Identifying classes of passengers in the Brisbane Translink Network

Introduction

Out of the 2 Million people in the city of Brisbane, close to 750,000 people used the public transit system in the month of March, 2013. This amounted to more than 14 Million transactions in the month alone.

Different types of users interact differently with the transportation system. By studying these interactions, it is possible to divide people into different groups.

While it is difficult to tell much about a person's lifestyle if he/she is a very infrequent user to the transportation system, the way that a frequent user interacts with the transportation system can tell us much about his travel preferences and can be indicative of the type of work he is involved in. An example in this context would be school students. The anticipated behavior for them would be to reach school at a fixed time in the morning and leave at almost the same time every weekday. Even though school students might not be the only ones with such travel patterns, there is a high probability of a person with such travel patterns to be a student.

Once we are able to get a classification of transit users into different groups, we can use it to study the potential impact of policy changes to the transportation network. For example, if we have a class-wise time series of transit usage, we can attempt to study (say) the impact of changing school timings on the network as a whole (network's time series).

Other than this, OD studies for the different classes can help us to identify the distribution of different classes of people in different analysis zones. This might be useful in studying the geographic distribution of some demographic characteristic.

This report is a preliminary effort towards determining the classes of people using the Brisbane transportation network.

Outline

The average user of a transportation system is expected to make an AM trip to work and a PM trip to home. The regularity/flexibility of a person to do so depends upon the nature of work he/she engages in. Hence, for majority of the frequent users of the transportation system, 4 variables- mean AM departure time, standard deviation in AM time, PM mean departure time and standard deviation in PM mean departure times- can be good indicators of a person's job requirement and hence can be used to classify people.

As a starting point, we need to set a criterion to consider a user to be 'frequent' to the system. The time period under consideration is the month of March, 2013. The number of weekdays during this period was 21. Within these 21 days, **any user who made 20 or more journeys was defined to be 'frequent'**. This limit was based on intuition, attempting to capture only those people who went to work for at least half the number of working days.

Once these users were subset, the 4 defined variables were used to cluster them into different classes.

Detailed methodology:

1. Defining the 'frequent' users:

Out of the 741,376 users in the month of March, 192,362 users were classified as frequent based on the 20 journeys criterion. Out of these 192,362 users, 272 were found never to have travelled in the PM period and 1088 did not travel in the AM period.

Primary Approach:

2. Calculation of the parameters for classification:

For each of these users, the timestamp for the trips with ID=1 for each of the journeys was noted. These timestamps were then segregated into AM and PM timestamps based on the traditional AM and PM definition. For each of the two groups (AM and PM) for each of the users, the mean and the standard deviation were calculated. These were then used as the variables for clustering.

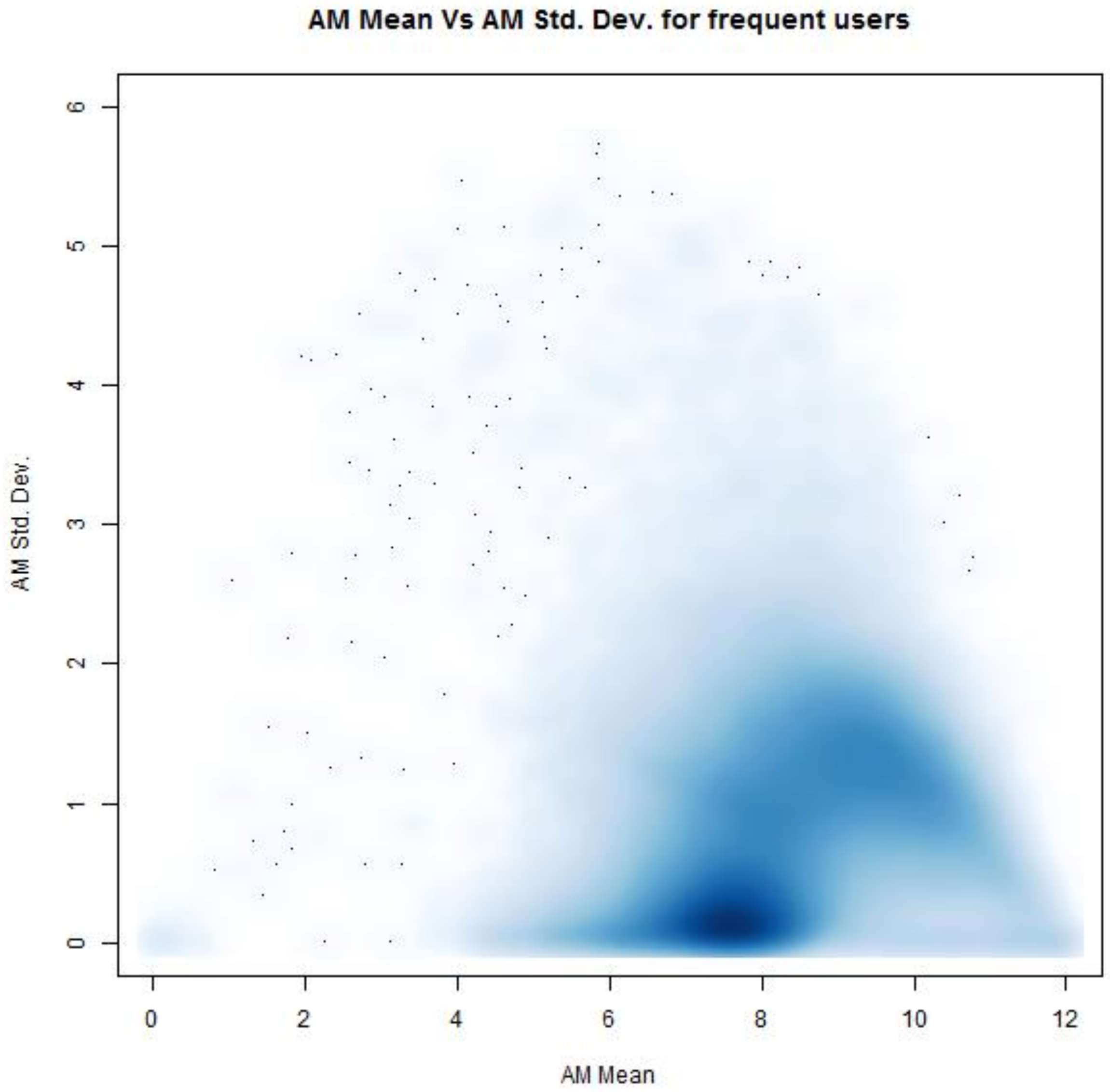
3. Clustering details:

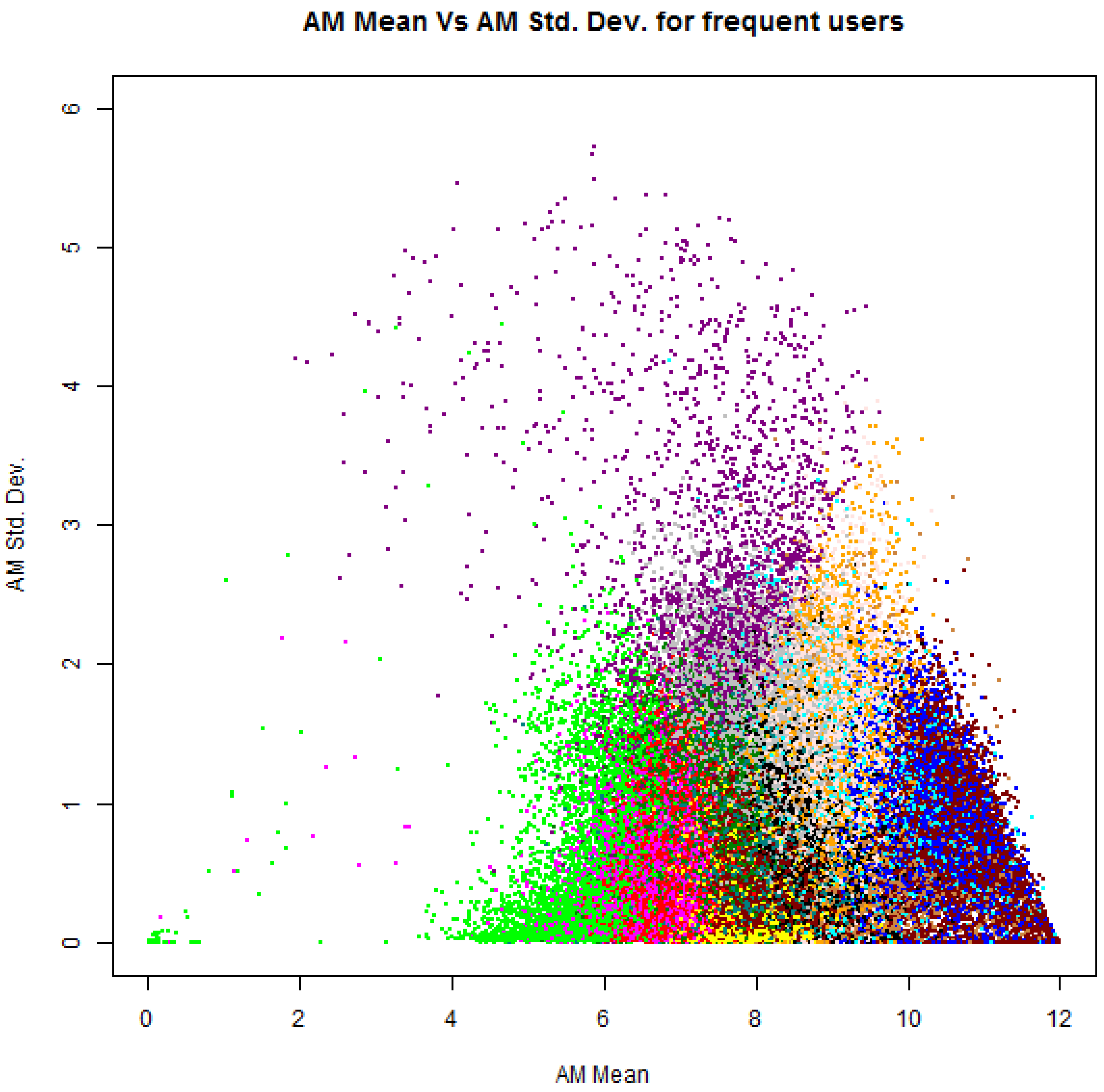
K-means clustering was then employed to identify the types of users amongst the 192,362 frequent users. As there are 4 variables used for the clustering, the number of classes of users (clusters) was set to 16 (the 4th power of 2 to accommodate high and low values for each of the variables).

