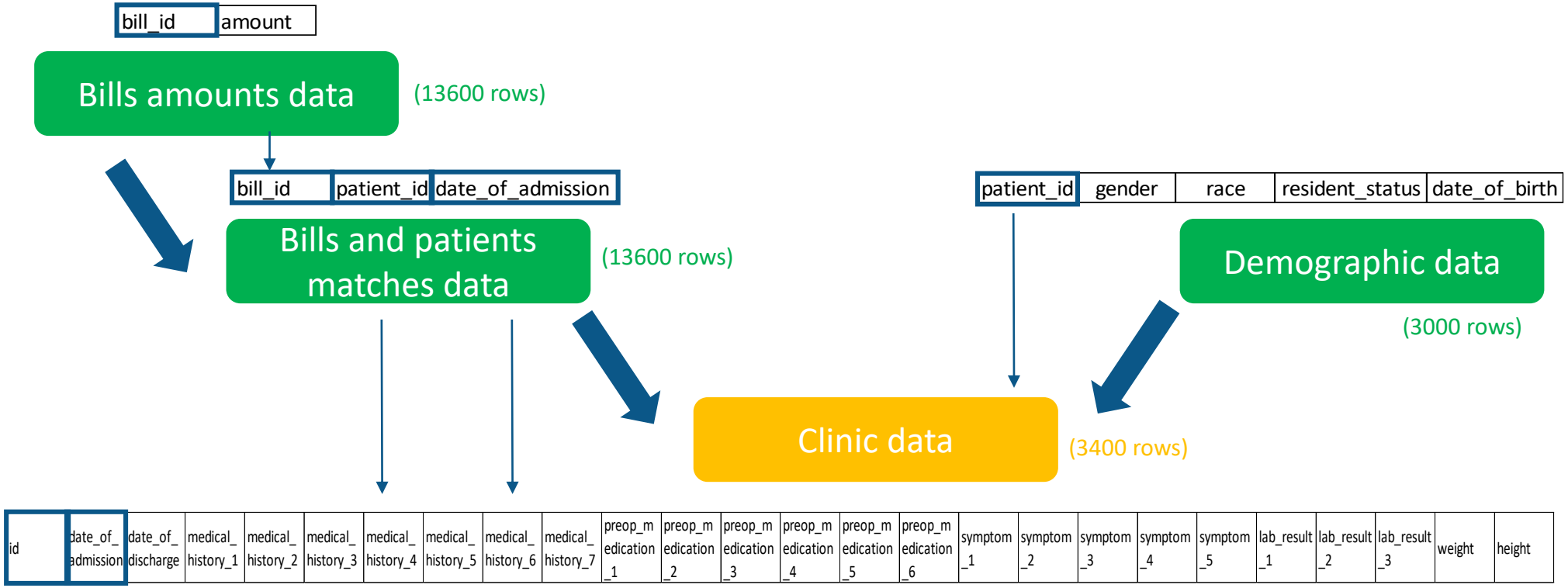
Holmusk Data Science Test

By Pavel Troshenkov

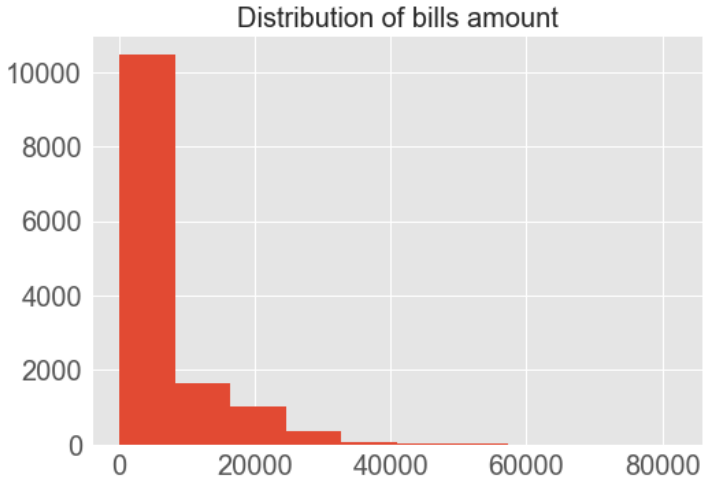
We have data that comes from 4 sources: bills and patients matches (i.e. transactions data), bills amounts data, clinic data with the description of care of patients and demographic data describing who patients are



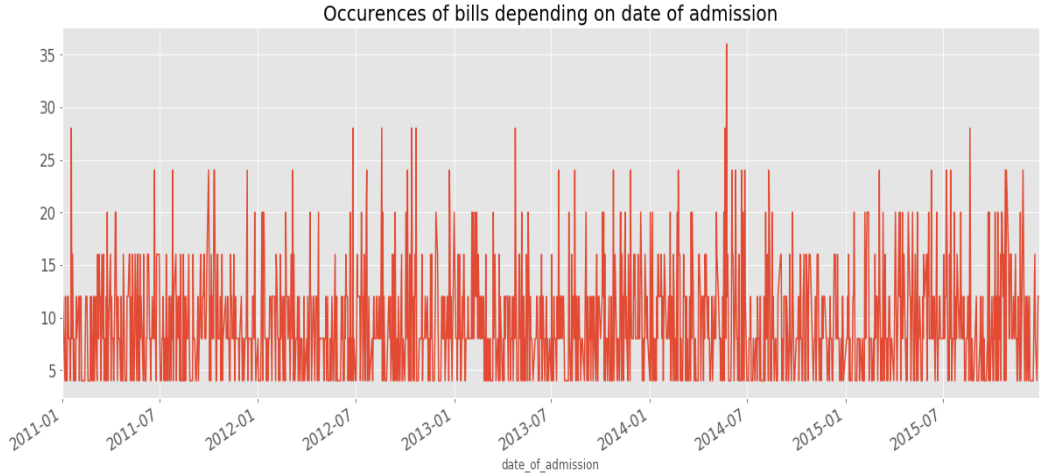
- **We have clinic data for 3400 cases of medical treatment and only 3000 of them are unique cases**
- **For each case we have 4 different bills which means we need to aggregate bills data**

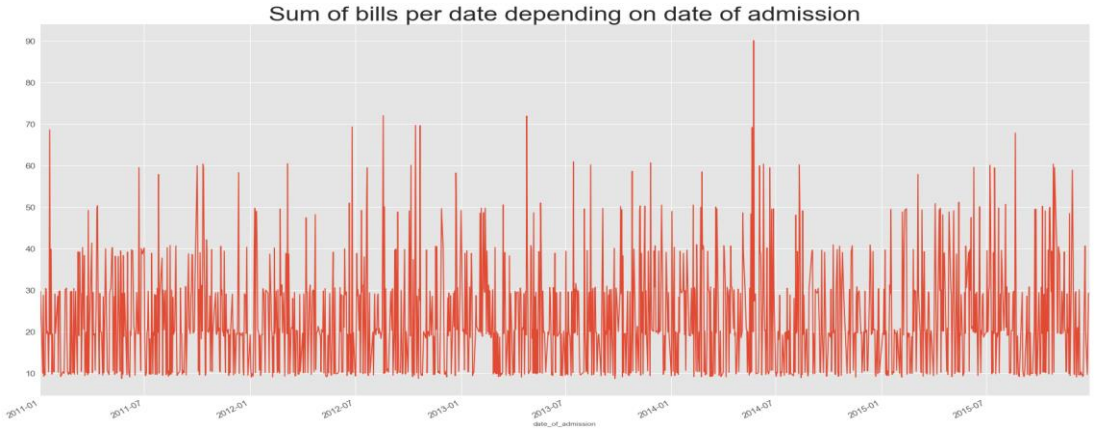
Bills data overview
Part 1/2

- **Most of the unique bills are small. So, a patient pays 3 little bills and 1 big one**



