Holmusk Data Science Assessment Task

Karan J Khanna

# Introduction

* We have been given 4 datasets, each in CSV format.
* **clinical\_data.csv** contains medical information, such as symptoms and past medical records of the patient. Its primary key is **id**.
* **demographics.csv** contains personal information of the patient, like age, gender, race etc. Its primary key is **patient\_id**.
* **bill\_id.csv** contains a unique identifier for each medical bill, and the respective date of admission of the patient. Its primary key is **bill\_id**. Its foreign key is **patient\_id**.
* **bill\_amount.csv** contains the corresponding amount in dollars for each **bill\_id**, which also happens to be its primary key.
* Our end **objective** is as follows:
  + To build a solitary **comprehensive dataframe** using python, containing actionable information from all the above datasets. We need to figure out a way to join all the tables together in a meaningful, neat, and organised manner. This will facilitate smoother analysis of the data.
  + To conduct exploratory data analysis on the newly generated comprehensive dataframe. The **end goal** is to obtain insights behind the **drivers** responsible for medical costs.

# Data Pre-processing

## 2.01 Importing required libraries and modules

* We will be using the following python libraries / modules for this task:
  + Pandas
  + Numpy
  + Python Operating System
  + Matplotlib
  + Seaborn
  + Bokeh
  + Scikit-learn
  + Warnings

## 2.02 Load the datasets using pandas read\_csv function

* Let’s load in the datasets now using the read\_csv function of pandas.
* Each CSV file will go into its own dataframe.
* We finally obtain 4 new pandas dataframes:
  + bill\_id
  + bill\_amount
  + demographics
  + clinical\_data
* It’s always recommended to conduct a sanity check after loading in foreign datasets. We do this by analysing the head (first 5 rows) of each new dataframe.

## 2.03 Merge **bill\_id** with **bill\_amount**: **bill**

* It makes sense to combine all the bill information together into one dataframe.
* The two datasets share the primary key, **bill\_id**.
* We do the following to join them:
  + Use pandas merge function.
  + Left side is bill\_id, and bill\_amount goes on the right side.
  + On bill\_id (the column), because it’s the primary key in both dataframes.
  + How? We will use INNER join, which happens to be the default anyway.
  + Why Inner Join? Because every bill must have a corresponding amount, and vice-versa
* The newly formed dataframe is called **bill**.
* Conduct sanity check.

## 2.04 Convert the date\_of\_admission column into pandas date-time

* Date columns need to specially processed in python.
* Fortunately, pandas makes this easier with its to\_datetime function.
* Conduct sanity check.

## 2.05a Rename primary key **(id)** column in **clinical\_data**

* Take a good look at clinical\_data[‘id’], and demographics[‘patient\_id’].
* Don’t they look alike? That’s because they’re twins! 😊
* We will now rename the column, clinical\_data[‘id’] to clinical\_data[‘patient\_id’].
* We’re doing this to make our upcoming merging process, and further analysis, simpler.
* You can use the following code to do this:



* As always, conduct sanity check, and don’t let your data go insane.

## 2.05b Merge **demographics** with **clinical\_data**: **patient**

* Let’s now come to the information about patients.
* It’s a good idea to have all the patients’ information in one place.
* The two datasets share the primary key, **patient\_id**.
* We do the following to join them:
  + Use pandas merge function.
  + Left side is demographics, and clinical\_data goes on the right side.
  + On patient\_id (the column), because it’s the primary key in both dataframes.
  + How? We will use RIGHT join.
  + Why Right Join? Because it is possible for a patient to be admitted more than once.
* The newly formed dataframe is called **patient**.
* Conduct sanity check.

## 2.06 Convert all date columns in **patient** to pandas datetime

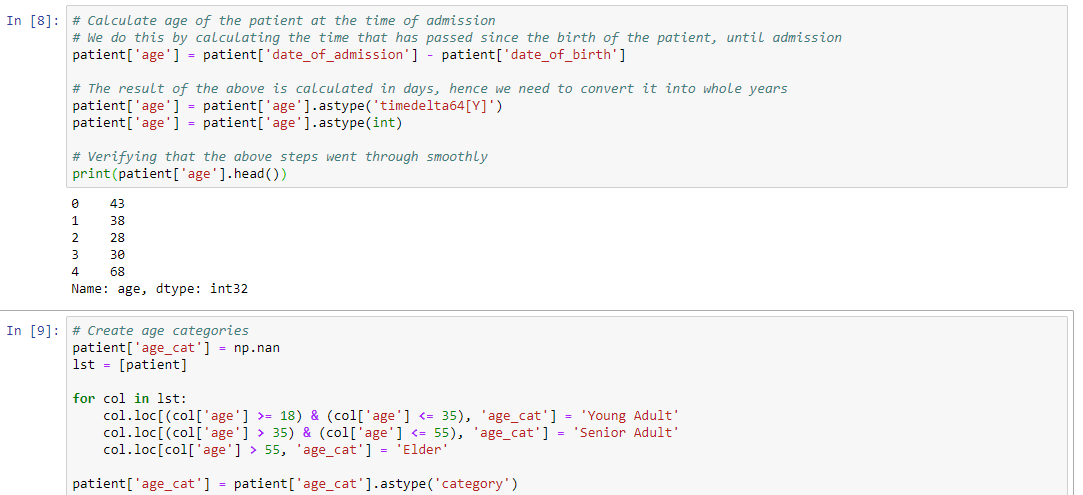
* As we did in the case of the bill dataset, we will now covert ALL date columns in the patient dataframe.
* To do this, we will use pandas to\_datetime function.
* Conduct sanity check.

## 2.07 Generate **age** column in **patient**

* We need to find out the age of the patients. This is crucial for analysis.
* We have been given the following:
  + date\_of\_admission
  + date\_of\_birth
* Subtract date\_of\_birth from date\_of\_admission, and store this data in a new column called age.
* The age column is nice, but it gives us data in number of days.
* Convert age to **timedelta64[Y]** using the **.astype** function on the column of the dataframe.
* Convert it once more to **int** (integer) using the .astype function. This will give us whole numbers in terms of years.
* Conduct sanity check.

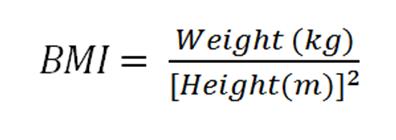
## 2.08 Generate **age categories** from age: **age\_cat**

* We need to cluster the different age groups. This allows us to perform better and more meaningful analysis.
* Let’s assume 3 clusters of patients:
  + Between 18 and 35 (both included) = Young Adult
  + Older than 35, but younger than or currently 55 = Senior Adult
  + Above 55 = Elder
* Create a column in patient called age\_cat. Equate it to numpy nan (null).
* Create a list with patient as the only entity in it.
* Using a for loop, observe the above age criterion, and locate the data from patient[‘age’]. Then according to the condition, store the new information in patient[‘age\_cat’].
* Conduct sanity checks.
* See attached screenshot for reference, if required.



## 2.09 Generate **BMI (Body Mass Index)** column in **patient: bmi\_cat**

* BMI, or Body Mass Index, is a measurement of a person's leanness or corpulence based on their height and weight, and is intended to quantify tissue mass.
* Formula:



* Parameters:

Underweight: BMI is less than 19.

Healthy: BMI is 19 to 29.9.

