BIKE SHARING DEMAND PREDICTION: AN EXPLORATION OF WEATHER, TEMPORAL, AND SEASONAL FACTORS USING REGRESSION

By

Yoseph Feleke, Subhodh Kally, and Max Christian

December, 2024

**INTRODUCTION**

Bike-sharing systems have gained significant popularity in urban areas due to their convenience and environmental benefits. Accurately predicting bike rental demand based on various factors such as weather conditions, seasons, and temporal patterns is crucial for optimizing bike availability and improving service efficiency. This project aims to develop a regression model that predicts the daily or hourly demand for bike rentals in Washington D.C.'s Capital Bikeshare system using weather data, time features, and other relevant variables.

This project aims to create a regression model that predicts the total bike rental count in a day based on various quantitative and categorical variables such as temperature, humidity, wind speed, season, holiday status, and weather conditions. By analyzing these factors, we aim to uncover patterns and relationships that influence bike demand and develop an accurate predictive model. Using our model, we intend to analyze the following questions: *How do weather conditions (temperature, humidity) affect bike rentals?* And *Can the model accurately predict demand for a specific day and/or weather condition?*

These questions are crucial for optimizing bike-sharing systems by predicting demand based on weather conditions. This ensures efficient bike allocation, enhances customer satisfaction, improves resource planning, boosts revenue, and promotes sustainable transportation through reliable and accessible services.

**DATASET**

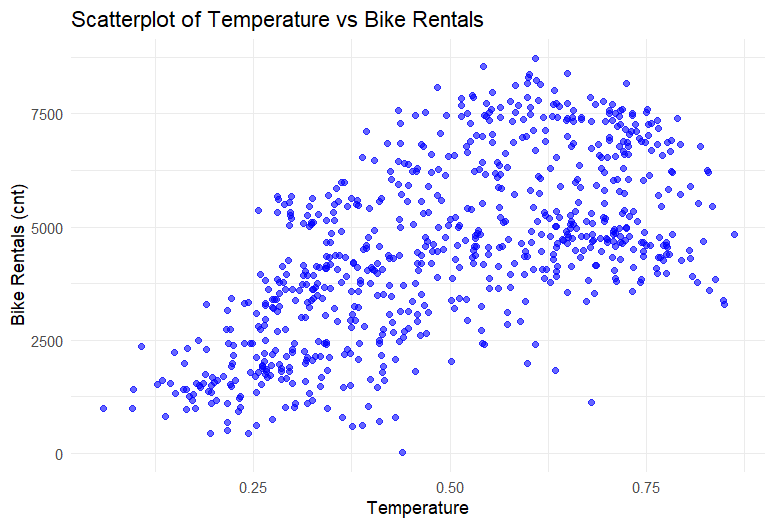
The dataset for this project is obtained from the UCI Machine Learning Repository and contains two years of data (2011-2012) taken from the Capital Bikeshare system in Washington D.C. The dataset includes 731 observations and 16 total variables, both quantitative and categorical. The units of observation for the rows are the dates of the year spanning from 01/01/2011 to 12/31/2012. The variables are defined as follows:

| **Variable Name** | **Role** | **Type** | **Description** |
| --- | --- | --- | --- |
| *instant* | ID | Integer | Record index |
| *dteday* | Feature | Date | Date |
| *season* | Feature | Categorical | Season (1:winter, 2:spring  3:summer, 4:fall) |
| *yr* | Feature | Categorical | Year (0: 2011, 1: 2012) |
| *mnth* | Feature | Categorical | Month (1 to 12) |
| *holiday* | Feature | Binary | Whether the day is a holiday  (0: nonholiday, 1: holiday) |
| *weekday* | Feature | Categorical | Day of the week |
| *workingday* | Feature | Binary | Whether day is a working day  (1: neither weekend nor holiday, 0: otherwise) |
| *weathersit* | Feature | Categorical | 1**:** Clear or partly cloudy (favorable weather conditions)  2**:** Mist and cloudy  3**:** Light rain, light snow, or other moderate conditions  4**:** Heavy rain, heavy snow, or other severe weather conditions |
| *temp* | Feature | Continuous | Normalized temperature in Celsius |
| *atemp* | Feature | Continuous | Normalized feeling temperature in Celsius |
| *hum* | Feature | Continuous | Normalized humidity (divided by max, 100) |
| *windspeed* | Feature | Continuous | Normalized wind speed (divided by max, 67) |
| *casual* | Other | Integer | Count of casual users |
| *registered* | Other | Integer | Count of registered users |
| *cnt* | Target | Integer | Count of total rental bikes, including both casual and registered user components |

