Weekly report: July 23, 2015

Ivan Rios S.

July 24, 2015

1 Goals for the week

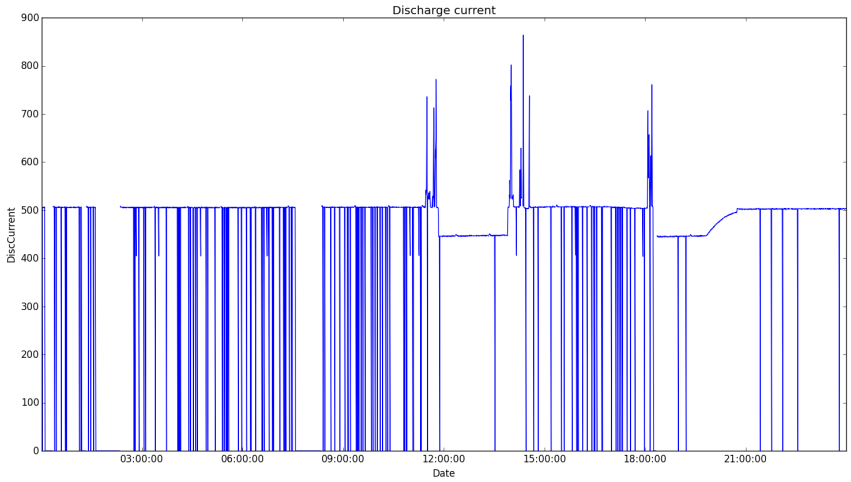
- Define the final version of the trip detection algorithm based on recent analysis of all the trips in the DB
- Obtain all trips available since the beginning of the project
- Make an aggregated analysis of all the trips and an analysis based on the demographics of each participant

2 Activities

- Implement new tables in the DB and read this data from the beta site
- Implement discharging current in the algorithm
- Implement linear acceleration as part of the algorithm.
- Fix bugs in the trip detection algorithm based on the results obtained from the analysis of all the data in the DB
- Tabulate mine and professor Golab's trips to analyze the accuracy of the algorithm
- Determine accuracy of the algorithm based on the logs available.
- Determine potential errors of the algorithm by looking at trips not detected from the logs.
- Make an aggregated analysis of the trips.
- Perform a sanity check analysis of the results and data cleaning.

3 What I learnt

Initially, I worked on implementing the discharging current as part of the algorithm. This requirement originally came from the idea that we can filter out trips where the participant took the bus, and put the bicycle on the rack, or the case when the participant walked with the bike. I implemented this change, and the number of trips detected radically went down. I analyzed some of the trips that were filtered out and the main problem is that charging current is not a very reliable measurement. The following plot will show how changing it is:



The previous graph shows 3 trips in the chosen date, one for each peak of the graph; however, each one of the peaks do not start at the time when the actual trips started, and end earlier compared to the trip logs. When analyzed closely, the data is very variable and because of that, the algorithm was not able to determine that there was a trip. Here, it is important to remember that the algorithm uses a sliding window approach, adding one record at the time, which requires the sensors to be sensitive and immediately catch variations. However, from the previous analysis it is possible to see that the data from the sensor is not stable enough to be used.

My next approach to this problem was to get an average of the discharging current during each trip, and based on that, determine if the battery was under use or not. With this solution, approximately 50% of the previously filtered out trips were included again but the high variation of the data was still filtering out a considerable percentage of data. I experimented with different threshold values but it was not possible to find a value that would prevent this variable to filter out trips that were successfully detected by the original algorithm.

In a similar manner, I worked on determining if the linear acceleration should be used instead of the gyroscope measurements for the trip detection algorithm

since both presented the lowest variability and a high sensitivity to find trips. I replaced this variable in the algorithm and determined that the number of trips detected are very similar with some differences on specific cases. In some rare occasions, data from one of the sensors was not available which means that the trip at that time could not be detected. For this reason, I decided to join both variables as part of the algorithm, and use either one of the two in order to determine trips (whichever detects movement). This not only incremented the number of trips, but also was able to determine the initial and final times of the trips more accurately.

Next, I found several problems on the algorithm that were causing some problems when obtaining all the trips available in the database. The following were found:

- GPS data not available: I found that many trips were being filtered out because the distance calculated when determining the average speed. This problem is related to the poor accuracy of GPS measurements which was reflected in a considerable amount of trips not being detected when the first (Tommy's) algorithm was written. The solution I found for this problem is to determine if the total distance calculated was using a high number of noisy data or there was not GPS data available at all. If it was the case that the GPS data was not being collected at the time of a trip, then we are not considering that distance and the algorithm adds that trip to the list.

