Improving Fetal Health Outcomes: Data Cleaning Process and Exploratory Data Analysis

A closer look into the technical process..

Fetal Health Data Set

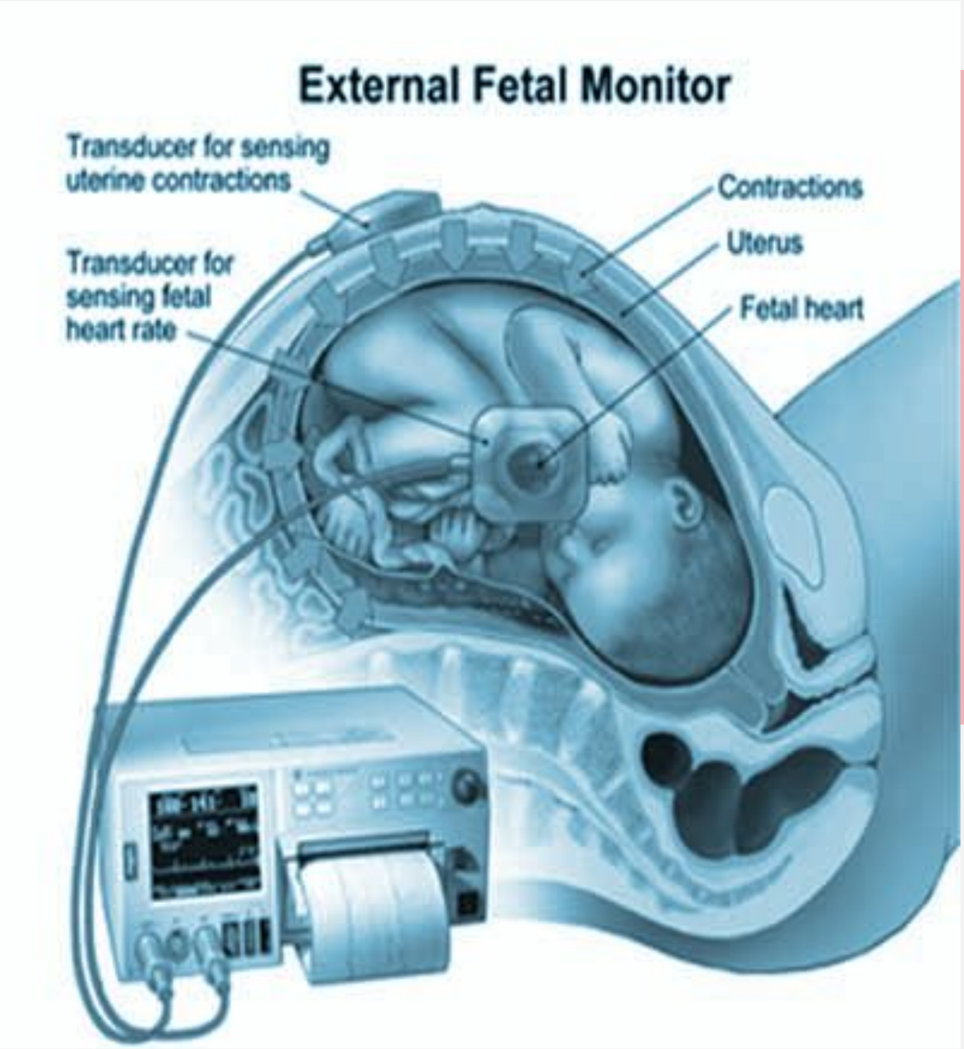
“Reduction of child mortality is reflected in several of the United Nations' Sustainable Development Goals and is a key indicator of human progress.

The UN expects that by 2030, countries end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce under 5 mortality to at least as low as 25 per 1,000 live births.

Parallel to notion of child mortality is of course maternal mortality, which accounts for 295,000 deaths during and following pregnancy and childbirth (as of 2017). The vast majority of these deaths (94%) occurred in low-resource settings, and most could have been prevented.

In light of what was mentioned above, Cardiotocograms (CTGs) are a simple and cost accessible option to assess fetal health, allowing healthcare professionals to take action in order to prevent child and maternal mortality. The equipment itself works by sending ultrasound pulses and reading its response, thus shedding light on fetal heart rate (FHR), fetal movements, uterine contractions and more.”

