Bikes

Sean Jin
seanj

Introduction

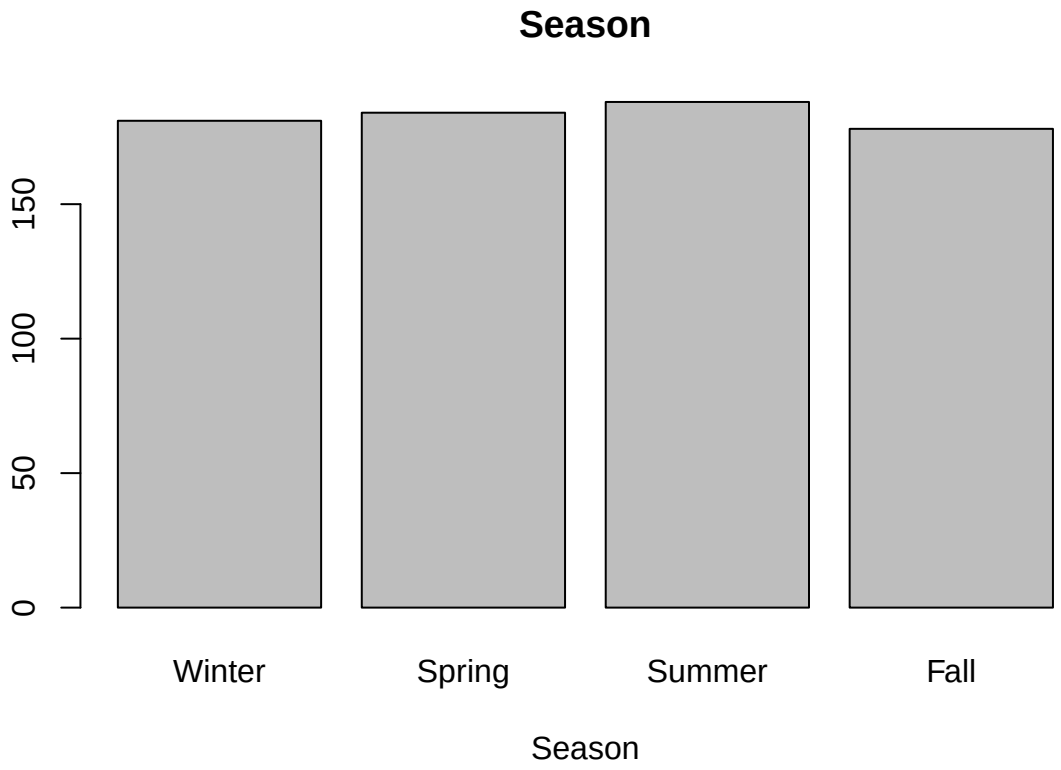
Bike sharing systems is the current generation of bike rentals as rental and return became automatic. The systems allow its user to easily rent a bike from one place and be able to return it at any other place. There are more than 500 bike-sharing programs that contain 500,00 bicycles. We want to have a better understanding of the use of these systems in how they play a role in traffic, evironmental, and health issues. A way in order to indentify changes in regard to duration of travel, departure, and arrival position is recorded in the systems. But, it is important for us to figure out how these bike sharing systems are catagorizing the information and understand the information that they supply and find out the total number of users of any given day.

Exploratory Data Analysis and Data Cleaning

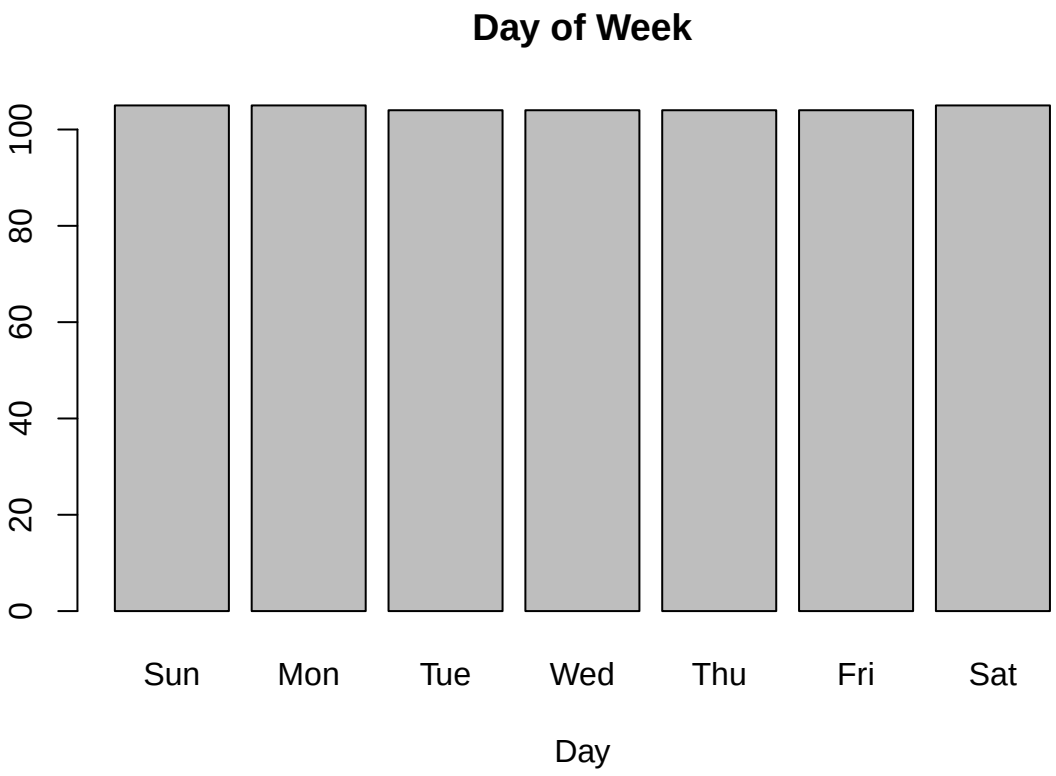
Variables

Season	DayOfWeek	TempFeel	TotalUsers
Winter	Sat	18	985
Winter	Sun	18	801
Winter	Mon	9	1349
Winter	Tue	11	1562
Winter	Wed	11	1600
Winter	Thu	12	1606

