Machine learning - Capstone project

Dora Ma

Data

Data from OKCupid, containing the following columns:

age
body_type
diet
drinks
drugs
education
ethnicity
height
income
job
offspring
orientation
pets
religion
sex
sign
smokes
speaks
status

Key
Continuous
Discrete
Categorical

Data also contains short word based answers:

- essay0 - My self summary
- essay1 - What I'm doing with my life
- essay2 - I'm really good at
- essay3 - The first thing people usually notice about me
- essay4 - Favorite books, movies, show, music, and food
- essay5 - The six things I could never do without
- essay6 - I spend a lot of time thinking about
- essay7 - On a typical Friday night I am
- essay8 - The most private thing I am willing to admit
- essay9 - You should message me if…

Combining all the short answers to derive:

Length of essays
No. of words
Frequency of "I", "me", "myself"

Data cleaning & processing

Graduated from…	Mapped to
High school	1
Two-year college	2
College/university	3
Masters program or Law school	4
Ph.D program or Med school	5

age ← Dropped entries beyond 100

body_type
diet
drinks
drugs

education → Focused only on entries that were "graduated from…" for fair comparisons between education levels, then mapped to range 1 to 5

ethnicity

height ← Data entries likely to be in inches, therefore discard entries below 50inches and above 90inches

income ← Too many entries of 1,000,000 – likely to be lying so discard. Also discard "-1" entries

job
offspring
orientation
pets
religion

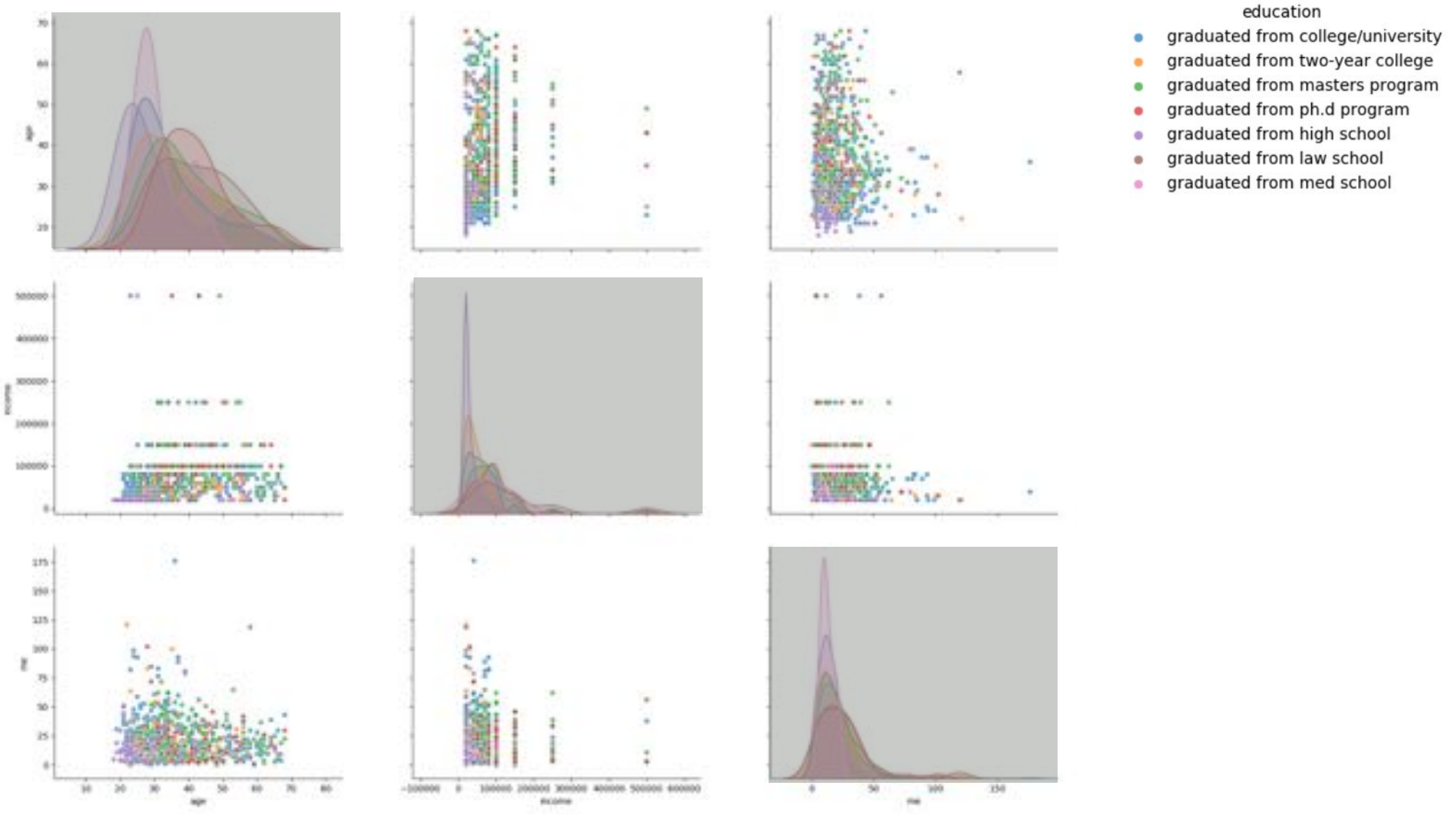
sex → Created new column by mapping m=0, f=1

