Práctica 2: Limpieza y validación de los datos

Paul Zambrano

5 January, 2019

Contents

Introducción

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

El objetivo de esta actividad será el tratamiento de un dataset obtenido desde Kaggle. La información completa del dataset puede obtenerse desde el siguiente enlance: https://www.kaggle.com/ramkumarr02/deodorant-instant-liking-data

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset seleccionado se compone de resultados de encuestas realizadas a usuarios de desodorantes. En las encuestas se realizó varias preguntas a los usuarios donde clasificaban las características de un desodorante y

al final respondían si el desodorante les gustó inmediatamente o no. El problema presentado por este conjunto de dato es determinar si un desodorante será o no instantáneamente del agrado del usuario. El problema planteado por el dueño del dataset es crear algún modelo de predicción con un conjunto de entrenamiento donde se tiene las encuestas de 5 desodorantes. El conjunto de datos posee 64 variables y 2500 registros. No todas las variables tienen una descripción, aquí se enumeran las que si tienen:

1. **Respondent.ID** – Un identificador del usuario que respondió la encuesta
2. **Product.ID** – Identificador del desodorante
3. **Product** – Nombre del desodorante
4. **Instant.Liking** – Variable dependiente. 1 si es instantáneamente del agrado o 0 en caso opuesto

Las siguientes variables corresponden a las respuestas a las preguntas de la encuesta. Las respuestas son valores numéricos que miden que tan de acuerdo están con la pregunta. La mayoría tiene valores del 1 al 5. Por ejemplo, la pregunta 4.9 pregunta al usuario que defina en una escala del 1 al 5 qué tan elegante es el desodorante.

5. q1_1.personal.opinion.of.this.Deodorant
6. q2_all.words
7. q3_1.strength.of.the.Deodorant
8. q4_1.artificial.chemical
9. q4_2.attractive
10. q4_3.bold
11. q4_4.boring
12. q4_5.casual
13. q4_6.cheap
14. q4_7.clean
15. q4_8.easy.to.wear
16. q4_9.elegant
17. q4_10.feminine
18. q4_11.for.someone.like.me
19. q4_12.heavy
20. q4_13.high.quality
21. q4_14.long.lasting
22. q4_15.masculine
23. q4_16.memorable
24. q4_17.natural
25. q4_18.old.fashioned
26. q4_19.ordinary
27. q4_20.overpowering
28. q4_21.sharp
29. q4_22.sophisticated
30. q4_23.upscale
31. q4_24.well.rounded
32. q5_1.Deodorant.is.addictive
33. q7
34. q8.1
35. q8.2
36. q8.5
37. q8.6
38. q8.7
39. q8.8
40. q8.9
41. q8.10

42. q8.11
43. q8.12
44. q8.13
45. q8.17
46. q8.18
47. q8.19
48. q8.20
49. q9.how.likely.would.you.be.to.purchase.this.Deodorant
50. q10.prefer.this.Deodorant.or.your.usual.Deodorant
51. Q13_Liking.after.30.minutes
52. q14.Deodorant.overall.on.a.scale.from.1.to.10
53. s13.2
54. s13a.b.most.often
55. s13b.bottles.of.Deodorant.do.you.currently.own
56. ValSegb

Las siguientes variables también están en formato numérico, pero representan categorías como el estado civil o la educación.

57. q11.time.of.day.would.this.Deodorant.be.appropriate
58. q12.which.occasions.would.this.Deodorant.be.appropriate
59. s7.involved.in.the.selection.of.the.cosmetic.products
60. s8.ethnic.background
61. s9.education
62. s10.income
63. s11.marital.status
64. s12.working.status

