PREDICTING SEX BASED ON INFORMATION IN OK CUPID USER PROFILES

MICHELLE A CALER

2021 MARCH 29 (UPDATED 2021 MAY 14)

OUTLINE

1. Main Questions
2. The Data
3. Overview of All Users
4. Associations of Features with Sex
5. Supervised Machine Learning Models
6. Conclusions

MAIN QUESTIONS:

1. Can sex be predicted using a multinomial naïve Bayes classifier trained on OK Cupid user essay texts?

2. Can sex be predicted using a machine learning algorithm trained on age, height, drinking habits, drug use habits, smoking habits, essay lengths, and average lengths of words in essays?

THE DATA

THE DATA:

One .csv file containing:

age	essay0	essay6	income	pets	status
body type	essay1	essay7	job	religion	
diet	essay2	essay8	last online	sex	
drinks	essay3	essay9	location	sign	
drugs	essay4	ethnicity	offspring	smokes	
education	essay5	height	orientation	speaks	

59,946 users

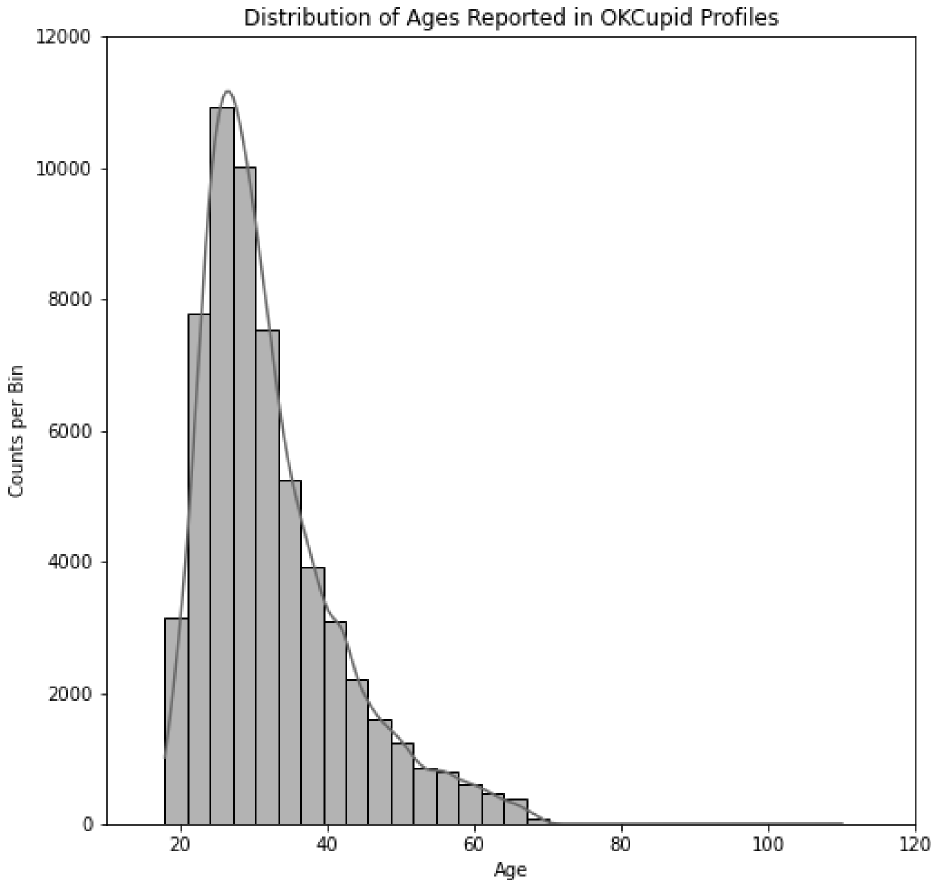
all located in the San Francisco area

Essay prompts:

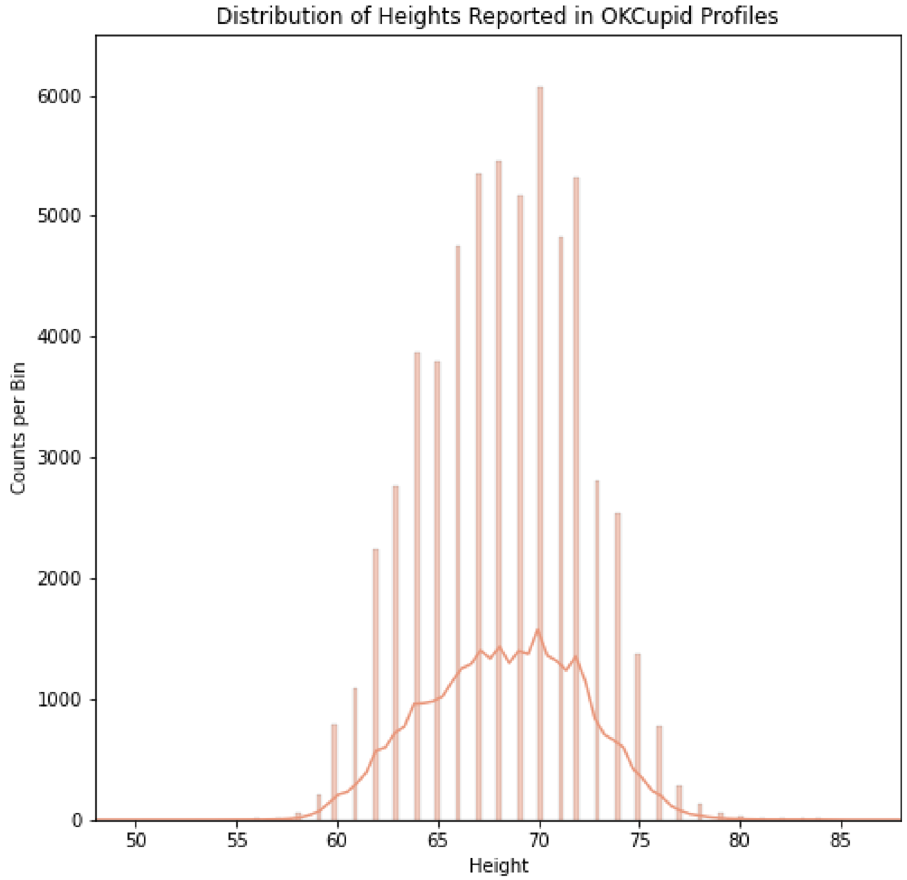
essay0 - My self summary
essay1 - What I'm doing with my life
essay2 - I'm really good at
essay3 - The first thing people usually notice about me
essay4 - Favorite books, movies, show, music, and food
essay5 - The six things I could never do without
essay6 - I spend a lot of time thinking about
essay7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

