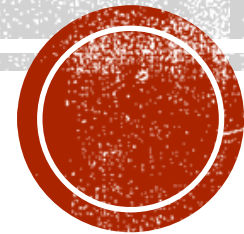


INTRODUCCIÓN A R

Christian Camilo Urcuqui López, MSc

Github: urcuqui

<https://github.com/urcuqui>



PRESENTACIÓN

Christian Camilo Urcuqui López

Ing. Sistemas, Magister en Informática y Telecomunicaciones

Big Data Professional

Big Data Scientist

Deep Learning Specialization

Grupo de investigación i2t

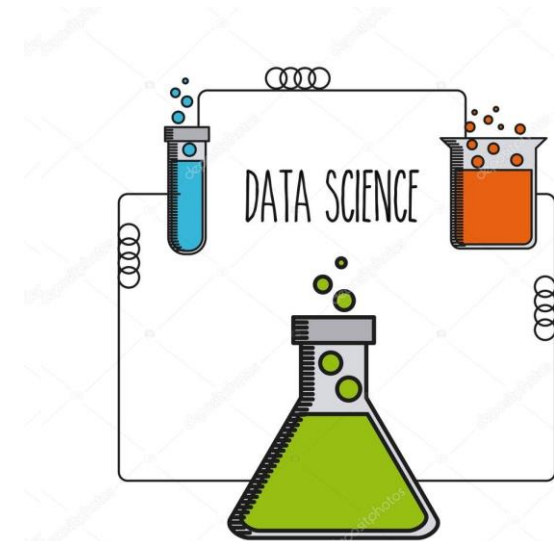
Líder de investigación y desarrollo

Ciberseguridad y ciencia de datos aplicada

ccurcuqui@icesi.edu.co

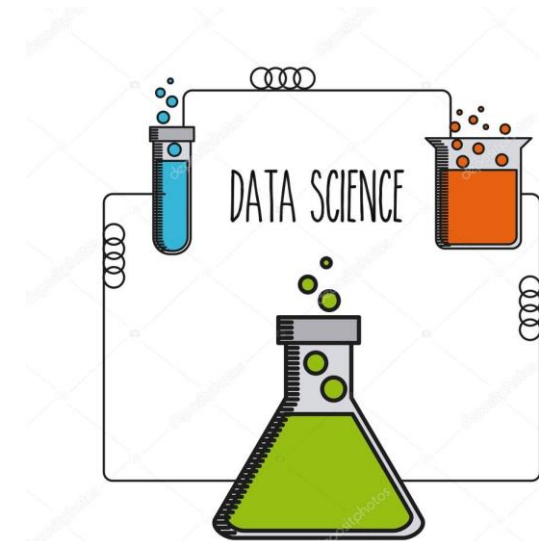
COMPETENCIAS

- Describir el lenguaje de programación R y su aplicación en proyectos de ciencia de datos
- Aplicar los conceptos básicos de codificación en R
- Explorar un *data.frame* en R



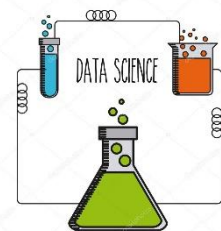
CONTENIDO

1. Introducción al lenguaje de programación R
2. Flujo de trabajo en R - Nombres de objetos, llamada a funciones
3. Explorando las estructuras de datos en R
4. Importando y trabajando con datos
5. Resolución de dudas

