GraphLab Project

The given data had .txt files and there were no feature names specified for the columns. I browsed the official website and got the feature names from there. Then I converted all the files from txt to csv with the addition of relevant column names for the data.

Task 1: Most congested communication period of the day in Milan.

For this task I tried two different approaches for November data and December data. I wanted to compare the data for both months. First I read each csv file from the given path one by one, changing the time stamp into hours only (00, 01, 02…..23) and adding up all the activities (sms in and out, call in and out and internet activity).

Then I am returning back only top 10 busy hours out of 24 and creating a join of days. This way I am finding the common hours for everyday.

The results are given. It shows that in November 17, 16, 15, 14, 11, 13 were the six busy hours on average per day.

```
In [83]: busy_time_nov
```

Out[83]:

time_interval	total	total.1	total.2	total.3	total.4	total.5
17	5135660.08051	5475050.58491	5653881.00195	7537485.45641	7675379.87141	7688563.20297
16	5108601.48883	5432765.37218	5639881.11101	7783366.18071	7886586.58171	7873889.89774
15	4916011.1817	5270597.51245	5384235.66732	7726240.95824	7749695.35856	7776260.84738
14	4808330.37342	5113098.18215	5162637.00401	7570427.28075	7713313.48314	7640262.32014
11	5038443.5945	5097979.32841	4900827.56305	7612407.22497	7827607.34668	7750840.89816
13	4774330.25905	4993293.46322	4974695.99066	7737459.37291	7900837.84844	7824463.39527

total.6	total.7	total.8	total.9	total.10	total.11	total.12
7623365.47894	7726414.31201	6422268.9555	6018489.23578	7440687.09754	7541960.87457	7483245.85117
7815218.97232	7950660.02076	6456644.28276	6077217.24781	7693471.7008	7752870.02363	7743503.77044
7786919.50181	7848903.71144	6231259.53503	5859213.15079	7621004.91984	7592468.28782	7615521.06131
7671635.4241	7738200.47115	6116837.19746	5693156.18941	7536042.95439	7605359.84713	7467909.3683
7809167.19021	7886134.29797	6304948.50548	5504056.46452	7587411.34332	7639319.35322	7657588.20319
7941600.06009	7819001.23273	6033966.3925	5438935.36338	7606570.71216	7782476.45107	7648805.32578

total.13	total.14	total.15	total.16	total.17	total.18	total.19
7564512.77015	7230341.92396	5954474.7358	5675167.88673	7290975.5532	7410877.80537	7393813.82302
7809647.60402	7511538.63982	5956377.39122	5658036.21688	7490549.63988	7625693.6756	7575514.51863
7725522.35193	7582202.24975	5798269.59256	5443236.93497	7415862.86782	7519984.63313	7511062.46023
7594854.58494	7539018.24703	5667474.97626	5236525.77214	7351333.75418	7524528.39097	7096769.0301
7755786.00793	7654989.60136	5916961.20389	5122845.21205	7470871.67341	7662914.12054	7520067.75335
7744813.72267	7627886.4613	5598980.44695	5067317.11394	7533022.87534	7649281.93362	7361937.91261

Fig 1: Milano Congested Hours in November

The above results are not only showing the busy hours but also describing the total telecommunication usage for each of these hours on everyday.

Following are the most busy time intervals for the month of November.

['17', '16', '15', '14', '11', '13']

After this approach, I tried to read the whole data of December as a single frame.

Now instead of reading each csv one by one, I read all the csv files and made them one frame. The purpose was to compare the time consumption of two different process.

I installed GraphLab on a virtual machine, so the performance of the machine was slow, but above process was bit faster than the one where graphlab was reading each csv performing some actions and then reading the next one.

The output here shows one column of total activity against the time stamp. The busy hours in December were almost the same as in November. Here the most busy hour for the whole month was 16 o' clock.

```
In [9]: busy_time_dec
```

Out[9]:

time_interval	total
16	180527600.249
15	177432343.443
17	176560607.78
11	176293952.415
10	174132366.805
14	173684038.15
13	172406383.517
12	172283352.479
09	165507494.499
18	163960799.034

[10 rows x 2 columns]

Fig 2: Milano Congested Hours in December

Task 2: Top 5 Italian provinces which are most called by residents of Milano

For this task I read all the files of the month of December and got the top 10 provinces where people of Milan called on a particular day. Then I joined the common provinces. Following are the top 5 provinces where people of Milan called. The count column shows total calls made per day.

['MILANO', 'MONZA E DELLA BRIANZA', 'PAVIA', 'BERGAMO', 'VARESE', 'COMO']

```
In [789]: most_called_prov
```

Out[789]:

province	count	count.1	count.2	count.3	count.4	count.5	count.6	count.7	count.8
MILANO	1031525	1081075	1092867	1099411	1093993	1103324	1075520	1023327	1089411
MONZA E DELLA BRIANZA	225989	404946	411135	414198	407608	407133	268650	223599	407468
PAVIA	201319	349931	352170	356598	356207	357295	241336	200306	341119
BERGAMO	85604	216297	215071	219295	224826	226046	121389	93116	212771
VARESE	102692	214295	222611	225204	223454	219923	119625	100857	217715
COMO	98069	193697	199782	201627	200413	205642	113985	93437	191748

count.9	count.10	count.11	count.12	count.13	count.14	count.15	count.16	count.17	count.18	count.19
1093254	1100634	1107377	1113268	1092326	1033055	1098134	1100907	1104789	1112211	1118814
405096	417316	410544	421246	294987	231746	422166	420736	421378	430194	430928
359792	364799	358887	365851	263299	202508	364844	360978	366196	373966	370197
219226	214785	225826	227249	117646	88815	226499	225289	228007	237021	228599
226123	234131	224838	226585	129652	103120	232150	229990	230528	235907	227327

Fig 3: Most Called Provinces from Milano

Task3: List top 5 languages tweeted by distinct users in Milano. How popular is Finnish as a tweeting language in Milano?

