

Bike Share Linear Regression



Elizabeth Do

Data Background

- University of California Irvine Machine Learning Repository:
<https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset>
- 17,379 Instances & 17 Variables
- Bike share records by the hour from 2011 to 2012 based in Washington, D.C
- Dataset contains environmental and seasonal settings (windspeed, temperature, working day, season, etc.)

Questions

- What are the main predictors for bike share count?
- How does the environment and weather affect the bike share count?
- When is the best time to rent a bike share?

Exploratory Data Analysis

- Load dataset and libraries
 - library(caret)
 - library(ggplot2)
 - library(dplyr)
 - library(GGally)
- Data cleaning
 - Removed columns
 - Renamed columns
 - Changed data types to factors

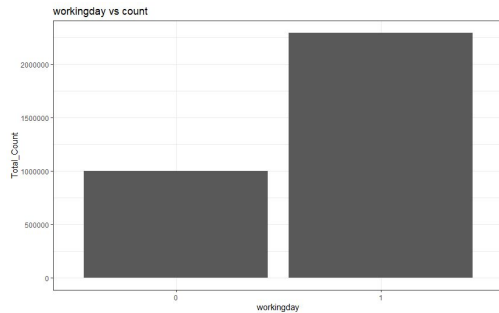
```
instant      dteday      season  yr      mnth      hr      holiday  weekday  workingday
Min.   : 1      Min. :2011-01-01  1:4242  0:8645  5      :1488  Min. : 0.00  0:16879  0: 5514
1st Qu.: 4346   1st Qu.:2011-07-04  2:4409  1:8734  7      :1488  1st Qu.: 6.00  1:2479  1:11865
Median : 8690   Median :2012-01-02  3:4496          12     :1483  Median :12.00  2:2453
Mean   : 8690   Mean   :2012-01-02  4:4232          8      :1475  Mean   :11.55  3:2475
3rd Qu.:13034   3rd Qu.:2012-07-02          3      :1473  3rd Qu.:18.00  4:2471
Max.   :17379   Max.   :2012-12-31          10     :1451  Max.   :23.00  5:2487
                        (other):8521          6:2512

weathersit    temp      atemp      hum      windspeed      casual      registered
Min.   :1.000  Min. :0.020  Min. :0.0000  Min. :0.0000  Min. :0.0000  Min. : 0.00  Min. : 0.0
1st Qu.:1.000  1st Qu.:0.340  1st Qu.:0.3333  1st Qu.:0.4800  1st Qu.:0.1045  1st Qu.: 4.00  1st Qu.: 34.0
Median :1.000  Median :0.500  Median :0.4848  Median :0.6300  Median :0.1940  Median :17.00  Median :115.0
Mean   :1.425  Mean   :0.497  Mean   :0.4758  Mean   :0.6272  Mean   :0.1901  Mean   :35.68  Mean   :153.8
3rd Qu.:2.000  3rd Qu.:0.660  3rd Qu.:0.6212  3rd Qu.:0.7800  3rd Qu.:0.2537  3rd Qu.: 48.00  3rd Qu.:220.0
Max.   :4.000  Max.   :1.000  Max.   :1.0000  Max.   :1.0000  Max.   :0.8507  Max.   :367.00  Max.   :886.0

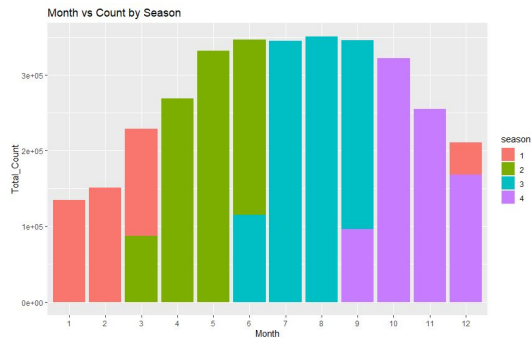
cnt
Min.   : 1.0
1st Qu.: 40.0
Median :142.0
Mean   :189.5
3rd Qu.:281.0
Max.   :977.0
```

