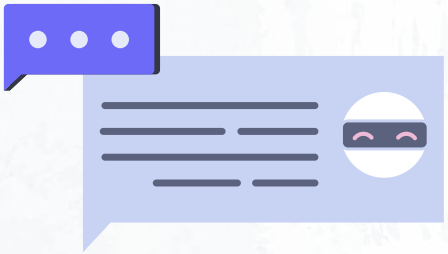
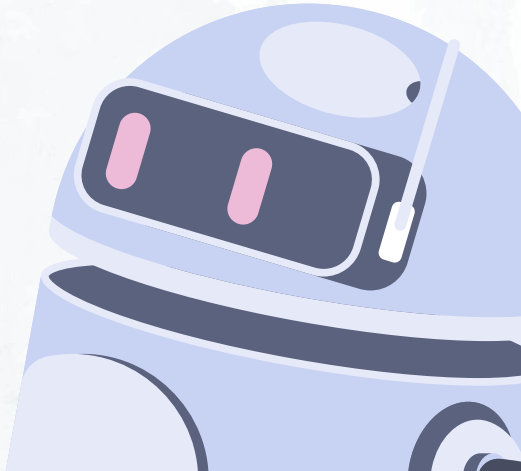
Analítica de Datos (Aprendizaje de máquina)





UNAL

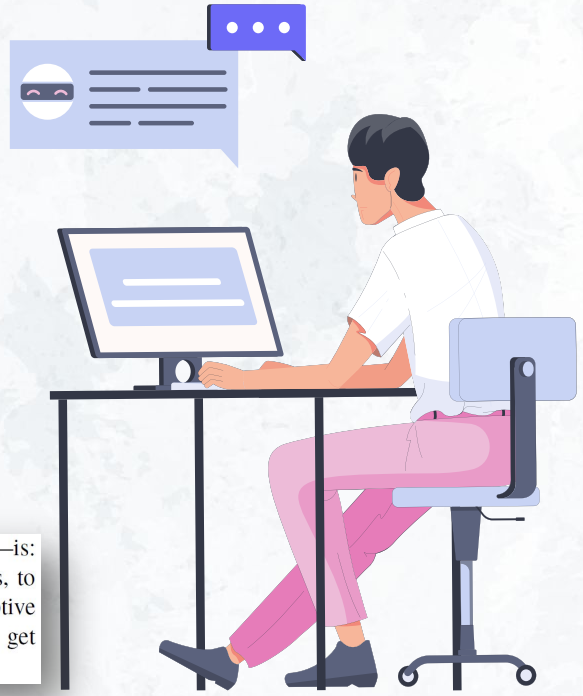
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD NACIONAL DE COLOMBIA

Datasets

- Un dataset (conjunto de datos) es una colección de descriptores de un **mismo fenómeno**.

- Estos descriptores pueden tomar una variedad de formas distintas, **pero lo importante es que todos describen el mismo fenómeno.**



The single most important question for a working scientist—perhaps the single most useful question anyone can ask—is: "what's going on here?" Answering this question requires creative use of different ways to make pictures of datasets, to summarize them, and to expose whatever structure might be there. This is an activity that is sometimes known as "Descriptive Statistics". There isn't any fixed recipe for understanding a dataset, but there is a rich variety of tools we can use to get insights.

Datasets, ¿dónde encontrarlos?

Actualmente, la oferta de conjuntos de datos es muy variada:

Datasets para data processing

https://registry.opendata.aws

https://cloud.google.com/bigquery/public-data/

https://archive.ics.uci.edu/ml/datasets.php

https://www.kaggle.com

https://www.quandl.com/search

https://data.world

https://www.dataquest.io/blog/free-datasets-for-projects/
https://towardsdatascience.com/26-datasets-for-your-data-science-projects-658601590a4c

Datasets, ¿dónde encontrarlos?

Actualmente, la oferta de conjuntos de datos es muy variada:

Datasets para data processing

DATA.GOV https://www.data.gov

THE WORLD BANK https://data.worldbank.org

reddit Datasets https://www.reddit.com/r/datasets/

Academic Torrents https://academictorrents.com

Datos de AGA/OGP

AGA = Alianza para el Gobierno Abierto
OGP = Open Government Partnership

https://www.opengovpartnership.org/es/

En Colombia la iniciativa de Datos Abiertos la maneja el **MINTIC** (***Ministerio de las Tecnologías de la Información y las Comunicaciones***)

GOV.CO | DATOS ABIERTOS

https://www.datos.gov.co/browse

Datasets, ¿qué encontramos en ellos?

Veamos dos ejemplos y aprendamos a interpretar los formatos en los que se presenta la información.

