Evaluating Bike Share Ridership

MGT286A Final Project

Matthew Barclay Riley Baumgarten Arvind Kamboh

2024-03-11

Contributions

Arvind Kamboh: Formatted powerpoint, wrote data description, summary and research questions for presentation and final report. Created Polynomial Regression model and Scatter Plot models: Humidity vs Bike Riders, Wind Speed vs Bike Riders, and Temperature vs Bike Riders. Wrote descriptions for models mentioned.

Matthew Barclay: Cleaned data, created ridership distribution visualizations and neural network models, and wrote conclusions.

Riley Baumgarten: Simple Linear regression, Trained Multilinear Regression, Akaike Information Criterion (AIC), Lasso Regression, Random Forests, above model descriptions.

About the Data

This dataset contains daily and hourly ridership levels on the Washington, DC Capital Bikeshare with weather information and additional context about the date. The dataset was obtained from the UCI machine learning repository. The daily dataset comprises 731 observations with 13 predictor variables. The dataset is complete, containing no missing values, which ensures the reliability of analyses conducted. The primary predictor value within this dataset is the number of bike riders, offering insights into the trends and patterns of bike sharing over the observed period. With a comprehensive set of variables, ranging from weather conditions to temporal factors, this dataset presents a robust foundation for predictive modeling and exploratory analysis in the realm of bike sharing systems.

Research Question

What factors influence when people choose to borrow bikes for riding?

Data Source

The dataframes are made up of the following columns:

- instant: record index (num)
- dteday: date (chr)
- season: season (1:spring, 2:summer, 3:fall, 4:winter) (num)
- yr: year (0: 2011, 1:2012) (num)
- mnth: month (1 to 12) (num)
- hr: hour (0 to 23) (num)
- holiday: weather day is holiday or not (num)
- weekday: day of the week
- workingday: if day is neither weekend nor holiday is 1, otherwise is 0. (num)

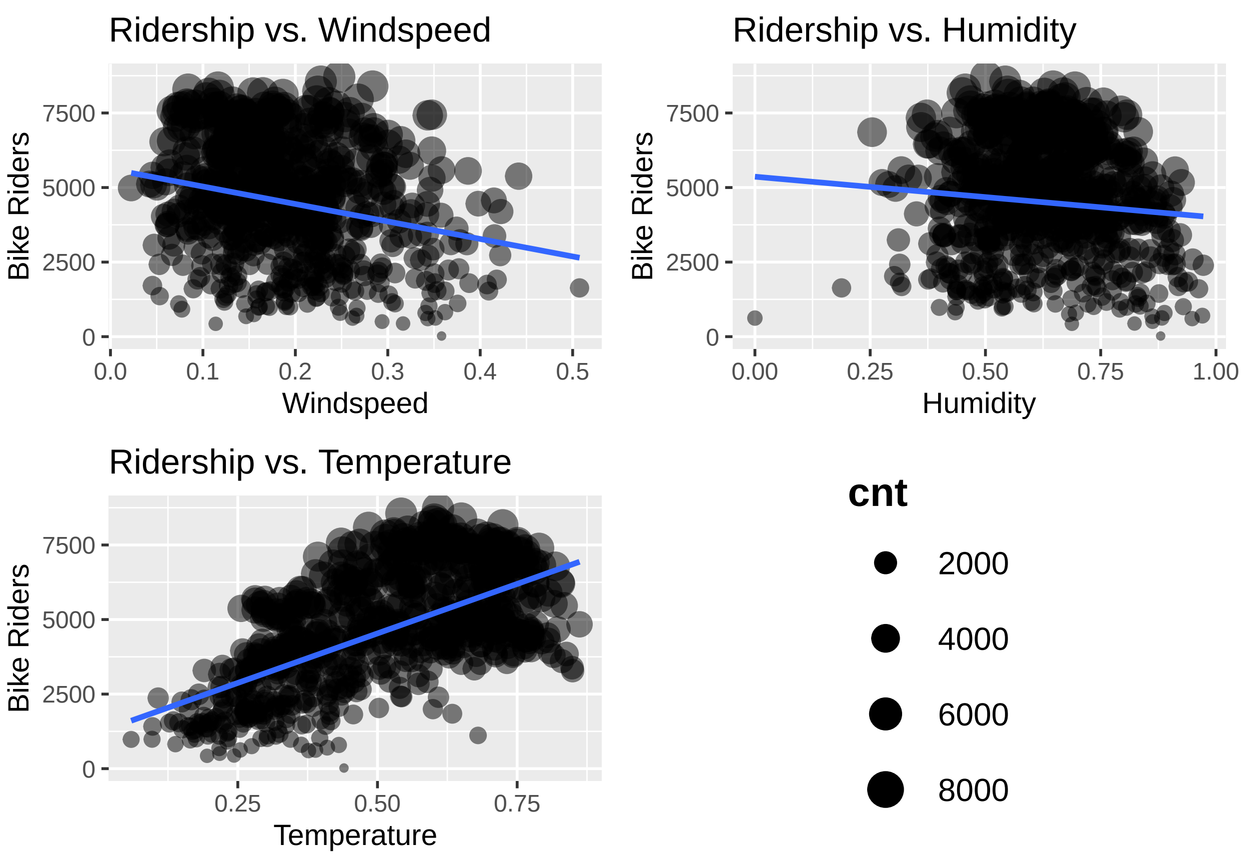
- weathersit: (num)
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp: Normalized temperature in Celsius. The values are divided to 41 (max) (num)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max) (num)
- hum: Normalized humidity. The values are divided to 100 (max) (num)
- windspeed: Normalized wind speed. The values are divided to 67 (max) (num)
- casual: count of casual users (num)
- registered: count of registered users (num)
- cnt: count of total rental bikes including both casual and registered (num)

Data Cleaning

The dataset had no missing values, so the only cleaning needed was converting the date character to date objects and the categorical variables from numeric to factors.

Visualizations

Weather Influence on Ridership

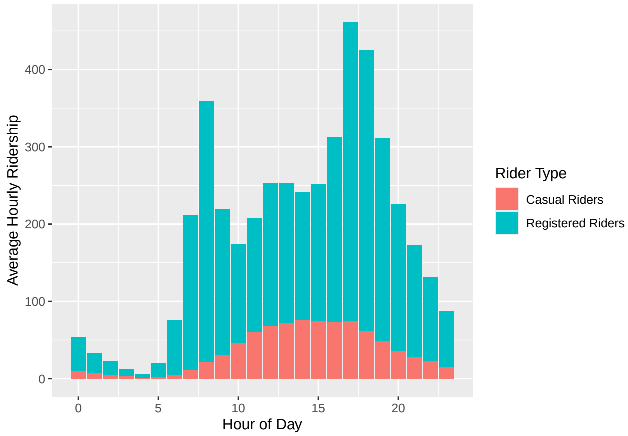


Windspeed: This scatter plot shows the relationship between wind speed and the number of bike riders. It shows a linear relationship as wind speed increases, the number of riders decreases.

