Healthcare Data Challenge

KEVIN GOH

AUGUST 2021

The Problem

- We have clinical and financial data of patients hospitalized for a certain condition
- There are several factors that can affect the cost of care
- As such, we want to find insights about the drivers of cost of care for patients of this condition

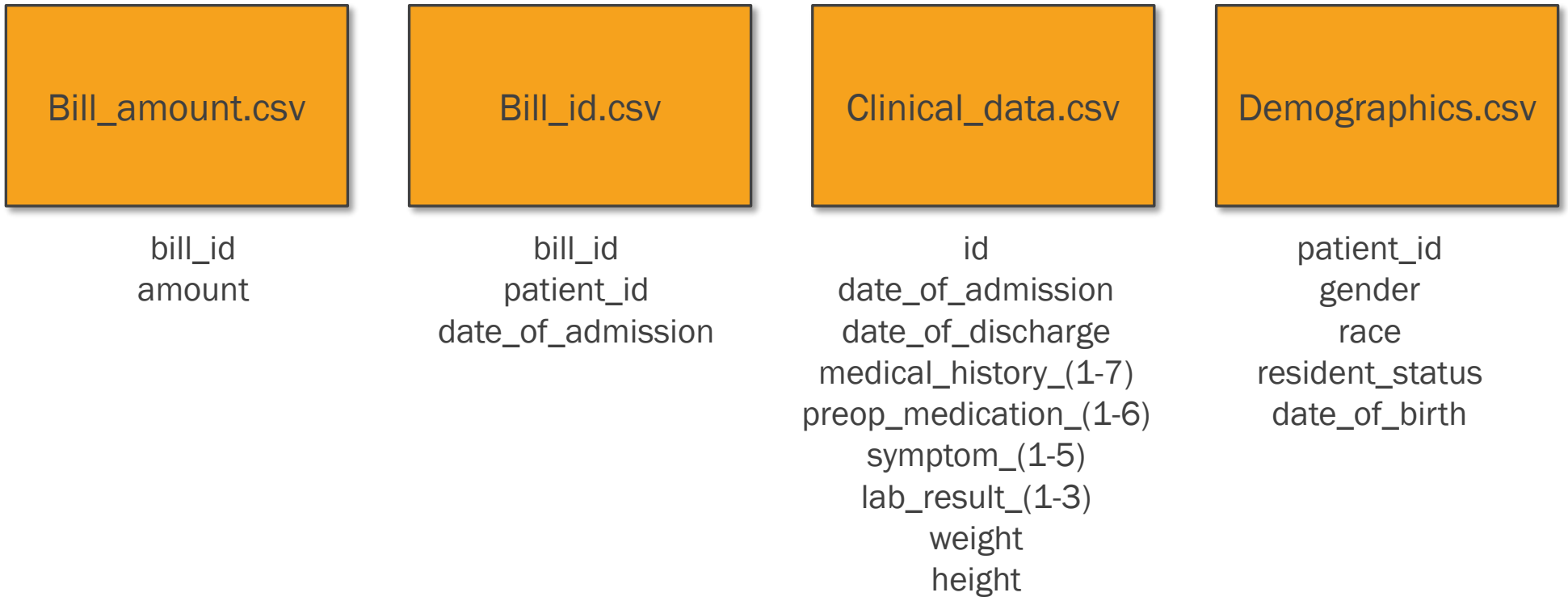
Objective: To determine the **key drivers of cost of care** for patients hospitalized

Proposed Approach

Exploratory Data Analysis	Dimensionality Reduction	Model Development	Hyperparameter Tuning	Conclusions & Recommendations
Perform EDA on the dataset to understand the underlying distributions and visualize data	Perform PCA and tSNE on dataset to find clusters in the data and understand which features contribute most to the variance of the cost of care (the variable **amount**)	Train machine learning models (e.g. linear regression model, decision tree, random forest etc) that predicts the numerical output variable **amount**	Tune hyperparameters to improve the model	• Conclude on the **key drivers** of the cost of care. • Provide recommendations on the deployment of a model for prediction and the way ahead for further data collection for model improvement

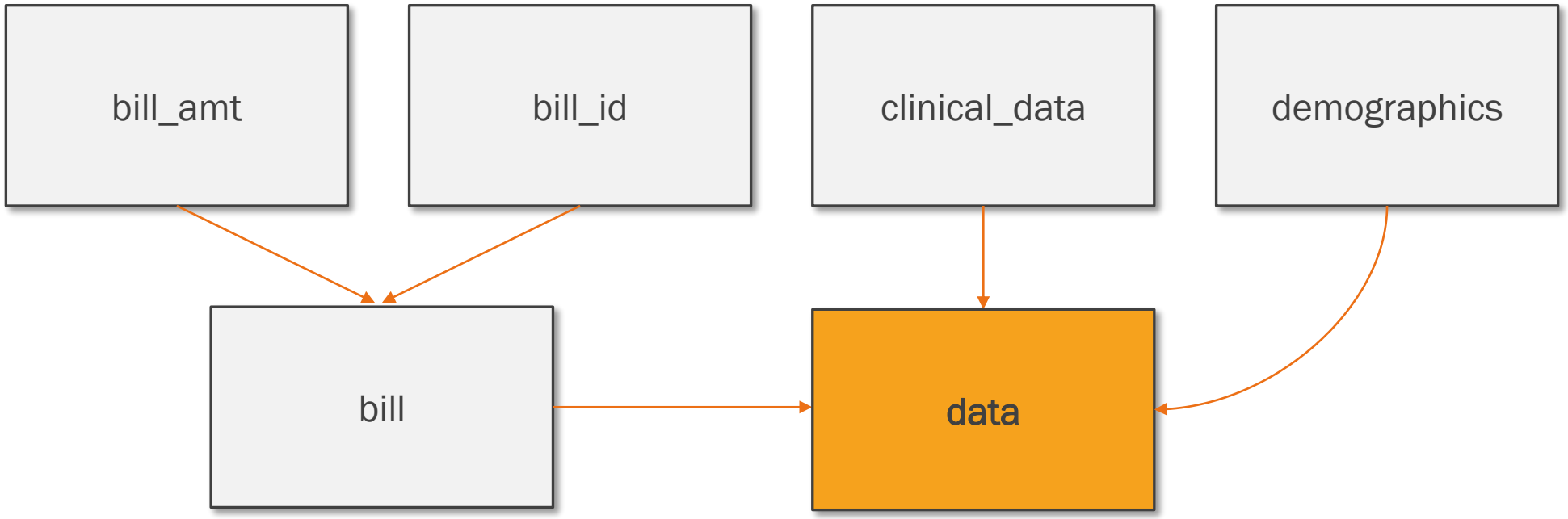
Exploratory Data Analysis
(Part 1 – Data Preparation)

- Data was taken from **four** csv files with its variables listed below



Exploratory Data Analysis
(Part 1 – Data Preparation)

- Dataframes from the four csvs were read and merged into a final dataframe named **data**



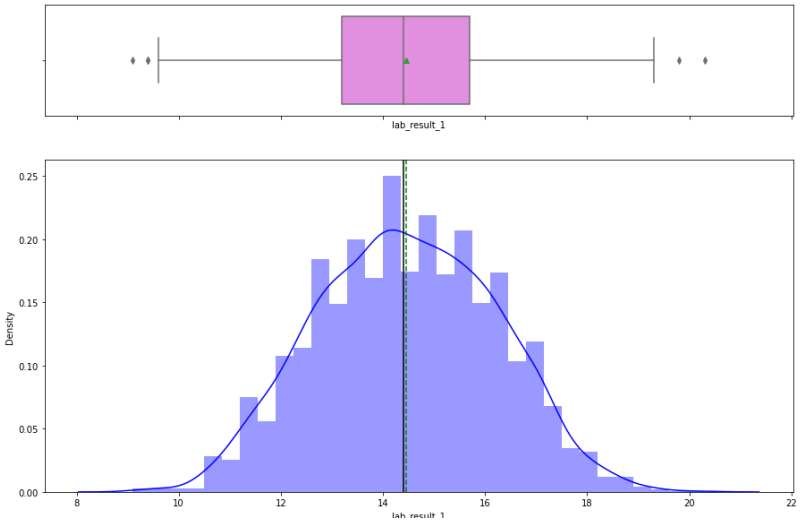
Exploratory Data Analysis
(Part 1 – Data Preparation)

- **Observation 1 (Mapping between bill and clinical_data):**
  - There is no one-to-one mapping between **bill** and **clinical_data**. For certain patients, there were multiple bill records for a particular date of admission while there was only one record in the clinical data for that date.
  - In order to merge the two dataframes, **bill** was grouped by 'patient_id' and 'date_of_admission', taking the sum of the total bill for the entire duration of hospital treatment.
- **Observation 2 (Missing values):**
  - The were missing values in 'medical_history_2' and 'medical_history_5' in **clinical_data.** The corresponding rows of the dataset were removed instead of imputing values.
  - After removal of missing values, there were a total of **2898** data entries in the dataframe **data**.
- **Observation 3 (Non-essential features):**
  - Once all the dataframes were merged, the features 'id', 'bill_id', and 'patient_id' were removed from **data** as they did not provide useful information

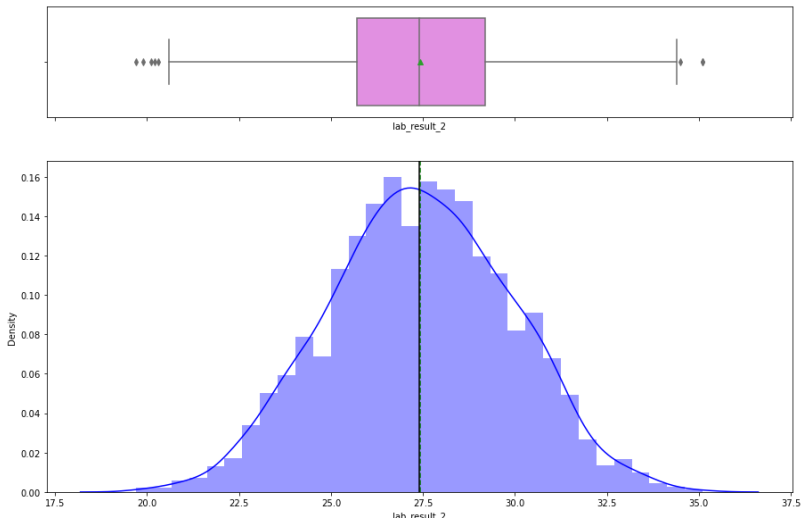
Exploratory Data Analysis
(Part 1 – Data Preparation)

- **Observation 4 (Non-consistent Data):**
  - There were non-consistent data in ‘gender’, ‘race’ and ‘resident_status’ as categories were entered in different representations (e.g. ‘Male’ and ‘m’, ‘Female’ and ‘f’, ‘Chinese’ and ‘chinese’, ‘Indian’ and ‘India’)
  - Data cleaning was performed to ensure consistency
- **Observation 5 (Date Features):**
  - There were three date features which could be engineered to provide better insights
  - Two features were created – ‘days_admitted’ and ‘age’
    - ‘days_admitted’ – difference between date of admission and discharge
    - ‘age’ – difference between date of admission and date of birth
  - With the two new features created, the original date features were dropped