- Threshold used for each variable causing problems: With the original values used for the variables of the trip detection algorithm and more ground truth, I was able to measure the accuracy of the algorithm in a better way. Based on this analysis, I revised each one of the records in the database to determine potential reasons why some trips were not being detected. Approximately 10% of the trips from the logs were not being detected because the thresholds to determine if there was movement were too high. With this information, I lowered the values but reached a point where the algorithm was giving false negatives (which is generally not the case in this type of algorithms). After iterating with changes and determining the new accuracy obtained with them, the solution I came up with is to use the average value of the records in the trips that were not detected. With this solution, I was able to increment the accuracy in approximately 7% of the logged trips and duplicated the total number of trips detected.

- Lower number of records were collected per minute: a common problem is the variable number of records collected per minute. This number varies very often which becomes a challenge since the sliding window approach relies on a fixed size. Additionally, the change of number of values collected every minute with version 16 of the software in the phone caused to have several problems with the algorithm since the sliding window started

covering approximately 20 minutes at the time instead of 4 as initially designed. For this reason, I experimented with several new sizes of the window (number of records analyzed at the time) and was able to go down to 12 (instead of 90) with no impact on the accuracy of the algorithm. 12 records uses at most 4 minutes of data which is consistent with the initial design. Also, this change did not affect the accuracy when analyzing trips with data collected with former versions of the software.

Next, to determine the accuracy of the algorithm, the following data was used:

IMEI: 3410

Date	Logged trips	Identified trips	Comments
27-Apr	3	3	
29-Apr	3	4	Additional pause in work trip
01-May	3	3	
06-May	2	2	
07-May	3	2	
08-May	2	2	Additional pause in work trip
11-May	2	2	
13-May	2	2	
14-May	2	2	
19-May	3	3	
22-May	2	3	
23-May	2	3	Additional trip at 7:36 pm
25-May	2	2	
01-Jun	2	2	
02-Jun	3	2	
05-Jun	3	3	
08-Jun	3	5	Two trips divided due to long pauses in between
TOTAL	42	45	

IMEI: 5233

Date	Logged trips	Identified trips	Comments
25-May	2	2	
26-May	2	4	Two trips divided due to long pauses in between
29-May	2	2	
01-Jun	5	5	
02-Jun	2	0	
05-Jun	2	2	
06-Jun	5	5	
07-Jun	2	2	
09-Jun	3	3	
16-Jun	2	2	
20-Jun	2	2	
23-Jun	2	2	
24-Jun	3	3	
27-Jun	3	3	
09-Jul	2	0	
13-Jul	3	3	
14-Jul	1	1	
TOTAL	43	41	

As shown in the table, in the case of IMEI 3410, all the trips were detected by the algorithm; however, additional trips were also identified. Analyzing this data in more detail showed that some trips were divided into two because there were longer pauses than allowed (4 minutes) in between. This is consistent with problems that were initially identified where the participant considers that there was only one trip but actually it does not meet the parameters defined in the

algorithm so the trips are divided into two.

Similarly, for the case of IMEI 5233, most of the trips were detected and some of them were divided into two. However, for this IMEI we find trips that were not detected at all in 2 specific dates. I analyzed each one of the records and in two, out of the four cases, the duration of the movement detected did not reach the minimum duration of a trip (5 minutes) so it was disposed. From this results, the following summary was obtained:

Parameter	Value
Total number of bicycles	2
Total number of trips (ground truth)	87
Total number of detected trips	86
Number of additional trips detected	7
Number of trips not detected	4
Accuracy	˜91%

Both of the previous tables present the final results obtained with the latest version of the algorithm. Since this is the highest accuracy obtained, the aggregated analysis was performed on the trips detected with this algorithm. The following table presents the total a summary of the results obtained from the analysis of all the database:

Parameter	Value
Total number of trips detected	1351
Average number of trips per participant	45
Number of different days of trips	657
Number days with odd number of trips	346
Number of days with 1 trip	223

From the previous table we can see that there is a considerable amount of trips to be analyzed even though the average per participant is still low. Similarly, an important value to consider is the number of days with only one trip since it means that approximately 17% of the trips are for exercising or only for pleasure; this means that in 83% of the cases the participant used their bike for commuting.

