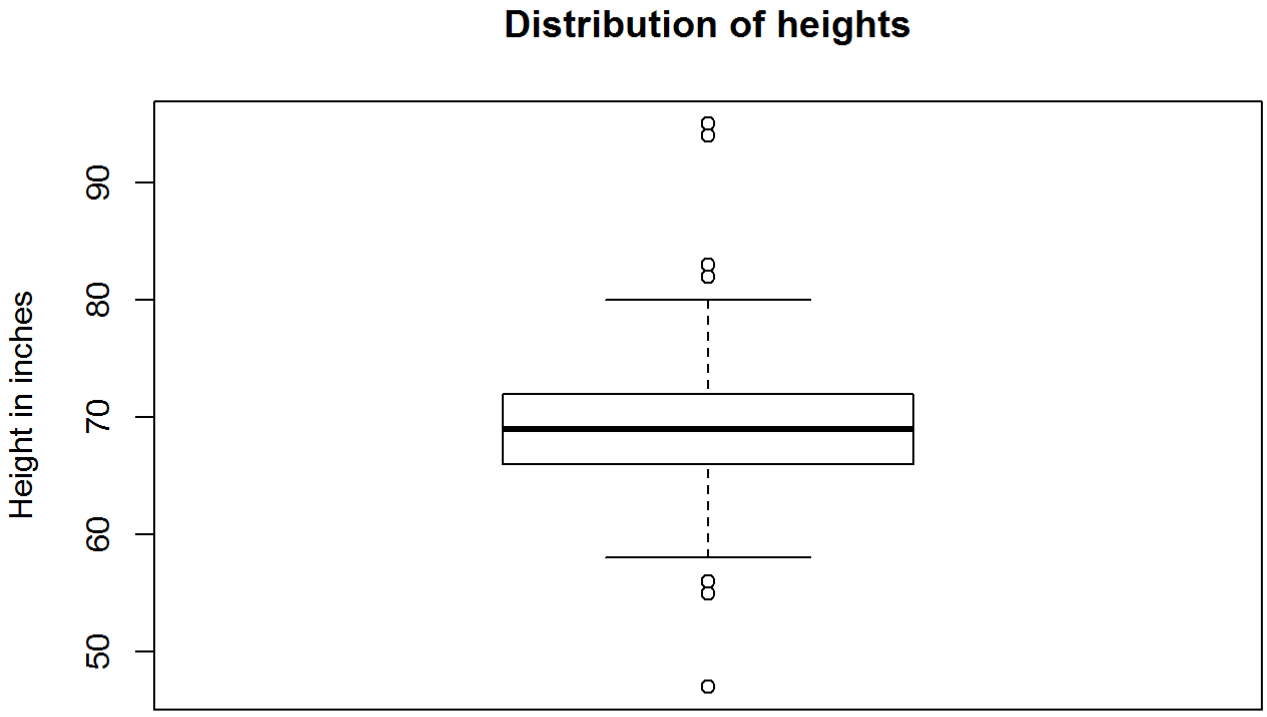
ADS Project5 Decode Online Dating

In this project, we conduct extensive exploratory analysis using various visualization tools, clustering, and regression models on datasets obtained from Kaggle and OkCupid, one of the most widely used dating apps to derive meaningful insights into online dating. We hope to help everyone in the fray to increase their opportunity of finding the one.

First of all, we will explore the two data sets and get a general sense of the dating scene.

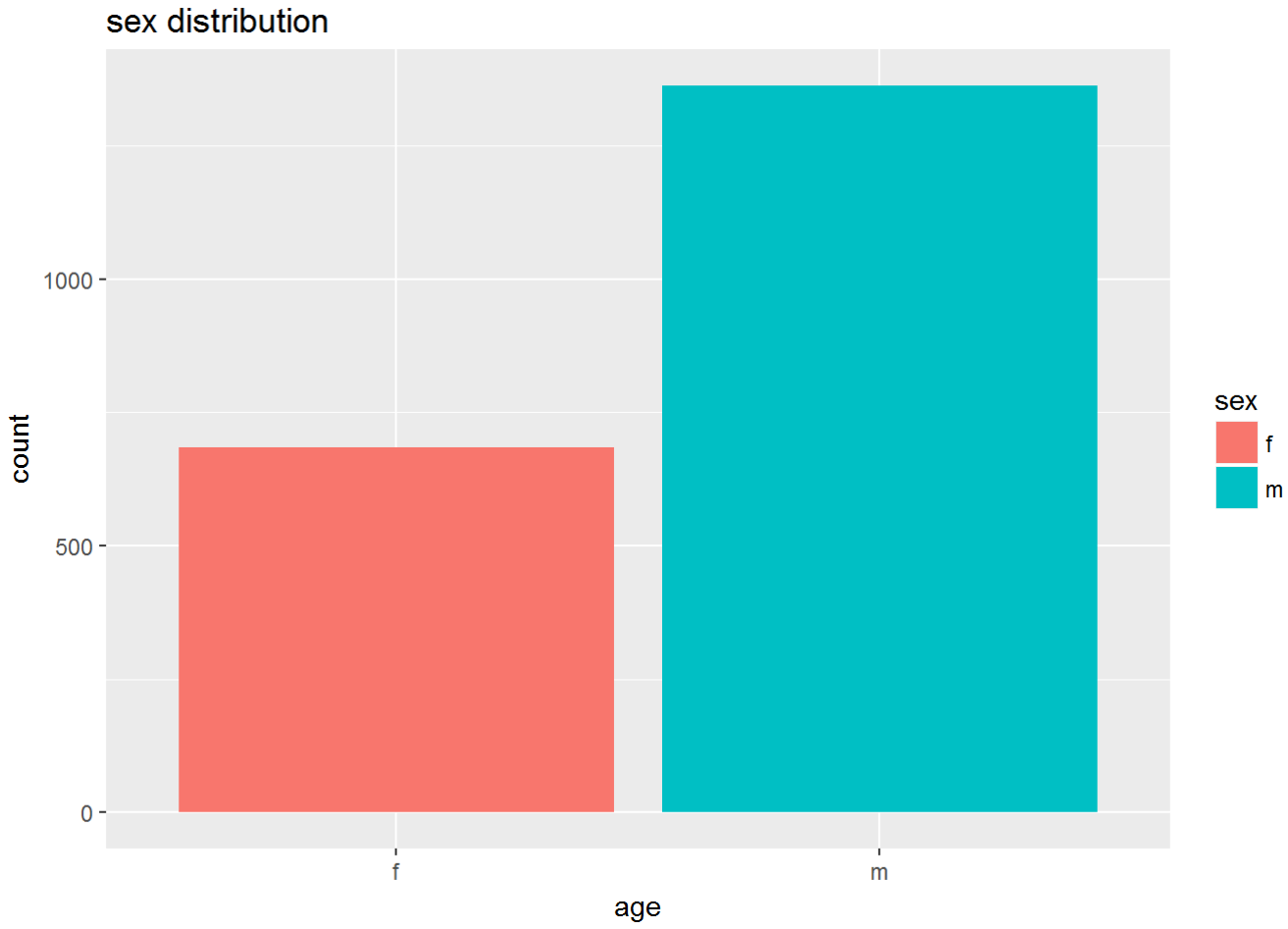
First, let's take a look at the distribution of height, which is converted into inches (1 foot = 12 inches). We can see the distribution is widely spreaded and some of them do not make sense in terms of heights. Therefore, we removed those rows that have extreme heights from our data.

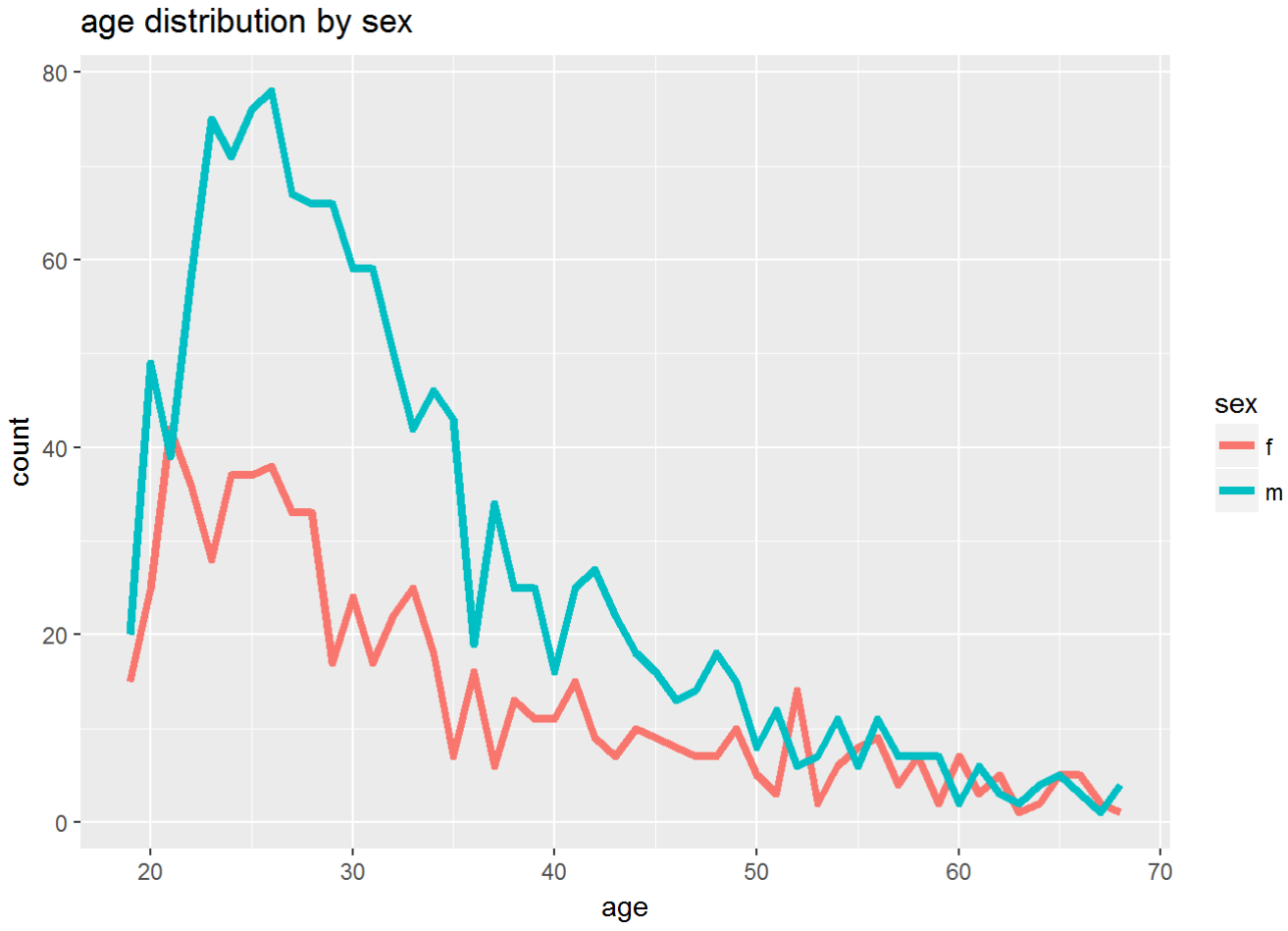



```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```

Let's start from some general information of the users.

Sex and age distributions


sex distribution
1000
500
0
count
f
m
age
sex
f
m

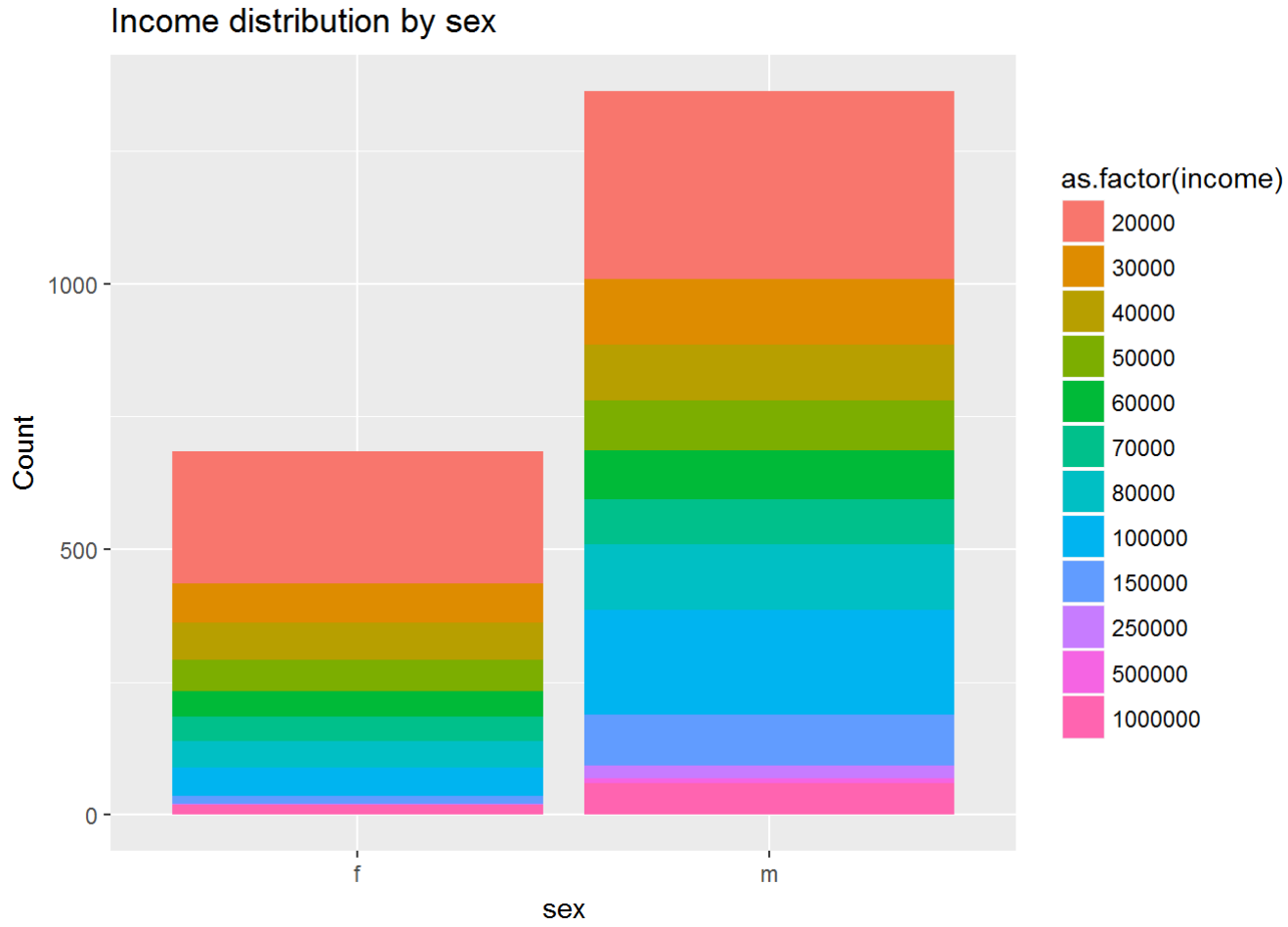

age distribution by sex
80
60
40
20
0
count
20
30
40
50
60
70
age
sex
f
m

Surprisingly there are a lot more male using the dating app. Most of the users are in their mid-20s.

Income difference by sex.

We also want to know the users' income information.

