Ultimate Data Science Challenge: Results

Part 1: Exploratory Data Analysis

The details of the analysis are given in the file: Ultimate_challenge_part1.ipynb

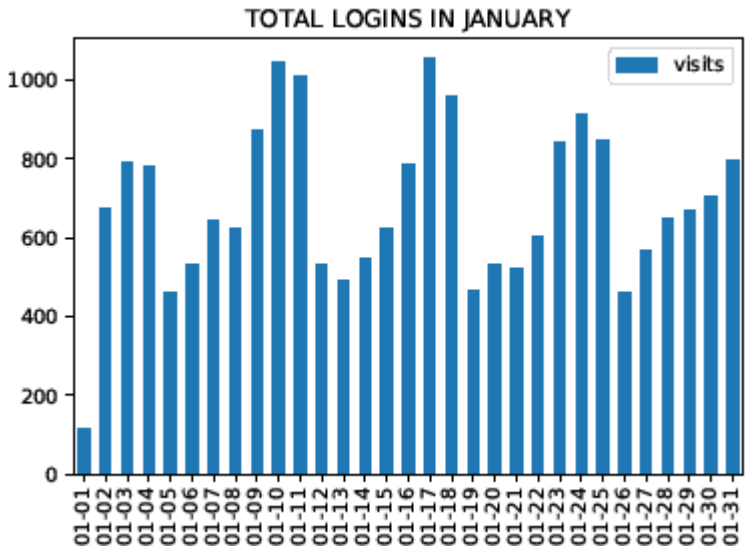
After reading the file and running its description, we see that the series contains 93142 values, but only 92265 are unique. The number of non-unique entries is quite large, and we will assume that these are the results of inadequate quality of data rather than the fact that different users made their logins at the same moment of time with one-second precision.

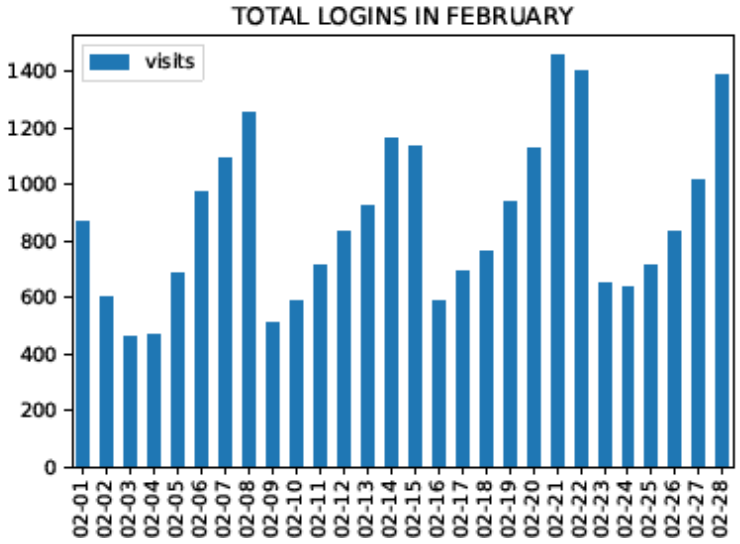
Let's aggregate the login counts based on the 15-minutes time interval. To do this, we introduce the column 'visits' and assign value of 1 to all rows in it. After that, we make 'login_time' our index column and resample taking the sum with respect to visits. The new data-frame contains 9788 entries in total and 407 NaN's corresponding to the absence of logins; we replace the latter entries by zeroes.

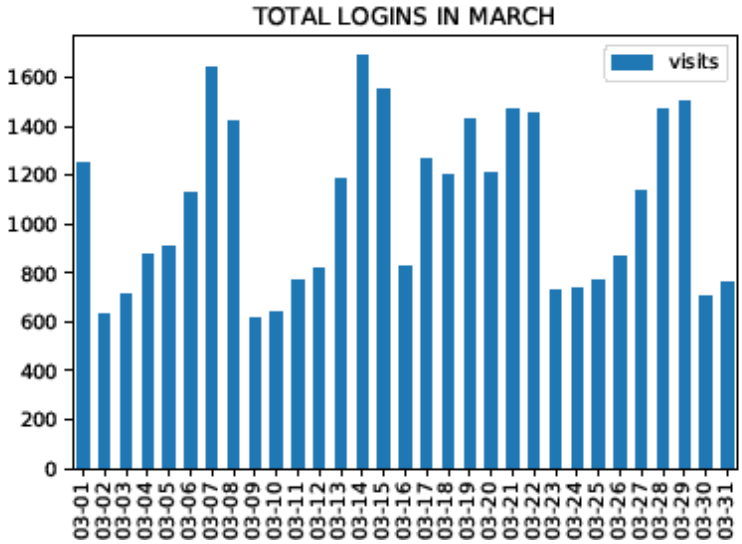
1) Let's aggregate the data over months and look at the distribution. We see that the largest number of logins occurred in March, while the smallest in April. This is because April is represented only by 13 days.

