FINAL PROJECT

Jiayu Liu, Lan Sun, Azra Pita

Table of Contents

1 Executive Summary

In this study, we observed the bike-sharing rental system by Capital bike sharing company in order for the company to arrange its bike mechanics employees better at different time. Our main objective is to predict the people's renting behavior based on time. Some ideas to enhance the analysis will be listed. The study rests on the fact that the rental process is highly correlated to the environmental, seasonal settings and time. For instance, weather conditions, precipitation, the day of the week, season, the hour of the day, etc. can affect the rental behaviors. After conducting the multiple regression analysis, we listed out our prediction for bikes demand by different time intervals and found that people rent bikes for different usages on weekdays and weekends.

2 Introduction

2.1 Background

Bike sharing systems are the new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Since 2010 the bike-sharing rental process has become popular in the major metropolitan U.S. cities, we think it would be useful to do some regression analysis on the data to get some insight on this industry.

The desired goal of the analysis is to predict people's renting behavior based on certain time intervals. The core dataset contains the hourly and daily count of bike rental between 2011 and 2012 with corresponding weather and seasonal information collected in Washington DC. Data is collected from Capital bike-share system.

2.2 Purpose of the Report

Our main purpose is to predict people's renting behavior by an hour of a day and day of a week and thus enhance the understanding of the bike rental demand.

The customer rental behavior will provide a framework for the analysis with a focus on the spring season of 2011. The particular research questions that will be investigated are listed below:

1. How do we choose the predictors in the original model?
2. Does the linear relationship between each predictor and response is appropriate/Do we need to include polynomial terms?
3. Why there is a gap in the Residual against Hours plot? How to improve this plot? Test whether the interaction term is needed (ANOVA)?
4. Can we reduce our predictor to a subset?
5. Does there any multicollinearity still exist after subsetting?
6. Does our final model satisfy the 4 assumptions related to the linear model assumption?
7. How to make assumptions based on our final model?

3 Research Methodology

3.1 Explanation of the Variables

To conduct our analysis, we used data of each bike rental, collected from the bike sharing automated system for the spring season in 2011 that summed to 2207 observations. The dependent variable for this data set was the count of total rental bikes that includes both casual and registered users. Since our dependent variable is highly correlated to the time and the environmental settings. So we decided to choose hour, weathersit, atemperature, temperature, humidity, windspeed, holiday, month, weekday and workingday as our predictors. The detailed explanations of the variables are listed below:

Cnt: Count of total rental bikes including both casual and registered

Hour: Dataset is based on recording data hourly (0 to 23)

Weathersit: 1: Clear, Few clouds, Partly cloudy, Partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain

Atemperature: Normalized feeling temperature in Celsius.

Temperature: Normalized temperature in Celsius. The values are divided to 41 (max)

Humidity: Normalized humidity.

Windspeed: Normalized wind speed.

Holiday: weather day is a holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)

Month: March, April, May, and June

Weekday: day of the week

Workingday: if day is neither weekend nor holiday is 1, otherwise is 0.

3.2 Variable Selection

In order to make our model more useful for prediction, we used several methods to add or drop the predictors. First, we draw the partial regression plot to see whether polynomial terms are needed. Then, we noticed from the Residual against Time plot that there is gap in the residual. So we added the interaction term in the model and did ANOVA test to verify that the interaction term is needed. Finally, since there are some redundant variables in our model, we use stepwise to derive a subset of the model. (The detailed procedure will be discussed in the Result Part)

3.3 Transformation

After fitting the model, we want to see if the four assumptions related to the linear model are satisfied or not. Because our goal is to make predictions, so the linear relationship is important. Since from the Diagnostic plots, we can see that the assumptions are not satisfied, so we decide to do the Box-Cox transformation. (The detailed procedure will be discussed in the Result Part)

3.4 Challenges

1. The challenge was to determine the best criteria to choose our variables. Due to the size of our predictor variables to consider regression outcomes best subset analysis was not possible. Therefore, to determine the best variables in our models we used stepwise regression.
2. There is a gap in the Residuals against Hour plot and we have solved it by adding the interaction term which has explained in the Variable Selection part.
3. After we create the original model and did the summary to see the t-test and p-value, we found that the line related with Workingday predictors gives us NA values. This may be due to the Workingday and Weekday variables have exact collinearity relationship. However, we didn't drop this predictor at first, because it can be dropped in the stepwise procedure automatically.