- Ejemplo 1)

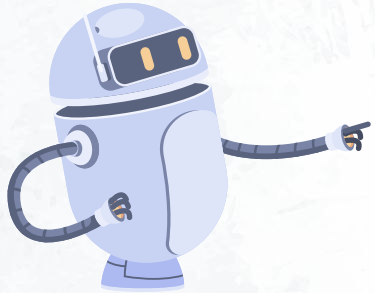
  Repositorio: Kaggle

  Base de datos: Calidad del aire en Madrid (2001-2018)

- Ejemplo 2)

  Repositorio: Datos abiertos – MINTIC Colombia

  Base de datos: Suscriptores y asociados de televisión cerrada (histórico)



Tipos de formatos

(+) Antes de iniciar cualquier estrategia de análisis de datos debemos pensar en qué **tipo de formato** vienen los datos y cómo podremos acceder a ellos desde el código que desarrollemos.

Tipos de formatos más comunes

RAW TEXT

Los formatos más comunes para ficheros de texto son Unicode, ASCII o UTF-8. Si hay necesidad de caracteres internacionales los más comunes serán Unicode o UTF-8. Hay que tener en cuenta que ficheros codificados o binarios (ficheros de Word, PDF, Matlab, etc.) no son Raw Text

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse
eget metus quis erat tempor hendrerit. Vestibulum turpis ante, bibendum
vitae nisi non, euismod blandit dui. Maecenas tristique consectetur est
nec elementum. Maecenas porttitor, arcu sed gravida tempus, purus tellus
lacinia erat, dapibus euismod felis enim eget nisl. Nunc mollis volutpat
ligula. Etiam interdum porttitor nulla non lobortis.
```

Tipos de formatos más comunes

Comma Separated Values - CSV

El formato CSV es de los más empleados en bases de datos. Se pueden encontrar otros delimitadores, tales como espacios de tabulación (TSV), o la barra vertical - pipe symbol (|) (PSV).

```
2,male,Mr.,Daniel,J,Carpenter,51 Guildford Rd,EAST
DRAYTON,,DN22 3GT,GB,United Kingdom,DanielCarpenter@teleworm.
us,Reste1990,Eich1Kiegie,079 2890 2948,Harris,3/26/1990,MasterCard
,5353722386063326,717,7/2018,KL 50 03 59 C,1Z 895 362 50 0377 620
2,Blue,Corporate administrative assistant,Hit or Miss,2000 Jeep Grand
Cherokee,BiologyConvention.co.uk,AB+,175.3,79.7,5' 7",169,ac907a59-a091-
4ba2-9b0f-a1276b3b5ada,52.801024,-0.719021

3,male,Mr.,Harvey,A,Hawkins,37 Shore Street,STOKE TALMAGE,,OX9
4FY,GB,United Kingdom,HarveyHawkins@armyspy.com,Spicionly,UcheeGh9xoh,077
7965 0825,Rees,3/1/1974,MasterCard,5131613608666799,523,7/2017,SS 81 32
33 C,1Z Y11 884 19 7792 722 8,Black,Education planner,Monsource,1999 BMW
740,LightingShadows.co.uk,A-,224.8,102.2,6' 1",185,6cf865fb-81ae-42af-
9a9d-5b86d5da7ce9,51.573674,-1.179834
```

Tipos de formatos más comunes

JSON

El formato JSON (JavaScript Object Notation) es empleado principalmente para comunicar datos entre máquinas y la web. Se basa en una notación pareada de key/value.

Inicialmente fue diseñado como una alternativa al formato XML, aunque su uso ya está muy extendido. El nombre JavaScript no implica que sea solo usado por entornos de JavaScript. Existen parsers (intérpretes gramaticales) de JSON para distintos lenguajes.

Tipos de formatos más comunes

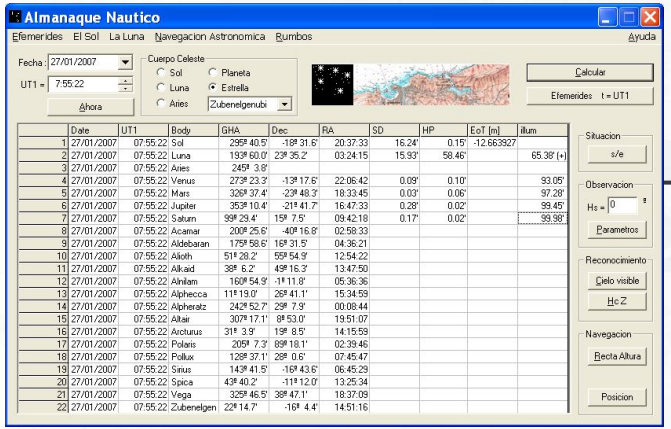
XML

