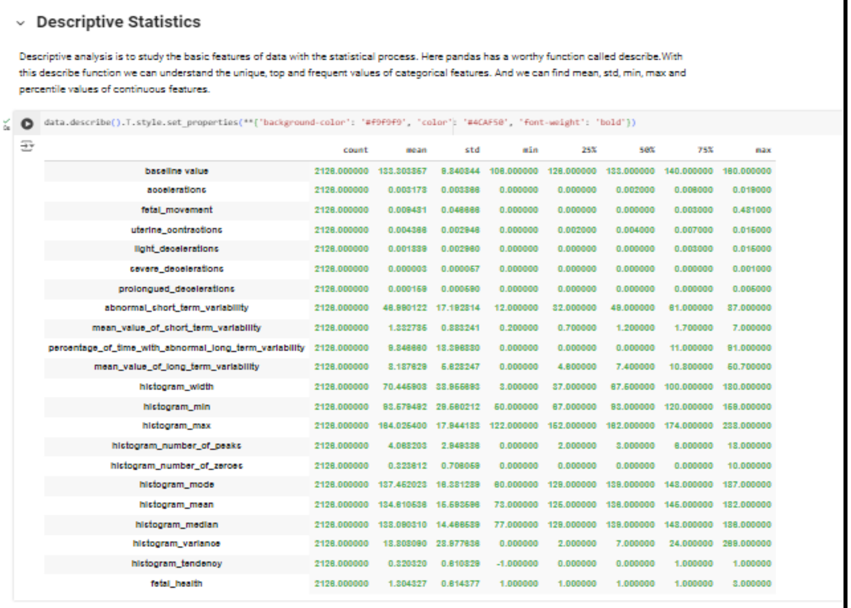
Data Collection and Preprocessing Phase

Date	10 JULY 2024
Team ID	FACULTY
Project Title	Fetal AI: Using Machine Learning To Predict And Monitor Fetal Health.
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

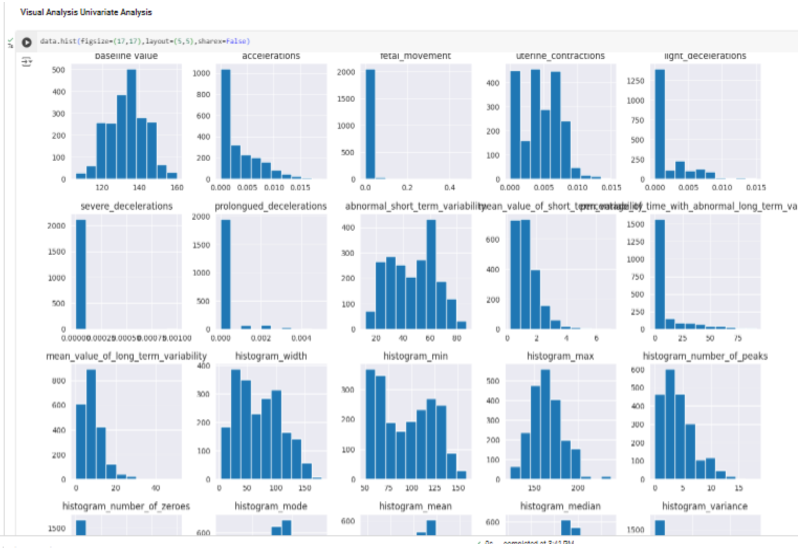
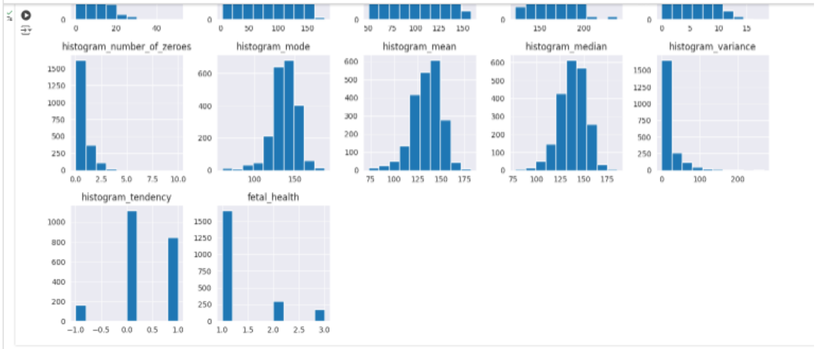
Section	**Description**
Data Overview	Dimension: 2126 rows × 22columns Descriptive statistics: 

