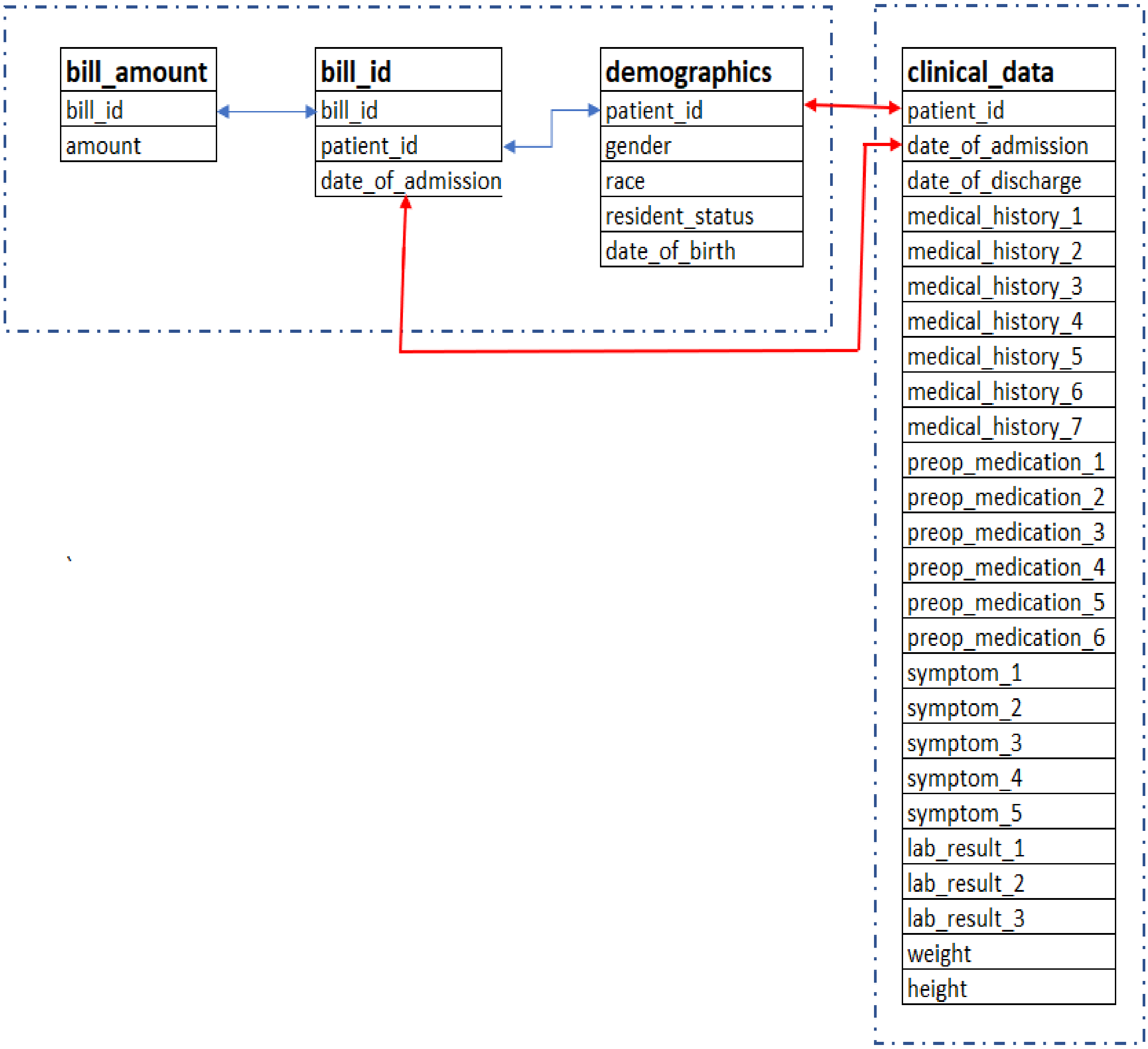
Holmusk - Data Science Task

Combining CSVs

On carefully eyeballing the four CSVs, the 4 tables were joined (LEFT JOIN) in the following manner to create a consolidated dataset. The blue arrows show the columns on which the tables were joined. The red arrows indicate the columns on which the consolidated table formed from bill_amount, bill_id and demographics, was joined with clinical_data table.



Data Cleaning

The data was cleaned by renaming categories in the following categorical variables to maintain consistency.

