The State of Masculinity

Shirley Yu, sey025@ucsd.edu
Reid Doctor, rhdoctor@ucsd.edu
Qi Goh, qgoh@ucsd.edu
Angela Zhang, anz020@ucsd.edu

I. Abstract

In this paper, multiple analyses are performed on a dataset which depicts men's views on masculinity. These analyses use PCA, logistic regression, polynomial regression, and k-means clustering. K-means clustering with PCA provided the most interpretable results. We found that there were 3 main ideologies about masculinity, though the population was not strongly divided among them.

II. Introduction

Over the past few decades, there has been much evolution of people's conceptions of gender and gender roles. We decided to study a dataset containing men's responses to questions relating to masculinity, sexuality and relationships to see if we could quantify the current role of masculinity in society as thought by men. In our modern political climate, with constant headlines of sexual assault stories, sexism in the workforce being brought in the spotlight, and general friction about issues relating to men and women, we believe this information is important to study in order to understand the reasons behind the bigger problems.

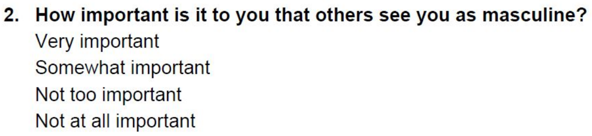
III. Dataset

The dataset comes from a SurveyMonkey survey on behalf of FiveThirtyEight titled "What Do Men Think It Means To Be A Man?" with 35 questions total. It begins with questions on ideas about masculinity such as "How masculine do you feel?" and "Where have you gotten your ideas about what it means to be a good man?". It then follows with questions about general aspects of lifestyle, such as asking for advice and worrying about physical health. Then, if a respondent said they were employed, it followed with questions about gender disparities and sexual harassment in the workplace. The survey ends with questions on intimacy and romantic relationships, asking questions relating to the typical role men play in relationships such as "How often do you try to pay when on a date?" and "Do you feel like you need to make the first move?". It also collects some demographic information, including relationship status, number of children, sexual orientation, age, race, and education level which allowed us sort the data into initial categories. There were a total of 1615 participants in the dataset, all male but otherwise randomly distributed.

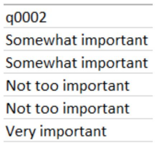
IV. Data Preprocessing

The data set we chose to work with was structured in a purely categorical way, so we had to do quite a bit of preprocessing before it was ready to work with and perform analysis on. There were 3 types of questions in the survey: multiple choice, multiple answer, and open ended. Open ended responses were left out of the raw responses dataset and multiple choice responses were filled in in their corresponding columns. Multiple answer responses were already semi one-hot encoded in the raw data set. Each choice for the question was given its own column to fill out. If the participant had some choice selected, its corresponding column would be filled out with the response in word form.

Multiple Choice:

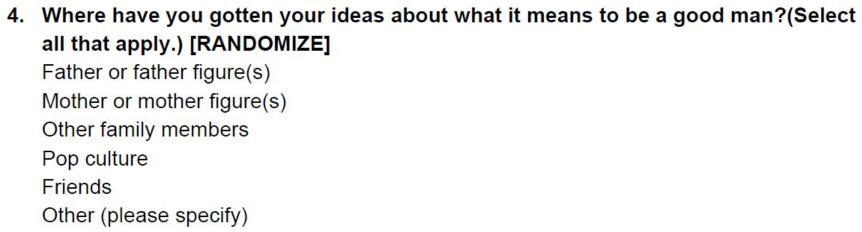


2. **How important is it to you that others see you as masculine?**
   Very important
   Somewhat important
   Not too important
   Not at all important

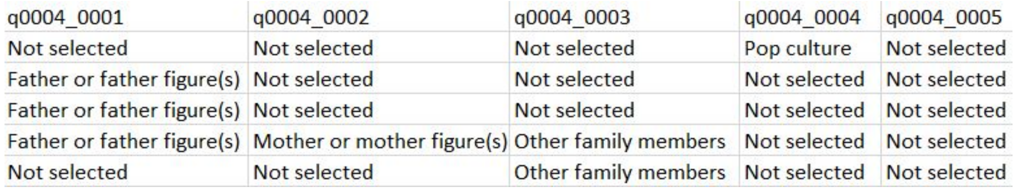


q0002
Somewhat important
Somewhat important
Not too important
Not too important
Very important

Multiple Answer:



4. **Where have you gotten your ideas about what it means to be a good man?(Select all that apply.) [RANDOMIZE]**
   Father or father figure(s)
   Mother or mother figure(s)
   Other family members
   Pop culture
   Friends
   Other (please specify)



q0004_0001	q0004_0002	q0004_0003	q0004_0004	q0004_0005
Not selected	Not selected	Not selected	Pop culture	Not selected
Father or father figure(s)	Not selected	Not selected	Not selected	Not selected
Father or father figure(s)	Not selected	Not selected	Not selected	Not selected
Father or father figure(s)	Mother or mother figure(s)	Other family members	Not selected	Not selected
Not selected	Not selected	Other family members	Not selected	Not selected

We think regression and clustering will both be useful approaches in analyzing this data. Regression will allow us to understand the relationships between responses to the different questions and see which answers are the most important in regards to feelings of masculinity. Clustering will be able to sort the surveyees into different categories, based on different predictors and groups of predictors, to see if we can further understand how combinations of features affect masculinity. We also think PCA will be useful to reduce the dimensionality of the data since there are a large number of features to begin with. The modeling and analysis algorithms we wished to employ on our data are unable to interpret categorical variables, therefore we applied one-hot encoding to the data in order to transform the categorical features of our dataset into numerical features so that our algorithms could more easily interpret them.

Categorical variables, such as the ones in our dataset, are less than ideal when it comes to fitting models and finding relationships between predictors and responses. This is largely due to the fact that categorical variables often consist of many more levels of predictors compared to numerical variables. These extraneous levels only serve to obscure the more significant variables that capture greater amounts of variance within the data set and negatively impact the final model that we are trying to fit.

Prior to any transformations, our data set consisted of 29 predictors, one for each non-open ended question in the survey. However, after performing one-hot encoding, the number of predictors increased to 261 since each question had a varying number of possible answers. It does not take an expert to tell that this is a superfluous amount of data and that it is more than likely that most of this data will have little to no effect on the models we fit. Consequently, we must find a way to reduce the dimensionality of our data such that we can fit models without capturing so much noise that we begin overfitting the dataset.

V. Model Selection and Estimation

We decided to first used Principal Component Analysis as a form of dimensionality reduction so the data would be easier to work with, then perform different modeling algorithms and compare them to see which one gave us the most information.

A. Principal Component Analysis

We used sklearn.decomposition.PCA for this part. We typically begin PCA by standardizing all the predictors by dividing each by its standard deviation, but because our predictors are all dummy variables for the one-hot encodings, we performed this with and without and compared the two. We made this decision based on research online, since we never learned about what to do in this situation in class, eventually going with the unstandardized PCA using Occam's Razor. We decided to use the top 50 PCs out of the 261, which we believed would explain enough variance in the data to still be accurate in future models and give us enough information to understand when the PCs began to be unnecessary. Using pca.explained_variance_ratio_, we found that the first principal component explained 13% of the variance, the second explained 5.24%, third explained 4.47% and then rest explained less than 3% each.

PC1	PC2	PC3	PC4	PC5	PC6
0.13063776	0.05241703	0.0446671	0.02854428	0.02477324	0.02168846

B. Multiple Logistic Regression with PCA

We used statsmodels.formula.api.smf for this part. To further test out the accuracy of our reduced-dimensionality data, we decided to fit a series of logistic regression models using the top k=1, 2, 3, …, 50 PCs to predict whether someone felt masculine or not masculine. This dependent variable was based on the data from the question "In general, how masculine or manly do you feel?" where the answers "Very masculine" and "Somewhat masculine" were encoded as 1 and "Not very masculine" and "not at all masculine" were encoded as 0. We used a 10-fold cross-validation to estimate the testing MSE and accuracy for each logistic model. We found that the number of PCs used with the lowest testing error and highest accuracy was 10, having an MSE of 0.0278638 and accuracy of 0.972755.

Number of PCs	Mean Training Error	Mean Testing Error	Mean Training Accuracy	Mean Testing Accuracy
1	0.028483	0.028483	0.971517	0.971517
10	0.027245	**0.027245**	0.972755	**0.972755**
20	0.028638	0.032198	0.971362	0.967802
30	0.027709	0.029721	0.972291	0.970279
40	0.025851	0.034056	0.974149	0.965944
50	0.024303	0.035913	0.975697	0.964087

