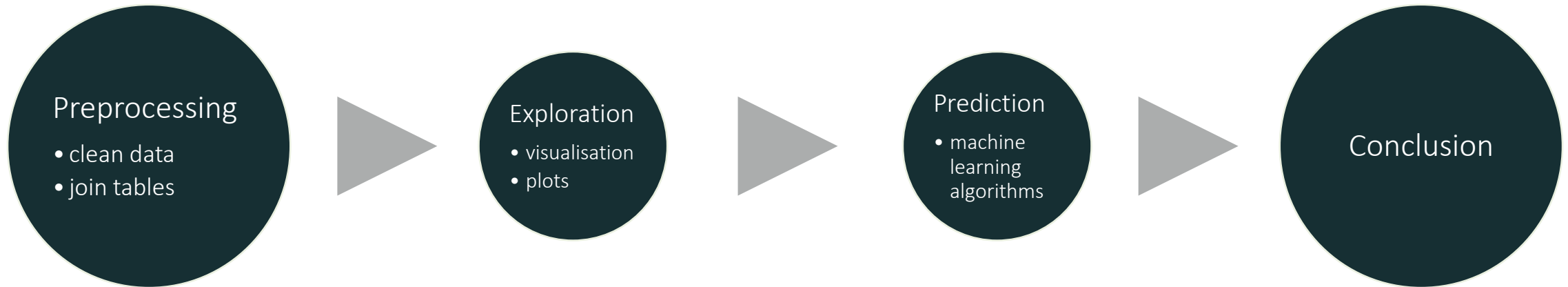


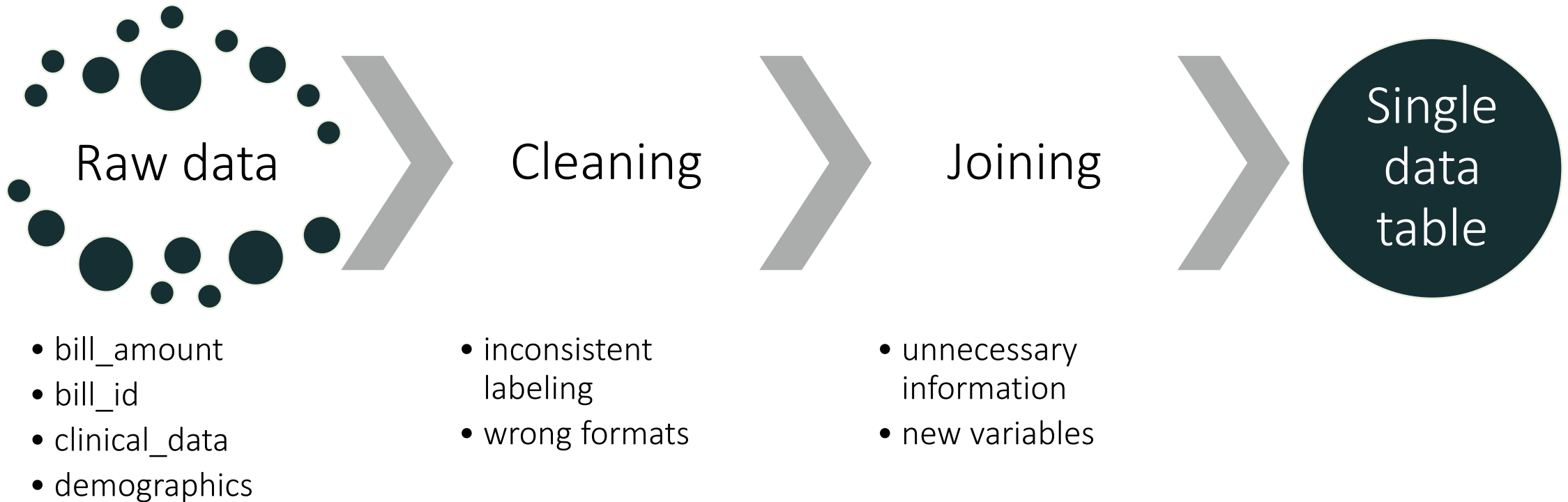
Analysis on Clinical and Financial Data of Patients with a Certain Condition

