HEALTHCARE ANALYTICS REPORT
MARIA ATHENA B. ENGESAETH (CID 01159179)

Q 1: PREDICTIVE MODELLING OF PATIENTS' READMISSION

I. DATA

The dataset used to answer the below questions consisted of the patient profiles of ~100 000 diabetic patients. The data contained information on race, gender, age group in addition to number of lab procedures received at the hospital, time spent there, number of medications and whether the patient was readmitted more or less than 30 days after the first admission. For my model I made the following modifications to the dataset (Appendix: *MAE_healtchare_parser.py*):

- The **age** groups were split into three categories:
  - Children and young adults: 0 - 30
  - Adults: 30 - 70
  - Elderly: 70 - 100
- Minority **races** including Asian, Hispanic and unknowns were grouped to '*other'.*
- The two variables of **readmission** before and after 30 days of release were combined.

II. PREDICTIVE MODEL

For the model I used the method of stepwise regression on a logit model. The final combination of independent variables to include gave us the following predictive model for the dependent variable *Y {0 ; 1} being the likelihood of a patient's readmission* after release from the hospital:

-0.934 + **0.002** num_lab_procedures **- 0.086** num_procedures + **0.013** num_medications

+ 0.022 time_in_hospital + **0.122** adult + **0.193** elderly + **0.043** Female

- 0.037 AfricanAmerican - **0.381** Other

Pseudo R.sq: 0.008
Sample size: 101 766

For the age group variable: *children and young adults* is the omitted/benchmark category.
For the race variable: *Male* is the omitted/benchmark category.
For the race variable: *Caucasian* is the omitted/benchmark category.

III. RESULTS

All the included independent variables are statistically significant, below are some points that summarise the most interesting findings concluded by the model:

- Both adult (~13%) and elderly (~20%) patients are more likely to be readmitted.
- Having a medical procedure that was not related to lab tests reduced the likelihood of patient readmission by ~9%.

For the full output of the statistical model:

Logit Regression Results

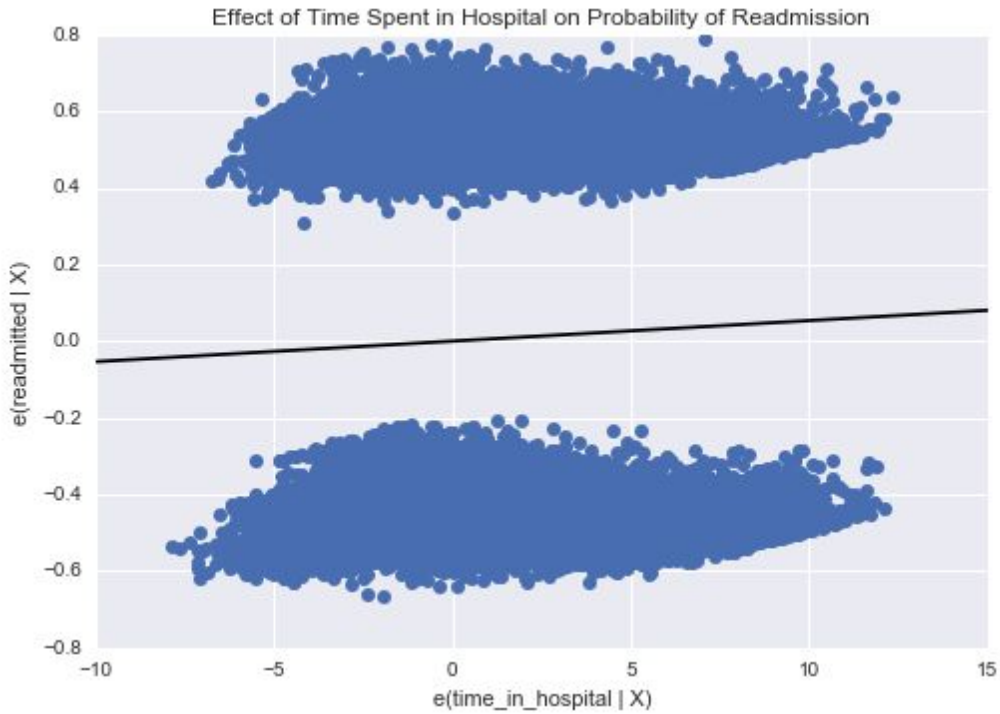
Dep. Variable:	readmitted	No. Observations:	101766
Model:	Logit	Df Residuals:	101756
Method:	MLE	Df Model:	9
Date:	Sat, 02 Apr 2016	Pseudo R-squ.:	0.007985
Time:	15:57:56	Log-Likelihood:	-69666.
converged:	True	LL-Null:	-70227.
		LLR p-value:	1.058e-235

