Nate Silver 538 Masculinity Survey

538.com published an article on masculinty based on a survey they took with WNYC Studios. Since most of the article focuses on visualization (and they were kind enough to make the data available), I wanted to take a look and see what trends might come out through ML-based analytics.

https://fivethirtyeight.com/features/what-do-men-think-it-means-to-be-a-man/

The workflow I used can be broken up into 3 sections:

1. **Data wrangling**

2. **Unsupervised analysis using kmeans**

3. **Supervised analysis: What variables drive an idea (or lack of) masculinity?**

Section 1: Setup and Data Import

Load data

https://data.fivethirtyeight.com/

Review unique responses on questions

Get some idea of where to start cleanup

```
## # A tibble: 5 x 2
##   q0002                    n
##   <chr>                <int>
## 1 No answer                9
## 2 Not at all important   240
## 3 Not too important      541
## 4 Somewhat important     628
## 5 Very important         197
```

Change "No answer" to NA across dataframe

```
## # A tibble: 5 x 2
##   q0002                    n
##   <chr>                <int>
## 1 Not at all important   240
## 2 Not too important      541
## 3 Somewhat important     628
## 4 Very important         197
## 5 <NA>                     9
```

Change "Not selected" to 0 in one-hot columns

```
## # A tibble: 5 x 5
##   q0001              q0002              q0004_0001 q0004_0002 q0004_0003
##   <chr>              <chr>                   <dbl>      <dbl>      <dbl>
## 1 Somewhat masculine Somewhat important          0          0          0
## 2 Somewhat masculine Somewhat important          1          0          0
## 3 Very masculine     Not too important           1          0          0
## 4 Very masculine     Not too important           1          1          1
## 5 Very masculine     Very important              0          0          1
```

Encode ordinal columns

Function to encode ordinal columns

Call function to encode ordinal columns on individual questions

```
## # A tibble: 5 x 5
##   q0001 q0002 q0004_0001 q0004_0002 q0004_0003
##   <chr> <chr>      <dbl>      <dbl>      <dbl>
## 1 3     3              0          0          0
## 2 3     3              1          0          0
## 3 4     2              1          0          0
## 4 4     2              1          1          1
## 5 4     4              0          0          1
```

Encode remaining dummy columns

```
## # A tibble: 5 x 5
##   q0013_Other_.pl~ q0013_You_didn.~ q0013_You_didn.~ q0013_You_weren~
##              <dbl>            <dbl>            <dbl>            <dbl>
## 1               NA               NA               NA               NA
## 2               NA               NA               NA               NA
## 3               NA               NA               NA               NA
## 4               NA               NA               NA               NA
## 5               NA               NA               NA               NA
## # ... with 1 more variable: q0013_You_weren.t_sure_who_to_contact <dbl>
```

Impute NA's using missRanger

```
## # A tibble: 5 x 5
##   q0013_Other_.pl~ q0013_You_didn.~ q0013_You_didn.~ q0013_You_weren~
##              <dbl>            <dbl>            <dbl>            <dbl>
## 1                0                0                1                1
## 2                1                0                0                0
## 3                0                1                0                0
## 4                0                0                0                0
## 5                0                0                0                0
## # ... with 1 more variable: q0013_You_weren.t_sure_who_to_contact <dbl>
```

Center_scale, don't change weightings column

Weightings column is meant to weight the responses to US demographics.