Matthew Zakharia Hadimaja

Analysis



Preprocessing



Exploration

variables

- numerical
 - continuous
 - binary
- categorical

between predictors

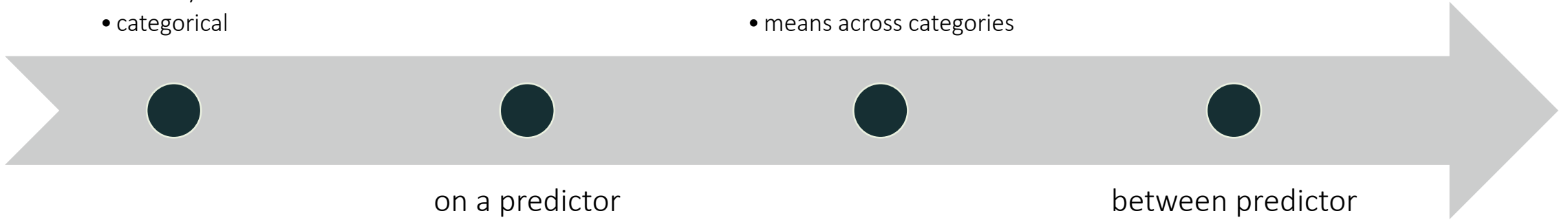
- correlation
- means across categories

on a predictor

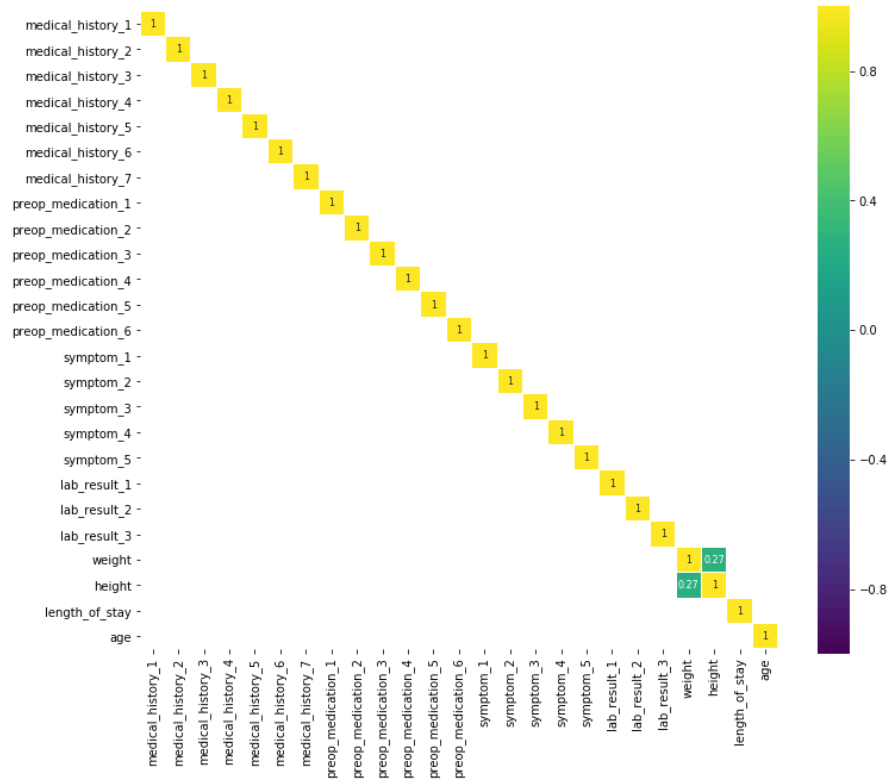
- histograms
- bar plot

between predictor and response