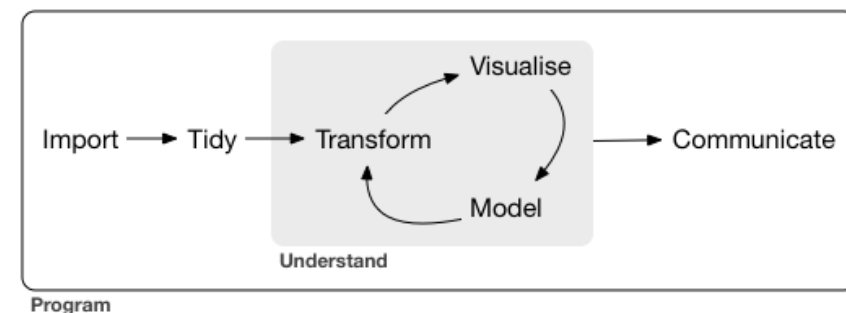
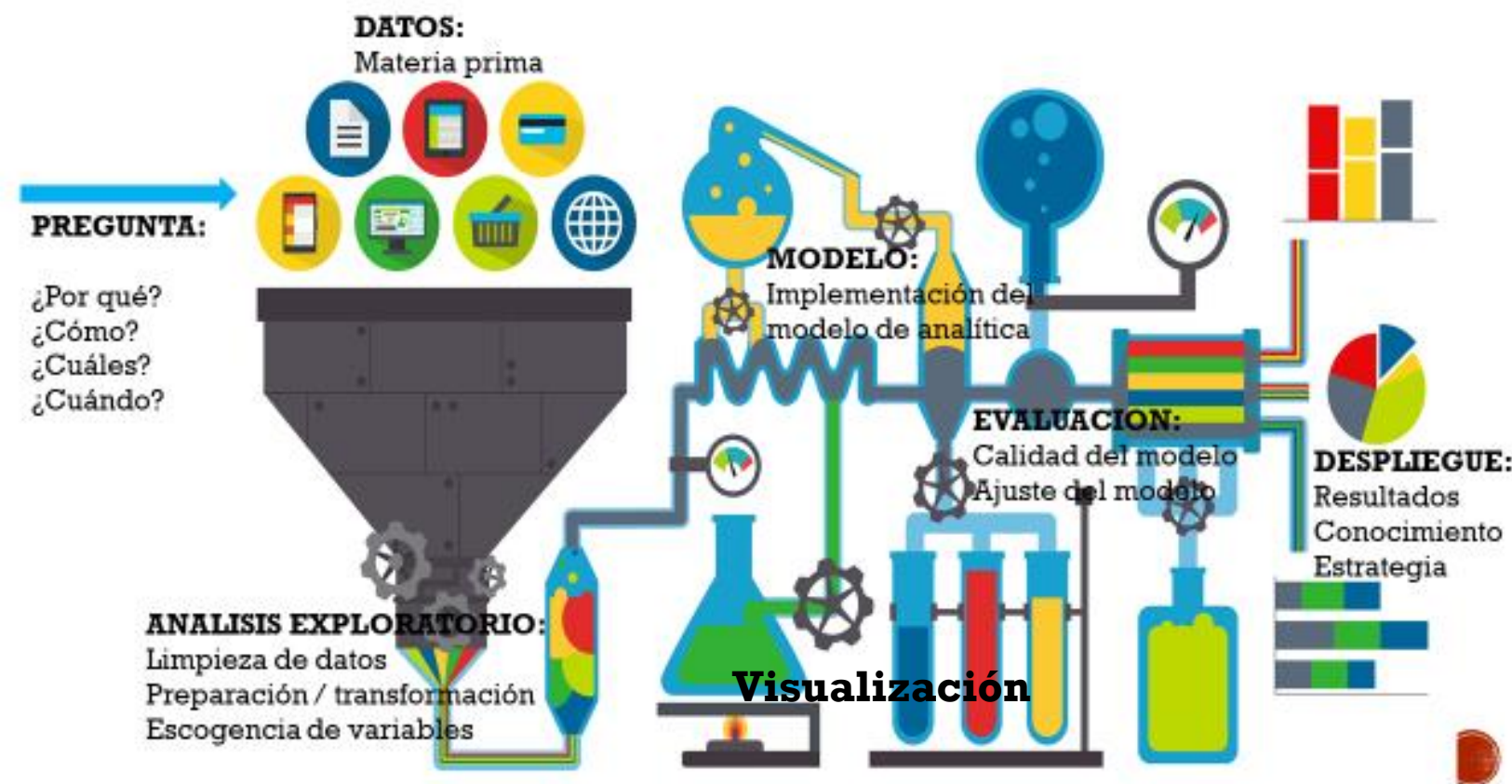


INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

- La forma de analizar los datos ha estado cambiando durante los últimos años.
- “Los datos son el petróleo del siglo xxi”.
- La ciencia de datos a partir de sus técnicas (por ejemplo, estadísticas, visuales, econométricas y de aprendizaje de máquina) han permitido descubrir y explotar la información.
- Antiguamente, los investigadores solían publicar sus resultados en prestigiosas revistas y la implementación de sus descubrimientos en software tomaba mucho tiempo.
- Actualmente, los investigadores y la industria han mejorado sus métodos en conjunto con la implementación de software, estos resultados ahora se encuentran en sitios web de fácil acceso (en muchos casos con licencia *open source*).

