1，Table join

Tables used				
patients	icustays	chartevents	outputevents	d_items

Personal Info				
Feature Name	**subject_id**	**icustay_id**	**age**	**gender**
Explain	Number for each patient	Number for each icu case	age	gender
Feature Name	**marital_status**	**ethnicity**	**icu_duration_hour**	**icu_times**
Explain	marital_status	ethnicity	Time spend in ICU	The ith time being admitted to ICU
Feature Name	**icu_times_total**	**icu_times_sepsis**	**icu_times_total_sepsis**	**hospital_expire_flag**
Explain	How many times being admitted to ICU	The ith time being admitted to ICU because of sepsis	How many times being admitted to ICU because of sepsis	Died in hospital
Feature Name	**expire_flag**	**died_immediately**		
Explain	Died eventually	Died in hospital or within 24h left hospital		

Test labels selection

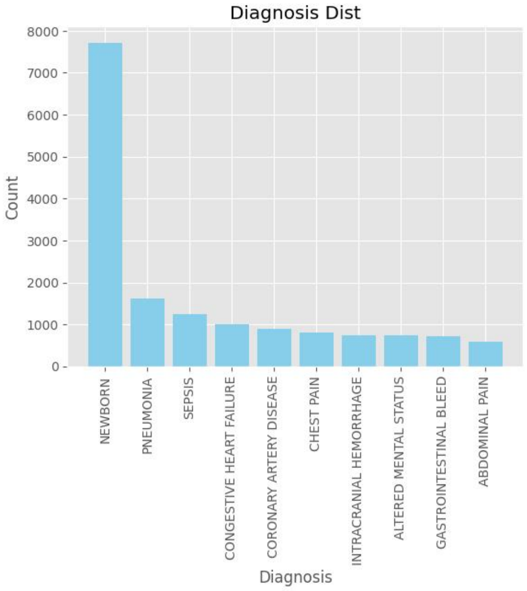
Initial label	Merged label	values
Glouse	glouse	min
Glucose (70-105)		max
		avg
Fingerstick Glucose	Fingerstick Glucose	min
		max
		avg
potassium	potassium	min
Potassium (3.5-5.3)		max
Potassium (3.5-5.3)		avg
Sodium (135-148)	Sodium	min
		max
		avg
Hematocrit	Hematocrit	min
Hematocrit (35-51)		max
		avg

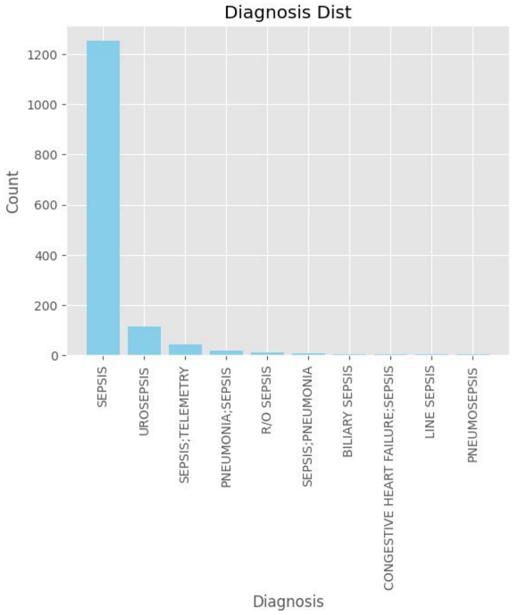
Chloride (100-112)	Chloride	min
		max
		avg
BUN (6-20)	BUN	min
BUN		max
		avg
Creatinine (0-1.3)	Creatinine	min
		max
		avg
Hemoglobin	Hemoglobin	min
		max
		avg
Carbon Dioxide	Carbon Dioxide	min
		max
		avg
RBC	RBC	min
RBC(3.6-6.2)		max
		avg
Platelets	Platelets	min
		max
		avg
WBC (4-11,000)	WBC	min
WBC (4-11,000)		max
WBC 4.0-11.0		avg
Heart Rate	Heart Rate	min
		max
		avg
Heart Rhythm	Heart Rhythm	mode
		last
Magnesium (1.6-2.6)	Magnesium	min
		max
		avg
Respiratory Rate	Respiratory Rate	min
		max
		avg
Temperature F	Temperature (if C then F)	min
		max
		avg

Why sepsis?

Except for newborn, sepsis is the second largest disease in this dataset. Also it’s one of the most dangerous diseases in USA.

All diagnosis with ‘sepsis’ will all be considered as sepsis.





Comorbidity?

Will not be considered.

Sepsis is caused by other diseases and is not a spontaneous condition. Therefore, sepsis inevitably accompanies comorbidity. In the original dataset, only a few medical records have noted the cause, so it is not considered for the time being.

