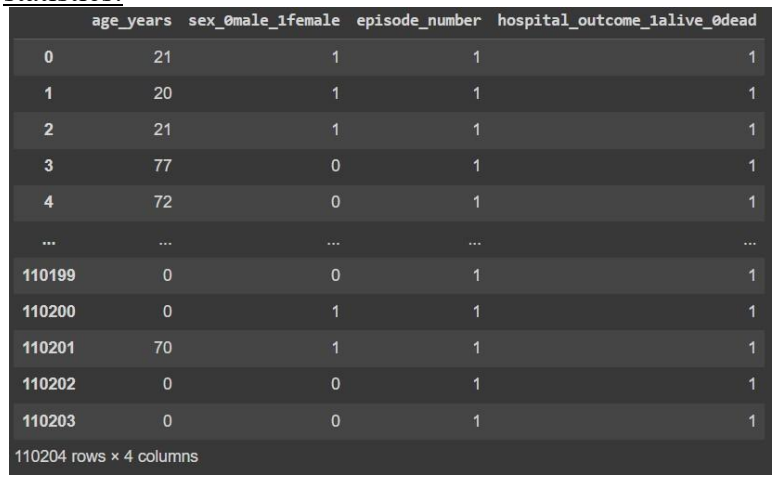
Data Collection and Preprocessing Phase

Date	15 July 2024
Team ID	740116
Project Title	Sepsis Survival Minimal Clinical Records
Maximum Marks	6 Marks

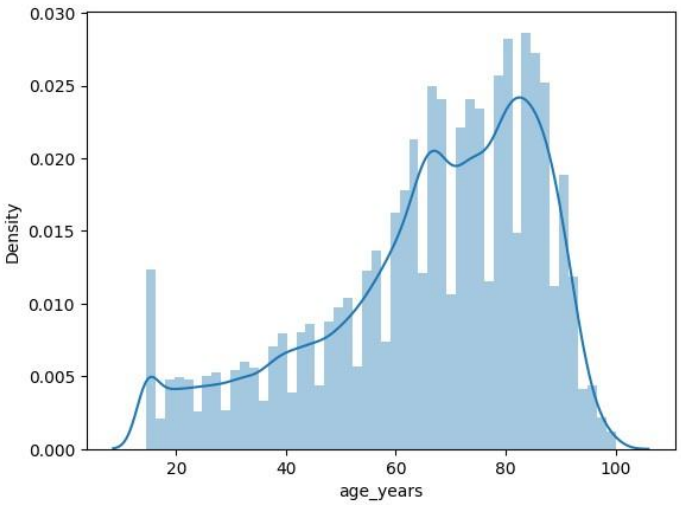
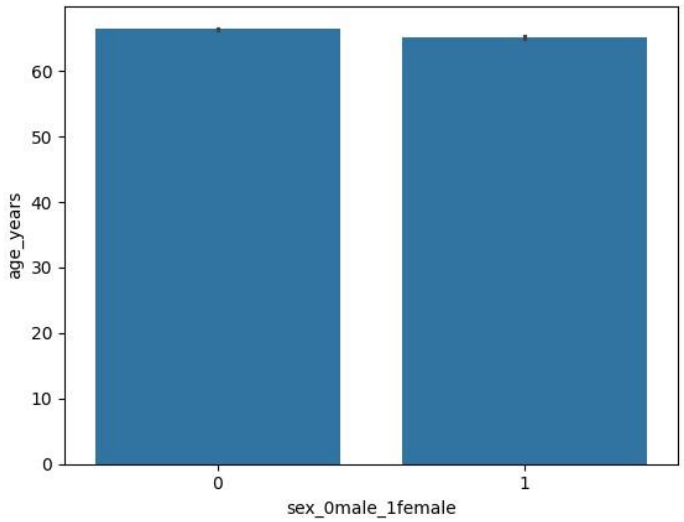
Data Exploration and Preprocessing Report