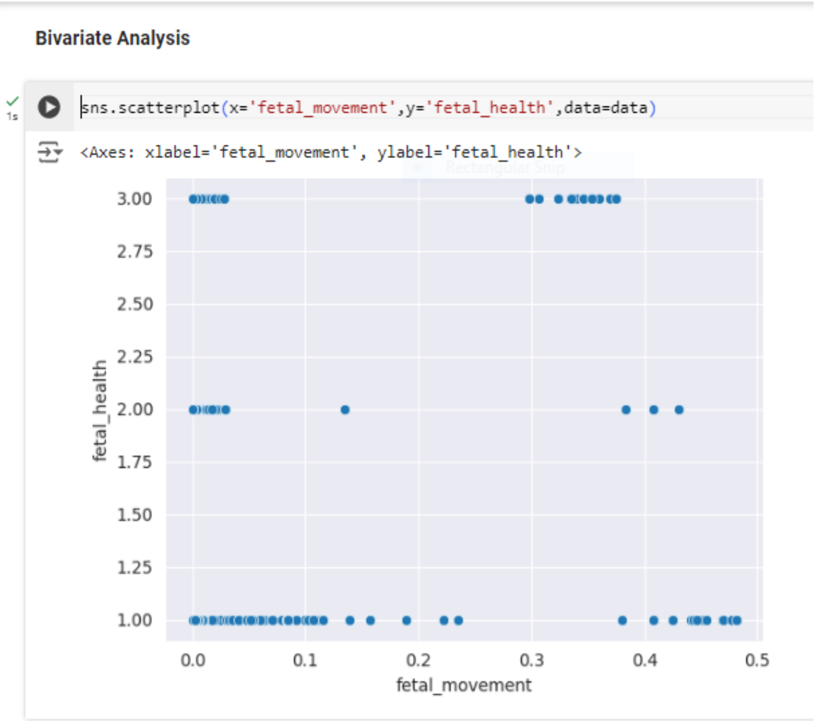
Descriptive Statistics

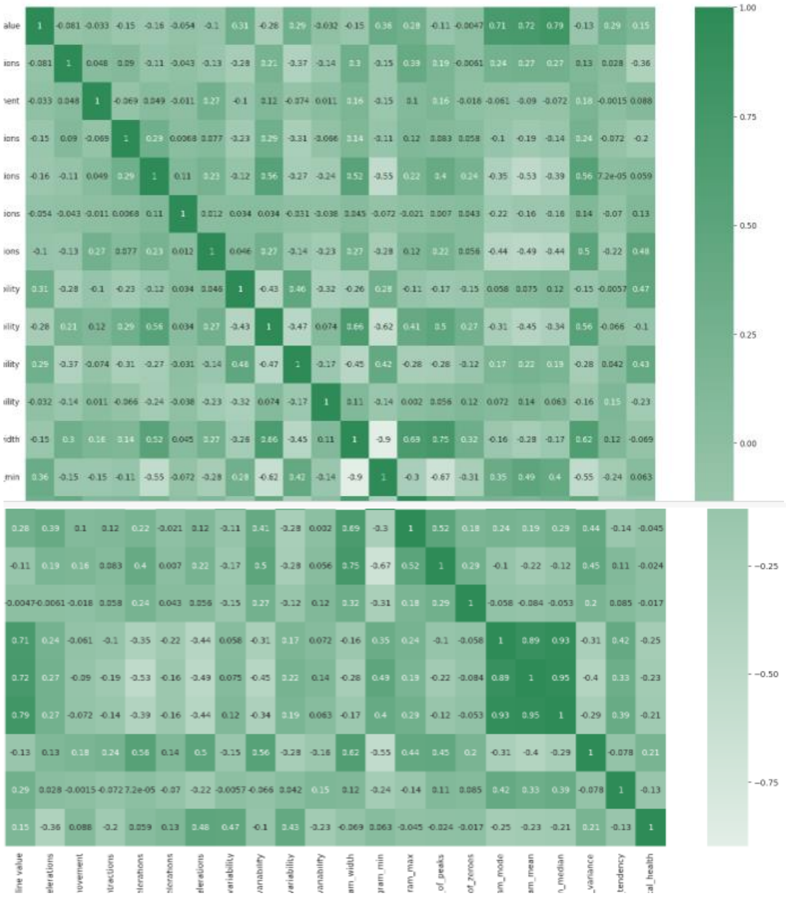
Descriptive analysis is to study the basic features of data with the statistical process. Here pandas has a worthy function called describe.With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.

```
data.describe().T.style.set_properties(**{'background-color': '#f9f9f9', 'color': '#4CAF50', 'font-weight': 'bold'})
```

