Economía Computacional: Tarea 1

Alfredo Lefranc, Marco Ramos, Rafael Sandoval y Cynthia Valdivia

2021

```
library(tidyverse)
library(data.table)
library(RCT)
library(knitr)
library(lfe)
library(broom)
library(stargazer)
library(kableExtra)
library(naniar)
library(nnet)
```

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 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"))
```

Columnas
quantity
price_paid_deal
price_paid_non_deal
coupon_value
promotion_type
size1_descr
flavor_descr
formula_descr
household_id
household_size
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.

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. De tal forma, podemos usar el identificador de y concatenarlo con una nueva variable para identificar la transacción por hogar para crear un identificador único a nivel de la base.

```
base2 <- base %>% group_by(household_id) %>%
          mutate(idper = sequence(n()))
base$idper <- base2$idper

base$id <- base$idper + base$household_id*100
```

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("Variable","Cantidad","%"))
```

Variable	Cantidad	%
promotion_type	12980	59.0698098
male_head_birth	5317	24.1967780
scantrack_market_identifier	4068	18.5127878
female_head_occupation	2267	10.3167380
female_head_birth	2267	10.3167380
tv_items	34	0.1547283

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 si hay un patrón en los valores vacíos y las características de cada variable podemos proponer una estrategia, por ejemplo imputación o 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
```

```
susp<-base%>% select(fips_state_code,fips_county_code,type_of_residence,scantrack_market_identifier)
```

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 varaibles, 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.

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 estás características con el número 0, la cual imputamos a los valores faltantes. Por su parte, dejamos como NAs los valores faltantes de la variable *female_head_birth*.

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

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.

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')
```

variable	mean	n	mín	0.05	0.25	0.5	0.75	
quantity	1.28	21974	1	1	1	1.00	1.00	
price_paid_deal	1.74	21974	0	0	0	0.00	3.34	
price_paid_non_deal	2.45	21974	0	0	0	2.99	3.56	
coupon_value	0.16	21974	0	0	0	0.00	0.00	
promotion_type	1.44	8994	1	1	1	1.00	2.00	
household_id	16612005.04	21974	2000358	2054762	8142253	8401573.00	30183891.00	3033
household_size	2.46	21974	1	1	2	2.00	3.00	
household_income	21.47	21974	3	11	17	23.00	26.00	2
age_of_female_head	5.51	21974	0	0	4	6.00	8.00	
age_of_male_head	4.76	21974	0	0	2	5.00	8.00	
age_and_presence_of_children	7.40	21974	1	2	6	9.00	9.00	
male_head_employment	3.09	21974	0	0	1	3.00	3.00	
female_head_employment	4.20	21974	0	0	2	3.00	9.00	
male_head_education	3.32	21974	0	0	2	4.00	5.00	
female_head_education	3.98	21974	0	0	3	4.00	5.00	
marital_status	1.94	21974	1	1	1	1.00	3.00	
male_head_occupation	5.11	21974	1	1	1	4.00	8.00	1
female_head_occupation	5.80	19707	1	1	1	3.00	12.00	1
household_composition	2.57	21974	1	1	1	1.00	5.00	
race	1.24	21974	1	1	1	1.00	1.00	
hispanic_origin	1.95	21974	1	2	2	2.00	2.00	
region	2.63	21974	1	1	2	3.00	4.00	
scantrack_market_identifier	23.05	17906	1	1	11	20.00	36.00	4
fips_state_code	27.20	21974	1	6	12	26.00	39.00	4
fips_county_code	79.67	21974	1	3	25	59.00	101.00	1
type_of_residence	2.08	21974	1	1	1	1.00	3.00	
kitchen_appliances	3.81	21974	1	1	4	4.00	4.00	
tv_items	1.93	21940	1	1	1	2.00	3.00	
household_internet_connection	1.16	21974	1	1	1	1.00	1.00	

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 las siguientes son categóricas: *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* y *household_internet_connection.*

