# Economía Computacional: Tarea 1

Isidoro Garcia

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 podriamos 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 si hay una manera implícita y es el índice de cada fila. Sin embargo, podemos crear una variable que contenga la información del índice de fila.

```
base<- base %>% rowid_to_column("ID")
```

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

```
kable( (base %>% select_if(~sum(is.na(.)) > 0) %>% miss_var_summary()), booktabs=T, align = 'c', col.nam
```

| Variable | Cantidad | % |
|---|---|---|
| promotion_type | 12980 | 59.0698098 |
| scantrack_market_identifier | 4068 | 18.5127878 |
| female_head_occupation | 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 5 o 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**

```
aux<-base %>% select(female_head_occupation,female_head_education,female_head_employment)%>%
  filter (is.na(female_head_occupation))
aux$female_head_education<-as.factor(aux$female_head_education)
aux$female_head_employment<-as.factor(aux$female_head_employment)
```

```
summary((aux))
```

```
##  female_head_occupation female_head_education female_head_employment
##  Min.   : NA            0:2267                0:2267
##  1st Qu.: NA
##  Median : NA
##  Mean   :NaN
##  3rd Qu.: NA
##  Max.   : NA
##  NA's   :2267
```

Notamos que todas las observaciones con valor faltante coinciden con los valores 0 de las columnas female_head_employment y female_head_education. Es muy probable que sean mujeres sin ocupación (tal vez ama de casa no lo están contemplando como ocupación). En este sentido crear una nueva categoría de mujeres con estás características con el número de 0 la cuál imputaremos a los valores faltantes.

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

**tv_items**

En este caso, puede que la variable indíque una cantidad de *items* o bien que indique una categoría. En el caso primero, parecería que no contemplaron una cantidad de 0s o de más de 3, bien podríamos imputar el valor de 0. En el segundo caso, no tenemos manera de saber que 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).

```
summary(factor(base$tv_items))
```

```
##    1    2    3 NA's
## 7986 7530 6424   34
```

**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<- summary_statistics(base,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)
kable(b,booktabs=T, align = 'c')
```

| variable | mean | n | mín | 0.05 | 0.25 | 0.5 | 0.75 |
|---|---|---|---|---|---|---|---|
| ID | 10987.50 | 21974 | 1 | 1099.65 | 5494.25 | 10987.50 | 16480.75 |
| quantity | 1.28 | 21974 | 1 | 1.00 | 1.00 | 1.00 | 1.00 |
| price_paid_deal | 1.74 | 21974 | 0 | 0.00 | 0.00 | 0.00 | 3.34 |
| price_paid_non_deal | 2.45 | 21974 | 0 | 0.00 | 0.00 | 2.99 | 3.56 |
| coupon_value | 0.16 | 21974 | 0 | 0.00 | 0.00 | 0.00 | 0.00 |
| promotion_type | 0.59 | 21974 | 0 | 0.00 | 0.00 | 0.00 | 1.00 |
| household_id | 16612005.04 | 21974 | 2000358 | 2054762.00 | 8142253.00 | 8401573.00 | 30183891.00 |
| household_size | 2.46 | 21974 | 1 | 1.00 | 2.00 | 2.00 | 3.00 |
| household_income | 21.47 | 21974 | 3 | 11.00 | 17.00 | 23.00 | 26.00 |
| age_of_female_head | 5.51 | 21974 | 0 | 0.00 | 4.00 | 6.00 | 8.00 |
| age_of_male_head | 4.76 | 21974 | 0 | 0.00 | 2.00 | 5.00 | 8.00 |
| age_and_presence_of_children | 7.40 | 21974 | 1 | 2.00 | 6.00 | 9.00 | 9.00 |
| male_head_employment | 3.09 | 21974 | 0 | 0.00 | 1.00 | 3.00 | 3.00 |
| female_head_employment | 4.20 | 21974 | 0 | 0.00 | 2.00 | 3.00 | 9.00 |
| male_head_education | 3.32 | 21974 | 0 | 0.00 | 2.00 | 4.00 | 5.00 |
| female_head_education | 3.98 | 21974 | 0 | 0.00 | 3.00 | 4.00 | 5.00 |
| marital_status | 1.94 | 21974 | 1 | 1.00 | 1.00 | 1.00 | 3.00 |
| male_head_occupation | 5.11 | 21974 | 1 | 1.00 | 1.00 | 4.00 | 8.00 |
| female_head_occupation | 5.20 | 21974 | 0 | 0.00 | 1.00 | 3.00 | 12.00 |
| household_composition | 2.57 | 21974 | 1 | 1.00 | 1.00 | 1.00 | 5.00 |
| race | 1.24 | 21974 | 1 | 1.00 | 1.00 | 1.00 | 1.00 |
| hispanic_origin | 1.95 | 21974 | 1 | 2.00 | 2.00 | 2.00 | 2.00 |
| region | 2.63 | 21974 | 1 | 1.00 | 2.00 | 3.00 | 4.00 |
| scantrack_market_identifier | 23.05 | 17906 | 1 | 1.00 | 11.00 | 20.00 | 36.00 |
| fips_state_code | 27.20 | 21974 | 1 | 6.00 | 12.00 | 26.00 | 39.00 |
| fips_county_code | 79.67 | 21974 | 1 | 3.00 | 25.00 | 59.00 | 101.00 |
| type_of_residence | 2.08 | 21974 | 1 | 1.00 | 1.00 | 1.00 | 3.00 |
| kitchen_appliances | 3.81 | 21974 | 1 | 1.00 | 4.00 | 4.00 | 4.00 |
| tv_items | 1.93 | 21940 | 1 | 1.00 | 1.00 | 2.00 | 3.00 |
| household_internet_connection | 1.16 | 21974 | 1 | 1.00 | 1.00 | 1.00 | 1.00 |

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

Sin duda las siguientes variables son categóricas: *male_head_employment.female_head_employment,marital_status,male_he household_internet_connection.*

Sin embargo, las siguientes podrían ser o no ser categóricas, tendríamos que hacer un breve análisis para determinar: *tv_items,kitchen_appliances,age_and_presence_of_children,male_head_education,female_head_education.*

```
variables_seguras<-c("male_head_employment","female_head_employment","marital_status","male_head_occupa

variables_no_seguras<-c("tv_items","kitchen_appliances","age_and_presence_of_children","male_head_educa