4. We found that the Normal Q-Q plot is not close to the line at tails which indicates that normal distribution assumption for the residuals isn't satisfied. The reason may be that the values for the response fall into two different scales. During the daytime, the scale is large, during the nighttime, the scale is small. Thus, we created a new variable called Day to reflect these two scales. We set the Day variable into two values. When Hour is between 7 and 21, we set the value of Day variable equals to 1, otherwise, equals to 0. Using the new model to derive the Normal Q-Q plot, we found the plot improves a lot. However, we didn't include the Day variable into our model due to the third challenge (exact collinearity between Day and Hour variables and gives NA value in summary). We can only choose one variable in two of them. Since our goal is to predict the response by hour, so we still use Hour variable in our model.

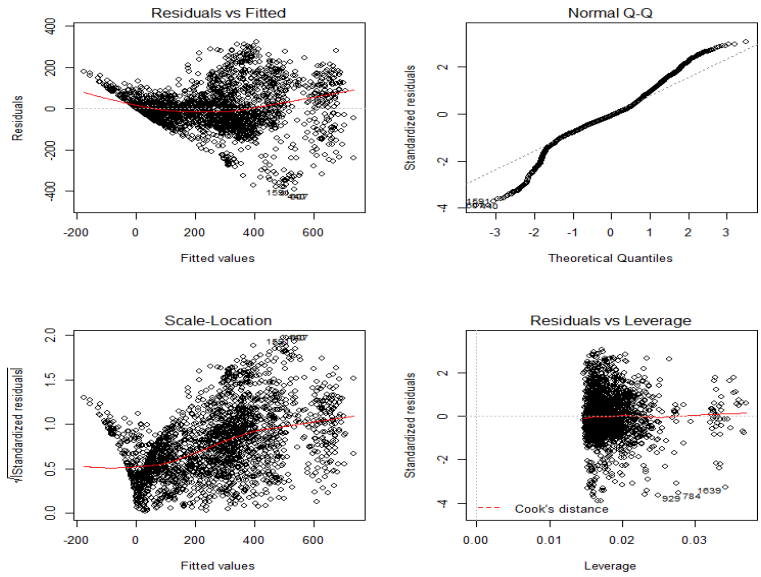
4 Results

4.1 Variable Selection

Our original model without any transformation is:

```
lm(cnt~temp+atemp+hum+windspeed+factor(weathersit))+factor(mnth)+factor(hr)+factor(holiday)+
factor(workingday)+factor(weekday)
```

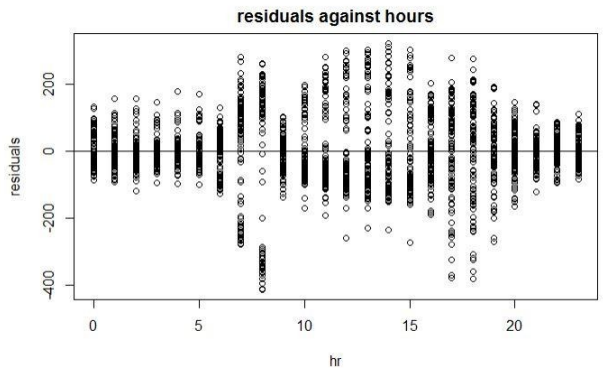
Due to the curve in Residuals vs Fitted plot (see Plot 1), we built the partial regression plot for e(Y|Xothers) verses e(Xi|Xothers) to see whether we have to include polynomial terms. The result turns out that there does not appear to be a strong nonlinear pattern in any of these plots, indicating that the assumed linear relationship between each predictor and the response is appropriate.



