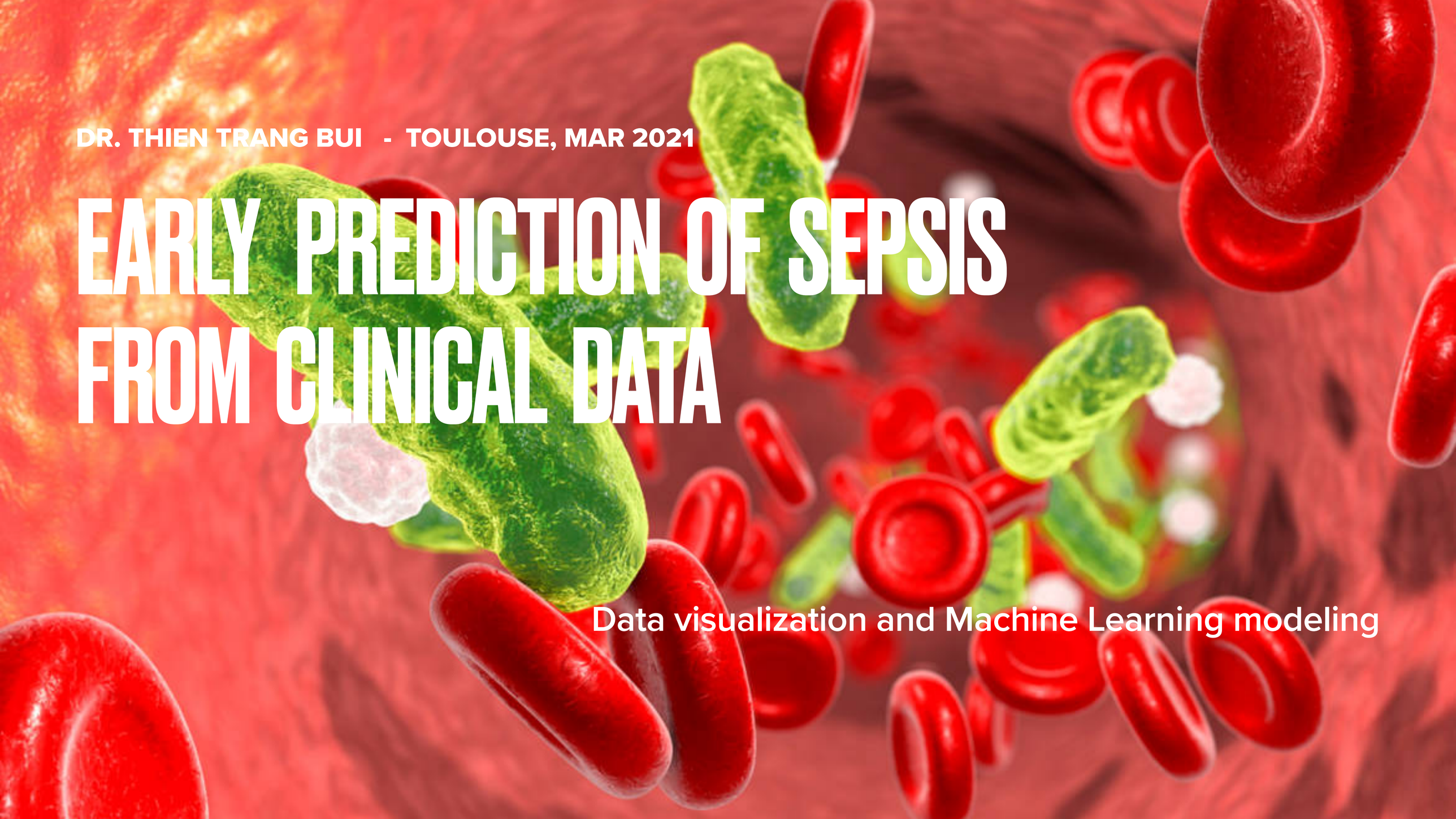


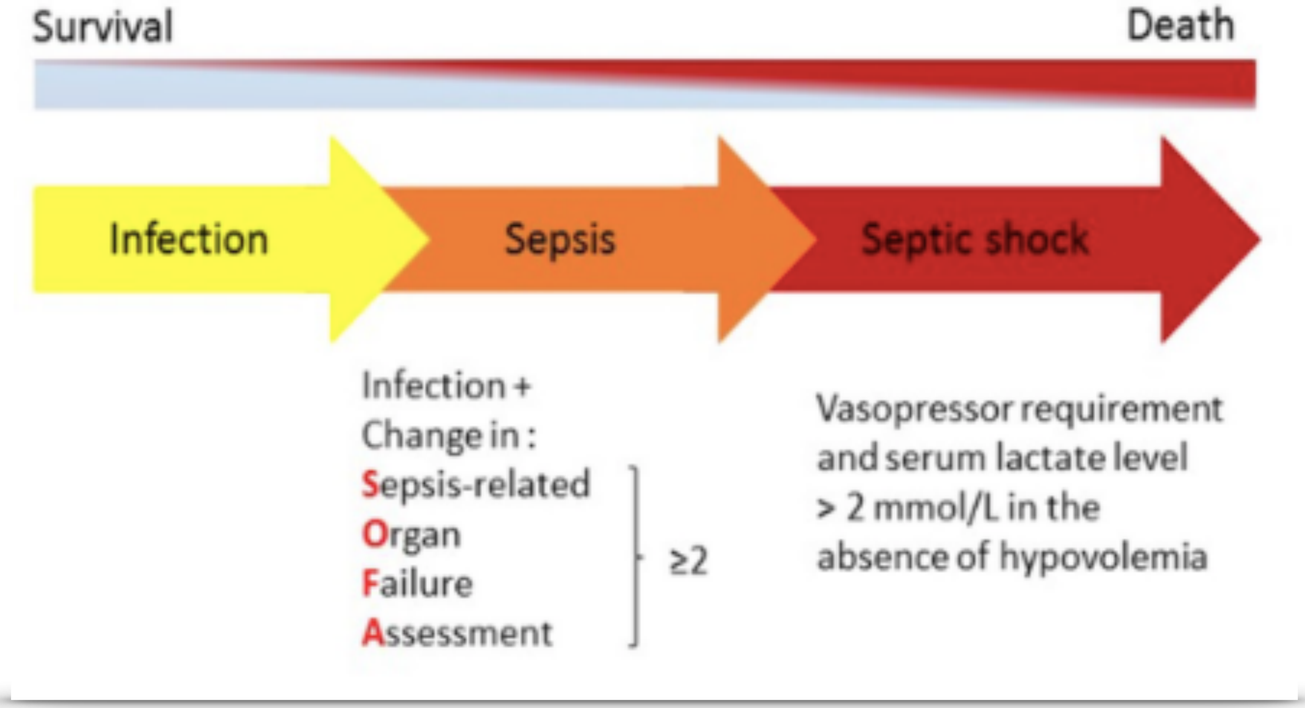
DR. THIEN TRANG BUI - TOULOUSE, MAR 2021

EARLY PREDICTION OF SEPSIS FROM CLINICAL DATA

Data visualization and Machine Learning modeling

WHAT IS SEPSIS?

