

Ciencia de datos para Políticas Públicas

Universidad Diego Portales
Pablo Javier Aguirre Hörmann

Sobre mí...

- Consultor Data Science/Business Intelligence – Superintendencia del Medio Ambiente
- Profesor Universitario (UAI, PUC)
- Magíster en Políticas Públicas – Universidad de Chicago
- Ingeniero Agrónomo - PUC



[@pjaguirreh](https://github.com/pjaguirreh)



[@pjaguirreh](https://www.linkedin.com/in/pjaguirreh)



[@PAguirreH](https://twitter.com/PAguirreH)

Ahora ustedes...

Sobre este curso...

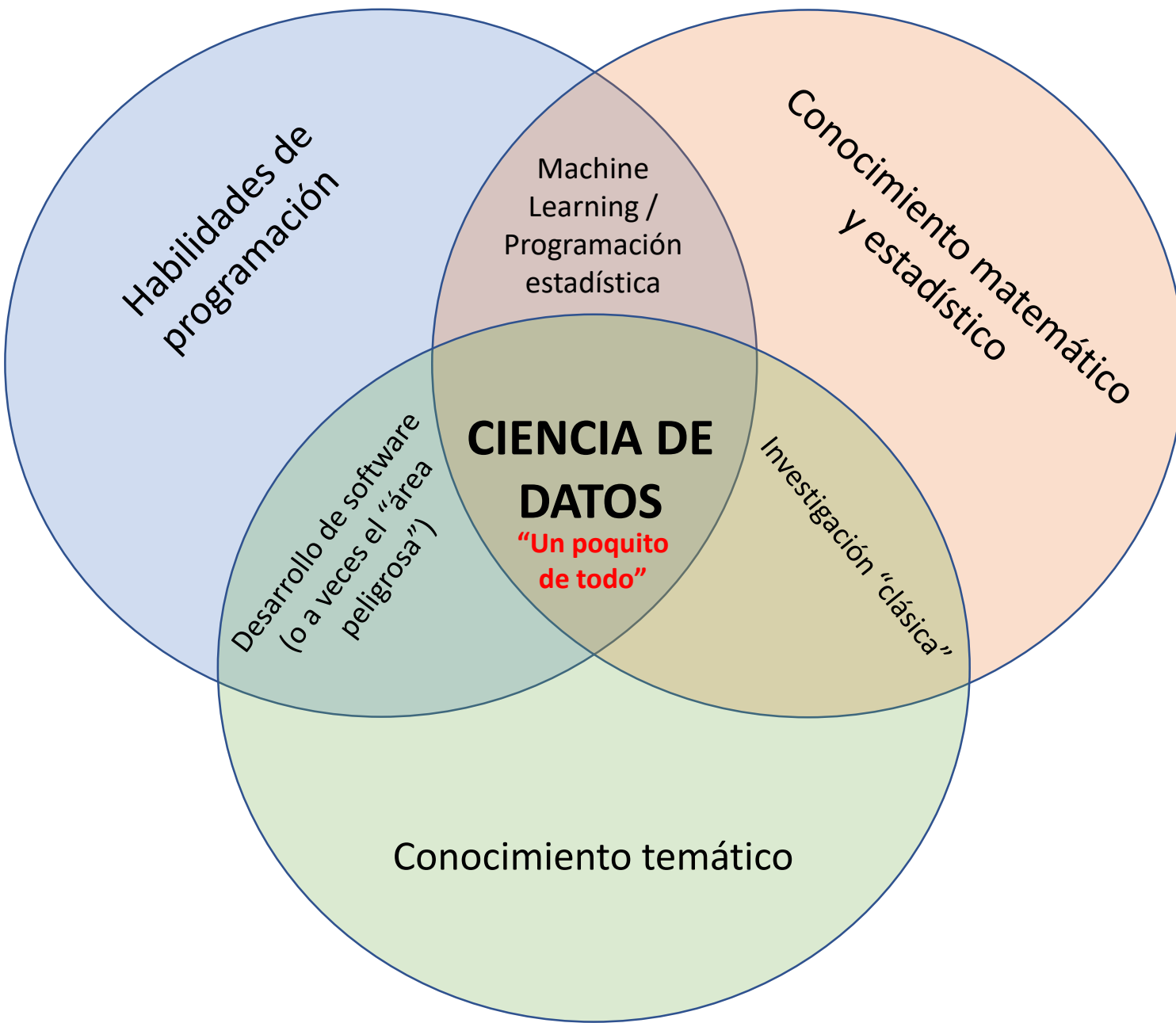
- **Objetivos:** El fin de este curso es hacer de los estudiantes mejores productores y consumidores de herramientas relacionadas a lo que conoce como *“big data”*, *“data science”*, y *“machine learning”* en el ámbito de problemas vinculados a las políticas públicas.
- **Metodología:** Clases expositivas y demostraciones prácticas. La primera parte del curso se concentrará en enseñar a utilizar el lenguaje y ambiente de programación R mientras que el resto del curso se revisarán distintas herramientas relacionadas a *“big data”*, *“data science”*, y/o *“machine learning”* y cómo implementar estas a través de R.

- Big Data:
“el trabajo duro y sucio”
 - Volumen: planilla con MUCHÍSIMAS filas y MUCHÍSIMAS columnas lo cual dificulta su manejo con softwares tradicionales (por ej. Excel)
 - Velocidad: la planilla suma rápidamente (por ej. cada segundo) nuevas filas con información y, por ende, hace más lento cualquier cálculo/manejo que se quieran hacer
 - Variedad: la planilla no solo suma filas nuevas si no que columnas particulares para algunas de estas... también de vez en cuando agrega una foto, o un video... o ambas. Lo anterior dificulta como manejar la información.