I read the social pulse file and aggregated the sum of all languages.
The result shows that Italian is most tweeted, then English, and then Spanish. Finnish is on 15th number with a total count of 1594.

```
+----------+--------+
| language | count  |
+----------+--------+
|    it    | 163889 |
|    en    | 47830  |
|    es    |  7745  |
|    tl    |  6815  |
|    pt    |  5162  |
|    so    |  4013  |
|    fr    |  3654  |
|    id    |  3183  |
|   und    |  3132  |
|    de    |  2183  |
|    tr    |  2065  |
|    ar    |  1886  |
|    nl    |  1754  |
|    ro    |  1746  |
|    fi    |  1594  |
+----------+--------+
[45 rows x 2 columns]
```

Task4: Compare call and internet activity between 24th, 25th and 26th December to 26th, 27th, 28th November for Milano. Plot the distribution.

I read 3 files from November for the given dates and 3 files from December.

The results gave calls in and out, sms in and out, and internet activity for 3 days of each month. The following result shows total calls and internet activity for 24 hours for the months of November(26, 27, and 28) and December (24, 25, and 26).

```
In [79]: compare_months
```

Out[79]:

time	nov_calls	nov_internet	dec_calls	dec_internet
00	58244.7409143	7728274.33017	118741.64179	6633322.19294
01	36185.9910265	6610044.89525	60736.3994432	5679718.17914
02	31428.1219474	6060435.11328	38123.8401479	5064958.32263
03	33935.1610568	5671226.68821	32236.4930624	4651150.24702
04	48388.4868479	5733181.36276	36732.4742017	4524129.17286
05	128593.977075	6725404.8912	61433.8635946	4651954.70225
06	535095.380621	10027939.5712	130271.905069	5179836.90202
07	1506788.70643	13636624.2293	405424.900151	6142928.96878
08	2378205.76396	15010332.653	1027483.19831	7386525.02885
09	2639849.46475	15453651.3748	1719930.31221	8642876.34494

[24 rows x 5 columns]
Note: Only the head of the SFrame is printed.
You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.

Fig 4: Comparison of Call & Internet Activities for Nov and Dec in Milano

Then I used Graphlab's feature Sframe.show to plot the charts.

Following are the graphs of 4 columns Nov calls, Dec calls, Nov internet, and Dec internet. The results clearly shows that calls made in November are higher than December. In November, the most busy time periods were from 9 to 10 and then from 15 to 17. If we look at the results of December we can see that the most calls made were between 9 and 11. At 10, it was the most busy hour but in total the calls made in December were a lot lower in numbers than November.

Similarly, the internet usage as per the results, was higher in November than December. The results are drawn for 24 hours and it recorded high usage of internet from 10 to 17 in November. If we compare the results of November with December, the high use was recorded at 11 and than from 14 to 17 in December.

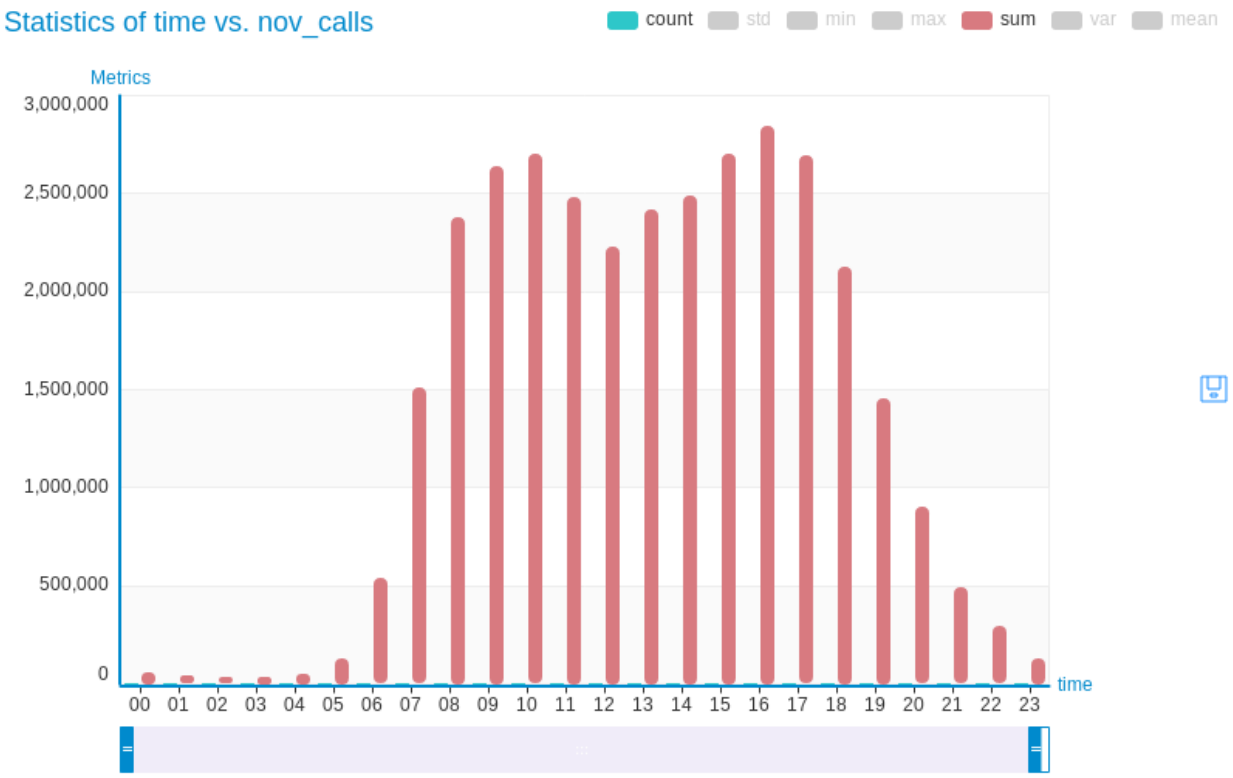

Fig 5: Busy Hours for Calls on Average in Milano (26, 27, 28 Nov)