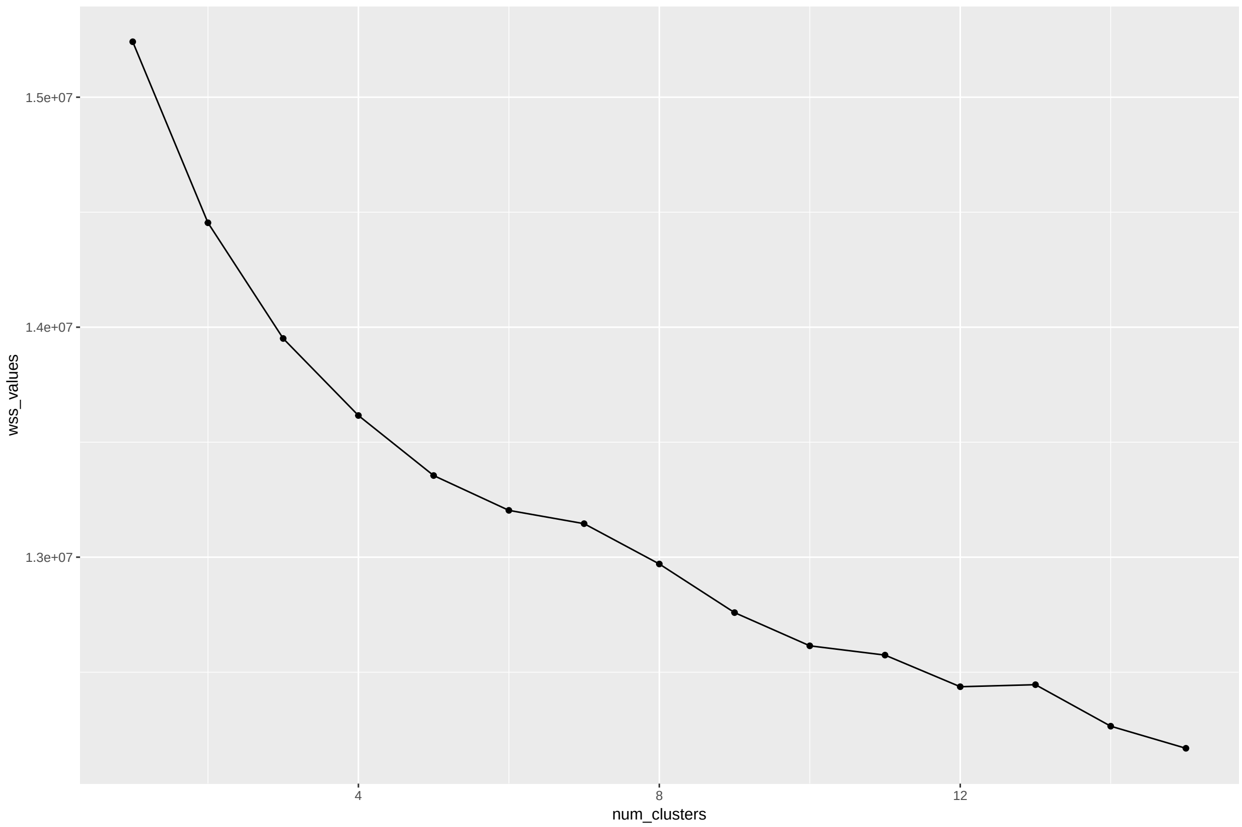
2. Unsupervised clustering with kmeans