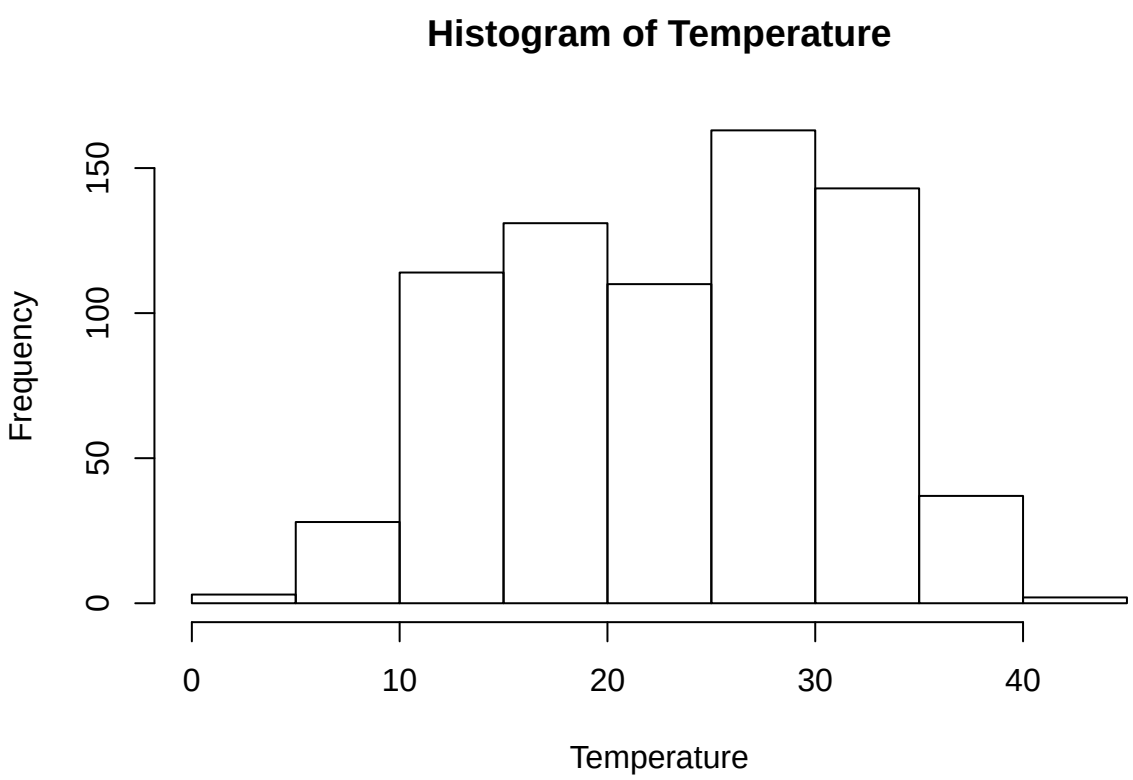
Univariate Analysis



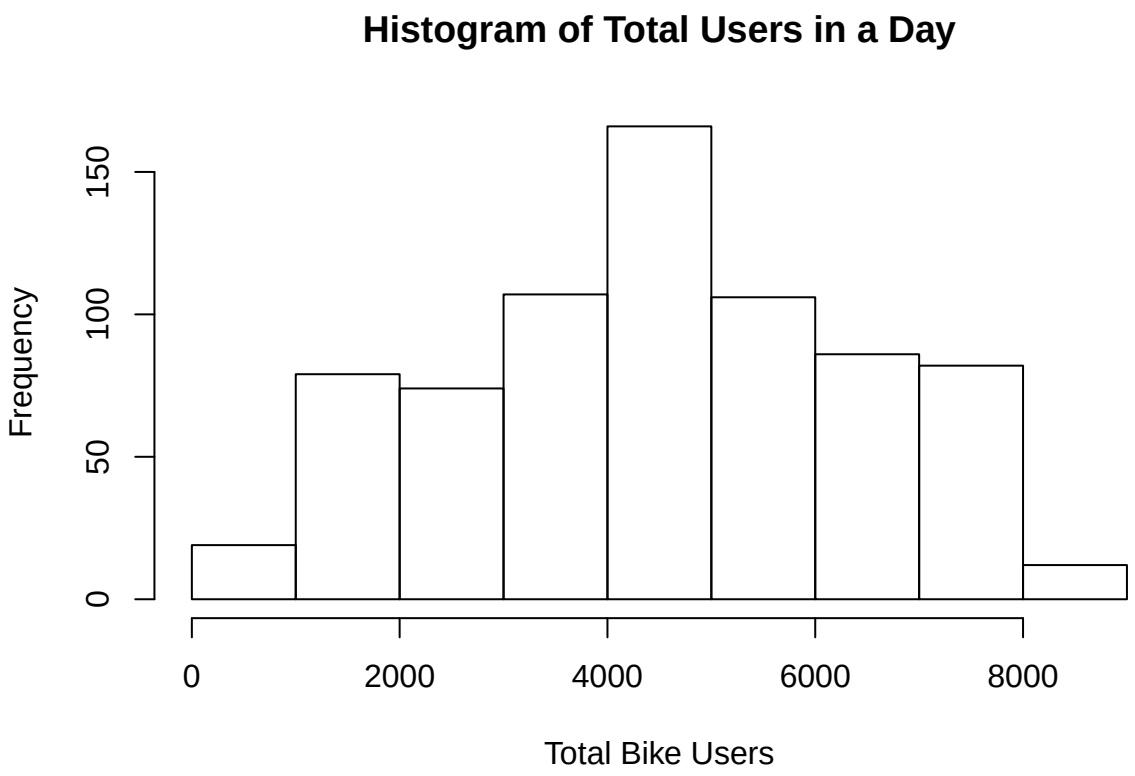
Winter	Spring	Summer	Fall
181	184	188	178



