Instituto Tecnológico Autónomo de México

Economía Computacional

Tarea 1

Prof. Isidoro García Urquieta

Alfredo Lefranc Flores

144346

Cynthia Raquel Valdivia Tirado

81358

Rafael Sandoval Fernández

143689

Marco Antonio Ramos Juárez

142244

Ciudad de México 2021

Contents

En esta tarea pondrán en práctica los conceptos de High Dimensional Inference y Regresión. La base de datos muestra las compras de helados Ben & Jerry. Cada fila es una compra. Cada columna es una característica del helado comprado o de la persona que compró.

Limpieza de datos

Carga los datos en BenAndJerry.csv.

```
# Carga las librerias
library(ggplot2)
library(dplyr)
library(RCT)
library(knitr)
library(broom)
library(stargazer)
library(kableExtra)
library(naniar)
library(ggthemes)

# Carga la base de datos
base<-read.csv("BenAndJerry.csv")
```

1. Cuales son las columnas de la base? Muestra una tabla con ellas

```
columnas <- (as.data.frame(colnames(base)))

kable(columnas, booktabs=T, align = 'c', col.names = c("Columnas"), longtable=T) %>%
  kable_styling(position = "center",latex_options = "repeat_header")
```

Columnas
quantity
price_paid_deal
price_paid_non_deal
coupon_value
promotion_type
size1_descr
flavor_descr
formula_descr
household_id
household_size

(continued)

Columnas
household_income
age_of_female_head
age_of_male_head
age_and_presence_of_children
male_head_employment
female_head_employment
male_head_education
female_head_education
marital_status
male_head_occupation
female_head_occupation
household_composition
race
hispanic_origin
region
scantrack_market_identifier
fips_state_code
fips_county_code
type_of_residence
kitchen_appliances
tv_items
female_head_birth
male_head_birth
household_internet_connection

2. A qué nivel está la base? Esto es, cuál es la variable que define la base de manera única. Si no la hay, crea una y muestra que es única a nivel de la base (Muestra el código)

Así como está la base sin niguna modificación, el nivel es la compra. Es decir, cada fila representa una transacción realizada por un hogar. Esto lo podríamos modificar para que la unidad sea el hogar o cualquier otra variable. Sin embargo, notamos que hay una cantidad alta de filas que están repetidas.

```
sum(as.numeric(duplicated(base)))
```

```
## [1] 2460
```

Esto puede deberse a un error de registro o a que el mismo hogar previamente registrado realizó la misma compra más de una vez. Por lo pronto asumirémos lo segundo.

En cuanto a la segunda parte de la pregunta, no hay una variable explícita que identifique cada observación de manera única pero sí hay una manera implícita y es el índice de cada fila. En este sentido, una manera fácil de crear una variable idenfiticadora sería simplemente crear una variable que "clone" el indice de cada fila. Por otro lado, una alternativa más nutritiva para el análisis de datos sería crear alguna variable que combine información del identificador del hogar y del número de transacción. Para esto podemos concatenar el identificador de cada hogar con una variable que lleva el conteo del número de transacciones que cada hogar lleva separadas con un guión, de la siguiente manera:

$$id = household\_id - conteo\_transacción$$

```
base <- base %>% group_by(household_id) %>%
          mutate(id_trans = sequence(n())) %>% ungroup()
base$id<-paste(base$household_id, base$id_trans, sep="-")

#Ejemplo:
kable(head(base$id), col.names = c("Ejemplo"), booktabs=T)%>%

  kable_styling(position = "center",latex_options = "repeat_header")
```

Ejemplo
2001456-1
2001456-2
2001456-3
2001637-1
2002791-1
2002791-2

3. Que variables tienen valores vacíos? Haz una tabla con el porcentaje de vacíos para las columnas que tengan al menos una observación vacía

Los NAs de las variables numéricas son identificables mediante un summary.

```
summary(base)
```

Las variables *promotion_type, scantrack_market_identifier, female_head_occupation* y *tv_items*

tienen valores faltantes. Sin embargo, es posible que las variables de caracteres también tengan valores vacíos.

Al revisar estas variables, notamos que *male_head_birth* y *female_head_birth* también tienen valores vacíos. En general encontramos lo siguiente:

```
kable( (base %>% select_if(~sum(is.na(.)) > 0) %>%
          miss_var_summary()), booktabs=T, align = 'c',
       col.names = c("Variale", "Cantidad","%"),digits = 2)%>%
  kable_styling(position = "center")
```

Variale	Cantidad	%
promotion_type	12980	59.07
male_head_birth	5317	24.20
scantrack_market_identifier	4068	18.51
female_head_occupation	2267	10.32
female_head_birth	2267	10.32
tv_items	34	0.15

4. Haz algo con los valores vacíos (Se deben reemplazar por algún valor? Eliminar de la base?). Justifica tu respuesta.

Pues dependiendo de la cantidad de valores vacíos, de las características de cada variable y de que exista algún patrón en los valores podemos proponer una estrategia, por ejemplo imputación o simplemente quitar esas observaciones. En este sentido tenemos que realizar un análisis por variable:

promotion_type

```
summary(factor(base$promotion_type))
```

```
##     1     2     3     4  NA's
##  6509  1106  1258   121 12980
```

En esta variable podría ser que los NAs nos indiquen que sencillamente no hubo ninguna promoción (y eso podría explicar que casi el 60% de sus valores sean NAs). En este caso podemos suponer eso e imputarle un valor de 0 a cada `NA`.

```
base$promotion_type[is.na(base$promotion_type)] <- 0
```

scantrack_market_identifier

```
summary(factor(base$scantrack_market_identifier))
```

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
##  960  609  269  196  122  118  988  559  310  229  259  802  650  468  136  345
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
##  442  666  567  424  137  394  187  569  318  332  199  382  350  240  105  337
##   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
##  406  128  102  138  137  472  311  200  392  499  208  404   79  259  117   72
##   49   50   51   52 NA's
##  251  468  403  191 4068
```

En este caso es más complejo porque es muy probable que cada valor corresponda a un producto, a una clasificación de cliente o a cualquier otra cosa. En este caso, lo que podríamos hacer es ver si podemos inferir está información de otras variables, de lo contrario imputar sería una muy mala idea pues estaríamos creando ruido en nuestra información. Investigando un poco nos dimos cuenta que se trata de una clasificación del posicionamiento en el mercado. En este sentido, es probable que depende del lugar geográfico. Por ello, decidimos investigar si existe una relación entre la variable scantrack_market_identifier y las variables que indican al estado, condado y tipo de residencia.

En primer lugar notamos que cada combinación de estado con condado solo permite una categoría de scantrack_market_identifier, en este sentido nuestra intuición era correcta. El problema es que los valores faltantes abarcan condados completos, es decir que en dichos condados no hay ninguna observación con scantrack_market_identifier, por lo que no podemos saber en realidad cuál categoría debería tener.

```
susp<-base%>% select(fips_state_code,fips_county_code,
                     scantrack_market_identifier)%>%
  mutate(estado_condado=paste(fips_state_code,"-",fips_county_code))%>%
  group_by(estado_condado,scantrack_market_identifier)%>%
  summarize(n())
kable(head(susp),booktabs=T,
      col.names = c("Ejemplo Estado-Condado","scantrack_market_identifier",
                    "Observaciones"))%>%
  kable_styling(position = "center",latex_options = "repeat_header")
```

Ejemplo Estado-Condado	scantrack_market_identifier	Observaciones
1 - 1	31	4
1 - 101	31	5
1 - 103	31	9
1 - 111	14	7
1 - 117	31	5
1 - 121	31	8

```
length(unique((susp$estado_condado)))
```

```
## [1] 1256
```

Sin embargo, podemos analizar la variedad de categorías en cada estado. En este sentido, en la siguiente tabla agrupamos las categorías de *scantrack_market_identifier* de cada estado y la cantidad de observaciones que tuvo cada combinación. Como podemos observar, en realidad las categorías de cada estado son pocas, (en promedio 2.7 y con mayorías claras), por lo que bien podríamos imputar la moda. Sin embargo, no podríamos hacer mucho más. Si quisiéramos mejorar la calidad de la imputación necesitaríamos tener más información o más dimensiones (por ejemplo, la ubicación de los condados para calcular una correlación espacial).

```
susp_2<-base%>% select(fips_state_code,scantrack_market_identifier)%>%
  group_by(fips_state_code,scantrack_market_identifier)%>%
  summarize(n())

kable(head(susp_2,10),booktabs=T,
      col.names = c("Ejemplo Estado","scantrack_market_identifier",
                    "Observaciones"))%>%
  kable_styling(position = "center",latex_options = "repeat_header")
```

Ejemplo Estado	scantrack_market_identifier	Observaciones
1	14	8
1	31	105
1	36	7
1	NA	22
4	38	472
5	34	127
5	35	7
5	NA	49
6	7	988
6	12	802

```
susp_2<-base%>% select(fips_state_code,scantrack_market_identifier)%>%
  group_by(fips_state_code,scantrack_market_identifier)%>%

  summarize(n()) %>% ungroup() %>%
  group_by(fips_state_code)%>%
  summarize(n())

kable(head(susp_2,10),booktabs=T,
      col.names = c("Ejemplo Estado",
                    "cantidad de scantrack_market_identifier diferentes"))%>%
  kable_styling(position = "center",latex_options = "repeat_header")
```

Ejemplo Estado	cantidad de scantrack_market_identifier diferentes
1	4
4	1
5	3
6	5
8	2
9	3
10	2
11	1
12	5
13	3

```
#promedio de categoría de scantrack_market_identifier por estado
mean(susp_2$`n()`)
```

```
## [1] 2.795918
```

En conclusión, la mejor imputación posible es imputar la moda por estado, aunque no es la imputación que qusieramos. Sin embargo, debido a que son el 18% de las observaciones, es preferible intentar conservarlas a simplemente descartarlas.