- **A LIFE-THREATENING ILLNESS CAUSED BY YOUR BODY'S RESPONSE TO AN INFECTION**
- **CAUSES TISSUE DAMAGE, ORGAN FAILURE, OR DEATH**
- **MORE THAN 1.5 MILLION CASES OF SEPSIS EACH YEAR (CDC)**
- **KILLS MORE THAN 250,000 AMERICANS A YEAR.**



DATA AND OBJECTIVES

Objective:

- Predicting sepsis or non-sepsis at each hour of any patients.
- Target feature: SepsisLabel.

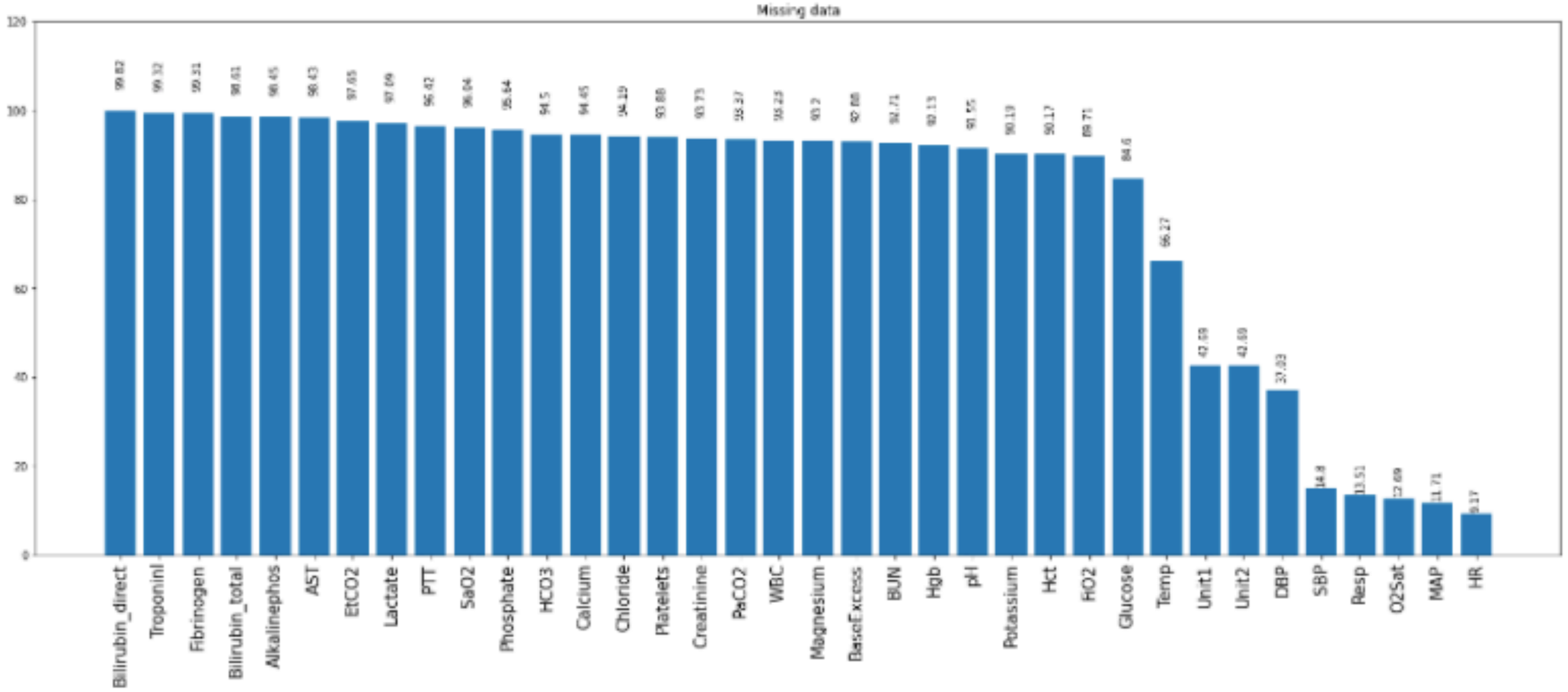
Data: 1170446 observations and 41 features.

- Vital Signs : Heart Rate, Temperature , Blood Pressure, Respiratory rate, etc.
- Laboratory Values : Platelet Count, Glucose , Calcium, etc.
- Demographics : Age, Gender, Time in ICU , Hospital Admit time, Unit1, Unit2.

HCO3	...	WBC	Fibrinogen	Platelets	Age	Gender	Unit1	Unit2	HospAdmTime	ICULOS	SepsisLabel
NaN	...	NaN	NaN	NaN	83.14	0	NaN	NaN	-0.03	1	0
NaN	...	NaN	NaN	NaN	83.14	0	NaN	NaN	-0.03	2	0
NaN	...	NaN	NaN	NaN	83.14	0	NaN	NaN	-0.03	3	0
NaN	...	NaN	NaN	NaN	83.14	0	NaN	NaN	-0.03	4	0
NaN	...	NaN	NaN	NaN	83.14	0	NaN	NaN	-0.03	5	0

DATA PREPARATION - MISSING VALUES

- The proportion of missing data are larger than 90% ===> REMOVE
- REMOVE also FiO2 cause the relative with PaCO2 (removed feature).



- Numerical NAs values: replace by the mean values, corresponding to each group of Sepsis or no-Sepsis.
- Categorical NAs values: replace by “No_info”.

DATA PREPARATION - UNDERSTANDING FEATURES

Numerical features:

- Glucose: Serum glucose (mg/dL)
- **Age:** Years (100 for patients 90 or above)
- HospAdmTime: Hours between hospital admit and ICU admit
- **ICULOS**: ICU length-of-stay (hours since ICU admit)
- **Temp:** Temperature (Deg C)
- **HR:** Heart rate (beats per minute)
- MAP: Mean arterial pressure (mm Hg)
- O2Sat: Pulse oximetry (%)

Categorical features:

- Gender: Female (0) or Male (1)
- Unit1: Administrative identifier for ICU unit (Medical Intensive Care Unit)
- Unit2: Administrative identifier for ICU unit (Surgical Intensive Care Unit)
- SepsisLabel = 1 (Sepsis)/ SepsisLabel = 0 (no sepsis)

Engineering feature:

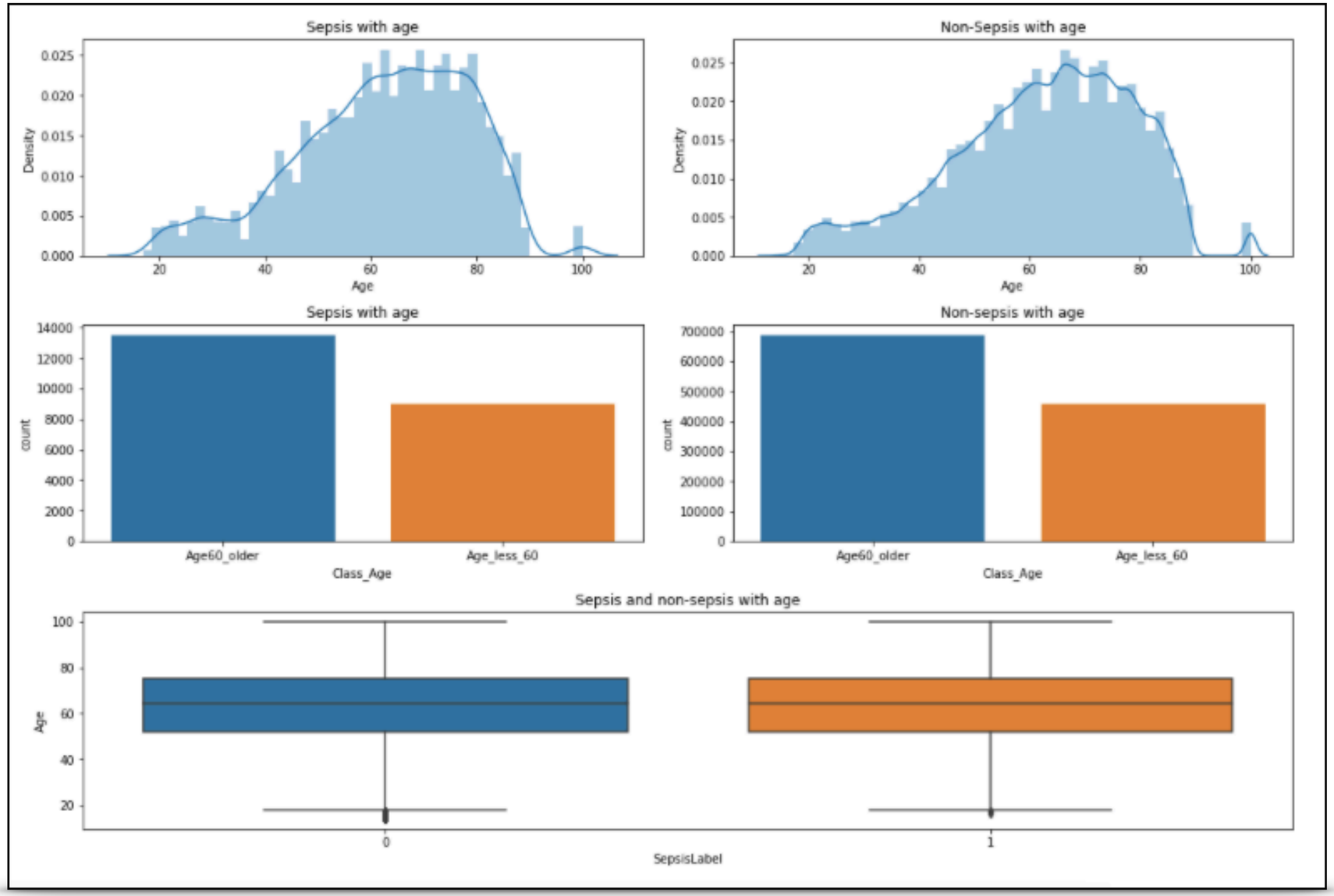
- Age class:
  - Age less than 60

And

  - Age greater than 60

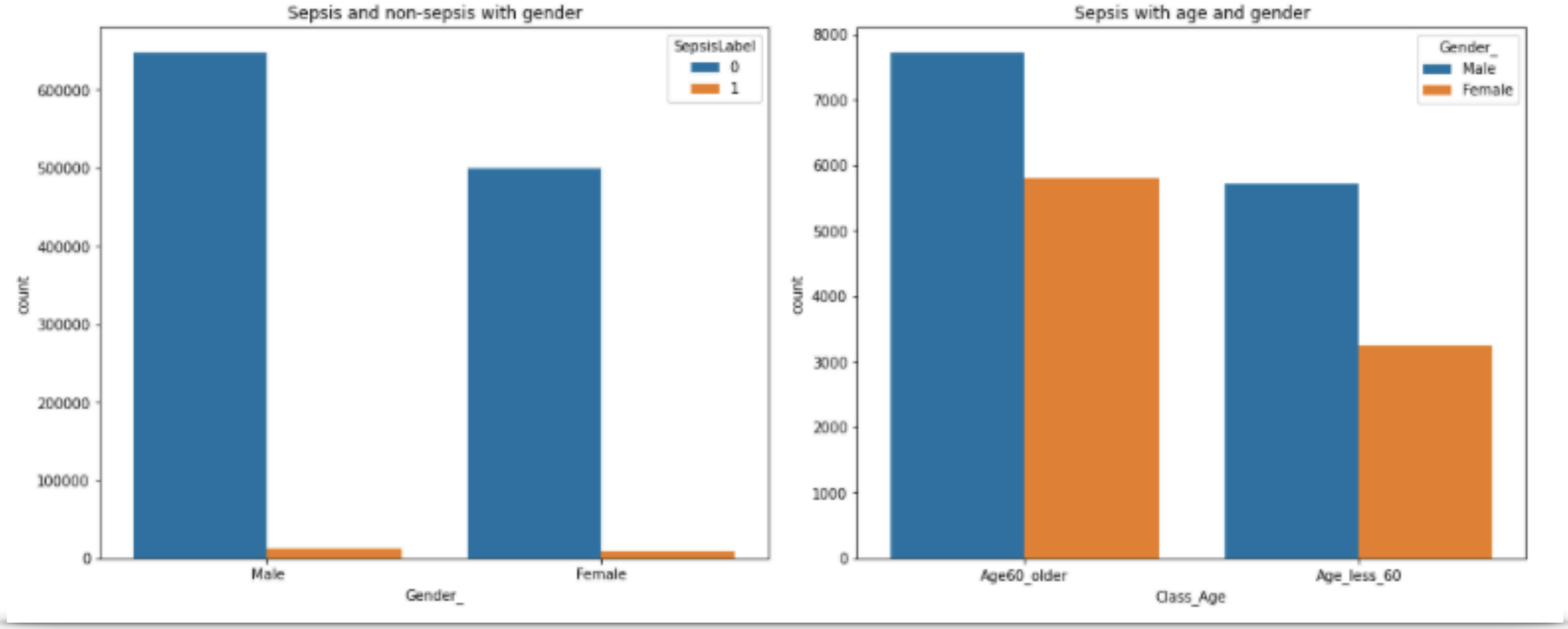
DATA VISUALISATION - SEPSIS AND AGE

* seniors can be at risk for sepsis (immune system weakens as we age)



