

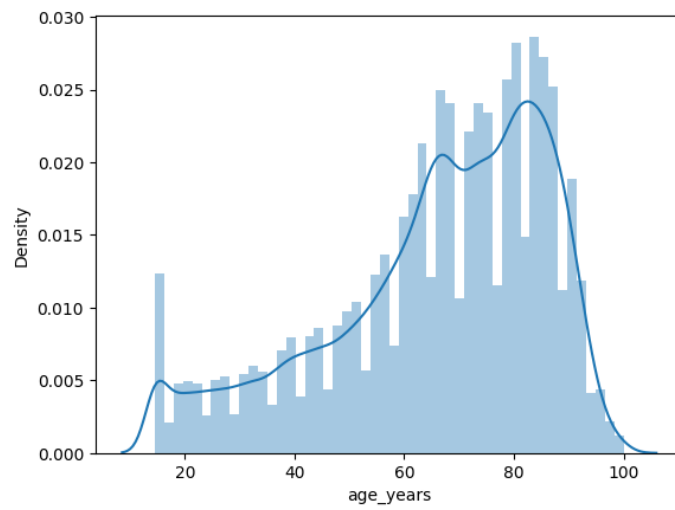
## Data Collection and Preprocessing Phase

Date	09 July 2024
Team ID	739801
Project Title	Sepsis Survival Minimal Clinical Records
Maximum Marks	6 Marks

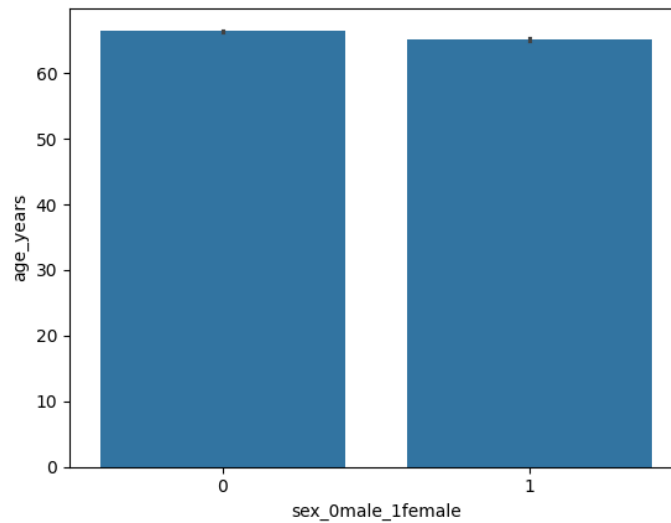
### Data Exploration and Preprocessing Report

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.

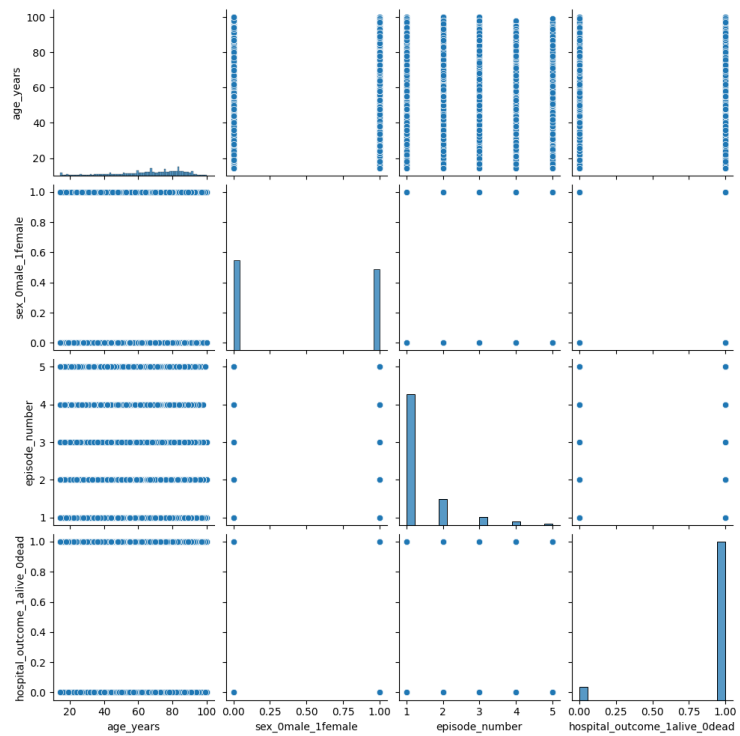
Section	Description																																																												
Data Overview	<p><u>Dimension:</u> 110204 rows × 4 columns</p> <p><u>Descriptive statistics:</u></p> <table><thead><tr><th></th><th>age_years</th><th>sex_0male_1female</th><th>episode_number</th><th>hospital_outcome_1alive_0dead</th></tr></thead><tbody><tr><td>0</td><td>21</td><td>1</td><td>1</td><td>1</td></tr><tr><td>1</td><td>20</td><td>1</td><td>1</td><td>1</td></tr><tr><td>2</td><td>21</td><td>1</td><td>1</td><td>1</td></tr><tr><td>3</td><td>77</td><td>0</td><td>1</td><td>1</td></tr><tr><td>4</td><td>72</td><td>0</td><td>1</td><td>1</td></tr><tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr><tr><td>110199</td><td>0</td><td>0</td><td>1</td><td>1</td></tr><tr><td>110200</td><td>0</td><td>1</td><td>1</td><td>1</td></tr><tr><td>110201</td><td>70</td><td>1</td><td>1</td><td>1</td></tr><tr><td>110202</td><td>0</td><td>0</td><td>1</td><td>1</td></tr><tr><td>110203</td><td>0</td><td>0</td><td>1</td><td>1</td></tr></tbody></table> <p>110204 rows × 4 columns</p>		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
		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																																																									
Univariate Analysis	Analysis of individual variables to understand their distribution and identify any anomalies.																																																												



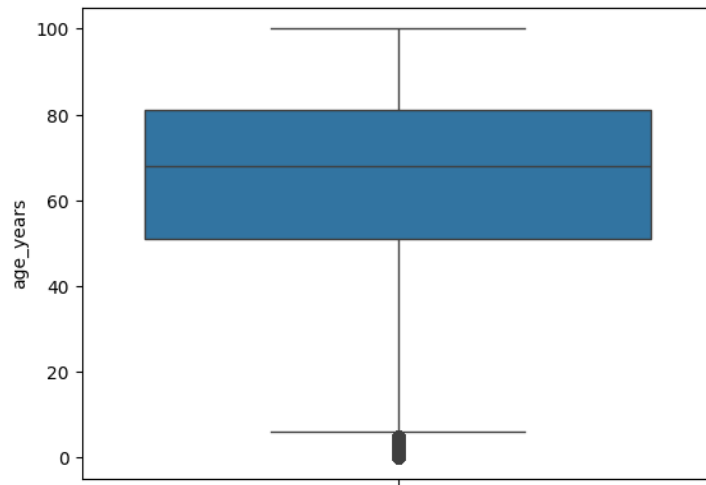
## Bivariate Analysis



## Multivariate Analysis



## Outliers and Anomalies



## Data Preprocessing Code Screenshots

### Loading Data

```
[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

Data Transformation	<pre> 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) </pre>
Feature Engineering	Attached the codes in final submission.
Save Processed Data	-