PI DATA STRATEGY & CONSULTING

VISUALIZACIÓN EN PYTHON

PI DATA STRATEGY & CONSULTING

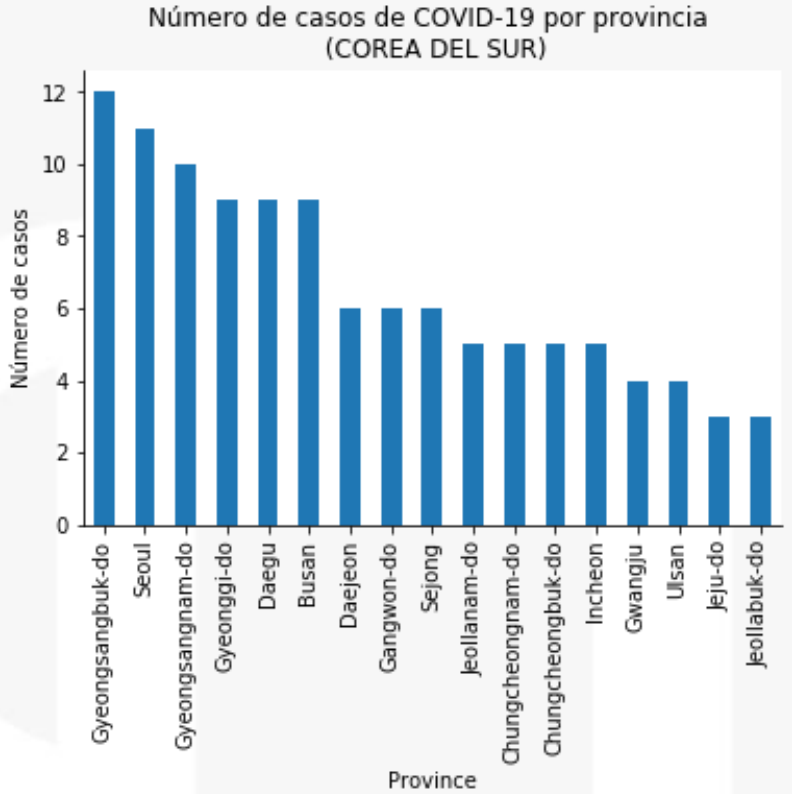
Visualización en Python

Johanna Frau
Data Scientist
Pi Data Strategy & Consulting

GRÁFICOS

BARPLOT (Gráfico de barras)

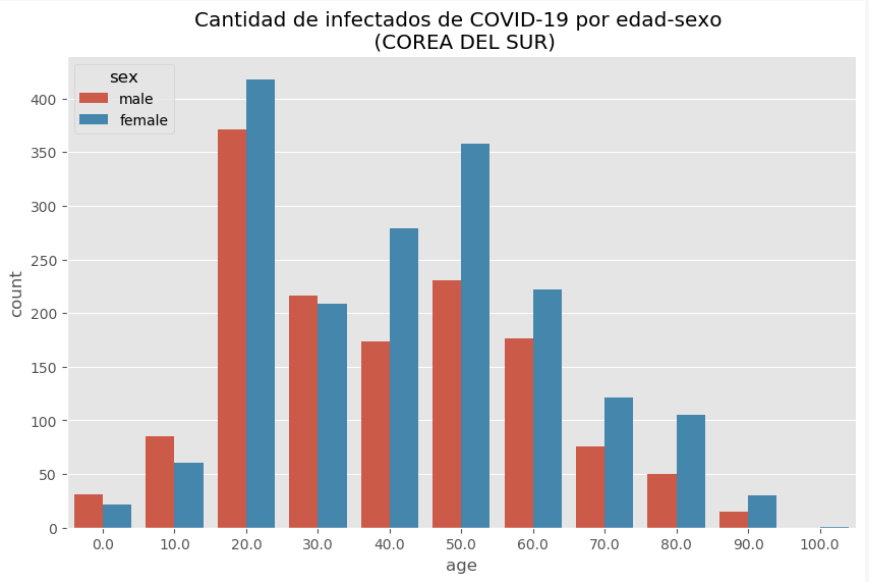
- Muestra la relación entre una variable numérica y una variable categórica usando barras.
- Cada valor categórico se asocia a una barra y la longitud de la barra se corresponde con su valor numérico.
- Se utiliza cuando se quiere mostrar una cierta distribución de los datos o realizar una comparación de valores de cierta variable a lo largo de diferentes subgrupos.
- La variable primaria suele ser la variable categórica.