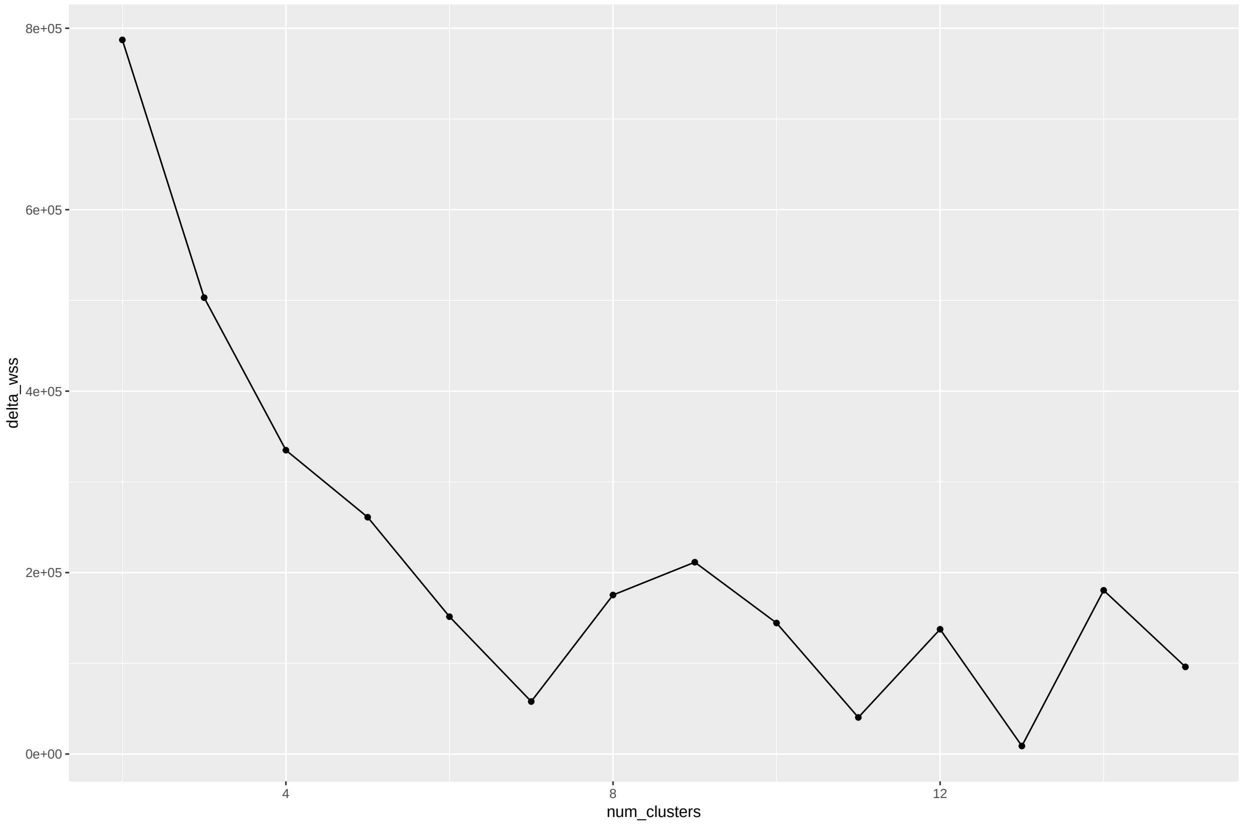
Prepare data: resample rows by the US demographic weightings.

