Capstone Interim Report

Problem Statement:

Motivation:

Understanding human behavior is important for both psychologists and businesses. Psychologists knowing how people may react to certain comments may choose to treat patients differently based on their personalities. Additionally knowing a person's personality can help match which psychologist would best be able to help a patient. Aside from treatment, knowing information about people can be used as a preventative measure. Finally, knowing what influences human behavior can help understand how the brain works.

Furthermore, businesswise, being able to classify people in different categories is important because certain personalities may be more likely to buy certain goods or partake in certain events. It also can be useful in an interview process, to understand how potential candidates would fit in a team without necessarily asking questions directly where an interviewee may be pressured to not tell the truth.

Goals:

This project will look at how people's backgrounds including the number of siblings that they have, the area that they live in (urban vs rural), and education affect their

personalities and things they like. It will further look at how people who like certain types of music/movies may fit into certain personalities.

This project will come up with a model to predict who is most likely to be afraid of public speaking. This is useful to know both for interviewers looking to identify strong public speakers and for people who may need to practice more to overcome their fears. It also will look at workaholism to help identify people who are very task oriented, a highly desired skill in the workplace.

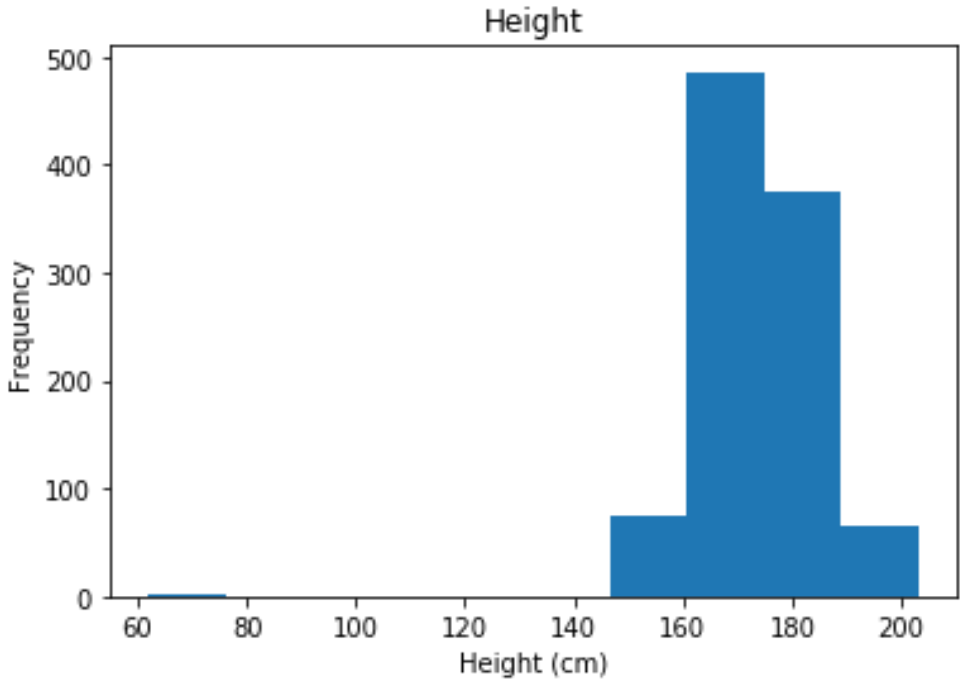
The dataset can be found on Kaggle here. This will be delivered with a slide deck and code.

Data Wrangling - see code in filtering.ipynb and DataWranglingOutliers.ipynb

After importing the data, I selected which features to use in my model. I kept the variables of interest, namely public speaking and workaholism. I then selectively chose around 30 other features that are general and not related to each other. While doing this, I filled in null values with the avg of the features. Next, I noticed that while most data points were numerical values between 1 and 5 describing the person's affinity to the variable in question, others were strings. To correct for this, I mapped the string

values to numeric. For example, I mapped the string ‘social drinker’ value under ‘Alcoholism’ to 5 since that indicates a strong affinity.

After doing this, I checked for outliers. The only columns not rated 1 -5 were height and weight. Upon plotting I saw that there was 1 height between 60-80 cm (see below).

