Paris-Dauphine PSL University

Generalized Linear Model

M1–MIDO 2024-2025

Final Project - Prediction/Explanation of the number of bicycle rentals, based on various characteristics

Authors :
Arthur DANJOU
Antonin DUROUSSEAU

24th January 2025

Table des matières

Introduction

With the rise of super-metropolises, significant population growth has been accompanied by an increasing need for mobility. The ecological challenges associated with transportation modes pose a major issue for policymakers. Among the proposed solutions, this study focuses on a particularly promising option : bike-sharing systems. These systems, requiring minimal space and offering unmatched flexibility, are attracting a growing number of users for daily commutes. In this context, we analyze the dataset "projet.csv" to gain insights into the bike rental process. The goal is to develop a predictive model capable of anticipating demand, thereby avoiding bike shortages. In our case, it is preferable to slightly overestimate the number of required bikes rather than risk a shortfall, which could compromise user satisfaction.
The dataset comprises 1,817 observations characterized by environmental and temporal measures, detailed as follows :

- **saison** : Winter, Spring, Summer, Autumn.
- **météo** : Clear, Cloudy/Foggy, Rain/Snow.
- *humidité* : Air humidity rate (percentage).
- *vent* : Wind speed (km/h).
- *température1* : Average measured temperature (°C).
- *température2* : Average perceived temperature (°C).
- **mois** : From 1 (January) to 12 (December).
- **jour_mois** : From 1 to 31 (day of the month).
- **jour_semaine** : From 1 (Sunday) to 7 (Saturday).
- **vacances** : 1 if the day is during holidays, 0 otherwise.
- **jour_travail** : 1 if the day is a workday, 0 otherwise.
- **horaires** :
  - 1 : From 0 :00 to 7 :00,
  - 2 : From 7 :00 to 11 :00,
  - 3 : From 11 :00 to 15 :00,
  - 4 : From 15 :00 to 19 :00,
  - 5 : From 19 :00 to 24 :00.
- *vélos* : Number of bike rentals.

Quantitative variables are displayed in *italics*, while qualitative variables are in **bold**. This dataset will serve as the foundation for constructing a reliable predictive model based on these variables.
It is clear that weather and seasonal conditions play a crucial role in the decision to use a bike, but which variables are truly significant ? For instance, summer and spring days with sunny skies, moderate temperatures (neither too hot nor too cold), low wind, and low humidity levels are logically more conducive to bike use than days with less favorable weather conditions.
Similarly, peak hours — from $7:00$ to $11:00$ and $15:00 to 19:00$ — are expected to have higher demand than off-peak hours. Additionally, holidays, when more people are likely to engage in leisure activities, appear to be another significant factor. However, it is hypothesized that variables such as the day of the week or the month may not significantly impact bike rentals, as regular users tend to maintain consistent habits across weeks and months.
To validate these hypotheses, we will conduct a systematic data analysis. This will begin with an examination of the quantitative (*humidité*, *vent*, *température1*, *température2*) and qualitative (*saison*, *météo*, *mois*, *jour_mois*, *jour_semaine*, *jour_travail*, *vacances*, *horaires*) variables to identify explanatory variables and those to be explained. Outlier detection will also be performed.
We will analyze correlations between variables to better understand their interactions. Different linear models will then be compared to identify the one that best meets the goals defined above.

Finally, the performance and limitations of the selected final model will be thoroughly analyzed.

1 Data Analysis

In this section, we will focus on analyzing the variables and their potential correlations. First, we will transform the quantitative variables in the dataset into factors to make them usable for our analysis and readable by *R*.
Next, we will separately study the quantitative and qualitative variables. For the quantitative variables, we will examine their distribution, existing correlations, and potential transformations to apply, using common functions such as the square root or squared power.
For the qualitative variables, we will analyze their distribution as well as potential correlations between them.
Finally, we will detect the outliers present in the data and discuss whether it is more appropriate to remove them or keep them in the dataset.

1.1 Data Preprocessing

After importing our data, we first checked its quality by looking for missing values (`NA`) in the columns and any potential duplicate rows. This verification showed that no missing values or duplications were present, confirming the initial integrity of the data.
Next, we examined the types of variables (quantitative or qualitative) to ensure their format was suitable for analysis. We also used the `summary` function to get an overview of the main characteristics of our variables, such as descriptive statistics (mean, median, minimum, maximum, quartiles) for quantitative variables, and the distribution of modalities for qualitative variables.
This preliminary step allowed us to better understand the data distributions, identify potential anomalies, and lay the groundwork for further stages of the analysis.

1.2 Study of Quantitative Variables

1.2.1 Distribution of Variables

We visualized the histograms of the quantitative variables by setting the number of bins to 30, which provided a detailed representation of their distributions (see Figure 1).
We decided to add transformations of the variables using common functions such as the exponential, logarithm, square root, and squared power. However, for the variables *température1* and *température2*, we could not apply logarithmic and square root functions because the values of these variables can be negative.

