Finding Factors Driving Cost Of Care

Holmusk Challenge

Ray Tan

Outline

01 Introduction

02 Data Preparation

03 Exploration & Visualisation

04 Pre-Processing & Modelling

05 Finale

Outline

01 Introduction

02 Data Preparation

03 Exploration & Visualisation

04 Pre-Processing & Modelling

05 Finale

Background

- Healthcare costs in Singapore **have been rising** and are **expected to keep increasing** over the coming years
- Health insurance **policies with full riders** that **cover hospital bills in full** → **Over-consumption** by policyholders and **over-charging** of medical services
- Health insurers have had to **adjust the terms** for full-rider health insurance policies to **require co-payment** of hospital bills in order to **keep healthcare costs down**

Problem Statement

Against a backdrop of **rising healthcare costs**,

ABC Health Insurance would like to understand the **factors driving cost of care** in order to **structure their policy offerings** in a way that remains **competitive, attractive, and sustainable** in the long run

Objectives

A **Explore** the data and **visualise** individual variables as well as the relationships between independent variables and bill amount

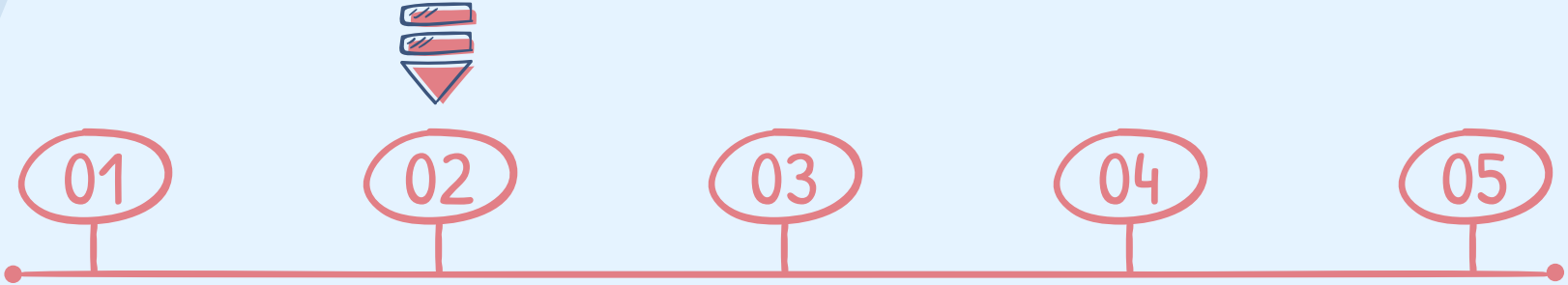
B **Predict bill amount** by training **3 regression models (linear, ridge, lasso)** and evaluate them using **2 metrics (R^2, RMSE)** to select the best performing model

C **Analyse** the final model for the **accuracy of its predictions** and the **weights of its features**

D **Find** the **main factors driving cost of care**

E **Make recommendations** that address the issue of rising healthcare costs

Outline



Introduction

Data Preparation

Exploration & Visualisation

Pre-Processing & Modelling

Finale

Data Sources

bill_amount.csv
Bill Amount Dataset
13,600 rows x 2 columns

bill_id.csv
Bill ID Dataset
13,600 rows x 3 columns

clinical_data.csv
Clinical Dataset
3,400 rows x 26 columns

demographics.csv
Demographics Dataset
3,000 rows x 5 columns

CLEAN

CLEAN

MERGE on bill_id

CLEAN

CLEAN

MERGE on patient_id

Bill Dataset

patient_id	date_of_admission	bill_id	bill_amount
00225710a878eff524a1d13be817e8e2	2014-04-10	4692776325	325.319345
00225710a878eff524a1d13be817e8e2	2014-04-10	8461069832	1074.885913
00225710a878eff524a1d13be817e8e2	2014-04-10	5175703971	79.496707
00225710a878eff524a1d13be817e8e2	2014-04-10	7746811189	3710.864731

patient_id	date_of_admission	bill_id	bill_amount
00225710a878eff524a1d13be817e8e2	2014-04-10	26076361317	5190.566695

Patient Dataset

MERGE on patient_id & date_of_admission

Cleaned Dataset

Feature Engineering

Engineered Features

Time Features

Year Of Admission
Month Of Admission

Hospitalisation Features

Days Hospitalised
Hospitalisation Count
Days Since Last Hospitalisation

Physical Features

Body Mass Index
Age At Admission

Clinical Features

Number Of Medical Histories
Number Of Pre-Op Medications
Number Of Symptoms

Retained Features

Gender, Race
Resident Status
Medical History 1-7
Pre-Op Medication 1-6
Symptom 1-5
Lab Result 1-3
Weight, Height