	count	mean	std	min	25%	50%	75%	max
baseline value	2126.000000	133.303857	9.840844	106.000000	126.000000	133.000000	140.000000	160.000000
accelerations	2126.000000	0.003178	0.003866	0.000000	0.000000	0.002000	0.006000	0.019000
fetal_movement	2126.000000	0.009481	0.046666	0.000000	0.000000	0.000000	0.003000	0.481000
uterine_contractions	2126.000000	0.004366	0.002946	0.000000	0.002000	0.004000	0.007000	0.015000
light_decelerations	2126.000000	0.001889	0.002960	0.000000	0.000000	0.000000	0.003000	0.015000
severe_decelerations	2126.000000	0.000003	0.000057	0.000000	0.000000	0.000000	0.000000	0.001000
prolongued_decelerations	2126.000000	0.000159	0.000590	0.000000	0.000000	0.000000	0.000000	0.005000
abnormal_short_term_variability	2126.000000	46.990122	17.192814	12.000000	32.000000	49.000000	61.000000	87.000000
mean_value_of_short_term_variability	2126.000000	1.332785	0.883241	0.200000	0.700000	1.200000	1.700000	7.000000
percentage_of_time_with_abnormal_long_term_variability	2126.000000	9.846660	18.396880	0.000000	0.000000	0.000000	11.000000	91.000000
mean_value_of_long_term_variability	2126.000000	8.187629	5.628247	0.000000	4.600000	7.400000	10.800000	50.700000
histogram_width	2126.000000	70.445908	38.955693	3.000000	37.000000	67.500000	100.000000	180.000000
histogram_min	2126.000000	93.579492	29.560212	50.000000	67.000000	93.000000	120.000000	159.000000
histogram_max	2126.000000	164.025400	17.944183	122.000000	152.000000	162.000000	174.000000	238.000000
histogram_number_of_peaks	2126.000000	4.068203	2.949386	0.000000	2.000000	3.000000	6.000000	18.000000
histogram_number_of_zeroes	2126.000000	0.323612	0.706059	0.000000	0.000000	0.000000	0.000000	10.000000
histogram_mode	2126.000000	137.452023	16.381289	60.000000	129.000000	139.000000	148.000000	187.000000
histogram_mean	2126.000000	134.610536	15.593596	73.000000	125.000000	136.000000	145.000000	182.000000
histogram_median	2126.000000	138.090310	14.466589	77.000000	129.000000	139.000000	148.000000	186.000000
histogram_variance	2126.000000	18.808090	28.977636	0.000000	2.000000	7.000000	24.000000	269.000000
histogram_tendency	2126.000000	0.320320	0.610829	-1.000000	0.000000	0.000000	1.000000	1.000000
fetal_health	2126.000000	1.304327	0.614377	1.000000	1.000000	1.000000	1.000000	3.000000

Univariate Analysis	 Visual Analysis Univariate Analysis `data.hist(figsize=(17,17),layout=(5,5),sharex=False)` baseline value accelerations fetal_movement uterine_contractions light_decelerations severe_decelerations prolongued_decelerations abnormal_short_term_variability mean_value_of_short_term_variability percentage_of_time_with_abnormal_long_term_va mean_value_of_long_term_variability histogram_width histogram_min histogram_max histogram_number_of_peaks histogram_number_of_zeroes histogram_mode histogram_mean histogram_median histogram_variance  histogram_number_of_zeroes histogram_mode histogram_mean histogram_median histogram_variance histogram_tendency fetal_health

