Ciencia de datos: R

Conferencista: Felipe Calvo Cepeda
fcalvoc@unal.edu.co – fe.calvo@uniandes.edu.co

educación continua

UNIVERSIDAD NACIONAL DE COLOMBIA

1

Información

2

Fechas y horario


```
Martes, jueves: 6pm a 9pm
Sábado: 9am a 12pm
```

3

Metodología

- Clases magistrales teóricas
- Participación de las y los estudiantes
- Actividades de práctica en R
- Break intermedio

4

Evaluación

- Quices: 40%
- Taller: 30%
- Evaluación individual: 30%

5

Certificados

- La Facultad de Ciencias Económicas de la Universidad Nacional de Colombia, otorgará un certificado de asistencia y/o aprobación del programa de Educación Continua, así:

- El certificado de asistencia se otorga a los estudiantes que cumplan con mínimo el ochenta por ciento (80%) de asistencia a los mismos.
- Los certificados de aprobación se entregan únicamente a quienes, además de cumplir con el mínimo de asistencia establecida obtengan un promedio de calificación final igual o superior a tres punto cero (3.0). Los certificados de aprobación son obligatorios para los Diplomados y para los cursos correspondientes a Formación a escala.

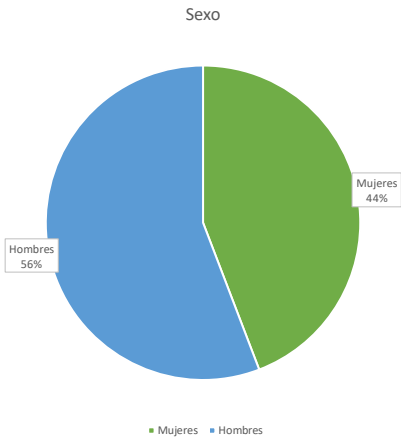
6

Temario

- Introducción, qué es R, instalación, paquetes, informes automáticos, proyectos y documentación. Carga de datos.
- Programación. Objetos y clases.
- Programación. Operaciones y funciones.
- Programación. Loops. Limpieza de datos, datos faltantes, datos atípicos, discretización de variables, trabajo con fechas y horas. Transformación de tablas de datos, crear nuevas columnas, generar resúmenes, desplegar y colapsar tablas. Operaciones entre tablas de datos. Inner join, left join., right join, full join.
- Datos univariados. Promedio, mediana, moda, varianza, cuartiles, rango intercuartílico.
- Datos multivariados. Covarianza, correlación, matriz de varianzas y covarianzas.
- Valor esperado y probabilidad condicionales.

7

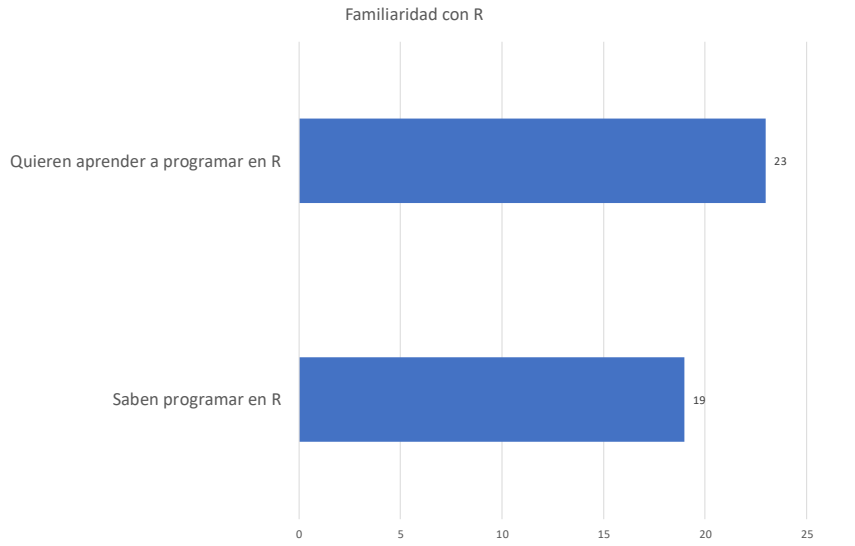
Una 'foto" de ustedes



Profesión	Conteo
Economía	18
Ingeniería	6
Administración	5
Contaduría	2
Finanzas	2
No reportan	2
Bibliotecología	1
Biología	1
Comercio Internacional	1
Estadística	1
Periodismo	1
Ciencias Políticas	1
Química	1

