REDUCTION OF HOSPITAL BED OCCUPANCY RATE

ISSS621 DATA SCIENCE FOR BUSINESS

G2 Group 4: Huang Weiju, Lee Chiayin, Lin Lin, Tan Hui Wen, Zhu Fangyuan

AGENDA

Overview

Business Problem

Translation into Data Science Task

Dataset

Frequent Pattern Mining

Discriminative Pattern Discovery

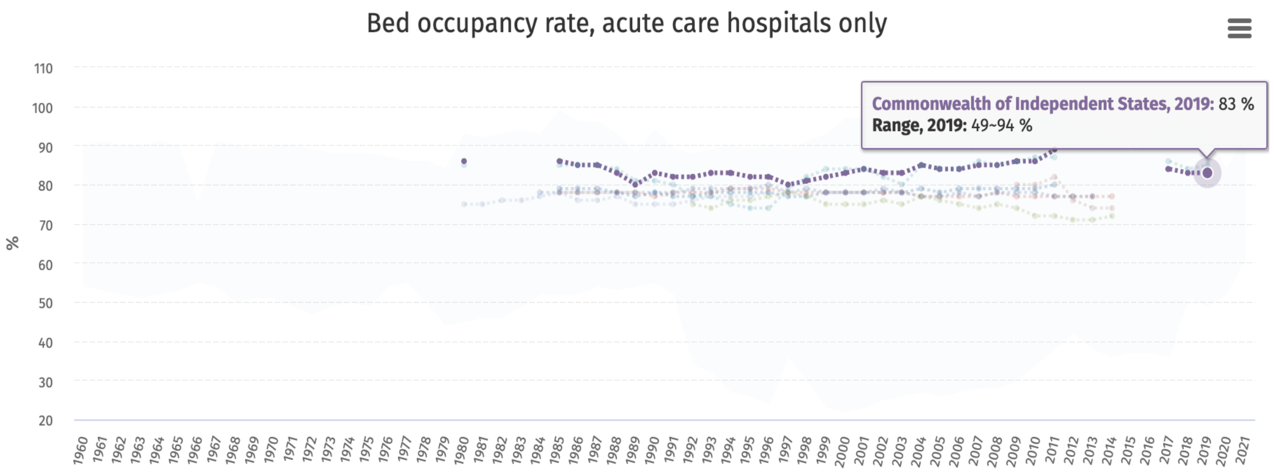
Predictive Analysis – Decision Tree

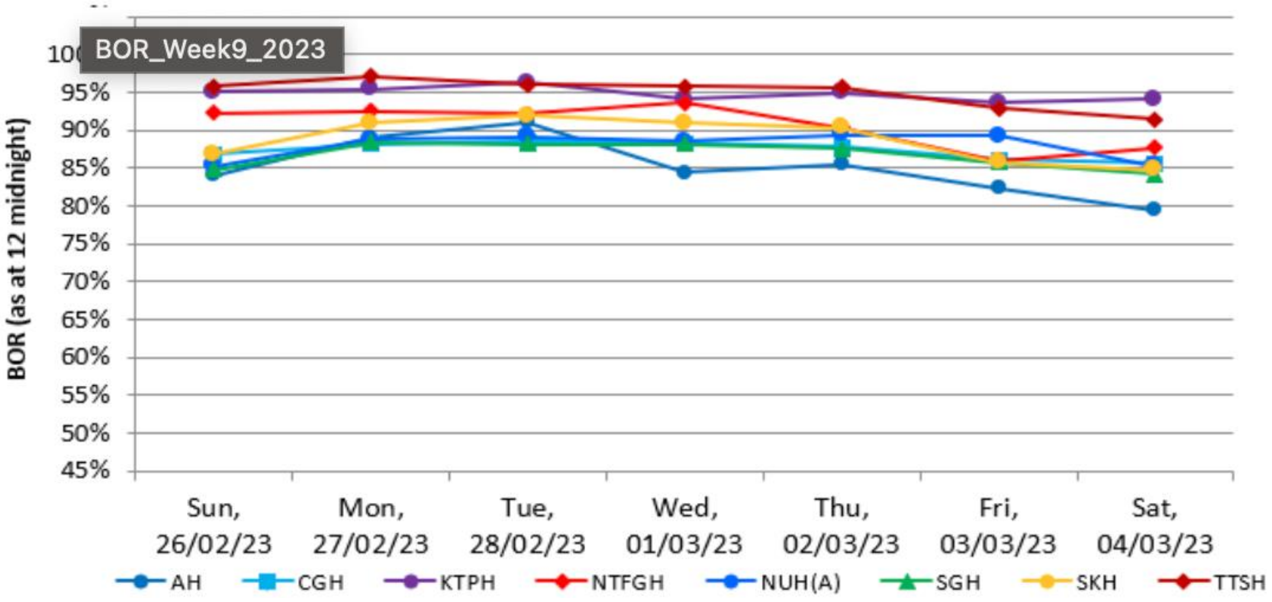
Closed-Loop & Future Work

Data Governance

OVERVIEW

- Worldwide,
  high Bed Occupancy Rate has been a problem especially in commonwealth countries and central Asia while below 85% is recommended
- Singapore has an average of around 92% BOR
- MOH aims to lower the BOR to 80% in the next 5 years
- COVID-19's impact in the hospital industry





BUSINESS PROBLEM DATA SCIENCE APPROACH

The Problem

- High BOR
- Different factors influencing BOR
- Inefficient allocation of resources in hospitals
- Long waiting time due to low bed turnover rates
- Future complications caused by high BOR

Our Approach

- Goal: optimize the use of hospital resources, lower BOR
- Analyze and predict LOS
- Frequent Pattern Mining
- Discriminative Pattern Discovery
- Decision Tree & other prediction methods

DATA SOURCES

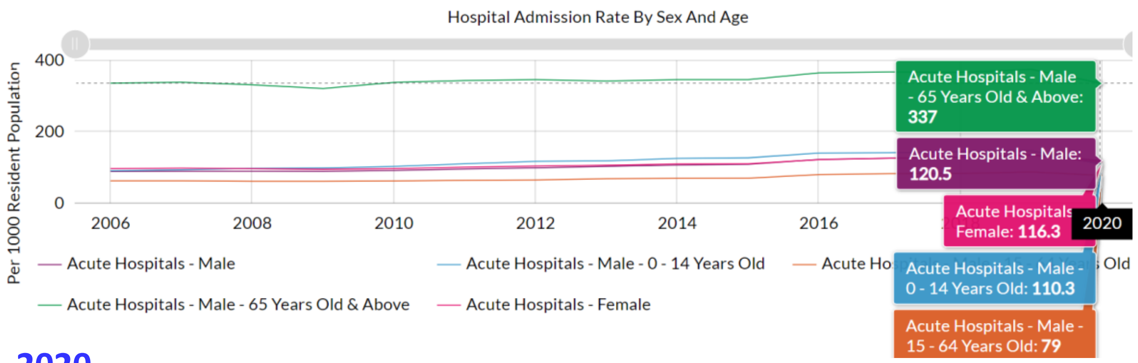
MINISTRY OF HEALTH
SINGAPORE

Data.gov.sg

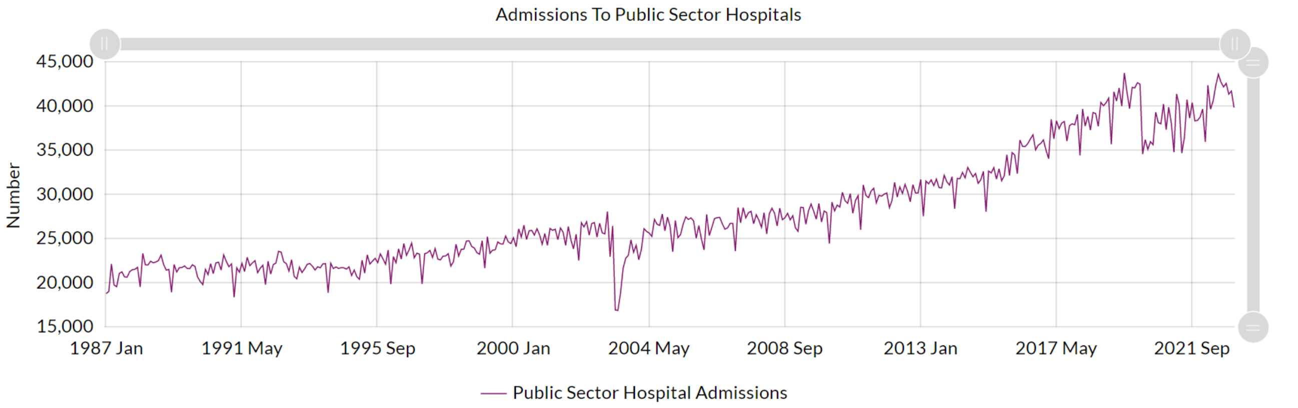
DEPARTMENT OF
STATISTICS
SINGAPORE
Empowering You with Trusted Data

SINGAPORE'S STATISTICS

(All the charts are obtained from the data sources.)



[2020](#)
Hospital Admissions per 1000 resident population is highest for Male – 65 years old & above.



