

Health Admission

Presented by Ok Gas

Muhammad Fachrudin Firdaus
2206819331

Muhammad Hilal Darul F.
2206830542

Muhammad Mariozulfandy
2206041404

Patrick Samuel Evans S.
2206028251

DAFTAR ISI

01.

Latar Belakang
dan Tujuan

02.

Dataset dan
Eksplorasi

03.

Task

03.

Hasil dan
Pembahasan

Latar Belakang dan Tujuan

Latar Belakang

Kesehatan merupakan salah satu aspek paling penting dalam kehidupan manusia. Kesehatan seseorang sangat mempengaruhi kesejahteraan orang tersebut dan juga orang-orang terdekatnya karena merawat orang sakit perlu tenaga yang tidak sedikit dan mengikuti logika yang sama, juga masyarakat.





Tujuan

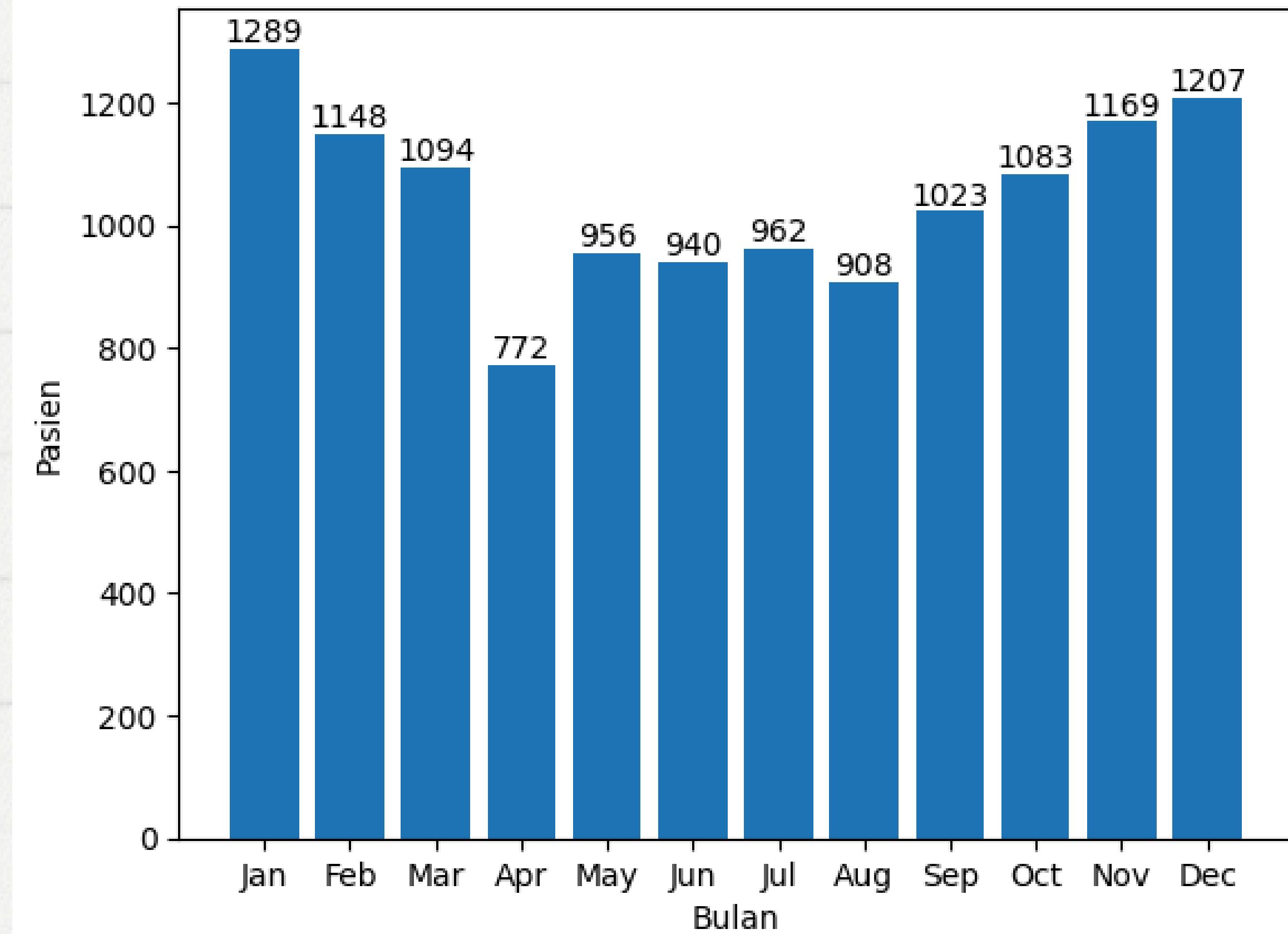
Tujuan dari proyek ini adalah:

- **Memprediksi Outcome dari pasien rumah sakit.**
Menentukan apakah pasien akan keluar dengan rekomendasi medis, keluar tanpa rekomendasi medis , atau meninggal selama perawatan.
- **Memprediksi seberapa lama pasien dirawat secara intensif di rumah sakit.** Mengestimasi berapa lama pasien akan dirawat secara intensif di rumah sakit berdasarkan data yang ada.
- **Membantu Identifikasi Profil Pasien.** Mengelompokkan pasien berdasarkan karakteristik yang serupa.

Dataset dan Eksplorasi

**Pada bulan apakah jumlah pasien
terbanyak yang dirawat di rumah sakit?**

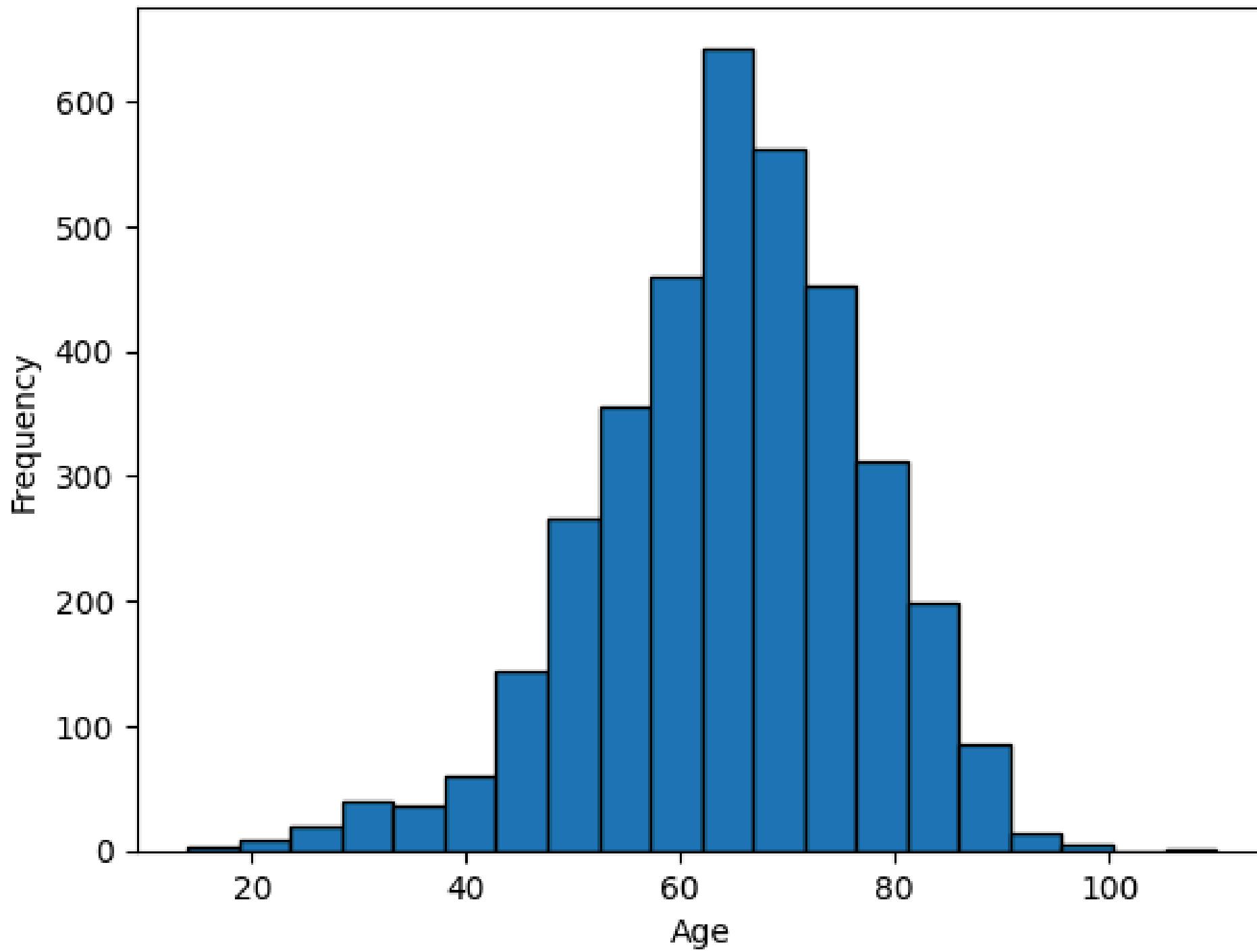
Pasien Setiap Bulan



Bagaimana karakteristik pasien yang menderita *heart failure*?