GROUPED BARPLOT (Gráfico de barras agrupadas)

- Considera una segunda variable categórica que se muestra como barras agrupadas una al lado de otra dentro de cada nivel de la primera variable categórica.

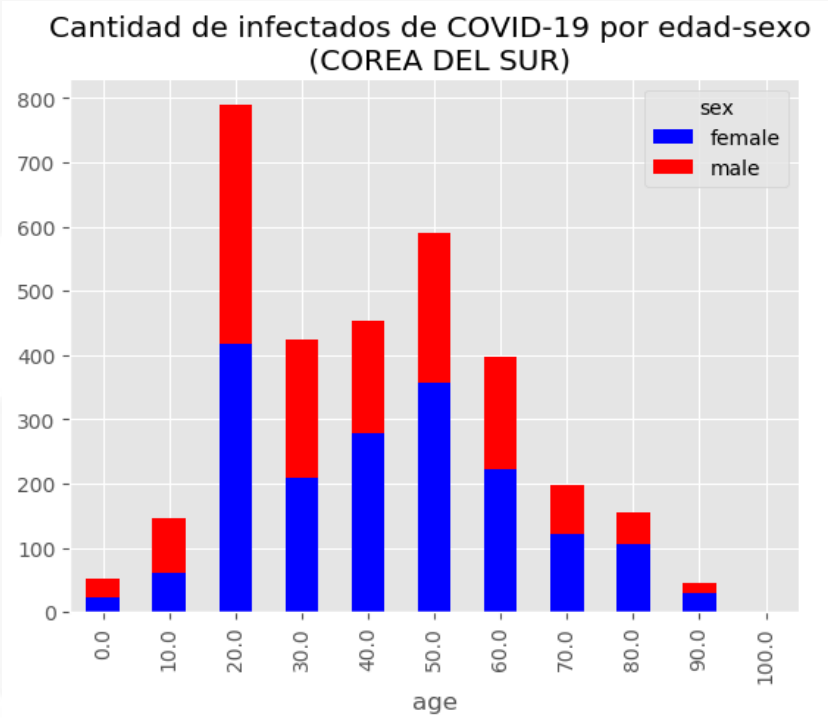
- Es usado cuando se quiere observar como cambia la segunda variable categórica dentro de la primera (***comparación dentro de un grupo***) o bien cuando se quiere observar la variación de la primera variable a lo largo de diferentes niveles de la segunda (***comparación entre grupos***).



STACKED BARPLOT (Gráfico de barras apiladas)

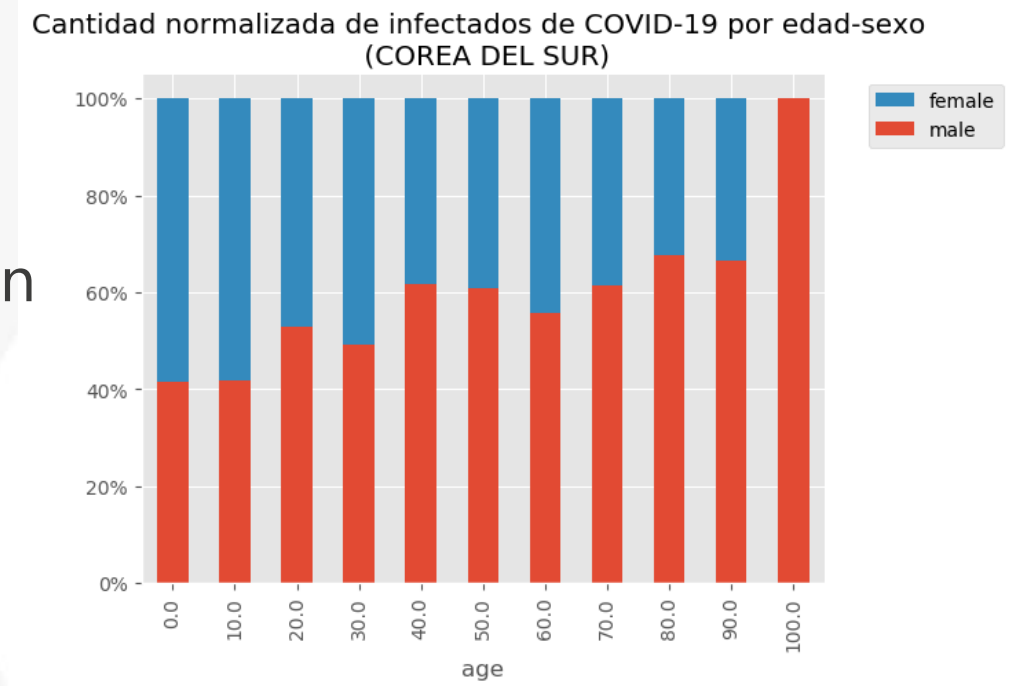
- También considera una segunda variable categórica en el análisis. En este caso, cada barra se divide en un número de subbarras apiladas que corresponden a niveles en la segunda variable categórica.