- **gender** – Replace "m" by "Male" and "f" by "Female".
- **race** – Replace "India" by "Indian" and "chinese" by "Chinese"
- **resident_status** – Replace "Singapore citizen" by "Singaporean"
- **medical_history_3** – Replace "No" by "0" and "Yes" by "1"

Handling Missing Values

Missing values were handled in the following manner:

- **medical_history_2** – 932 missing values – Replaced blank entry by "Not Specified"
- **medical_history_5** – 1216 missing values – Replaced blank entry by "Not Specified"

Feature Engineering

Four new features were engineered to make the analysis more comprehensive. Also, these features helped in delving deeper into the data.

- **number_of_days** – equal to (date_of_discharge – date_of_admission)
- **age_at_admission** – equal to (date_of_admission – date_of_birth)
- **bmi** – equal to (weight/(height^2))
- **weight_classification** – As per following table:

BMI	WEIGHT_CLASSIFICATION
LESS THAN 18.5	Underweight
18.5 TO 24.9	Normal
24.9 TO 29.9	Overweight
29.9 TO 39.9	Obese
MORE THAN 39.9	Extremely obese

Final Columns Data Types

COLUMNS		
Categorical	**Numerical**	**Others**
gender	amount	bill_id
race	lab_result_1	patient_id
resident_status	lab_result_2	date_of_admission
medical_history_1	lab_result_3	date_of_birth
medical_history_2	weight	date_of_discharge
medical_history_3	height	
medical_history_4	number_of_days	
medical_history_5	age_at_admission	
medical_history_6	bmi	
medical_history_7		
preop_medication_1		
preop_medication_2		
preop_medication_3		
preop_medication_4		
preop_medication_5		
preop_medication_6		
symptom_1		
symptom_2		
symptom_3		
symptom_4		
symptom_5		
weight_classification		

BRIEF OVERVIEW

1. In the combined dataset, each row is a unique bill having bill amounts, date of birth, admission and discharge. There are 13600 bills spread across 3000 unique patients and the time horizon in terms of date of admission is from January 2011 till December 2015.
2. The patients are spread across 2 genders – male and female, 4 races – Indian, Malay, Chinese and Others, 3 resident statuses – Singaporean, PR and Foreigner and 5 weight_classifications – underweight, normal, overweight, obese and extremely obese.
3. Lab results, weight and height of each patient with respect to a particular bill are also listed.
4. Finally, we have binary responses (Yes/No = 1/0) of 7 questions related to medical history, 6 questions related to preop medication and 5 questions related to symptom.

ANALYSIS

First we perform our analysis at patient level.

We draw a count plot of different races of patients across different resident_statuses.

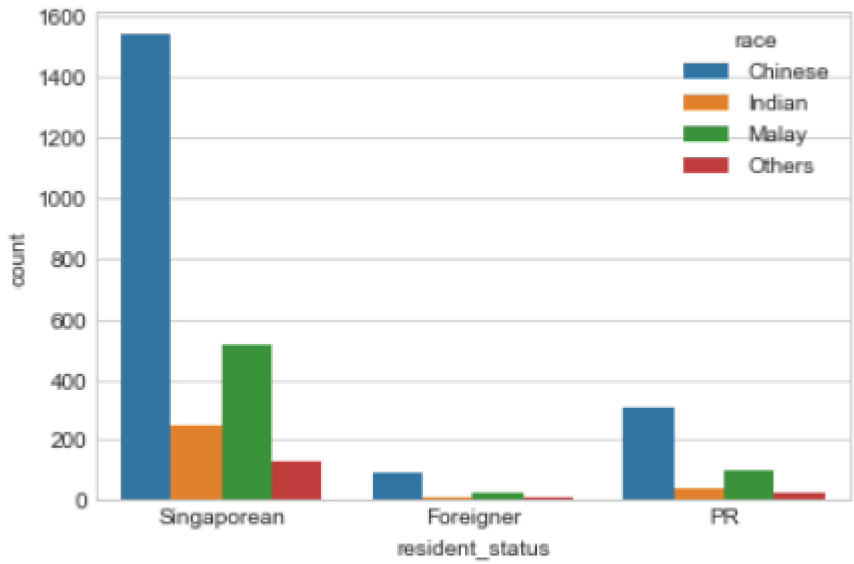


Fig1.

