CAPITAL BIKESHARE

Arfa Tahir, Kirill Kim, Matthew Ng, Riwaaz B. Sijapati, Sangmo Lama, Tugba Yildirim

Team 5

BACKGROUND

- Real time rental data was collected by Capital Bikeshare from 2011–2012
  - D.C. based bike sharing company that currently has"… 4,300 bikes and 500+ stations across 7 jurisdictions: Washington, DC.; Arlington, VA; Alexandria, VA; Montgomery, MD; Prince George's County, MD; Fairfax County, VA; and the City of Falls Church, VA."[1]
  - Capital bikeshare provides their rental data from 2011 to 2019 for public use.
- Rental data was collated with time metrics [2] and weather records [3] to and uploaded to Kaggle by user Mark Kaghazgarian.[4]

1. http://capitalbikeshare.com/system-data
2. http://dchr.dc.gov/page/holiday-schedule
3. http://www.freemeteo.com
4. https://www.kaggle.com/marklvl/bike-sharing-dataset

MOTIVATION

- Number of bikes operating through a bike share company in the US alone has increased from 42,500 in 2016 to about 100,000 by the end of 2017.[1]
- Companies like Capital Bikeshare, Citi Bike, Hubway and Divvy who operate in cities with a large population have proven successful in providing a inexpensive and flexible way to commute.
- Great way for city inhabitants and tourists to explore the city.
- Environmentally friendly alternatives for transportation in a dense city such as Washington DC or New York City.

1. https://nacto.org/wp-content/uploads/2018/05/NACTO-Bike-Share-2017.pdf

CAN WE PREDICT BIKE RENTALS BASED ON TIME AND WEATHER METRICS?

RAW DATASET

Variable Name	Value example	Description
instant	1, 2, 3 … 730, 731	Record Index
cnt	4035	Total rental bike count including both registered and casual riders
season	1	Season: spring, summer, fall, and winter
	2	
	3	
	4	
yr	0	Year from January 1st, 2011 to December 31st, 2012
	1	
mnth	1, 2, … 12	Month from January to December
holiday	0	Whether the day was a holiday or not
	1	
weekday	0, 1, … 6	Day of the week
workingday	0	If the day is neither a holiday nor a weekend
	1	
weatheritis	1	Physically what type of weather it was that day
	2	
	3	
	4	
temp	0.344167	Feels-Like Temperature in Celsius
atemp	0.363625	Actual Temperature in Celsius
humidity	80.5833%	Humidity
windspeed	0.160446	Windspeed in kilometer per hour

POST CLEANUP DATASET

Post Clean Variable Name	Post-Clean Up Value Example	Description
instant	1, 2, 3 … 730, 731	Record Index
count	4036	Total rental bike count including both registered and casual riders
season	Spring	Season: spring, summer, fall, and winter
	Summer	
	Fall	
	Winter	
year	Year 2011	Year from January 1st, 2011 to December 31st, 2012
	Year 2012	
month	Jan, Feb … Dec	Month from January to December
holiday	Not a Holiday	Whether the day was a holiday or not
	Holiday	
weekday	Sun, Mon … Sat	Day of the week
workingday	Not A Working Day	If the day is neither a holiday nor a weekend
	Working Day	
weathertype	Clear: Good (Clear or clear with few or partly cloudy	Physically what type of weather it was that day
	Cloudy: Adequet (Misty, Cloudy and Misty, Broken clouds	
	LightRain: Bad (Light Snow/Rain Slight thunderstorm	
	HeavyRain: Terrible (Heavy Rain/Snow, Ice Pallets, Thunderstorm, Dense Fog)	
tempdenorm	18.86°	Feels-Like Temperature in Celsius
acttempdenorm	22.50605°	Actual Temperature in Celsius
humiditydenorm	88.7917%	Humidity
windspeeddenorm	12.875725 Km/h	Windspeed in kilometer per hour

REMOVED FROM CONSIDERATION

Removed from Consideration		
Variable Name	**Value Description/Example**	**Description**
dteday	2011-01-01, 2011-01-02 … 2012-12-30, 2012-12-31	Date from January 1st, 2011- December 31rd, 2012
casual	1710	Count of casual riders
registered	2481	Count of registered riders

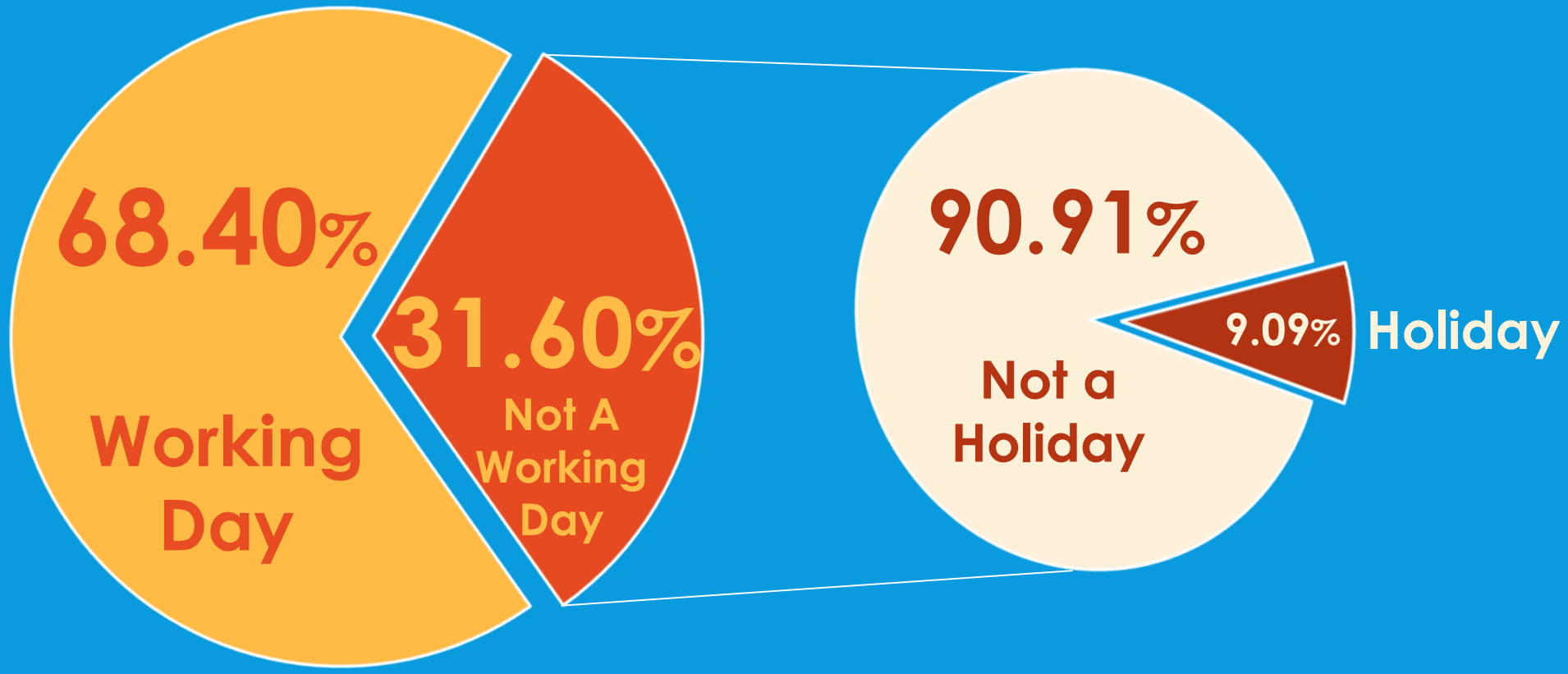
DATA SUMMARY STATISTICS

Metric Type	Statistics	Standard Deviation	Minimum	25th Percentile	Mean	75th Percentile	Maximum
Rider Break Down	Casual	686.60	2.00	315.50	848.20	1096.00	3410.00
	Registered	1560.30	20.00	2497.00	3656.20	4776.50	6946.00
	Count	1937.20	22.00	3152.00	4504.30	5956.00	8714.00
Weather Metrics	Temperature	7.50	2.40	13.80	20.30	26.90	35.30
	Actual temperature	8.10	4.00	16.90	23.70	30.40	42.00
	Humidity	14.20	0.00	52.00	62.80	73.00	97.20
	Windspeed	5.20	1.50	9.00	12.80	15.60	34.00