Sin embargo, las siguientes podrían ser o no ser categóricas: *tv_items, kitchen_appliances, age_and_presence_of_children, male_head_education, female_head_education.*

```
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.

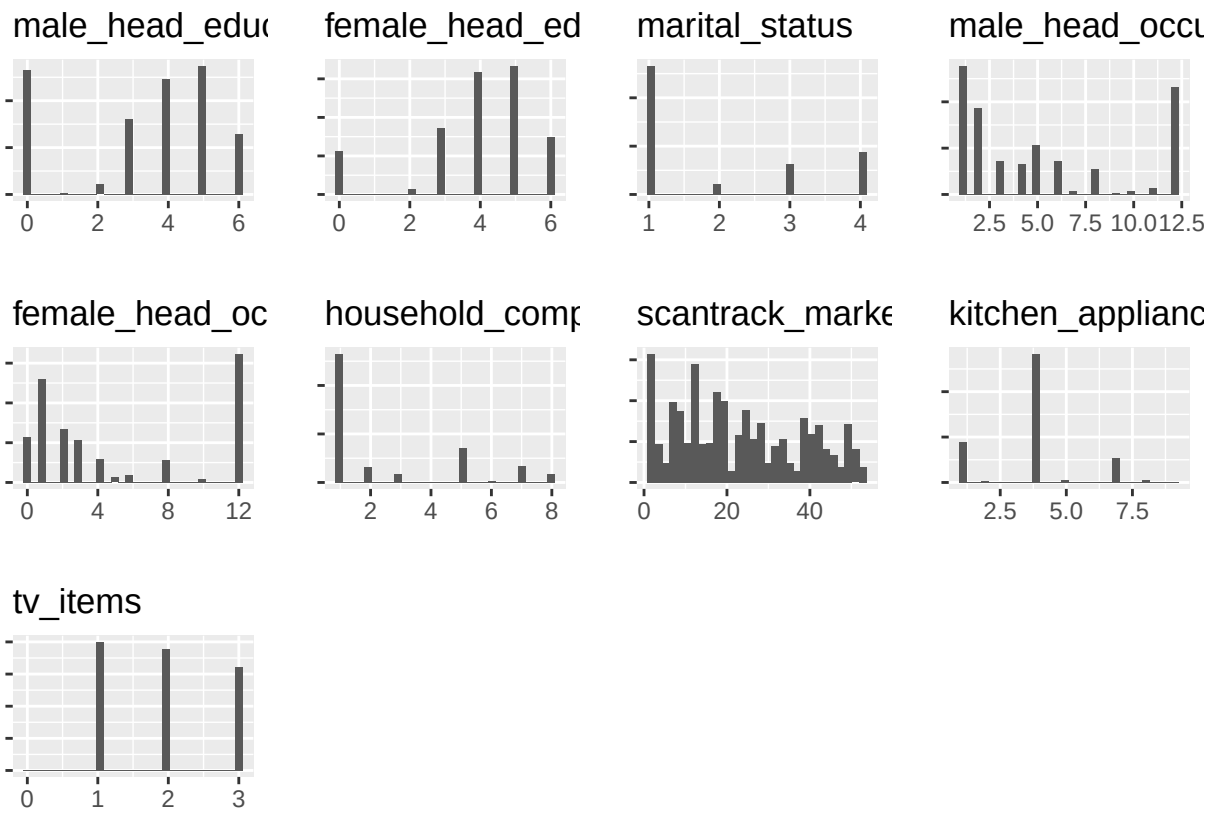
```
variables_numericas<-c("tv_items","kitchen_appliances","age_and_presence_of_children")
base[,variables_numericas] <- lapply(base[,variables_numericas] , as.numeric)
```

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)
```



No. Las edades 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).

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.

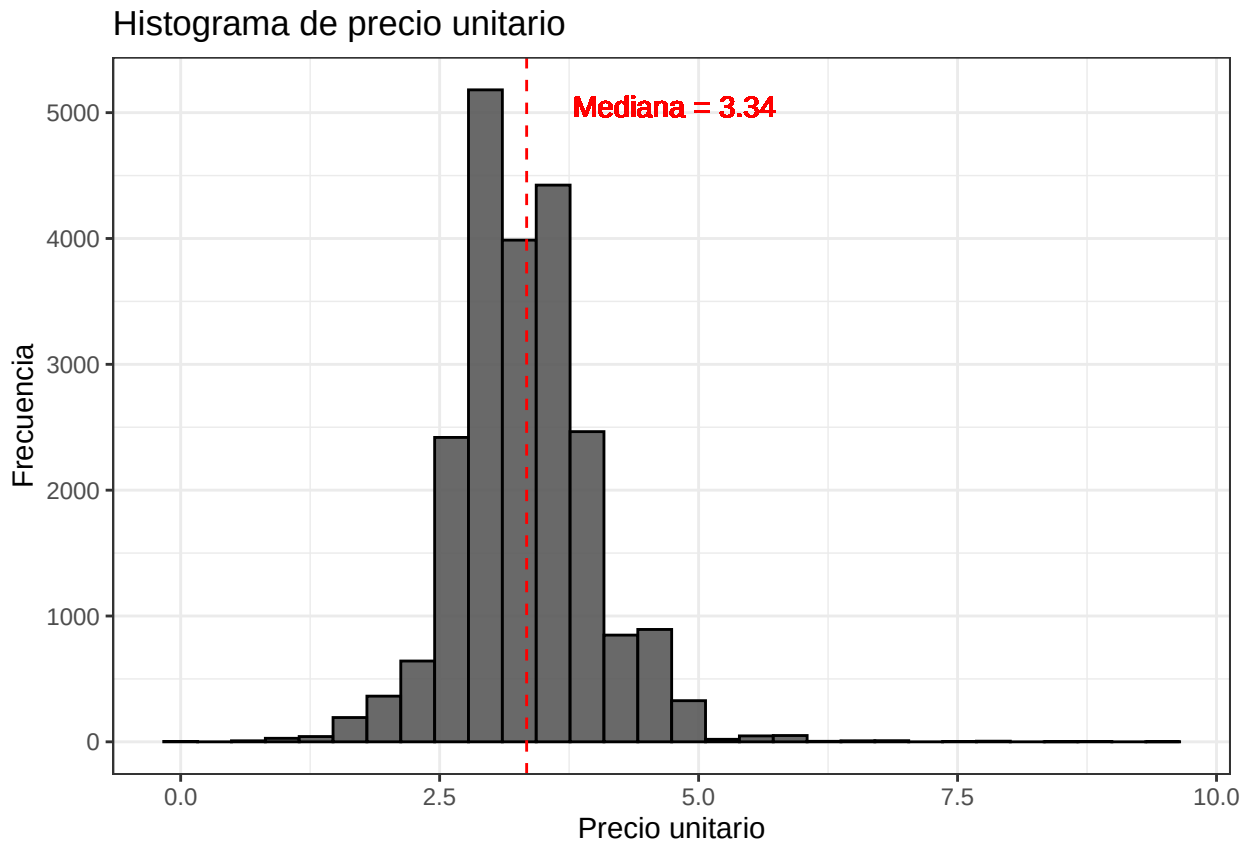
Algo importante es que siempre debemos mirar primero las variables más relevantes de cerca y su relación en:

- Relación univariada
- Relaciones bivariadas
- Relaciones trivariadas

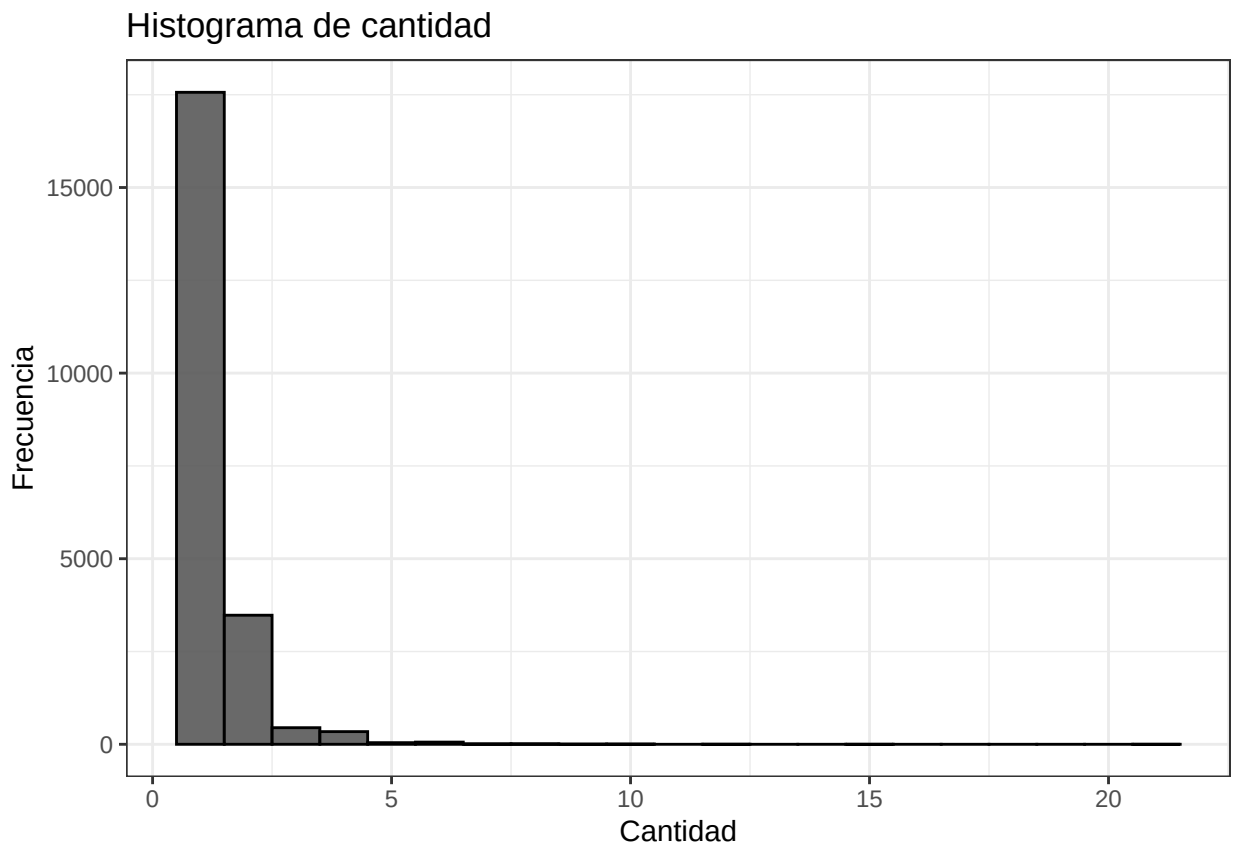
Importante: Las gráficas deben estar bien documentadas (título, ejes con etiquetas apropiadas, etc). Cualquier gráfica que no cumpla con estos requisitos les quitaré algunos puntos.

