Bicycle Ride Sharing

Donald Butler

12/6/2021

DATA 606 Final Project

Part 1 - Introduction

Abstract

I wanted to examine the impact of weather on an individual's desire to ride a bicycle. I utilized a dataset containing two years (2011-12) of bicycle ride sharing data from Washington DC to model the impact of associated weather attributes as predictors for the daily number of rides completed. Each observation in the dataset was for a single day and contained the number of rides completed along with historical weather data (temperature, humidity, wind speed, precipitation). A surprising result was that the day of the week, or rather weekday vs weekend, was not statistically significant. Using linear regression I identified that all of the weather attributes in the dataset were statistically significant and along with the season, were able to account for 55% of the variability in the dataset.

Introduction

I serve on the board for my local cycling club and for the past 6 years I have been collecting the sign-in sheets from our club rides. We hold about 200 rides per year. We regularly discuss falling membership and ride attendance, and often speculate that some rides are less popular for various reasons, but without any concrete reasoning behind it. I have thought about analyzing the sign-in sheet data that I have, but it's not currently in a format that would be conducive to reading digitally.

Part 2 - Data

I obtained a dataset from https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset, which was originally published in 2013. (Fanaee-T and Gama 2013) The researchers collected data from http://capitalbikeshare.com/system-data, which operates a bike share service in Washington DC and correlated it with historical weather data. The dataset contains the number of rides that were completed each day for two years along with temperature, humidity, wind speed, and weather situation (precipitation).

I modified the original dataset by adding labels to a few of the categorical variables and converted the temperature from a normalized celcius scale into Fahrenheit. Here is a snapshot of the data.

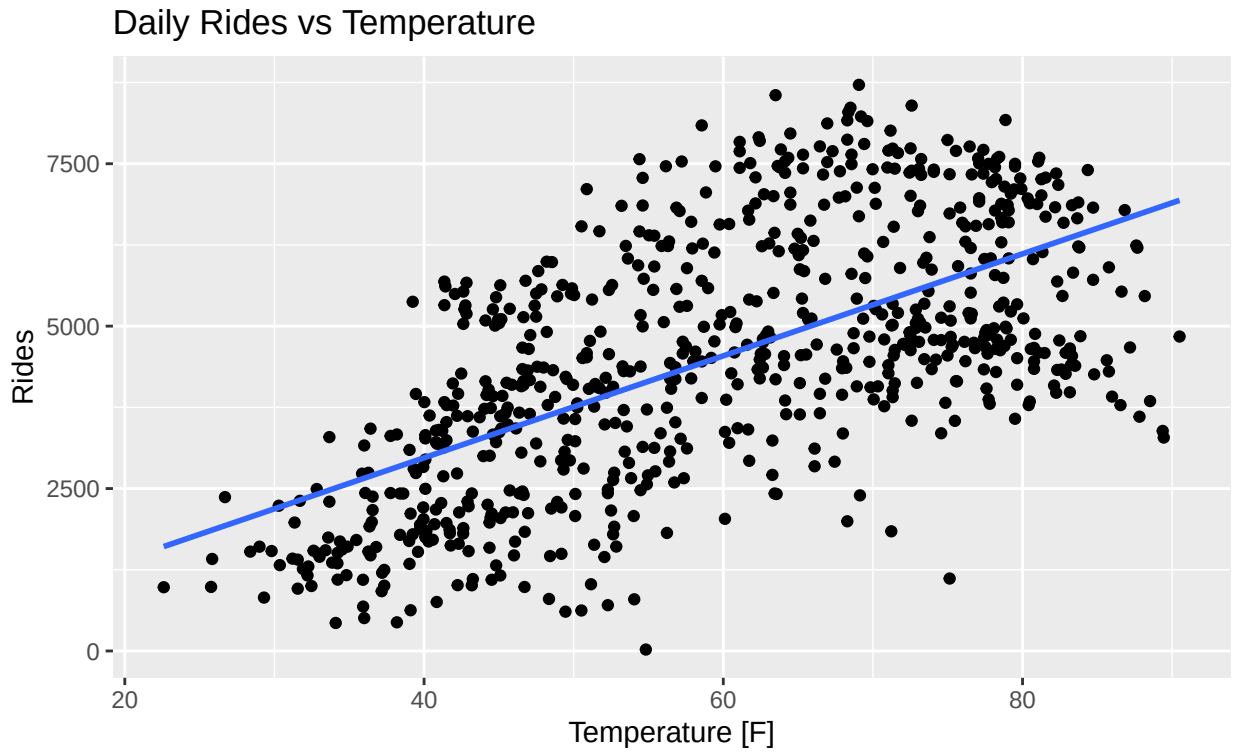
```
## Rows: 731
## Columns: 9
## $ dteday       <date> 2011-01-01, 2011-01-02, 2011-01-03, 2011-01-04, 2011-01-~
## $ season_name  <chr> "Winter", "Winter", "Winter", "Winter", "Winter", "Winter~
## $ dayofweek    <chr> "Sat", "Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "~
## $ workday      <chr> "Weekend", "Weekend", "Weekday", "Weekday", "Weekday", "W~
## $ weather_type <chr> "Mist", "Mist", "Clear", "Clear", "Clear", "Clear", "Mist~
## $ tempF        <dbl> 46.71653, 48.35024, 34.21239, 34.52000, 36.80056, 34.8878~
## $ hum          <dbl> 0.805833, 0.696087, 0.437273, 0.590435, 0.436957, 0.51826~
## $ windspeed    <dbl> 0.1604460, 0.2485390, 0.2483090, 0.1602960, 0.1869000, 0.~
## $ cnt          <dbl> 985, 801, 1349, 1562, 1600, 1606, 1510, 959, 822, 1321, 1~
```