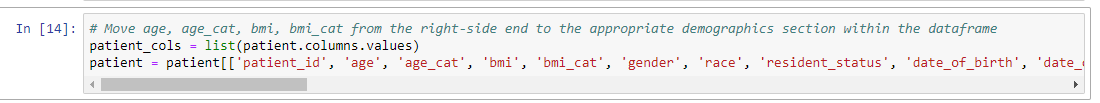
Obese: BMI is 30 or more.

* We have been given height in centimetres. We need it in metres.
* Divide the height column by 100, and store these records in a new column height\_m.
* Get the square of height\_m. Store these records in a new column as height\_m\_squared.
* Divide the weight column by the height\_m\_squared column. This is our true BMI.
* Just as we did for age above, we need to create categories for BMI as well. Take a look at the aforementioned parameters.
* Store the BMI categories as a new column bmi\_cat.
* Conduct sanity checks.
* See screenshot if in doubt.



## 2.10 Position the dataframe appropriately

* Move age, age\_cat, bmi, and bmi\_cat from the right-side end to the appropriate demographics section within the dataframe.
* Simply list the column values.
* Parse the column values as a list within a list to update the existing patient dataframe.
* Neat and tidy. 😊



## 2.11 Convert categorical columns, and transform dirty data

* Gender, race, resident\_status, age\_cat, bmi\_cat are all categorical variables.
* Convert these columns into categorical by using the .astype function.
* Examine the head and tail of these columns to check for discrepancies.
* Gender has two columns each for both male and female patients, ‘Male’ and ‘m’. And ‘Female’ and ‘f’. Replace all occurrences of ‘m’ with ‘Male’. And ‘f’ with ‘Female’. Use .cat.remove\_unused\_categories() on resident\_status column to do this.

Remove the redundant categories. Use .cat.remove\_unused\_categories() on gender column to do this.

* Resident\_status has two columns for Singapore citizens, ‘Singapore citizen’ and ‘Singaporean’. Replace all occurrences of ‘Singapore citizen’ with ‘Singaporean’.

Remove the redundant category, ‘Singapore ciziten’. Use .cat.remove\_unused\_categories() on resident\_status column to do this.

* Race has two columns each for both Indian patients and Chinese patients. ‘India’ and ‘Indian’ along with ‘chinese’ (lowercase) and ‘Chinese’, respectively. Replace all occurrences of ‘India’ with ‘Indian’. And replace ‘chinese’ with ‘Chinese’.

Remove the redundant categories. Use .cat.remove\_unused\_categories() on resident\_status column to do this.

* See attached screenshot for reference.
* Conduct sanity checks.

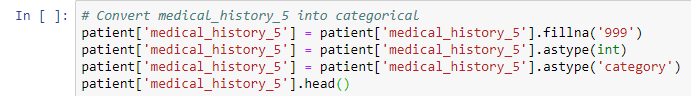


## 2.12 Convert **medical\_history** and **symptom** data into categorical

* Medical\_history is the specific past medical record of the patient. It is a binary variable with only 0 or 1 as possible values.
* 0 means the patient does not have that history. 1 means they do.
* There are 7 types of medical history data given to us. Each with medical\_history\_x as the filename, where x ranges from 1 to 7 (included).
* Convert all 7 of these columns to categorical. Use. astype(‘category’) function.
* You could use an appropriate for loop to do the above, to save you time. But I personally chose to do it manually, because I wanted to work with one column at a time.
* Inspect all medical history dataframes.
* In medical\_history\_3, we have 4 categories. 0, 1, Yes, and No. This is dirty data. In this case, Yes corresponds to 1. And 0 corresponds to No. Replace the same accordingly.
* Drop unwanted categories.
* Next, we have pre-operation medication. There are 6 columns.
* The meaning of pre-operation medication is pretty much self-explanatory.
* Convert all to categorical.
* Inspect for sanity.
* Similarly, for symptom data, we have 5 columns.
* Symptoms are the different types of ailments faced by the patient at the time of admission.
* Convert all to categorical.
* Inspect accordingly for sanity.

## 2.13 Dealing with **missing values**

* By now you should have noticed that there are two columns in the entire environment that have missing values, medical\_history\_2 and medical\_history\_5, respectively.
* Medical\_history\_2 has 233 missing values out of 3400.
* Medical\_history\_5 has 304 missing values out of 3400.
* We replace both with 999.
* This way, we’ll have 3 categories in total, 0, 1 and 999.
* But, you will see later, when we encode the categorical variables with sklearn, that this won’t have an impact on analysis.
* patient['medical\_history\_5'].fillna('999')
* After you have done this, you’ll need to convert the column into integer.
* Once again, you’ll have to convert the integer column into category.
* The above applies to both columns.

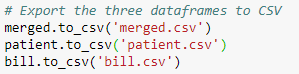


## 2.14 Merge **patient** and **bill: merged**

* Time to merge everything together.
* We’ll again use pandas merge.
* Using RIGHT Join.
* On both ‘**patient\_id**’ and ‘**date\_of\_admission’**.
* It’s important to include date\_of\_amission in the ON criterion of the RIGHT join. A patient can have multiple admissions that is, on different dates.
* Why Right Join?
* Because a patient can, and in fact, is likely to have multiple bills.
* Conduct sanity checks.

## 2.15 Export **patient**, **bill**, and **merged** to CSV

* Always nice to have clean datasets handy. 😊



## 2.16 **Encode** categorical variables

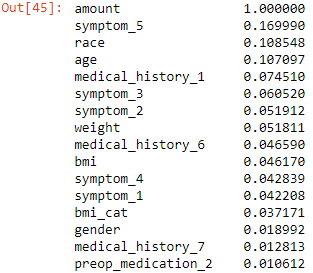
* Before analysing correlations, we must encode all categorical variables.
* If we don’t, pandas will not on its own be able to correlate categorical drivers with amount (amount of bill).
* To encode, we use the scikit-learn library.
* We must manually do the encoding (as shown in the screenshot) for gender, age\_cat, bmi\_cat, race, and resident status.
* For medical\_history\_x, symptom\_x, and preop-medication\_x, we will use a for loop to save time, as shown.



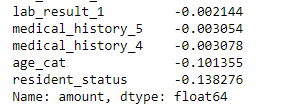
# Analysis

## 3.01 Calculate correlation of **amount**

* Pandas has a built-in method called .corr to calculate correlation of each column in a dataframe with all the other columns.
* We want to calculate the correlation of **amount** (which is the amount paid for each bill). After all, that has always been our objective. To analyse the drivers behind the cost of medical care in Singapore.
* merged.corr()['amount'].sort\_values(ascending=False)
* It’s wise to sort the values in descending order.
* **Note:** Don’t forget about the negative correlations. They are just as important and relevant. They will be located at the bottom in this case.
* Positive correlation- two variables move in same direction. Negative- two variables move in the opposite direction.

****

* We can observe from the above table that, in terms of **positive** correlation, symptom\_5, race, age, medical\_history\_1, symptom\_3, symptom\_2, weight (we will use bmi as it’s a better indicator of a patient’s status of being in shape) are the key drivers of amount.
* **Symptom\_5** has a whopping **17%** correlation with the amount. Certain ailments like fever, headaches, weakness are common in a large variety of medical conditions. This needs to be investigated accordingly. We will do this later on.
* **Race** has almost **11%** correlation with the amount. We’ll look into this as well.
* **Age** has almost **11%** correlation with the amount. Let’s investigate!
* **Medical\_history\_1** has over **7%** correlation. Maybe the cost of taking care of this particular kind of medical history is generally expensive in Singapore. One reason could be that the medication to treat the same is imported from a monopoly or a patented manufacturer.
* **Symptom\_3** with **6%**.
* **Symptom\_2, weight, medical\_history\_6,** and **bmi** with close to **5%** each.

