Holmusk Data Assignment

Michelle Ng

Problem Statement

- How is costing affected by length stay in the hospital?
- If not, what are the major determinants of cost?

Datasets

- Clinical Data
  - 22 continuous data columns (medical history (7), preop medication (6), symptoms(5), lab_results(3), weight, height)
  - 4 categorical data columns (id, dates (2) , medical_history_3)
- Transaction Data
  - 1 continuous data column (bill id)
  - 2 categorical data columns (patient id, date of admission)
- Demographic Data
  - 5 categorical data columns (patient id, gender, race, resident_status, date of birth)
- Cost Data
  - 2 continuous data columns (bill id, amount)

Missing data and changes in data

- 502 rows of missing data
  - Decided to drop the records, as there is no meaningful way to impute the data
- Mapped errors in recording
  - F, m to female, male
  - India to Indian
  - Etc.
- Date to datetime variables
- Billing data - sum of cost per patient per admission

Feature Engineering

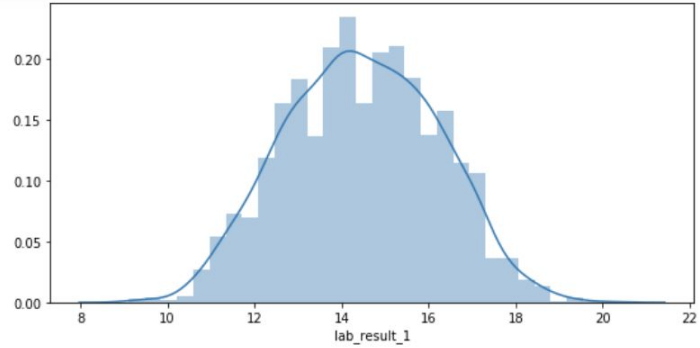
- Calculated hospital stay in days
  - (discharge - admission)

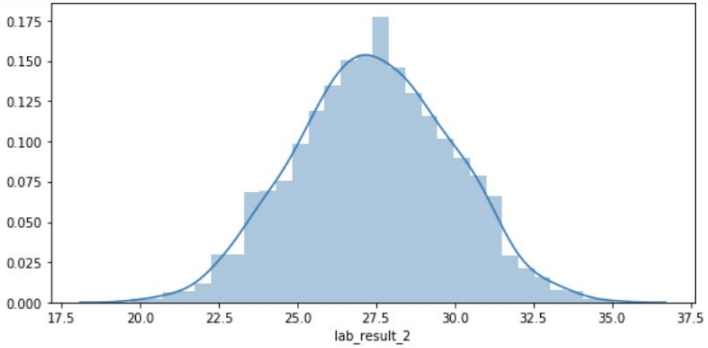
- Calculated age of patient upon admission
  - (admission - date of birth)

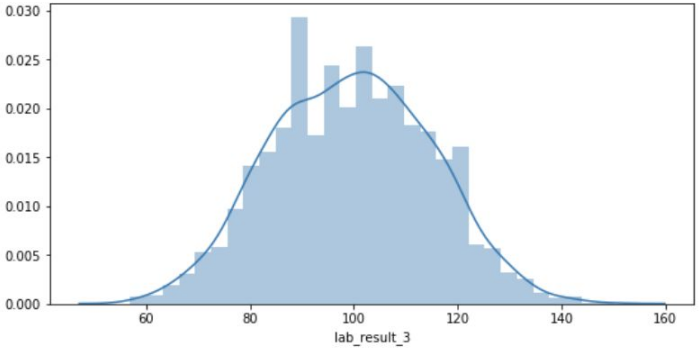
Assumptions

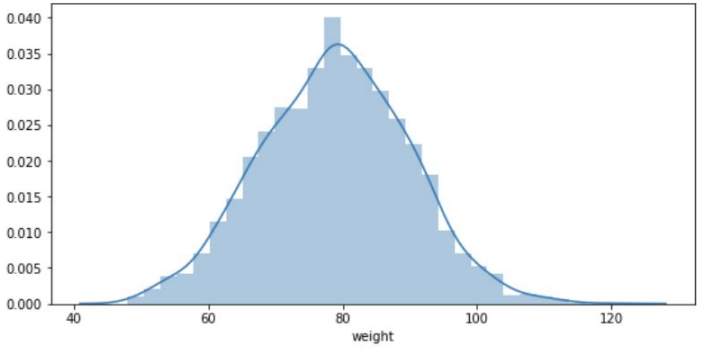
- Central Limit Theorem
  - Large dataset (n= 2898)
- Medical history, preop medications, symptoms
  - Dummy coded (0 no, 1 yes)

