Pregunta 3

Javier Saavedra

20-10-2021

1. Preparación de la data

Cargamos la información y separamos la data entre características del encuestado y los puntajes de cada pregunta del cuestionario

```
Datos_prueba_Fix <- read.csv2(file = "Datos prueba Fix.csv",
                              sep = ";")

dataD <- Datos_prueba_Fix[, c(paste0("D", 1:8))]
dataP <- Datos_prueba_Fix[, c(paste0("P", 1:28))]
```

Podemos apreciar que existe una cantidad importante de valores atipicos para las respuestas de cada pregunta del diccionario que no están dentro del rango contemplado (de 1 a 5). Corregiremos algunos de estos valores en base a ciertas reglas de interpretación.

```
valores_unicosP <- unique(unlist(unname(apply(dataP, 2, unique))))
valores_unicosP <- unique(gsub(" ", "", valores_unicosP))
valores_unicosP
```

```
##  [1] "4"       "3"       "2"       "1"       NA        "5"       "10000"
##  [8] "a"       "b"       ""        "2000"    "4000000" "7"       "4000"
## [15] "5000"    "-1"      "500"     "8"       "400"     "-3"      "-2"
## [22] "12"      "400000"  "-4"      "10"      "50000"   "9"       "40000"
```

Creamos una función para reemplazar valores atípicos en base a un mapeo entre los caracteres extraños y el valor interpretado por nuestra parte. Posteriormente se imputaran los valores faltantes por el valor 3, ya que dentro de las encuestas en la escala de Likert este valor representa una respuesta neutra por parte del participante. Finalmente, filtraremos un total de 23 filas que presentan valores nulos para las caracteristicas de los encuestados (18 observaciones) o que presenten valores fuera de un rango definido (5 observaciones en el campo de edad), quedando así con un conjunto final de 1975 observaciones (23 menos con respecto al original).

```
reemplaza_valores <- function(x) {

  x.char <- as.character(x)
  x.char <- stringr::str_replace(x.char, " ", "")
  x.replace <- str_replace_all(x.char, vector_reemplazo)
  x.numeric <- as.numeric(x.replace)
```

```
  return(x.numeric)
}

vector_reemplazo <- c("4000000$" = "4",
                      "400000$" = "4",
                      "40000$" = "4",
                      "4000$" = "4",
                      "400$" = "4",
                      "50000$" = "5",
                      "5000$" = "5",
                      "500$" = "5",
                      "2000$" = "2",
                      "10000$" = "1",
                      "5$" = "5",
                      "4$" = "4",
                      "3$" = "3",
                      "2$" = "2",
                      "1$" = "1",
                      "a$" = "2",
                      "b$" = "2",
                      "-1$" = "1",
                      "-3$" = "3",
                      "-2$" = "2",
                      "-4$" = "4",
                      "7$" = "",
                      "8$" = "",
                      "12$" = "",
                      "10$" = "",
                      "9$" = "")
```