2. Integración y selección de los datos de interés a analizar.

El fichero 'Data_train_reduced.csv' se lo puede bajar del sitio de Kaggle para proceder con el análisis. Lo primero será leer el fichero desde R y vamos a descartar las variables que no tienen una descripción.

```
deodorants <- read.csv("Data_train_reduced.csv")
deodorants[,33:48] <- NULL
deodorants$ValSegb <- NULL
deodorants$s13.2 <- NULL
deodorants$s13a.b.most.often <- NULL

#renombrar columnas muy extensas
colnames(deodorants)[colnames(deodorants)==
    "s7.involved.in.the.selection.of.the.cosmetic.products"] <-
      "s7.involved"

colnames(deodorants)[colnames(deodorants)==
    "q1_1.personal.opinion.of.this.Deodorant"] <-
      "q1.personal.opinion"

colnames(deodorants)[colnames(deodorants)==
    "q9.how.likely.would.you.be.to.purchase.this.Deodorant"] <-
      "q9.likely.purchase.deodorant"
```

```r
colnames(deodorants)[colnames(deodorants)==
    "q10.prefer.this.Deodorant.or.your.usual.Deodorant"] <-
      "q10.prefer.this.Deodorant"

colnames(deodorants)[colnames(deodorants)==
    "q11.time.of.day.would.this.Deodorant.be.appropriate"] <-
      "q11.time.of.day"

colnames(deodorants)[colnames(deodorants)==
    "q12.which.occasions.would.this.Deodorant.be.appropriate"] <-
      "q12.occasions.appropriate"

colnames(deodorants)[colnames(deodorants)==
    "q14.Deodorant.overall.on.a.scale.from.1.to.10"] <-
      "q14.scale.1.to.10"

colnames(deodorants)[colnames(deodorants)==
    "s13b.bottles.of.Deodorant.do.you.currently.own"] <-
      "s13b.bottles.owned"

str(deodorants)
```

```
## 'data.frame':    2500 obs. of  45 variables:
##  $ Respondent.ID                : int  3800 3801 3802 3803 3804 3805 3806 3807 3808 3809 ...
##  $ Product.ID                   : int  121 121 121 121 121 121 121 121 121 121 ...
##  $ Product                      : Factor w/ 5 levels "Deodorant B",..: 1 1 1 1 1 1 1 1 1 ...
##  $ Instant.Liking               : int  1 0 0 1 1 0 0 0 0 1 ...
##  $ q1.personal.opinion          : int  4 5 6 4 4 5 7 5 6 4 ...
##  $ q2_all.words                 : int  1 1 1 0 1 1 0 2 0 1 ...
##  $ q3_1.strength.of.the.Deodorant: int  4 4 3 4 2 5 2 3 3 3 ...
##  $ q4_1.artificial.chemical     : int  2 4 2 5 1 5 3 1 2 2 ...
##  $ q4_2.attractive              : int  5 2 5 5 3 5 2 4 5 5 ...
##  $ q4_3.bold                    : int  4 2 2 4 1 2 3 3 3 4 ...
##  $ q4_4.boring                  : int  2 1 4 3 1 1 2 1 4 1 ...
##  $ q4_5.casual                  : int  3 3 2 5 3 4 4 1 5 2 ...
##  $ q4_6.cheap                   : int  5 2 4 2 3 5 2 1 3 1 ...
##  $ q4_7.clean                   : int  5 4 3 5 5 5 4 5 3 4 ...
##  $ q4_8.easy.to.wear            : int  5 4 5 3 3 2 4 2 4 3 ...
##  $ q4_9.elegant                 : int  4 4 4 5 5 4 5 3 2 5 ...
##  $ q4_10.feminine               : int  5 3 4 5 5 4 3 4 1 5 ...
##  $ q4_11.for.someone.like.me    : int  3 1 4 5 5 3 5 1 5 3 ...
##  $ q4_12.heavy                  : int  1 1 3 1 1 2 3 2 1 3 ...
##  $ q4_13.high.quality           : int  5 3 1 4 4 4 3 3 5 1 ...
##  $ q4_14.long.lasting           : int  1 4 2 3 4 5 4 5 5 4 ...
##  $ q4_15.masculine              : int  2 4 1 3 2 2 2 2 1 3 ...
##  $ q4_16.memorable              : int  4 5 4 5 3 3 3 5 5 4 ...
##  $ q4_17.natural                : int  5 3 2 5 5 4 4 5 1 5 ...
##  $ q4_18.old.fashioned          : int  4 3 4 4 1 2 1 1 2 4 ...
##  $ q4_19.ordinary               : int  5 4 3 2 2 2 4 3 2 4 ...
##  $ q4_20.overpowering           : int  1 2 2 5 4 3 3 5 3 1 ...
##  $ q4_21.sharp                  : int  1 2 5 3 2 1 1 3 1 1 ...
##  $ q4_22.sophisticated          : int  4 5 4 3 3 5 5 4 4 3 ...
##  $ q4_23.upscale                : int  1 4 4 5 1 5 3 4 3 3 ...
```

```
##  $ q4_24.well.rounded            : int  4 4 3 4 5 4 5 3 5 4 ...
##  $ q5_1.Deodorant.is.addictive   : int  1 4 4 4 3 1 1 3 3 2 ...
##  $ q9.likely.purchase.deodorant  : int  2 3 5 5 5 5 5 2 4 5 ...
##  $ q10.prefer.this.Deodorant     : int  1 5 1 4 3 4 4 1 4 3 ...
##  $ q11.time.of.day               : int  1 3 3 1 3 2 3 2 3 1 ...
##  $ q12.occasions.appropriate     : int  2 3 3 3 2 2 3 1 1 2 ...
##  $ Q13_Liking.after.30.minutes   : int  1 3 2 6 5 6 6 6 7 2 ...
##  $ q14.scale.1.to.10             : int  7 8 5 8 4 7 4 5 6 7 ...
##  $ s7.involved                   : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ s8.ethnic.background          : int  1 1 1 1 1 1 1 2 1 3 ...
##  $ s9.education                  : int  4 4 3 4 3 4 3 3 4 4 ...
##  $ s10.income                    : int  3 3 5 9 5 5 2 7 5 6 ...
##  $ s11.marital.status            : int  1 1 1 1 1 2 2 1 3 2 ...
##  $ s12.working.status            : int  1 1 1 3 2 2 6 2 1 1 ...
##  $ s13b.bottles.owned            : int  3 4 2 3 3 2 5 3 4 2 ...
```

