Regression Analysis of Bike Sharing Demand

Chance Robinson, Jayson Barker and Neha Dixit

Master of Science in Data Science, Southern Methodist University, USA

1 Introduction

[Intro]

2 Data Description

The dataset we chose for this project was a publicly shared, hourly bike sharing dataset made available through Kaggle in csv format.

This data set is divided into two distinct sets – a train and test set. The train set consists of 10,886 rows (titled "train.csv") and the test set consists of 6,493 rows (titled "test.csv"). Within the training set, the first 19 days of each month are captured whereas in the test data set, the 20th day to the end of each month is present. The entirity of the data spans from 1/1/2011 through 12/31/2012 - encompassing two full years of bike sharing data. Interstingly, the time component of this analysis is captured in hours of each day meaning we have a calendar date represented 24 times (for each hour of that day) in the data, along with it's associated attribute values.

In train data set, there are a total of 12 attributes which capture multiple variables related to bike rentals. Some of these attributes are categorical, and others are continuous. All attributes are summarized in the table below:

Column Name	Type	Description
1. datetime	Date	YYYY-MM-DD HH24 (example: 2011-01-01 04:00:00)
2. season	Integer	(1-4)
3. holiday	Integer	(0 or 1)
4. workingday	Integer	(0 or 1)
5. weather	Integer	(1-4)
6. temp	Float	temparture in Celcius
7. atemp	Float	"feels like" temperature in Celsius
8. humidity	Integer	relative humidity
9. windspeed	Float	wind speed
10. casual	Integer	count of casual users
11. registered	Integer	count of registered users
12. count	Integer	count of total users `response variable`

** The test data set lacks the casual, registered and count variables.

3 Exploratory Data Analysis

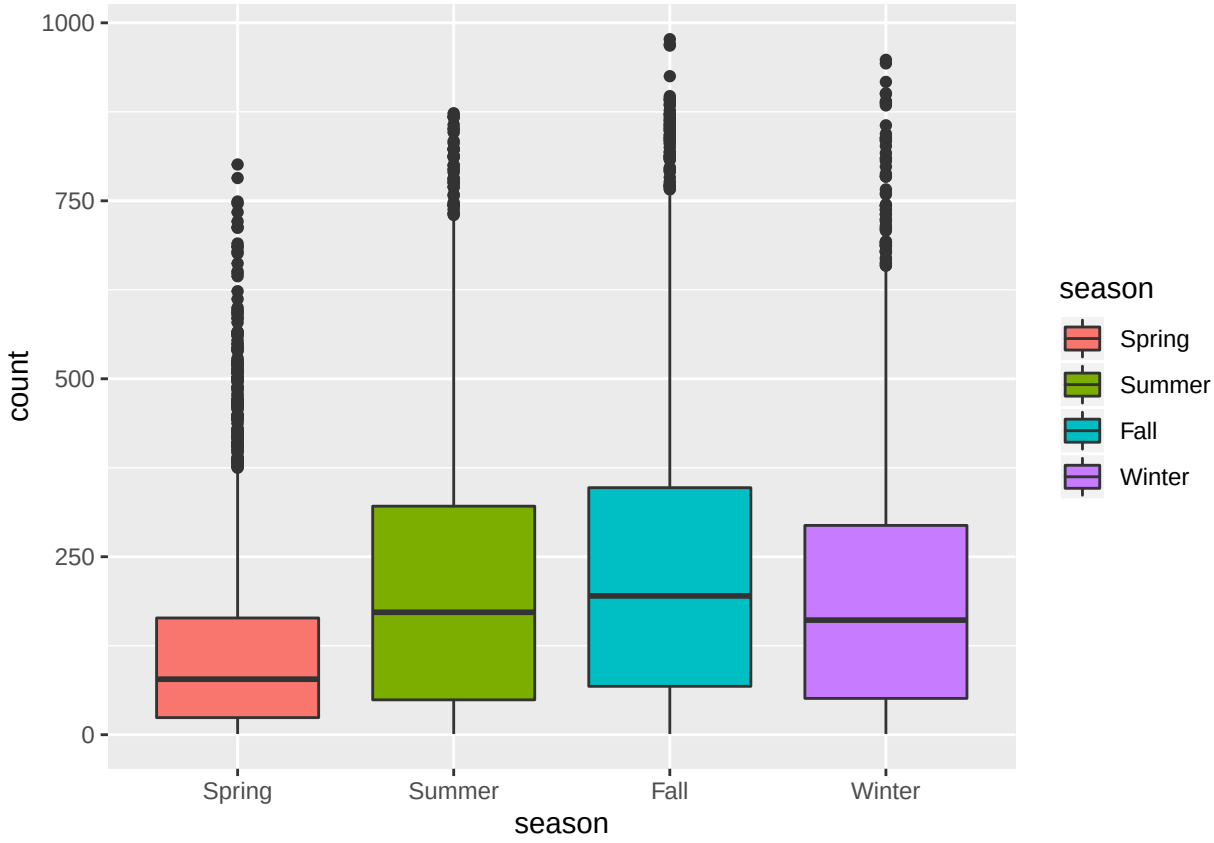
- Plotting Explanatory vs. Response Variables