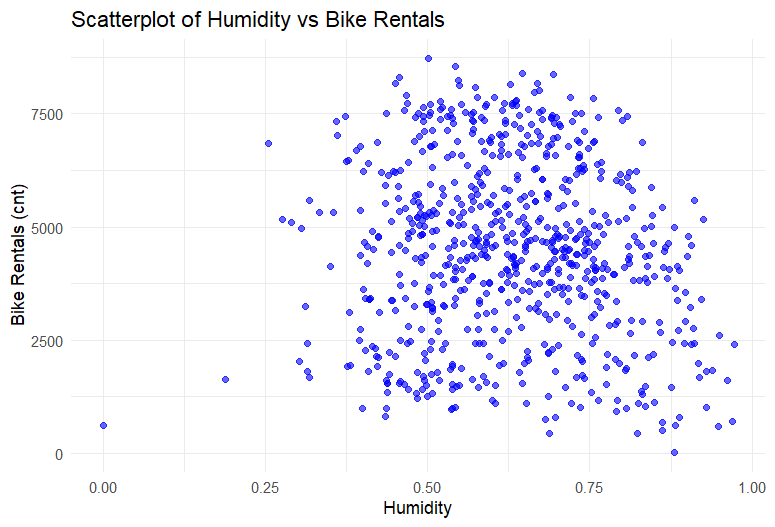
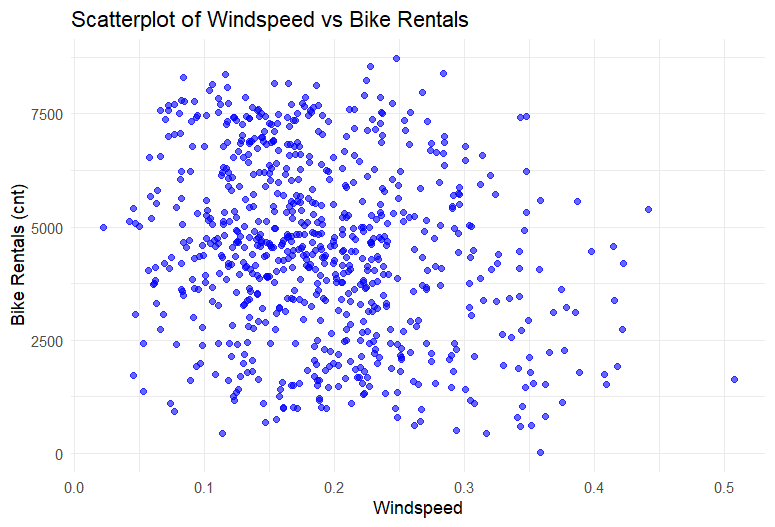
Of the 15 variables that are not the target (*cnt*), we see that *instant, dteday, casual,* and *registered* do not provide any meaningful information in predicting the response. In particular, *instant* and *dteday* serve as identifier-esque variables, while *casual* and *registered* are the total count of bike rentals on the day broken down into casual and registered users. Because the goal of our analysis is to accurately predict the **total** number of bike rentals in a day, we are not concerned as to how they are split between casual and registered users.

**EXPLORATORY ANALYSIS**

We look at initial relationships between some of the various predictors and the response.

