M2.951 - Tipologia i cicle de vida de les dades - PRA2 Neteja i anàlisi de les dades

Autor: Daniel Rubio Baena

Dec 2019

# **Presentació**

**En aquesta pràctica s’elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d’integració, neteja, validació i anàlisi de les mateixes. Per fer aquesta pràctica haureu de treballar en grups de 2 persones. Haureu de lliurar un sol fitxer amb l’enllaç Github (**[**https://github.com**](https://github.com)**) on es trobin les solucions incloent els noms dels components de l’equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius que corresponen a la vostra entrega. Cada membre de l’equip haurà de contribuir amb el seu usuari Github. Malgrat que no es tracta del mateix enunciat, els següents exemples d’edicions anteriors us poden servir com a guia:**

* **Exemple:** [**https://github.com/Bengis/nba-gap-cleaning**](https://github.com/Bengis/nba-gap-cleaning)
* **Exemple complex (fitxer adjunt).**

# **Competències**

**En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:**

* **Capacitat d’analitzar un problema en el nivell d’abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre’l.**
* **Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.**

# **Objectius**

**Els objectius concrets d’aquesta pràctica són:**

* **Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.**
* **Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.**
* **Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.**
* **Identificar la millor representació dels resultats per tal d’aportar conclusions sobre el problema plantejat en el procés analític.**
* **Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l’àmbit d’aplicació.**
* **Desenvolupar les habilitats d’aprenentatge que els permetin continuar estudiant d’una manera que haurà de ser en gran manera autodirigida o autònoma.**
* **Desenvolupar la capacitat de cerca, gestió i ús d’informació i recursos en l’àmbit de la ciència de dades.**

# **Descripció de la pràctica a realitzar**

**L’objectiu d’aquesta activitat serà el tractament d’un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (**[**https://www.kaggle.com**](https://www.kaggle.com)**). Alguns exemples de dataset amb els que podeu treballar són:**

* **Red Wine Quality (**[**https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009**](https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009) **).**
* **Titanic: Machine Learning from Disaster (**[**https://www.kaggle.com/c/titanic**](https://www.kaggle.com/c/titanic) **).**

**L’últim exemple correspon a una competició activa a Kaggle de manera que, opcionalment, podeu aprofitar el treball realitzat durant la pràctica per entrar en aquesta competició.**

**Seguint les principals etapes d’un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:**

1. **Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?**
2. **Integració i selecció de les dades d’interès a analitzar.**
3. **Neteja de les dades.**  
   3.1. **Les dades contenen zeros o elements buits? Com gestionaries aquests casos?**  
   3.2. **Identificació i tractament de valors extrems.**
4. **Anàlisi de les dades.**  
   4.1. **Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).**  
   4.2. **Comprovació de la normalitat i homogeneïtat de la variància.**  
   4.3. **Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l’objectiu de l’estudi, aplicar proves de contrast d’hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d’anàlisi diferents.**
5. **Representació dels resultats a partir de taules i gràfiques.**
6. **Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?**
7. **Codi: Cal adjuntar el codi, preferiblement en R, amb el que s’ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.**

# **Solució**

## **Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?**

El joc de dades triat es pot trobar al següent enllaç (<https://www.kaggle.com/neuromusic/avocado-prices>). Aquest representa l’evolució del preu mijà als EEUU dels advocats en funció del tipus, regió de venda, volum de venda. Les raons de triar el joc de dades són:

1. El tema que representen les dades em resulta interessant. Moltes vegades no veiem tot el que comporta que un bé concret arribi als prestatges dels nostres supermercats (feines, intermediaris, logistica, km, recursos, etc.). Com a ferm defensor del comerç de proximitat m’agrada aprofundir sobre el nostre model de consum actual.
2. El joc de dades té un número d’atributs mig (no molt gran, no molt petit) que em permet treballar amb comoditat. El meu objectiu principal era agafar un dataset que sigues un repte assumible i amb el qual estigués còmode treballant.
3. El dataset té un volum de registres prou elevat per poder treballar, practicar i explotar (preparació, neteja i anàlisi).
4. El dataset té atributs amb diferents formats/tipus (numèrics, strings, date)

El domini es descriu mitjançant els atributs següents:

* **Index** - Index del registre/identificador numèric
* **Date** - Data de l’observació
* **AveragePrice** - Preu mitjà d’un alvocat
* **Total Volume** - Número total d’alvocats venuts
* **4046** - Número total d’alvocats venuts amb referència PLU 4046
* **4225** - Número total d’alvocats venuts amb referència PLU 4225
* **4770** - Número total d’alvocats venuts amb referència PLU 4770
* **Total Bags** - Número total de bosses venudes
* **Small Bags** - Número total de bosses petites venudes
* **Large Bags** - Número total de bosses grans venudes
* **XLarge Bags** - Número total de bosses extra grans venudes
* **Type** - Tipus de alvocat (organic o convencional)
* **Year** - Any de l’observació
* **Region** - Regió dels Estatus Units

De totes les referències d’alvocats existents, la base de dades només contempla els de tipus Hass:

* **4046 - Hass - small**
* **4225 - Hass - large**
* **4770 - Hass Extra Large**
* 4224 - Green - Large from West of the Mississippi River
* 4222 - Green - Small from West of the Mississippi River
* 4221 - Green - Small from East of the Mississippi River
* 4223 - Green - Large from East of the Mississippi River
* 4771 - Green - Medium from East of the Mississippi River
* 3080 - Pinkerton variety - all sizes

## **Integració i selecció de les dades d’interès a analitzar**

Primerament carreguem les llibreries, creem el dataset a partir de l’arxiu csv i generem un resum de cadascun dels atributs del dataset.