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"
  theme_bw()+
  labs(title="Histograma de precio unitario",x="Precio unitario",y="Frecuencia")
```



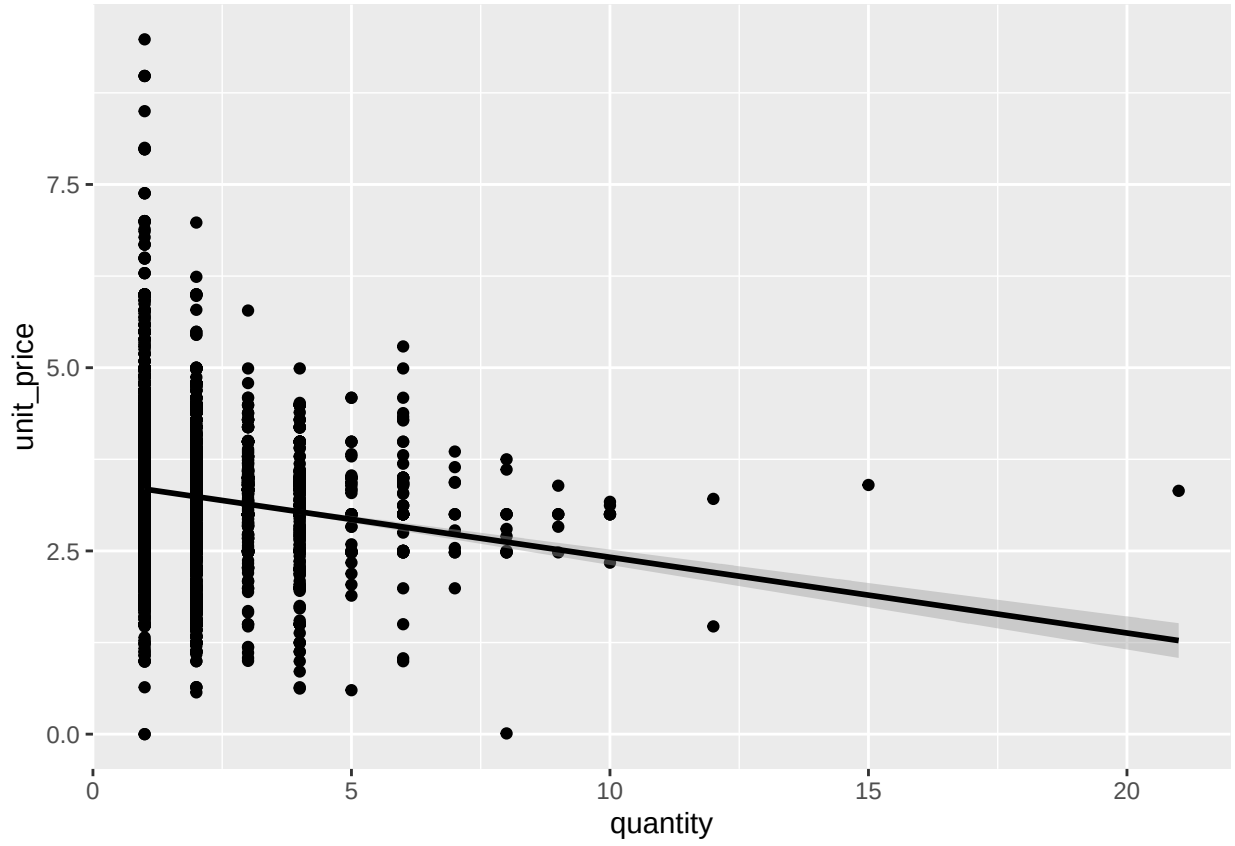