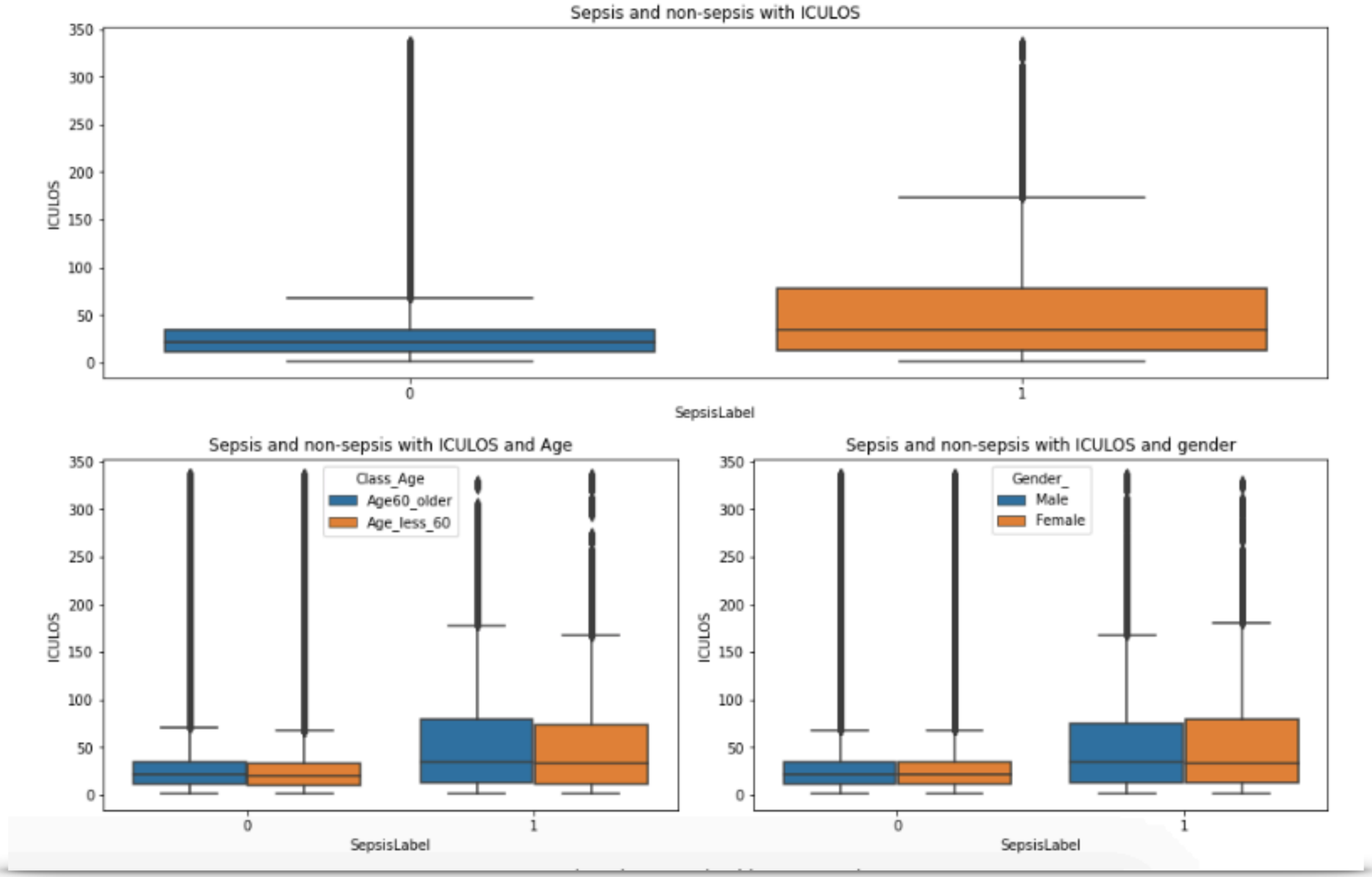
DATA VISUALISATION - SEPSIS AND GENDER

* Male > Female (ability)



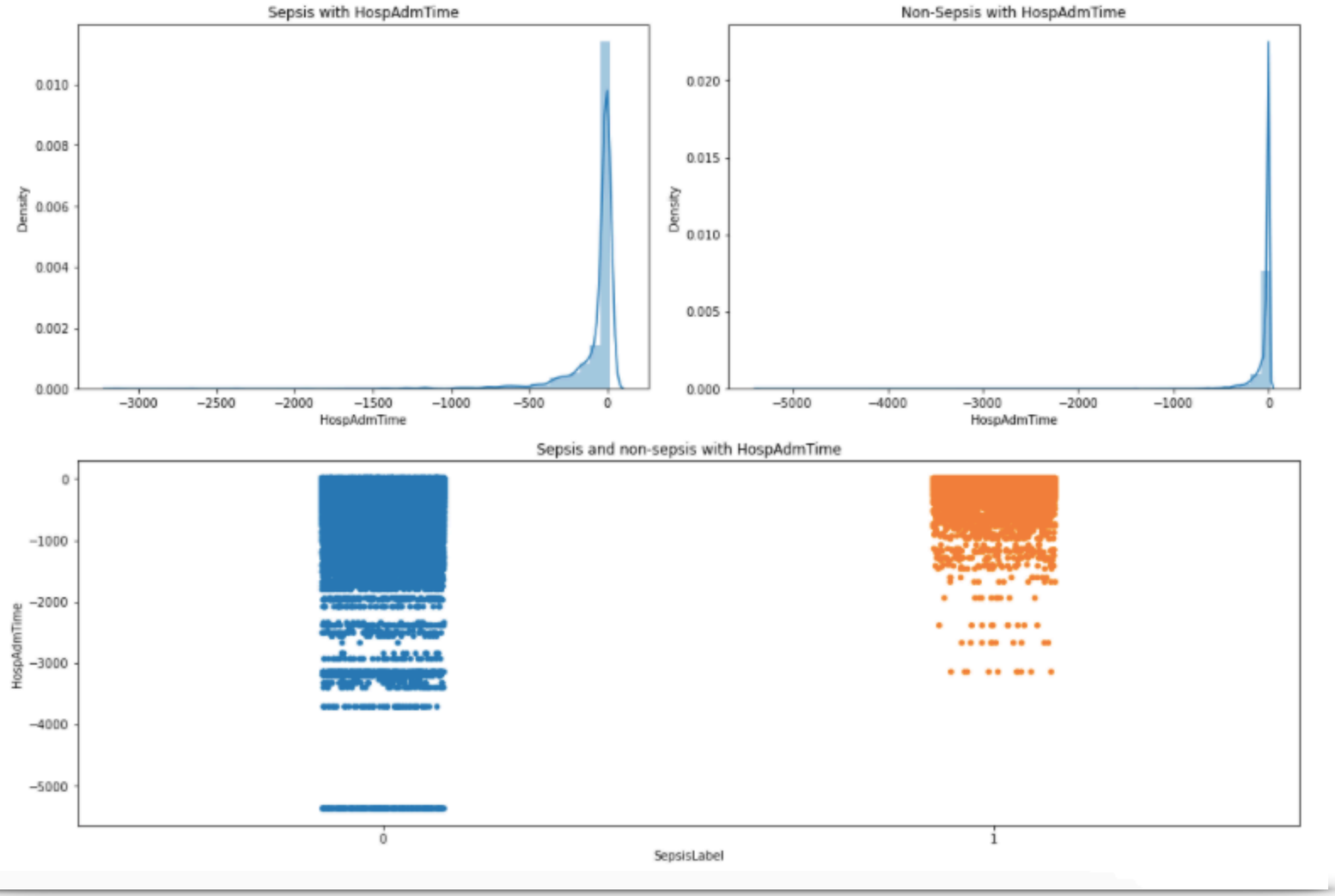
DATA VISUALISATION - SEPSIS AND ICULOS

- The clear effect of the intensive case unit length of stay hours since ICU admit
- ICULOS (sepsis) > ICULOS (no-sepsis)



