Capstone Project 1 - Final Report

Berenice Dethier
Mentor: Dhiraj Khanna
Springboard cohort of October 2019

Contents

Congratulations, you are expecting a baby!

It's exciting and scary, and you are trying to be as prepared as possible for all the changes in your family. Your doctor gave you an estimation for the due date, but you know that women deliver babies anywhere between 25 and 42 weeks of gestation[1]. That's a wide window for parents-to-be, and you probably want to know more accurately and precisely when your baby will come. In addition, premature babies have higher risks of having lungs or heart problems, as well as other conditions. Is your pregnancy at risk for preterm delivery?

As it turns out, the tools to estimate a delivery date are limited. Researchers have been trying to improve the estimate, some by measuring certain biomarkers[2]. Instead, can we improve the estimate using data?

1 Context

The advancement of a pregnancy is inferred by two techniques. The first technique measures the size of the fetus by ultrasound and estimates how far along the pregnancy is. This is called the obstetric estimate of gestation. The second technique estimates the time of conception from the date of the last menstrual period (LMP). In both cases, the due date is established based on a normal pregnancy length of 40 weeks[3].

There are two main sources of difference between the **actual week of birth** and the **estimated due date**:

- The advancement of pregnancy might have been estimated poorly.
  - By ultrasound: not all fetuses grow at the same pace and the measurement itself includes error (associated with the technique).
  - Last menstrual period start date: menstrual cycles can vary in length, both between women and over one person's life.
- Pregnancy is a physiological process, therefore its length can vary slightly.

The CDC publishes yearly datasets on birth and infancy health called "Natality Public Use files". They include demographics on the mother, difference between obstetric due date and actual due date (which would be the predicted variable of the study) as well as newborn health and other outcomes of the pregnancy. The US dataset for 2018 contains almost 4 million entries. There are over 75 recorded variables, but some can be immediately eliminated.

[1] Babies can be born before 25 weeks, but they often have health conditions as a consequence of the early birth. Doctors usually induce labor past 42 weeks for health reasons

[2] For example in this PLoS One study.

[3] Normal gestation is 38 weeks from **conception** to delivery. However, because the conception date is usually unknown, the date of the **last menstrual period** is employed as the initial time of pregnancy. Ovulation occurs on average two weeks after that, therefore the normal pregnancy length is 38 + 2 = 40 weeks. Health professionals and the general public use the "40 weeks from last period" system, not the "38 weeks from conception" one.

After gathering and cleaning the data (section 3), we will filter it to keep spontaneous births only because inductions and scheduled c-sections do not reflect the natural delivery date. During the exploratory data analysis (section 4), we drew interesting insights on what impacts the length of a pregnancy. We created plots to tell the story of delivery dates and what impacts them.

The next step will be to split the dataset into a training and test set, then train a model on the training set. A linear regression would be the first step, but other models can be used. Alternatively, we could use a logistic regression to predict if a mother has higher risks of delivering prematurely. Once a model has been selected, we can assess its accuracy using the test set. If the model is suitable, a prediction tool could be created for future parents in the form of a web app. We would need to provide ample resources for interpreting the results.

Some of the terminology needs to be specified before we dig in. The **date of birth** is not available in the file for the sake of anonymity. The duration of pregnancy in weeks is provided. This will be referred to as the **delivery week**. Following the recommendation of the CDC, we will use the delivery week calculated from the obstetric estimate of gestation (estimation by ultrasound), and not the estimate from the last menstrual period.

We chose to use the delivery week, not the difference between the actual and the estimate delivery week, to avoid confusion related to negative numbers. When looking at delivery weeks, the reader should keep in mind the normal duration of 40 weeks used by health professionals. For example, a woman who delivered a baby at 37 weeks delivered 3 weeks ahead of her due date. **Full term** is used to qualify births at 37 weeks of gestation or more. **Preterm births or premature births** are births occurring at 36 weeks of gestation and under. A more detailed classification used in obstetric is as follows:

- Late preterm, born between 34 and 36 completed weeks of pregnancy
- Moderately preterm, born between 32 and 34 weeks of pregnancy
- Very preterm, born at less than 32 weeks of pregnancy
- Extremely preterm, born at or before 25 weeks of pregnancy

2 In short

2.1 Method

Spontaneous births were the only observations considered (about 1.9 million records). Scheduled c-sections and inductions were filtered out (one half of 3.8 million births in 2018).

Data collection and cleaning was performed with R Studio. Exploratory data analysis was done in two parts: visual EDA with Tableau and statistical analysis in Python Jupyter notebook. A full report of the work, code and details are available on github. Interactive visualizations are available on Tableau Public. Machine learning was performed with algorithms from scikit learn and Catboost. We approached the question as a regression (predicting the week of birth) and as a classification (predicting the term of birth).

2.2 Data Analysis

The differences between premature and full term babies presented in this document have been confirmed to be of statistical significance.

2.2.1 Biological factors

The factors are described Figure 1.

There was an optimal range for most physical features in which women would have a lower risk of having a premature baby.

The race of the parents had a significant impact: Asian and White parents have the longest pregnancies.

Female newborns have lower prematurity rates, as well as first born children.

Previous premature births, and use of infertility treatments were correlated with more premature births.

Unsurprisingly, expecting multiple babies, risk factors, infections, and tobacco also increase the prematurity rates.

Only one biological factor was not significant in predicting the delivery week: the age of the father.

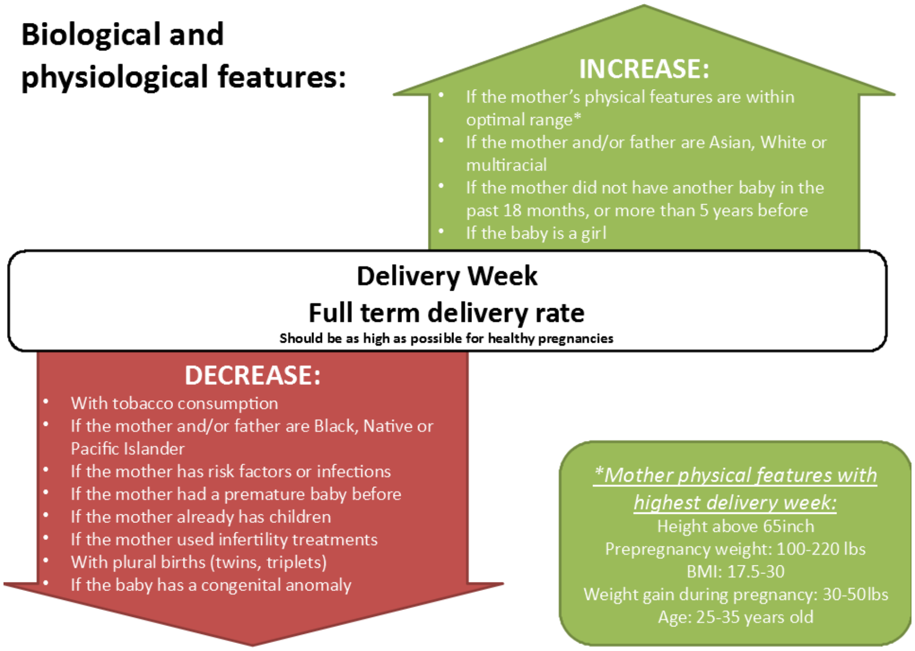

Biological and physiological features:
INCREASE:
If the mother's physical features are within optimal range*
If the mother and/or father are Asian, White or multiracial
If the mother did not have another baby in the past 18 months, or more than 5 years before
If the baby is a girl
Delivery Week
Full term delivery rate
Should be as high as possible for healthy pregnancies
DECREASE:
With tobacco consumption
If the mother and/or father are Black, Native or Pacific Islander
If the mother has risk factors or infections
If the mother had a premature baby before
If the mother already has children
If the mother used infertility treatments
With plural births (twins, triplets)
If the baby has a congenital anomaly
*Mother physical features with highest delivery week:
Height above 65inch
Prepregnancy weight: 100-220 lbs
BMI: 17.5-30
Weight gain during pregnancy: 30-50lbs
Age: 25-35 years old

Figure 1: Summary of how the biological and physical variables affect the average delivery week and rate of full term pregnancies in the population.

2.2.2 Socioeconomic and other factors

These factors are summarized Figure 2.

Factors associated with higher economic status are correlated with fewer premature births: education of the parents, health insurance, not on WIC.

One interesting correlation was the marital status of the mother: a single mother has almost twice as much risk to deliver prematurely than a married mother, even when controlling for age or race.

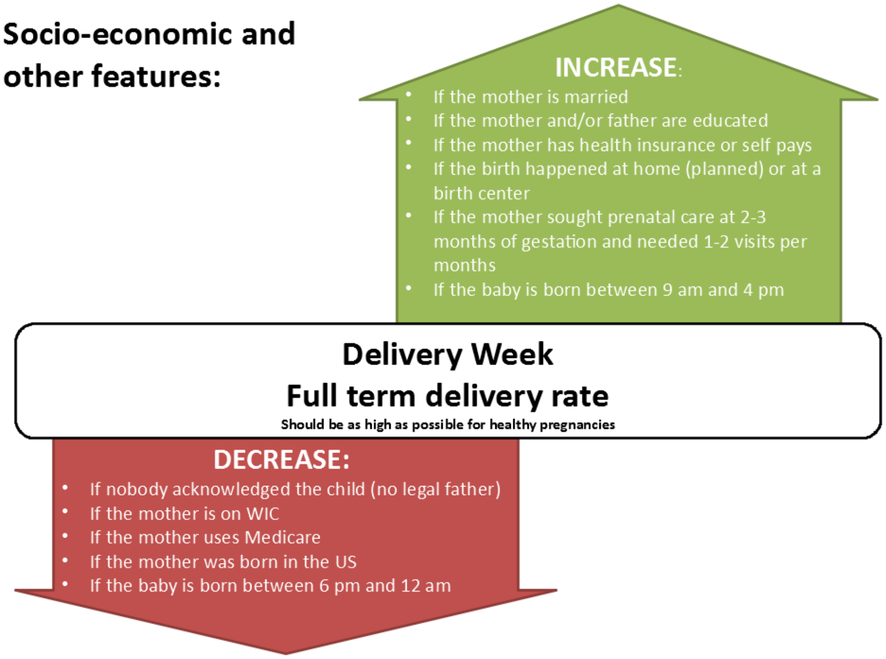

Socio-economic and other features:
INCREASE:
If the mother is married
If the mother and/or father are educated
If the mother has health insurance or self pays
If the birth happened at home (planned) or at a birth center
If the mother sought prenatal care at 2-3 months of gestation and needed 1-2 visits per months
If the baby is born between 9 am and 4 pm
Delivery Week
Full term delivery rate
Should be as high as possible for healthy pregnancies
DECREASE:
If nobody acknowledged the child (no legal father)
If the mother is on WIC
If the mother uses Medicare
If the mother was born in the US
If the baby is born between 6 pm and 12 am

Figure 2: Summary of how the socio-economic and other variables affect the average delivery week and rate of full term pregnancies in the population.

The place of birth of the mother (in the US or not) also impacts the prematurity rates.

Surprisingly, the month of birth and time of birth are correlated with the rate of premature births. Babies born between 6 pm and 12 am are more likely to be premature.

The outcomes for the baby and mother were analyzed. The birth weight, maternal morbidity and breastfeeding rates were on average lower for premature babies. The rate of abnormal conditions on the other hand was higher for premature babies.

2.3 Machine learning

We were able to predict the week of birth with a preicison of 96.9% and an average error of 1.13 weeks. This is unfortunately not very good, since over 70% of the observations consisted of births at 38-40 weeks. When we tried to predict the term of birth, we reached an accuracy of 92.4%, which is disappointing as well, given the class imbalance (8.3% of premature births).

We suggest using a neural network to obtain a better model.

3 Data collection and wrangling

In agreement with my mentor, the data wrangling portion of the project was handled in RStudio©.

3.1 Data collection

The datasets for the past few years are available on the CDC website. We focus on the US 2018 public file and related documentation, which are available for download. The file is a fixed width format file: the dataset comprises 3,801,534 records, each saved as a string of 1330 characters. Variable names (column names) are matched to the values by using the documentation: each (group of) character(s) is extracted and saved in the corresponding column. Due to the file size (5+ Go), we start by looking at the first 100,000 records and extract only columns that are thought to impact the dependent variable. Variables such as the baby APGAR score or characteristics of labor are not imported to the dataframe.

3.2 Data cleaning

The dataset is mostly clean. Because all variables are coded with numbers/letters, there are no visible typos nor string manipulations needed.

3.3 Filtering relevant data

One important step was to filter the data to keep only the spontaneous births (Figure 3 and 4). Inductions are not considered spontaneous. They made up 27.1% of the observations. C-sections can be scheduled, therefore not spontaneous, or the final route of delivery after attempted labor (unplanned or emergency c-sections). We kept only the c-sections for which labor had been attempted, and we called them "unscheduled c-sections". Roughly 22.2% of the dataset consisted in scheduled c-sections. The dataset was reduced from 3,801,534 observations to 1,921,127 (50.7% left).

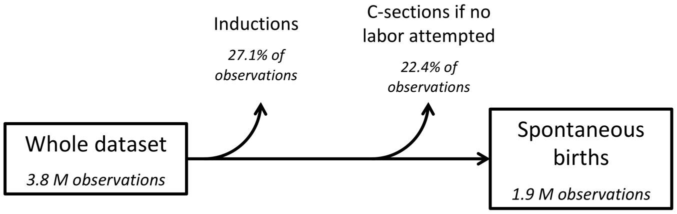


Figure 3: Filtering out non-spontaneous births

It would be interesting to investigate the demographics of these two populations (births following an induction and scheduled c-sections) to understand why they are so substantial, possibly as a classification problem. This study is out of the scope of this project.

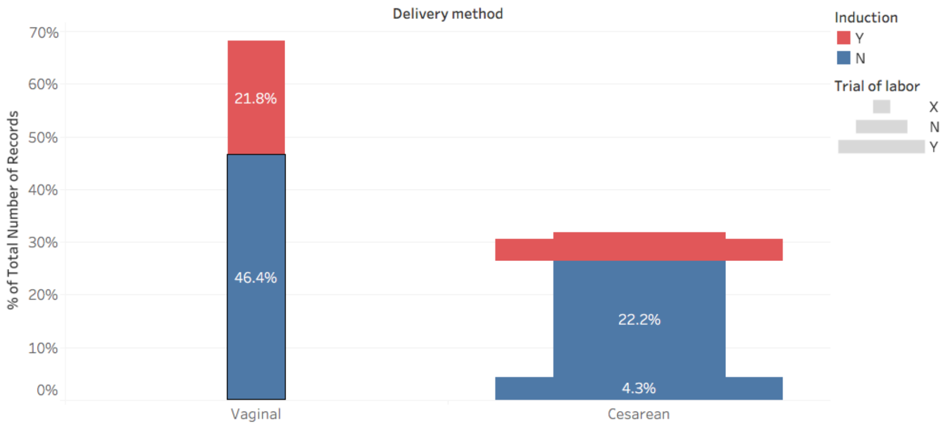


Figure 4: Overview of the observations in the 2018 Natality dataset. We keep the vaginal, non-induced births (left bar, blue) and non-induced c-sections where labor has been attempted (right bar, wide, blue).

3.4 Missing values

In the CDC dataset, missing values were coded with 9s. The breakdown of missing values computed with the function `plot_missing()` of R `DataExplorer` package is presented Figure 5.

The dataset documentation comments on missing information regarding the father: missing age, race, hispanic origin and education of the father often occurs when the mother is a single mother. This seems to be important information that should not be eliminated by imputing the data. All missing values (9-coded) were replaced by ‘NA’, except the categorical variables regarding the father (father race, father hispanic origin and father education), where missing values were kept as value 9. Father age is left as is.

3.5 Outliers

Outliers were identified with boxplots on the non-imputed values of the dataset. Numeric variables and how they were handled are described below. This section also provides an overview of the numeric variables in the dataset.

The biological/physiological variables, **age of the mother, height of the mother, prepregnancy** and **delivery weight**, follow an expected distribution, as well as **birth weight**. The **BMI** 1st quartile of 21.7, median of 24.7 and 3rd quartile of 29.2 is lower than the national distribution of 23.8, 28.3 and 33.7, respectively [4]. This makes sense given the age range and health level required for pregnancy.

Total birth order ranges between one (the current birth is the first child) to 8, with a median at 2 and most mothers having between one and three children.

The **number of previous c-sections** shows unlikely extreme values of 4 and above. Our dataset is featuring only attempted vaginal births, yet it is unusual to have a vaginal birth after a c-section

[4] Source of BMI data: DQYDJ

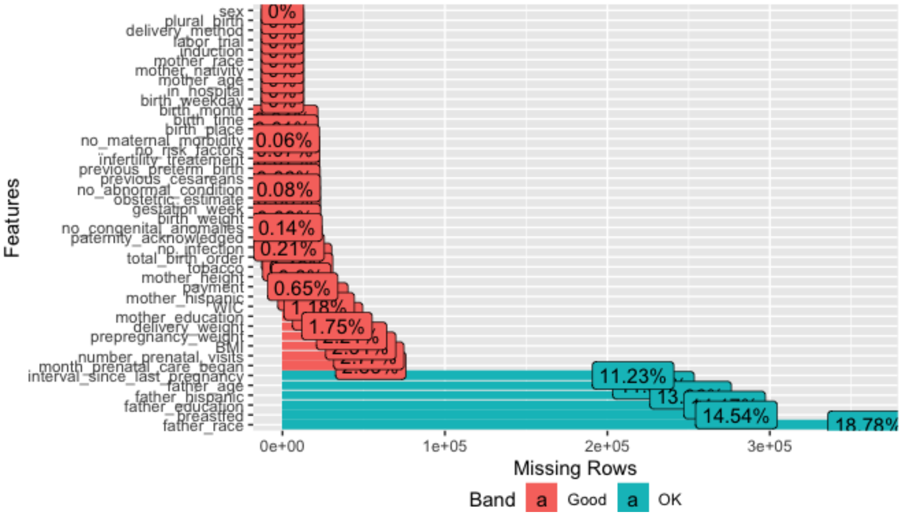


Figure 5: Overview of the missing values in the dataset. According to the documentation, missing values regarding the father are often related to the absence of father. This constitutes information, therefore these variables will not be imputed.

(VBAC), and it is even less recommended after multiple c-sections. Although not plausible, the consequences of an error in this field and its proportions don't justify replacing the values.

Prenatal care began usually between the first and third month of pregnancy, with a few outliers past 5 and up to 10 months or before the first month (value 0). This was calculated based on the obstetric estimated due date. As discussed above (section 1), the estimate could be off, therefore 0 and 10 are not to be excluded.