Exploratory Data Analysis

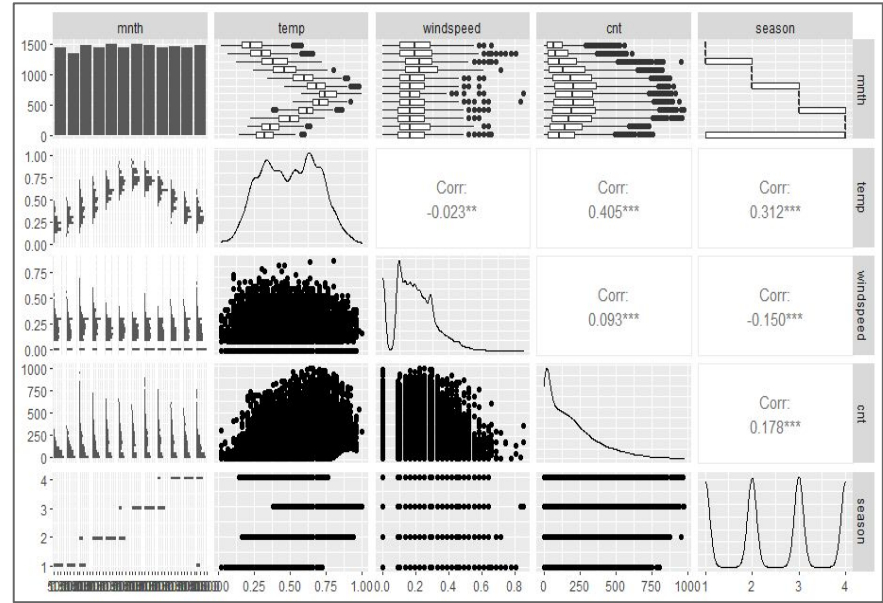
- Visualize distribution of counts
 - Working day vs. Count: increased bike rentals during working days



- Month vs Count by Season: increased bike rentals during Spring and Summer months

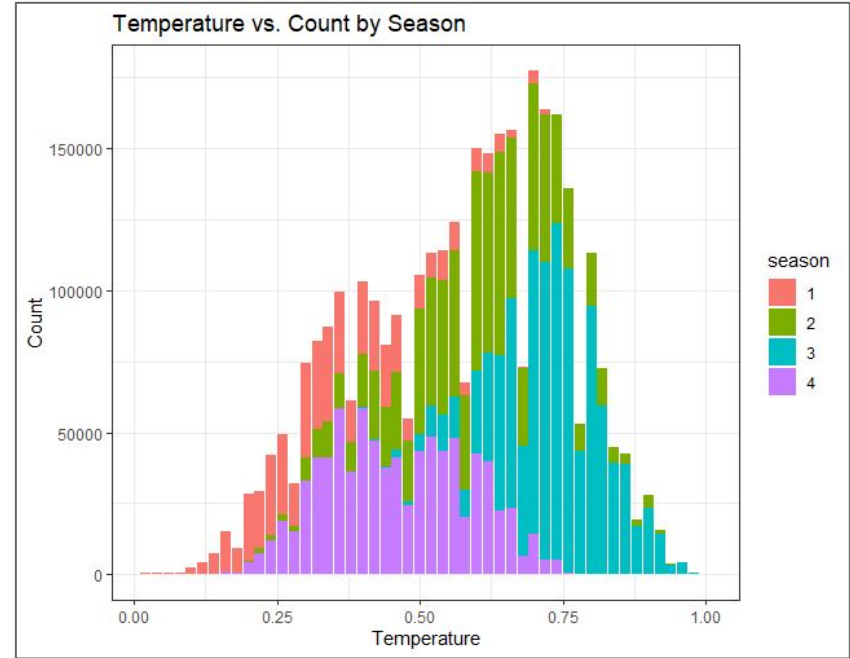


Variable Selection



- ggpairs to explore correlation between variables
- Highly positive correlation between count and temp at 0.405
- Determined that count and wind speed correlation is low at 0.093

Variable Selection



- Focused on temperature as the variable vs. count of bike rentals
- Increased in Spring and Summer
- Decreased in Fall and Winter

