Questions answers

VLADIMIR DOBRODEEV

Q1.1: How many different apps contain the dataset?

There are 1543 different applications

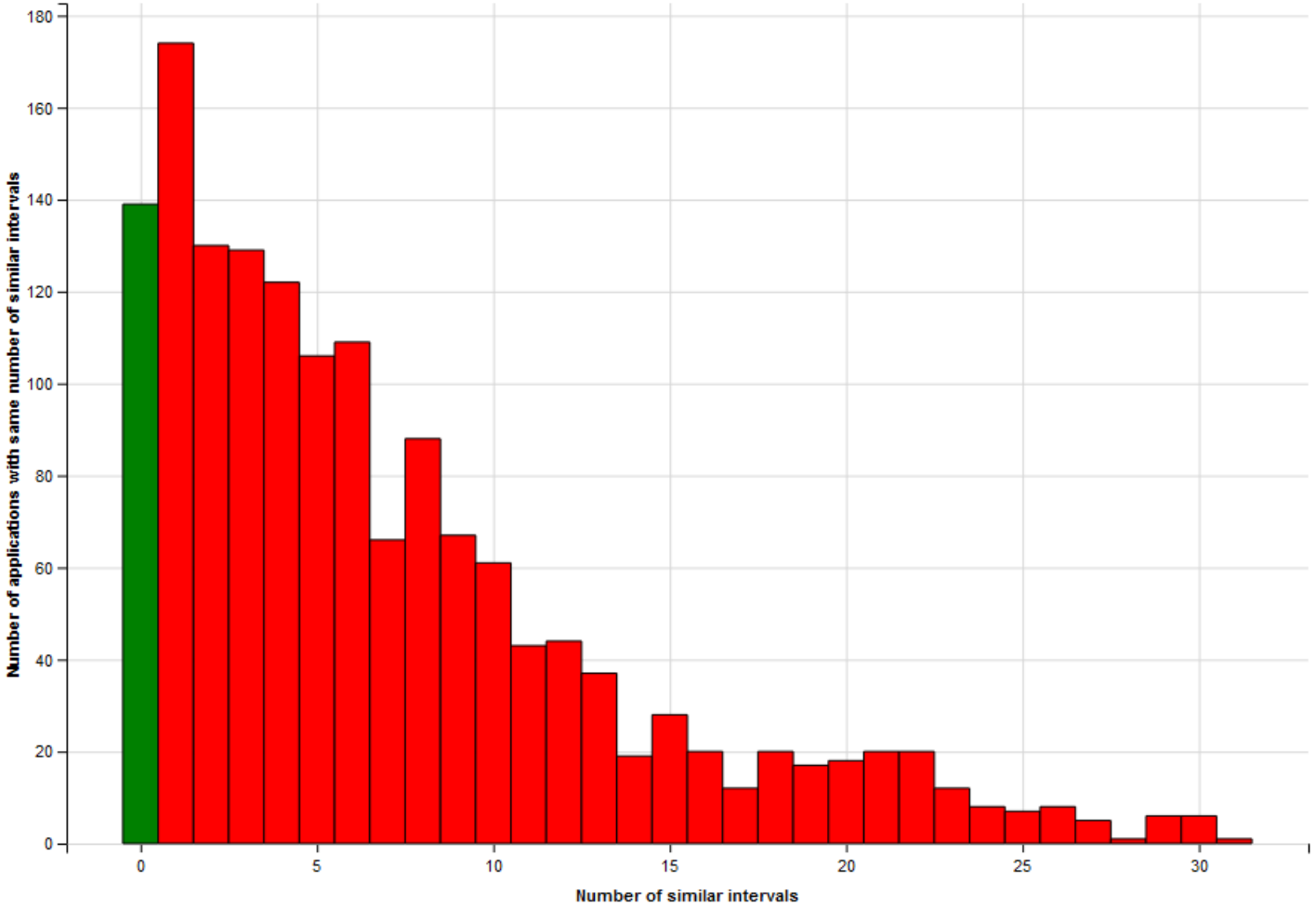
Q1.2: What is the max and min data volume interval of the dataset?