3.1 Categorical Variables

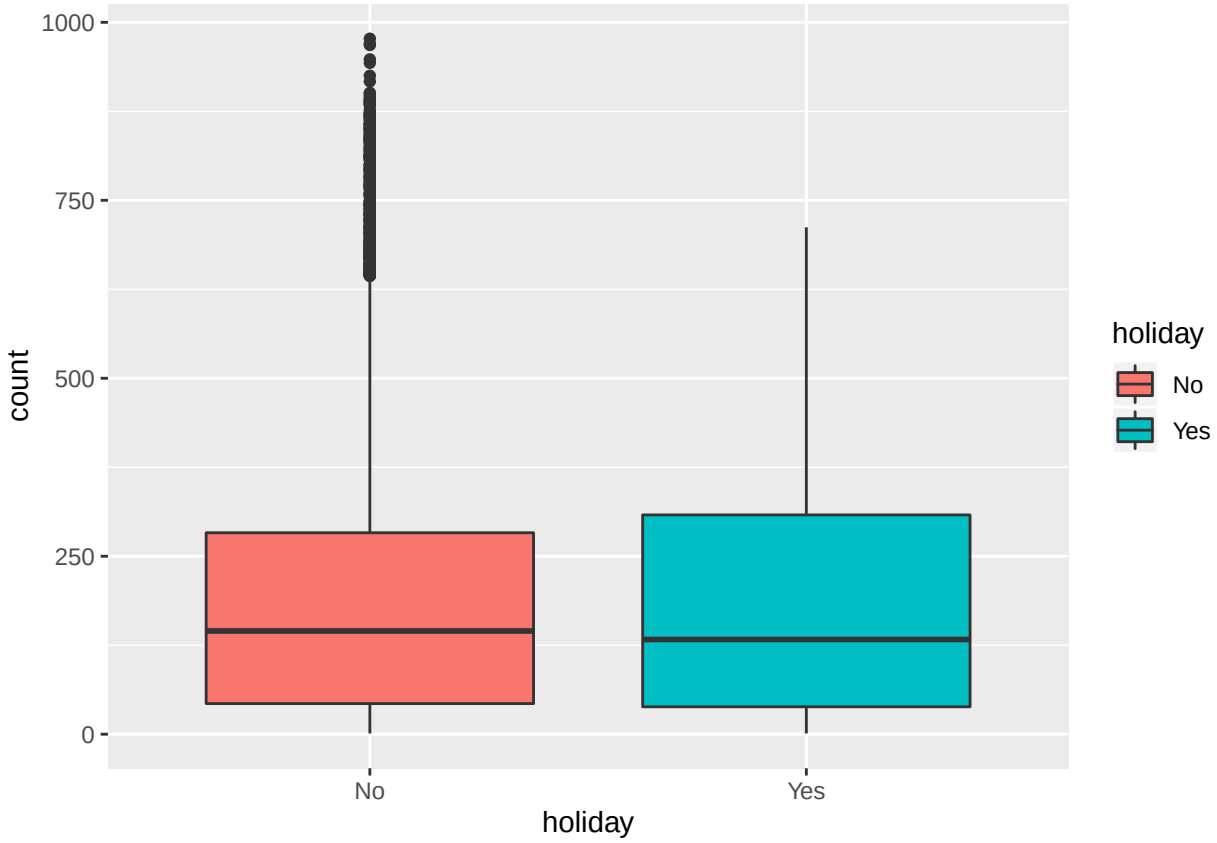
Several numeric variables were found to be better suited to categorization and were converted to factors.

3.1.1 Season

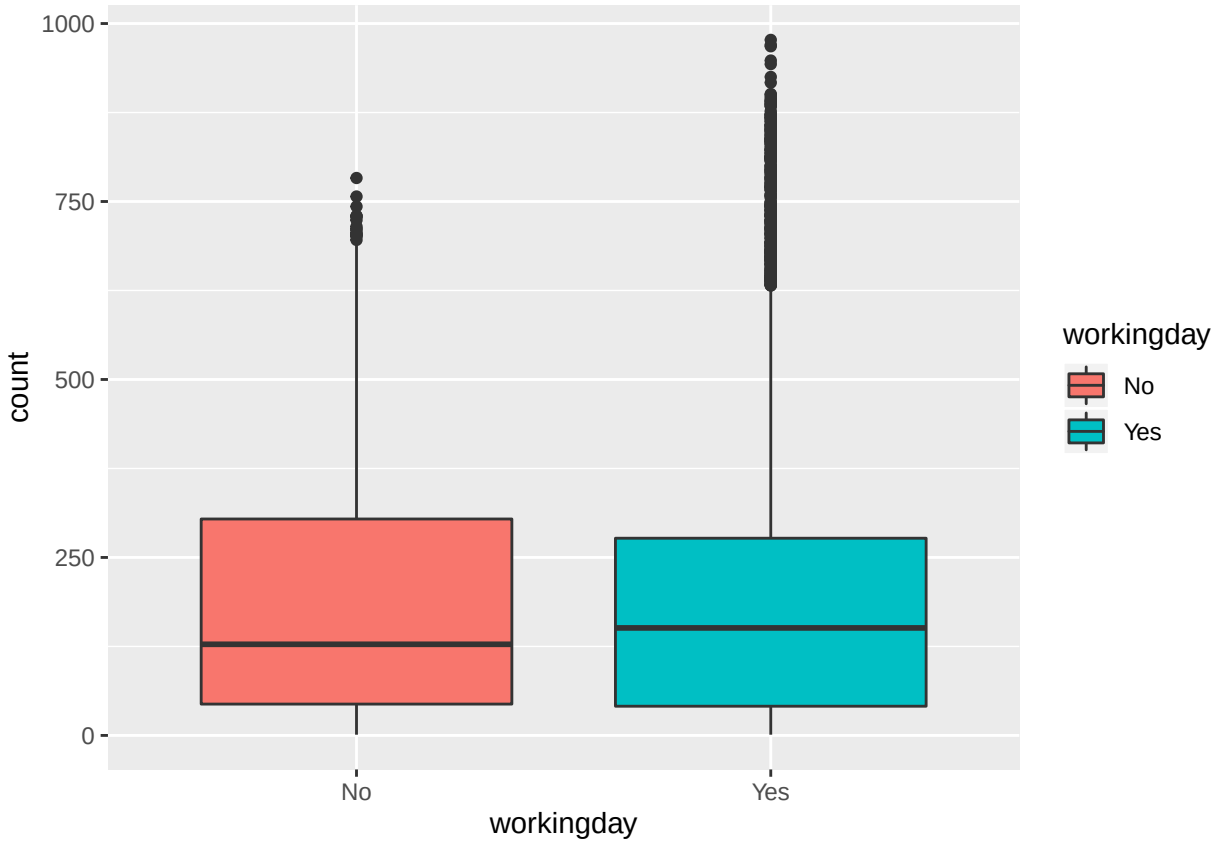
Season	Label	Description
1	Spring	Dec 21 ~ Mar 20
2	Summer	Mar 21 ~ Jun 20
3	Fall	Jun 21 ~ Sep 20
4	Winter	Sep 21 ~ Dec 20