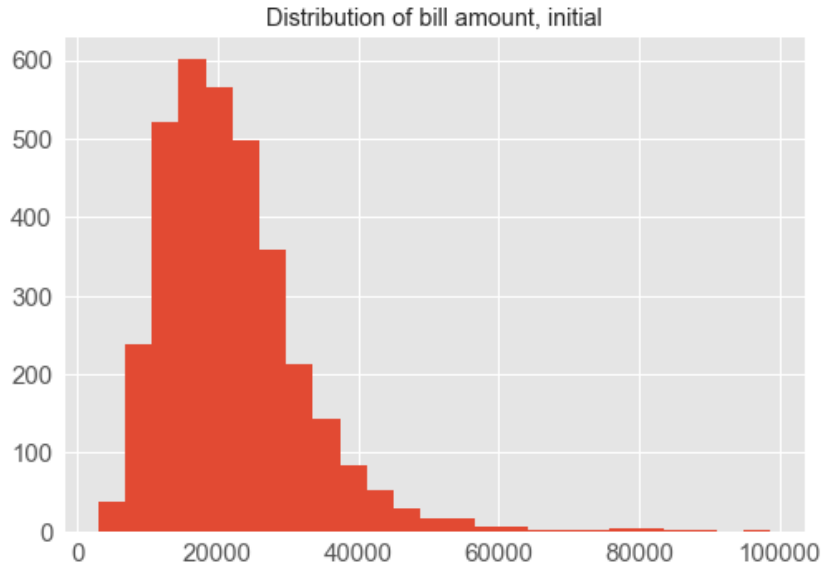
- **Bills come with similar regularity and amount of bills doesn't have a trend over dates**

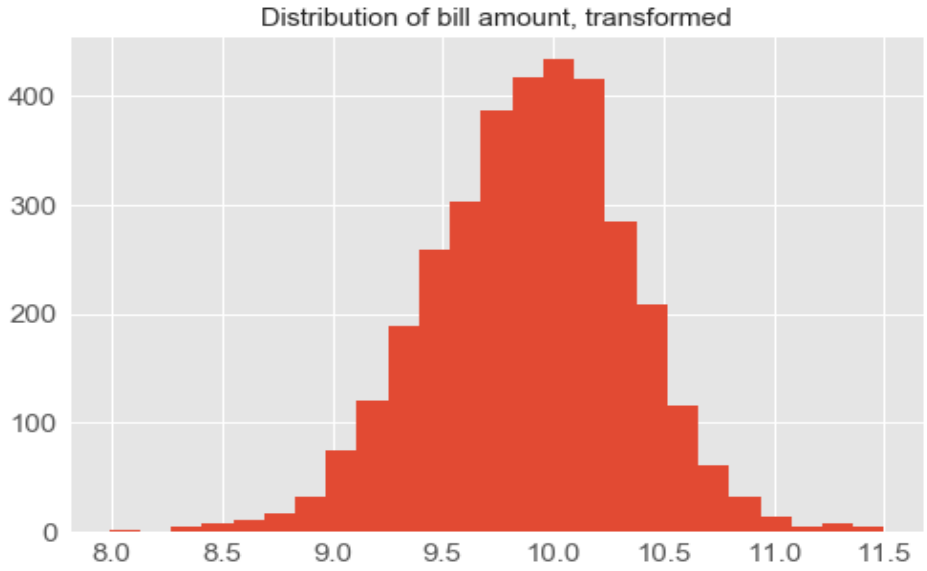




Bills data overview
Part 2/2

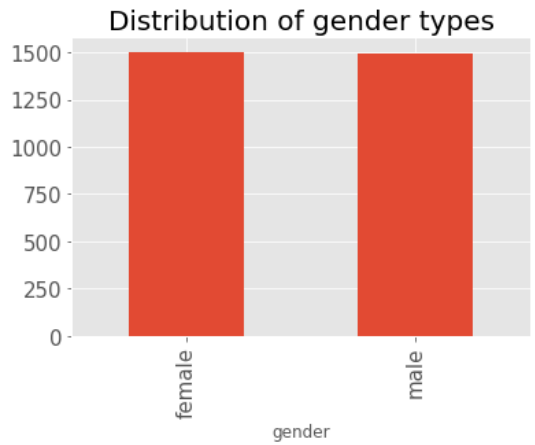
- **The distribution of sums of bills per client is right skewed. Applying log-transformation we can fit it to a normal distribution to be able to work with confidence intervals for predictors**



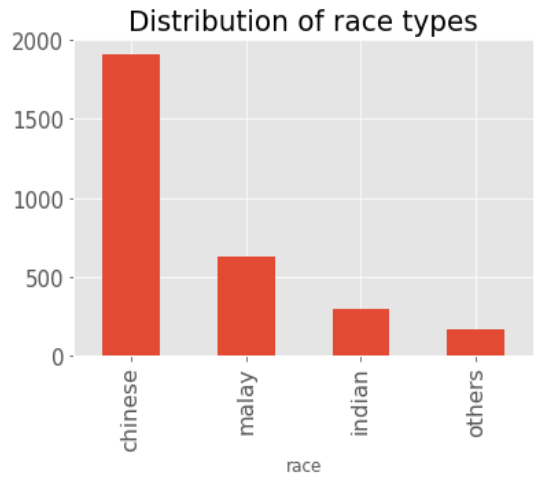


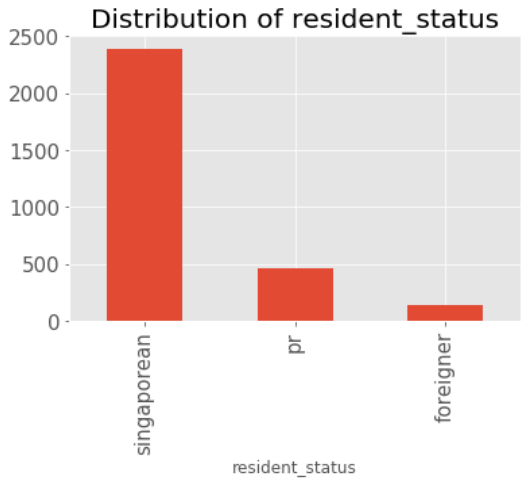
Demographic overview
Part 1/1

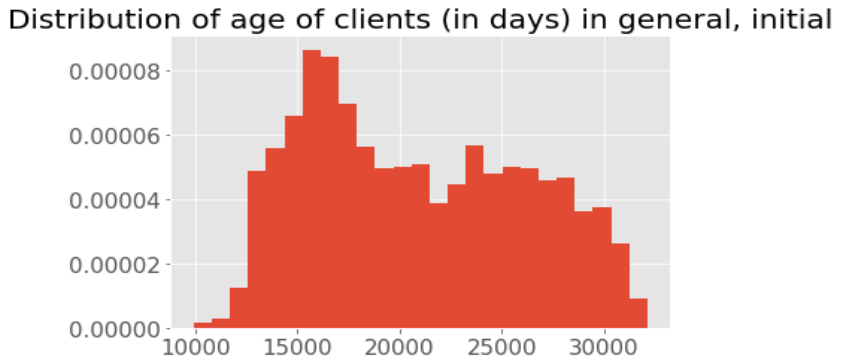
- **The proportion of male and female patients is equal**



