# Rental Bike Demand Analysis

William Lei, Irene Na, Safi Aharoni, Kelly Short

## Contents

# 1 Introduction

Rental bikes are an essential transportation service in many cities, supporting both work commutes and leisure activities. One city in which rental bikes are a key public service is Seoul, South Korea. It is important for the city to ensure rental bike availability and accessibility, lessening wait times during peak demand. To do so, the city needs to be able to predict how many rental bikes need to be available on different days and at different times. In this report, we aim to provide decision makers with insights into the factors influencing rental bike usage and demand, facilitating availability for the public by controlling for relevant factors such as weather conditions, seasonality, and time of day. We hypothesize that these factors, particularly temperature, season, and time of day, significantly impact rental bike demand, with specific conditions like higher temperatures and rush hours leading to increased usage.

# 2 Data

## 2.1 Description

The dataset contains the count of public bicycles rented per hour in the Seoul Bike Sharing System, with corresponding weather data and holiday information The data consists of 8760 entries, and 13 features in total. Unfortunately, the data collection and sampling process is not explicitly detailed in the provided dataset documentation, which limits our understanding of potential biases or limitations in the data.

The features include Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall, the number of bikes rented per hour and date information. In this report we used different regression models to explore the connections between hourly bike demand and various weather conditions to predict the bike demand in a statistically robust way.

## 2.2 EDA

Below is our approach for exploratory data analysis. For numerical/metric variables, including the response variable, we explore both their distributions and summary statistics. First, we found that the response variable is right skewed, with a wider span than its mean. Second, the variables 'Visibility', 'Solar_radiation', 'Rainfall', and 'Snowfall', by their nature, have relatively sparse and skewed distributions Third, other metric variables, such as 'Temperature', are relatively well-behaved, with fewer significant outliers and less skew. Table 1 shows summary statistics for a selection of the metric variables. For categorical variables (namely 'Hour', 'Seasons', 'Holiday', and 'Functioning_day'), we looked at their counts and noted that except for certain hours (such as 5am) and non-functioning days, all groups are relatively well represented. In pre-processing, we found no missing values (NA) in the dataset, which simplified the pre-processing.

Table 1: Summary Statistics of Selective Numerical Variables

| Variable | Observations | Mean | Std. Dev | Median |
|---|---|---|---|---|
| Rented_bike_count(Y) | 8760 | 705 | 645 | 504 |
| Temperature | 8760 | 13 | 12 | 14 |
| Dew_point_temp | 8760 | 4.1 | 13 | 5.1 |
| Solar_radiation | 8760 | 0.57 | 0.87 | 0.01 |
| Visibility | 8760 | 1437 | 608 | 1698 |
| Humidity | 8760 | 58 | 20 | 57 |

## 2.3 Data visualization - numerical/metric variables

We use Pearson correlation heatmap to examine how numerical variables are correlated with rental bike counts, and how they are correlated with each other. Further, we also examine the pair-wise relationship

between rented bike count (Y) and numerical variables through a scatter plot to discern any obvious non-linear relationships.