Kmeans clustering

Determine best number of clusters

https://uc-r.github.io/kmeans_clustering

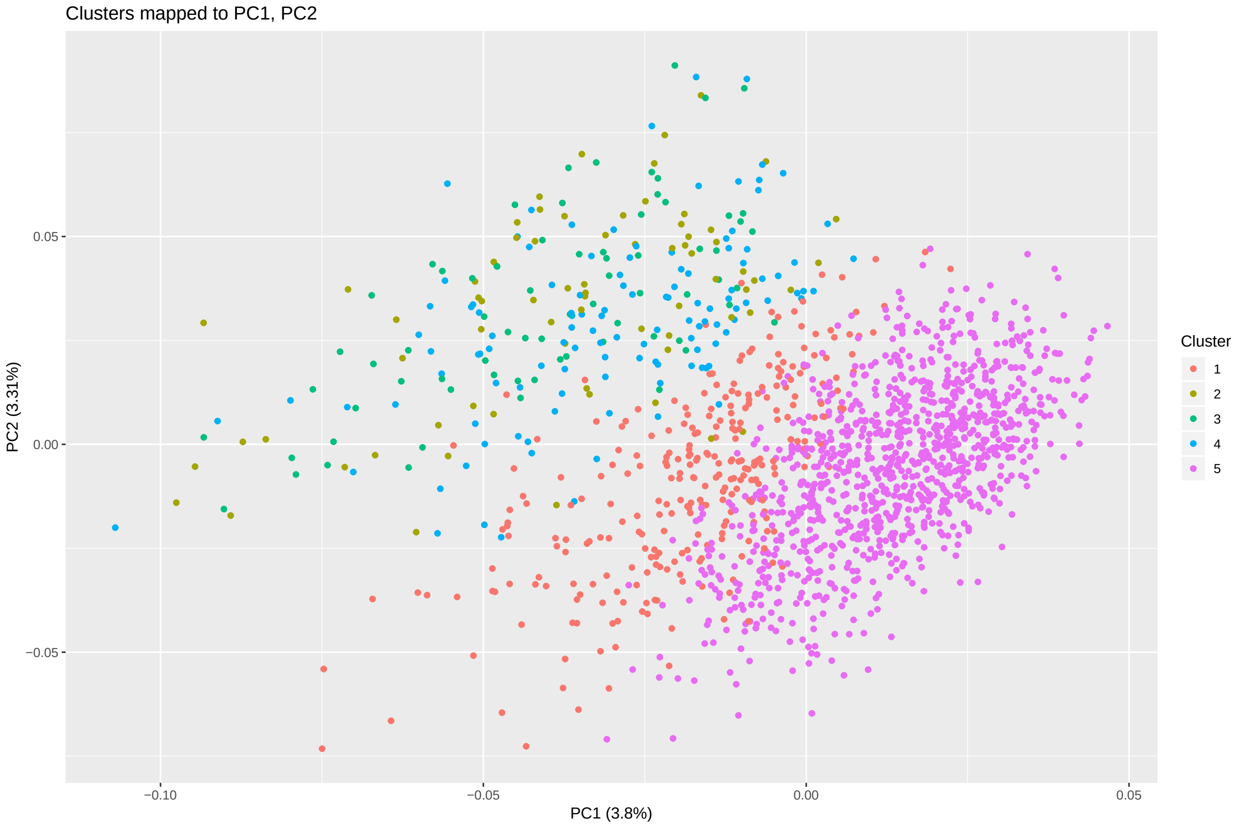

1.5e+07
1.4e+07
1.3e+07
wss_values
4
8
12
num_clusters

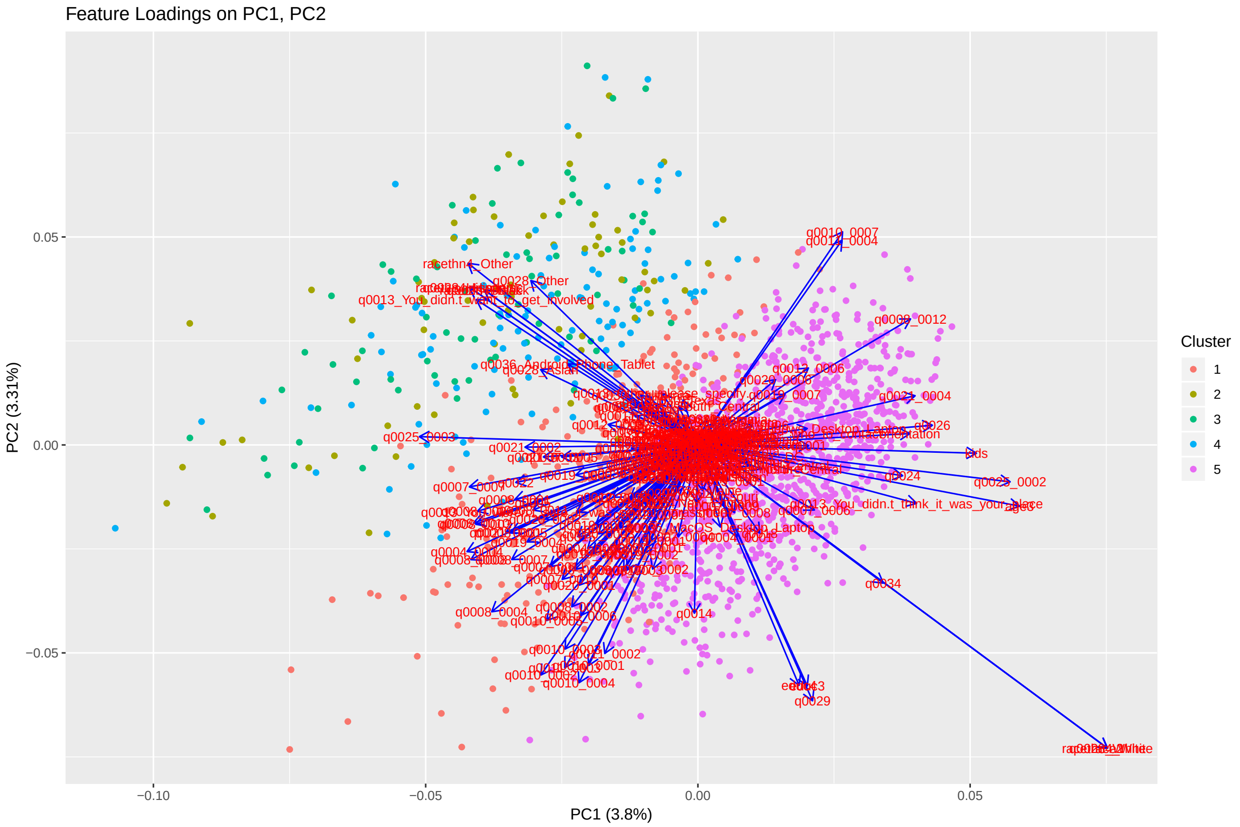

8e+05
6e+05
4e+05
2e+05
0e+00
delta_wss
4
8
12
num_clusters

Cluster data by 5 clusters and visualize

Determine distinctive features

https://cran.r-project.org/web/packages/ggfortify/vignettes/plot__pca.html





Graphs are a little hard to read but

Trend is observed between low-PC1/hi-PC2 :: hi-PC1/low-PC2

Determine the questions driving the features

```
## # A tibble: 167 x 4
##        PC1     PC2 cols            distance
##      <dbl>   <dbl> <chr>              <dbl>
##  1  0.244  -0.236  q0028_White        0.339
##  2  0.244  -0.236  racethn4_White     0.339
##  3  0.244  -0.236  race2              0.339
##  4  0.0682 -0.199  q0029              0.211
##  5 -0.0933 -0.179  q0010_0002         0.202
##  6  0.0651 -0.188  educ3              0.199
##  7 -0.0707 -0.185  q0010_0004         0.198
##  8  0.191  -0.0476 age3               0.197
##  9 -0.137   0.141  racethn4_Other     0.197
## 10  0.0602 -0.187  educ4              0.197
## # ... with 157 more rows
```

So just from an unsupervised perspective, the largest breaks in the survey population (reweighted to US demographics) are race and education level.

The biggest respondents group appear to be highly educated, non-minority.

Supervised learning: Predicting what does masculinity look like?

Set target variable: "How masculine are you?"