Finalmente, generamos un conjunto de datos compuesto por las dos variables dependientes de interés (satisfacción y recomendación), variables descriptivas de cada sujeto y una última variable que correspondera a la suma del resultado de cada pregunta del encuestado escalado entre 0 y 100 (sin contar la P18 y P19). El enfoque de sumar los puntajes de todas las preguntas es algo común en los campos de psicometría y permite generar variables continuas, lo cuál tiene sus ventajas al momento de usar los modelos típicos. Una mejora que se podría realizar en este caso es agrupar las preguntas en base a ciertos tópicos, por ejemplo preguntas relacionadas a la calidad de los productos ofrecidos por la tienda, preguntas relacionadas a la calidad del servicio o preguntas relacionadas a la calidad de la postventa. Por el momento solo generaremos una única variable que represente una evaluación de la experiencia del cliente.

```
dataP.clean <- as.data.frame(sapply(dataP, reemplaza_valores))
dataP.clean.nonan <- dataP.clean %>% replace(is.na(.), 3)

data.clean <- cbind(dataD, dataP.clean.nonan)
data.clean.na <- na.omit(data.clean)

Y1 <- data.clean.na$P18
Y2 <- data.clean.na$P19

D <- data.clean.na[, 1:8]
D <- as.data.frame(sapply(D, as.factor))

Q <- data.clean.na[, 9:36]
```

```r
Q <- Q[, -c(18,19)]
dataQ <- apply(Q, 1, sum)/(ncol(Q)*5)

data.final <- cbind(Y1, Y2, D, dataQ)
data.final <- data.final %>%
  dplyr::filter(D1 != 3) %>%
  dplyr::mutate(
    dataQ = dataQ*100
  )
data.final$Y1 <- factor(data.final$Y1, levels = c("1", "2", "3", "4", "5"))
data.final$Y2 <- factor(data.final$Y2, levels = c("1", "2", "3", "4", "5"))
colnames(data.final) <- c("satisf",
                          "recomen",
                          "genero",
                          "edad",
                          "nacion",
                          "resid",
                          "sosten",
                          "ingreso",
                          "frec",
                          "horario",
                          "puntaje")
head(data.final)
```

```
##   satisf recomen genero edad nacion resid sosten ingreso frec horario  puntaje
## 1      5       4      1    2      1    14      2       1    3       3 76.15385
## 2      4       4      1    2      1     7      2       3    4       3 63.84615
## 3      4       5      1    6      1     7      2       4    1       2 65.38462
## 4      5       5      1    5      1     7      1       3    3       3 69.23077
## 5      4       5      1    2      1     3      2       5    3       1 69.23077
## 6      4       5      1    3      1     7      1       1    3       1 63.07692
```

