Enabling the Supply Chain Optimization for the Citi Bike Station, New York using Machine Learning

Ketan Walia and Swetha Vemula

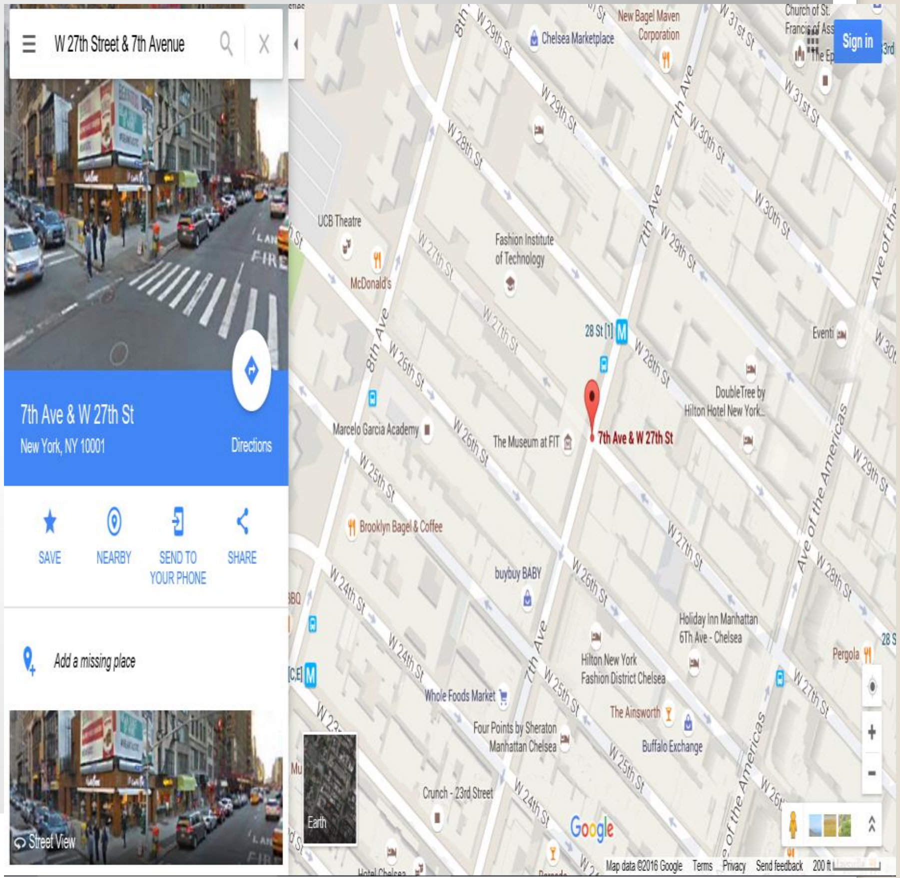
Introduction:

- Citi Bike is New York City's bike sharing system with an efficient network

- A model that could predict the number of rental bikes per hour and span of a particular trip for the bike station–“**The Seventh Avenue at 27 Street Citi Bike station**”

- Depends on the customer’s behavior

- A range of potential factors affecting the customer’s behavior are the gender, age, membership, weather conditions, season, the total number of bikes available at a station, commuting on which day of the week.



Description of concept to be learnt:

- Objective of this study is to enable the supply chain optimization for a Citi bike station in New York City.

- An imperative measure to manage the smooth supply chain is to ensure that optimum number of bikes are available at a particular station at given point in time.

- One of the major problems faced by the bike stations is of re-balancing(maintain a reasonable distribution across docking stations)

- In order to achieve above objective we decided to utilize machine learning techniques to learn two concepts:

    1) To estimate the number of trips per hour for the mentioned bike station

    2) To estimate whether a particular trip would end at the start station and if not, predict the station trip will terminate at.

Problem Statement Redefined:

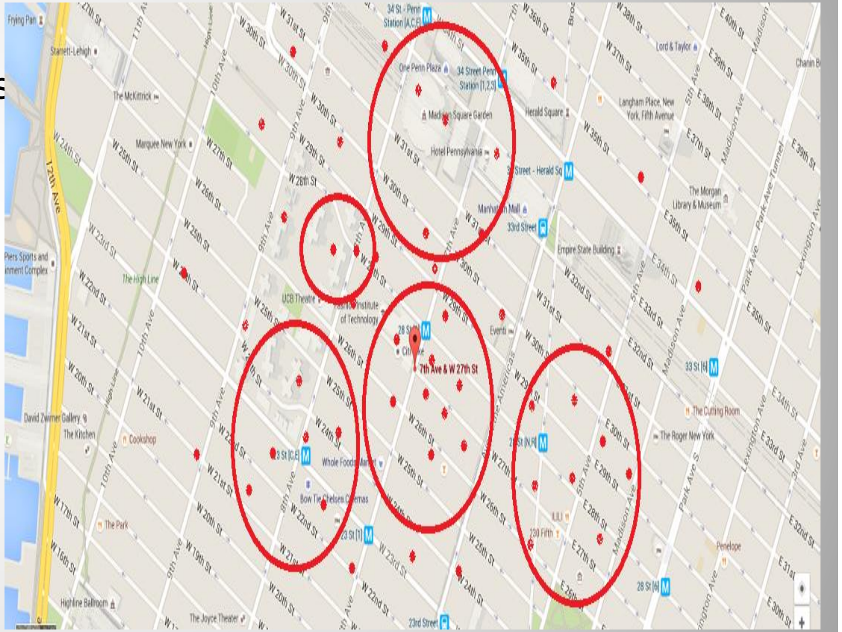
Reason:

- Close proximity of near by stations~0.2miles
- Very less proportion of trips ending at the start station-”7th Ave 27th Street.
- “Problem of Rebalancing persists” from a Business point of view.

Approach: Divide the stations into clusters.

Problem Statement:

“To estimate whether a particular trip would end at the cluster it started from, if not predict which cluster the trip will terminate at.”



Obtaining and Cleansing of Data

- Obtaining data-39,000 instances(CitiBike Data), 10,000(Weather).
- Merging the three datasets- Normalization of data
- Data Validation (weather.com)
- Cleansing of data-

1. Removing Outliers- Box Plots
2. Imputing/Removing Missing values- Predicting missing values, taking Average, Removing irrelevant instances
3. Treating inconsistent data- Manual checking, Validating with internet
4. Formatting/Structuring of dataset- Bringing the merged data to same format, formatting the user data

Attributes	**Source**	**Extraction Method**
Trip information	Citibike Website	API extraction
Holiday Information	Office Holiday website	Extraction data directly
	Time and date website	Extraction data directly
Weather Information	Wunderground.com	Web Scraping
User Information	Citibike Website	API extraction

Outlier Treatment		
Attribute	Upper Cutoff	LowerCutoff
Distance	Mean+2SD	Mean-1.5SD
Age	Mean+1.5SD	Mean-1.5SD
Temperature	Mean+1.5SD	Mean-2SD
Humidity	Mean+1.7SD	Mean-1.5SD
Trip Duration	Mean+SD	N/A
No. of trips	Mean+SD	N/A

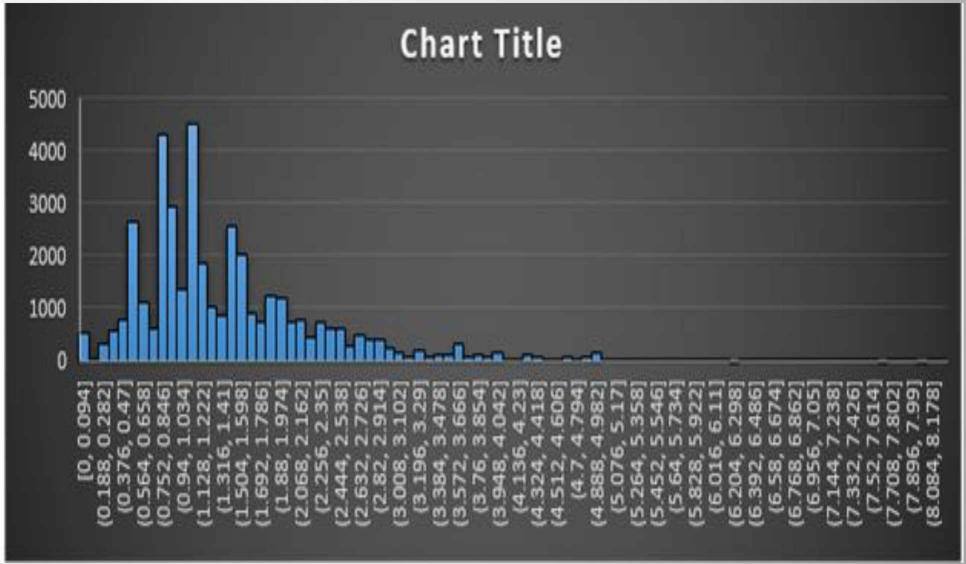
Summary	
Tool	Weka, Python, R, Excel, Mysql
Learning	Supervised & Unsupervised
Data Split	Cross Validation
File	CSV
Attribute Selection Technique	Wrapper Subset Eval Method