The following data was obtained considering the gender of each participant (including the ISS4E team):

Gender	# of Participants	# of Trips	Average
Male	18	786	44
Female	13	565	44

In this case, we can see that there is no difference between the number of trips calculated from male and female participants. Both genders have been using the bikes the same amount of times.

Also, the following data was obtained considering the profession of each participant (without ISS4E team):

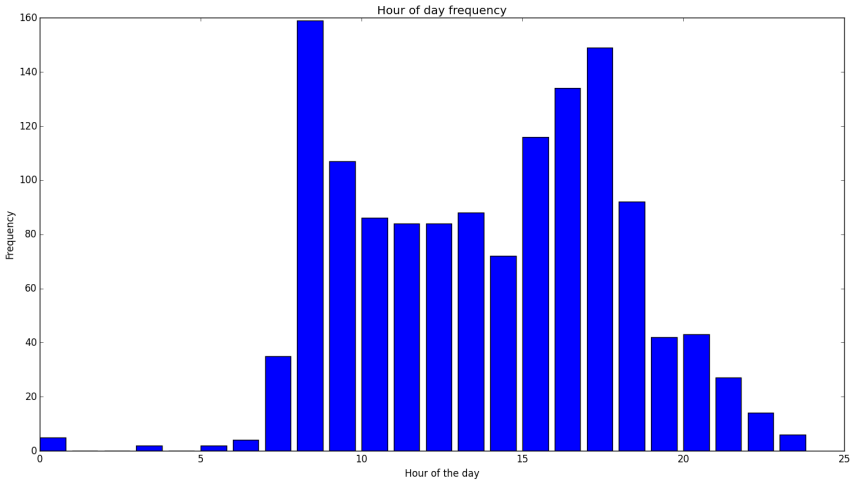
Profession	# of Participants	# of Trips	Average
Staff/Faculty	13	506	39
Student	12	568	48

The previous table shows that there are more trips obtained from students than from staff. This is consistent with the demographics considering that most of staff and faculty have a car and generally a bike is a more common way of transportation for students.

The following are the results obtained from the aggregated analysis of all the trips:

Important note: all the graphs where there is an analysis divided by gender or profession have been modified to use percentages instead of the actual values. This was done because the ratios of male/female and staff/students is not 1.

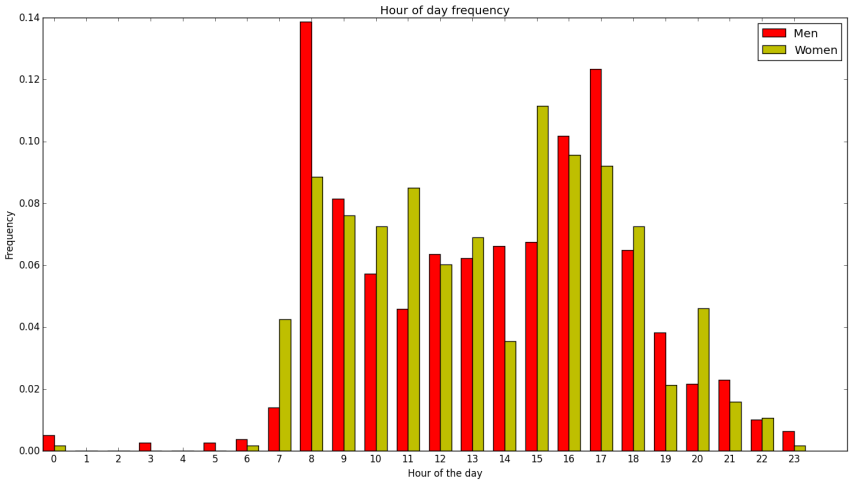
FREQUENCY OF TRIPS PER HOUR (Data includes all the participants and the ISS4E team):



The previous plot shows the number of trips identified at every hour of the day. Here, we can see that most of the trips happen between 9 am and 6 pm which is consistent with the schedule followed by both students and staff. The plot also shows 2 peaks at 9 am, and 5 pm, which is expected because that is

