Natural Cycles

Assignment for Senior Data Scientist position

-

Jeroen Buil

Contents

- **Introduction**

- **Exploratory Data Analysis**

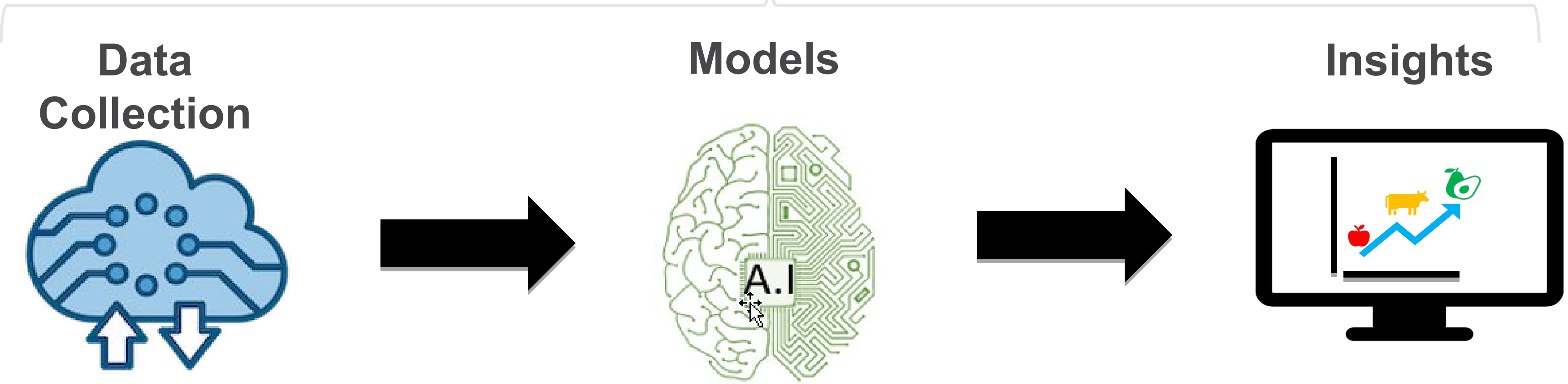
- **Questions:**
    1. *What is the chance of getting pregnant within 13 cycles?*
    2. *How long does it usually take to get pregnant?*
    3. *What factors impact the time it takes to get pregnant?*
    4. *ML vs non-ML methods approach?*

Quick introduction

- **Jeroen Buil**
  - Senior Data Scientist / Biomedical Engineer

Focus:

Practically applying AI for



Data Collection → Models → Insights

2013 → 2024

MAX PLANCK GESELLSCHAFT

imec

OnePlanet Research Center

ABN·AMRO

FOOD & BIOBASED RESEARCH WAGENINGEN UR

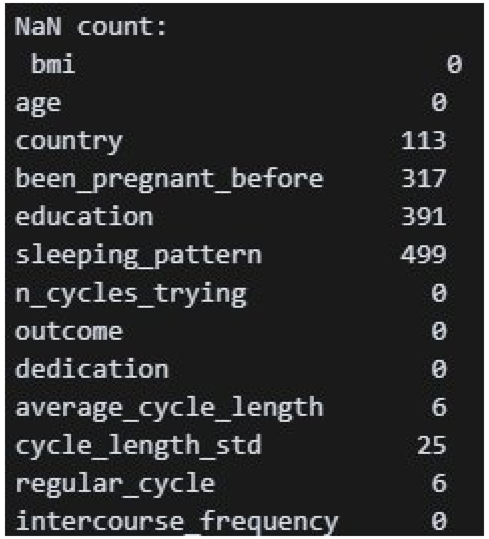
EDA: Exploratory Data Analysis

=> First step before any analysis

- Q: Is the data suitable?

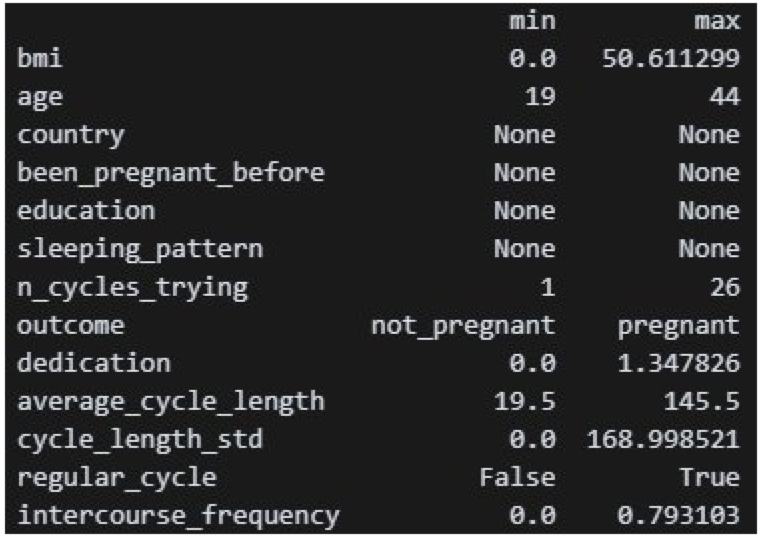
EDA: Quick glance

- Missing data:
  - ~40% of samples miss some data points
  - => Need to remove (or ideally fill) depending on analysis



```
NaN count:
 bmi                       0
age                       0
country                 113
been_pregnant_before    317
education               391
sleeping_pattern        499
n_cycles_trying           0
outcome                   0
dedication                0
average_cycle_length      6
cycle_length_std         25
regular_cycle             6
intercourse_frequency     0
```

- Data range:
  - BMI of 0 => not possible
  - Big spread in (average) cycle length
  - Dedication +100% => not possible



```
                                 min         max
bmi                              0.0   50.611299
age                               19          44
country                         None        None
been_pregnant_before            None        None
education                       None        None
sleeping_pattern                None        None
n_cycles_trying                    1          26
outcome                 not_pregnant    pregnant
dedication                       0.0    1.347826
average_cycle_length            19.5       145.5
cycle_length_std                 0.0  168.998521
regular_cycle                  False        True
intercourse_frequency            0.0    0.793103
```

