot Zones The results of the model are robust and may be utilised to derive a comparative quantity for bike station activity that is corrected from monthly and weather-related effects. These figures may be employed as a basis for activity ranking, station grouping and hot zone calculation, and the interpretation of these figures is that the higher the number, the higher the demand at that specific station. An overview of the top and bottom five stations in terms of activity is accessible in Table 6, and a comprehensive map of the existing hot zones within Washington D.C. can be accessed via the online webpage.

Recommendations for Management To enhance operational efficiency, weather-responsive strategies should be adopted by means of model-based predictions, such as dynamically reallocating resources based on weather forecasts, reducing operations during days of heavy rain or snow, and focusing on high-demand conditions like moderate temperatures. For station performance optimization, one should prioritize investments in high-demand locations like Lincoln Memorial and New Hampshire Ave & T St NW, adding capacity to meet peak usage. The activity ranking and hot zones can be utilised to tailor promotional efforts to underperforming neighbourhoods, fostering broader engagement and increased bike usage. Lastly, insights from the model can support strategic expansion efforts by identifying neighbourhoods with untapped potential, leading to new station locations that are strategically positioned in the city to maximize impact and broaden the network's reach.

2 Exploratory Data Analysis

Capital Bikeshare is a public bike-sharing service operating across the Washington D.C. metropolitan area. It provides a sustainable and affordable transportation alternative by offering over 6.000 bikes and 700 docking stations distributed across seven jurisdictions that are used for commuting purposes year-round. In order to conduct this study, a dataset comprising a subset of 100 selected stations was utilised, with a particular focus on those located in the central and inner-city areas of Washington D.C. that are strategically positioned to allow for the analysis of areas of high demand. Consequently, these stations represent some of the most significant nodes in the Capital Bikeshare network and reflect the core dynamics of its operations. The initial stage of the analysis is concerned with an examination of exploratory data analysis, and is therefore divided into two distinct phases: data preparation and data checking. To be more precise, we will initially explore and outline the principal characteristics of the panel dataset and its measurement variables. This stage of the process is largely concerned with the handling of missing values on the time dimension and the variables themselves. Subsequently, an in-depth inspection of each specific variable contained in the dataset is conducted. This enables the formulation of preliminary hypotheses regarding the variables that may have an impact on bike rental, the role of bike stations and months, and the extent to which these differ from each other. Following this procedure will enable us to ascertain which variables should be considered for modelling purposes and what properties a suitable candidate model should possess, thus allowing us to model the number of bikes rented for each day per station accurately.

2.1 Data structure & Missing values

The Capital Bikeshare dataset employed for the analysis is structured as a panel dataset, comprising 100 different bicycle stations observed over a period of 334 consecutive days between

1st January and 30th November 2022, thus having stations on the individual and days on time dimension. With respect to the individual dimension, the dataset comprises station locations organised by street intersections and presented in the form of station names, which typically consist of a numbered street and directional alignment, e.g. "1st & K St NE". Furthermore, the dataset comprises a range of information about each station, as well as a variety of weather data for each day, which will be essential for the subsequent analysis. The dataset includes the number of rented bikes for each station, as well as station-independent measurements of wind speed, precipitation, snowfall, snow depth, and the mean, maximum, and minimum temperature of Washington, D.C. for the day, that were provided by the National Centers for Environmental Information (NCEI). An early assessment of the dataset reveals some inconsistencies, particularly with regard to missing data at a two-level structure and anomalies in certain variables. Firstly, while a substantial proportion of 60% of the stations exhibit a complete range of day observations, a considerable proportion of the remaining 40% are characterised by the absence of data on the temporal level. The extent and severity of these issues are further elucidated in Table 1. It is noteworthy that the missing days occurred only

Table 1: Summary of Observation Completeness for Bikeshare Stations

	Complete	**Missing Days**		
		1-3 Days	**4-8 Days**	**143 Days**
Number of Stations	60	35	4	1