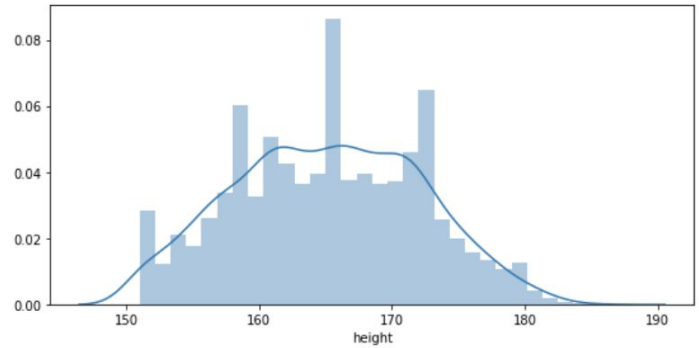
Visualisations

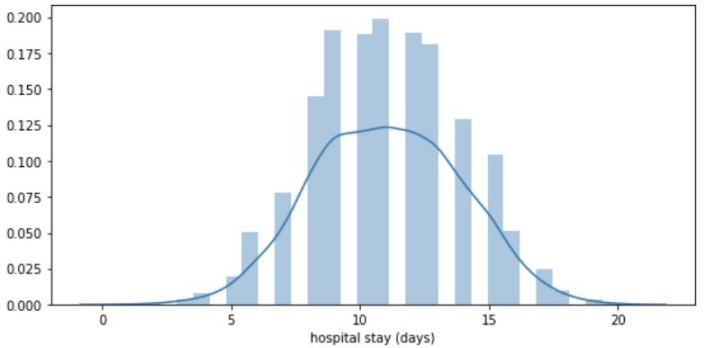

0.20
0.15
0.10
0.05
0.00
8
10
12
14
16
18
20
22
lab_result_1

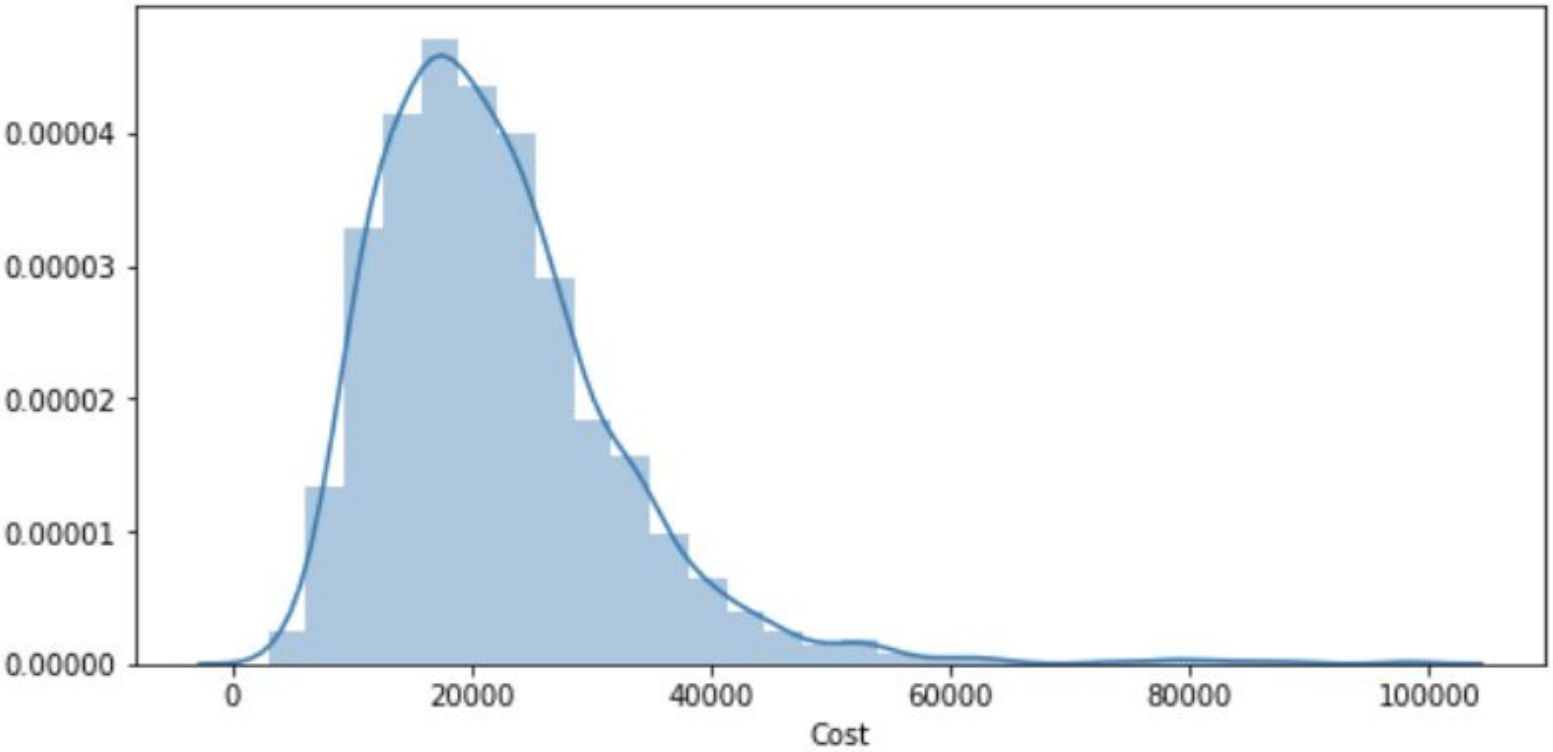

0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