Jan 2023
Public Sector Hospital Admissions = 39,878

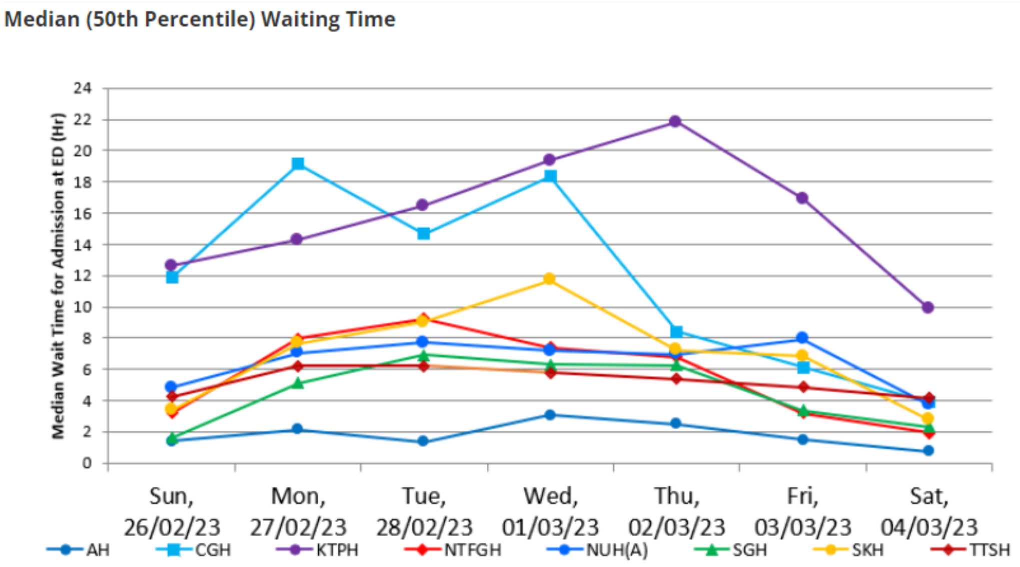
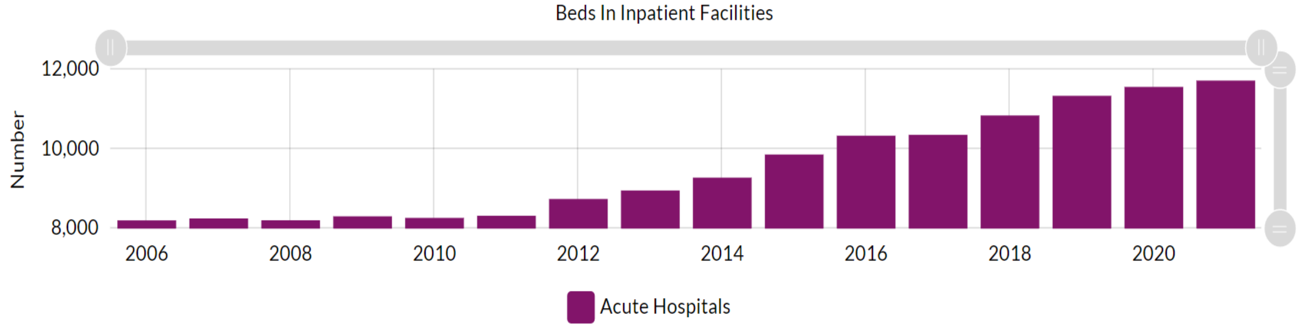


Chart: Daily Median Waiting Time for Admission at ED (26 Feb 2023 - 4 Mar 2023)

2023
Median waiting time for admission typically within half a day.



2021
Number of hospital beds = 11,704

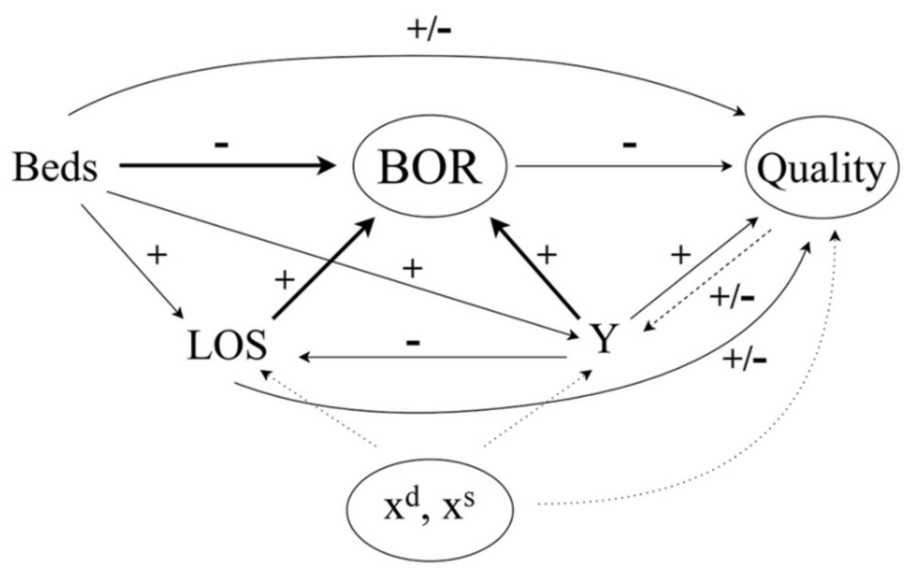
BED OCCUPANCY RATE (BOR)

- Defined as the average number of days when hospital bed was occupied as % of available 365 days.

- Formula to calculate BOR (%):
  BOR = (Utilized bed-days / available bed-days during the calendar year) x 100%

FACTORS INFLUENCING BOR

- Hospital size measured by number of hospital beds
- Locations of hospitals
- Demographics of the growing populations
- Allocations of hospital beds by type of facilities (e.g., Orthopaedic, Ear, Nose & Throat, etc) within the hospitals
- Length of inpatient stay



Conceptual framework. BOR bed occupancy rate; LOS length of stay, Y volume, xd demand shifters, xs supply shifters

Source: The European Journal of Health Economics

LENGTH OF STAY (LOS)

- Defined as a clinical metric that measures the length of time elapsed between a patient's hospital admittance and discharge.

- Critical parameter used to measure the efficiency of the healthcare management in a hospital and can help hospitals identify patients who will need to stay longer at the time of admission.

- LOS correlates to BOR and is one of the quantifiable measures of BOR.

- Parameter to focus on for this business problem is **<u>Average LOS of an inpatient</u>**.

- 'Average' is used with the underlying assumption that all hospital stays by an inpatient are treated to be the same.

TRANSLATING BUSINESS PROBLEM TO DATA SCIENCE TASK

Business Problem

Optimize the usage of hospital resources, specifically looking at the reduction of bed occupancy rate.

Data Science Task

Analyse and predict average length of inpatient stay.

DATA SOURCES

World Health Organization
European Region

kaggle

THE WORLD BANK
IBRD • IDA
Data

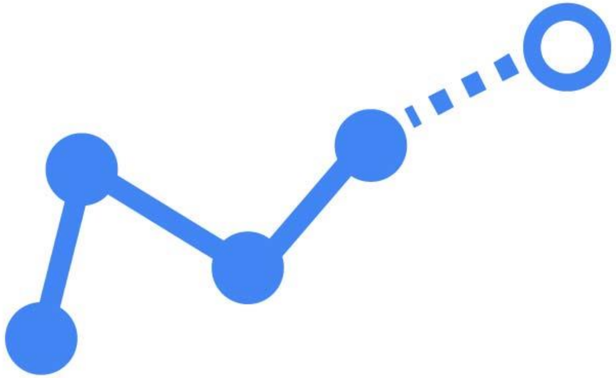
DATASET

Attribute	Description	Data Type
case_id	Case_ID registered in Hospital	Numerical
Hospital_code	Unique code for the Hospital	Numerical
Hospital_type_code	Unique code for the type of Hospital	Categorical
City_Code_Hospital	City Code of the Hospital	Numerical
Hospital_region_code	Region Code of the Hospital	Categorical
Available Extra Rooms in Hospital	Number of Extra rooms available in the Hospital	Numerical
Department	Department overlooking the case	Categorical
Ward_Type	Code for the Ward type	Categorical
Ward_Facility_Code	Code for the Ward Facility	Categorical
Bed Grade	Condition of Bed in the Ward	Numerical
patientid	Unique Patient Id	Numerical
City_Code_Patient	City Code for the patient	Categorical
Type of Admission	Admission Type registered by the Hospital	Categorical
Severity of Illness	Severity of the illness recorded at the time of admission	Categorical
Visitors with Patient	Number of Visitors with the patient	Numerical
Age	Age of the patient	Categorical
Admission_Deposit	Deposit at the Admission Time	Numerical
Stay	Stay Days by the patient	Categorical

DATA ANALYSIS WITH DATA SCIENCE CONCEPT/NOTION

Frequent Pattern Mining

Discriminative Pattern Discovery



Predictive Modeling - Decision Tree

DATA CLEANING - Dealing with data attributes & missing values

Dealing with ID data

Our raw data has 318438 records and 18 columns in total, in which 2 columns ***patientid*** and ***case_id*** are ID numbers.

Resolution

We directly remove them because they carry no information.

Dealing with missing data

There are 2 variables that have missing values. One is ***Bed Grade*** with 113 (0.04% of the total records) missing values and ***City_Code_Patient*** with 4532 (1.42% of the total records) missing values.

Resolution

since the low rate of missing values to the whole records, we choose to remove the records with missing values of ***Bed Grade.***

For ***City_Code_Patient,*** we chose to remove this column because the number of missing is more than the amount of one of its categories, so it's hard to interpolate them with other values.

The **reason why** we don't take **missing value** as a **category**?
Our method to clean the data is **mainly Frequent Pattern Mining**, which has been developed for transaction database where no missing value exists.

```
In [4]: hel.isnull().sum()

Out[4]: case_id                              0
        Hospital_code                        0
        Hospital_type_code                   0
        City_Code_Hospital                   0
        Hospital_region_code                 0
        Available Extra Rooms in Hospital    0
        Department                           0
        Ward_Type                            0
        Ward_Facility_Code                   0
        Bed Grade                          113
        patientid                            0
        City_Code_Patient                 4532
        Type of Admission                    0
        Severity of Illness                  0
        Visitors with Patient                0
        Age                                  0
        Admission_Deposit                    0
        Stay                                 0
        dtype: int64
```

DATA CLEANING - Recoding data

- Dealing with continuous data

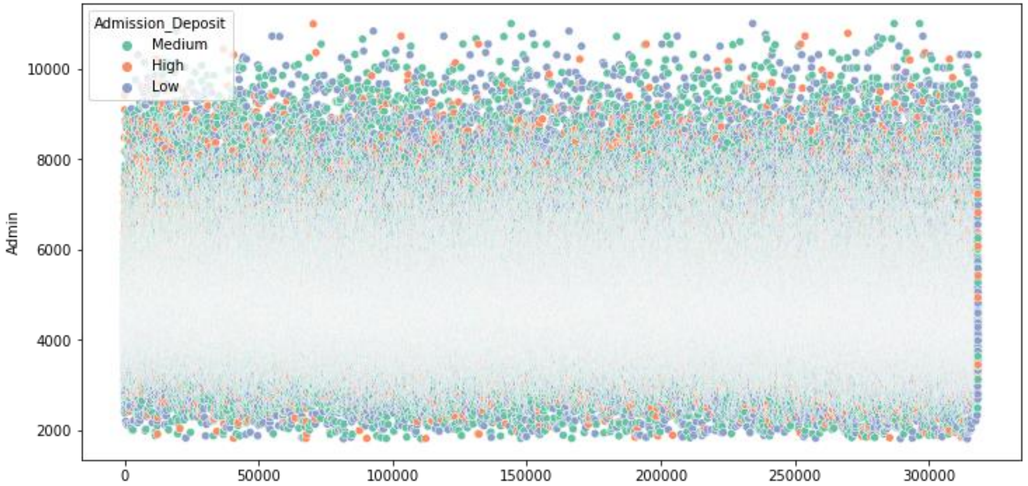
In order to use Frequent Pattern Mining we need to divide continuous variable ***Admission_Deposit*** into categories.

