

# Introduction to Data Science

Be a RockStar with data :)



# Data Scientist?

Forbes ADVISOR

## Fastest-Growing Tech Careers

### Data Scientists

**Growth Rate (2021-31):** +36%

**Median Pay:** \$100,910 per year

**Education Requirements:** Bachelor's degree

**Career Overview:** Data scientists extract insights and knowledge from large, complex data sets. They leverage that data to make intelligent, informed decisions to help organizations improve their performance and achieve their goals.

Conducting surveys or scraping the web to collect data is a key component of a data scientist's job. From there, data scientists clean and classify raw data, using machine learning and data visualization software to demonstrate their findings. It's paramount that data scientists know how to communicate their findings effectively and in a way that's accessible to a general audience.

TOI Times of India

## How data science is ushering in a new era of radiology

In September 2022, an Apple Watch saved the life of a 54-year-old man in the United Kingdom. The watch's built-in ECG sensor detected...

18 hours ago

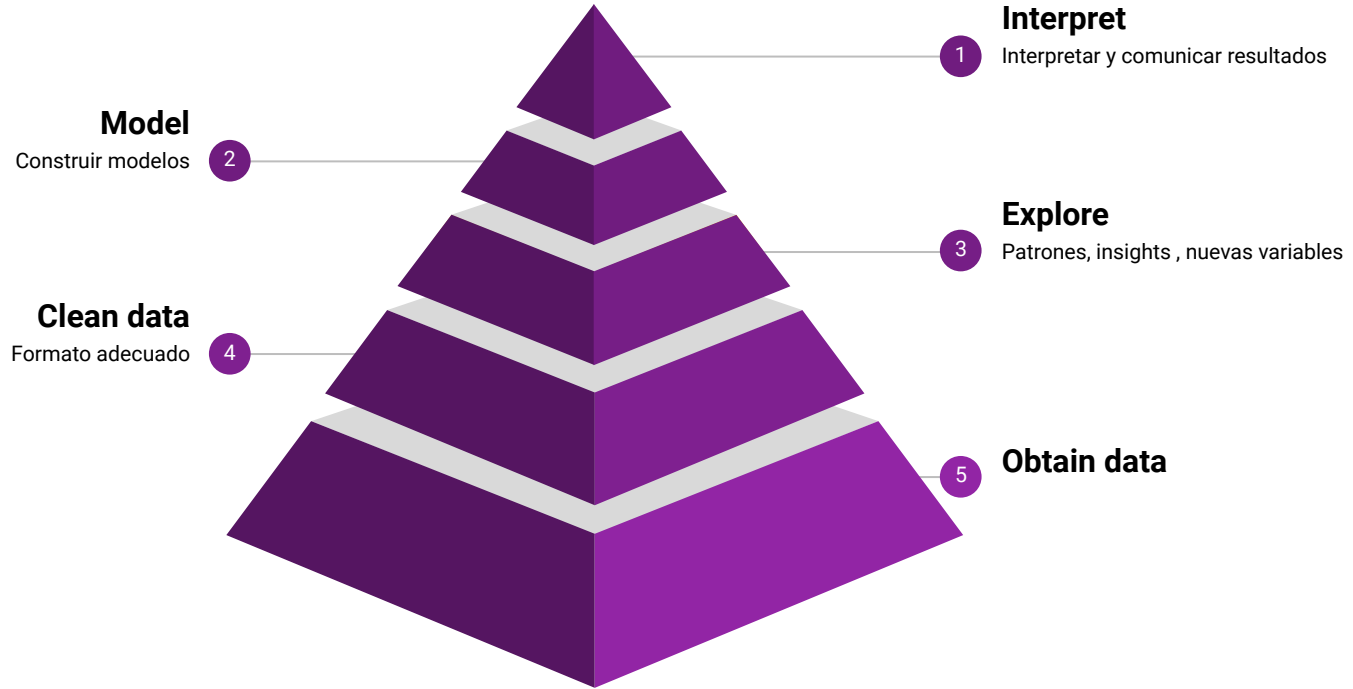
TOI  
BLOGS



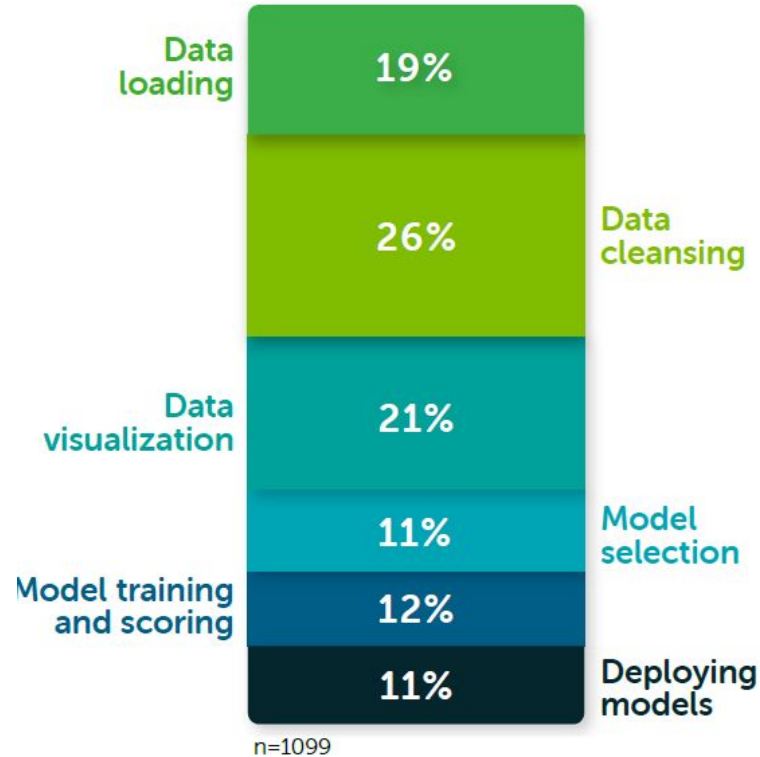
## Aplicaciones



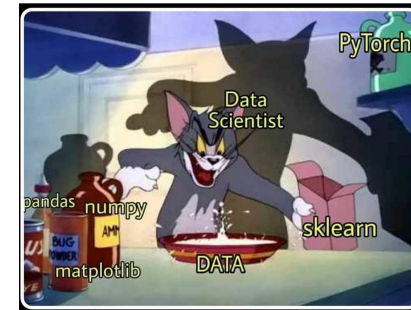
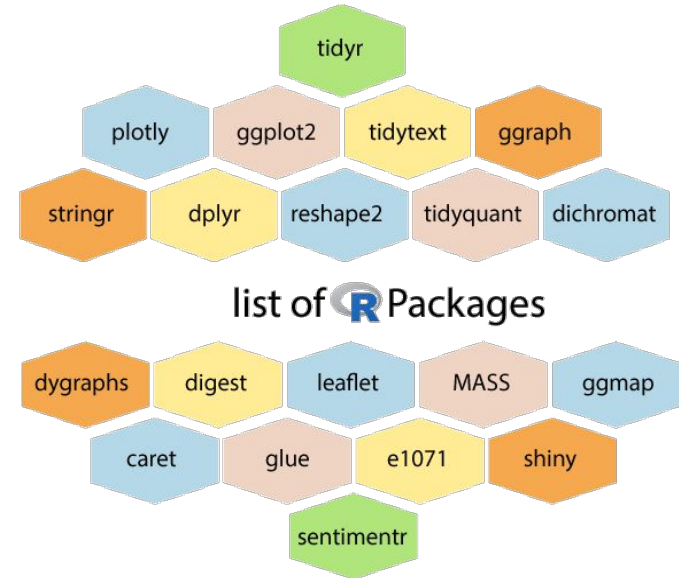
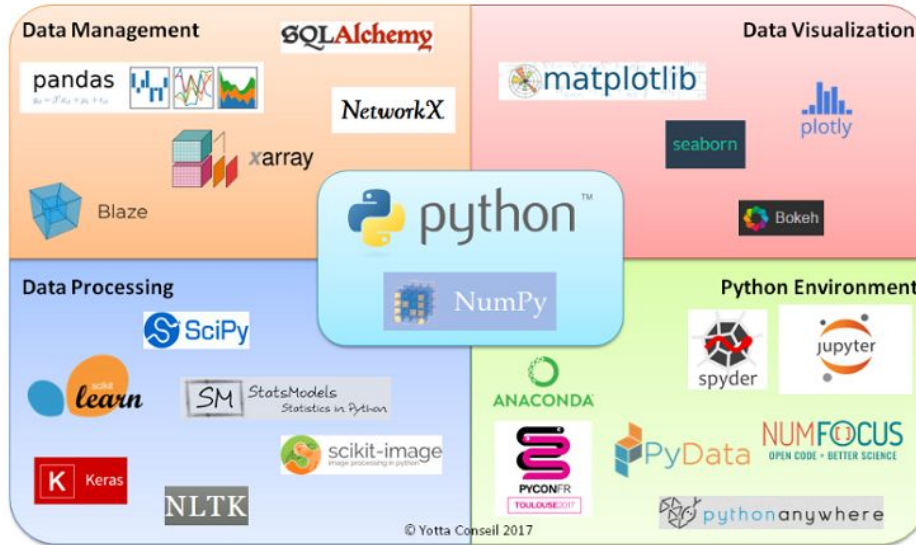
# Data Science process