Part 3 - Exploratory data analysis

Before examining the data, I predict that temperature and percipitation should be strong predictors of the number of daily rides.

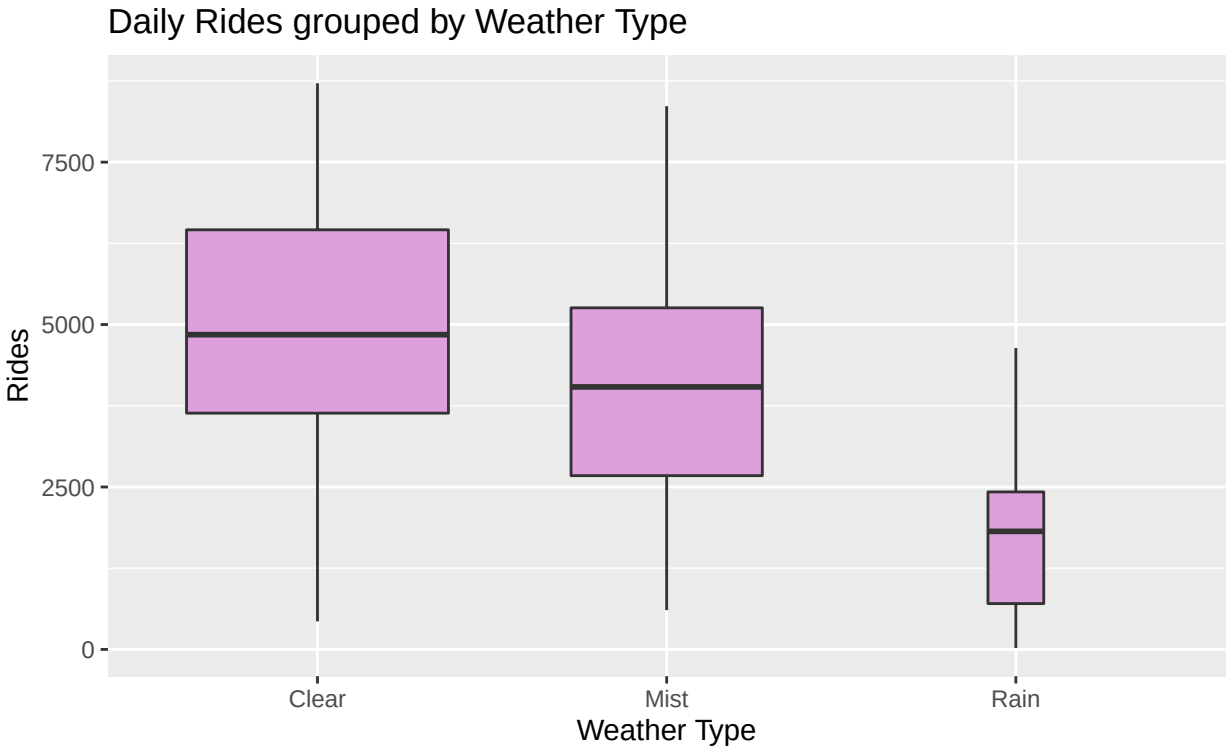
Impact of Temperature

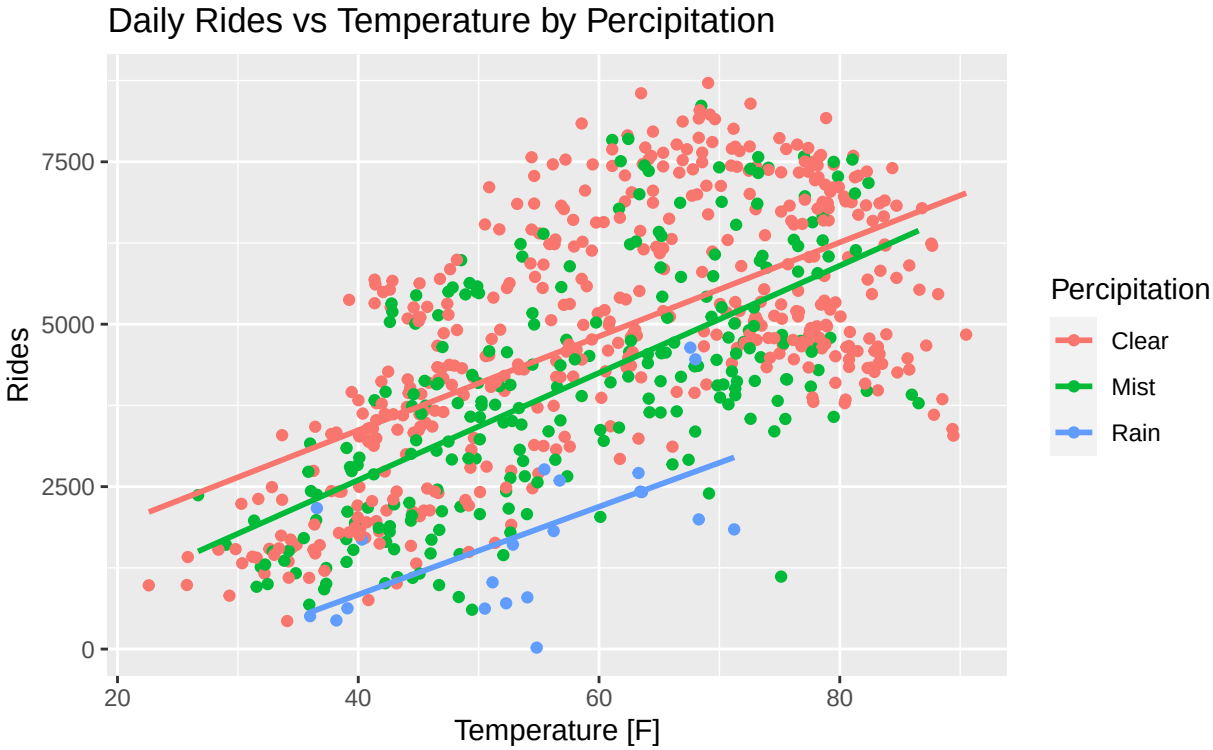