The **number of prenatal visits** has a few extremes values, which is also plausible: high risk pregnancies could have up to 89 visits (more than two visits a week). **Interval since the last pregnancy** has two types of NA values: truly not applicable when it's the first pregnancy (coded 888) or a missing value (coded 999). We might have to group the values in bins and treat the variable as an ordered discrete variable for machine learning, but information would be lost.

Delivery week calculated from the **LMP estimate** and **obstetric estimate of gestation** seem to have incorrect extreme values on the upper side, between 43 and 46 weeks (considering the fact that a normal pregnancy is 40 weeks and there are risks for the baby past 42 weeks). This likely comes from the fact that they are both estimates and could be erroneous. No changes will be made to these variables since they are the dependent variables.

4 Exploratory Data Analysis

In agreement with my mentor, visual EDA was performed in Tableau, and statistical Data Analysis with Python (packages numpy, scipy.stats, pandas, pymc3, matplotlib.pyplot, seaborn and statsmodels).

The focus of this study is the delivery week for pregnancies in the US in 2018. The variable itself is an estimate based on an ultrasound measurement. The expected delivery week is 40 (a normal pregnancy is 40 weeks long). A baby born at 38 weeks is therefore 2 weeks early. Although a normal pregnancy is 40 weeks, the median for this dataset is 39 weeks (Figure 6).

Interpretations of the trends have not been evaluated scientifically and are provided for informational purposes only.

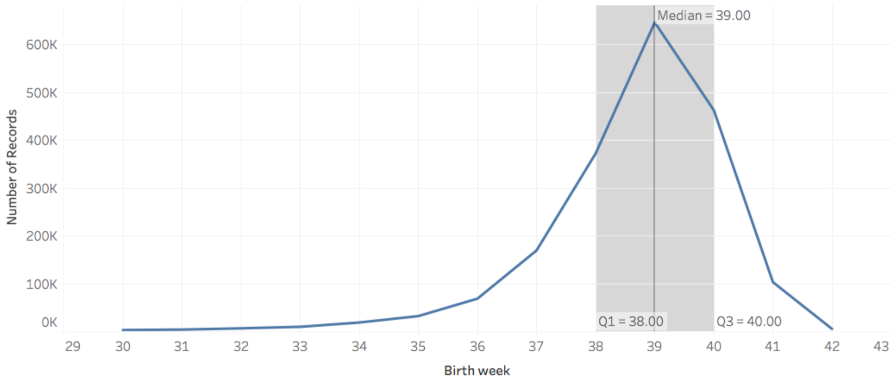


Figure 6: Number of births per delivery week. Even though a normal gestation is thought to be 40 weeks long, the median and mode for this dataset is 39 weeks.

For all variables, we check the impact on the delivery week and the term of the pregnancy: if the baby was born at 36 weeks of pregnancy or less, she is premature/preterm, if she was born at 37 weeks or later she is full term. Statistical Data Analysis was performed by ANOVA unless stated otherwise. For continuous variables, the term of birth was used to divide into two groups. For categorical variables, we looked at the delivery week (continuous variable) in each category. The proportion of premature births in the dataset is 8.3%, and the average delivery week is 38.55.

4.1 Biological and physiological factors

The weight of the baby at birth is a consequence of the length of the pregnancy, so this factor will not be explored.

4.1.1 Height

The mother height appears to be normally distributed (bell curve, Figure 7, top). The average delivery week tends to increase with the mother height, but the spread is small: the difference between the highest and lowest average is 0.5 week (Figure 7, bottom).

Statistical exploration (ANOVA) indicates a difference in mother height between preterm and full term babies. The mean mother height for premature babies is 63.9 inch and for full-term babies 64.1. The rate of premature babies decreases with the mother height, and is lower than the general average for mothers 65 inch and above (Figure 8).

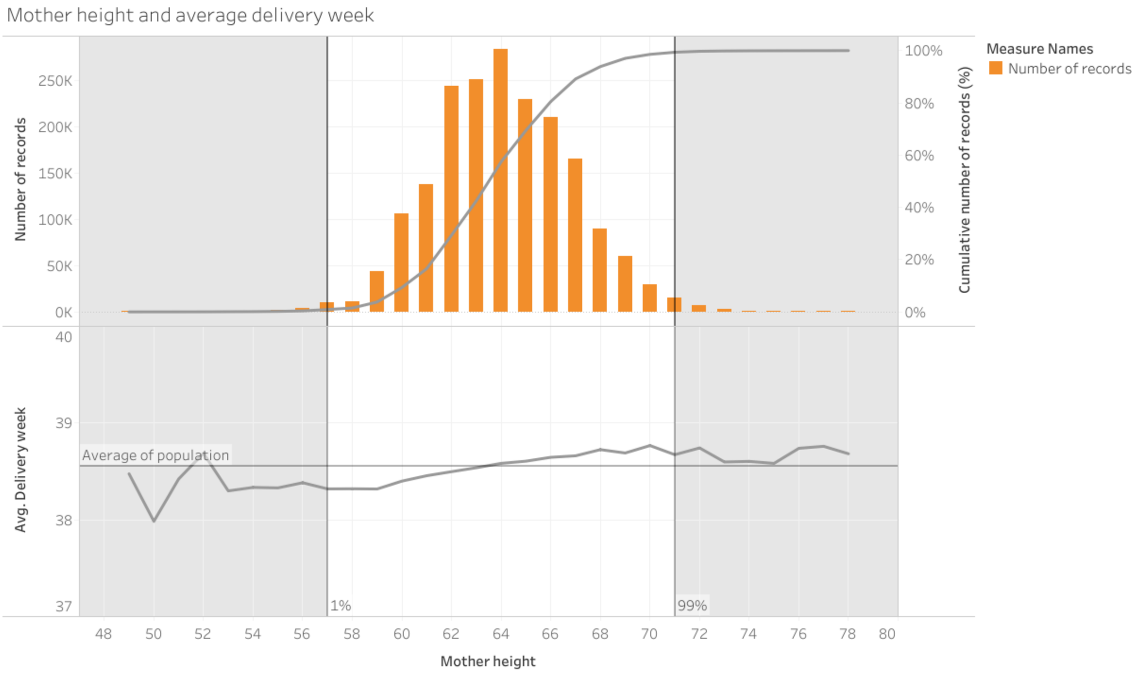


Figure 7: The average delivery week tends to increase with the mother height, but the spread is small: the difference between the highest and lowest average is 0.5 week.

4.1.2 Prepregnancy weight

The distribution of prepregnancy wieghts is skewed to the right and will probably be transformed (logarithmically) for modeling. The prepregnancy weight is a significant predictor of the term of birth. The average mother weight for premature and full term deliveries are 155.7 and 152.2 lbs, respectively. Furthermore, the rate of premature babies for mothers between 100 and 169 lbs is lower than the general average (Figure 9).

4.1.3 Weight gain

The weight gain is normally distributed. Average delivery week increases with the weight gain. A weight gain of 20 lbs and above is correlated with a longer pregnancy, on average (see appendix).

There is an optimum for full term deliveries at a weight gain of 30-50 lbs (more than 92% of babies born full term, Figure 10). The ANOVA also indicated a difference in average weight gain for the two groups (at 24.7 lbs and 29.1 lbs for premature and full term pregnancies, respectively). In addition, there is a weight gain range between 30 and 49 pounds where the average risk for preterm birth is lower than the global average.

4.1.4 BMI

The BMI combines height and pre-pregnancy weight. BMI distribution is skewed like the prepregnancy weight.

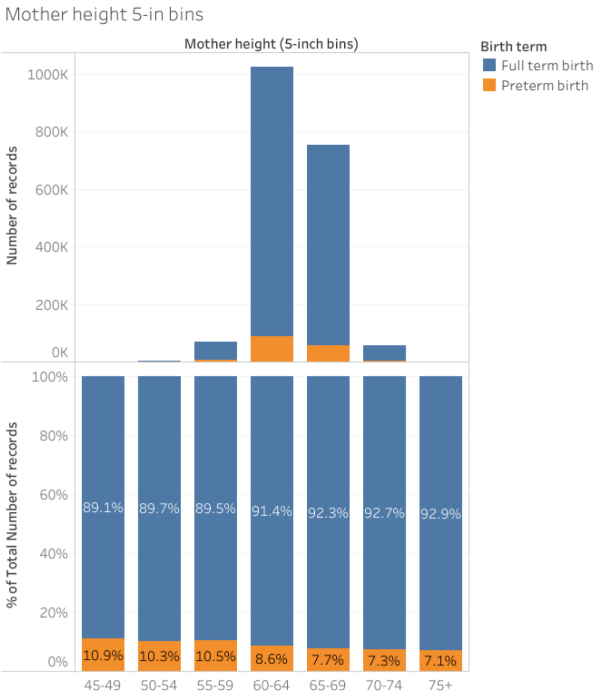


Figure 8: The rate of premature babies for mothers 65 inch and above is lower than the general average.

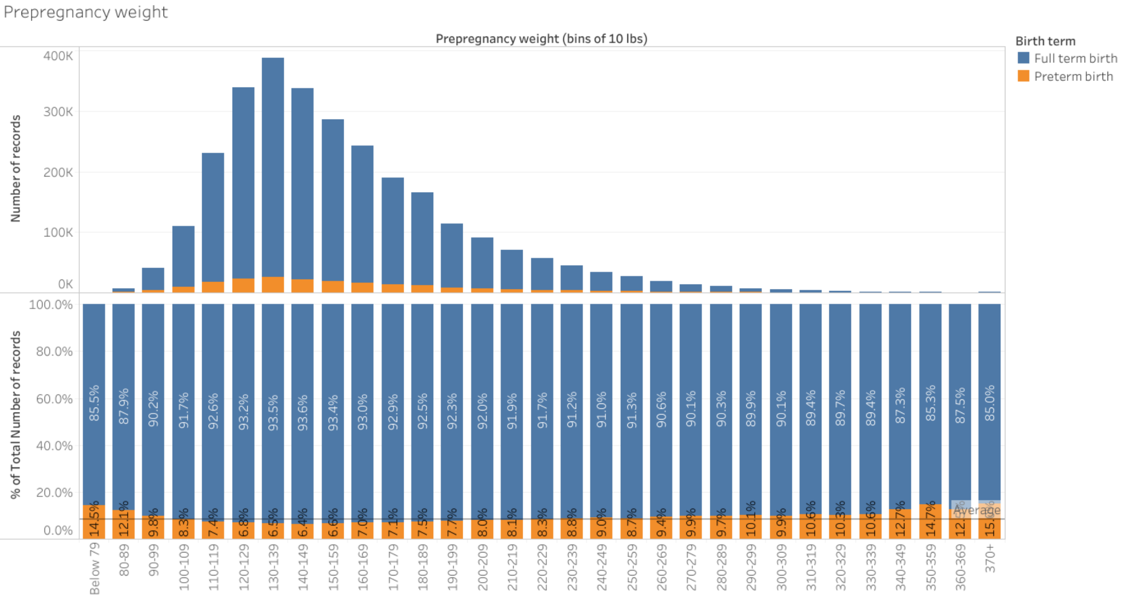


Figure 9: There is an optimum between 100 and 220 lbs for which there are less premature births.

In addition, BMIs between 17.5 and 30 were associated with less premature births on average than the rest of the population (Figure 11). ANOVA confirms that we reject the equality of means

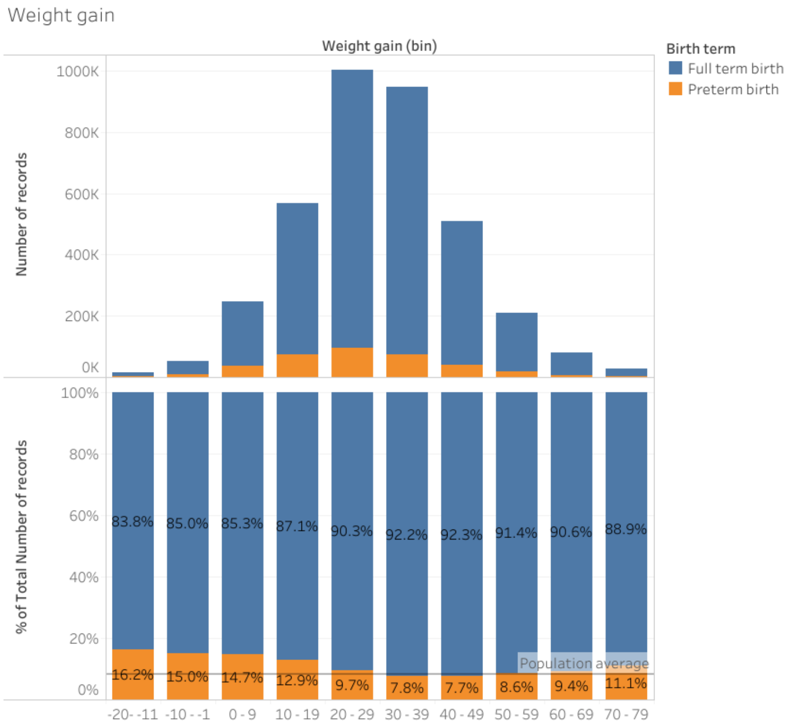


Figure 10: Distribution of preterm and full-term birth per weight gain. There seems to be a weight gain range between 30 and 49 pounds where the risk for preterm birth is lower than the global average of premature births (8.3%).

between BMIs of mothers having preterm and full term deliveries (average BMI of 26.0 and 26.8, respectively).

4.1.5 Age of the mother

We tested the age of the mother in a similar fashion. The plot of the average delivery week per age indicates that the average length of pregnancy reaches a maximum at 32 years old, and mothers between 17 and 39 years of age deliver later on average than the global average (Figure 12).

Based on Figure 13, there is an optimum with less premature births around 25-34 years old. The ANOVA confirm that the two populations (premature and full term babies) have mothers of different ages: 28.4 and 28.5, respectively.

It is known that older mothers have more complications during pregnancy, which can lead to early delivery. Observations on younger mothers might include pregnancies that were unplanned/unexpected, and in this context mothers could deliver early due to stress or lack of proper care.

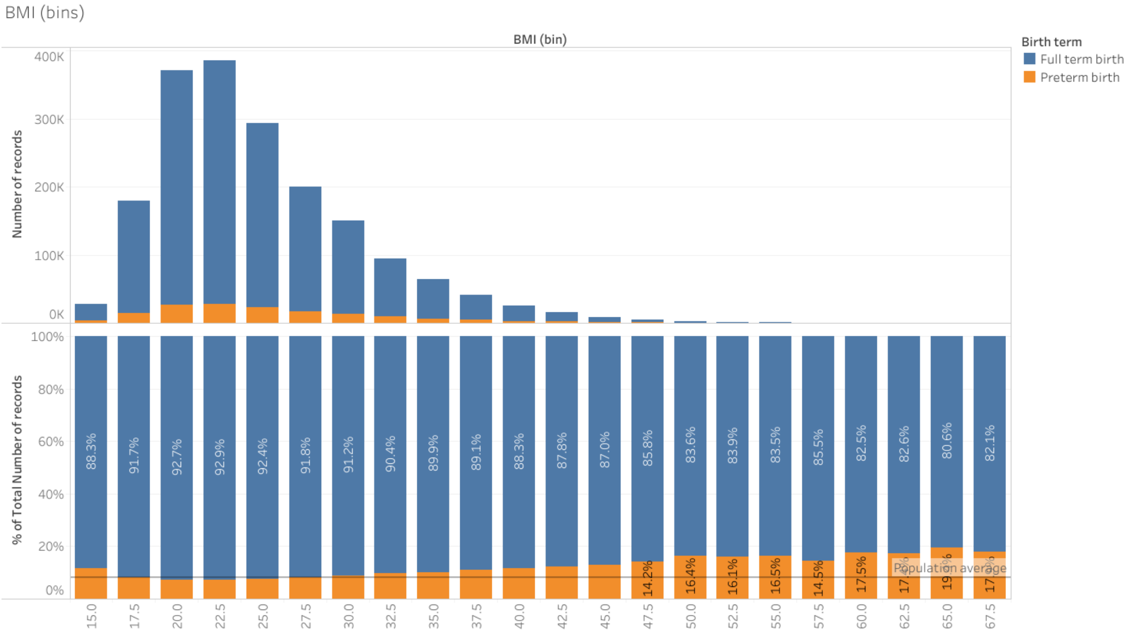


Figure 11: Mother BMI (bins of 2.5) with the corresponding prematurity rate. There are less premature births between BMI 17.5 and 30.

4.1.6 Tobacco consumption

Smoking during pregnancy increases the risk of health problems for developing babies, including preterm birth, low birth weight, and birth defects of the mouth and lip[5]. Without surprises, the ANOVA indicates that the smoker and non-smoker populations are not equal, with the mean delivery week equal to 38.0 for smokers and 38.6 for non-smokers. The difference in rate of premature births between smoking and non-smoking mothers is quite large, with smoking mothers having 14.7% of premature babies and non-smoking mothers having only 7.8% of them (Figure 14, left).

4.1.7 Mother race

Mean delivery weeks range between 38.6 (White mothers) and 38.2 (Black mothers). Races listed in addition to these two are AIAN (American Indian and Alaskan Native, 38.4 weeks on average), Asian (38.6 weeks on average), NHOPI (Native Hawaiian and other Pacific Islander, 38.5 weeks on average), and more than one race (38.5 weeks on average). Figure 15 shows that, proportionally, more Black mothers and Native mothers deliver early compared to the other races. The race is a significant factor in predicting the delivery week based on the ANOVA. There could be biological factors at play, but also socio-economic factors. Reports indicate that Blacks and AIANs are proportionally poorer, less educated, and have poorer health [6].

The rates of premature births follow the same trend (Figure 14, right): Asian women have the lowest premature rate at 6.8%, and Black mothers have the highest at 11.6%.

[5] See CDC website.