Metric Type	Statistics	Statistics	Count
Weather Metrics	**Season**	Fall	188
		Spring	181
		Summer	184
		Winter	178
	Weather Type	Clear: Good	463
		Cloudy: Adequet	247
		LightRain:Bad	21
		HeavyRain:Terrible	0

Metric Type	Statistics	Statistics	Count
Time Metrics	**Year**	Year 2011	365
		Year 2012	366*
	Month	Jan	62
		Feb	57
		Mar	62
		Other	550
	Weekday	Mon	550
		Tues	696.9
		Thr	843.8
		Other	990.7

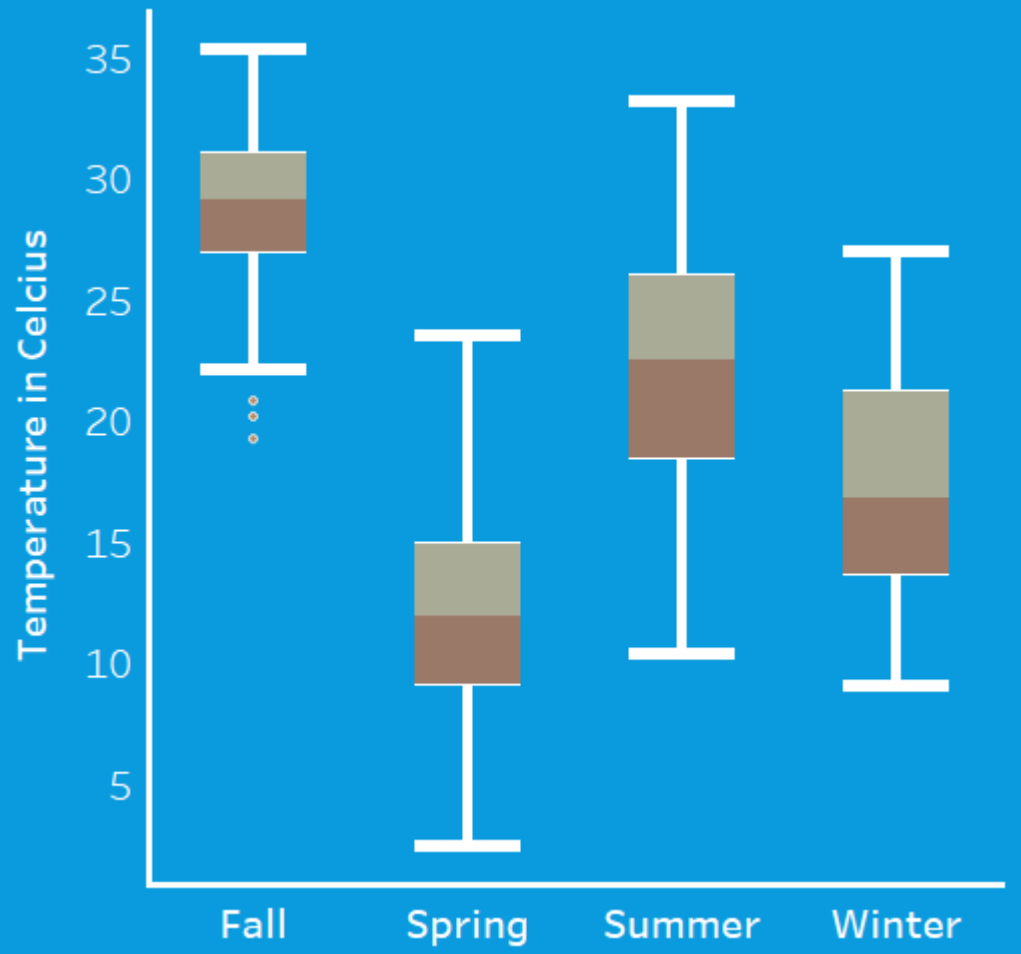
TYPE OF DAY


68.40%
Working Day
31.60%
Not A Working Day
90.91%
Not a Holiday
9.09%
Holiday

TEMPERATURE-SEASON BOX PLOT IN FAHRENHEIT

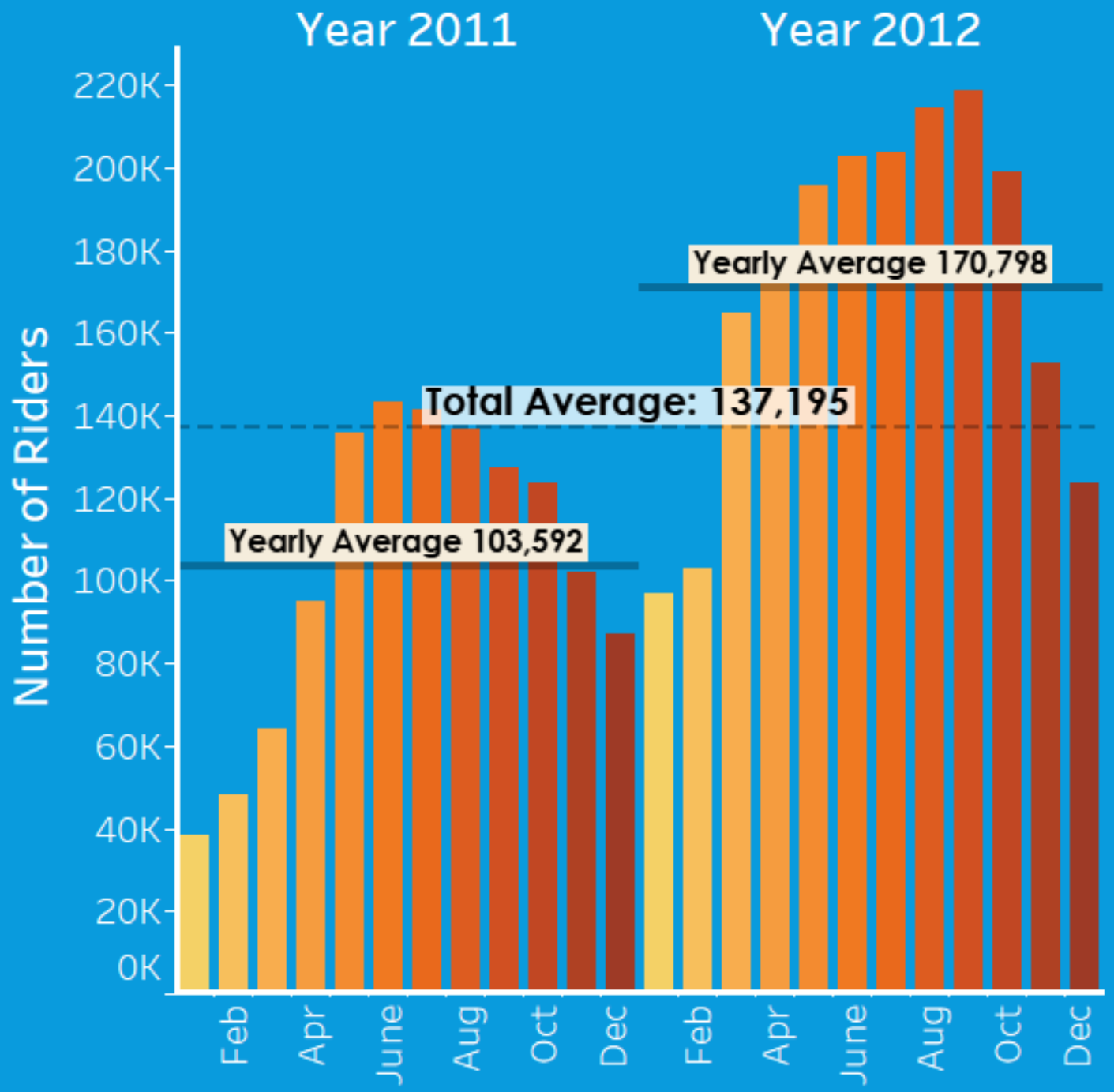
Fall	
Upper Whisker	35.33
Upper Quartile	30.99
Median	29.14
Lower Quartile	26.92
Lower Whisker	22.14