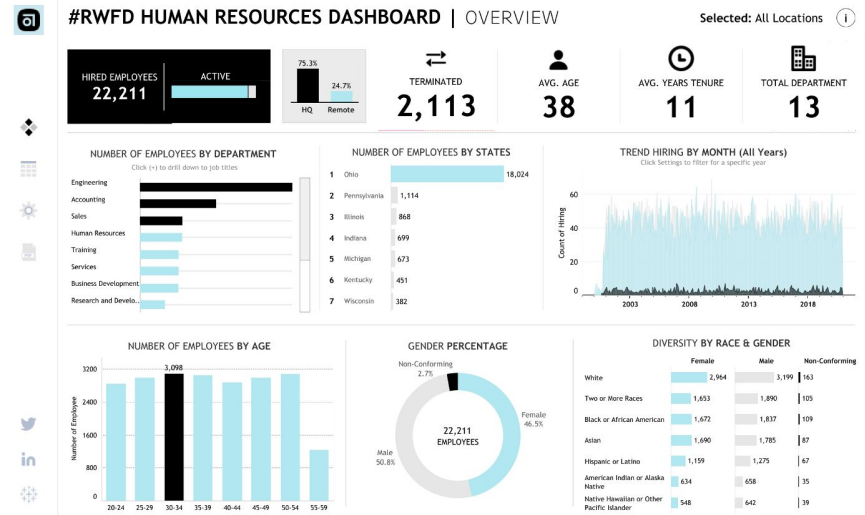
# Data Scientists' Time



# Programming Languages

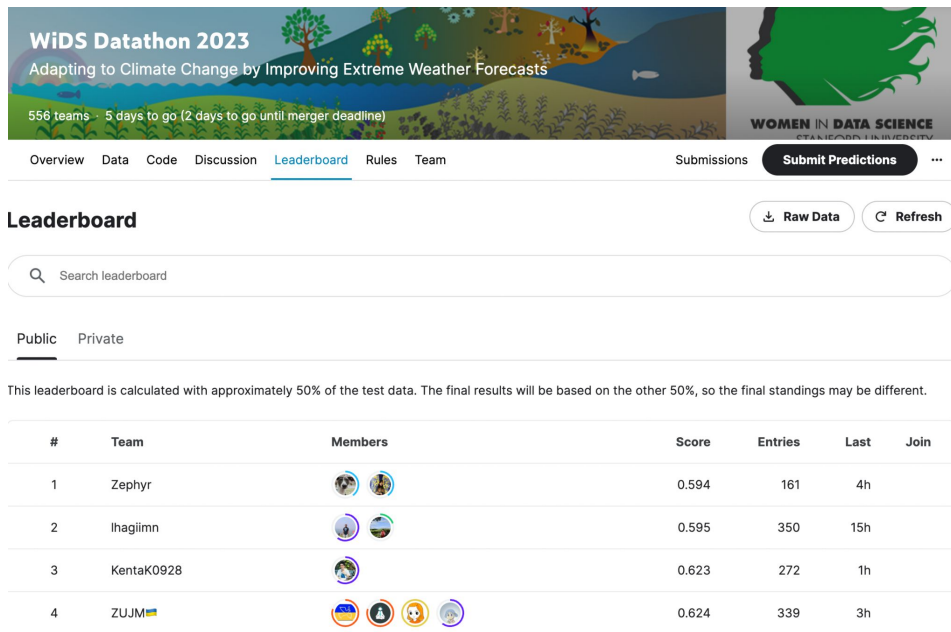


# Visualization tools



# Intro to kaggle

## Leaderboard



**WiDS Datathon 2023**  
Adapting to Climate Change by Improving Extreme Weather Forecasts  
556 teams 5 days to go (2 days to go until merger deadline)

Overview Data Code Discussion **Leaderboard** Rules Team Submissions **Submit Predictions** ...

**Leaderboard** [Raw Data](#) [Refresh](#)

Search leaderboard

Public Private

This leaderboard is calculated with approximately 50% of the test data. The final results will be based on the other 50%, so the final standings may be different.

#	Team	Members	Score	Entries	Last	Join
1	Zephyr		0.594	161	4h	
2	lhagiimn		0.595	350	15h	
3	Kentak0928		0.623	272	1h	
4	ZUJM		0.624	339	3h	

<https://www.kaggle.com/c/widsdatathon2023/>

## Entornos de trabajo



# Data Explorer

651.4 MB

sample\_solution.csv

test\_data.csv

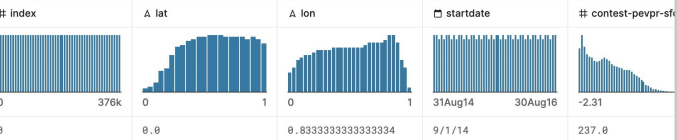
train\_data.csv

train\_data.csv (600.76 MB)



Detail Compact Column

10 of 246 columns

