Assignment-1

Yizhe Qu

January 19, 2019

Q1

a.

```
library(gclus)
```

```
## Loading required package: cluster
```

```
data(bank)
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.5.2
```

```
ldamod <- lda(Status ~ Diagonal + Top + Length + Left +
                Right, data=bank, CV=TRUE)
table(bank$Status, ldamod$class)
```

```
##
##       0   1
##   0  98   2
##   1   0 100
```

As we can see, we misclassify two genuine bank notes as a counterfeit, giving us a misclassification rate of 0.01. Let's look at some of our other metrics.

```
library(MLmetrics)
```

```
## Warning: package 'MLmetrics' was built under R version 3.5.2
```

```
##
## Attaching package: 'MLmetrics'
```

```
## The following object is masked from 'package:base':
##
##     Recall
```

```
Sensitivity(bank$Status, ldamod$class)
```

```
## [1] 0.98
```

```
Recall(bank$Status, ldamod$class) #same thing!
```

```
## [1] 0.98
```

```
Precision(bank$Status, ldamod$class)
```

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

```
Specificity(bank$Status, ldamod$class)
```

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

```
F1_Score(bank$Status, ldamod$class)
```

```
## [1] 0.989899
```

```
LogLoss(ldamod$posterior[,2], bank$Status)
```

```
## [1] 0.06772524
```

b.

```
qdamod <- qda(Status ~ Diagonal + Top + Length + Left +
                Right, data=bank, CV=TRUE)
table(bank$Status, qdamod$class)
```

```
##
##       0   1
##   0  98   2
##   1   0 100
```

Once more, we misclassify two genuine bank notes as a counterfeit, giving us a misclassification rate of 0.01.

```
Sensitivity(bank$Status, qdamod$class)
```

```
## [1] 0.98
```

```
Recall(bank$Status, qdamod$class) #same thing!
```

```
## [1] 0.98
```

```
Precision(bank$Status, qdamod$class)
```

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

```
Specificity(bank$Status, qdamod$class)
```

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

```
F1_Score(bank$Status, qdamod$class)
```

```
## [1] 0.989899
```

```
LogLoss(qdamod$posterior[,2], bank$Status)
```

```
## [1] 0.09395232
```

Logloss is slightly worse than from LDA.

c.

```
library(class)
knnmod <- knn.cv(bank[,-c(1,5)],  cl=bank$Status, k=3, prob=TRUE)
table(bank$Status, knnmod)
```

```
##    knnmod
##       0   1
##   0  99   1
##   1   0 100
```

This time, we only misclassify one genuine bank note as a counterfeit, giving us a misclassification rate of 0.005.

```
Sensitivity(bank$Status, knnmod)
```

```
## [1] 0.99
```

```
Recall(bank$Status, knnmod)
```

```
## [1] 0.99
```

```
Precision(bank$Status, knnmod)
```

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

```
Specificity(bank$Status, knnmod)
```

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

```
F1_Score(bank$Status, knnmod)
```

```
## [1] 0.9949749
```

Note that calculating LogLoss is a bit of a mess here (just like it was in lab).

```
probs <- attr(knnmod, "prob")
probs[probs==0] <- 1e-15
missedprobs <- 1-probs[bank$Status!=knnmod]
missedprobs[missedprobs==0] <- 1e-15
(sum(-log(probs[bank$Status==knnmod])) + sum(-log(missedprobs)))/200
```

```
## [1] 0.1822416
```

d. This is fairly open ended, as the performance is pretty similar. KNN gets heavily penalized for its one misclassification having happened with probability 1, so by logloss LDA is the winner. That said, if we view the misclassification probabilities associated with LDA.

```
ldamod$posterior[ldamod$class!=bank$Status,]
```

```
##               0         1
## 1  1.151861e-01 0.8848139
## 70 2.665765e-05 0.9999733
```

the worst misclassification at .9999733 is pretty darn close to probability 1. So perhaps the capped logloss is differentiating between 1 and .9999733 a bit too much. For example, if we truncate at 1e-5 instead of 1e-15.

```
probs <- attr(knnmod, "prob")
probs[probs==0] <- 1e-5
missedprobs <- 1-probs[bank$Status!=knnmod]
missedprobs[missedprobs==0] <- 1e-5
(sum(-log(probs[bank$Status==knnmod])) + sum(-log(missedprobs)))/200
```

