Cocktail Recommendation Engine

Byteflow Dynamics

10/22/2017

Contents

```
Warning: package 'tidyr' was built under R version 3.4.1
```

0.1 Introduction

While there are many recommendation based engines in the market, such as Netflix and Amazon, it is hard to find a good one for cocktails. We aim to create such an engine which in addition to recommending a cocktail based on previous ratings, also suggests replacement for missing ingredients needed to make the drink.

The data is scraped from thecocktaildb which has a great API.

Current works

Relationship between ingredients. This is useful for ingredient replacement.

Clusters of cocktails. Useful for recommending new drinks within cluster.

Future Works

1. Scrape more data from boutique bars/hotels for better cocktails. (Start with New York based bars)

2. *Drink-a-gram* Create an app for users to rate cocktails when checking in at a bar.
   - search engine for specific cocktails in bars by distance.
   - sell the ranking when tied, which ones to show first. (when searching for a cousine on yelp for example, what shows first is not random or by ranking…the default option is the highest bidder. One can however filter by ranking.)

0.2 Teaching Topics

- data cleaning. dealing with white spaces, douple white spaces, “/n”, “TAB” etc.
  - TASK: Detect all anomalities and replace with zero.
- use domain knowledge in dim reduction. i.e: nobody wants to make a cocktail with more then eight ingredients.
  - TASK: Select cocktails with seven or less ingredients.
- tidy data: how to make this data tidy.
  - TASK: make the data tidy.
- Exploratory Data Analysis (EDA). Explore the data, what are the most common ingridents, least common ingredients.
  - TASK: list the 10 most common and 10 least common ingredients.
- Vector Multiplication: To calculate how close two cocktails are, we convert the data into a matrix and take inner products. If inner product is zero then they have no ingridients in common, if one..it is the same cocktail but different name.
- Unsupervised Learning: K-means Clustering.

1 Data Cleaning. Dealing with missing values

Load packages we are using.

```r
library(tidyverse)
library(stringr)
```

First let’s load and look at the data file. Explore the data. Are there any missing values? (3 min)

[redacted]

```r
# read the data, set blank spaces and whitespaces to NAs. The data lives locally.
data_base <- read.csv("database.csv", header=T,
                     stringsAsFactors=FALSE)
head(data_base)
```

```
  X drinks.idDrink         drinks.strDrink drinks.strCategory
1 1          11000      A Furlong Too Late     Ordinary Drink
2 2          11001 A Night In Old Mandalay     Ordinary Drink
3 3          11002                   A. J.     Ordinary Drink
4 4          11003          Abbey Cocktail     Ordinary Drink
5 5          11004                 Abilene     Ordinary Drink
6 6          11005                Acapulco     Ordinary Drink
  drinks.strAlcoholic     drinks.strGlass
1           Alcoholic      Highball glass
2           Alcoholic      Highball glass
3           Alcoholic      Cocktail glass
4           Alcoholic      Cocktail glass
5           Alcoholic      Highball glass
6           Alcoholic Old-fashioned glass

1
2 In a shaker half-filled with ice cubes, combine the light rum, a\xf1ejo rum, orange juice, and lemon
3
4
5
6                                                                                          Combine and
  drinks.strDrinkThumb drinks.strIngredient1 drinks.strIngredient2
1                 <NA>             Light rum           Ginger beer
2                 <NA>             Light rum          A\xf1ejo rum
3                 <NA>             Applejack      Grapefruit juice
4                 <NA>                   Gin        Orange bitters
5                 <NA>              Dark rum          Peach nectar
6                 <NA>             Light rum            Triple sec
  drinks.strIngredient3 drinks.strIngredient4 drinks.strIngredient5
1            Lemon peel
2          Orange juice           Lemon juice            Ginger ale
3
4                Orange                Cherry
5          Orange juice
6            Lime juice                 Sugar             Egg white
  drinks.strIngredient6 drinks.strIngredient7 drinks.strIngredient8
1
2            Lemon peel
3
4
5
6                  Mint
  drinks.strIngredient9 drinks.strIngredient10 drinks.strIngredient11
1
2
3
4
5
6
```

```
  drinks.strIngredient12 drinks.strIngredient13 drinks.strIngredient14
1                                            NA                     NA
2                                            NA                     NA
3                                            NA                     NA
4                                            NA                     NA
5                                            NA                     NA
6                                            NA                     NA
  drinks.strIngredient15 drinks.strMeasure1 drinks.strMeasure2
1                     NA               2 oz               4 oz
2                     NA               1 oz               1 oz
3                     NA           1 1/2 oz               1 oz
4                     NA           1 1/2 oz             1 dash
5                     NA           1 1/2 oz               2 oz
6                     NA           1 1/2 oz          1 1/2 tsp
  drinks.strMeasure3 drinks.strMeasure4 drinks.strMeasure5
1         1 twist of
2               1 oz             1/2 oz               3 oz
3
4       Juice of 1/4                  1
5               3 oz
6            1 tblsp              1 tsp                  1
  drinks.strMeasure6 drinks.strMeasure7 drinks.strMeasure8
1
2         1 twist of
3
4
5
6                  1
  drinks.strMeasure9 drinks.strMeasure10 drinks.strMeasure11
1
2
3
4
5
6
  drinks.strMeasure12 drinks.strMeasure13 drinks.strMeasure14
1                                      NA                  NA
2                                      NA                  NA
3                                      NA                  NA
4                                      NA                  NA
5                                      NA                  NA
6                                      NA                  NA
  drinks.strMeasure15 drinks.dateModified
1                  NA                <NA>
2                  NA                <NA>
3                  NA                <NA>
4                  NA                <NA>
5                  NA                <NA>
6                  NA                <NA>
```

Looking at the structure of the dataset there is a lot of cleaning to do. This is common with most data. In this case it seems like it comes from different sources combined into one.

Let’s check how many NAs are in the data.

```
# To check the number of missing values

# For the entire data frame.

total_na <- sum(is.na(data_base))
total_na
```

```
[1] 25399
```

```
# thats a lot of missing values.

# total entries

total_entries <- nrow(data_base)*ncol(data_base)
total_entries
```

```
[1] 124878
```

```
# only a fraction of the data is complete.

100 * total_na/total_entries
```

```
[1] 20.33905
```

20% of data is missing. Let’s handle these missing data first.

1.0.1 Replace missing values with NAs

The problem here is that there are multiple forms of missing values: “”(empty cell), " “(single space),”/t“,”\n“, NA. First let’s convert all of them into NAs. The easiest way to do this is to add na.strings=( ) argument when reading the file

```
# read the data, set blank spaces and whitespaces to NAs. The data lives locally.
data_base<- read.csv("database.csv", header=T,
                     na.strings=c(""," ", "/t","\n", NA), stringsAsFactors=FALSE)
```

Check all empty cells are now filled with NAs.

