

# **LAB 1 Y 2**

**Alejandro David Arzola Saveedra**

**2- Ingeniería informática**

**EJERCICIO 1: ANALIZAR CON EL COMANDO SEARCH() LOS PAQUETES PRESENTES EN EL ENTORNO DE TRABAJO. CON LIBRARY(HELP=PACKAGE), SELECCIONAR EL PAQUETE DATASETS, Y, DENTRO DE LOS DISTINTOS CONJUNTOS DE DATOS, VISUALIZAR EN LA CONSOLA LOS CONTENIDOS DE VARIOS DE ELLOS CON DISTINTAS CARACTERÍSTICAS (TIPOS DE VARIABLES, SERIES, ETC.).**

```
1 {r}
2 search()
3
```

[1]	".GlobalEnv"	"tools:rstudio"	"package:stats"
[4]	"package:graphics"	"package:grDevices"	"package:utils"
[7]	"package:datasets"	"package:methods"	"AutoLoads"
[10]	"package:base"		

Como podemos observar al ejecutar el comando “search” en el R-Studio se nos muestra los paquetes que tenemos activos en nuestro programa.

El primer paquete que podemos observar es el “**stats**” que es un paquete de estadística general que podemos ver en Windows o Linux y básicamente nos ayuda en el análisis y para poder manipular los datos y del SPSS que es un programa estadístico. Este paquete tiene funciones para cálculos estadísticos y generación de números aleatorios

El segundo paquete es el “**graphics**” contiene funciones para los graficos “base” .Los graficos base son tipo S.

El paquete “**grDevices**” que sirve para poder crear dispositivos grafico que son utiles para crear gráficos. Es una tipología de grafico en R. Este paquete contiene funciones que admiten graficos base y de cuadrícula

El paquete “**utils**” contiene una colección de funciones de la utilidad

El paquete “**datasets**” sirve para ver una lista de conjunto de datos y funciones en el paquete. Contiene una variedad de conjunto de datos.

El paquete “**methods**” sirve para poder ver una lista de las diferentes funciones de conversión que contiene R, es decir las formas en las que podemos trabajar en R. Son todos los métodos disponibles para una clase S3 o S4, el s3 es un esquema de envio de metodos

Y por último el paquete “**Base**” cuando se instala r R nos ofrece unas funcionalidades mínimas aunque la mayoría se encuentra en paquetes adicionales. Son las construcciones básicas del control de flujo en lenguaje R. Este paquete contiene las funciones básicas como las de aritmética, entrada/salida, soporte básico de programación etc...

# Information on package 'datasets'

## Description:

Package: datasets  
Version: 3.6.1  
Priority: base  
Title: The R Datasets Package  
Author: R Core Team and contributors worldwide  
Maintainer: R Core Team <R-core@r-project.org>  
Description: Base R datasets.  
License: Part of R 3.6.1  
Built: R 3.6.1; ; 2019-07-05 08:06:38 UTC; windows

## Index:

AirPassengers	Monthly Airline Passenger Numbers 1949-1960
BJSales	Sales Data with Leading Indicator
BOD	Biochemical Oxygen Demand
CO2	Carbon Dioxide Uptake in Grass Plants
ChickWeight	Weight versus age of chicks on different diets
DNase	Elisa assay of DNase
EuStockMarkets	Daily Closing Prices of Major European Stock Indices, 1991-1998
Formaldehyde	Determination of Formaldehyde
HairEyeColor	Hair and Eye Color of Statistics Students
Harman23.cor	Harman Example 2.3
Harman74.cor	Harman Example 7.4
Indometh	Pharmacokinetics of Indomethacin
InsectSprays	Effectiveness of Insect Sprays
JohnsonJohnson	Quarterly Earnings per Johnson & Johnson Share
LakeHuron	Level of Lake Huron 1875-1972
LifeCycleSavings	Intercountry Life-Cycle Savings Data
Loblolly	Growth of Loblolly pine trees
Nile	Flow of the River Nile
Orange	Growth of Orange Trees
OrchardSprays	Potency of Orchard Sprays
PlantGrowth	Results from an Experiment on Plant Growth
Purromycin	Reaction Velocity of an Enzymatic Reaction
Theoph	Pharmacokinetics of Theophylline
Titanic	Survival of passengers on the Titanic
ToothGrowth	The Effect of Vitamin C on Tooth Growth in Guinea Pigs
UCBAdmissions	Student Admissions at UC Berkeley
UKDriverDeaths	Road Casualties in Great Britain 1969-84
UKLungDeaths	Monthly Deaths from Lung Diseases in the UK
UKgas	UK Quarterly Gas Consumption
USAccDeaths	Accidental Deaths in the US 1973-1978
USArrests	Violent Crime Rates by US State
USJudgeRatings	Lawyers' Ratings of State Judges in the US Superior Court
USPersonalExpenditure	Personal Expenditure Data
VADeaths	Death Rates in Virginia (1940)
WWWusage	Internet Usage per Minute
WorldPhones	The World's Telephones
ability.cov	Ability and Intelligence Tests
airmiles	Passenger Miles on Commercial US Airlines, 1937-1960
airquality	New York Air Quality Measurements
anscombe	Anscombe's Quartet of 'Identical' Simple Linear Regressions
attenu	The Joyner-Boore Attenuation Data
attitude	The Chatterjee-Price Attitude Data
austres	Quarterly Time Series of the Number of Australian Residents
beavers	Body Temperature Series of Two Beavers
cars	Speed and Stopping Distances of Cars
chickwts	Chicken Weights by Feed Type
co2	Mauna Loa Atmospheric CO2 Concentration
crimtab	Student's 3000 Criminals Data
datasets-package	The R Datasets Package
discoveries	Yearly Numbers of Important Discoveries
esoph	Smoking, Alcohol and (O)esophageal Cancer

euro	Conversion Rates of Euro Currencies
eurodist	Distances Between European Cities and Between US Cities
faithful	Old Faithful Geyser Data
freeny	Freeny's Revenue Data
infert	Infertility after Spontaneous and Induced Abortion
iris	Edgar Anderson's Iris Data
islands	Areas of the World's Major Landmasses
lh	Luteinizing Hormone in Blood Samples
longley	Longley's Economic Regression Data
lynx	Annual Canadian Lynx trappings 1821-1934
morley	Michelson Speed of Light Data
mtcars	Motor Trend Car Road Tests
nhtemp	Average Yearly Temperatures in New Haven
nottem	Average Monthly Temperatures at Nottingham, 1920-1939
npk	Classical N, P, K Factorial Experiment
occupationalStatus	Occupational Status of Fathers and their Sons
precip	Annual Precipitation in US Cities
presidents	Quarterly Approval Ratings of US Presidents
pressure	Vapor Pressure of Mercury as a Function of Temperature
quakes	Locations of Earthquakes off Fiji
randu	Random Numbers from Congruential Generator RANDU
rivers	Lengths of Major North American Rivers
rock	Measurements on Petroleum Rock Samples
sleep	Student's Sleep Data
stackloss	Brownlee's Stack Loss Plant Data
state	US State Facts and Figures
sunspot.month	Monthly Sunspot Data, from 1749 to "Present"
sunspot.year	Yearly Sunspot Data, 1700-1988
sunspots	Monthly Sunspot Numbers, 1749-1983
swiss	Swiss Fertility and Socioeconomic Indicators (1888) Data
treering	Yearly Treering Data, -6000-1979
trees	Diameter, Height and Volume for Black Cherry
uspop	Populations Recorded by the US Census
volcano	Topographic Information on Auckland's Maunga Whau Volcano
warpbreaks	The Number of Breaks in Yarn during Weaving
women	Average Heights and Weights for American Women

Como podemos observar el paquete Dentro del paquete datasets nos encontramos el "Formalhyde". Que nos muestra una "dataframe" de distintos elementos. Estos datos están referidos a un experimento químico para preparar una curva estándar.

El primer elemento es carbohidratos que es un elemento numérico y el segundo optden que es también un elementos numérico que es la densidad óptica.

R Console

```
search()
library(help="datasets")
Formaldehyde |
```

	carb	optden
1	0.1	0.086
2	0.3	0.269
3	0.5	0.446
4	0.6	0.538
5	0.7	0.626
6	0.9	0.782

6 rows

Si ponemos “Titanic” nos salen distintas listas con los supervivientes que hubo después de el Titanic .E s una análisis de la probabilidad de supervivencia de los pasajeros del Titanic .Es una matriz de resultados , son observación individuales divididas e edad, sexo ,hijos, la clase en la que viajaban si eran tripulantes o no y si sobrevivieron o no

```
{r}
Titanic
```

```
, , Age = Child, Survived = No
```

	Sex	
Class	Male	Female
1st	0	0
2nd	0	0
3rd	35	17
Crew	0	0

```
, , Age = Adult, Survived = No
```

	Sex	
Class	Male	Female
1st	118	4
2nd	154	13
3rd	387	89
Crew	670	3

```
, , Age = Child, Survived = Yes
```

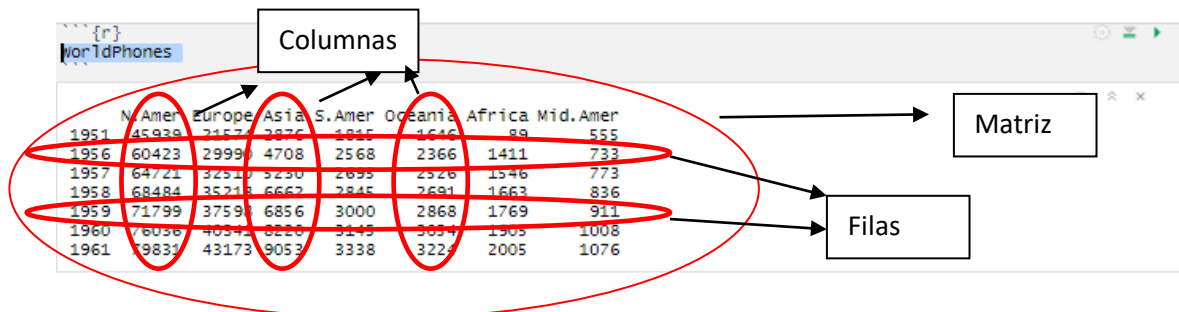
	Sex	
Class	Male	Female
1st	5	1
2nd	11	13
3rd	13	14
Crew	0	0

```
, , Age = Adult, Survived = Yes
```

