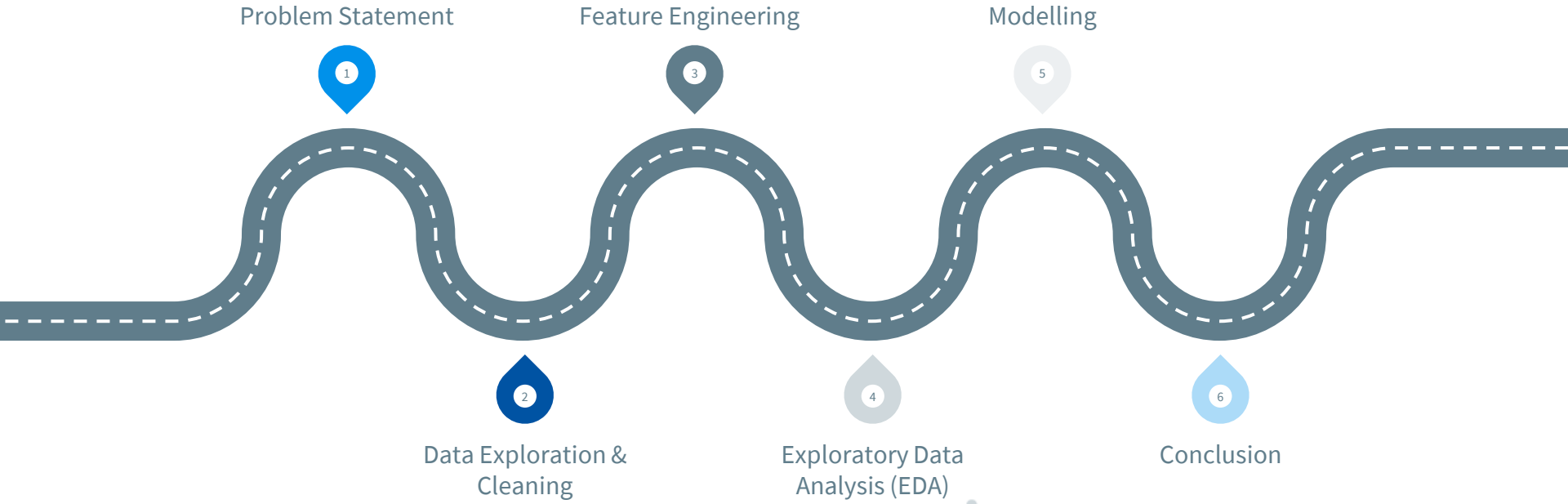
Hospitalisation Cost Drivers

Rifqi Alkhatib

Holmusk Healthcare Data Challenge

Agenda


Problem Statement
1
Feature Engineering
3
Modelling
5
2
Data Exploration & Cleaning
4
Exploratory Data Analysis (EDA)
6
Conclusion

Problem Statement

Problem Statement

In order to combat the issue of high hospitalisation bills, the Ministry of Health (MOH) wants to understand the drivers of cost of care for patients hospitalised for a certain condition

Data Provided:

Clinical & financial data of patients hospitalised for a certain condition (Jan 2011 – Jan 2016)

Data Exploration & Cleaning

Datasets

- 4 separate datasets
  - Bill Amount
  - Bill ID
  - Clinical Data
  - Demographics
- Clean & merge into 1 dataframe

Datasets – Bill Amount

- Bill ID
  - 6 to 10-digit number
- Bill Amount
  - SGD
  - Target Variable

Datasets – Bill ID

- Bill ID
- Patient ID
  - Anonymised
  - Alphanumeric ID
- Date of Admission
  - YYYY-MM-DD Format

	bill_id	patient_id	date_of_admission
0	7968360812	1d21f2be18683991eb93d182d6b2d220	2011-01-01
2	7512568183	1d21f2be18683991eb93d182d6b2d220	2011-01-01
4	7654730355	1d21f2be18683991eb93d182d6b2d220	2011-01-01
5	1692196063	1d21f2be18683991eb93d182d6b2d220	2011-01-01
12792	6466235037	1d21f2be18683991eb93d182d6b2d220	2015-09-17
12793	7809633370	1d21f2be18683991eb93d182d6b2d220	2015-09-17
12794	5607272671	1d21f2be18683991eb93d182d6b2d220	2015-09-17
12795	5776306727	1d21f2be18683991eb93d182d6b2d220	2015-09-17

- Multiple bill IDs for each patient & date of admission

Datasets – Clinical Data

- Patient ID
- Date of Admission
- Date of Discharge
- Medical History – 7 features (Categorical: 0 or 1)
- Preop Medication - 6 features (Categorical: 0 or 1)
- Symptoms - 5 features (Categorical: 0 or 1)
- Lab Results – 3 features (Float)
- Weight (kg)
- Height (cm)

Datasets – Clinical Data

- Multiple entries for same patient
  - One for each individual admission
  - Clinical data inconsistent across admissions
    - Medical history
    - Preop Medication
    - Symptoms
    - Lab Results

	id	date_of_admission	date_of_discharge	medical_history_1	medical_history_2	medical_history_3
88	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	2011-03-08	0	1	0
273	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	2011-06-08	0	0	1
986	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	2012-06-29	1	0	0

Datasets – Demographics

- Patient ID
- Gender
  - Male / Female
- Race
  - CMIO
- Resident Status
  - SG Citizen / PR / Foreigner
- Date of Birth
  - YYYY-MM-DD Format

Datasets – Merging into 1 Dataframe

- Merge Bill Amount & Bill ID (on bill_id)
  - 13600 rows each
  - ‘bill’
- Left join Demographics onto Clinical Data (on patient_id)
  - Multiple rows for same patient in clinical data
  - 3400 rows
  - ‘patient’
- Left join ‘patient’ onto ‘bill’ (on patient_id)
  - 13600 rows, 32 variables

