Exploration and Data Analysis

Arturo Valdivia Glez.

2023

Agenda

01. Entorno de trabajo

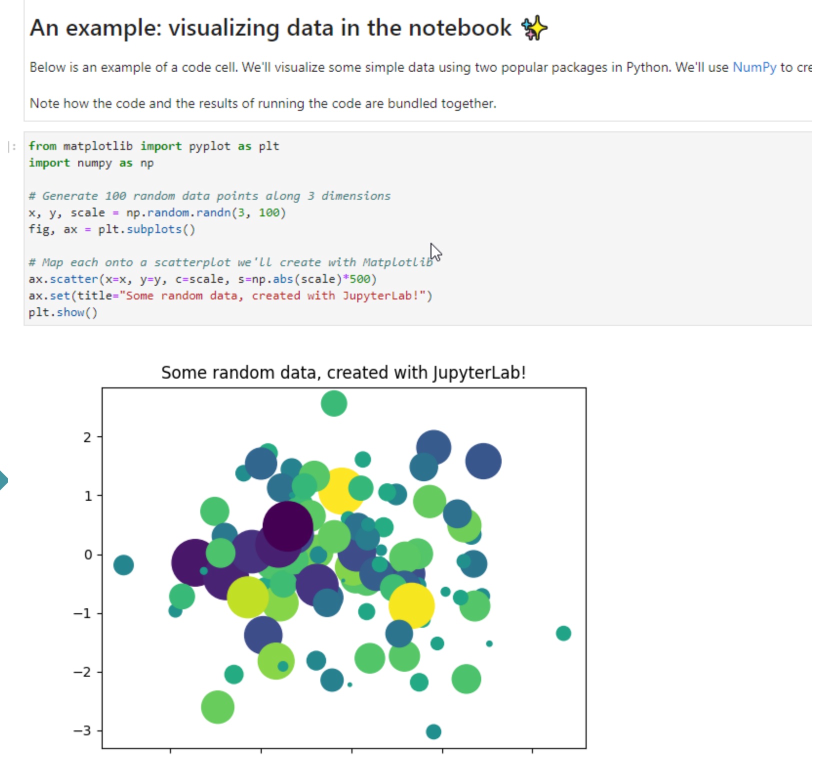
02. Google Colab

03. Dataset

04. Pandas y los DF

05. Estadística Aplicada

Jupyther LAB



An example: visualizing data in the notebook ✨

Below is an example of a code cell. We'll visualize some simple data using two popular packages in Python. We'll use NumPy to cre