Todas las variables son de tipo numérico excepto el nombre del producto. Veremos los primeros diez elementos con unas pocas columnas de ejemplo

```
deodorants[1:10,c(3,4,9:13) ]
```

```
##        Product Instant.Liking q4_2.attractive q4_3.bold q4_4.boring
## 1  Deodorant B              1               5         4           2
## 2  Deodorant B              0               2         2           1
## 3  Deodorant B              0               5         2           4
## 4  Deodorant B              1               5         4           3
## 5  Deodorant B              1               3         1           1
## 6  Deodorant B              0               5         2           1
## 7  Deodorant B              0               2         3           2
## 8  Deodorant B              0               4         3           1
## 9  Deodorant B              0               5         3           4
## 10 Deodorant B              1               5         4           1
##    q4_5.casual q4_6.cheap
## 1            3          5
## 2            3          2
## 3            2          4
## 4            5          2
## 5            3          3
## 6            4          5
## 7            4          2
## 8            1          1
## 9            5          3
## 10           2          1
```

Para la resolución del problema seleccionaremos los datos correspondientes a la encuesta más la variable dependiente "Instant Liking". Los datos de identificadores no son necesarios.

```
deodorants$Respondent.ID <- NULL
deodorants$Product.ID <- NULL
```

Nos quedan 43 variables de las cuales 42 son numéricas. Veremos un resumen de cada una las variables numéricas

```r
fiveNums <- t(sapply(deodorants[,-1], fivenum))
colnames(fiveNums) <- c("Min","1.C","Med","3.C","Max")
fiveNums
```

```
##                                Min 1.C Med 3.C Max
## Instant.Liking                   0   0   0   0   1
## q1.personal.opinion              1   5   5   6   7
## q2_all.words                     0   0   1   2   5
## q3_1.strength.of.the.Deodorant   1   3   3   4   5
## q4_1.artificial.chemical         1   1   2   4   5
## q4_2.attractive                  1   3   4   5   5
## q4_3.bold                        1   3   4   5   5
## q4_4.boring                      1   1   2   3   5
## q4_5.casual                      1   3   4   5   5
## q4_6.cheap                       1   1   2   3   5
## q4_7.clean                       1   3   4   5   5
## q4_8.easy.to.wear                1   3   4   5   5
## q4_9.elegant                     1   3   4   5   5
## q4_10.feminine                   1   4   4   5   5
## q4_11.for.someone.like.me        1   3   4   5   5
## q4_12.heavy                      1   1   3   4   5
## q4_13.high.quality               1   3   4   5   5
## q4_14.long.lasting               1   3   4   5   5
## q4_15.masculine                  1   1   2   3   5
## q4_16.memorable                  1   3   4   5   5
## q4_17.natural                    1   3   4   5   5
## q4_18.old.fashioned              1   1   2   4   5
## q4_19.ordinary                   1   2   3   4   5
## q4_20.overpowering               1   2   3   4   5
## q4_21.sharp                      1   2   4   4   5
## q4_22.sophisticated              1   3   4   5   5
## q4_23.upscale                    1   3   4   5   5
## q4_24.well.rounded               1   3   4   5   5
## q5_1.Deodorant.is.addictive      1   2   4   4   5
## q9.likely.purchase.deodorant     1   3   4   5   5
## q10.prefer.this.Deodorant        1   2   3   4   5
## q11.time.of.day                  1   2   3   3   3
## q12.occasions.appropriate        1   1   2   3   3
## Q13_Liking.after.30.minutes      1   4   5   6   7
## q14.scale.1.to.10                1   5   7   9  10
## s7.involved                      4   4   4   4   4
## s8.ethnic.background             1   1   1   2   5
## s9.education                     2   2   3   4   7
## s10.income                       2   3   4   7  10
## s11.marital.status               1   1   2   2   5
## s12.working.status               1   1   1   2   7
## s13b.bottles.owned               1   2   3   4   6
```

Vemos que la variable s7 tiene un solo valor lo que no aporta en nada para un modelo predictivo. Eliminamos esa variable

```
deodorants$s7.involved <- NULL
```