Data Cleaning – Duplicate Bill Amounts

	bill_id	patient_id	date_of_admission	bill_amount	date_of_discharge	r
8820	2367848755	88daa1492f00862c8cdeb8ed181df22e	2014-04-03	1012.028954	2014-04-16	
10109	3180485821	88daa1492f00862c8cdeb8ed181df22e	2014-09-13	1012.028954	2014-09-30	

- Same patient
  - Different admissions
  - Different bill_id
- Exact same bill amount
  - Combined bill or human error
- Dropped
  - 9 data points

Merging Bills from Same Admission

- Multiple bills for same patient for each hospitalization
  - Bills from different departments

	patient_id	date_of_admission	bill_amount	date_of_discharge
366	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	2444.80	2011-03-08
367	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	1455.54	2011-03-08
368	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	19943.02	2011-03-08
371	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	1447.26	2011-03-08
1124	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	1045.39	2011-06-08
1126	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	1460.12	2011-06-08
1127	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	1426.59	2011-06-08
1128	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	9087.35	2011-06-08
3972	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	1516.63	2012-06-29
3975	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	1188.14	2012-06-29
3977	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	6502.11	2012-06-29
3979	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	8472.64	2012-06-29

	patient_id	date_of_discharge	bill_amount	date_of_admission
90	b2d15cda8c4e1f86ba43356434df6718	2011-03-08	25290.62	2011-02-26
273	b2d15cda8c4e1f86ba43356434df6718	2011-06-08	13019.45	2011-06-02
986	b2d15cda8c4e1f86ba43356434df6718	2012-06-29	17679.52	2012-06-21

Additional Data Preprocessing

- Adjusting bill amount for inflation
  - To Jan 2016 Consumer Price Index
- Dropping bill ID
  - No relation to patient clinical & demographic data

Feature Engineering

Feature Engineering

- Length of Hospitalisation (days)
  - Longer hospitalization → higher costs from daily charges
- Patient Age (years)
  - Older patients → more complications
- Body Mass Index, BMI
  - Estimate risk for obesity-related diseases

Feature Engineering

- Extract year & month of hospitalization
  - Study trends year-to-year & month-to-month

- Number of times hospitalised (on admission date)
  - Study effect of repeated hospitalisations on the bill amount
  - Incremental (1 → 2 → 3)

	patient_id	date_of_admission	hosp_no
90	b2d15cda8c4e1f86ba43356434df6718	2011-02-26	1
273	b2d15cda8c4e1f86ba43356434df6718	2011-06-02	2
986	b2d15cda8c4e1f86ba43356434df6718	2012-06-21	3

Feature Engineering

- Summed clinical features
  - Medical History, Preop Medication, Symptoms
  - Study relationship between total occurrences and bill amount
    - More history / meds / symptoms → higher costs

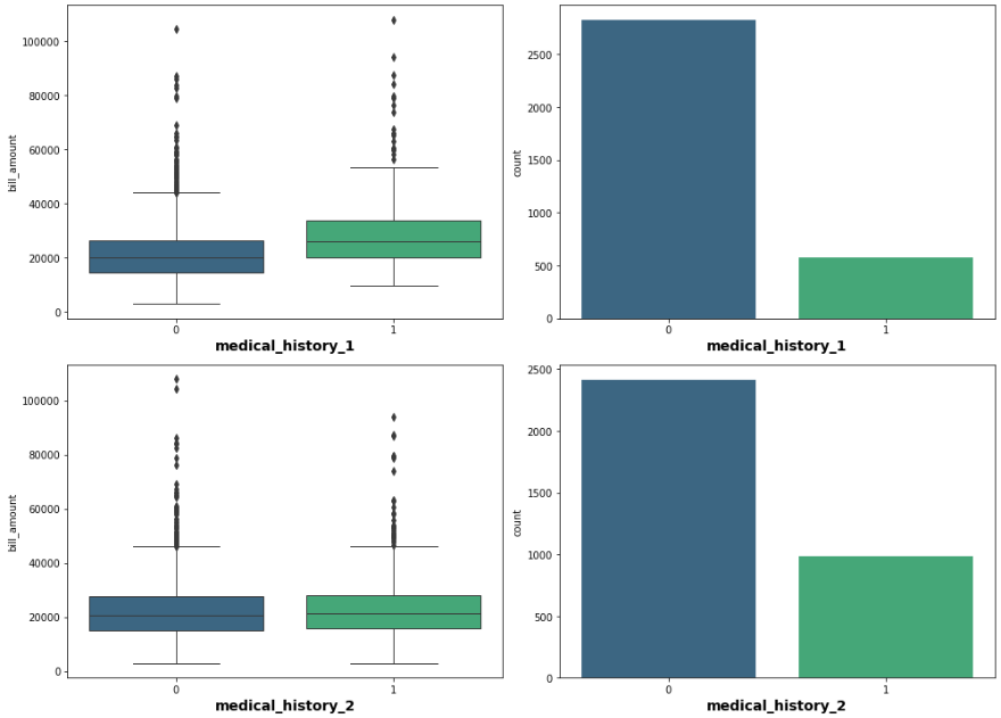
- Initial Feature Elimination
  - Patient ID
  - Date of admission & discharge
  - DOB

EDA

EDA (Categorical) – Medical History

Individual medical histories

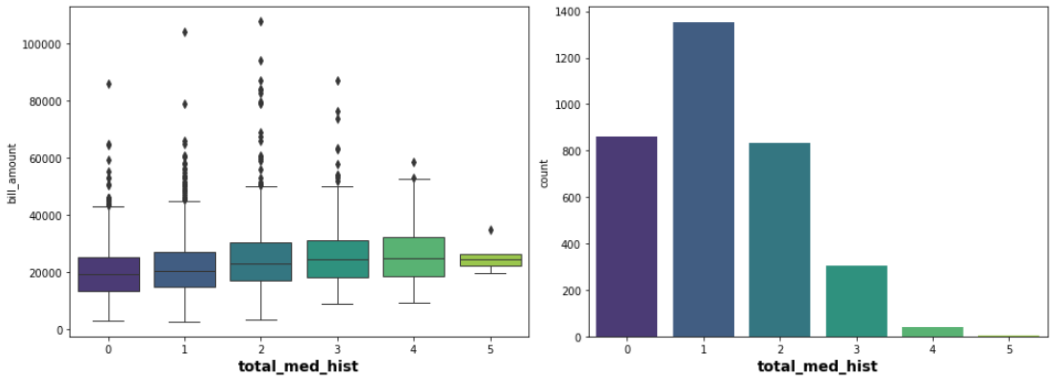
- More patients *without* each medical history
- Median bill amount slightly higher for patients *with* medical history