# Carreguem els paquets de R que utilitzarem  
library(dplyr)  
library(ggplot2)  
library(nortest)  
# Carreguem l'arxiu csv i guardem les dades en un data frame de nom 'data'  
data <- read.csv('/home/drb/Desktop/PRA1/Solucio/avocado.csv', stringsAsFactors = FALSE)  
# Resum del data frame  
summary(data)

## X Date AveragePrice Total.Volume   
## Min. : 0.00 Length:18249 Min. :0.440 Min. : 85   
## 1st Qu.:10.00 Class :character 1st Qu.:1.100 1st Qu.: 10839   
## Median :24.00 Mode :character Median :1.370 Median : 107377   
## Mean :24.23 Mean :1.406 Mean : 850644   
## 3rd Qu.:38.00 3rd Qu.:1.660 3rd Qu.: 432962   
## Max. :52.00 Max. :3.250 Max. :62505647   
## X4046 X4225 X4770 Total.Bags   
## Min. : 0 Min. : 0 Min. : 0 Min. : 0   
## 1st Qu.: 854 1st Qu.: 3009 1st Qu.: 0 1st Qu.: 5089   
## Median : 8645 Median : 29061 Median : 185 Median : 39744   
## Mean : 293008 Mean : 295155 Mean : 22840 Mean : 239639   
## 3rd Qu.: 111020 3rd Qu.: 150207 3rd Qu.: 6243 3rd Qu.: 110783   
## Max. :22743616 Max. :20470573 Max. :2546439 Max. :19373134   
## Small.Bags Large.Bags XLarge.Bags type   
## Min. : 0 Min. : 0 Min. : 0.0 Length:18249   
## 1st Qu.: 2849 1st Qu.: 127 1st Qu.: 0.0 Class :character   
## Median : 26363 Median : 2648 Median : 0.0 Mode :character   
## Mean : 182195 Mean : 54338 Mean : 3106.4   
## 3rd Qu.: 83338 3rd Qu.: 22029 3rd Qu.: 132.5   
## Max. :13384587 Max. :5719097 Max. :551693.7   
## year region   
## Min. :2015 Length:18249   
## 1st Qu.:2015 Class :character   
## Median :2016 Mode :character   
## Mean :2016   
## 3rd Qu.:2017   
## Max. :2018

Selecció - Eliminem l’atribut **X/Index** que conté els identificador númerics dels registres/observacions. Aquest atribut no aporta valor a l’estudi del dataset.

# Eliminem la primera columna (identificadors numèrics) del data frame   
data = select(data, -1)

Selecció/Transformació - Creem dos atributs nous a partir de l’atribut ‘Date’. Per estudiar la tendencia del preu mig de l’alvocat és molt més interessant l’any i/o mes de l’observació que el dia de l’observació.

# Nou atribut amb el mes de l'observació  
data$DateMonth <- format(as.POSIXct(data$Date), "%m")  
# Nou atribut amb l'any i mes de l'observació  
data$DateYearMonth <- format(as.POSIXct(data$Date), "%Y-%m")

Selecció - Després de generar els dos nous atribut **DateMonth** i **DateYearMonth**, eliminem l’atribut **Date**

# Eliminem la primera columna (Date) del data frame   
data = select(data, -1)  
# Resum del data frame  
summary(data)

## AveragePrice Total.Volume X4046 X4225   
## Min. :0.440 Min. : 85 Min. : 0 Min. : 0   
## 1st Qu.:1.100 1st Qu.: 10839 1st Qu.: 854 1st Qu.: 3009   
## Median :1.370 Median : 107377 Median : 8645 Median : 29061   
## Mean :1.406 Mean : 850644 Mean : 293008 Mean : 295155   
## 3rd Qu.:1.660 3rd Qu.: 432962 3rd Qu.: 111020 3rd Qu.: 150207   
## Max. :3.250 Max. :62505647 Max. :22743616 Max. :20470573   
## X4770 Total.Bags Small.Bags Large.Bags   
## Min. : 0 Min. : 0 Min. : 0 Min. : 0   
## 1st Qu.: 0 1st Qu.: 5089 1st Qu.: 2849 1st Qu.: 127   
## Median : 185 Median : 39744 Median : 26363 Median : 2648   
## Mean : 22840 Mean : 239639 Mean : 182195 Mean : 54338   
## 3rd Qu.: 6243 3rd Qu.: 110783 3rd Qu.: 83338 3rd Qu.: 22029   
## Max. :2546439 Max. :19373134 Max. :13384587 Max. :5719097   
## XLarge.Bags type year region   
## Min. : 0.0 Length:18249 Min. :2015 Length:18249   
## 1st Qu.: 0.0 Class :character 1st Qu.:2015 Class :character   
## Median : 0.0 Mode :character Median :2016 Mode :character   
## Mean : 3106.4 Mean :2016   
## 3rd Qu.: 132.5 3rd Qu.:2017   
## Max. :551693.7 Max. :2018   
## DateMonth DateYearMonth   
## Length:18249 Length:18249   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##

## **Neteja de les dades**

### **Les dades contenen zeros o elements buits? Com gestionaries aquests casos?**

A continuació, observem que no hi ha valors nulls

colSums(is.na(data))

## AveragePrice Total.Volume X4046 X4225 X4770   
## 0 0 0 0 0   
## Total.Bags Small.Bags Large.Bags XLarge.Bags type   
## 0 0 0 0 0   
## year region DateMonth DateYearMonth   
## 0 0 0 0

Tampoc observem valors buits

colSums(data=="")

