Date-a-scientist

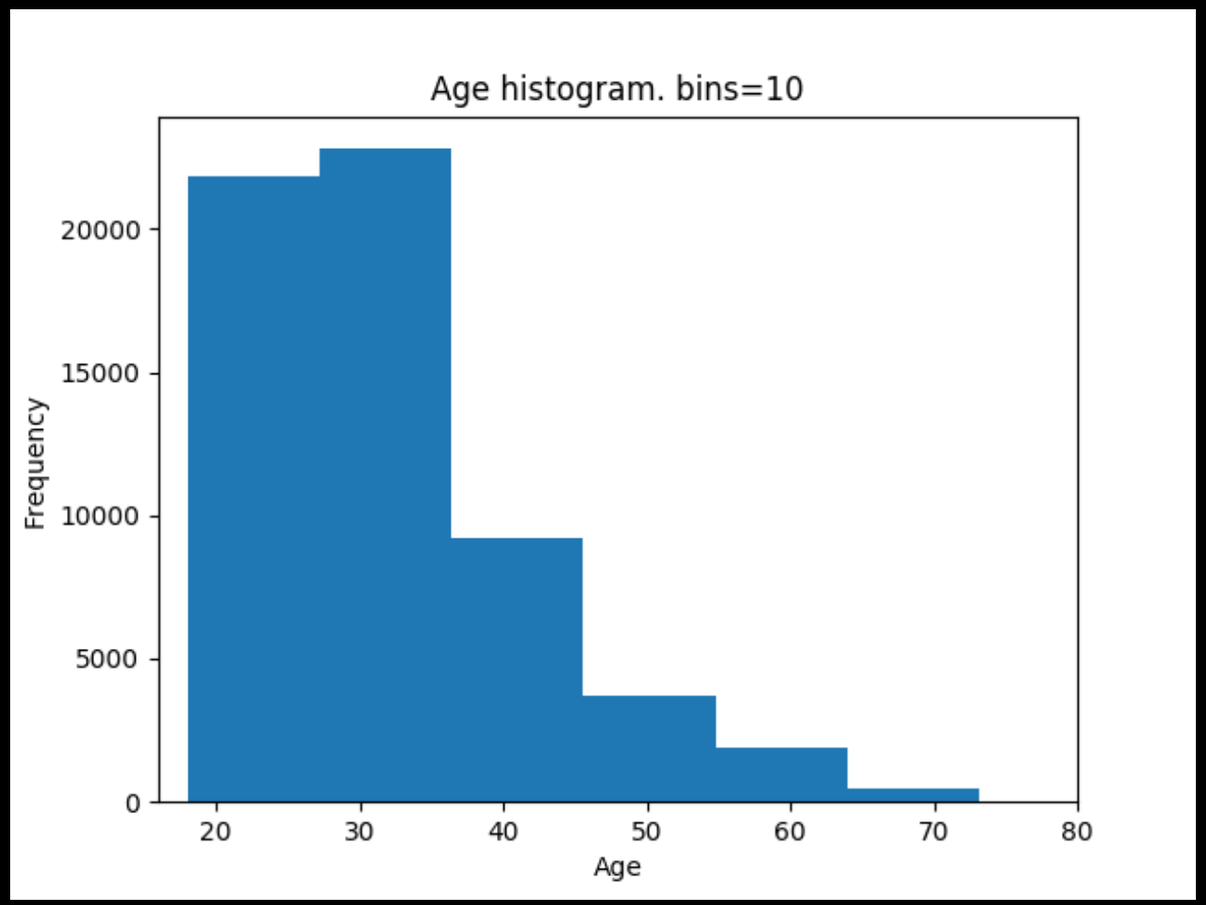
Codecademy - Machine Learning Fundamentals