Mon	Tue	Wed	Thur	Fri	Sat	Sun
105	104	104	104	104	105	105



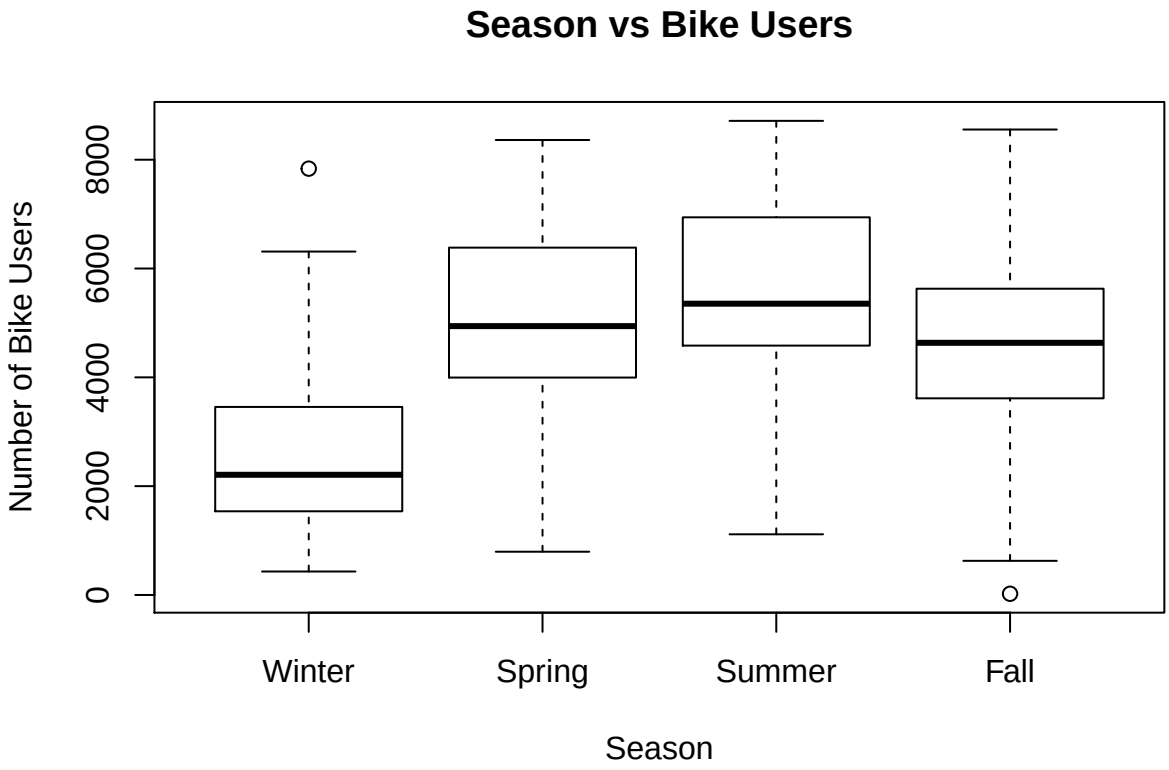
Min	4.00
1Q	17.00
Med	24.00
Mean	23.74
3Q	30.00
Max	42.00



Min	22
1Q	3152
Med	4548
Mean	4504
3Q	5956
Max	8714

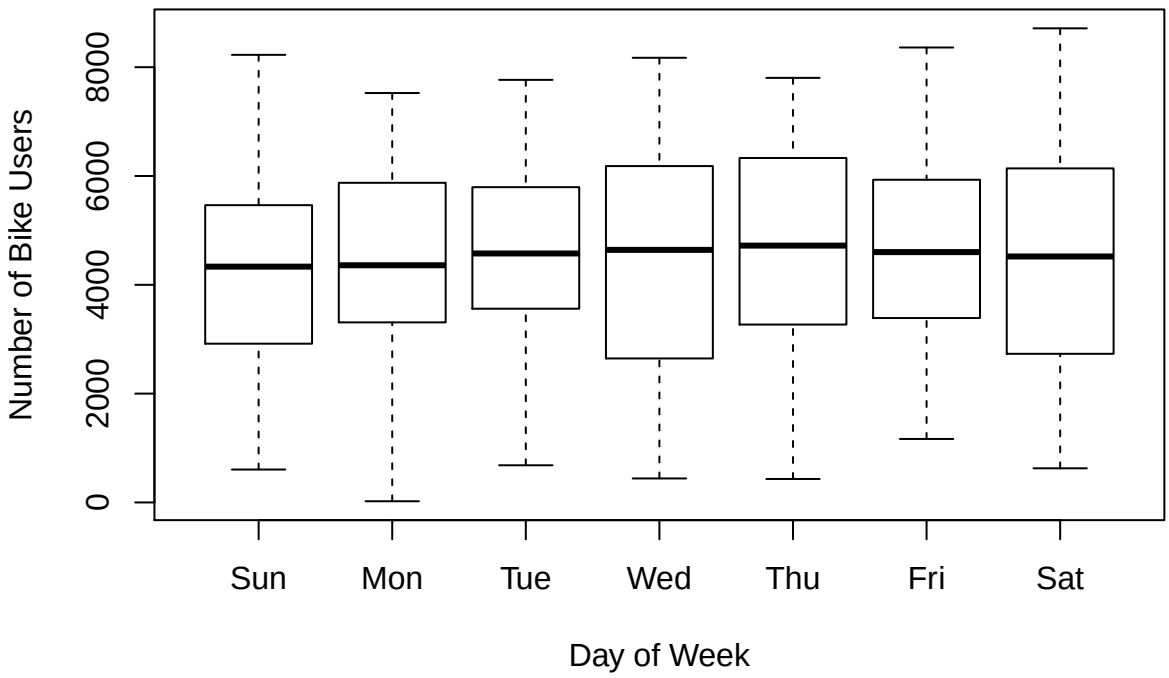
There is an even distribution between all of the seasons as well all of the days of the weekas and that there is enough data of each season for normality since there is a huge sample size. Notice that for temperature we have a range of 38 with a min of 4 and a max of 42. The median and mean are similarly close and there is an IQR of 13; this data is slightly bimodal, normally distrubted, and without skew. Finally, for the total number of users, they have a range of 8692 with a minimum of 22 and a maximum of 8714. The median and mean are again very closeby and there is an IQR of 2804; the data is unimodal, normally distributed, and has no skew.