2. Análisis exploratorio

Relación entre edad, satisfacción y posibilidad de recomendación

Podría existir una relación entre la edad del cliente y la evaluación de satisfacción y recomendación. La siguiente tabla muestra un promedio de cada evaluación por rango de edad y la diferencia con respecto a un promedio general. Se puede observar que para la evaluación de satisfacción los puntajes promedio de los rangos de edad 1, 2 y 3 están por sobre el promedio general y para el resto de los rangos de edad los puntajes estan por debajo de lo general, lo mismo ocurre para la evaluación de recomendación.

```
tabla1 <- data.final %>%
  dplyr::group_by() %>%
  dplyr::mutate(promedio.general.satisf = round(mean(as.numeric(satisf)), 3),
                promedio.general.recomen = round(mean(as.numeric(recomen)), 3)) %>%
  dplyr::group_by(edad) %>%
  dplyr::summarise(casos = n(),
                   promedio.satisf = round(mean(as.numeric(satisf)), 3),
                   diff.satisf = promedio.satisf - mean(promedio.general.satisf),
                   promedio.recomen = round(mean(as.numeric(recomen)), 3),
                   diff.recomen = promedio.recomen - mean(promedio.general.recomen))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

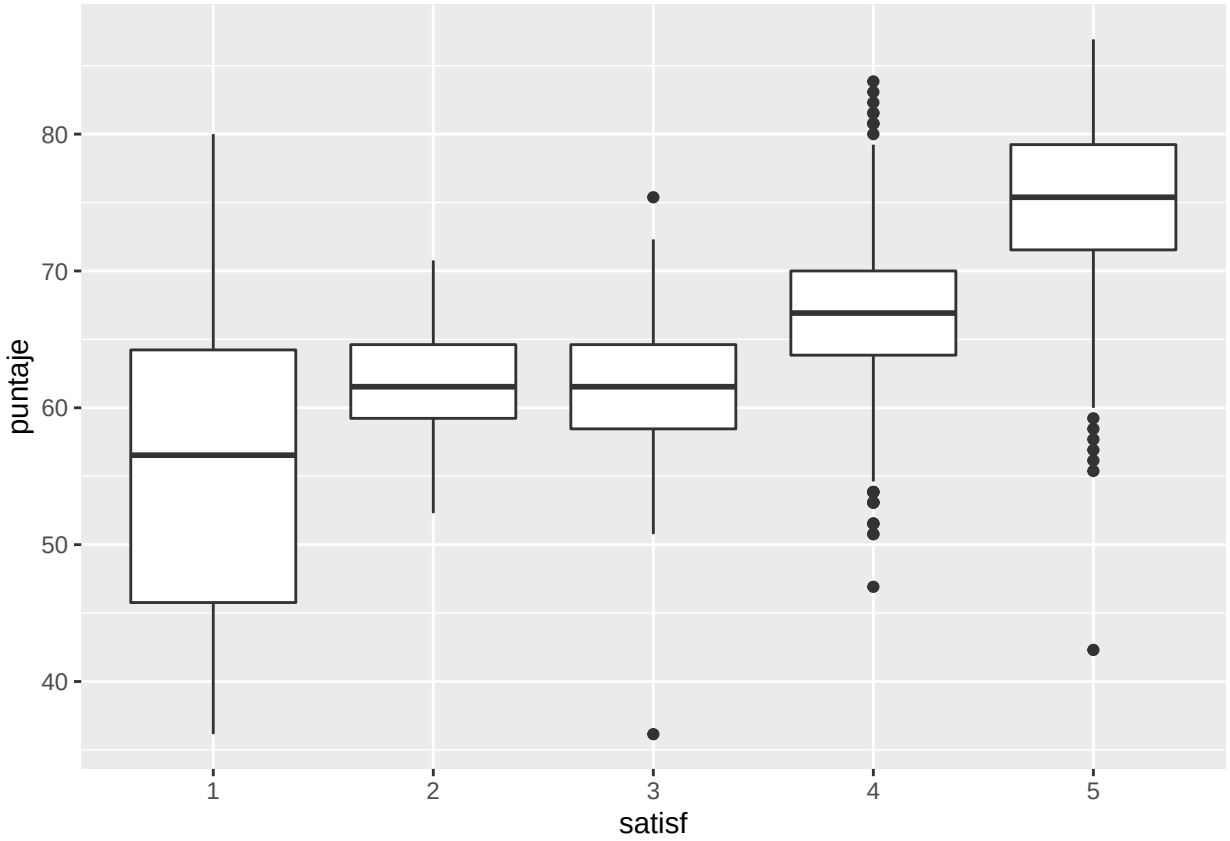
```
knitr::kable(tabla1)
```

edad	casos	promedio.satisf	diff.satisf	promedio.recomen	diff.recomen
1	135	4.452	0.303	4.622	0.274
2	471	4.268	0.119	4.499	0.151
3	417	4.151	0.002	4.372	0.024
4	558	3.993	-0.156	4.213	-0.135
5	279	4.118	-0.031	4.219	-0.129
6	83	4.084	-0.065	4.313	-0.035
7	32	4.250	0.101	4.250	-0.098

Relación entre el puntaje de la encuesta y satisfacción

Otro hallazgo interesante es la relación que puede existir entre el puntaje total de la encuesta, lo cual es bastante intuitivo ya que si una persona responde positivamente al resto de las preguntas lo más probable es que su puntaje de recomendación y de satisfacción sea elevado

```
data.final %>%
  ggplot(aes(x = satisf, y = puntaje)) +
  geom_boxplot()
```




```
  # facet_grid(.~genero)
```

Relación entre satisfacción y frecuencia de compra

Finalmente concluimos con una interacción interesante que ocurre entre la evaluación de satisfacción y la frecuencia de compra, donde apreciamos que la frecuencia 1 concentra los mejores puntajes de satisfacción y estos van disminuyendo a medida que la frecuencia aumenta.