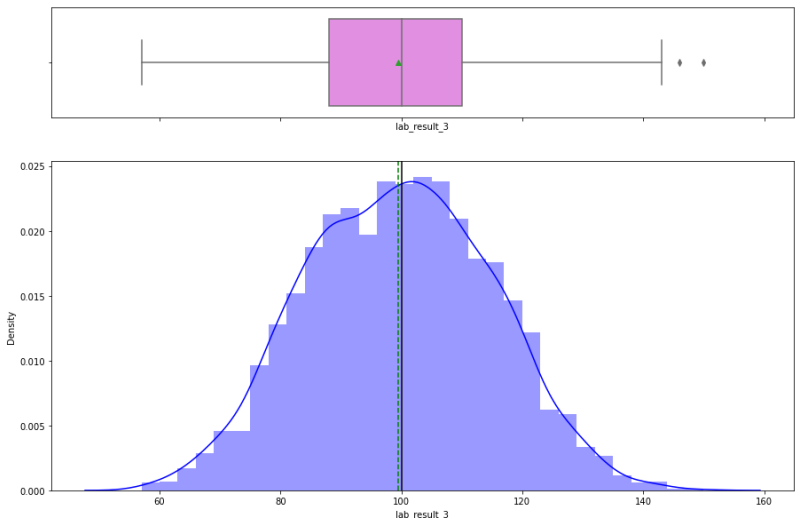
Exploratory Data Analysis
(Part 2 – Univariate Analysis [Numerical Variables])



lab_result_1



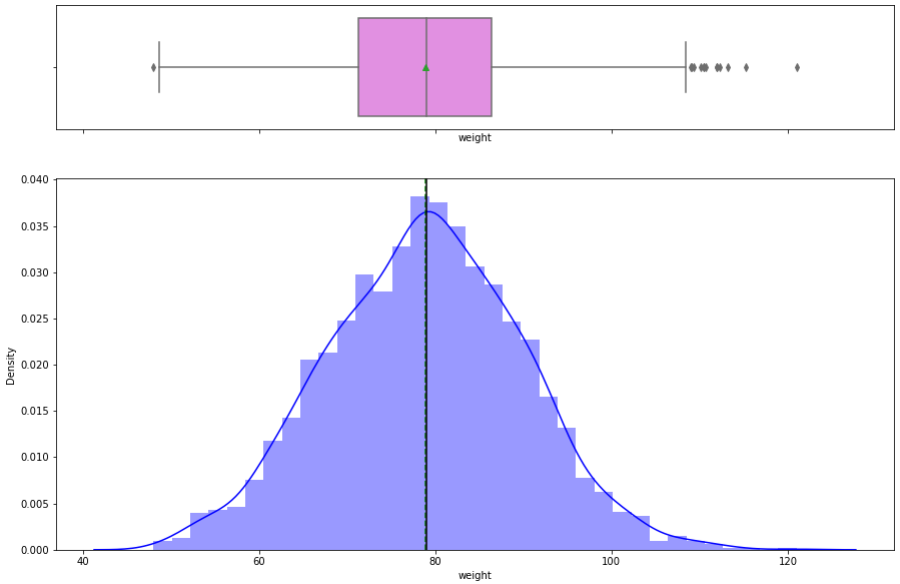
lab_result_2



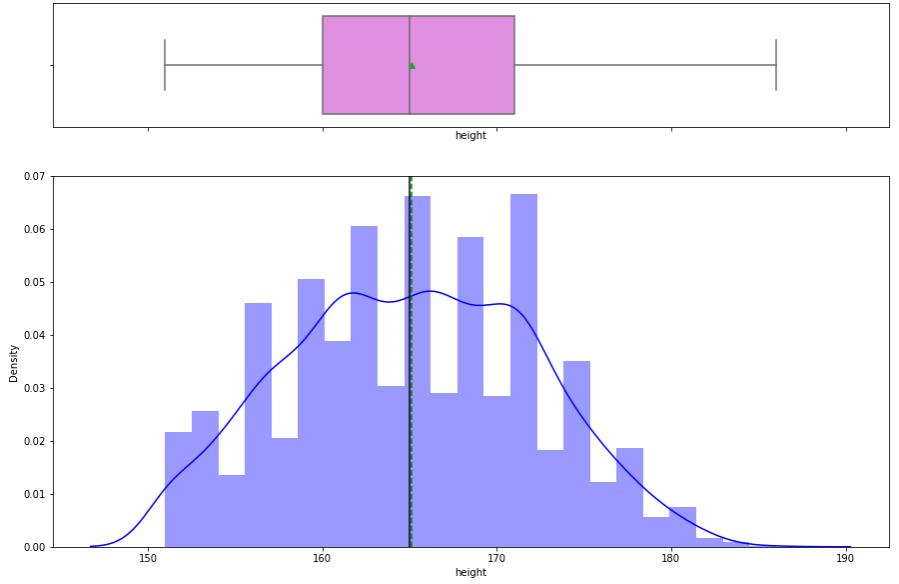
lab_result_3

Observation: Lab results are normally distributed with few outliers

Exploratory Data Analysis
(Part 2 – Univariate Analysis [Numerical Variables])



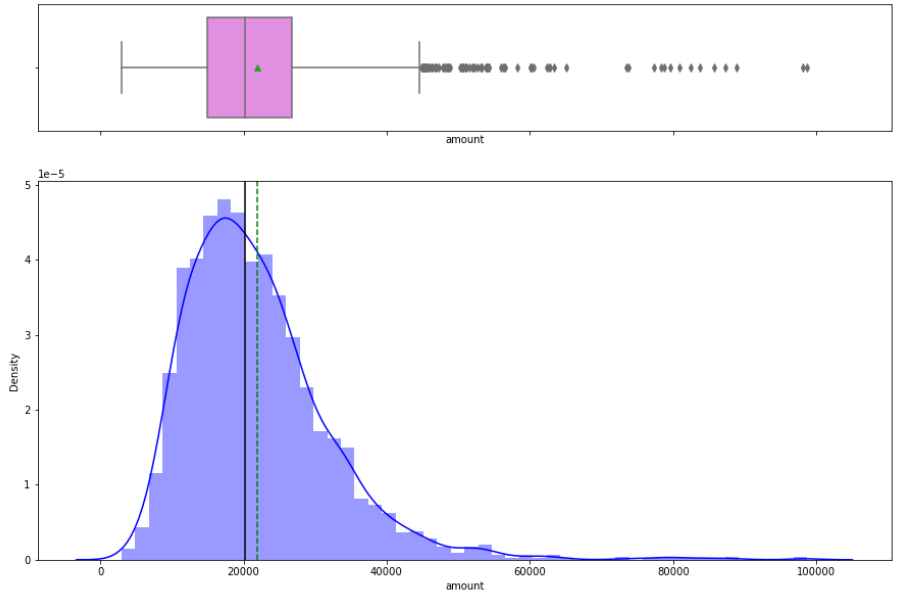
weight



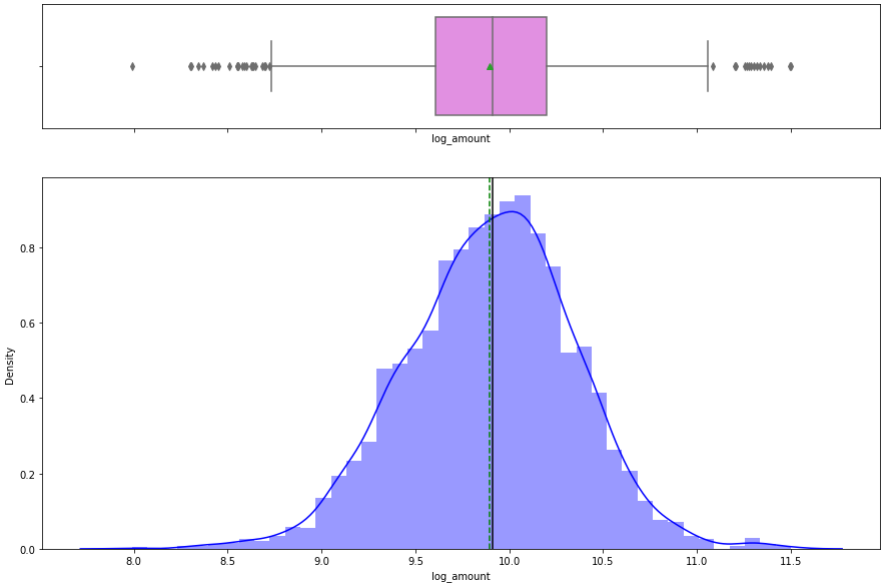
height

Observation: Weight and Height are normally distributed with few outliers

Exploratory Data Analysis
(Part 2 – Univariate Analysis [Numerical Variables])



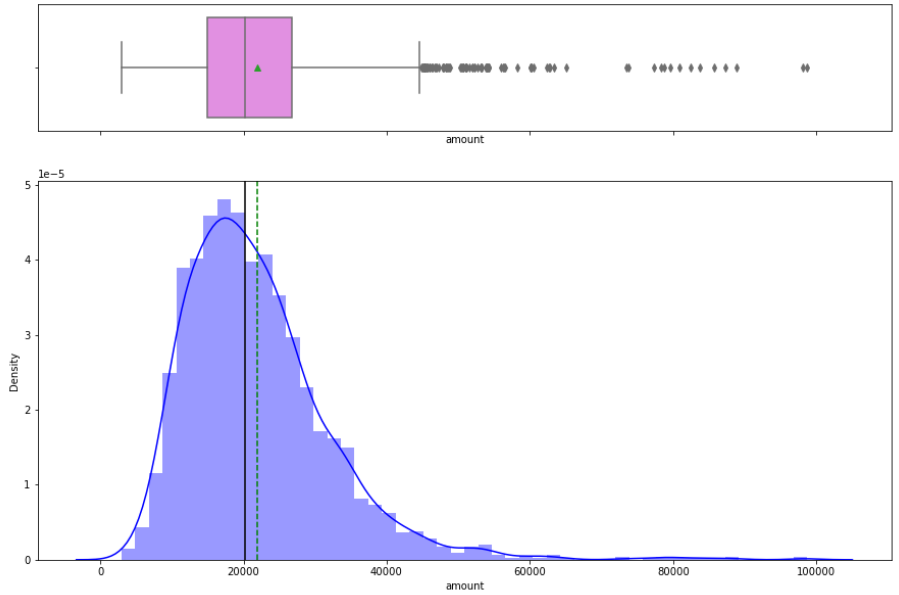
amount



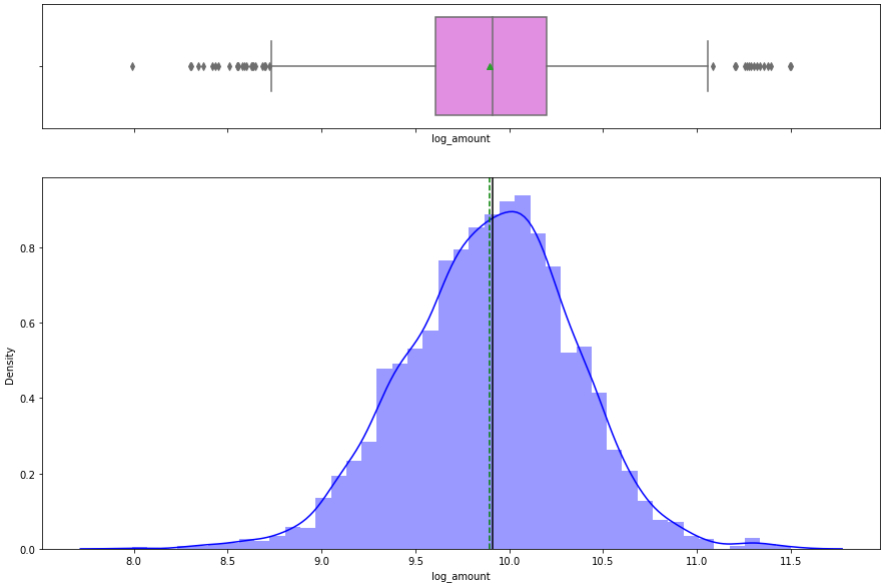
log_amount

Observation: Amount is right-skewed. After log transformation, the skew is reduced.