8

Una “foto” de ustedes



9

Una “foto” mía

10

11

Algunos retos que ustedes tienen

- Aprender a programar
- Habilidades de análisis de datos
- Investigación
- Modelado
- Visualización de datos
- Estadística
- Machine learning
- Automatización de procesos
- Ir más allá de Excel

12

Sus retos son importantes

13

"Every one of us begins life with an open mind, a driving curiosity, a sense of wonder."

Carl Sagan

14

En la ciencia de datos, los datos son la segunda cosa más importante

- Lo más importante es una pregunta
- Lo segundo más importante son los datos
- Generalmente los datos limitan o permiten las preguntas
- Sin embargo, tener datos no habilita nada si detrás no hay una pregunta
- Los métodos estadísticos no sustituyen un buen diseño de investigación

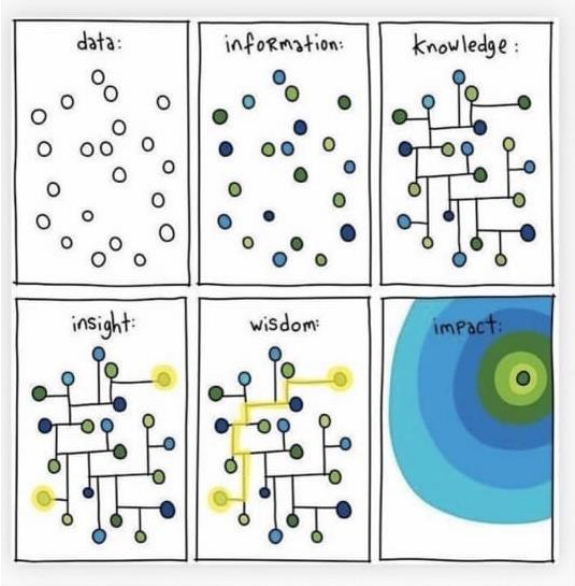
15

Preguntas del mundo hoy

- ¿Donald Trump será reelegido?
- ¿Qué podemos hacer con la epidemia de noticias falsas?
- ¿Está la democracia liberal en crisis? Si sí, ¿por qué?
- ¿Se aproxima una nueva guerra mundial?
- ¿Qué civilización domina el mundo?
- ¿Tendría Europa que abrir sus puertas a los inmigrantes?
- ¿Puede el nacionalismo resolver los problemas de desigualdad y de cambio climático?

16

Una metáfora



17

Hipótesis para algunas preguntas del mundo hoy

- ¿Donald Trump será reelegido? Sí
- ¿Qué podemos hacer con la epidemia de noticias falsas? No censurarlas
- ¿Está la democracia liberal en crisis? Si sí, ¿por qué? Sí
- ¿Se aproxima una nueva guerra mundial? No
- ¿Qué civilización domina el mundo? NS/NR
- ¿Tendría Europa que abrir sus puertas a los inmigrantes? Sí
- ¿Puede el nacionalismo resolver los problemas de desigualdad y de cambio climático? No

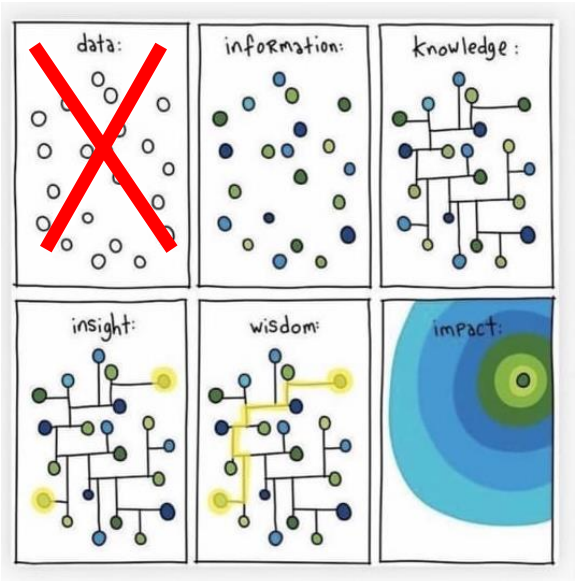
18

¿Tenemos datos para resolver esas preguntas?