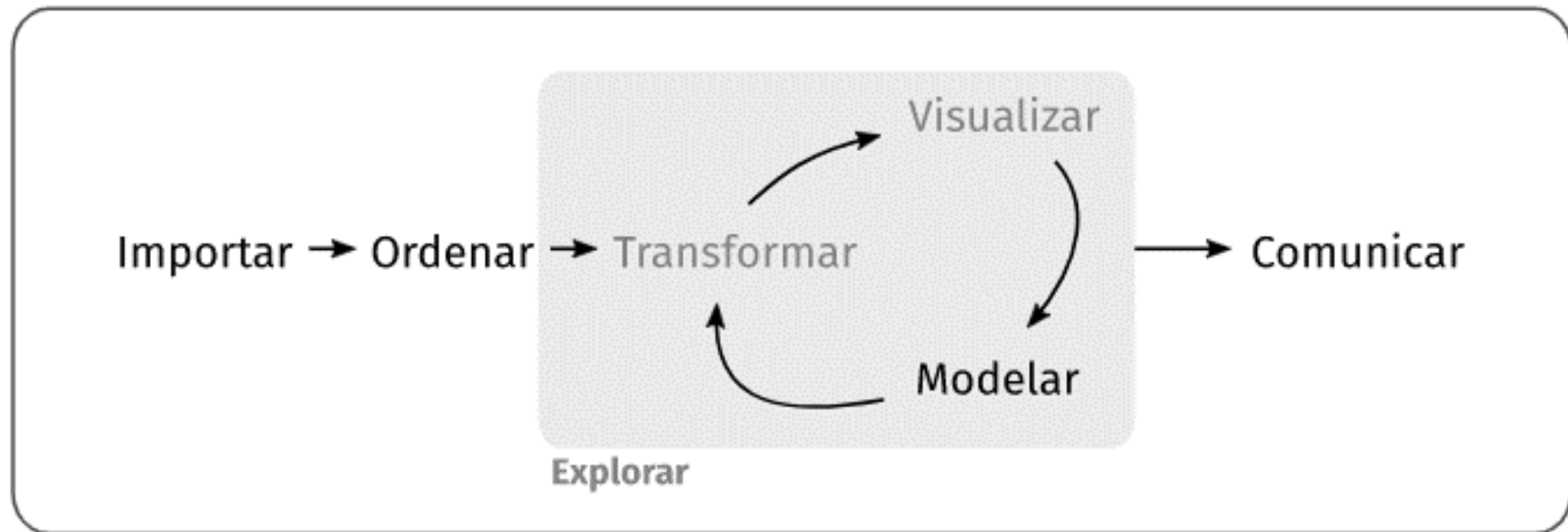
- Machine Learning:
“la parte más estadística”
 - Predicción!
 - Ejemplos:
 - Netflix recomendándote una serie o película a ver
 - Filtro del correo electrónico (¿es spam o no?)
 - Not Mayo (¿?)
 - Modelos (regresiones, árboles, redes neuronales, etc), datos de entrenamiento, datos de prueba, datos de validación.





Los/as científicos/as de datos (“data scientists”) son profesionales con las habilidades para “bucear” a través de grandes volúmenes de información, encontrar patrones, y extraer hallazgos que permitan entender mejor un problema, proponer soluciones y/u optimizar algún proceso

El proceso de “data science” y lo que veremos en este curso



El proceso de "data science" y lo que veremos en este curso



En específico...

Semana	Fecha	Contenidos	Lectura previa	Evaluación
Parte I: Introducción a la programación / Manejo de datos				
1	30/10	Descripción del curso e introducción al uso de datos para políticas públicas Introducción a R: R y RStudio	Shmueli, 2010 ISL: 2.1 laR: 2 ADP: 2.1 y 2.2	
2	06/11	Introducción a R: Tipo de datos y sintaxis	laR: 2	
3	13/11	Manejo de datos 1	R4DS: 12	Tarea 1
4	20/11	Manejo de datos 2	R4DS: 5	
5	27/11	Visualización de datos	laR: 12 R4DS: 3	Tarea 2
Parte II: Modelos/Machine Learning				
6	04/12	Regresión Lineal y logística	ISL: 3.1 a 3.3; 4.1 a 4.3 ADP: 6.1 a 6.4; 7.1 y 7.2	Tarea 3
7	11/12	Dilema varianza sesgo y Métodos de remuestreo	ISL: 2.2; 5.1 y 5.2	Informe preliminar
8	18/12	Regularización de modelos lineales Regresión “stepwise” Análisis de componentes principales	ISL: 6.1 a 6.4 y 10.2 ADP: 10	Tarea 4
9	08/01	Árboles de decision, Bagging y Random Forest	ISL: 8.1 y 8.2	Tarea 5
10	13/01	Presentación trabajo		Informe Final Presentación

¿Qué es ?

Lenguaje y plataforma

- Lenguaje de programación estadística
- Herramienta de visualización de datos
- Gratuito

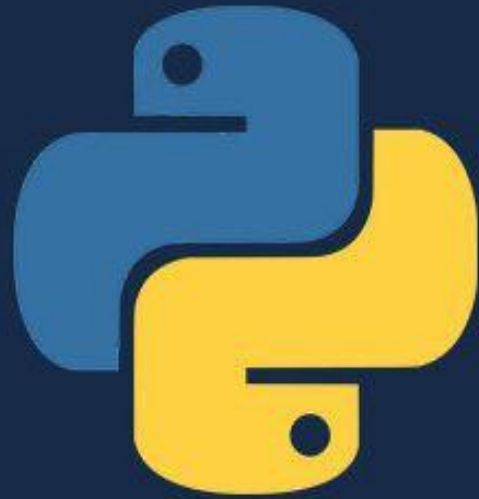
Ecosistema

- Muchas aplicaciones e integraciones con otras plataformas
- 12.000+ librerías gratuitas disponibles

Comunidad

- 2.5+ millones de usuarios
- Muchos y diversos grupos de usuarios a nivel mundial





R Vs Python: What's the Difference?

R VS PYTHON 2018 ¿CUÁL ES EL MEJOR?

Publicado por: Rosana Ferrero

Categoría: Data Science +R

No hay comentarios



R y Python son dos de los lenguajes favoritos de los Cientistas.

Preguntas relacionadas

Is Python or R better for data science?

Should I learn Python or R first?

Is R similar to Python?