```
<?xml version="1.0" encoding="UTF-8" ?>
    <Customer>
        <Number>1</Number>
        <Gender>male</Gender>
        <Title>Mr.</Title>
        <GivenName>Joe</GivenName>
        <MiddleInitial>L</MiddleInitial>
        <Surname>Perry</Surname>
        <StreetAddress>50 Park Row</StreetAddress>
        <City>EDERN</City>
        <State></State>
        <ZipCode>LL53 2SQ</ZipCode>
        <Country>GB</Country>
        <CountryFull>United Kingdom</CountryFull>
```

XML (Extensible Markup Language) hace parte de los formatos SGML (Standar Generalized Markup Language). La filosofía detrás de este lenguaje es que sea entendible tanto por máquinas como por personas.

Es una generalización del formato HTML. Su uso puede ser complejo a la hora de enfrentarse con grandes estructuras. Los sistemas más comunes usan dos tipos de parsers distintos (el DOM - Document Object Model o el SAX - Simple API for XML).

Spreadsheets =



	Date	UT1	Body	GHA	Dec	RA	SD	HP	EoT [m]	illum
1	27/01/2007	07:55:22	Sol	295º 40.5'	-18º 31.6'	20:37:33	16.24'	0.15'	-12.663927	
2	27/01/2007	07:55:22	Luna	193º 60.0'	23º 35.2'	03:24:15	15.93'	58.46'		65.38' (+)
3	27/01/2007	07:55:22	Aries	245º 3.8'						
4	27/01/2007	07:55:22	Venus	273º 23.3'	-13º 17.6'	22:06:42	0.09'	0.10'		93.05'
5	27/01/2007	07:55:22	Mars	326º 37.4'	-23º 48.3'	18:33:45	0.03'	0.06'		97.28'
6	27/01/2007	07:55:22	Jupiter	353º 10.4'	-21º 41.7'	16:47:33	0.28'	0.02'		99.45'
7	27/01/2007	07:55:22	Saturn	99º 29.4'	15º 7.5'	09:42:18	0.17'	0.02'		99.98'
8	27/01/2007	07:55:22	Acamar	200º 25.6'	-40º 16.8'	02:58:33				
9	27/01/2007	07:55:22	Aldebaran	175º 58.6'	16º 31.5'	04:36:21				
10	27/01/2007	07:55:22	Alioth	51º 28.2'	55º 54.9'	12:54:22				
11	27/01/2007	07:55:22	Alkaid	38º 6.2'	49º 16.3'	13:47:50				
12	27/01/2007	07:55:22	Alnilam	160º 54.9'	-1º 11.8'	05:36:36				
13	27/01/2007	07:55:22	Alphecca	116º 19.0'	26º 41.1'	15:34:59				
14	27/01/2007	07:55:22	Alpheratz	242º 52.7'	29º 7.9'	00:08:44				
15	27/01/2007	07:55:22	Altair	307º 17.1'	8º 53.0'	19:51:07				
16	27/01/2007	07:55:22	Arcturus	31º 3.9'	19º 8.5'	14:15:59				
17	27/01/2007	07:55:22	Polaris	205º 7.3'	89º 18.1'	02:39:46				
18	27/01/2007	07:55:22	Pollux	128º 37.1'	28º 0.6'	07:45:47				
19	27/01/2007	07:55:22	Sirius	143º 41.5'	-16º 43.6'	06:45:29				
20	27/01/2007	07:55:22	Spica	43º 40.2'	-11º 12.0'	13:25:34				
21	27/01/2007	07:55:22	Vega	325º 46.5'	38º 47.1'	18:37:09				
22	27/01/2007	07:55:22	Zubenelgen	22º 14.7'	-16º 4.4'	14:51:16				

- Las hojas de cálculo son el formato más común para almacenar y procesar información al interior de las empresas.
- Hay que tener en cuenta no sólo los contenidos a la hora de procesarlas, sino también las fórmulas en su interior.
- Muchos lenguajes de programación no hacen parsing a las fórmulas contenidas en ficheros de excel.

UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

Tipos de formatos más comunes

Las más comunes:
- MySQL
- Postgres
- Microsoft SQL Server
- Oracle

Las no tan comunes (NoSQL)
- MongoDB
- CouchDB
- Cassandra
- Redis
- HBase

DATABASES

- **Las bases de datos** dan una solución estructurada a la forma de almacenar información.
- Se basan en asociaciones (relaciones) entre atributos y valores.
- Si bien la curva de aprendizaje en el tema de bases de datos puede llegar a ser lenta, la sugerencia es darles una oportunidad para la organización de la información dentro de un sistema de análisis de datos.

Parser sintáctico

```xml
<?xml version='1.0' encoding='UTF-8'?>
<Filas>
<Fila><TenenciaDeVivienda>Propietario vivienda y
terreno</TenenciaDeVivienda> <departamento>Montevideo
</departamento> <año>2006.0</año> <valor>56.4
</valor></Fila>

<Fila><TenenciaDeVivienda>Inquilino
</TenenciaDeVivienda> <departamento>Montevideo
</departamento> <año>2006.0</año> <valor>20.9
</valor></Fila>

<Fila><TenenciaDeVivienda>Ocupante gratuito con
permiso u ocupante en relación de dependencia
</TenenciaDeVivienda> <departamento>Montevideo
</departamento> <año>2006.0</año> <valor>12.2
</valor></Fila>
```

XML

