

Unidad 1.1 Bases de datos

Módulo 1

Daniel Enrique González Gómez
Universidad Javeriana Cali

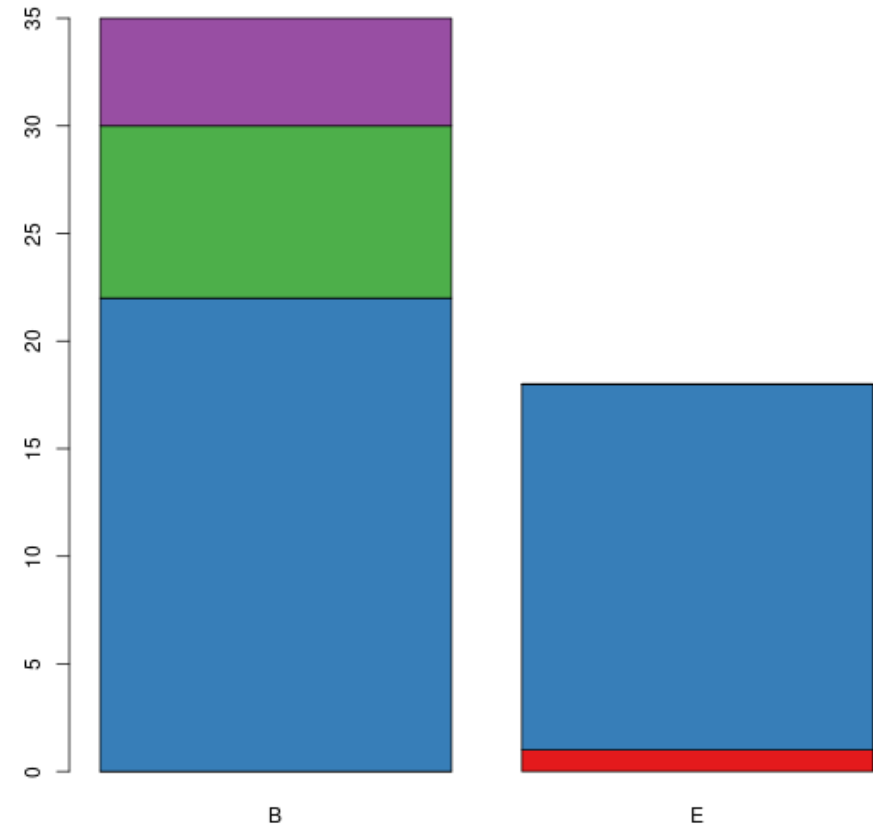
2021-07-26

Que es Estadística ?