- Es útil para observar la descomposición relativa de cada barra primaria respecto a la segunda categoría considerada.



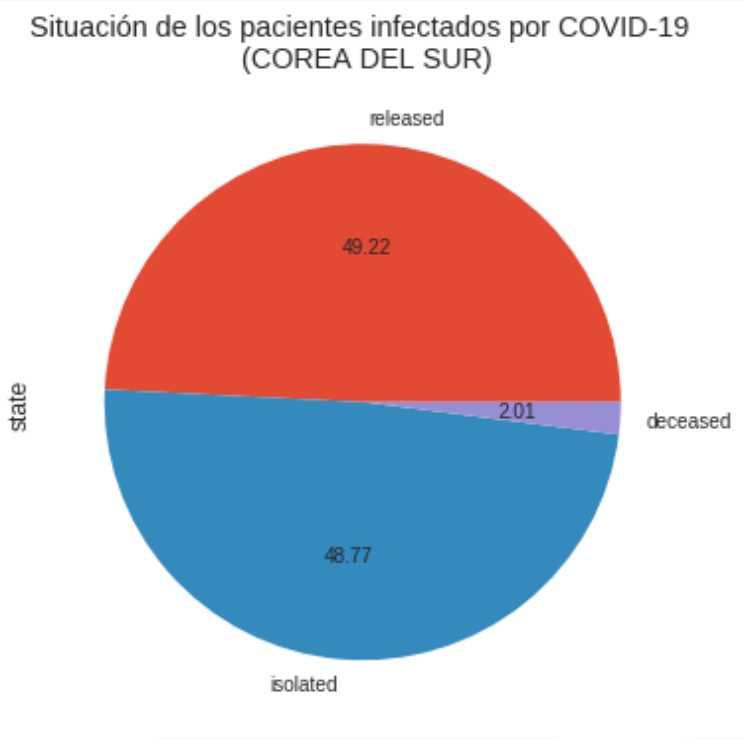
PERCENTAGE STACKED BARPLOT (Gráfico de barras apiladas por porcentajes)

- Esta versión utiliza porcentajes (o frecuencias relativas) en lugar de valores con el objetivo de observar la distribución relativa de la segunda categoría. Por esta razón todas las barras poseen la misma altura.



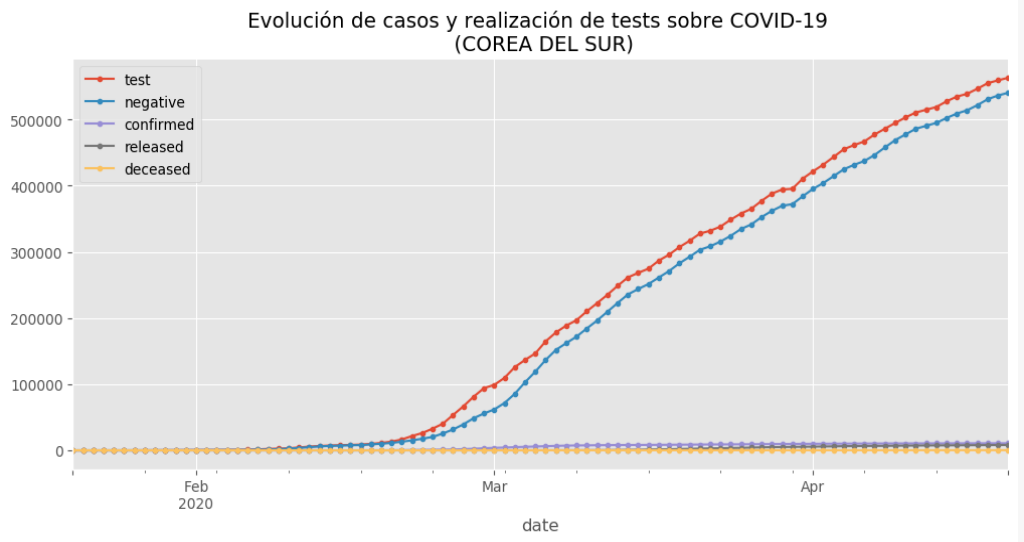
PIEPLOT (Gráfico de torta)