Resolution

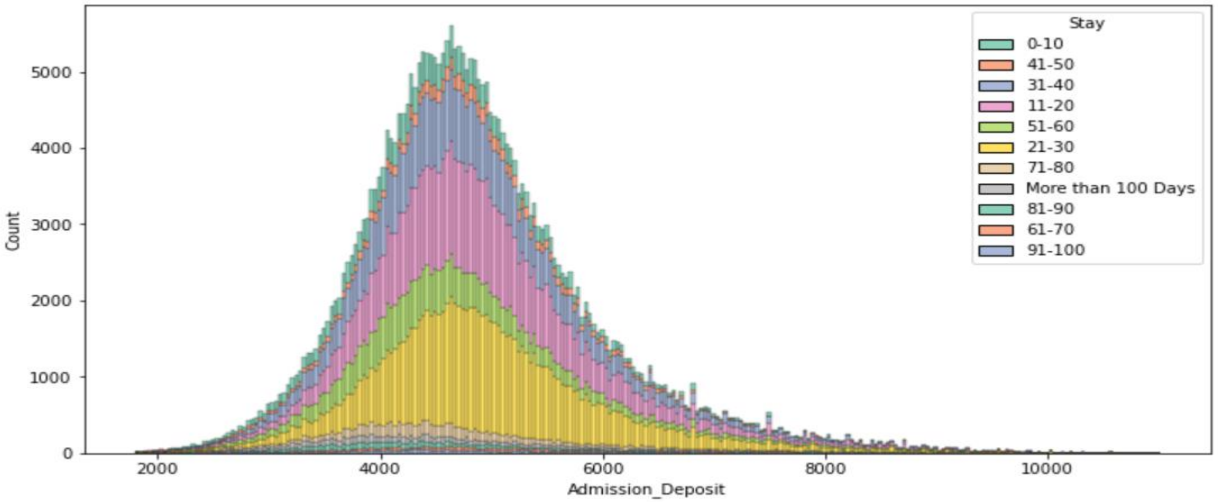
We divide this variable *Admission_Deposit* into 3 categories – "High", "Medium" and "Low".

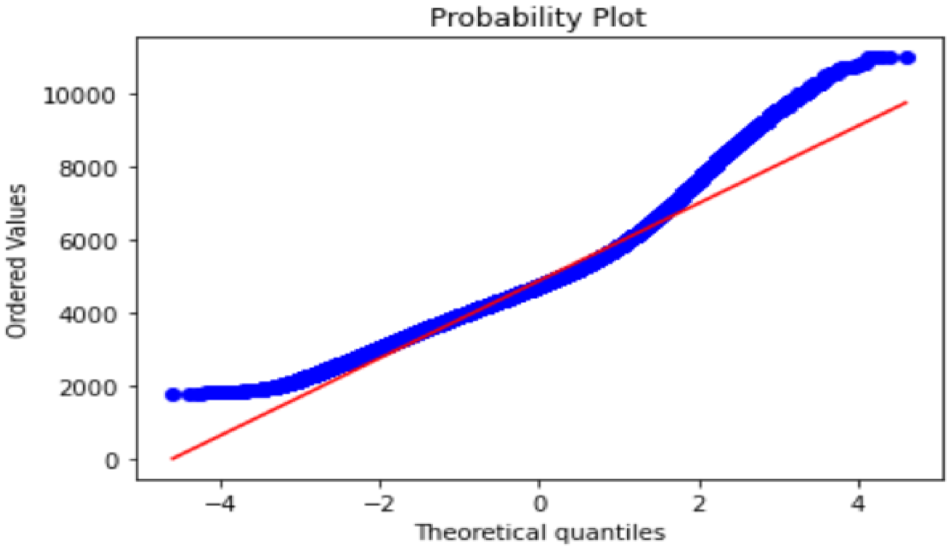
We finally choose k-means because the distribution of this variable is not highly skewed and close to normal distribution.

K-Means clustering outcome



Distribution of Admission Deposit





FREQUENT PATTERN MINING- Recoding data

- Recode Discrete/nominal data

We found that there are discrete/nominal variables ***Hospital_code, City_Code_Hospital, Bed_Grade_mapping*** and ***Visitors with Patient.*** There are many overlapped values which cannot be used in Frequent Pattern Mining.

Resolution

We recode *Hospital_code, City_Code_Hospital* and *Visitors with Patient* data by changing their data scales, to make them no overlaps with each other. And recode *Bed_Grade_mapping* by adding a string to each value. Lastly we change these discrete/nominal data into string to be sorted in the later used algorithm.

- Recode ariables with same values

We found that there are same items in variables ***Stay*** and ***Age***, which also cannot be used in Frequent Pattern Mining.

Resolution

We recode *Stay* by adding a "d" to the end of each item to differentiate items of this variable from those of *Age*.

```
In [55]: tr_hel.describe()
Out[55]:
```

	Hospital_code	City_Code_Hospital	Available Extra Rooms in Hospital	Visitors with Patient
count	318325.000000	318325.000000	318325.000000	318325.000000
mean	118.323163	44.771297	63.197509	3.284153
std	8.632207	3.102999	1.168208	1.764190
min	101.000000	41.000000	60.000000	0.000000
25%	111.000000	42.000000	62.000000	2.000000
50%	119.000000	45.000000	63.000000	3.000000
75%	126.000000	47.000000	64.000000	4.000000
max	132.000000	53.000000	84.000000	32.000000

```
Stay
0-10                   23602
11-20                  78120
21-30                  87454
31-40                  55137
41-50                  11735
51-60                  35005
61-70                   2740
71-80                  10250
81-90                   4837
91-100                  2764
More than 100 Days      6681
Name: Stay, dtype: int64
Number of types: 11
```

```
Age
0-10      6254
11-20    16763
21-30    40828
31-40    63613
41-50    63716
51-60    48497
61-70    33681
71-80    35784
81-90     7887
91-100    1302
Name: Age, dtype: int64
Number of types: 10
```

FREQUENT PATTERN MINING - Filter out L2 frequent records

- Expected outcome of FP in our project

Generally in the process of modeling, we will check the outliers to reduce some anomalies. In our business case, our target variable is ***Length of Stay***, so it is meaningless and impractical to focus on those cases with fairly long length of hospitalization. Therefore, we choose to use the **Frequent Pattern Mining** to filter out those records with low frequent stay of length.

```
frozenset({'11-20 d', '41-50'})
frozenset({'4', '11-20 d'})
frozenset({'21-30 d', '41-50'})
frozenset({'31-40 d', '63'})
frozenset({'21-30 d', '4'})
frozenset({'21-30 d', '41'})
frozenset({'21-30 d', 'E'})
frozenset({'21-30 d', '3'})
frozenset({'21-30 d', 'Q'})
frozenset({'E', '51-60 d'})
```

```
Stay
0-10     22873
11-20    78120
21-30    87454
31-40    55137
51-60    35001
Name: Stay, dtype: int64
```

L2 frozenset{(Item , Stay)}

Resolution

- **Step1** - Setting a **minimal support threshold**

Given that the data has totally 318325 records and 15 columns, and the total number of items of all variables is 146. the running speed is quite slow if we set the minimal support threshold too small. But high threshold will miss a lot of information. After trying for several times, we chose 0.04 as a minimal support threshold, meaning that items more than 318325*0.04 = 12733 are frequent.

Resolution

- **Step2** – Filtering out records by target variable in **frequent L2 items**

We believe that records with frequent *Length of Stay* and at least one item other variables are meaningful for predictive modeling. So we only focus on L2 frequent sets, where there are 2 items in a frozenset, and only retain those containing Length of *Stay*. Thus we filtered totally 278585 records and 5 out of 11 kinds of *Length of Stay*.

FREQUENT PATTERN MINING - Conclusion

Variable	Initial values	Filtered values	Category changes
Available Extra Rooms in Hospital	0,…,14, 20, ~~21~~,24	0,…,14, 20, 24	18 → 17
Visitors with Patient	0,…, ~~23, 24, 25, 30, 32~~	0,…,18, 20, 22	28 → 20
Admission_ Deposit	Min = 1800 Max = 11008 Mean = 4881	Min = 1801 Max = 11008 Mean = 4900	
Stay	"0-10","11-20", …, ~~"61-70",…, "91-100","More than100 Days"~~	"0-10","11-20", "21-30","31-40","41-50", "51-60"	11 → 5
			Records: 318438 → 278585

- Data category is narrowed down

After being filtered by L2 frequency, we have narrowed down categories of the left 4 attributes. Significantly, value category of ***Visitors with Patient*** is narrowed to 0-22, which makes sense because it is rare to have more than 20-30 visitors to the patient.

And the ***stay of length*** is narrowed to less than 60 days. That's also rational because it is infeasible to forecast more than 60 stay of length.

- Total records is narrowed down

The total records of the data are also narrowed down for sure after being filtered by L2 frequency.

