Natural Cycles Data Challenge Report - Giulia Gonella
May, 10th 2021

Overview and strategy

This report summarises the analysis of a dataset provided by *Natural Cycles* (NC) in the context of a data challenge for a Junior Data Scientist position.
The answers to the three main questions will be provided, together with a general explanation of the work-flow and approach to the analysis.

The general strategy for approaching the task is that of exploring and understanding the dataset first. This is the zero-step to any analysis, especially if the kind of data is not well known. This will be summarised at the beginning, after a short description of the tools used. The workflow for the analysis is divided into three parts corresponding to the three main questions, reflecting the procedure followed for approaching the task. However, the three questions trigger the curiosity for many others on the way, which could find answers with a more detailed analysis.

Tools

The analysis is developed using iPython 3. notebooks, supported by JupyterLab[1] environment.

Dataset

The dataset used for the task is a collection of features relative to women using *Natural Cycles Plan Pregnancy*, provided in form of a *.cvs* file. It is important to take some time at the beginning to understand the type of data provided, its quantity and quality.

The dataset has 13 columns (*features*), with both numerical (N) and non-numerical (NN) values, including boolean (B). In particular it contains information on:

- BMI (N)
- age (N)
- country (NN)
- woman been pregnant before (NN)
- education (NN)
- woman sleeping pattern (NN)
- number of menstrual cycles trying to conceive (N)
- outcome (NN)
- dedication (N)
- average cycles length (N)
- standard deviation (STD) of average cycle length (N)
- cycle being regular (B)
- frequency of intercourse (N).

[1] Ref: Jupyter homepage

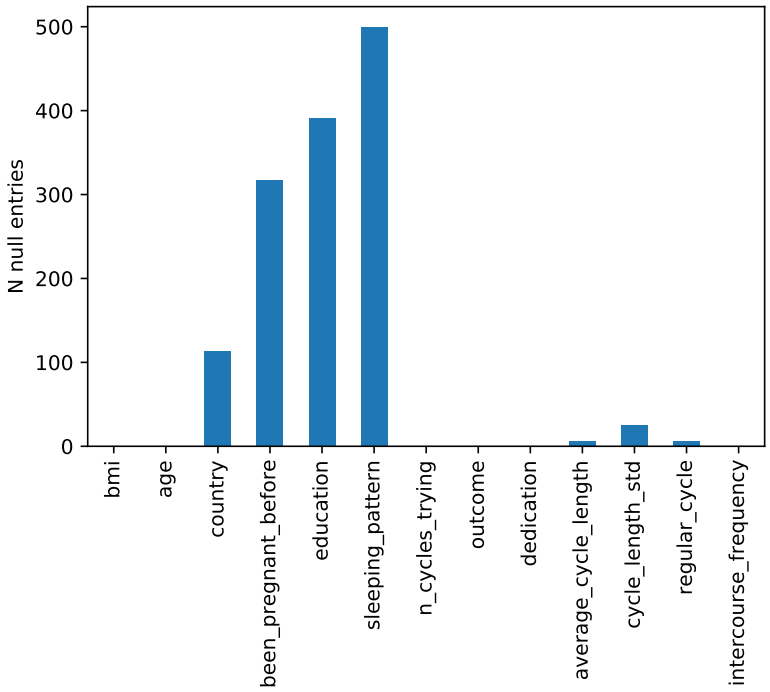


Figure 1: Amount of null values for each feature.

The dataset has namely 2000 entries, but actually 6 indices are missing, so the dataset collects data from 1995 women. Not all the features are available for all the women. In general the PANDAS library, used to parse the input data, recognises missing data and does not account for it in the specific operations, but it is good to know the quantity of missing data for each feature.

In particular the percentage of missing data in decreasing order is about 25% for sleeping pattern, 20% for education, 16% for woman been pregnant before, 6% for country, 0.3% for average cycles length and 0.3% for cycle being regular. These are represented in Fig. 1, and their impact can be considered negligible with respect to the whole size of the dataset.

