Bike Sharing Systems Multiple Regression & Diagnostic Analysis

Steven Broussard

November 22th, 2023

Table of Contents

[1 Introduction 2](#_Toc151589633)

[2 Problem Statement 2](#_Toc151589634)

[3 Exploratory Data Analysis 3](#_Toc151589635)

[3.1 Check Missing Data 3](#_Toc151589636)

[3.2 Cleaning Methods 3](#_Toc151589637)

[3.3 Correlation & Diagnostics/Partitioning The Data 4](#_Toc151589638)

[4 First Model & Insights 5](#_Toc151589639)

[5 Second Model & Third Model 6](#_Toc151589640)

[5.1 Second Model & Linearity 6](#_Toc151589641)

[5.2 Third Model & y-Transformations 6](#_Toc151589642)

[6 Results 7](#_Toc151589643)

[7 Appendix 8](#_Toc151589644)

# 1 Introduction

This research stems from the analysis of multi-regression models utilizing bikeshare datasets, culminating in a strategic recommendation for bike share companies. The findings propose the optimization of future business development initiatives. The study underscores the significance of these variables in predicting bike demand, offering actionable insights that can enhance the precision and efficacy of business predictions in the realm of bike-sharing systems. These systems empower users to effortlessly rent bikes from specific locations and return them at alternate positions, fostering unparalleled convenience. Globally, the proliferation of bike-sharing programs exceeds 500, spanning a vast network of over 500 thousand bicycles. The profound interest in these systems emanates from their pivotal role in mitigating traffic concerns and addressing critical environmental and health issues. As we delve into the intricate dynamics of bike-sharing, our focus narrows to a critical analysis of the data derived from these systems, aiming to unravel insights that can further enhance their impact on business development and predictive modeling.

I explore the potential of leveraging temperature and time of day as influential variables for shaping future initiatives and predicting bike demand. The research underscores the profound significance of these variables in accurately predicting bike demand, providing actionable insights that can elevate the precision and effectiveness of business predictions within the realm of bike-sharing systems. The Bike Share dataset offers intricate data on bike rental demand in Washington, D.C., where individuals utilize the bike-sharing system to rent bikes from specific locations and return them either to the same place or a different location based on their needs—mainly casual users or through membership for regular users. This process is facilitated by a network of automated kiosks distributed across the city. The dataset encompasses hourly rental data spanning two years (2011 and 2012) for the first 19 days of each month, featuring a range of variables that contribute to the comprehensive understanding of bike rental dynamics.

# 2 Problem Statement

Capital Bikeshare stands among the numerous enterprises reshaping global transportation systems. When confronted with decisions aimed at demand for relatively novel infrastructures, harnessing current datasets and employing predictive modeling becomes paramount. Capital Bikeshare strategically tailors marketing approaches, retaining valuable customers by delving into predictive modeling for fresh insights and forecasting future behaviors. My ultimate recommendation, informed by multi-regression models on bikeshare datasets, advises the integration of temperature and hour of the day for forthcoming business development initiatives.

To arrive at this solution, an exhaustive exploration of the dataset. A correlation study among dependent variables further aided in selecting pertinent variables. Subsequently, the modeling iterative process commenced, focusing on predicting the most influential variables in bike rental demand. Multiple diagnostic tests and model modifications were undertaken, guided by key assumptions—namely, independence of observations, linearity, normality, absence of multicollinearity, and, notably, heteroscedasticity. The refined model, featuring temperature and hour of the day, exhibited a multiple R-squared value of 0.77, meeting assumptions post-outlier removal.

The primary objective in developing the multiple linear regression model for Capital Bikeshare was to establish a robust predictive framework capable of discerning and quantifying the key variables that wield the most significant influence on the demand for bike rentals. By delving into this multifaceted analysis, the goal was to equip Capital Bikeshare with a predictive tool that not only identifies the critical drivers of demand but also provides valuable insights for strategic decision-making in optimizing bike-sharing services.