As expected, temperature appears to be correlated to the number of rides completed in the bicycle sharing service.



Impact of Percipitation

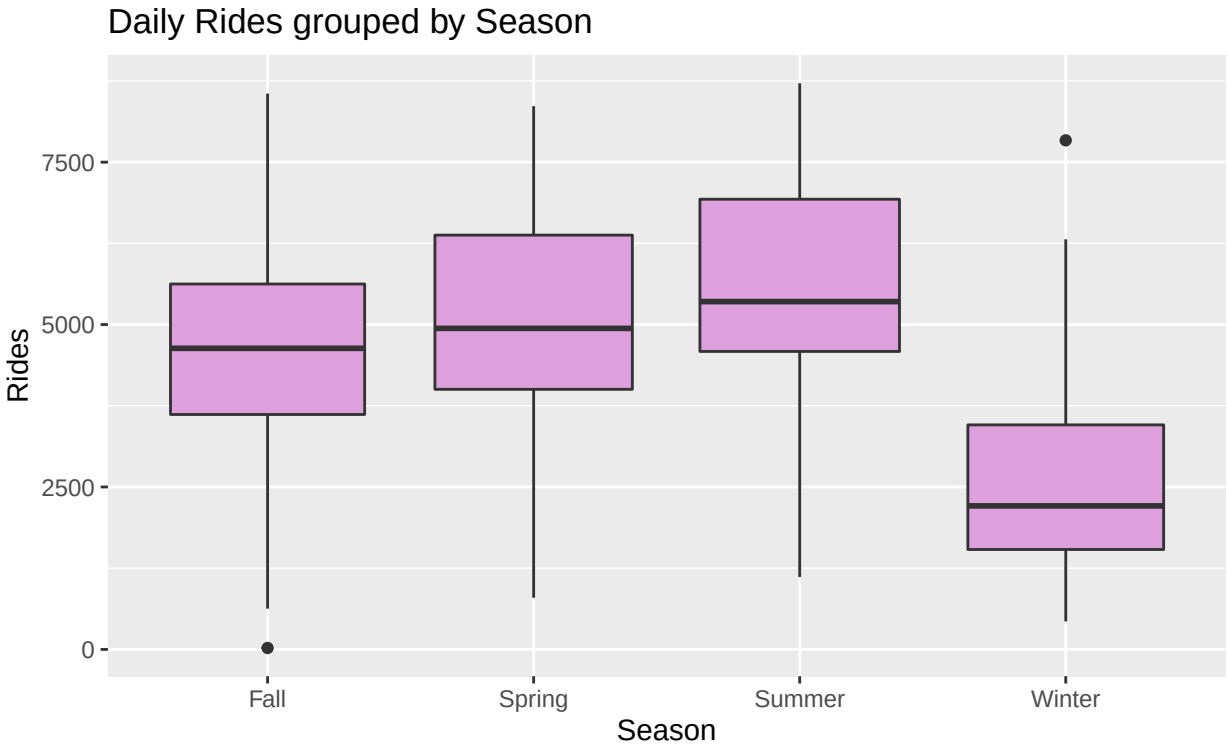
Again, rain has a negative relationship with cycling.