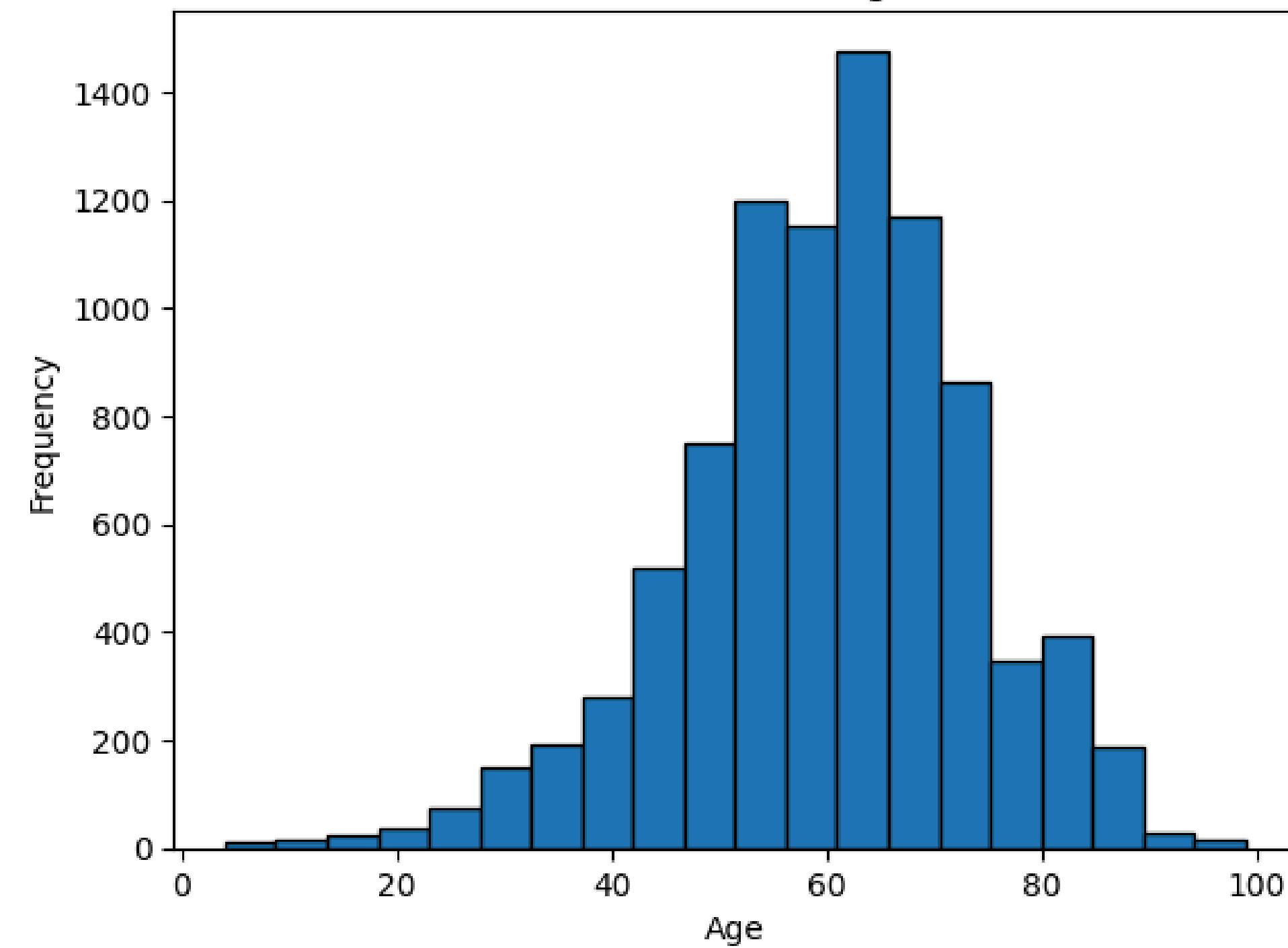
Usia Heart Failure

Distribution of Age

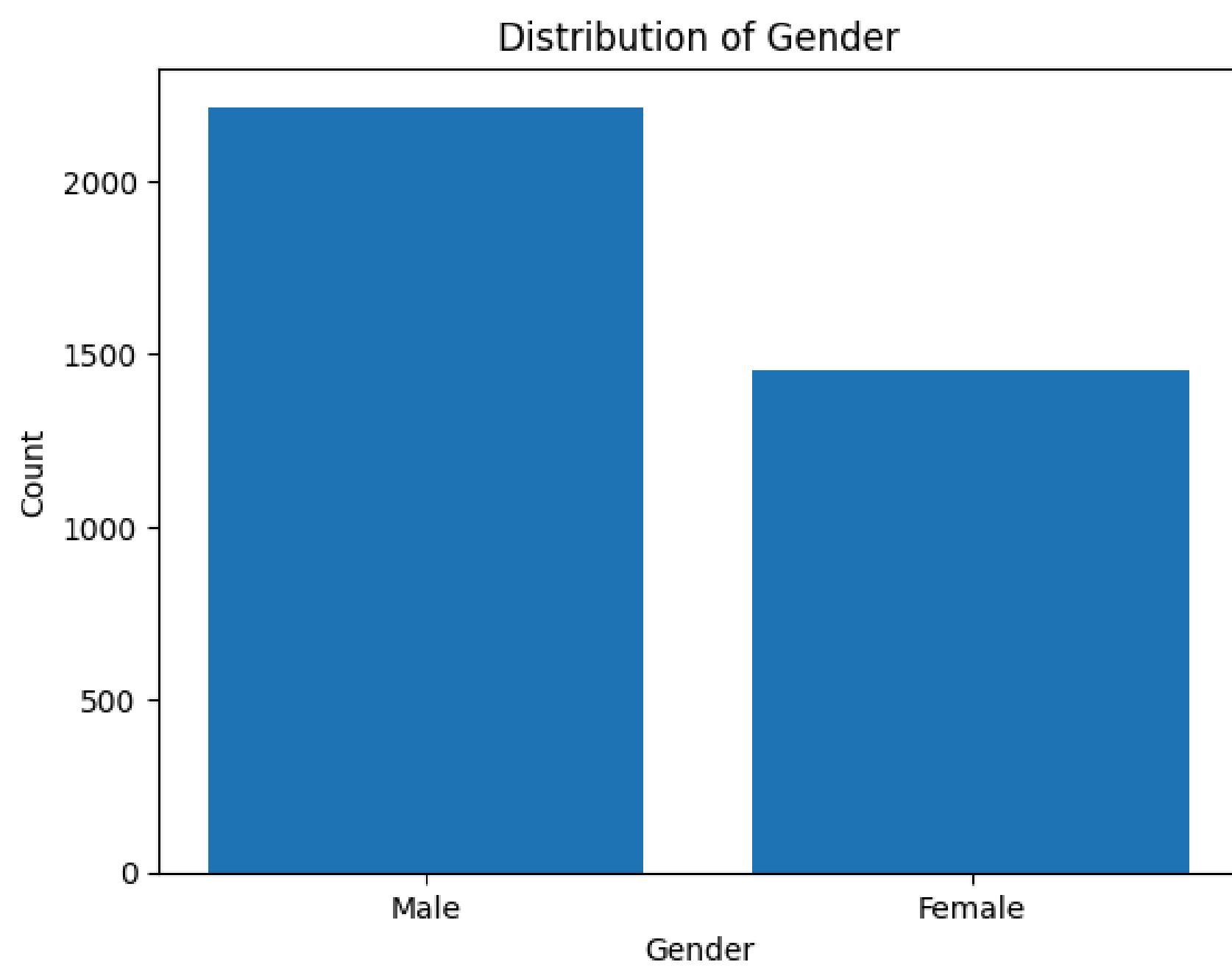


Usia Non Heart Failure

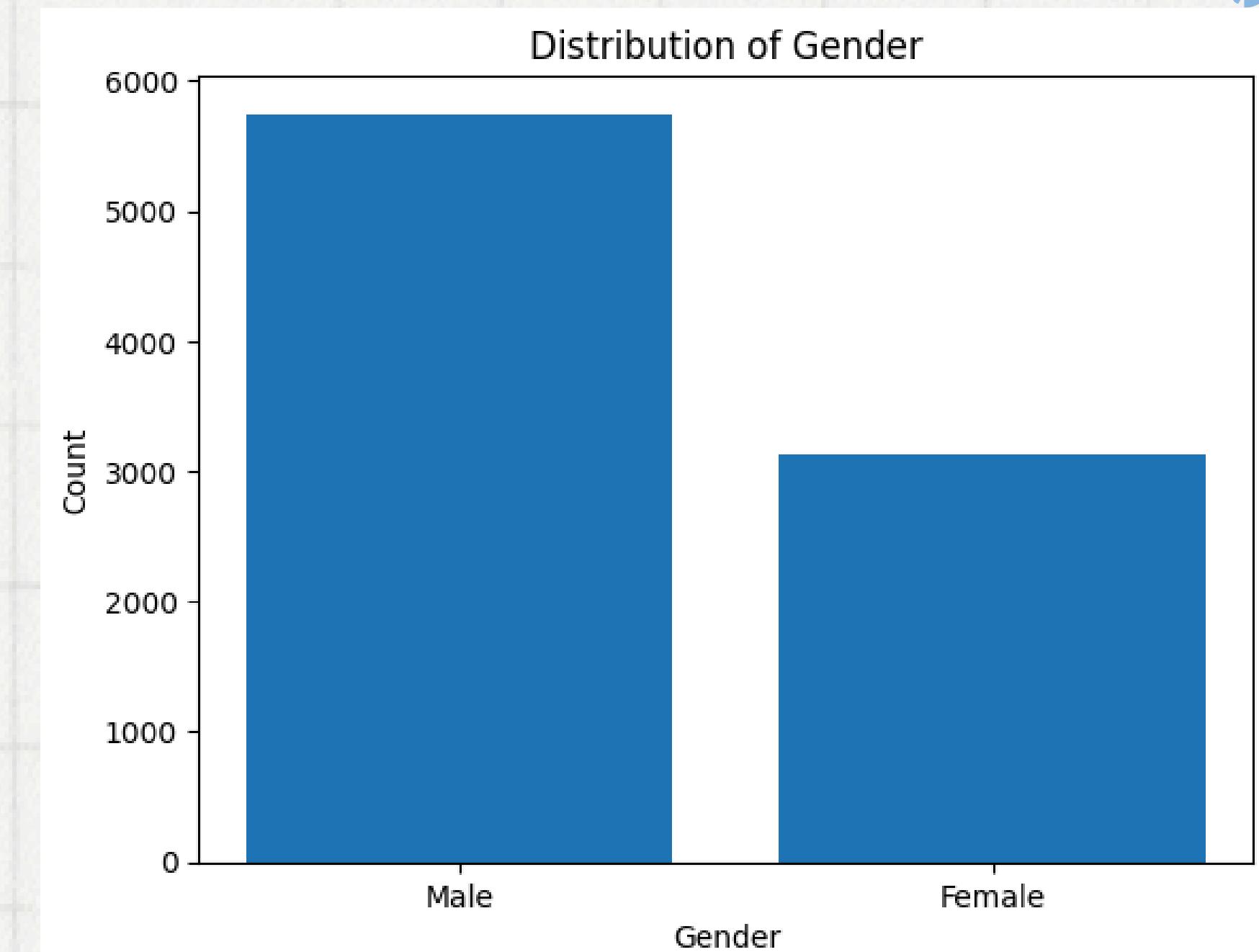
Distribution of Age



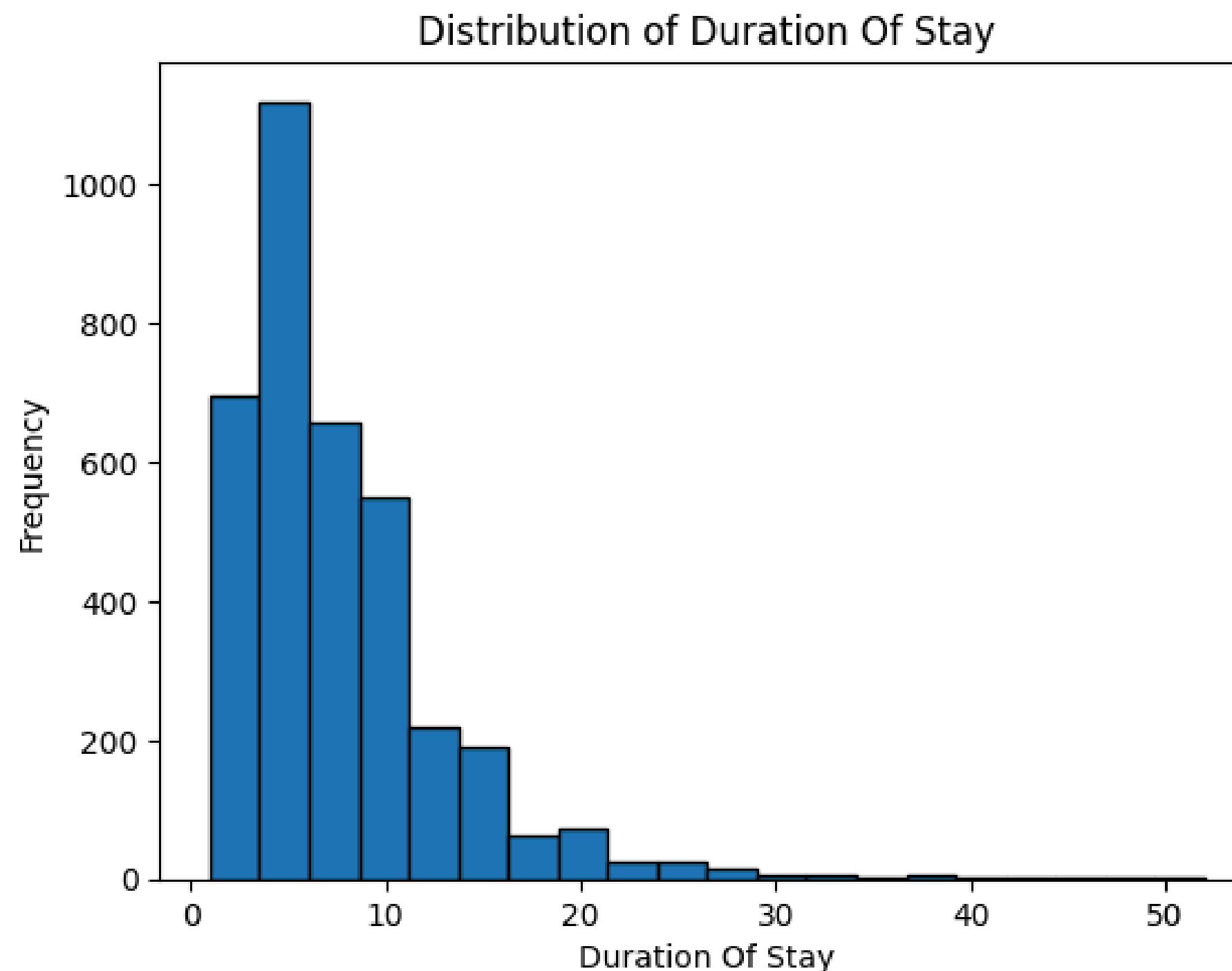