Application 1988 has the largest data volume interval: [375, 11954282]

The following applications has the minimal interval (zero interval):

Application	Interval
939	[717, 717]
1079	[1194, 1194]
1080	[1205, 1205]
1478	[537, 537]
1586	[53451, 53451]
1780	[1444, 1444]
1810	[1496, 1496]
1899	[86440, 86440]

Q1.3: How many apps share similar data volume interval?



First approach:

- **Interval is similar, when minimum and maximum values of interval differ less, than by 10%**
- **E.g. applications 12: [136, 720630] and 145: [136, 722586] are quite similar**
- **The graphics shows, how many applications have same number of similar intervals**
- ***The answer is: 1404 applications share one or more similar intervals***
- **E.g. 6 applications has 30 similar intervals**

Q1.3

The second approach was to measure overlaps between application intervals

Applications 1: [136, 11903890] and 10: [204, 568323] has interval overlaps with all applications in the set

There is no applications, which intervals do not overlap

Q1.4: If two apps have similar data volume interval, how can you make them different during the analysis

The approach, that I tried was to take the maximal possible packet and change all traffic volumes in the following way:

traffic_volume = app_id * max_traffic_volume + traffic_volume

If we analyze this concrete set of data, this approach provides very good result width 99,999% accuracy

Most likely, this approach is not applicable for analyzing other sets of data

Q1.5: What is the max and min amount of samples for a particular app in the dataset?

For app 2 samples number is 1028845

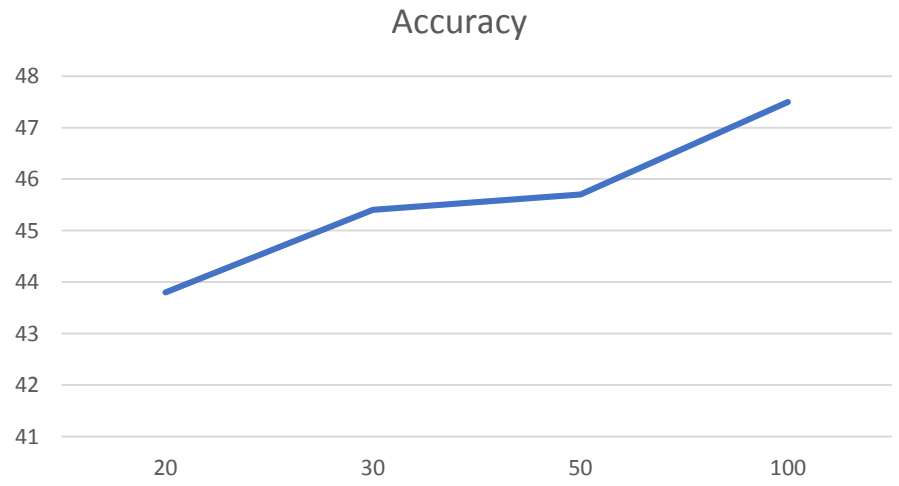
For applications 939**,** 1079, 1080, 1478, 1586, 1780, 1810, 1899 there is only one sample

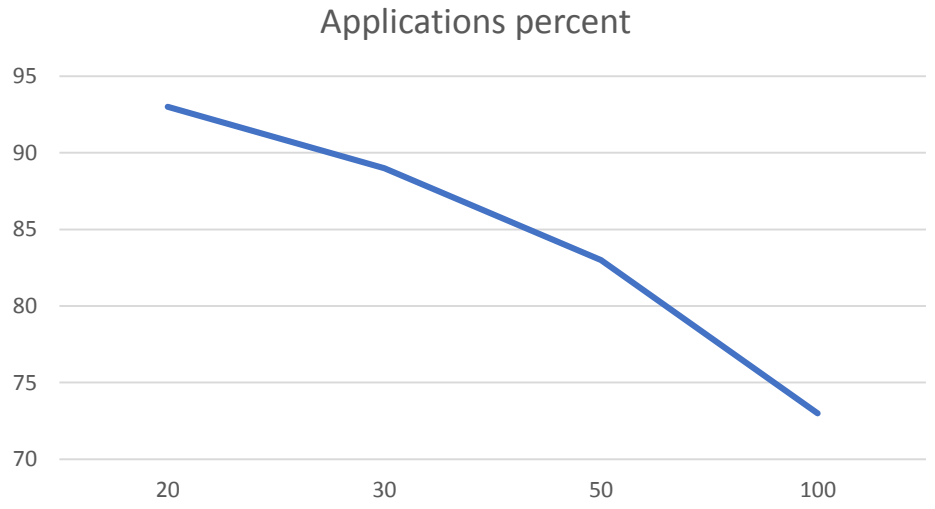
Q1.6: What is the min amount of samples needed for an app, such that it can be considered in the analysis?