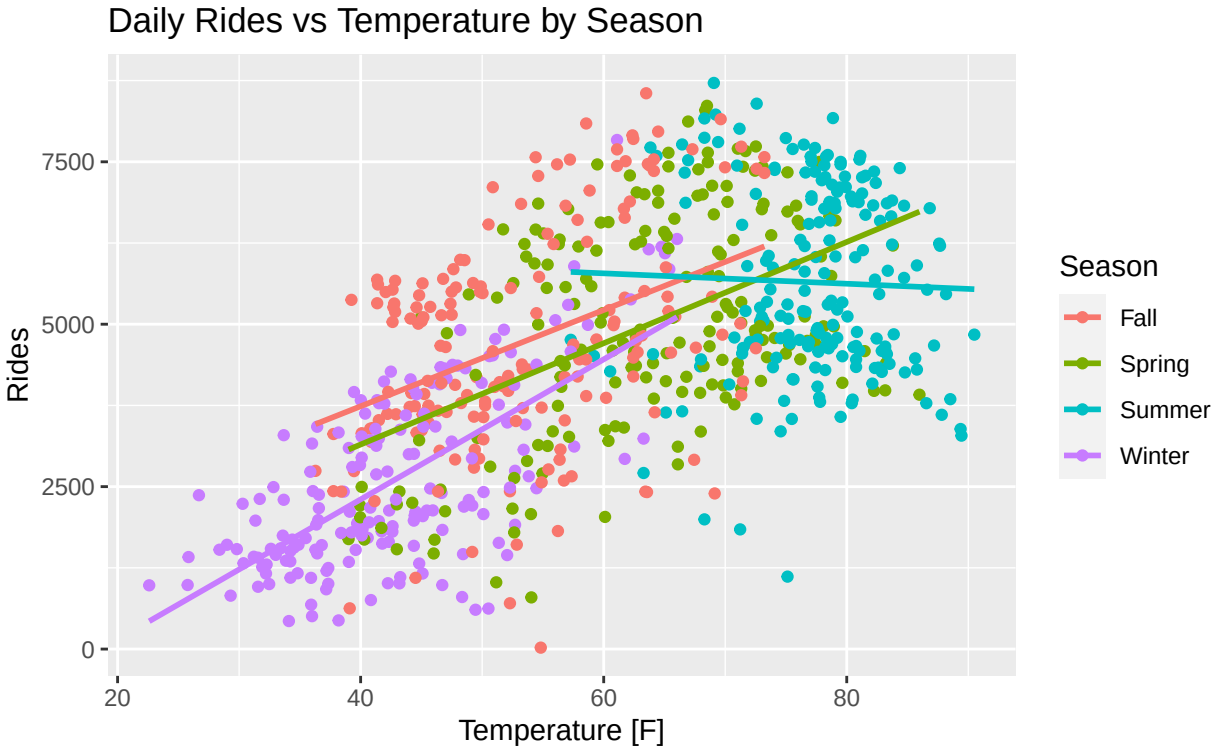




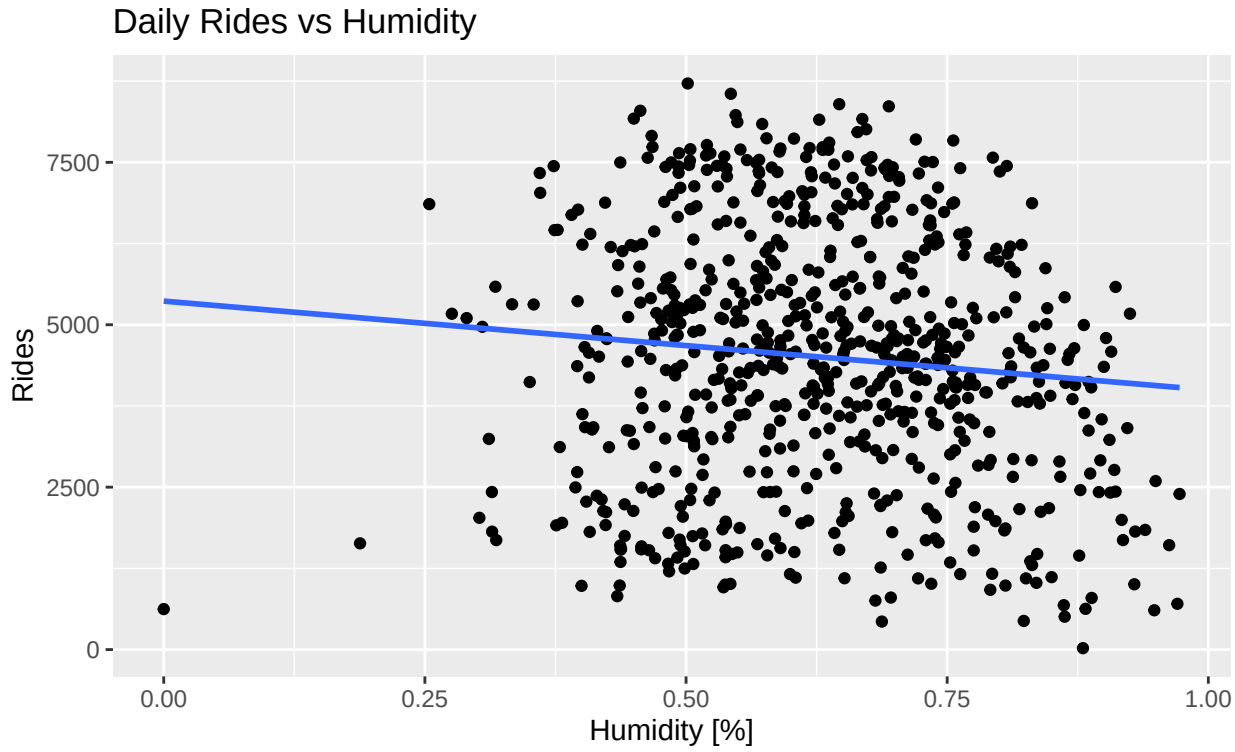
Impact of Season