Now we can replace NAs with 0s.

```
data_base[is.na(data_base)]<-0
```

Check the data frame again to make sure all missing values are 0 now.

2 Remove some Columns

Now that we have all the data let us explore it and standardize it.

In this case we can use domain knowledge in dimension reduction. i.e: nobody wants to make a cocktail with more than seven ingredients.

2.0.1 Can you think of an elegant and efficient way to remove all the cocktails that have more than 7 ingredients?

We can use filter to keep only cocktails with less than 8 ingredients (Ingredient8 column should be 0)

```
db <- data_base %>%
  filter(drinks.strMeasure8 == 0) # drop rows that have 8th ingredient
```

2.0.2 Subset data base to drop irrelevant columns

Now we only keep cocktail names, ingredients and measures.

```
db <- db %>%
  select(cocktail.name = drinks.strDrink,
         drinks.strIngredient1:drinks.strIngredient7,
         drinks.strMeasure1:drinks.strMeasure7)
```

3 Make data tidy

TASK: 10 min

Sketch out on a piece of paper how tidy data should look. How would you go about it. One reason why: get the top 10 most common ingredient. Very easy to do with tidy data.

```
db2 <- db %>%
  gather(-cocktail.name, key = "key", value = "value") %>% #gather all ingredient and measure columns
  mutate(type = str_replace(key, "\\d+", "")) %>% #make a new column type: ingredient or measure
  mutate(type = str_sub(type, start=11)) %>%
  mutate(key = str_replace_all(key, "[^0-9]", "")) %>% #make key column only digits
  arrange(cocktail.name, key) %>% # sort by cocktail name and key
  spread(key = type, value = value) # finally, spread ingredient and measure
```

3.0.1 Drop rows where Measure == 0

```
db_tidy <- db2 %>%
  filter(Measure != 0)
```

4 EDA

Explore the data, what are the most common ingredients, least common ingredients.

TASK: (15 min)

List the 10 most and least common ingredients.

```
db_tidy %>% head()
```

```
                       cocktail.name key      Ingredient
1 '57 Chevy with a White License Plate   1 Creme de Cacao
2 '57 Chevy with a White License Plate   2          Vodka
3              \xd6xn\xe4s Temptation   1          Vodka
4              \xd6xn\xe4s Temptation   2 Banana liqueur
```

```
5                \xd6xn\xe4s Temptation   3         Sprite
6                \xd6xn\xe4s Temptation   4   Orange juice
                  Measure
1              1 oz white
2                    1 oz
3                    6 cl
4                    2 cl
5 Nearly fill glass with
6                1 splash
```

```
# we want to group by ingredient then sum the frequency of apparence

db_top <- db_tidy %>%
  group_by(Ingredient) %>%
  summarise(N = n()) %>%
  arrange(desc(N))


db_top %>% head(10)
```

```
# A tibble: 10 x 2
        Ingredient     N
             <chr> <int>
 1           Vodka   621
 2             Gin   453
 3    Orange juice   356
 4     Lemon juice   261
 5       Grenadine   232
 6           Sugar   228
 7      Triple sec   218
 8 Pineapple juice   209
 9             Ice   204
10       Light rum   194
```

```
db_bot <- db_tidy %>%
  group_by(Ingredient) %>%
  summarise(N = n()) %>%
  arrange(N)


db_bot %>% head(10)
```

```
# A tibble: 10 x 2
              Ingredient     N
                   <chr> <int>
 1               Acerola     1
 2 Apple-cranberry juice     1
 3            Asafoetida     1
 4            Banana rum     1
 5          Blackberries     1
 6        Bloody mary mix     1
 7           Blueberries     1
 8            Cantaloupe     1
 9        Caramel liqueur     1
10                Celery     1
```

Cool we see that Vodka rules, Gin is number two, then comes the citrus: lemon and orange juice. On the other hand, some fresh fruit and vegetable are least common. However these numbers do not tell the correct total. The same liquor in this data set are sometimes represented with different names. i.e Vodka and Absolute Vodka. For now we skip this part.

5 Standardize the unit of measure

In the column Measure, we have different units of measure, such as ounces, table spoons, teaspoons etc. We need to make this uniform. We do this so that when taking the inner products to gauge how similar two cocktails are, the ingredients have appropriate weights.

Let's separate Measure into numbers and units. We create 3 number columns: integer, decimal, and fraction.

```
db_tidy %>% head()
```

```
                         cocktail.name key     Ingredient
1 '57 Chevy with a White License Plate   1 Creme de Cacao
2 '57 Chevy with a White License Plate   2          Vodka
3               \xd6xn\xe4s Temptation   1          Vodka
4               \xd6xn\xe4s Temptation   2 Banana liqueur
5               \xd6xn\xe4s Temptation   3         Sprite
6               \xd6xn\xe4s Temptation   4   Orange juice
                  Measure
1             1 oz white
2                   1 oz
3                   6 cl
4                   2 cl
5 Nearly fill glass with
6               1 splash
```

```
db_tidy2 <- db_tidy %>%
  mutate(num = str_extract(Measure, "^[:digit:]+ ")) %>%
  mutate(num = str_replace_na(num, "1")) %>% # we want to keep measures that don't have digits
  mutate(num_dec = str_extract(Measure, "[:digit:][.,][:digit:]")) %>%
  mutate(num_dec = str_replace(num_dec, ",", ".")) %>%
  mutate(num_dec = str_replace_na(num_dec, "0")) %>%
  mutate(frac = str_extract(Measure, "[:digit:]/[:digit:]")) %>%
  mutate(frac = str_replace_na(frac, "0")) %>%
  mutate(unit = str_replace_all(Measure, "[:digit:]", "") ) %>%
  mutate(unit = str_replace_all(unit, "[^[:alpha:]]", " ")) %>%
  mutate(unit = str_replace_all(unit, "^\\s+$", ""))

#Check unique values
unique(db_tidy2$units)
```

```
NULL
```

We can now convert units into mL with the proper conversion.

```
# Now we can drop Measure column
db_tidy3 <- db_tidy2 %>%
   select(-Measure)

db_tidy3 %>% head()
```

```
                         cocktail.name key     Ingredient num num_dec frac
```