```
#Imputación de la moda

#Creo mi tabla de modas
llave_imputaciones<-base %>% select(fips_state_code,scantrack_market_identifier)%>%
  group_by(fips_state_code) %>% count(scantrack_market_identifier) %>% top_n(1) %>%
  rename(moda=scantrack_market_identifier)
llave_imputaciones$n<-NULL

#Imputo
base<-merge(base,llave_imputaciones, by = "fips_state_code")
base<-base%>%mutate(scantrack_market_identifier=
      ifelse(is.na(scantrack_market_identifier),moda,scantrack_market_identifier))
```

```
sum(is.na(base$scantrack_market_identifier))
```

```
## [1] 1569
```

De los 4068 valores faltantes los redujimos a tan solo 1569 (debido a que para algunos estados no hay nada de información). Sin embargo, para no perder las filas les agregamos una etiqueta 0 para identificarlas y seguir con nuestro análisis.

```
base$scantrack_market_identifier[is.na(base$scantrack_market_identifier)] <- 0
```

female_head_occupation y female_head_birth

```
aux<-base %>% select(age_of_female_head,
                     female_head_occupation,
                     female_head_education,
                     female_head_employment,
                     female_head_birth) %>%
  filter (is.na(female_head_occupation))

summary((aux))
```

```
##  age_of_female_head female_head_occupation female_head_education
```

```
##  Min.   :0          Min.   : NA            Min.   :0
##  1st Qu.:0          1st Qu.: NA            1st Qu.:0
##  Median :0          Median : NA            Median :0
##  Mean   :0          Mean   :NaN            Mean   :0
##  3rd Qu.:0          3rd Qu.: NA            3rd Qu.:0
##  Max.   :0          Max.   : NA            Max.   :0
##                     NA's   :2267
##  female_head_employment female_head_birth
##  Min.   :0              Length:2267
##  1st Qu.:0              Class :character
##  Median :0              Mode  :character
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
##
```

```
summary(aux$age_of_female_head[aux$female_head_birth==""])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      NA      NA      NA     NaN      NA      NA    2267
```

Explorando los datos, notamos que todos los NAs de las variables *female_head_occupation* y *female_head_birth* coinciden, y además corresponden a observaciones en que se registra una edad, educación y ocupación de la jefa del hogar de cero. Esto nos lleva a concluir que en los hogares que hicieron esas compras no hay una jefa de hogar femenina. En este sentido creamos una nueva categoría de ocupación de mujeres con estas características con el número 0, la cual imputamos a los valores faltantes. Por su parte, dejamos como desconocidos los valores faltantes de la variable *female_head_birth* sin embargo les ponemos una etiqueta para no perder las observaciones

```
base$female_head_occupation[is.na(base$female_head_occupation)] <- 0
base$female_head_birth[is.na(base$female_head_birth)] <- "unknown"
```

male_head_birth

```
aux2<-base %>% select(age_of_male_head,
                      male_head_occupation,
                      male_head_education,
                      male_head_employment,
                      male_head_birth) %>%
  filter (is.na(male_head_birth))

summary((aux2))
```

```
##  age_of_male_head male_head_occupation male_head_education male_head_employment
##  Min.   :0        Min.   : 1.000       Min.   :0           Min.   :0
##  1st Qu.:0        1st Qu.: 1.000       1st Qu.:0           1st Qu.:0
##  Median :0        Median : 3.000       Median :0           Median :0
##  Mean   :0        Mean   : 5.073       Mean   :0           Mean   :0
##  3rd Qu.:0        3rd Qu.:12.000       3rd Qu.:0           3rd Qu.:0
##  Max.   :0        Max.   :12.000       Max.   :0           Max.   :0
##  male_head_birth
##  Length:5317
##  Class :character
##  Mode  :character
##
##
##
```

Los valores faltantes de *male_head_birth* coinciden con ceros en edad, educación y empleo del jefe del hogar masculino, aunque curiosamente sí se tiene registro de su ocupación. Concluimos, como en el caso de las mujeres, que se trata de casos, en los que la compra corresponde a hogares sin un jefe del hogar masculino, y decidimos ignorar estos valores vacíos (pero agregándoles una etiqueta para no perder las observaciones).

```
base$male_head_birth[is.na(base$male_head_birth)] <- "unknown"
```

tv_items

En este caso, puede que la variable indique una cantidad de *items* o bien que indique una categoría. En el caso primero, parecería que no contemplaron una cantidad de ceros o de más de 3, bien podríamos imputar el valor de 0. En el segundo caso, no tenemos manera de saber el tipo de categorías son, en ese caso no podríamos imputar tan facilmente: podríamos agregar un valor para identificarlas (como un 0) o bien simplemente prescindir de dichas observaciones (lo cuál no afectaría nuestro análisis debido a que son tan solo 34 observaciones). Optamos por imputarles el valor de cero, dado que esa opción es congruente sea la variable categórica o numérica.

```
base$tv_items[is.na(base$tv_items)] <- 0
summary(factor(base$tv_items))
```

```
##    0    1    2    3
##   34 7986 7530 6424
```

5. Muestra una tabla de estadisticas descriptivas de la base. Esta debe tener cada columna númerica con algunas estadísticas descriptivas (N, media, min, p05, p25, p50, p75, p90, p95, max).

Sin hacer ninguna adecuación en el tipo de variables, la tabla es la siguiente:

```
b <- read.csv("BenAndJerry.csv")
b<- summary_statistics(b,probs=c(0,0.05,0.25,0.5,0.75,0.9,0.95,1),na.rm=T)
b<- b %>% mutate_at(vars(-variable),funs(round(.,2))) %>%
  rename(mín=4) %>%
  rename(máx=11)

options(scipen=999) # quitamos notación científica
kable(b,booktabs=T, align = 'c')%>%
  kable_styling(position = "center")%>%
  kable_styling(latex_options="scale_down")
```

variable	mean	n	mín	0.05	0.25	0.5	0.75	0.9	0.95	máx
quantity	1.28	21974	1	1	1	1.00	1.00	2.00	2.00	21.00
price_paid_deal	1.74	21974	0	0	0	0.00	3.34	4.50	6.86	28.88
price_paid_non_deal	2.45	21974	0	0	0	2.99	3.56	4.99	6.86	69.72
coupon_value	0.16	21974	0	0	0	0.00	0.00	0.50	1.00	12.95
promotion_type	1.44	8994	1	1	1	1.00	2.00	3.00	3.00	4.00
household_id	16612005.04	21974	2000358	2054762	8142253	8401573.00	30183891.00	30338638.00	30387781.05	30440689.00
household_size	2.46	21974	1	1	2	2.00	3.00	4.00	5.00	9.00
household_income	21.47	21974	3	11	17	23.00	26.00	27.00	28.00	30.00
age_of_female_head	5.51	21974	0	0	4	6.00	8.00	8.00	9.00	9.00
age_of_male_head	4.76	21974	0	0	2	5.00	8.00	8.00	9.00	9.00
age_and_presence_of_children	7.40	21974	1	2	6	9.00	9.00	9.00	9.00	9.00
male_head_employment	3.09	21974	0	0	1	3.00	3.00	9.00	9.00	9.00
female_head_employment	4.20	21974	0	0	2	3.00	9.00	9.00	9.00	9.00
male_head_education	3.32	21974	0	0	2	4.00	5.00	6.00	6.00	6.00
female_head_education	3.98	21974	0	0	3	4.00	5.00	6.00	6.00	6.00
marital_status	1.94	21974	1	1	1	1.00	3.00	4.00	4.00	4.00
male_head_occupation	5.11	21974	1	1	1	4.00	8.00	12.00	12.00	12.00
female_head_occupation	5.80	19707	1	1	1	3.00	12.00	12.00	12.00	12.00
household_composition	2.57	21974	1	1	1	1.00	5.00	7.00	7.00	8.00
race	1.24	21974	1	1	1	1.00	1.00	2.00	3.00	4.00
hispanic_origin	1.95	21974	1	2	2	2.00	2.00	2.00	2.00	2.00
region	2.63	21974	1	1	2	3.00	4.00	4.00	4.00	4.00
scantrack_market_identifier	23.05	17906	1	1	11	20.00	36.00	45.00	50.00	52.00
fips_state_code	27.20	21974	1	6	12	26.00	39.00	48.00	53.00	56.00
fips_county_code	79.67	21974	1	3	25	59.00	101.00	163.00	201.00	810.00
type_of_residence	2.08	21974	1	1	1	1.00	3.00	5.00	6.00	7.00
kitchen_appliances	3.81	21974	1	1	4	4.00	4.00	7.00	7.00	9.00
tv_items	1.93	21940	1	1	1	2.00	3.00	3.00	3.00	3.00
household_internet_connection	1.16	21974	1	1	1	1.00	1.00	2.00	2.00	2.00

No obstante, algunas de estas variables en realidad no son numéricas, por lo que sus estadísticas descriptivas podrían ser engañosas.

6. Hay alguna númerica que en verdad represente una categorica? Cuales? Cambialas a factor

De las variables numéricas, por su nombre y rango de valores, podemos inferir que algunas son categóricas con seguridad y algunas otras pueden o no ser categóricas. En este sentido en el siguiente código realizamos el análisis de la situación y convertimos a las variables categóricas pertinentes.