```
##  [1]   20000   30000   40000   50000   60000   70000   80000  100000
##  [9]  150000  250000  500000 1000000
```

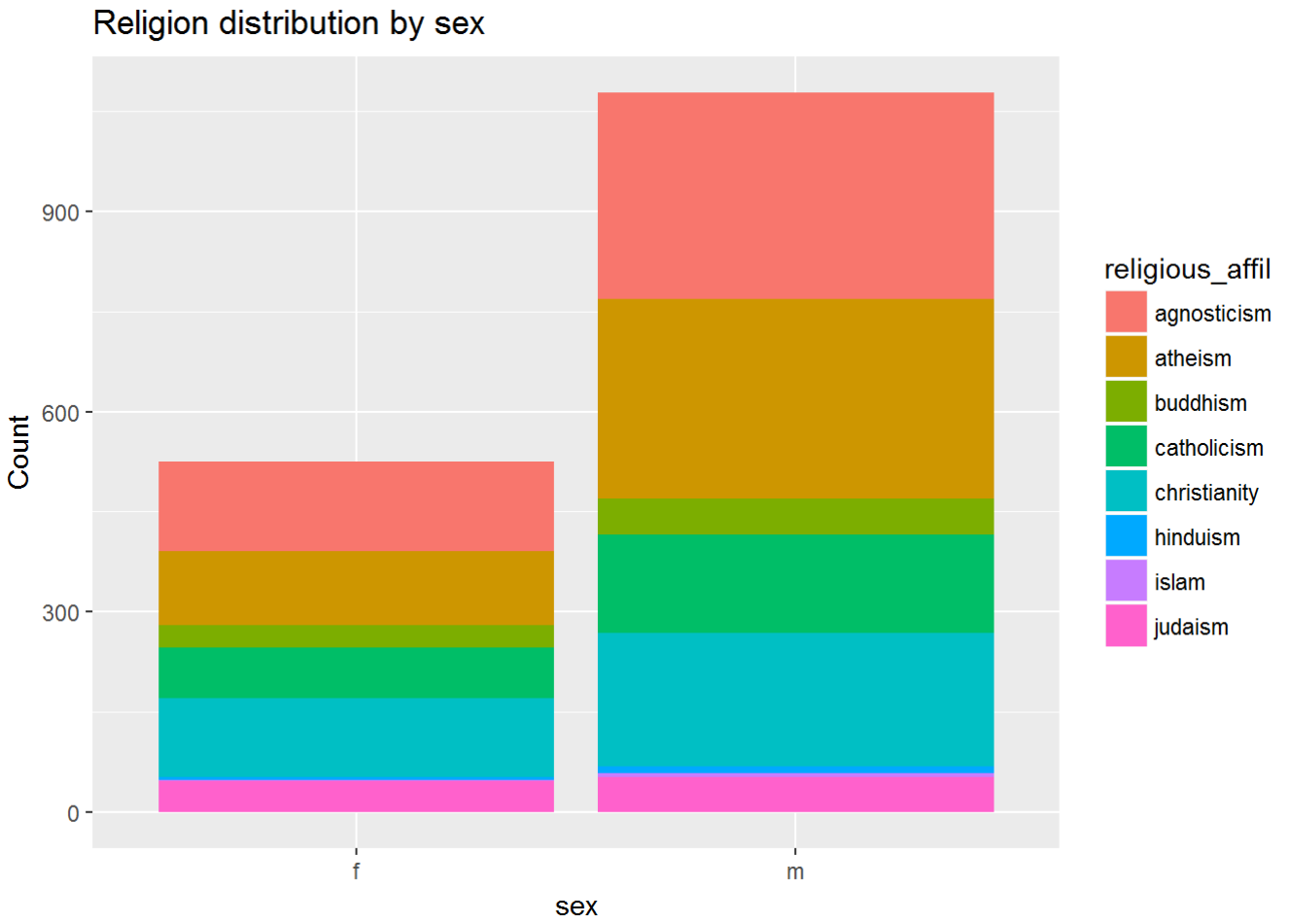


Women's incomes are lower than men's. A larger percentage of women, for example, report an income of 20,000. Men also have a much higher representation incomes above 100,000.

Religions difference

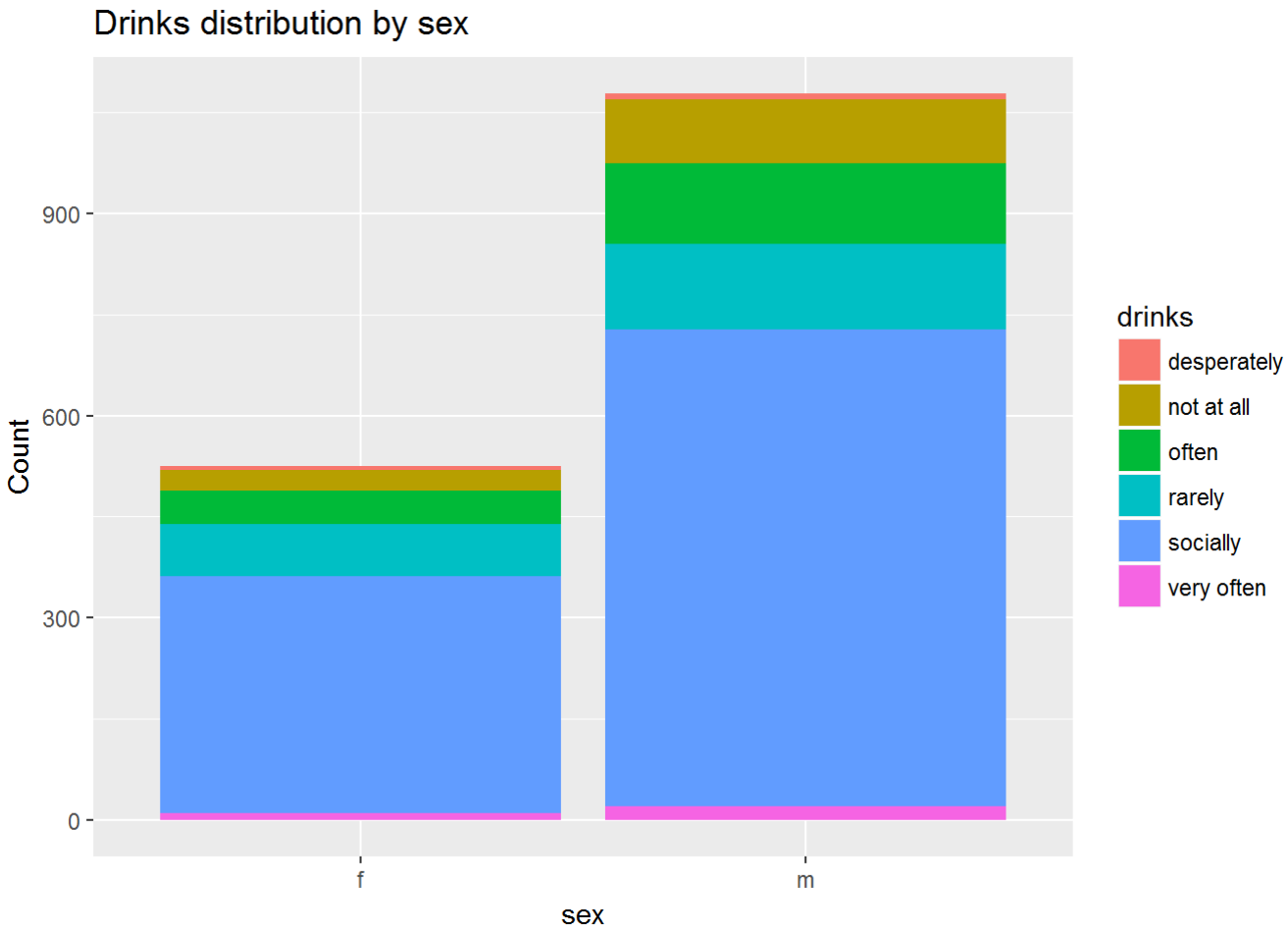
The religion data is a little complicated, as it includes both the affiliation and how serious they are about their belif. For simplicity, we will only focus on the affiliation attributes in this case.

