DS 803 Final Project: Bikeshare Rental Data

Lisa Olsson and Matthew Clarke

Capital Bikeshare system in Washington D.C., USA has shared data from their hourly bike rental usage of the years 2011 and 2012. This data has then been joined with the current weather data for each hour in the dataset to allow for exploring the relationship between weather and bike usage.

In this report we will investigate the expected value of the number of bikes used during rush hour in Washington D.C. How many bikes are needed to accommodate rush hour usage in Washington D.C and what is the relationship between the experienced temperature, windspeed and number of bikes used?

Our dataset was accesses from UCI Machine Learning Repository: Data Sets .

Data Variables: (Variables used are underlined)

- date
- season (springer, summer, fall, winter)
- year
- month
- hour
- holiday (weather day is holiday or not)
- Weekday
- working day: if day is neither weekend nor holiday is 1, otherwise is 0.
- Weather (Clear, Mist, Light snow or rain, heavy snow, or rain)
- Temperature in Celsius
- Feeling temperature in Celsius
- Humidity
- Windspeed
- Count of casual users of bikes
- Count of registered users of bikes
- Count of total users of bikes

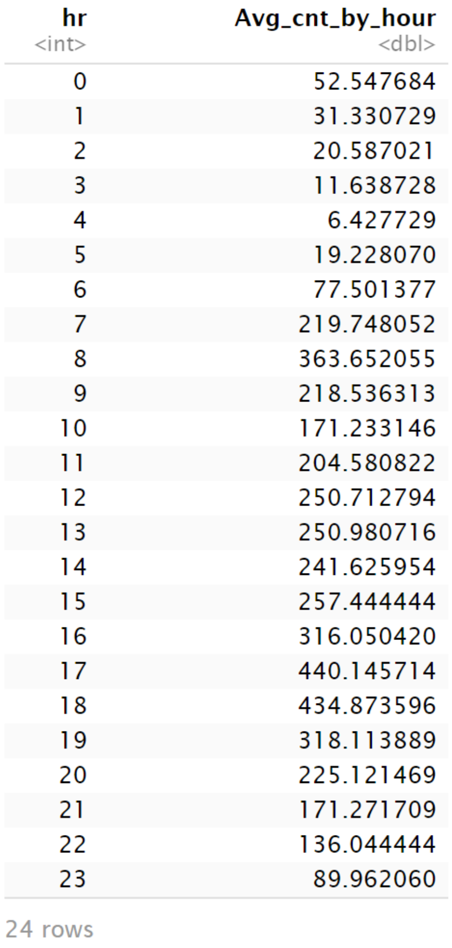
Prior to performing any explanatory analysis, the dataset was split into a sample of 50%, which left us with 8689 hourly observations for our investigation.

Explanatory analysis

The sampling was performed by the following:

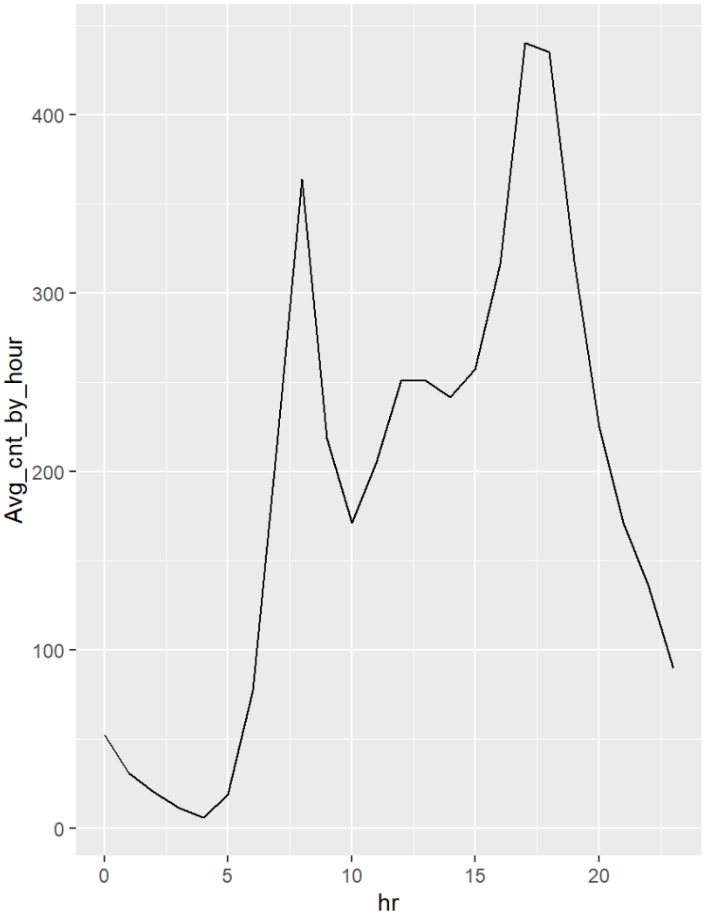
```
smp_size <- floor(0.5 * nrow(data))
set.seed(123)
train_ind <- sample(seq_len(nrow(data)), size = smp_size)

train <- data[train_ind, ]
test = data[-train_ind, ]

nrow(data)
nrow(train)
nrow(test)
```

After sampling 50% of the available dataset, we performed some descriptive data analysis of the dataset. One of our initial assumptions with this dataset was that the number of bikes used would vary heavily depending on the time of the day. The summary statistic for each subset further supports this claim. Therefore, we explored the average number of bikes used by grouping for each hour of the day. Below is a table and chart visualizing the trend of bikes used based on the hour of the day.



hr	Avg_cnt_by_hour
<int>	<dbl>
0	52.547684
1	31.330729
2	20.587021
3	11.638728
4	6.427729
5	19.228070
6	77.501377
7	219.748052
8	363.652055
9	218.536313
10	171.233146
11	204.580822
12	250.712794
13	250.980716
14	241.625954
15	257.444444
16	316.050420
17	440.145714
18	434.873596
19	318.113889
20	225.121469
21	171.271709
22	136.044444
23	89.962060