## AveragePrice Total.Volume X4046 X4225 X4770   
## 0 0 0 0 0   
## Total.Bags Small.Bags Large.Bags XLarge.Bags type   
## 0 0 0 0 0   
## year region DateMonth DateYearMonth   
## 0 0 0 0

Observem que els registres del dataset han de complir les següents expressions:

* Total.Volume = X4046 + X4225 + X4770 + Total.Bags
* Total.Bags = Small.Bags + Large.Bags + XLarge.Bags

De les dos anteriors expressions també se’n desprèn que:

* Total.Volume = X4046 + X4225 + X4770 + Small.Bags + Large.Bags + XLarge.Bags

A partir de les 3 expressions anteriors, creem 3 atributs **temporals** per verificar que els valors dels registres són correctes. D’aquesta manera tenim confirmació que les observacions que es van fer són correctes.  
Els 3 nous atributs s’obtenen a partir de les següents expressions; el resultat de totes 3 ha de ser 0 per garantir que els registres són correctes:

* checkTotalVolume1 = Total.Volume - X4046 - X4225 - X4770 - Total.Bags
* checkTotalBags1 = Total.Bags - Small.Bags - Large.Bags - XLarge.Bags
* checkTotalVolume2 = Total.Volume - X4046 - X4225 - X4770 - Small.Bags - Large.Bags - XLarge.Bags

Els valors dels atributs obtinguts, els arrodonim (sense decimals)

# Nous atributs Temporals (checkTotalVolume1, checkTotalBags1 i checkTotalVolume2)  
data$checkTotalVolume1 <- data$Total.Volume - data$X4046 - data$X4225 - data$X4770 - data$Total.Bags  
data$checkTotalVolume1 <- round(data$checkTotalVolume1,0)  
data$checkTotalBags1 <- data$Total.Bags - data$Small.Bags - data$Large.Bags - data$XLarge.Bags  
data$checkTotalBags1 <- round(data$checkTotalBags1,0)  
data$checkTotalVolume2 <- data$Total.Volume - data$X4046 - data$X4225 - data$X4770 - data$Small.Bags - data$Large.Bags - data$XLarge.Bags  
data$checkTotalVolume2 <- round(data$checkTotalVolume2,0)  
# Resum del data frame  
summary(data)

## AveragePrice Total.Volume X4046 X4225   
## Min. :0.440 Min. : 85 Min. : 0 Min. : 0   
## 1st Qu.:1.100 1st Qu.: 10839 1st Qu.: 854 1st Qu.: 3009   
## Median :1.370 Median : 107377 Median : 8645 Median : 29061   
## Mean :1.406 Mean : 850644 Mean : 293008 Mean : 295155   
## 3rd Qu.:1.660 3rd Qu.: 432962 3rd Qu.: 111020 3rd Qu.: 150207   
## Max. :3.250 Max. :62505647 Max. :22743616 Max. :20470573   
## X4770 Total.Bags Small.Bags Large.Bags   
## Min. : 0 Min. : 0 Min. : 0 Min. : 0   
## 1st Qu.: 0 1st Qu.: 5089 1st Qu.: 2849 1st Qu.: 127   
## Median : 185 Median : 39744 Median : 26363 Median : 2648   
## Mean : 22840 Mean : 239639 Mean : 182195 Mean : 54338   
## 3rd Qu.: 6243 3rd Qu.: 110783 3rd Qu.: 83338 3rd Qu.: 22029   
## Max. :2546439 Max. :19373134 Max. :13384587 Max. :5719097   
## XLarge.Bags type year region   
## Min. : 0.0 Length:18249 Min. :2015 Length:18249   
## 1st Qu.: 0.0 Class :character 1st Qu.:2015 Class :character   
## Median : 0.0 Mode :character Median :2016 Mode :character   
## Mean : 3106.4 Mean :2016   
## 3rd Qu.: 132.5 3rd Qu.:2017   
## Max. :551693.7 Max. :2018   
## DateMonth DateYearMonth checkTotalVolume1 checkTotalBags1   
## Length:18249 Length:18249 Min. : -1.000 Min. :-1.0000000   
## Class :character Class :character 1st Qu.: 0.000 1st Qu.: 0.0000000   
## Mode :character Mode :character Median : 0.000 Median : 0.0000000   
## Mean : 2.082 Mean : 0.0007124   
## 3rd Qu.: 0.000 3rd Qu.: 0.0000000   
## Max. :3939.000 Max. : 1.0000000   
## checkTotalVolume2   
## Min. : -2.000   
## 1st Qu.: 0.000   
## Median : 0.000   
## Mean : 2.083   
## 3rd Qu.: 0.000   
## Max. :3939.000

En aquest punt verifiquem quins registres/observacions del dataset no cumpleixen les 3 expressions anteriors:

* **256** registres no compleixen l’expressió 1 (**checkTotalVolume1**). Representa un **1.42%** del total dels registres (**256/18249**)
* **75** registres no compleixen l’expressió 2 (**checkTotalBags1**). Representa un **0.41%** del total dels registres (**75/18249**)
* **281** registres no compleixen l’expressió 3 (**checkTotalVolume2**). Representa un **1.54%** del total dels registres (**281/18249**)

S’observa que el nombre de registres amb errors és molt petit.

colSums(data!=0)

## AveragePrice Total.Volume X4046 X4225   
## 18249 18249 18007 18188   
## X4770 Total.Bags Small.Bags Large.Bags   
## 12752 18234 18090 15879   
## XLarge.Bags type year region   
## 6201 18249 18249 18249   
## DateMonth DateYearMonth checkTotalVolume1 checkTotalBags1   
## 18249 18249 256 75   
## checkTotalVolume2   
## 281