Bivariate Exploration



	Winter	Spring	Summer	Fall
Min	431	795	1115	22
Med	2209	4942	5354	4634
Mean	2604	4992	5644	4728
SD	1400	1696	1460	1700
IQR	1918	2374	2343	2009
Max	7836	8362	8714	8555

Notice that Summer had the highest amount of users whereas winter has the lowest. Spring and Fall were similarly close to Summer which makes sense since Biking during cold weather, especially in the winter, unfeasible. Thus, winter had a drastic drop compared to the other seasons. There was a large range between the total user where the values for each season in regards of range was between 7405 and 8533 while the IQR for each season was between 1918 and 2374. This graph indicates that people are less likely to bike in the winter than any of the other seasons.

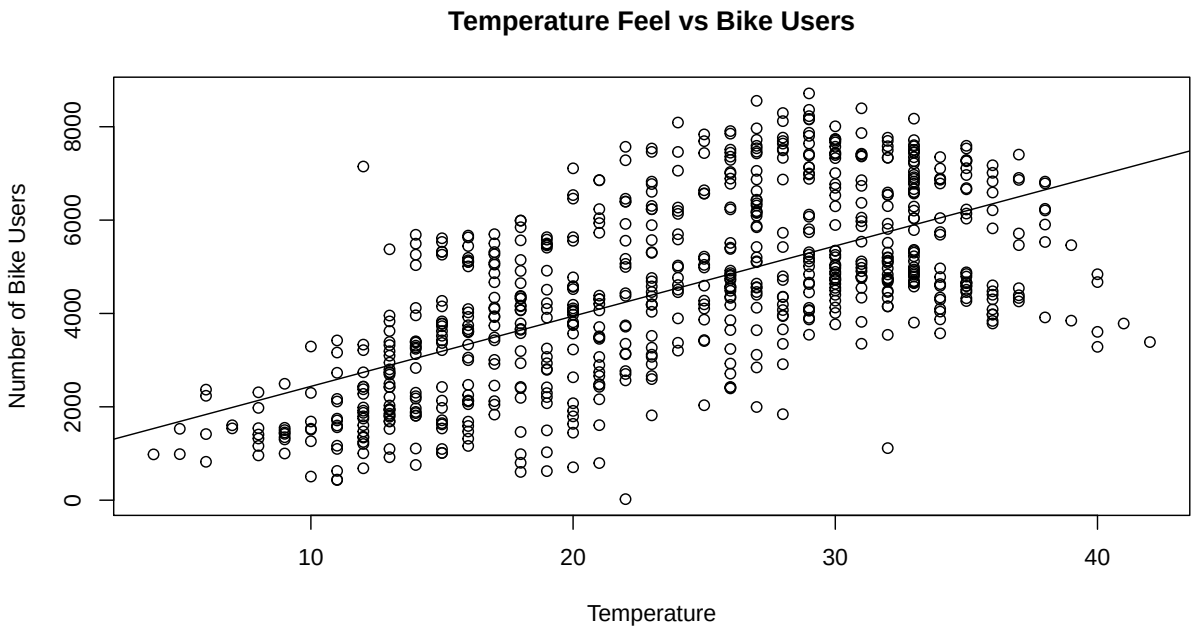


	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
Min	605	22	683	441	431	1167	627
Med	4334	4359	4576	4642	4721	4602	4521
Mean	4229	4338	4511	4549	4667	4690	4551
SD	1872	1793	1827	2038	1939	1875	2197
IQR	2546	2565	2190	3522	2510	2510	3408
Max	8227	7525	7767	8173	8362	8362	8714

There seems no distinguishable difference in the total users for any day of the week as the means and medians for all of the days were within a close range of one another. The values for the range for each day of the week are between 7084 and 8087 while the IQR for each day of the week are between 2100 and 3400. There is no evident relationship between the day of the week and the total users.

```
##
## Call:
## lm(formula = (TotalUsers) ~ TempFeel, data = bikes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4632.6 -1104.1  -104.7  1068.9  4409.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  933.381    171.208   5.452 6.83e-08 ***
## TempFeel     150.445      6.823  22.050  < 2e-16 ***
```