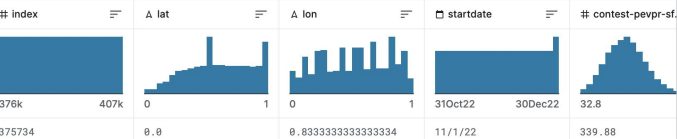


test\_data.csv (49.83 MB)



Detail Compact Column

10 of 245 columns



**Task:** Predecir la media aritmética de la temperatura máxima y mínima durante los próximos 14 días, para cada lugar y fecha de inicio.

## Data Dictionary

The WIDS 2023 Datathon is using a subset of a pre-prepared dataset in which the variables were gathered from the following datasets (*source of the WIDS Datathon dataset will be revealed after the competition closes*):

- Temperature:** Daily maximum and minimum temperature measurements at 2 meters from 1979 onwards were obtained from NOAA's Climate Prediction Center (CPC) Global Gridded Temperature dataset and converted to Celsius. The official contest target temperature variable is `tmp2m = tmax+tmin / 2`.

[ftp://ftp.cpc.ncep.noaa.gov/precip/PEOPLE/wd52ws/global\\_temp/](ftp://ftp.cpc.ncep.noaa.gov/precip/PEOPLE/wd52ws/global_temp/)

- Global precipitation:** Daily precipitation data from 1979 onward were obtained from NOAA's CPC Gauge-Based Analysis of Global Daily Precipitation [42] and converted to mm.

