AN EXPLORATORY DATA ANALYSIS FOR DRIVERS OF COST OF CARE

LINYE CHEN

11.02.2022

Outline

- **Dataset Overview**
    - Demographic Data
    - Clinical Data
    - Bill Data
    - Master Data

- **Univariate Analysis**
    - Continuous Variables
    - Categorical Variables

- **Multi-Variable Linear Regression**

Section 1 – Dataset Overview

Demographic Data

- 3000 observations of 5 variables, each observation belonging to one unique patient (identified by patient ID)

- New variable generated: Age

Variable	Patient ID	Gender	Race	Resident Status	Date of Birth	Age
Type		Binary - Female - Male	Categorical - Chinese - Malay - Indian - Others	Categorical - Singaporean - PR - Foreigner	Date	Discrete (treated as continuous)

Clinical Data

- 3400 observations of 26 variables, each observation belonging to one unique admission (identified by patient ID and date of admission)

- Based on the number of unique patients (3000), a small portion of patients are admitted more than once

- Assumption: from Client's perspective, the expense of each unique admission is of greater interest than the overall expense of each patient

- Medical history 2 and 5 have missing values (7% and 9%), NA treated as another level

- New variables generated: Length of stay and BMI

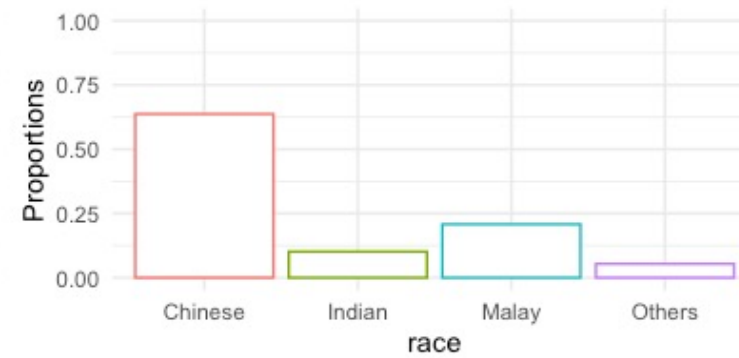| Variable	Patient ID	Date of Admission/Discharge	Medical History 1 -7	Preop Medication 1 - 6	Symptom 1 - 5	Lab Result 1 - 3	Weight	Height	Length of Stay	BMI
Type		Date	Categorical (1/0/NA)	Binary (1/0)			Continuous			

Bill Data

- Combined Bill ID with Bill Amount, 13600 observations of 5 variables

- Based on the number of unique admission record (3400), each admission correspond with multiple bills

- Sum the bill amount for each unique admission, generating 3400 observations of 3 variables

Variable	Patient ID	Date of Admission	Bill Amount
Type		Date	Continuous

Master Data

- Combined Demographic, Clinical and Bill data by patient ID and date of admission

- 3400 observations of 31 variables

- Categorical variables:

Variable	Readmitted Status	Gender	Race	Resident Status	Medical History 1 - 7	Preop Medication 1 - 6	Symptom 1 - 5
Levels	(1/0)	- Female - Male	- Chinese - Indian - Malay - Others	- Singaporean - PR - Foreigner	(1/0/NA)	(1/0)	(1/0)

- Continuous variables:

Variable	Amount	Lab Result 1 - 3	Weight	Height	BMI	Length of Stay	Age

Section 2 – Univariate Analysis

# Categorical Variables – Demographics
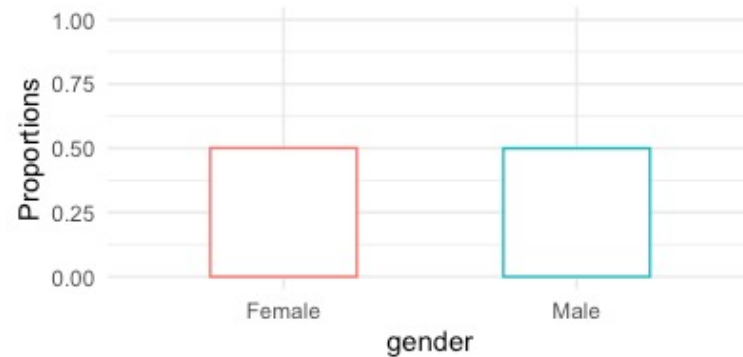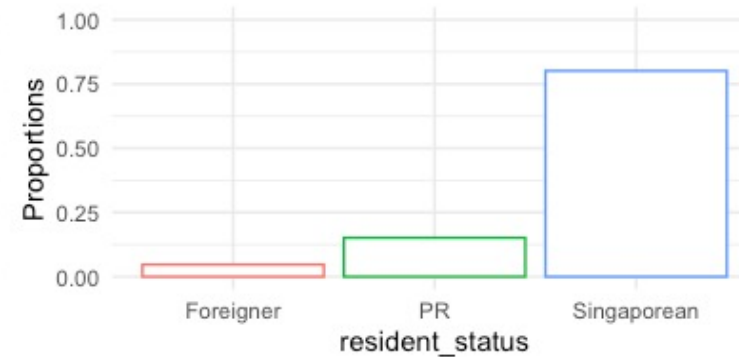
- Readmitted status: majority being first-time admission with 12% being re-admitted

