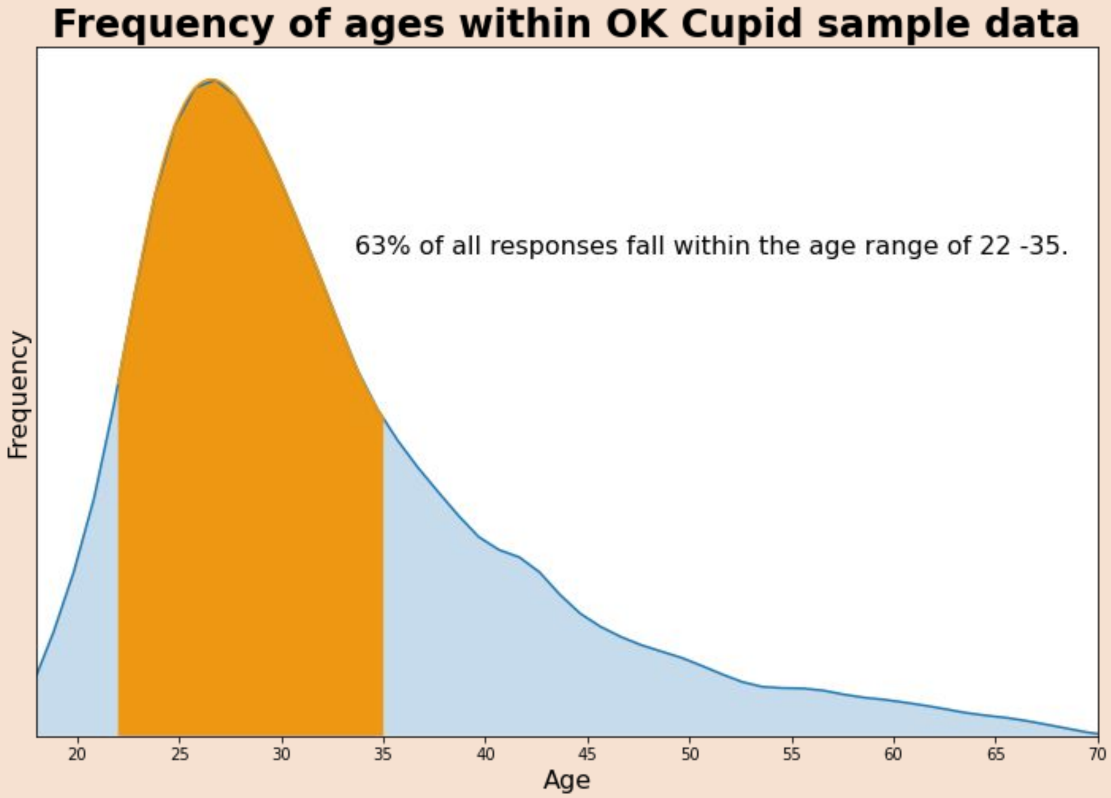
Date-A-Scientist Capstone Project

Machine Learning Fundamentals
Kevin Sullivan
18 July 2020

Table of Contents

- Exploration of the Dataset
- Classification Question to Answer
- Classification Approaches
- Regression Questions to Answer
- Regression Approaches
- Conclusions/Next steps