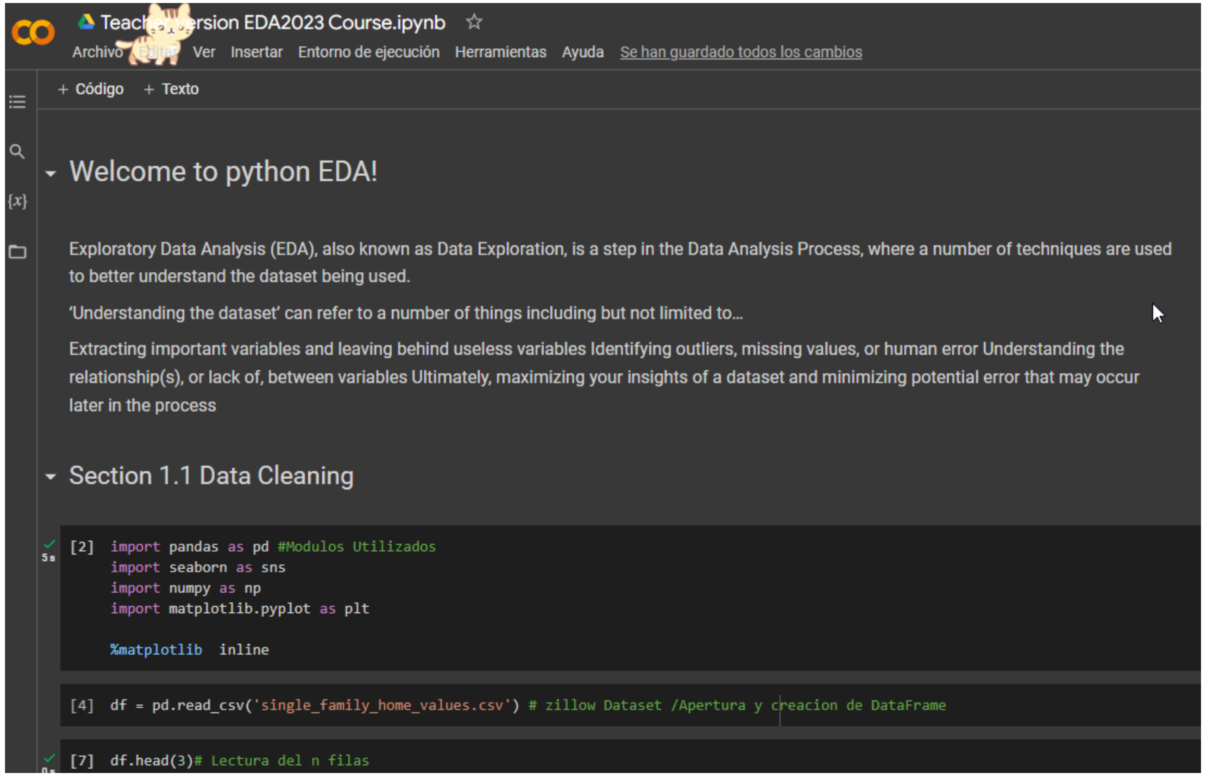
Note how the code and the results of running the code are bundled together.

```
from matplotlib import pyplot as plt
import numpy as np

# Generate 100 random data points along 3 dimensions
x, y, scale = np.random.randn(3, 100)
fig, ax = plt.subplots()

# Map each onto a scatterplot we'll create with Matplotlib
ax.scatter(x=x, y=y, c=scale, s=np.abs(scale)*500)
ax.set(title="Some random data, created with JupyterLab!")
plt.show()
```

Google Colab



Welcome to python EDA!

Exploratory Data Analysis (EDA), also known as Data Exploration, is a step in the Data Analysis Process, where a number of techniques are used to better understand the dataset being used.

'Understanding the dataset' can refer to a number of things including but not limited to...

Extracting important variables and leaving behind useless variables Identifying outliers, missing values, or human error Understanding the relationship(s), or lack of, between variables Ultimately, maximizing your insights of a dataset and minimizing potential error that may occur later in the process

Section 1.1 Data Cleaning

```
[2] import pandas as pd #Modulos Utilizados
    import seaborn as sns
    import numpy as np
    import matplotlib.pyplot as plt

    %matplotlib  inline
```

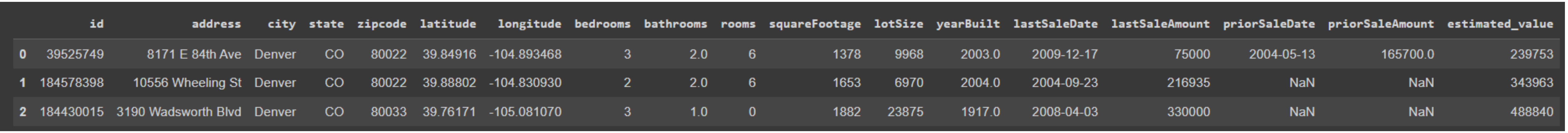
```
[4] df = pd.read_csv('single_family_home_values.csv') # zillow Dataset /Apertura y creacion de DataFrame
```

```
[7] df.head(3)# Lectura del n filas
```