Gender Heart Failure



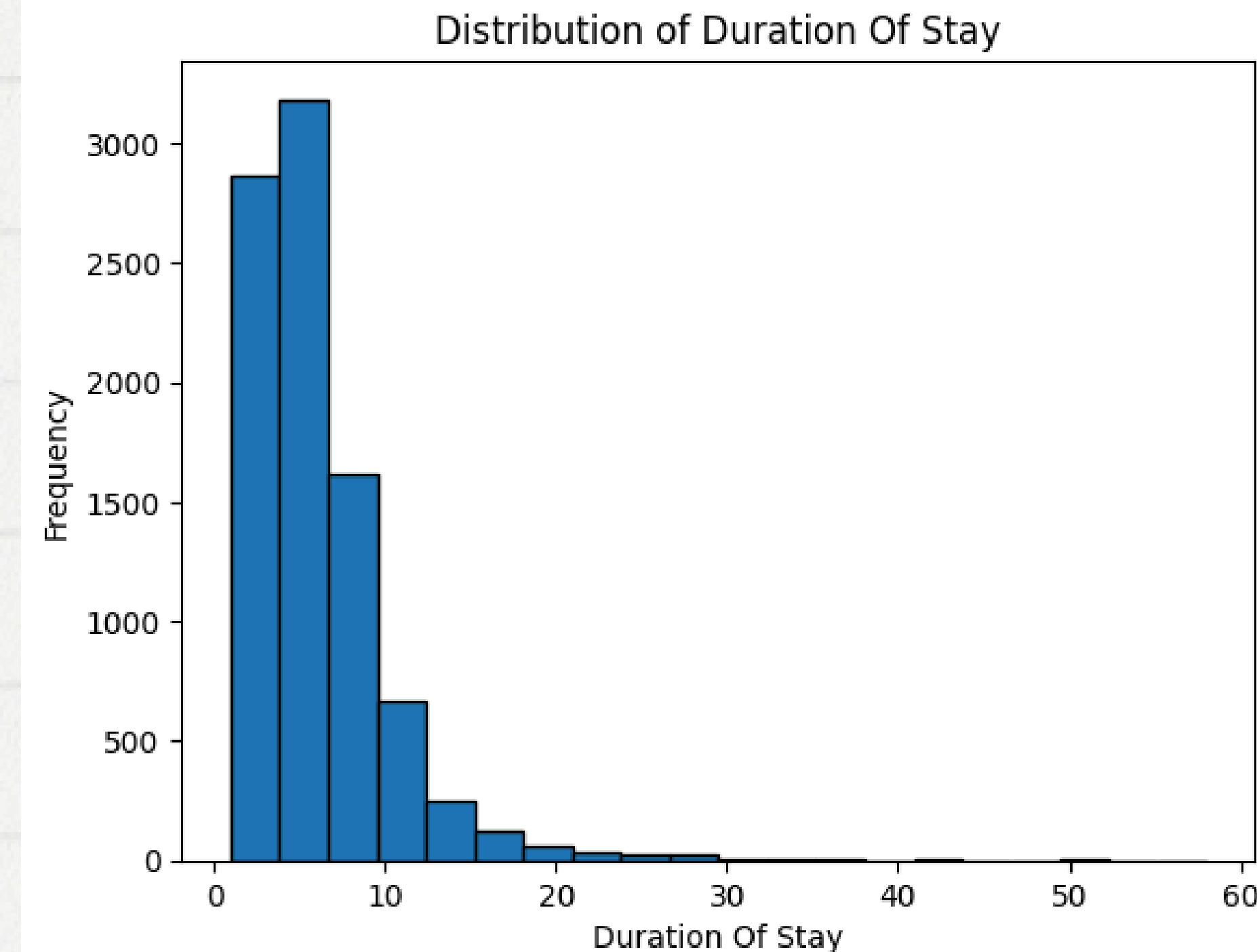
Gender Non Heart Failure



Durasi Rawat Inap Heart Failure



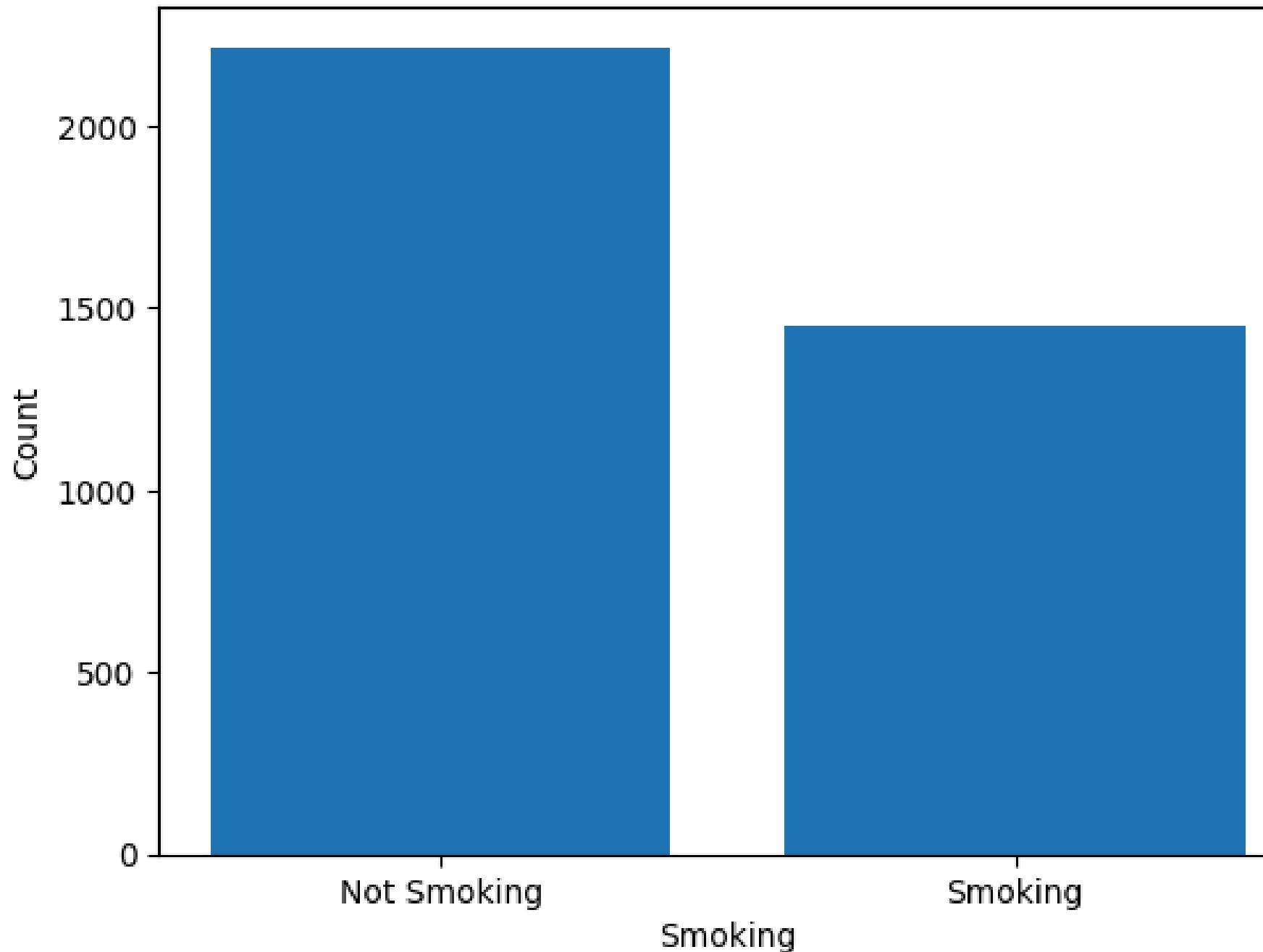
Durasi Rawat Inap Non Heart Failure



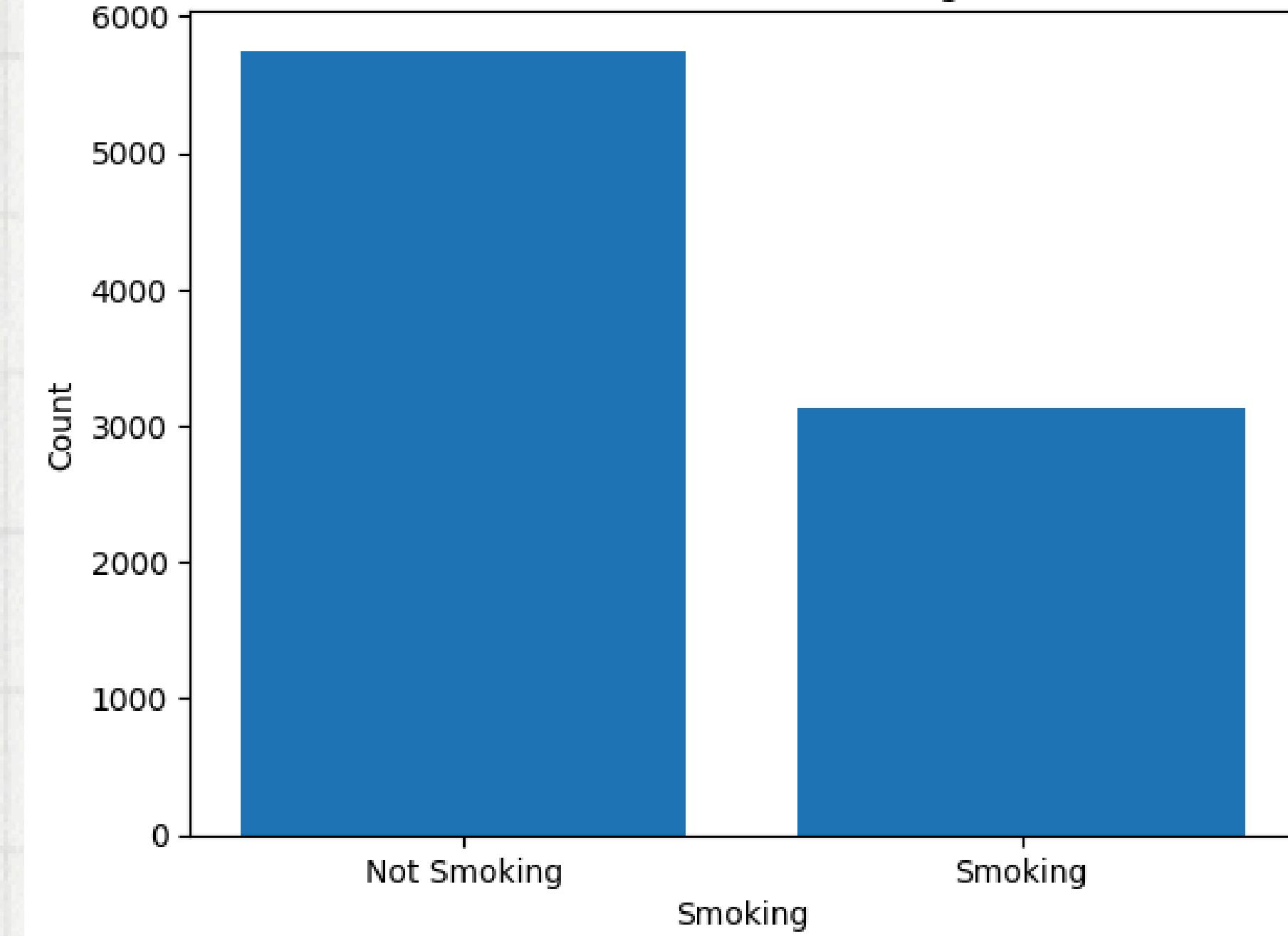