A continuació, verifiquem quin és el marge d’error dels registres afectats a les 3 expressions anteriors.  
Per les **expressions 1 (checkTotalBags1) i 3 (checkTotalBags2)** obtenim un marge d’error que va de **-2 fins a 3939**. En canvi per **l’expressió 2 (checkTotalBags1)** tenim un marge d’error que va de **-1 a 1**  
Observem que és un marge d’error molt petit en comparació amb les valors màxims/promitjos de cadascún dels atributs **Total.Volume, X4046, X4225, X4770, Small.Bags, Large.Bags, XLarge.Bags i Total.Bags**.  
Degut a que el nombre de registres amb errors és petit i el marge d’error també ho es, decideix no realitzar cap acció de neteja/correcció/transformació sobre els registres amb errors.

table(round(data$checkTotalVolume1,0))

##   
## -1 0 1 2 3 4 5 6 7 8 9 10 11   
## 47 17993 28 6 7 2 2 2 1 3 3 1 4   
## 12 13 15 16 17 18 19 21 22 23 24 25 26   
## 3 3 3 1 2 1 1 3 1 2 1 1 1   
## 27 28 29 32 33 35 36 37 40 41 42 43 44   
## 3 2 1 1 1 1 1 2 1 1 1 1 2   
## 46 50 51 53 54 56 57 61 62 69 70 72 74   
## 1 1 2 2 1 1 1 2 1 1 1 1 1   
## 75 77 84 92 93 98 104 110 113 117 127 130 133   
## 1 1 1 1 1 1 1 2 1 3 2 1 1   
## 138 139 151 152 157 161 162 165 173 178 182 183 186   
## 1 1 2 2 1 1 1 1 1 1 1 2 2   
## 192 203 205 208 209 211 217 221 226 230 240 242 243   
## 1 4 1 1 1 2 2 1 1 1 1 1 1   
## 245 248 259 278 290 293 303 308 311 319 326 356 454   
## 1 1 2 2 1 1 2 2 1 1 2 2 1   
## 470 480 481 491 496 500 508 528 542 548 549 563 856   
## 1 1 1 1 1 1 1 1 1 1 1 2 1   
## 859 1030 1175 3886 3939   
## 2 2 2 1 1

table(round(data$checkTotalBags1,0))

##   
## -1 0 1   
## 31 18174 44

table(round(data$checkTotalVolume2,0))

##   
## -2 -1 0 1 2 3 4 5 6 7 8 9 10   
## 3 51 17968 44 5 10 2 2 2 1 3 3 1   
## 11 12 13 15 16 17 18 19 21 22 23 24 25   
## 4 3 3 3 1 2 1 1 3 1 2 1 1   
## 26 27 28 29 32 33 35 36 37 40 41 42 43   
## 1 3 2 1 1 1 1 1 2 1 1 1 1   
## 44 46 50 51 53 54 56 57 61 62 69 70 72   
## 2 1 1 2 2 1 1 1 2 1 1 1 1   
## 74 75 77 84 92 93 98 104 110 113 117 127 130   
## 1 1 1 1 1 1 1 1 2 1 3 2 1   
## 133 138 139 151 152 157 161 162 165 173 178 182 183   
## 1 1 1 2 2 1 1 1 1 1 1 1 2   
## 186 192 203 205 208 209 211 217 221 226 230 240 242   
## 2 1 4 1 1 1 2 2 1 1 1 1 1   
## 243 245 248 259 278 290 293 303 308 311 319 326 356   
## 1 1 1 2 2 1 1 2 2 1 1 2 2   
## 454 470 480 481 491 496 500 508 528 542 548 549 563   
## 1 1 1 1 1 1 1 1 1 1 1 1 2   
## 856 859 1030 1175 3886 3939   
## 1 2 2 2 1 1

Eliminem els atributs temporals que haviem generat per verificar les observacions amb errors.

# Eliminem els 3 atributs temporals  
data$checkTotalVolume1 <- NULL  
data$checkTotalBags1 <- NULL  
data$checkTotalVolume2 <- NULL  
# Resum del data frame  
summary(data)

## AveragePrice Total.Volume X4046 X4225   
## Min. :0.440 Min. : 85 Min. : 0 Min. : 0   
## 1st Qu.:1.100 1st Qu.: 10839 1st Qu.: 854 1st Qu.: 3009   
## Median :1.370 Median : 107377 Median : 8645 Median : 29061   
## Mean :1.406 Mean : 850644 Mean : 293008 Mean : 295155   
## 3rd Qu.:1.660 3rd Qu.: 432962 3rd Qu.: 111020 3rd Qu.: 150207   
## Max. :3.250 Max. :62505647 Max. :22743616 Max. :20470573   
## X4770 Total.Bags Small.Bags Large.Bags   
## Min. : 0 Min. : 0 Min. : 0 Min. : 0   
## 1st Qu.: 0 1st Qu.: 5089 1st Qu.: 2849 1st Qu.: 127   
## Median : 185 Median : 39744 Median : 26363 Median : 2648   
## Mean : 22840 Mean : 239639 Mean : 182195 Mean : 54338   
## 3rd Qu.: 6243 3rd Qu.: 110783 3rd Qu.: 83338 3rd Qu.: 22029   
## Max. :2546439 Max. :19373134 Max. :13384587 Max. :5719097   
## XLarge.Bags type year region   
## Min. : 0.0 Length:18249 Min. :2015 Length:18249   
## 1st Qu.: 0.0 Class :character 1st Qu.:2015 Class :character   
## Median : 0.0 Mode :character Median :2016 Mode :character   
## Mean : 3106.4 Mean :2016   
## 3rd Qu.: 132.5 3rd Qu.:2017   
## Max. :551693.7 Max. :2018   
## DateMonth DateYearMonth   
## Length:18249 Length:18249   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##