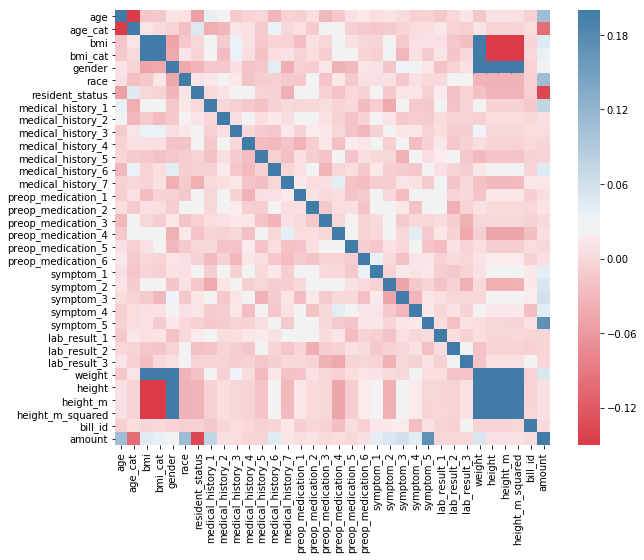
****

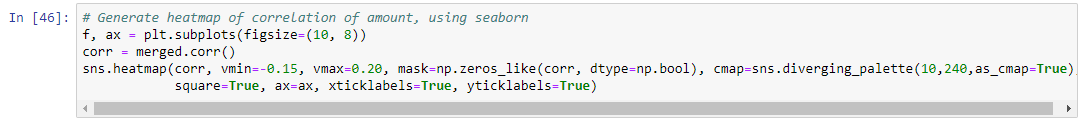
* **Resident status** has a **negative** correlation of **-14%** with amount. I suspect healthcare is cheaper for locals in Singapore, as compared to that for foreigners. It is obvious that the government would subsidize its own country’s citizens’ healthcare. And not those of other countries.
* We will **not** consider **age\_cat**, as we are already taking the numerical column, age, having a positive correlation into consideration.



## 3.02 **Visualize** the correlation using **heatmap**

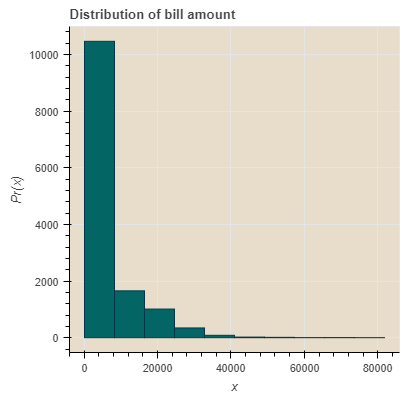
* We created this heatmap using seaborn.
* We have set the minimum correlation to **-0.15 (-15%)** and maximum to **0.20 (20%)**. This is done because the next highest correlation is 1.00 (100%) ie. the variable itself.
* A variable has a correlation of 1 with itself, which is irrelevant for analysis.
* Negative values are labelled in **red**, and positive correlations are labelled in **green**.
* You may refer to the exact code in the IPython notebook enclosed.



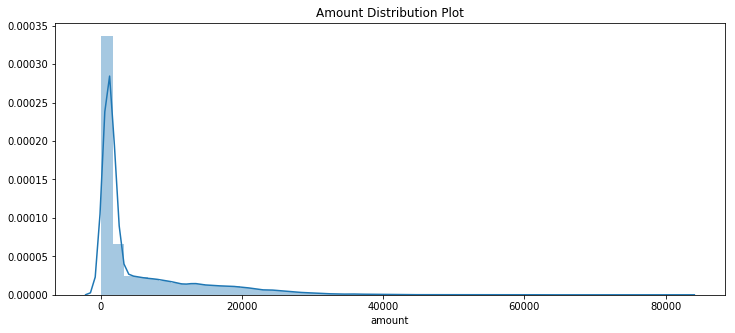


* As we can observe, the **darkest** of shades (meaning the correlation is greater), are observed on (except for amount itself, and other discarded and/or irrelevant variables like age\_cat), **symptom\_5** (**17%**), **medical\_history\_1** (**7.5%**), **age** (**11%**), **race** (**11%**), **weight** (**5.2%**), **resident\_status** (-14%) are among the variables that stand out.
* We will investigate all of these.

## Generate a **distribution table** for **amount**

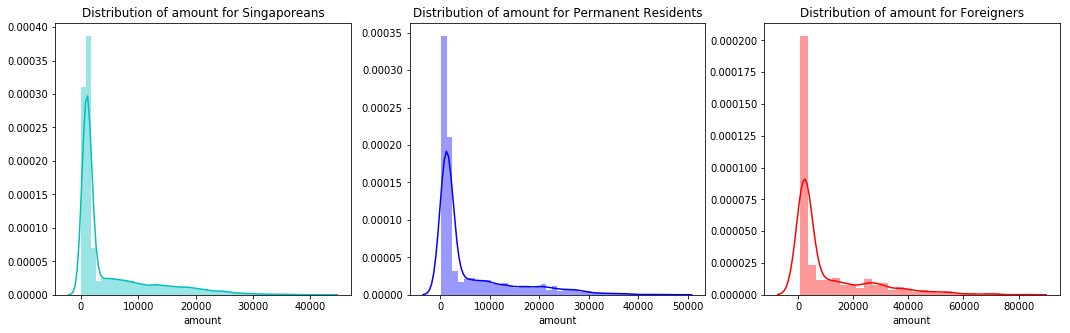


* The majority of the bills amount to below $10,000 each.
* Distribution plot made using Bokeh

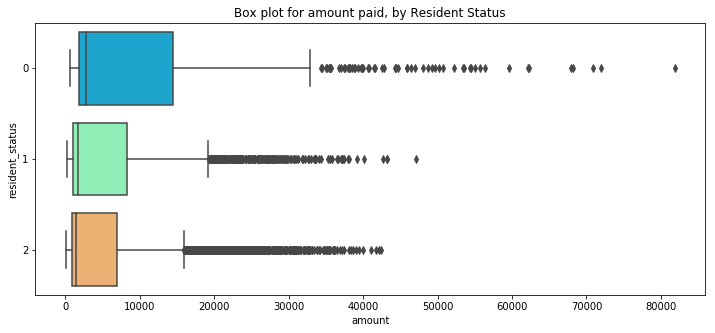


* Distribution plot made using Seaborn

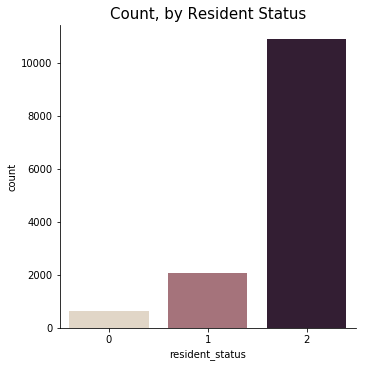
## **Are foreigners paying more than locals?**



* Indeed, the data confirms that foreigners pay far more than locals- both Singaporeans and Permanent Residents.
* This is probably because the government of Singapore may be subsidizing the cost of healthcare for its citizens and permanent residents.
* Let’s see exactly where each category of inhabitants stands in terms of shelling out for healthcare in Singapore, with the help of a boxplot:



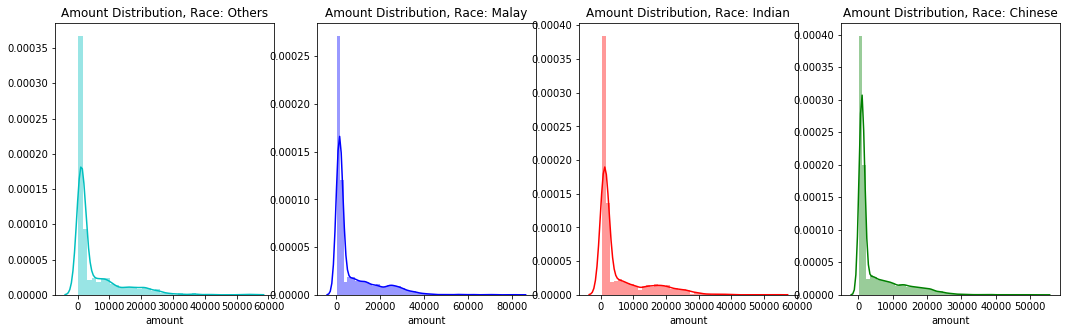
* ***Legend: 0 = Foreigners 1 = PRs 2 = Singaporeans***
* It is clear that the baseline spend (minimum bill amount) for **Foreigners** is nearly twice that of **PR**s.
* The baseline spend (minimum bill amount) for **Foreigners** is more than twice that of **Singaporeans**.
* Even as far as outliers are concerned, **Foreigners** lead the way, as can be seen above.

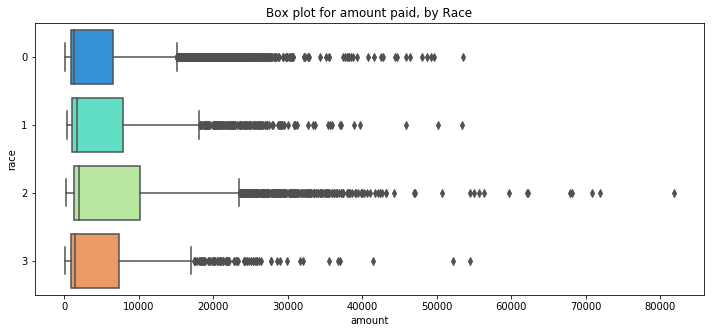


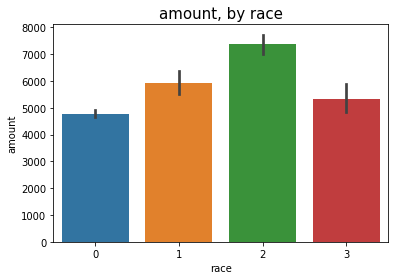
* Having said that, in our dataset, there are far fewer foreigner patients than Singapore Citizens and PRs. So, in purely cumulative totality terms, perhaps the impact of the premium imposed on foreigners isn’t quite as overbearing as it initially seemed to be.
* However still, **Resident Status** is a very crucial driver behind the bill amount.



## **Is there any disparity between bill amount and patient’s race?**

****

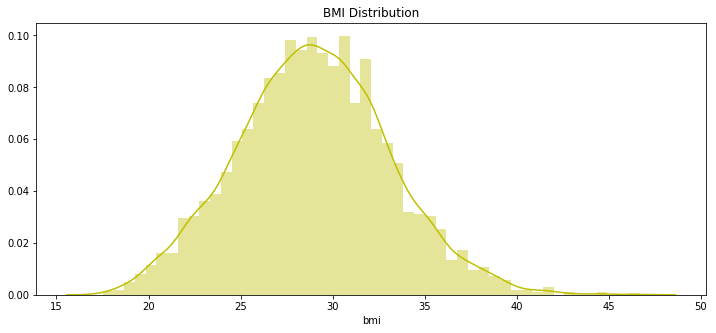
****

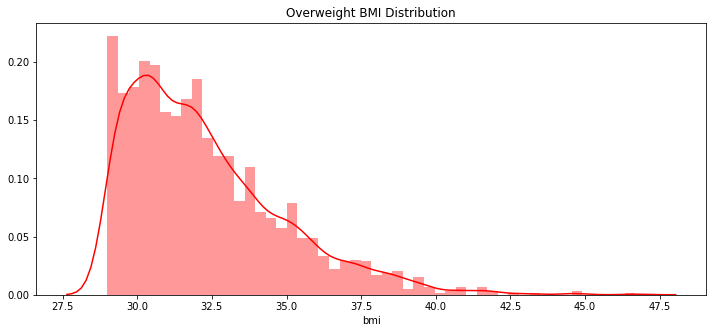


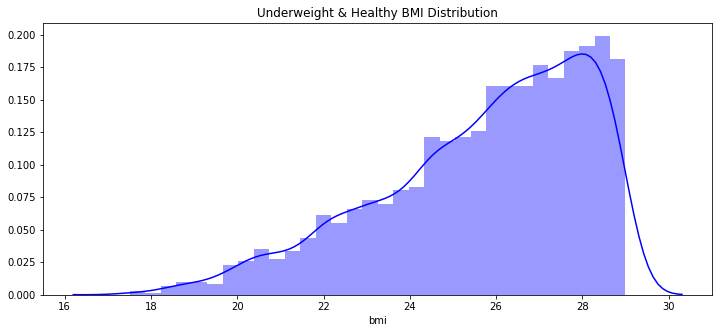
* **Label:**
  + **0 = Chinese**
  + **1 = Indian**
  + **2 = Malay**
  + **3 = Others**
* There is a clear disparity observed here.
* **The amount for Malay patients is far higher than all the other races.**
* The reasons for the same could be various- genetic proneness to certain medical conditions, being overweight, diet habits etc.

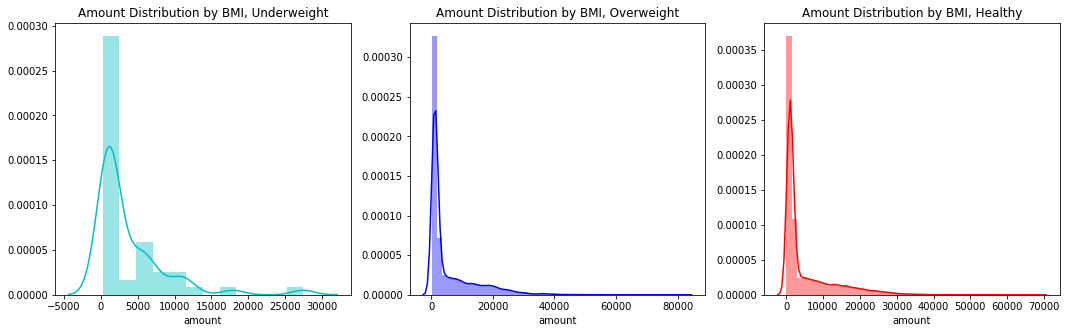
****

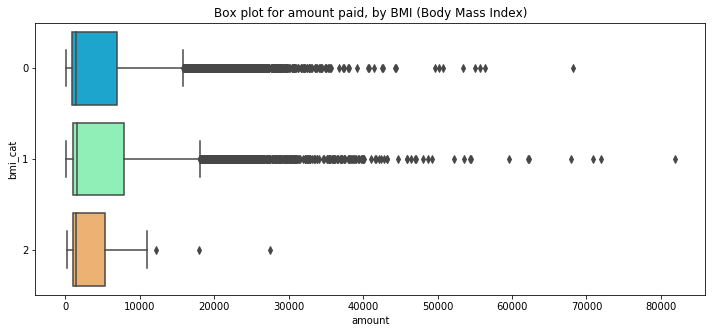
## **Is the amount higher for overweight patients?**