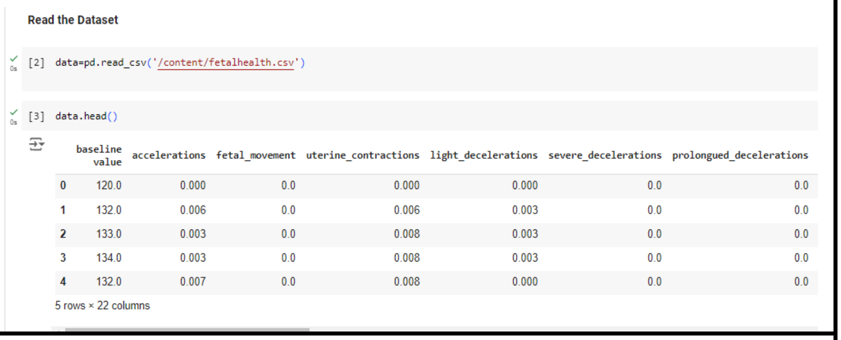
Bivariate Analysis	 Bivariate Analysis
	`sns.scatterplot(x='fetal_movement',y='fetal_health',data=data)`
	`<Axes: xlabel='fetal_movement', ylabel='fetal_health'>`
	 3.00
	2.75
	2.50
	2.25
	2.00
	1.75
	1.50
	1.25
	1.00
	fetal_health
	0.0
	0.1
	0.2
	0.3
	0.4
	0.5
	fetal_movement

Multivariate Analysis	

Outliers and Anomalies	--

Data Preprocessing Code Screenshots

Loading Data



Read the Dataset

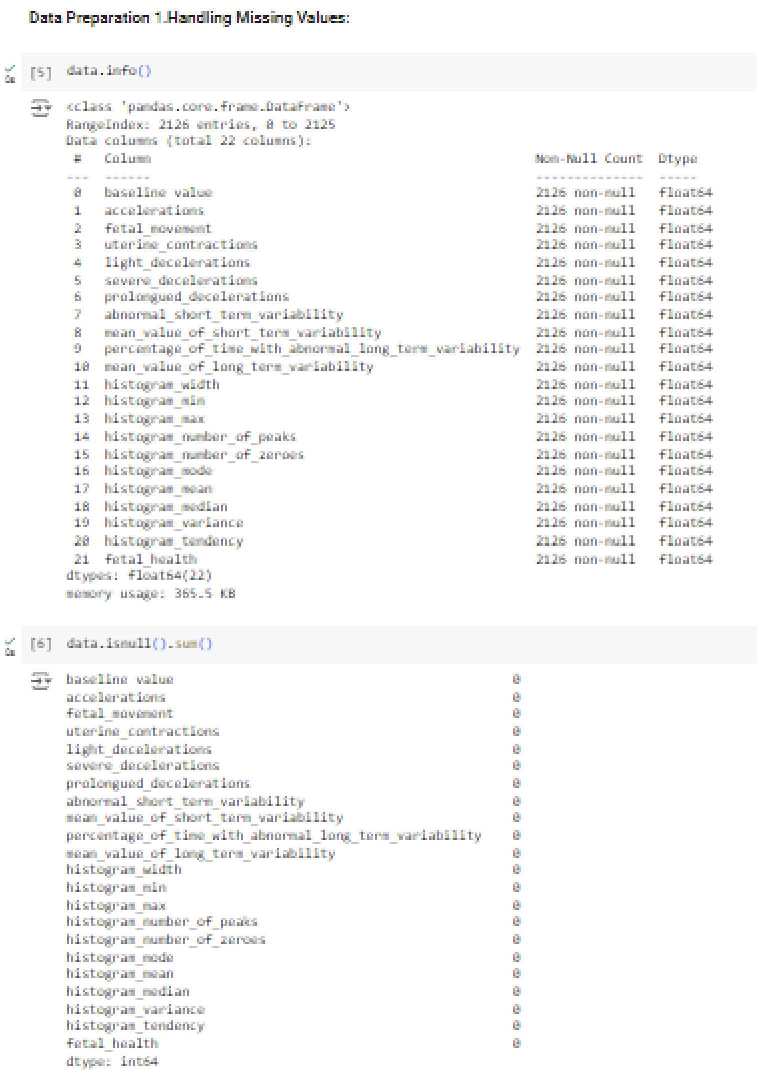
```
[2] data=pd.read_csv('/content/fetalhealth.csv')
```

```
[3] data.head()
```