in patterns of consecutive measured days and that no time period was entirely dropped for all stations, indicating that the primary source of information loss occurred in the absence of bike rental count data. This is due to the fact that the collected weather data is measured on a daily basis and therefore still present for at least one other station in the dataset at any given time, thus allowing for potential recovery for their respective variables. Moreover, the station associated with the address 1st & I St SE was the only station that exhibited a significantly higher number of missing observation days, with a total of 143 consecutive days being absent, resulting in the station's first observation being recorded on 24 May, up until the final day. This suggests that the specific station may be either the newest addition to the fleet of bike rental stations or that a larger-scale disturbance may have occurred during the data collection process for that particular station. Secondly, the dataset contains a total of 300 missing values in the bike rental count and 301 missing values in the mean temperature measurement on the variable level. Each station displays a total of approximately three missing values per measurement variable, and these missing values occur at seemingly random time points with no discernible systematic structure. Similarly to the missing days, the absence of data relating solely to the bike rental count could not be recovered. However, the aforementioned characteristics of the weather data permitted the recovery of some observations, whereby the missing mean temperature value could be derived from other available stations, resulting in the recovery of 298 rows. Finally, it was noted that some wind speed measurements displayed irregularities, as they were recorded at negative values. These were promptly rectified in a manner analogous to that employed for the mean temperature values, whereby a valid value from another station on the same day was used as a substitute. In conclusion, each measurement variable has been examined and any issues, such as the inaccurate wind speed data, have been rectified. However, while the missingness in the time dimension of some of the stations renders the dataset an unbalanced panel, it does not threaten the validity of the subsequent analysis and thus does not require explicit addressing. The combination of properties inherent to the panel data structure and the estimation method selected, namely in fixed effects, allows

for the reliable estimation of data points per station, despite the unbalanced nature of the panel. This results in a panel dataset comprising 32.879 rows, which is close to the optimal number of rows estimated at 33.400, and marks the end of the preliminary data preparation phase.

2.2 Data Inspection & Imputation

The objective of this data inspection is to analyse the relationship between the outcome variable, represented by bike rental count per station, and the weather-related variables contained in the provided dataset. As previously stated in Section 2, the primary objective is to examine the potential correlation between these two variables in quantifying the degree of this relationship and determining the anticipated direction of influence. Consequently, this enables us to hypothesise which entities are the primary driving factors. Prior to this, however, we highlight the occurrence of an anomaly in the bike rental counts, which we resolve through an imputation and replacement of the respective count data for that period. Subsequently, each explanatory variable for the counts will be considered in consecutive order, accompanied by a brief summary of the variable's primary findings and illustrative examples highlighted at a specific station. The weather data is independent of the stations, thus allowing for a straightforward examination of the variables for a single station under the full range of day observations. For a more comprehensive analysis across multiple stations, additional figures and statistics, please refer to the "Inspection of data frame variables" Section in the attached code, which is available in the project repository.

August anomaly & Bike rental counts We will begin by examining the observations of bike rental counts from August, as illustrated in Figure 1 for a set of four stations representing the ranks of 1, 36, 76 and 100 in terms of mean bike rental counts over the entire time span. These data points illustrate a rather atypical behaviour in the figure for all stations, as illustrated here by the four aforementioned stations, exhibiting values that are considerably higher than what would be expected based on the preceding and subsequent time periods. These data points, which appear to be artificially inflated, were initially presumed to be associated with increased summer holiday activity, resulting in anomalously high bike rental counts. However, following consultation with the relevant personnel and a more in-depth analysis, it can be inferred that summer holidays are unlikely to be the underlying cause of this phenomenon. Instead, it is more probable that measurement error throughout the entire month of August is the primary factor contributing to the observed overinflation of these count values.

Consequently, the issue was addressed by means of imputation for the entire August count observations, with replacement values being randomly generated based on a regression-based lambda prediction derived from a Poisson process. The model used for these predictions was a similar model to that used in the model formulation Section 3.1, utilising the same set of explanatory variables, where the idea behind this approach was that the final model already provides the most accurate representation of the underlying procedure at hand. The model was fitted to the dataset, with the entire August month removed in an initial step. This was employed to estimate the month-specific fixed effect on the remaining months. These were subsequently employed for a polynomial interpolation of the fixed effects for each August month day, with the resulting effects ultimately being combined with the August day predictions obtained in the initial step. For further details, please refer to the Imputation.R file. The imputation results are robust and are presented in Figure 2, which depicts the same stations as the previous Figure 1, with the red dots now representing the imputed count values.