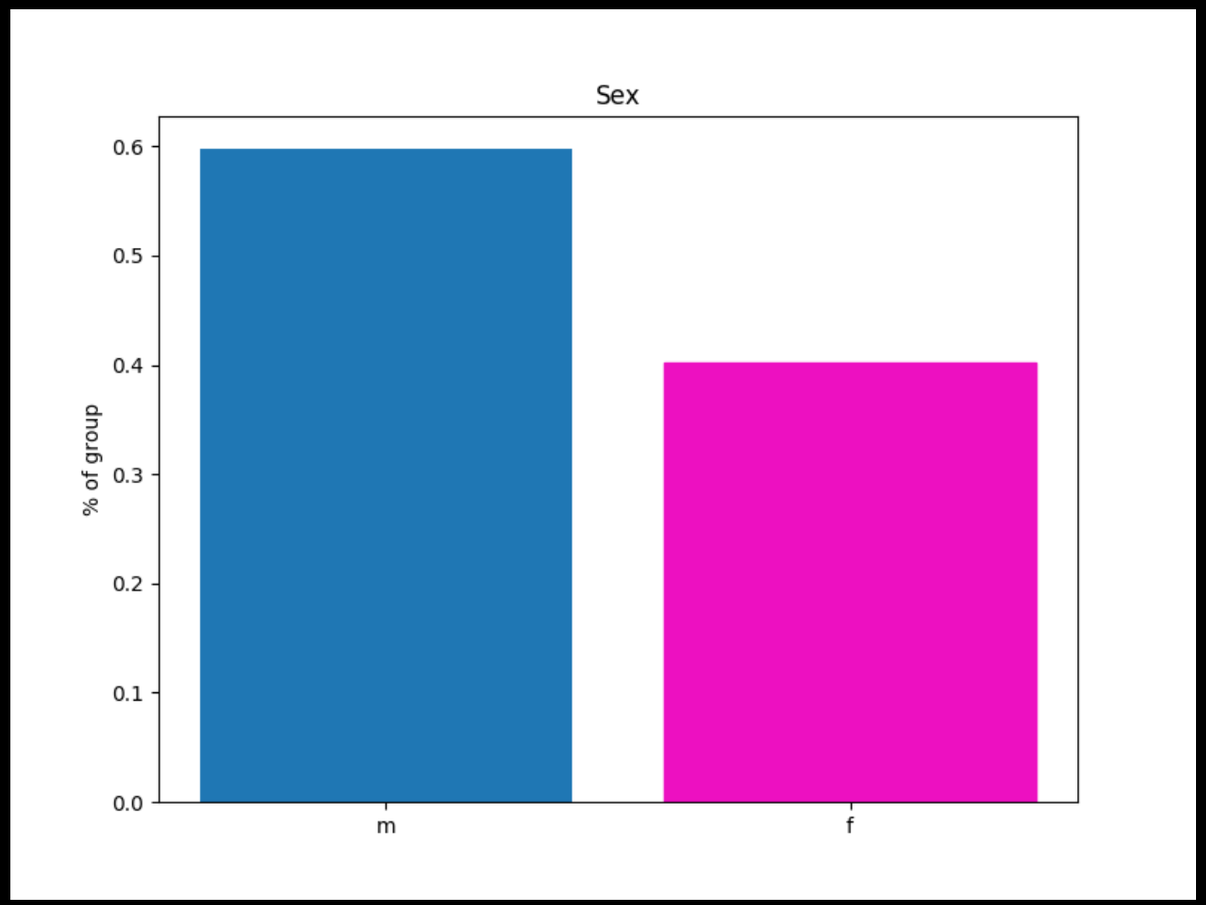
Sahle A. Alturaigi

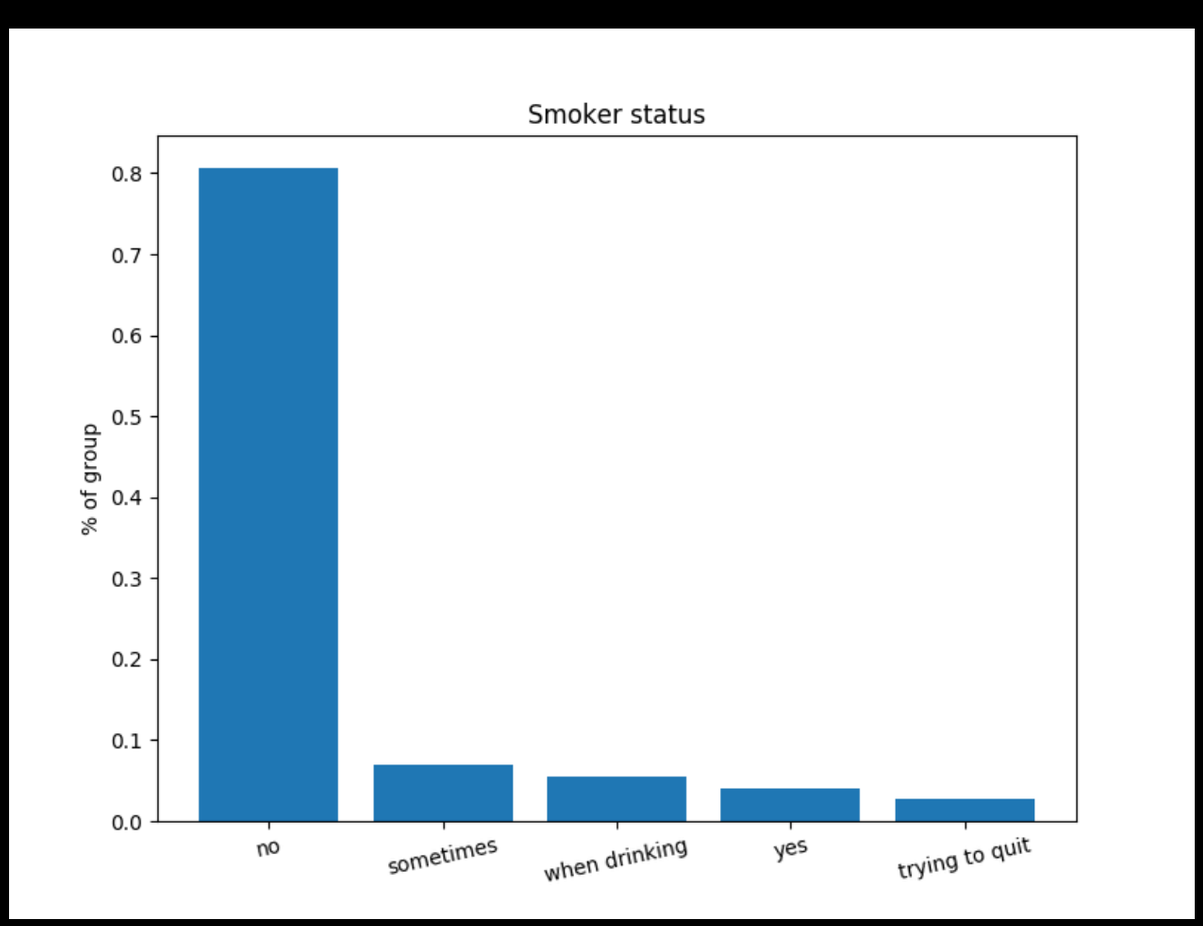
Exploring the data

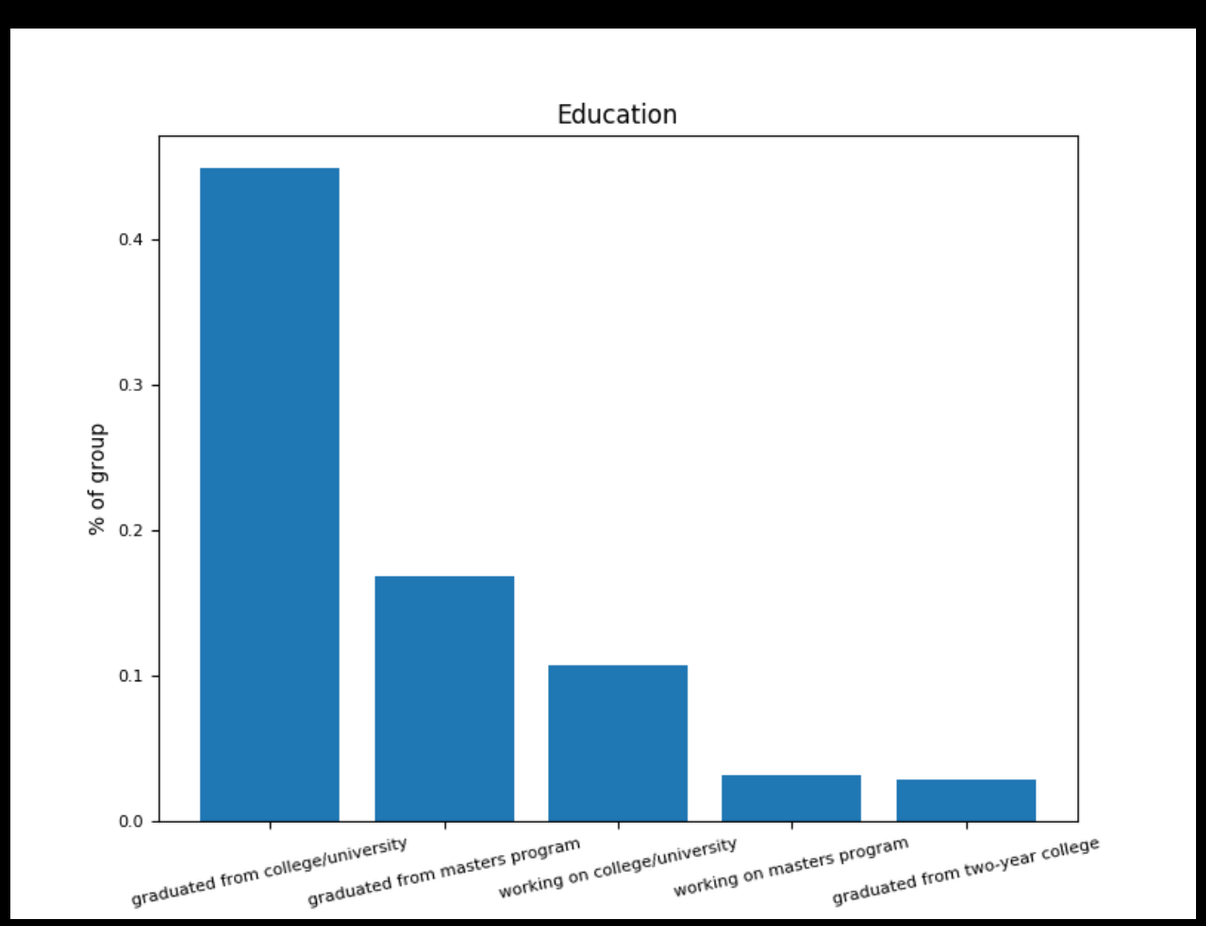
Exploring the data

- Most participants are educated, non-smokers, English speaking, with average to fit body types, diets consisting of any kind of food, in their 20s or 30s.
- There were 59,946 profiles, 60% male, 40% females.
- Numerous attributes and essays were either left blank or given -1, such as average income (>1/2 of profiles had -1).


Age histogram. bins=10
Frequency
20000
15000
10000
5000
0
20
30
40
50
60
70
80
Age


Sex
% of group
0.6
0.5
0.4
0.3
0.2
0.1
0.0
m
f


Smoker status
% of group
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
no
sometimes
when drinking
yes
trying to quit


Education
% of group
0.4
0.3
0.2
0.1
0.0
graduated from college/university
graduated from masters program