Able to identify Type of Admission with Bed Grade & Stay

- Type of Admission
- Stay
- Stay (Recoded)
- Department
- Admission_Deposit (kmean = 4)
- Admission_Deposit (quantile)
- Age
- Age (Recoded)

```
Frequent itemsets of Type of Admission in Trauma :
frozenset({'gynecology'})
frozenset({3.0})
frozenset({2})
```

Bed Grade

```
Frequent itemsets of Type of Admission in Emergency :
frozenset({'gynecology'})
frozenset({2.0})
frozenset({3})

Frequent itemsets of Type of Admission in Urgent :
frozenset({'gynecology'})
frozenset({3})
frozenset({'< 1m'})
frozenset({2})
frozenset({2, 'gynecology'})
```

Stay

Patients staying < 1 m tend to get Bed 2.0 and not necessary the least visitors

- Type of Admission
- Stay
- Stay (Recoded)
- Department
- Admission_Deposit (kmean = 4)
- Admission_Deposit (quantile)
- Age
- Age (Recoded)

```
Frequent itemsets of Stay in 2 - 3m :
frozenset({'gynecology'})
frozenset({3.0})
frozenset({2})

Frequent itemsets of Stay in > 3m :
frozenset({2})
frozenset({'gynecology'})
frozenset({3})

Frequent itemsets of Stay in 1 - 2m :
frozenset({'gynecology'})
frozenset({3.0})
frozenset({4})
frozenset({2})

Frequent itemsets of Stay in < 1m :
frozenset({'gynecology'})
frozenset({2.0})
frozenset({3})
frozenset({2.0, 'gynecology'})
```

Visitors with Patient

Bed Grade

Patients coming for anesthesia stay < 1 m, and for surgery are admitted in emergency

- Type of Admission
- Stay
- Stay (Recoded)
- Department
- Admission_Deposit (kmean = 4)
- Admission_Deposit (quantile)
- Age
- Age (Recoded)

```
Frequent itemsets of Department in surgery :
frozenset({'Emergency'})
frozenset({2.0})
frozenset({2.0, 'Emergency'})
```

Type of Admission

```
Frequent itemsets of Department in gynecology :
frozenset({3.0})
frozenset({2})

Frequent itemsets of Department in TB & Chest disease :
frozenset({2})

Frequent itemsets of Department in radiotherapy :
frozenset({2.0})

Frequent itemsets of Department in anesthesia :
frozenset({3.0})
frozenset({'< 1m'})
frozenset({'Trauma'})
frozenset({2})
```

Stay, Type of Admission

Patients w/ 50-75% admission deposit tend to stay < 1 m

- Type of Admission
- Stay
- Stay (Recoded)
- Department
- Admission_Deposit (kmean = 4)
- Admission_Deposit (quantile)
- Age
- Age (Recoded)

```
Frequent itemsets of Admission_Deposit in 0–25% :
frozenset({2})
frozenset({'gynecology'})

Frequent itemsets of Admission_Deposit in 75–100% :
frozenset({'gynecology'})
frozenset({3.0})
frozenset({2})

Frequent itemsets of Admission_Deposit in 25–50% :
frozenset({'gynecology'})
frozenset({3})
frozenset({2})

Frequent itemsets of Admission_Deposit in 50–75% :
frozenset({'gynecology'})
frozenset({'< 1m'})
frozenset({2.0})
frozenset({3})
```

Stay

Less possible for patients > 80 y coming for gynecological problems, more possible for patients < 20 y having visitors and staying < 1 m

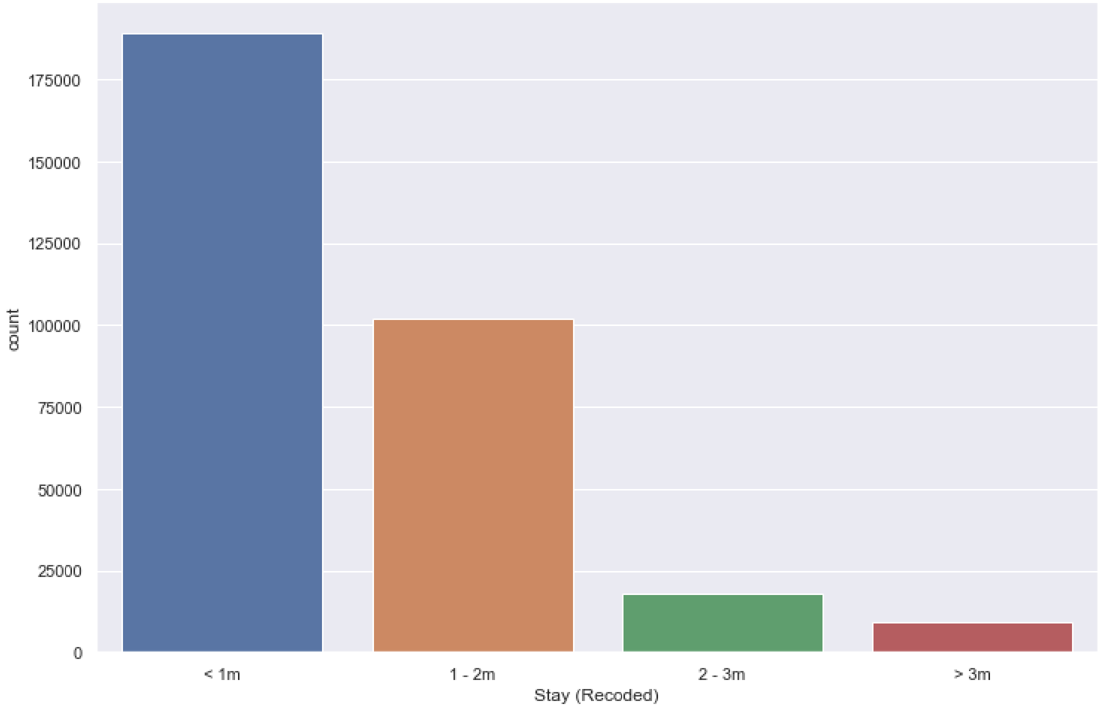
- Type of Admission
- Stay
- Stay (Recoded)
- Department
- Admission_Deposit (kmean = 4)
- Admission_Deposit (quantile)
- Age
- Age (Recoded)

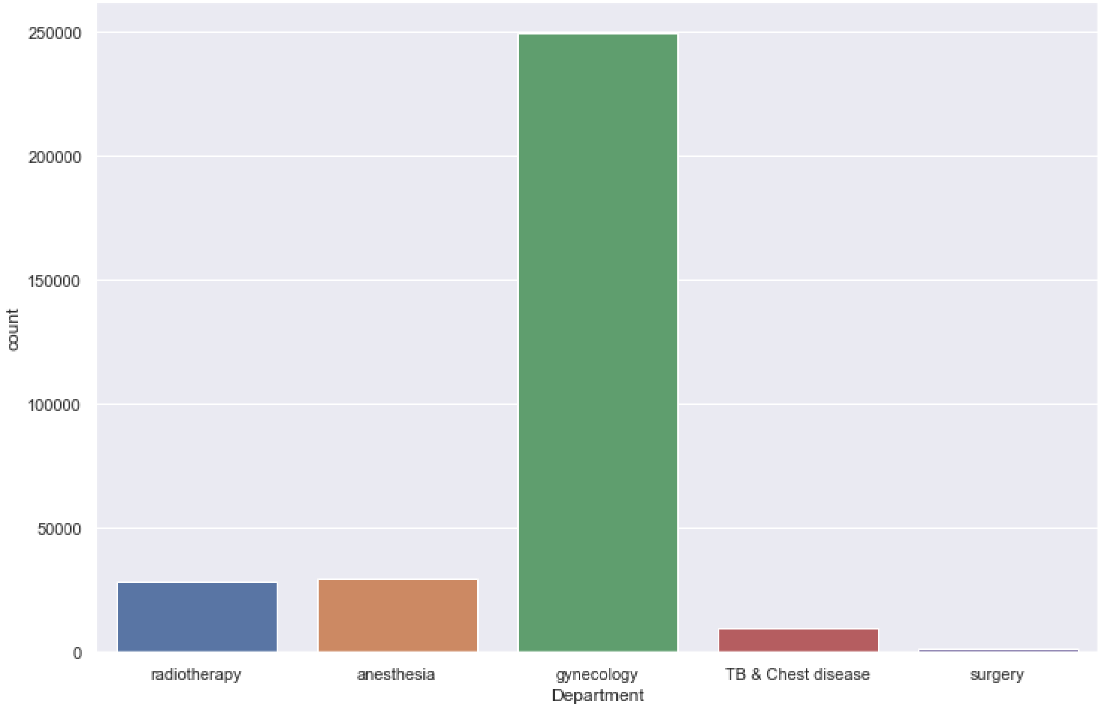
```
Frequent itemsets of Age in mature :
frozenset({2})
frozenset({'gynecology'})

Frequent itemsets of Age in senior :
frozenset({2.0})

Frequent itemsets of Age in middle age :
frozenset({'gynecology'})
frozenset({2.0})

Frequent itemsets of Age in child :
frozenset({'< 1m'})
frozenset({'gynecology'})
frozenset({3})
frozenset({2})

Frequent itemsets of Age in young adult :
frozenset({3.0})
frozenset({2.0})
frozenset({'gynecology'})
frozenset({2.0, 'gynecology'})
```