```
tabla2 <- table(data.final$satisf, data.final$frec, dnn = c("Satisfacción",
                                                            "Frecuencia"))
tabla2 <- round(prop.table(tabla2, 2)*100, 1)
tabla2
```

```
##             Frecuencia
## Satisfacción    1    3    4
##            1  0.6  0.2  0.9
##            2  0.2  1.0  3.5
##            3  7.2  9.3 17.5
##            4 49.3 64.8 67.1
##            5 42.8 24.6 11.0
```

3. Modelamiento de satisfacción y recomendación

3.1. Satisfacción - Selección de variables

Dado que la evaluación de satisfacción y recomendación son de carácter categórico y ordinal, donde la categoría "Muy de acuerdo" es mayor que "De acuerdo" y así con el resto de las categorías, ajustaremos un modelo logístico para categorías ordenadas (probit). Este modelo compara razones de chance de pertenecer a una categoria con respecto a la siguiente (ordenadamente).

Partimos ajustando un modelo con todas las variables para detectar cuales son o no significativas para el modelo. A partir de la siguiente tabla podemos concluir que las variables significativas son el género, la edad, frecuencia y puntaje, por lo que ajustaremos un modelo con estas variables e interpretaremos los resultados.

```
formula1 <- as.formula(
  "satisf~genero+edad+nacion+resid+sosten+ingreso+frec+horario+puntaje")
modelo.polr <- MASS::polr(data = data.final,
                          formula = formula1,
                          Hess = TRUE)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
coef.table <- coef(summary(modelo.polr))
p <- pnorm(abs(coef.table[, "t value"]), lower.tail = FALSE) * 2
coef.table <- cbind(coef.table, "p value" = p)
coef.table <- as.data.frame(coef.table)
coef.table$significant <- ifelse(coef.table$`p value` > 0.05, FALSE, TRUE)
knitr::kable(coef.table)
```

	Value	Std. Error	t value	p value	significant
genero2	-0.4533886	0.1579154	-2.8710846	0.0040907	TRUE
edad2	-0.4326648	0.2382621	-1.8159194	0.0693827	FALSE
edad3	-0.5654454	0.2507298	-2.2551986	0.0241209	TRUE
edad4	-0.8606906	0.2547913	-3.3780215	0.0007301	TRUE
edad5	-0.4792253	0.2701994	-1.7735984	0.0761296	FALSE
edad6	-0.5073193	0.3503120	-1.4481927	0.1475632	FALSE
edad7	0.0131672	0.4736193	0.0278012	0.9778207	FALSE
nacion2	0.3566082	0.2383637	1.4960674	0.1346361	FALSE
resid10	0.2092453	0.5738729	0.3646196	0.7153954	FALSE
resid11	1.2550223	0.7200689	1.7429197	0.0813476	FALSE
resid12	0.3039006	0.6118997	0.4966509	0.6194352	FALSE
resid13	0.7171407	0.6838503	1.0486809	0.2943250	FALSE
resid14	1.0522599	0.6264083	1.6798307	0.0929903	FALSE
resid15	1.8323481	1.3224621	1.3855581	0.1658819	FALSE
resid16	1.2236636	0.9004586	1.3589339	0.1741675	FALSE
resid2	0.6985775	0.6848388	1.0200611	0.3076995	FALSE
resid3	0.5686061	0.6377822	0.8915365	0.3726414	FALSE
resid4	0.7273023	0.7441235	0.9773946	0.3283738	FALSE
resid5	0.9527433	0.6127750	1.5548012	0.1199934	FALSE
resid6	0.7413317	0.5645209	1.3132051	0.1891138	FALSE
resid7	0.6147418	0.5458983	1.1261105	0.2601188	FALSE
resid8	0.7104443	0.5861769	1.2119963	0.2255138	FALSE
resid9	1.0033061	0.5924837	1.6933902	0.0903812	FALSE

	Value	Std. Error	t value	p value	significant
sosten2	0.0702899	0.1246025	0.5641128	0.5726774	FALSE
ingreso2	-0.2531987	0.1350213	-1.8752504	0.0607583	FALSE
ingreso3	-0.2252740	0.1559256	-1.4447534	0.1485272	FALSE
ingreso4	-0.3741999	0.1956827	-1.9122787	0.0558405	FALSE
ingreso5	0.1312923	0.2726059	0.4816195	0.6300763	FALSE
ingreso6	-0.1120429	0.3106480	-0.3606748	0.7183426	FALSE
frec3	-0.4557005	0.1234253	-3.6921157	0.0002224	TRUE
frec4	-0.9146732	0.1915537	-4.7750222	0.0000018	TRUE
horario2	0.0814477	0.1655506	0.4919806	0.6227330	FALSE
horario3	0.1505342	0.1477484	1.0188546	0.3082720	FALSE
horario4	0.1953771	0.1769841	1.1039249	0.2696257	FALSE
puntaje	0.2604226	0.0105252	24.7427875	0.0000000	TRUE
1|2	10.2328409	0.9633224	10.6224467	0.0000000	TRUE
2|3	11.7396019	0.9163369	12.8114476	0.0000000	TRUE
3|4	14.3016551	0.9179164	15.5805637	0.0000000	TRUE
4|5	19.0047010	0.9781968	19.4283000	0.0000000	TRUE

3.2. Satisfacción - Ajuste e interpretación del modelo

Este ajuste nos proporcionará razones de chance de pertenecer a una categoria mayor con respecto a una menor. Algunos ejemplos concretos:

- La chance de dar una mejor evaluación en la encuesta de satisfacción aumenta en un factor de 1.29 por cada punto del total del puntaje en la encuesta.
- La chance de dar una peor evaluación en la encuesta de satisfacción disminuye en un factor de 0.64 cuando el genero es 2.
- La chance de dar una peor evaluación en la encuesta de satisfacción disminuye en un factor de 0.39 cuando el rango de edad es 4.