Aunque las variables categóricas están en formato numérico no tiene sentido tratarlas como tal, los convertiremos en factores

```
deodorants$q11.time.of.day <-
  as.factor(deodorants$q11.time.of.day)

deodorants$q12.occasions.appropriate <-
  as.factor(deodorants$q12.occasions.appropriate)

deodorants$s8.ethnic.background <- as.factor(deodorants$s8.ethnic.background)
deodorants$s9.education <- as.factor(deodorants$s9.education)
deodorants$s10.income <- as.factor(deodorants$s10.income)
deodorants$s11.marital.status <- as.factor(deodorants$s11.marital.status)
deodorants$s12.working.status <- as.factor(deodorants$s12.working.status)
```

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Veremos si en los datos existen valores vacíos representados con NA

```
sum(is.na(deodorants))
```

```
## [1] 0
```

Por el resumen de los cuartiles que vimos antes, sabemos que la variable dependiente tiene valores de 0 que es normal ya que el cero representa que el desodorante no fue instantáneamente del agrado del usuario. La otra variable que contiene ceros es la pregunta 2 ("q2_all.words"), calculamos el número de valores con 0:

```
nrow(deodorants[deodorants$q2_all.words == 0,])
```

```
## [1] 709
```

La variable q2_all.words tiene la misma escala de valores de 1 al 5 sin embargo, tiene 709 valores con 0. Este valor de 709 es elevado porque solo tenemos 2500 observaciones entonces no sería conveniente imputarlos. Descartaremos esta variable antes de calcular el modelo.

```
deodorants$q2_all.words <- NULL
```

3.2. Identificación y tratamiento de valores extremos.

En este ejercicio de los desodorantes donde existe un rango de valores definido para cada variable no existe el problema de los valores extremos. Realizaremos el cálculo de los que son considerados valores atípicos para todas las variables independientes solo como demonstración, sin embargo, todos los valores serán aceptados ya que ser considerado atípico en este ejercicio solo denota que pocas personas utilizaron algunos valores en la escala.

```r
#Leer las variables numéricas excepto la variable dependiente
#en segunda posición y las variables de tipo factor.
for (i in 3:ncol(deodorants)){
  if( is.factor(deodorants[,i]) == FALSE ) {
    res <- boxplot.stats(deodorants[,i])$out

    if(length(res) > 0){
      print(colnames(deodorants)[i])
      print(res)
    }
  }
}
```

```
## [1] "q1.personal.opinion"
##   [1] 3 1 3 1 1 2 2 3 3 2 1 3 1 2 3 2 2 1 1 1 3 1 1 2 1 2 2 1 3 3 1 3 2 1 3
##  [36] 1 1 3 1 1 3 2 1 1 3 3 2 2 1 2 1 3 1 3 1 2 3 2 2 1 1 1 1 2 2 3 2 3 3 3
##  [71] 2 3 2 1 3 1 1 1 2 3 1 1 3 2 3 1 1 1 2 3 1 3 2 3 3 1 3 2 3 2 2 2 3 3 3
## [106] 3 2 1 2 2 1 3 3 2 1 1 2 2 2 2 2 3 1 2 3 2 1 3 1 2 1 2 3 3 3 3 2 2 3 1
## [141] 2 3 3 1 2 3 2 2 1 2 3 2 1 1 2 1 1 2 1 2 3 3 1 3 3 3 1 2 1 2 1 1 3 2 1
## [176] 3 1 2 3 1 2 3 1 1 2 1 3 2 3 3 3 3 2 1 2 3 3 3 2 3 1 2 1 3 1 1 1 1 1 2
## [211] 1 2 2 3 1 3 3 3 1 1 2 1 2 3 1 3 1 1 3 3 3 3 1 2 2 2 2 3 1 3 1 1 1 1 2
## [246] 1 2 1 1 3 2 3 1 2 3 2 1 3 3 2 1 3 2 3 2 3 1 2 2 1 2 2 3 1 1 3 2 1 3 3
## [281] 3 3 3 3 2 3 1 1 1 3 1 1 3 3 3 1 3 3 3 2 2 2 3
## [1] "q3_1.strength.of.the.Deodorant"
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1] "q4_10.feminine"
##   [1] 1 2 2 2 2 2 2 2 1 2 1 2 1 1 2 1 2 1 2 1 1 2 2 2 1 2 2 2 2 1 1 2 2 2 2
##  [36] 2 1 1 2 1 1 2 1 2 2 1 2 2 2 2 1 1 2 2 1 2 1 2 1 1 1 1 1 2 2 2 1 2 1 2
##  [71] 2 1 2 2 2 2 2 2 1 2 1 2 1 2 2 1 2 1 2 2 1 2 1 2 1 1 2 2 1 1 2 1 2 2 1
## [106] 2 2 1 2 1 1 1 1 1 1 1 2 2 1 2 1 1 2 2 2 1 2 2 1 1 1 2 2 1 2 1 1 2 1 2
## [141] 1 1 1 2 2 2 1 1 2 1 2 1 1 2 2 1 2 1 2 1 2 2 1 2 1 1 1 2 2 2 2 1 2 1 2
## [176] 2 2 1 2 2 2 1 2 2 1 1 1 1 1 2 2 1 1 1 1 2 1 2 1 1 2 2 2 2 2 1 1 2 2 1
## [211] 2 1 2 2 1 2 1 1 1 2 2 1 2 2 2 2 1 1 1 2 1 2 2 2 1 1 2 2 2 1 1 1 1 1 1
## [246] 1 2 1 2 1 2 2 1 1 2 1 2 2 2 2 1 1 1 2 2 1 1 2 2 1 2 1 2 2 1 2 2 2 2 2
## [281] 2 2 2 1 1 1 1 1 2 1 1 1 2 1 2 2 1 1 1 2 2 1 2 1 2 1 1 2 2 1 2 2 1 2 2
## [316] 1 1 1 1 1 2 2 1 1 2 2 2 2 1 1 1 2 2 2 2 2 2 1 2 2 2 1 1
```