Exploratory Data Analysis
(Part 2 – Univariate Analysis [Numerical Variables])



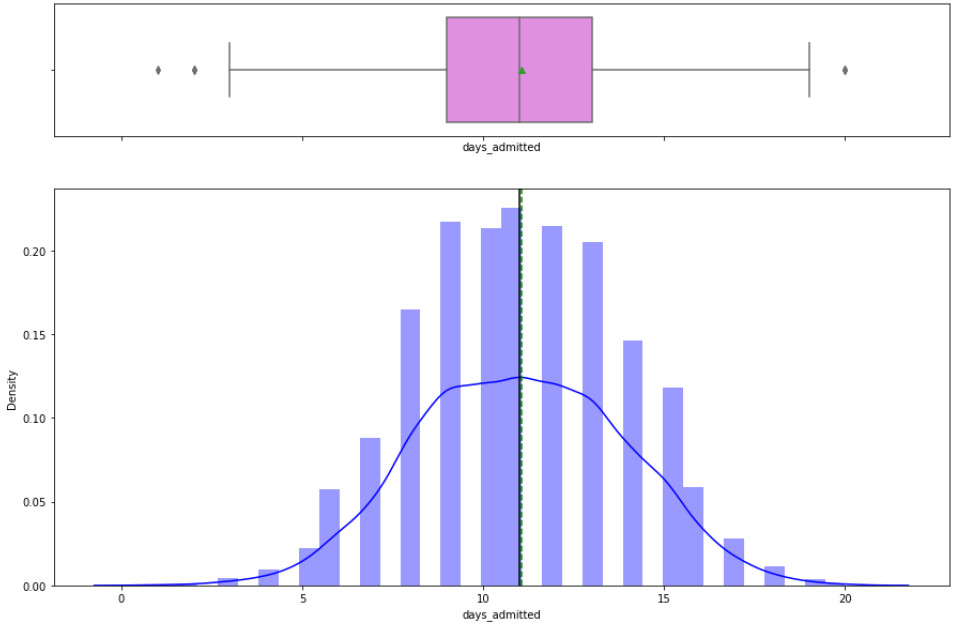
amount



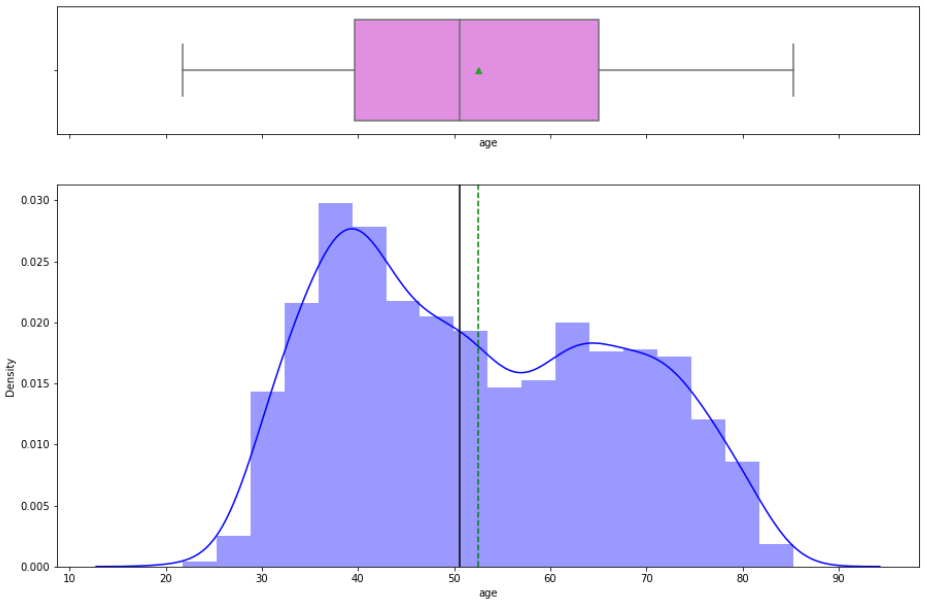
log_amount

Observation: Amount is right-skewed. After log transformation, the skew is reduced.

Exploratory Data Analysis
(Part 2 – Univariate Analysis [Numerical Variables])



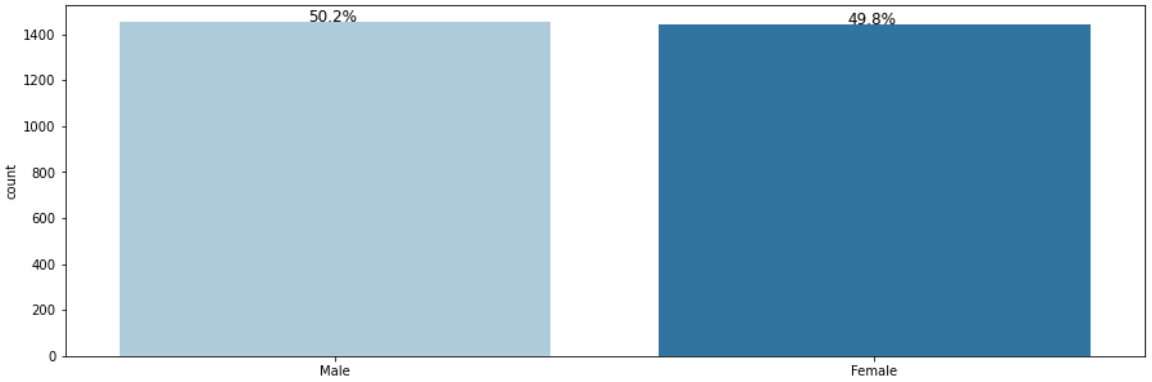
days_admitted



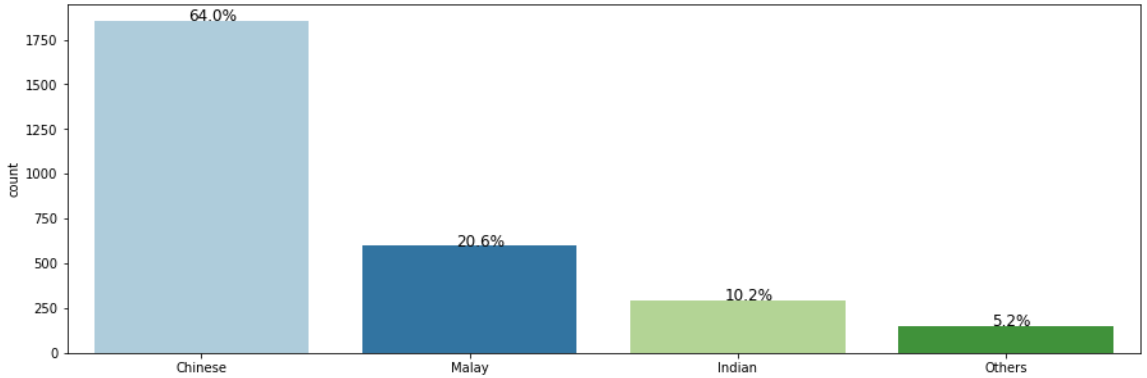
age

Observation: days_admitted is normally distributed while age is not.

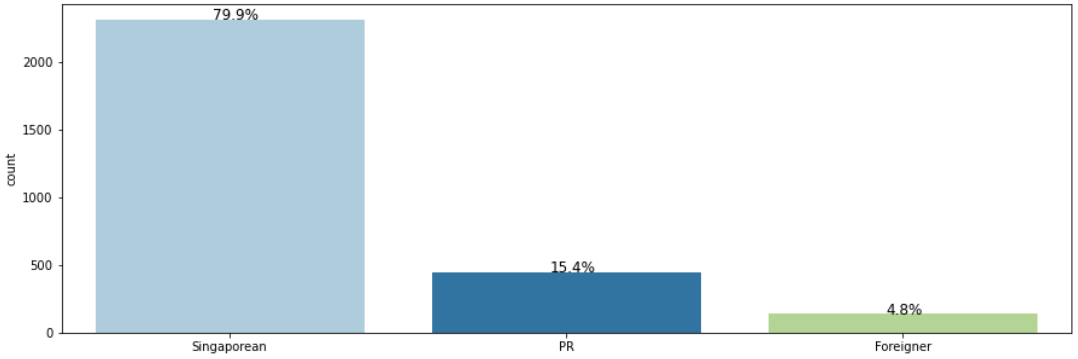
Exploratory Data Analysis
(Part 2 – Univariate Analysis [Categorical Variables])



gender



race

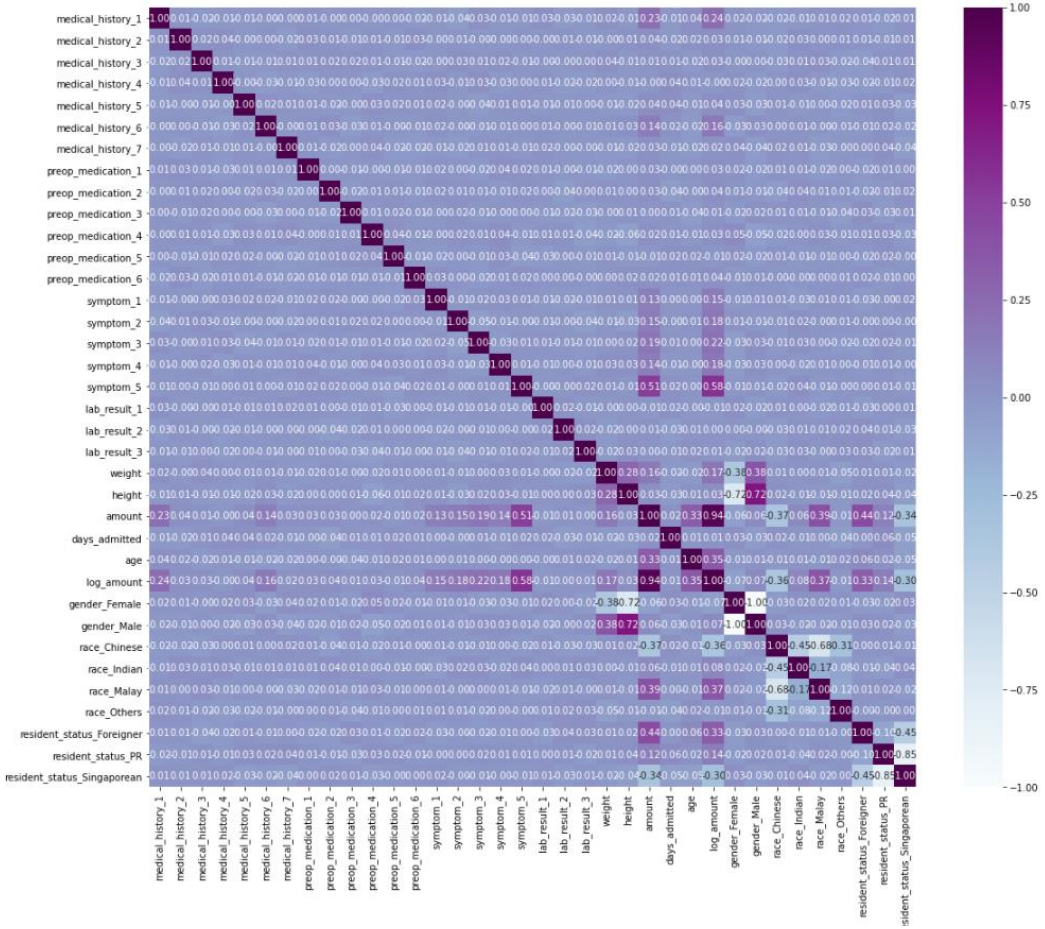


resident_status

Observation (of proportions):

1. Male and females are almost equal
2. Majority race is ‘Chinese’
3. Majority resident status is Singaporean

Exploratory Data Analysis
(Part 3 – Bivariate Analysis [Numerical Variables])

