DSO 530 – Bay Area Bike Share Program – Demand Forecasting Model

Project Team:


Xuan Huang


Chun Yang


Xiaoqiu Yu


Alok Abhishek


Yuzhou Dou

SFMTA Municipal Transportation Agency

BayAreaBikeShare

Agenda

- Executive Summary
- Data Description
- Process Overview
- Data Imputation and Outlier handling
- Clustering
- Statistical Models
- Results

Executive Summary

Objective

The Bay Area Bike Share enables quick, easy, and affordable bike trips around the San Francisco Bay Area. We plan to develop insights of bike usage patterns and predict the renting demand to help the company better allocate the bikes and optimize the operations. We hope to make a generic model so that it can then be adapted by bike share program in other cities in countries with final aim of helping companies optimize their operations and improve return on investment. This model can also serve a starting reference for infrastructure planning of autonomous cars.

Data Analysis Methods

Visualization of data to find correlations
Imputing data with average values to make up for missing data
Analyzing test error and validation error rate to find the best fit model

Conclusion

We found some results which were very different from our initial assumptions such as:

- Bike demand is higher on weekdays and during business hours Vs weekend or non business hours
- Bike demand is higher within downtown area Vs areas where educations institutes are

We found three main clusters of bike stations – Stations which were near commercial centers, Tourists spots or educational institutes.
Demand was highly correlated to whether the day was holiday or not, and also on the temperature and visibility of the day

BayAreaBikeShare

Data Description

2 years of bike share program operations including demand and supply provided by Bay are bike share program.
2 GB of data
4 data sets

Stations

Represents a station where users can pickup or return bikes.
7 variables

Status

Represents the number of bikes and docks available for given station and minute.
4 variables

Trip

Represents an individual bike trip.
11 variables

Weather

Represents the weather for a specific day and zip code in the bay area.
24 variables

Process Overview

- Data Discovery and initial analysis
- Went through complete data set to analyze all data and understand every feature
- Finalize imputation methods to cover for missing data

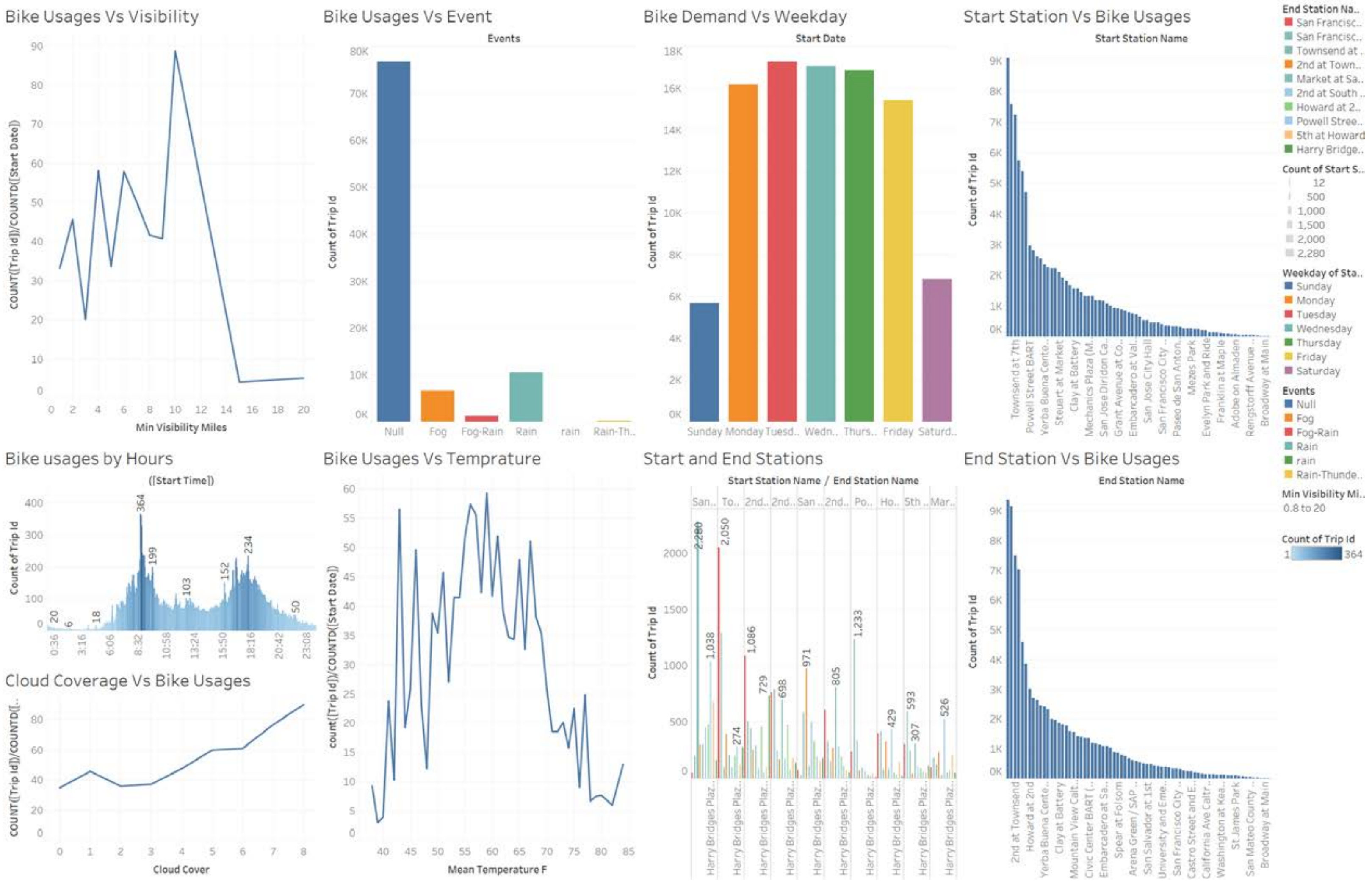
- Data Visualization and understanding
- Used Tableau to visualize relation in between different features and predictors to spot correlation and prediction capabilities

- Data correlation identification, clustering and model ideation

- Data Partition – Training, validation and Testing
- 70% for training
- 30% for testing

- Predictive Modeling
- Using lineal regression, LASSO, PRC, Random Forest and Decision Tree

Data Visualization



Data Imputation & Outlier Treatment

- Some of the weather data like – Temperature, Visibility, Cloud Coverage etc. was not available for all zip codes for everyday.
- Since the stations and ZIP codes were geographically close by and weather conditions do not vary too much across the area, we took average across all the zip codes to impute the missing values.
- For all predictors – we checked data set for outliers and removed them before modeling.

Variables	# of NA
max/mean/min_temperature_f	4
max/mean/min_dew_point_f	54
max/mean/min_humidity	54
max/mean/min_sea_level_pressure_inches	1
max/mean/min_visibility_miles	22
max/mean/min_wind_Speed_mph	1
wind_dir_degrees	1

Duration	
Min.	60
1st Qu.	347
Median	518
Mean	1019
3rd Qu.	748
Max.	17270400

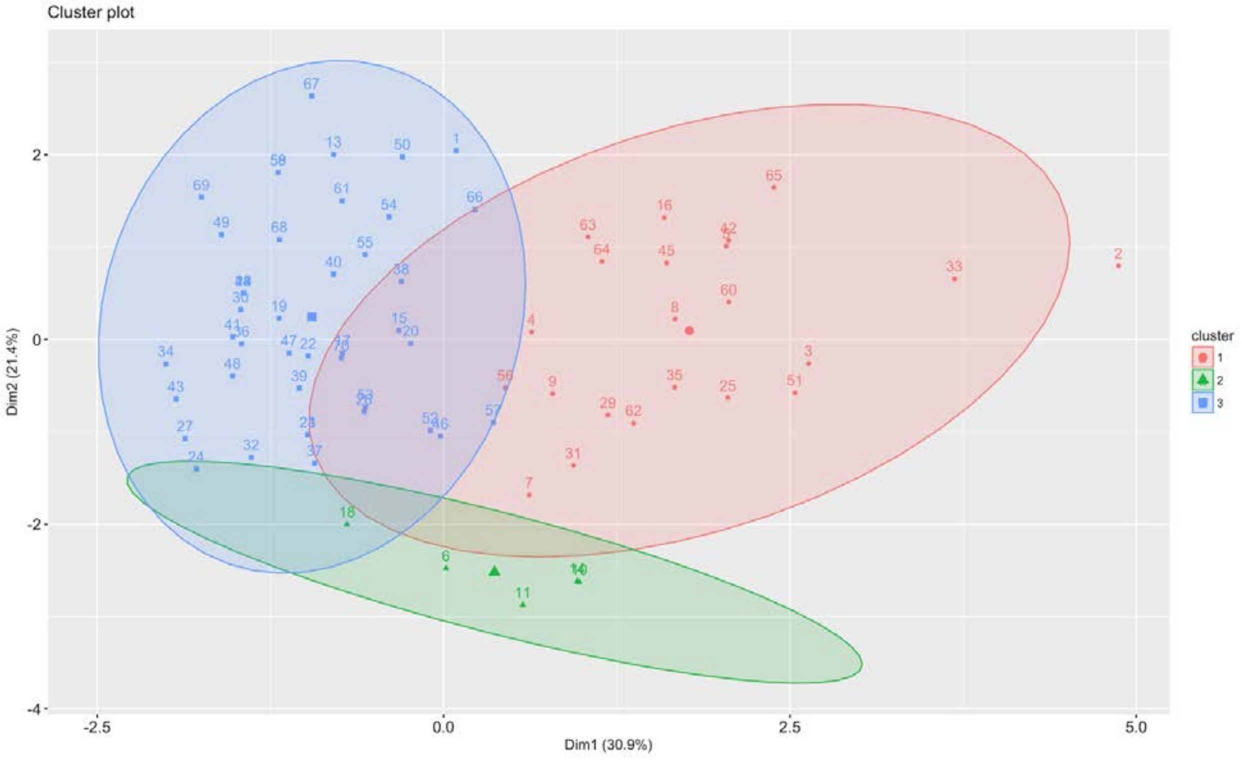
Filter

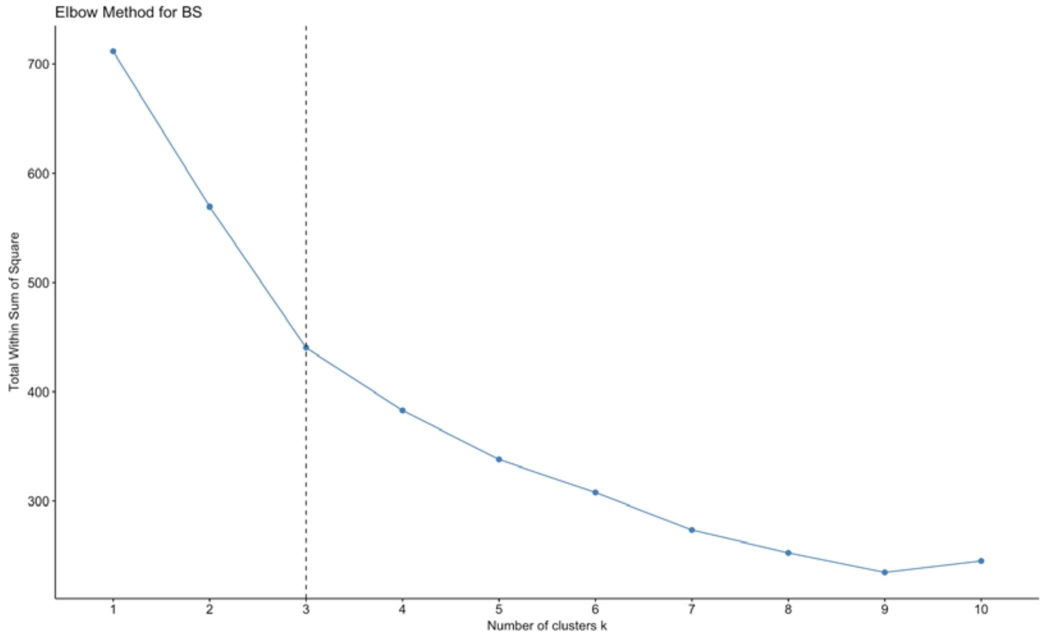
	id	duration	start_date	start_station_name	start_station_id	end_date	end_station_name
573567	568474	17270400	12/6/2014 21:59	South Van Ness at Market	66	6/24/2015 20:18	2nd at Folsom
382719	825850	2137000	6/28/2015 21:50	Market at Sansome	77	7/23/2015 15:27	Yerba Buena Center of the .
440340	750192	1852590	5/2/2015 6:17	San Antonio Shopping Center	31	5/23/2015 16:53	Castro Street and El Camin
371067	841176	1133540	7/10/2015 10:35	University and Emerson	35	7/23/2015 13:27	University and Emerson
80511	111309	722236	11/30/2013 13:29	University and Emerson	35	12/8/2013 22:06	University and Emerson
606064	522337	720454	10/30/2014 8:29	Redwood City Caltrain Station	22	11/7/2014 15:36	Stanford in Redwood City
223017	323594	716480	6/13/2014 16:57	Harry Bridges Plaza (Ferry Building)	50	6/21/2014 23:59	Civic Center BART (7th at M
195380	361321	715339	7/13/2014 5:50	Arena Green / SAP Center	14	7/21/2014 12:32	Adobe on Almaden
421840	774999	688899	5/20/2015 15:27	Palo Alto Caltrain Station	34	5/28/2015 14:49	California Ave Caltrain Stati
524522	635260	655939	2/8/2015 3:05	San Jose Civic Center	3	2/15/2015 17:17	SJSU 4th at San Carlos
287338	237942	644771	4/6/2014 3:37	South Van Ness at Market	66	4/13/2014 14:44	Clay at Battery