	baseline value	accelerations	fetal_movement	uterine_contractions	light_decelerations	severe_decelerations	prolongued_decelerations
0	120.0	0.000	0.0	0.000	0.000	0.0	0.0
1	132.0	0.006	0.0	0.006	0.003	0.0	0.0
2	133.0	0.003	0.0	0.008	0.003	0.0	0.0
3	134.0	0.003	0.0	0.008	0.003	0.0	0.0
4	132.0	0.007	0.0	0.008	0.000	0.0	0.0

5 rows × 22 columns

Handling Missing Data



Data Preparation 1.Handling Missing Values:

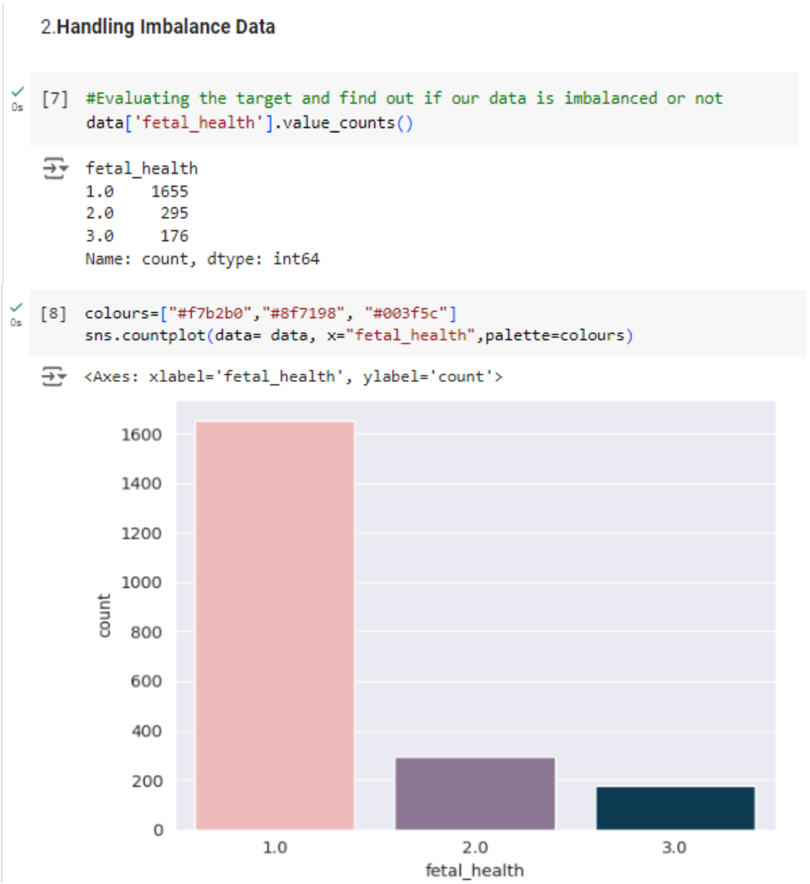
```
[5] data.info()
```

```
<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
```

```
[6] data.isnull().sum()
```

```
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
histogram_width                                           0
histogram_min                                             0
histogram_max                                             0
histogram_number_of_peaks                                 0
histogram_number_of_zeroes                                0
histogram_mode                                            0
histogram_mean                                            0
histogram_median                                          0
histogram_variance                                        0
histogram_tendency                                        0
fetal_health                                              0
dtype: int64
```