Is Perokok Heart Failure

Is Perokok Non Heart Failure

Distribution of Smoking

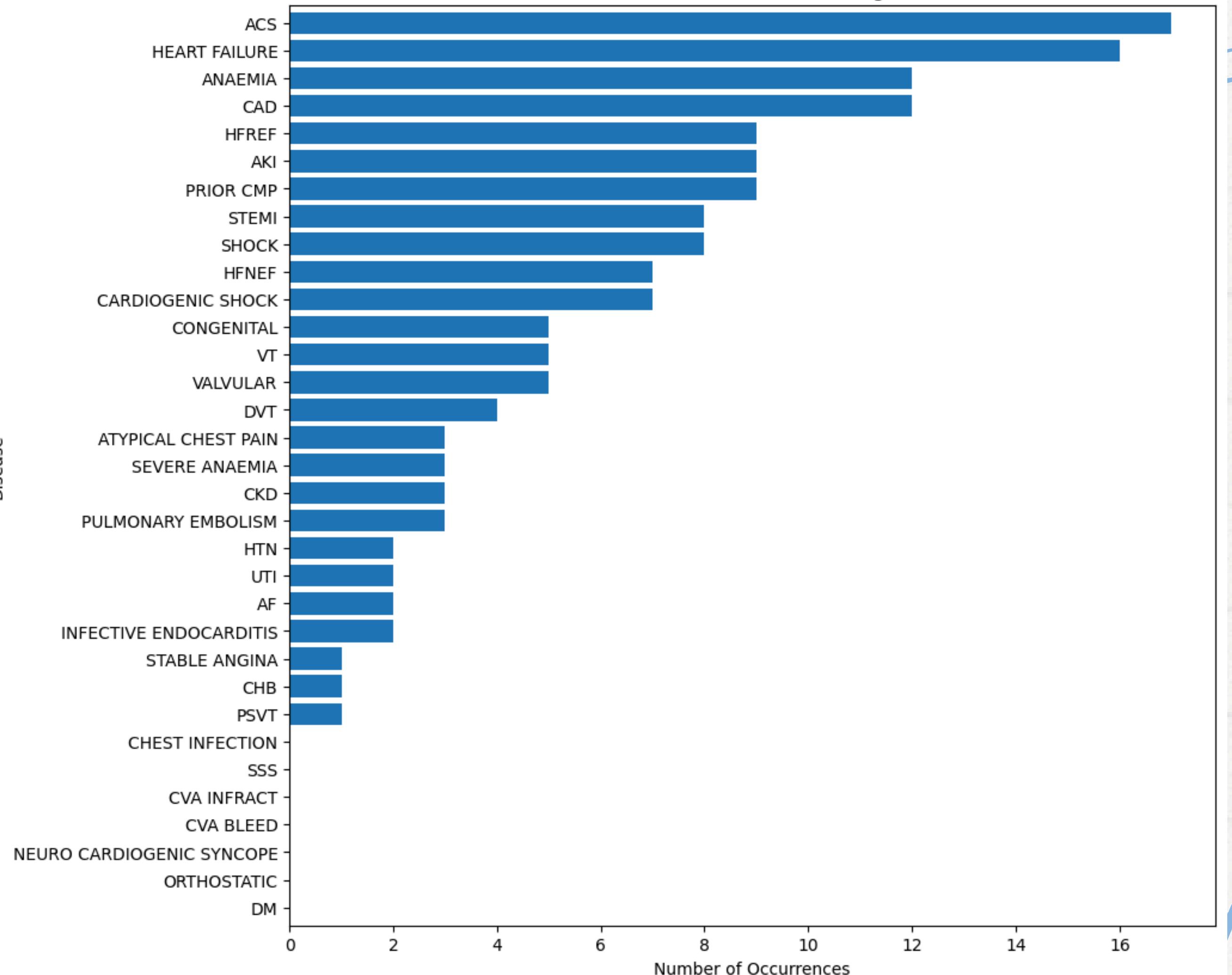


Distribution of Smoking



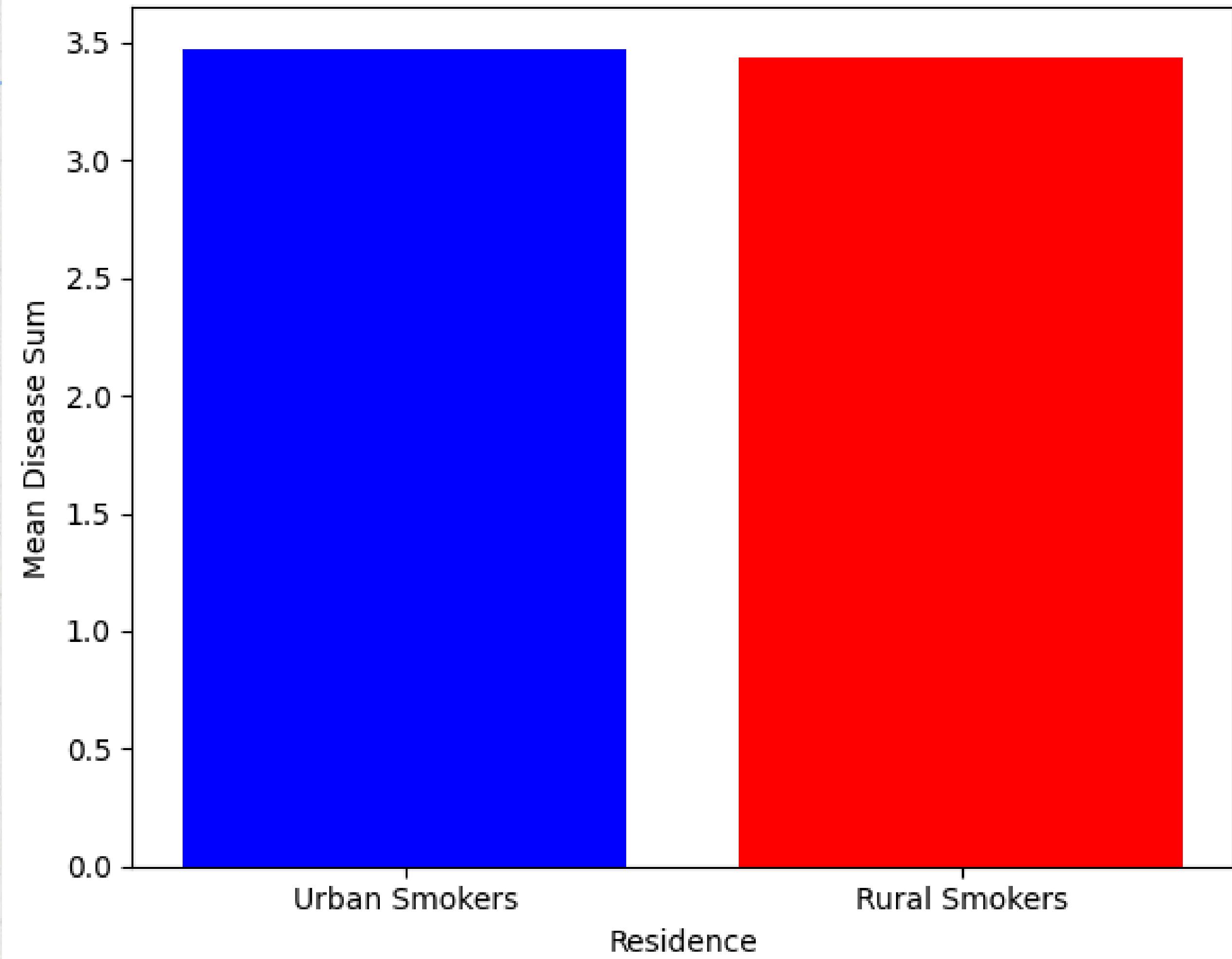
**Penyakit apa yang terbanyak diderita
pasien yang berumur 19 – 25 tahun?**

