```
#################
# Association Rules for Yelp
# Author: Ravi Makhija
# Version 1
#
# Description:
# We explore the Yelp dataset using association rule mining.
#
# File Dependencies:
#   'data/yelp_data.Rdata'
#
# How to run:
#    Source this script (no need to set wd beforehand if directory structure is
#    maintained as downloaded).
#
# References
#   1) Set working directory to the file path of a script:
#      http://stackoverflow.com/questions/13672720/r-command-for-setting-working-dire
#   2) Tutorial on association rules in R:
#      http://www.rdatamining.com/examples/association-rules
#   3) Renaming levels of a factor:
#      http://www.cookbook-r.com/Manipulating_data/Renaming_levels_of_a_factor/
#   4) Installing package from a source file:
#      https://cran.r-project.org/web/packages/arules/index.html

require("arules")   # version 2.2 is needed, which required installing from source

## Loading required package:  arules
## Warning:  package 'arules' was built under R version 3.2.2
## Loading required package:  Matrix
##
## Attaching package:  'arules'
##
## The following objects are masked from 'package:base':
##
##    %in%, abbreviate, write

require("arulesViz")

## Loading required package:  arulesViz
```

```
## Warning:  package 'arulesViz' was built under R version 3.1.3
## Loading required package:  grid
##
## Attaching package:  'arulesViz'
##
## The following object is masked from 'package:arules':
##
##    abbreviate
##
## The following object is masked from 'package:base':
##
##    abbreviate
```