- ¿Donald Trump será reelegido?
- ¿Qué podemos hacer con la epidemia de noticias falsas?
- ¿Está la democracia liberal en crisis? Si sí, ¿por qué?
- ¿Se aproxima una nueva guerra mundial?
- ¿Qué civilización domina el mundo?
- ¿Tendría Europa que abrir sus puertas a los inmigrantes?
- ¿Puede el nacionalismo resolver los problemas de desigualdad y de cambio climático?

19

Otra metáfora



20

En la ciencia de datos

- Podemos
  - aprender
  - tomar decisiones
  - presentar conclusiones

si tenemos datos.

- Pero a veces (bastantes veces) no los tenemos.

21

El camino largo

- ¡Hay que recolectarlos!
- Diseñar una estrategia de recolección

22

A, B y C

A. El 80% de la ciencia de datos tiene que ver con tener datos y tenerlos bien estructurados
B. Para el curso vamos a trabajar con algunos datos bien estructurados
C. Les invito a poner en práctica los temas vistos con datos de su propio interés profesional o académico o recreativo

23

Cómo se ven los datos

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a```^\___`_```^a``a``a^a_^___]a_]\]`a_______`_^^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaabaabbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
```

24
```

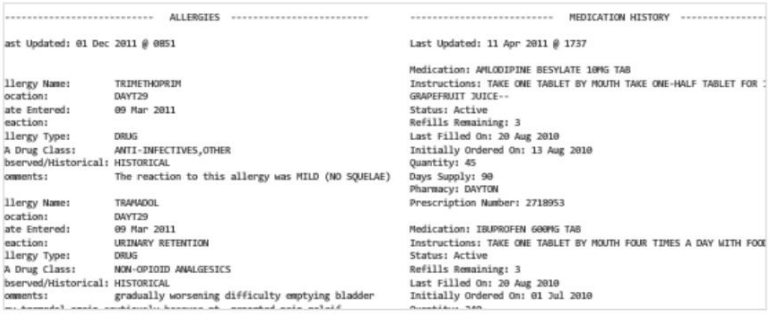
Cómo se ven los datos

Example Response


```
{
    "contributors_enabled": true,
    "created_at": "Sat May 09 17:58:22 +0000 2009",
    "default_profile": false,
    "default_profile_image": false,
    "description": "I taught your phone that thing you like.  The Mobile Partner Engineer @Twitter. ",
    "entities": {
        "description": {
            "urls": []
        }
    },
    "favourites_count": 586,
    "follow_request_sent": false,
    "followers_count": 10622,
    "following": false,
    "friends_count": 1181,
    "geo_enabled": true,
    "id": 38895958,
    "id_str": "38895958",
    "is_translator": false,
    "lang": "en",
    "listed_count": 190,
    "location": "San Francisco",
    "name": "Sean Cook",
    "notifications": false,
    "profile_background_color": "1A1B1F",
    "profile_background_image_url": "http://a0.twimg.com/profile_background_images/495742332/purty_wood.png
    "profile_background_image_url_https": "https://si0.twimg.com/profile_background_images/495742332/purty_
```


