Universidad Mayor de San Andrés

Facultad de Ciencias Puras y Naturales - Club de Ciencia de Datos[1]

Análisis de Conglomerados, Encuesta Nacional de Opinión sobre TIC

Por: Ayar Yuman Paco Sanizo[2] y Marco Antonio Vino Chipana[3]

La Paz, Bolivia 22 de abril de 2018

1 https://bit.ly/NuevosMiembrosCCD
2 Carrera de Estadística
3 Carrera de Informática

Tabla de Contenidos

Análisis de Conglomerados, Encuesta Nacional de Opinión sobre TIC

1. Introducción

1.1 Descripción del proyecto

A partir del lanzamiento de la base de datos abiertos de la encuesta de nacional de opinión sobre TIC en Bolivia desarrollada por la Agencia de gobierno electrónico y tecnologías de la información (AGETIC), iniciamos un proceso de exploración de la base de datos a fin de identificar un producto adecuado. En este proceso se identificó una gran cantidad de variables con valores perdidos que dificultaría y afectaría la representatividad de cualquier análisis posterior. Por tanto, se planteó un procedimiento de imputación que dio como resultado una base de datos reducida respecto a las variables que mantiene el total de observaciones. Por otra parte, se identificó una gran cantidad de variables categóricas en contraste a la cantidad de variables numéricas. Finalizada la exploración, se encontró útil identificar conglomerados entre los encuestados respecto a la temática de las TIC.

Bajo esta línea inicialmente se trató de identificar variables que permitan discriminar conglomerados considerando teorías económicas. Sin embargo esta era una tarea laboriosa y en muchos casos no se identificaban diferencias marcadas entre los conglomerados. Por tanto decidimos considerar un problema de aprendizaje no supervisado tomando una gran cantidad de variables para que bajo una perspectiva multivariante se puedan identificar conglomerados que estén considerablemente separados. El modelo elegido para este fin fue K-Prototypes, una combinación de K-means y K-modes, que permite considerar variables categóricas y numéricas.

Una vez identificados los conglomerados resultaba importante seleccionar las variables más importantes en su definición. Es decir, encontrar las variables que permitan identificarlos con mayor facilidad. De esta forma se consideró un problema de aprendizaje supervisado para identificar las variables más importantes a la hora de predecir la pertenencia de las observaciones a los conglomerados. Para esta tarea se consideró el modelo Random Forest por la facilidad que tiene para identificar variables importantes. Luego, tomando estas variables se volvieron a identificar conglomerados para finalmente realizar un análisis descriptivo cruzado respecto a las variables que ingresaron al modelo final y a otras de interés para la interpretación de los conglomerados.

1.2 Objetivos

1.2.1 Objetivo General

- Identificar conglomerados entre las personas encuestadas en la Encuesta Nacional de Opinión sobre TIC y describir su comportamiento.

1.2.2 Objetivos Específicos

- Pre-procesar la base de datos de la Encuesta Nacional de Opinión sobre TIC para poder identificar conglomerados representativos entre los encuestados.
- Identificar conglomerados en la base de datos considerando variables categóricas y numéricas representativas.
- Identificar las variables más importantes que permitan una mejor identificación de los conglomerados.
- Realizar un análisis descriptivo de los conglomerados respecto a las variables que mejor los definen.
- Realizar un análisis descriptivo de los conglomerados respecto a otras variables de interés.
- Definir perfiles conceptuales de los conglomerados identificados.

1.3 Consideraciones

Es importante notar que el presente proyecto tiene como resultado un análisis descriptivo de la muestra respecto a los conglomerados identificados. Esto es, no se puede asegurar que los resultados reflejen a la población. Este tratamiento inferencial es una tarea posterior que requiere un análisis minucioso considerando el método de muestreo usado en la encuesta.

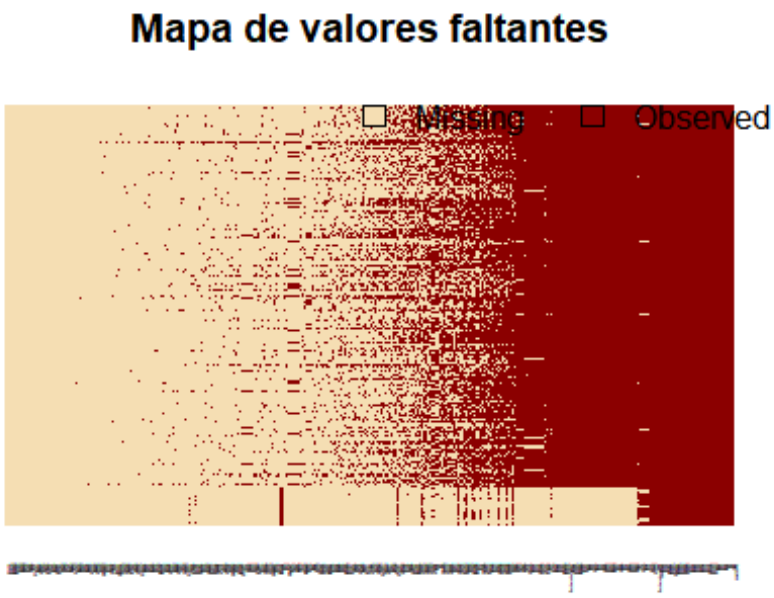
2. Pre-procesamiento

A continuación se describe la etapa de pre-procesamiento de la base de datos. Partimos describiendo el análisis de valores faltantes, luego el proceso de imputación y finalmente la pre-selección de variables para la etapa de modelación.

2.1 Análisis de valores faltantes

Empezando el análisis mapeamos los valores faltantes para tener una idea visual de la cantidad de datos perdidos en la encuesta. En esta primera etapa solo se consideran como valores faltantes a las celdas vacías. Esto es, aún no se consideran como faltantes a las celdas con codificaciones correspondientes a la etiqueta "No sabe/No responde". Este tratamiento se efectuará posteriormente.

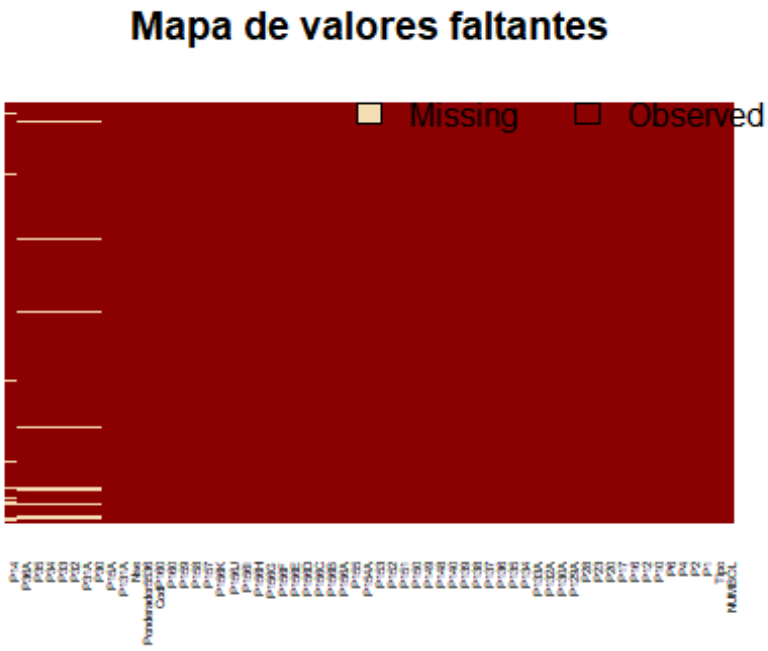
En el siguiente gráfico, en el eje vertical tenemos filas y en horizontal columnas de la base de datos. Las celdas con observaciones tienen un color guindo mientras que las celdas vacías tienen color beige. A partir de este mapeo se identifica que de las 451 variables, la gran mayoría tiene bastantes datos faltantes.



A continuación ordenamos las variables de la base de datos respecto a su porcentaje de valores perdidos y consideramos únicamente las variables que tienen hasta un 5% de valores faltantes como máximo. Las variables que pasan este filtro son:

```
##   [1] "NUMBOL"            "Tipo"              "P1"                "P2"
##   [5] "P4"                "P6"                "P10"               "P12"
##   [9] "P14"               "P15A"              "P16"               "P17"
##  [13] "P20"               "P23"               "P28"               "P30"
##  [17] "P31A"              "P32"               "P33"               "P34"
##  [21] "P35"               "P36A"              "P129A"             "P130A"
##  [25] "P131A"             "P132A"             "P133A"             "P134"
##  [29] "P135"              "P136"              "P137"              "P138"
##  [33] "P139"              "P140"              "P148"              "P149"
##  [37] "P150"              "P151"              "P152"              "P153"
##  [41] "P154A"             "P155"              "P156A"             "P156B"
##  [45] "P156C"             "P156D"             "P156E"             "P156F"
##  [49] "P156G"             "P156H"             "P156I"             "P156J"
##  [53] "P156K"             "P157"              "P158"              "P159"
##  [57] "P160"              "CodP160"           "Ponderador5536"    "Nse"
```

Luego filtrando la base de datos con estas 60 variables y volviendo a mapear los valores faltantes obtenemos el siguiente mapeo de valores faltantes.



2.2 Definición de tipos de variables y codificación de valores perdidos

Ahora, si bien al parecer la base de datos resultante tiene pocos datos faltantes, debemos considerar que varias variables tienen codificaciones para la etiqueta "No sabe/No responde". Por tanto, resulta importante no solo definir si una variable es categórica o numérica, sino también codificar adecuadamente estas etiquetas como valores perdidos.

Luego de efectuar este procedimiento, obtenemos el siguiente resumen de variables y el mapeo de valores faltantes actualizado.

```
##      NUMBOL       Tipo           P1        P2       P4       P6
##  Min.   :    1   1:5033   Min.   :14.00   1:5066   1:1098   1:2736
##  1st Qu.: 2023   2: 503   1st Qu.:21.00   2: 470   2:4438   2:2800
##  Median : 9066            Median :30.00
##  Mean   : 7856            Mean   :32.62
##  3rd Qu.:11298            3rd Qu.:42.00
##  Max.   :17514            Max.   :90.00
##
##       P10        P12            P14           P15A      P16      P17
##  9      :2455   1:5297   Min.   :   70   1    :3883   1:2277   1: 625
##  7      : 667   2: 239   1st Qu.: 1200   2    : 500   2:3259   2:4911
##  8      : 467            Median : 1800   3    :  98
##  3      : 444            Mean   : 2083   4    : 811
##  2      : 440            3rd Qu.: 2650   5    :  20
##  5      : 321            Max.   :21000   6    : 222
##  (Other): 742            NA's   :2276    NA's:   2
##
##  P20      P23      P28           P30            P31A        P32
##  1:4501   1: 572   1:5345   Min.   :  10   1      :5019   1   :4861
##  2:1035   2: 911   2: 191   1st Qu.: 600   2      :  97   2   : 436
##           3:4053            Median : 960   3      :  91   3   :  48
##                             Mean   :1117   6      :  59   NA's: 191
##                             3rd Qu.:1500   4      :  29
##                             Max.   :8001   (Other):  50
##                             NA's   :937    NA's   : 191
##
##    P33            P34            P35         P36A          P129A
##  1   :1820   Min.   :  1.00   1   : 127   1   :4351   1      :3951
##  2   :2650   1st Qu.: 35.00   2   : 333   2   : 169   2      : 468
##  3   : 875   Median : 70.00   3   :1479   3   :  60   6      : 450
##  NA's: 191   Mean   : 90.41   4   :2117   4   : 161   5      : 231
##              3rd Qu.:120.00   5   :1289   NA's: 795   3      : 192
##              Max.   :900.00   NA's: 191               (Other): 152
##              NA's   :249                              NA's   :  92
##
```

```
##      P130A          P131A          P132A          P133A
##  1      :3272   1      :2279   1      :3008   1      :1681
##  6      : 653   5      : 972   6      : 823   3      : 860
##  2      : 505   6      : 687   5      : 434   2      : 623
##  5      : 490   2      : 356   2      : 331   6      : 449
##  3      : 244   3      : 250   3      : 198   5      : 381
##  (Other): 178   (Other): 307   (Other): 226   (Other): 255
##  NA's   : 194   NA's   : 685   NA's   : 516   NA's   :1287
##
##      P134           P135           P136           P137
##  1      :2697   1      :2550   1      :2536   1      :2654
##  6      :1189   2      : 901   3      :1054   2      : 815
##  2      : 617   3      : 575   2      : 862   3      : 761
##  5      : 608   7      : 356   7      : 247   5      : 329
##  4      : 142   5      : 291   5      : 221   7      : 245
##  (Other): 156   (Other): 429   (Other): 294   (Other): 370
##  NA's   : 127   NA's   : 434   NA's   : 322   NA's   : 362
##
##      P138           P139           P140      P148          P149
##  1      :1862   1      :2202   6      :1566   1:2774   3      :1409
##  2      : 964   2      : 737   1      :1512   2:2762   2      :1231
##  3      : 645   3      : 581   2      : 990            6      :1061
##  7      : 594   5      : 410   5      : 345            7      : 625
##  6      : 371   7      : 383   7      : 272            4      : 611
##  (Other): 478   (Other): 556   (Other): 257            1      : 392
##  NA's   : 622   NA's   : 667   NA's   : 594            (Other): 207
##
##  P150     P151          P152           P153      P154A         P155
##  1:2726   1:4959   ES     :1539   3      :1411   1:3733   2      :2198
##  2:2360   2: 143   AS     :1160   1      : 931   2:1256   1      :1353
##  3: 298   3: 415   CP     : 850   4      : 848   3:  88   3      :1197
##  4: 152   4:   5   NP     : 781   7      : 811   4: 221   4      : 465
##           5:   9   AC     : 623   2      : 793   5: 238   5      : 215
##           6:   5   PI     : 232   6      : 373            6      :  85
##                    (Other): 351   (Other): 369            (Other):  23
##
##  P156A    P156B    P156C    P156D    P156E    P156F    P156G    P156H
##  1:5183   1:4046   1:5345   1:1740   1:1195   1:2244   1: 863   1:3835
##  2: 353   2:1490   2: 191   2:3796   2:4341   2:3292   2:4673   2:1701
##
##      P156I      P156J    P156K          P157           P158      P159
##  1      :3604   1: 730   1: 267   4      :1416   4      : 741   1:2824
##  2      :1345   2:4806   2:5269   5      :1197   1      : 663   2:1306
##  3      : 296                    3      : 983   6      : 661   3:1406
##  0      : 129                    6      : 735   9      : 647
##  4      : 108                    2      : 390   2      : 615
##  5      :  39                    7      : 356   3      : 598
##  (Other):  15                    (Other): 459   (Other):1611
##
```

```
##          P160         CodP160     Ponderador5536    Nse
##  Cobija    : 422   16     : 422   Min.   :0.04898   1:1533
##  Santa Cruz: 420   70     : 420   1st Qu.:0.33448   2:1738
##  Sucre     : 360   78     : 360   Median :0.41612   3:1351
##  Oruro     : 308   44     : 308   Mean   :0.97589   4: 611
##  Cochabamba: 251   17     : 251   3rd Qu.:1.52909   5: 303
##  El Alto   : 244   24     : 244   Max.   :7.26183
##  (Other)   :3531   (Other):3531
```

2.3 Imputación de datos faltantes

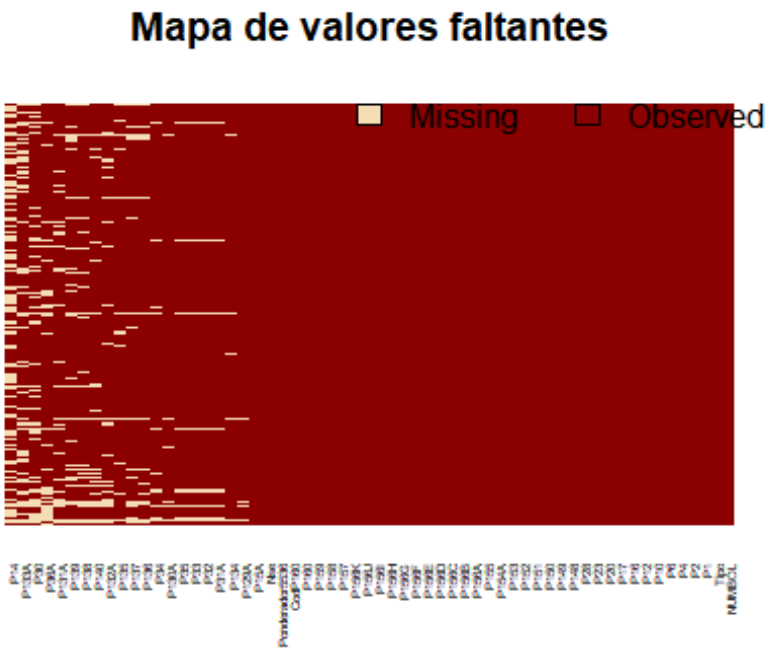
A partir de la base de datos resultante efectuamos un proceso de imputación para completar los valores de las variables que tienen valores perdidos. Este proceso de imputación considera dos métodos según el tipo de variable:

- Predictive mean matching (pmm), para variables cuantitativas
- Polytomous logistic regression (polyreg), para variables categóricas

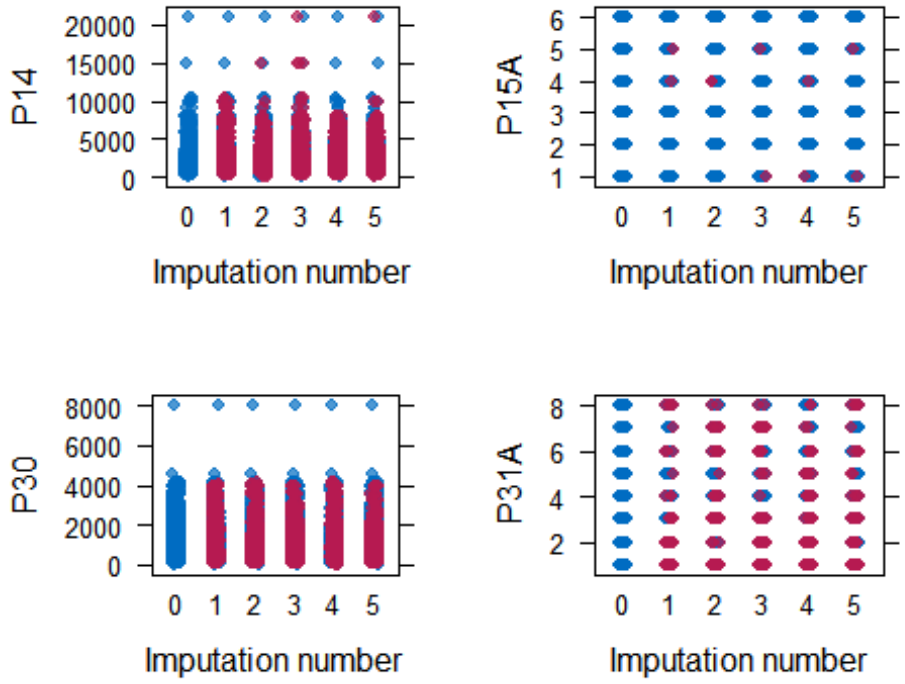
Luego, los métodos de imputación correspondientes a cada variable con valores faltantes son:

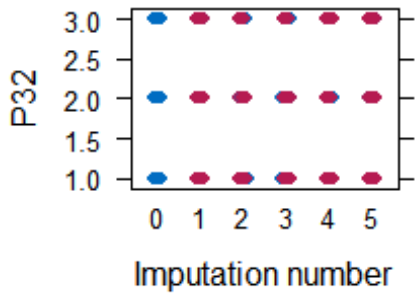
```
##       P14      P15A       P30      P31A       P32       P33       P34
##     "pmm" "polyreg"     "pmm" "polyreg" "polyreg" "polyreg"     "pmm"
##
##       P35      P36A     P129A     P130A     P131A     P132A     P133A
## "polyreg" "polyreg" "polyreg" "polyreg" "polyreg" "polyreg" "polyreg"
##
##      P134      P135      P136      P137      P138      P139      P140
## "polyreg" "polyreg" "polyreg" "polyreg" "polyreg" "polyreg" "polyreg"
```

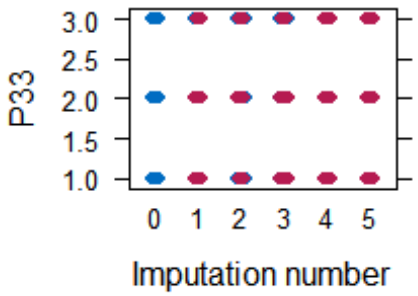
Ahora, el proceso de imputación genera cinco bases de datos imputadas. Las cinco son diferentes porque en cada caso los métodos de imputación simulan valores para los valores perdidos. De esta forma, a fin de validar este proceso de imputación, se muestran los siguientes gráficos de dispersión. En cada gráfico, el valor 0 en el eje X representa la distribución original y el resto de valores (1, 2, 3, 4 y 5) representan las distribuciones en las bases de datos imputadas, donde los puntos celestes son valores reales y los puntos rojos valores imputados. Analizando estos gráficos puede observarse que las imputaciones mantienen los comportamientos de distribución y por tanto podemos validar el procedimiento usado.