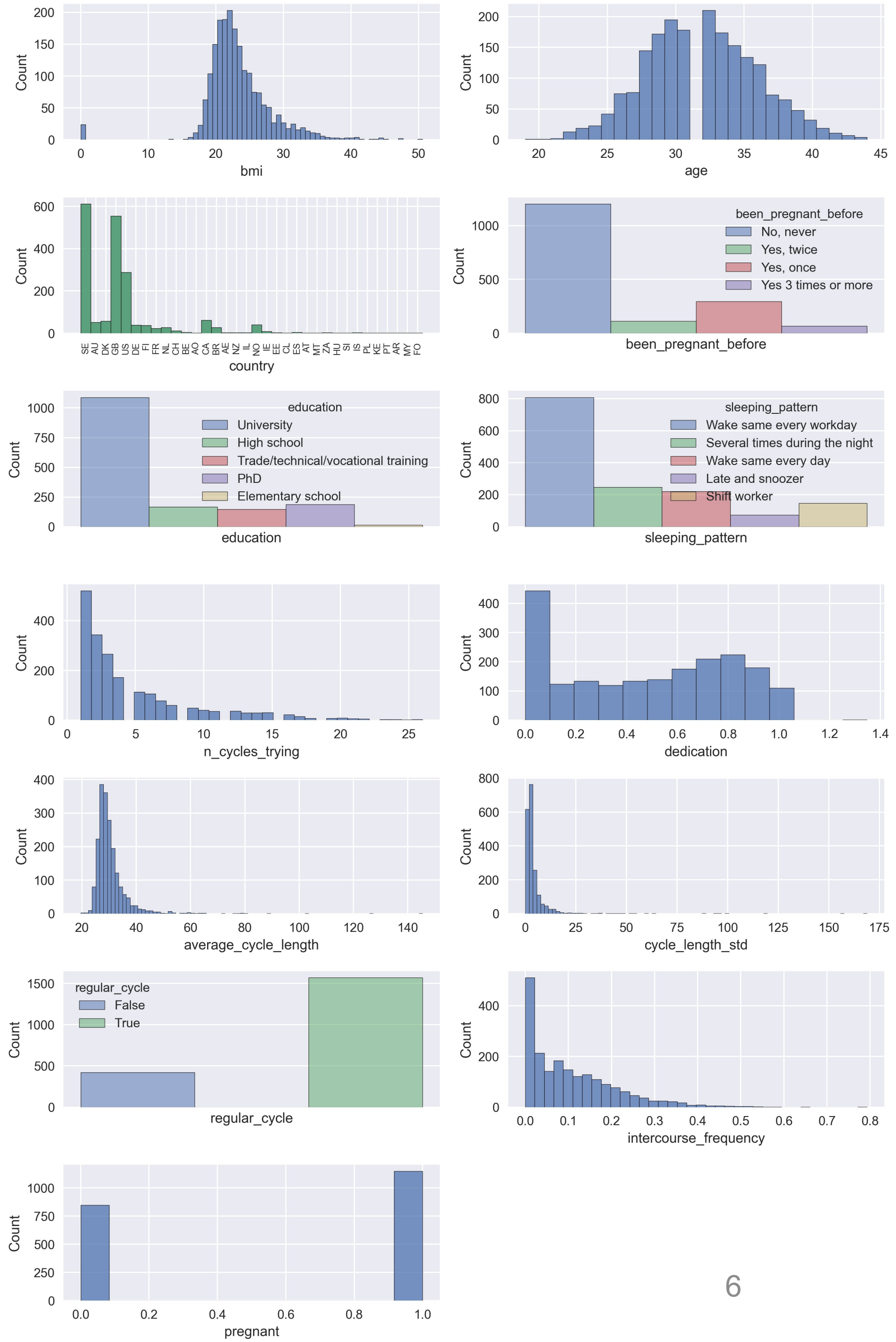
EDA: Histograms

- BMI:
  - Contains missing data (BMI of 0) => remove!
  - Underweight (BMI <16) + Morbidly Obese (BMI > 40) present => Keep or consider outliers?

- Cycle length:
  - (Very) high cycle lengths (>35 - 145 days)
    => users with PCOS*? => Keep or consider outliers?

- Unbalanced variables:
  - Country
  - Been pregnant before
  - Education
  - Sleeping Pattern
  - Cycle regularity
  - Regular_cycle

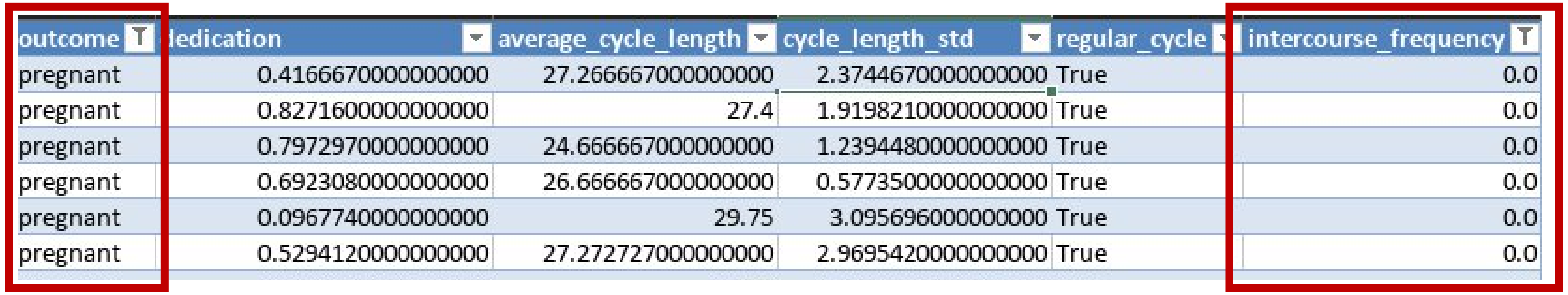
  => Makes them harder to use as predictors!



*Polycystic ovary syndrome (PCOS) is characterised by irregular menstrual cycles, higher chance of diabetis type II and difficulty getting pregnant - https://en.wikipedia.org/wiki/Polycystic_ovary_syndrome

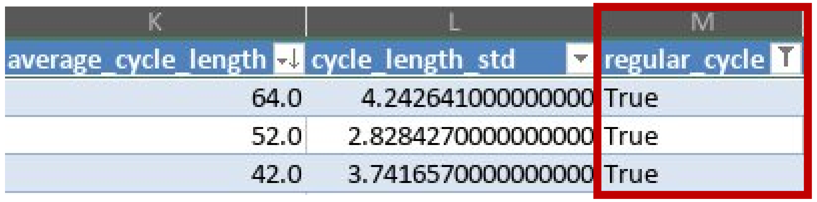
EDA: Data Inconsistencies?

- No intercourse, but still pregnant?
  - => remove samples



outcome	ledication	average_cycle_length	cycle_length_std	regular_cycle	intercourse_frequency
pregnant	0.4166670000000000	27.266667000000000	2.3744670000000000	True	0.0
pregnant	0.8271600000000000	27.4	1.9198210000000000	True	0.0
pregnant	0.7972970000000000	24.666667000000000	1.2394480000000000	True	0.0
pregnant	0.6923080000000000	26.666667000000000	0.5773500000000000	True	0.0
pregnant	0.0967740000000000	29.75	3.0956960000000000	True	0.0
pregnant	0.5294120000000000	27.272727000000000	2.9695420000000000	True	0.0

- cycles_length > 50 days are NOT regular cycles!
  - => Regularity of cycle is determined by cycle_length_std (< 5 days)