https://www.youtube.com/watch?v=BE7QneKWkPs&t=43s

https://www.youtube.com/watch?v=8VFYs3Ot_aA&t=447s

Dataset



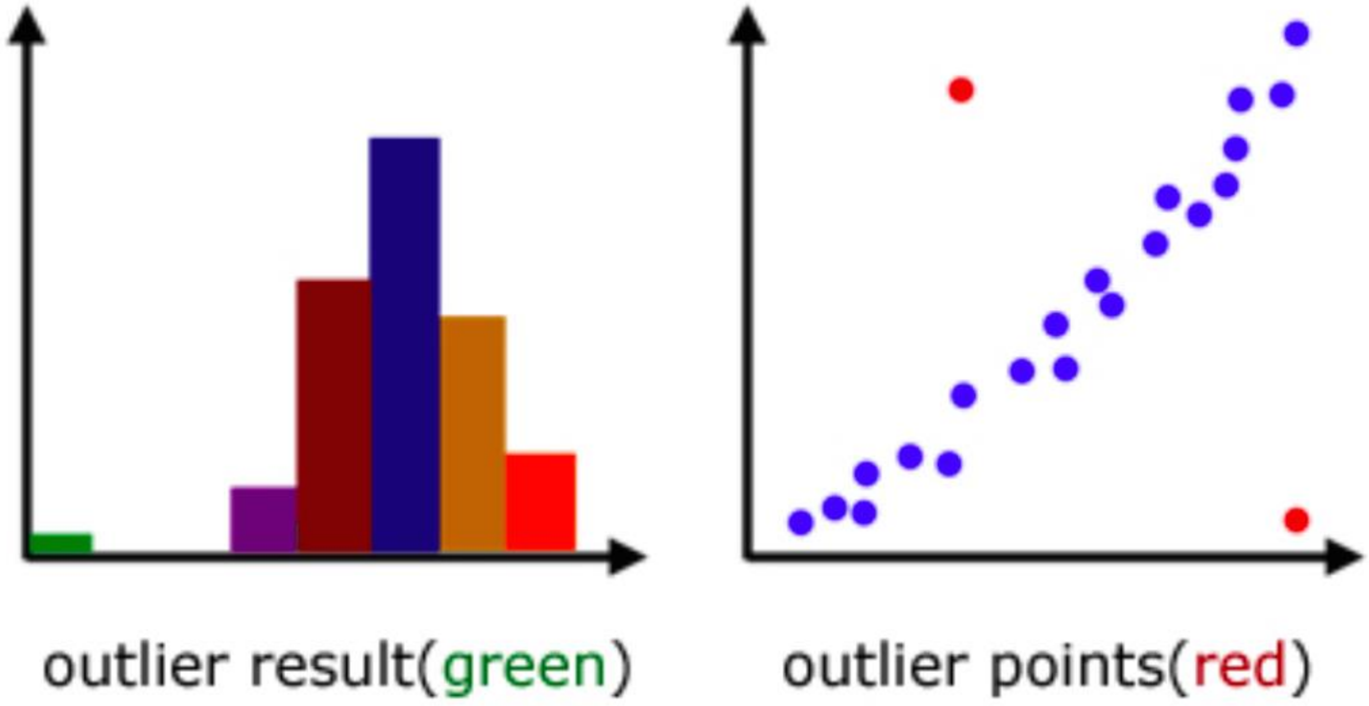
	id	address	city	state	zipcode	latitude	longitude	bedrooms	bathrooms	rooms	squareFootage	lotSize	yearBuilt	lastSaleDate	lastSaleAmount	priorSaleDate	priorSaleAmount	estimated_value
0	39525749	8171 E 84th Ave	Denver	CO	80022	39.84916	-104.893468	3	2.0	6	1378	9968	2003.0	2009-12-17	75000	2004-05-13	165700.0	239753
1	184578398	10556 Wheeling St	Denver	CO	80022	39.88802	-104.830930	2	2.0	6	1653	6970	2004.0	2004-09-23	216935	NaN	NaN	343963
2	184430015	3190 Wadsworth Blvd	Denver	CO	80033	39.76171	-105.081070	3	1.0	0	1882	23875	1917.0	2008-04-03	330000	NaN	NaN	488840

id	address	city	state	zipcode	latitude	longitude	bedrooms	bathrooms	rooms	squareFootage	lotSize	yearBuilt	lastSaleDate	lastSaleAmount	priorSaleDate	priorSaleAmount	estimated_value

EDA

EDA (Exploración y Análisis de Datos) es un proceso de análisis de datos que se utiliza para entender mejor la naturaleza de los datos y las relaciones entre las diferentes variables en un conjunto de datos.