[6] See minority health's first and second report, featuring data from the Census Bureau

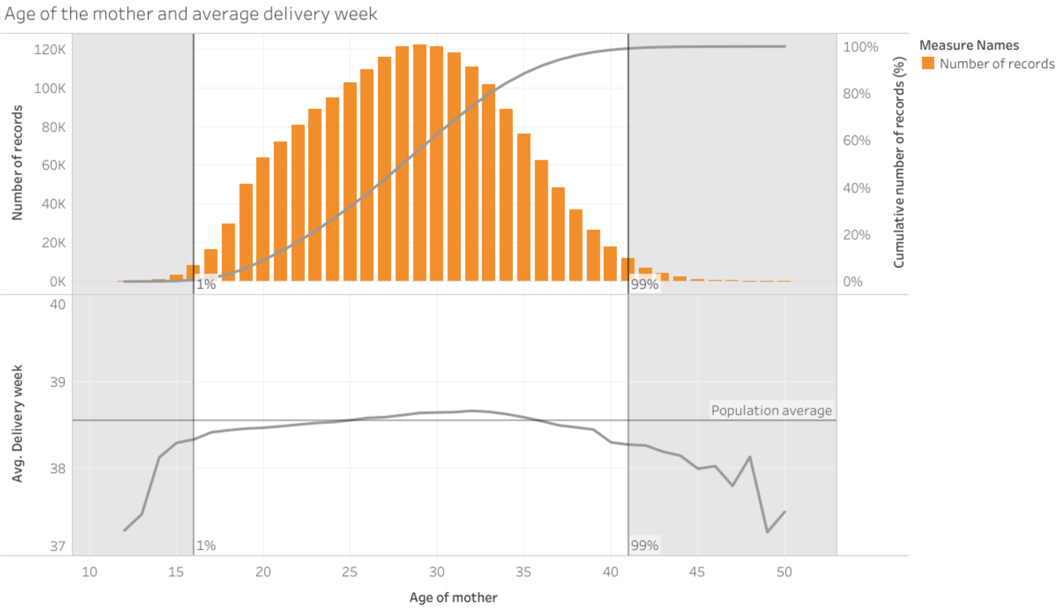


Figure 12: The average length of pregnancy reaches a maximum at 32 years old, and pregnancies of mothers between 17 and 39 years of age are on average longer than the global average.

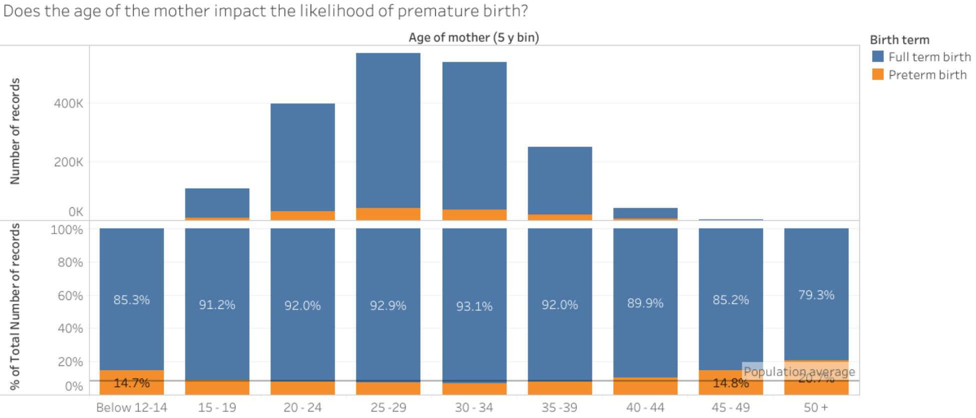


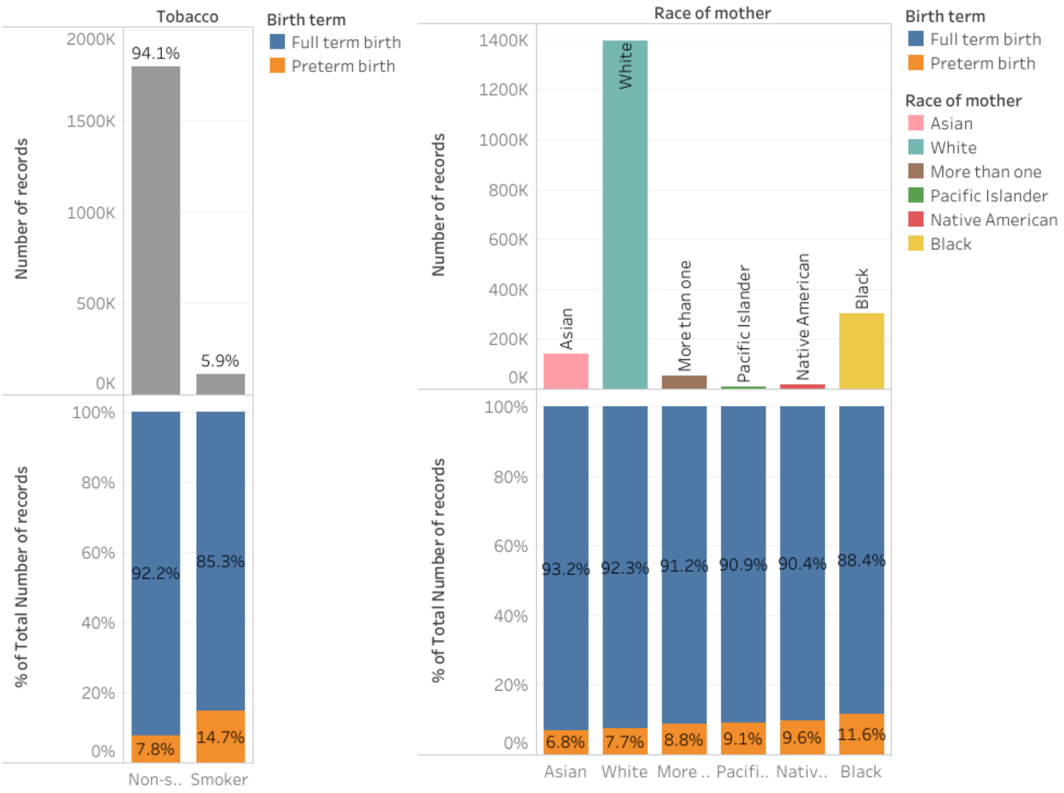
Figure 13: There are more preterm deliveries among the younger (below 20) and older mothers (over 40) compared to the global average of 8.3%. The optimum is between 25 and 34 years old.

4.1.8 Mother hispanic origin

This variable is significant based on the ANOVA, but visual EDA did not suggest a strong trend (See link in appendix).

4.1.9 Risk factors and infections

The mean delivery weeks (Table 1) are different with or without risk factors, and with or without infection(s). This is expected, risk factors and infections lead to earlier births, likely because of



(a)

(b)

Figure 14: (a) 14.7% of smokers deliver prematurely, a sharp contrast with 7.8% of non-smokers. (b) The proportion of premature babies varies by mother races. The top portion of the graph indicates the distribution of mother races in the 2018 Natality file. The bottom part shows the proportion of premature within each category. Races have been ordered from the lowest premature rate (Asian, left) to the highest (Black, right). Black mothers have on average 60% more risks of having a premature baby than Asian mothers.

complications in the pregnancy. The proportions of premature births follow the same trend in the two groups (Figures 16).

4.1.10 Previous preterm birth

The distribution (ECDF Figure 17, left) and mean delivery week are different with or without a previous preterm birth (at 37.1 and 38.6 weeks, respectively). The premature rates are consistent with this at 27.9% and 7.6%, respectively (Figure 17, right). Previous preterm births could indicate an underlying condition or a biological predisposition for preterm births.

4.1.11 Number of previous children (birth order)

About a third of the dataset consists of first time mothers. The more children one already has, the earlier they are born (Figure 18). The average delivery week goes from 38.7 (first born) to 38.2

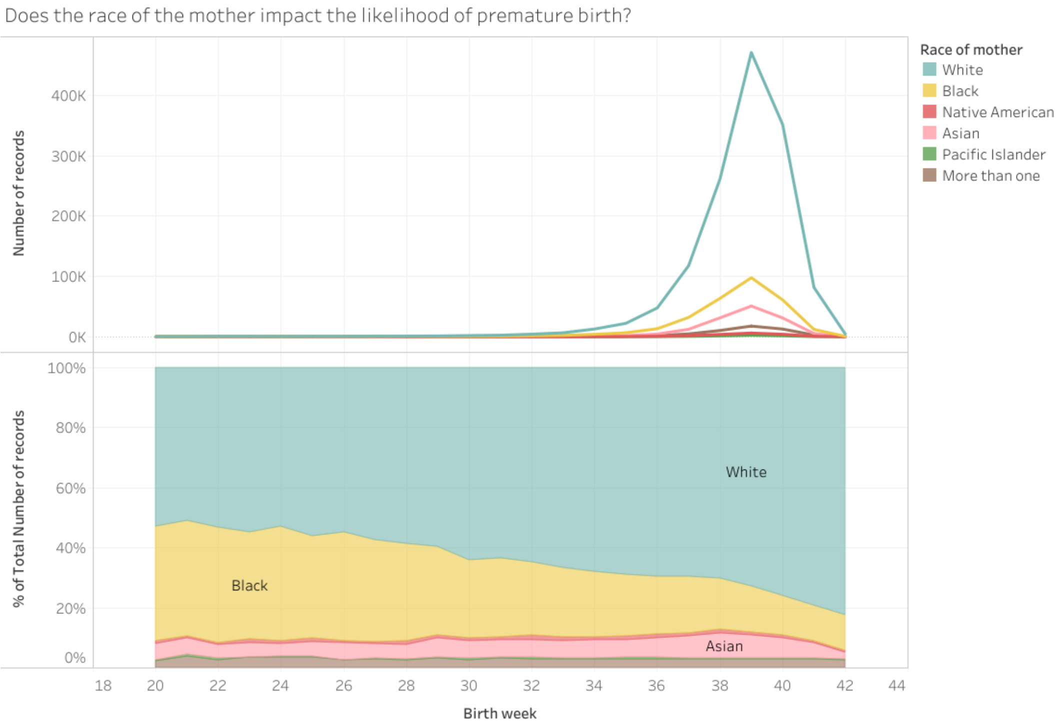


Figure 15: Distribution of births per delivery week by mother race. The top chart indicates the number of births, and the bottom chart the proportion (% per race at each delivery week). Overall, Black mothers deliver their babies earlier than White and Asian mothers: 38% of the mothers who delivered at 20 weeks were Black, 6% were Asian and 53% were White, but at 40 weeks, only 13% were Black, 7% were Asian and 76% were White. In absolute numbers (top chart), this effect is hard to detect due to the lower occurrences of premature births.

(eight or more previous children) in an almost linear fashion. In addition, they are more likely to be premature have they have more siblings (Figure 19).

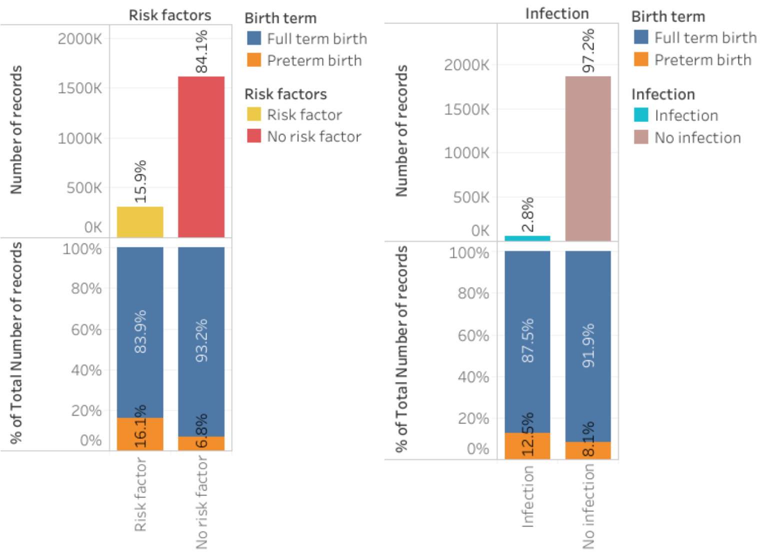
4.1.12 Interval since last pregnancy

The distribution of intervals since the last pregnancy is skewed to the right. A large percentage of mothers have another child one to two years after the previous one.

Very short or long intervals between pregnancies seem to be linked with more premature births (Figure 20). The optimal interval between pregnancies (interval that minimizes the risk of premature delivery) is between 18 and 35 months. For larger intervals, an increasing number of premature births is recorded. They could be related to the age of the mother: a woman having a child 6+ years after the previous one is likely older. For these plots, first time mothers are not taken into account, which increases the average premature rate to 8.6%. This variable has a significant impact in predicting the delivery week (based on the ANOVA).

Mean Delivery Week		
Variable	Yes	No
Risk factor	37.9	38.7
Infection	38.2	38.6

Table 1: Delivery week is lower on average with risk factor(s) or infection(s)



(a) (b)

Figure 16: The proportion of preterm births is higher with risk factors or infection.

4.1.13 Sex of the baby

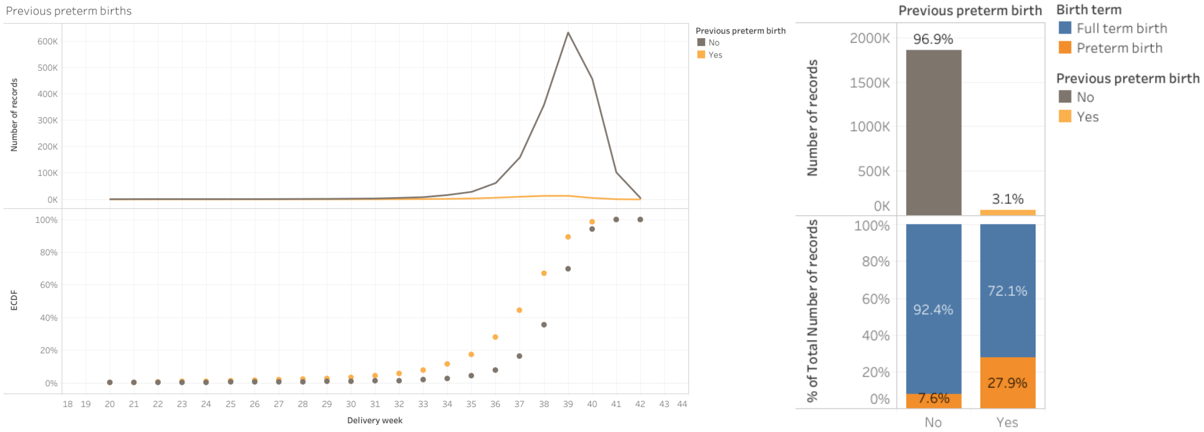
There are more premature boys than girls (Figure 21, left). The ANOVA confirms that the two populations, boys and girls, don't have an equal week of birth, with girls being born on average at 38.6 weeks and boys at 38.5 weeks.

4.1.14 Infertility treatment

Mothers who used infertility treatments tend to deliver earlier, with a mean delivery week of 37.8 (38.6 for parents not using infertility treatment). ANOVA on delivery week with and without infertility treatment confirms that the populations are not equal. There are more premature deliveries among the women who used assistance to conceive, with a premature rate of 17.2% (as opposed to 8.2% without treatment, Figure 21, right). This effect could be due to underlying conditions unfavorable for childbearing, or the stress related to treatments.

4.1.15 Plural birth

As expected, the delivery week goes down with the number of plural babies, and the rates of premature births go up (Figure 22, note that the sample size goes down and only 15 records of



(a) (b)

Figure 17: Mothers who delivered prematurely in previous pregnancies tend to give birth earlier than mothers who did not: the yellow dot (previous preterm birth) are always higher than the brown points (no previous preterm birth), meaning that a higher proportion of the previous preterm birth population gave birth at every given delivery week.

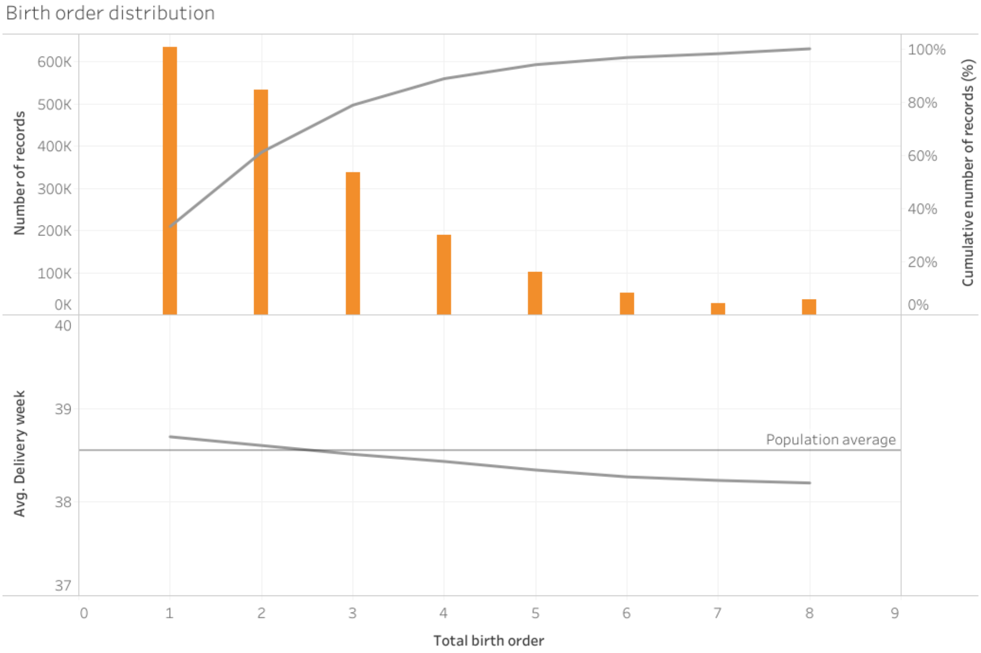


Figure 18: About one third of the observations are first-time mothers (top chart). The average delivery week decreases with each additional child a mother previously has (bottom chart).

the dataset are quadruplets). The mean delivery week is 38.62, 34.34, 28.29 and 30.13 for parents having singleton, twins, triplets, and quadruplets, respectively. Plural birth is a significant factor based on the ANOVA results. It is known that mothers expecting multiples go into labor earlier,

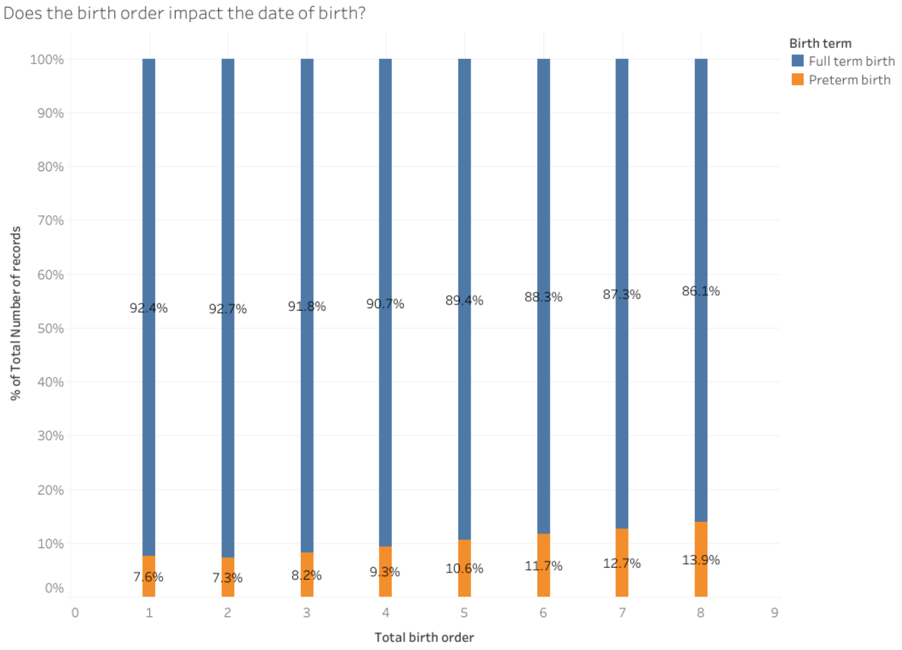


Figure 19: The proportion of preterm births increases with the number of children already born. There are more preterm deliveries for the second and subsequent children than for the first.