# 3 Exploratory Data Analysis

## 3.1 Check Missing Data

The original dataset held 17 variables and 17,379 observations. In this dataset, each row represents a specific record, and the columns hold different pieces of information. The "Instant" column serves as the record index, allowing for easy reference. "Dteday" represents the date of the record, while "Season" categorizes the season into four types: Winter (1), Spring (2), Summer (3), and Fall (4). The "Yr" column indicates the year (0 for 2011, 1 for 2012), and "Mnth" denotes the month (ranging from 1 to 12). "Hr" signifies the hour of the day (0 to 23). The dataset includes information on whether the day is a holiday ("Holiday"), the day of the week ("Weekday"), and whether it is a working day ("Workingday"). The "Weather" column classifies weather conditions into four categories. Temperature-related variables include "Temp" for hourly temperature in Celsius (normalized) and "Atemp," representing the "feels like" temperature. "Hum" stands for relative humidity, while "Windspeed" indicates wind speed (both normalized). The dataset also captures the number of registered users ("Registered"), casual users ("Casual"), and the total count of rentals ("Count"). Each variable's definition helps users interpret and analyze the dataset effectively.

The primary step necessitated a thorough examination of the dataset to identify any potential missing, duplicate, or erroneous values. Fortunately, the data did not present any instances of missing values. However, if such gaps were present, an in-depth analysis would have been conducted to determine their nature—whether they were missing completely at random, missing at random, or missing not at random, wherein the absence of values depended on other variables.

## 3.2 Cleaning Methods

Upon embarking on the analysis of the subsequent data retrieved from the Excel file, the primary focus was to look at which numeric values necessitated conversion into categorical variables for optimal model compatibility. Within the dataset comprising 17 variables, a selection was made, resulting in the identification of 8 variables ready for transition from numeric to categorical status. These pivotal variables encompassed season, year, month, hour, holiday, weekday, working day, and weathersit. To enhance interpretability and align with conventional naming conventions, each variable was categorized accordingly. For instance, the "season" variable assumed values from 1 to 4, representing "Winter," "Spring," "Summer," and "Fall," respectively. Employing this approach, all 8 variables were seamlessly transformed into categorical counterparts.

During this analysis, a discerning observation led to the identification of redundancy in the variable’s "day" and "instant," with the latter serving as an index. Consequently, a strategic decision was made to remove these redundant variables from the new dataset, resulting in a refined dataset comprising 13 variables while retaining 17,379 observations.

The subsequent step involved the creation of dummy variables for the categorical variables. The 8 numeric variables that underwent encoding were subject to the dummy encoding process, culminating in the formation of a new dataset ready for comprehensive analysis and eventual utilization for my linear regression model. This new bike-sharing dataset now had 56 variables, with the initial 8 variables serving as the first column being systematically omitted. This meticulous process lays the foundation for a robust analytical framework, ensuring that my subsequent linear regression model is ready and curated for an optimal predictive modeling outcome.

## 3.3 Correlation & Diagnostics/Partitioning The Data

A graph of a graph

Description automatically generatedIn the process of transforming numerical values into categorical variables within the training data, a meticulous examination of the correlation among each data point in the training set was undertaken. The overarching objective of this scrutiny was to delve into the intricacies of the relationships inherent within the set of numerical variables. As visually represented in ***Figure 1***, variables positioned in the middle of the bar graph are construed as less significant and probably unrelated to the number of registered bikes. On the contrary, the presence of larger bars, each associated with numerical values, signifies potential correlations with the target variable, namely, "bikes rented." Worth noting is that a substantial portion of these variables manifests as dummy variables, specifically those pertaining to "weekday," "hour," and "month." These variables, when grouped together, exhibit a lower likelihood of correlation with "cnt." To further refine the model and mitigate multicollinearity, certain variables such as "registered" and "casual" were omitted. These variables demonstrated a pronounced correlation with "cnt" due to the inherent grouping of bike rentals. Following this refinement, the train-test split method was strategically employed for its expeditious execution, facilitating a prompt and insightful comparison of machine learning

***Figure 1 :*** Correlation Bar Plot of each continuous data point in the dataset

algorithm effectiveness. The dataset was meticulously partitioned, with 70 percent designated for training and 30 percent for rigorous testing purposes, ensuring a robust evaluation of predictive modeling outcomes.

# 4 First Model & Insights

A graph of a diagram

Description automatically generatedA close-up of a number

Description automatically generatedApproaching the initiation of our model, the set of 56 variables were ready for analysis. I opted to employ a simple linear regression model as a preliminary approach. This initial exploration yielded great results, with approximately 33 variables demonstrating significance concerning the total count of registered bikes, regarded as "cnt." Given the desire to streamline our model and focus on variables exhibiting high correlation with the predictive variable, I started analyzing the variables.