OVERVIEW OF ALL USERS

AGE AND HEIGHT DISTRIBUTIONS, ALL USERS

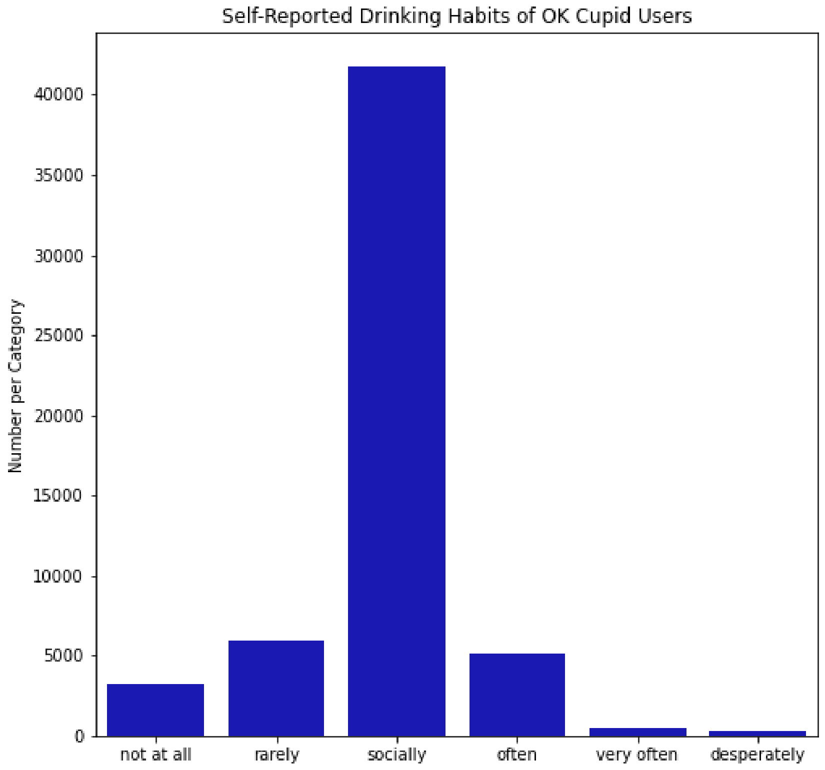


Skew right, multimodal?