login_time	visits
1970-01-31 00:00:00	20308
1970-02-28 00:00:00	23855
1970-03-31 00:00:00	34000
1970-04-30 00:00:00	14102

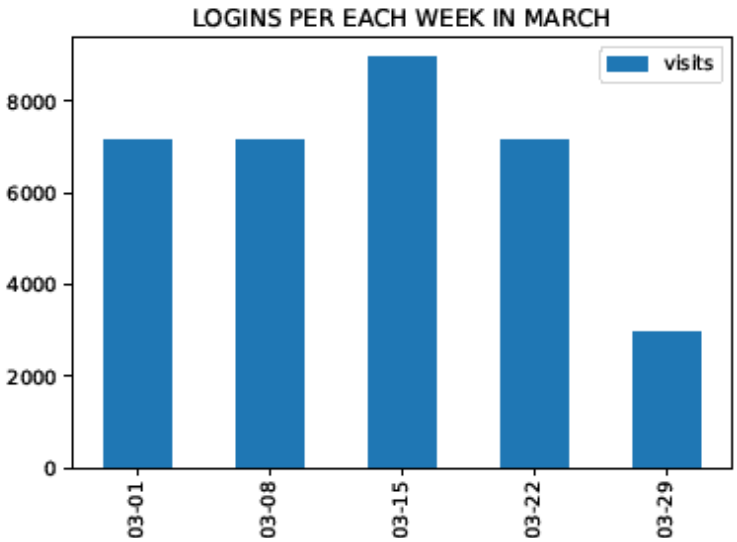
2) We then look at the distributions over days for January, February and March. The plots below clearly demonstrate the presence of periods approximately equal to 7 days. The graphs also suggest that the number of daily logins increases towards the end of the week and reaches maximum during weekend.







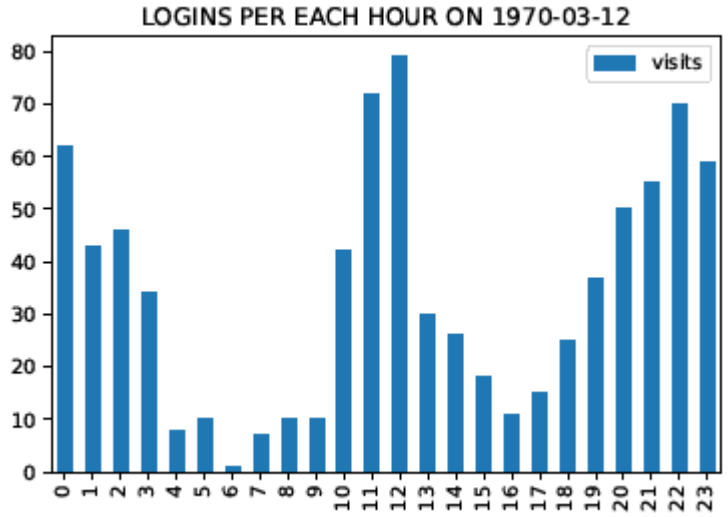
To ascertain the latter statement, we plot the weekly distribution of logins for month March, and then compute the average number of logins during weekends and during weekdays for the whole dataset. The first graph below shows that the total number of logins per each full week doesn't change much.

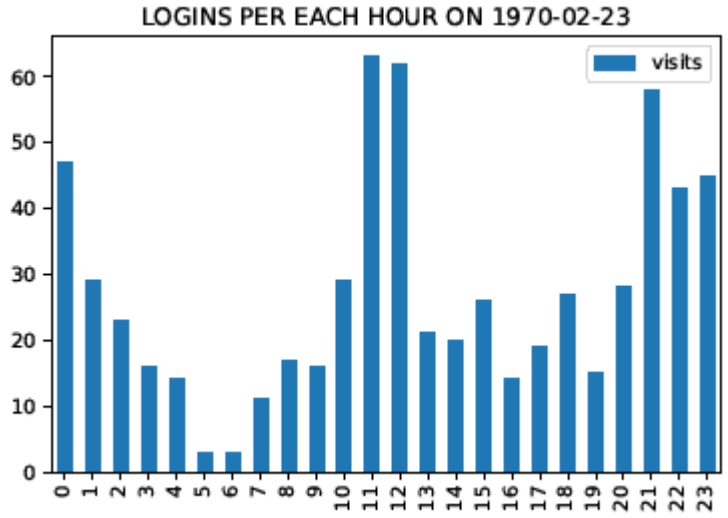


At the same time, the table presented next shows that the average number of logins during weekends is almost 1.5 larger than during weekdays.