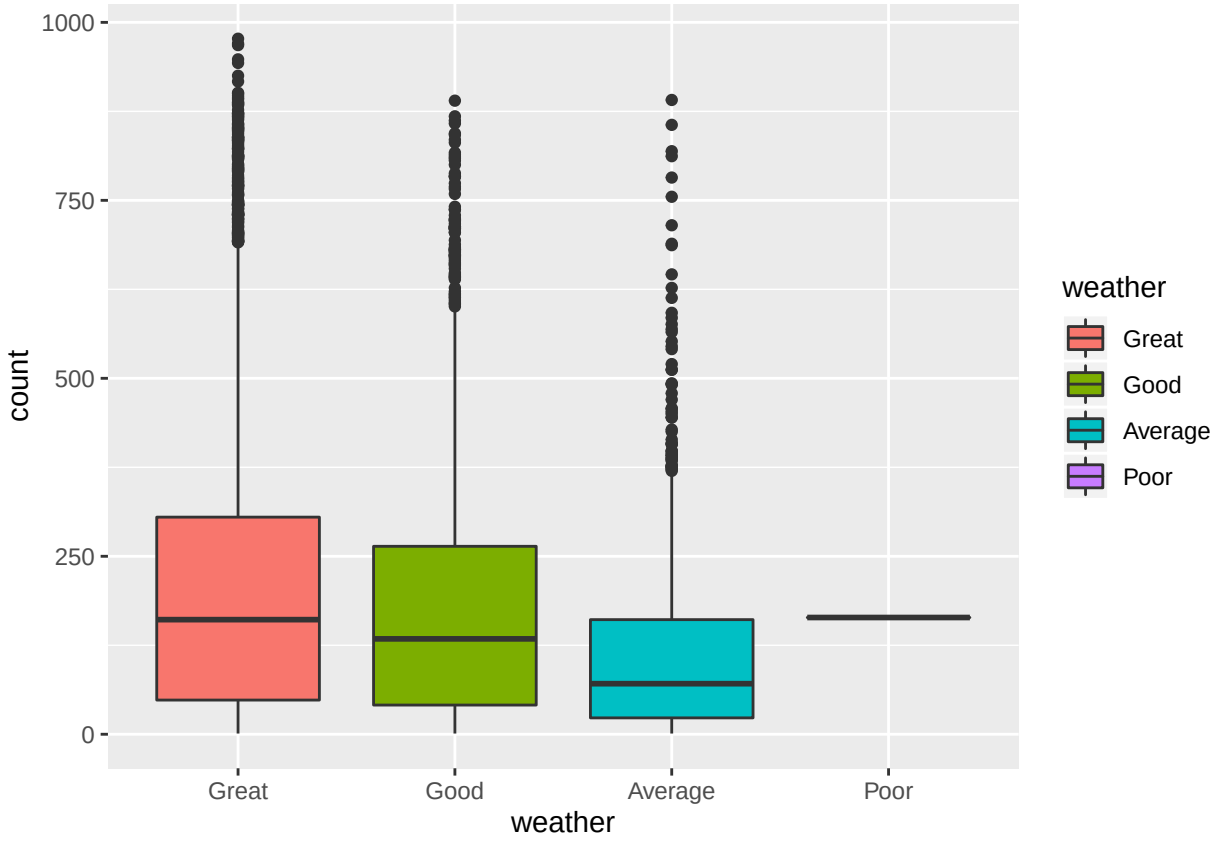
3.1.1.1 Holiday



3.1.2 Working Day



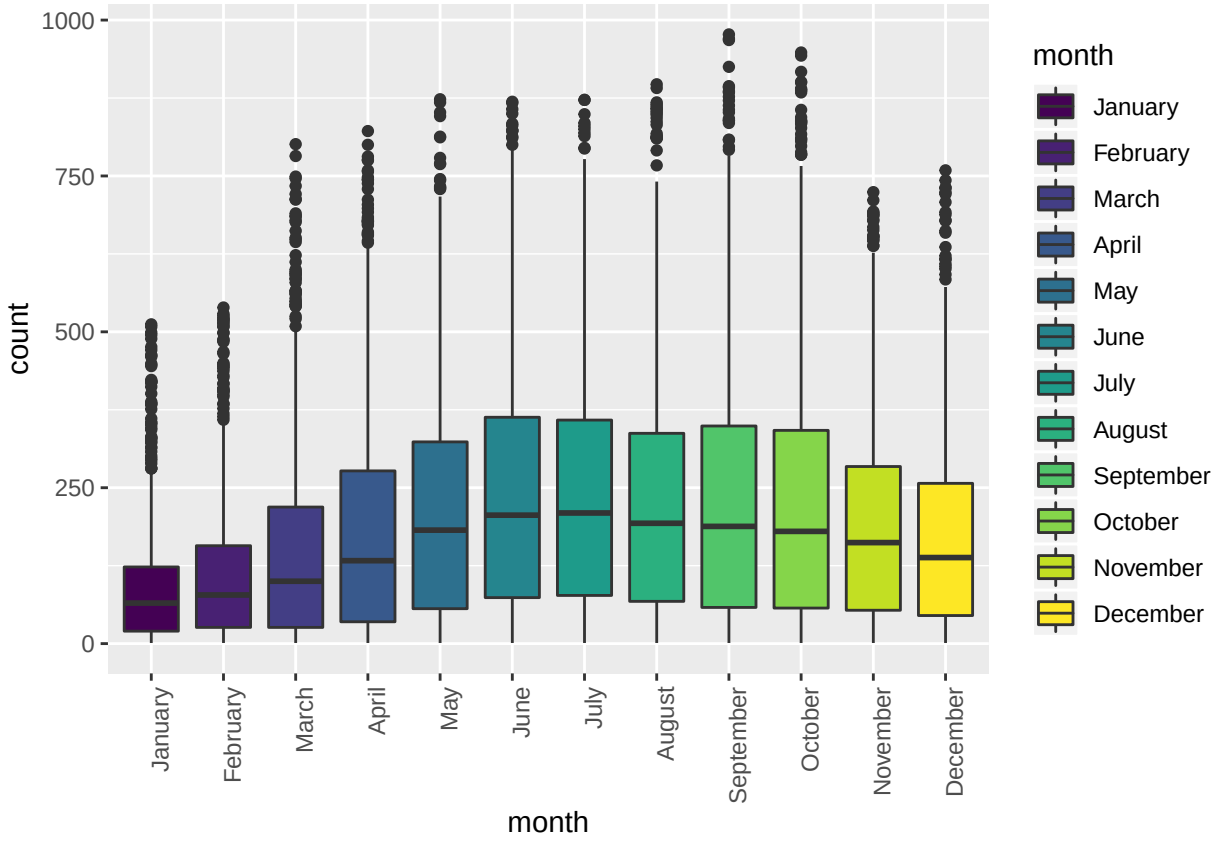
3.1.3 Weather



The datetime column was broken out into multiple factors as well so that we could visualize the components of each date and aggregate by different dimensions of the timestamp. We felt this was also necessary due to the nature of how the train/ test data sets had been pre-split. (i.e. . . with the first 19 days of the month holding the only true counts to validate our models against.)