```
variables_seguras<-c("promotion_type",
                     "household_income",
                     "age_of_female_head",
                     "age_of_male_head",
                     "male_head_employment",
                     "female_head_employment",
                     "marital_status",
                     "male_head_occupation",
                     "female_head_occupation",
                     "household_composition",
                     "race",
                     "hispanic_origin",
                     "region",
                     "scantrack_market_identifier",
                     "fips_state_code",
                     "fips_county_code",
                     "type_of_residence",
                     "household_internet_connection")

variables_no_seguras<-c("tv_items",
                        "kitchen_appliances",
                        "age_and_presence_of_children",
                        "male_head_education",
                        "female_head_education")

base[,variables_seguras] <- lapply(base[,variables_seguras] , factor)
base[,variables_no_seguras] <- lapply(base[,variables_no_seguras] , factor)

summary(base[,variables_no_seguras])
```

```
##  tv_items kitchen_appliances age_and_presence_of_children male_head_education
##  0:  34   4      :14130      9      :15945                0:5317
##  1:7986   1      : 4430      3      : 2107                1:  59
```

```
##  2:7530   7      : 2698      2      : 1181                2: 425
##  3:6424   5      :  309      1      : 1016                3:3213
##           8      :  247      6      :  807                4:4922
##           2      :  132      4      :  588                5:5475
##           (Other):   28      (Other):  330                6:2563
##  female_head_education
##  0:2267
##  1:  15
##  2: 267
##  3:3453
##  4:6351
##  5:6659
##  6:2962
```

Parece que *tv_items, kitchen_appliances, age_and_presence_of_children* no son categóricas después de todo. Las regresamos a numéricas otra vez, Por el contrario *male_head_education* y *female_head_education* parece que sí son categóricas.

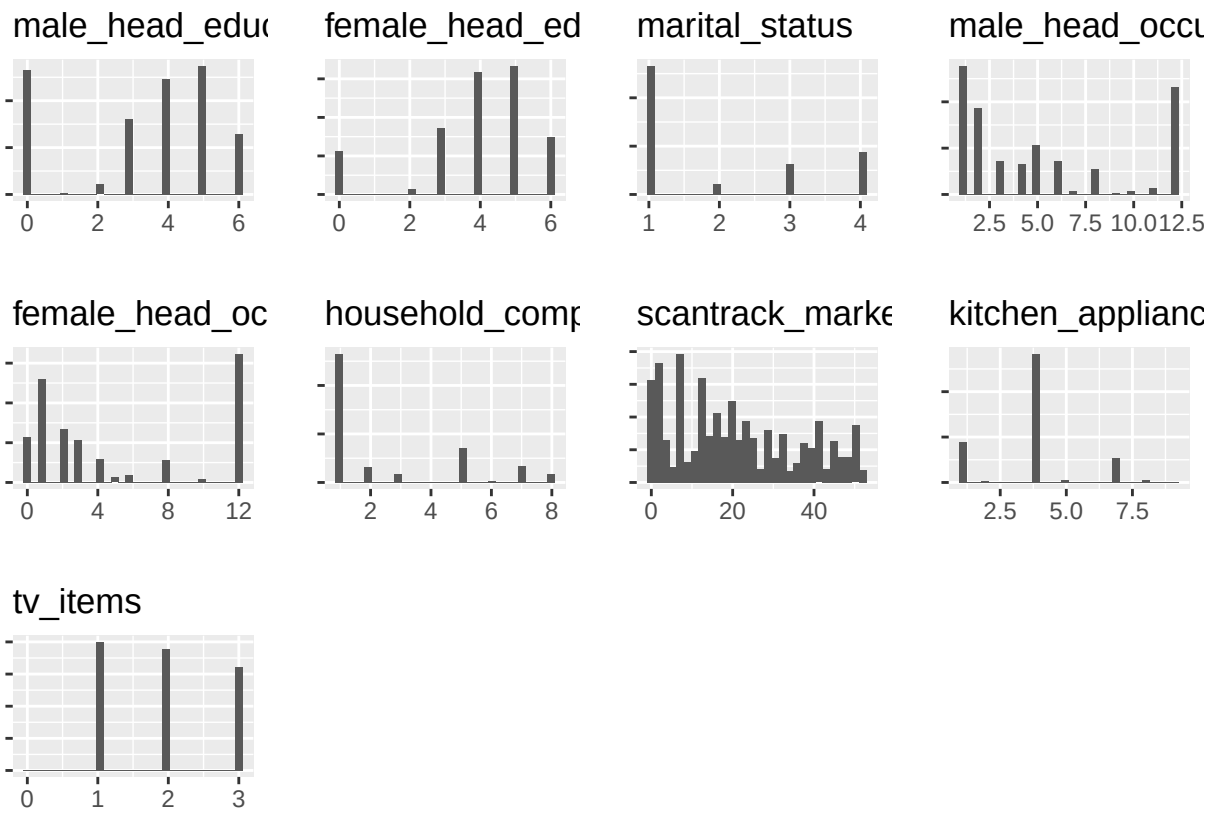
```
variables_numericas<-c("tv_items",
                       "kitchen_appliances",
                       "age_and_presence_of_children")
base[,variables_numericas] <- lapply(base[,variables_numericas] , as.numeric)
```

7. Revisa la distribución de algunas variables. Todas tienen sentido? Por ejemplo, las edades?

```
myhist <- function(yvar){
  ggplot(numericas, aes_(x=as.name(yvar)))+
    geom_histogram()+
    ggtitle(paste0(as.name(yvar)))+
    xlab("")+
    ylab("")+
    theme(axis.text.y = element_blank())
}
hists<- numericas %>% select(price_paid_deal,
                             price_paid_non_deal,
                             coupon_value,
                             household_size:household_composition,
                             scantrack_market_identifier,
                             kitchen_appliances,
                             tv_items) %>%
```

```
        names() %>%
        lapply(myhist)

library(gridExtra)

grid.arrange(grobs=hists[1:10],ncol=4)
```


```
grid.arrange(grobs=hists[11:19],ncol=4)
```



Definitvamente vemos comportamientos atípicos en las edades. Las de los jefes del hogar y el ingreso del hogar tienen valores muy bajos, lo que nos hace pensar que estas variables son categóricas (transformadas en el inciso anterior). Por otra parte, es implausible que la educación de los jefes del hogar, masculino y femenino, conste de 0 a 6 años, razón por la cual también decidimos transformarlas a categóricas en el inciso anterior.

8. Finalmente, crea una variable que sea el precio total pagado y el precio unitario

```
# precio total pagado
base <- base %>% mutate(total_price=price_paid_deal+price_paid_non_deal)
# precio unitario
base <- base %>% mutate(unit_price= (total_price)/quantity)
```

Exploración de los datos

Intentaremos comprender la elasticidad precio de los helados. Para ello, debemos entender:

- La forma funcional base de la demanda (i.e. como se parecen relacionarse q y p).

- Qué variables irían en el modelo de demanda y cuáles no para encontrar la elasticidad de manera 'insesgada'.
- Qué variables cambian la relacion de q y p. Esto es, que variables alteran la elasticidad.

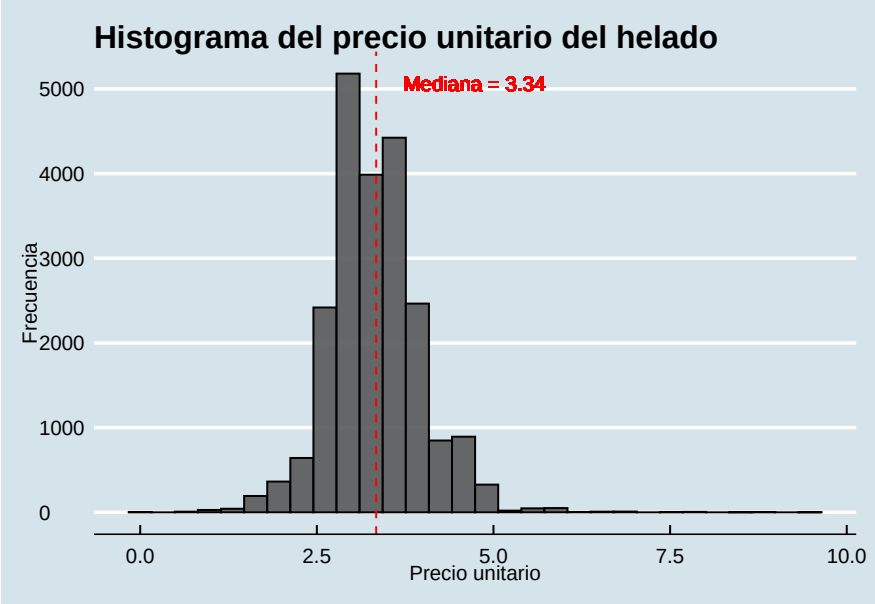
Algo importante es que siempre debemos mirar primero las variables más relevantes de cerca y su relación en:

- Relación univariada
- Relaciones bivariadas
- Relaciones trivariadas

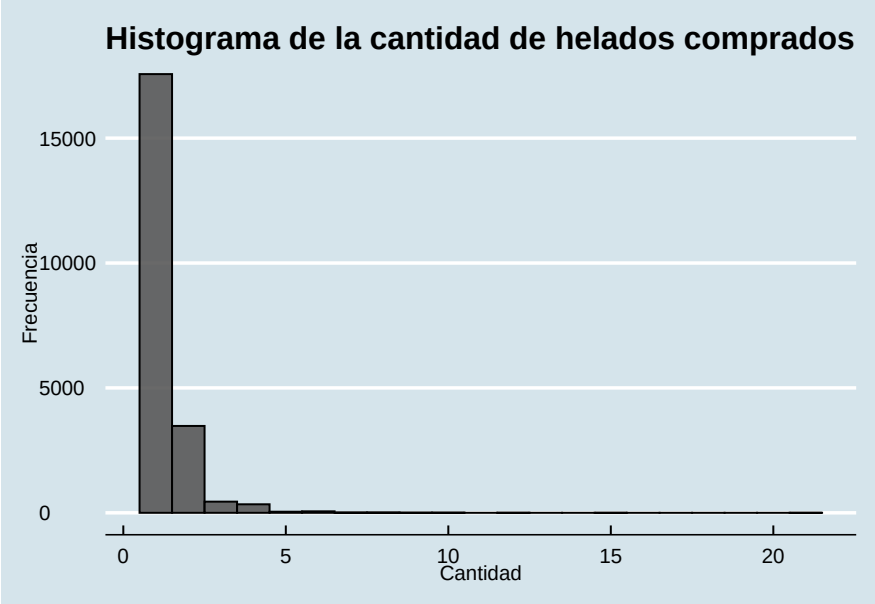
Importante: Las gráficas deben estar bien documentadas (título, ejes con etiquetas apropiadas, etc). Cualquier gráfica que no cumpla con estos requisitos les quitaré algunos puntos.

9. Cómo se ve la distribución del precio unitario y de la cantidad demandada. Haz un histograma.