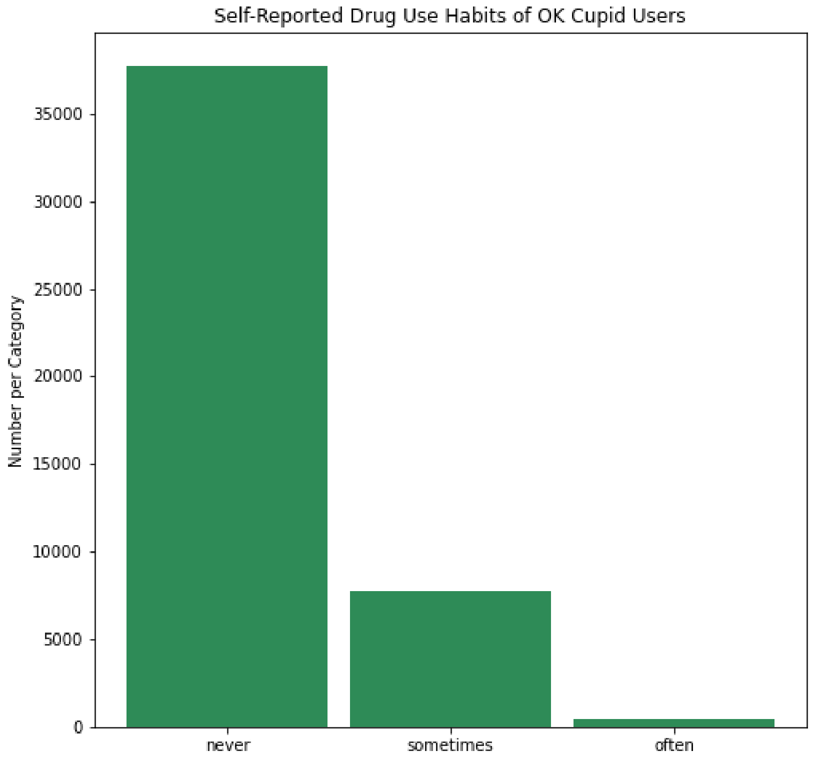


Approx. Normal Distribution

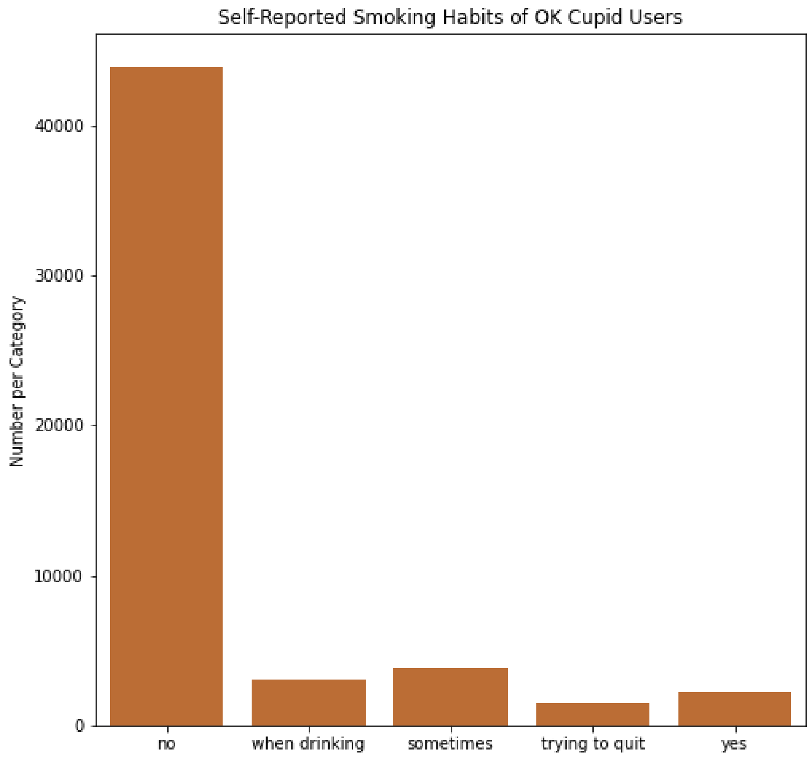
DRINK, DRUG, SMOKE HABITS, ALL USERS



Majority social drinkers



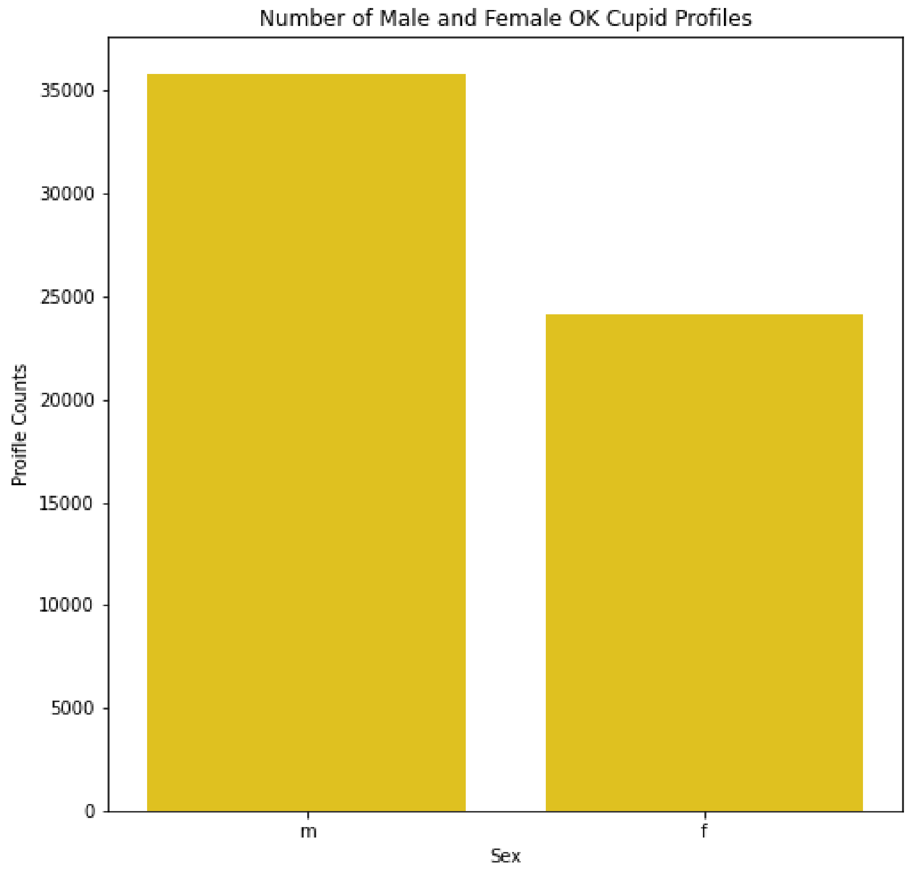
Majority no drug use



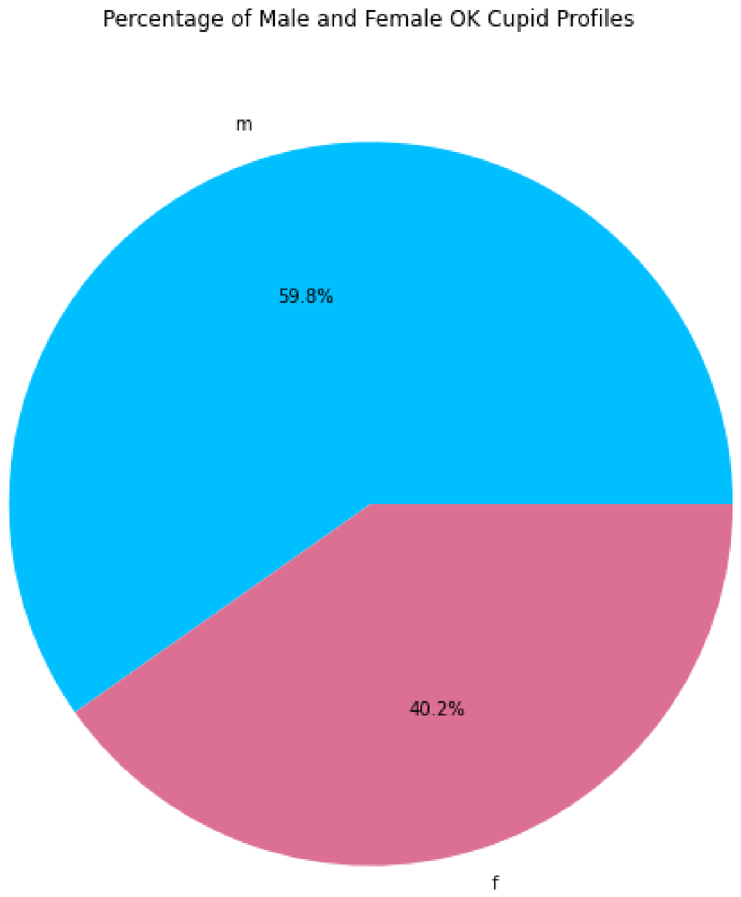
Majority non-smokers

ASSOCIATIONS OF FEATURES WITH SEX

FRACTION OF USERS OF EACH SEX

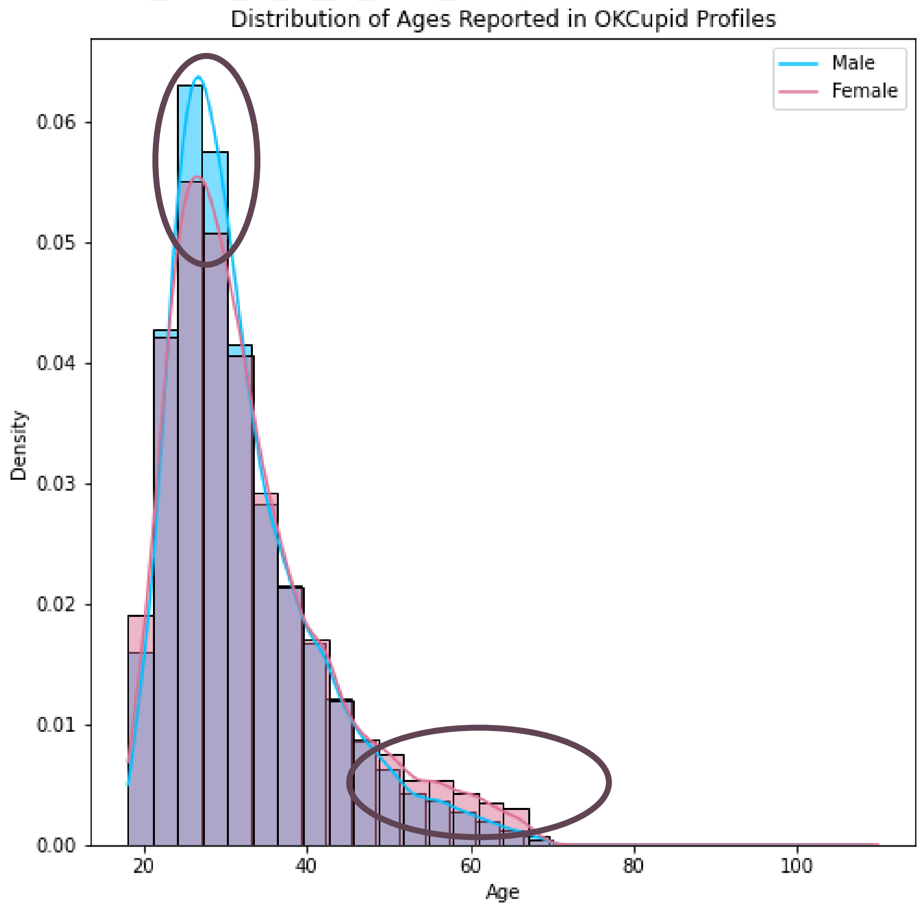


Raw Counts

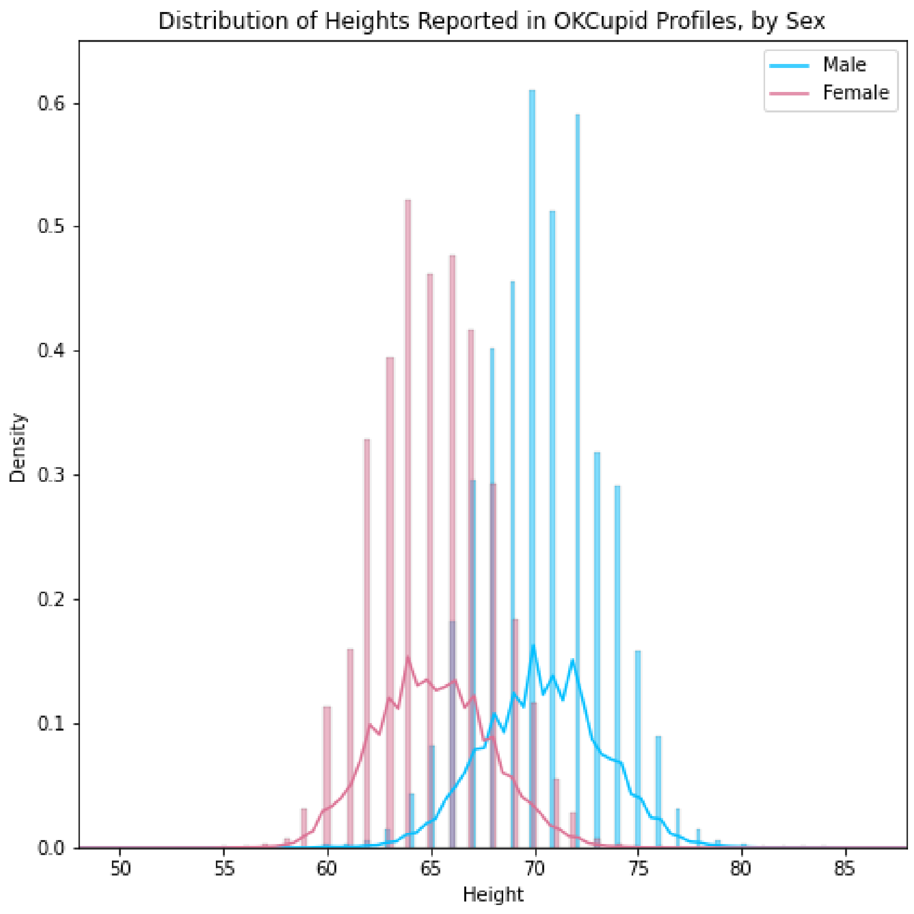