lab_result_2

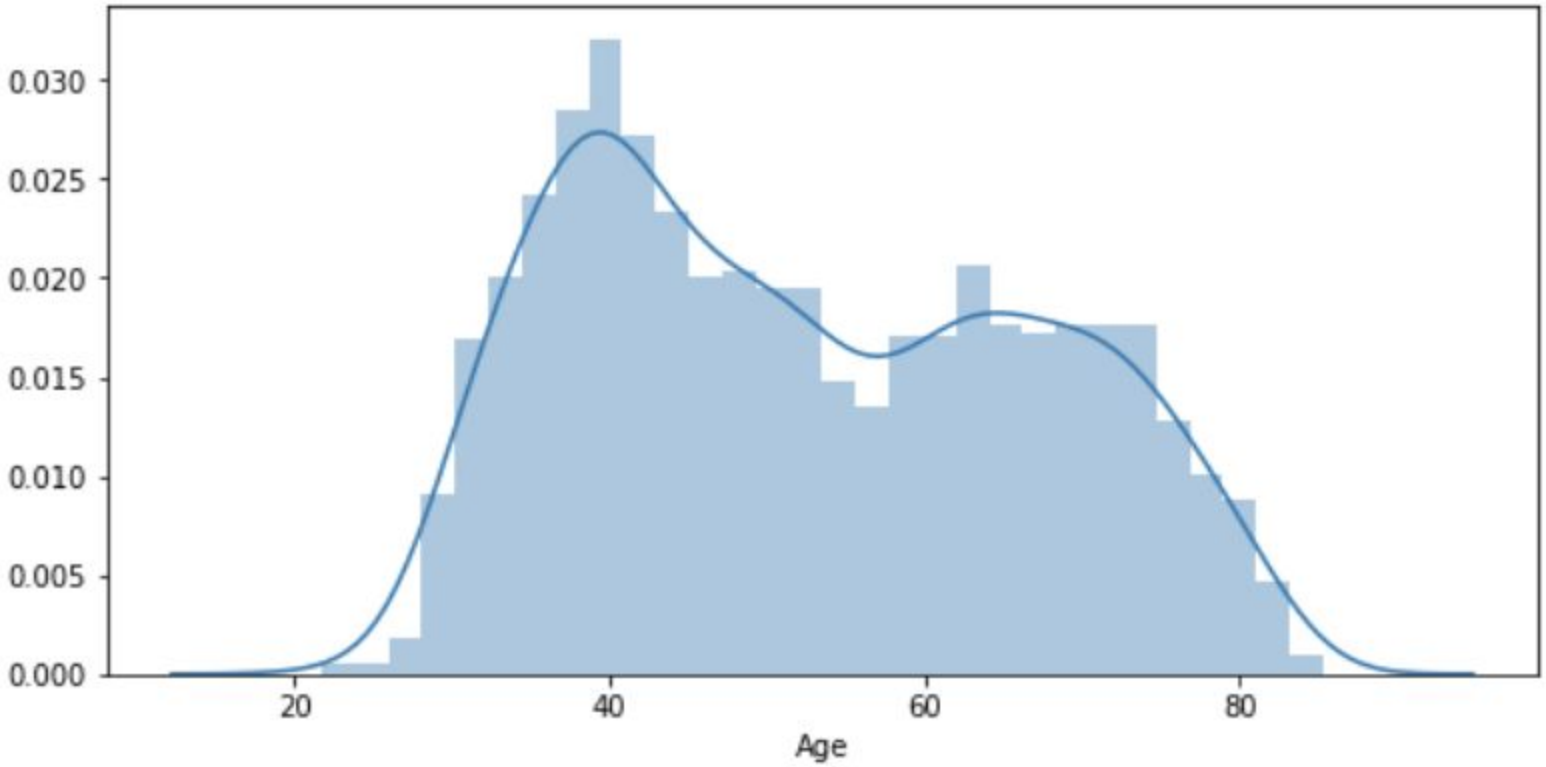

0.030
0.025
0.020
0.015
0.010
0.005
0.000
60
80
100
120
140
160
lab_result_3

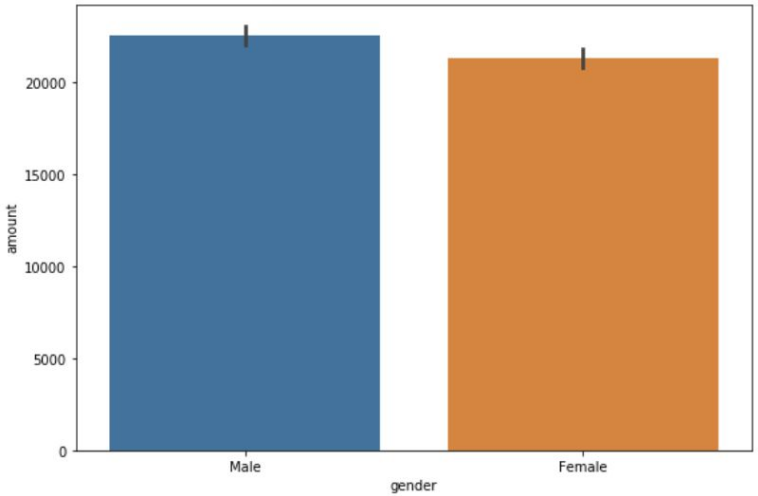

0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
40