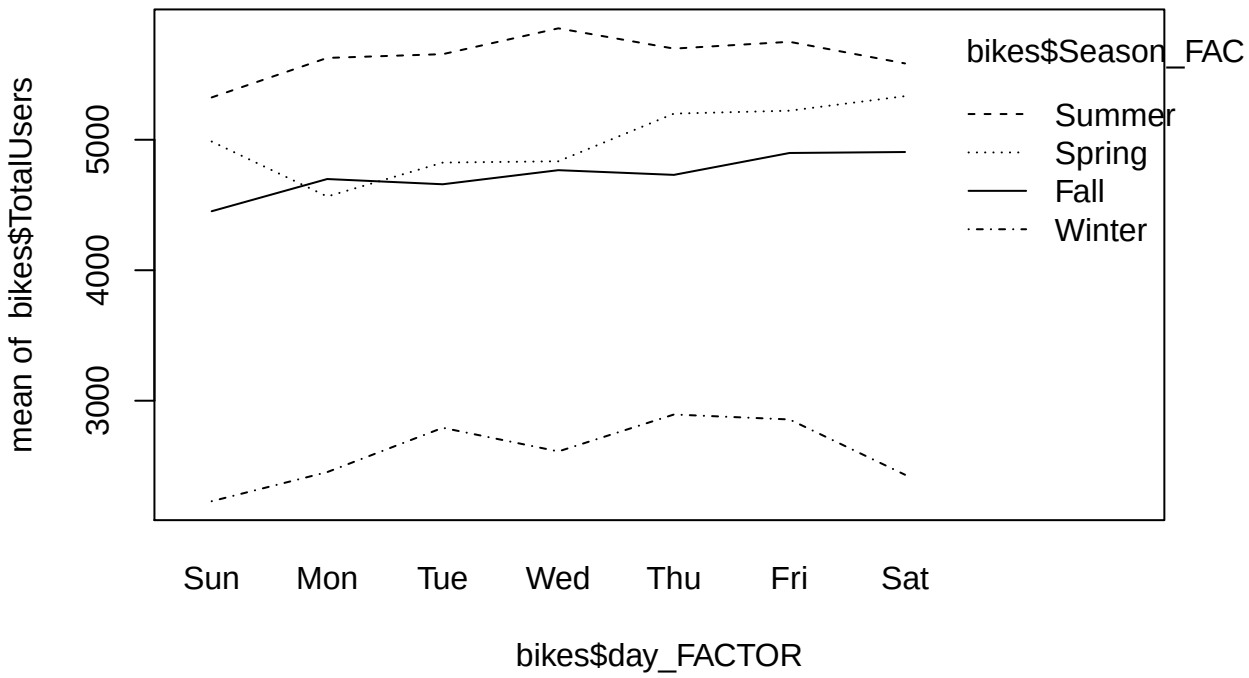
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1501 on 729 degrees of freedom
## Multiple R-squared:  0.4001, Adjusted R-squared:  0.3993
## F-statistic: 486.2 on 1 and 729 DF,  p-value: < 2.2e-16
```

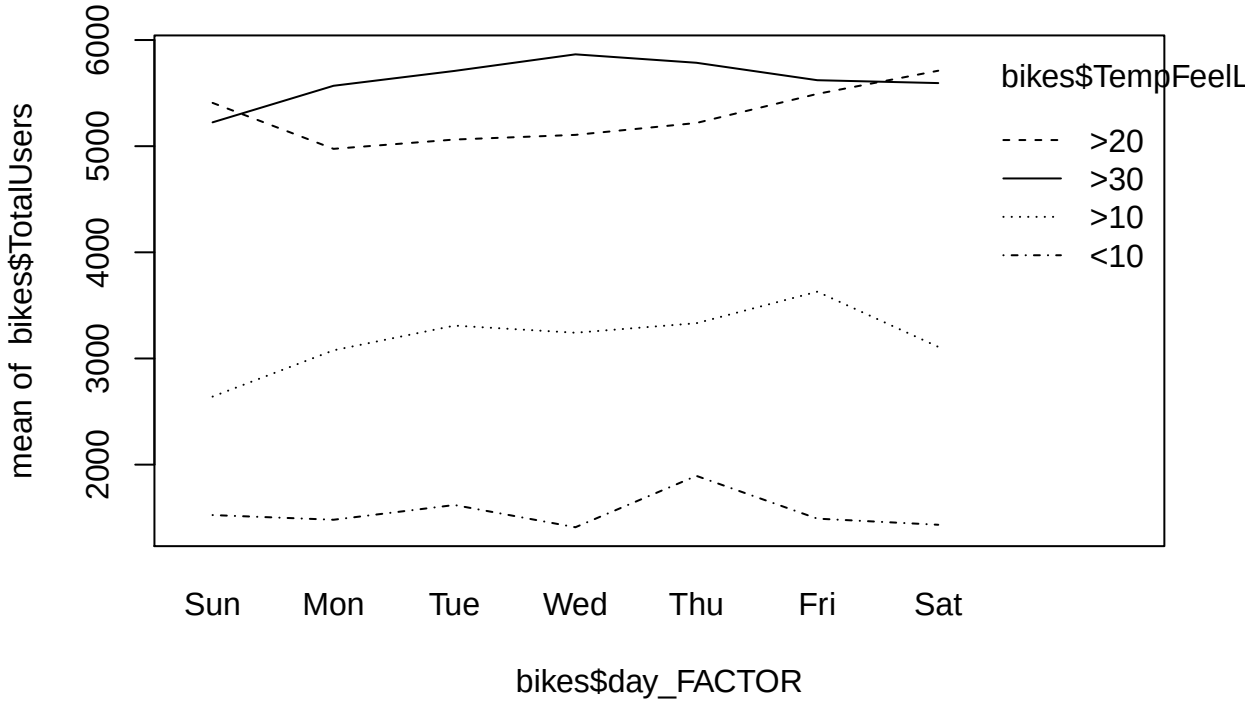


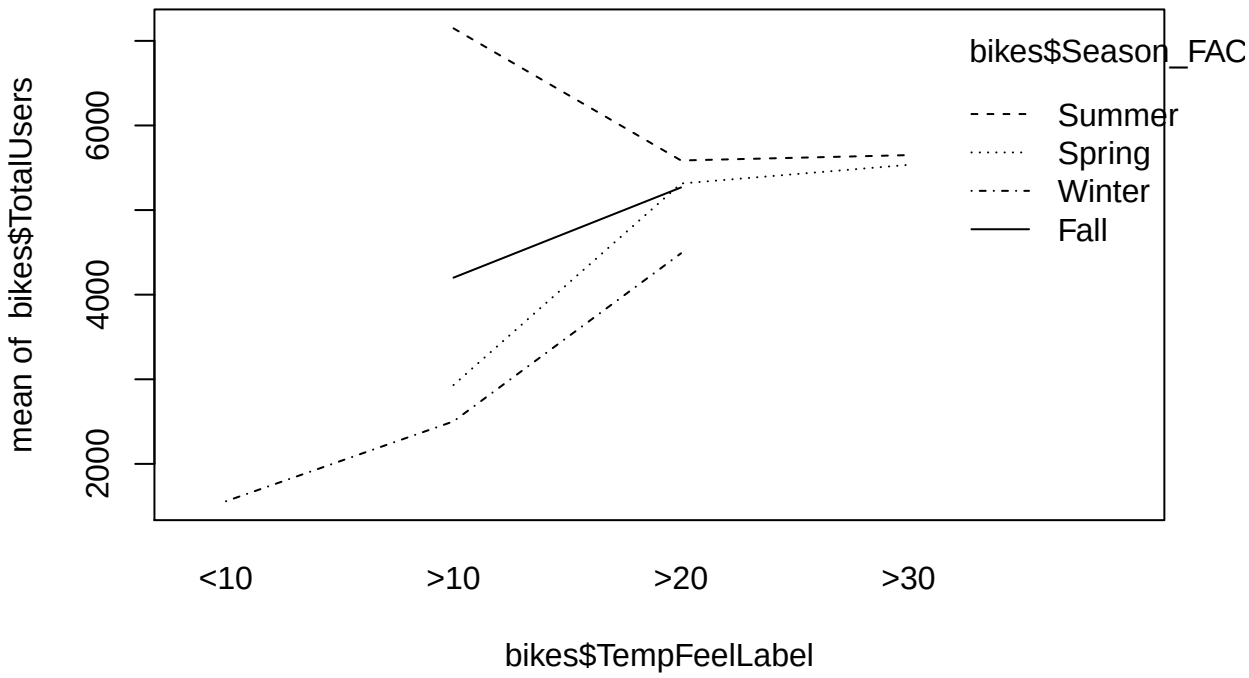
According to the graph above, there is a positive relationship between the temperature and the total number of users. Note that as the weather becomes warmer that people are more likely to use the bike systems.

