Date-A-Scientist

Machine Learning Fundamentals

Nabyl Belgrade 12-Nov-2018

Table of Contents

- Variables considered
- Question to Answer
- Basic Statistics
- Cleaning of Data
- Codification of Data
- Classification Approaches
- Conclusions

Variables considered

- ***Age***
- ***Income***
- ***Body type***
- ***Drugs***
- ***Essay length***
- ***Number of languages spoken***
- ***Education level***

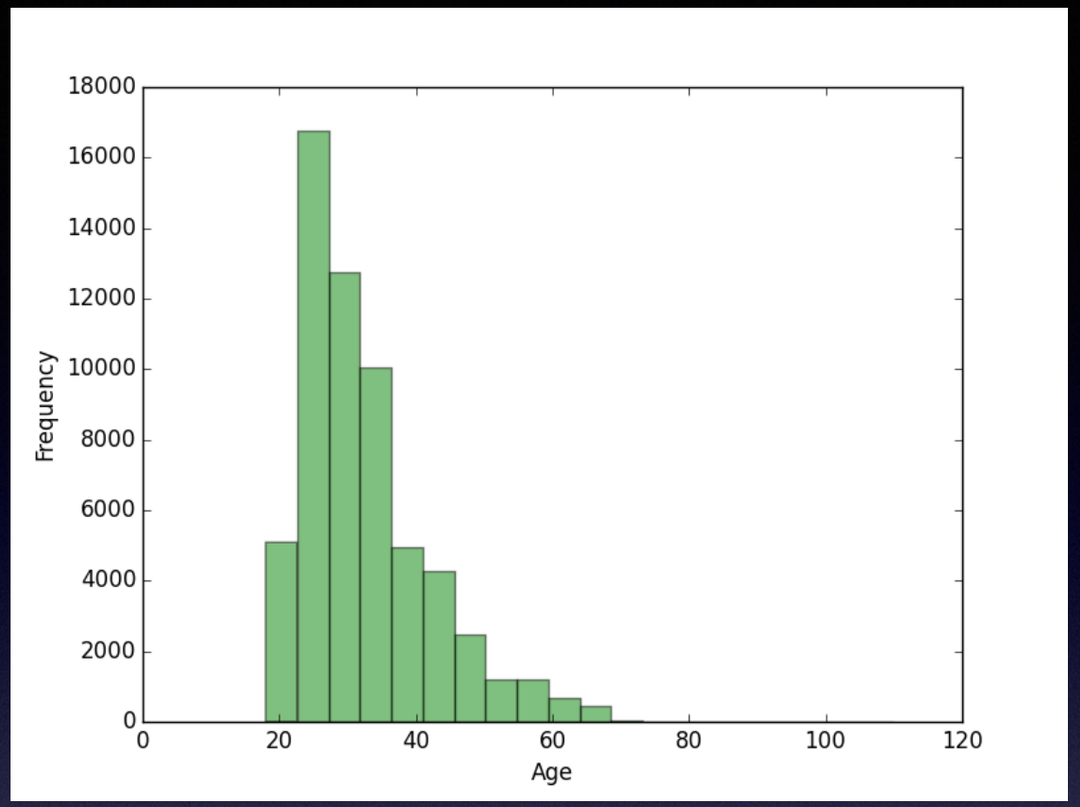
Variables considered

- We added two variables:
  - ***Essay_len:*** that counts the words used in the essays
  - ***Nb_Languages:*** that counts the languages spoken

Question to Answer

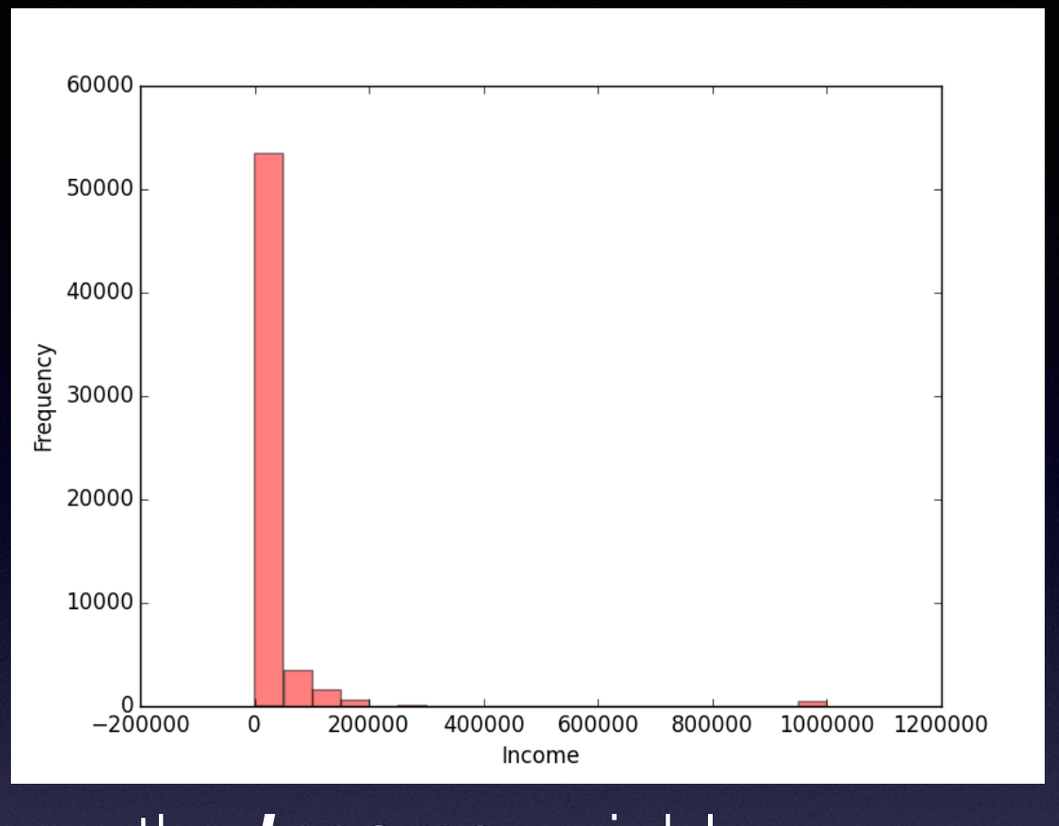
- Can we predict the education level of an individual based on his/her level of income, age, self conscious of body and sophistication of vocabulary?
  - ***Education level:*** described by the variable ***Education***
  - ***Income and Age:*** described by the variables ***Income*** and ***Age*** directly respectively
  - ***Self maintaining:*** described by the variable ***Body_type*** and ***Drugs***
  - ***Sophistication of vocabulary:*** described by the variables ***Essay length*** and ***Number of languages spoken***

Basic Statistics: Histogram of Age



We can see from the histogram that the distribution of the ***Age*** is “balanced” and we can ignore values outside the interval 16 and 80 years old.

Basic Statistics: Histogram of Income

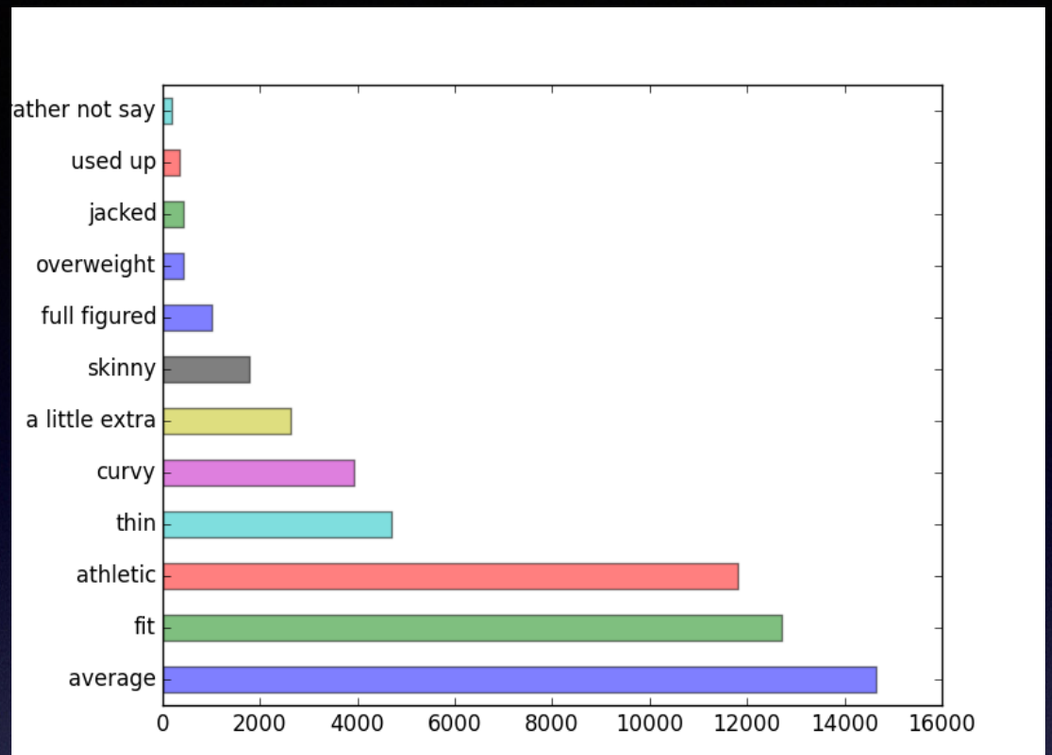


We can see from the following stats on the ***Income*** variable:

- ***Average:*** 20033.22
- ***Median:*** -1.0
- ***Variance:*** 9476281117.08

That we need to filter the Income data as the median is “negative” and the variance is huge.