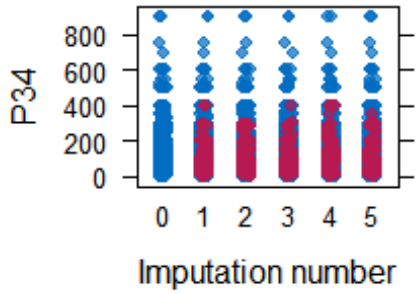


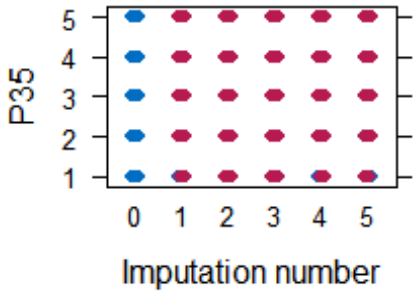
Dispersión de variables en base de datos original y bases de datos imputadas

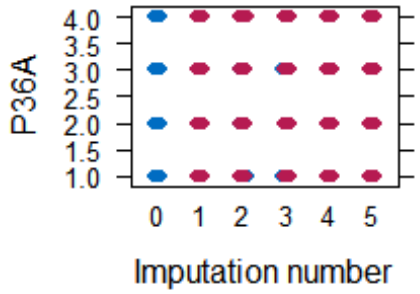


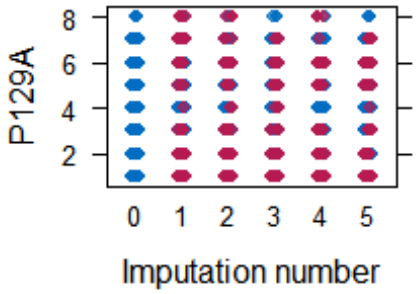

P32
3.0
2.5
2.0
1.5
1.0
0 1 2 3 4 5
Imputation number

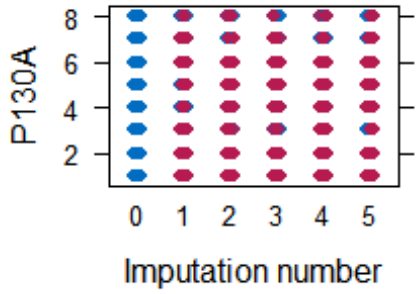

P33
3.0
2.5
2.0
1.5
1.0
0 1 2 3 4 5
Imputation number

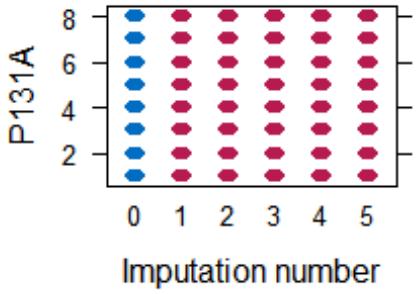

P34
800
600
400
200
0
0 1 2 3 4 5
Imputation number

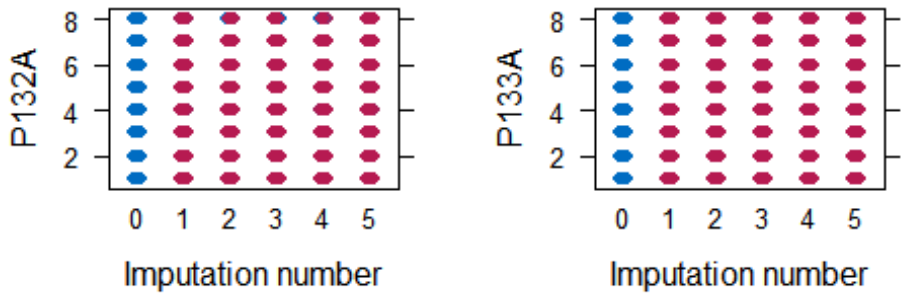

P35
5
4
3
2
1
0 1 2 3 4 5
Imputation number


P36A
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0 1 2 3 4 5
Imputation number

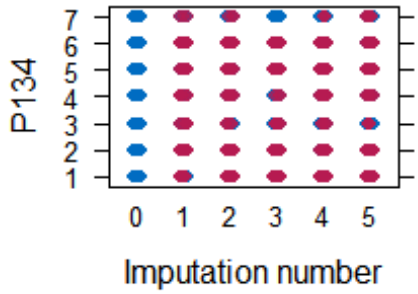

P129A
8
6
4
2
0 1 2 3 4 5
Imputation number

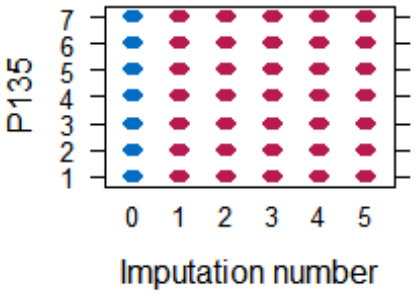

P130A
8
6
4
2
0 1 2 3 4 5
Imputation number

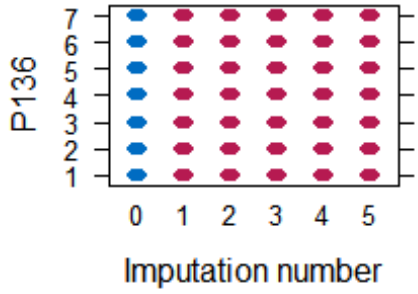

P131A
8
6
4
2
0 1 2 3 4 5
Imputation number

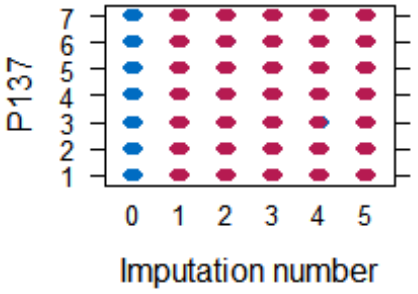

P132A
8
6
4
2
0 1 2 3 4 5
Imputation number

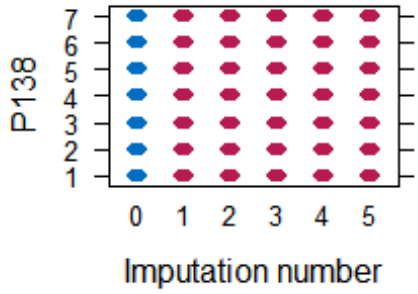
P133A
8
6
4
2
0 1 2 3 4 5
Imputation number

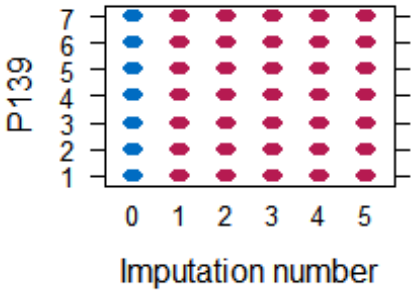

P134
7
6
5
4
3
2
1
0 1 2 3 4 5
Imputation number

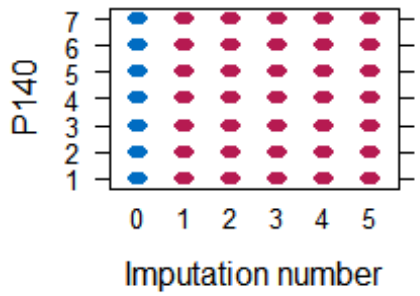

P135
7
6
5
4
3
2
1
0 1 2 3 4 5
Imputation number


P136
7
6
5
4
3
2
1
0 1 2 3 4 5
Imputation number

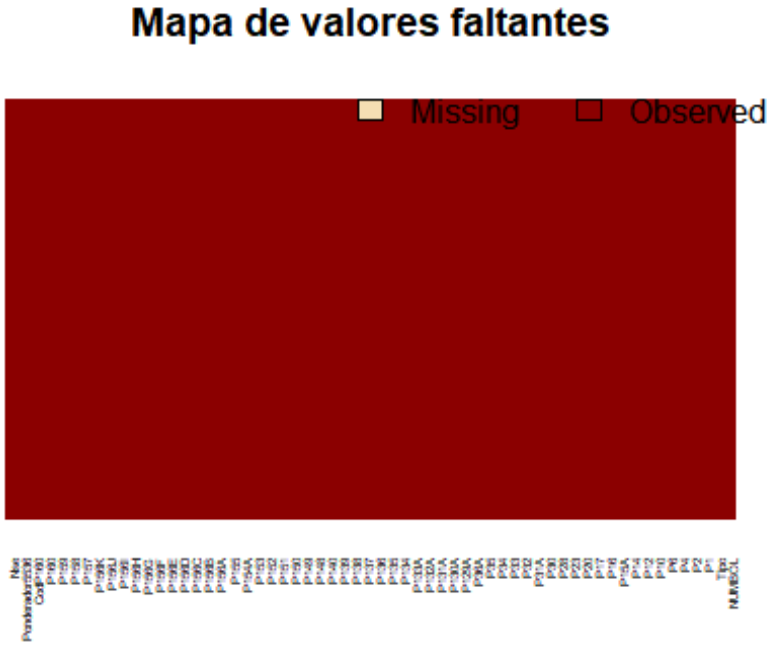

P137
7
6
5
4
3
2
1
0 1 2 3 4 5
Imputation number


P138
7
6
5
4
3
2
1
0 1 2 3 4 5
Imputation number


P139
7
6
5
4
3
2
1
0 1 2 3 4 5
Imputation number



Concluido el proceso de imputación, claramente el mapeo de valores faltantes nos muestra que ya no existen valores faltantes.



2.4 Pre-selección de variables para la etapa de modelación

Ahora, dado que en la etapa de modelación usaremos un modelo de conglomerados debemos dejar de lado variables que no aporten a la separación de los mismos. De esta forma, no tiene sentido considerar el número de encuestado (NUMBOL), el ponderador (Ponderador5536) ni la variable (Nse). Por otra parte, la pregunta 31 de la encuesta a internautas: ¿Qué usos le da a su celular? permite varias respuestas. Luego, la variable P31A es tan solo una de las respuestas de los encuestados y no

necesariamente representa el uso más importante, y en todo caso es probable que tan solo represente el primer valor digitado en la base de datos respecto a esta pregunta. Por este motivo, tampoco resultaría beneficioso considerar la pregunta P31A. Finalmente, las variables P160 y CodP160 representan el nombre y código de las comunidades a las que pertenecen los encuestados, cada una con un total de 97 alternativas posibles y donde en varios casos se tienen menos de 10 observaciones. Luego, este total de alternativas genera un problema técnico en la segunda etapa de modelación donde se aplicará el modelo Random Forest, motivo por el cual no se consideran ambas variables.

Tomando esta pre-selección de variables, a continuación se muestra el resumen de la base de datos que ingresará a la etapa de modelación.

```
##  Tipo           P1        P2       P4       P6             P10
##  1:5033   Min.   :14.00   1:5066   1:1098   1:2736   9      :2455
##  2: 503   1st Qu.:21.00   2: 470   2:4438   2:2800   7      : 667
##           Median :30.00                              8      : 467
##           Mean   :32.62                              3      : 444
##           3rd Qu.:42.00                              2      : 440
##           Max.   :90.00                              5      : 321
##                                                      (Other): 742
##
##  P12           P14        P15A     P16      P17      P20      P23
##  1:5297   Min.   :   70   1:3883   1:2277   1: 625   1:4501   1: 572
##  2: 239   1st Qu.: 1200   2: 500   2:3259   2:4911   2:1035   2: 911
##           Median : 1800   3:  98                             3:4053
##           Mean   : 2082   4: 812
##           3rd Qu.: 2650   5:  21
##           Max.   :21000   6: 222
##
##  P28           P30       P32      P33           P34          P35
##  1:5345   Min.   :  10   1:5030   1:1873   Min.   :  1.00   1: 137
##  2: 191   1st Qu.: 600   2: 450   2:2765   1st Qu.: 35.00   2: 357
##           Median : 950   3:  56   3: 898   Median : 70.00   3:1533
##           Mean   :1102                     Mean   : 89.99   4:2175
##           3rd Qu.:1500                     3rd Qu.:120.00   5:1334
##           Max.   :8001                     Max.   :900.00
##
##  P36A         P129A          P130A          P131A          P132A
##  1:4846   1      :4004   1      :3383   1      :2572   1      :3294
##  2: 216   2      : 479   6      : 670   5      :1062   6      : 895
##  3: 116   6      : 456   2      : 529   6      : 790   5      : 465
##  4: 358   5      : 234   5      : 503   2      : 420   2      : 386
##           3      : 198   3      : 251   3      : 292   3      : 225
##           4      : 108   4      : 145   4      : 268   7      : 153
##           (Other):  57   (Other):  55   (Other): 132   (Other): 118
##
```

```
##      P133A       P134      P135      P136      P137      P138      P139
##  1      :2191   1:2734   1:2713   1:2662   1:2808   1:2013   1:2398
##  3      :1049   2: 647   2: 971   2: 927   2: 874   2:1095   2: 835
##  2      : 813   3: 119   3: 640   3:1109   3: 792   3: 701   3: 645
##  6      : 579   4: 149   4: 168   4: 162   4: 193   4: 235   4: 215
##  5      : 498   5: 624   5: 318   5: 239   5: 359   5: 339   5: 473
##  7      : 210   6:1217   6: 314   6: 166   6: 223   6: 440   6: 469
##  (Other): 196   7:  46   7: 412   7: 271   7: 287   7: 713   7: 501
##  P140      P148          P149      P150      P151          P152
##  1:1650   1:2774   3      :1409   1:2726   1:4959   ES     :1539
##  2:1107   2:2762   2      :1231   2:2360   2: 143   AS     :1160
##  3: 186            6      :1061   3: 298   3: 415   CP     : 850
##  4: 129            7      : 625   4: 152   4:   5   NP     : 781
##  5: 394            4      : 611            5:   9   AC     : 623
##  6:1732            1      : 392            6:   5   PI     : 232
##  7: 338            (Other): 207                      (Other): 351
##
##       P153       P154A          P155       P156A    P156B    P156C
##  3      :1411   1:3733   2      :2198   1:5183   1:4046   1:5345
##  1      : 931   2:1256   1      :1353   2: 353   2:1490   2: 191
##  4      : 848   3:  88   3      :1197
##  7      : 811   4: 221   4      : 465
##  2      : 793   5: 238   5      : 215
##  6      : 373            6      :  85
##  (Other): 369            (Other):  23
##
##  P156D    P156E    P156F    P156G    P156H         P156I      P156J
##  1:1740   1:1195   1:2244   1: 863   1:3835   1      :3604   1: 730
##  2:3796   2:4341   2:3292   2:4673   2:1701   2      :1345   2:4806
##                                               3      : 296
##                                               0      : 129
##                                               4      : 108
##                                               5      :  39
##                                               (Other):  15
##
##  P156K          P157           P158      P159
##  1: 267   4      :1416   4      : 741   1:2824
##  2:5269   5      :1197   1      : 663   2:1306
##           3      : 983   6      : 661   3:1406
##           6      : 735   9      : 647
##           2      : 390   2      : 615
##           7      : 356   3      : 598
##           (Other): 459   (Other):1611
```

3. Modelación

A continuación se describe la etapa de modelación. Partimos describiendo el análisis de conglomerados por K-Prototypes, luego se explica el proceso de identificación de variables importantes por Random Forest y finalmente se resume el segundo análisis de conglomerados por K-Prototypes considerando el subconjunto de variables importantes.

3.1 Análisis de Conglomerados por K-Prototypes

Para empezar resolvemos un problema de aprendizaje no supervisado aplicando el modelo de conglomerados K-Prototypes. El objetivo es identificar tres conglomerados, y si bien es posible tratar de identificar un número óptimo, se define tres como parámetro para fines de interpretación.