- regression line
- box plot

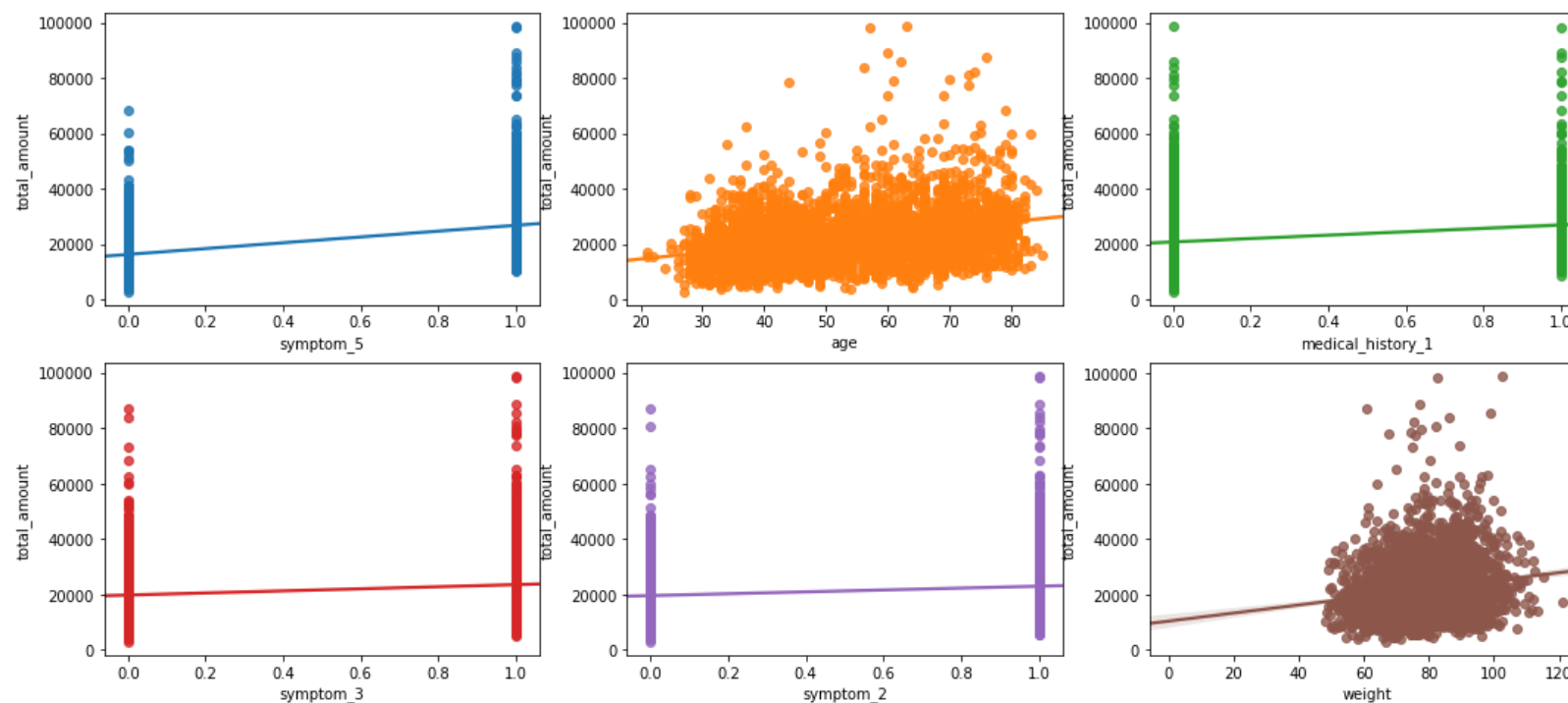


Exploration - numerical



Predictors are uncorrelated of each other. Only one predictor pair (weight-height) has correlation more than 0.05!

Exploration - numerical

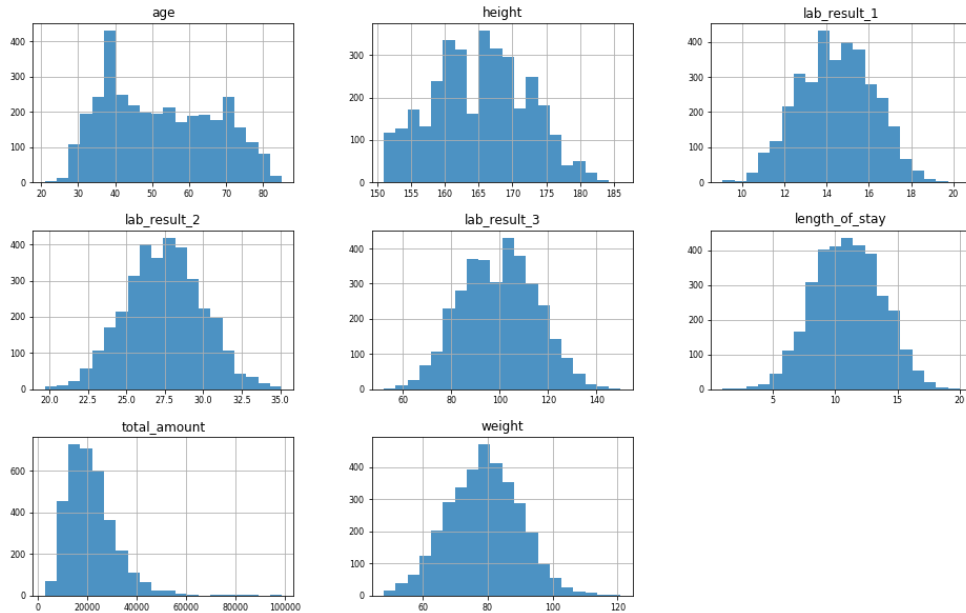


Top 6 predictors with highest correlation with total_amount:

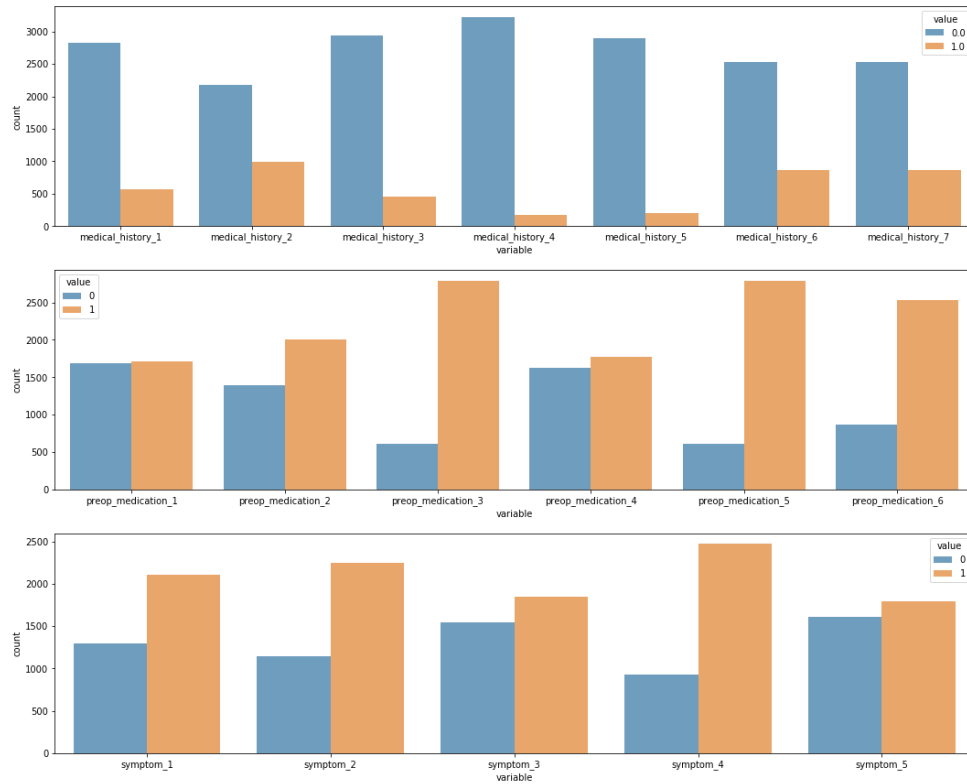
variable	correlation
symptom_5	0.517
age	0.326
medical_history_1	0.227
symptom_3	0.184
symptom_2	0.158
weight	0.158

Exploration - continuous

Most distribution have Gaussian shape, with some following bimodal distributions.