- **The half of patients are Chinese with Singaporean resident status. Age distribution has two peaks**

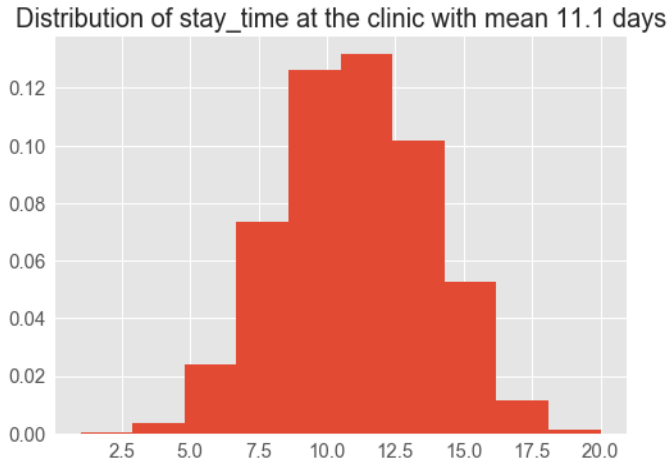


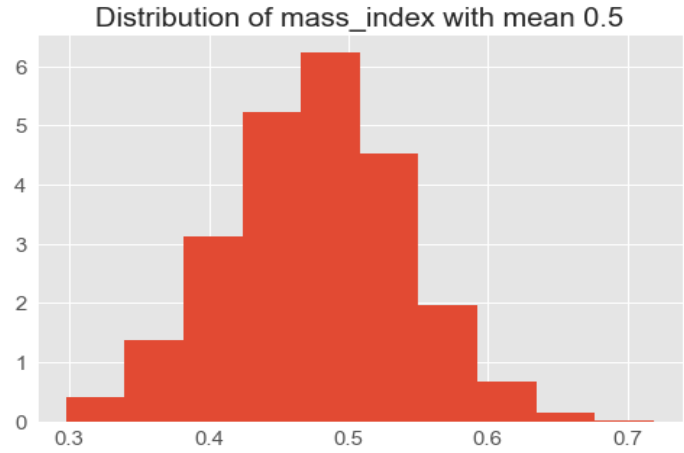


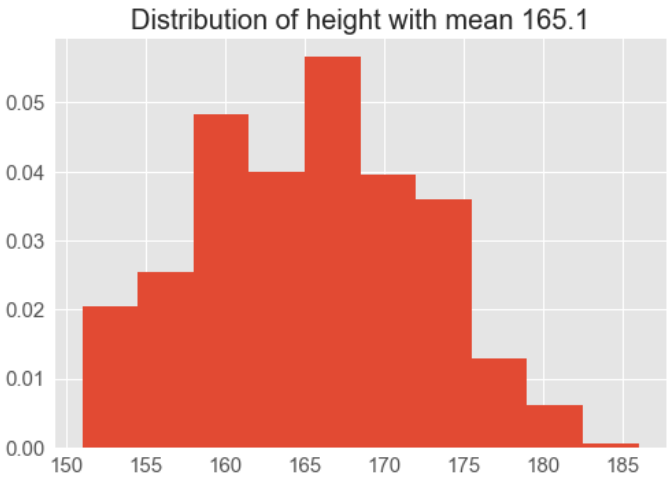


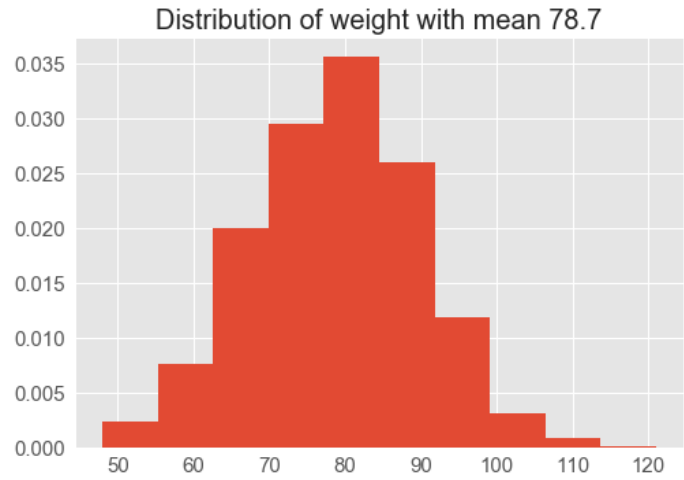
Clinic data overview
Part 1/2

- **Distributions of stay time at clinic (in days), weight, height, mass index (weight / height) look normal**



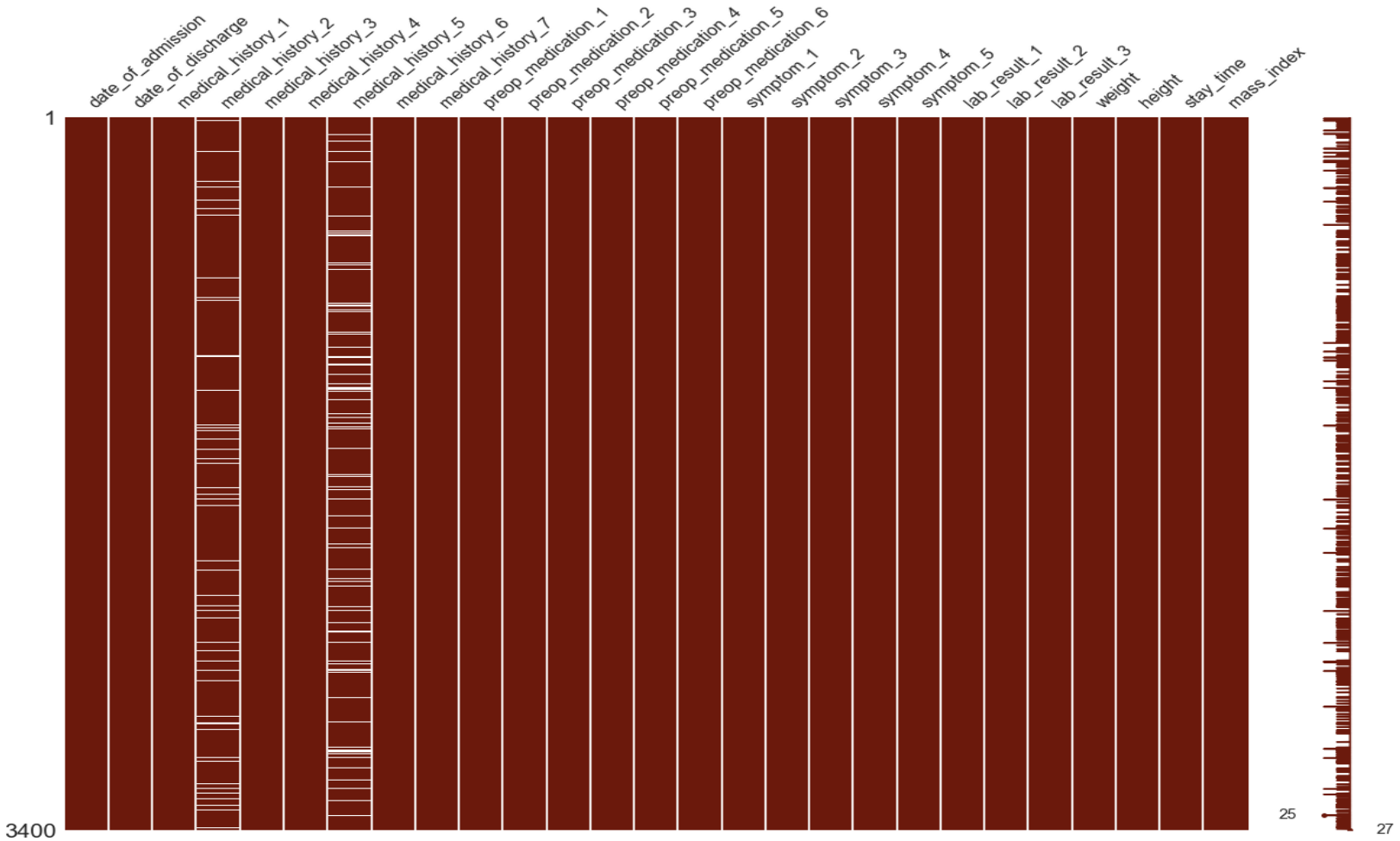






Clinic data overview
Part 2/2

- **Some data about medical histories is missing. Since data in such columns is binary we can fill it with 0 values**