Luego de aplicar este primer modelo podemos ver que el 36.25% de las observaciones pertenece al conglomerado 1, 21.35% al conglomerado 2 y 42.40% al conglomerado 3.

```
##     1     2     3
## 36.25 21.35 42.40
```

Ahora revisamos el resumen del modelo K-Prototypes que muestra comparaciones entre los conglomerados a partir de distintos estadísticos, para variables categóricas proporciones y para variables cuantitativas medidas de posición.

```
## Tipo
##
## cluster     1     2
##       1 0.828 0.172
##       2 0.983 0.017
##       3 0.941 0.059
## ------------------------------------------------------------
## P1
##   Min. 1st Qu. Median  Mean 3rd Qu. Max.
## 1   14      20     29 33.01      44   89
## 2   14      24     33 34.50      43   82
## 3   14      21     28 31.33      40   90
## ------------------------------------------------------------
## P2
##
## cluster     1     2
##       1 0.838 0.162
##       2 0.986 0.014
##       3 0.945 0.055
## ------------------------------------------------------------
## P4
##
## cluster     1     2
##       1 0.049 0.951
##       2 0.426 0.574
##       3 0.211 0.789
## ------------------------------------------------------------
```

```
## P6
##
## cluster     1     2
##       1 0.191 0.809
##       2 0.828 0.172
##       3 0.585 0.415
##
## ------------------------------------------------------------------
## P10
##
## cluster     1     2     3     4     5     6     7     8     9
##       1 0.052 0.065 0.049 0.029 0.027 0.012 0.029 0.077 0.658
##       2 0.057 0.080 0.106 0.069 0.100 0.066 0.297 0.080 0.146
##       3 0.047 0.091 0.094 0.054 0.063 0.038 0.110 0.093 0.410
##
## ------------------------------------------------------------------
## P12
##
## cluster     1     2
##       1 0.918 0.082
##       2 0.978 0.022
##       3 0.979 0.021
##
## ------------------------------------------------------------------
## P14
##   Min. 1st Qu. Median Mean 3rd Qu.  Max.
## 1   70     900   1400 1494    2000  5250
## 2  450    2500   3225 3645    4250 21000
## 3  250    1200   1700 1798    2225  5250
##
## ------------------------------------------------------------------
## P15A
##
## cluster     1     2     3     4     5     6
##       1 0.670 0.124 0.012 0.126 0.003 0.066
##       2 0.730 0.090 0.025 0.110 0.007 0.039
##       3 0.714 0.062 0.019 0.183 0.003 0.019
##
## ------------------------------------------------------------------
## P16
##
## cluster     1     2
##       1 0.119 0.881
##       2 0.672 0.328
##       3 0.530 0.470
##
## ------------------------------------------------------------------
## P17
##
## cluster     1     2
##       1 0.116 0.884
##       2 0.125 0.875
##       3 0.104 0.896
## ------------------------------------------------------------------
```

```
## P20
##
## cluster     1     2
##       1 0.841 0.159
##       2 0.833 0.167
##       3 0.779 0.221
##
## ------------------------------------------------------------------
## P23
##
## cluster     1     2     3
##       1 0.019 0.057 0.924
##       2 0.255 0.282 0.463
##       3 0.098 0.198 0.704
##
## ------------------------------------------------------------------
## P28
##
## cluster     1     2
##       1 0.934 0.066
##       2 0.986 0.014
##       3 0.983 0.017
##
## ------------------------------------------------------------------
## P30
##   Min. 1st Qu. Median   Mean 3rd Qu. Max.
## 1   10     450    700  799.5    1050 4000
## 2   60     900   1500 1585.0    2000 4200
## 3   50     680   1000 1117.0    1500 8001
##
## ------------------------------------------------------------------
## P32
##
## cluster     1     2     3
##       1 0.955 0.039 0.006
##       2 0.810 0.180 0.010
##       3 0.919 0.068 0.013
##
## ------------------------------------------------------------------
## P33
##
## cluster     1     2     3
##       1 0.295 0.598 0.107
##       2 0.366 0.423 0.211
##       3 0.361 0.454 0.185
##
## ------------------------------------------------------------------
## P34
##   Min. 1st Qu. Median   Mean 3rd Qu. Max.
## 1    3      30     50  73.12     100  600
## 2    1      50    100 118.40     160  900
## 3    2      40     75  90.08     120  900
##
## ------------------------------------------------------------------
```

```
## P35
##
## cluster     1     2     3     4     5
##       1 0.026 0.075 0.277 0.381 0.240
##       2 0.031 0.060 0.280 0.393 0.235
##       3 0.020 0.058 0.275 0.403 0.245
##
## ------------------------------------------------------------------
## P36A
##
## cluster     1     2     3     4
##       1 0.819 0.026 0.024 0.131
##       2 0.910 0.063 0.013 0.014
##       3 0.906 0.037 0.023 0.034
##
## ------------------------------------------------------------------
## P129A
##
## cluster     1     2     3     4     5     6     7     8
##       1 0.701 0.140 0.027 0.017 0.027 0.076 0.010 0.002
##       2 0.684 0.066 0.058 0.036 0.060 0.085 0.011 0.001
##       3 0.763 0.052 0.032 0.013 0.046 0.087 0.007 0.001
##
## ------------------------------------------------------------------
## P130A
##
## cluster     1     2     3     4     5     6     7     8
##       1 0.588 0.150 0.044 0.016 0.067 0.124 0.008 0.002
##       2 0.538 0.068 0.070 0.052 0.127 0.135 0.008 0.003
##       3 0.668 0.063 0.034 0.022 0.093 0.112 0.007 0.002
##
## ------------------------------------------------------------------
## P131A
##
## cluster     1     2     3     4     5     6     7     8
##       1 0.467 0.114 0.047 0.038 0.159 0.139 0.020 0.015
##       2 0.364 0.052 0.074 0.076 0.273 0.148 0.008 0.005
##       3 0.513 0.055 0.047 0.043 0.179 0.143 0.015 0.004
##
## ------------------------------------------------------------------
## P132A
##
## cluster     1     2     3     4     5     6     7     8
##       1 0.579 0.112 0.038 0.010 0.070 0.145 0.038 0.007
##       2 0.563 0.043 0.040 0.029 0.111 0.195 0.019 0.002
##       3 0.625 0.047 0.043 0.017 0.082 0.159 0.023 0.003
##
## ------------------------------------------------------------------
```

```
## P133A
##
## cluster     1     2     3     4     5     6     7     8
##       1 0.412 0.229 0.145 0.015 0.058 0.096 0.036 0.010
##       2 0.295 0.099 0.213 0.032 0.166 0.128 0.047 0.019
##       3 0.433 0.101 0.216 0.022 0.079 0.101 0.035 0.014
## ------------------------------------------------------------
## P134
##
## cluster     1     2     3     4     5     6     7
##       1 0.436 0.206 0.018 0.023 0.106 0.198 0.013
##       2 0.459 0.070 0.030 0.041 0.157 0.233 0.010
##       3 0.560 0.064 0.020 0.023 0.096 0.232 0.003
##
## ------------------------------------------------------------
## P135
##
## cluster     1     2     3     4     5     6     7
##       1 0.428 0.273 0.086 0.032 0.052 0.051 0.079
##       2 0.438 0.155 0.149 0.039 0.073 0.067 0.080
##       3 0.570 0.103 0.124 0.025 0.054 0.056 0.068
##
## ------------------------------------------------------------
## P136
##
## cluster     1     2     3     4     5     6     7
##       1 0.421 0.279 0.154 0.022 0.043 0.030 0.051
##       2 0.471 0.107 0.242 0.041 0.048 0.027 0.063
##       3 0.537 0.103 0.219 0.029 0.041 0.032 0.040
##
## ------------------------------------------------------------
## P137
##
## cluster     1     2     3     4     5     6     7
##       1 0.450 0.250 0.113 0.026 0.053 0.047 0.061
##       2 0.496 0.102 0.163 0.047 0.095 0.036 0.061
##       3 0.562 0.108 0.159 0.036 0.060 0.037 0.040
## ------------------------------------------------------------
## P138
##
## cluster     1     2     3     4     5     6     7
##       1 0.264 0.329 0.101 0.033 0.058 0.087 0.130
##       2 0.321 0.157 0.122 0.071 0.069 0.082 0.178
##       3 0.471 0.106 0.151 0.036 0.060 0.072 0.104
##
## ------------------------------------------------------------
## P139
##
## cluster     1     2     3     4     5     6     7
##       1 0.353 0.252 0.088 0.042 0.076 0.100 0.089
##       2 0.385 0.110 0.145 0.046 0.107 0.082 0.126
##       3 0.526 0.085 0.127 0.033 0.083 0.073 0.074
## ------------------------------------------------------------
```

```
## P140
##
## cluster     1     2     3     4     5     6     7
##       1 0.217 0.338 0.020 0.021 0.064 0.277 0.063
##       2 0.200 0.183 0.045 0.023 0.096 0.375 0.080
##       3 0.417 0.091 0.039 0.026 0.065 0.313 0.050
## ------------------------------------------------------------
## P148
##
## cluster     1     2
##       1 0.626 0.374
##       2 0.612 0.388
##       3 0.338 0.662
##
## ------------------------------------------------------------
## P149
##
## cluster     1     2     3     4     5     6     7     8
##       1 0.141 0.370 0.185 0.093 0.016 0.143 0.051 0.002
##       2 0.014 0.105 0.122 0.136 0.019 0.227 0.300 0.078
##       3 0.040 0.156 0.381 0.112 0.016 0.215 0.072 0.009
##
## ------------------------------------------------------------
## P150
##
## cluster     1     2     3     4
##       1 0.499 0.397 0.057 0.047
##       2 0.398 0.552 0.038 0.013
##       3 0.535 0.389 0.059 0.018
##
## ------------------------------------------------------------
## P151
##
## cluster     1     2     3     4     5     6
##       1 0.793 0.051 0.149 0.001 0.004 0.001
##       2 0.961 0.012 0.027 0.000 0.000 0.000
##       3 0.951 0.011 0.035 0.001 0.000 0.001
##
## ------------------------------------------------------------
## P152
##
## cluster    AS    CP    NP    PI    EM    ES    AC    RJ    SB    SN
##       1 0.177 0.215 0.110 0.015 0.005 0.292 0.114 0.026 0.036 0.010
##       2 0.321 0.107 0.150 0.118 0.008 0.190 0.059 0.021 0.019 0.005
##       3 0.182 0.124 0.164 0.026 0.014 0.310 0.138 0.014 0.024 0.004
##
## ------------------------------------------------------------
```

```
## P153
##
## cluster     1     2     3     4     5     6     7     8
##       1 0.328 0.230 0.188 0.121 0.026 0.049 0.051 0.006
##       2 0.049 0.073 0.145 0.138 0.041 0.072 0.377 0.106
##       3 0.091 0.104 0.367 0.188 0.040 0.081 0.112 0.016
##
## ------------------------------------------------------------
## P154A
##
## cluster     1     2     3     4     5
##       1 0.542 0.330 0.017 0.054 0.056
##       2 0.821 0.111 0.008 0.024 0.036
##       3 0.713 0.197 0.018 0.036 0.036
## ------------------------------------------------------------
## P155
##
## cluster     1     2     3     4     5     6     7     8
##       1 0.368 0.455 0.123 0.038 0.010 0.005 0.000 0.000
##       2 0.115 0.205 0.354 0.157 0.104 0.047 0.017 0.001
##       3 0.204 0.444 0.226 0.087 0.030 0.008 0.001 0.000
##
## ------------------------------------------------------------
## P156A
##
## cluster     1     2
##       1 0.890 0.110
##       2 0.983 0.017
##       3 0.952 0.048
##
## ------------------------------------------------------------
## P156B
##
## cluster     1     2
##       1 0.589 0.411
##       2 0.915 0.085
##       3 0.759 0.241
##
## ------------------------------------------------------------
## P156C
##
## cluster     1     2
##       1 0.942 0.058
##       2 0.995 0.005
##       3 0.971 0.029
##
## ------------------------------------------------------------
## P156D
##
## cluster     1     2
##       1 0.093 0.907
##       2 0.684 0.316
##       3 0.317 0.683
## ------------------------------------------------------------
```

```
## P156E
##
## cluster     1     2
##       1 0.043 0.957
##       2 0.608 0.392
##       3 0.166 0.834
## ------------------------------------------------------------------
## P156F
##
## cluster     1     2
##       1 0.235 0.765
##       2 0.742 0.258
##       3 0.381 0.619
##
## ------------------------------------------------------------------
## P156G
##
## cluster     1     2
##       1 0.051 0.949
##       2 0.370 0.630
##       3 0.138 0.862
##
## ------------------------------------------------------------------
## P156H
##
## cluster     1     2
##       1 0.699 0.301
##       2 0.735 0.265
##       3 0.666 0.334
##
## ------------------------------------------------------------------
## P156I
##
## cluster     0     1     2     3     4     5     6     9    10    15
##       1 0.038 0.796 0.134 0.024 0.006 0.001 0.000 0.000 0.000 0.000
##       2 0.009 0.340 0.474 0.111 0.044 0.016 0.005 0.001 0.000 0.000
##       3 0.017 0.684 0.220 0.049 0.019 0.007 0.002 0.001 0.000 0.000
##
## ------------------------------------------------------------------
## P156J
##
## cluster     1     2
##       1 0.059 0.941
##       2 0.254 0.746
##       3 0.133 0.867
##
## ------------------------------------------------------------------
## P156K
##
## cluster     1     2
##       1 0.018 0.982
##       2 0.116 0.884
##       3 0.040 0.960
## ------------------------------------------------------------------
```

```
## P157
##
## cluster     1     2     3     4     5     6     7     8     9    10    11
##       1 0.040 0.077 0.177 0.220 0.216 0.137 0.075 0.040 0.012 0.003 0.000
##       2 0.030 0.068 0.179 0.296 0.226 0.111 0.046 0.031 0.009 0.005 0.000
##       3 0.030 0.066 0.178 0.266 0.211 0.141 0.064 0.032 0.007 0.003 0.001
##
## cluster    12    13
##       1 0.001 0.000
##       2 0.000 0.000
##       3 0.002 0.000
##
## ------------------------------------------------------------
## P158
##
## cluster     1     2     3     4     5     6     7     8     9
##       1 0.113 0.117 0.103 0.077 0.110 0.204 0.081 0.104 0.093
##       2 0.185 0.113 0.149 0.108 0.073 0.094 0.103 0.050 0.125
##       3 0.093 0.106 0.092 0.195 0.092 0.060 0.111 0.118 0.133
##
## ------------------------------------------------------------
## P159
##
## cluster     1     2     3
##       1 0.234 0.260 0.506
##       2 0.757 0.182 0.061
##       3 0.622 0.242 0.135
##
## ------------------------------------------------------------
```

Notemos que en los resultados existen variables donde la diferencia entre conglomerados es mínima. De esta forma, a continuación se plantea un método para seleccionar las variables más importantes para la identificación de los conglomerados.

3.2 Identificación de Variables Importantes por Random Forest

Para identificar las variables más importantes ahora consideramos un problema de aprendizaje supervisado donde trataremos de predecir el conglomerado al que pertenece cada observación a partir de las variables que ingresaron al modelo K-Prototypes. Para este fin, primero partiremos la base de datos en dos partes: un conjunto de observaciones de entrenamiento y otro de prueba.

A continuación se preparan los datos para la modelación y se aplica el modelo Random Forest. Se toma el conjunto de entrenamiento y se define la selección aleatoria de dos mil árboles de decisión que en promedio estimarán la pertenencia de cualquier observación a alguno de los conglomerados.

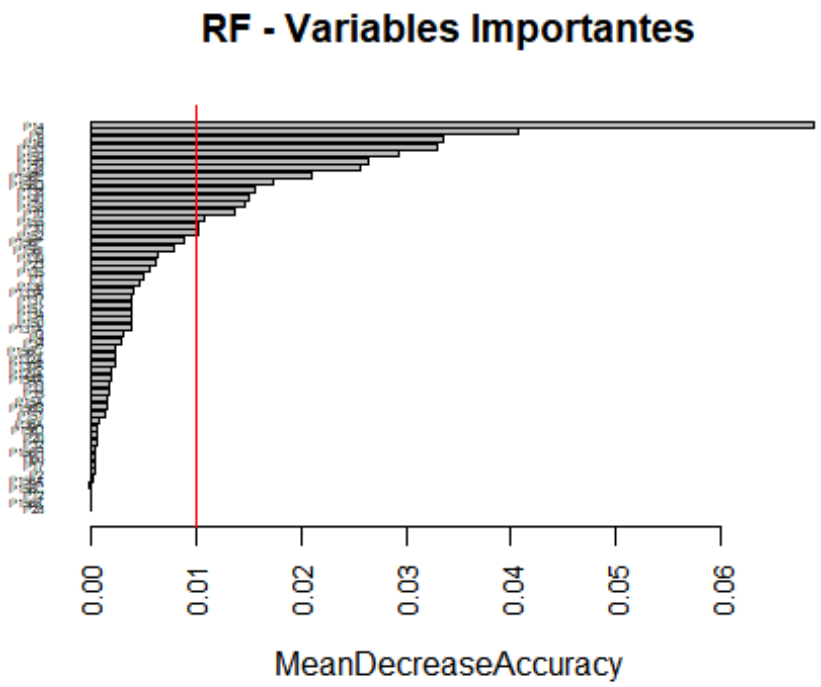
Una vez entrenado el modelo, ahora nos interesa identificar las variables más importantes. Para este fin primero analizamos la exactitud en la predicción de los

conglomerados a partir de la tabla de confusión considerando el conjunto de observaciones de prueba. Esta tabla no es más que una tabla de contingencia donde en las filas se tienen los conglomerados predichos y en las columnas los conglomerados efectivos. Luego, la exactitud de predicción es la relación entre la sumatoria de los valores de la diagonal de la tabla y el total de observaciones de prueba. A partir de este cálculo, obtenemos una exactitud de predicción de 88.19%. Con este primer paso validamos la idea de que podemos predecir la pertenencia de los conglomerados a partir de las variables que ingresaron al modelo K-Prototypes.

```
## Tabla de confusión
## clust.i_rf.hat    1    2    3
##              1  874   17   57
##              2    7  477   34
##              3  114   98 1090

## [1] "EP = 88.19"
```

Como segundo paso, ordenamos las variables que ingresaron al modelo Random Forest por su reducción promedio en la exactitud de predicción si se dejaran de considerar en el modelo.



Filtrando aquellas variables que tienen una reducción promedio en la exactitud de predicción mayor a 0.01, identificamos las siguientes variables importantes para la definición de los conglomerados.

Variable	Pregunta	MeanDecrease Accuracy
P14	¿Cuánto pagaron por el televisor?	0.06883702
P6	¿Tiene computadora de escritorio, computadora portátil o tablet?	0.04061848
P16	¿Tiene servicio de TV cable?	0.03361939
P153	Nivel de instrucción Jefe del hogar	0.03303252
P159	Información que busca en internet	0.02932124
P149	Nivel de instrucción del entrevistado	0.02647323
P148	Sexo	0.02561235
P156E	¿Tiene microondas?	0.02097771
P156D	¿Tiene lavadora de ropa?	0.01728703
P138	¿Qué medio considera el más imparcial?	0.01556413
P140	¿Qué medio considera el más abierto a su participación?	0.01508431
P158	Departamento	0.01477315
P10	¿Cuántos días a la semana utiliza la computadora o tablet?	0.01368472
P155	¿Ingreso promedio mensual del hogar?	0.01075884
P30	¿Cuánto pagó la última vez que se compró un celular?	0.01029748
P152	Categoría ocupacional	0.01023634

3.3 Análisis de Conglomerados por K-Prototypes tomando variables importantes

Ahora volvemos a obtener conglomerados a partir de K-Prototypes considerando las variables importantes. En este caso podemos ver que el 56.59% de las observaciones pertenece al conglomerado 1, 11.96% al conglomerado 2 y 31.45% al conglomerado 3. Es importante notar que si bien se ha presentado una rotación en los conglomerados, aún se preserva una distribución de proporciones similar a la inicial, con una mayoría, una minoría y un intermedio.

```
## Estimated lambda: 1845168
##
##     1     2     3
## 56.59 11.96 31.45
```

Por otra parte, a fin de validar la selección de variables veamos el efecto en la exactitud de predicción considerando nuevamente un modelo Random Forest.

```
## Tabla de confusión
## clust.i_rf_f.hat    1    2    3
##                1 1466    7   79
##                2   16  300   12
##                3   79   23  786

## [1] "EP = 92.2"
```

Notemos que considerando las variables importantes de hecho se ha incrementado la exactitud en las predicciones. Esto se debe a que los nuevos conglomerados tienen mayor separación en sus características. Esto es, la pérdida en la exactitud de predicción por dejar de considerar variables menos importantes resulto menor a la ganancia por considerar variables que separan mejor a los conglomerados.