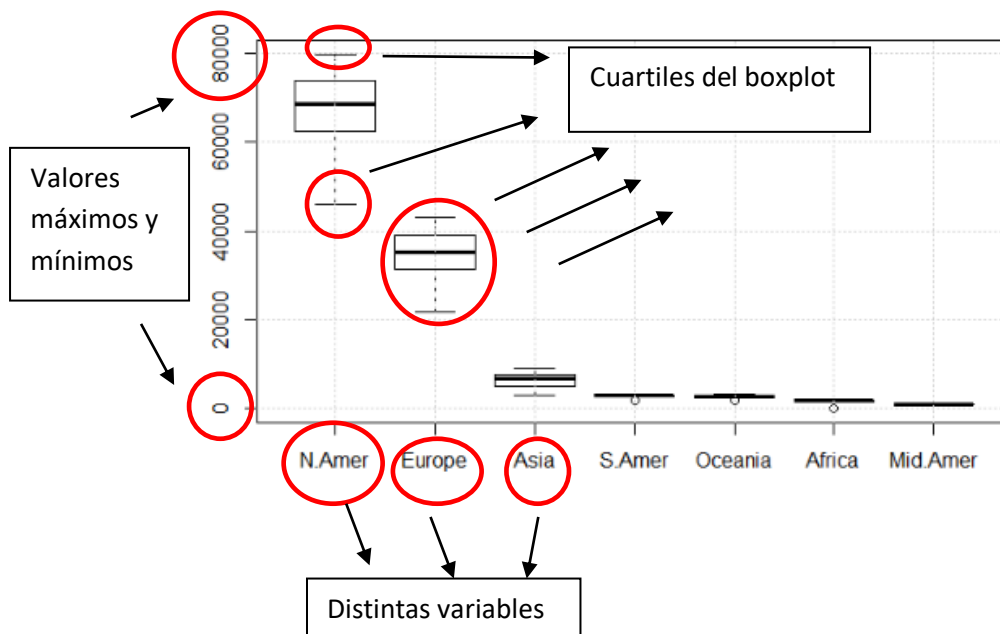
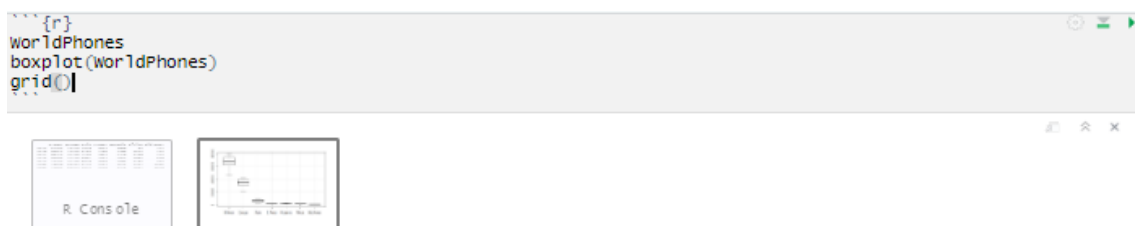
	Sex	
Class	Male	Female
1st	57	140
2nd	14	80
3rd	75	76
Crew	192	20

Es una matriz de 7 filas y 8 columnas con las cifras de distintas zonas y en las filas pos las cifras de distintos años .

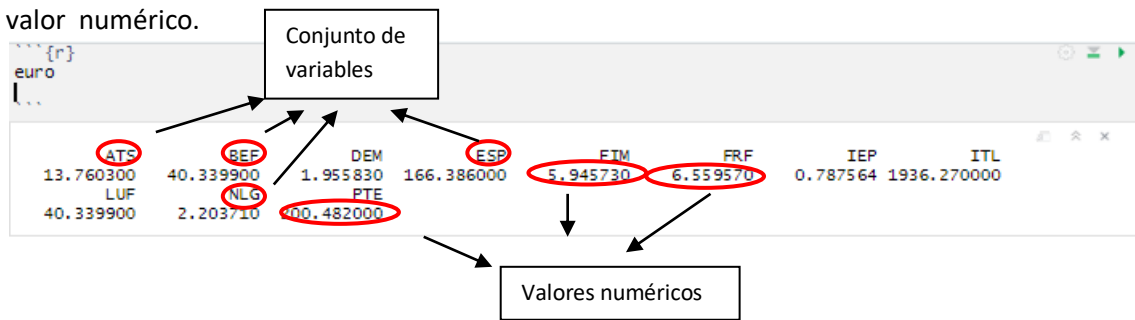
Representa el número de teléfonos en distintas zonas en distintos años



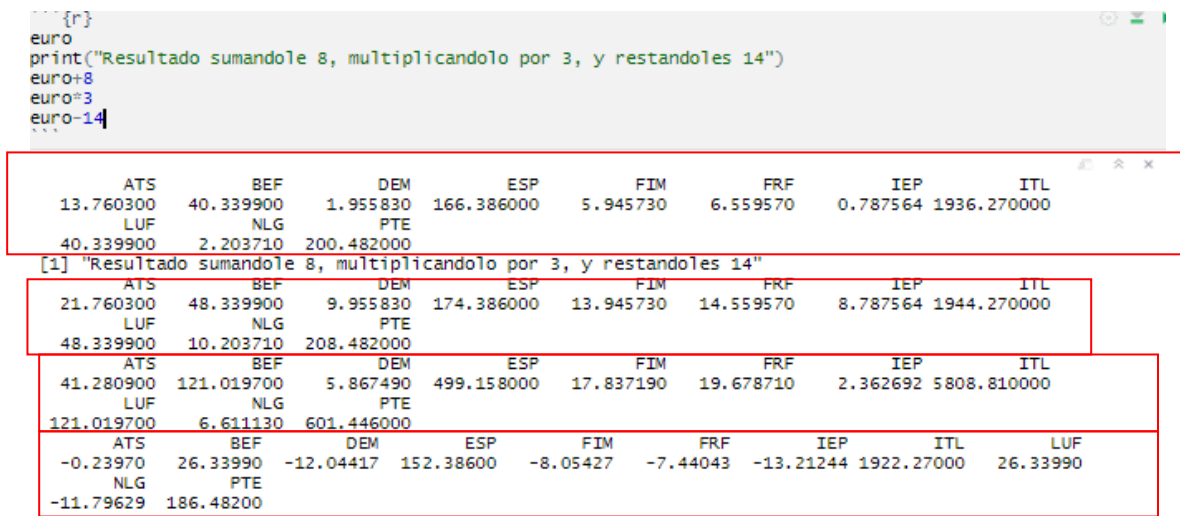
A través de un boxplot podemos observar como la cantidad de móviles en los distintos continentes .Y cómo podemos observar N. América es uno de los continentes que mas teléfonos tiene durante todos los años y que luego lo sigue Europa y Asia. También nos fijamos que el mínimo de móviles en N. América es mayor que el máximo número de móviles en Europa lo cual dice que N. América tiene un gran número de teléfonos.



Con el comando "euro" como observamos podemos ver las distintas conversiones de las monedas .A las distintas variables que son las monedas de cada sitio pos se le ha asignado un valor numérico.



Aquí podemos observar distintas conversiones de los euros como al multiplicarlo por 3 tenemos y al sumarle o restarle distintos valores





- a) Analizar cómo están estructurados los datos para familiarizarse con ellos.
- b) Distinguir claramente en su contenido aquellos que contengan factores y vectores.
- c) Visualizar y direccionar su contenido y realizar algunos cálculos sencillos sobre el mismo.
- d) Generar, utilizando R Markdown, un report de laboratorio que recoja la sesión y explicar en él los resultados que se han obtenido. Utilizar aquellos trozos de código R empotrados (code chunks) con sintaxis knitr que se consideren necesarios para este fin. Alguno de estos “data sets” pueden ser utilizados como parte experimental del proyecto o trabajo de curso. Para otros paquetes puede consultarse el siguiente link

<https://vincentarelbundock.github.io/Rdatasets/datasets.html>

Ejercicio 2: El Data Set “MplsStops” de la librería carData contiene datos de incidencias de personas implicadas en actuaciones policiales por el Departamento de Policía de Minneapolis en 2017. Se pide:

a) Analizar su contenido y visualizar los factores y vectores.

b) Explicar el uso del comando subset() y emplearlo para obtener un subconjunto de este data set que contenga los vectores race, gender y neighborhood para el caso de actuaciones derivadas de accidentes de tráfico: `datos_seleccionados <- subset(datos[problem=="traffic",], select = c(race, gender, neighborhood))`

c) Utilizando el comando ftable() analizar los diferentes porcentajes de accidentes de tráfico según raza y género.

d) Visualizar con el comando gráfico pie() los resultados del apartado anterior.

e) Encontrar en qué zona de Minneapolis se registraron más accidentes.

Como dice el enunciado se trata de las paradas realizadas por la policía de Minneapolis en 2017. Disponemos de 51920 datos de los cuales como podemos ver se encuentran divididos en 14 columnas.

Se trata sobre un “dataframe” son tablas con columnas, las filas son los casos y las columnas son las variables. Es una tabla donde cada columna corresponde a un valor y las filas tratan sobre objetos.

Las variables son:

El “**id numero**” hace referencia a un vector que nos identifica los incidentes

El “**date**” hace referencia a la fecha y la hora de la parada

Luego por otra parte el “**problema**” trata sobre los vehículos sospechosos y paradas de persona o tráfico.

También tenemos el “**citationIssued**” es un factor que nos indica si se emitió la citación.

Por otra parte tenemos el “**personSearch**” que nos indica si se busco a la persona detenida a la persona o no.

También tenemos el “**vehicleSearch**” que nos dice si se busco el vehículo o no.