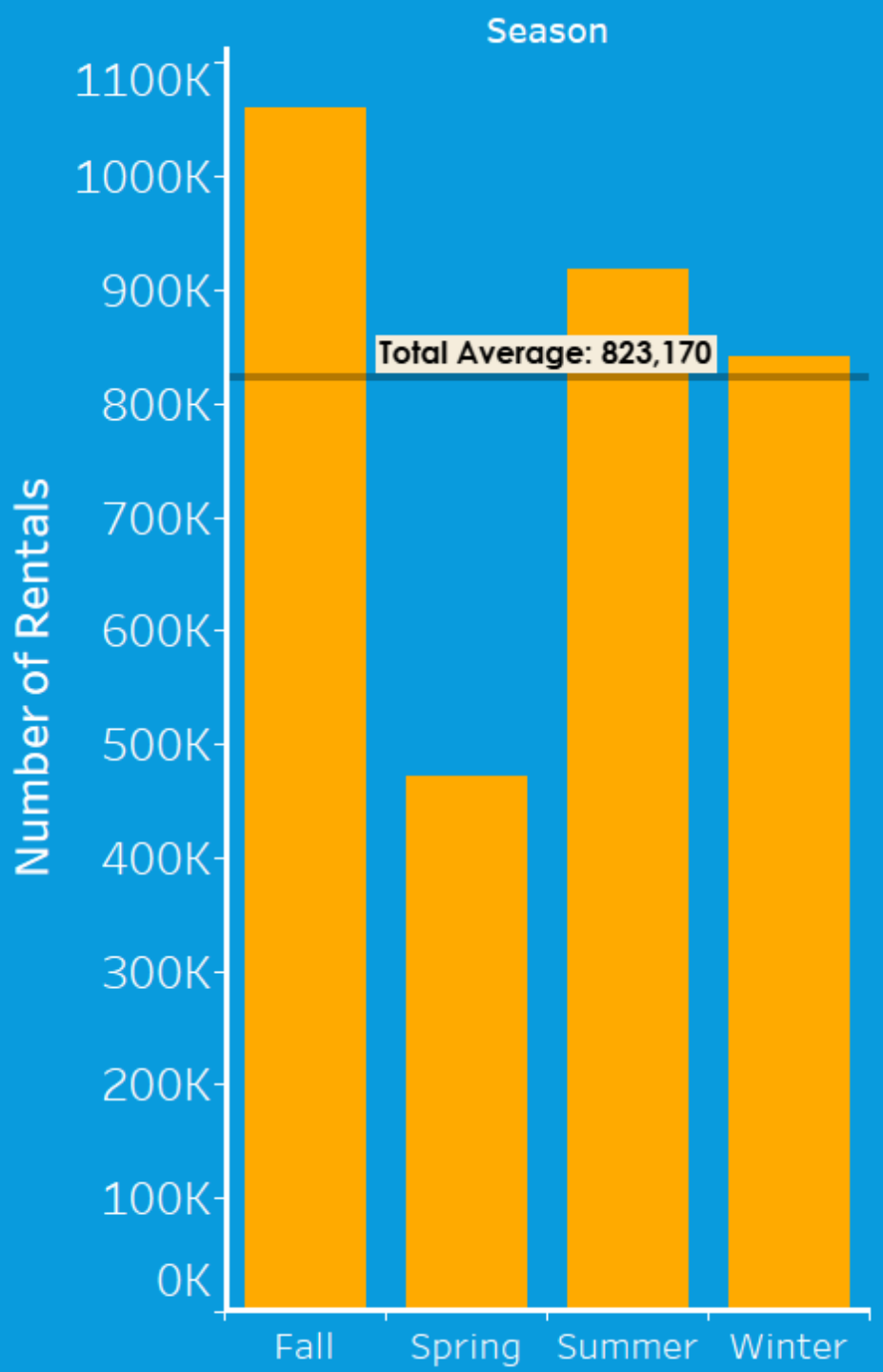
Spring	
Upper Whisker	23.47
Upper Quartile	14.92
Median	11.96
Lower Quartile	9.07
Lower Whisker	2.42



Summer	
Upper Whisker	33.14
Upper Quartile	25.97
Median	22.53
Lower Quartile	18.42
Lower Whisker	10.37

Winter	
Upper Whisker	26.96
Upper Quartile	21.22
Median	16.83
Lower Quartile	13.55
Lower Whisker	9.05


Year 2011
Year 2012
Number of Riders
220K
200K
180K
160K
140K
120K
100K
80K
60K
40K
20K
0K
Yearly Average 170,798
Total Average: 137,195
Yearly Average 103,592
Feb
Apr
June
Aug
Oct
Dec
Feb
Apr
June
Aug
Oct
Dec


Season
1100K
1000K
900K
800K
700K
600K
500K
400K
300K
200K
100K
0K
Number of Rentals
Total Average: 823,170
Fall
Spring
Summer
Winter

DATA MINING METHOD

Multilinear Regression	**Regression Tree**
We used multi linear regression, as there are multiple independent variables that could affect the number of bikes rented. For instance, weather conditions, precipitation, day of week, season and other factors can affect the rental behaviors. We expect to see multilinear relationship between the total number of bike rental users and the variables of interest We used the multiple linear regression model, then evaluate the adequacy of the fitted model by doing a cross validation	They are very interpretable. Making predictions is fast (no complicated calculations, just looking up constants in the tree. It's easy to understand what variables are important in making the prediction. The internal nodes (splits) are those variables that most largely reduced the SSE.

MULTILINEAR REGRESSION

- Method used to determine important variables: Backward Selection.
- Removed working day, weekday and actual temperature from model
- Original model issues:
  - Year variable had a high effect on model, but was not important for our purposes so it was removed.