60
80
100
120
weight

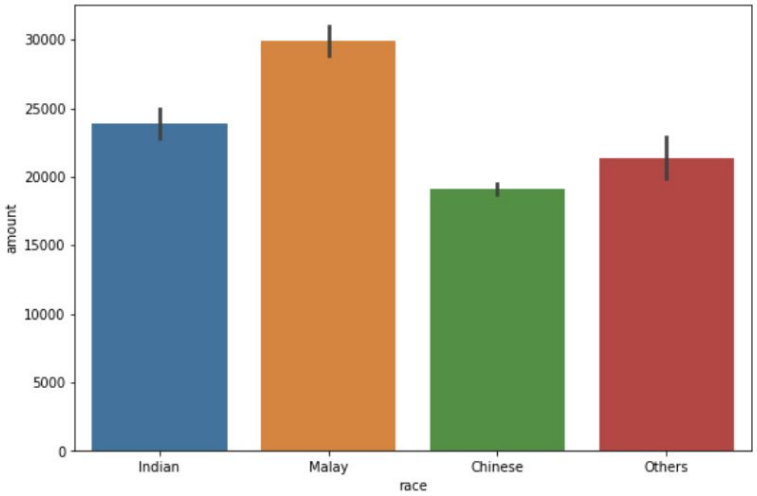

0.08
0.06
0.04
0.02
0.00
150
160
170
180
190
height

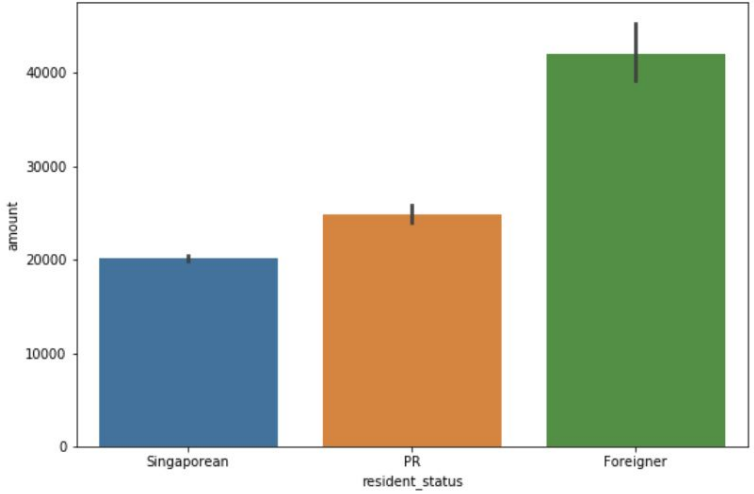

0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
5
10
15
20
hospital stay (days)