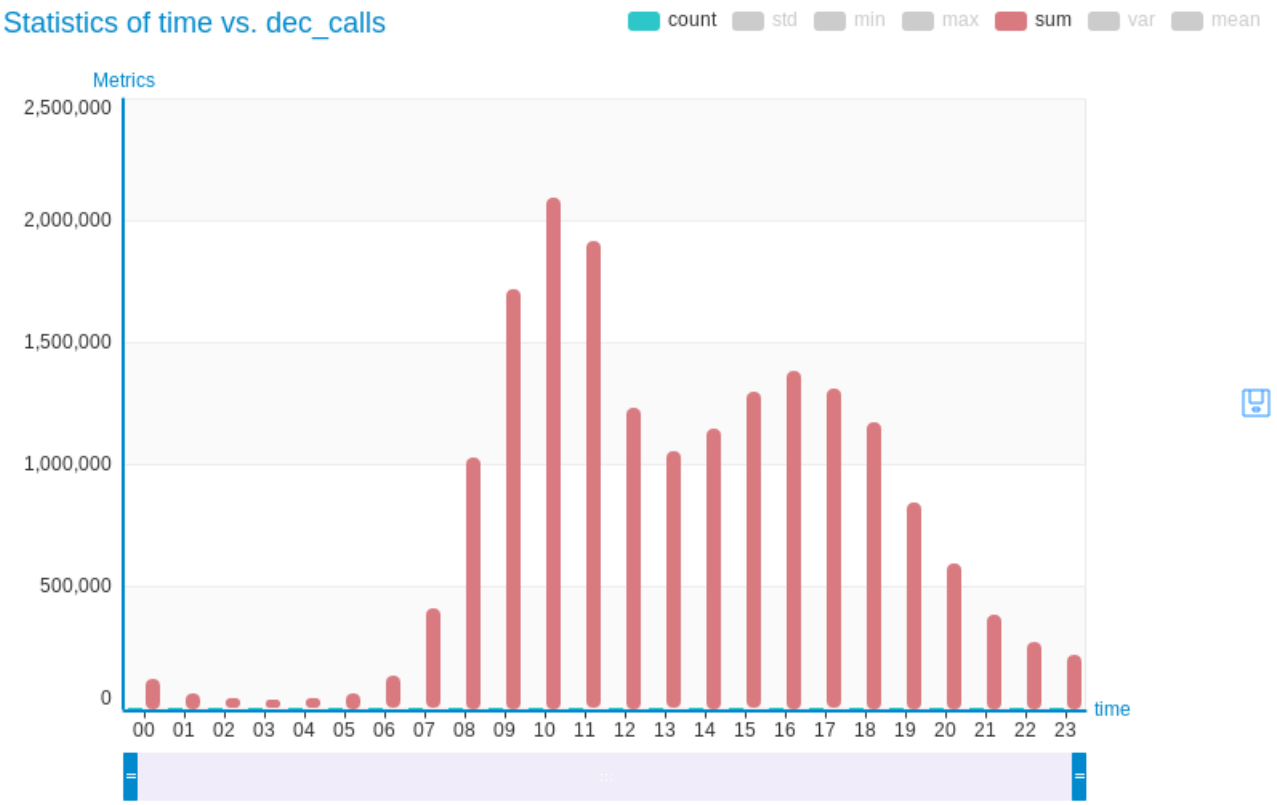

Fig 6: Busy Hours for Calls on Average in Milano (24, 25, 26 Dec)

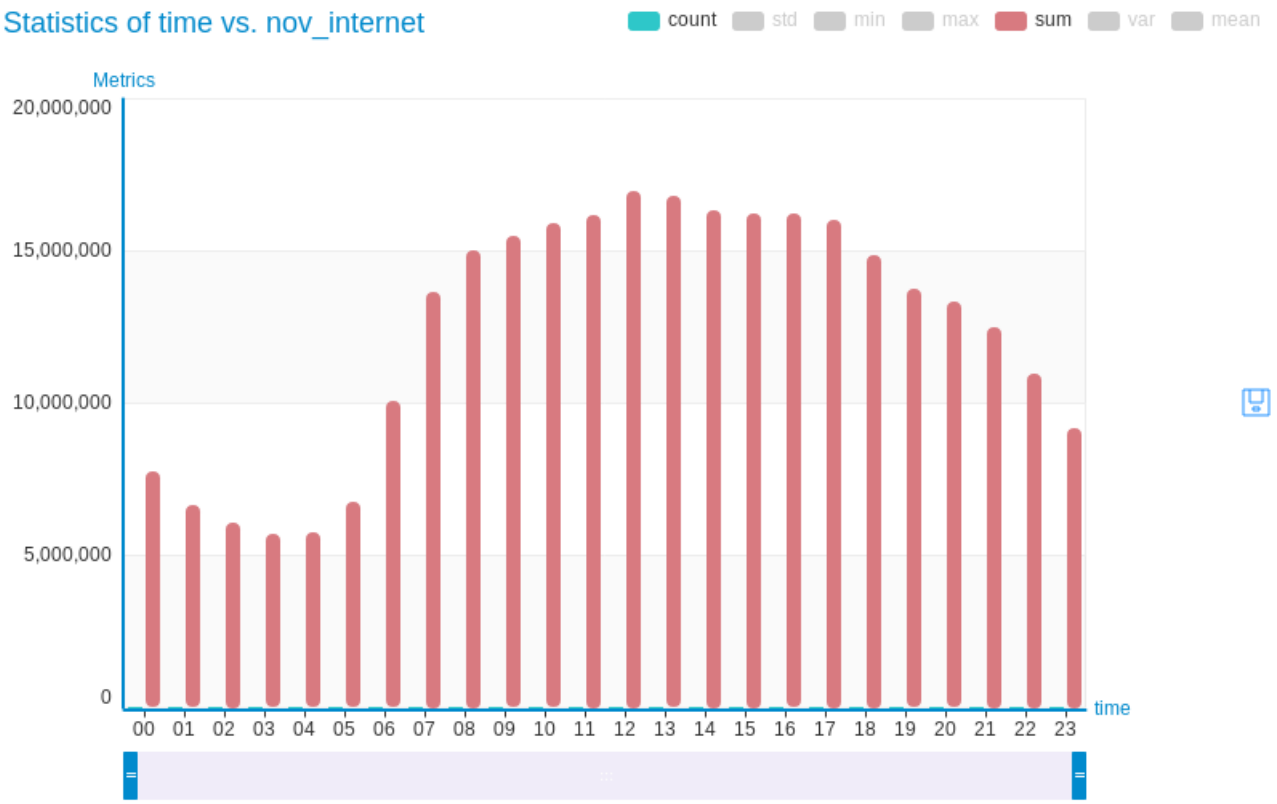

Fig 7: Internet Activity on Average in Milano (26, 27, 28 Nov)