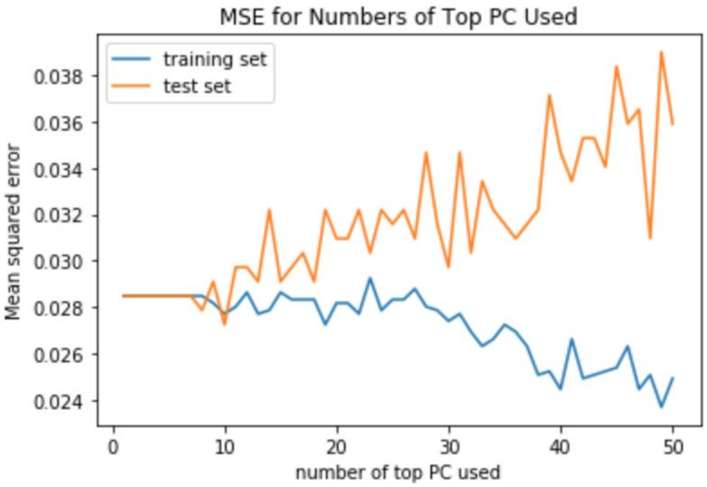


Figure 2: Mean squared error of logistic regression for different numbers of top PCs used

As shown in the graph above, we found that after using more than 10 PCs, although the training error continued to decrease, the testing error slowly began to increase. This may be a sign that using increasingly more PCs may lead to overfitting the training data, which would result in lower training error but high testing error. This does make sense intuitively, since those additional PCs only explain around 1% or less of the variance each, they have little to no effect on the data as a whole.

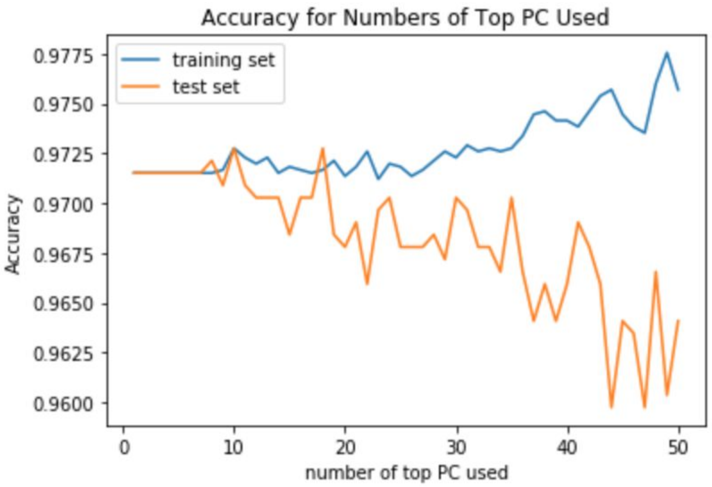


Figure 3: Accuracy of logistic regression for different numbers of top PCs used

The same can be said for the accuracy, once we went past 10 PCs the training accuracy continued to climb steadily while the testing accuracy begin to decrease with a steeper slope. In this particular run of our series of logistic regression models, we found that 10 PCs was the best model both in terms of mean squared error and accuracy. We ran the algorithm a few more times to check if this was a consistent result and found that it was.

C. Polynomial Regression

As a comparison to our logistic regression models, we decided to fit 3 polynomial regressions (degrees 1, 2, and 3) using the top 10 PCs to see if we could achieve even more accurate predictions. Although polynomial regression isn't typically used for binary outputs, we modified the predictions in a way that would be comparable to the output of logistic regression. We decided to use a threshold of 0.8 on the output of the model to determine whether it would predict 1 or 0 for masculinity, if the prediction was above the threshold it would be 1 and if it was under it would be 0. We also used a 10-fold cross-validation on the 3 different degree models, and observed that the mean squared error decreased for both training and testing sets as degree increased. The MSE increased from degree 2 to 3 most likely due to overfitting which would cause the model to capture noise and undermine the accuracy of predictions.

Degree	**Mean Training Error**	**Mean Testing Error**
1	0.024974	0.025393
2	0.004059	0.021682
3	0.000000	0.059451

D. K-Means Clustering with PCA

Finally, to gain a clearer understanding of the distribution of participants, we fit k-means clustering with PCA (clustering was also done without PCA but did not yield any meaningful additional results). This was done with 3 clusters and 5 PCs. 5 PCs were chosen for this to make more interpretable graphs, and 3 clusters were chosen as more clusters did not show clear and distinct differences between clusters. This model discovered something that was not immediately apparent with earlier PCA analysis - that there were 2 distinct clusters, those who answered most of the questions and those who didn't answer most of the questions (the non-answerers were also likely to be >65, be unemployed, or not exercise). This appears in PC1 in the pair graph on the right. Note that despite being a very clear cluster in PC1, the green group is completely indistinct in the other PCs, making it clear that it is not a meaningful cluster.