This number was chosen in a way to provide the most accurate analysis with reasonable time

Finally it was chosen to use applications with 30 samples minimum

This number was selected to provide a good enough accuracy with larger percent of application involved in analysis





Q1.7: How many apps fulfil the previous requirement?

1387 applications fulfil this requirement

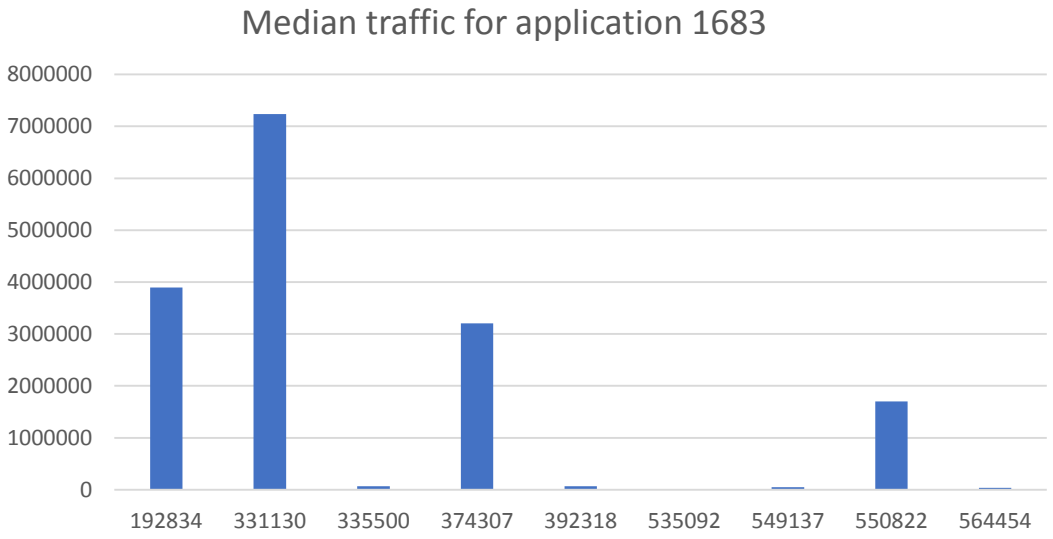
Q1.8: Is there any relation between app usage and location (base station)?

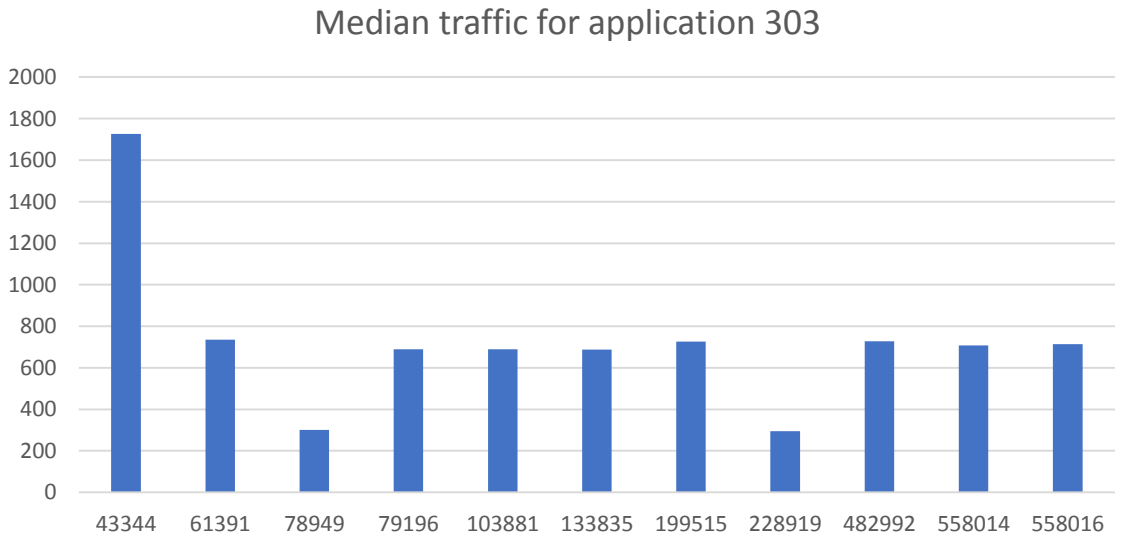
Yes, there is a relation for majority of application: 91,4% of applications show at least some difference in medians through different base station. For 88,4% this deviation is larger than minimal possible packet