```
median_price <- quantile(base$unit_price)[3]

ggplot(base)+
  geom_histogram(aes(x=unit_price),alpha=0.9,col = 'black')+
  geom_vline(xintercept = median_price,size=0.5,colour="red", linetype = "dashed")+
  geom_text(aes(x=median_price+2.8, label=paste("Mediana =",median_price),
                y=4800),size=4, colour="red", vjust = -1, hjust = 1.2)+
  labs(title="Histograma del precio unitario del helado",x="Precio unitario",y="Frecuencia")+
  theme_economist() + scale_fill_economist()
```



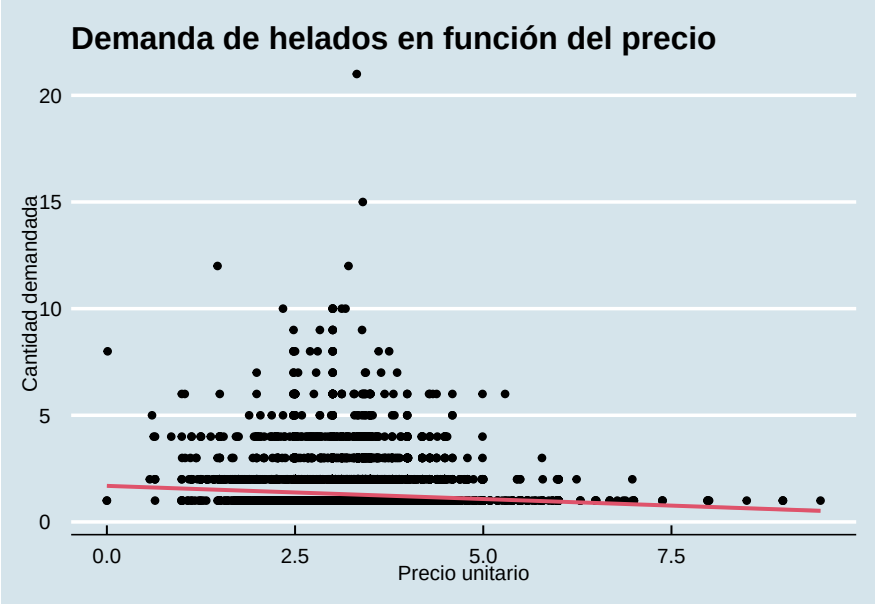
```
ggplot(base)+
  geom_histogram(aes(x=quantity),binwidth=1,alpha=0.9,col = 'black')+
  labs(title="Histograma de la cantidad de helados comprados",x="Cantidad",y="Frecuencia")+
  theme_economist() + scale_fill_economist()
```



10. Grafica la $q(p)$. Que tipo de relación parecen tener?

Aunque parece haber una relación negativa, marcada por las compras de 1 a 4 productos, esta no es tan clara para mayores cantidades.

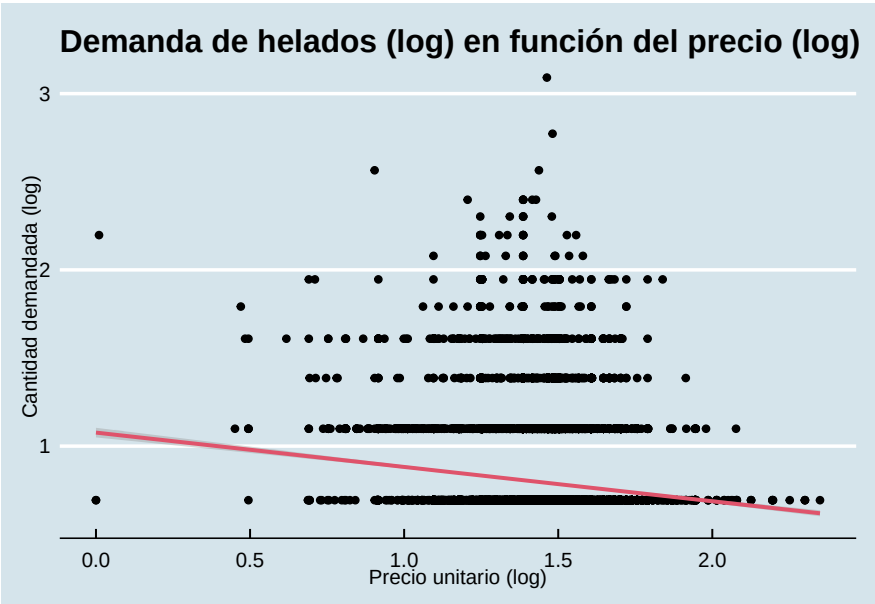
```
ggplot(base)+
  geom_point(aes(y=quantity, x=unit_price))+
  geom_smooth(formula=y~x,method=lm, color='2',aes(y=quantity, x=unit_price))+
  labs(title="Demanda de helados en función del precio",
       x="Precio unitario",y="Cantidad demandada")+
  theme_economist() + scale_fill_economist()
```



11. Grafica la misma relación pero ahora entre $log(p+1)$ y $log(q+1)$

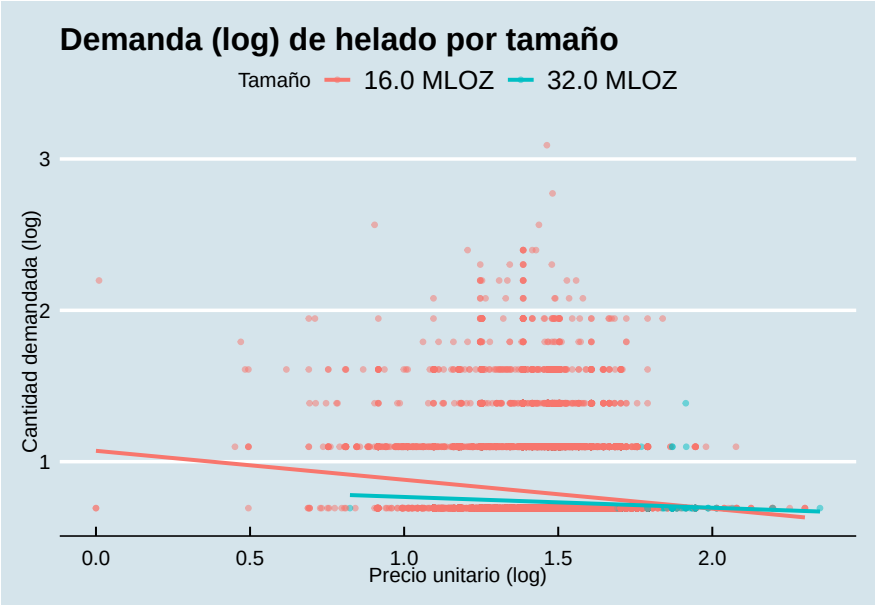
Cuando hacemos la transformación, la relación negativa es más evidente:

```
ggplot(base)+
  geom_point(aes(y=log(quantity+1),x=log(unit_price+1)))+
  geom_smooth(formula=y~x,method=lm, color='2',
              aes(y = log(quantity+1), x = log(unit_price+1)))+
  labs(title = "Demanda de helados (log) en función del precio (log)",
       x="Precio unitario (log)", y="Cantidad demandada (log)")+
  theme_economist() + scale_fill_economist()
```



12. Grafica la curva de demanda por tamaño del helado. Parece haber diferencias en la elasticidad precio dependiendo de la presentación del helado? (2 pts)

```
ggplot(data = base, aes(y=log(quantity+1), x=log(unit_price+1),
                        col=as.factor(size1_descr))) +
  geom_point(size=1, alpha=0.5) +
  geom_smooth(aes(group=size1_descr), method='lm',
              formula= y~(x),
              se=FALSE, size=1)+
  labs(title="Demanda (log) de helado por tamaño",
       y="Cantidad demandada (log)",
       x="Precio unitario (log)",
       col="Tamaño")+
  theme_economist() + scale_fill_economist()
```



De observar la gráfica, hay indicios de que es probable que ambas curvas sean diferentes. Sin embargo, para tener más argumentos planteamos la siguiente prueba de hipótesis:

$$H_0 : \beta_{3,32MLOZ} = 0$$

$$H_a : \beta_{3,32MLOZ} \neq 0$$

Donde el modelo a estimar es el siguiente:

$$log(q_i + 1) = \beta_0 + \beta_1 log(p_i + 1) + \beta_2 32MLOZ + \beta_3 log(p_i + 1)32MLOZ_i + \upsilon_i$$

```
# Prueba de hipotesis
elast_size <- lm(log(quantity+1)~log(unit_price+1)*size1_descr,
                 data=base)

hip_size <- c(0,0,0,1)

library(car)
linearHypothesis(elast_size,hip_size, rhs = NULL, white.adjust="hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(unit_price  + size1_descr32.0 MLOZ = 0
##
## Model 1: restricted model
## Model 2: log(quantity + 1) ~ log(unit_price + 1) * size1_descr
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1  21971
## 2  21970  1 4.5407 0.03311 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al realizar la prueba de hipótesis se obtiene un valor p de 0.03, por lo que se rechazaría la hipótesis nula con un nivel de significancia de 5% en favor de la alternativa (que son diferentes de manera significativa).

13. Grafica la curva de demanda por sabor. Crea una variable con los 3 sabores más populares y agruga el resto de los sabores como 'otros'. Parece haber diferencias en la elasticidad precio dependiendo del sabor?

El primer paso es averiguar cuáles son estos 3 sabores más populares. Para ello elaboramos la siguiente tabla:

```
kable(
  head(base%>% select(flavor_descr)%>% group_by(flavor_descr)%>%
  summarize(n())%>% arrange(desc(`n()`))),
  col.names = c("Sabor","Observaciones"),booktabs=T)%>%
  kable_styling(position = "center")
```

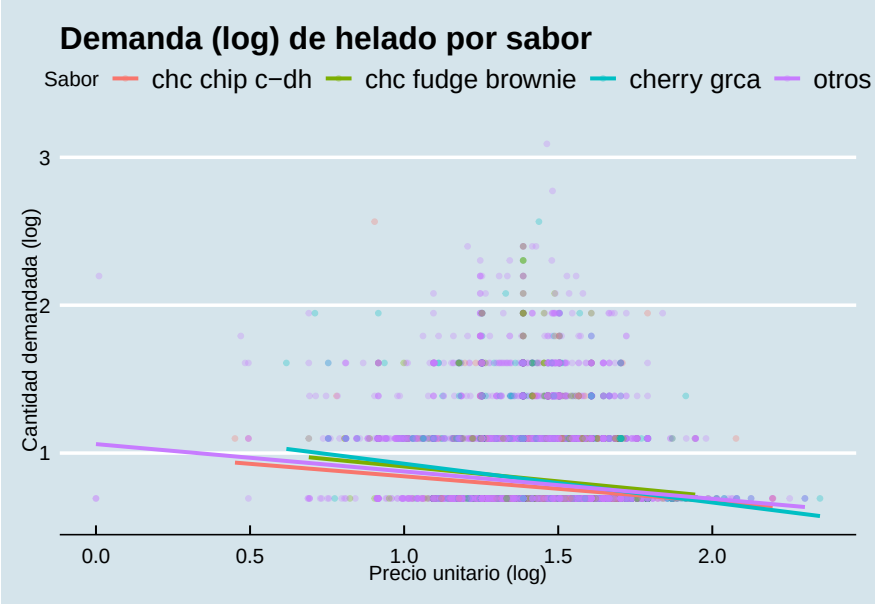
Sabor	Observaciones
CHERRY GRCA	2097
CHC FUDGE BROWNIE	1235
CHC CHIP C-DH	1070
HEATH COFFEE CRUNCH	1070
CHUNKY MONKEY	1064
PHISH FOOD	968

Parece que los 3 sabores más populares son *CHERRY GRCA*, *CHC FUDGE BROWNIE* y *CHC CHIP C-DH*.

El siguiente paso es crear la variable que distinga entre estos 3 sabores y agrupe al resto para después graficar y hacer una prueba de hipótesis:

```
base<-base%>%
  mutate(sabores_pop= ifelse(flavor_descr=='CHERRY GRCA','cherry grca',
(ifelse(flavor_descr=='CHC FUDGE BROWNIE','chc fudge brownie',
(ifelse(flavor_descr=='CHC CHIP C-DH','chc chip c-dh','otros'))))))
```

```
ggplot(data = base, aes(y=log(quantity+1),
                        x=log(unit_price+1),
                        col=as.factor(sabores_pop))) +
  geom_point(size=1, alpha=0.3) +
  geom_smooth(method='lm',
              formula= y~(x),
              se=FALSE, size=1)+
  labs(title="Demanda (log) de helado por sabor",
      y="Cantidad demandada (log)",
      x="Precio unitario (log)",
      col="Sabor")+ theme_economist() + scale_fill_economist()
```



A primera vista, parece que las 4 curvas están muy empalmadas y parece que tienen una pendiente y ordenada al origen similar. A diferencia del caso del tamaño del helado, las diferencias en la elasticidad precio demanda observadas para los sabores de helado aparentan ser muy pequeñas. Sin embargo, realizamos las siguientes pruebas de hipótesis para tener una conclusión más certera:

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

donde evaluamos una prueba para el coeficiente de cada interacción de términos (β_5, β_6 y β_7), con base en el siguiente modelo:

$$log(q_i + 1) = \beta_0 + \beta_1 log(p_i + 1) + \beta_2 chc_c hip_c - dh + \beta_3 chc_f udge_b rownie + \beta_4 cherry_g rca + \beta_5 log(p_i + 1) chc_c hip_c - dh + \beta_6 log(p_i + 1) chc_f udge_b rownie + \beta_7 log(p_i + 1) cherry_g rca + \upsilon_i$$

```
elast_sabores <- lm(log(quantity+1)~log(unit_price+1)*relevel(as.factor(sabores_pop),
                                                              ref='otros'),
                    data=base)

hipchip <- c(0,0,0,0,0,1,0,0)
hipfudge <- c(0,0,0,0,0,0,1,0)
hipcherry <- c(0,0,0,0,0,0,0,1)

linearHypothesis(elast_sabores,hipchip, rhs = NULL, white.adjust="hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(unit_price  + relevel(as.factor(sabores_pop), ref = "otros")chc chip c - dh = 0
##
## Model 1: restricted model
## Model 2: log(quantity + 1) ~ log(unit_price + 1) * relevel(as.factor(sabores_pop),
##     ref = "otros")
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1  21967
## 2  21966  1 0.1194 0.7297
```

```
linearHypothesis(elast_sabores,hipfudge, rhs = NULL, white.adjust="hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(unit_price  + relevel(as.factor(sabores_pop), ref = "otros")chc fudge brownie = 0
##
## Model 1: restricted model
## Model 2: log(quantity + 1) ~ log(unit_price + 1) * relevel(as.factor(sabores_pop),
##     ref = "otros")
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1  21967
## 2  21966  1 0.1352 0.7131
```

```
linearHypothesis(elast_sabores,hipcherry, rhs = NULL, white.adjust="hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(unit_price  + relevel(as.factor(sabores_pop), ref = "otros")cherry grca = 0
##
## Model 1: restricted model
## Model 2: log(quantity + 1) ~ log(unit_price + 1) * relevel(as.factor(sabores_pop),
##     ref = "otros")
```

```
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1  21967
## 2  21966  1 5.7615 0.01639 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al evaluar la hipótesis nula de si los coeficientes estimados de la elasticidad precio demanda entre las submuestras contempladas eran iguales, solamente se rechaza la de *cherry grca* con un $p-value$ de 1.6%. Es decir, la elasticidad precio demanda de los helados *cherry grca* tiene un comportamiento diferenciado del resto de los sabores, y lo mismo no se puede decir para los otros dos sabores más populares.

Por último, también observamos las diferencias entre los tres sabores más populares. Para hacer esto, usamos el mismo modelo para evaluar las hipótesis de diferencias en los coeficientes de interacciones de los tres sabores más populares:

$$H_0 : \beta_j - \beta_k = 0$$

$$H_a : \beta_j - \beta_k \neq 0$$