[ftp://ftp.cpc.ncep.noaa.gov/precip/CPC\\_UNI\\_PRCP/GAUGE\\_GLB/RT/](ftp://ftp.cpc.ncep.noaa.gov/precip/CPC_UNI_PRCP/GAUGE_GLB/RT/)



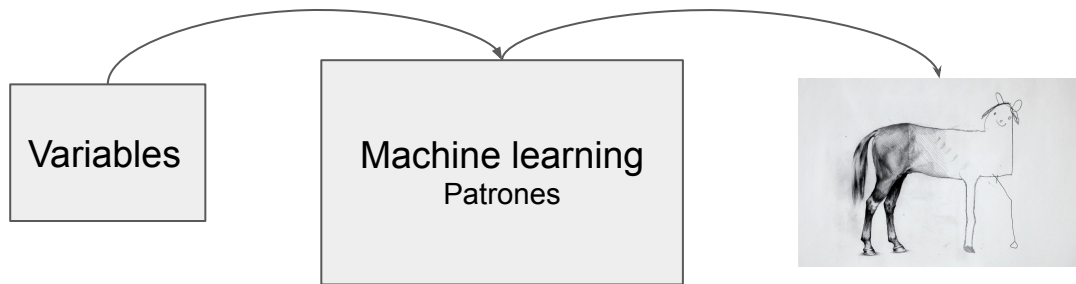
# TIPOS DE VARIABLES

NUMERICAS

CATEGORICAS

# Feature engineering

¿Por qué debemos darle mucha importancia a este paso?



¿Qué variables podemos crear a partir de otras?

- Edad
- Fecha

# Handling missing Values

Depende del tipo de valor faltante!

ID	AGE	GENDER	ANNUAL INCOME	MARITAL STATUS	NUMBER OF CHILDREN	JOB	BUY?
1		A	150000		1	ENGINEER	NO
2	27	B	50000			TEACHER	NO
3		A	10000	MARRIED	2		YES
4	40	B			2	ENGINEER	YES
5	35	B		SINGLE	0	DOCTOR	YES
6		A	50000		0	TEACHER	NO
7	33	B	60000	SINGLE		TEACHER	NO
8	20	B	10000			STUDENT	NO

Designing Machine Learning Systems, Chip Huyen

- Eliminar las observaciones faltantes
- Imputación
  - Promedio, media, mediana, 0, -9999

01	Missing not at random	<ul style="list-style-type: none"><li>• Falta por el valor verdadero en sí Ej, ingresos</li></ul>
02	Missing at random	<ul style="list-style-type: none"><li>• Falta a causa de otra característica Ej, edad con mujeres</li></ul>
03	Missing completely at random	<ul style="list-style-type: none"><li>• No existe un patrón identificado para ese valor faltante</li></ul>

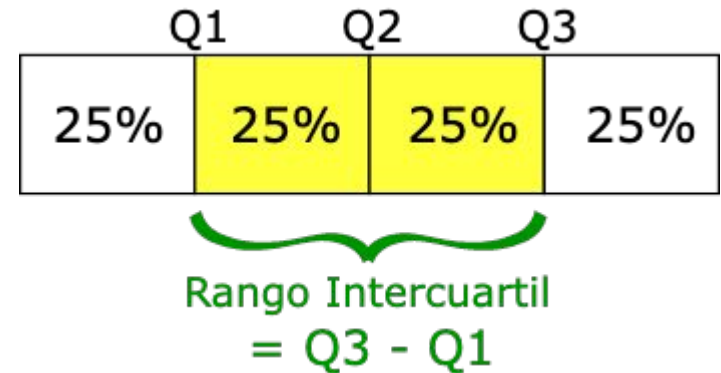


# Outliers

- Eliminarlos? No es una opción factible

La mejor opción es quitarle peso a esas observaciones atípicas mediante técnicas robustas!

- Imputar con la mediana.
- Imputar con extremos del rango intercuartil.



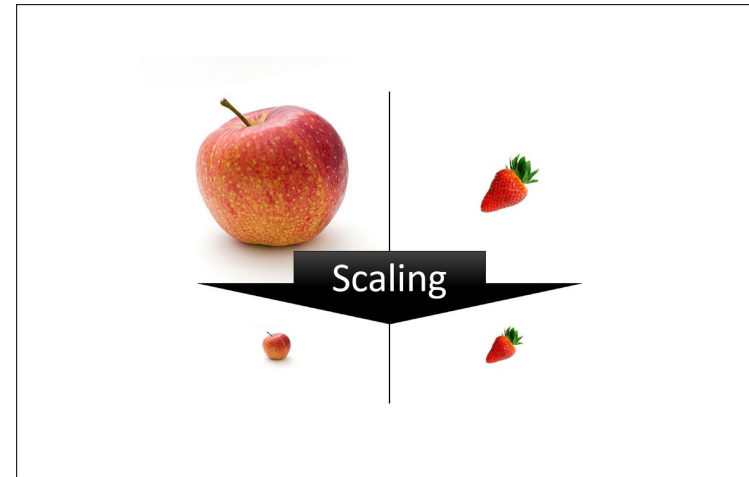
# Variable transformations

Scaling ( Valores muy grandes pueden llegar a generalizar!)