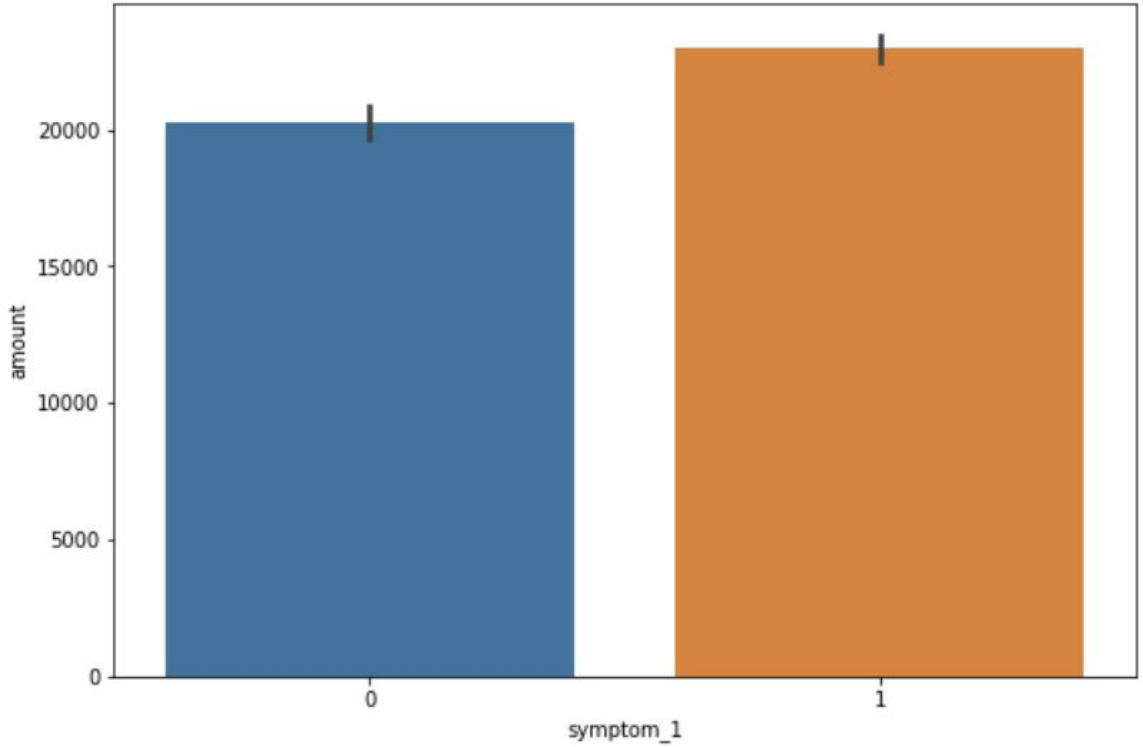

0.00004
0.00003
0.00002
0.00001
0.00000
0
20000
40000
60000
80000
100000
Cost

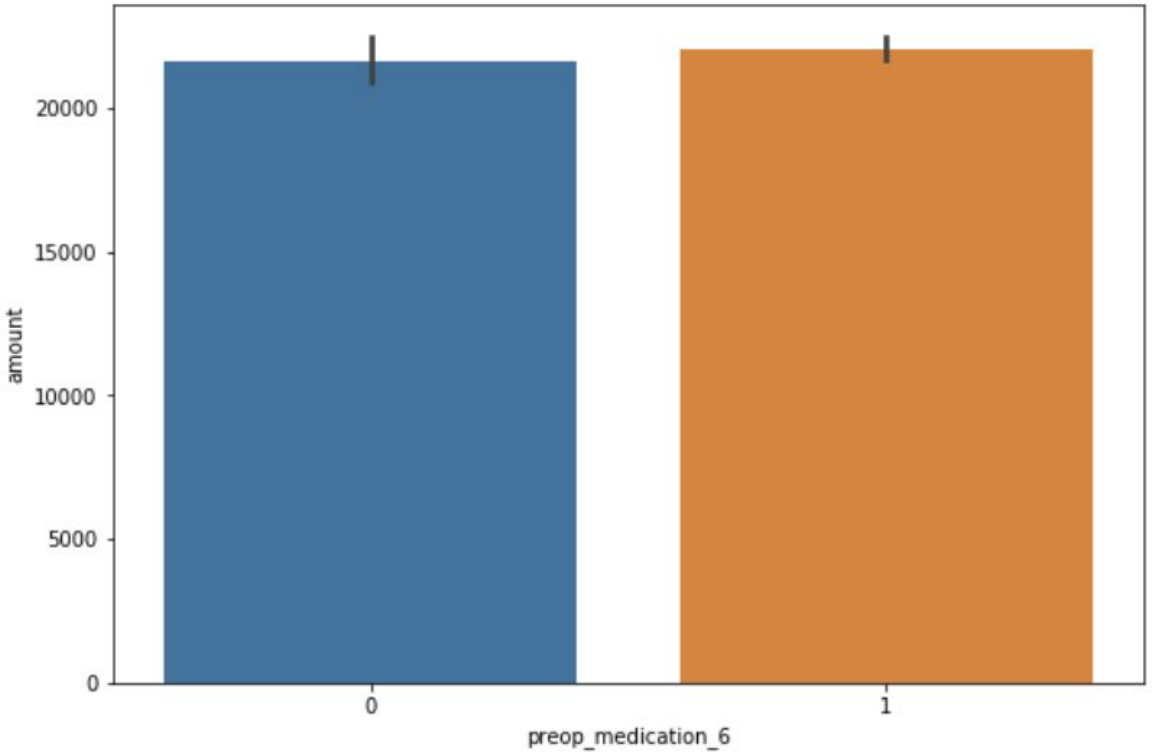

0.030
0.025
0.020
0.015
0.010
0.005
0.000
20
40
60
80
Age


amount
20000
15000
10000
5000
0
Male
Female
gender


amount
30000
25000
20000
15000