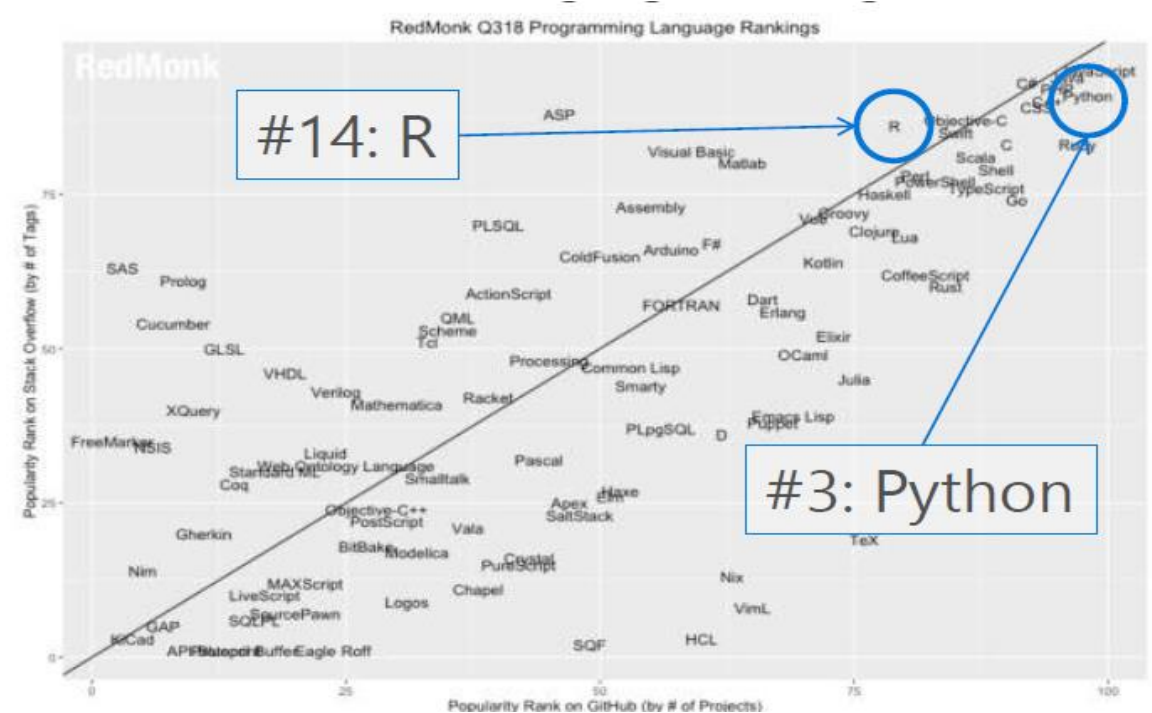


¿Qué dicen los usuarios?

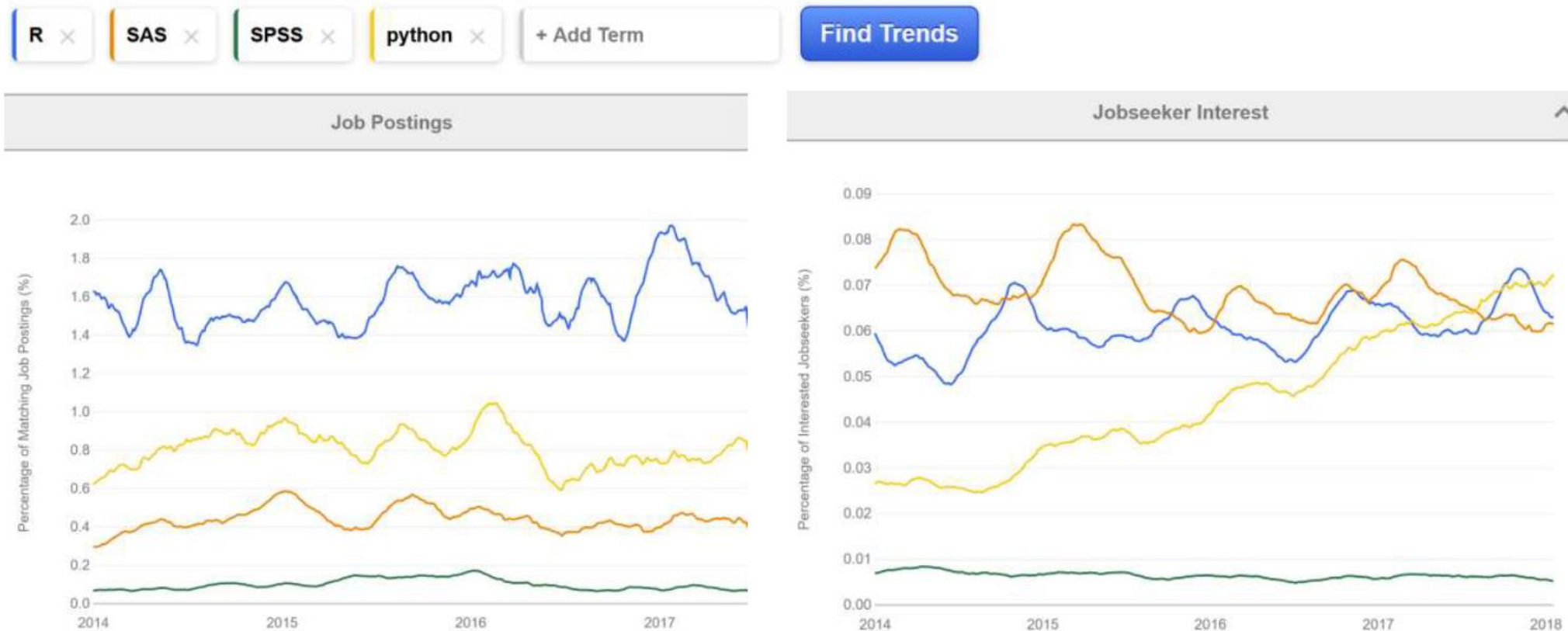
[IEEE Spectrum, Julio 2018](#)

Language Rank	Types	Spectrum Ranking
1. Python	  	100.0
2. C++	  	99.7
3. Java	  	97.5
4. C	  	96.7
5. C#	  	89.4
6. PHP		84.9
7. R		82.9
8. JavaScript	 	82.6
9. Go	 	76.4
10. Assembly		74.1

[Redmonk, Junio 2018](#)



¿Qué dicen los empleadores?



<https://www.indeed.cl/?r=us>

The reality is that learning both tools and using them for their respective strengths can only improve you as a data scientist. Versatility and flexibility are traits any data scientist at the top of their field. The Python vs R debate confines you to one programming language. **You should look beyond it and embrace both tools for their respective strengths. Using more tools will only make you better as a data scientist.**

[Medium, Junio 2018](#)

From 'R vs Python' to 'R and Python'

Leveraging the best of both 'Python and R' in a single project.

[Medium, Marzo 2019](#)

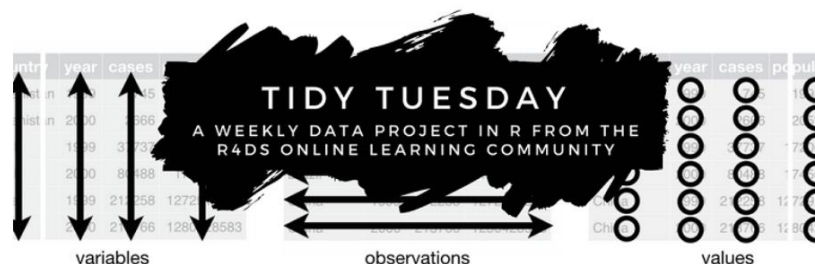
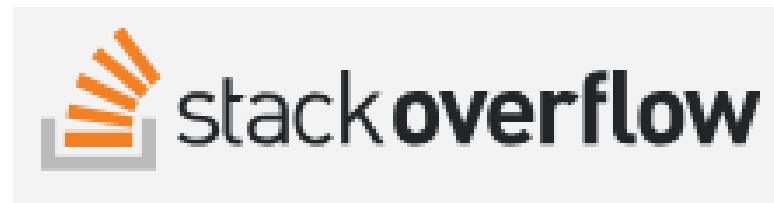
Is SQL Needed to be a Data Scientist?