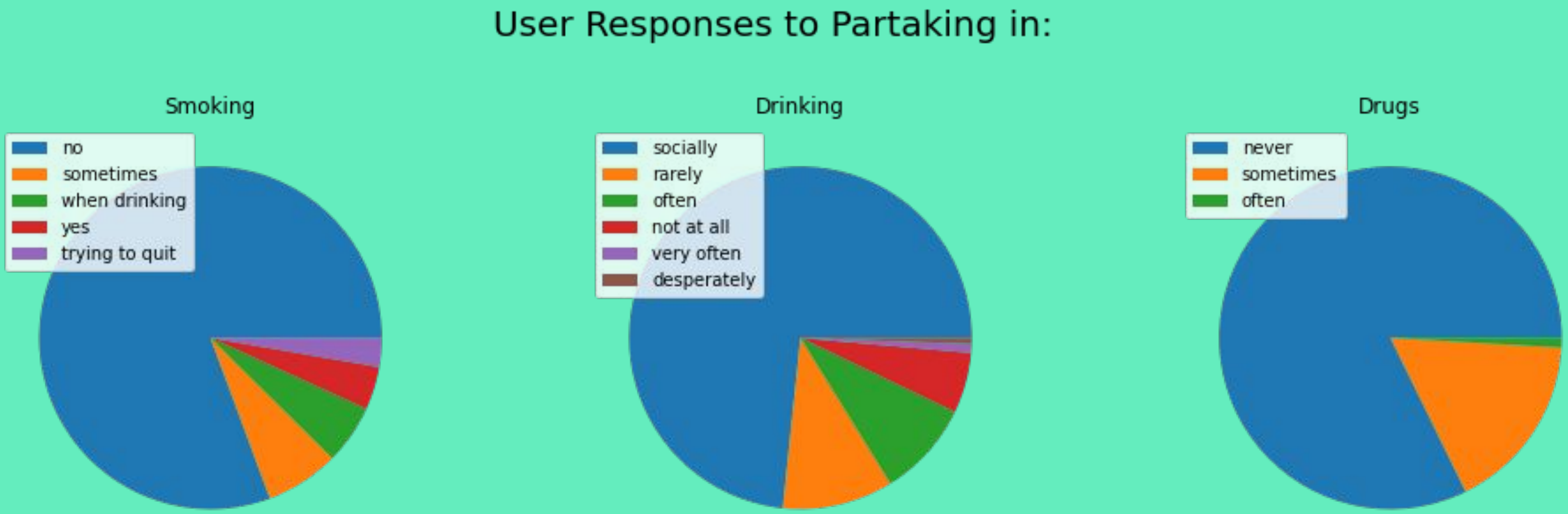
Age Distribution Within Dataset



Taking a look at the data we can look first into the age range

- Using a probability distribution of the data we can determine that 62.9% of users fall between the age ranges of 22-35 indicating that this dataset skews very strongly toward young adults
- The max listed age is 110, but beyond age 70, there are very few responses

Attributes of the Dataset



We'll use the data for responses about smoking, drinking and drugs as part of our classification so what are the actually responses.?

These data are placed on a numeric scale starting from 0 for not partaking in order to be able to enter a classification algorithm

Attributes of the Dataset

Long Form Essays

The responses also include long form essay responses to various personal questions.

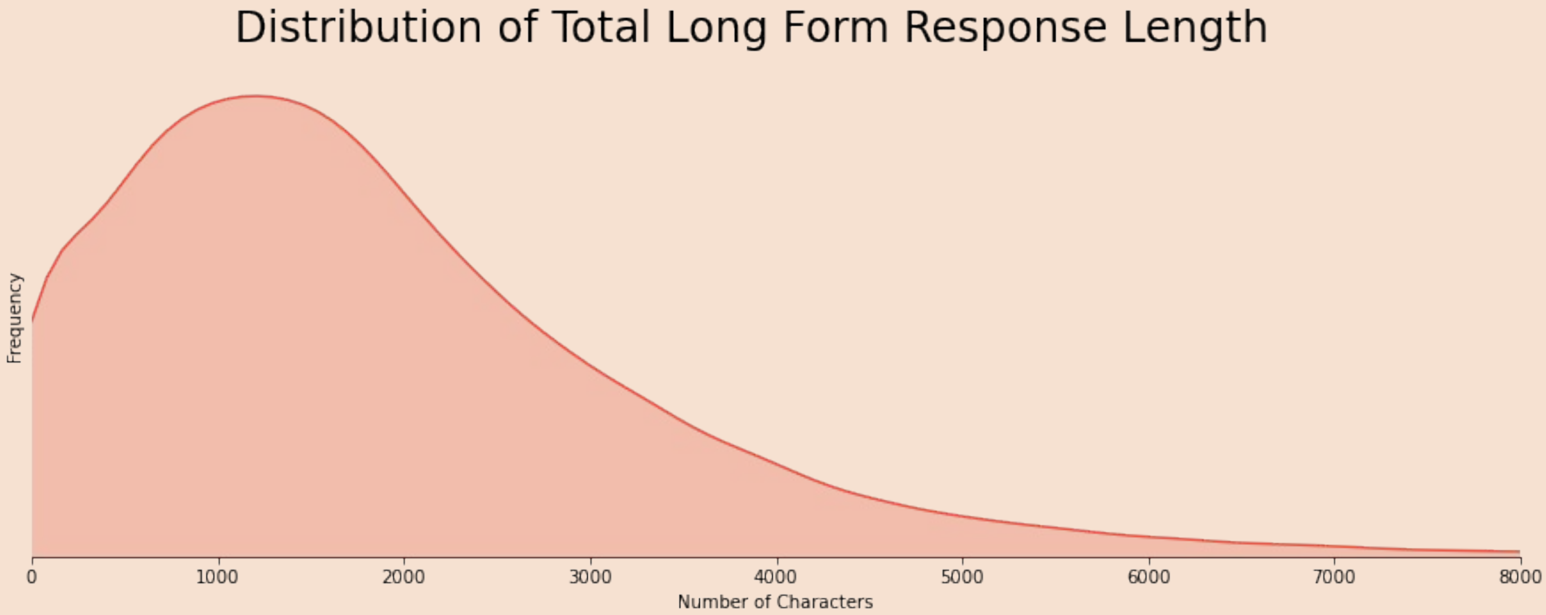
In order to analyze these responses we should clean the data after merging all responses together in order to:

- Place all words as lowercase
- Remove HTML tags included in the responses
- Remove punctuation

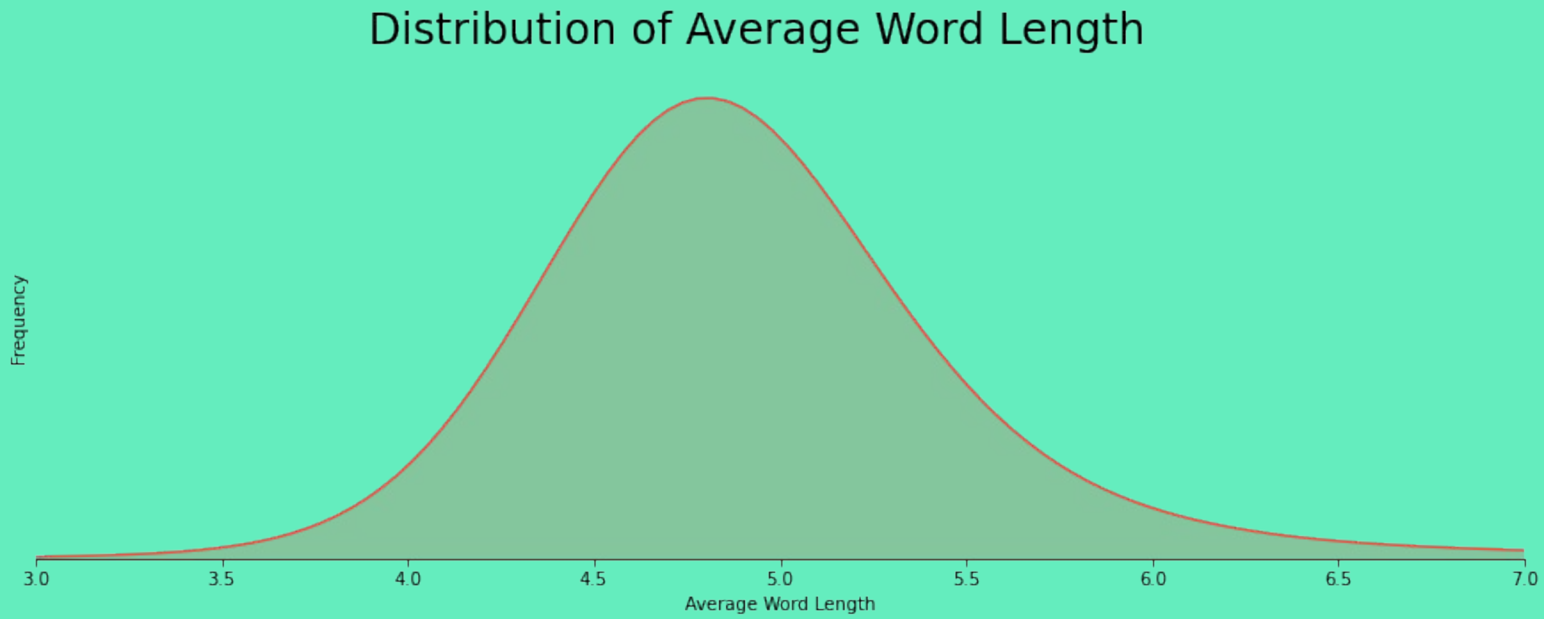
From this we will make columns for:

- Total length of essay
- Average word length
- The number of times “I” or “me” appears divided by total number of words

Attributes of the Dataset

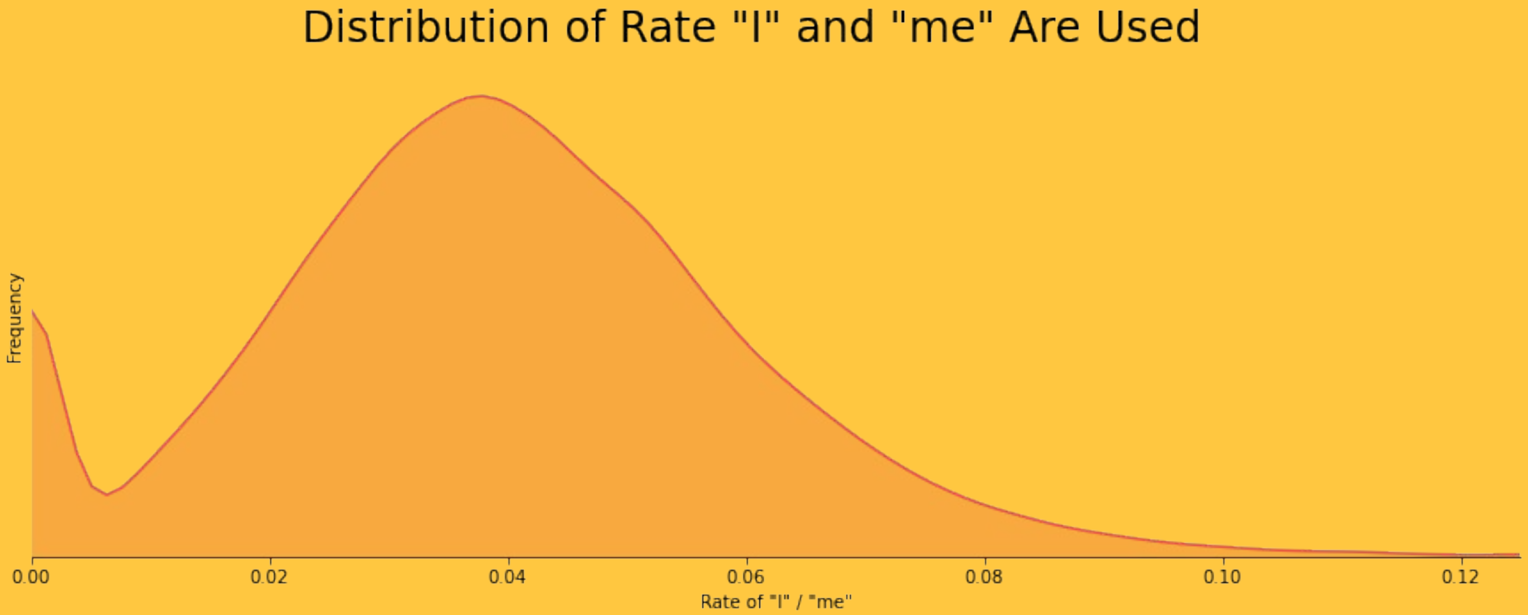


Attributes of the Dataset



Note: There is a large peak at 0 from empty responses

Attributes of the Dataset

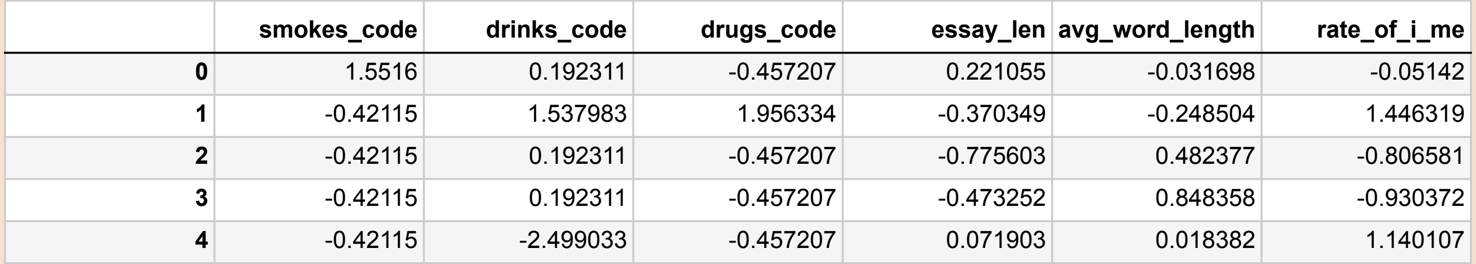


Normalizing the Dataset

The data were normalized using a StandardScalar object from Scikit Learn.

This was used over MinMax due to the presence of many extreme outliers in the essay responses

The first 5 normalized responses are shown below