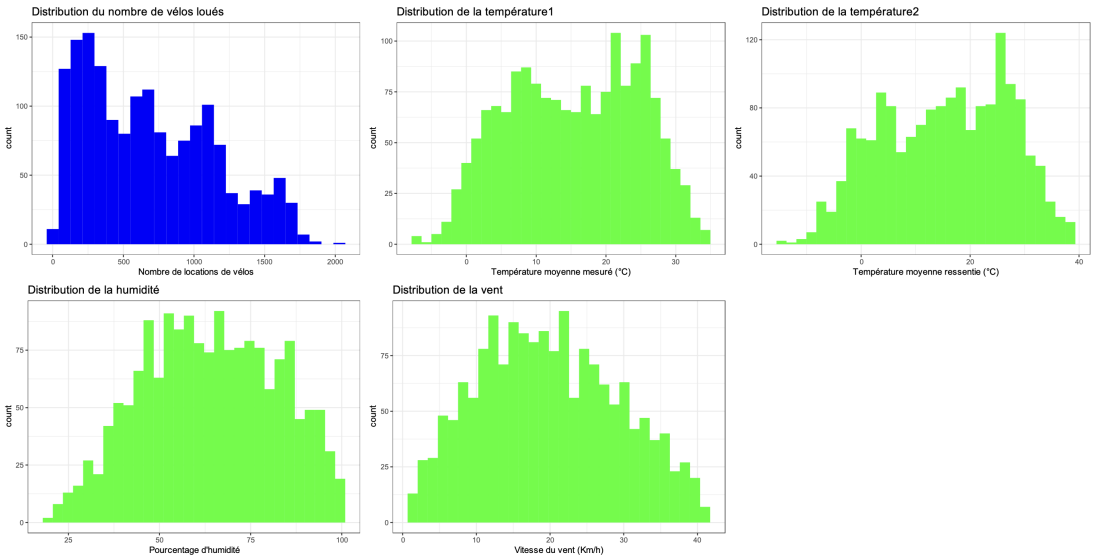

FIGURE 1 – Histogram of the distribution of quantitative variables

Based on this visualization, we formulated an initial hypothesis regarding the distribution followed by our target variable, *vélos*. Since the support of this variable is $\mathbb{N}$, it is likely that it follows a Poisson distribution or a Negative Binomial distribution, both of which are commonly used for discrete counting data.
We will confirm and precisely define the distribution followed by *vélos* in section 2.2 : Choice of the distribution for *vélos*.

1.2.2 Correlation between Quantitative Variables

To analyze the correlations between the quantitative variables, we used the `corrplot` function to generate a graphical matrix of correlations (see Figure 2), making it easier to identify strong or weak relationships between the variables. We observe a logical relationship between the variables and their transformations, but further investigation into this relationship is not necessary. However, we also note a correlation between *température1* and *température2*, which is an expected relationship given the definitions of measured and felt temperature.

We then deepened the analysis by performing a Pearson test to validate the correlation between *température1* and *température2*, getting a *p-value* of $< 2.2 \times 10^{-16}$, (see Figure 3). This confirms a significant correlation between these two variables. The strong redundancy between them implies that a choice will need to be made to retain only one of the models. We will justify this decision in section 2.3 : Model Selection.

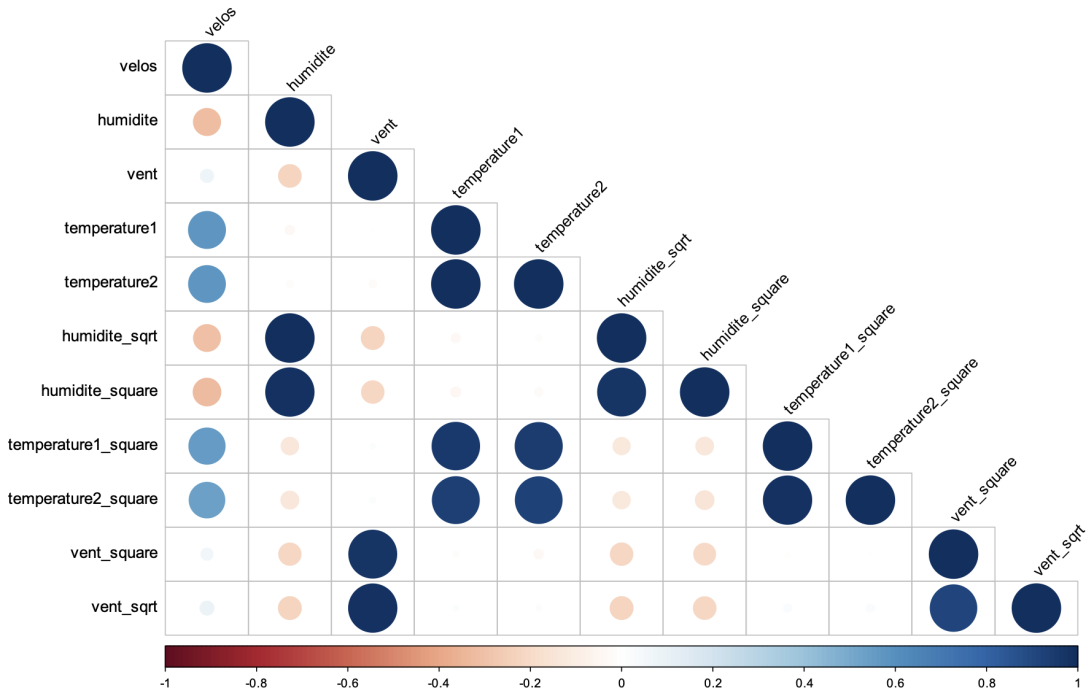

FIGURE 2 – Correlation Plot of the Qualitative Variables

```
    Pearson's product-moment correlation

data:  data$temperature1 and data$temperature2
t = 393.29, df = 1663, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9941306 0.9951549
sample estimates:
      cor
0.9946672
```

FIGURE 3 – Pearson Test for *température1* and *température2*

Finally, using the `pairs` plot (see Figure 4), we visualized the linear relationships between the different variables, providing a better understanding of the potential links and dependency patterns within the data.