- Race: majority being Chinese and Malay, with 10% Indian and 5% Others

- Gender: equal proportion of female and male

- Resident status: majority being Singaporean, with 15% PR and 5% foreigner
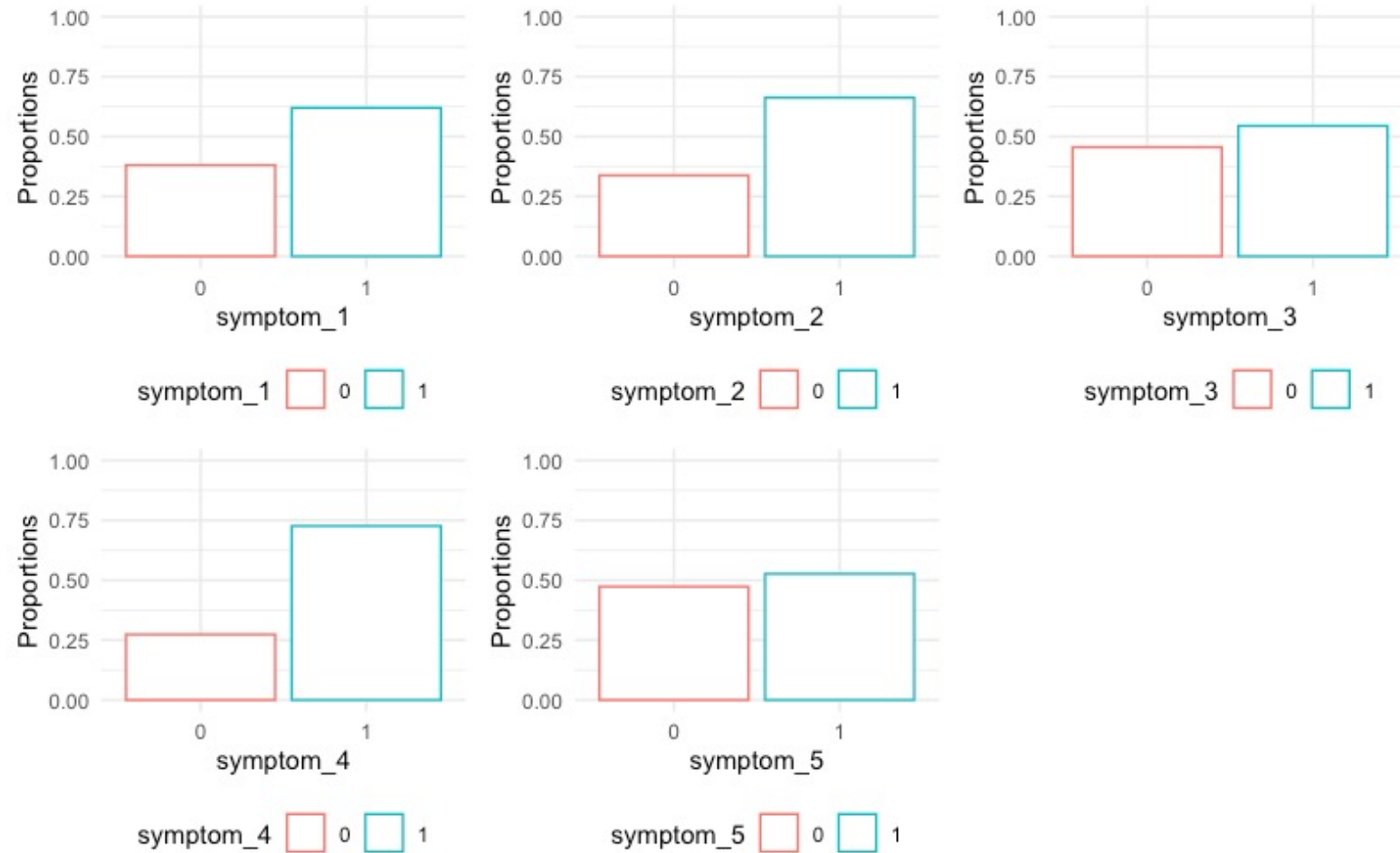
Categorical Variables – Medical History 1 - 7
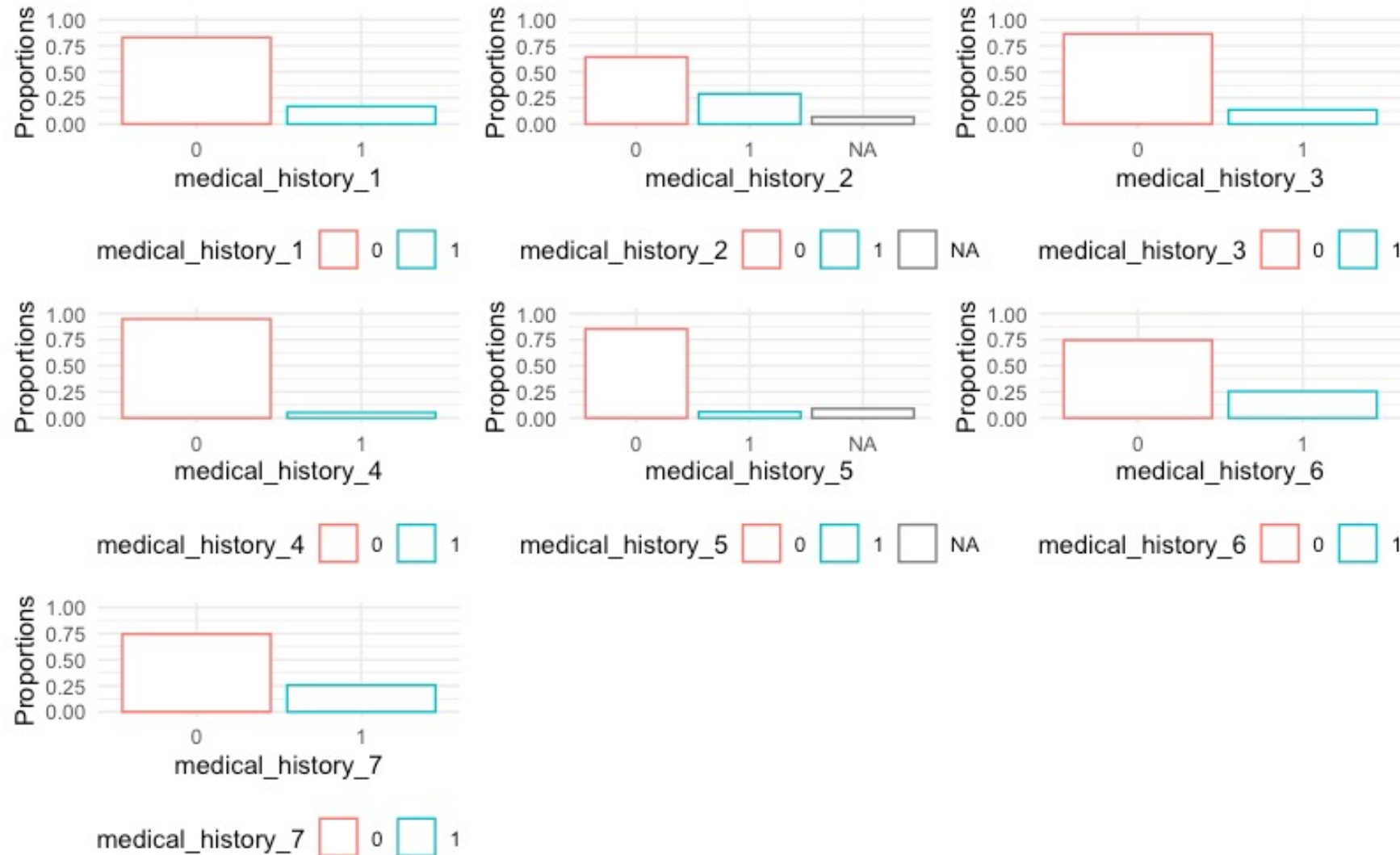
For all 7 medical histories, only a small portion is YES, unbalanced data to be noted

Categorical Variables – Preop Medication 1 - 6
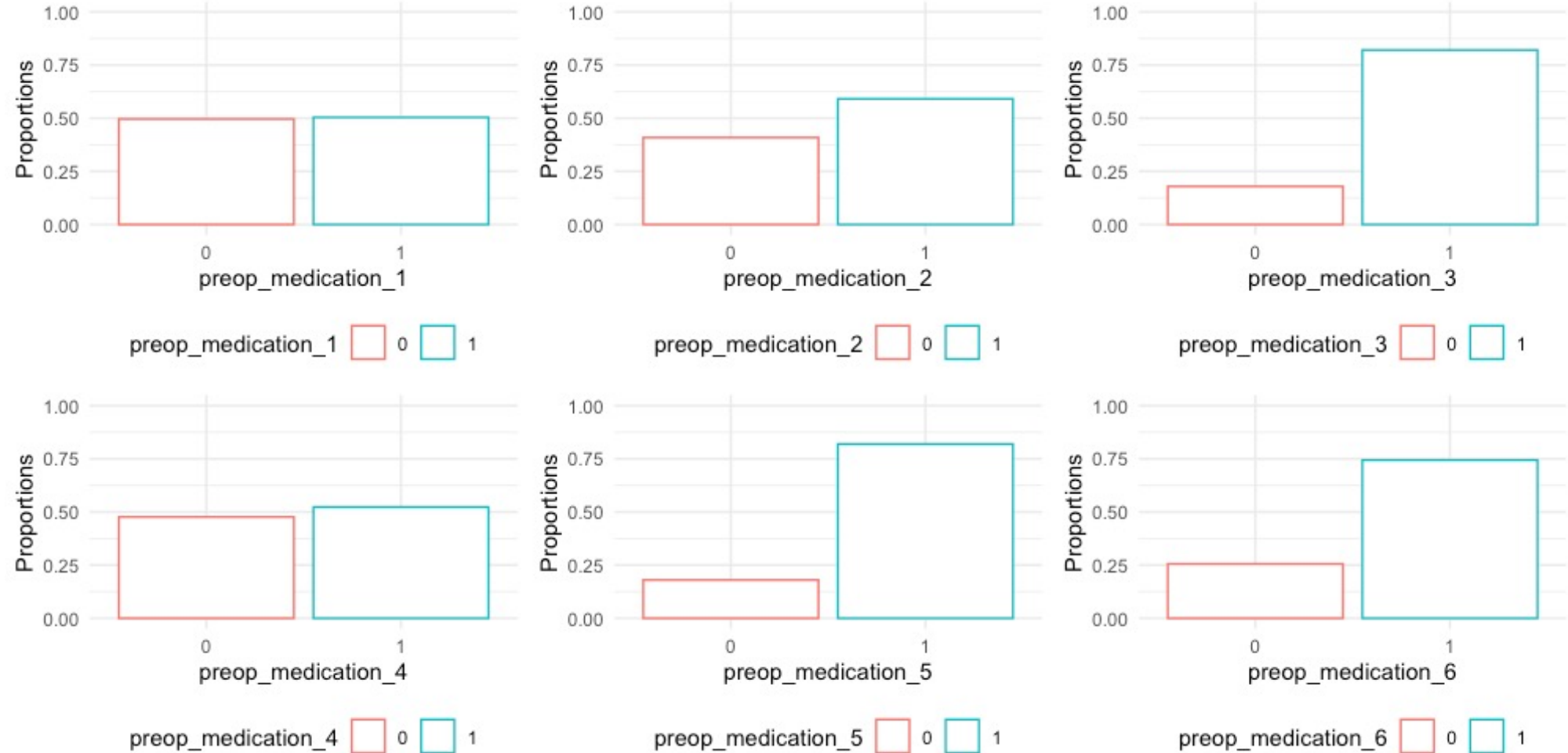
Most of the admission records show usage of preop medication 1, 2, 3, 4, 5, or 6