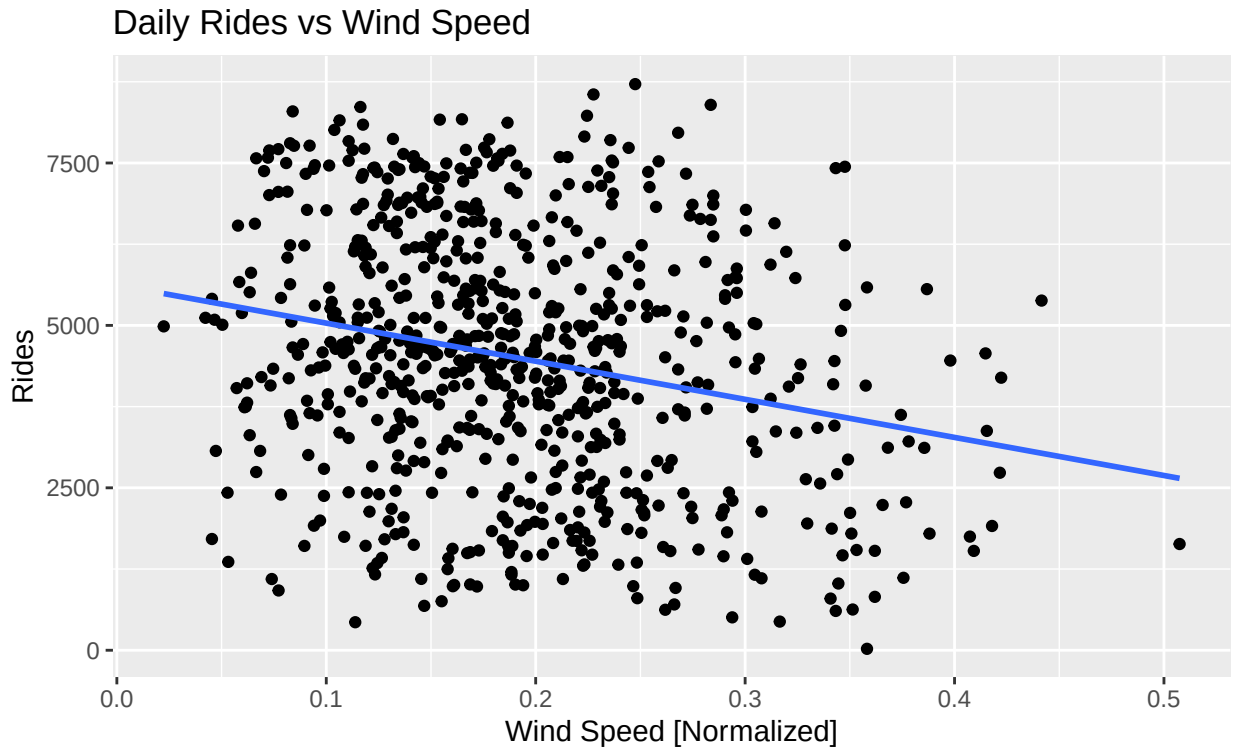
Not surprisingly, there is not as much bicycling in the winter.