```
formula2 <- as.formula(
  "satisf~genero+edad+frec+puntaje")
modelo.polr.sig <- MASS::polr(data = data.final,
                              formula = formula2,
                              Hess = TRUE)
tabla3 <- cbind(OR = exp(coef(modelo.polr.sig)))
tabla3
```

```
##                OR
## genero2 0.6437312
## edad2   0.6374116
## edad3   0.5493324
## edad4   0.3903074
## edad5   0.5549351
## edad6   0.5168238
## edad7   0.9179967
## frec3   0.6451769
## frec4   0.4111817
## puntaje 1.2957103
```

3.3. Recomendación - Selección de variables

Haciendo un procedimiento similar al del apartado 3.1 podemos seleccionar las variables significativas para el modelamiento de la variable recomendación. Las variables son género, edad, residencia, frecuencia y puntaje.

```
formula3 <- as.formula(
  "recomen~genero+edad+nacion+resid+sosten+ingreso+frec+horario+puntaje")
modelo.polr.r <- MASS::polr(data = data.final,
                            formula = formula3,
                            Hess = TRUE)
coef.table.r <- coef(summary(modelo.polr.r))
p <- pnorm(abs(coef.table.r[, "t value"]), lower.tail = FALSE) * 2
coef.table.r <- cbind(coef.table.r, "p value" = p)
coef.table.r <- as.data.frame(coef.table.r)
coef.table.r$significant <- ifelse(coef.table.r$`p value` > 0.05, FALSE, TRUE)
knitr::kable(coef.table.r)
```

	Value	Std. Error	t value	p value	significant
genero2	-0.4650575	0.1506877	-3.0862332	0.0020271	TRUE
edad2	-0.2315184	0.2453027	-0.9438070	0.3452683	FALSE
edad3	-0.3540825	0.2567222	-1.3792435	0.1678197	FALSE
edad4	-0.5544974	0.2584636	-2.1453599	0.0319241	TRUE
edad5	-0.6521521	0.2725691	-2.3926116	0.0167289	TRUE
edad6	-0.1580032	0.3444433	-0.4587203	0.6464350	FALSE
edad7	-0.7649530	0.4520224	-1.6922901	0.0905907	FALSE
nacion2	0.1490945	0.2298216	0.6487403	0.5165062	FALSE
resid10	0.9216590	0.5216692	1.7667500	0.0772701	FALSE
resid11	1.3161116	0.6568034	2.0038137	0.0450900	TRUE
resid12	0.7794480	0.5587797	1.3949110	0.1630427	FALSE
resid13	1.1795200	0.6253026	1.8863187	0.0592520	FALSE
resid14	1.3266549	0.5813506	2.2820220	0.0224880	TRUE
resid15	0.9568605	1.2156900	0.7870925	0.4312277	FALSE
resid16	2.2833189	0.8307309	2.7485664	0.0059857	TRUE
resid2	0.6513230	0.6132317	1.0621157	0.2881832	FALSE
resid3	0.8030202	0.5806432	1.3829839	0.1666698	FALSE
resid4	0.9121025	0.6986307	1.3055574	0.1917031	FALSE
resid5	1.3902566	0.5653303	2.4591937	0.0139249	TRUE
resid6	1.1657803	0.5095384	2.2879143	0.0221425	TRUE
resid7	1.1297391	0.4900891	2.3051709	0.0211570	TRUE
resid8	1.0373087	0.5300005	1.9571844	0.0503258	FALSE
resid9	0.9952336	0.5391887	1.8457985	0.0649215	FALSE
sosten2	0.1498885	0.1198235	1.2509103	0.2109672	FALSE
ingreso2	-0.1596483	0.1304605	-1.2237291	0.2210544	FALSE
ingreso3	-0.1504935	0.1499601	-1.0035569	0.3155923	FALSE
ingreso4	-0.3029170	0.1851914	-1.6356973	0.1019030	FALSE
ingreso5	0.0244657	0.2576944	0.0949407	0.9243619	FALSE
ingreso6	0.1298869	0.2885122	0.4501953	0.6525696	FALSE
frec3	-0.6747949	0.1235946	-5.4597441	0.0000000	TRUE
frec4	-1.6018259	0.1842694	-8.6928468	0.0000000	TRUE
horario2	0.2967304	0.1593769	1.8618150	0.0626292	FALSE
horario3	0.2225370	0.1400493	1.5889907	0.1120625	FALSE
horario4	0.2039178	0.1669279	1.2215926	0.2218617	FALSE
puntaje	0.2105937	0.0097276	21.6492041	0.0000000	TRUE

	Value	Std. Error	t value	p value	significant
1|2	8.3537134	0.8879611	9.4077468	0.0000000	TRUE
2|3	9.7241559	0.8660374	11.2283323	0.0000000	TRUE
3|4	11.1338586	0.8667320	12.8457911	0.0000000	TRUE
4|5	14.8568393	0.9020330	16.4703935	0.0000000	TRUE

3.4. Recomendación - Ajuste e interpretación del modelo

Este ajuste nos proporcionará razones de chance de pertenecer a una categoria mayor con respecto a una menor. Algunos ejemplos concretos:

- La chance de dar una mejor evaluación en la encuesta de satisfacción aumenta en un factor de 9.75 cuando la residencia es 16.
- La chance de dar una peor evaluación en la encuesta de satisfacción disminuye en un factor de 0.21 cuando la frecuencia es 4.