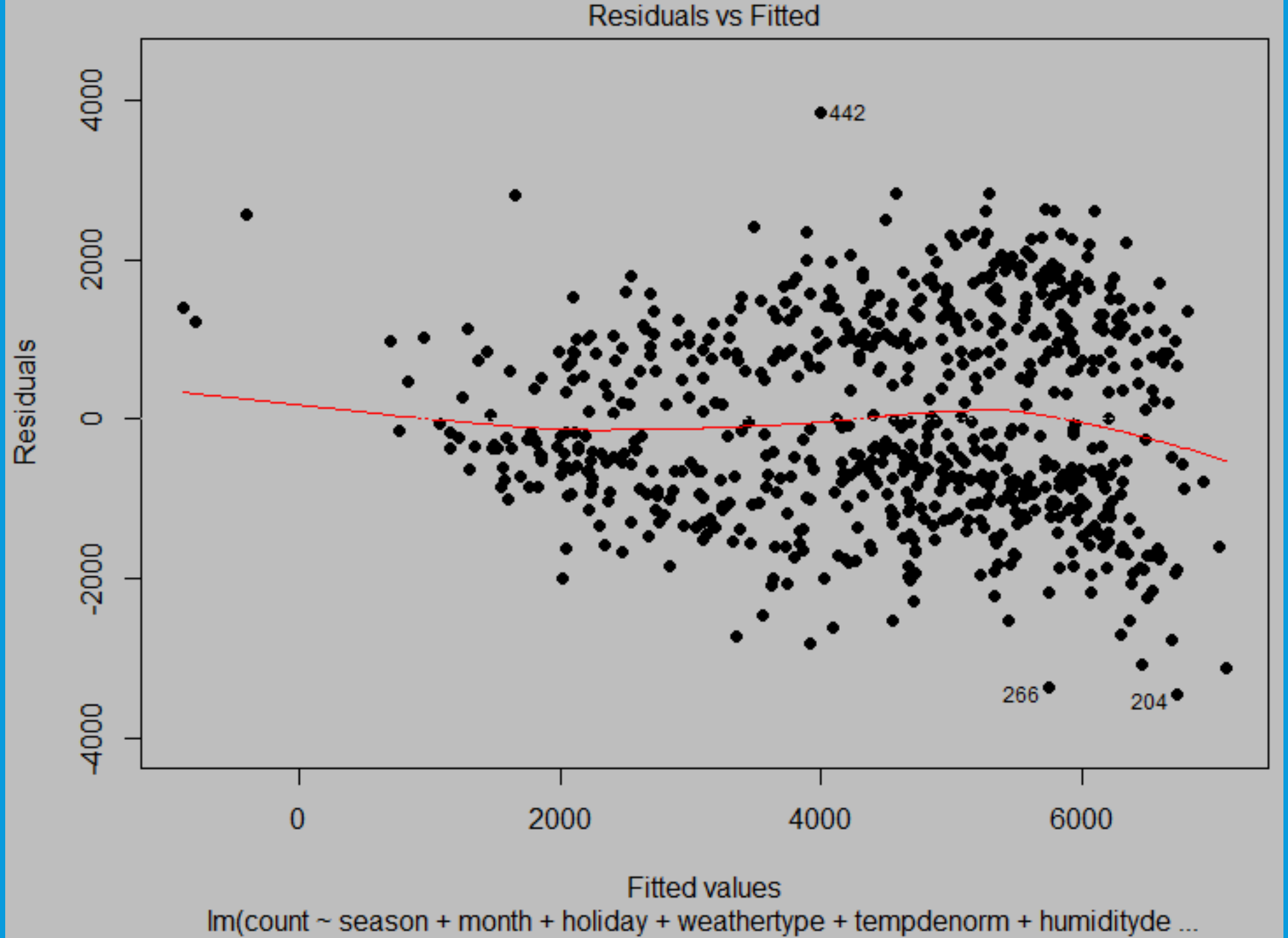
MULTILINEAR REGRESSION ANALYSIS

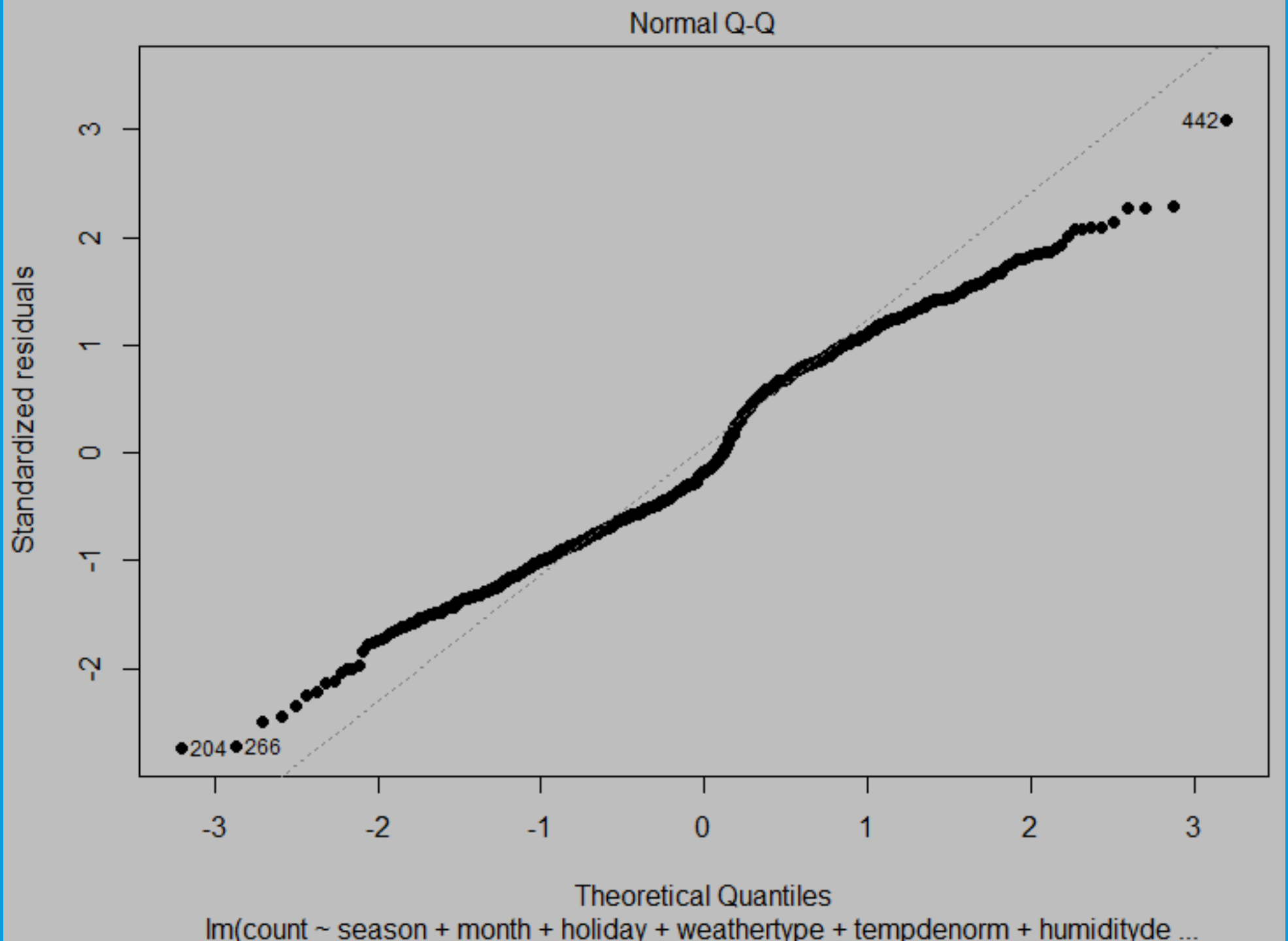
Variable type	Variables	Coefficients	P Value	Significance
Intercept	*Intercept*	3279.43	2.00E-07	***
Season	Spring	-744.55	0.034	*
	Summer	210.65	0.489	
	Winter	748.05	0.018	
Month	Aug	-391.56	0.323	
	Dec	7.28	0.985	
	Feb	123.82	0.735	
	Jan	115.18	0.777	
	Jul	-942.16	0.021	*
	June	425.64	0.135	
	Mar	307.36	0.311	
	May	100.19	0.684	
	Nov	-119.85	0.781	
	Oct	299.16	0.477	
	Sept	495.52	0.184	
Holiday	Not A Holiday	672.16	0.018	*