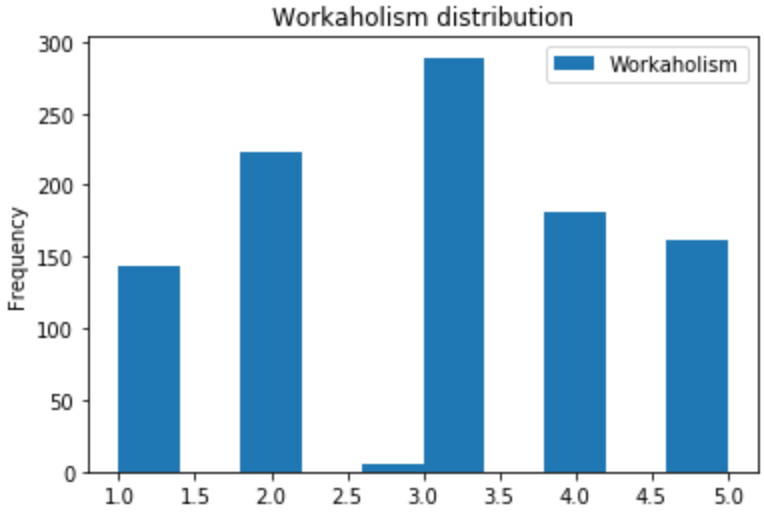


I corrected this point by noticing that it’s corresponding weight was in the normal range. I therefore filled the point with the avg height of people with that given weight.

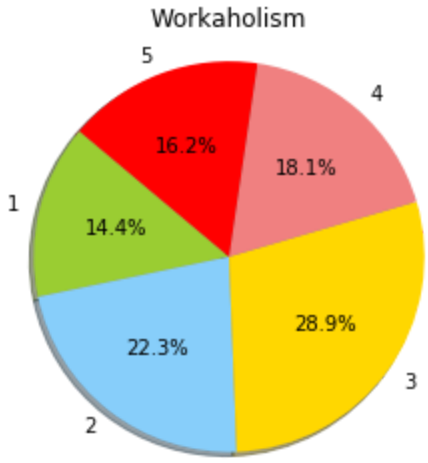
Finally, I confirmed that all the other columns had values between 1 and 5.

Exploratory Data Analysis - see code in Statistics.ipynb and DataExploring.ipynb

I next visualized my data to find trends in the relationship between variables. First, I plotted the distribution of our target parameter, Workaholism.

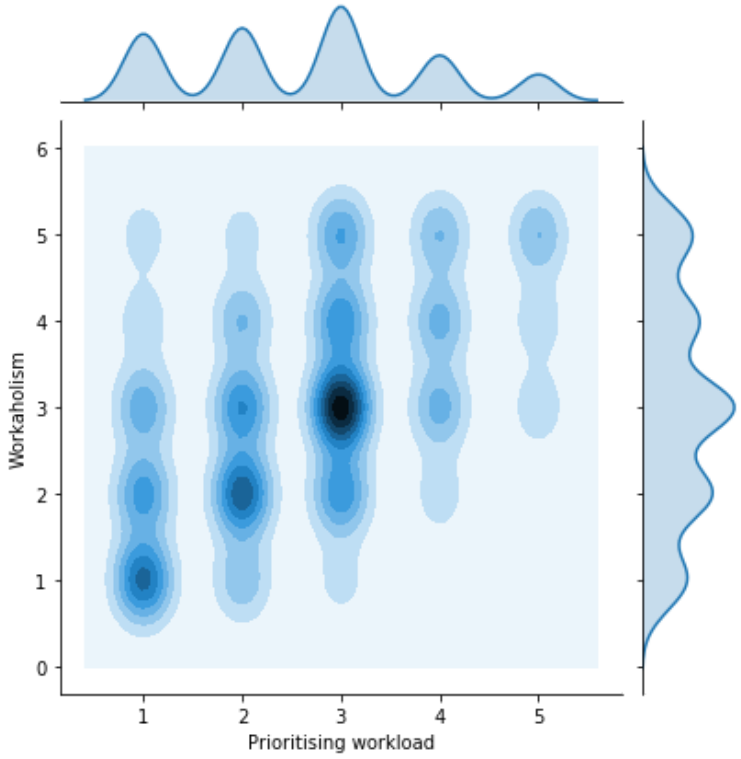


The distribution looks normal with most of the values being 3.