- Es utilizado para comparar la distribución de valores o categorías dentro de una sola categoría o grupo.
- No es recomendable para comparar distribuciones entre diferentes grupos.
- Tampoco se lo recomienda cuando una categoría puede tomar muchos valores diferentes pues dificulta la separación entre porciones a nivel visual.



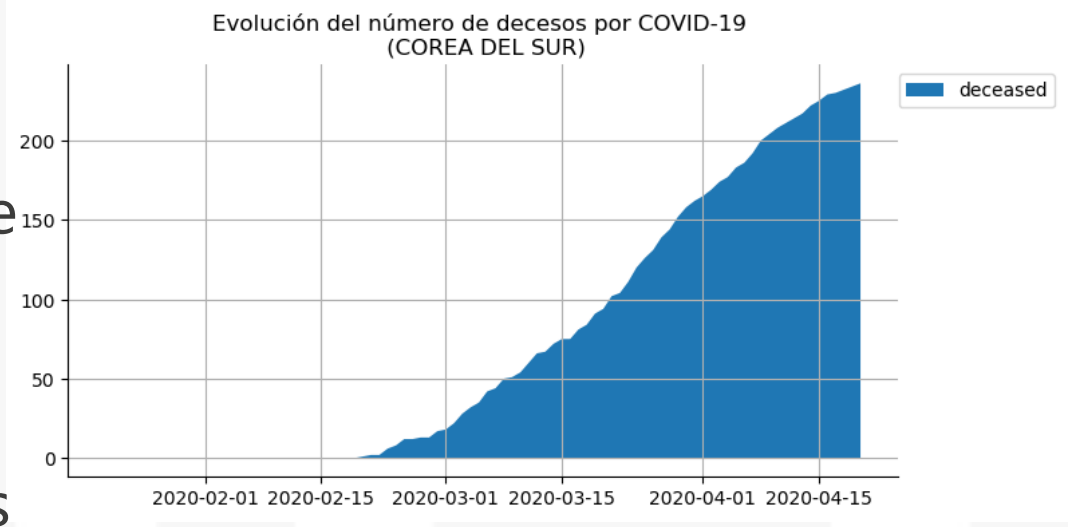
LINEPLOT (Gráfico de línea)

- Es utilizado para mostrar o enfatizar los cambios de valores de una variable con respecto a una segunda variable continua.
- En el eje horizontal la variable, además de ser continua tiene que tener un intervalo regular de medida. Conmúmente es temporal, generando observaciones por minutos, horas, días, meses, etc.
- El eje vertical usualmente es un valor estadístico como suma de frecuencias, promedios, etc.
- Es posible dibujar varias líneas en un mismo gráfico para comparar tendencias entre series.