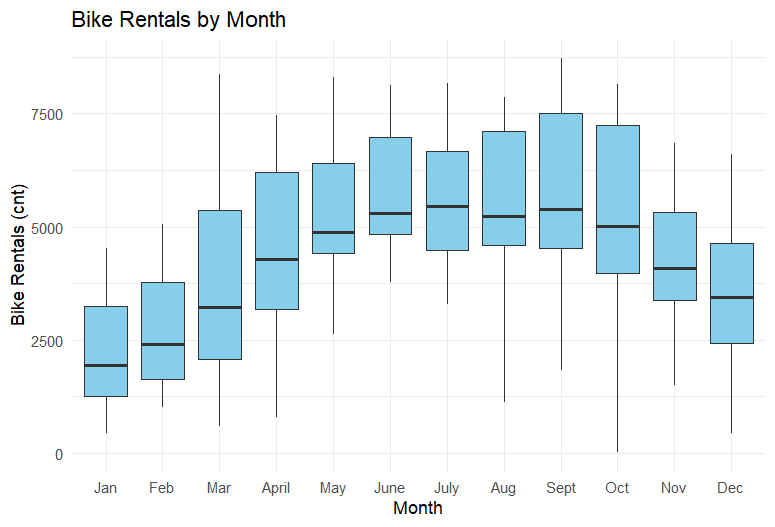
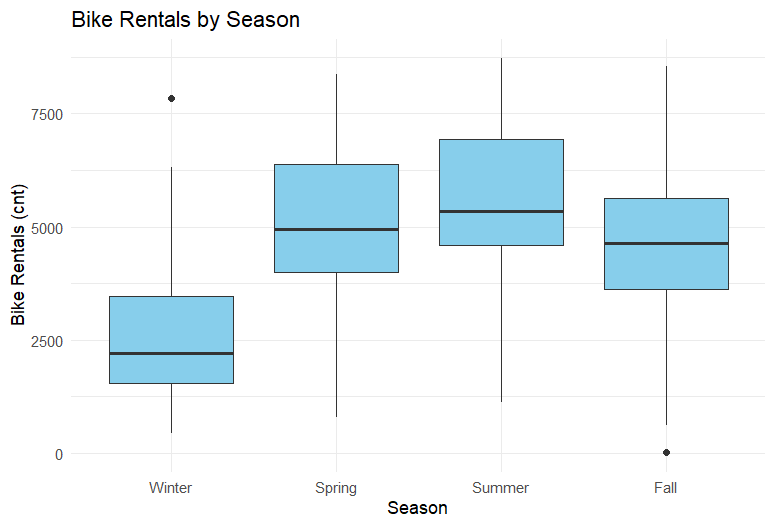


Temperature seems to indicate a nonlinear, and possibly quadratic relationship with the response. It appears that bike rentals increase with temperature, but decrease towards the highest and lowest temperatures, which is to be expected.

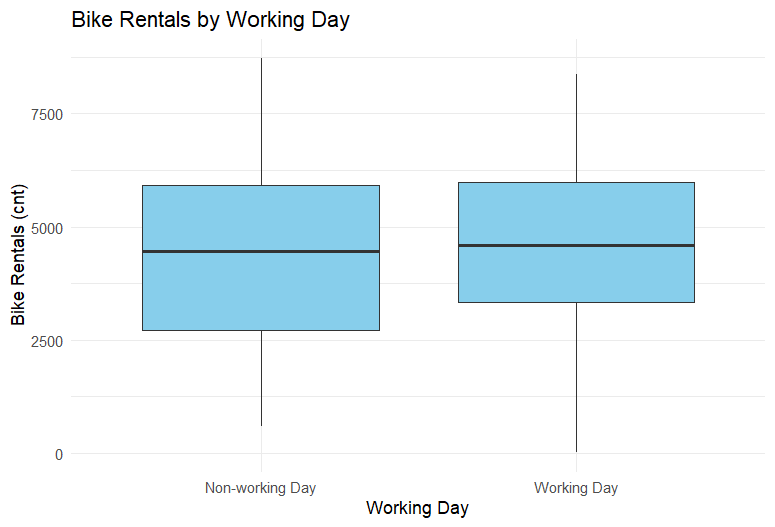
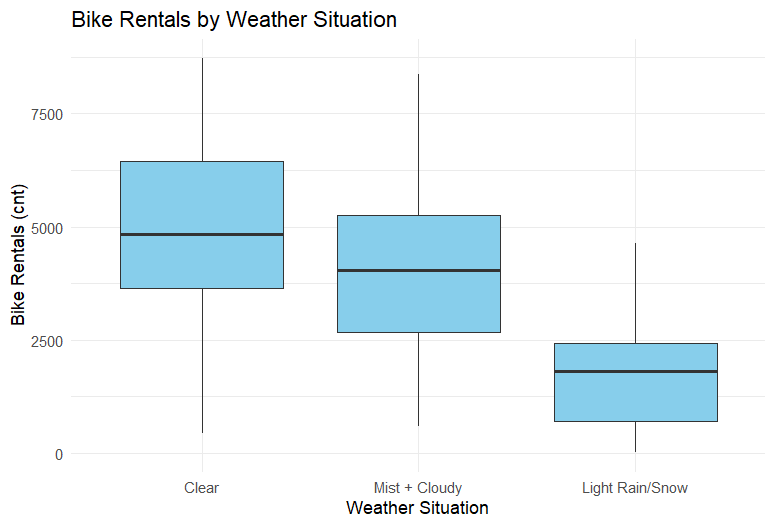


Meanwhile, wind speed and humidity don’t show any kind of strong relationship with the response and are spread like “confetti-in-a-box”.