We see that more than half of the total patients have "Singaporean" resident_status and belong to "Chinese" race. Also, across all resident_statuses, count of Chinese is highest, followed by Malay.

Next, we inspect further whether this trend continues across different weight classifications.

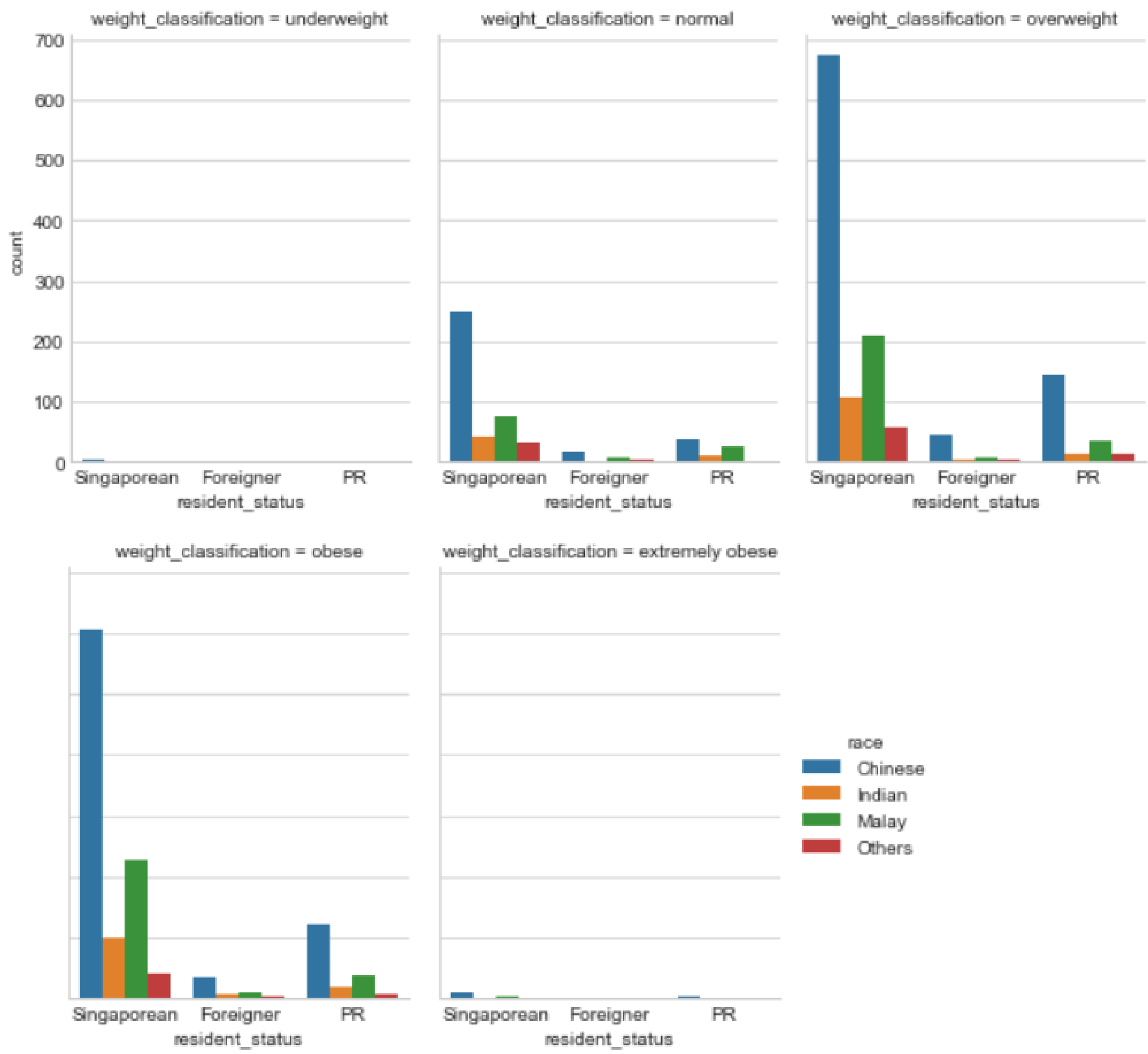
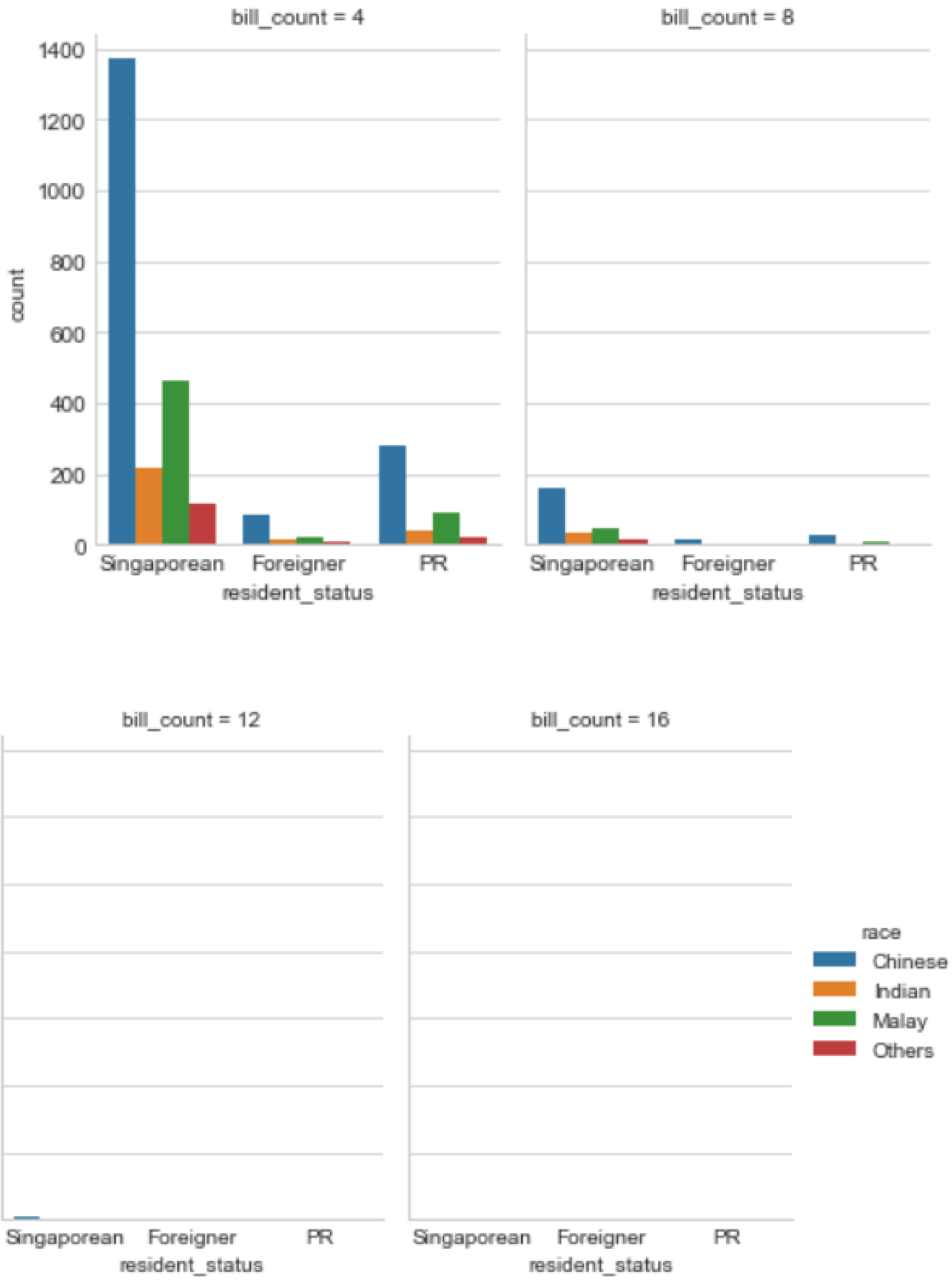