Is SQL needed to be a Data Scientist? the answer is Yes, SQL (Structured Query Language) is Needed for Data Scientists to get the data and to work with that data. Everyone is busy to Learn R or Python for Data Science, but without Database Data Science is

[Datacamp, 2018](#)



“La comunidad de R”



R-statistics blog

Statistics with R, and open source stuff (software, data, community)



¿Qué es Studio?

IDE para R: *Entorno de desarrollo integrado*

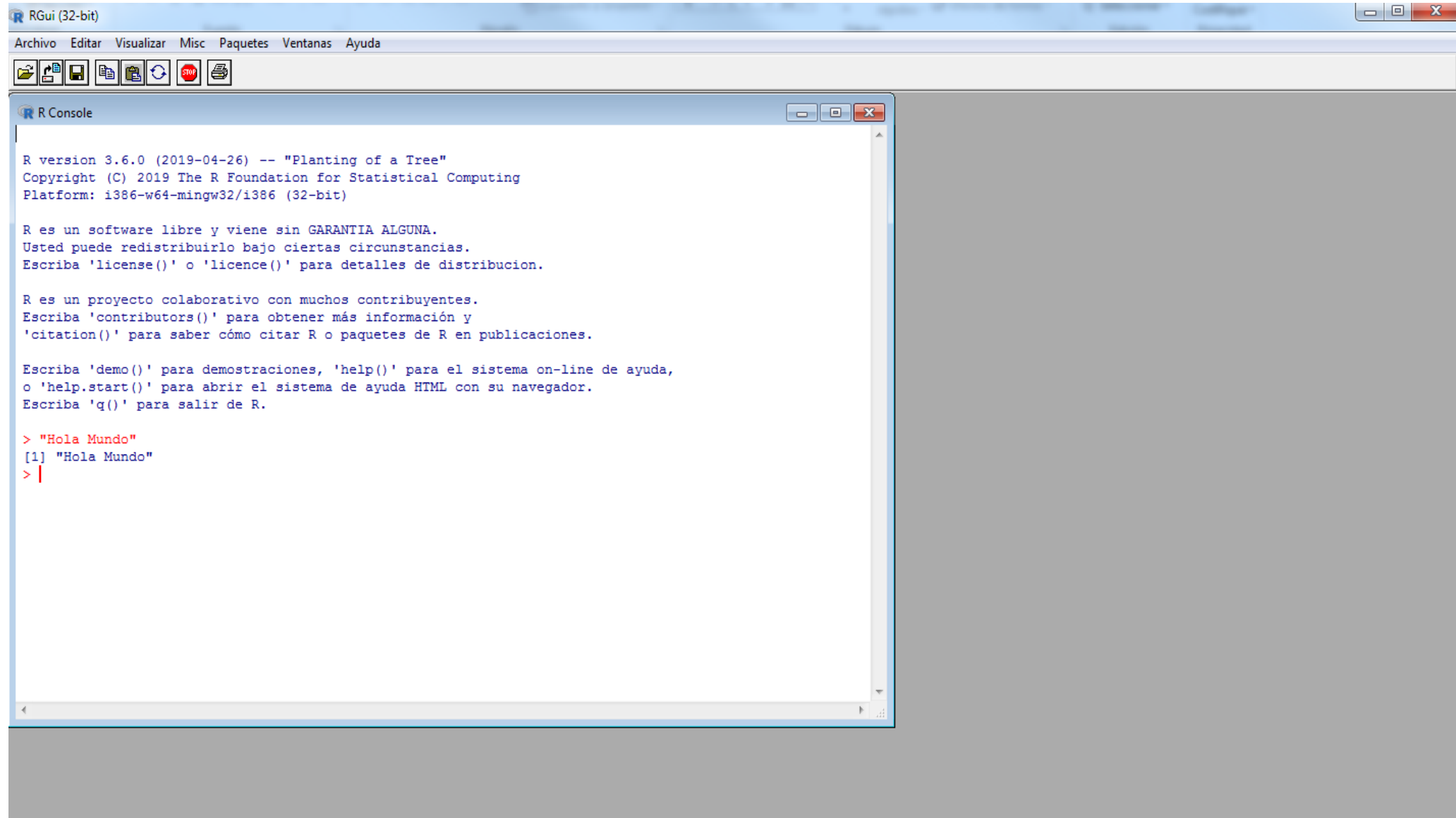
Consola

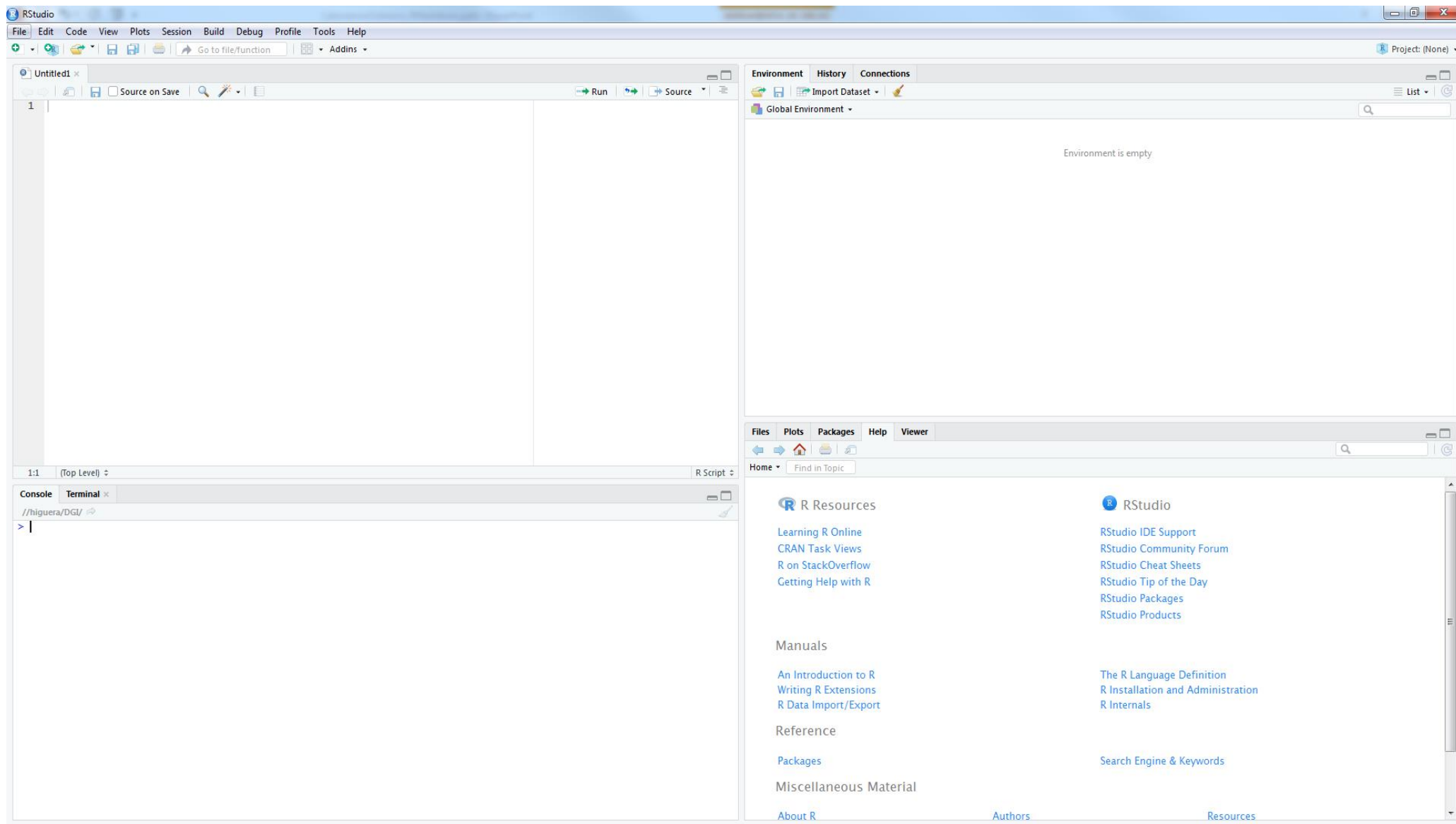
Editor de código