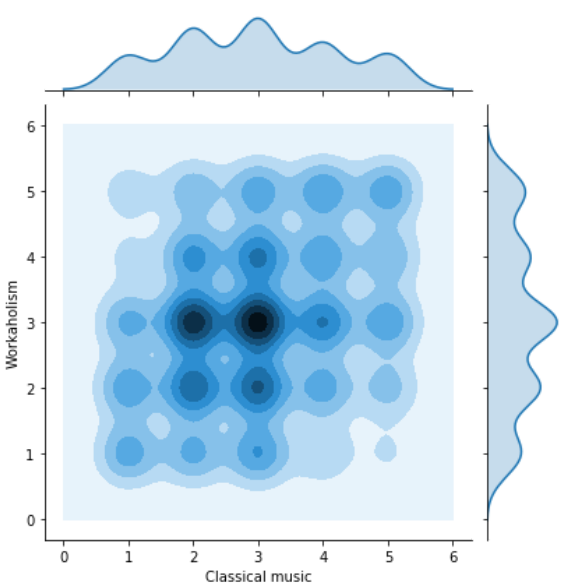
Next I visualized correlations by creating join plots to get an idea of which features to include. Most of the graphs did not show any strong correlations. A couple

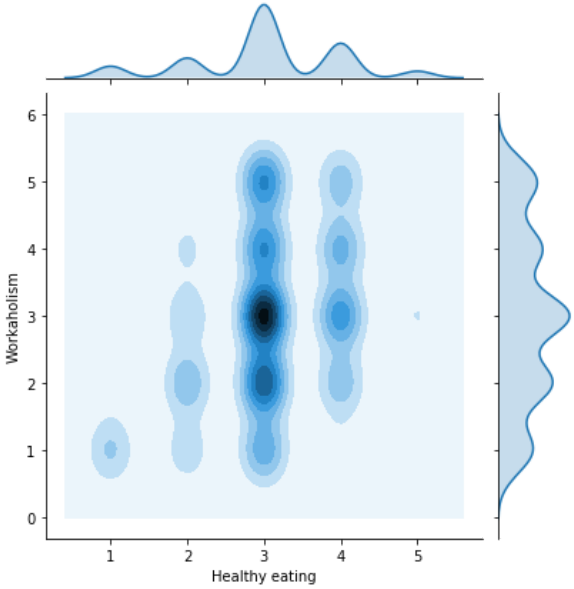
did though. One correlation that made a lot of sense intuitively was with ‘prioritizing workload’.