Existe 3 variables que contienen valores que fueron escasamente seleccionados por los usuarios, pero aún así son valores totalmente válidos.

```r
# Exportación de los datos en .csv
write.csv(deodorants, "Deodorants.csv", row.names = F)
```

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Para tener una idea de qué grupos son interesantes calcularemos las frecuencias de cada categoría de las variables de tipo factor respecto a la variable dependiente.

```r
factorData <- deodorants[,sapply(deodorants, is.factor)]

for(i in 1:ncol(factorData)){
  print(names(factorData)[i])
  print(table(deodorants$Instant.Liking, factorData[,i]))
  writeLines("\n")
}
```

```
## [1] "Product"
##
##     Deodorant B Deodorant F Deodorant G Deodorant H Deodorant J
##   0         373         373         378         387         371
##   1         127         127         122         113         129
##
##
## [1] "q11.time.of.day"
##
##        1    2    3
##   0  401  464 1017
##   1  137  165  316
##
##
## [1] "q12.occasions.appropriate"
##
##       1   2   3
##   0 544 431 907
##   1 161 172 285
##
##
## [1] "s8.ethnic.background"
##
##        1    2    3    4    5
##   0 1209  305  234   89   45
##   1  378  106   87   25   22
##
##
## [1] "s9.education"
##
##       2   3   4   5   6   7
##   0 521 634 578 119  27   3
##   1 174 214 186  34  10   0
##
##
## [1] "s10.income"
##
##       2   3   4   5   6   7   8   9  10
##   0 249 350 365 268 179 143 114  60 154
##   1  90 113 113  85  56  54  40  25  42
##
##
## [1] "s11.marital.status"
##
##       1   2   3   4   5
```

```
##   0 734 976 130  26  16
##   1 226 333  39  14   6
##
##
## [1] "s12.working.status"
##
##       1    2    3    4    5    6    7
##   0 1099  418  129   64   75   88    9
##   1  363  126   44   19   21   42    3
```

No parece haber una categoría que tenga un claro comportamiento diferente. Tomaremos el caso del Producto Deodorant H que parece tener menor agrado instantáneo comparado con los otros desodorantes. También intentaremos el caso de la categoría 2 de la pregunta 12 "q12.which.occasions.would.this.Deodorant.be.appropriate" que parece tener mayor agrado instantáneo que el resto. Probaremos las dos hipótesis en las siguientes secciones.

```
deodorants.productH <- deodorants[deodorants$Product == "Deodorant H",]
deodorants.notProductH <- deodorants[deodorants$Product != "Deodorant H",]


deodorants.q12_cat2 <- deodorants[
  deodorants$q12.occasions.appropriate == "2",]
deodorants.q12_notCat2 <- deodorants[
  deodorants$q12.occasions.appropriate != "2",]
```

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Todas nuestras variables, a pesar de ser numéricas, son valores discretos por lo que una prueba de que las variables pertenecen a una distribución normal no tiene lugar. Ejecutamos la prueba de Shapiro-Wilk como ejemplo a dos de nuestras variables y veremos que están lejos de ser normales.

```
shapiro.test(deodorants$q1.personal.opinion)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  deodorants$q1.personal.opinion
## W = 0.87083, p-value < 2.2e-16
```

El valor p es muy cercano a cero lo que indica un rechazo contundente a la hipótesis nula que indica que la variable pertenece a una distribución normal. Esto es porque nuestras variables no son continuas.

```
shapiro.test(deodorants$q4_8.easy.to.wear)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  deodorants$q4_8.easy.to.wear
## W = 0.82875, p-value < 2.2e-16
```

Lo mismo sucede con esta otra variable.