Handling Imbalance Data



2.Handling Imbalance Data

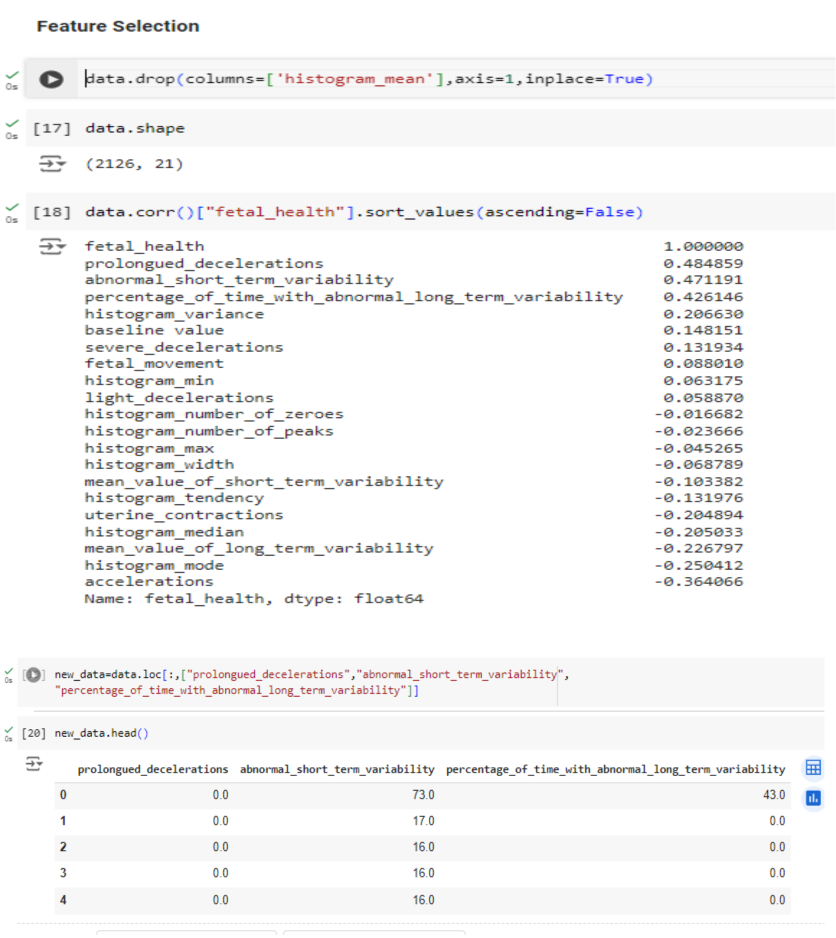
```
[7] #Evaluating the target and find out if our data is imbalanced or not
    data['fetal_health'].value_counts()
```

```
fetal_health
1.0    1655
2.0     295
3.0     176
Name: count, dtype: int64
```

```
[8] colours=["#f7b2b0","#8f7198", "#003f5c"]
    sns.countplot(data= data, x="fetal_health",palette=colours)
```

```
<Axes: xlabel='fetal_health', ylabel='count'>
```

Feature Engineering



Feature Selection

```
data.drop(columns=['histogram_mean'],axis=1,inplace=True)
```

```
[17] data.shape
```

```
(2126, 21)
```