sign
smokes
speaks
Status

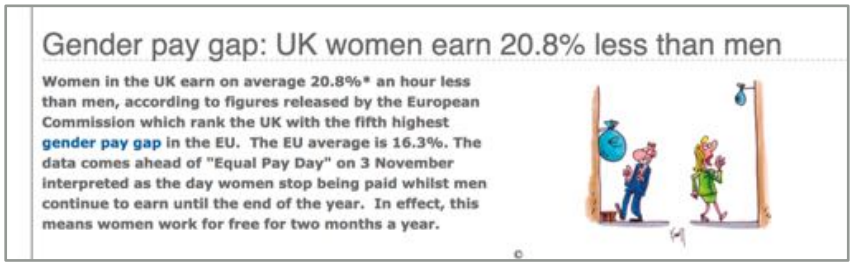
Essays:
Length of answer
No. of words

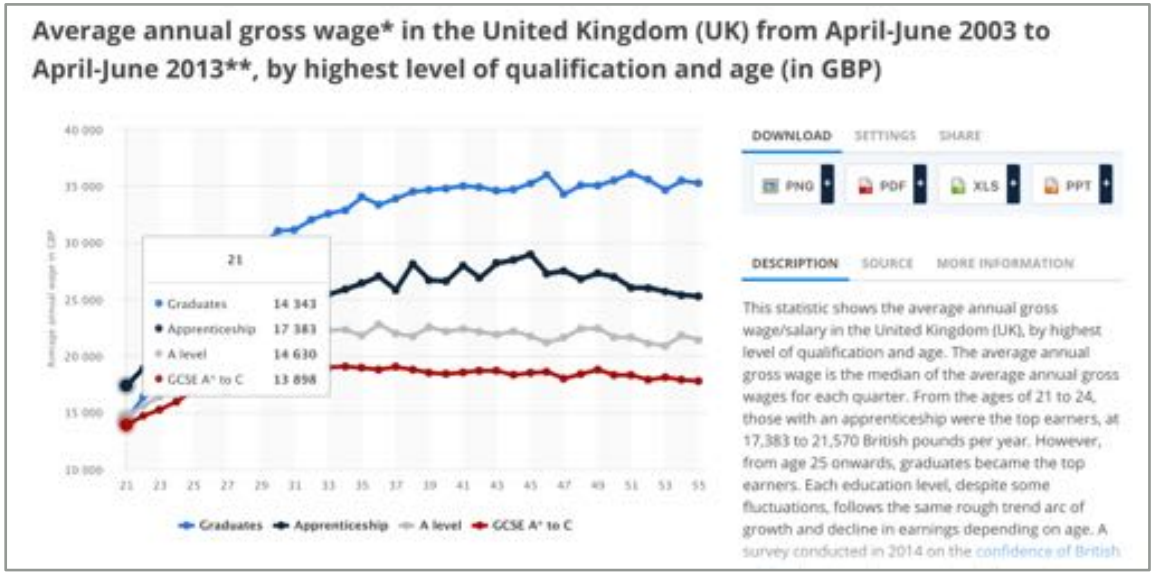
Frequency of "self" related words → Created new column by counting number of occurrence on words: "I", "me", "myself"