However, for some applications difference in behavior is not significant

The chart, showing variance of median traffic volume per application:

Q1.8


Median traffic for application 1683
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
192834
331130
335500
374307
392318
535092
549137
550822
564454


Median traffic for application 303
2000
1800
1600
1400
1200
1000
800
600
400
200
0
43344
61391
78949
79196
103881
133835
199515
228919
482992
558014
558016

Q1.9: How many different users has the dataset?

There are 998 users in dataset

Q1.10: Is there any relation between app usage, location (base station) and a particular?

Yes, it is

1410 application observe deviations for different users at different locations

However, same as in Q1.8 is applicable here

Q1.11: Is there any relation between app usage, location and time?

Yes, it is

1535 (all, except apps having 1 sample) observed some difference

However, same as in Q1.8 is applicable here

Comparison of analysis result with respect to different field combinations

Accuracy	Traffic volume	Base station	User	Time (as is)	Time (minute)	Time (hour)
45,4%	+	-	-	-	-	-
81,7%	+	+	-	-	-	-
86,4%	+	-	+	-	-	-
86,7%	+	+	+	-	-	-
88,3%	+	+	+	+	-	-
88%	+	+	+	-	+	-
87,2	+	+	+	-	-	+
99,999%	+ (if changed)	-	-	-	-	-

Q2.1: How significant is the data transferred rate of an app for distinguish it among the rest?

To answer this question distributions densities of application samples were compared

t. test was used to distinguish traffic volume distributions and, therefore, only applications with more than 30 samples were chosen (Mann-Witney-Wilcox test might better fit here)

Null hypothesis was that population means of two distribution are same (i.e. two-sided test was employed), and, therefore, alternative hypothesis is that they are not same

All samples were compared to application 2 (as it has the largest number of samples and second highest traffic rate) with different significance levels

Minimal significance level at which alternative hypothesis is true for all apps should show how significant traffic rate is to distinguish apps

The result was, that the maximal significance with which apps are distinguishable is **0,005%** $(\alpha = 0{,}9975)$

Q2.2: Is it possible to categorize application usage based on data transferred rate?