Plot 1: Diagnostic Plots for the original model

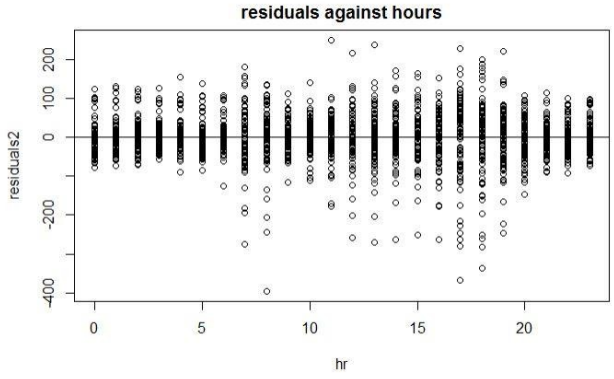
Then we drew the Residuals against Time plot. Because of the gap at around 7am and 8am, which is the time when people go to work, we guess that the variable Hour has different influence on the response at different level of variable weekday. Thus, we tested by adding the added Interaction term of hour and weekday. Based on the results in Plot 2 & 3, we could conclude that our model drastically changed the interpretation of residuals in the model.

```{r}
residuals<-m1$residuals
plot(hr,residuals,main="residuals against hours")
abline(h=0)
```



Plot 2: Residuals against hours plot (before)

```{r}
residuals2<-m2$residuals
plot(hr,residuals2,main="residuals against hours")
abline(h=0)
```



Plot 3: Residuals against hours plot (after)

By using the ANOVA function, we compare the models with and without interaction terms. The p-value is very small, so we reject the Null Hypothesis and conclude that it is reasonable to build the model with the interaction term. (see Table 1)

```
```{r}
anova(m2,m1)
```
```

```
Analysis of Variance Table

Model 1: cnt ~ temp + atemp + hum + windspeed + factor(mnth) + factor(hr) +
    factor(holiday) + factor(weekday) + factor(weathersit) +
    factor(workingday) + factor(hr) * factor(weekday)
Model 2: cnt ~ temp + atemp + hum + windspeed + factor(mnth) + factor(hr) +
    factor(holiday) + factor(weekday) + factor(weathersit) +
    factor(workingday)
  Res.Df      RSS   Df Sum of Sq     F    Pr(>F)
1   2028  7374844
2   2166 24613064 -138 -17238220 34.35 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 1: ANOVA analysis of interaction term

We then summarized our model with the interaction term. The p-value of some variables are larger than 0.05, which indicates that these variables either have no significance influence to the response variable or there is some multicollinearity between the variables. Thus, we decided to reduce the variables in the model. (see Appendix, Table 2)

Due to the size of our predictor variables to consider regression outcomes, best subset analysis was not possible. Therefore, to determine the most important predictors in our models we used stepwise regression. The procedure displays that hour, weathersit, atemp, humidity, weekday, windspeed, holiday, month and hour:weekday variables should be included in our model. (see Table 3)

```
Step:  AIC=18255.72
cnt ~ factor(hr) + factor(weathersit) + atemp + hum + factor(weekday) +
    windspeed + factor(holiday) + factor(mnth) + factor(hr):factor(weekday)

                             Df Sum of Sq      RSS   AIC
<none>                                     7377517 18256
- factor(mnth)                3     21349  7398866 18256
+ temp                        1      2673  7374844 18257
- factor(holiday)             1     15772  7393289 18258
- windspeed                   1     29180  7406697 18262
- hum                         1    341829  7719347 18354
- atemp                       1    638133  8015650 18437
- factor(weathersit)          2   1562491  8940009 18676
- factor(hr):factor(weekday) 138  17236073 24613590 20638
```

Table 3: Stepwise regression

The VIF test helps us to check that there is no severe multicollinearity exists in the stepwise model.