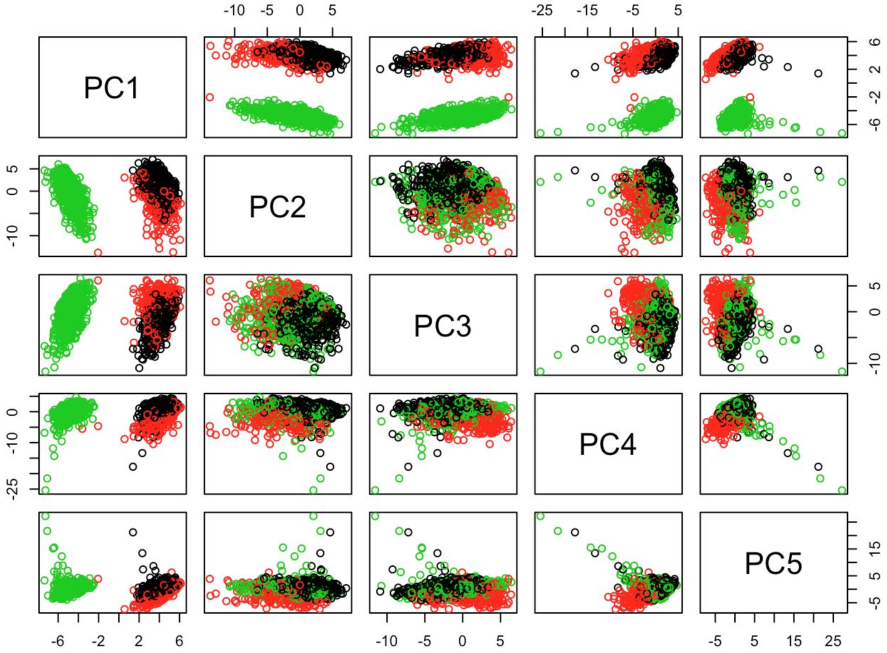


Figure 4: Initial k-means clustering with PCA

A second k-means PCA cluster analysis was performed (displayed below). In this one, the dataset was modified to not include the cluster of participants that did not answer most questions. This analysis formed clusters that held up to much closer scrutiny, but it is still clear that the clustering gets murky for the lower-variance PCs. Defining characteristics of these clusters were found by looking at the highest weights of the PCs that separated out the groups - for instance, the black cluster gets a low score on PC2, and thus the defining characteristics must be highly negatively weighted to separate it out - and predictors where the clusters' means were significantly different.

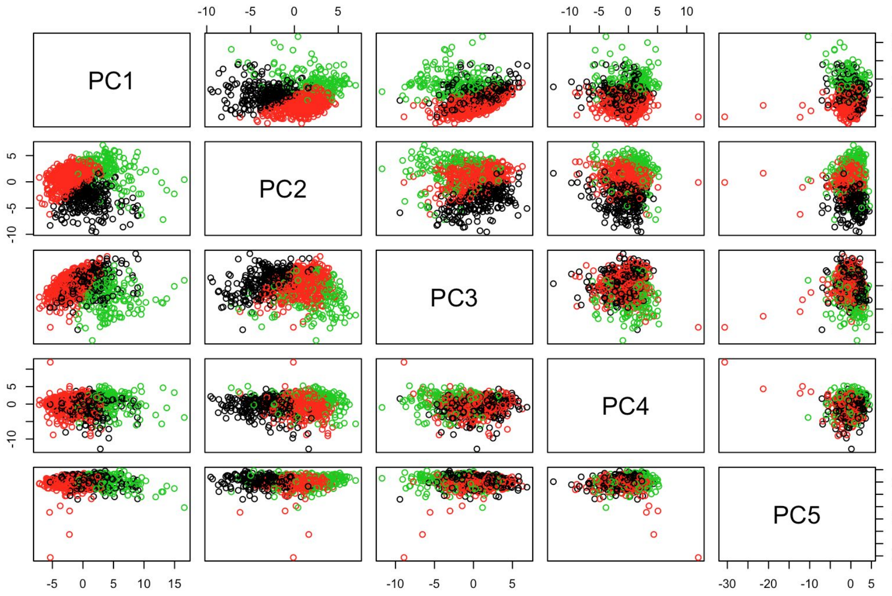


Figure 5: k-means clustering w/PCA after removal of participants not answering majority of questions

VI. Results and Conclusion

A. Feature Importance

From the logistic regression results, we found that the best model with the lowest average test MSE used 10 PCs. Below is the variance explained by each principal component and the most important feature contributing to the principal component.