Exploration - binary

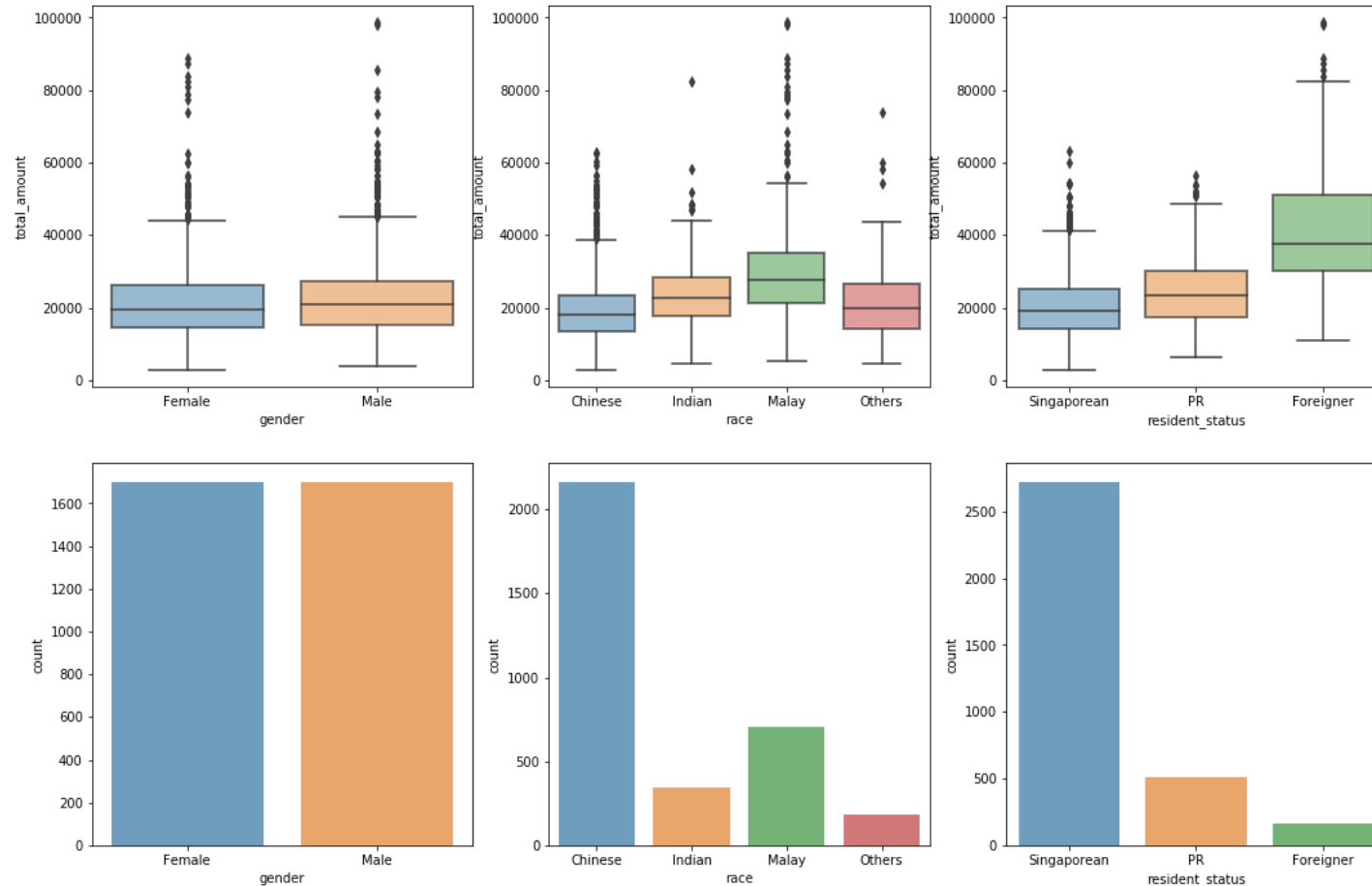


Medical history variables (top row) are unbalanced, but it is expected.

Patients under this condition are more likely to receive certain preop medications.

Some symptoms are more common than the other under this condition.