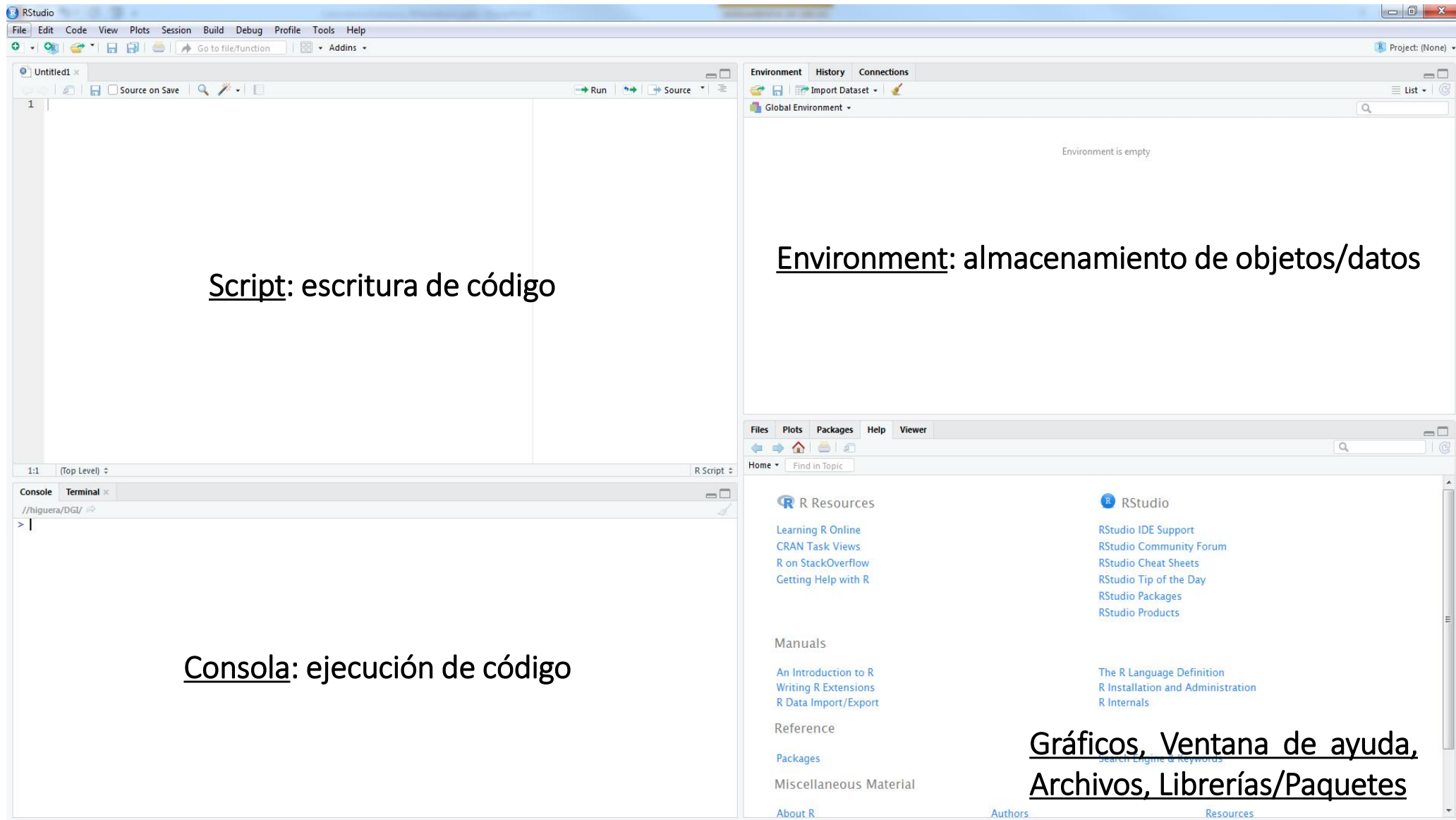
- Auto-completación
- Ayuda de sintaxis
- Ejecución directa

Herramientas para distintas tareas

- Visualización
- Conexión con otras plataformas
- Depuración de código
- Manejo del ambiente de trabajo








Script: escritura de código

Environment: almacenamiento de objetos/datos


Consola: ejecución de código


Gráficos, Ventana de ayuda,
Archivos, Librerías/Paquetes

<https://rstudio.cloud/>

 rstudio.cloud

☆ M S G

 Studio Cloud

Log In Sign Up 

Welcome to RStudio Cloud^{alpha}

Do, share, teach and learn data science with R.