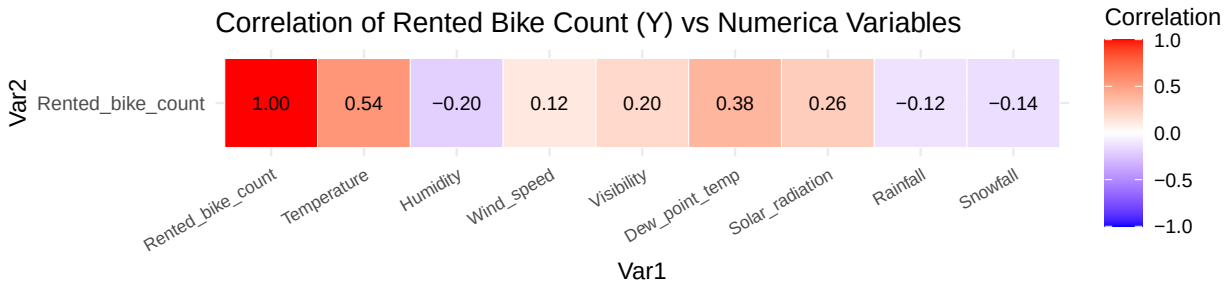



Figure 1: Correlation of Rented Bike Count (Y) vs Numerica Variables

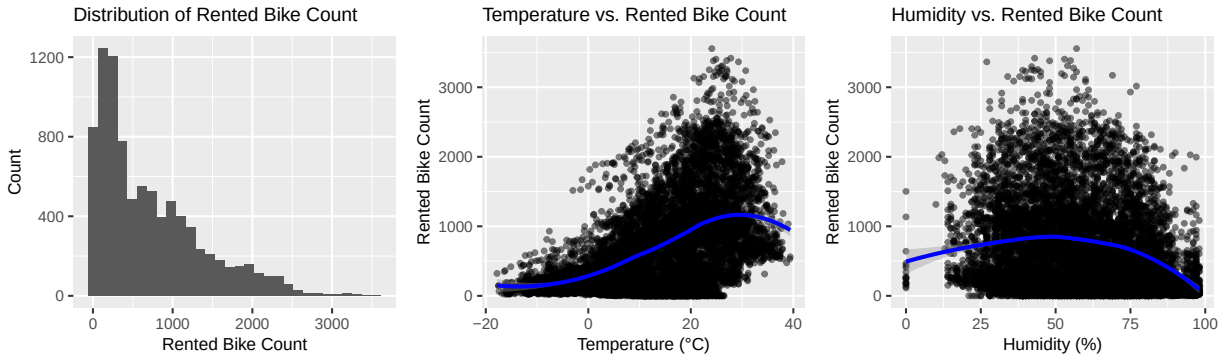



Figure 2: Scatter plot - Rented Bike Count (Y) vs Variables with Potential Non-linear Relationship

The key obervations from the correlation check (**Figure 1**) and scatter plot (**Figure 2**) for numerical variables are: **1.** Temperature, dew point temperature, solar radiation, visibility, and humidity appear to have higher correlation with rental bike count than other variables (absolute correlation >=20%). **2.** Temperature and dew point temperature are highly correlated (~90%), and therefore we consider dropping dew point temperature as a variable in our modeling to avoid serious multi-collinearity. **3.** We also note from the scatter plots that temperature and humidity may have a nonlinear relationship with rented bike count. **4.** The wide rented bike count dispersion compared to its mean may cause issues for a Poisson regression fit.

## 2.4 Data visualization - categorical/factor variables

We use boxplots to examine how each factor variable is correlated with rental bike counts (**Figure 3**).

The key observations from the boxplots analysis regarding the relationship between factor variables and rented bike count are:

**1.** We see visibly notable differences in median rented bike counts among different levels for all four factor variables. Therefore we believe that it is reasonable to include all four of them in our regression analysis.

**2.** More specifically, we note that hour and seasons show more notable level wise differences and less overlapping dispersion. For the holiday variable, the rented bike count median difference by level is notable but the level wise dispersion is highly overlapping, which suggests the holiday variable may be less significant in later analysis.