EDA (Categorical) – Medical History

Total medical histories

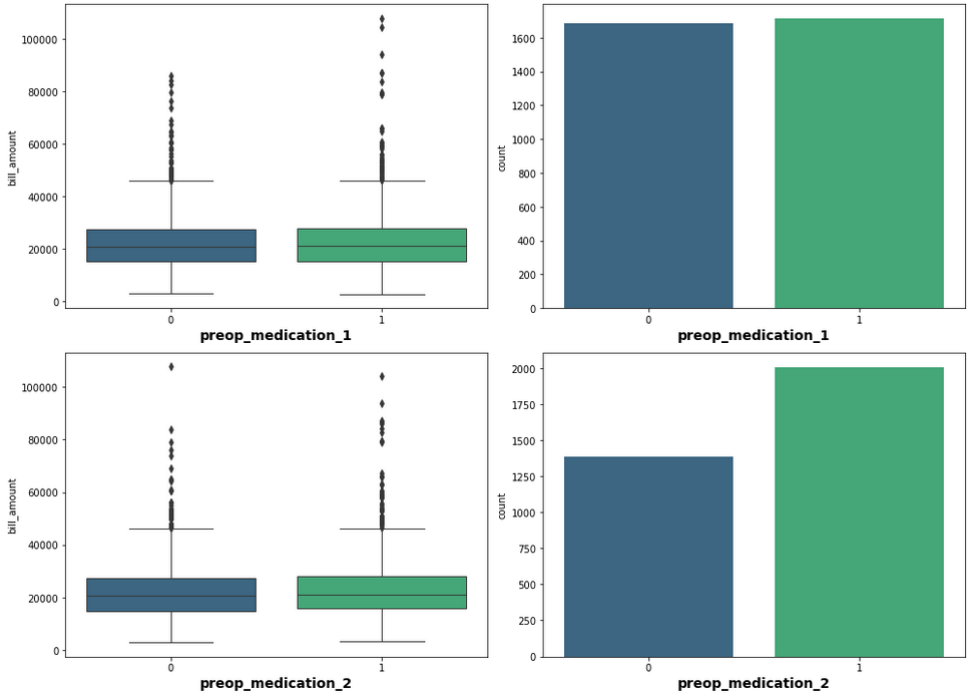
- Right skew in distribution
- Positive trend for total medical history
- Ordinal variable



EDA (Categorical) – Preop Medication

Individual preop medications

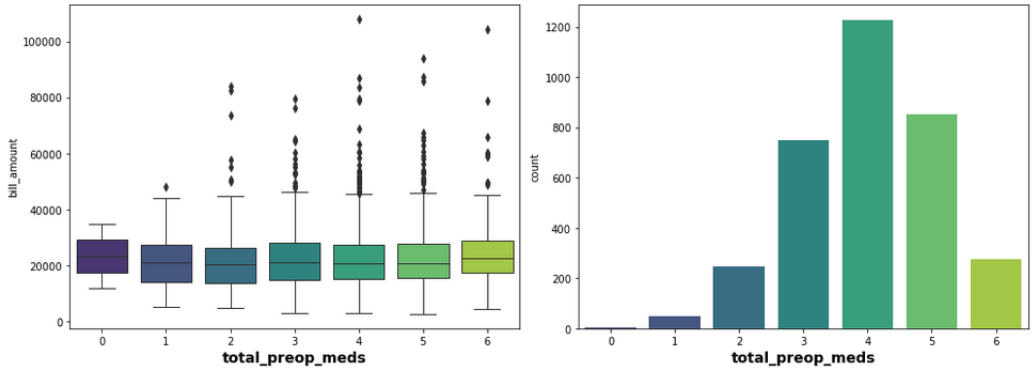
- More patients *given* each individual medication
- No clear difference in medians



EDA (Categorical) – Preop Medication

Total preop medications

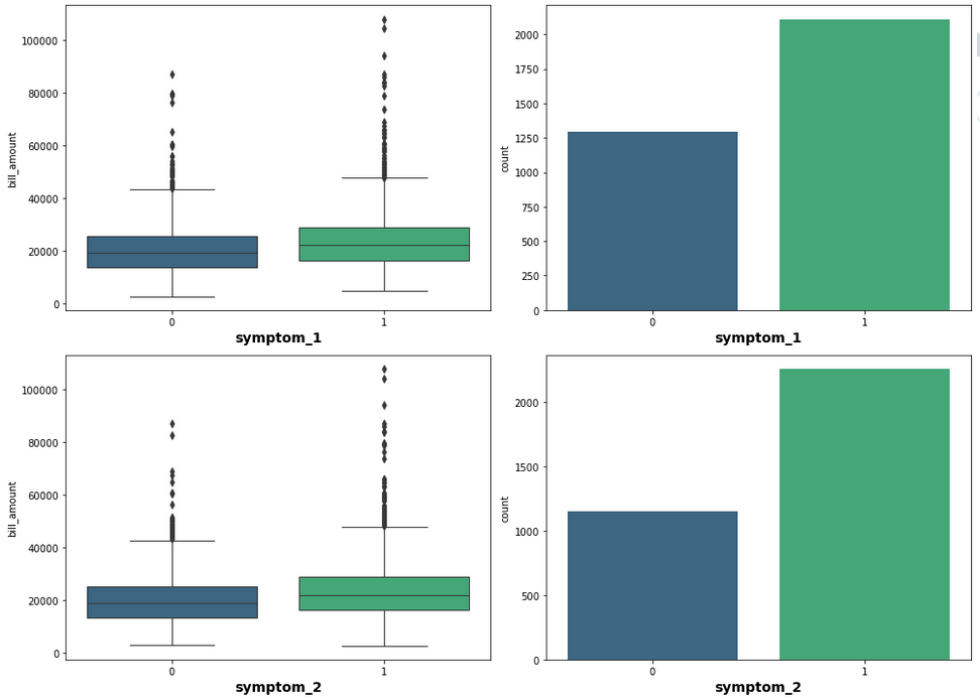
- Left skew in distribution
- No clear trend for total number of preop meds given
- Dropped