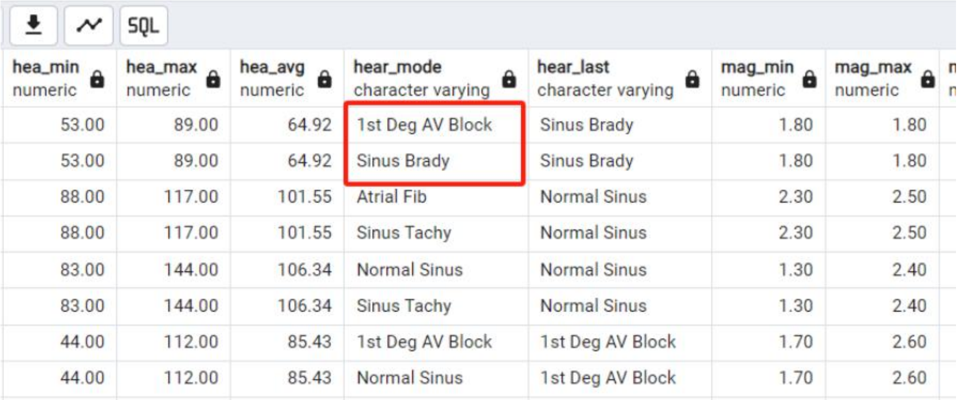
2 Pre-process

2.1 the last case

Since the same patient will have similar test values, which might affect the model, we select the last time one patient get admitted to ICU because of sepsis if the patient get admitted to ICU for multiple times.

2.2 Drop dup

Since the mode of heart rhythm might not be a unique value if two or more values have the same amount. We decided to delete all of them completely. Only 10 patients (20 records) will be removed, which might not affect the data set too much. Also, only one of the duplicated records ends up died. If we are doing an outlier detection, this also does not remove too much of the positive samples.



hea_min numeric	hea_max numeric	hea_avg numeric	hear_mode character varying	hear_last character varying	mag_min numeric	mag_max numeric
53.00	89.00	64.92	1st Deg AV Block	Sinus Brady	1.80	1.80
53.00	89.00	64.92	Sinus Brady	Sinus Brady	1.80	1.80
88.00	117.00	101.55	Atrial Fib	Normal Sinus	2.30	2.50
88.00	117.00	101.55	Sinus Tachy	Normal Sinus	2.30	2.50
83.00	144.00	106.34	Normal Sinus	Normal Sinus	1.30	2.40
83.00	144.00	106.34	Sinus Tachy	Normal Sinus	1.30	2.40
44.00	112.00	85.43	1st Deg AV Block	1st Deg AV Block	1.70	2.60
44.00	112.00	85.43	Normal Sinus	1st Deg AV Block	1.70	2.60

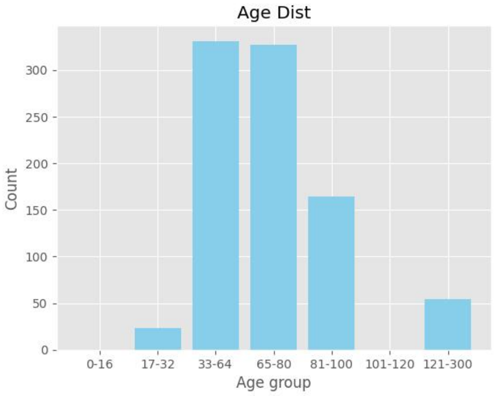
2.3 Drop null

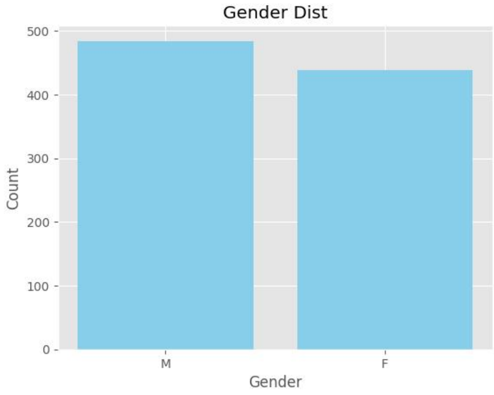
Top 20 most common tests among sepsis patients.If a patient has more than 25% of tests not conducted, then remove that patient.

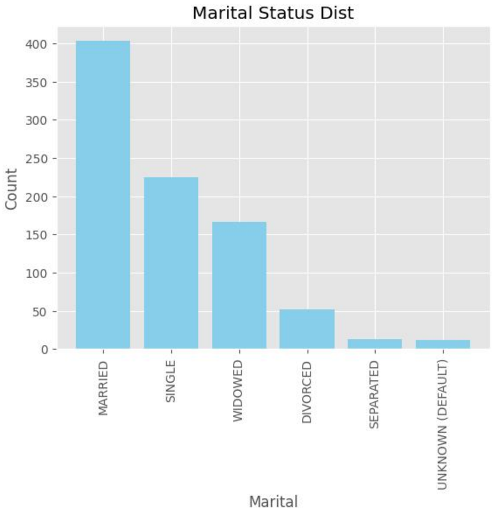
End up in 922 cases in total.

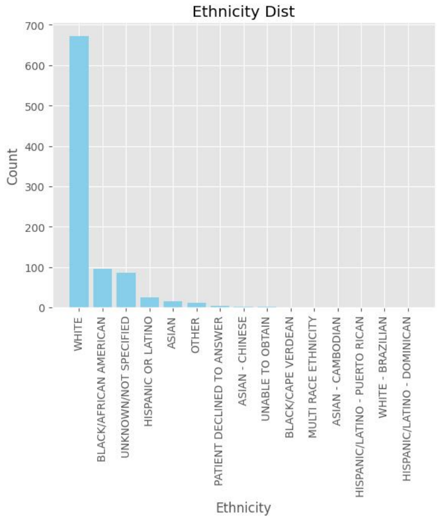
2.3 Outliers

Age: 121-300 outliers. Delete









Gender/marital_status/ethnicity/icu_duration_hour/icu_times/icu_times_total/icu_times_total_sepsis 的分布

Test value 的最大最小平均值的分布
针对数值类型的 test value，用平均值填补空值。