Analysis

Q1 What is the chance of getting pregnant within 13 cycles?

The probability to get pregnant within 13 cycles can be estimated in this data sample by counting the fraction of women that got pregnant within 13 cycles over all the women in the sample:

$$P(\text{pregnant within 13 cycles}) = \frac{N_{\text{pregnant within 13 cycles}}}{N_{\text{tot}}}.$$

The total number of women in the sample is 1995, and out of them 1148 managed to get pregnant within 13 cycles.

The error on this measurement is retrieved by considering the statistical fluctuation of the samples. Assuming the counts used to compute the probability to be Poisson-distributed, their uncertainty is estimated with their square-root. The uncertainty on their ratio is then propagated with the standard error propagation. The final result is therefore:

$$P(\text{pregnant within 13 cycles}) = 0.57 \pm 0.02.$$

By looking at the distribution of the number of cycles for women that managed to get pregnant, it is interesting to notice that all the women that got pregnant, did it within 13 cycles. So based on this sample, 13 cycles is the maximum length for a woman to get pregnant by using NC. However the maximum value for the number of cycles length is 26 cycles, which means that women that did not get pregnant within 13 cycles

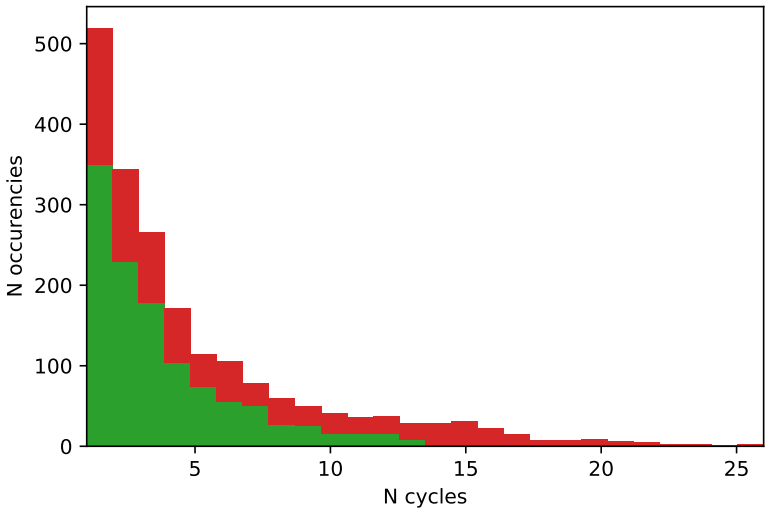


Figure 2: Histogram of number of cycles for women that got pregnant (green) and that did not get pregnant (red) within the time of use of NC.

- so not pregnant at all within the use of NC - (847, i.e. 42%) continued using NC beyond the 13$^{\text{th}}$ cycle, for a rather longer period up to the 26th. Fig. 2 shows the distribution of the number of cycles for women that got pregnant and did not get pregnant, where the number of cycles in the latter case indicates after how many cycles the negative outcome of the trials brought them to resume contraception or just interrupt the usage of NC.

Q2 How long does it usually take to get pregnant?

To address this question the distribution of the number of cycles trying to conceive (*number of cycles* in the following) is considered, for women that managed to get pregnant. The number of cycles it usually takes to get pregnant can be estimated via the mean value of this distribution.

A fit of the distribution is performed, assuming it to be Poisson-distributed[2]:

$$P(n;\mu) = \frac{\mu^n e^{-\mu}}{n!},$$

where the only parameter μ is the mean value, and its square-root its uncertainty.

The fit is performed and shown in Fig. 3. The number of cycles it usually takes to get pregnant is hence

$$\hat{\mu} =< N^{cycles} >= 1.8 \pm 1.4.$$

The result obtained by retrieving the mean directly from the dataframe is

$$< N_{\text{df}}^{cycles} >= 3.4 \pm 2.7.$$

This is a weighted mean over the series and not a proper mean value of the occurrences. However, the two values are compatible within the uncertainties, hence the second approach is used in the following to obtain an estimation of the mean values in first approximation, due to time constraints.