```
"Tenencia De Vivienda","departamento",año,valor

"Propietario vivienda y terreno","Montevideo",2006,56.4

"Inquilino","Montevideo",2006,20.9

"Ocupante gratuito con permiso u ocupante en relación
de dependencia","Montevideo",2006,12.2
```

CSV

Tenencia De Vivienda	departamento	año	valor
Propietario vivienda y terreno	Montevideo	2006	56.4
Inquilino	Montevideo	2006	20.9
Ocupante gratuito con permiso u ocupante en relación de dependencia	Montevideo	2006	12.2

Spreadsheet

Calidad de los datos

Se debe verificar siempre que los datos cumplen con los requerimientos necesarios. Esta etapa es particularmente necesaria cuando los datos son ingresados de manera manual.

Verificador de existencia

	Nombre	Apellidos	Correo	Edad
Correcto	Juan	Pérez	jp@utp.edu.co	35
Incorrecto	Juan	Pérez	jp@utp.edu.co	

Verificador de tipo de datos

Con las bases de datos relacionales uno ya sabe qué tipo de dato esperar.

```
nombre, apellido, correo, edad
Juan, Pérez, jp@utp.edu.co, 35
35, Juan, Pérez, jp@utp.edu.co
```

Calidad de los datos

Chequeo de rangos

- Días de la semana [1 - 7] , [Lunes, Martes, ... , Domingo]
- Meses del año [1 - 12] , [Enero, Febrero, ... , Diciembre]
- Edad [0 - 120]
- Género [M / F / B / L / G / ...]

Chequeo de formatos

Cuando se sabe que una entrada particular debe cumplir un formato, las expresiones regulares son una buena herramienta a la hora de verificar los datos.

Por ejemplo, las direcciones de e-mail deben cumplir con la norma RFC 5322.

Los **códigos postales** deben cumplir también con una codificación especial. Muchas veces es preferible dejarle esta tarea a software especializado (verificador de barrios, fichas catastrales, distrito, comunas, etc.).

El dilema Britney


```
brittany spears
brittney spears
britany spears
britny spears
briteny spears
britteny spears
briney spears
brittny spears
brintey spears
britanny spears
britiny spears
britnet spears
britiney spears
britaney spears
britnay spears
brithney spears
brtiney spears
birtney spears
brintney spears
briteney spears
bitney spears
brinty spears
brittaney spears
brittnay spears
```

Imaginemos que en nuestros datos aparece un registro escrito de diversas formas, sin embargo todas esas entradas hacen referencia al MISMO REGISTRO.

¿Qué hacer con tantas entradas repetidas?

P.ej.: Corrección mediante técnicas de mínima distancia de edición

Calidad de los datos

Desambiguación de registros

Por ejemplo, nombres de países

- Irlanda
- República de Irlanda
- Eire
- Rep. de Irlanda

- Uruguay
- República Oriental del Uruguay
- Rep. del Uruguay

Apellidos

- Echeverry
- Echeverri
- Echeberri
- Etxeberri

Cuando el problema lo generan instancias conocidas como nombres de países, es relativamente fácil construir una tabla de mapeo.