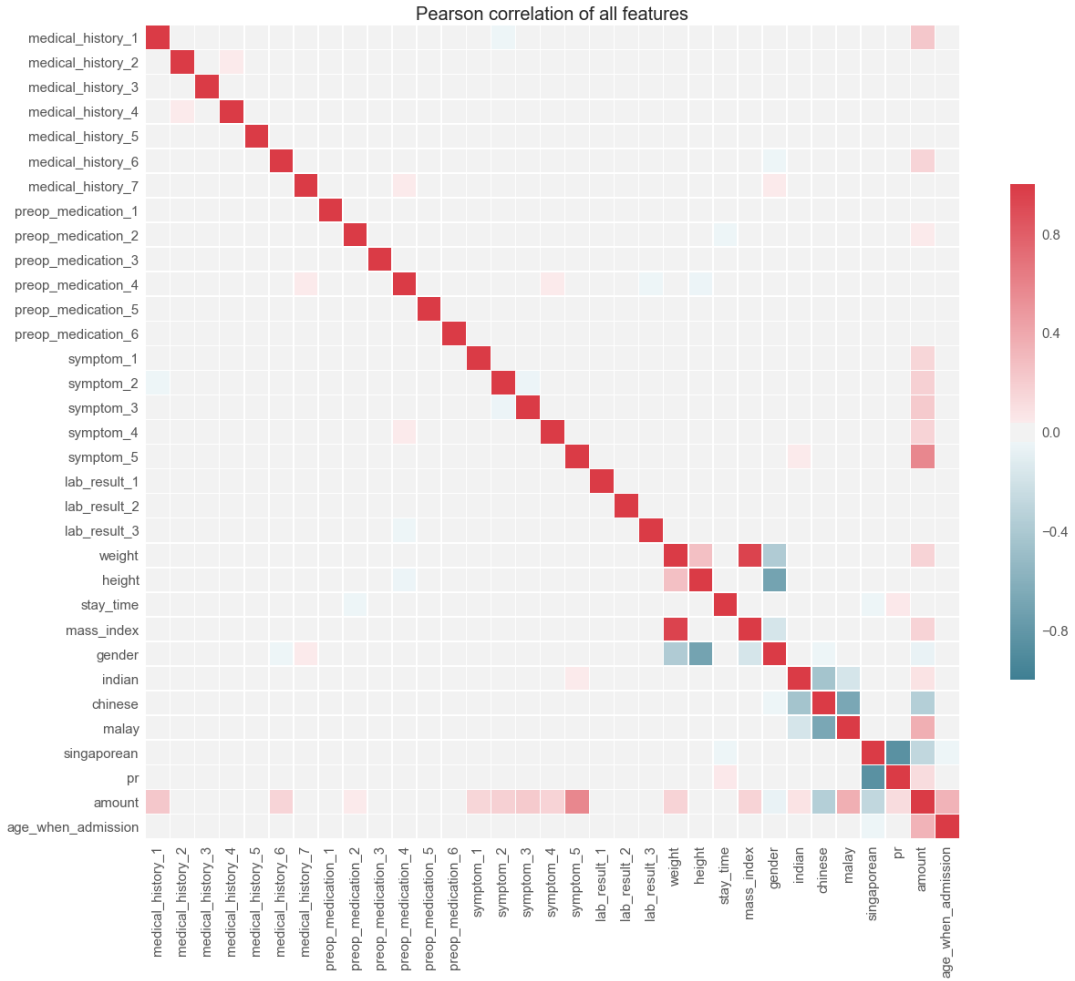


- In this work I will be looking for a linear dependency between **AMOUNT (sum of bills per client per clinical case, i.e. total cost of care)** and all other both binary and continuous predictors
- For the purpose of interpretation of regression coefficients, I will make log transformation of **AMOUNT** since it has skewed distribution