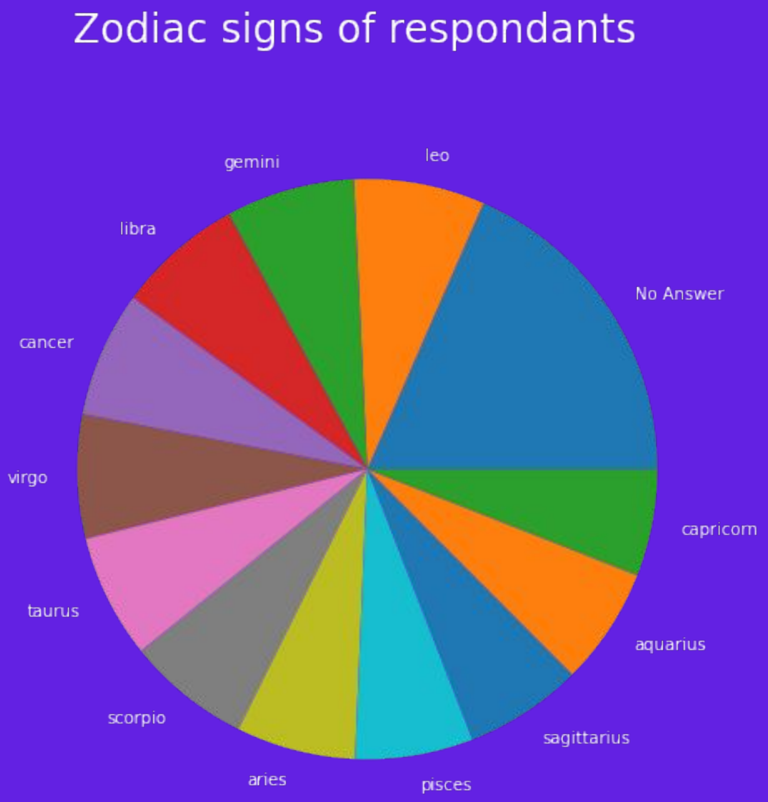
	smokes_code	drinks_code	drugs_code	essay_len	avg_word_length	rate_of_i_me
0	1.5516	0.192311	-0.457207	0.221055	-0.031698	-0.05142
1	-0.42115	1.537983	1.956334	-0.370349	-0.248504	1.446319
2	-0.42115	0.192311	-0.457207	-0.775603	0.482377	-0.806581
3	-0.42115	0.192311	-0.457207	-0.473252	0.848358	-0.930372
4	-0.42115	-2.499033	-0.457207	0.071903	0.018382	1.140107

Analyzing the Question

We will look to see how predictive our existing traits are for Zodiac symbols

Let’s look at the responses to the zodiac sign.