Dropped Features

Bill ID, Patient ID
Date Of Admission
Date Of Discharge
Date Of Birth

Outline

01 Introduction

02 Data Preparation

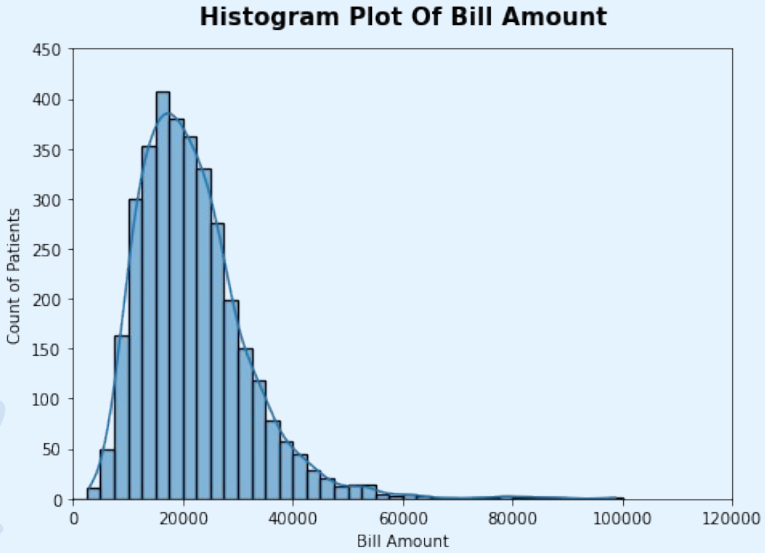
03 Exploration & Visualisation

04 Pre-Processing & Modelling

05 Finale

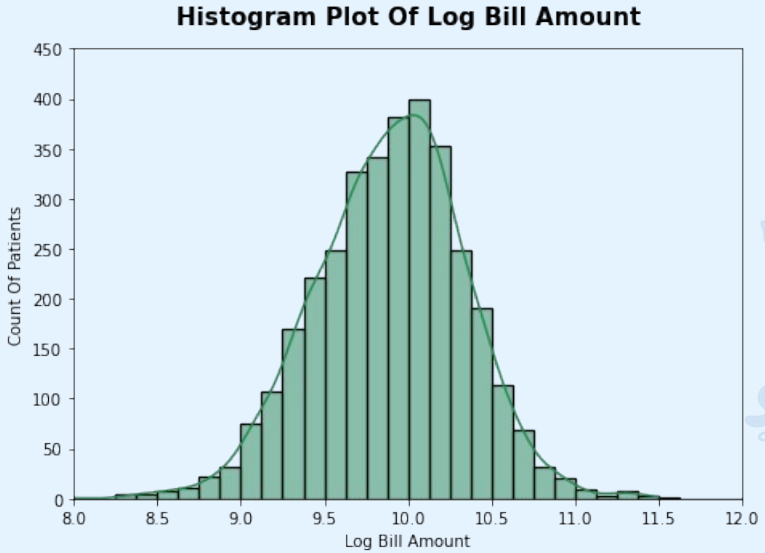
Target Visualisations

Bill Amount



The distribution of bill_amount is asymmetrical with a positive or right skew indicating several patients who have extremely large bill amounts.

Log Bill Amount

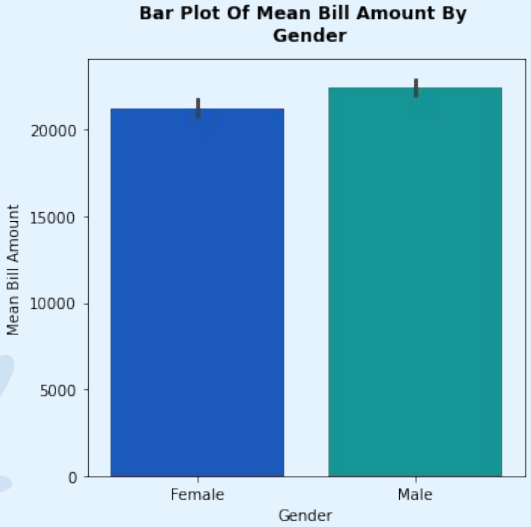


The distribution of log_bill_amount is more symmetrical with less positive or right skew.

Categorical Feature Visualisations

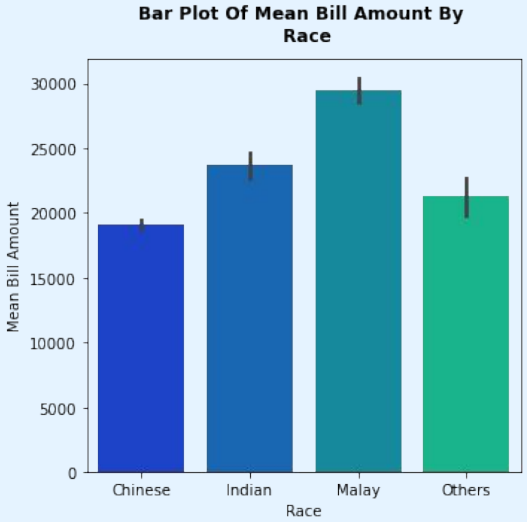
Demographic Features

Gender



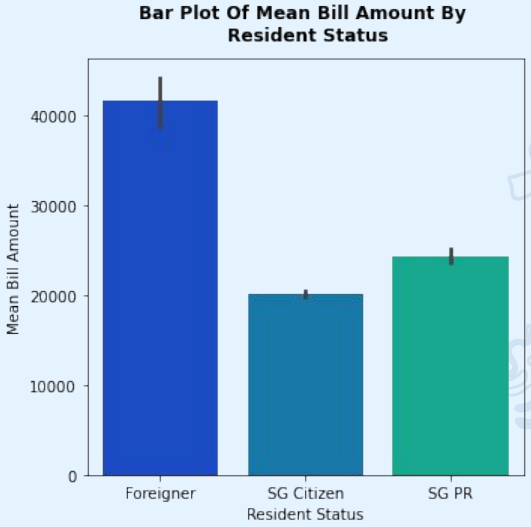
Male patients have a slightly higher mean bill amount than female patients.

Race



Malay patients have the highest mean bill amount whereas Chinese patients have the lowest.

Resident Status

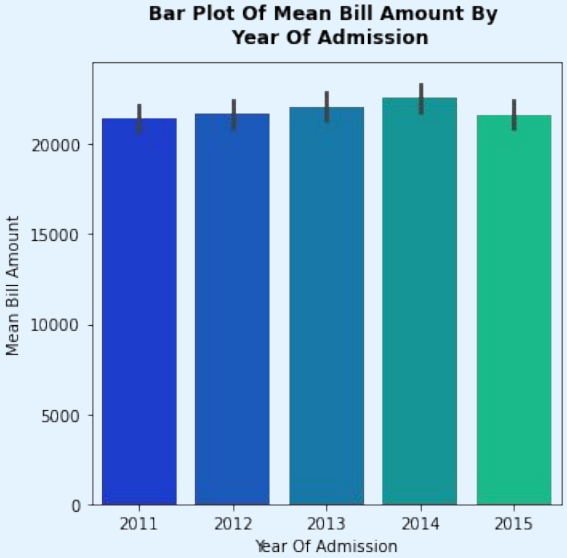


Foreigners have the highest mean bill amount, followed by Singapore PRs, and then Singapore Citizens.