Sum of Disease Occurrences for Ages 19-25



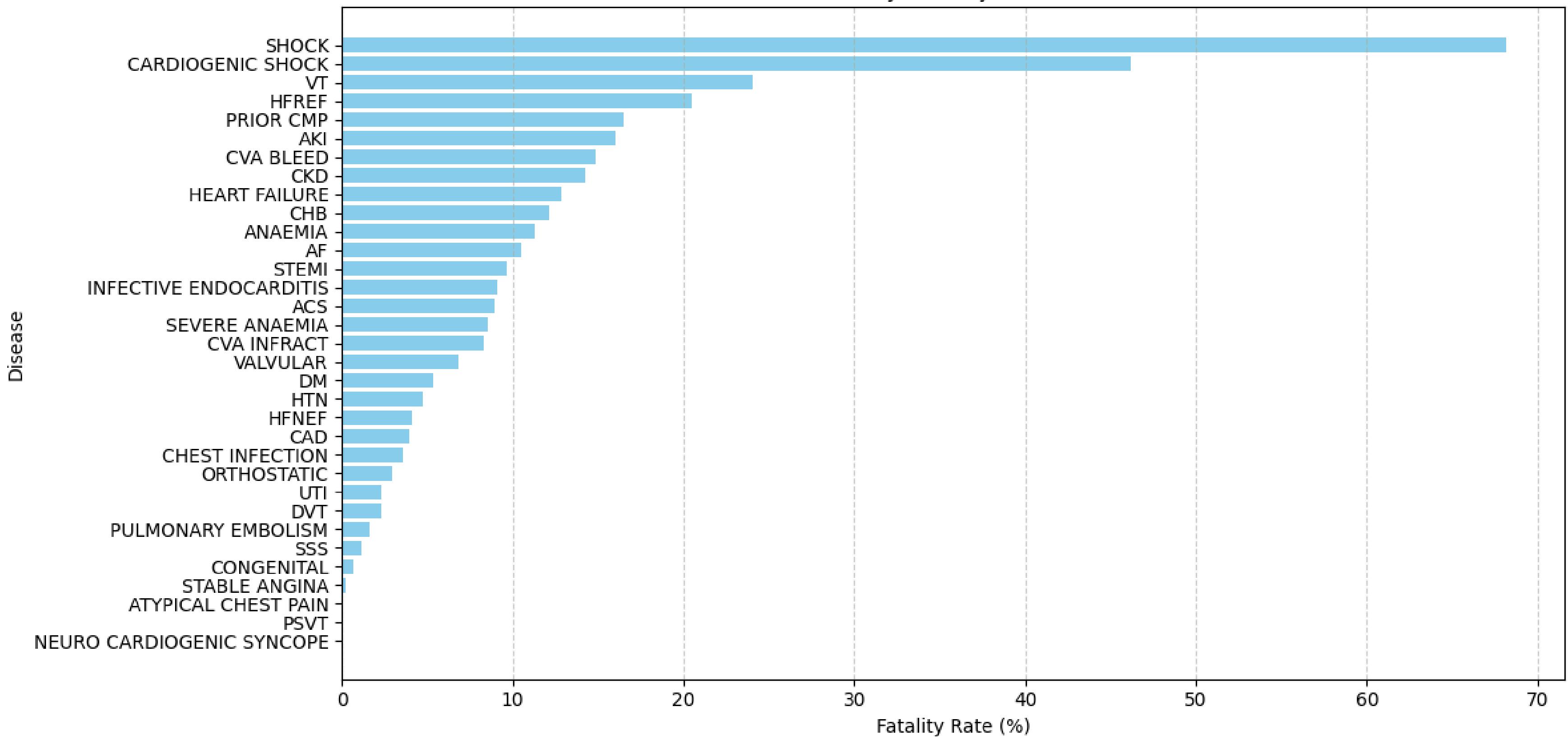
Apakah pasien yang memiliki kebiasaan merokok dan tinggal di daerah urban memiliki rata-rata komplikasi penyakit yang lebih banyak dibandingkan pasien yang memiliki kebiasaan merokok dan tinggal di daerah rural?

Mean Disease Sum by Residence for Smokers



Apa jenis penyakit yang memiliki tingkat fatalitas tertinggi?

Fatality Rate by Disease



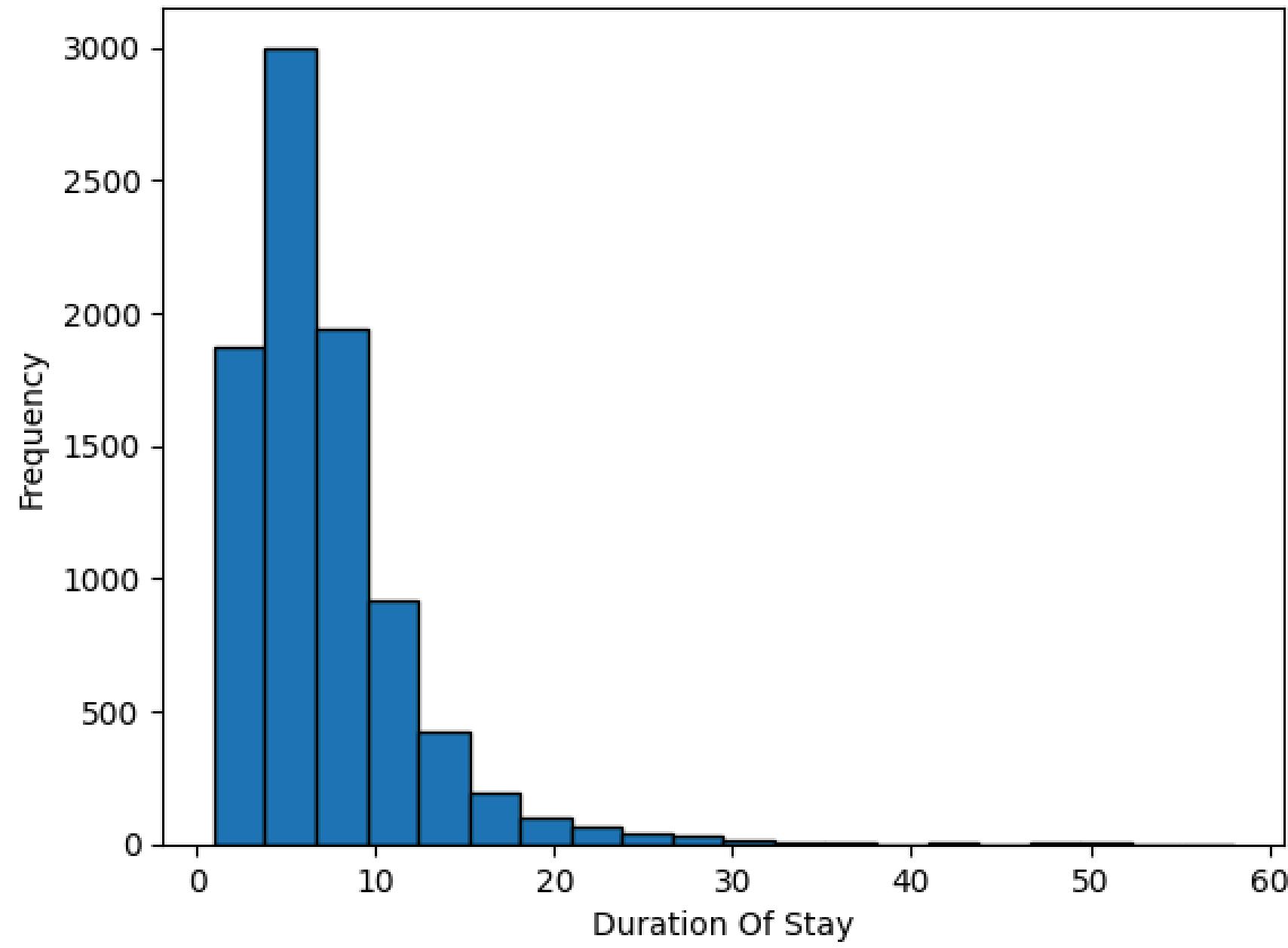
Apakah ada hubungan antara jenis emergensi perawatan dengan durasi pasien dirawat di rumah sakit?

Durasi Inap Pasien Emergensi

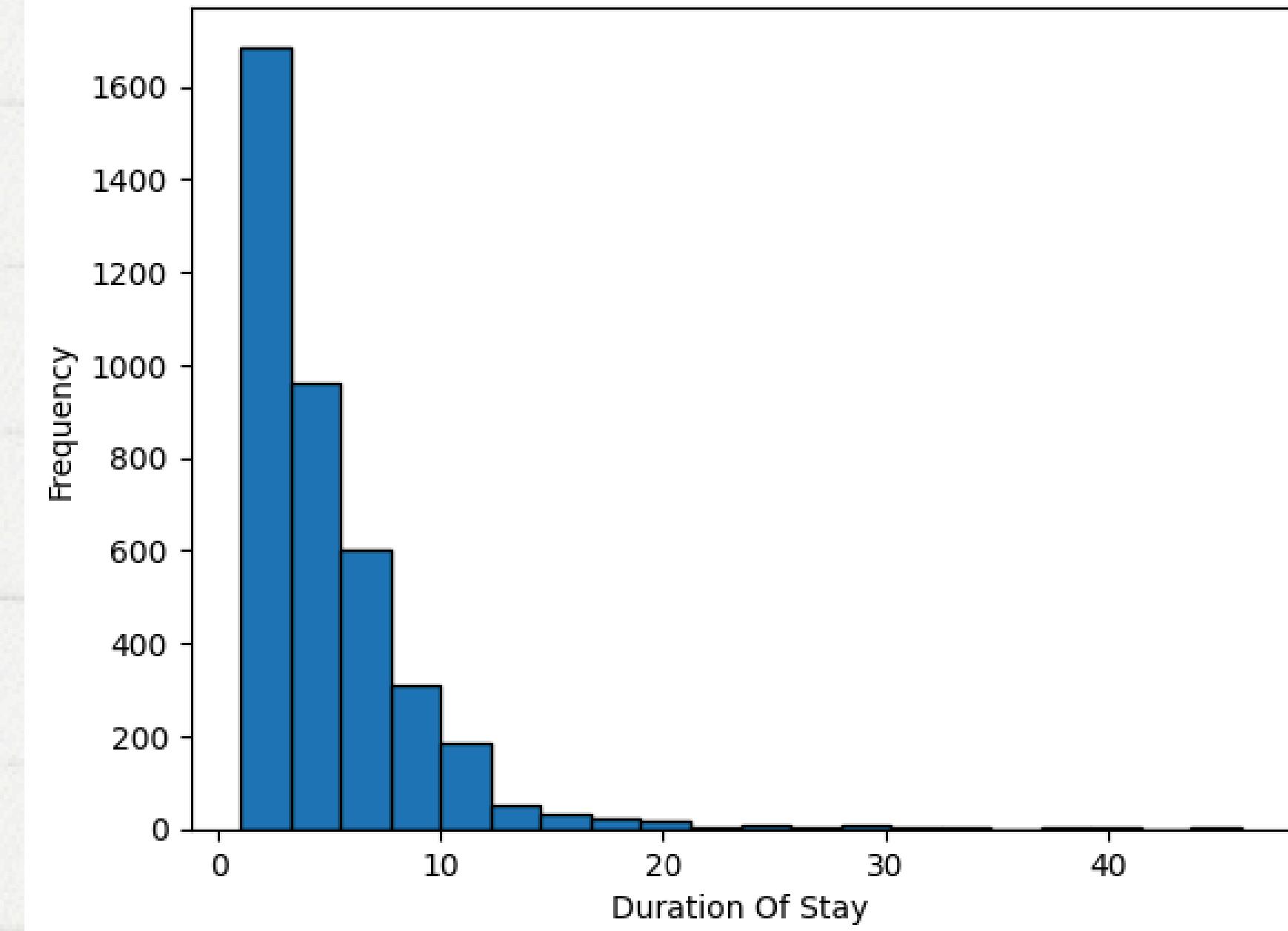
Durasi Inap Pasien Non Emergensi

Korelasi: -0.18908369439578634

Distribution of Duration Of Stay



Distribution of Duration Of Stay



Data Cleaning

Missing & Duplicate Values

Dapat terlihat bahwa tidak ada entry duplikat pada dataset dan missing values pada kolom BNP terlalu banyak, kami memutuskan untuk melakukan drop kolom tersebut. Untuk kolom sisanya yang masih terdapat missing values akan dihandle ketika data preprocessing.