With 13 responses, random chance should be 7.7% accurate

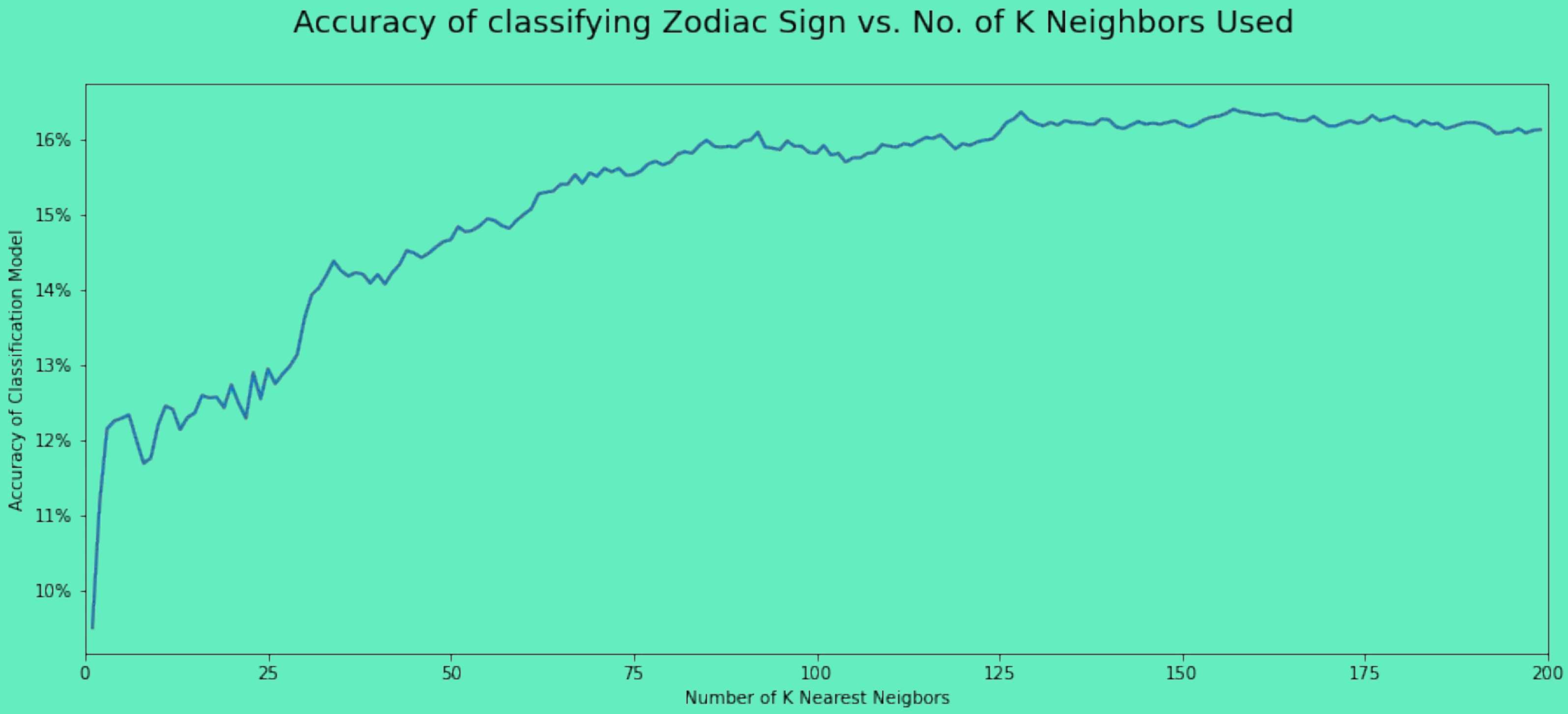


Classification Results

Obviously there’s nothing to astrology so it’s got to be close to random chance.

So….what did we find?

Classification Results



Classification Results

Did we just prove there’s something to Astrology?

Well….not yet. We should come up with a few alternative explanations.