Q3 What factors impact the time it takes to get pregnant?

In order to analyse the factors that play a major role in the time it takes to get pregnant, a different procedure is followed depending on the factor being expressed as a numerical value or not. The sub-sample of women that got pregnant is used for the study.

For numerical values the most straightforward approach is to check the correlation among each feature and the number of cycles. As first approach the Pearson correlation coefficient[3] is obtained, which quantifies

[2] A test on the goodness of fit should be performed, this is not done for the limited amount of time available.

[3] This is defined as the covariance between two variables divided by their standard deviation.

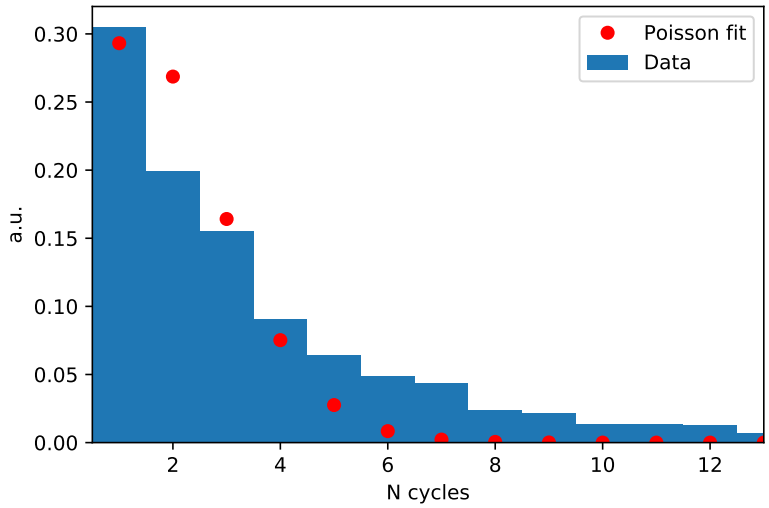


Figure 3: Histogram of the number of cycles trying to get pregnant for women that got pregnant, and fitted values from Poisson distribution. The distribution is normalised and uncertainties are not displayed.

Table 1: Correlation coefficient for each numerical feature with the number of cycles trying to conceive.

	BMI	Age	Dedication	Average cycle length	Intercourse frequency
Correlation with N cycles	-0.006	0.134	-0.176	-0.048	-0.195

the linear correlation and has values between -1 and 1, where -1 indicates a maximum anti-correlation while 1 a maximum correlation. Table 1 shows the correlation values for each numerical feature with the number of cycles.

It is already possible to see that the BMI with its very low (anti-)correlation, and the cycle length, with slightly higher but still low (anti-)correlation, play a very minor role in impacting the time to conceive. The age is the only feature with a positive correlation, but still not so enhanced. As expected, both the dedication in using NC and the intercourse frequency have slightly higher impact. In particular, the higher is the fraction of days with logged temperatures or the frequency of intercourses, the shorter is the time needed to conceive. This can also be seen in a visual way in Fig. 4 for the dedication feature. Here the occurrences for the dedication is displayed in 2-dimensional way with respect to the number of cycles: from the darker hexagons around 0.9 it is visible how the majority of women that had a higher dedication in logging data into NC app needed a shorter time to conceive.

Finally, as further test, the average value of the two most (anti-)correlated features is computed for each bin of N cycles. This allows to check and fit the trend with respect to the number of cycles. This is shown in Fig. 5. The result of the linear fit confirms the anti-correlation between each of the two features and the time needed to conceive expressed as N cycles in the plots.

The approach chosen to test possible dependencies in the non-numerical categories is to explore the distribution of the number of cycles trying to conceive for the different *classes* (corresponding to different answers) in each feature. The aim here is to search for any differences in the time needed to conceive for different classes, i.e. performing a sort of binned analysis for non numerical values.

The mean number of cycles is extracted for each class and its trend analysed. This is done for all the features, and an example can be seen in Fig. 6 where the distribution of the number of cycles is shown overlaid for each class of the *being pregnant before* feature. The mean number of cycles is retrieved for each class, and its trend shown in Fig. 7.