25

Cómo se ven los datos

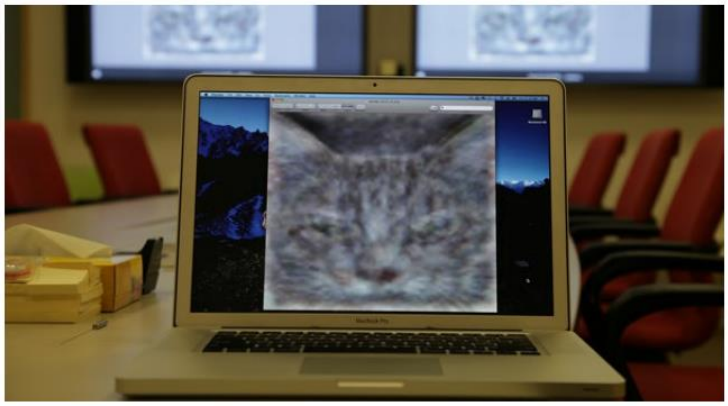


```
--------------------------  ALLERGIES  --------------------------      --------------------------  MEDICATION HISTORY  ----------
ast Updated: 01 Dec 2011 @ 0851                                       Last Updated: 11 Apr 2011 @ 1737

                                                                      Medication: AMLODIPINE BESYLATE 10MG TAB
llergy Name:        TRIMETHOPRIM                                      Instructions: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR
ocation:            DAYT29                                            GRAPEFRUIT JUICE--
ate Entered:        09 Mar 2011                                       Status: Active
eaction:                                                              Refills Remaining: 3
llergy Type:        DRUG                                              Last Filled On: 20 Aug 2010
A Drug Class:       ANTI-INFECTIVES,OTHER                             Initially Ordered On: 13 Aug 2010
bserved/Historical: HISTORICAL                                        Quantity: 45
omments:            The reaction to this allergy was MILD (NO SQUELAE)  Days Supply: 90
                                                                      Pharmacy: DAYTON
llergy Name:        TRAMADOL                                          Prescription Number: 2718953
ocation:            DAYT29
ate Entered:        09 Mar 2011                                       Medication: IBUPROFEN 600MG TAB
eaction:            URINARY RETENTION                                 Instructions: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOO
llergy Type:        DRUG                                              Status: Active
A Drug Class:       NON-OPIOID ANALGESICS                             Refills Remaining: 3
bserved/Historical: HISTORICAL                                        Last Filled On: 20 Aug 2010
omments:            gradually worsening difficulty emptying bladder   Initially Ordered On: 01 Jul 2010
```

26

Cómo se ven los datos

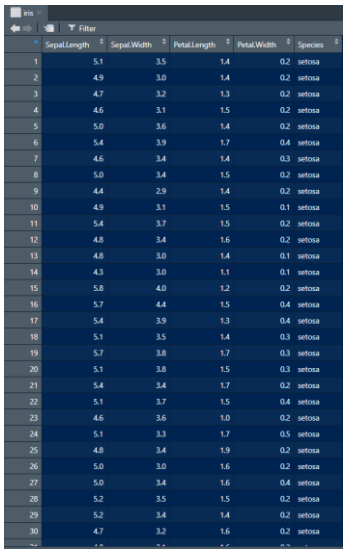


27

Cómo se ven los datos

28

Cómo se ven los datos *(rara vez)*



	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
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa
23	4.6	3.6	1.0	0.2	setosa
24	5.1	3.3	1.7	0.5	setosa
25	4.8	3.4	1.9	0.2	setosa
26	5.0	3.0	1.6	0.2	setosa
27	5.0	3.4	1.6	0.4	setosa
28	5.2	3.5	1.5	0.2	setosa
29	5.2	3.4	1.4	0.2	setosa
30	4.7	3.2	1.6	0.2	setosa

29

Roles en proyectos con datos

- Ingeniería de datos
  - Obtener los datos
  - Limpiarlos y estructurarlos para posterior análisis
  - Crear pipelines de análisis automatizado
  - Utilización de herramientas en la nube
  - Análisis descriptivo de los datos
- Ciencia de datos
  - Análisis matemático de los datos
  - Identificación de variables relevantes / features
  - Generación de modelos predictivos y prescriptivos
- Profesionales de modelado ML
  - Creación de sistemas predictivos y prescriptivos de gran escala
  - Mantenimiento y ajuste del modelo

30

Practica

- Configuración del ambiente de trabajo
- Programación básica

31