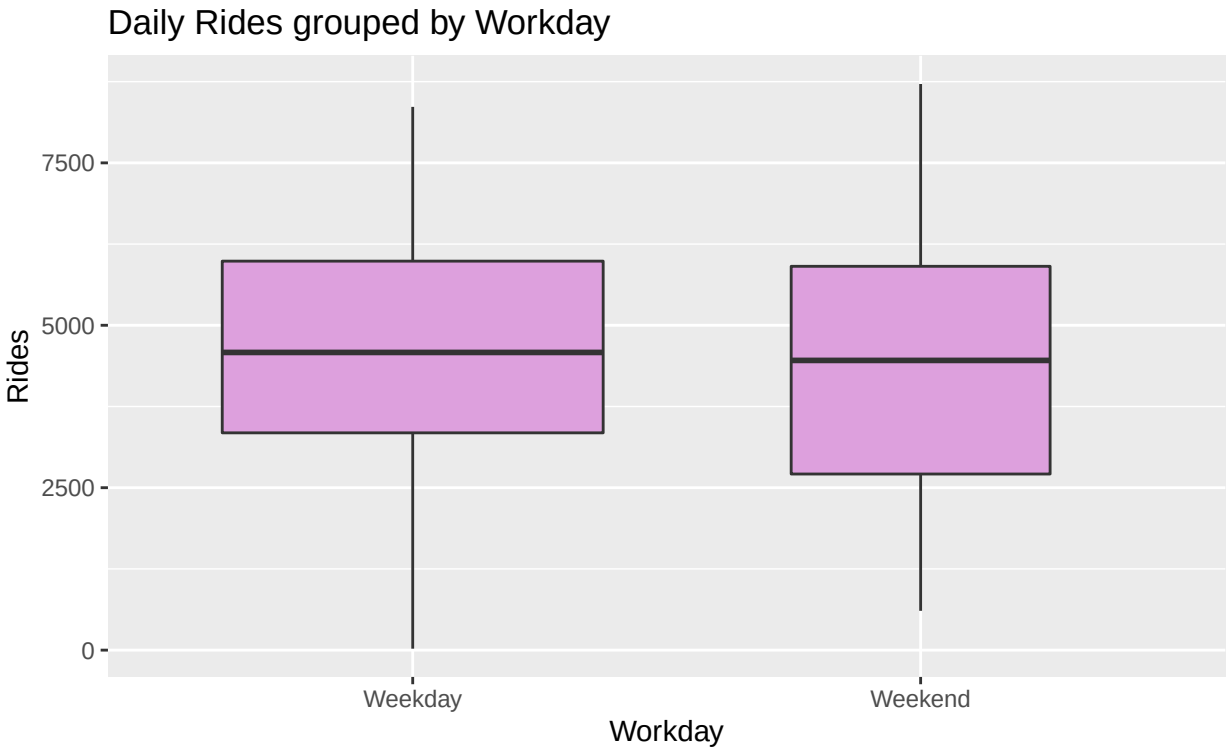
Impact of other weather variables

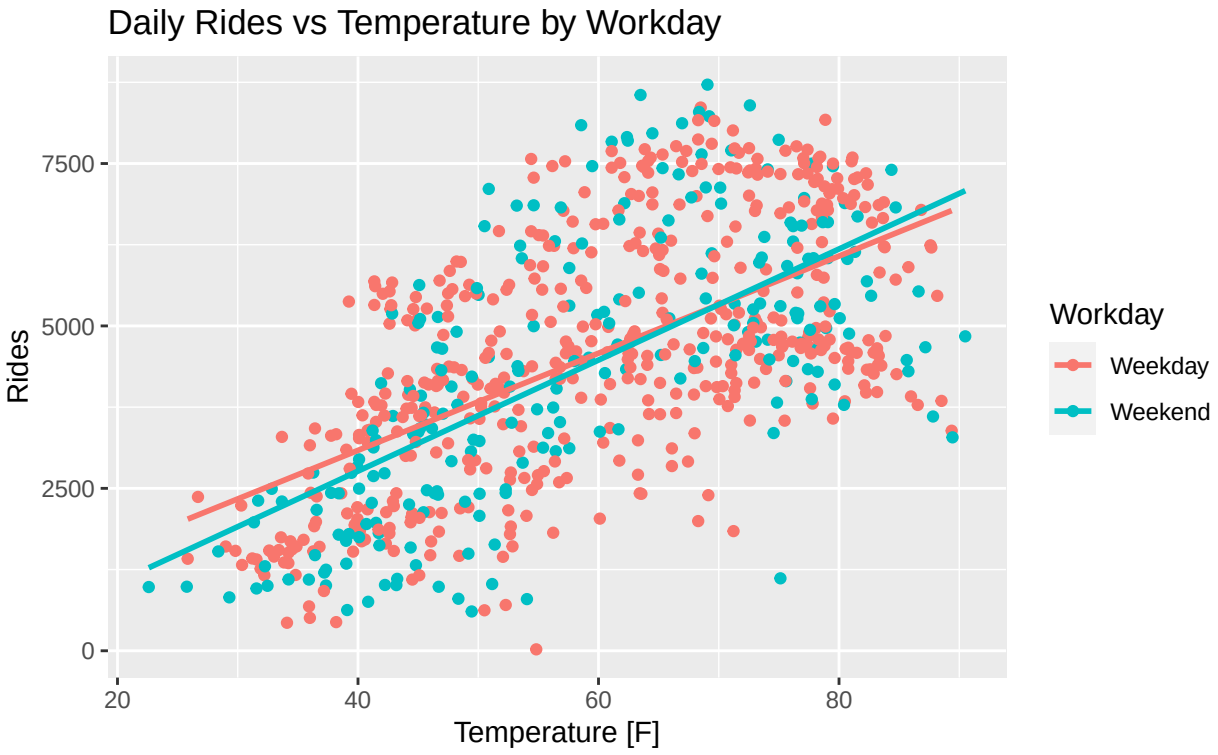




Impact of Day of Week

I was expecting to see some difference between weekday and weekend rides with the high number of commuter users.