Por otro lado tenemos “**PreRace**” trata sobre la nacionalidad de la persona característica, negro, blanco, asiático, americano, latino es la evaluación antes de hablar con la persona

Después tenemos “**Race**” que trata sobre lo mismo que “**PreRace**” pero con determinación de la raza después de que haya ocurrido dicho incidente.

Además tenemos “**gender**” que se refiere a si es genero femenino, masculino, o desconocido

Luego tenemos “**lat**” que es la referencia a la latitud de la ubicación del incidente “**long**” se refiere a la longitud del accidente.

“**PolicePrecinct**” Trata sobre el registro policial,

“**neighborhood**” se refiere al vecindario de los incidentes de Minneapolis. Y por ultimo “**MDC**” es los datos a través del ordenador del vehículo enviados por los policías q no están en un vehículo,

Variables del dataframe

idNum	date	problem	MDC	citationIssued	personSearch	vehicleSearch	preRace	race
6823	17-000003	2017-01-01 00:00:42	suspicious	MDC	N/A	N/A	Unknown	Unknown
6824	17-000007	2017-01-01 00:03:07	suspicious	MDC	N/A	N/A	Unknown	Unknown
6825	17-000073	2017-01-01 00:23:15	traffic	MDC	N/A	N/A	Unknown	White
6826	17-000092	2017-01-01 00:33:48	suspicious	MDC	N/A	N/A	Unknown	East African
6827	17-000098	2017-01-01 00:37:58	traffic	MDC	N/A	N/A	Unknown	White
6828	17-000111	2017-01-01 00:46:48	traffic	MDC	N/A	N/A	Unknown	East African
6829	17-000114	2017-01-01 00:48:46	suspicious	MDC	N/A	N/A	Unknown	Black
6830	17-000120	2017-01-01 00:50:55	traffic	MDC	N/A	N/A	Unknown	Other
6831	17-000127	2017-01-01 00:57:10	traffic	MDC	N/A	N/A	Unknown	White
6832	17-000139	2017-01-01 01:05:50	traffic	MDC	N/A	N/A	Unknown	Black

citationIssued	personSearch	vehicleSearch	preRace	race	gender	lat	long	policePrecinct	neighborhood
N/A	NO	NO	Unknown	Unknown	Unknown	44.96662	-93.24646	1	Cedar Riverside
N/A	NO	NO	Unknown	Unknown	Male	44.98045	-93.27134	1	Downtown West
N/A	NO	NO	Unknown	White	Female	44.94835	-93.27538	5	Whittier
N/A	NO	NO	Unknown	East African	Male	44.94836	-93.28135	5	Whittier
N/A	NO	NO	Unknown	White	Female	44.97908	-93.26208	1	Downtown West
N/A	NO	NO	Unknown	East African	Male	44.98054	-93.26363	1	Downtown West
N/A	NO	NO	Unknown	Black	Male	44.98081	-93.27314	1	Downtown West
N/A	NO	NO	Unknown	Other	Female	44.98209	-93.23816	2	Marcy Holmes
N/A	NO	NO	Unknown	White	Male	44.99032	-93.25204	2	Nicollet Island - East Bank
N/A	NO	NO	Unknown	Black	Male	45.01327	-93.30824	4	Folwell

El comando subset() se utiliza básicamente para devolver subconjuntos de los vectores o matrices con y sin condiciones

Como podemos ver con el comando ftable() el mayor porcentaje de personas que intervienen en las paradas policiales de Mineapolis son Hombres que son negros, luego le sigue los hombres blancos, y después de ahí tenemos desconocidos y mujeres que hay un gran porcentaje por no tan grande como el de los hombres . Tambien como curiosidad se puede observar que hay una parte que se desconoce si es hombre o mujer y que los menores incidentes son causante por asiáticos.

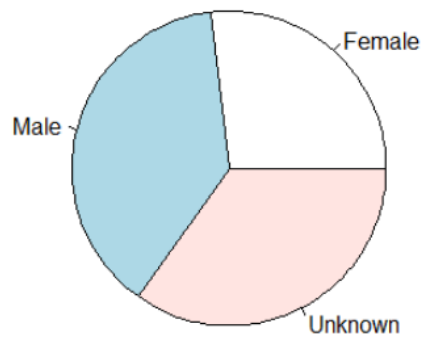
	gender	Female	Male	Unknown
race				
Black		3510	11630	64
White		4036	7635	20
Unknown		447	2521	6235
East African		481	1694	12
Latino		396	1453	8
Native American		631	865	17
Other		295	909	134
Asian		219	424	2

Female	Male	Unknown
11724	16755	15220

```
datos_seleccionados<-subset(datos[problem=="traffic",],select = c(race,gender,neighborhood))
ftable(race,gender)
pie(table(gender[race]))
```

```
data.frame
51920 x 14
```

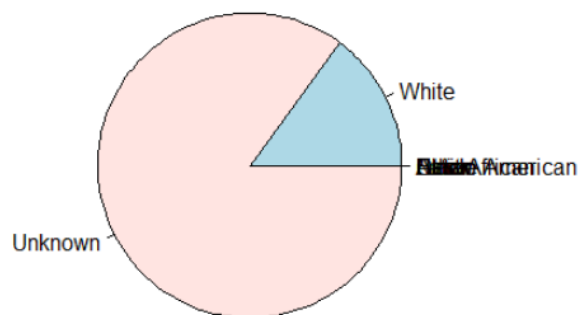
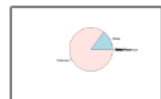
```
R Console
```



```
datos_seleccionados<-subset(datos[problem=="traffic",],select = c(race,gender,neighborhood))
ftable(race,gender)
pie(table(race[gender]))
```

```
data.frame
51920 x 14
```

```
R Console
```



year	year	year	year	year
2000	2001	2002	2003	2004
2005	2006	2007	2008	2009
2010	2011	2012	2013	2014
2015	2016	2017	2018	2019
2020	2021	2022	2023	2024

data.frame  
51920 x 14

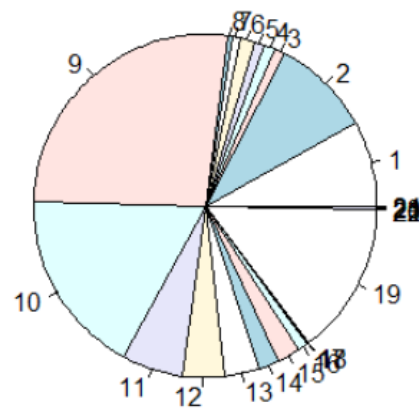
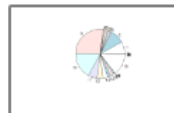
The following appears to be the first column of the data frame.

```

# A tibble: 51920 x 14
  year  year  year  year  year
  <dbl> <dbl> <dbl> <dbl> <dbl>
1 2000  2001  2002  2003  2004
2 2005  2006  2007  2008  2009
3 2010  2011  2012  2013  2014
4 2015  2016  2017  2018  2019
5 2020  2021  2022  2023  2024

```

R Console



data.frame

year	year	year	year	year
2000	2001	2002	2003	2004
2005	2006	2007	2008	2009
2010	2011	2012	2013	2014
2015	2016	2017	2018	2019
2020	2021	2022	2023	2024

data.frame  
51920 x 14

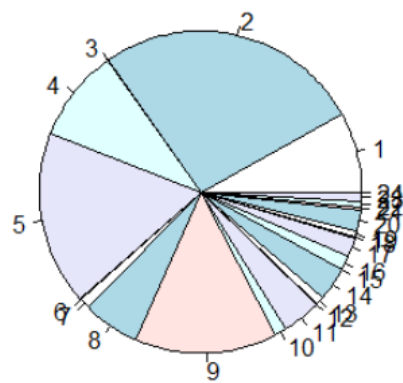
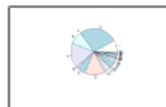
The following appears to be the first column of the data frame.

```

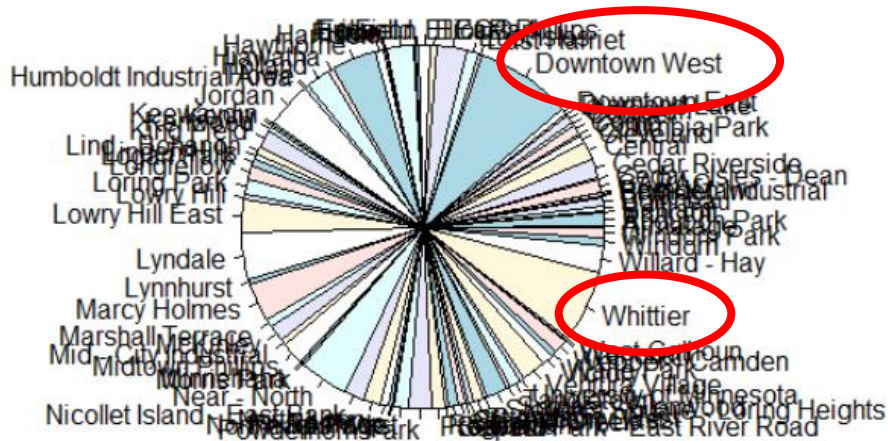
# A tibble: 51920 x 14
  year  year  year  year  year
  <dbl> <dbl> <dbl> <dbl> <dbl>
1 2000  2001  2002  2003  2004
2 2005  2006  2007  2008  2009
3 2010  2011  2012  2013  2014
4 2015  2016  2017  2018  2019
5 2020  2021  2022  2023  2024

```

R Console



Como podemos ver Downtown West es la zona mas peligrosa o donde pueden ocurrir mas incidentes en Minneapolis y también se puede ver que Whittier es la segunda zona mas peligrosa.

neighborhood

Armatage	Audubon Park
77	554
Bancroft	Beltrami
134	211
Bottineau	Bryant
377	96
Bryn - Mawr	Camden Industrial
125	34
CARAG	Cedar - Isles - Dean
559	153
Cedar Riverside	Central
825	832
Cleveland	Columbia Park
356	151
Como	Cooper
452	112
Corcoran	Diamond Lake
360	149
Downtown East	Downtown West
262	4409
East Harriet	East Isles
169	530
East Phillips	ECCO
1387	308