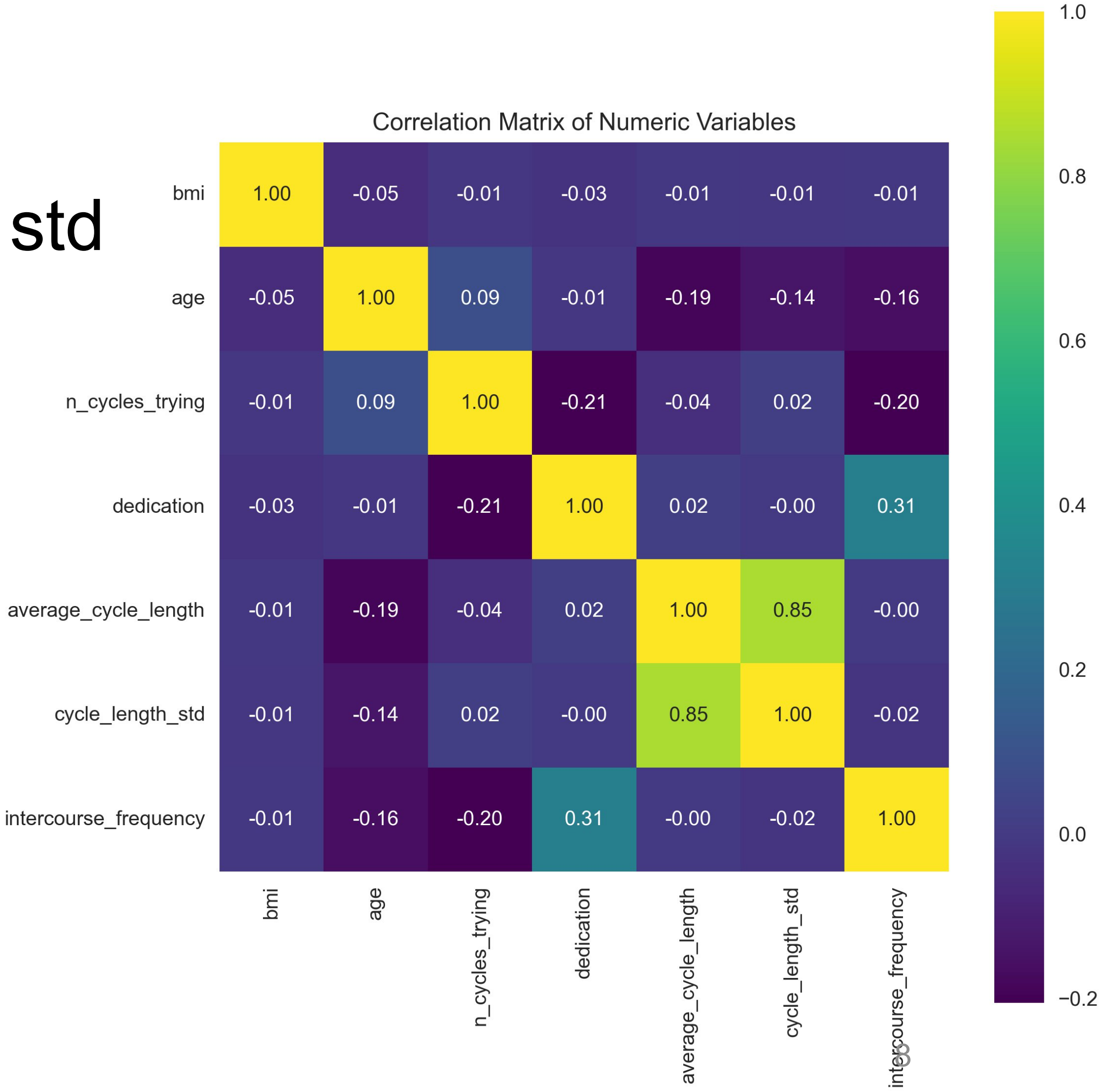
K	L	M
average_cycle_length	cycle_length_std	regular_cycle
64.0	4.242641000000000	True
52.0	2.828427000000000	True
42.0	3.741657000000000	True

EDA: Data correlation

High correlation:

- Average cycle length <=> Cycle length std
  - This is to be expected

  => consider keeping only one of the two



Correlation Matrix of Numeric Variables

	bmi	age	n_cycles_trying	dedication	average_cycle_length	cycle_length_std	intercourse_frequency
bmi	1.00	-0.05	-0.01	-0.03	-0.01	-0.01	-0.01
age	-0.05	1.00	0.09	-0.01	-0.19	-0.14	-0.16
n_cycles_trying	-0.01	0.09	1.00	-0.21	-0.04	0.02	-0.20
dedication	-0.03	-0.01	-0.21	1.00	0.02	-0.00	0.31
average_cycle_length	-0.01	-0.19	-0.04	0.02	1.00	0.85	-0.00
cycle_length_std	-0.01	-0.14	0.02	-0.00	0.85	1.00	-0.02
intercourse_frequency	-0.01	-0.16	-0.20	0.31	-0.00	-0.02	1.00

EDA: Conclusion

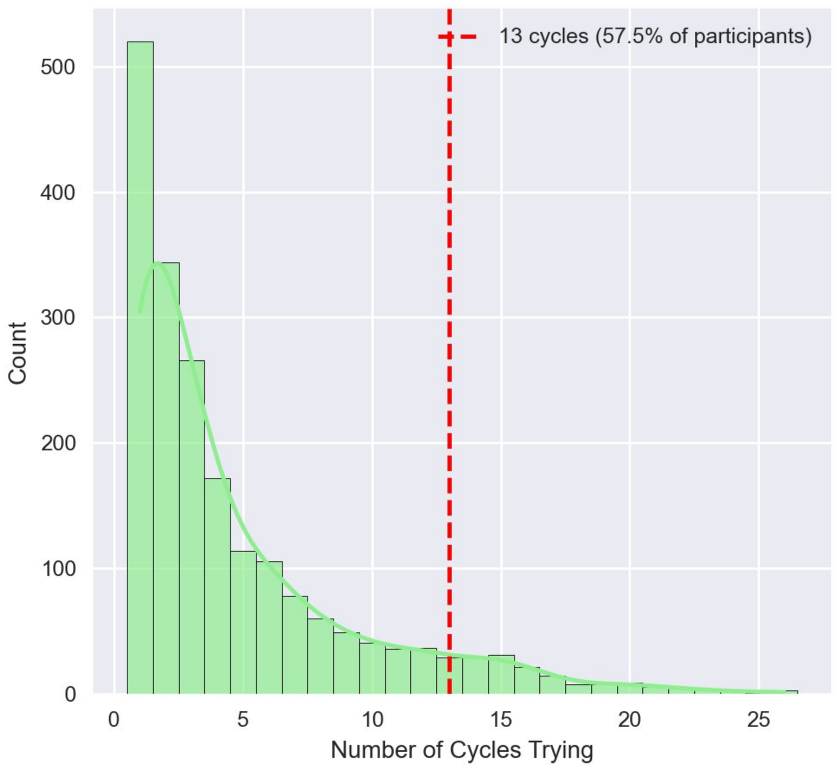
- Data seems usable!

- Data requires some clean-up
    - Not all samples + variable are useable for modelling

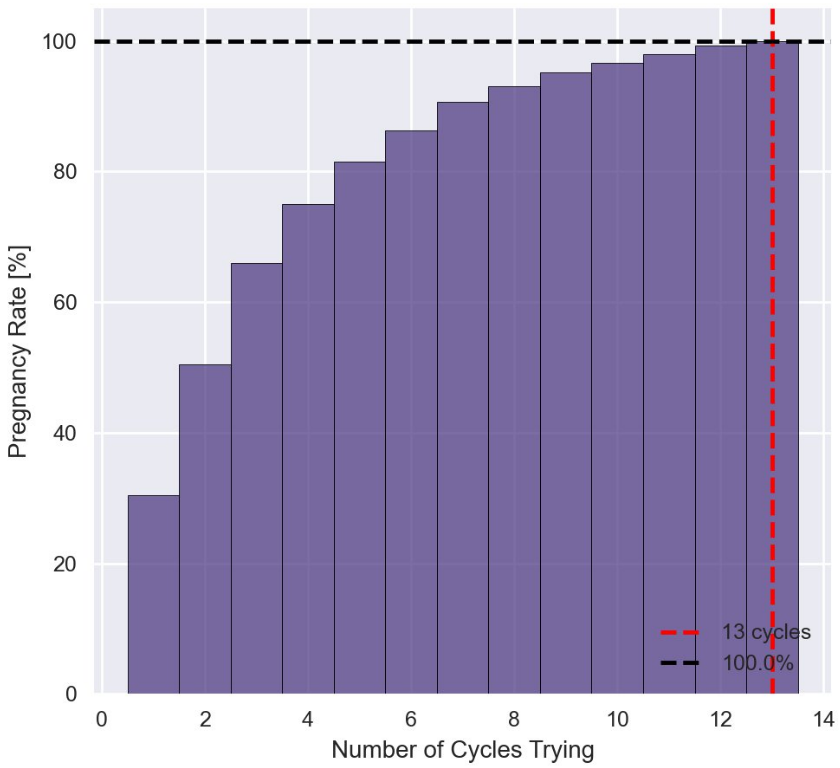
Q1: What is the chance of getting pregnant within 13 cycles?

