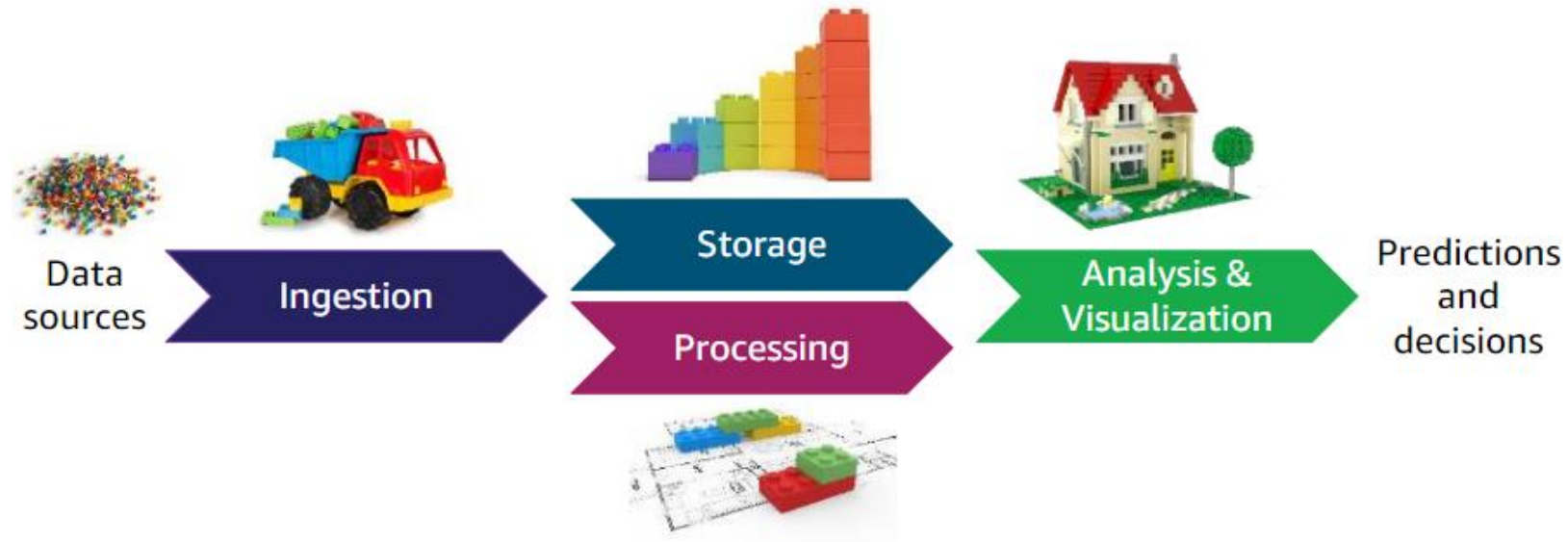


CIENCIA DE DATOS

GRADO EN BIOMEDICINA

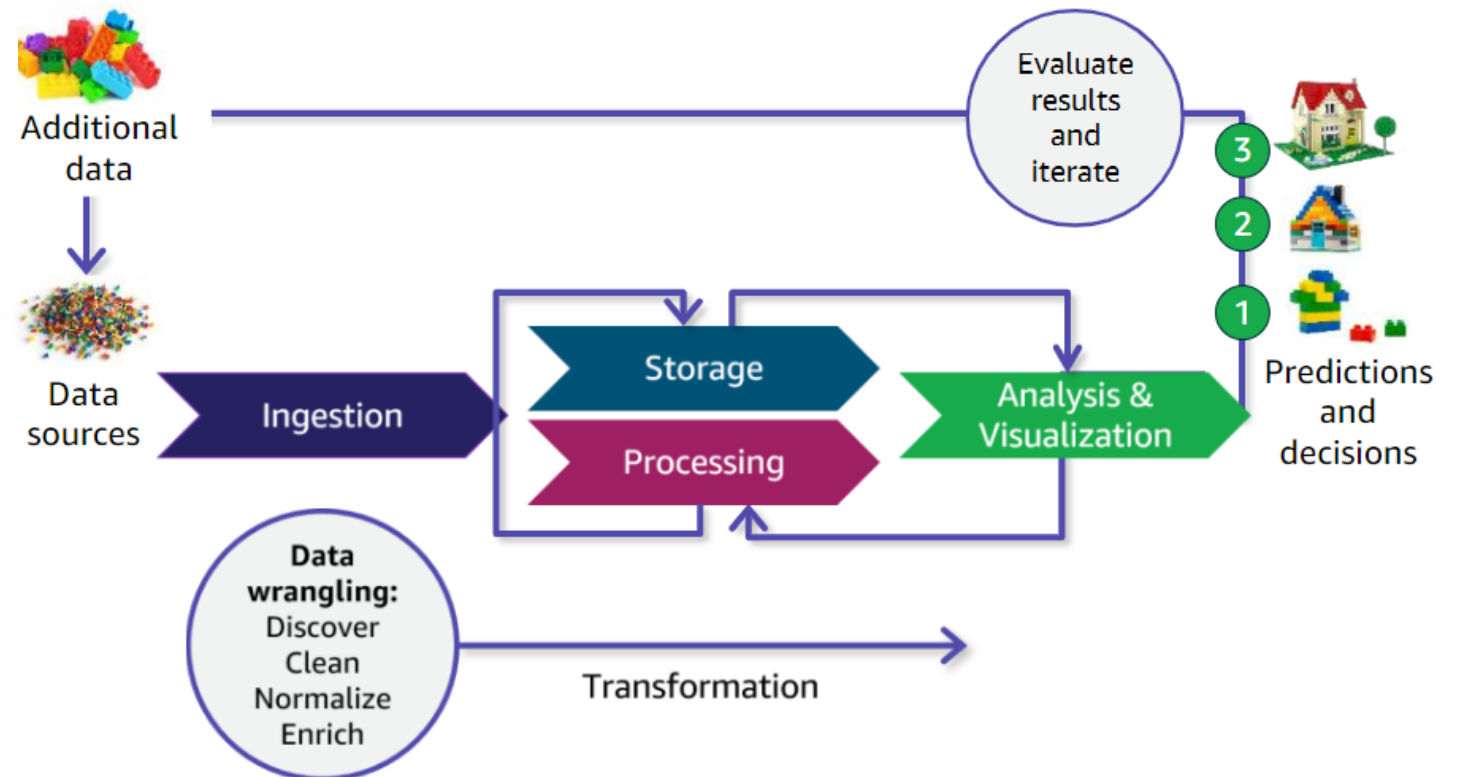
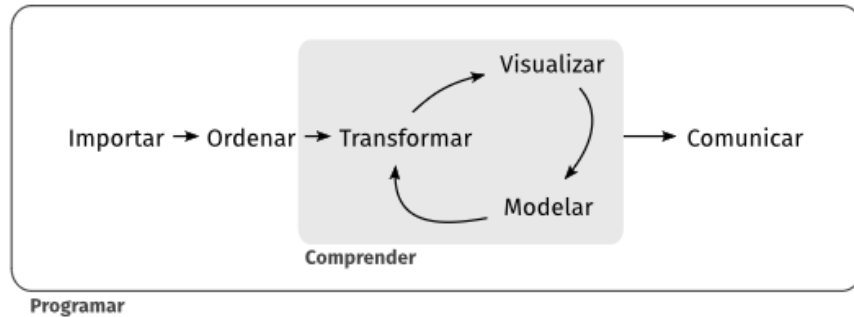
DR. ARIEL CARIAGA-MARTÍNEZ

Ciencia de datos: fases → PIPELINES



Ciencia de datos: fases

CONTEXTO + MÉTODO CIENTÍFICO



ADQUIRIR → ANALIZAR → EXPLICAR

ANÁLISIS EXPLORATORIO DE LOS DATOS

ES EL "ARTE DE MIRAR" LOS DATOS DE UNA FORMA CUIDADOSA Y ESTRUCTURADA (SISTEMÁTICA).