```
## [1] "agnosticism and laughing about it"
## [2] "judaism and somewhat serious about it"
## [3] "islam but not too serious about it"
## [4] "christianity and very serious about it"
```



A large proportion of men users reporting to be atheists and agnostics as their affiliation, and women users have similar distribution with a slightly lower percentage.

Drinking habit

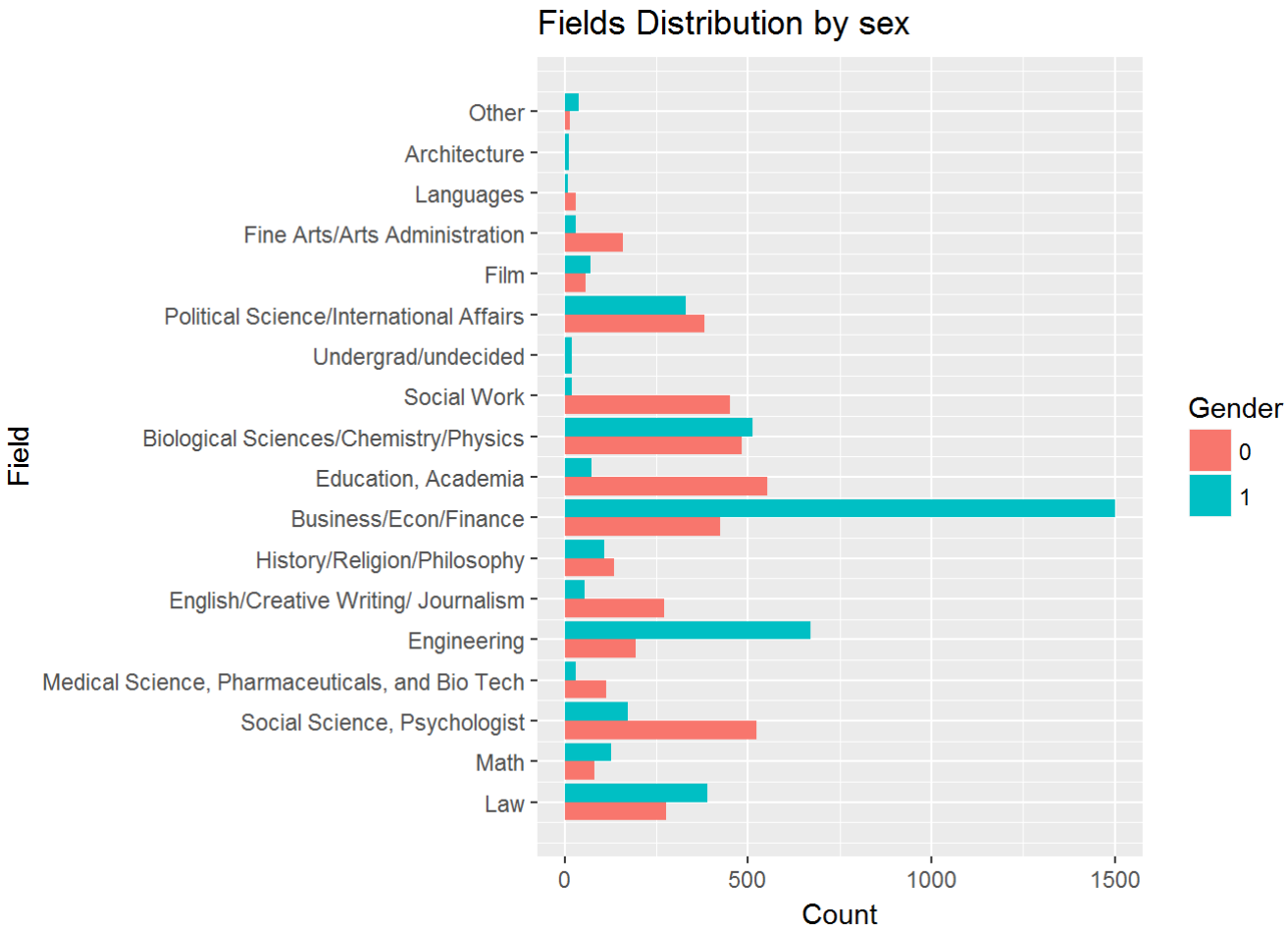


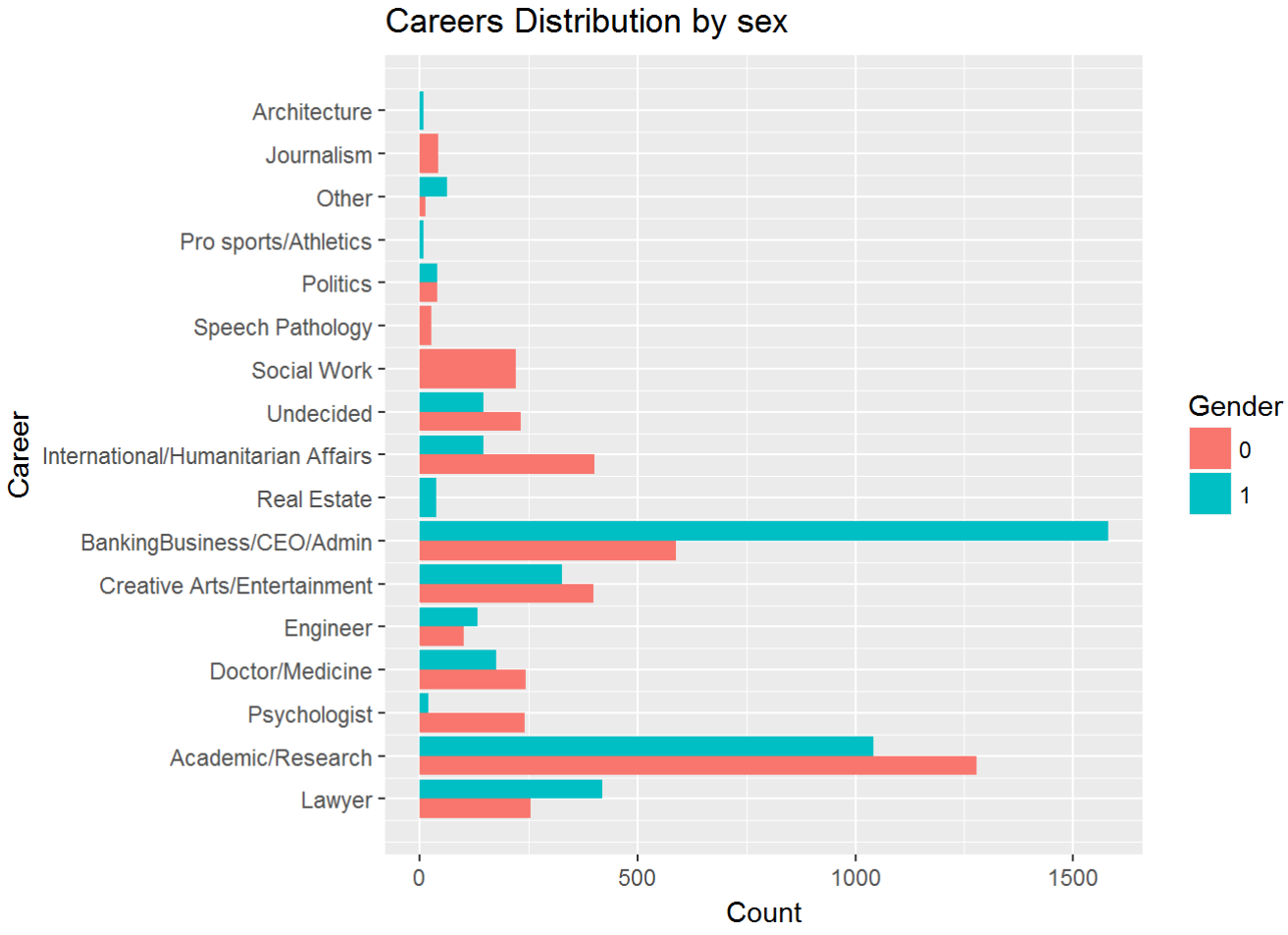
More than half of the okcupid users are socially drinkers, and men users have a slightly greater tendency to drink more often than women users.

Speed Dating Data

We used the speed dating data to get a sense of the relationship between different attributes and number of matches.

Field and Career distribution by gender

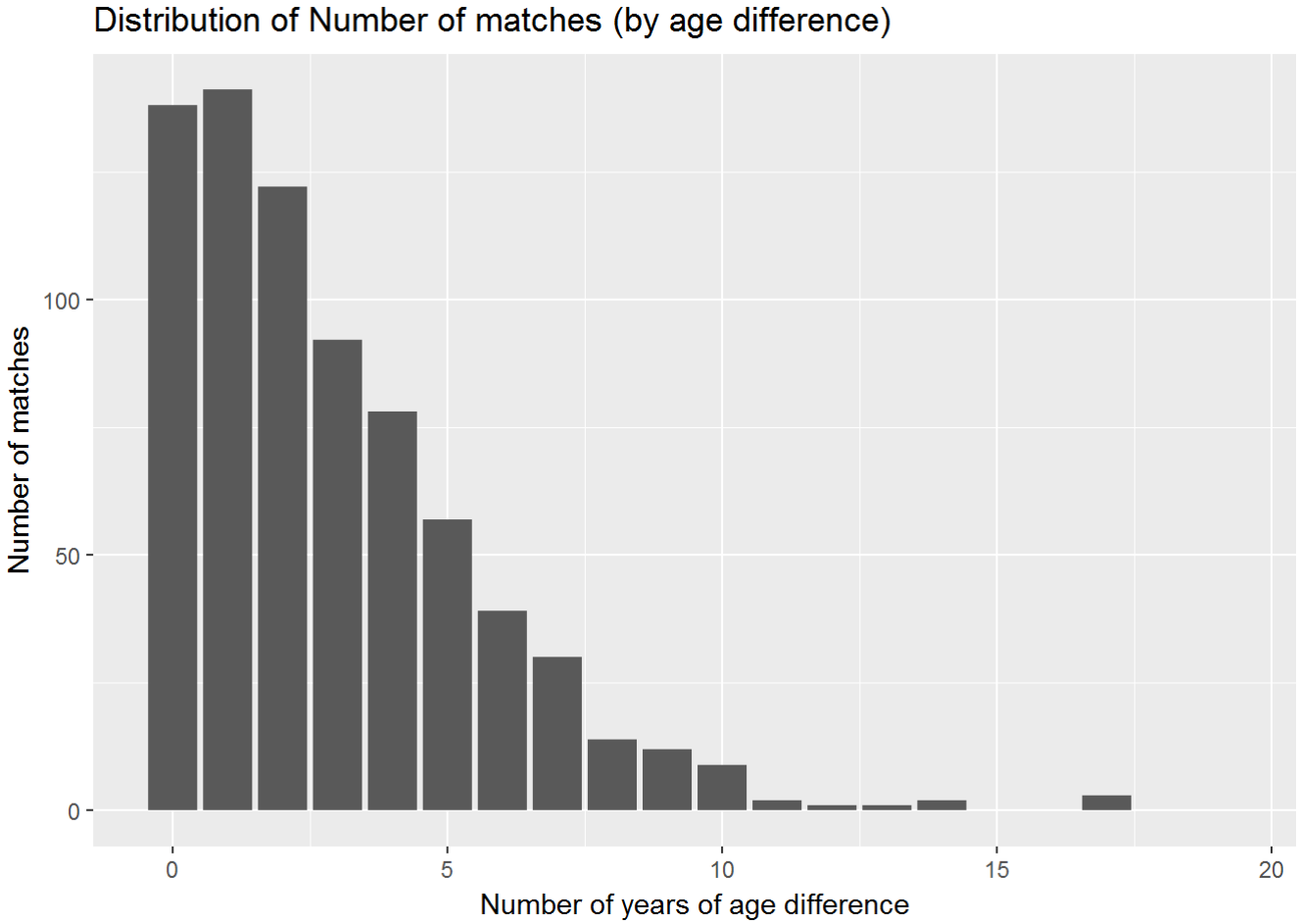

Fields Distribution by sex
Other
Architecture
Languages
Fine Arts/Arts Administration
Film
Political Science/International Affairs
Undergrad/undecided
Social Work
Biological Sciences/Chemistry/Physics
Education, Academia
Business/Econ/Finance
History/Religion/Philosophy
English/Creative Writing/ Journalism
Engineering
Medical Science, Pharmaceuticals, and Bio Tech
Social Science, Psychologist
Math
Law
Field
0
500
1000
1500
Count
Gender
0
1


Careers Distribution by sex
Architecture
Journalism
Other
Pro sports/Athletics
Politics
Speech Pathology
Social Work
Undecided
International/Humanitarian Affairs
Real Estate
BankingBusiness/CEO/Admin
Creative Arts/Entertainment
Engineer
Doctor/Medicine
Psychologist
Academic/Research
Lawyer
Career
0
500
1000
1500
Count
Gender
0
1

For OkCupid users, the majority proportion of the male users are major in Business/Finance field, and working in banking or other academic careers.

Does age really matter? Relation between Age difference & no. of matches.

Here we draw the barplot of number of matches by number of years of age difference of the participants. Clearly, age does matter to most people! Most of the matches were made when the two participants have 0 to 3 years of age difference.

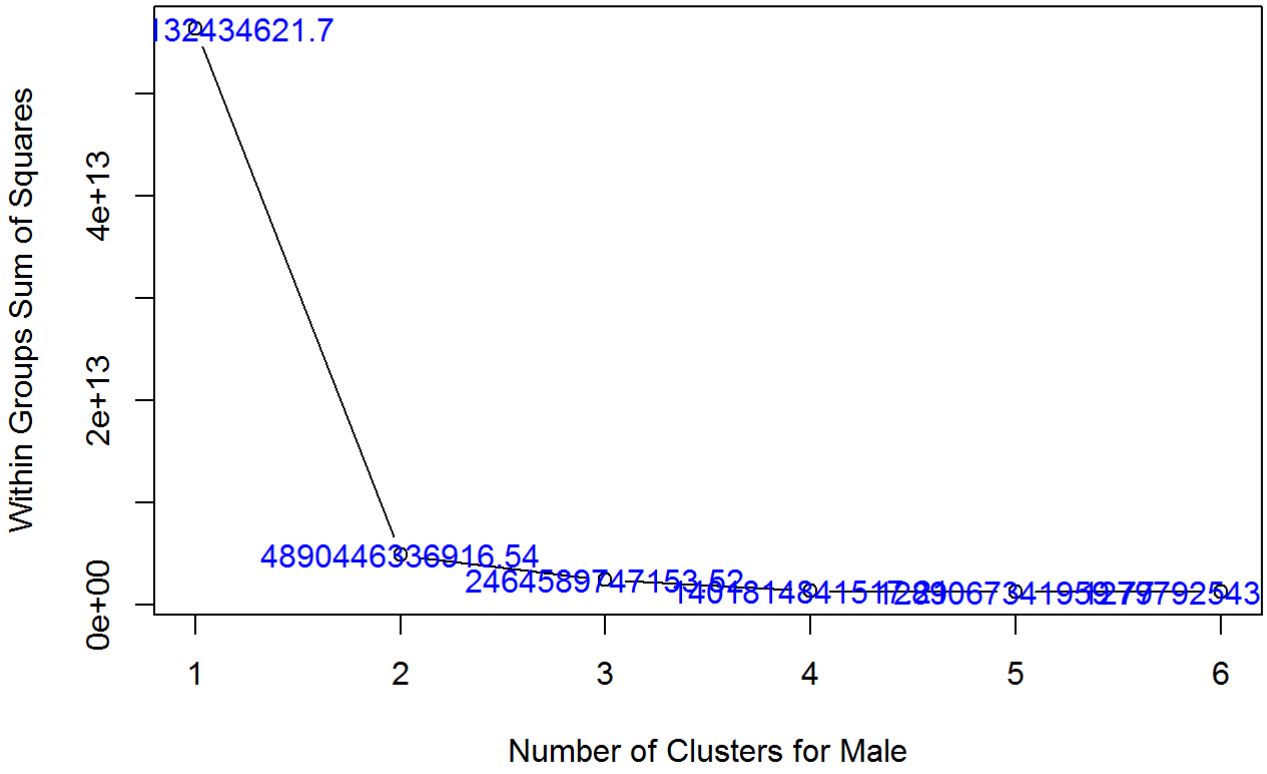


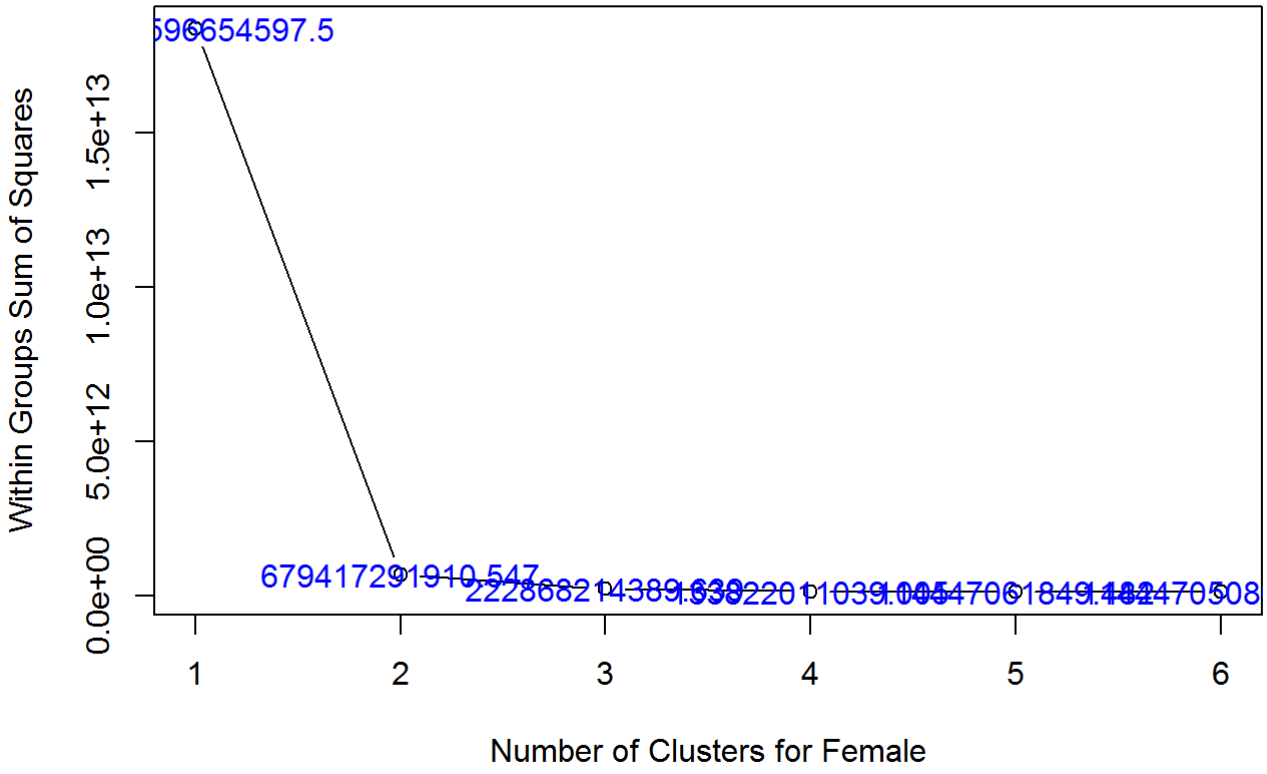