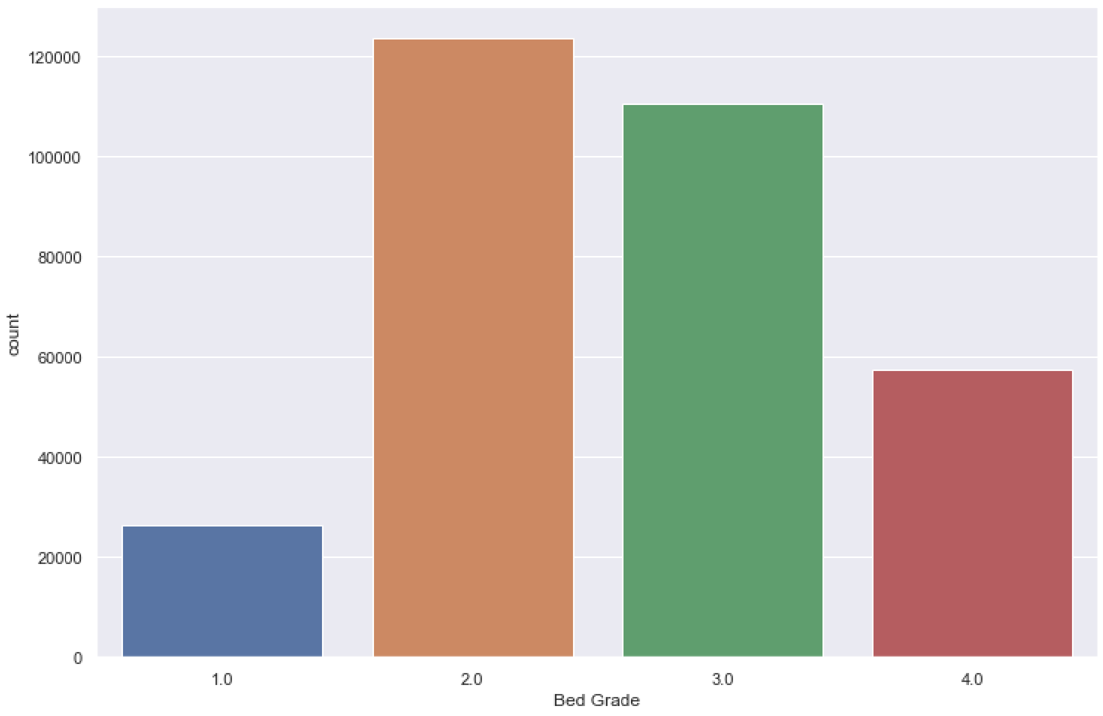
Department

Stay

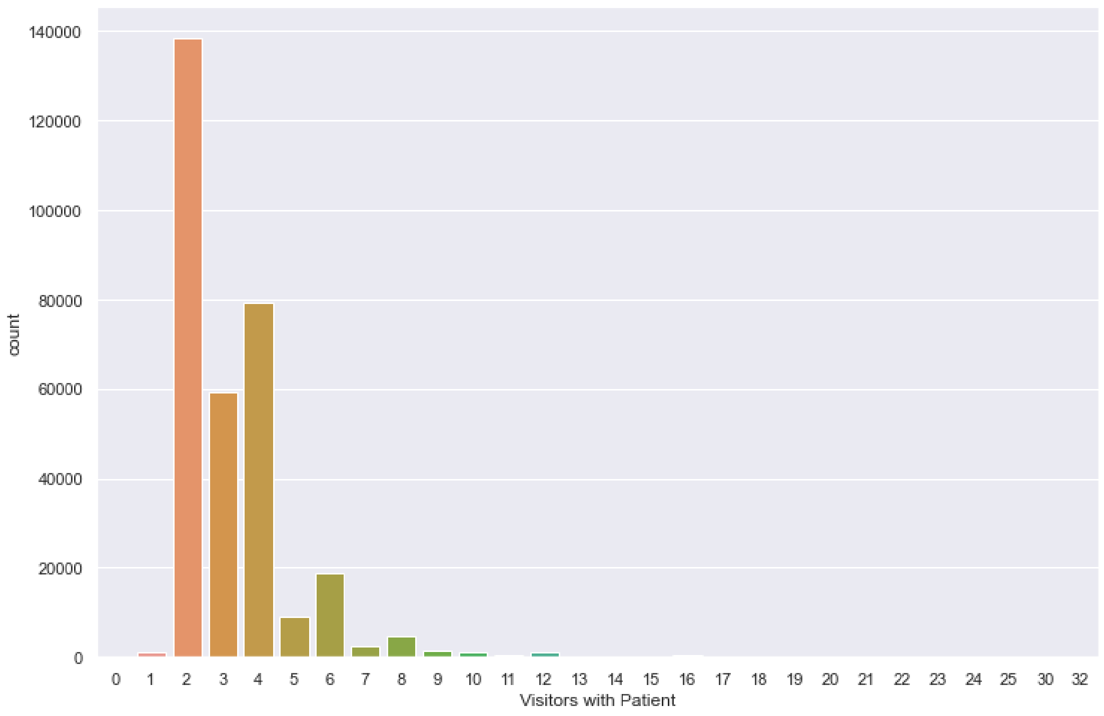
Visitors with Patient


175000
150000
125000
100000
75000
50000
25000
0
count
< 1m
1 - 2m
2 - 3m
> 3m
Stay (Recoded)


250000
200000
150000
100000
50000
0
count
radiotherapy
anesthesia
gynecology
TB & Chest disease
surgery
Department

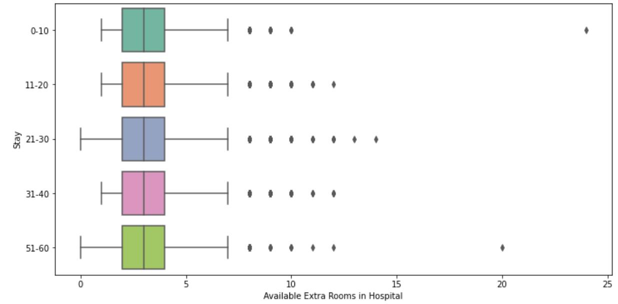


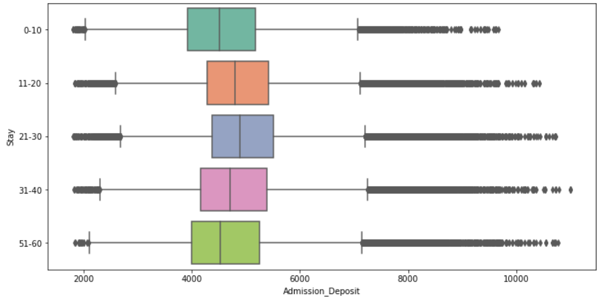
count
0
20000
40000
60000
80000
100000
120000
1.0
2.0
3.0
4.0
Bed Grade

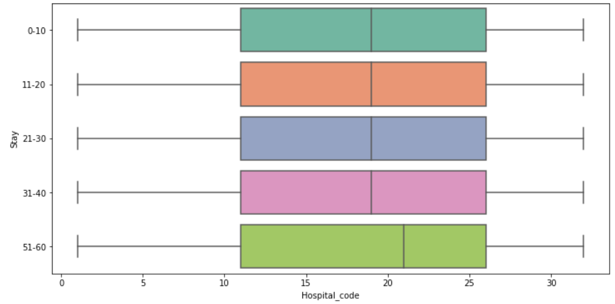


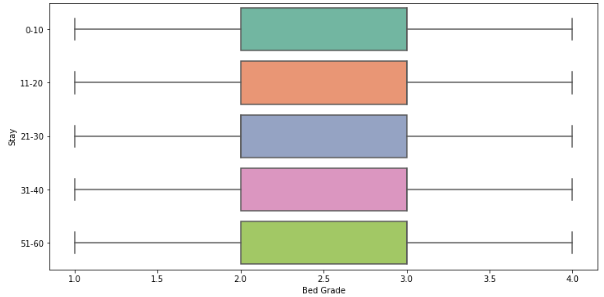
count
0
20000
40000
60000
80000
100000
120000
140000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 30 32
Visitors with Patient

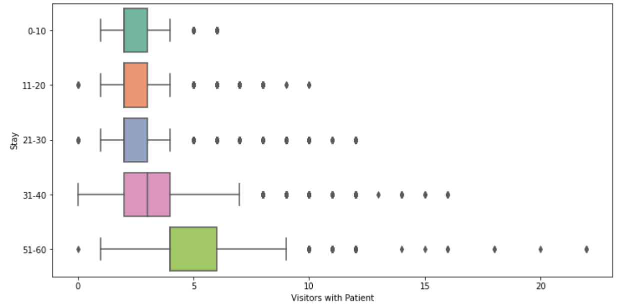
CATEGORICAL DATA ARE ALL SIGNIFICANT

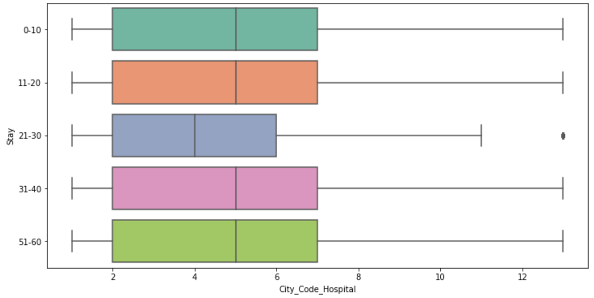






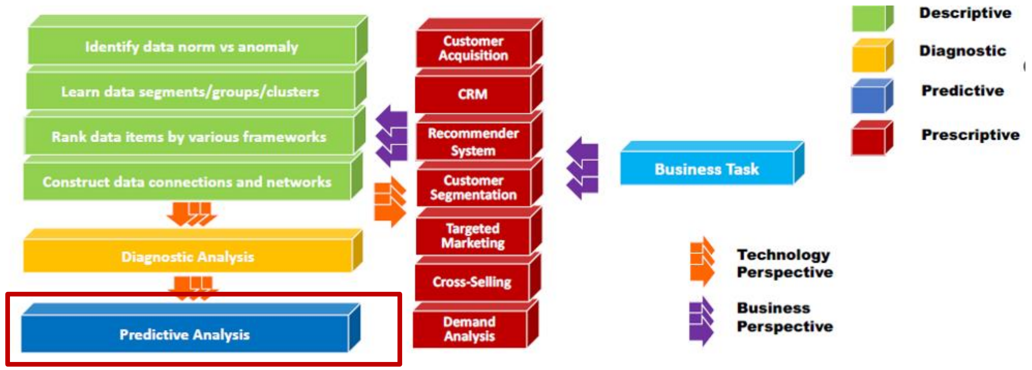




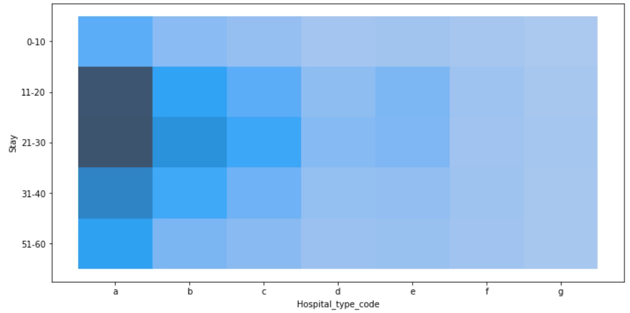


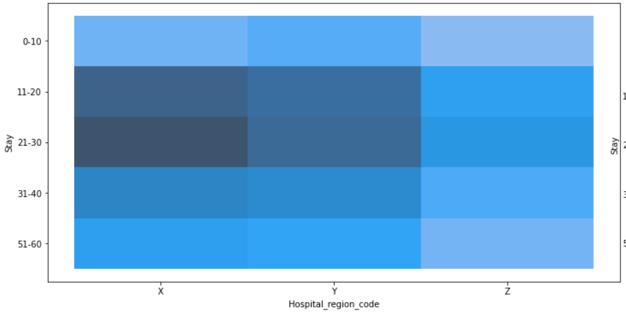

```
Stay                  0-10   11-20  21-30  31-40  51-60
Severity of Illness
Extreme               3261   10515  15487  10076   7772
Minor                 7691   27073  21526  14447   7124
Moderate             11921   40532  50441  30614  20105
P value for Severity of Illness 0.0
P-value for Hospital_type_code vs "Stay": 0.0000
P-value for Hospital_region_code vs "Stay": 0.0000
P-value for Department vs "Stay": 0.0000
P-value for Ward_Type vs "Stay": 0.0000
P-value for Ward_Facility_Code vs "Stay": 0.0000
P-value for Type of Admission vs "Stay": 0.0000
P-value for Severity of Illness vs "Stay": 0.0000
P-value for Age vs "Stay": 0.0000
```