East Harriet	East Isles
169	530
East Phillips	ECCO
1387	308
Elliot Park	Ericsson
544	136
Field	Folwell
87	1230
Fulton	Hale
130	61
Harrison	Hawthorne
401	2031
Hiawatha	Holland
235	1169
Howe	Humboldt Industrial Area
196	10
Jordan	Keewaydin
2075	115
Kenny	Kenwood
118	193
King Field	Lind - Bohanon
846	344
Linden Hills	Logan Park
218	355
Longfellow	Loring Park
603	741
Lowry Hill	Lowry Hill East
243	1491

Lyndale	Lynnhurst
2154	245
Marcy Holmes	Marshall Terrace
1798	355
McKinley	Mid - City Industrial
772	278
Midtown Phillips	Minnehaha
1019	113
Morris Park	Near - North
74	2256
Nicollet Island - East Bank	North Loop
945	799
Northeast Park	Northrop
326	189
Page	Phillips West
41	726
Powderhorn Park	Prospect Park - East River Road
1055	594
Regina	Seward
142	510
Sheridan	Shingle Creek
318	132
St. Anthony East	St. Anthony West
218	475
Standish	Steven's Square - Loring Heights
212	1006

---

Sumner - Glenwood	Tangletown
123	547
University of Minnesota	Ventura Village
218	1096
Victory	Waite Park
498	244
Webber - Camden	Wenonah
656	112
West Calhoun	Whittier
80	3328
Willard - Hay	Windom
1207	404
Windom Park	
461	

Pág. 2 Ejercicio 3: Utilizar el Data Set “Davis” de la librería carData, que proporciona los datos de hombres y mujeres que realizan ejercicio regularmente de peso y altura, tanto medidos como comunicados por los/las afectados/as. El Data Set contiene datos no disponibles (NA´s). Analizar la estructura de los datos correspondientes y:

- Estudiar y aplicar posibles soluciones para los NA´s.
- Encontrar las variaciones de altura y peso reales en función del género. Calcular las medias, medianas y desviación estándar correspondientes.
- Analizar las variaciones de altura y peso comunicadas en función del género. Calcular las medias, medianas y desviación estándar correspondientes.
- Visualizar gráficamente, utilizando boxplot(), una comparativa de los datos de peso medido y peso declarado por un lado y de la altura medida y la altura declarada por otro. Establecer justificadamente las conclusiones.
- Encontrar si hay diferencias significativas entre lo medido y declarado según el género y analizar las posibles formas de corregirla

a) Una posible solución para eliminar los NA's puede ser aplicar la función `na.omit`, que se encarga de omitir todos los na de este `dataFrame`

```
{r}
Davis_sin_na=na.omit(Davis)
print(Davis_sin_na)
```

	sex <fctr>	weight <int>	height <int>	repwt <int>	repht <int>
176	M	71	178	68	178
178	M	66	170	67	165
179	M	81	178	82	175
180	M	68	174	68	173
181	M	80	176	78	175
184	F	63	165	59	160
185	M	70	173	70	173
186	F	56	162	56	160
187	F	60	172	55	168
188	F	58	169	54	166

161-170 of 181 rows

Previous 1 ... 14 15 16 17 18 19 Next

Otra forma de hacerlo seria aplicando el `[is.na(Davis)]` pero esto haría que nuestro `dataframe` se convirtiera a un vector

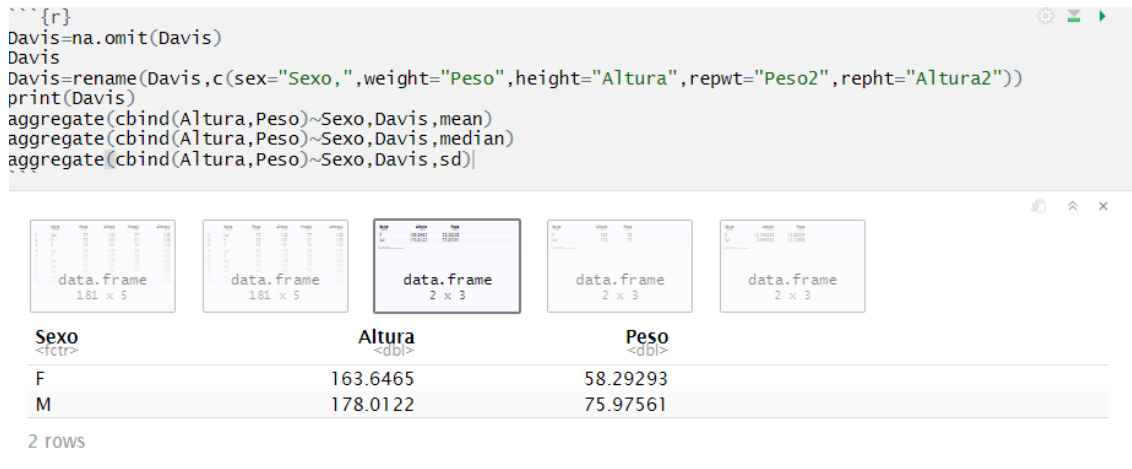
```
{r}
Davis=Davis[!is.na(Davis)]
print(Davis)
```

[1]	"M"	"F"	"F"	"M"	"F"	"M"	"M"	"M"	"M"	"M"	"M"	"F"	"F"	"F"	"F"
[16]	"F"	"M"	"F"	"M"	"F"	"M"	"F"	"M"	"M"	"F"	"F"	"F"	"F"	"F"	"M"
[31]	"F"	"M"	"M"	"F"	"F"	"M"	"F"	"M"	"M"	"F"	"M"	"F"	"M"	"M"	"M"
[46]	"F"	"M"	"F"	"F"	"F"	"M"	"F"	"M"	"M"	"M"	"M"	"F"	"M"	"M"	"M"
[61]	"M"	"M"	"M"	"F"	"M"	"F"	"F"	"F"	"M"	"F"	"M"	"M"	"F"	"F"	"F"
[76]	"F"	"F"	"F"	"M"	"M"	"F"	"M"	"F"	"F"	"M"	"M"	"F"	"F"	"F"	"F"
[91]	"M"	"F"	"M"	"M"	"M"	"F"	"M"	"F"	"F"	"F"	"F"	"M"	"F"	"F"	"F"
[106]	"F"	"F"	"F"	"F"	"F"	"M"	"M"	"F"	"M"	"F"	"F"	"M"	"M"	"M"	"M"
[121]	"M"	"M"	"F"	"F"	"M"	"F"	"F"	"F"	"F"	"F"	"F"	"M"	"F"	"F"	"M"
[136]	"F"	"F"	"F"	"M"	"M"	"M"	"F"	"F"	"F"	"F"	"F"	"F"	"F"	"F"	"M"
[151]	"F"	"F"	"F"	"F"	"F"	"M"	"M"	"F"	"F"	"F"	"F"	"F"	"F"	"F"	"M"
[166]	"F"	"F"	"F"	"M"	"F"	"M"	"F"	"M"	"M"	"M"	"F"	"M"	"M"	"M"	"M"
[181]	"M"	"F"	"M"	"M"	"F"	"F"	"F"	"M"	"F"	"M"	"M"	"F"	"F"	"F"	"F"
[196]	"M"	"M"	"M"	"M"	"M"	"77"	"58"	"53"	"68"	"59"	"76"	"76"	"69"	"71"	"65"
[211]	"70"	"166"	"51"	"64"	"52"	"65"	"92"	"62"	"76"	"61"	"119"	"61"	"65"	"66"	"54"
[226]	"50"	"63"	"58"	"39"	"101"	"71"	"75"	"79"	"52"	"68"	"64"	"56"	"69"	"88"	"65"
[241]	"54"	"80"	"63"	"78"	"85"	"54"	"73"	"49"	"54"	"75"	"82"	"56"	"74"	"102"	"64"
[256]	"65"	"66"	"73"	"75"	"57"	"68"	"71"	"71"	"78"	"97"	"60"	"64"	"64"	"52"	"80"
[271]	"62"	"66"	"55"	"56"	"50"	"50"	"50"	"63"	"69"	"69"	"61"	"55"	"53"	"60"	"56"
[286]	"59"	"62"	"53"	"57"	"57"	"70"	"56"	"84"	"69"	"88"	"56"	"103"	"50"	"52"	"55"
[301]	"55"	"63"	"47"	"45"	"62"	"53"	"52"	"57"	"64"	"59"	"84"	"79"	"55"	"67"	"76"
[316]	"62"	"83"	"96"	"75"	"65"	"78"	"69"	"68"	"55"	"67"	"52"	"47"	"45"	"68"	"44"
[331]	"63"	"87"	"56"	"50"	"82"	"63"	"64"	"63"	"80"	"85"	"66"	"53"	"53"	"54"	"64"

b)

Esta es la media calculada

```
{r}
Davis=na.omit(Davis)
Davis
Davis=rename(Davis,c(sex="Sexo",weight="Peso",height="Altura",repwt="Peso2",repht="Altura2"))
print(Davis)
aggregate(cbind(Altura,Peso)~Sexo,Davis,mean)
aggregate(cbind(Altura,Peso)~Sexo,Davis,median)
aggregate(cbind(Altura,Peso)~Sexo,Davis,sd)
```



Sexo <fctr>	Altura <dbl>	Peso <dbl>
F	163.6465	58.29293
M	178.0122	75.97561

2 rows

Esta es la mediana

```
{r}
Davis=na.omit(Davis)
Davis
Davis=rename(Davis,c(sex="Sexo",weight="Peso",height="Altura",repwt="Peso2",repht="Altura2"))
print(Davis)
aggregate(cbind(Altura,Peso)~Sexo,Davis,mean)
aggregate(cbind(Altura,Peso)~Sexo,Davis,median)
aggregate(cbind(Altura,Peso)~Sexo,Davis,sd)
```