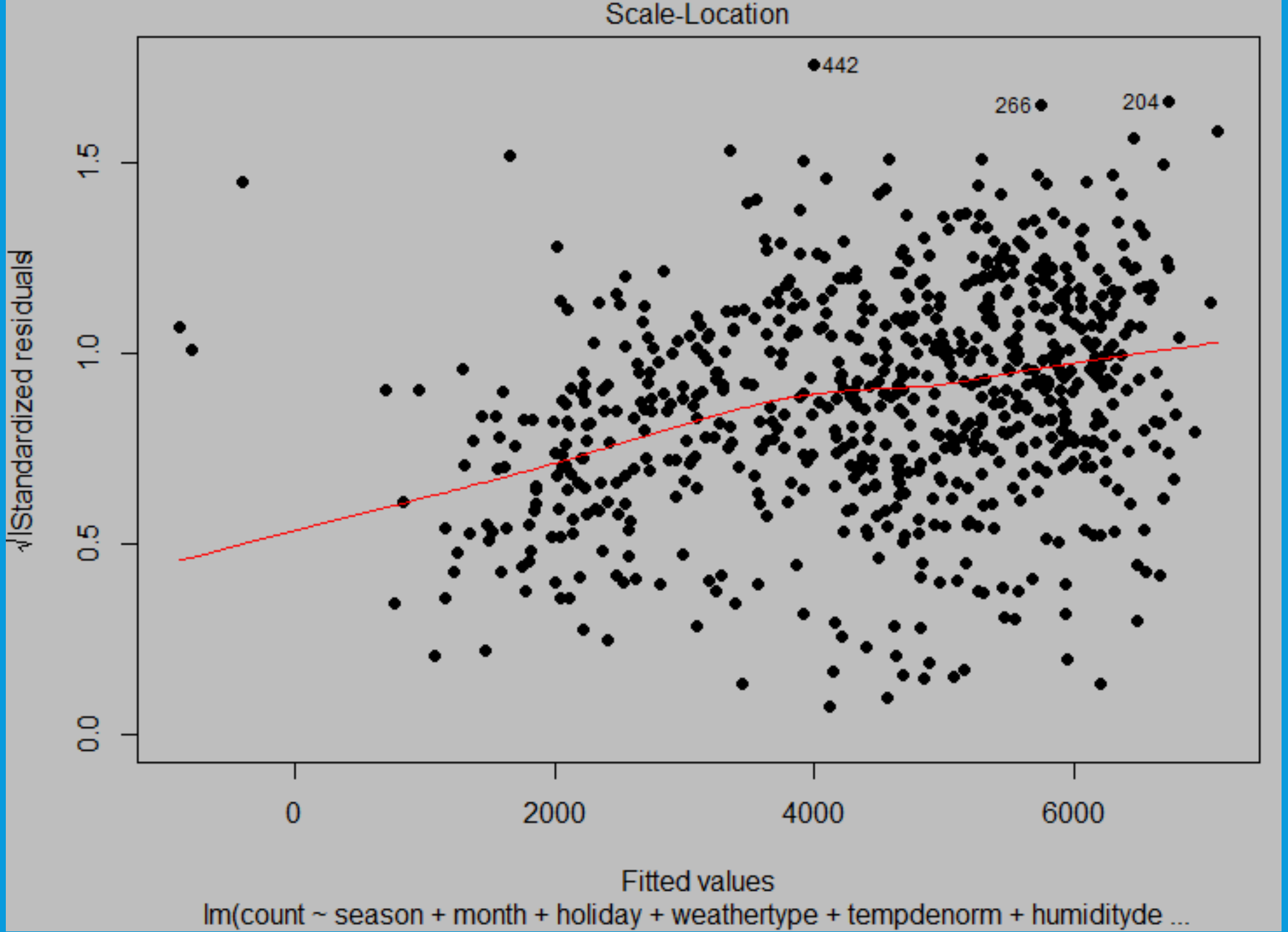
MULTILINEAR REGRESSION ANALYSIS

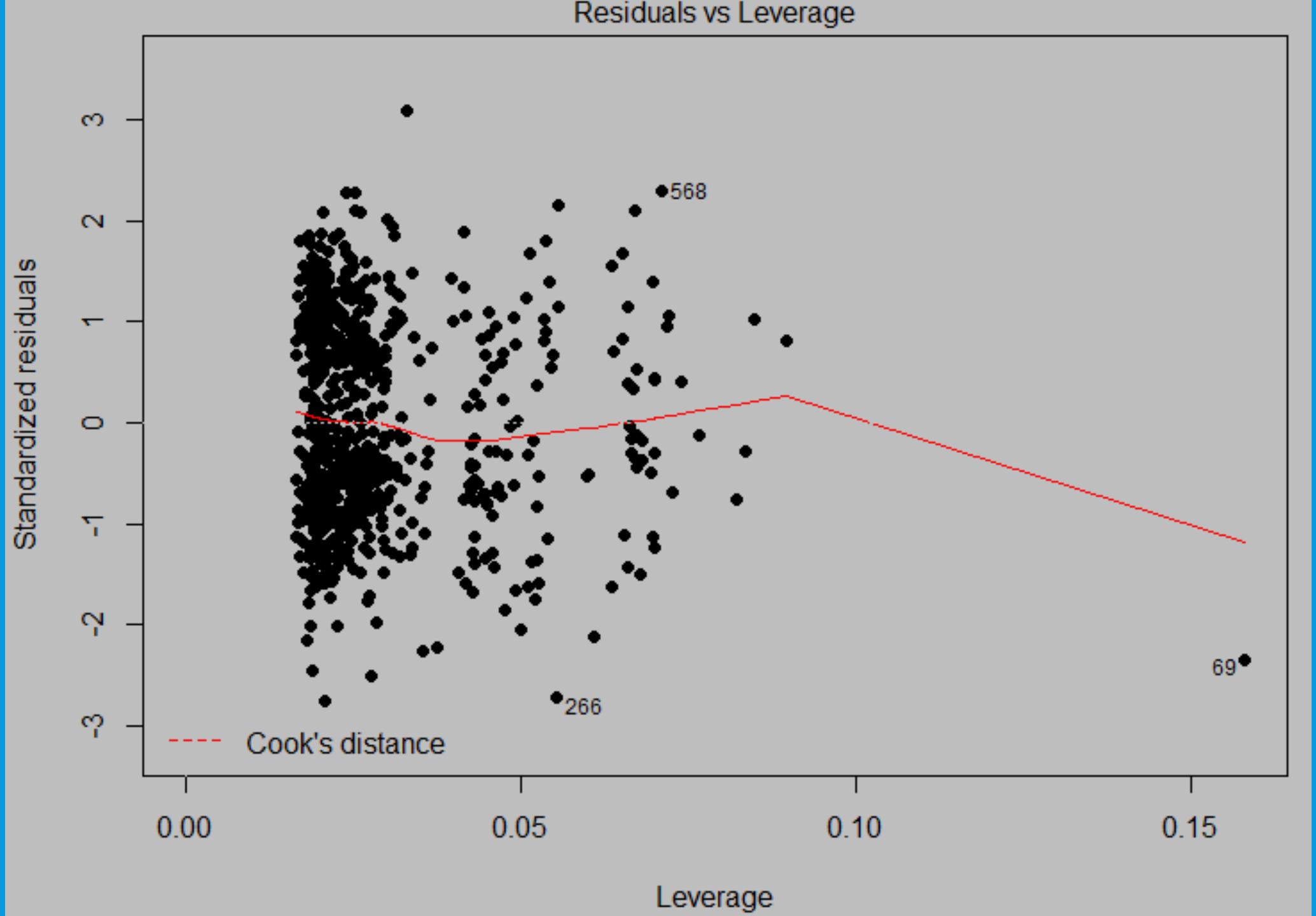
Variable type	Variables	Coefficients	P Value	Significance
Weather type	Cloudy:Adequet	-209.18	0.096	.
	LightRain: Bad	-1878.72	7.80E-09	***
Weather metrics	Temp	166.17	2.00E-16	***
	Humidity	-31.97	2.50E-11	***
	Windspeed	-53.55	1.00E-07	***