EDA (Categorical) – Symptoms

Individual symptoms

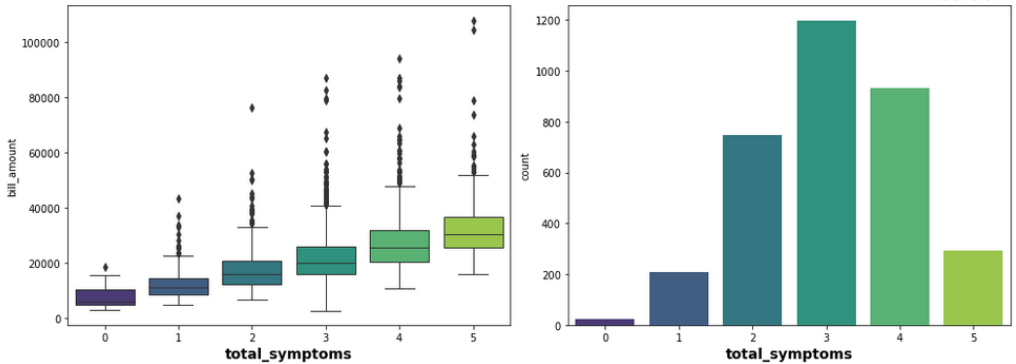
- More patients *with* each individual symptom
- Median bill price slightly higher for patients *with* symptom



EDA (Categorical) – Symptoms

Total symptoms

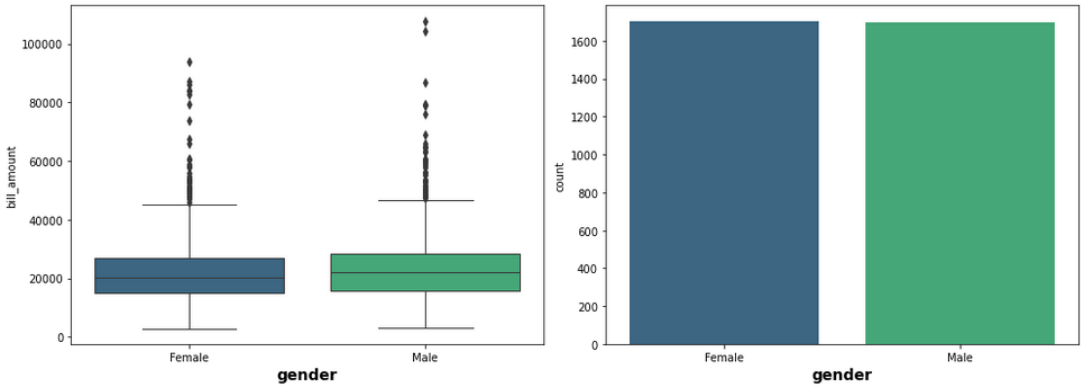
- Slight left skew in distribution
- Positive trend for total symptoms
- Ordinal variable



EDA (Categorical) – Demographic Data

Gender

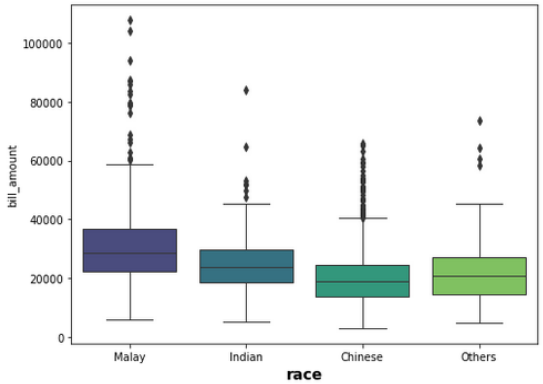
- Even balance of males & females
- Slightly higher median bill amount for males

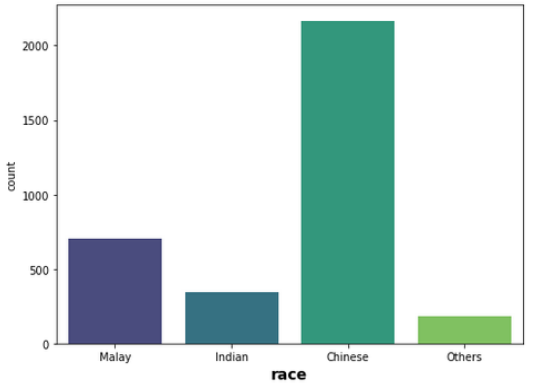


EDA (Categorical) – Demographic Data

Race

- Minority races slightly overrepresented
- Median bill amount: Malay > Indian > Others > Chinese
- Nominal variable

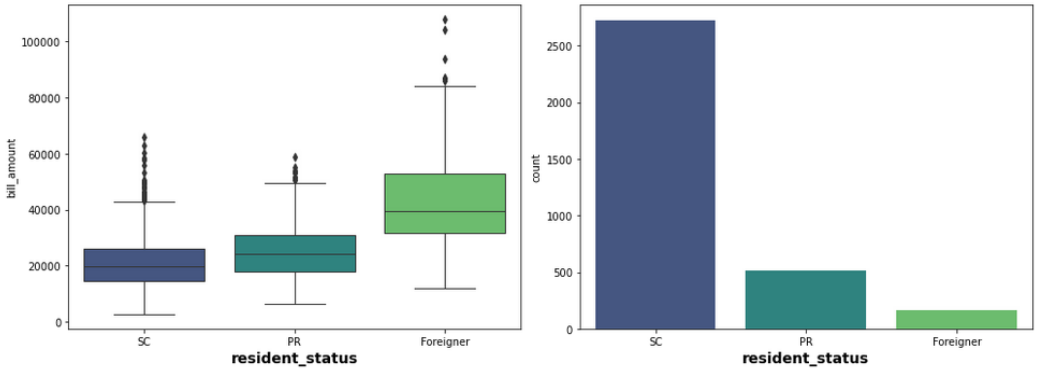




EDA (Categorical) – Demographic Data

Resident Status

- Foreigners underrepresented
- Median bill amount: Foreigner > PR > SC
- Nominal variable