```
```{r}
library(car)
vif(stepwise1)
```
```

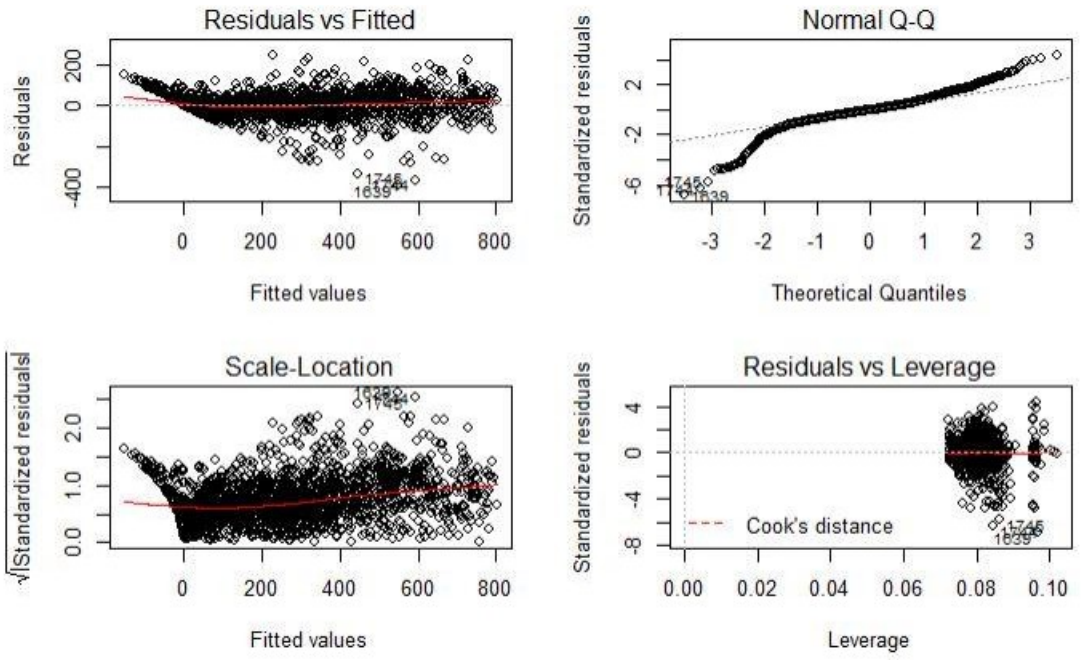
```
                                GVIF  Df GVIF^(1/(2*Df))
factor(hr)                4.038227e+19  23        2.668219
factor(weathersit)        1.935285e+00   2        1.179468
atemp                     2.862823e+00   1        1.691988
hum                       2.275517e+00   1        1.508482
factor(weekday)           1.955259e+08   6        4.908329
windspeed                 1.304398e+00   1        1.142103
factor(holiday)           1.331368e+00   1        1.153849
factor(mnth)              2.554951e+00   3        1.169222
factor(hr):factor(weekday) 1.565513e+27 138        1.254675
```

Table 4: Multicollinearity test

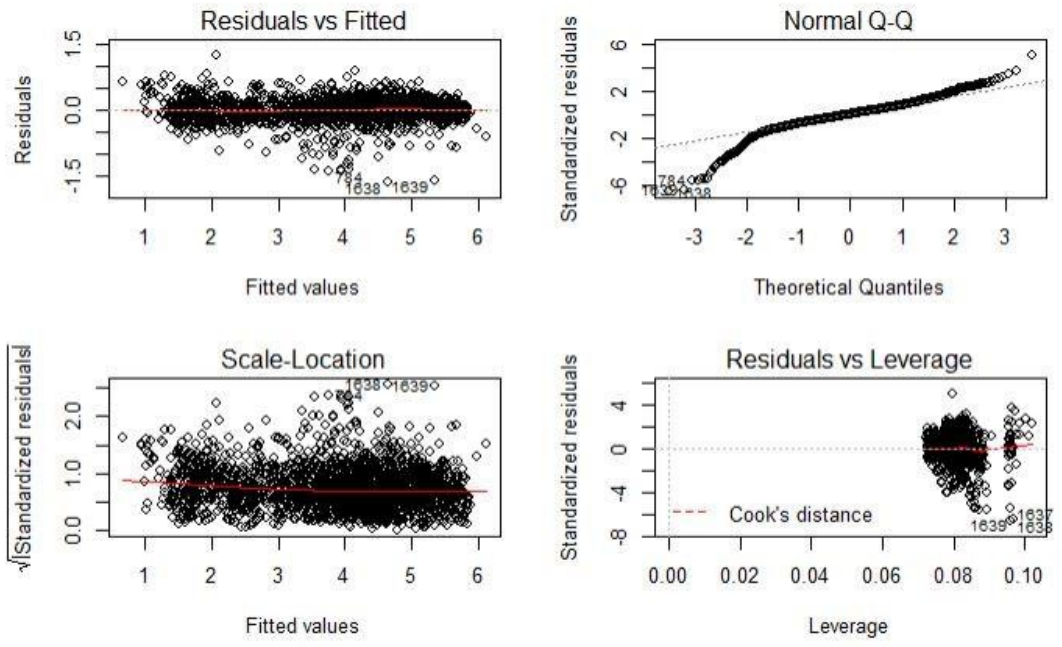
4.2 Transformation

After variable selection, we draw the four diagnostic plots in R and found that the assumptions related to the linear regression model are not satisfied. The Residuals vs Fitted plot showed a slightly curved pattern of residuals. The errors are not normally distributed because the dots shown

in the Normal Q-Q plot deviate slightly from a straight line towards the bottom and end of the regression line. Scale location plot shows an increasing pattern for the square root of the absolute standardized residuals. So we decide to use Box-Cox transformation to transform our model. The best transformation for our model is Box-Cox Transformation with the best lambda of 0.2626263. From the diagnostics plot after using Box-Cox transformation, we can see that the multiple regression model fitted the data better after adding the best lambda into our model. (see Plot 3 & Plot 4)



Plot 3: Diagnostic Plots before Box-Cox transformation

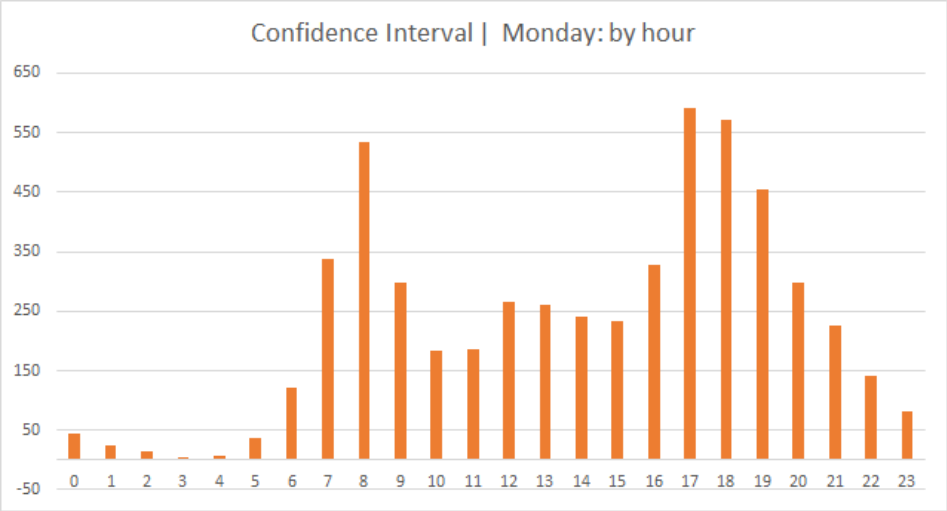


Plot 4: Diagnostic Plots after Box-Cox transformation

4.3 Predictions

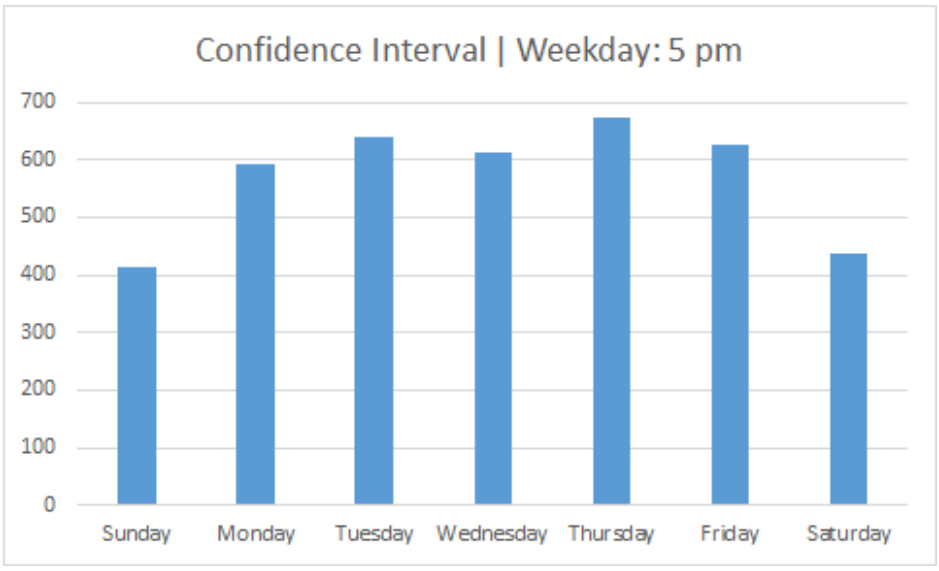