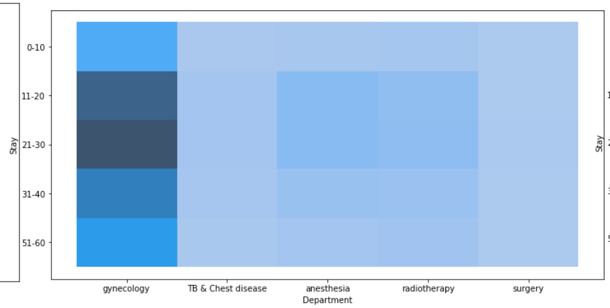
- Chi Square test: $P < 0.05$

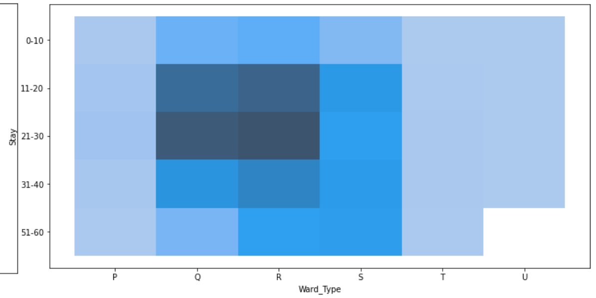


HEATMAPS OF CATEGORICAL DATA – HOW DO THEY AFFECT DISTRIBUTION









- ***Hospital_type_code***

The majority of records happened at hospitals of type *a*, with most patients staying for 21-30 days in the hospitals.

- ***Hospital_region_code***

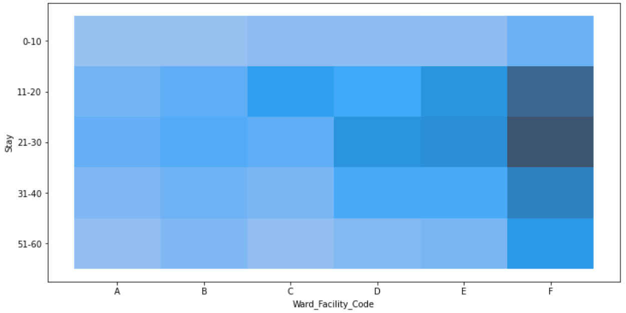
The majority of records happened at hospitals in region x, with most of the patients staying for 21-30 days in hospitals.

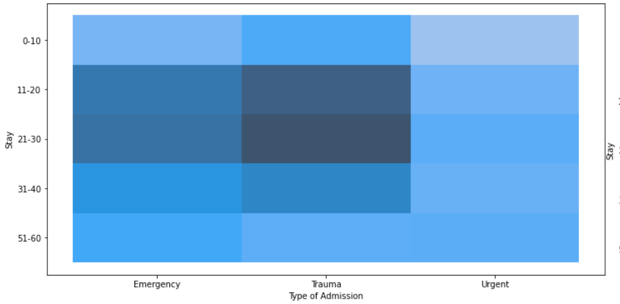
- ***Department***

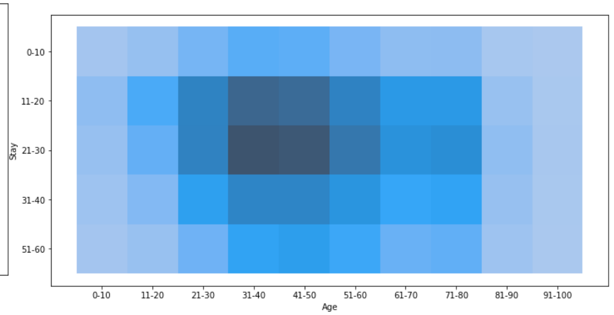
The majority of records happened in the department of gynecology, with most of the patients staying for 21-30 days in hospitals.

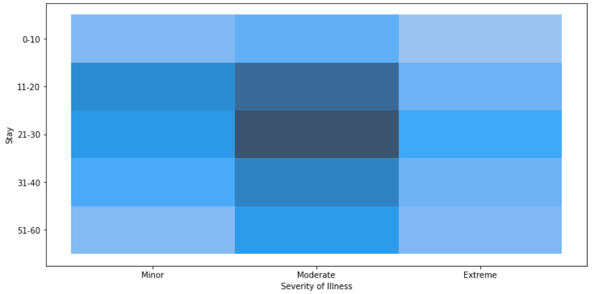
- ***Ward Type***

The majority of records happened in Wards of type Q, R and S, with most of the patients staying for 21-30 days in hospitals.









- ***Ward_Facility_Code***

The majority of records happened with Ward Facility Code of E and F, with most of the patients staying for 21-30 days in hospitals.

- ***Type of Admission***

The majority of records happened with patients in Emergency and Trauma admissions, with most of the patients staying for 21-30 days in hospitals.