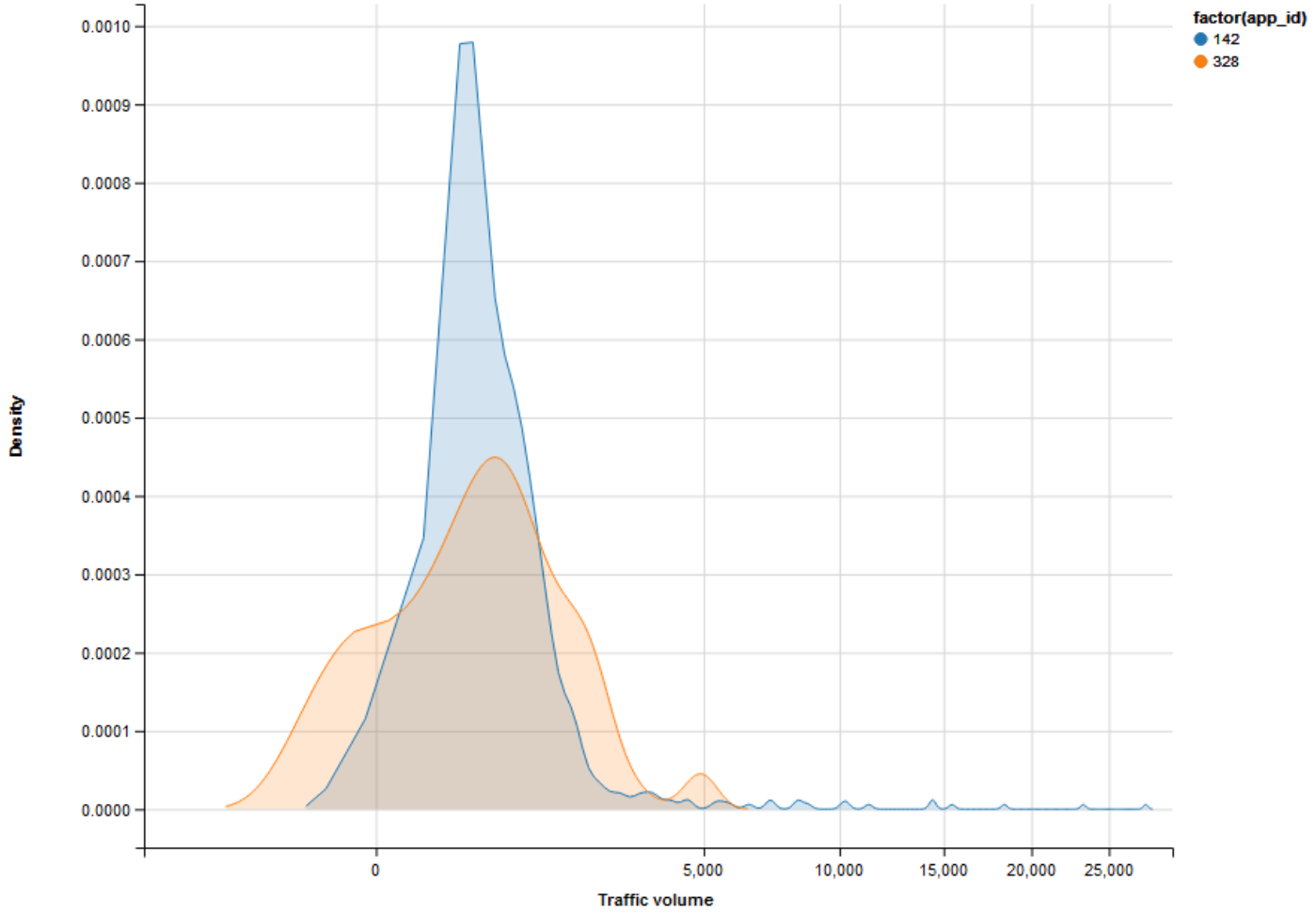
An one approach might be to compare different applications behavior at different hours and put application with same behavior to the same category

As a behavior characteristic we might consider mean traffic; so categorization might look like “we are 95% confident that these two apps share same mean at given hour”

As an example 14:00 was taken as it has the largest number of samples

Categorization was done with 30 samples and 95% confidence

The answer is that it is possible to categorize applications with this criteria



Q2.2

Example of category: applications 142 and 328

Q2.3: What is the minimal amount of samples to characterize data transferred rate?

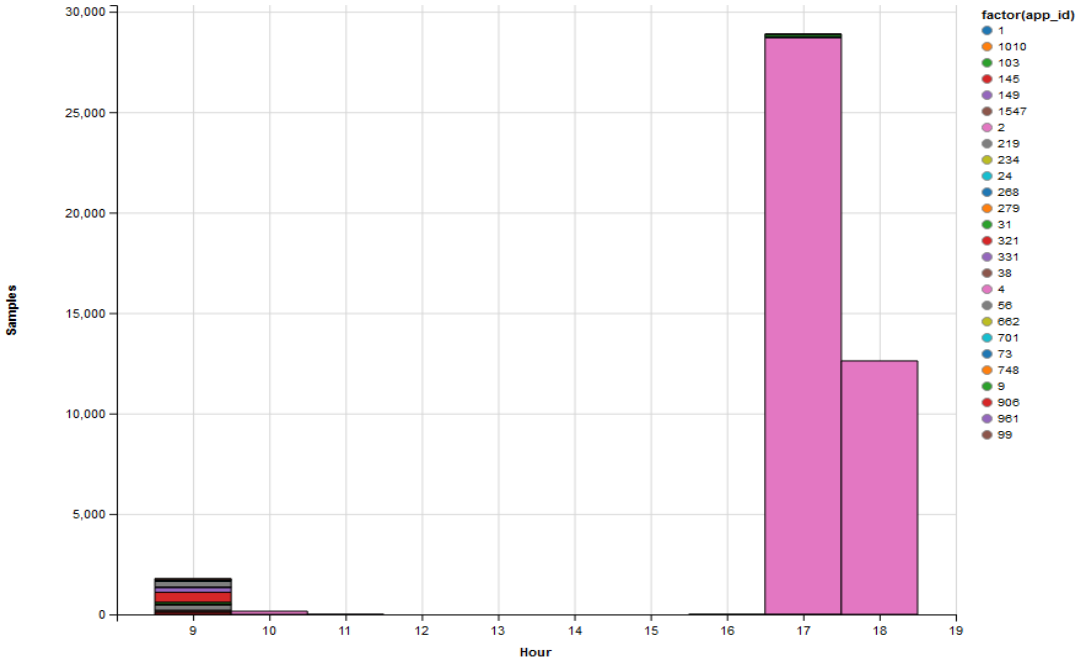
The following table presents minimal number of samples, that allows to distinguish all given applications