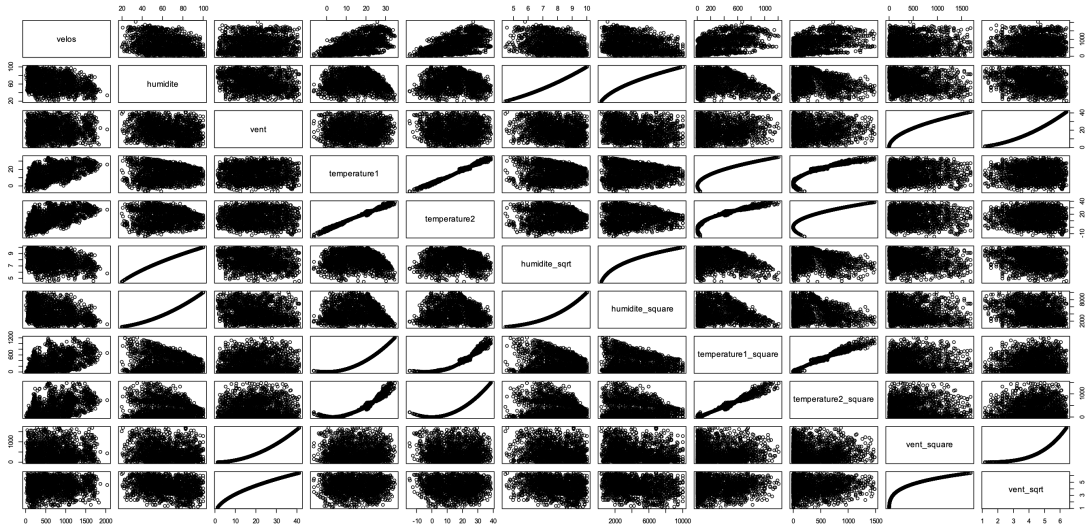

FIGURE 4 – Pairs Plot of the Quantitative Variables

1.3 Study of Qualitative Variables

Once the analysis of the quantitative variables was completed, we proceeded with the analysis of the qualitative variables. We used boxplots to visualize the distribution of *velos* based on the different qualitative variables (see Figure 5). These plots allowed us to identify the variables that appear to have a significant impact on the target variable. Based on these visualizations, we selected the following variables as potentially important for predicting *velos* : *horaires*, *jour_mois*, *mois*, *saison*, *vacances*, and *meteo*.

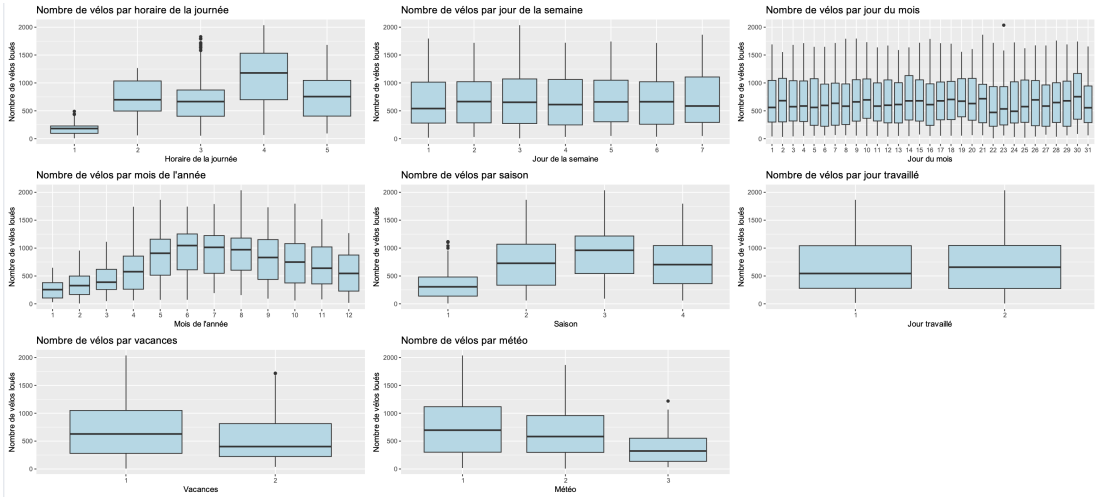

FIGURE 5 – BoxPlot of the Qualitative Variables

To go further, we conducted chi-square tests to evaluate the probability of observing the distribution differences between the categories, assuming that these variables are independent in the distribution process. The tests were carried out for the variables *saison*, *meteo*, and *mois*, and the results are presented below, see Figure 6.

```
    Pearson's Chi-squared test

data:  data$mois and data$saison
X-squared = 3974.1, df = 33, p-value < 2.2e-16


    Pearson's Chi-squared test

data:  data$meteo and data$saison
X-squared = 25.923, df = 6, p-value = 0.0002301


    Pearson's Chi-squared test

data:  data$meteo and data$mois
X-squared = 88.106, df = 22, p-value = 7.185e-10
```

FIGURE 6 – Chi-square test of the Qualitative Variables *saison*, *meteo* and *mois*

These tests allowed us to validate the hypothesis of an association between the variables *météo*, *mois*, and *saison*.

1.4 Outliers Detection

Once the analysis of all our variables was completed, we proceeded with the analysis of outliers using `boxplot` for the quantitative variables.