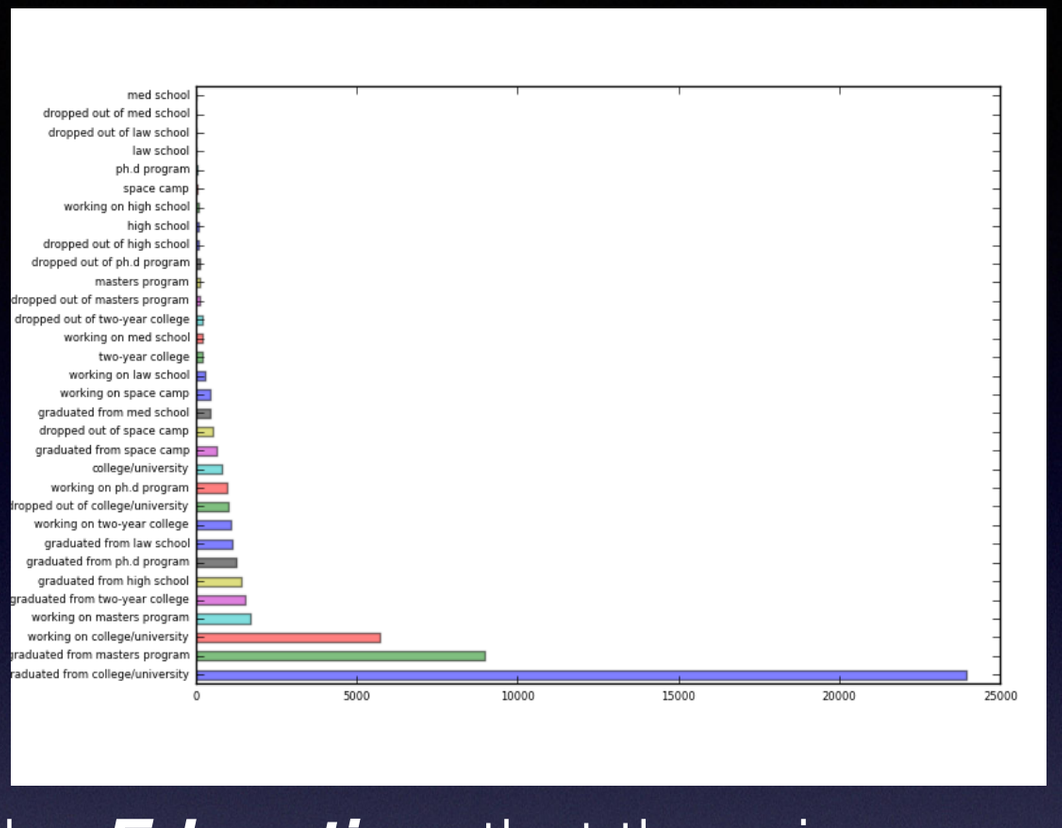
Basic Statistics: Bar chart of Body type



From the Bar chart of the ***Body_type*** data, the fact that the most frequent value is average is comforting.

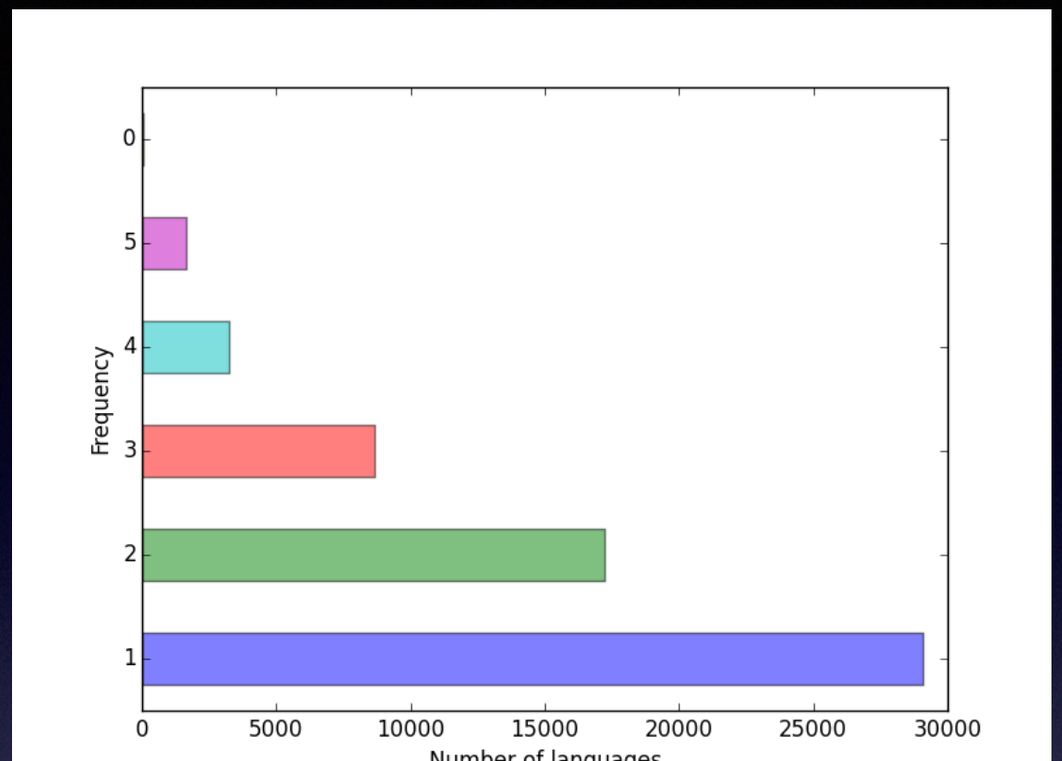
We have to ignore the unknown values as *“rather not say”.*

Basic Statistics: Bar chart of Education



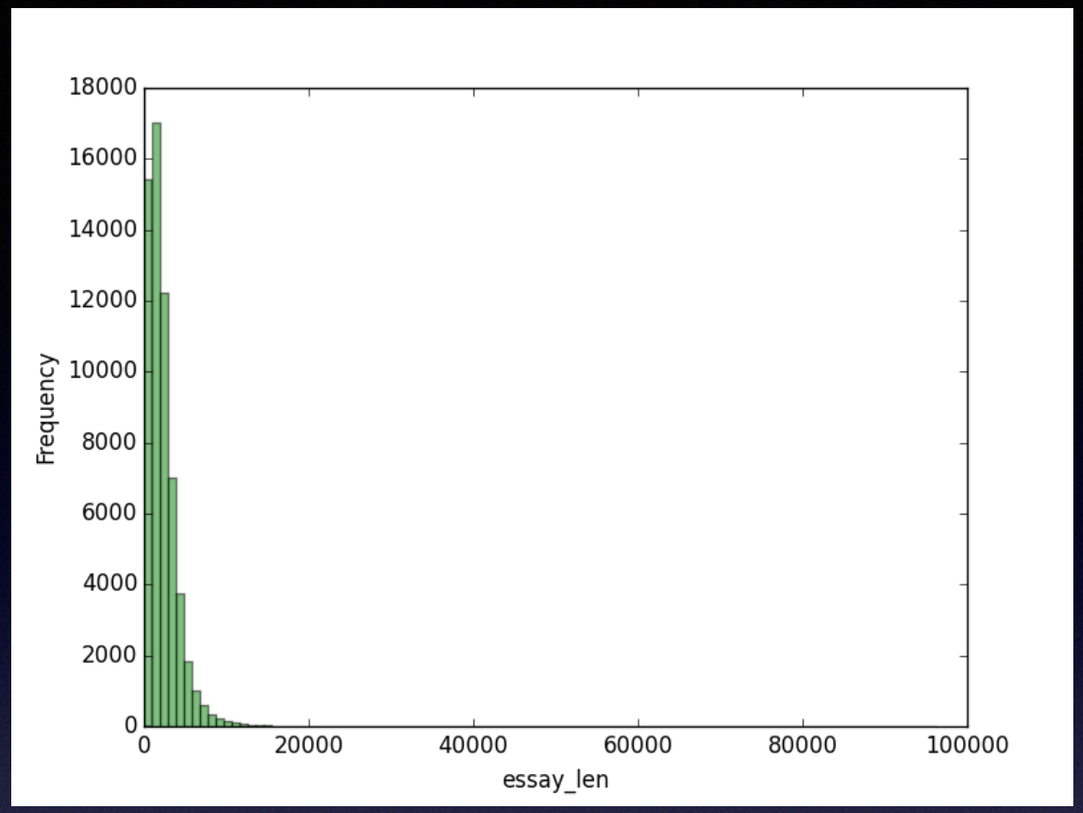
We can see from the Bar chart of the ***Education***, that there is a need to group the data in sub-categories. We can explain the rational in a further slide.

Basic Statistics: Bar chart of Number of Languages



We counted the number languages *decently* (other than reported as *"poorly"*) and we included English as we don't know where the data has been collected. We created a new variable called ***Nb_languages***.

Basic Statistics: Bar chart of Essay_Len



We should consider the data less than 20000 words as there is a huge skew on the right side.

Cleaning of data

- We proceeded to the following filtering of the data with the corresponding rationals:

Variable	Action	Why
All variables	Remove NA values	Standard
Age	Consider data within [16Y, 80Y]	As explained in the slide 5
Income	Consider data within [0$, $80000]	As explained in the slide 6
Body_type	Exclude "rather not say”	Not relevant as a value
Nb_Languages	Exclude 0 values	Not relevant as a value

Cleaning of data

- We proceeded to the following filtering of the data with the corresponding rationals:

Variable	Action	Why
Essay_len	Consider data within [0, 20000]	As explained in the slide 10
Education	Exclude educations with “Working on”	Status pending
Education	Exclude "rather not say”	Not relevant as a value

Codification of qualitative data

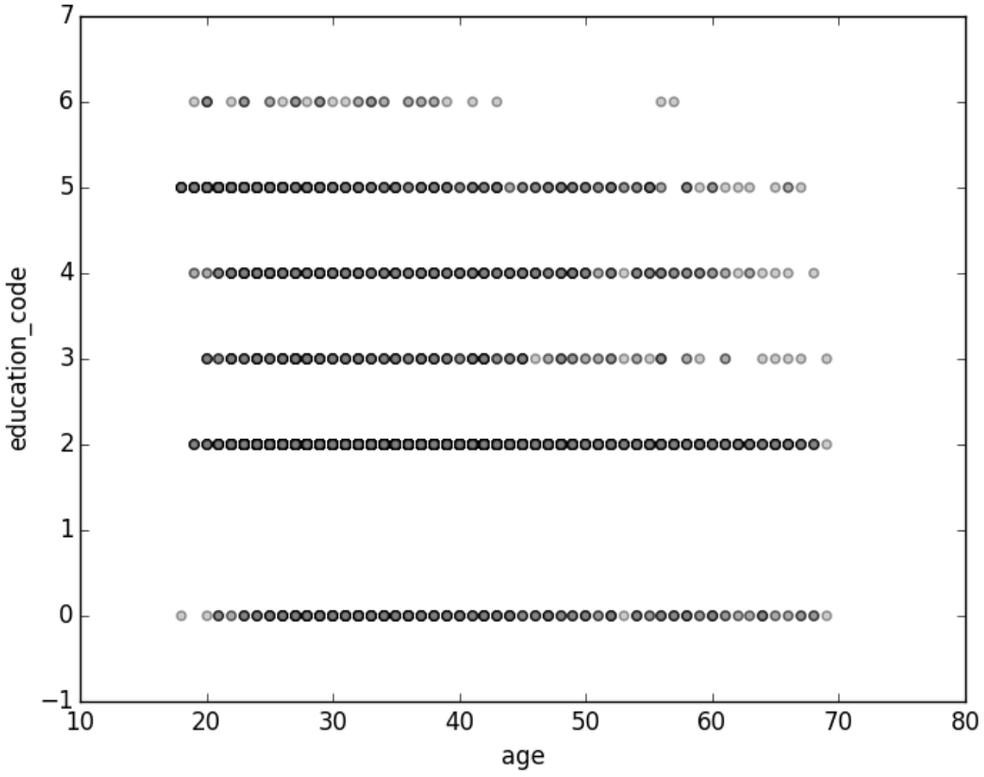
- We gathered the ***Education*** data into categories based on the last institution that an individual has reached and mapped the results in a variable named ***Education_code***