Forward Selection Model

Linear Regression with Forward Selection

13904 samples
16 predictor

No pre-processing

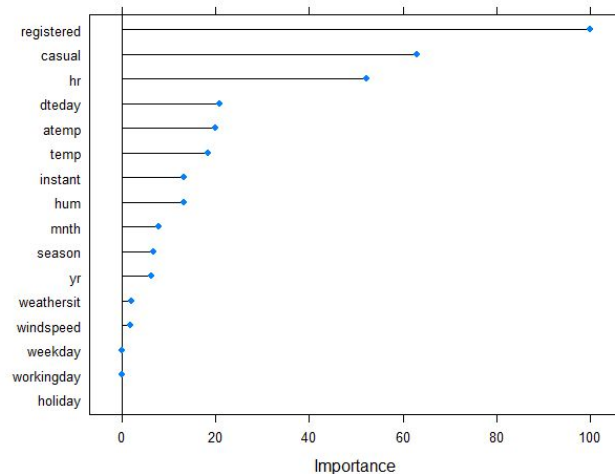
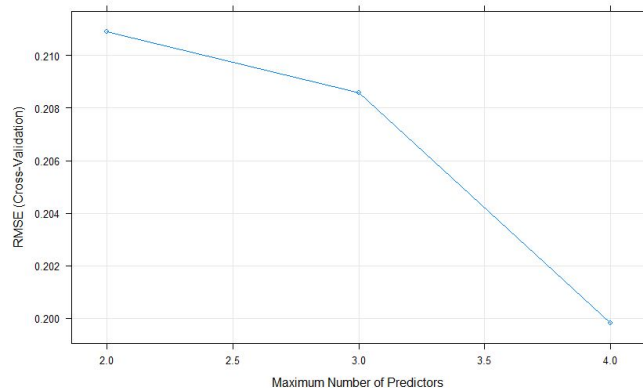
Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 12514, 12515, 12512, 12513, 12513, 12513, ...

Resampling results across tuning parameters:

| nvmax | RMSE | Rsquared | MAE |
|-------|-----------|-----------|-----------|
| 2 | 0.2109028 | 0.9556167 | 0.1398529 |
| 3 | 0.2085877 | 0.9565088 | 0.1398499 |
| 4 | 0.1998162 | 0.9599774 | 0.1352395 |

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was nvmax = 4.



Ridge Model

Ridge Regression

13904 samples
16 predictor

No pre-processing

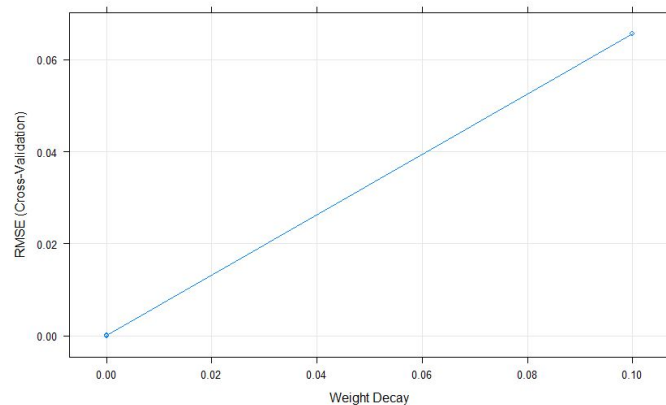
Resampling: Cross-validated (10 fold)

Summary of sample sizes: 12513, 12514, 12513, 12514, 12514, 12514, ...

Resampling results across tuning parameters:

| lambda | RMSE | Rsquared | MAE |
|--------|--------------|-----------|--------------|
| 0e+00 | 1.678867e-11 | 1.0000000 | 1.205701e-11 |
| 1e-04 | 7.280988e-05 | 1.0000000 | 5.892592e-05 |
| 1e-01 | 6.558327e-02 | 0.9965326 | 5.332291e-02 |

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was lambda = 0.



Full Model

```
> full_model
Linear Regression

13904 samples
  16 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 12514, 12513, 12513, 12513, 12513, 12514, ...
Resampling results:
```

| RMSE | Rsquared | MAE |
|--------------|----------|--------------|
| 5.863997e-15 | 1 | 5.059537e-15 |

Tuning parameter 'intercept' was held constant at a value of TRUE

```
> postResample(pred = pred_full, obs = data_test_proc$cnt)
```