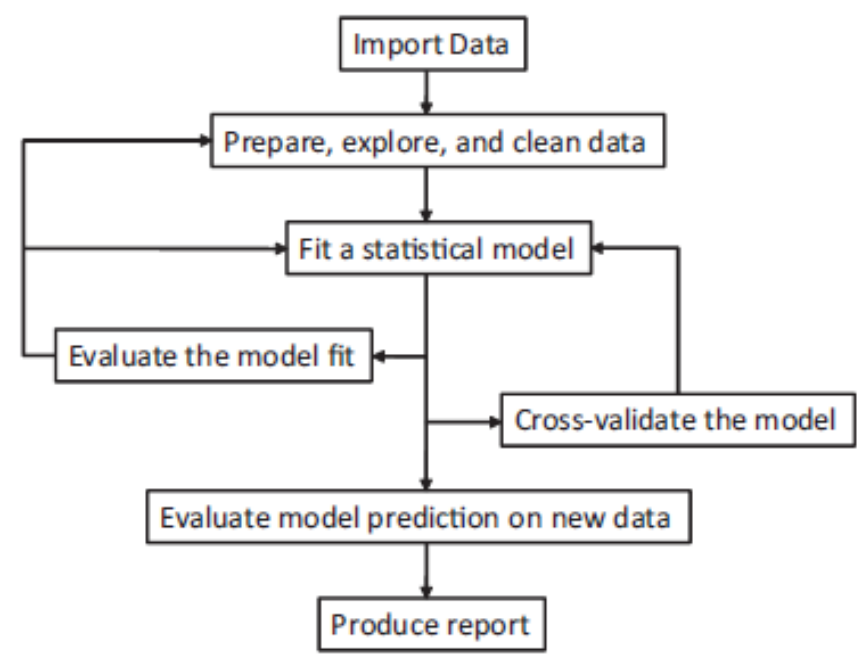
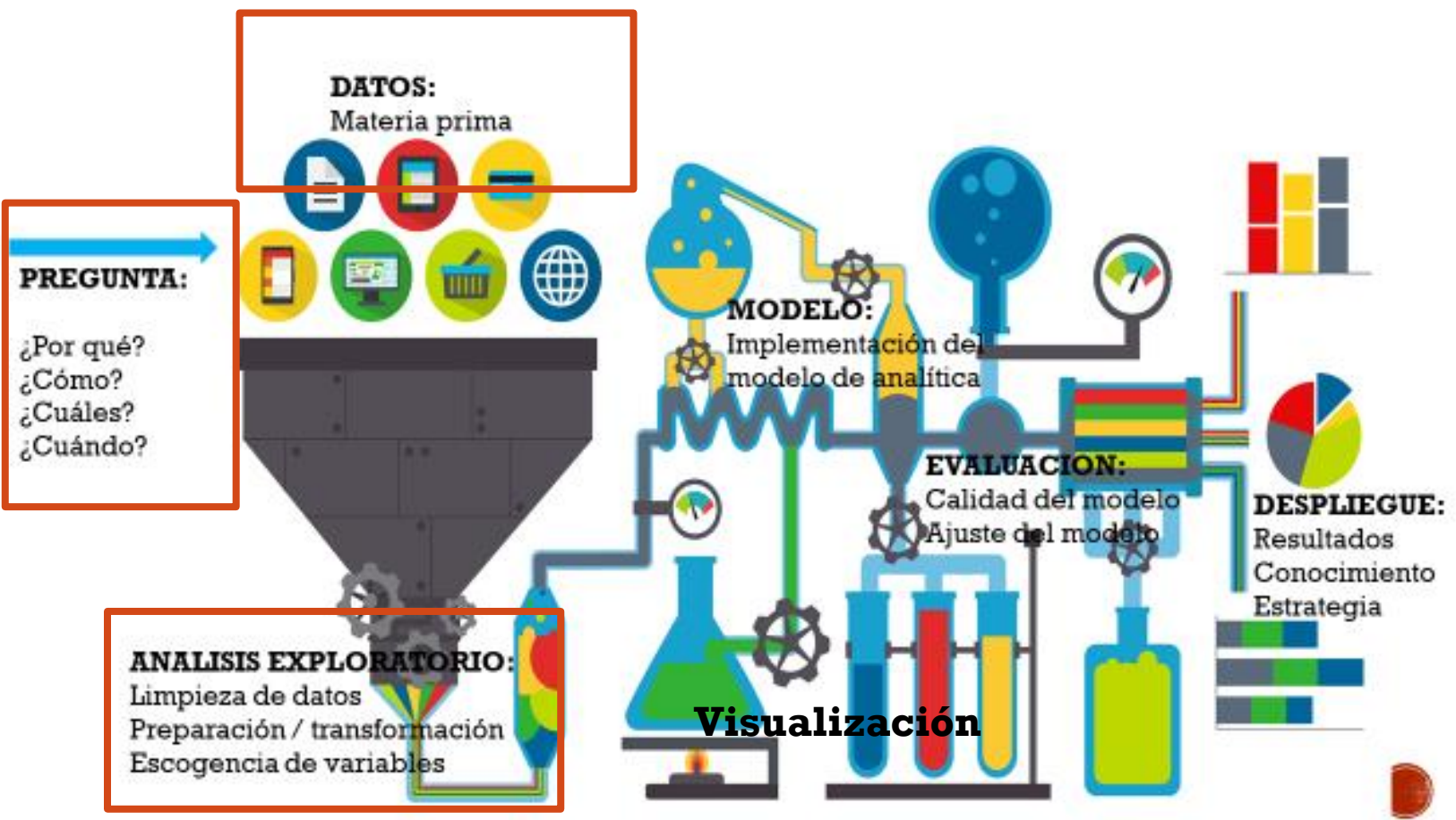