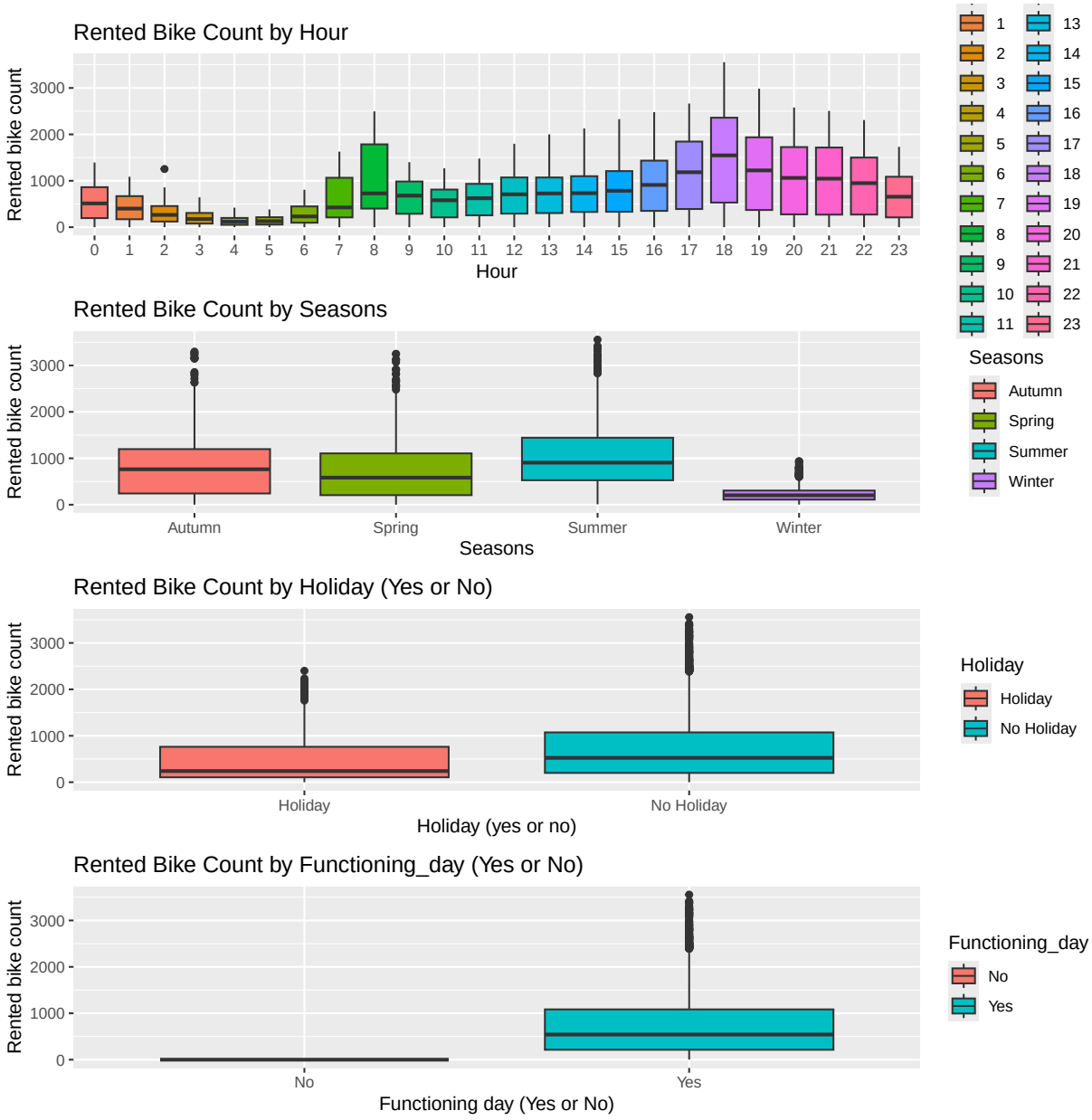



Figure 3: Boxplots of Rented Bike Count (Y) vs Categorical Variables

**3.** We also note that when functioning_day is 'No', rented bike counts are always zero. That means when functioning day is 'No', we could learn little about the demand from the dataset. Thus we decided to only focus on the entries with functioning_day being 'Yes', which reduced the observationas from 8760 to 8465.

# 3 Model Development

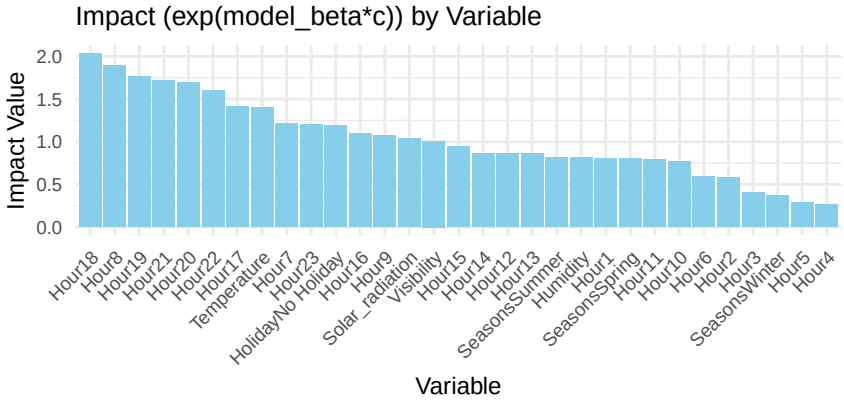
## 3.1 Poisson regression

Given the rented bike count for each hour is a count variable, our initial hypothesis is that it would be appropriate to model with a Poisson regression.Based on the above EDA, we build the initial poisson model with the following variables: temperature, solar radiation, humidity, visibility, hour, season, holiday. To start, we include only linear terms for each of these explanatory variables.

```
mod1.poisson <- glm(Rented_bike_count ~ Temperature + Solar_radiation + Humidity + Visibility
                    + Hour + Seasons + Holiday, family = poisson(link=log),
                    data = rentb_used)
# summary(mod1.poisson)
```

```
##
## Model1 Anova LR test results:

## Analysis of Deviance Table (Type II tests)
##
## Response: Rented_bike_count
##                  LR Chisq Df Pr(>Chisq)
## Temperature       130213  1  < 2.2e-16 ***
## Solar_radiation     1958  1  < 2.2e-16 ***
## Humidity           98776  1  < 2.2e-16 ***
## Visibility           244  1  < 2.2e-16 ***
## Hour             1063105 23  < 2.2e-16 ***
## Seasons           316425  3  < 2.2e-16 ***
## Holiday             6364  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




From the Model 1 summary and LR test results, we note that most of the variables that we selected are statistically significant with a high confidence level (99.9%), with both the Wald-test and LR chi-square test agreeing.

In terms of practical significance, we examine the magnitude of each variable on rented bike count. More specifically, we examine the impact on rented bike count for one standard deviation change for each explanatory variable, holding everything else constant. We noted that evening hours and morning hours have the biggest impact on bike count, such as Hour18 increasing the bike demand by 1.03x (vs Hour24, the reference hour), and Hour8 increasing the bike demand by 89% (vs Hour24), all else equal. This makes sense, given those are rush hour times when people go to and get off work. Other than that, among the metric variables, temperature also has a sizable impact: every 12 degree increase in temperature is associated with a 41% increase in bike demand, all else equal.