Looking at the relationship between the categorical variables and the number of bike rentals, we observe the following.



Similar to the relationship between temperature and bike rental count, the coldest months and seasons (the winter and fall months, but particularly winter) observe lower amounts of bike rentals, while bike rentals rise in the warmer seasons (the summer and spring months). The responses between *season* and *mnth* show very similar trends, and so these predictors would appear to be correlated, which is not surprising as the seasons are defined by groups of months. Therefore, we might exclude *month* as a predictor when we fit the model.

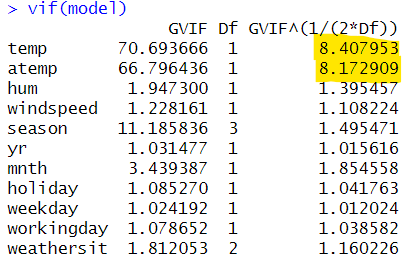
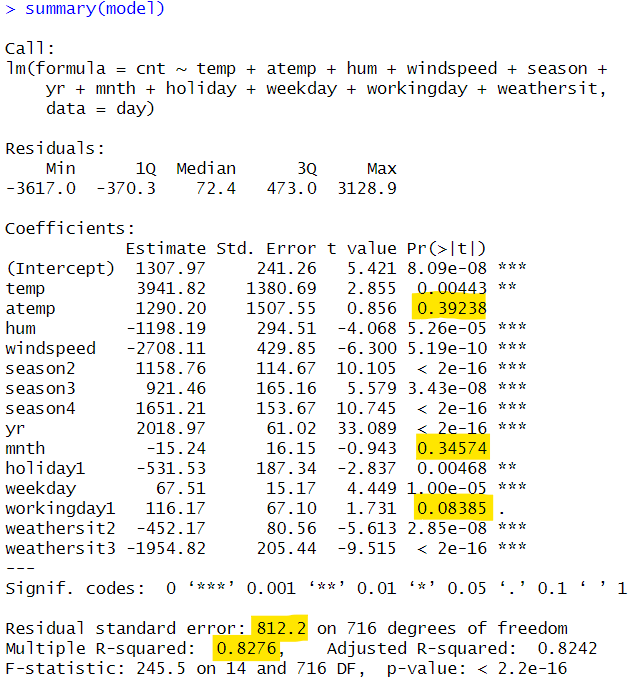


We see that the weather situation on a given day exerts some influence on the amount of bike rentals; clear days experience the highest count of rentals and rainy/snowy days the lowest count, with moderate weather conditions like mist and cloudiness falling in the middle of the two ranges. Once again, this makes sense as people would probably be most likely to ride a bike in clearer weather conditions.

Finally, we see that there is no significant difference in bike rental demand between a working day and non-working day, though there is a higher spread for demand on non-working days. This indicates that a day being either a working or non-working day may not be very significant in predicting the total amount of bike rentals on that day.

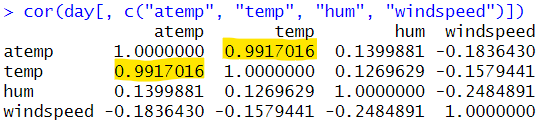
**MODEL FITTING**

We first fit a “full” model with all predictors except for identifiers *instant* and *dteday,* and *cnt* components *casual* and *registered*. Categorical predictors are converted to factors.

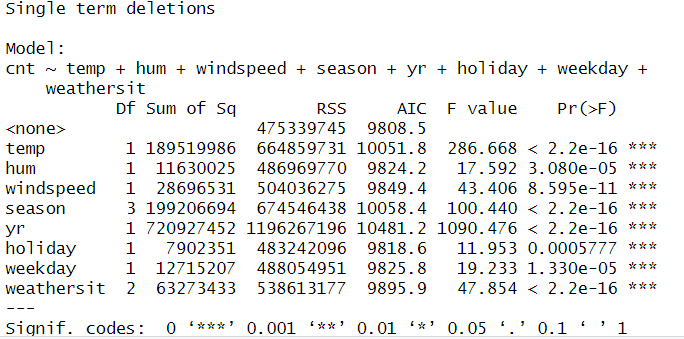
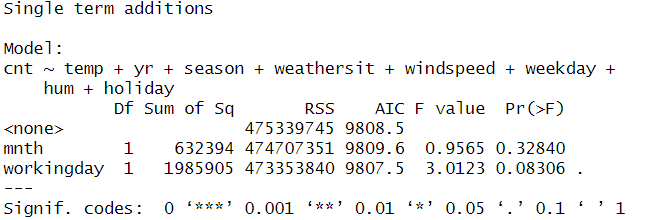