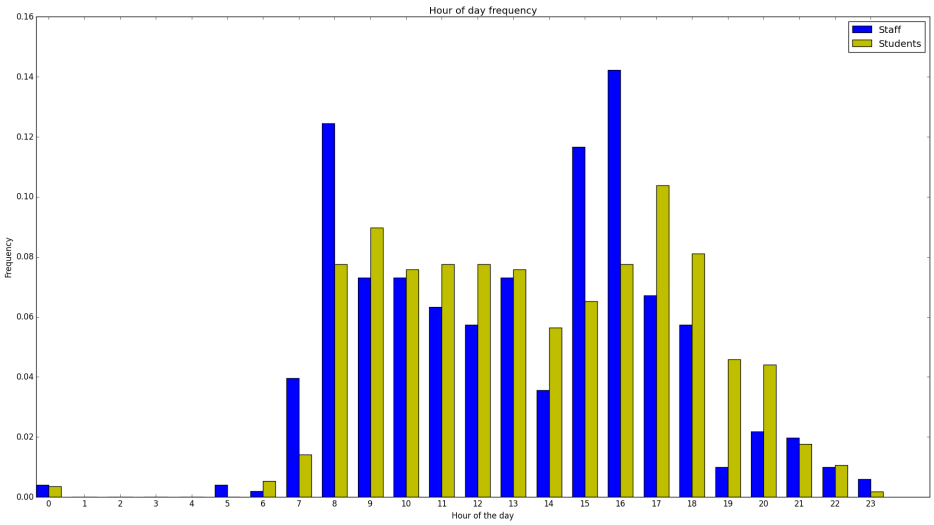
generally the schedule followed by people who use the bike to go to work. Now, if we divide this between genders, the following graph will be obtained:

FREQUENCY OF TRIPS PER HOUR DIVIDED BY GENDER (Data includes all the participants and the ISS4E team):



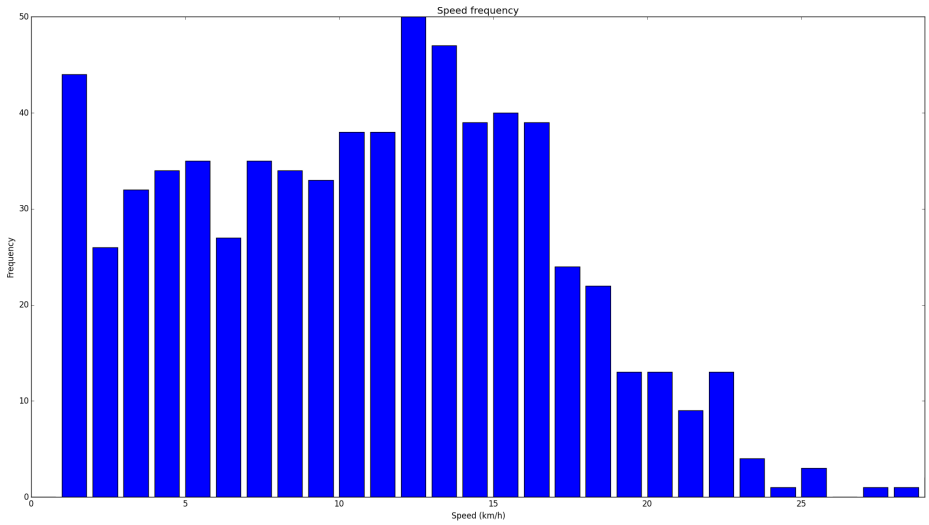
The previous plot shows the same results but divided by gender. As we can see, men seem to follow a more defined schedule when using their bike, especially on the peak hours while women have more scattered values with a peak at 3 pm. Women's data has more flat values while men's shows a 'U' shape with the lowest value at 11 am. Similarly, the analysis was done by dividing staff/faculty and students:

FREQUENCY OF TRIPS PER HOUR DIVIDED BY PROFESSION (Data does not include the ISS4E team):