RECORDEMOS



Marco de trabajo típico de un proyecto de ciencia de datos.
R for Data Science

RECORDEMOS

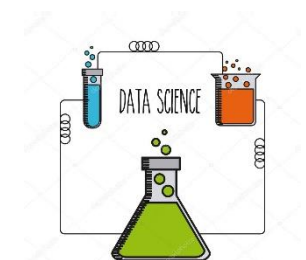


Pasos típicos de un proyecto de ciencia de datos.
R IN ACTION: Data analysis and graphics with R



INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

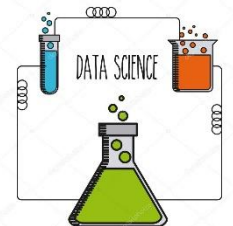
- R es un entorno y un lenguaje para computación estadística y gráfica desarrollado en Bell Labs, un proyecto que nace del software libre S.
- Una solución *open source* para análisis de datos soportado por varias comunidades científicas en todo el mundo.
- ¿Por qué utilizar R en vez de otras soluciones populares para estadística y gráficos (por ejemplo, Microsoft Excel, SAS, IBM SPSS, Stata, y Minitab)?



INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

Excel

- **Cálculo básicos.** Excel provee una interfaz (entorno) más amigable para cálculos simples (por ejemplo, estadísticas descriptivas) o algunas manipulaciones sencillas (por ejemplo, filtros y búsquedas).
- **Ver los datos continuamente.** Excel es una herramienta que nos permite constantemente ver la estructura y el contenido de los datos.
- **Presentación de datos y resumen.** Excel nos da un contenido estético más agradable de las hojas de cálculo.
- **Menor curva de aprendizaje.** Requiere un menos tiempo para llegar a manejar gran parte de sus funcionalidades. n



INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

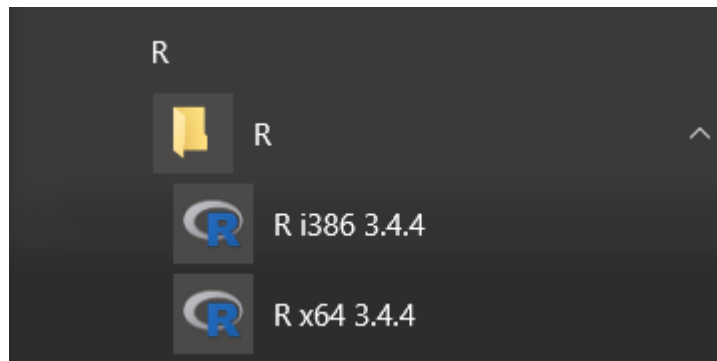
R

- Es una tecnología gratuita a diferencia de otros software comerciales de altos costos.
- Tiene comunidades muy activas, casi semanalmente se proponen nuevos paquetes estadísticos y actualizaciones, lo cual a llevado a los sistemas comerciales a integrar R.
- Es más fácil la automatización ya que se pueden desarrollar *scripts* (líneas de código con un propósito específico) que permiten ejecutar el análisis varias veces.
- Leer casi cualquier tipo de datos (.txt, .csv, .dat), también, existen paquetes que permiten leer información de archivos JSON, Excel, STATA, SAS. E incluso utilizar datos de sitios web y de sistemas de base de datos (Por ejemplo, MySQL y, PostgreSQL)

INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

Nuestro primer hello world

- Una vez instalado R en nuestro equipo, nosotros podemos acceder a la consola de R para que podamos escribir los código. Procedamos ha abrir una consola y digitemos el comando **print("hello world")**



```

RGui (64-bit)
Archivo  Editar  Visualizar  Misc  Paquetes  Ventanas  Ayuda

R Console

R version 3.4.4 (2018-03-15) -- "Someone to Lean On"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

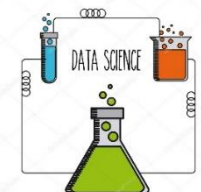
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.

R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.

Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.

[Previously saved workspace restored]

> print("hello world")
hello world
  
```



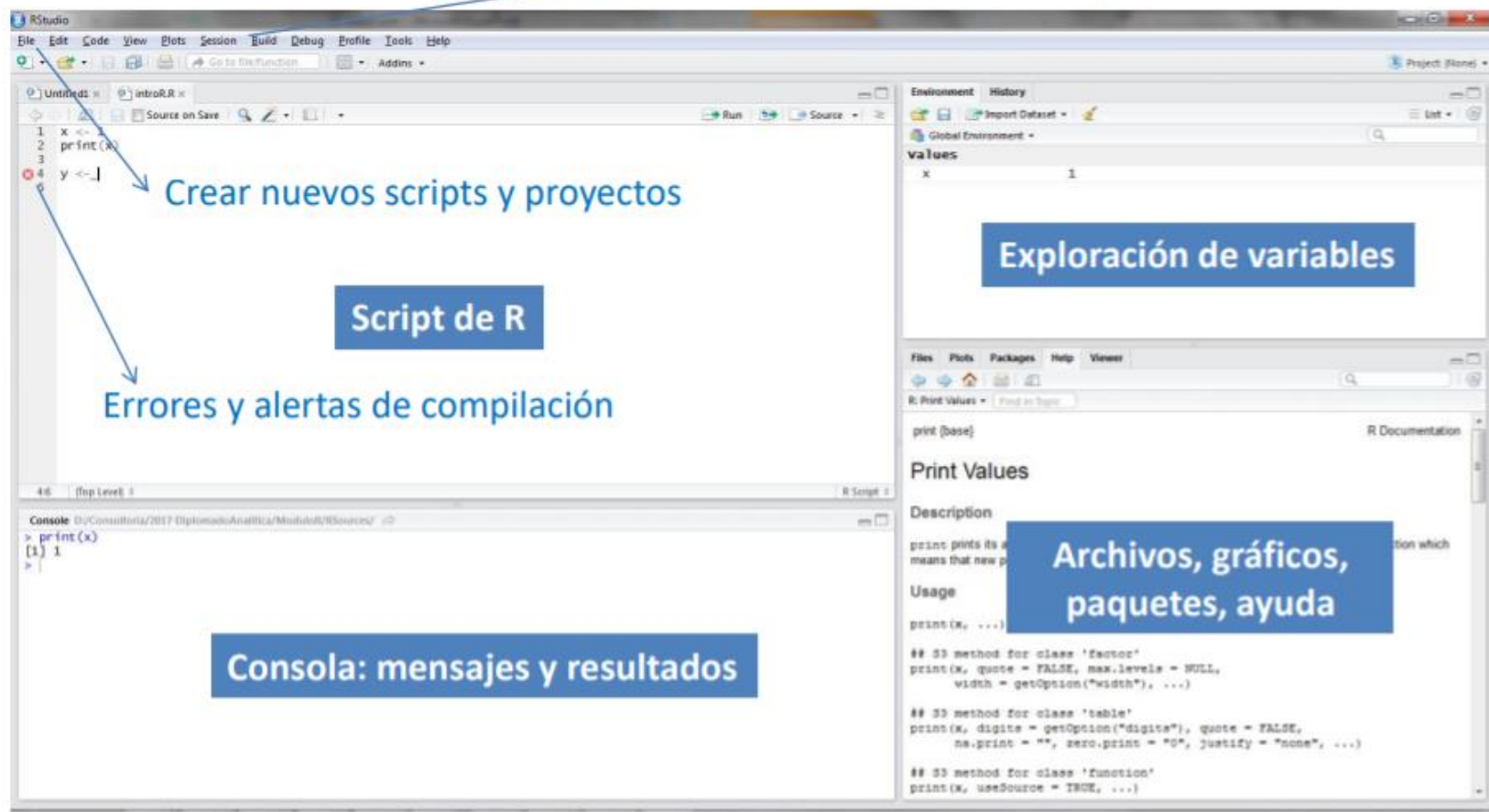
INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