[redacted]

```
1 '57 Chevy with a White License Plate   1 Creme de Cacao  1        0    0
2 '57 Chevy with a White License Plate   2          Vodka  1        0    0
3               \xd6xn\xe4s Temptation   1          Vodka  6        0    0
4               \xd6xn\xe4s Temptation   2 Banana liqueur  2        0    0
5               \xd6xn\xe4s Temptation   3         Sprite   1       0    0
6               \xd6xn\xe4s Temptation   4   Orange juice  1        0    0
                     unit
1               oz white
2                     oz
3                     cl
4                     cl
5 Nearly fill glass with
6                 splash
```

```
# Replace units with proper conversion to mL
db_tidy3 <- db_tidy3 %>%
   mutate(unit=str_replace(unit,"ozjamaican","oz")) %>%
   mutate(unit=str_replace(unit,"oz","29.5")) %>%
   mutate(unit=str_replace(unit,"shot","29.5")) %>%
   mutate(unit=str_replace(unit,"jigger","44.5")) %>%
   mutate(unit=str_replace(unit,"cup","257")) %>%
   mutate(unit=str_replace(unit,"tblsp","11.1")) %>%
   mutate(unit=str_replace(unit,"tsp","3.7")) %>%
   mutate(unit=str_replace(unit,"ts p","3.7")) %>%
   mutate(unit=str_replace(unit,"teaspoon","3.7")) %>%
   mutate(unit=str_replace(unit,"cl","10")) %>%
   mutate(unit=str_replace(unit,"dl","100")) %>%
   mutate(unit=str_replace(unit,"litre","1000")) %>%
   mutate(unit=str_replace(unit,"liter","1000")) %>%
   mutate(unit=str_replace(unit,"dash","0.9")) %>%
   mutate(unit=str_replace(unit,"splash","3.7")) %>%
   mutate(unit=str_replace(unit,"twist","15")) %>%
   mutate(unit=str_replace(unit,"twistof","15")) %>%
   mutate(unit=str_replace(unit,"can","355")) %>%
   mutate(unit=str_replace(unit,"cube","12")) %>%
   mutate(unit=str_replace(unit,"part","29.5")) %>%
   mutate(unit=str_replace(unit,"pint","473")) %>%
   mutate(unit=str_replace(unit,"glass","473"))

# Check if missing something like glass or a pint add it to the code on top
unique(db_tidy3$unit)
```

```
  [1] " 29.5 white "
  [2] " 29.5 "
  [3] " 10 "
  [4] "Nearly fill 473 with "
  [5] " 3.7 "
  [6] "A tiny 3.7 "
  [7] "  29.5 "
  [8] " 29.5s "
  [9] "  29.5 Bacardi "
 [10] "  29.5 Koskenkorva "
 [11] "   29.5 "
 [12] "Fill with "
 [13] " 29.5 dry "
```

```
[14] " 15 of "
[15] ""
[16] " 12s "
[17] " 44.5 "
[18] "Juice of   "
[19] " 0.9 "
[20] " 473 "
[21] " 3.7 crumbled "
[22] "  3.7 "
[23] " bottle "
[24] " gr "
[25] " ml pure "
[26] " 11.1 "
[27] "Top it up with "
[28] " slice "
[29] "Add "
[30] "Fill to top "
[31] "   3.7 "
[32] " 29.5 Bacardi "
[33] "  473 "
[34] "  355 "
[35] "   29.5 dry "
[36] " 0.9es "
[37] "Juice of  "
[38] " 10 hot "
[39] "  L Cava "
[40] "  L "
[41] " 100 "
[42] "Twist of "
[43] " 29.5 Stefanoffs "
[44] "  29.5 white "
[45] "  3.7 grated "
[46] "  257 "
[47] " 257s "
[48] "   11.1 "
[49] "   257 "
[50] " 257 "
[51] "  257s "
[52] " scoops "
[53] "Add 3.7 "
[54] "  10 "
[55] "  44.5 "
[56] "Add   257 "
[57] " 29.5 hot "
[58] "Chilled "
[59] "Pour in  29.5 "
[60] "Add  0.9es "
[61] "Fill 473 with "
[62] " 29.5 red "
[63] " drop "
[64] " 11.1 fresh "
[65] " fr29.5en ripe "
[66] "Top with "
[67] " 29.5 fresh "
```

```
 [68] " 29.5 Smirnoff "
 [69] "  slice "
 [70] " chopped "
 [71] " ml "
 [72] " large "
 [73] "A 0.9 of "
 [74] "0.9 "
 [75] " fresh "
 [76] " qt "
 [77] " fifth "
 [78] " L "
 [79] "  gal "
 [80] "Fill   473 "
 [81] "Fill "
 [82] " Cubes "
 [83] " 29.5 Green Ginger "
 [84] "lots "
 [85] " 29.5 blue "
 [86] " 29.5 lemon "
 [87] "   29.5 white "
 [88] "  473 crushed "
 [89] " 10 Smirnoff "
 [90] " 29.5s Finlandia "
 [91] " 355 "
 [92] "About  bottle "
 [93] "   10 "
 [94] "   29.5 Bacardi "
 [95] "3.7 Bacardi "
 [96] "Very little granulated "
 [97] " 10 Bacardi "
 [98] "One 473 "
 [99] " 29.5 Jamai355 "
[100] " whole "
[101] " squeeze "
[102] "Half fill "
[103] "Slice of  "
[104] "cracked "
[105] " scoop "
[106] " 3.7 crushed "
[107] "fill with "
[108] " 10 cold "
[109] " sprigs "
[110] "   257s "
[111] " 29.5 frozen "
[112] "  lb fr29.5en "
[113] " fr29.5en "
[114] " 257 plain "
[115] " to taste "
[116] " 257s fresh "
[117] "  473 Bacardi "
[118] " 29.5 cold "
[119] "crushed "
[120] "   29.5 Cuervo premium or "
[121] "Float "
```

```
[122] "Splash "
[123] "Whipped "
[124] "  3.7 granulated "
[125] " 257 crushed "
[126] " 3.7 granulated "
[127] " very ripe "
[128] " 29.5s Stoli "
[129] " 3.7 superfine "
[130] " 29.5 Barbados "
[131] "  29.5 fresh "
[132] " 29.5 strawberry "
[133] "Handfull "
[134] " bottle Boone Strawberry Hill "
[135] " gal Tropical Berry "
[136] " 29.5 Triple Berry "
[137] "Lots "
[138] " wedges "
[139] "Part  "
[140] "  Absolut  "
[141] " 10 Hammer "
[142] " 10 apricot "
[143] "  Farris  Perrier  "
[144] "  Burgundy  "
[145] " 29.5 Bass pale "
[146] " black "
[147] "Pour Over "
[148] " 29.5 strong  black "
[149] "Fifty fifty with "
[150] "And "
[151] "  drops "
[152] " 29.5 boiling "
[153] " 29.5 chilled "
[154] " ring with fruits  pineapple  lemon  grapes  "
[155] "  3.7 superfine "
[156] " to fill "
[157] "  gal premium "
[158] " medium 355 "
[159] "Some Cherry "
[160] " crushed "
[161] "A float of "
[162] "Several drop "
[163] " wedge "
[164] " 12s crushed "
[165] " ever10ear rum  "
[166] "Fill rest of 473 "
[167] " package Strawberry "
[168] "Coarse "
[169] " 29.5 chilled blue "
[170] "0.9 crushed "
[171] "Several 0.9es of "
[172] "  orange pekoe  "
[173] "Rim 473 "
[174] " package "
[175] "Orange "
```

```
[176] "  29.5 Smirnoff "
[177] " 29.5 bottled "
[178] "   3.7 superfine "
[179] "   29.5 blended "
[180] "  3.7 ground "
[181] " fifth Smirnoff red label "
[182] " small bottle "
[183] " 29.5 10ear "
[184] " ml white "
[185] " gal hic berry "
[186] " measures "
[187] " or  29.5 "
[188] "Full 473 of "
[189] "A few squirt "
[190] " 29.5 sweet "
[191] "Layer   29.5 "
[192] "Sprinkle fresh ground "
[193] "   29.5s Strawberry Kiwi "
[194] " 29.5 hard "
[195] "Mostly "
[196] "Healthy 3.7 "
[197] "Small 3.7 "
[198] "Sprinkling "
[199] "  Coco Lopez  "
[200] "  12s "
[201] "Blend with "
[202] "To fill "
[203] "Fill With "
[204] "Crushed "
[205] "wedge "
[206] "Add to taste "
[207] "   29.5 Canadian "
[208] " stick "
[209] " drops "
[210] " 29.5 soft "
[211] "   257 superfine "
[212] " scoop crushed "
[213] " pieces "
[214] "Some "
[215] " piece "
[216] " bottle chilled "
[217] "  bottle "
[218] " 10 indian "
[219] " 10 cheap "
[220] " x 29.5 355s "
[221] "Layered on   29.5 "
[222] "Add  29.5 "
[223] " 29.5 Mexi355 "
[224] " 3.7 powdered "
[225] " gr semi sweet "
[226] " 11.1 shaved sweet "
[227] "  473 dry "
[228] "  29.5 Muscatel "
[229] "Shredded "
```

```
[230] " 11.1 instant "
[231] " 257s white "
[232] "  257 instant "
[233] " bottle Cold Duck "
[234] " 355 fr29.5en "
[235] "  gal rainbow "
[236] " 3.7es "
[237] "Float Bacardi "
[238] "As many wedge "
[239] " 29.5 evaporated "
[240] "  sticks "
[241] " 29.5 Stoli "
[242] "Small 0.9 "
[243] " 355s "
[244] " 355s fr29.5en "
[245] " or  29.5s "
[246] "  kg coarsely chopped "
[247] " lb "
[248] " slices "
[249] " 29.5 sweetened "
[250] "   257 almond  mint  orange or  "
[251] "3.7 "
[252] " scoop vanilla fr29.5en "
[253] " 29.5 cream "
[254] " 3.7 instant "
[255] " 3.7 boiled "
[256] " 29.5 heavy "
[257] " 29.5 pure "
[258] " 11.1 green "
[259] " 44.5s "
[260] " or  Up "
[261] "Squeeze "
[262] " bottle cold "
[263] " 100 cold "
[264] "fill "
[265] "Layer  29.5 "
[266] "Fill with   "
[267] "Fill with  29.5 "
[268] " pinch "
[269] "fill 473 "
[270] "Tahiti Treat or "
[271] " 29.5 Blue Label Smirnoff "
[272] " 29.5 Skyy "
[273] " ml Blue label Smirnoff "
[274] " ml Red Label "
[275] " or  3.7es "
[276] "Half mug "
[277] " 29.5 finely chopped dark "
[278] "Fresh "
[279] "Float   29.5 "
[280] " 11.1 carob or "
[281] " 3.7 shaved sweet "
[282] "  257 crushed "
[283] "Lots of "
```

```
[284] " 3.7 Jamai355 "
[285] " 257s"
[286] " 3.7"
[287] " 1000"
[288] " 29.5 high proof "
[289] " package Orange "
[290] " package Lemon Lime "
[291] " 29.5 whole "
[292] " 29.5 skimmed "
[293] "  257 skimmed "
[294] " separated "
[295] " 257 granulated "
[296] "Grated "
[297] "  257 peach or apricot "
[298] "Freshly ground "
[299] " qt Egg Nog "
[300] "  Stoli "
[301] " 44.5 red "
[302] " 355s light "
[303] " inch Russian "
[304] "Add  0.9 "
[305] " cola "
[306] "  29.5 cream "
[307] " green "
[308] "  red "
[309] " 44.5 Stoli "
[310] "Bacardi "
[311] " 3.7 white "
[312] " 10 crushed "
[313] " cocktail "
[314] " 10 Finlandia "
[315] "   3.7 Fino "
[316] " 29.5s red "
[317] "   44.5 "
[318] " 3.7 freshly squeezed "
[319] "  or  UP  "
[320] "  as desired "
[321] " 29.5 amber "
[322] "  257 strong black "
[323] " 29.5 Chilled "
[324] "  473 Orangina "
[325] "Fill 473 "
[326] " 12 "
[327] "   257 fresh "
[328] " 257s granulated "
[329] " drops green "
[330] " drops blue "
[331] " 29.5s blue "
[332] "Granulated "
[333] " 3.7 coarse "
[334] " chunks "
[335] " gal "
[336] " packages unsweetened red "
[337] " as needed "
```

```
[338] " 257 fruit "
[339] "  piece textural "
[340] " crate "
[341] "Top with Bacardi "
[342] "Dash "
[343] " bottle Smirnoff "
[344] " bottles "
[345] "   gal fresh squeezed "
[346] " bag "
[347] "  kg "
[348] "Mix in  11.1 "
[349] "And  29.5 "
[350] "  measure "
[351] " ml green "
[352] " 257s distilled "
[353] " 257 white "
[354] "  drops yellow "
[355] " 29.5s hot "
[356] " 10 fresh "
[357] "   29.5s "
[358] " bottle diced "
[359] " 11.1 medium dry "
[360] " pinch ground "
[361] "  473 heavy "
[362] "  pieces minced crystallized "
[363] "float "
[364] "  piece "
[365] "  or vodka or schnapps  "
[366] " sticks "
[367] "  100 "
[368] " thing "
[369] "  11.1 "
[370] " qt chilled "
[371] " 473 lemon or orange "
[372] " ring "
[373] "Fill remainder with "
[374] " measure "
[375] " unbroken "
[376] "   scoop "
[377] " 29.5 Gill "
[378] "  packages "
[379] " bags "
[380] "  257 granulated "
[381] " 10 Koskenkorva salmiac "
[382] " 10 red "
[383] "  slices "
[384] " or Chambourd "
[385] " or   473 "
[386] " or   473 Bacardi "
[387] " 29.5s Bacardi "
[388] "Mango "
[389] "Appx  10 "
[390] " drop blue "
[391] " leaves "
```

```
[392] " 29.5 yellow "
[393] "  optional  "
[394] "Some drop "
[395] "  473 hard "
[396] " 29.5 green "
[397] " small "
[398] "Add  29.5s "
[399] " 10 skimmed "
[400] " 29.5 ruby red "
[401] "Fill up "
[402] "   29.5  proof "
[403] "  0.9es "
[404] "Garnish "
[405] "Topper "
[406] "On top "
[407] "Little "
[408] "very sweet "
[409] "Fill with 473 "
[410] "  orange "
[411] " 29.5 peeled   crushed "
[412] "Peel of  small "
[413] "   257s boiling "
[414] "  257 lukewarm "
[415] "   29.5 instant "
[416] " 257s boiling "
[417] " 257 strong "
[418] " 257 cold "
[419] " to add tartness  optional  "
[420] " gr pure "
[421] " packages Ameri355 "
[422] " 355s fr29.5en lemon lime "
[423] " long strip "
[424] "  lb salted "
[425] " 3.7 ground "
[426] " 3.7 ground white "
[427] "  257 white "
[428] "  lb "
[429] " 473 good quality "
[430] " 29.5 fine "
[431] "mini "
[432] "A few whole "
[433] "   257 hot "
[434] "Add  29.5 hot "
[435] "  washed "
[436] " gal good "
[437] "  inch "
[438] " slice fresh "
[439] "Fill with hot "
[440] " 44.5 light or dark "
[441] "A lot of "
[442] "  257 hot "
[443] " 257s cold "
[444] "Strong cold "
[445] " 355 sweetened "
```

```
[446] "Fill  12s "
[447] "Ground green "
[448] "   29.5 Stoli "
[449] "To taste "
[450] "  29.5 amber "
[451] " 29.5s blond "
[452] " 29.5s dry "
[453] " drop Red "
[454] "  473 strong black "
[455] "  473 cold "
[456] "Add  3.7 "
[457] "  29.5  small boxI  "
[458] " 257 boiling "
[459] " packages "
[460] " large package Black Cherry "
[461] " 257s hot "
[462] "A little "
[463] "   29.5 Finlandia "
[464] "Around rim put  pinch "
[465] "Fill to top with "
[466] "Swirl of "
[467] " 11.1 fr29.5en "
[468] " bottle Bacardi "
[469] " 29.5 instant "
[470] " 11.1 good fresh coarsely ground "
[471] " 29.5 strong "
[472] "  gal cheap "
[473] " scoops fudge "
[474] " raw "
[475] "Juice of  wedges "
[476] " handful "
[477] " inch "
[478] " 257 hot "
[479] "Optional   29.5 "
[480] "  29.5s "
[481] "A few drops "
[482] " or Sprite "
[483] " 10 finlandia "
[484] " 29.5 light "
[485] "  3.7 Tropical "
[486] "  29.5 Grape "
[487] "    29.5 "
[488] "Till with  29.5 "
[489] "Turkish apple "
[490] "  if needed  "
[491] " 29.5 Finlandia "
[492] "To fill blender "
[493] " Caguamas tecate "
[494] " pinches "
[495] "      29.5 "
[496] "Ground "
[497] " 10 dry "
[498] " 10 boiling "
[499] "Full 473 "
```