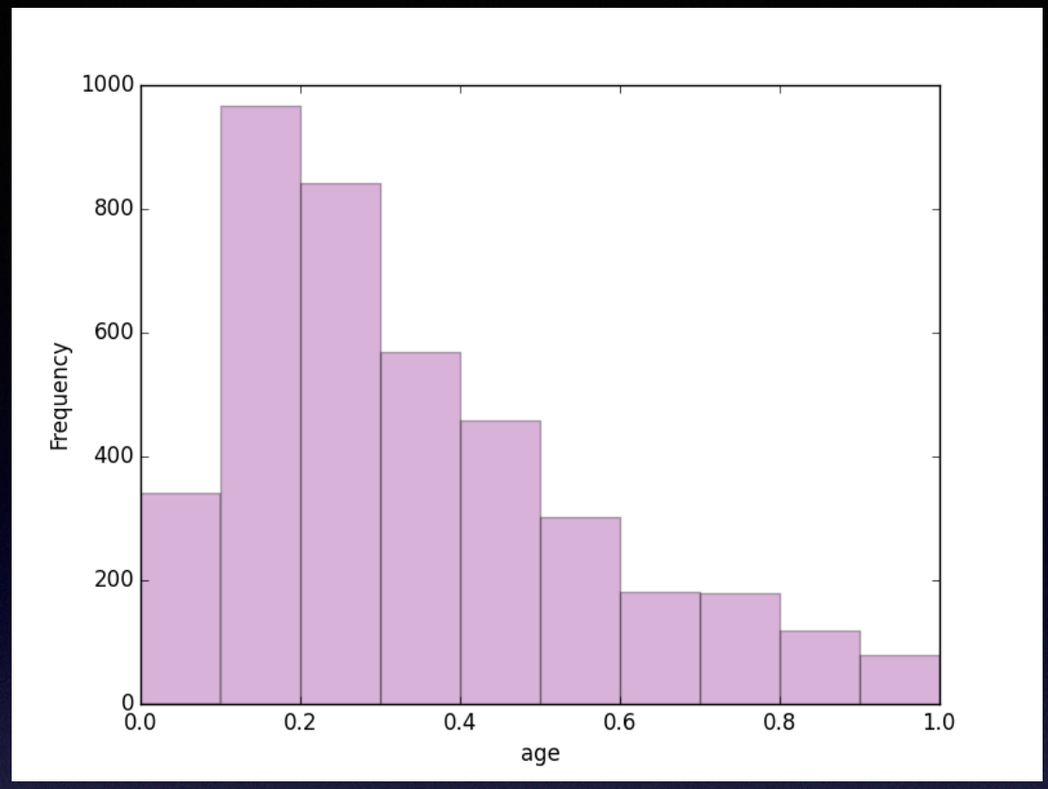
- We mapped one-two-one each of the variables ***Body_type*** and ***Drugs*** separately

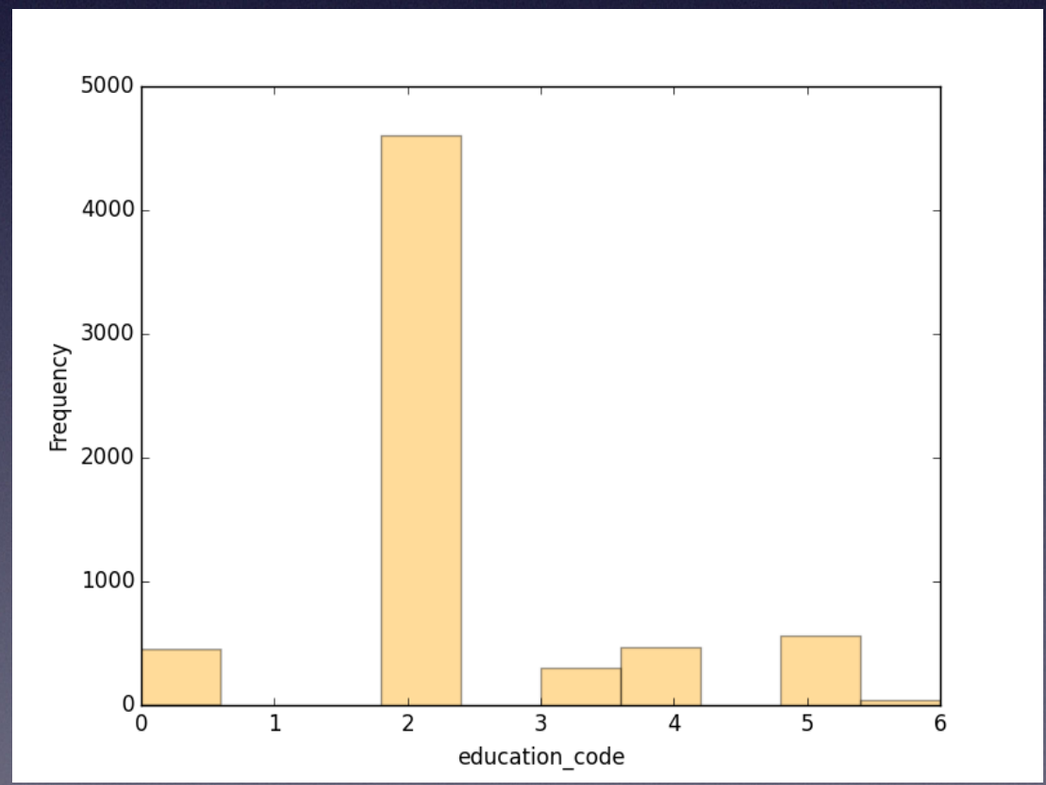
Normalisation

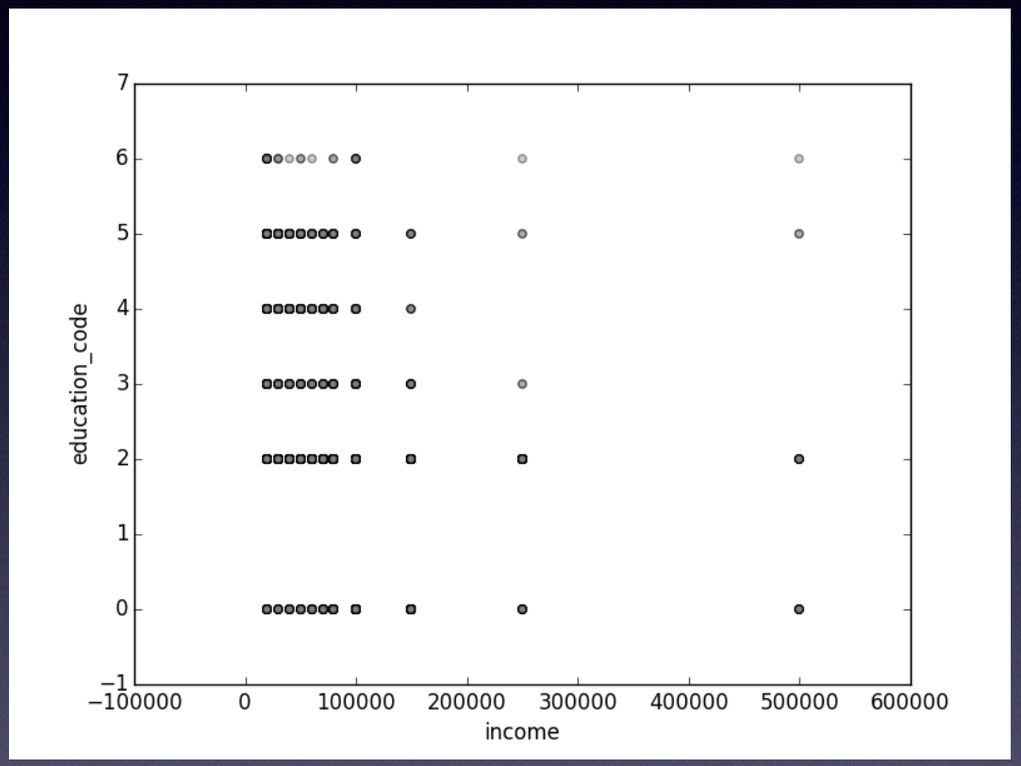
- As prior to every data classification or regression, we proceeded to normalisation steps following a min-max procedure. In the following slides, we expose some charts to show that the data has been well prepared to the analysis.

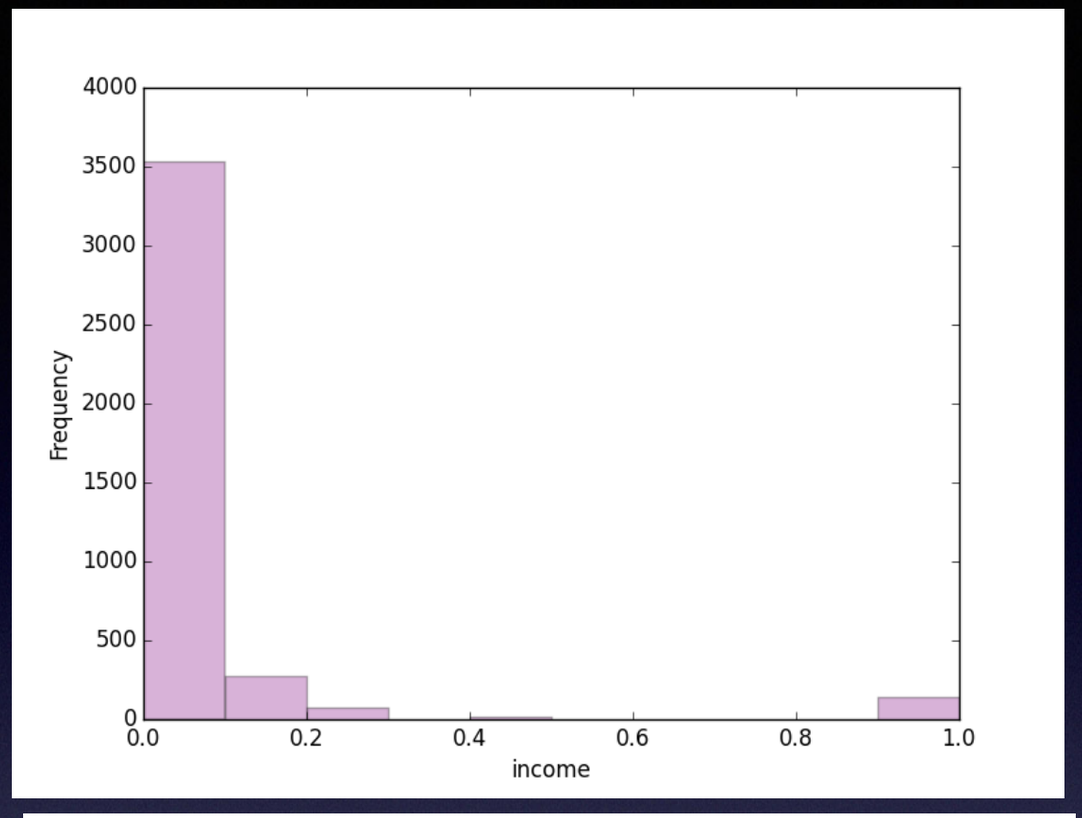
- Also we expose the scatter plot of the “explicative” variables vs the ***Education*** to visualise their cross dependancy