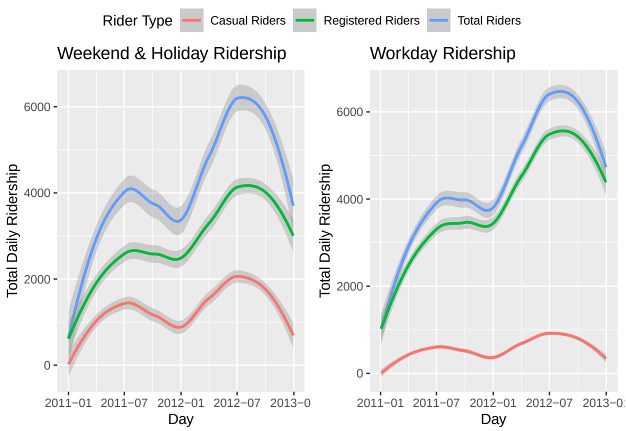
Humidity: This scatter plot shows the relationship between humidity and the number of bike riders. It shows a linear relationship as humidity increases, the number of bike riders decreases.

Temperature: This scatter plot shows the relationship between temperature and the number of bike riders. It shows a linear relationship as temperature increases, the number of bike riders increases.

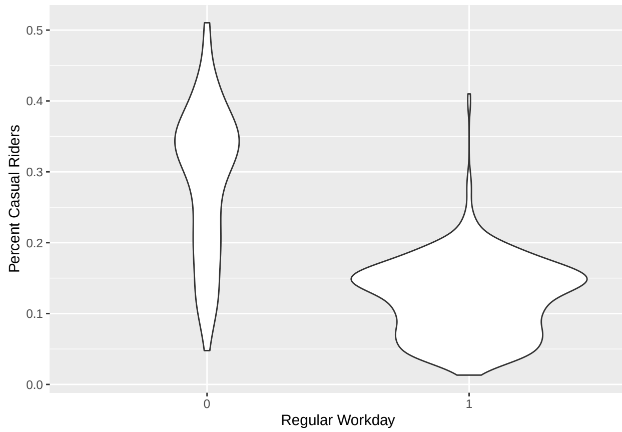
Ridership Distributions



With hourly ridership data available, we were interested in seeing what an average day would look like. There appears to be peaks in registered user ridership during typical commuting times, while casual riders fill in the gaps between commuting hours. While we did not use the hourly data in our modeling, it may be an interesting task for future development.



This plot shows the daily ridership smoothed average over the two years collected by the dataset split by working day status. While there does not seem to be a difference in total ridership, there does seem to be more casual riders on weekends and holidays while working days are almost entirely registered riders.



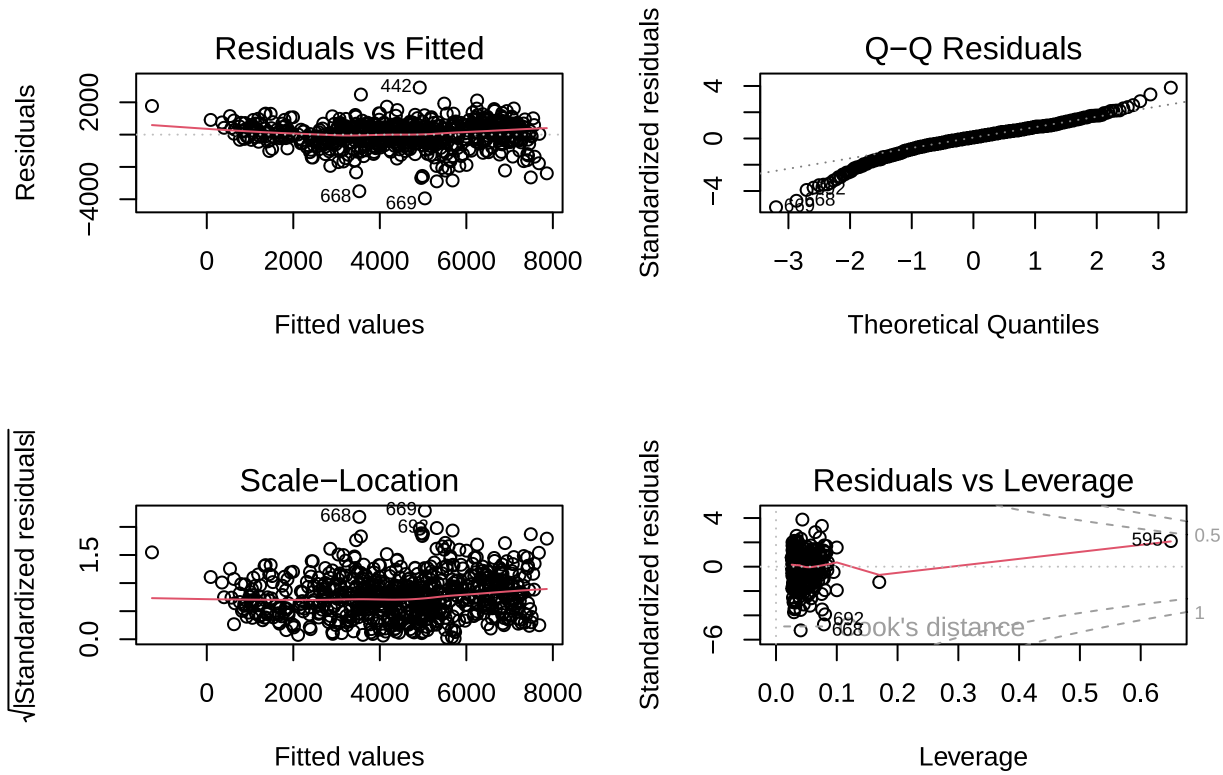
To further explore how working days impact casual or registered ridership, this plot calculates the casual rider percentage and groups by working day. The same pattern emerges where there is very little casual ridership on working days while rivaling registered riders on weekends and holidays.