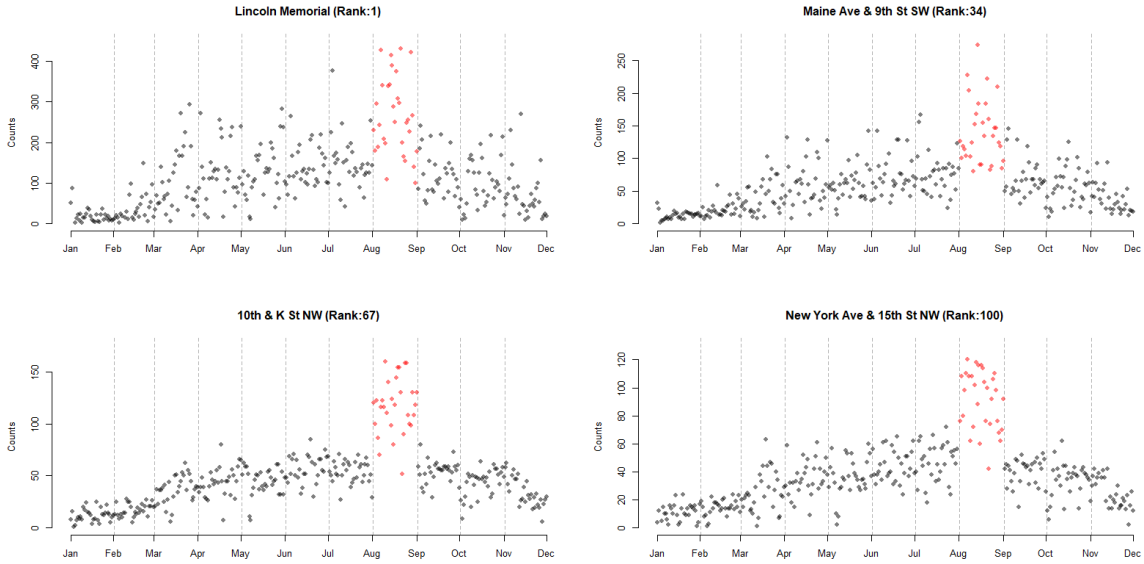


Figure 1: Bike Rental Count over observation time frame, highlighting abnormality in over-inflated August Observation, here emphasized in red.

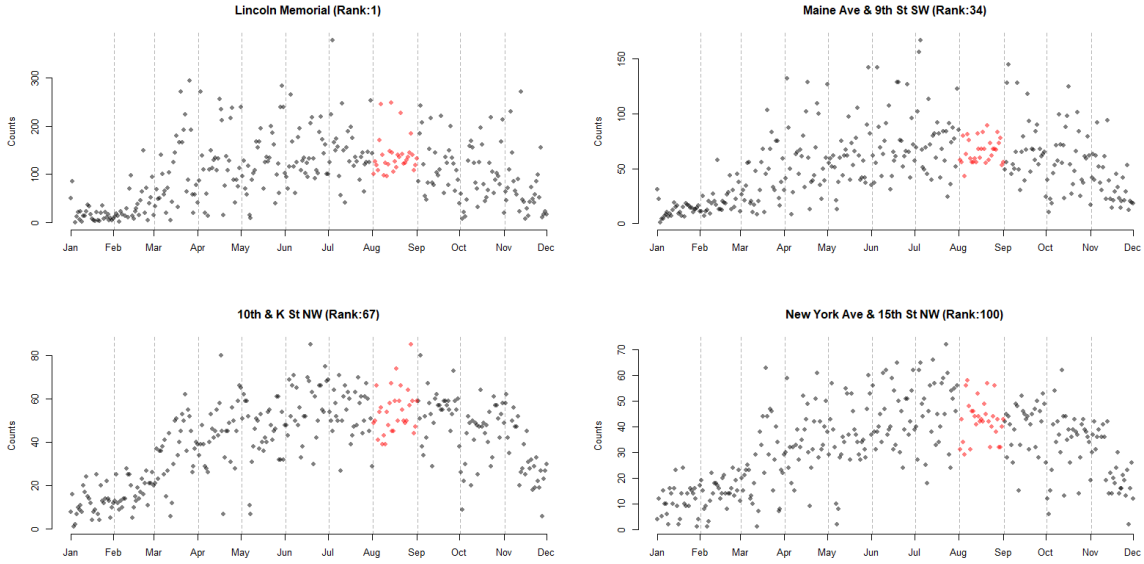


Figure 2: Seasonal Trends in Bike Rental Count after performing imputation of August Observation, here emphasized in red.

Furthermore, there is a noticeable heterogeneity in bike rental counts, which is reflected in the variety of rental stations and comes from the fact that different characteristics affect the rental count in different ways. It therefore seems reasonable to assume that location characteristics will also play one of the most crucial roles in explaining bike rental counts, although these are not present here in the provided dataset. Such characteristics are typically linked to factors such as population density, the socioeconomic profile of the neighbourhood, the presence of nearby landmarks, or the degree of expansion of the local cycling infrastructure. This suggests that a fixed effects approach is a fitting method of capturing these station-specific

effects that are not present in the dataset. Similarly, monthly and weekly seasonality effects are likely to occur, which will be captured later on with the relevant dummies. Therefore, we will proceed with our analysis using a dataset with imputed August count values, rather than the original dataset. Moreover, it can be stated that there are no days on which no bikes are rented at any of the stations in the dataset, and this finding holds true before and after the imputation process, with the minimum number of bike rentals on any given day and station being one. The mean bike rental count over all days for each station, calculated separately and then ranked accordingly, is presented in Table 2. This provides an initial impression without controlling for any explanatory factor. Using the median as the ranking criterion yields a slightly different ordering overall, and the complete list can be found in the "Inspect imputation results" section of the Imputation.R file.