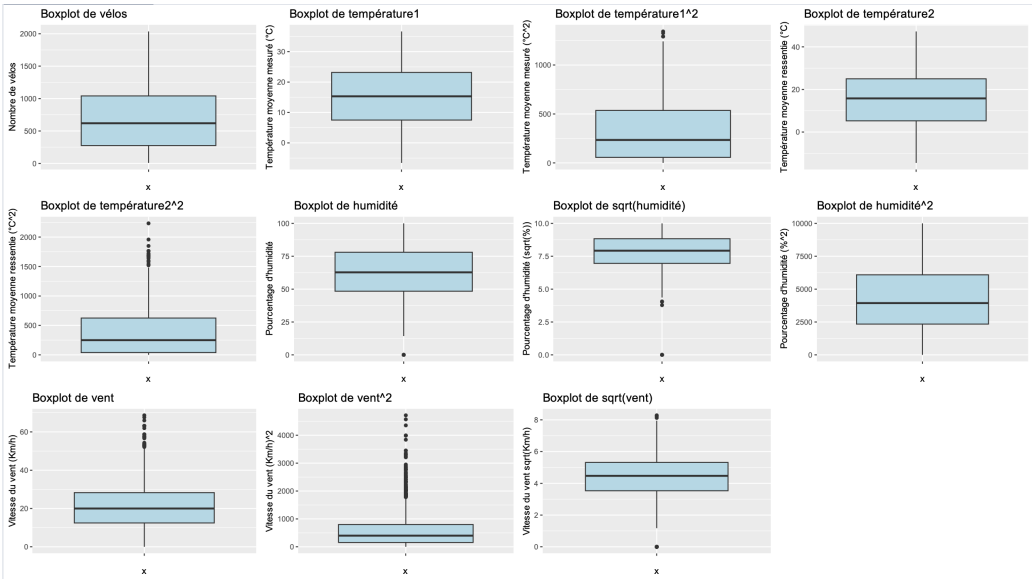

FIGURE 7 – BoxPlot of the Qualitative Variable highlighting Outliers

During this analysis, we identified outliers for the variables *humidite*, *vent*, *temperature2* as well as their transformations, totaling **152** outlier values (see Figure 8). After careful consideration, we decided to remove all the outliers for these variables. This choice is justified by the fact that humidity percentages close to 0 are extremely rare and are generally encountered in extremely dry areas such as the Sahara or Death Valley, where the use of bike-sharing is virtually

nonexistent. Additionally, cycling under wind speeds of at least $55km/h$ becomes dangerous, so we also chose to remove these observations. For the variables derived from transformations, we removed the observations where the value was outside of $[Q1-1.5*(Q_3-Q_1); Q1+1.5*(Q_3-Q_1)]$, where $Q1$ and $Q3$ are the 0.25 and 0.75 quantiles of the variables.

```
Number of outliers for the variable velos : 0
Number of outliers for the variable humidite : 5
Number of outliers for the variable vent : 24
Number of outliers for the variable temperature1 : 0
Number of outliers for the variable temperature2 : 0
Number of outliers for the variable humidite_sqrt : 1
Number of outliers for the variable humidite_square : 0
Number of outliers for the variable temperature1_square : 4
Number of outliers for the variable temperature2_square : 11
Number of outliers for the variable vent_square : 80
Number of outliers for the variable vent_sqrt : 27
Number of outliers removed : 152
Data length after removing outliers : 1665
```

FIGURE 8 – Outlier Verification Output and Removal

2 Model Creation and Comparison

Now that we have prepared and cleaned our data, we can begin constructing several models and comparing them to select the one that best meets our objectives.
First, we will discuss the *train/test* split method for the dataset, which is essential for evaluating the performance of our models on unseen data. We will then discuss the model comparison criteria we covered in class, to eliminate irrelevant variables.
After that, we will incorporate interactions between variables in our models to identify those that truly add value to the prediction. This step will help us better understand the combined influence of certain variables.
Finally, we will justify the choice of the final model, explaining its relevance and validity. We will also provide a detailed interpretation of the estimated regression coefficients to understand the role of each variable in predicting bike-sharing demand.

2.1 Data Split

Using the `rsample` library, we split our dataset into two subsets : the `train` set, representing *80%*, and the `test` set, containing the remaining *20%*.

2.2 Choice of the Distribution *vélos*

Before beginning the model selection, we must choose the correct distribution for our *vélos* variable among : the **Poisson** distribution, the **Negative Binomial** distribution, and the **Gaussian** distribution. We created three complete models using the `glm()` and `glm.nb()` functions on our *data_train* set. By comparing the AICs of the three full models and selecting the

smallest, we chose the **Negative Binomial** distribution (see Figure 9).
We also performed a dispersion test to validate our choice : this test evaluates whether a statistical model, such as a Poisson or Negative Binomial regression model, exhibits overdispersion. Overdispersion occurs when the variance of the data is larger than what the model assumes. The *p-value* $< 2.2e^{-16}$ confirms that overdispersion is present in the model's residuals.
Finally, by directly calculating the mean and variance of the *vélos* variable, we confirm our final choice of distribution : we will definitely use the **Negative Binomial** distribution.

```
    model_poisson model_nb model_gaussian
AIC      82988.75  17532.5       18420.81

    Overdispersion test

data:  model_poisson
z = 20.793, p-value < 2.2e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  54.85188

Mean : 693.4317 Variance : 214261.1
```

FIGURE 9 – Output of AIC comparisons, Dispersion test and the Mean and the Variance of the Variable to be explained

2.3 Model Selection

The first model built includes only the quantitative variables to identify those that significantly explain our target variable : the number of *vélos*. To do this, we used `ANOVA Type I` and `Type II` tests, which allowed us to confirm the choice of relevant variables. After analysis, we retained the following variables based on their *p-values* being below 5% :

- *temperature1*
- *temperature1_square*
- *humidite_square*
- *humidite_sqrt*
- *humidite*
- *vent*
- *vent_square*

Next, we added the qualitative variables identified as interesting in section 1.3 : Study of Qualitative Variables. These variables include :

- *mois*
- *saison*
- *horaire*
- *météo*
- *vacances*
- *jour_mois*

The `ANOVA` tests allowed us to remove some irrelevant variables : *vent_sqrt*, *jour_mois*, *jour_travail* and *jour_semaine*.

Then, we performed model selection using the `AICStep` function with the *forward*, *backward*, and *bothward* approaches. All three methods converged to the same optimal model, with the minimal AIC. This final model includes the following variables :

- *horaire*
- *mois*
- *météo*
- *température1*
- *saison*
- *température1_square*
- *humidité_square*
- *vacances*
- *humidité*
- *vent*

2.4 Addition of Interactions

Once the variables were selected, we needed to test the interactions between them. Many trials were performed to calculate the AIC of different models, while keeping the issue of *overfitting* in mind.
Thus, for each combination of variables and their potential interactions, we fitted several statistical models. The primary goal was to identify the best combination that minimizes the AIC, while ensuring that overfitting was avoided, as it could compromise the model's ability to generalize.
Additionally, we performed first order `ANOVA` tests (using the `anova` function in `R`) to retain only the significant interactions. This step helped reduce model complexity while ensuring the relevance of the selected interactions.
To do this, we followed an iterative approach : after each adjustment, we compared the performance of the models using criteria such as the AIC, but also by examining the residuals and calculating the *Mean Squared Error* score, which we will discuss further in section 3.1 : Predictions and *Mean Square Error*. Models that were too complex, despite having a low AIC, were discarded if they showed signs of *overfitting*.