```
require("Hmisc")
```

```
## Loading required package:  Hmisc
## Warning:  package 'Hmisc' was built under R version 3.1.3
## Loading required package:  lattice
## Loading required package:  survival
## Loading required package:  splines
## Loading required package:  Formula
## Warning:  package 'Formula' was built under R version 3.1.3
## Loading required package:  ggplot2
##
## Attaching package:  'Hmisc'
##
## The following objects are masked from 'package:base':
##
##    format.pval, round.POSIXt, trunc.POSIXt, units
```

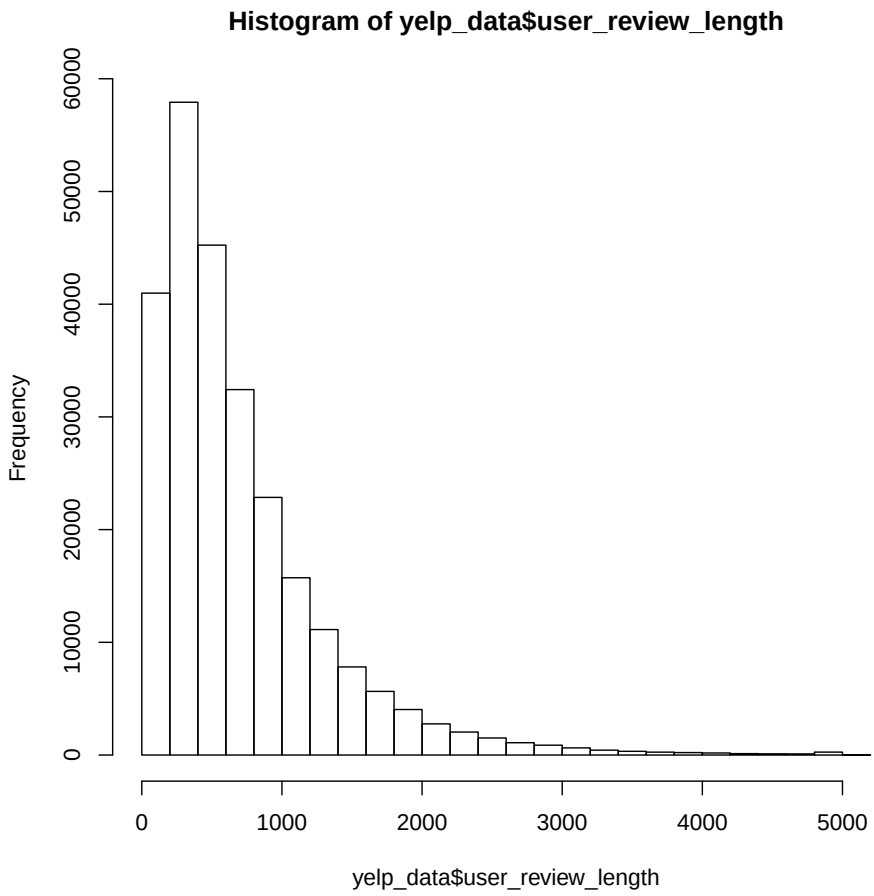
```
require("data.table")
```

```
## Loading required package:  data.table
## Warning:  package 'data.table' was built under R version 3.1.3
```

```
require("plyr")
```

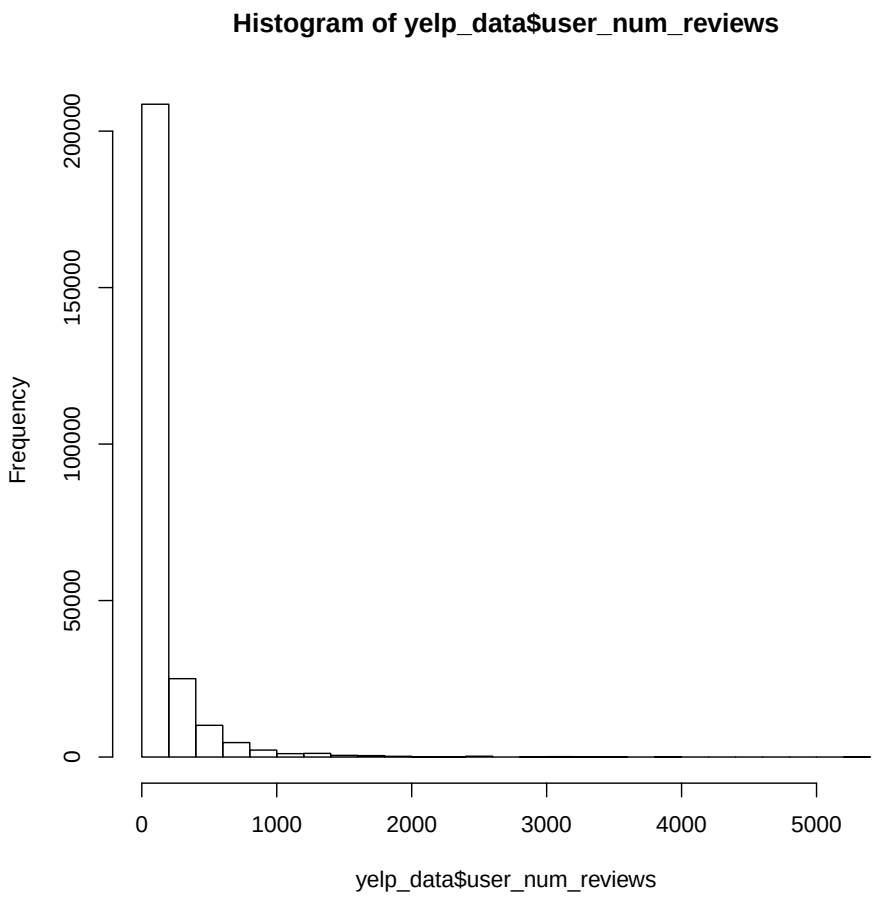
```
## Loading required package:  plyr
##
## Attaching package:  'plyr'
##
## The following objects are masked from 'package:Hmisc':
```

```
##
##    is.discrete, summarize

#################
# Load in data
#################

load("tripadvisor_data.Rdata")
load("yelp_data.Rdata")

#################
# Prep data for association rules
#################

# Create a new data.frame for the data set we want to use association rules on.
# We want to create categorical variables for this purpose.

yelp_data_categorical <- data.frame(user_is_local = as.factor(yelp_data$user_is_local