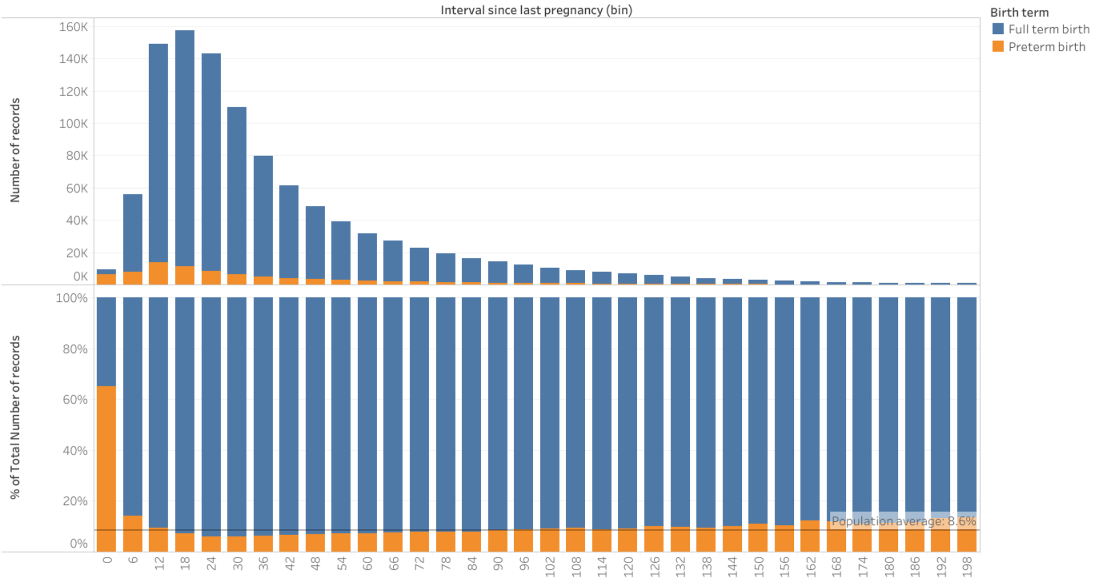
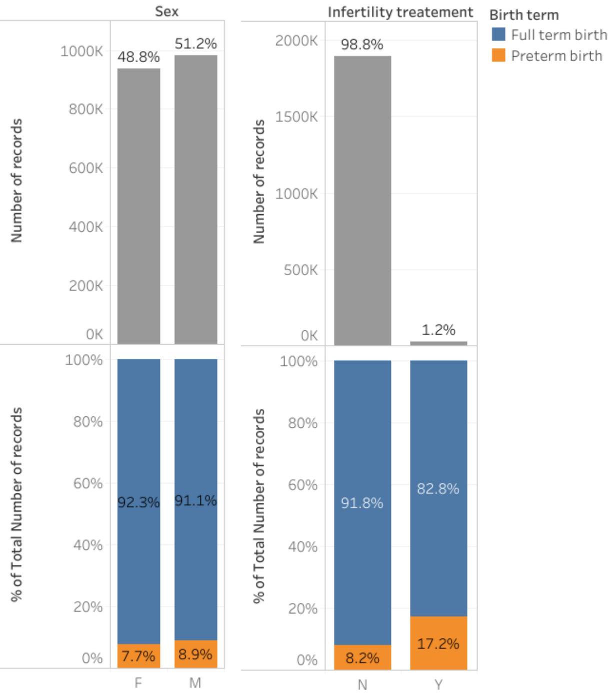


Figure 20: Intervals in months since last pregnancy have been grouped in 6-month bins. And interval of 0 indicates multiplets. There appears to be an ideal interval where the risk of premature birth is lower 18 to 83 months after the previous birth, with an optimum at 24-35 months. The effect on the right side might simply be due to the age of the mother.



(a) (b)

Figure 21: (a) Boys have a higher prematurity rate than girls. (b) Infertility treatment is correlated with more premature births.

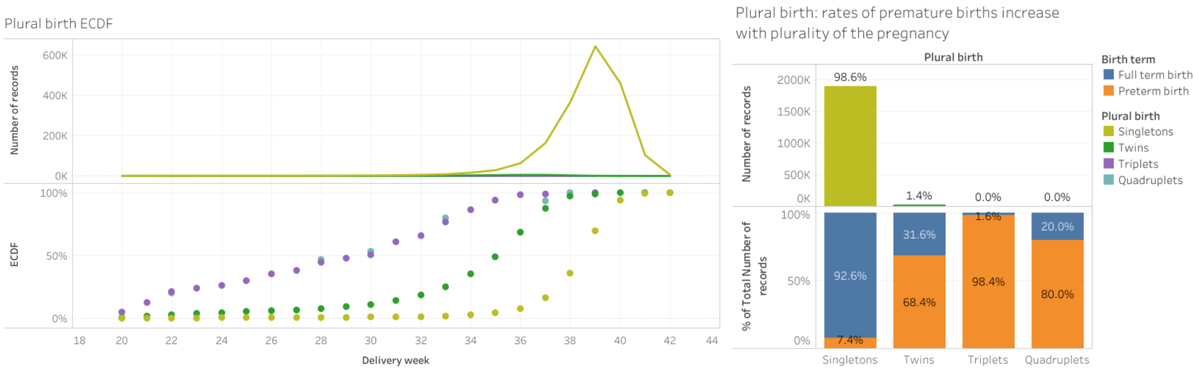
although the causes are unclear. It could be related to the size of multiple fetuses being a limitation.

4.1.16 Congenital anomalies of the baby

Most babies don't present congenital anomalies (99.7%). There is a significant difference in delivery week between babies presenting or not presenting congenital anomalies (37.5 weeks vs. 38.6 weeks). Amongst the babies presenting an anomaly, 23.1% were premature, but only 8.2% of the children without anomalies were premature (Figure 23, left).

4.1.17 Characteristics of the father

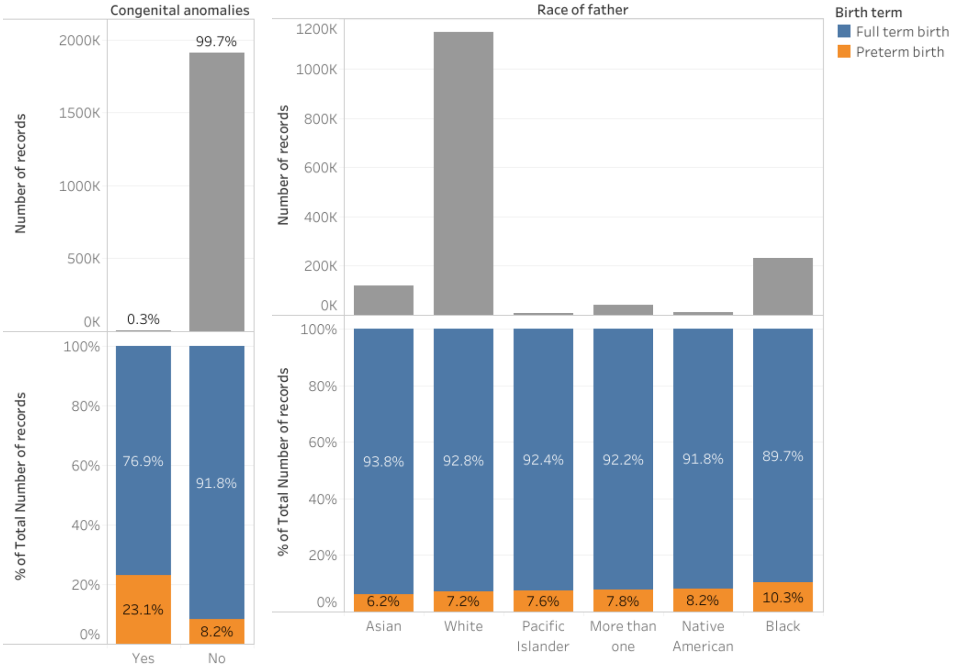
The age of the father is not a significant predictor for the week of birth, while his race and hispanic origin are. Only the plots for race are presented Figure 23 (right), since the hispanic origin doesn't follow a trend (See A). Race and hispanic origin have a physiological impact because of the genes associated with race and origin, and there is also the fact that people often marry within their race. In this dataset, only 10.5% of the couples were interracial (data used only when a race was reported for both parents, Figure 24). The reasons developed for the impact of the mother race might therefore be valid here as well. To investigate this, we ran a two-way ANOVA checking for the father race, the mother race, and the interaction between the two. Each variable brings significant variability to the delivery week, as well as the interaction between the two (significant



(a) (b)

Figure 22: Pregnancy of multiples is strongly correlated with earlier delivery weeks and more premature deliveries. The lower incidence of triplets and quadruplets makes the percentages less consistent.

interaction means that the combination of the parents' races is also correlated with the delivery week, these variables are not independent).



(a) (b)

Figure 23: (a) More premature deliveries are recorded when the babies present a congenital anomaly. (b) The premature rates by father race are very similar to the ones for the mother race. This might be due to genetic traits, especially with most couples being from the same race (Figure 24).

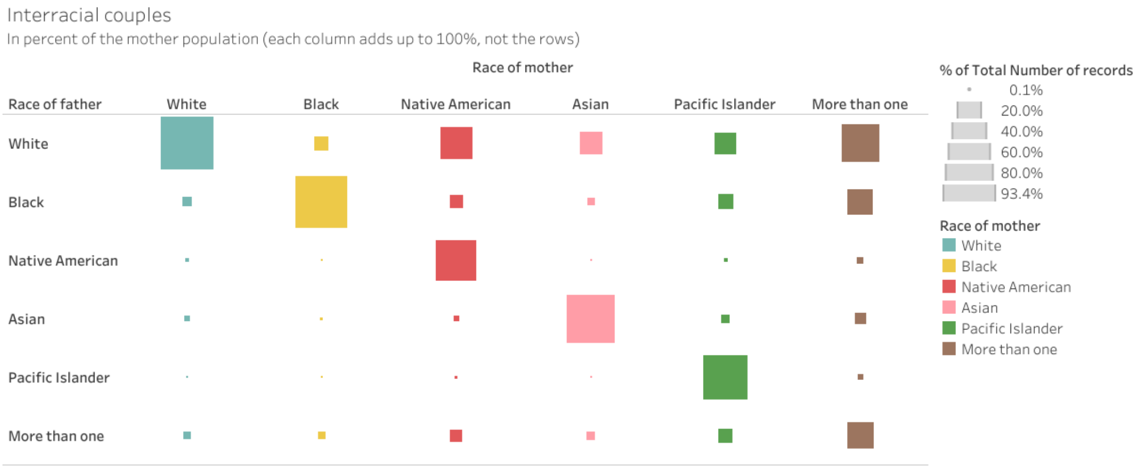


Figure 24: Race of the father in percentage of the mother race (columns add up to 100%). The diagonal presents the biggest squares, which indicates that most people have children within their race.

4.2 Socio-economic factors

4.2.1 Marital status and father involvement

For this variable (referred to as ‘father involvement’), there are three statuses: Married (the husband is the legal father), paternity acknowledged (the person acknowledging is the legal father), and no paternity acknowledged (no legal father). Visual EDA identified an effect from the paternity acknowledgement on the delivery week and prematurity rates. Figure 25 presents the rates per group: the more involved the father is, the lower the risks of premature birth. The effect is quite large: babies of married couples are almost half less likely of being premature than babies with no legal father (7.2% versus 13.4% of premature births, respectively)! This is confirmed by inferential statistics: the ANOVA indicates that the means between the groups are not equal at 38.7 weeks, 38.4 weeks and 38.1 weeks for married parents, paternity acknowledged, and no father, respectively.

These findings are interesting: they are surprising, and the father involvement can be acted upon, unlike biological factors. Before drawing conclusions, however, we need to make sure that the effect we observe is not correlated with a second variable, and that it is the second variable that is actually responsible for the effect. For example, one could argue that teenagers are less likely to be married, and teenagers have a higher prematurity rate than women in their 20-30s. Do we observe more prematurity for single mothers because they are single, or because they might be younger on average?

To control for the age variable, we ran an ANCOVA to capture separately the variability associated with the involvement of the father, the age of the mother and the interaction between the two. This way, we isolate the effect of each variable. If age is solely responsible for the higher prematurity rates, the ‘age’ F-value would be the only significant F-value among the three.

When we apply the procedure to age and father involvement, the F-value of the latter is still significant (the variability in delivery week is partially explained by the father involvement when

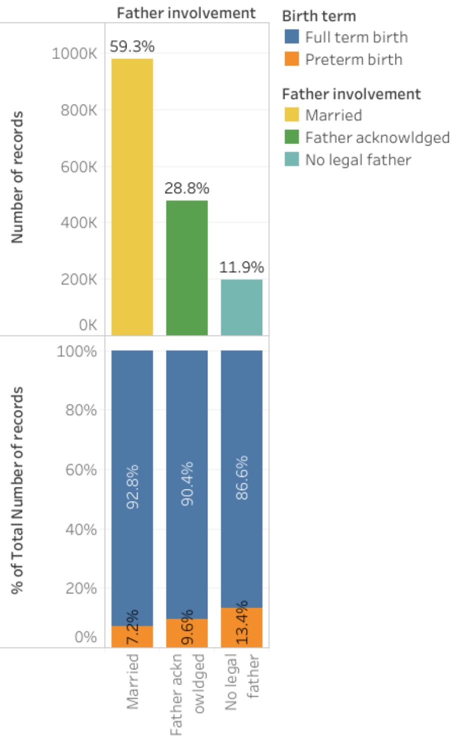


Figure 25: The involvement of the father has an impact on the prematurity rate: babies of married parents are less at risk for prematurity than babies who have been acknowledged by a father, who are less at risk than babies with no legal father. There are almost twice as many premature babies for single mothers compared to married ones.

we control for age). We can visualize it by splitting the age of the mother in bins and looking at the histogram of premature babies depending on the father involvement (Figure 26). The increase in premature rate when the father is less involved is seen throughout all age groups, despite differences in the distribution of observations between married, father acknowledged and no legal father groups (top part).

Similarly, we tested the impact of father involvement when controlling for race. Compared to other races, Black mothers are less married, and Asian mothers are more married (Figure 27, top). Since Black women deliver earlier than other races, the trend in father involvement could be a consequence of the effect seen between mother races, not a real trend. The ANCOVA resulted in a significant F-value for the father involvement when controlling for mother race. The Visual EDA leads to the same conclusion (Figure 27, bottom), with the exception of the Pacific Islander group.

These two two-way ANOVAs also indicated that the interaction between father involvement on one hand and age or race on the other hand is significant. A combined effect of each pair of variables on delivery week is therefore present. In other words, the variability in delivery week is explained by the father involvement, age and race, as well as by a combination of father involvement and age and a combination of father involvement and race.

After controlling for age and race, we believe the father involvement is a significant variable in predicting the delivery week, independent of other variables. The early delivery could be triggered

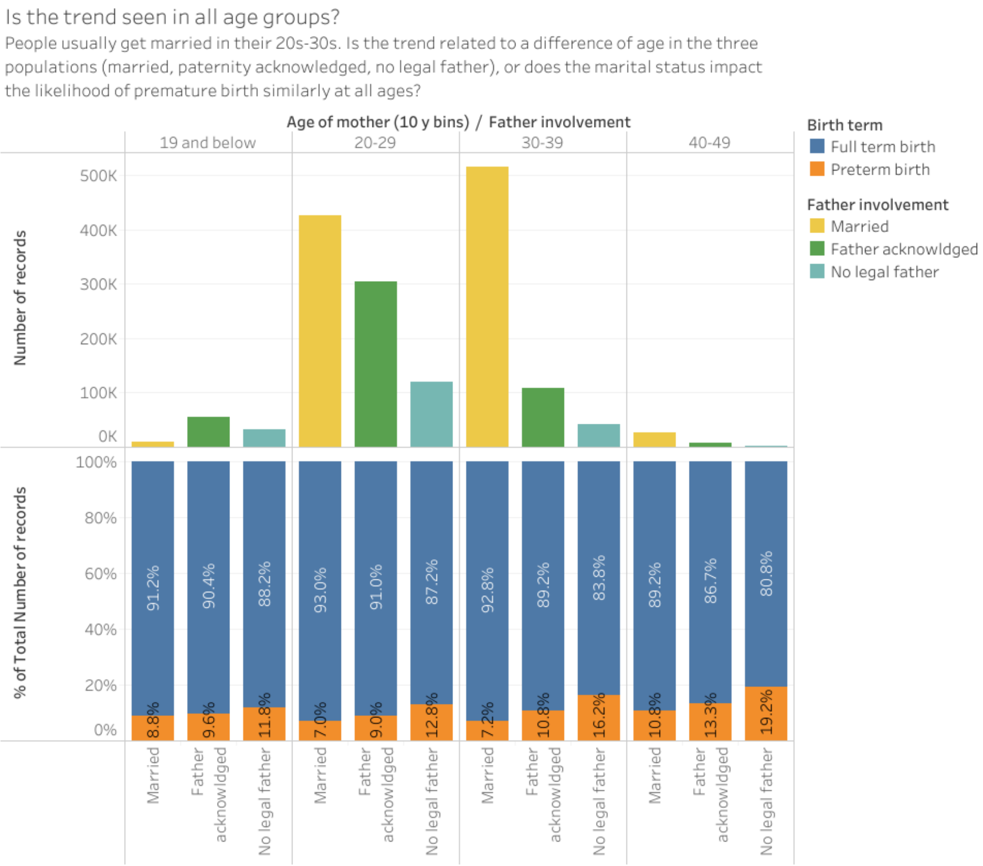


Figure 26: Prematurity rates by father involvement by age groups. Top: number of observations per group. Teenage mothers are proportionally less married than mothers in their 20s or older. Bottom: prematurity rates are different in each age group. However, they all follow the same trend: married mothers have lower prematurity rates than non-married mothers if a partner acknowledged the child, who have lower prematurity rates than single mothers with no father.

by the stress of the perspective of raising a child alone (financial, emotional), physical fatigue during pregnancy if there is no partner to help, or other psychological and physical factors.