EDA (Categorical) – Hospitalisation Length & Number

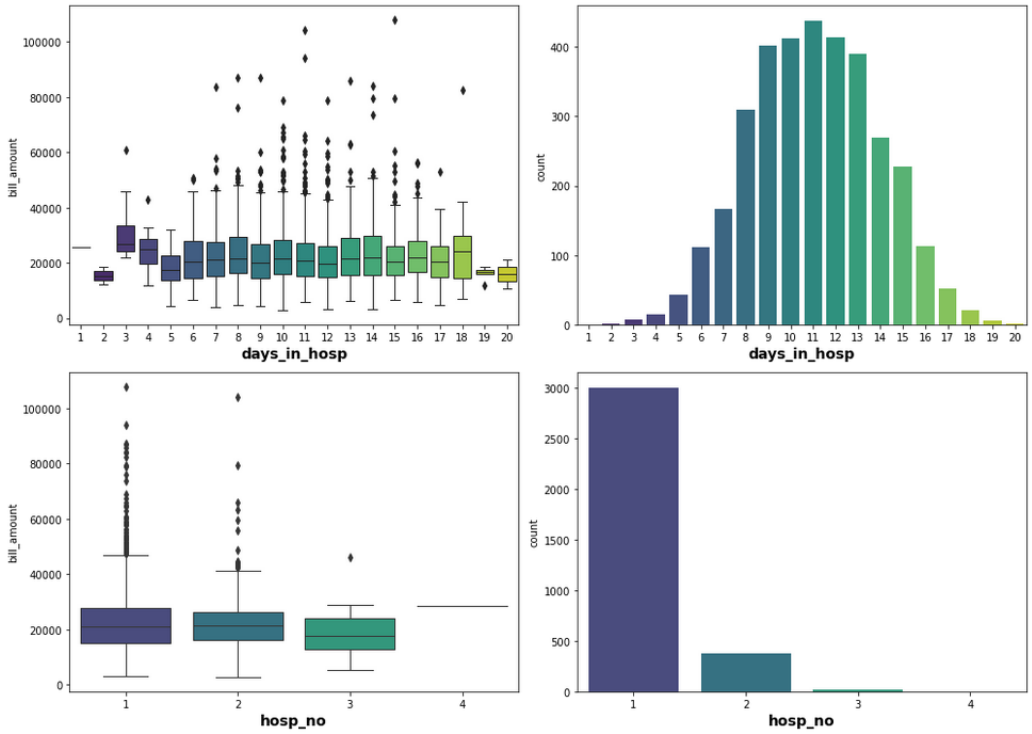
Days in Hospital

- Approximately normal
- No clear trend

Hospitalisation Number

- Data from Jan 2011
- No clear trend

- Dropped



EDA (Categorical) – Month & Year of Hospitalisation

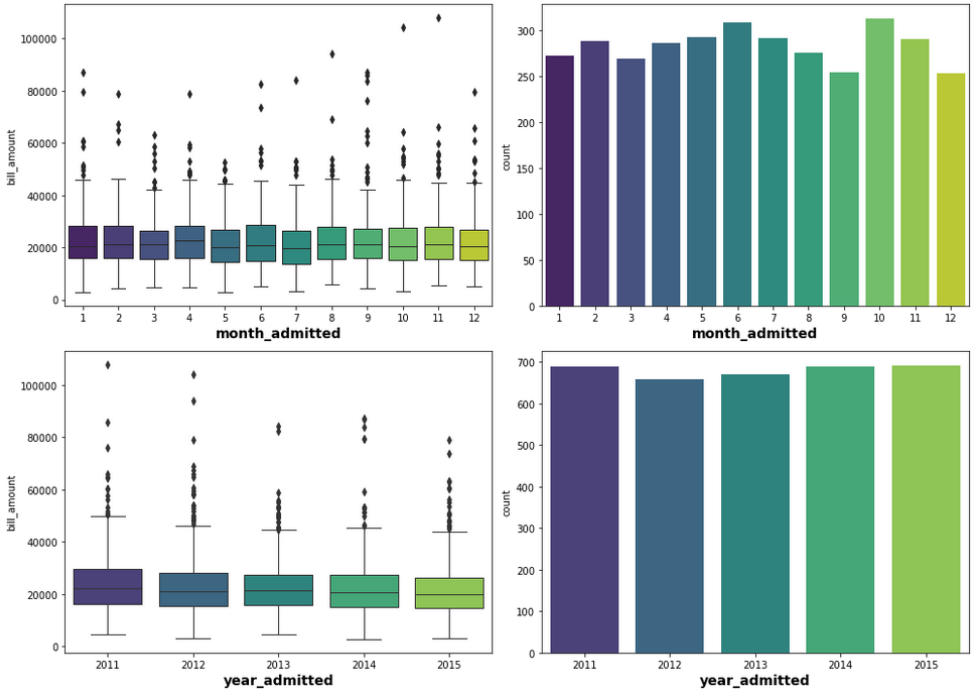
Month Admitted

- No imbalance
- No clear trend

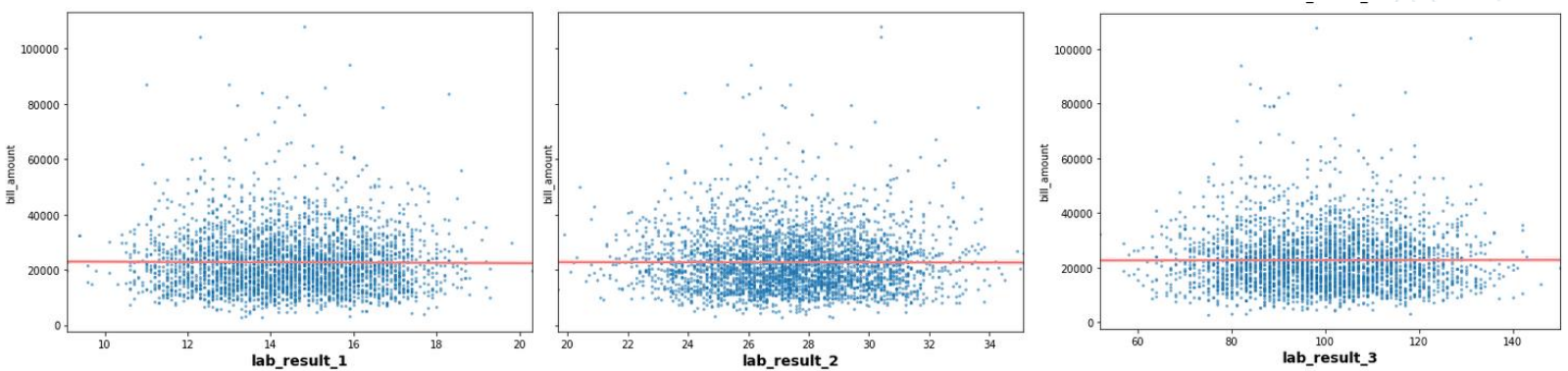
Year Admitted