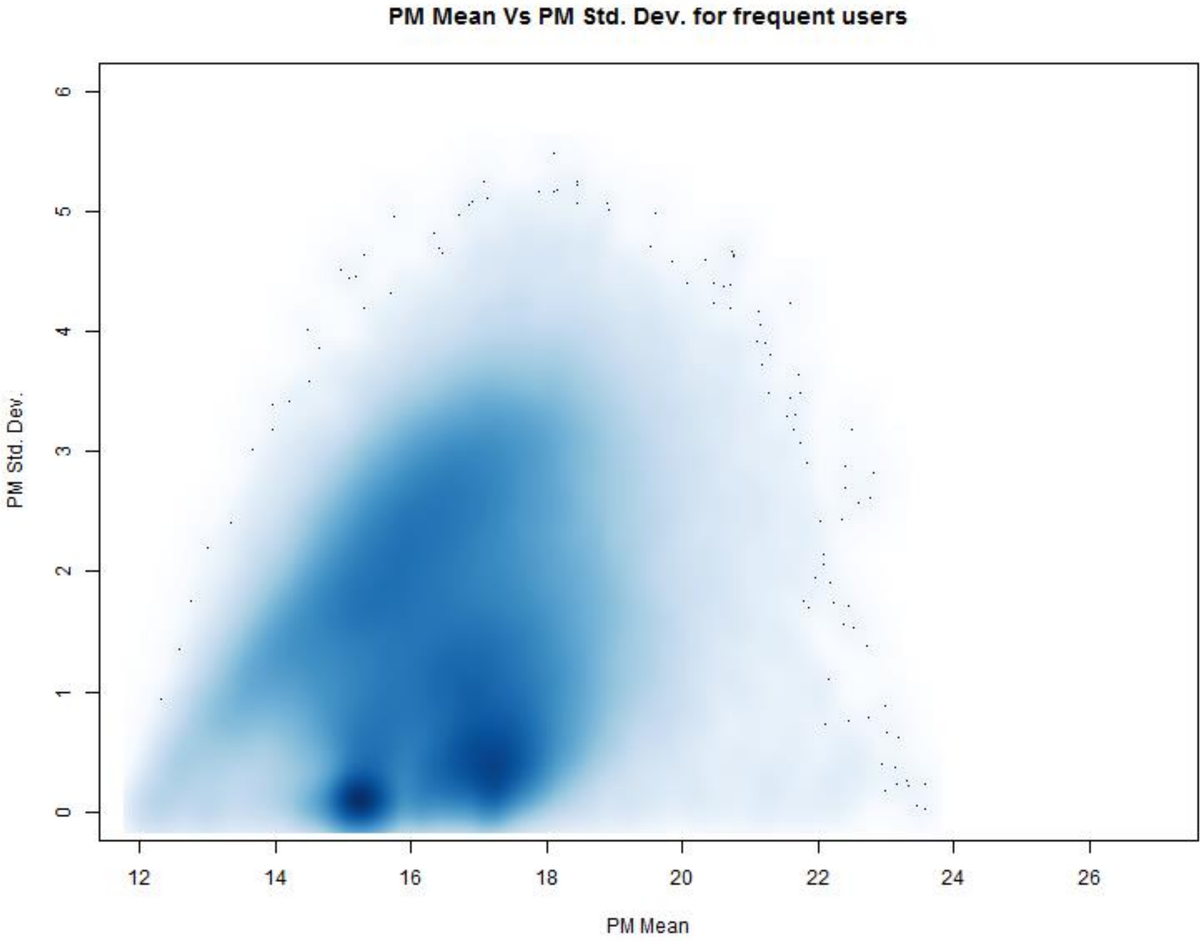
The results for the cluster centers from this clustering are shown below-

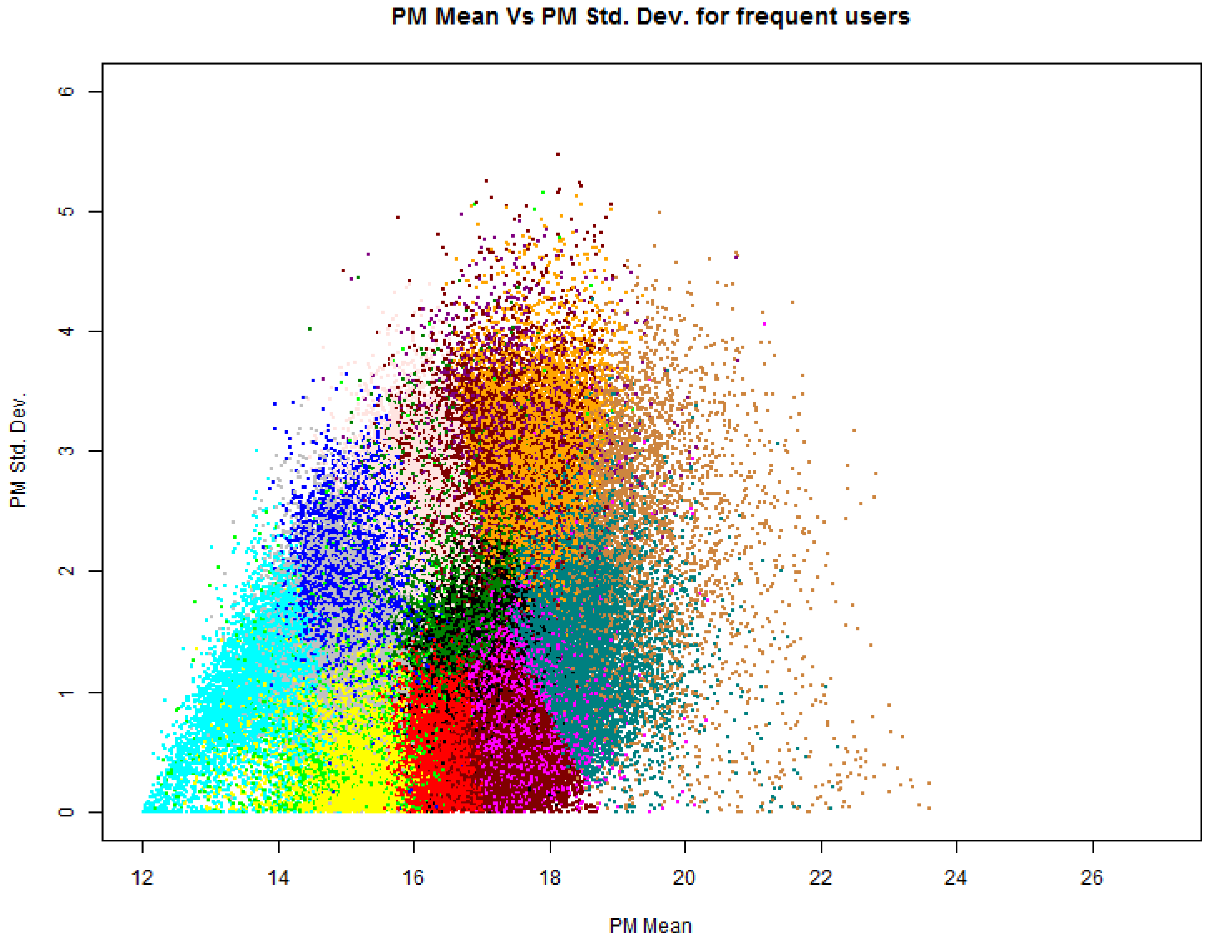
	AM Mean	AM Std. Dev.	PM Mean	PM Std. Dev.	# Points	Within SS
1	7.053	0.369	16.397	0.599	17167	9027.468
2	9.454	1.171	13.808	1.058	7148	8927.533
3	10.034	1.075	15.292	1.886	11562	10578.244
4	10.430	0.892	17.016	2.775	9361	10493.469
5	7.739	0.214	15.240	0.361	20083	10669.615
6	5.901	0.471	15.428	0.864	6968	11815.058
7	6.884	0.355	17.444	0.814	12127	8601.299
8	8.262	1.327	15.113	1.510	14652	14644.261
⬛	8.665	0.913	16.872	1.390	10826	9305.035
10	9.035	1.287	17.640	2.850	10413	10875.865
11	7.903	0.370	17.308	0.557	22378	10197.960
12	7.542	0.687	16.504	1.833	13779	12628.500
13	7.413	2.484	17.217	2.891	3163	8881.625
14	9.078	1.410	16.110	2.467	14796	11103.741
15	9.513	0.852	19.458	2.238	3852	9885.960
16	7.886	0.517	18.334	1.463	12727	12946.915

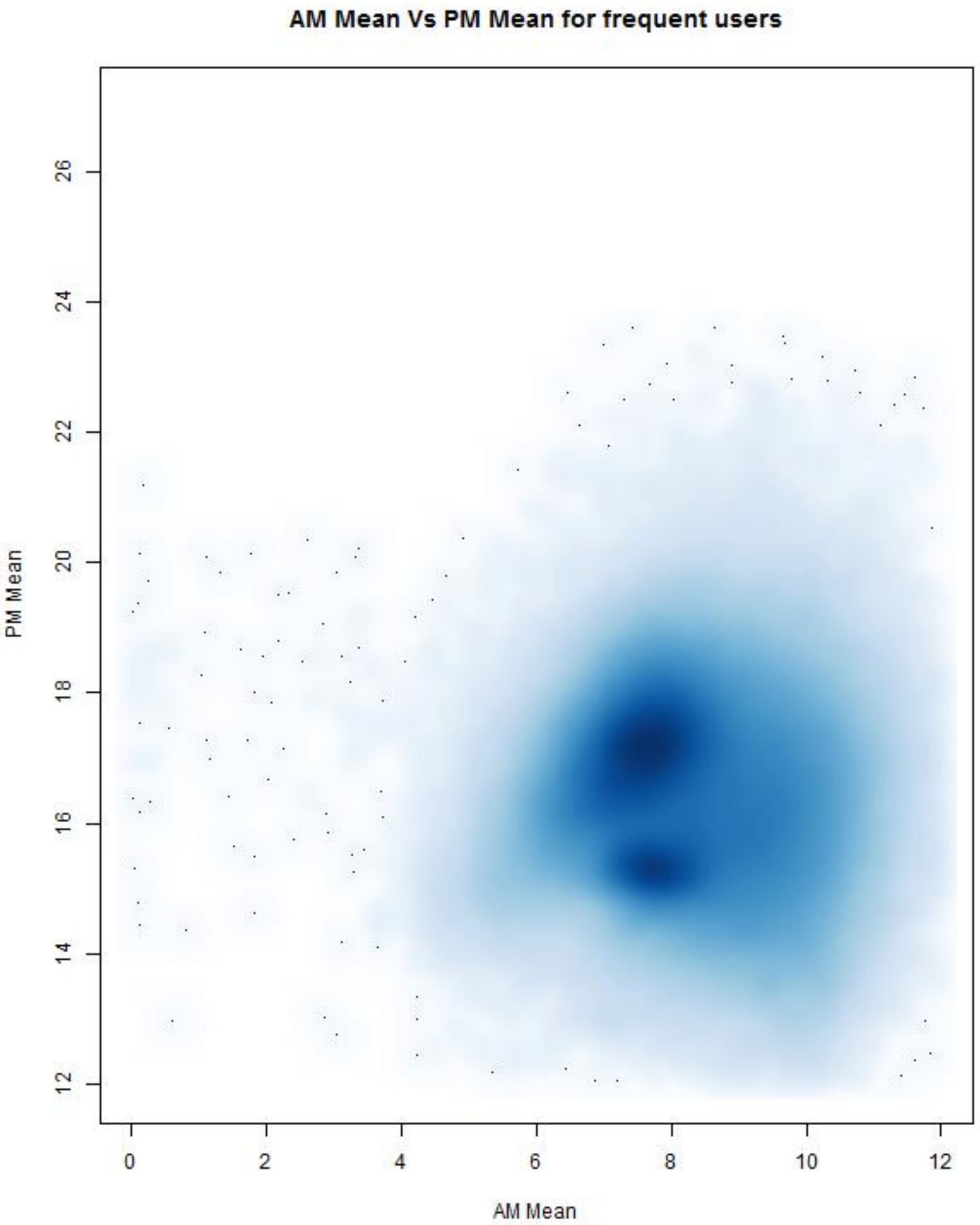
In order to interpret these results, it would be useful to see the variation of each of the 4 parameters with the other. Hence the following 12 plots (density plots and cluster-wise scatter plots for each of the 6 possible combinations)-