```
formula4 <- as.formula(
  "recomen~genero+edad+resid+frec+puntaje")
modelo.polr.r.sig <- MASS::polr(data = data.final,
                                formula = formula4,
                                Hess = TRUE)
tabla4 <- cbind(OR = exp(coef(modelo.polr.r.sig)))
tabla4
```

```
##                OR
## genero2 0.6038513
## edad2   0.7596633
## edad3   0.6383416
## edad4   0.4897335
## edad5   0.4468327
## edad6   0.7001123
## edad7   0.4003244
## resid10 2.5793704
## resid11 3.7532388
## resid12 2.1709899
## resid13 3.2952010
## resid14 3.7430271
## resid15 3.1076179
## resid16 9.7515978
## resid2  1.9564730
## resid3  2.2353857
## resid4  2.4447442
## resid5  4.0542974
## resid6  3.3219909
## resid7  3.0655862
## resid8  2.7949634
## resid9  2.7934173
## frec3   0.5149186
## frec4   0.2064483
## puntaje 1.2352149
```

4. Indicaciones y recomendaciones

Una vez que identificamos y entendemos cuales son los factores que afectan la probabilidad de recibir una cierta evaluación, estamos en posición de poder tomar acciones en base a estos hallazgos. Tomando como referencia algunas de las interpretaciones del apartado anterior, un camino de acción puede ser enfocar los esfuerzos en personas del genero 2 o de rango de edad 4 y entender que aspectos de la encuesta tienen un menor puntaje. En la siguiente tabla se muestra un puntaje promedio de las 10 preguntas peor evaluadas para encuestados en el rango de edad 4 ordenadas de menor a mayor puntaje. Es interesante ver como las 7 primeras preguntas peor evaluadas están relacionadas con al precio, calidad y variedad de productos en la tienda, lo cual segun los encuestados en este rango de edad es bastante deficiente, por lo que se podrían tomar medidas como incluir productos nuevos que sean de interés de este grupo de personas.

Este tipo de enfoque permite tomar acciones concretas sobre cómo se podría mejorar y si es que es rentable para el negocio llevar a cabo estas mejoras.

```
data.final.desglose <- as.data.frame(cbind(Y1, Y2, D, Q))
data.final.desglose <- data.final.desglose %>%
  dplyr::filter(D2 == 4) %>%
  dplyr::select(-c(Y1, Y2, D1, D2, D3, D4, D5, D6, D7, D8))

puntajes.promedio.pregunta <- cbind(
  puntaje.promedio = apply(data.final.desglose, 2, mean))

puntajes.promedio.pregunta <- cbind(
  puntaje.promedio = puntajes.promedio.pregunta[order(puntajes.promedio.pregunta),])
head(puntajes.promedio.pregunta, 10)
```

```
##     puntaje.promedio
## P5          2.345259
## P1          2.361360
## P4          2.361360
## P6          2.382826
## P2          2.438283
## P7          2.445438
## P3          2.465116
## P21         3.282648
## P28         3.323792
## P23         3.438283
```

Si ahora vemos un ejemplo similar pero para el modelo de recomendación podremos corroborar que la tendencia es la misma que para el caso de satisfacción, donde los ítems de la encuesta con menor puntaje estan relacionados al precio, calidad y variedad de productos en la tienda.

```
data.final.desglose.r <- as.data.frame(cbind(Y1, Y2, D, Q))
data.final.desglose.r <- data.final.desglose.r %>%
  dplyr::filter(D7 == 4) %>%
  dplyr::select(-c(Y1, Y2, D1, D2, D3, D4, D5, D6, D7, D8))

puntajes.promedio.pregunta.r <- cbind(
  puntaje.promedio = apply(data.final.desglose.r, 2, mean))

puntajes.promedio.pregunta.r <- cbind(
  puntaje.promedio = puntajes.promedio.pregunta.r[order(puntajes.promedio.pregunta.r),])
head(puntajes.promedio.pregunta.r, 10)
```

```
##     puntaje.promedio
## P5          2.362445
## P7          2.379913
## P4          2.388646
## P1          2.414847
## P6          2.419214
## P2          2.441048
## P3          2.541485
## P28         3.091703
## P21         3.126638
## P23         3.231441
```