****

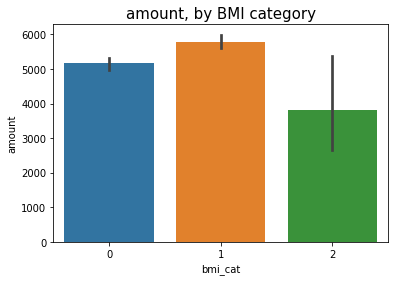
****

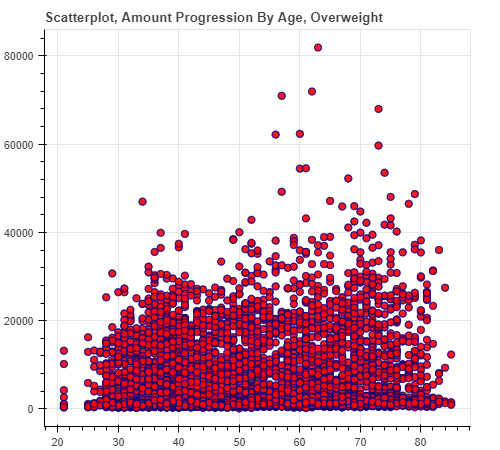
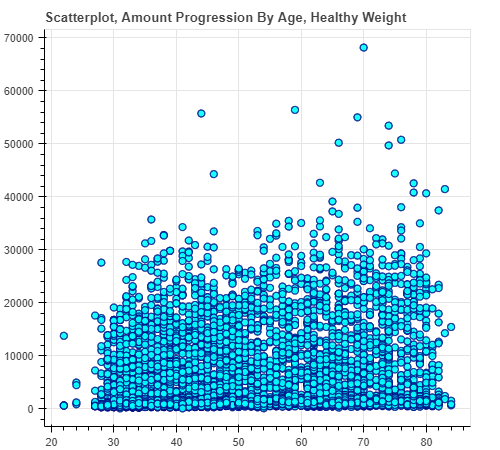
****

****

****

* **Label**
  + **Underweight BMI = 2**
  + **Overweight BMI = 1**
  + **Healthy BMI = 0**
* It can be seen from the above figures that yes, in fact, overweight patients do have to pay a higher amount. Let’s see by how much:

****

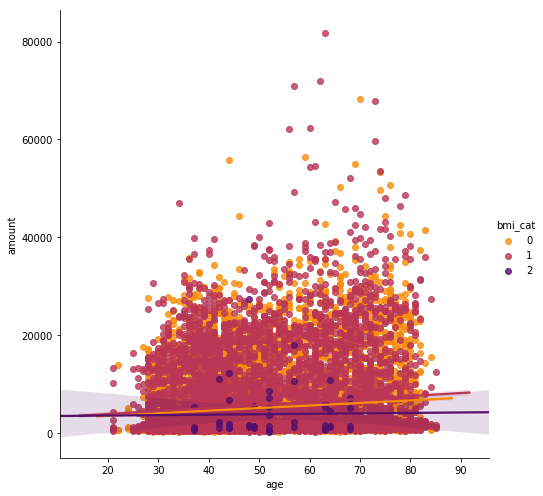
****

**Scatterplot of Amount Progression Scatterplot of Amount Progression**

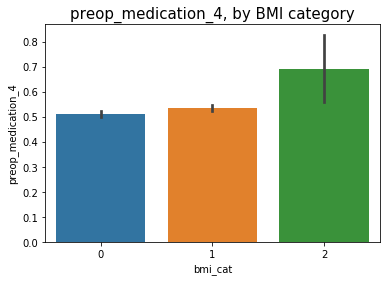
**By Age, If Patient Stays Healthy By Age, If Patient Stays Overweight**