Categorical Feature Visualisations

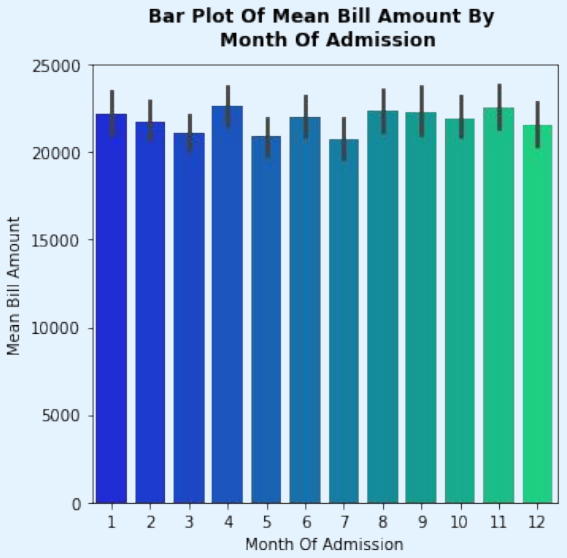
Time Features

Year Of Admission



The mean bill amount has stayed uniform across all years.

Month Of Admission

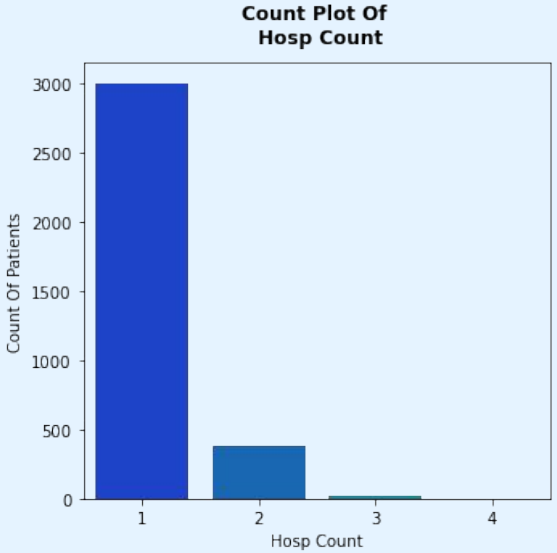


The mean bill amount does not seem to follow any trend across all months.

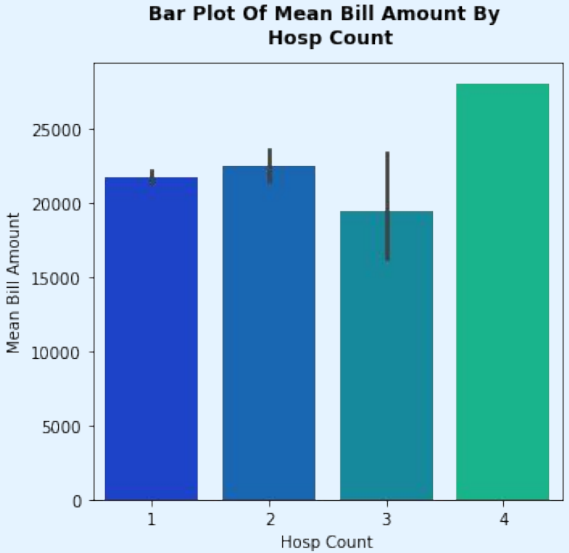
Categorical Feature Visualisations

Hospitalisation Features

Hospitalisation Count



The vast majority of patients have been admitted to the hospital only once.

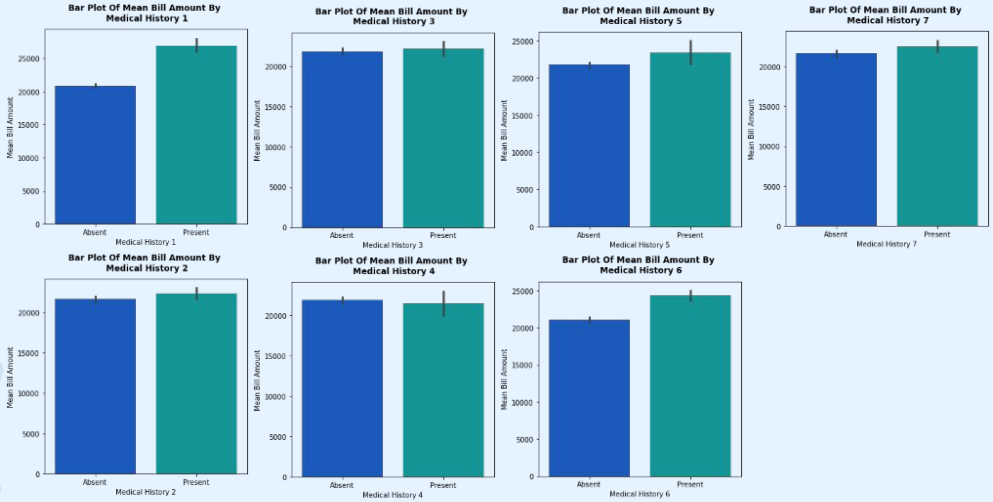


The mean bill amount appears to increase as hospitalisation count increases.

Categorical Feature Visualisations

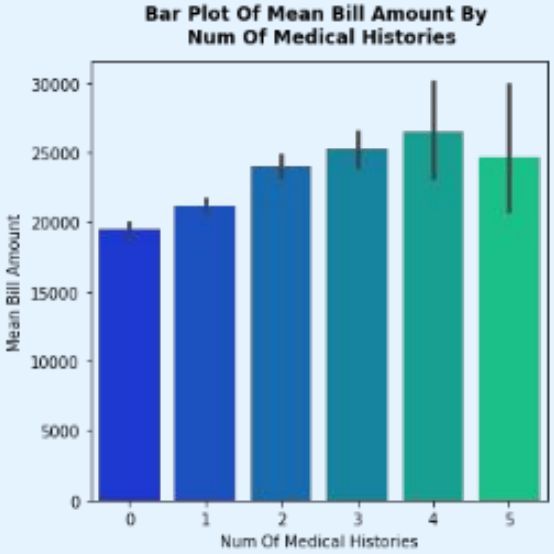
Clinical Features

Medical History 1-7



In all medical histories except medical_history_4, patients who have that medical history have a higher mean bill amount than patients who do not. In particular, patients with medical_history_1 or medical_history_6 pay considerably higher bills than patients without them.