- People may adjust their responses to fit their sign
- There is probably a huge correlation with no response on the essays and no response for zodiac sign

That second explanation seems much more probable and our classifier can just say no answer for one means no answer for another. Let’s test it out by removing the rows with “No Answer” from our zodiac sign list.

We don’t want to remove empty essays because that still *could* be a hint for zodiac signs

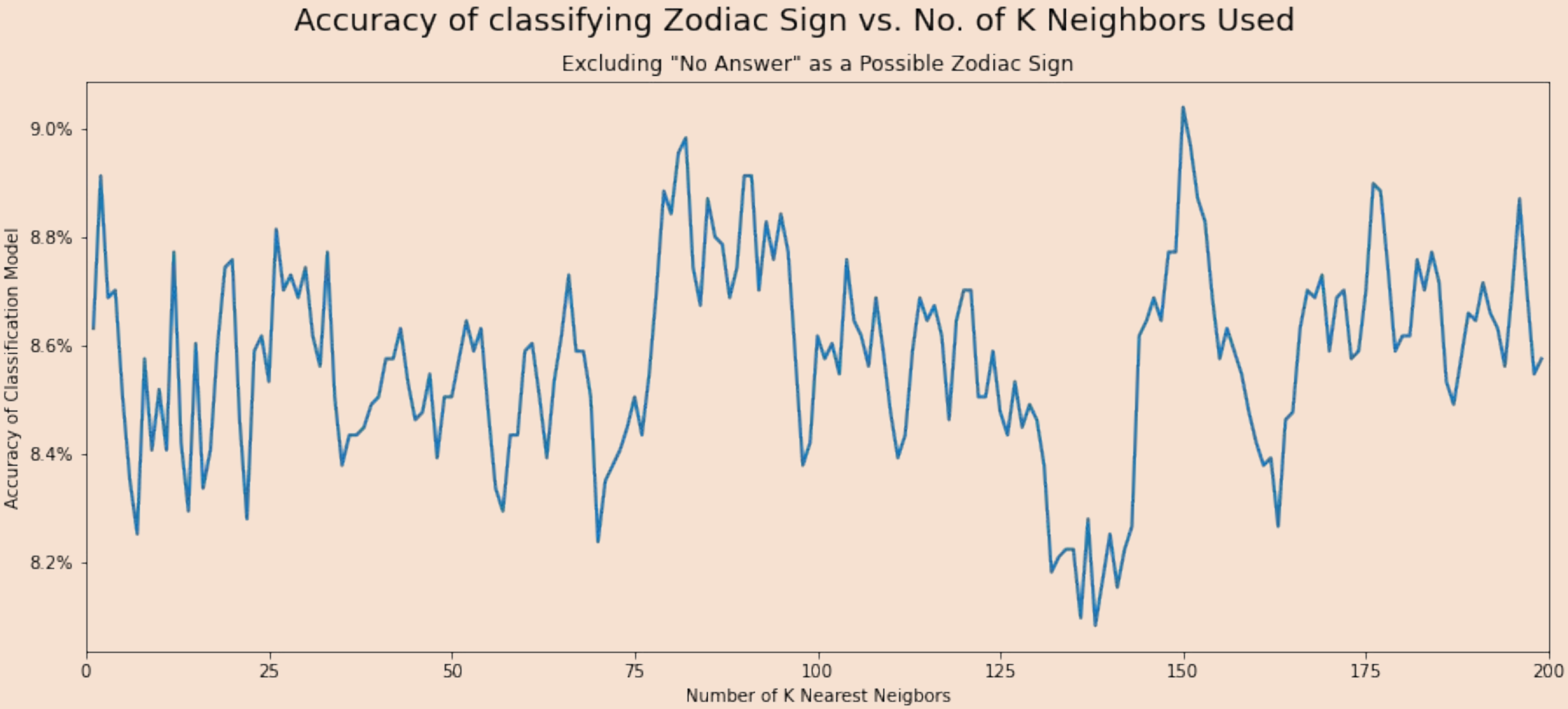
Classification Results

Sooooo…..

Let’s re-run the data!

Remember that now with 12 possible responses, the accuracy to beat is 8.3%

Classification Results



Classification Results

It looks like our classification did beat random chance…

Just not by much.