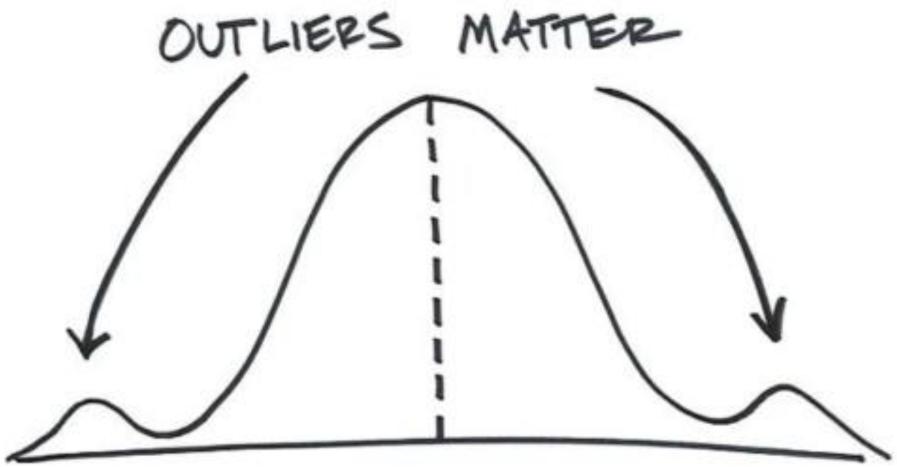
El objetivo de EDA es descubrir patrones, identificar valores atípicos (outliers), detectar posibles errores o inconsistencias en los datos, y determinar qué variables son importantes y cómo se relacionan entre sí.



EDA

- Algunas de las técnicas que se utilizan en EDA incluyen gráficos, estadísticas descriptivas, análisis de correlación, técnicas de visualización de datos, y herramientas de minería de datos.

- En resumen, EDA es una fase crucial en el proceso de análisis de datos, ya que permite a los analistas de datos obtener una mejor comprensión de los datos y prepararlos para el modelado y análisis más profundos.