Next, we explore the different categories within the online dating population with K-Means Clustering, so that people can identify themselves with a category and understand what kind of competition they are facing.

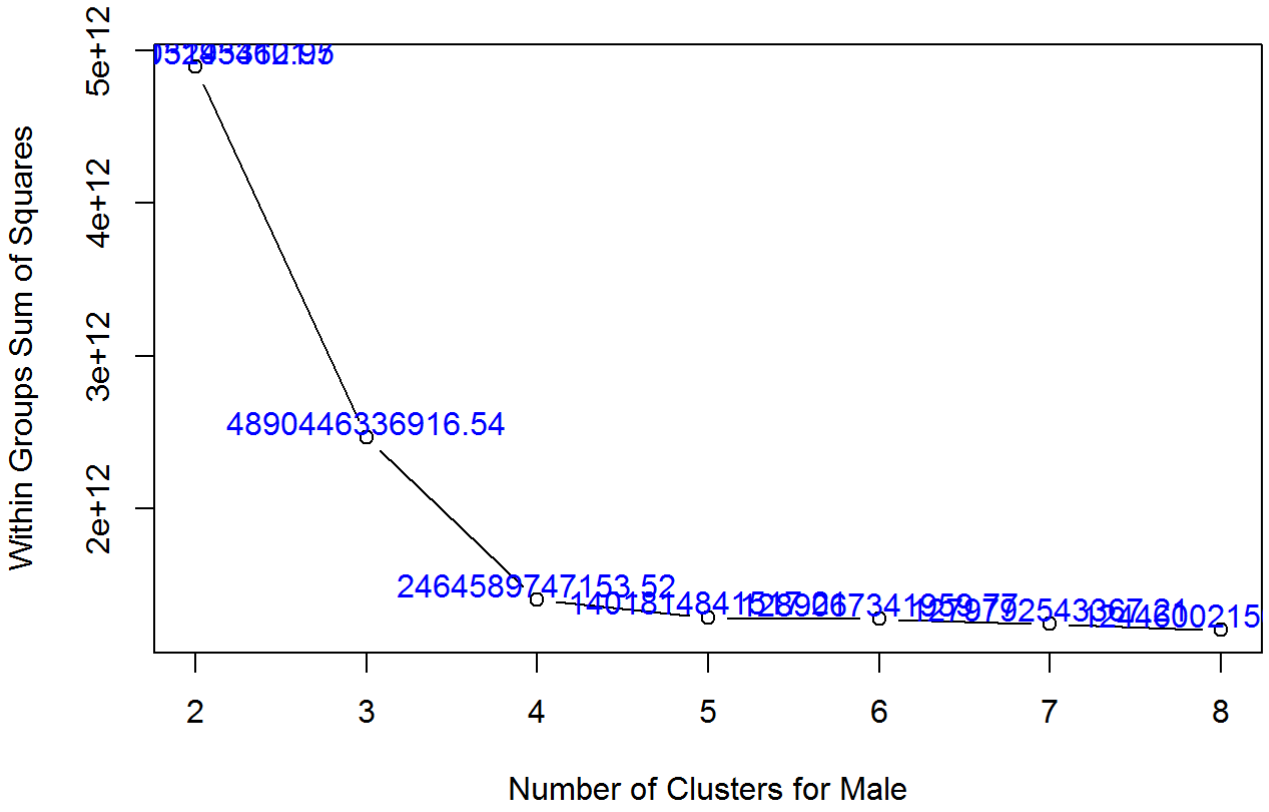
We performed K-Means on three of the continuous variables: age, height, and income, and did it on male and female respectively.

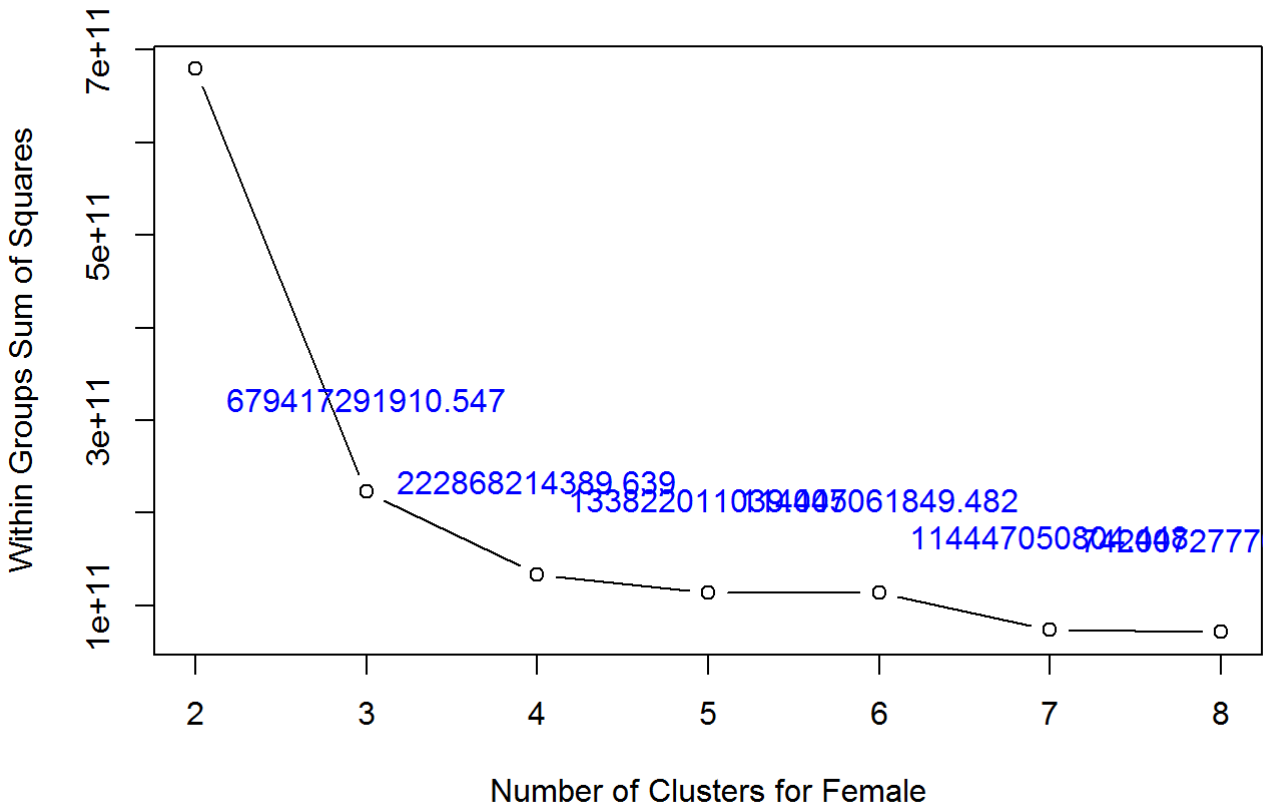
Optimize the Number of Clusters for K-Means

The plot trails off after 3, suggesting 3 or 4 might be a good choice for K.


Within Groups Sum of Squares
0e+00
2e+13
4e+13
1
2
3
4
5
6
Number of Clusters for Male

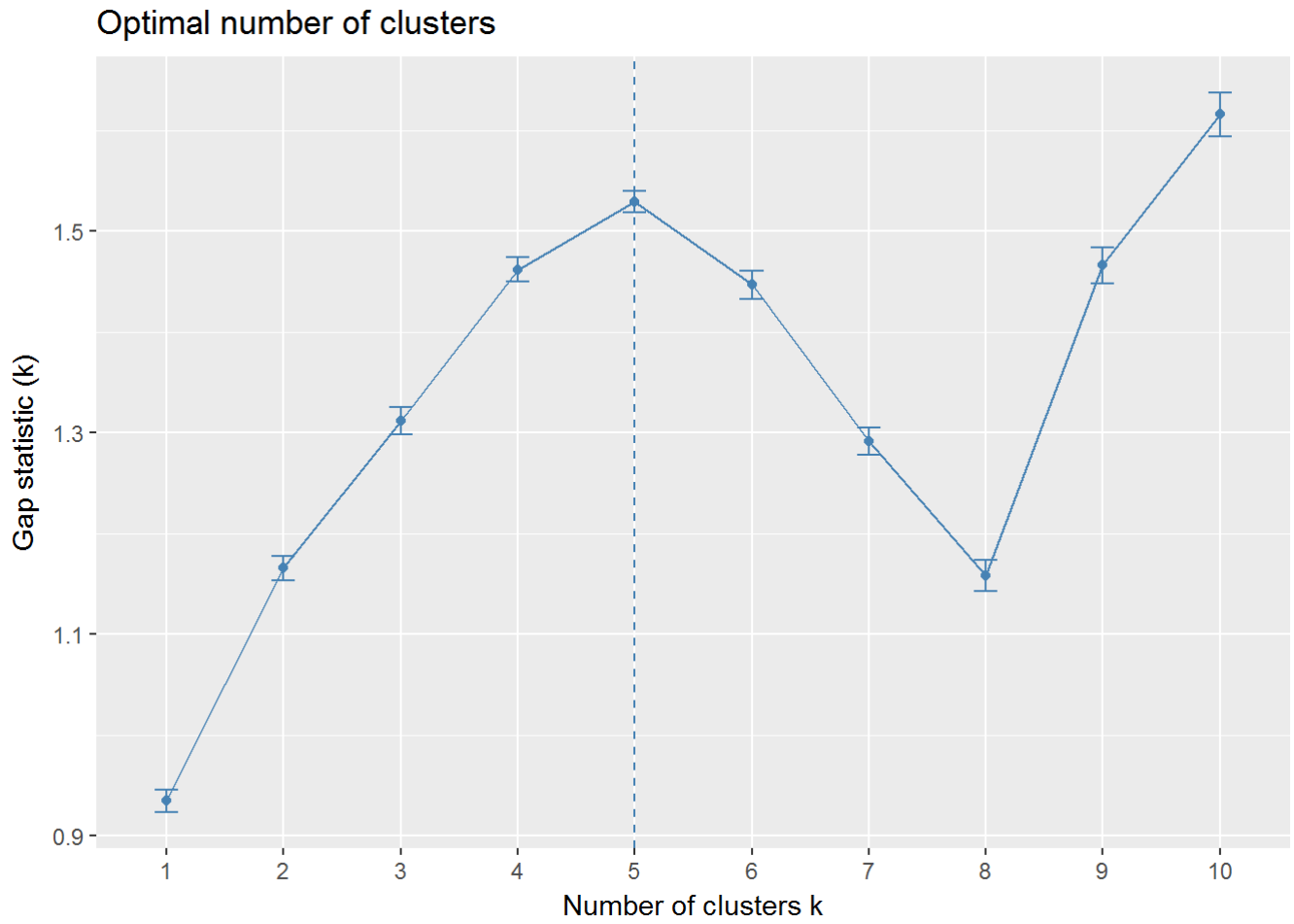

Within Groups Sum of Squares
0.0e+00
5.0e+12
1.0e+13
1.5e+13
1
2
3
4
5
6
Number of Clusters for Female

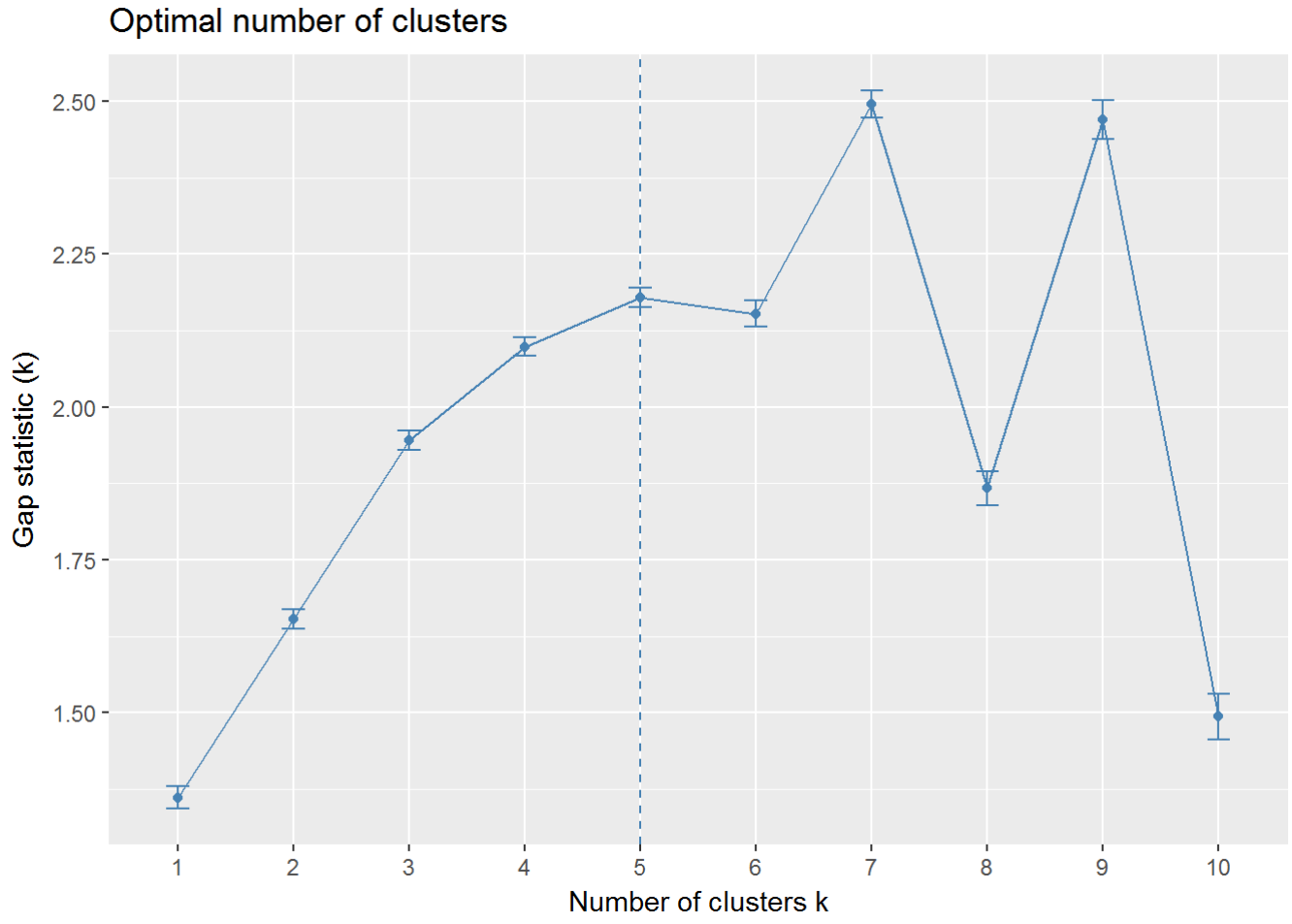




Gap staistics: A more rigorous method to help us find the number of clusters.

Since the larger the gap statistics the better and we should choose the smallest value of K such that the gap stat is within one standard deviation of the next gap stat, the output plot suggests K = 5.


Optimal number of clusters
Gap statistic (k)
0.9
1.1
1.3
1.5
1
2
3
4
5
6
7
8
9
10
Number of clusters k

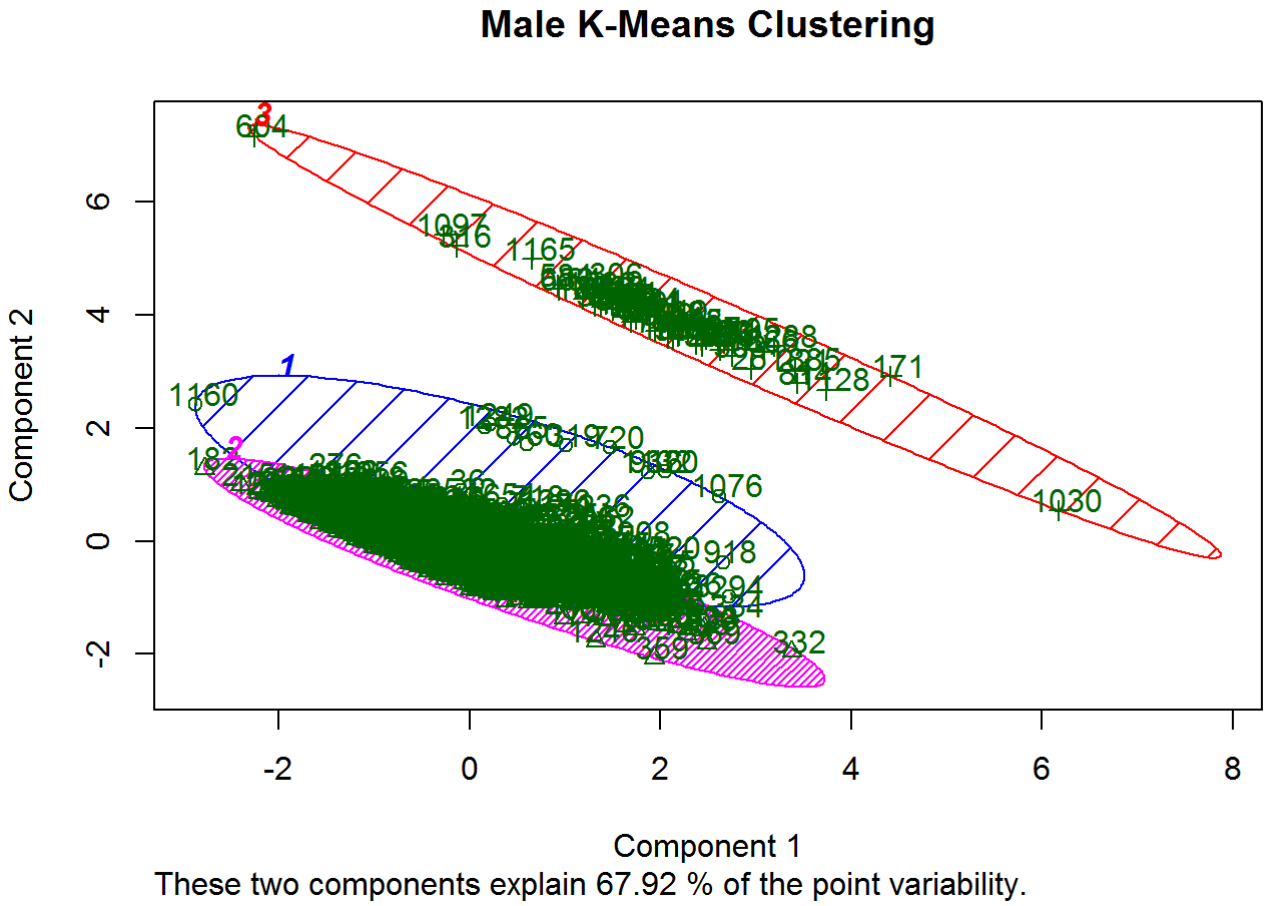

Optimal number of clusters
Gap statistic (k)
1.50
1.75
2.00
2.25
2.50
1
2
3
4
5
6
7
8
9
10
Number of clusters k

$$\hat{k} = \text{smallest } k \text{ such that } \text{Gap}(k) \geqslant \text{Gap}(k+1) - s_{k+1}.$$

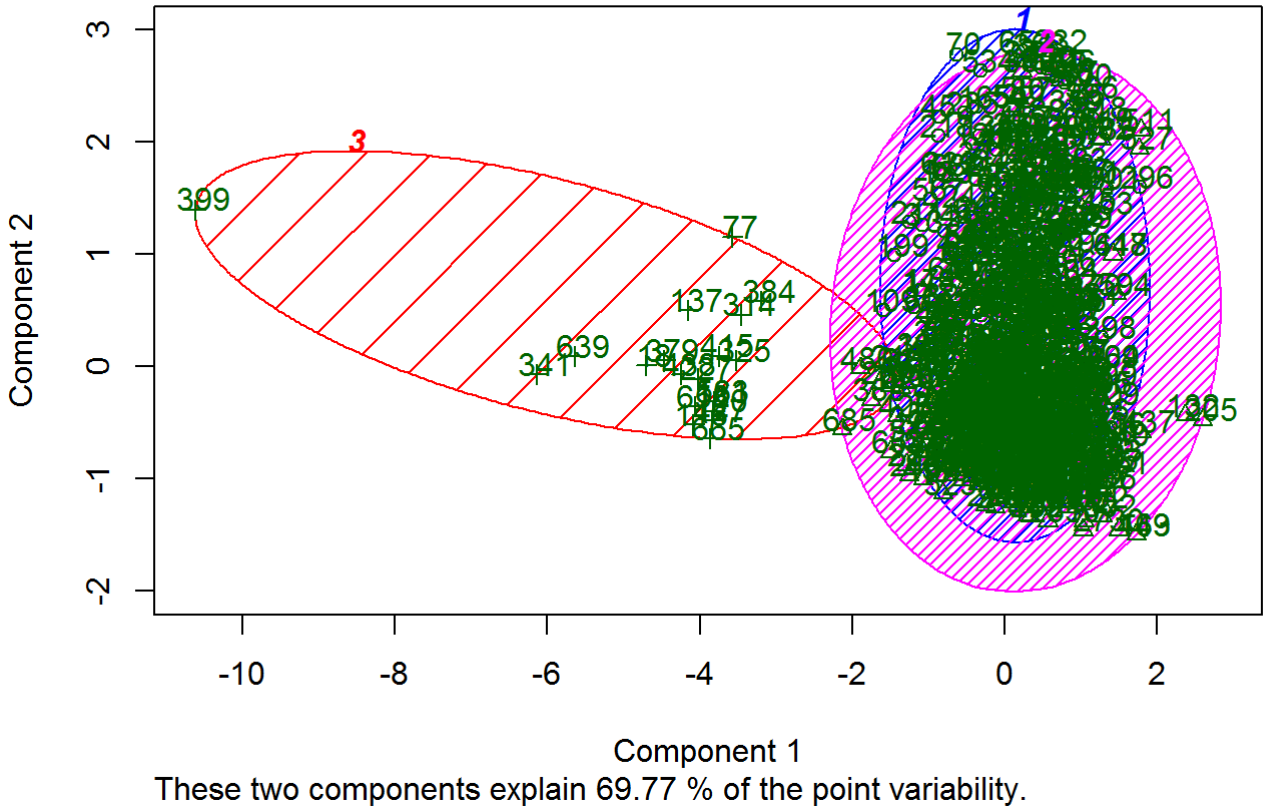
Choose the optimal k according to the above

Visualize the clustering

We then visualize the clustering solution and find that there seems to be two distinct groups when visualizing with a 2-dimension representation. This result is repeated when K = 3,4,5.



These two components explain 67.92 % of the point variability.



These two components explain 69.77 % of the point variability.

Cluster Sizes

It seems that for both male and female, there are three distinct groups and one of them stand out. What are the characteristics of this outstanding group? We further investigate using within cluster statistics.