Dataset Snippet:

end station name	Index	Time	tripduration(mins)	usertype	age	gender	TemperatureF	Dew PointF	Humidity	Wind SpeedMPH	Conditions	Season	Holiday (Y/N)	Day of week	weekend/weekday	Working day(true/false)	Event(Y/N)	Distance B/w Stations (Miles)
Broadway & E 22 St	1/1/2014	late night	5.5	Subscriber	39	1	25	5	43	9.2	Clear	WINTER	Y	Wednesday	weekday	FALSE	N	0.491823
W 15 St & 7 Ave	1/1/2014	late night	4.65	Subscriber	54	1	25	5	43	9.2	Clear	WINTER	Y	Wednesday	weekday	FALSE	N	0.577759
9 Ave & W 45 St	1/1/2014	morning	8.033333	Subscriber	49	1	28	12	51	5.8	Clear	WINTER	Y	Wednesday	weekday	FALSE	N	0.946203
W 20 St & 7 Ave	1/1/2014	morning	0	Subscriber	55	2	28	12	51	9.2	Clear	WINTER	Y	Wednesday	weekday	FALSE	N	0.342495
Broadway & W 32 St	1/1/2014	morning	4.116667	Subscriber	41	1	28.9	12	49	10.4	Clear	WINTER	Y	Wednesday	weekday	FALSE	N	0.332292
Lexington Ave & E 26 St	1/1/2014	morning	4.583333	Subscriber	62	1	30	12	47	6.9	Mostly Cloudy	WINTER	Y	Wednesday	weekday	FALSE	N	0.664734
W 4 St & 7 Ave S	1/1/2014	afternoon	7.866667	Subscriber	23	1	32	15.1	50	8.1	Partly Cloudy	WINTER	Y	Wednesday	weekday	FALSE	N	0.992646
Park Pl & Church St	1/1/2014	afternoon	23	Subscriber	56	1	32	16	52	8.1	Clear	WINTER	Y	Wednesday	weekday	FALSE	N	2.438139
W 26 St & 8 Ave	1/1/2014	afternoon	199.9	Subscriber	36	1	32	16	52	0	Mostly cloudy	WINTER	Y	Wednesday	weekday	FALSE	N	0.180439
W 22 St & 10 Ave	1/1/2014	afternoon	10.11667	Subscriber	34	1	32	16	52	0	Mostly cloudy	WINTER	Y	Wednesday	weekday	FALSE	N	0.555381
W 45 St & 6 Ave	1/1/2014	afternoon	8.483333	Subscriber	24	1	32	16	52	0	Mostly cloudy	WINTER	Y	Wednesday	weekday	FALSE	N	0.907646
W 41 St & 8 Ave	1/1/2014	afternoon	4.5	Subscriber	25	1	32.5	15.5	50	0	Clear	WINTER	Y	Wednesday	weekday	FALSE	N	0.7043

Approach for clustering:

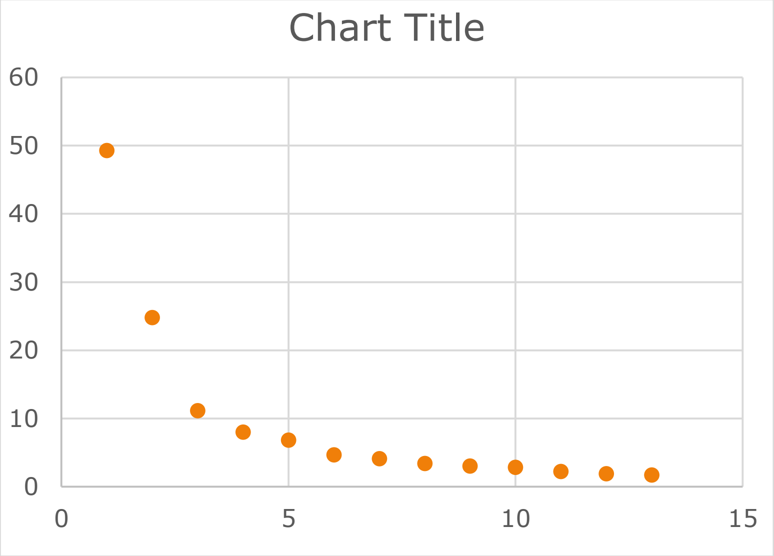
- Business Problem-Supply chain optimization through re-balancing
- Relevant attribute- Distance and Location(longitude & latitude): Introduced Bias
- Considered Attribute: Distance