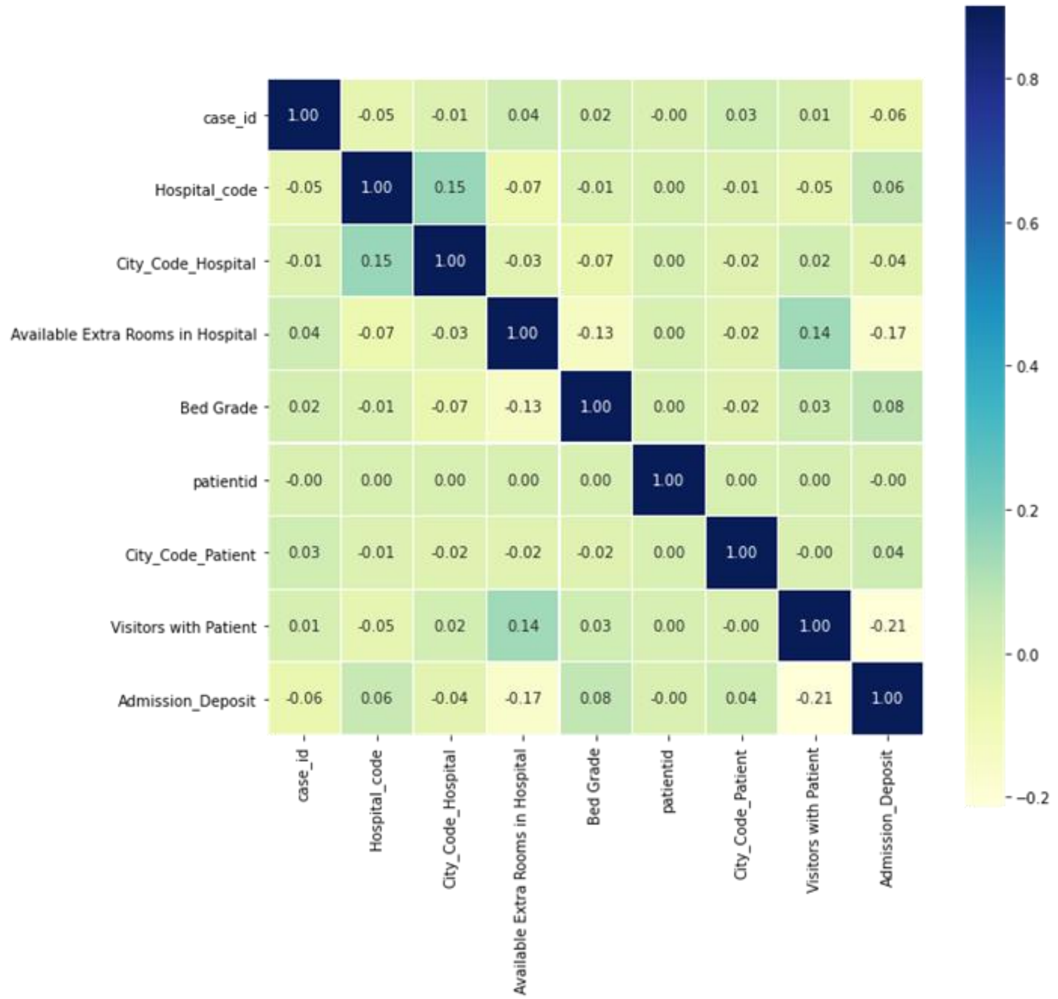
- ***Age***

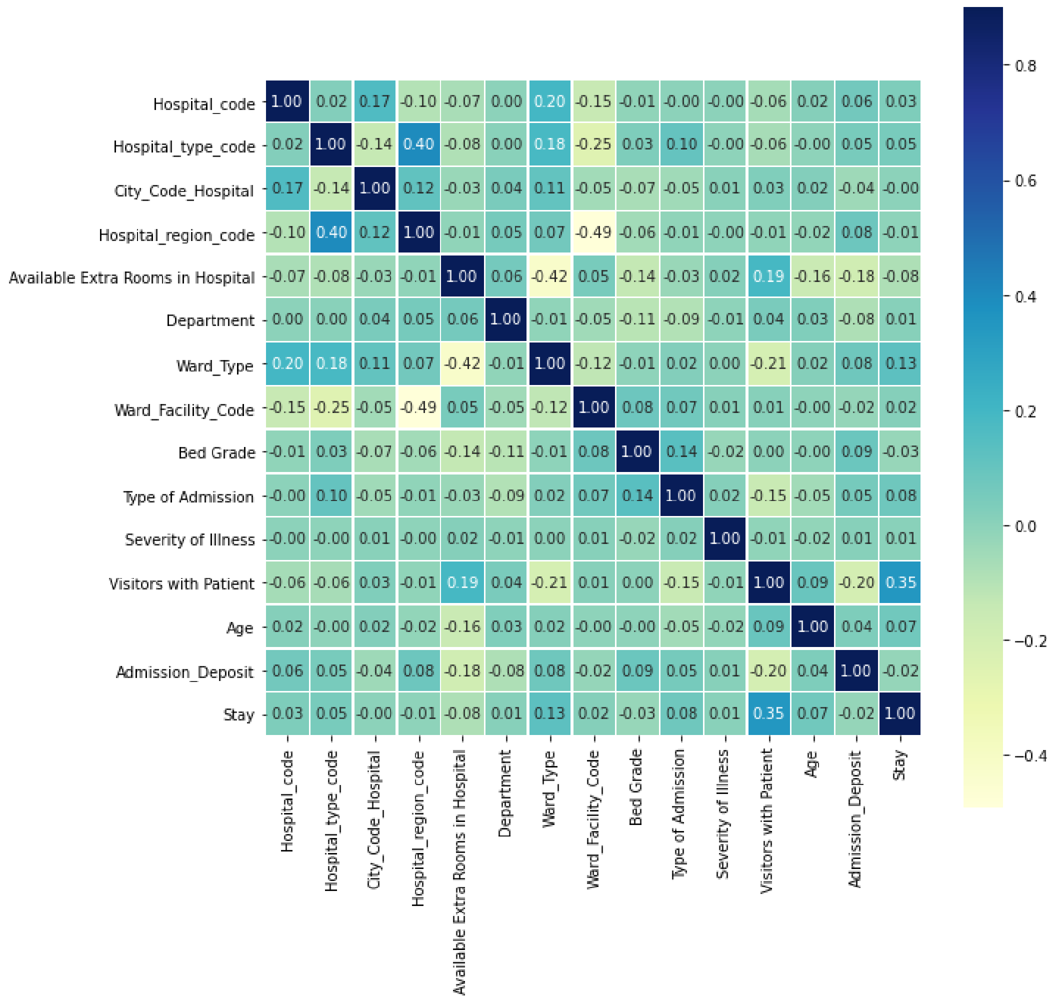
The majority of records happened with patients aging 21-80 years, with most of the patients staying for 21-30 days in hospitals.

- ***Severity of Illness***

The majority of records happened with patients in Minor and Moderate severity of illness, with most of the patients staying for 11-30 days in hospitals.

PAIRWISE CORRELATIONS FOR VARIABLES SHOWS ALL ARE NOT CORRELATED



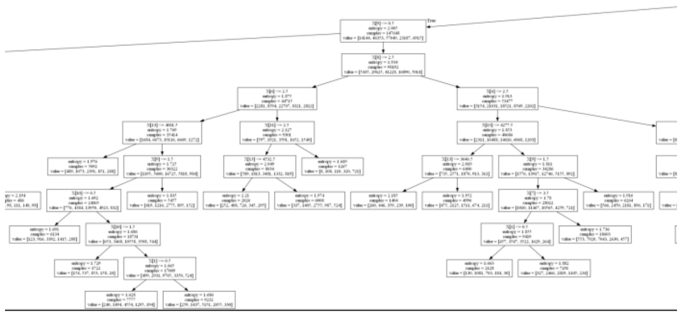


DECISION TREE CLASSIFIER – 0.424 SCORE

Entropy based classifier

All 14 features in, no other limits

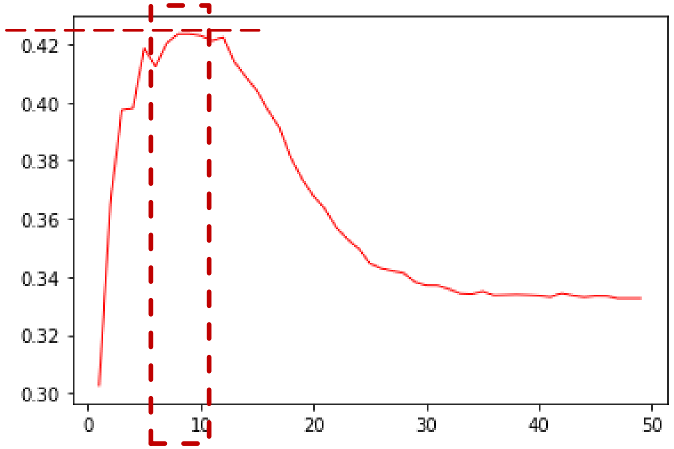
Score: 0.4182



Maximum Tree Depth

The maximum score 0.423533

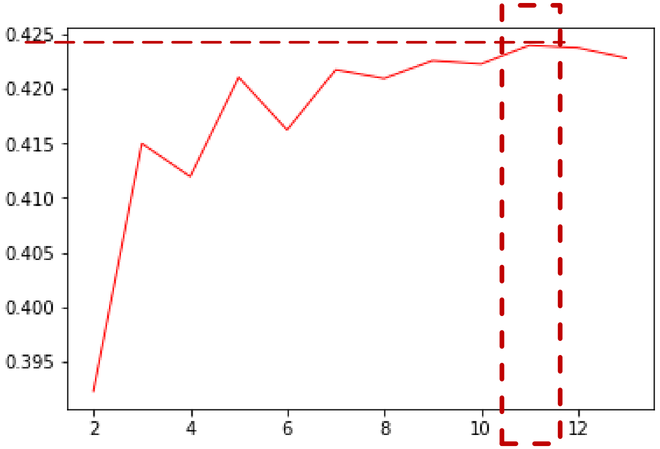
obtained at tree depth 9



Feature Number optimization

The maximum score 0.4239639

obtained at tree depth 9 feature number 11.




```
['Hospital_code' 'Hospital_type_code'
'City_Code_Hospital' 'Hospital_region_code'
'Available Extra Rooms in Hospital'
'Department' 'Ward_Type'
'Ward_Facility_Code' 'Bed Grade' 'Type of
Admission' 'Severity of Illness' 'Visitors
with Patient' 'Age' 'Admission_Deposit']
```

OTHER PREDICTIVE ANALYSIS - STEPWISE REGRESSION

Data Prep for Predictive analysis

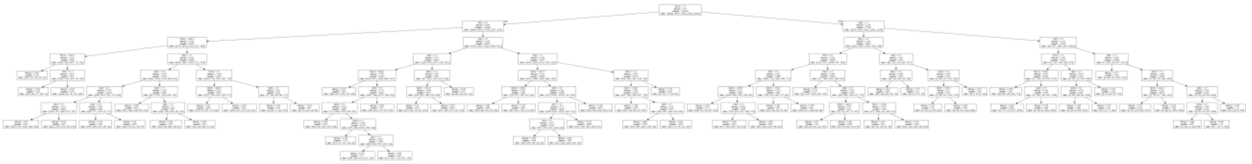
- Data encoding ➔ categorical to numerical
- Correlation checking ➔ all 14 variables are not correlated

Decision Tree Classifier

- Finetuned: from 14 to 11 variables, max depth 9 produce decision tree
- Decision tree score 0.4239639
- Entropy is high, not ideal

Stepwise regression to optimize

- Most variables are significant, Severity of Illness excluded
- R squared from 0.226 improve to 0.809



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Stay   R-squared:                       0.226
Model:                            OLS   Adj. R-squared:                  0.226
Method:                 Least Squares   F-statistic:                     4639.
Date:                Thu, 23 Mar 2023   Prob (F-statistic):               0.00
Time:                        20:52:18   Log-Likelihood:            -3.1786e+05
No. Observations:              222868   AIC:                         6.358e+05
Df Residuals:                  222853   BIC:                         6.359e+05
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
             t      P>|t|      [0.025      0.975]
---------------------------------------------------
         6.458      0.000       0.101       0.189
        12.399      0.000       0.003       0.004
        12.111      0.000       0.015       0.021
       -13.927      0.000      -0.011      -0.008
         2.332      0.020       0.001       0.015
       -38.467      0.000      -0.085      -0.077
         5.159      0.000       0.012       0.027
        83.059      0.000       0.254       0.266
        18.897      0.000       0.026       0.032
       -36.298      0.000      -0.097      -0.087
        65.165      0.000       0.200       0.212
        -1.434      0.152      -0.010       0.001
       236.371      0.000       0.399       0.405
        17.646      0.000       0.018       0.023
        23.122      0.000    4.41e-05    5.22e-05
=======================
                  1.998
):             1845.182
                   0.00
               5.31e+04
=======================
```