10000
5000
0
Indian
Malay
Chinese
Others
race


amount
40000
30000
20000
10000
0
Singaporean
PR
Foreigner
resident_status

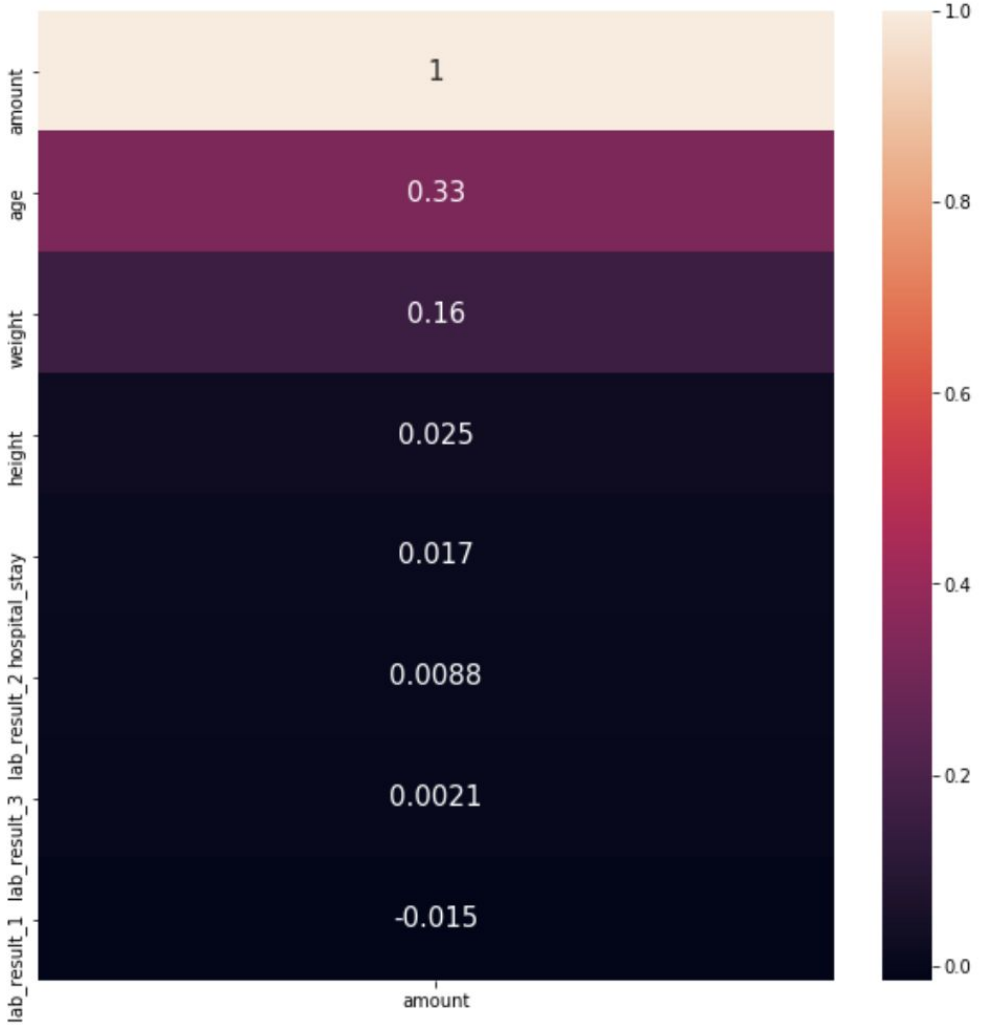

20000
15000
10000
5000
0
amount
0
1
symptom_1


20000
15000
10000
5000
0
amount
0
1
preop_medication_6

Statistical Tests

-