Number Of Medical Histories

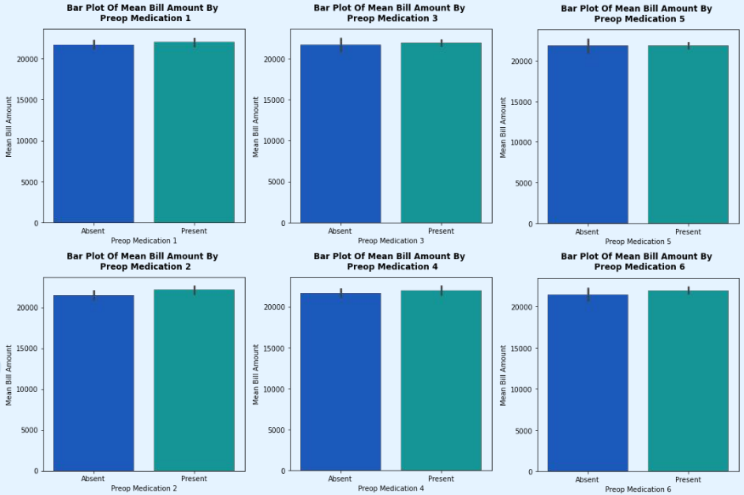


There appears to be a subtle positive relationship between the number of medical histories and the bill amount.

Categorical Feature Visualisations

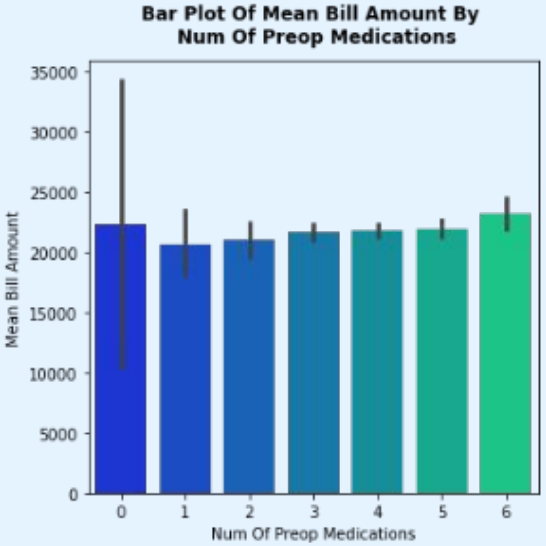
Clinical Features

Pre-Op Medication 1-6



For each pre-op medication, the mean bill amount for patients who received that pre-op medication is roughly the same as that of patients who did not.

Number Of Pre-Op Medications

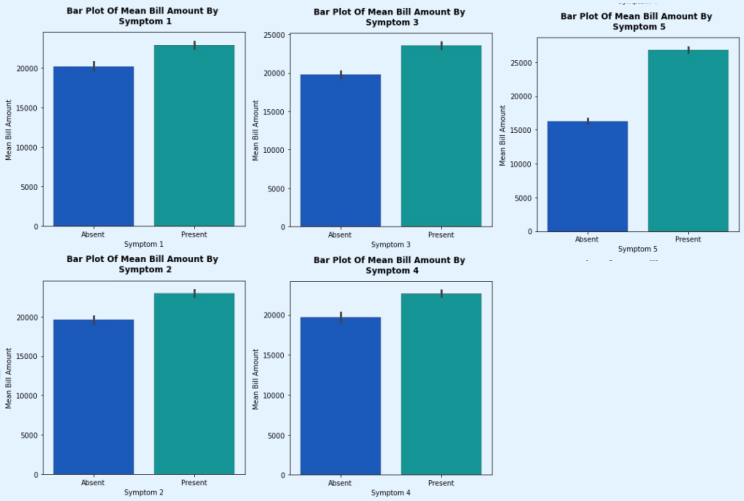


The mean bill amount across the different number of pre-op medications appears to show a slight increasing trend.

Categorical Feature Visualisations

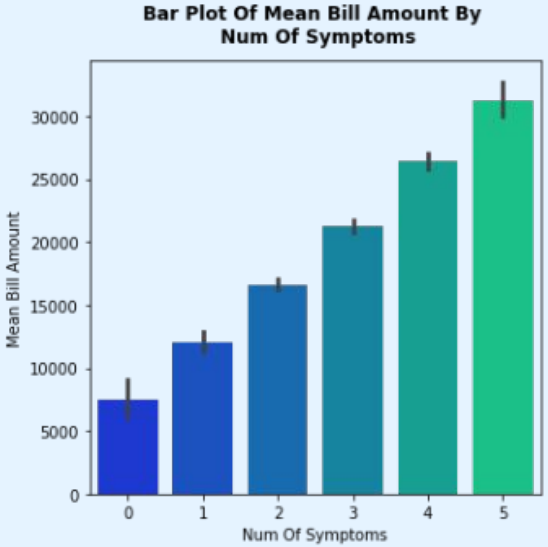
Clinical Features

Symptom 1-5



Across all symptoms, the mean bill amount of patients who have that symptom are higher than that of patients who do not. Of note, patients with symptom_5 have a mean bill amount that is drastically higher than patients without it.

Number Of Symptoms



There appears to be a strong positive correlation between the number of symptoms and the bill amount.

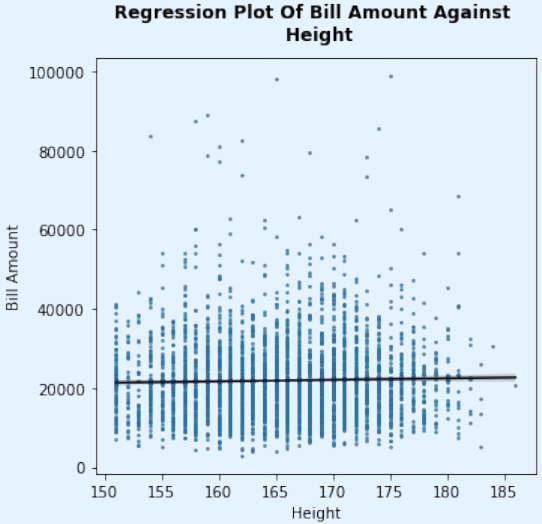
Numerical Feature Visualisations

Physical Features

Weight

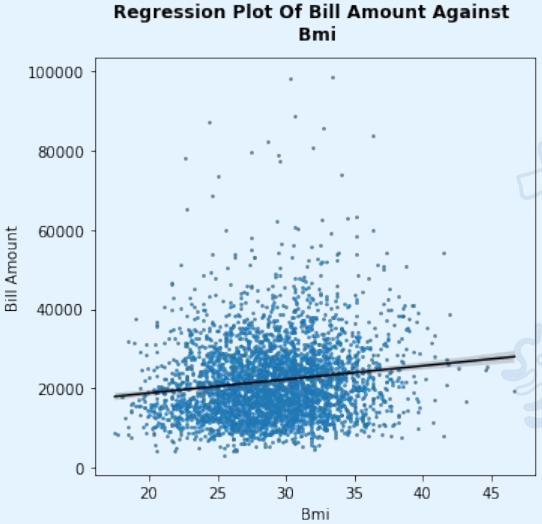
There is a positive correlation between weight and bill amount.