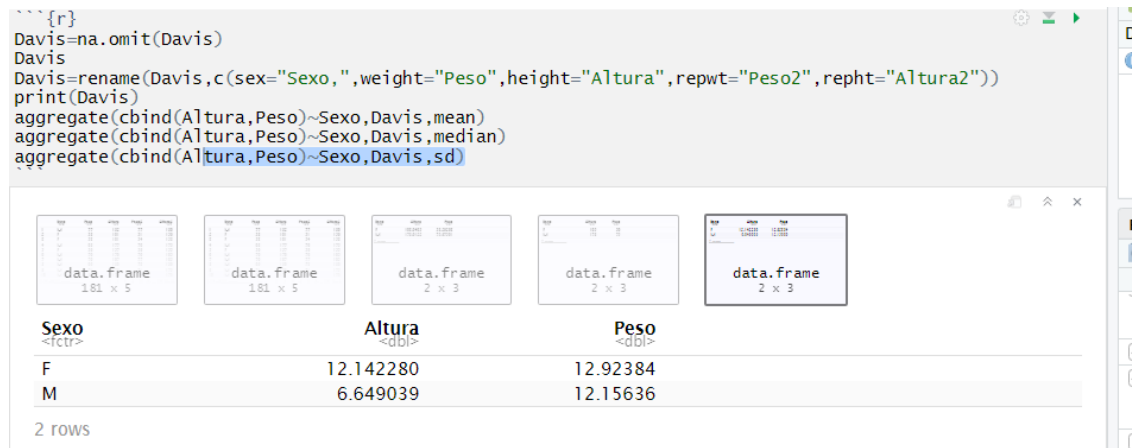


Sexo <fctr>	Altura <dbl>	Peso <dbl>
F	165	56
M	178	75

2 rows

Esta es la desviacion estandar:

```
{r}
Davis=na.omit(Davis)
Davis
Davis=rename(Davis,c(sex="Sexo",weight="Peso",height="Altura",repwt="Peso2",repht="Altura2"))
print(Davis)
aggregate(cbind(Altura,Peso)~Sexo,Davis,mean)
aggregate(cbind(Altura,Peso)~Sexo,Davis,median)
aggregate(cbind(Altura,Peso)~Sexo,Davis,sd)
```



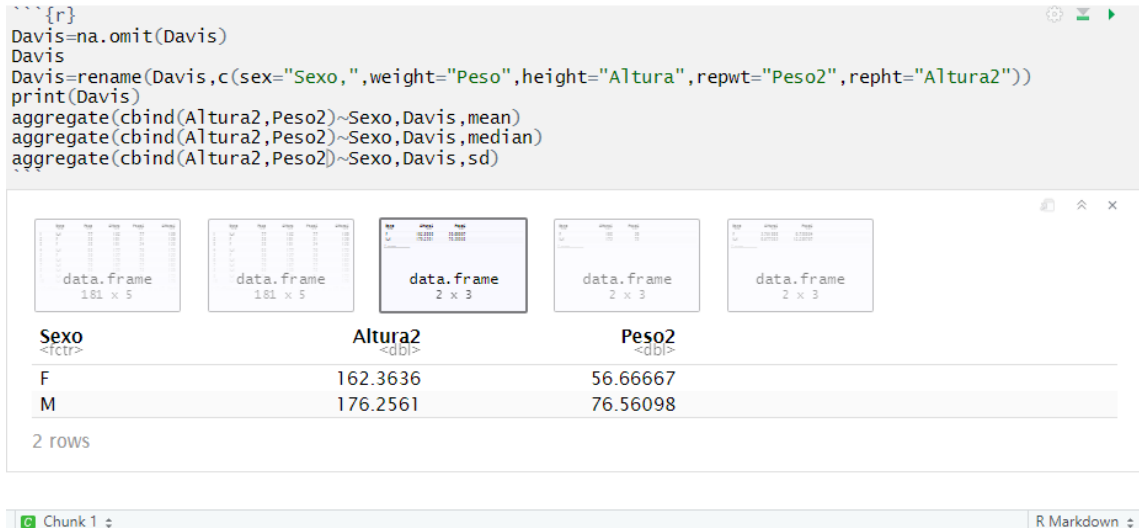
Sexo <fctr>	Altura <dbl>	Peso <dbl>
F	12.142280	12.92384
M	6.649039	12.15636

2 rows

c)

Esta es la media

```
{r}
Davis=na.omit(Davis)
Davis
Davis=rename(Davis,c(sex="Sexo",weight="Peso",height="Altura",repwt="Peso2",repht="Altura2"))
print(Davis)
aggregate(cbind(Altura2,Peso2)~Sexo,Davis,mean)
aggregate(cbind(Altura2,Peso2)~Sexo,Davis,median)
aggregate(cbind(Altura2,Peso2)~Sexo,Davis,sd)
```



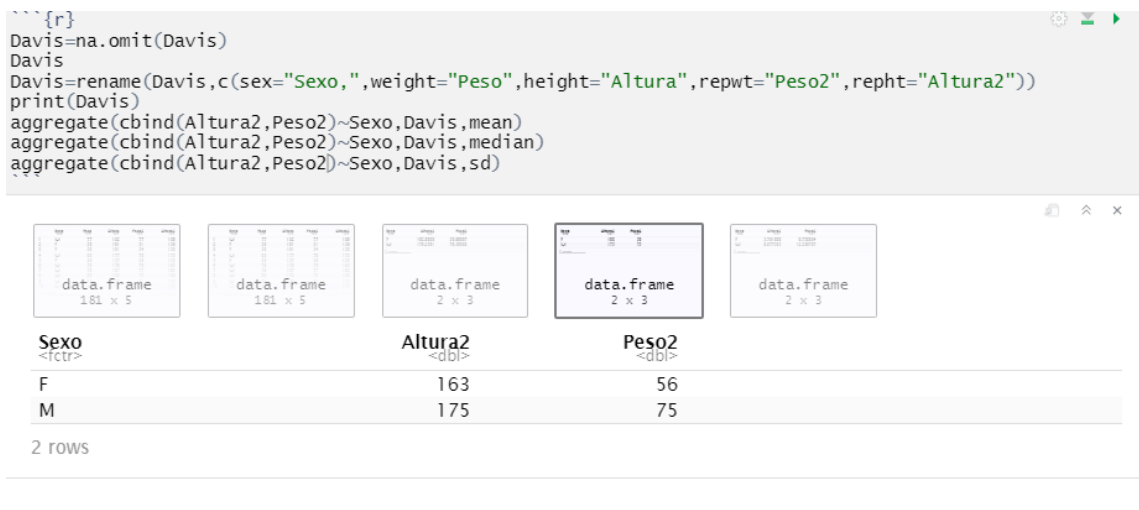
Sexo <fctr>	Altura2 <dbl>	Peso2 <dbl>
F	162.3636	56.66667
M	176.2561	76.56098

2 rows

Chunk 1 R Markdown

Esta es la mediana

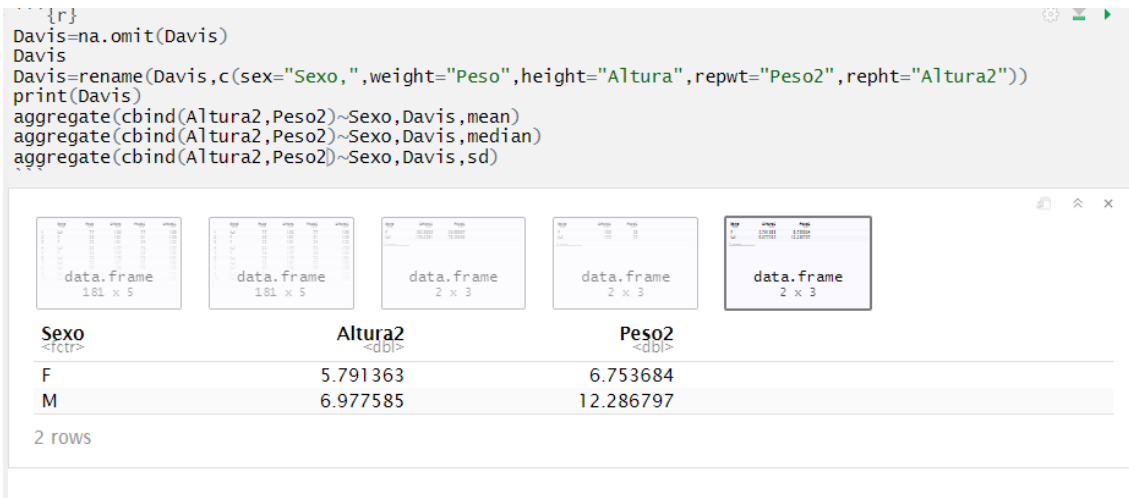
```
{r}
Davis=na.omit(Davis)
Davis
Davis=rename(Davis,c(sex="Sexo",weight="Peso",height="Altura",repwt="Peso2",repht="Altura2"))
print(Davis)
aggregate(cbind(Altura2,Peso2)~Sexo,Davis,mean)
aggregate(cbind(Altura2,Peso2)~Sexo,Davis,median)
aggregate(cbind(Altura2,Peso2)~Sexo,Davis,sd)
```