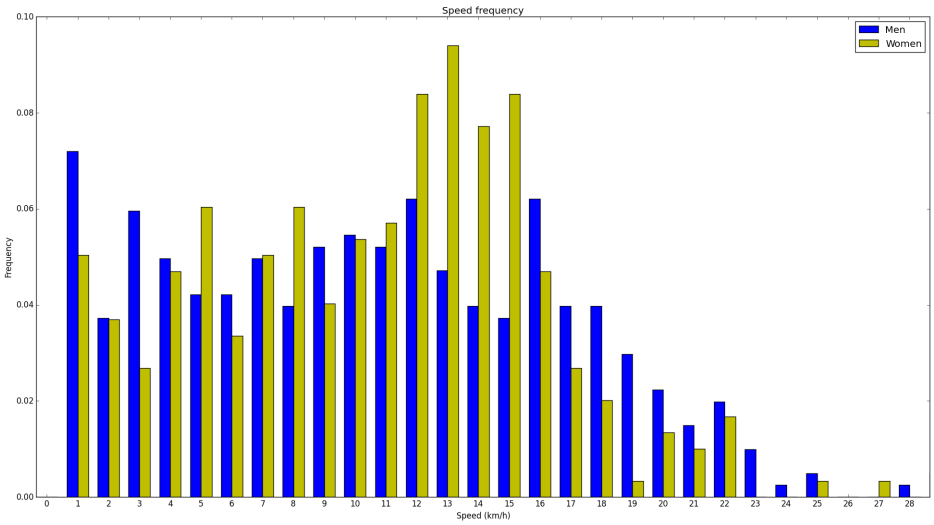
Here, we can see that students have a very flat plot which is consistent with their schedule of classes scattered at different times of the day. On the other hand, Faculty and Staff have a more defined schedule where the higher peaks happen at 8 pm and 4 pm which describes a typical schedule of a working person.

FREQUENCY OF AVERAGE SPEED (KM/H) (Data includes all the participants and the ISS4E team):



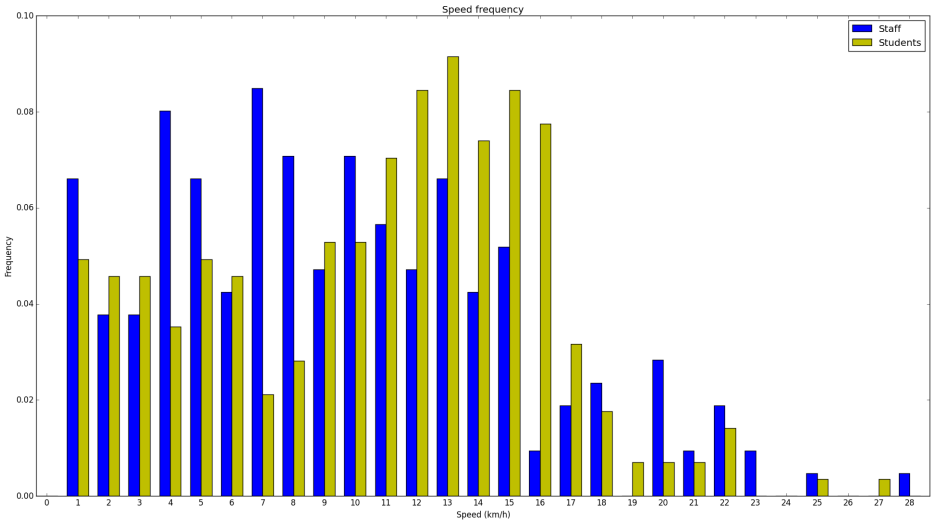
The previous plot shows the frequency of the average speed. Here, we can see that the average speed has a peak of 12-13 km/h which is a normal speed generally obtained in a trip where there are several stops (traffic lights, among others). Also, there is a considerable number of trips with low speed which could be considered normal when using the bike for pleasure. Similarly, this analysis was divided into genders:

FREQUENCY OF AVERAGE SPEED (KM/H) DIVIDED BY GENDER (Data includes all the participants and the ISS4E team):



Here, it is possible to see that women generally ride faster than men but there is more men reaching values above the average. The higher average reached by women is not generally expected since men tend to drive/ride faster. Similarly, the analysis was done dividing the population by profession:

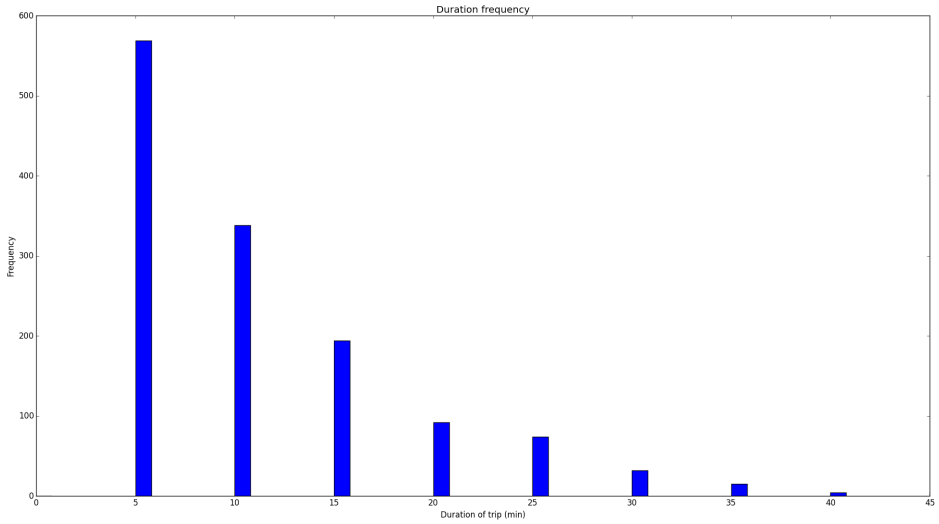
FREQUENCY OF AVERAGE SPEED (KM/H) DIVIDED BY PROFESSION (Data does not include the ISS4E team):