```
==========================================================================================
Dep. Variable:                   Stay   R-squared (uncentered):                      0.809
Model:                            OLS   Adj. R-squared (uncentered):                 0.809
Method:                 Least Squares   F-statistic:                             7.278e+04
Date:                Thu, 23 Mar 2023   Prob (F-statistic):                           0.00
Time:                        20:52:18   Log-Likelihood:                        -3.1789e+05
No. Observations:              222868   AIC:                                     6.358e+05
Df Residuals:                  222855   BIC:                                     6.359e+05
Df Model:                          13
Covariance Type:            nonrobust
======================================================================================================
                                        coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------------
Hospital_code                         0.0036      0.000     13.988      0.000       0.003       0.004
Hospital_type_code                    0.0185      0.002     12.322      0.000       0.016       0.021
City_Code_Hospital                   -0.0092      0.001    -13.192      0.000      -0.011      -0.008
Hospital_region_code                  0.0126      0.003      3.749      0.000       0.006       0.019
Available Extra Rooms in Hospital    -0.0747      0.002    -40.479      0.000      -0.078      -0.071
Department                            0.0290      0.004      8.106      0.000       0.022       0.036
Ward_Type                             0.2675      0.003     93.146      0.000       0.262       0.273
Ward_Facility_Code                    0.0322      0.001     22.232      0.000       0.029       0.035
Bed Grade                            -0.0871      0.002    -36.193      0.000      -0.092      -0.082
Type of Admission                     0.2079      0.003     66.181      0.000       0.202       0.214
Visitors with Patient                 0.4049      0.002    246.399      0.000       0.402       0.408
Age                                   0.0221      0.001     19.676      0.000       0.020       0.024
Admission_Deposit                  5.471e-05    1.8e-06     30.413      0.000    5.12e-05    5.82e-05
==============================================================================
Omnibus:                     2991.162   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1799.838
Skew:                          -0.031   Prob(JB):                         0.00
Kurtosis:                       2.564   Cond. No.                     8.90e+03
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 8.9e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
By using stepwise, the Mean Squared Error:  1.0052631444321234
```

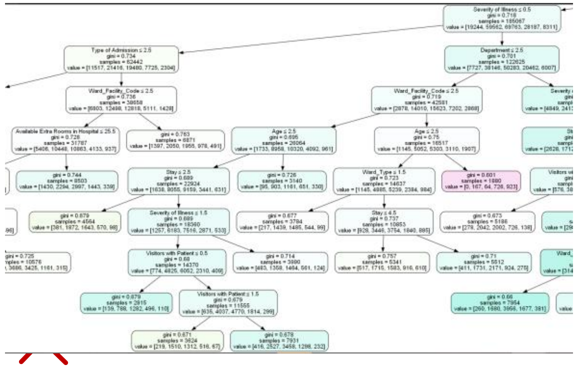
Highest co-efficients:

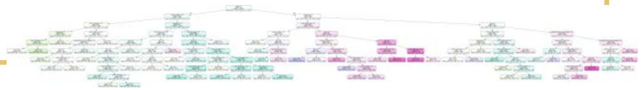
Visitors with Patient	0.4049
Ward_Type	0.2675
Type of Admission	0.2079

MANUAL REMOVING ‘GENERIC’ FEATURES & OTHER MODELS

Decision Tree manual remove of features

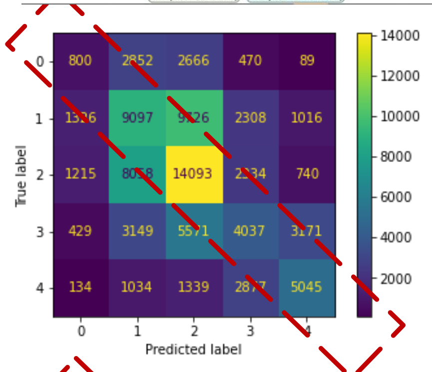
- Explore how useful are the high level data?
- Features dropped: 'Hospital_code', 'City_Code_Hospital','Hospital_region_code', 'Bed Grade',
- Experiment to manual remove some "generic" features
- Gini value around 0.4 – 0.6, not ideal, Decision tree score 0.3311 ➔ Value dropped as compared to 0.4239639
- Generic level of features are useful





RandomForestClassifier

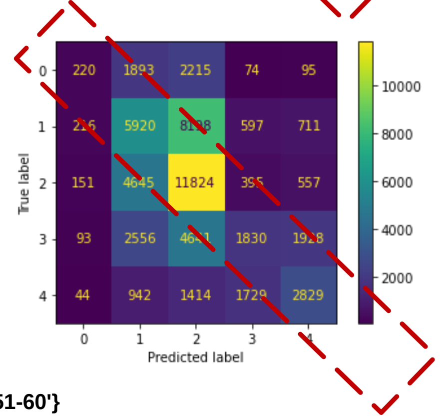
- Multiple decision trees are created using different random subsets of the data and features.
- Accuracy 0.3957



Naïve Bayes

- well-suited for multi-class classification problems
- Robust to isolated noise points
- Robust to irrelevant attributes
- feature independence
- Accuracy 0.41

	precision	recall	f1-score	support
0	0.30	0.05	0.08	4497
1	0.37	0.38	0.37	15642
2	0.42	0.67	0.52	17572
3	0.40	0.17	0.23	11048
4	0.46	0.41	0.43	6958
accuracy			0.41	55717
macro avg	0.39	0.33	0.33	55717
weighted avg	0.40	0.41	0.37	55717



Length of Stay {0: '0-10', 1: '11-20', 2: '21-30', 3: '31-40', 4: '51-60'}

CLOSED-LOOP & FUTURE WORK

Data Needed

- Singapore local data (utilization of bed)
- Breakdown of BOR information
- Region specific hospital functions and bed numbers
- Population demographic (age, gender, race, BMI)
- Insurance details
- Diagnosis of patient

Data Collection

- Keep track of the length of stay of each patient
- Collaboration with hospital functions/government agencies
- Waiting time in each function in the hospital
- Tracking app system within the hospital ecosystem that contains patient information and length of stay

Further Analysis

- Predictive analysis to predict the length of stay of each patient
- Clustering analysis to determine location allocation
- Include other parameters into the current analysis (service time, available resource, waiting time, waiting list)

Implementation

- Define and monitor factors that cause high BOR
- Collaborate and share data information among hospitals to relocate patients for optimal allocation
- Collect LOS data report and make relevant prediction to improve the bed allocation system

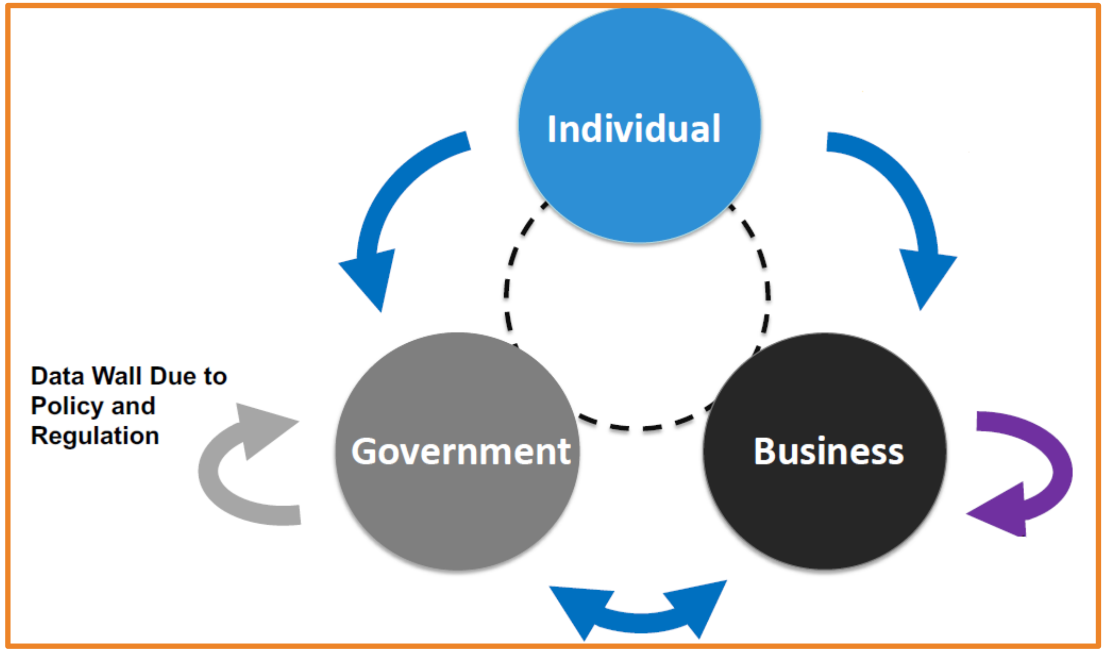
DATA GOVERNANCE – ISSUE & PROPOSED SOLUTION

Data Wall due to Policy and Regulation

- ✓ The Health Information Bill will be tabled in the second half of 2023 to better safeguard the sharing of patient information across different healthcare providers to pave the way for Singapore's switch to a new healthcare model based on prevention.

- ✓ The Bill will enable the collection of patients' selected health data from healthcare providers, and to allow healthcare providers to share health and administrative data with one another for specific purposes.

Data Ecosystem



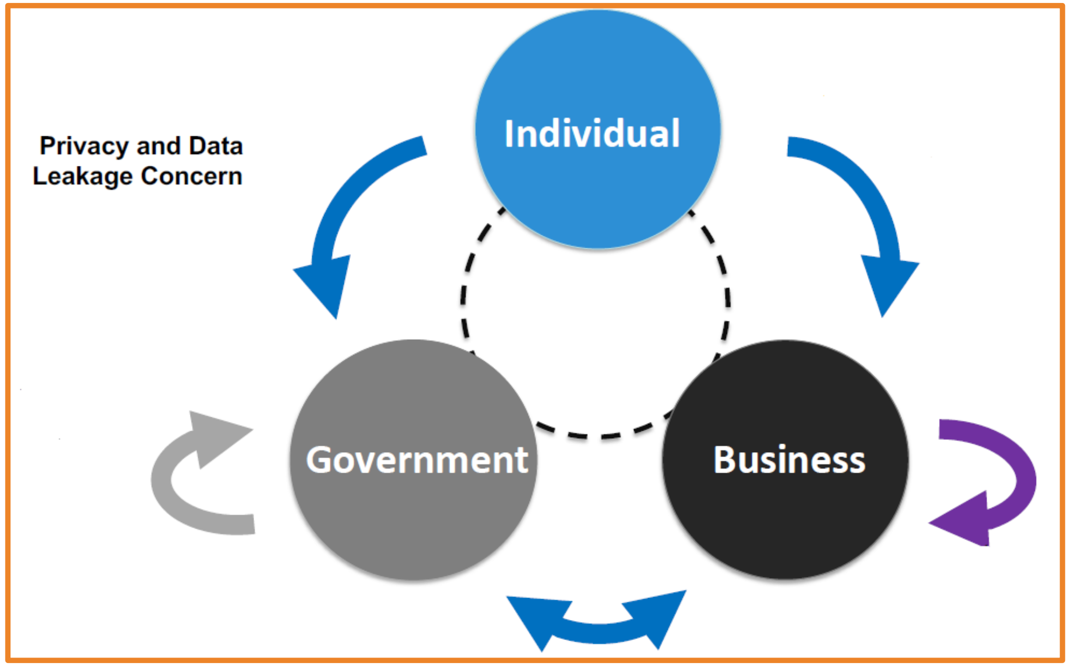
- ✓ Data sharing would allow for the seamless provision of care, and early detection and intervention to keep high risk illnesses at bay ➔ Lower hospital bed demand, reduce BOR.

DATA GOVERNANCE – ISSUE & PROPOSED SOLUTION

Privacy and Data Leakage Concern

- ✓ While protections are already in place under the Personal Data Protection and Cybersecurity Acts governing the use of health information, more will still need to be done.
- ✓ MOH is working with the Personal Data Protection Commission and the Cyber Security Agency of Singapore to identify areas where safeguards would need to be strengthened.

Data Ecosystem



- ✓ MOH has consulted healthcare professionals, patients and IT vendors on data privacy and sharing issues. The views gathered will help to shape the Bill to address our policy intent, the needs of patients, and the administrative and operational costs to providers.

THANK YOU

Brought to you by:

ISSS621 G2 Group 4