It results that the means for each class in all features are compatible with each other, and no great deviation is shown for a specific class. In other words, the trend of the mean values is quite flat with respect to each class. This applies to all the other features investigated as well[4], i.e. the education, sleeping pattern and even the cycle being regular or not. Some variations are present in the country feature, but this is not enough to

[4]The full set of plots can be found in the GitHub space, where also the code is located (`gonella_NC_data_science_challenge_code.ipynb`).

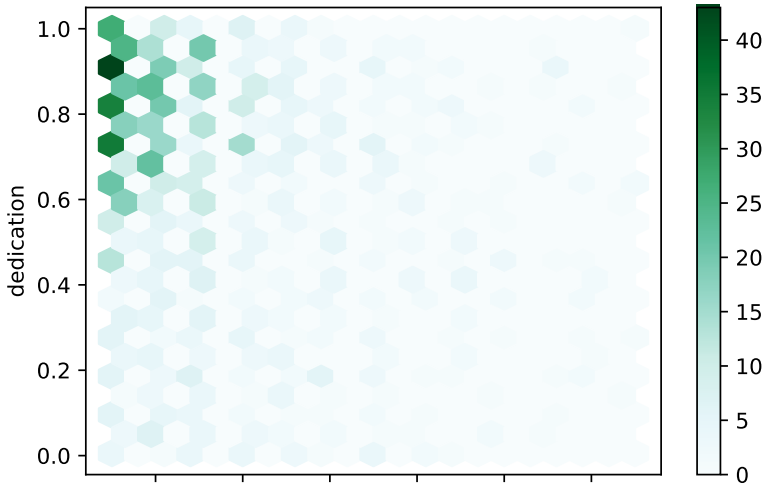
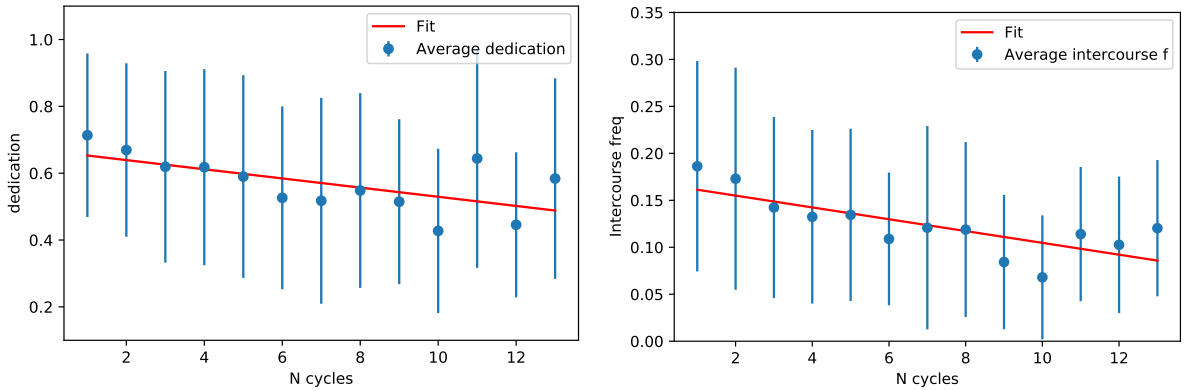


Figure 4: Bi-dimensional histogram of the dedication with respect to the number of cycles.



(a) (b)

Figure 5: Average values for dedication (a) and intercourse frequency (b) in each bin of number of cycles. The uncertainties are the STD on the average value and the linear fit line is also shown.

draw conclusions on a dependency which would need additional information to be discussed.

Finally, it appears that the main impact in the time needed to conceive for women getting pregnant comes from the dedication in using NC and the frequency of intercourses. The age has a small impact, while the other features does not seem to play a major role in the time.