4.2.2 Parents' education

Visual EDA indicates that the higher the education of the mother, the less likely she was to deliver prematurely (Figure 28, left). ANOVA confirms that the groups are different. This could be explained by the type of jobs women have access to based on their degrees: more education is usually correlated with less hard, physical labour. Physical fatigue could cause early deliveries. Another explanation lays in higher income, which leads to a healthier pregnancy through access to quality care, nutrition and services, as well as less anxiety regarding financial aspects of parenthood. It could also be that more education leads to better understanding of information about healthy pregnancy.

The same trend is observed for the level of education of the father (Figure 28, right), but this

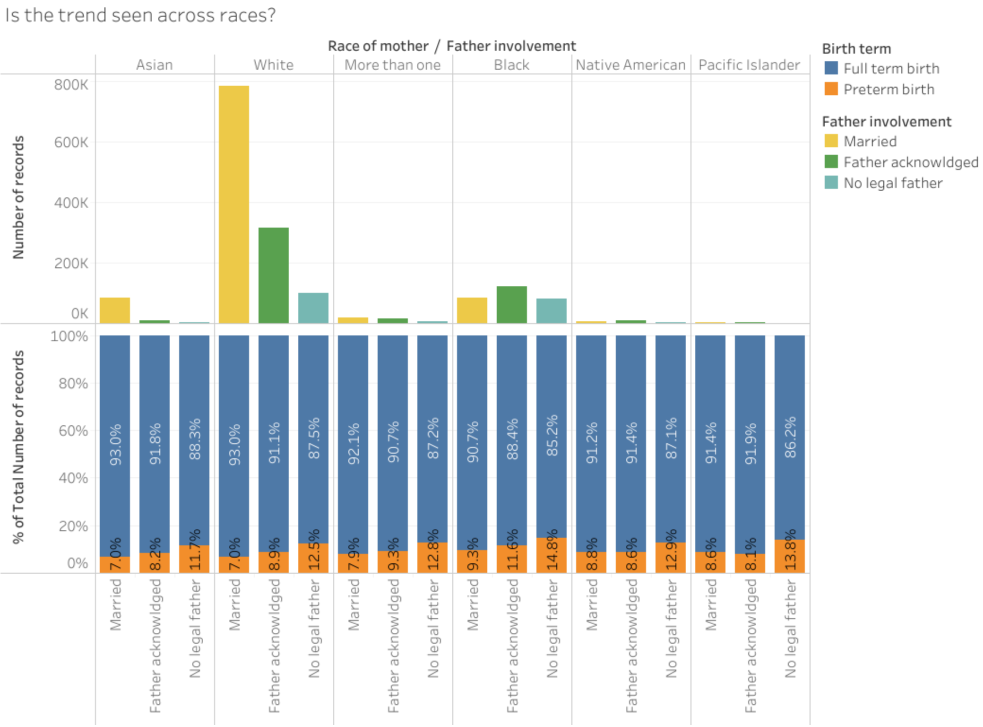
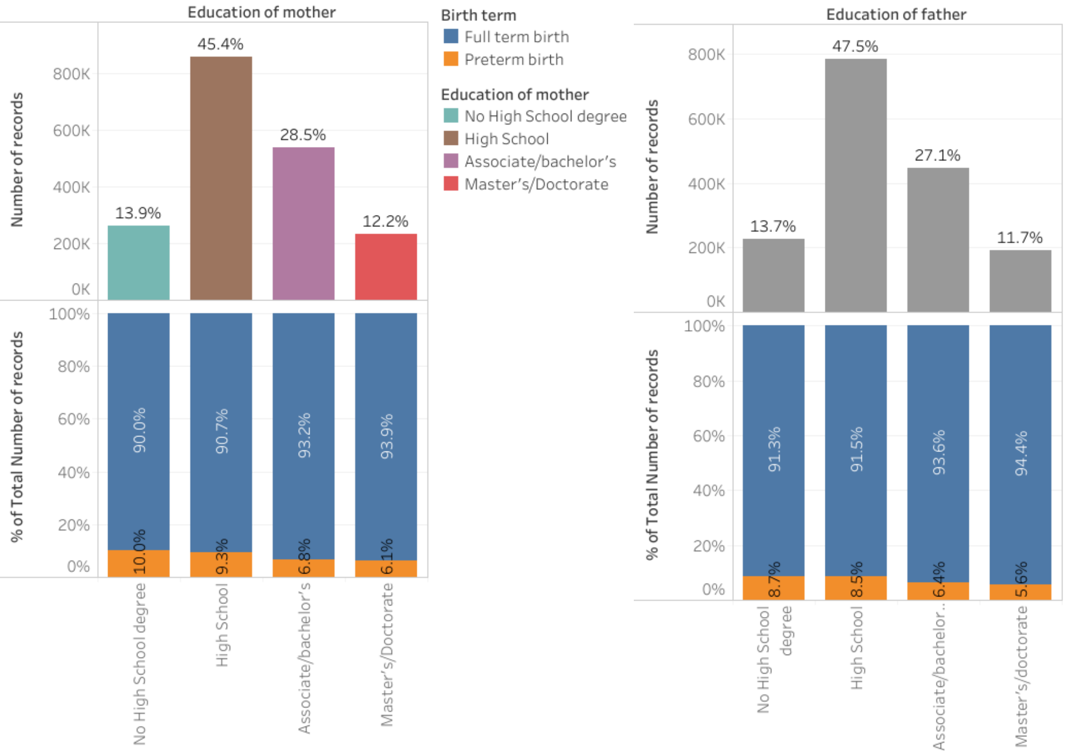


Figure 27: Premature rates by father involvement broken down by race. The top part depicts the frequency for each group. In the bottom graph, for each race group except Pacific Islanders, the premature rates follow the same trend as the general population: the rate increases with less father involvement.

observation could be related to the fact that the majority of people choose partners of the same education level (Figure 29).

Once again, we could argue that the level of education of the mother depends on her age, or her race. Even when we control for age or race, the education level is a significant variable in predicting the delivery week as seen on Figure 30 and 31 (only three races represented, the trend is the same for the other three). Analysis of the interaction of the sets of two variables indicate that they are not independent, and there is a combined effect of education and age, and education and race, respectively.

Now that we confirmed the significance of the level of education, we went back and controlled for the level of education of the mother when looking at the involvement of the father. Couples with higher education are more often married than the population average, and this could be the real reason why we see such a difference in prematurity rates by father involvement. The two-way ANOVA indicates that the the mother's education level is significant even when controlling for father involvement (Figure 32). It also indicates that the two variables, father involvement and mother education level, are not independent. These last tests (two-way ANOVAs) were not conducted on the father's education level.



(a)

(b)

Figure 28: (a) The level of education of the mother impacts the delivery week: mothers who are less educated have higher prematurity rates. (b) Father education is correlated with premature rate the same way as mother education.

4.2.3 WIC status

WIC is an income-based program aimed at pregnant women, infants, and children through age 5. The goal is to provide proper nutrition by supplying vouchers for food, nutrition counseling, health care screenings and referrals. It is administered by the U.S. Department of Agriculture.

The mean delivery week for mothers on WIC is 38.5 (38.6 for their counterparts not on WIC), and the distributions are different for the two populations. The difference is significant as determined per ANOVA. Women on WIC have a higher premature rate (Figure 33, left). This difference could be explained by overall economical difficulties (less access to good nutrition, medical and psychological support, financial stress over the new baby, etc.).

4.2.4 Payment method

The mean delivery week is 38.4, 38.7, 38.8, and 38.6 for women using Medicare, private insurance, self pay, or other, respectively. These differences are significant. The prematurity rates confirm this trend (Figure 33, right). Again, financial hardship may increase stress and limit access to care, explaining why some women deliver earlier.

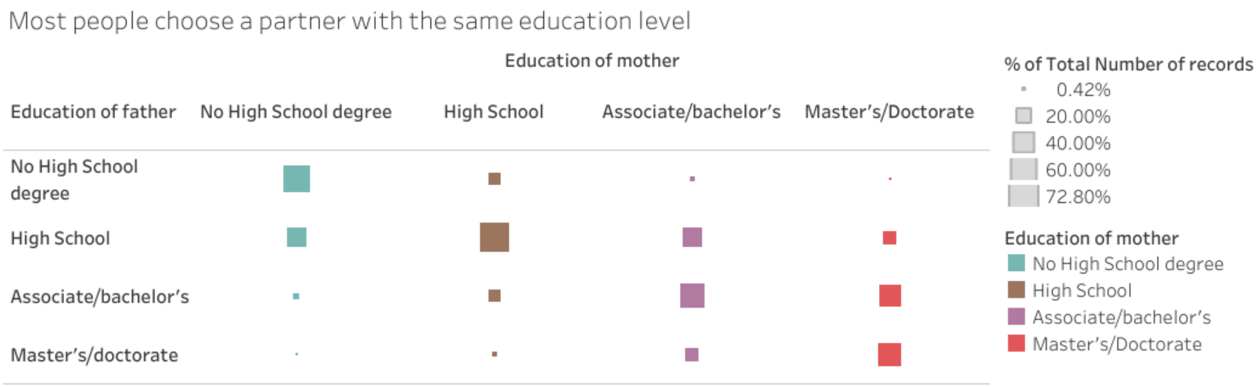


Figure 29: The correlation between mother and father education could be linked to couples having in general the same education level, as indicated by the large size of the diagonal.

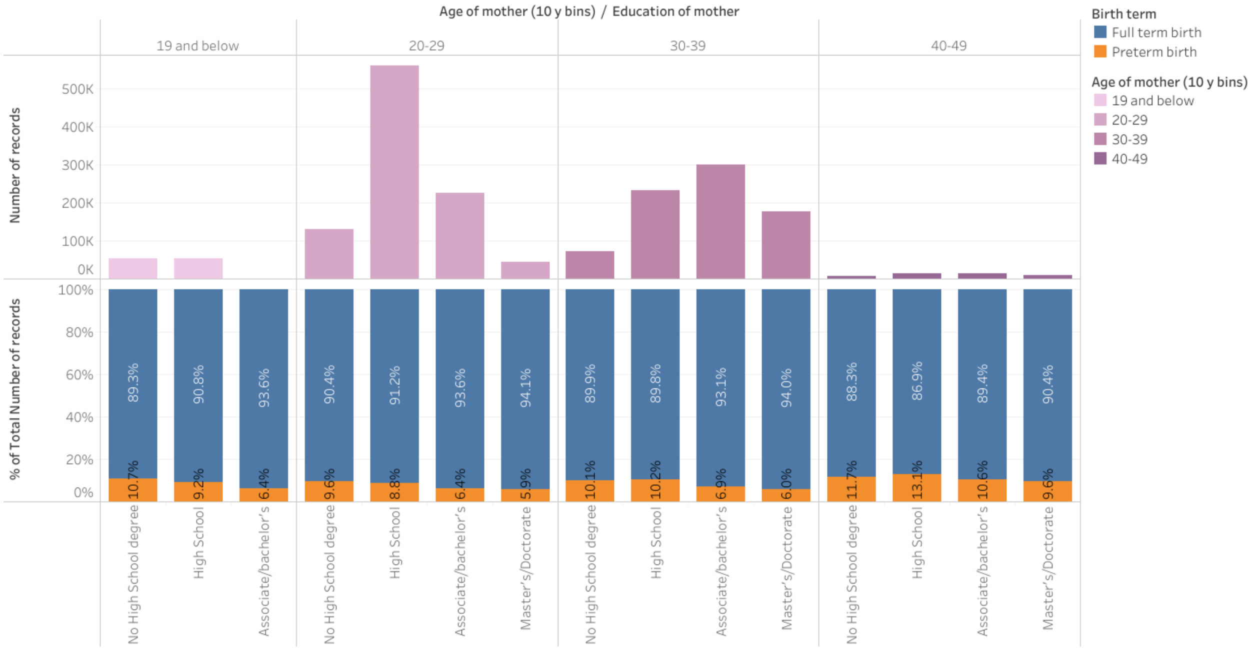


Figure 30: Prematurity rates by mother education broken down by age groups. The top graph depicts the number of records for each group. In the bottom graph, for each age group, the premature rates follow the same trend as the general population: the rate decreases with higher level of education. There is an exception for women in their forties who have a high school degree: they have higher risks for premature birth than other women in their forties.

4.2.5 Place of birth

Most mothers (96.8%) choose a hospital to deliver their babies. The mean delivery week is 38.5, 39.6, 39.7, 37.5, 39.2, 39.1, and 38.3 for babies born in a hospital, a birth center, at home (intended), at home (not intended), at home (unknown if intended), at a clinic/doctor's office, and other place, respectively. These differences are significant.

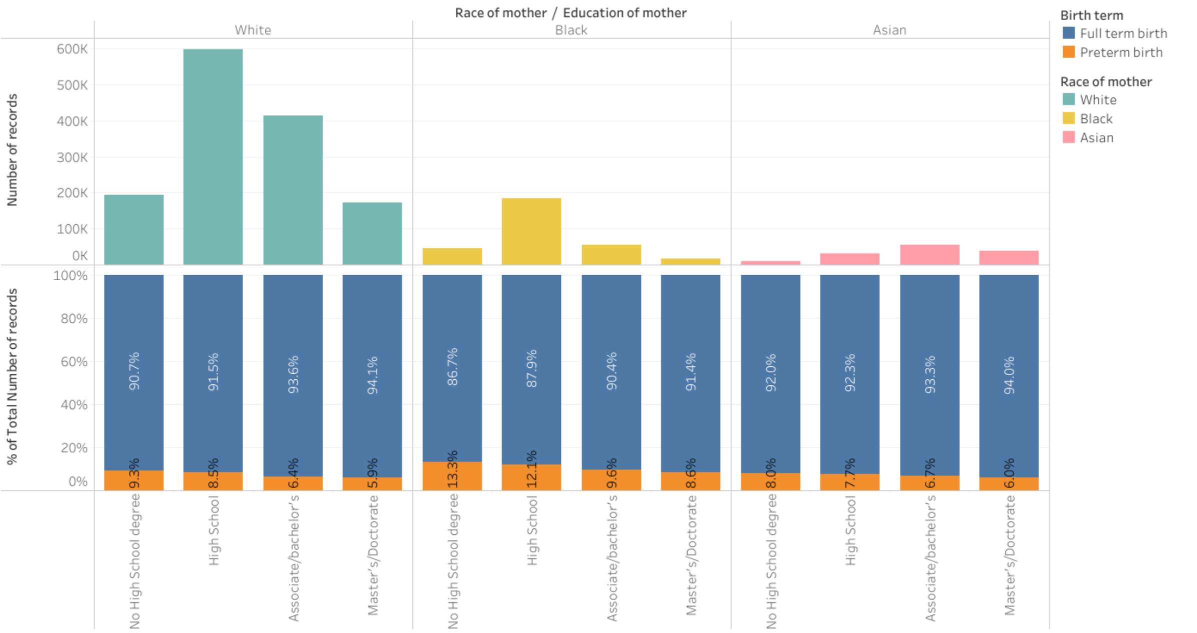


Figure 31: Rates of premature by mother education by race. For each race group, the prematurity rates follow the same trend as the general population: the rate decreases with higher level of education. The races with the most observations are presented, others show the same trend.

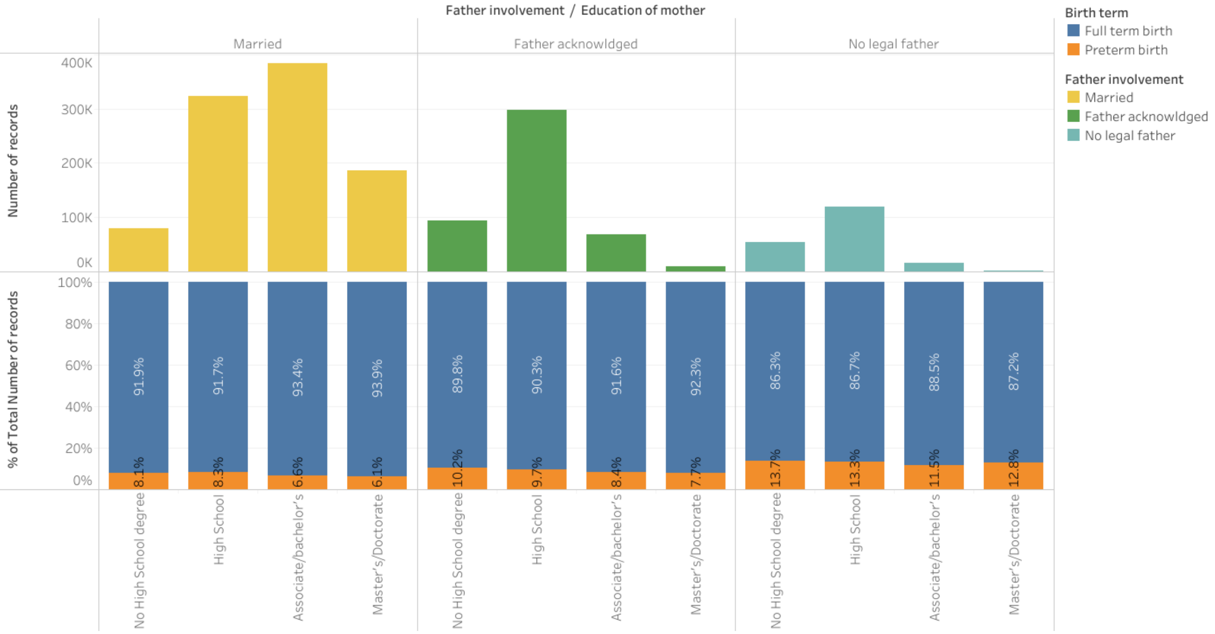
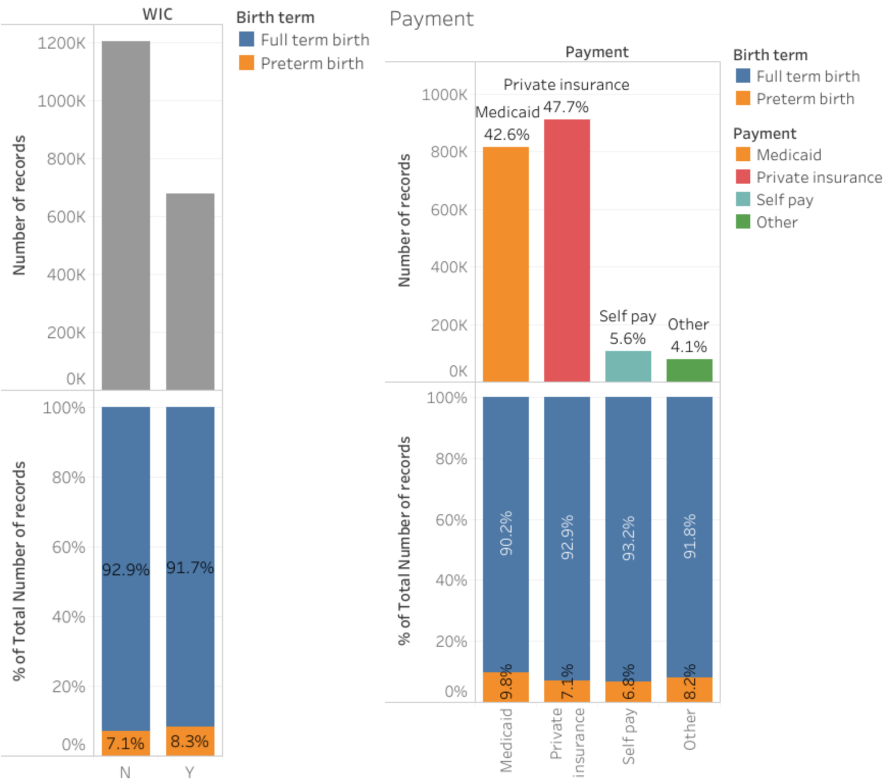


Figure 32: Prematurity rates by mother education broken down by father involvement. In the bottom graph, for each father involvement group, the premature rates follow the same trend as the general population: the rate decreases with higher level of education.