4. Análisis de resultados

Para el análisis de resultados previamente renombramos los conglomerados identificados. El conglomerado mayoría tendrá la etiqueta A, el conglomerado intermedio la etiqueta B y la minoría la etiqueta C.

```
##
##     A     B     C
## 56.59 31.45 11.96
```

4.1 Descripción de conglomerados respecto a las variables seleccionadas

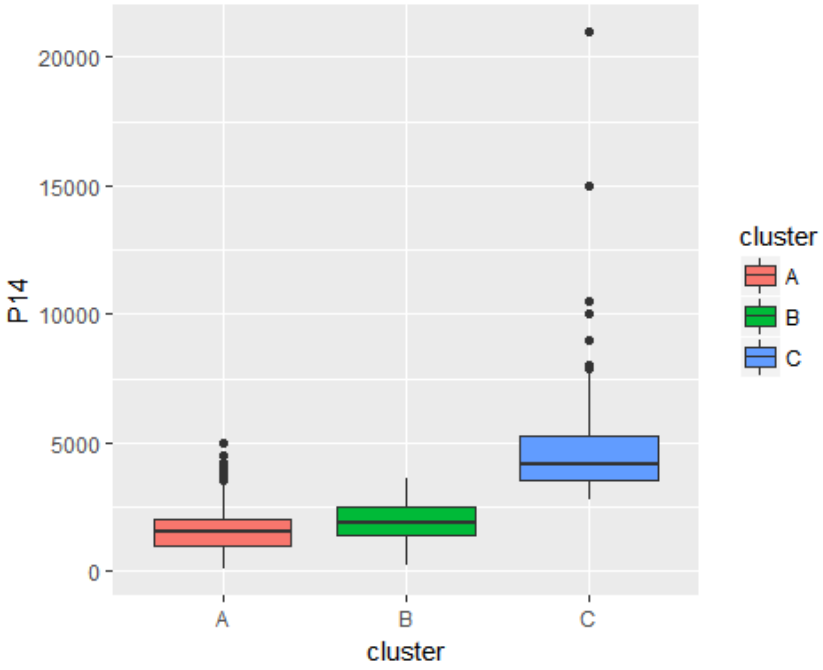
En esta sección haremos una descripción de las variables que hallamos más importantes para la clasificación de los conglomerados.

4.1.1 Costo del último televisor (P14)

En el análisis anterior pudimos observar que la variable que más aporta a la distinción entre los diferentes grupos, es el costo del último televisor que se compró en el hogar (P14). Respecto a esta variable podemos ver que las diferencias de los promedios entre los grupos son significativas, el grupo C gastó en promedio Bs 4.745, mientras que el grupo A y B gastaron en promedio Bs 1.622 y Bs 1.897, respectivamente.

```
## Resumen de estadísticos de posición
##                       $A
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      70    1000    1500    1622    2000    5000
##
##                       $B
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     250    1390    1900    1897    2500    3630
##
##                       $C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2800    3500    4150    4745    5250   21000
```

Costo de último televisor por conglomerado (Bs)

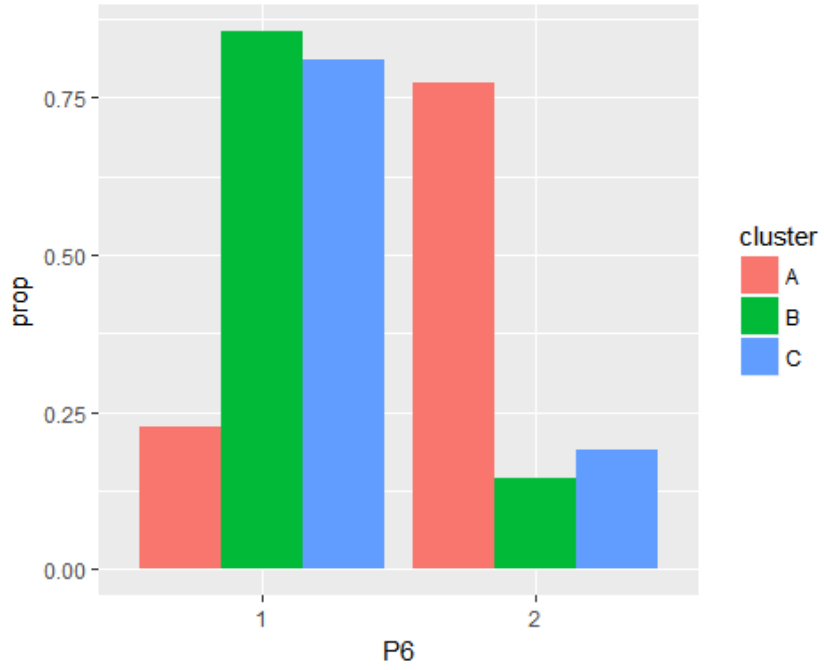


4.1.2 Tenencia de computadora de escritorio, computadora portátil o tablet en casa (P6)

Podemos observar claras diferencias entre el grupo A, respecto a los grupos B y C. Tenemos que tan solo el 23% de las personas del grupo A tienen computadoras de escritorio, portátiles o tablets en su hogar, mientras que los grupos B y C los tienen en un 86% y 81%, respectivamente. Esto nos puede mostrar que la tenencia de estos bienes influye en que las personas de los grupos B y C tengan mayor facilidad de acceso al uso de la tecnología frente al grupo A.

```
##   Frecuencias condicionales por conglomerado
##
##        A    B    C
##   1 0.23 0.86 0.81
##   2 0.77 0.14 0.19
##
##   1: Si; 2: No
```

Tenencia de computadora de escritorio, computadora portátil o tablet en casa por conglomerado

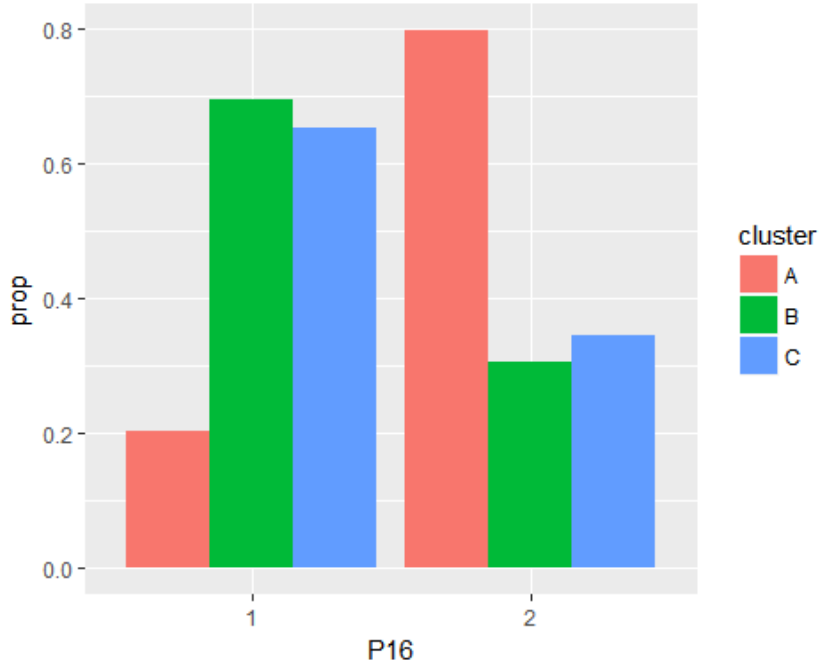


4.1.3 Tenencia de servicio de TV Cable (P16)

El comportamiento de la variable de tenencia de servicio de TV cable en el hogar tiene un comportamiento similar. Se muestra que una minoría del grupo A (20%) tiene TV cable, frente a una mayoría para los grupos B y C (70% y 65%).

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.20 0.70 0.65
##   2 0.80 0.30 0.35
##
##   1: Si; 2: No
```

Tenencia de servicio de TV Cable por conglomerado



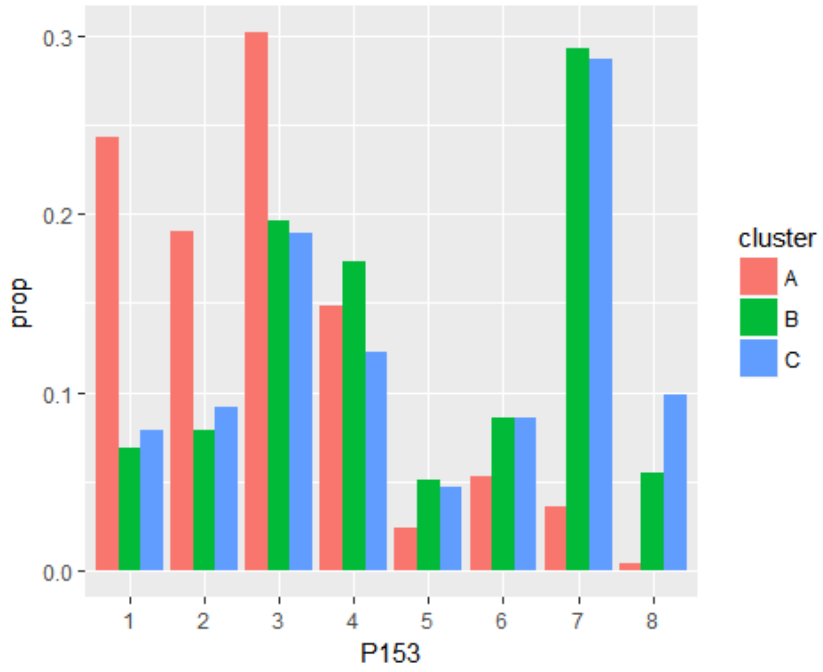
4.1.4 Nivel de instrucción del jefe de hogar (P153)

En lo que concierne al nivel de instrucción del jefe del hogar podemos observar una concentración importante diferenciada por grupos. En el caso del grupo A esta concentración está en las categorías de primaria o menos, secundaria incompleta y secundaria completa (1, 2, 3) llegando a representar el 73% de las personas del grupo A, con el 30% de las personas de este grupo donde el(la) jefe(a) de hogar tiene la secundaria completa. Mientras que en los grupos B y C el jefe de hogar con mayor frecuencia tiene la universidad completa, representando el aproximadamente el 29% para los 2 grupos.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.24 0.07 0.08
##   2 0.19 0.08 0.09
##   3 0.30 0.20 0.19
##   4 0.15 0.17 0.12
##   5 0.02 0.05 0.05
##   6 0.05 0.09 0.09
##   7 0.04 0.29 0.29
##   8 0.00 0.06 0.10
```

```
##   Frecuencias condicionales por categoría
##        A    B    C
##   1 0.82 0.13 0.06
##   2 0.75 0.17 0.08
##   3 0.67 0.24 0.09
##   4 0.55 0.35 0.10
##   5 0.39 0.45 0.16
##   6 0.45 0.40 0.15
##   7 0.14 0.63 0.23
##   8 0.07 0.55 0.37
##
##   1: Primaria o menos    ; 2: Secundaria incompleta
##   3: Secundaria completa ; 4: Técnico
##   5: Educación Terciaria ; 6: Universidad incompleta
##   7: Universidad completa; 8: Posgrado
```

Nivel de instrucción del jefe de hogar por conglomerado

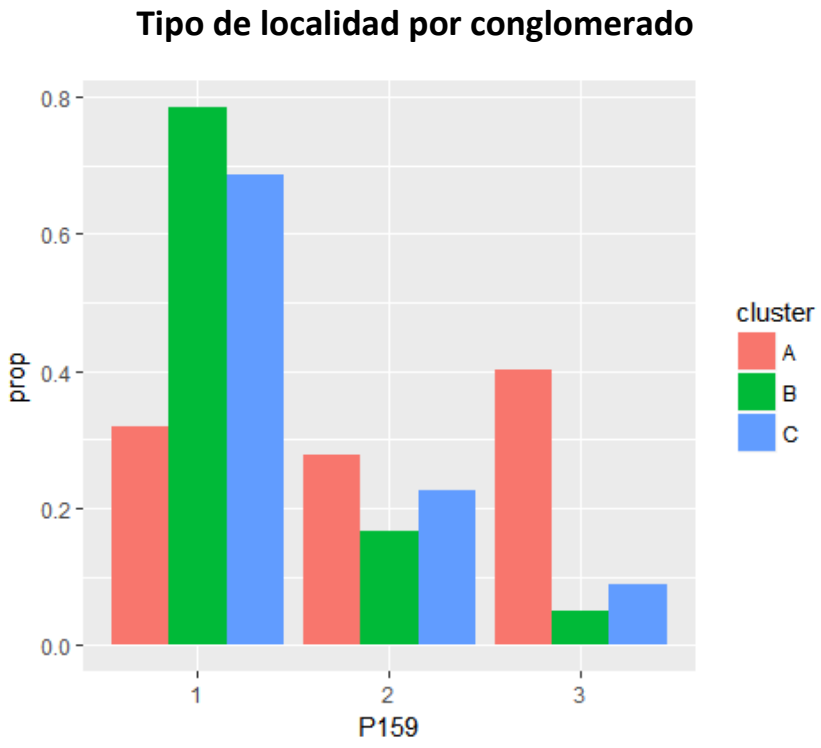


4.1.5 Tipo de localidad (P159)

Respecto a la distribución del tipo de localidad en el que habitan las personas de los grupos, podemos observar que los grupos B y C están fuertemente concentrados en las ciudades capitales siendo que el 79% de las personas del grupo B viven en ciudades capitales y el 69% de las personas del grupo C. Al mismo tiempo, observamos que la distribución del grupo A es más dispersa y concentra al 40% de las personas del grupo en centros poblados de entre 2.000 y 10.000 habitantes, y al 38% en ciudades capitales.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.32 0.79 0.69
##   2 0.28 0.16 0.23
##   3 0.40 0.05 0.09

##   Frecuencias condicionales por categoría
##        A    B    C
##   1 0.36 0.48 0.16
##   2 0.67 0.22 0.11
##   3 0.90 0.06 0.04
##
##   1: Ciudades capitales; 2: Ciudades intermedias; 3: Centros poblados
```



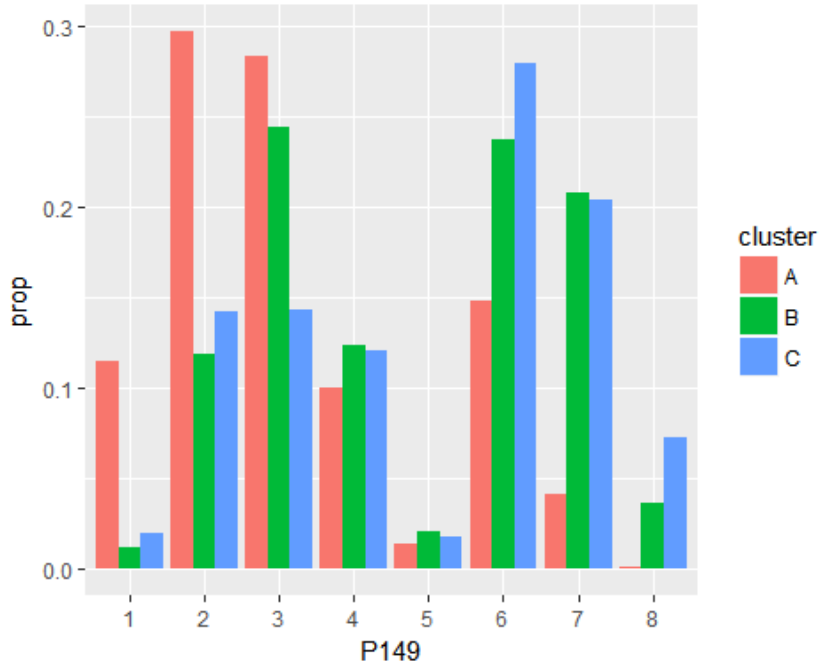
4.1.6 Nivel de instrucción del entrevistado (P149)

Respecto al nivel de instrucción del entrevistado podemos observar que la mayoría del grupo A tiene la secundaria incompleta (30%) o la secundaria completa (28%). Por otra parte la mayoría del grupo B tiene la secundaria completa (24%) o la universidad incompleta (24%). Finalmente la mayoría del grupo C tiene la universidad incompleta (28%) o la universidad completa (20%).

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.11 0.01 0.02
##   2 0.30 0.12 0.14
##   3 0.28 0.24 0.14
##   4 0.10 0.12 0.12
##   5 0.01 0.02 0.02
##   6 0.15 0.24 0.28
##   7 0.04 0.21 0.20
##   8 0.00 0.04 0.07

##   Frecuencias condicionales por categoría
##        A    B    C
##   1 0.92 0.05 0.03
##   2 0.76 0.17 0.08
##   3 0.63 0.30 0.07
##   4 0.52 0.35 0.13
##   5 0.47 0.40 0.13
##   6 0.44 0.39 0.17
##   7 0.20 0.58 0.22
##   8 0.04 0.54 0.41
##
##   1: Primaria o menos    ; 2: Secundaria incompleta
##   3: Secundaria completa ; 4: Técnico
##   5: Educación terciaria ; 6: Universidad incompleta
##   7: Universidad completa; 8: Posgrado
```

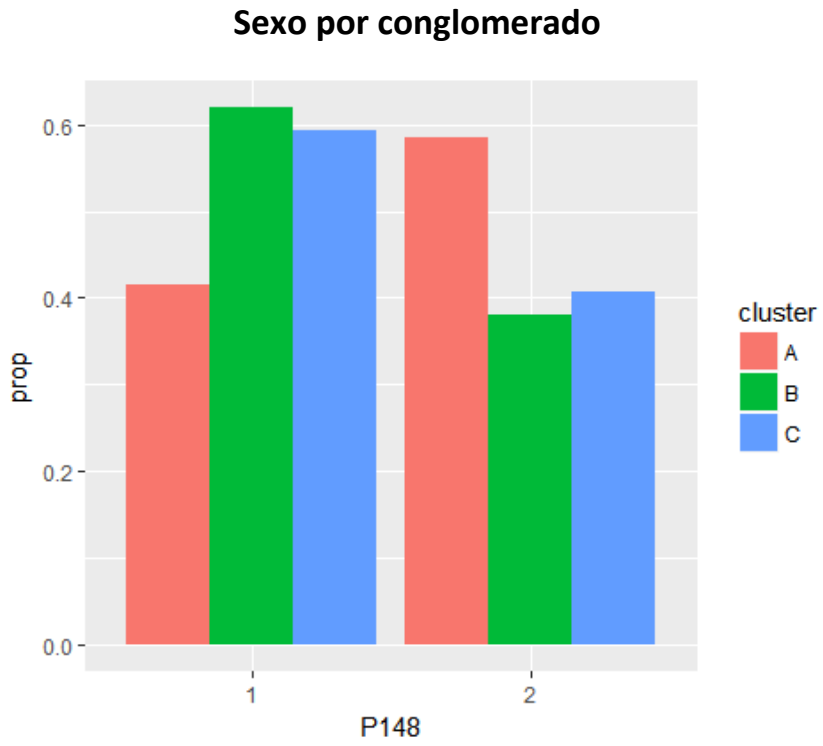
Nivel de instrucción del entrevistado por conglomerado



4.1.7 Sexo (P148)

Del total de datos en la encuesta el 50% se la realizó a hombres, y otro 50% a mujeres, pero en nuestra clasificación por categorías podemos observar que hay diferencias, no tan grandes, pero estas existen. Por ejemplo el grupo A esta en su mayoría compuesto por mujeres, llegando a representar el 58% del grupo. El grupo que tiene más diferencias es el B, con 62% de hombres frente a un 38% de mujeres. Mientras que el grupo C tiene el comportamiento inverso al de A, en este el 59% son hombres y un 41% son mujeres.