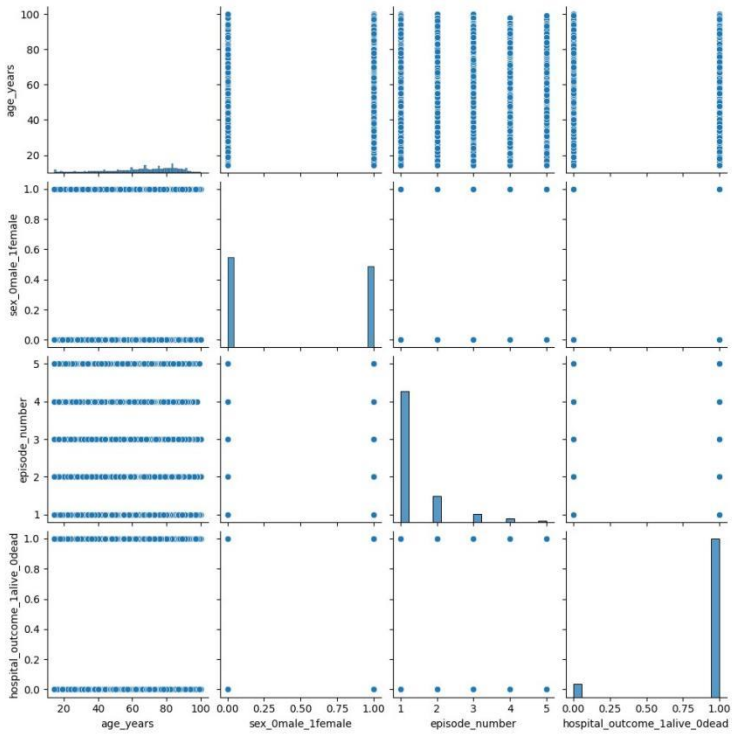
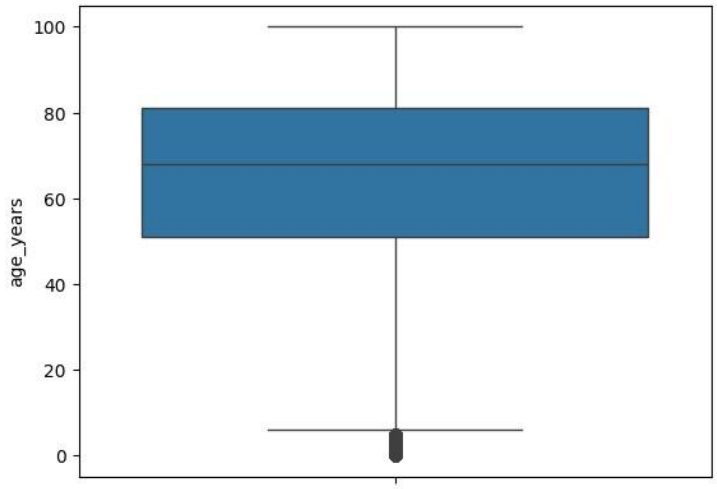
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong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

Section	Description
Data Overview	Dimension: 110204 rows × 4 columns Descriptive statistics: 
Univariate Analysis	Analysis of individual variables to understand their distribution and identify any anomalies.

	age_years	sex_0male_1female	episode_number	hospital_outcome_1alive_0dead
0	21	1	1	1
1	20	1	1	1
2	21	1	1	1
3	77	0	1	1
4	72	0	1	1
...	...	...	...	...
110199	0	0	1	1
110200	0	1	1	1
110201	70	1	1	1
110202	0	0	1	1
110203	0	0	1	1

110204 rows × 4 columns

	
Bivariate Analysis	

Multivariate Analysis	
Outliers and Anomalies	
Data Preprocessing Code Screenshots	

<table>
<tr><td>Loading Data</td><td>

```
[4] data = pd.read_csv(r'/content/s41598-020-73558-3_sepsis_survival_primary_cohort.csv')

[5] data.head()
```

	age_years	sex_0male_1female	episode_number	hospital_outcome_1alive_0dead
0	21	1	1	1
1	20	1	1	1
2	21	1	1	1
3	77	0	1	1
4	72	0	1	1

</td></tr>
<tr><td>Data Transformation</td><td>

```
data.drop(['case_admission_id','ethnicity','hospital_admission_id','icu_admission_id'],axis=1,inplace=True)
data['sex'].replace({'M':1,'F':0},inplace=True)
data['hospital_discharge_status'].replace({'Alive':1,'Died':0},inplace=True)
age_group = data.groupby('age_group')
mean_age = age_group['age_years'].mean()
data['age_years'].fillna(data['age_group'].map(mean_age), inplace=True)
data['age_years'] = data['age_years'].astype(int)
data.drop(['age_group'],axis=1,inplace=True)
data = data.dropna()
```

```
X = data.drop('hospital_discharge_status', axis=1)
y = data['hospital_discharge_status']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

</td></tr>
<tr><td>Feature Engineering</td><td>Attached the codes in final submission.</td></tr>
<tr><td>Save Processed Data</td><td>-</td></tr>
</table>