Group:	Chance	Participants
Got pregnant within the study	57.5%	1148 / 1995participants
Got pregnant within the study within 13 cycles	**57.5%**	**1148 / 1995 participants**
Got pregnant within 13 cycles out of all pregnant participants	100%	1148 / 1148 pregnant participants

All Participants: Cycles Needed to Conceive



Pregnant Participants: Cumulative Pregnancy Rate



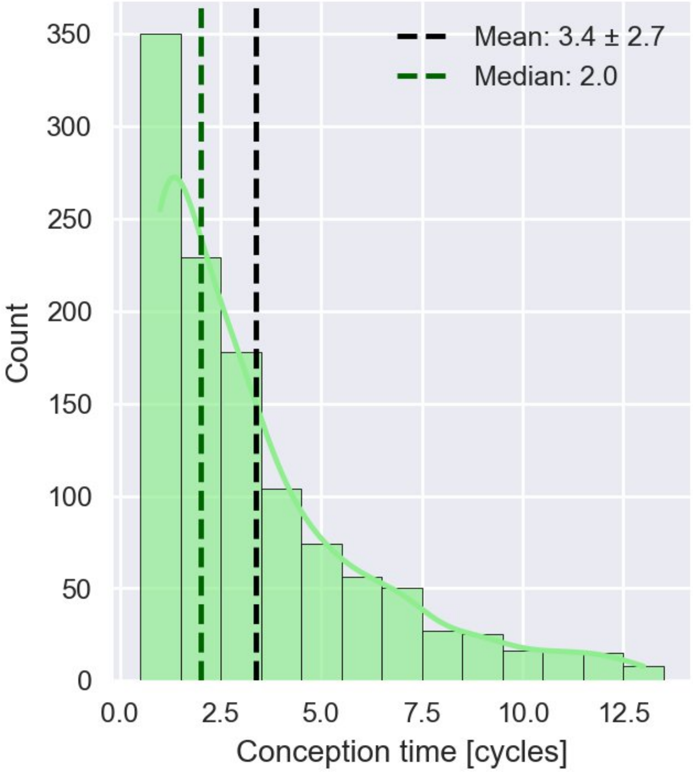
Q2: How long does it usually take to get pregnant?

- What is “usually”? => 50%? 90%?
- Answer expressed in Days/Cycles?
- Note: only pregnant participants are included

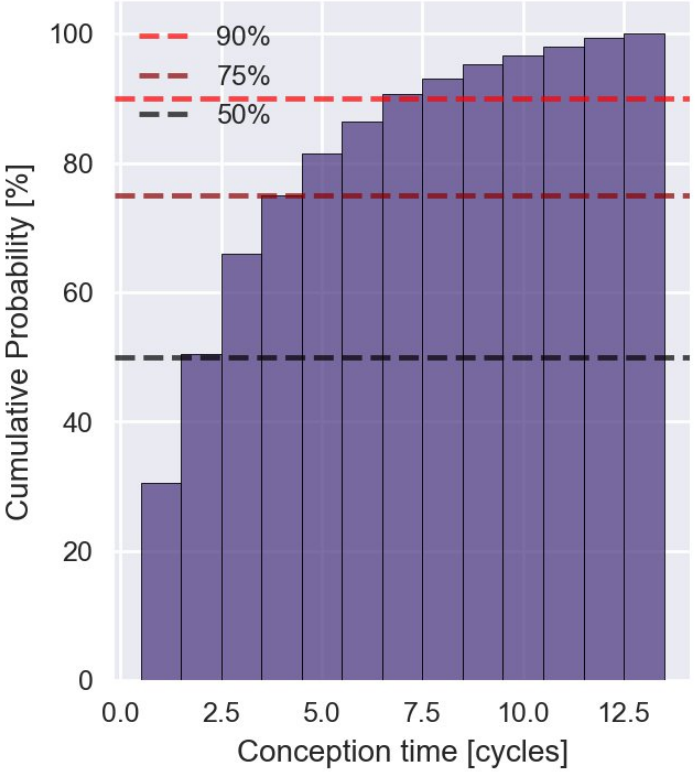
Answer:

- Majority of participants (>50%) got pregnant ≤ 2 cycles
- 90% did so ≤ 7 cycles

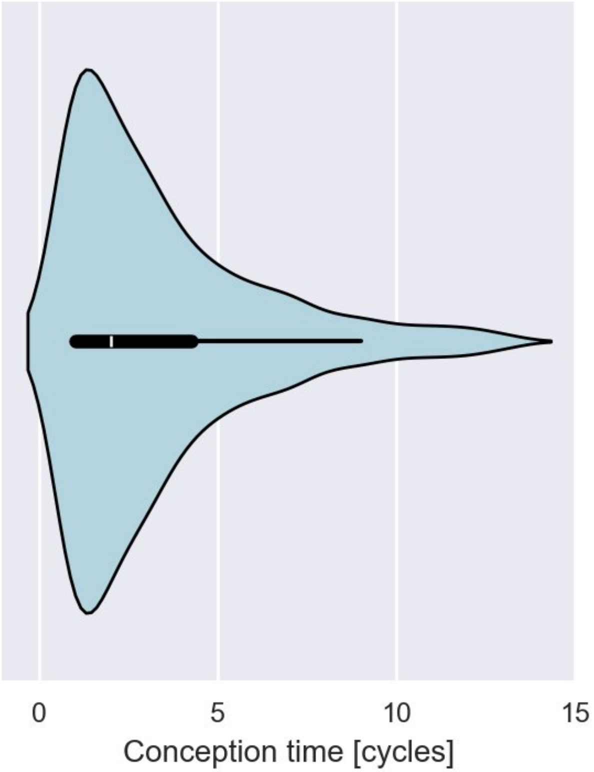
Histogram of conception time



Cumulative Histogram of conception time



Violin Plot of conception time



Q2: How long does it usually take to get pregnant?