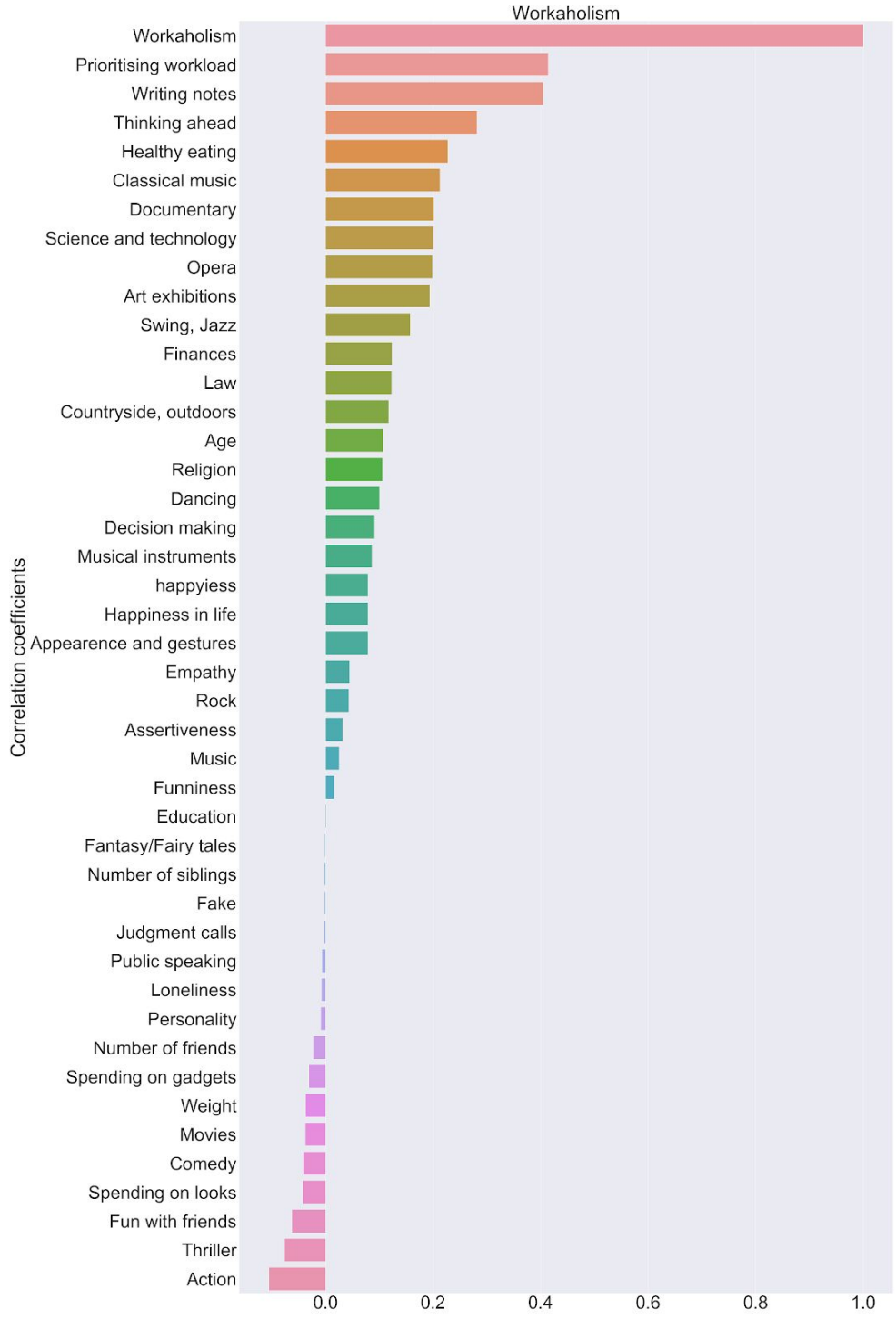
We see that people who don’t prioritise their workload tend not to be as workaholic.

We can find more interesting correlations though.





We see that people who eat healthier tend to be more workaholic and tend to like classical music more. I then used statistical methods to be able to more easily identify the relationships. To do this, I calculated the Pearson coefficients



This graph gives the correlation between workaholism and our various variables.

We see as expected that there is a strong correlation with ‘Prioritising workload’, and

‘thinking ahead’. We also see some intuitive negative correlations with ‘fun with friends’. Finally the lack of correlation between some of the features like ‘education’ is somewhat surprising.

After calculating these coefficients I evaluated whether these values were significant or just chance. First I set up a hypothesis test. The null hypothesis was that there was no significant relationship between ‘Workaholism’ and ‘ Prioritising workload The alternative was that there was. I set a significance level of .05 and split up the dependant feature by the young people who said they strongly had it (rated themselves higher than 3) and those who said they didn’t have it (had less than 3). I found that the p value was very close to 0, allowing me to conclude that there was a significant relationship. Instead of repeating this for each feature, I set up the same test in a loop for all of the features. I found that 23 of the 42 columns I was looking at had significant correlations.

After doing this, I also found a 95 percent confidence interval for mean difference of workaholism and ‘fun with friends’ as another way to double check for significance. The fact that 0 was not included in the CI, (-.144,-.03) shows that it is unlikely that there is no correlation. That being said though, -.03 would be a weak correlation. It made sense therefore that we obtained a p value close to 0.