[redacted]

```
[500] "Remainder "
[501] "  257 plain "
[502] "   257 cold "
[503] "  3.7 ground roasted "
[504] "  3.7 dried "
[505] " 257 iced "
[506] "pinch "
[507] " 29.5 Cruzan "
[508] " 29.5 Coco Lopez "
[509] "One or more whole "
[510] " or lemon lime juice  to cover eggs  "
[511] " 29.5 brewed "
[512] " 11.1 granulated "
[513] "Mix with  29.5 "
[514] "Add  11.1 "
[515] "Juice of  wedge "
[516] "Fr29.5en "
[517] "cold "
[518] "  355s "
[519] "ground "
[520] " 29.5 crushed "
[521] "  seltzer water  "
[522] "  Bacardi "
[523] "  scoops vanilla or "
[524] "   29.5 Black Cherry "
[525] "Fill with      355 "
[526] " 100 Schweppes "
[527] " 29.5 Cherry "
[528] "  29.5 Fino or dry "
[529] "drop "
[530] "  29.5 dry "
[531] "Less than   29.5 "
[532] " syrup "
[533] "Add   bottle indian "
[534] " package Peach Passion Fruit "
[535] "  Dole  "
[536] "  257 Hawaiian Plantations Lilikoi "
[537] "  257 Hawaiian Plantations "
[538] "   29.5 Blended "
[539] " or  "
[540] "  Makers Mark  "
[541] " 0.9 white "
[542] " ml Fresh "
[543] "Unsweetened "
[544] " 29.5 cherry "
[545] "fr29.5en "
[546] " 11.1 hot "
[547] " 29.5 Hazlenut "
[548] "  29.5 double "
[549] "  29.5 freshly squeezed "
[550] "Float  ml "
[551] " 3.7 sweetened "
[552] "Fill up  10 fresh "
[553] " 29.5 mint flavored "
```