Get Started

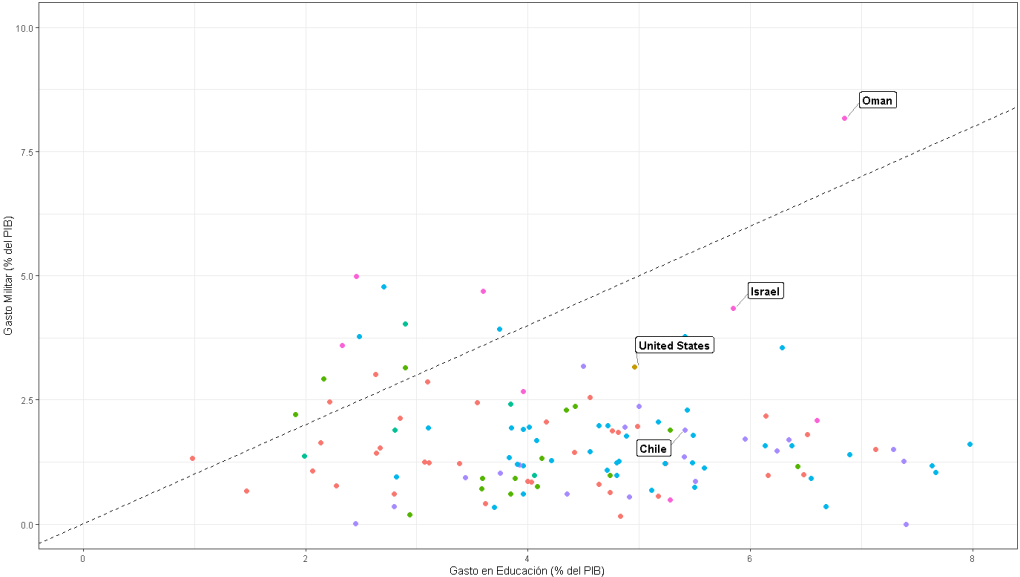
If you already have an RStudio shinyapps.io account, you can log in using your existing credentials.

	A	B	C	D	E	F	G	H	I	J	K
1	Country Name	Country Code	Indicator Name	Indicator Code	2012	2013	2014	2015	2016	2017	2018
2	Arab World	ARB	2005 PPP conversion factor, GDI PA,NUS,PPP.05								
3	Arab World	ARB	2005 PPP conversion factor, priv PA,NUS,PRVT,PP.05								
4	Arab World	ARB	Access to clean fuels and technic EG.CFT.ACCTS.ZS		83.12030269	83.53345682	83.89759605	84.17159902	84.51017125		
5	Arab World	ARB	Access to electricity (% of popul EG.ELC.ACCTS.ZS		87.07057557	88.17683639	87.34273859	89.13012075	89.67985806	90.2736866	
6	Arab World	ARB	Access to electricity, rural (% of EG.ELC.ACCTS.RU.ZS		75.24410407	77.16230459	75.53897632	78.74115241	79.66563504	80.74929273	
7	Arab World	ARB	Access to electricity, urban (% of EG.ELC.ACCTS.UR.ZS		95.96216637	96.35293045	95.99783283	96.64991605	96.83418418	97.00397447	
8	Arab World	ARB	Account ownership at a financia FX.OWN.TOTL.ZS				30.27713013			37.16521072	
9	Arab World	ARB	Account ownership at a financia FX.OWN.TOTL.FE.ZS				22.07934952			25.63540268	
10	Arab World	ARB	Account ownership at a financia FX.OWN.TOTL.MA.ZS				37.79076385			48.32851791	
11	Arab World	ARB	Account ownership at a financia FX.OWN.TOTL.OL.ZS				34.21658325			42.54204559	
12	Arab World	ARB	Account ownership at a financia FX.OWN.TOTL.40.ZS				22.77989006			27.72478104	
13	Arab World	ARB	Account ownership at a financia FX.OWN.TOTL.PL.ZS				21.27804184			26.45811081	
14	Arab World	ARB	Account ownership at a financia FX.OWN.TOTL.60.ZS				35.23748016			43.44695282	
15	Arab World	ARB	Account ownership at a financia FX.OWN.TOTL.SG.ZS				38.90061188			48.66697693	
16	Arab World	ARB	Account ownership at a financia FX.OWN.TOTL.YG.ZS				21.25614166			20.95479965	
17	Arab World	ARB	Adequacy of social insurance pr per_sl_allis.adq_pop_tot								
18	Arab World	ARB	Adequacy of social protection ar per_allsp.adq_pop_tot								
19	Arab World	ARB	Adequacy of social safety net pr per_sa_allsa.adq_pop_tot								
20	Arab World	ARB	Adequacy of unemployment ben per_fm_allim.adq_pop_tot								
21	Arab World	ARB	Adjusted net enrollment rate, pri SE.PRM.TENR		85.20714	84.21832	84.2543	84.03523	84.53258	85.14375	85.38422
22	Arab World	ARB	Adjusted net enrollment rate, pri SE.PRM.TENR.FE		84.11878	83.21839	83.34494	83.18996	83.82028	83.99478	84.25278
23	Arab World	ARB	Adjusted net enrollment rate, pri SE.PRM.TENR.MA		86.30059	85.22583	85.18359	84.89517	85.25464	86.28836	86.50601
24	Arab World	ARB	Adjusted net national income (ar NY.ADJ.NNTY.KD.ZG		6.090651928	3.251583234	1.437907034	-5.363871141	1.4662486	1.610269936	
25	Arab World	ARB	Adjusted net national income (cc NY.ADJ.NNTY.KD		2.01E+12	2.08E+12	2.11E+12	2.00E+12	2.03E+12	2.06E+12	
26	Arab World	ARB	Adjusted net national income (cc NY.ADJ.NNTY.CD		2.19E+12	2.26E+12	2.34E+12	2.16E+12	2.14E+12	2.16E+12	
27	Arab World	ARB	Adjusted net national income pe NY.ADJ.NNTY.PC.KD.ZG		3.724358293	1.004889815	-0.706924203	-7.304377833	-0.54189691	-0.332277606	