NAN

Not a Number

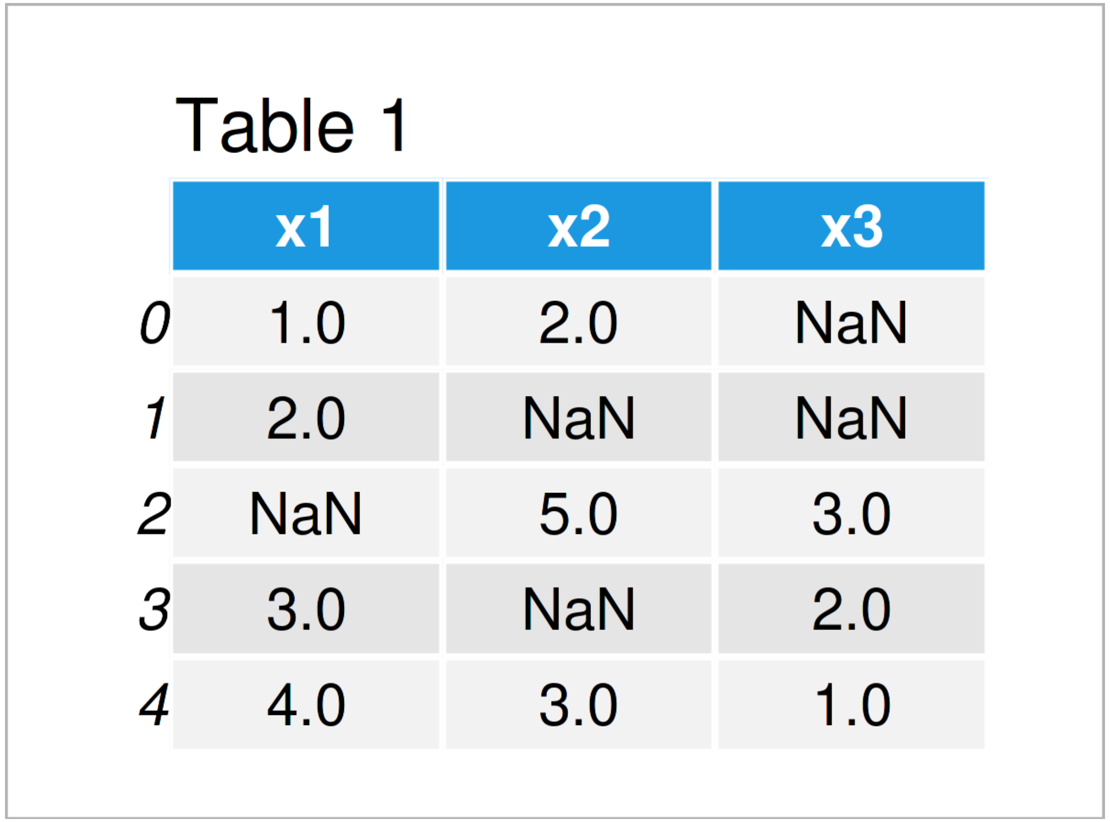


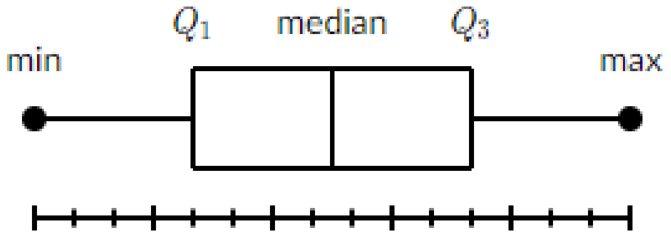
Table 1

	x1	x2	x3
0	1.0	2.0	NaN
1	2.0	NaN	NaN
2	NaN	5.0	3.0
3	3.0	NaN	2.0
4	4.0	3.0	1.0

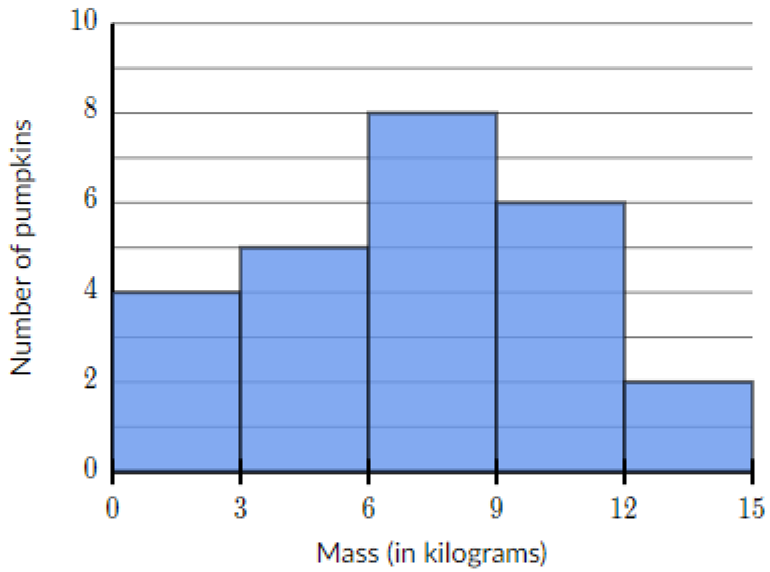
What is a box and whisker plot?

A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.