***Figure 2:*** Model 1 R Squared and Adjusted R Squared

***Figure 3:*** Histogram of Residuals in Model 1

The variables chosen to propel my model forward included temperature, humidity, season, year, and hour. These variables stood out as the most highly correlated within the dataset, setting the stage for a more refined and targeted analysis. To delve deeper into the model's performance, considerations of linearity, homoscedasticity, and distribution were meticulously examined, laying the groundwork for the subsequent evolution into Model 2.

Upon reviewing the R-squared and adjusted R-squared values, I realized that it was registering at 0.6772 and 0.6764, respectively, as illustrated in ***Figure 2***. My analysis gauged the model's explanatory power and its ability to accommodate the underlying dataset's complexity. As I delved into the normal distribution aspect of the model, an unexpected revelation surfaced – the model exhibited normality, albeit with a subtle leftward skewness indicative of a positive skew, as represented in ***Figure 3***. This examination of the model's characteristics serves as a pivotal foundation for guiding the trajectory toward subsequent model iterations.

# 5 Second Model & Third Model

## 5.1 Second Model & Linearity

A graph of a graph showing a line

Description automatically generated with medium confidenceA black and white diagram with numbers and a red line

Description automatically generatedIn the transition to my second model iteration, a pivotal consideration was the imperative to mitigate multicollinearity and eliminate variables that were potentially contributing to the curvature observed in my linearity graph (Residual vs Fitted). As vividly portrayed in ***Figure 4***, this curve was notably pronounced at both ends of the model, prompting a closer look. It became evident that extreme variability in the numerical values at the model's extremes might be introducing distortions in the predictive accuracy, leading to the observed curve.

***Figure 4:*** Linearity of Model 1

Taking proactive measures to enhance the model, I opted to strategically remove the variables "year" and "season," recognizing their potential impact on the model's linearity. The transformation in the linearity dynamics is strikingly evident in ***Figure 5***. With the exclusion of these variables, the linearity graph underwent an alteration, represented by the red line, which now closely aligns with the dotted line in the middle of the graph. This adjustment not only contributes to the model's overall coherence but also sets the stage for a more refined predictive framework. As we go through the model refinement, such discerning modifications play a crucial role in shaping the model's accuracy and predictive efficacy.

***Figure 5:*** Linearity of Model 2

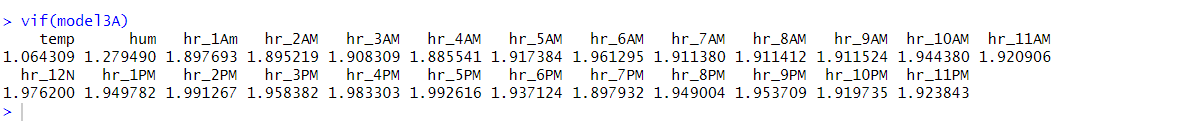
## 5.2 Third Model & y-Transformations

When it came to the point to transform my y-axis, I ran through 3 y-transformations methodologies: log, sqrt, and cube root. I was testing the linearity, homoscedastic, distribution and skewedness. I ran through these models with the following variables: temperature, humidity, and hours. Upon conducting the Breusch-Pagan test, it was reassuring to observe that the models not only satisfied the homoscedasticity assumption, as indicated by a p-value exceeding 0.05 but also yielded me with the best R squared and adjusted R squared value and it also yielded me the most fitted line when plotting it on the Residuals vs Fitted line. You will see the progress from ***Figure 3***, when I first plotted the A graph of black dots

Description automatically generatedgraph, versus now, in ***Figure 6***. This linearity assumption holds a lot of significance. It presupposes a linear relationship between the parameters of independent variables and the dependent variable (Y). Delving into the residual vs. fitted values plot, depicted in ***Figure 6*** of Model 3, the horizontal alignment of the red line at zero corroborates the model's adherence to linearity. This validation process ensures that our model aligns with fundamental assumptions, instilling confidence in its accuracy and efficacy in predicting bike rental demand for Capital Bikeshare. Additionally, My R squared for the logarithmic transformation was now .7718 vs when I first started the model, in ***Figure 1***, and it was .6764. This signifies the reliability of my regression models.