Height



There is no correlation between height and bill amount.

Body Mass Index

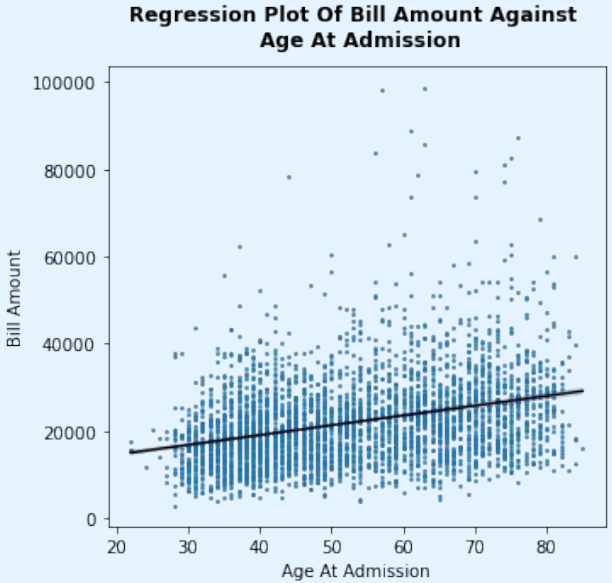


There is a positive correlation between body mass index and bill amount.

Numerical Feature Visualisations

Physical Features

Age At Admission



There is a positive correlation between age at admission and bill amount.

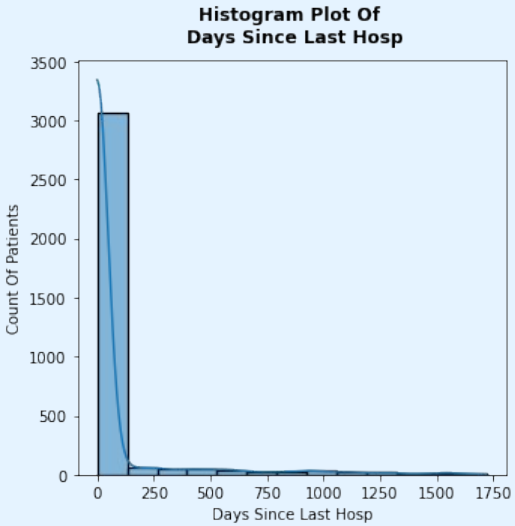
Numerical Feature Visualisations

Hospitalisation Features

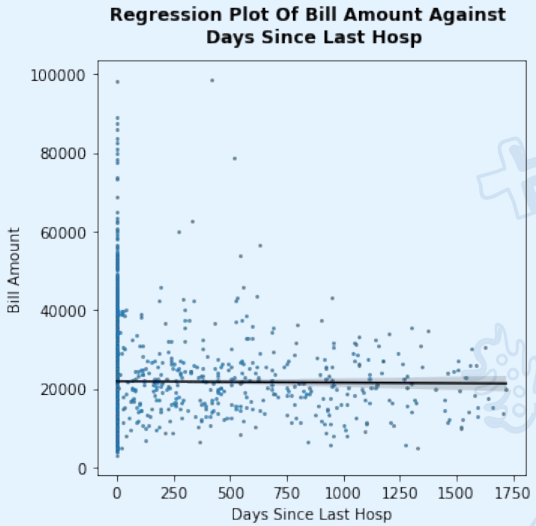
Days Hospitalised

There appears to be no correlation between days hospitalised and bill amount.

Days Since Last Hospitalisation



Days since last hospitalisation is predominantly 0.

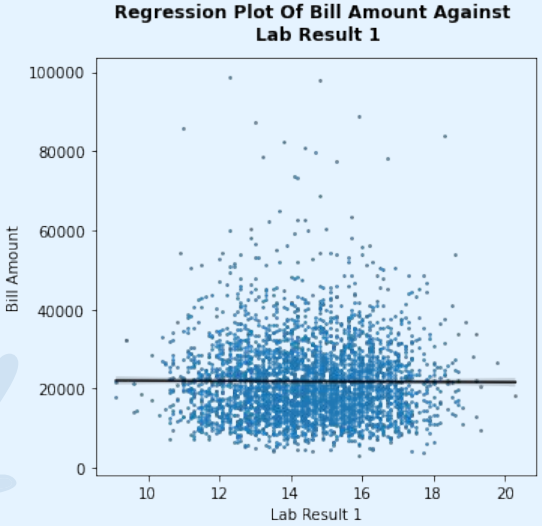


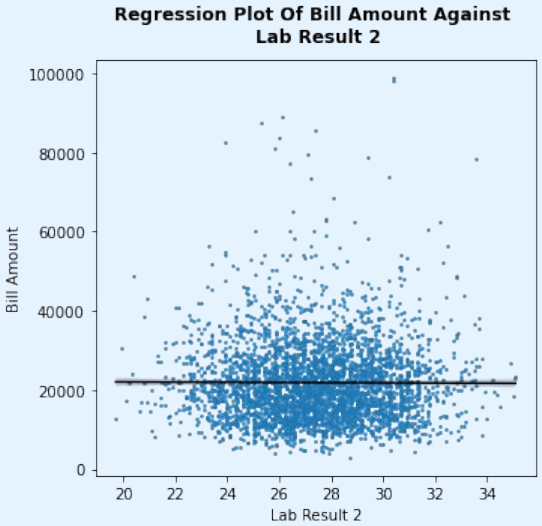
Days since last hospitalisation and bill amount have no correlation.