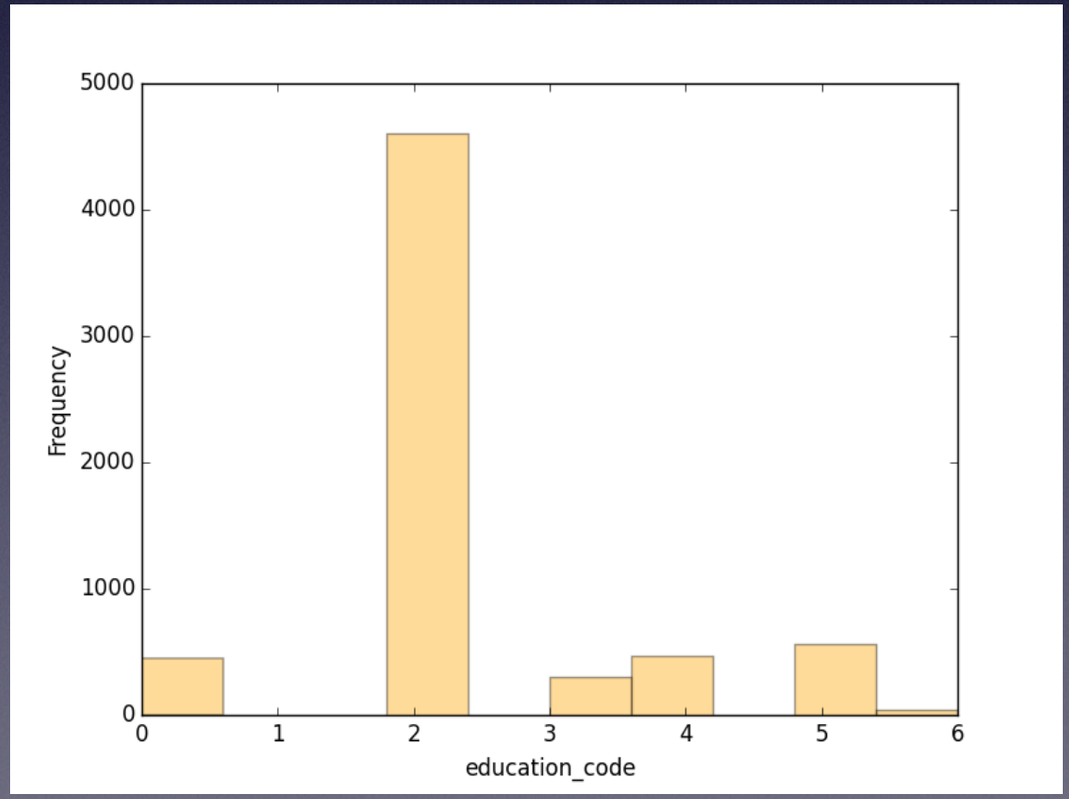

7
6
5
4
3
2
1
0
−1
education_code
10
20
30
40
50
60
70
80
age

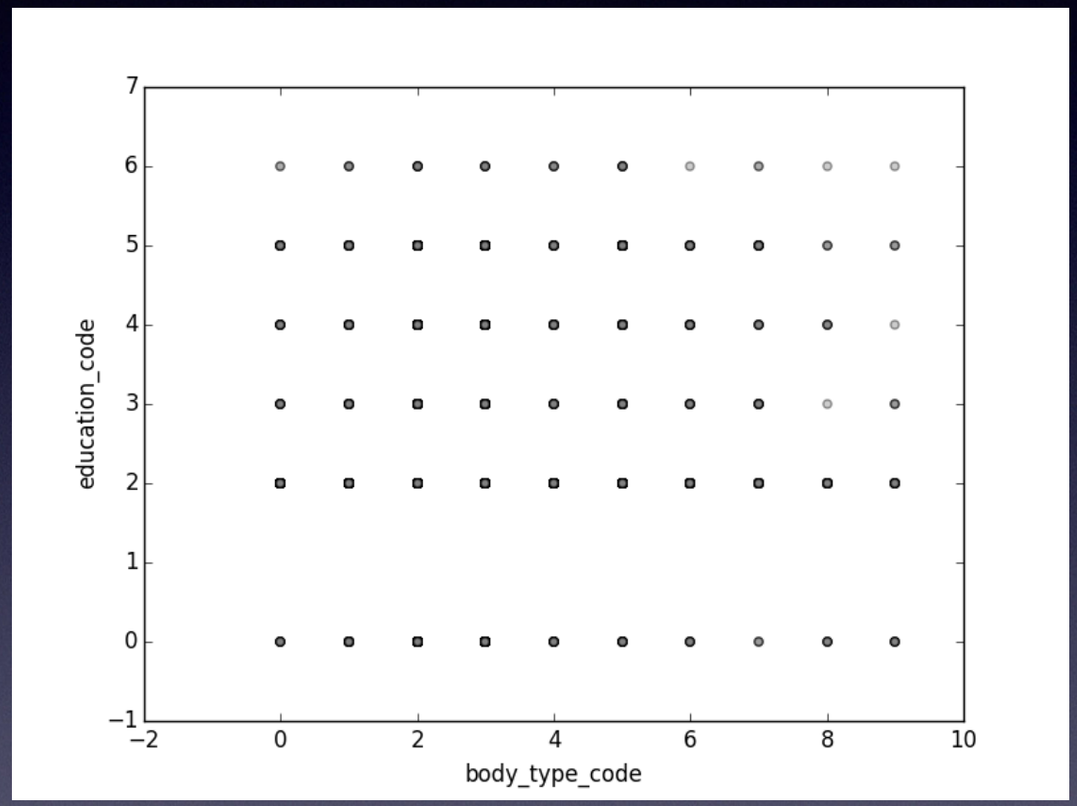

1000
800
600
400
200
0
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
age

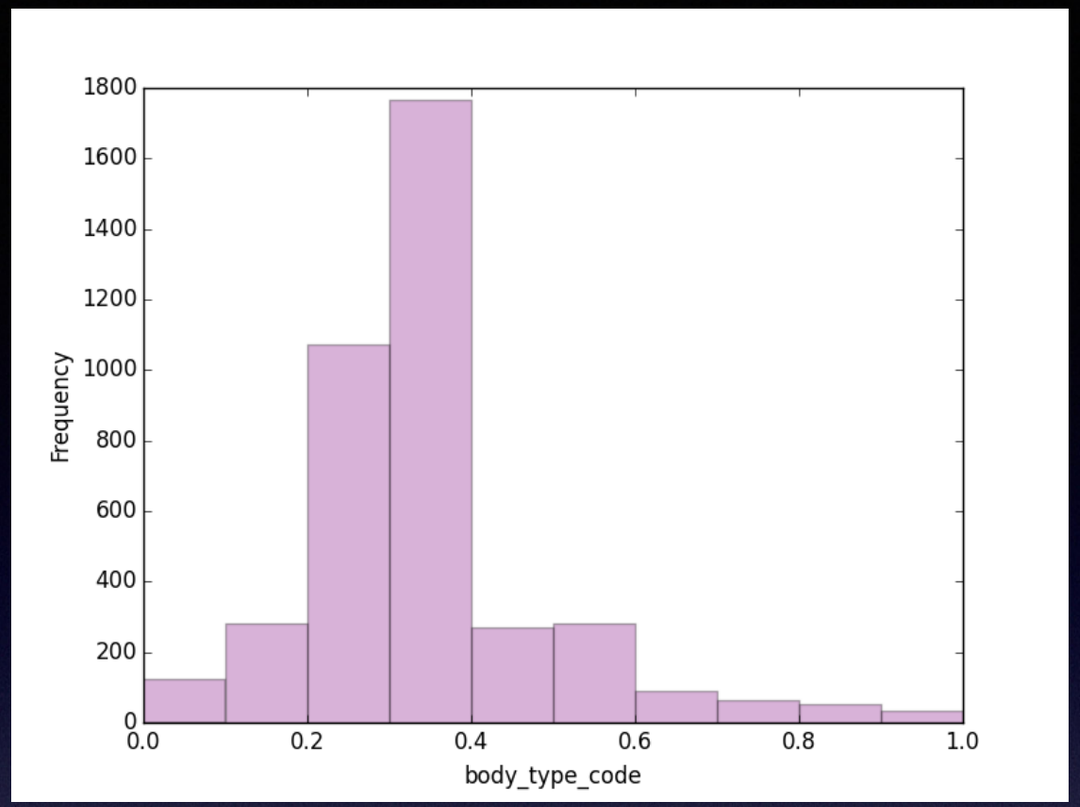

5000
4000
3000
2000
1000
0
Frequency
0
1
2
3
4
5
6
education_code

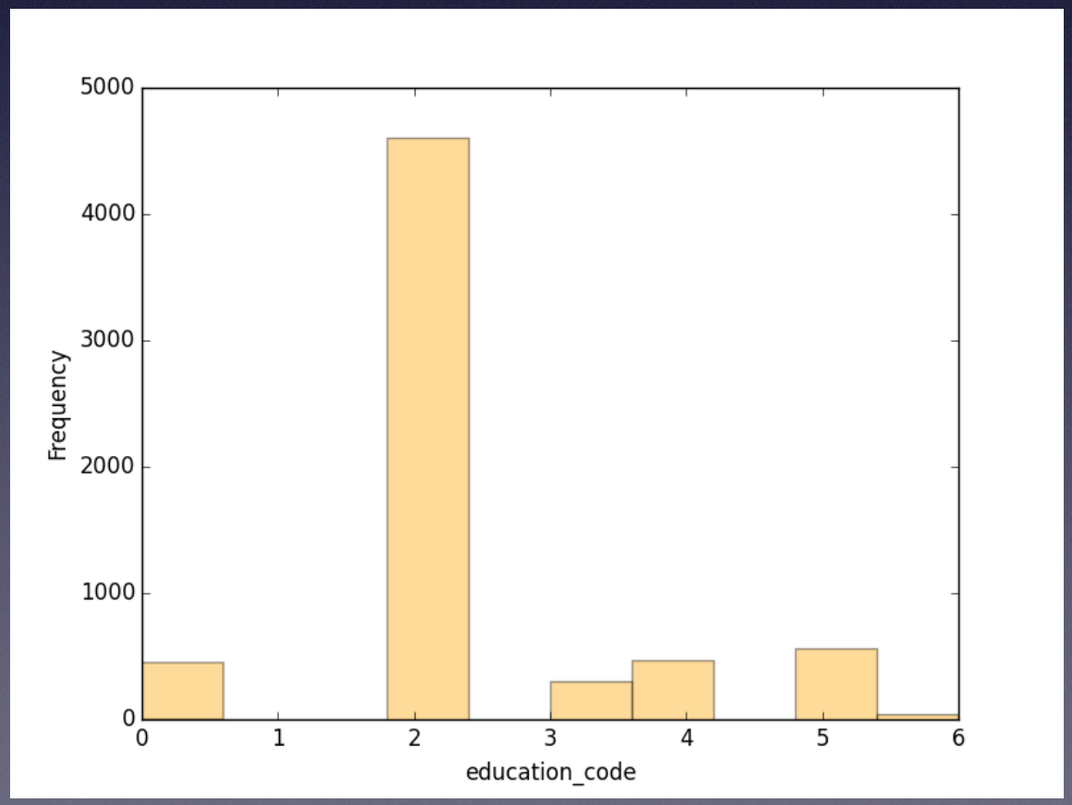

education_code
income
7
6
5
4
3
2
1
0
−1
−100000
0
100000
200000
300000
400000
500000
600000