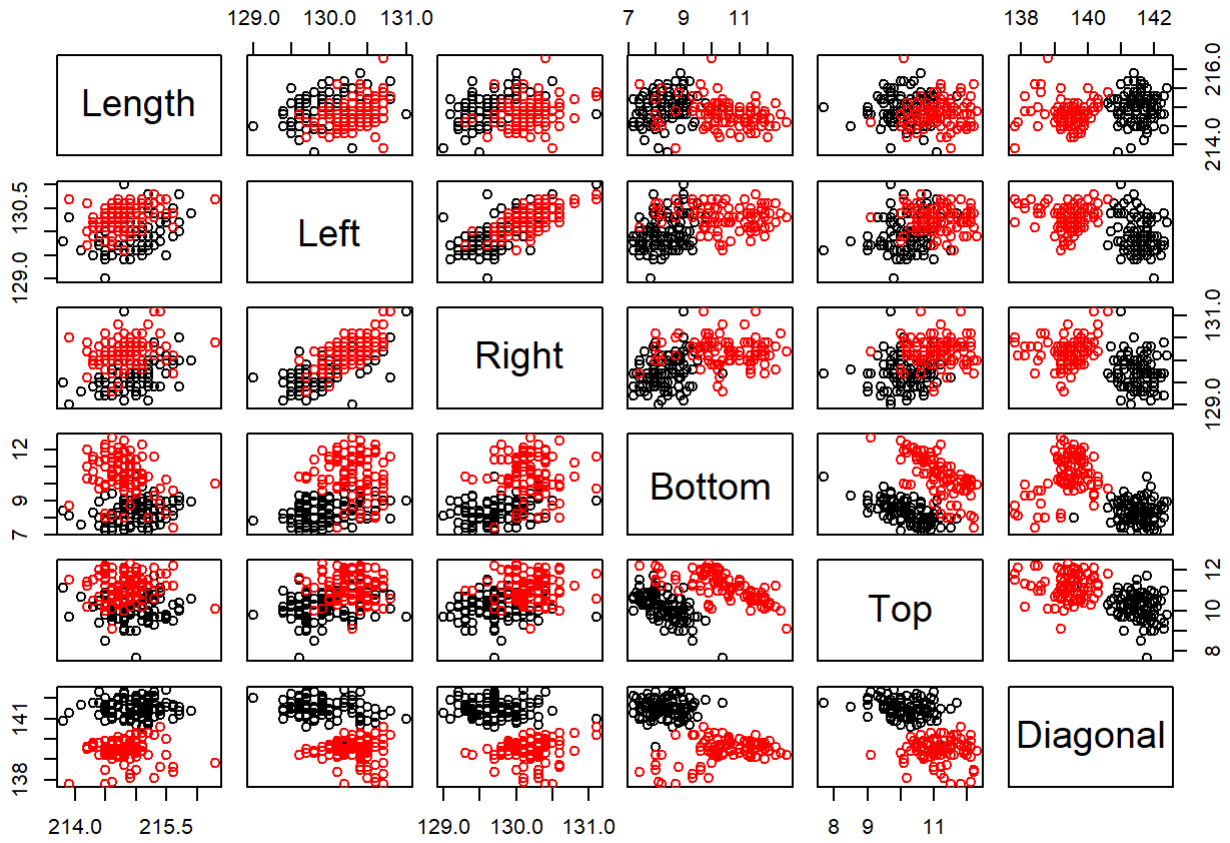
```
## [1] 0.06711234
```

Now KNN would be considered just slightly better than LDA.

With this in mind, I'm inclined to lean towards misclassification rate or F1 score and declare KNN as a better classification model on this data. Note that if we want to factor in inference abilities of the model, this would certainly swing over to a preference for LDA.

e. If we look at the data, along with the results we see in the earlier analyses, it is clear that this data set is easy from a classification standpoint. We have two well-separated groups:

```
plot(bank[,-1], col=bank[,1]+1)
```



If we consider a cost associated with taking each of these measurements, it behooves us to simplify the measuring process. The scatterplot matrix provides a clear way to go about this: the Diagonal measure on its own will likely serve us just as well to differentiate between the genuine and counterfeit notes as the full data set! Any of the systematic approaches for variable selection that we have covered should (hopefully) come to the same conclusion.