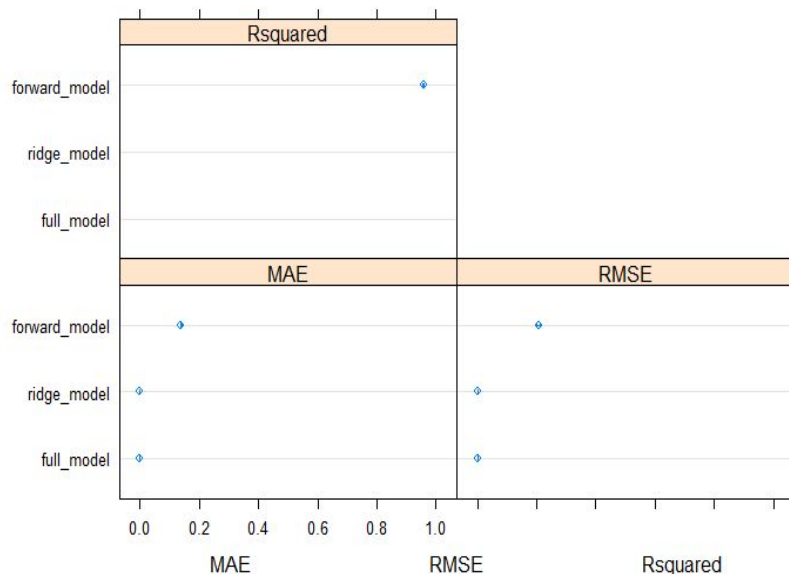
| RMSE | Rsquared | MAE |
|--------------|--------------|--------------|
| 8.650100e-15 | 1.000000e+00 | 7.506199e-15 |

```
> postResample(pred = pred_fit, obs = data_test_proc$cnt)
```

| RMSE | Rsquared | MAE |
|-----------|-----------|-----------|
| 0.9139233 | 0.1560137 | 0.6955093 |

- RMSE from the Full Model improved from the model fit from 0.9139233 to 8.650100e-15

Model Comparison



```
Call:  
summary.resamples(object = results)
```

```
Models: forward_model, ridge_model, full_model  
Number of resamples: 10
```

| MAE | Min. | 1st Qu. | Median | Mean |
|---------------|--------------|--------------|--------------|--------------|
| forward_model | 1.351373e-01 | 1.373775e-01 | 1.400522e-01 | 1.394791e-01 |
| ridge_model | 2.058812e-15 | 4.812974e-13 | 8.032226e-13 | 6.160524e-10 |
| full_model | 3.345588e-15 | 4.215177e-15 | 4.661787e-15 | 5.059537e-15 |

| | 3rd Qu. | Max. | NA's |
|---------------|--------------|--------------|------|
| forward_model | 1.415159e-01 | 1.426219e-01 | 0 |
| ridge_model | 2.318225e-11 | 5.465598e-09 | 0 |
| full_model | 4.776140e-15 | 1.136381e-14 | 0 |

| RMSE | Min. | 1st Qu. | Median | Mean |
|---------------|--------------|--------------|--------------|--------------|
| forward_model | 1.966329e-01 | 2.019988e-01 | 2.068263e-01 | 2.057752e-01 |
| ridge_model | 2.704072e-15 | 6.765759e-13 | 1.124692e-12 | 8.425584e-10 |
| full_model | 3.982160e-15 | 5.115196e-15 | 5.407496e-15 | 5.863997e-15 |

| | 3rd Qu. | Max. | NA's |
|---------------|--------------|--------------|------|
| forward_model | 2.091535e-01 | 2.174141e-01 | 0 |
| ridge_model | 3.211450e-11 | 7.459150e-09 | 0 |
| full_model | 5.754910e-15 | 1.177975e-14 | 0 |

| Rsquared | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---------------|-----------|-----------|-----------|-----------|-----------|-----------|
| forward_model | 0.9541719 | 0.9555462 | 0.9582487 | 0.9577666 | 0.9599802 | 0.9605839 |
| ridge_model | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 |
| full_model | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 |

| | NA's |
|---------------|------|
| forward_model | 0 |
| ridge_model | 0 |
| full_model | 0 |

- The Full Model had a significantly lower RMSE mean than both the Forward Selection Model & Ridge Model

Conclusion & Recommendations

Conclusion:

- Temperature and working day are the main predictors for bike rental share count
- The environment and weather affect the bike share count, since there is an increase in count as temperature increases
- The best time to rent a bike share is when it is a warmer temperature (Summer and Spring) and on a working day.

Recommendations:

- Increase the number of bike rentals in the warmer seasons for more usage
- Increase the number of bike rentals near offices on working days for commuters
- Schedule maintenance for bike rentals on weekend evenings when it is being used less

Questions?