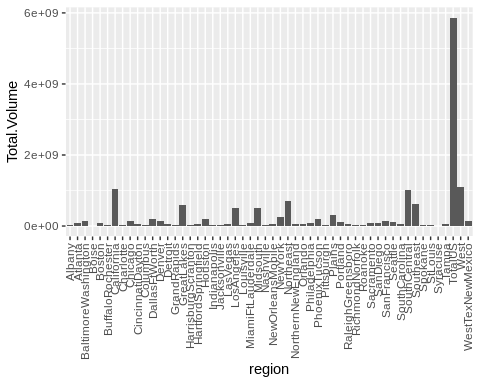
Després de carregar les dades, realitzar el tractament, la validació i la neteja de les dades, procedim a guardar el dataset amb el nom **avocado\_clean.csv**

# Exportación de los datos limpios en .csv  
write.csv(data, "avocado\_clean.csv")

### **Identificació i tractament de valors extrems**

Si representem el volum total venut per regió observem que existeix una **regió** anomenada **TotalUS** que agrega totes les altres regions i que actua com a valor extrem quan volem estudiar el dataset en funció de les regions.

# Volum total venut per regió  
ggplot(data) +   
 geom\_bar(aes(region, Total.Volume), position = "dodge", stat = "summary", fun.y = "sum") +   
 theme(axis.text.x = element\_text(angle = 90, hjust = 1, vjust = 0.5))

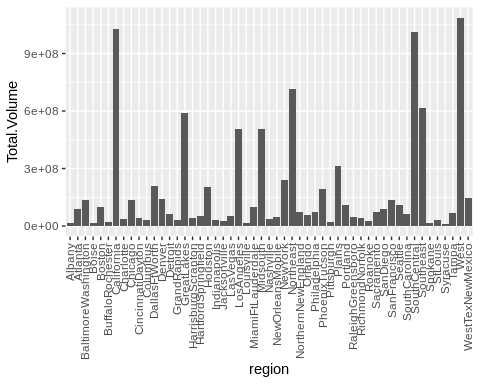


Per tal de solventar el problema del valor extrem de la regió **TotalUS**; es crea un nou sub-dataset (**dataTotalUS**) amb els registres del dataset **data** amb valor de regió igual a **TotalUS**.  
Eliminem els registres del dataset **data** amb valor de regió igual a **TotalUS**.

# Nou dataset amb els registres de les dades totals del United States  
dataTotalUS <- data %>%   
 filter(region == "TotalUS")  
  
# Nou dataset amb tots els registres menys els de les dades totals del United States  
data <- data %>%   
 filter(region != "TotalUS")

Representació del volum total venut per regió després dels canvis anteriors.

# Volum total venut per regió  
ggplot(data) +   
 geom\_bar(aes(region, Total.Volume), position = "dodge", stat = "summary", fun.y = "sum") +   
 theme(axis.text.x = element\_text(angle = 90, hjust = 1, vjust = 0.5))



## **Anàlisi de les dades**

### **Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar)**

A continuació es generen diversos sub-datasets que poden respondre preguntes sobre l’evoluació del consum d’alvocats als Estats Units:

* **dataConvetional** - Seleccionem els registres amb alvocat de tipus **conventional**
* **dataOrganic** - Seleccionem els registres amb alvocat de tipus **organic**
* **data2015** - Seleccionem els registres de l’any **2015**
* **data2016** - Seleccionem els registres de l’any **2016**
* **data2017** - Seleccionem els registres de l’any **2017**
* **data2018** - Seleccionem els registres de l’any **2018**
* **dataBestConsumers** - Seleccionem els registres de les regions amb un major **Total.Volume** (**California, GreatLakes, LosAngeles, Midsouth, Northeast, SouthCentral, Southeast i West**)
* **dataAverageConsumers** - Seleccionem els registres de les regions amb un consum **Total.Volume** normal (**totes les regions menys California, GreatLakes, LosAngeles, Midsouth, Northeast, SouthCentral, Southeast i West**)

dataConvetional <- data[data$type == "conventional",]  
dataOrganic <- data[data$type == "organic",]  
  
data2015 <- data[data$year == 2015,]  
data2016 <- data[data$year == 2016,]  
data2017 <- data[data$year == 2017,]  
data2018 <- data[data$year == 2018,]  
  
dataBestConsumers <- data %>%   
 filter(region == "California" | region == "GreatLakes" | region == "LosAngeles" | region == "Midsouth" | region == "Northeast" | region == "SouthCentral" | region == "Southeast" | region == "West")  
  
dataAverageConsumers <- data %>%   
 filter(region != "California" | region != "GreatLakes" | region != "LosAngeles" | region != "Midsouth" | region != "Northeast" | region != "SouthCentral" | region != "Southeast" | region != "West")

### **Comprovació de la normalitat i homogeneïtat de la variància**

Per estudiar la **normalitat** dels atributs del dataset s’utilitza la prova de normalitat d’**Anderson-Darling**.

Per a cadascun dels atributs es verifica si el valor **p-value** obtingut és major al valor del nivell de significació prefixat (0.05). En cas de ser-ho l’atribut segueix una distribuació normal.