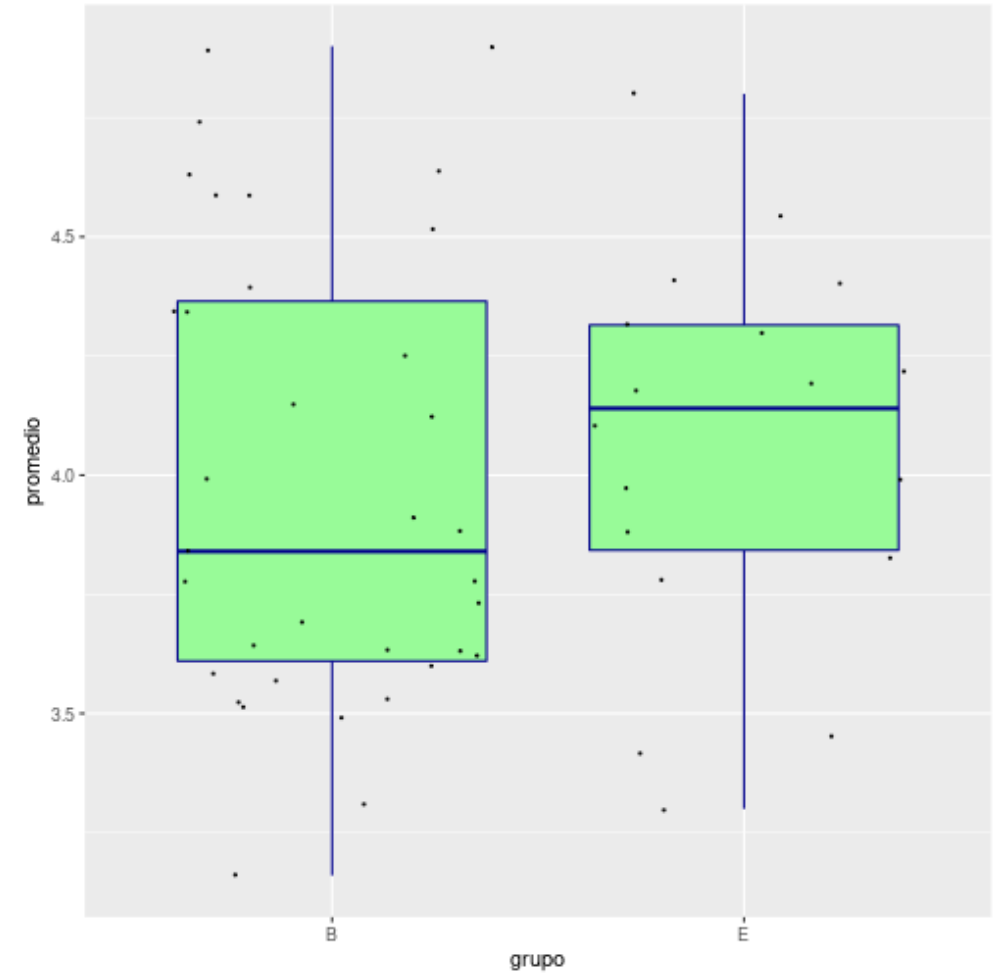
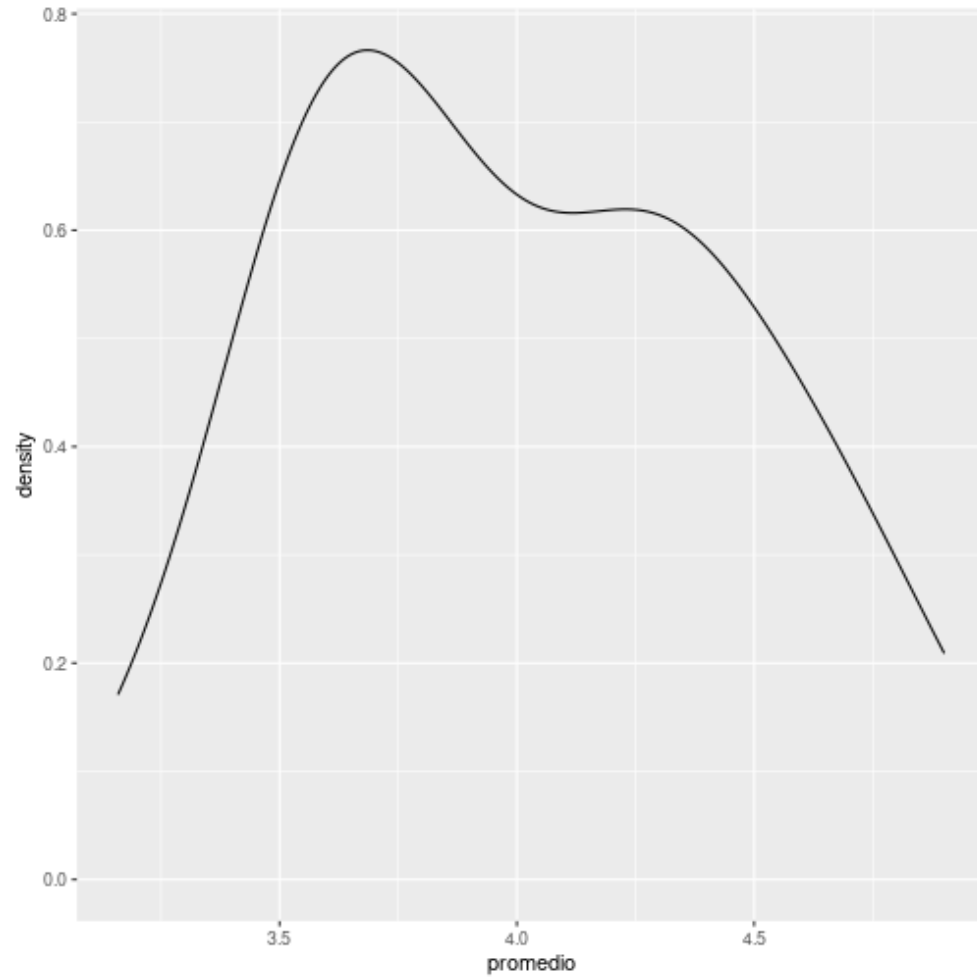
Análisis de datos para la toma de decisiones