```
ggplot(base)+
  geom_histogram(aes(x=quantity),binwidth=1,alpha=0.9,col = 'black')+
  theme_bw()+
  labs(title="Histograma de cantidad",x="Cantidad",y="Frecuencia")
```



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

Aunque parece haber una relación negativa, esta no es tan clara.

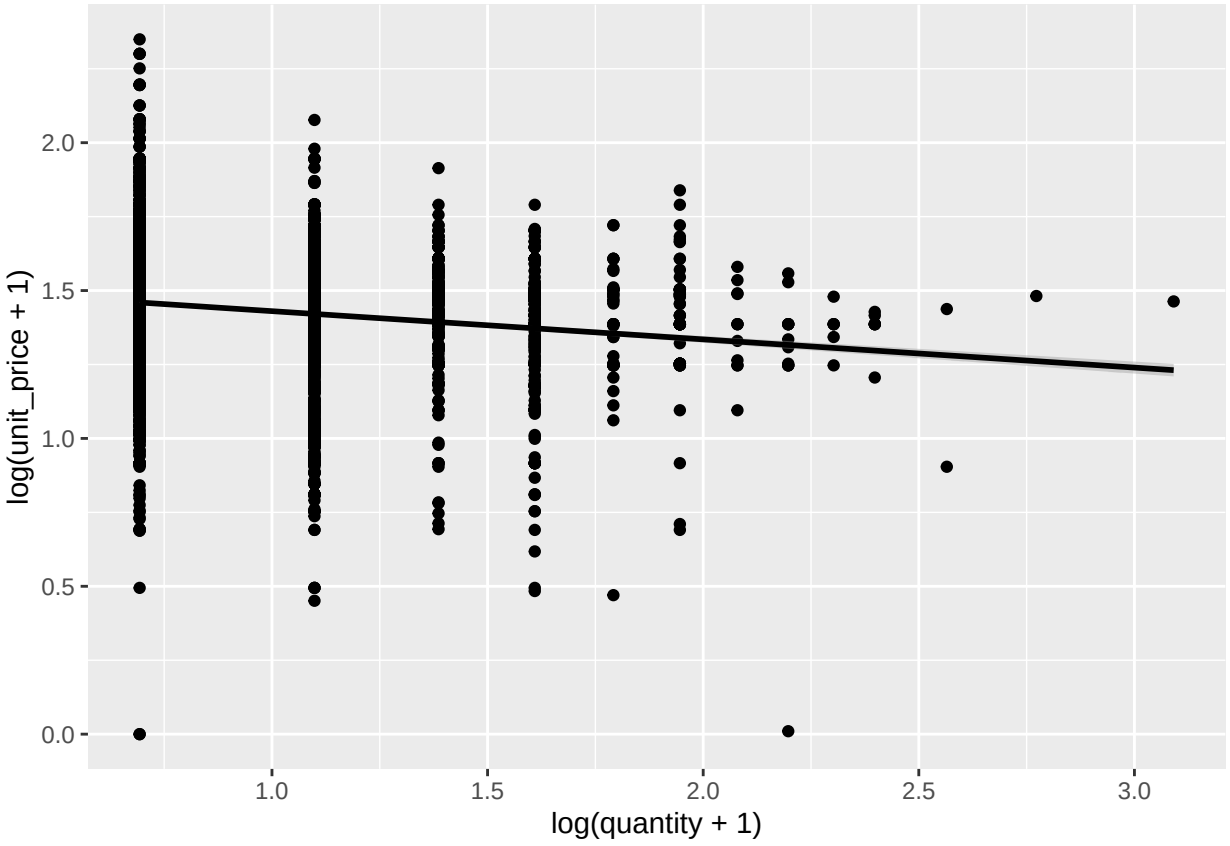
```
ggplot(base)+
  geom_point(aes(x=quantity,y=unit_price))+
  geom_smooth(formula=y~x,method=lm, color='1',aes(x = quantity, y = unit_price))
```



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

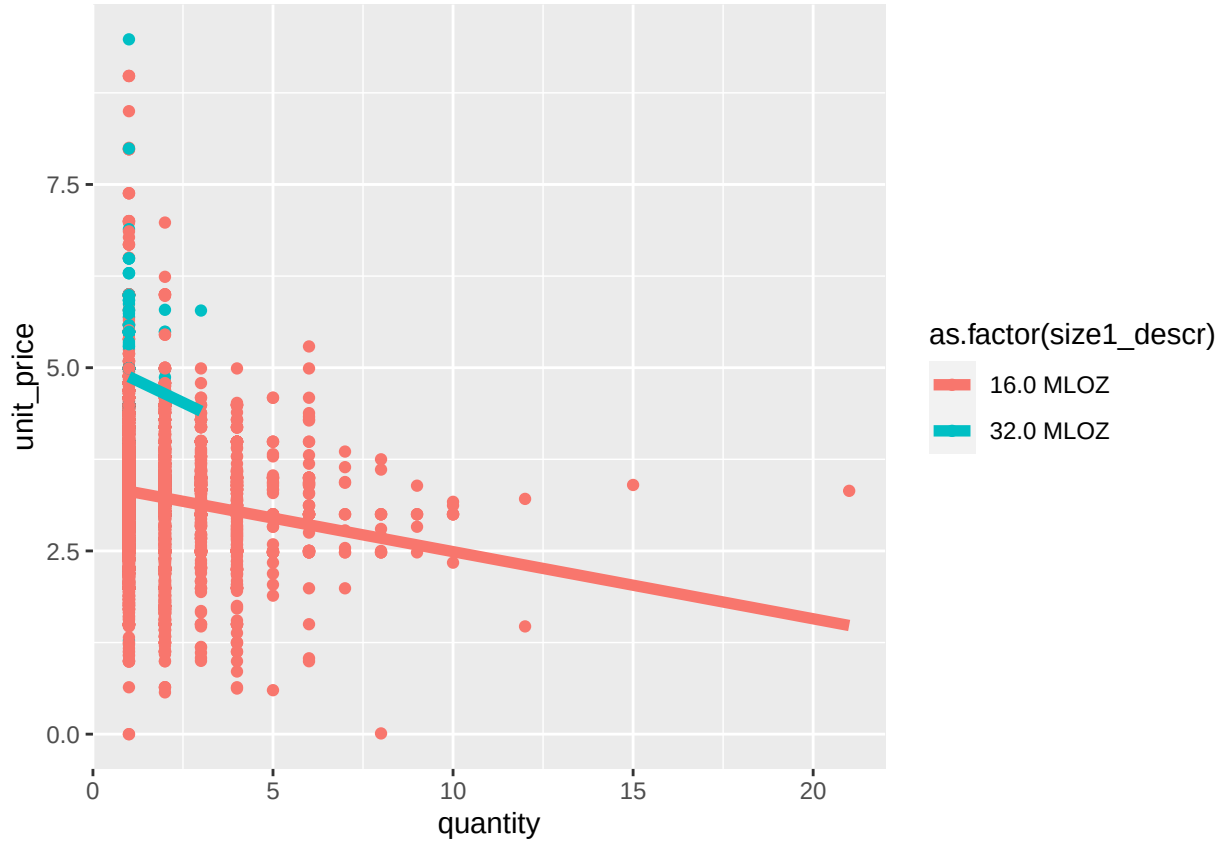
Cuando hacemos la transformación, la relación es más evidente:

```
ggplot(base)+
  geom_point(aes(x=log(quantity+1),y=log(unit_price+1)))+
  geom_smooth(formula=y~x,method=lm, color='1',aes(x = log(quantity+1), y = log(unit_price+1)))
```



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(x=quantity, y=unit_price,col=as.factor(size1_descr))) +
  geom_point() +
  geom_smooth(method='lm',
              formula= y~(x),
              se=FALSE, size=2)
```



HACER UNA PRUEBA DE HIPOTESIS

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?