$$Log\ (Y) = a + bX$$

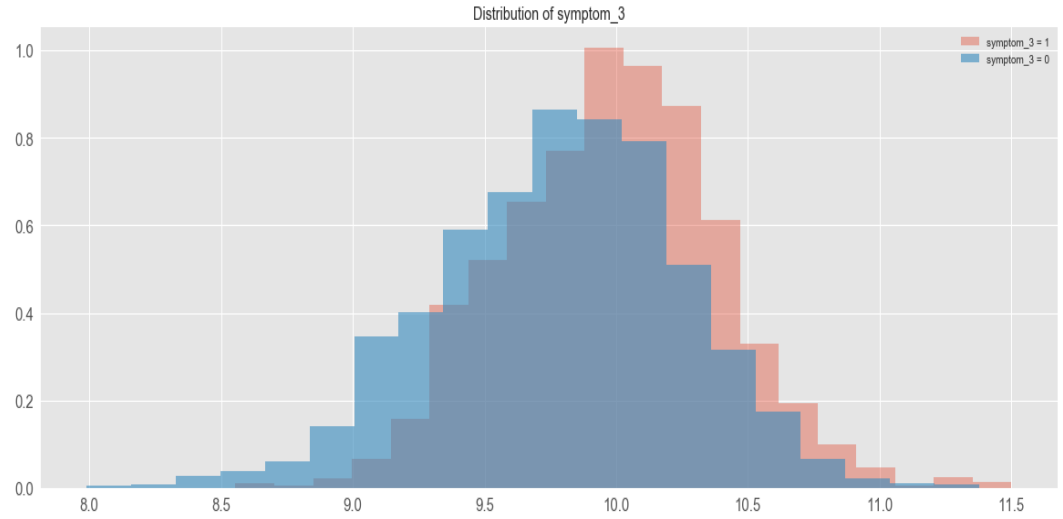
Aggregated data overview
Part 1/1

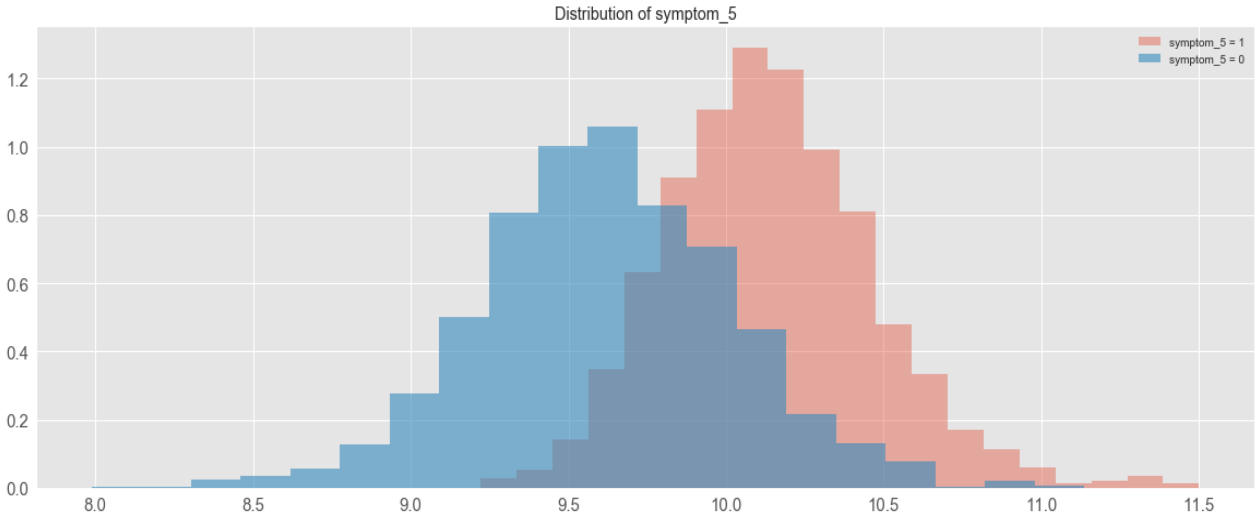
- **At the correlation heatmap we see correlation between amount and symptom_5, age, malay race, chinese race, medical_history_1**
- **We see correlation among some predictors. In next steps we should check VIF values to get rid of redundant features**

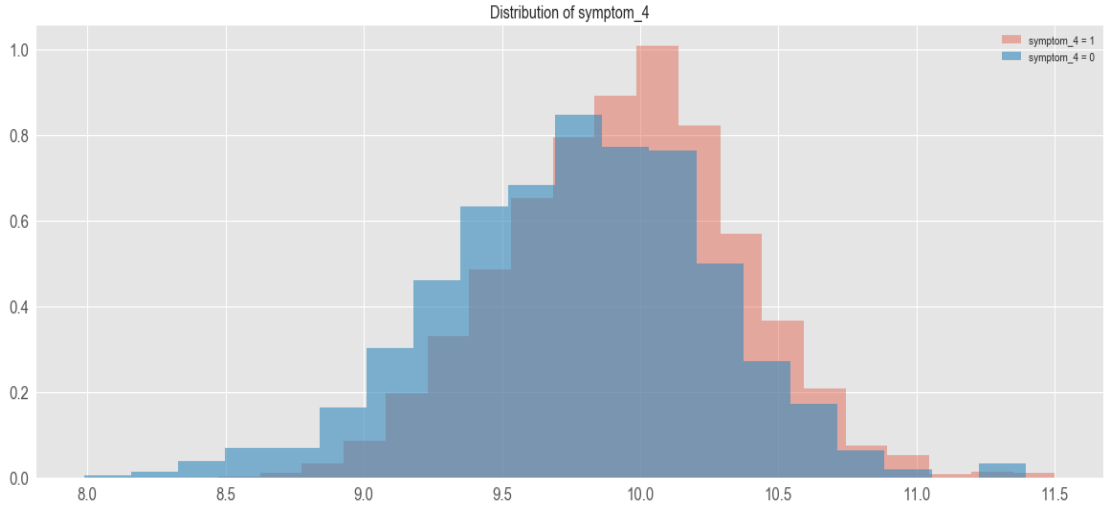


Aggregated data overview
Let's look at the predictors highly correlated with the target variable separately on graphs. Part 1/3