Exploration - categorical



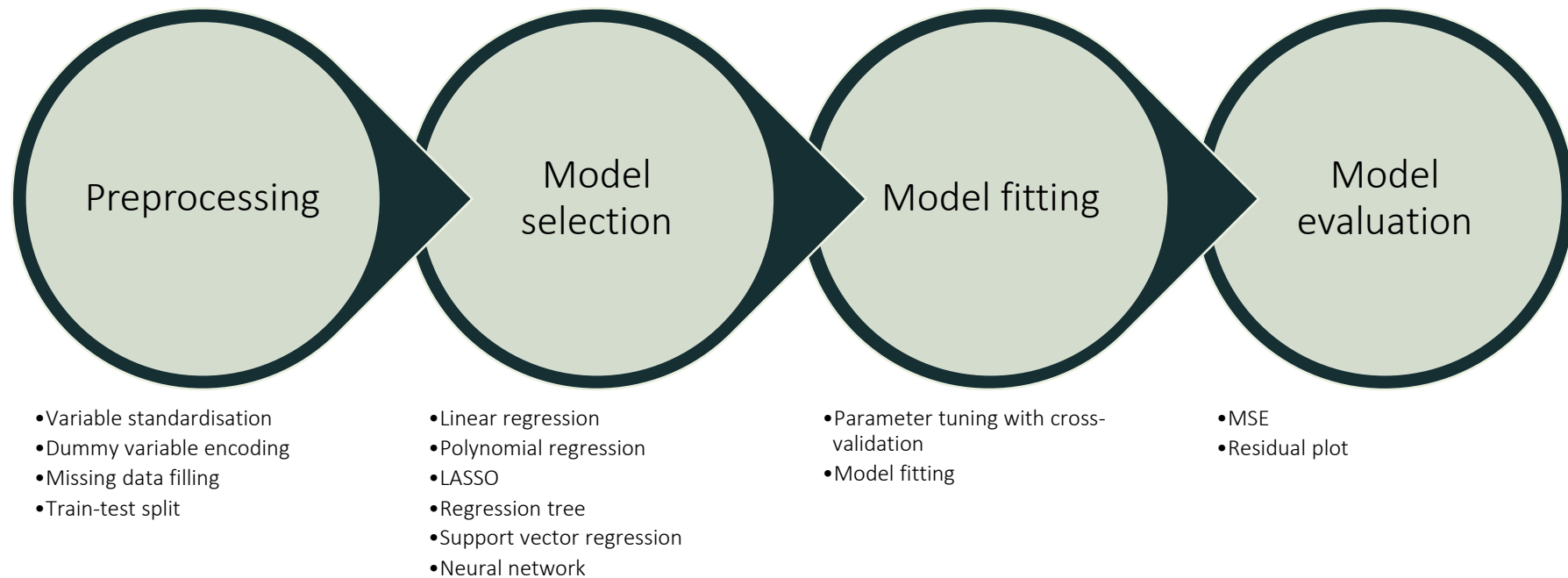
No difference in cost across gender. The condition also affects each gender equally.

Malays and Indians have higher cost. Does this condition affect them more?

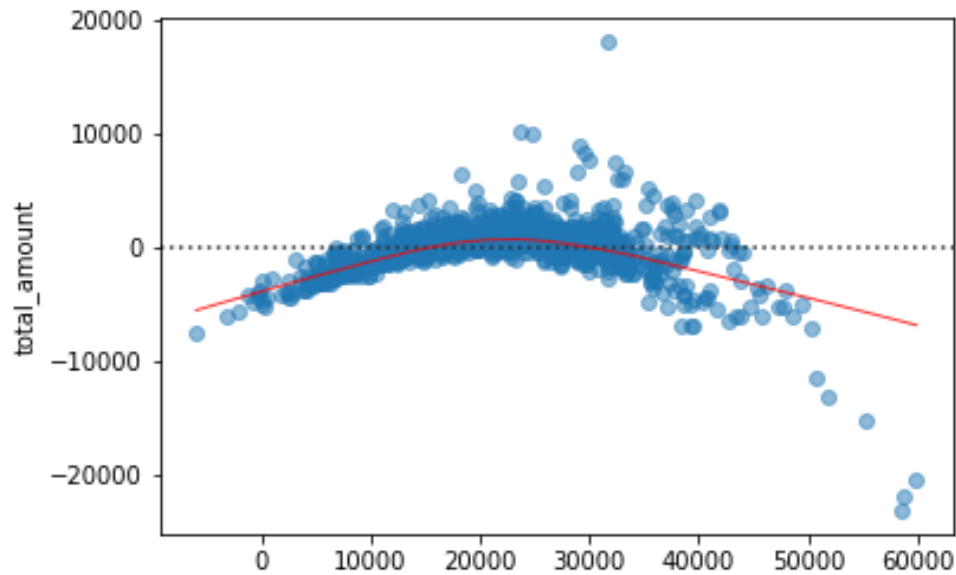
Foreigners and PRs pay more than Singaporeans do.

race and resident_status are not distributed equally in our data.

Prediction



Prediction – linear regression



Baseline model

MSE = 2473.793

Important variables:

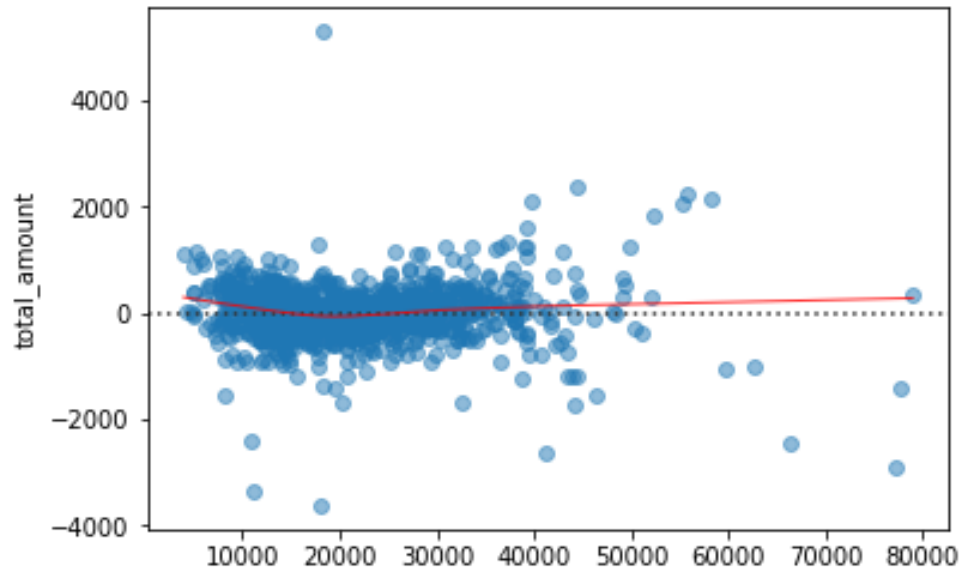
symptom_5, symptom_3, symptom_2
medical_history_1, medical_history_6
race and resident_status

Prediction – polynomial regression

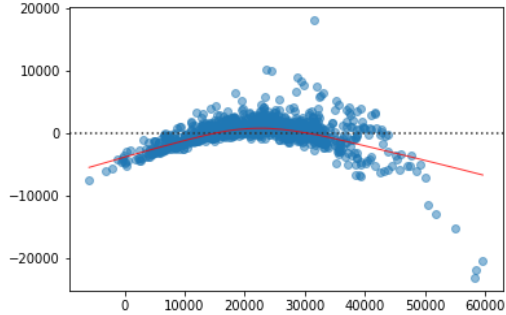
Best model

MSE = 520.633

Higher performance, more variables. May be difficult to explain.



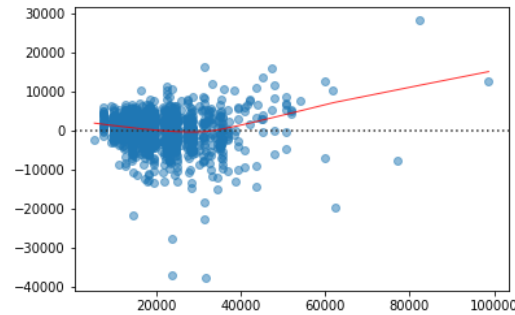
Prediction – other models



LASSO

- Small penalty from CV
- Similar to OLS

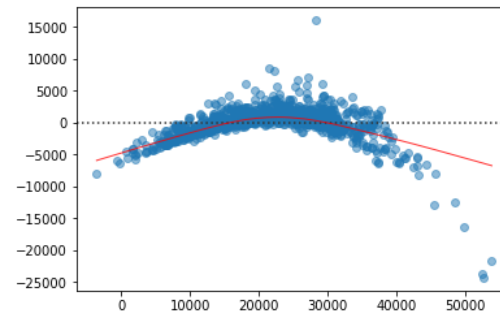
MSE: 2466.897



Regression Tree

- Worst model, even with deep tree

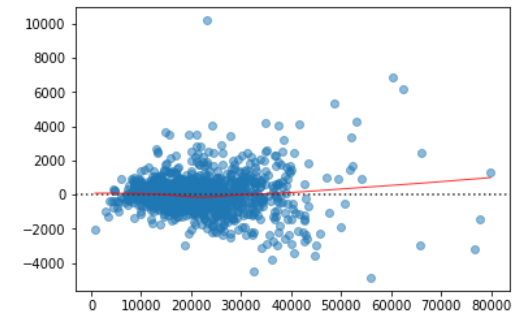
MSE: 4868.979



Support Vector Regression

- CV chose linear kernel
- Similar to OLS

MSE: 2555.064



Neural Network

- Better than OLS, but worse than polynomial regression

MSE: 1122.375

Conclusion

Important cost drivers

- symptom variables
- race, residential status
- age, weight
- some medical history variables

Less relevant variables

- preop medication variables
- lab results
- some medical history variables

Model selection

- simple models with regularized parameter