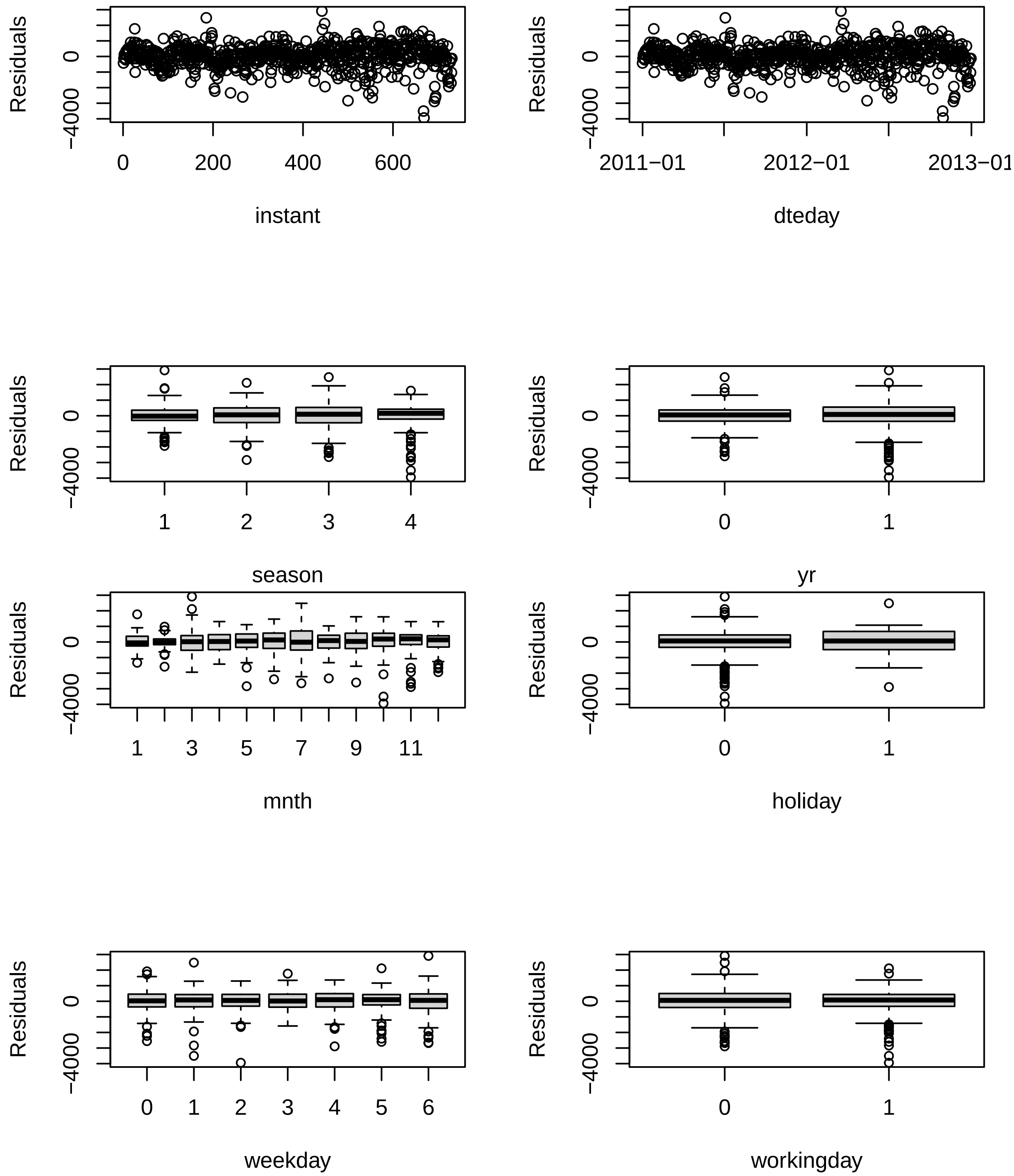
Models

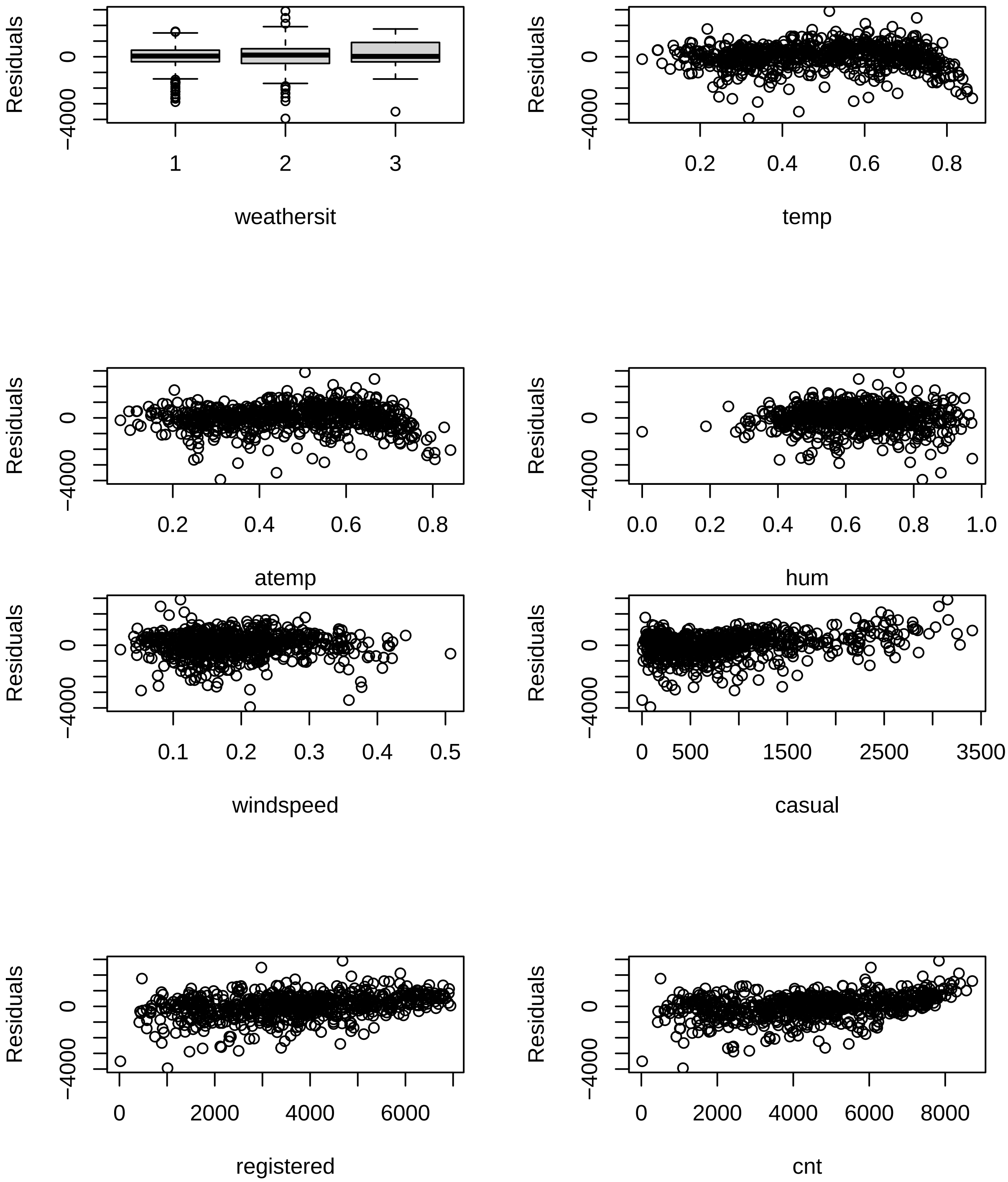
Simple Linear Model

```
##
## Call:
## lm(formula = cnt ~ . - cnt - instant - dteday - casual - registered,
##     data = myData_daily)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3944.7  -348.2    63.8   457.4  2912.7
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1485.84     239.75   6.198 9.77e-10 ***
## season2       884.71     179.49   4.929 1.03e-06 ***
## season3       832.70     213.13   3.907 0.000102 ***
## season4      1575.35     181.00   8.704  < 2e-16 ***
## yr1          2019.74      58.22  34.691  < 2e-16 ***
## mnth2         131.03     143.78   0.911 0.362443
## mnth3         542.83     165.43   3.281 0.001085 **
## mnth4         451.17     247.57   1.822 0.068820 .
## mnth5         735.51     267.63   2.748 0.006145 **
## mnth6         515.40     282.41   1.825 0.068423 .
## mnth7          30.80     313.82   0.098 0.921854
## mnth8         444.95     303.17   1.468 0.142639
## mnth9        1004.17     265.12   3.788 0.000165 ***
## mnth10        519.67     241.55   2.151 0.031787 *
## mnth11       -116.69     230.78  -0.506 0.613257
## mnth12        -89.59     182.21  -0.492 0.623098
## holiday1     -589.70     180.36  -3.270 0.001130 **
## weekday1      212.05     109.49   1.937 0.053187 .
## weekday2      309.53     107.13   2.889 0.003982 **
## weekday3      381.36     107.48   3.548 0.000414 ***
## weekday4      386.34     107.53   3.593 0.000350 ***
```

```
## weekday5       436.98     107.44   4.067 5.30e-05 ***
## weekday6       440.46     106.56   4.133 4.01e-05 ***
## workingday1        NA         NA      NA       NA
## weathersit2  -462.54      77.09  -6.000 3.16e-09 ***
## weathersit3 -1965.09     197.05  -9.972  < 2e-16 ***
## temp         2855.01    1398.16   2.042 0.041526 *
## atemp        1786.16    1462.12   1.222 0.222261
## hum         -1535.47     292.45  -5.250 2.01e-07 ***
## windspeed   -2823.30     414.55  -6.810 2.09e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 769.2 on 702 degrees of freedom
## Multiple R-squared:  0.8484, Adjusted R-squared:  0.8423
## F-statistic: 140.3 on 28 and 702 DF,  p-value: < 2.2e-16
```