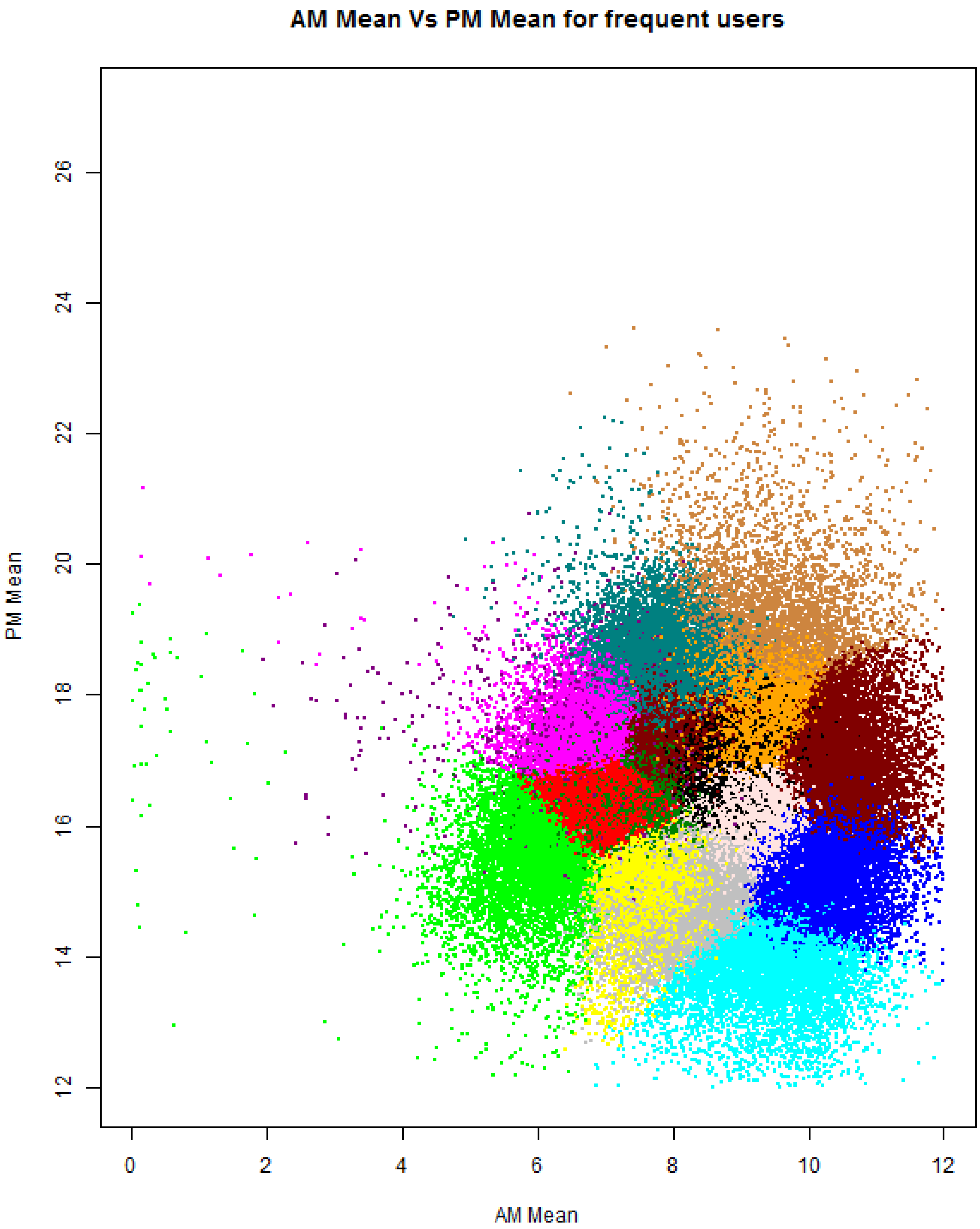


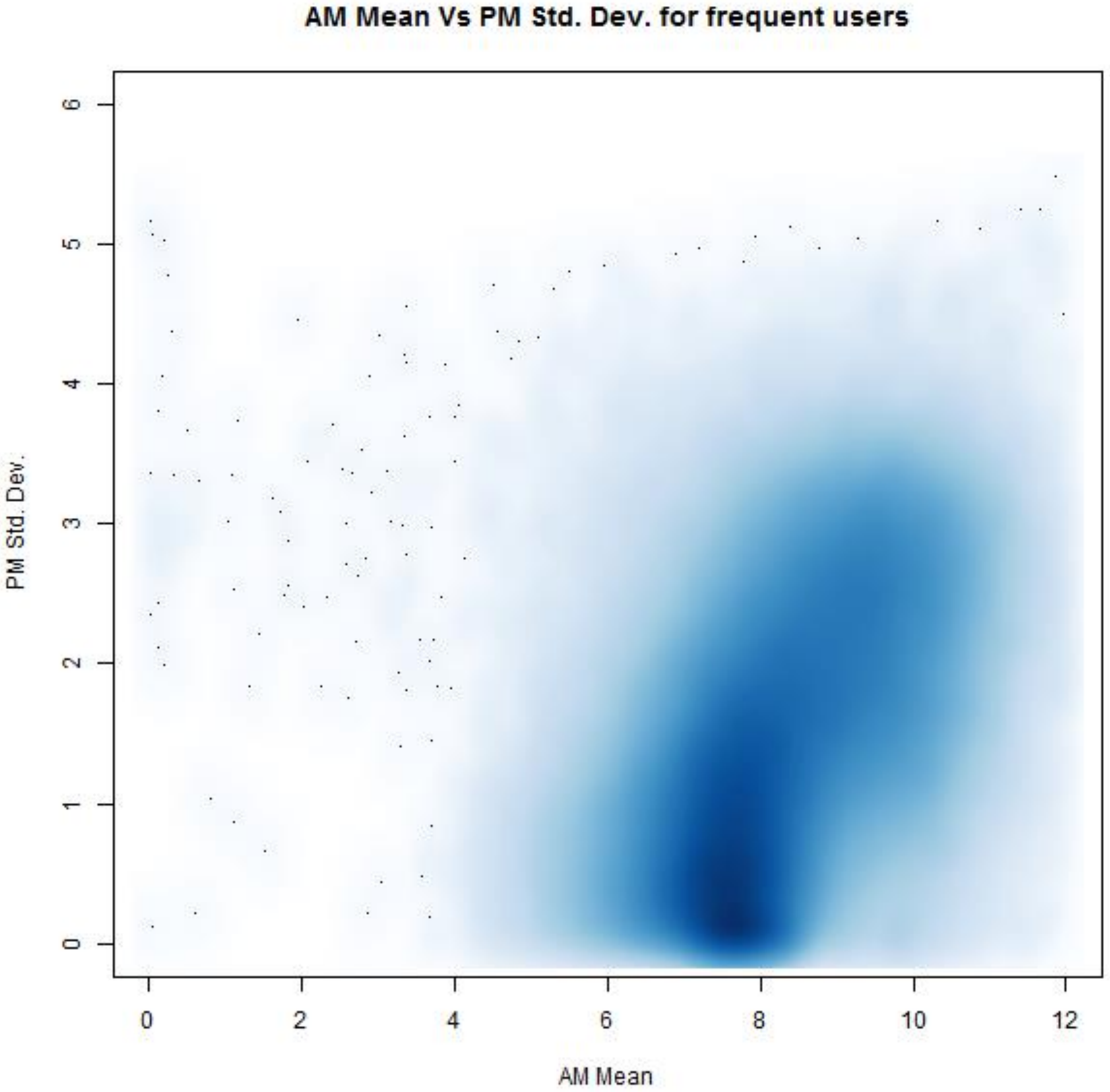

AM Mean Vs AM Std. Dev. for frequent users
AM Std. Dev.
0
1
2
3
4
5
6
0
2
4
6
8
10
12
AM Mean

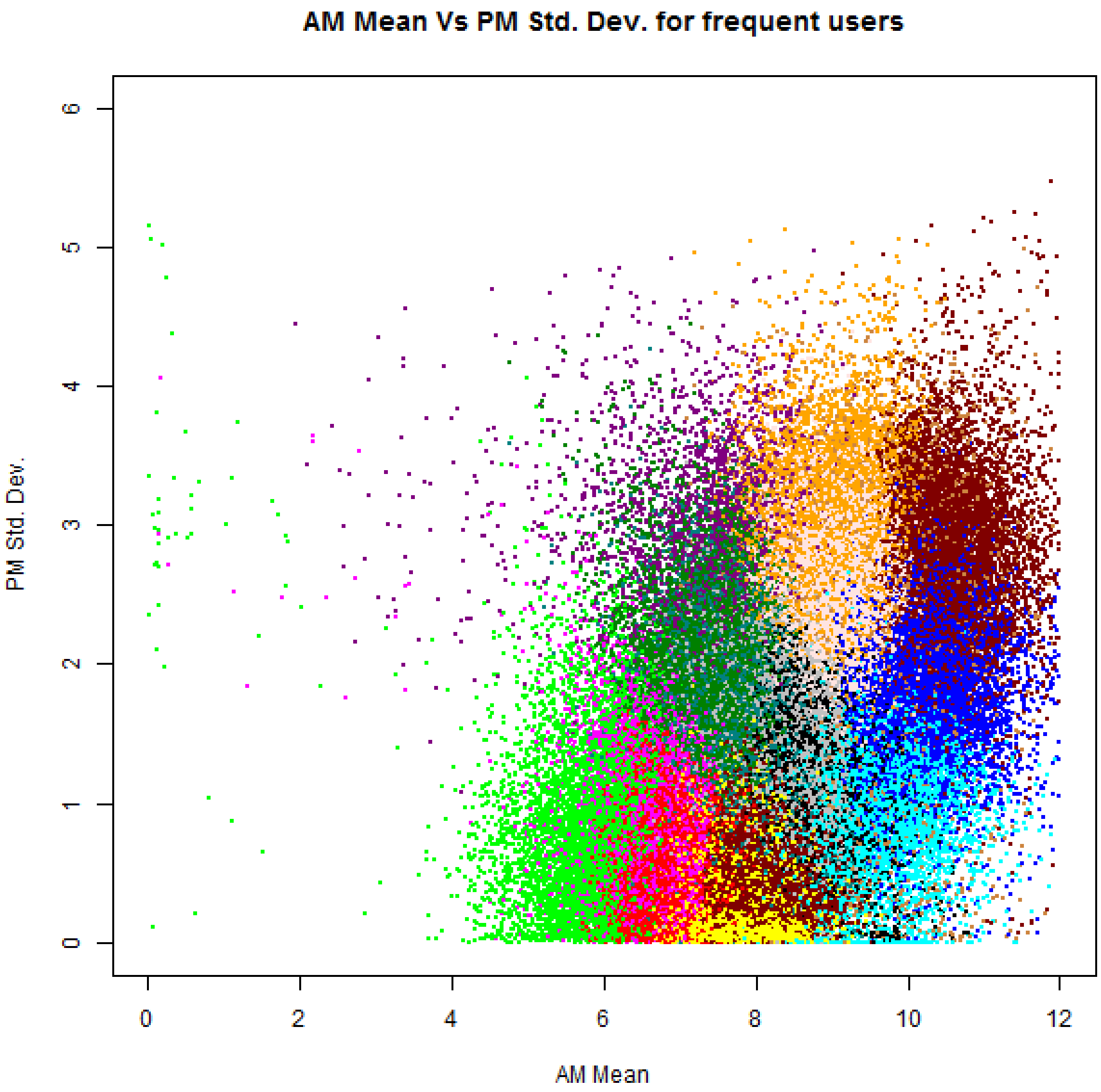

PM Mean Vs PM Std. Dev. for frequent users
PM Std. Dev.
0
1
2
3
4
5
6
PM Mean
12
14
16
18
20
22
24
26