24 rows

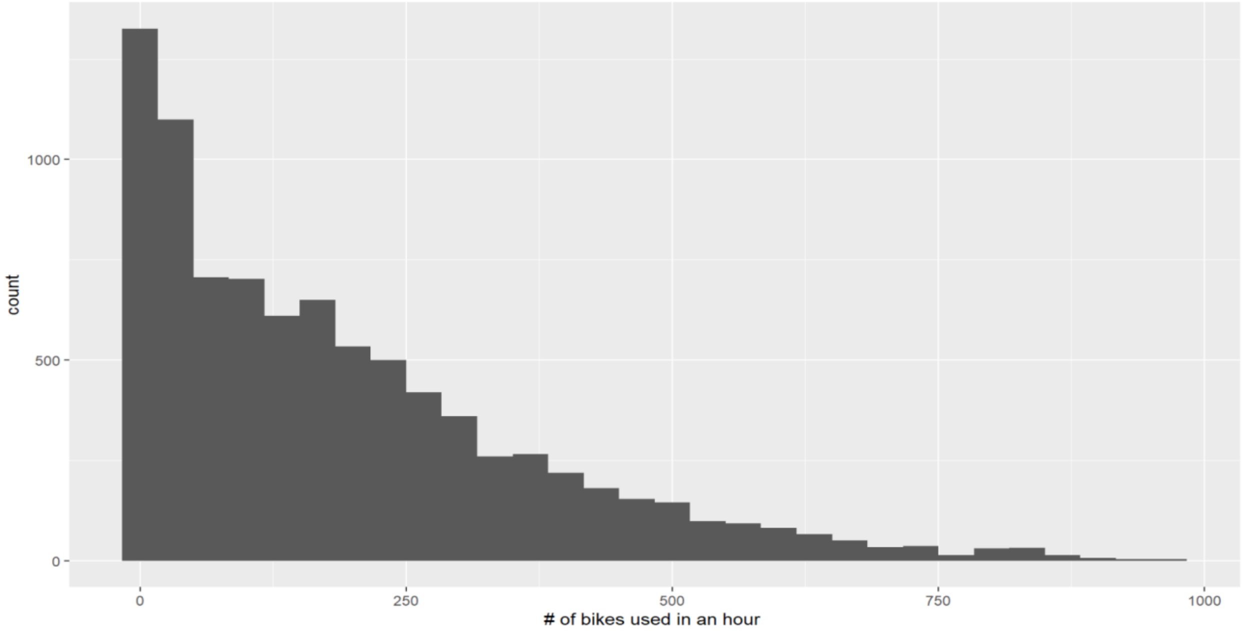


As expected, we can see that the hour of the day matters notably for the number of bikes used. Between 11pm to 6am, the night hours, there is particularly low usage of bikes and during the rush hours of 8am and 5-6pm the demand peaks.

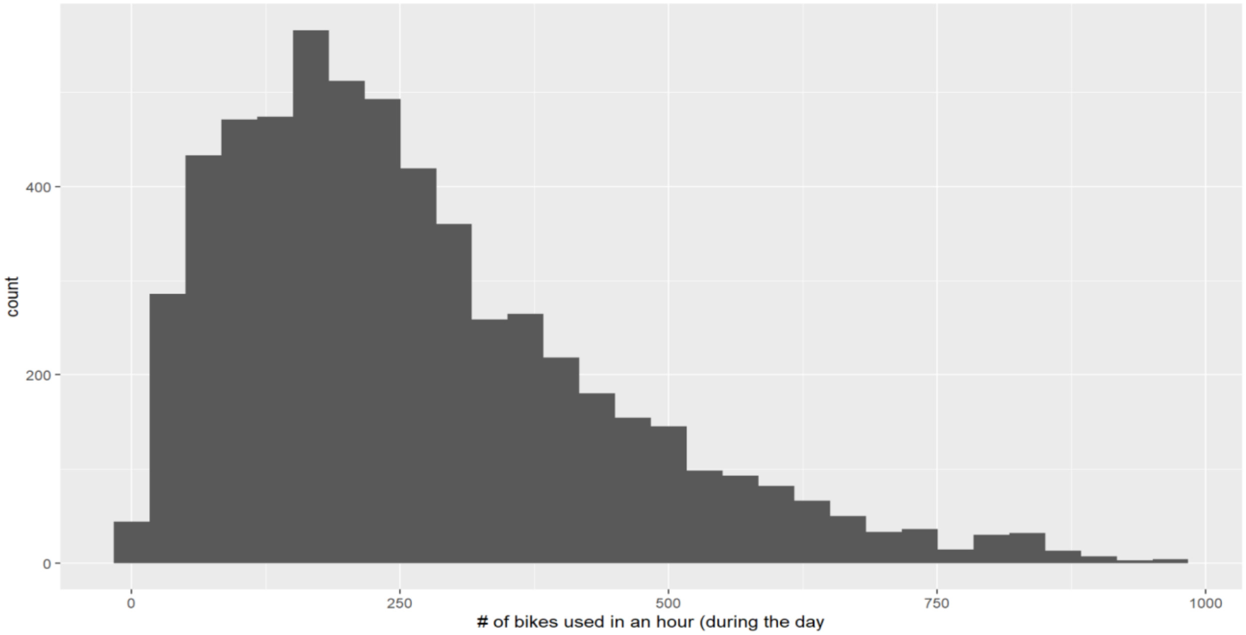
In our research we are most interested in the effect of temperature on bike usage during the day. As this relationship might differ during the night when bike usage is low, we decided to create a new subset of the data sample which filtered out the hour of the night between 11pm and 6am. Furthermore, we created a new binary variable called “rush hour” where 1 indicated the datapoint was either at 8, 17 or 18 o'clock and 0 otherwise.

The histograms below show the distribution of count of bikes by the hour before and after we created the daytime sample dataset.