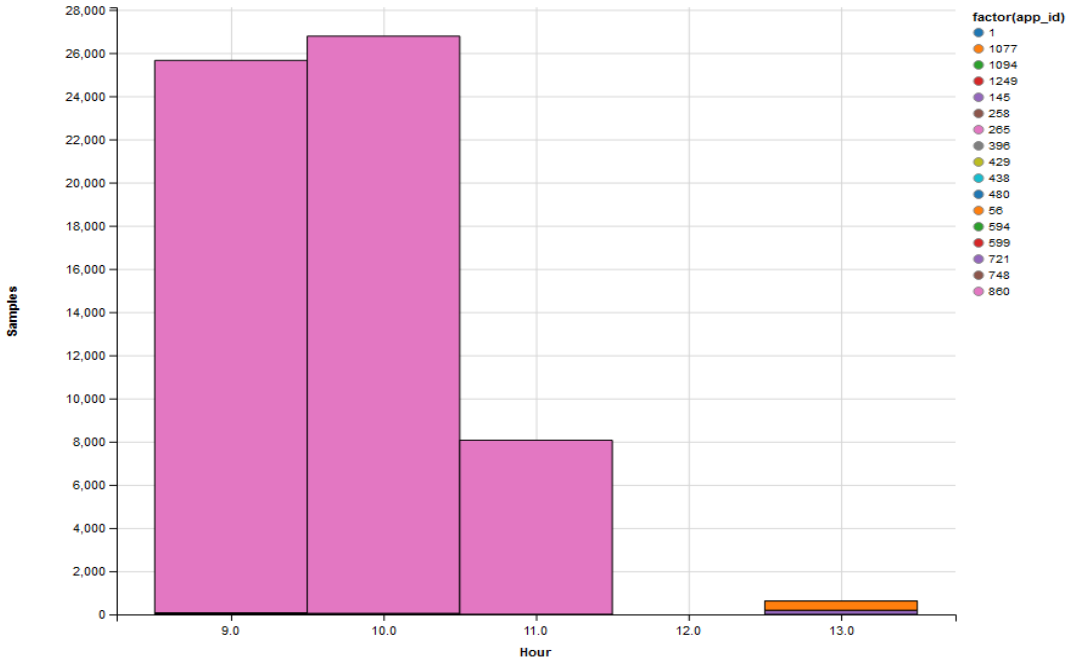
Number of applications, which we can distinguish differs from hour to hour

Hour	Minimal number of samples	Number of characterized apps
0	2837	18
1	7589	3
2	2417	13
3	290	50
4	31	118
5	31	132
6	31	196
7	9646	6
8	1686	48
9	8637	21
10	4182	39
11	5378	34
12	10392	12
13	1320	95
14	3231	59
15	2625	54
16	1438	67
17	1968	55
18	13095	9
19	12781	7
20	1674	47
21	8814	10
22	1861	36
23	1941	25
Whole set	18300	100

Q2.4: Are peak times of a base station influenced by a particular app?