[redacted]

```
[554] " medium size "
[555] "   29.5 frozen "
[556] "12 "
[557] "  29.5s fr29.5en "
[558] "  fresh "
[559] "  Sunny Delight  "
[560] "  or lime slice  "
[561] "Slices of   "
[562] "  handful "
[563] "  L Jamai355 "
[564] " or lemon  with skin  "
[565] " or vodka "
[566] " 29.5 Berry Blue "
[567] " 29.5 premium "
[568] "Fill with  "
[569] "   29.5 Smirnoff "
[570] " or Cherries "
[571] "Equal amount "
[572] "Fill up with "
[573] "Dash of "
[574] " 29.5 black brewed "
[575] "Fill   "
[576] "  Claret  "
[577] " 29.5 fr29.5en strawberry "
[578] " 257 steamed "
[579] " 29.5s grapefruit "
[580] "A little freshly squeezed "
[581] "As much as you wish "
[582] " count "
[583] "A few  drops "
[584] " 10 champagne flavored "
[585] "  Farris  "
[586] " 29.5 chopped bittersweet or semi sweet "
[587] "  inch strips "
[588] "  3.7 instant "
[589] "  presweetened "
[590] "Over "
[591] "Strawberry or "
[592] "  single "
[593] "Sprinkle "
[594] "About  "
[595] " 29.5 pear "
[596] " 10 Koskenkorva "
[597] "With "
[598] "Then "
[599] " squirt "
[600] " Mer  non carbonated "
[601] "  or Sprite  "
[602] "   L "
[603] " 3.7 Bacardi "
[604] "wedge fresh "
[605] " 29.5 strawberry kiwi "
[606] "  scoops "
[607] " 0.9 grape and apple "
```

```
[608] " 11.1 white "
[609] "  29.5 cold semi skimmed "
[610] "     29.5 "
[611] " case Molson Canadian "
[612] "  29.5 29.5 "
[613] " 10 strawberry kiwi "
[614] "Plenty of "
[615] "  29.5 red "
[616] "Add  ml "
[617] " 10  proof "
[618] "  29.5 sweet "
[619] "   11.1 raspberry "
[620] " ml Bacardi "
[621] " 29.5  proof "
[622] "Zest "
[623] " mix  Mr    Mrs  T  "
[624] " 29.5 Russian "
[625] " 29.5 White "
[626] "Slice of   "
[627] " and or lemon slices "
[628] " 29.5 Early Times straight Kentucky "
[629] "Add   12s "
[630] "Fill   29.5 "
[631] " 10 strong  black "
[632] " counts "
[633] "  or  Up  "
[634] " on top "
[635] "Top with  3.7 "
[636] "Splash in "
[637] " 10 blended "
[638] " ml hot "
[639] " 29.5 light or dark "
[640] "mikey bottle "
[641] "large bottle "
[642] " ml fr29.5en "
[643] "  355 silver "
[644] "   355s iced "
[645] " 3.7 fresh "
[646] "  29.5 Ruby red "
[647] " 0.9es Russian "
[648] "Top off  29.5 "
[649] " 3.7 whole "
[650] "  29.5 plain "
[651] "A few drops of "
[652] " bottle Chablis "
[653] "   257 mild "
[654] " 3.7 dried and chopped "
[655] "  3.7 crushed "
[656] "Add a bit "
[657] "A 0.9 "
[658] " 29.5s sweet non alcoholic "
[659] "The rest "
[660] "A handful of crushed "
[661] " 3.7 blue "
```