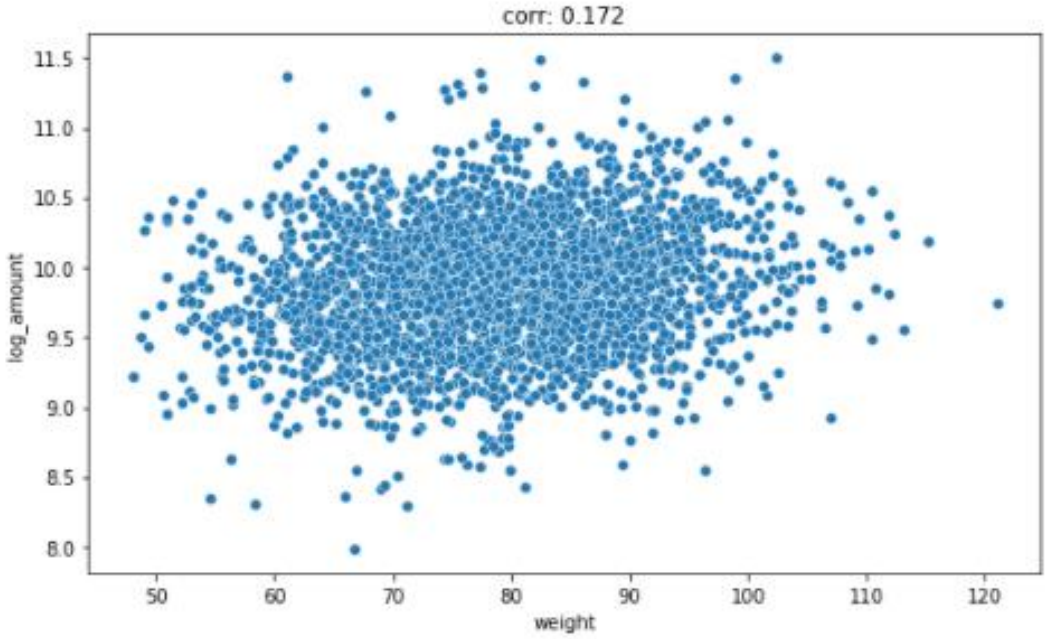


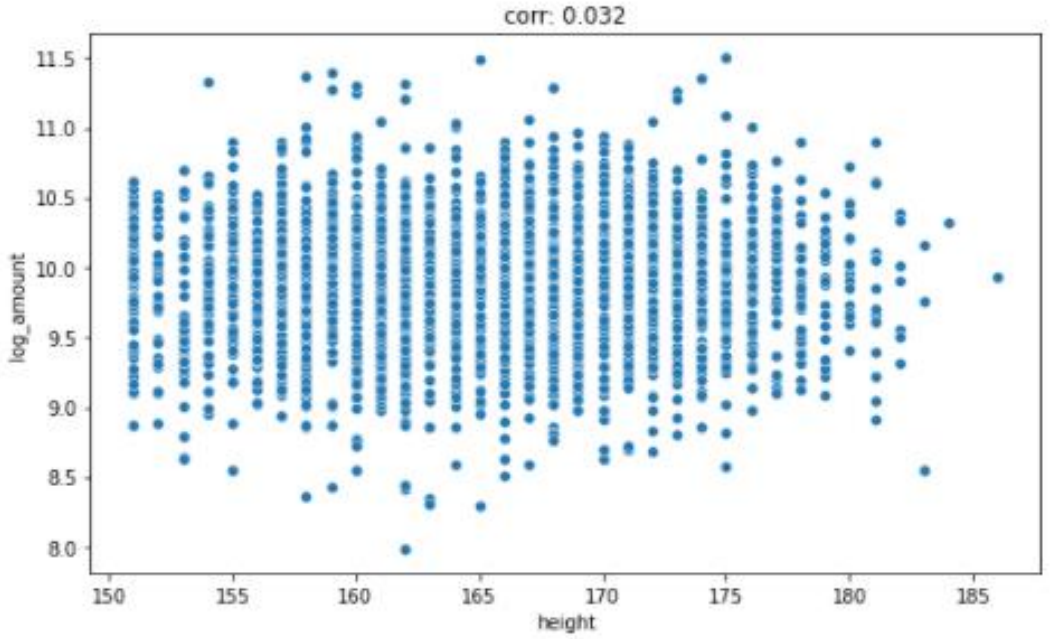
Postive correlation (with Log_Amount)	Negative correlation (with Log_Amount)
Amount (0.94)*	Race_Chinese (-0.36)
Symptom_5 (0.51)	Resident_status_Singaporean (-0.30)
Race_Malay (0.37)	
Age (0.35)	
Resident_status_Foreigner (0.33)	
Medical_history_1 (0.24)	
Symptom_3 (0.22)	
Symptom_4 (0.18)	
Symptom_2 (0.18)	
Weight (0.17)	
Symptom_1 (0.15)	

** Expected high correlation due to log transformation*

Observation: The top five variables with high correlation with log_amount are ‘symptom_5’, ‘race_Malay’, ‘race_Chinese’, ‘age’, ‘resident_status_Foreigner’

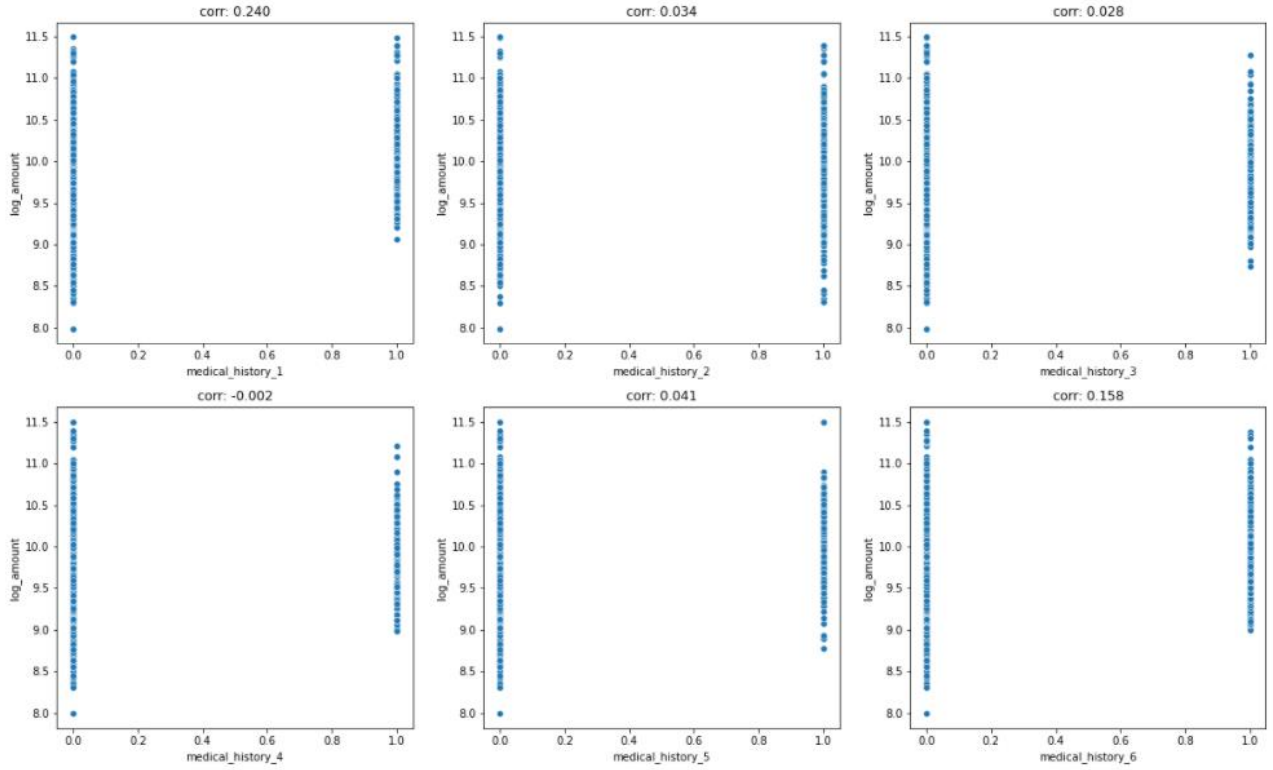
Exploratory Data Analysis
(Part 3 – Bivariate Analysis [Numerical Variables])





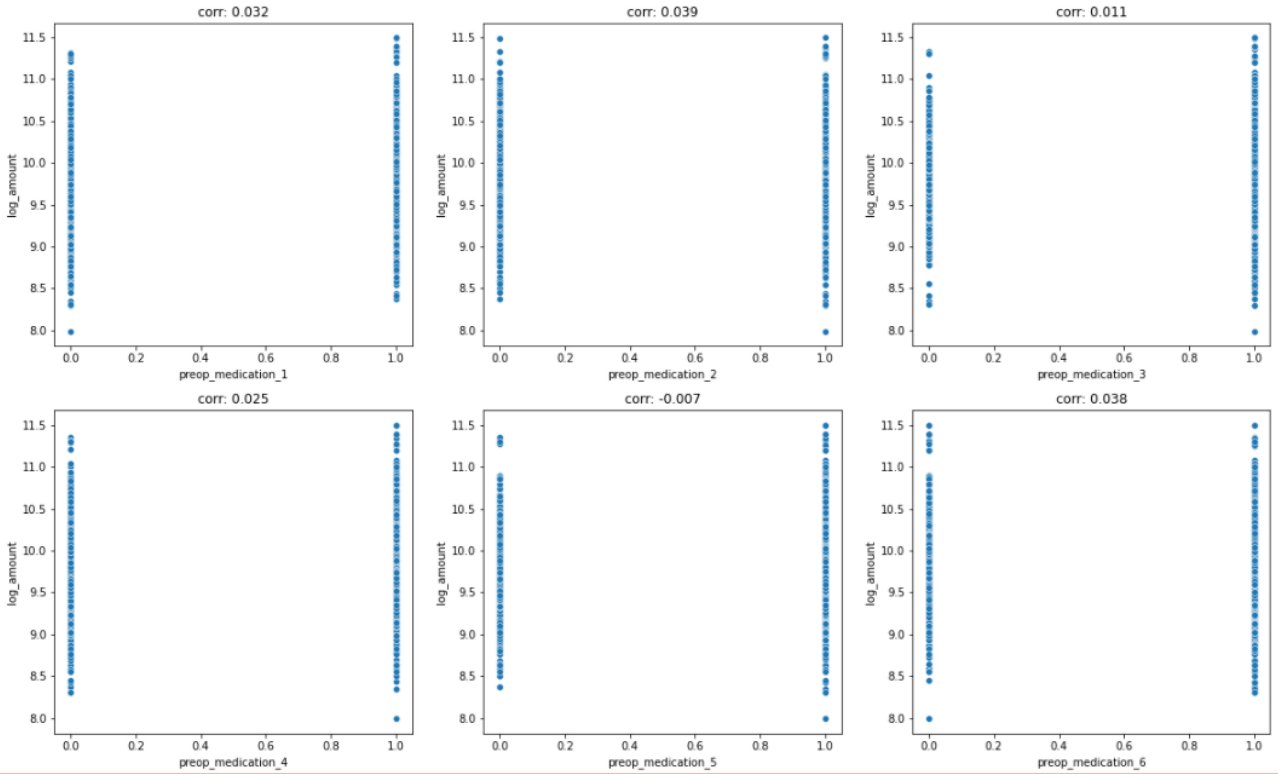
Observation: weight has low-moderate correlation (0.172) with log_amount, height has low correlation (0.032) with log_amount

Exploratory Data Analysis
(Part 3 – Bivariate Analysis [Binary Variables])