Additional remarks:

- Not all participants got pregnant during the study
    - Only 1139/1975 participants (57,8%)
- Longer study time might show longer ‘average’ conceptions times

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

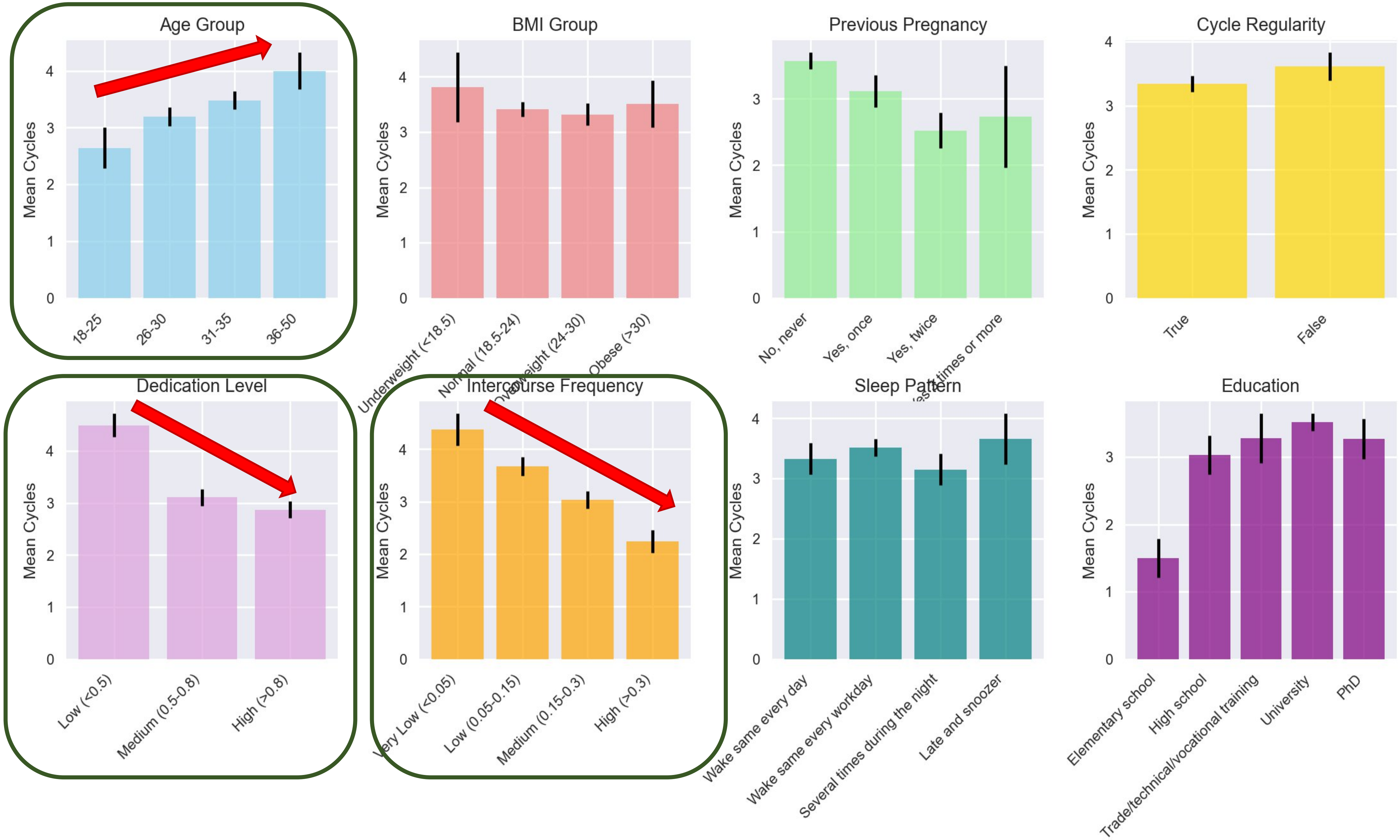
- Two approaches:
  - **Non-ML** → K.I.S.S. | *Keep It Simple Stupid*
  - ML:
- **Why?**
  - (Relatively) Low amount of variables
  - Many categorical variables are unbalanced, can cause bias
  - Simpler method => easier interpretable results

First more data cleaning needed

Removed “outliers”	Why
Non-pregnant participants	We want to know how quickly, not IF they got pregnant.
Pregnant participants with intercourse frequency 0	Can’t get pregnant without sexual intercourse.
BMI < 12	Unrealistic value
Dedication > 1.00	You can’t log more than 100% of days
Remove samples with NaN’s / empty values	We want to investigate all parameters

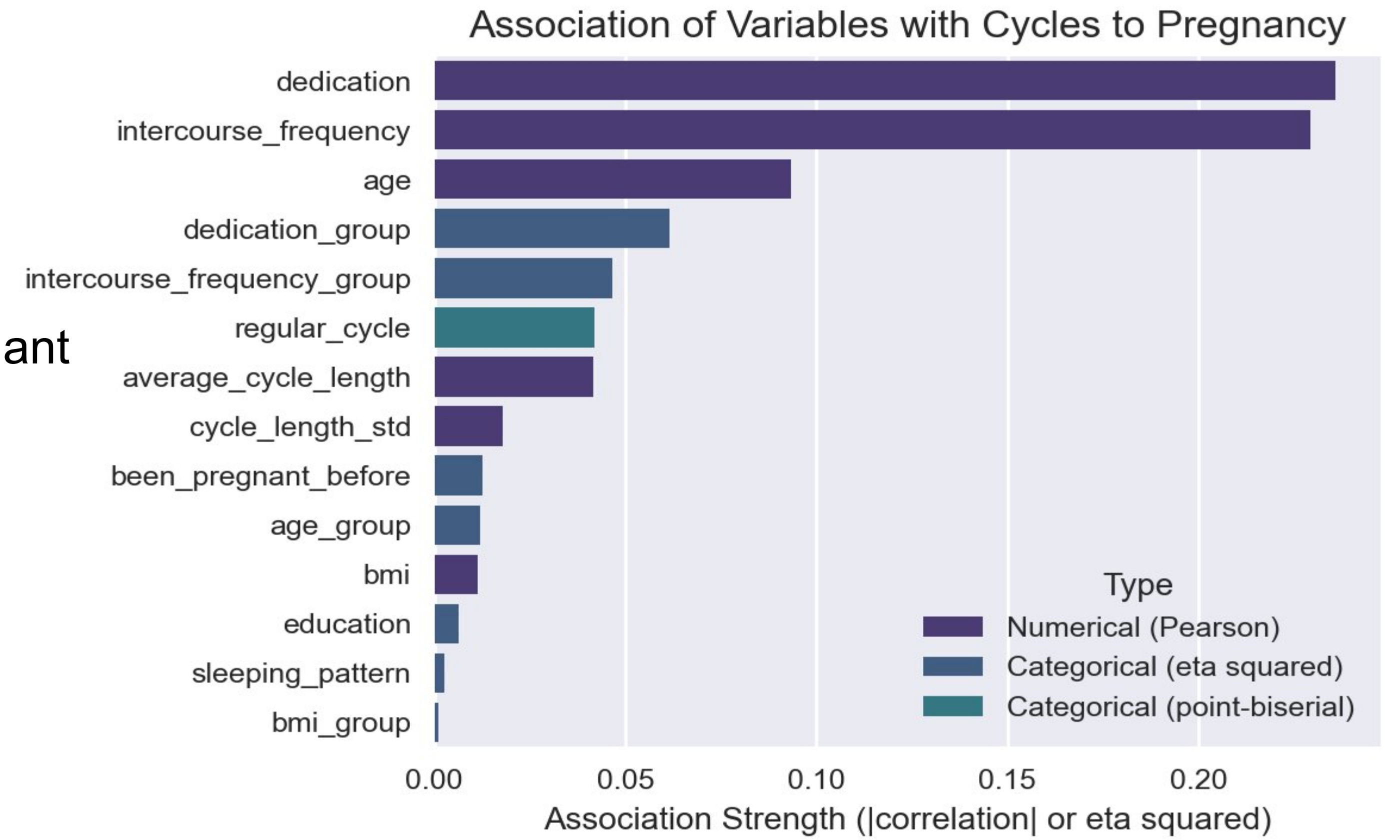
Clean samples: **1121** (out of 1995)