## 3.2 Model Comparison

In the second model, we add the variable Wind_speed to capture potential weather-related effects, as well as Snowfall and Rainfall, which are both weather factors that can influence the likelihood of people renting bikes. These variables were selected because logically speaking, they should have some impact on bike demand at a given hour due to their direct influence on comfort and safety when cycling.

The third model introduces non-linear and interaction terms. From our previous exploratory analysis, we observed potential non-linear relationships between Temperature and Humidity with rented bike count. Therefore, we included their quadratic terms (I(Temperature^2) and I(Humidity^2)) to capture these effects. Additionally, we introduced an interaction term between Temperature and Humidity because these two factors can jointly influence the perceived comfort for outdoor activities. Additionally, the interaction between Holiday and Hour is included in the model to capture how the effect of the time of day on bike rentals changes based on whether the day is a holiday or not. Without this interaction term, the model assumes that the effect of Hour on bike rentals is the same on both holidays and non-holidays.

We conducted LR tests (Anova) for both Model2 and Model3, which show statistical significance for all parameters at the 99.9% confidence level. Due to limited page space, we do not show them one by one here.

Next, we compare all three models (**Table 2**). The key observation are:

**1.** Comparing the three models' results, we note that all three models show significance for most of their parameter coefficients, with a few exceptions in the third model: **1.1** Select hours and the interaction terms for those 'hour' variables are not significant (e.g., Hour 11 and Hour16). However, the fact that the majority of the hours show meaningful significance supports the conclusion that 'hour' as a variable is significant. **1.2** The 'Holiday' dummy loses its significance after we include the holiday-hour interactions in Model 3. This suggests that the significance of the holiday dummy is mainly through its interaction effect with hours.

**2.** We also considered three information criteria (AIC, AIC_corrected and BIC) to compare the three models. As a rule of thumb, the lower the information criteria, the better the model. Although we know that BIC tends to favor simpler models and AIC tends to favor more sophisticated models, all three criteria decrease as we move from Model 1 to Model 3. Based on the AIC, AICc, and BIC values, Model 3 is the preferred model with the best trade-off between model complexity and goodness of fit.

We further compare the three models using an LR test (anova) to see if adding sophistication as a whole results in a statistically significant improvement in the modeling.