Showing 1 to 12 of 983,648 entries

Data Clustering

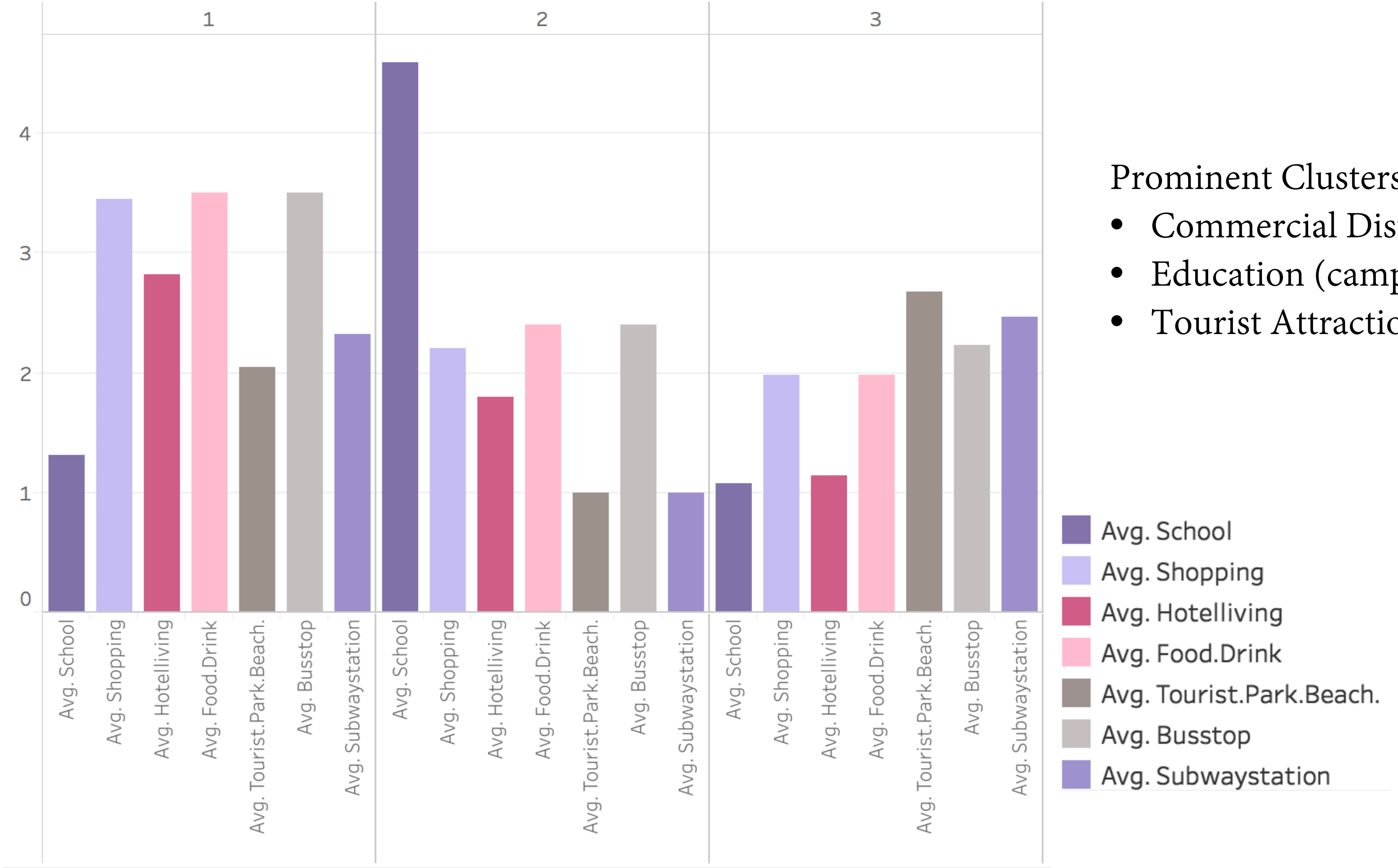
- Food & Drink - Count the restaurants and cafes within 200 meters
- School - Calculate the distance to the border of campus and library
- Hotel - Count the hotels within 200 meters
- Shopping - Count the shopping malls and supermarkets within 200 meters
- Tourist - Calculate the distance to the border of parks and beach
- Bus Stops - Count the bus stops within 200 meters
- Subway Station -Count the subway stations within 200 meters





BayAreaBikeShare

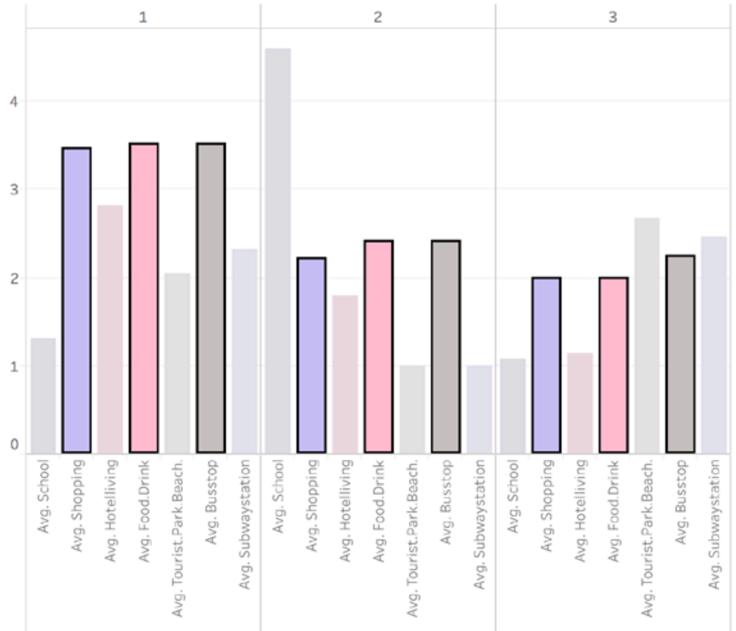
Cluster Analysis



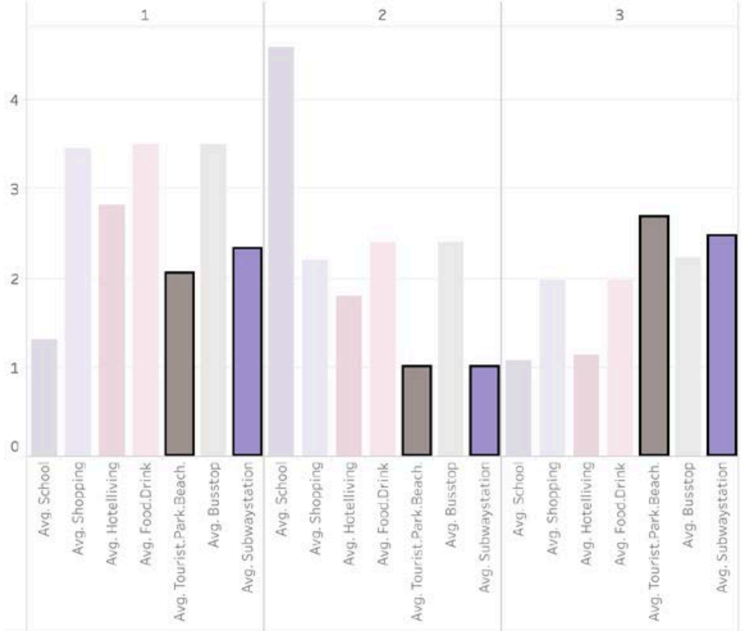
Prominent Clusters:

- Commercial Districts (office, shopping, restaurants)
- Education (campus, library)
- Tourist Attraction Spots (pikes, beaches)