Categorical Variables – Symptom 1 - 5

Most of the admission records present symptom 1, 2, 3, 4 or 5

# Association between Categorical Variables and Amount

■ Applying one-way ANOVA, 13 categorical variables identified to be associated with amount

❑ Gender

❑ Race

❑ Resident status

❑ Medical history 1

❑ Medical history 6

❑ Medical history 7

❑ Preop medication 2

❑ Preop medication 6

❑ Symptom 1

❑ Symptom 2

❑ Symptom 3

❑ Symptom 4

❑ Symptom 5

Continuous Variables - Amount

Histogram of amount

Histogram of log(amount)
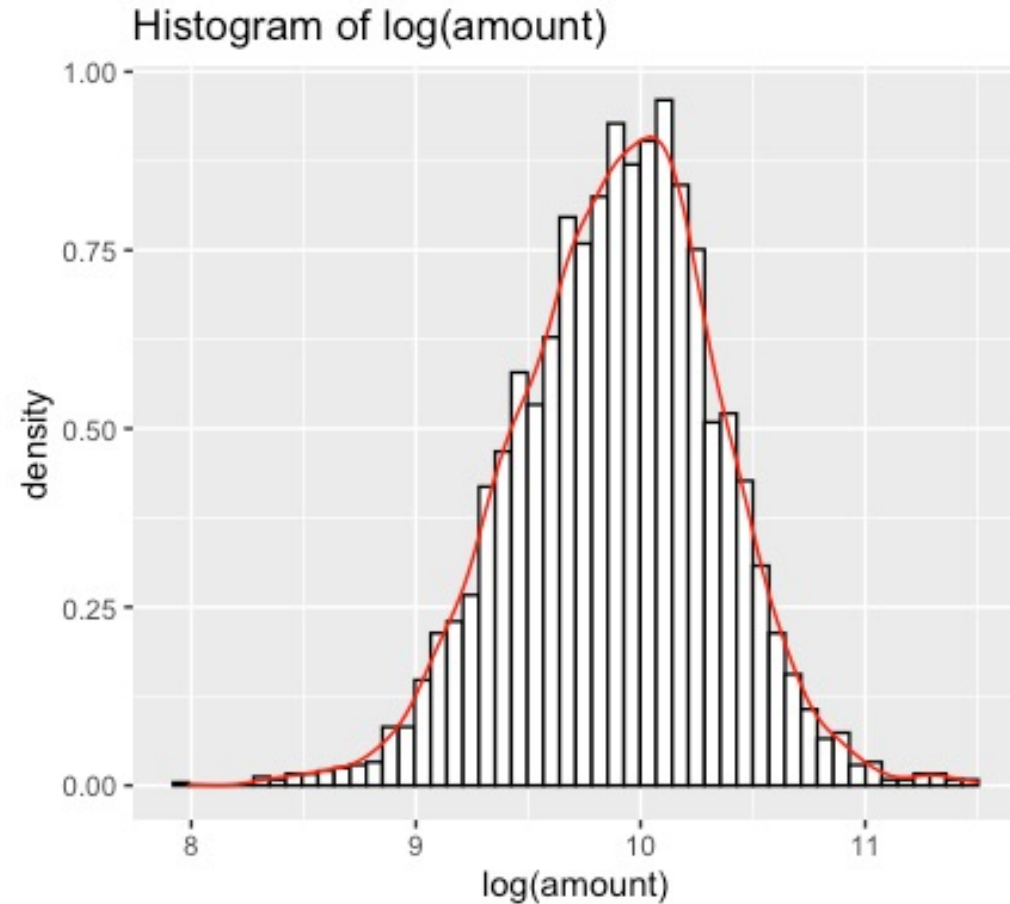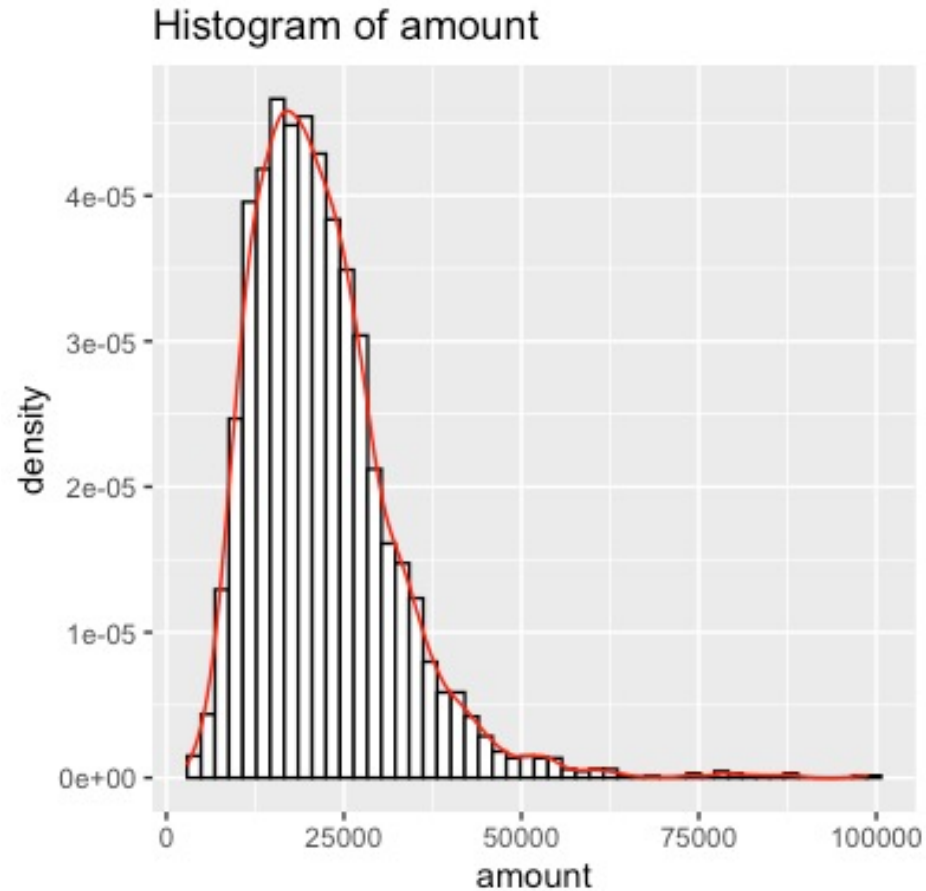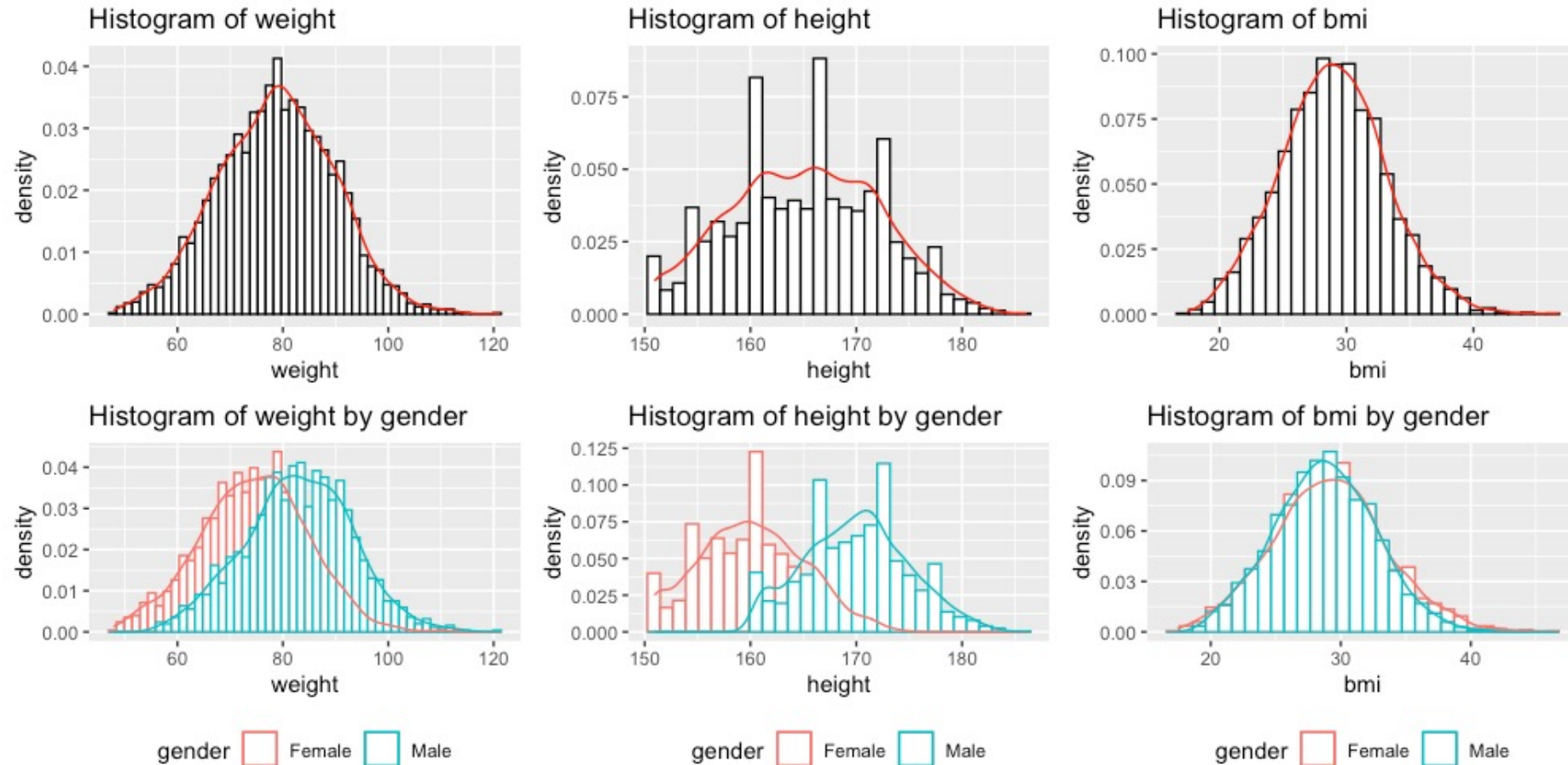
Due to the right skewedness of bill amount, log(amount) is taken to approximate normal distribution