	Total	Percent
UREA	25	0.199187
PLATELETS	29	0.231057
TLC	30	0.239025
CREATININE	31	0.246992
GLUCOSE	580	4.621146
EF	1194	9.513186
BNP	6615	52.704964

```
df.duplicated().any()  
✓ 0.0s  
False  
  
Tidak ada entry duplikat pada dataset.
```

Outliers

Outlier pada tiap atribut:

SNO	0
AGE	267
DURATION OF STAY	585
duration of intensive unit stay	542
SMOKING	631
ALCOHOL	836
DM	0
HTN	0
CAD	0
PRIOR CMP	1968
CKD	1239
RAISED CARDIAC ENZYMES	2506
SEVERE ANAEMIA	245
ANAEMIA	2238
STABLE ANGINA	1048
ACS	0

STEMI	1775
ATYPICAL CHEST PAIN	333
HEART FAILURE	0
HFREF	1964
HFNEF	1711
VALVULAR	439
CHB	314
SSS	85
AKI	2807
CVA INFRACT	372
CVA BLEED	47
AF	617
VT	415
PSVT	98
CONGENITAL	138
UTI	772
NEURO CARDIOGENIC SYNCOPE	104
ORTHOSTATIC	102
INFECTIVE ENDOCARDITIS	22
DVT	172
CARDIOGENIC SHOCK	730
SHOCK	534
PULMONARY EMBOLISM	184

Menurut kami, nilai-nilai outliers dari kolom-kolom yang ada mungkin saja terjadi di dunia nyata. Mengingat pentingnya tidak kehilangan informasi dalam konteks medis, kami memilih untuk biarkan saja outlier ini karena tidak ingin kehilangan potensi informasi penting tentang pasien dengan nilai ekstrem yang mungkin menyebabkan durasi pasien dirawat secara intensif cukup lama dan outcome dari pasien.



Data Preprocessing

Encoding

Kami mengekstrak kolom "month year" menjadi kolom "month" dan "year". Mengingat kolom ini kemungkinan besar relevan.

Ada 5 Kolom yang merupakan data kategorikal dan perlu diencode. Yaitu Gender, Rural, Type of Admission, Month, dan Outcome. Semua Encoding dilakukan secara manual. Untuk kolom Gender, Rural, Type of Admission karena unique valuenya hanya ada 2 maka tipe encoding apapun akan bekerja. Untuk Kolom Month dipetakan sesuai urutan bulan pada satu tahun (sesuai konteks), dan untuk Outcome kami mengikuti sesuai dengan petunjuk kaggle.

```
# Split month and year and change it as int
train[['month', 'year']] = train['month year'].str.split('-', expand=True)
test[['month', 'year']] = test['month year'].str.split('-', expand=True)

month_mapping = {
    'Jan': 1, 'Feb': 2, 'Mar': 3, 'Apr': 4,
    'May': 5, 'Jun': 6, 'Jul': 7, 'Aug': 8,
    'Sep': 9, 'Oct': 10, 'Nov': 11, 'Dec': 12
}

train['year'] = train['year'].astype(int)
test['year'] = test['year'].astype(int)

train['month'] = train['month'].map(month_mapping)
test['month'] = test['month'].map(month_mapping)

train.drop(['month year'], axis=1, inplace=True)
test.drop(['month year'], axis=1, inplace=True)

# Outcome mapping
outcome_mapping = {
    'DAMA': 0,
    'DISCHARGE': 1,
    'EXPIRY': 2
}

train['OUTCOME'] = train['OUTCOME'].map(outcome_mapping)
```

```
# Encoding
train['GENDER'] = train['GENDER'].map({'M': 1, 'F': 0})
test['GENDER'] = test['GENDER'].map({'M': 1, 'F': 0})
train['RURAL'] = train['RURAL'].map({'R': 1, 'U': 0})
test['RURAL'] = test['RURAL'].map({'R': 1, 'U': 0})
train['TYPE OF ADMISSION-EMERGENCY/OPD'] = train['TYPE OF ADMISSION-EMERGENCY/OPD'].map({'E': 1, 'O': 0})
test['TYPE OF ADMISSION-EMERGENCY/OPD'] = test['TYPE OF ADMISSION-EMERGENCY/OPD'].map({'E': 1, 'O': 0})
```

Drop Kolom

Kami melakukan drop untuk kolom yang tidak relevan. Seperti Admission Number, Serial Number. Kami juga melakukan drop ke D.O.A (Date Of Admission) dan D.O.D (Date Of Discharge) karena menurut kami informasi pada dua kolom tersebut sudah dicakup oleh kolom DURATION OF STAY.

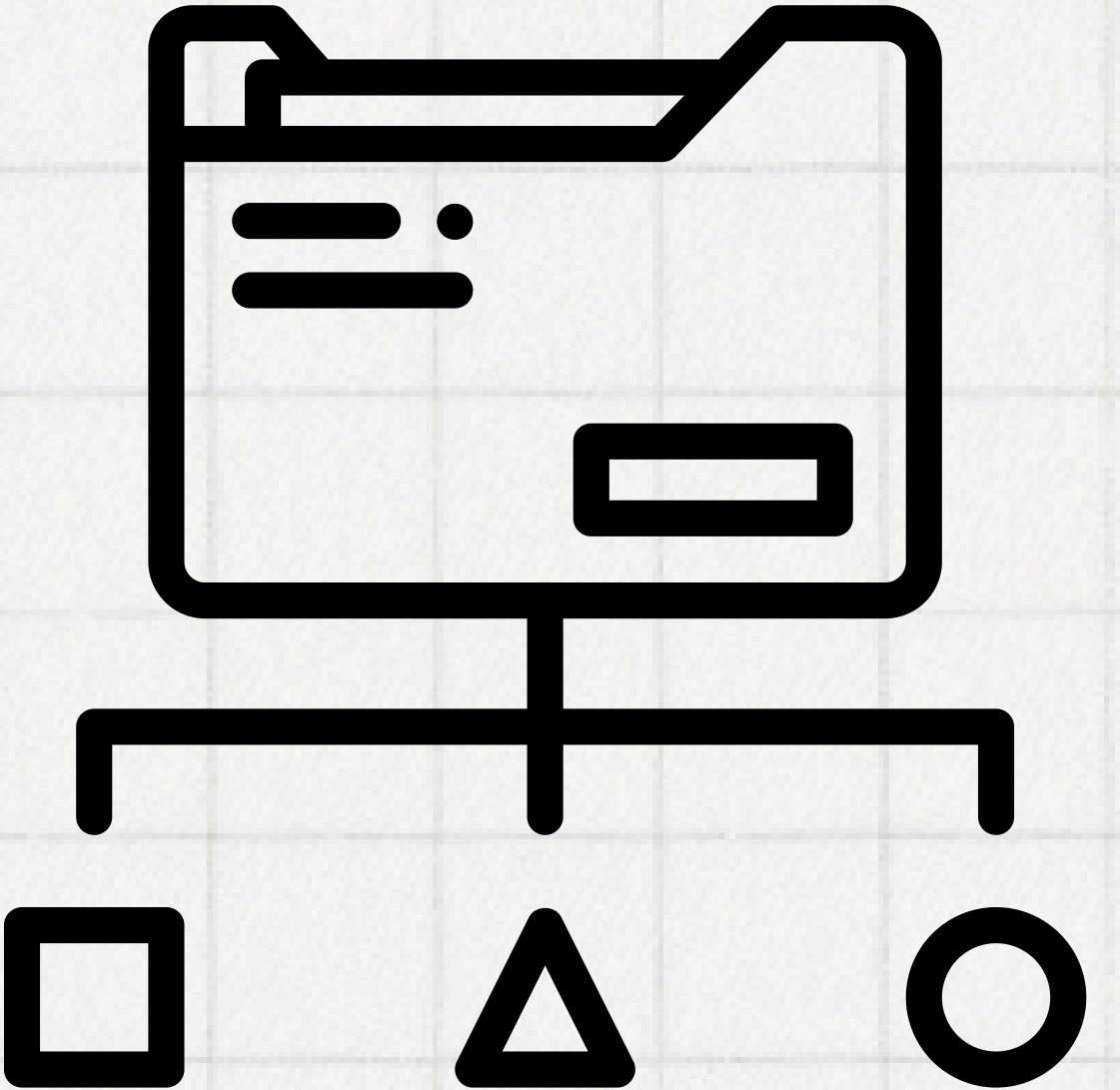
```
# Drop Admission Number karena tidak berisi informasi apapun
train.drop(['MRD No.'],axis=1, inplace=True)
test.drop(['MRD No.'],axis=1, inplace=True)

# Drop Serial Number karena tidak berisi informasi apapun
train.drop(['SNO'],axis=1, inplace=True)
test.drop(['SNO'],axis=1, inplace=True)

# Drop D.O.A karena sudah direpresentasikan melalui kolom lain
train.drop(['D.O.A'],axis=1, inplace=True)
test.drop(['D.O.A'],axis=1, inplace=True)

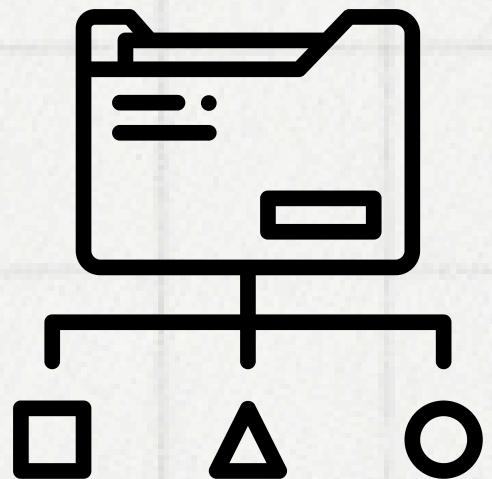
# Drop D.O.D karena sudah direpresentasikan melalui kolom lain
train.drop(['D.O.D'],axis=1, inplace=True)
test.drop(['D.O.D'],axis=1, inplace=True)
```

Hasil Dan Pembahasan



Klasifikasi





Klasifikasi

Melakukan fine tuning beberapa model classifier.

```
param_grid = {
    'logistic_regression': {
        'C': [0.1, 1.0, 10.0],
        'penalty': ['l2']
    },
    'decision_tree': {
        'max_depth': [None, 10, 20, 30],
        'min_samples_split': [2, 5, 10]
    },
    'random_forest': {
        'n_estimators': [100, 200, 300],
        'max_depth': [None, 10, 20, 30],
        'min_samples_split': [2, 5, 10]
    },
    'adaboost': {
        'n_estimators': [50, 100, 200],
        'learning_rate': [0.01, 0.1, 1.0]
    },
    'catboost': {
        'iterations': [100, 200, 300],
        'depth': [6, 8, 10],
        'learning_rate': [0.01, 0.05, 0.1]
    },
    'xgboost': {
        'n_estimators': [100, 200, 300],
        'max_depth': [3, 6, 9],
        'learning_rate': [0.01, 0.05, 0.1]
    }
}
```