# Now, we bin some continuous variables and add to this new data set.

####################
# user_review_length

# Explore data
summary(yelp_data$user_review_length)



hist(yelp_data$user_review_length)
```




```
# Bin and add to new data set:
# small: [0 to 300)
# medium: [300 to 1000)
# large: [1000 and up)
yelp_data_categorical$user_review_length <- cut2(x=yelp_data$user_review_length, cut=

####################
# user_rating

# Explore data
table(yelp_data$user_rating)

##
```

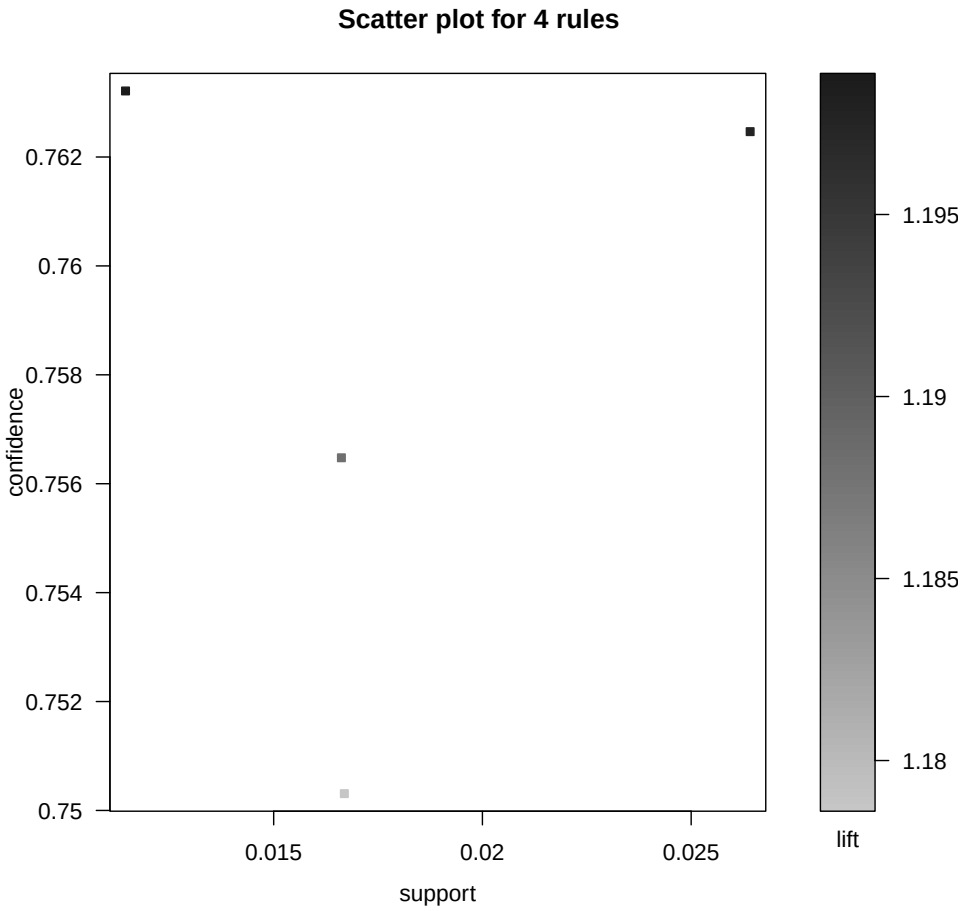
```
##     1     2     3     4     5
## 14612 24792 45103 93302 76917