Bikes in an hour full sample data:

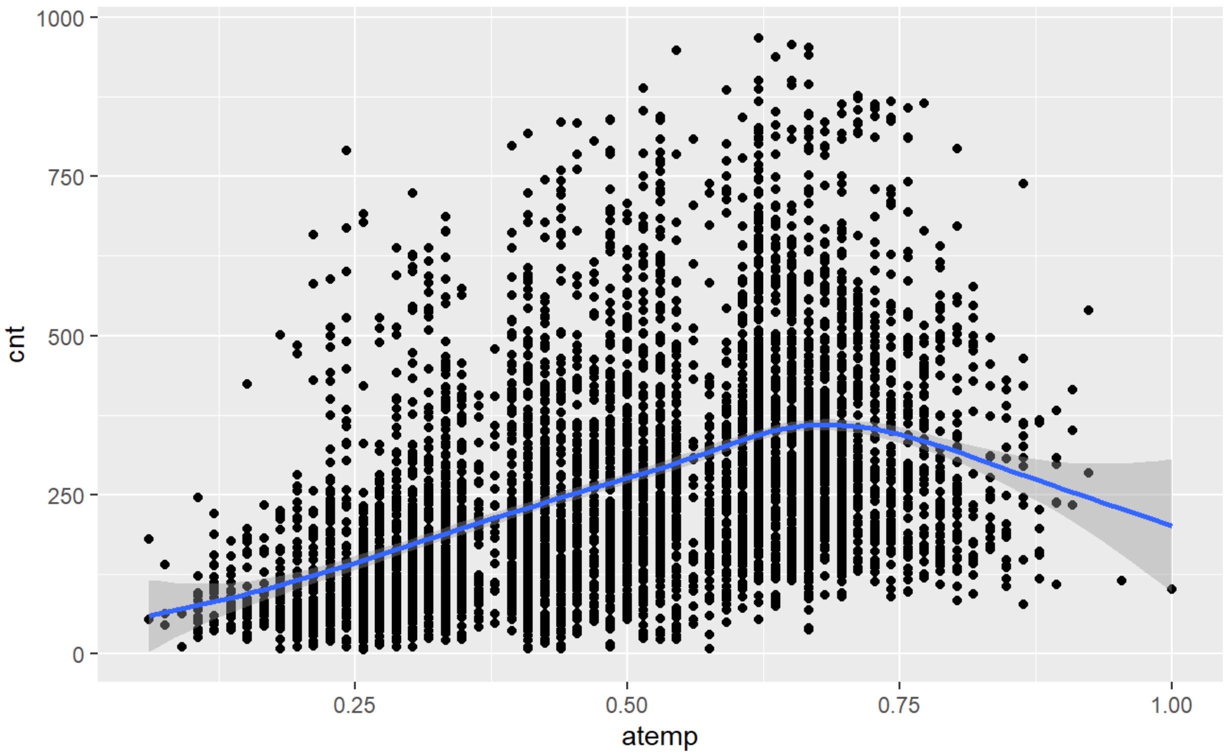


Bikes used in an hour daytime sample:

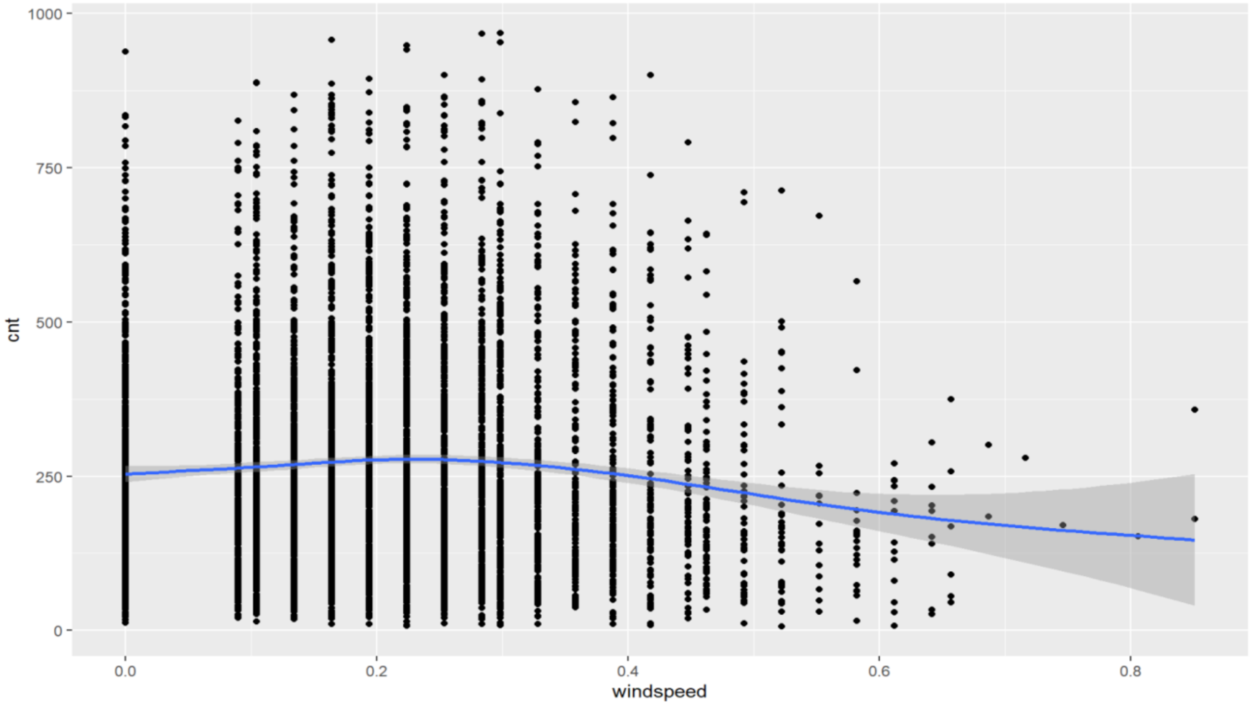


The difference between the histograms shows that the nightly hours were heavily represented in our initial data and would have skewed our estimations. The full sample dataset resembles the characteristics of an exponential distribution while the daytime sample looks more like a gamma distribution.