Percentage Breakdown

AGE AND HEIGHT DISTRIBUTIONS, SPLIT BY SEX

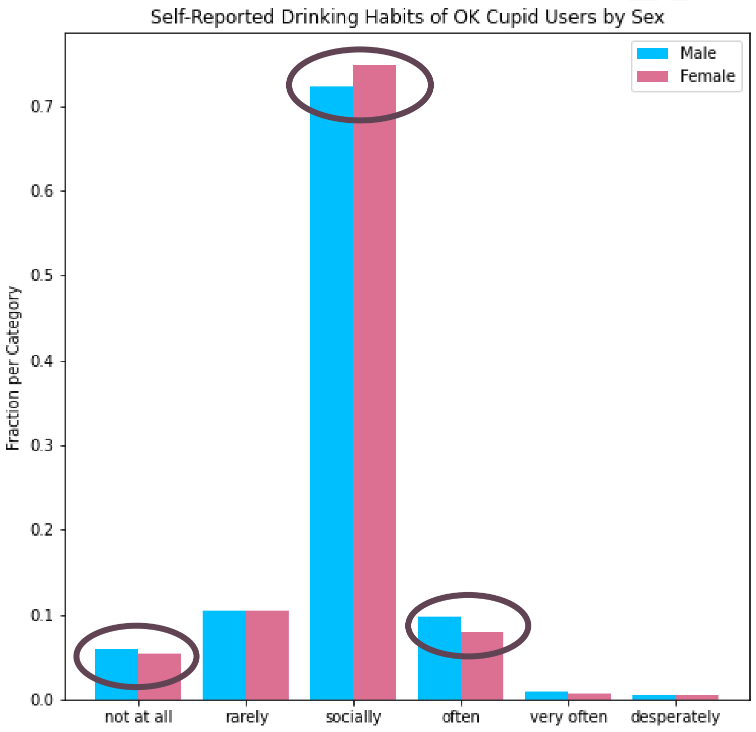


Peak Differences

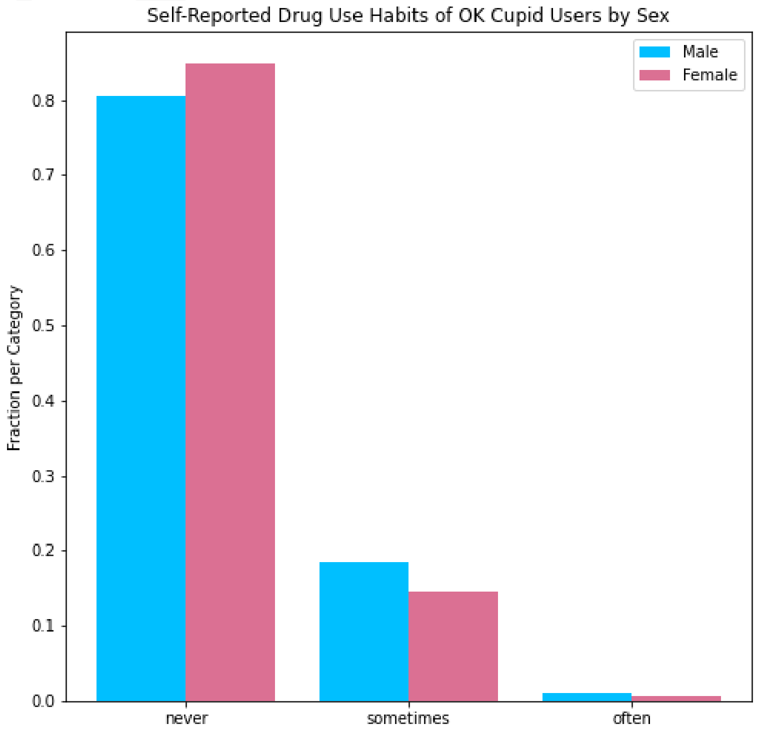


Clear separation

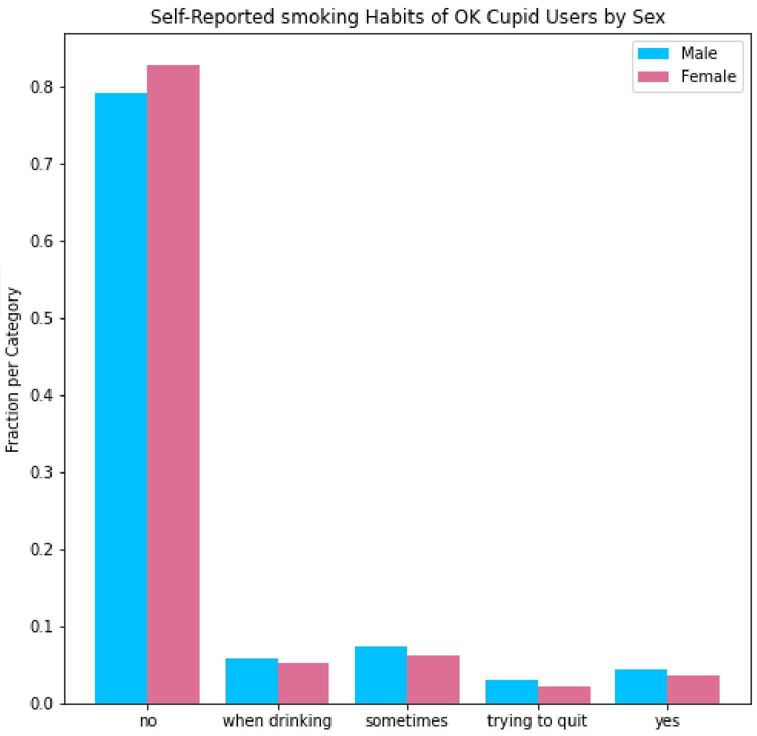
DRINK, DRUG, AND SMOKE HABITS, SPLIT BY SEX



Small differences



More male drug use



More male smokers

FEATURE ASSOCIATIONS WITH SEX

Feature	Associated with sex?	Hypothesis Test	Significance threshold	p-value