^	pais	anio	grupo_ingresos	region	OECD	plb_percapita	gini	gasto_educacion	gasto_militar	gasto_I&D	desempleo
1	Afghanistan	2018	Ingreso Bajo	Asia del Sur	No	1951.5585	NA	4.05887	0.984560504	NA	1.542
2	Albania	2018	Ingreso Medio-Alto	Europa & Asia Central	No	13325.5546	29.0	3.95464	1.178900558	NA	13.898
3	Algeria	2018	Ingreso Medio-Alto	Medio Este & Africa del Norte	No	15621.8594	NA	NA	5.271414046	0.53347	12.145
4	American Samoa	2018	Ingreso Medio-Alto	Asia del Este & Pacifico	No	NA	NA	NA	NA	NA	NA
5	Andorra	2018	Ingreso Alto	Europa & Asia Central	No	NA	NA	3.19556	NA	NA	NA
6	Angola	2018	Ingreso Medio-Bajo	Africa Subsahariana	No	6440.9763	NA	NA	1.777137554	NA	7.253
7	Antigua and Barbuda	2018	Ingreso Alto	Latinoamérica y el Caribe	No	26739.4682	NA	NA	NA	NA	NA
8	Argentina	2018	Ingreso Medio-Alto	Latinoamérica y el Caribe	No	20567.3018	41.2	5.50825	0.854560979	0.53274	9.483
9	Armenia	2018	Ingreso Medio-Alto	Europa & Asia Central	No	10324.9351	33.6	2.70545	4.778337587	0.22770	17.712
10	Aruba	2018	Ingreso Alto	Latinoamérica y el Caribe	No	39454.6298	NA	6.18990	NA	NA	NA
11	Australia	2018	Ingreso Alto	Asia del Este & Pacifico	Si	51601.7839	35.8	5.28031	1.891559767	1.92296	5.387
12	Austria	2018	Ingreso Alto	Europa & Asia Central	Si	55509.5931	30.5	5.50070	0.735992539	3.15925	4.786
13	Azerbaijan	2018	Ingreso Medio-Alto	Europa & Asia Central	No	18012.3155	NA	2.48097	3.773694092	0.18521	5.220
14	Bahamas, The	2018	Ingreso Alto	Latinoamérica y el Caribe	No	31581.1044	NA	NA	NA	NA	11.850
15	Bahrain	2018	Ingreso Alto	Medio Este & Africa del Norte	No	47219.8411	NA	2.32721	3.595485391	0.10116	0.962
16	Bangladesh	2018	Ingreso Medio-Bajo	Asia del Sur	No	4364.0453	32.4	1.98602	1.364917496	NA	4.308
17	Barbados	2018	Ingreso Alto	Latinoamérica y el Caribe	No	18526.0086	NA	4.65652	NA	NA	9.568
18	Belarus	2018	Ingreso Medio-Alto	Europa & Asia Central	No	19959.5427	25.4	4.82003	1.265447172	0.58716	5.708
19	Belgium	2018	Ingreso Alto	Europa & Asia Central	Si	50366.6790	27.7	6.54428	0.925790635	2.60617	6.323
20	Belize	2018	Ingreso Medio-Alto	Latinoamérica y el Caribe	No	8786.4948	NA	7.37876	1.262456192	NA	9.368
21	Benin	2018	Ingreso Bajo	Africa Subsahariana	No	2420.4797	47.8	3.99469	0.863004045	NA	2.125
22	Bermuda	2018	Ingreso Alto	América del Norte	No	52547.3331	NA	1.50039	NA	0.21916	NA

Relación entre Gasto Militar y Gasto en Educación (2018)

La gran parte de los países del mundo gasta más en educación



	A	B	C	D	E	F	G	H	I	J	K
1	Country Name	Country Code	Indicator Name	Indicator Code	2012	2013	2014	2015	2016	2017	2018
2	Arab World	ARB	2005 PPP conversion factor, GDI PA,NUS,PPP.05								
3	Arab World	ARB	2005 PPP conversion factor, priv PA,NUS,PRVT,PP.05								
4	Arab World	ARB	Access to clean fuels and techn EG,CFT,ACCS.ZS		83.12030269	83.53345682	83.89759605	84.17159902	84.51017125		
5	Arab World	ARB	Access to electricity (% of popul EG,ELC,ACCS.ZS		87.07057557	88.17683639	87.34273859	89.13012075	89.67968506	90.2736866	
6	Arab World	ARB	Access to electricity, rural (% of EG,ELC,ACCS.RU.ZS		75.24410407	77.16230459	75.53897632	78.74115241	79.66563504	80.74929273	
7	Arab World	ARB	Access to electricity, urban (% of EG,ELC,ACCS.UR.ZS		95.96216637	96.35293045	95.99783283	96.64991605	96.83418418	97.00397447	
8	Arab World	ARB	Account ownership at a financia FX,OWN,TOTL,ZS				30.27713013			37.16521072	
9	Arab World	ARB	Account ownership at a financia FX,OWN,TOTL,FE,ZS				22.07934952			25.63540268	
10	Arab World	ARB	Account ownership at a financia FX,OWN,TOTL,MA,ZS				37.79076385			48.32851791	
11	Arab World	ARB	Account ownership at a financia FX,OWN,TOTL,OL,ZS				34.21658325			42.54204559	
12	Arab World	ARB	Account ownership at a financia FX,OWN,TOTL,40,ZS				22.77989006			27.72478104	
13	Arab World	ARB	Account ownership at a financia FX,OWN,TOTL,PL,ZS				21.27804184			26.45811081	
14	Arab World	ARB	Account ownership at a financia FX,OWN,TOTL,60,ZS				35.23748016			43.44695282	
15	Arab World	ARB	Account ownership at a financia FX,OWN,TOTL,SO,ZS				38.90061188			48.66697693	
16	Arab World	ARB	Account ownership at a financia FX,OWN,TOTL,YG,ZS				21.25614166			20.95479965	
17	Arab World	ARB	Adequacy of social insurance pri per_sa_adi.adq_pop_tot								
18	Arab World	ARB	Adequacy of social protection ar per_sa_adi.adq_pop_tot								
19	Arab World	ARB	Adequacy of social safety net pri per_sa_adi.adq_pop_tot								
20	Arab World	ARB	Adequacy of unemployment ben per_fm_allim.adq_pop_tot								
21	Arab World	ARB	Adjusted net enrollment rate, pri SE,PRM,TENR		85.20714	84.21832	84.2543	84.03523	84.53258	85.14375	85.38422
22	Arab World	ARB	Adjusted net enrollment rate, pri SE,PRM,TENR,FE		84.11878	83.21839	83.34494	83.18996	83.82028	83.99478	84.25278
23	Arab World	ARB	Adjusted net enrollment rate, pri SE,PRM,TENR,MA		86.30059	85.22583	85.18359	84.89517	85.25464	86.28836	86.50601
24	Arab World	ARB	Adjusted net national income (ar NY,ADI,NNTY,KD,ZG		6.090651928	3.251583234	1.437907034	-5.363871141	1.4662486	1.610269936	
25	Arab World	ARB	Adjusted net national income (co NY,ADI,NNTY,KD		2.01E+12	2.08E+12	2.11E+12	2.00E+12	2.03E+12	2.06E+12	
26	Arab World	ARB	Adjusted net national income (co NY,ADI,NNTY,CD		2.19E+12	2.26E+12	2.34E+12	2.16E+12	2.14E+12	2.16E+12	
27	Arab World	ARB	Adjusted net national income pe NY,ADI,NNTY,PC,KD,ZG		3.724358293	1.004889815	-0.706924203	-7.304377833	-0.54189691	-0.332277606	