```
TempFeel.vect <- rep(0, nrow(bikes))

bikes$TempFeelLabel <- TempFeel.vect


bikes$TempFeelLabel <- ifelse(bikes$TempFeel > 30,
                              ">30",
                                ifelse(bikes$TempFeel > 20,
                                ">20",
                                  ifelse(bikes$TempFeel > 10,
                                  ">10", "<10")))

crossMeans(bikes, f1 = "Season_FACTOR", f2 = "day_FACTOR", y = "TotalUsers")
```

```
##           Sun    Mon    Tue    Wed    Thu    Fri    Sat
## Winter 2229.4 2452.7 2792.5 2611.1 2894.2 2856.3 2432.3
## Spring 4986.5 4565.0 4825.0 4835.1 5200.3 5222.8 5334.2
## Summer 5324.3 5627.2 5656.1 5853.9 5698.6 5750.6 5585.0
## Fall   4452.0 4698.8 4658.9 4766.8 4730.9 4898.7 4906.0
```


mean of bikes$TotalUsers
3000
4000
5000
Sun
Mon
Tue
Wed
Thu
Fri
Sat
bikes$day_FACTOR
bikes$Season_FAC
Summer
Spring
Fall
Winter


mean of bikes$TotalUsers
2000
3000
4000
5000
6000
Sun
Mon
Tue
Wed
Thu
Fri
Sat
bikes$day_FACTOR
bikes$TempFeelL
>20
>30
>10
<10



Data Cleaning

The data is normal with no major outliers. Thus, there is no need for any data cleaning for the given model.

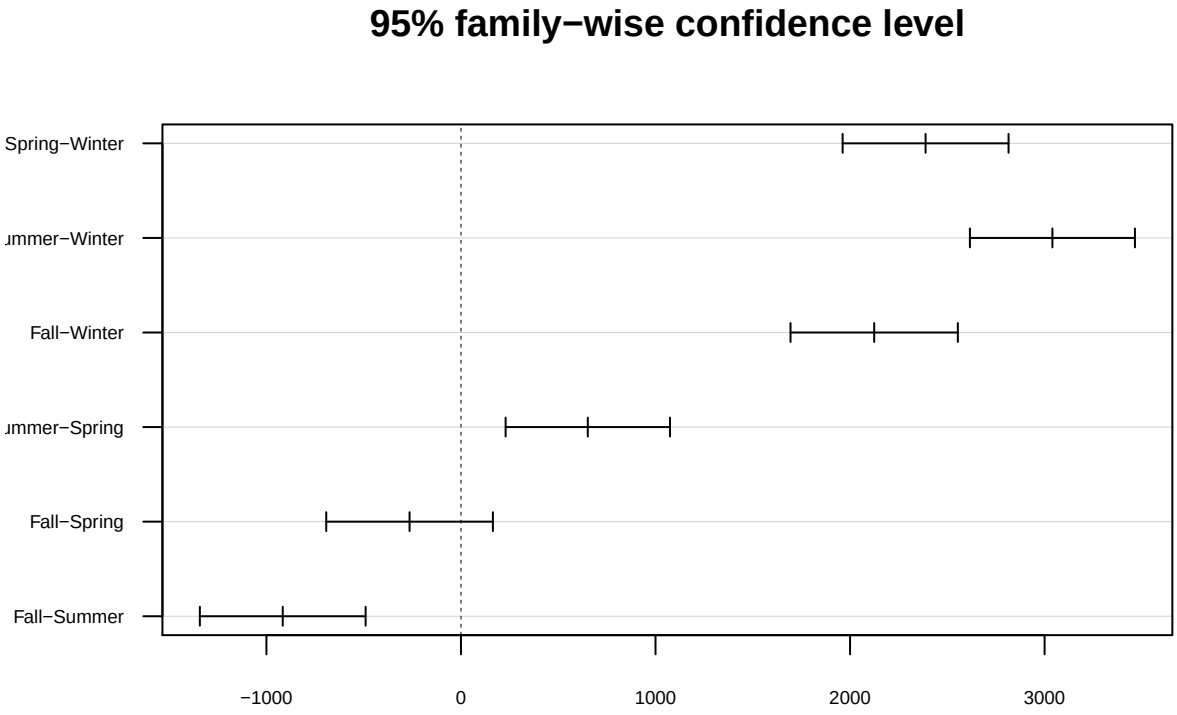
Modeling