age	yes	K-S	0.01	8.5×10^{-16}
height	yes	2-Sample t-test	0.01	$< 3.4 \times 10^{-33}$
drinks	yes	Chi-squared	0.01	5.9×10^{-14}
drugs	yes	Chi-squared	0.01	3.4×10^{-33}
smokes	yes	Chi-squared	0.01	3.5×10^{-25}

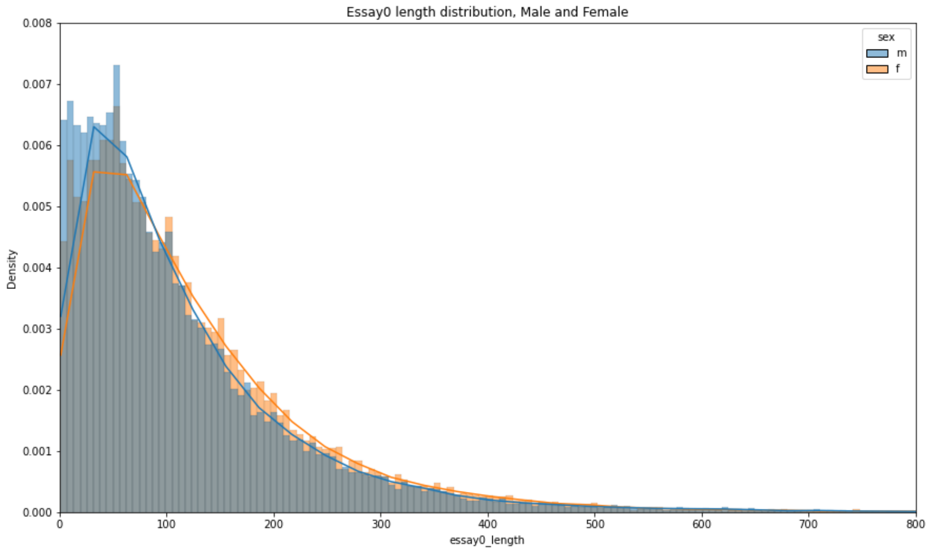
TEXT PRE-PROCESSING

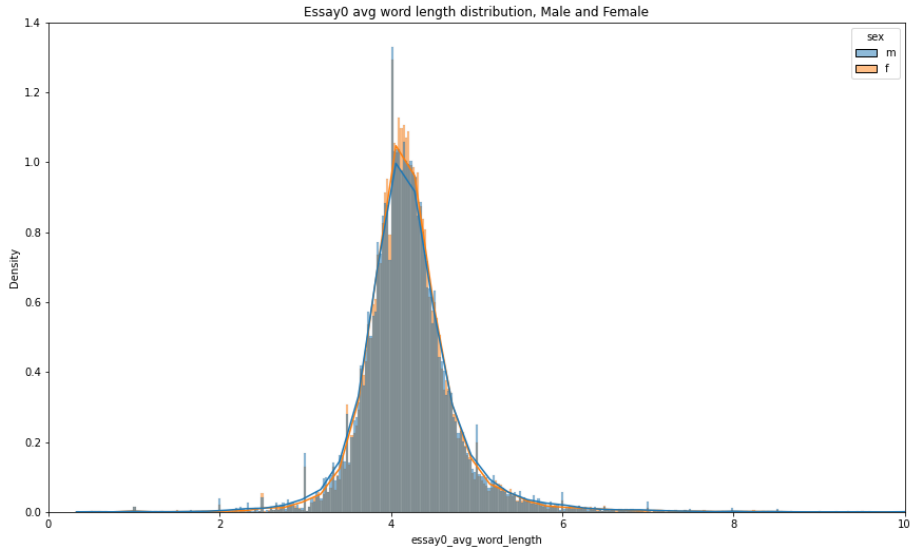
FOR ALL ESSAYS:

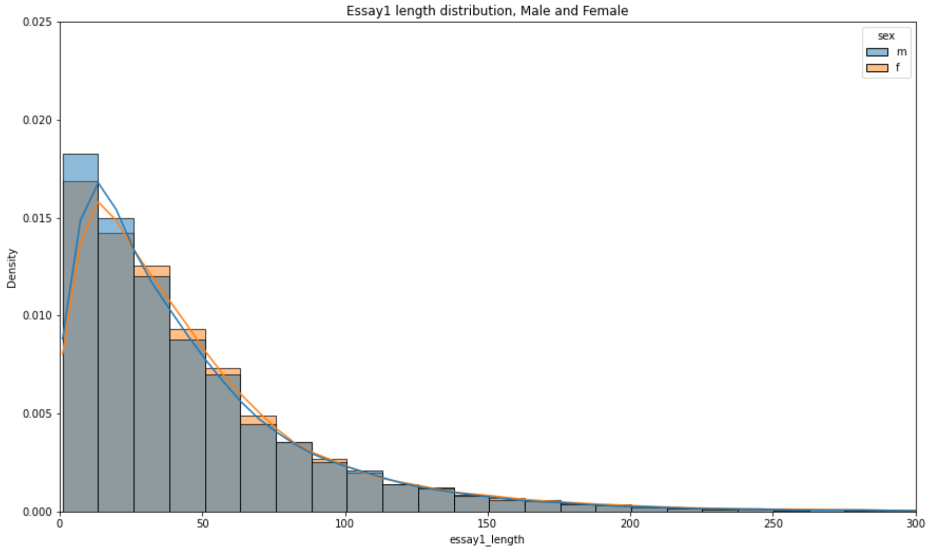
- ❖ Remove HTML links
- ❖ Remove HTML characters (, \n, &lt, &gt)
- ❖ Remove Punctuation
- ❖ All Lowercase