Residuals
0
−4000
0
200
400
600
instant
2011−01
2012−01
2013−01
dteday
1
2
3
4
season
0
1
yr
1
3
5
7
9
11
mnth
0
1
holiday
0
1
2
3
4
5
6
weekday
0
1
workingday



Here is just a simple linear regression model along with some reasonably evenly distributed residual plots. We took out cnt, instant, dteday, casual, registered because they are directly correlated. We can see that the data is indeed linear with a high R^2 however we can do better with a trained model.

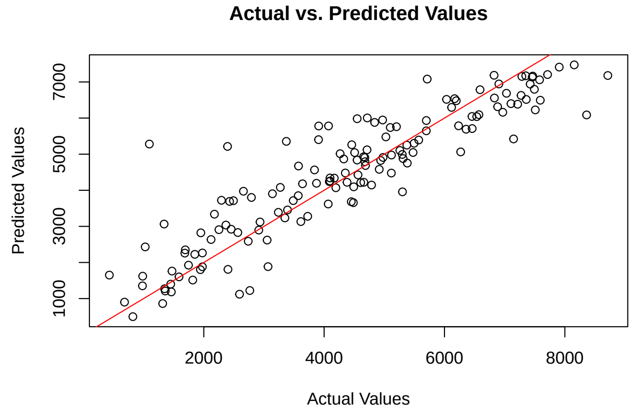
Polynomial Regression

```
##
## Call:
## lm(formula = cnt ~ . - instant - dteday - casual - registered +
##     poly(temp, 3), data = myData_daily)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3256.8  -311.4    53.9   383.4  2449.2
##
## Coefficients: (2 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2209.010    231.432   9.545  < 2e-16 ***
## season2            796.151    156.228   5.096 4.47e-07 ***
## season3           1132.356    186.537   6.070 2.09e-09 ***
## season4           1581.142    157.519  10.038  < 2e-16 ***
## yr1               1977.140     50.956  38.801  < 2e-16 ***
## mnth2               66.368    130.021   0.510  0.60990
## mnth3              266.839    152.035   1.755  0.07968 .
## mnth4               -7.863    221.145  -0.036  0.97165
## mnth5              199.596    235.575   0.847  0.39714
## mnth6              239.547    246.521   0.972  0.33153
## mnth7              274.714    273.705   1.004  0.31588
## mnth8               42.080    265.235   0.159  0.87399
## mnth9              170.392    237.181   0.718  0.47275
## mnth10             -65.911    216.036  -0.305  0.76039
## mnth11            -354.058    207.890  -1.703  0.08899 .
## mnth12            -156.055    164.947  -0.946  0.34443
## holiday1          -437.115    157.204  -2.781  0.00557 **
## weekday1           135.044     95.448   1.415  0.15756
## weekday2           290.148     93.194   3.113  0.00192 **
## weekday3           371.336     93.487   3.972 7.86e-05 ***
## weekday4           380.563     93.536   4.069 5.27e-05 ***
## weekday5           462.426     93.519   4.945 9.55e-07 ***
## weekday6           454.512     92.688   4.904 1.17e-06 ***
## workingday1             NA         NA      NA       NA
## weathersit2       -443.421     67.140  -6.604 7.90e-11 ***
## weathersit3      -1950.430    171.447 -11.376  < 2e-16 ***
## temp              3570.787   1224.726   2.916  0.00366 **
## atemp              869.858   1300.577   0.669  0.50383
## hum              -2031.399    256.508  -7.919 9.38e-15 ***
## windspeed        -3224.771    362.660  -8.892  < 2e-16 ***
## poly(temp, 3)1          NA         NA      NA       NA
## poly(temp, 3)2 -11015.017   1004.616 -10.964  < 2e-16 ***
## poly(temp, 3)3  -8524.255    818.410 -10.416  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 669.1 on 700 degrees of freedom
## Multiple R-squared:  0.8856, Adjusted R-squared:  0.8807
## F-statistic: 180.7 on 30 and 700 DF,  p-value: < 2.2e-16
```

The output displays the results of a polynomial regression model applied to predict bike ridership based on various factors including temperature, season, and weather conditions. Notably, the inclusion of a polynomial transformation of temperature up to the third degree allows for capturing potential non-linear relationships. The significant negative effect of the third degree polynomial term suggests that extreme temperatures may deter individuals from using bikes, providing valuable insights for bike sharing system operators to adjust service provision accordingly.