working on college/university
working on masters program
graduated from two-year college

Question Statement

- I use to be a social smoker, but curbed the habit to the point where I smoke perhaps a handful of times in a year.

- Naturally, I became interested in seeing what attributes or topics (essay 6, “I spend a lot of time thinking about”) had a positive correlation in smokers vs. non-smokers.

- To keep things relatively simple, I:
  - Ran two separate classification models on essay 6
  - Implemented two separate regression models on *age, income,* and *body_type,* but found that the too few profiles had income data so I dropped that attribute
- For fun, asked my friends if they could tell me phrases only smokers or non-smokers would say and see if my classifier could guess correctly.

Assumptions

- I had to take some shortcuts to try and make this work since our dataset really captures mostly the educated, healthy, type of individual. About 80% of profiles were non-smokers.

- Assumption: Anyone who smokes, even those trying to quit, will be categorized as **smoker**. Anyone who has listed "no" to smoking is a **non-smoker**.

Augmenting Data

- I created a new boolean column called “is_smoker” which indicated if a person is a smoker or not based on aforementioned my assumption.

- Later on, in the regression model, I encoded the different body_types into a dictionary.

Classification

Topics for smokers

- This section took the most of my time.
- Firstly, I ran a Naive Bayes' Classifier over all the non-NaN essays in essay 6 "I spend a lot of time thinking about".
- Since there were many non-smokers, I took a random subset of non-smoker essays from essay 6 that is equal in size to the smoker's set.
- Unfortunately, the results were far from stellar.