[redacted]

```
[662] "   29.5 cream "
[663] "    0.9es "
[664] " stir "
[665] "Mix of  29.5s "
[666] "Fill with Purplesaurus Rex "
[667] "   proof  "
[668] "  29.5 tropical "
[669] "  0.9 "
[670] "   29.5 blue "
[671] "  473 sweet or dry "
[672] "A little bit of "
[673] " 29.5s Mango Madness "
[674] "  29.5 cold "
[675] "Add a few "
[676] "Fill half  10 "
[677] "Fill rest  10 "
[678] "  29.5 amontillado "
[679] " beaten "
[680] "About  29.5 "
[681] "Fill whith "
[682] "Grape "
[683] "  29.5 oz white "
[684] "Sweet "
[685] " bottles chilled "
[686] " kg "
[687] "  257 fr29.5en "
[688] "Jucie of  "
[689] "  ripe "
[690] " 473 lemon lime "
[691] "Top With  11.1 "
[692] "  fr29.5en "
[693] "Float   ml "
[694] " 29.5s white "
[695] "    gal "
[696] "Juice of  gal "
[697] "  29.5  proof "
[698] "By taste "
[699] "Half Fill With "
[700] "  10 fresh "
[701] "  10 red "
[702] " 29.5 pureed frozen "
[703] "Add   10 "
[704] "Fizz on top   10 "
[705] "Fill 473 sweet "
[706] " 11.1 grated "
[707] " handfuls "
[708] " 29.5s cold aromatic "
[709] "full 473 "
[710] "About  drops "
[711] " 11.1 Fine ground whole  rich "
[712] "  whole green "
[713] "Strong  black ground "
[714] " pods "
[715] "  257 strong Thai "
```

```
[716] "  257 boiling "
[717] "  handfuls "
[718] "Add crushed "
[719] " 29.5 fr29.5en "
[720] "some chunk "
[721] "Top with fresh "
[722] " L Orangina "
[723] "  257"
[724] " whole"
[725] " chunk "
[726] " 29.5 Genny  horse "
[727] "  473 strong "
[728] " 0.9 dry "
[729] "   29.5 pure "
[730] "Add a 0.9 of "
[731] " pieces fr29.5en "
[732] " 257s crushed "
[733] " add ice "
[734] " 29.5s Smirnoff "
[735] " gal high proof "
[736] "Juice of     "
[737] "Fill with       gal ice cold "
[738] "  fifth "
[739] "    L "
[740] "  sliced "
[741] "    11.1 "
[742] "  ml "
[743] "  L unflavored "
[744] " 29.5 unsweetened "
[745] "Fill up with Schweppes "
[746] "  packet Tropical punch or Incrediberry "
[747] " 29.5s chopped "
[748] " 29.5 Cinnamon "
[749] "  conserved "
[750] " 10 conserved "
[751] " 29.5 white or "
[752] "Hot "
[753] "Slice  "
[754] " 29.5 Grape "
[755] " or lime "
[756] "  29.5 blood "
[757] "  29.5 dark "
```

Now we deal with fractions and multiply numbers and units to get total amount.

```
db_tidy4 <- db_tidy3 %>%
   # Select all numbers that match the following patterns
   mutate(unit=str_extract(unit,"[:digit:]+[:punct:]*[:digit:]*"))

# set NAs to 1
db_tidy4[is.na(db_tidy4)]<-1

db_tidy4 <- db_tidy4 %>%
  mutate(num = as.numeric(num),
         num_dec = as.numeric(num_dec),
```

[redacted]

```
         frac2 = sapply(frac, function(x) eval(parse(text=x))),
         unit = as.numeric(unit)) %>%
  mutate(value = (num+frac2+num_dec)*unit)

#  Select only cocktail name, ingredient and value
db_tidy5 <- db_tidy4 %>%
   select(cocktail.name,key,Ingredient,value)

db_tidy5 %>% head(10)
```

```
                          cocktail.name key      Ingredient  value
1   '57 Chevy with a White License Plate   1 Creme de Cacao  29.50
2   '57 Chevy with a White License Plate   2          Vodka  29.50
3                 \xd6xn\xe4s Temptation   1          Vodka  60.00
4                 \xd6xn\xe4s Temptation   2 Banana liqueur  20.00
5                 \xd6xn\xe4s Temptation   3         Sprite 473.00
6                 \xd6xn\xe4s Temptation   4   Orange juice   3.70
7                 \xd6xn\xe4s Temptation   5      Grenadine   3.70
8                       110 in the shade   1          Lager 472.00
9                       110 in the shade   2        Tequila  73.75
10                           155 Belmont   1       Dark rum  29.50
```

6 Unsupervised Learning

Hard work is done. Let us now do some machine learning. We are interested in calculating how similar two cocktails are. To do this we convert the dataframe into a term matrix (this is why we did all the work so far) and then calculate the inner products.

```
# Use tidyr to spread the data the same way we did in the beginning of class.

db_spread <- db_tidy5 %>%
   spread(Ingredient, value)

# Replace NAs with 0s
db_spread[is.na(db_spread)]<-0

# Drop key column
db_spread <- db_spread %>%
  select(-key)

# Group by cocktail name
db_spread_comb <- db_spread %>%
   group_by(cocktail.name) %>%
   summarise_all(funs(sum))



# Check ingredients for a random cocktail
i=37
db_spread_comb$cocktail.name[i]
```

```
[1] "Acapulco"
```