Q2

```
library(arules)
```

```
## Warning: package 'arules' was built under R version 3.5.2
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'arules'
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
okc <- read.csv("C:/Users/yizhe/Desktop/MDS/Term4/data_572/data/okcupidprofiles.csv")
summary(okc)
```

```
##       age           body_type                   diet
##  26     : 3724   average :14652                  :25132
##  27     : 3685   fit     :12711   mostly anything  :16585
##  28     : 3583   athletic:11819   anything         : 6183
##  25     : 3531           : 5771   strictly anything: 5113
##  29     : 3295   thin    : 4711   mostly vegetarian: 3444
##  24     : 3242   curvy   : 3924   mostly other     : 1007
##  (Other):39683   (Other) : 7155   (Other)          : 3279
##         drinks                          drugs
##  socially  :41780   never                :37724
##  rarely    : 5957                        :14866
##  often     : 5164   sometimes            : 7732
##            : 3754   often                :  410
##  not at all: 3267    and everything between:    1
##  very often:  471    easygoing even keeled :    1
##  (Other)   :  350   (Other)              :    9
##                              education                ethnicity
##  graduated from college/university:23959   white           :32828
##  graduated from masters program   : 8961                   : 6477
##                                   : 7418   asian           : 6134
##  working on college/university    : 5712   hispanic / latin: 2823
##  working on masters program       : 1683   black           : 2008
##  graduated from two-year college  : 1531   other           : 1706
##  (Other)                          :11479   (Other)         : 8767
##      height          income                                  job
##  70     : 6072   -1     :48440                                 : 8995
##  68     : 5448   20000  : 2952   other                         : 7588
##  67     : 5354   100000 : 1621   student                       : 4882
##  72     : 5315   80000  : 1111   science / tech / engineering  : 4848
##  69     : 5179   30000  : 1048   computer / hardware / software: 4709
##  (Other):33372   40000  : 1004   artistic / musical / writer   : 4438
##  NA's   :    3   (Other): 4567   (Other)                       :25283
##            last_online                          location
##                  :  797   san francisco, california:31064
##  2012-06-29-22-56:   24   oakland, california      : 7214
##  2012-06-30-21-51:   23   berkeley, california     : 4210
##  2012-06-30-22-09:   23   san mateo, california    : 1331
##  2012-06-30-22-56:   23   palo alto, california    : 1064
##  2012-06-30-23-27:   23   alameda, california      :  910
##  (Other)         :59830   (Other)                 :14950
##                                         offspring
##                                              :36359
##  doesn&rsquo;t have kids                     : 7559
##  doesn&rsquo;t have kids, but might want them: 3875
##  doesn&rsquo;t have kids, but wants them     : 3565
##  doesn&rsquo;t want kids                     : 2926
##  has kids                                    : 1883
##  (Other)                                     : 4576
##                                      orientation
##                                              :  799
##  bisexual                                    : 2767
##  english, spanish (poorly), german (poorly):    1
##  gay                                         : 5573
```

```
##  straight                                :51603
##
##
##                         pets
##                           :20719
##  likes dogs and likes cats:14813
##  likes dogs               : 7224
##  likes dogs and has cats  : 4313
##  has dogs                 : 4133
##  has dogs and likes cats  : 2333
##  (Other)                  : 7208
##                                     religion     sex
##                                         :21026    :  800
##  agnosticism                            : 2723   f:24117
##  other                                  : 2691   m:35826
##  agnosticism but not too serious about it: 2636
##  agnosticism and laughing about it       : 2496
##  catholicism but not too serious about it: 2318
##  (Other)                                :26853
##                                       sign                  smokes
##                                         :11856                : 6312
##  gemini and it&rsquo;s fun to think about : 1782   no            :43893
##  scorpio and it&rsquo;s fun to think about: 1772   sometimes     : 3787
##  leo and it&rsquo;s fun to think about    : 1692   trying to quit: 1480
##  libra and it&rsquo;s fun to think about  : 1648   when drinking : 3040
##  taurus and it&rsquo;s fun to think about : 1640   yes           : 2231
##  (Other)                                :40353
##                                    speaks                status
##  english                              :21828                :  800
##  english (fluently)                   : 6627   available    : 1864
##  english (fluently), spanish (poorly)  : 2059   married      :  310
##  english (fluently), spanish (okay)    : 1917   seeing someone: 2064
##  english (fluently), spanish (fluently): 1288   single       :55695
##  english, spanish                     :  859   unknown      :   10
##  (Other)                              :26165
```