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
##  1   :7986   4      :14130      9      :15945
##  2   :7530   1      : 4430      3      : 2107
##  3   :6424   7      : 2698      2      : 1181
```

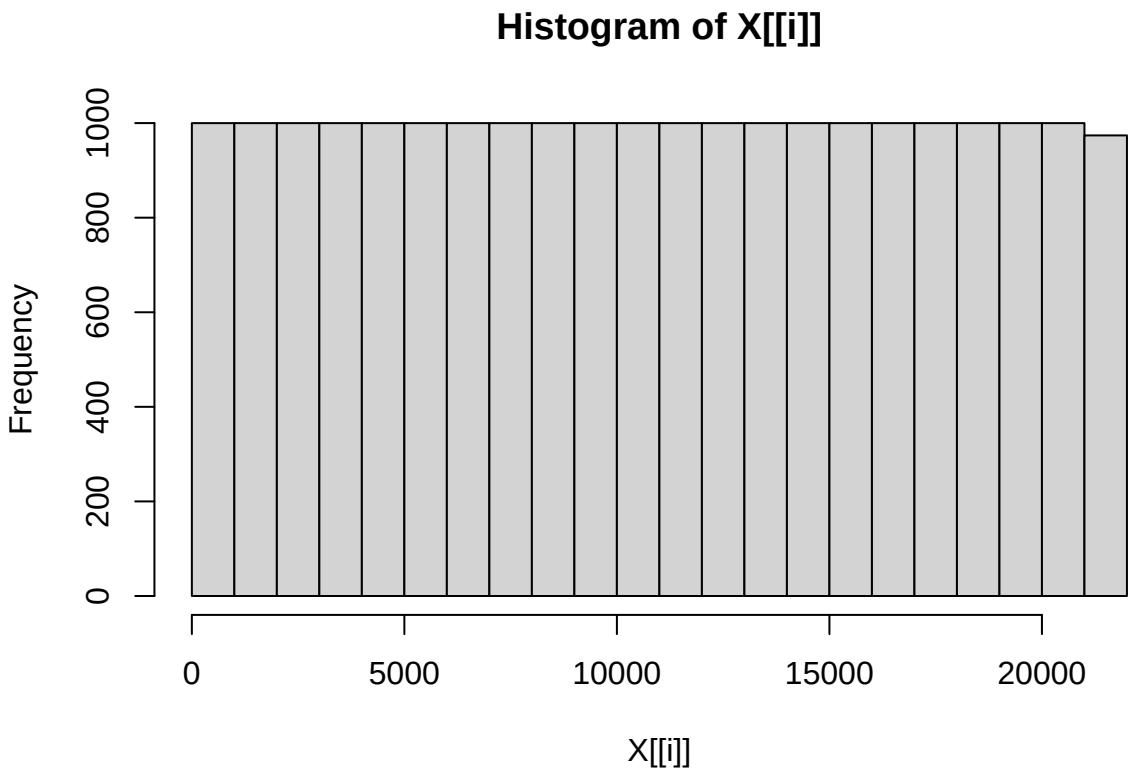
```
##  NA's:  34   5      :  309      1      : 1016
##              8      :  247      6      :  807
##              2      :  132      4      :  588
##             (Other):   28      (Other):  330
##  male_head_education female_head_education
##  0:5317              0:2267
##  1:  59              1:  15
##  2: 425              2: 267
##  3:3213              3:3453
##  4:4922              4:6351
##  5:5475              5:6659
##  6:2563              6:2962
```

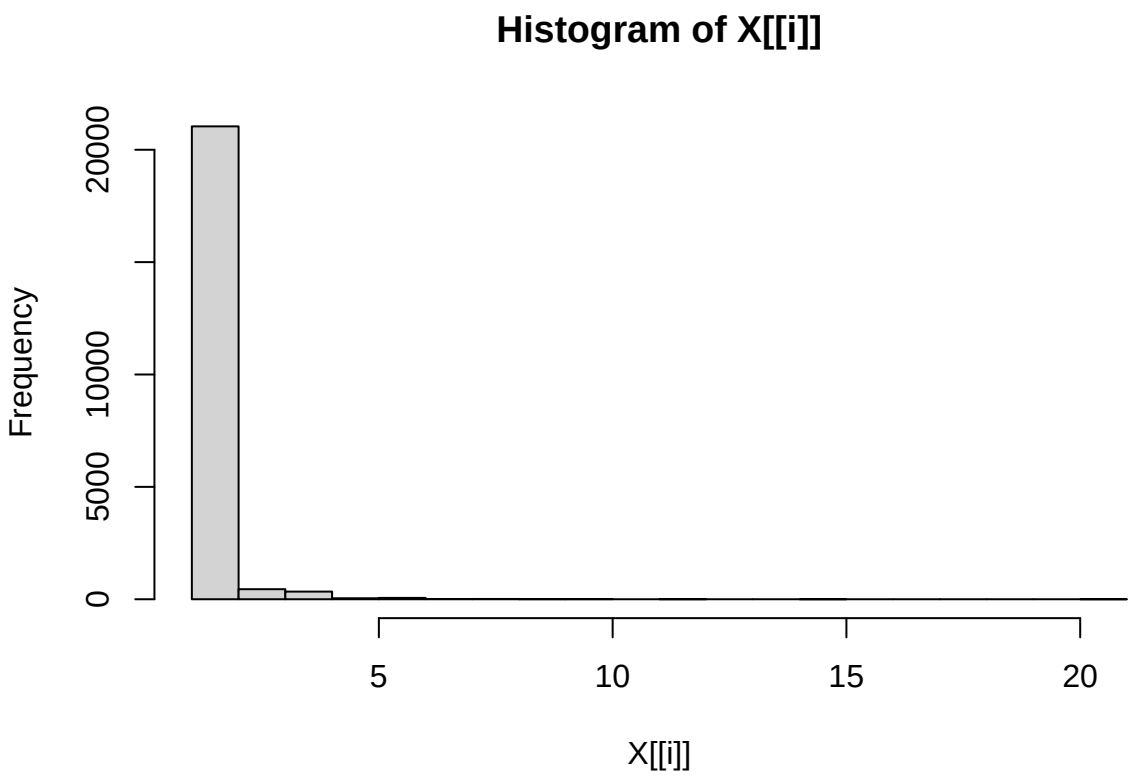
Parece que *tv_items, kitchen_appliances, age_and_presence_of_children* no son categóricas después de todo. Las regresamos a numéricas otra vez, Por el contrario *male_head_education* y *female_head_education* parece que sí son categóricas.

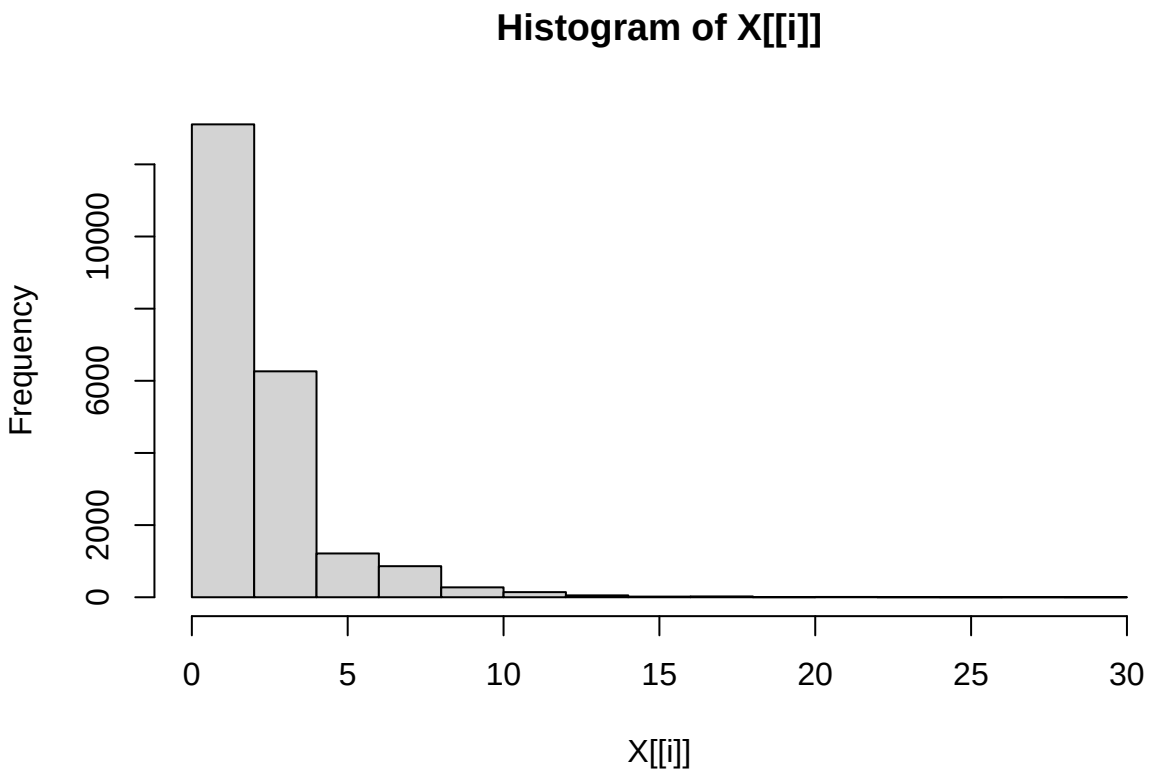
```
variables_numericas<-c("tv_items","kitchen_appliances","age_and_presence_of_children")
base[,variables_numericas] <- lapply(base[,variables_numericas] , as.numeric)
```

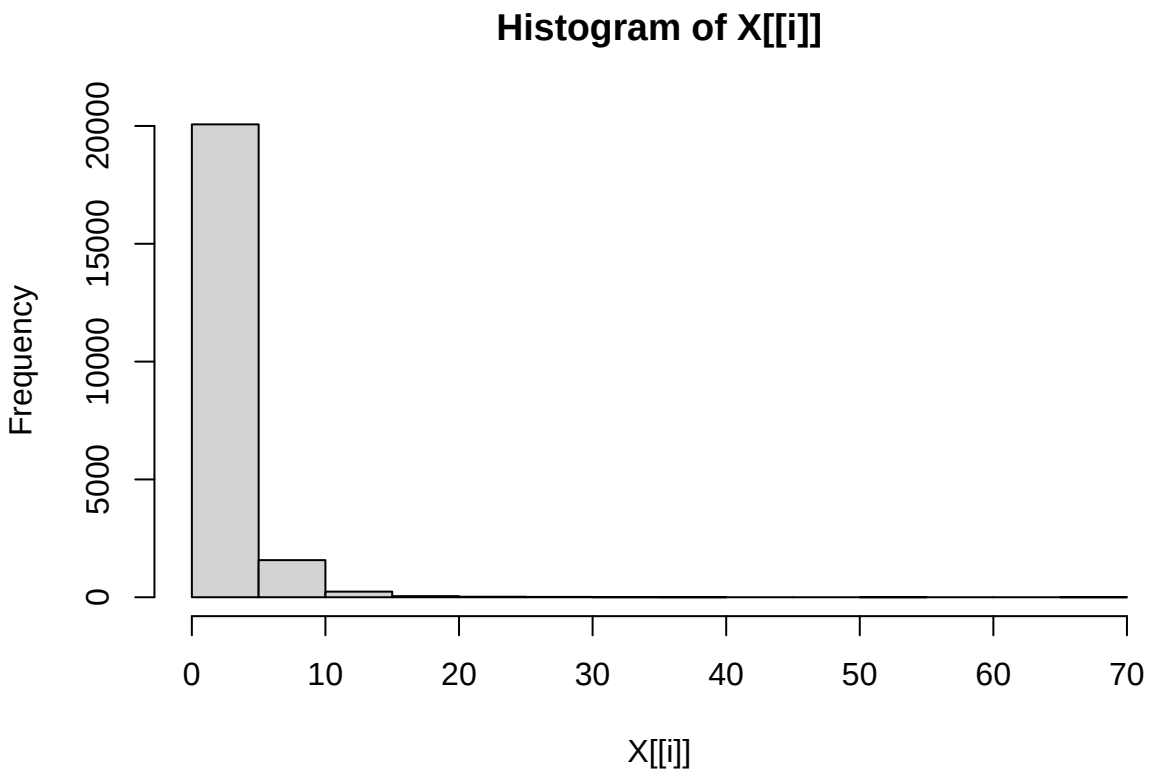
**7. Revisa la distribución de algunas variables. Todas tienen sentido? Por ejemplo, las edades?**

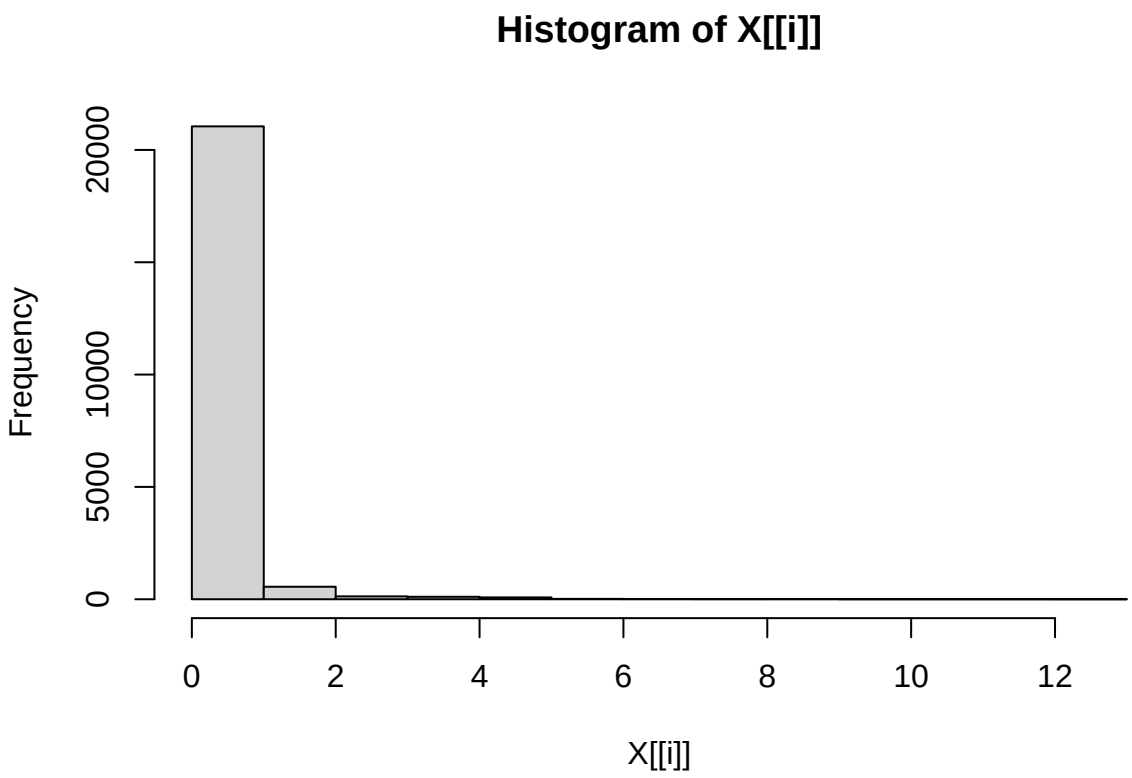