[redacted]

```
colnames(db_spread_comb)[which(db_spread_comb[i,]!=0)]
```

```
[1] "cocktail.name" "Egg white"     "Light rum"     "Lime juice"
[5] "Mint"          "Sugar"         "Triple sec"
```

```
# Spread is ready! Let's save it so when getting back to this file we can start working just on Cluster

write.csv(db_spread_comb,"db_spread_comb.csv")
```

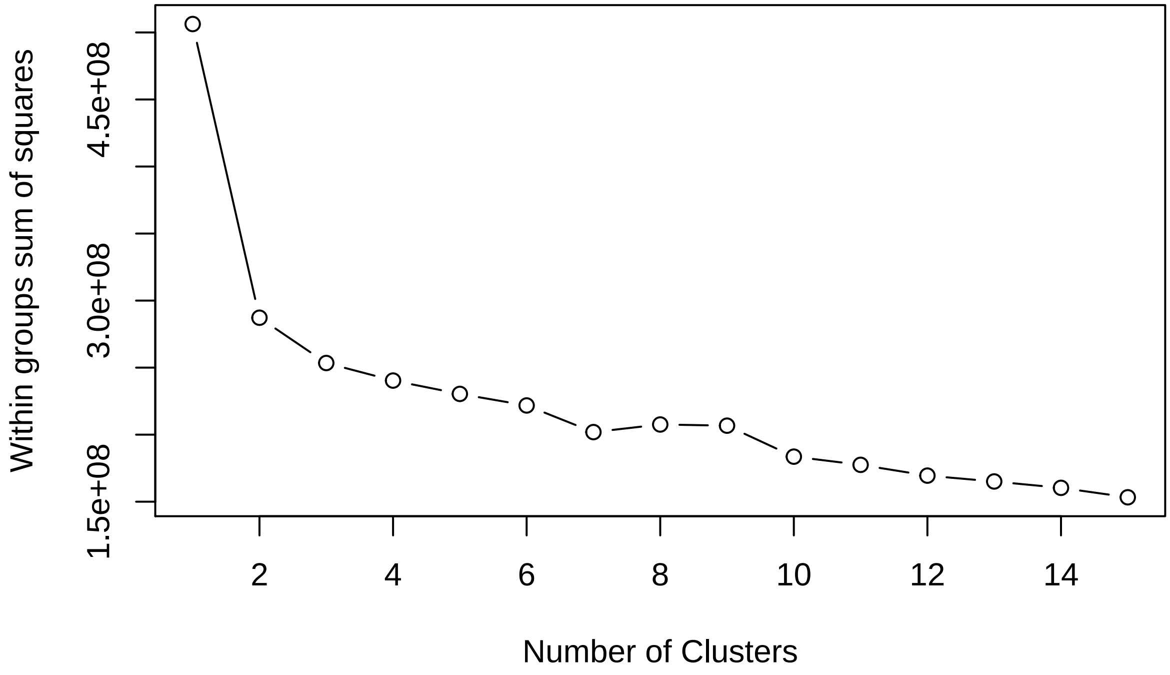
6.1 K-means Clustering

We will use the function we defined in week 3 to find the optimal number of clusters.

```
wssplot <- function(data, nc=15, seed=1234){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
     set.seed(seed)
     wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
  ylab="Within groups sum of squares")}
```

We cluster the data frame we just spread.

```
db_spread_comb <- read.csv("db_spreaded_comb.csv", stringsAsFactors = FALSE)

dtf <- db_spread_comb[, -1] # drop X

# First determine number of clusters

wssplot(dtf[,-1]) # drop cocktail name
```



```
# Perform K-means clustering with 6 possible groups, add clusters as additional column
set.seed(20)
```

```
km.out <- kmeans (dtf[,-1], 6, nstart=20, iter.max=50)
dclust<-data.frame(km.out$cluster,db_spread_comb$cocktail.name)

head(dclust)
```

```
  km.out.cluster         db_spread_comb.cocktail.name
1              6 '57 Chevy with a White License Plate
2              6                \xd6xn\xe4s Temptation
3              6                      110 in the shade
4              6                           155 Belmont
5              6                         24k nightmare
6              6                                   252
```

Explore the cluster results.

7 Inner products

Declare a function to calculate inner product between rows (cocktails) and represent it in degrees.

```
angle <- function(x,y){
  dot.prod <- x%*%y
  norm.x <- norm(x,type="2")
  norm.y <- norm(y,type="2")
  theta <- acos(dot.prod / (norm.x * norm.y))
  as.numeric(theta/3.14*180)
}


# Prepare ingredient term matrix
dtf <- sapply(dtf[,-1], as.numeric)

db_spread_comb$cocktail.name[100]
```

```
[1] "Amer Picon Punch"
```

```
db_spread_comb$cocktail.name[220]
```

```
[1] "Barcardi Volcano"
```

```
# Test if angle function works, answer is in degrees
angle(dtf[100,],dtf[220,])
```

```
[1] 90.04565
```