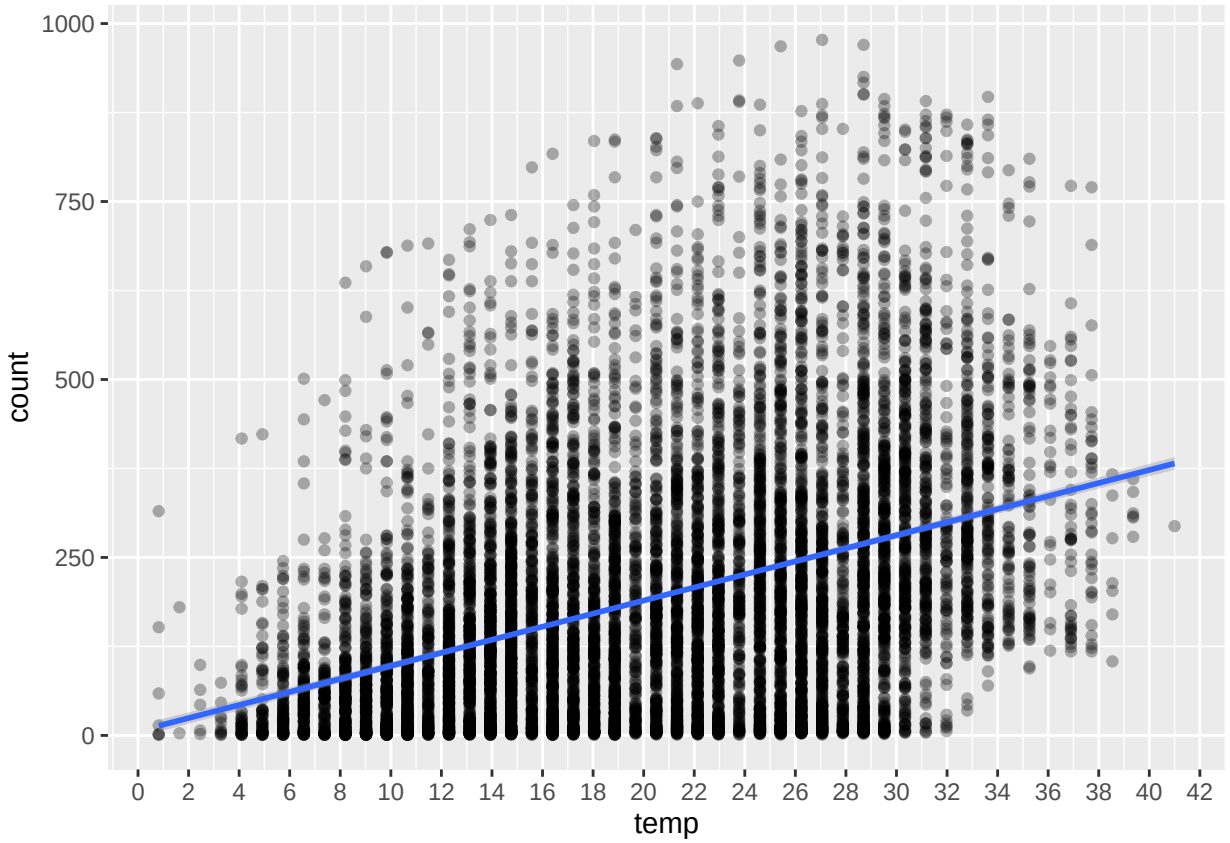
- Year
- Month
- Day
- Hour

3.1.4 Month

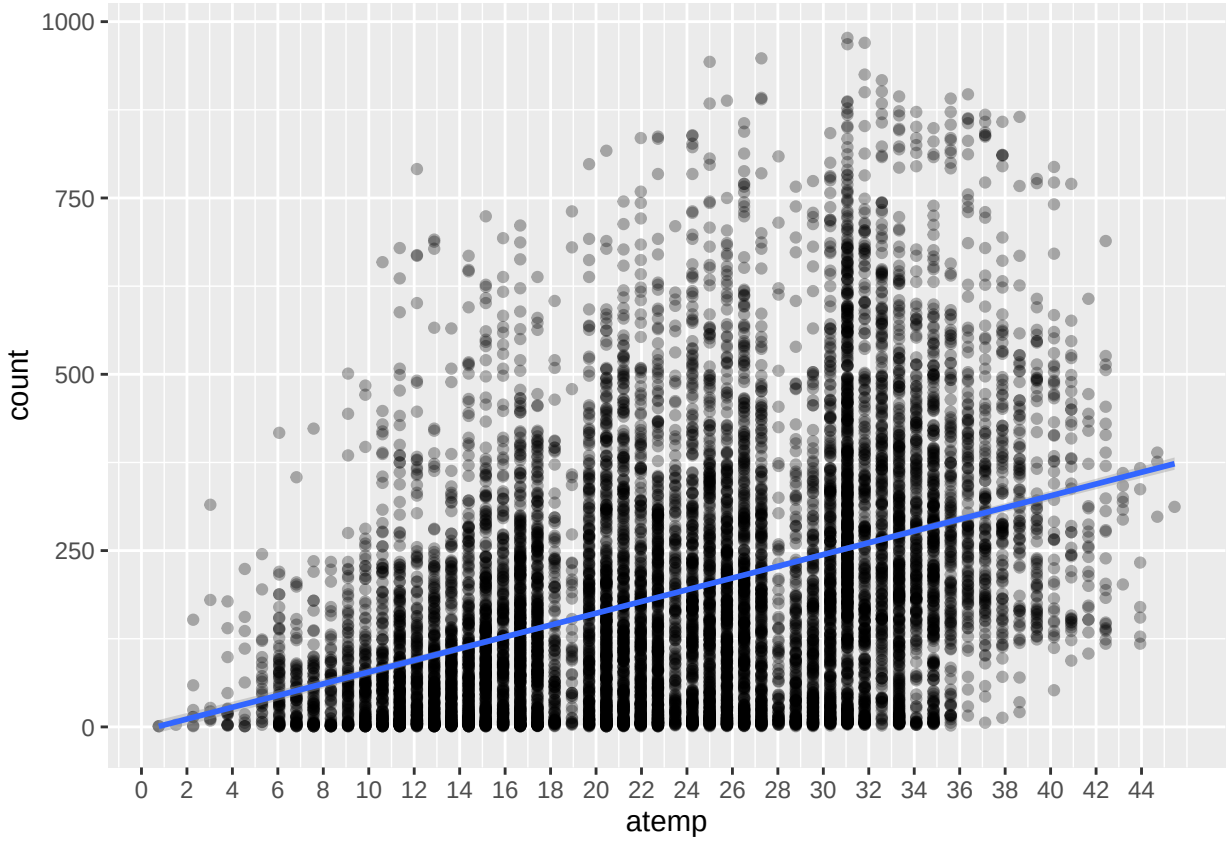

1000
750
500
250
0
count
January
February
March
April
May
June
July
August
September
October
November
December
month
month
January
February
March
April
May
June
July
August
September
October
November
December