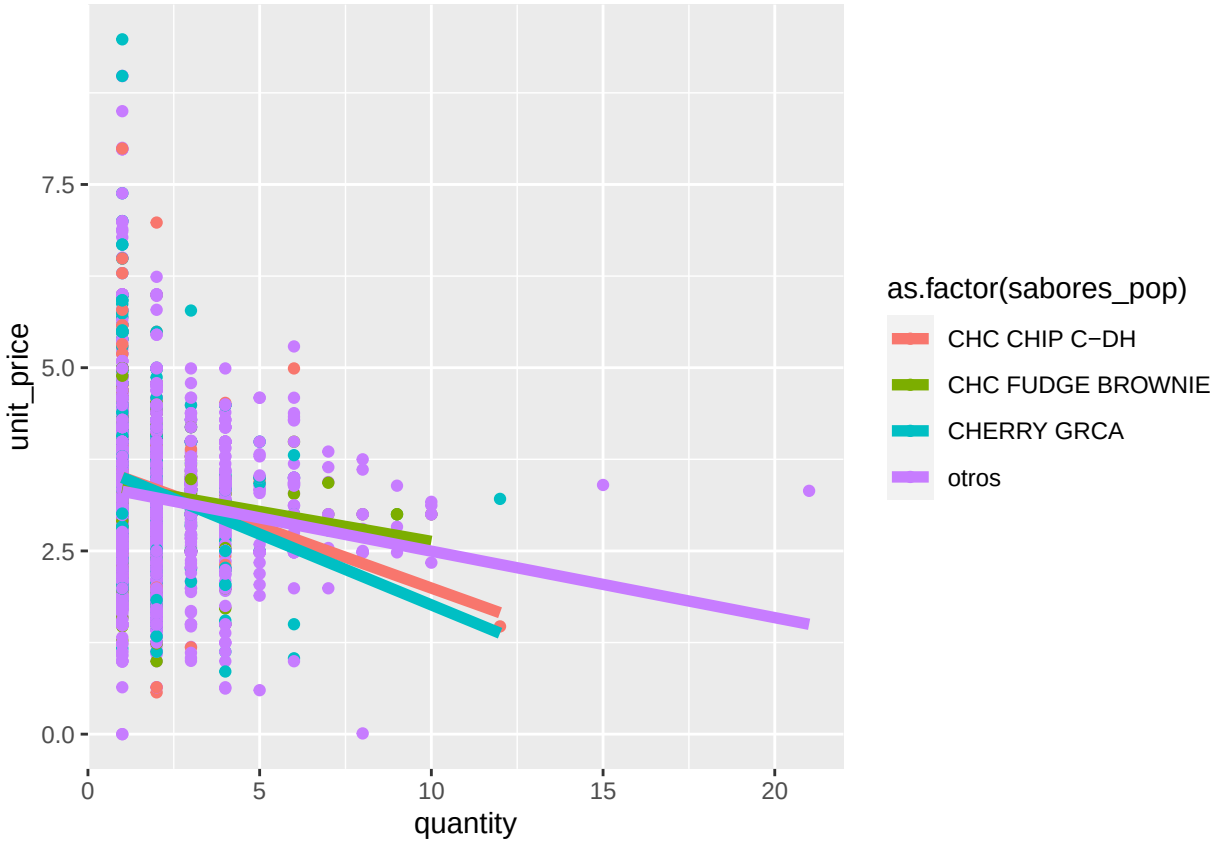
```
summary(factor(base$flavor_descr))
```

```
##                AMERICONE DREAM                   BANANA SPLIT
##                            865                            599
##                   BLACK & TAN                 BROWNIE BATTER
##                             25                            146
##                   BUTTER PECAN                   CAKE BATTER
##                            241                            409
##                            CHC              CHC ALMOND NOUGAT
##                             97                            120
##                  CHC CHIP C-DH              CHC FUDGE BROWNIE
##                           1070                           1235
##                    CHERRY GRCA                  CHUBBY HUBBY
##                           2097                            318
##                  CHUNKY MONKEY                 CINNAMON BUNS
##                           1064                            614
##                         COFFEE                   CREME BRULEE
##                             56                            455
##            DOUBLE CHC FUDGE SWR                DUBLIN MUDSLIDE
##                              1                            370
##                    FOSSIL FUEL                    HALF BAKED
```

```
##                              84                          704
## HEATH CANDY EVERYTHING BUT THE            HEATH COFFEE CRUNCH
##                             527                         1070
##                    HEATH CRUNCH          IMAGINE WHIRLED PEACE
##                             493                          612
##                  KARAMEL SUTRA                MAGIC BROWNIES
##                             738                          199
##                 MINT CHC CHUNK            NEAPOLITAN DYNAMITE
##                             146                          190
##     NEW YORK SUPER FUDGE CHUNK           OATMEAL COOKIE CHUNK
##                             932                          184
##                ONE CSK BROWNIE         OXFORD MINT CHC COOKIE
##                             557                          326
##                          PB CUP                    PB TRUFFLE
##                             828                            1
##                     PHISH FOOD            PISTACHIO PISTACHIO
##                             968                          723
##                    PUMPKIN CSK                 RSP CHC CHUNK
##                             143                           79
##                          SMORES                          STR
##                             200                           11
##                         STR CSK           STRAWBERRIES & CREAM
##                             515                           13
##          SWEET CREAM & COOKIES           TRIPLE CARAMEL CHUNK
##                              17                           87
##                    TURTLE SOUP                          VAN
##                             204                          517
##               VAN CARAMEL FUDGE                VERMONTY PYTHON
##                             290                          134
##                       W-N-C-P-C                 WHITE RUSSIAN
##                             699                            1
```

