The approach I took was from a standpoint of a company that is offering the same service to the patients where the data was from. I wanted to look at the my intended demographic and the time and cost that these patients spend.

I also assumed that we are looking at Singapore data. As such, I researched past and current population demographics of Singapore and compared it to the demographics that are in the csv files.

From singstat.com, the ratio of males to females shown is almost 1:1. The dominant race in Singapore is Chinese, amounting to 74% of the population. Permanent residents have the least population in terms of resident status and the median age population of Singapore is around 50.

Data Analysis

I first had to look at each of the csv files and check if there are null values. The only file that had null values is the clinical_data.csv. I decided to drop the rows with null values after checking that I cannot map those out. They also account for a small percentage of the data.

After checking for null values, I checked the data types for every feature. I decided to change all date-related columns to datetime so that I can make calculations for the age and number of days admitted. I looked at the binary columns and made sure they were all 0s or 1s and changed them if they weren't. Finally, I changed Singapore citizen to Singaporean for resident status, changed all female, f, male and m to binary and fixed capitalizations and grouped same races.

I merged bill_amount and bill_id first on bill_id and merged clinical_data and demographics on patient_id. I then grouped the patients that had separate bill amount during the same duration of their admission and dropped the duplicates after.

I made the age column by subtracting the date of birth to the current timestamp and I created total_days to represent how long a patient stayed in the clinic.

I merged the merged dataframes and made some data visualizations to see the distribution of the data. It was expected to see the ratio of males to females and the distribution of the patients by race. What was surprising was to see was the resident status of the patients. It might have been expected that permanent residents will be the least, following the population in singstat. It was also interesting to see that Malay foreigners spend the most amount among all other patients.

Modeling
I wanted to know if given the demographics that we are looking at, we can estimate how much the patient would spend. As such, I ran a regression model using the demographics, medical history and symptoms as features to see if these will be indicative of how much a patient would spend. The cross val score showed an R squared score of .89 for random forest regression when the features were filtered based on feature importances.

I also wanted to try and see if we can know whether a person would have a certain symptom, if we can predict if the lab results will be higher or lower than the mean and if we can find out if a patient will stay longer or shorter than 11 days. I tried running classification models but the accuracy score was either lower or marginally better than random guessing. Even with feature selection, the scores did not improve.