Note that there are a lot of issues here with the read-in. For example, one of the answers for orientation appears to be languages. A deeper dive will show issues for several consecutive rows.

```
okc[27939:27944,1]
```

```
## [1] 36
## [2] +smile+on+a+passerby%2c+a%0a%22ladybug%22%2c+a+friend%2c+the+sun+shining+etc%29."">
## [3] there is so much horror in the world
## [4] light must be celebrated (a flower
## [5] ladybug""
## [6] <br />
## 744 Levels:  ...
```

```
okc[28325:28330,1]
```

```
## [1] i am a spontaneously creatively adventuring two spirit."
## [3] 34                                31
## [5] 42                                38
## 744 Levels:  ...
```

There are lots of other rows that are corrupted with additional (html code type) information in here as well. I'm just going to remove them all by removing any age entry that has more than 3 characters

```
okc <- okc[nchar(as.character(okc$age))<=2,]
okc <- droplevels(okc)
summary(okc)
```

```
##       age           body_type                    diet
##  26     : 3724   average :14652                   :24403
##  27     : 3685   fit     :12711   mostly anything  :16585
##  28     : 3583   athletic:11818   anything         : 6183
##  25     : 3531           : 5304   strictly anything: 5113
##  29     : 3295   thin    : 4711   mostly vegetarian: 3444
##  24     : 3242   curvy   : 3924   mostly other     : 1006
##  (Other):38893   (Other) : 6833   (Other)          : 3219
##          drinks            drugs
##             : 2992            :14088
##  desperately:  322   never    :37723
##  not at all : 3267   often    :  410
##  often      : 5164   sometimes: 7732
##  rarely     : 5957
##  socially   :41780
##  very often :  471
##                               education                ethnicity
##  graduated from college/university:23959   white           :32828
##  graduated from masters program   : 8961   asian           : 6134
##                                   : 6636                   : 5690
##  working on college/university    : 5712   hispanic / latin: 2823
##  working on masters program       : 1682   black           : 2008
##  graduated from two-year college  : 1531   other           : 1706
##  (Other)                          :11472   (Other)         : 8764
##      height          income                                  job
##  70     : 6072   -1     :48438                                 : 8209
##  68     : 5448   20000  : 2952   other                         : 7588
##  67     : 5353   100000 : 1621   student                       : 4881
##  72     : 5315   80000  : 1111   science / tech / engineering  : 4848
##  69     : 5179   30000  : 1048   computer / hardware / software: 4709
##  (Other):32583   40000  : 1004   artistic / musical / writer   : 4438
##  NA's   :    3   (Other): 3779   (Other)                       :25280
##            last_online                          location
##  2012-06-29-22-56:   24   san francisco, california:31063
##  2012-06-30-21-51:   23   oakland, california      : 7214
##  2012-06-30-22-09:   23   berkeley, california     : 4210
##  2012-06-30-22-56:   23   san mateo, california    : 1331
##  2012-06-30-23-27:   23   palo alto, california    : 1064
##  2012-06-30-10-15:   22   alameda, california      :  910
##  (Other)         :59815   (Other)                  :14161
##                                          offspring       orientation
##                                               :35571           :   12
##  doesn&rsquo;t have kids                      : 7559   bisexual: 2767
##  doesn&rsquo;t have kids, but might want them: 3875   gay     : 5573
##  doesn&rsquo;t have kids, but wants them     : 3565   straight:51601
##  doesn&rsquo;t want kids                      : 2926
##  has kids                                     : 1883
##  (Other)                                      : 4574
##                          pets
##                            :19930
##  likes dogs and likes cats:14813
##  likes dogs               : 7224
##  likes dogs and has cats  : 4313
```

```
##  has dogs                : 4133
##  has dogs and likes cats  : 2333
##  (Other)                 : 7207
##                                    religion     sex
##                                        :20237    :   12
##  agnosticism                           : 2723   f:24116
##  other                                 : 2691   m:35825
##  agnosticism but not too serious about it: 2636
##  agnosticism and laughing about it       : 2496
##  catholicism but not too serious about it: 2318
##  (Other)                               :26852
##                                       sign                 smokes
##                                         :11067                : 5523
##  gemini and it&rsquo;s fun to think about : 1782   no            :43893
##  scorpio and it&rsquo;s fun to think about: 1772   sometimes     : 3787
##  leo and it&rsquo;s fun to think about    : 1692   trying to quit: 1480
##  libra and it&rsquo;s fun to think about  : 1648   when drinking : 3039
##  taurus and it&rsquo;s fun to think about : 1640   yes           : 2231
##  (Other)                               :40352
##                                    speaks                status
##  english                              :21827                :   12
##  english (fluently)                   : 6627   available    : 1863
##  english (fluently), spanish (poorly)  : 2059   married      :  310
##  english (fluently), spanish (okay)    : 1917   seeing someone: 2064
##  english (fluently), spanish (fluently): 1288   single       :55694
##  english, spanish                     :  859   unknown      :   10
##  (Other)                              :25376
```