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

```
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(x=quantity, y=unit_price,col=as.factor(sabores_pop))) +
  geom_point() +
  geom_smooth(method='lm',
              formula= y~(x),
              se=FALSE, size=2)
```



PRUEBA DE HIPOTESIS

Estimación

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

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

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: dom., feb. 07, 2021 - 04:52:45 p. m.

```
model_a2<-lm(quantity~unit_price,data = base)
stargazer(model_a2, type = "latex", title="Regresión", digits=1)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: dom., feb. 07, 2021 - 04:52:45 p. m.

CORREGIR

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.

Table 1: Regresión

	Dependent variable:
	unit_price
quantity	-0.1^{***}
	(0.01)
Constant	3.4^{***}
	(0.01)
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

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

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 número.

Quitemos las fechas

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

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)

```
# para model_a2
(names(model_a2))
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
glance(model_a2)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1    0.0128        0.0127 0.723      284. 2.83e-63     1 -24043. 48092. 48116.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
RSS <- sum(model_a2$residuals^2)
TSS<- sum((base$quantity-mean(base$quantity))^2)
n<- length(base$unit_price)
k<- 1

(F <- ((TSS-RSS)/k)/(RSS/(n-k-1)))
```

```
## [1] 283.7687
```

```
summary(model_a2)
```

```
##
## Call:
## lm(formula = quantity ~ unit_price, data = base)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6909 -0.3207 -0.2591 -0.1369 19.7187
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  1.690909   0.024763   68.28 <0.0000000000000002 ***
## unit_price  -0.123388   0.007325  -16.84 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7227 on 21972 degrees of freedom
## Multiple R-squared:  0.01275,    Adjusted R-squared:  0.01271
## F-statistic: 283.8 on 1 and 21972 DF,  p-value: < 0.00000000000000022
```

```
# sí es igual
```

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 * (x/y)
# A primera vista sí es significativo
paste("La elasticidad precio de la demanda es de",round(elasticity,2),"%")
```

```
## [1] "La elasticidad precio de la demanda es de -0.32 %"
```

```
paste("Ante un aumento de 1% en el precio, la cantidad demandada se reduce en",round(elasticity,2),"%")
```

```
## [1] "Ante un aumento de 1% en el precio, la cantidad demandada se reduce en -0.32 %"
```

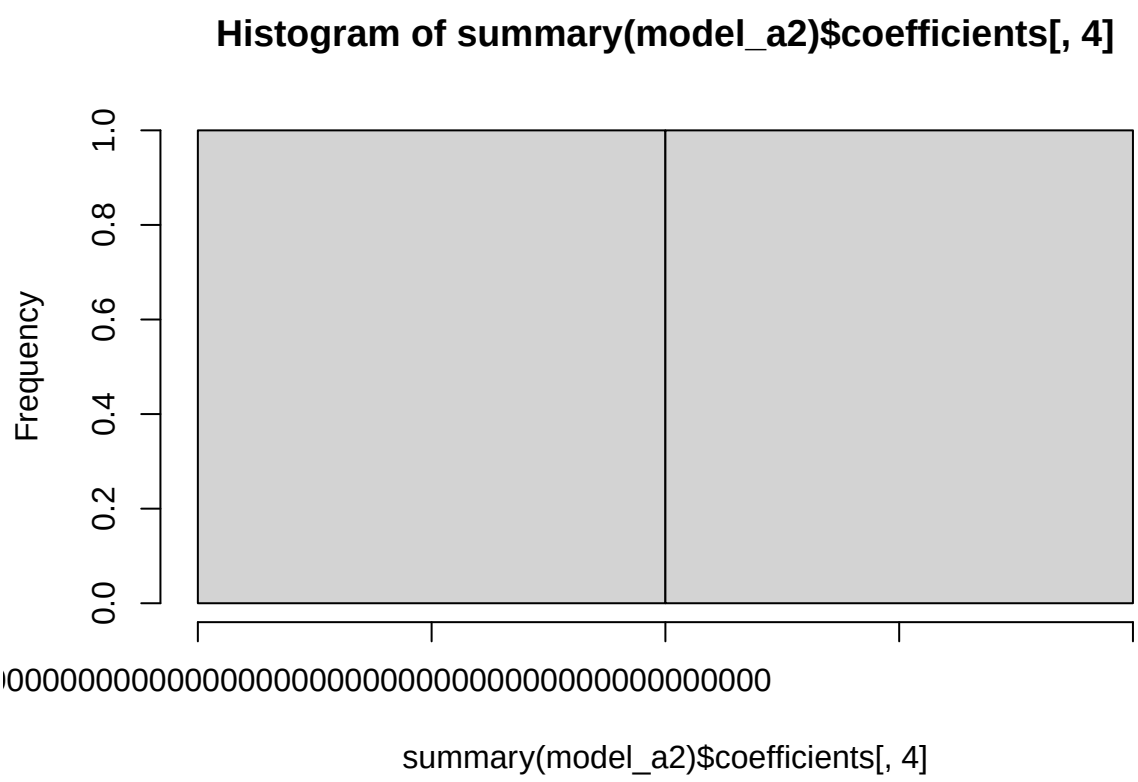
```
beta2 <- model_a$coefficients[2]
beta2*(y/x)
```

```
##    quantity
## -0.03996474
```