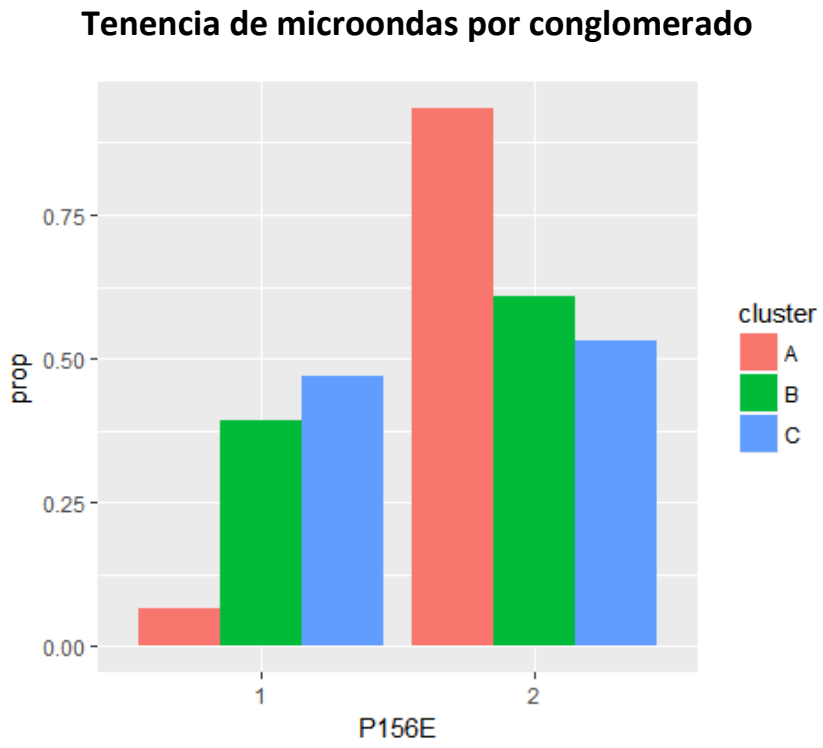
```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.42 0.62 0.59
##   2 0.58 0.38 0.41
##
##   1: Hombre; 2: Mujer
```



4.1.8 Tenencia de microondas (P156E)

Siguiendo la línea de la tenencia de bienes podemos observar claramente las diferencias que existen entre los grupos. Del total de la encuesta tan solo el 22% de los hogares tienen microondas, y haciendo la división por grupos, tenemos que el 94% de las personas del grupo A no poseen microondas, mientras que en los grupos B y C las diferencias entre los que poseen microondas y los que no, no son tan grandes, llegando a 22 y 6 puntos porcentuales, frente a la diferencia de 88 puntos porcentuales del grupo A.

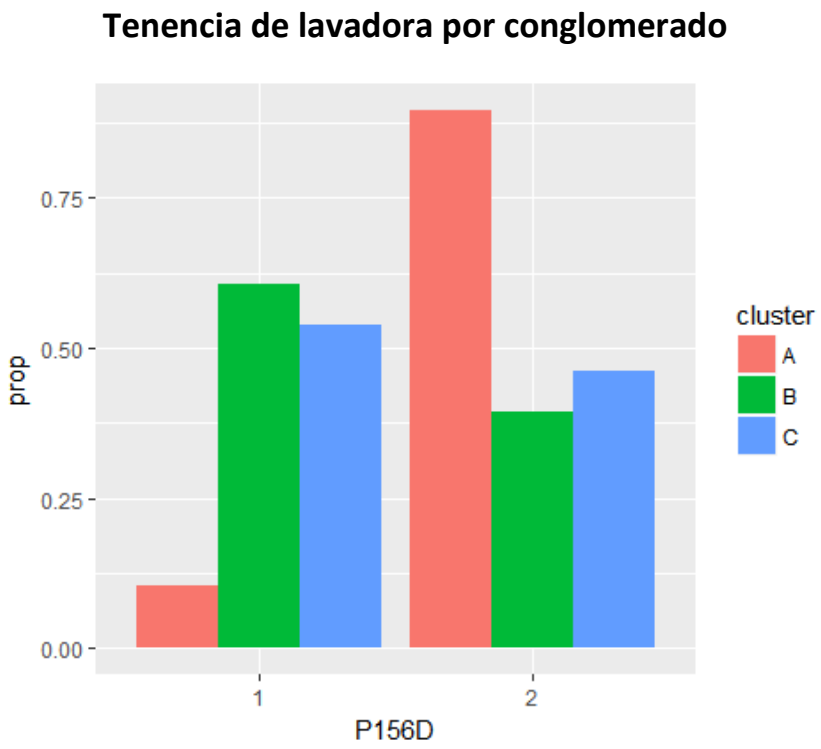
```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.06 0.39 0.47
##   2 0.94 0.61 0.53
##
##   1: Si; 2: No
```



4.1.9 Tenencia de lavadora de ropa (P156D)

Respecto a la lavadora de ropa, el comportamiento es similar, la gran mayoría de las personas del grupo A no tienen lavadora de ropa (90%) mientras que los grupos B y C si poseen este bien, con un 61% y 54% de hogares que poseen lavadora de ropa, respectivamente.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.10 0.61 0.54
##   2 0.90 0.39 0.46
##
##   1: Si; 2: No
```

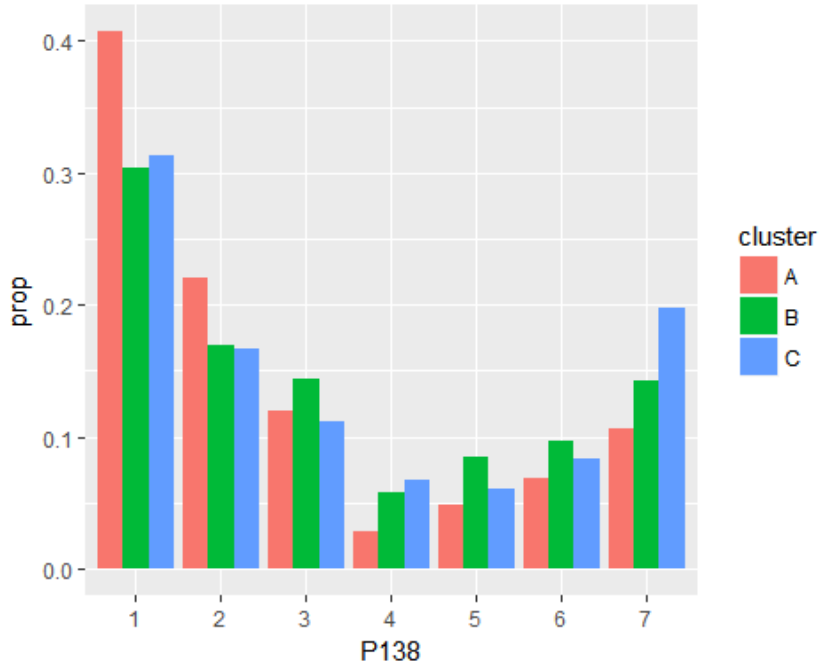


4.1.10 Consideración de medio más imparcial (P138)

Podemos observar que para los 3 grupos el medio de comunicación más imparcial es la TV, con proporciones de los grupos entre 41 y 30 porciento. Al mismo tiempo podemos observar que el grupo C es el más escéptico respecto a la imparcialidad de los medios de comunicación, siendo que el 20% de este grupo considera que ningún medio de comunicación es imparcial.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.41 0.30 0.31
##   2 0.22 0.17 0.17
##   3 0.12 0.14 0.11
##   4 0.03 0.06 0.07
##   5 0.05 0.09 0.06
##   6 0.07 0.10 0.08
##   7 0.11 0.14 0.20
##
##   1: TV             ; 2: Radio      ; 3: Periódicos Imp.
##   4: Periódicos Dig.; 5: Páginas Web; 6: Redes Sociales
##   7: Ninguno
```

Consideración de medio más imparcial por conglomerado

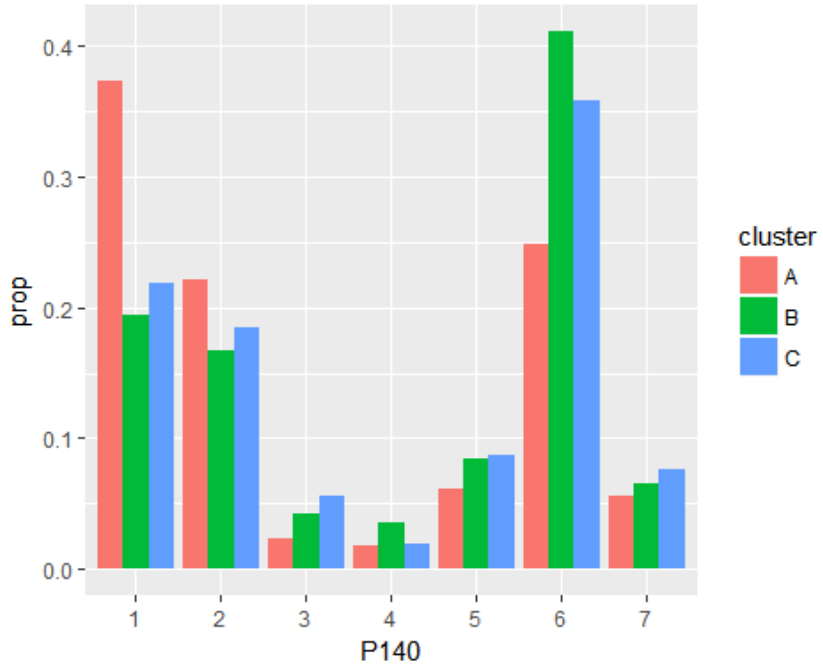


4.1.11 Consideración de medio más abierto a la participación (P140)

La mayoría de las personas del grupo A consideran a la televisión como el medio de comunicación más participativo, mientras que los grupos B y C tienen una clara preferencia por las redes sociales. Es interesante observar la interacción que tienen las personas con los medios y lo imparciales que creen que son. En este caso, si bien las redes sociales son las más participativas ocupaban el cuarto lugar en la escala de medios imparciales.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.37 0.19 0.22
##   2 0.22 0.17 0.18
##   3 0.02 0.04 0.06
##   4 0.02 0.04 0.02
##   5 0.06 0.08 0.09
##   6 0.25 0.41 0.36
##   7 0.06 0.07 0.08
##
##   1: TV             ; 2: Radio      ; 3: Periódicos Imp.
##   4: Periódicos Dig.; 5: Páginas Web; 6: Redes Sociales
##   7: Ninguno
```

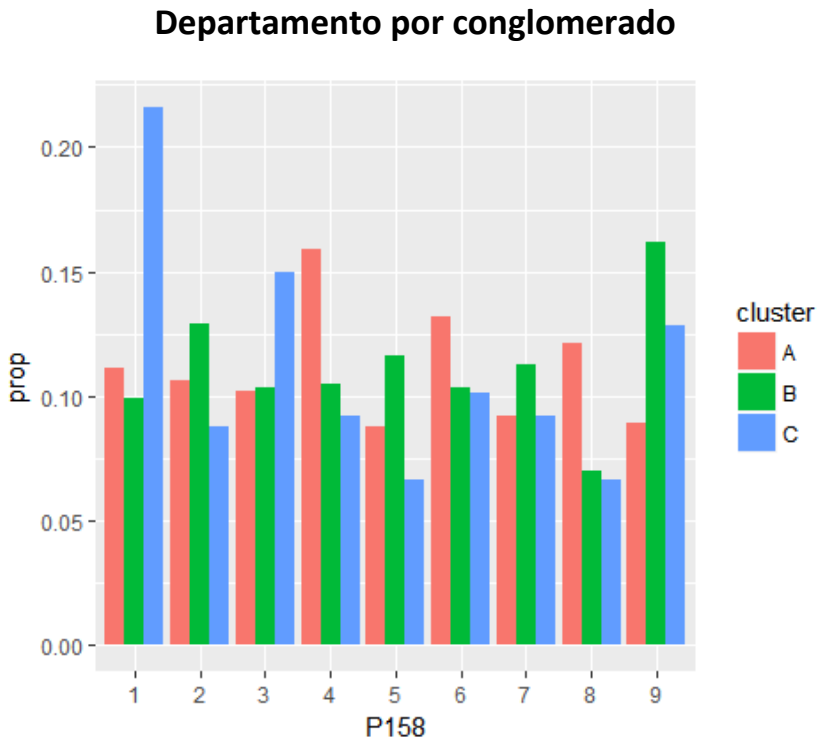
Consideración de medio más abierto a la participación por conglomerado



4.1.12 Departamento (P158)

Podemos notar una concentración importante del grupo C en La Paz (22%), Cochabamba (15%) y Pando (13%), mientras que los demás grupos están distribuidos de manera más regular entre todos los departamentos.

```
##    Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.11 0.10 0.22
##   2 0.11 0.13 0.09
##   3 0.10 0.10 0.15
##   4 0.16 0.10 0.09
##   5 0.09 0.12 0.07
##   6 0.13 0.10 0.10
##   7 0.09 0.11 0.09
##   8 0.12 0.07 0.07
##   9 0.09 0.16 0.13
##
##   1: La Paz    ; 2: Chuquisaca; 3: Cochabamba
##   4: Santa Cruz; 5: Oruro     ; 6: Potosí
##   7: Tarija    ; 8: Beni      ; 9: Pando
```



4.1.13 Días a la semana de uso de computadora de escritorio, computadora portátil o tablet (P10)

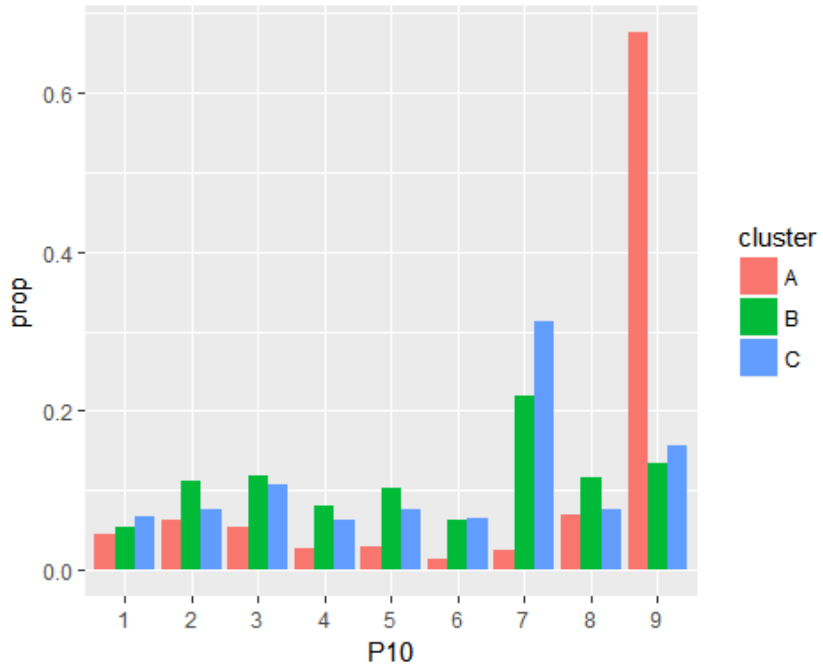
En lo que concierne a la intensidad de uso de computadoras de escritorio, portátiles o tablets, podemos observar claramente un comportamiento diferenciado. Para el grupo A, el 68% de sus integrantes no usan los equipos mostrados. Esto va en coherencia que justamente el grupo A eran aquellas personas que no tenían acceso a estos equipos en su hogar. Al mismo tiempo, de las personas que pertenecen a este grupo y si utilizan estos equipos, podemos observar que hay una concentración en frecuencias bajas, es decir el 16% de las personas del grupo A utilizan estos equipos entre 1 y 3 veces a la semana, y estas personas representan cerca del 42% de las personas que si utilizan estos equipos dentro del grupo A.

El comportamiento de los otros grupos es cualitativamente diferente, ya que estos están más "conectados". En el grupo C, el 62% de las personas utilizan estos equipos entre 3 a 7 días a la semana, siendo que el 31% de las personas del grupo C los usan todos los días. El grupo B tiene un comportamiento similar pero con valores un poco más bajos, el 58% de las personas del grupo B utilizan estos equipos entre 3 a 7 días a la semana, con un 22% de personas que los usan todos los días.

Este comportamiento nos hace pensar que las personas del grupo C son aquellas que más interactúan con las TICS.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.05 0.05 0.07
##   2 0.06 0.11 0.08
##   3 0.05 0.12 0.11
##   4 0.03 0.08 0.06
##   5 0.03 0.10 0.08
##   6 0.01 0.06 0.06
##   7 0.03 0.22 0.31
##   8 0.07 0.12 0.08
##   9 0.68 0.13 0.16
##
##   1: 1 día   ; 2: 1 días                      ; 3: 3 días;
##   4: 4 días  ; 5: 5 días                      ; 6: 6 días;
##   7: 7 días  ; 8: Menos de una vez por semana; 9: No usa estos equipos
```

Días a la semana de uso de computadora de escritorio, computadora portátil o tablet por conglomerado



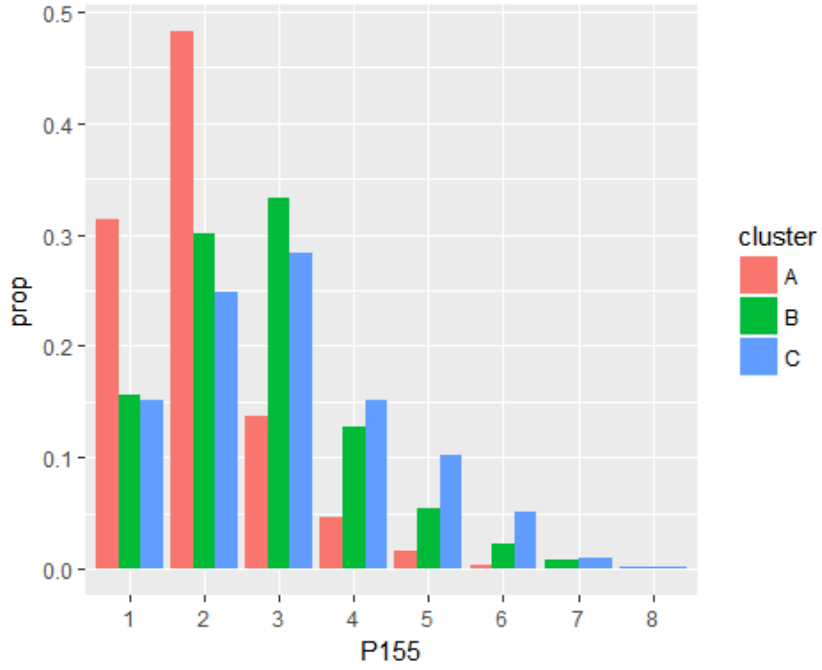
4.1.14 Ingreso mensual promedio en el hogar (P155)

Viendo la distribución de los ingresos mensuales promedio declarados en la encuesta, podemos ver que el grupo A tiene una concentración en las 2 primeras categorías, es decir el 79% de los hogares del grupo A tienen un ingreso mensual promedio de Bs 3500 o menos. En este mismo intervalo, los grupos B y C concentran el 46% y 40% respectivamente, lo cual es significativamente menor. Entre los grupos B y C podemos observar que el intervalo donde están mayormente concentrados es el de Bs 3.501 a Bs 5.000 representando el 33% del grupo B y el 28% del grupo C.