Clustering

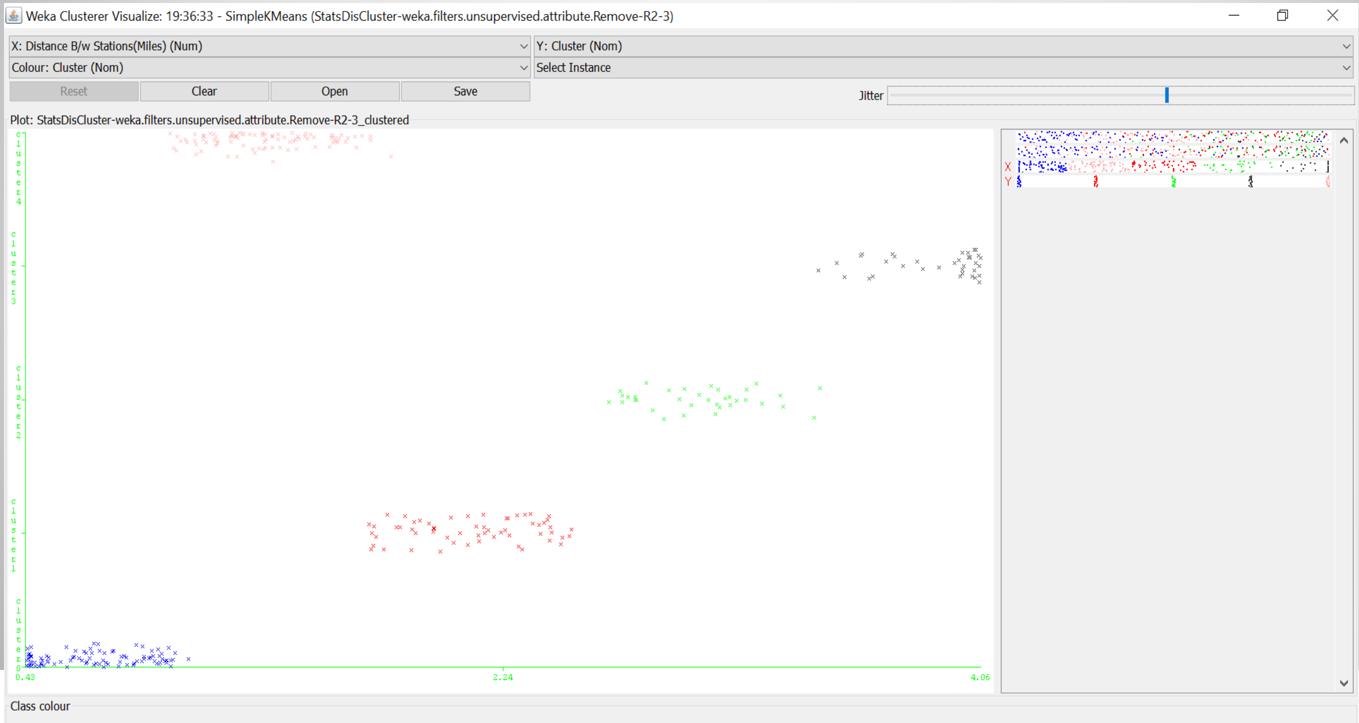
- Simple K means
- Bias: significant fluctuation with random seeds
  - Initialization method: Random
- Approach: K means with initialization method K++

K	Within cluster sum of squared errors
1	49.26829776
2	24.78157677
3	11.10708489
4	7.980840837
5	6.807196242
6	4.646752241
7	4.073448372
8	3.353900433
9	2.976969237
10	2.786784985
11	2.220590131
12	1.88433489
13	1.67360513



cluster	percentage of stations
1	29%
2	19%
3	11%
4	12%
5	29%

Clustering:


Weka Clusterer Visualize: 19:36:33 - SimpleKMeans (StatsDisCluster-weka.filters.unsupervised.attribute.Remove-R2-3)
X: Distance B/w Stations(Miles) (Num)
Y: Cluster (Nom)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: StatsDisCluster-weka.filters.unsupervised.attribute.Remove-R2-3_clustered
cluster4
cluster3
cluster2
cluster1
cluster0
0.43
2.24
4.06
X
Y
Class colour

Class Attribute generation:

Season	Holiday	Day of week	weekend/day	working day	event	popularity	D-C5
WINTER	N	Tuesday	weekday	TRUE	N	High	cluster5
SPRING	N	Tuesday	weekday	TRUE	N	Good	cluster1
AUTUMN	Y	Monday	weekday	FALSE	N	Low	cluster3
WINTER	N	Monday	weekday	TRUE	N	Very High	cluster5
WINTER	N	Monday	weekday	TRUE	N	Very High	cluster5
SUMMER	N	Friday	weekday	TRUE	N	Very High	cluster5
SUMMER	N	Tuesday	weekday	TRUE	N	Very High	cluster5
WINTER	N	Friday	weekday	TRUE	N	Low	cluster4
AUTUMN	N	Monday	weekday	TRUE	N	Low	cluster1
SUMMER	N	Monday	weekday	TRUE	N	Good	cluster5
WINTER	N	Thursday	weekday	TRUE	N	Low	cluster4
AUTUMN	N	Tuesday	weekday	TRUE	N	Low	cluster1
AUTUMN	N	Tuesday	weekday	TRUE	N	Very High	cluster5
SUMMER	N	Thursday	weekday	TRUE	N	Good	cluster3
SPRING	N	Thursday	weekday	TRUE	N	High	cluster1

Classification:

Algorithm	Accuracy	F-measure
naïve bayes	59.10%	45.5
J48	59.90%	54.3
IBK	53.30%	48.6
Logistic	59.20%	44.6
Zero R	59.10%	44

Naïve Bayes

```
=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 7197  110    0    2    0 |    a = cluster5
 3681  105    0    3    0 |    b = cluster3
  248    8    0    0    0 |    c = cluster4
  896   30    0    1    0 |    d = cluster1
   69    2    0    0    0 |    e = cluster2
```

KNN

```
=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5739 1422   17  130    1 |    a = cluster5
 2887  821    5   75    1 |    b = cluster3
  197   54    1    4    0 |    c = cluster4
  692  204    2   29    0 |    d = cluster1
   53   17    0    1    0 |    e = cluster2
```

Logistic Regression

```
=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 7282   27    0    0    0 |    a = cluster5
 3750   39    0    0    0 |    b = cluster3
  250    6    0    0    0 |    c = cluster4
  913   14    0    0    0 |    d = cluster1
   70    1    0    0    0 |    e = cluster2
```

J48

```
    a    b    c    d    e   <-- classified a
 6414  766   24  102    3 |    a = cluster5
 2791  915   11   68    4 |    b = cluster3
  198   31   15   12    0 |    c = cluster4
  722  143    4   58    0 |    d = cluster1
   60   10    0    1    0 |    e = cluster2
```

Why low performance?

- Algorithms are not learning the concept for cluster 4, cluster 2 and cluster 1
- Weather attributes effect the trip after a certain distance of ~0.8 miles
- Under representation of classes

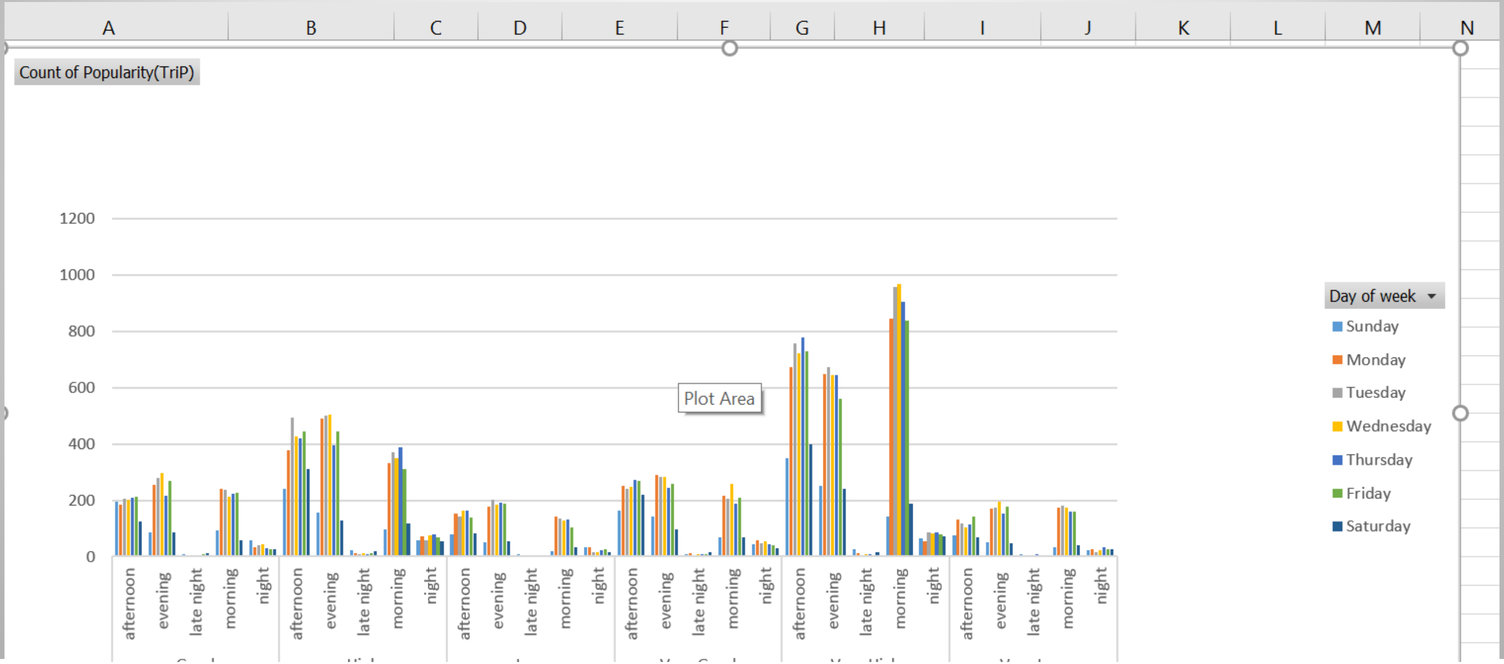
  -More information required for learning the concept