Training Multilinear Regression




```
##
## Call:
## lm(formula = cnt ~ . - instant - casual - registered, data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3463.5  -359.0    65.5   447.6  2862.0
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1485.47     272.21   5.457 7.31e-08 ***
## season2       822.91     198.17   4.153 3.81e-05 ***
## season3       724.46     232.25   3.119 0.001907 **
## season4      1616.73     193.22   8.367 4.80e-16 ***
## yr1          2032.14      64.15  31.678  < 2e-16 ***
## mnth2          27.84     157.94   0.176 0.860154
## mnth3         536.40     183.25   2.927 0.003560 **
## mnth4         568.70     276.37   2.058 0.040081 *
## mnth5         826.34     299.91   2.755 0.006056 **
## mnth6         803.29     317.11   2.533 0.011579 *
## mnth7         345.25     351.16   0.983 0.325945
## mnth8         772.56     335.71   2.301 0.021746 *
## mnth9        1295.03     295.74   4.379 1.43e-05 ***
## mnth10        532.99     265.42   2.008 0.045115 *
## mnth11       -270.15     251.88  -1.073 0.283931
## mnth12       -267.99     197.27  -1.359 0.174851
## holiday1     -598.33     212.00  -2.822 0.004940 **
## weekday1      184.39     119.75   1.540 0.124158
## weekday2      329.50     116.51   2.828 0.004851 **
## weekday3      465.04     117.61   3.954 8.68e-05 ***
## weekday4      394.69     118.12   3.341 0.000889 ***
## weekday5      438.63     117.46   3.734 0.000208 ***
## weekday6      484.16     114.73   4.220 2.85e-05 ***
## workingday1       NA         NA      NA       NA
## weathersit2  -447.30      83.65  -5.347 1.31e-07 ***
## weathersit3 -2172.86     217.93  -9.970  < 2e-16 ***
## temp          553.24    2519.05   0.220 0.826245
## atemp        3534.70    2709.17   1.305 0.192529
## hum         -1181.15     316.01  -3.738 0.000205 ***
```

```
## windspeed   -2653.50     478.29  -5.548 4.48e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 754 on 555 degrees of freedom
## Multiple R-squared:  0.8526, Adjusted R-squared:  0.8451
## F-statistic: 114.6 on 28 and 555 DF,  p-value: < 2.2e-16
```

For Multilinear regression we partitioned the data to 80 train 20 test so we can get a more reasonable look at the data. Here we see the R^2 decreases only by a little and the coefficients have sharper magnitudes. Looking at the Actual vs. Predicted values scatter plot, we can see our model is quite efficient.

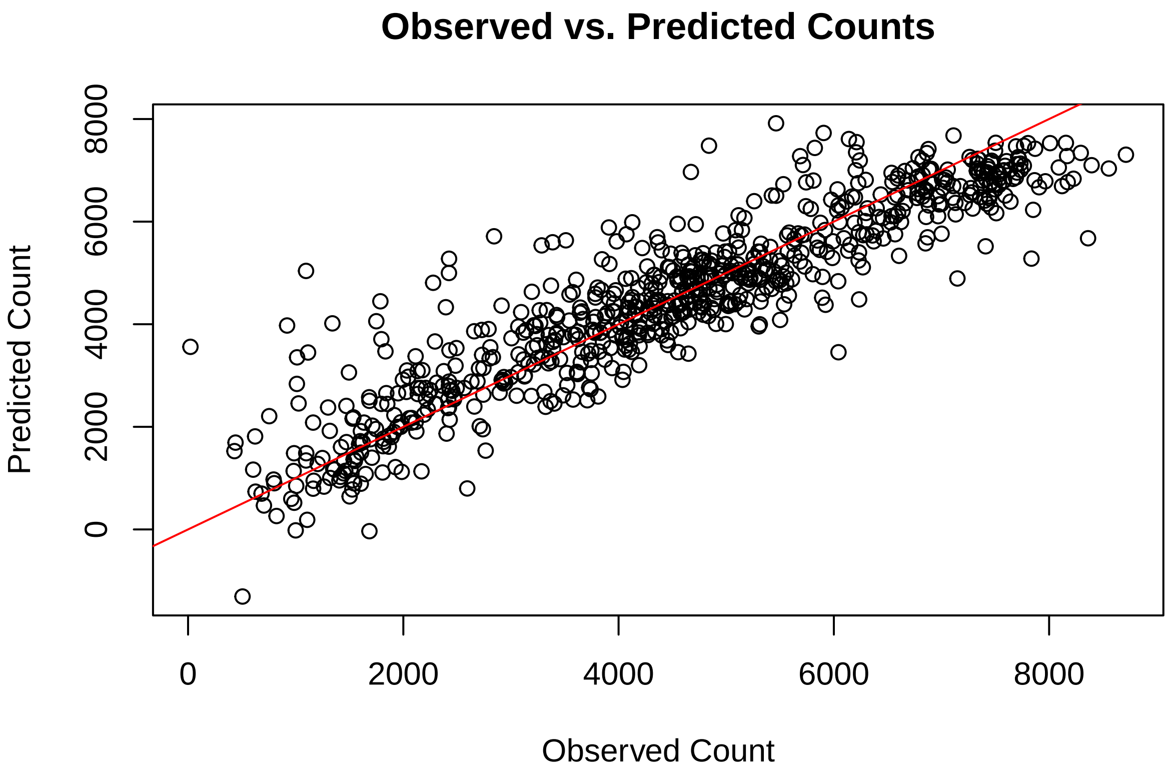
AIC

```
##  [1] "atemp"      "yr"         "season"     "weathersit" "mnth"
##  [6] "weekday"    "hum"        "holiday"    "workingday" "cnt"
## [11] "temp"
```

```
##
## Call:
## lm(formula = cnt ~ atemp + yr + season + weathersit + mnth +
##     weekday + hum + holiday + cnt + temp, data = myData_daily)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4109.2  -364.2    71.6   482.4  3185.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   592.73     207.08   2.862 0.004330 **
## atemp        3794.53    1477.57   2.568 0.010432 *
## yr1          2039.07      60.00  33.985  < 2e-16 ***
## season2       916.69     185.13   4.952 9.23e-07 ***
## season3       889.16     219.73   4.047 5.78e-05 ***
## season4      1687.63     185.98   9.074  < 2e-16 ***
## weathersit2  -533.28      78.81  -6.767 2.78e-11 ***
## weathersit3 -2244.61     198.85 -11.288  < 2e-16 ***
## mnth2         113.83     148.32   0.767 0.443066
## mnth3         507.58     170.61   2.975 0.003029 **
## mnth4         366.59     255.11   1.437 0.151169
## mnth5         766.50     276.09   2.776 0.005645 **
## mnth6         610.56     291.03   2.098 0.036267 *
## mnth7         161.55     323.19   0.500 0.617330
## mnth8         553.25     312.37   1.771 0.076967 .
## mnth9        1059.09     273.42   3.874 0.000117 ***
## mnth10        483.27     249.17   1.940 0.052834 .
## mnth11       -181.32     237.91  -0.762 0.446221
## mnth12       -108.77     187.98  -0.579 0.563037
## weekday1      207.25     112.97   1.835 0.067001 .
## weekday2      310.28     110.54   2.807 0.005140 **
## weekday3      401.11     110.85   3.618 0.000318 ***
## weekday4      401.57     110.92   3.620 0.000315 ***
## weekday5      467.35     110.76   4.220 2.77e-05 ***
## weekday6      437.88     109.95   3.983 7.53e-05 ***
## hum         -1003.36     290.77  -3.451 0.000593 ***
```