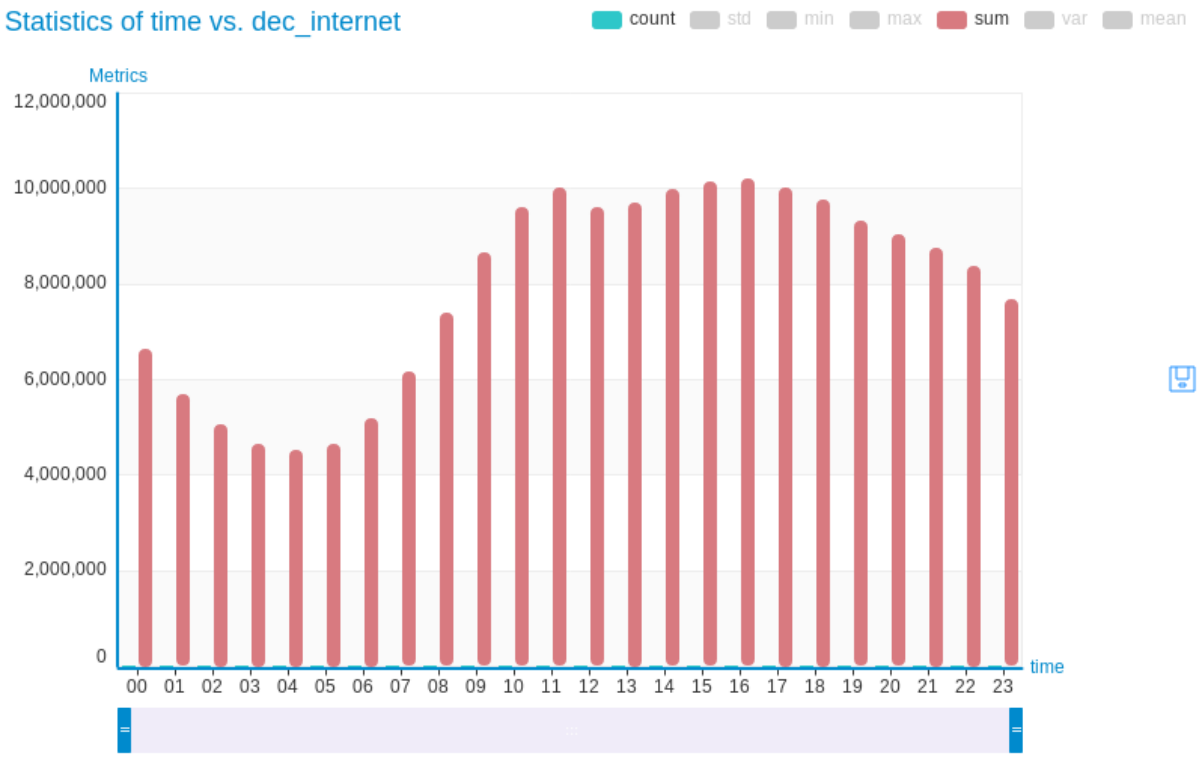

Fig 8: Internet Activity on Average in Milano (24, 25, 26 Dec)

Task 6: User telecommunication activity Milano.

I read the data for the November and December for Milan and summed all the data which includes calls in and out, and internet activity. Then I added data of both months and finally I had time and total activity. I plotted the heat map but since I didn't have enough data (as I turned whole data into 24 hours) so the heat map did not clear anything.

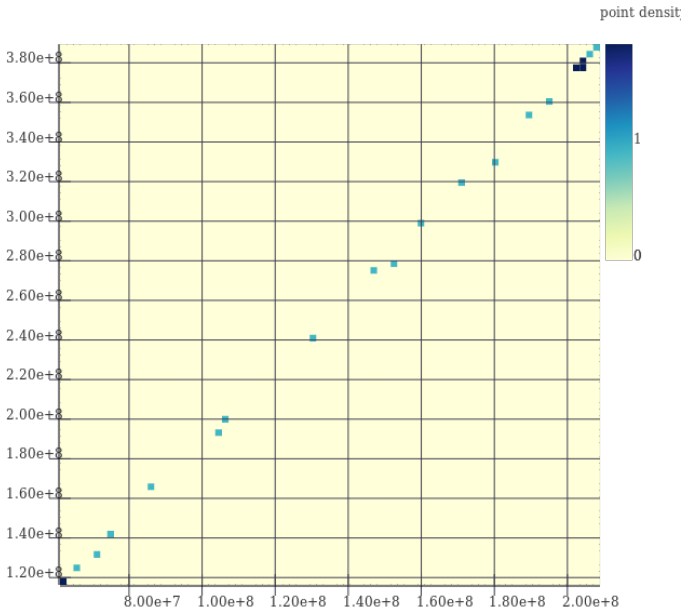

Fig 9: User Telecommunication Activity Milano

Then I plotted a line chart to understand the total telecommunication activity during the 24 hours for the data of two months November and December. The result shows that the highest telecommunication activity is recorded at 16 o' clock. After 16, its a steady decrease.