3.2 Continuous Variables

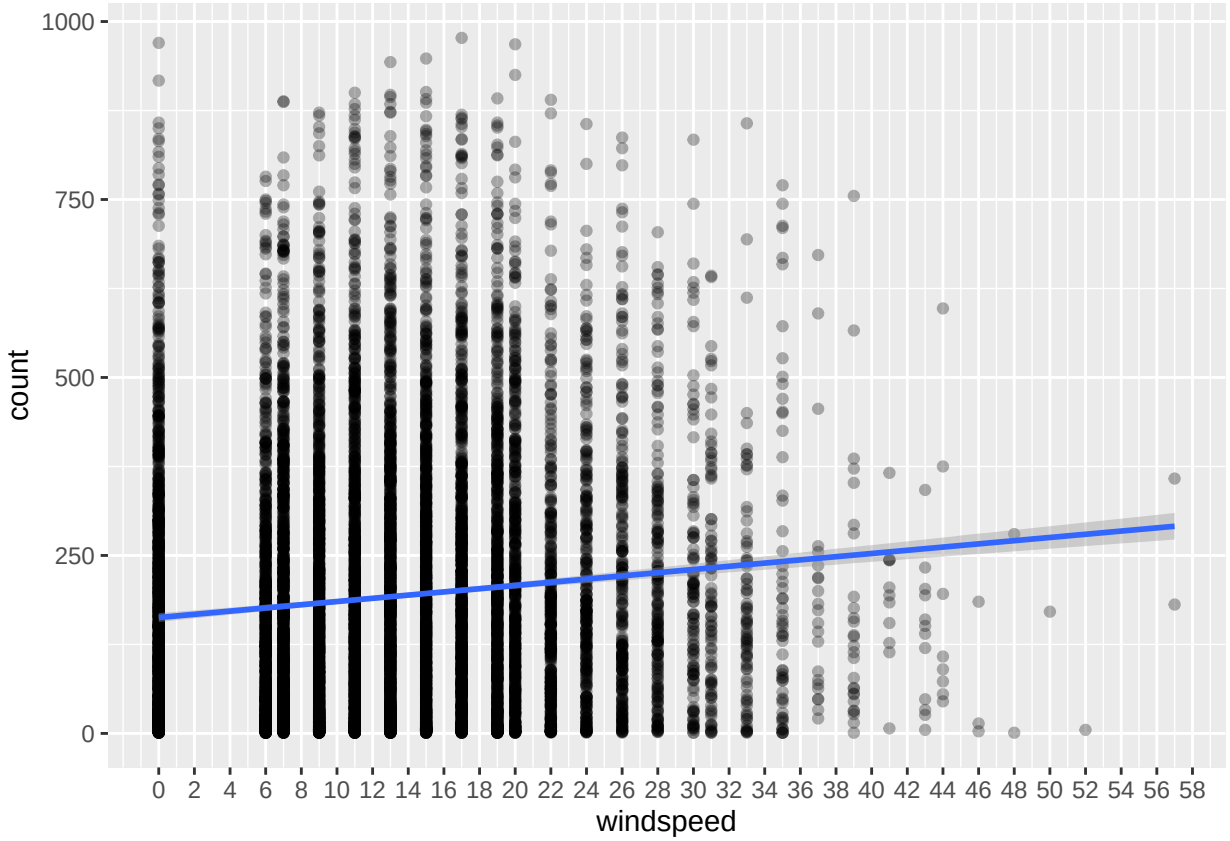
3.2.1 Count by Temperature



3.2.2 Count by "Feels like" Temperature



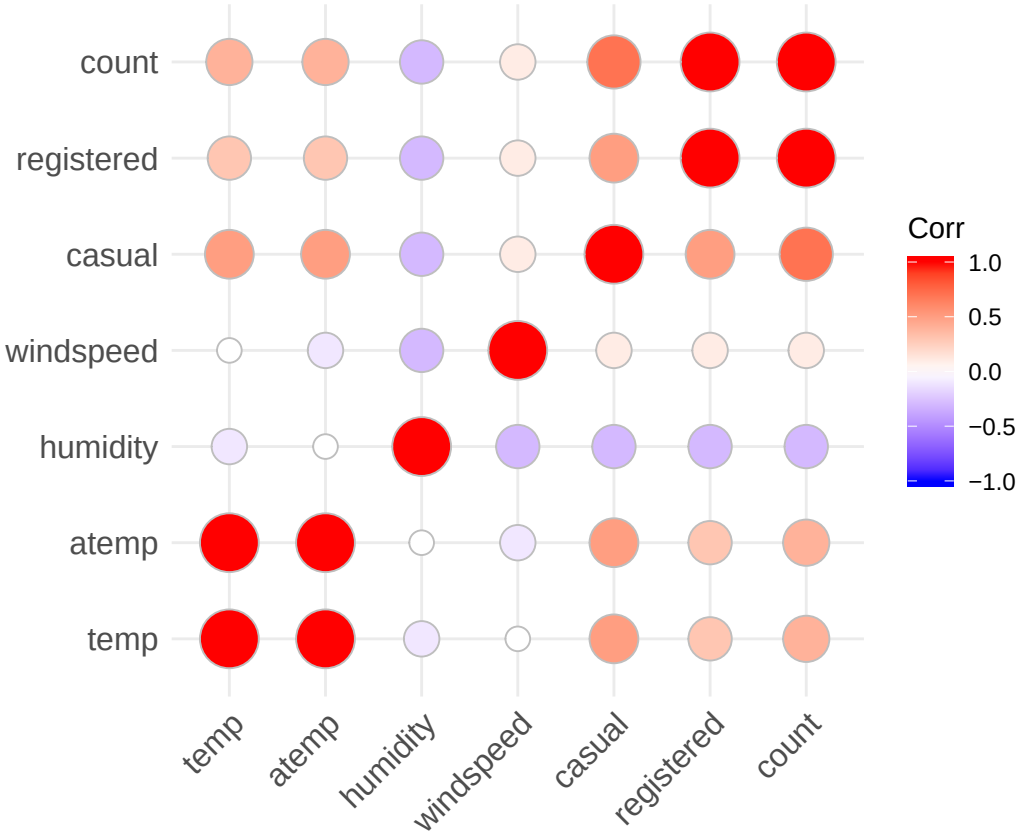
3.2.3 Count by Wind Speed



3.3 Correlation Matrix

There are several variables with a relatively high level of covariance from the training set. The following columns should therefore be removed so as not to be picked up by any automated modelling techniques.

- atemp
- causual
- registered



4 Objective I Analysis

4.1 Question of Interest

The team be utilizing the multiple linear regression techniques we've learned up to this point in the program to predict the bike rental deman on a given date and time. The model will be evaluated on the Root Mean Squared Logarithmic Error, or RMSLE. As this data represents hourly data collected, there is an obvious time component associated with this particular competition. We wanted to gauge how effective mutiple linear regression would be when the assumption of indepdence is clearly violated.

```