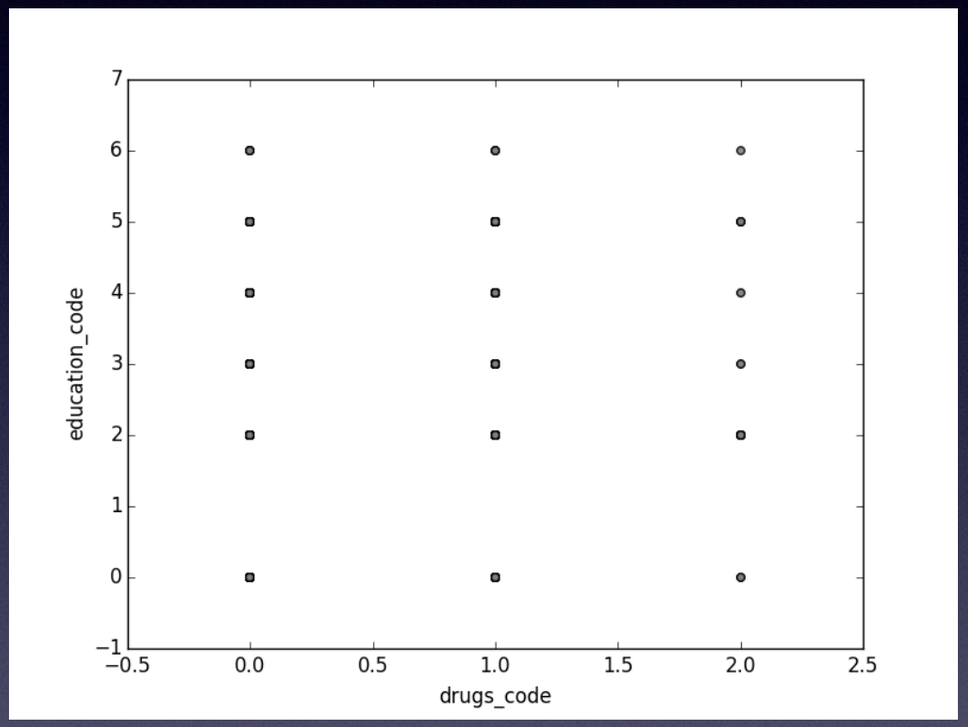

Frequency
income
4000
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0

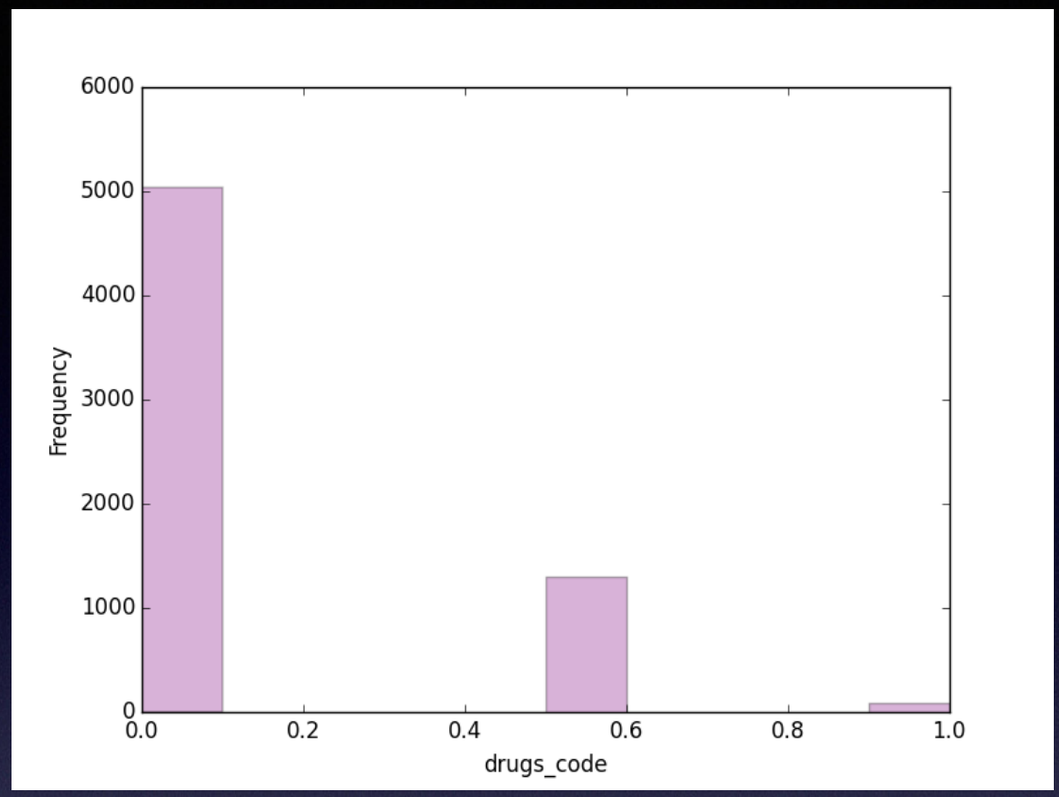

Frequency
education_code
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6

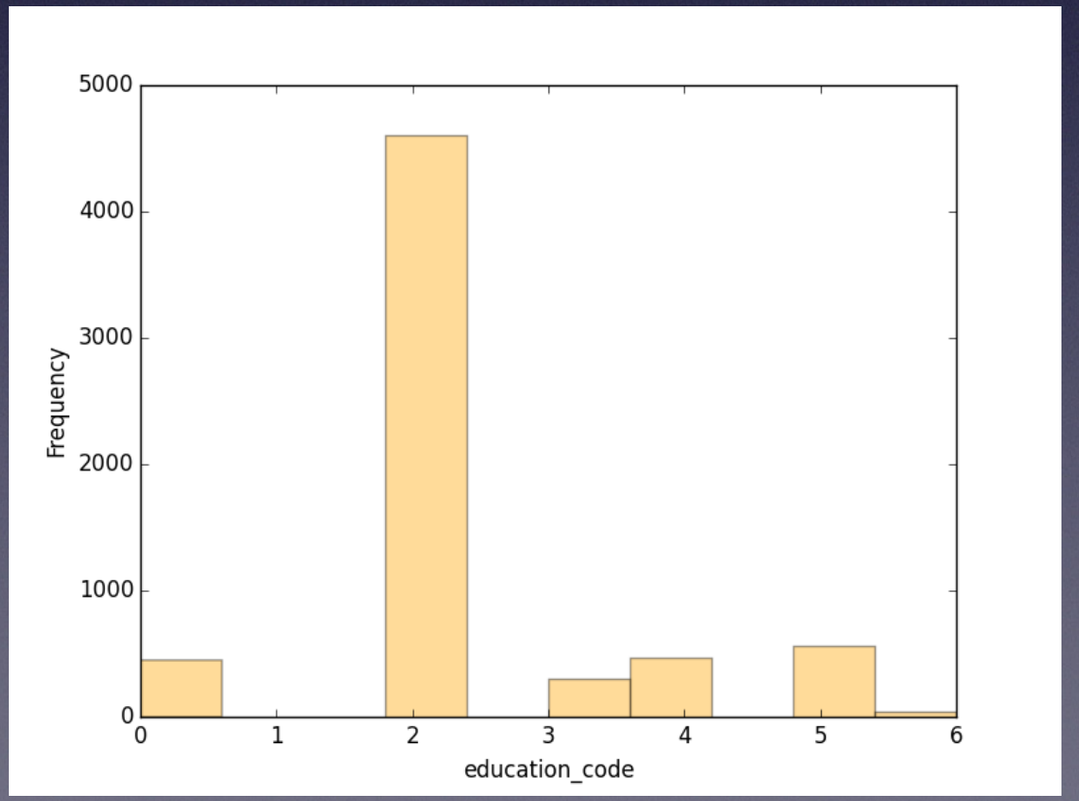

education_code
body_type_code
7
6
5
4
3
2
1
0
−1
−2
0
2
4
6
8
10

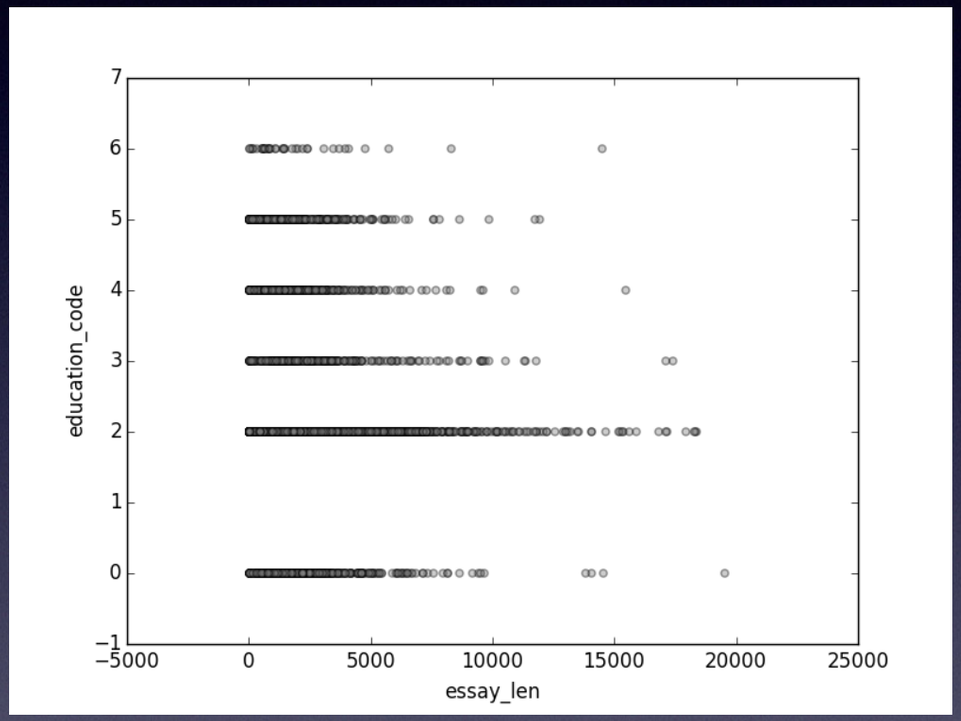