Per al dataset de l’estudi es verifica que **cap atribut quantitatiu segueix una distribució normal**.

alpha = 0.05  
col.names = colnames(data)  
for (i in 1:ncol(data)) {  
 if (i == 1) cat("Variables que no segueixen una distribució normal:\n")  
 if (is.integer(data[,i]) | is.numeric(data[,i])) {  
 p\_val = ad.test(data[,i])$p.value  
 if (p\_val < alpha) {  
 cat(col.names[i])  
 # Format output  
 if (i < ncol(data) - 1) cat(", ")  
 if (i %% 3 == 0) cat("\n")  
 }  
 }  
}

## Variables que no segueixen una distribució normal:  
## AveragePrice, Total.Volume, X4046,   
## X4225, X4770, Total.Bags,   
## Small.Bags, Large.Bags, XLarge.Bags,   
## year,

Per estudiar l’**homoscedasticitat** de les dades no podem utilitzar el test de **Levene** degut a que les dades **no** segueixen una distribució normal. Per al cas que estem estudiant utilitzem el test de **Fligner-Killeen** per permet estudiar l’homogeneïtat de la variància quan les dades no segueixen una distribució normal.

A continuació s’estudia l’homogeneïtat per al grups de tipus **conventional** en front dels que són de tipus **organic**. En aquest cas el valor de **p-value** obtingut és inferior a 0.05, per tant **no** es presenta homogeneïtat.

fligner.test(AveragePrice ~ type, data = data)

##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: AveragePrice by type  
## Fligner-Killeen:med chi-squared = 553.35, df = 1, p-value < 2.2e-16

Seguidament s’estudia l’homogeneïtat per al grups de regions. En aquest cas el valor de **p-value** obtingut és inferior a 0.05, per tant **no** es presenta homogeneïtat.

fligner.test(AveragePrice ~ region, data = data)

##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: AveragePrice by region  
## Fligner-Killeen:med chi-squared = 1784.5, df = 52, p-value < 2.2e-16

Per últim s’estudia l’homogeneïtat per al grups d’anys. En aquest cas el valor de **p-value** obtingut és inferior a 0.05, per tant **no** es presenta homogeneïtat.

fligner.test(AveragePrice ~ year, data = data)

##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: AveragePrice by year  
## Fligner-Killeen:med chi-squared = 154.15, df = 3, p-value < 2.2e-16

### **Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l’objectiu de l’estudi, aplicar proves de contrast d’hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d’anàlisi diferents.**

#### **Wilcoxon (comparació entre dos grups de dades)**

¿Hi ha diferencia de preu mig entre els alvocats de tipus **conventional** i **organic**?

Donat que la normalitat i l’homoscedasticitat no es compleixen apliquem la prova no paramètrica de Wilcoxon. S’observen diferències en el preu mig dels alcovats de tipus conventional i organic.

wilcox.test(AveragePrice ~ type, data = data)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: AveragePrice by type  
## W = 10376686, p-value < 2.2e-16  
## alternative hypothesis: true location shift is not equal to 0

¿Hi ha diferencia de preu mig per les vendes fetes durant **Desembre** i **Gener**?

Aplicant Wilcoxon no s’observen diferències en el preu mig dels alvocats entre els mesos de Desembre i Gener.

dataDecJan <- data[data$DateMonth == "12" | data$DateMonth == "01",]  
wilcox.test(AveragePrice ~ DateMonth, data = dataDecJan)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: AveragePrice by DateMonth  
## W = 1274323, p-value = 0.1425  
## alternative hypothesis: true location shift is not equal to 0

#### **Kruskal-Wallis (comparació entre més de dos grups)**

¿S’ha produït un increment del consum d’alvocats de tipus X4046 a la regió de Louisville durant els anys?

Aplicant el test de Kruskal-Wallis (atributs **X4046** i **year**) obtenim un valor de p-value superior al nivell de significació; es pot concloure que el consum d’alvocats de tipus 4046 no mostra diferències significatives durant els anys.

dataLouisville <- data %>%   
 filter(region == "Louisville")  
  
kruskal.test(X4046 ~ year, data = dataLouisville)

##   
## Kruskal-Wallis rank sum test  
##   
## data: X4046 by year  
## Kruskal-Wallis chi-squared = 4.93, df = 3, p-value = 0.177

¿S’ha produït un increment del consum d’alvocats de tipus X4046 a la regió de Nashville durant els anys?

Aplicant el test de Kruskal-Wallis (atributs **X4046** i **year**) obtenim un valor de p-value superior al nivell de significació; es pot concloure que el consum d’alvocats de tipus 4046 no mostra diferències significatives durant els anys.

dataNashville <- data %>%   
 filter(region == "Nashville")  
  
kruskal.test(X4046 ~ year, data = dataNashville)

##   
## Kruskal-Wallis rank sum test  
##   
## data: X4046 by year  
## Kruskal-Wallis chi-squared = 3.1171, df = 3, p-value = 0.3739

#### **Correlació**

¿Quins atributs influeixen més en el preu dels alvocats de tipus **conventional**?

Per respondre a aquesta pregunta analitzem la correlació. Per fer-ho fem ús del coeficient de correlació d’Spearman (per dades que no compleixen la condició de normalitat).  
S’observa que l’any i el mes de l’any són els atributs que més influeixen en el preu dels alvocats de tipus conventional.