```
numericas<-base%>% select(where(is.numeric))
names(numericas) <- names(numericas)
lapply(numericas,hist)
```

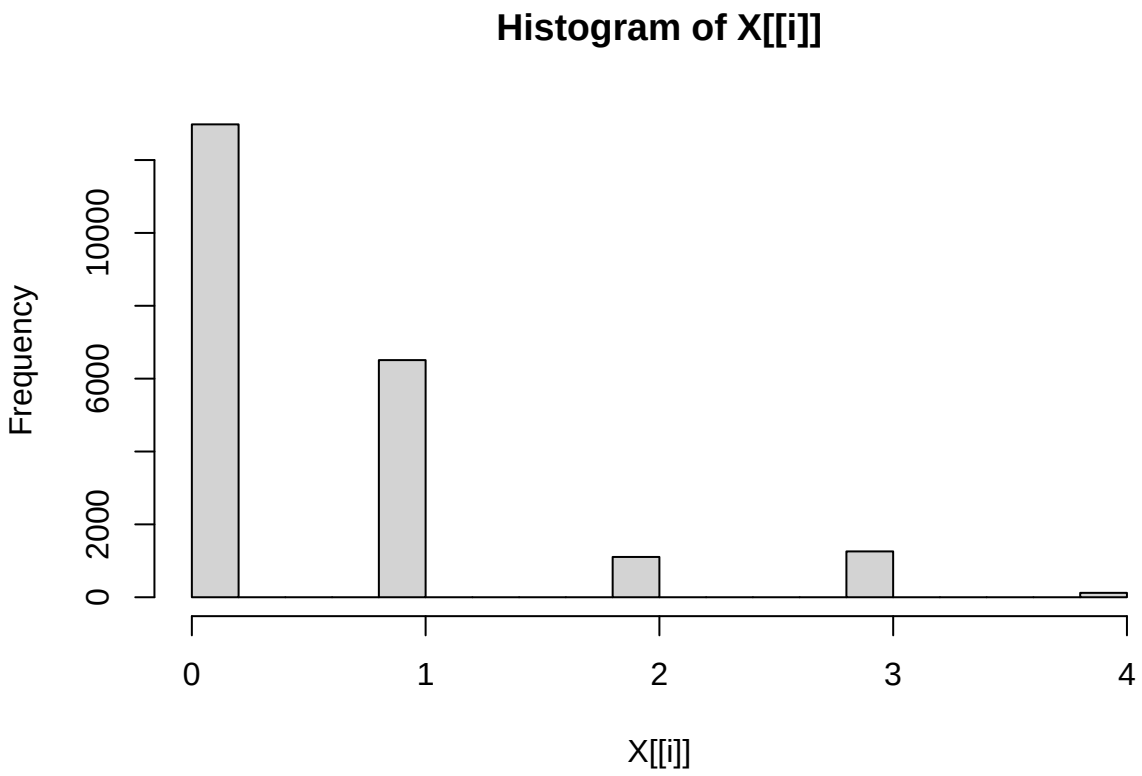


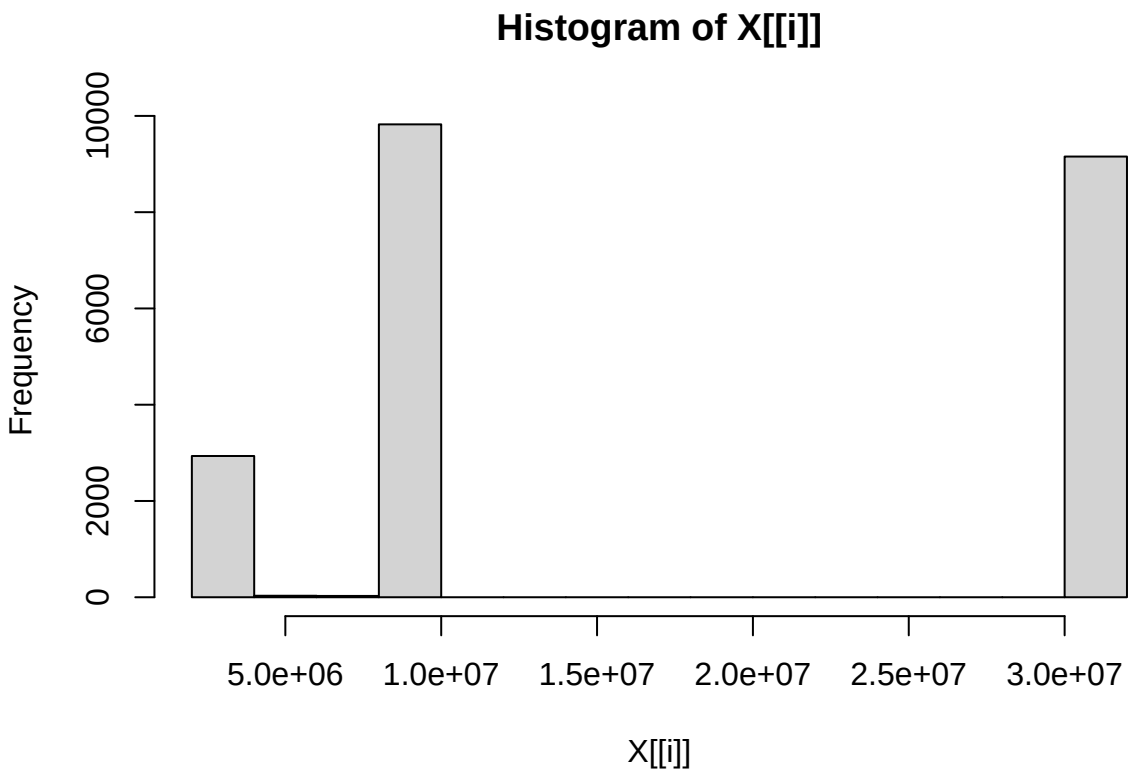

Histogram of X[[i]]
Frequency
0
5000
10000
20000
5
10
15
20
X[[i]]

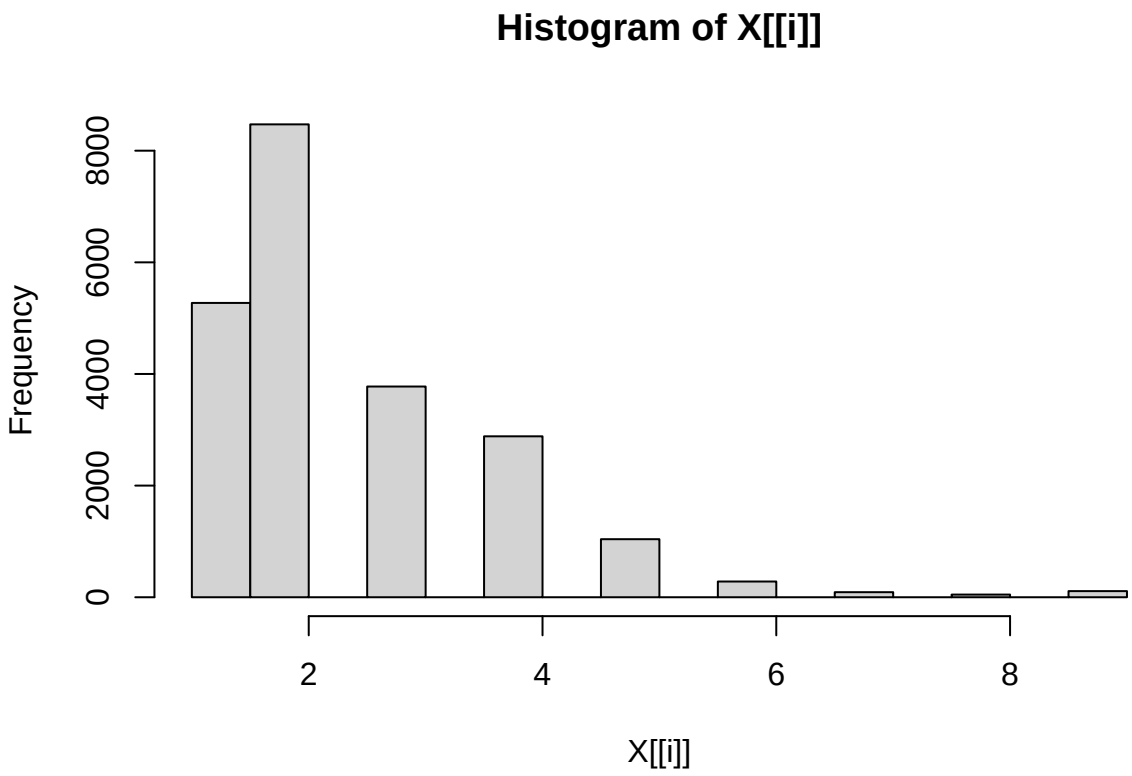

Histogram of X[[i]]
Frequency
0
2000
6000
10000
0
5
10
15
20
25
30
X[[i]]

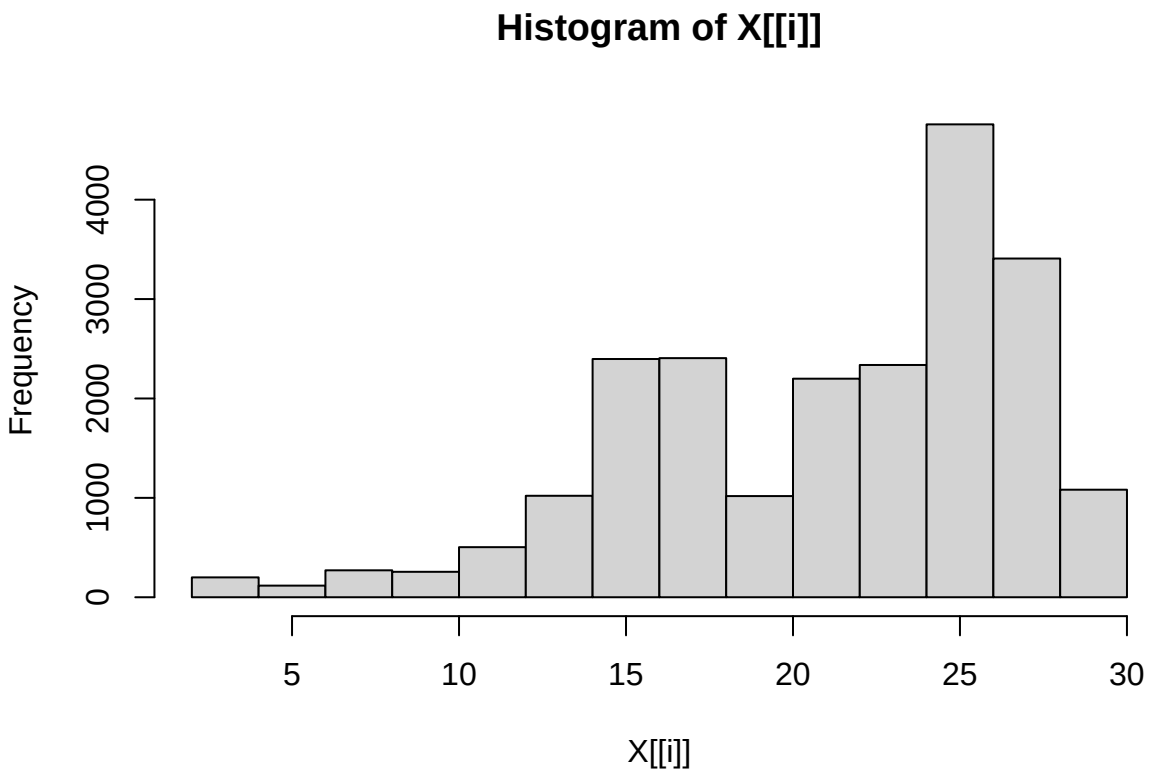

Histogram of X[[i]]
Frequency
0
5000
10000
15000
20000
0
10
20
30
40
50
60
70
X[[i]]

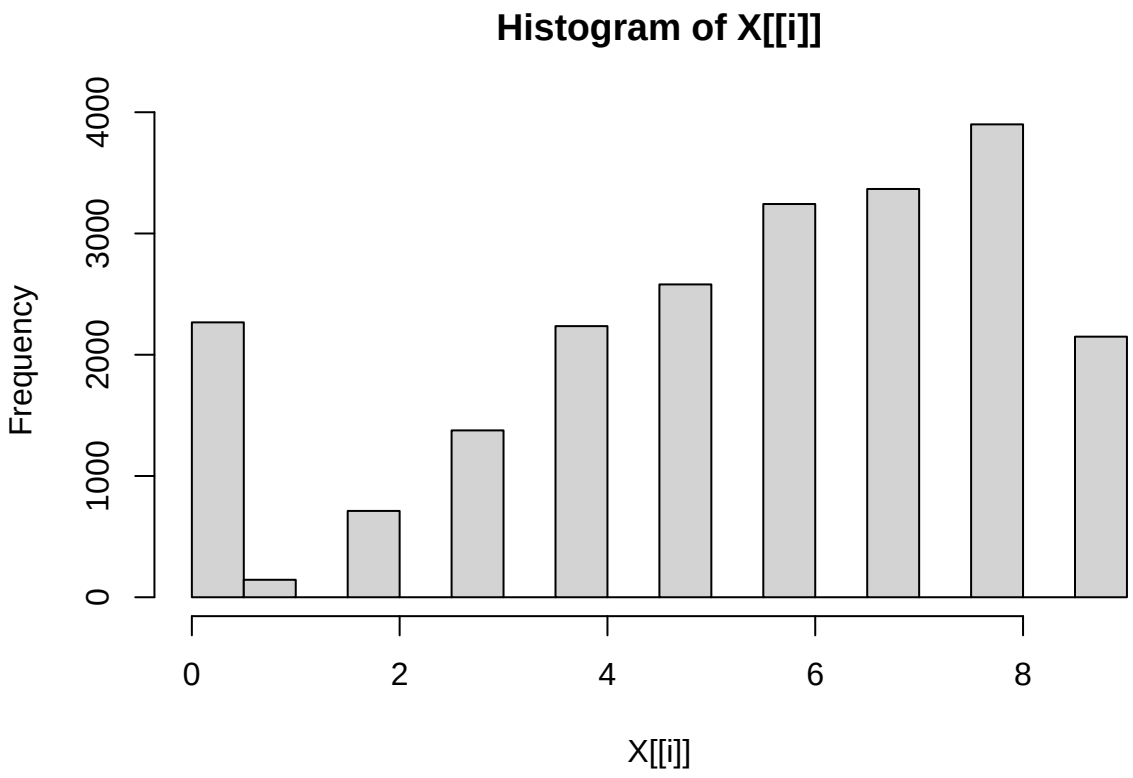

Histogram of X[[i]]
Frequency
0
5000
10000
20000
0
2
4
6
8
10
12
X[[i]]

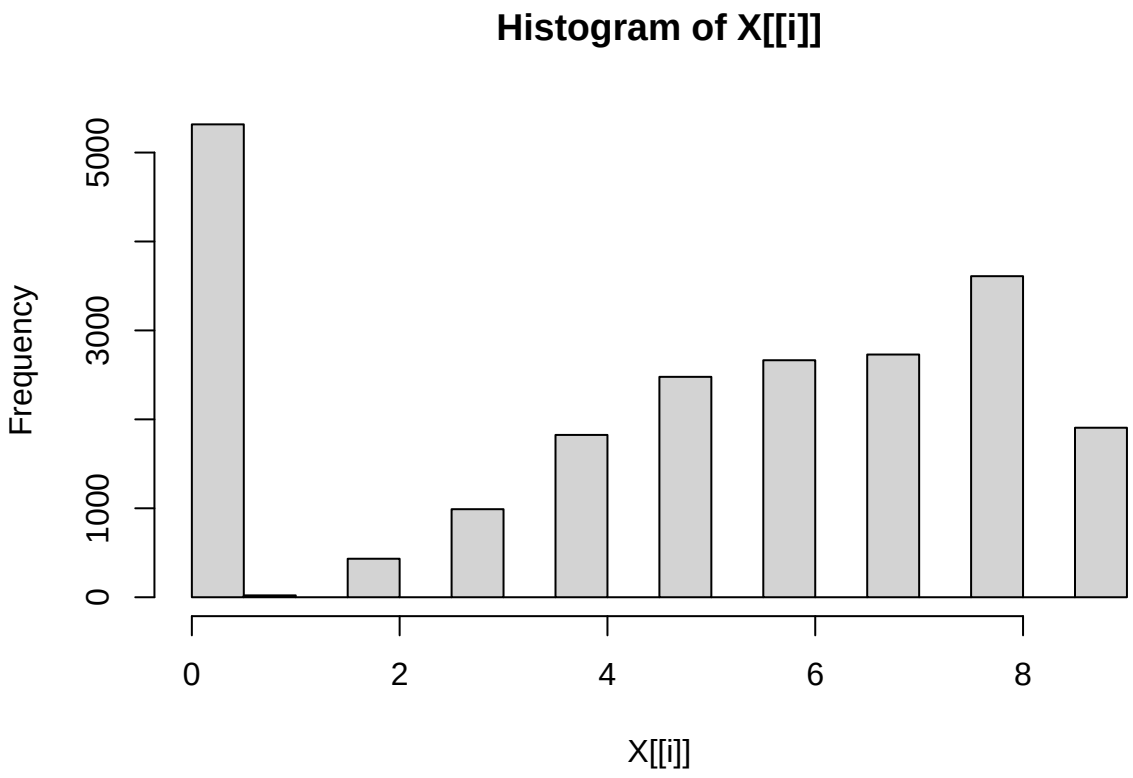

Histogram of X[[i]]
Frequency
0
2000
6000
10000
0
1
2
3
4
X[[i]]

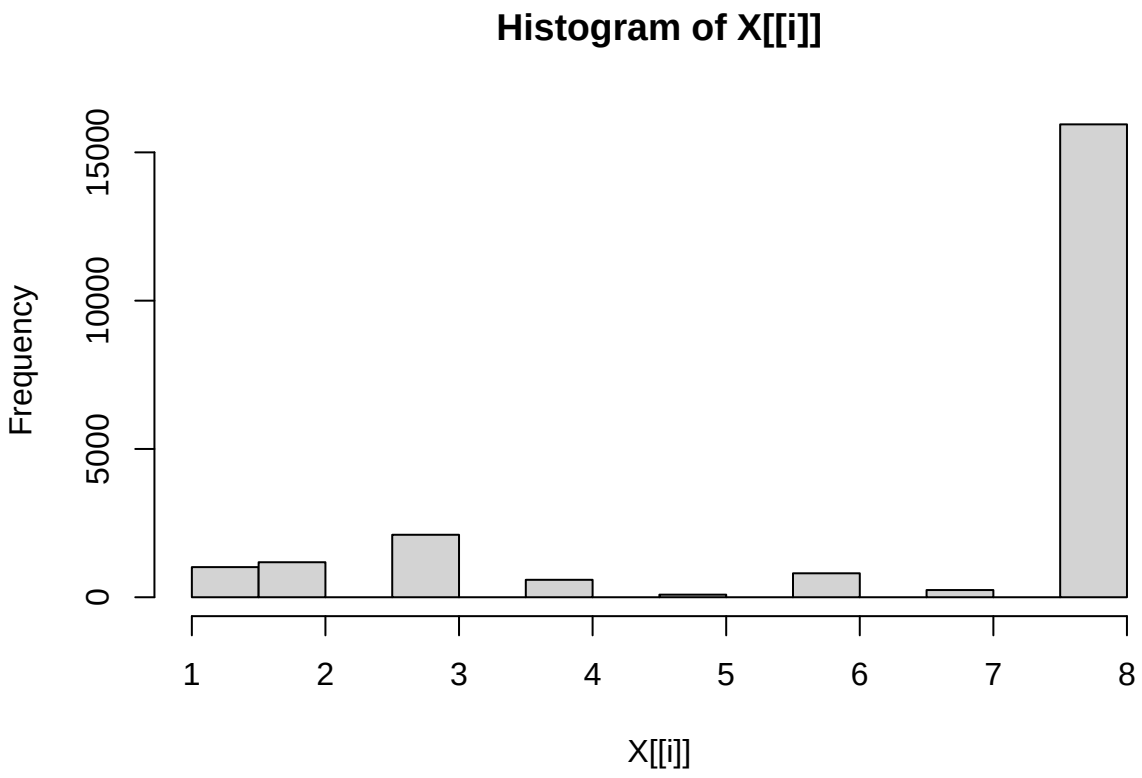

Histogram of X[[i]]
Frequency
0
2000
6000
10000
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
X[[i]]

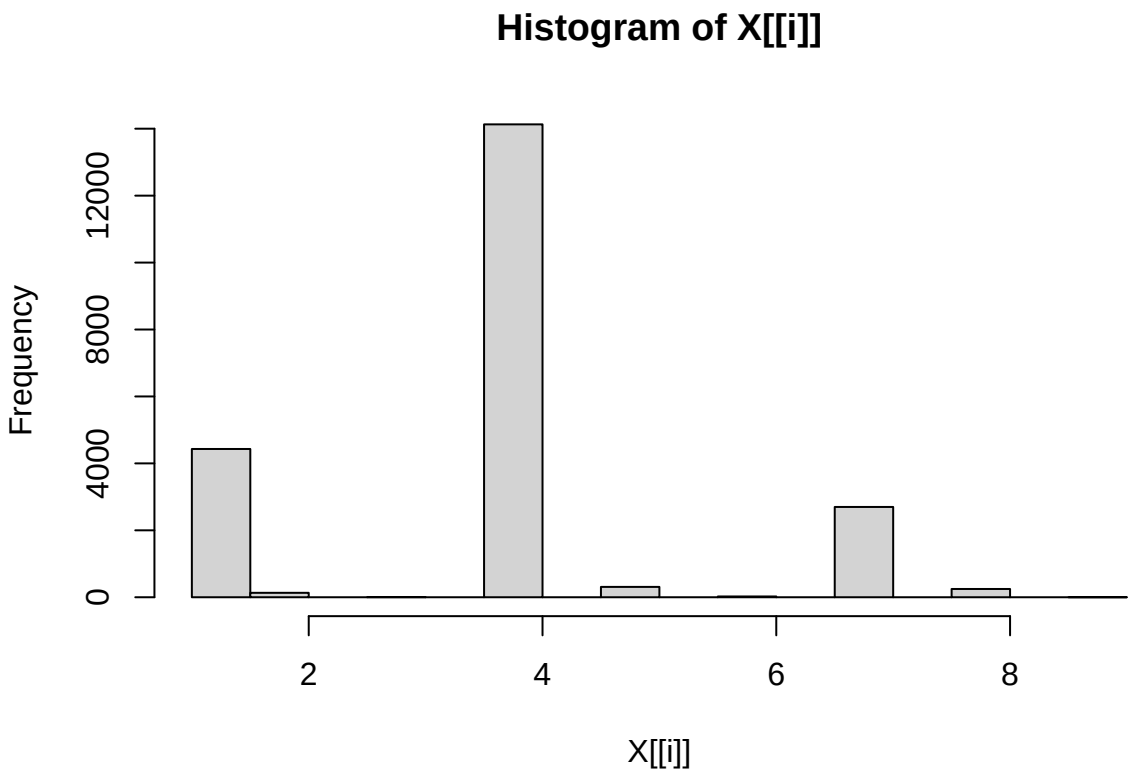

Histogram of X[[i]]
Frequency
0
2000
4000
6000
8000
2
4
6
8
X[[i]]

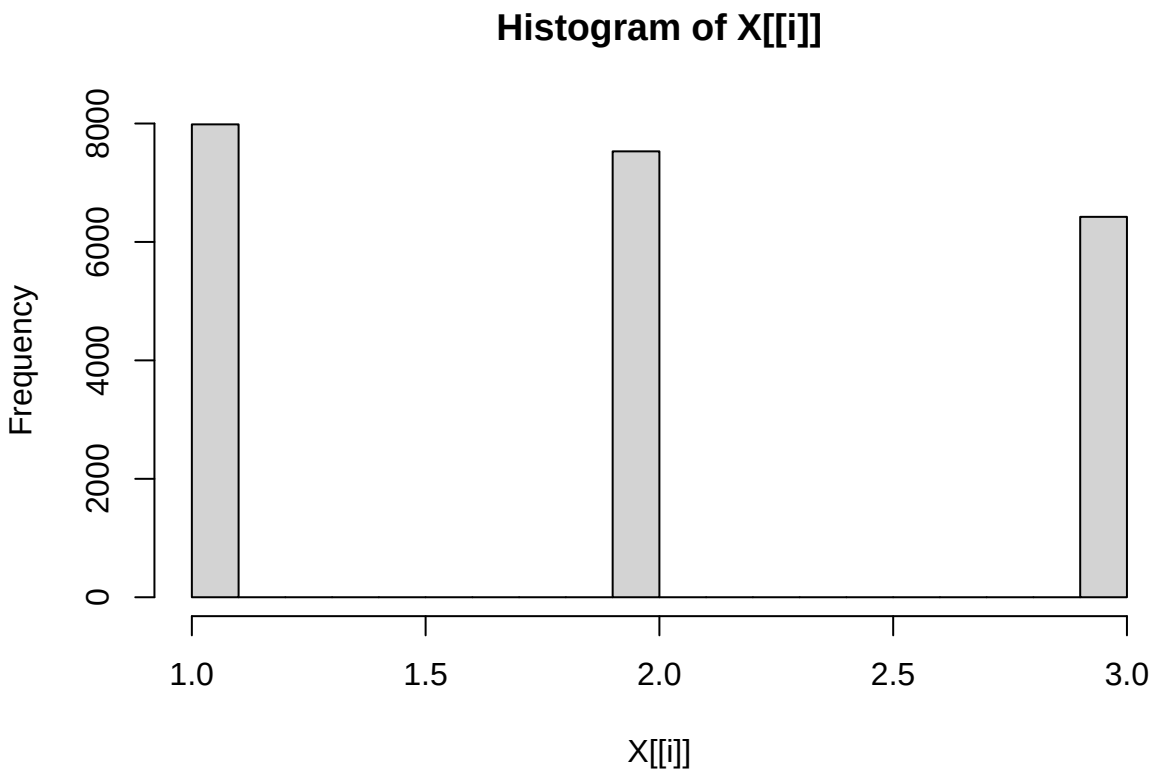

Histogram of X[[i]]
Frequency
0
1000
2000
3000
4000
5
10
15
20
25
30
X[[i]]


Histogram of X[[i]]
Frequency
0
1000
2000
3000
4000
0
2
4
6
8
X[[i]]


Histogram of X[[i]]
Frequency
0
1000
3000
5000
0
2
4
6
8
X[[i]]


Histogram of X[[i]]
Frequency
0
5000
10000
15000
1
2
3
4
5
6
7
8
X[[i]]


Histogram of X[[i]]
Frequency
0
4000
8000
12000
2
4
6
8
X[[i]]