Data visualisation



Formulating questions





Classification:

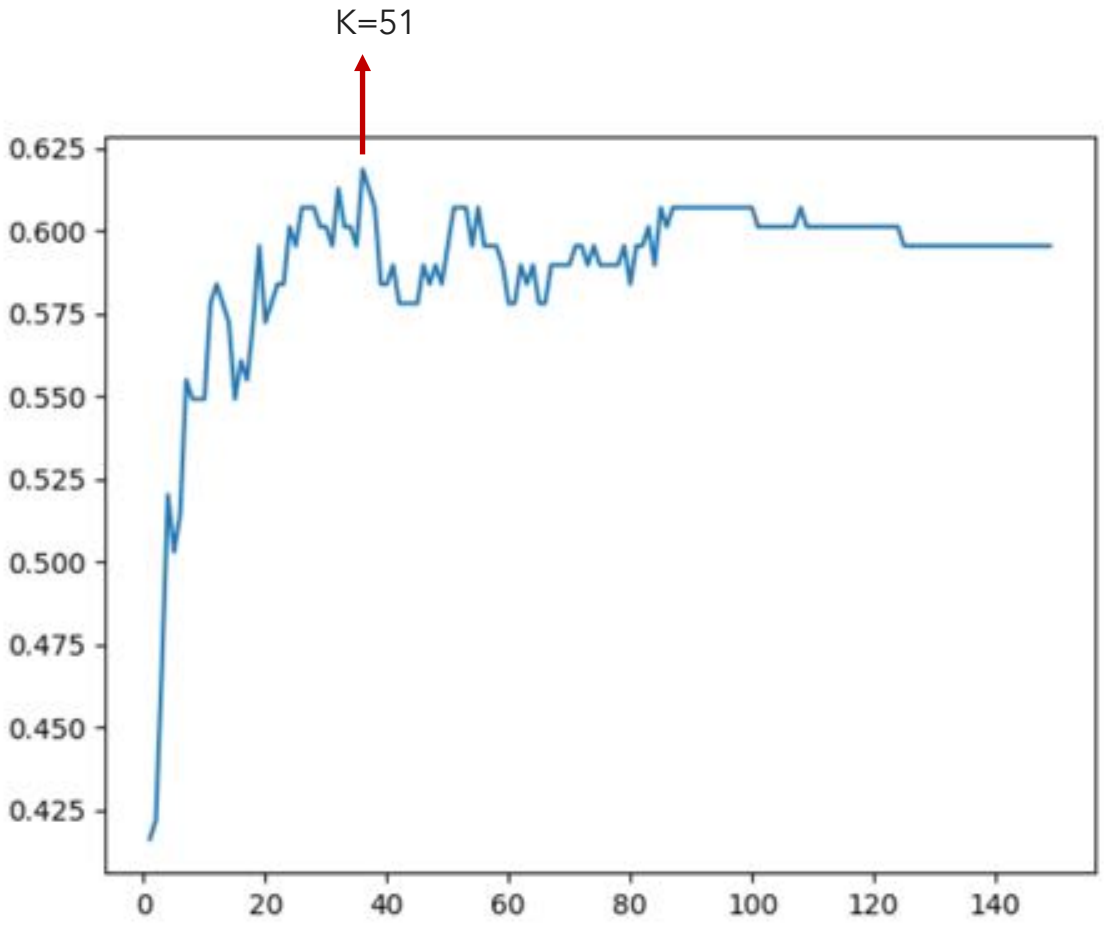
Can we predict level of education based on age, gender, income?

Linear regression:

Can we predict height with gender and income?