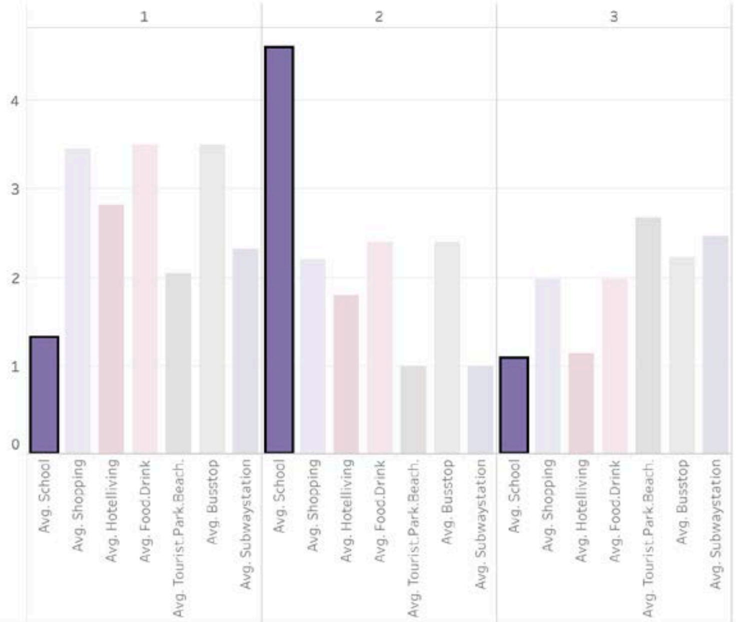
Three Types of Stations



Commercial Stations



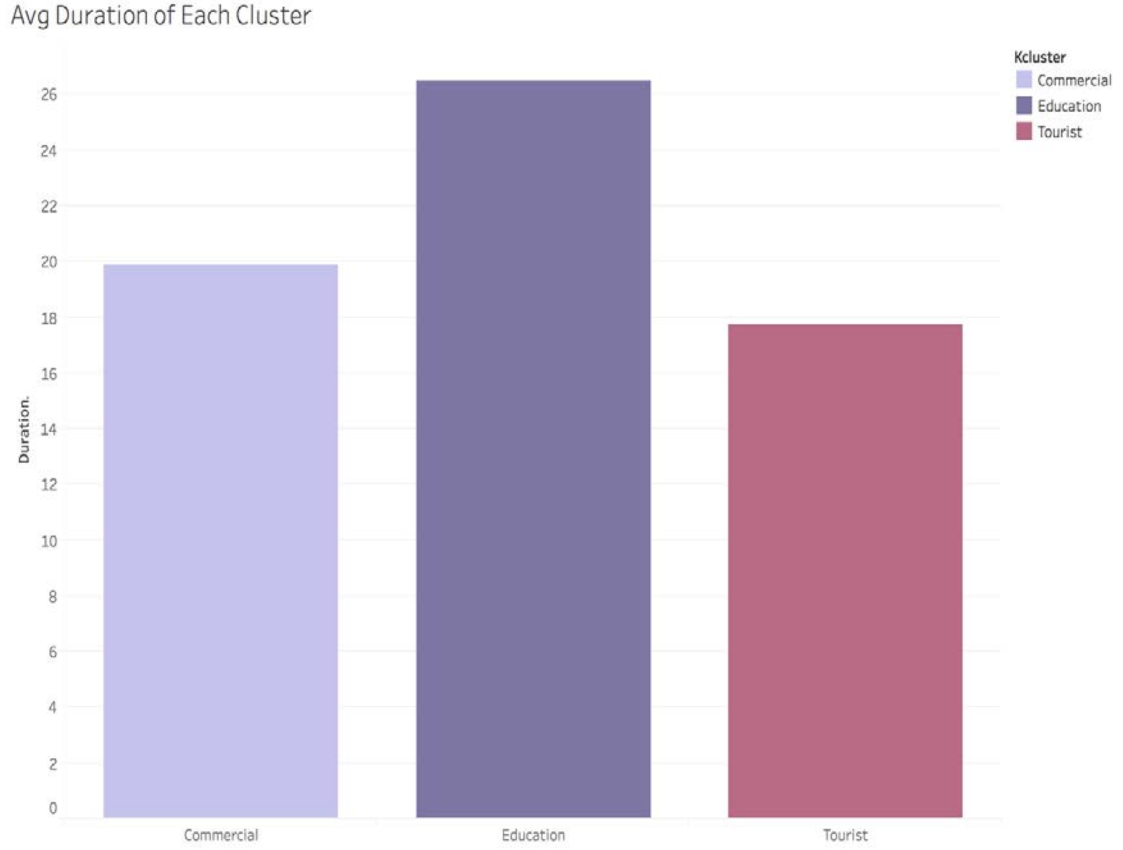
Tourist



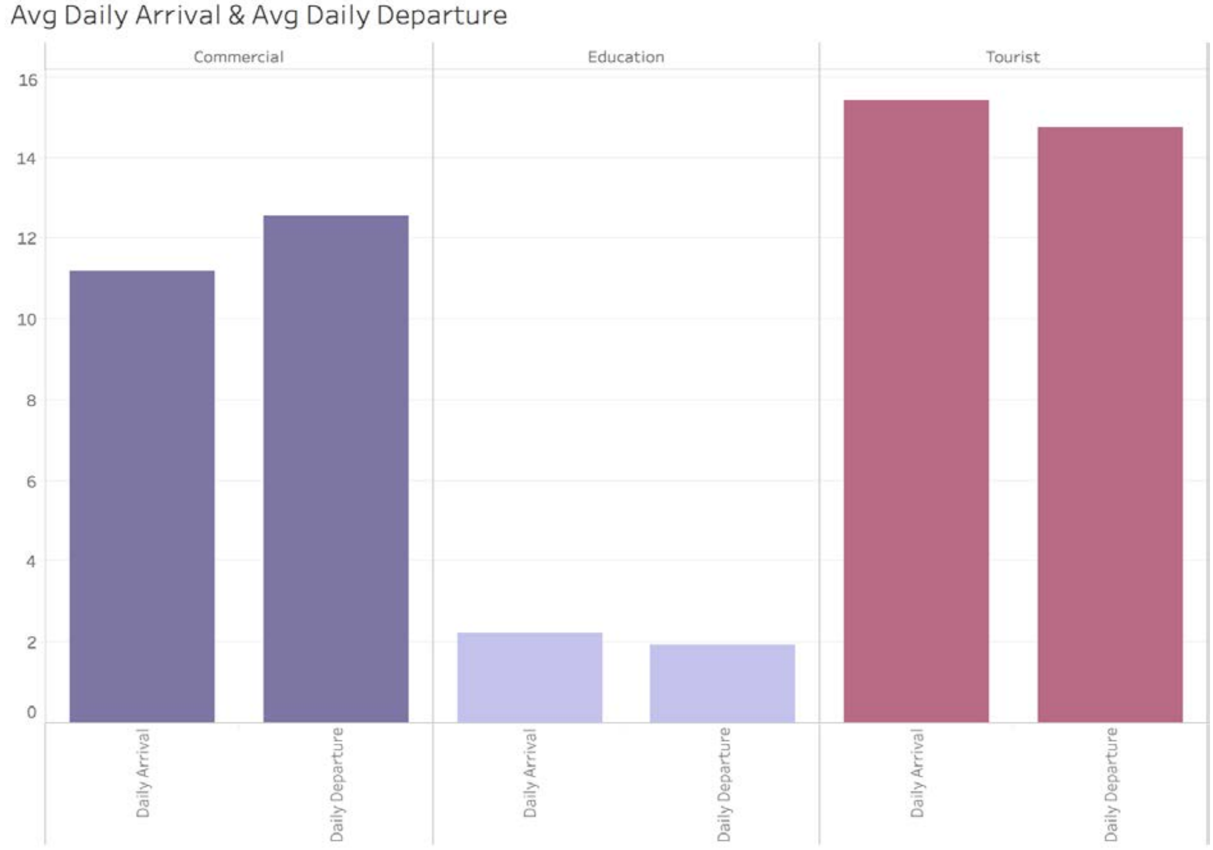
Educations

BayAreaBikeShare

Cluster Analysis

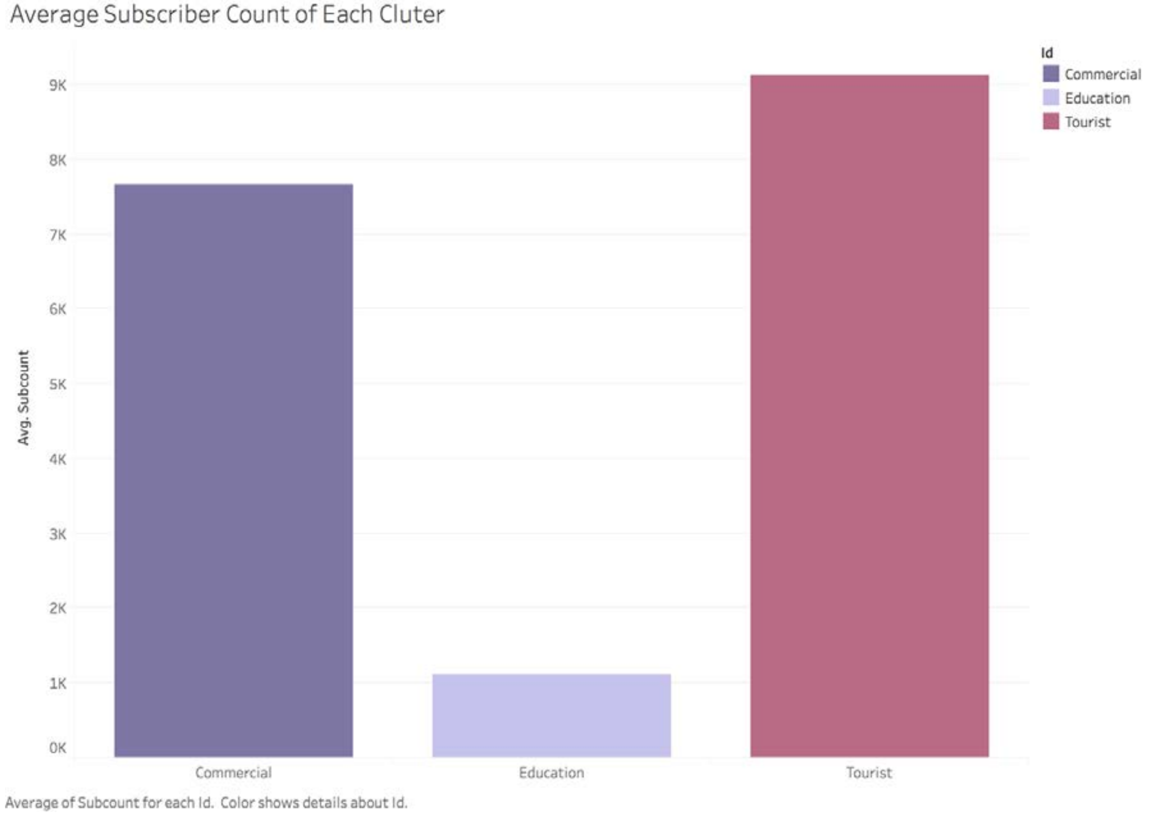


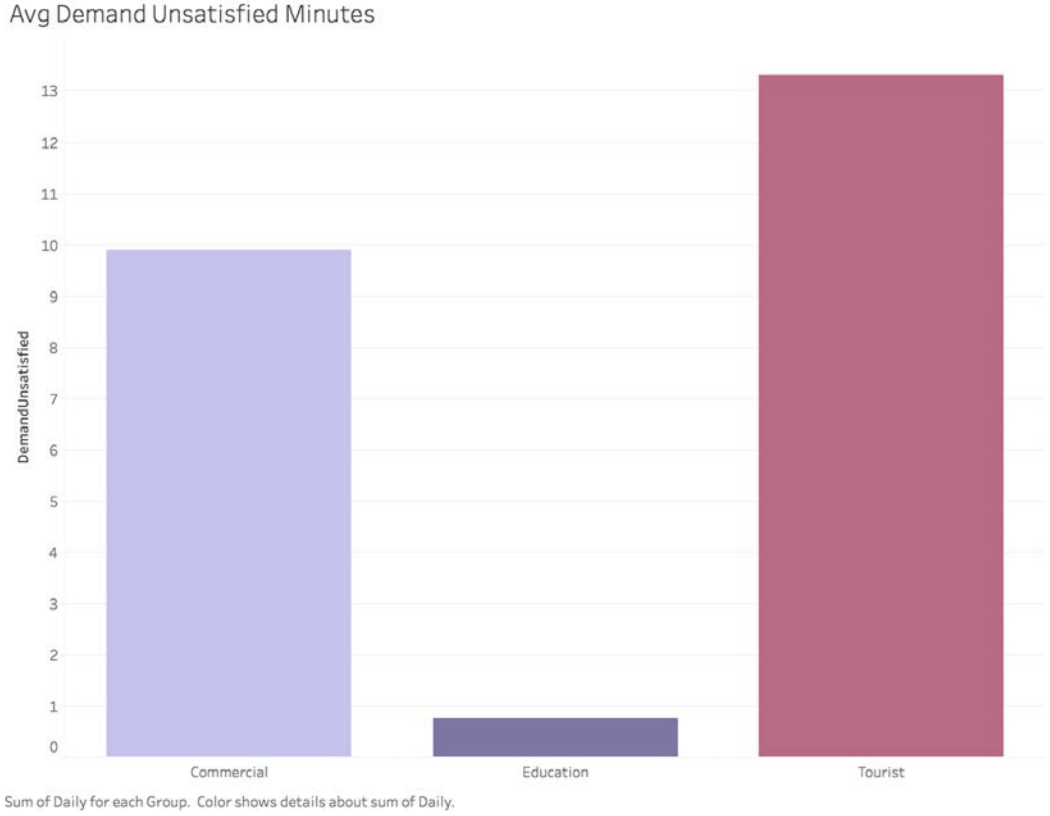
Sum of Duration. for each Kcluster. Color shows details about Kcluster.



Daily Arrival and Daily Departure for each Id. Color shows details about Id.

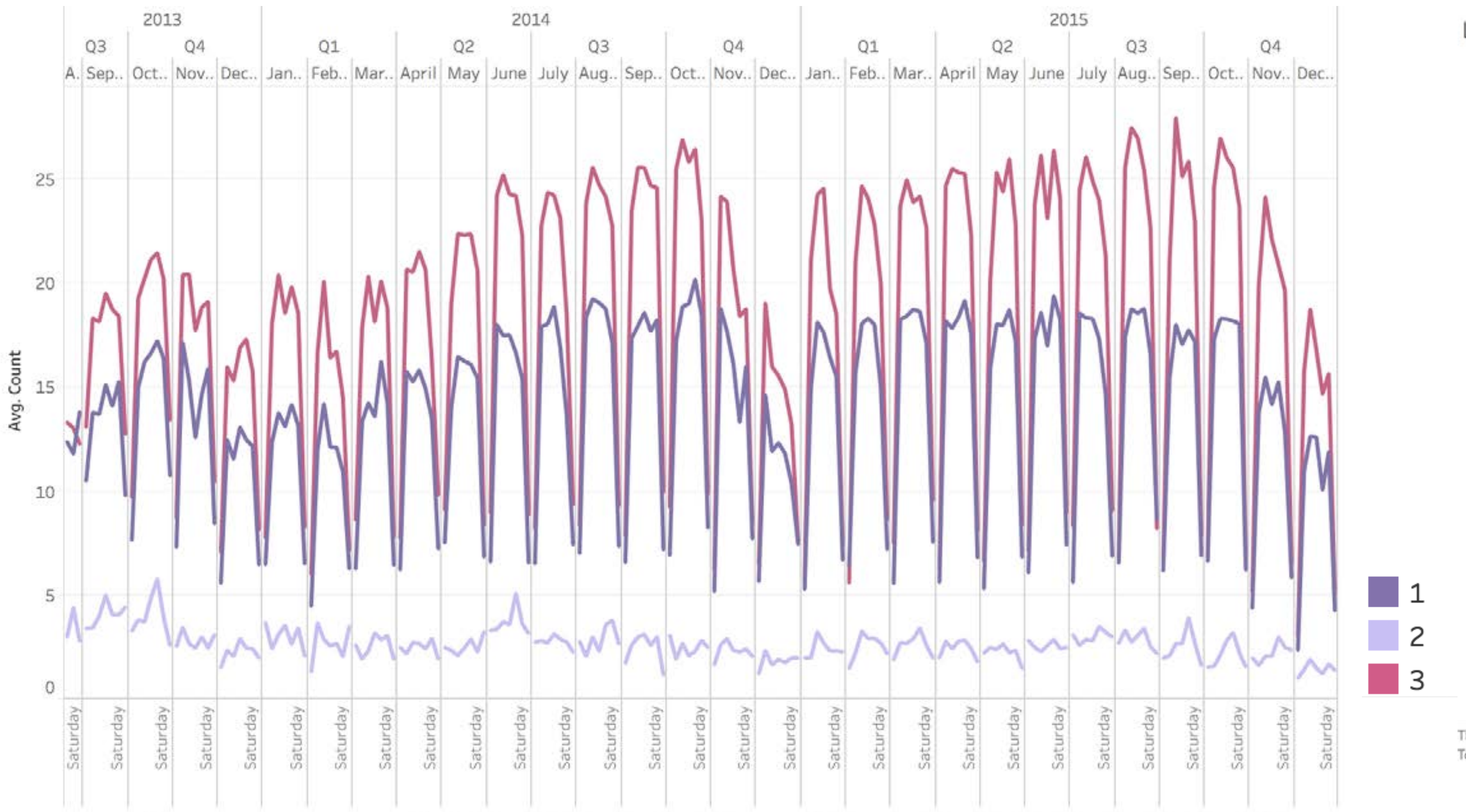
Cluster Analysis

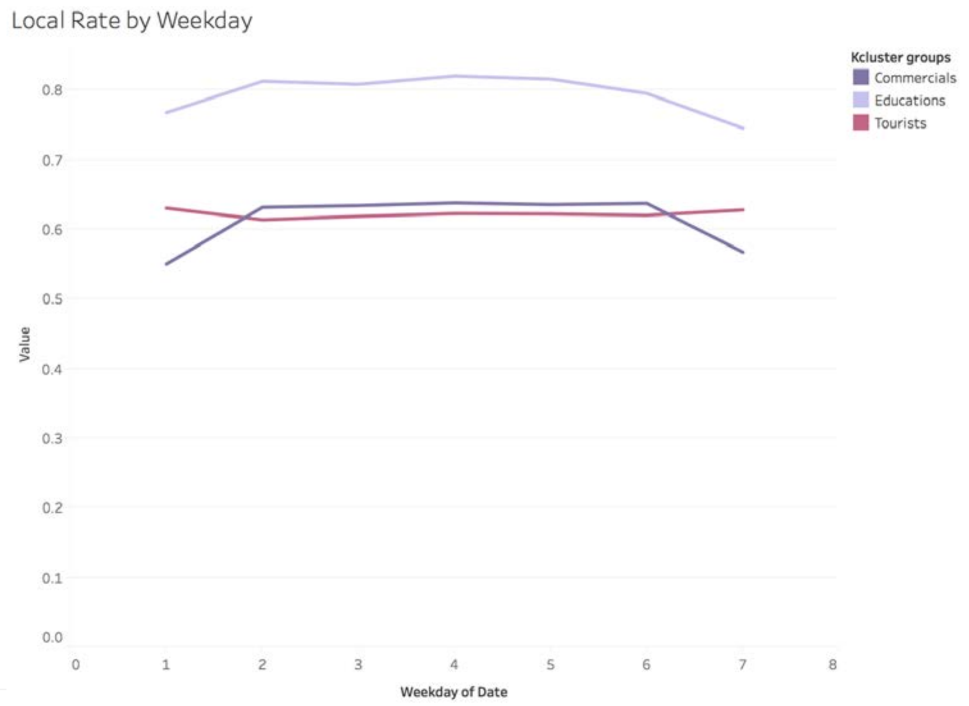




Cluster Analysis

Bike rental pattern for stations around shopping centers and Tourist spots showed similar patterns whereas bike rental pattern in areas close to school showed very low usages probably indicating students saw value in owning their own bikes and using skateboards to travel instead of using bike share program.

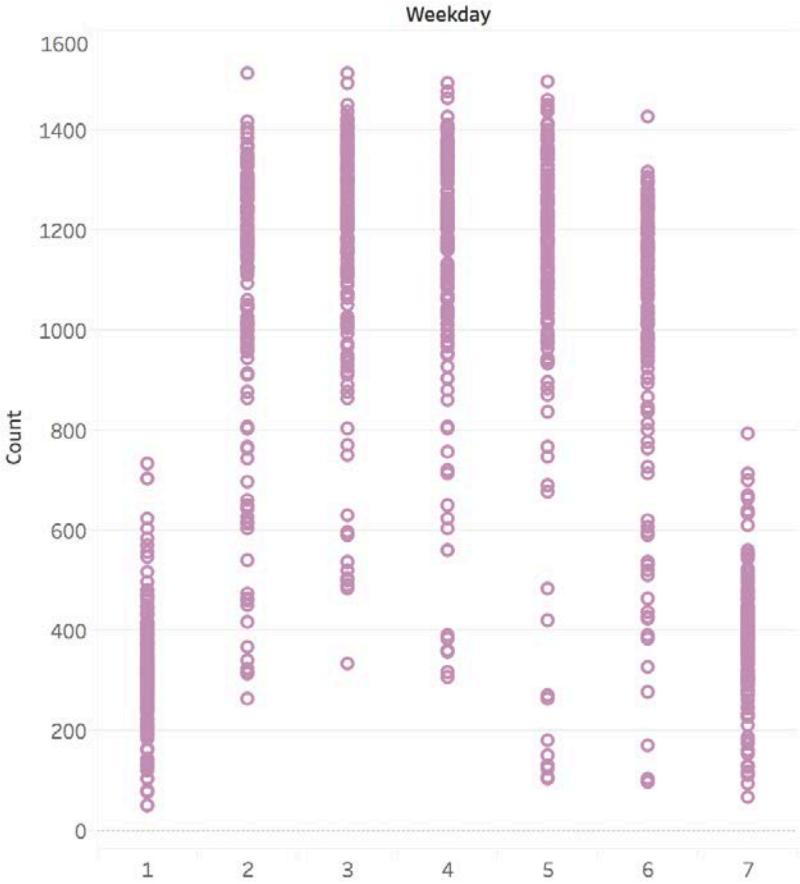


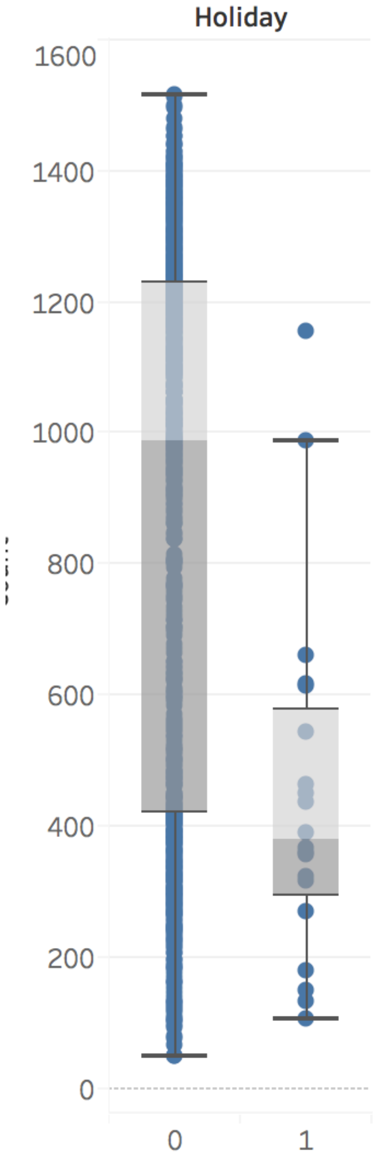


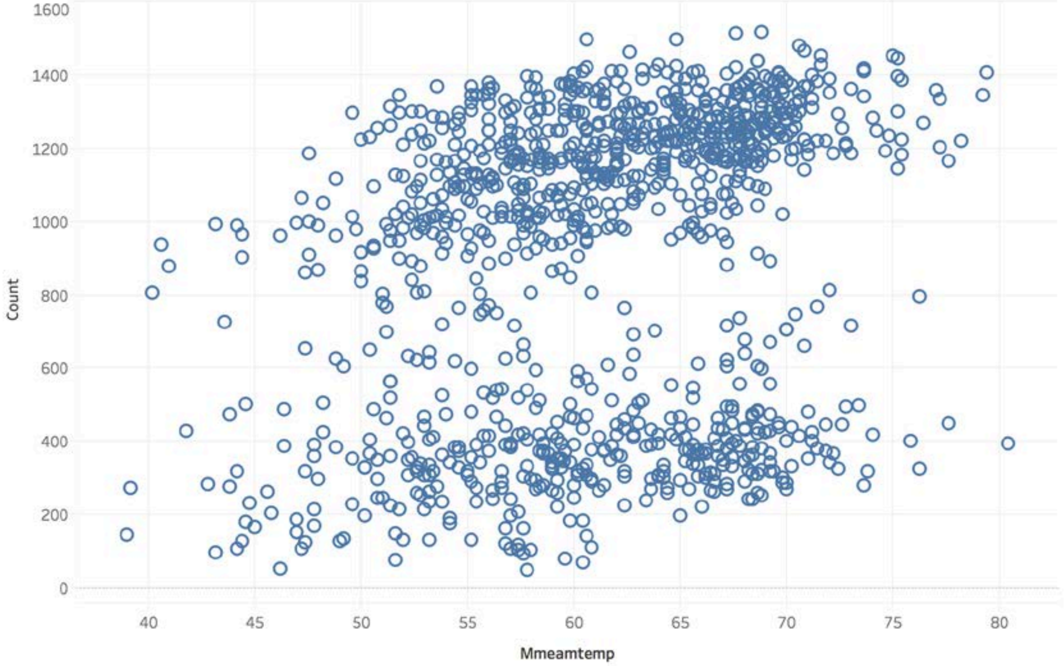
The trends of Commercials, Educations and Tourists for Date Weekday. Color shows details about Commercials, Educations and Tourists.