2.5 Final Model

2.5.1 Choice and Validation of the Model

After adding, testing, and comparing the possible interactions among the selected variables, we get the following final model :

- *horaire*
- *mois*
- *meteo*
- *temperature1*
- *saison*
- *température1_square*
- *humidité_square*
- *vacances*
- *humidité*
- *vent*
- *horaire* `:` *température1*
- *température1_square* `:` *mois*

We performed an `LRTest` between the final model and the final model without interactions. The *p-value* was $< 5\%$ (see Figure 10), so we retained the model with interactions as the definitive model.
We computed the dispersion ratio based on the Pearson residuals and got a value of 0.99, which is close to 1. This indicates that the model does not exhibit overdispersion or underdispersion. Additionally, we plotted the histogram of the deviance residuals, which exhibits a pattern closely resembling a **Normal** distribution (see Figure 11).

```
Likelihood ratio test

Model 1: velos ~ temperature2 + humidite + humidite_sqrt + saison + horaire *
    meteo + mois * temperature2_square
Model 2: velos ~ temperature2 + temperature2_square + humidite_sqrt +
    humidite + horaire + saison + mois + meteo
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  45 -8681.6
2  26 -8728.8 -19 94.278  5.739e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 10 – LRTest on the final models with and without interactions

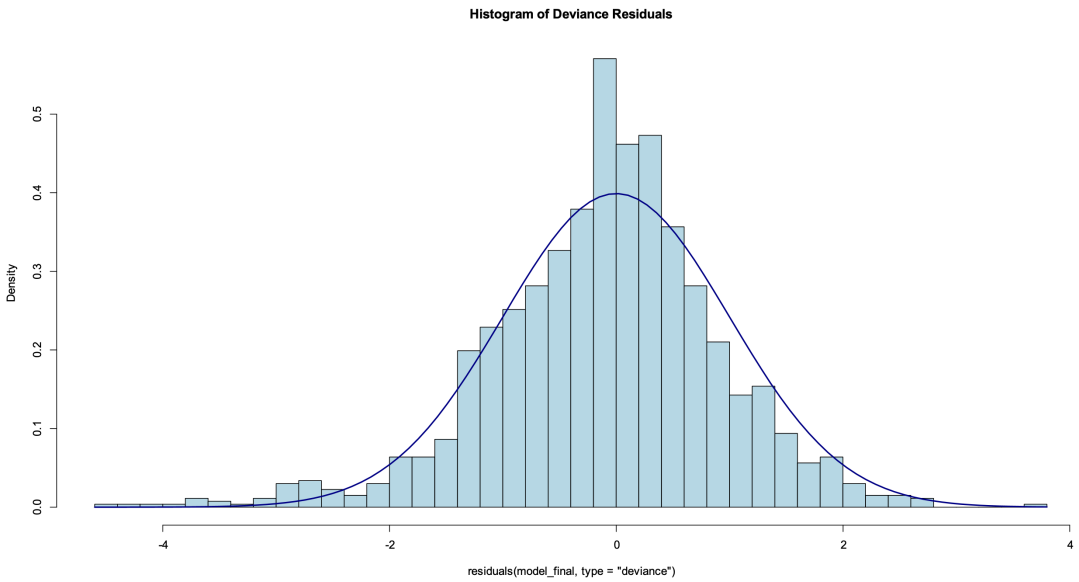

FIGURE 11 – Histogram of the deviance Residuals

2.5.2 Interpretation of Estimated Regression Coefficients

We can thus write the mathematical equation of our model. Since the distribution of the response variable is a negative binomial distribution, the canonical link function is the **log** function.

We have :

$$\ln(\mathbb{E}[\text{velos}]) = \beta_0 + \sum_{i=2}^{5} \beta_{\text{horaire}_i} \cdot \text{horaire}_i + \sum_{j=2}^{12} \beta_{\text{mois}_j} \cdot \text{mois}_j + \sum_{k=2}^{3} \beta_{\text{meteo}_k} \cdot \text{meteo}_k$$

$$+\beta_{\text{temperature1}} \cdot \text{temperature1} + \sum_{l=2}^{4} \beta_{\text{saison}_l} \cdot \text{saison}_l + \beta_{\text{temperature1\_square}} \cdot \text{temperature1\_square}$$

$$+\beta_{\text{humidite\_square}} \cdot \text{humidite\_square} + \beta_{\text{vacances}_2} \cdot \text{vacances}_2 + \beta_{\text{humidite}} \cdot \text{humidite} + \beta_{\text{vent}} \cdot \text{vent}$$

$$+ \sum_{i=2}^{5} \beta_{\text{horaire}_i:\text{temperature1}} \cdot (\text{horaire}_i \cdot \text{temperature1})$$

$$+ \sum_{j=2}^{12} \beta_{\text{mois}_j:\text{temperature1\_square}} \cdot (\text{mois}_j \cdot \text{temperature1\_square})$$

We can interpret the coefficients from the `summary` as follows :

— β_0 is the intercept of the model.
— $\beta_{\text{temperature1}}$ represents the effect of the variable *temperature1*.
— $\beta_{\text{temperature1\_square}}$ represents the effect of the variable *temperature1_ square*.
— $\beta_{\text{humidite\_square}}$ represents the effect of the variable *humidite_ square*.
— β_{humidite} represents the effect of the variable *humidite*.
— β_{vent} represents the effect of the variable *vent*.
— The terms β_{horaire_i} represent the effects of the different levels of the *time slot* variable, with $i = 2, \ldots, 5$, and 1 being the reference level.
— The terms β_{mois_j} represent the effects of the different levels of the *mois* variable, with $j = 2, \ldots, 12$, and 1 being the reference level.
— The terms β_{meteo_k} represent the effects of the different levels of the *meteo* variable, with $k = 2, \ldots, 3$, and 1 being the reference level.
— The terms $\beta_{\text{vacances}_l}$ represent the effects of the different levels of the *vacances* variable, with $l = 2$, and 1 being the reference level.
— The terms $\beta_{\text{horaire}_i:\text{temperature1}}$ represent the interactions between the *temperature1* and *horaire* variables, with $i = 2, \ldots, 5$, and 1 being the reference level.
— The terms $\beta_{\text{mois}_j:\text{temperature1\_square}}$ represent the interactions between the *temperature1_ square* and *mois* variables, with $j = 2, \ldots, 12$, and 1 being the reference level.

We get the following estimations of the coefficients :

— $\hat{\beta}_0 = 4.1688$: Represents the logarithm of the expected number of bikes when all the explanatory variables (and interactions) are at their reference level.
— $\hat{\beta}_{\text{horaire}_2} = 1.6137$: Means that the logarithm of the expected number of *bikes* increases by 1.6137 (or approximately +400% in exponential terms) for time slot 2 compared to time slot 1.