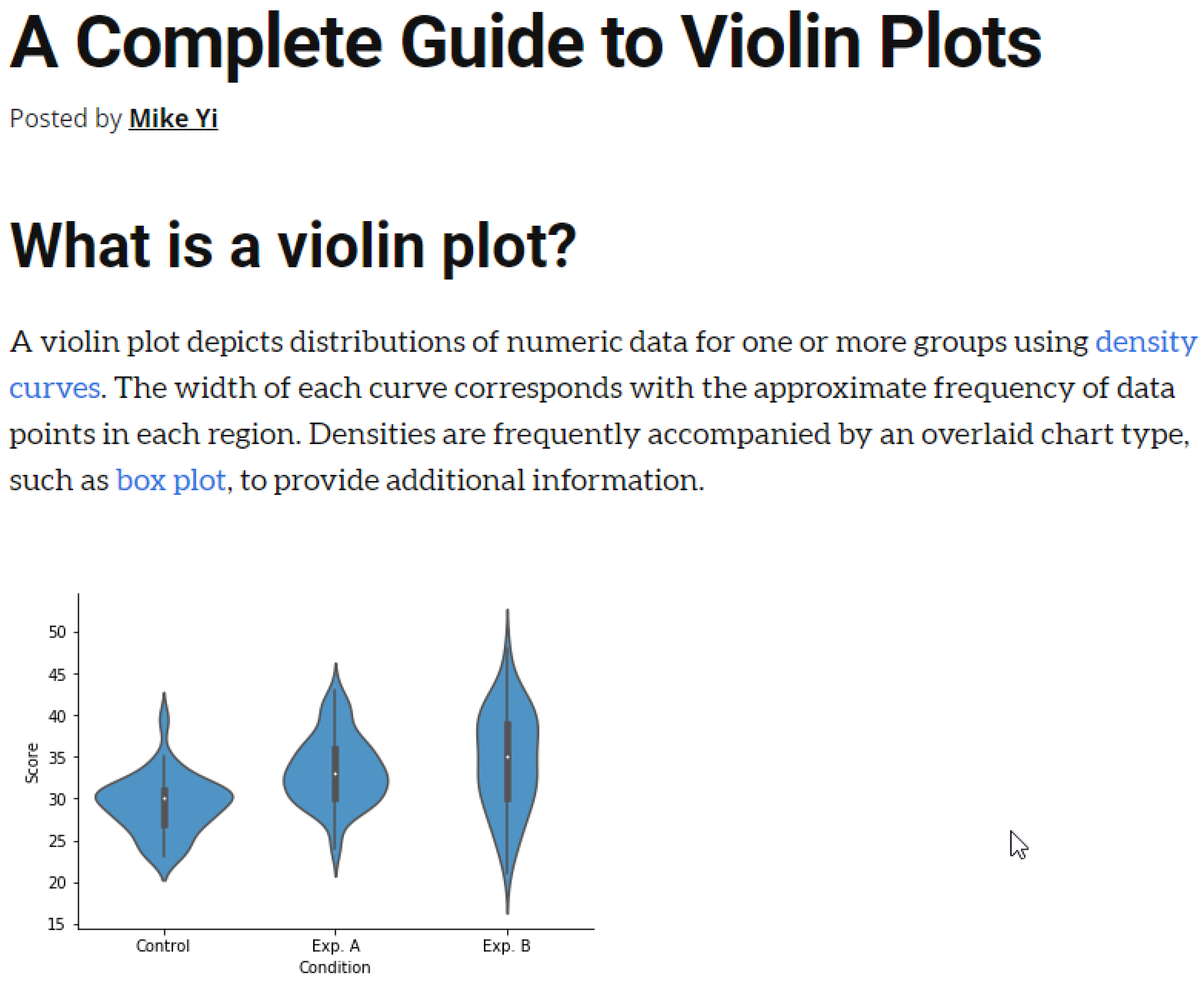
In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.



HISTOGRAM



A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data.