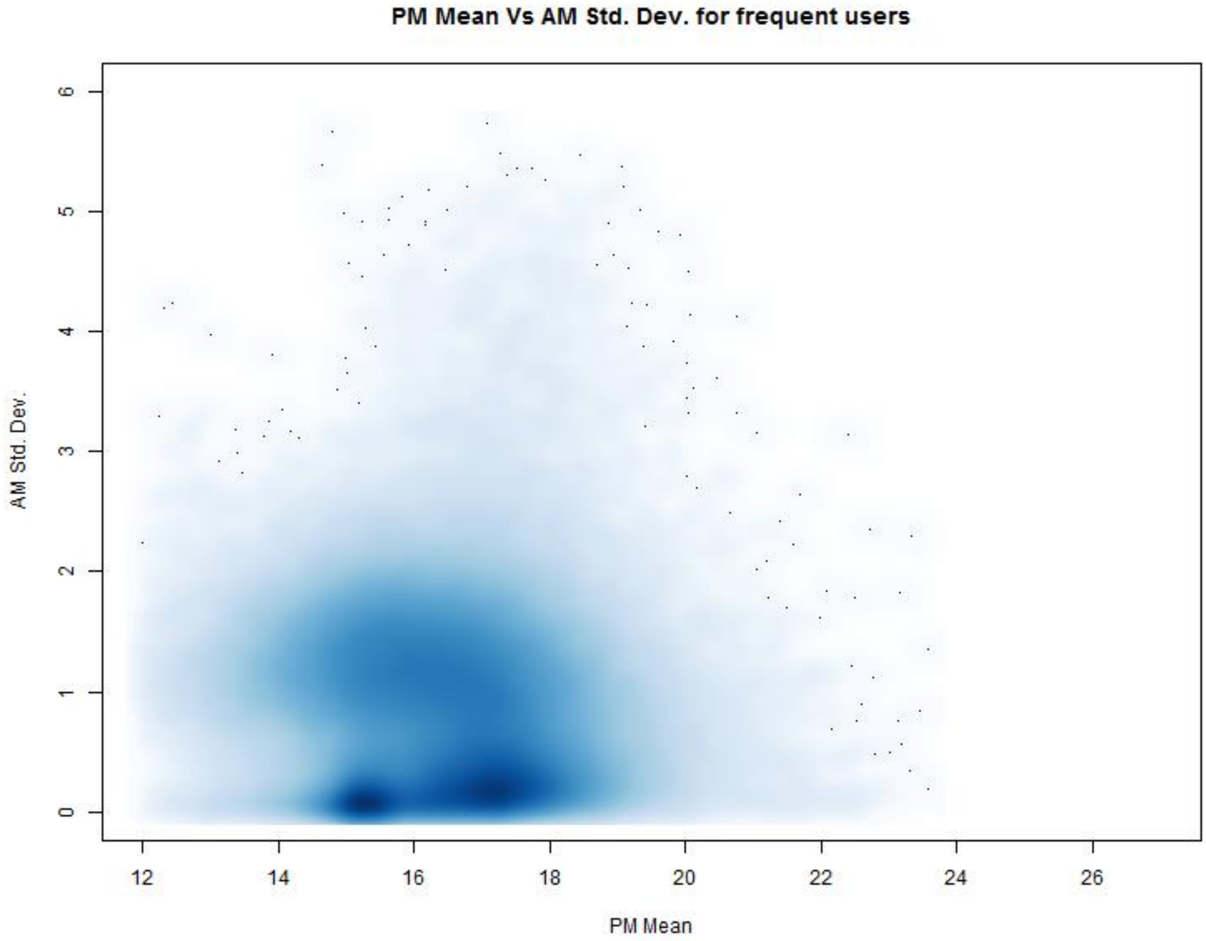

PM Mean Vs PM Std. Dev. for frequent users
PM Std. Dev.
0
1
2
3
4
5
6
PM Mean
12
14
16
18
20
22
24
26

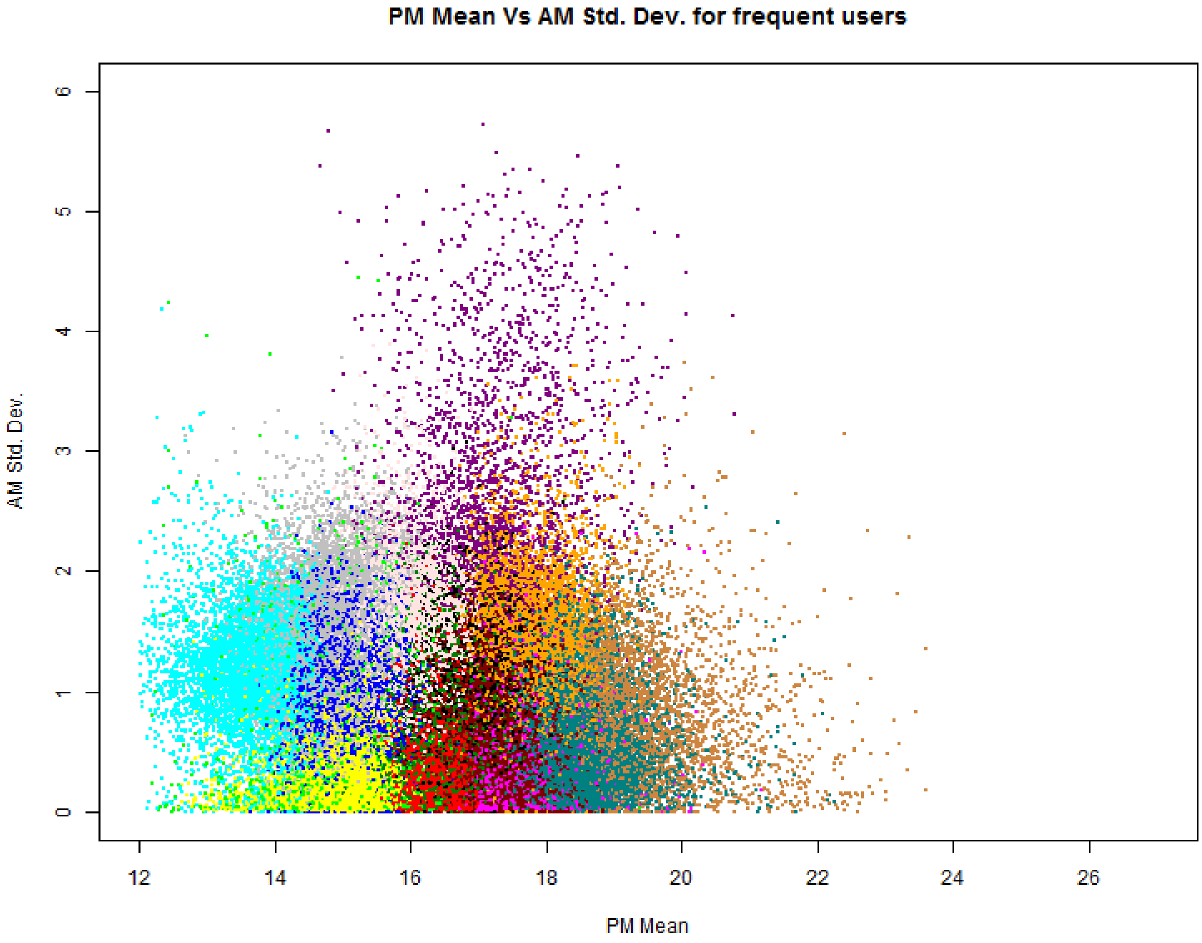

AM Mean Vs PM Mean for frequent users
PM Mean
26
24
22
20
18
16
14
12
0
2
4
6
8
10
12
AM Mean

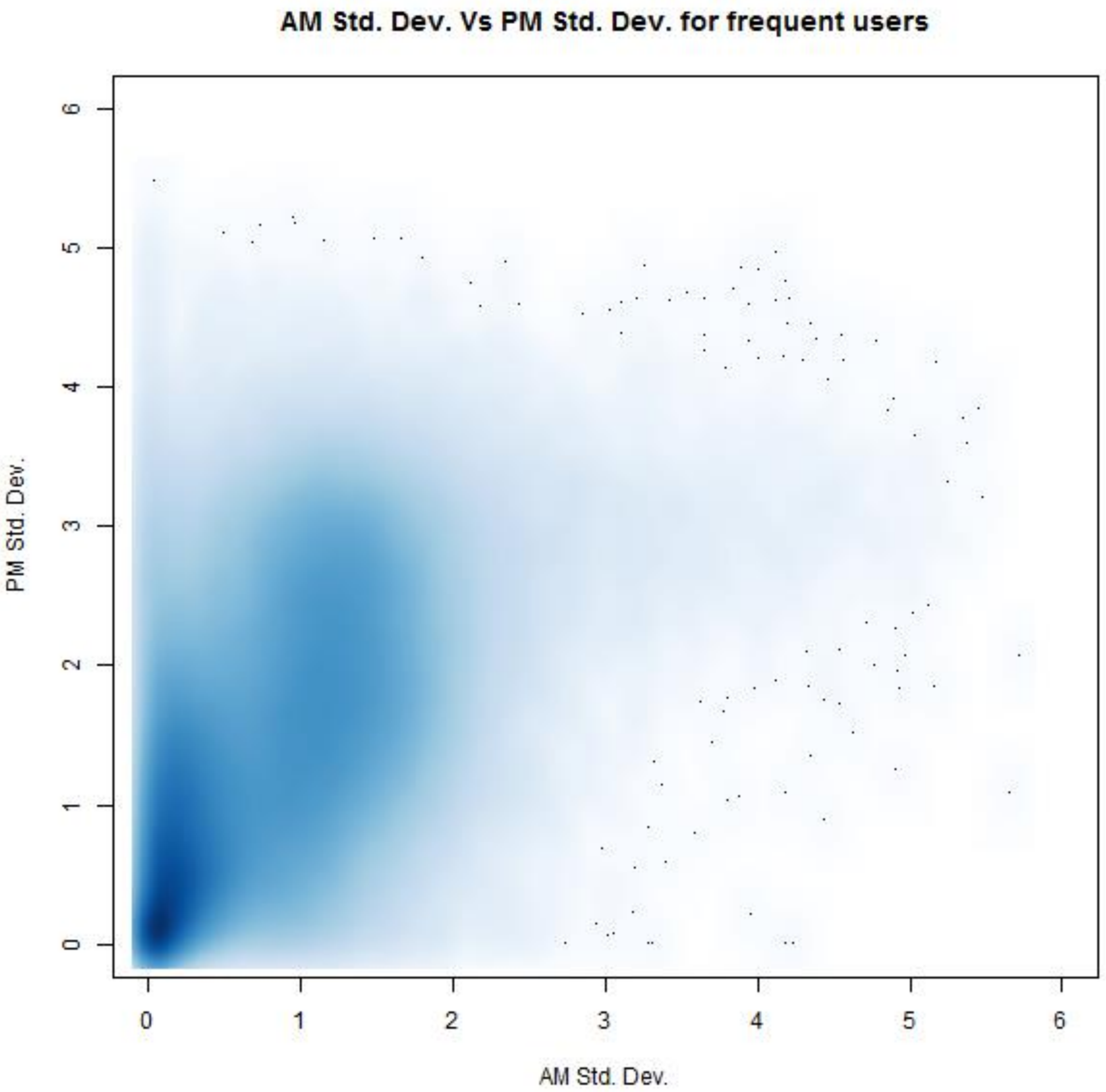

AM Mean Vs PM Mean for frequent users
PM Mean
12
14
16
18
20
22
24
26
AM Mean
0
2
4
6
8
10
12