Sexo <fctr>	Altura2 <dbl>	Peso2 <dbl>
F	163	56
M	175	75

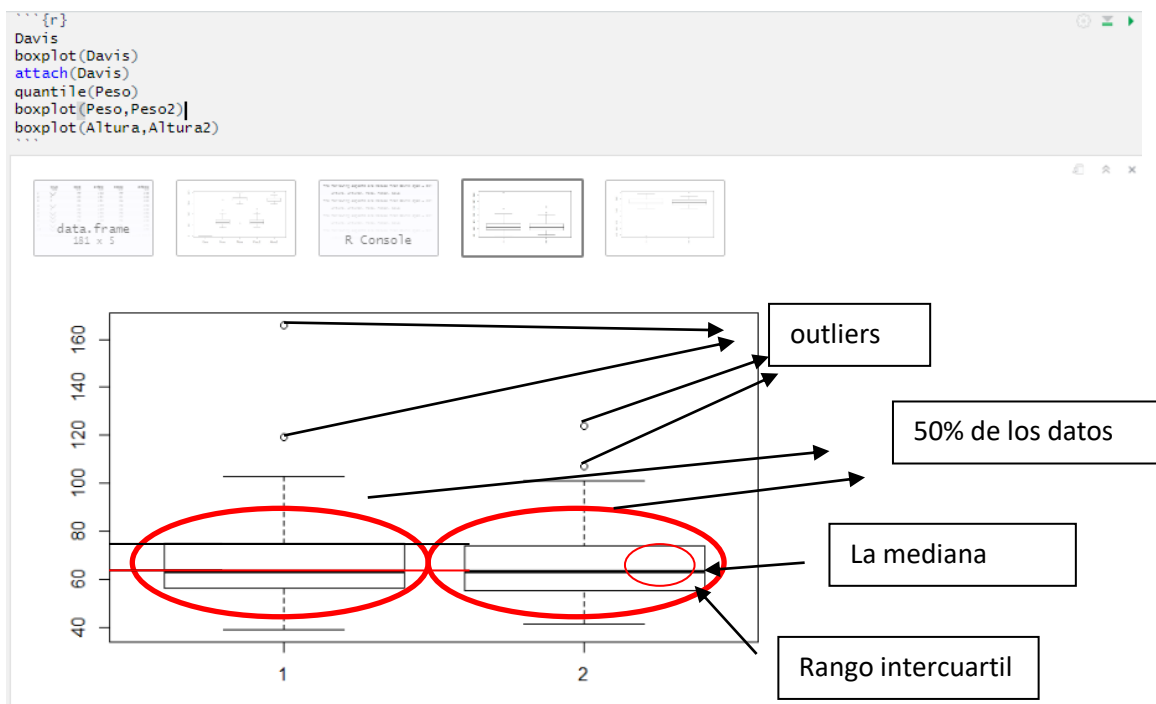
2 rows

## Esta es la desviación estandar



### d)boxplot

Aquí podemos ver la comparación encunto al peso real y el peso declarado de las personas que hacen ejercicio regularmente, Como podemos podemos ver las personas suelen poner un poco más del peso que realmente tienen , el peso es algo que no influye mucho al en estas personas



Aquí podemos ver una comparación de la altura real y de la altura medida como podemos ver las personas suelen ponerse más altura de la que tienen, en estas personas influye mucho la altura y en la sociedad en la que vivimos igual ,hay estudios q declaran que la mayoría de personas altas tienen más posibilidades de conseguir parejas, por otra parte hay muchas personas que se ven acomplejadas por su altura debido a que al sentirse bajos se sienten más inferiores que los demás



e) Decir las diferencias y analizar las formas de corregirla según el genero

Si hay diferencia las personas se ponen más altura de la que en realidad tienen y por otro lado las mujeres se ponen más peso del que tienen pero los hombres se ponen menos del q en realidad tienen

Ejercicio 4 (Opcional): Utilizar la siguiente secuencia de comandos para leer los ficheros “empleados.txt” y “salarios.txt”.

```
setwd("C:/Users/Antonio/Documents/R/Scripts R")
empleados <- read.table("empleados.txt",sep = ",",
header = TRUE) salarios <-
read.table("salarios.txt",sep = "\t",header = TRUE)
names(empleados) [1] "Num_Empleado"
"Fecha_nacimiento" "Nombre" "Apellido" "Genero"
"Fecha_Contrato" names(salarios) [1]
"Num_Empleado" "Salario" "Desde_Fecha"
"Hasta_Fecha" Estos ficheros contienen los datos de
los empleados y salarios de una empresa de
Ingeniería vinculados por un campo común
"Num_Empleado".
```

- a) Analizar el contenido de los Data Frames con los comandos tail() y head()
- b) Razonar sobre los tipos de datos que lo integran (factores y vectores).
- c) Encontrar las medias, medianas y desviaciones estándar de la variable “Salario” agrupada por la variable “Num\_empleado” y encontrar el empleado que más cobra y el que menos en promedio.
- d) Visualizar utilizando boxplot() las variaciones de “Salario” dependiendo del empleado.



e) Utilizar el comando `merge()` para unir los dos data frames unificados por “Num\_empleado” y repetir los apartados c) y d) para el data frame resultante.

f) Con los comandos `interval()` , `now()` y `ymd()` del paquete `lubridate`, determinar la edad de los diez empleados y añadir una nueva columna con el campo “Edad” al data frame resultante del apartado anterior g) Anadir un nuevo registro al data frame del apartado e). Explicar en detalle el proceso.

El comando head nos pone los primeros archivos de arriba como podemos ver el prime empleado es el numero 10001 y llega hasta el 10006 esto es porque el comando head nos indica los 6 primeros líneas desde arriba. Podemos ver que está dividido entre su nacimiento, nombre, apellidos, género y fecha de contrato. Y el salario desde la fecha hasta su fecha final que se le ha estado pagando

Los 6  
primeros  
empezand  
o desde el  
1

```
{r}
setwd("C:/Users/34636/Desktop/Data_Labs_R")
empleados <- read.table("empleados.txt",sep = ",", header = TRUE)
salarios <- read.table("salarios.txt",sep = "\t",header = TRUE)
names(empleados)
names(salarios)
tail(empleados)
head(empleados)
wc(empleados)
```

	Num_Empleado <int>	Fecha_nacimiento <fctr>	Nombre <fctr>	Apellido <fctr>	Genero <fctr>	Fecha_Contrato <fctr>
1	10001	1983-09-02	Mario	Rodriguez	M	2018-06-23
2	10002	1984-06-02	Berta	Santana	F	2011-08-02
3	10003	1986-12-03	Pedro	Brito	M	2011-12-01
4	10004	1984-05-01	Carmelo	Ortega	M	2018-11-28
5	10005	1991-01-21	Oscar	Suarez	M	2018-09-10
6	10006	1983-04-20	Ana	Priego	F	2011-08-02

6 rows

```
{r}
setwd("C:/Users/34636/Desktop/Data_Labs_R")
empleados <- read.table("empleados.txt",sep = ",", header = TRUE)
salarios <- read.table("salarios.txt",sep = "\t",header = TRUE)
names(empleados)
names(salarios)
tail(empleados)
head(salarios)
```

	Num_Empleado <int>	Salario <int>	Desde_Fecha <fctr>	Hasta_Fecha <fctr>
1	10001	60117	2006-06-26	2007-06-26
2	10001	62102	2007-06-26	2008-06-25
3	10001	66074	2008-06-25	2009-06-25
4	10001	66596	2009-06-25	2010-06-25
5	10001	66961	2010-06-25	2011-06-25
6	10001	71046	2011-06-25	2012-06-24

6 rows

El comando tail() pos se encarga de poner los ultimos de la lista, como podemos observar se vuelven a repetir algunos

del anterior y tambien se encuentra dividido de la misma forma que el anterior.

Vuelve a repetir el 5 y el 6 y añade 3 últimos mas

```
{r}
setwd("C:/Users/34636/Desktop/Data_Labs_R")
empleados <- read.table("empleados.txt",sep = ",", header = TRUE)
salarios <- read.table("salarios.txt",sep = "\\t",header = TRUE)
names(empleados)
names(salarios)
tail(empleados)
head(empleados)
wc(empleados)
```

	Num_Empleado	Fecha_nacimiento	Nombre	Apellido	Genero	Fecha_Contrato
5	10005	1991-01-21	Oscar	Suarez	M	2018-09-10
6	10006	1983-04-20	Ana	Priego	F	2011-08-02
7	10007	1987-05-23	Cristina	Bautista	F	2018-02-08
8	10008	1988-02-19	Samuel	Lopez	M	2010-03-10
9	10009	1982-04-19	Sara	Perez	F	2018-02-15
10	10010	1993-06-01	Dori	Pelaez	F	2011-11-23

Como podemos ver tenemos el vector de la tabla en este caso son los encabezados y trata sobre el número ki de empleados su fecha de nacimiento, nombre apellidos su salario y desde una fecha hasta otra


```
{r}
setwd("C:/Users/34636/Desktop/Data_Labs_R")
empleados <- read.table("empleados.txt",sep = ",", header = TRUE)
salarios <- read.table("salarios.txt",sep = "\\t",header = TRUE)
names(empleados)
names(salarios)
tail(empleados)
head(empleados)
wc(empleados)
```

	Num_Empleado	Fecha_nacimiento	Nombre	Apellido	Genero	Fecha_Contrato
[1]	"Num_Empleado"	"Fecha_nacimiento"	"Nombre"	"Apellido"	"Genero"	"Fecha_Contrato"
[1]	"Num_Empleado"	"Salario"	"Desde_Fecha"	"Hasta_Fecha"		

c)

## La media


```
{r}
setwd("C:/Users/34636/Desktop/Data_Labs_R")
empleados <- read.table("empleados.txt",sep = ",", header = TRUE)
salarios <- read.table("salarios.txt",sep = "\t",header = TRUE)
names(salarios)
Total<-aggregate(Salario~(Num_Empleado),salarios,mean)
print(Total)
```



Num_Empleado	Salario
10001	72117.54
10002	68854.50
10003	43030.29
10004	53055.31
10005	85790.90
10006	50514.92
10007	65945.20
10008	49307.67
10009	74006.79
10010	76723.00

## La mediana

```
{r}
setwd("C:/Users/34636/Desktop/Data_Labs_R")
empleados <- read.table("empleados.txt",sep = ",", header = TRUE)
salarios <- read.table("salarios.txt",sep = "\t",header = TRUE)
names(salarios)
Total<-aggregate(Salario~(Num_Empleado),salarios,mean)
Total2<-aggregate(Salario~(Num_Empleado),salarios,median)
Total3<-aggregate(Salario~(Num_Empleado),salarios,sd)
print(Total)
print(Total2)
print(Total3)
```