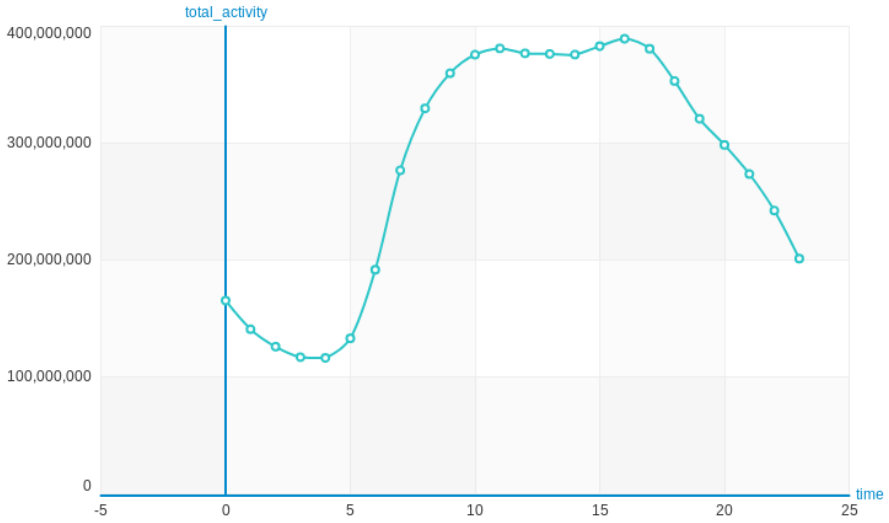

Fig 9: User Telecommunication Pattern Milano

To answer the question about communication pattern of users during day and night, the following bar chart gives a clear sight. It shows a continuous increase in usage from 8AM. There is a significant difference in usage during the day time and night time. From 8AM, the user's activity in increasing but after 8PM its very low until 3AM. At 3 and 4 Am it is at its lowest.

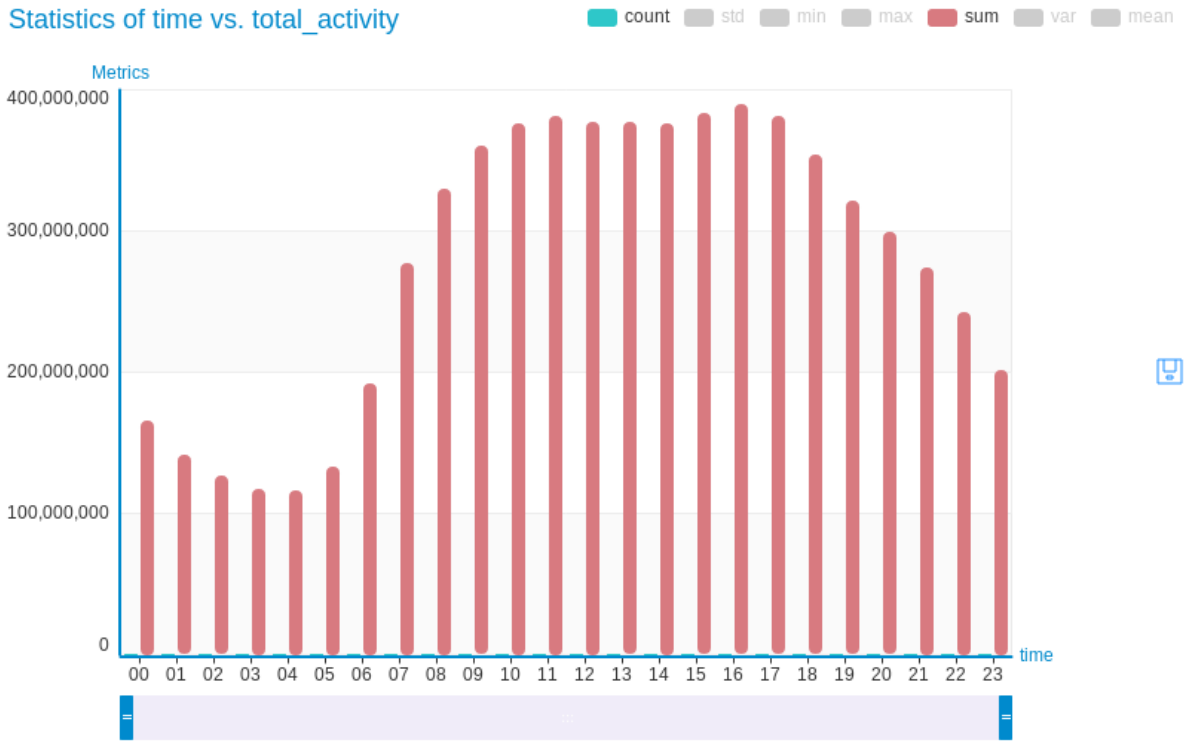

Fig 10: User Telecommunication Pattern Milano

After Milano, I used almost the same methods to find answers of the above questions with Trento Dataset. The complete code is attached in the zip file.

Here are the results of Trento dataset:

Task1: Find the most congested communication period of the day in Trento.

Following are the most busy hours in Trento along the total telecommunication activity over the period of November and December.

```
In [70]: trento_busy_hours.head(5)
```

Out[70]:

time_interval	total
16	65419694.1227
15	63504899.9281
17	63038743.3513
10	62050186.3225
11	61580351.91

[5 rows x 2 columns]

Fig 11: Most Congested Hours Trento

Task2:List top 5 Italian provinces which are most called by residents of Milano and Trentino on average.

Following are the top 5 provinces called by the people of Trento along the total calls.

```
In [72]: call_prov(torento_prov)
```

Out[72]:

province	count
TRENTO	33107842
BOLZANO/BOZEN	10027477
BRESCIA	6423139
VERONA	5929072
MILANO	3969455

[5 rows x 2 columns]

Fig 12: Most Called Provinces From Trento

Task 5: correlation between user communication activity and different weather conditions (e.g. rain, snow etc.)

On the official site, it states that the weather data has a feature named 'Type' with three different values 0,1, and 2 whereas 0 indicates clear weather, 1 shows rain, and 2 represents snow. So I read that data and filtered it out for three different frames as per their types. Then I joined this data with telecommunication usage data based on time column. I used that data to plot the frames individually and together as distribution plots.

I am first attaching individual graphs as they gave me a better understanding in the beginning.