Table 2: Mean and Median Bike Rental Counts for Selected Stations

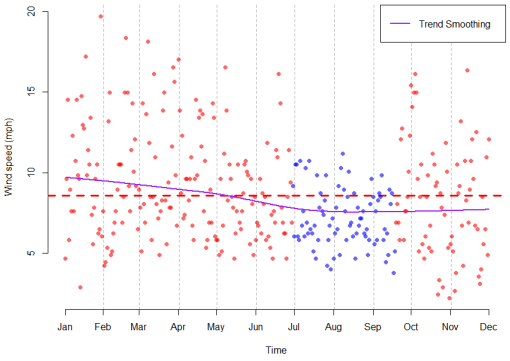
Rank	**Station**	**Mean Count**	**Median Count**
1	Lincoln Memorial	104.0	102
2	Columbus Circle / Union Station	102.0	104
3	New Hampshire Ave & T St NW	101.0	104
4	15th & P St NW	99.2	106
5	Jefferson Dr & 14th St SW	94.2	94
⋮	⋮	⋮	⋮
96	1st & M St SE	34.1	35
97	Rhode Island & Connecticut Ave NW	33.7	35
98	13th & H St NE	33.2	34
99	16th & R St NW	32.7	34
100	New York Ave & 15th St NW	32.4	33

Wind speed The Figure 3a of wind speeds over the entire observation time window demonstrates a markedly disparate behaviour of wind speeds occurring between the beginning of July and the middle of September, where the variance is considerably less prominent than at other periods. The data points highlighted in blue also coincide with a period of exceptionally high mean temperatures, as depicted later in Figure 4a. A slight negative correlation is evident in the attached Figure 3b of wind speed vs. counts for the particular station, as is the case for a majority of stations, indicating that higher wind speeds are associated with reduced bike rental counts. This is additionally also supported by the included kernel density estimate. Furthermore, low wind speeds appear to be associated with increased count variance, and the majority of the blue data points within the aforementioned July to September window coincide with the largest count data occurrences for most stations.

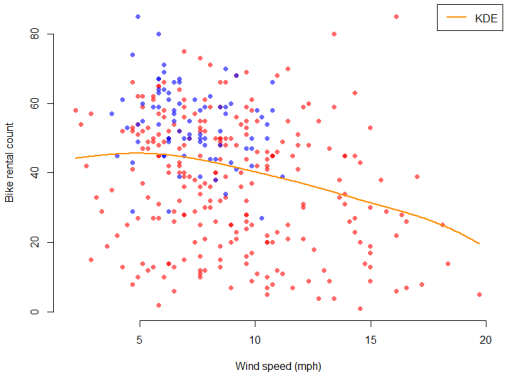
A calculation of the mean bike rental count over the specified period (table 3), in comparison with the figures for the preceding and subsequent periods at the same station, as well as the overall station total, provides further evidence to support the aforementioned claim.

Table 3: Mean Bike Rental Counts by Time Period for Single Station and All Stations

	Mean Bike Rental Count		
Calculated on	**1 Jan. - 30. Jun.**	**1 Jul. - 15 Sep.**	**16 Sep. - 30. Nov.**
In-Plot Station	34.9	77.7	41.8
All Stations	43.8	87.2	52.5

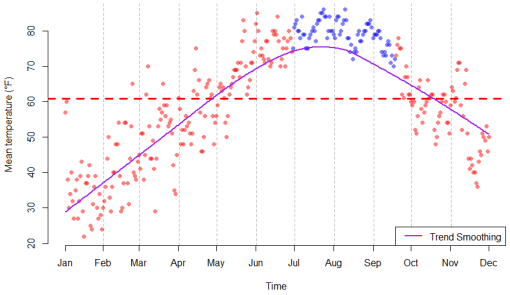


(a) Seasonal Trends in Wind speeds. Blue dots highlight wind speeds for July to September, and the purple line an applied Trend smoothing.