Num_Empleado	Salario
10001	74333.0
10002	68450.0
10003	43478.0
10004	52119.0
10005	85811.0
10006	50086.0
10007	64019.0
10008	48584.0
10009	73023.0
10010	76799.5

## Y la desviación estándar

```

'''{r}
setwd("C:/Users/34636/Desktop/Data_Labs_R")
empleados <- read.table("empleados.txt",sep = ",", header = TRUE)
salarios <- read.table("salarios.txt",sep = "\t",header = TRUE)
names(salarios)
Total<-aggregate(Salario~(Num_Empleado),salarios,mean)
Total2<-aggregate(Salario~(Num_Empleado),salarios,median)
Total3<-aggregate(Salario~(Num_Empleado),salarios,sd)
print(Total)
print(Total2)
print(Total3)
'''

```




Num_Empleado	Salario
10001	7153.5493
10002	2932.2706
10003	1339.9334
10004	9033.3284
10005	3666.0696
10006	7154.6668
10007	5891.2429
10008	3063.2944
10009	8621.0080
10010	3118.0512

## El máximo y el minimo salario

```

'''{r}
setwd("C:/Users/34636/Desktop/Data_Labs_R")
empleados <- read.table("empleados.txt",sep = ",", header = TRUE)
salarios <- read.table("salarios.txt",sep = "\t",header = TRUE)
names(salarios)
Total3<-aggregate(Salario~(Num_Empleado),salarios,mean)
Total3[which.max(Total3[,2]),]
'''

```



Num_Empleado	Salario
17	87064.7

1 row

```

####{r}
setwd("C:/Users/34636/Desktop/Data_Labs_R")
empleados <- read.table("empleados.txt",sep = ",", header = TRUE)
salarios <- read.table("salarios.txt",sep = "\t",header = TRUE)
names(salarios)
Total3<-aggregate(Salario~(Num_Empleado),salarios,mean)
Total3[which.min(Total3[,2]),]
####

```

R Console

data.frame  
1 x 2

	Num_Empleado <int>	Salario <dbl>
15	10015	40000

1 row

d)

```

####{r}
setwd("C:/Users/34636/Desktop/Data_Labs_R")
empleados <- read.table("empleados.txt",sep = ",", header = TRUE)
salarios <- read.table("salarios.txt",sep = "\t",header = TRUE)
names(salarios)
Total3<-aggregate(Salario~(Num_Empleado),salarios,mean)
Total3[which.max(Total3[,2]),]
boxplot(salarios$Num_Empleado,ylab="Empleados",xlab="Salarios")
####

```

R Console

data.frame  
1 x 2

Ejercicio 5: Ejercicio: Leer el fichero “casas.txt” que incluye el precio medio de viviendas en miles de euros por localizaciones en España. Generar un vector “Precios” a partir de los datos indicados en el fichero. Realizar a continuación las siguientes operaciones:

```
A<-rank(Precios)
```

```
B<- sort(Precios)
```

```
C<- order(Precios) Comparativa<data.frame(Precios,A,B,C)
```

Comparativa Explicar la diferencia entre las diferentes columnas que resultan en cada caso y obtener las casas de precio medio superior a 190.000 €

Las diferencias son que en sort ordena el vector, devuelve los índices del vector ordenado ya el rango de los números del vector el mas pequeño es el rango 1

```
2
3 ~~~{r}
4 getwd()
5 h=read.table("C:/Users/34636/Desktop/Data_Labs_R/casas.txt" , sep = "\t", header=TRUE)
6 h
7 attach(h)
8 names(h)
9 a=rank(Precio)
10 b=sort(Precio)
11 c=order(Precio,decreasing=FALSE)
12 comparativa=data.frame(Precio,a,b,c)
13 comparativa
14 h=data.frame(Localizacion[Precio>190])
15 h
16 ~~~
```

R Console

data.frame  
12 x 2

data.frame  
12 x 4

data.frame  
3 x 1

Localizacion.Precio...190.  
<factor>

Madrid  
Salamanca  
Malaga

3 rows

```

3 ~~~{r}
4 getwd()
5 h=read.table("C:/Users/34636/desktop/Data_Labs_R/casas.txt" , sep = "\t", header=TRUE)
6 h
7 attach(h)
8 names(h)
9 a=rank(Precio)
10 b=sort(Precio)
11 c=order(Precio,decreasing=FALSE)
12 comparativa=data.frame(Precio,a,b,c)
13 comparativa
14 h=data.frame(Localizacion[Precio>180])_
15 h
16 ~~~

```

R Console

data.frame  
12 x 2

data.frame  
12 x 4

data.frame  
5 x 1

Localizacion <fctr>	Precio <int>
Madrid	325
Salamanca	201
Barcelona	157
Castellon	162
Badalona	164
Zaragoza	101
Malaga	211
Teruel	188
Cadiz	95
Albacete	117

```

~~~~~{r}
getwd()
h=read.table("C:/Users/34636/Desktop/Data_Labs_R/casas.txt" , sep = "\t", header=TRUE)
h
attach(h)
names(h)
a=rank(Precio)
b=sort(Precio)
c=order(Precio,decreasing=FALSE)
comparativa=data.frame(Precio,a,b,c)
comparativa
h=data.frame(Localizacion[Precio>180])_
h
~~~~~

```

R Console

data.frame  
12 x 2

data.frame  
12 x 4

data.frame  
5 x 1

Precio <int>	a <dbl>	b <int>	c <int>
325	12.0	95	9
201	10.0	101	6
157	5.0	117	10
162	6.0	121	12
164	7.0	157	3
101	2.0	162	4
211	11.0	164	5
188	8.5	188	8
95	1.0	188	11
117	3.0	201	2

Chunk 1

D. Marin

R Console

data.frame  
12 x 2

data.frame  
12 x 4

data.frame  
5 x 1

Precio <int>	a <dbl>	b <int>	c <int>
188	8.5	211	7
121	4.0	325	1



El fichero "Ventas\_Provincia.txt" contiene datos de ventas en euros de una empresa productora de cereales a distintas provincias españolas durante el año 2012. Se desea realizar un análisis de estos datos para valorar los procesos. Se pide:

- Cantidades totales y las medias anuales de ventas por provincia.
- Provincia en la que más se vende y en la que menos.
- Estudiar la evolución de las ventas de las provincias de Cáceres, Madrid y Barcelona en el segundo semestre de 2012.
- Utilizando los comandos gráficos de base de R, visualizar la evolución temporal de los datos del apartado c)
- Alternativamente, utilizando ggplot2() realizar una visualización de la evolución mensual de los datos del apartado c), tanto absolutos como relativos al total de ventas de la empresa. Explicar las distintas soluciones adoptadas.

a) Las cantidades totales de ventas y las de media anual por provincia son:



Provincia <fctr>	Total_Ventas <dbl>
Albacete	728212.56
Alicante	99064.40
Almeria	450594.81
Asturias	429942.21
Avila	207869.08
Badajoz	440368.13
Barcelona	416216.34
Caceres	368265.55
Gerona	161298.07
Huelva	29392.34

1-10 of 13 rows

Previous 1 2 Next



Provincia <fctr>	Medias <dbl>
Albacete	60684.380
Alicante	9005.855
Almeria	37549.568
Asturias	35828.517
Avila	17322.423
Badajoz	36697.344
Barcelona	34684.695
Caceres	30688.796
Gerona	13441.506
Huelva	2449.362

1-10 of 13 rows

B) la provincia que vende mas es Albacete y la que menos es Oviedo



	Provincia <fctr>	Total_Ventas <dbl>
1	Albacete	728212.6

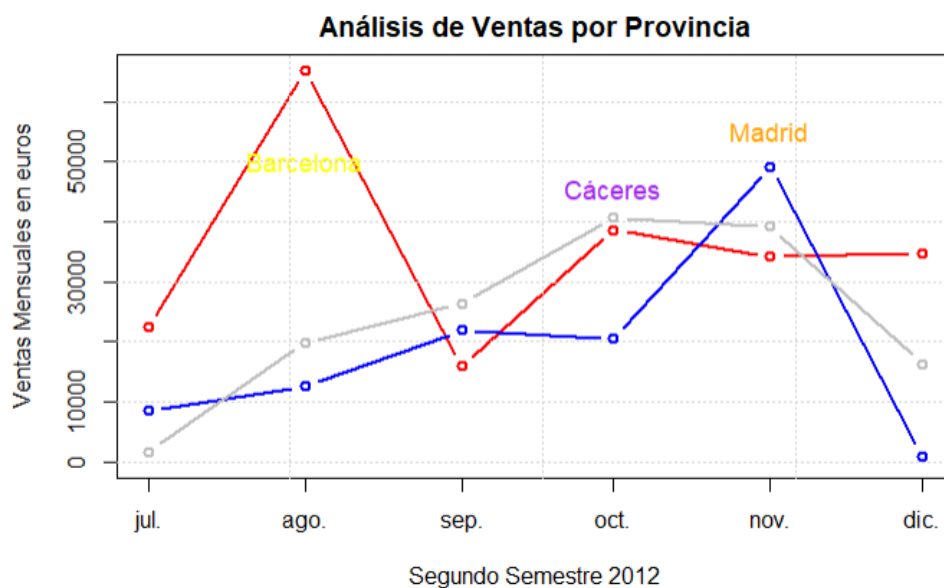
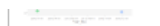
1 row