```
mapeo.pais("República de Irlanda","IE");
mapeo.pais("Irlanda","IE");
mapeo.pais("Eire","IE");
mapeo.pais("Rep. de Irlanda","IE");
```

Calidad de los datos

DD/MM/YY vs MM/DD/YY vs YYYY/MM/DD etc.

12 vs 24 hour clocks

Fecha y hora

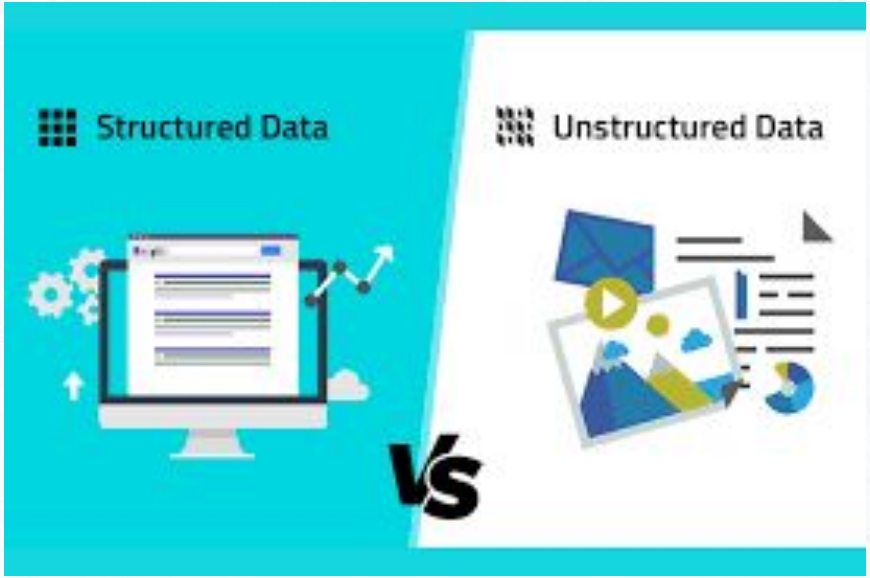
Si estamos trabajando con **series de tiempo**, los formatos de fecha y hora deben ser consistentes.

En teoría, el estándar ISO 8601 establece la norma respecto a los formatos de fecha y hora, pero cada uno puede trabajar con el que quiera (a partir del 19 de enero de 2038 se quedará obsoleto - Problema Y2038)

Nuevos estándares como el TAI (Temps Atomique International) están ajustando soluciones ante este problema.

Independientemente del lenguaje o herramienta de programación, se debe tener mucho cuidado con el manejo temporal.

https://blog.exploratory.io/5-most-common-date-time-data-wrangling-operations-in-exploratory-97b41d299934
https://towardsdatascience.com/dates-times-calendars-the-universal-source-of-data-science-trauma-92a887fdedd1



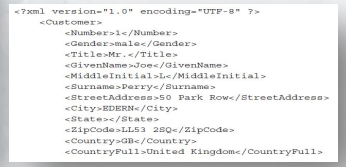
DATOS ESTRUCTURADOS VS. NO ESTRUCTURADOS

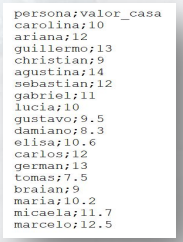
UNAL
Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales
UNIVERSIDAD NACIONAL DE COLOMBIA

Datos estructurados vs. NO estructurados

Los datos estructurados son los datos típicos que encontramos en la mayoría de bases de datos relacionales.

Estas **bases de datos** se caracterizan por tener un esquema determinado que define cómo son las tablas en las que se almacenan los datos, qué tipo de campos tienen y cómo se relacionan entre ellas.





Tenencia De Vivienda	departamento	año	valor
Propietario vivienda y terreno	Montevideo	2006	56.4
Inquilino	Montevideo	2006	20.9
Ocupante gratuito con permiso u ocupante en relación de dependencia	Montevideo	2006	12.2

Los datos que hemos visto hasta ahora son ejemplos de datos estructurados.

Datos estructurados vs. NO estructurados

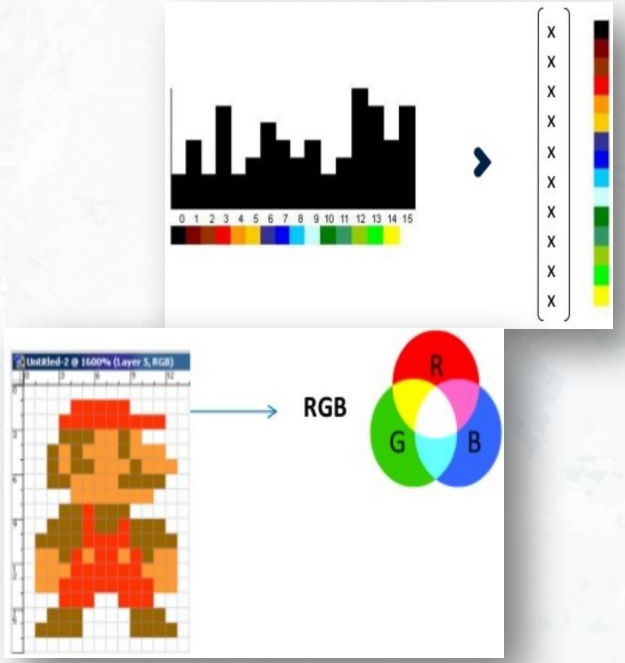
Los datos no estructurados son prácticamente todo lo demás que podamos encontrar. Se estima que estos datos suponen un 80% del volumen de todos los datos generados.