— $\hat{\beta}_{\text{mois}_5} = 0.5272$: Means that for month 5 (Mai), the logarithm of the expected number of bikes is higher by 0.5272 (or approximately +69%) compared to month 1 (January).
— $\hat{\beta}_{\text{meteo}_3} = 0.4136$: Weather condition 3 (Pluie/Neige) decreases the logarithm of the expected number of *vélos* by 0.4136 (or approximately −33%) compared to weather condition 1.
— $\hat{\beta}_{\text{temperature1}} = 0.0509$: For each additional degree Celsius of the measured *average temperature*, the logarithm of the expected number of bikes increases by 0.0509 (approximately +5%).
— $\hat{\beta}_{\text{humidite}} = 0.0101$: An increase of 1 unit of humidity leads to an increase of 0.0101 in the logarithm (approximately +1%).

— $\hat{\beta}_{\text{horaire}_2:\text{temperature1}} = 0.0157$: For time slot 2, each additional degree Celsius decreases the logarithm of the expected number of bikes by 0.0157.
— $\hat{\beta}_{\text{mois}_5:\text{temperature1\_square}} = 0.0028$: In May (month 5), each additional unit of the square of the *température1* increases the logarithm of the expected number of *vélos* by 0.0028.

3 Performance and Limitations of the Final Model

In this section, we will begin by calculating the *Mean Square Error (MSE)* score and plotting the `binnedplot` based on our model predictions to demonstrate the performance of the final model. Next, we will categorize the predictions to facilitate the interpretation of results and to assess both prediction quality and accuracy. Finally, we will discuss the limitations of the model and potential improvements. It is important to note that in our case, overestimating the number of *vélos* is preferable to underestimating it.

3.1 Predictions and *Mean Square Error*

The first step is to predict the number of *vélos* using our final model on the *test dataset*. This dataset is also used to compute the confidence intervals for our predictions. We use the following formula to calculate the MSE :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Our model produces an MSE around 45000 *vélos*^{2}, leading to a *Root Mean Square Error (RMSE)* around 215 *vélos*. Since the RMSE is significantly lower than the actual mean (693.4317), this indicates that the model is relatively accurate. Considering that our model is intended for general estimations rather than highly precise predictions, this level of error is acceptable. Additionally, the confidence intervals allow us to quantify the uncertainty of the predictions, providing further insight into the model's reliability.
Moreover, given that the goal of the model is to support decision-making where overestimation is less problematic than underestimation, the observed error patterns align with our practical needs. We can state that our model is biased by about 200 *vélos* on average.

3.2 Evaluation of Model Performance

To evaluate the performance of the final model, we compared the null deviance and the deviance of our model with `summary`. The deviance of our model is significantly lower than the null deviance. This indicates that our model provides a better fit than the null model.
Then, we divided our variable *vélos* into three categories : *Low*, *Mid*, and *High* affluence. These categories are defined as follows : below 200 *vélos* for the *Low* category, between 201 and 650 *vélos* for the *Mid* category, and above 651 *vélos* for the *High* category. The last category is significantly broader because our model struggles to accurately predict a high number of *vélos*, an issue we will revisit in 3.3 : Model limitations and improvements. Using these categories, we computed the `confusion matrix` (see Figure 12).
We first analyze the `accuracy` score, as it provides a global assessment of the model's performance, ensuring it behaves correctly in most cases. With an `accuracy` of 83%, the model demonstrates strong reliability for decision-making purposes.
Next, we focus on the `recall` score, which measures the model's ability to correctly identify all positive cases (in this context, critical or underestimated predictions). A high recall is particularly important as it reduces the risk of underestimating the number of *vélos*, which could result

in service shortages. For the *Low* category, the model achieves a `recall` of 90%, while the *Mid* and *High* categories exhibit scores of 77% and 86%, respectively. These results highlight the model's strong performance, particularly in detecting low and high-affluence scenarios, while indicating room for improvement in the *Mid* category.