```
hipcherryfudge <- c(0,0,0,0,0,0,1,-1)
hipcherrychip <- c(0,0,0,0,0,1,0,-1)
hipchipfudge <- c(0,0,0,0,0,1,-1,0)

linearHypothesis(elast_sabores,hipchipfudge, rhs = NULL, white.adjust="hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(unit_price  + relevel(as.factor(sabores_pop), ref = "otros")chc chip c - dh - log(unit_
##
## Model 1: restricted model
## Model 2: log(quantity + 1) ~ log(unit_price + 1) * relevel(as.factor(sabores_pop),
##     ref = "otros")
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1  21967
## 2  21966  1 0.2769 0.5988
```

```
linearHypothesis(elast_sabores,hipcherryfudge, rhs = NULL, white.adjust="hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(unit_price  + relevel(as.factor(sabores_pop), ref = "otros")chc fudge brownie - log(unit
##
## Model 1: restricted model
## Model 2: log(quantity + 1) ~ log(unit_price + 1) * relevel(as.factor(sabores_pop),
##     ref = "otros")
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1  21967
## 2  21966  1 1.4144 0.2343
```

```
linearHypothesis(elast_sabores,hipcherrychip, rhs = NULL, white.adjust="hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(unit_price  + relevel(as.factor(sabores_pop), ref = "otros")chc chip c - dh - log(unit_p
##
## Model 1: restricted model
## Model 2: log(quantity + 1) ~ log(unit_price + 1) * relevel(as.factor(sabores_pop),
##     ref = "otros")
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1  21967
## 2  21966  1 3.0389 0.08131 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estas pruebas no se rechazan a un nivel de significancia de 5, lo que indica que, vistas por separado, las demandas de helados sabor *cherry grca*, *chc fudge brownie* y *chc chip c-dh* no son distintas entre sí.

Estimación

14. Estima la regresión de la curva de demanda de los helados. Reporta la tabla de la regresión

Algunos tips:

- No olvides borrar la variable que recien creamos de sabores. Incluirla (dado que es perfectamente colineal con flavor), sería una violación a supuesto GM 3 de la regresión.
- No olvides quitar `quantity`, `price_unit`, `price_deal` y otras variables que sirven como identificadora. Tambien quitar `fips_state_code` y `fips_county_code`.
- Empecemos con una regresión que incluya a todas las variables.

Nota: La regresión en `R` entiende que si le metes variables de texto, debe convertirlas a un factor. En algunos otros algoritmos que veremos durante el curso, tendremos que convertir manualmente toda la base a una númerica.

Quitemos las fechas

```
base$female_head_birth<-NULL
base$male_head_birth<-NULL
```

En primer lugar estimamos el modelo más sencillo posible.

```
model_a<-lm(quantity~unit_price,data = base)
stargazer(model_a, type = "latex", title="Regresión", digits=1,header=FALSE)
```

En segundo lugar, estimamos el modelo con todas las variables, siguiendo los tips. En el siguiente código mostramos el proceso previo a la estimación.

```
# convertimos sabores en dummies y ponemos como base los de vainilla
base$flavor_descr <- relevel(factor(base$flavor_descr),"VAN")

# uso de cupon / no uso de cupon
base$coupon <- factor(base$coupon_value>0)

# región
levels(base$region) <- c("Region1","Region2","Region3","Region4")

# estado civil
base$married <- factor(base$marital_status==1)

# raza
base$race <- factor(base$race, levels= 1:4)
```

Table 2. Regresión

	Dependent variable:
	quantity
unit_price	−0.1***
	(0.01)
Constant	1.7***
	(0.02)
Observations	21,974
R^2	0.01
Adjusted R^2	0.01
Residual Std. Error	0.7 (df = 21972)
F Statistic	283.8*** (df = 1; 21972)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
levels(base$race) <- c("Race1","Race2","Race3","Race4")

# hispano
base$hispanic_origin <- factor(base$hispanic_origin==1)

# fórmula
base$formula_descr <- as.factor(base$formula_descr)

# tamaño
base$size1_descr <- as.factor(base$size1_descr)

# tvs
base$tvs <- base$tv_items>1

# internet
base$internet <- base$household_internet_connection==1

base$log_price <- log(base$unit_price+1)
```

```r
# quitamos columnas irrelevantes y que sirven como identificadoras, como
# sabores_pop, price_paid_deal, price_paid_non_deal, total_price
# id, id_trans, household_id,
# fips_county_code, fips_state_code)

# guardamos la selección final de variables en una nueva base
 base_reg<- base[,c("quantity",
                    "log_price",
                    "flavor_descr",
                    "size1_descr",
                    "household_income",
                    "household_size",
                    "coupon",
                    "region",
                    "married",
                    "race",
                    "hispanic_origin",
                    "promotion_type",
                    "age_of_female_head",
                    "age_of_male_head",
                    "age_and_presence_of_children",
                    "male_head_employment",
                    "female_head_employment",
                    "male_head_education","female_head_education",
                    "male_head_occupation",
                    "female_head_occupation",
                    "household_composition",
                    "type_of_residence",
                    "kitchen_appliances",
                    "tvs",
                    "internet")]

# estimamos el modelo
model_a2<-lm(log(quantity+1) ~ .,data = base_reg)
```

Por cuestiones de espacio y de comodidad lectora, los resultados los publicamos ***en este enlace***.

15 (2 pts). Cuales son los elementos que guarda el objecto de la regresión? Listalos. Cual es el F-test de la regresión? Escribe la prueba de manera matemática (i.e. como la vimos en clase). (Tip: `summary(fit)` te arroja algo del F-test)

En cuanto a elementos del objeto, se guardan los coeficientes de las 161 variables regresoras (incluyendo las omitidas por colinealidad), junto con su error estándar, estadístico t y valor-p. Esta información puede ser accesada de la siguiente forma:

```
tidy(model_a2)
```

```
## # A tibble: 161 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                     0.924     0.0535    17.3    2.35e-66
##  2 log_price                      -0.170     0.00993  -17.1    3.51e-65
##  3 flavor_descrAMERICONE DREAM     0.0288    0.0122     2.36   1.81e- 2
##  4 flavor_descrBANANA SPLIT        0.0233    0.0131     1.77   7.69e- 2
##  5 flavor_descrBLACK & TAN         0.142     0.0446     3.19   1.44e- 3
##  6 flavor_descrBROWNIE BATTER     -0.00152   0.0204    -0.0746 9.40e- 1
##  7 flavor_descrBUTTER PECAN        0.0337    0.0170     1.99   4.71e- 2
##  8 flavor_descrCAKE BATTER        -0.00419   0.0145    -0.290  7.72e- 1
##  9 flavor_descrCHC                 0.00536   0.0244     0.220  8.26e- 1
## 10 flavor_descrCHC ALMOND NOUGAT -0.00658   0.0221    -0.298  7.66e- 1
## # ... with 151 more rows
```

En cuanto a los elementos del modelo en general, R guarda diversos estimados:

```
glance(model_a2)
```

Adicionalmente, pueden consultarse:

- vectores del tamaño de las 21974 filas de la base con los residuales, valores predichos,
- los grados de libertad (k, incluyendo el intercepto) y el residual de los mismos (n-k)
- los elementos auxiliares usador por R para llegar a la solución de la regresión mediante descomposición QR.
- el comando y la fórmula del modelo
- una lista de las variables tomadas como categóricas con más de dos niveles, así como el número de niveles de cada una de ellas
- una lista de las variables dicotómicas
- las proyecciones ortogonales en los subespacios producidos por la descomposición QR (*effects*).
- un vector de enteros que indica la columna a la que corresponde cada regresor del modelo.

En cuanto a la prueba F, recordemos que:

$$F = \frac{\frac{TSS-RSS}{k}}{\frac{RSS}{n-k-1}}$$

y en R la podemos calcular de la siguiente forma.

```
base<-base%>% mutate(y=log(quantity+1), pred=predict(model_a2))
RSS <- sum((base$y-base$pre)^2)
TSS<- sum((base$y-mean(base$y))^2)
n<- length(base$unit_price)
k<- sum(!is.na(model_a2$coefficients))
#como nuestra k incluye el intercepto, modificamos un poco la formula
(F <- ((TSS-RSS)/(k-1))/(RSS/(n-k)))
```

```
## [1] 11.05489
```

El valor coincide con el estadístico F que arroja R. Debido a que el valor es 1463.597 evidentemente rechazamos la hipótesis nula de que todas las β son cero.

16. Cuál es la elasticidad precio de los helados Ben and Jerry ? Es significativo? Interpreta el coeficiente

```
# para model_a2
x <- mean(base$unit_price)
y <- mean(base$quantity)
beta <- model_a2$coefficients[2]

elasticity <- beta



#p-value
tidy(model_a2)$p.value[2] %>% round(2)
```

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

Dado que el p-value es un valor muy cercano a cero, podemos rechazar la Hipótesis nula (beta=0) prácticamente para cualquier nivel de significancia y el coeficiente (elasticidad) es estadísticamente significativo.

```
# Significancia t
tidy(model_a2)$statistic[2]
```

```
## [1] -17.10683
```

Como el estadístico t es -17.1, podemos rechazar la Hipótesis nula para cualquier nivel de significancia.