The number of neighbors doesn't appear to affect the accuracy much and our mean was 8.6%

If we assume each score value is independent, we can determine that it is indeed statistically significantly higher than random chance*

This is a very slight if real result so our other hypothesis may be valid that people may choose responses they think are fitting.

* resulting p-value was 5.5 x 10^{-46}

Regression Questions

We will also see how well we can evaluate the following questions using regression analysis:

- Can we predict income with length of essays and average word length?
- Can we predict age with the prevalence of “I” or “me” in an essay

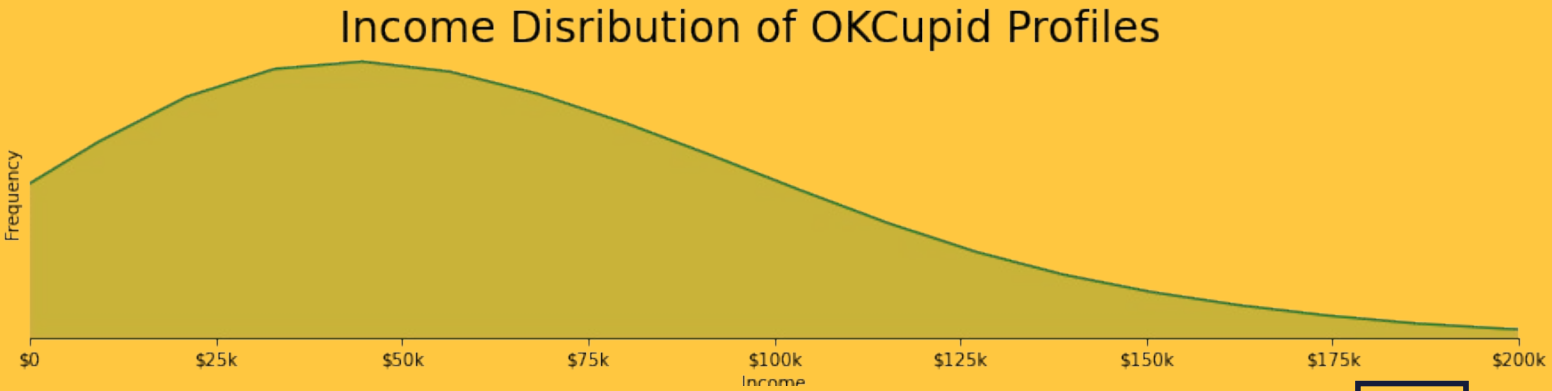
Let’s get started!

Regression Preparation

Upon looking at the income data we can see that a value of -1 corresponds to no answer so we remove all of those rows.

Unfortunately for our analysis only 19.2% of respondents have a valid response

The distribution may be seen below:



*Note: There is also a peak of people reporting income of $1,000,000

Regression Preparation

We can imagine there might be some selection bias in who is choosing to place their income as well as the possibility that people may be lying to inflate their income.

The asterisk in the graph noted that there is a peak around $1,000,000 that seems very unnatural in a true distribution

Fortunately, we don’t care about the “Real” distribution since we only want to compare within the dataset.

Time to model!

Regression Analysis

We will use standard linear regression and K nearest neighbors regression to see how well the data fit our questions.

On the standard linear regression model we get the following results for our two questions

	Impact of essay length on income (char/$)	Impact of average word length on income (char/$)	R^2 value of fit
Standard Linear Regression	-2.19	10100	**-0.00306**

K Neighbors Regression returned an R^2 value of -0.0144

Regression Analysis

Time to move on to asking about Age

With a standard linear regression model we get a coefficient of -16.8 meaning that for a rate difference of 10% more (e.g. from overall 5% to 15%) we get would expect age to decrease by 1.6 years.

Considering our data range is almost entirely under 10% this doesn't make much sense. Indeed the R^2 value is .000875.

Using our K nearest Neighbors approach the R^2 value is -0.0117

Regression Conclusion

Our data appears to not be very useful for linear regression as all of our methods resulted in a very poor fit.

GOOD LUCK AND HAVE FUN!

Thank You For Your Time

codecademy