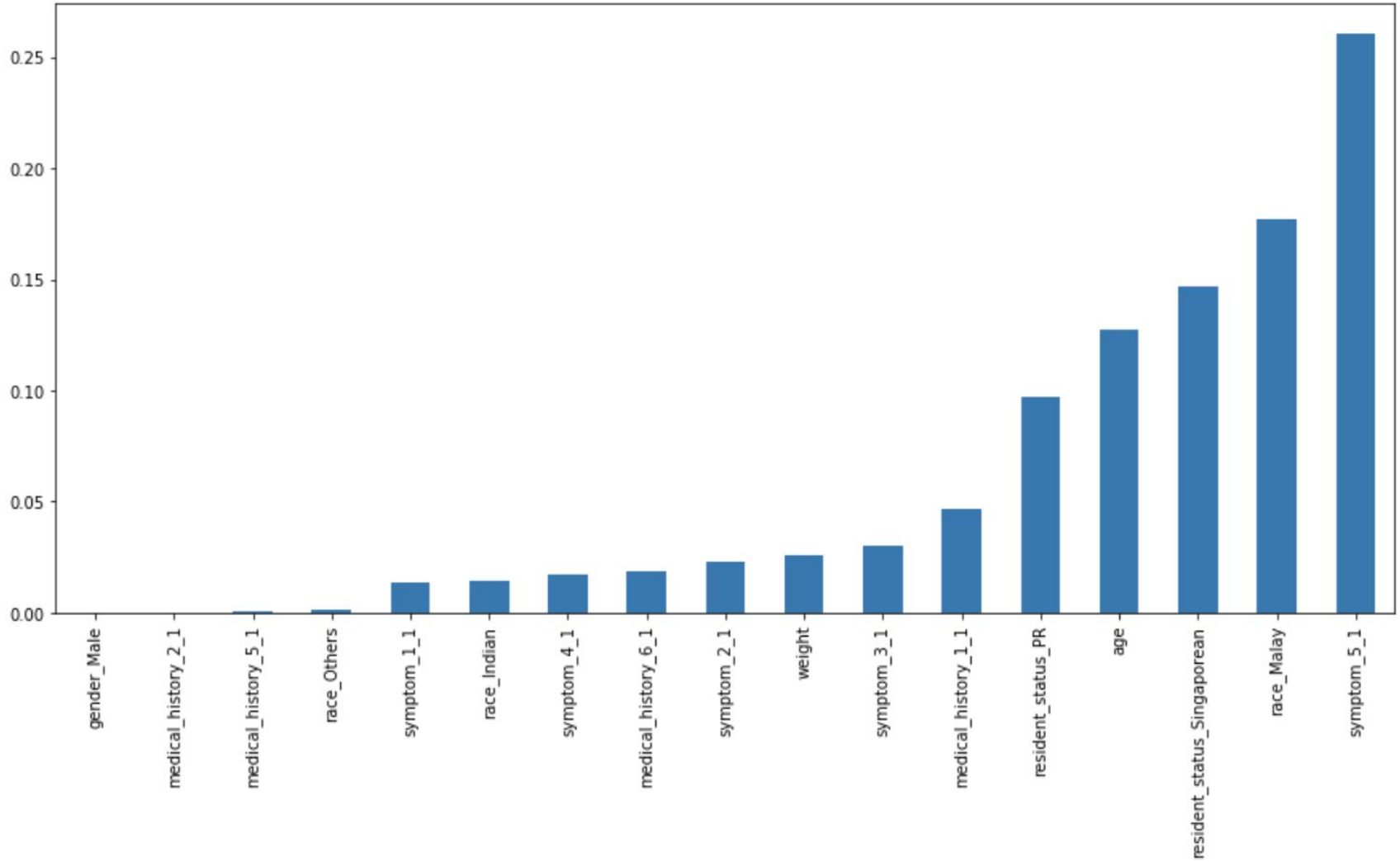
Correlation test



T-test/ANOVA

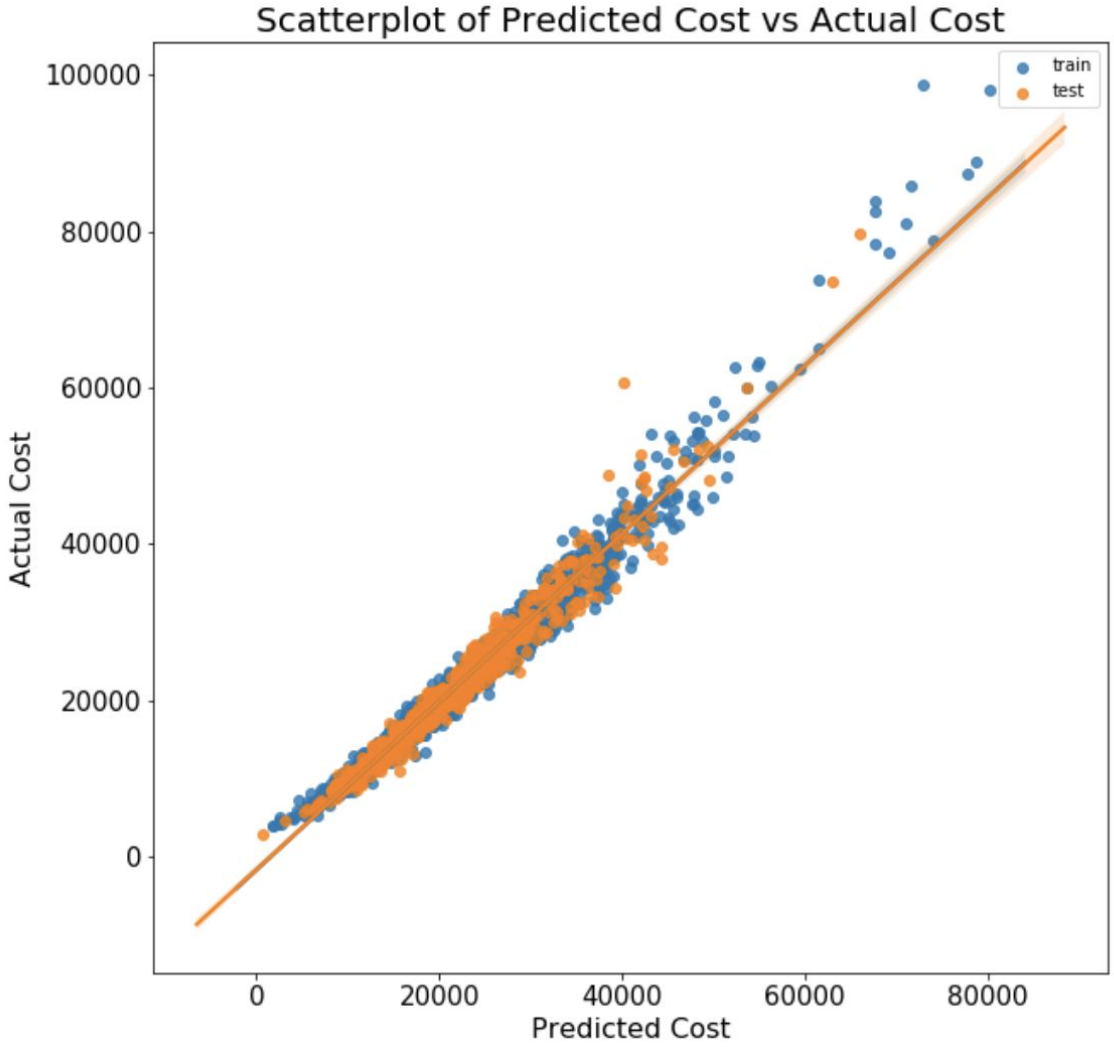
- Mean cost of treatment compared between categories of the same variable
- Significant difference in cost means:
  - Medical history 1, 2, 5, 6
  - Symptom 1, 2, 3, 4, 5
  - Gender
  - Race
  - Resident status