Grupos Probabilidad y Estadística 2021-2

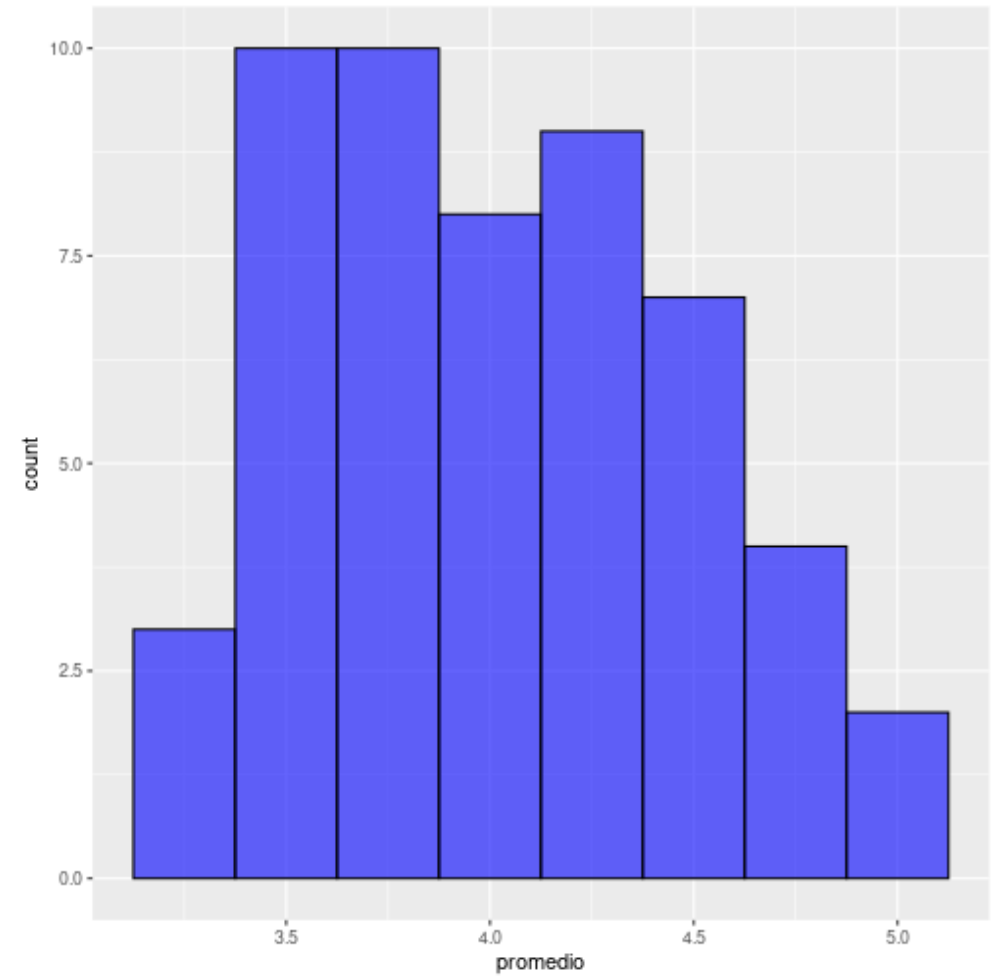
	B	E
bio	0	1
civ	22	17
mec	8	0
sis	5	0



Promedio académico



	promedio
Mean	4.011
Std.Dev	0.454
Min	3.160
Q1	3.630
Median	3.970
Q3	4.340
Max	4.900
MAD	0.549
IQR	0.710
CV	0.113
Skewness	0.202
SE.Skewness	0.327
Kurtosis	-1.051
N.Valid	53.000
Pct.Valid	100.000





Base de datos

Una base de datos es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso.

Wikipedia

Una base de datos en estadística es un conjunto de información relacionada con una población organizada en filas y columnas. Las columnas corresponden a las variables y las filas están relacionadas con los individuos u objetos de estudio.

Existen repositorio de bases de datos para uso general

- dataset en RStudio
- [Portal Bases de datos abiertos Colombia](#)
- [Datos Banco mundial](#)
- [Portal de Datos Abiertos de Esri España](#)

[*] Open Data Barometer : <https://opendatabarometer.org/4thedition/report/?lang=es>

Base de datos

Base datos iris (dataset R)

```
head(iris)
```

##	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
## 1	5.1	3.5	1.4	0.2	setosa
## 2	4.9	3.0	1.4	0.2	setosa
## 3	4.7	3.2	1.3	0.2	setosa
## 4	4.6	3.1	1.5	0.2	setosa
## 5	5.0	3.6	1.4	0.2	setosa
## 6	5.4	3.9	1.7	0.4	setosa

Datos de iris (de Fisher o Anderson)

- longitud y ancho del sépalo
- largo y ancho de pétalos
- especies: setosa, versicolor y virginica.

Base de datos estadísticos : arreglo de filas y columnas (matriz) donde por lo general las columnas representan las variables y las filas los registros de los objetos de estudio

Base de datos

Una base de datos es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso.

Wikipedia

Una base de datos en estadística es un conjunto de información relacionada con una población organizada en filas y columnas. Las columnas corresponden a las variables y las filas están relacionadas con los individuos u objetos de estudio.