Ayres de Campos et al. (2000) SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318



Cardiotocogram (CTGs)

Fetal Health Dataset

```
filename = '/content/fetal_health.csv'
df = pd.read_csv(filename)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2126 entries, 0 to 2125
Data columns (total 22 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   baseline value                                          2126 non-null   float64
 1   accelerations                                           2126 non-null   float64
 2   fetal_movement                                          2126 non-null   float64
 3   uterine_contractions                                    2126 non-null   float64
 4   light_decelerations                                     2126 non-null   float64
 5   severe_decelerations                                    2126 non-null   float64
 6   prolongued_decelerations                                2126 non-null   float64
 7   abnormal_short_term_variability                         2126 non-null   float64
 8   mean_value_of_short_term_variability                    2126 non-null   float64
 9   percentage_of_time_with_abnormal_long_term_variability  2126 non-null   float64
 10  mean_value_of_long_term_variability                     2126 non-null   float64
 11  histogram_width                                         2126 non-null   float64
 12  histogram_min                                           2126 non-null   float64
 13  histogram_max                                           2126 non-null   float64
 14  histogram_number_of_peaks                               2126 non-null   float64
 15  histogram_number_of_zeroes                              2126 non-null   float64
 16  histogram_mode                                          2126 non-null   float64
 17  histogram_mean                                          2126 non-null   float64
 18  histogram_median                                        2126 non-null   float64
 19  histogram_variance                                      2126 non-null   float64
 20  histogram_tendency                                      2126 non-null   float64
 21  fetal_health                                            2126 non-null   float64
dtypes: float64(22)
memory usage: 365.5 KB
```

Data Cleaning

```
df.isnull().sum()

#No missing values
```

```
baseline value                                            0
accelerations                                             0
fetal_movement                                            0
uterine_contractions                                      0
light_decelerations                                       0
severe_decelerations                                      0
prolongued_decelerations                                  0
abnormal_short_term_variability                           0
mean_value_of_short_term_variability                      0
percentage_of_time_with_abnormal_long_term_variability    0
mean_value_of_long_term_variability                       0
fetal_health                                              0
dtype: int64
```

Resolving duplicated rows

```
[11] df.duplicated().sum()

13

[12] #13 Duplicated rows will be dropped

[13] df.drop_duplicates(inplace = True)

[14] df.duplicated().sum()

0
```

Eliminating unnecessary columns

Removing unnecessary columns

```
[15] df.drop(columns = ['histogram_width', 'histogram_min', 'histogram_max', 'histogram_number_of
```

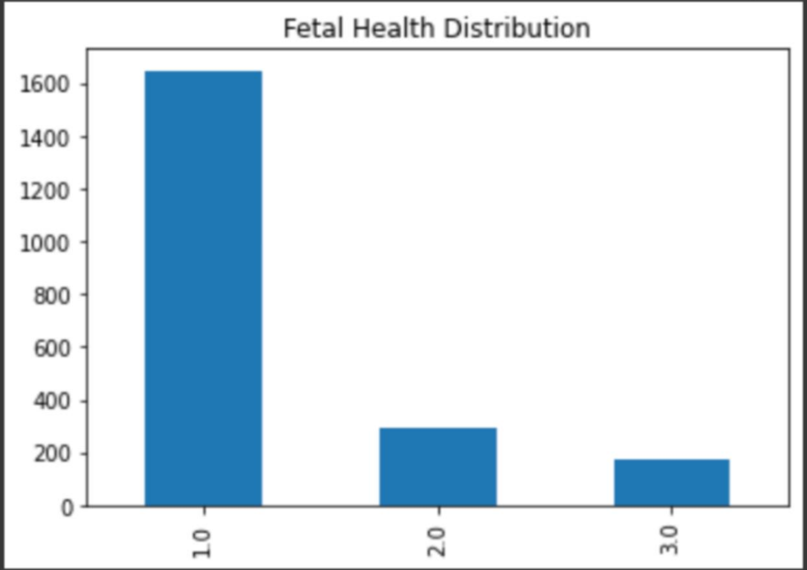
```
df.head().T
```

	0	1	2	3	4
baseline value	120.0	132.000	133.000	134.000	132.000
accelerations	0.0	0.006	0.003	0.003	0.007
fetal_movement	0.0	0.000	0.000	0.000	0.000
uterine_contractions	0.0	0.006	0.008	0.008	0.008
light_decelerations	0.0	0.003	0.003	0.003	0.000
severe_decelerations	0.0	0.000	0.000	0.000	0.000
prolongued_decelerations	0.0	0.000	0.000	0.000	0.000
abnormal_short_term_variability	73.0	17.000	16.000	16.000	16.000
mean_value_of_short_term_variability	0.5	2.100	2.100	2.400	2.400
percentage_of_time_with_abnormal_long_term_variability	43.0	0.000	0.000	0.000	0.000
mean_value_of_long_term_variability	2.4	10.400	13.400	23.000	19.900
fetal_health	2.0	1.000	1.000	1.000	1.000