- **Symptoms**

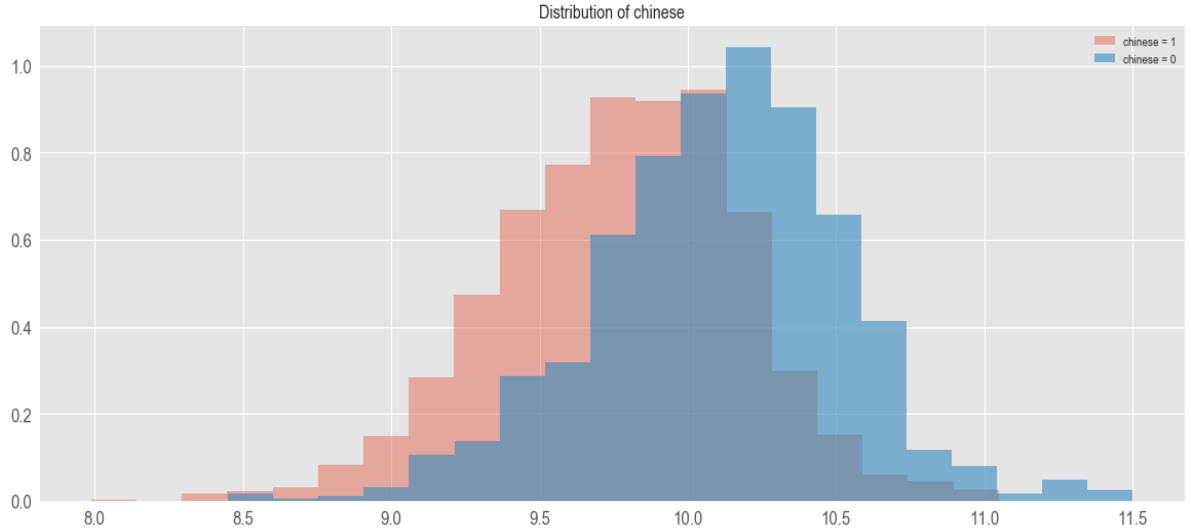


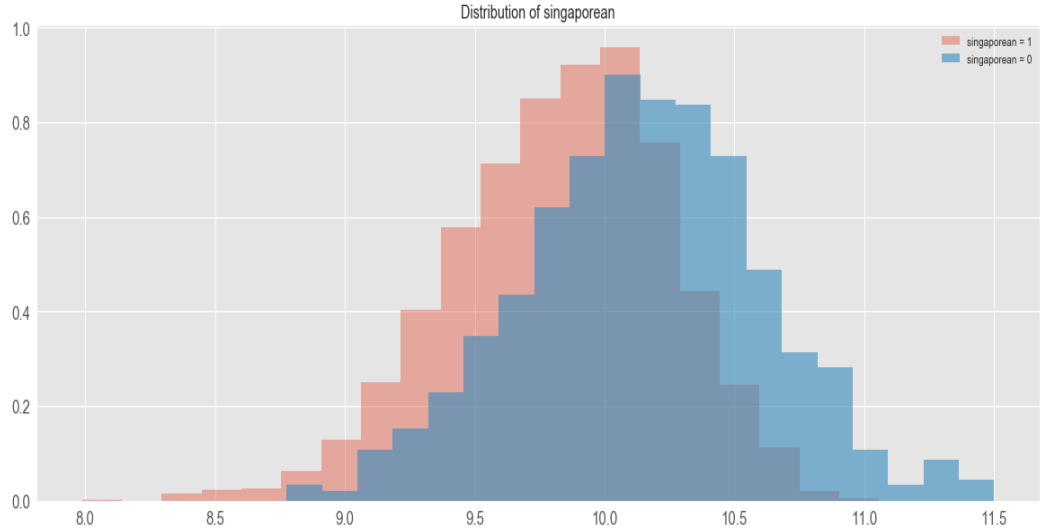


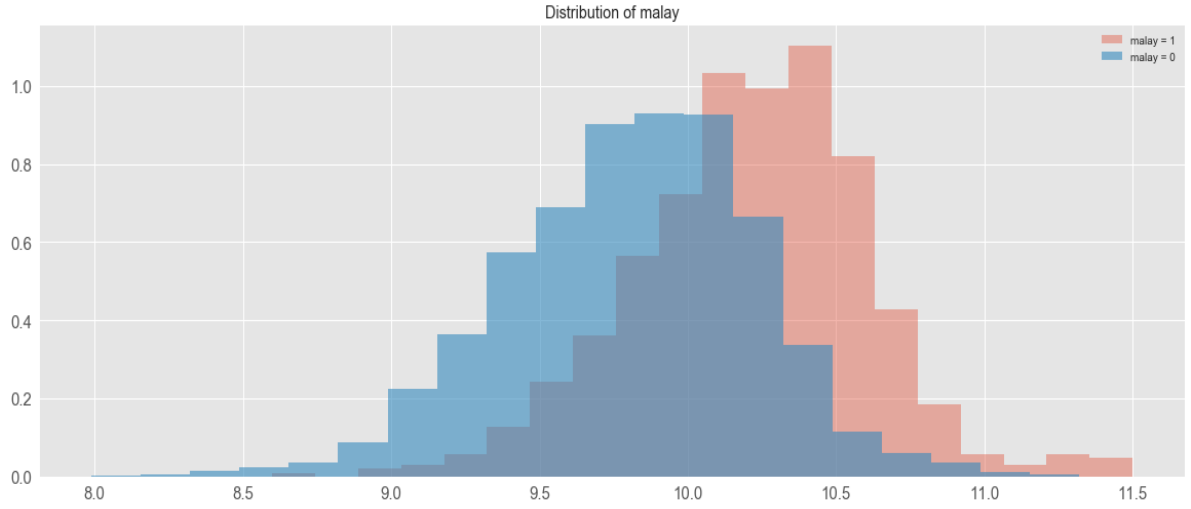


Aggregated data overview
Let's look at the predictors highly correlated with the target variable separately on graphs. Part 2/3

- **Races**

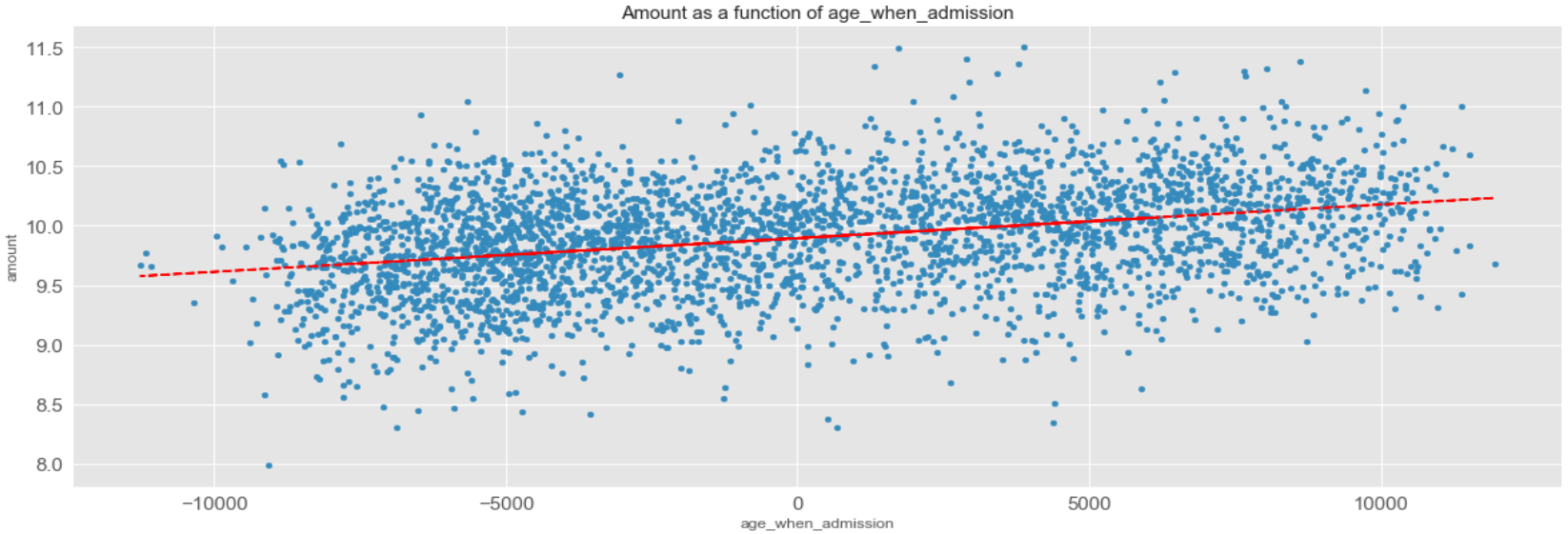




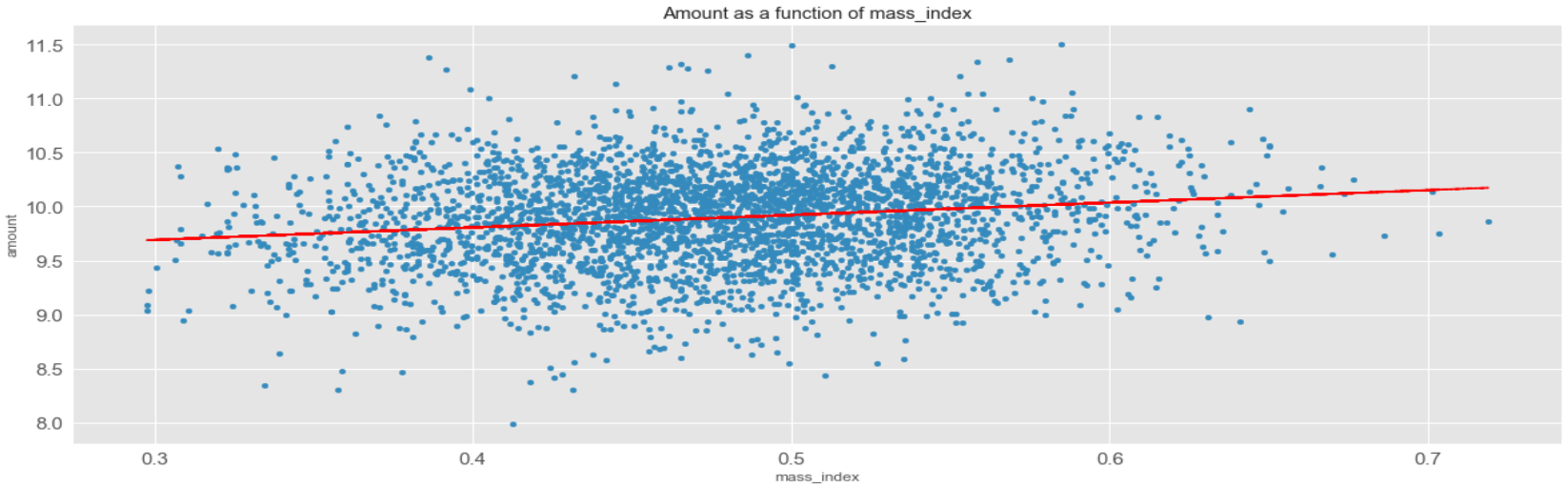


Aggregated data overview
Let's look at the predictors highly correlated with the target variable separately on graphs. Part 2/3