Model 1

*R2* (0.8276) looks agreeable here, but *MSRes* (812.22) seems a bit high. We see that month and working day have large *p*-values at significance level of 0.05, so we may decide to remove them from our final model. Additionally, when checking variance inflation factors, we observe the variables temperature (*temp*) and feels-like temperature (*atemp*) have high VIF values, suggesting significant multicollinearity between the models. Checking the correlation matrix, we can confirm they have an extremely high pairwise correlation coefficient (>0.99).

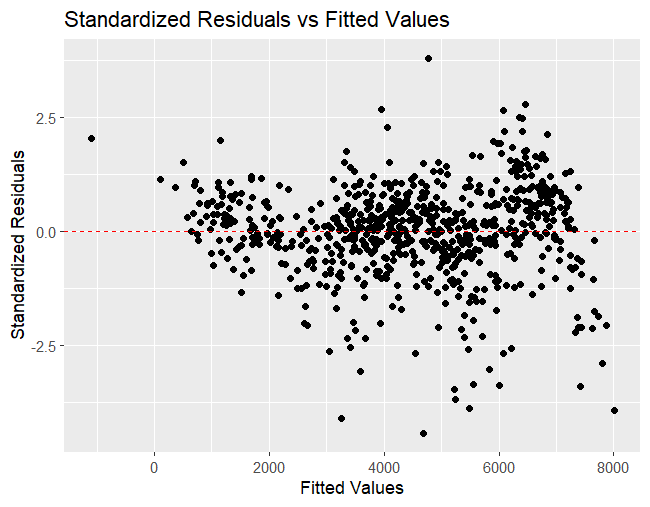
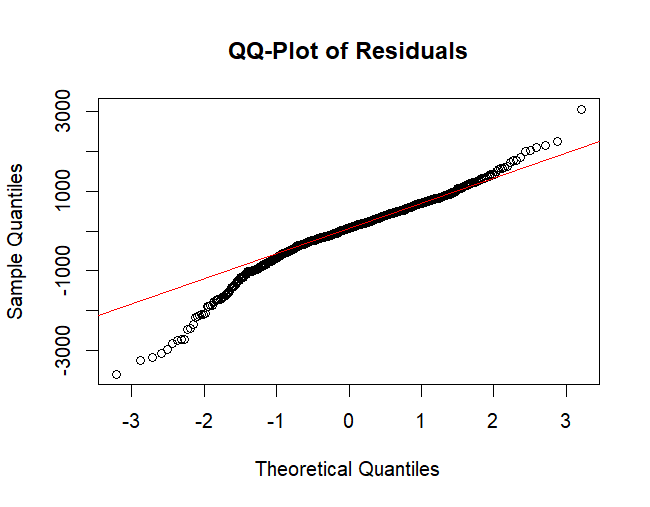


Therefore, we can remove *atemp* as a predictor when refitting the model. We also remove *mnth* and *workingday*, as their high *p*-values support the initial conclusions made in the exploratory analysis - namely, that they are not very significant in predicting total bike rental count. The decision to remove these variables is supported by the suggested models given by forward and backward variable selection at a significance level of 0.05.