Fig2.

Next, we see that majority of total patients are either overweight or obese. Also, across different weight classifications, in different resident_statuses, the count of Chinese is highest, followed by Malay. Hence the trend is similar.

Now as we see that a particular patients can have multiple bills. Here we see that a patient either has 4, 8 12 or 16 bills.

Fig3.



We see that majority of patients have 4 bills and across different bill counts, in different resident_statuses, the count of Chinese is highest, followed by Malay.

Next, we create a boxplot depicting how bill amount varies across races and resident statuses.

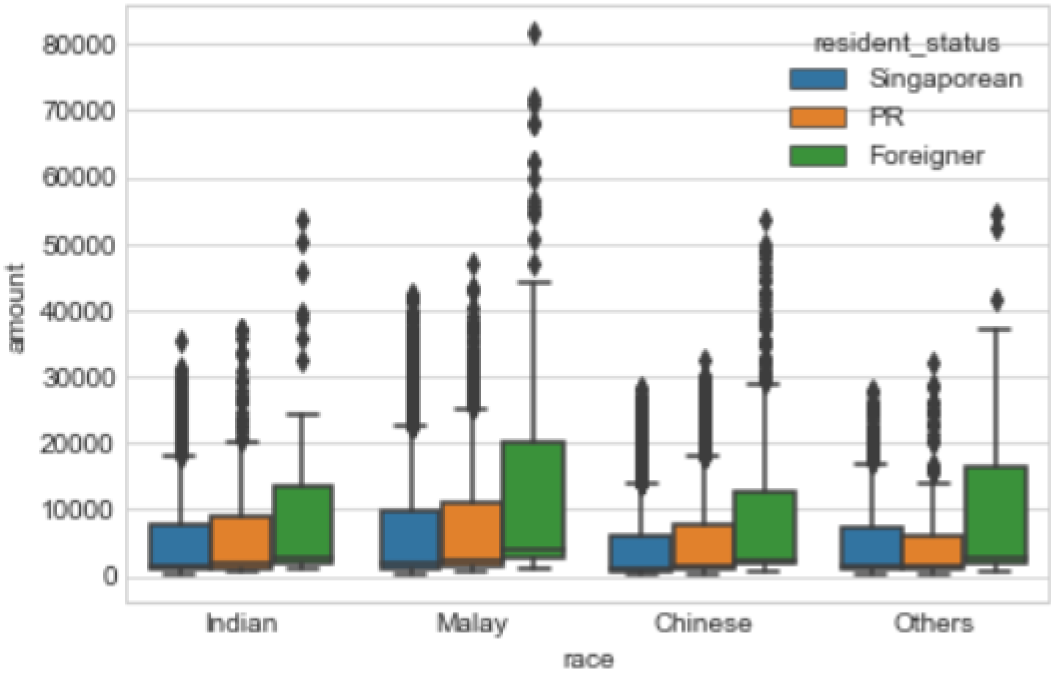


Fig4.

We can clearly see that the range of amount, as well as median amount is higher for foreigners than PRs and Singaporeans even though we saw earlier in Fig1. that count of foreigners is quite low as compared to PRs and Singaporeans. Thus we can hypothesize that foreigners are charged at a premium as compared to other resident statuses.

Next, we plot a boxplot depicting how bill amount varies across different weight classifications and genders.