- Min max scaler [0,1]
- Estandarización ( variables siguen una distribución normal)

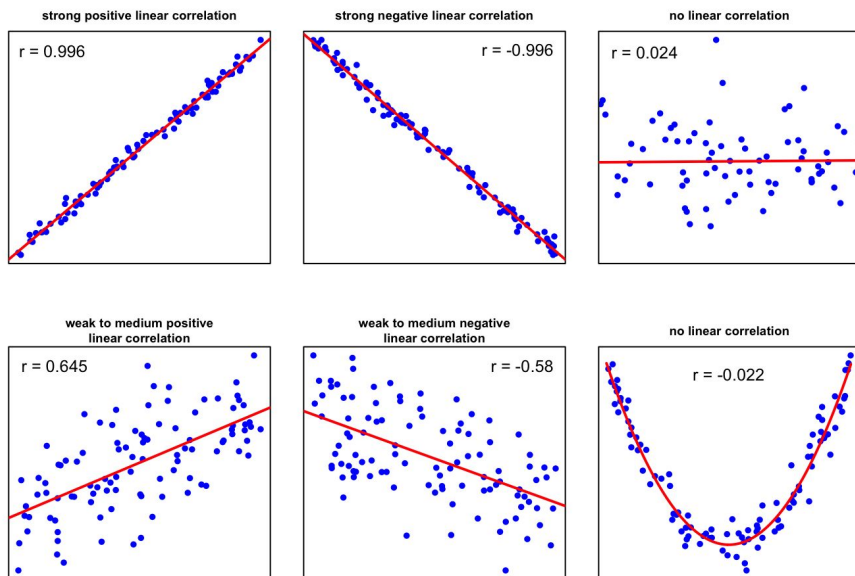
$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$$x_{standardized} = \frac{x - x_{mean}}{x_{standard\ deviation}}$$

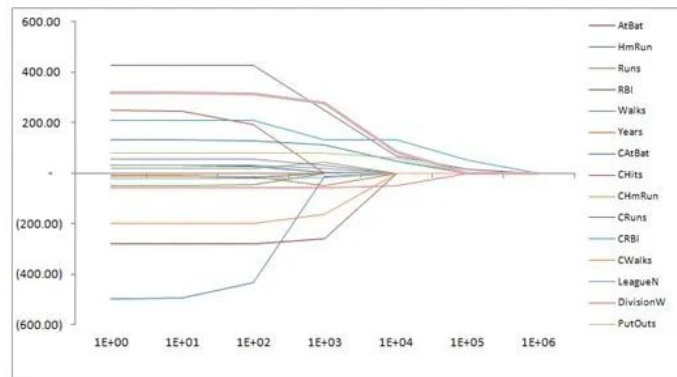


# Feature Selection

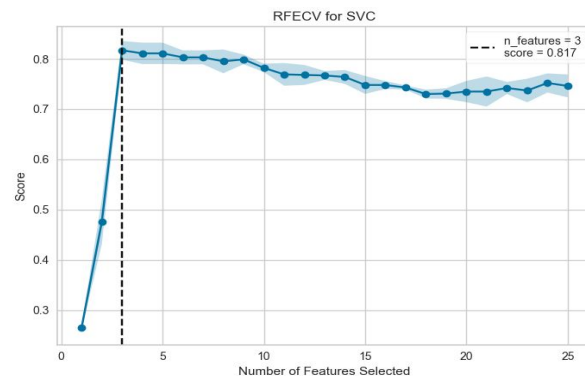
## Correlación



## Lasso



## Eliminación recursiva de características (RFE)



# Machine Learning Tasks