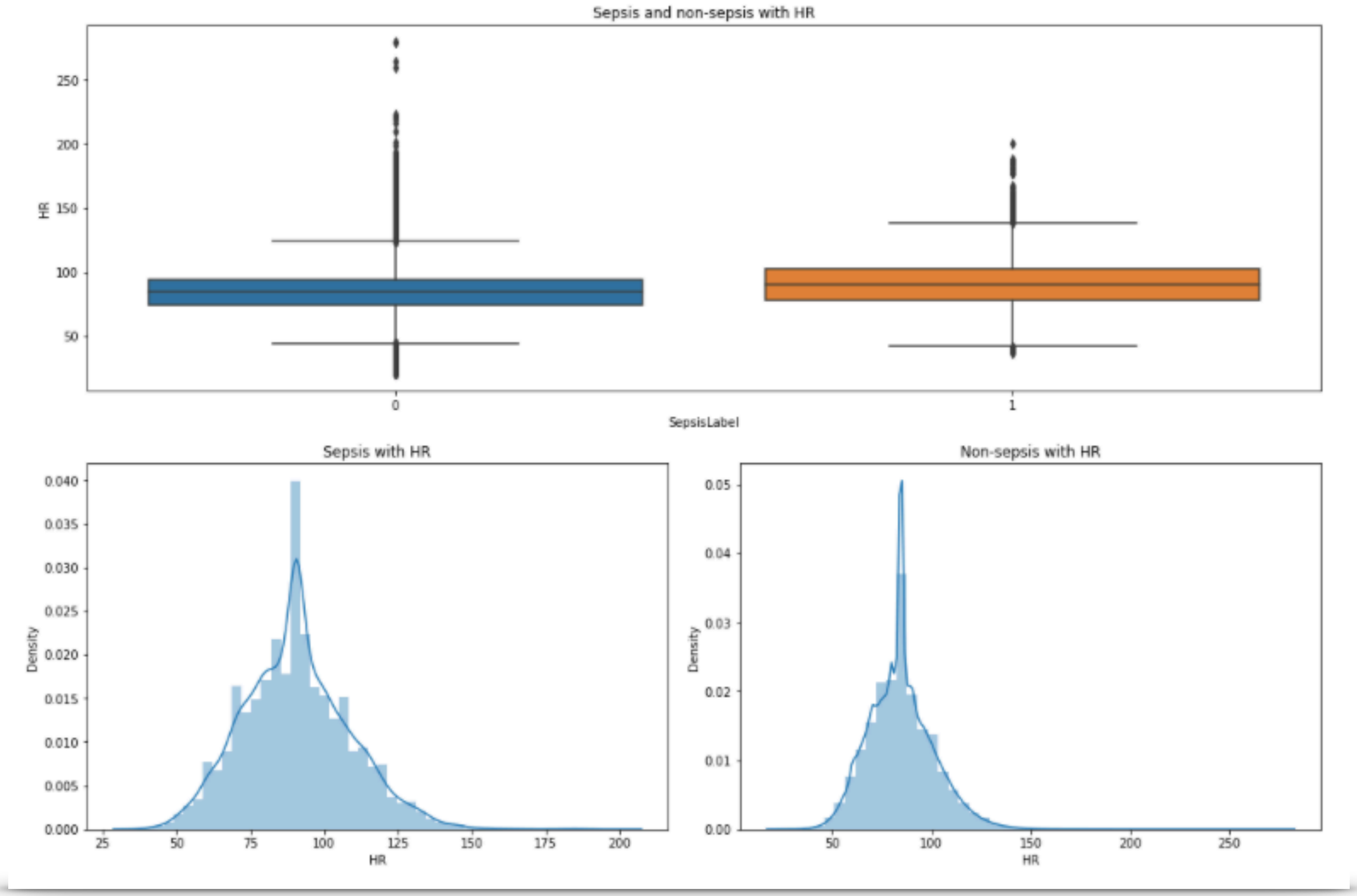
DATA VISUALISATION - SEPSIS AND HOSPADMTIME

- The effect of hospital length of stay was investigated.
- P/s: ...need discussion this point



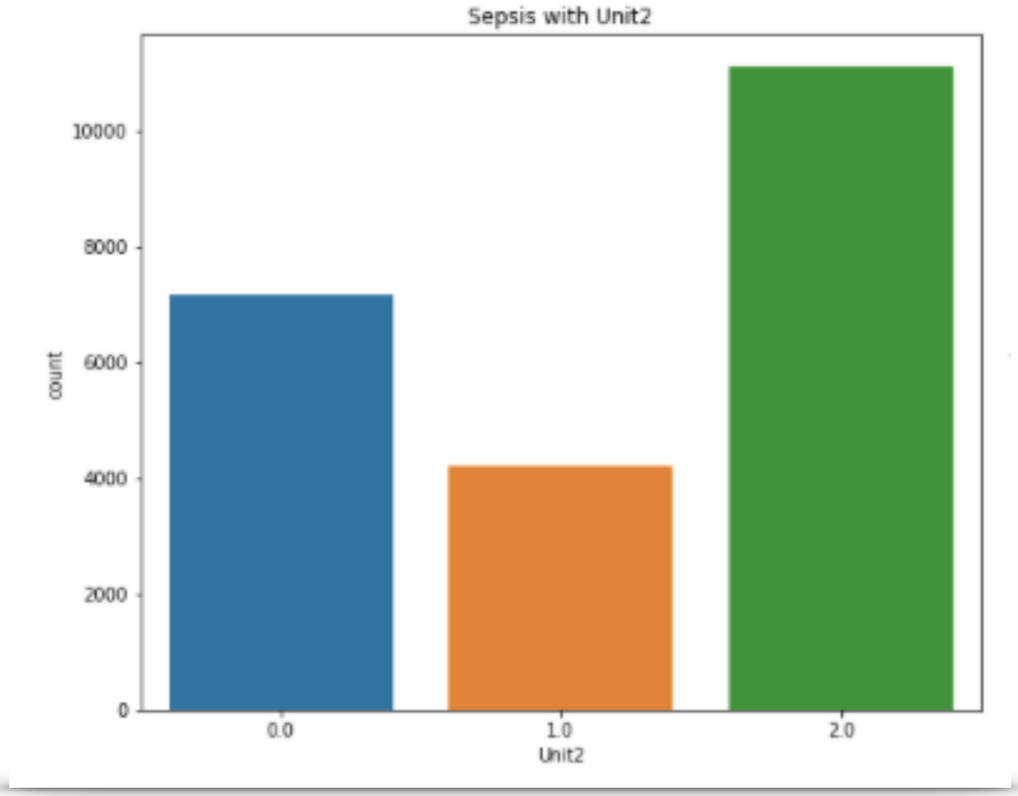
DATA VISUALISATION - SEPSIS AND HEART RATE