Key
Continuous
Discrete
Categorical

Find the optimum k

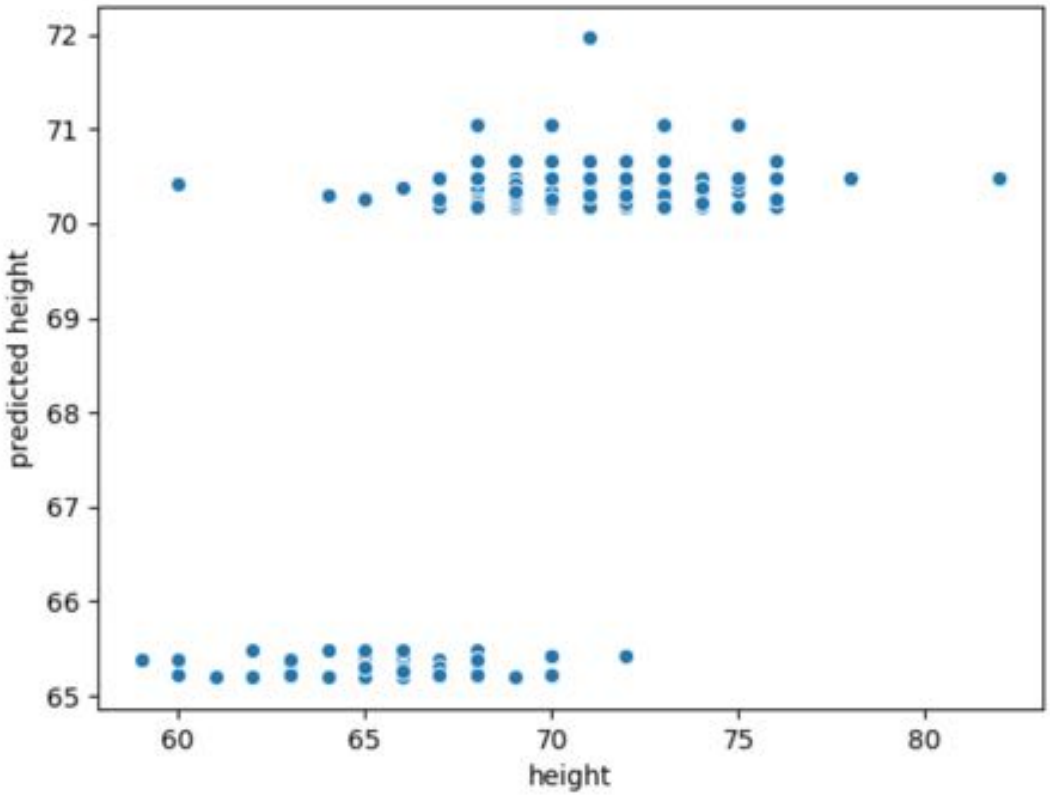


Comparison of Bayes & K-nearest neighbour

	Bayes	K-nearest neighbour	
Time	0.00230	0.00229	← Both methods seem to take around the same time
Accuracy %	29%	63%	← KNN significantly more accurate
Recall			
High school	31%	23%	← KNN do not classify any items in 2 of the output classes at all
Two-year college	57%	0	
College/university	31%	93%	
Masters / law school	15%	24%	
Ph.D / med school	22%	0	
Precision			
High school	18%	75%	
Two-year college	17%	0	
College/university	65%	64%	
Masters / law school	15%	50%	
Ph.D / med school	9%	0	
F1 score			
High school	23%	35%	← KNN generally has higher precision and recall then Bayes
Two-year college	26%	0	
College/university	42%	77%	
Masters / law school	15%	32%	
Ph.D / med school	13%	0	

Visualisations for linear regression

Linear regression model



Inputs = gender, income
Output = height

Accuracy = 39%

Insights & next steps

Can we predict level of education based on income, gender and age?

Depending on the method, there may be potential to use income, gender and age to classify the level of education.

I ran some variations of the model to include average word length or length of short answers, but did not provide improvements in accuracy.

Can we predict height with gender and income?

There appears to be some correlation, but not strong.

Height maybe correlated with ethnicity? And perhaps a stronger correlation for certain age range.

Lessons learnt:

Features engineering and investigating correlations between features requires much time and research.

Not all data within the field should be used to build the model.

Translating categorical data to numerical data requires thinking.

Formulating meaningful questions and building the corresponding model is tricky. For example, frequency of “self” related words can correlate very well with length of essay answers, but do not provide meaningful insights?