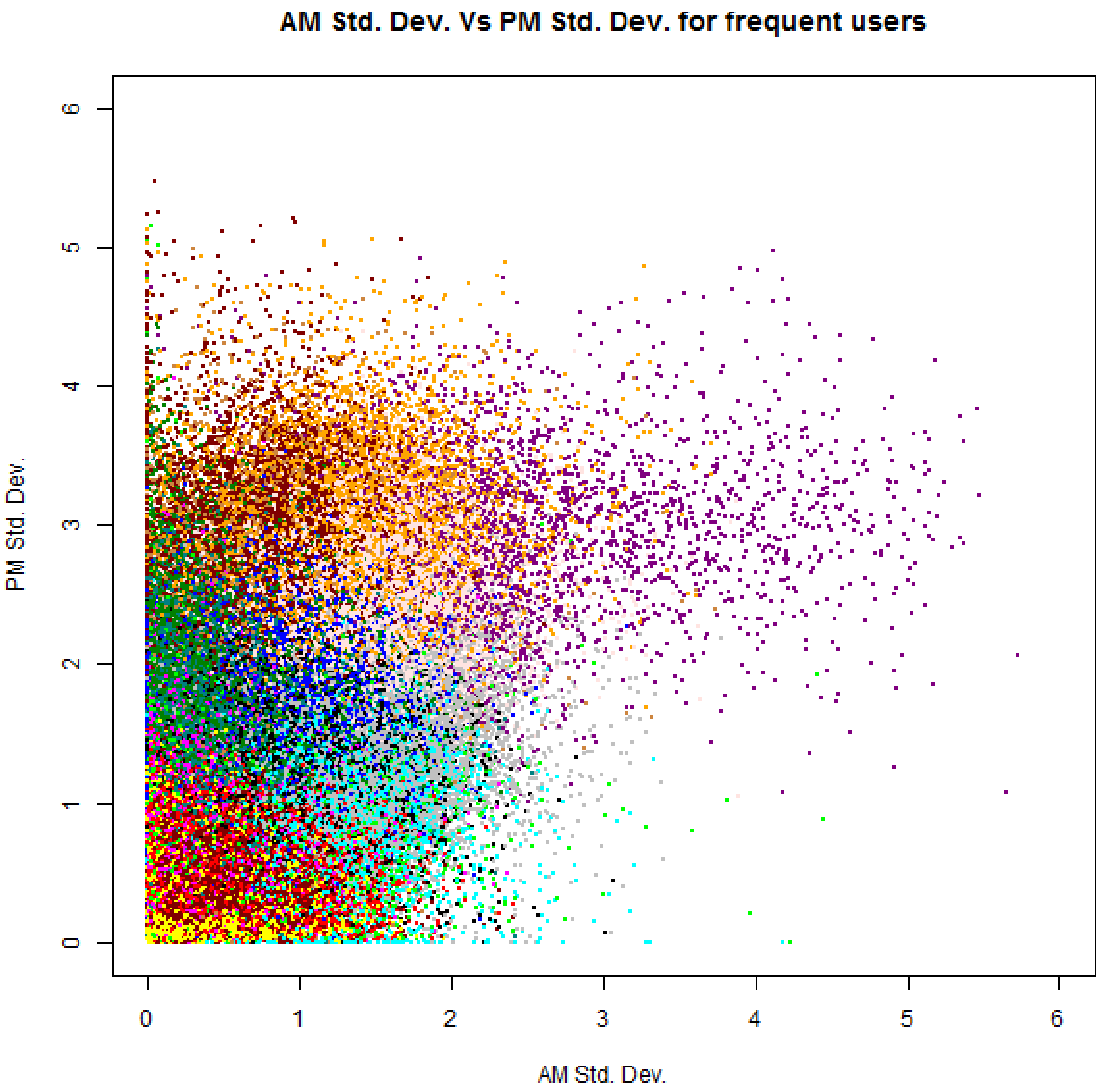

AM Mean Vs PM Std. Dev. for frequent users
PM Std. Dev.
0
1
2
3
4
5
6
AM Mean
0
2
4
6
8
10
12


AM Mean Vs PM Std. Dev. for frequent users
PM Std. Dev.
0
1
2
3
4
5
6
AM Mean
0
2
4
6
8
10
12


PM Mean Vs AM Std. Dev. for frequent users
AM Std. Dev.
0
1
2
3
4
5
6
12
14
16
18
20
22
24
26
PM Mean


PM Mean Vs AM Std. Dev. for frequent users
AM Std. Dev.
0
1
2
3
4
5
6
12
14
16
18
20
22
24
26
PM Mean


AM Std. Dev. Vs PM Std. Dev. for frequent users
PM Std. Dev.
0
1
2
3
4
5
6
AM Std. Dev.
0
1
2
3
4
5
6


AM Std. Dev. Vs PM Std. Dev. for frequent users
PM Std. Dev.
0
1
2
3
4
5
6
AM Std. Dev.
0
1
2
3
4
5
6

Analysis of the results:

From the plots shown above, some classes of users stand out immediately. As an example, the yellow cluster probably indicates the student community going to school from 8-3 with low standard deviations.
The 2 dense spots observed on most of the plots would probably indicate the school-going students and the office goers.

Most of the plots are observed to have a number of outliers with means and deviations which are not likely to be true for many users. This might be attributed to the fact that these users are making multiple trips in the AM and/or PM period.

Accounting for multiple AM/PM trips:

As multiple AM and PM trips can significantly affect the values for the mean and standard deviation for a user, it would be useful to quantify how many users are making multiple trips and what is the number of days in the month on which they make multiple trips.

The table below shows the count of the users based on the number of days that they performed multiple trips.

PM Repetitions			
No. of Days	No. of Users	%	Cum %
0	101268	52.644	52.644
1	34204	17.781	70.426
2	17715	9.209	79.635
3	11706	6.085	85.720
4	8352	4.342	90.062
5	5550	2.885	92.947
6	3973	2.065	95.013
7	2733	1.421	96.433
8	1971	1.025	97.458
9	1376	0.715	98.173
10	1001	0.520	98.694
11	713	0.371	99.064
12	513	0.267	99.331
13	362	0.188	99.519
14	273	0.142	99.661
15	216	0.112	99.773
16	125	0.065	99.838

AM Repetitions			
No. of Days	No. of Users	%	Cum %
0	122362	63.610	63.610
1	37710	19.604	83.214
2	14166	7.364	90.578
3	7092	3.687	94.265
4	4065	2.113	96.378
5	2355	1.224	97.602
6	1469	0.764	98.366
7	955	0.496	98.863
8	671	0.349	99.211
9	431	0.224	99.435
10	303	0.158	99.593
11	210	0.109	99.702
12	128	0.067	99.769
13	99	0.051	99.820
14	86	0.045	99.865
15	88	0.046	99.911
16	45	0.023	99.934

17	101	0.053	99.891
18	76	0.040	99.930
19	85	0.044	99.975
20	46	0.024	99.998
21	3	0.002	100.000

17	34	0.018	99.952
18	37	0.019	99.971
19	30	0.016	99.986
20	22	0.011	99.998
21	4	0.002	100.000

From the table above we can see that a large fraction of the users do not tend to make too many trip repetitions within the day. If we consider people with 4 or less instances of multiple AM or PM trips, we see that about 90% of the people had 4 or lesser PM repetitions and about 96.5% had 4 or lesser AM repetitions. These numbers are in keeping with the intuition that people would generally not prefer making multiple AM trips and have a higher chance of making multiple PM trips.

For the subsequent analysis, we shall consider that subset of users who make multiple trips on 4 or fewer days in the month.
Note: Though this might lead us into excluding the class of travellers who make multiple AM and/or PM trips, and we should probably come up with a way to include the people who frequently make multiple trips as well (Maybe by considering the additional trips as those made by a separate user), the current analysis does not go into this.

The intersection of those users who had 4 or fewer days with multiple trips in the AM with those who had 4 or fewer days with multiple trips in the PM period resulted in 167525 users (approximately 87.1%).

For these users, the days on which the multiple trips were made shall not be considered in the computation. Note that we might even go in with using the latter AM trip and the former PM trip for the purpose of analysis but this has been neglected for the time being.

Another factor that has to be considered in the analysis is the **definition of the AM and PM periods.**

With the traditional definition of the AM period, the trips with late-night timestamps (12am-3am approximately) would be considered as AM trips. But these are not essentially perceived as first trips of the morning.
Similarly, for the PM period, a large number of trips are lunchtime-trips. These should not be considered as they are not the 'returning home' trips.

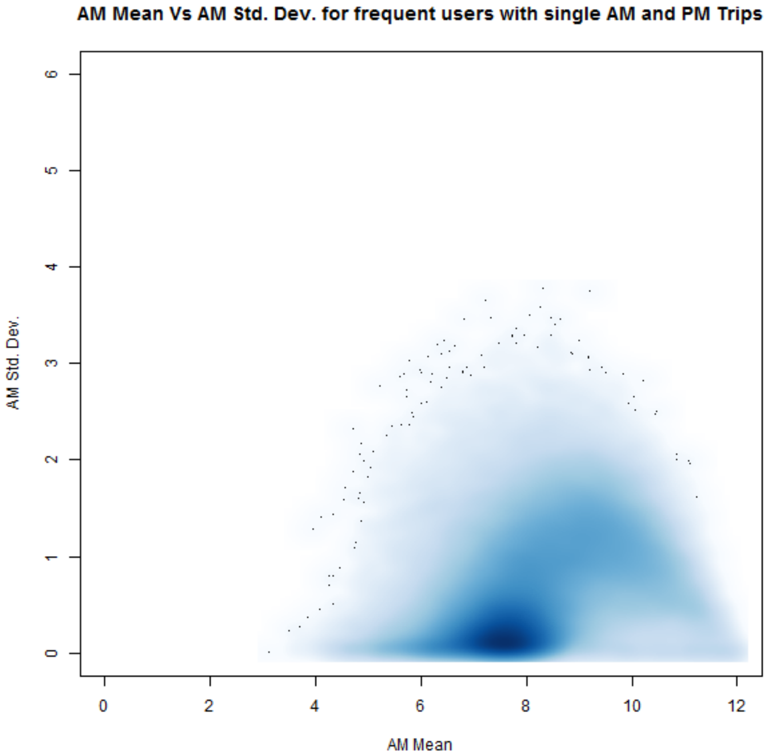
(Note- the 87.1% figure is associated with the newly defined AM and PM periods)

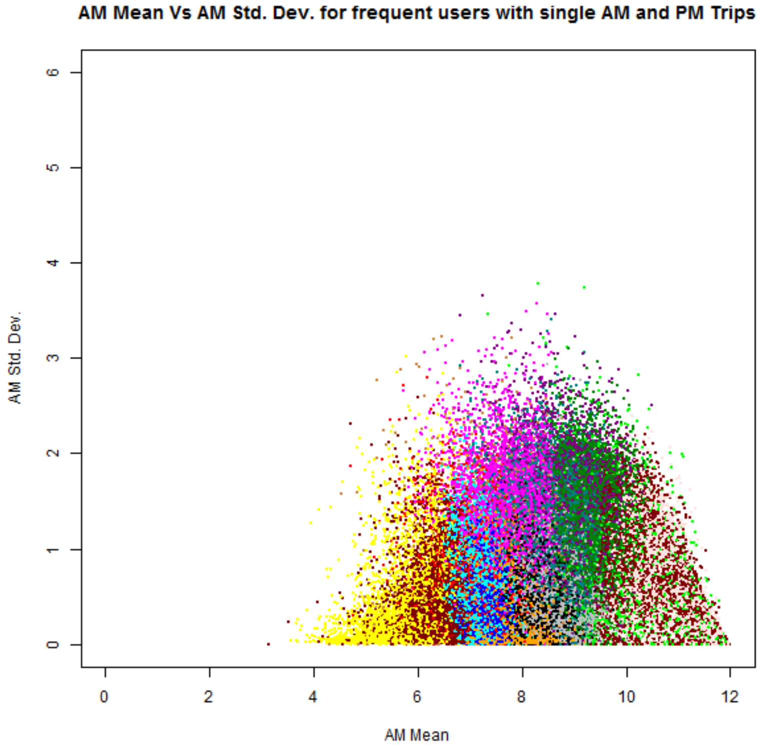
Hence, the definition of the AM period for further analysis is 3am to 12pm and that of PM period is from 2pm to 3am.