**X = Age | Y = Amount**

**Scatterplot, Amount Progression By Age, All BMI Categories**

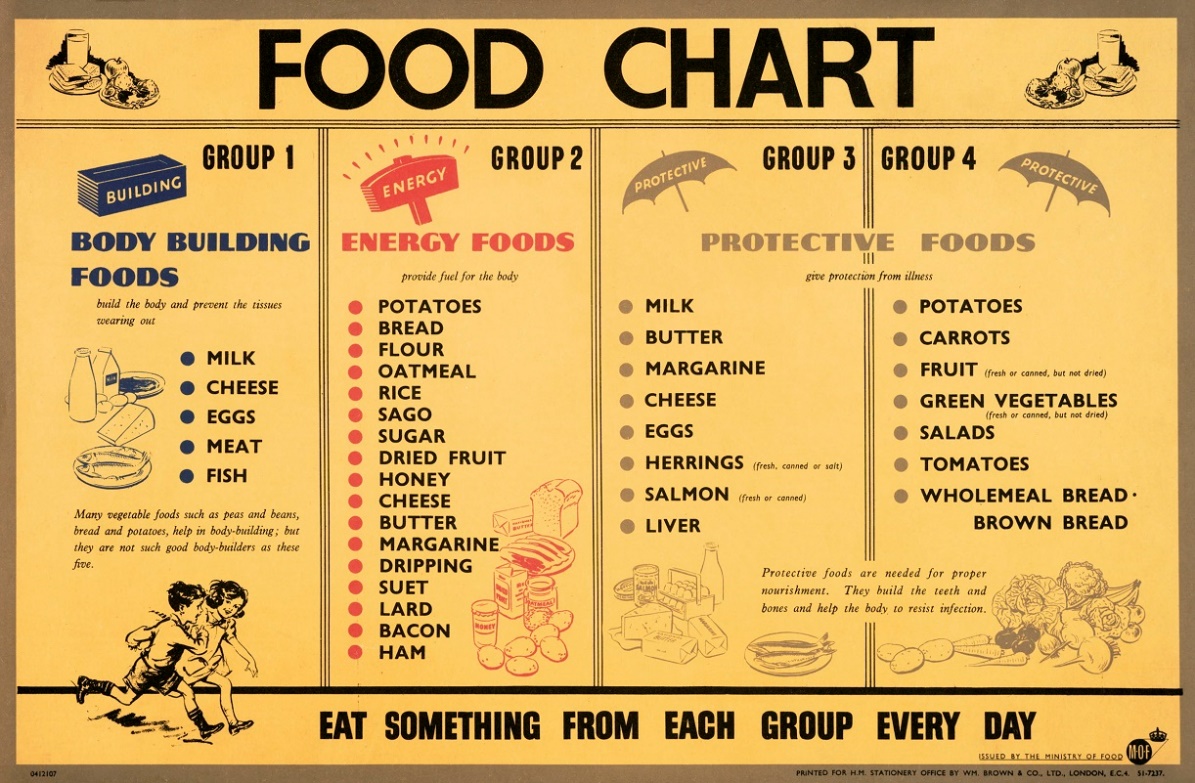
****

**0 = Healthy Weight | 1 = Overweight | 2 = Underweight**

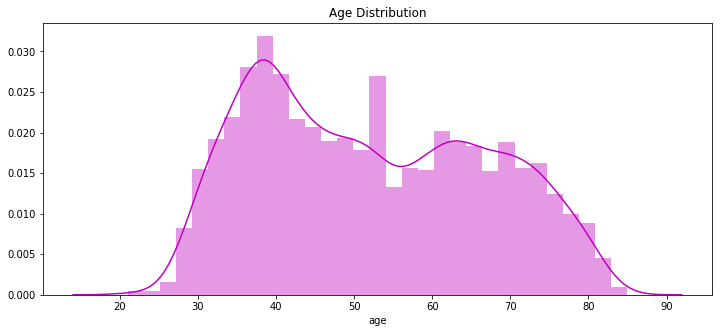
****

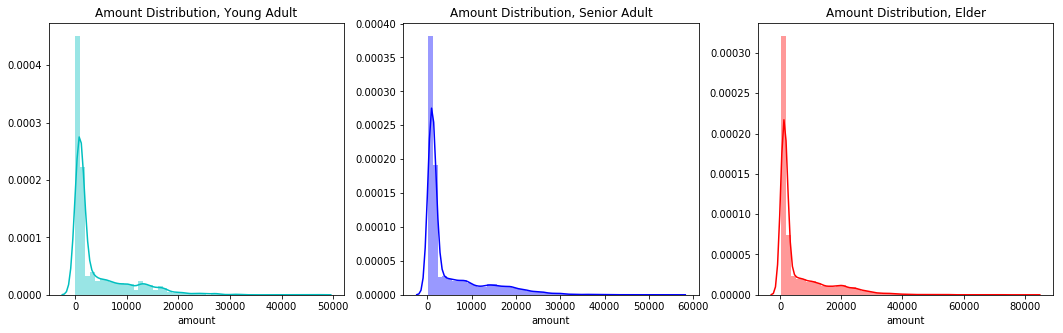
**0 = Healthy Weight | 1 = Overweight | 2 = Underweight**

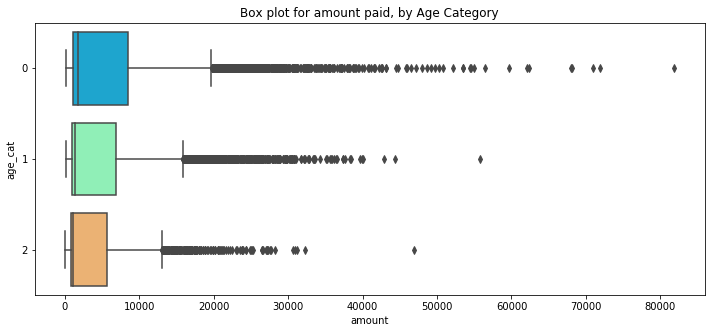
* The amount paid by overweight patients is absolutely a notch higher than that paid by healthy ones.
* We have established that healthcare costs rise with age, but we now have an insight that costs rise even higher with age if the patient is overweight. Refer to both lines moving upwards. The one with bmi\_cat = 1 (red) has the highest upward moving slope.
* It must be noted that observations for underweight are very few, hence this is not taken into consideration.
* Speaking of the underweight division, a very intriguing insight popped up while analysing the data. Interestingly, preop\_medication\_4 is most prevalent among the underweight. Its relationship barely varies among the healthy and the overweight.
* Perhaps, preop\_medication\_4 is a type of a weight booster given to underweight patients, so that their weight can be elevated to a point which would allow for a smoother operation. An example of this is liquid glucose which is pumped into a patient’s body using tubes.

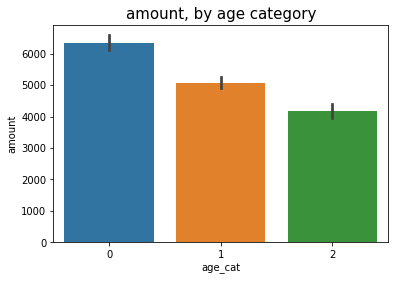


## **Do healthcare costs increase with age?**







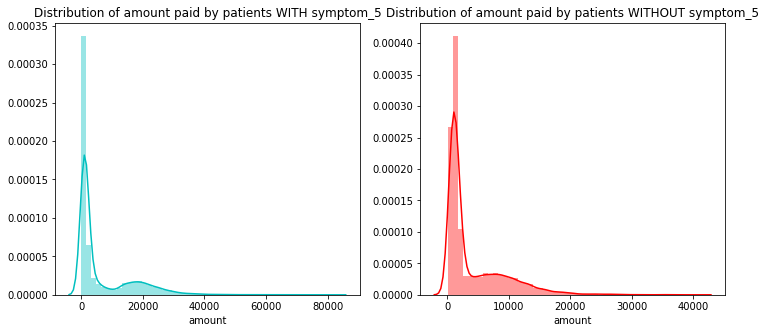
****

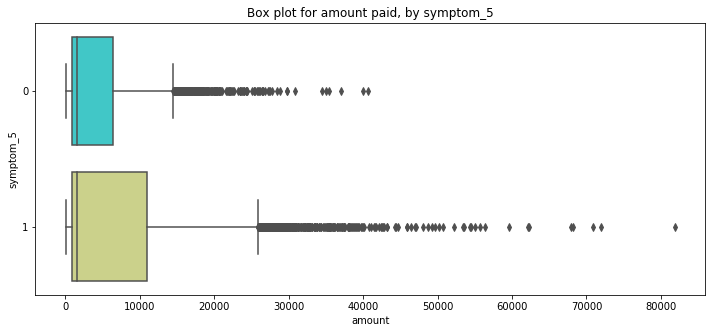
* **Label**
  + **Young Adult (between 18, until 35, included) = 2**
  + **Senior Adult (over 35, until 55, included) = 1**
  + **Elder (over 55) = 0**
* **It’s true!**
* **Healthcare costs rise significantly as a person gets older.**

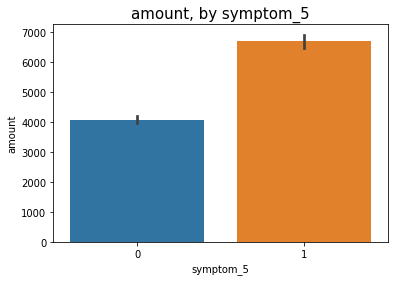
****

## **What’s the deal with symptom\_5?**

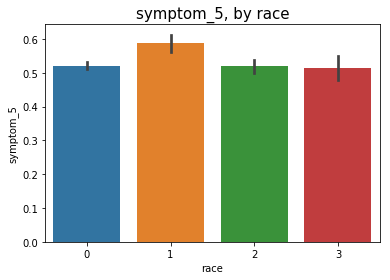
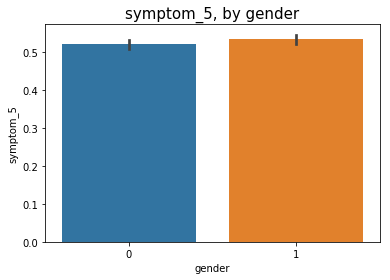
* Symptom\_5 has a **17%** positive correlation with amount.
* Let’s take a look further:



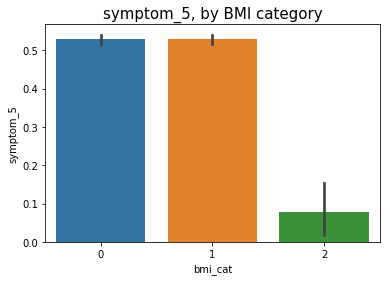
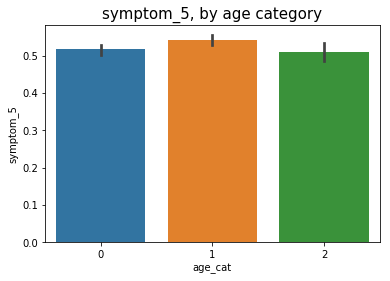