```
## $ID
## $breaks
##  [1]     0  1000  2000  3000  4000  5000  6000  7000  8000  9000 10000 11000
## [13] 12000 13000 14000 15000 16000 17000 18000 19000 20000 21000 22000
##
## $counts
##  [1] 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000
## [16] 1000 1000 1000 1000 1000 1000  974
##
## $density
##  [1] 4.550833e-05 4.550833e-05 4.550833e-05 4.550833e-05 4.550833e-05
##  [6] 4.550833e-05 4.550833e-05 4.550833e-05 4.550833e-05 4.550833e-05
## [11] 4.550833e-05 4.550833e-05 4.550833e-05 4.550833e-05 4.550833e-05
## [16] 4.550833e-05 4.550833e-05 4.550833e-05 4.550833e-05 4.550833e-05
## [21] 4.550833e-05 4.432511e-05
##
## $mids
##  [1]   500  1500  2500  3500  4500  5500  6500  7500  8500  9500 10500 11500
## [13] 12500 13500 14500 15500 16500 17500 18500 19500 20500 21500
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
```

```
## attr(,"class")
## [1] "histogram"
##
## $quantity
## $breaks
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
##
## $counts
##  [1] 21041   447   341    44    59    12    13     6     7     0     2     0
## [13]     0     1     0     0     0     0     0     1
##
## $density
##  [1] 9.575407e-01 2.034222e-02 1.551834e-02 2.002366e-03 2.684991e-03
##  [6] 5.460999e-04 5.916083e-04 2.730500e-04 3.185583e-04 0.000000e+00
## [11] 9.101666e-05 0.000000e+00 0.000000e+00 4.550833e-05 0.000000e+00
## [16] 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 4.550833e-05
##
## $mids
##  [1]  1.5  2.5  3.5  4.5  5.5  6.5  7.5  8.5  9.5 10.5 11.5 12.5 13.5 14.5 15.5
## [16] 16.5 17.5 18.5 19.5 20.5
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $price_paid_deal
## $breaks
##  [1]  0  2  4  6  8 10 12 14 16 18 20 22 24 26 28 30
##
## $counts
##  [1] 13109  6263  1214   861   274   142    52    18    22     4     8     3
## [13]     1     2     1
##
## $density
##  [1] 2.982843e-01 1.425093e-01 2.762356e-02 1.959134e-02 6.234641e-03
##  [6] 3.231091e-03 1.183217e-03 4.095750e-04 5.005916e-04 9.101666e-05
## [11] 1.820333e-04 6.826249e-05 2.275416e-05 4.550833e-05 2.275416e-05
##
## $mids
##  [1]  1  3  5  7  9 11 13 15 17 19 21 23 25 27 29
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
##
## $price_paid_non_deal
## $breaks
##  [1]  0  5 10 15 20 25 30 35 40 45 50 55 60 65 70
##
## $counts
##  [1] 20069  1575   239    48    21    15     4     1     0     0     1     0
## [13]     0     1
##
## $density
##  [1] 1.826613e-01 1.433512e-02 2.175298e-03 4.368799e-04 1.911350e-04
##  [6] 1.365250e-04 3.640666e-05 9.101666e-06 0.000000e+00 0.000000e+00
## [11] 9.101666e-06 0.000000e+00 0.000000e+00 9.101666e-06
##
## $mids
##  [1]  2.5  7.5 12.5 17.5 22.5 27.5 32.5 37.5 42.5 47.5 52.5 57.5 62.5 67.5
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $coupon_value
## $breaks
##  [1]  0  1  2  3  4  5  6  7  8  9 10 11 12 13
##
## $counts
##  [1] 21050   557   128   109    82    17     9     7     8     2     2     2
## [13]     1
##
## $density
##  [1] 9.579503e-01 2.534814e-02 5.825066e-03 4.960408e-03 3.731683e-03
##  [6] 7.736416e-04 4.095750e-04 3.185583e-04 3.640666e-04 9.101666e-05
## [11] 9.101666e-05 9.101666e-05 4.550833e-05
##
## $mids
##  [1]  0.5  1.5  2.5  3.5  4.5  5.5  6.5  7.5  8.5  9.5 10.5 11.5 12.5
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $promotion_type
## $breaks
##  [1] 0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6
```