Continuous Variables – Lab Results 1 - 3

Histogram of lab result 1

Histogram of lab result 2

Histogram of lab result 3
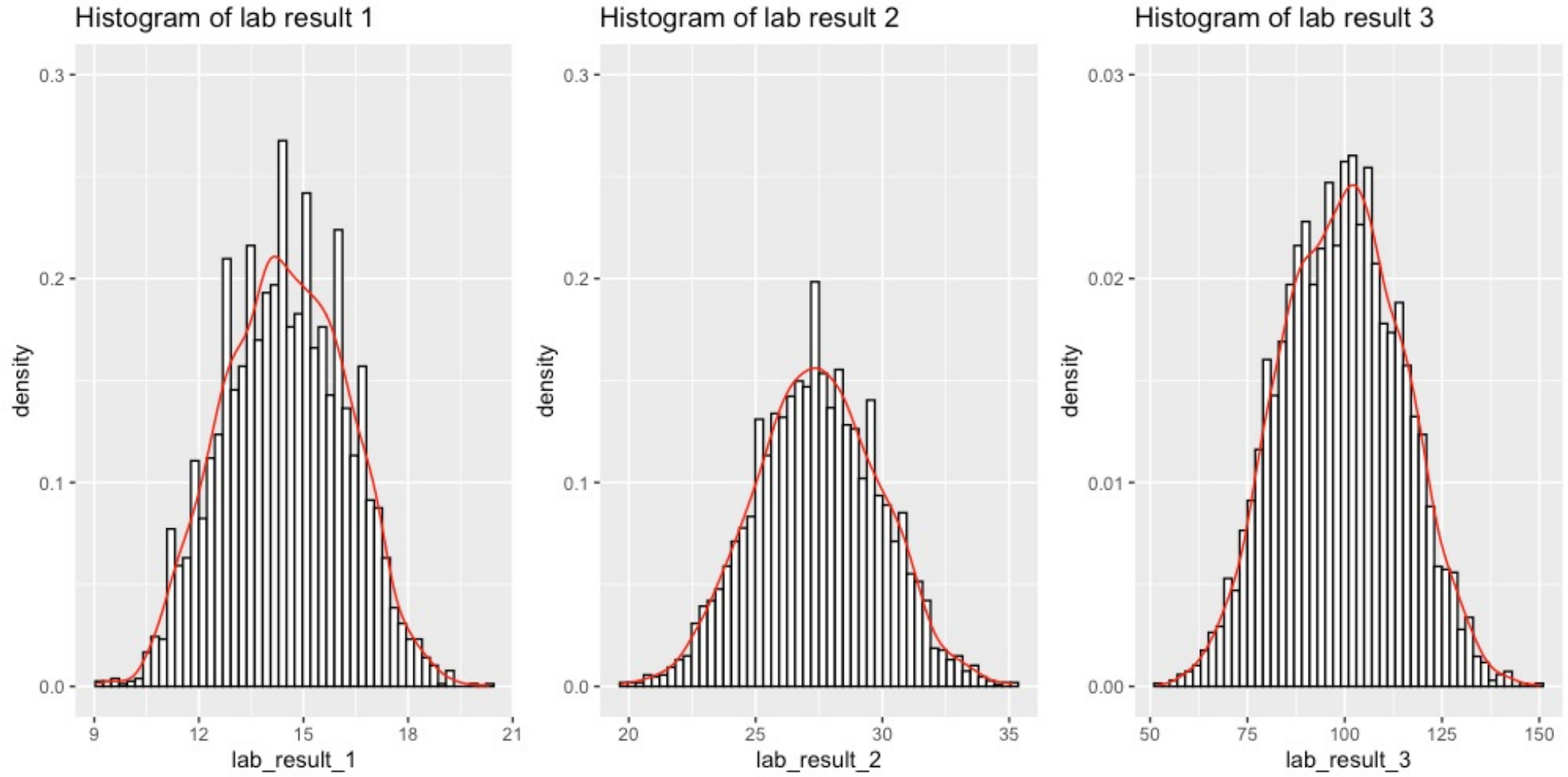
Lab results 1 to 3 all approximate normal distribution

# Continuous Variables – Weight, Height & BMI

- Weight and height are both significantly associated with gender, whereas BMI between female and male shows no significant difference