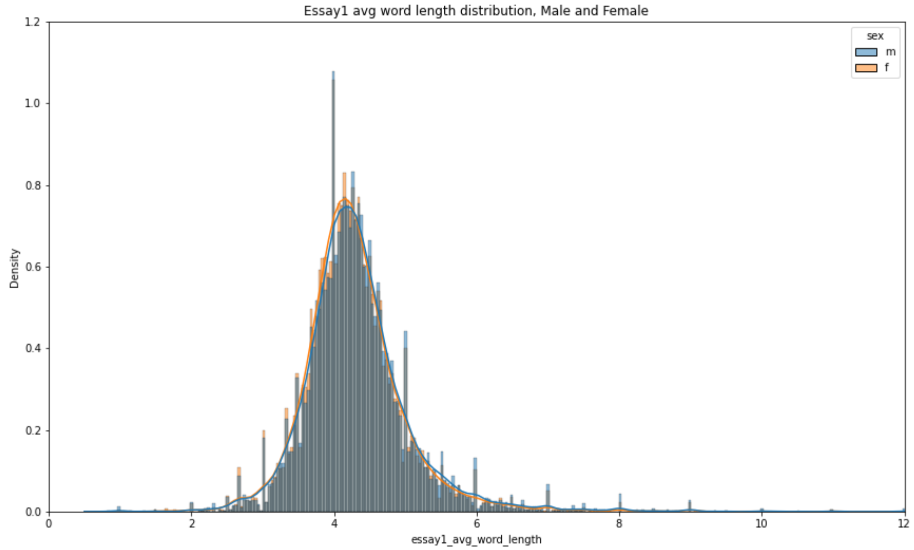
FOR ESSAYS 0, 1, 4, 5, 6, 7:

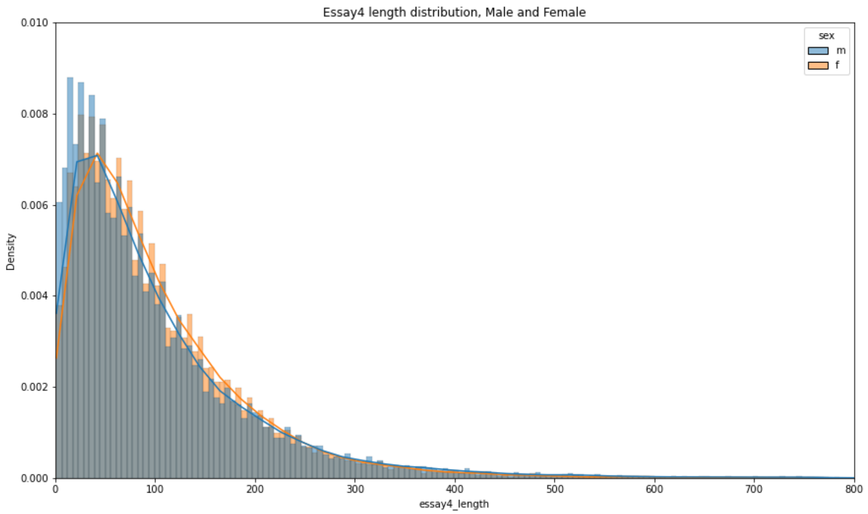
- ❖ Split essays into individual words
  - ❖ Count number of words
  - ❖ Calculate average length of words in essay

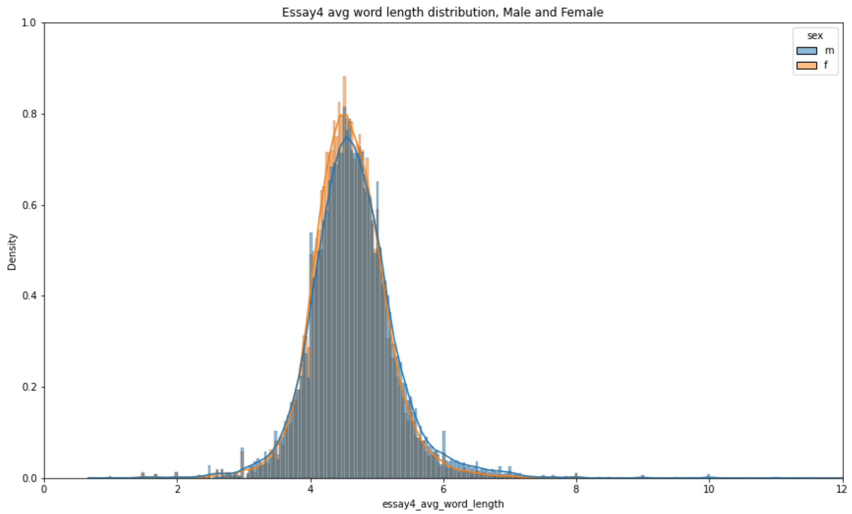

Essay0 length distribution, Male and Female
Density
essay0_length
sex
m
f

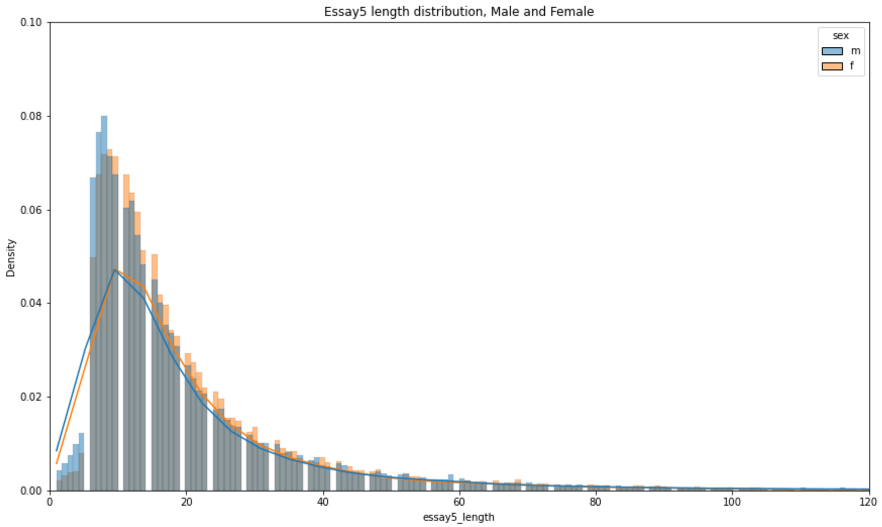

Essay0 avg word length distribution, Male and Female
Density
essay0_avg_word_length
sex
m
f