To get an initial idea of the relationship between the experienced temperature and count of bikes we created a scatterplot for a visual inspection. The plot supports the hypothesis that there is a correlation between temperature and count of bikes used. There also appears to be a peak in the relationship from which point the inverse applies. Our theory is that the declining relationship occurs when the temperature gets uncomfortably warm, and it is no longer enjoyable for people to be biking outside.



Plotting the windspeed scatterplot showed us that windspeed appears to have been measured as a categorical variable, with ranges of windspeed. The relationship between count of bikes used and windspeed is not as immediately apparent as the one with temperature. Nonetheless there might be a negative relationship between the variables which we can explore further in the regression. As windspeeds increases we see somewhat of a decline in the number of bikes used.



Summary Statistics

Mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

The Sample Mean for the:

Full sample => 189.78.

Daytime sample => 266.18

Rush hour sample => 412.32

Confidence Intervals

We can calculate the confidence interval of the mean through:

$$M_E = t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}$$

$$Lower\ bound\ =\ (\overline{x} - M_E),\ Upper\ bound\ =\ (\overline{x} + M_E)$$

The 95% confidence interval for the:

Full sample => (185.9939, 193.5759)

Daytime sample => (261.5016, 270.8657)

Rush hour sample => (398.3691, 426.2788)

As expected, we can see that the confidence interval increases as the sample size decreases.

We can calculate the sample standard deviation through:

$$s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

The standard deviation for the

Full sample => 180.2739

Daytime sample => 176.3964

Rush hour sample => 232.745

The confidence interval of the sample standard deviation is calculated as

$$\sqrt{\frac{(n-1)\cdot s^2}{x^2_{\frac{\alpha}{2},n-1}}} < \sigma < \sqrt{\frac{(n-1)\cdot s^2}{x^2_{1-\frac{\alpha}{2},n-1}}}$$

Degrees of freedom = n-1

The 95% confidence interval for the:

Full sample => (177.633, 182.995)

Daytime sample => (173.1476, 179.7703)

Rush hour sample => (223.2888, 243.0438)

Fitting the data towards the Daytime sample:

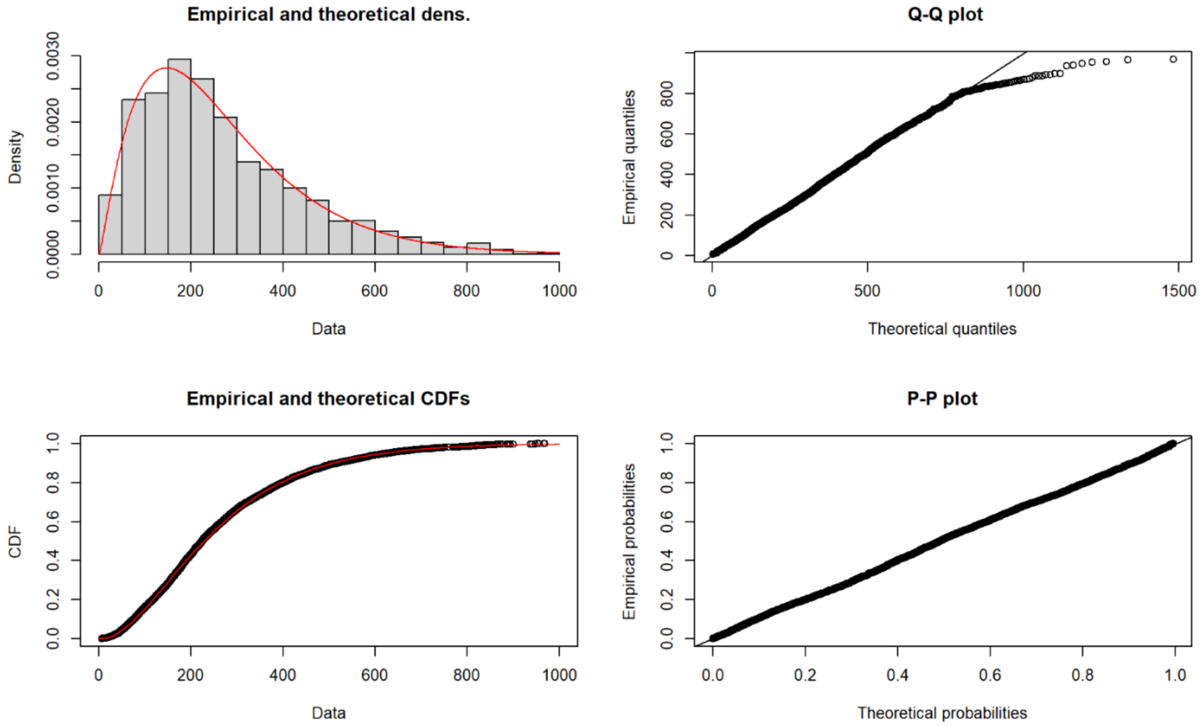
After examining the data's distribution in a histogram, we were curious to find which distribution best fit our data. From visualizing the shape of our histogram, we assumed the distribution could be represented as a gamma distribution. Using the fitdist function from the fitdistplus package to estimate the alpha and beta parameters of a gamma distribution led to the following best fit for our data.

We used MOM method to estimate the parameters alpha and beta which does so through the following.

$$E(x) = \frac{\alpha}{\beta},\ V(x) = \frac{\alpha}{\beta^2} \Longrightarrow \beta = \frac{E(x)}{V(x)},\ \alpha = \beta E(x) = \frac{\left(E(x)\right)^2}{V(x)} \Longrightarrow \overline{\beta} = \frac{\overline{x}}{s^2},\ \overline{\alpha} = \frac{(\overline{x})^2}{s^2}$$

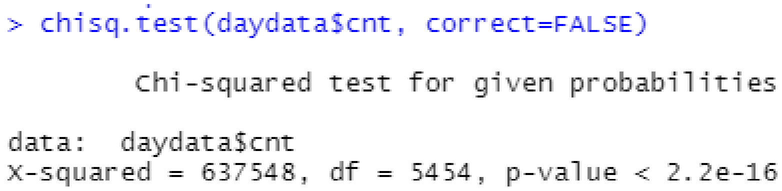
```
Parameters:
      estimate
shape 2.277525178
rate  0.008556216
```