Now, we have a useful model to make some predictions for our response. We use confidence interval instead prediction interval because we want to see the average value of our response.

First, we set other variables at a constant level and only change the hour variable to get predictions by hour. The confidence interval of 95% showed that hour of the day has an impact on the customer rental behavior. Furthermore, we could conclude that during night bike rental is low and during the day bike rental significantly increases. The highest demand for bike rental is from 8:00 to 8:59 am and 17:00 to 18:59 pm. The analysis showed that bike rentals are primarily used as a transportation for work. (see Appendix, Table 5) (see Plot 5)



Plot 5: Predicted value by hour

We also performed a 95% confidence interval for each day at 5 pm and we noticed that the demand for bike rental is lower during the weekend. However, data shows that people are using bikes for other purposes besides work. (see Appendix, Table 6) (see Plot 6)



Plot 6: Predicted value by day

5 Summary and Recommendations

In conclusion, there is a significant relationship between several independent variables and the Bike-sharing rental behavior. The developed regression model can be used to predict hourly and daily demand in the spring season for the bike-sharing system. We found that at a specific hour of the day there is a high demand for the bike rental which can further be explained if we look at the relationship between casual and registered bike users. Our research offers the reference for the bike sharing company to arrange the number of staff for operation and maintenance at different hour and weekday. In order to improve our analysis, the next step would be to compare the effect of the rental behavior based on the environmental and all four seasonal settings.

Appendix