Also, there are lots of empty answers that should probably be recorded as NA's for us. For example,

```
names(table(okc$smokes))
```

```
## [1] ""               "no"             "sometimes"      "trying to quit"
## [5] "when drinking"  "yes"
```

Shows one of the answers is just "". Let's replace all of those with NA

```
okc[okc==""] <- NA
summary(okc)
```

```
##       age           body_type                   diet
##  26     : 3724   average :14652   mostly anything  :16585
##  27     : 3685   fit     :12711   anything         : 6183
##  28     : 3583   athletic:11818   strictly anything: 5113
##  25     : 3531   thin    : 4711   mostly vegetarian: 3444
##  29     : 3295   curvy   : 3924   mostly other     : 1006
##  (Other):42126   (Other) : 6833   (Other)          : 3219
##  NA's   :    9   NA's    : 5304   NA's             :24403
##         drinks            drugs
##  socially  :41780            :    0
##  rarely    : 5957   never    :37723
##  often     : 5164   often    :  410
##  not at all: 3267   sometimes: 7732
##  very often:  471   NA's     :14088
##  (Other)   :  322
##  NA's      : 2992
##                              education                ethnicity
##  graduated from college/university:23959   white           :32828
##  graduated from masters program   : 8961   asian           : 6134
##  working on college/university    : 5712   hispanic / latin: 2823
##  working on masters program       : 1682   black           : 2008
##  graduated from two-year college  : 1531   other           : 1706
##  (Other)                          :11472   (Other)         : 8764
##  NA's                             : 6636   NA's            : 5690
##      height          income                                  job
##  70     : 6072   -1     :48438   other                         : 7588
##  68     : 5448   20000  : 2952   student                       : 4881
##  67     : 5353   100000 : 1621   science / tech / engineering  : 4848
##  72     : 5315   80000  : 1111   computer / hardware / software: 4709
##  69     : 5179   30000  : 1048   artistic / musical / writer   : 4438
##  (Other):32571   (Other): 4771   (Other)                       :25280
##  NA's   :   15   NA's   :   12   NA's                          : 8209
##            last_online                          location
##  2012-06-29-22-56:   24   san francisco, california:31063
##  2012-06-30-21-51:   23   oakland, california      : 7214
##  2012-06-30-22-09:   23   berkeley, california     : 4210
##  2012-06-30-22-56:   23   san mateo, california    : 1331
##  2012-06-30-23-27:   23   palo alto, california    : 1064
##  (Other)         :59825   (Other)                  :15059
##  NA's            :   12   NA's                     :   12
##                                        offspring       orientation
##  doesn&rsquo;t have kids                     : 7559           :    0
##  doesn&rsquo;t have kids, but might want them: 3875   bisexual: 2767
##  doesn&rsquo;t have kids, but wants them     : 3565   gay     : 5573
##  doesn&rsquo;t want kids                     : 2926   straight:51601
##  has kids                                    : 1883   NA's    :   12
##  (Other)                                     : 4574
##  NA's                                        :35571
##                          pets
##  likes dogs and likes cats:14813
##  likes dogs               : 7224
##  likes dogs and has cats  : 4313
##  has dogs                 : 4133
```

```
##  has dogs and likes cats  : 2333
##  (Other)                  : 7207
##  NA's                     :19930
##                                    religion       sex
##  agnosticism                           : 2723       :    0
##  other                                 : 2691   f   :24116
##  agnosticism but not too serious about it: 2636   m   :35825
##  agnosticism and laughing about it     : 2496   NA's:   12
##  catholicism but not too serious about it: 2318
##  (Other)                               :26852
##  NA's                                  :20237
##                                       sign                 smokes
##  gemini and it&rsquo;s fun to think about : 1782                :    0
##  scorpio and it&rsquo;s fun to think about: 1772   no           :43893
##  leo and it&rsquo;s fun to think about    : 1692   sometimes    : 3787
##  libra and it&rsquo;s fun to think about  : 1648   trying to quit: 1480
##  taurus and it&rsquo;s fun to think about : 1640   when drinking : 3039
##  (Other)                               :40352   yes          : 2231
##  NA's                                  :11067   NA's         : 5523
##                                  speaks                status
##  english                            :21827                :    0
##  english (fluently)                 : 6627   available    : 1863
##  english (fluently), spanish (poorly)  : 2059   married      :  310
##  english (fluently), spanish (okay)    : 1917   seeing someone: 2064
##  english (fluently), spanish (fluently): 1288   single       :55694
##  (Other)                            :26173   unknown      :   10
##  NA's                               :   62   NA's         :   12
```