Using the shape and rate, the expected value for the count of bikes in the day comes out to be approximately 266.18369 bikes

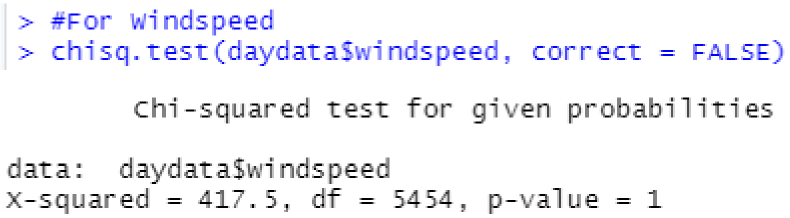


The distribution parameters do not give us too much insight too our research questions at this stage. However, it helped us understand how well the data we used was fitting to a distribution which potentially could be used for future research and more advanced predictive models that could further benefit Capital Bikeshares.

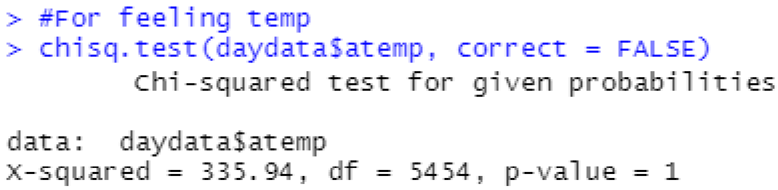
Chi Squared:



```
> chisq.test(daydata$cnt, correct=FALSE)

        Chi-squared test for given probabilities

data:  daydata$cnt
X-squared = 637548, df = 5454, p-value < 2.2e-16
```



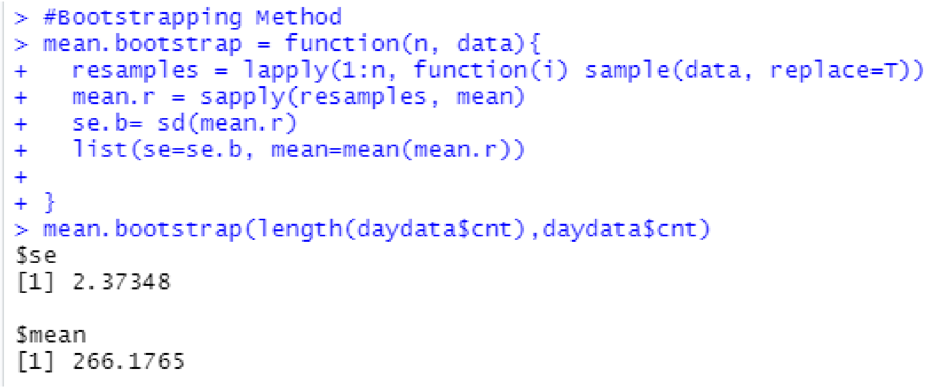
```
> #For Windspeed
> chisq.test(daydata$windspeed, correct = FALSE)

        Chi-squared test for given probabilities

data:  daydata$windspeed
X-squared = 417.5, df = 5454, p-value = 1
```



```
> #For feeling temp
> chisq.test(daydata$atemp, correct = FALSE)
        Chi-squared test for given probabilities

data:  daydata$atemp
X-squared = 335.94, df = 5454, p-value = 1
```

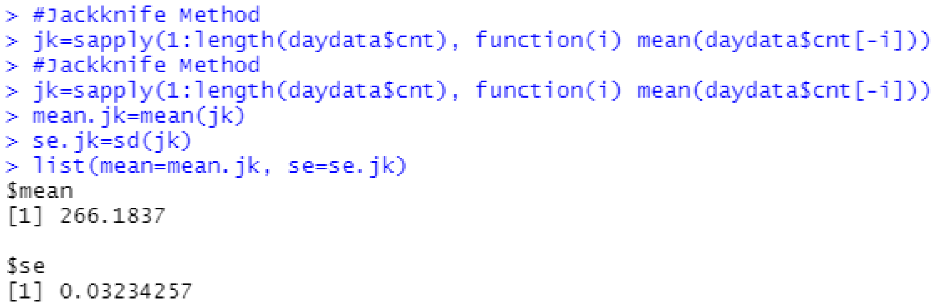
After running the Chi Squared test for daytime bike count, windspeed, and feeling temperature, the X-Squared value, or error squared were rather high for the data used. And the p-values derived from testing windspeed and feeling temperature were 1, which identifies extremely high insignificance. Therefore, we found this test to not be helpful in further analysis of our research question. The reason for using the Chi Squared test was to see if our expectations of the data set following a gamma distribution to be incorrect, which they were not.

Bootstrapping Method:



```
> #Bootstrapping Method
> mean.bootstrap = function(n, data){
+   resamples = lapply(1:n, function(i) sample(data, replace=T))
+   mean.r = sapply(resamples, mean)
+   se.b= sd(mean.r)
+   list(se=se.b, mean=mean(mean.r))
+ 
+ }
> mean.bootstrap(length(daydata$cnt),daydata$cnt)
$se
[1] 2.37348

$mean
[1] 266.1765
```

Jackknife Method:



```
> #Jackknife Method
> jk=sapply(1:length(daydata$cnt), function(i) mean(daydata$cnt[-i]))
> #Jackknife Method
> jk=sapply(1:length(daydata$cnt), function(i) mean(daydata$cnt[-i]))
> mean.jk=mean(jk)
> se.jk=sd(jk)
> list(mean=mean.jk, se=se.jk)
$mean
[1] 266.1837

$se
[1] 0.03234257
```

The Bootstrap method was our first computational estimation method that we ran, resampling the entire size of the day data bike count dataset and the mean was extremely close to our expected sample mean, with a standard error of 2.37348. This seemed reasonable and a good estimator for our data, but when we used the jackknife method, the results were significantly more precise. The mean derived from the jackknife method was similar to the mean produced by the bootstrap method, but the standard error was minimized to 0.03234257, an improvement.