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
Intercept	-0.5780	0.044	-13.057	0.000	-0.665	-0.491
num_lab_procedures	0.0020	0.000	5.819	0.000	0.001	0.003
num_procedures	-0.0852	0.004	-20.757	0.000	-0.093	-0.077
num_medications	0.0128	0.001	13.401	0.000	0.011	0.015
time_in_hospital	0.0216	0.002	8.753	0.000	0.017	0.026
Age_adult	0.1281	0.042	3.055	0.002	0.046	0.210
Age_elderly	0.1954	0.042	4.645	0.000	0.113	0.278
Gender_Female	0.0436	0.013	3.404	0.001	0.018	0.069
Race_AfricanAmerican	-0.0414	0.017	-2.503	0.012	-0.074	-0.009
Race_Other	-0.3813	0.027	-14.150	0.000	-0.434	-0.328

Q 2: RELATIONSHIP BETWEEN TIME IN HOSPITAL VS PROBABILITY OF READMISSION

Although statistically significant, the ***time*** spent in the hospital did not have an economically significant impact on the likelihood of a patient's readmission. The parameter associated with this variable is **+ 0.022**, below it the partial regression plot of this relationship.

In contrast the **number of procedures** (- 0.086) as well as the demographic characteristics had a more significant impact on this and is explored in the last section of this report.



Q 3: READMISSION *UNDER* 30 DAYS *vs OVER* 30 *DAYS*

46% of the patients in the sample were readmitted after being released from the hospital

- ~ 11% are readmitted less than 30 after being released.
- ~ 35% are readmitted 30 after being released.

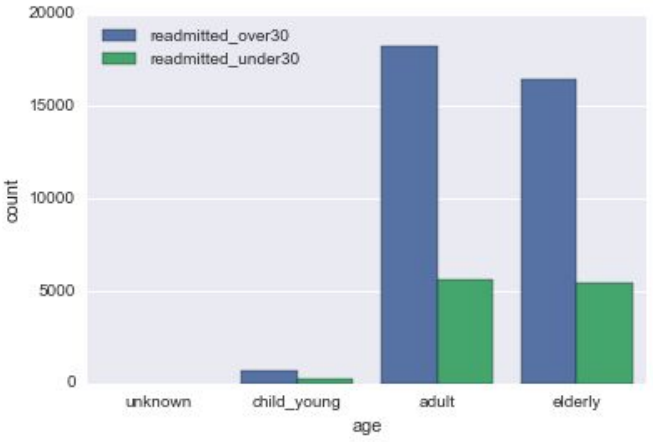
The patients in the dataset suffer from a chronic disease. Achieving improvements in the above rates can lead to improved quality of life for the patients in addition to higher economic efficiency of the hospital's operations.

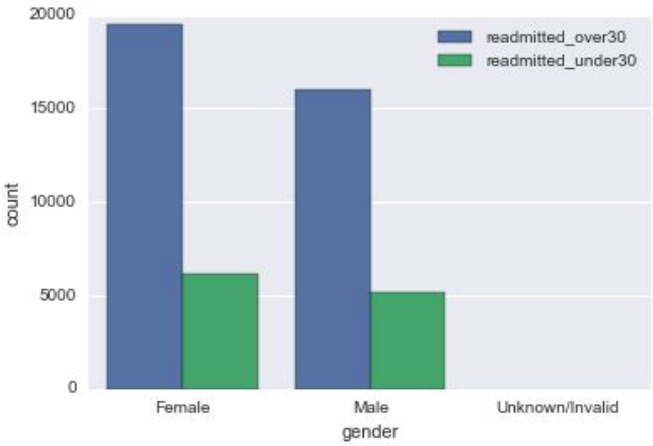
Therefore, given the above statistics it is worth exploring what might be the characteristics of the patients that are readmitted so as to either:

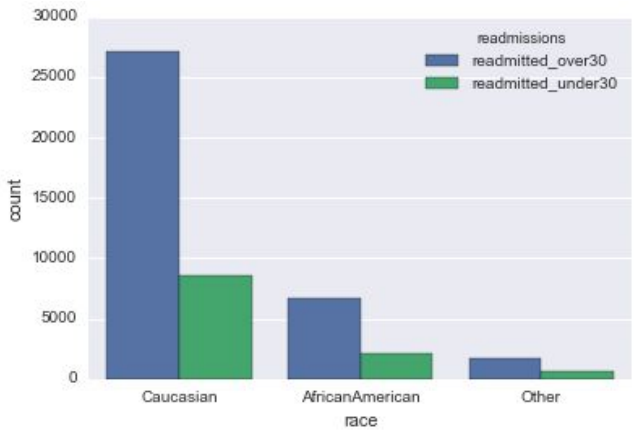
i. reduce the rate of patients readmitted shortly after being released
Ii. reduce the overall rate of patients readmitted

PATIENT PROFILE BY DEMOGRAPHICS

In the below plots we have only included only the patients that were readmitted. We can conclude that the proportion of readmittance readmittance *over* 30 and *under* 30 days is proportional regardless of demographic differences.



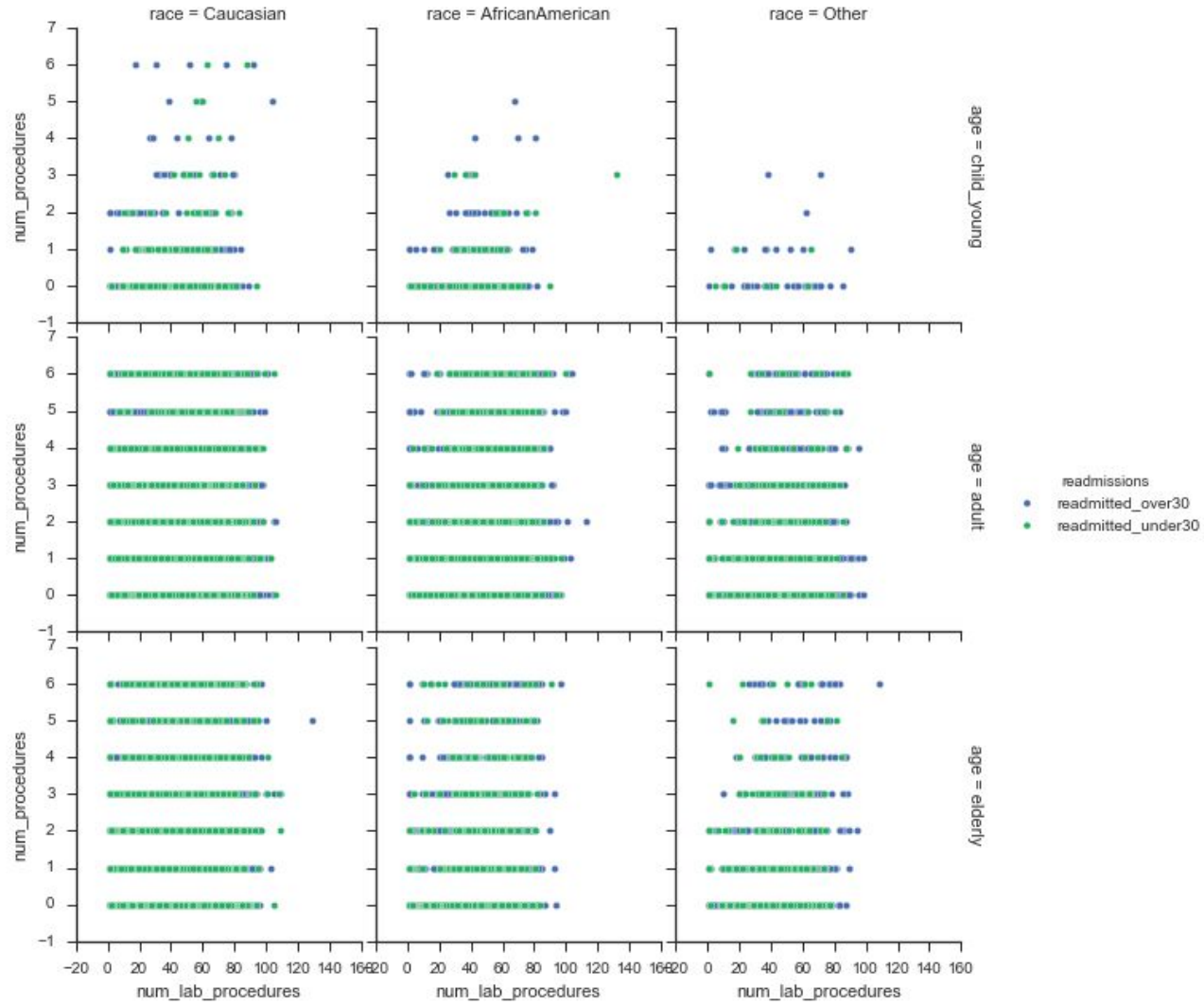




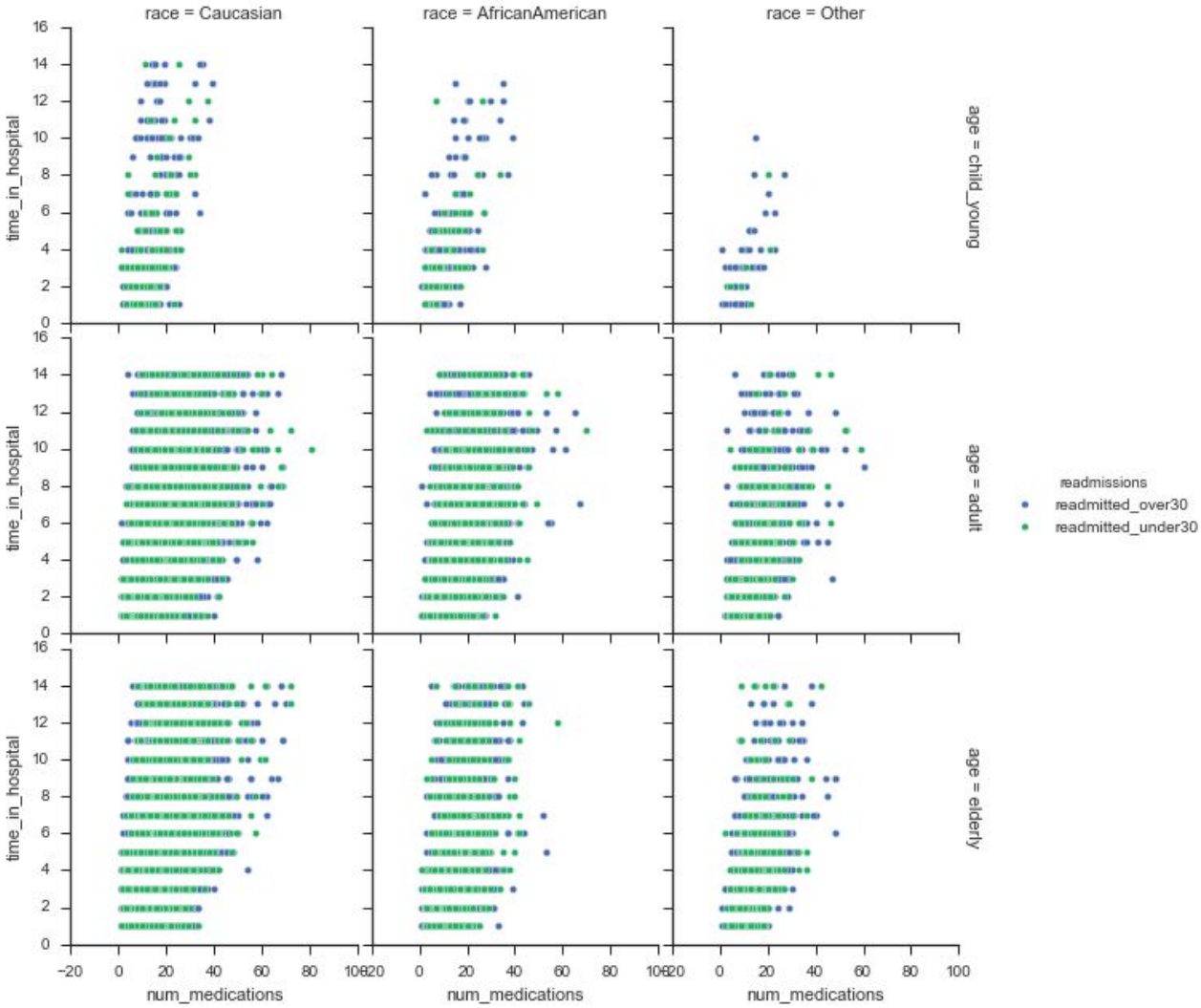
PATIENT PROFILE BY DEMOGRAPHICS **AND** ILLNESS SEVERITY

Lastly I wanted to look at a more granular combination of both demographics and illness severity. Below are two grids of scatter plots that have the patient's race on the columns and the age group on the rows.