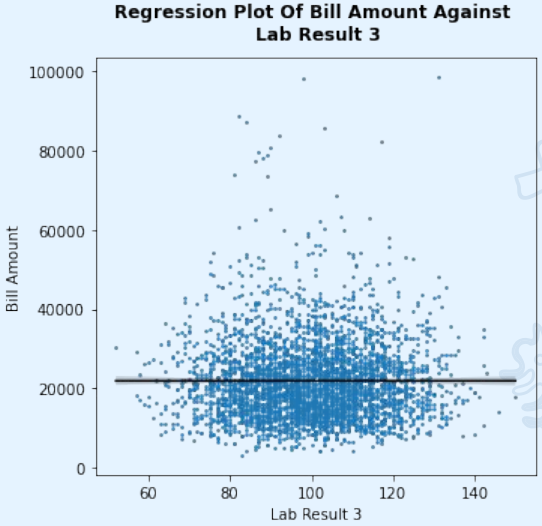
Numerical Feature Visualisations

Clinical Features

Lab Result 1-3

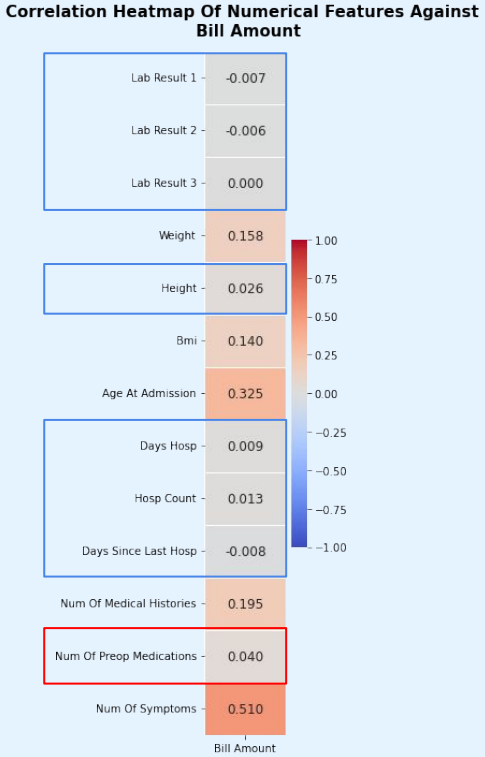






There is no correlation between any of the lab results and bill amount.

Correlation Heatmaps



Linearity

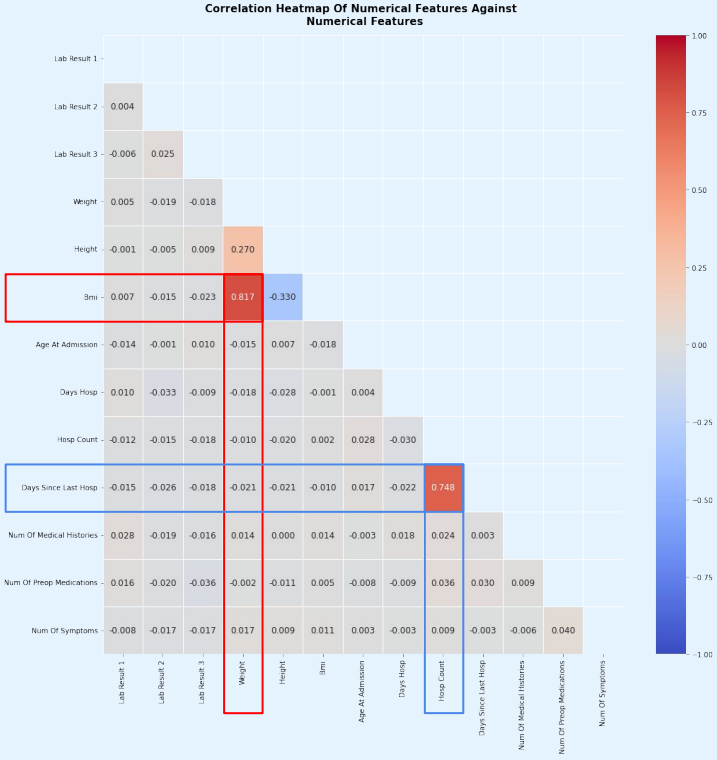
We want to eliminate the features that are **weakly correlated with the target** to **improve linearity**.

The numerical features which are weakly correlated with the target (**absolute correlation coefficient < 0.1**) include:

- lab_result_1
- lab_result_2
- lab_result_3
- height
- days_hosp
- hosp_count
- days_since_last_hosp
- num_of_preop_medications

Correlation Heatmaps

Multicollinearity



We want to eliminate pairs of features that are **strongly correlated with each other** to **reduce multicollinearity**.

The pairs of numerical features which are strongly correlated with each other (**absolute correlation coefficient > 0.7**) include:

- hosp_count and days_since_last_hosp
- weight and bmi

Outline

Introduction

Data Preparation

Exploration & Visualisation

Pre-Processing & Modelling

Finale

Modelling Workflow

Final Features

Gender
Race
Resident Status
Medical History 1-7
Number Of Medical Histories
Symptom 1-5
Number Of Symptoms
Weight
Age At Admission

Final Targets

Bill Amount
Log Bill Amount

Feature One-Hot Encoding

↓

Train Test Split

↓

Feature Scaling

LINEAR REGRESSION
- bill_amount
- log_bill_amount

+

RIDGE REGRESSION
- bill_amount
- log_bill_amount

+

LASSO REGRESSION
- bill_amount
- log_bill_amount

TRAIN

TEST

BEST MODEL and BETTER TARGET

R^2

RMSE

Model Evaluation & Selection

No	Model	Target	Train R^2	Test R^2	R^2 Difference	Train RMSE	Test RMSE	RMSE Difference
1	Linear Regression	bill_amount	0.929	0.921	-0.008	2674.948	2934.107	259.159
		log_bill_amount	0.973	0.973	-0.000	1633.142	1700.389	67.246
2	Ridge Regression	bill_amount	0.929	0.921	-0.008	2674.312	2929.862	255.550
		log_bill_amount	0.973	0.973	-0.000	1639.241	1713.660	74.420
3	Lasso Regression	bill_amount	0.929	0.921	-0.008	2674.413	2931.671	257.259
		log_bill_amount	0.974	0.973	-0.000	1628.886	1702.975	74.089

Best Model: Lasso Regression

Better Target: Log Bill Amount

Model Intercept & Coefficients

Regression Model Equation: $\log y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$

where $\log y$ is log_bill_amount or the **natural log** of bill_amount

Intercept

$\beta_0 = 9.894$

$e^{\wedge}(\beta_0) = 19{,}815$

The intercept gives the **target** when **all the features are equal to 0**.