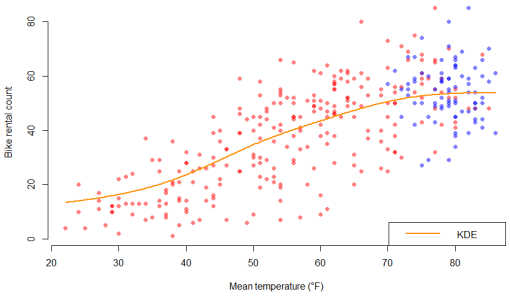


(b) Bike rental count vs. Wind speeds. Blue dots highlight the unusual sight for wind speeds in the respective time frame, and the orange line an applied Kernel density estimate.

Mean Temperature Upon examination of the mean temperature plotted over the entire time span in Figure 4a, an inverted U-shape is clearly visible, which is not unexpected given the context. The blue points highlighted in the figure refer again to the time frame between the beginning of July and the middle of September, which coincides with the period when wind speeds were observed to be anomalous and when temperatures are at their peak.



(a) Seasonal Trends in Mean Temperature over observation time frame along an applied Trend smoothing in purple.

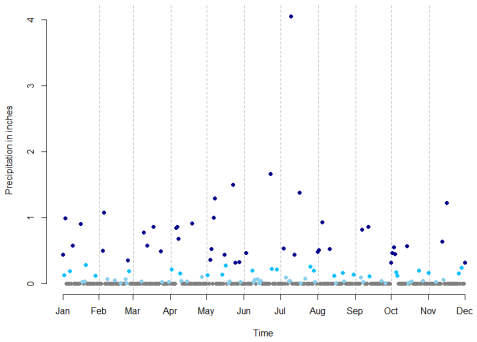


(b) Mean Temperature vs. Bike Rental Count along an applied Kernel density estimate in orange.

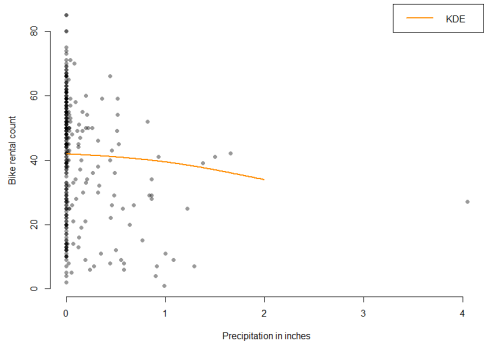
Furthermore, a strong positive correlation is evident between the two variables, as illustrated by the particular station in Figure 4b, and this correlation between mean temperature and bike rental counts represents the highest calculated correlation among the explanatory variables. The effect appears to be most accurately captured by a polynomial of degree 2, given the inverse u-shape. However, employing such an approach also entails the potential for underfitting in regions with extreme temperatures, in here particularly the area of extremely low temperatures. It is also noteworthy that the variance around these extremes at very low to low and very high temperatures is smaller, and there is considerable heterogeneity in the spread between stations. In general, it can be stated that temperature is one of the most significant explanatory variables in the context of model formulation. In order to gain a more accurate insight into the effects associated with low temperatures, another variable deduced from the mean temperature will also be used in the form of a dummy variable named Is-

Freezing. As the variable's name implies, it measures whether the mean temperature is below 32°F, which is the temperature at which water freezes, thus allowing for a more accurate representation of the effects associated with low temperatures.

Precipitation In general, large amount of precipitation in Washington, D.C. is relatively infrequent, as demonstrated by Figure 5a, which depicts the observed precipitation over the specified time period. However, there are a few exceptions, such as the 9th of July, which recorded 4 inches of precipitation. By filtering the dataset according to the precipitation categories provided by the Glossary of Meteorology over the specified time span, it was determined that 226 days exhibited no precipitation, 40 days were classified as light rain, while only 25 days were classified as moderate rain and 43 days were classified as heavy rain.



(a) Seasonal Trends in Precipitation over observation time frame along an applied Trend smoothing in purple.



(b) Precipitation vs. Bike Rental Count along an applied Kernel density estimate in orange.

Moreover, any precipitation has an observable negative effect when considering mean count calculations and a small negative correlation with regards to counts. However, as highlighted by Table 4 and Figure 5b, the final effect seems similar to wind speeds in size, particularly as precipitation becomes heavier.

Table 4: Mean Bike Rental Counts by Rain Intensity for Single Station and All Stations

	Mean Bike Rental Count			
Rain Intensity	**Non**	**Light**	**Moderate**	**Heavy**
In-Plot Station	44.4	42.1	38.3	29.7
All Stations	55.1	53.2	45.0	37.0

Snowfall & Snow depth In contrast to precipitation, there are significantly fewer days with snowfall, with only five recorded days out of 334, mostly due to the exclusion of the entire December. Four of these days fall within January, and only one day in March, which is not uncommon for that time of year. Similarly, there are only six days where snow is lying on the ground, which are all consecutive days following a snowfall in January. Although the sample size is small, the recorded effect of snowfall and snow depth on bike rental appears sever when using calculated mean bike rental count on the days of recorded snowfall and snow depth over all stations as a reference.