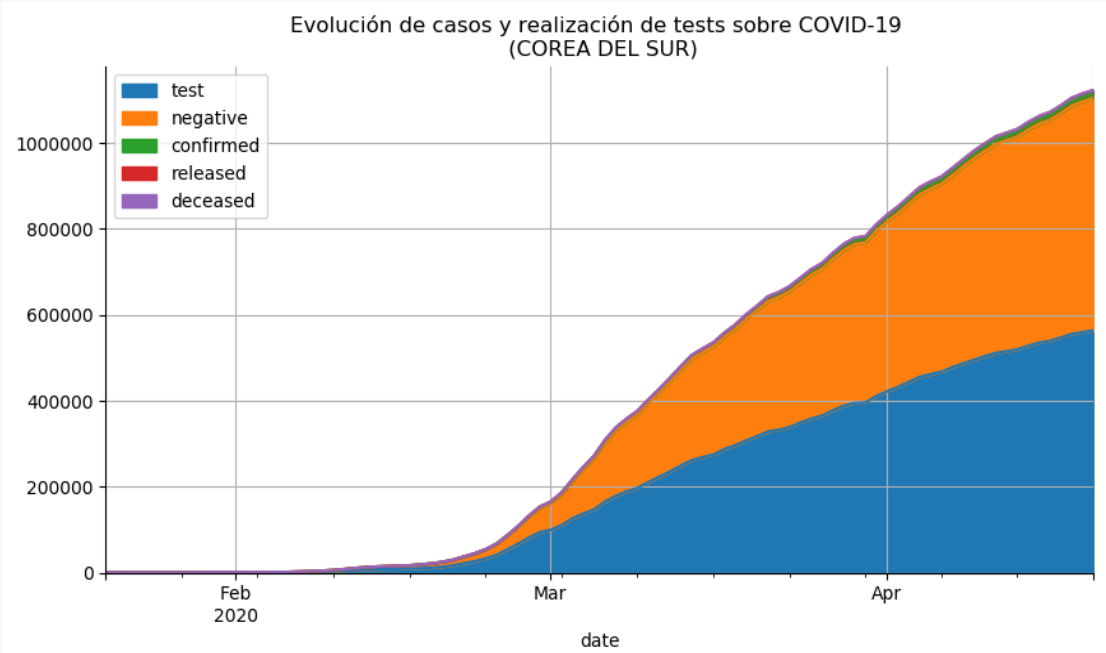
AREA CHART (Gráfico de área)

- Combina un gráfico de líneas y un gráfico de barras para mostrar como cambian los valores numéricos de un grupo sobre la progresión de una segunda variable que usualmente es el tiempo.
- Uno de los mayores usos es para representar totales acumulados (porcentajes) o tendencias sobre el tiempo.



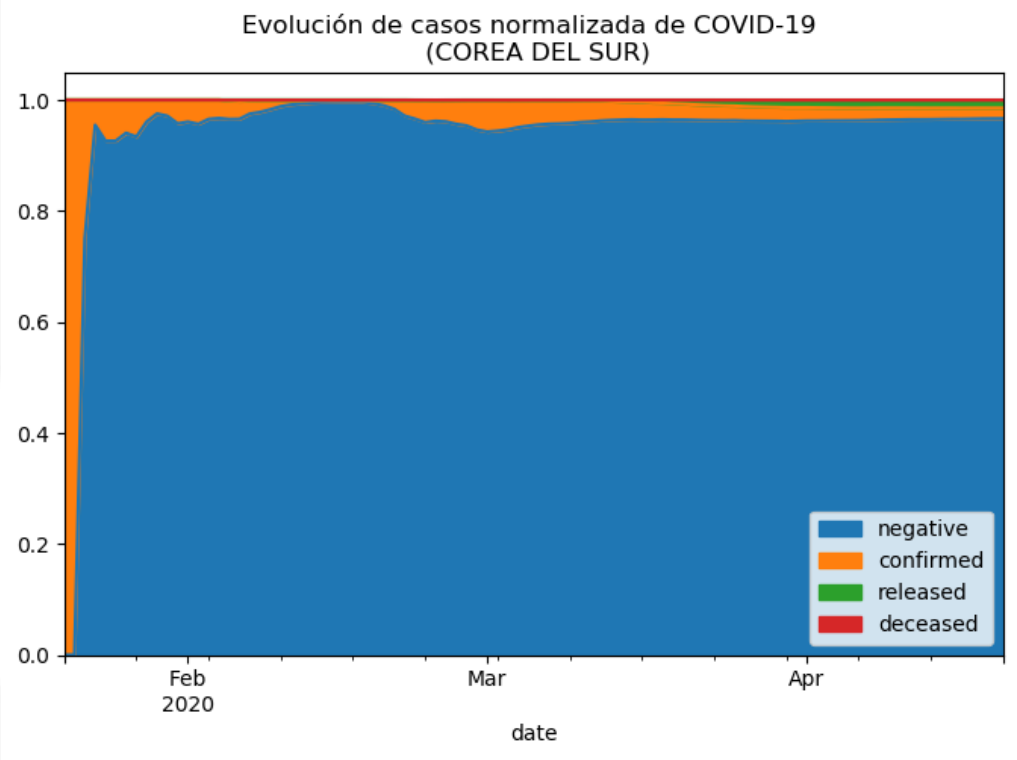
STACKED AREA CHART (Gráfico de área apilado)

- Es utilizado para representar la evolución a lo largo del tiempo para diferentes grupos. Cada grupo es mostrado uno encima del otro.
- Con este gráfico la contribución de cada grupo a la cantidad total es fácil de observar. No es apropiado para estudiar la evolución de cada grupo individual (en este caso es preferible el gráfico de línea).