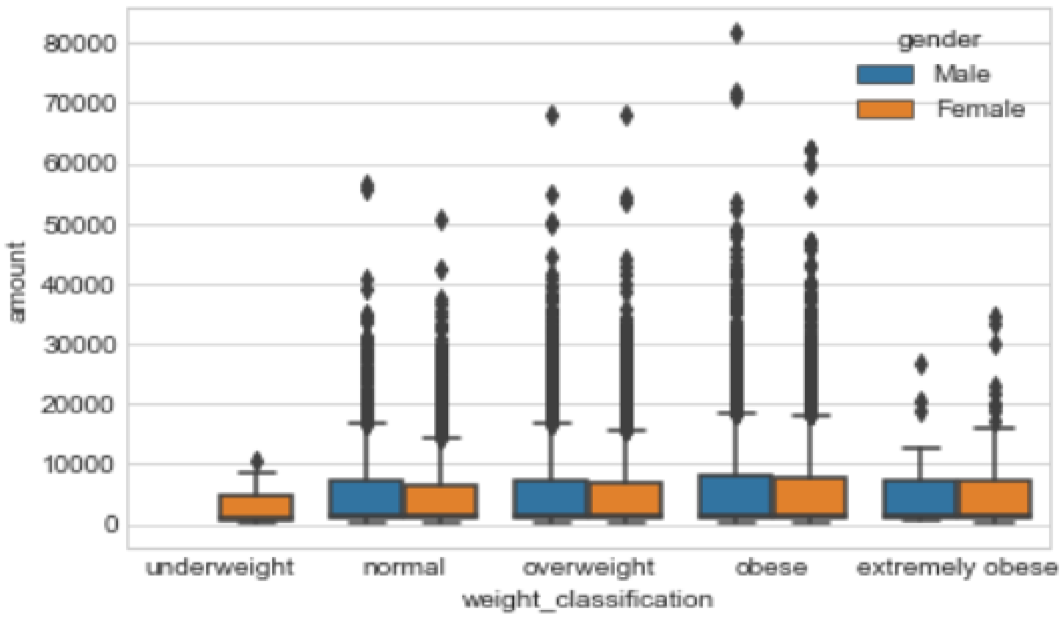
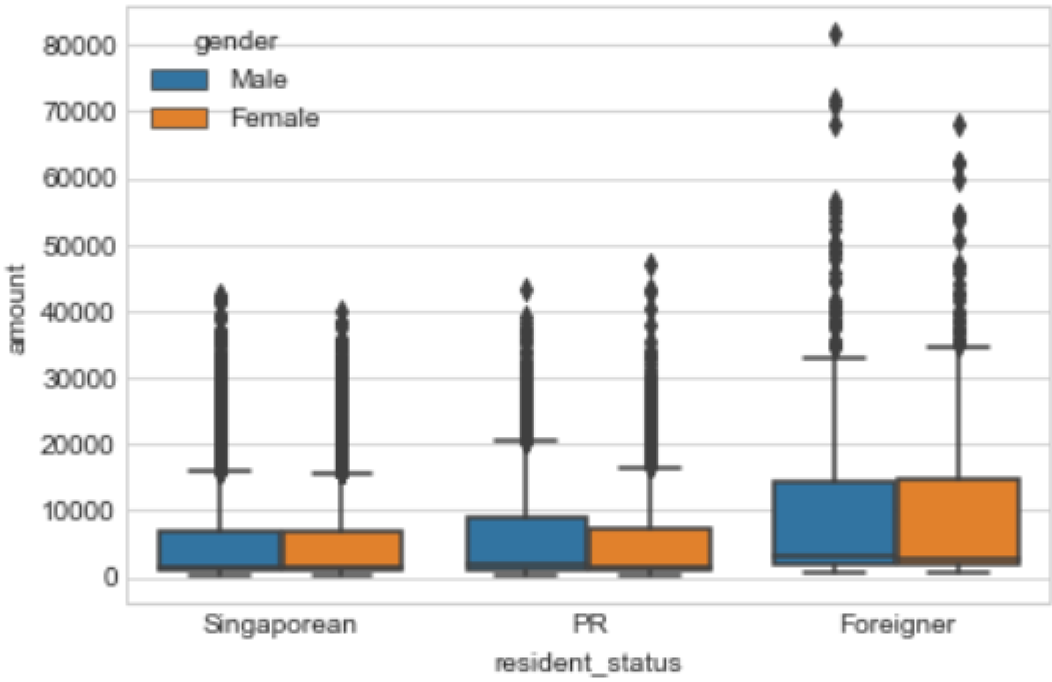


Fig 5.

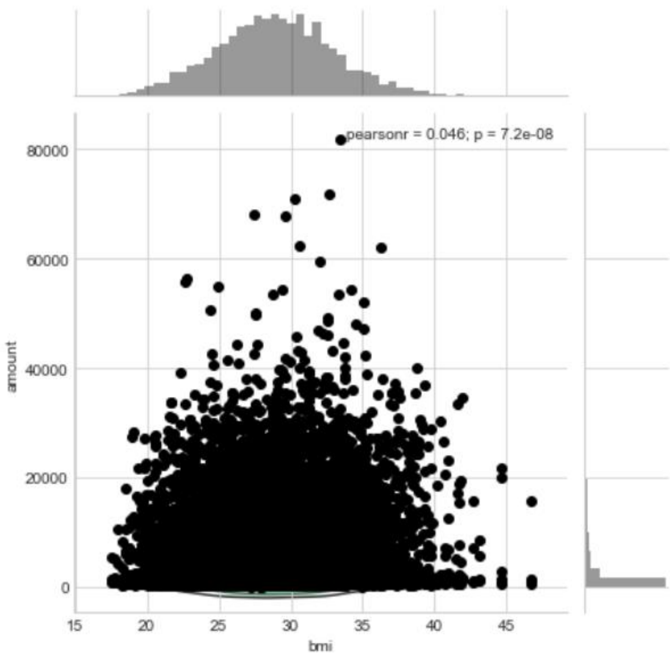
Here, we clearly see that the median amount does not vary significantly across genders and weight classifications.