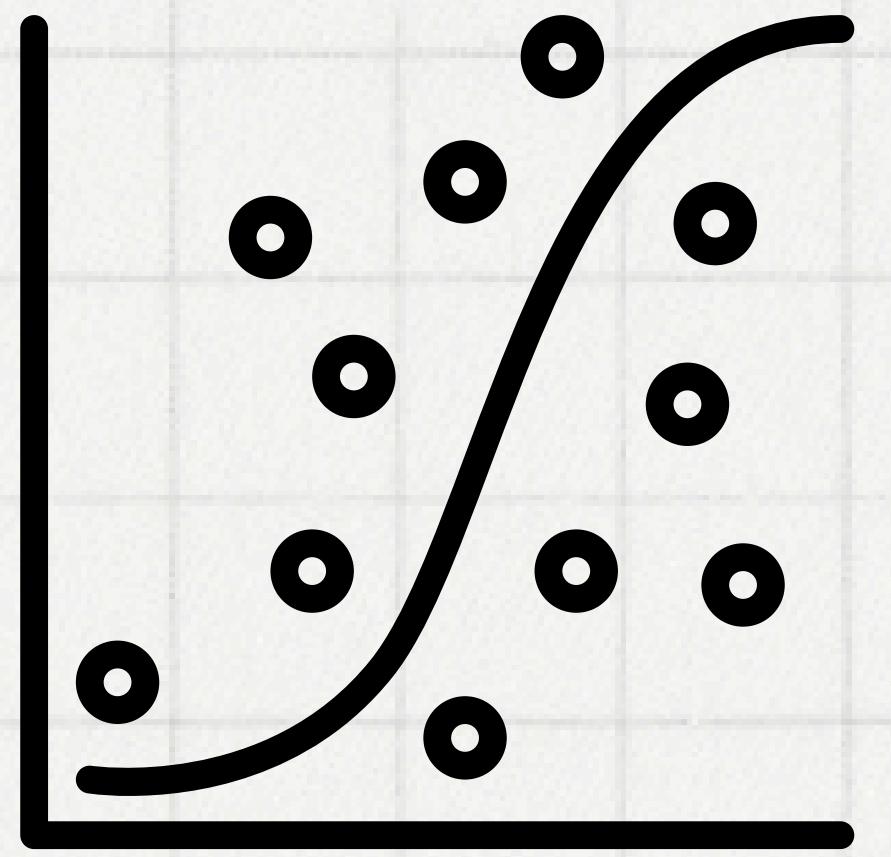
Hasil Klasifikasi

Model	F1_micro	F1_macro	Recall	Precision	Accuracy
Logistic Regression	0.882517	0.401322	0.882517	0.882517	0.882517
Decision Tree	0.898049	0.596919	0.898049	0.898049	0.898049
Random Forest	0.916368	0.613974	0.916368	0.916368	0.916368
AdaBoost	0.908801	0.597575	0.908801	0.908801	0.908801
CatBoost	0.930705	0.715098	0.930705	0.930705	0.930705
XGBoost	0.931103	0.729718	0.931103	0.931103	0.931103

Analisis

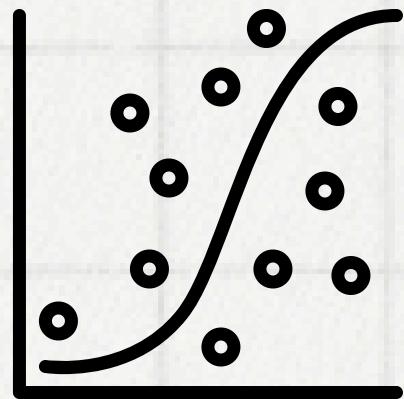
XGBoost dengan Hypertuning muncul sebagai model dengan performa terbaik dibandingkan dengan model lain, mencapai F1-mikro sebesar 0.931 dan F1-makro sebesar 0.729. Ini menunjukkan kemampuan yang kuat dalam menangkap hubungan dalam dataset dan membuat prediksi yang akurat.

XGBoost sangat efektif untuk klasifikasi multikelas yang tidak seimbang karena kerangka peningkatan gradiennya, yang secara berurutan membangun pohon yang memperbaiki kesalahan sebelumnya, sehingga meningkatkan akurasi. Teknik regularisasinya mencegah overfitting, dan kemampuan bawaannya menangani nilai yang hilang dan data yang jarang secara efisien. XGBoost juga mendapat manfaat dari pemrosesan paralel, yang mempercepat pelatihan, dan pemangkasan pohon yang canggih untuk mengontrol kompleksitas model. Selain itu, skalabilitasnya untuk kumpulan data besar dan fleksibilitas dengan penyetelan hyperparameter dan tujuan khusus menjadikannya kuat dan mudah beradaptasi dengan berbagai skenario klasifikasi yang tidak seimbang, seringkali mengungguli algoritme lainnya.



Regresi





Regresi

Melakukan fine tuning beberapa model regressor.

```
param_grid = {
    'logistic_regression': {
        'C': [0.1, 1.0, 10.0],
        'penalty': ['l2']
    },
    'decision_tree': {
        'max_depth': [None, 10, 20, 30],
        'min_samples_split': [2, 5, 10]
    },
    'random_forest': {
        'n_estimators': [100, 200, 300],
        'max_depth': [None, 10, 20, 30],
        'min_samples_split': [2, 5, 10]
    },
    'adaboost': {
        'n_estimators': [50, 100, 200],
        'learning_rate': [0.01, 0.1, 1.0]
    },
    'catboost': {
        'iterations': [100, 200, 300],
        'depth': [6, 8, 10],
        'learning_rate': [0.01, 0.05, 0.1]
    },
    'xgboost': {
        'n_estimators': [100, 200, 300],
        'max_depth': [3, 6, 9],
        'learning_rate': [0.01, 0.05, 0.1]
    }
}
```

Hasil Regresi

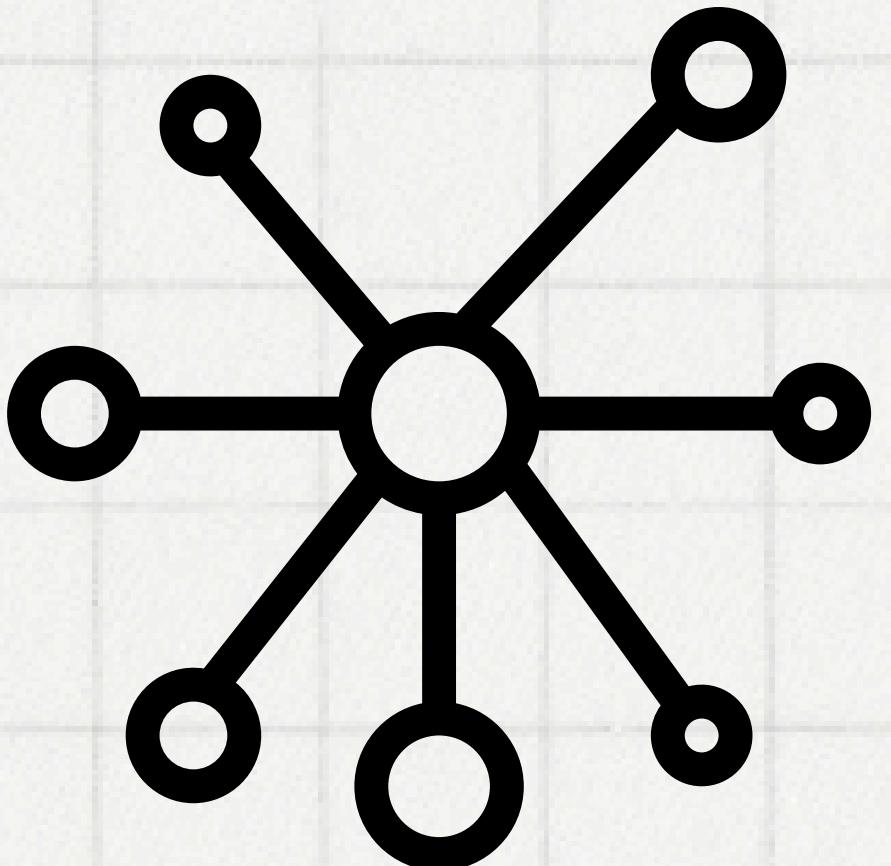
Model	R_squared	MAE	MSE	RMSE
Linear Regression	0.622189	1.559608	5.527651	2.351096
Ridge	0.622809	1.557078	5.518583	2.349166
Lasso	0.616477	1.569595	5.611225	2.368802
Decision Tree	0.546113	1.572977	6.640701	2.576956
Random Forest	0.645944	1.431099	5.180091	2.275981
AdaBoost	0.574033	1.746618	6.232208	2.496439
CatBoost	0.655253	1.454391	5.043896	2.245862
XGBoost	0.623223	1.497932	5.512526	2.347877

Analisis

CatBoost dengan Hypertuning muncul sebagai model dengan performa terbaik dibandingkan dengan model lain, mencapai R-squared sebesar 0.66 dan MAE sebesar 1.45. Ini menunjukkan kemampuan yang kuat dalam menangkap hubungan dalam dataset dan membuat prediksi yang akurat.

Karena dataset Health Admission merupakan data yang mayoritas bersifat kategorikal (Apakah pasien menderita penyakit ini, Ya/Tidak). CatBoost sangatlah cocok untuk dataset ini karena CatBoost memang didesain khusus untuk data kategorikal. Mengingat Cat pada CatBoost berasal dari kata Categorical.