- In the following analysis, BMI will be used instead of weight and height

Continuous Variables – Age & Length of Stay
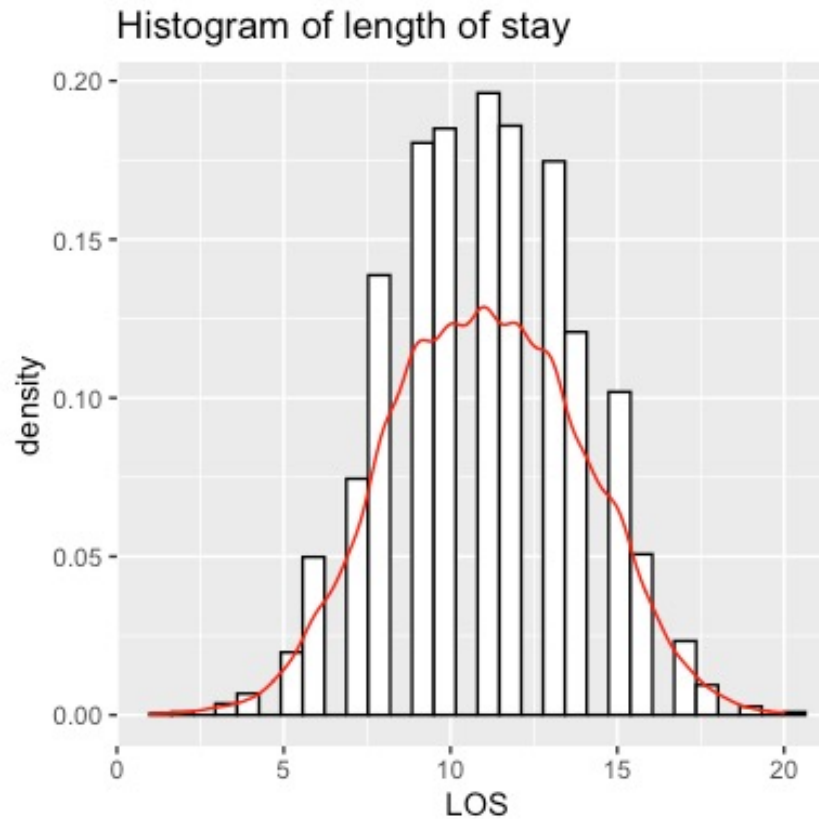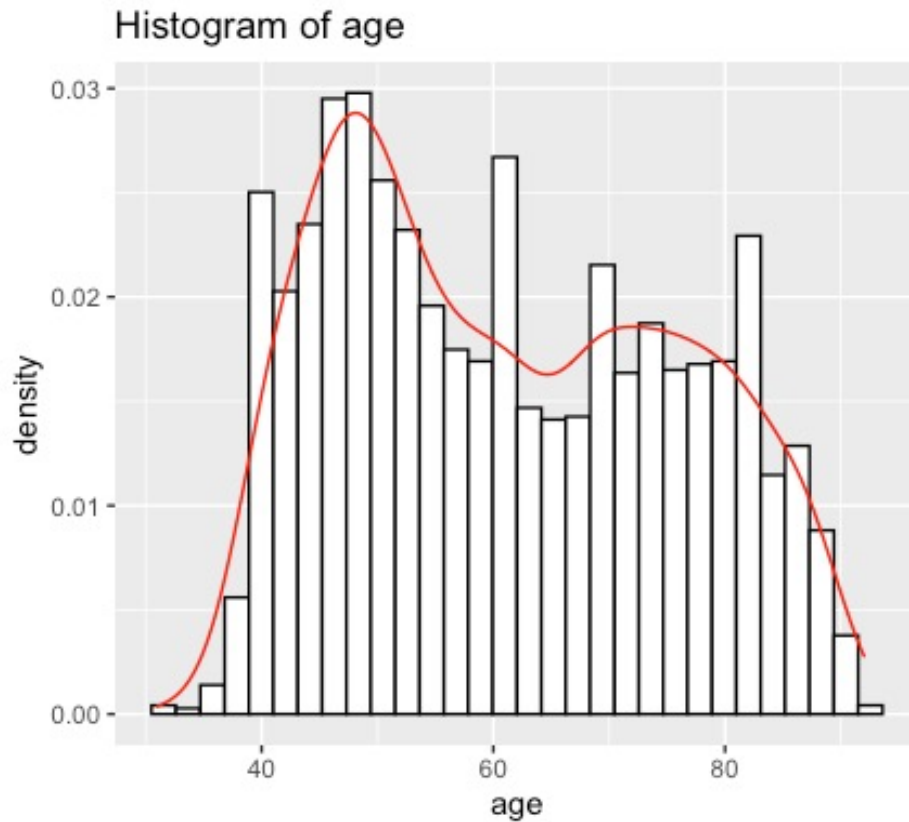
- Age approximates bi-modal normal distribution
- Length of stay approximates normal distribution when treated as continuous
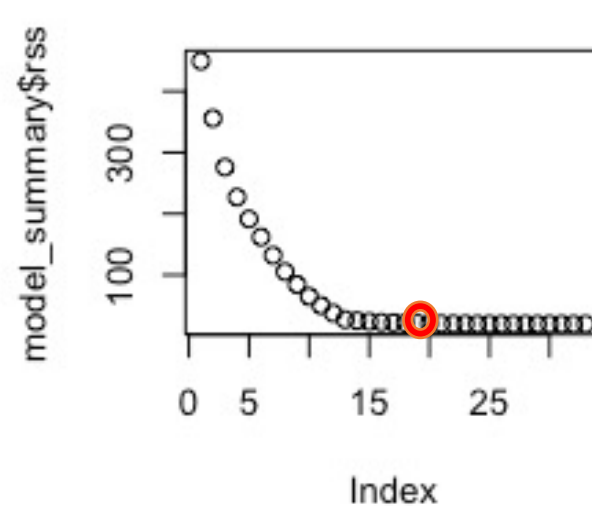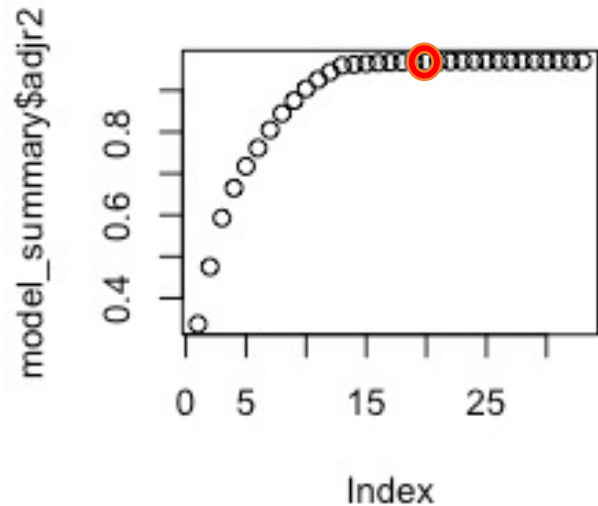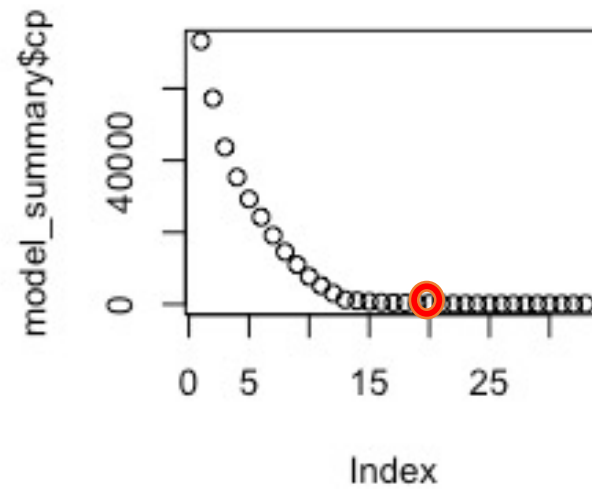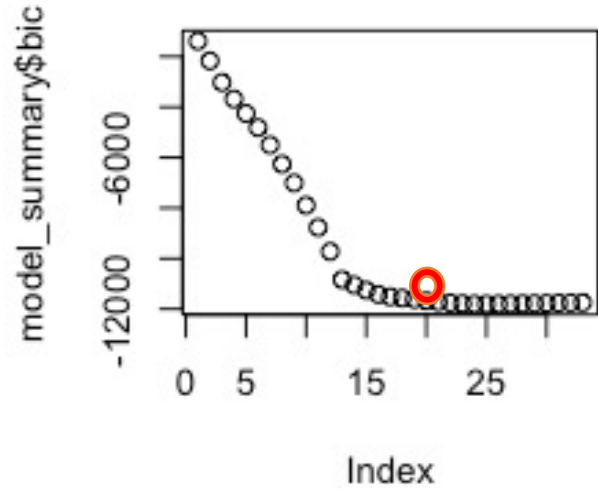
Correlation Among Continuous Variables
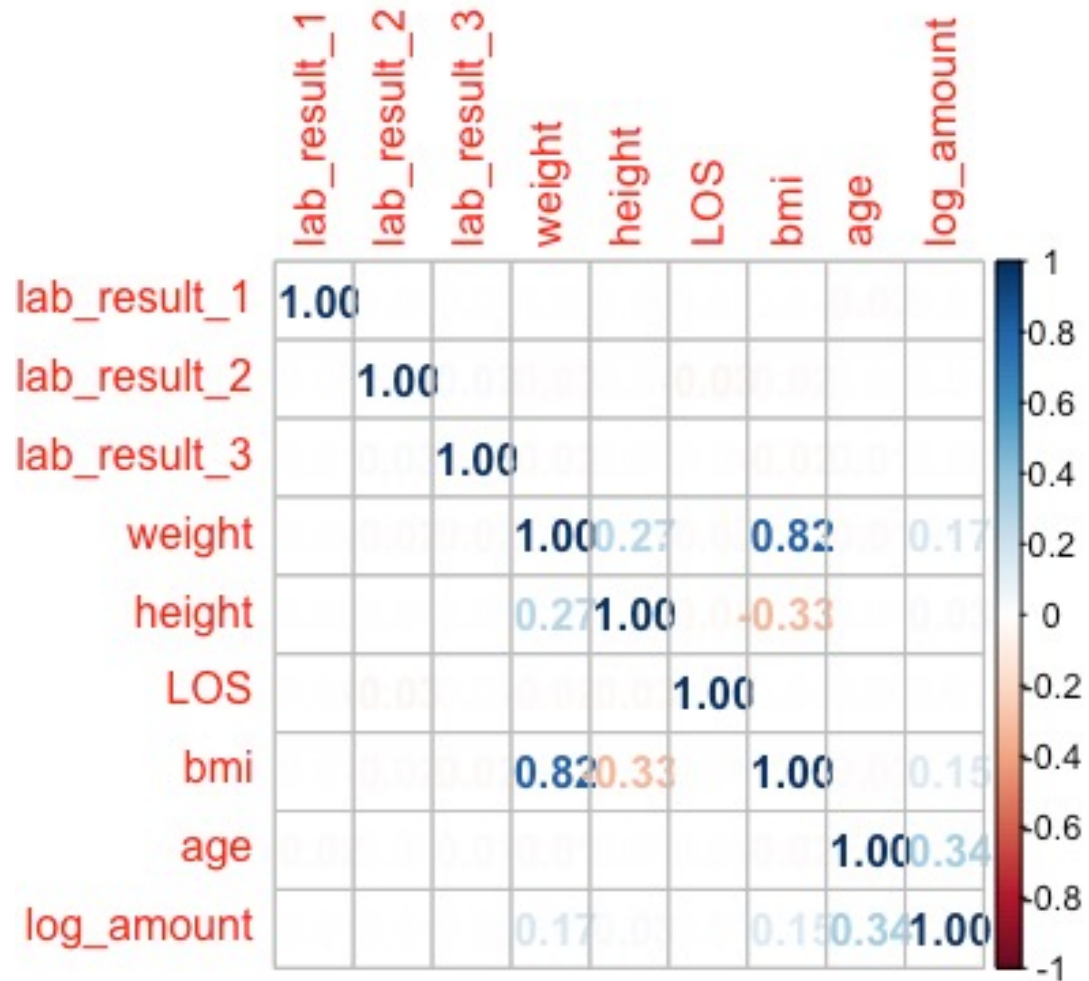
- Age is moderately associated with amount