Next, we plot a boxplot depicting how bill amount varies across gender and different resident statuses.



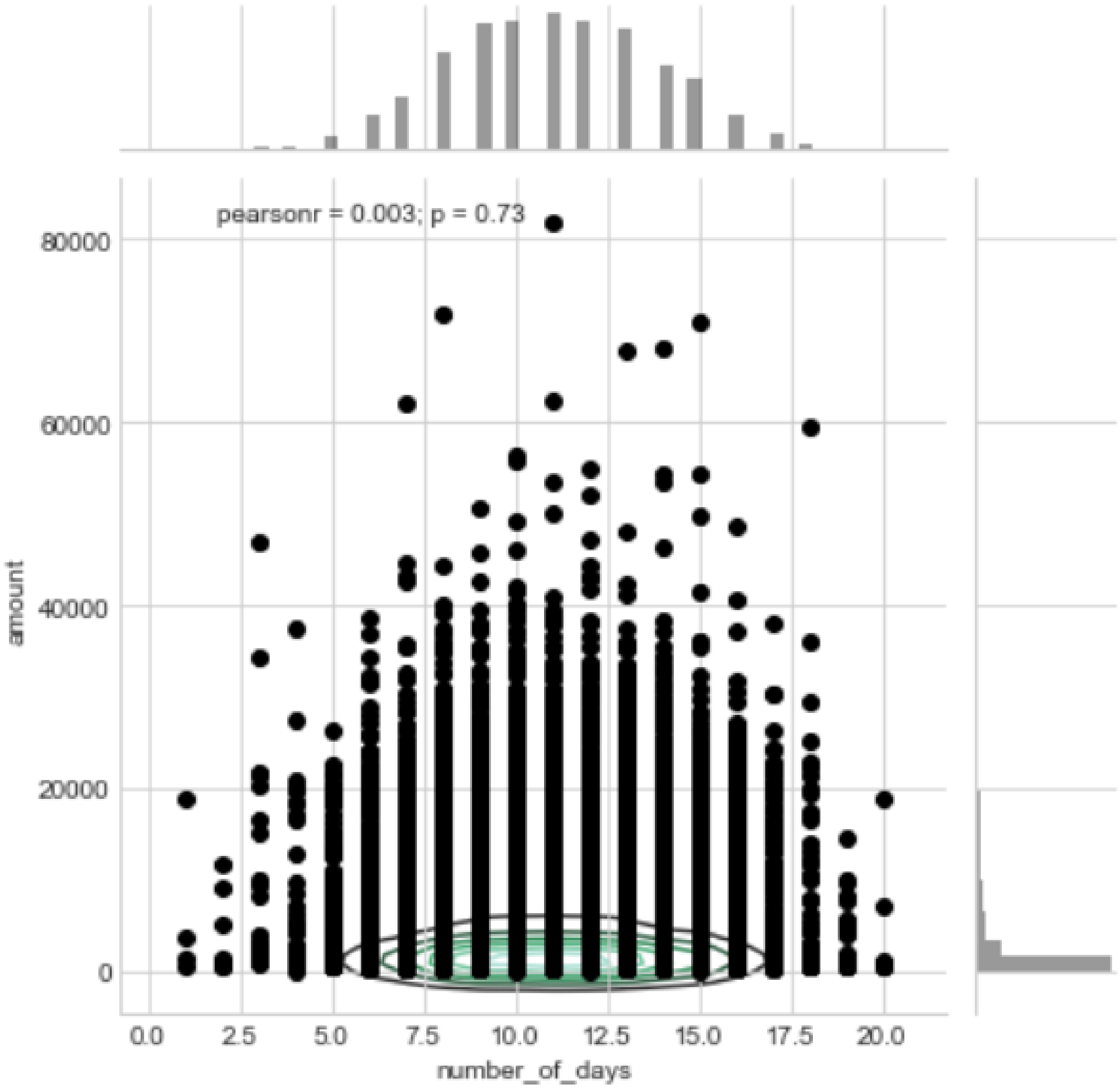
We clearly observe that foreigners have higher median bill amount but there is no significant difference in bill amount with respect to gender.

Next, we plot a scatter plot of amount versus bmi.