Las pruebas de homogeneidad de la varianza las realizaremos con los grupos pertenecientes a los dos factores que seleccionamos en la sección anterior.

```
fligner.test(deodorants$Instant.Liking, deodorants$Product)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  deodorants$Instant.Liking and deodorants$Product
## Fligner-Killeen:med chi-squared = 1.7962, df = 4, p-value = 0.7732
```

Para los 5 tipos de desodorante la prueba con un alto valor p indica que las varianzas son homogéneas.

```
fligner.test(deodorants$Instant.Liking,
             deodorants$q12.occasions.appropriate)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  deodorants$Instant.Liking and deodorants$q12.occasions.appropriate
## Fligner-Killeen:med chi-squared = 6.4507, df = 2, p-value =
## 0.03974
```

Para las 3 categorías de la pregunta 12 en cambio el valor p es bajo. Si la significancia de la prueba es de 0.05 entonces nos indica que las varianzas no son homogéneas. Por esta razón las pruebas estadísticas de comparación no las ejecutaremos para esta variable.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

Contraste de Hipótesis

Realizamos una prueba de contraste planteada en la sección 4.1. donde los resultados del desodorante H parecen tener menor agrado instantáneo comparado con los otros desodorantes

La hipótesis nula será que las medias de las dos muestras son iguales. La hipótesis alternativa será que la diferencia de las muestras es menor a cero

```
t.test(deodorants.productH$Instant.Liking, deodorants.notProductH$Instant.Liking, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  deodorants.productH$Instant.Liking and deodorants.notProductH$Instant.Liking
## t = -1.2563, df = 789.71, p-value = 0.1047
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf 0.008237735
## sample estimates:
## mean of x mean of y
##    0.2260    0.2525
```

El valor p es de 0.1047. Si tenemos un nivel de significación menor a 10% que sería apropiado tendríamos que aceptar la hipótesis nula, es decir que **no** se puede afirmar que el desodorante H es de menor agrado instanténeo que el resto.

Correlaciones

Para buscar correlaciones aplicaremos el método de Spearman que funciona para distribuciones que no son normales. Primero seleccionamos las variables numéricas y a cada una de ellas aplicamos el método junto con la variable dependiente que es Instant.Liking

```
numData <- deodorants[,sapply(deodorants, is.integer)]

#Aplicar a cada variable el método Spearman
spearmanCorr <- lapply(numData[,-1], cor.test, numData$Instant.Liking, method = "spearman")

#Construir tabla con resultados del valor p y el estimado
output <- data.frame(variable=colnames(numData[,-1]))
pVals <- c()
estimates <- c()
for(i in 1:length(spearmanCorr)){
  pVals[i] <- spearmanCorr[[i]]$p.value
  estimates[i] <- spearmanCorr[[i]]$estimate
}
output <- cbind(output, p.Value=pVals, estimate=estimates)

#Mostrar valores de variables que tienen valor p menor que 0.2
output[output$p.Value < 0.2,]
```

```
##                      variable     p.Value    estimate
## 1         q1.personal.opinion 0.000000000 -0.77120964
## 10          q4_8.easy.to.wear 0.173599318  0.02722315
## 15         q4_13.high.quality 0.111354311  0.03185056
## 18            q4_16.memorable 0.196213508  0.02585691
## 19              q4_17.natural 0.119751006  0.03112442
## 29 q10.prefer.this.Deodorant 0.001476025  0.06355462
```

Para un nivel de significación de 0.2 se ven 6 variables. Siendo más estrictos, existen en realidad 2 variables con una correlación importante: “q1.personal.opinion” y “q10.prefer.this.Deodorant.or.your.usual.Deodorant” La primera tiene una relación inversa y la segunda una relación directa.