```
# Elasticidad
round(elasticity,2)
```

```
## log_price
##     -0.17
```

La elasticidad precio de la demanda evaluada en la media es de -0.17. Por lo tanto, ante un aumento de 1% en el precio, la cantidad demandada se reduce en 0.17%.

Dado que el p-Value es menor a .001, podemos rechazar la Hipótesis nula ($\beta = 0$) prácticamente para cualquier nivel de significancia y el coeficiente (elasticidad) es estadísticamente significativo. De manera, paralela, como el estadístico t es menor a -17 podemos rechazar la Hipótesis nula para cualquier nivel de significancia.

Finalmente, la elasticidad precio de la demanda es de -.17%, esto quiere decir que ante un aumento de 1% en el precio, la cantidad demandada se reduce en .17%.

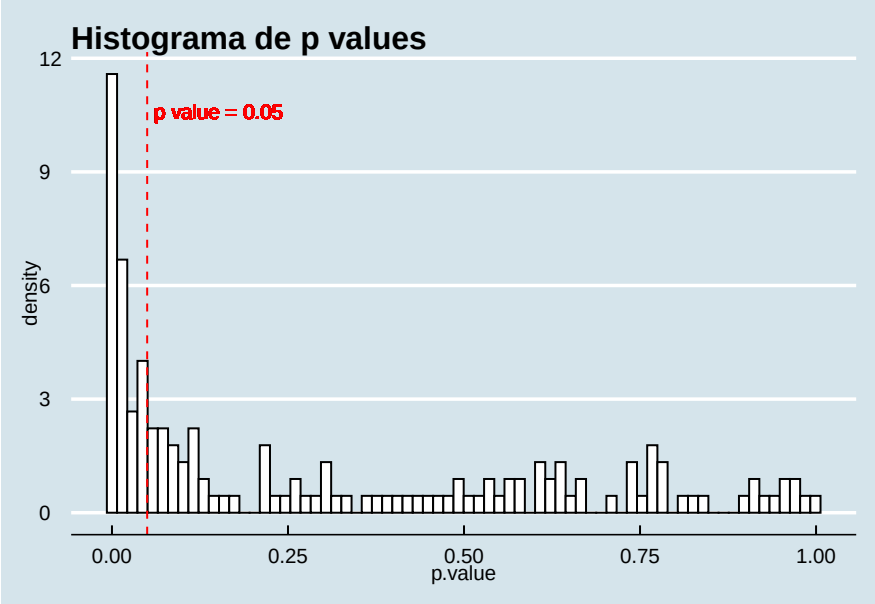
17. Cuántos p-values tenemos en la regresión. Haz un histograma de los p-values.

```
#Creo un dataframe con los pvalues del modelo
pvalues<-as.data.frame(tidy(model_a2))
#Quito los pvalues que aparecen como NAs
pvalues <- pvalues %>% filter(p.value != "NA") %>%
  select(term,p.value)%>%
  arrange(desc(p.value))

nrow(pvalues)
```

```
## [1] 155
```

Tenemos 155 p values. Notamos que el principio deberíamos tener más pero debido a la colinealidad perfecta entre algunas variables (por ejemplo las relacionadas con male_head y female_head), se redujeron de 161 a 155.

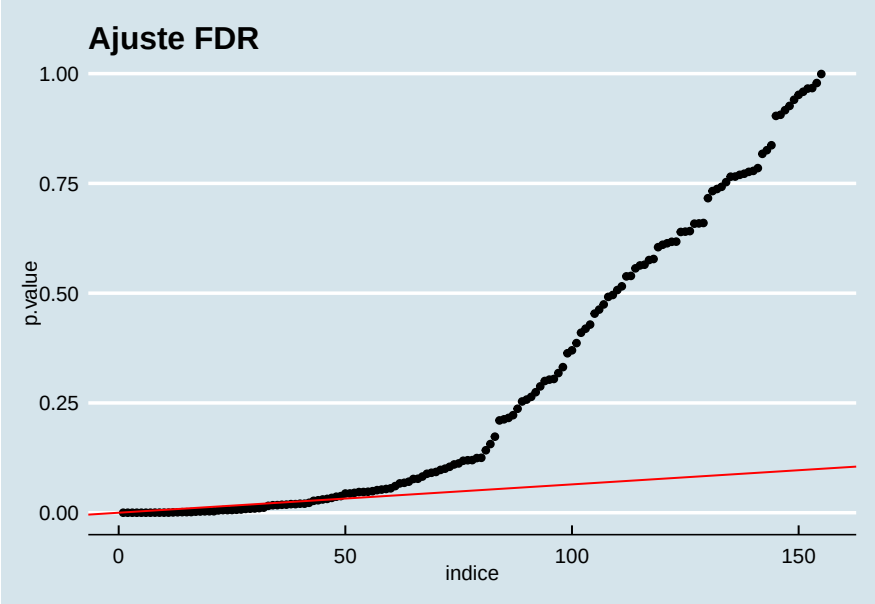
```
ggplot(pvalues, aes(x=p.value)) +
 geom_histogram(aes(y=..density..), bins=70, colour="black", fill="white")+
  geom_vline(xintercept = .05,size=0.5,colour="red", linetype = "dashed")+
  geom_text(aes(x=.28, label=paste("p value =",.05), y=10),
            size=4, colour="red", vjust = -1, hjust = 1.2)+
  labs(title="Histograma de p values") +
  theme_economist() +
  scale_fill_economist()
```



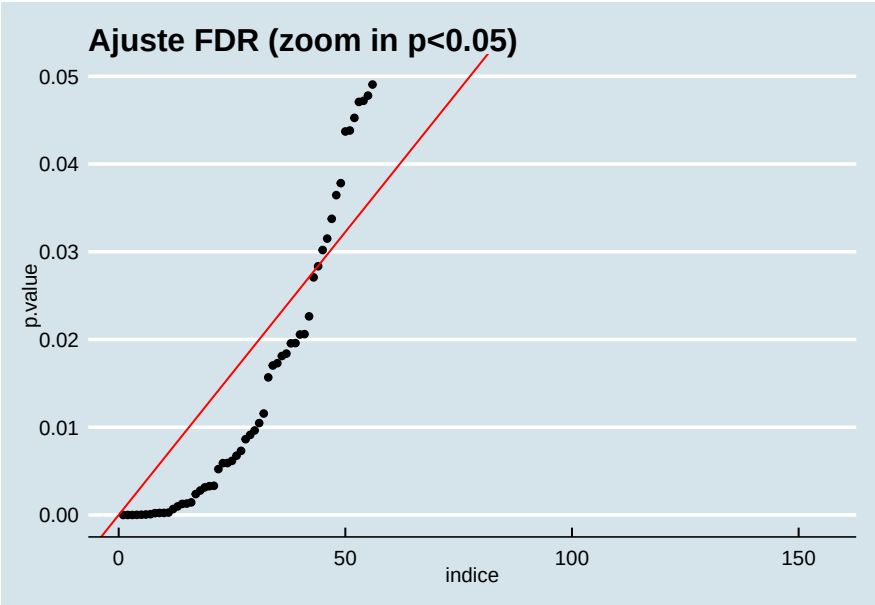
18 (4pts). Realiza un ajuste FDR a una $q = 0.10$. Grafica el procedimiento (con y sin zoom-in a p-values<0.05). Cuantas variables salían significativas con $\alpha = 0.05$? Cuantas salen con FDR?

Tip: crea el ranking de cada p-value como `resultados %>% arrange(p.value) %>% mutate(ranking = row_number)`

```
# FDR
q <- 0.1
  fdr <- pvalues %>% arrange(p.value) %>%
    mutate(indice = 1:length(p.value)) %>%
    mutate(`q*k/n` = q*(indice/length(p.value))) %>%
    mutate(significancia = p.value <= `q*k/n`)


ggplot(data =fdr, aes(x=indice,y=p.value))+
  geom_point()+
  geom_abline(slope=q/length(pvalues$p.value),col='red')+
  labs(title="Ajuste FDR") +
  theme_economist() +
  scale_fill_economist()
```




```
# con zoom
ggplot(data =fdr, aes(x=indice,y=p.value))+
  geom_point()+
  geom_abline(slope=q/length(pvalues$p.value),col='red')+
  labs(title="Ajuste FDR (zoom in p<0.05)") +
  theme_economist() +
  scale_fill_economist()+
  ylim(c(0,0.05))
```




```
# prueba
# abc <- c(0.52, 0.07, 0.013, 0.0001, 0.26, 0.04, 0.01, 0.15, 0.03, 0.0002)
```

```
# alternativa
# fdr2 <- p.adjust(p=pvalues$p.value,method="BH")
# names(fdr2) <- fdr2 <=0.05
```

¿Cuántas variables salían significativas? ¿Cuántas salen con FDR?

Antes 56 variables salían significativas. Después, con FDR sólo 44.