Multiple R-Squared Value = .584
Lowest Cross Validation MSE achieved = 1649346


Residuals vs Fitted
Residuals
4000
2000
0
-2000
-4000
442
266
204
0
2000
4000
6000
Fitted values
lm(count ~ season + month + holiday + weathertype + tempdenorm + humidityde ...


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(count ~ season + month + holiday + weathertype + tempdenorm + humidityde ...
442
204
266
3
2
1
0
-1
-2
-3


Scale-Location
442
266
204
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
2000
4000
6000
Fitted values
lm(count ~ season + month + holiday + weathertype + tempdenorm + humidityde ...

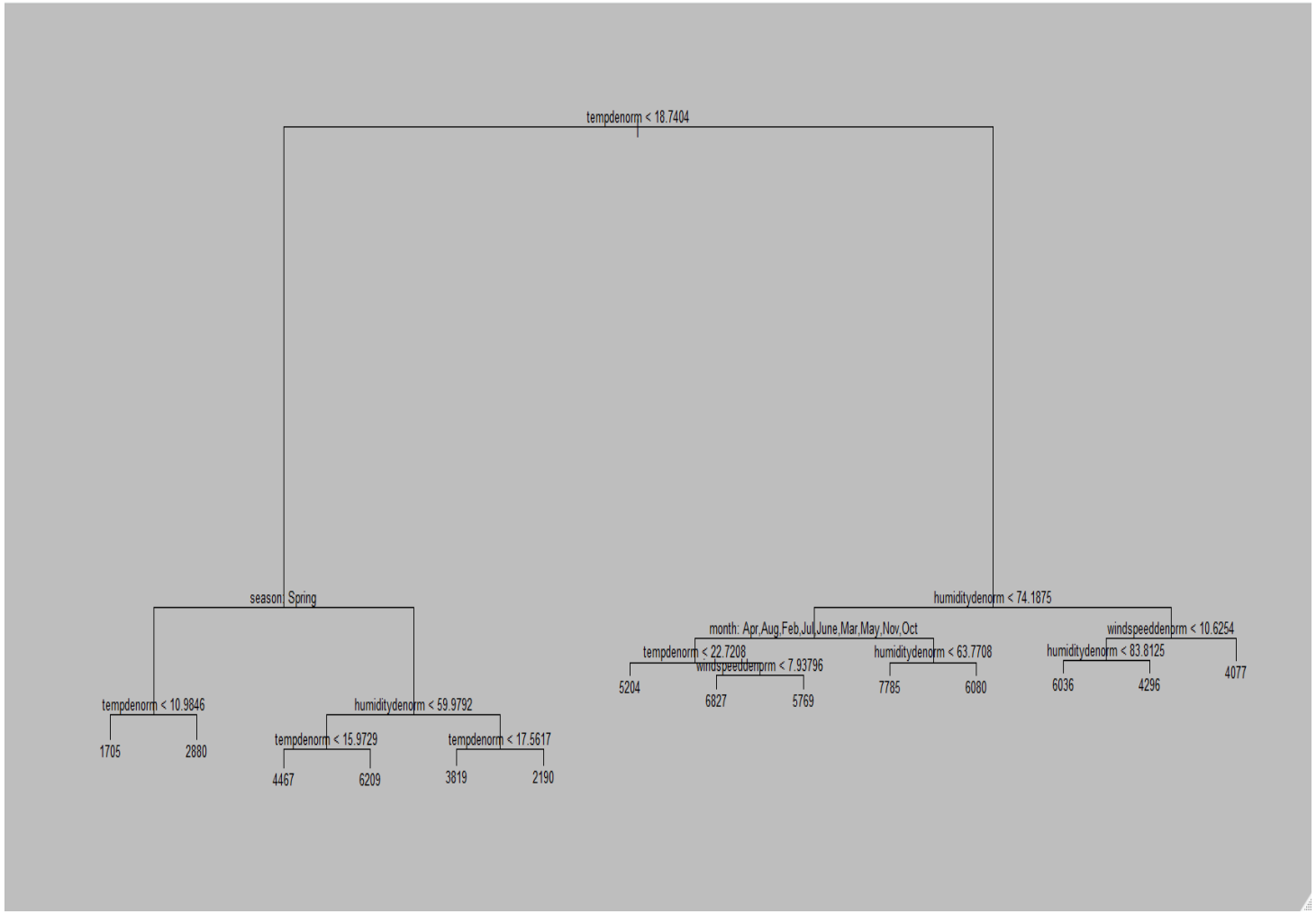

Residuals vs Leverage
Standardized residuals
3
2
1
0
-1
-2
-3
568
266
69
Cook's distance
0.00
0.05
0.10
0.15
Leverage