```
## Analysis of Deviance Table
##
## Model 1: Rented_bike_count ~ Temperature + Solar_radiation + Humidity +
##     Visibility + Hour + Seasons + Holiday
## Model 2: Rented_bike_count ~ Temperature + Solar_radiation + Humidity +
##     Visibility + Hour + Seasons + Holiday + Wind_speed + Snowfall +
##     Rainfall
## Model 3: Rented_bike_count ~ Temperature + Solar_radiation + Humidity +
##     Visibility + Hour + Seasons + Holiday + Wind_speed + Snowfall +
##     Rainfall + I(Temperature^2) + I(Humidity^2) + Temperature:Humidity +
```

Table 2: Rented Bike Count Analysis - Poisson Regresion

| | *Dependent variable:* | | |
|---|---|---|---|
| | Rented_bike_count | | |
| | (1) | (2) | (3) |
| Constant | 6.693*** (0.004) | 6.570*** (0.004) | 5.823*** (0.012) |
| Temperature | 0.028*** (0.0001) | 0.027*** (0.0001) | 0.072*** (0.0002) |
| Solar_radiation | 0.043*** (0.001) | 0.031*** (0.001) | 0.090*** (0.001) |
| Humidity | −0.010*** (0.00003) | −0.007*** (0.00003) | 0.016*** (0.0001) |
| Visibility | 0.00001*** (0.00000) | −0.00000*** (0.00000) | 0.00002*** (0.00000) |
| Hour1 | −0.215*** (0.003) | −0.229*** (0.003) | −0.158*** (0.016) |
| Hour2 | −0.539*** (0.004) | −0.550*** (0.004) | −0.378*** (0.017) |
| Hour3 | −0.914*** (0.004) | −0.915*** (0.004) | −0.694*** (0.020) |
| Hour4 | −1.327*** (0.005) | −1.343*** (0.005) | −1.178*** (0.023) |
| Hour5 | −1.263*** (0.005) | −1.288*** (0.005) | −1.528*** (0.027) |
| Hour6 | −0.523*** (0.004) | −0.535*** (0.004) | −1.105*** (0.023) |
| Hour7 | 0.194*** (0.003) | 0.182*** (0.003) | −0.678*** (0.019) |
| Hour8 | 0.639*** (0.003) | 0.645*** (0.003) | −0.323*** (0.017) |
| Hour9 | 0.065*** (0.003) | 0.118*** (0.003) | −0.363*** (0.016) |
| Hour10 | −0.257*** (0.003) | −0.192*** (0.003) | −0.373*** (0.016) |
| Hour11 | −0.229*** (0.004) | −0.144*** (0.004) | −0.261*** (0.015) |
| Hour12 | −0.152*** (0.004) | −0.040*** (0.004) | −0.084*** (0.015) |
| Hour13 | −0.155*** (0.004) | −0.037*** (0.004) | 0.007 (0.014) |
| Hour14 | −0.146*** (0.003) | −0.028*** (0.003) | 0.051*** (0.014) |
| Hour15 | −0.057*** (0.003) | 0.064*** (0.003) | 0.125*** (0.014) |
| Hour16 | 0.088*** (0.003) | 0.195*** (0.003) | 0.230*** (0.014) |
| Hour17 | 0.349*** (0.003) | 0.454*** (0.003) | 0.339*** (0.014) |
| Hour18 | 0.708*** (0.003) | 0.794*** (0.003) | 0.390*** (0.014) |
| Hour19 | 0.571*** (0.003) | 0.634*** (0.003) | 0.355*** (0.014) |
| Hour20 | 0.531*** (0.003) | 0.575*** (0.003) | 0.318*** (0.014) |
| Hour21 | 0.543*** (0.003) | 0.577*** (0.003) | 0.303*** (0.014) |
| Hour22 | 0.469*** (0.003) | 0.483*** (0.003) | 0.201*** (0.015) |
| Hour23 | 0.182*** (0.003) | 0.193*** (0.003) | −0.040** (0.016) |
| SeasonsSpring | −0.217*** (0.001) | −0.197*** (0.001) | −0.191*** (0.001) |
| SeasonsSummer | −0.203*** (0.001) | −0.185*** (0.001) | −0.014*** (0.001) |
| SeasonsWinter | −0.996*** (0.002) | −0.987*** (0.002) | −0.549*** (0.002) |
| HolidayNo Holiday | 0.172*** (0.002) | 0.168*** (0.002) | 0.014 (0.011) |
| Wind_speed | | −0.016*** (0.0005) | −0.015*** (0.0005) |
| Snowfall | | −0.138*** (0.002) | −0.023*** (0.002) |
| Rainfall | | −0.567*** (0.002) | −0.453*** (0.002) |
| I(Temperature^2) | | | −0.002*** (0.00000) |
| I(Humidity^2) | | | −0.0002*** (0.00000) |
| Temperature:Humidity | | | 0.0002*** (0.00000) |
| Hour1:HolidayNo Holiday | | | −0.065*** (0.017) |
| Hour2:HolidayNo Holiday | | | −0.161*** (0.018) |
| Hour3:HolidayNo Holiday | | | −0.204*** (0.020) |
| Hour4:HolidayNo Holiday | | | −0.129*** (0.024) |
| Hour5:HolidayNo Holiday | | | 0.305*** (0.028) |
| Hour6:HolidayNo Holiday | | | 0.640*** (0.024) |
| Hour7:HolidayNo Holiday | | | 0.922*** (0.020) |
| Hour8:HolidayNo Holiday | | | 0.978*** (0.017) |
| Hour9:HolidayNo Holiday | | | 0.431*** (0.017) |
| Hour10:HolidayNo Holiday | | | 0.099*** (0.016) |
| Hour11:HolidayNo Holiday | | | 0.027 (0.015) |
| Hour12:HolidayNo Holiday | | | −0.041** (0.015) |
| Hour13:HolidayNo Holiday | | | −0.107*** (0.014) |
| Hour14:HolidayNo Holiday | | | −0.115*** (0.014) |
| Hour15:HolidayNo Holiday | | | −0.072*** (0.014) |
| Hour16:HolidayNo Holiday | | | −0.025 (0.014) |
| Hour17:HolidayNo Holiday | | | 0.137*** (0.014) |
| Hour18:HolidayNo Holiday | | | 0.438*** (0.014) |
| Hour19:HolidayNo Holiday | | | 0.291*** (0.014) |
| Hour20:HolidayNo Holiday | | | 0.259*** (0.014) |
| Hour21:HolidayNo Holiday | | | 0.271*** (0.014) |
| Hour22:HolidayNo Holiday | | | 0.285*** (0.015) |
| Hour23:HolidayNo Holiday | | | 0.237*** (0.016) |
| AIC | 1233384.13 | 1066086.74 | 878336.08 |
| Corrected AIC | 1233384.38 | 1066087.04 | 878336.98 |
| BIC | 1233609.53 | 1066333.27 | 878765.74 |
| Observations | 8,465 | 8,465 | 8,465 |
| Log Likelihood | −616,660.100 | −533,008.400 | −439,107.000 |

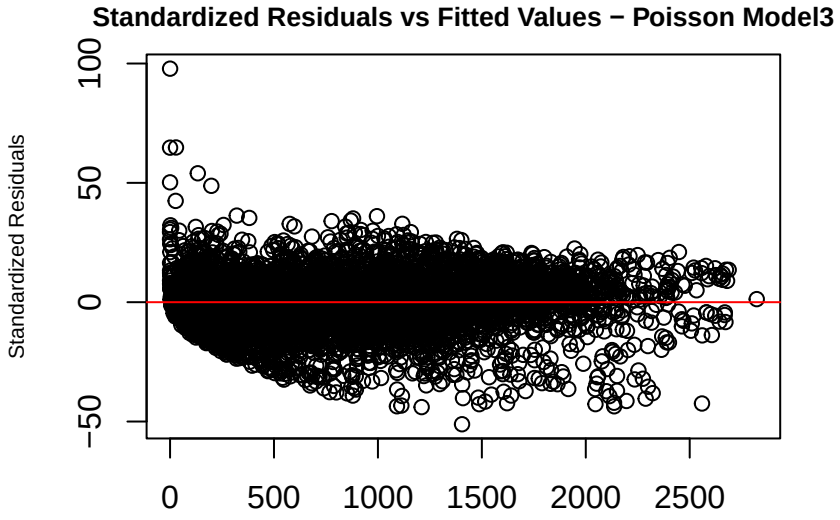
*Note:* *p<0.05; **p<0.01; ***p<0.001