Regression Analysis:

Simple linear regression:

Feeling temperature regression

```
Call: lm(formula = cnt ~ atemp, data = daydata)
Residuals:
   Min      1Q  Median      3Q     Max 
-382.10 -106.92  -37.72   70.69  661.83 

Coefficients:
           Estimate Std. Error t value Pr(>|t|)    
(Intercept)   49.819      6.506   7.658 2.22e-14 ***
atemp        433.277     12.292  35.249  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 159.2 on 5453 degrees of freedom
Multiple R-squared:  0.1856,	Adjusted R-squared:  0.1854 
F-statistic:  1243 on 1 and 5453 DF,  p-value: < 2.2e-16
```

By running a regression only using the feeling temperature (atemp) as an independent variable we are able to explain 18,5% of the variation in count of bikes (cnt). The feeling temperature is statistically significant at 99% confidence level and has a strong positive correlation to counts of bikes used during the daytime. Our model produced an overall large F statistic and a p-value close to zero which would lead us to reject H0 in favor of the alternative, that there is significant evidence of a relationship between feeling temperature and bike count.

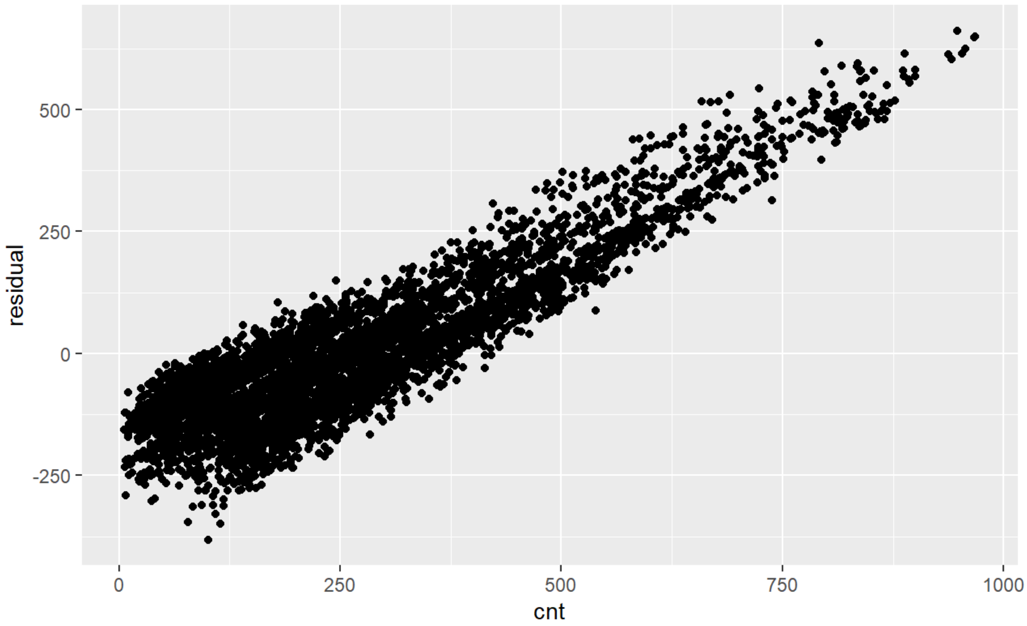
Residuals testing model 1:

Durbin Watson Test

```
lag Autocorrelation D-W Statistic p-value
  1     0.008971492       1.98175   0.528
Alternative hypothesis: rho != 0
```

The Durbin Watson test fails to reject H0, thereby supporting the claim that the residuals are not autocorrelated.

Constant Variance

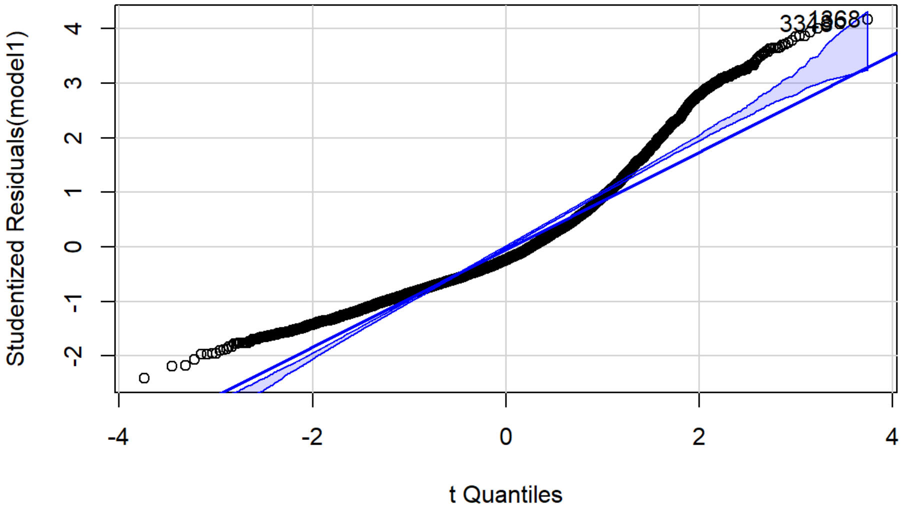


As seen by the visual above there might be a problem with constant variance. A Breusch-Pagan test was run to examine this further.

```
	studentized Breusch-Pagan test

data:  model1
BP = 136.81, df = 1, p-value < 2.2e-16
```

The Breusch-Pagan test rejects H0 in favor of the alternative and thereby supports the claim that there is a heteroskedasticity problem in the regression.

Normality



The residuals somewhat follow a normal distribution, but the model does not appear to be fully capturing a linear representation of normality. In this instance, further analysis should be conducted to see which distribution the residuals follow.