3 Estimation and Results

3.1 Model Formulation

The previous sections have already touched in parts on our primary goal of formulating a model that captures the relationship between the bike rental counts, our outcome variable, and the set of explanatory variables, which were discussed prior. Given that our outcome variable consists of non-negative integer counts, a Poisson regression within the generalized linear model (GLM) framework is a natural choice for modelling this type of data. Such an approach is particularly well-suited for situation where one aims to model the frequency of events, in this case bike rental counts. Therein, one models the expected number of bike rentals as a function of the explanatory variables under assumption of equidispersion, while also accounting for the count nature of the data. While this is the typical approach, we, however, decided against the usage of a Poisson GLM and instead for the usage of a log-normal LM. This approach essentially resembles an equivalent idea as the Poisson GLM pursues, but allows working in the realm of the classical LM setting, thus allowing for more modelling flexibility and possible utilization of a wide range of test statistics specifically developed for the panel data setting. Moreover, inference, particularly hypothesis testing and the interpretation of test statistics, is generally more straightforward in the log-normal case compared to the Poisson GLM where one mostly relies on deviance and the likelihood. Nonetheless, it also entails a series of disadvantages, in that coefficient interpretation represents the change in the logarithm of the outcome and that the model cannot handle counts of zero directly due to the logarithmic transformation, although the latter is here (for now) not problematic since there are no zeroes present in the data. More important however is that the former problem also influences predictions under a log-normal, because simply applying a retransformation by exponentiating the mean is not an option due to leading to a bias that underestimates the expected count. Hence, in order to correct for this bias, we have to multiply the prediction with an adjustment factor, which is derived from the variance of the error term. This factor is called a smearing estimate and was calculated based on the works of Wooldridge, p. 207, yielding a consistent, but not unbiased estimate of expected count.

In accordance with our selected LM approach, we previously advocated for the incorporation of fixed effects due to the pronounced heterogeneity evident in the plots presented in the preceding section, particularly across stations. By conducting initial Chow- and F-tests, as well as Wooldridge's test for unobserved individual effects, with the objective of comparing the pooled and station fixed effects models, we can ascertain that these fixed effects should be incorporated into the model as a fundamental component. Similarly, the month structure introduces a substantial difference between the observed and unobserved factors in the plots, while weekdays also illustrate a form of seasonality that must be accounted for. In light of the test results, it can be deduced that these factors should also constitute part of the final model. The inclusion of a lagged count term by one week, in addition to the observation of notable autocorrelation following the formulation of preceding models, similarly indicates that this should be included.

For the time being, the set of weather-related explanatory variables with its time-dependent and station-independent structure will be denoted as follows:

$$\begin{aligned}\mathbf{Weather}_t =& \beta_1 \text{Wind Speed}_t + \beta_2 \text{Mean Temperature}_t + \beta_3 \text{Mean Temperature}_t^2 + \\ & \beta_4 \text{IsFreezing}_t + \beta_5 \text{Precipitation}_t + \beta_6 \text{Snowfall}_t + \beta_7 \text{Snow depth}_t, \\ \text{with} \quad & \text{Station } i = 1, \ldots, 100, \qquad \text{Observation Day } t = 1, \ldots, T_i,\end{aligned}$$

Consider the case where the log-transformed response variable $\text{Count}_{i,t}$ is modelled as a linear function of covariates under fixed effects transformation, then the final log-normal

regression model is formulated as:

$$\mathbb{E}\left[\log(\text{Count}_{i,t})\right] = \mathbf{Weather}_t + \beta_8 \log(\text{Count}_{i,t-1}) + \alpha_{\text{Station}} + \alpha_{\text{Month}} + \alpha_{\text{Weekday}} + \epsilon_{i,t},$$
$$\text{with} \quad \text{Station } i = 1, \ldots, 100, \qquad \text{Observation Day } t = 2, \ldots, T_i,$$

where the dependent variable is independently drawn, with $\alpha_{\text{Station}}, \alpha_{\text{Month}}$, and α_{Weekday} representing the respective station-, month- and weekday-specific fixed effect and idiosyncratic error ϵ_{it}.