Label	Percentage proportion
Cluster 5	59.6%
Cluster 3	30%
Cluster 4	2.1%
Cluster 1	7.6%
Cluster 2	0.5%

Problem 1: Weather attributes effect the trip after a certain distance of ~0.8 miles

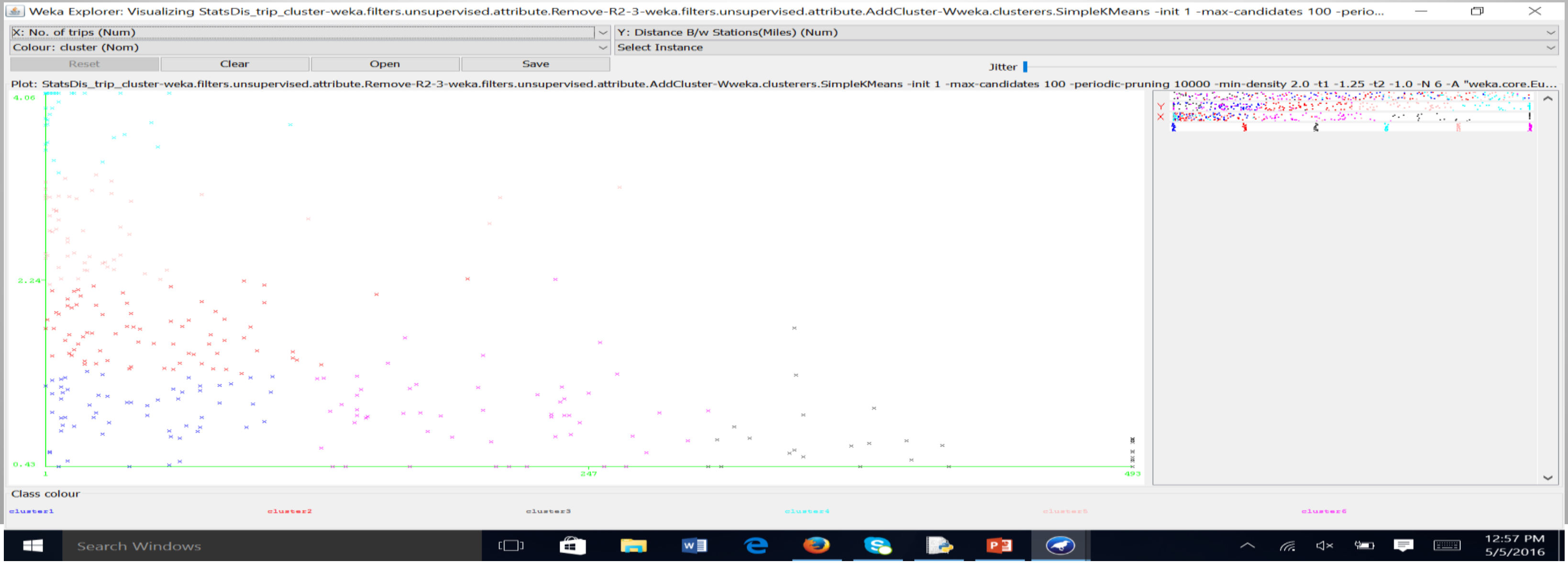
Trend noticed based on popularity:

Irrespective of the season or timing, there exists a similar frequency(95%) of trips per hour for the various stations



Attribute generation:

- There exists no generalized relationship between distance and trip. Popularity in itself represents other attributes associated with a particular station(schools, offices, bus stops)
- Attribute generation based on frequency of trips received.