# Bin and add to new data set:
# For the time being, we keep all five categories, since we have plenty of
# observations.
yelp_data_categorical$user_rating <- as.factor(yelp_data$user_rating)

####################
# user_num_reviews

# Explore data
summary(yelp_data$user_num_reviews)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    12.0    37.0   129.9   134.0  5267.0

hist(yelp_data$user_num_reviews)
```




```
# Bin and add to new data set:
# low: [1 to 12)
# medium: [12 to 134)
# high: [134 and up)
yelp_data_categorical$user_num_reviews <- cut2(x=yelp_data$user_num_reviews,
                                               cut=c(1, 12, 134))

####################
# user_review_date

# Explore data
table(substring(yelp_data$user_review_date, first=1, last=4))
```

```
##
##  2005  2006  2007  2008  2009  2010  2011  2012  2013  2014  2015
##    59  1136  3526  8105 14003 20076 28230 30920 40039 54478 54154

# Bin and add to new data set:
# old: [2005, 2008]
# less_recent: [2009, 2012]
# recent: [2013, 2015]
yelp_data_categorical$user_review_time_period <- cut2(x=as.numeric(substring(yelp_dat

###################
# Check out the new data set
head(yelp_data_categorical)

##   user_is_local user_review_length user_rating user_num_reviews
## 1          TRUE        [1000,5060]           5      [  12, 134)
## 2          TRUE        [1000,5060]           4      [ 134,5267]
## 3          TRUE        [ 300,1000)           4      [ 134,5267]
## 4         FALSE        [1000,5060]           4      [  12, 134)
## 5         FALSE        [ 300,1000)           5      [ 134,5267]
## 6          TRUE        [1000,5060]           5      [ 134,5267]
##   user_review_time_period
## 1             [2013,2015]
## 2             [2013,2015]
## 3             [2013,2015]
## 4             [2013,2015]
## 5             [2013,2015]
## 6             [2013,2015]

#################
# Start association rules mining for Yelp
#################

attach(yelp_data_categorical)

# Since a central question we are asking is whether or not local or non_local
# ratings are higher, we start association rule mining with the binary
# user_is_local on the right, to see if we can find any implications. We
# adjust the minimum support and confidence levels to obtain the most
# meaningful rule set.
```