- The other continuous variables do not show significant correlation with amount

# Section 3 – Multi-Variable Linear Regression

Optimal Linear Regression Model

- Using regsubsets() function to identify the best linear models with different number of independent variables

- The 14th model with 12 variables is chosen due to a good balance of low BIC, low Cp, high Adjusted R square and low RSS

  ☐ Gender   ☐ BMI

  ☐ Race   ☐ Age

  ☐ Resident status

  ☐ Medical history 1

  ☐ Medical history 6

  ☐ Symptom 1

  ☐ Symptom 2

  ☐ Symptom 3

  ☐ Symptom 4

  ☐ Symptom 5

# Amount versus Race grouped by Gender and Resident Status
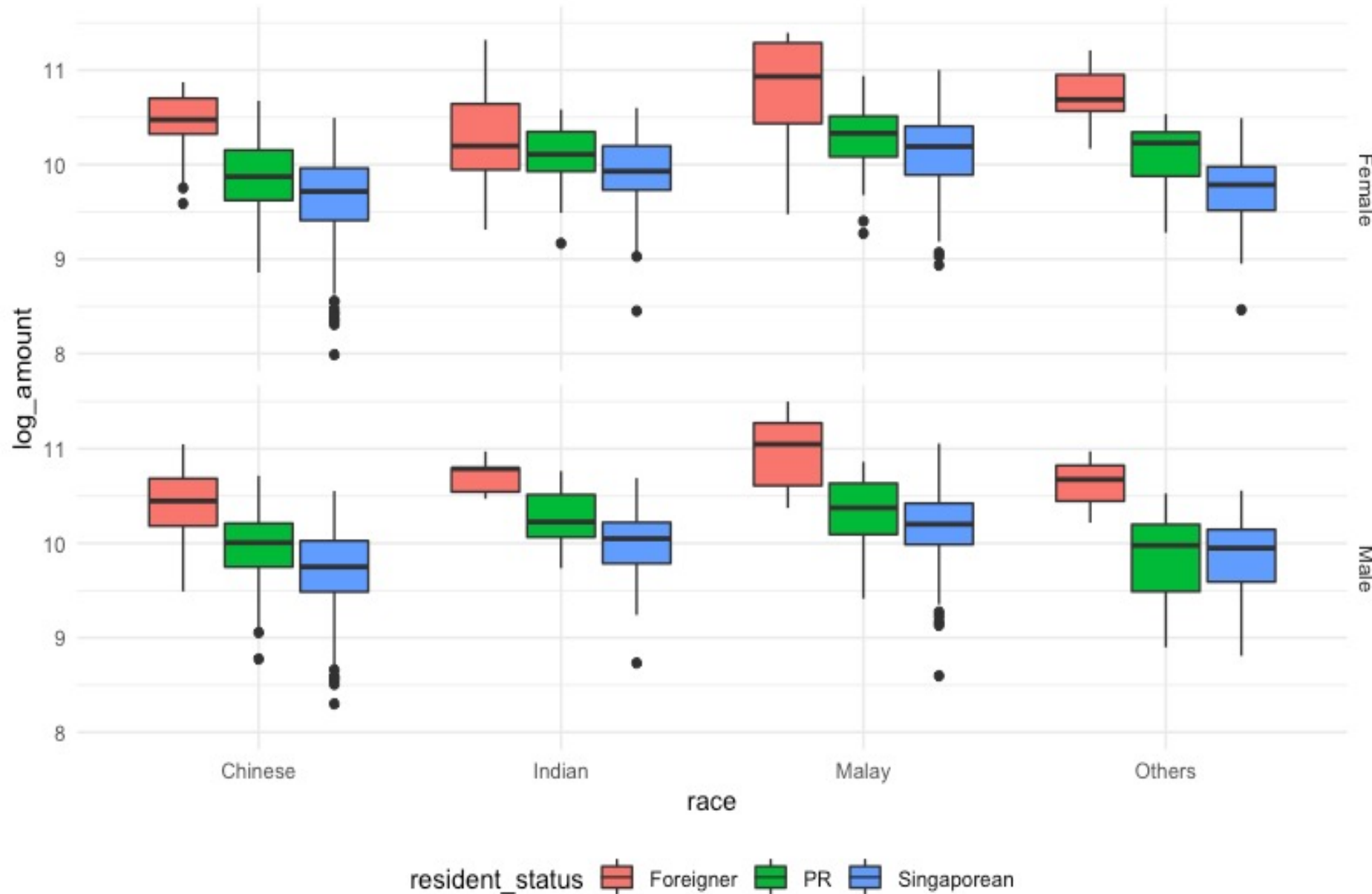
- The median bill amount of Malays is the largest among all races

- Male has slightly larger median bill amount than female

- Foreigner has the largest median bill among, followed by PR. Singaporean has the smallest median bill amount regardless of gender or race

# Amount versus Race grouped by Gender and Resident Status

- If government subsidy for Singaporean's and PR's healthcare service is taken into consideration, the larger bill amount of Foreigners could potentially be explained by the applicability/accessibility of subsidy scheme.