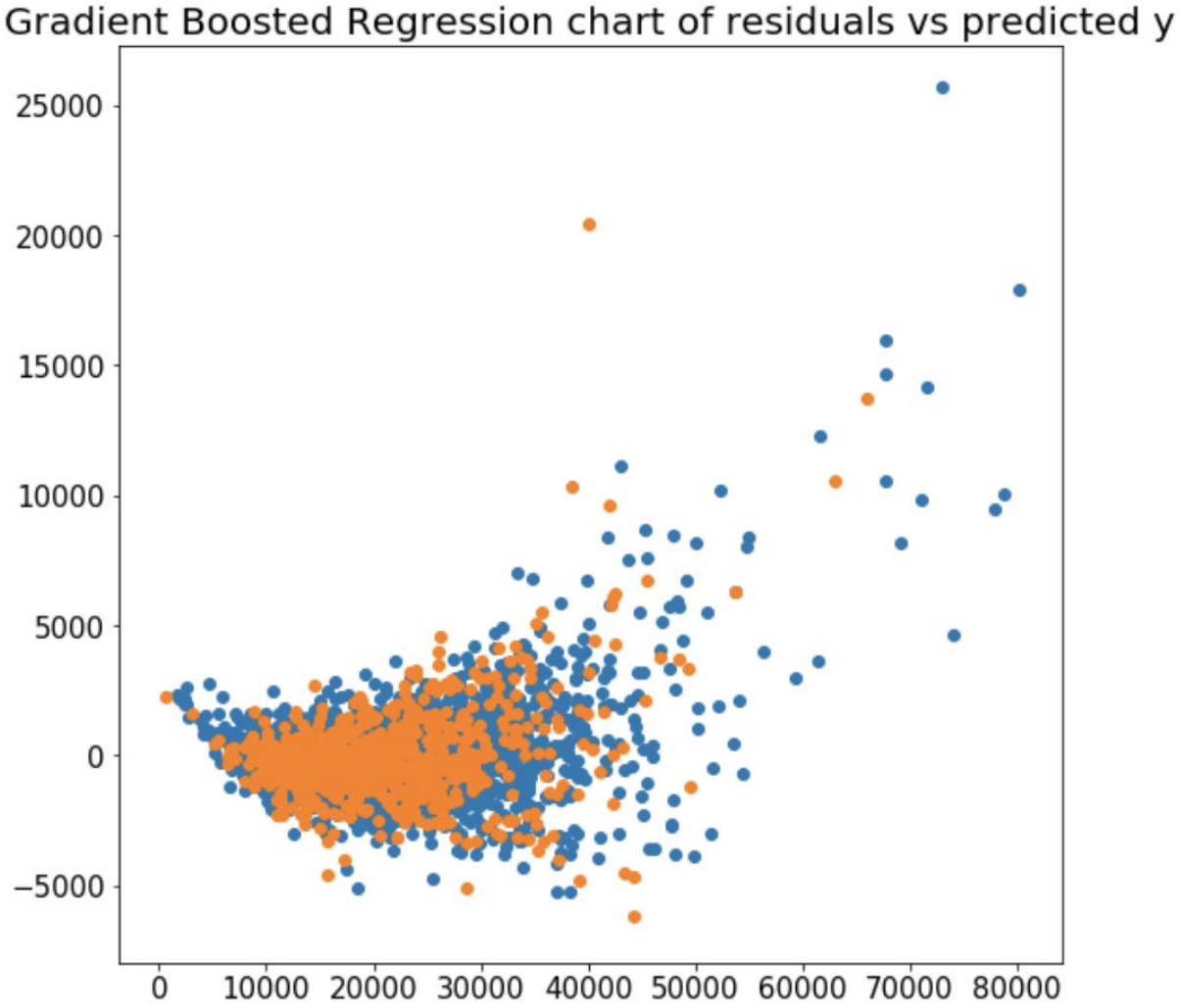
Models	Train Score	Test Score	RMSE
Linear Regression	0.921	0.927	2557.7
Lasso	0.921	0.927	2553.7
Ridge	0.921	0.927	2557.6
ElasticNet	0.751	0.799	4240.7
Random Forest	0.981	0.893	3015.6
Gradient Boost	0.970	0.958	1947.4
AdaBoost	0.532	0.388	7624.6


0.25
0.20
0.15
0.10
0.05
0.00
gender_Male
medical_history_2_1
medical_history_5_1
race_Others
symptom_1_1
race_Indian
symptom_4_1
medical_history_6_1
symptom_2_1
weight
symptom_3_1
medical_history_1_1
resident_status_PR
age
resident_status_Singaporean
race_Malay
symptom_5_1

	ranking
symptom_5_1	0.260842
race_Malay	0.176679
resident_status_Singaporean	0.146738
age	0.127064
resident_status_PR	0.097446

Problems with Model


Scatterplot of Predicted Cost vs Actual Cost
train
test
Actual Cost
Predicted Cost
100000
80000
60000
40000
20000
0
0
20000
40000
60000
80000


Gradient Boosted Regression chart of residuals vs predicted y
25000
20000
15000
10000
5000
0
−5000
0
10000
20000
30000
40000
50000
60000
70000
80000

Conclusions

- Stay in hospital is not highly correlated to cost of treatment (r = 0.017)
- Gradient Boosting model was best able to account for the training and test dataset.
- Symptom 5 is the most important feature, followed by race and age.

Limitations and Improvements

Limitations

- Model is not accurate for extreme ends of cost
- Model is not homoscedastic
- Each treatment is considered as independent from each other

Improvements to be made

- Exploring the effect of previous treatments on the cost
- Exploring other parameters for GridSearch
- Exploring what is the main cause of error at the extreme ends