Además, en los ingresos más altos, en el intervalo de más de Bs 5.001, el grupo B agrupa el 21 % de sus integrantes, mientras que el grupo C en el mismo intervalo tiene un 31% de personas.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.31 0.16 0.15
##   2 0.48 0.30 0.25
##   3 0.14 0.33 0.28
##   4 0.05 0.13 0.15
##   5 0.02 0.05 0.10
##   6 0.00 0.02 0.05
##   7 0.00 0.01 0.01
##   8 0.00 0.00 0.00
##
##   1:     Hasta 1.400 Bs; 2: 1.401 - 3.500 Bs; 3:  3.501 -  5.000 Bs
##   4:  5.001 -  7.000 Bs; 5: 7.001 -10.000 Bs; 6: 10.000 - 14.000 Bs
##   7: 14.001 - 21.000 Bs; 8: Más de 21.000 Bs
```

Ingreso mensual promedio en el hogar por conglomerado

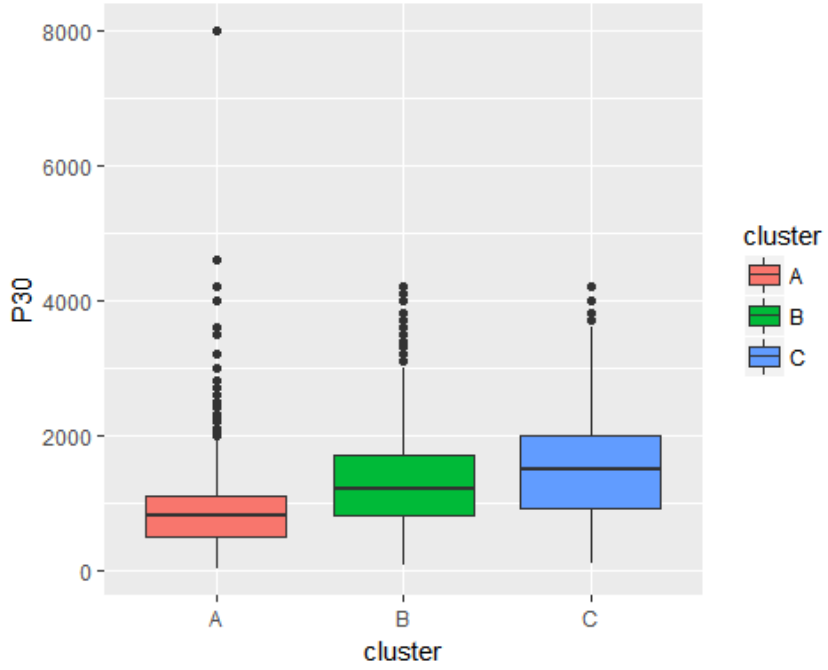


4.1.15 Costo último celular (P30)

En cuanto al costo del último celular que adquirieron, podemos observar que hay diferencias significativas en el gasto promedio que hicieron los diferentes grupos, siendo que el grupo A es el que en promedio gasto menos, mientras que los grupos B y C gastaron en promedio, montos más parecidos. Lo interesante en esta parte es observar que si bien en promedio el grupo A gastó menos, la variabilidad de este gasto es mayor en este grupo, llegando a tener observaciones altas, con valores parecidos a los de los otros grupos. Esto puede estar sugiriéndonos que hay personas en el grupo A que, a pesar de sus restricciones en los ingresos (vistas en un punto anterior), están dispuestas a gastar proporcionalmente más en celulares que los otros grupos. Esto puede deberse a que, si bien una característica del grupo A es que su mayoría no posee computadoras ni tablets, su principal medio de interacción con las redes sociales y la información sea justamente el teléfono celular, y por eso, probablemente, es que invierten proporcionalmente más.

```
## Resumen de estadísticos de posición
## $A
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.0   500.0   800.0   862.9  1100.0  8001.0
##
## $B
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      60     800    1200    1346    1700    4200
##
## $C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     100     900    1500    1591    2000    4200
```

Costo último celular por conglomerado

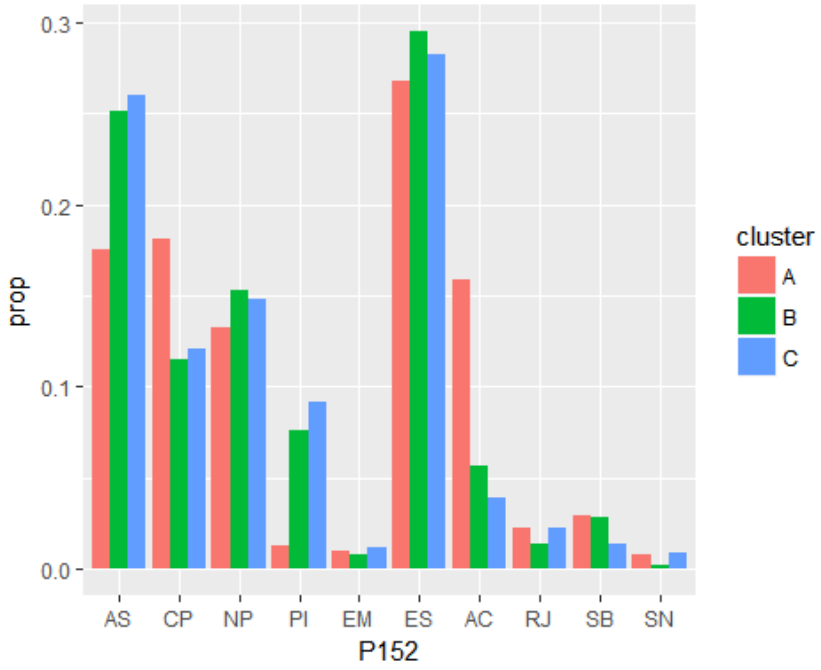


4.1.16 Categoría ocupacional (P152)

La conformación de los grupos en base a la categoría ocupacional, nos muestra que los 3 grupos están compuestos en su mayoría por estudiantes (A 27%, B 30%, C 28%). Al mismo tiempo es interesante ver algunas diferencias que existen entre los grupos respecto a algunas categorías ocupacionales. Por ejemplo para los trabajadores por cuenta propia el grupo A tiene al 18% de sus integrantes, contra un 12% para los grupos B y C, viendo una diferencia significativa. De manera similar una categoría ocupacional que diferencia a los grupos es justamente la de profesional independiente, donde el grupo A cuenta con solo el 1% de sus miembros, mientras que el B y C tienen al 8 y 9 por ciento de las personas respectivamente. Al mismo tiempo es interesante ver que para el grupo A la categoría de ama de casa representa al 16% de las entrevistadas, que es significativamente diferente a los otros grupos. Esta composición puede servirnos para explicar las diferencias en la estructura por sexos entre los grupos.

```
##   Frecuencias condicionales por conglomerado
##         A    B    C
##   AS 0.18 0.25 0.26
##   CP 0.18 0.12 0.12
##   NP 0.13 0.15 0.15
##   PI 0.01 0.08 0.09
##   EM 0.01 0.01 0.01
##   ES 0.27 0.30 0.28
##   AC 0.16 0.06 0.04
##   RJ 0.02 0.01 0.02
##   SB 0.03 0.03 0.01
##   SN 0.01 0.00 0.01
##
##   AS: Asalariado          ; CP: Cuenta propia
##   NP: Negocio propio      ; PI: Profesional indep.
##   EM: Empleador           ; ES: Estudiante
##   AC: Ama de casa         ; RJ: Rentista/Jubilado
##   SB: Sin empleo, buscando; SN: Sin empleo, no busca
```

Categoría ocupacional por conglomerado



4.2 Descripción de conglomerados respecto a otras variables de interés

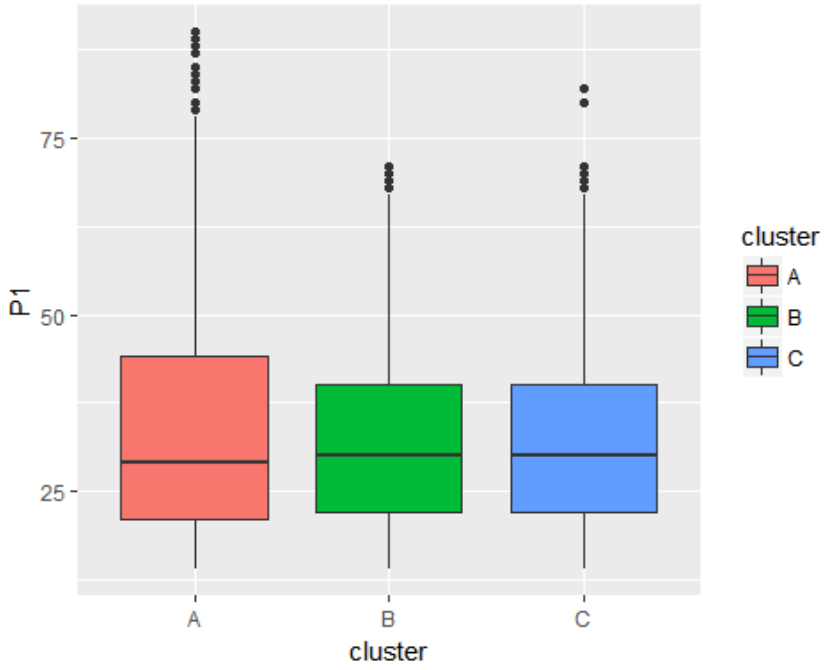
En la subsección anterior vimos algunas diferencias de las características y comportamientos de los grupos, principalmente con aquellas variables que se identificaron como importantes al momento de hacer la clasificación por K-Prototypes. Ahora vamos a analizar el comportamiento con otras variables de interés.

4.2.1 Edad (P1)

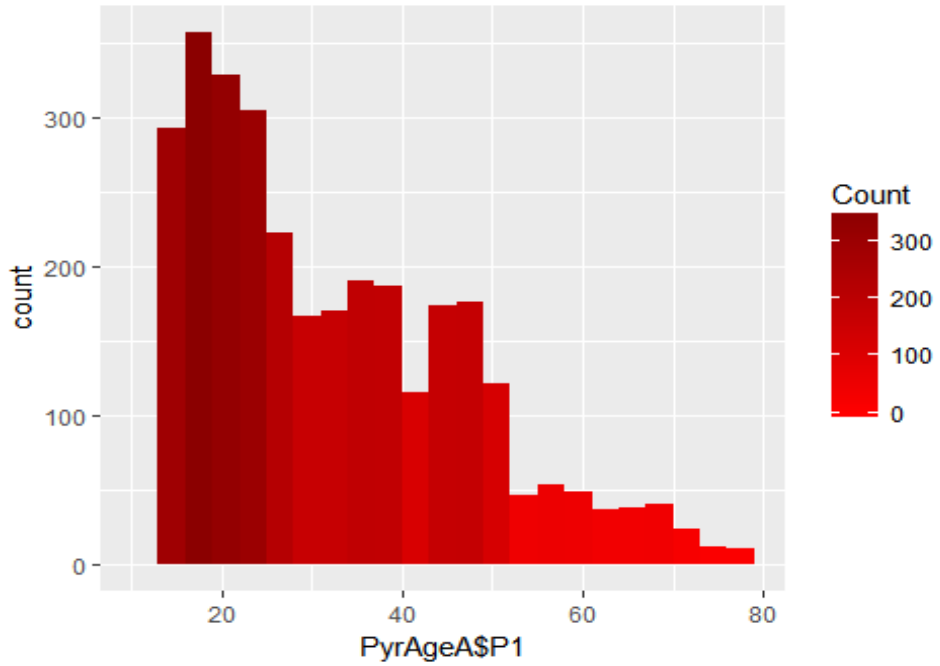
Como se puede observar en la gráfica, las edades promedio de los grupos son muy parecidas, rondan los 32 años, lo que nos muestra que tienen una estructura demográfica parecida. Al mismo tiempo, alguna de las diferencias que podemos obtener al observar los histogramas de las edades de cada grupo, es que dentro del grupo A hay más personas de más de 70 años, mientras que en los grupos B y C, este grupo etario es más reducido.

```
## Resumen de estadísticos de posición
## $A
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   21.00   29.00   33.16   44.00   90.00
##
## $B
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   22.00   30.00   31.86   40.00   71.00
##
## $C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   22.00   30.00   32.03   40.00   82.00
```

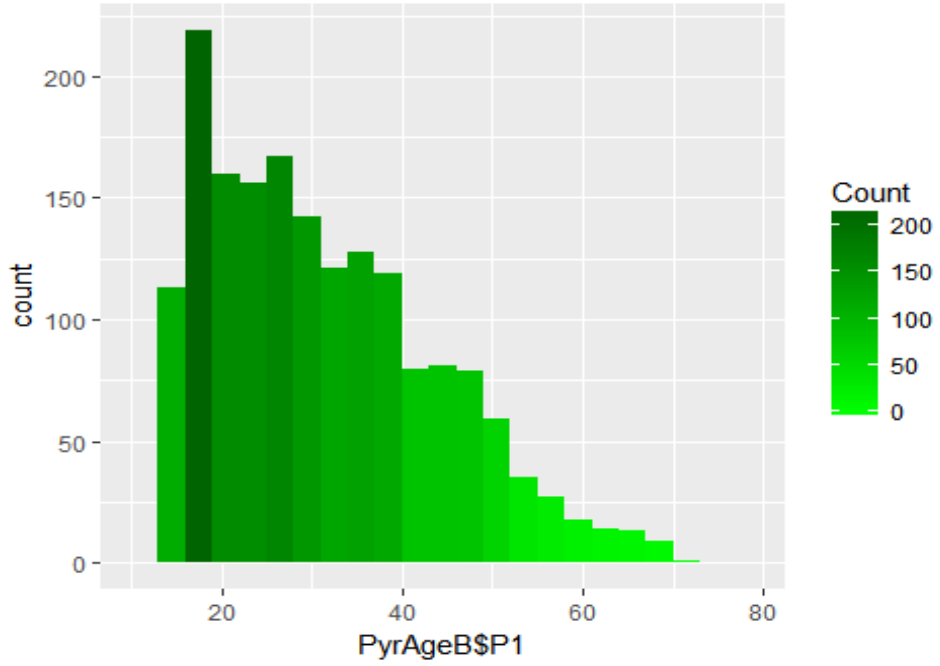
Edad por conglomerado



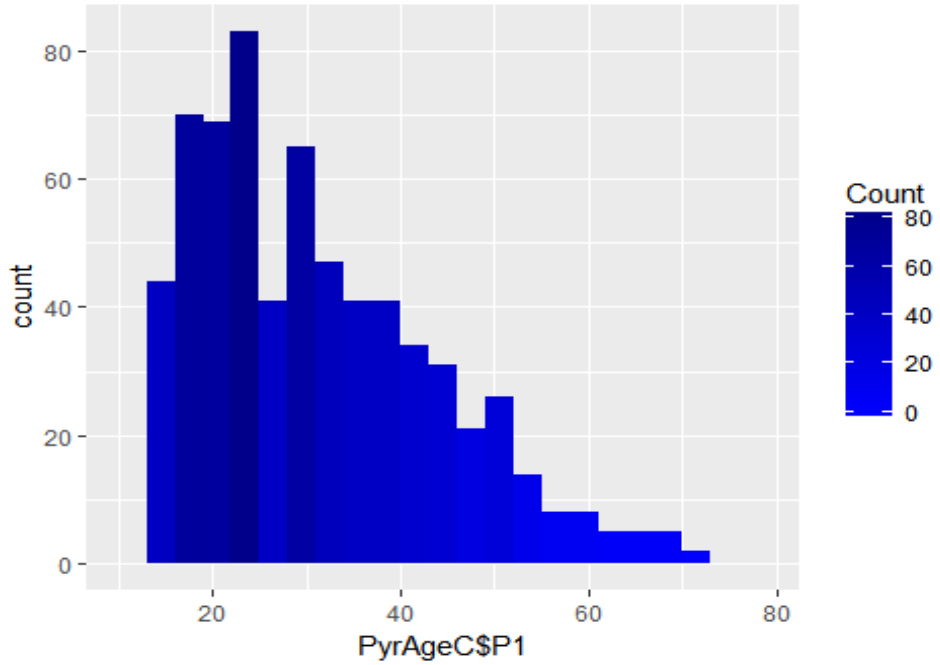
Histograma de Edad en el conglomerado A



Histograma de Edad en el conglomerado B



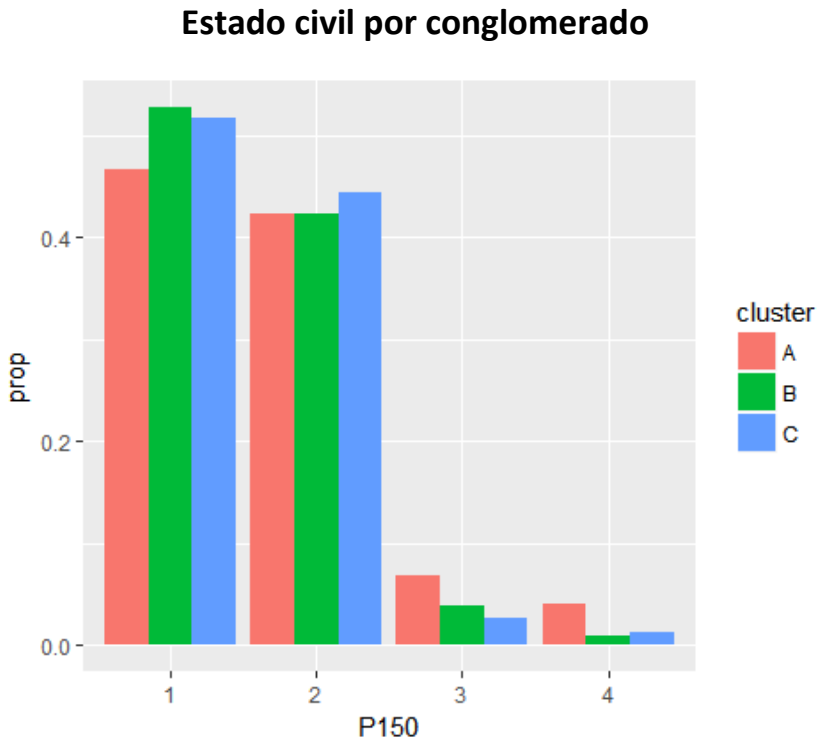
Histograma de Edad en el conglomerado C



4.2.2 Estado civil (P150)

Respecto al estado civil, podemos identificar que los tres grupos se conforman de una manera similar, siendo que el estado civil más representativo es ser soltero, seguido de casado o conviviente. El tema de soltero tiene coherencia con la estructura de edades y la ocupación principal de la mayoría de los encuestados, estudiantes, mientras que las segunda categoría, la de casado o conviviente, agrupa casi a todo el resto de encuestados.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.47 0.53 0.52
##   2 0.42 0.42 0.44
##   3 0.07 0.04 0.03
##   4 0.04 0.01 0.01
##
##   1: Soltero; 2: Casado/Conviviente; 3: Separado/Divorciado; 4: Viudo
```