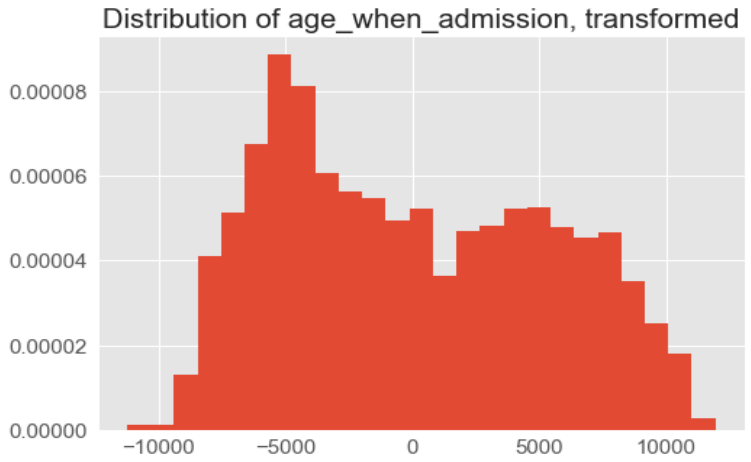
- **Age of patient (in days)**



- **Mass index (weight / height)**



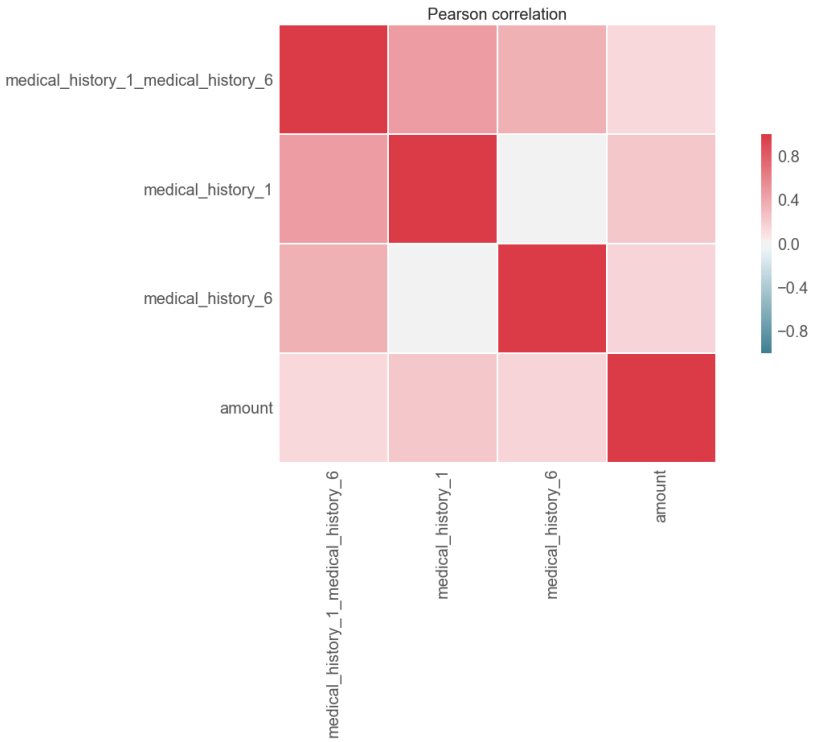
- **Mass index we transform applying (X – X_mean) transformation to separate two distribution peaks by 0 value**



- **Feature elimination by VIF value. Performing auxiliary regression we exclude weight feature**

VIF Factor	features
642.892827	weight
61.034605	height
596.365870	mass_index

- **We create additional feature medical_history_6 that is a multiplication of medical_history_1 and medical_history_6**



- **We exclude outliers of bills amount which are out of 3 standard deviations range**

- **R squared looks promissing. F-statistic says our model is significant**

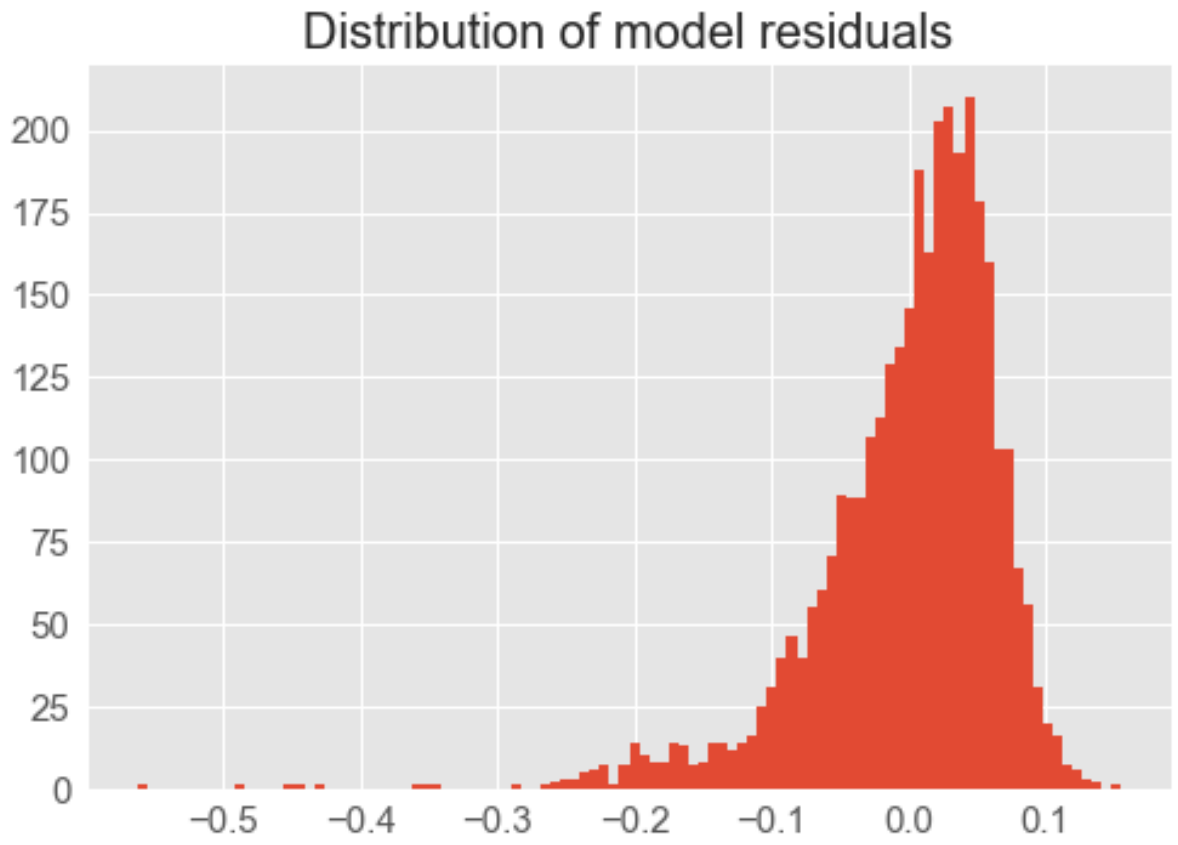
Dep. Variable:	amount	**R-squared:**	0.976
Model:	OLS	**Adj. R-squared:**	0.976
Method:	Least Squares	**F-statistic:**	4266.
Date:	Thu, 25 Jan 2018	**Prob (F-statistic):**	0.00
Time:	15:45:33	**Log-Likelihood:**	4361.9
No. Observations:	3375	**AIC:**	-8658.
Df Residuals:	3342	**BIC:**	-8456.
Df Model:	32		