```
## [1]  130 1190   62
```

```
## [1]   72 1310
```

```
## [1] 213 462  20
```

```
## [1]  20 675
```

Mean for each variable within each cluster for male and femal respectively

```
##   cluster      age   height     income
## 1       1 38.75385 71.17692  194615.38
## 2       2 31.88151 70.50420   51411.76
## 3       3 29.32258 70.51613 1000000.00
```

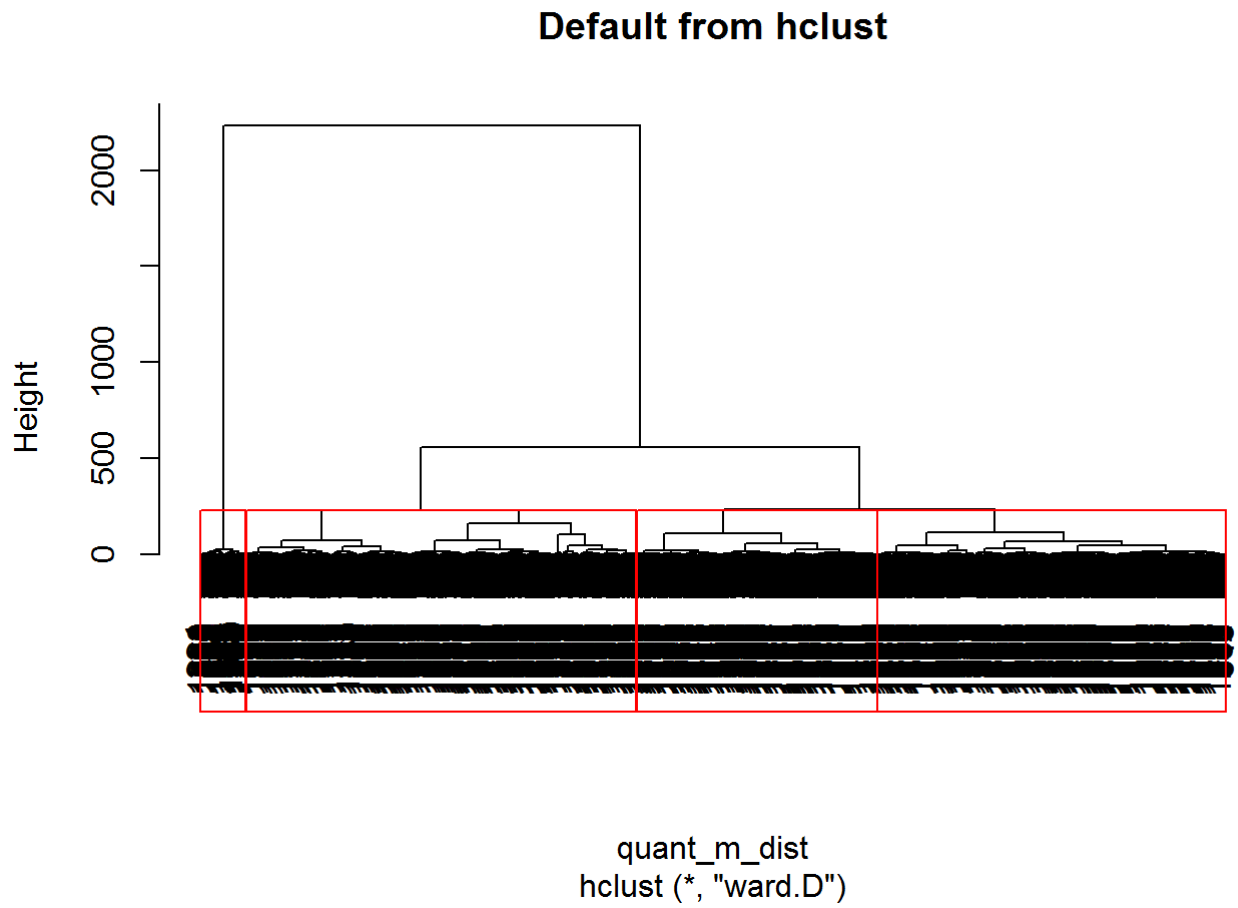
```
##   cluster      age   height     income
## 1       1 41.35681 65.28638   84553.99
## 2       2 29.85714 65.30303   28593.07
## 3       3 26.10000 66.90000 1000000.00
```

Median for each variable within each cluster for male and femal respectively

```
##   cluster  age height  income
## 1       1 36.0   71.0  150000
## 2       2 29.0   70.0   40000
## 3       3 27.5   70.5 1000000
```

```
##   cluster  age height income
## 1       1 41.0   65.0  8e+04
## 2       2 26.5   65.0  2e+04
## 3       3 25.0   64.5  1e+06
```

Agglomerative Hierarchical Clustering: starts with each individual observation as a cluster; then the two closest points as a new cluster



```
## m_groups_4
##   1   2   3   4
## 526 324  62 470
```

With a broad understanding of clusters and dataset, we continued to add more categorical variables into clustering. Here, we used drinks habits, diet habits, drugs habits, and body type for people

Since we had both numeric and categorical variables in the model, we could not use k-means as clustering method. Instead, we used two-step clustering.

Two-step clustering includes following major steps:

1. Calculating distance
2. Choosing a clustering algorithm
3. Selecting the number of cluster
4. Cluster and explain the cluster

In the first step, we want to calculate distance to get measurement of similarity between observations. Since the data type is mixed data types, we could not work with Euclidean distance. Here we chose Gower distance.

Gower distance is a distance metric that scales the data between 0 and 1. And then, the distance matrix is calculated based on weights (e.g. average). The metrics used for each data type are described below:

1. quantitative: range-normalized Manhattan distance
2. categorical: variables of k categories are first converted into k binary columns and then the Dice coeffcient is used

We calculated Gower distance with daisy function in R and checked out what is the most similar and dissimilar pair to do the sanity check.

Here is the most similar pair for female users:

```
##     sex age height income   drinks              diet drugs body_type
## 321   f  26     64  40000 socially mostly vegetarian never     curvy
## 53    f  27     64  40000 socially mostly vegetarian never     curvy
```

Here is the most dissimilar pair for female users:

```
##     sex age height  income   drinks            diet     drugs    body_type
## 500   f  65     64   20000   rarely mostly anything     never      average
## 341   f  21     74 1000000 socially        anything sometimes full figured
```

In the second step, we selected partitioning around medoids (PAM). PAM is similar with k-means, but the cluster center for PAM are restricted as the observations and could handle a custom distance matrix (i.e. Gower distance). The detail pf the algorithm is following:

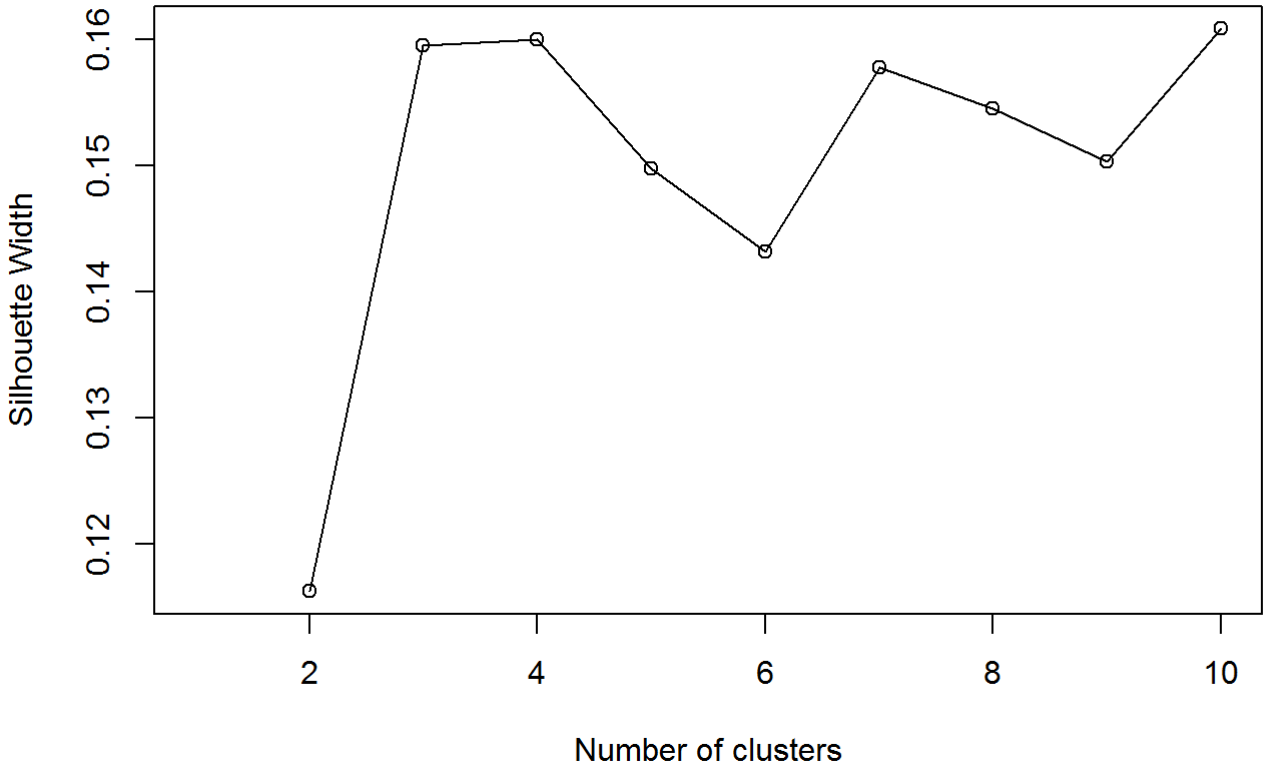
1. Choose k random entities to become the medoids
2. Assign every entity to its closest medoid (using our custom distance matrix in this case)

3. For each cluster, identify the observation that would yield the lowest average distance if it were to be re-assigned as the medoid. If so, make this observation the new medoid.

4. If at least one medoid has changed, return to step 2. Otherwise, end the algorithm.

After choosing PAM as clustering method, we used silhouette width, an internal validation metric which is an aggregated measure of how similar an observation is to its own cluster compared its closest neighboring cluster. The width ranges from -1 to 1 and the higher the better.

Below is the result for female users with different numbers of clusters. The figure suggests that number of cluster is 3.



With number of cluster = 3, we calculated cluster and interpretate cluster results via summary and visualization.

Below is variable summary for each cluster of female users. We could conclude that female users in the cluster 1 tend to have curvy body type and never take drugs. Female users in the cluster 2 tend to little bit aged and have average body type, higer income and never take drugs. Female users in the cluster 3 sometimes take drugs.