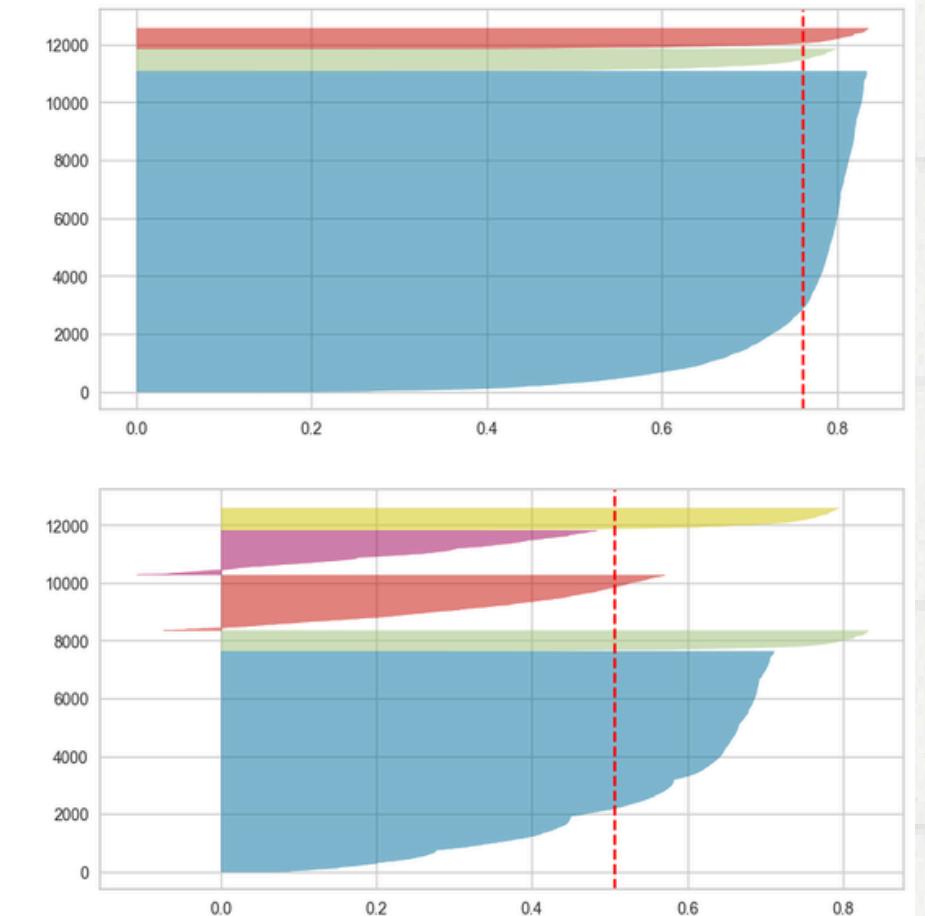
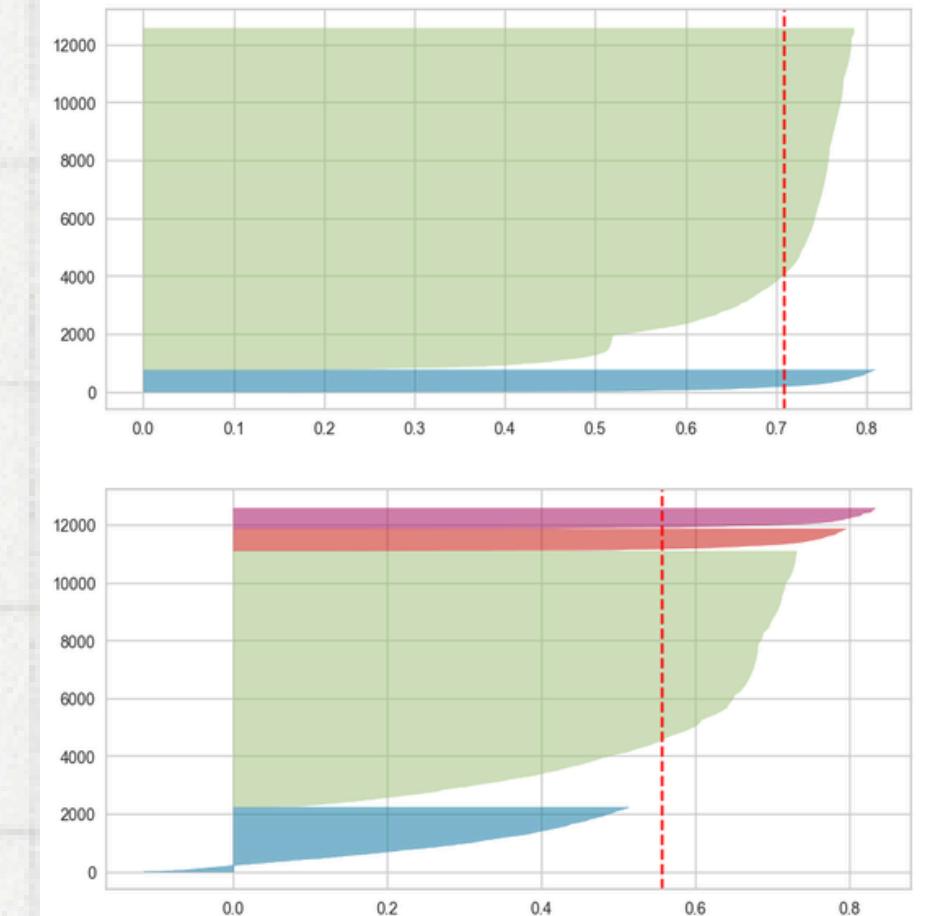
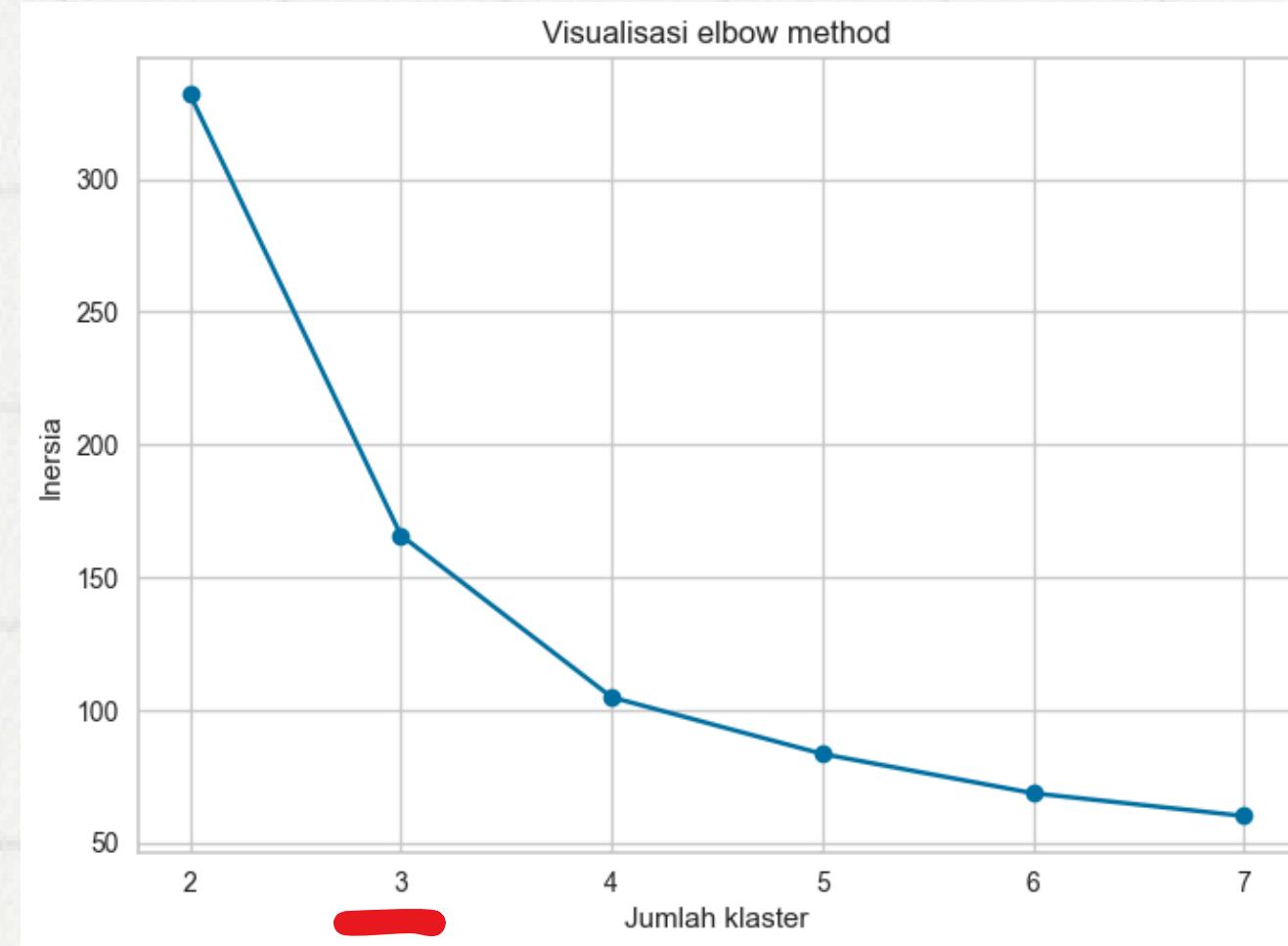
CatBoost juga memiliki berbagai mekanisme untuk mencegah overfitting, seperti ordered boosting dan shrinkage, yang dapat menghasilkan metrik performa yang lebih baik seperti R-squared.



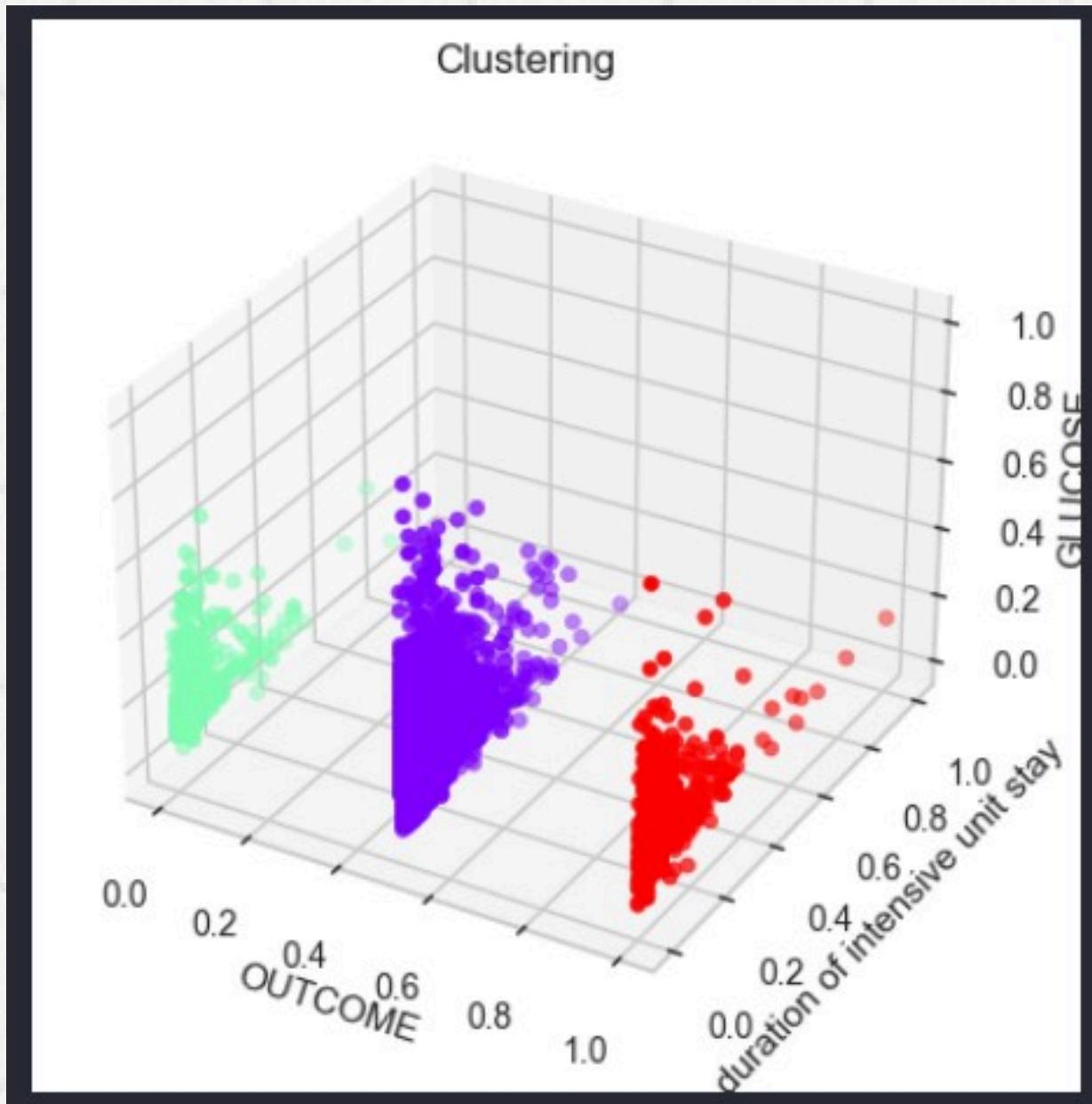
Clustering



Hasil Clustering



Hasil Clustering



Didapat k=3

```
For n_clusters = 2 The average silhouette_coefficient is : 0.7095441406018378  
For n_clusters = 3 The average silhouette_coefficient is : 0.761171030991885  
For n_clusters = 4 The average silhouette_coefficient is : 0.5573447705840525  
For n_clusters = 5 The average silhouette_coefficient is : 0.5063175499943765
```

Interpretasi Clustering

Dapat dilihat bahwa terdapat 3 kluster yang ditandai dengan warna yang berbeda. Pembeda utama antara cluster adalah Outcome. 2 Variabel lain yaitu glukosa dan duration of intensive unit care berbeda.

Cluster hijau merepresentasikan Outcome DAMA

Cluster Ungu merepresentasikan Outcome Discharge

Cluster merah merepresentasikan Outcome Expiry

Variabel lain kurang berpengaruh dalam clustering ini.

Terima Kasih