Frequency
body_type_code
1800
1600
1400
1200
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0

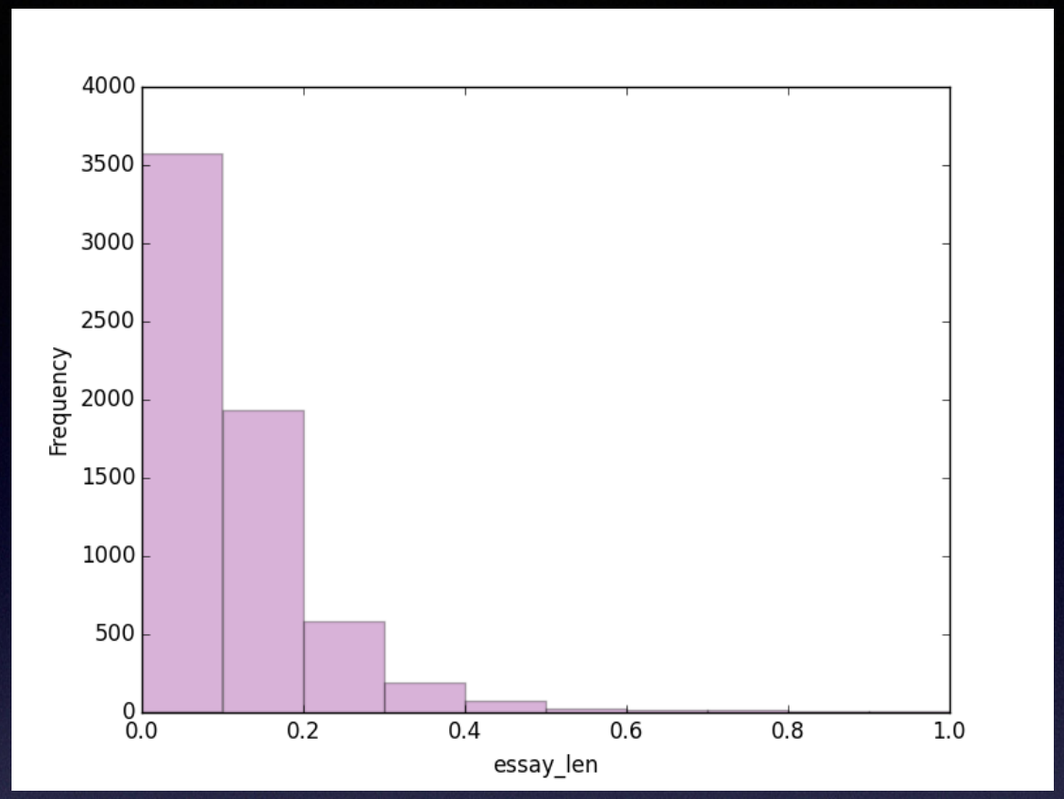

Frequency
education_code
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6

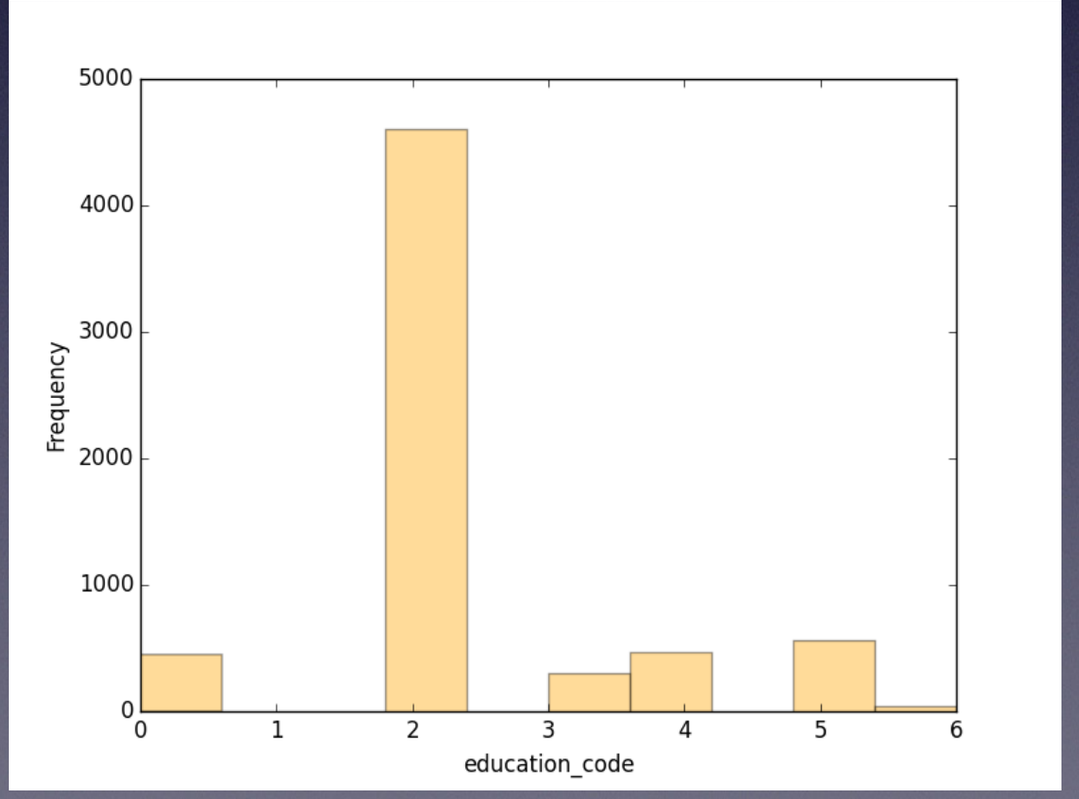

education_code
drugs_code
7
6
5
4
3
2
1
0
−1
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