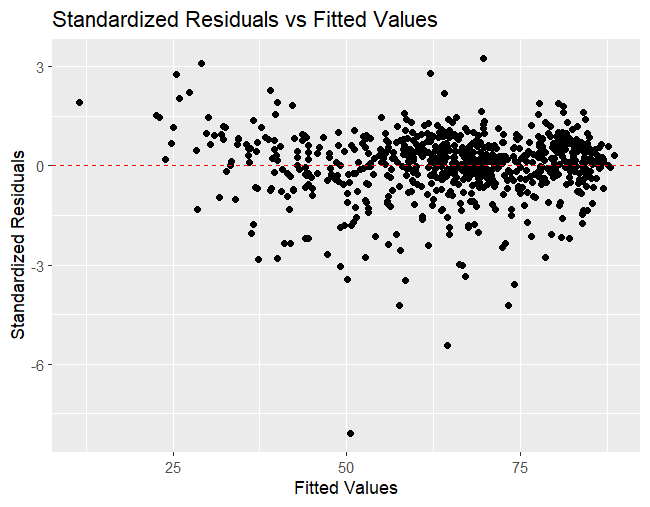
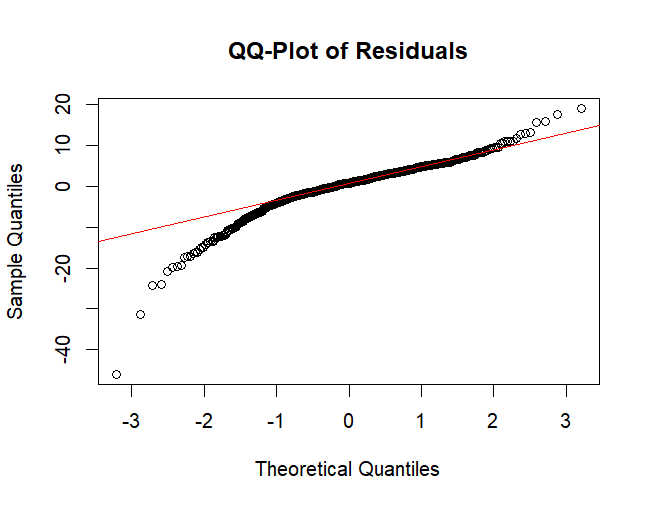
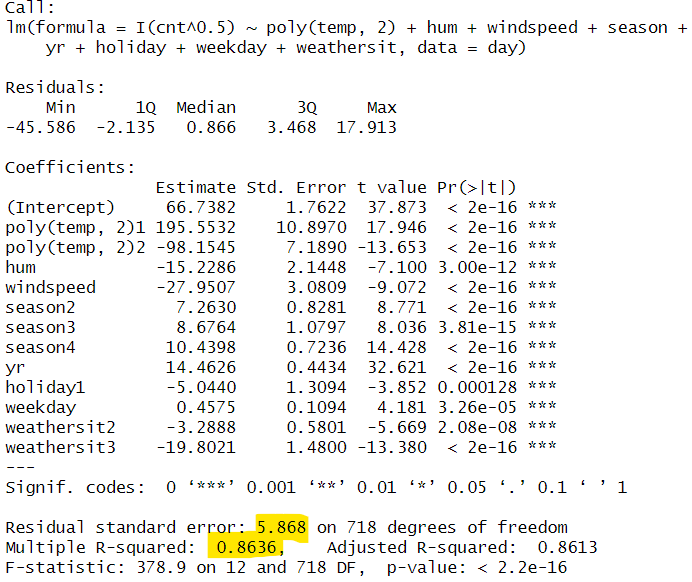


Forward and backward selection

Refitting the model with the insignificant predictors removed yields no significant changes in *R2* or *MSRes*. Below, we see that the refitted model roughly satisfies the normal distribution assumption placed on the residuals, but when examining the residuals vs. fitted values plot, we notice a U or W shape pattern, which violates the constant variance assumption placed on the residuals. When analyzing BoxCox for an appropriate transformation on the response, we see that the bound of the 95% confidence interval for the lambda value falls between approximately 0.55 and 0.7. We then transform the response *cnt* variable with an easily interpretable square root transformation, as 0.5 is not too far from the interval.



Additionally, recall that we observed temperature seems to have some kind of polynomial, and seemingly quadratic, relationship with the response. We also transform the predictor *temp* and refit the model with both the transformed *cnt* and *temp* variables.



Model 3

We notice that we have improved significantly on *R2* (+4%) and also have significantly decreased *MSRes*. We also notice that the residual assumptions are better satisfied, with a stabilized, more constant variance as indicated by the residual plot and a closer Normal distribution shown in the *qq*-plot.

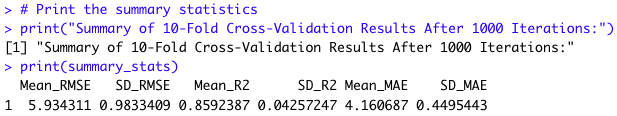
**MODEL VALIDATION**

**PRESS Statistic**

|  | **Model 1**  *(initial, full model)* | **Model 2**  *(sqrt trans. on count)* | **Model 3**  *(final model with both trans.)* |
| --- | --- | --- | --- |
| **PRESS Statistic** | 1242828459 | 32649.86 | 26095.76 |

We see that the PRESS statistic drastically lowers between our initial and final models, which indicates that our final model has the best predictive accuracy of the three. In other words, our final model has a good ability to generalize to unseen data, which is one of the key goals of our project- to accurately predict the total number of bike rentals in a day. Additionally, since the Model 3, the most simple model, has the lowest PRESS statistic, we can conclude that our model is well-regularized and is not a victim of significant overfitting.