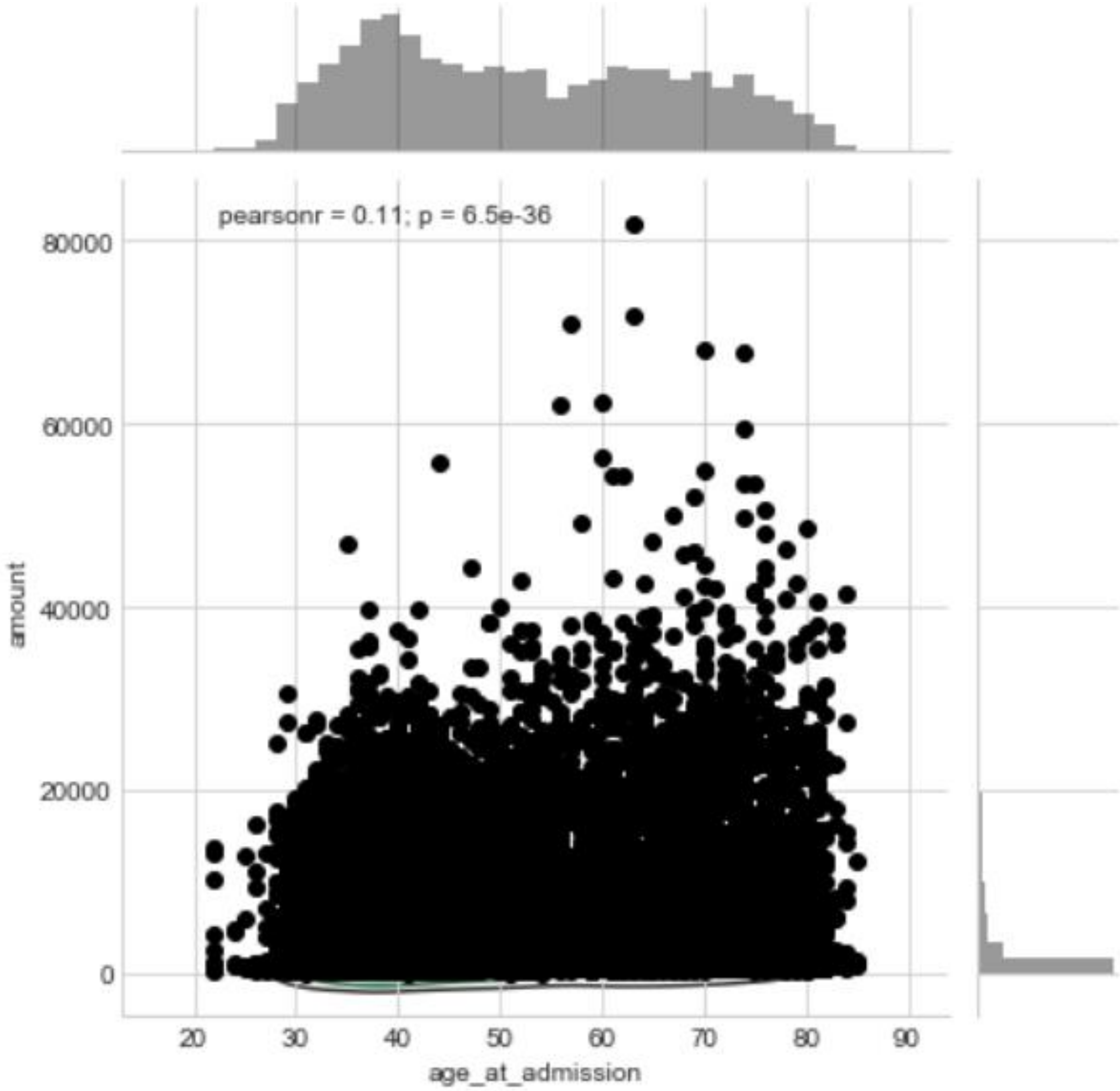
We observe that for underweight, normal and overweight patients, the bill amount increases with increment in bmi. But, for obese patients, the bill amount decreases with increase in bmi.

Next, we plot a scatter plot of amount versus number of days in hospital.



We see that from 2 till 12 days, the bill amount increases with number of days in hospital.

Next, we plot a scatter plot of amount versus age_at_admission.



As expected, we do not see any trend. Hence bill amount does not depend significantly on age at admission.