## # A tibble: 12 x 5
##    month      mean    sd median observations
##    <ord>     <dbl> <dbl>  <dbl>        <int>
##  1 January    3.78  1.47   4.17          884
##  2 February   3.98  1.51   4.36          901
##  3 March      4.19  1.61   4.61          901
##  4 April      4.46  1.53   4.89          909
##  5 May        4.76  1.43   5.20          912
##  6 June       4.89  1.38   5.33          912
##  7 July       4.90  1.35   5.34          912
##  8 August     4.86  1.37   5.26          912
##  9 September  4.81  1.42   5.24          909
## 10 October    4.79  1.42   5.19          911
## 11 November   4.65  1.40   5.09          911
## 12 December   4.52  1.43   4.93          912
```

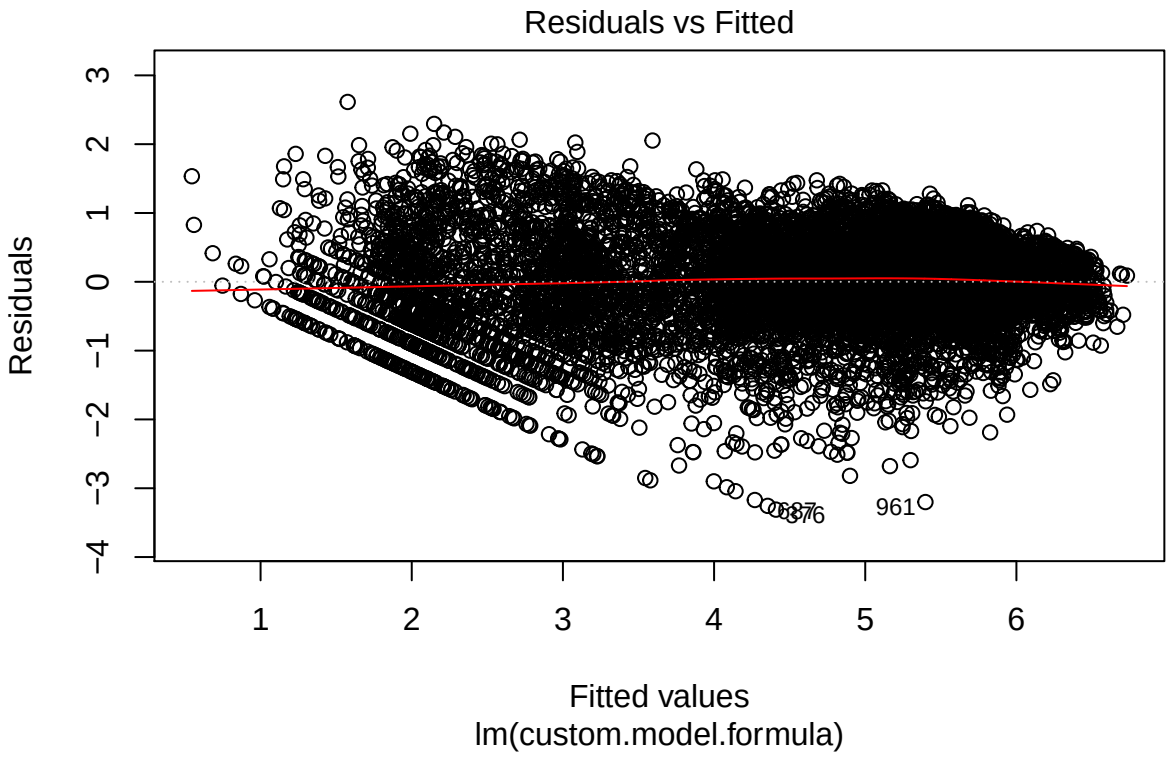
```
## # A tibble: 1 x 12
##   datetime            season holiday workingday weather  temp atemp
##   <dttm>               <dbl>   <dbl>      <dbl>   <dbl> <dbl> <dbl>
## 1 2012-01-06 13:00:00      1       0          1       1  16.4  20.5
## # ... with 5 more variables: humidity <dbl>, windspeed <dbl>,
## #   casual <dbl>, registered <dbl>, count <dbl>
```