BayAreaBikeShare

Linear Regression







Linear Regression – Correlation matrix

	count	weekday	month	day	mmaxtemp	mmeamtemp	mmintemp	mmaxdew	mmeandew	mmindew
count	1.000	-0.001	-0.015	-0.021	0.262	0.259	0.211	0.167	0.163	0.155
weekday	-0.001	1.000	-0.009	0.014	0.006	0.007	0.007	0.020	0.022	0.025
month	-0.015	-0.009	1.000	0.012	0.104	0.111	0.103	0.123	0.093	0.074
day	-0.021	0.014	0.012	1.000	0.022	0.017	0.008	0.013	0.007	-0.005
mmaxtemp	0.262	0.006	0.104	0.022	1.000	0.933	0.698	0.658	0.554	0.450
mmeamtemp	0.259	0.007	0.111	0.017	0.933	1.000	0.908	0.813	0.759	0.677
mmintemp	0.211	0.007	0.103	0.008	0.698	0.908	1.000	0.855	0.868	0.828
mmaxdew	0.167	0.020	0.123	0.013	0.658	0.813	0.855	1.000	0.964	0.882
mmeandew	0.163	0.022	0.093	0.007	0.554	0.759	0.868	0.964	1.000	0.968
mmindew	0.155	0.025	0.074	-0.005	0.450	0.677	0.828	0.882	0.968	1.000
mmaxhum	-0.114	0.033	0.015	0.006	-0.283	-0.182	-0.037	0.340	0.392	0.385
mmeanhum	-0.133	0.028	0.005	-0.016	-0.417	-0.202	0.081	0.344	0.465	0.526
mminhum	-0.127	0.020	-0.031	-0.033	-0.464	-0.199	0.143	0.284	0.430	0.529
mmaxsea	-0.109	-0.038	-0.139	0.037	-0.471	-0.595	-0.639	-0.529	-0.534	-0.506
mmeansea	-0.079	-0.039	-0.140	0.043	-0.418	-0.542	-0.593	-0.498	-0.497	-0.468
mminsea	-0.056	-0.039	-0.139	0.047	-0.376	-0.488	-0.536	-0.459	-0.450	-0.418
maxvis	0.085	-0.019	0.011	0.040	0.195	0.244	0.257	0.195	0.210	0.207
mmeanvis	0.153	-0.003	0.016	0.035	0.331	0.285	0.180	0.014	0.002	0.003
mminvis	0.164	-0.025	0.039	0.020	0.387	0.294	0.135	-0.061	-0.074	-0.070
mmaxwind	0.021	-0.008	-0.066	0.048	0.125	0.264	0.377	0.225	0.232	0.213
mmeanwind	0.017	-0.022	-0.097	0.030	0.068	0.290	0.490	0.262	0.315	0.328
mmaxgust	-0.037	-0.021	-0.001	0.022	0.029	0.155	0.273	0.145	0.148	0.128
cloud	-0.127	-0.006	-0.120	-0.058	-0.376	-0.107	0.223	0.236	0.325	0.370

	mmaxhum	mmeanhum	mminhum	mmaxsea	mmeansea	mminsea	maxvis	mmeanvis	mminvis	mmaxwind
count	-0.114	-0.133	-0.127	-0.109	-0.079	-0.056	0.085	0.153	0.164	0.021
weekday	0.033	0.028	0.020	-0.038	-0.039	-0.039	-0.019	-0.003	-0.025	-0.008
month	0.015	0.005	-0.031	-0.139	-0.140	-0.139	0.011	0.016	0.039	-0.066
day	0.006	-0.016	-0.033	0.037	0.043	0.047	0.040	0.035	0.020	0.048
mmaxtemp	-0.283	-0.417	-0.464	-0.471	-0.418	-0.376	0.195	0.331	0.387	0.125
mmeamtemp	-0.182	-0.202	-0.199	-0.595	-0.542	-0.488	0.244	0.285	0.294	0.264
mmintemp	-0.037	0.081	0.143	-0.639	-0.593	-0.536	0.257	0.180	0.135	0.377
mmaxdew	0.340	0.344	0.284	-0.529	-0.498	-0.459	0.195	0.014	-0.061	0.225
mmeandew	0.392	0.465	0.430	-0.534	-0.497	-0.450	0.210	0.002	-0.074	0.232
mmindew	0.385	0.526	0.529	-0.506	-0.468	-0.418	0.207	0.003	-0.070	0.213
mmaxhum	1.000	0.861	0.649	0.061	0.034	0.011	-0.054	-0.426	-0.548	-0.107
mmeanhum	0.861	1.000	0.935	0.007	-0.011	-0.017	-0.045	-0.455	-0.578	-0.034
mminhum	0.649	0.935	1.000	-0.027	-0.039	-0.035	-0.026	-0.403	-0.519	0.046
mmaxsea	0.061	0.007	-0.027	1.000	0.982	0.933	-0.203	-0.136	-0.077	-0.378
mmeansea	0.034	-0.011	-0.039	0.982	1.000	0.980	-0.186	-0.097	-0.023	-0.387
mminsea	0.011	-0.017	-0.035	0.933	0.980	1.000	-0.159	-0.062	0.018	-0.378
maxvis	-0.054	-0.045	-0.026	-0.203	-0.186	-0.159	1.000	0.488	0.165	0.135
mmeanvis	-0.426	-0.455	-0.403	-0.136	-0.097	-0.062	0.488	1.000	0.822	0.150
mminvis	-0.548	-0.578	-0.519	-0.077	-0.023	0.018	0.165	0.822	1.000	0.070
mmaxwind	-0.107	-0.034	0.046	-0.378	-0.387	-0.378	0.135	0.150	0.070	1.000
mmeanwind	-0.116	0.041	0.172	-0.432	-0.442	-0.425	0.181	0.157	0.065	0.778
mmaxgust	-0.059	-0.004	0.052	-0.340	-0.362	-0.365	0.119	0.080	-0.023	0.727
cloud	0.412	0.653	0.736	-0.104	-0.129	-0.133	-0.007	-0.348	-0.493	0.156

	mmeanwind	mmaxgust	cloud
count	0.017	-0.037	-0.127
weekday	-0.022	-0.021	-0.006
month	-0.097	-0.001	-0.120
day	0.030	0.022	-0.058
mmaxtemp	0.068	0.029	-0.376
mmeamtemp	0.290	0.155	-0.107
mmintemp	0.490	0.273	0.223
mmaxdew	0.262	0.145	0.236
mmeandew	0.315	0.148	0.325
mmindew	0.328	0.128	0.370
mmaxhum	-0.116	-0.059	0.412
mmeanhum	0.041	-0.004	0.653
mminhum	0.172	0.052	0.736
mmaxsea	-0.432	-0.340	-0.104
mmeansea	-0.442	-0.362	-0.129
mminsea	-0.425	-0.365	-0.133
maxvis	0.181	0.119	-0.007
mmeanvis	0.157	0.080	-0.348
mminvis	0.065	-0.023	-0.493
mmaxwind	0.778	0.727	0.156
mmeanwind	1.000	0.678	0.308
mmaxgust	0.678	1.000	0.180
cloud	0.308	0.180	1.000

Linear Regression

Full Model

```
lm.fit=lm(count~.-date-duration-sub, data=data[train,])

vif(lm.fit)
   weekday       month         day       event     weekend     holiday    mmaxtemp
  1.026024    1.202142    1.044923    1.846574    1.046579    1.059185  462.925027
 mmeamtemp    mmintemp    mmaxdew    mmeandew     mmindew     mmaxhum    mmeanhum
1295.424991  318.323474  47.956358  167.751547   44.492718   17.693517   95.831976
   mminhum    mmaxsea    mmeansea     mminsea      maxvis    mmeanvis     mminvis
 42.679265   75.021175  218.311572   58.703600    1.827680    5.878593    6.602745
  mmaxwind   mmeanwind    mmaxgust       cloud
  3.302532    4.703713    2.623314    3.536410
```

Smaller Model

```
lm.fit=lm(count~weekday + event + weekend + holiday + mmeamtemp +
mmeandew + mmeanhum + mmeansea + mminvis + mmeanwind + mmaxgust +
cloud, data[train, ])

vif(lm.fit)
 weekday     event   weekend   holiday mmeamtemp  mmeandew  mmeanhum  mmeansea
1.013967  1.699207  1.024605  1.026191 34.534470 44.657349 22.580461  1.682451
 mminvis mmeanwind  mmaxgust     cloud
2.298920  2.493138  2.012138  2.232380
```

Linear Regression

Smaller Model

```
lm.fit=lm(count~ weekday +event + weekend + holiday + mmeamtemp +
mmeanhum + mmeansea + mminvis + mmeanwind + mmaxgust + cloud,
data[train, ])
```

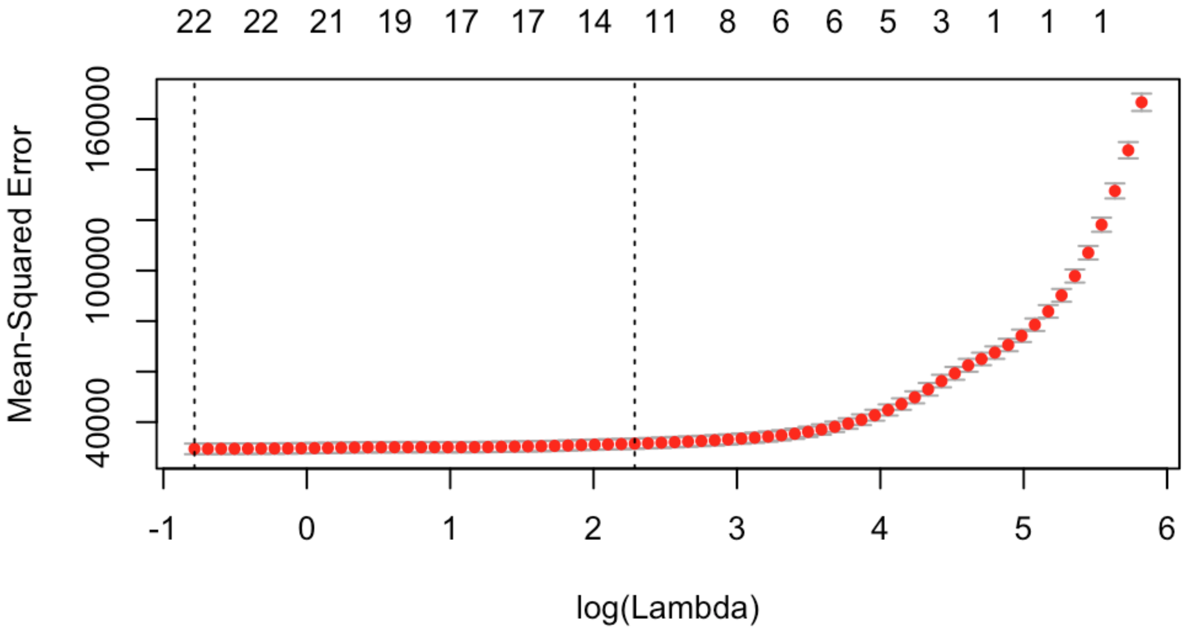
```
vif(lm.fit)
weekday     event    weekend   holiday mmeamtemp  mmeandew  mmeanhum  mmeansea
1.013967  1.699207  1.024605  1.026191 34.534470 44.657349 22.580461  1.682451
 mminvis mmeanwind  mmaxgust     cloud
2.298920  2.493138  2.012138  2.232380
```

Final Model

```
lm.fit=lm(count~ event + weekend + holiday + mmeamtemp + mminvis +
mmaxgust + cloud, data[train, ])
```

```
Residual standard error: 174.1 on 761 degrees of freedom Multiple R-squared: 0.82, Adjusted R-
squared: 0.8184  F-statistic: 495.4 on 7 and 761 DF,  p-value: < 2.2e-16
> vif(lm.fit)
 event   weekend   holiday mmeamtemp   mminvis  mmaxgust     cloud
1.688162  1.014836  1.023197  1.161850  1.799213  1.071834  1.454505
MSE: 185.6062
```

LASSO



bestlam
[1] 0.4569004

Variables dropped:
Mean temperature
Mean sea level pressure

Model Results :
MSE: 181.2726

Best Subset

Best subset selection—all 27 variables included

1: weekend

2: weekend mmaxtemp

3: weekend holiday mmaxtemp

4: weekend holiday mmeantemp mminvis….

Based on r^2, 25 variables are selected

Based on adjusted r^2, 18 variables are selected

Fit the model again using 18 variables, the test MSE is 182.5887

PCR

Smallest CV happens when the model has 24 components and has explained 100 %

VALIDATION: RMSEP

Cross-validated using 10 random segments.