Simple binning in groups already shows obvious effects



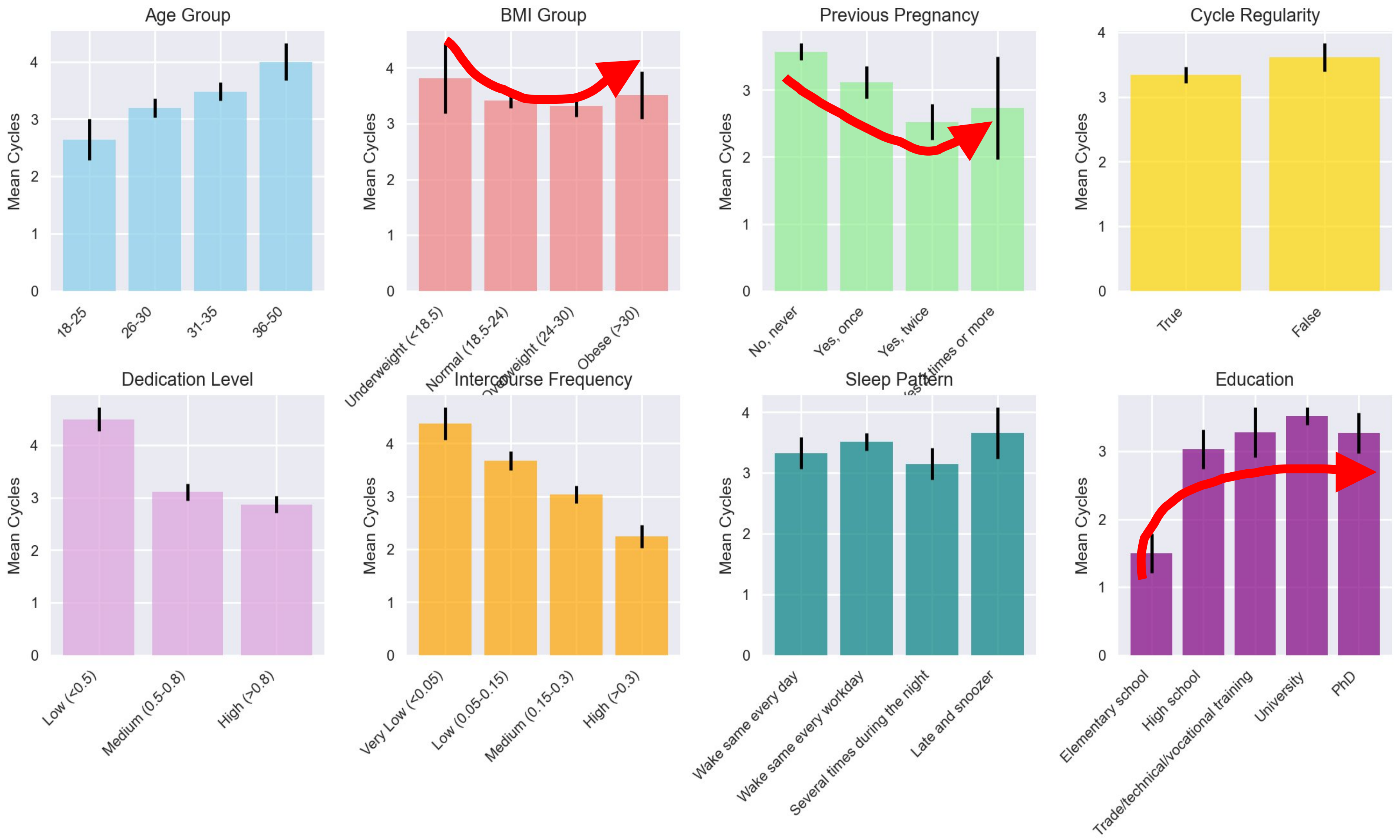
Statistical analysis confirms (linear) effect

- All 3 seem quite logical:
  - **Dedication**
    - Better tracking => quicker pregnant
  - **Intercourse frequency**
    - Regular intercourse => quicker pregnant
  - **Age**
    - Younger => more fertile => quicker pregnant



Also non-linear effects visible

Limitation of linear analysis => More advanced modelling needed?



Q4: How would your approach change if you were to use different techniques (e.g., ML or non-ML methods)?

- Two approaches:
  - Non-ML
  - **ML:**

- **Why?**
  - Easier to to investigate complex (non-linear) patterns
  - Resulting model (potentially) useable for predicting conception time