Accuracy	~0.55
Precision	~0.56
Recall	~0.47
F1 Score	~0.51

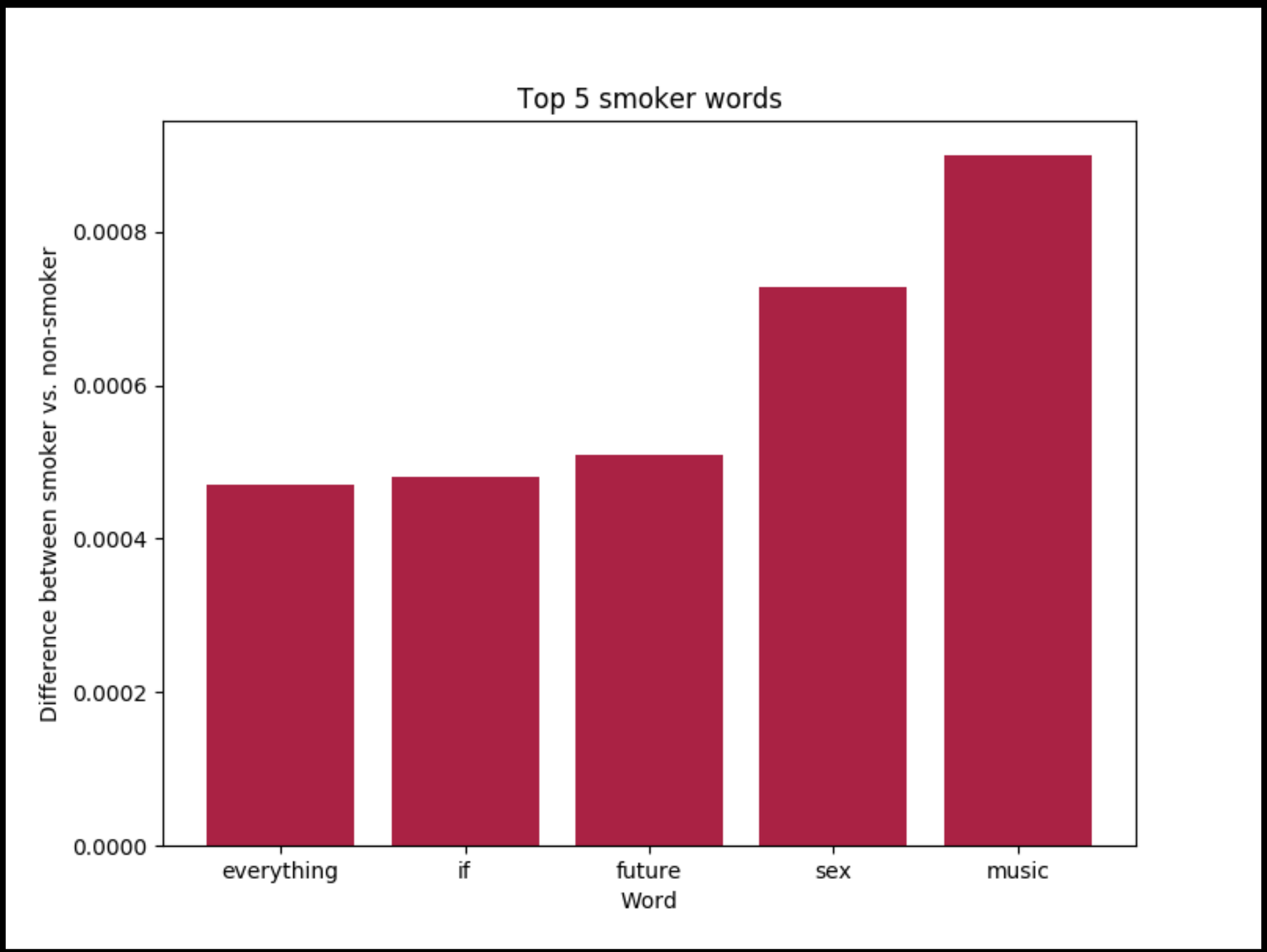
Topics for smokers

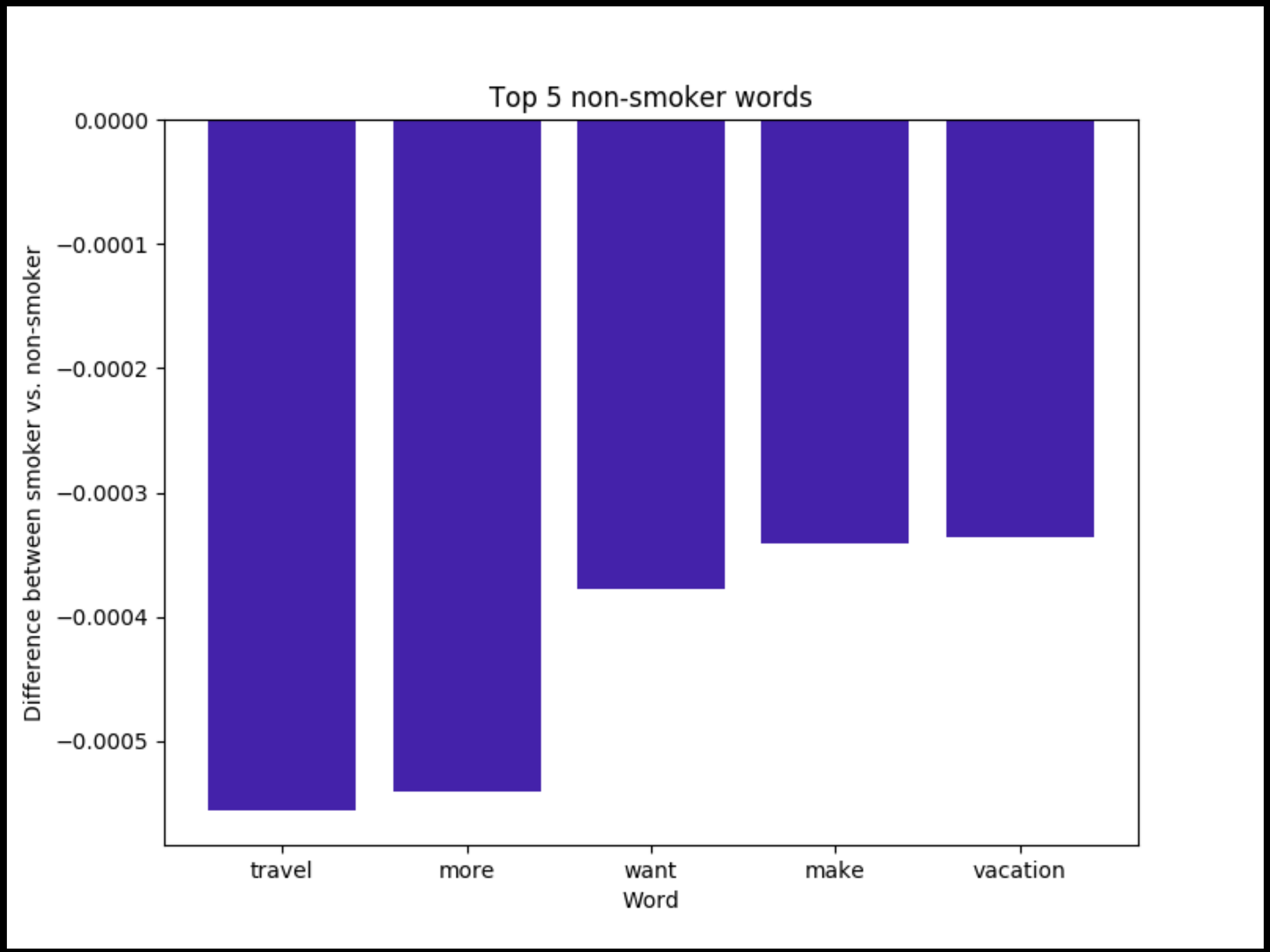
- Since each run uses, virtually, a different subset of non-smoker essay data, the numbers change slightly, but not by more than 1-3%.

- Having said that, certain phrases my friends gave me yielded favorable results from the classifier. Most of my friends agreed with the results too.

- Next slide is a small sample of the ones I found humorous.

Phrase	Classification
“You got a lighter?”	Is a smoker 🚬
“I gotta cut down”	Is a smoker 🚬
“Heaven or Hell”	Is a smoker 🚬
“Kids”	Not a smoker 🚭
“Travel”	Not a smoker 🚭


Top 5 smoker words
Difference between smoker vs. non-smoker
0.0008
0.0006
0.0004
0.0002
0.0000
everything
if
future
sex
music
Word


Top 5 non-smoker words
0.0000
−0.0001
−0.0002
−0.0003
−0.0004
−0.0005
Difference between smoker vs. non-smoker
travel
more
want
make
vacation
Word