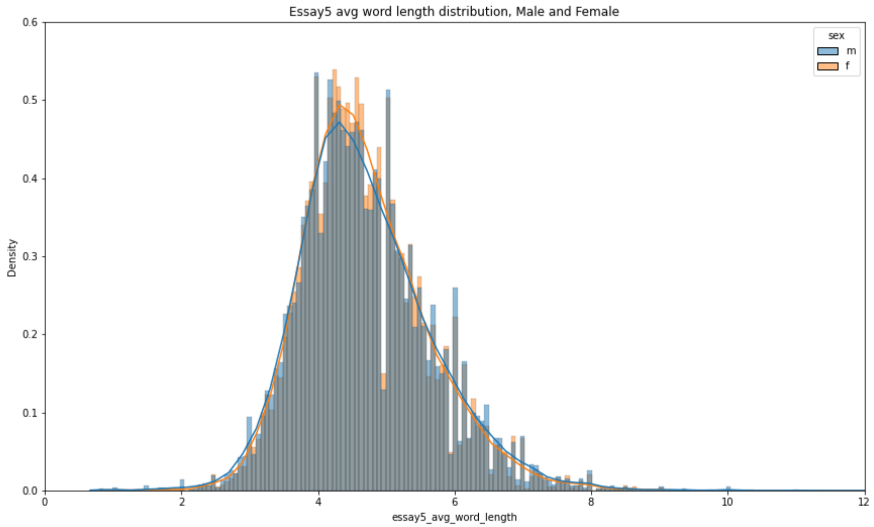

Essay1 length distribution, Male and Female
Density
essay1_length
sex
m
f

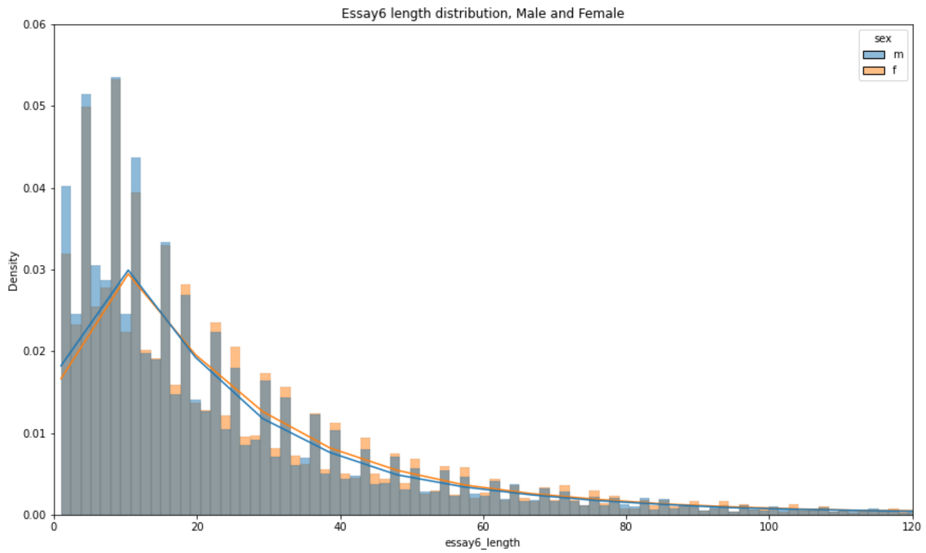

Essay1 avg word length distribution, Male and Female
Density
essay1_avg_word_length
sex
m
f

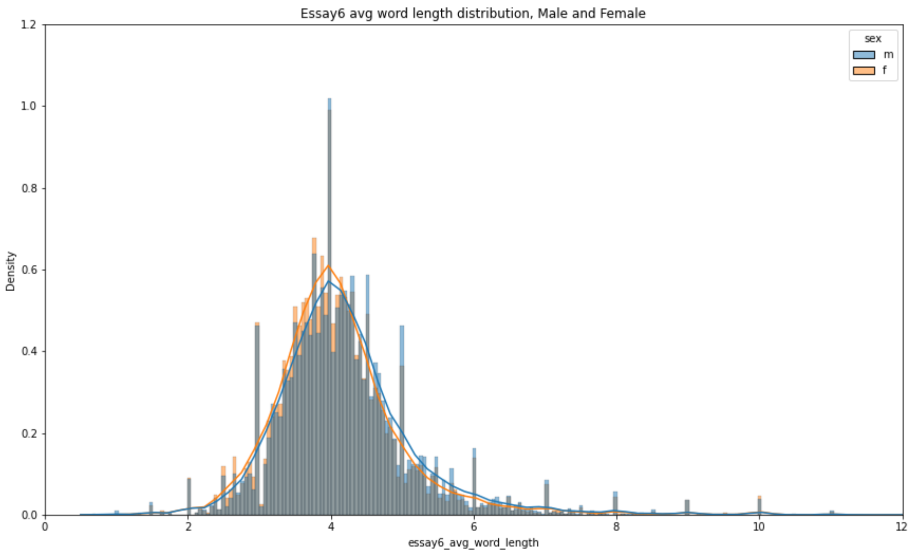

Essay4 length distribution, Male and Female
sex
m
f
Density
essay4_length

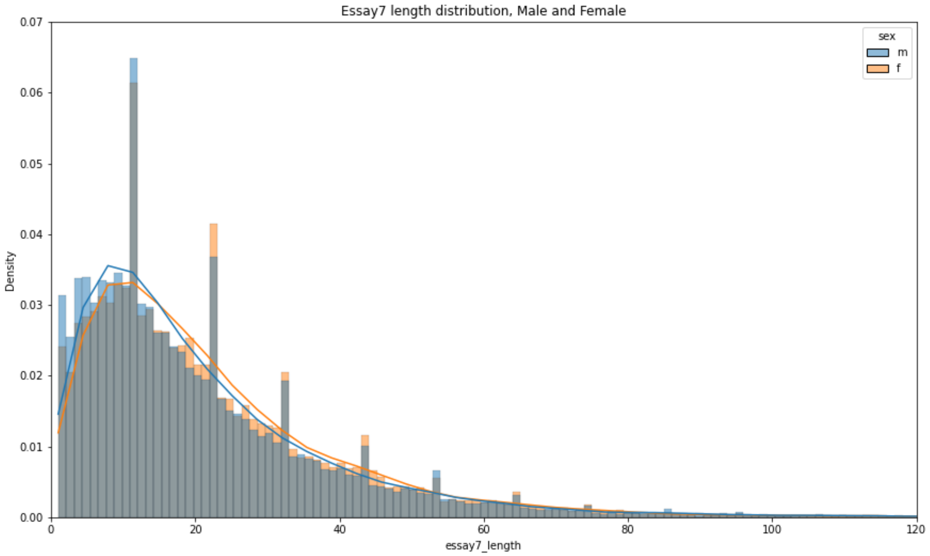

Essay4 avg word length distribution, Male and Female
sex
m
f
Density
essay4_avg_word_length