- No imbalance
- Slight negative trend
  - Due to CPI adjustment

- Dropped

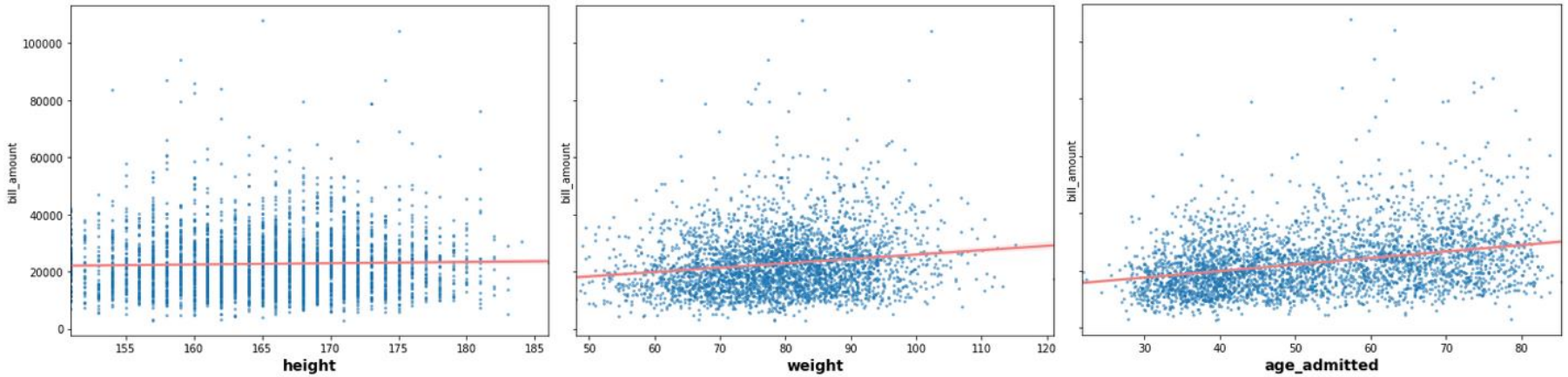


EDA (Continuous) – Lab Results



- No clear relationship
- Patients charged for lab test regardless of result
- Dropped

EDA (Continuous) – Physical Features

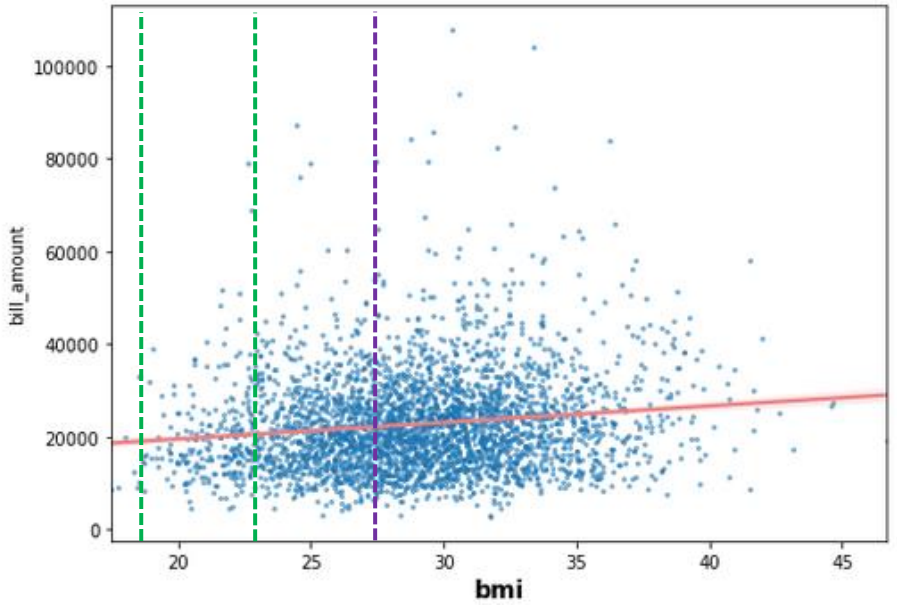


- No clear relationship for height – dropped
- Slight positive correlation for weight & age