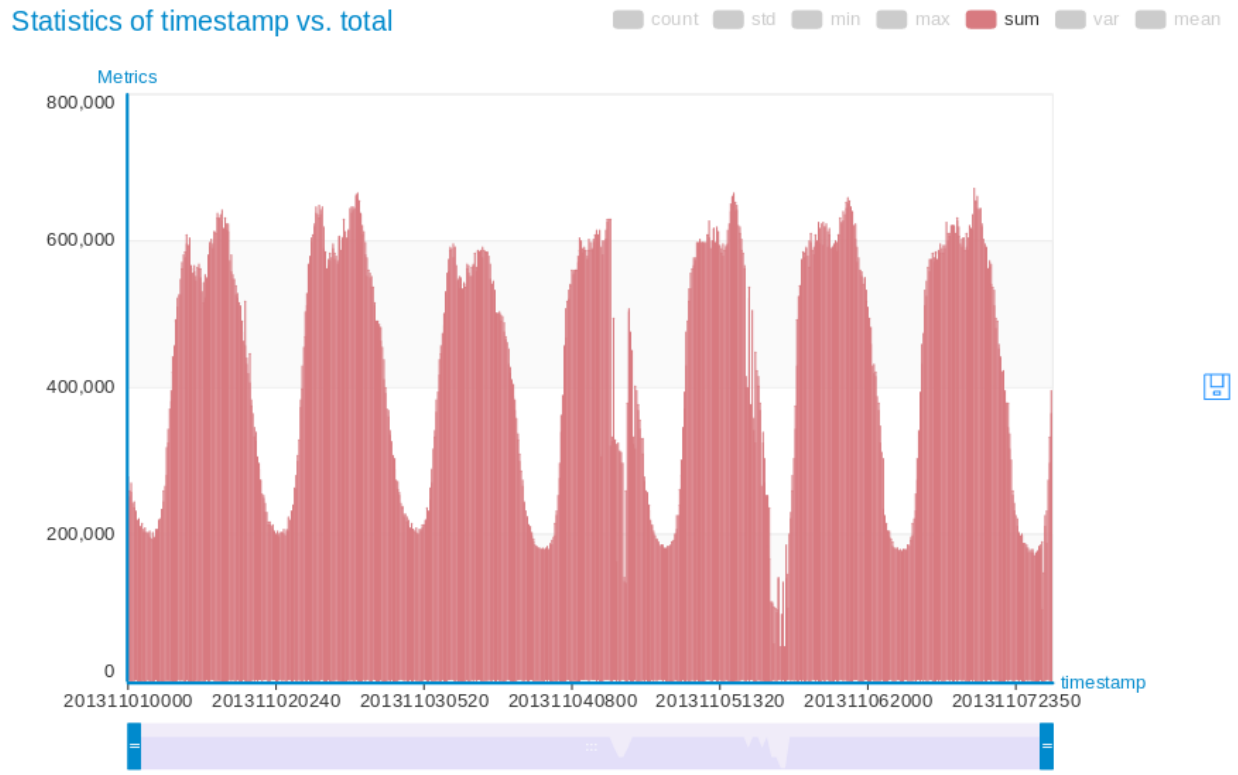

Fig 13: Telecommunication Usage for Clear Weather

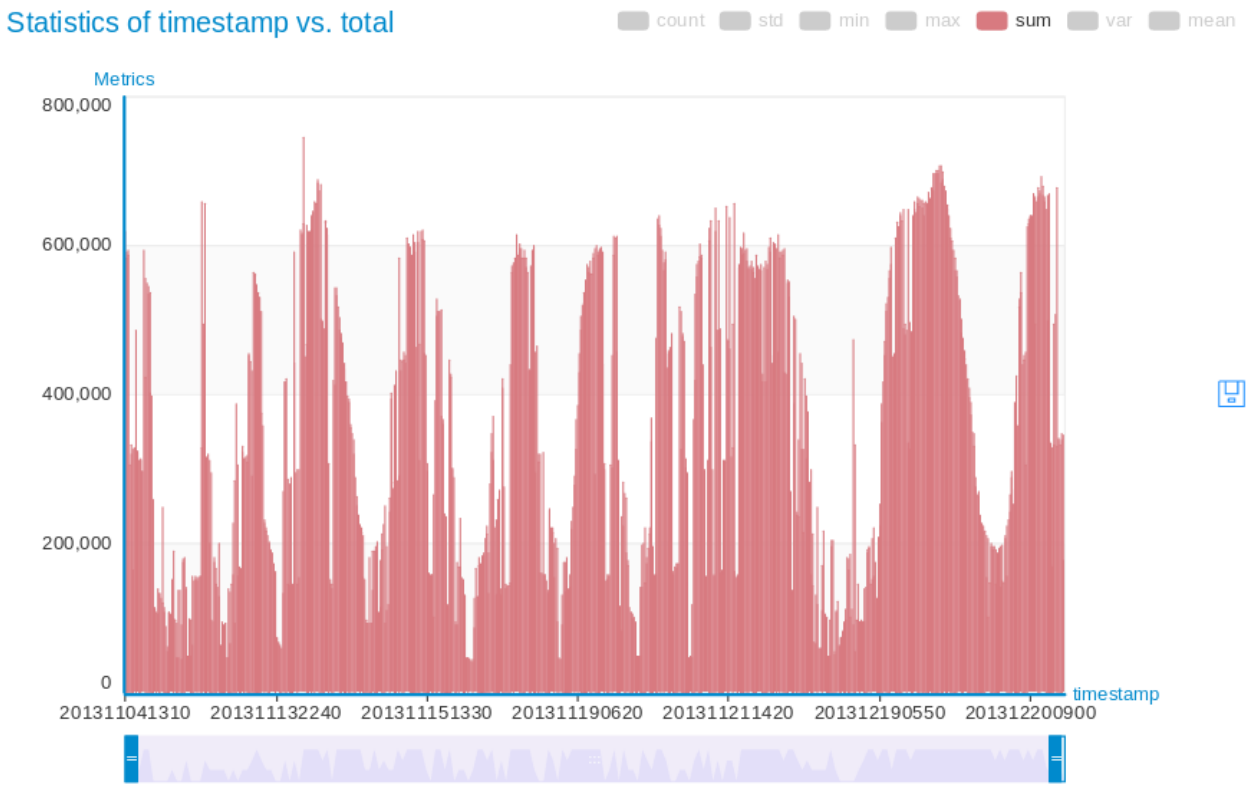

Fig 14: Telecommunication Usage for Rain

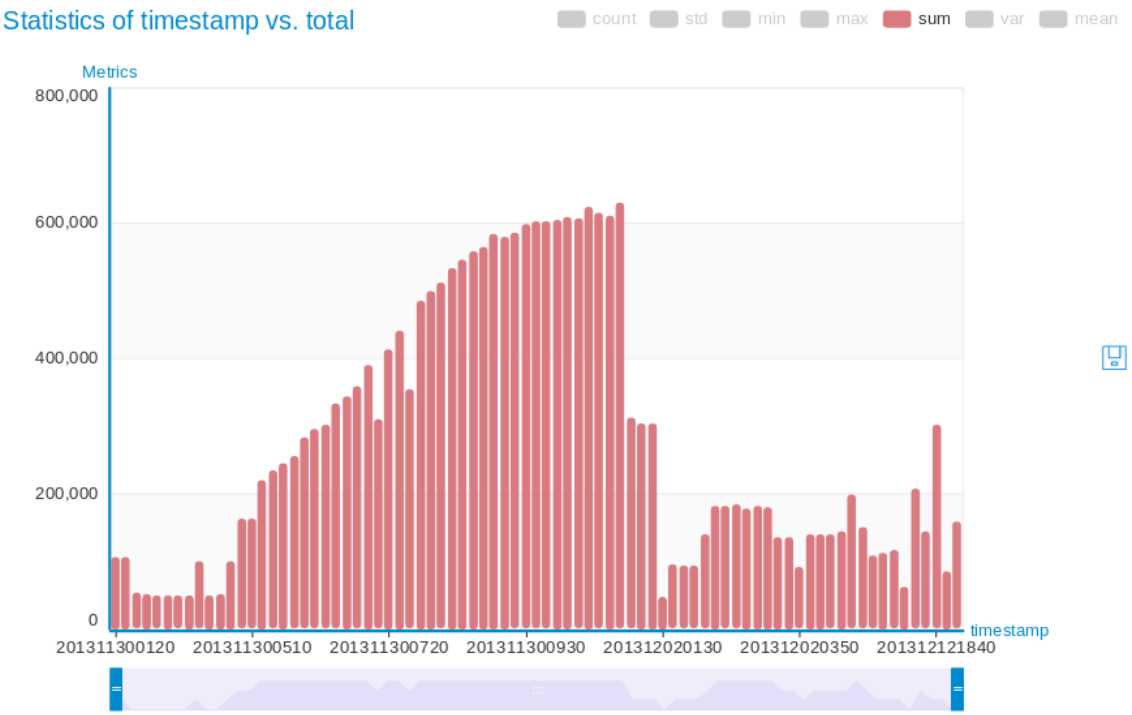

Fig 15: Telecommunication Usage for Snow

The results indicate that for the clear weather overall usage of calls, sms, internet activity is higher than rain days. Similarly, the data suggests that the telecommunication usage for rainy days is higher than the days when it snow.