```
##     Holiday:Hour
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      8433    1166226
## 2      8430     998923  3   167303 < 2.2e-16 ***
## 3      8404     811120 26   187803 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We conducted anova tests to compare each model sequentially from Model 1 to Model 3. The results of these tests indicate that each increase in complexity, from Model 1 to Model 2, and from Model 2 to Model 3, adds a statistically significant improvement to the overall fit of the model. This statistical evidence, together with the comparison of the AIC and BIC values earlier, consistently ranks Model 3 as the best, further supporting our decision to use that as the preferred poisson model.

### 3.3 Model Assessment

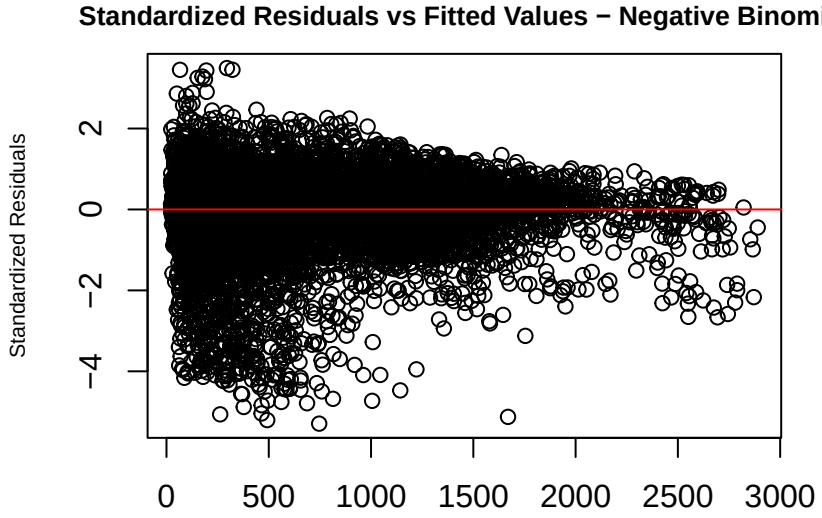
To examine the model fit for Model 3, we constructed the standard residual versus fitted value test to examine any outliers or overdispersion issues.




```
##
##  Overdispersion test
##
## data:  mod3.poisson
## z = 1.0369, p-value = 0.1499
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   53462.25
```

Observation/note: The residual plot for Model 3 reveals several issues. First, there is evidence of heteroscedasticity, as the spread of residuals varies with fitted values, indicating non-constant variance. This violates a key assumption of Poisson regression. Additionally, some significant outliers are visible, which suggests that certain observations may disproportionately influence the model's results. Further investigation shows that the skewness in rainfall is the main reason for outliers. Finally, besides outliers, the slight

skew in the residuals suggests that the error distribution is not perfectly symmetrical. To improve model fit, we decided to drop rainfall from the model to improve outlier issue, and make other model adjustments such as employing a quasi-Poisson or Negative Binomial regression, to attempt to address overdispersion.




```
##      Poisson_Model_3 Negative_Binomial_Model
## AIC         878336.1                116649.5
## AICc        878337.0                116650.4
## BIC         878765.7                117079.2
```

The Negative Binomial model (mod.nb) leads to meaningfully lower values for AIC, AICc, and BIC compared to the Poisson model (mod3.poisson). This suggests that the Negative Binomial model is a better fit for the data, likely due to its ability to handle overdispersion, which was evident previously in the Poisson model.

The standardized residual plot for the Negative Binomial model also shows a better distribution of residuals, which indicates that the overdispersion issue seen in the Poisson model is improved.