EXPLORATORIO VS EXPLICATIVO

Las 4 "R"s:

- Revelaciones: visualización
- Residuos: (futura) validación de modelos aplicados (aunque sean triviales).
- Re-Expresión: "ingeniería de variables" (transformaciones matemáticas para mejorar procesos posteriores). ¿Correlaciones?
- Resistencia: análisis de outliers, "anormalidades"

CONOCIENDO LOS DATOS

Características de formatos de datos:

- **Independiente del lenguaje (de programación).**
- **Soporte de estructuras complejas (como anidamiento)**
- **Eficiente/Dinámico.**
- **Formato completo que permita la división (separación/compresión)**

Según el formato del almacenamiento:

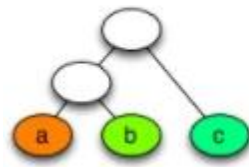
- **Texto/Binario**
- **Filas vs columnas**

Característica	CSV	XML / JSON	SequenceFile
Independencia del lenguaje	👍	👍	👎
Expresivo	👎	👍	👍
Eficiente	👎	👎	👍
Dinámico	👍	👍	👎
<i>Standalone</i>	?	👍	👎
Divisible	?	?	👍

<https://manoli-iborra.github.io/BigData/index.html>

Organización de datos

Row storage		Column storage	
Row 1	1	user_id	1
	US		2
	Free		3
Row 2	2	country	US
	UK		UK
	Paid		ES
Row 3	3	subscription_type	Free
	ES		Paid
	Paid		Paid



Nested schema

Logical table representation

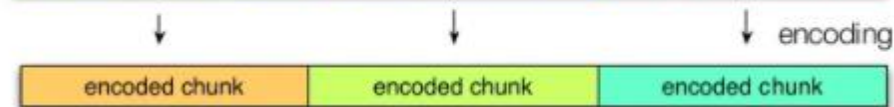
a	b	c
a1	b1	c1
a2	b2	c2
a3	b3	c3
a4	b4	c4
a5	b5	c5

Row layout

a1	b1	c1	a2	b2	c2	a3	b3	c3	a4	b4	c4	a5	b5	c5
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

Column layout

a1	a2	a3	a4	a5	b1	b2	b3	b4	b5	c1	c2	c3	c4	c5
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----



“Datos ordenados”

tabla1

```
# A tibble: 6 × 4
  pais      anio  casos poblacion
  <chr>    <dbl> <dbl>    <dbl>
1 Afganistán 1999    745  19987071
2 Afganistán 2000   2666  20595360
3 Brasil     1999  37737  172006362
4 Brasil     2000  80488  174504898
5 China      1999 212258 1272915272
6 China      2000 213766 1280428583
```

tabla2

```
# A tibble: 12 × 4
  pais      anio tipo      cuenta
  <chr>    <dbl> <chr>    <dbl>
1 Afganistán 1999 casos        745
2 Afganistán 1999 población  19987071
3 Afganistán 2000 casos        2666
4 Afganistán 2000 población  20595360
5 Brasil     1999 casos        37737
6 Brasil     1999 población  172006362
7 Brasil     2000 casos        80488
8 Brasil     2000 población  174504898
9 China      1999 casos        212258
10 China     1999 población 1272915272
11 China     2000 casos        213766
12 China     2000 población 1280428583
```

tabla3

```
# A tibble: 6 × 3
  pais      anio tasa
  <chr>    <dbl> <chr>
1 Afganistán 1999 745/19987071
2 Afganistán 2000 2666/20595360
3 Brasil     1999 37737/172006362
4 Brasil     2000 80488/174504898
5 China      1999 212258/1272915272
6 China      2000 213766/1280428583
```

Dividido en dos tibbles

tabla4a # casos

```
# A tibble: 3 × 3
  pais      `1999` `2000`
  <chr>    <dbl> <dbl>
1 Afganistán    745    2666
2 Brasil       37737  80488
3 China       212258  213766
```

tabla4b # poblacion

```
# A tibble: 3 × 3
  pais      `1999` `2000`
  <chr>    <dbl> <dbl>
1 Afganistán 19987071 20595360
2 Brasil    172006362 174504898
3 China    1272915272 1280428583
```

Datos ordenados

- Cada variable en su propia columna.
- Cada observación en su propia fila
- Cada valor en su propia celda.