Regresión Binomial

Trataremos un modelo de regresión logístico con todas las variables. Luego compararemos el modelo con otro donde solo utilizaremos las variables sugeridas por la correlación

```
glmModel1 <- glm( Instant.Liking~., family=binomial(link='logit'), numData)

summary(glmModel1)
```

```
##
## Call:
```

```
## glm(formula = Instant.Liking ~ ., family = binomial(link = "logit"),
##     data = numData)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.028e-05  -7.561e-06  -2.110e-08  -2.110e-08   1.310e-05
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     2.146e+02  5.416e+04   0.004    0.997
## q1.personal.opinion            -4.767e+01  5.993e+03  -0.008    0.994
## q3_1.strength.of.the.Deodorant  4.945e-02  3.803e+03   0.000    1.000
## q4_1.artificial.chemical        5.465e-03  2.156e+03   0.000    1.000
## q4_2.attractive                 8.701e-04  2.334e+03   0.000    1.000
## q4_3.bold                      -9.642e-03  2.454e+03   0.000    1.000
## q4_4.boring                    -1.672e-02  2.519e+03   0.000    1.000
## q4_5.casual                     1.475e-02  2.322e+03   0.000    1.000
## q4_6.cheap                     -1.309e-02  2.327e+03   0.000    1.000
## q4_7.clean                     -3.378e-03  2.516e+03   0.000    1.000
## q4_8.easy.to.wear              -1.145e-02  2.326e+03   0.000    1.000
## q4_9.elegant                   -7.822e-03  2.396e+03   0.000    1.000
## q4_10.feminine                 -6.542e-03  2.479e+03   0.000    1.000
## q4_11.for.someone.like.me      -5.847e-02  2.214e+03   0.000    1.000
## q4_12.heavy                    -2.071e-02  2.005e+03   0.000    1.000
## q4_13.high.quality              3.817e-02  2.518e+03   0.000    1.000
## q4_14.long.lasting              7.087e-03  2.807e+03   0.000    1.000
## q4_15.masculine                -2.979e-02  2.373e+03   0.000    1.000
## q4_16.memorable                -1.532e-02  2.568e+03   0.000    1.000
## q4_17.natural                   3.542e-02  2.283e+03   0.000    1.000
## q4_18.old.fashioned            -5.393e-03  2.126e+03   0.000    1.000
## q4_19.ordinary                 -1.575e-02  2.317e+03   0.000    1.000
## q4_20.overpowering             -2.291e-02  2.115e+03   0.000    1.000
## q4_21.sharp                    -3.341e-02  2.324e+03   0.000    1.000
## q4_22.sophisticated            -1.114e-02  2.355e+03   0.000    1.000
## q4_23.upscale                   5.035e-03  2.389e+03   0.000    1.000
## q4_24.well.rounded             -2.516e-02  2.485e+03   0.000    1.000
## q5_1.Deodorant.is.addictive     5.124e-03  2.327e+03   0.000    1.000
## q9.likely.purchase.deodorant    8.301e-03  2.184e+03   0.000    1.000
## q10.prefer.this.Deodorant       1.951e-02  2.152e+03   0.000    1.000
## Q13_Liking.after.30.minutes     3.909e-03  1.708e+03   0.000    1.000
## q14.scale.1.to.10              -2.827e-03  1.219e+03   0.000    1.000
## s13b.bottles.owned              1.239e-03  1.859e+03   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.7962e+03  on 2499  degrees of freedom
## Residual deviance: 8.4279e-08  on 2467  degrees of freedom
## AIC: 66
##
## Number of Fisher Scoring iterations: 25
```

Ahora calculamos un segundo modelo y veremos el valor del AIC. Mientras menor sea el valor de AIC mejor será el modelo

```r
glmModel2 <- glm( Instant.Liking ~ q1.personal.opinion +
                  q10.prefer.this.Deodorant, family=binomial(link='logit'), deodorants)

summary(glmModel2)
```

```
##
## Call:
## glm(formula = Instant.Liking ~ q1.personal.opinion + q10.prefer.this.Deodorant,
##     family = binomial(link = "logit"), data = deodorants)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -8.527e-06  -8.142e-06  -2.110e-08  -2.110e-08   1.084e-05
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.143e+02  2.839e+04   0.008    0.994
## q1.personal.opinion        -4.769e+01  5.988e+03  -0.008    0.994
## q10.prefer.this.Deodorant   2.312e-02  2.131e+03   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.7962e+03  on 2499  degrees of freedom
## Residual deviance: 8.4291e-08  on 2497  degrees of freedom
## AIC: 6
##
## Number of Fisher Scoring iterations: 25
```

El modelo con dos variables es mucho mejor. Incluso podríamos suponer que este problema se puede resolver solo con una variable

```r
glmModel3 <- glm( Instant.Liking ~ q1.personal.opinion,
                  family=binomial(link='logit'), deodorants)

summary(glmModel3)
```

```
##
## Call:
## glm(formula = Instant.Liking ~ q1.personal.opinion, family = binomial(link = "logit"),
##     data = deodorants)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -8.317e-06  -8.317e-06  -2.110e-08  -2.110e-08   1.060e-05
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)           214.36   27573.29   0.008    0.994
## q1.personal.opinion   -47.69    5985.62  -0.008    0.994
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 2.7962e+03  on 2499  degrees of freedom
## Residual deviance: 8.4306e-08  on 2498  degrees of freedom
## AIC: 4
##
## Number of Fisher Scoring iterations: 25
```

Efectivamente, este último modelo es mejor y solo tiene una variable.

5. Representación de los resultados a partir de tablas y gráficas.

Realizaremos las predicciones con este modelo tal de poder crear una matriz de confusión.