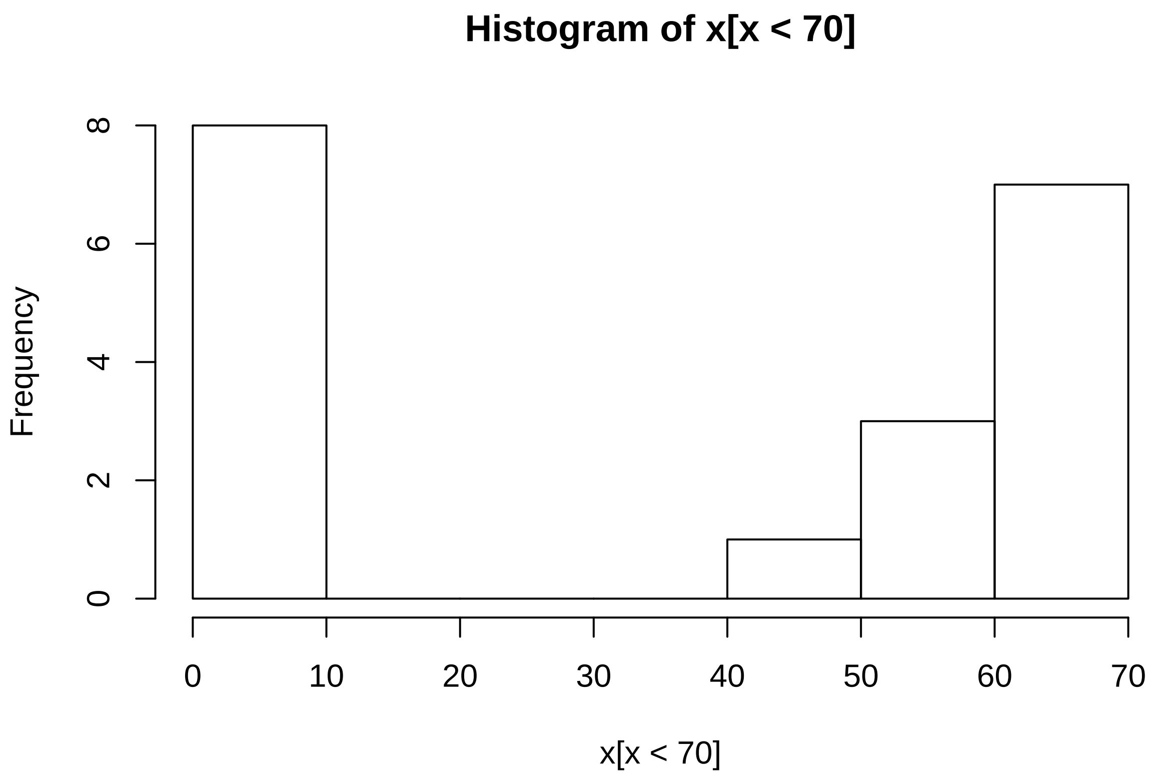
```
# Create dummy inner product matrix. We will take a small sample of whole data for demonstration
sumi<-matrix(nrow=100,ncol=100)

# Apply angle function on ingredient term matrix (without cocktail names), row by row

#for (i in 1:100){
#   for (j in 1:100){
#   sumi[i,j]<-angle(dtf[i,],dtf[j,])
#   print(i)
#   print(j)
#   }
```

[redacted]

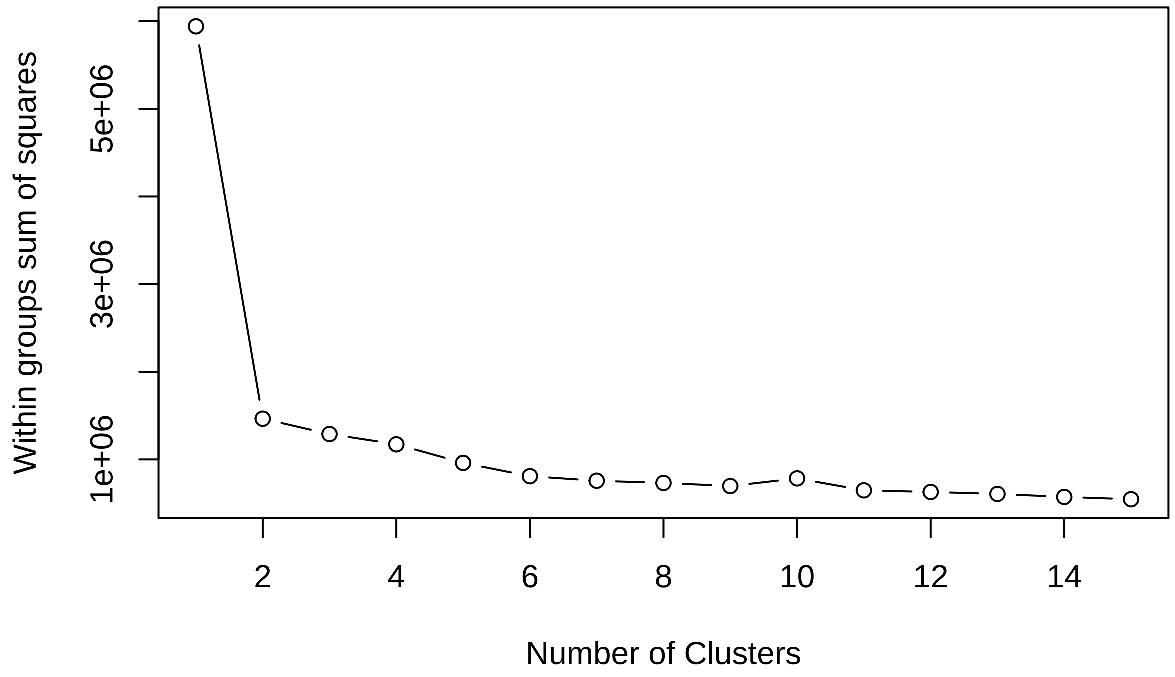
```
#}


# Replace NAs with 0
#sumi[is.na(sumi)]<-0

# Get histogram
#hist(sumi)

# Convert to dataframe set rows and columns names to the names of cocktails
#sumidf<-as.data.frame(sumi)
#colnames(sumidf)<-db_spread_comb$cocktail.name[1:100]
#rownames(sumidf)<-colnames(sumidf)

# Replace NAs with 0s
#sumidf[is.na(sumidf)]<-0

# Save as db_innerproduct_matrix.csv
#write.csv(sumi,"cocktailz/db_innerproduct_matrix.csv")
#library(readr)
#write_csv(sumidf,"db_innerproduct_matrix_100.csv")
```

7.1 Explore inner products matrix

```
# Load Inner Products Matrix, set rownames the same as column names
dfsumi<-read.csv("db_innerproduct_matrix_100.csv", header = TRUE, stringsAsFactors = FALSE)

colnames(dfsumi) <- db_spread_comb$cocktail.name[1:100]
rownames(dfsumi)<-colnames(dfsumi)

# set NAs to zero
dfsumi[is.na(dfsumi)] <- 0

# Choose acapulco
x<-dfsumi %>%
   select(Acapulco)
rownames(x)<-colnames(dfsumi)
x<-sapply(x, as.numeric)

y<-colnames(dfsumi)
hist(x[x<70])
```



```
# Perform unsupervised kmeans clustering on inner products matrix

wssplot(dfsumi)
```



```
km.out.ip <- kmeans(dfsumi, 6, nstart =20,iter.max=50)


# previous cluster
```

[redacted]

```
ip.x <-as.data.frame(km.out.ip$cluster)

ip.x <- bind_cols(ip.x, as.data.frame(rownames(ip.x)))


# lets compare the clusters.
head(ip.x)
```

```
  km.out.ip$cluster                       rownames(ip.x)
1                 2 '57 Chevy with a White License Plate
2                 4               \xd6xn\xe4s Temptation
3                 4                     110 in the shade
4                 1                          155 Belmont
5                 4                        24k nightmare
6                 4                                  252
```

```
head(dclust)
```

```
  km.out.cluster         db_spread_comb.cocktail.name
1              6 '57 Chevy with a White License Plate
2              6               \xd6xn\xe4s Temptation
3              6                     110 in the shade
4              6                          155 Belmont
5              6                        24k nightmare
6              6                                  252
```

7.1.1 CURRENT DOCUMENT ENDS HERE