- Existen ambientes de programación más amigables y con una serie de utilidades que nos facilitan el desarrollo de proyectos de software, para nuestro caso utilizaremos RStudio.
- RStudio puede descargarse de la siguiente página web:
<https://www.rstudio.com/products/rstudio/download/>
- RStudio nos facilita el trabajo con R
 - Editor de código
 - Depurador (permite probar y depurar errores en tiempo de ejecución)
 - Herramientas de visualización
- Algunas de sus versiones son de uso libre, otras son licenciadas.



INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

Definir directorio de trabajo



INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

- R es un lenguaje de programación interpretado, es decir, el código será ejecutado instrucción por instrucción.
- Muchos de los datos y variables son almacenadas en memoria durante una sesión. Nosotros podemos guardar una sesión con la finalidad de conservar nuestro trabajo para futuras sesiones.
- R utiliza el símbolo `<-` para detonar una asignación, a diferencia del típico `=` utilizado en muchos otros lenguajes de programación.
- ¡Veamos el ejemplo en el notebook!

INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

- Como hemos mencionado, la aplicación de la función `rnorm` nos dio un vector de cinco valores, ahora si deseamos crear un vector y asignarlo a una variable, debemos proceder a digitar la función `_c_` seguido de los valores que queremos tener.
- ¡Veamos el ejemplo en el notebook!
- Ahora, si deseamos obtener mayor información sobre una función podemos utilizar la opción de ayuda con R a través de `?`,
- ¡Veamos el ejemplo en el notebook!

INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

- Los comentarios en el código podemos agregarlos con #
- Procedamos a un tercer ejercicio, vamos a crear una tabla a través de la asignación de vectores, utilizaremos unas funciones predefinidas en R para estadística descriptiva y vamos a crear un gráfico (conocidos como *-plot*).