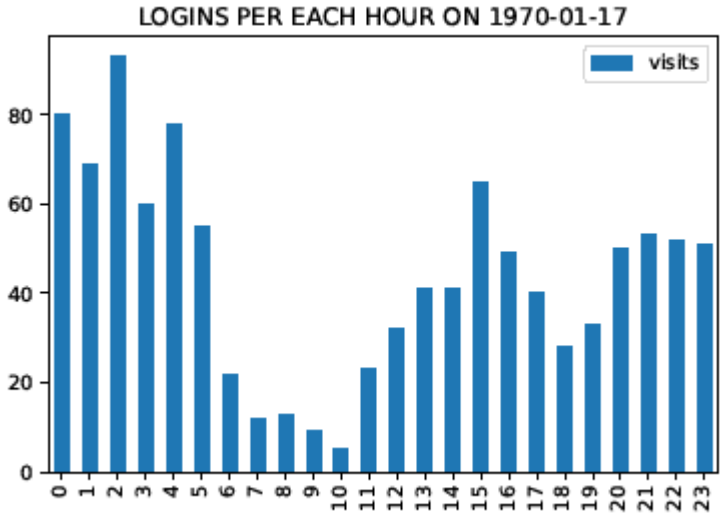
weekend	visits
0	7,984366
1	12,88507

3) Finally, we look at the hourly distribution of the number of logins. Let's first take two arbitrary business days, do the hourly aggregation and plot the results. We see that the number of logins increases after business hours and has a peak in the middle of the day, presumably during the lunch time.





From the other side, it is interesting to look at the typical hourly distribution of logins during a weekend day, say Saturday. '1970-01-17' is Saturday. We see that the number of logins is high during the late night/early morning hours, but the smallest during the late morning hours. During the day/evening hours the number of logins is relatively evenly distributed. This picture suggests that many users prefer to spend night before computers and take a sleep closer to morning hours on weekends.



Part 2 – Experiment and metrics design

1) As a key metric, I would choose the percentage of the trips that begin in one city and end in the other during a certain period of time. Since there are two cities Gotham and Metropolis, I would introduce two quantities p_G and p_M. They both in fact have the meaning of probabilities. p_G is the probability of occurrence of a trip that begins in Gotham and ends in Metropolis. p_M is the probability of occurrence of a trip that begins in Metropolis and ends in Gotham. I can think of two possible types of underlying reasons of the experiment, non-commercial and commercial. If the experiment has nothing to do with monetary gains or losses (for example, to ensure a better communication between the residents of two cities), one can define p_G and p_M as follows:

p_M = n_M/N_M, p_G = n_G/N_G,

where N_M and N_G are the total numbers of trips made by Metropolis and Gotham partners respectively, and n_M and n_G are respectively the numbers of any trips that end in the other city (Gotham and Metropolis in our case). However, if the experiment has commercial component, then n_M and n_G are not the numbers of arbitrary trips ending in the opposite city, but rather the numbers of the trips that bring some income (are long enough to cover the toll costs or meet some other requirements). As a key measure of success of the experiment, I would propose looking at whether observed values of p_M and p_G are equal or greater than certain target values p0_G and p0_M. For example, one can choose p0_G and p0_M both equal to 0.3 or 0.5.

2) In general, p_M and p_G are the parameters of the binary distribution. However, if n_M, N_M, n_G and N_G are large enough (greater than 10 each), estimators of p_M, p_G obey the normal distribution with the averages p_M, p_G given above and standard deviations sqrt[p_M (1 - p_M)/N_M], sqrt[p_G (1 - p_G)/N_G]. Since it is always convenient to work with normal

distributions, one can propose the following. Let's take a long enough time interval, so that we have enough instances to ensure that quantities n_M, N_M, n_G and N_G well exceed 10, and measure these quantities. After that we compute p_M and p_G and compare with the target values p0_M and p0_G. If at least one of p_G or p_M is considerably smaller than the corresponding target value, the experiment should be regarded unsuccessful and stopped; things should be left unchanged in such a case. If both p_G and p_M are much greater than the corresponding target values, the experiment should be deemed successful and the corresponding changes should be implemented. However, if the values p_M and p_G are close to the target values, we should resort to statistical testing.

Suppose p_M is close to p0_M. Let's formulate the null-hypothesis H0 that p_M = p0_M, and alternative hypothesis H1 stating that p_M is not equal to p0_M. We then apply the goodness of fit chi-squared fit to compute the corresponding p-value and decide whether the null-hypothesis should be rejected. If the p-value exceeds say 0.01, H0 should be retained and the part of experiment related to p_M should be considered successful since we decide that the target is met. We should note that we could formulate H1 in other ways, namely that p_M > p0_M and p_M < p0_M and apply the one-sided z-tests.

Similar analysis, if necessary, can be done for p_G. In addition, if we expect p_M to be close to p_G (and this is a reasonable expectation if two cities are similar), we can also apply the chi-squared test of homogeneity to test the null-hypothesis p_M = p_G regardless of whether both quantities are large or small.

Part 3 – Predictive modelling

All details, comments and conclusions can be found in the notebook: Ultimate_challenge_part3.ipynb