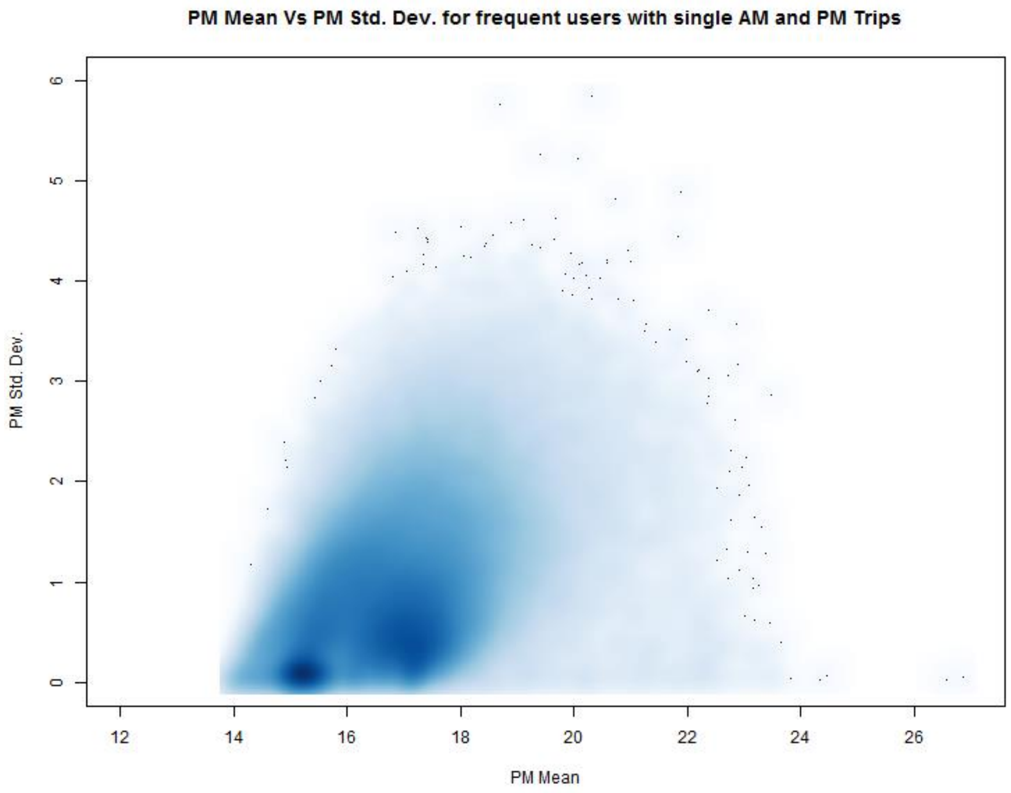
The results of the analysis with these two changes are shown below-

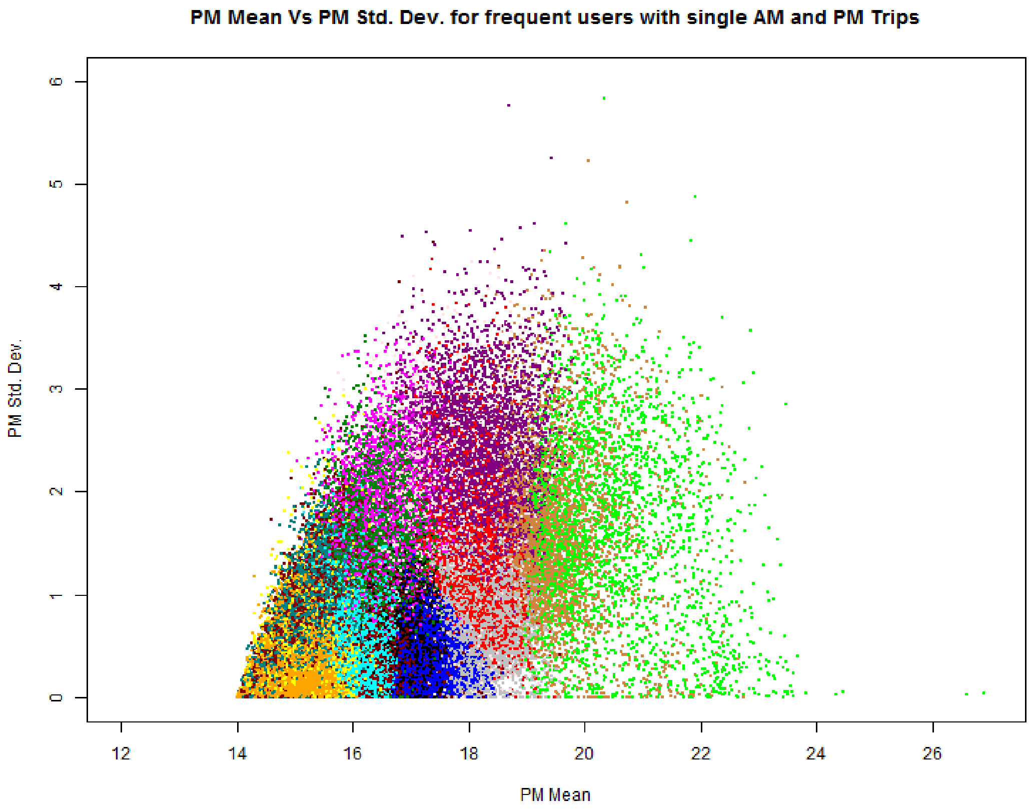
	AM Mean	AM Std. Dev.	PM Mean	PM Std. Dev.	# Points	Within SS
1	7.340	0.451	17.985	1.288	9600	7318.934
2	7.282	0.370	16.229	0.596	15640	7126.608
3	7.424	0.281	17.317	0.454	20615	6109.426
4	6.475	0.293	16.826	0.586	10593	6118.965
5	5.920	0.304	15.496	0.471	6524	5890.426
6	9.912	0.853	20.410	1.602	2887	8072.218
7	7.905	1.188	16.681	1.501	8790	8393.404
8	8.445	0.511	18.033	0.792	10578	6640.730
■	8.162	0.416	17.008	0.602	14850	5794.058
10	7.727	0.191	15.228	0.222	20480	7461.817
11	10.269	0.845	15.521	0.790	7147	6891.758
12	9.334	1.223	16.608	1.331	11268	8258.147
13	9.006	1.289	18.128	2.162	7170	8711.469
14	10.436	0.830	17.526	1.679	5847	7408.689
15	7.801	0.699	19.500	1.609	3632	6679.406
16	8.603	1.065	15.462	0.733	10585	8129.489

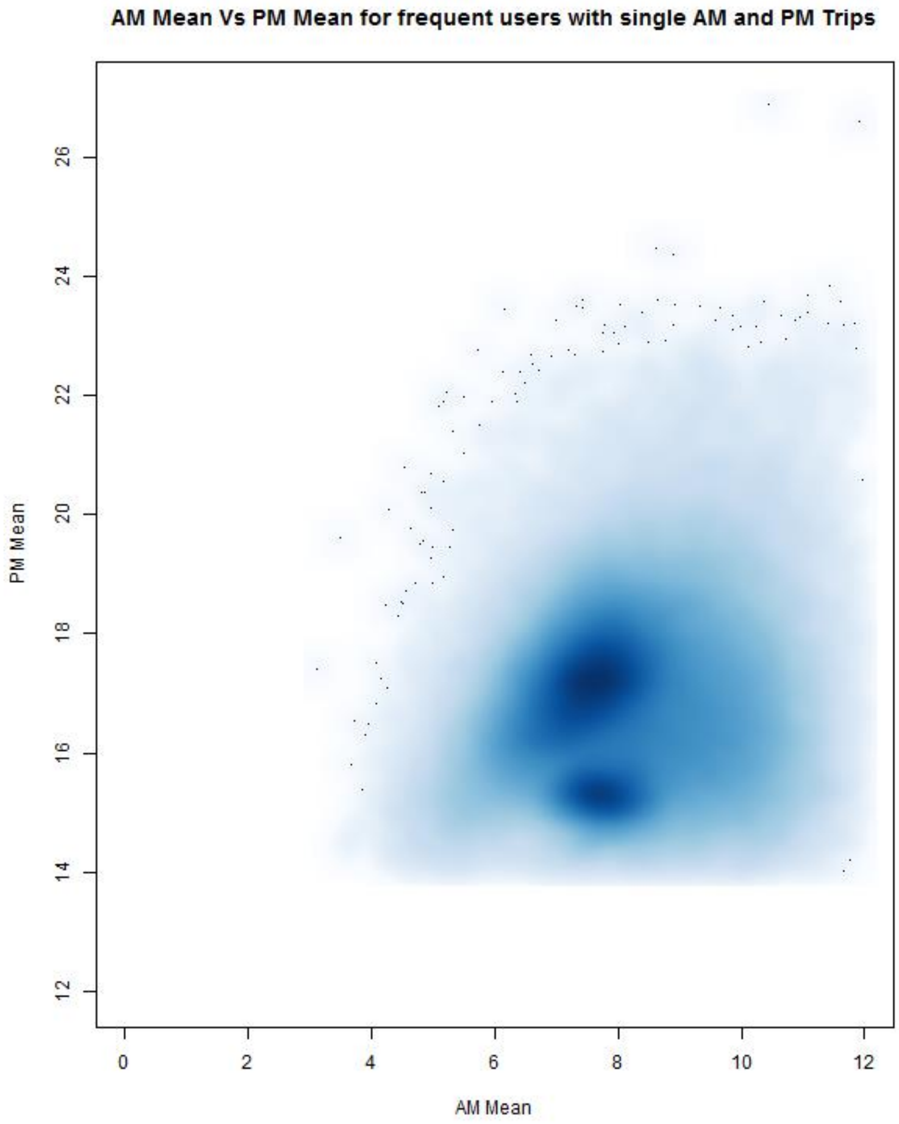
The density plots and the cluster-wise scatter plots are shown below.

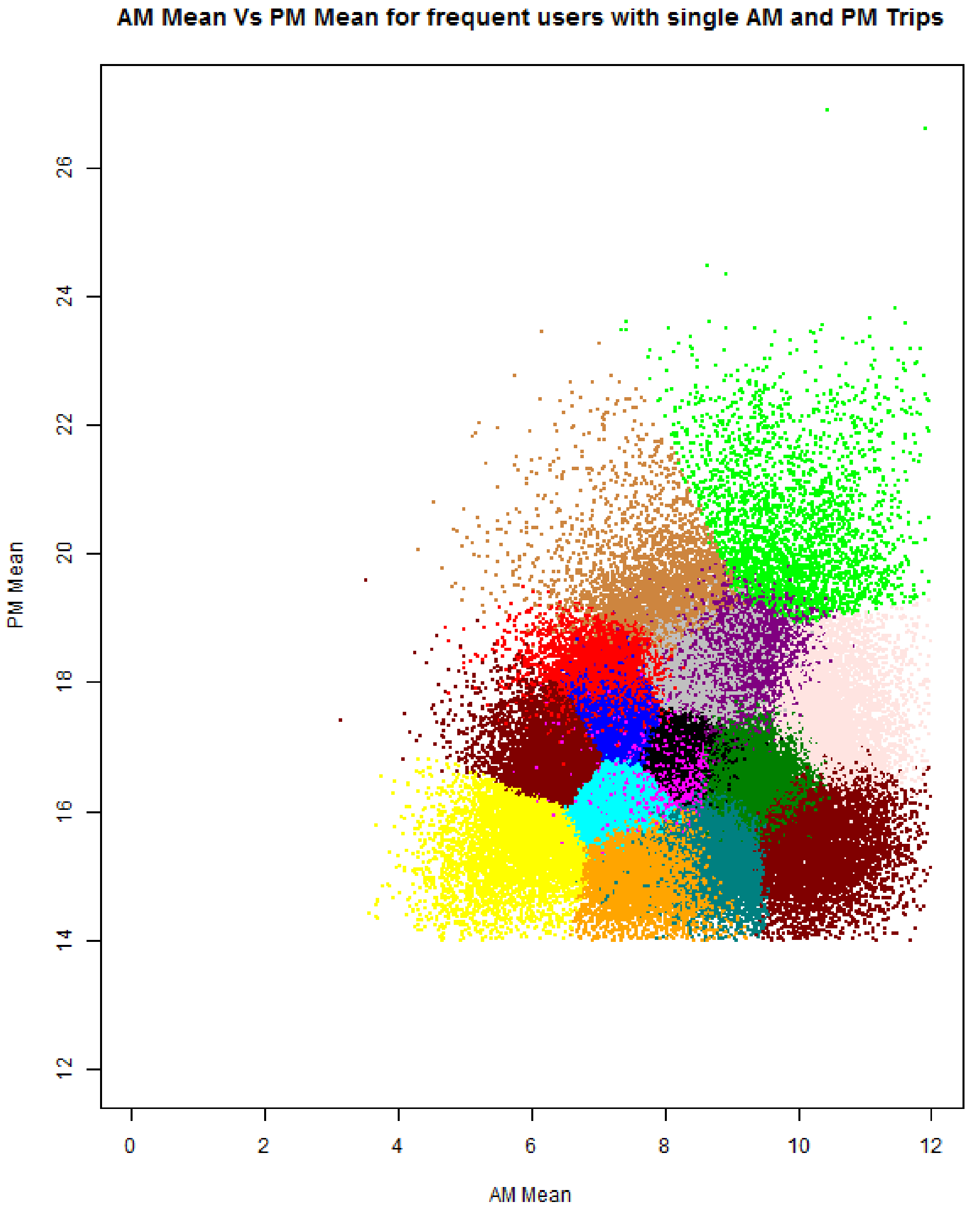

AM Mean Vs AM Std. Dev. for frequent users with single AM and PM Trips
AM Std. Dev.
0
1
2
3
4
5
6
0
2
4
6
8
10
12
AM Mean