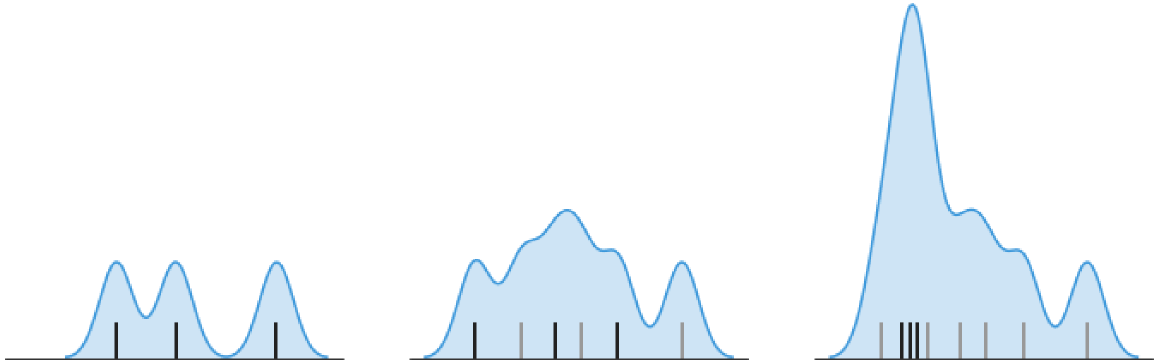


A Complete Guide to Violin Plots

Posted by **Mike Yi**

What is a violin plot?

A violin plot depicts distributions of numeric data for one or more groups using density curves. The width of each curve corresponds with the approximate frequency of data points in each region. Densities are frequently accompanied by an overlaid chart type, such as box plot, to provide additional information.



Kernel density estimation is best used when a fair amount of data is available, resulting in more stable density estimates. With few data points available, it can be easy to be misled by the smoothness of the curve or the length of the tails past the largest and smallest points.

In a violin plot, individual density curves are built around center lines, rather than stacked on baselines. Other than this difference in display pattern, curves in a violin plot follow the exact same construction and interpretation.

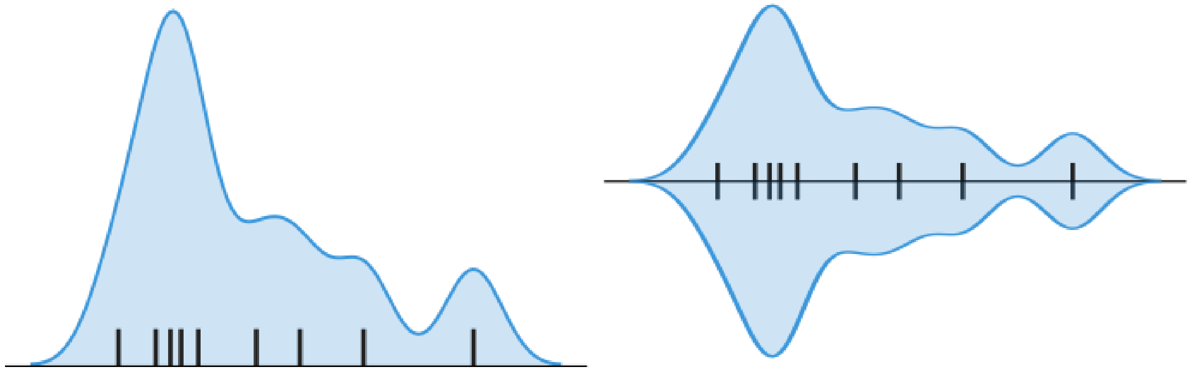


Tabla de crecimiento por sector

	P1	T2	T3	T4
Serie 1	4.3	2.5	3.5	4.5
Serie 2	2.4	4.4	1.8	2.8
Serie 3	2	2	3	5

Referencias

ttps://chartio.com/learn/charts/violin-plot-complete-guide/

ttps://www.khanacademy.org/math/statistics-probability

ttps://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e

Gracias