Let's get age into a continuous format and then discretize it into age groups

```
okc$age <- as.numeric(as.character(okc$age))
okc$age <- discretize(okc$age, method="interval", breaks=4)
```

Now, there are a number of variables that are probably uninteresting - I'm going to remove last_online and location

```
okc <- okc[,-c(11:12)]
```

Okay, now we run apriori under default specification.

```
aok <- apriori(okc)
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.8    0.1    1 none FALSE            TRUE       5     0.1      1
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 5995
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[8170 item(s), 59953 transaction(s)] done [0.15s].
## sorting and recoding items ... [27 item(s)] done [0.01s].
## creating transaction tree ... done [0.03s].
## checking subsets of size 1 2 3 4 5 6 7 done [0.14s].
## writing ... [2000 rule(s)] done [0.00s].
## creating S4 object  ... done [0.01s].
```

```
summary(aok)
```

```
## set of 2000 rules
##
## rule length distribution (lhs + rhs):sizes
##   1   2   3   4   5   6   7
##   3  62 312 691 643 262  27
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   4.000   4.402   5.000   7.000
##
## summary of quality measures:
##     support         confidence          lift            count
##  Min.   :0.1002   Min.   :0.8001   Min.   :0.9296   Min.   : 6010
##  1st Qu.:0.1164   1st Qu.:0.8562   1st Qu.:1.0124   1st Qu.: 6977
##  Median :0.1455   Median :0.8874   Median :1.0257   Median : 8723
##  Mean   :0.1789   Mean   :0.8909   Mean   :1.0410   Mean   :10726
##  3rd Qu.:0.1998   3rd Qu.:0.9400   3rd Qu.:1.0428   3rd Qu.:11977
##  Max.   :0.9290   Max.   :0.9747   Max.   :1.3637   Max.   :55694
##
## mining info:
##  data ntransactions support confidence
##   okc         59953     0.1        0.8
```

```
inspect(sort(aok, by="support")[1:10])
```

```
##      lhs                         rhs                     support   confidence
## [1]  {}                       => {status=single}        0.9289610 0.9289610
## [2]  {}                       => {orientation=straight} 0.8606909 0.8606909
## [3]  {orientation=straight} => {status=single}        0.8124197 0.9439158
## [4]  {status=single}        => {orientation=straight} 0.8124197 0.8745466
## [5]  {}                       => {income=-1}            0.8079329 0.8079329
## [6]  {income=-1}            => {status=single}        0.7544243 0.9337710
## [7]  {status=single}        => {income=-1}            0.7544243 0.8121162
## [8]  {income=-1}            => {orientation=straight} 0.6927927 0.8574879
## [9]  {orientation=straight} => {income=-1}            0.6927927 0.8049263
## [10] {smokes=no}            => {status=single}        0.6798492 0.9285991
##      lift      count
## [1]  1.0000000 55694
## [2]  1.0000000 51601
## [3]  1.0160984 48707
## [4]  1.0160984 48707
## [5]  1.0000000 48438
## [6]  1.0051778 45230
## [7]  1.0051778 45230
## [8]  0.9962786 41535
## [9]  0.9962786 41535
## [10] 0.9996104 40759
```

```
inspect(sort(aok, by="confidence")[1:10])
```

```
##      lhs                                   rhs              support confidence     lift cou
nt
## [1]  {offspring=doesn&rsquo;t have kids,

##       orientation=straight}             => {status=single} 0.1136057  0.9746709 1.049205  68
11
## [2]  {diet=mostly anything,

##       drugs=never,

##       income=-1,

##       orientation=straight,

##       smokes=no}                        => {status=single} 0.1069504  0.9737282 1.048191  64
12
## [3]  {drinks=socially,

##       drugs=never,

##       income=-1,

##       orientation=straight,

##       speaks=english}                   => {status=single} 0.1143729  0.9735908 1.048043  68
57