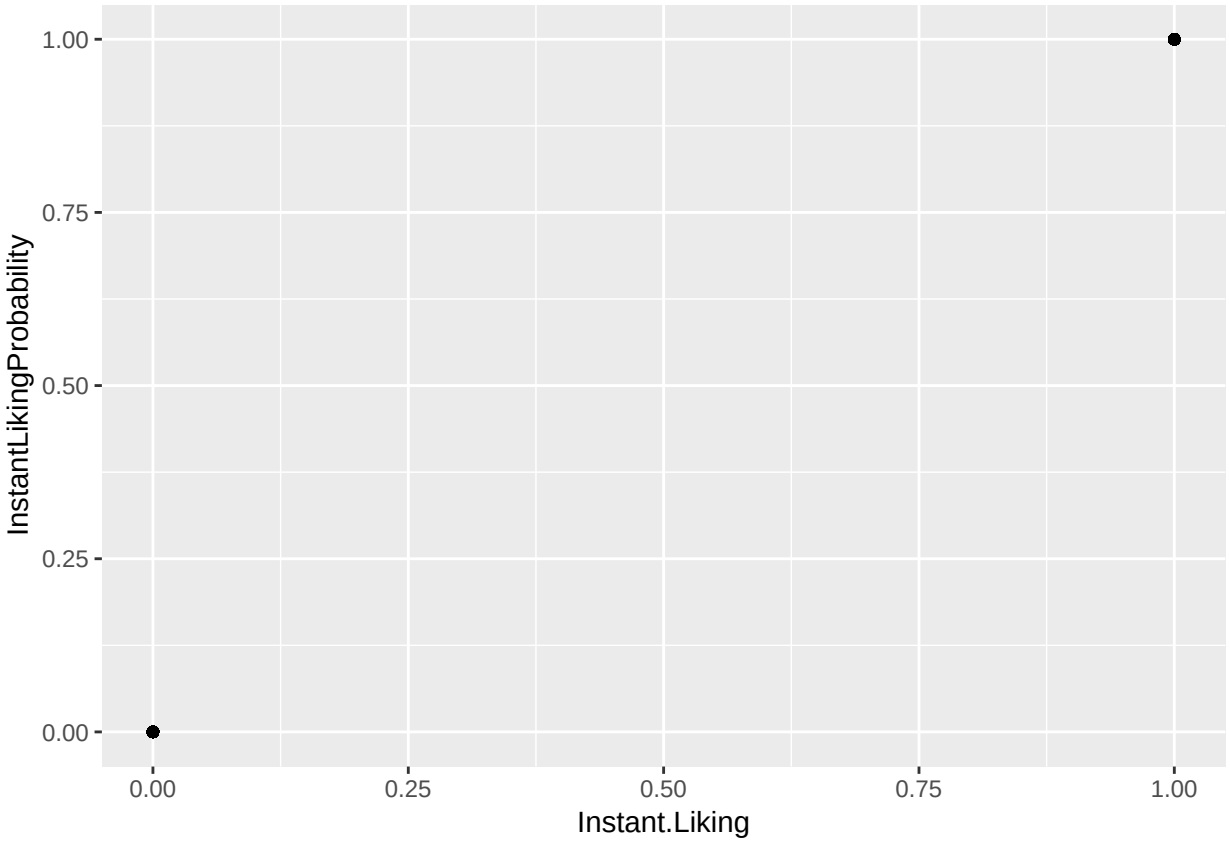
```
#realizamos las predicciones del mejor modelo
deodorants$InstantLikingProbability <- predict(glmModel3, type = 'response')

#matriz de confusion, tomando el valor 1 (Instant Liking) los valores mayor a 0.7
table(real=deodorants$Instant.Liking, "predicción"=
        ifelse(deodorants$InstantLikingProbability > 0.7, 1, 0))
```

```
##     predicción
## real    0    1
##    0 1882    0
##    1    0  618
```

En esta matriz de confusión hemos tomado un umbral de discriminación del 70% para determinar el resultado. No hay falsos positivos ni falsos negativos. Sin embargo, veremos en un gráfico que en realidad el umbral de discriminación no es importante ya que las predicciones son exactas a 0 y 1.

```
library(ggplot2)
ggplot(deodorants, aes(x=Instant.Liking, y=InstantLikingProbability)) + geom_point()
```



No hay valores de probabilidades entre 0 y 1 por eso solo se ven dos puntos en (0,0) y (1,1).

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Utilizando la información que obtuvimos de las correlaciones pudimos determinar que, aunque el conjunto de datos tenía un gran número de variables, estás no eran necesarias para resolver el problema planteado. Se quería un modelo capaz de predecir, en base a las respuestas de una encuesta a usuarios, si un desodorante sería de agrado instantáneo. El conjunto de datos requería un modelo de regresión binomial con función logística que ha resultado perfecto en la resolución del problema. Los resultados obtenidos han determinado que para tener una respuesta al problema únicamente se necesitaba la información de una pregunta de la encuesta que es la opinión personal del desodorante.