(a)

(b)

Figure 33: (a) WIC status is positively correlated with the premature rate (being on WIC means more risks of having a premature baby). (b) Families with more means (private insurance) tend to have a lower prematurity rate.

Scheduled home births and birth centers have the highest average delivery weeks of all the subgroups of the study, and a low prematurity rate (Figure 34). These births are associated with parents seeking a more natural, less medicated approach to childbirth. A better mental preparation or the willingness to let things happen naturally could explain the difference. Families who had originally planned an at-home birth might also have changed their plans if the baby was early, causing this group to be underrepresented.

4.3 Other factors

4.3.1 Mother nativity

This variable studies whether the mother was born inside or outside the US. The mean delivery week for mothers born in the US is 38.5, and 38.6 for mothers not born in the US. The one-way ANOVA indicates that this factor is significant. Non-US native mothers have slightly lower prematurity rate (Figure 35, left). One explanation could be that women who decide to immigrate are generally healthy, which could contribute to a healthy, full-term pregnancy. Mothers of unknown

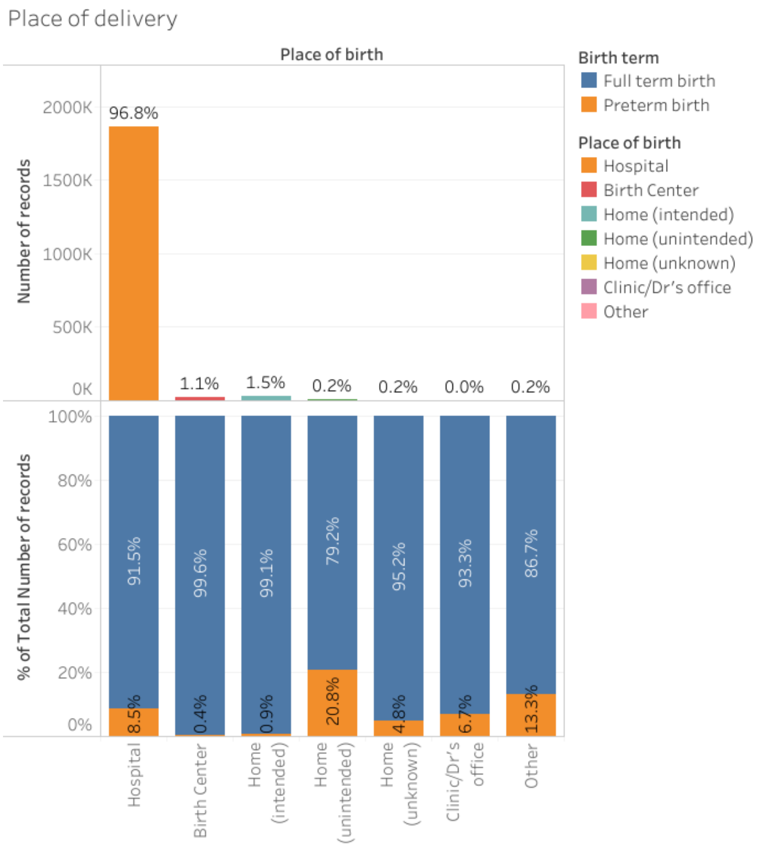


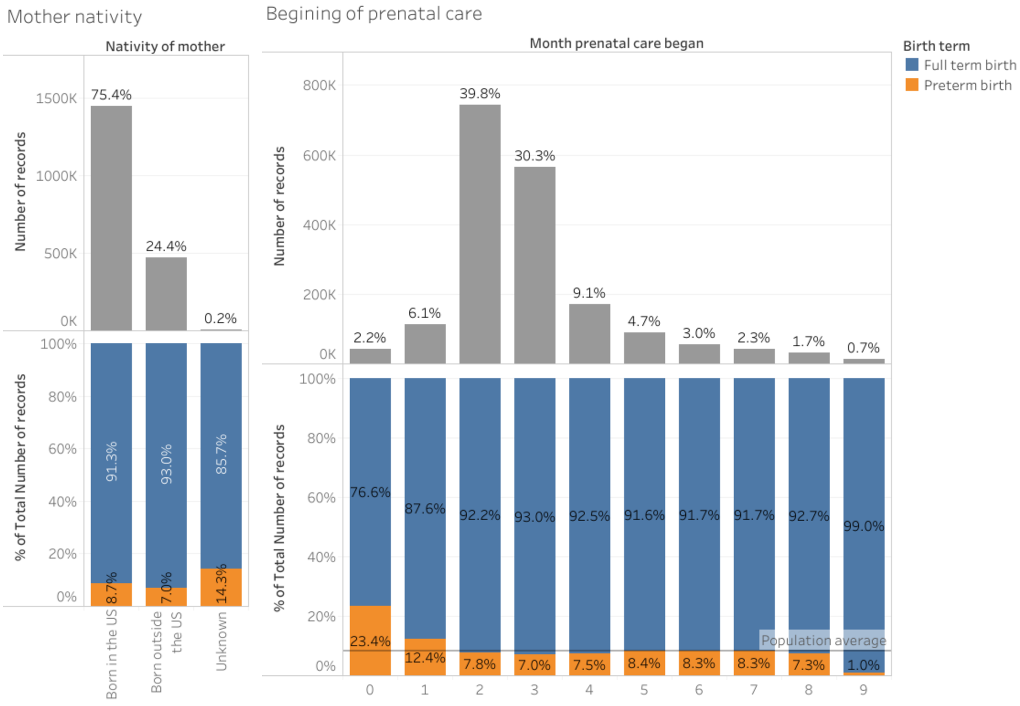
Figure 34: Babies born at home (scheduled) or a birth center have the lowest prematurity rate.

nativity deliver earlier and have more risks of prematurity. We might not know the nativity of these women because they are in a precarious situation, which could cause stress and ultimately an early delivery.

4.3.2 Month prenatal care began

Women usually start seeing a doctor by the end of the first trimestre. Overall, 70% of expecting women started their prenatal visits during month 2 or 3 (Figure 35, right top). The average month prenatal care began for full-term premature babies is 2.9, and 2.7 for premature babies. There are less premature babies when the mothers started to see a doctor after 2 months or more (Figure 35, right bottom). This effect might be related to what was seen with the infertility treatments: medically assisted pregnancies would be monitored from the very beginning and are correlated with more premature deliveries. The effect might also be seen because high risk pregnancies, which would be monitored earlier, are correlated with preterm delivery. This factor is not easy to interpret.

Finally, the increase in prematurity rate at 9 months is inevitable: a women who started seeing a physician at 9 months of gestation could not have delivered a premature baby (8 months of gestation or less).



(a) (b)

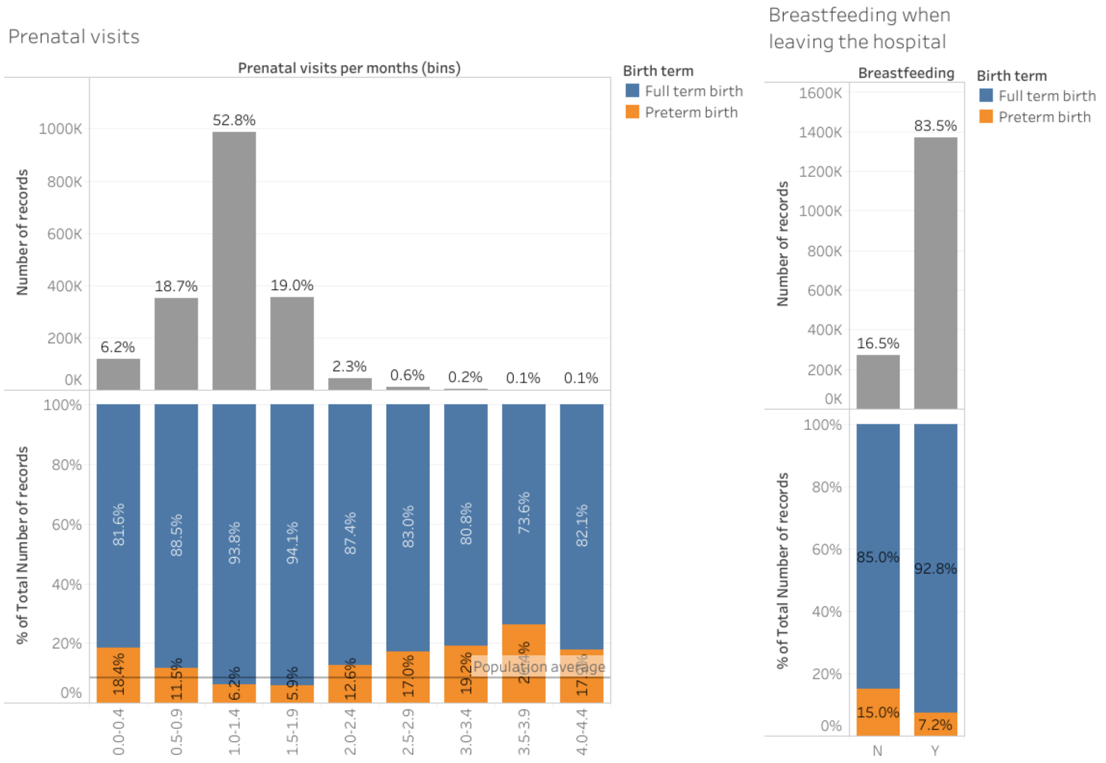
Figure 35: (a) Mothers of unknown nativity deliver earlier than women born in the US, who deliver earlier than women born outside the US. There are proportionally more preterm babies born from mothers of unknown nativity, but this group is very small (0.2% of the population). (b) Most mothers consult for the first time at 2-3 months of gestation. The prematurity rates are constant except for mothers who consulted very early (maybe related to health conditions) or very late (the average can only be high then).

4.3.3 Number of prenatal visits

The number of prenatal visits was not a representative variable in itself. Since preterm pregnancies imply less opportunities for prenatal visits, delivery week and the number of visits were strongly correlated. Instead, we divided the number of prenatal visits by the delivery week, yielding a number of visits per week then multiplied by 4.33, the number of weeks per months, to have a more relatable number.

The average number of prenatal visits per month for full-term babies is 1.26, and 1.13 for preterm babies. This variable is significantly correlated with the delivery week. There are more premature deliveries among women who have lower (less than 1 per months) and higher average numbers of prenatal visits (more than 2 per month, Figure 36). The former could be related to poor access to healthcare, denial of pregnancy or pregnancy being not known, while the latter can be linked to preexisting conditions/risk factors.

This variable cannot be used for prediction: how can a person know in advance how many visits she will have? The number of visits will not be used in the model but is reported here to present a comprehensive picture.



(a) (b)

Figure 36: (a) Over 70% of mothers see a doctor 1-2 times a month on average. The prematurity rate is higher for mothers who have a low rates (below 1 per month) or high rate of prenatal visits (2 visits per months and more). (b) Full term babies are more likely to be breastfed than preterm babies.

4.3.4 Delivery method

Women delivering by c-sections deliver at an average week of 38.5, slightly earlier than vaginal births (38.6). The difference is statistically significant. Once again, the information won't be available for prediction and will not be developed further.

4.3.5 Breastfeeding

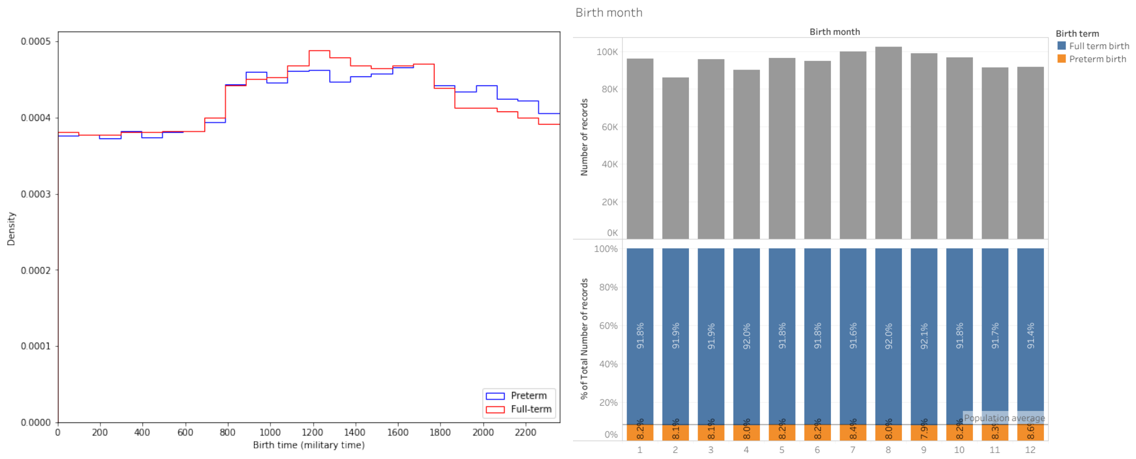
This variable is only available post-delivery, but can be interesting in the general context of perinatal care. About 84% of mothers are breastfeeding when they leave the hospital. Breastfeeding mothers have usually delivered closer to term (38.7 weeks) compared to mothers who were not breastfeeding (37.9 weeks). The difference is significant and rather large: there is almost a week of difference between the two groups. Visual EDA indicates that formula-fed babies have a high risk of being premature than their breastfed counterparts (Figure 36, right). This interesting relation could be studied further but is outside the scope of this project. Possible explanations are: not enough preparation if the baby is early, baby in the NICU making nursing complicated.

4.4 Time of birth variables

These variables were thought to be random, but will be investigated to be certain.

4.4.1 Time of birth

The time of birth should be random. The mean birth time for premature babies is 12:14, but the mean birth time for full-term babies is 12:04. This difference is significant, premature babies are born later in the day. The density distribution (Figure 37, left) indicates more full term births than preterm births between 9 AM and 4 PM, and more preterm births between 6 PM and midnight. We also notice less births during the night between midnight and 8 AM. These trends will not be investigated in the context of this project, and the variable will not be used for prediction (not available before delivery).



(a)

(b)

Figure 37: (a) The proportion of full term births is higher around noon (red trace higher between 9:00, 9 AM, and 16:00, 4 PM), and the proportion of preterm births is higher at night (blue trace higher past 18:00 or 6 PM) (b) .

4.4.2 Birth month

We expect a few more births in 30-day months and even less in February, which is the case (Figure 37, right). August and September see more births than other comparable months, and December has less. The prematurity rate varies slightly, and the difference is significant. This variable will be used for prediction because the month of birth can be roughly estimated.

4.4.3 Birth weekday

We expect this variable to be distributed uniformly between the seven days of the week, since we removed non-spontaneous births. On Figure 38, we see that there are less births on Saturday and Sunday. We might have missed some scheduled births, or the pace of weekdays trigger labor more than the weekend one. The difference is not significant between full term and premature babies, so the variable will not be used in the model (in addition, the variable is not available at the time of prediction).

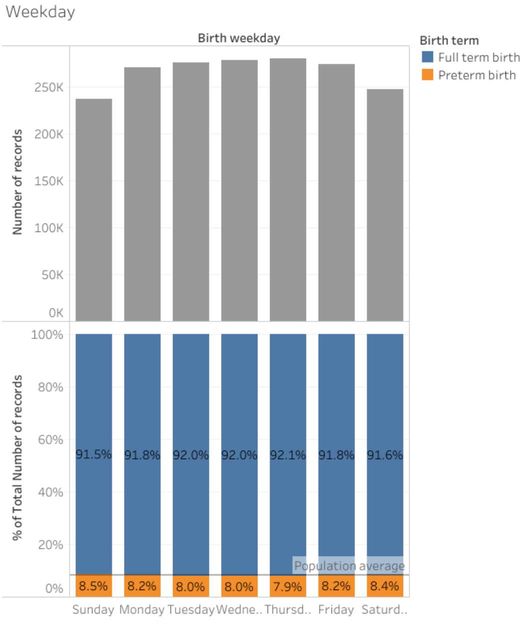


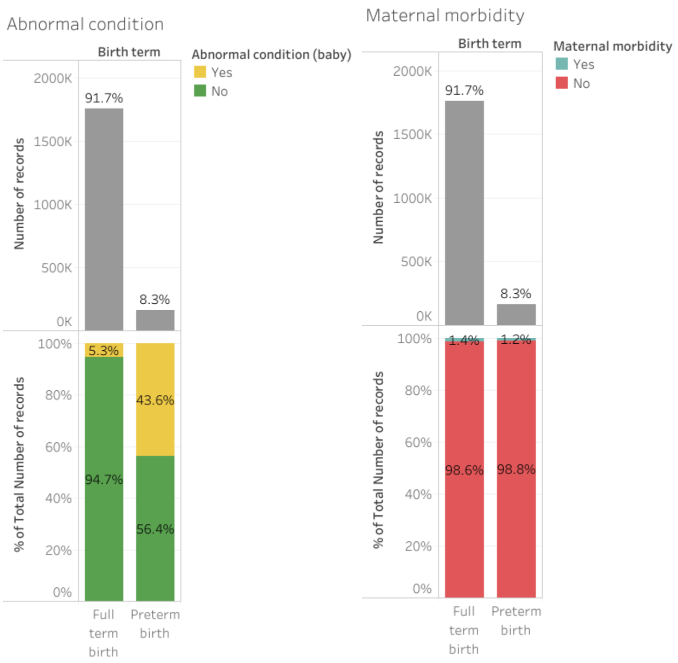
Figure 38: There are less births on Sunday and Saturday, but the premature rates are not statistically different between the seven days.