Estos datos pueden tener una estructura interna, pero no siguen ningún esquema o modelo de datos predefinido.

Ejemplos de datos no estructurados:

- Ficheros de texto (documentos de Word, PDF, presentaciones).
- Correos electrónicos (el cuerpo del mensaje es un dato no estructurado).
- Imágenes.
- Videos.
- Audios.
- Datos de sensores.

Datos estructurados vs. NO estructurados



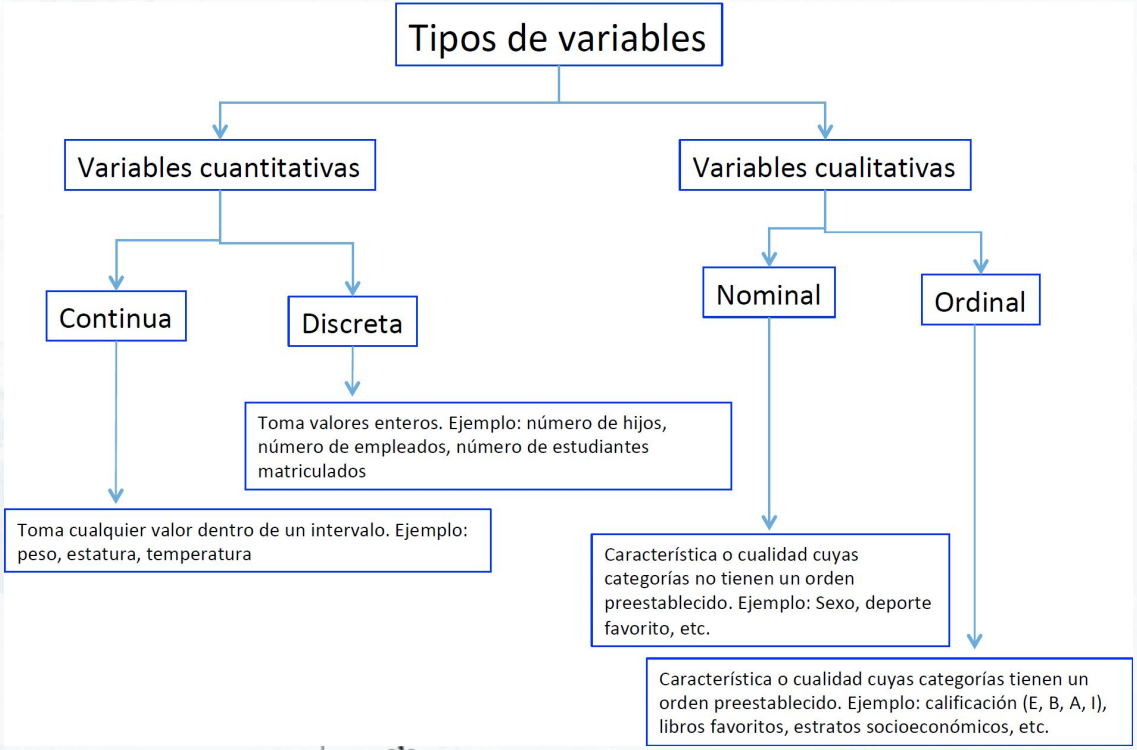
Dependiendo del tipo de datos existen distintas técnicas para extraer y procesar información.

Textos – Minería de textos.

Imágenes
- Transformación de espacios.
- Características de textura.
- Características de forma:
  * Tamaño, área, perímetro, altura, anchura.
  * Curvatura, circularidad, contorno.

Audio
- Análisis en el dominio del tiempo
- Análisis en el dominio de la frecuencia.

Tipos de variables



Tipos de variables

AÑO	NIT	NOMBRE	ID_MUNICIPIO	MUNICIPIO	ASOCIADOS	INGRESOS	VALOR_APORTE_MENSUAL	TARIFA_POR_INSTALACION
2020	900137741	ASOCIACION CABLE CIMA TV	11001	BOGOTÁ, D.C.	353	10590000	30000	30000
2020	811009220	ASOCIACION ANTENA PARABOLICA CIUDAD BOLIVAR APACIBOL	5101	CIUDAD BOLÍVAR	2198	62139343	27000	43400
2020	900000135	ASOC.USUARIOS TV SATELITE MIRAFLORES	15455	MIRAFLORES	820	13120000	16000	70000
2020	900028671	ASOCIACION DE TELEVIDENTES DE GUACHETA	25317	GUACHETÁ	1068	47098000	18000	180000
2020	900319058	ASOCIACION COMUNITARIA PARABOLICA GRANADA	5313	GRANADA	570	9788000	16000	160000
2020	832005664	ASOCIACION DE COPROPIETARIOS DE LA ANTENA PARABOLICA DE GACHANCIPA	25295	GACHANCIPÁ	1000	26066000	17000	80000
2020	811015974	ASOCIACION CIVICA LA SIERRA ACISIERRA	5585	PUERTO NARE	564	10699600	17000	30000
2020	800240299	ASOCIACION POR RECREACION Y CULTURA DE ZIPAQUIRA APRECUZ	25899	ZIPAQUIRÁ	797	16543000	19900	70000
2020	900032684	ASOCIACION DE TELEVIDENTES DE SAN MIGUEL ASOTELMI	86757	SAN MIGUEL	224	3360000	15000	30000
2020	811010541	ASOCIACION DE USUARIOS DE LA ANTENA PARABOLICA DE SALGAR	5642	SALGAR	1540	30554665	20000	50000

¿ Qué tipos de variables tenemos en esta tabla?

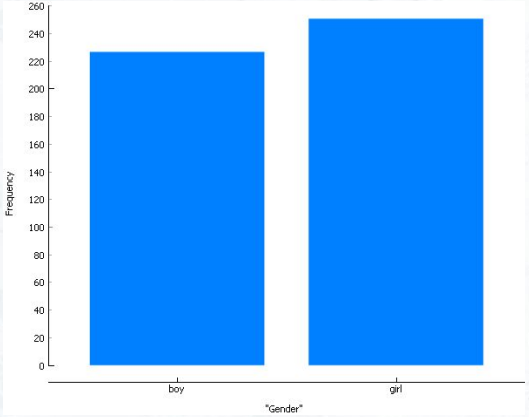
Graficando los datos

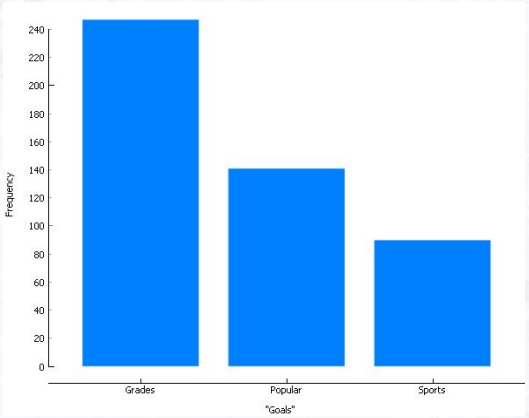
Gráfico de barras

La forma más sencilla de presentar o visualizar un conjunto de datos es por medio de una **tabla**.