**Cross Validation**



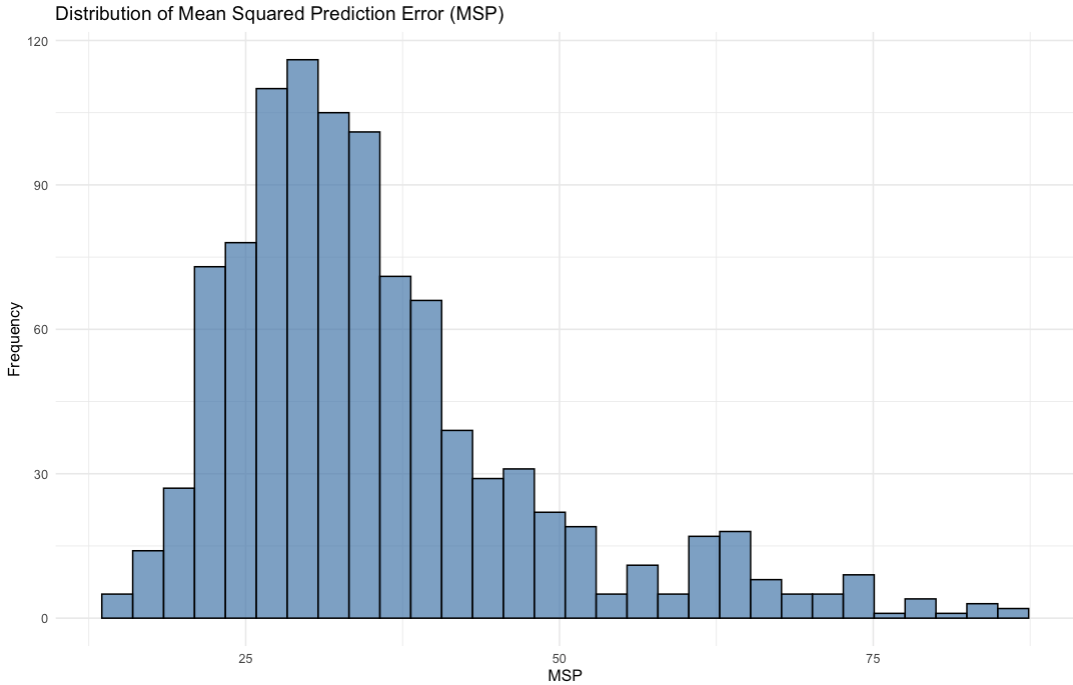
The regression model was evaluated using 10-fold cross-validation repeated 1000 times, with the following results:

* Root Mean Squared Error (RMSE): 5.91591
* R-squared (R2R2): 0.860232
* Mean Absolute Error (MAE): 4.139796

These metrics indicate that the model explains approximately 86% of the variability in the square root of bike rentals (*cnt*0.5), with a mean prediction error of approximately 5.92 units on the transformed scale. The low variability in these metrics across iterations demonstrates the consistency of the model. The overall cross- validation performance supports the model's reliability for predicting bike rentals.

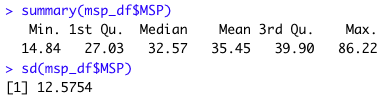
The *R2* value of the final model without cross-validation was 0.8602, which is very close to the mean *R2* value of 0.8592 obtained from cross-validation. This similarity demonstrates that the final model's performance generalizes well across different subsets of the data, validating its robustness and reliability and indicating low bias present in the fitted model.