^	pais	anio	grupo_ingresos	region	OECD	plb_percapita	gini	gasto_educacion	gasto_militar	gasto_I&D	desempleo
1	Afghanistan	2018	Ingreso Bajo	Asia del Sur	No	1951.5585	NA	4.05887	0.984560504	NA	1.542
2	Albania	2018	Ingreso Medio-Alto	Europa & Asia Central	No	13325.5546	29.0	3.95464	1.178900558	NA	13.898
3	Algeria	2018	Ingreso Medio-Alto	Medio Este & Africa del Norte	No	15621.8594	NA	NA	5.271414046	0.53347	12.145
4	American Samoa	2018	Ingreso Medio-Alto	Asia del Este & Pacifico	No	NA	NA	NA	NA	NA	NA
5	Andorra	2018	Ingreso Alto	Europa & Asia Central	No	NA	NA	3.19556	NA	NA	NA
6	Angola	2018	Ingreso Medio-Bajo	Africa Subsahariana	No	6440.9763	NA	NA	1.777137554	NA	7.253
7	Antigua and Barbuda	2018	Ingreso Alto	Latinoamérica y el Caribe	No	26739.4682	NA	NA	NA	NA	NA
8	Argentina	2018	Ingreso Medio-Alto	Latinoamérica y el Caribe	No	20567.3018	41.2	5.50825	0.854560979	0.53274	9.483
9	Armenia	2018	Ingreso Medio-Alto	Europa & Asia Central	No	10324.9351	33.6	2.70545	4.778337587	0.22770	17.712
10	Aruba	2018	Ingreso Alto	Latinoamérica y el Caribe	No	39454.6298	NA	6.18990	NA	NA	NA
11	Australia	2018	Ingreso Alto	Asia del Este & Pacifico	Si	51601.7839	35.8	5.28031	1.891599767	1.92296	5.387
12	Austria	2018	Ingreso Alto	Europa & Asia Central	Si	55509.5931	30.5	5.50070	0.735992539	3.15925	4.786
13	Azerbaijan	2018	Ingreso Medio-Alto	Europa & Asia Central	No	18012.3155	NA	2.48097	3.773694092	0.18521	5.220
14	Bahamas, The	2018	Ingreso Alto	Latinoamérica y el Caribe	No	31581.1044	NA	NA	NA	NA	11.850
15	Bahrain	2018	Ingreso Alto	Medio Este & Africa del Norte	No	47219.8411	NA	2.32721	3.59545391	0.10116	0.962
16	Bangladesh	2018	Ingreso Medio-Bajo	Asia del Sur	No	4364.0453	32.4	1.98602	1.364917496	NA	4.308
17	Barbados	2018	Ingreso Alto	Latinoamérica y el Caribe	No	18526.0086	NA	4.65652	NA	NA	9.568
18	Belarus	2018	Ingreso Medio-Alto	Europa & Asia Central	No	19959.5427	25.4	4.82003	1.265447172	0.58716	5.708
19	Belgium	2018	Ingreso Alto	Europa & Asia Central	Si	50366.6790	27.7	6.54428	0.925790635	2.60617	6.323
20	Belize	2018	Ingreso Medio-Alto	Latinoamérica y el Caribe	No	8786.4948	NA	7.37876	1.262456192	NA	9.368
21	Benin	2018	Ingreso Bajo	Africa Subsahariana	No	2420.4797	47.8	3.99469	0.863004045	NA	2.125
22	Bermuda	2018	Ingreso Alto	América del Norte	No	52547.3331	NA	1.50039	NA	0.21916	NA

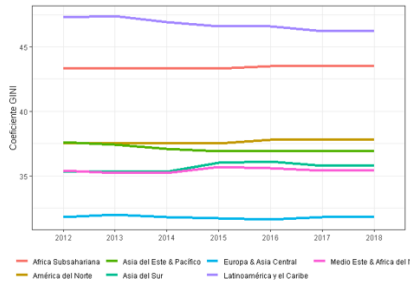
Análisis de la desigualdad en el mundo (coeficiente GINI)

Sobre el documento

En el siguiente documento haremos un breve análisis sobre la desigualdad en el mundo (medida a través del **coeficiente de GINI**). Los datos utilizados provienen del sitio web de indicadores del **Banco Mundial**.

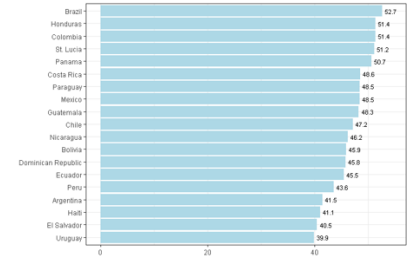
Descripción de desigualdad a nivel regional y país

En primer lugar analizaremos como se comporta la desigualdad en el período estudiado (2012-2018) para las distintas regiones del mundo. Cabe mencionar que a **menor coeficiente GINI, menor desigualdad**.



En el gráfico se puede percibir que la desigualdad tiende a mantenerse constante en el período estudiado, destacando **Latinoamérica y el Caribe** como la región más desigual pero la que pareciera tener la mayor disminución. El máximo valor corresponde a Latinoamérica y el Caribe el año 2013 mientras que el menor valor corresponde a Europa & Asia Central el año 2016.

Considerando que Latinoamérica y el Caribe es la región con mayor desigualdad, haremos un zoom en los países que lo componen. El siguiente gráfico muestra el promedio del coeficiente GINI para cada uno de los países de la región en el período estudiado.



Análisis estadístico

A continuación, estimaremos algunos modelos para ver si podemos aprender algo más sobre la desigualdad a nivel mundial para el período 2012-2018. Para esto, estimaremos - para todo país (i) y año (j) - los siguientes modelos:

- $GINI_{i,j} = PIB_{percapita_{i,j}} + GrupoIngresos_{i,j} + Region_i + Año_j$
- $GINI_{i,j} = PIB_{percapita_{i,j}} + GastoEducacion_{i,j} + Desempleo_{i,j} + GrupoIngresos_{i,j} + Region_i + Año_j$
- $GINI_{i,j} = PIB_{percapita_{i,j}} + GastoEducacion_{i,j} + Desempleo_{i,j} + GastoMilitar_{i,j} + GastoI\&D_{i,j} + GrupoIngresos_{i,j} + Region_i + Año_j$

Los resultados de los modelos (tablas de regresión) se pueden ver en el **Anexo I**. Los modelos estimados parecieran no ser muy concluyentes, lo más consistente es que a medida que los países tienen mayor desempleo también aumenta el nivel de desigualdad. Sería interesante poder incluir otras variables para tratar de explicar de mejor manera la desigualdad a nivel país.