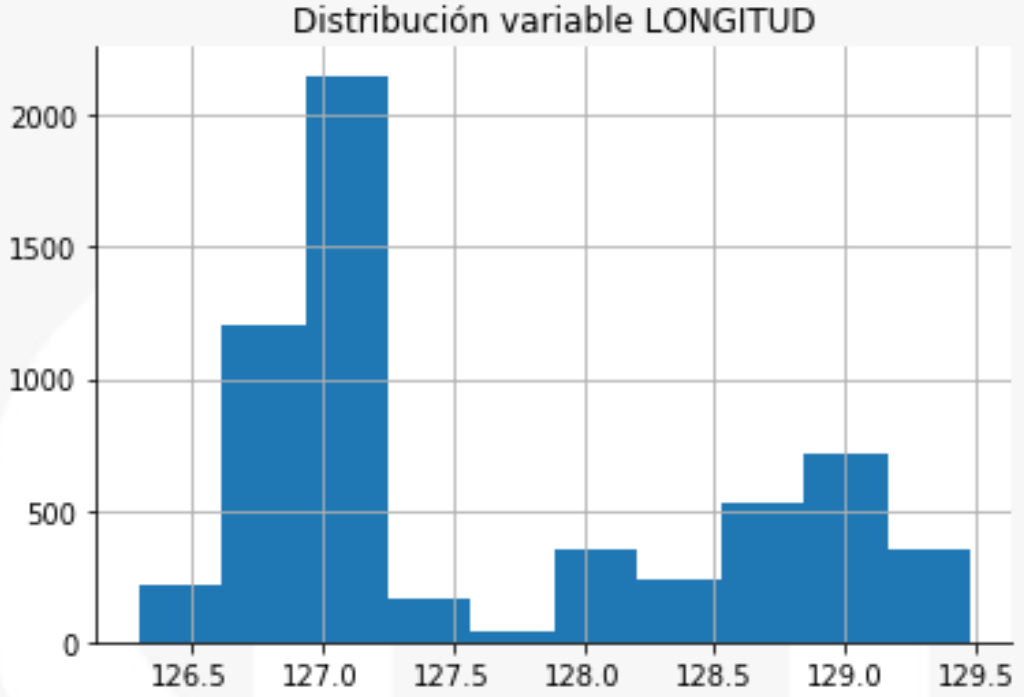
PERCENTAGE STACKED AREA CHART (Gráfico de área apilado por porcentajes)

- En este caso, cada grupo es normalizado sobre los puntos del tiempo.
- Este tipo de gráfico pierde información acerca de la tendencia de los absolutos totales pero ayuda a realizar comparaciones de contribuciones relativas entre los grupos