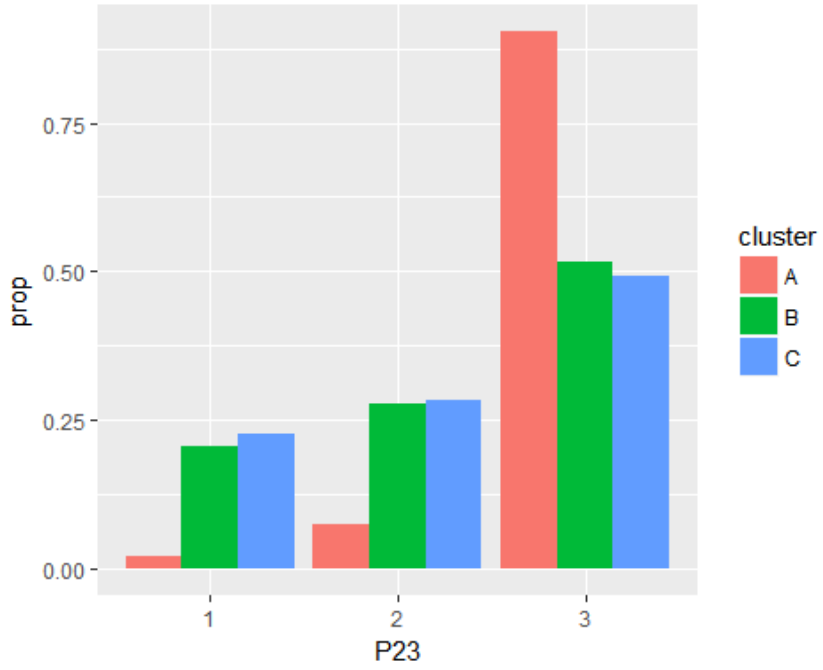


4.2.3 Tenencia de internet fijo en casa o módem (P23)

El acceso a internet en casa, tiene coherencia con la posesión de equipos de computación y además es una de las características donde encontramos más diferencias. EL 90% de las personas del grupo A no tienen acceso a internet en sus hogares juntamente con el 52% del grupo B y el 49% del grupo C. Este aspecto es interesante y puede ser uno de los factores que ayuden a explicar porque las personas de los grupos B y C están más conectados que las del grupo A. Además, si bien las diferencias entre las disponibilidades de internet en sus hogares de los grupos B y C no son grandes, observamos que algo que si diferencia a estos grupos es el uso que le dan a esta conexión, siendo que el grupo C está más conectado.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.02 0.21 0.23
##   2 0.08 0.28 0.28
##   3 0.90 0.52 0.49
##
##   1: Internet fijo; 2: Módem; 3: Ninguno
```

Tenencia de internet fijo en casa o módem por conglomerado

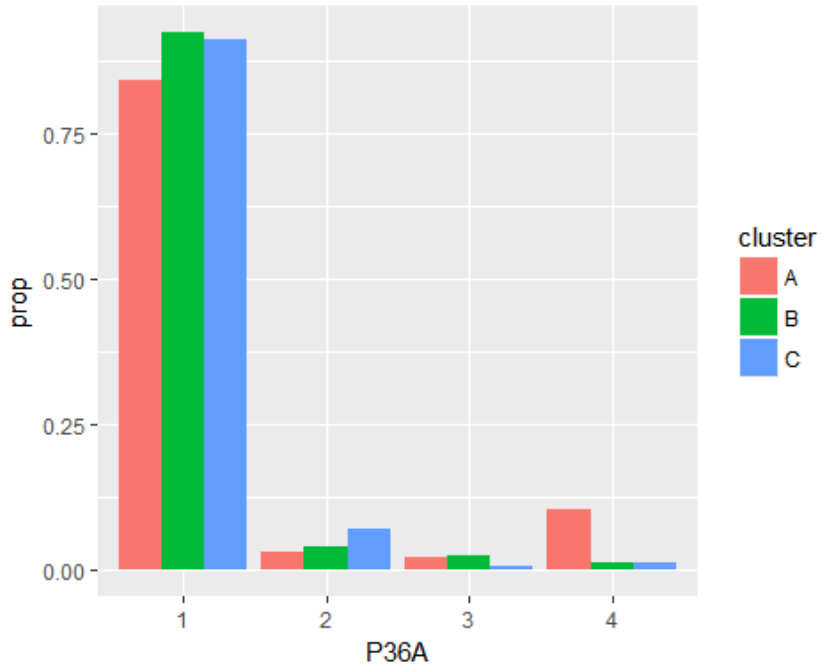


4.2.4 Sistema Operativo en celular (P36A)

El sistema operativo que utilizan los teléfonos móviles de los entrevistados nos muestra un claro dominio de Android, siendo la categoría más importante para los 3 grupos y llegando a abarcar el 88% del total de la encuesta. Por otra parte, se observa que el conglomerado A tiene menos participación en el uso de Android que los otros grupos, y al mismo tiempo llega a tener el 10% de sus integrantes en la categoría de otros. Esto nos muestra que en el grupo A hay más personas que utilizan celulares convencionales, es decir aquellos que no son smartphones, que utilizan otros sistemas operativos diferentes a Android y iOS. Esto también es coherente con el hecho de que en promedio este grupo gasta menos en sus teléfonos. Finalmente, si bien la cuota de mercado de iOS es baja respecto a los demás sistemas operativos, podemos observar que para el grupo C esta categoría es importante, abarcando el 7% del conglomerado.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.84 0.92 0.91
##   2 0.03 0.04 0.07
##   3 0.02 0.02 0.01
##   4 0.10 0.01 0.01
##
##   1: Android; 2: Apple; 4: Otros
```

Sistema Operativo en celular por conglomerado

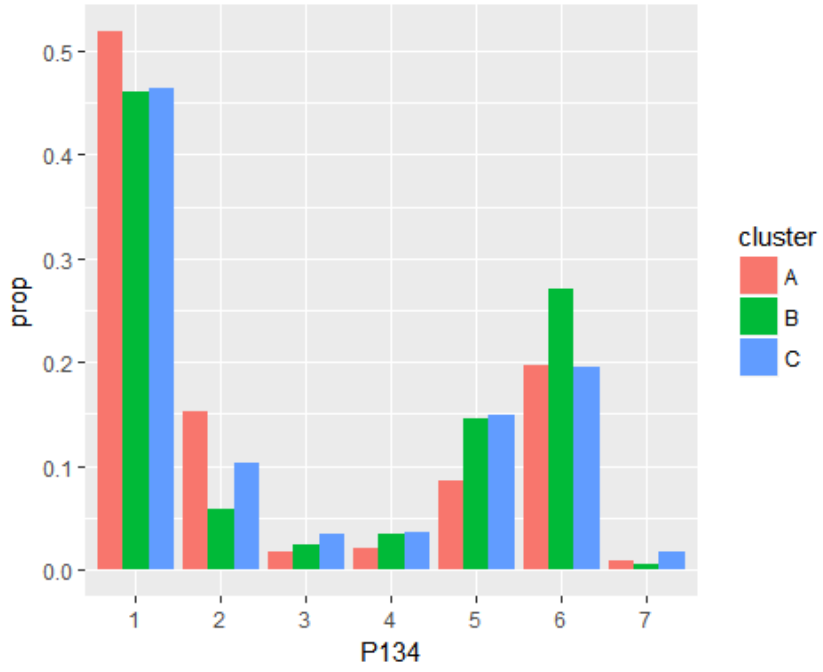


4.2.5 Consideración de medio más rápido (P134)

Las preferencias de los grupos respecto a qué medio de comunicación consideran más rápido son parecidas en rasgos generales. En los tres grupos existe una mayoría que considera que la televisión (categoría 1) es el medio más rápido, donde el grupo A resalta en esta opinión. Una segunda proporción importante en los tres grupos considera que las redes sociales (categoría 6) son el medio más rápido, y en esta opinión resalta el grupo B. Por otra parte, una tercera proporción importante de A considera que la radio es el medio más rápido mientras que las respectivas proporciones en los grupos B y C consideran más rápidas a las páginas web (categoría 5). Finalmente, se puede notar que si bien las cuartas proporciones más importantes de los grupos B y C consideran a la radio como medio más rápido, el grupo C resalta en esta opinión respecto a B.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.52 0.46 0.46
##   2 0.15 0.06 0.10
##   3 0.02 0.02 0.03
##   4 0.02 0.04 0.04
##   5 0.09 0.15 0.15
##   6 0.20 0.27 0.19
##   7 0.01 0.00 0.02
##
##   1: TV              ; 2: Radio      ; 3: Periódicos Imp.
##   4: Periódicos Dig.; 5: Páginas Web; 6: Redes Sociales
##   7: Ninguno
```

Consideración de medio más rápido por conglomerado



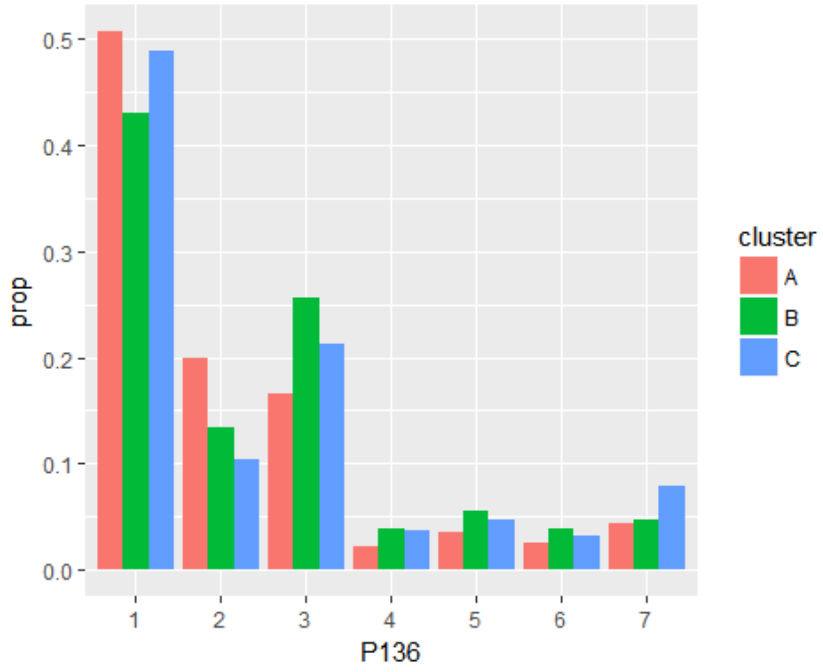
4.2.6 Consideración de medio más serio (P136)

Si bien las redes sociales son consideradas, después de la TV, como un medio rápido, estas en general no se consideran un medio serio entre los encuestados. Las perspectivas de las mayorías en los grupos nos muestran que la TV sigue siendo el medio de comunicación más serio. Por otra parte, se resalta que los periódicos son una fuente seria de información para proporciones de segunda importancia en los

grupos B y C. En cuanto a la radio, el 20% del grupo A la considera una fuente de información seria, mientras que esta proporción en los demás grupos ronda el 12%. Al mismo tiempo es interesante resaltar que el grupo C, se muestra en general más escéptico que los otros grupos respecto a la seriedad de los medios de comunicación, siendo que el 8% de las personas del grupo C consideran que ningún medio de comunicación es serio.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.51 0.43 0.49
##   2 0.20 0.13 0.10
##   3 0.17 0.26 0.21
##   4 0.02 0.04 0.04
##   5 0.04 0.06 0.05
##   6 0.03 0.04 0.03
##   7 0.04 0.05 0.08
##
##   1: TV             ; 2: Radio      ; 3: Periódicos Imp.
##   4: Periódicos Dig.; 5: Páginas Web; 6: Redes Sociales
##   7: Ninguno
```

Consideración de medio más serio por conglomerado

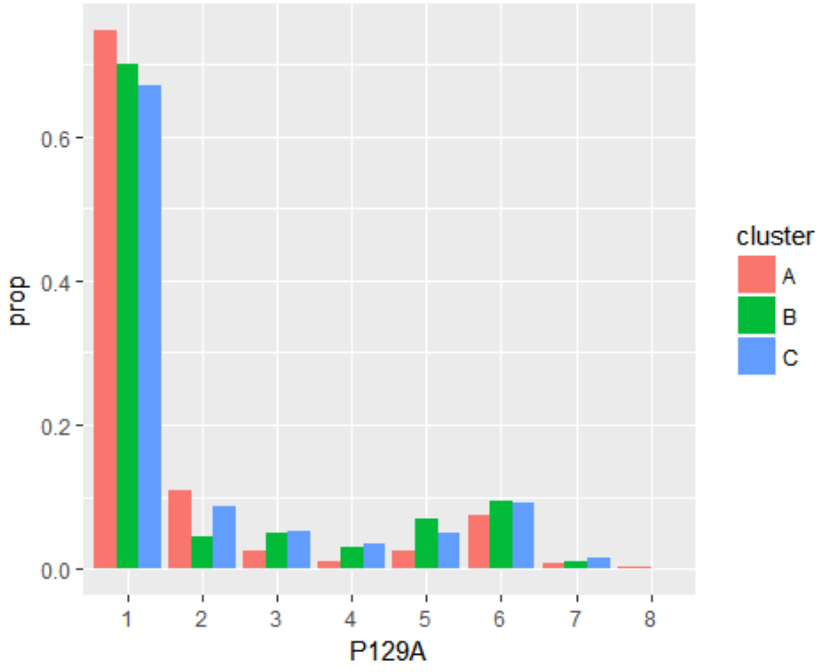


4.2.7 Medio preferido para informarse sobre noticias nacionales (P129A)

La televisión es uno de los medios mejor posicionados en varios aspectos, y en este caso, es el medio preferido para obtener noticias nacionales para los 3 grupos. También podemos observar que las personas del grupo A, después de la televisión prefieren la radio y como tercera fuente a las redes sociales, mientras que el grupo B tiene un comportamiento diferente, prefiere informarse por las redes sociales en vez de por la radio. El grupo C mantiene una relación muy parecida respecto a tener como fuentes de información nacional a las redes sociales y la radio.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.75 0.70 0.67
##   2 0.11 0.05 0.09
##   3 0.02 0.05 0.05
##   4 0.01 0.03 0.03
##   5 0.03 0.07 0.05
##   6 0.07 0.09 0.09
##   7 0.01 0.01 0.02
##   8 0.00 0.00 0.00
##
##   1: TV                ; 2: Radio         ; 3: Periódicos Imp.
##   4: Periódicos dig.   ; 5: Páginas Web   ; 6: Redes Sociales
##   7: charlas con amigos; 8: Otro
```

Medio preferido para informarse sobre noticias nacionales por conglomerado

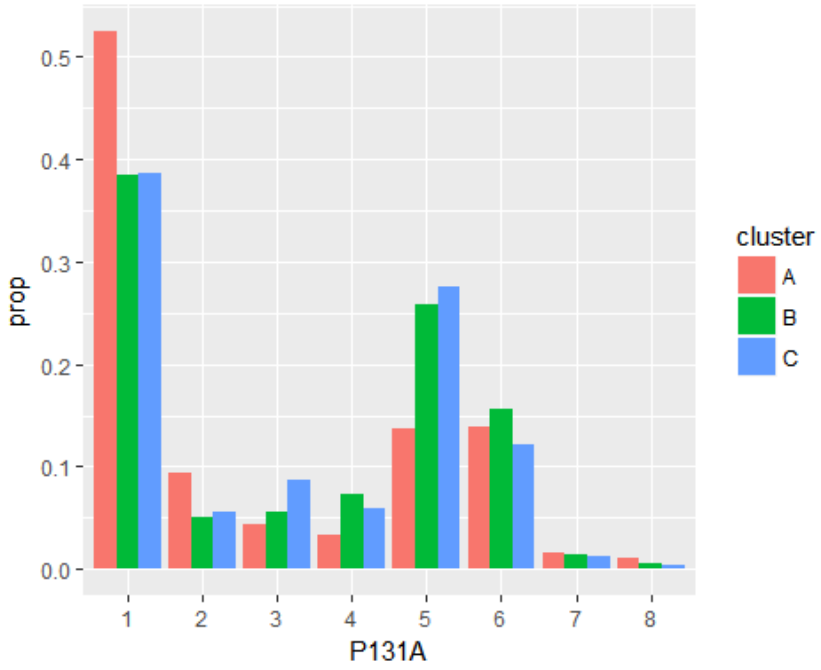


4.2.8 Medio preferido para informarse sobre ciencia y tecnología (P131A)

Después de la televisión, el medio preferido para informarse de ciencia y tecnología son las páginas web, seguido de las redes sociales. Esta preferencia por las páginas web puede deberse al hecho de que en el contexto nacional el contenido de ciencia y tecnología es muy escaso en los diferentes medios de comunicación, mientras que en las páginas web se pude acceder a este tipo de contenido de manera más fácil ya que a nivel internacional hay sitios especializados en estos aspectos. En cuanto a las redes sociales, se puede notar que estas suelen atraer la atención de los interesados a través de distintas comunidades sobre temas de ciencia y tecnología para luego funcionar como un puente hacia páginas web.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.53 0.39 0.39
##   2 0.09 0.05 0.06
##   3 0.04 0.06 0.09
##   4 0.03 0.07 0.06
##   5 0.14 0.26 0.27
##   6 0.14 0.16 0.12
##   7 0.02 0.01 0.01
##   8 0.01 0.01 0.00
##
##   1: TV                ; 2: Radio        ; 3: Periódicos Imp.
##   4: Periódicos dig.   ; 5: Páginas Web  ; 6: Redes Sociales
##   7: charlas con amigos; 8: Otro
```