## 3.4 Alternative Specification

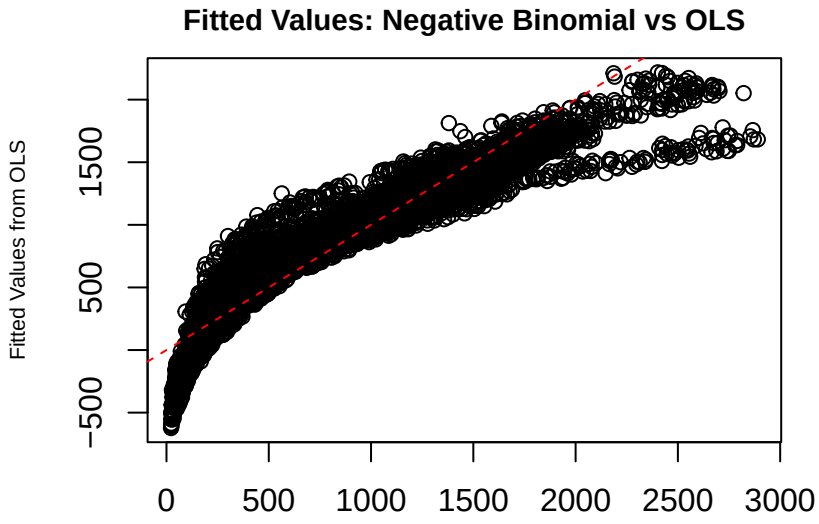
To offer a different perspective, we also attempt to fit the data with an alternative model family: OLS. We inherit all the parameter choices from Poisson regression Model 3 above, and refit the model with OLS instead. We then compare the two models side by side in terms of variable significance as well as overall fit.

The model summary (**Table 3**) shows that both the Negative Binomial and OLS models offer a very different interpretation in terms of parameter significance compared to the Poisson regression. For example, wind_speed, visibility, the hour dummies, and the holiday dummy show little significance in the OLS model, whereas the Negative Binomial suggests wind_speed is important but visibility and snowfall are not. We also note that both the Negative Binomial and OLS AIC and BIC are much lower than those results for the Poisson regression. The Negative Binomial model has the lowest information criteria overall.

In order to compare the fit between the two models, we constructed the fitted value versus residual comparison for the two models.

Table 3: Rented Bike Count Analysis - Poisson Model 3 vs OLS Regresion

| | *Dependent variable:* | |
|---|---|---|
| | Rented_bike_count | |
| | *negative binomial* | *OLS* |
| | (1) | (2) |
| Constant | 5.571*** (0.138) | 233.122* (99.185) |
| Temperature | 0.057*** (0.002) | 37.557*** (1.594) |
| Solar_radiation | 0.114*** (0.015) | 58.756*** (10.548) |
| Humidity | 0.031*** (0.002) | 17.132*** (1.129) |
| Visibility | 0.00001 (0.00001) | 0.015 (0.009) |
| Hour1 | −0.041 (0.175) | −61.229 (125.422) |
| Hour2 | −0.211 (0.175) | −133.462 (125.426) |
| Hour3 | −0.537** (0.175) | −216.723 (125.448) |
| Hour4 | −1.058*** (0.176) | −295.374* (125.444) |
| Hour5 | −1.392*** (0.177) | −331.642** (125.443) |
| Hour6 | −1.008*** (0.176) | −277.075* (125.445) |
| Hour7 | −0.595*** (0.175) | −207.795 (125.442) |
| Hour8 | −0.212 (0.175) | −138.928 (125.477) |
| Hour9 | −0.359* (0.175) | −182.893 (125.635) |
| Hour10 | −0.446* (0.176) | −229.128 (126.013) |
| Hour11 | −0.344 (0.176) | −208.647 (126.420) |
| Hour12 | −0.184 (0.176) | −128.997 (126.613) |
| Hour13 | −0.082 (0.176) | −71.614 (126.758) |
| Hour14 | −0.019 (0.176) | −45.082 (126.687) |
| Hour15 | 0.037 (0.176) | −14.393 (126.401) |
| Hour16 | 0.120 (0.175) | 67.671 (126.036) |
| Hour17 | 0.218 (0.175) | 161.757 (125.789) |
| Hour18 | 0.259 (0.175) | 200.041 (125.602) |
| Hour19 | 0.158 (0.175) | 153.348 (125.520) |
| Hour20 | 0.167 (0.175) | 153.679 (125.482) |
| Hour21 | 0.184 (0.175) | 139.470 (125.477) |
| Hour22 | 0.100 (0.175) | 72.546 (125.463) |
| Hour23 | −0.121 (0.175) | −64.036 (125.460) |
| SeasonsSpring | −0.251*** (0.017) | −146.111*** (12.079) |
| SeasonsSummer | −0.038 (0.023) | −76.545*** (16.743) |
| SeasonsWinter | −0.600*** (0.025) | −314.419*** (18.088) |
| HolidayNo Holiday | 0.106 (0.127) | −23.247 (90.983) |
| Wind_speed | −0.026*** (0.006) | 4.640 (4.591) |
| Snowfall | −0.013 (0.014) | 29.046** (9.857) |
| I(Temperature^2) | −0.001*** (0.0001) | −0.262*** (0.037) |
| I(Humidity^2) | −0.0004*** (0.00001) | −0.193*** (0.010) |
| Temperature:Humidity | 0.0001** (0.00003) | −0.171*** (0.023) |
| Hour1:HolidayNo Holiday | −0.126 (0.179) | −41.117 (128.568) |
| Hour2:HolidayNo Holiday | −0.260 (0.179) | −77.975 (128.573) |
| Hour3:HolidayNo Holiday | −0.308 (0.180) | −75.785 (128.593) |
| Hour4:HolidayNo Holiday | −0.179 (0.181) | −54.450 (128.584) |
| Hour5:HolidayNo Holiday | 0.253 (0.181) | 7.916 (128.588) |
| Hour6:HolidayNo Holiday | 0.601*** (0.180) | 122.176 (128.585) |
| Hour7:HolidayNo Holiday | 0.942*** (0.180) | 374.277** (128.573) |
| Hour8:HolidayNo Holiday | 1.021*** (0.179) | 676.417*** (128.576) |
| Hour9:HolidayNo Holiday | 0.502** (0.179) | 222.622 (128.575) |
| Hour10:HolidayNo Holiday | 0.168 (0.179) | 42.169 (128.574) |
| Hour11:HolidayNo Holiday | 0.086 (0.179) | 19.430 (128.580) |
| Hour12:HolidayNo Holiday | 0.002 (0.179) | −12.562 (128.579) |
| Hour13:HolidayNo Holiday | −0.078 (0.179) | −66.050 (128.591) |
| Hour14:HolidayNo Holiday | −0.111 (0.179) | −75.660 (128.599) |
| Hour15:HolidayNo Holiday | −0.065 (0.179) | −26.319 (128.596) |
| Hour16:HolidayNo Holiday | −0.003 (0.179) | 25.825 (128.591) |
| Hour17:HolidayNo Holiday | 0.157 (0.179) | 201.774 (128.600) |
| Hour18:HolidayNo Holiday | 0.482** (0.179) | 619.452*** (128.593) |
| Hour19:HolidayNo Holiday | 0.380* (0.179) | 390.693** (128.589) |
| Hour20:HolidayNo Holiday | 0.298 (0.179) | 300.217* (128.585) |
| Hour21:HolidayNo Holiday | 0.299 (0.179) | 308.619* (128.583) |
| Hour22:HolidayNo Holiday | 0.319 (0.179) | 286.728* (128.585) |
| Hour23:HolidayNo Holiday | 0.267 (0.179) | 182.962 (128.592) |
| AIC | 116649.52 | 124000.34 |
| Corrected AIC | 116650.42 | 124001.24 |
| BIC | 117079.18 | 124430.01 |
| Observations | 8,465 | 8,465 |
| Adjusted $R^2$ | | 0.676 |
| $\theta$ | 3.908*** (0.059) | |
| *Note:* | | *p<0.05; **p<0.01; ***p<0.001 |