Attribution generation based on frequency of trips:

Snippet of data:

Attribute value: very high, high, very good, good, low, very low

Conditions	Season	Holiday	Day of Week	Weekend/Weekday	Working Day	Event	Popularity	D C-5
Clear	WINTER	N	Tuesday	weekday	TRUE	N	High	cluster5
Mostly cloudy	SPRING	N	Tuesday	weekday	TRUE	N	Good	cluster1
Mostly cloudy	AUTUMN	Y	Monday	weekday	FALSE	N	Low	cluster3
Mostly cloudy	WINTER	N	Monday	weekday	TRUE	N	Very High	cluster5
Clear	WINTER	N	Monday	weekday	TRUE	N	Very High	cluster5
Mostly cloudy	SUMMER	N	Friday	weekday	TRUE	N	Very High	cluster5
Fog	SUMMER	N	Tuesday	weekday	TRUE	N	Very High	cluster5
Clear	WINTER	N	Friday	weekday	TRUE	N	Low	cluster4
Clear	AUTUMN	N	Monday	weekday	TRUE	N	Low	cluster1
Clear	SUMMER	N	Monday	weekday	TRUE	N	Good	cluster5
Clear	WINTER	N	Thursday	weekday	TRUE	N	Low	cluster4
Clear	AUTUMN	N	Tuesday	weekday	TRUE	N	Low	cluster1
Clear	AUTUMN	N	Tuesday	weekday	TRUE	N	Very High	cluster5
Mostly Cloudy	SUMMER	N	Thursday	weekday	TRUE	N	Good	cluster3
Clear	SPRING	N	Thursday	weekday	TRUE	N	High	cluster1

Classification:

Testing option: Cross Validation

Algorithm	Accuracy	F-measure
naïve bayes	69.40%	67.8
J48	71.8%	71
IBK	63.2%	63
Logistic	70.2%	67.3
Zero R	59.10%	44

Logistic Regression

```
=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 6048 1261    0    0    0 |    a = cluster5
 1163 2626    0    0    0 |    b = cluster3
    2  254    0    0    0 |    c = cluster4
   31  896    0    0    0 |    d = cluster1
    0   71    0    0    0 |    e = cluster2
```

Naïve Bayes

```
=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5966 1310    1   32    0 |    a = cluster5
 1157 2504    3  123    2 |    b = cluster3
    5  201    1   47    2 |    c = cluster4
   31  792    0  102    2 |    d = cluster1
    0   60    0    9    2 |    e = cluster2
```

KNN

```
=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5758 1285   54  197   15 |    a = cluster5
 1452 1791  113  398   35 |    b = cluster3
   53  102   30   68    3 |    c = cluster4
  243  377   66  223   18 |    d = cluster1
   12   31    9   11    8 |    e = cluster2
```

J48

```
=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 6255  931   23   95    5 |    a = cluster5
 1140 2318   62  261    8 |    b = cluster3
   28  125   49   52    2 |    c = cluster4
  186  458   37  239    7 |    d = cluster1
    9   30    9   13   10 |    e = cluster2
```

Problem 2:Under representation of classes

- Generating synthetic instances for under-represented classes
- Supervised learning
- Technique: Synthetic minority Oversampling Technique(SMOTE) & Randomized sampling

Label	Percentage proportion
Cluster 5	45.7%
Cluster 3	23%
Cluster 4	9.7%
Cluster 1	11.6%
Cluster 2	9.8%

Algorithm performance

Algorithm	Accuracy	F-measure
naïve bayes	64.6%	64.2
J48	74.03%	73.37
IBK	69.8%	69.6
Logistic	66.09%	65.9
Zero R	45.7%	28.7

=== Confusion Matrix === Naive Bayes:

a	b	c	d	e	<-- classified as
17680	523	2915	329	235	a = cluster5
187	1463	1977	1157	750	b = cluster1
3170	1074	5209	911	554	c = cluster3
20	552	498	2462	1082	d = cluster4
0	232	175	399	3850	e = cluster2

=== Confusion Matrix === J48

a	b	c	d	e	<-- classified as
18451	563	2484	130	54	a = cluster5
633	2774	1468	500	159	b = cluster1
3322	1283	5856	364	93	c = cluster3
133	369	371	3640	101	d = cluster4
32	92	79	77	4376	e = cluster2