## [4]  {diet=mostly anything,

##       drugs=never,

##       income=-1,

##       orientation=straight}             => {status=single} 0.1234300  0.9729161 1.047316  74
00
## [5]  {diet=mostly anything,

##       drinks=socially,

##       drugs=never,

##       orientation=straight}             => {status=single} 0.1205611  0.9722895 1.046642  72
28
## [6]  {diet=mostly anything,

##       drinks=socially,

##       drugs=never,

##       orientation=straight,

##       smokes=no}                        => {status=single} 0.1044652  0.9719119 1.046235  62
63
## [7]  {drinks=socially,
```

```
##       drugs=never,

##       orientation=straight,

##       speaks=english}                    => {status=single} 0.1386419  0.9713685 1.045650  83
12
## [8]  {body_type=athletic,

##       drugs=never,

##       orientation=straight}              => {status=single} 0.1146231  0.9711702 1.045437  68
72
## [9]  {body_type=athletic,

##       drinks=socially,

##       income=-1,

##       orientation=straight}              => {status=single} 0.1002452  0.9710777 1.045337  60
10
## [10] {diet=mostly anything,

##       drinks=socially,

##       drugs=never,

##       income=-1}                         => {status=single} 0.1031475  0.9709531 1.045203  61
84
```

```
inspect(sort(aok, by="lift")[1:10])
```

```
##      lhs                                          rhs           support confidence     li
ft count
## [1]  {body_type=athletic,

##       orientation=straight}                    => {sex=m}     0.1445466  0.8148566 1.3636
59  8666
## [2]  {body_type=athletic,

##       orientation=straight,

##       status=single}                           => {sex=m}     0.1391090  0.8128655 1.3603
27  8340
## [3]  {body_type=athletic,

##       drinks=socially,

##       orientation=straight}                    => {sex=m}     0.1062499  0.8081705 1.3524
70  6370
## [4]  {body_type=athletic,

##       drinks=socially,

##       orientation=straight,

##       status=single}                           => {sex=m}     0.1025136  0.8063500 1.3494
24  6146
## [5]  {body_type=athletic,

##       status=single}                           => {sex=m}     0.1519690  0.8048587 1.3469
28  9111
## [6]  {body_type=athletic}                      => {sex=m}     0.1585909  0.8045355 1.3463
87  9508
## [7]  {body_type=athletic,

##       drinks=socially,

##       status=single}                           => {sex=m}     0.1116208  0.8011493 1.3407
20  6692
## [8]  {body_type=athletic,

##       drinks=socially}                         => {sex=m}     0.1162244  0.8009195 1.3403
36  6968
## [9]  {drinks=socially,

##       drugs=never,

##       education=graduated from college/university,

##       income=-1,

##       orientation=straight}                    => {smokes=no} 0.1182093  0.8885406 1.2136
49  7087
## [10] {drugs=never,
```

```
##       education=graduated from college/university,

##       ethnicity=white,

##       orientation=straight}                       => {smokes=no} 0.1083015  0.8879923 1.2129
00  6493
```

With more cleaning, organization, and tuning, we might be able to remove some of the more uninteresting results. But I think the higher lift associations do have some level of interestingness about them. The top 8 rules according to lift all result in an association with male profiles for the rhs, and on the lhs they all show body_type = athletic - this begins to suggest that males might be more likely to describe their bodies as “athletic” than females do. The last two top lift rules show that folks that say they never do drugs, have higher education, and are straight, are more likely than expected to also not smoke.