Visualizing Fetal Health Distribution in this dataset:

```
[17] df['fetal_health'].value_counts()

1.0    1646
2.0     292
3.0     175
Name: fetal_health, dtype: int64
```

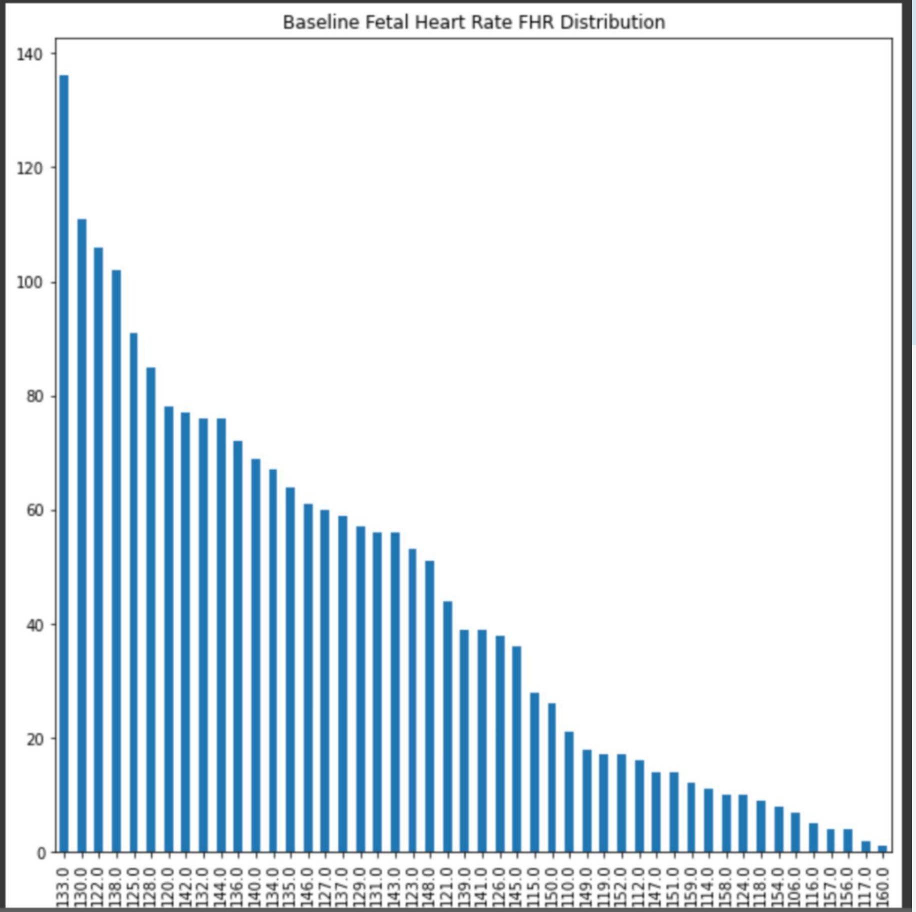
```
[18] df['fetal_health'].value_counts().plot(kind='bar');
     plt.title('Fetal Health Distribution')

Text(0.5, 1.0, 'Fetal Health Distribution')
```



Visualizing the Baseline Fetal Heart Rate bpm Distribution:

```
[19] df['baseline value'].value_counts().plot(kind='bar');
     plt.title('Baseline Fetal Heart Rate FHR Distribution')
     #Baseline Fetal Heart Rate (FHR)
```

```
Text(0.5, 1.0, 'Baseline Fetal Heart Rate FHR Distribution')
```