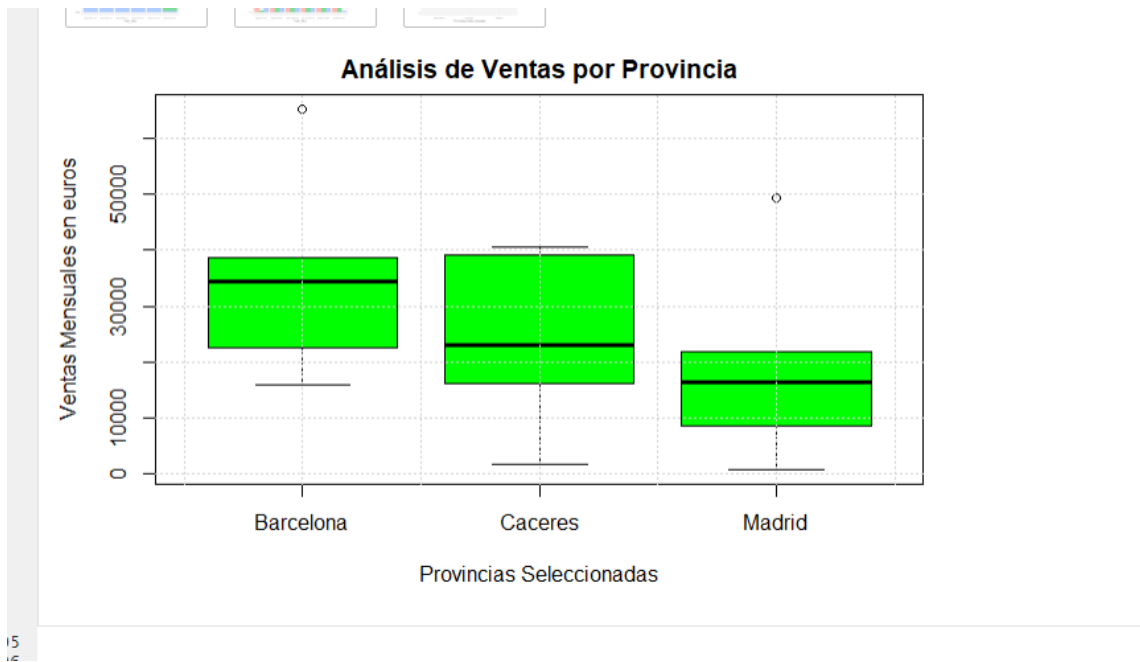


	Provincia <fctr>	Total_Ventas <dbl>
13	Oviedo	26204.42

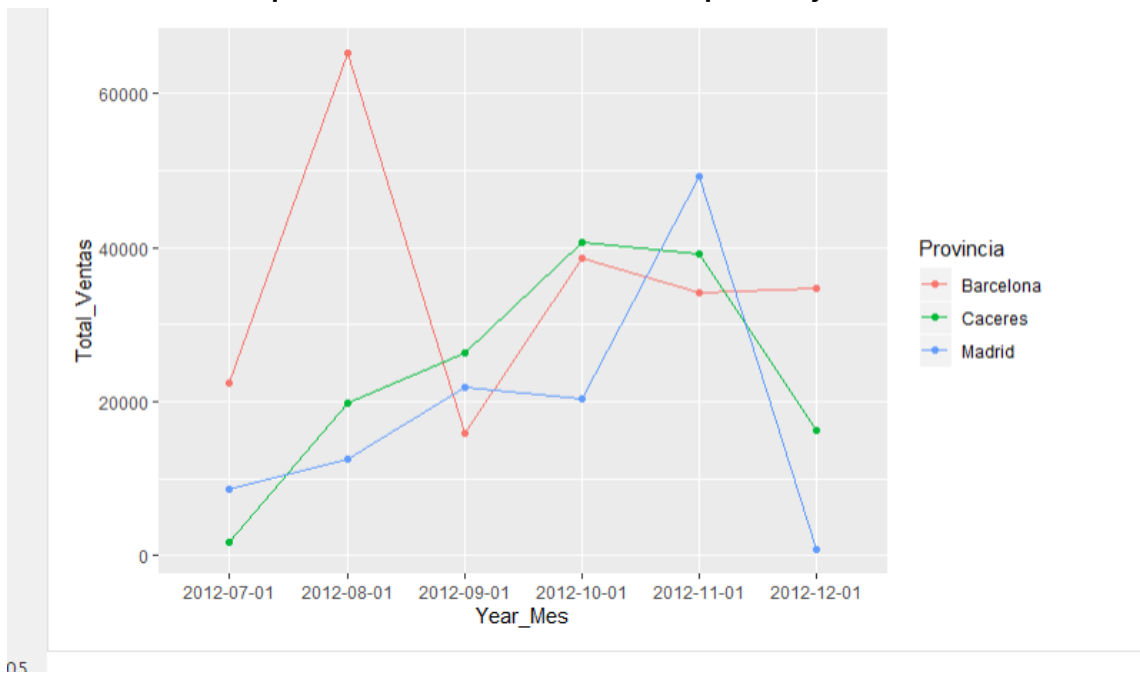
1 row

c) Como podemos ver Barcelona ha sido la que mas ha crecido en el segundo semestre seguido de Madrid han sufrido cambios severos de subidas y bajas altas y Madrid al final acabo decayendo mientras q caceres mas o menos se ha mantenido estable

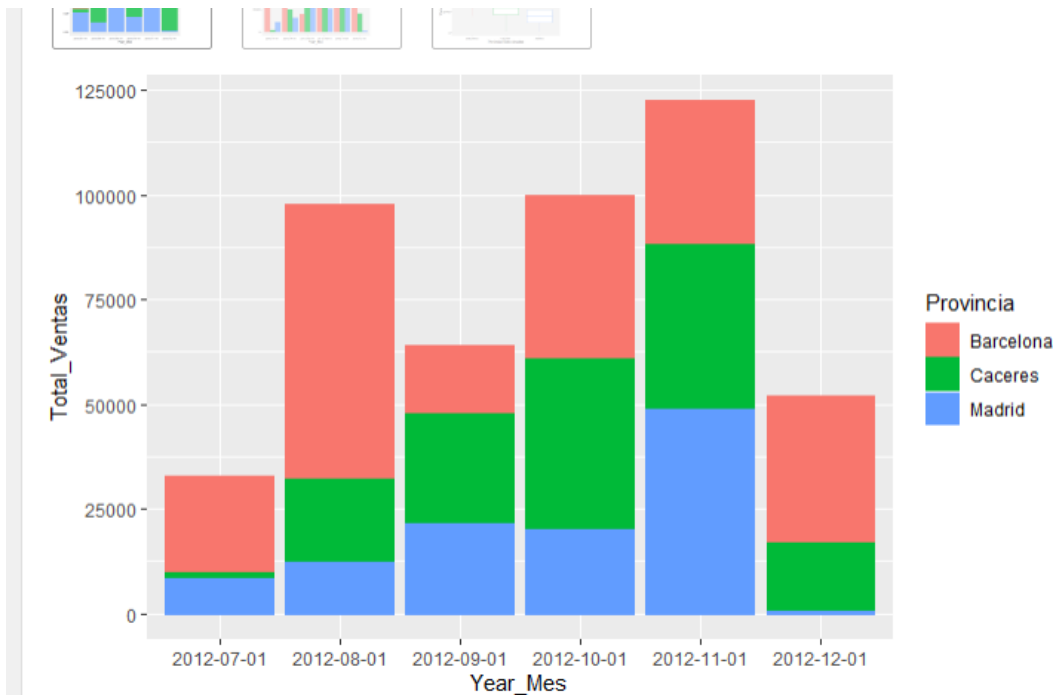




d) como vimos atrás Barcelona fue una de las que mas vendio pero luego bajo mucho y caceres al mantenerse estable es la que al final mas ventas produjo

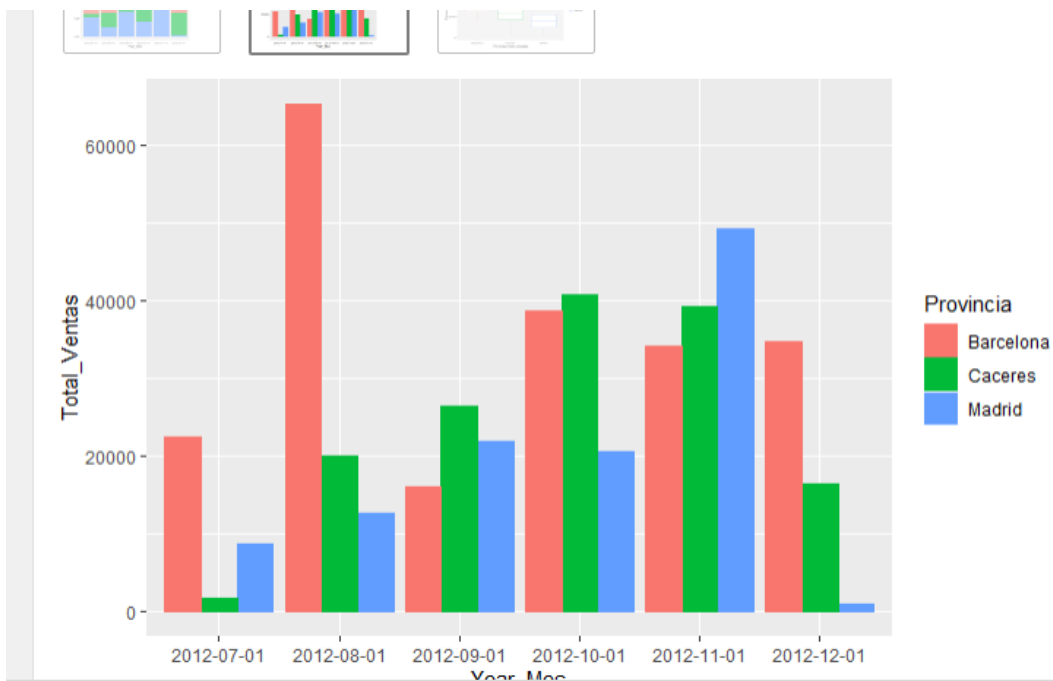


d) Nos encontramos ante un histograma geométrico de barras



En el podemos observar muy claramente cuanto ha sido el crecimiento de cada provincia y cuanto han representado sus ventas en comparación con las otras durante el segundo semestre

Aquí podemos ver de otra manera también como Barcelona es su segundo mes tuvo su mayor cantidad de ventas y saber mas o menos aproximando la cantidad que se vendió



En este boxplot geométrico observamos q caceres ocupa la mayoría de las ventas que su su mayor cantidad de ventas se encuentra en el tercer cuartil y que no ha producido ningún outlier solo en Barcelona y Madrid