To get a more detail understanding, following figures explicits the above described detail in more significant way.

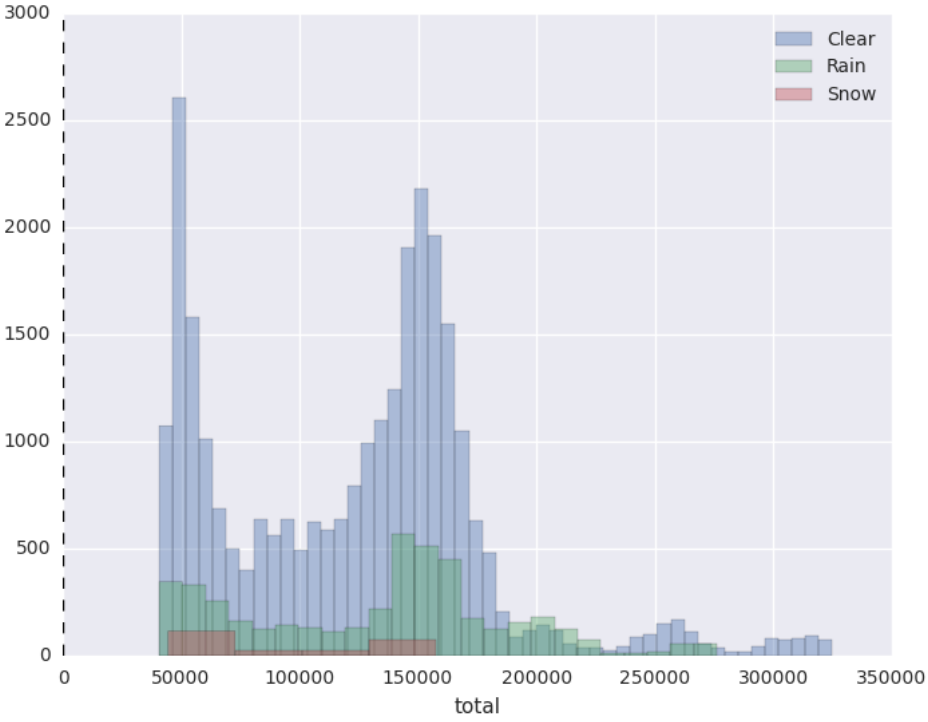

Fig 16: Telecommunication Usage for different weather conditions

Task 6: Communication Pattern of the users during the day and night in Trento.

I have got the heat map but it was not explaining in any detail, so I have drawn the line chart and bar chart to understand the behavior.