```
## [[1]]
##       age            height          income                drinks
##  Min.   :18.00   Min.   :56.00   Min.   :  20000   desperately:  4
##  1st Qu.:23.00   1st Qu.:64.00   1st Qu.:  20000   not at all : 20
##  Median :27.00   Median :65.00   Median :  30000   often      : 18
##  Mean   :28.45   Mean   :65.53   Mean   :  69964   rarely     : 47
##  3rd Qu.:31.00   3rd Qu.:67.00   3rd Qu.:  50000   socially   :185
##  Max.   :65.00   Max.   :94.00   Max.   :1000000   very often :  7
##
##                   diet           drugs              body_type
##  mostly anything    :141   never    :277   curvy         :139
##  mostly vegetarian  : 39   often    :  4   thin          : 35
##  strictly anything  : 30   sometimes:  0   fit           : 34
##  anything           : 26                   a little extra: 18
##  mostly other       : 13                   full figured  : 18
##  strictly vegetarian:  7                   athletic      : 13
##  (Other)            : 25                   (Other)       : 24
##     cluster
##  Min.   :1
##  1st Qu.:1
##  Median :1
##  Mean   :1
##  3rd Qu.:1
##  Max.   :1
##
##
## [[2]]
##       age            height          income                drinks
##  Min.   :19.00   Min.   :59.00   Min.   :  20000   desperately:  0
##  1st Qu.:32.00   1st Qu.:63.00   1st Qu.:  40000   not at all : 25
##  Median :42.00   Median :65.00   Median :  60000   often      : 12
##  Mean   :42.03   Mean   :65.13   Mean   :  81627   rarely     : 47
##  3rd Qu.:52.00   3rd Qu.:67.00   3rd Qu.:  80000   socially   :168
##  Max.   :67.00   Max.   :72.00   Max.   :1000000   very often :  0
##
##                   diet           drugs              body_type
##  mostly anything    :148   never    :247   average       :123
##  mostly vegetarian  : 28   often    :  1   fit           : 52
##  anything           : 25   sometimes:  4   athletic      : 21
##  strictly anything  : 17                   thin          : 18
##  mostly other       : 11                   full figured  : 16
##  strictly vegetarian:  8                   a little extra: 14
##  (Other)            : 15                   (Other)       :  8
##     cluster
##  Min.   :2
##  1st Qu.:2
##  Median :2
##  Mean   :2
##  3rd Qu.:2
##  Max.   :2
##
##
## [[3]]
```

```
##       age             height          income                drinks
##  Min.   :18.00   Min.   :55.00   Min.   :  20000   desperately:  5
##  1st Qu.:22.00   1st Qu.:63.00   1st Qu.:  20000   not at all :  2
##  Median :25.00   Median :65.00   Median :  20000   often      : 35
##  Mean   :28.01   Mean   :65.35   Mean   :  67840   rarely     : 10
##  3rd Qu.:31.00   3rd Qu.:67.00   3rd Qu.:  40000   socially   :103
##  Max.   :68.00   Max.   :74.00   Max.   :1000000   very often :  7
##
##                   diet          drugs              body_type     cluster
##  mostly anything    :56   never    :  0   curvy          :48   Min.   :3
##  strictly anything  :31   often    :  4   average        :28   1st Qu.:3
##  mostly vegetarian  :27   sometimes:158   thin           :23   Median :3
##  anything           :25                   fit            :20   Mean   :3
##  strictly vegetarian: 7                   full figured   :11   3rd Qu.:3
##  strictly other     : 5                   a little extra:10   Max.   :3
##  (Other)            :11                   (Other)        :22
```

Based on the results analysis, it seems like cluster 1 and 2 are more simliar and cluster 3 is more seperate from other two groups. Will the visualization result follow the analysis? Here, we used t-distributed stochastic neighborhood embedding, or t-SNE, to visualize many variables in a lower dimensional space.This method is a dimension reduction technique that tries to preserve local structure so as to make clusters visible in a 2D or 3D visualization.

The below is the visualization result. The visualization follows the analysis. We could seer the cluster 3 is little special from other two clusters.

We went through the cluster for female users above. We also used the same method to cluster male users with 3 clusters.

Below is the clustering results. Cluster 1 for male users tend to have average body type and sometimes take drugs. For cluster 2, those male users have higher income. Most of them never take drugs and have athletic or fit body type. Male users in cluster 3 tend to have lower income compared to users in other two clusters. Most of them have average body type and never take drugs.

```
## [[1]]
##       age            height          income                drinks
##  Min.   :18.00   Min.   :63.00   Min.   :  20000   desperately:  5
##  1st Qu.:23.00   1st Qu.:69.00   1st Qu.:  20000   not at all : 12
##  Median :27.00   Median :71.00   Median :  30000   often      : 71
##  Mean   :29.36   Mean   :70.91   Mean   : 106853   rarely     : 35
##  3rd Qu.:32.00   3rd Qu.:72.00   3rd Qu.:  70000   socially   :204
##  Max.   :66.00   Max.   :80.00   Max.   :1000000   very often : 13
##
##                   diet           drugs              body_type
##  mostly anything    :161   never    :  0   average       :115
##  strictly anything  : 72   often    : 26   fit           : 69
##  anything           : 31   sometimes:314   athletic      : 37
##  mostly vegetarian  : 29                   a little extra: 34
##  mostly other       : 12                   thin          : 31
##  strictly vegetarian: 12                   skinny        : 25
##  (Other)            : 23                   (Other)       : 29
##     cluster
##  Min.   :1
##  1st Qu.:1
##  Median :1
##  Mean   :1
##  3rd Qu.:1
##  Max.   :1
##
##
## [[2]]
##       age            height          income                drinks
##  Min.   :18.00   Min.   :47.00   Min.   :  20000   desperately:  4
##  1st Qu.:26.00   1st Qu.:69.00   1st Qu.:  60000   not at all : 48
##  Median :32.00   Median :71.00   Median :  80000   often      : 32
##  Mean   :34.55   Mean   :70.68   Mean   : 146767   rarely     : 66
##  3rd Qu.:42.00   3rd Qu.:73.00   3rd Qu.: 100000   socially   :370
##  Max.   :68.00   Max.   :95.00   Max.   :1000000   very often : 12
##
##                 diet           drugs              body_type      cluster
##  mostly anything  :285   never    :488   athletic      :326   Min.   :2
##  strictly anything: 76   often    : 14   fit           :117   1st Qu.:2
##  anything         : 62   sometimes: 30   a little extra: 29   Median :2
##  mostly vegetarian: 45                  thin          : 22   Mean   :2
##  mostly other     : 21                  jacked        : 12   3rd Qu.:2
##  strictly other   : 11                  overweight    : 12   Max.   :2
##  (Other)          : 32                  (Other)       : 14
##
## [[3]]
##       age            height          income                drinks
##  Min.   :18.00   Min.   :60.00   Min.   :  20000   desperately:  3
##  1st Qu.:25.00   1st Qu.:68.00   1st Qu.:  20000   not at all : 64
##  Median :29.00   Median :70.00   Median :  40000   often      : 44
##  Mean   :32.22   Mean   :70.22   Mean   :  66804   rarely     : 81
##  3rd Qu.:37.75   3rd Qu.:72.00   3rd Qu.:  60000   socially   :314
##  Max.   :68.00   Max.   :83.00   Max.   :1000000   very often :  4
##
```

```
##                   diet           drugs              body_type
##  mostly anything    :297   never    :505   average       :296
##  anything           : 64   often    :  5   fit           : 90
##  strictly anything  : 53   sometimes:  0   a little extra: 57
##  mostly vegetarian  : 38                   thin          : 24
##  mostly other       : 21                   skinny        : 23
##  strictly vegetarian: 10                   overweight    :  9
##  (Other)            : 27                   (Other)       : 11
##     cluster
##  Min.   :3
##  1st Qu.:3
##  Median :3
##  Mean   :3
##  3rd Qu.:3
##  Max.   :3
##
```

The analysis indicates that cluster 1 and 3 might have overlaps due to similar body type and other features. Similarly, cluster 2 and cluster 3 may have overlaps, so we used visualization to confirm whether the analysis is right or not.

And the below figure clearly shows that three clusters are seperated from each other, there are overlaps between different clusters, which follows the analysis.

Now, we decided to analyze, in parallel, the Speed Dating data set, to understand who are those that get matches more frequently. What are their characteristics? What do they have in common?

Due to several differences in both dataset, crossing the data would not be possible, so we only focused on providing as much information as possible for the user to understand both the user clusters, their characteristics and what he/she should have in mind if he wants to increase his matches (on online dating).

Read data set:

The full raw data set has the following dimensions:

```
## [1] 8378  195
```

Below we will filter the data set, selecting only the variables we considered important for this study. Also, we break it into 2 data sets, one for analysing the individual who participated on the speed dating event, and the other data set for analyzing paired matches.

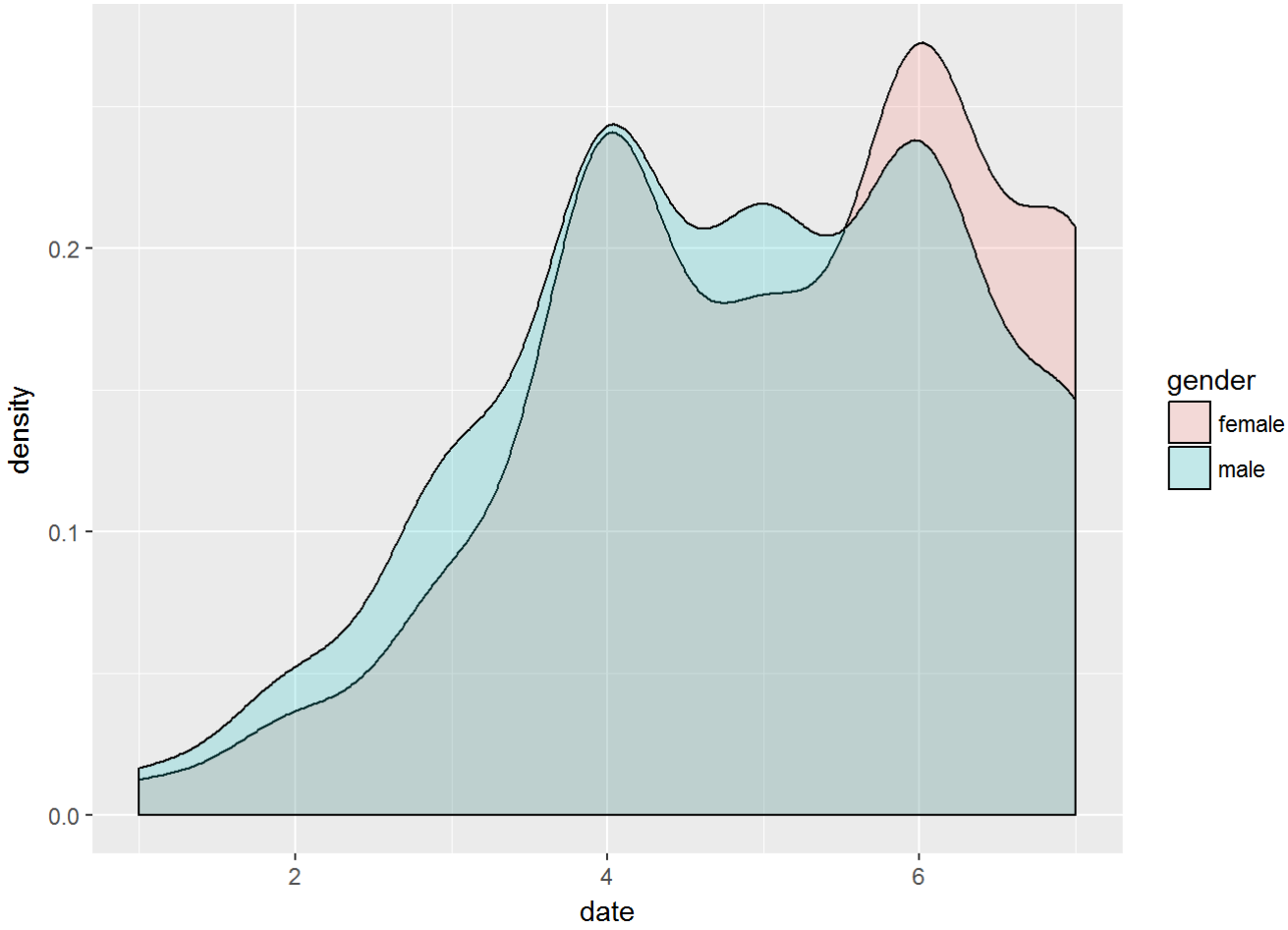
To have matching information on the user level, we decided to calculate matching rates for the individuals. The matching rate is simply the percentage of how many matches did the person have for the total indivuals he/she met that day.

Below we calculate the matching rates and merge it to the individual data set.

Preliminary Analysis:

One initial question that might arise is: who goes more often on dates, men or women? Let's see.