```
## Call:
##    aov(formula = TotalUsers ~ Season_FACTOR + day_FACTOR + Season_FACTOR:day_FACTOR,
##     data = bikes)
##
## Terms:
##                 Season_FACTOR day_FACTOR Season_FACTOR:day_FACTOR
## Sum of Squares      950595868   15208002                  14520331
## Deg. of Freedom             3          6                        18
##                  Residuals
## Sum of Squares  1759211191
## Deg. of Freedom        703
##
## Residual standard error: 1581.908
## Estimated effects may be unbalanced
```

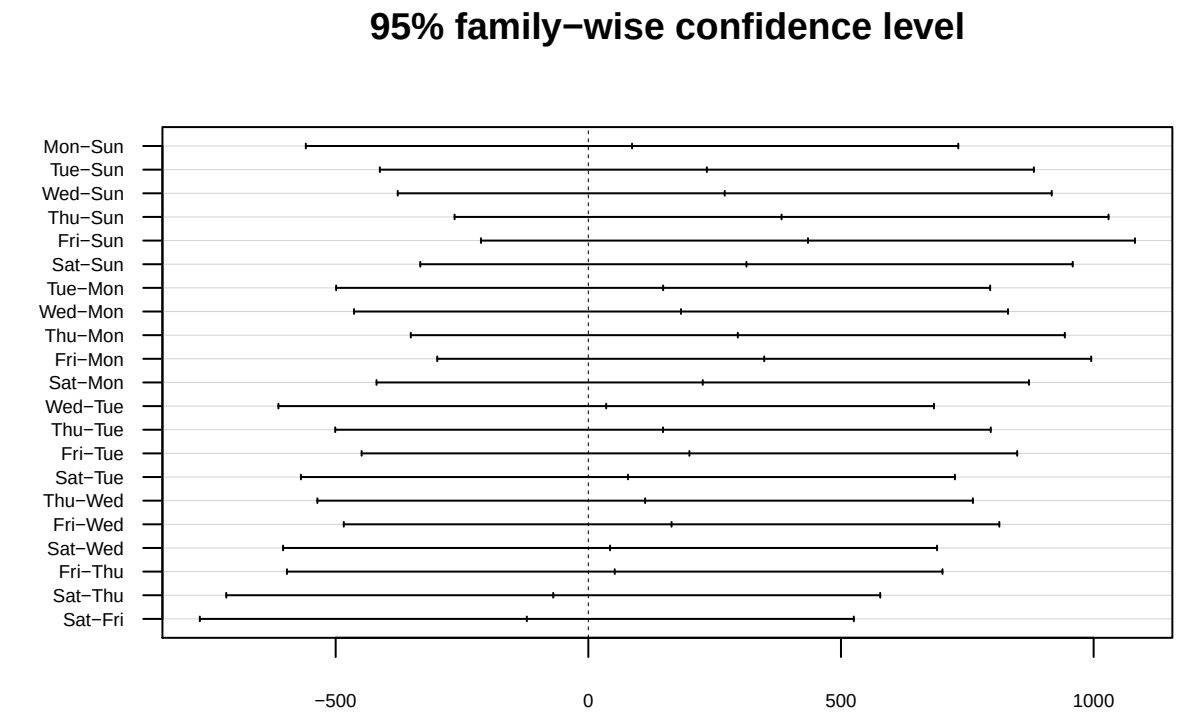
```
##                            Df    Sum Sq   Mean Sq F value Pr(>F)
## Season_FACTOR               3 9.506e+08 316865289 126.623 <2e-16 ***
## day_FACTOR                  6 1.521e+07   2534667   1.013  0.416
```

```
## Season_FACTOR:day_FACTOR  18 1.452e+07    806685   0.322  0.997
## Residuals                703 1.759e+09   2502434
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here is the ANOVA model comparing Season and Day of the Week. Originally, three 2-way ANOVA models were devised. But because there was a significant correlation between Temperature and Season, the ANOVA model included the Season and Day of the Week.