Windspeed regression

```
Call:
lm(formula = cnt ~ factor(windspeed), data = daydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-273.38 -127.73  -39.38   91.20  690.16 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              255.428      7.322  34.886   <2e-16 ***
factor(windspeed)0.0896   -8.123     12.210  -0.665   0.5059    
factor(windspeed)0.1045   18.762     11.281   1.663   0.0963 .  
factor(windspeed)0.1343    1.807     10.915   0.166   0.8685    
factor(windspeed)0.1642   24.192     10.640   2.274   0.0230 *  
factor(windspeed)0.194    24.559     10.546   2.329   0.0199 *  
factor(windspeed)0.2239   24.948     10.476   2.381   0.0173 *  
factor(windspeed)0.2537   20.107     11.007   1.827   0.0678 .  
factor(windspeed)0.2836   24.072     11.853   2.031   0.0423 *  
factor(windspeed)0.2985   22.410     12.411   1.806   0.0710 .  
factor(windspeed)0.3284   -5.897     13.890  -0.425   0.6712    
factor(windspeed)0.3582    6.714     16.205   0.414   0.6787    
factor(windspeed)0.3881    3.609     15.714   0.230   0.8183    
factor(windspeed)0.4179    4.370     18.026   0.242   0.8085    
factor(windspeed)0.4478  -12.037     22.403  -0.537   0.5911    
factor(windspeed)0.4627  -36.542     22.260  -1.642   0.1007    
factor(windspeed)0.4925  -42.408     26.170  -1.620   0.1052    
factor(windspeed)0.5224  -39.399     31.039  -1.269   0.2044    
factor(windspeed)0.5522  -68.642     47.572  -1.443   0.1491    
factor(windspeed)0.5821  -92.323     41.008  -2.251   0.0244 *  
factor(windspeed)0.6119 -106.197     49.326  -2.153   0.0314 *  
factor(windspeed)0.6418  -94.803     62.611  -1.514   0.1300    
factor(windspeed)0.6567  -90.095     72.174  -1.248   0.2120    
factor(windspeed)0.6866  -12.428    124.579  -0.100   0.9205    
factor(windspeed)0.7164   24.572    176.029   0.140   0.8890    
factor(windspeed)0.7463  -84.428    176.029  -0.480   0.6315    
factor(windspeed)0.806  -103.428    176.029  -0.588   0.5569    
factor(windspeed)0.8507   14.072    124.579   0.113   0.9101    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 175.9 on 5427 degrees of freedom
```

```
Multiple R-squared:  0.01081,     Adjusted R-squared:  0.005884
F-statistic: 2.196 on 27 and 5427 DF,  p-value: 0.0003463
```

Through this regression model, we see that overall windspeed is significant, but we observed only six variables of windspeed that are significant (0.1642, 0.194, 0.2239, 0.2836, 0.5821, 0.6119). Compared to earlier models ran for linear regression, we see that the F-statistic for windspeed factor is rather low and the adjusted R-Square suggests that we are only explaining ~0.5% of the variation in our data. This leads us to believe that the overall impact of windspeed is less impactful than atemp. With such a minimal Adjusted R-Squared value for windspeed, we decided not to perform residual diagnostic tests as we couldn’t be certain that the variation is true.

Multiple variable regression.

```
lm(formula = cnt ~ atemp + atemp2 + factor(rushhour) + factor(windspeed) + 
    atemp * factor(rushhour), data = daydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-487.42  -88.93  -20.08   73.14  556.61 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -103.066     15.128  -6.813 1.06e-11 ***
atemp                  980.841     59.102  16.596  < 2e-16 ***
atemp2                -607.122     59.174 -10.260  < 2e-16 ***
factor(rushhour)1       48.527     14.077   3.447  0.00057 ***
factor(windspeed)0.0896    4.652      9.577   0.486  0.62716    
factor(windspeed)0.1045   12.763      8.851   1.442  0.14938    
factor(windspeed)0.1343    1.964      8.560   0.229  0.81849    
factor(windspeed)0.1642   24.364      8.351   2.918  0.00354 ** 
factor(windspeed)0.194    13.697      8.275   1.655  0.09794 .  
factor(windspeed)0.2239   20.409      8.219   2.483  0.01306 *  
factor(windspeed)0.2537   16.278      8.641   1.884  0.05966 .  
factor(windspeed)0.2836   22.544      9.298   2.425  0.01536 *  
factor(windspeed)0.2985   17.781      9.758   1.822  0.06849 .  
factor(windspeed)0.3284   18.571     10.920   1.701  0.08908 .  
factor(windspeed)0.3582   28.202     12.727   2.216  0.02674 *  
factor(windspeed)0.3881   17.629     12.326   1.430  0.15271    
factor(windspeed)0.4179   24.288     14.135   1.718  0.08581 .  
factor(windspeed)0.4478    8.706     17.574   0.495  0.62033    
factor(windspeed)0.4627    8.815     17.493   0.504  0.61432    
factor(windspeed)0.4925  -26.014     20.527  -1.267  0.20510    
factor(windspeed)0.5224   -3.749     24.361  -0.154  0.87769    
factor(windspeed)0.5522   -4.000     37.336  -0.107  0.91469    
factor(windspeed)0.5821  -55.815     32.161  -1.735  0.08272 .  
factor(windspeed)0.6119  -76.453     38.681  -1.977  0.04815 *  
factor(windspeed)0.6418  -65.034     49.100  -1.325  0.18539    
factor(windspeed)0.6567   -1.593     56.608  -0.028  0.97755    
factor(windspeed)0.6866   -4.822     97.658  -0.049  0.96062
factor(windspeed)0.7164   38.885    137.987   0.282  0.77810
factor(windspeed)0.7463  -39.697    137.989  -0.288  0.77360
factor(windspeed)0.806   -51.758    137.989  -0.375  0.70761
factor(windspeed)0.8507 -278.632     97.946  -2.845  0.00446 **
atemp:factor(rushhour)1  281.672     26.974  10.442  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 137.9 on 5423 degrees of freedom
Multiple R-squared:  0.3927,	Adjusted R-squared:  0.3892
F-statistic: 113.1 on 31 and 5423 DF,  p-value: < 2.2e-16
```

In our second model we included more variables. In addition to both windspeed and the feeling temperature we added the squared feeling temperature in order to capture decreasing bike use in very warm temperatures, rush hour was also added to capture the effect of demand peaks and the interaction between temperature and rushhour.