```
## [20] 3.8 4.0
##
## $counts
##  [1] 12980     0     0     0  6509     0     0     0     0  1106     0     0
## [13]     0     0  1258     0     0     0     0   121
##
## $density
##  [1] 2.95349049 0.00000000 0.00000000 0.00000000 1.48106854 0.00000000
##  [7] 0.00000000 0.00000000 0.00000000 0.25166105 0.00000000 0.00000000
## [13] 0.00000000 0.00000000 0.28624738 0.00000000 0.00000000 0.00000000
## [19] 0.00000000 0.02753254
##
## $mids
##  [1] 0.1 0.3 0.5 0.7 0.9 1.1 1.3 1.5 1.7 1.9 2.1 2.3 2.5 2.7 2.9 3.1 3.3 3.5 3.7
## [20] 3.9
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $household_id
## $breaks
##  [1]  2000000  4000000  6000000  8000000 10000000 12000000 14000000 16000000
##  [9] 18000000 20000000 22000000 24000000 26000000 28000000 30000000 32000000
##
## $counts
##  [1] 2935   31   27 9825    0    0    0    0    0    0    0    0    0    0 9156
##
## $density
##  [1] 6.678347e-08 7.053791e-10 6.143624e-10 2.235597e-07 0.000000e+00
##  [6] 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## [11] 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 2.083371e-07
##
## $mids
##  [1] 3.0e+06 5.0e+06 7.0e+06 9.0e+06 1.1e+07 1.3e+07 1.5e+07 1.7e+07 1.9e+07
## [10] 2.1e+07 2.3e+07 2.5e+07 2.7e+07 2.9e+07 3.1e+07
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $household_size
## $breaks
##  [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0
```