3.2 Regression Results

Table 5 displays the regression coefficient table of our model for the analysis, which were estimated in version 2.6.4 of the *plm* R-package and version 4.2.2 of R. In this table, we can inspect that every variable is significant and the signs are as expected for the most part, the exception here being the IsFreezing dummy variable which one would moreso expect to be negative. This, however, is most likely due to the variable basically taking on the role of an adjustment intercept for the levels of extremely low temperatures compared to the rest, which also tends to be the area with worse fit overall, indicated by diagnostic residual plots. Additionally, when comparing the final model with a restricted one that doesn't include the variable, most coefficients stay very similar except for the mean temperature that shows substantially higher values in both the slope and second degree polynomial term, thus points also more towards the former explanation of why there is a positive sign.

Table 5: Regression Summary of Bike Rental Count Data Model

Dependent Variable:	$\log(\text{Count}_{i,t})$			
Variables	**Estimate**	**Std. Error**	**t-value**	**Pr(>|t|)**
Wind Speed	-2.5205e-02	7.1995e-04	-35.0091	< 2.2e-16 ***
Mean Temperature	6.8460e-02	1.9006e-03	36.0195	< 2.2e-16 ***
Mean Temperature^2	-4.4825e-04	1.5997e-05	-28.0212	< 2.2e-16 ***
IsFreezing	2.6840e-01	1.5392e-02	17.4379	< 2.2e-16 ***
Precipitation	-4.0577e-01	6.6034e-03	-61.4491	< 2.2e-16 ***
Snowfall	-1.3321e-01	1.5234e-02	-8.7443	< 2.2e-16 ***
Snow depth	-2.2577e-01	2.4102e-02	-9.3671	< 2.2e-16 ***
log(Count)	2.1943e-01	4.6610e-03	47.0774	< 2.2e-16 ***
Fixed-effects				
Unbalanced Panel: $i = 100$, $T = 179 - 325$, $N = 31920$				
Station (100 Stations), Month (11 Months), Weekdays (7 Days)				
Fit statistics				
Residual Sum of Squares	4641.4			
R-Squared	0.59278			
Adj. R-Squared	0.5912			
F-statistic	1928.51 on 24 and 31796 DF, p-value: < 2.22e-16			

The R^2 and adj. R^2 both indicate an explained correlation of approximately 60%, which suggests that the model is relatively accurate in terms of its explanatory power, despite being limited to a few explanatory variables. Nevertheless, this level of accuracy is insufficient for performing high-precision predictions that allow for narrow 95% confidence interval boundaries in the out-of-sample case. This is due to the fact that the standard errors used in their calculation are largely, but not exclusively, dependent on the size of residual variance. The consequence of this can be observed in Figure 6, depicting the predicted mean bike rental count for station 10th & K St NW, which placed 65th in the subsequent Activity Ranking

in Section 3.4. The figure serves to illustrate the aforementioned concerns regarding the generally excessively large interval ranges, which can be attributed to the large standard errors. In this particular case, the interval width covers a range of approximately 61 units for an average August day prediction at 80°F mean temperature (-22 and +39 around the expected value at 45 predicted mean bike counts). Furthermore, it also demonstrates the heightened uncertainty in regions with lower temperatures, which can be observed by comparing the interval widths before and after the discontinuity created by the IsFreezing variable. It should be noted that while Figure 6 clearly highlights the discussed problem regarding prediction, the 95% confidence interval for in-sample predictions is notably absent. This is due to the interval boundaries being too narrow in the in-sample case, meaning that including them in the figure would make them indistinguishable from the mean predictions.