Topics for smokers

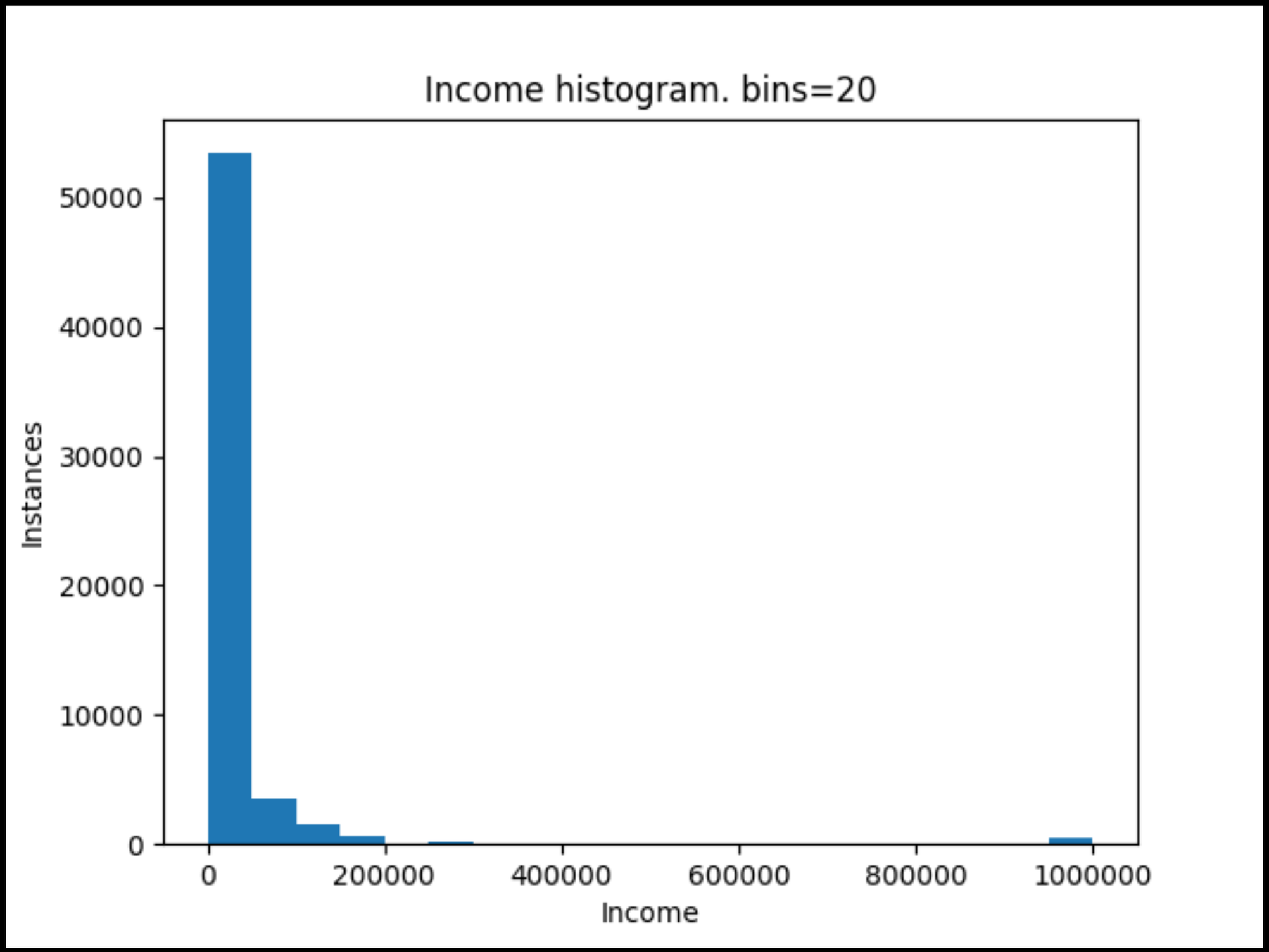
- Overall, the NBC didn’t perform well, and same goes for the Support Vector Classifier, with some caveats:
    - Firstly, the SVC did score better (about 1-5% higher on average compared to the NBC).
    - Secondly, I could NOT run the SVC overall all the data due to memory issues on my machine. As a workaround, I took random samples of 2,000 words and ran this process 10 times over. Then I took the average scores of those 10 iterations.
    - The average F1 score on my last run was: ~0.545

Regression

Age and Body

- Using *age* and *body_type* I wanted to see the regressor could predict if someone was a smoker or not.

- I wanted to include income as well, but there were too few profiles that had given their income.