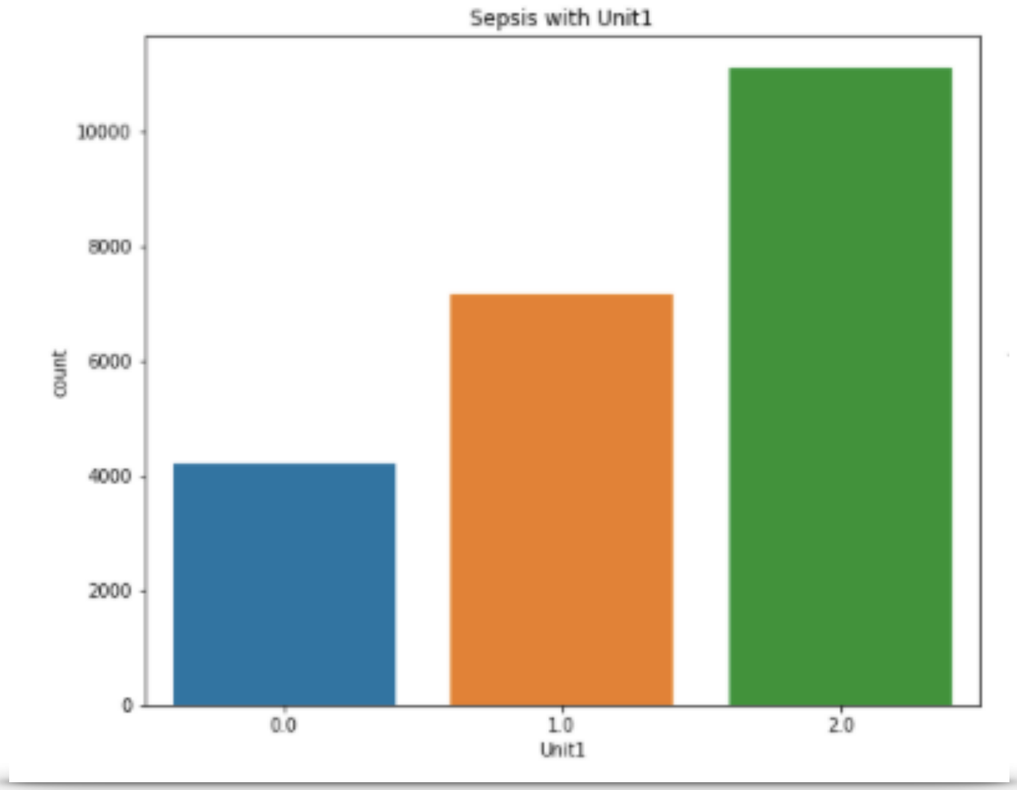
- heart rate of more than 90 beats per minute
- increased heart rate often persists in sepsis



DATA VISUALISATION - SEPSIS AND UNIT1 AND UNIT2

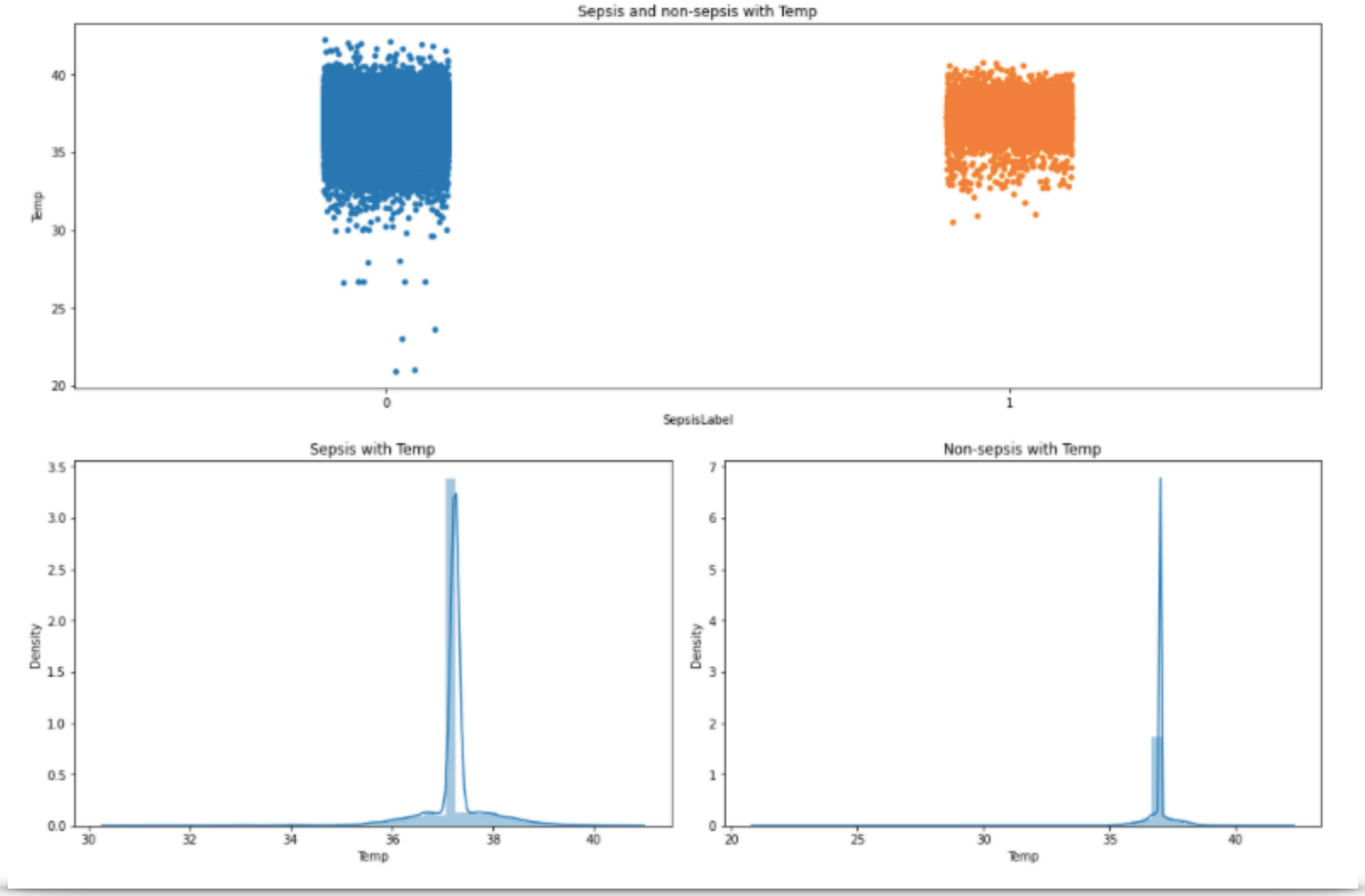
* the impact of the intervention of mobile intensive care unit (Unit1) and surgical intensive care unit (Unit2) on mortality of patients with sepsis
* P/s: need more clinical data understanding to comment!!!!