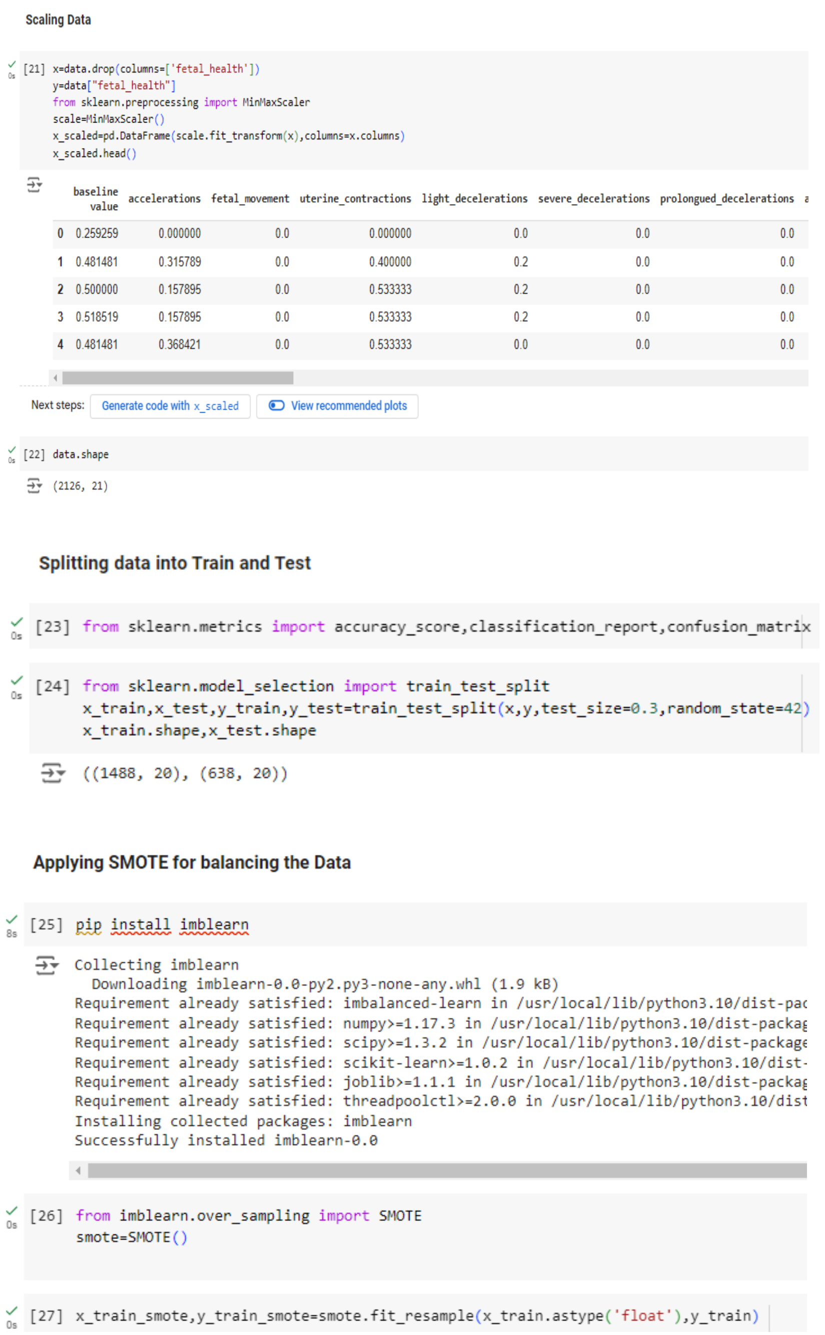
```
[18] data.corr()["fetal_health"].sort_values(ascending=False)
```

```
fetal_health                                              1.000000
prolongued_decelerations                                  0.484859
abnormal_short_term_variability                           0.471191
percentage_of_time_with_abnormal_long_term_variability    0.426146
histogram_variance                                        0.206630
baseline value                                            0.148151
severe_decelerations                                      0.131934
fetal_movement                                            0.088010
histogram_min                                             0.063175
light_decelerations                                       0.058870
histogram_number_of_zeroes                               -0.016682
histogram_number_of_peaks                                -0.023666
histogram_max                                            -0.045265
histogram_width                                          -0.068789
mean_value_of_short_term_variability                     -0.103382
histogram_tendency                                       -0.131976
uterine_contractions                                     -0.204894
histogram_median                                         -0.205033
mean_value_of_long_term_variability                      -0.226797
histogram_mode                                           -0.250412
accelerations                                            -0.364066
Name: fetal_health, dtype: float64
```

```
new_data=data.loc[:,["prolongued_decelerations","abnormal_short_term_variability",
"percentage_of_time_with_abnormal_long_term_variability"]]
```

```
[20] new_data.head()
```

	prolongued_decelerations	abnormal_short_term_variability	percentage_of_time_with_abnormal_long_term_variability
0	0.0	73.0	43.0
1	0.0	17.0	0.0
2	0.0	16.0	0.0
3	0.0	16.0	0.0
4	0.0	16.0	0.0