```
Confusion Matrix and Statistics

          Reference
Prediction Low Mid High
      Low   38  12    0
      Mid    4  96   24
      High   0  17  142

Overall Statistics

               Accuracy : 0.8288
                 95% CI : (0.784, 0.8677)
    No Information Rate : 0.4985
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7163

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Low Class: Mid Class: High
Sensitivity              0.9048     0.7680      0.8554
Specificity              0.9588     0.8654      0.8982
Pos Pred Value           0.7600     0.7742      0.8931
Neg Pred Value           0.9859     0.8612      0.8621
Prevalence               0.1261     0.3754      0.4985
Detection Rate           0.1141     0.2883      0.4264
Detection Prevalence     0.1502     0.3724      0.4775
Balanced Accuracy        0.9318     0.8167      0.8768
```

FIGURE 12 – Confusion Matrix for the *vélos*'s categories

Finally, we display the barplot of our observed categories (see Figure 13), as well as the plot of the predicted *vélos* versus the actual numbers (see Figure 14). In the *High* category, we observe significantly larger confidence intervals, which can be attributed to the model's difficulty in accurately predicting a large number of *vélos*.
The plot also shows greater dispersion in predictions for higher values, indicating that the model struggles to perfectly capture large counts. However, if this overestimation remains within an acceptable margin (e.g., within a high confidence bound), they are preferable to underestimations, as they ensure the system can adequately respond to high demand for *vélos*.

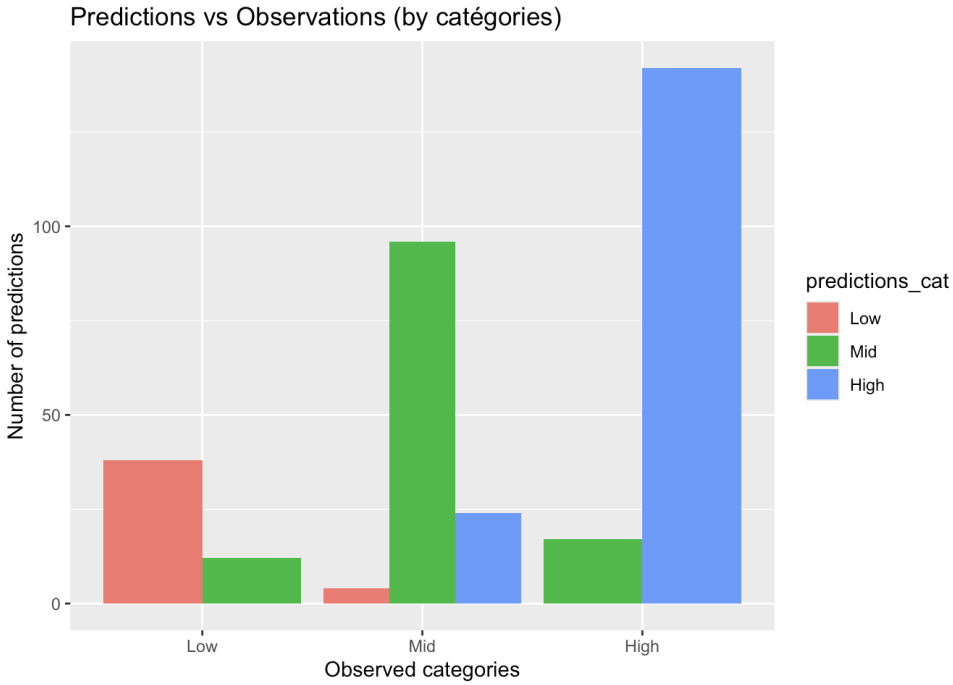

FIGURE 13 – Barplot of the predictions observed by categories

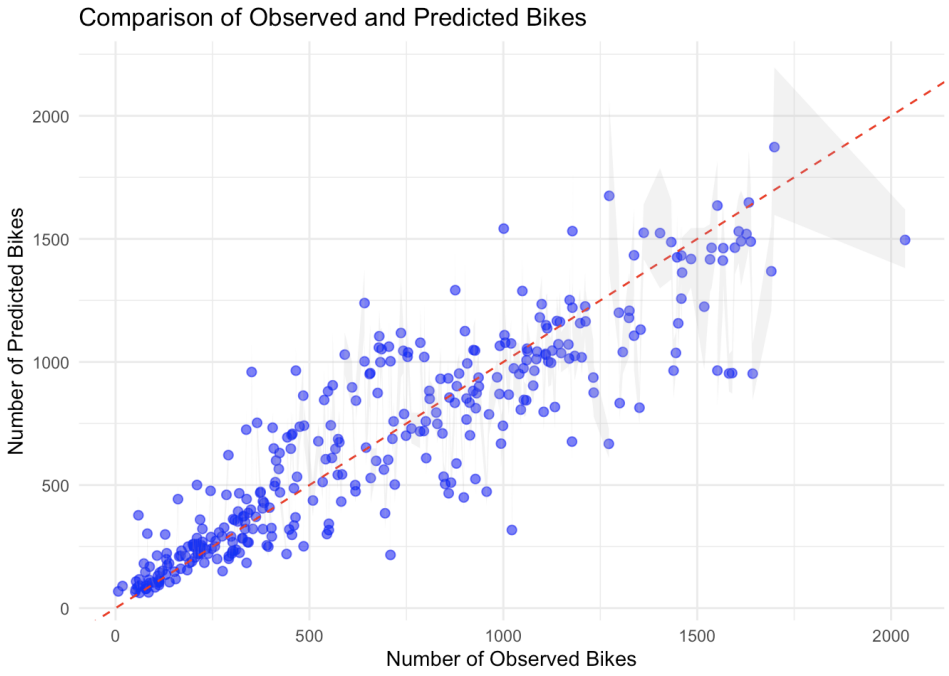

FIGURE 14 – Plotting predictions against the actual number of *vélos*

3.3 Model Limitations and Improvements

As defined earlier, our model tends to overestimate the predicted number of *vélos* rather than underestimate it. However, it struggles to predict very high numbers of *vélos*, leading to significantly larger confidence intervals in the `binnedplot` (see Figure 16) for high prediction values.
Moreover, a larger dataset with more observations of high *vélos* counts would allow us to adjust the model and reduce prediction errors. Additionally, we could focus on making new predictions for values close to the profiles of outliers to better understand the final impact of such cases.
Future improvements could involve refining the model to reduce prediction variance while maintaining its conservative bias. Investigating whether external factors, such as weather anomalies or seasonal effects, could explain the observed outliers in the predictions would also be a valuable avenue for further research.

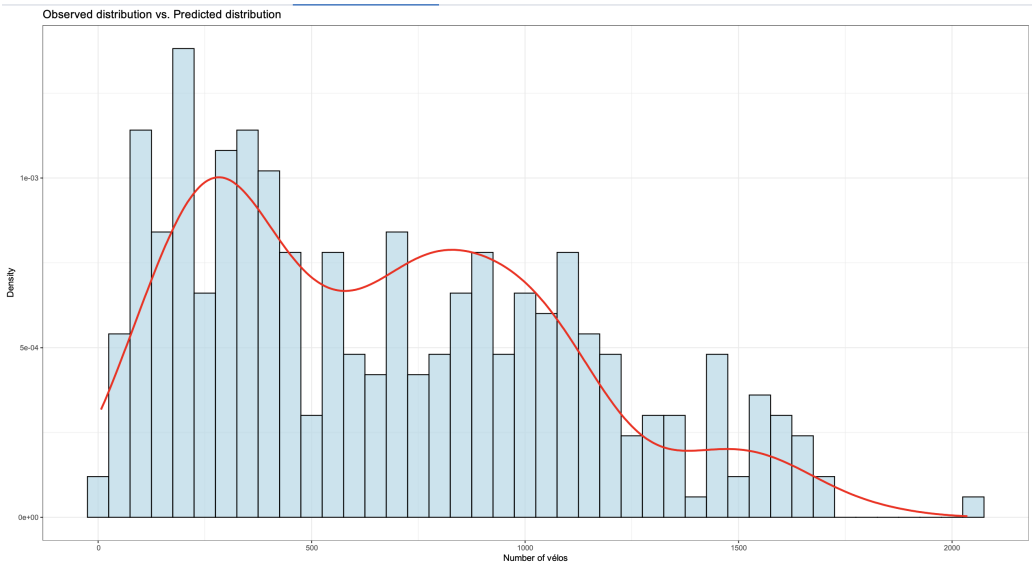

FIGURE 15 – Final distribution of our model

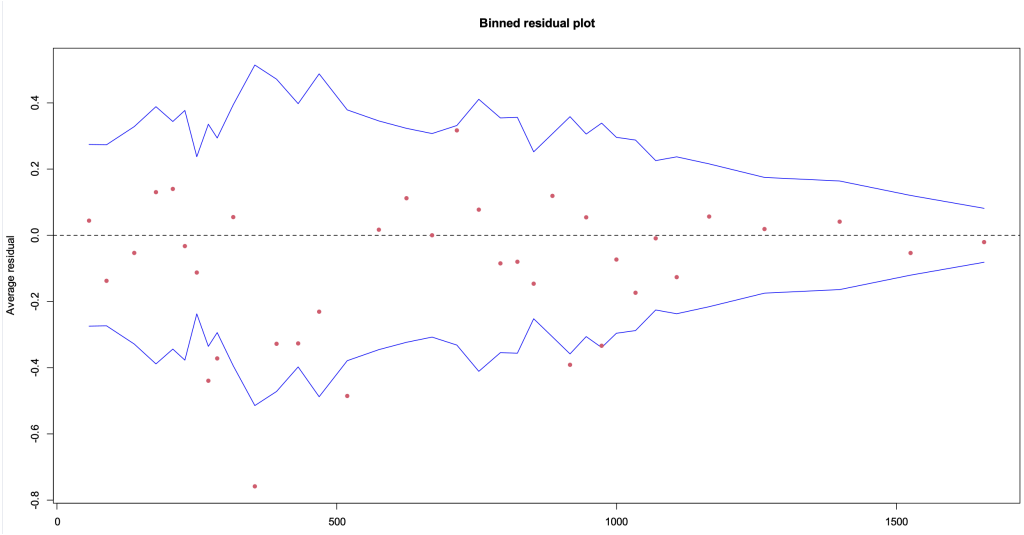

FIGURE 16 – Binnedplot of the residuals

Conclusion of the Analysis