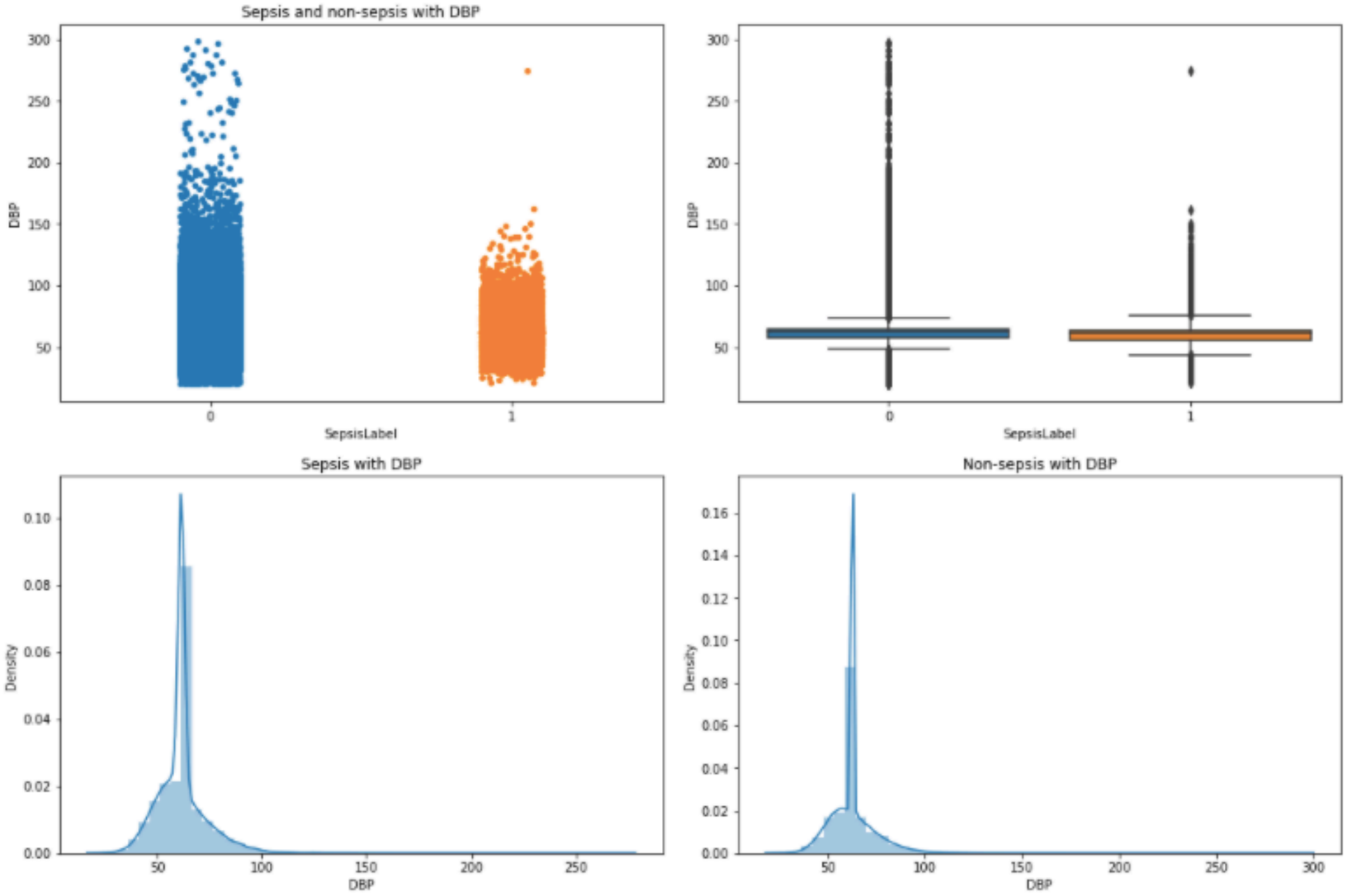
DATA VISUALISATION - SEPSIS AND TEMPERATURE

- a fever above 38°C or a temperature below 36°C



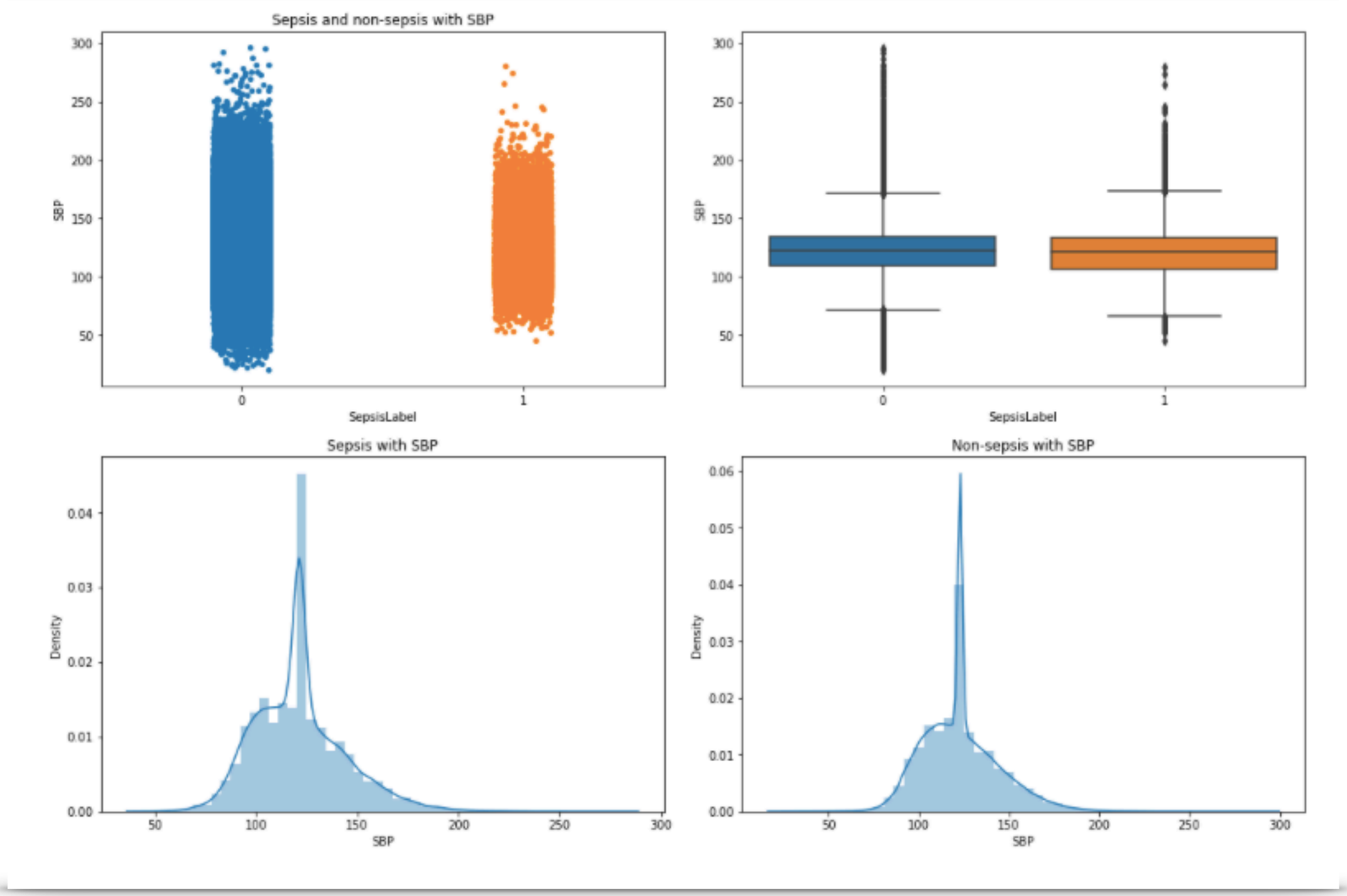
DATA VISUALISATION - SEPSIS AND DBP

- The values of diasystolic blood pressure was researched higher in patients with sepsis.
- *P/s: it isn't clear in the nude eyes, need more discussions.*



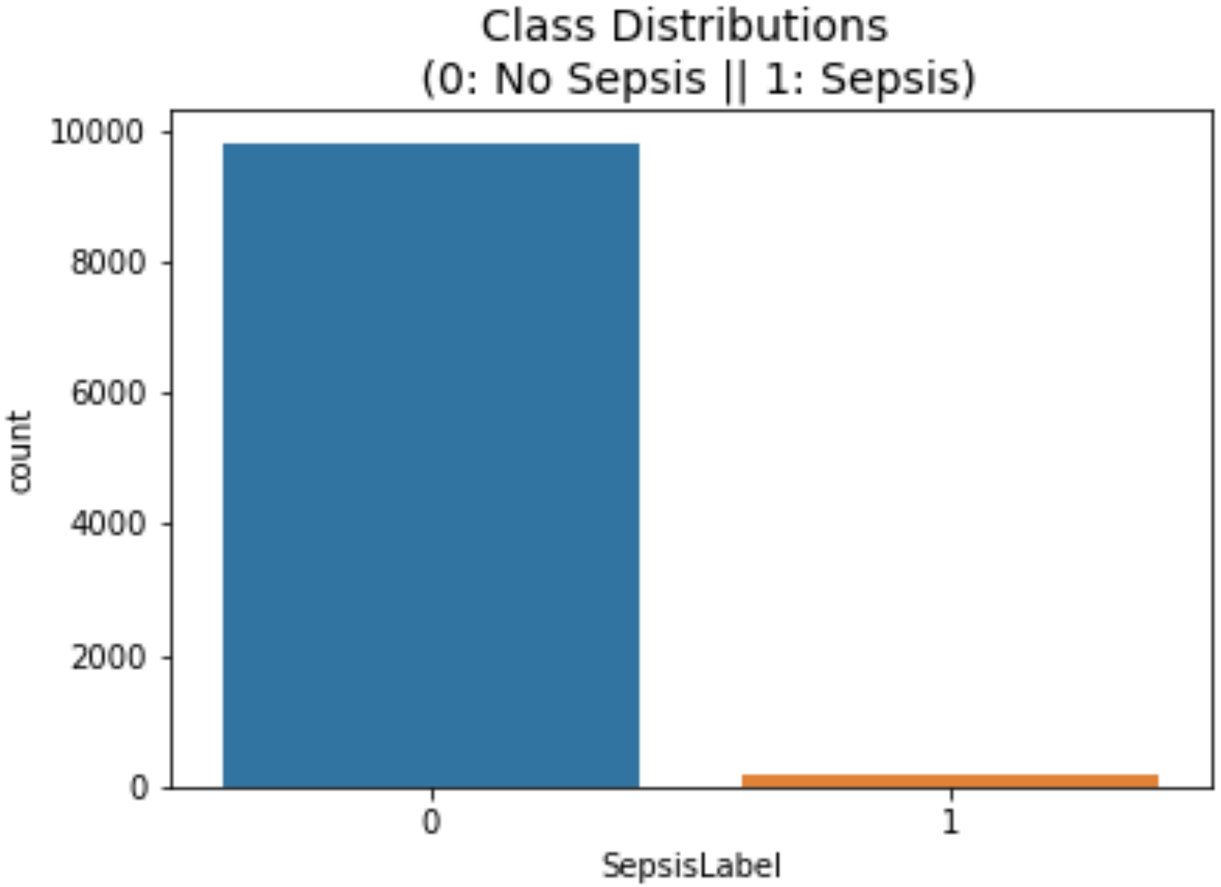
DATA VISUALISATION - SEPSIS AND SBP

- The values of systolic blood pressure was researched higher in patients with sepsis.
- *P/s: it isn't clear in the nude eyes, need more discussions.*



MODEL TRAINING AND EVALUATION

* **Imbalanced data:** No Sepsis 98.08 % and Sepsis 1.92 %



* **Random resampling technique:**

SMOTE-NC

MODEL TRAINING AND EVALUATION

- **Label encoding:** convert the categorical text data into model-understandable OnehotEncoding,

- Splitting data into **testing** and **training sets,**

```
Label Distributions:

Train distribution:  [0.981125 0.018875]
Test distribution:  [0.9795 0.0205]
```

- **Random resampling on the training set** and **evaluate models on the original testing set,**

- **The main goal:** fit the model either with the dataframes that were resampled in order for our models to detect the patterns, and test it on the original testing set.

MODEL TRAINING AND EVALUATION

* Comparison of machine learning models

	model	precision	recall	f1-score	AUC-ROC
0	CatBoost	1.000000	0.902439	0.948718	0.997784
1	LGBM	1.000000	0.853659	0.921053	0.997373
2	Logistic	0.040367	0.536585	0.075085	0.684209
3	RandomForest	0.972222	0.853659	0.909091	0.997466
4	DecisionTree	0.878049	0.878049	0.878049	0.937748

```
============================================================
CatBoost SMOTENC
============================================================
Confusion Matrix
============================================================
[[1959    0]
 [   4   37]]

Classification Report
============================================================
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1959
           1       1.00      0.90      0.95        41

    accuracy                           1.00      2000
   macro avg       1.00      0.95      0.97      2000
weighted avg       1.00      1.00      1.00      2000
```

* The best model - CatBoost

CONCLUSIONS

- Data visualization is quite correlated with medical research literature.
- Symptoms of sepsis are significant with data, especially for ICULOS, Age, HR, Temp.
- SMOTE-NC handles mixed numerical and categorical data well on an imbalanced dataset.
- Machine learning models can detect sepsis or non-sepsis ability based on patient's medical data per any hour.

PERPECTIVES

- Research more about the correlation between SepsipLabel and other patients' features.
- Improve and try other ML algorithms, such as XGboost, SVM, etc.
- Consider the time component for the data, particularly, sepsis is diagnosed for each patient at each hour using the past data. Hence, we can such use popular ML/DL models: LSTM, Prophet, ARIMA, etc.

THANK YOU FOR YOUR ATTENTION!