4.5 Outcomes of the pregnancy

These outcomes will not be used for prediction (not available before delivery), but are reported to emphasize the importance of a full term delivery.

4.5.1 Abnormal conditions

Premature birth is a source of abnormal conditions. The average delivery week for infants presenting an abnormal condition is 36.3. In contrast, infants without abnormal conditions were born at 38.8 weeks on average. This difference is significant. However, since causality between the two is established, it is more interesting to look at the numbers the other way: 5.3% f full-term infants have an abnormal condition, far less than the 43.6% of infants born prematurely with an abnormal condition (Figure 39, left).



(a) (b)

Figure 39: Outcomes of the pregnancy: (a) premature babies have on average eight times more risks of being born with an abnormal condition. (b) Mothers of premature babies are slightly less likely to suffer from conditions related to the delivery (ruptured uterus, excessive bleeding, etc.).

4.5.2 Maternal morbidity

The same reasoning holds for maternal morbidity. Maternal morbidity include possible negative outcomes of delivery on the mother: ruptured uterus, excessive bleeding leading to blood transfusion, perineal laceration, etc. The difference is less striking, with an average delivery week of 38.7 without maternal morbidity and 38.6 with maternal morbidity. On the other hand, 1.2% of mothers of preterm babies have maternal morbidity, and 1.4% of mothers of full-term babies (Figure 39, right). Giving birth to a smaller infant appears to limit the bad outcomes for the mother.

4.6 EDA Conclusions

4.6.1 Impact of biological and physiological factors on the delivery week and the prematurity rates

The impact of these variables is summarized Figure 40.

The physical attributes of the mother seem to impact the week of birth, with the extremes in each category leading to more premature births. In other words, there seems to be an optimal range of age, BMI, weight and weight gain for a full term pregnancy. The height of the mother is positively correlated with the week of birth, which means that the taller a woman is, the later she is likely to deliver.

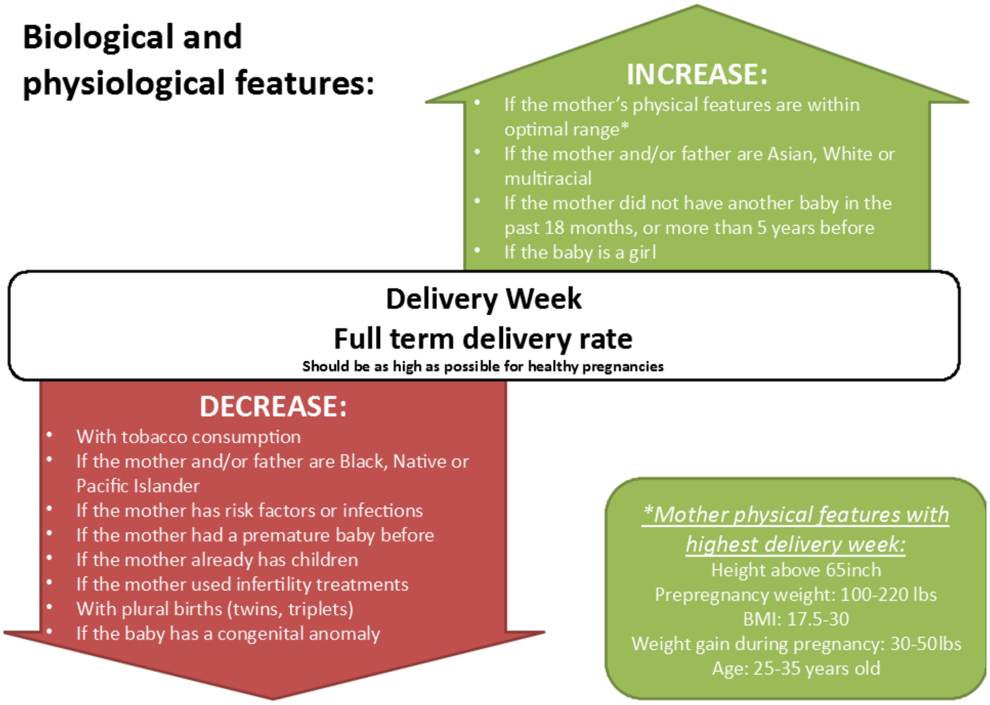

Biological and physiological features:
INCREASE:
If the mother's physical features are within optimal range*
If the mother and/or father are Asian, White or multiracial
If the mother did not have another baby in the past 18 months, or more than 5 years before
If the baby is a girl
Delivery Week
Full term delivery rate
Should be as high as possible for healthy pregnancies
DECREASE:
With tobacco consumption
If the mother and/or father are Black, Native or Pacific Islander
If the mother has risk factors or infections
If the mother had a premature baby before
If the mother already has children
If the mother used infertility treatments
With plural births (twins, triplets)
If the baby has a congenital anomaly
*Mother physical features with highest delivery week:
Height above 65inch
Prepregnancy weight: 100-220 lbs
BMI: 17.5-30
Weight gain during pregnancy: 30-50lbs
Age: 25-35 years old

Figure 40: Summary of how the biological and physical variables affect the average delivery week and rate of full term pregnancies in the population. This figure was already presented in section 2.

Asian, White and multiracial mothers and/or fathers have later deliveries than the global average, while Black, Native American and Pacific Islander mothers and/or fathers have earlier deliveries than the global average. Hispanic origin of the parents is a significant factor, however it is not clear in which direction it impacts the delivery date.

Smoking, risk factors (diabetes, preeclampsia, etc.) and infections during pregnancy are all correlated with earlier births and a higher prematurity rate.

The family history is also significant. Mothers who already had a preterm baby are more at risk of having another one. Furthermore, with each pregnancy, the delivery week tends to decrease and the risk of preterm birth increases. For comparison, a first born has 6.9% risks of being premature, whereas a child 8th or more in birth order has 12.6% of risks of being premature. Mothers who deliver a baby between 18 and 60 months after the previous pregnancy tend to deliver later and have less risks of premature delivery than mothers having babies very close to another, or with a wider gap than 5 years.

Boys are generally born earlier than girls, and are more likely to be premature.

Mothers who used infertility treatments tend to deliver earlier and are at higher risks for premature delivery than those who did not (prematurity rate of 14.9% as opposed to 7.5% without treatment).

Plural births are unsurprisingly correlated with preterm births, with 98.4% of triplets born before 37 weeks, 68.4% of twins and only 7.4% of singletons born before 37 weeks.

Finally, congenital anomalies of the baby are correlated with earlier deliveries and higher prematurity rates (21.1% as opposed to 7.5% without anomalies).

4.6.2 Impact of socioeconomic and other factors on the delivery week and the prematurity rates

The impact of these variables is summarized Figure 41.

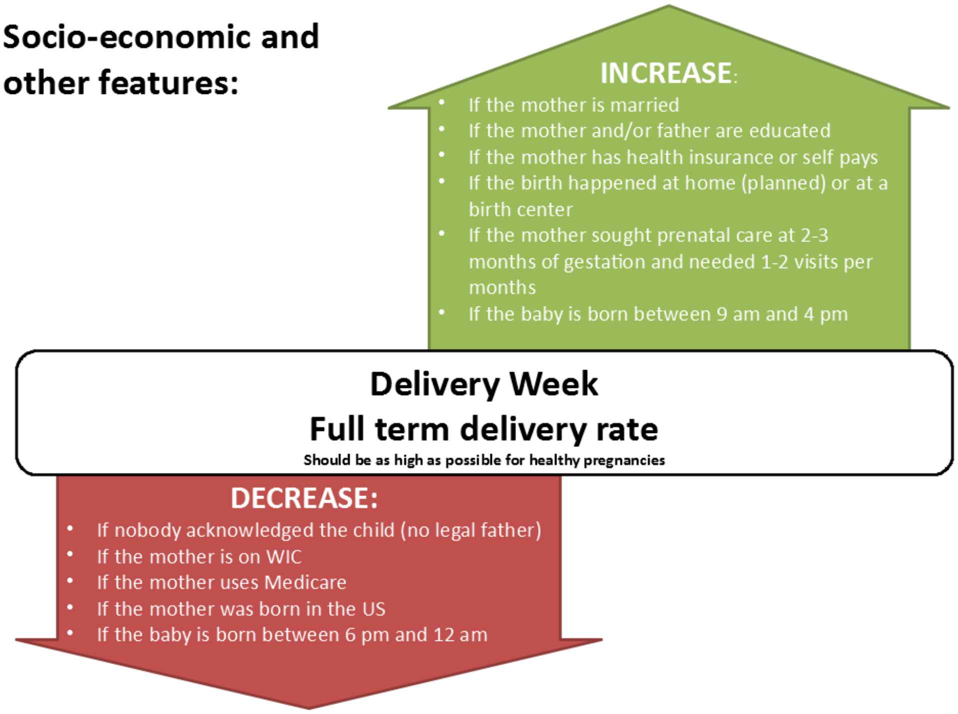

Socio-economic and other features:
INCREASE:
If the mother is married
If the mother and/or father are educated
If the mother has health insurance or self pays
If the birth happened at home (planned) or at a birth center
If the mother sought prenatal care at 2-3 months of gestation and needed 1-2 visits per months
If the baby is born between 9 am and 4 pm
Delivery Week
Full term delivery rate
Should be as high as possible for healthy pregnancies
DECREASE:
If nobody acknowledged the child (no legal father)
If the mother is on WIC
If the mother uses Medicare
If the mother was born in the US
If the baby is born between 6 pm and 12 am

Figure 41: Summary of how the socio-economic and other variables affect the average delivery week and rate of full term pregnancies in the population. This figure was already presented in section 2.

To our surprise, the involvement of the father plays a significant and rather large role in predicting the prematurity rate. There are three situations: the mother is married, her husband is the legal father of the child, the mother is not married but a person acknowledged the child, becoming the legal father, or nobody acknowledged the child, and there is no legal father. The proportion of premature deliveries in the three groups is 7.2%, 9.6% and 13.4%, respectively. A single mother has almost twice as much risk as a married mother to give birth to a premature baby!

This variable was investigated further: we needed to make sure the effect was really related to the involvement of the father, not to a correlation of the father involvement with another variable. The effect was still there when controlled for age and race.

The education of the parents displays a similar trend: if the parents are more educated, there is less risk for premature birth. We controlled for age, race, and father involvement, and still observed the same trend.

WIC status impacts the delivery week negatively, and prematurity rate positively (proportionally more premature babies if the mother is on WIC).

Payment method and place of birth indicated that mothers on Medicaid had higher risks for a premature delivery than others, and that giving birth at a birth center or at home were correlated with the lowest prematurity rates. Mother nativity shows a difference for mothers born outside of the U.S.: they have less risks of prematurity.

Prenatal care factors are probably correlated with other factors. We see that there is an optimal time to seek prenatal care (during the second or third month), and an optimal number of prenatal visits (1-2 per months, not necessarily distributed homogeneously). These variables are more likely outcomes of other factors (high risk pregnancy, or denial o pregnancy for example)

The delivery method is not known at the time of prediction (we only predict spontaneous births), and this reasoning goes also for breastfeeding, abnormal condition(s) of the newborn and maternal morbidity.

Finally, to our surprise, the time of birth was a significant variable in predicting if a baby was premature, as well as the month of birth. The day of the week was not.

4.6.3 Summary

Out of all the variables tested, the father's age and weekday of birth were the only variable with a non-significant correlation with the dependent variable. The variables that are correlated and can be used for the model (because they are available at the time of prediction) are:

- Height of the mother
- Age of the mother
- Prepregnancy weight (as log)
- Tobacco consumption
- Mother race
- Mother hispanic origin
- Risk factors
- Infections
- Previous preterm birth
- Birth order
- Interval since last pregnancy (as log)
- Sex of the baby
- Infertility treatment
- Plural birth
- Race of the father
- Hispanic origin of the father
- Father involvement/Marital status
- Mother's education
- Father's education
- WIC status
- Payment method
- Place of birth
- Mother nativity
- Month prenatal care began
- Month of birth

Factors which are correlated but cannot be used for prediction because they are not known at the time of prediction are:

- Time of birth
- Weight gain
- Number of prenatal visits
- Delivery method
- Breastfeeding
- Abnormal condition(s) (sometimes detected in utero but we cannot generalize)
- Congenital abnomaly(ies) (sometimes detected in utero but we cannot generalize)
- Maternal morbidity
- Birth weight
- Trial of labor

Some other factors are not used for prediction for reasons described below:

- BMI of the mother (because it is correlated with weight and height)
- Father age (not a significant predictor)
- Weekday of birth (not a significant predictor and not available at the time of prediction)

5 Machine learning models and predictions

This section describes the machine learning process applied in Python (scikit-learn, adaboost and catboost packages). All notebooks are uploaded on Github.

The imputed dataset from data analysis was adapted as follows, except for Catboost models:

- All categorical variables were dummified. Birth month was turned into a categorical variable.
- Variable 'interval since last pregnancy' was transformed in two steps:
  - First, we took the log of the value in months (because the distribution was skewed to the right). The new distribution appeared to be normal. For about a third of the observations, the interval since last pregnancy was coded 888 for women who did not have a previous pregnancy. After applying the log function, the value was far enough from the other values (log of interval in months) to be clearly identified.

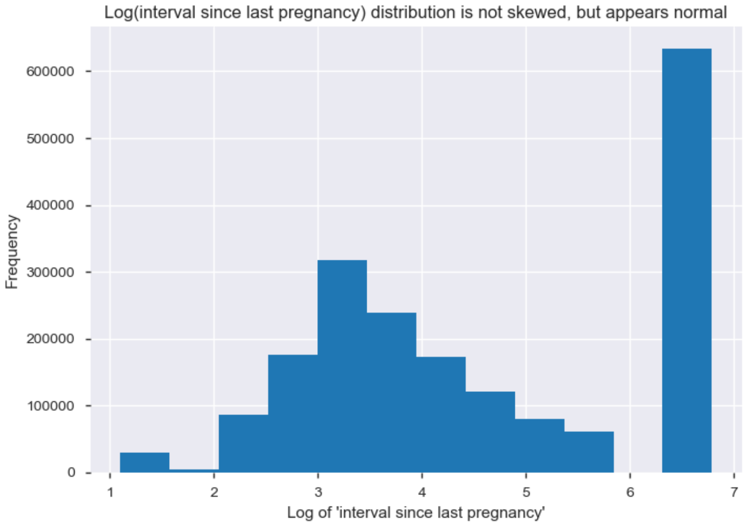


Figure 42: Log-transform of the interval since last pregnancy: the distribution was skewed and now appears normal. The bin with large count at 6.5 represents NaN values, coded 888, for first-time mothers.

- Then, we split the distribution into 10 bins, and grouped the observations in 10 categories. One category was the transformed 888 for first-time mothers. The histogram Figue 42 describe the categories after the log transformation:

- Prepregnancy weight' was also transformed by the log function to obtain a normal distribution.

Catboost models were trained on the unimputed, undummified dataset. No variables were transformed.

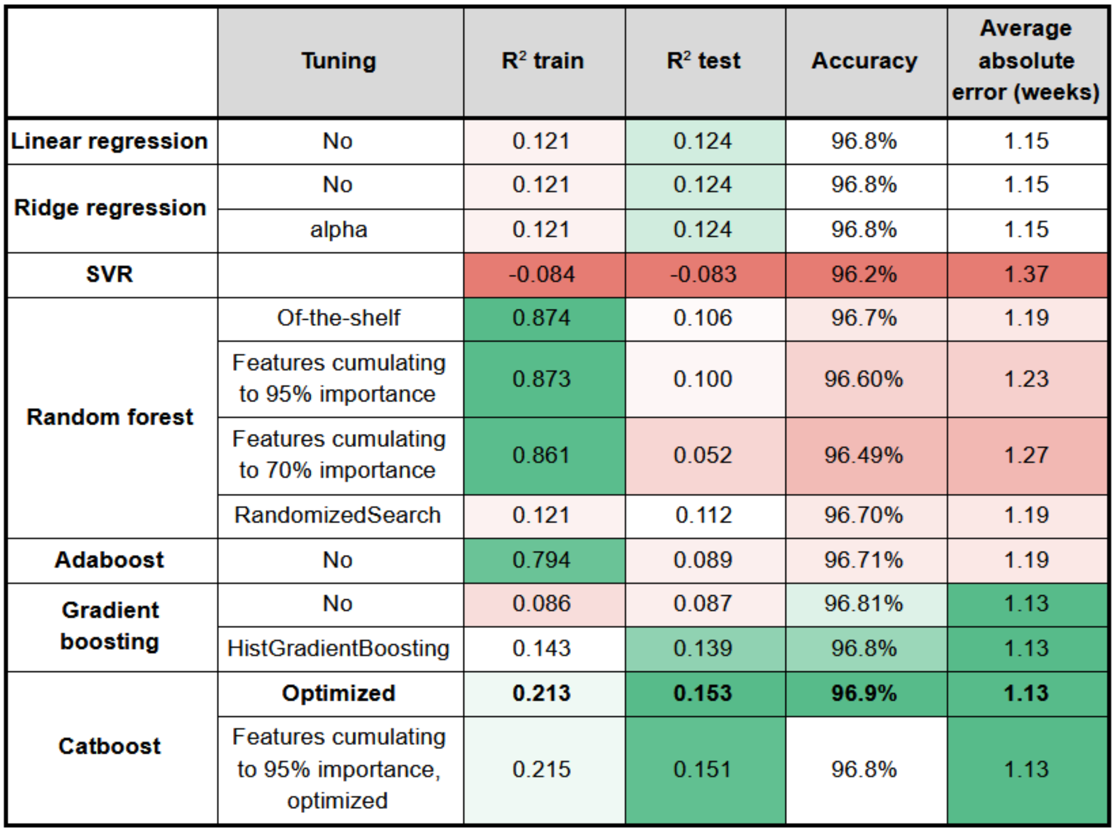
Initially, we tried to predict the delivery week (regression problem). However, the granularity of the problem made it difficult to predict relevant information: since we only knew the week of birth, not the day, we could only predict delivery up to the week. The range we were targeting is 20-42 weeks, and, with a mean absolute error of 1-2 weeks, the prediction was not very informative. In addition, all R^2 were not very promising. The best regression model we obtained was with Catboost, where the test R^2 was 0.153, the average absolute error was 1.13 weeks, and the accuracy was 96.9%. The models are described below (section 5.1).