The comparison plot demonstrates that the Negative Binomial model provides higher and more accurate fitted values compared to the OLS model, particularly for higher counts. The OLS model tends to underpredict for larger values, as seen in the divergence from the red reference line at higher fitted values relative to Negative Binomial. In contrast, the OLS model produces some unrealistic negative fitted values and struggles to accommodate the count nature of the data. This leads to even greater deviations from the red reference line.

The Negative Binomial model is better suited for predicting bike rental counts as it handles overdispersion in the data, which the previous Poisson model fails to capture effectively. This makes the Negative Binomial model a more appropriate choice when the variance is much greater than the mean, which is the case for bike rental counts. Thus, the Negative Binomial regression provides more reliable predictions, particularly for higher bike rental counts, as it better aligns with the distribution of the data.

# 4 Conclusion

In this analysis, we aimed to investigate the factors influencing bike rental demand using the SeoulBikeData dataset. Our primary hypothesis was that various weather conditions—such as temperature, humidity, and rainfall—along with time-related factors like the hour, seasons, and holidays, significantly impact bike rental counts. In the process we have modeled both OLS and Poisson regressions. The Poisson model revealed significant overdispersion, which led us to adopt the Negative Binomial regression model.

Our key findings show that the Negative Binomial model is more appropriate for this dataset as it handles non-negative count data and accounts for the overdispersion observed in the Poisson model. The OLS model produces some negative fitted values and overpredicts for larger values, making it unreliable for this type of data. The Negative Binomial regression, particularly derived from our Model 3 Poisson, indicates that temperature, humidity, rainfall, and wind speed, along with interactions between time and holiday variables, are important factors in predicting bike rentals.

The implications of these findings suggest that bike-sharing systems can optimize bike availability by considering these factors in real-time demand forecasts. While the Negative Binomial model improves upon the Poisson model, the residual plot still shows some large residuals and outliers, suggesting that further refinement may be necessary. However, the flexibility of the Negative Binomial model in addressing overdispersion makes it a better fit for this dataset compared to the OLS and Poisson models. Lastly, limited documentation on the sampling process may introduce biases or affect the representativeness of our findings in a way that we are not aware of.