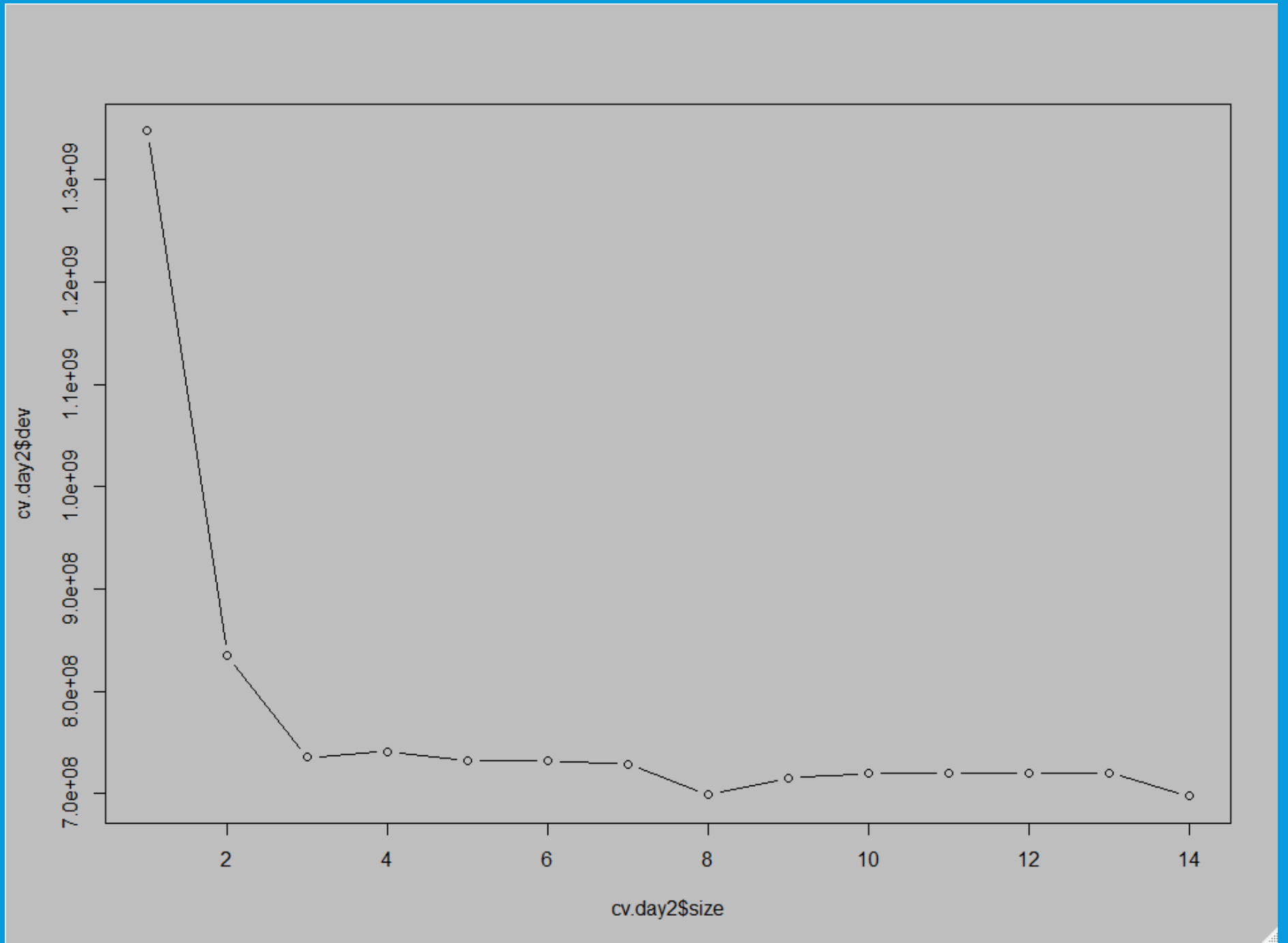
REGRESSION TREE

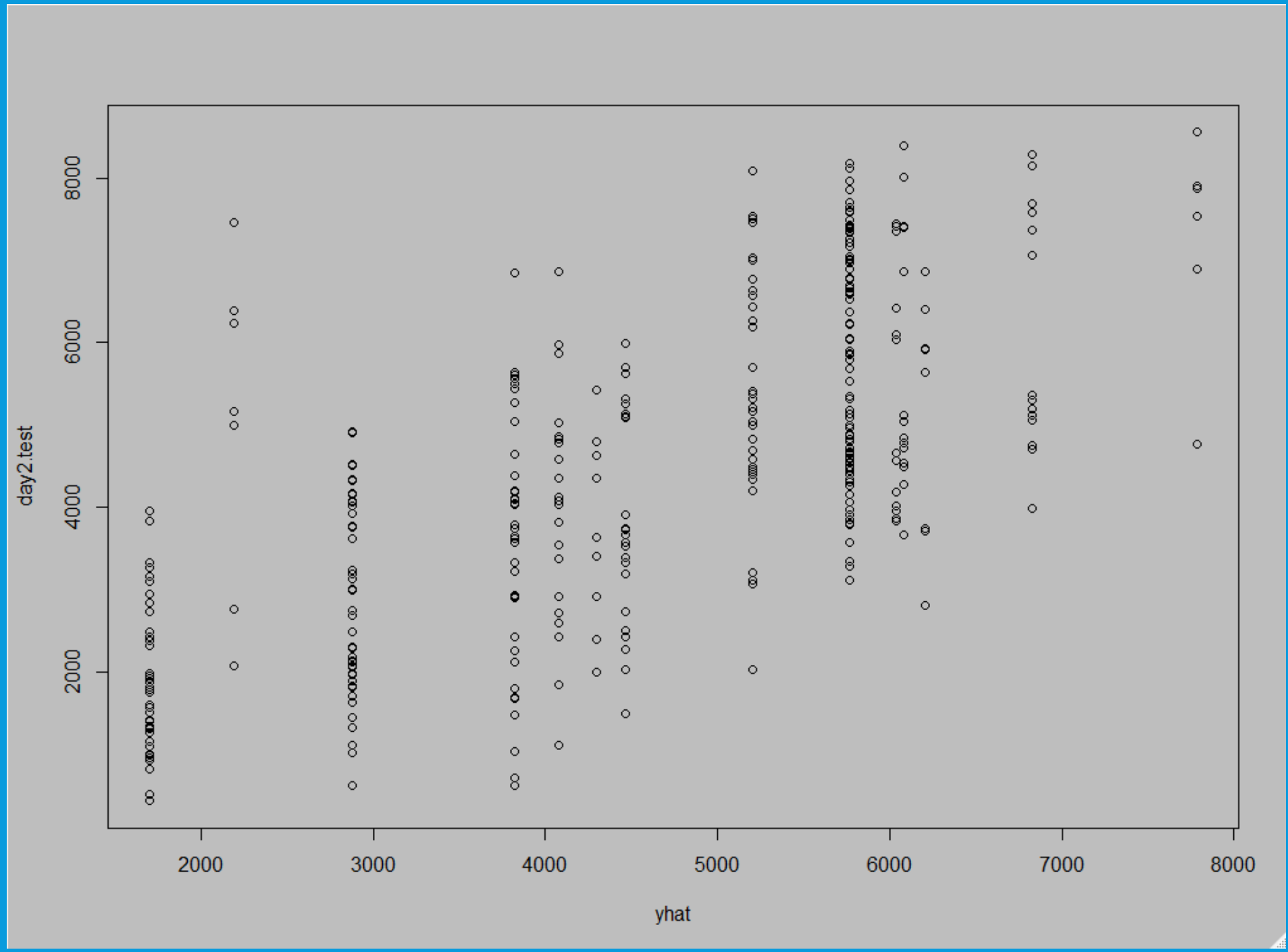
- we create a training set, and fit the tree to the training data
- We find the variables actually used

"tempdenorm“, "season“,"humiditydenorm“, "month“, "windspeeddenorm“

- Terminal Nodes : 14
- residual mean deviance: 1298000 = 455600000 / 351




cv.day2$dev
7.0e+08
8.0e+08
9.0e+08
1.0e+09
1.1e+09
1.2e+09
1.3e+09
2
4
6
8
10
12
14
cv.day2$size