pais	anio	casos	poblacion
Afganistán	1999	745	19987071
Afganistán	2000	2666	20595360
Brasil	1999	37737	172006362
Brasil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

variables

pais	anio	casos	poblacion
Afganistán	1999	745	19987071
Afganistán	2000	2666	20595360
Brasil	1999	37737	172006362
Brasil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

observaciones

pais	anio	casos	poblacion
Afganistán	1999	745	19987071
Afganistán	2000	2666	20595360
Brasil	1999	37737	172006362
Brasil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

valores

Datos ordenados: cada variable en su columna, cada observación en su fila, cada valor único en su celda

tabla1

```
# A tibble: 6 × 4
  pais      anio  casos poblacion
<chr>    <dbl> <dbl>    <dbl>
1 Afganistán 1999    745  19987071
2 Afganistán 2000   2666  20595360
3 Brasil     1999  37737  172006362
4 Brasil     2000  80488  174504898
5 China      1999 212258 1272915272
6 China      2000 213766 1280428583
```

tabla2

```
# A tibble: 12 × 4
  pais      anio tipo      cuenta
<chr>    <dbl> <chr>    <dbl>
1 Afganistán 1999 casos      745
2 Afganistán 1999 población 19987071
3 Afganistán 2000 casos      2666
4 Afganistán 2000 población 20595360
5 Brasil     1999 casos      37737
6 Brasil     1999 población 172006362
7 Brasil     2000 casos      80488
8 Brasil     2000 población 174504898
9 China      1999 casos      212258
10 China     1999 población 1272915272
11 China     2000 casos      213766
12 China     2000 población 1280428583
```

tabla3

```
# A tibble: 6 × 3
  pais      anio tasa
<chr>    <dbl> <chr>
1 Afganistán 1999 745/19987071
2 Afganistán 2000 2666/20595360
3 Brasil     1999 37737/172006362
4 Brasil     2000 80488/174504898
5 China      1999 212258/1272915272
6 China      2000 213766/1280428583
```

Dividido en dos tibbles

tabla4a # casos

```
# A tibble: 3 × 3
  pais      `1999` `2000`
<chr>    <dbl> <dbl>
1 Afganistán    745    2666
2 Brasil       37737  80488
3 China       212258 213766
```

tabla4b # poblacion

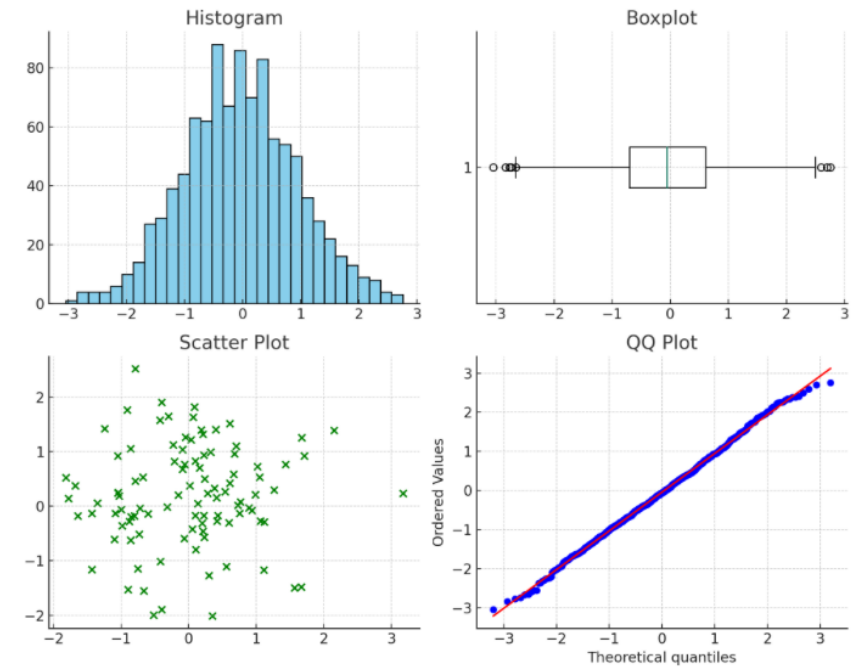
```
# A tibble: 3 × 3
  pais      `1999` `2000`
<chr>    <dbl> <dbl>
1 Afganistán 19987071 20595360
2 Brasil    172006362 174504898
3 China    1272915272 1280428583
```

Intro-to-R.R x new_metadata x

Filter

	genotype	celltype	replicate	samplemeans	age_in_days
sample1	Wt	typeA	1	10.266102	40
sample2	Wt	typeA	2	10.849759	32
sample3	Wt	typeA	3	9.452517	38
sample4	KO	typeA	1	15.833872	35
sample5	KO	typeA	2	15.590184	41
sample6	KO	typeA	3	15.551529	32
sample7	Wt	typeB	1	15.522219	34
sample8	Wt	typeB	2	13.808281	26
sample9	Wt	typeB	3	14.108399	28
sample10	KO	typeB	1	10.743292	28
sample11	KO	typeB	2	10.778318	30
sample12	KO	typeB	3	9.754733	32