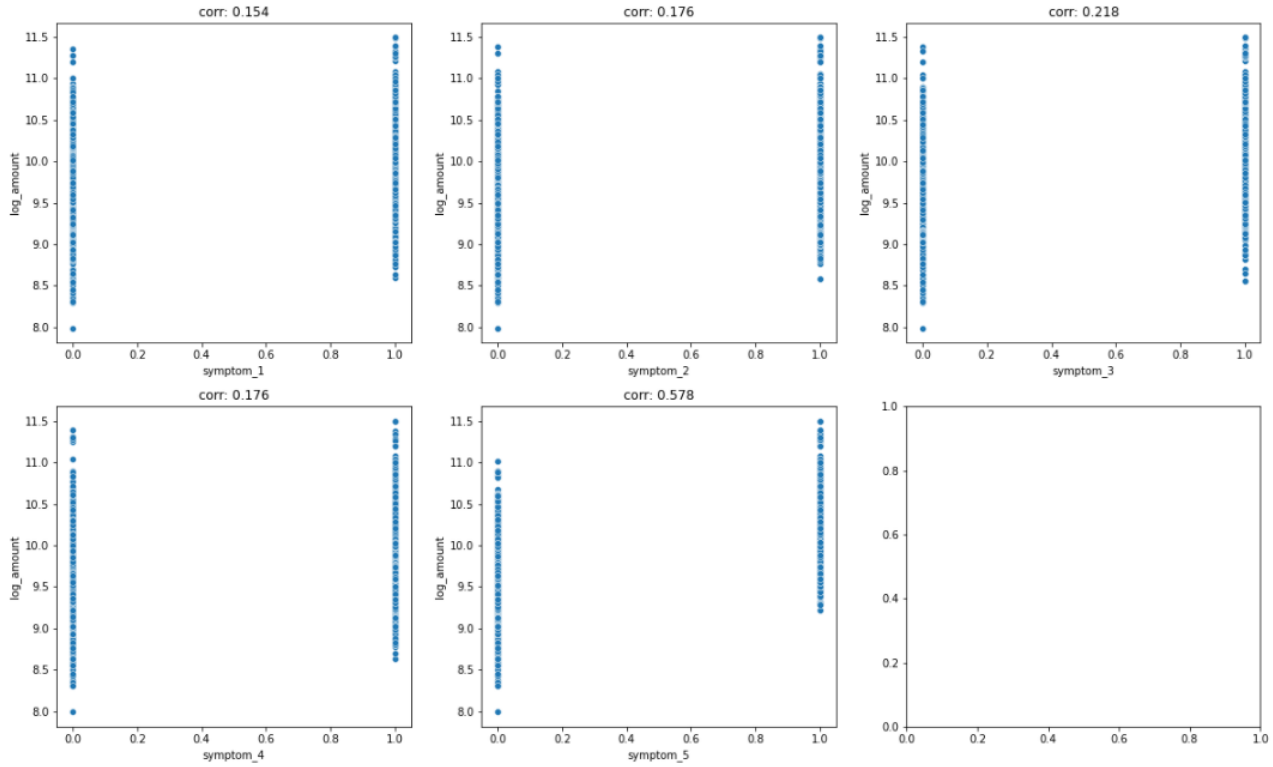
Observation: Order of correlation (descending): medical_history_1, medical_history_6, medical_history_5, medical_history_2, medical_history_3, medical_history_4

Exploratory Data Analysis
(Part 3 – Bivariate Analysis [Binary Variables])



Observation: Order of correlation (descending): preop_medication_2, preop_medication_6, preop_medication_1, preop_medication_4, preop_medication_3, preop_medication_5

Exploratory Data Analysis
(Part 3 – Bivariate Analysis [Binary Variables])



Observation: Order of correlation (descending): symptom_5, symptom_3, symptom_2, symptom_4, symptom_1

Exploratory Data Analysis
Key Insights

- **Insight 1 (Distributions of Numerical Variables):**
  - Numerical variables are observed to be normally distributed except for ‘age’ and ‘amount’
  - ‘Amount’ was log transformed and ‘log_amount’ will be used as the output variable
- **Insight 2 (Proportions of Race and Resident Status):**
  - Majority race is ‘Chinese’ (64.0%)
  - Majority resident status is Singaporean (79.9%)
- **Insight 3 (Medical history, Preop Medication, and Symptoms):**
  - Symptoms are generally low to moderately correlated with log_amount (0.154 to 0.578)
  - ‘Medical_history_1’ (0.240) and ‘Medical_history_6’ (0.158) have low correlation with log_amount

Dimensionality Reduction
(Part 1 – PCA)

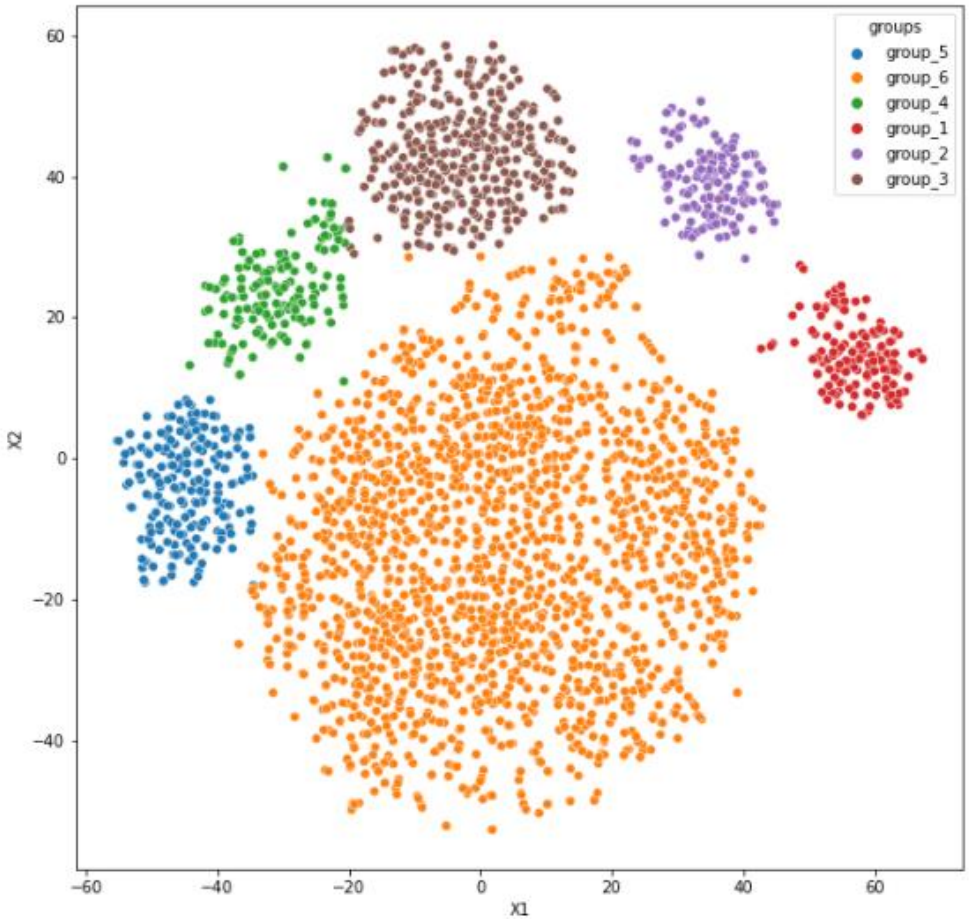
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
medical_history_1	0.130000	-0.050000	-0.120000	0.030000	-0.160000	0.250000	0.290000	-0.060000	0.330000	-0.120000	-0.120000	-0.010000	0.220000	0.240000	-0.150000	0.230000	0.110000
medical_history_2	0.020000	-0.010000	-0.050000	0.100000	0.010000	-0.070000	-0.220000	-0.160000	0.150000	0.170000	0.420000	0.290000	0.060000	0.290000	-0.240000	0.200000	0.160000
medical_history_3	0.020000	0.000000	-0.010000	-0.130000	0.210000	-0.140000	0.070000	-0.200000	-0.100000	0.160000	0.270000	-0.120000	-0.190000	0.250000	0.510000	0.100000	0.150000
medical_history_4	-0.010000	-0.020000	-0.030000	0.070000	0.300000	0.330000	-0.290000	0.000000	0.210000	0.060000	0.230000	0.040000	-0.130000	0.050000	0.140000	-0.090000	0.060000
medical_history_5	0.030000	-0.050000	0.040000	-0.100000	0.080000	-0.140000	-0.120000	0.430000	0.020000	-0.140000	-0.150000	0.160000	-0.070000	0.080000	-0.040000	0.370000	0.070000
medical_history_6	0.090000	0.020000	-0.020000	0.040000	0.030000	-0.160000	0.140000	0.280000	-0.030000	0.340000	-0.300000	0.360000	0.170000	-0.040000	0.030000	-0.210000	0.030000
medical_history_7	0.010000	-0.050000	0.080000	0.080000	0.080000	-0.060000	0.430000	0.100000	-0.020000	-0.160000	0.160000	-0.120000	-0.110000	-0.110000	0.140000	0.310000	-0.300000
preop_medication_1	0.020000	-0.020000	-0.010000	0.100000	-0.060000	-0.290000	0.110000	0.170000	0.080000	0.050000	0.190000	-0.050000	0.110000	0.540000	-0.150000	-0.300000	-0.150000
preop_medication_2	0.020000	-0.010000	-0.060000	0.100000	0.060000	-0.100000	0.070000	-0.150000	0.120000	0.570000	-0.200000	0.030000	-0.210000	-0.050000	-0.040000	0.180000	0.180000
preop_medication_3	-0.000000	0.010000	-0.030000	-0.110000	0.290000	-0.060000	-0.020000	-0.120000	0.040000	-0.290000	-0.040000	-0.390000	0.150000	0.010000	-0.230000	-0.050000	0.320000
preop_medication_4	0.020000	-0.090000	0.040000	-0.180000	0.130000	-0.290000	0.220000	-0.070000	0.230000	-0.110000	0.070000	0.270000	-0.110000	-0.270000	0.000000	0.180000	-0.040000
preop_medication_5	-0.010000	-0.010000	0.030000	-0.130000	0.060000	-0.070000	-0.150000	-0.160000	0.400000	-0.230000	-0.330000	0.240000	-0.290000	0.030000	0.060000	-0.150000	0.210000
preop_medication_6	0.020000	0.010000	-0.030000	0.000000	-0.060000	-0.020000	-0.380000	0.220000	0.010000	-0.030000	0.270000	0.040000	0.040000	-0.290000	-0.150000	0.360000	0.000000
symptom_1	0.080000	-0.010000	-0.080000	-0.050000	-0.030000	-0.010000	-0.130000	0.350000	0.240000	0.260000	0.190000	-0.200000	-0.200000	-0.210000	-0.010000	-0.250000	-0.220000
symptom_2	0.080000	-0.060000	-0.050000	-0.020000	0.110000	-0.380000	-0.260000	-0.310000	-0.040000	0.060000	-0.120000	-0.130000	0.350000	-0.160000	0.270000	0.000000	-0.200000
symptom_3	0.120000	0.000000	-0.060000	0.110000	0.170000	0.380000	0.260000	-0.090000	0.120000	0.030000	0.120000	0.040000	-0.190000	-0.230000	-0.050000	-0.220000	-0.060000
symptom_4	0.090000	0.010000	-0.040000	-0.140000	-0.070000	-0.370000	0.100000	0.140000	0.190000	-0.210000	0.170000	-0.050000	-0.310000	0.120000	0.030000	-0.150000	0.020000
symptom_5	0.310000	-0.080000	-0.200000	0.220000	0.080000	-0.080000	-0.050000	0.170000	-0.230000	-0.060000	-0.010000	-0.200000	-0.020000	-0.090000	-0.060000	-0.030000	0.330000
lab_result_1	-0.010000	-0.020000	0.000000	0.090000	0.010000	0.090000	0.260000	0.150000	0.110000	0.020000	0.260000	0.210000	0.430000	-0.070000	0.340000	0.040000	0.260000
lab_result_2	0.000000	-0.010000	0.030000	0.040000	-0.180000	-0.100000	0.000000	-0.150000	-0.300000	-0.260000	0.260000	0.420000	-0.030000	-0.210000	-0.000000	-0.330000	0.210000
lab_result_3	-0.000000	0.030000	-0.030000	0.190000	-0.220000	0.120000	-0.020000	0.080000	-0.370000	-0.060000	-0.140000	0.050000	-0.400000	0.230000	0.200000	0.170000	0.100000
weight	0.130000	0.410000	-0.060000	-0.080000	0.110000	-0.040000	0.050000	-0.030000	0.070000	0.000000	-0.010000	0.020000	0.020000	-0.000000	0.050000	0.100000	0.040000
height	0.080000	0.610000	0.010000	0.020000	-0.020000	-0.000000	-0.020000	0.030000	0.020000	-0.030000	0.020000	0.030000	0.020000	0.010000	0.010000	0.010000	-0.020000
amount	0.540000	-0.090000	-0.140000	-0.010000	-0.020000	0.010000	0.010000	-0.020000	-0.010000	-0.020000	-0.010000	0.000000	0.010000	-0.000000	0.000000	0.010000	-0.000000
days_admitted	0.020000	-0.060000	0.090000	-0.060000	0.280000	0.170000	-0.170000	0.350000	0.060000	-0.150000	-0.130000	0.080000	0.140000	0.140000	0.380000	-0.100000	0.040000
age	0.200000	-0.050000	-0.080000	0.160000	-0.240000	0.100000	-0.230000	-0.200000	0.170000	-0.190000	-0.060000	0.140000	0.030000	0.120000	0.210000	0.060000	-0.440000
log_amount	0.550000	-0.080000	-0.160000	0.020000	0.000000	-0.020000	0.000000	-0.000000	0.000000	-0.020000	-0.010000	-0.010000	-0.000000	-0.020000	0.020000	-0.010000	0.010000
gender_Male	0.100000	0.630000	-0.010000	-0.000000	-0.010000	-0.020000	-0.010000	0.000000	-0.000000	-0.030000	-0.010000	0.000000	0.020000	-0.020000	0.030000	-0.000000	-0.020000
race_Indian	0.020000	-0.010000	-0.090000	0.530000	0.410000	-0.140000	0.020000	-0.090000	-0.080000	-0.090000	-0.050000	0.160000	-0.080000	0.010000	-0.110000	0.050000	-0.130000
race_Malay	0.210000	-0.080000	-0.100000	-0.600000	-0.080000	0.160000	0.020000	-0.090000	-0.230000	0.130000	0.060000	0.110000	-0.010000	0.070000	-0.050000	0.040000	-0.020000
race_Others	-0.020000	-0.020000	0.020000	0.240000	-0.500000	-0.100000	-0.030000	-0.010000	0.260000	0.030000	0.010000	-0.230000	-0.010000	-0.170000	0.260000	0.000000	0.310000
resident_status_PR	0.200000	-0.010000	0.660000	0.050000	0.020000	0.020000	-0.010000	-0.030000	0.010000	0.070000	0.020000	-0.040000	-0.010000	0.020000	-0.050000	-0.010000	0.030000
resident_status_Singaporean	-0.280000	0.020000	-0.630000	-0.050000	0.000000	-0.020000	0.030000	0.050000	0.010000	-0.030000	-0.010000	0.030000	0.000000	-0.020000	0.060000	-0.010000	-0.020000