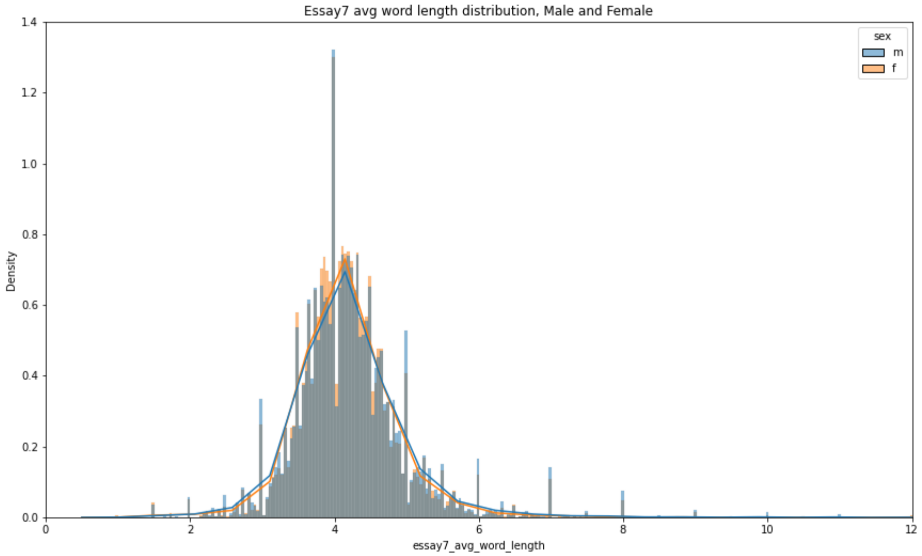

Essay5 length distribution, Male and Female
sex
m
f
Density
essay5_length


Essay5 avg word length distribution, Male and Female
sex
m
f
Density
essay5_avg_word_length


Essay6 length distribution, Male and Female
sex
m
f
Density
essay6_length


Essay6 avg word length distribution, Male and Female
sex
m
f
Density
essay6_avg_word_length


Essay7 length distribution, Male and Female
sex
m
f
Density
essay7_length


Essay7 avg word length distribution, Male and Female
sex
m
f
Density
essay7_avg_word_length

ESSAY LENGTH ASSOCIATION WITH SEX

Essay	Associated with sex?	Hypothesis Test	Significance threshold	p-value
0	yes	K-S	0.01	9.9×10^{-35}
1	yes	K-S	0.01	2.8×10^{-08}
4	yes	K-S	0.01	8.8×10^{-24}
5	yes	K-S	0.01	1.9×10^{-19}
6	yes	K-S	0.01	9.1×10^{-11}
7	yes	K-S	0.01	2.3×10^{-17}

AVERAGE LENGTH OF WORDS IN ESSAYS ASSOCIATION WITH SEX

Essay	Associated with sex?	Hypothesis Test	Significance threshold	p-value
0	NO	2-Sample t-test	0.01	0.995
1	yes	2-Sample t-test	0.01	3.2×10^{-11}
4	yes	2-Sample t-test	0.01	5.9×10^{-24}
5	NO	2-Sample t-test	0.01	0.42
6	yes	2-Sample t-test	0.01	1.0×10^{-22}
7	yes	2-Sample t-test	0.01	0.00049

SUPERVISED MACHINE LEARNING MODELS

ORDINAL CATEGORICAL VARIABLE MAPPING

drinks:

Description	Value
not at all	0
rarely	1
socially	2
often	3
very often	4
desperately	5

drugs:

Description	Value
never	0
sometimes	1
often	2

smokes:

Description	Value
no	0
when drinking	1
sometimes	2
trying to quit	3
yes	4

MULTINOMIAL NAÏVE BAYES CLASSIFICATION MODEL

MODEL 1: Essay 0 contents only

Metric	Value	Extremum?
Accuracy	0.715	
Precision	0.663	
Male misclassifications	0.225	YES
Female misclassifications	0.369	

MODEL 2: Contents of Essays 0, 1, 2, 3, 4, 5, 6, and 7

Metric	Value	Extremum?
Accuracy	0.764	YES
Precision	0.680	YES
Male misclassifications	0.268	
Female misclassifications	0.190	YES

Model 2 is the better model.

K-NEAREST NEIGHBOURS CLASSIFICATION MODEL

MODEL 1: age, height, drinks, drugs, smokes, essay 0, 1, 4, 5, 6, 7 lengths, essay 1, 4, 6, 7, avg word lengths

Metric	Value
Accuracy	0.783
Precision	0.778
Male misclassifications	0.124
Female misclassifications	0.354

MODEL 2: age, height, drinks, drugs, smokes

Metric	Value
Accuracy	0.80
Precision	0.775
Male misclassifications	0.138
Female misclassifications	0.292

LOGISTIC REGRESSION CLASSIFICATION MODEL

age, height, drinks, drugs, smokes, essay 0, 1, 4, 5, 6, 7 lengths, essay 1, 4, 6, 7, avg word lengths

Metric	Value
Accuracy	0.824
Precision	0.809
Male misclassifications	0.117
Female misclassifications	0.262

SUPPORT VECTOR MACHINE CLASSIFICATION MODEL

age, height, drinks, drugs, smokes, essay 0, 1, 4, 5, 6, 7 lengths, essay 1, 4, 6, 7, avg word lengths

Metric	Value
Accuracy	0.819
Precision	0.752
Male misclassifications	0.182
Female misclassifications	0.180

OVERALL PREDICTIVE MODEL PERFORMANCE

Model	Accuracy	Precision	Male misclassifications	Female misclassifications
Multinomial Naïve Bayes	0.764	0.680	0.268	0.190
k-Nearest Neighbours[2]	0.80	0.775	0.138	0.292
Logistic Regression[1]	0.824	0.809	0.117	0.262
Support Vector Machine	0.819	0.752	0.182	0.180

1: shortest computational time

2: longest computational time

CONCLUSIONS

1. Can sex be predicted using a multinomial naïve Bayes classifier trained on OK Cupid user essay texts?
   Answer: YES, but it doesn't perform as well as a supervised machine learning model trained on quantitative features.

2. Can sex be predicted using a machine learning algorithm trained on age, height, drinking habits, drug use habits, smoking habits, essay lengths, and average lengths of words in essays?
   Answer: YES. Use a logistic regression classifier if higher precision is desired; use a support vector machine classifier if higher recall is desired.

ACKNOWLEDGEMENTS

This project is a Codecademy "Portfolio Project" which fulfills a requirement of the Data Science learning path.

I would like to thank Codecademy for providing the data used, as well as for hints provided in an earlier version of the project (when it was called a "Capstone Project").

Codecademy gives no indication of where they got the data from, but it stands to reason that they had to interact with OK Cupid at some point to get it. So, I would like to thank OK Cupid for the role they played in allowing Codecademy to compile the data.