The lasso regression model predicts the **bill amount to be $19,815** for a patient who is:

- Female, Chinese, Singapore Citizen
- No medical history, No symptom
- Weight of 0kg, Age at admission of 0 years

Coefficients

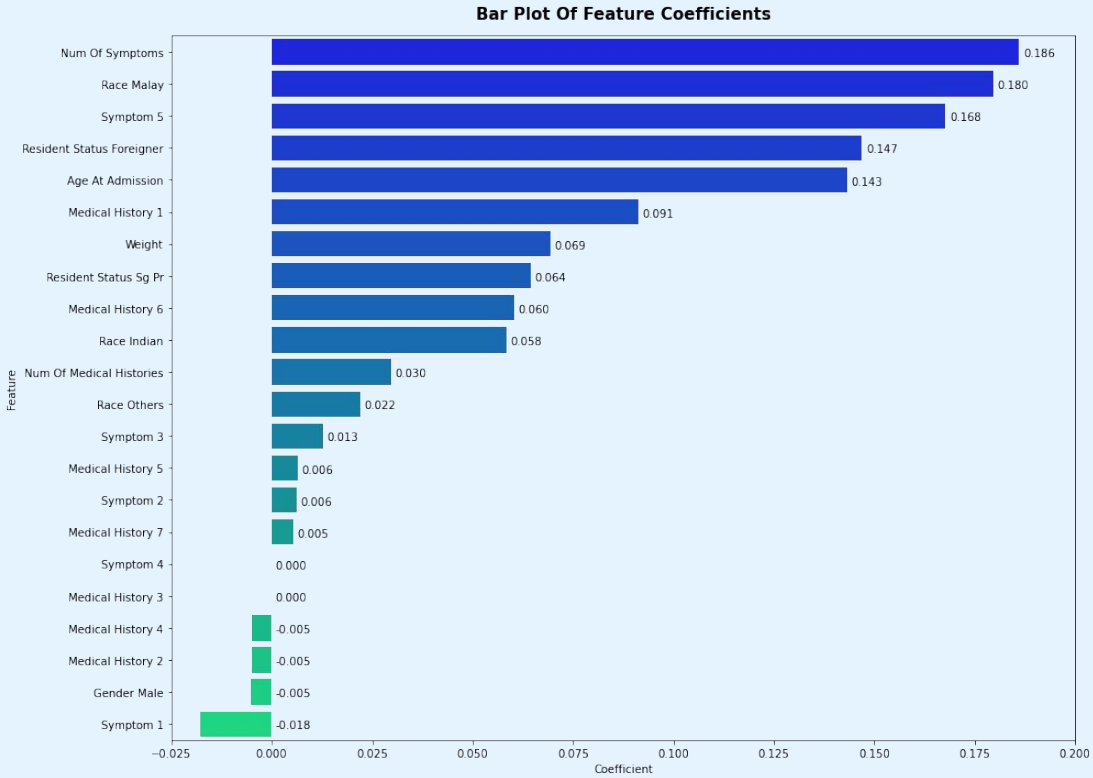
	Feature	Coefficient	$e^{Coefficient}$	Standard Deviation
15	num_of_symptoms	0.186062	1.204497	1.068724
18	race_Malay	0.179578	1.196712	0.405894
11	symptom_5	0.167748	1.182638	0.499357
20	resident_status_Foreigner	0.146943	1.158288	0.212424
13	age_at_admission	0.143180	1.153937	14.710728
0	medical_history_1	0.091343	1.095645	0.374911
12	weight	0.069473	1.071943	10.975076
21	resident_status_SG PR	0.064482	1.066606	0.358560
5	medical_history_6	0.060337	1.062195	0.435760
17	race_Indian	0.058457	1.060200	0.301607
14	num_of_medical_histories	0.029677	1.030122	0.971992
19	race_Others	0.022125	1.022371	0.225702
9	symptom_3	0.012837	1.012920	0.498071
4	medical_history_5	0.006442	1.006463	0.234222
8	symptom_2	0.006233	1.006252	0.472977
6	medical_history_7	0.005476	1.005491	0.435594
10	symptom_4	0.000000	1.000000	0.445836
2	medical_history_3	0.000000	1.000000	0.343026
3	medical_history_4	-0.004852	0.995160	0.222178
1	medical_history_2	-0.004921	0.995091	0.453829
16	gender_Male	-0.005144	0.994869	0.500073
7	symptom_1	-0.017802	0.982356	0.485530

The interpretation of the features with respect to the target is **complicated by 2 factors**:

- log transformation of the target
- standard scaling of the features

In here, the coefficients represent how much a **1 standard deviation change** in the **feature** would lead to a **certain percentage change** in the **target**.

Feature Coefficients



The effect of gender on the bill amount is negligible.

Differences in race have a significant effect on the bill amount.

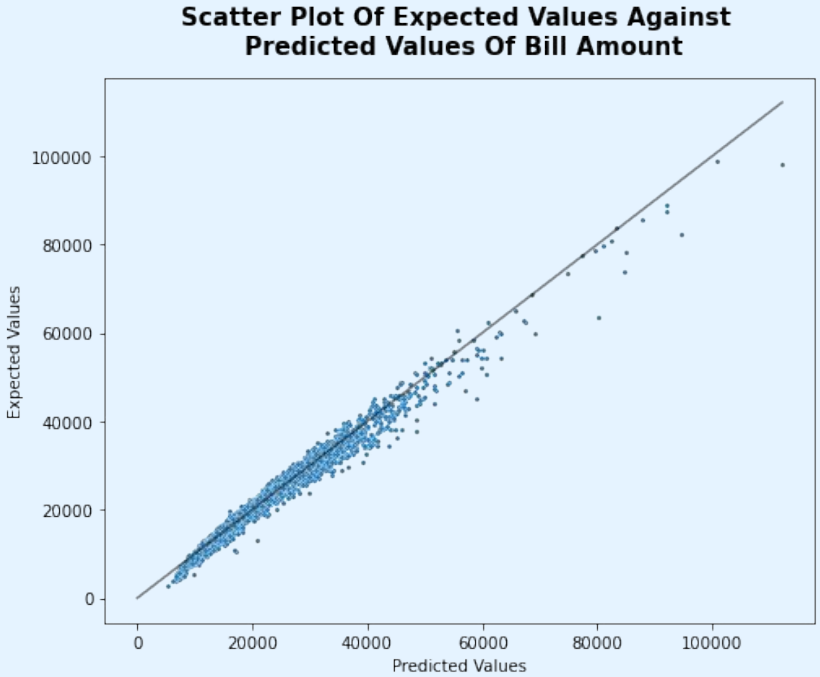
Differences in resident_status have a significant effect on the bill amount.

symptom_5, medical_history_1, and medical_history_6 have a greater impact on the bill amount.

num_of_symptoms is the most important feature explaining bill amount but num_of_medical_histories is not an important feature explaining bill amount.

age_at_admission and weight are relatively important features affecting bill amount.