day2.test
2000
4000
6000
8000
2000
3000
4000
5000
6000
7000
8000
yhat

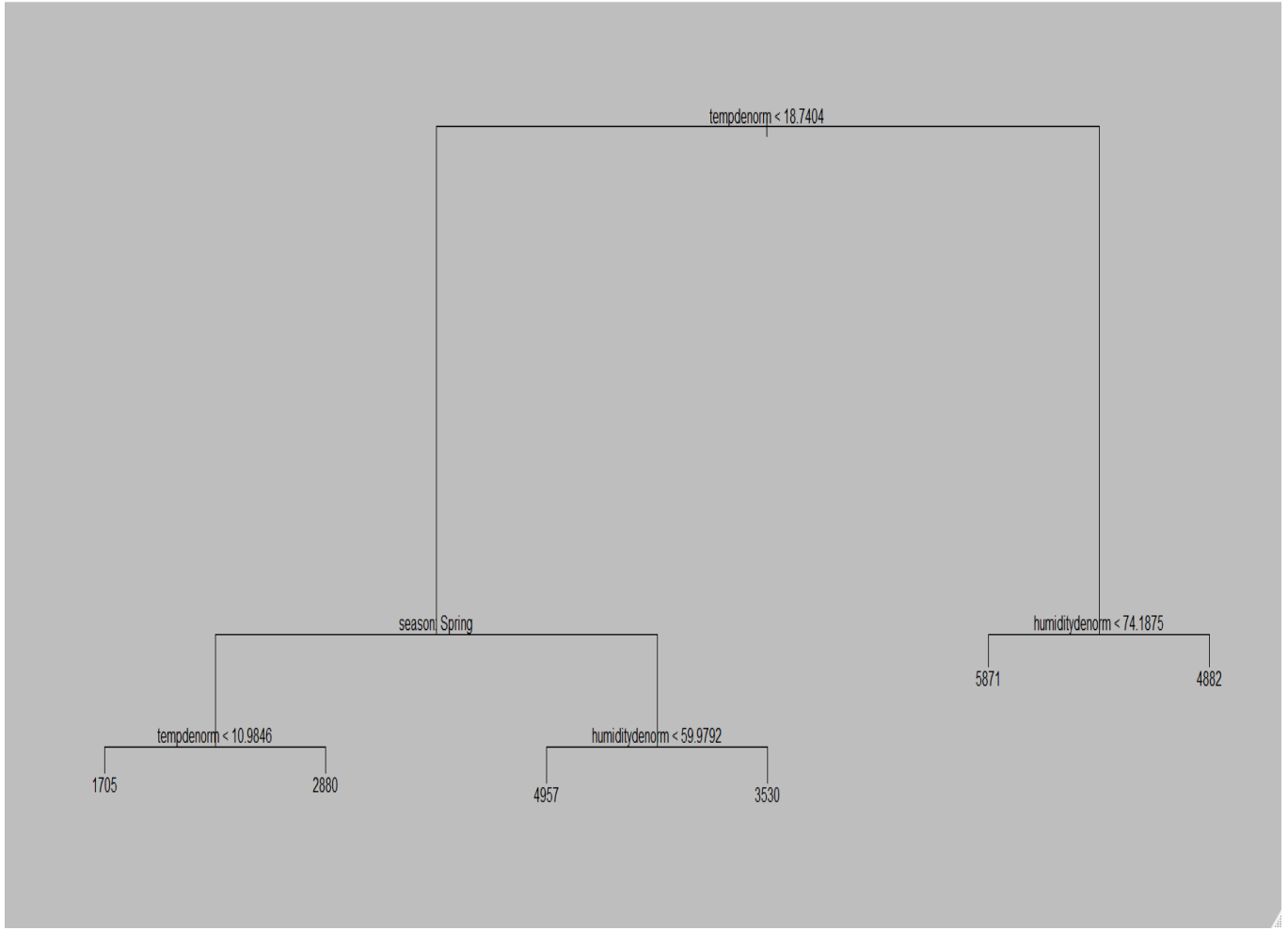
PRUNE TREE

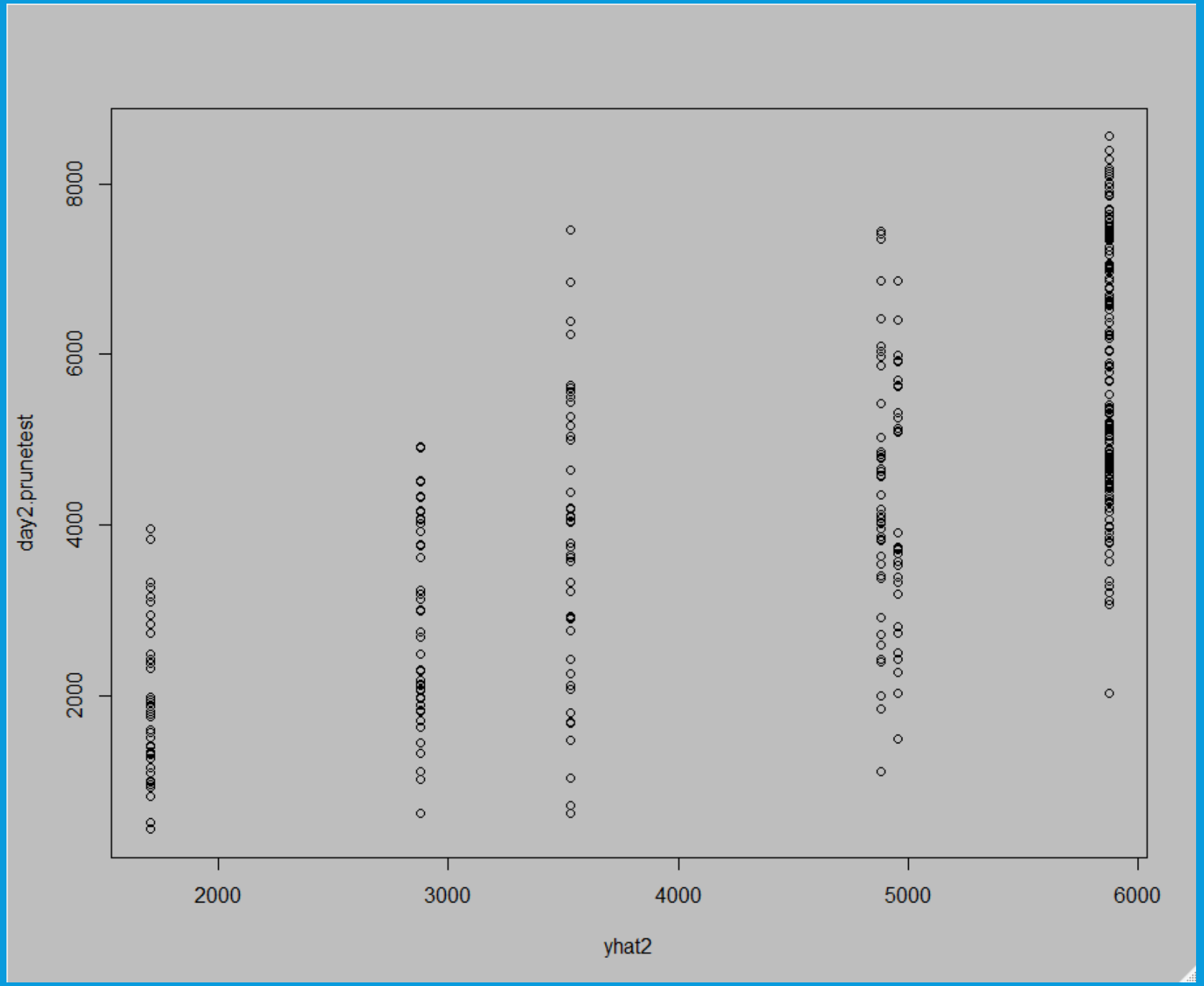
We pruned the tree (by doing cross validation on the training dataset to determine the best tree)

We use pruned tree to make prediction on the test data

we calculated the predicted value of DV on the test data using the trained model

We Calculated the Mean Squared Error to be 2052060




day2.prunetest
2000
4000
6000
8000
2000
3000
4000
5000
6000
yhat2

CONCLUSION

- Choice of model: Multi Linear Regression
  - Attained lower MSE
- To make the model stronger we should probably consider other variables
  - Such as hourly rentals

PRACTICAL IMPLICATIONS

- The major usage of this project would be to forecast bike rentals for the coming years based on the independent variables, 'Time Metrics' and 'Weather Records'.
- Also, this analysis can be used in helping a company interested in starting a bike sharing service in cities with similar weather conditions as D.C.
- China Bike Pile
  - There is a oversupply of bikes in China that ends up in bike graveyard
  - These bikes can be repurposed and used for bike sharing services.