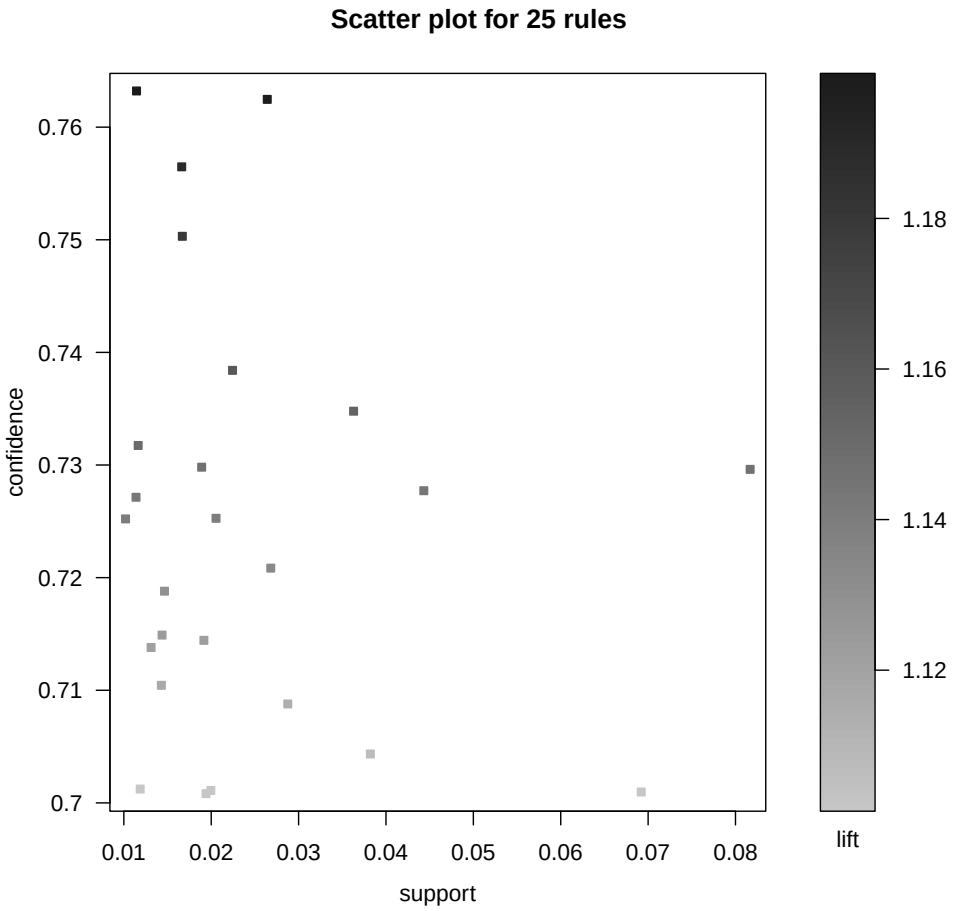
```r
# A first look shows that the four rules returned all suggest that longer
# reviews may be associated with local reviewers. Rule 2 and 3 also
# may inadvertently be suggesting that user ratings of 2 and 3 are associated
# with longer reviews --- something to keep in mind and consider exploring
# further.

yelp_rules_1 <- apriori(yelp_data_categorical,
                        parameter = list(minlen=1, supp=.01, conf=0.75),
                        appearance = list(rhs=c("user_is_local=FALSE", "user_is_local
                        control = list(verbose=F))

## Warning in match(x, table, nomatch = 0):  bytecode version mismatch;
using eval

inspect(yelp_rules_1)

##   lhs                                        rhs                     support confide
## 1 {user_review_length=[1000,5060],
##    user_num_reviews=[   1,  12)}        => {user_is_local=TRUE} 0.02641269  0.7624
## 2 {user_review_length=[1000,5060],
##    user_rating=2,
##    user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01145152  0.7632
## 3 {user_review_length=[1000,5060],
##    user_rating=3,
##    user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01669245  0.7503
## 4 {user_review_length=[1000,5060],
##    user_num_reviews=[   1,  12),
##    user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01662178  0.7564

plot(yelp_rules_1)
```




```
# Bringing the confidence level down, we get many more rules. We can see that
# among the top 17 rules, all rules have ratings of at most 3. E.g. the lower
# half of ratings imply local reviews with the highest confidence. This is
# in accordance with the t-tests conducted, which showed local reviewer mean
# rating is less than that of non-local reviewers.

yelp_rules_2 <- apriori(yelp_data_categorical,
                        parameter = list(minlen=1, supp=.01, conf=0.70),