The previous plot shows how students have the tendency to reach faster

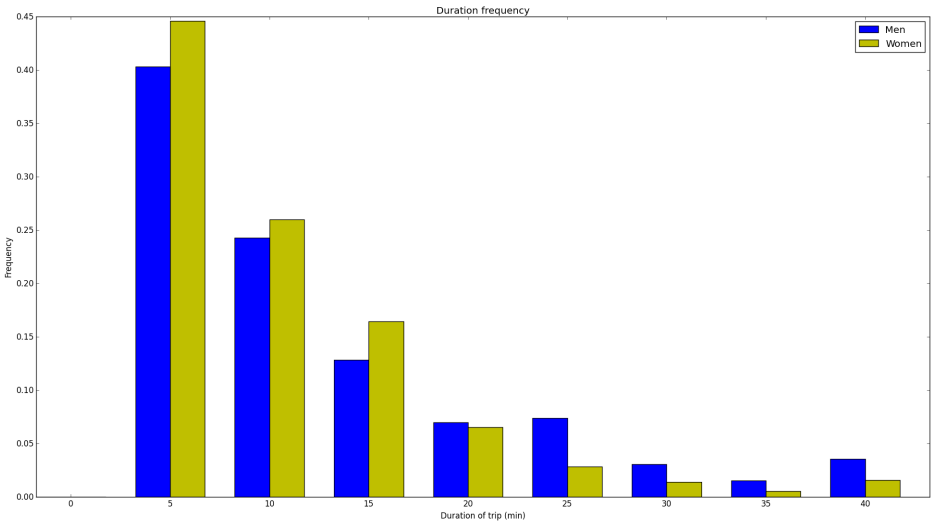
speeds than staff/faculty. This results are expected since younger people generally tend to drive/ride faster. Students in this case show a considerable lower amount of trips at low speed which could lead to conclude that they do not use their bikes for pleasure but generally to commute. The peak of speed by the staff is 7 km/h while it is 13 km/h for students.

FREQUENCY OF TRIP DURATION (Minutes) (Data includes all the participants and the ISS4E team):



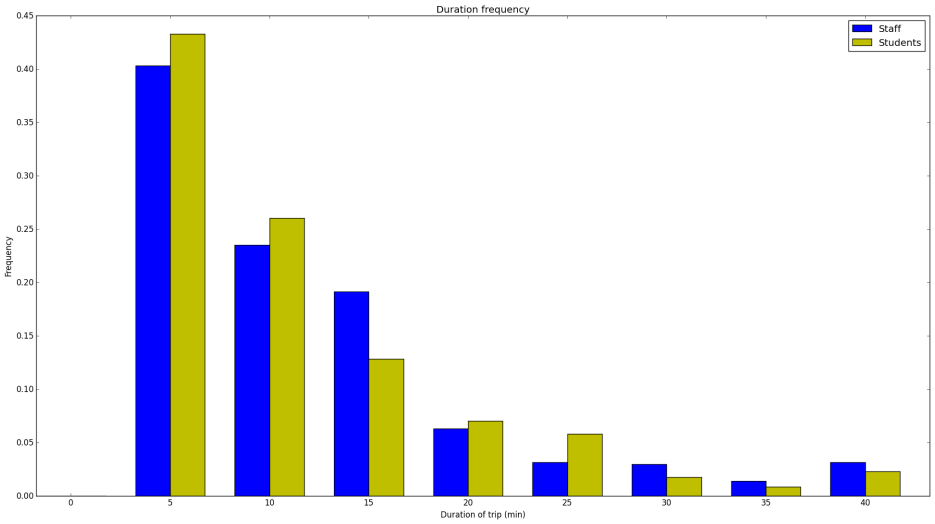
As we can see in the plot, the duration of trips shows a perfect exponential curve that tends to zero trips going over 40 minutes. This is consistent with a regular pattern of use of the bikes since generally people decide to drive in cases when the distance (and, hence the duration) is too big. As before, the analysis was also performed dividing the population by gender:

FREQUENCY OF TRIP DURATION (Minutes) DIVIDED BY GENDER (Data includes all the participants and the ISS4E team):



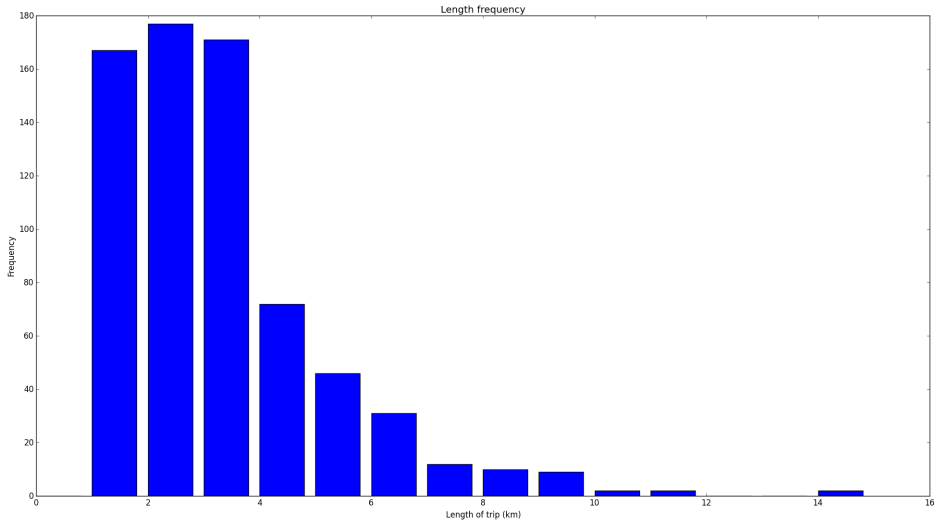
Here we can see that there is not a clear difference between man and women since the values obtained are very similar. Again, the following plot divides the population based on their profession:

FREQUENCY OF TRIP DURATION (Minutes) DIVIDED BY PROFESSION (Data does not include the ISS4E team):



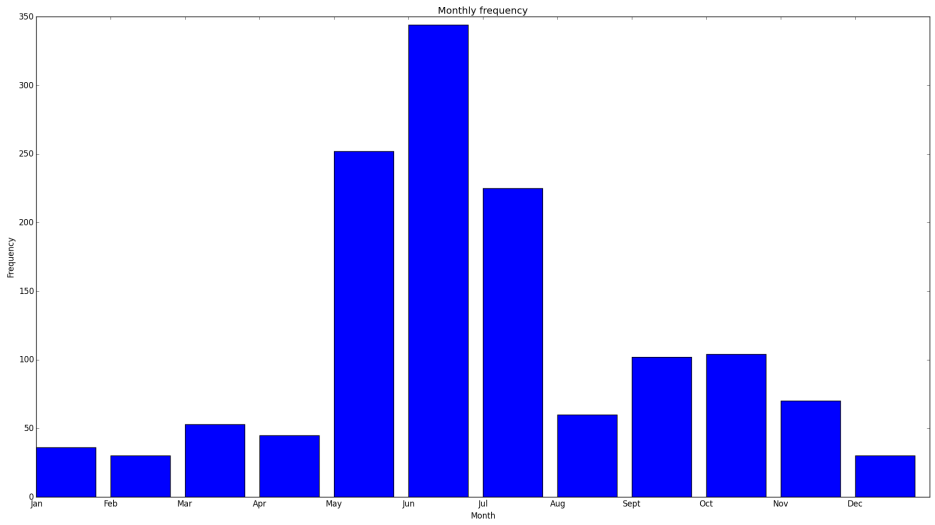
Again, the results do not show a considerable difference between both populations.