Age (mo.)	Weight (kg.)	Age (mo.)	Weight (kg.)
01	4.4	09	7.3
03	5.3	03	6.0
05	7.2	09	10.4
02	5.2	12	10.2
11	8.5	03	6.1

INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

- Existen numerosas funciones de R en distintos paquetes que nos pueden facilitar la vida. La instalación de un paquete se realiza a través de:

`install.packages("nombre del paquete")`

- Como hemos mencionado, constantemente las comunidades lanzan nuevas versiones de sus paquetes ya sea para incorporar nuevas funcionalidades o para corregir algún error; si deseamos estar en la última versión podemos utilizar el siguiente comando:

`update.packages()`

- Para cargar el paquete que necesitamos utilizar en nuestro entorno de trabajo debe utilizar el siguiente comando en conjunto con el nombre del paquete

`library(gclus)`

INTRODUCCIÓN AL LENGUAJE DE PROGRAMACIÓN R

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Introduccion_R.Rmd

Knit Insert Run

AA	1141	N619AA	JFK	MIA	160	1089	5	40
AA	301	N3ALAA	LGA	ORD	138	733	6	0
AA	707	N3DUAA	LGA	DFW	257	1389	6	0
AA	1895	N633AA	EWB	MIA	152	1085	6	10
AA	1837	N3EMAA	LGA	MIA	153	1096	6	10
AA	413	N3BAAA	JFK	SJU	192	1598	6	30
AA	303	N3CYAA	LGA	ORD	140	733	6	30
AA	711	N3GKAA	LGA	DFW	248	1389	6	35
AA	305	N4WNAA	LGA	ORD	143	733	7	0
AA	1815	N5FMAA	JFK	MCO	142	944	6	59

1-10 of 32,729 rows | 10-18 of 19 columns Previous 1 2 3 4 5 6 ... 100 Next

```

301
302 ##cargar archivos y bases de datos
303
304 R tienen la capacidad de cargar conjuntos de información en distintos formatos a nuestro entorno de
305 trabajo. En el siguiente ejemplo veremos distintas formas de cargar archivos .csv y de otro tipo a
306 través de comandos y del mismo Rstudio, es muy importante conocer previamente el formato del archivo
307 antes de cargarlo (por ejemplo, si los datos vienen separados por ";", ",", " " y " ").
308
309 _cargar archivos con Rstudio_
310
311 En la sección de "environment" podemos encontrar una opción "import Dataset" que nos abrirá una
312 segunda ventana para cargar los archivos
313
314 Cargar archivos y bases de datos
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

Environment History Connections

Global Environment

Data

200 obs. of 4 variables

storiesdb Formal class MySQLConnection

values

age	num [1:10]	1 3 5 2 11 9 3 9 12 3
localuserpassword		"SomethingDifficult"
resultado_diferencia	logi [1:6]	TRUE TRUE FALSE TRUE TRUE TRUE
resultado_igualdad	logi [1:6]	FALSE TRUE FALSE FALSE FALSE TRUE
resultado_multiplicacion	num [1:6]	4 8 16 28 36 0
resultado_negativo	num [1:6]	-0.5 -2 -3 -5 -1 0
resultado_raiz		2
resultado_suma	num [1:6]	5 6 8 11 13 4
user_password		"SomethingDifficult"
vector	int [1:10]	1 2 3 4 5 6 7 8 9 10
weight	num [1:10]	4.4 5.3 7.2 5.2 8.5 7.3 6 10.4 10.2 6.1
x	num [1:6]	1 2 4 7 9 0
y		4
z	num [1:6]	0.5 2 3 5 1 0

Files Plots Packages Help Viewer

R: Flights data

Find in Topic

Date of departure

dep_time, arr_time

Actual departure and arrival times (format HHMM or HMM), local tz.

sched_dep_time, sched_arr_time

Scheduled departure and arrival times (format HHMM or HMM), local tz.

dep_delay, arr_delay

Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.

hour, minute

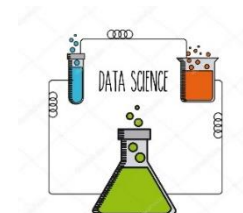
Time of scheduled departure broken into hour and minutes.

carrier

Two letter carrier abbreviation. See [airlines\(\)](#) to get name

tailnum

Plane tail number



REFERENCIAS

- Kabacoff, R. (2015). R IN ACTION: Data analysis and graphics with R.
- Wickham, H., & Grolemund, G. (2016). R for Data Science.