                        appearance = list(rhs=c("user_is_local=FALSE", "user_is_local
                        control = list(verbose=F))
inspect(yelp_rules_2)

##    lhs                                    rhs                    support confid
```

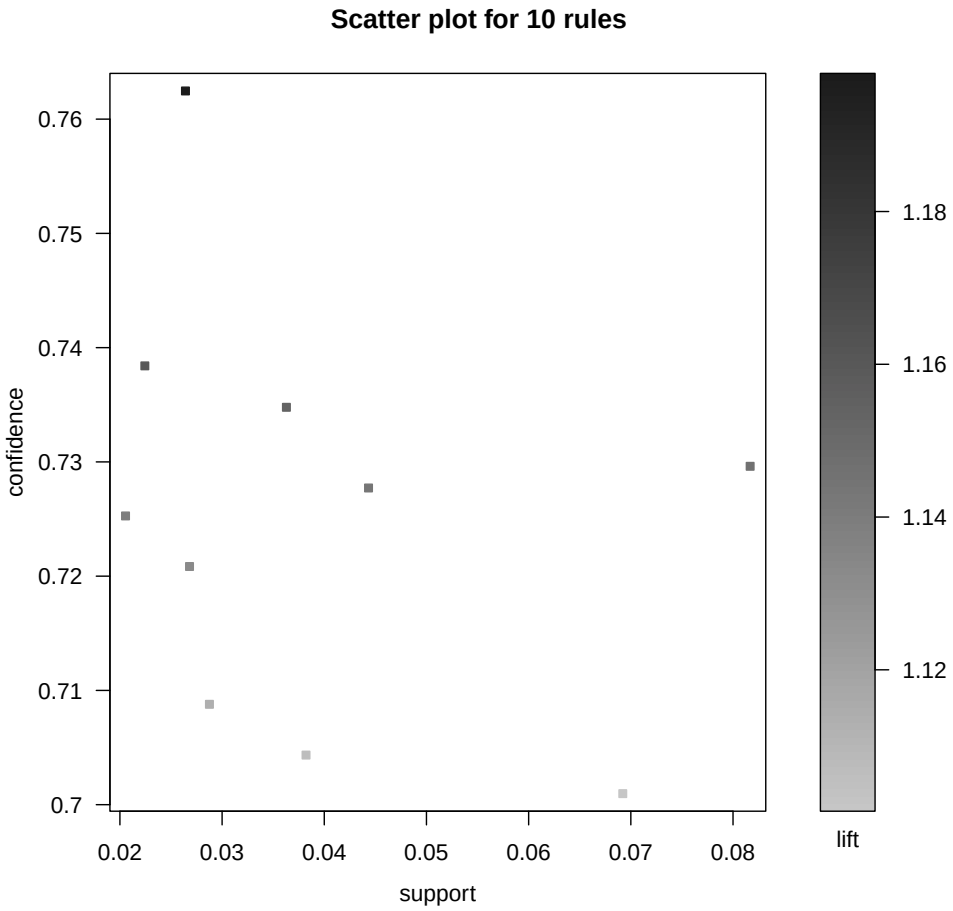
```
## 1  {user_review_length=[1000,5060],
##     user_rating=1}                       => {user_is_local=TRUE} 0.01139263  0.727
## 2  {user_review_length=[1000,5060],
##     user_rating=2}                       => {user_is_local=TRUE} 0.02055149  0.725
## 3  {user_rating=2,
##     user_num_reviews=[   1,  12)}        => {user_is_local=TRUE} 0.01890659  0.729
## 4  {user_rating=2,
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.03821361  0.704
## 5  {user_review_length=[1000,5060],
##     user_num_reviews=[   1,  12)}        => {user_is_local=TRUE} 0.02641269  0.762
## 6  {user_review_length=[1000,5060],
##     user_num_reviews=[  12, 134)}        => {user_is_local=TRUE} 0.06921162  0.700
## 7  {user_review_length=[1000,5060],
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.08167600  0.729
## 8  {user_num_reviews=[   1,  12),
##     user_review_time_period=[2009,2013)} => {user_is_local=TRUE} 0.04432999  0.727
## 9  {user_review_length=[1000,5060],
##     user_rating=2,
##     user_num_reviews=[  12, 134)}        => {user_is_local=TRUE} 0.01019527  0.725
## 10 {user_review_length=[1000,5060],
##     user_rating=2,
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01145152  0.763
## 11 {user_rating=2,
##     user_num_reviews=[   1,  12),
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01311998  0.713
## 12 {user_rating=2,
##     user_num_reviews=[  12, 134),
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01916177  0.714
## 13 {user_review_length=[ 300,1000),
##     user_rating=2,
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01941302  0.700