***Figure 6:*** New Linearity from Model 3

Model 3 has demonstrated homoscedasticity and improved fitting, to a meticulous multicollinearity examination. The VIF serves as a crucial metric, denoting the absence of correlation between the variables in the regression model and after running this code in my regression model – all of my variables below the threshold of 3 as seen in ***Figure 7.***



***Figure 7:*** VIF from Model 3; everything is under 3

# 6 Results

After drawing insights from the multi-regression models my conclusive recommendation to the esteemed bike share company centers on the variables of temperature, hour, and humidity as pivotal components for steering future business development initiatives. The discerned significance of these variables lies in the company’s potential for invaluable insights and the overall landscape of bike demand. This strategic integration of predictive variables stands as a testament to our commitment to delivering forward-thinking solutions, creating the symbiotic relationship between data-driven foresight and the strategic evolution of bike-sharing systems. As we embark on the nexus of data analytics and industry foresight, our recommendation shows how your company can grow.

# 7 Appendix

***Figure 1)***

Code: # Calculating correlations of continuous variables with SalePrice

continuous\_vars <- bikedf2 %>% select\_if(is.numeric)

correlations <- cor(continuous\_vars)

cnt\_correlations <- correlations['cnt',]

cnt\_correlations\_df <- data.frame(Variable = names(cnt\_correlations),

Correlation = cnt\_correlations) %>%

arrange(desc(Correlation))

# Plotting these correlations

ggplot(data = cnt\_correlations\_df, aes(x = reorder(Variable, Correlation), y = Correlation)) +

geom\_bar(stat = "identity") +

coord\_flip() +

labs(title = "Correlation of Continuous Variables with Cnt", x = "", y = "Correlation") +

theme\_minimal()

Picture: A graph of a graph

Description automatically generated

***Figure 2)***

Code: model1 <- lm(cnt ~ +temp+hum+season\_Spring+season\_Summer+season\_Fall+yr\_2012+hr\_1Am+hr\_2AM+hr\_3AM+hr\_4AM+hr\_5AM+hr\_6AM+hr\_7AM+hr\_8AM+hr\_9AM+hr\_10AM+hr\_11AM+hr\_12N+hr\_1PM+hr\_2PM+hr\_3PM+hr\_4PM+hr\_5PM+hr\_6PM+hr\_7PM+hr\_8PM+hr\_9PM+hr\_10PM+hr\_11PM, data = train) # Land.Slope

summary(model1)

A close-up of a number

Description automatically generatedPicture:

***Figure 3)***

Code: CheckNormal=function(model){

hist(model$residuals,breaks=30)

shaptest=shapiro.test(sample(model$residuals,4000))

print(shaptest)

skewness(model$residuals)

if(shaptest$p.value==.05){

print("H0 rejected : the residuals are Not distributed normally")

} else{

print("H0 failed to reject : the residuals Are distributed normally")

}

}

CheckNormal(model= model1)

Picture: A graph of a diagram

Description automatically generated

***Figure 4)***

Code: plot(model1)

Picture:

A graph of a graph showing a line

Description automatically generated with medium confidence

***Figure 5)***

Code: model2 <- lm(cnt ~ +temp+hum+hr\_1Am+hr\_2AM+hr\_3AM+hr\_4AM+hr\_5AM+hr\_6AM+hr\_7AM+hr\_8AM+hr\_9AM+hr\_10AM+hr\_11AM+hr\_12N+hr\_1PM+hr\_2PM+hr\_3PM+hr\_4PM+hr\_5PM+hr\_6PM+hr\_7PM+hr\_8PM+hr\_9PM+hr\_10PM+hr\_11PM, data = train)

summary(model2)

plot(model2)

Picture: A black and white diagram with numbers and a red line

Description automatically generated

***Figure 6)***

Code: model3A=lm(log(cnt)~temp+hum+hr\_1Am+hr\_2AM+hr\_3AM+hr\_4AM+hr\_5AM+hr\_6AM+hr\_7AM+hr\_8AM+hr\_9AM+hr\_10AM+hr\_11AM+hr\_12N+hr\_1PM+hr\_2PM+hr\_3PM+hr\_4PM+hr\_5PM+hr\_6PM+hr\_7PM+hr\_8PM+hr\_9PM+hr\_10PM+hr\_11PM,data=train)

summary(model3A)

CheckNormal(model3A)

skewness(model3A$residuals)

plot(model3A)

Picture: A graph of black dots

Description automatically generated

***Figure 7)***

Code:vif(model3A)

Picture: 