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

Tenemos 2, uno para $beta_0$ y otro para β_1 OJO: CREO QUE SE REFIERE A OTRA COSA, HAY QUE CORRER REGRESION CON TOOOOOOODAS LAS VARAIABLES

```
valores_p <- summary(model_a2)$coefficients[,4]
hist(summary(model_a2)$coefficients[,4])
```

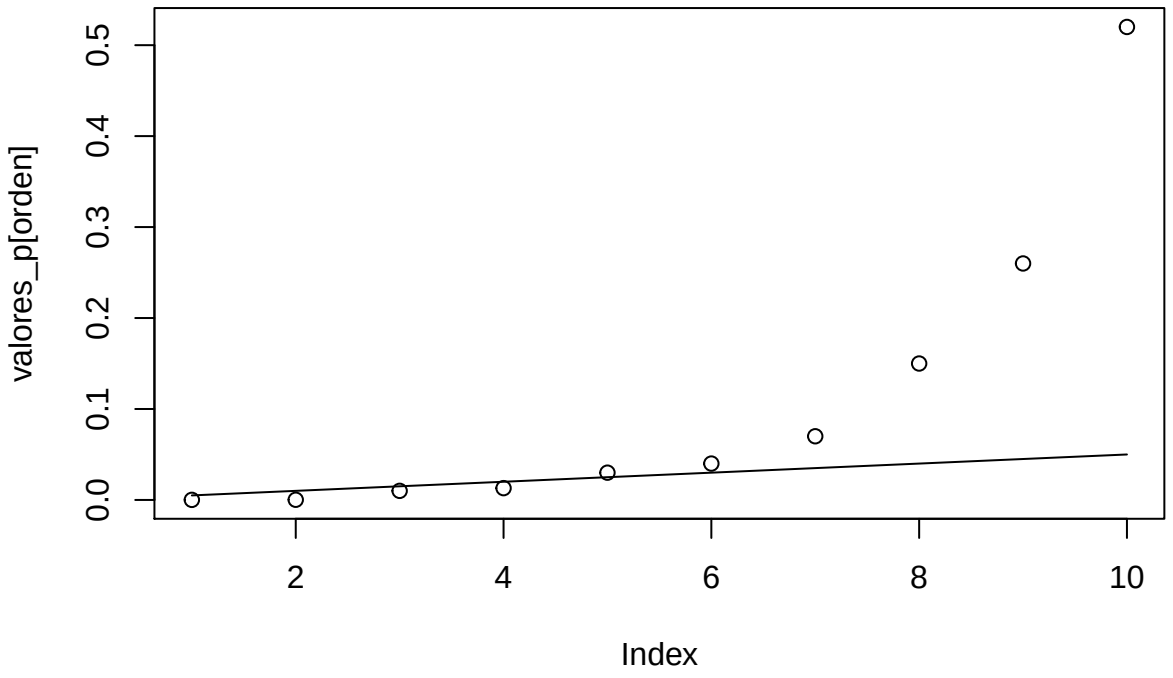


18 (4pts). Realiza un ajuste FDR a una $q = 0.10$. Grafica el procedimiento (con y sin zoom-in a p-values<0.05). Cuantas variables salían significativas con $\alpha = 0.05$? Cuantas salen con FDR?

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

```
# Función(vector de valores p, q)

fdr <- function(valores_p,q){
  valores_p <- valores_p[!is.na(valores_p)]
  n <- length(valores_p)

  k <- rank (valores_p,ties.method="min")
  t <- valores_p <= q*k/n
  t[valores_p<max(valores_p[t])] <- TRUE

  orden <- order(valores_p)
  plot(valores_p[orden])
  lines(1:n, q*(1:n)/n)

  return(max(valores_p[t]))

}

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

```
fdr(abc,0.05)
```




```
## [1] 0.013
```

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

```
##      FALSE      FALSE       TRUE       TRUE      FALSE      FALSE       TRUE
## 0.52000000 0.10000000 0.03250000 0.00100000 0.28888889 0.06666667 0.03250000
##      FALSE      FALSE       TRUE
## 0.18750000 0.06000000 0.00100000
```

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)