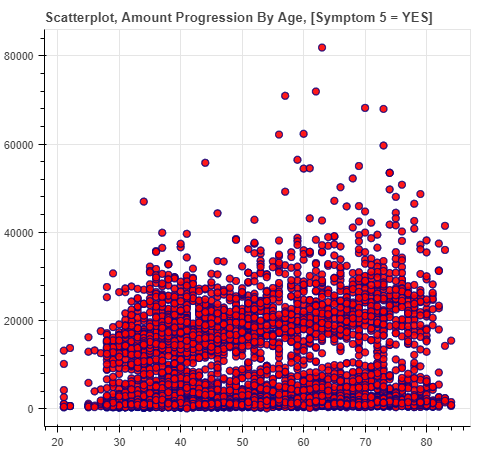
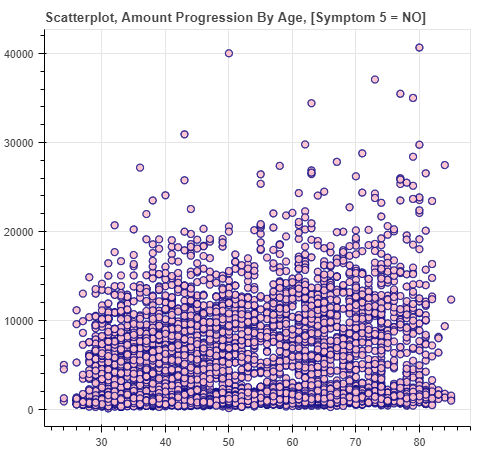
* By this point, it is clear that patients showing symptom\_5 pay a much higher price than those who don’t have that particular ailment.
* One possible explanation is that it could be something as common as fever, or headache, which are prevalent in a large variety of illnesses and medical conditions.
* Another explanation is that it is something very expensive to treat, at least in Singapore. Some examples of such expensive treatments would be chemotherapy, and plastic surgery.
* Since we have no discernable information about medical\_history or symptoms, it is pointless to speculate on what the significance of the symptom is. Please note that the above theories are purely speculative in nature and are not backed by any kind of data whatsoever.
* What we as analysts can do, however, is correlate this symptom with other accessible variables- like weight, age, gender etc. Let’s get started on that!



Label- 0 = Female, 1 = Male 0 = Chinese, 1 = Indian, 2 = Malay, 3 = Others

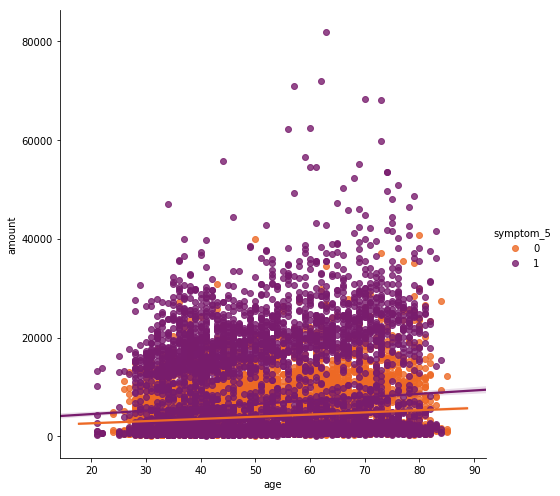


Label- 0 = Elder, 1 = Senior, 2 = Young Label- 0 = Healthy, 1 = Overweight, 2 = Underweight



**X = Age | Y = Amount X = Age | Y = Amount**

**Scatterplot, Amount Progression By Age, [Symptom 5 = YES and NO]**

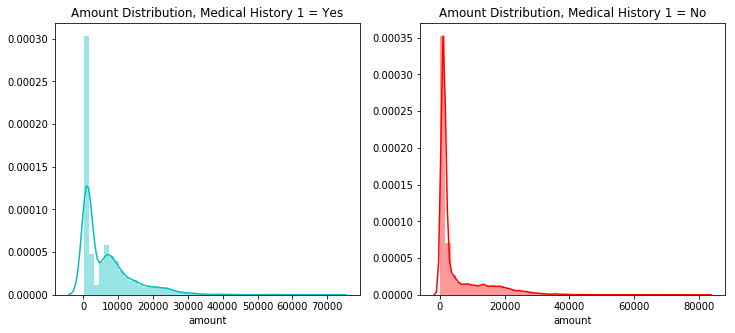


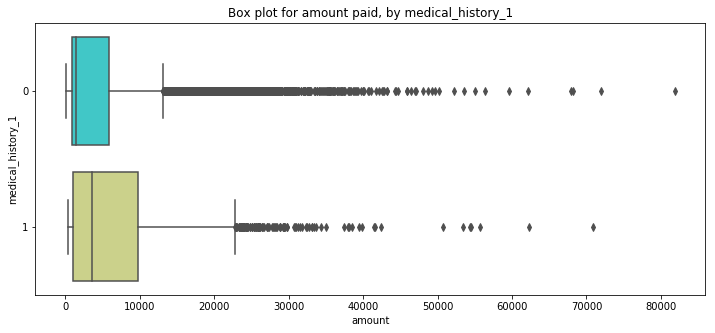
* So far, we have gathered the following information about symptom\_5
  + **It affects both genders equally.**
  + **It is relatively higher in Indian patients, than those of other races.**
  + **It is marginally higher in Senior Adults (35 to 55).**
  + **We have established that healthcare costs rise with age, but we now have an insight that costs rise even higher with age if the patient has symptom 5. Refer to both lines moving upwards. The one with symptom\_5 = 1 (purple) has a higher upward moving slope.**
  + **Could it be a symptom of a stress-related illness like clinical depression? It’s possible.**

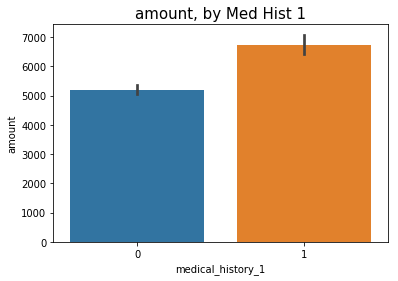
****

## **Will history (medical\_history\_1) repeat itself?**

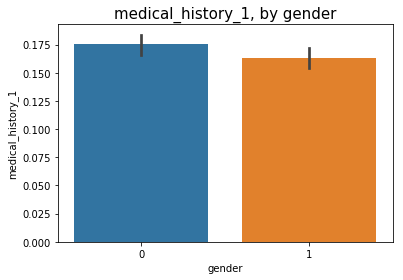
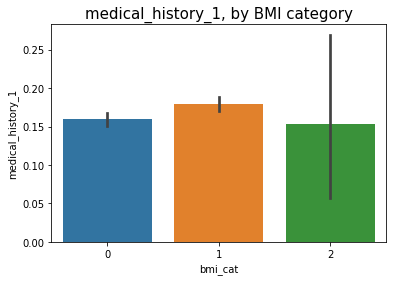
* Medical\_history\_1 has a correlation of approximately **7.5%** with amount.
* Let us investigate:



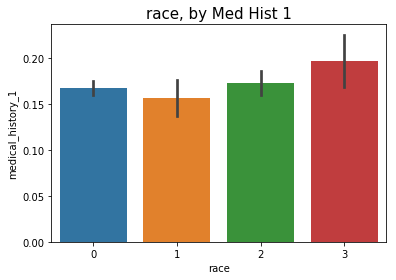
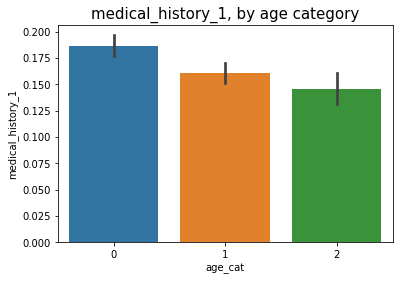




0 = No | 1 = Yes



0 = Healthy | 1 = Over | 2 = Under 0 = Female | 1 = Male

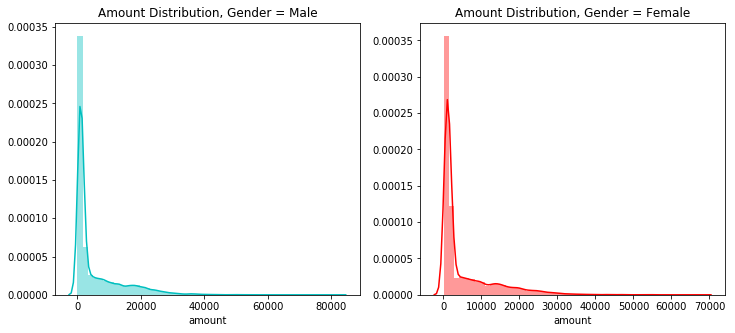


0 = Elder | 1 = Senior | 2 = Young 0 = Chinese | 1 = Indian | 2 = Malay | 3 = Others

* Amount is much higher, almost 40%, for patients with medical\_history\_1.
* This medical history has a direct positive correlation with the patient being overweight.
* This medical history has a direct positive correlation with the patient’s age. It is most prevalent in the eldest age bin, and least in the youngest.
* In our dataset, among race, this history is most prevalent amongst ‘Other’ races apart from the 3 most densely populated ones. Although this shouldn’t be relied upon because of the low number of observations for this race category.
* It is least popular amongst the Indian race.
* It is relatively more prevalent amongst female patients.
* I am by no means a doctor! But it is very possible that this history is related to heart condition. As it is positively correlated with being overweight, increases in likelihood with age, and we all know how expensive heart condition treatments can be. Again, just an innocent speculation. 😊

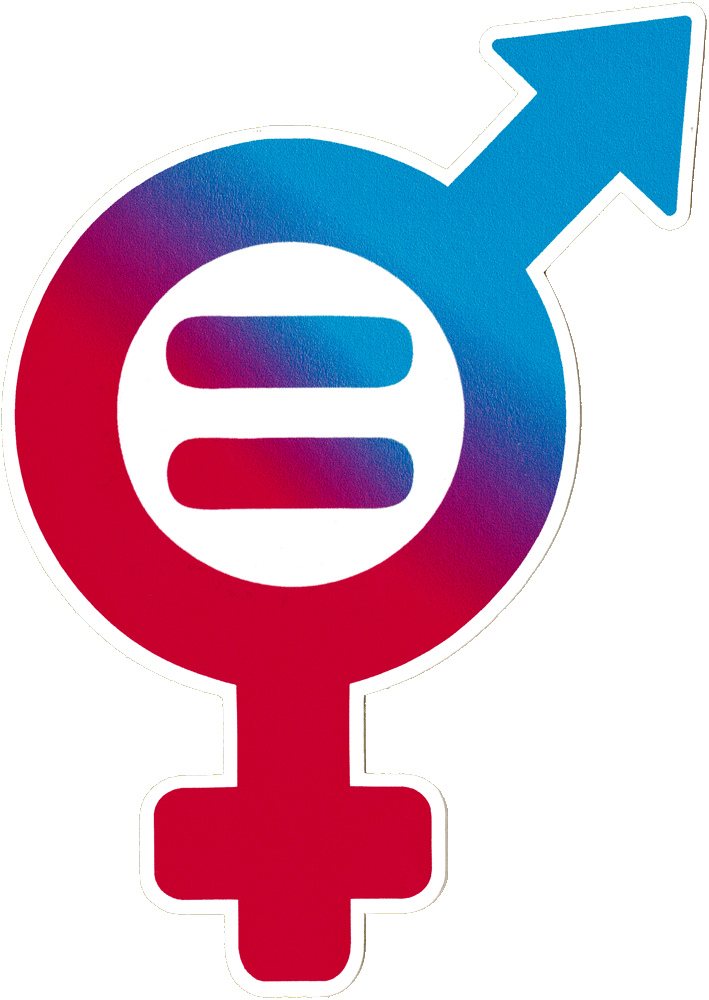


## **3.10 Is there disparity between bill amount for men and women?**



## 

* **Legend:**
  + **Male: 1**
  + **Female: 0**
* There is a very subtle disparity between the amount paid by each sex.
* Having said that, males do pay a slightly higher bill amount than females.

. 

## **Is medical\_history\_3 = type 2 diabetes?**

## 

0 = Healthy 1 = Overweight 2 = Underweight

* It affects the overweight far more than it does other BMI categories.
* Type 2 Diabetes is known for being prevalent in overweight patients.

## **Is medical\_history\_6 a children’s disease (like chicken pox, measles etc.)?**

## 

0 = Healthy 1 = Senior 2 = Young

* It affects the young far more than it does other age categories.
* Diseases like chicken pox, measles and rashes are known for being prevalent in children.

## **4.00 Concluding Notes**

* Challenges faced during data pre-processing:
  + Disjointed datasets
  + Tedious parsing of date-time objects
  + Cumbersome derivation of patient’s age from two separate date-time objects
  + Generating BMI column
  + Manual conversion of categorical data
  + Transformation of dirty data within multiple categorical variables
  + Missing values in two columns
  + Encoding of categorical variables to allow for correlation analysis
* Let’s summarize our analysis of the drivers of cost, one by one:
  + **resident\_status** => high negative correlation of almost **-14%** | foreigners pay the highest amount | citizens pay the lowest | reason behind is straightforward, government policies
  + **symptom\_5** => high positive correlation of almost **17%** | patients who have this symptom pay almost 75% more than those who don’t | reasons behind are ambiguous, speculative at best
  + **age** => high positive correlation of **10.70%** | as one gets older, healthcare costs increase | interestingly there are various combinations, such as being overweight, or having symptom\_5 persistently, that can raise the bill amount even higher over time | check relevant scatterplots
  + **bmi** => relatively high positive correlation of **4.60%** | the more overweight the person, the higher the amount | teams up well with other factors contributing to higher amount
  + **race** => high positive correlation of almost **11%** | astonishingly, malay patients pay around 17% higher than Indian patients, and a whopping 46% higher than chinese patients.
  + **medical\_history\_1** => high positive correlation of almost **7.50%** | patients with this history pay around 40% more | direct positive correlation with being overweight, age | more popular among females | relatively lower frequency among Indians | possible heart condition indicator
  + **gender** => low positive correlation of **1.9%** | minor disparity between male / female | males pay slightly more
  + **medical\_history\_3** => there is a high chance this is associated with type 2 diabetes | it affects the overweight far more than it does other BMI categories
  + **medical\_history\_6** => there is a high chance this is associated with a children’s disease (like chicken pox, measles, rashes etc.) | it affects the young far more than it does other age categories
* Last but not least, thank you Sankha for shortlisting me for the Data Science position at Holmusk! 😊