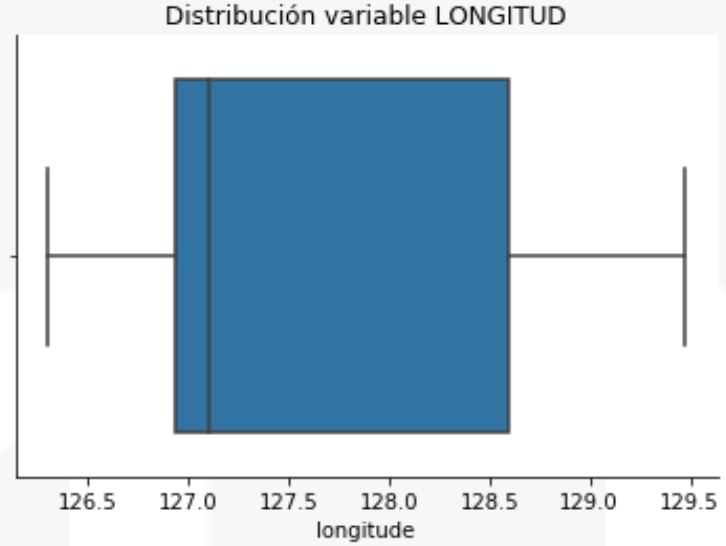
HISTOGRAMS

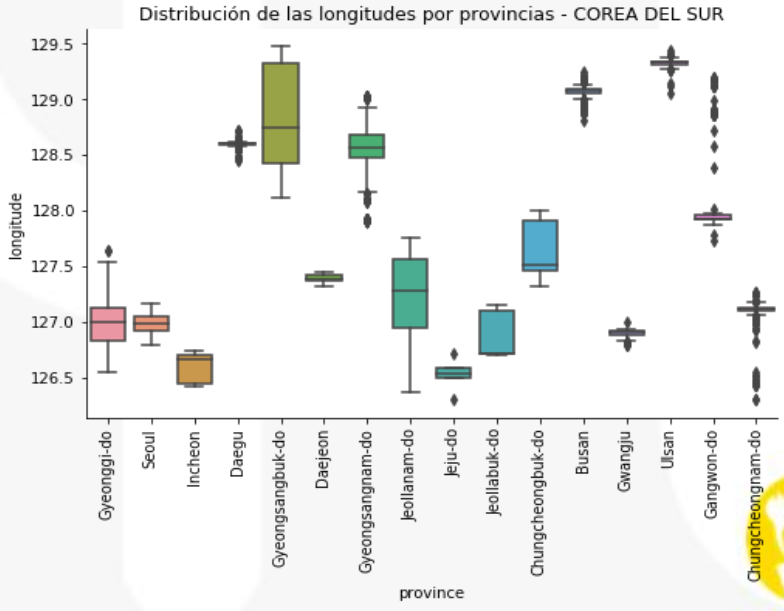
- Muestra la distribución de los valores de una variable numérica como una serie de barras una al lado de la otra.
- Cada barra generalmente cubre un rango de valores numéricos llamados *bines* o clases. La altura de las barras es la frecuencia de los puntos de datos dentro de cada bin.
- Son útiles para mostrar las distribuciones generales de las *variables numéricas* dentro de un conjunto de datos.



BOXPLOTS (Gráficos de cajas)

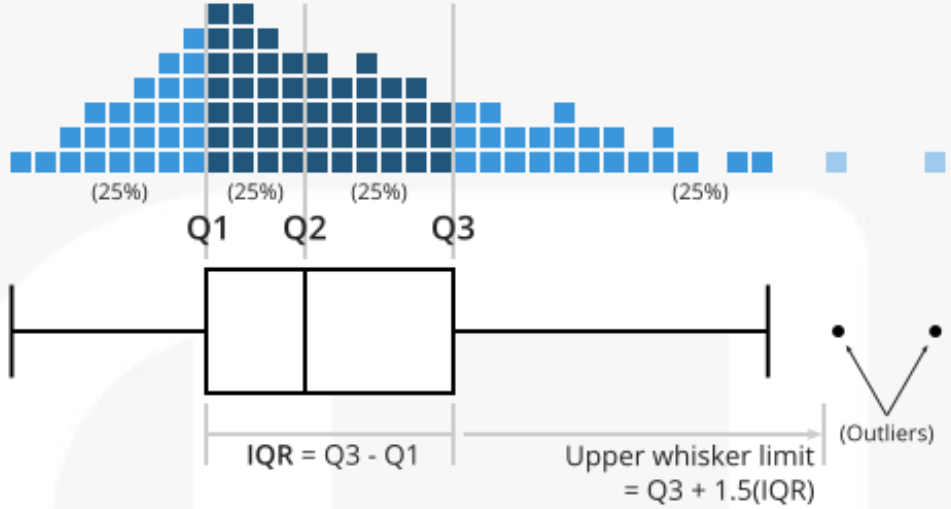
- Utiliza cajas y líneas para representar las distribuciones de *datos numéricos* de un grupo (o más de un grupo).
- Los límites de la caja indican el rango del 50% central de los datos.
- Proporciona un alto nivel de información ofreciendo información general acerca de la simetría de los grupos de datos, varianza y outliers.
- Ideal para comparar distribuciones entre diferentes grupos.
- Se pierde el detalle de la forma que tiene la distribución de los datos (modas, picos y vallles, etc.)