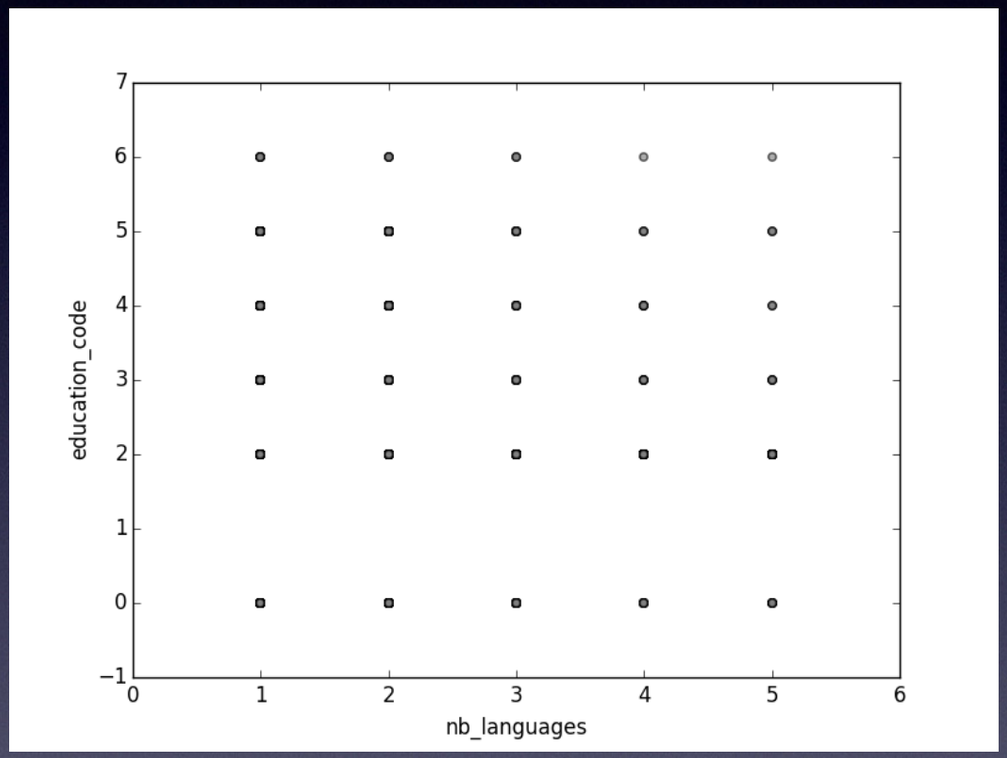

Frequency
drugs_code
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0

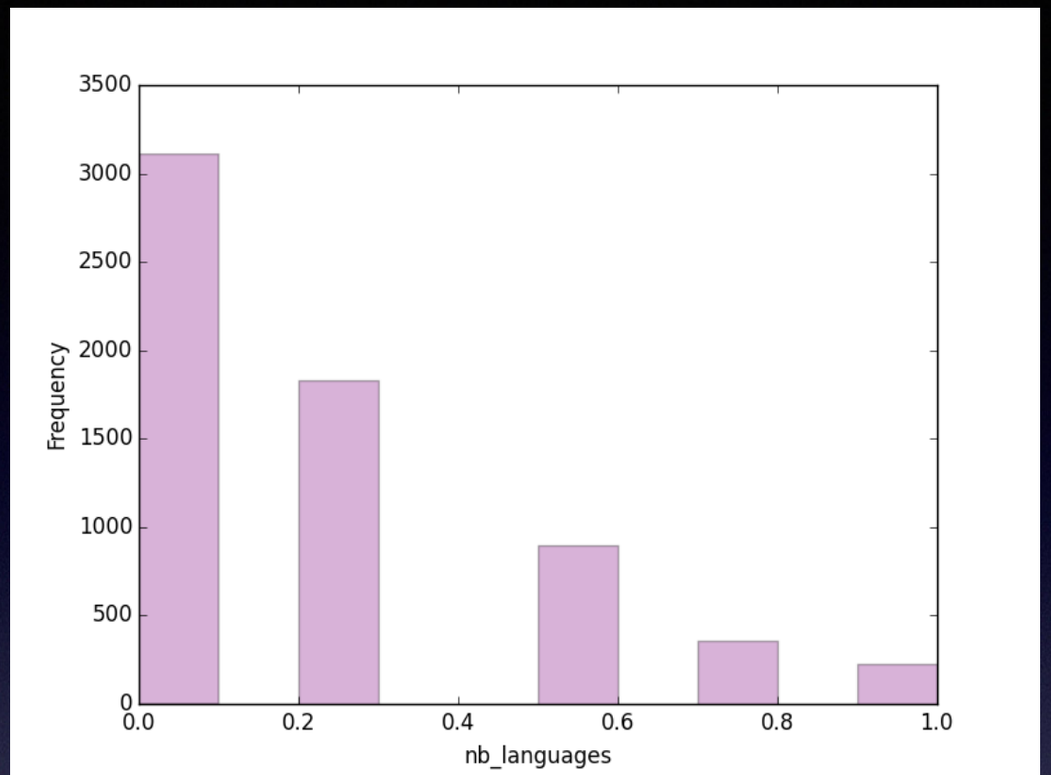

Frequency
education_code
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6

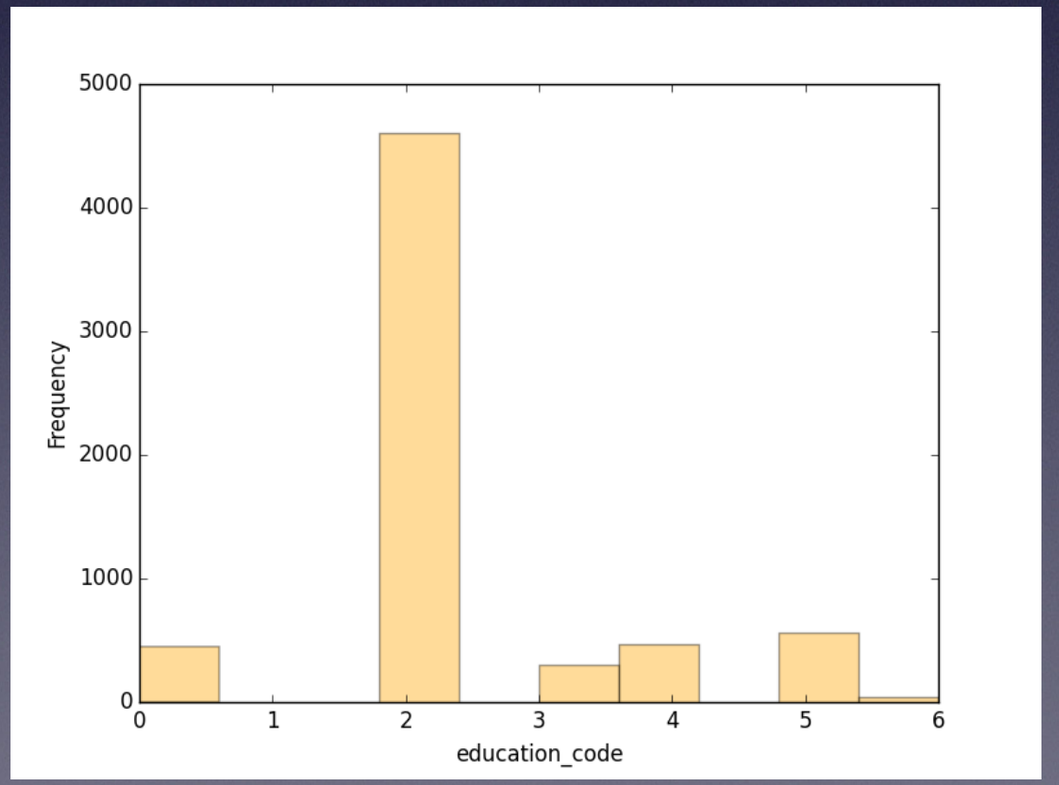

education_code
essay_len
7
6
5
4
3
2
1
0
−1
−5000
0
5000
10000
15000
20000
25000


Frequency
essay_len
4000
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0


Frequency
education_code
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6


education_code
nb_languages
7
6
5
4
3
2
1
0
−1
0
1
2
3
4
5
6


Frequency
nb_languages
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0


Frequency
education_code
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6

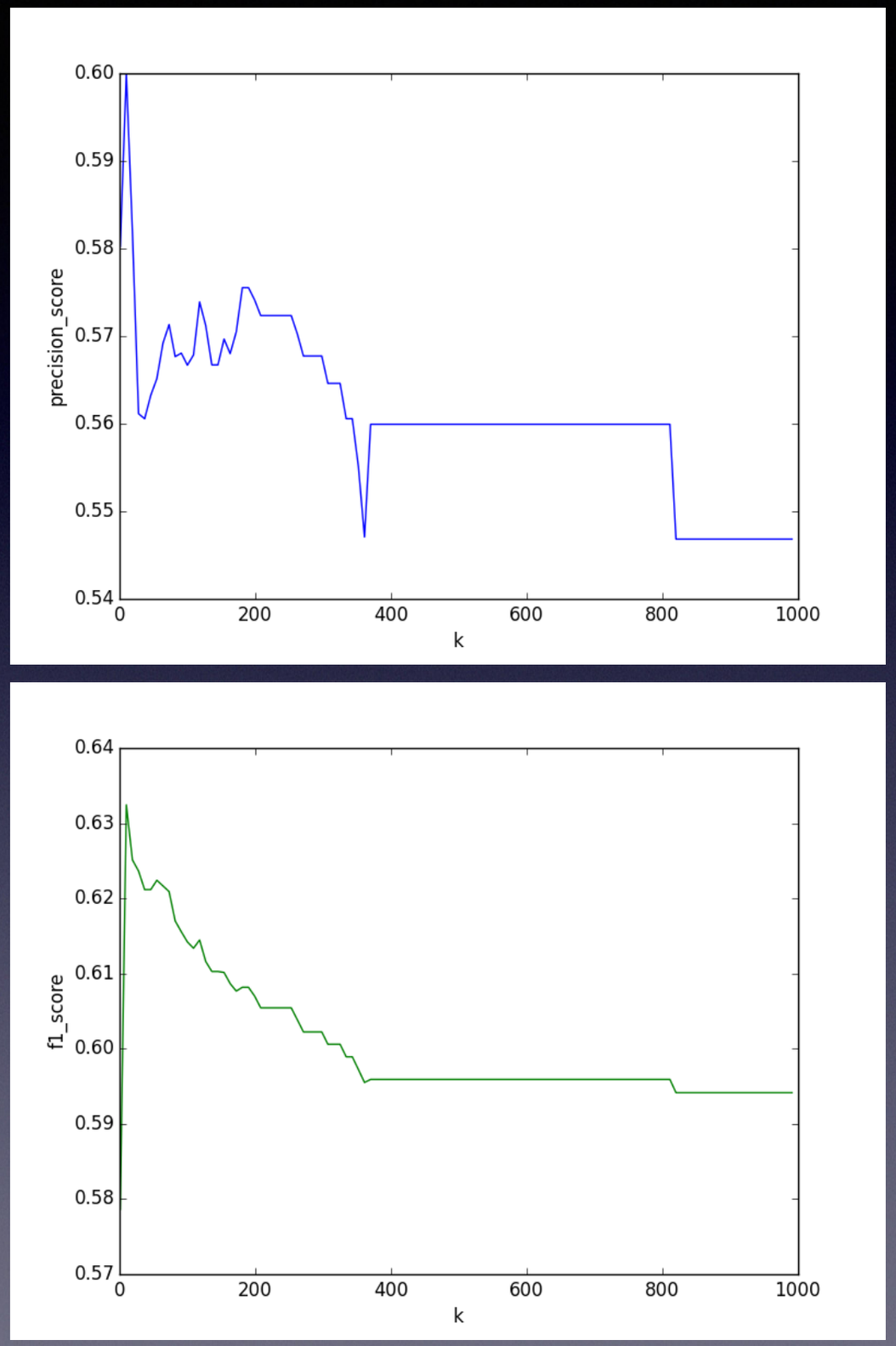
Classification Approaches

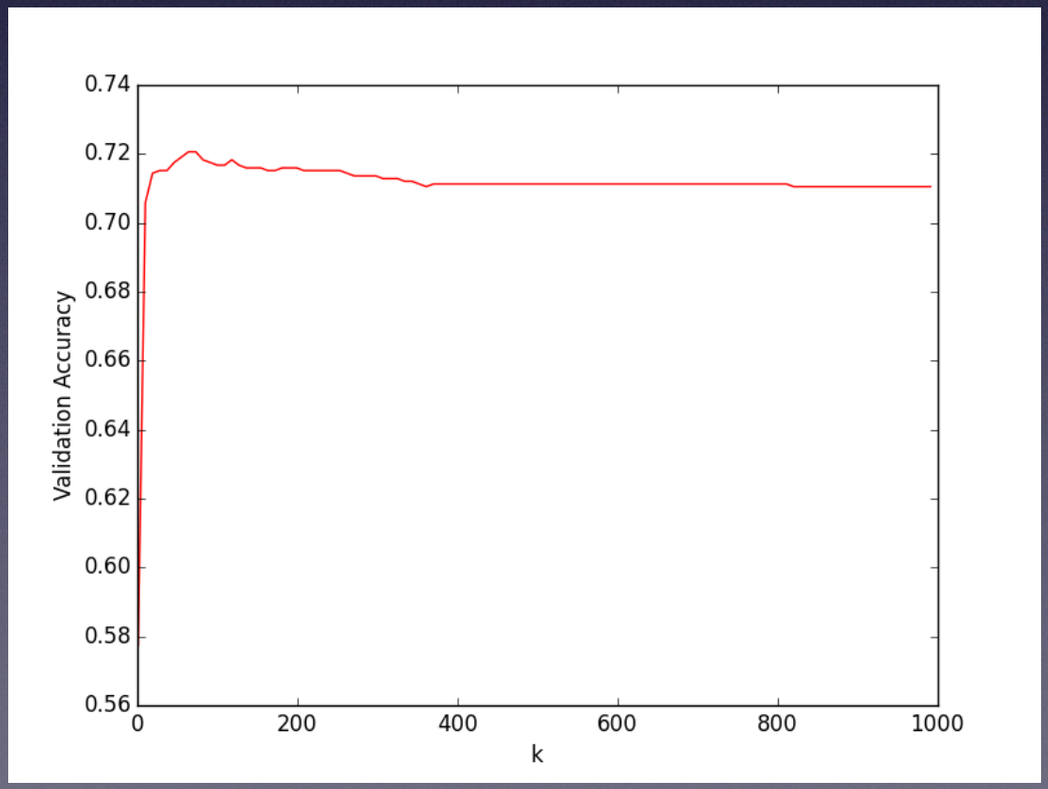
- We used two different classification methods from sklearn:
  - ***KNeighbors Classifier***
  - ***SVM Classifier***

where we used the usual ***Training Set vs Validation Set vs Test Set*** to N-Fold Cross-Validate our results.