Data Background

- University of California Irvine Machine Learning Repository:
<https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset>
- 17,379 Instances & 17 Variables
- Bike share records by the hour from 2011 to 2012 based in Washington, D.C
- Dataset contains environmental and seasonal settings (windspeed, temperature, working day, season, etc.)

Questions

- What are the main predictors for bike share count?
- How does the environment and weather affect the bike share count?
- When is the best time to rent a bike share?

Exploratory Data Analysis

- Load dataset and libraries
 - library(caret)
 - library(ggplot2)
 - library(dplyr)
 - library(GGally)
- Data cleaning
 - Removed columns
 - Renamed columns
 - Changed data types to factors

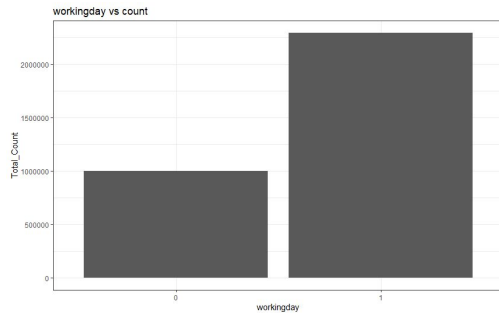
```
instant      dteday      season  yr      mnth      hr      holiday  weekday  workingday
Min.   : 1      Min. :2011-01-01  1:4242  0:8645  5      :1488  Min. : 0.00  0:16879  0: 5514
1st Qu.: 4346   1st Qu.:2011-07-04  2:4409  1:8734  7      :1488  1st Qu.: 6.00  1:2479  1:11865
Median : 8690   Median :2012-01-02  3:4496          12     :1483  Median :12.00  2:2453
Mean   : 8690   Mean   :2012-01-02  4:4232          8      :1475  Mean   :11.55  3:2475
3rd Qu.:13034  3rd Qu.:2012-07-02          3      :1473  3rd Qu.:18.00  4:2471
Max.   :17379   Max.   :2012-12-31          10     :1451  Max.   :23.00  5:2487
                    (other):8521                    6:2512

weathersit    temp      atemp      hum      windspeed      casual      registered
Min.   :1.000  Min. :0.020  Min. :0.0000  Min. :0.0000  Min. :0.0000  Min. : 0.00  Min. : 0.0
1st Qu.:1.000  1st Qu.:0.340  1st Qu.:0.3333  1st Qu.:0.4800  1st Qu.:0.1045  1st Qu.: 4.00  1st Qu.: 34.0
Median :1.000  Median :0.500  Median :0.4848  Median :0.6300  Median :0.1940  Median :17.00  Median :115.0
Mean   :1.425  Mean   :0.497  Mean   :0.4758  Mean   :0.6272  Mean   :0.1901  Mean   :35.68  Mean   :153.8
3rd Qu.:2.000  3rd Qu.:0.660  3rd Qu.:0.6212  3rd Qu.:0.7800  3rd Qu.:0.2537  3rd Qu.: 48.00  3rd Qu.:220.0
Max.   :4.000  Max.   :1.000  Max.   :1.0000  Max.   :1.0000  Max.   :0.8507  Max.   :367.00  Max.   :886.0

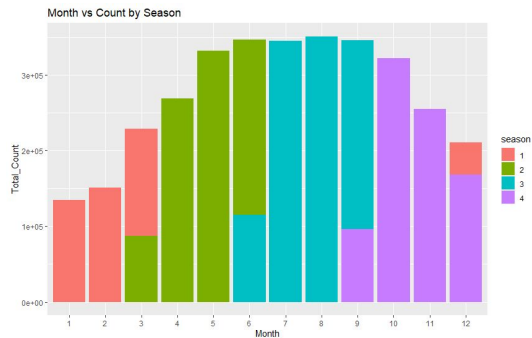
cnt
Min.   : 1.0
1st Qu.: 40.0
Median :142.0
Mean   :189.5
3rd Qu.:281.0
Max.   :977.0
```