Medio preferido para informarse sobre ciencia y tecnología por conglomerado

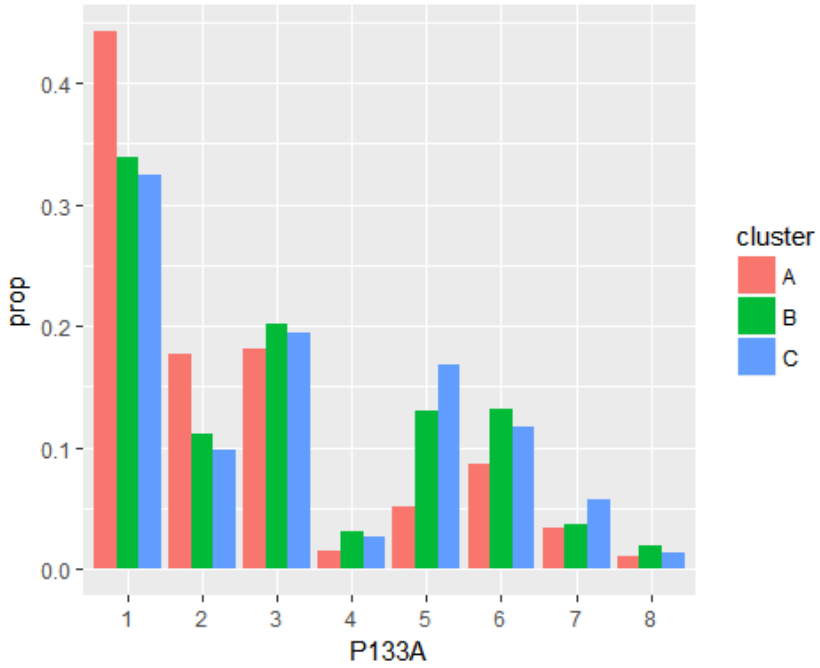


4.2.9 Medio preferido para informarse sobre negocios/oportunidades laborales/bienes o servicios (P133A)

En este aspecto, cabe resaltar la importancia que toman los periódicos impresos, siendo que llega a ser el segundo medio de comunicación preferido para informarse sobre negocios, oportunidades laborales, bienes o servicios. Otra vez un medio importante para el grupo A es la radio (18%), mientras que los grupos B y C prefieren ver páginas web y redes sociales.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.44 0.34 0.32
##   2 0.18 0.11 0.10
##   3 0.18 0.20 0.19
##   4 0.02 0.03 0.03
##   5 0.05 0.13 0.17
##   6 0.09 0.13 0.12
##   7 0.03 0.04 0.06
##   8 0.01 0.02 0.01
##
##   1: TV                ; 2: Radio         ; 3: Periódicos Imp.
##   4: Periódicos dig.   ; 5: Páginas Web   ; 6: Redes Sociales
##   7: charlas con amigos; 8: Otro
```

Medio preferido para informarse sobre negocios/oportunidades laborales/bienes o servicios por conglomerado

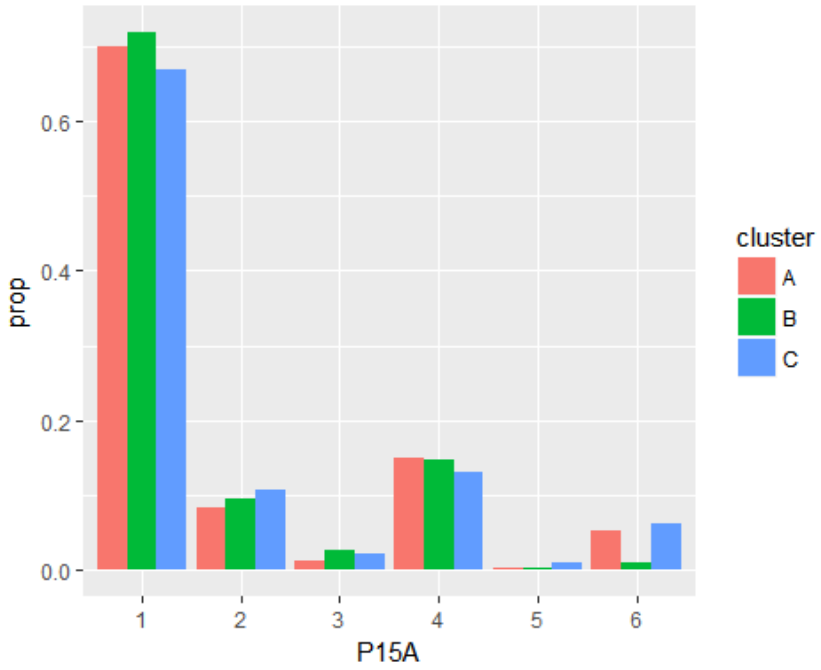


4.2.10 Tipo de programa preferido en televisión (P15A)

Podemos observar que uno de los tipos de programas más preferidos para ver en la televisión es un programa de noticias, para los 3 grupos. Un siguiente tipo de contenido que les interesa a las personas es el de entretenimiento, y como tercero programación de deportes. El comportamiento general se aplica a los 3 grupos, pero es interesante observar que tan solo el 1% del grupo B no ve televisión, frente a un 5% del grupo A y un 6% del grupo C, siendo este último grupo uno que más bienes duraderos posee, es decir que no ven televisión por elección propia y no por restricciones de no tener televisor.

```
##   Frecuencias condicionales por conglomerado
##        A    B    C
##   1 0.70 0.72 0.67
##   2 0.08 0.10 0.11
##   3 0.01 0.03 0.02
##   4 0.15 0.15 0.13
##   5 0.00 0.00 0.01
##   6 0.05 0.01 0.06
##
##   1: Noticias       ; 2: Deportes   ; 3: Salud y educación
##   4: Entretenimiento; 5: Otro       ; 6: No mira televisión
```

Tipo de programa preferido en televisión por conglomerado



4.3 Interpretación de los conglomerados

Haciendo una división de algunas de las variables que consideramos en nuestro análisis podemos observar lo siguiente:

- **División de bienes duraderos**

  Los bienes duraderos son aquellos que no se renuevan constantemente, por ejemplo un computador, este, se espera que tenga alrededor de unos 4 años de vida útil en un contexto doméstico, mientras que los bienes no duraderos son aquellos que por sus características, se renuevan con mayor frecuencia, por ejemplo los alimentos. Hacemos énfasis en esta división porque notamos que las características de nuestros conglomerados aparentemente tienen una relación con estas variables. En los trabajos de Cortinovis 1993 y Filmer & Pritchett 2001 se utilizan algunos bienes duraderos para hacer una clasificación de estatus socio económico de las familias. En este aspecto el grupo A se diferencia por no poseer varios de los bienes duraderos considerados, mientras que los grupos B y C tienen indicadores muy parecidos y estos si poseen (en una proporción significativamente mayor que A), bienes duraderos.

- **División de servicios TIC**

  Al mismo tiempo los servicios de TIC como ser TV cable en el hogar o internet domiciliario también marcan diferencias significativas entre las características del grupo A respecto a los grupos B y C.

- **Características hogar**

  En las características del hogar, como ser el nivel de instrucción del jefe del hogar, o los ingresos promedios reportados, tenemos diferencias importantes entre el grupo A frente a los grupos B y C. Otra vez hay una marcada división.

- **Intensidad en uso de la TIC**

  Respecto al uso de las TIC tenemos diferencias entre los grupos, tanto A como B y C presentan comportamientos diferentes en esta "intensidad" en el uso de las TIC. Unos tienen restricciones para poder utilizarlas, otros prefieren utilizar redes sociales, y los demás prefieren diversificar sus fuentes de información.

Dado que la clasificación que hicimos a través de K-Prototypes es no supervisada, los grupos encontrados no tenían a priori una interpretación. Pero una vez analizadas las variables podemos identificar que parece haber 2 niveles de clasificación, el primero divide al grupo A de los grupos B y C, esta primera división podemos interpretarla como una clasificación de personas por "estatus socio-económico". En esta división, siguiendo la línea del trabajo de Cortinovis 1993, se tiene en el grupo A personas de escasos recursos respecto a los de la categoría B y C. Esto se ve reflejado en variables como la posesión de bienes durables, niveles de educación del jefe del hogar, gasto en el último televisor, etc.

La segunda división, aquella que separa a los grupos B y C, parece ser la interacción y uso que hacen de las TIC. Es decir, si bien el grupo B y C tienen proporciones de acceso a las TIC parecidas (a favor del grupo C), es justamente el conglomerado C el que utiliza estos medios con una mayor intensidad (frecuencia de uso), y con preferencias más diversas respecto a los fines para cada una de estas tecnologías. Mientras que el grupo B está más enfocado en las redes sociales.

En los siguientes puntos detallaremos estás diferencias, características y perfiles encontrados entre los grupos.

4.3.1 Conglomerado A: Mayoría

El conglomerado A agrupa al 56.59% de la población, este grupo está caracterizado por tener limitantes al momento de interactuar con las TIC. Uno de los motivos de estas diferencias es que en su mayoría no poseen en su hogar ni una computadora de escritorio, portátil o tablet, con la que podrían acercarse más a estas tecnologías. Al mismo tiempo, el 90% de las personas del grupo no tienen acceso a internet fijo ni

modem en su hogar. Probablemente uno de los principales motivos de estas restricciones sea el ingreso y riqueza del hogar; la mayoría de las personas de este grupo no tiene bienes durables (qué suelen ser proxys de riqueza) como, microondas (94% no tiene), lavadora de ropa (90% no tiene), etc. En promedio son los que menos gastaron en la compra de sus últimos celulares, al igual que al adquirir su último televisor. Además el nivel de instrucción de la mayoría de jefes(as) de hogar es menor que de los otros grupos, siendo que el 73% de los jefes de hogar del grupo tienen la secundaria completa o algún grado menor de instrucción. Respecto a los ingresos mensuales promedio del hogar declarados acumulan el 79% de familias que mensualmente tienen Bs 3.500 o menos.

Estas personas tienen preferencias por medios de comunicación más tradicionales como ser la TV y la radio. No interactúan mucho con la tecnología (el 68% no usa equipos de computación ni tablets); y aquellos del grupo que si lo hacen, tienen una frecuencia baja, el 42% de las personas que si utilizan estos equipos lo hacen de 1 a 3 veces por semana, nuevamente por la limitante que no tienen estos equipos en casa y para acceder a ellos deben hacer esfuerzos como acudir a cafés internet. También es importante resaltar que es el grupo que cuenta relativamente con más cuenta propistas respecto al total del grupo. Respecto a su distribución geográfica, si bien este grupo esta principalmente concentrado en centros poblados de 2.000 a 10.000 habitantes (40% del conglomerado), su participación no es trivial en las ciudades capitales (32% de conglomerado) ni en ciudades intermedias (28% del conglomerado).

Uno de los comportamientos más interesantes es que si bien las personas de este grupo gastaron en promedio menos en la compra de su último celular, que los otros grupos, este conglomerado en particular presenta una mayor variabilidad en esta variable. Una de las hipótesis que podría explicar esto, es que justamente las personas de este grupo que quieren interactuar con los TIC, lo hacen preferentemente a través de sus teléfonos celulares, y para tener una mejor experiencia es que invierten más en este bien ya que en su mayoría no tienen acceso a computadores ni internet fijo en su hogar.

4.3.2 Conglomerado B: Intermedio

Este grupo aglutina al 31.45% de la población, dentro de este conglomerado, el 86% de las personas tienen al menos un computador de escritorio, portátil o tablet en su hogar. Respecto al nivel de instrucción del jefe del hogar tenemos que la categoría más importante es la de universidad terminada, pero algo que diferencia a este grupo es el hecho que tiene un 17% de jefes de hogar con nivel de instrucción técnico. Consideran a las redes sociales como uno de sus principales medios de comunicación después de la televisión, y las consideran el medio más participativo. El 27% de las personas del grupo B consideran a las redes sociales como el medio de comunicación más rápido después de la TV. Además, para informarse de noticias nacionales, después de la televisión como principal fuente, el 9% prefiere las redes sociales y un 7% las páginas web.

En este grupo el 58% de las personas utiliza los equipos de computación o tablets entre 3 a 7 días a la semana, con un 22% de personas que los usan todos los días.

Es curioso encontrar que justamente el grupo que más énfasis hace en las redes sociales como un medio de comunicación importante, pase menos tiempo usando equipos de computación respecto al grupo C, a pesar de que tienen oportunidades de acceso parecidas. Parece ser que las personas del grupo B están más "entusiasmadas" con la conexión a las redes sociales, y a pesar de que se conectan menos, le dan más importancia a este medio en general.

4.3.3 Conglomerado C: Minoría

El conglomerado C tiene al 11.96% de la población, al igual que el grupo B, este grupo tiene acceso a los bienes durables, el 81% de las personas del grupo tienen un computador de escritorio, portátil o tablet en su hogar. En el grado de instrucción del jefe del hogar, si bien la mayor frecuencia es la de universidad completa, el comportamiento interesante es que en este grupo están más personas con posgrados, el 10% de los jefes de hogar de las familias del grupo tiene un posgrado, frente a un 6% del grupo B. Otra característica de este conglomerado es el hecho de que son los más escépticos respecto a la imparcialidad de los medios de comunicación, siendo que el 20% de las personas del grupo consideran que ningún medio de comunicación es imparcial. Al mismo tiempo, el 8% de las personas del grupo consideran que ningún medio de comunicación es serio. Para informase de noticias nacionales prefieren a la televisión al igual que los otros grupos, pero se diferencian del grupo B porque además prefieren a la par otras fuentes para informarse de noticias nacionales, redes sociales (9%) y radio (9%), lo que nos muestra una diversificación en las fuentes que consultan las personas de este grupo. En otros aspectos, para este grupo los periódicos impresos y la radio son también medios de comunicación alternativos a la televisión y las redes sociales.

En la distribución geográfica del grupo se observaron concentraciones en La Paz (22%) y Cochabamba (15%), un aspecto interesante ya que los demás grupos no mostraban una predominancia de locación geográfica por departamentos. Justamente Cochabamba es uno de los centros tecnológicos más importantes de Bolivia y La Paz, al ser sede de gobierno, reúne gran parte del capital humano del país.

Finalmente, una de las diferencias más importantes respecto al grupo B es la intensidad del uso de las TIC. El 62% de las personas de este grupo utilizan equipos de computación o tablets entre 3 a 7 días a la semana, siendo que el 31% de las personas los usan todos los días.

5. Conclusiones

Fue posible identificar conglomerados en la base de datos siguiendo en primera instancia una etapa de pre-procesamiento seguida por una etapa de modelación donde se consideraron variables categóricas y numéricas. En la etapa de modelación se lograron identificar tres conglomerados considerando un problema de aprendizaje no supervisado, y posteriormente, planteando un problema de aprendizaje supervisado, se logró encontrar un subconjunto de variables importantes en la definición de los mismos. Para resolver el problema de aprendizaje no supervisado se consideró el modelo K-Prototypes y para el problema de aprendizaje supervisado el modelo Random Forest.

En cuanto a los resultados, a pesar de que nuestra clasificación por conglomerados no tenía un criterio separador a priori, pudimos identificar 3 tipos de categorías claramente diferenciadas y detalladas en la sección anterior.

La clasificación obtenida fue a 2 niveles, el primero, diferenciando a aquellas personas que tienen restricciones estructurales para poder acceder a la información y la tecnología. Probablemente estas restricciones se deban a un tema de escasos recursos, pero dentro de este grupo, dadas estas restricciones, se identifican personas que hacen un esfuerzo en acceder a medios que les permitan conectarse a las TIC, y en este contexto invierten en "smartphones".

El segundo nivel de clasificación se da entre aquellas personas que tienen la posibilidad de tener acceso a tecnologías como computadoras y tablets dentro de sus hogares, además de complementarlas con conexiones a internet. Dentro de este conjunto se diferencian dos grupos, uno de ellos que tiene una preferencia marcada por las redes sociales como medios de comunicación, frente a otro grupo que es más cauteloso y escéptico respecto a la confiabilidad de las fuentes y prefiere diversificarlas para obtener mayor información.

6. Recomendaciones

6.1. Recomendaciones metodológicas.

Es recomendable que en la base de datos se aplique un proceso de ingeniería de variables para incorporar variables que aporten información o incluso puedan marcar mayores diferencias entre los conglomerados.

Para la selección de variables importantes se recomienda obtener medidas de importancia condicionales. En el presente estudio no se logró hacer este análisis por una falta de capacidad computacional en los equipos usados.

En cuanto al análisis de valores faltantes, es recomendable hacer un tratamiento más minucioso al respecto. Haciendo referencia a la primera recomendación, podrían generarse variables indicatrices que aporten información respecto al grado de confianza o voluntad de los encuestados a la hora de responder preguntas.

En cuanto a los resultados, se recomienda hacer un contraste entre distribuciones de variables que también se consideran en las encuestas de hogares realizadas por el INE, ya que en general, se esperaría encontrar similitudes. Un hallazgo que permite plantear esta recomendación es la alta concentración de estudiantes entre los encuestados.

6.2. Recomendaciones de política.

El conglomerado que presenta restricciones en el acceso a medios para interactuar con la tecnología e información de las TIC, representa una mayoría dentro del contexto nacional (56.59%). La labor pendiente en este tema es llevar a cabo políticas públicas para poder democratizar el acceso a estos medios. Dar la posibilidad a las personas de este grupo a que puedan acceder a oportunidades para desarrollar habilidades tecnológicas.

La política de democratización de las TIC, debe ser desarrollada con un enfoque crítico, para que las personas que se interioricen en el tema, puedan consultar varias fuentes al momento de tomar decisiones, y al mismo tiempo no se sientan abrumadas con la cantidad de información disponible en la red.

6.3 Líneas de investigación.

Dentro de la exploración de esta base de datos y con el análisis de los conglomerados, se plantearon hipótesis interesantes como: el nivel de instrucción influye en el escepticismo sobre la imparcialidad y seriedad de los medios de comunicación; las personas de escasos recursos interesadas en temas de tecnología hacen un esfuerzo mayor para poder acceder a medios con los que puedan conectarse y navegar por internet, principalmente celulares; la televisión es el medio de comunicación más importante para las personas, pero medios alternativos como las redes sociales van cobrando fuerza. Por otra parte, se plantean interrogantes interesantes como: ¿Serán

representativos los resultados en la población?; ¿Cuál sería el comportamiento de un cuarto conglomerado?; y ¿Cuáles serían los resultados si se consideran otros métodos de selección de variables importantes? Estás hipótesis e interrogantes pueden ser desarrolladas en profundidad en posteriores trabajos y aportar a la línea de investigación sobre TIC en Bolivia.

7. Referencias

- Cortinovis I. , Vella V. & Ndiku J. (1993). “Construction of a socio-economic index to facilitate analysis of health data in development countries”. Social Science and Medicine 36: 1087 – 1097.
- Filmer D, Pritchett LH. (2001). “Estimating wealth effect without expenditure data – or tears: an application to educational enrollments in states of India”. Demography 38: 115–32.
- Ismaili O.A., Lemaire V., Cornuéjols A. (2014). "A supervised methodology to measure the variables contribution to a clustering". International Conference on Neural Information Processing: 159-166.
- Strobl, Carolin & Hothorn, Torsten & Zeileis, Achim. (2009). Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package. The R Journal. 1. 14-17.
- Van Buuren, Stef & Groothuis-Oudshoorn, Catharina. (2011). MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 45. 10.18637/jss.v045.i03.
- Zhexue Huang. (1998). "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values". Data Mining and Knowledge Discovery 2: 283 – 304.