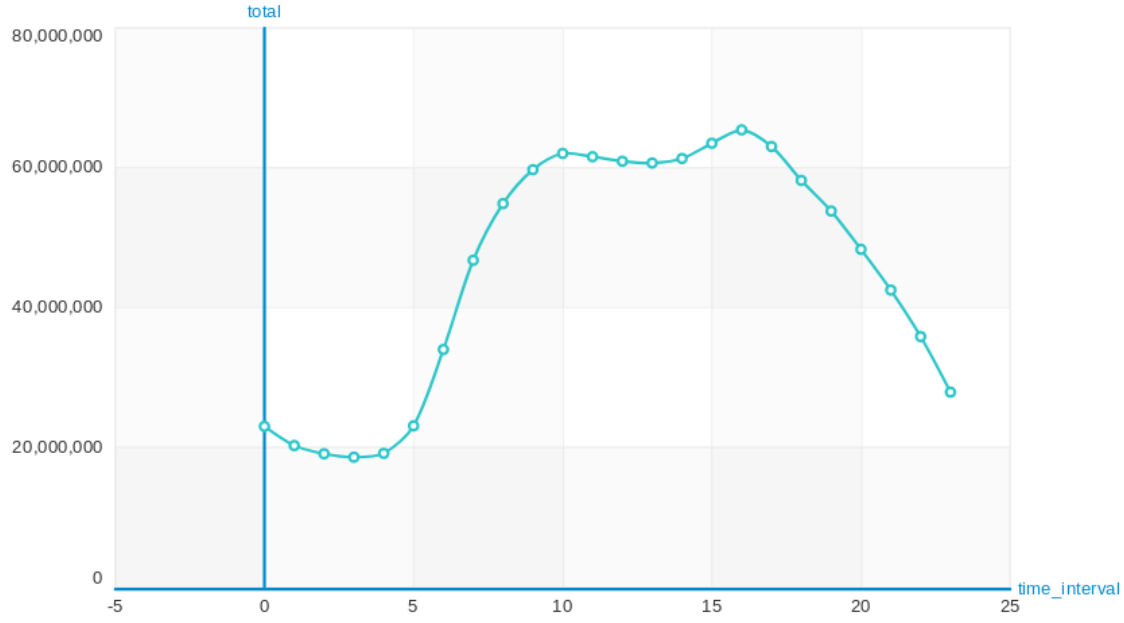

Fig 17: Communication Pattern during the day and night Trento

If we look at this line chart it shows that the maximum activity was recorded around 16. Overall the telecommunication activity was in an increasing order during the day time. After 16 its in a continuous decrease but overall the communication pattern suggests that the users were more active during the day time (8AM-8PM) as compare to the night.The following bar chart shows the same results.

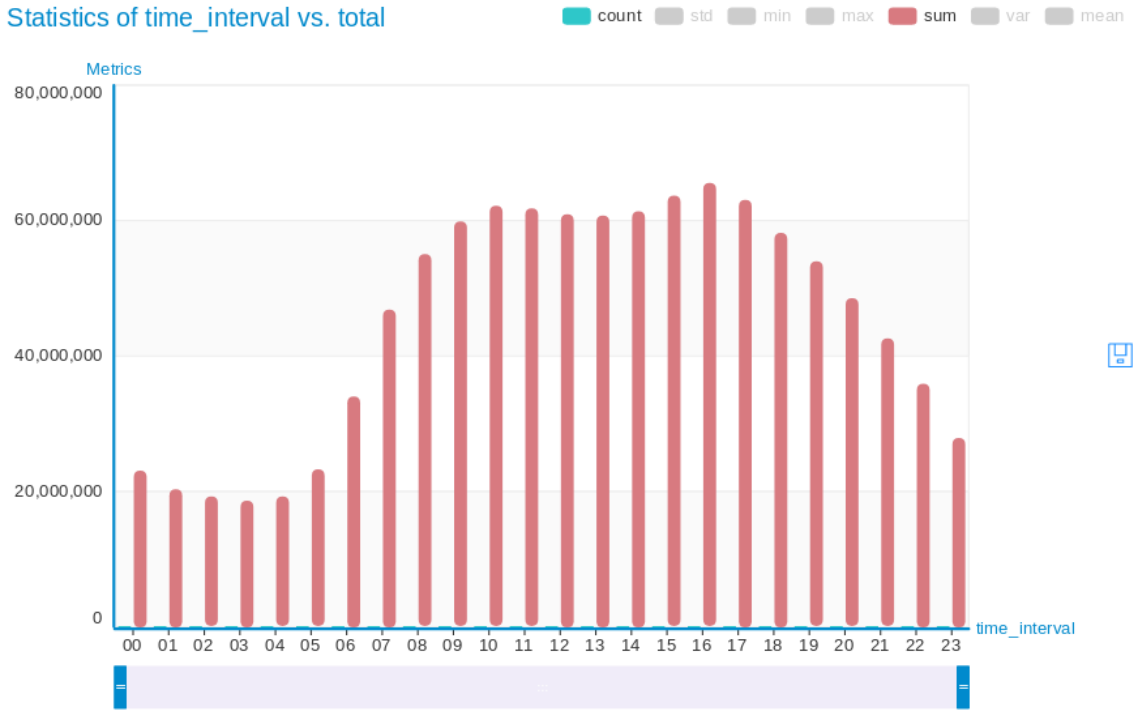

Fig 18: Communication Pattern during the day and night Trento

It was interesting to see that the total telecommunication activity of Trento is significantly lower than Milano. If we look at Figures 17 & 18, it is clear that the total telecommunication never come closer to the total activity of Milano. The highest telecommunication activity in Trento was recorded a bit higher than 6e7, whereas in Milano the highest activity recorded was almost 4e8.

Task 7: correlation between air quality and weather.

I read all the data for air quality, and filtered it out based on different categories given on the official site as (1 = VERY GOOD until 5 = VERY POOR). The idea was to read the weather data with different categories such as rain, sunshine, wind, temperature and then join the air quality data with this weather data and plot air quality categories against this weather data. This way we could have seen the weather representation when the air quality is good or very poor. Getting the weather data from grid's data was a hurdle for me. I somehow managed to get the data but because of the shortage of time I can not complete this task.