Exploratory Data Analysis

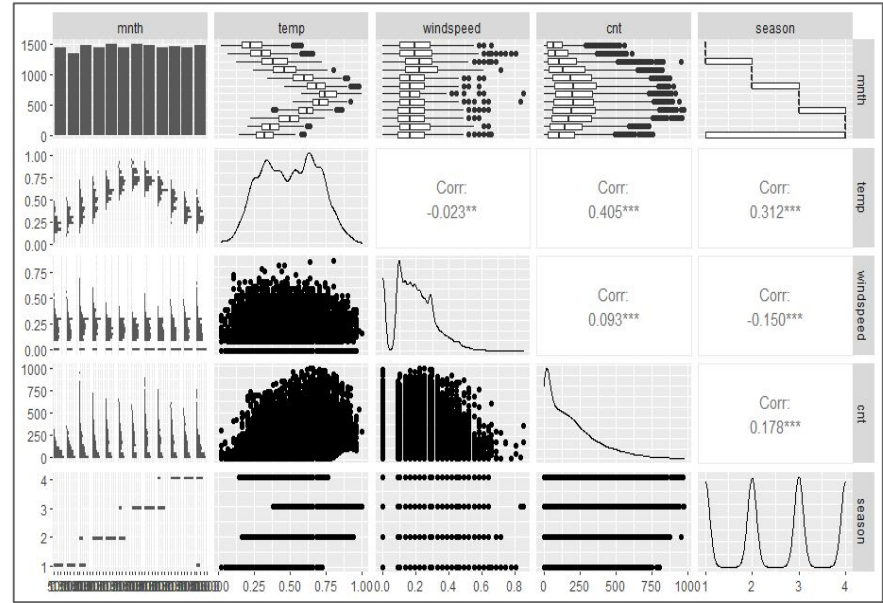
- Visualize distribution of counts
 - Working day vs. Count: increased bike rentals during working days



- Month vs Count by Season: increased bike rentals during Spring and Summer months

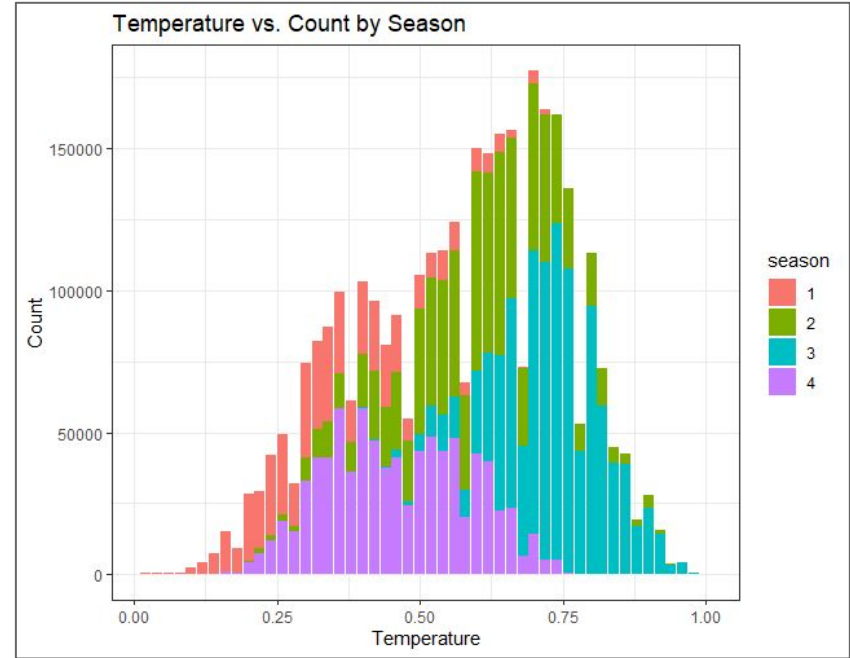


Variable Selection



- ggpairs to explore correlation between variables
- Highly positive correlation between count and temp at 0.405
- Determined that count and wind speed correlation is low at 0.093

Variable Selection



- Focused on temperature as the variable vs. count of bike rentals
- Increased in Spring and Summer
- Decreased in Fall and Winter