```
##
## $counts
##  [1] 5273 8472    0 3775    0 2883    0 1040    0  282    0   91    0   48    0
## [16]  110
##
## $density
##  [1] 0.479930827 0.771093110 0.000000000 0.343587877 0.000000000 0.262401019
##  [7] 0.000000000 0.094657322 0.000000000 0.025666697 0.000000000 0.008282516
## [13] 0.000000000 0.004368799 0.000000000 0.010011832
##
## $mids
##  [1] 1.25 1.75 2.25 2.75 3.25 3.75 4.25 4.75 5.25 5.75 6.25 6.75 7.25 7.75 8.25
## [16] 8.75
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $household_income
## $breaks
##  [1]  2  4  6  8 10 12 14 16 18 20 22 24 26 28 30
##
## $counts
##  [1]  200  117  271  256  504 1021 2397 2406 1018 2199 2337 4758 3408 1082
##
## $density
##  [1] 0.004550833 0.002662237 0.006166378 0.005825066 0.011468099 0.023232001
##  [7] 0.054541731 0.054746519 0.023163739 0.050036407 0.053176481 0.108264312
## [13] 0.077546191 0.024620005
##
## $mids
##  [1]  3  5  7  9 11 13 15 17 19 21 23 25 27 29
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $age_of_female_head
## $breaks
##  [1] 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0
##
## $counts
##  [1] 2267  144    0  712    0 1376    0 2236    0 2580    0 3243    0 3367    0
## [16] 3900    0 2149
```

```
##
## $density
##  [1] 0.20633476 0.01310640 0.00000000 0.06480386 0.00000000 0.12523892
##  [7] 0.00000000 0.20351324 0.00000000 0.23482297 0.00000000 0.29516702
## [13] 0.00000000 0.30645308 0.00000000 0.35496496 0.00000000 0.19559479
##
## $mids
##  [1] 0.25 0.75 1.25 1.75 2.25 2.75 3.25 3.75 4.25 4.75 5.25 5.75 6.25 6.75 7.25
## [16] 7.75 8.25 8.75
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $age_of_male_head
## $breaks
##  [1] 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0
##
## $counts
##  [1] 5317   21    0  433    0  990    0 1825    0 2478    0 2665    0 2729    0
## [16] 3610    0 1906
##
## $density
##  [1] 0.48393556 0.00191135 0.00000000 0.03941021 0.00000000 0.09010649
##  [7] 0.00000000 0.16610540 0.00000000 0.22553927 0.00000000 0.24255939
## [13] 0.00000000 0.24838445 0.00000000 0.32857013 0.00000000 0.17347775
##
## $mids
##  [1] 0.25 0.75 1.25 1.75 2.25 2.75 3.25 3.75 4.25 4.75 5.25 5.75 6.25 6.75 7.25
## [16] 7.75 8.25 8.75
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $age_and_presence_of_children
## $breaks
##  [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
##
## $counts
##  [1]  1016  1181     0  2107     0   588     0    87     0   807     0   243
## [13]     0 15945
##
## $density
```

```
##  [1] 0.092472923 0.107490671 0.000000000 0.191772094 0.000000000 0.053517794
##  [7] 0.000000000 0.007918449 0.000000000 0.073450441 0.000000000 0.022117047
## [13] 0.000000000 1.451260581
##
## $mids
##  [1] 1.25 1.75 2.25 2.75 3.25 3.75 4.25 4.75 5.25 5.75 6.25 6.75 7.25 7.75
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $kitchen_appliances
## $breaks
##  [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0
##
## $counts
##  [1]  4430   132     0     4     0 14130     0   309     0    23     0  2698
## [13]     0   247     0     1
##
## $density
##  [1] 4.032038e-01 1.201420e-02 0.000000e+00 3.640666e-04 0.000000e+00
##  [6] 1.286065e+00 0.000000e+00 2.812415e-02 0.000000e+00 2.093383e-03
## [11] 0.000000e+00 2.455629e-01 0.000000e+00 2.248111e-02 0.000000e+00
## [16] 9.101666e-05
##
## $mids
##  [1] 1.25 1.75 2.25 2.75 3.25 3.75 4.25 4.75 5.25 5.75 6.25 6.75 7.25 7.75 8.25
## [16] 8.75
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
##
## $tv_items
## $breaks
##  [1] 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8
## [20] 2.9 3.0
##
## $counts
##  [1] 7986    0    0    0    0    0    0    0    0 7530    0    0    0    0    0
## [16]    0    0    0    0 6424
##
## $density
##  [1] 3.639927 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
```

```
##  [9] 0.000000 3.432088 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
## [17] 0.000000 0.000000 0.000000 2.927985
##
## $mids
##  [1] 1.05 1.15 1.25 1.35 1.45 1.55 1.65 1.75 1.85 1.95 2.05 2.15 2.25 2.35 2.45
## [16] 2.55 2.65 2.75 2.85 2.95
##
## $xname
## [1] "X[[i]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