The first grid combines *num_procedures* against *num_lab_procedures.* There does not seem to be a clear visual separation between patients.



The second grid combines *time_in_hospital* against *num_medications.* Again there does not seem to be a clear visual separation between patients.



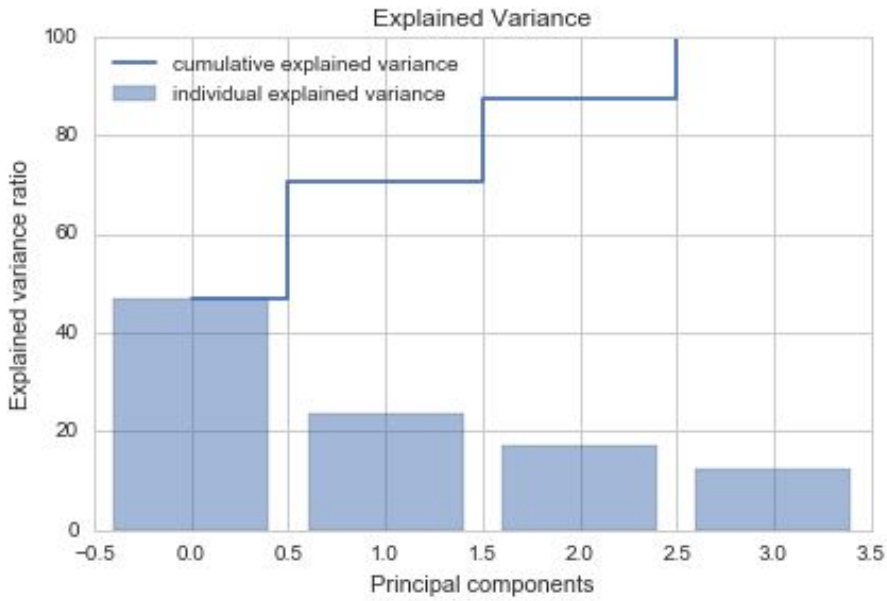
PATIENT PROFILE BY ILLNESS SEVERITY

As no clear profile emerged of patients that would be submitted before vs after 30 days I wanted to look at alternative methods of distinguishing these patient groups. First a PCA and then a attempt to cluster the data using the k-means algorithm and then computing the confusion matrix.

1. PRINCIPAL COMPONENT ANALYSIS

Treating the variables of *num_lab_procedures (0), num_procedures (1), num_medications (2)* and *time_in_hospital (3)* as proxies of the severity of the patient's illness: we can test to see which of the variables are most responsible for explaining the variance in the data.

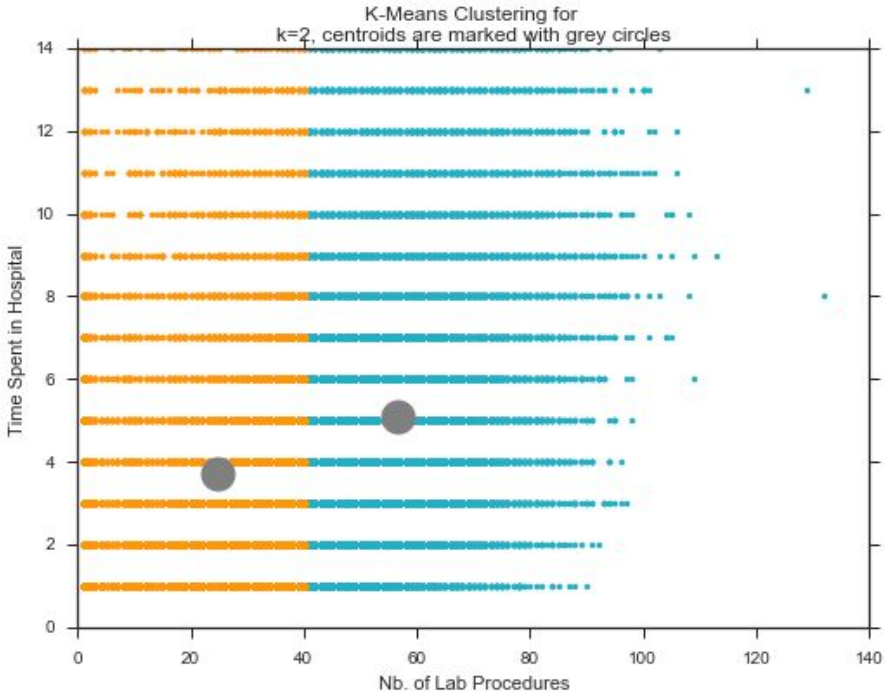
PCA analysis can be error-prone for categorical variables so the demographic variables were excluded. It was deduced that including 2-3 of the variables would account for the most significant variability in the data.