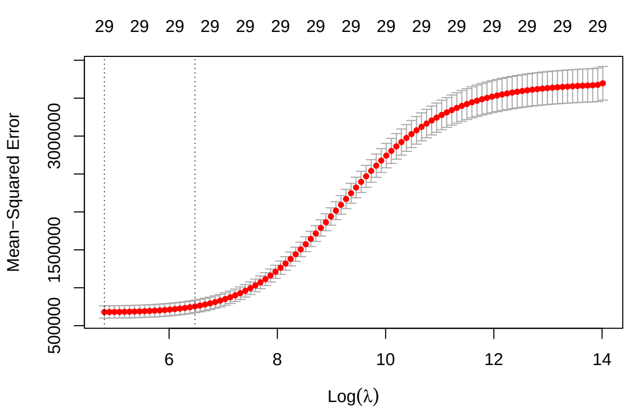
```
## holiday1     -572.70     186.08  -3.078 0.002166 **
## temp          872.93    1410.98   0.619 0.536333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 793.7 on 703 degrees of freedom
## Multiple R-squared:  0.8383, Adjusted R-squared:  0.8321
## F-statistic:   135 on 27 and 703 DF,  p-value: < 2.2e-16

## Start:  AIC=9825.17
## cnt ~ atemp + yr + weathersit + mnth + weekday + windspeed +
##     hum + holiday
##
##              Df Sum of Sq        RSS     AIC
## <none>                     469292413  9825.2
## - holiday     1   8896865  478189278  9836.9
## - weekday     6  15562615  484855028  9837.0
## - hum         1  15286213  484578627  9846.6
## - windspeed   1  32628663  501921077  9872.3
## - weathersit  2  62183230  531475643  9912.1
## - atemp       1  75738298  545030711  9932.5
## - mnth       11 199013401  668305814 10061.6
## - yr          1 720175288 1189467702 10503.0

##
## Call:
## lm(formula = cnt ~ atemp + yr + weathersit + mnth + weekday +
##     windspeed + hum + holiday, data = myData_daily)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3943.2  -377.2    62.5   484.9  2686.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1430.17     251.52   5.686 1.90e-08 ***
## atemp        4820.52     451.60  10.674  < 2e-16 ***
## yr1          2025.85      61.55  32.915  < 2e-16 ***
## weathersit2  -457.35      81.65  -5.601 3.05e-08 ***
## weathersit3 -1952.37     208.59  -9.360  < 2e-16 ***
## mnth2         129.66     152.29   0.851 0.394833
## mnth3         862.65     160.48   5.375 1.04e-07 ***
## mnth4        1346.66     177.79   7.575 1.13e-13 ***
## mnth5        1643.12     203.61   8.070 3.03e-15 ***
## mnth6        1444.63     231.74   6.234 7.83e-10 ***
## mnth7         928.48     253.12   3.668 0.000263 ***
## mnth8        1362.97     233.06   5.848 7.60e-09 ***
## mnth9        2074.14     209.22   9.914  < 2e-16 ***
## mnth10       2086.34     178.67  11.677  < 2e-16 ***
## mnth11       1446.62     158.08   9.151  < 2e-16 ***
## mnth12        903.32     151.93   5.945 4.33e-09 ***
## weekday1      229.83     116.01   1.981 0.047966 *
## weekday2      323.05     113.49   2.847 0.004548 **
## weekday3      393.05     113.73   3.456 0.000581 ***
## weekday4      398.38     113.83   3.500 0.000495 ***
```

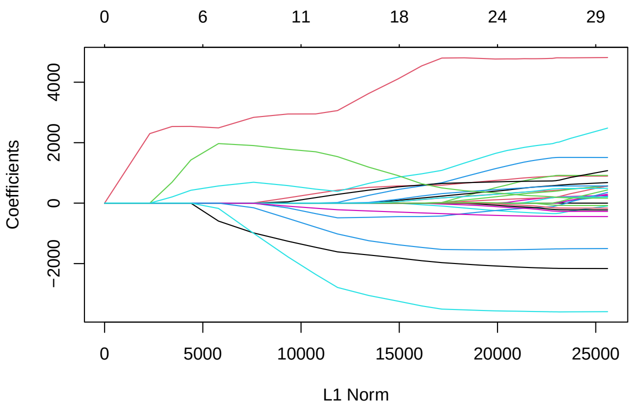
```
## weekday5      444.64     113.52   3.917 9.85e-05 ***
## weekday6      435.77     112.89   3.860 0.000124 ***
## windspeed   -2997.33     427.81  -7.006 5.72e-12 ***
## hum         -1483.22     309.30  -4.795 1.98e-06 ***
## holiday1     -694.61     189.86  -3.658 0.000273 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 815.3 on 706 degrees of freedom
## Multiple R-squared:  0.8287, Adjusted R-squared:  0.8229
## F-statistic: 142.3 on 24 and 706 DF,  p-value: < 2.2e-16
```

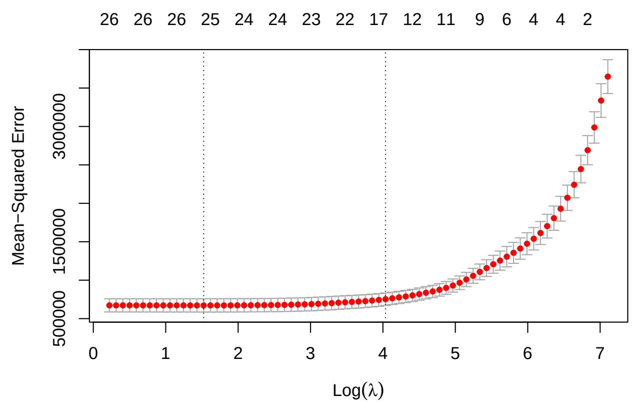


We implemented 2 separate variable selection approaches. One being Akaike Information Criterion (AIC) which selects the variables that minimize AIC value and two, by using stepwise regression. We see that the Rˆ2 is a high value and the residuals plot indicates the model is predicting the data correctly.