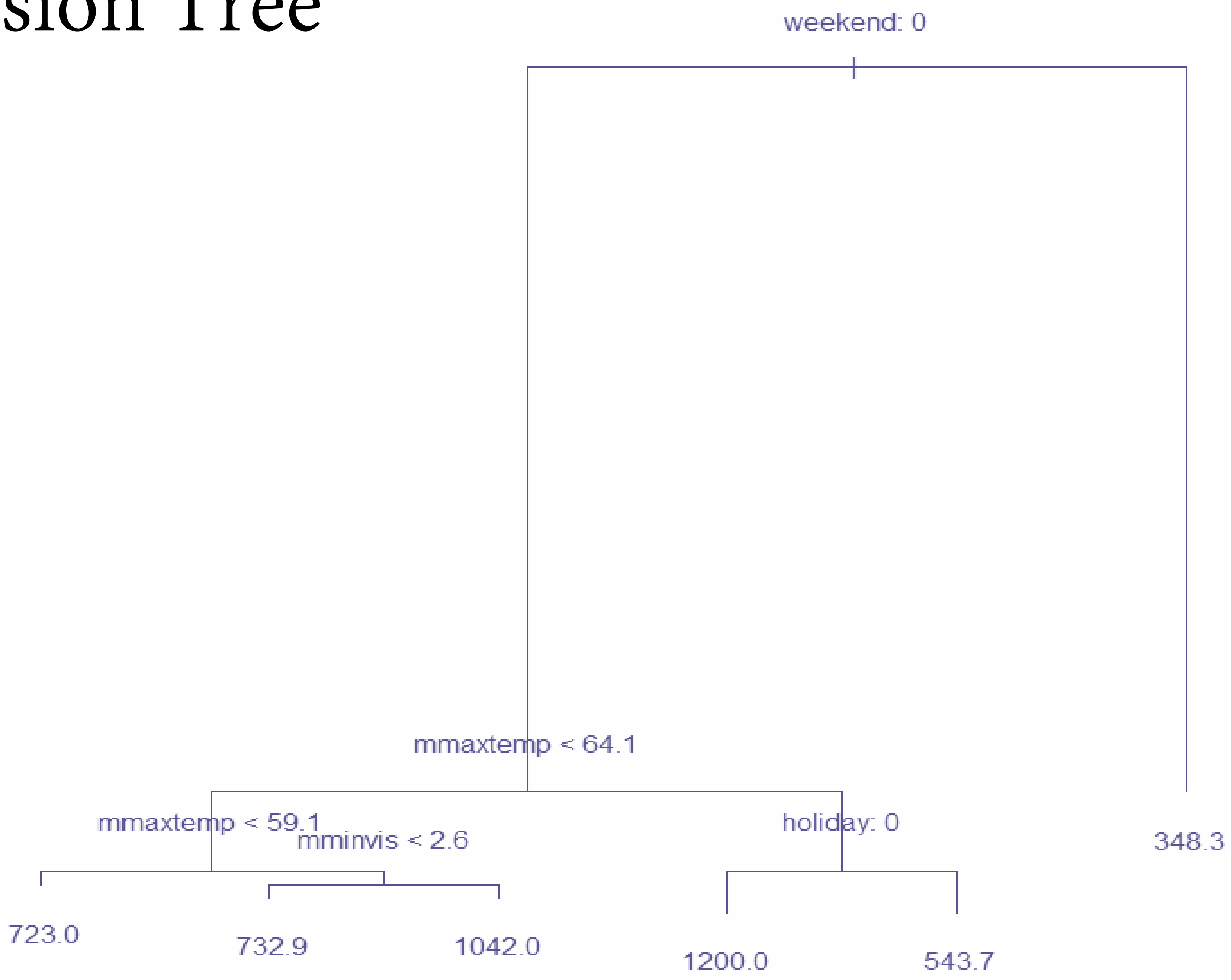
```
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
CV           408.7    403.2    394.6    393.1    393.5    360.2    363.9    339.5    328.3
adjCV        408.7    403.2    394.5    393.0    393.4    359.2    363.7    338.6    330.1
       9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps  16 comps  17 comps
CV       311.9     197.3     182.3     175.7     175.3     175.8     176.2     176.2     175.8
adjCV    311.7     191.0     181.9     175.4     175.0     175.5     175.9     175.8     175.5
       18 comps  19 comps  20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
CV        175.1     175.4     175.8     174.0     173.4     173.3     173.3     173.5
adjCV     174.7     175.0     175.4     173.5     173.0     172.8     172.9     173.1
```

TRAINING: % variance explained

```
       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps
X       27.878   47.987   57.922   64.323    69.04    73.42    77.54    81.43    85.13     88.62
count    2.735    7.268    8.267    8.336    22.69    23.35    34.66    40.67    48.97     79.88
       11 comps  12 comps  13 comps  14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
X         91.87     94.28     96.21     97.26     98.15     98.89     99.23     99.46     99.66
count     81.19     82.54     82.62     82.63     82.63     82.66     82.82     83.01     83.03
       20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
X         99.83     99.94     99.97     99.99    100.00    100.00
count     83.03     83.41     83.58     83.63     83.63     83.63
```

The test error is 182.2452

Decision Tree

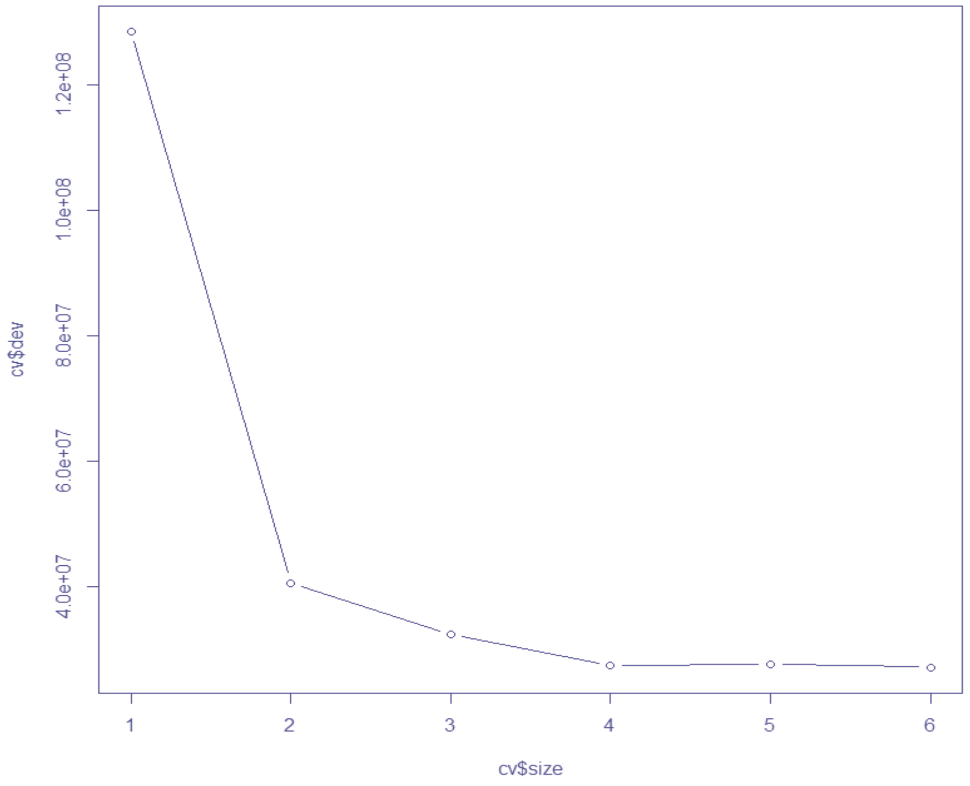


The simple decision tree includes only 6 terminal node. It splits on weekend mmaxtemp holiday and mminvis

Decision Tree

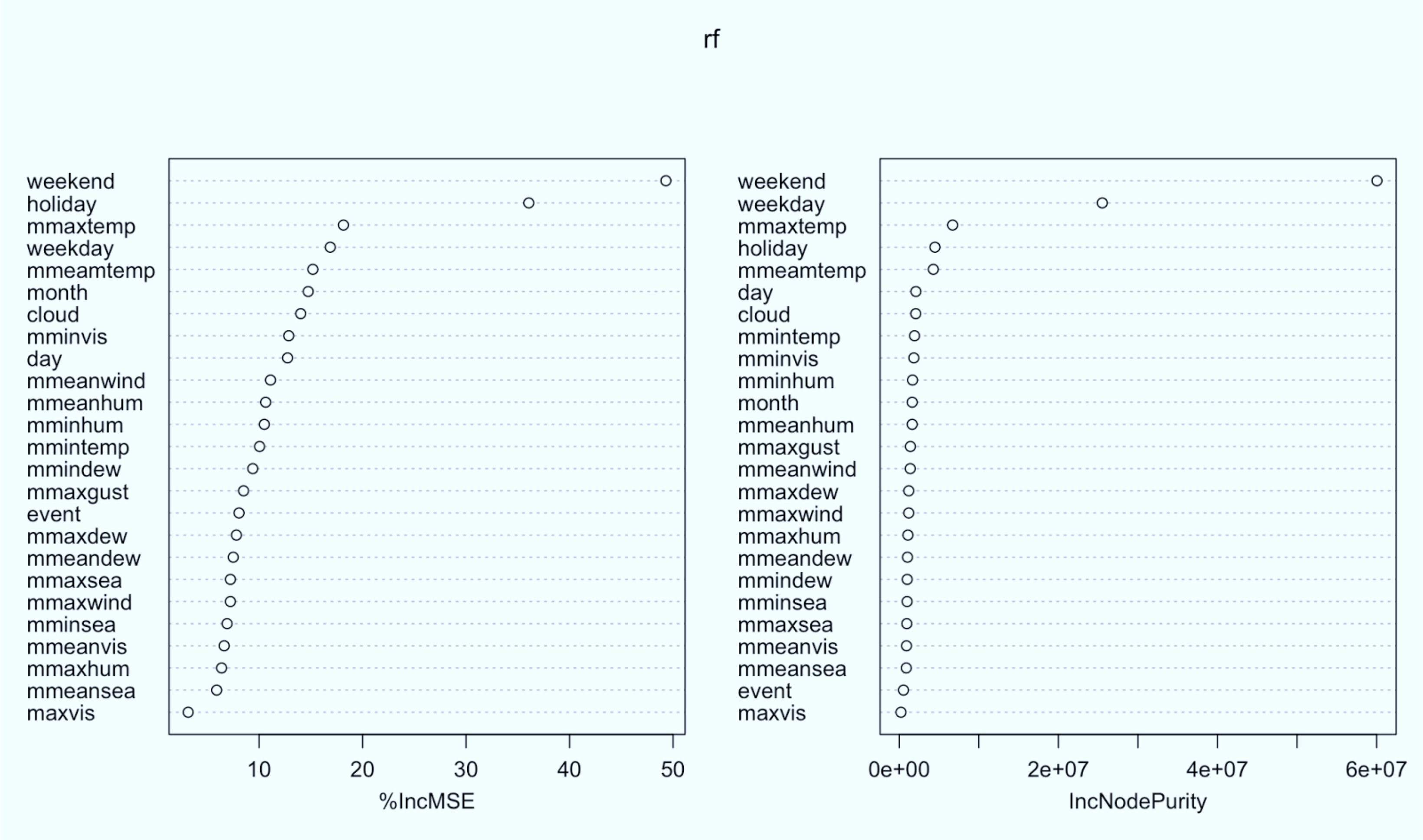
Based on the deviance, there is no need to prune the tree

The test error is 189.8095



Random Forest

Weekend, holiday, temperature are the most important variables to predict the demand.



Model Results :
MSE: 159.9379

Statistical Models

Linear Regression

MSE – 185.6062

LASSO

MSE – 181.2726

λ - 0.4569004

Best Selection

MSE – 182.5887

Variables# 18

Random Forest

MSE - 159.9379

Decision Tree

MSE – 189.8095

Terminal Nodes# 6

PCR

MSE – 182.2452

Components# 24

Important Predictors

Weekends
Weekday and Weekends have very different bike demand patterns

Holidays
Bike demand pattern is very different for holidays Vs non Holidays

Temperature
Both extremely high temperature and low temperature negatively impacts bike demands
Temperature in between 55^{o} F to 65^{o} F is the sweet spot for high bike usages.