Forward Selection Model

Linear Regression with Forward Selection

13904 samples
16 predictor

No pre-processing

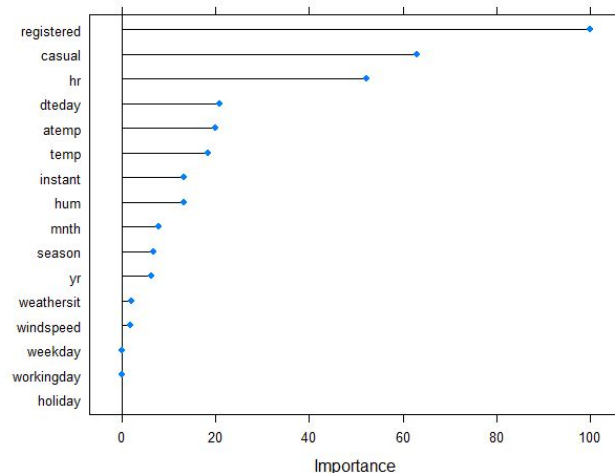
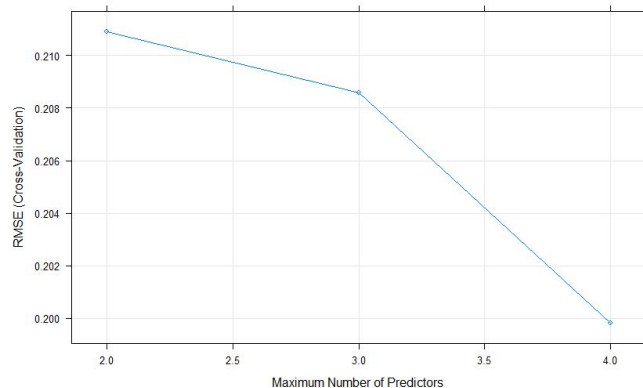
Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 12514, 12515, 12512, 12513, 12513, 12513, ...

Resampling results across tuning parameters:

| nvmax | RMSE | Rsquared | MAE |
|-------|-----------|-----------|-----------|
| 2 | 0.2109028 | 0.9556167 | 0.1398529 |
| 3 | 0.2085877 | 0.9565088 | 0.1398499 |
| 4 | 0.1998162 | 0.9599774 | 0.1352395 |

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was nvmax = 4.



Ridge Model

Ridge Regression

13904 samples
16 predictor

No pre-processing

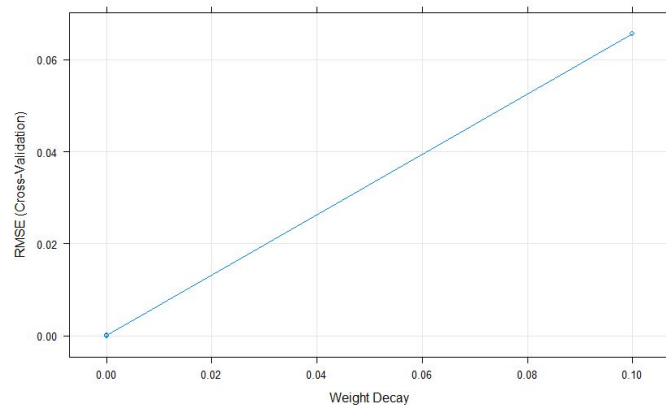
Resampling: Cross-validated (10 fold)

Summary of sample sizes: 12513, 12514, 12513, 12514, 12514, 12514, ...

Resampling results across tuning parameters:

| lambda | RMSE | Rsquared | MAE |
|--------|--------------|-----------|--------------|
| 0e+00 | 1.678867e-11 | 1.0000000 | 1.205701e-11 |
| 1e-04 | 7.280988e-05 | 1.0000000 | 5.892592e-05 |
| 1e-01 | 6.558327e-02 | 0.9965326 | 5.332291e-02 |

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was lambda = 0.



Full Model

```
> full_model
Linear Regression

13904 samples
  16 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 12514, 12513, 12513, 12513, 12513, 12514, ...
Resampling results:
```

| RMSE | Rsquared | MAE |
|--------------|----------|--------------|
| 5.863997e-15 | 1 | 5.059537e-15 |

Tuning parameter 'intercept' was held constant at a value of TRUE

```
> postResample(pred = pred_full, obs = data_test_proc$cnt)
```