FREQUENCY OF TRIP LENGTH (km) (Data includes all the participants and the ISS4E team):



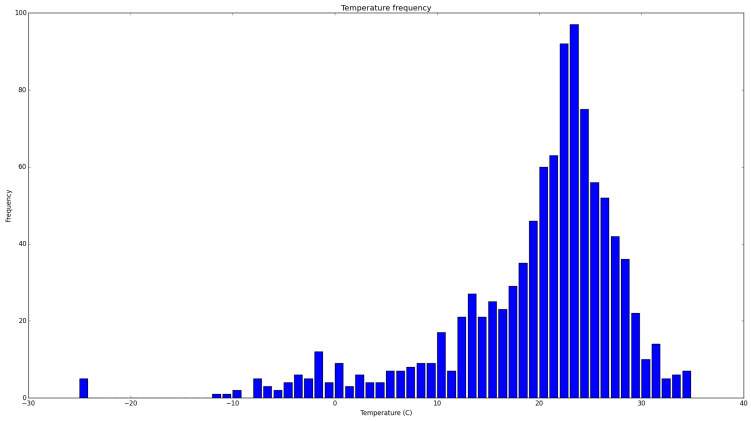
The plot shows the frequency of the length (total distance) of each trip in kilometers. Here, we can see that the distance of trips behaves in a very similar way to the duration, presenting an exponential decrease of the frequency which is consistent with previous conclusions. This is consistent with the fact that people tend to use the bike only for short distances while they prefer to take the bus or drive when the total distance is too long.

FREQUENCY OF TRIPS PER MONTH (Data includes all the participants and the ISS4E team):



The previous plot shows the frequency of trips for each month of the year. As expected, there is a peak during the summer months. Not many conclusions can be obtained from this graph due to the problems with data obtained during last summer because it would be expected to have more trips during August. Once August and September 2015 comes, it is expected to obtain a higher number of trips during those months. Similarly, it is possible to see that the lowest values happen during the winter months. The graph seems has its lowest value in December which could mean that people stop using the bike this month, and then it goes up again after they have adapted to the typical low temperatures.

FREQUENCY OF AVERAGE TEMPERATURE (Data includes all the participants and the ISS4E team):



This plot shows the frequency of the average temperature of each trip. As expected, there is not that many people biking when the temperature is under zero degrees Celsius. The graph shows a clear peak between 20 and 30 degrees which is consistent with an preferred temperature to exercise or commute using a bike. Also, the decreasing number of trips after 30 degrees could be related to the fact that participants are using the bike to commute to work, which means they do not want to have to change after riding under the heat.

Finally, here is a summary of the number of trips per IMEI number per month:

IMEI	Total trips	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
7303	139	0	0	0	2	34	51	50	0	0	0	2	0
8664	121	0	0	0	0	0	10	1	2	0	0	0	0
3469	100	0	0	0	0	18	29	0	0	0	2	0	0
2910	93	0	0	0	2	0	1	6	1	3	0	0	0
9050	84	0	0	0	3	14	10	4	21	0	0	3	0
3410	61	0	0	0	0	3	17	7	0	0	0	4	0
0603	59	0	0	0	0	0	1	5	1	0	2	1	1
6473	55	0	0	0	0	2	2	4	0	14	4	3	0
0665	55	0	0	0	5	9	1	0	0	0	0	1	1
5233	50	0	0	0	0	6	3	0	10	0	2	0	0
0669	49	0	0	3	0	5	26	38	0	0	6	6	0
1473	47	10	12	19	4	2	0	0	0	0	5	2	5
6904	46	0	0	0	7	26	13	1	0	6	7	0	1
9519	43	0	0	0	0	8	11	14	0	0	10	0	0
1210	31	0	0	0	1	3	0	2	0	0	3	0	0
7459	29	0	0	0	0	5	5	0	0	8	1	0	0
7517	29	0	0	1	0	0	16	4	0	3	4	0	0
3215	29	0	0	0	0	0	0	2	0	0	0	0	0
4381	28	0	0	0	0	0	16	0	0	7	0	0	0
6089	26	0	0	0	0	33	4	6	0	0	3	0	0
8508	25	0	0	0	6	5	5	0	0	4	2	2	0
7710	24	2	10	2	0	9	8	14	0	0	0	2	0
3014	23	0	0	3	0	5	9	4	2	0	2	0	0
0636	21	0	0	0	0	0	18	25	18	24	8	0	0
5432	19	0	0	0	8	28	9	3	0	0	7	0	0
8870	17	0	0	0	1	0	8	6	0	11	3	0	0
6994	13	24	8	2	2	0	1	0	0	22	7	15	19
0657	13	0	0	0	0	10	35	5	0	0	0	0	0
6097	11	0	0	0	0	3	1	4	4	0	14	0	0
9407	9	0	0	0	0	19	7	1	0	0	2	0	0
0587	2	0	0	23	4	5	27	19	1	0	10	29	3