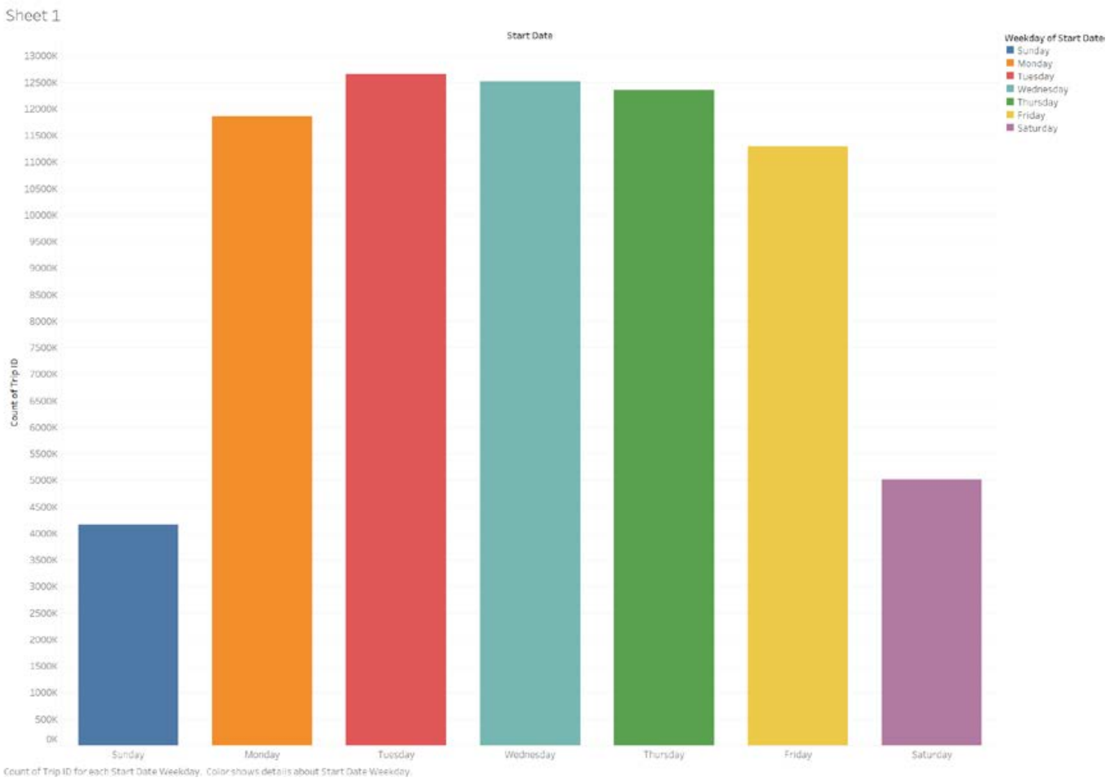
Q&A

BayAreaBikeShare

Appendix

Data Visualization

Bike usages pattern on different weekdays

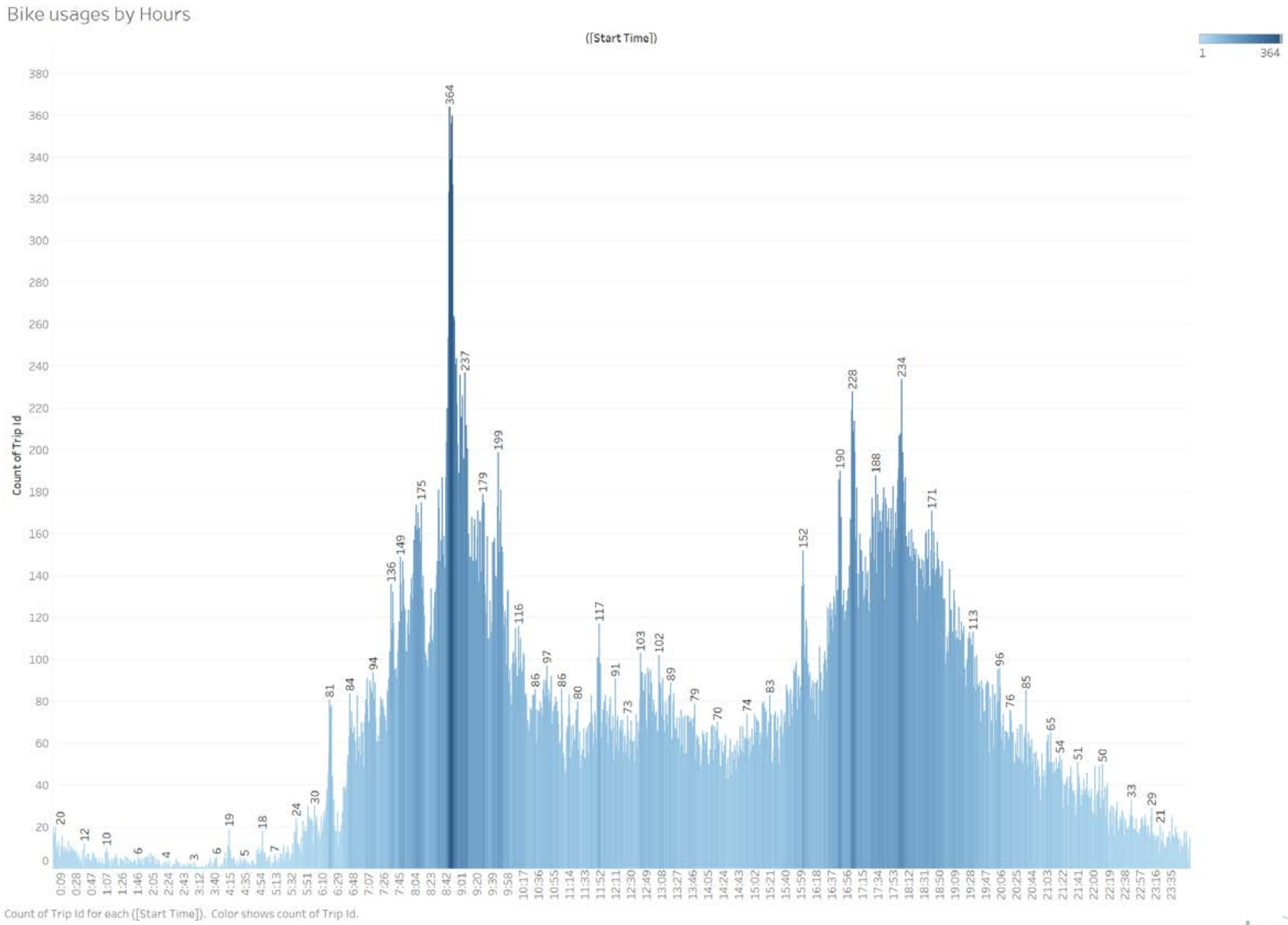


Bike demand on weekends in significantly lower than bike demand on weekdays

BayAreaBikeShare

Data Visualization

Bike demand by hours of the day

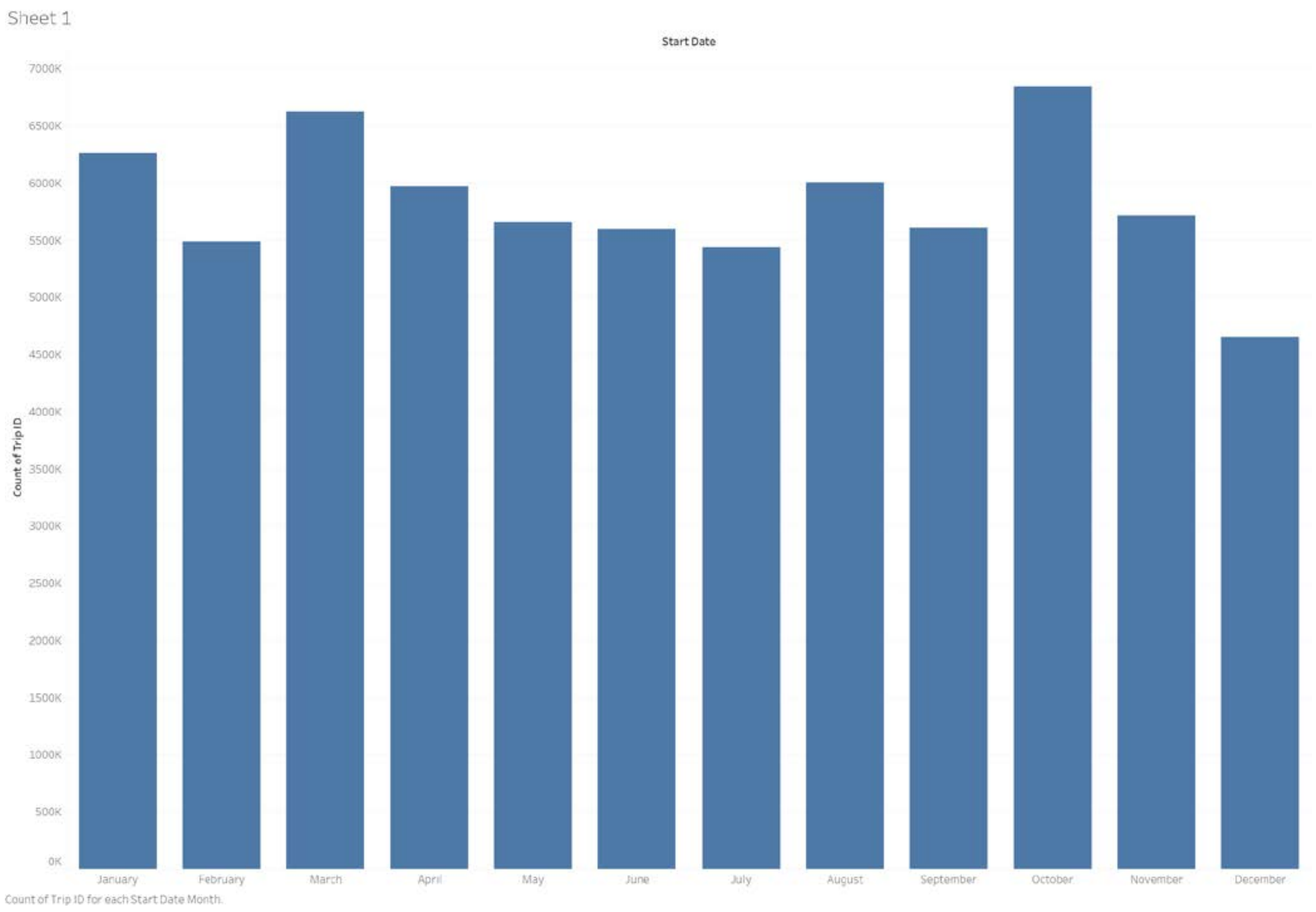


Bi-Model usages pattern with two spikes:

1. One at 9 am in the morning and
2. Second at 6 pm in the evening

Data Visualization

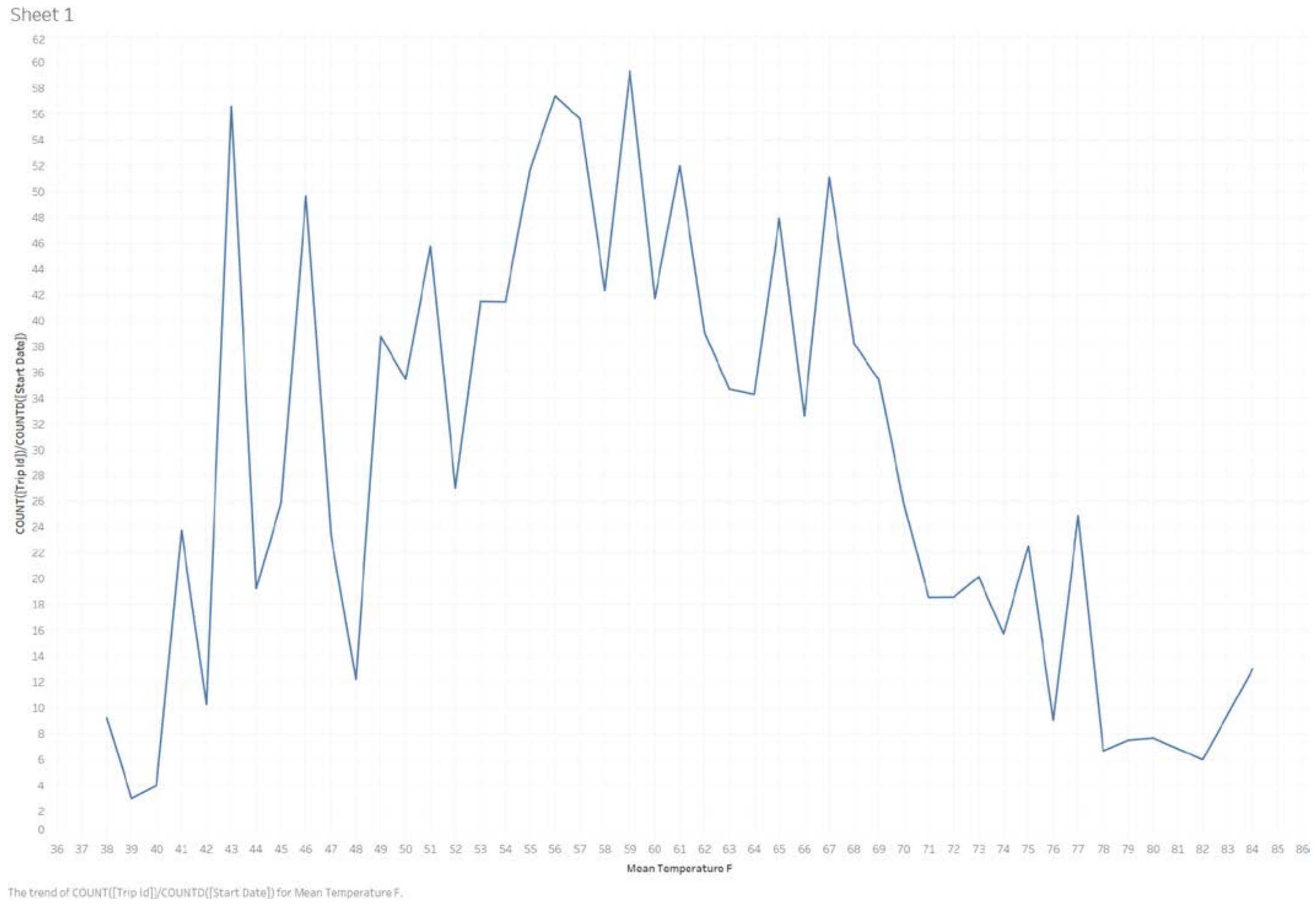
Bike usages pattern for different months



Except for December there is not a lot of variability in demand of bikes across different month.
The drop in December could be contributed to the holiday season.

Data Visualization

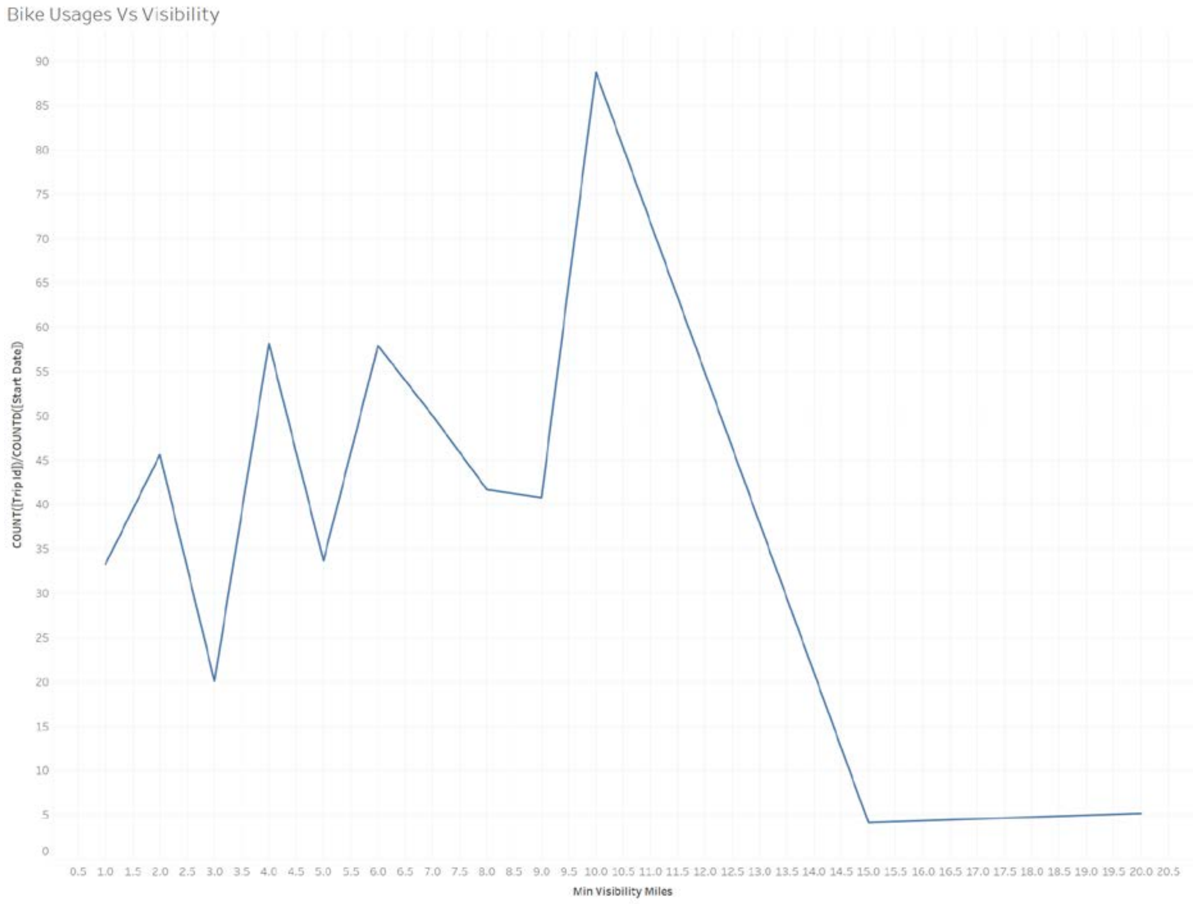
Effect of temperature on bike demand



High and Low temperature both negatively impacts the bike demand.
Low temperature has higher variability in the demand