EDA (Continuous) – Physical Features

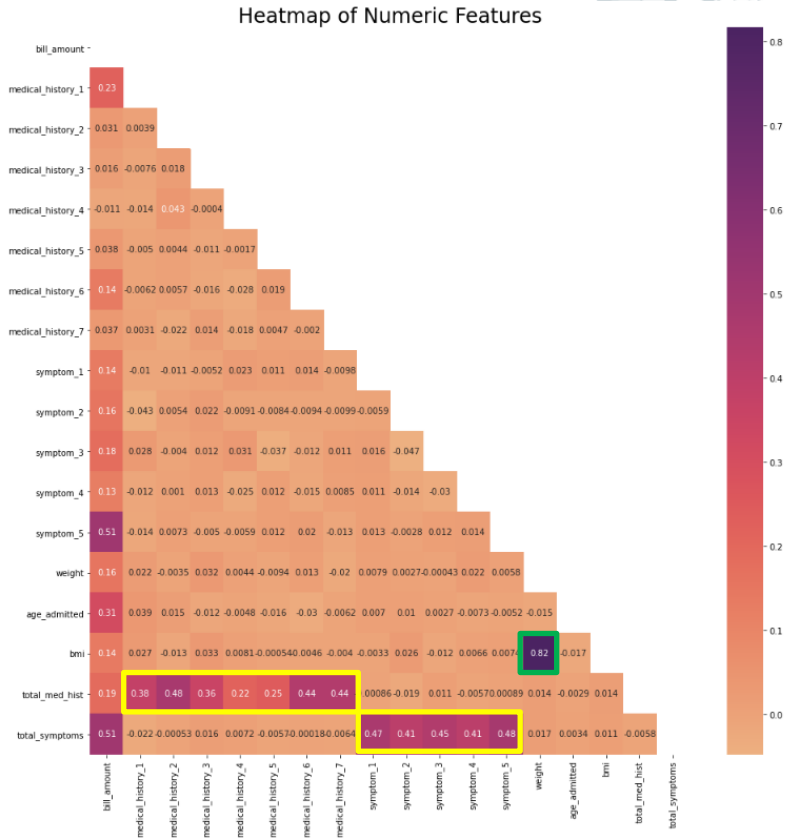
BMI

- Singapore healthy range: 18.5 – 22.9 kg / m^2
- Most patients overweight
  - Many at high risk
- Very few underweight
- Slight positive correlation



EDA – Heatmap of Numeric Features

- Look out for multicollinearity
- High correlation between weight and BMI
  - Dropped BMI
- Moderate correlation between total_med_hist, total_symptoms and their components
  - Keep



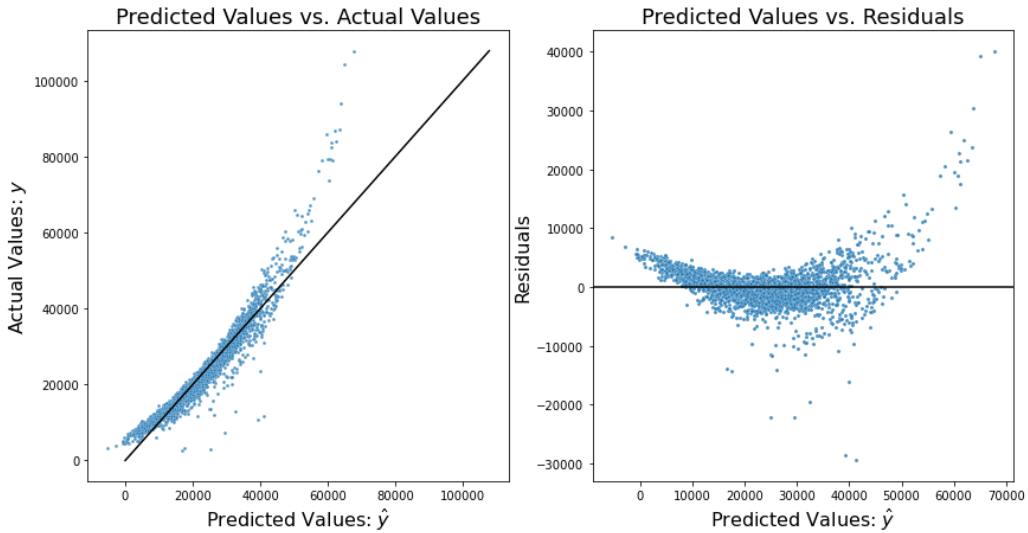
Final Feature Set

- 19 independent variables
  - Medical History (1 – 7 & total)
  - Symptom (1 – 5 & total)
  - Weight
  - Gender
  - Race
  - Resident Status
  - Age

Modelling

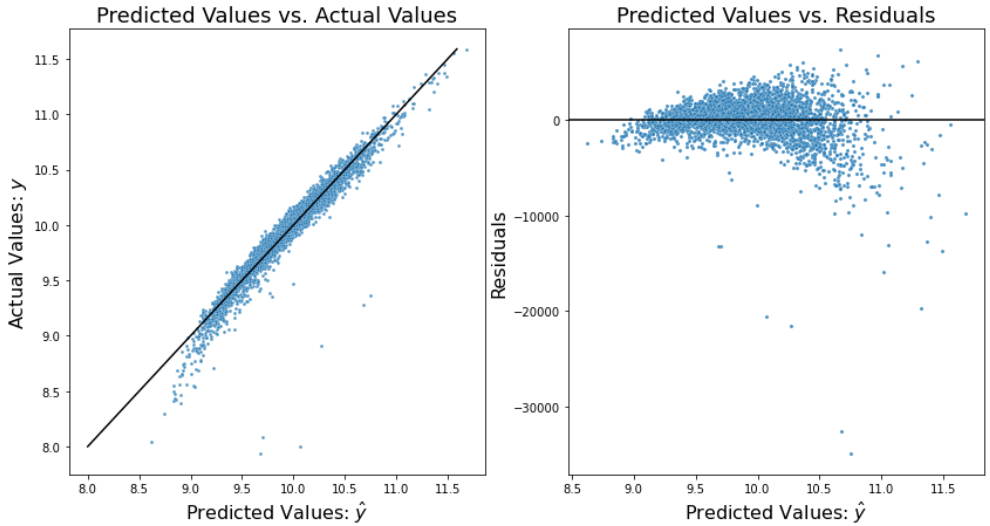
Initial Model

- Fit in all features with bill amount as y
- Clear non-linearity of predictions
- Heteroscedasticity of residuals
- RMSE: 3180.9