```
# Antes
sum(pvalues$p.value<=0.05)
```

```
## [1] 56
```

```
# Con FDR
pestrella <- fdr %>% filter(significancia == 'TRUE')
pestrella <- max(pestrella$`q*k/n`)

sum(pvalues$p.value<=pestrella)
```

```
## [1] 44
```

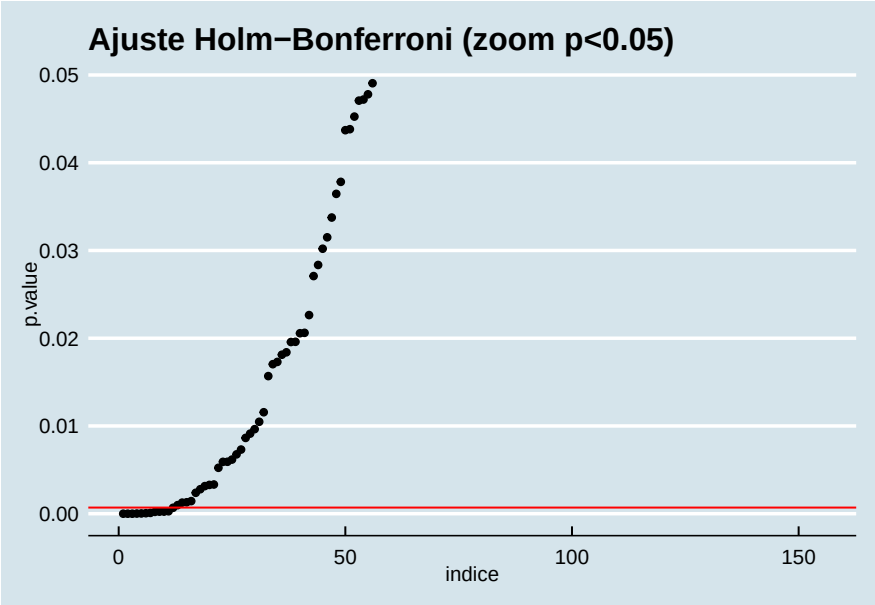
19 (2pts). Repite el ejercicio pero ahora con Holm-Bonferroni. Comparalo vs FDR. En este caso cuantas variables son significativas? Haz la grafica comparativa (solo con zoom-in)

```
q <- 0.1
  bonferroni <- pvalues %>% arrange(p.value) %>%
    mutate(indice = 1:length(p.value)) %>%
    mutate(`alpha/(m+1-k)`= q /(length(p.value)+1-indice)) %>%
    mutate(significancia = p.value <= `alpha/(m+1-k)`)

  sum(bonferroni$significancia=='TRUE')
```

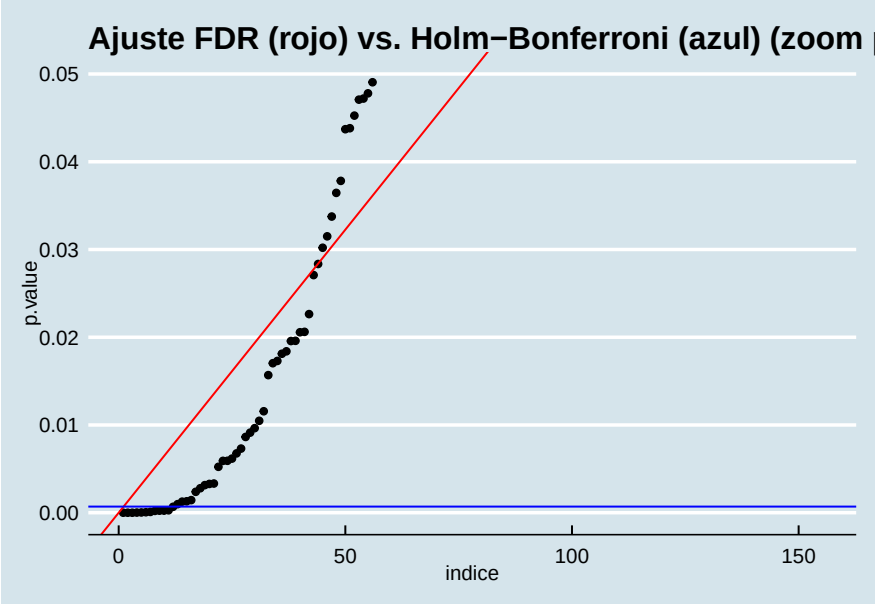
```
## [1] 12
```

```
  # Tenemos únicamente 12 variables que resultan ser significativas
  # con el ajuste de Holm-Bonferroni


corte <- bonferroni %>% filter(significancia == 'FALSE')
corte <- min(corte$`alpha/(m+1-k)`)
```

```
# Gráfica con zoom
ggplot(data =bonferroni, aes(x=indice,y=p.value))+
  geom_point()+
  geom_hline(yintercept= corte,col='red')+
  labs(title="Ajuste Holm-Bonferroni (zoom p<0.05)")+
  theme_economist() + scale_fill_economist()+
  ylim(c(0,0.05))
```




```
# comparación FDR vs. Holm-Bonferroni
ggplot()+
  geom_point(data=fdr, aes(x=indice,y=p.value))+
  geom_abline(slope=q/length(pvalues$p.value),col='red')+
  geom_point()+
  geom_hline(aes(x=indice,y=p.value, linetype="H-B"),
             yintercept= corte,col='blue',show.legend=TRUE)+
  labs(title="Ajuste FDR (rojo) vs. Holm-Bonferroni (azul) (zoom p<0.05)") +
  theme_economist() + scale_fill_economist()+
  ylim(c(0,0.05))
```



Tenemos únicamente 12 variables que resultan ser significativas con el ajuste de Holm-Bonferroni. Como vimos en clase, el ajuste Holm-Bonferroni es más estricto que el ajuste FDR. Controla más por el error tipo 1. A continuación, se muestran las variables que resultan ser significativas para cada ajuste:

```
coefs <- merge(fdr,bonferroni,by=c("term","p.value")) %>%
          mutate(significancia_clasica = case_when(p.value<0.05 ~ TRUE,
                                                   TRUE~ FALSE)) %>%
          arrange(p.value)


coefs <- coefs %>%
  select(term,p.value,significancia.y,significancia.x,significancia_clasica)
coefs$p.value <- round(coefs$p.value,5)
names(coefs) <- c("término","p_value","bonferroni","fdr","significancia_clasica")

coefs <- coefs[(coefs$fdr==TRUE |
                  coefs$bonferroni==TRUE |
                  coefs$significancia_clasica==TRUE),]

kable(coefs,
  col.names =
    c("Término","p-value","Bonferroni","FDR","Por significancia"),booktabs=T)%>%
  kable_styling(position = "center")
```

Término	p-value	Bonferroni	FDR	Por significancia
(Intercept)	0.00000	TRUE	TRUE	TRUE
log_price	0.00000	TRUE	TRUE	TRUE
household_size	0.00000	TRUE	TRUE	TRUE
type_of_residence5	0.00002	TRUE	TRUE	TRUE
age_and_presence_of_children	0.00003	TRUE	TRUE	TRUE
type_of_residence2	0.00006	TRUE	TRUE	TRUE
flavor_descrCREME BRULEE	0.00009	TRUE	TRUE	TRUE
flavor_descrPUMPKIN CSK	0.00021	TRUE	TRUE	TRUE
flavor_descrHEATH COFFEE CRUNCH	0.00023	TRUE	TRUE	TRUE
age_of_female_head8	0.00024	TRUE	TRUE	TRUE
flavor_descrCINNAMON BUNS	0.00027	TRUE	TRUE	TRUE
flavor_descrCHC FUDGE BROWNIE	0.00067	TRUE	TRUE	TRUE
promotion_type1	0.00098	FALSE	TRUE	TRUE
flavor_descrVAN CARAMEL FUDGE	0.00126	FALSE	TRUE	TRUE
female_head_occupation6	0.00130	FALSE	TRUE	TRUE
flavor_descrBLACK & TAN	0.00144	FALSE	TRUE	TRUE
promotion_type3	0.00238	FALSE	TRUE	TRUE
household_income29	0.00279	FALSE	TRUE	TRUE
flavor_descrPISTACHIO PISTACHIO	0.00314	FALSE	TRUE	TRUE
male_head_occupation5	0.00328	FALSE	TRUE	TRUE
household_income27	0.00333	FALSE	TRUE	TRUE
age_of_female_head7	0.00523	FALSE	TRUE	TRUE
household_income10	0.00591	FALSE	TRUE	TRUE
male_head_occupation2	0.00592	FALSE	TRUE	TRUE
female_head_occupation3	0.00617	FALSE	TRUE	TRUE
type_of_residence4	0.00676	FALSE	TRUE	TRUE
female_head_occupation7	0.00731	FALSE	TRUE	TRUE
age_of_female_head6	0.00864	FALSE	TRUE	TRUE
male_head_occupation9	0.00913	FALSE	TRUE	TRUE
female_head_education4	0.00964	FALSE	TRUE	TRUE
flavor_descrCHERRY GRCA	0.01048	FALSE	TRUE	TRUE
male_head_occupation8	0.01156	FALSE	TRUE	TRUE
size1_descr32.0 MLOZ	0.01568	FALSE	TRUE	TRUE
male_head_education1	0.01704	FALSE	TRUE	TRUE
household_income15	0.01730	FALSE	TRUE	TRUE
flavor_descrAMERICONE DREAM	0.01812	FALSE	TRUE	TRUE
female_head_occupation9	0.01839	FALSE	TRUE	TRUE
type_of_residence6	0.01957	FALSE	TRUE	TRUE
household_income28	0.01959	FALSE	TRUE	TRUE
household_income8	0.02057	FALSE	TRUE	TRUE

Como comentario final, era de esperarse que el precio fuera la variable con mayor significancia y que se incluyera bajo los tres criterios de selección. Otras variables con alto poder explicativo fueron algunos sabores de helado, dos tipos de promoción, la edad y presencia de hijos y características del hogar y su composición.