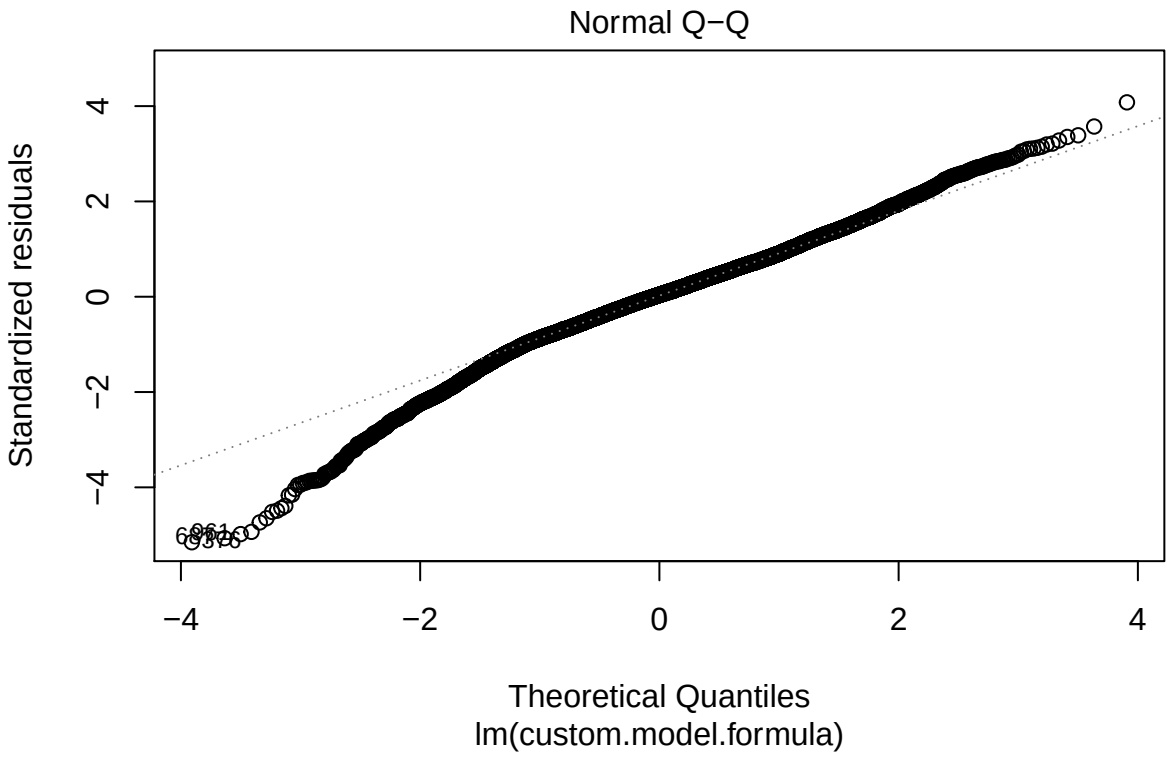
4.2 Model Selection

4.2.1 Custom Variable Selection

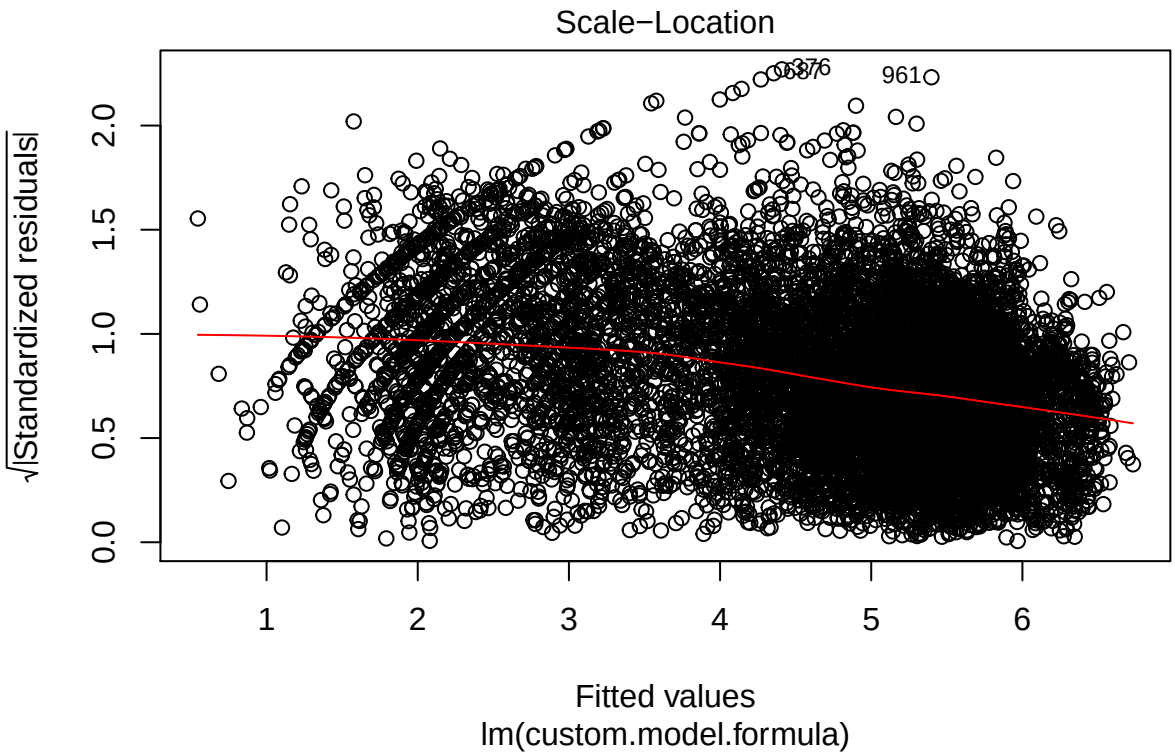
We developed the best fitting model with a custom selection of variables after having accounted for those with high correlation and adding interactive terms such as month/ hour based on the seasonal nature that the box plots exhibited.