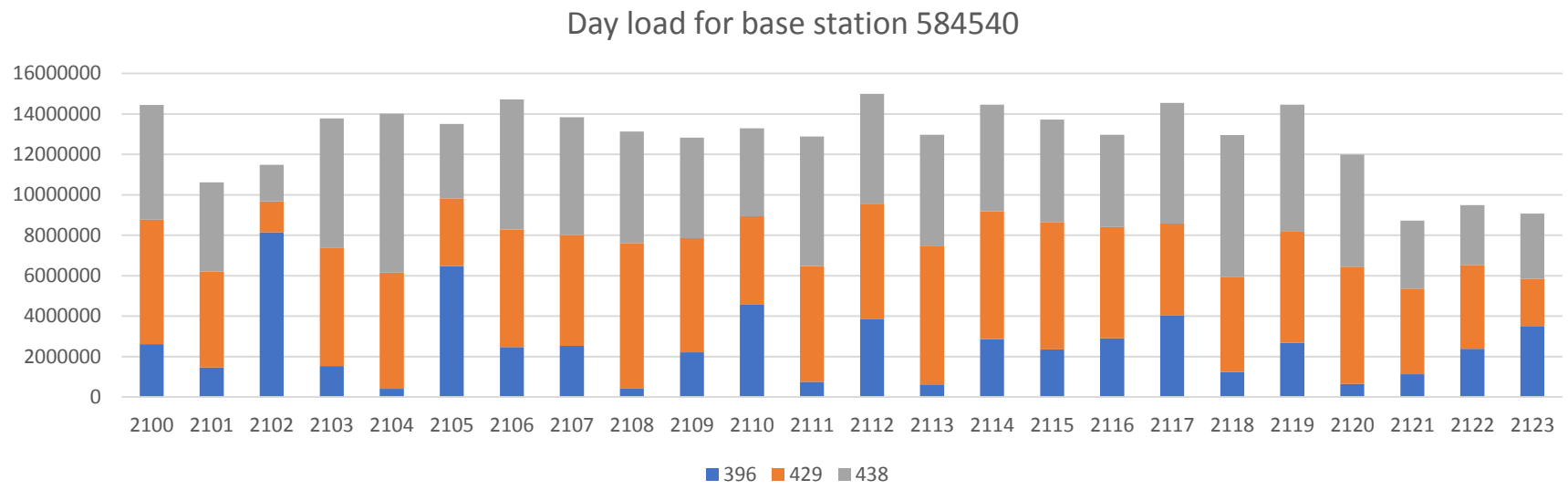
Yes, for some base station we can observe peak hours, influenced by one application. Examples:





Q2.4

Example of base station, which has no obvious peak, caused by one application



Q2.5: Categorize users based on app usage

Users use different application and, therefore, they can not be categorized based on same app usage

However, we can compare user activity (total number of samples, for example) at different hours

Some users are active in the morning, i.e. they show most of samples between 6 and 10

Other categories: night users (from 0 to 6), day users (from 10 to 18, working day is implied) and evening users (from 18 to 0)

Q2.6: Is it possible to identify users' mobility (trajectory) based on base station access

Yes, it is possible

906 (90,8%) of users are trackable (i.e. they visited more than one base station)

However, building an actual trajectory is not a correct task as:

1) times we observe show just when sample was captured;

2) if time is large it might not show the fact, that stations are distant from each other;