FALTA ANALISIS

**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

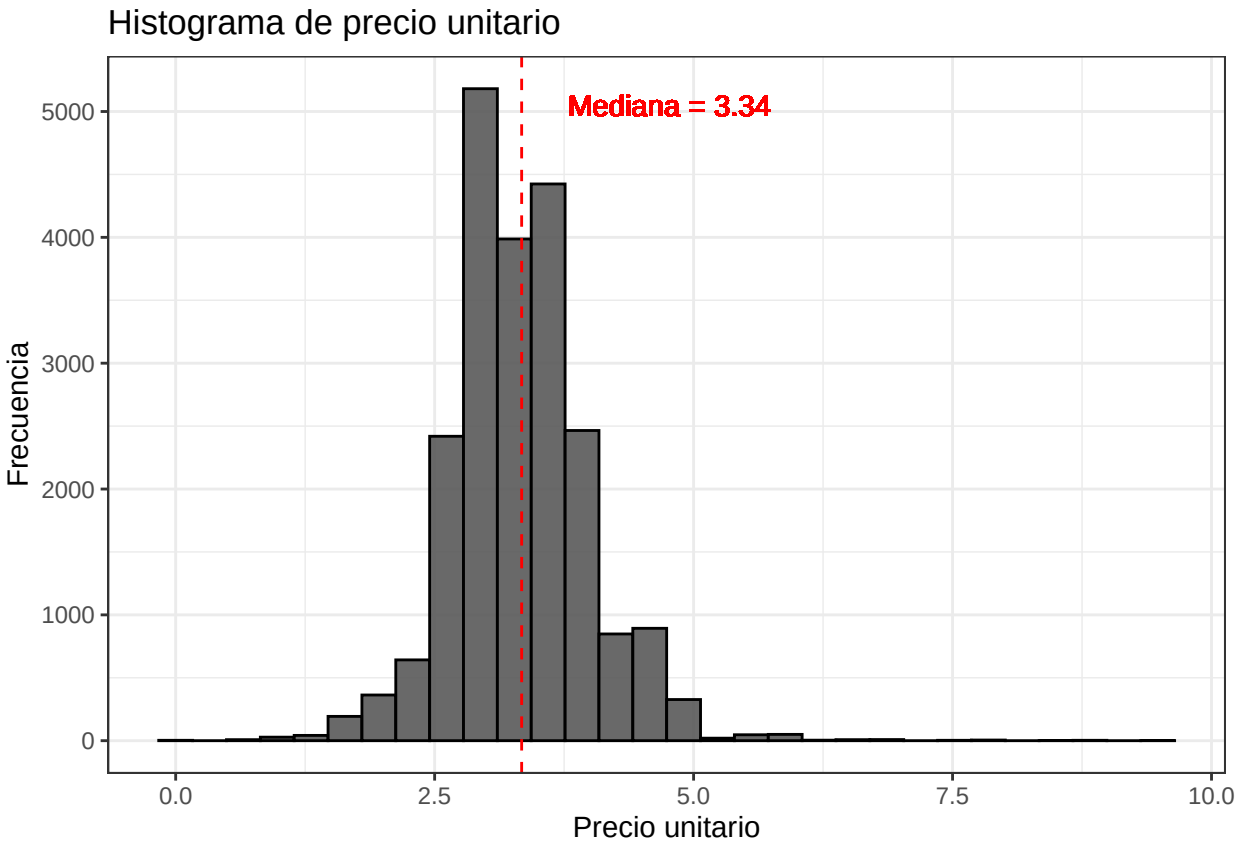
Importante: Las gráficas deben estar bien documentadas (título, ejes con etiquetas apropiadas, etc). Cualquier gráfica que no cumpla con estos requisitos les quitaré algunos puntos.

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

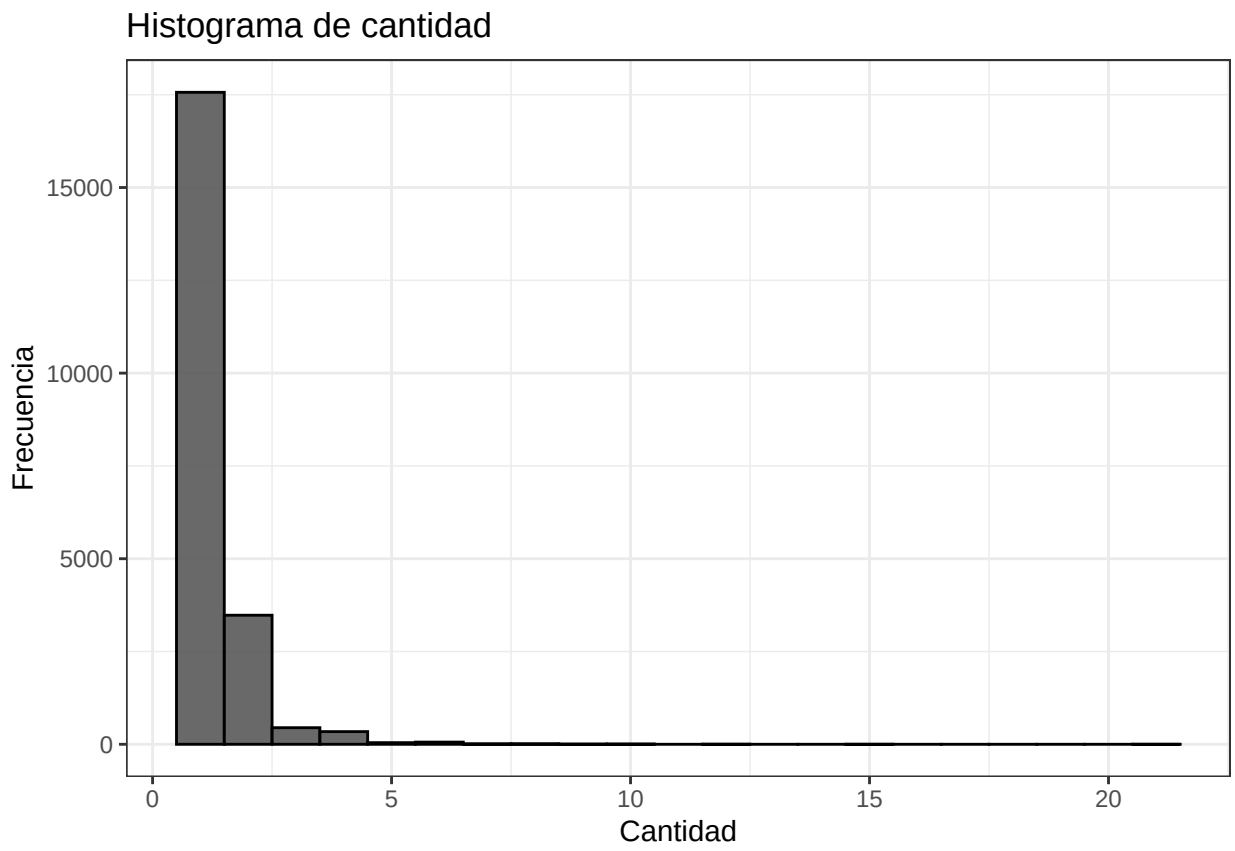
```
median_price <- quantile(base$unit_price)[3]

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")
```



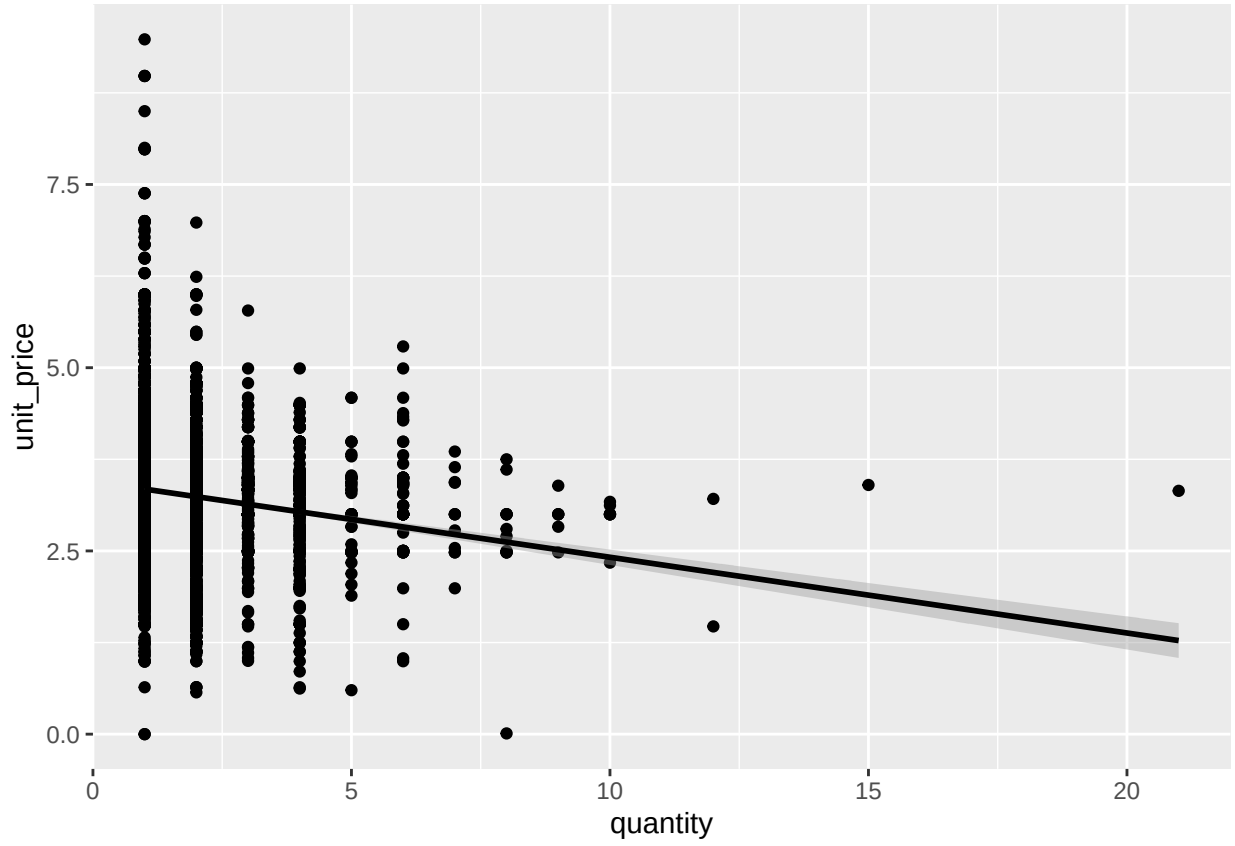

```
ggplot(base)+
  geom_histogram(aes(x=quantity),binwidth=1,alpha=0.9,col = 'black')+
  theme_bw()+
  labs(title="Histograma de cantidad",x="Cantidad",y="Frecuencia")
```