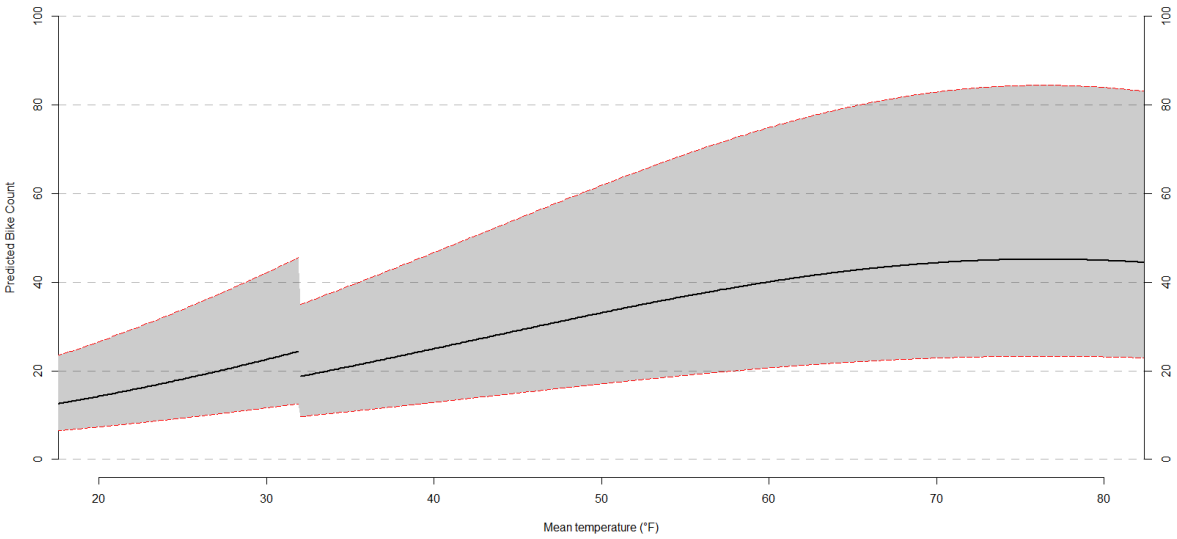


Figure 6: Predicted Mean Bike Rental Count as a function of Mean Temperature (black line) and its interaction with the IsFreezing dummy variable (discontinuity at 32°F). The predictions were calculated based on the 10th & K St NW station on a Monday in August and taking the mean values of the other variables, which were exclusively calculated on August entries. The outer red lines highlight a 95% out-of-sample Prediction Confidence Interval.

Furthermore, a diagnostic analysis was performed on the residuals and fitted values to ascertain whether the assumptions inherent to the fixed effects estimator were valid, thus ensuring the reliability of the results obtained through the within transformation. The illustration of the expected residuals in dependence on specific stations demonstrated the aforementioned lack of fit in terms of mean temperature at extremely low temperatures and, on occasion, also at extremely high temperatures. It also revealed a noticeable lack of fit in the months of January, February and November, as well as a lack of fit for the IsFreezing variable, where residuals are not mean-centred in the case where the dummy is equal to one. In conclusion, a significant proportion of the underfitted areas correspond to periods of low bike rental activity, while peak times are well represented in the model. Additionally, residuals versus fitted plots for each station indicate that heteroscedasticity is not a concern in this model. However, the Wooldridge test for an AR(1) error component yielded a significant result, suggesting the potential for serial correlation.

3.3 Predictions

The final model also offers the possibility of in-sample prediction and out-of-sample forecasting, which may be performed on any of the 100 stations between the beginning of January and the end of November, along with a calculation of the smearing estimate. This functionality is available within the PredictionModel.R file via the *get_ interval()*-command, which also lists the corresponding confidence intervals and offers the option of additional tuning parameter specification. At this time, only singular predictions can be made on the provided *plm.model* and the *plm.data.frame* object x, which consists of a single row and represents the necessary specified arguments. Additionally, one may pass down an already *cleaned* plm.data.frame row that only contains the necessary variable information, as in the final model. This is achieved by setting the *cleaned* argument to TRUE, otherwise the cleaning procedure will be run by default. Furthermore, the *type* argument allows the user to switch between in-sample "prediction" and out-of-sample "forecast". Moreover, the smearing estimate and residual standard error are calculated based on the residuals from all stations using the *sigma* and *smearing* arguments by the "default" setting. Nevertheless, the aforementioned calculations may also be performed on the residuals, limited to the specific station for which prediction is desired, which can be achieved by setting the relevant argument to "specific" for both the sigma and smearing calculations. Lastly, the default t-test values are calculated at a significance level of 0.05, but this can be again modified by passing down a custom alpha value via the *alpha* argument.

3.4 Hot Zones & Activity Ranking

Once the station-dependent fixed effects have been extracted from the final model, an analysis can be conducted to determine which stations exhibited the greatest activity and which contributed to the formation of hot zones in Washington, D.C. These serve as station-specific baselines for activity that are independent of all aforementioned variables, which are controlled for within the final model. The results are presented in a map of Washington, D.C. created within Google My Maps, which can be accessed via the online webpage. Lastly, a ranking based on the station-dependent fixed effect size yields the Table 6 highlighted below, which differs significantly from the previous one in Table 2. This concludes our analysis. For further details on this ranking, please refer to the "Fixed effects ranking" section in the Imputation.R file.

Table 6: Ranking of Station activity based on station-specific fixed effects (rounded)

Rank	**Station**	**Fixed effect**
1	New Hampshire Ave & T St NW	1.250
2	15th & P St NW	1.244
3	Columbus Circle / Union Statio	1.231
4	1st & M St NE	1.140
5	Lincoln Memorial	1.061
⋮	⋮	⋮
96	17th & K St NW / Farragut Square	0.339
97	10th & E St NW	0.317
98	1st & M St SE	0.302
99	New York Ave & 15th St NW	0.274
100	Ohio Dr & West Basin Dr SW / MLK & FDR Memorials	0.218