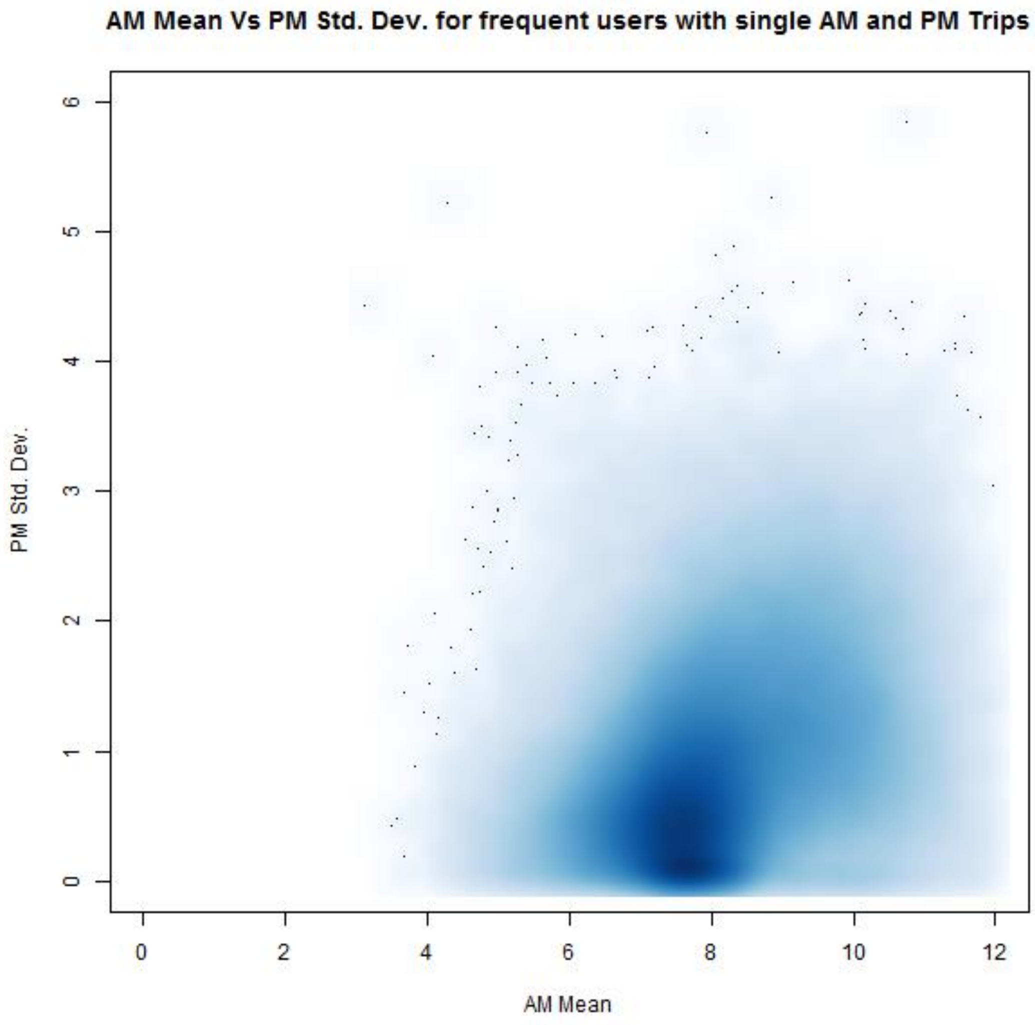

AM Mean Vs AM Std. Dev. for frequent users with single AM and PM Trips
AM Std. Dev.
0
1
2
3
4
5
6
0
2
4
6
8
10
12
AM Mean

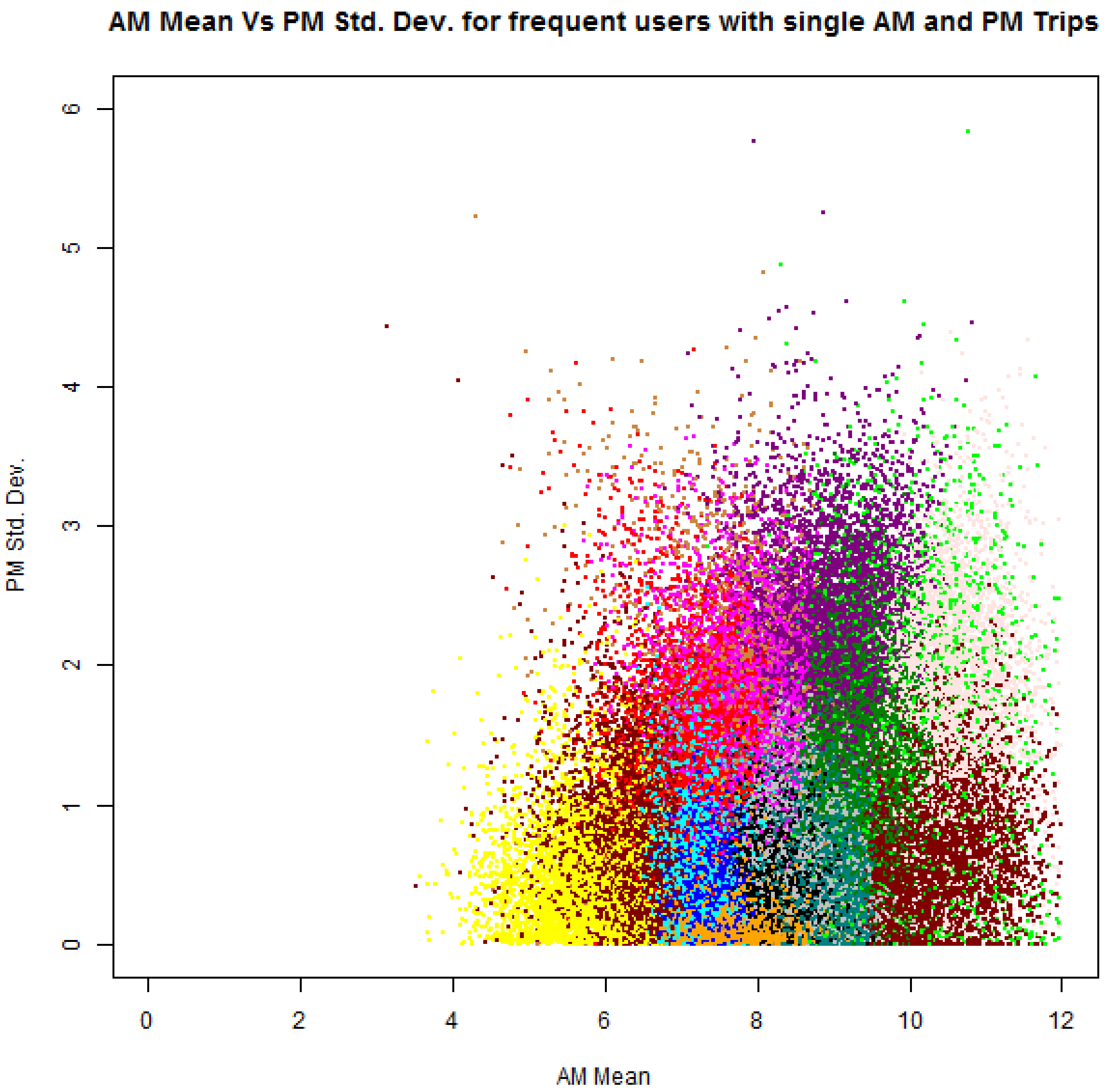

PM Mean Vs PM Std. Dev. for frequent users with single AM and PM Trips
PM Std. Dev.
0
1
2
3
4
5
6
PM Mean
12
14
16
18
20
22
24
26

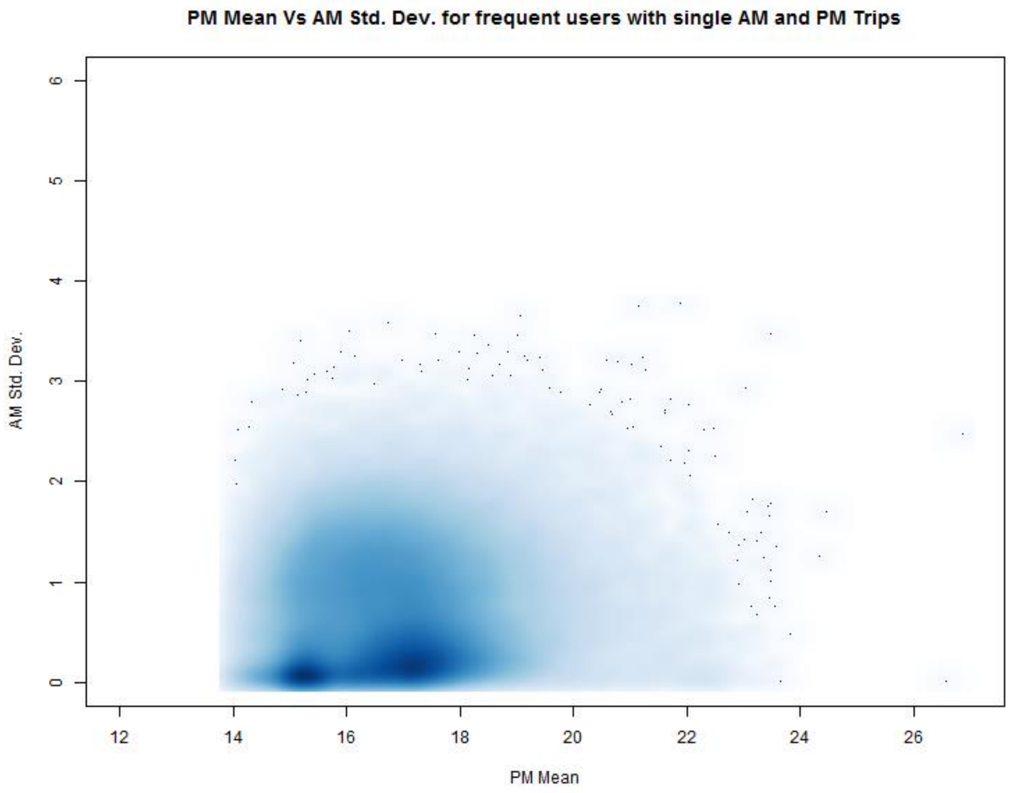

PM Mean Vs PM Std. Dev. for frequent users with single AM and PM Trips
PM Std. Dev.
0
1
2
3
4
5
6
PM Mean
12
14
16
18
20
22
24
26