Our study of the *projet.csv* dataset was carried out in three major phases. First, we explored and processed the data to ensure its readiness for analysis. This phase involved transforming variables into usable formats and studying their behavior with respect to the target variable, *vélos*, as well as their interrelationships. Outliers were identified and removed based on clear justifications, ensuring a cleaner and more reliable dataset for modeling.
Second, we developed and compared multiple predictive models to identify the most effective one. This process began with the selection of an appropriate distribution for the target variable. We then adopted an iterative approach, constructing models step by step : starting with only quantitative variables, followed by the inclusion of qualitative variables, and ultimately incorporating interactions between variables. Each step was carefully validated to optimize the model's predictive performance.
Finally, we evaluated the chosen model by thoroughly analyzing its accuracy and limitations. To this end, we validated its predictions, created new categories to better assess its performance, and visualized its limitations through comparisons of predicted values against actual observations. This analysis revealed a model that aligns with the specific approach of our study : favoring overestimation to avoid critical shortages in service, particularly during periods of high demand. While this conservative approach may sacrifice precision in certain cases, it ensures a robust safeguard against underestimations that could disrupt service reliability. This aligns with the overarching goal of maintaining continuous and reliable access for users, even in scenarios of high demand.
Nevertheless, the model does exhibit limitations in accurately predicting extreme values, particularly in high-demand cases. Addressing these issues would require a larger dataset, especially with more observations of peak usage patterns, to refine its precision. Additionally, the model shows a systematic bias of approximately 200 bikes, which remains reasonable given the maximum observed number of bikes : 2,036. Future iterations of the model could focus on reducing this bias while maintaining the conservative approach to avoid underestimation.
This model offers significant potential for practical applications. Policymakers, municipalities, urban planning authorities, and bike rental companies could leverage its insights to optimize the planning and installation of new bike-sharing stations. It can also aid in monitoring the number of bikes in circulation on existing networks, identifying demand trends, and ensuring proper allocation of resources. Furthermore, the model could support decisions related to scaling operations, maintaining equipment, or replacing bikes as needed. Ultimately, this study provides a solid foundation for enhancing urban mobility strategies. By ensuring that bike-sharing services can expect and respond to demand fluctuations, this model contributes to building more sustainable, efficient, and user-focused transportation systems. As cities continue to evolve, this tool could become a critical asset in addressing the challenges of modern urban mobility.