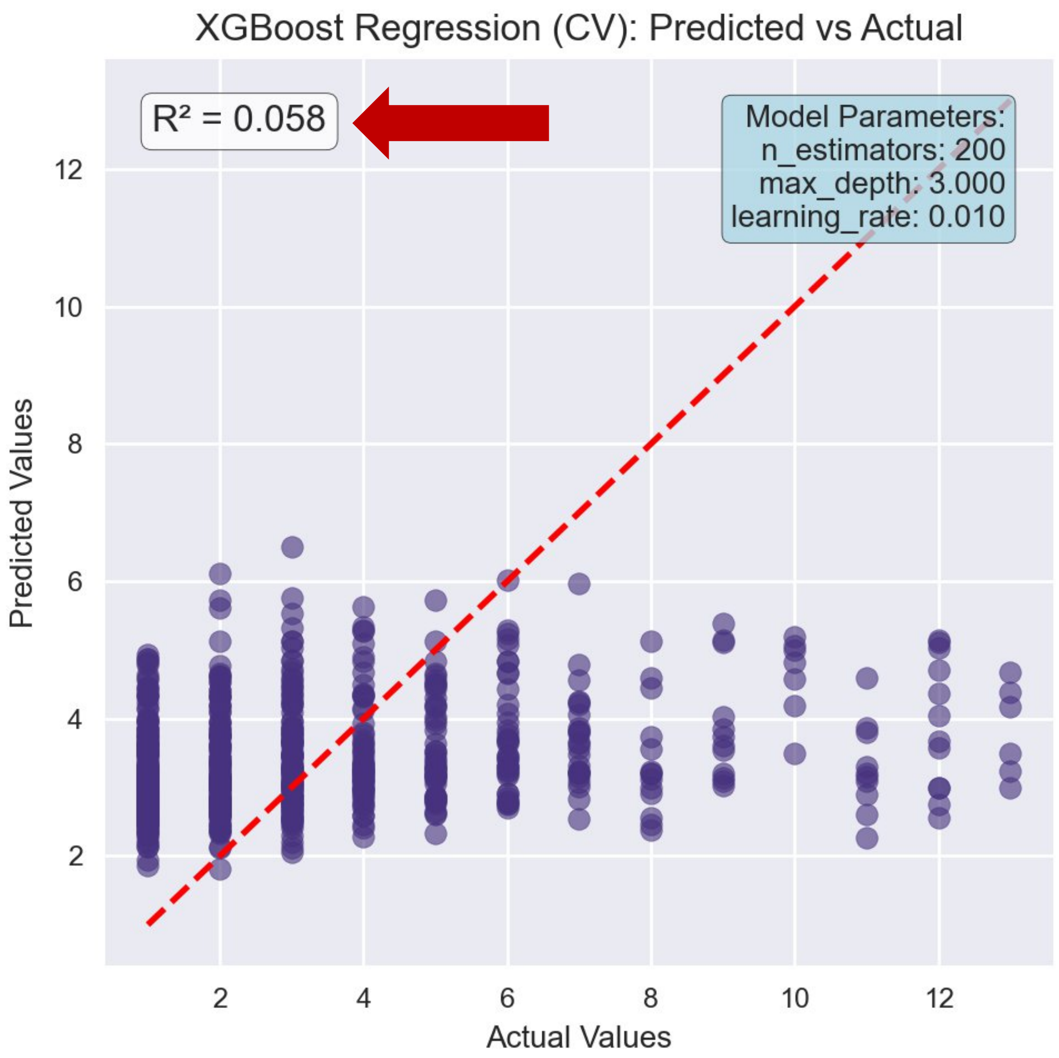
Model choice

XGBoost

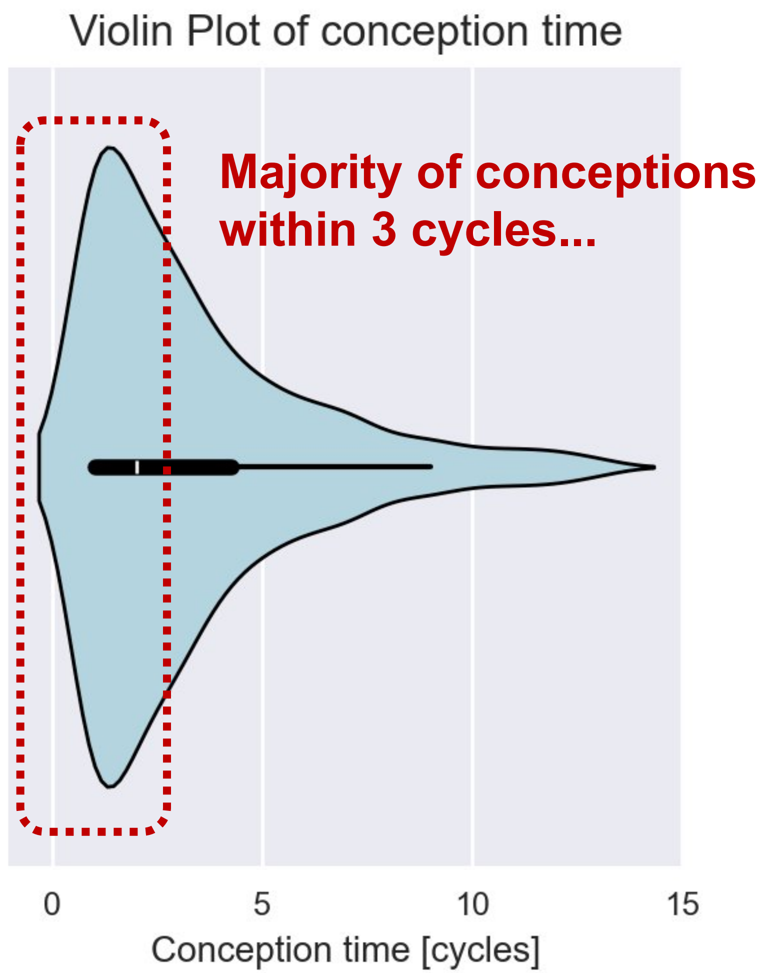
- When in doubt: **XGBoost!**
  - Easy implementation (especially compared to neural networks)
  - Accepts categorical + numerical data
  - Allows for classification + regression modeling
  - Build-in explainability tools (feature importance)

- Note: average performance over 5-fold cross validation is shown

Poor performance XGBoost regression model...