Model with Log Transformed Target Variable

- Fit in all features with log(bill amount) as y
- Improved linearity of predictions
- Reduced heteroscedasticity of residuals
- RMSE: 2236.6



Model Analysis

- R-squared
  - Model able to explain 94.1% of changes in target variable
- Adj. R-squared
  - Almost all variables are contributing properly
- Prob (F-statistic)
  - At least one independent variable has significant effect

OLS Regression Results

Dep. Variable:	np.log(bill_amount)	**R-squared:**	0.941
Model:	OLS	**Adj. R-squared:**	0.941
Method:	Least Squares	**F-statistic:**	2708.
Date:	Mon, 30 Aug 2021	**Prob (F-statistic):**	0.00
Time:	20:16:34	**Log-Likelihood:**	2700.6
No. Observations:	3400	**AIC:**	-5359.
Df Residuals:	3379	**BIC:**	-5230.
Df Model:	20		
Covariance Type:	nonrobust		

Model Analysis – Coefficients

- Equation for MLR model:

$$\log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

  - 1 unit increase in X_1 → β_1 increase in log(y)

- log(bill amount) does not make sense
  - Exponentiate

$$y = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}$$

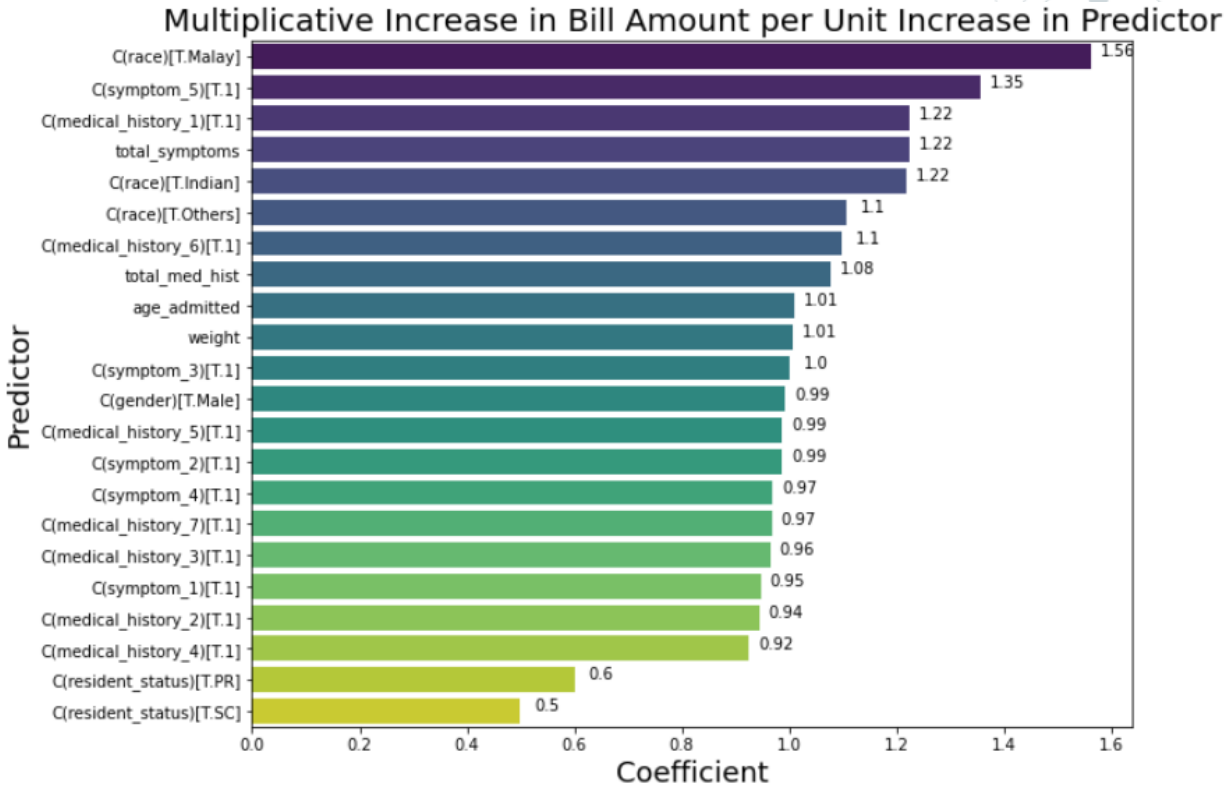
  - 1 unit increase in X_1 → e^{β_1} times increase in y compared to 'baseline'

Model Analysis – Coefficients

- ‘Baseline’
  - Bill amount when all other coefficients set to 0 → $y = e^{\beta_0}$
    - Gender = Female
    - Race = Chinese
    - Resident Status = Foreigner
    - No medical history
    - No symptoms
    - Hypothetical weight & age = 0
  - Bill Amount = $5607.82
    - Add 1kg → $5643.15

Model Analysis – Coefficients

- Race an important feature
- Certain symptoms & medical histories have greater impact
- Resident status also important
- Gender, age & weight not very important
- Total medical histories & symptoms have greater impact



Conclusion

Recommendations

- Conduct further studies into race-specific differences
  - Results indicate race plays a huge role in patient's cost of care
  - Studies to identify underlying causes
  - Develop targeted measures to equalise cost of care
- Target symptom_5, medical_history_1 & medical_history_6 for early intervention
  - Studies show that early intervention and prevention highly effective at saving costs
  - Mass media campaigns targeting these 3 features
  - Too late once hospitalised

Limitations

- Ambiguity of bills
  - Multiple bills per hospitalisation
  - Nett or gross amounts
    - Subsidies, insurance, etc
- Lack of context
  - Clinical features difficult to understand without knowing how data is collected
  - Inconsistencies in data
- Addressing anonymity
  - Inevitable in healthcare
  - More domain knowledge
    - Enables formulation of more reasonable assumptions

Thank You