- **Perorming 10-folds cross validation across 3 random seeds gives us**
  - R squared mean value after back transformation value among folds and seeds 0.974 with standard deviation 0.0046
  - MAPE mean value after back transformation among folds and seeds 0.05 with standard deviation 0.003

Results

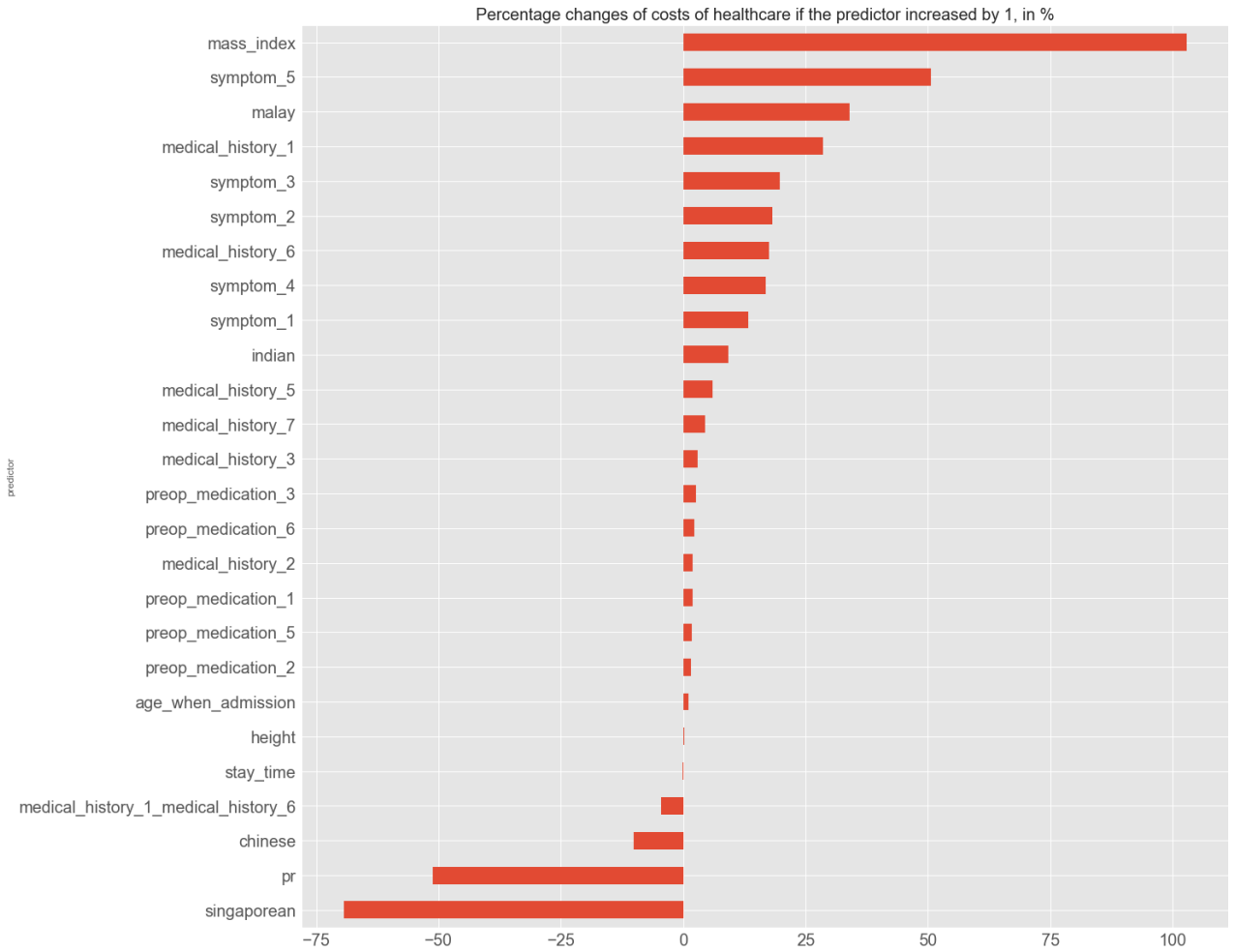
Performing regression with elimination of features with low pvalues of coefficients gave good results

- **The distribution of residuals looks normal but our model tends to understate costs of care. Since our MAPE is very low, there is no problem**



Results
Factor influence (for more details check appendix)

- **Mass index, symptom 5, malay race and medical history 1 are the main factors that increase costs of care**
- **Singaporean citizenship, pr and chinese race are the main factors that decrease costs of care**



THANK YOU VERY MUCH

- **The code is available here**

  https://github.com/paveltr/tests/blob/master/20180123_holmusk.ipynb

- **My profile**

  https://www.linkedin.com/in/paveltr/

Appendix 1/1
Regression coefficients

coeficient	Coeficient value	std error	t	P>|t|	[0.025	0.975]
const	8.8926	0.046	194.118	0.000	8.803	8.982
medical_history_1	0.2855	0.004	80.172	0.000	0.279	0.293
medical_history_2	0.0197	0.003	7.733	0.000	0.015	0.025
medical_history_3	0.0295	0.003	8.791	0.000	0.023	0.036
medical_history_4	0.0074	0.005	1.424	0.154	-0.003	0.018
medical_history_5	0.0596	0.005	12.095	0.000	0.050	0.069
medical_history_6	0.1755	0.003	60.601	0.000	0.170	0.181
medical_history_7	0.0445	0.003	16.764	0.000	0.039	0.050
preop_medication_1	0.0193	0.002	8.345	0.000	0.015	0.024
preop_medication_2	0.0158	0.002	6.716	0.000	0.011	0.020
preop_medication_3	0.0252	0.003	8.341	0.000	0.019	0.031
preop_medication_5	0.0167	0.003	5.560	0.000	0.011	0.023
preop_medication_6	0.0232	0.003	8.774	0.000	0.018	0.028
symptom_1	0.1325	0.002	55.706	0.000	0.128	0.137
symptom_2	0.1816	0.002	74.155	0.000	0.177	0.186
symptom_3	0.1974	0.002	84.948	0.000	0.193	0.202
symptom_4	0.1680	0.003	64.723	0.000	0.163	0.173
symptom_5	0.5064	0.002	219.105	0.000	0.502	0.511
lab_result_3	6.65e-05	7.56e-05	0.880	0.379	-8.17e-05	0.000
height	0.0015	0.000	6.295	0.000	0.001	0.002
stay_time	-0.0010	0.000	-2.546	0.011	-0.002	-0.000
mass_index	1.0281	0.019	54.781	0.000	0.991	1.065
gender	-0.0031	0.003	-0.911	0.362	-0.010	0.004
indian	0.0924	0.006	15.001	0.000	0.080	0.104
chinese	-0.1009	0.005	-19.459	0.000	-0.111	-0.091
malay	0.3406	0.006	60.941	0.000	0.330	0.352
singaporean	-0.6936	0.006	-122.502	0.000	-0.705	-0.682
pr	-0.5120	0.006	-81.861	0.000	-0.524	-0.500
age_when_admission	2.656e-05	2.15e-07	123.384	0.000	2.61e-05	2.7e-05
month_1	0.0068	0.004	1.589	0.112	-0.002	0.015
month_9	0.0074	0.004	1.668	0.095	-0.001	0.016
month_10	-0.0052	0.004	-1.285	0.199	-0.013	0.003
medical_history_1_medical_history_6	-0.0462	0.007	-6.484	0.000	-0.060	-0.032