We then switched to predicting the term of birth (premature or full term), turning the problem into a classification problem (section 5.2). The best model was obtained with Catboost. The recall was 58% and area under the ROC 0.73.

5.1 Regression

For the sake of the exercise, we tried multiple models, summarized in Table 43.

The accuracy is generally high since most women deliver week 38-40. More specifically, 19.4% of women delivered at 38 weeks, 33.7% at 39 weeks and 24.1% at 40 weeks, for a total of 77.2%



	Tuning	R² train	R² test	Accuracy	Average absolute error (weeks)
Linear regression	No	0.121	0.124	96.8%	1.15
Ridge regression	No	0.121	0.124	96.8%	1.15
	alpha	0.121	0.124	96.8%	1.15
SVR		-0.084	-0.083	96.2%	1.37
Random forest	Of-the-shelf	0.874	0.106	96.7%	1.19
	Features cumulating to 95% importance	0.873	0.100	96.60%	1.23
	Features cumulating to 70% importance	0.861	0.052	96.49%	1.27
	RandomizedSearch	0.121	0.112	96.70%	1.19
Adaboost	No	0.794	0.089	96.71%	1.19
Gradient boosting	No	0.086	0.087	96.81%	1.13
	HistGradientBoosting	0.143	0.139	96.8%	1.13
Catboost	**Optimized**	**0.213**	**0.153**	**96.9%**	**1.13**
	Features cumulating to 95% importance, optimized	0.215	0.151	96.8%	1.13

Figure 43: Summary of the performances of regression models. Catboost did best with an accuracy of 96.9%, but the R^2 values were very low.

of births occurring at week 38-40. The models are decribed below:

1. Linear regression

   Training R^2 and testR^2 are very low. The distribution of the residuals (Figure 44) indicates that there is a pattern in the data that the model cannot capture. It is likely not a linear model.

2. Ridge/Lasso regression

   There is a small improvement with Lasso, but the predictions are still poor. Other models not based on linear regression might perform better.

3. SVR

   SVR performs very poorly on this dataset. We used the linear kernel however, which is the only one suitable for such a large dataset. SVM doesn't seem suitable for our problem.

4. Random forest

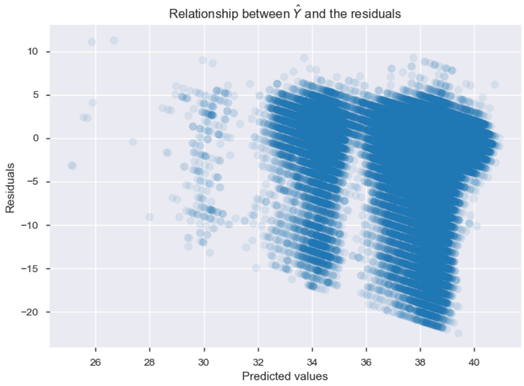


Figure 44: Residuals vs. predicted value of the delivery week. The distribution is not random: we see three vertical bands on the plot. This indicates that the model does not explain the data very well. If it were, only the random error would be left to plot, and the distribution would be random.

The first attempt with the default parameters of the model was not very successful, with a good training R2 (0.86) but a terrible R2 for the test data (0.08). This indicates that the model is overfitting, which is clear also from the scatter plots on Figure 45: the observed vs. predicted values are matching for the train set (45 degree line), but not for the test set.

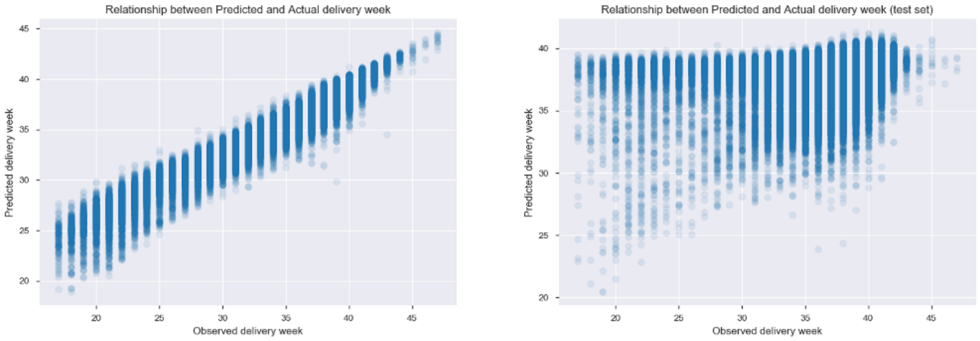


Figure 45: Predicted vs. actual value of the delivery week: the model performs well on the train set (left, plot forms a 45 degree line), but very poorly on the test set (right).

We selected the most important features (95% of importance, Figure 46) to avoid overfitting, but the performances went down. The 95% importance dataset was used to tune the Random Forest hyperparameters via a randomized search, and there was less overfitting for performances closer to the linear regression (training R^2 of 0.121 and testing R^2 of 0.112).

5. Adaboost

   Of-the-shelf Adaboost has a good training R^2 (0.79) but low test R^2 (0.09). This indicates overfit, just like the regular Random Forest. We will try two more gradient boosting algorithms to see if they are more suitable.

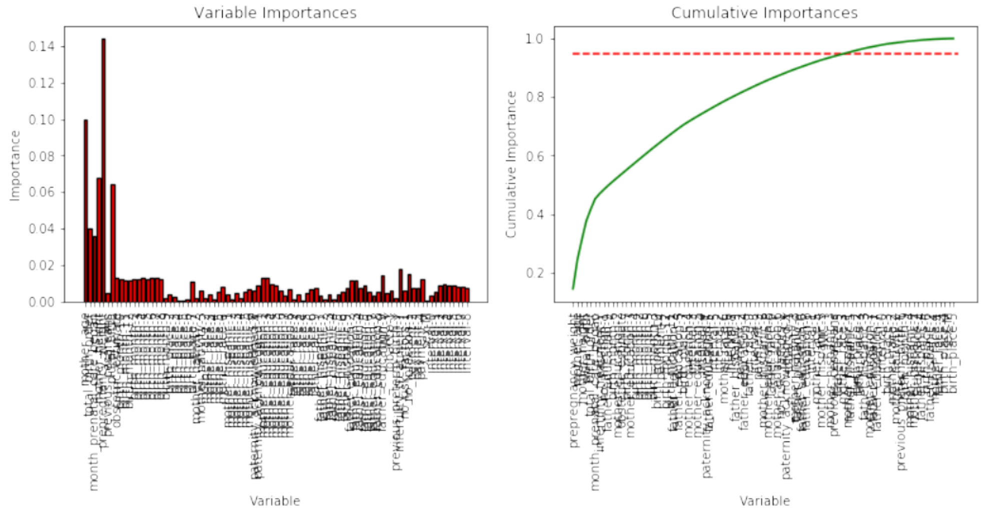


Figure 46: Feature importance: chart (left) and cumulative normalized importance (right).

6. Gradient boosting

   The regular gradient boosting shows poor performances but is not overfitting. We upscaled to a Histogram-based Gradient Boosting which has improved performances for large dataset, and improved the training R^2 (0.143) and test R^2 (0.139). These results are higher than the linear regression without overfitting, and therefore make HistGradientBoosting the best algorithm so far.

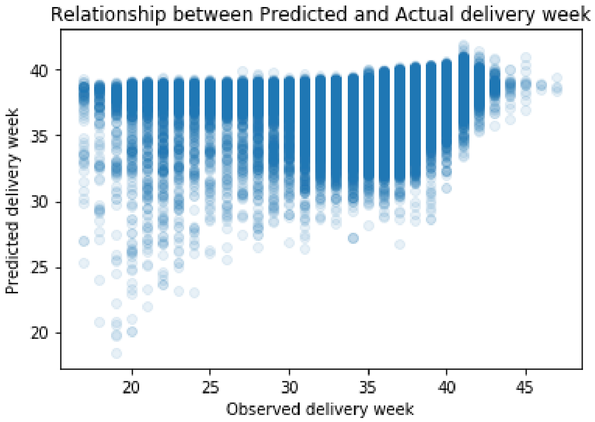


Figure 47: Predicted vs. actual value of the delivery week: test set only. The test predictions are not matching the true values, but are more aligned than previous models.

7. Catboost

   Catboost uses gradient boosting with special handling of categorical features. The first model with hyperparameters tuned had the highest test R^2 so far: 0.153. This is still not great, but there is little overfitting (training R^2 is 0.213). The scatter plots indicate that there are still lots of predictions misassigned in the test set (Figure 47), and the distribution of residuals is still not random (Figure 48).

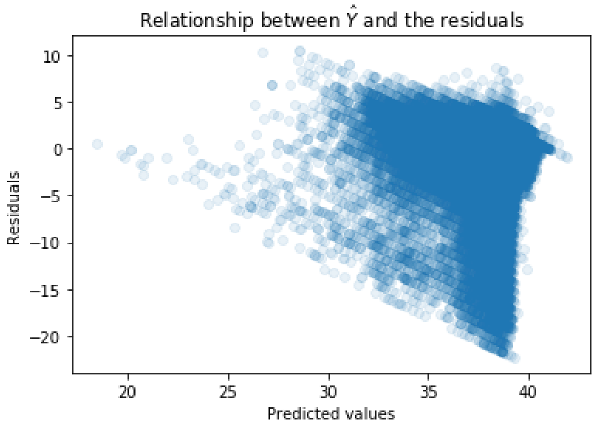


Figure 48: Residuals vs. predicted value of the delivery week. The distribution is not random: there is a higher concentration of observations on the top-right side, and the bottom left quadrant is completely empty. This indicates that the model does not explain the data very well. If it were, only the random error would be left to plot, and the distribution would be random.

The model does not capture the variability completely, but is the best we have so far. We tried to select important features accounting for 95% of importance before tuning again, but the new model wasn't performing better.

5.2 Classification

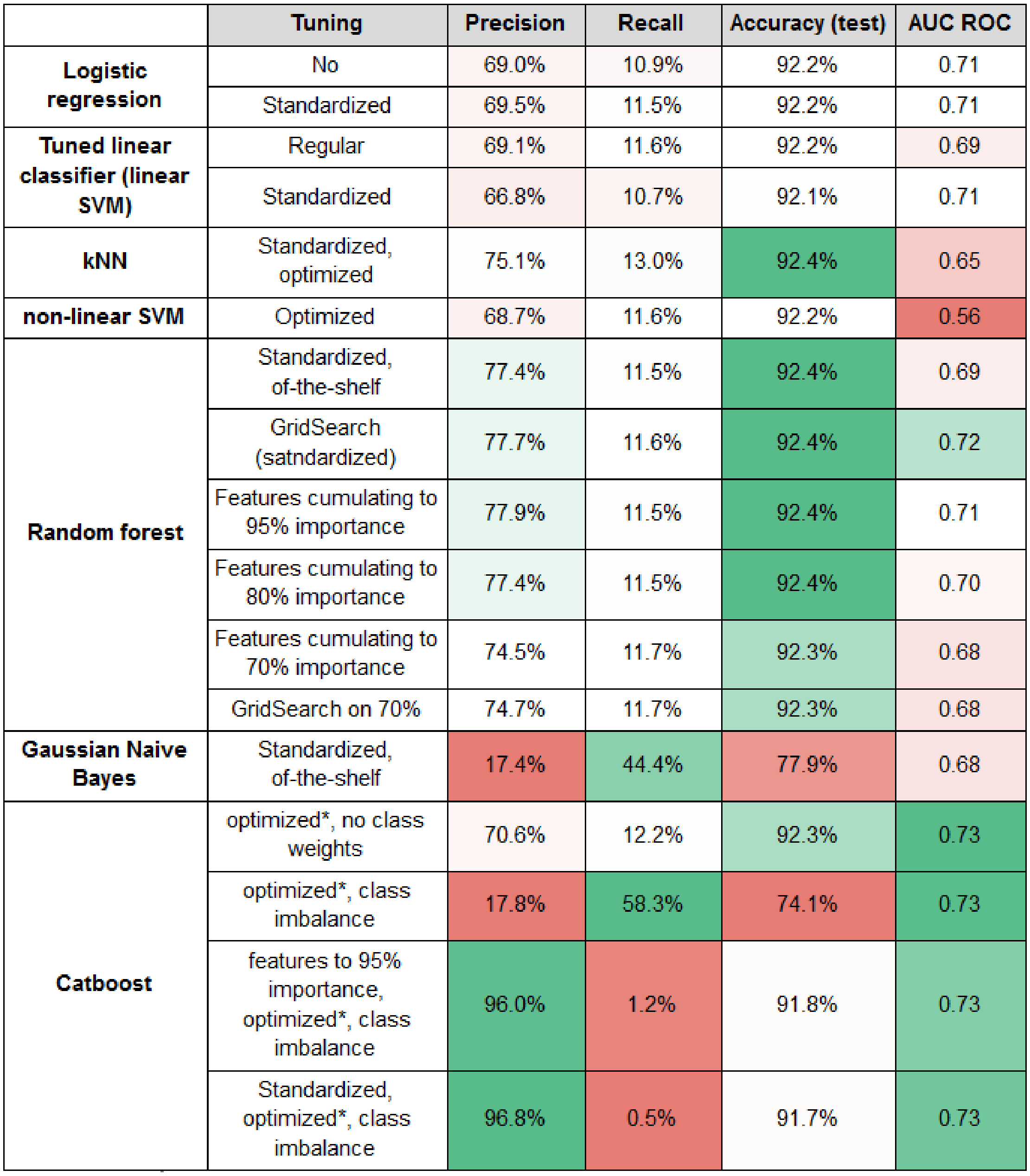
With the class imbalance that is present in the dataset (8.3% of premature births), accuracy is a poor metric for assessing how well a model performs (91.7% accuracy would be obtained by predicting only full term births, but that would not make it a great model). Instead, we should focus on optimizing the area under the ROC.

Unfortunately this is not what we did initially, so the work prior to Catboost used the accuracy as the metric for hyperparameter tuning. Another important consequence of the class imbalance is that we did a stratified split between train and test sets.

The dataset was too large for the tuning of hyperparameters of kNN and SVM, so we performed the tuning on 5% of the dataset only. The summary of the techniques applied and corresponding performances are presented in Table 49.

1. Logistic regression

   We tried to use a basic logistic regression on the raw data then on standardized data. The results are not promising, but standardization improved the model. From now on, we will use standardized data.



	Tuning	Precision	Recall	Accuracy (test)	AUC ROC
Logistic regression	No	69.0%	10.9%	92.2%	0.71
	Standardized	69.5%	11.5%	92.2%	0.71
Tuned linear classifier (linear SVM)	Regular	69.1%	11.6%	92.2%	0.69
	Standardized	66.8%	10.7%	92.1%	0.71
kNN	Standardized, optimized	75.1%	13.0%	92.4%	0.65
non-linear SVM	Optimized	68.7%	11.6%	92.2%	0.56
Random forest	Standardized, of-the-shelf	77.4%	11.5%	92.4%	0.69
	GridSearch (satndardized)	77.7%	11.6%	92.4%	0.72
	Features cumulating to 95% importance	77.9%	11.5%	92.4%	0.71
	Features cumulating to 80% importance	77.4%	11.5%	92.4%	0.70
	Features cumulating to 70% importance	74.5%	11.7%	92.3%	0.68
	GridSearch on 70%	74.7%	11.7%	92.3%	0.68
Gaussian Naive Bayes	Standardized, of-the-shelf	17.4%	44.4%	77.9%	0.68
Catboost	optimized*, no class weights	70.6%	12.2%	92.3%	0.73
	optimized*, class imbalance	17.8%	58.3%	74.1%	0.73
	features to 95% importance, optimized*, class imbalance	96.0%	1.2%	91.8%	0.73
	Standardized, optimized*, class imbalance	96.8%	0.5%	91.7%	0.73

Figure 49: Summary of the performances of classification models. Catboost did best with a precision of 96.8% or a recall of 58.3%, both with an area under ROC of 0.73. The results are still disappointing in terms of performance.

2. Ridge/Lasso regression

   SGD Classifier can use Lasso and/or Ridge to regularize a dataset. Hyperparameter tuning points towards a Lasso, not Ridge, for better accuracy. However, no improvement was achieved compared to the logistic regression by reducing the number of features from 87 to

45.

3. kNN

   The optimized number of neighbors was 19, and the weight calculated by distance. The model took a long time to train and predict due the size of the dataset and the high number of neighbors. Performances were disappointing.

4. Non-linear SVM

   Hyperparameter tuning returned a sigmoid kernel, a C of 1 and gamma of 0.001. Despite the optimization, the results were lower than the logistic regression.

5. Random forest

   The first attempt with the default parameters of the model was quite successful compared to previous classification models, with a AUC of 0.56 but a precision of . We selected the most important features to avoid overfitting, but the performances went down.

   This model was optimized via a grid search, but the performances were not improved. We removed features that do not contribute much to the model (keeping features sorted by importance cumulating to 95% of total importance, then 80 and 70%). Only at 70% of importance did we start to see a decrease in accuracy, indicating that we were removing features containing relevant information. We did one last optimization on the 70% dataset but the predictions were still not very accurate.

6. Naive Bayes - Gaussian

   We did a quick test with the out-of-the-box scikit-learn algorithm, and obtained promising results. By then we had started to use Catboost, and those results were even more promising, so we moved on.

7. Catboost classifier

   Catboost works almost on the raw data, just like we did in 5.1, 7. We specified the weights of the classes according to the documentation, and played with the most important features. In our case, we would trade off some precision for a better recall: we would rather warn too many people about possible premature births even if we are wrong rather than mislabel a risky pregnancy as 'no-risk'. Given that fact, we would settle on the model with optimized hyperparameters with a class imbalance. The AUC ROC is 0.73 and the recall is 58%.

5.3 Conclusions

Catboost performed best , but the predictions are still not very good. We should have addressed the class imbalance earlier, and used a better metric for hyperparameter tuning in classification problems. For future developments, we can work on:

- Resampling for class imbalance - play with over/undersampling
- Work with neural networks (we have enough data and could even add data from other years)

A Appendix

All the visualizations are available at an interactive format on Tableau Public. This includes visualisations not presented in the document. See the three documents here.