# Convertim l'atribut DataMonth a integer  
dataConvetional$DateMonth <- as.integer(dataConvetional$DateMonth)  
# Calcular el coeficient de correlació per a cada variable numèrica respecte l'atribut AveragePrice  
for (i in 2:(ncol(dataConvetional))) {  
 if (is.integer(dataConvetional[,i]) | is.numeric(dataConvetional[,i])) {  
 print(colnames(dataConvetional)[i])  
 spearman\_test <- cor.test(dataConvetional$AveragePrice, dataConvetional[,i], method = "spearman")  
 print(spearman\_test$estimate)  
 }  
}

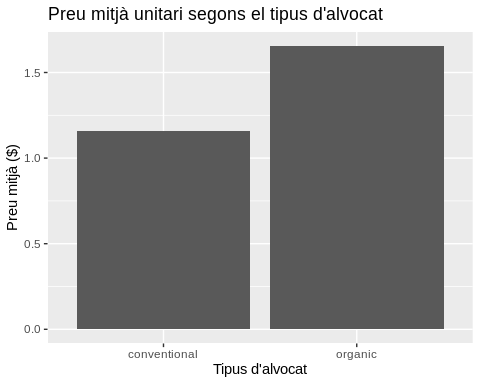
## [1] "Total.Volume"  
## rho   
## -0.2513245   
## [1] "X4046"  
## rho   
## -0.3767694   
## [1] "X4225"  
## rho   
## -0.07229206   
## [1] "X4770"  
## rho   
## -0.2611859   
## [1] "Total.Bags"  
## rho   
## -0.2126517   
## [1] "Small.Bags"  
## rho   
## -0.1446565   
## [1] "Large.Bags"  
## rho   
## -0.230496   
## [1] "XLarge.Bags"  
## rho   
## -0.006986144   
## [1] "year"  
## rho   
## 0.2587072   
## [1] "DateMonth"  
## rho   
## 0.2072692

## **Representació dels resultats a partir de taules i gràfiques**

¿Hi ha diferencia de preu mig entre els alvocats de tipus **conventional** i **organic**?

Tal i com es va poder veure amb la prova de **Wilcoxon**, la següent gràfica reflexa les diferències en el preu mig dels alcovats de tipus conventional i organic.

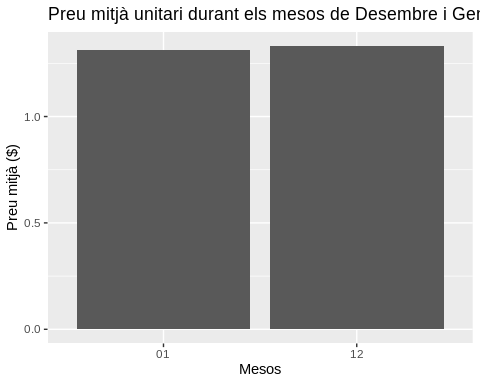
# Preu mitjà dels alvocats organics i convencionals  
ggplot(data, aes(x = type, y = AveragePrice)) +   
 stat\_summary(fun.y = "mean", geom = "bar") +   
 xlab("Tipus d'alvocat") +   
 ylab("Preu mitjà ($)") +   
 ggtitle("Preu mitjà unitari segons el tipus d'alvocat")



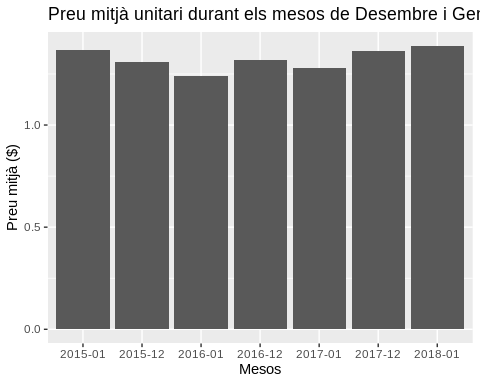
¿Hi ha diferencia de preu mig per les vendes fetes durant **Desembre** i **Gener**?

Tal i com es va poder veure amb la prova de **Wilcoxon**, les següents gràfiques reflexen com no hi ha diferència en el preu mig dels alcovats durant els mesos de Desembre i Gener.

# Preu mitjà dels alvocats durant els mesos de Desembre i Gener  
ggplot(dataDecJan, aes(x = DateMonth, y = AveragePrice)) +   
 stat\_summary(fun.y = "mean", geom = "bar") +   
 xlab("Mesos") +   
 ylab("Preu mitjà ($)") +   
 ggtitle("Preu mitjà unitari durant els mesos de Desembre i Gener")



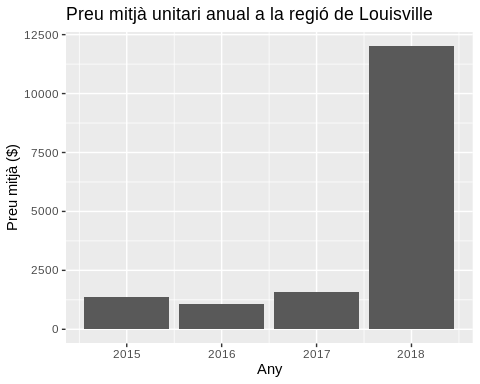
# Preu mitjà dels alvocats durant els mesos de Desembre i Gener  
ggplot(dataDecJan, aes(x = DateYearMonth, y = AveragePrice)) +   
 stat\_summary(fun.y = "mean", geom = "bar") +   
 xlab("Mesos") +   
 ylab("Preu mitjà ($)") +   
 ggtitle("Preu mitjà unitari durant els mesos de Desembre i Gener")



¿S’ha produït un increment del consum d’alvocats de tipus X4046 a la regió de Louisville durant els anys?