Data Visualization

Effect of visibility on bike demand



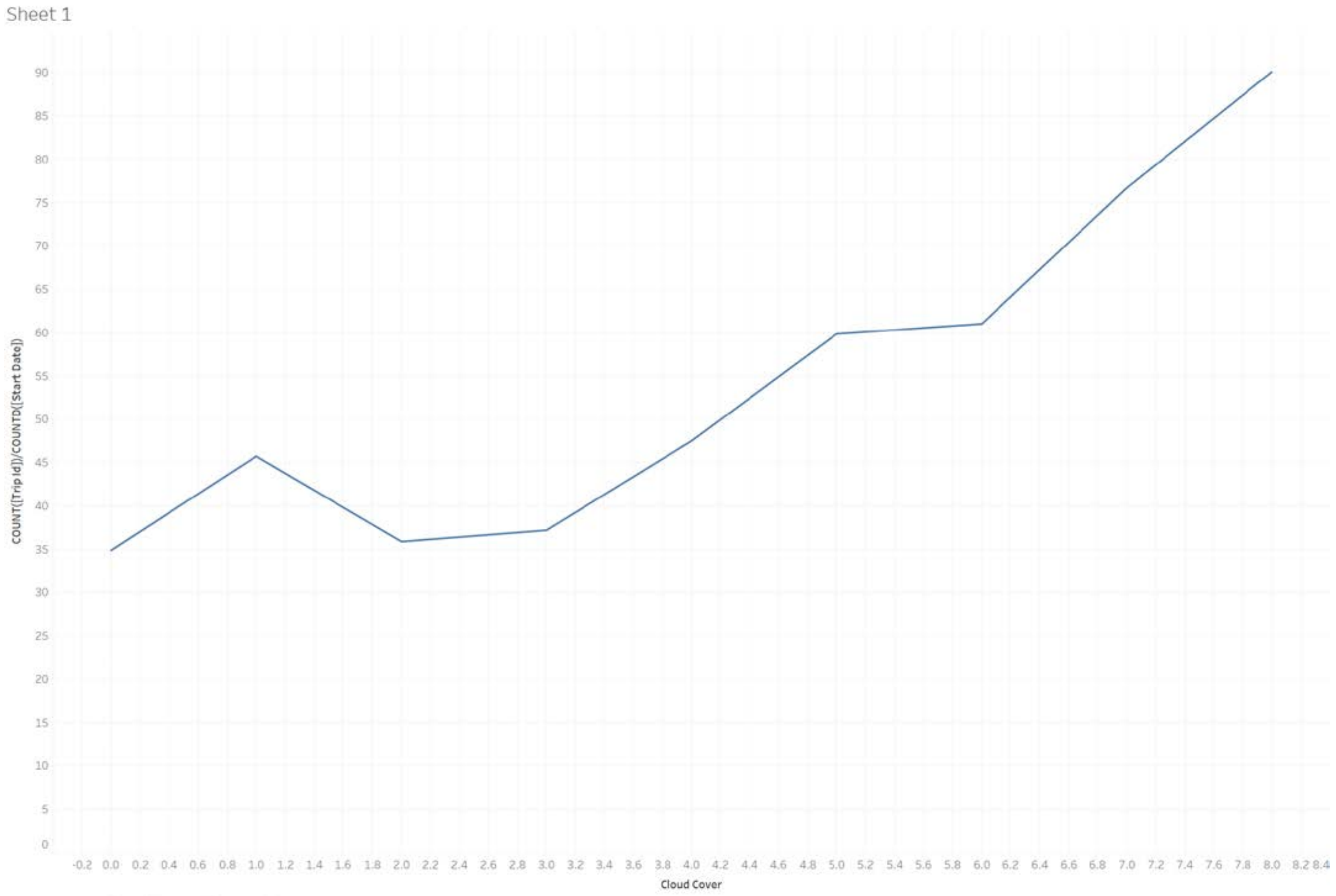
The trend of COUNT([Trip Id])/COUNTD([Start Date]) for Min Visibility Miles. The view is filtered on Min Visibility Miles, which ranges from 0.80 to 20.00.

**High visibility correlates to high temperature and therefore bike demand drops off after certain threshold

BayAreaBikeShare

Data Visualization

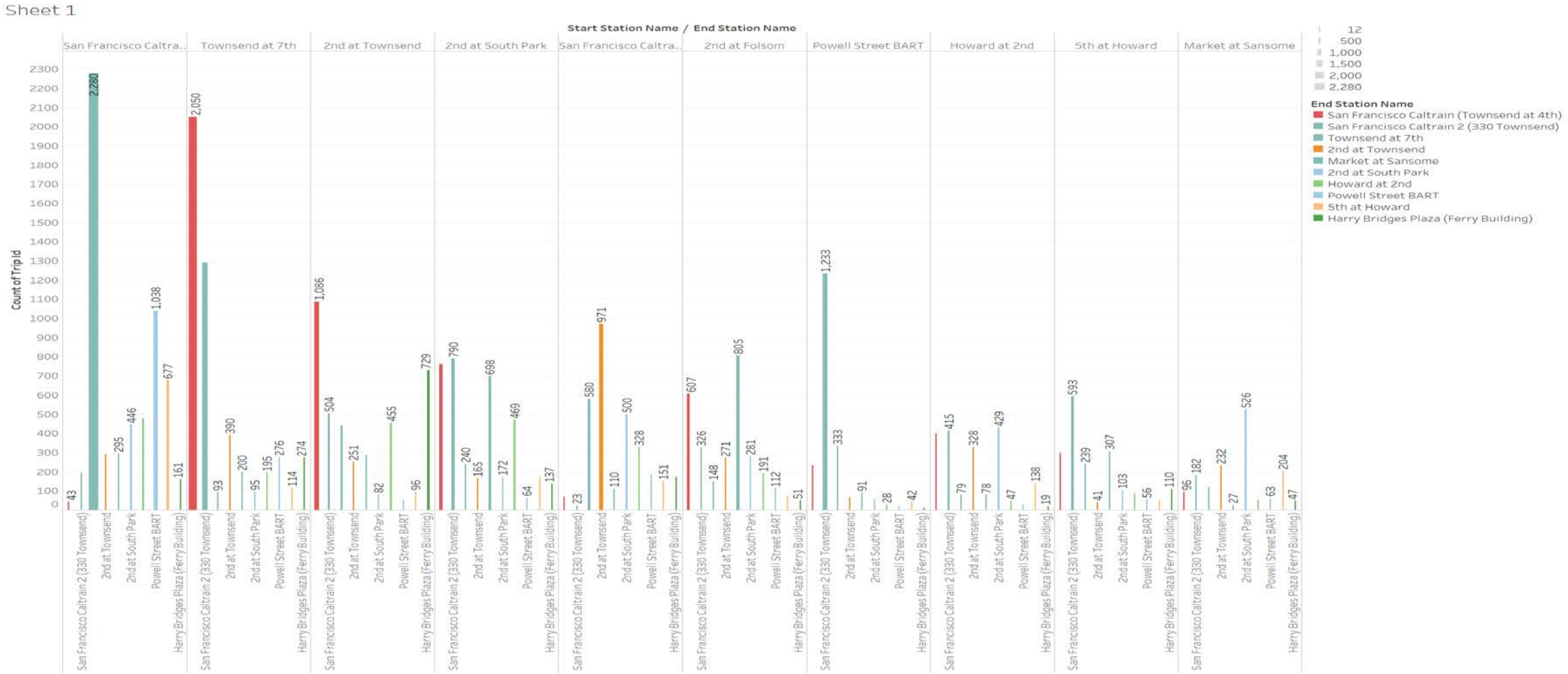
Bike usages and Cloud Cover – Higher Cloud Cover means lower wind and lower temperature.. Looks like very good indicator for bike usages



The trend of COUNT([Trip Id])/COUNTD([Start Date]) for Cloud Cover.

Data Visualization

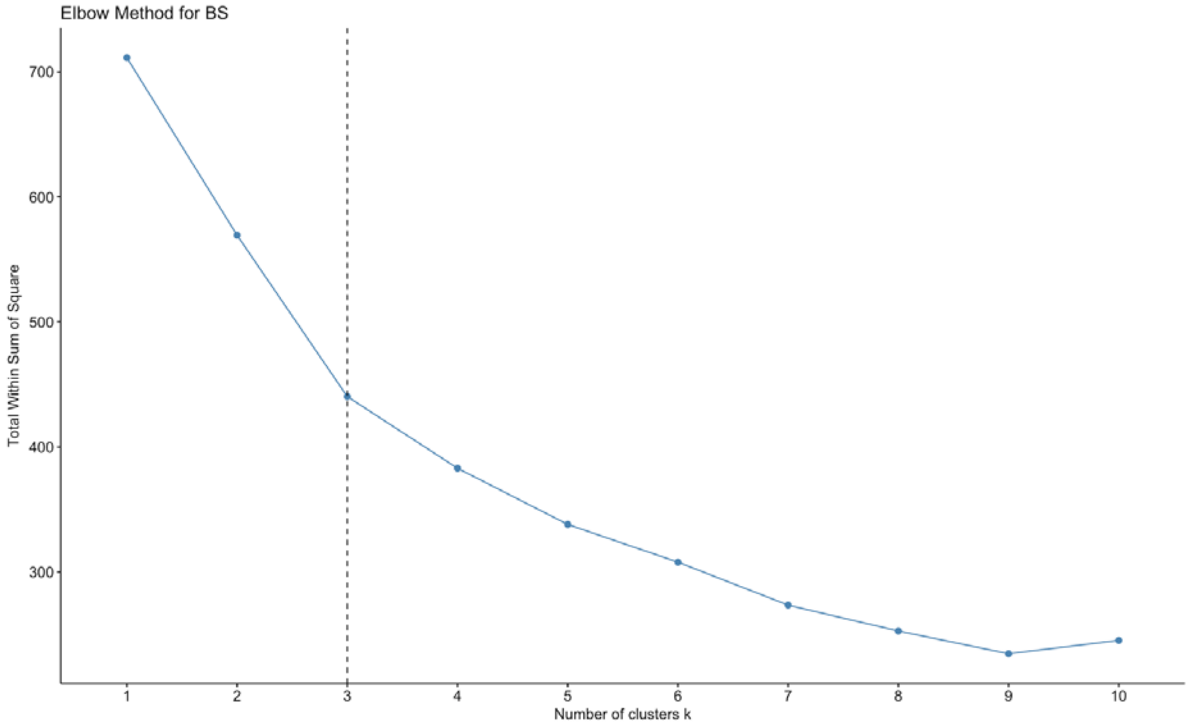
Bike demand in between top ten stations

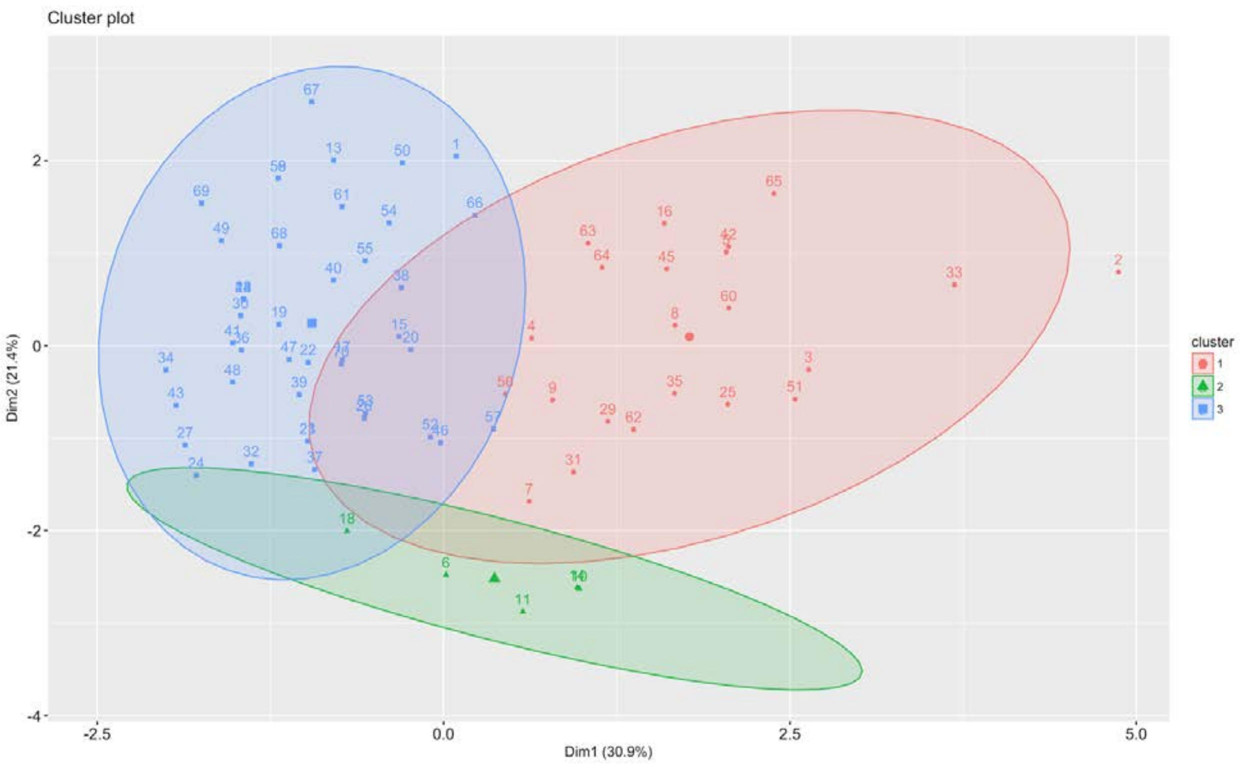


Count of Trip Id for each End Station Name broken down by Start Station Name. Color shows details about End Station Name. Size shows count of Start Station Name. The view is filtered on Start Station Name and End Station Name. The Start Station Name filter keeps 10 of 74 members. The End Station Name filter keeps 10 of 74 members.

BayAreaBikeShare

Data Clustering

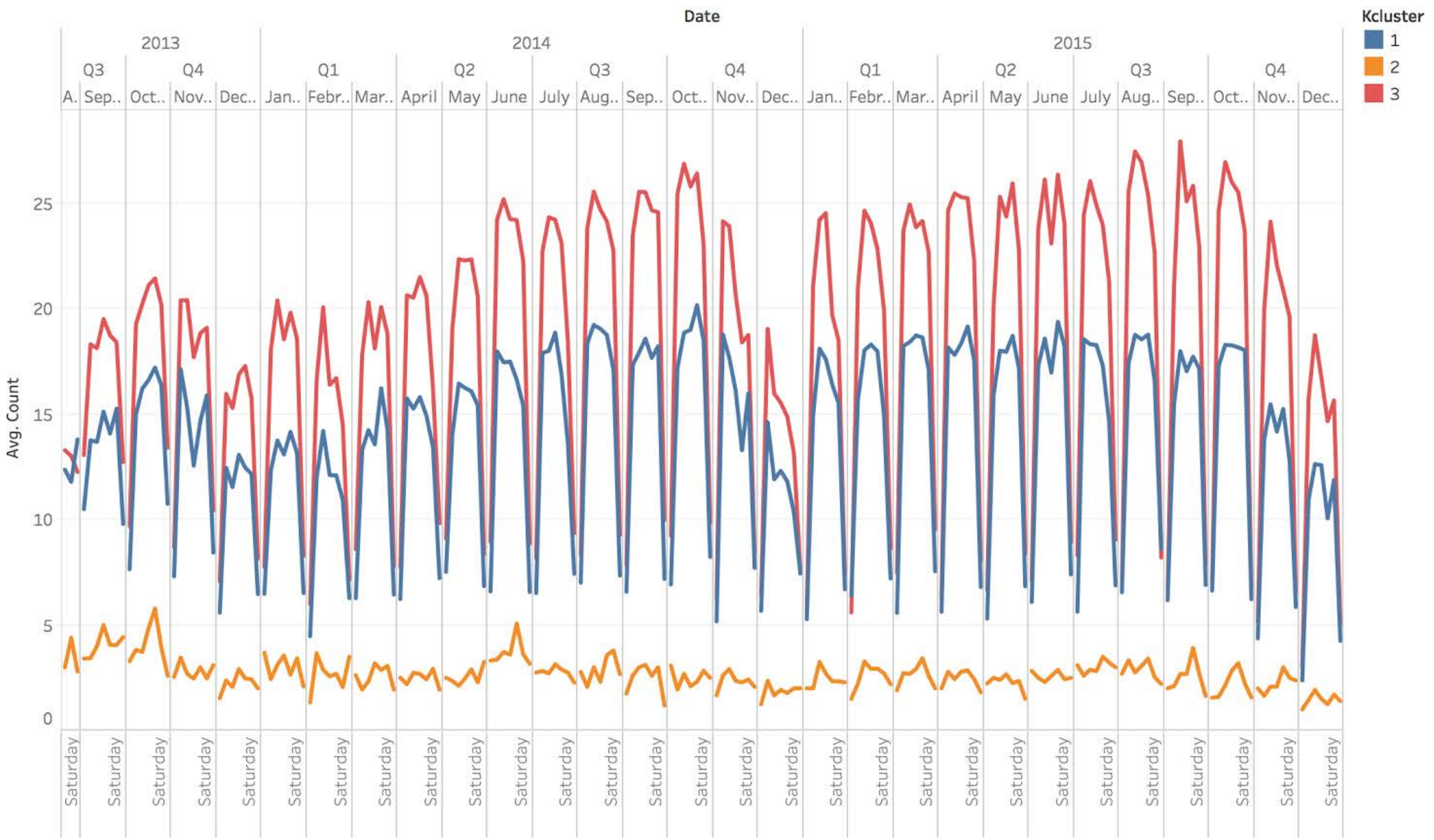




BayAreaBikeShare

Cluster Analysis

Sheet 4



The trend of average of Count for Date Weekday broken down by Date Year, Date Quarter and Date Month. Color shows details about Kcluster. The view is filtered on Date Year and Kcluster. The Date Year filter keeps 2013, 2014 and 2015. The Kcluster filter keeps 1, 2 and 3.