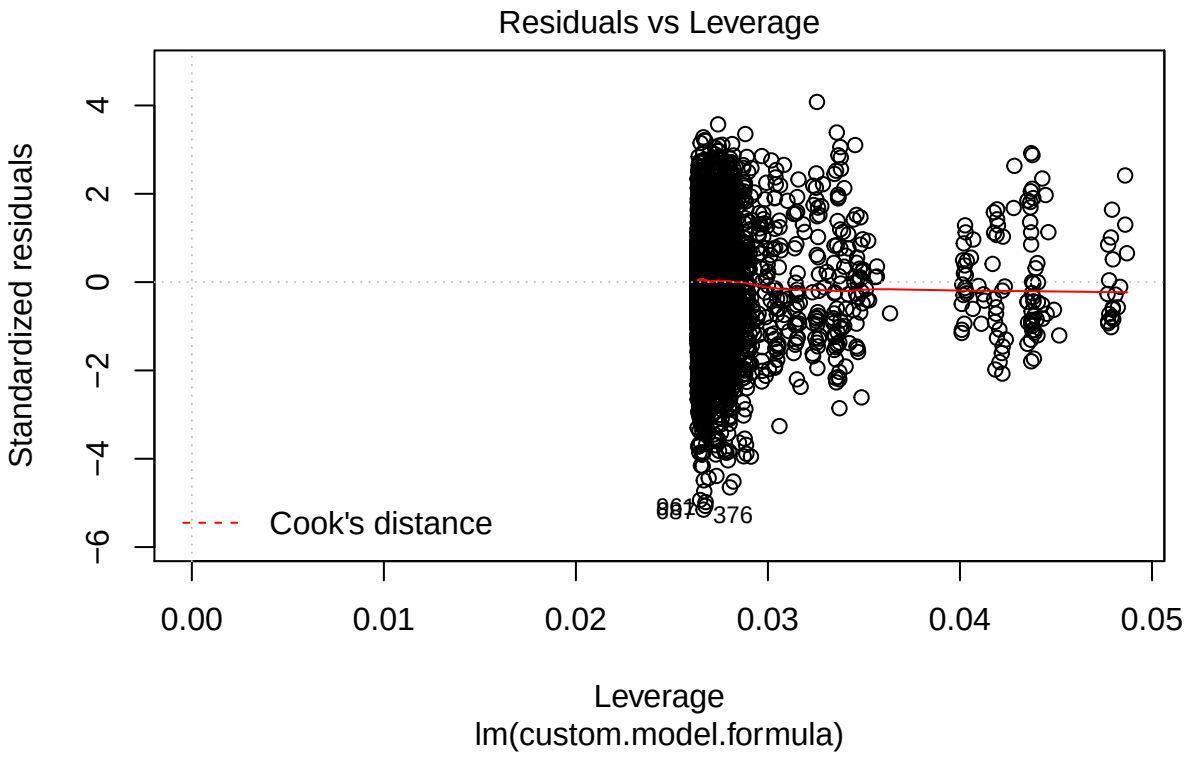
```
## Warning: not plotting observations with leverage one:
##   5555
```

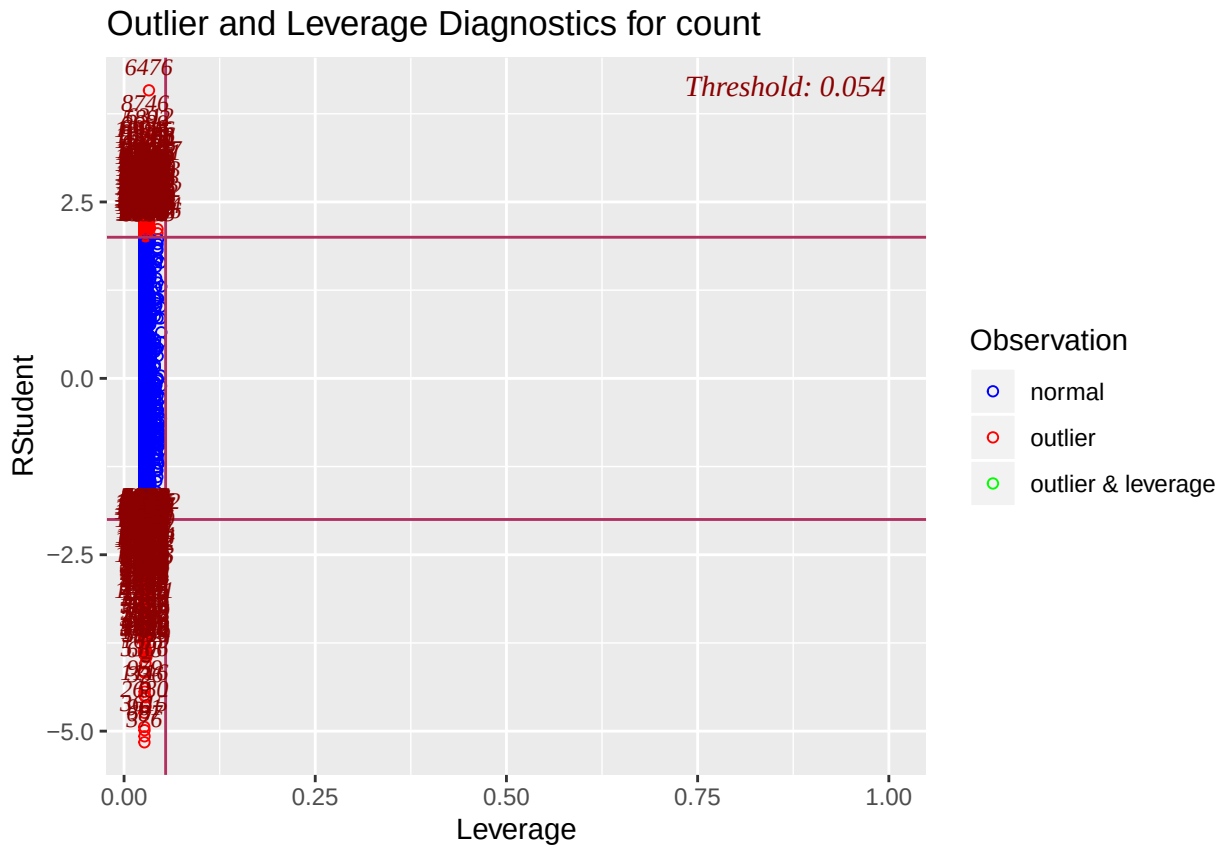

Residuals vs Fitted
Residuals
Fitted values
lm(custom.model.formula)
961




```
## Warning: not plotting observations with leverage one:
##   5555
```


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1
2
3
4
5
6
Fitted values
lm(custom.model.formula)
961


Residuals vs Leverage
Standardized residuals
4
2
0
−2
−4
−6
0.00
0.01
0.02
0.03
0.04
0.05
Cook's distance
376
Leverage
lm(custom.model.formula)



4.2.2 RMSLE: Root Mean Squared Logarithmic Error Loss

```
## [1] 0.1923972
```

4.2.3 Lasso

4.2.4 Stepwise

4.3 Model Assumptions Assessment

- The response variable is linear
- The data is normally distributed
- Independence

4.4 Comparing Competing Models

4.5 Parameters

Parameters

4.6 Model Interpretation

Model Interpretation

4.7 Conclusion

Conclusion

5 Objective II Analysis

5.1 Question of Interest

[Discuss time series method]

Time-Series Analysis

[Details goe here]

5.2 Model Assumption Assessment

Model Assumption Assessment

5.3 Comparing Competing Models

5.4 Conclusion

Conclusion

6 Appendix

6.1 Code