```r
summary(m2)
```

```
Call:
lm(formula = cnt ~ temp + atemp + hum + windspeed + factor(mnth) +
    factor(hr) + factor(holiday) + factor(weekday) + factor(weathersit) +
    factor(workingday) + factor(hr) * factor(weekday))

Residuals:
    Min      1Q  Median      3Q     Max
-395.22  -27.14   -0.83   26.68  249.79

Coefficients: (1 not defined because of singularities)
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  71.078     20.723   3.430 0.000616 ***
temp                         58.973     68.780   0.857 0.391317
atemp                       208.498     77.350   2.696 0.007086 **
hum                         -92.726      9.537  -9.722  < 2e-16 ***
windspeed                   -35.775     12.412  -2.882 0.003989 **
factor(mnth)4                -8.989      4.553  -1.975 0.048456 *
factor(mnth)5                -5.318      5.046  -1.054 0.292062
factor(mnth)6                -1.309      5.641  -0.232 0.816530
factor(hr)1                 -30.964     23.659  -1.309 0.190770
factor(hr)2                 -47.200     23.658  -1.995 0.046163 *
factor(hr)3                 -77.375     23.660  -3.270 0.001093 **
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.3 on 2028 degrees of freedom
Multiple R-squared:  0.9283,    Adjusted R-squared:  0.9221
F-statistic: 148.4 on 177 and 2028 DF,  p-value: < 2.2e-16
```

Table 2: Summary of the model with interaction term (Part)

	fit	lwr	upr
0	45.206	36.717	55.062
1	25.111	19.662	31.601
2	13.838	10.390	18.064
3	4.914	3.296	7.054
4	8.142	5.833	11.063
5	35.867	28.708	44.263
6	120.397	102.484	140.519
7	336.747	297.964	379.132
8	534.859	480.173	594.004
9	296.933	261.771	335.461
10	183.627	159.090	210.848
11	186.895	161.965	214.546
12	266.679	234.147	302.427
13	261.907	229.771	297.240
14	240.572	210.386	273.837
15	233.250	203.805	265.719
16	328.510	290.419	370.165
17	591.244	532.281	654.888
18	571.456	514.132	633.360
19	454.741	406.462	507.121
20	298.572	263.312	337.197
21	226.534	197.790	258.247
22	140.737	120.605	163.245
23	82.076	68.707	97.271

Table 5: Confidence Interval: Monday by hour

	fit	lwr	upr
Sunday	413.593	369.063	461.965
Monday	591.244	532.281	654.888
Tuesday	638.695	576.805	705.350
Wednesday	611.711	553.981	673.759
Thursday	673.082	608.818	742.216
Friday	626.343	565.050	692.404
Saturday	438.061	390.761	489.453

Table 6: Confidence Interval: Weekday 5pm