Las tablas pueden ser útiles, pero no lo son para grandes conjuntos de datos, ya que es difícil entender el significado de los datos a partir de una tabla.

"género"	"objetivo"
niño	Deportes
niño	Popular
niña	Popular
niña	Popular
niña	Popular
niña	Popular
niña	Popular
niña	Notas
niña	Deportes
niña	Deportes
niña	Deportes
niña	Notas
niño	Popular
niño	Popular
niño	Popular
niña	Notas
niña	Deportes
niña	Popular
niña	Notas
niña	Popular
niña	Popular
niña	Notas
niña	Popular



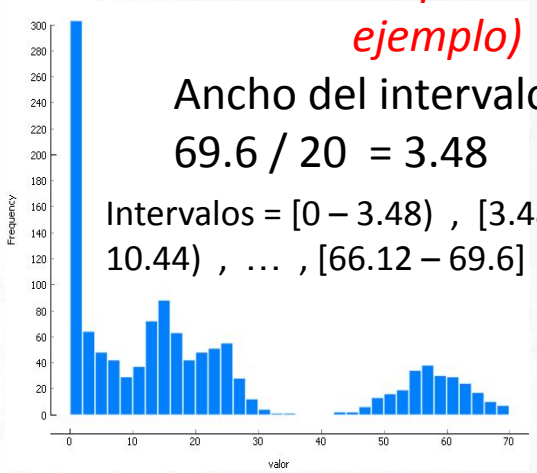


Graficando los datos

$\min_{valor} = 0$ $\max_{valor} = 69.6$

$\max_{valor} - \min_{valor} = 69.6 - 0 = 69.6$

Se quieren 20 intervalos (por ejemplo)



Ancho del intervalo =
69.6 / 20 = 3.48

Intervalos = [0 – 3.48) , [3.48 – 6.96) , [6.96 – 10.44) , … , [66.12 – 69.6]

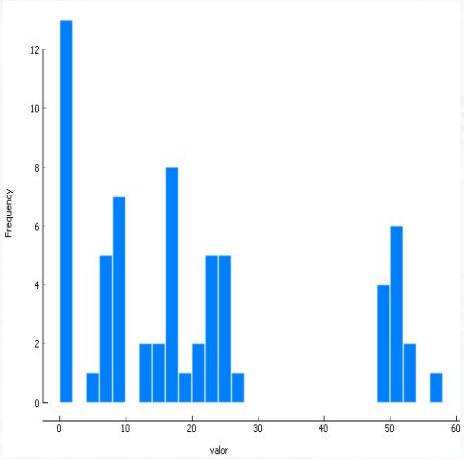
Histograma

Tenencia de vivienda	departamento	año	valor
Propietario vivienda	risaralda	2005	56,4
Inquilino	caldas	2005	20,9
Ocupante con residencia	valle	2005	12,2
Ocupante sin residencia	cauca	2005	1,3
Ocupante sin residencia	quindío	2005	9,2
Propietario vivienda	quindío	2005	54,8
Ocupante sin residencia	caldas	2005	9,3
Propietario vivienda	risaralda	2005	15,3
Inquilino	nariño	2005	20,5
Propietario vivienda	nariño	2005	0
Ocupante sin residencia	valle	2005	65,6
Propietario vivienda	caldas	2005	10,5
Ocupante sin residencia	quindío	2006	17,3
Propietario vivienda	risaralda	2006	1,7
Inquilino	valle	2006	4,9
Ocupante con residencia	cauca	2006	8,6
Ocupante sin residencia	quindío	2006	21,8
Inquilino	risaralda	2006	52,4
Ocupante sin residencia	valle	2006	72,1
Propietario vivienda	risaralda	2006	0
Inquilino	risaralda	2006	2,4
Ocupante con residencia	tolima	2006	5,6
…			

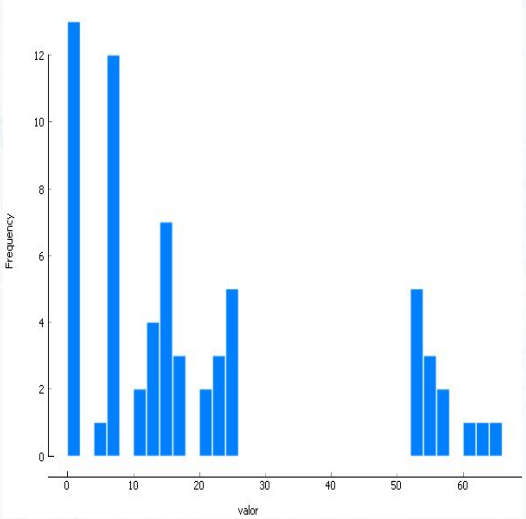
Graficando los datos

Histograma Condicional

Tenencia de vivienda	departamento	año	valor
Propietario vivienda	risaralda	2005	56,4
Inquilino	caldas	2005	20,9
Ocupante con residencia	valle	2005	12,2
Ocupante sin residencia	cauca	2005	1,3
Ocupante sin residencia	quindío	2005	9,2
Propietario vivienda	quindío	2005	54,8
Ocupante sin residencia	caldas	2005	9,3
Propietario vivienda	risaralda	2005	15,3
Inquilino	nariño	2005	20,5