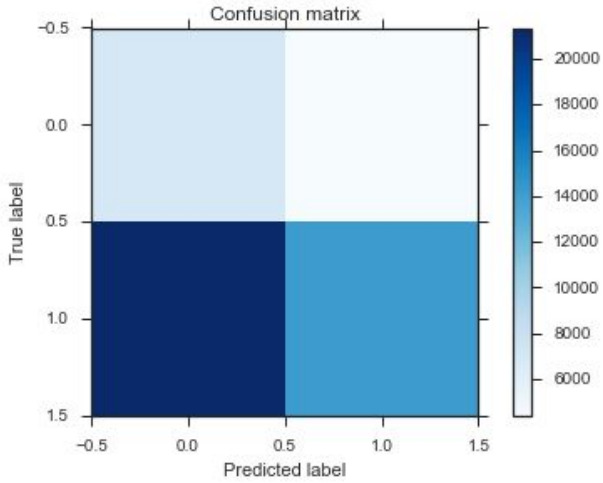


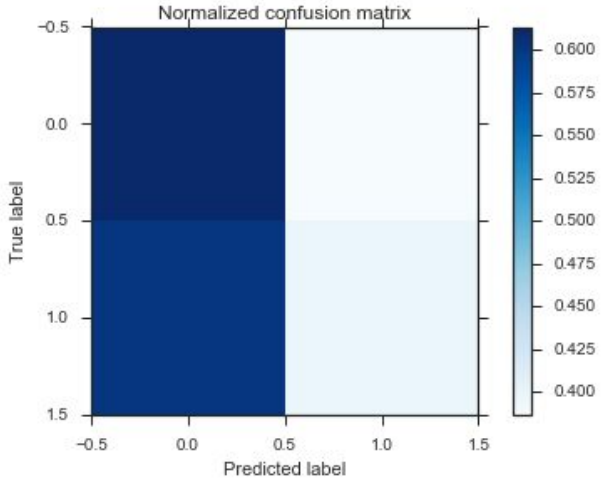
2. K-MEANS CLUSTERING

Multiple combinations of variables to use for clustering were attempted and none yielded particularly feasible results. Attached is an example from clustering the data into two groups using the *time_in_hospital* against *num_lab_procedures.* Also find attached the confusion matrix and normalized confusion matrix (heatmap) which in an optimal scenario would have the darkest/most populous squares on the diagonal, representing the true positives and true negatives.

```
Confusion matrix, without normalization
[[ 6963  4391]
 [21355 14164]]
Normalized confusion matrix
[[ 0.61  0.39]
 [ 0.6   0.4 ]]
```


K-Means Clustering for
k=2, centroids are marked with grey circles
Time Spent in Hospital
Nb. of Lab Procedures
0
2
4
6
8
10
12
14
0
20
40
60
80
100
120
140


Confusion matrix
True label
Predicted label
−0.5
0.0
0.5
1.0
1.5
20000
18000
16000
14000
12000
10000
8000
6000


Normalized confusion matrix
True label
Predicted label
−0.5
0.0
0.5
1.0
1.5
0.600
0.575
0.550
0.525
0.500
0.475
0.450
0.425
0.400

APPENDIX FOLDER CONTENTS

1. Python script for tidying the raw data: *MAE_healtchare_parser.py*
2. The parsed data*: parsed_healthcare_data.csv*
3. iPython Notebook containing:
    a. *Statistical analysis*
    b. *Code for visual exploration of patient profiles*