## 14 {user_review_length=[1000,5060],
##     user_rating=3,
##     user_num_reviews=[  12, 134)}        => {user_is_local=TRUE} 0.01465104  0.718
## 15 {user_review_length=[1000,5060],
##     user_rating=3,
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01669245  0.750
## 16 {user_review_length=[1000,5060],
##     user_num_reviews=[   1,  12),
```

```
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01662178  0.756
## 17 {user_review_length=[1000,5060],
##     user_num_reviews=[ 134,5267],
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.02876032  0.708
## 18 {user_review_length=[1000,5060],
##     user_rating=5,
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01995477  0.701
## 19 {user_review_length=[1000,5060],
##     user_rating=4,
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.02681312  0.720
## 20 {user_review_length=[1000,5060],
##     user_num_reviews=[  12, 134),
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.03629390  0.734
## 21 {user_rating=5,
##     user_num_reviews=[   1,  12),
##     user_review_time_period=[2009,2013)} => {user_is_local=TRUE} 0.01429379  0.710
## 22 {user_rating=4,
##     user_num_reviews=[   1,  12),
##     user_review_time_period=[2009,2013)} => {user_is_local=TRUE} 0.01437231  0.714
## 23 {user_review_length=[ 300,1000),
##     user_num_reviews=[   1,  12),
##     user_review_time_period=[2009,2013)} => {user_is_local=TRUE} 0.02243980  0.738
## 24 {user_review_length=[1000,5060],
##     user_rating=4,
##     user_num_reviews=[ 134,5267],
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01186765  0.701
## 25 {user_review_length=[1000,5060],
##     user_rating=4,
##     user_num_reviews=[  12, 134),
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.01163996  0.731

plot(yelp_rules_2)
```



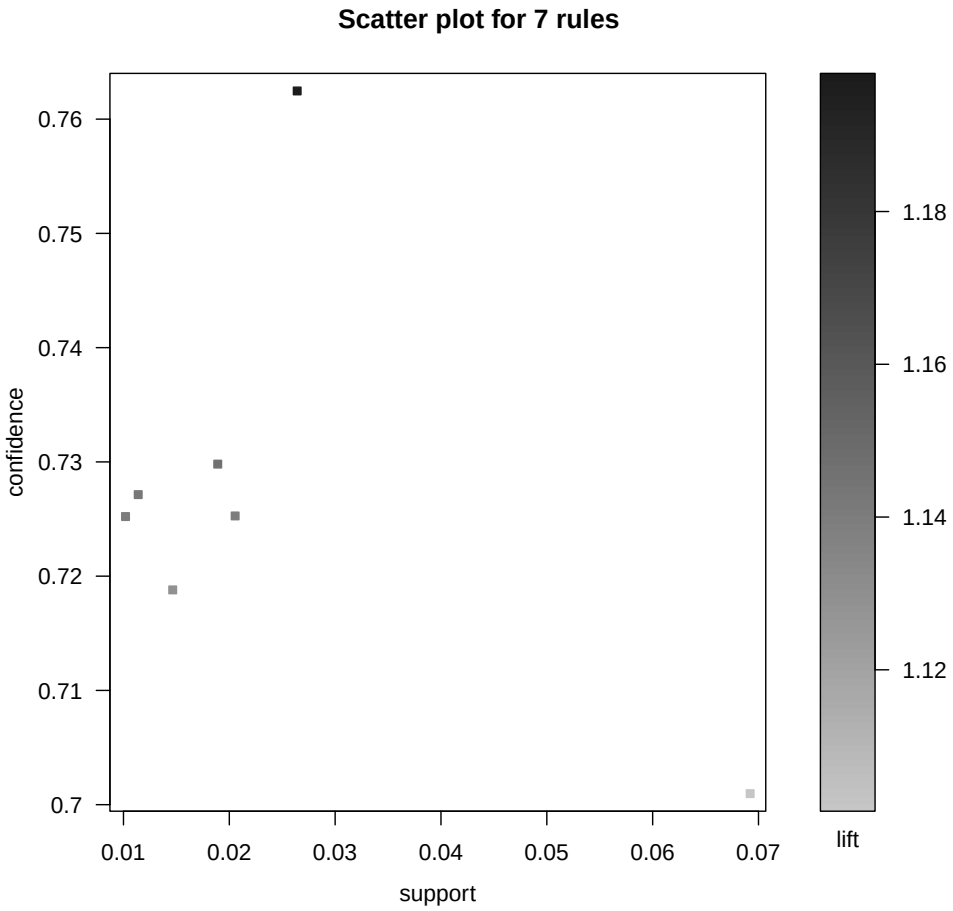