2) zero or relatively small interval, however, indicate neighbor base stations

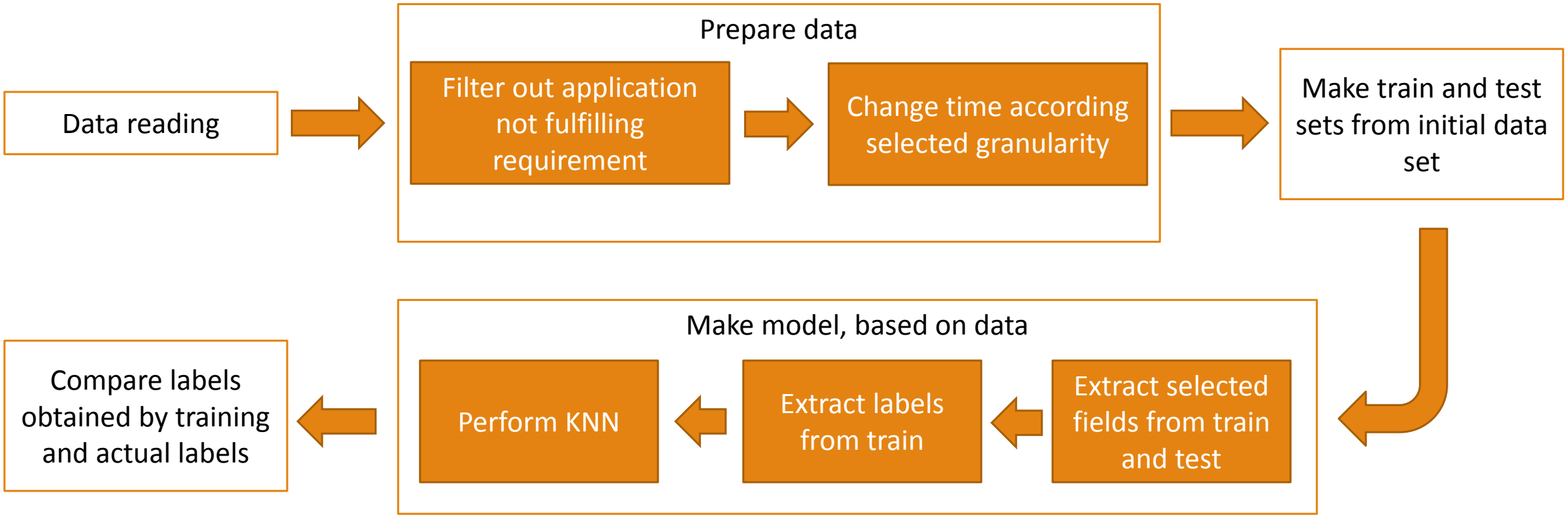
Q2.6

Examples: users 0095346 and 0057202

From	To	Time	Start	Finish
186327	207059	2	21101936	21101938
207059	186327	75	21102925	21103000
186327	207059	128	21104920	21105048
207059	140370	5	21110007	21110012
140370	186327	4	21111009	21111013
186327	140370	1	21111046	21111047
140370	186327	4	21111047	21111051
186327	207059	20118	21113031	21133149
207059	176396	8	21134108	21134116
176396	206352	6	21135109	21135115
206352	176396	0	21135115	21135123
176396	206352	1	21135123	21135124
206352	176396	0	21135124	21135124
176396	206352	1	21135124	21135125
206352	186327	69	21141049	21141118
186327	157144	6	21142109	21142115
157144	137643	112	21143014	21143126
137643	157144	0	21143126	21143133
157144	137643	1	21143133	21143134
137643	157144	0	21143134	21143134
157144	137643	4	21143134	21143138
137643	215255	5	21144109	21144114
215255	137643	2475	21152326	21154801
137643	215255	5673	21155328	21161001
215255	186327	4503	21161002	21165505
186327	207059	1399	21170045	21171444
207059	186327	782	21172043	21172825
186327	207059	7630	21184409	21192039

From	To	Time	Start	Finish
283213	283212	0	21080036	21124902
283212	283213	2	21124902	21124904
283213	283212	1	21124921	21124922
283212	283213	0	21124922	21125008
283213	283212	3	21125008	21125011
283212	283213	2	21125011	21125013
283213	283212	2	21125013	21125015
283212	283213	0	21125015	21125015
283213	283212	1	21125015	21125016
283212	283213	1	21130942	21130943
283213	283212	1	21130951	21130952
283212	283213	6	21131000	21131006
283213	283212	0	21131006	21131006
283212	283213	5	21131006	21131011
283213	283212	2	21131011	21131013
283212	283213	20	21131013	21131033
283213	283212	70	21131944	21132014
283212	283213	362	21132955	21133317
283213	283212	10	21150011	21150021
283212	283213	8	21150021	21150029
283213	283212	3	21150029	21150032
283212	283213	1	21152022	21152023

Q3.1: Pipeline



Q3.2: Most interesting findings

The most interesting things were:

1) applications making base station peak times: sometimes just one user might cause this peak;

2) large similarity, which different applications showed (same purpose applications?); however density is not exactly the same but it is difficult to distinguish it numerically

3) most likely there is, in fact, not enough data to properly characterize application

Q3.3: Most important findings

1) Traffic volume for different applications observed similarity and it only can not be used to distinguish applications

2) Information of user, time and location allows much better identify application; for this set time can be used with second preciseness; it is also important that at different hour behavior differs

3) The way to identify application is a construction of usage pattern for a given user this pattern can be built considering time and location; data for week might be very helpful