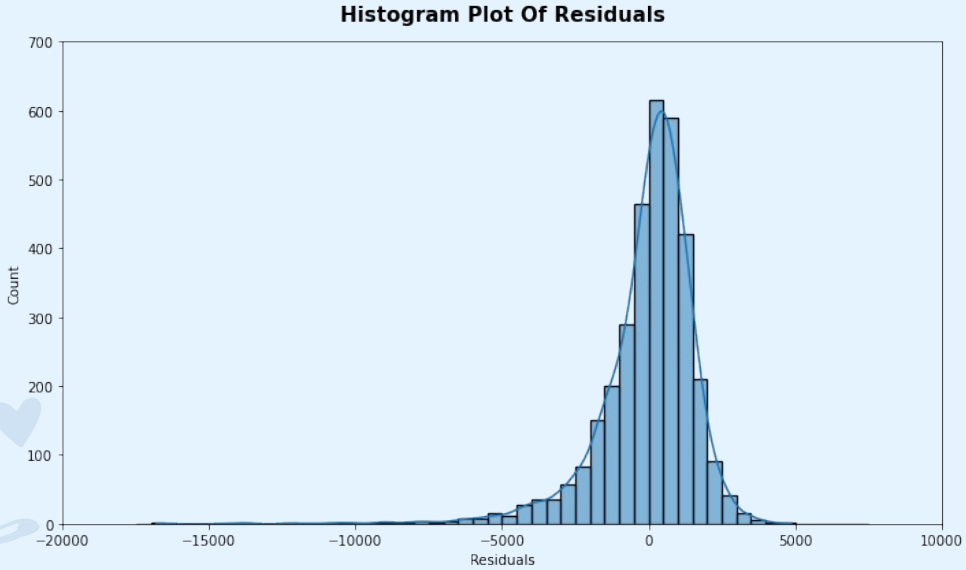
Expected Values Vs Predicted Values



The predicted values by the lasso regression model are very similar to the expected values in the dataset.

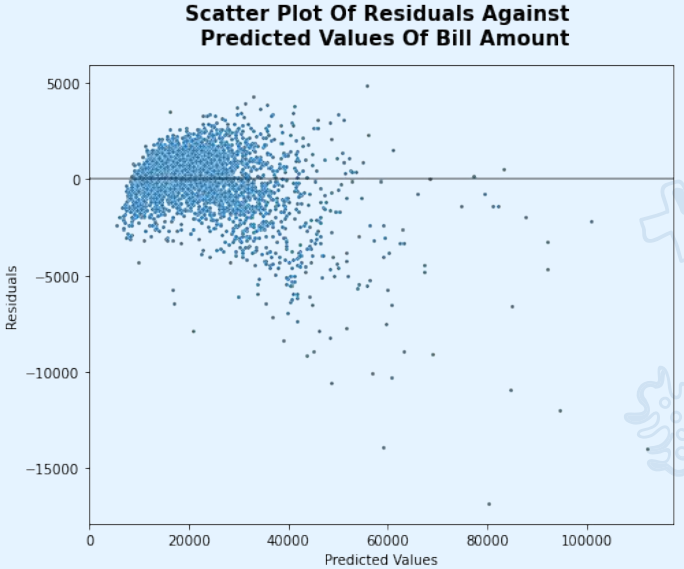
Residuals

Normality Check



The distribution of residuals is asymmetrical with a negative or left skew indicating there are a few predicted values that are way higher than the expected values.

Homoscedasticity Check



For smaller predicted values, there is a somewhat equal distribution of data points above and below the zero line. For bigger predicted values, the data points are mainly below the zero line.

Final Model Vs Null Model

Final Model

Lasso Regression

Null Model

$21,859
Baseline Prediction

0.973	Vs	0.000
Final R^2		Baseline R^2
1,654	Vs	10,153
Final RMSE		Baseline RMSE

Outline

Introduction

Data Preparation

Exploration & Visualisation

Pre-Processing & Modelling

Finale

Recommendations

Symptoms & Medical Histories

symptom_5, medical_history_1, and medical_history_6 are key features.

ABC Health Insurance could construct all their policies to include **mandatory entry-level coverage** of these conditions.

Race

race plays a huge role. Malay patients are predicted to have the highest bill amount > Indian > Other > Chinese.

ABC Health Insurance would need to investigate if there are **confounding variables** that could explain the **racial differences** in the bill amount.

Weight

weight has a moderately important place. Patients with heavier weights are predicted to have larger bill amounts.

ABC Health Insurance could begin to provide **coverage and benefits** for patients seeking **obesity treatment**.

Limitations

Absence of critical features

Information on some of the **major determinants of hospitalisation costs** such as type of surgery, experience of attending doctor, any imaging done, and ward class are **absent**.

The **presence** of these predictors could potentially lead to a **better estimation** of bill amount.

Inadequacy of existing features

Due to an **anonymisation** of the data, features such as medical_history, preop_medication, symptom, and lab_result have been left **vague** intentionally.

This makes it challenging to apply any **domain-specific knowledge** into the creation of **interaction features**.

Limited data on recurring hospitalisations

Of the 3,400 rows in the dataset, 3,000 rows were **first-time admissions** and 400 rows were **repeating admissions**.

To **better study** the effect of recurring hospitalisations on bill amount, **more data** on repeating admissions would be needed.

Conclusion

The final lasso regression model has performed well as it is
Able to explain 97.3% of the total variability in the bill amount with all the features
Generally off-the-mark by 7% on its predictions of the bill amount

It has found the main factors driving cost of care to be
num_of_symptoms (most important feature predicting bill amount)
race, resident_status, age_at_admission, weight (among the top few features explaining bill amount)

The future plans for this project could involve
Extending it to other cities and countries to find the factors driving their cost of care
Adapting it to predict hospital bills at pre-admission to help reduce bill shocks

Thank You