Conclusion and notes

Based on the dataset provided, the probability of getting pregnant within 13 cycles results to be 0.57 ± 0.02. This time, i.e. 13 cycles, is also the longest time reported for women that got pregnant, while it usually takes around 2 cycles to get pregnant. This value is estimated as the average time of getting pregnant through a fit to the distribution, in particular resulting to be $< N^{cycles} >= 1.8 \pm 1.4$. Finally, the main features that impact the time needed to conceive turn out to be the dedication of logging data into NC app and the frequency of intercourses, which both decrease the time needed. Besides those, the age has a very small correlation, in particular the higher it is, the slightly longer it takes to get pregnant. The other categories do not show a great correlation with the average time needed to conceive, which is stable with respect to all analysed classes in each feature.

The dataset offers way more paths of exploration which could not be pursued due to the lack of time, but can trigger further questions and can point to interesting properties. For example, the analysis is concentrated on the women that managed to get pregnant, but it worth exploring what impacts most the chance or not to

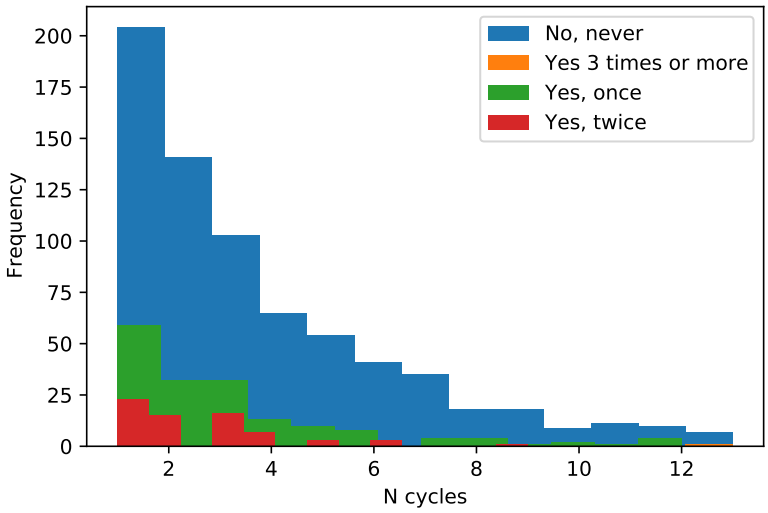


Figure 6: Distribution of number of cycles for all classes of the feature *being pregnant before*. The plot is not stacked.

get pregnant at all. Also, an interesting question that might find answer in this or a similar dataset, is to test which feature (if any) impacts the choice of dropping the use of NC after a certain time of unsuccessful tries. The data provided gives great boost to the imagination of different kinds of investigation.

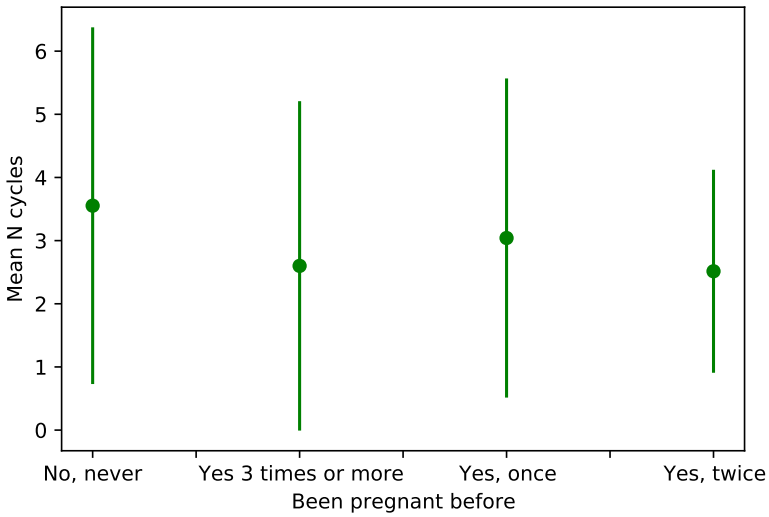


Figure 7: Mean values for each class of the feature *being pregnant before*. The uncertainties are the STD on the mean.