- **Methodology**
  - PCA was performed on the dataset to reduce dimensionality and determine key features through principal components
  - In order to capture 70% of the variance, it was found that 19 PCs were required

Dimensionality Reduction
(Part 1 – PCA)

- **Observation 1 (Principal Component with Amount and Log_Amount):**
  - The first Principal Component is positively related with ‘symptom_5’, ‘amount’ and ‘log_amount’ and negatively correlated with ‘resident_status_Singaporean’
  - This captures the observation that those with ‘symptom_5’ are related to higher amounts
  - It also captures that those who are Singaporean are related to lower amounts
  - Conclusion from PCA: symptom_5 and resident_status_Singaporean are important features to predict amount and log_amount
- **Observation 2 (No other PC with Amount and Log_Amount):**
  - There are no other PCs with ‘Amount’ and ‘Log_Amount’

Dimensionality Reduction
(Part 2 – tSNE)



- **Methodology**
  - tSNE was performed on the dataset to reduce dimensionality to 2 components and find clusters in the datset
  - Multiple values of the parameter ‘perplexity’ were used and it was observed that a value of 35 produced good clusters with a total of 6 groupings

Dimensionality Reduction
(Part 2 – tSNE)

- **Observation 1 (Characteristics of Groups):**
  - It was found that the groups could be differentiated by the resident status, race, medical history or symptoms
    - Group 1: No medical_history_6
    - Group 2: medical_history_4, no medical_history_6
    - Group 3: symptom_4
    - Group 4: medical_history_5
    - Group 5: No medical_history_6
    - Group 6: Largest variance in weight, height, amount and days_admitted
  - While Group 6 had the largest variance in amount, the group was not immediately differentiable by features

Dimensionality Reduction
Key Insights

- **Insight 1 (Symptom_5 and Resident_Status_Singaporean)**
  - The largest variance of the dataset, explained by PC1 in PCA, is positively related with ‘symptom_5’, ‘amount’ and ‘log_amount’ and negatively correlated with ‘resident_status_Singaporean’
  - ‘symptom_5’ and ‘resident_status_Singaporean’ are important features to predict ‘amount’ and ‘log_amount’

Model Development
Linear Regression Model

OLS Regression Results

Dep. Variable:	log_amount	R-squared:	0.975
Model:	OLS	Adj. R-squared:	0.975
Method:	Least Squares	F-statistic:	3249.
Date:	Mon, 23 Aug 2021	Prob (F-statistic):	0.00
Time:	11:02:17	Log-Likelihood:	2472.8
No. Observations:	2028	AIC:	-4896.
Df Residuals:	2003	BIC:	-4755.
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.0430	0.013	716.085	0.000	9.018	9.068
medical_history_1	0.2712	0.004	65.146	0.000	0.263	0.279
medical_history_2	0.0184	0.003	5.340	0.000	0.012	0.025
medical_history_3	0.0262	0.005	5.592	0.000	0.017	0.035
medical_history_5	0.0556	0.006	8.708	0.000	0.043	0.068
medical_history_6	0.1654	0.004	45.024	0.000	0.158	0.173
medical_history_7	0.0433	0.004	11.730	0.000	0.036	0.050
preop_medication_1	0.0230	0.003	7.169	0.000	0.017	0.029
preop_medication_2	0.0163	0.003	4.984	0.000	0.010	0.023
preop_medication_3	0.0311	0.004	7.402	0.000	0.023	0.039
preop_medication_5	0.0159	0.004	3.859	0.000	0.008	0.024
preop_medication_6	0.0247	0.004	6.674	0.000	0.017	0.032
symptom_1	0.1321	0.003	39.920	0.000	0.126	0.139
symptom_2	0.1814	0.003	53.479	0.000	0.175	0.188
symptom_3	0.1968	0.003	60.776	0.000	0.190	0.203
symptom_4	0.1705	0.004	47.191	0.000	0.163	0.178
symptom_5	0.5056	0.003	157.175	0.000	0.499	0.512
weight	0.4060	0.010	41.338	0.000	0.387	0.425
height	-0.0406	0.008	-4.900	0.000	-0.057	-0.024
age	0.6131	0.007	89.553	0.000	0.600	0.626
race_Indian	0.1917	0.005	35.622	0.000	0.181	0.202
race_Malay	0.4435	0.004	109.709	0.000	0.436	0.451
race_Others	0.0983	0.008	12.417	0.000	0.083	0.114
resident_status_PR	-0.5095	0.009	-59.649	0.000	-0.526	-0.493
resident_status_Singaporean	-0.6923	0.008	-89.084	0.000	-0.708	-0.677

Omnibus:	961.794	Durbin-Watson:	2.010
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7823.791
Skew:	-2.062	Prob(JB):	0.00
Kurtosis:	11.694	Cond. No.	26.0

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- **Methodology**
  - The dataset was split in a ratio of 70:30 for training and testing
  - The Linear Regression model was trained on the training data
  - A hypothesis test was conducted for each parameter β_i

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

  - Null hypotheses were rejected if p-values were<0.05. Features with p-values >= 0.05 were removed.
  - A linear regression model was re-trained using the remaining features and non-significant features were iteratively removed.

Model Development

Linear Regression Model

OLS Regression Results

Dep. Variable:	log_amount	R-squared:	0.975
Model:	OLS	Adj. R-squared:	0.975
Method:	Least Squares	F-statistic:	3249.
Date:	Mon, 23 Aug 2021	Prob (F-statistic):	0.00
Time:	11:02:17	Log-Likelihood:	2472.8
No. Observations:	2028	AIC:	-4896.
Df Residuals:	2003	BIC:	-4755.
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.0430	0.013	716.085	0.000	9.018	9.068
medical_history_1	0.2712	0.004	65.146	0.000	0.263	0.279
medical_history_2	0.0184	0.003	5.340	0.000	0.012	0.025
medical_history_3	0.0262	0.005	5.592	0.000	0.017	0.035
medical_history_5	0.0556	0.006	8.708	0.000	0.043	0.068
medical_history_6	0.1654	0.004	45.024	0.000	0.158	0.173
medical_history_7	0.0433	0.004	11.730	0.000	0.036	0.050
preop_medication_1	0.0230	0.003	7.169	0.000	0.017	0.029
preop_medication_2	0.0163	0.003	4.984	0.000	0.010	0.023
preop_medication_3	0.0311	0.004	7.402	0.000	0.023	0.039
preop_medication_5	0.0159	0.004	3.859	0.000	0.008	0.024
preop_medication_6	0.0247	0.004	6.674	0.000	0.017	0.032
symptom_1	0.1321	0.003	39.920	0.000	0.126	0.139
symptom_2	0.1814	0.003	53.479	0.000	0.175	0.188
symptom_3	0.1968	0.003	60.776	0.000	0.190	0.203
symptom_4	0.1705	0.004	47.191	0.000	0.163	0.178
symptom_5	0.5056	0.003	157.175	0.000	0.499	0.512
weight	0.4060	0.010	41.338	0.000	0.387	0.425
height	-0.0406	0.008	-4.900	0.000	-0.057	-0.024
age	0.6131	0.007	89.553	0.000	0.600	0.626
race_Indian	0.1917	0.005	35.622	0.000	0.181	0.202
race_Malay	0.4435	0.004	109.709	0.000	0.436	0.451
race_Others	0.0983	0.008	12.417	0.000	0.083	0.114
resident_status_PR	-0.5095	0.009	-59.649	0.000	-0.526	-0.493
resident_status_Singaporean	-0.6923	0.008	-89.084	0.000	-0.708	-0.677

Omnibus:	961.794	Durbin-Watson:	2.010
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7823.791
Skew:	-2.062	Prob(JB):	0.00
Kurtosis:	11.694	Cond. No.	26.0

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	R^2	MAE	RMSE
Linear Regression (on training data)	0.975	0.051	0.071
Linear Regression (on test data)	0.966	0.063	0.082

- **Findings**
  - The final linear regression model (ols_res_2) fits the data well with high R2 scores and low errors on both training and test data
  - There were a total of **24** features in the final model
  - Model was tested for assumptions and fulfilled:
    1. No multicollinearity between variables
    2. Zero mean of residuals
    3. Normality of residuals
    4. Linearity of residuals
    5. Homoskedasticity