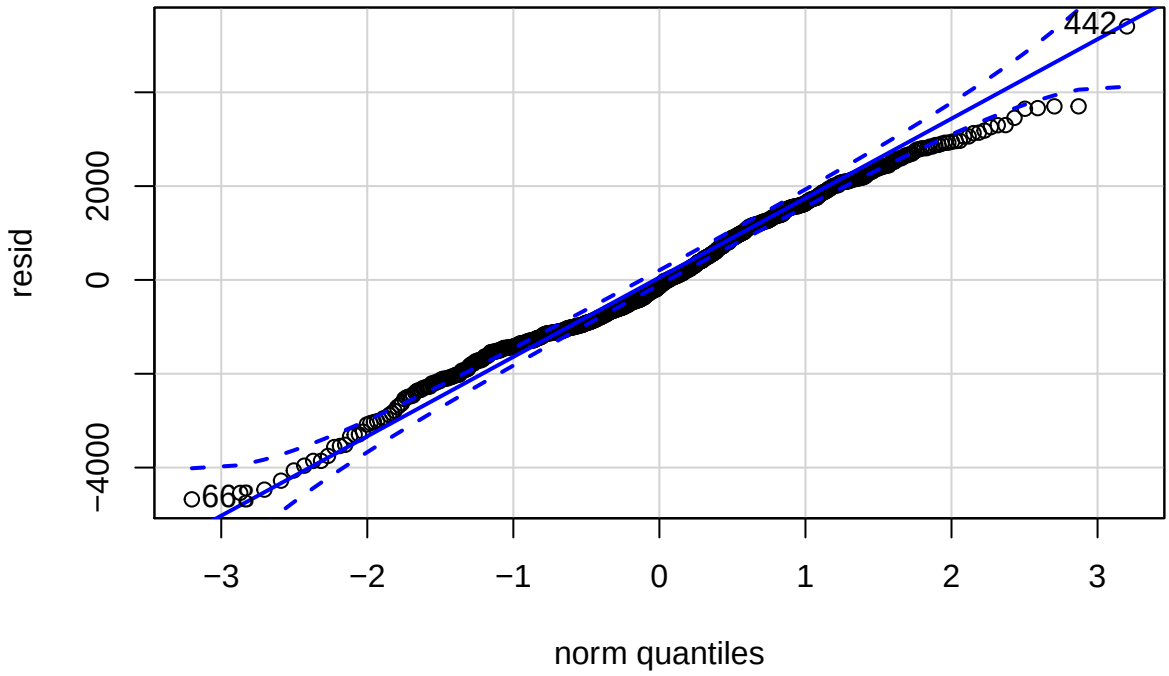


Differences in mean levels of Season_FACTOR

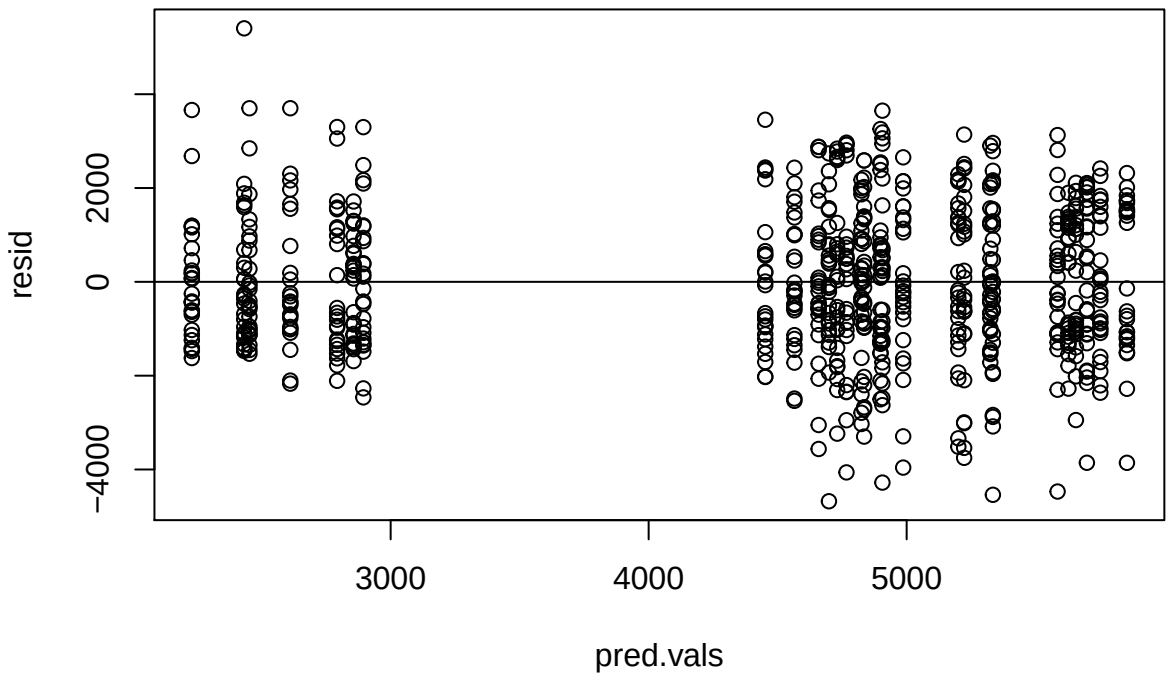


Differences in mean levels of day_FACTOR

Examining the comparison plot, all seasons have a significant difference in the total user number of users except for the relationship between Fall and Spring. While, there is no signficant difference in the total user for any particular day of the week.




```
## [1] 442 668
```




```
##
## Call:
## lm(formula = TotalUsers ~ Season_FACTOR + day_FACTOR, data = bikes)
##
## Coefficients:
##     (Intercept)  Season_FACTOR.L  Season_FACTOR.Q  Season_FACTOR.C
##         4492.73          1570.17         -1649.02            40.14
##    day_FACTOR.L     day_FACTOR.Q     day_FACTOR.C     day_FACTOR^4
##          337.01          -149.10           -74.70           -38.08
##    day_FACTOR^5     day_FACTOR^6
##          -37.15            34.30
```

```
##
## Call:
## lm(formula = TotalUsers ~ Season_FACTOR + day_FACTOR, data = bikes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4593.5 -1067.5  -140.5  1188.1  5163.0
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4492.73      58.02  77.428   <2e-16 ***
## Season_FACTOR.L  1570.17     116.88  13.434   <2e-16 ***
## Season_FACTOR.Q -1649.02     116.06 -14.208   <2e-16 ***
## Season_FACTOR.C    40.14     115.23   0.348   0.7277
## day_FACTOR.L      337.01     153.24   2.199   0.0282 *
```

```
## day_FACTOR.Q     -149.10     153.39  -0.972   0.3314
## day_FACTOR.C      -74.70     153.44  -0.487   0.6265
## day_FACTOR^4      -38.08     153.48  -0.248   0.8041
## day_FACTOR^5      -37.15     153.65  -0.242   0.8090
## day_FACTOR^6       34.30     153.78   0.223   0.8236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1568 on 721 degrees of freedom
## Multiple R-squared:  0.3525, Adjusted R-squared:  0.3445
## F-statistic: 43.62 on 9 and 721 DF,  p-value: < 2.2e-16
```

From the residual plots, the model assumptions of independence and mean zero assumptions are met. From the difference in variances observed in the EDA, I imagined homocedasticity could be an issue, but it is satisfied. At last, normality of the residuals is reasonably satisfied, with a slightly heavier left tail and right tail, the overall of the tails are not much heavier than a normal distribution tail. Since we have a sample of size 731, the central limit theorem will take care of the small deviances from normality, since t-tests are fairly robust to mildheavy tails.

Final Model:

Total Users = 4492.73 + 1570.17(Spring) -1649.02(Summer) + 40.14(Fall) + 337.01(Monday) - 149.10(Tuesday) - 74.70(Wednesday) - 38.08(Thursday) - 37.15(Friday) + 34.30(Saturday)

$\widehat{\beta_1}$ = 4492.73. The total users on Monday, knowing that the season is Winter, would be on average 4492.73

$\widehat{\beta_2}$ is dependent on the Season, with the base case being Winter.

$\widehat{\beta_3}$ is dependent on the day of the week, with the base case being Sunday.

35.25% of variation of the total users can be explained by the regression model above.

Prediction

Model: Number of Bike Users = 4492.74 + 40.14(Fall) + 337.01(Monday)

The condition for our prediction is a season of Fall, TempFeel of 25, and its on a Monday. Based on those conditions, the total user would on average be 4869.89.

Discussion

I would expect the days of the week to not affect the number of people to not affect the total users on a specific day. I only thought that either the temperature or the season may change the number of people that would use the bike systems. And that seems to be true, but I didn't include the values of the temperature as it that would usually correspond with the temperature of a given day, and there was a strong positive correlation between temperature and the season.

However, the data that was included needs to determine the weather of that day. As, more people would ride a bike if it was sunny, and less people would ride bikes if it was raining, and even less if it was snowing. Instead of casing based on the season and the temperature, it would be best to also include the weather, as the weather can depict the number of people would use a bike on nicer or harsher conditions on a given day. Note that the weather and season or temperature may not have a correlatoion as we have witnessed in Pittsburgh that it rains throughout all 4 seasons and can potentially rain or snow throughout at different ranges of temperature. This would help us be more specific on what days people would use the bike system.