- If datasets originate from different hospitals including private and public hospitals, the sampled Foreigners might have a higher tendency to visit a private hospital than a public one, compared with Singaporean and PR.

- Among all patients admitted by this particular condition, Malays might have a worse underlying health condition compared with the other races, which is not captured by any of the variable reported. As a result, more intensive care is needed for Malays and bill amount is increased.

- Among all patients admitted by this particular condition, Male might have an overall worse health condition compared with Female.

- Relatively small sample size of Foreigners and Malays should be noted.

# Amount versus Medical Histories and Symptoms
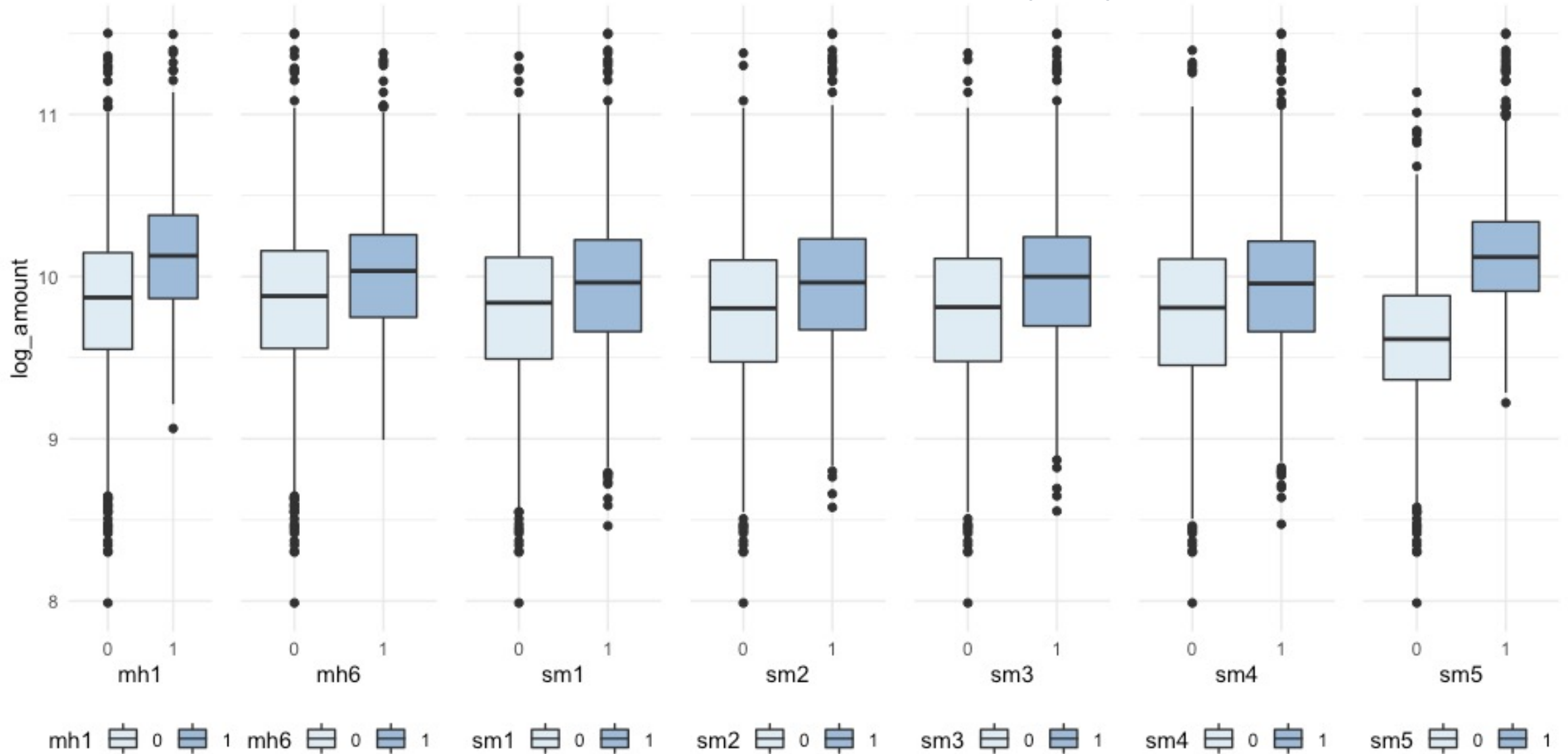
Presence of Medical History 1, Medical History 6 or any symptom from 1 to 5 would lead to a larger median bill amount