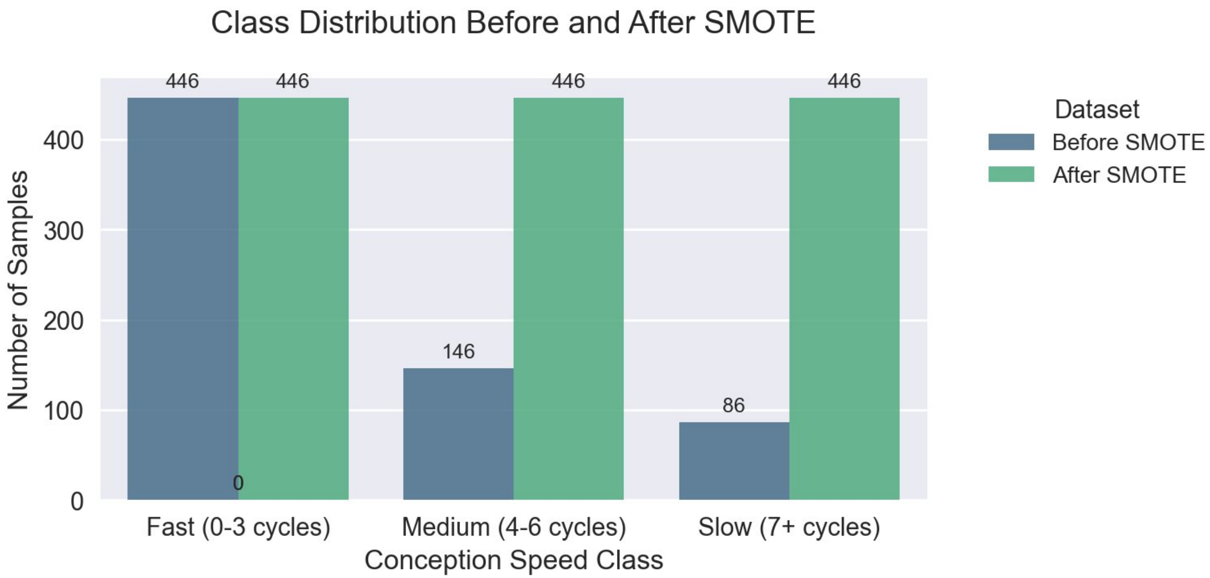
- Likely due to **data imbalance**

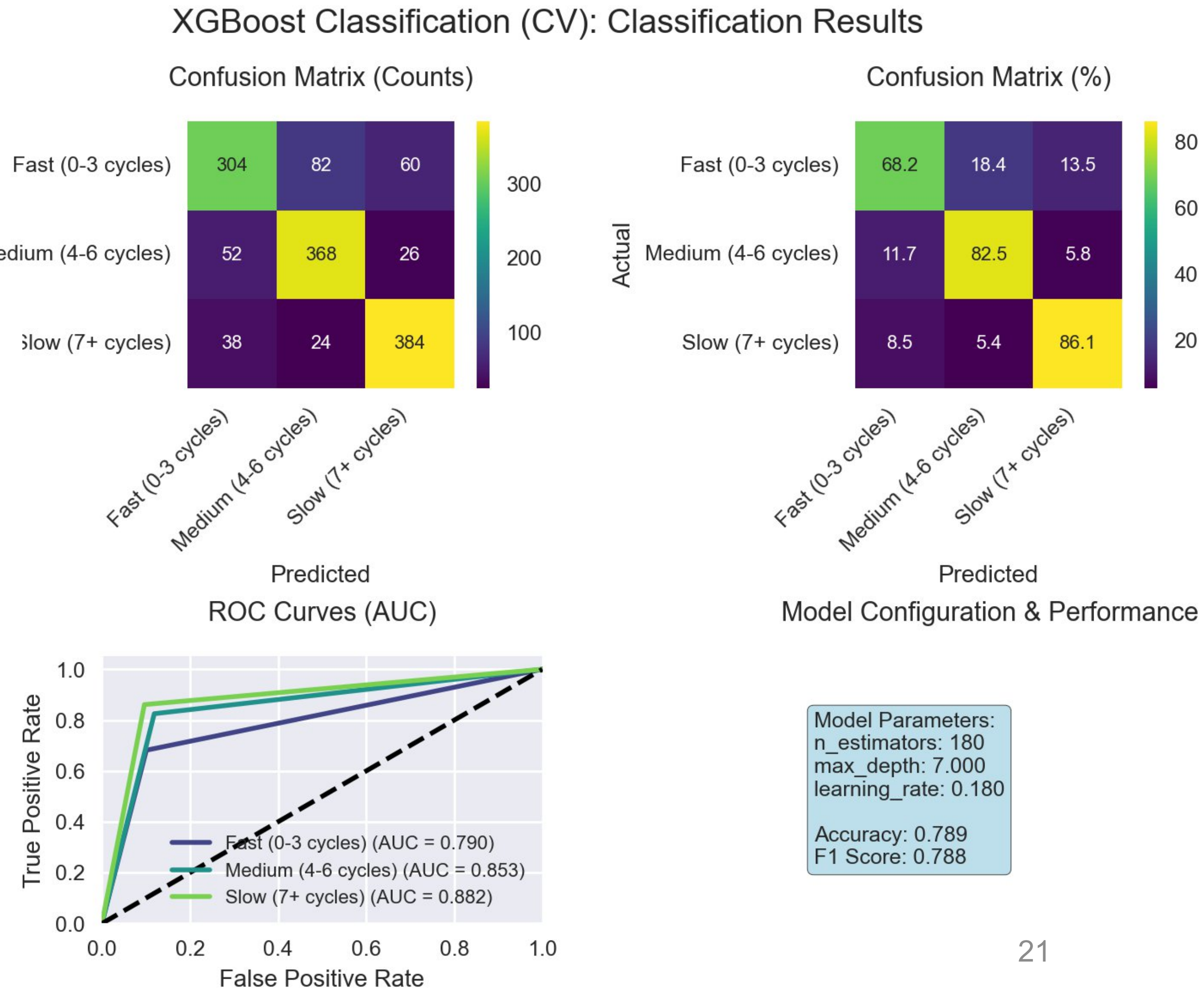


Majority of conceptions within 3 cycles...

Solution: bin, balance and classify!

1. Group conception times into 3 bins
2. Rebalance classes
3. Classification now gives **~79% accuracy**

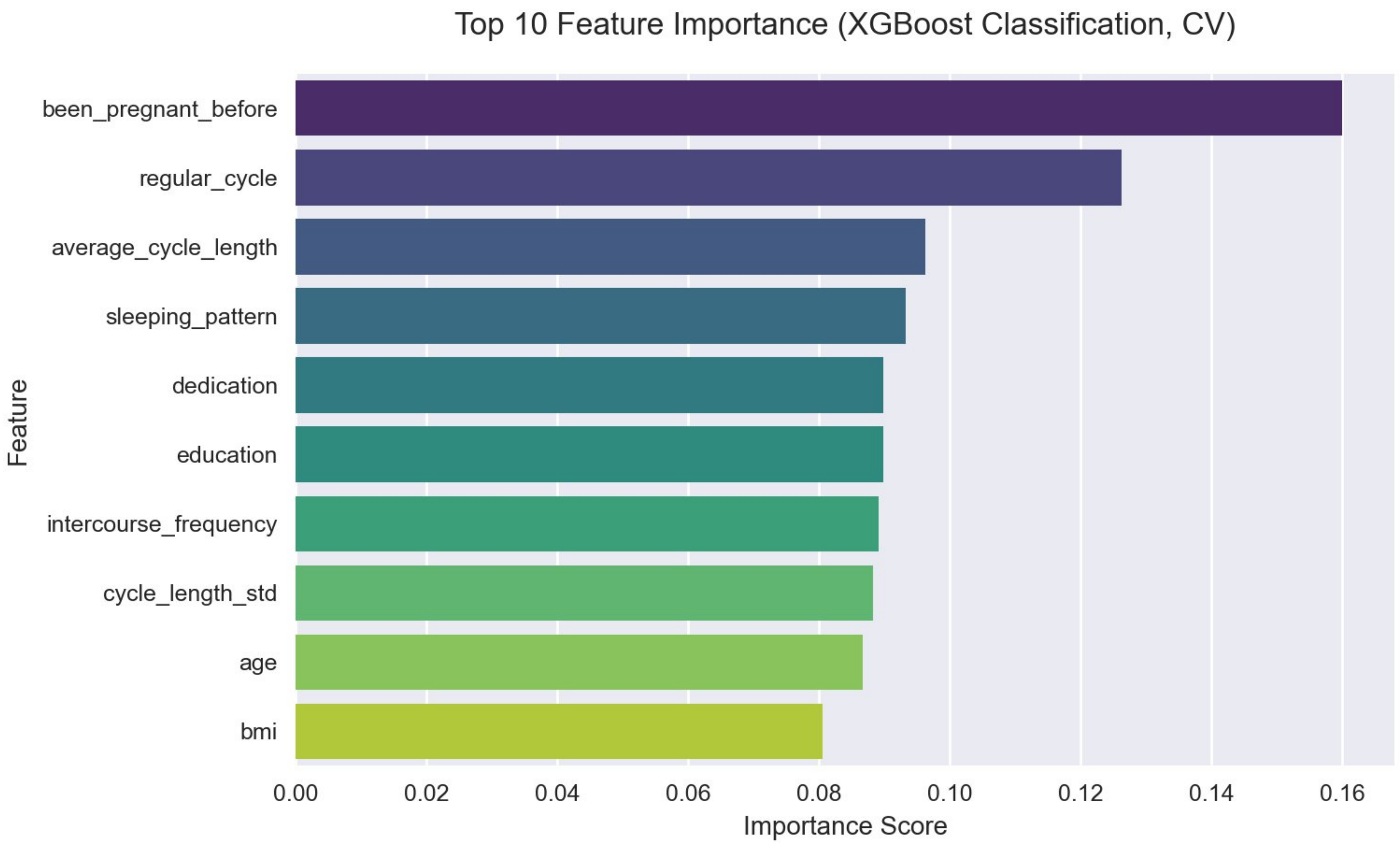




Model Parameters:
n_estimators: 180
max_depth: 7.000
learning_rate: 0.180

Accuracy: 0.789
F1 Score: 0.788

Feature importance shows new effects affecting conception time



Conclusion

- ML techniques require (more) tweaking to get working
  - But can reveal more influencing factors

- Factors affecting conception time (in order of importance):
  1. Nr of previous pregnancies
  2. Cycle regularity and average length
  3. Sleeping patterns
  4. Dedication to logging data on Natural Cycles
  5. Intercourse frequency
  6. Physical condition: Age + BMI

- (Further investigation needed **HOW** these affect conception time)