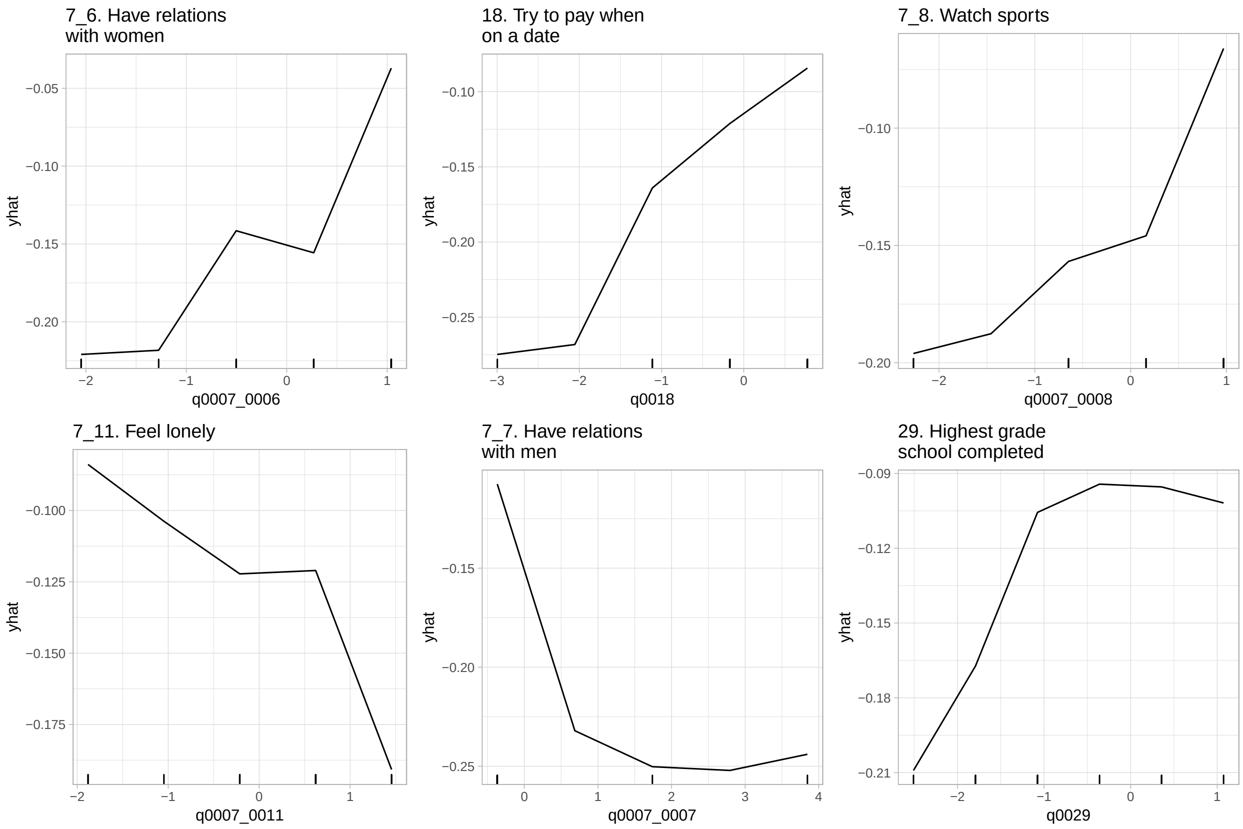
Remove other questions essentially re-asking the same question.