Part 4 - Inferences

I will be using linear regression to identify the variables that are statistically significant predictors for the number of rides completed each day.

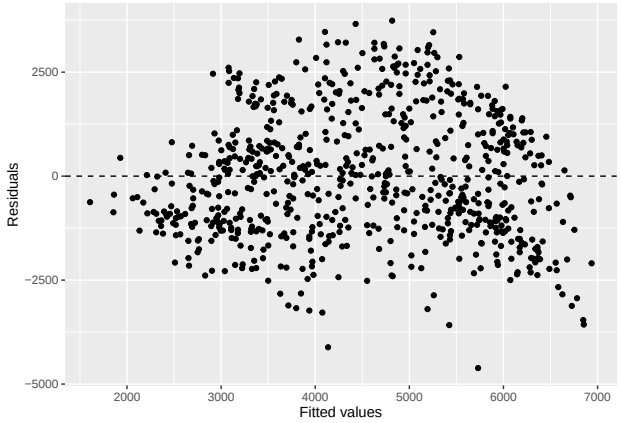
Temperature as a single predictor

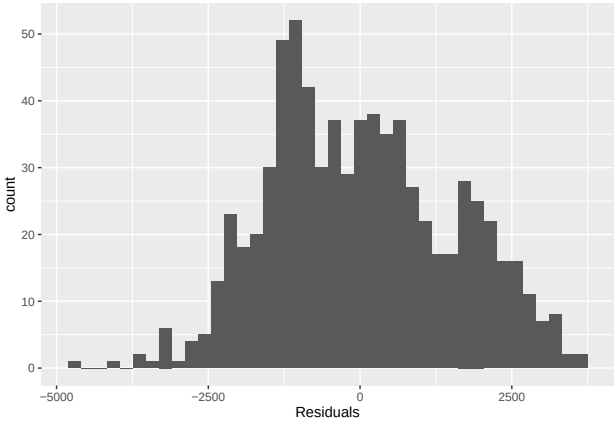
First I will fit a model using just temperature as a predictor. As a cyclist, I expect temperature to be the predominant factor in predicting the number of rides.

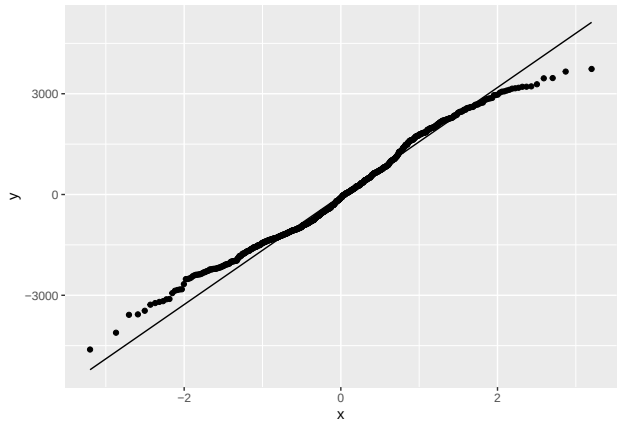
```
##
## Call:
## lm(formula = cnt ~ tempF, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4615.3 -1134.9  -104.4  1044.3  3737.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -166.877    221.816  -0.752    0.452
## tempF         78.495      3.607  21.759   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1509 on 729 degrees of freedom
## Multiple R-squared:  0.3937, Adjusted R-squared:  0.3929
## F-statistic: 473.5 on 1 and 729 DF,  p-value: < 2.2e-16
```

$$\hat{y} = -166.877 + 78.498 \times tempF$$

Temperature is accounting for 39% of the variability in the number of rides per day.