Lasso Regression








```
## [1] "Best Lamda:  121.921875136854"
```

```
## [1] "RMSE:  803.593245105102"
```

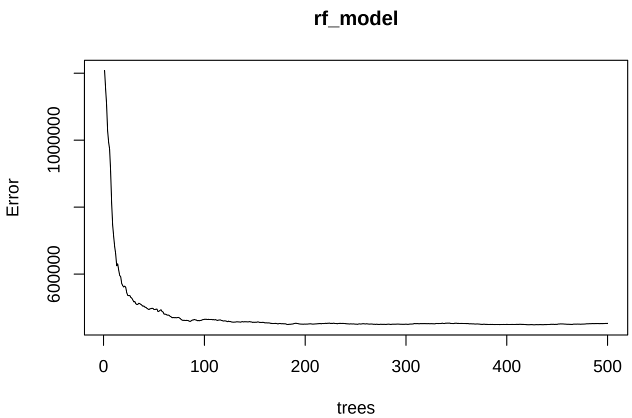
```
##  (Intercept)  (Intercept)       dteday      season2      season3      season4
## -33396.22956      0.00000      2.33446    629.90090    410.04843   1017.67660
##          yr1        mnth2        mnth3        mnth4        mnth5        mnth6
##   1100.42794    -40.59925    362.43775    359.46442    563.46898    354.12461
##        mnth7        mnth8        mnth9
##    -68.09200    258.72979    738.45603
```

For Lasso Regression, we also have ridge regression included (alpha = 0). We can see by plotting with the L1 Norm that only a few variables got pushed to zero, season3 (fall), workingday1, and holiday1 as well as some variables coming close to zero with respect to the data's

Random Forests

```
##
## Call:
##  randomForest(formula = cnt ~ . - instant - casual - registered,      data = trainData, ntree = 500)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 452888.7
##                     % Var explained: 87.64
```

```
## [1] "RMSE: 748.010280191426"
```




```
## [1] "Adjusted R-squared: 0.880675034867504"
```

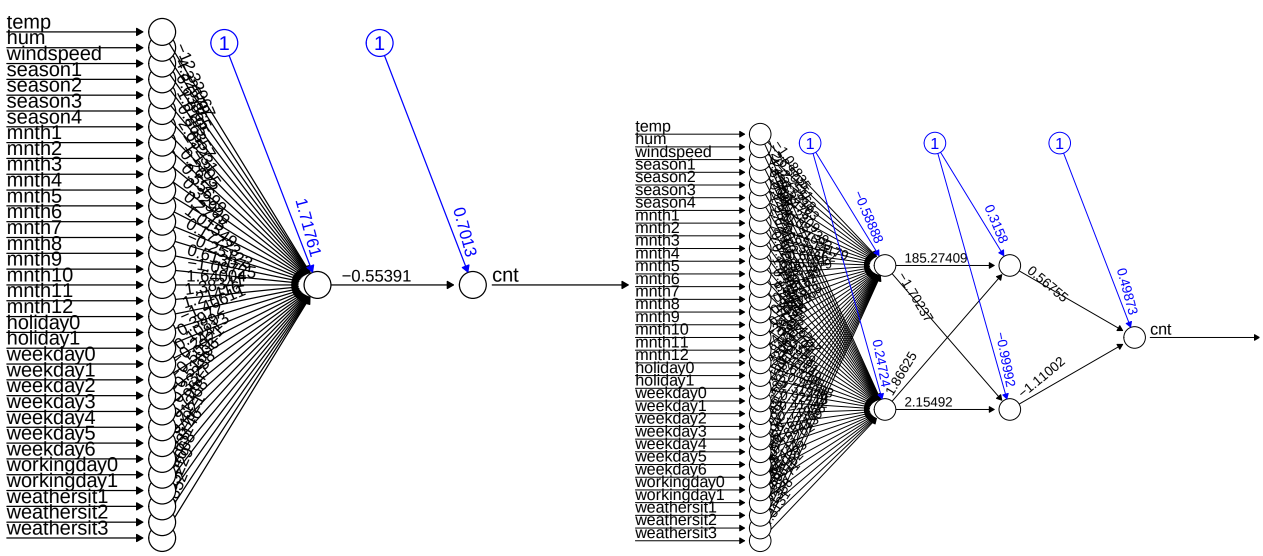
```
## [1] "Optimal number of trees: 350"
```

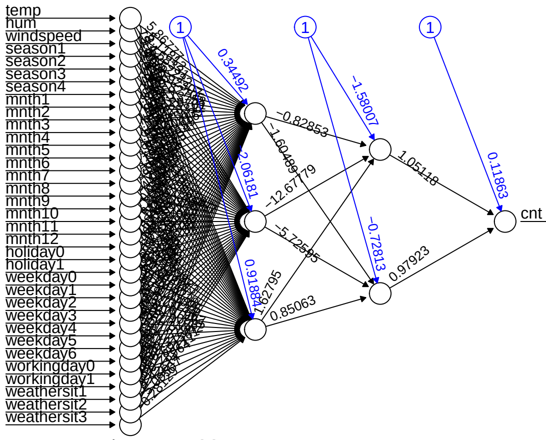
Artificial Neural Networks

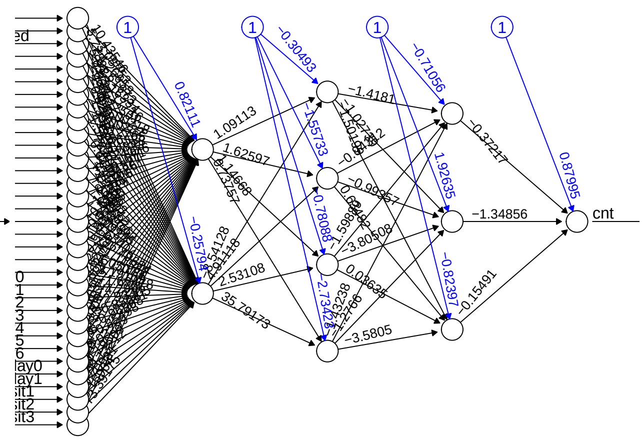
The linear models performed well, but there may be hidden nonlinear relationships that we did not identify, so training an artificial neural network may capture that hidden variation.