Existen repositorio de bases de datos para uso general

- dataset en RStudio
- [Portal Bases de datos abiertos Colombia](#)
- [Datos Banco mundial](#)
- [Portal de Datos Abiertos de Esri España](#)

[*] Open Data Barometer : <https://opendatabarometer.org/4thedition/report/?lang=es>



Base de datos

Base datos iris (dataset R)

```
DT::datatable(head(iris, 150), fillContainer = FALSE, options = list(pageLength = 8))
```

Show entries

Search:

	Sepal.Length 	Sepal.Width 	Petal.Length 	Petal.Width 	Species 
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa

Showing 1 to 8 of 150 entries

Base de datos estudiantiles Probabilidad y Estadística 2021-2

```
var1=c(4,5,6)
DT::datatable(head(bd0052[var1],53),fillContainer = FALSE, options = list(pageLength = 8))
```

Show entries

Search:

	grupo		promedio	programa
1	B		4.12	civ
2	B		3.84	civ
3	B		4.25	mec
4	B		3.31	civ
5	B		3.78	civ
6	B		3.52	sis
7	B		4.9	civ
8	B		4.63	civ

Showing 1 to 8 of 53 entries

Previous

1

2

3

4

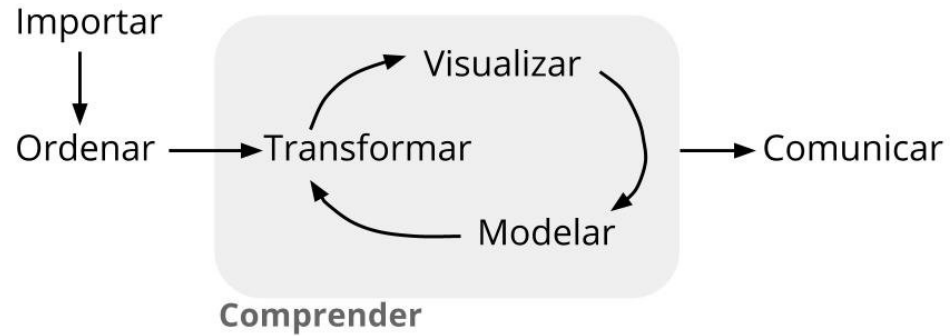
5

6

7

Next

Etapas del proceso de datos



[*] Imagen tomada de : https://bitsandbricks.github.io/ciencia_de_datos_gente_sociable/

Importar datos

Origen de los datos

- Encuesta personal (datos primarios)
 - Online
 - Entrevista cara a cara
 - Entrevista telefónica
- Investigación propia
- Sistema automático de recolección de datos
- Fuente externa (datos secundarios)
 - DANE
 - Cámara de Comercio
 - Agremiaciones
- Bancos de datos abiertos
- Otros medios...

Herramientas computacionales

- Excel
- SQL
- Oracle
- SAS
- R
- RStudio
- Python

Video: Importar datos en R
por Rafa Gonzalez Gouveia
https://youtu.be/Bi0PoYq_gjE

Limpieza de datos

Es importante después de haber importado la base de datos, hacer una revisión de cada una de las variables con el fin de poder detectar:

- Datos faltantes (NA)
- Datos anómalos o raros
- Etiquetas mal colocadas (minúsculas, MAYÚSCULAS, Título...)

Existen metodologías para corregir estos problemas sin afectar la información contenida en la data

Ficha técnica

Las bases de datos debe estar acompañadas de una ficha técnica donde se indican sus principales características :

- [Ficha tecnica](#)
- [Casos positivos de COVID-19 en Colombia](#)

Actividades a realizar

A1 Metodología estadística : Formular un problema que le permita desarrollar un ejercicio académico durante el semestre a través de la recolección de información (primaria o secundaria), Además deberá establecer los objetivos y las variables de interés , para las cuales deberá identificar el tipo de variable y su escala de medición. El resultado de esta actividad deberá se entregado en archivo pdf con nombre: **actividad1.pdf**

A2 Base de datos : Deberá buscar una base de datos de su interés en el portal <https://www.datos.gov.co>, depurarla y documentarla si es necesario. A partir de la información recolectada deberá construir la ficha técnica de la base. El resultado de esta actividad deberá se entregado en archivo pdf con nombre: **actividad2.pdf**

A3 Instalación de R y RStudio : Para el desarrollo de las actividades del curso deberá instalar las ultimas versiones de [\href{https://www.r-project.org/}](https://www.r-project.org/) y de <https://rstudio.com/products/rstudio/download/>.

Gracias

Daniel Enrique González Gómez



Imagen tomada de : <https://www.javerianacali.edu.co/noticias/la-javeriana-bogota-y-cali-1-de-colombia>