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

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

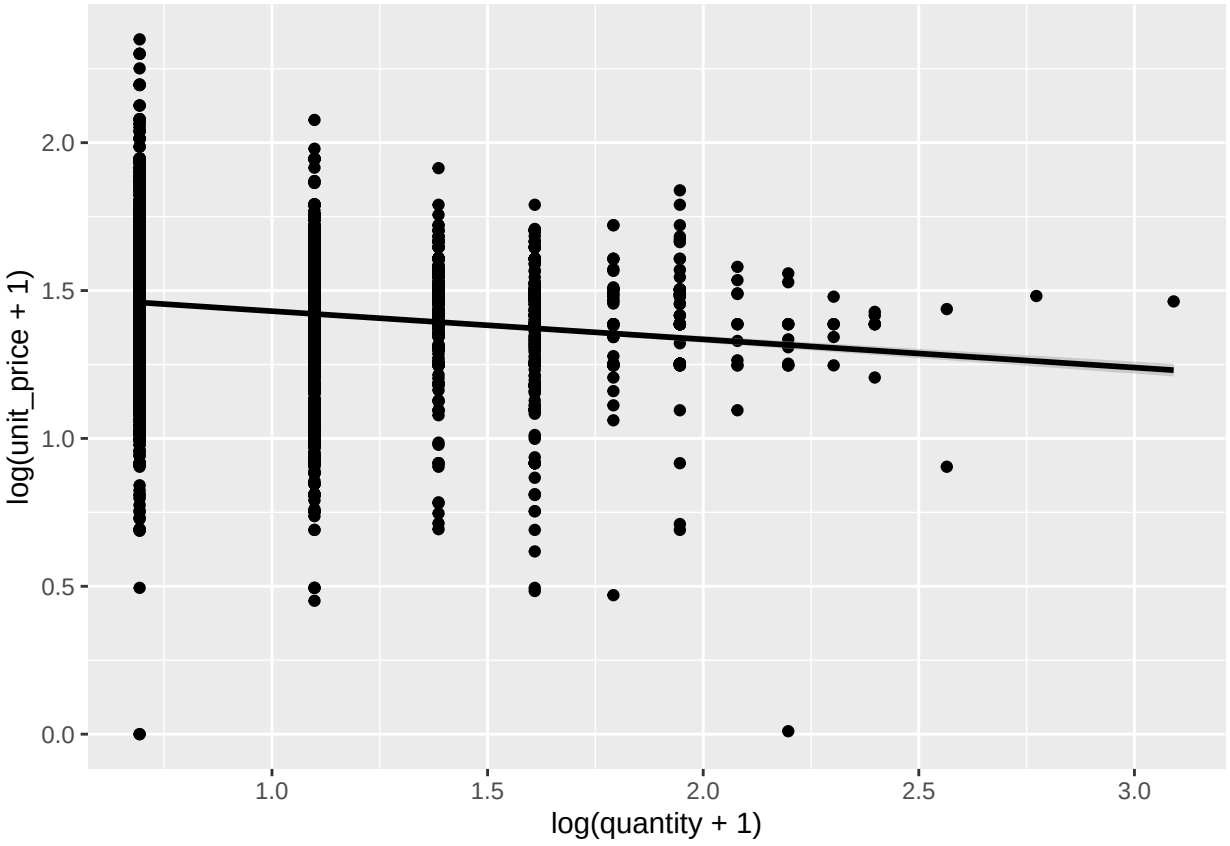
```
ggplot(base)+
  geom_point(aes(x=quantity,y=unit_price))+
  geom_smooth(formula=y~x,method=lm, color='1',aes(x = quantity, y = unit_price))
```




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

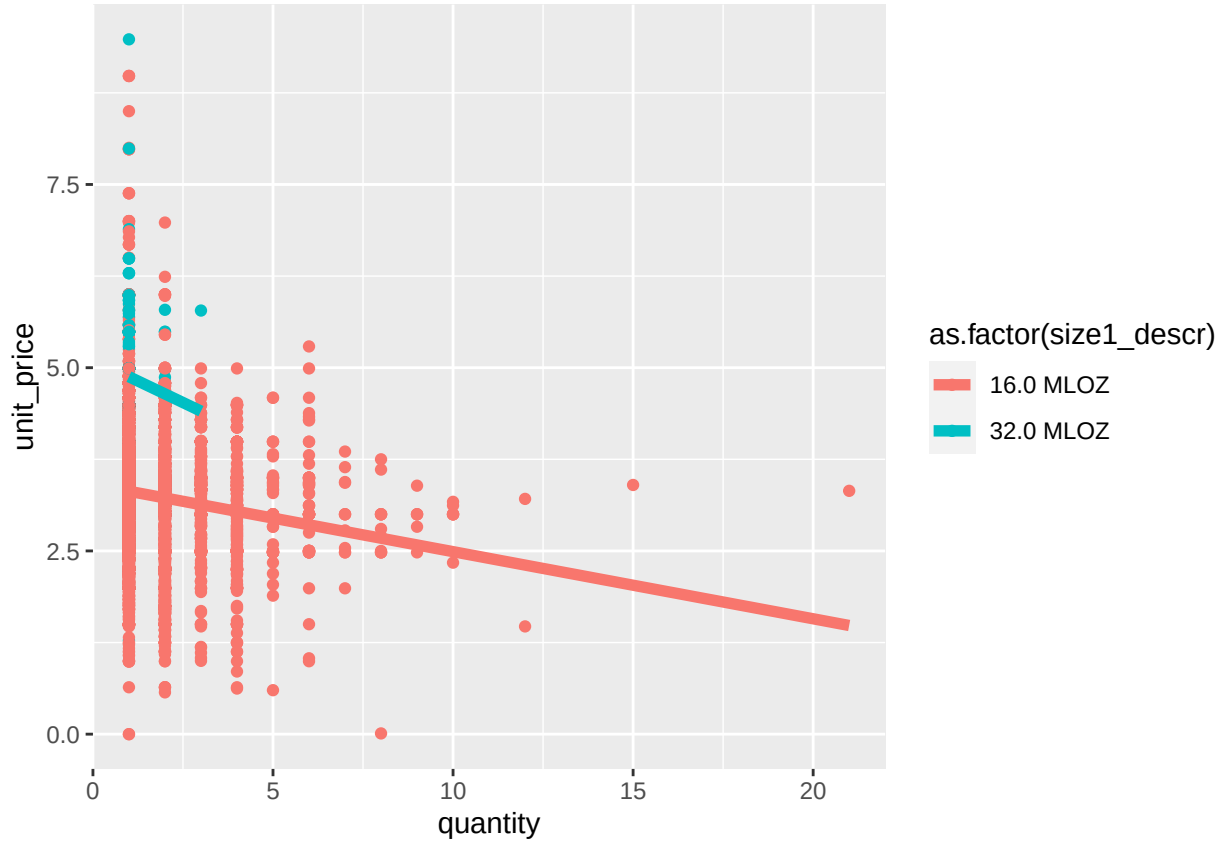
Cuando hacemos la transformación, la relación es más evidente:

```
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))
```

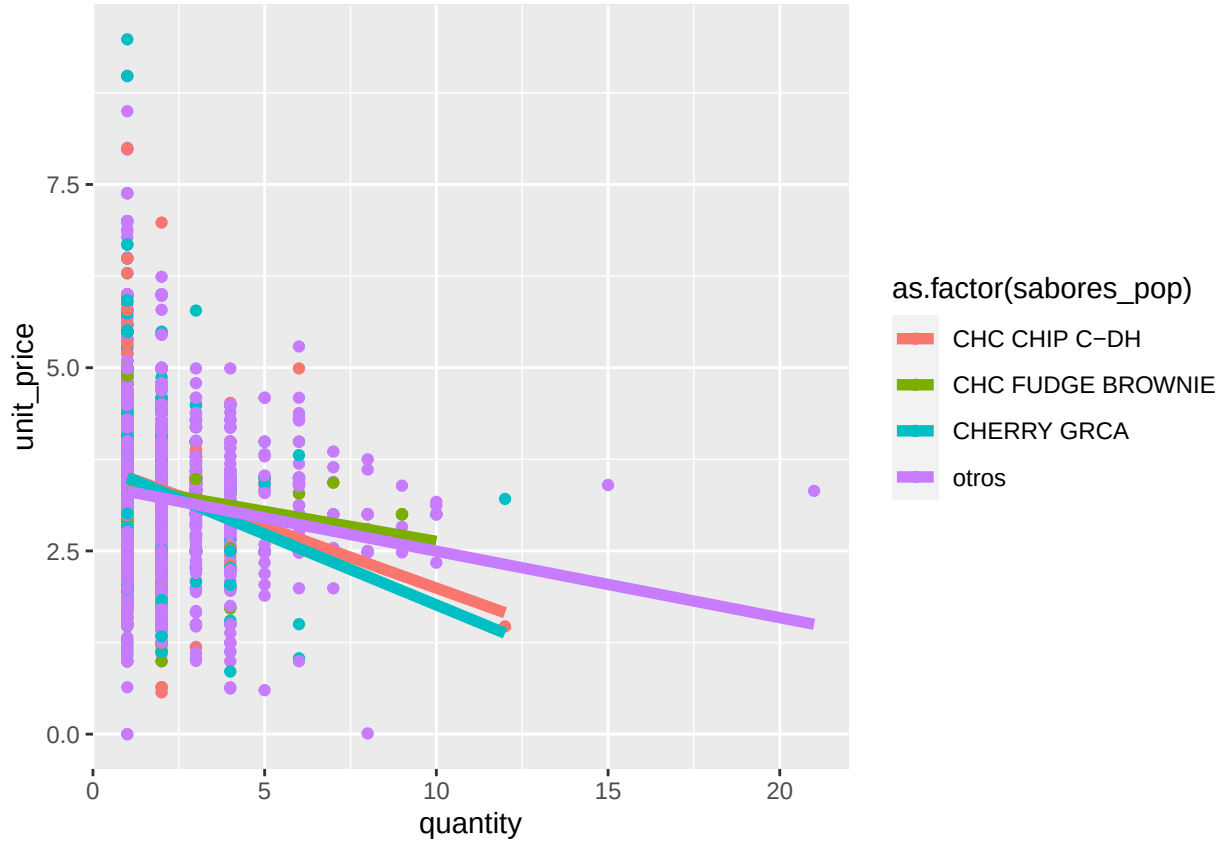
```
##                AMERICONE DREAM                   BANANA SPLIT
##                            865                            599
##                   BLACK & TAN                 BROWNIE BATTER
##                             25                            146
##                   BUTTER PECAN                   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: Sat, Feb 06, 2021 - 03:14:10

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.

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 ica.

Quitemos las fechas

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 |

```
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)**

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

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

**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)`

**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)**