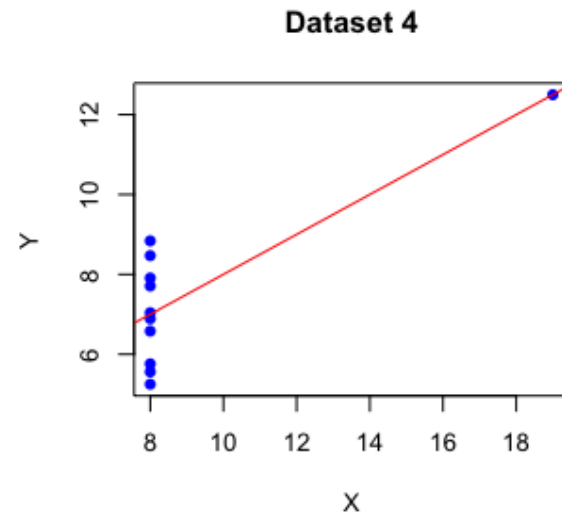
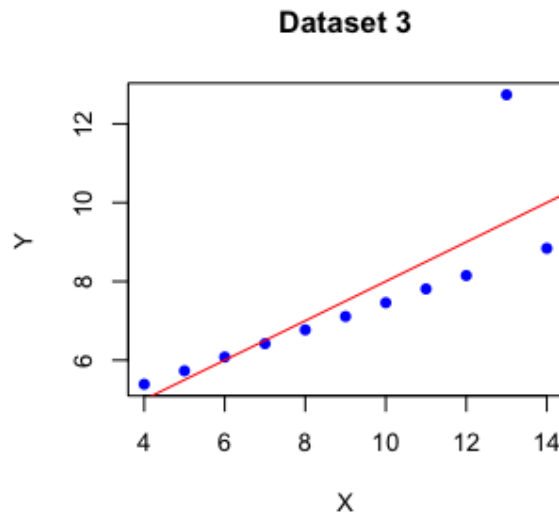
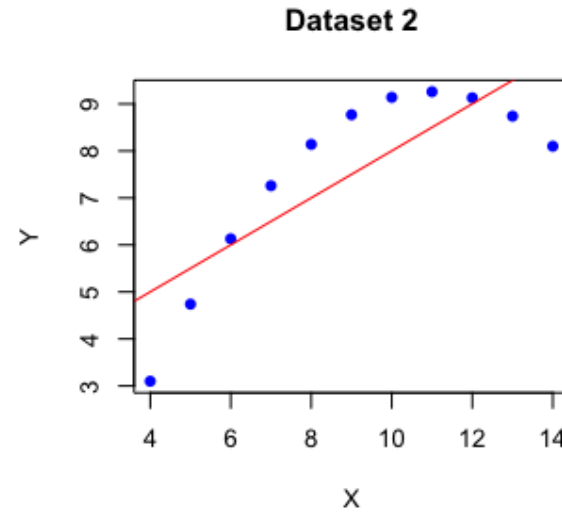
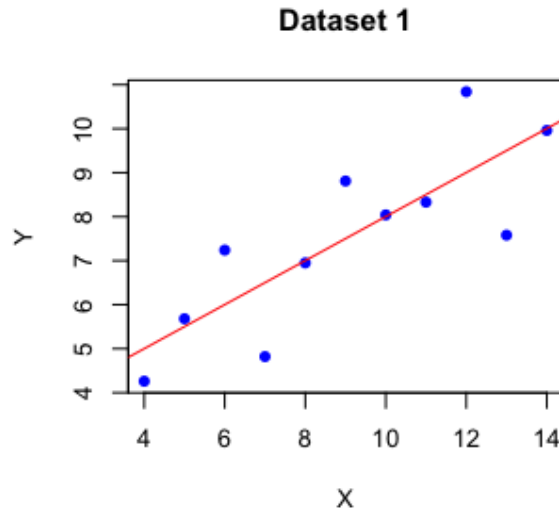
Showing 1 to 12 of 12 entries, 5 total columns



ALGUNAS PREGUNTAS PRELIMINARES → CONTEXTO

- ¿Cuántas observaciones hay?
- ¿Cuántos campos/variables se incluyen en cada observación?
- ¿Qué tipo de variables son?
- ¿Las variables se “observan siempre”? ¿Implica un problema la ausencia de datos?
- ¿Las variables que se incluyen son las esperadas? ¿Tienen sentido?
- ¿Las variables son consistentes con lo que se espera?
- ¿Qué relaciones esperaríamos que existieran entre las variables? ¿Por qué?