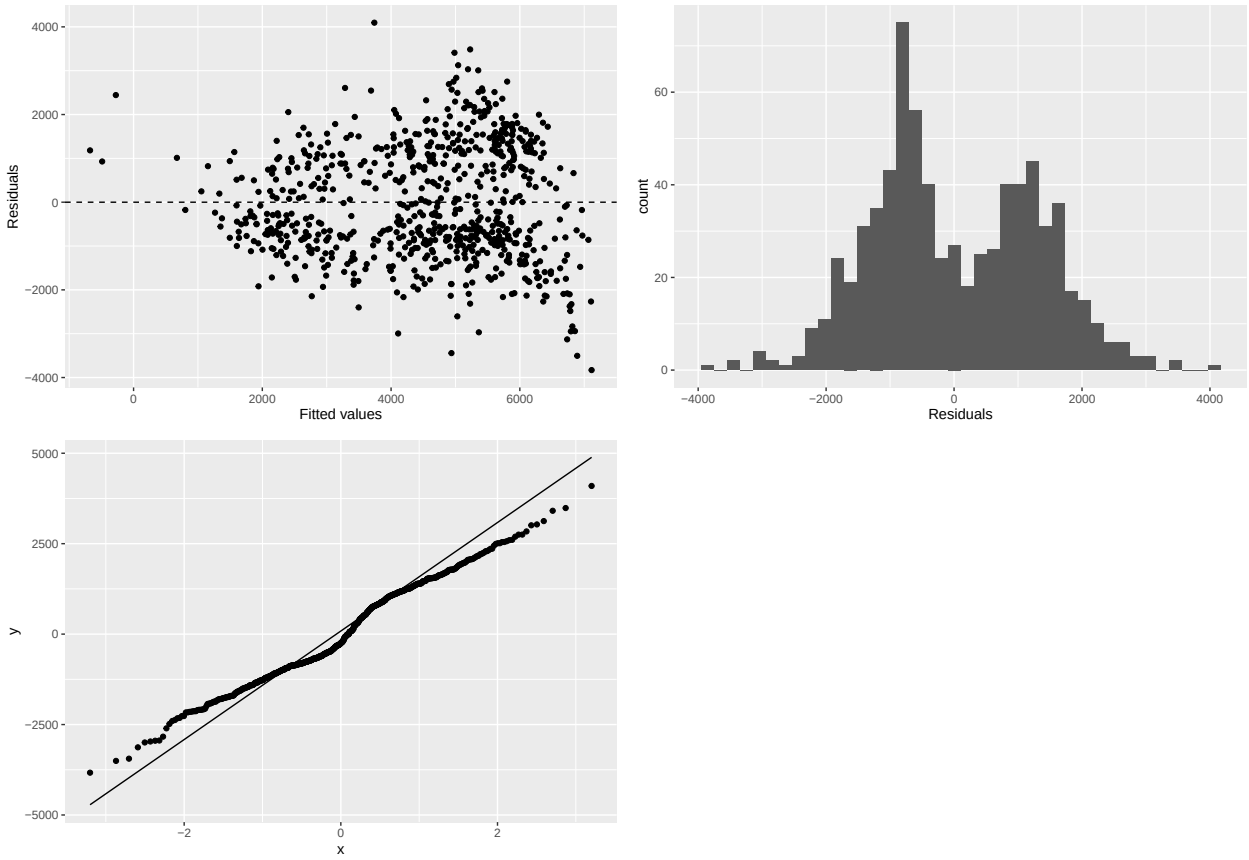
Conditions for the least squares line are met.
* Linearity:
* Nearly normal residuals:
* Constant variability:
* Independent observations:

Multiple Linear Regression

I then built a multiple linear regression model with all of the available variables, and removed the lowest p-value predictors.

```
##
## Call:
## lm(formula = cnt ~ tempF + season_name + weather_type + hum +
##     windspeed, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3830.0  -926.4  -253.8  1097.0  4095.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3334.75     425.17   7.843 1.58e-14 ***
## tempF                73.35       5.68  12.913  < 2e-16 ***
## season_nameSpring  -569.81     151.45  -3.762 0.000182 ***
## season_nameSummer -1025.46     192.41  -5.330 1.32e-07 ***
## season_nameWinter -1496.78     152.51  -9.814  < 2e-16 ***
## weather_typeMist   -218.77     127.54  -1.715 0.086716 .
## weather_typeRain  -1902.85     326.57  -5.827 8.51e-09 ***
## hum               -2635.17     461.15  -5.714 1.61e-08 ***
## windspeed         -3330.79     674.87  -4.935 9.94e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1297 on 722 degrees of freedom
## Multiple R-squared:  0.5564, Adjusted R-squared:  0.5514
## F-statistic: 113.2 on 8 and 722 DF,  p-value: < 2.2e-16
```

Along with temperature, the variables of season, weather_type (precipitation), humidity, and wind speed, are statistically significant in predicting the daily number of rides and have accounted for 55% of the variability in the number of rides per day.



It is hard to determine if the conditions for the model are met with the residuals not appearing to be nearly normal. Both the histogram and the QQ plot show some deviations from normal that may invalidate the model.

Part 5 - Conclusion

My results show how the weather and the season can affect the number of bicycle rides are completed each day. I found it surprising that day of the week, or rather, weekday vs weekend, was not statistically significant in this dataset. I suspect this is due to the high number of commuters using the bike sharing service.

I would like to expand this work with my cycling club's data. I may be able to add additional predictors for the ride (distance, hilliness, etc) and also include member attributes (age, sex, employment status, etc).

References

Fanaee-T, Hadi, and Joao Gama. 2013. "Event Labeling Combining Ensemble Detectors and Background Knowledge." *Progress in Artificial Intelligence*, 1–15. https://doi.org/10.1007/s13748-013-0040-3.