Model Development

Linear Regression Model

OLS Regression Results

Dep. Variable:	log_amount	R-squared:	0.975
Model:	OLS	Adj. R-squared:	0.975
Method:	Least Squares	F-statistic:	3249.
Date:	Mon, 23 Aug 2021	Prob (F-statistic):	0.00
Time:	11:02:17	Log-Likelihood:	2472.8
No. Observations:	2028	AIC:	-4896.
Df Residuals:	2003	BIC:	-4755.
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.0430	0.013	716.085	0.000	9.018	9.068
medical_history_1	0.2712	0.004	65.146	0.000	0.263	0.279
medical_history_2	0.0184	0.003	5.340	0.000	0.012	0.025
medical_history_3	0.0262	0.005	5.592	0.000	0.017	0.035
medical_history_5	0.0556	0.006	8.708	0.000	0.043	0.068
medical_history_6	0.1654	0.004	45.024	0.000	0.158	0.173
medical_history_7	0.0433	0.004	11.730	0.000	0.036	0.050
preop_medication_1	0.0230	0.003	7.169	0.000	0.017	0.029
preop_medication_2	0.0163	0.003	4.984	0.000	0.010	0.023
preop_medication_3	0.0311	0.004	7.402	0.000	0.023	0.039
preop_medication_5	0.0159	0.004	3.859	0.000	0.008	0.024
preop_medication_6	0.0247	0.004	6.674	0.000	0.017	0.032
symptom_1	0.1321	0.003	39.920	0.000	0.126	0.139
symptom_2	0.1814	0.003	53.479	0.000	0.175	0.188
symptom_3	0.1968	0.003	60.776	0.000	0.190	0.203
symptom_4	0.1705	0.004	47.191	0.000	0.163	0.178
symptom_5	0.5056	0.003	157.175	0.000	0.499	0.512
weight	0.4060	0.010	41.338	0.000	0.387	0.425
height	-0.0406	0.008	-4.900	0.000	-0.057	-0.024
age	0.6131	0.007	89.553	0.000	0.600	0.626
race_Indian	0.1917	0.005	35.622	0.000	0.181	0.202
race_Malay	0.4435	0.004	109.709	0.000	0.436	0.451
race_Others	0.0983	0.008	12.417	0.000	0.083	0.114
resident_status_PR	-0.5095	0.009	-59.649	0.000	-0.526	-0.493
resident_status_Singaporean	-0.6923	0.008	-89.084	0.000	-0.708	-0.677

Omnibus:	961.794	Durbin-Watson:	2.010
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7823.791
Skew:	-2.062	Prob(JB):	0.00
Kurtosis:	11.694	Cond. No.	26.0

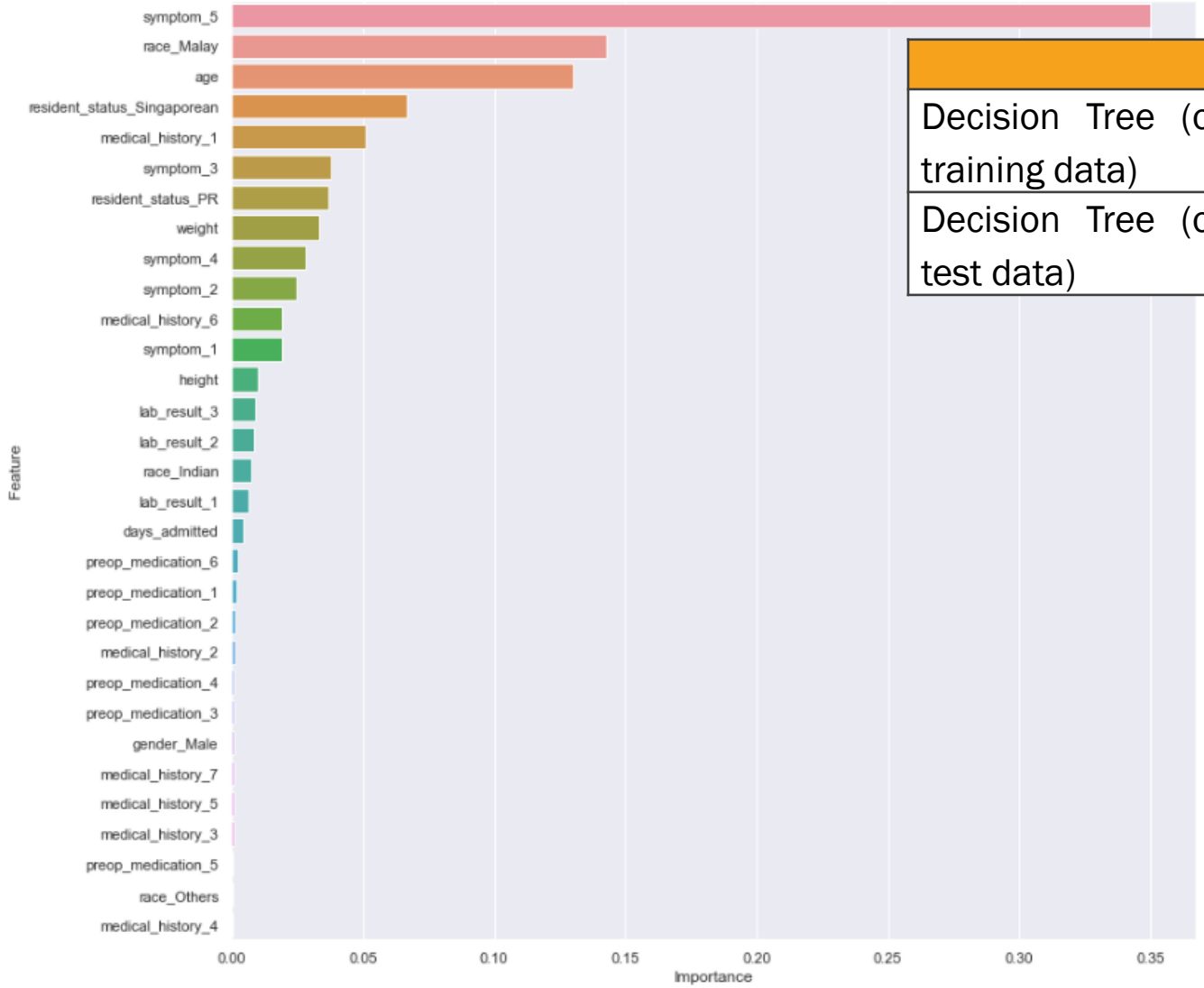
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	R^2	MAE	RMSE
Linear Regression (on training data)	0.975	0.051	0.071
Linear Regression (on test data)	0.966	0.063	0.082

- **Findings**
  - The most significant features of the model were (sorted in descending t stat values):
    1. Symptom_5
    2. Race_Malay
    3. Age
    4. Resident_status_Singaporean
    5. Medical_history_1

Model Development

Decision Tree Regressor



	R^2	MAE	RMSE
Decision Tree (on training data)	1.000	0.000	0.000
Decision Tree (on test data)	0.762	0.162	0.216

- **Methodology**
  - The dataset was split in a ratio of 70:30 for training and testing (using the same split in the linear regression model)
  - DecisionTreeRegressor was trained on the training data and tested on test data
- **Findings**
  - The model had a R2 of 1.000 and no errors on the training data.
  - The model did not perform well on test data with significantly lower R2 score of 0.762 and higher errors.
  - The model was assessed to be overfitting the data.

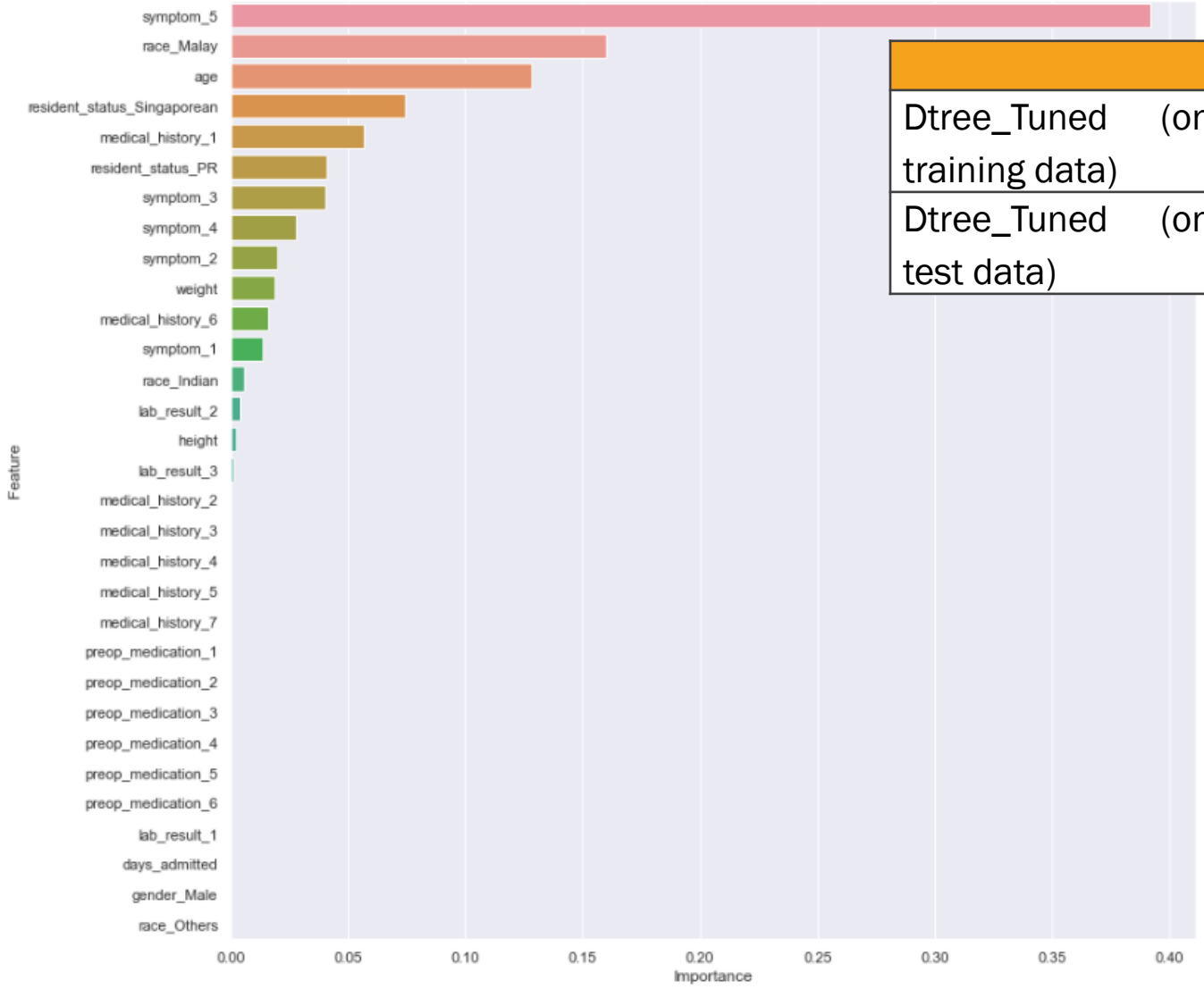
Model Development
Decision Tree Regressor (Tuned)

	R^2	MAE	RMSE
Dtree_Tuned (on training data)	0.893	0.117	0.148
Dtree_Tuned (on test data)	0.797	0.154	0.200

- **Methodology**
  - The Decision Tree was regularized through hyperparameter tuning by iterating through alphas from the cost complexity pruning path
  - The alpha corresponding to the highest test score was used for the tuned decision tree
- **Findings**
  - After regularization, the model performed better on the test data with R2 score of 0.797 and MAE of 0.154 and RMSE of 0.200
  - It also performed less well on training data