TÉTRADA DE ANSCOMBE



La caracterización numérica no es suficiente.

Ciencia de datos: fases estructurales

Pregunta global →

- **Análisis univariante.**
- **Análisis bivalente.**
- **Análisis “simples”.**
- **Outliers.**

Contrastar asunciones Gaussianas:

- Histogramas/density
- QQ-Plots

2º parte

- **¿Qué podemos hacer con texto/caracteres y datos no estructurados en general?**
- **Análisis complejos y modelado.**

QUICK R: TIPOS DE GRÁFICOS

Function	Object type
plot	Many
barplot	Numeric
boxplot	Formula, numeric, or list
hist	Numeric
sunflowerplot	Numeric + Numeric
mosaicplot	Formula or table
symbols	Multiple numeric

BOOK (EDA in R)

EDA: estrategia general.

1. **Evaluar el dataset de forma general.**
 1. **Nº de observaciones, variables (tipos), nombres de las variables (¿significado?)**
 2. **Tipo de variables (numérica, categórica, lógica/binaria)**
 3. **Valores únicos/valores frecuentes/observaciones faltantes**
2. **Estadística descriptiva mínima.**
3. **¿Es posible hacer alguna visualización? En especial para las variables de interés.**
4. **¿Se pueden buscar "datos anómalos"?**
5. **Resumir los resultados principales.**
6. **"Data dictionary" → Documentar el dataset y sus características, así como observaciones de interés para el "yo" futuro.**

GRACIAS

DR. ARIEL CARIAGA-MARTINEZ

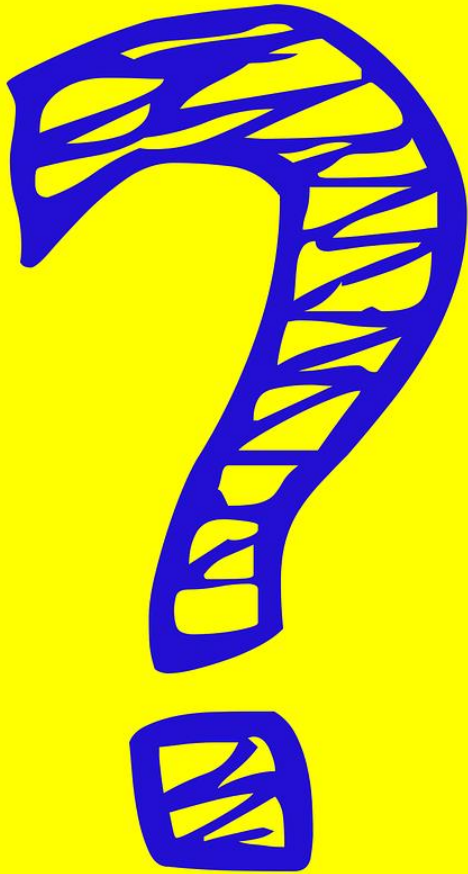
CIENCIA DE DATOS

ACARIMAR@UAX.ES

TRABAJO

GRADO EN BIOMEDICINA

DR. ARIEL CARIAGA-MARTÍNEZ



PREGUNTAS Y RESPUESTAS

- Formar equipos (2-3 personas).
- Buscar y organizar la información con respecto al análisis exploratorio de datos.
- Usar bibliografía específica y lo que hemos visto en clase.
- Organizar los conceptos teóricos en una “receta” (protocolo).
- PRESENTAR LOS RESULTADOS DE UNA FORMA VISUAL/INTERACTIVA.
- LA IDEA FINAL ES APLICARLO A UN CONJUNTO DE DATOS → PUEDES USAR EL DE TRABAJOS ANTERIORES.
- EL ENTREGABLE ES UN DOCUMENTO TIPO IMRD (INTRODUCCIÓN/MATERIALES-MÉTODOS/RESULTADOS (LA PARTE VISUAL/INTERACTIVA)/DISCUSIÓN): IDEALMENTE RMARKDOWN CON EXPLICACIÓN Y CÓDIGO.