Amount versus BMI and Age
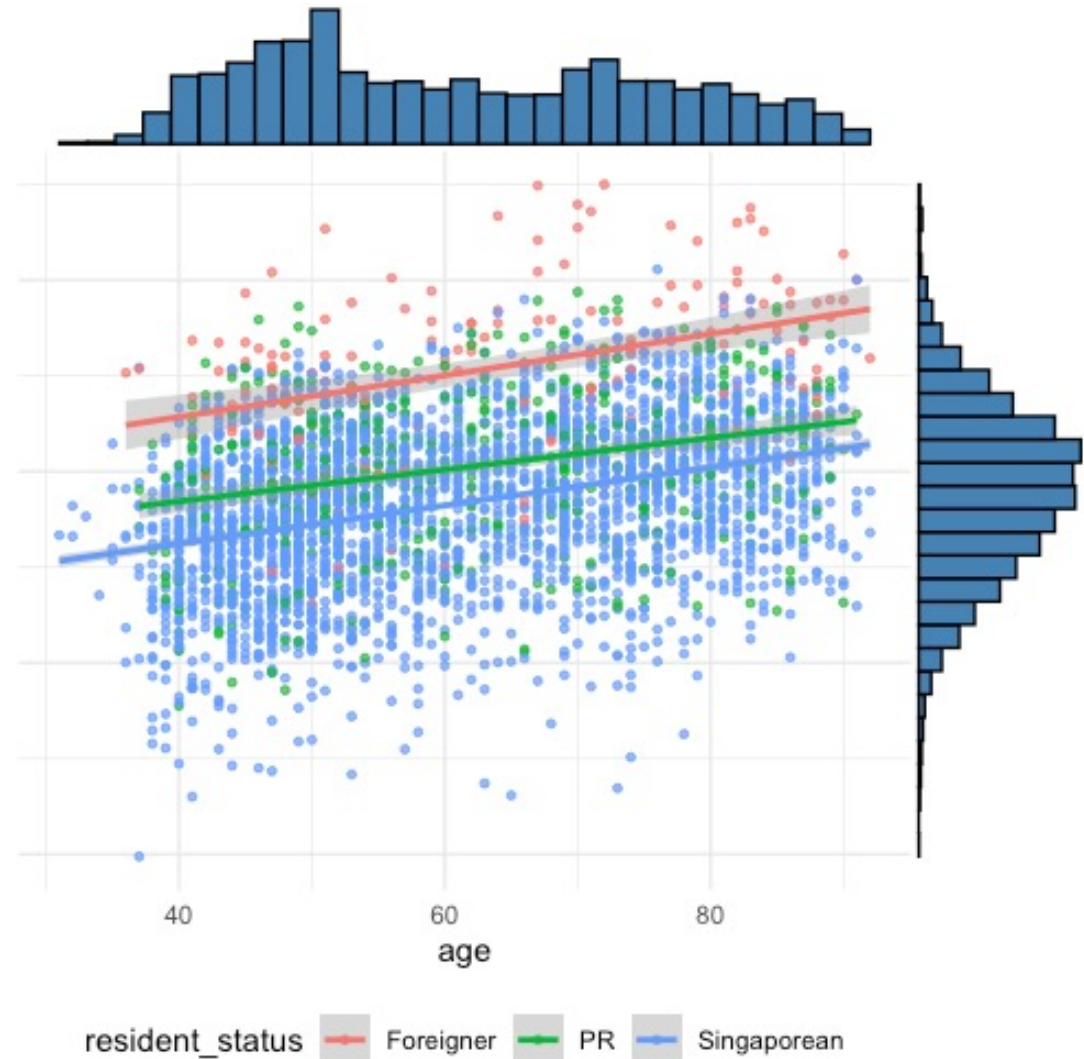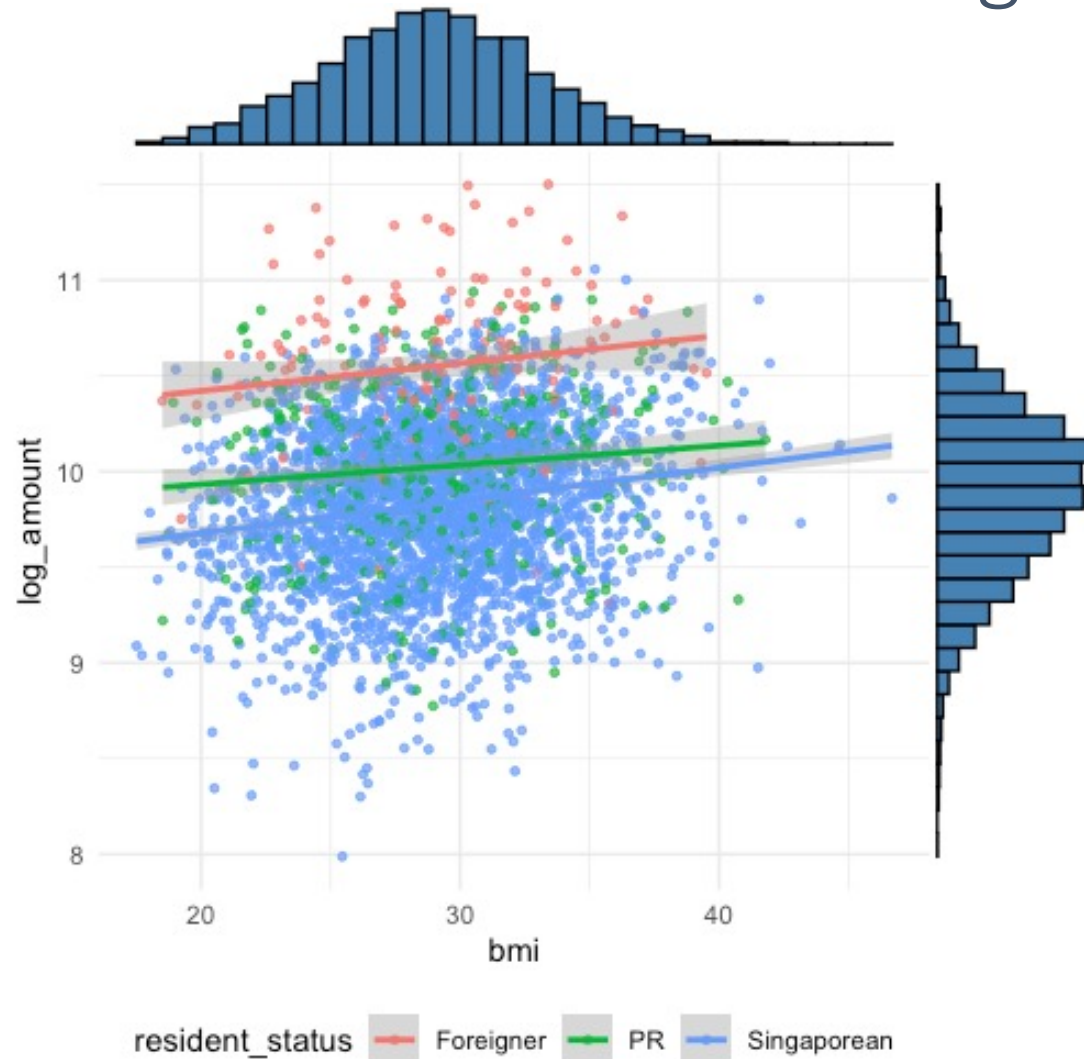
- BMI and Age are both positively correlated with bill amount.

- No significant difference observed in the linear trend of different Resident Status. Similar trend observed when grouped by Race and Gender.

Optimal Linear Regression Model - Limitation
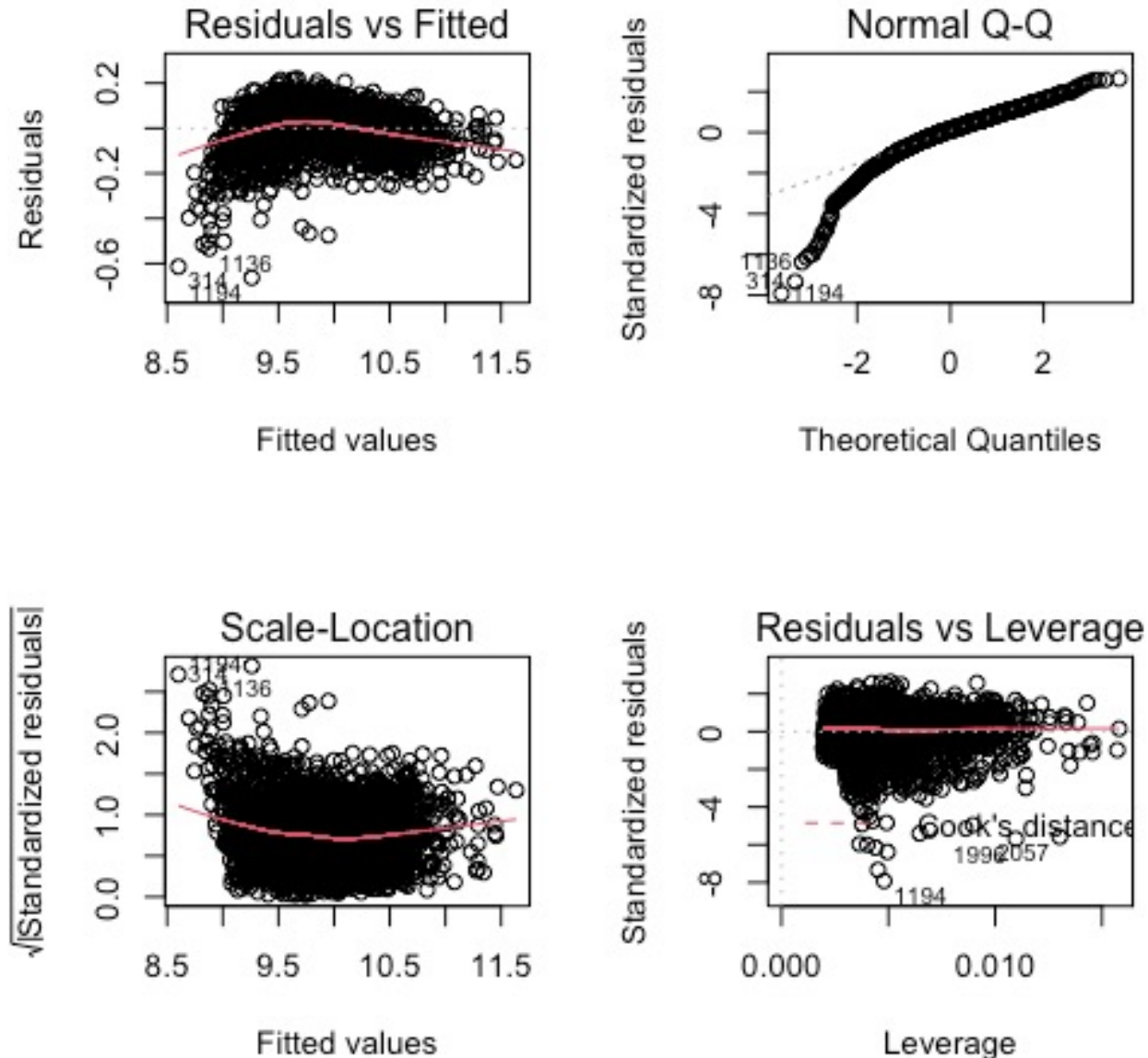
- Diagnostic graphs indicating that data do not fit well into a multi-variable linear model.

- Nevertheless, linear model is good starting point to identify the meaningful variables.