By adding these variables our model was able to explain 38.9% of the variation in the data. This is a big improvement to the initial models. All independent variables except for windspeed are statistically significant at 99% confidence level. Like the earlier model, wind speed is statistically significant however only for certain wind levels. Our second model also produced an overall large F statistic and a p-value close to zero. This would lead us to reject H0 and conclude that there is considerable evidence of a relationship between our independent variables and bike count.

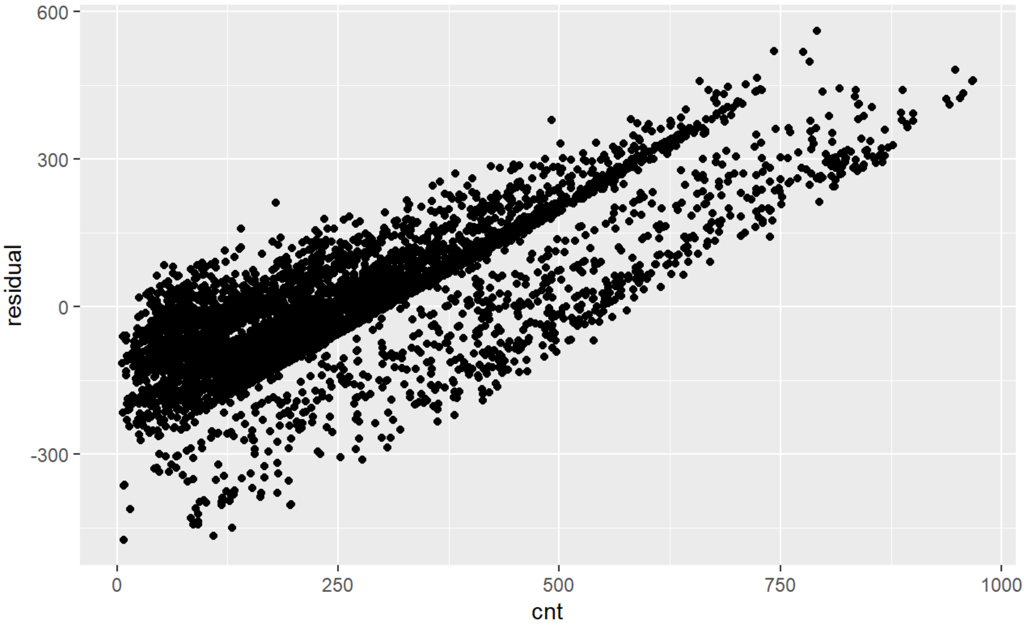
Residuals testing model 2:

Autocorrelation

```
lag Autocorrelation D-W Statistic p-value
   1     0.008971492       1.98175   0.528
Alternative hypothesis: rho != 0
```

The Durbin Watson test fails to reject H0, thereby supporting the claim that the residuals are not autocorrelated.

Constant Variance



Like the model above, the visual suggests there might be heteroskedasticity in our model

```
studentized Breusch-Pagan test

data:  model2
BP = 699.79, df = 31, p-value < 2.2e-16
```

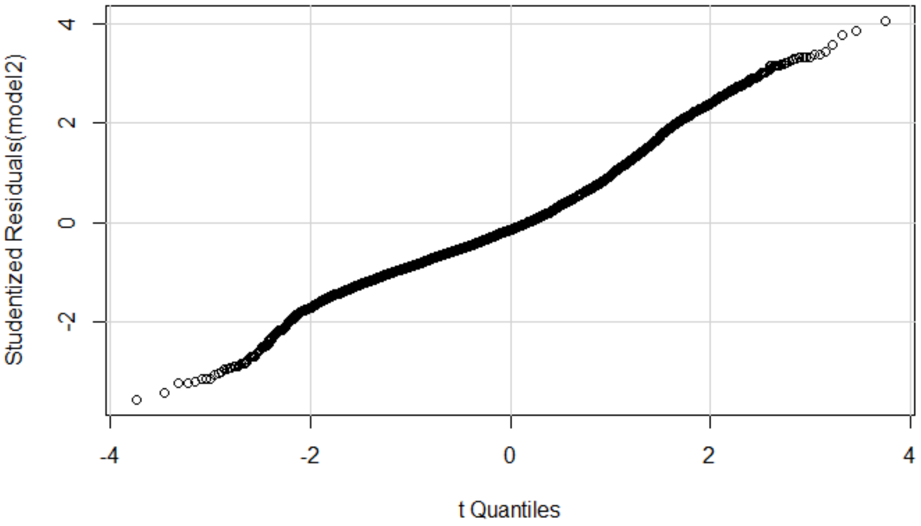
The Breusch –Pagan test rejects H0 for the second model as well, supporting the claim that the regression has an issue with non-constant variance.

Multicollinearity

```
      atemp             atemp2        factor(rushhour)
          30.833580           30.407359              8.974587
atemp:factor(rushhour)  factor(windspeed)
           9.151880                    1.053174
```

The VIF test suggests there is multicollinearity between the feeling temperature (atemp) and its squared variable. We expected this as one is dependent on the other. However, we still find it valuable to keep the squared version of atemp considering the variable helps us capture the negative effect on bike count which excessively warm temperatures entails.

Normality



Overall the residuals appear to have a normal distribution.

Conclusions:

When we refer to our original research question, the necessary range of bikes to accommodate rush hour usage based on a 95% confidence interval was (398.3691, 426.2788) bikes. With further research and additional resources, we would be able to boil this range down to a single number, but due to our current lack of knowledge on the purchase cost of bikes, cost of maintenance, location optimization, and system's revenue, this range allows the Captial Bikeshare system and future researchers to take our findings one step further. Additionally, we did find a positive relationship between bikes used and feeling temperature as well as windspeed and bikes used, however, the impact of feeling temperature was greater than windspeed. The relationship that we view for feeling temperature is the inverse when it's a very warm day, as indicated in earlier scatterplots of feeling temperature.

Our group determined that the future research of this data lies in understanding the financial costs associated with a bike share program and the optimal location for bikes to be held, to allow for the most common users to always have close-proximity access to a bike. With that knowledge and the predictions used, we are certain that an optimal bike count to accommodate rush hour can be achieved