```
## Warning: Removed 8 rows containing non-finite values (stat_density).
```

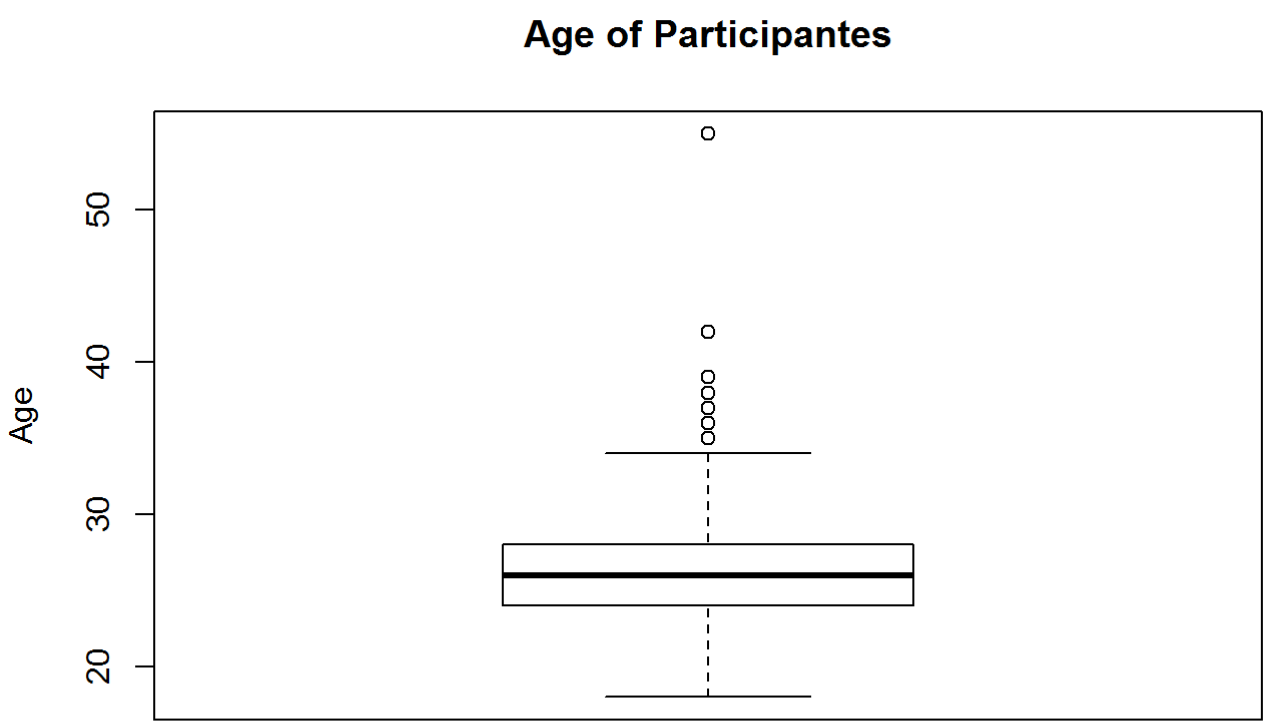



```
## [1] 5.133829
```

```
## [1] 4.835766
```

As we can see, apparently women go more often on dates than men.

Now, let's see how the age of the participants is distributed on this data set.

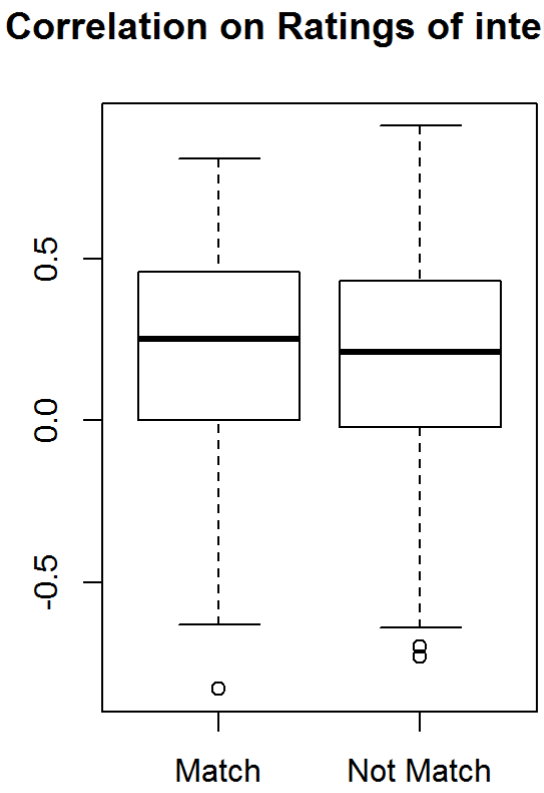


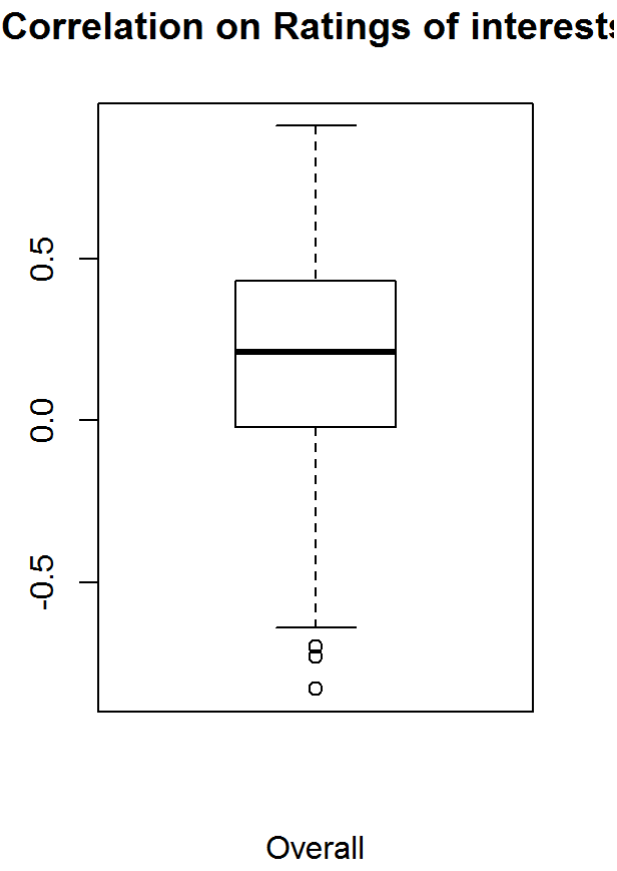

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   18.00   24.00   26.00   26.36   28.00   55.00       8
```

We see that most participants were on their mid-20's for this dataset. Let's check the overall matching rate.

```
## [1] "The overall matching rate in the speed dating event was: 16.47 %"
```

Let's see the behavior of the interest correlations.






```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -0.830  -0.020   0.210   0.196   0.430   0.910     158
```

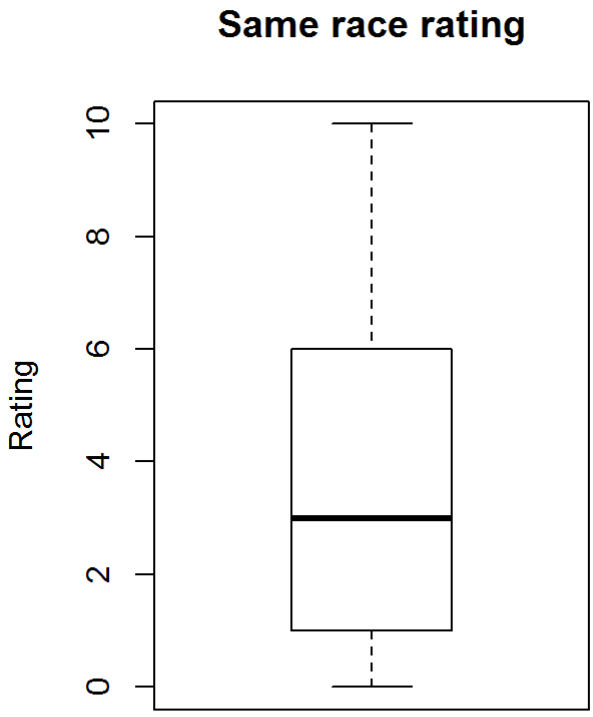
Looking at the interest ratings correlations above, we can see that having a high correlation did not seem to be relevant for having matches. It was also interesting to have such high correlation values (mean aprox 0.2), as someone could expect these to be overall around 0.

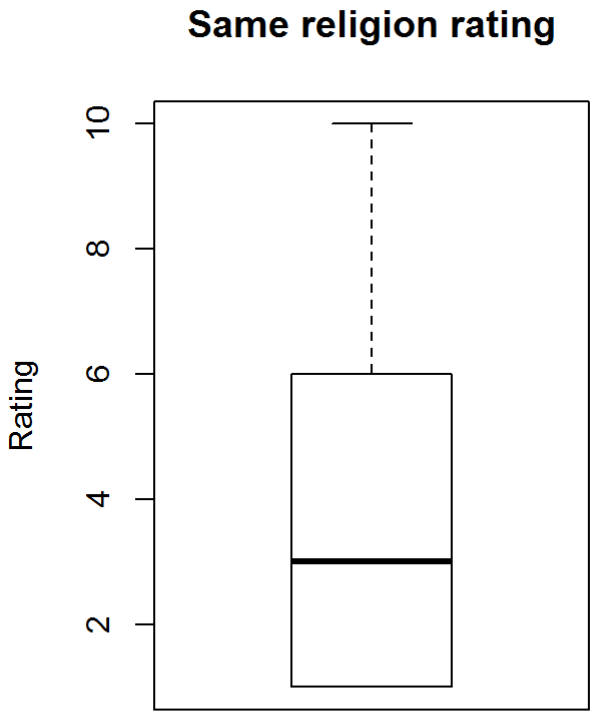
We have also the data on whether people had the same race (yes or no) and their ratings on same race preference and same religion preference, being 1 not important for the other person to have the same race and 10 very important that the other person is of the same race.

```
##            Dif Race Same Race
##
## Match           814       566
## Not Match      4248      2750
```

If we look at the matching rates, having the same race or not does seem to have little impact on the rate of matches. We get 17.06% matching rate if individuals have the same race and 16.08%% matching rate if they have different races.

Now looking at the importance rates for partner having the same religion and same race:





0 = Not Important & 10 = Very Important

0 = Not Important & 10 = Very Important

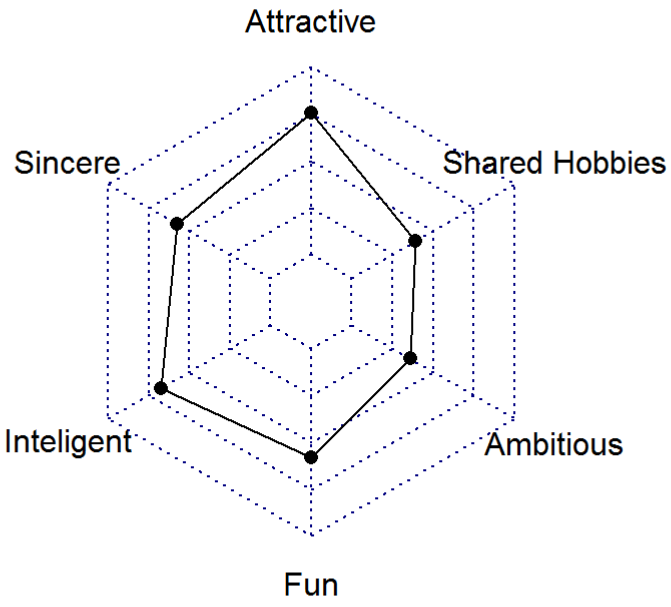
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.000   3.000   3.733   6.000  10.000       7
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   3.000   3.583   6.000  10.000       7
```

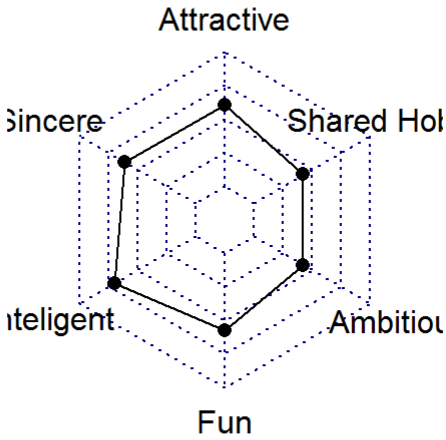
We can see that for most people, having the same race or religion is a bit important.

To continue our investigation, we are also analyzing which attributes are the participating individuals looking for:

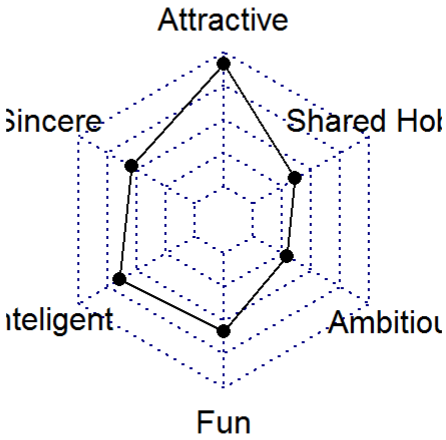
What people are looking for



What women are looking for

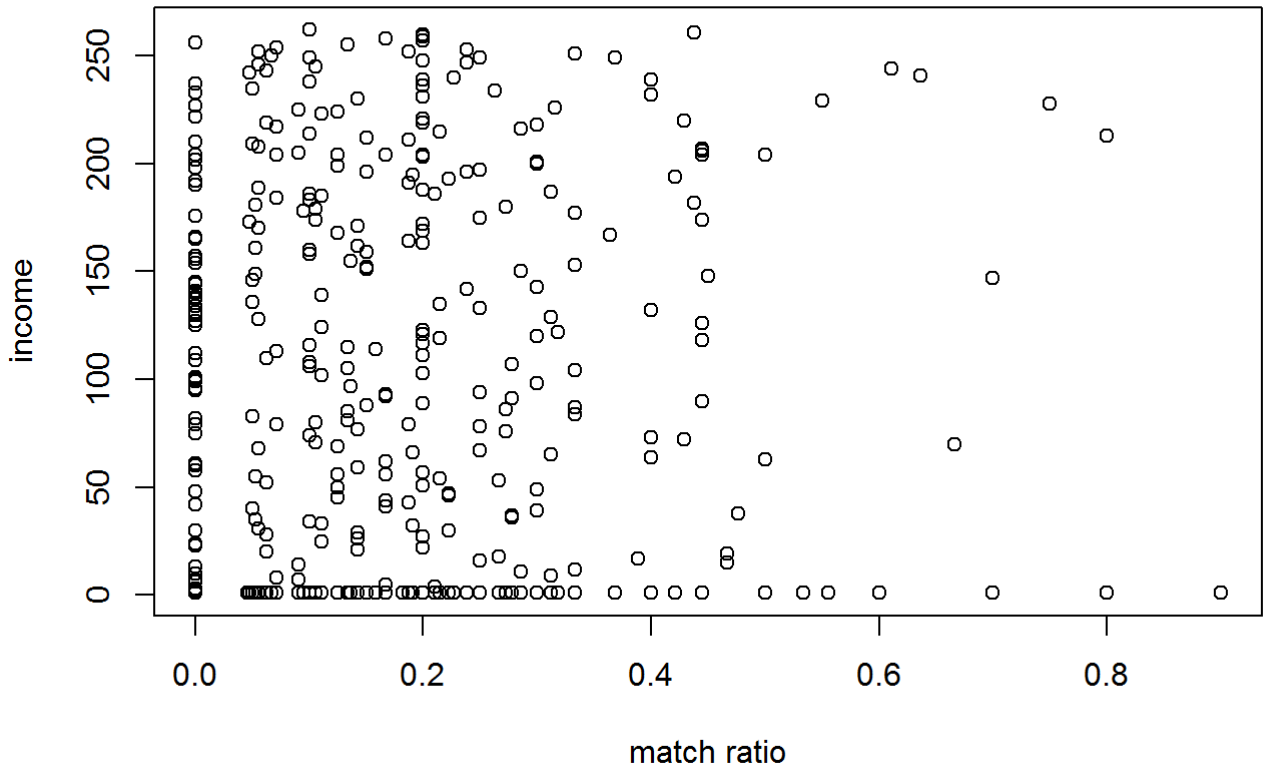


What men are looking for



As we can see, men seem to care more if women are attractive or not, when it comes to looking for a partner. Women tend to value more intelligence in their partners.

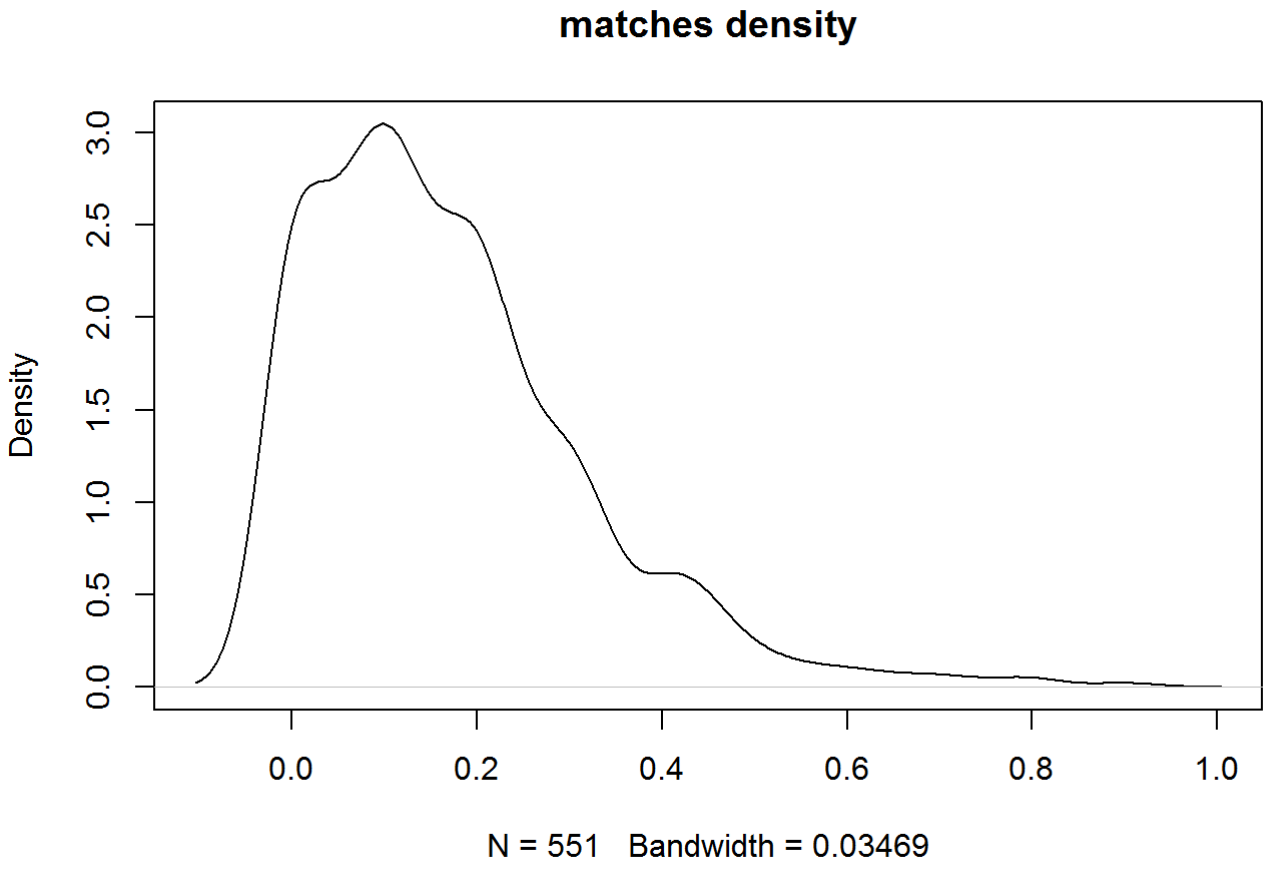
Now, let's analyse how relevant is the income level of an individual to the amount of matches he/she receives. Is there any correlation between a person's income and his percentage of matches?




```
## [1] 0.06785193
```

Apparently, it doesn't appear to exist a relation between these two factors.

Additionally, we can observe the matching rates distribution for the individuals below:




```
##        10%        25%        50%        75%        90%
## 0.00000000 0.05555556 0.13636364 0.23809524 0.36842105
```

```
## [1] 0.1689276
```

```
## [1] 0.1497309
```

Now, one of our goals is to understand which factors drive the matching rates for these individuals. For that, we decided to fit some linear models, not necessarily to predict the matching rates, but to narrow down which variables should we be looking at if we want to increase our matching rates:

```
##
## Call:
## lm(formula = match ~ ., data = speed.dating.individual[, -c(1,
##     7)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26974 -0.10105 -0.02791  0.07508  0.59124
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.958e-01  2.620e-01   1.511  0.13142
## gender      -2.717e-03  1.510e-02  -0.180  0.85726
## race        -3.807e-03  5.142e-03  -0.741  0.45933
## imprace     -6.397e-03  2.528e-03  -2.531  0.01167 *
## imprelig    -2.943e-04  2.551e-03  -0.115  0.90818
## age         -3.727e-03  1.743e-03  -2.139  0.03294 *
## goal        -3.217e-04  4.405e-03  -0.073  0.94182
## date        -1.530e-02  4.674e-03  -3.273  0.00114 **
## go_out      -1.116e-02  6.352e-03  -1.757  0.07944 .
## career_c    -2.925e-03  1.909e-03  -1.532  0.12605
## sports       1.023e-03  2.519e-03   0.406  0.68493
## dining       1.364e-03  4.011e-03   0.340  0.73391
## museums     -9.558e-03  6.444e-03  -1.483  0.13862
## art          9.382e-03  5.622e-03   1.669  0.09579 .
## gaming       7.907e-04  2.605e-03   0.303  0.76166
## clubbing     7.427e-03  2.673e-03   2.779  0.00566 **
## reading      4.581e-03  3.387e-03   1.352  0.17683
## music        1.828e-03  3.707e-03   0.493  0.62210
## yoga         3.577e-03  2.446e-03   1.462  0.14428
## attr1_1     -1.148e-03  2.523e-03  -0.455  0.64943
## sinc1_1     -1.910e-03  2.616e-03  -0.730  0.46554
## intel1_1     3.231e-05  2.668e-03   0.012  0.99035
## fun1_1       6.298e-04  2.681e-03   0.235  0.81435
## amb1_1      -1.321e-03  2.428e-03  -0.544  0.58661
## shar1_1     -3.186e-03  2.661e-03  -1.197  0.23177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.142 on 509 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.1456, Adjusted R-squared:  0.1053
## F-statistic: 3.615 on 24 and 509 DF,  p-value: 3.121e-08
```

```
## [1] 10.2632
```

From our fit, we can observe the main drivers of the matching rates are: Importance of Race, Age, How often the individual goes on dates, How often the individual "go out", interest in arts and interest in clubbing.

Let's fit another model removing the variables that seem not to be important:

```
##
## Call:
## lm(formula = match ~ imprace + age + date + go_out + art + clubbing,
##     data = speed.dating.individual[, -c(1, 7)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26580 -0.09998 -0.02715  0.07622  0.62313
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.341465   0.055550   6.147 1.55e-09 ***
## imprace     -0.007209   0.002184  -3.301 0.001028 **
## age         -0.004410   0.001670  -2.640 0.008522 **
## date        -0.015221   0.004495  -3.386 0.000762 ***
## go_out      -0.015926   0.005962  -2.671 0.007792 **
## art          0.005326   0.002739   1.945 0.052334 .
## clubbing     0.007845   0.002509   3.126 0.001867 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1423 on 534 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.1111, Adjusted R-squared:  0.1011
## F-statistic: 11.12 on 6 and 534 DF,  p-value: 1.039e-11
```

```
## [1] 10.80748
```

The variable “interest in arts” seems to be the less important one, and the most important drivers seem to be whether the person often goes on dates and whether the person likes to “go out”.

Even though the linear model may not be the best fit for our data, we can at least see what are the characteristics of people who get most matches.

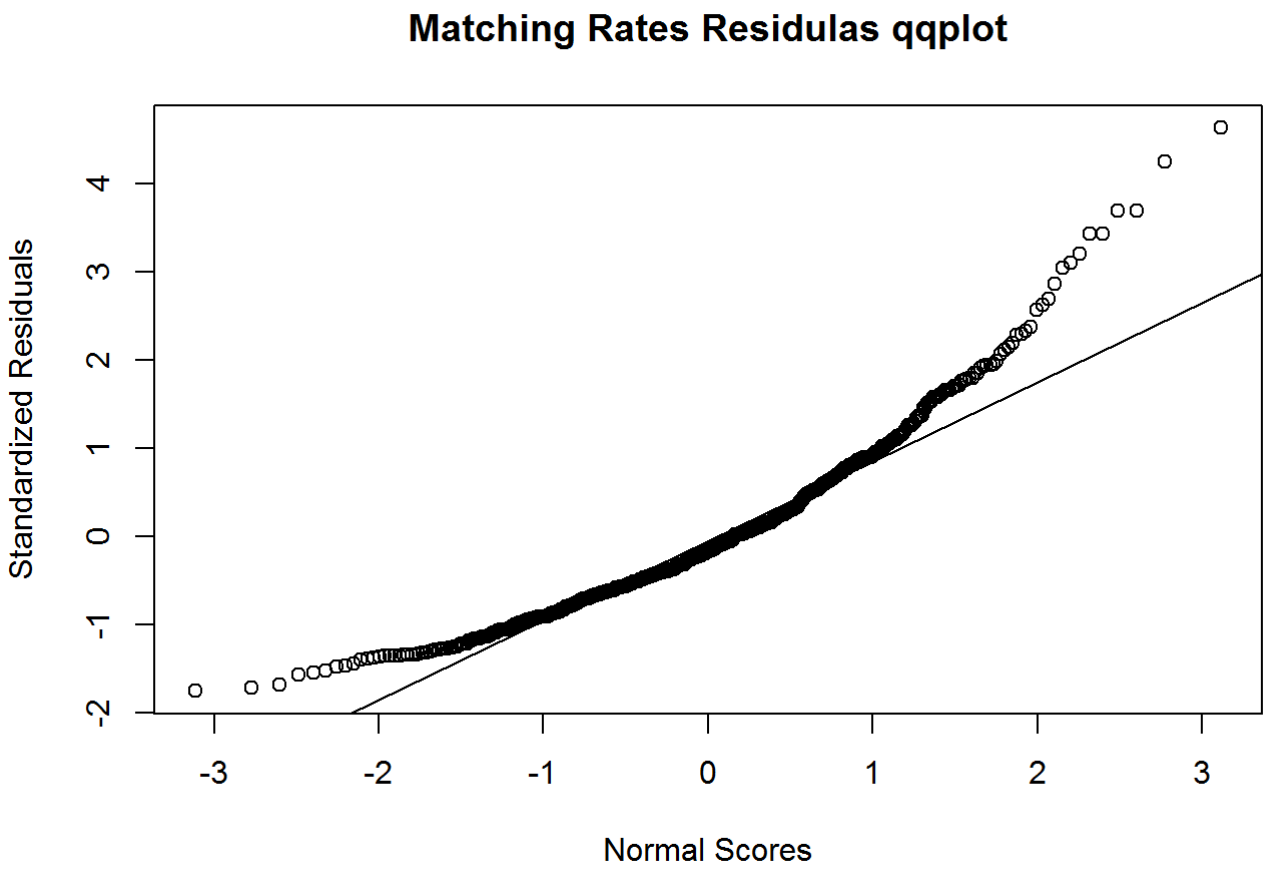
Out of curiosity, we decided to fit a third model without the intercept:

```
##
## Call:
## lm(formula = match ~ 0 + age + go_out + art + clubbing, data = speed.dating.individua
l[,
##     -c(1, 7)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25613 -0.09605 -0.02433  0.08193  0.67915
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## age       0.0028404  0.0008966   3.168 0.001621 **
## go_out   -0.0189068  0.0056935  -3.321 0.000959 ***
## art       0.0081926  0.0027496   2.980 0.003018 **
## clubbing  0.0132510  0.0024080   5.503 5.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1472 on 538 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.5775, Adjusted R-squared:  0.5744
## F-statistic: 183.8 on 4 and 538 DF,  p-value: < 2.2e-16
```

```
## [1] 11.65976
```

After several trials and errors, we decided that the best variables for the third model are: Age, “go out”, art and clubbing.

However, as mentioned, the linear model is not the best fit for our matching rates. We can see that by taking a quick look at the residuals:



Interesting: the 'date' variable was important on the second model, but not on the third... are they highly correlated? Or were people lying?

```
## [1] 0.35
```

They are fairly correlated, but it could be that people lie about that...

From the output above, we can see our model was no capable of identifying who will have a match.

Conclusion:

After testing these linear models, our conclusion was that: if you want to increase your matching rates, you better start to like going out more often, clubbing and arts. Also, have more practice with dates. Which makes total sense, since people with these characteristics are generally more extrovert, less shy and tend to feel more confortable among people they don't know very well.

Obs: We have tried boxcox transformations and taking the log out of the match rates, however we had similar results, and the residuals still did not present to be great fits. Some kind of generalized linear model could be a next step.