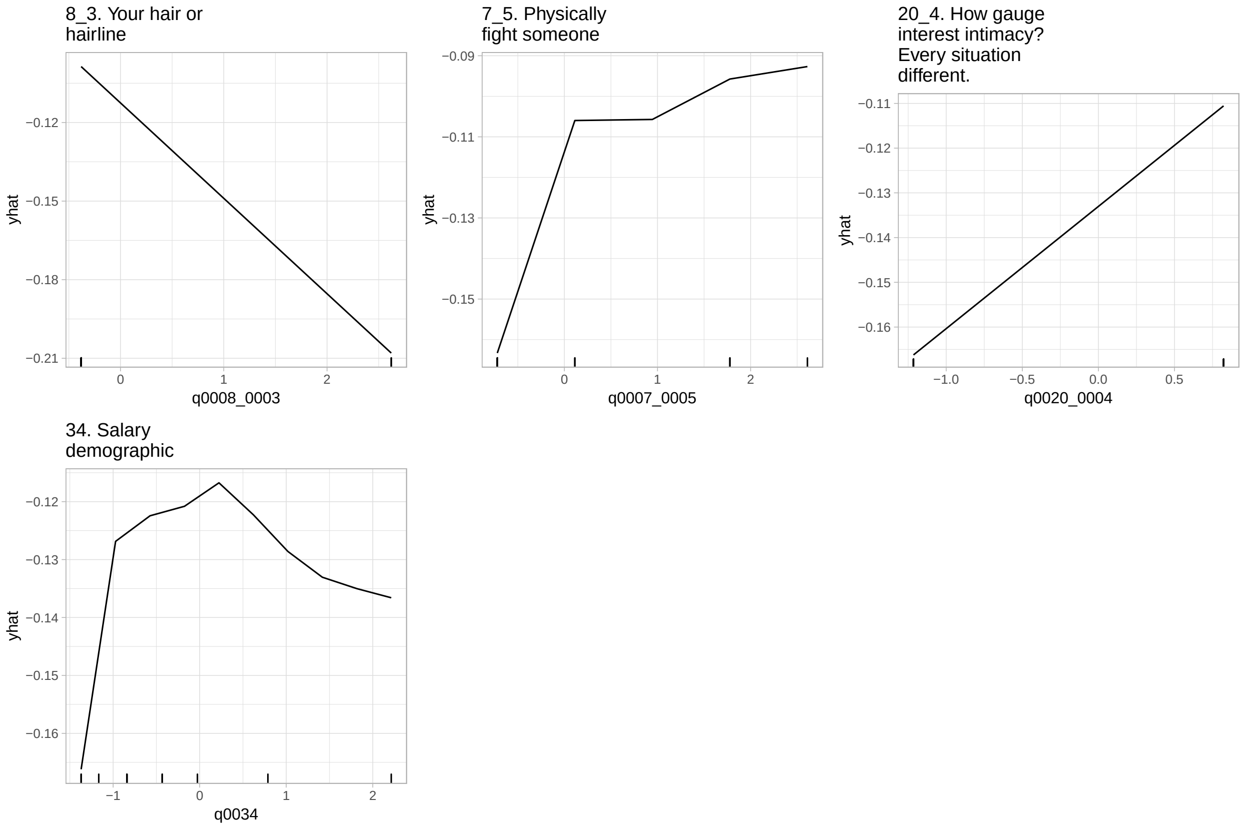
Random forest model will provide feature importances.

```
##    importance     pvalue      query
## 1   0.4322939 0.00990099 q0007_0006
## 2   0.3315004 0.00990099      q0018
## 3   0.3106693 0.00990099 q0007_0008
## 4   0.2633456 0.00990099 q0007_0011
## 5   0.2470338 0.00990099 q0007_0007
## 6   0.2382067 0.00990099      q0029
## 7   0.2229917 0.00990099 q0008_0003
## 8   0.2202866 0.00990099 q0007_0005
## 9   0.1964099 0.00990099 q0020_0004
## 10  0.1958545 0.00990099      q0034
```

Create partial dependency plots for the top 10 variables.

https://bgreenwell.github.io/pdp/articles/pdp.html





Questions, by importance:

7_6. Have relations with women

18. Try to pay when on a date

7_8. Watch sports

7_11. Feel lonely

7_7. Have relations with men

29. Highest grade school completed

8_3. Your hair or hairline

7_5. Physically fight someone

20_4. How do you gauge someone's interest in intimacy? Every situation different.

34. Salary demographic

Consider the shapes of the respones. Yes, more masculine men have more sexual relations with women, but compare which questions show a fairly linear response vs. other patterns.

For instance, any level of sexual contact with other men results in a complete lack of masculine identity.

Surprisingly: A moderate sports watching only results in moderate masculinity. I would have expected masculinity to increase faster with sports watching, similar to the physical fighting response.

Masculine men also admit to some level of loneliness.

Masculinity seems to require a moderate level of education. However, all but the lowest level of salary earnings feel masculine.