PC#	**Explained Variance Ratio**	**Highest Weighted Feature**
PC1	0.131	q0009_Not employed-retired (Which of the following categories best describes your employment status?)
PC2	0.052	q0025_0003_No children (Do you have any children?)
PC3	0.045	q0025_0003_No children (Do you have any children?)
PC4	0.029	q0025_0003_Not selected (Do you have any children? - Option: No children)
PC5	0.025	q0004_0002_Not selected (Where have you gotten your ideas about what it means to be a good man? - Option: Mother/mother figures)
PC6	0.022	q0011_0004_Not selected (In which ways would you say it's a disadvantage to be a man at your work right now? - Option: Other)

PC7	0.0196	q0020_0001_Read their physical body language to see if they are interested (When you want to be physically intimate with someone how would you gauge their interest?)
PC8	0.019	q0020_0005_Not selected (When you want to be physically intimate with someone, how do you gauge their interest? - Option: It’s not always clear)
PC9	0.018	q0005_Yes (Do you think society puts pressure on men in a way that is unhealthy or bad?)
PC10	0.018	q0020_0004_Not selected (When you want to be physically intimate with someone, how do you gauge their interest? - Option: Every situation is different)

Note that although only the most contributive feature for each PC is shown, there are various other features that contribute only marginally less, if not equally. What we found wasn’t entirely expected because we didn’t consider the nature of the data taken from the survey, which limits the interpretability of our results. For example, once we ran k-means clustering, we found that PC1 mostly identifies 65+ men who didn’t fill out a lot of the questions. We can observe that a lot of the most contributive features for each PC indicates that the question was not answered.

The feature that came up multiple times as most contributive to the top PCs was q0025_0003_No children, which was answering the question “Do you have any children?”. This does make sense, since having no children could be representative of many other features that would affect masculinity. For example, having no children could mean that a subject is younger, and therefore puts more importance on being masculine since that always seems important to the younger generations. It could also mean not having a “paternal instinct”, which is usually associated with being warm and comforting and less masculine. So it make sense that the predictor of having any children would be correlated with masculinity.

The other question that was highest weighted for a few of the first 10 PCs was “When you want to be physically intimate with someone how would you gauge their interest?”. Intuitively, this question does seem to be strongly linked to one’s thoughts on masculinity in romantic and sexual relationships. Someone who felt themselves to be more masculine would be more likely to not put as much importance on getting an explicit or verbal sign of interest and just go with how they feel in a physically intimate situation. This agrees with our findings, where the answer “Read their physical body language to see if they are interested” was more correlated with masculinity than the answer “Ask for a verbal confirmation of consent”.

B. Cluster Findings

From the k-means clustering with PCA, we found three distinct clusters (refer to figure 5). This corresponded to 3 different ideologies. The first group (black in the figure) is more likely to be single, more likely to be non-heterosexual, more likely to be older than 35, middle class, with children, and perceive men as getting the short stick in professional life. Most of these predictors are associated with a more conservative mindset. The second group (red) is is less likely to worry about their weight, hair, sexual prowess, etc. and think that men tend to make more money, but greater risk of being accused of sexual harassment and less likely to be promoted. These predictors mix attitudes from the first and third groups, making it clear that this is a moderate-ideology group. The third group (green) tended to be young men who worried about themselves and questioned societal expectations about masculinity.

In the cluster analysis, green and black clusters are always distinct from each other, while the red cluster tends to occupy either the space between them or be mixed in. This lends credence to the distinction between the three groups' ideologies. There is no hugely distinct clustering though, surprisingly. All of the groups border each other, and it is clear that perception of masculinity lies on a spectrum rather than having a few competing vastly different ideologies.

C. Next Steps and Possible Improvements

In the future, it might be useful for researchers to find additional concrete correlations between individual predictors, or use the findings of this paper to find correlations between ideologies about masculinity and other data. It is a useful way to see the ways ideologies change over time, as there is a strong correlation between age and ideology.

A limitation that we experienced while analyzing the data was the nature of the survey -- there were some significant open ended questions in the survey that had to be left out of the analysis, perhaps the answers could have been categorized into smaller subset of generalizations. Additionally, it seems that many surveyors left various questions unanswered which affected the interpretability of some of our results -- we don't exactly understand why these particular responses seem to be significant although we have a general hypothesis. Another limitation we experienced may be a result of the models we chose to implement. PCA works optimally in situations where the correlations are linear; however, self-perceived masculinity may be a much more complex relationship. We could explore this path by perhaps using a non-linear dimensional reduction method and comparing.

Links

Dataset: https://github.com/fivethirtyeight/data/tree/master/masculinity-survey

Code for this project: https://github.com/angela278/cogs109