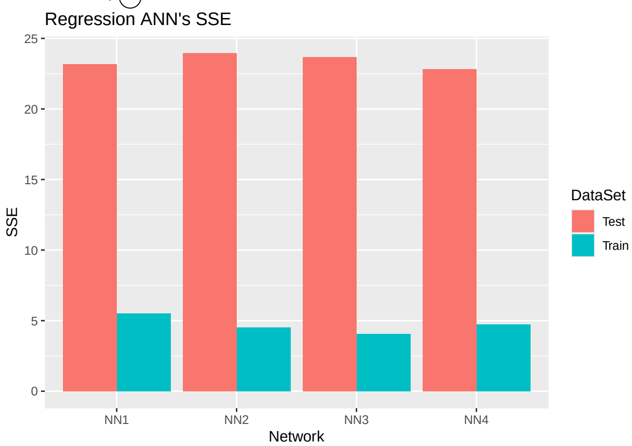
Predicting total ridership

To predict total ridership, the data was randomly split into an 80% train 20% test validation scheme. The predictors season and atemp were removed from the inputs as they greatly increased the error and decreased the model fitness. The neural network models were evaluated by their residual sum of squared error and many variations of the neuron structure were tested. Neural network number 4 with 2, 4, and 3 hidden layers was found to be the optimal model with the overall model having a respectable 0.71 adjusted R-squared value. With additional experimentation on layer structure this may improve, but the model is still overall good for predicting ridership.





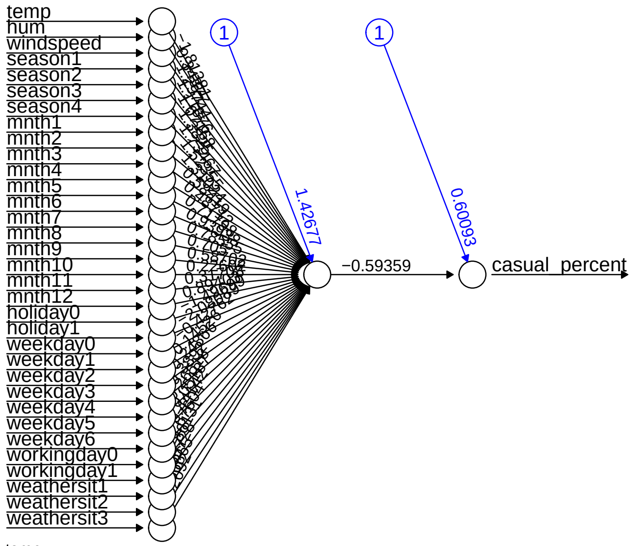


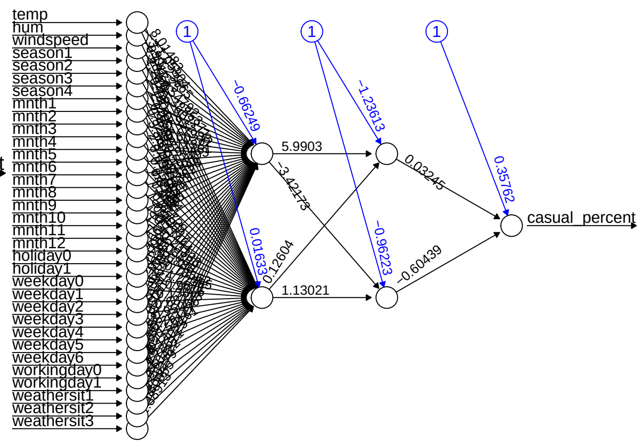


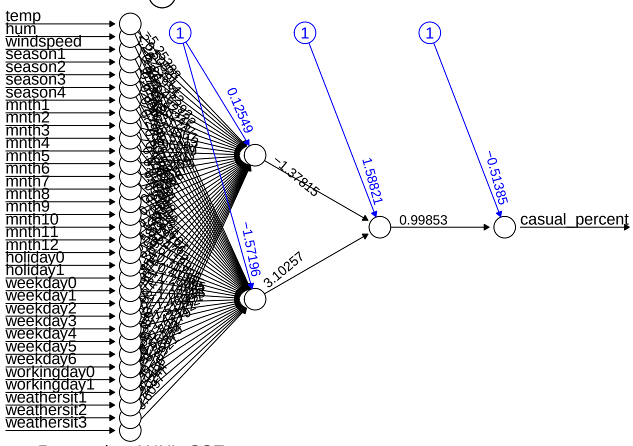

```
## [1] "Adjusted R Squared:  0.707377949393871"
```

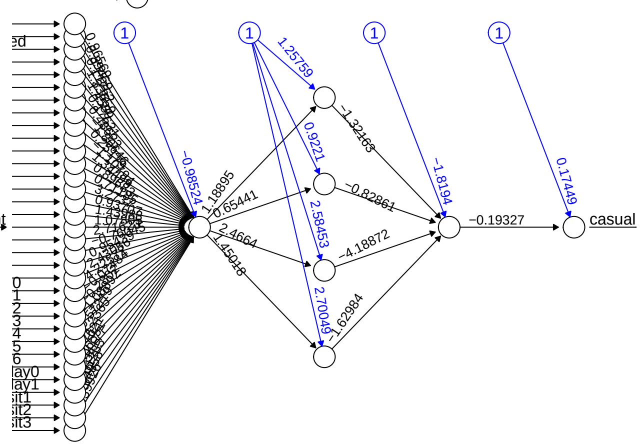
Predicting casual ridership percentage

This model was trained similarly to the total ridership neural network, but it used the calculated casual ridership percentage as the dependent variable. This relationship is important to understand when making business operational decisions as casual riders appear to be more weather and temporally dependent which may impact how system expansion plans are designed. This model was also trained on an 80/20 validation split and evaluated with residual SSE. The optimal layer structure was 1, 4, 1 and the overall adjusted R-squared was 0.824, which is about on par with the multiple linear regression.










```
## [1] "Adjusted R Squared:  0.824320033689755"
```

Conclusion

Model Performance

Model	Adj.R2
Polynomial	0.872

Model	Adj.R2
MLR	0.821
Stepwise AIC	0.779
Lasso	0.843
Random Forests	0.878
Neural Network Total	0.707
Neural Network % Casual	0.824

All models performed very well at explaining the variance in the original data as seen with the high adjusted R-squared metrics. The best model we created was the Random Forests model which makes sense as its decision tree structure might align with the decisions made by human customers looking at the weather and the day of the week when considering whether to rent the bicycles. The polynomial linear models' strong performance also indicates that ridership levels could have nonlinear effects that we can further model. The linear models all performed well and the coefficients may offer insights into the effects of weather and the date on ridership. Finally, the neural networks performed well but they were outperformed by the other models, which may indicate that there is not a nonlinear relationship those models did not take into account.

Applications

By better understanding the factors influencing ridership, the bike share company will be able to predict their future demand and account for trends in ridership that may not be obvious when looking only at daily ridership numbers. With further information about casual rider to registered rider conversion, we may be able to suggest increased marketing or coupons on holidays and weekends where casual riders are most prevalent. Finally, by understanding the relationship between weather and ridership, the bike share company may be able to utilize this information to put reserve bikes into service on days with optimal weather conditions or increase advertising when good conditions are forecast.

We believe there is a lot of information that can be used to inform these business decisions and many more with just the weather and temporal data, so we recommend that the bike share companies collect and utilize this information if they are not already doing so.