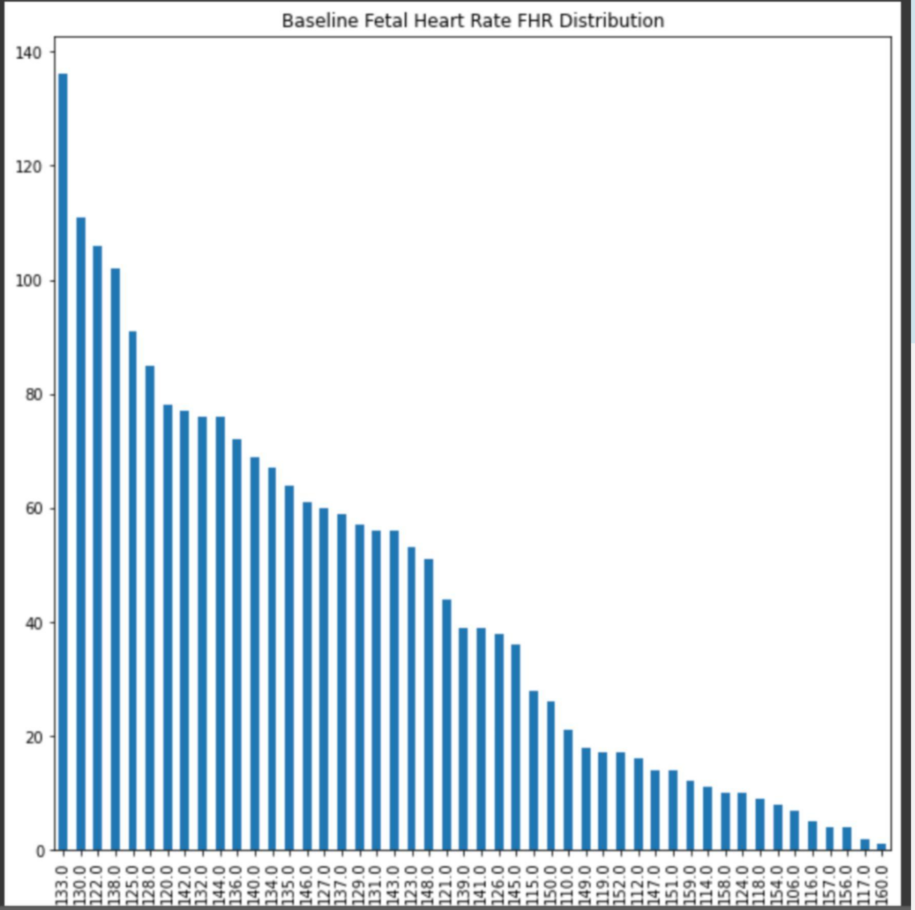


Visualizing the Baseline Fetal Heart Rate bpm Distribution:

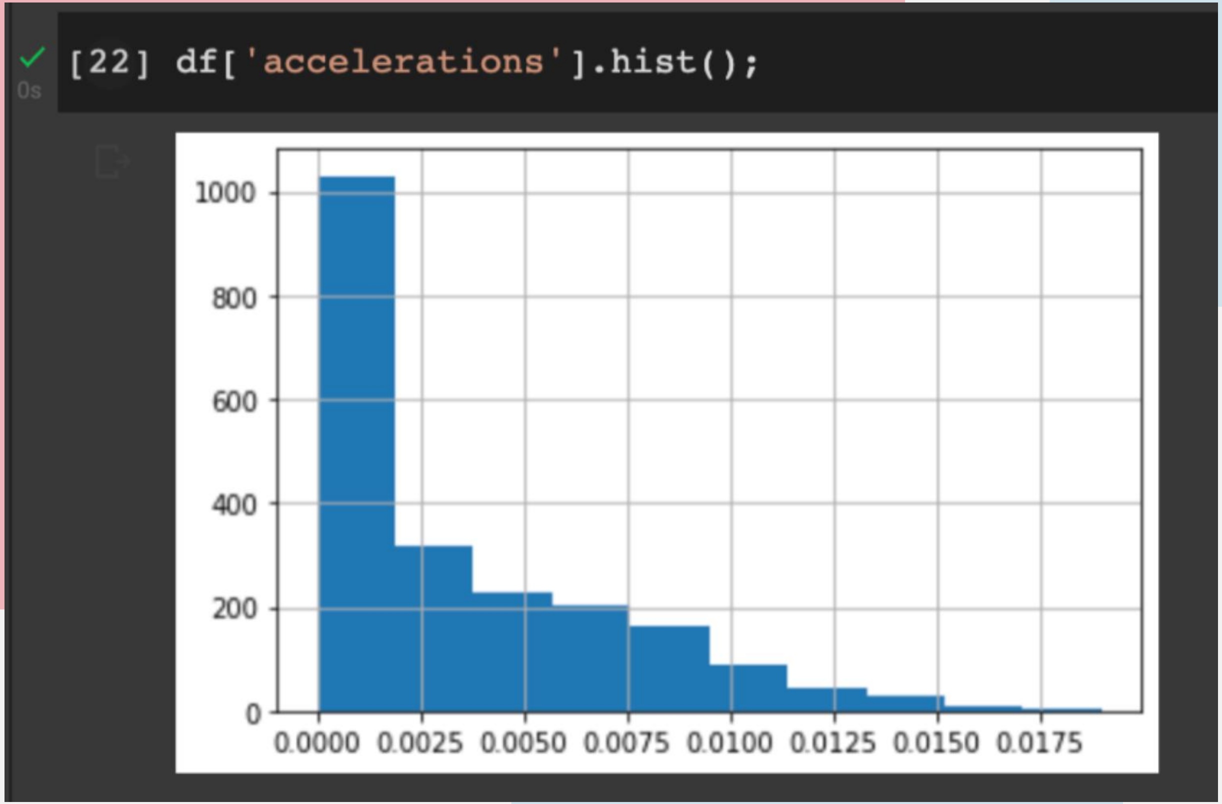
[19]

```
df['baseline value'].value_counts().plot(kind='bar');
plt.title('Baseline Fetal Heart Rate FHR Distribution')
#Baseline Fetal Heart Rate (FHR)
```

```
Text(0.5, 1.0, 'Baseline Fetal Heart Rate FHR Distribution')
```

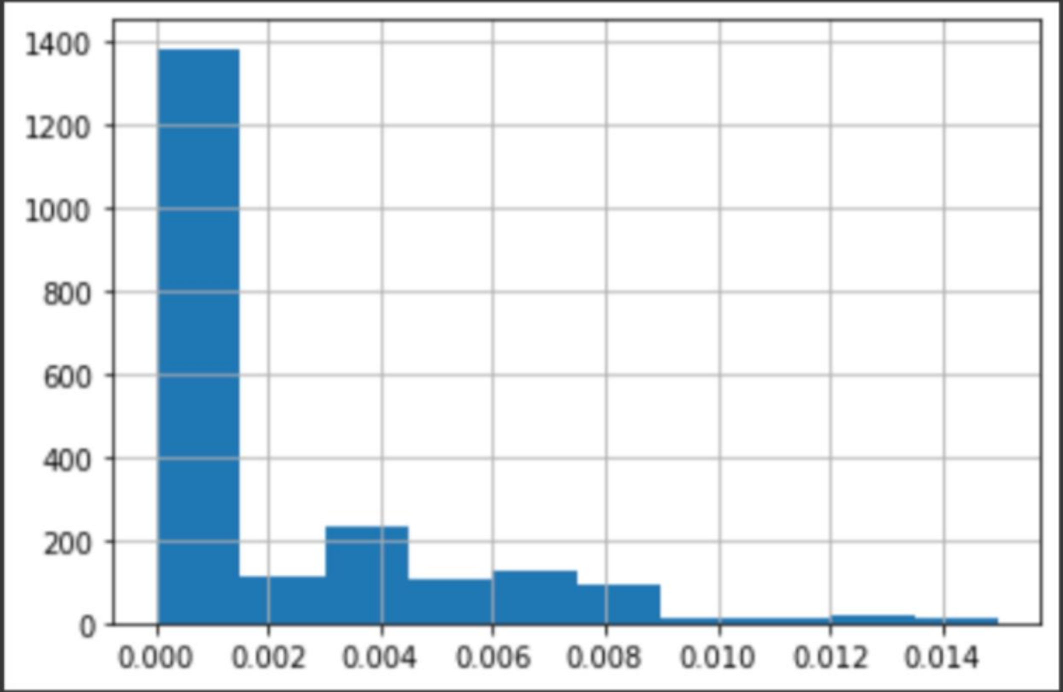


Distribution of FHR accelerations per second:

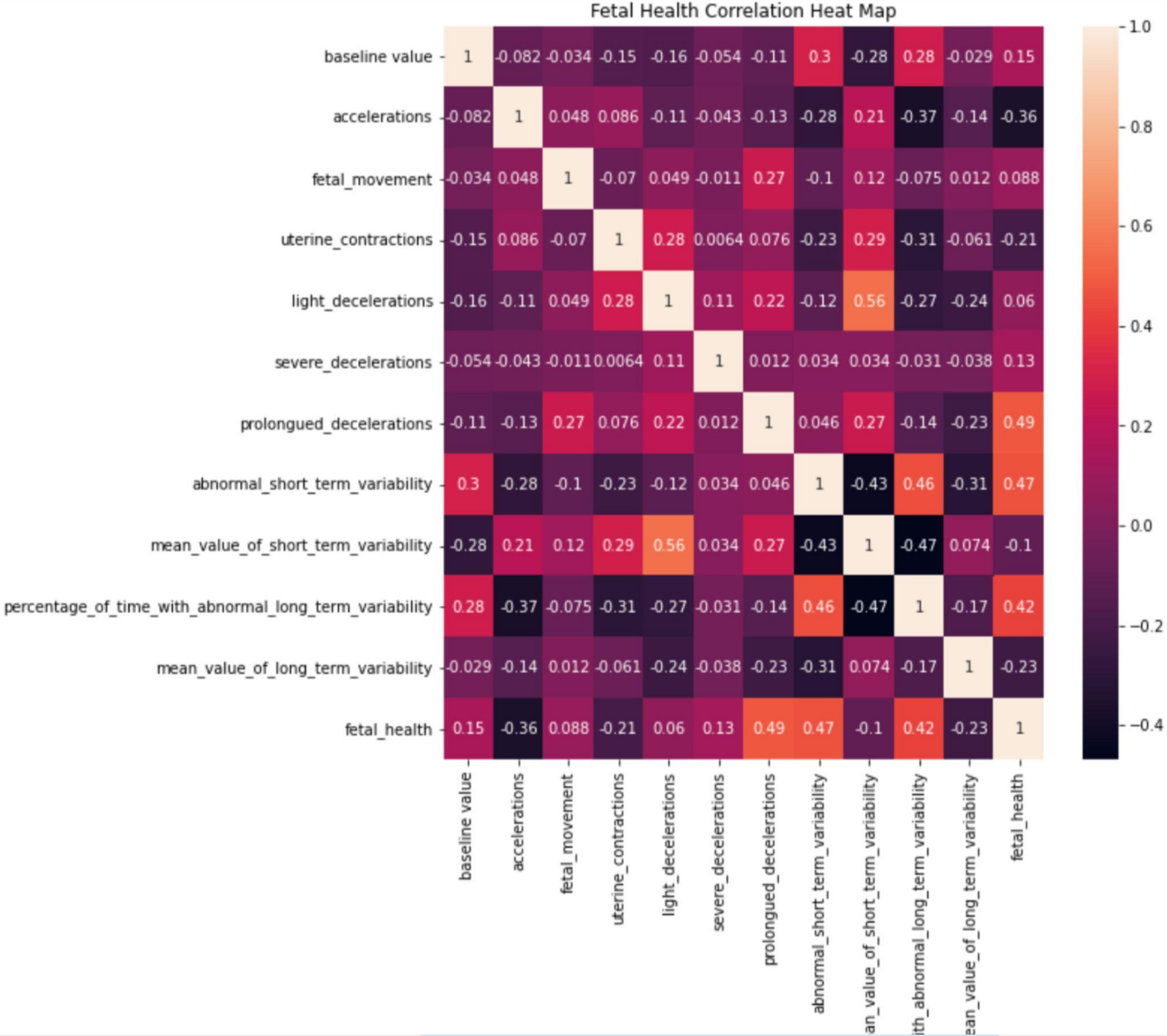


Distribution of FHR decelerations per second:

```
[28] df['light_decelerations'].hist();
```



Visualizing correlations of Fetal Health Data and emerging Fetal Health Predictors



Possible next steps..

01

Supervised Machine Learning

Fetal Health as target vector

02

Unsupervised Machine Learning

Cluster Fetal Health Indicators to identify essential health data for outcome prediction

03

Deliver results

Analyse machine learning models and deliver results to providers

Thank you!

Email questions:

eva.vukich@gmail.com