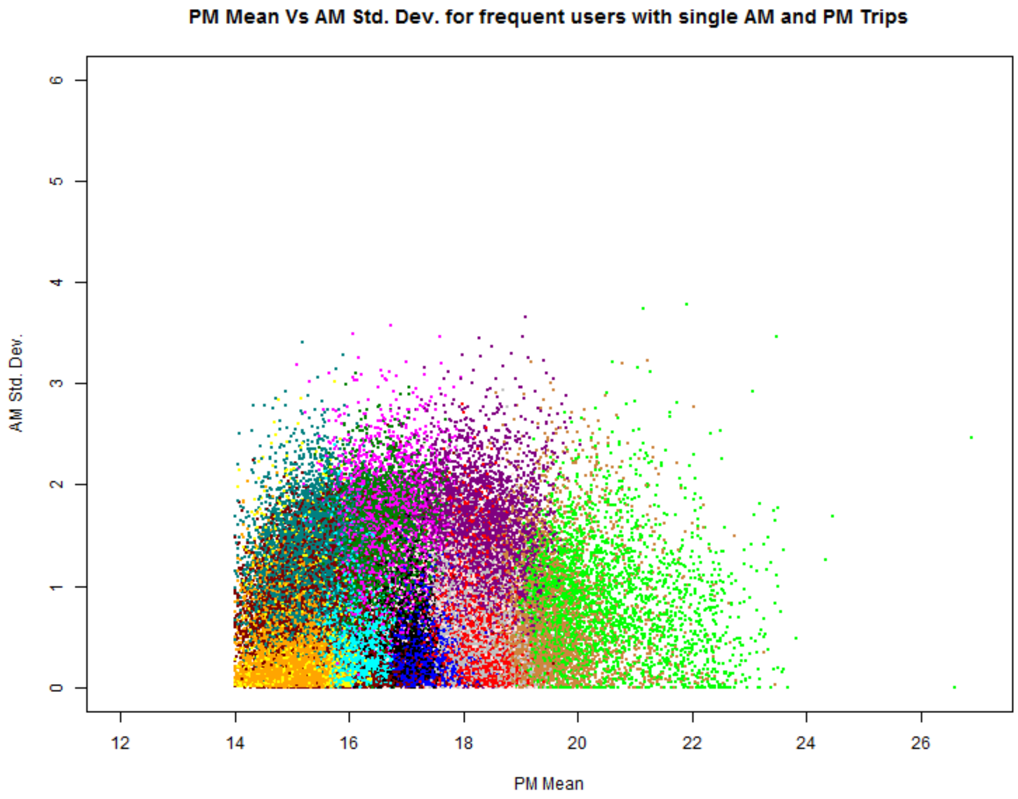

AM Mean Vs PM Mean for frequent users with single AM and PM Trips
PM Mean
12
14
16
18
20
22
24
26
AM Mean
0
2
4
6
8
10
12

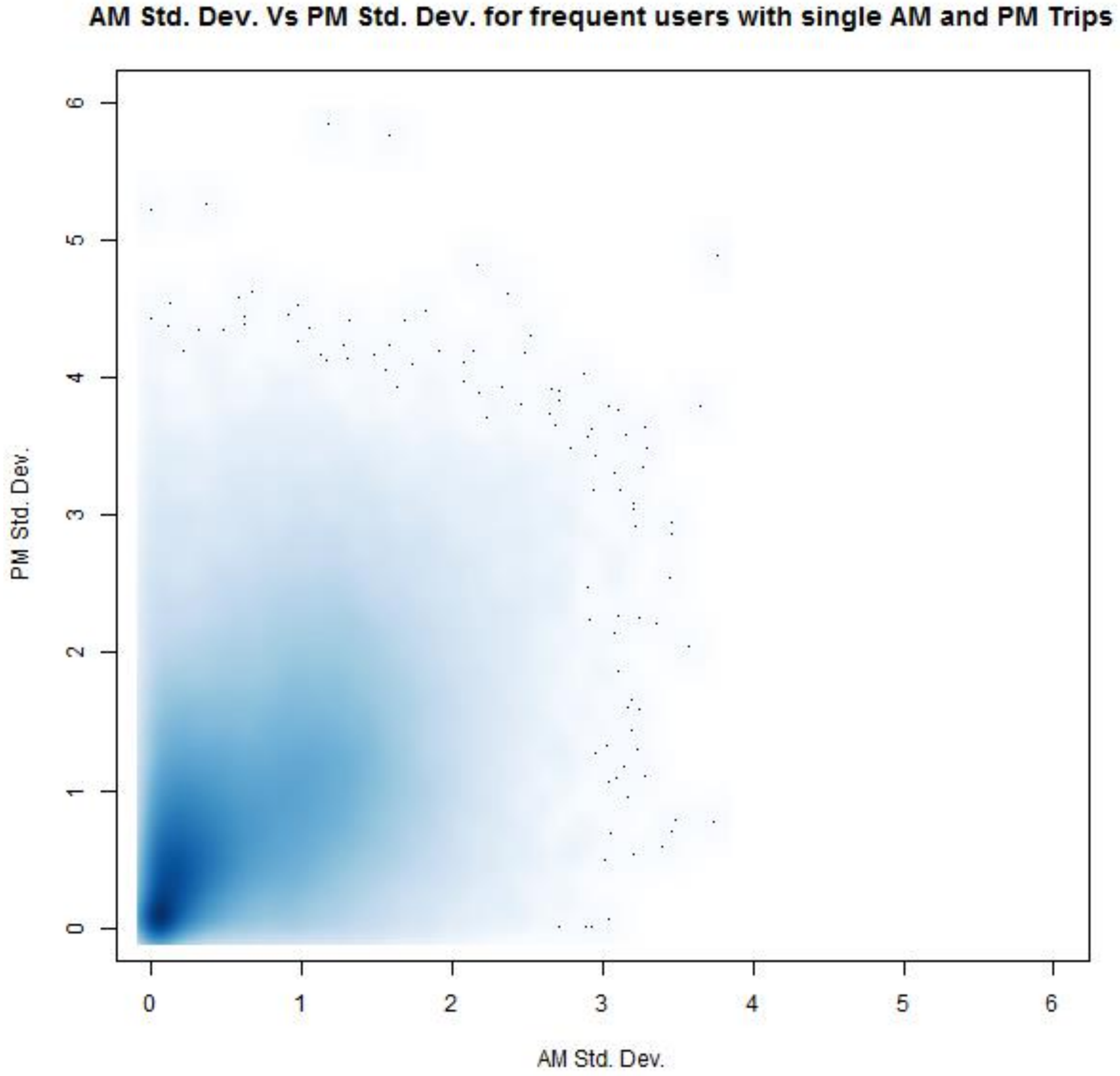

AM Mean Vs PM Mean for frequent users with single AM and PM Trips
PM Mean
12
14
16
18
20
22
24
26
AM Mean
0
2
4
6
8
10
12

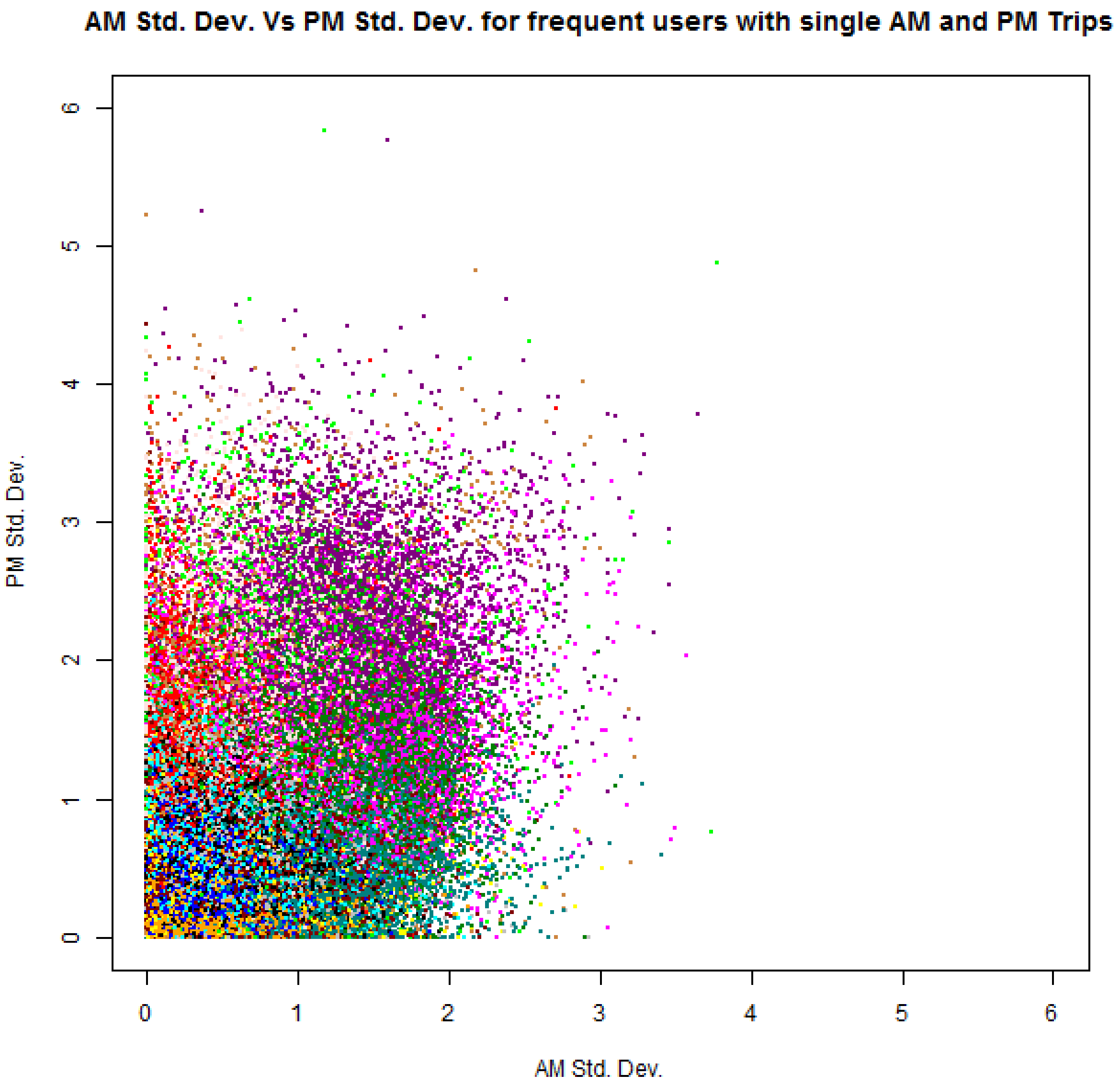

AM Mean Vs PM Std. Dev. for frequent users with single AM and PM Trips
PM Std. Dev.
0
1
2
3
4
5
6
AM Mean
0
2
4
6
8
10
12


AM Mean Vs PM Std. Dev. for frequent users with single AM and PM Trips
PM Std. Dev.
0
1
2
3
4
5
6
AM Mean
0
2
4
6
8
10
12


PM Mean Vs AM Std. Dev. for frequent users with single AM and PM Trips
AM Std. Dev.
0
1
2
3
4
5
6
PM Mean
12
14
16
18
20
22
24
26


PM Mean Vs AM Std. Dev. for frequent users with single AM and PM Trips
AM Std. Dev.
0
1
2
3
4
5
6
PM Mean
12
14
16
18
20
22
24
26


AM Std. Dev. Vs PM Std. Dev. for frequent users with single AM and PM Trips
PM Std. Dev.
0
1
2
3
4
5
6
AM Std. Dev.
0
1
2
3
4
5
6



AM Std. Dev. Vs PM Std. Dev. for frequent users with single AM and PM Trips

Closing Remarks

The results for the second analysis also show 2 distinct high density regions which might be attributed to the school students and the traditional office goers.

This analysis opens a variety of interpretations with regards to the different clusters. The use of the analysis depends on the objective of the study for which the results are referred.

Further interpretations of the results shall follow in a subsequent report.

Another useful exercise could be to analyze the results with fewer or greater number of clusters and hence make out the major and minor clusters.

Possible improvements-

1. The number of parameters being studied can be increased to encompass a broader range of transit characteristics. For eg. the average trip length or the travel time can help us say that a particular group of people had 'x' travel time at a mean time 't' in the AM period with a deviation of 'd'.
2. Differential weightage- a difference of some amount in the mean time of two users might not have the same implication as the difference on the same amount in the standard deviation. Hence it might be a good idea to define different multiplication factors for the means and the standard deviations.
3. Further, the set of users being analyzed can be broadened to encompass less frequent users and users with multiple trips as well.