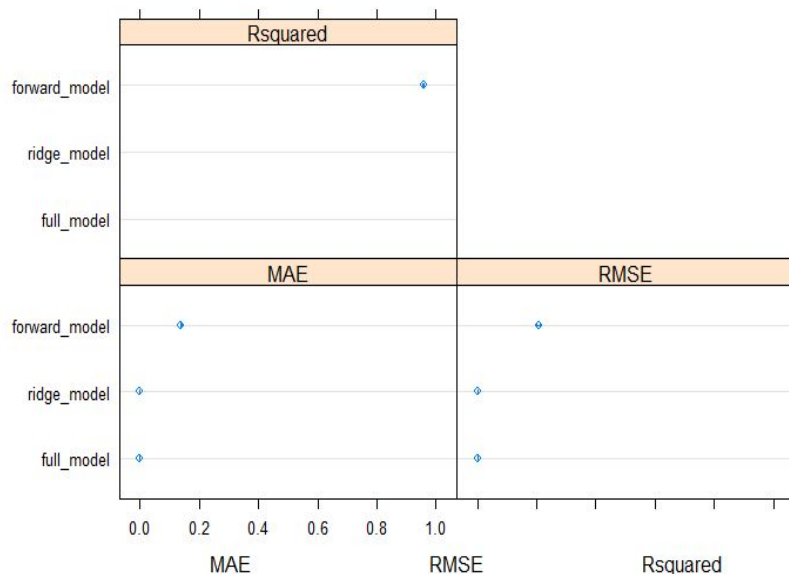
| RMSE | Rsquared | MAE |
|--------------|--------------|--------------|
| 8.650100e-15 | 1.000000e+00 | 7.506199e-15 |

```
> postResample(pred = pred_fit, obs = data_test_proc$cnt)
```

| RMSE | Rsquared | MAE |
|-----------|-----------|-----------|
| 0.9139233 | 0.1560137 | 0.6955093 |

- RMSE from the Full Model improved from the model fit from 0.9139233 to 8.650100e-15

Model Comparison



```
Call:  
summary.resamples(object = results)
```

```
Models: forward_model, ridge_model, full_model  
Number of resamples: 10
```

| MAE | Min. | 1st Qu. | Median | Mean |
|---------------|--------------|--------------|--------------|--------------|
| forward_model | 1.351373e-01 | 1.373775e-01 | 1.400522e-01 | 1.394791e-01 |
| ridge_model | 2.058812e-15 | 4.812974e-13 | 8.032226e-13 | 6.160524e-10 |
| full_model | 3.345588e-15 | 4.215177e-15 | 4.661787e-15 | 5.059537e-15 |

| | 3rd Qu. | Max. | NA's |
|---------------|--------------|--------------|------|
| forward_model | 1.415159e-01 | 1.426219e-01 | 0 |
| ridge_model | 2.318225e-11 | 5.465598e-09 | 0 |
| full_model | 4.776140e-15 | 1.136381e-14 | 0 |

| RMSE | Min. | 1st Qu. | Median | Mean |
|---------------|--------------|--------------|--------------|--------------|
| forward_model | 1.966329e-01 | 2.019988e-01 | 2.068263e-01 | 2.057752e-01 |
| ridge_model | 2.704072e-15 | 6.765759e-13 | 1.124692e-12 | 8.425584e-10 |
| full_model | 3.982160e-15 | 5.115196e-15 | 5.407496e-15 | 5.863997e-15 |

| | 3rd Qu. | Max. | NA's |
|---------------|--------------|--------------|------|
| forward_model | 2.091535e-01 | 2.174141e-01 | 0 |
| ridge_model | 3.211450e-11 | 7.459150e-09 | 0 |
| full_model | 5.754910e-15 | 1.177975e-14 | 0 |

| Rsquared | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---------------|-----------|-----------|-----------|-----------|-----------|-----------|
| forward_model | 0.9541719 | 0.9555462 | 0.9582487 | 0.9577666 | 0.9599802 | 0.9605839 |
| ridge_model | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 |
| full_model | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 |

| | NA's |
|---------------|------|
| forward_model | 0 |
| ridge_model | 0 |
| full_model | 0 |

- The Full Model had a significantly lower RMSE mean than both the Forward Selection Model & Ridge Model

Conclusion & Recommendations

Conclusion:

- Temperature and working day are the main predictors for bike rental share count
- The environment and weather affect the bike share count, since there is an increase in count as temperature increases
- The best time to rent a bike share is when it is a warmer temperature (Summer and Spring) and on a working day.

Recommendations:

- Increase the number of bike rentals in the warmer seasons for more usage
- Increase the number of bike rentals near offices on working days for commuters
- Schedule maintenance for bike rentals on weekend evenings when it is being used less