Modelling

Building first model with all predictors except date and duration of the rental.

```
lm.fit=lm(count~.-date-duration, data=data[train,])
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   578.42707  569.28005   1.016 0.309928
## weekday         4.34430    0.90165   4.818 1.76e-06 ***
## month           0.31506    0.56375   0.559 0.576418
## day             0.21710    0.20652   1.051 0.293499
## event1        -10.33674    5.99979  -1.723 0.085332 .
## weekend1       14.42102    9.68442   1.489 0.136887
## holiday1       37.90535   14.38891   2.634 0.008606 **
## sub             0.93281    0.01032  90.392  < 2e-16 ***
## mmaxtemp       14.86021    4.55701   3.261 0.001161 **
## mmeamtemp     -32.70388    8.97154  -3.645 0.000286 ***
## mmintemp       19.79481    4.51608   4.383 1.34e-05 ***
## mmaxdew        -1.55852    1.94859  -0.800 0.424071
## mmeandew        3.00123    3.16847   0.947 0.343837
## mmindew         0.66477    1.37813   0.482 0.629688
## mmaxhum        -0.35587    0.97179  -0.366 0.714321
## mmeanhum        0.28764    1.77564   0.162 0.871355
## mminhum        -1.50255    0.86563  -1.736 0.083015 .
## mmaxsea      -509.14426  115.92807  -4.392 1.29e-05 ***
## mmeansea      786.18124  201.30784   3.905 0.000103 ***
## mminsea      -295.99496  102.30668  -2.893 0.003925 **
## maxvis          8.46994    3.07204   2.757 0.005975 **
## mmeanvis       -0.58194    4.54638  -0.128 0.898183
## mminvis         1.80940    1.80468   1.003 0.316371
## mmaxwind        0.30649    0.67761   0.452 0.651174
## mmeanwind      -6.05022    1.46041  -4.143 3.83e-05 ***
## mmaxgust        1.21832    0.49858   2.444 0.014775 *
## cloud          -7.97502    1.61785  -4.929 1.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.51 on 742 degrees of freedom
## Multiple R-squared:  0.9864, Adjusted R-squared:  0.9859
## F-statistic:  2066 on 26 and 742 DF,  p-value: < 2.2e-16
```

```
vif(lm.fit)
```

```
##    weekday      month        day      event    weekend    holiday
##   1.027814   1.222056   1.050529   1.856674   6.012019   1.462918
##        sub   mmaxtemp  mmeamtemp   mmintemp    mmaxdew   mmeandew
##   6.447136 463.949558 1295.945348 318.503342  47.973284 168.174409
##    mmindew    mmaxhum   mmeanhum    mminhum    mmaxsea   mmeansea
##  44.646099  17.903077  97.775558  42.712696  75.168141 219.318652
##    mminsea     maxvis   mmeanvis    mminvis   mmaxwind  mmeanwind
##  59.161199   1.830159   5.892397   6.651169   3.302713   4.711032
##   mmaxgust      cloud
##   2.641716   3.539605
```

High VIF of certain some predictors indicate these predictors have collinearity

Modelling

Only include one variable for each weather dimension. Select min visibility and max wind, select mean values for all the remaining weather variables. Fit the model again.

```
lm.fit=lm(count~weekday+event+weekend+holiday+sub+mmeamtemp+
          +mmeandew+mmeanhum+mmeansea+mminvis+mmaxwind+mmaxgust+cloud+
          mmeanwind+cloud, data[train, ])
```

```
vif(lm.fit)
```

```
##    weekday     event   weekend   holiday       sub mmeamtemp  mmeandew
##   1.015807  1.715371  5.693402  1.406270  6.093044 35.096598 46.247453
##   mmeanhum  mmeansea   mminvis  mmaxwind  mmaxgust     cloud mmeanwind
## 23.200081  1.688764  2.312401  3.235083  2.576081  2.243650  3.324315
```

Max wind becomes insignificant. Replace it with mean wind, and fit the model.

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.939  -30.438   -2.475   23.563  286.842
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 430.83072  565.06479   0.762  0.44603
## weekday       4.58236    0.93322   4.910 1.12e-06 ***
## event1      -15.45540    6.00402  -2.574  0.01024 *
## weekend1     10.29027    9.81169   1.049  0.29462
## holiday1     30.78426   14.68745   2.096  0.03642 *
## sub           0.92929    0.01044  88.973  < 2e-16 ***
## mmeamtemp     1.70224    1.53709   1.107  0.26846
## mmeandew      2.50876    1.72985   1.450  0.14740
## mmeanhum     -0.89924    0.90049  -0.999  0.31830
## mmeansea    -14.85248   18.39085  -0.808  0.41957
## mminvis       2.93198    1.10784   2.647  0.00830 **
## mmaxwind      0.14922    0.69820   0.214  0.83082
## mmaxgust      1.34533    0.51259   2.625  0.00885 **
## cloud        -6.48441    1.34101  -4.835 1.61e-06 ***
## mmeanwind    -4.74191    1.27721  -3.713  0.00022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.5 on 754 degrees of freedom
## Multiple R-squared:  0.985,  Adjusted R-squared:  0.9847
## F-statistic:  3534 on 14 and 754 DF,  p-value: < 2.2e-16
```

Modelling

```
lm.fit=lm(count~weekday+event+weekend+holiday+sub+mmeamtemp+
           +mmeandew+mmeanhum+mmeansea+mminvis+mmeanwind+mmaxgust+cloud+
           mmeanwind+cloud, data[train, ])
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.769  -30.532   -2.443   23.642  286.702
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 426.74219  564.38384   0.756  0.44981
## weekday       4.58283    0.93262   4.914 1.10e-06 ***
## event1      -15.51768    5.99315  -2.589  0.00980 **
## weekend1     10.24132    9.80281   1.045  0.29648
## holiday1     30.69521   14.67226   2.092  0.03677 *
## sub           0.92925    0.01044  89.038  < 2e-16 ***
## mmeamtemp     1.71726    1.53452   1.119  0.26346
## mmeandew      2.49843    1.72808   1.446  0.14865
## mmeanhum     -0.89463    0.89966  -0.994  0.32034
## mmeansea    -14.71670   18.36825  -0.801  0.42327
## mminvis       2.92494    1.10665   2.643  0.00839 **
## mmeanwind    -4.60542    1.10538  -4.166 3.45e-05 ***
## mmaxgust      1.39536    0.45574   3.062  0.00228 **
## cloud        -6.50445    1.33689  -4.865 1.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.47 on 755 degrees of freedom
## Multiple R-squared:  0.985,  Adjusted R-squared:  0.9847
## F-statistic:  3811 on 13 and 755 DF,  p-value: < 2.2e-16
```

Modelling

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.994  -31.468   -3.113   24.972  292.347
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -93.163232  20.421763  -4.562 5.91e-06 ***
## weekday       4.598221   0.930578   4.941 9.56e-07 ***
## event1      -15.783744   5.990743  -2.635  0.00859 **
## holiday1     22.764633  12.711315   1.791  0.07371 .
## sub           0.919697   0.004411 208.487  < 2e-16 ***
## mmeamtemp     4.120414   0.290933  14.163  < 2e-16 ***
## mminvis       2.962860   0.993616   2.982  0.00296 **
## mmeanwind    -4.228595   1.072656  -3.942 8.82e-05 ***
## mmaxgust      1.303529   0.450951   2.891  0.00395 **
```

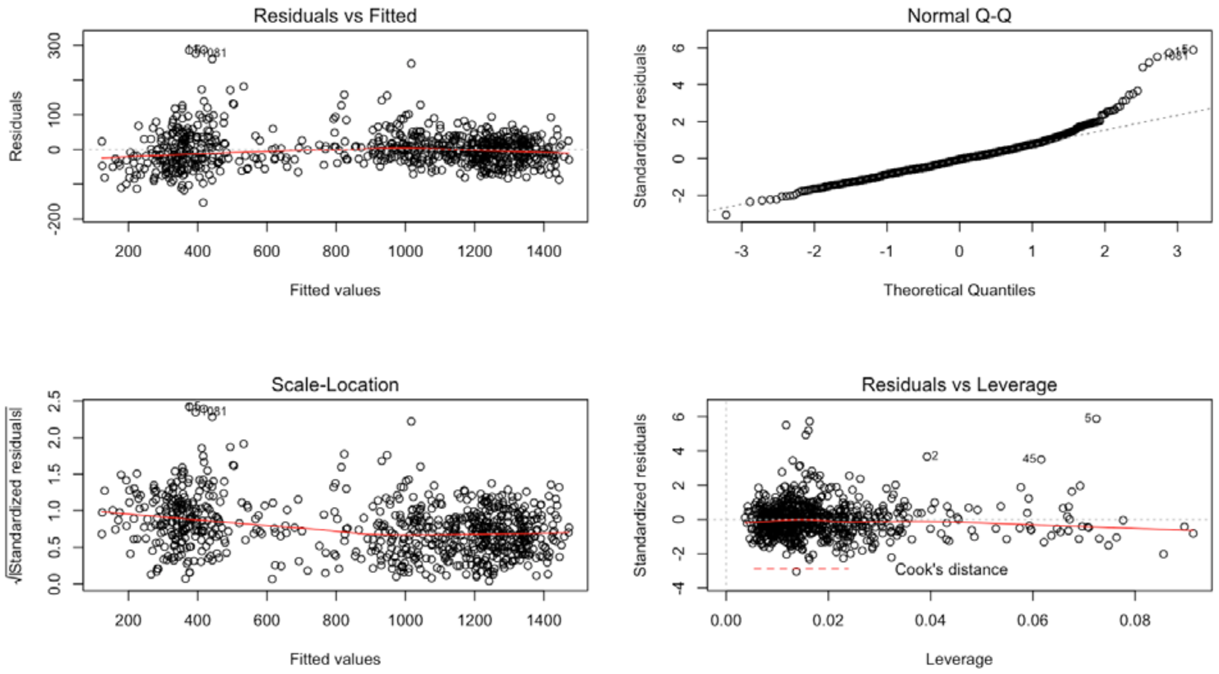
9

```
## cloud        -5.681391   1.164088  -4.881 1.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.54 on 759 degrees of freedom
## Multiple R-squared:  0.9849, Adjusted R-squared:  0.9847
## F-statistic:  5487 on 9 and 759 DF,  p-value: < 2.2e-16
```

Modelling

```
vif(lm.fit)
```

```
##   weekday     event   holiday       sub mmeamtemp   mminvis mmeanwind
##  1.008320  1.704833  1.051485  1.084990  1.255154  1.856926  2.340700
##  mmaxgust     cloud
##  1.990332  1.687742
```

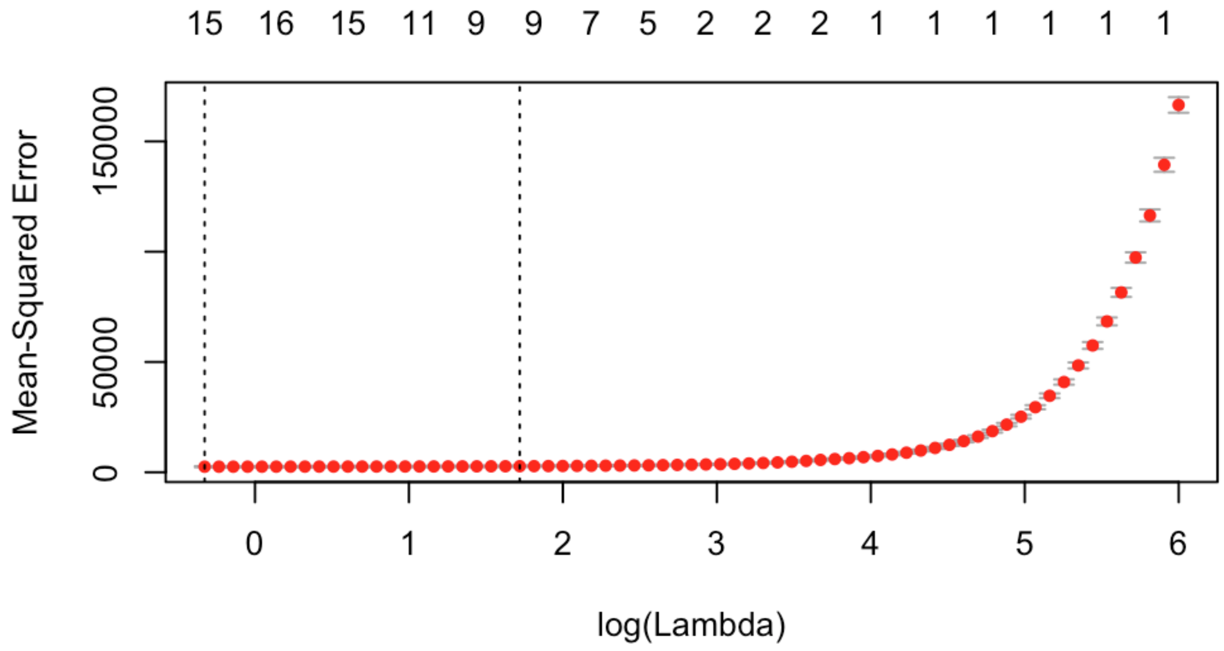


Modelling

```
# test error
pred.count=predict(lm.fit, newdata=data[-train, ])
sqrt(mean((data$count[-train]-pred.count)^2))
```

```
## [1] 63.11523
```

Modelling




```
> bestlam=cv.out$lambda.min
> bestlam
[1] 0.7209649
```

```
> lasso.pred=predict(lasso.mod,s=bestlam ,newx=x[-train,])
> sqrt(mean((lasso.pred-y[-train])^2))
[1] 62.47014
```

BayAreaBikeShare

Modelling

```
                        1
(Intercept) 413.84277667
weekday       4.03730568
month         .
day          -0.02229262
event1      -16.54263295
weekend1      .
holiday1      1.08864565
sub           0.91344106
mmaxtemp      0.16632623
mmeamtemp     .
mmintemp      4.18273827
mmaxdew       .
mmeandew      .
mmindew       0.40188342
mmaxhum      -0.02959267
mmeanhum      .
mminhum      -0.31141169
mmaxsea     -16.24131617
mmeansea      .
mminsea       .
maxvis        5.38296245
mmeanvis      .
mminvis       2.15081928
mmaxwind      .
mmeanwind    -4.81114118
mmaxgust      0.32461551
cloud        -9.59856436
```