=== Confusion Matrix === Logistic Regression

a	b	c	d	e	<-- classified as
17865	419	2881	308	209	a = cluster5
175	1571	1974	1178	636	b = cluster1
3243	1037	5277	913	448	c = cluster3
7	464	445	2746	952	d = cluster4
0	195	57	533	3871	e = cluster2

=== Confusion Matrix === IBK

a	b	c	d	e	<-- classified as
17095	726	3641	173	47	a = cluster5
746	2823	1425	402	138	b = cluster1
3926	1495	4943	421	133	c = cluster3
144	282	320	3802	66	d = cluster4
30	54	67	37	4468	e = cluster2

Classifier comparision:

Null Hypothesis: All 4 classifiers perform similarly

Dataset	Metric	J48	Logistic Regression	Naïve Bayes	IBK
Experimenter File	Accuracy	74.03	66.09	64.64	69.87
Random Seeds>10	F-Measure	73.37	65.9	64.2	69.6

- At 95% confidence level J48 significantly performs better than IBK, Logistic and Naïve Bayes
- Therefore, Null Hypothesis is rejected
- J48 is the better classifier for the considered data set

Way forward

- Adding the attribute, trip duration significantly improves the accuracy and F-measure for most of the classifiers
- And there is an improvement in learning the concept for all the 5 clusters.
- We will be exploring this attribute further for improving the model
- Identifying the attributes affecting the trip duration and building a regression model to predict the trip duration
- Further observing its implementation in our classification model

Algorithm	**Accuracy**	**F-measure**
naïve bayes	73.2%	71.7
J48	80.80%	80
IBK	65.1%	65
Logistic	75.3%	73.8
Zero R	59.10%	44

Task 2: Estimating the no. of trips/hour

Linear Regression:

Index	Time	TemperatureF	Dew PointF	Humidity	Wind SpeedMPH	Holiday (Y/N)	Day of week	weekend/ weekday	Working day(true/false)	Event(Y/N)	no. of trips
1/1/2014	1	25	5	43	9.2	Y	Wednesday	weekday	FALSE	N	2
1/1/2014	8	28	12	51	5.8	Y	Wednesday	weekday	FALSE	N	1
1/1/2014	9	28	12	51	9.2	Y	Wednesday	weekday	FALSE	N	1
1/1/2014	10	28.9	12	49	10.4	Y	Wednesday	weekday	FALSE	N	1
1/1/2014	11	30	12	47	6.9	Y	Wednesday	weekday	FALSE	N	1
1/1/2014	13	32	15.1	50	8.1	Y	Wednesday	weekday	FALSE	N	1
1/1/2014	14	32	16	52	8.1	Y	Wednesday	weekday	FALSE	N	1
1/1/2014	15	32	16	52	0	Y	Wednesday	weekday	FALSE	N	3
1/1/2014	16	32.5	15.5	50	0	Y	Wednesday	weekday	FALSE	N	1
1/1/2014	17	33.1	15.1	48	0	Y	Wednesday	weekday	FALSE	N	2

Regression:

- Linear Regression vs M5P

```
Analysing:  Correlation_coefficient
Datasets:   1
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       5/11/16 6:22 PM


Dataset                  (1) functio | (2) tree
------------------------------------------------
'9 months regression Bike(100)   0.46 |   0.74 v
------------------------------------------------
                             (v/ /*) |  (1/0/0)


Key:
(1) functions.LinearRegression '-S 0 -R 1.0E-8 -num-decimal-places 4' -3364580862046573600
(2) trees.M5P '-M 4.0' -6118439039768244200
```

```
Analysing:  Root_mean_squared_error
Datasets:   1
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       5/11/16 6:24 PM


Dataset                  (1) functio | (2) tree
------------------------------------------------
'9 months regression Bike(100)   3.44 |   2.57 *
------------------------------------------------
                             (v/ /*) |  (0/0/1)


Key:
(1) functions.LinearRegression '-S 0 -R 1.0E-8 -num-decimal-places 4' -3364580862046573600
(2) trees.M5P '-M 4.0' -6118439039768244200
```

Classifier comparision:

Null Hypothesis: Both the algorithms perform equally

Dataset	Metric	M5P	Linear Regression
Experimenter File	Accuracy	0.74	0.46
Random Seeds>10	F-Measure	2.57	3.44

- At 95% confidence level M5P significantly performs better than Linear Regression.
- Therefore, Null Hypothesis is rejected
- M5P is the better classifier for the considered data set

Thank you…………