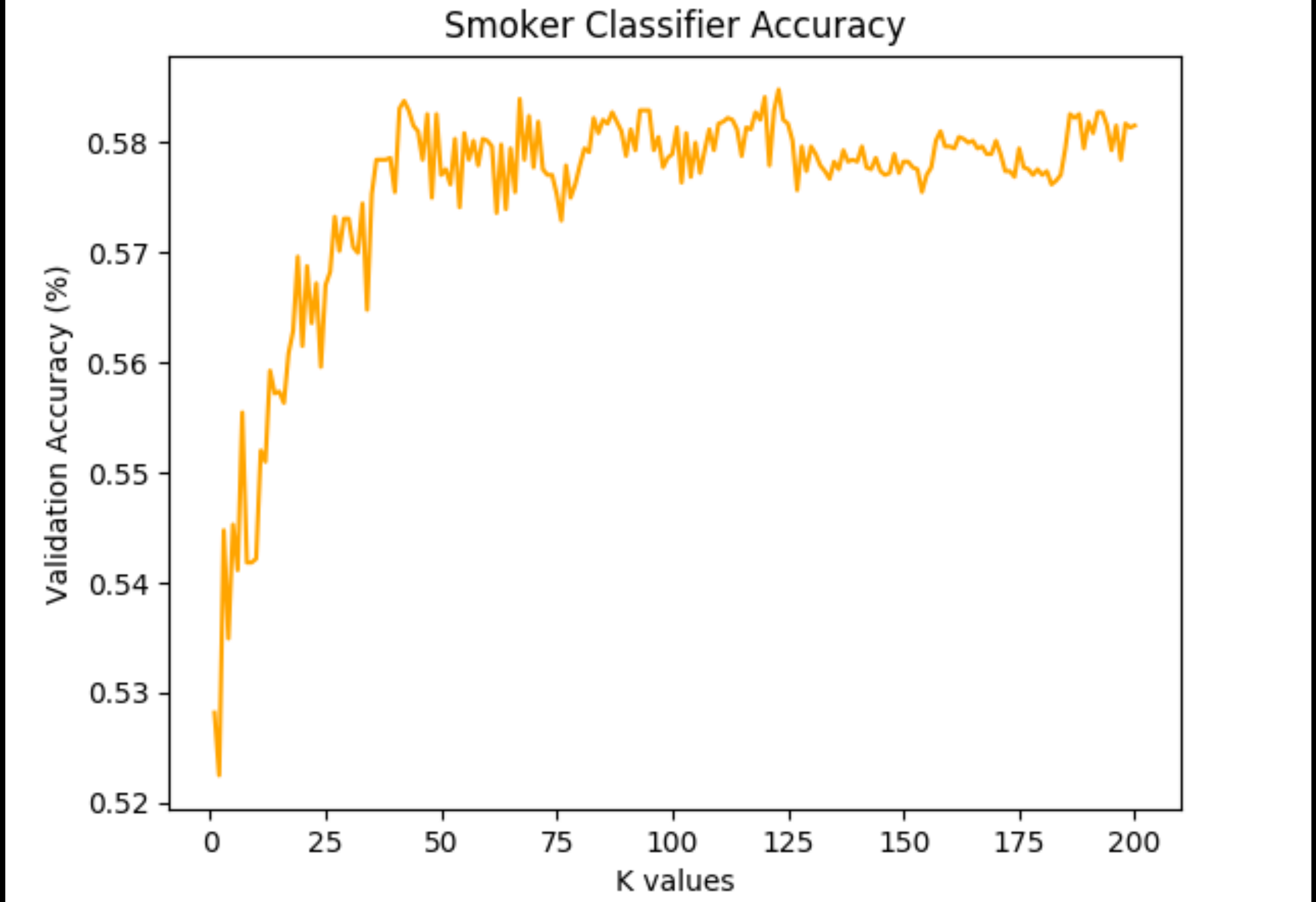

Income histogram. bins=20
Instances
50000
40000
30000
20000
10000
0
0
200000
400000
600000
800000
1000000
Income

- For the Multi-Linear Regression, the R2 score averaged out to about 0.04 for the training set and 0.037 for the testing set

- For K-Nearest Neighbors, hardly performed better either. It's accuracy rating was, at its highest, ~59%

- Best K value fluctuated quite a bit. Had instances of 44, 123, and 103 being the best k value.

- It appears that both models fair poorly at explaining the variability in the data and therefore are almost as good as flipping a coin to guess whether or not someone was a smoker based on their age and body type.


Smoker Classifier Accuracy
Validation Accuracy (%)
0.58
0.57
0.56
0.55
0.54
0.53
0.52
0
25
50
75
100
125
150
175
200
K values

Conclusions

- The dataset doesn't encompass enough different kinds of people.

- Most profiles appear to be of educated, mid 20-30s, healthy people. Naturally, our data set steers quite close to only a certain group of people. Therefore, if we wanted a more accurate model for predicting someone's likelihood of smoking based off of certain attributes, we would more data on people of different age groups, different backgrounds, different education levels, different diets, etc…

- As of now, all I've managed to get working is a classifier that can reasonably guess phrases common smoker phrases

Thank you