Scaling Data

```
[21] x=data.drop(columns=['fetal_health'])
     y=data["fetal_health"]
     from sklearn.preprocessing import MinMaxScaler
     scale=MinMaxScaler()
     x_scaled=pd.DataFrame(scale.fit_transform(x),columns=x.columns)
     x_scaled.head()
```

	baseline value	accelerations	fetal_movement	uterine_contractions	light_decelerations	severe_decelerations	prolongued_decelerations	a
0	0.259259	0.000000	0.0	0.000000	0.0	0.0	0.0	
1	0.481481	0.315789	0.0	0.400000	0.2	0.0	0.0	
2	0.500000	0.157895	0.0	0.533333	0.2	0.0	0.0	
3	0.518519	0.157895	0.0	0.533333	0.2	0.0	0.0	
4	0.481481	0.368421	0.0	0.533333	0.0	0.0	0.0	

```
[22] data.shape
```

```
(2126, 21)
```

Splitting data into Train and Test

```
[23] from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
```

```
[24] from sklearn.model_selection import train_test_split
     x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=42)
     x_train.shape,x_test.shape
```

```
((1488, 20), (638, 20))
```

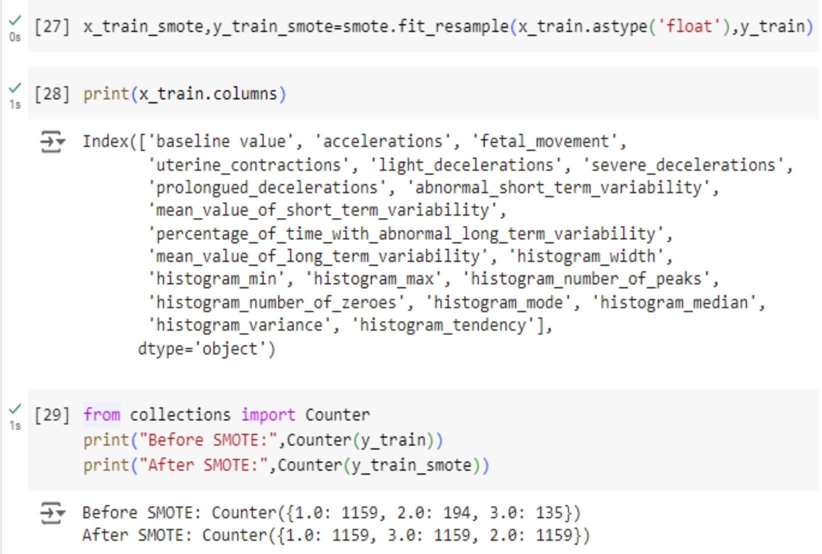
Applying SMOTE for balancing the Data

```
[25] pip install imblearn
```

```
Collecting imblearn
  Downloading imblearn-0.0-py2.py3-none-any.whl (1.9 kB)
Requirement already satisfied: imbalanced-learn in /usr/local/lib/python3.10/dist-pac
Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: scipy>=1.3.2 in /usr/local/lib/python3.10/dist-package
Requirement already satisfied: scikit-learn>=1.0.2 in /usr/local/lib/python3.10/dist-
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist
Installing collected packages: imblearn
Successfully installed imblearn-0.0
```

```
[26] from imblearn.over_sampling import SMOTE
     smote=SMOTE()
```

```
[27] x_train_smote,y_train_smote=smote.fit_resample(x_train.astype('float'),y_train)
```



```
[27] x_train_smote,y_train_smote=smote.fit_resample(x_train.astype('float'),y_train)
```

```
[28] print(x_train.columns)
```

```
Index(['baseline value', 'accelerations', 'fetal_movement',
       'uterine_contractions', 'light_decelerations', 'severe_decelerations',
       'prolongued_decelerations', 'abnormal_short_term_variability',
       'mean_value_of_short_term_variability',
       'percentage_of_time_with_abnormal_long_term_variability',
       'mean_value_of_long_term_variability', 'histogram_width',
       'histogram_min', 'histogram_max', 'histogram_number_of_peaks',
       'histogram_number_of_zeroes', 'histogram_mode', 'histogram_median',
       'histogram_variance', 'histogram_tendency'],
      dtype='object')
```

```
[29] from collections import Counter
     print("Before SMOTE:",Counter(y_train))
     print("After SMOTE:",Counter(y_train_smote))
```

```
Before SMOTE: Counter({1.0: 1159, 2.0: 194, 3.0: 135})
After SMOTE: Counter({1.0: 1159, 3.0: 1159, 2.0: 1159})
```