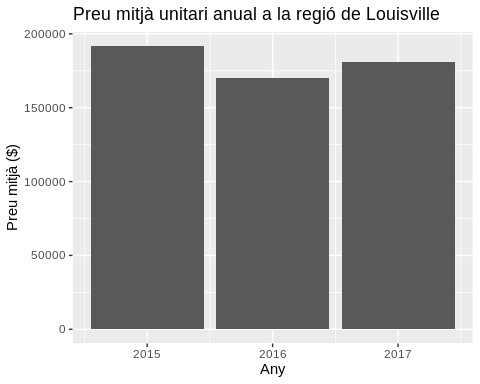
La següent gràfica no coincideix amb els resultats obtinguts amb el test de **Kruskal-Wallis**. S’observa un increment mitjà durant l’any 2018 d’alvocats de tipus X4046 en la regió de Louisville.

# Preu mitjà unitari anual a la regió de Louisville  
ggplot(dataLouisville, aes(x = year, y = X4046)) +   
 stat\_summary(fun.y = "mean", geom = "bar") +   
 xlab("Any") +   
 ylab("Preu mitjà ($)") +   
 ggtitle("Preu mitjà unitari anual a la regió de Louisville")



Per altra banda, si no tenim en consideració l’any 2018 la gràfica coincideix amb els resultats obtinguts amb el test de **Kruskal-Wallis**. El dataset no té dades completes de l’any 2018, només fins el més de Març. Per tant, podem obviar la representació gràfica de l’any 2018 ja que no tenim prou dades d’aquest any.

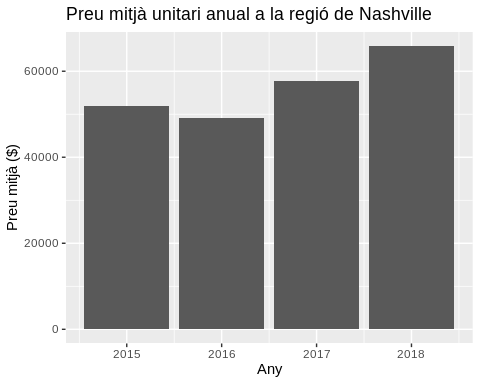
dataLouisville <- data %>%   
 filter(year == 2015 | year == 2016 | year == 2017)  
# Preu mitjà unitari anual a la regió de Louisville  
ggplot(dataLouisville, aes(x = year, y = X4046)) +   
 stat\_summary(fun.y = "mean", geom = "bar") +   
 xlab("Any") +   
 ylab("Preu mitjà ($)") +   
 ggtitle("Preu mitjà unitari anual a la regió de Louisville")



¿S’ha produït un increment del consum d’alvocats de tipus X4046 a la regió de Nashville durant els anys?

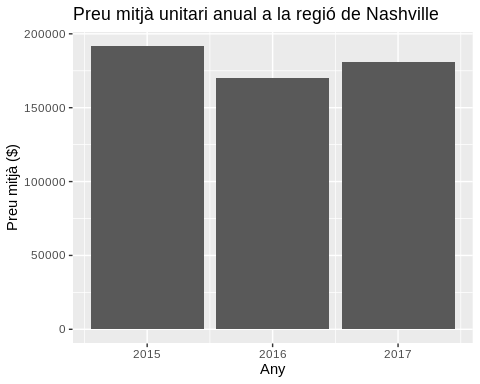
La següent gràfica no coincideix amb els resultats obtinguts amb el test de **Kruskal-Wallis**. S’observa un increment mitjà durant l’any 2018 d’alvocats de tipus X4046 en la regió de Nashville.

# Preu mitjà unitari anual a la regió de Nashville  
ggplot(dataNashville, aes(x = year, y = X4046)) +   
 stat\_summary(fun.y = "mean", geom = "bar") +   
 xlab("Any") +   
 ylab("Preu mitjà ($)") +   
 ggtitle("Preu mitjà unitari anual a la regió de Nashville")



Per altra banda, si no tenim en consideració l’any 2018 la gràfica coincideix amb els resultats obtinguts amb el test de **Kruskal-Wallis**. El dataset no té dades completes de l’any 2018, només fins el més de Març. Per tant, podem obviar la representació gràfica de l’any 2018 ja que no tenim prou dades d’aquest any.

dataNashville <- data %>%   
 filter(year == 2015 | year == 2016 | year == 2017)  
# Preu mitjà unitari anual a la regió de Nashville  
ggplot(dataNashville, aes(x = year, y = X4046)) +   
 stat\_summary(fun.y = "mean", geom = "bar") +   
 xlab("Any") +   
 ylab("Preu mitjà ($)") +   
 ggtitle("Preu mitjà unitari anual a la regió de Nashville")



## Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

La neteja i transformació del dataset, seguit dels tres estudis estadístics plantejats han permès respondre preguntes sobre el preu mig, i del volum d’alvocats venut a les diferents regions dels EEUU.

S’han realitzat tasques de transformació del dataset, generant nous atributs **DateMonth** i **DateYearMonth** que permeten analitzar millor la tendencia temporal de les dades. També s’han realitzat tasques de verificació de les dades durant l’etapa de neteja.

S’han tractat també els valors extrems, eliminant les dades agregades/totalitzades del Total dels EEUU. Tinc la sensació que les dades contenen més valors agregats/totalitzadors per regions, com són el cas de la regió **West**, però no he aconseguit esbrinar com estava plantejada l’organització territorial del dataset (quines regions contenien d’altres); els EEUU disposa de molts tipus d’organitzacions territorials diferents.

En referencia als valors extrems, i després de realitzar l’estudi/pràctica, tinc la sensació que podriem haver eliminat també les dades de l’any 2018, ja que no són dades completes de l’any, només fins al mes de Març.

# **Bibliografia**

* Database  
  <https://www.kaggle.com/neuromusic/avocado-prices>