**MSP**

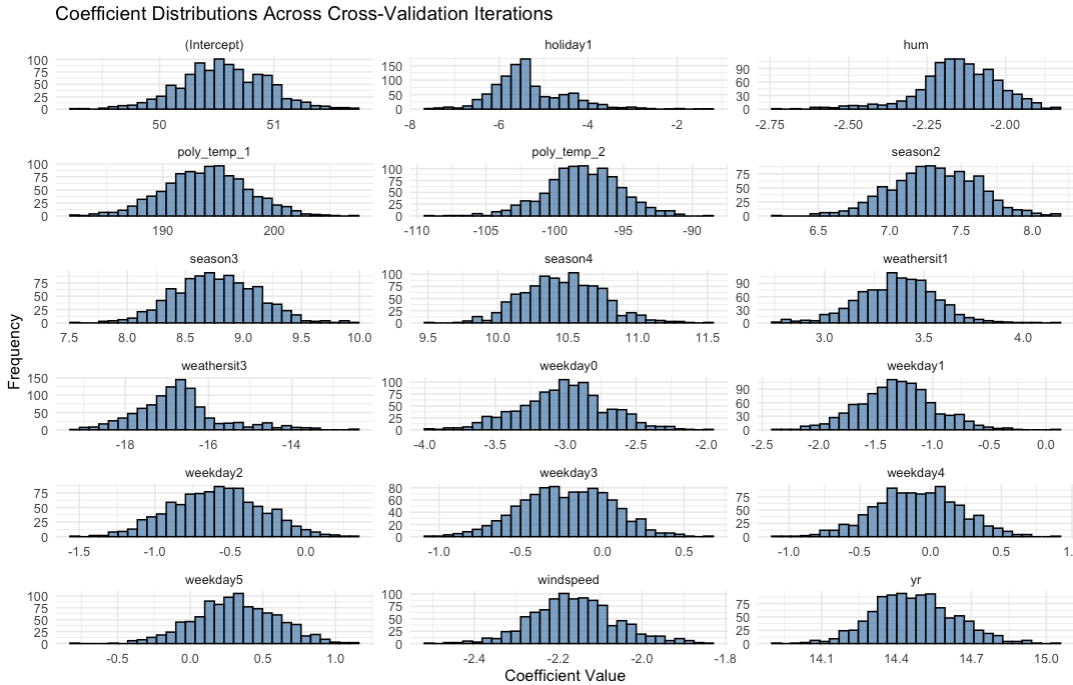


A peak frequency around 25-40 is where most MSP values are concentrated, indicating consistent performance across iterations. A long tail extending beyond 50 represents outliers where the model faced challenges with prediction accuracy

The *MSP* distribution highlights that the model performs consistently for the majority of iterations, with occasional higher errors. The overall results validate the model's robustness and provide insight into its predictive stability across varying data splits.



The mean *MSP* of 35.45 indicates the average squared prediction error across the 1000 iterations of cross-validation. The spread of *MSP* values (min: 14.84, max: 86.22) reflects variability in prediction performance, with most values concentrated between 27.03 (1st quartile) and 39.90 (3rd quartile). The standard deviation of 12.5754 suggests that while most iterations yielded similar prediction errors, a few outliers caused notably higher *MSP* values.



The histograms represent the distribution of coefficient estimates for each predictor variable across 1000 iterations of 10-fold cross-validation. We see that the coefficients of the model seem to be generally be Normally distributed, which is good.

Narrower distributions (e.g., *yr, poly\_temp\_1, season4*) indicate that these coefficients are stable across cross-validation folds and iterations. They have consistent contributions to the model's predictions. Wider distributions (e.g., *weekday3, windspeed*) suggest more variability in these coefficients, potentially due to their lesser importance or interactions with other predictors. Negative coefficients like *hum* and *windspeed* with moderate variability suggest these weather-related predictors consistently reduce bike rentals, which is reasonable as days with extreme humidity and weather conditions are likely to see people take transportation less. The overall consistency of coefficient signs and narrow distributions for key variables validates the robustness of the final model.

**CONCLUSION**

This project ultimately explores the factors influencing bike-sharing demand using regression analysis. Exploratory analysis revealed a quadratic relationship between temperature and bike demand, with peak rentals occurring at moderate temperatures. Clear weather saw the highest bike rentals, while adverse conditions like rain or snow saw the lowest. Initially, the full regression model yielded an *R2* value of 0.8276 and a mean squared residual (*MSRes*) of 812.22. After refining the model by removing insignificant predictors (e.g., month and working day) and addressing multicollinearity (e.g., high correlation between temperature and feels-like temperature), the adjusted model showed significant improvement, with a 4% increase in *R2* and a notable decrease in *MSRes*.

Transformations, such as a square root transformation of the target, bike rental count, and a quadratic transformation of the temperature predictor, enhanced model performance. The final model satisfies assumptions of residual normality and constant variance, demonstrating its reliability for accurately predicting bike-sharing demand. Our refined regression model answered the question: *Can the model accurately predict demand for a specific day and/or weather condition?* by proving to be effective at predicting bike rentals for specific kinds of days and weather conditions. These results provide insights that can be used in the future for optimizing bike availability and improving urban mobility services.

**REFERENCES**

H. Fanaee-T. "Bike Sharing," UCI Machine Learning Repository, 2013. [Online]. Available: https://doi.org/10.24432/C5W894.