Model Development

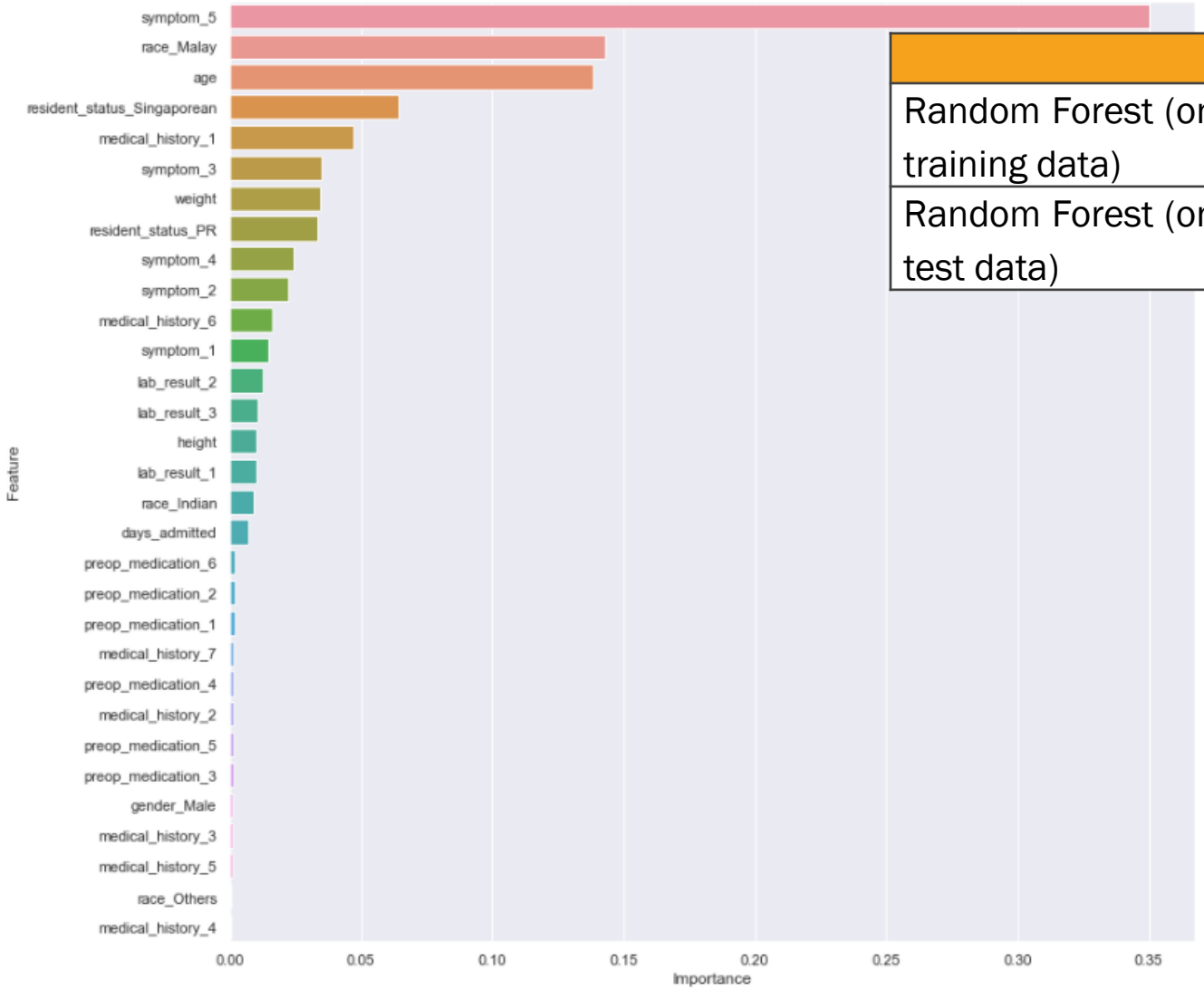
Decision Tree Regressor (Tuned)



	R^2	MAE	RMSE
Dtree_Tuned (on training data)	0.893	0.117	0.148
Dtree_Tuned (on test data)	0.797	0.154	0.200

- **Findings**
  - There were a total of 16 features in the tuned decision tree model
  - The top five features of Dtree_Tuned were:
    1. Symptom_5
    2. Race_Malay
    3. Age
    4. Resident_status_Singaporean
    5. Medical_history_1

Model Development
Random Forest Regressor

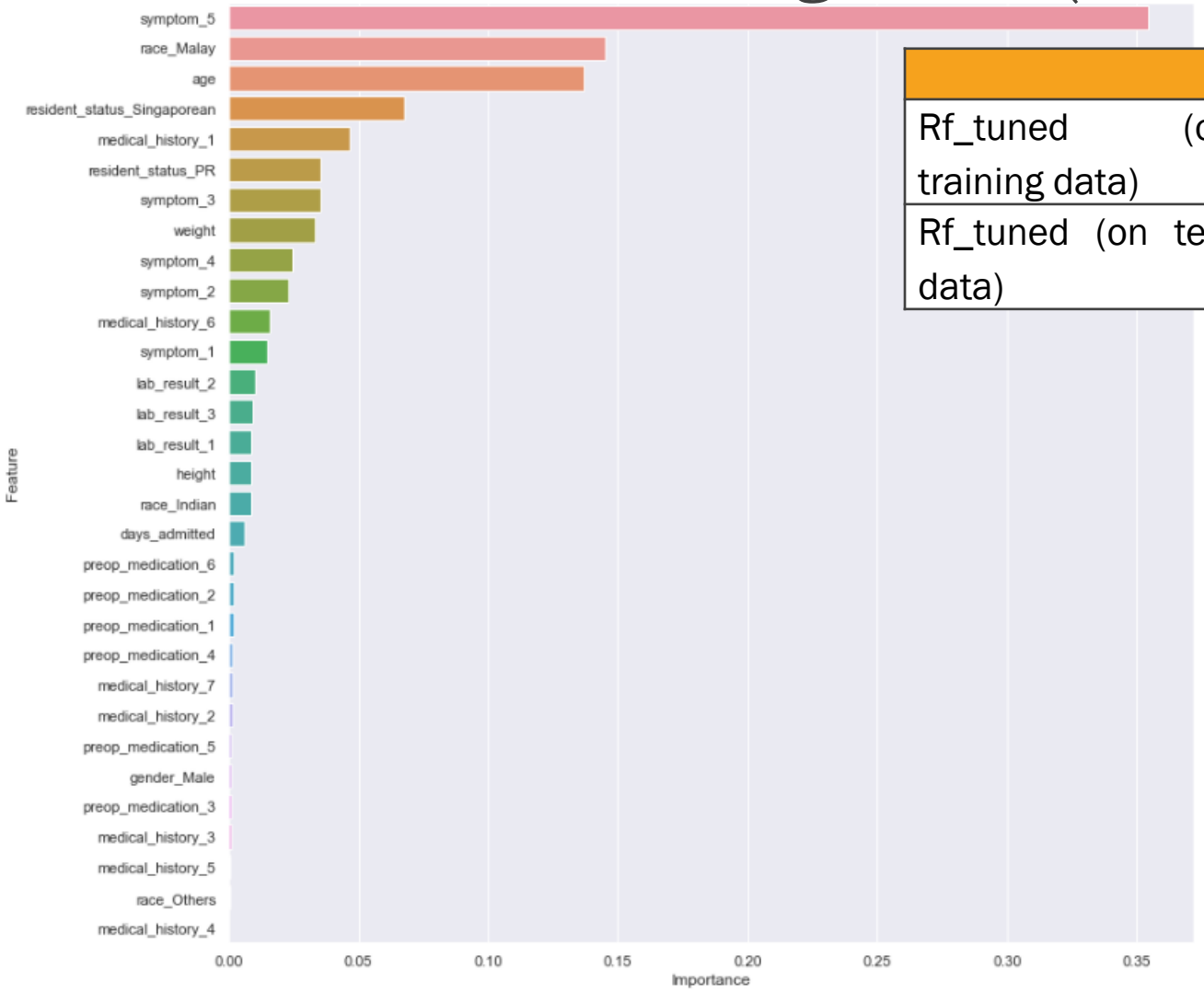


	R^2	MAE	RMSE
Random Forest (on training data)	0.985	0.041	0.056
Random Forest (on test data)	0.877	0.113	0.155

- **Methodology**
  - The dataset was split in a ratio of 70:30 for training and testing (using the same split in the linear regression model)
  - RandomForestRegressor was trained on the training data and tested on test data
- **Findings**
  - Compared to the decision tree model, the random forest before tuning already performs better than dtree_tuned
  - However, it does not perform as well as the linear regression model

Model Development

Random Forest Regressor (Tuned)



	R^2	MAE	RMSE
Rf_tuned (on training data)	0.980	0.046	0.064
Rf_tuned (on test data)	0.881	0.111	0.153

- **Methodology**
  - Randomised search was first performed using RandomisedSearchCV which indicated in the best parameters that ‘max_depth’ was ‘None’.
  - GridSearchCV was used for hyperparameter tuning over ‘min_samples_leaf’, ‘min_samples_split’ and ‘n_estimators’.
- **Findings**
  - The tuned random forest performs better than the decision tree (both tuned and untuned).
  - The top five features are the same as the decision tree

Model Development
Summary of Results

	R^2	MAE	RMSE	Features
TRAINING				
Linear Regression (on training data)	0.975	0.051	0.071	24
Decision Tree (on training data)	**1.000**	**0.000**	**0.000**	**31**
Dtree_Tuned (on training data)	0.893	0.117	0.148	16
Random Forest (on training data)	0.985	0.041	0.056	31
Rf_tuned (on training data)	0.980	0.046	0.064	24
TESTING				
Linear Regression (on test data)	**0.966**	**0.063**	**0.082**	24
Decision Tree (on test data)	0.762	0.162	0.216	31
Dtree_Tuned (on test data)	0.797	0.154	0.200	16
Random Forest (on test data)	0.877	0.113	0.155	31
Rf_tuned (on test data)	0.881	0.111	0.153	31

Conclusion: Decision Tree is overfitting the training data. The best overall model is the **linear regression model** which performs well on both training and test data.

Key Insights

- **Key Insight 1 (Medical history and Symptoms):**
  - Symptoms are generally low to moderately correlated with log_amount (0.154 to 0.578)
  - ‘Medical_history_1’ (0.240) and ‘Medical_history_6’ (0.158) have low correlation with log_amount
- **Key Insight 2 (Symptom_5 and Resident_Status_Singaporean)**
  - ‘symptom_5’ and ‘resident_status_Singaporean’ are important features to predict ‘amount’ and ‘log_amount’
- **Key Insight 3 (Linear Regression Model fits data best)**
  - The linear regression model is the best overall model
  - The data can be described by a linear model. Non-linear models such as decision trees and random forests do not perform very well on the data.

Key Insights

- **Key Insight 4 (The top 5 features)**
  - The top five features across models are same: ‘Symptom_5’, ‘Race_Malay’, ‘Age’, ‘Resident_status_Singaporean’, ‘Medical_history_1’
- **Key Insight 5 (Features excluded from Linear Regression model)**
  - There are 7 features excluded from the linear regression model:
    1. Medical_history_4
    2. Preop_medication_4
    3. Lab_result_1 to Lab_result_3
    4. Days_admitted
    5. Gender_Male

Conclusions

- **Key Drivers of Cost of Care** (in descending order)**:**
    1. Symptom_5
    2. Race_Malay
    3. Age
    4. Resident_status_Singaporean (drives cost negatively)
    5. Medical_history_1

Recommendations

- **Determining Causality:** It is important to note that correlation does not imply causation and further study must be done to understand if there are latent variables. While the linear regression model predicts that certain races or resident status have higher cost of care, there could be existence of **latent confounding variables** which are not featured in the data set (e.g. being of a particular resident status could mean additional taxes or fees, or possibly a different lifestyle that leads to different symptoms but this may not be captured in the dataset).
- **Data Harvesting:** More data is required to build better models as the current dataset is limited to only 2898 entries.
- **Data Inputs**: each bill could be correlated to certain treatments to provide a one-to-one mapping between **bill** and **clinical_data**
- **Adding features**: More features can be added to provide more insight
- **Model Development**: Further model refinement can be done on the random forest model to explore if it is possible to improve performance of non-linear models