```
# Raising the minimum support, we see that a large user_review_length seems to
# often imply a local reviewer. This includes two rules with support .06 and
# .08, higher than in the previous rule sets.

yelp_rules_3 <- apriori(yelp_data_categorical,
                        parameter = list(minlen=1, supp=.02, conf=0.70),
                        appearance = list(rhs=c("user_is_local=FALSE", "user_is_local
                        control = list(verbose=F))
inspect(yelp_rules_3)

##    lhs                                      rhs                    support confid
## 1  {user_review_length=[1000,5060],
```

```
##     user_rating=2}                       => {user_is_local=TRUE} 0.02055149  0.725
## 2  {user_rating=2,
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.03821361  0.704
## 3  {user_review_length=[1000,5060],
##     user_num_reviews=[   1,  12)}        => {user_is_local=TRUE} 0.02641269  0.762
## 4  {user_review_length=[1000,5060],
##     user_num_reviews=[  12, 134)}        => {user_is_local=TRUE} 0.06921162  0.700
## 5  {user_review_length=[1000,5060],
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.08167600  0.729
## 6  {user_num_reviews=[   1,  12),
##     user_review_time_period=[2009,2013)} => {user_is_local=TRUE} 0.04432999  0.727
## 7  {user_review_length=[1000,5060],
##     user_num_reviews=[ 134,5267],
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.02876032  0.708
## 8  {user_review_length=[1000,5060],
##     user_rating=4,
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.02681312  0.720
## 9  {user_review_length=[1000,5060],
##     user_num_reviews=[  12, 134),
##     user_review_time_period=[2013,2015]} => {user_is_local=TRUE} 0.03629390  0.734
## 10 {user_review_length=[ 300,1000),
##     user_num_reviews=[   1,  12),
##     user_review_time_period=[2009,2013)} => {user_is_local=TRUE} 0.02243980  0.738

plot(yelp_rules_3)
```




```
# It does not seem like the year provided any useful information here, so we
# also try mining rules without this variable. Doing this shows more clearly
# that longer reviews seem to imply local reviewers.

yelp_rules_4 <- apriori(yelp_data_categorical[ , -5],
                        parameter = list(minlen=1, supp=.01, conf=0.70),
                        appearance = list(rhs=c("user_is_local=FALSE", "user_is_local
                        control = list(verbose=F))
inspect(yelp_rules_4)

##   lhs                                    rhs                    support confidence
## 1 {user_review_length=[1000,5060],
```

```
##    user_rating=1}                  => {user_is_local=TRUE} 0.01139263  0.7271361 1
## 2 {user_review_length=[1000,5060],
##    user_rating=2}                  => {user_is_local=TRUE} 0.02055149  0.7252702 1
## 3 {user_rating=2,
##    user_num_reviews=[   1,  12)}   => {user_is_local=TRUE} 0.01890659  0.7298075 1
## 4 {user_review_length=[1000,5060],
##    user_num_reviews=[   1,  12)}   => {user_is_local=TRUE} 0.02641269  0.7624660 1
## 5 {user_review_length=[1000,5060],
##    user_num_reviews=[  12, 134)}   => {user_is_local=TRUE} 0.06921162  0.7009662 1
## 6 {user_review_length=[1000,5060],
##    user_rating=2,
##    user_num_reviews=[  12, 134)}   => {user_is_local=TRUE} 0.01019527  0.7252164 1
## 7 {user_review_length=[1000,5060],
##    user_rating=3,
##    user_num_reviews=[  12, 134)}   => {user_is_local=TRUE} 0.01465104  0.7187982 1

plot(yelp_rules_4)
```




```
# We use a third support level to look for the most frequent item sets overall.
# We see heree with rule 3 that user number of reviews in the medium category
# [12, 134) are commonly associated with local reviewers.
yelp_rules_5 <- apriori(yelp_data_categorical,
                        parameter = list(minlen=1, supp=.25, conf=0.1),
                        appearance = list(rhs=c("user_is_local=FALSE", "user_is_local
                        control = list(verbose=F))
inspect(yelp_rules_5)

##   lhs                                      rhs                     support confide
## 1 {}                                    => {user_is_local=FALSE} 0.3633394  0.3633
## 2 {}                                    => {user_is_local=TRUE}  0.6366606  0.6366
```

```
## 3 {user_num_reviews=[  12, 134)}        => {user_is_local=TRUE}  0.3276580  0.6553
## 4 {user_review_length=[ 300,1000)}      => {user_is_local=TRUE}  0.3203364  0.6356
## 5 {user_review_time_period=[2013,2015]} => {user_is_local=TRUE}  0.3816100  0.6538

plot(yelp_rules_5)
```