Propietario vivienda	nariño	2005	0
Ocupante sin residencia	valle	2005	65,6
Propietario vivienda	caldas	2005	10,5
Ocupante sin residencia	quindío	2006	17,3
Propietario vivienda	risaralda	2006	1,7
Inquilino	valle	2006	4,9
Ocupante con residencia	cauca	2006	8,6
Ocupante sin residencia	quindío	2006	21,8
Inquilino	risaralda	2006	52,4
Ocupante sin residencia	valle	2006	72,1
Propietario vivienda	risaralda	2006	0
Inquilino	risaralda	2006	2,4
Ocupante con residencia	tolima	2006	5,6
...			

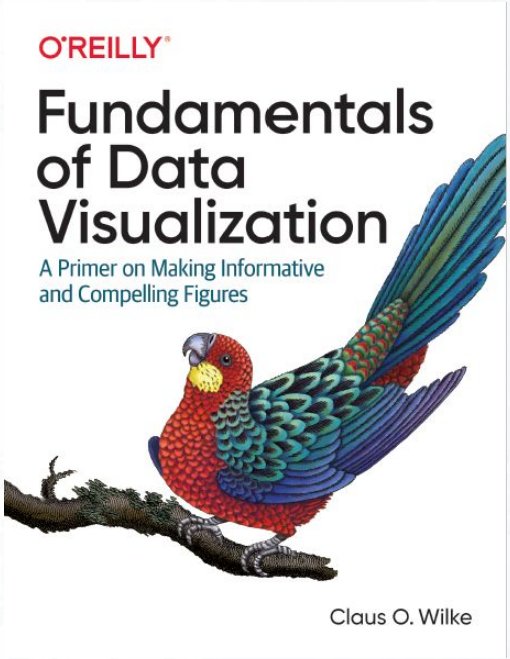


Histograma Risaralda



Histograma Quindío

Entre paréntesis (visualización de datos)



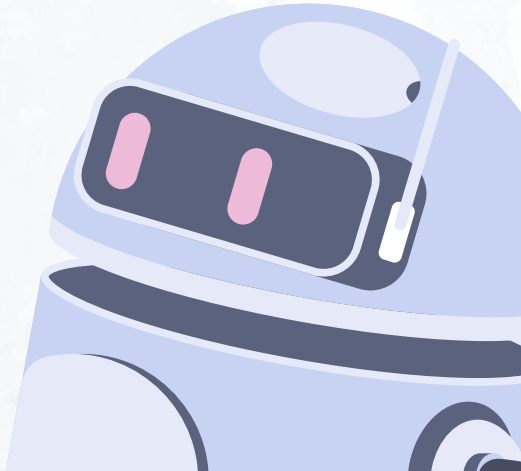
Type of variable	Examples	Appropriate scale	Description
Quantitative/ numerical continuous	1.3, 5.7, 83, 1.5×10^{-2}	Continuous	Arbitrary numerical values. These can be integers, rational numbers, or real numbers.
Quantitative/ numerical discrete	1, 2, 3, 4	Discrete	Numbers in discrete units. These are most commonly but not necessarily integers. For example, the numbers 0.5, 1.0, 1.5 could also be treated as discrete if intermediate values cannot exist in the given dataset.
Qualitative/ categorical unordered	dog, cat, fish	Discrete	Categories without order. These are discrete and unique categories that have no inherent order. These variables are also called *factors*.
Qualitative/ categorical ordered	good, fair, poor	Discrete	Categories with order. These are discrete and unique categories with an order. For example, "fair" always lies between "good" and "poor." These variables are also called *ordered factors*.
Date or time	Jan. 5 2018, 8:03am	Continuous or discrete	Specific days and/or times. Also generic dates, such as July 4 or Dec. 25 (without year).
Text	The quick brown fox jumps over the lazy dog.	None, or discrete	Free-form text. Can be treated as categorical if needed.

Tarea - Consulta

- Descriptores de datos (media, desviación estándar, varianza, mediana)
- Rangos intercuartiles
- Box Plots
- Normalización de datos

Gracias !

dfcollazosh@unal.edu.co



UNAL

Departamento de Eléctrica, Electrónica y Computación
Facultad de Ingeniería y Arquitectura
Sede Manizales

UNIVERSIDAD
NACIONAL
DE COLOMBIA