Classification Approaches: KNeighbors Classifier

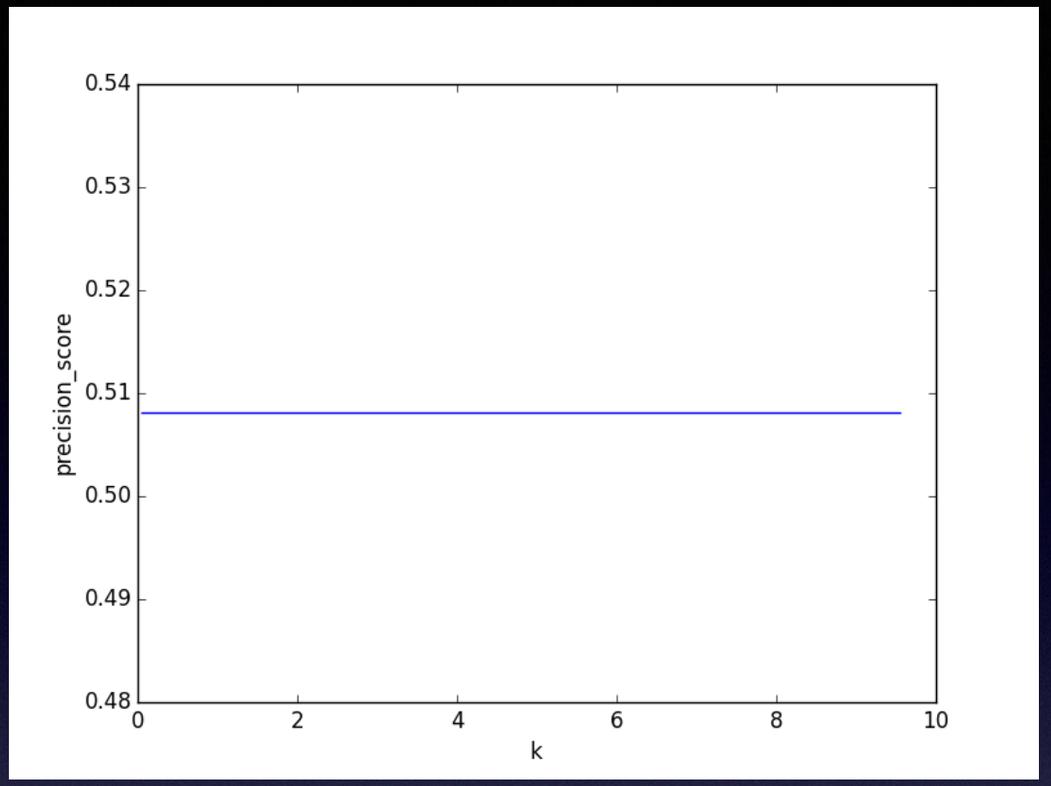
The results are just above average and the ideal range of k likes between [350, 800]

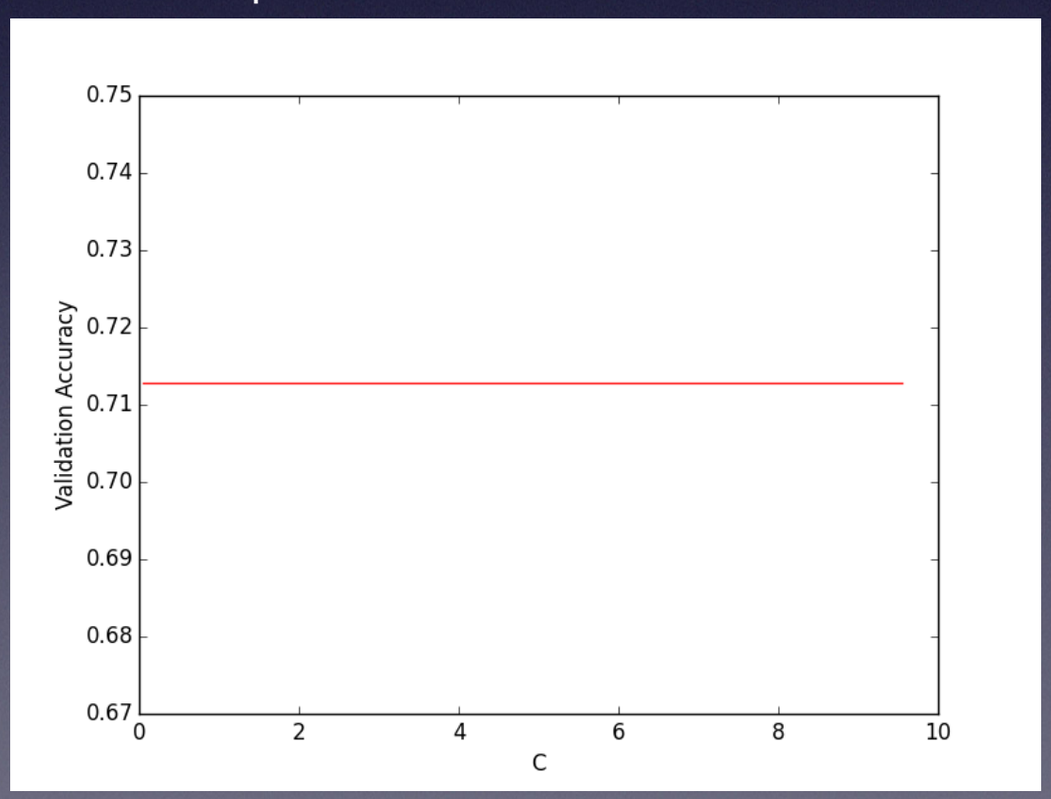


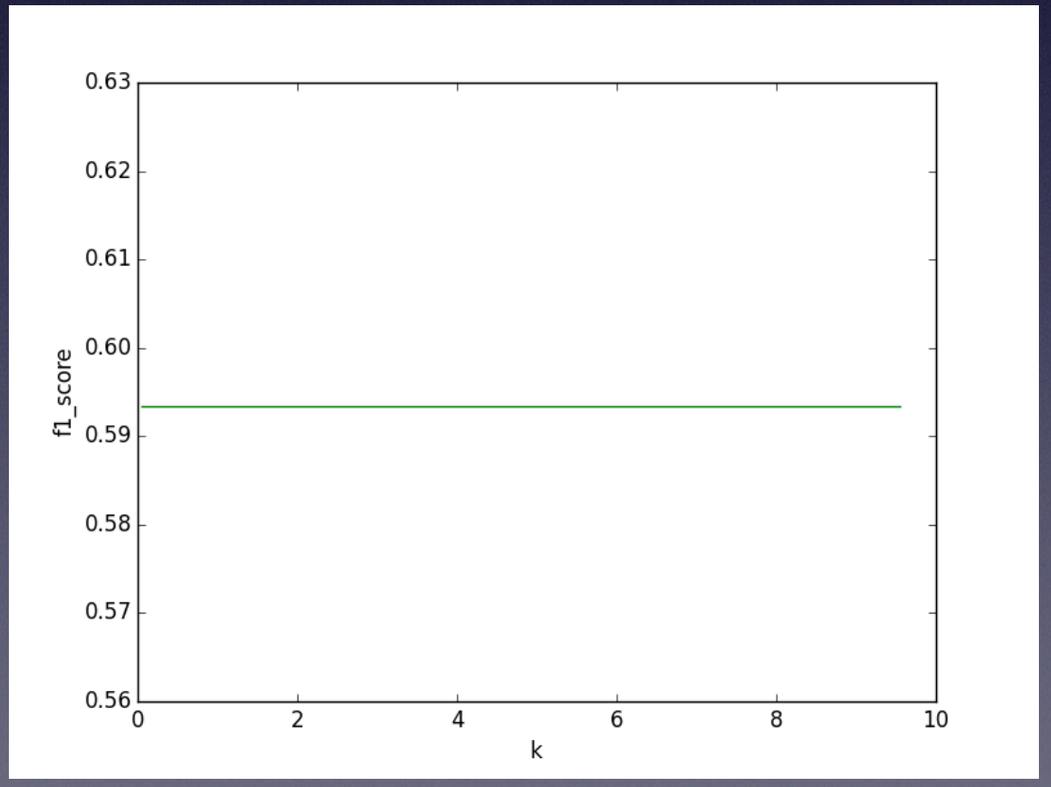


Classification Approaches: SVM Classifier

- The results are very stable and robust with different C parameters.
- We tried also with different gamma, the impact is minimal.







Classification Approaches: comparison of methods

- Both methods have similar structures to call and manipulate and they give very similar results.

- However, there are some structural differences:

- While the ***KNeighbors Classifier*** is intuitive to use and change the parameters, the convergence of the accuracy, scores, … can be visualised easily but can be very slow.

- The **SVM Classifier** is certainly more robust but slower in execution and difficult to visualise. There is a need to use PCA analysis to plot the boundaries in multidimensional analysis.

Conclusion

- Through the results, we can see some links between the education levels and the of self-body awareness and vocabulary sophistication

- However, there is still a need to refine the data and the data subsets or re-choice of variables because the ***f1_scores*** and ***accuracies*** are not at the levels expected

- There is maybe a beed to refine the problematic with more selected individuals: ***ethnicity*** / ***choice of words*** but it needs more advanced procedures (natural language, …)

- Next step would be to explore these possibilities

Conclusion

- The large data analysis is more challenging than it seems
- The data preparation is crucial to define and refine the problematic and obtain significant results
- The choice of variable has been an incremental exercise to obtain significant results
- The visualisation is necessary for more inputs and outputs understanding which leads me to use more advanced tools like PCA, …