BOXPLOTS (Gráficos de cajas)

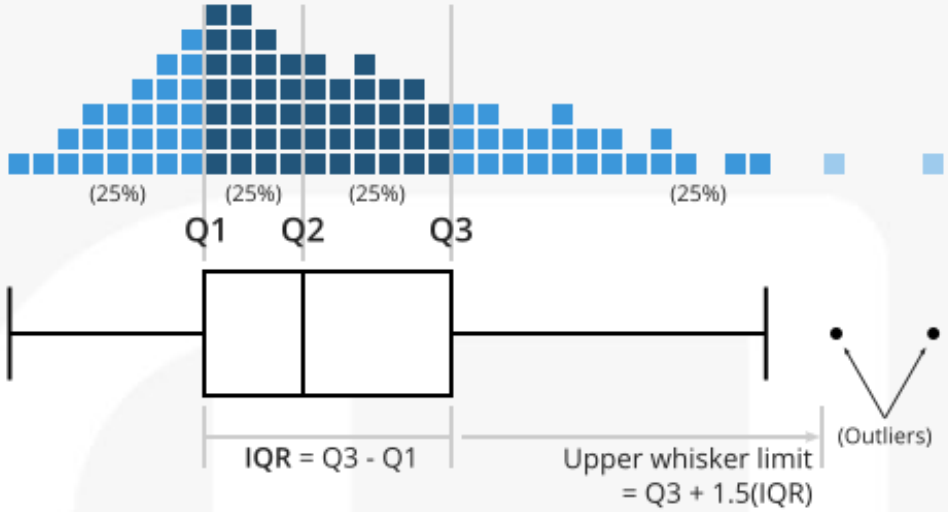
- El primer cuartil (Q1) es mayor que el 25% de los datos y menor que el otro 75%.
- El segundo cuartil (Q2) se encuentra en el medio, dividiendo los datos por la mitad. Q2 también se conoce como la ***mediana***.
- El tercer cuartil (Q3) es mayor que el 75% de los datos y menor que el 25% restante.
- La distancia entre Q3 y Q1 es conocida como el *rango intercuartil (IQR)*.



BOXPLOTS (Gráficos de cajas)

- Cada whisker se extiende hasta el punto de datos más alejado en cada ala que está dentro de 1,5 veces el IQR. Cualquier punto de datos más allá de esa distancia se considera un valor atípico y se marca con un punto.

- Cuando la distribución de datos es simétrica, se puede esperar que la media sea exacta en el centro de la caja.

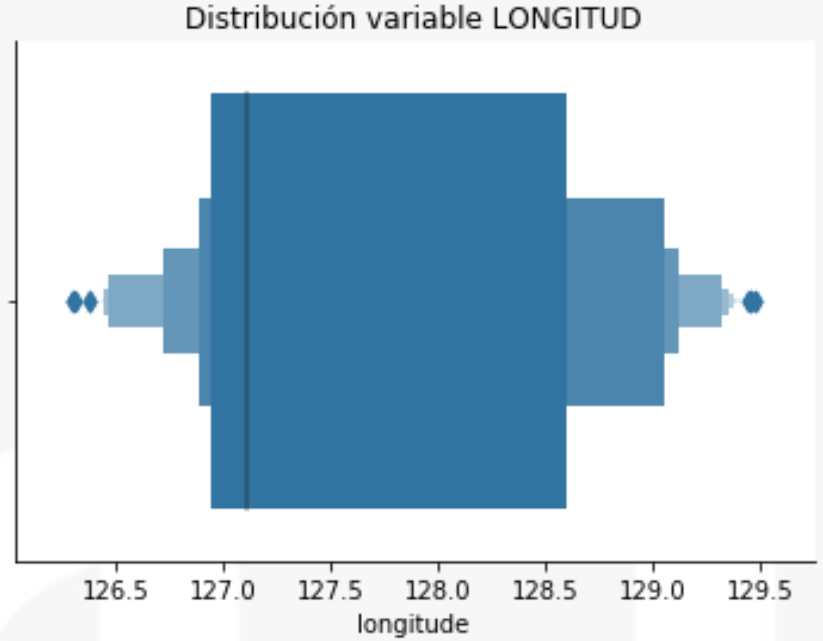


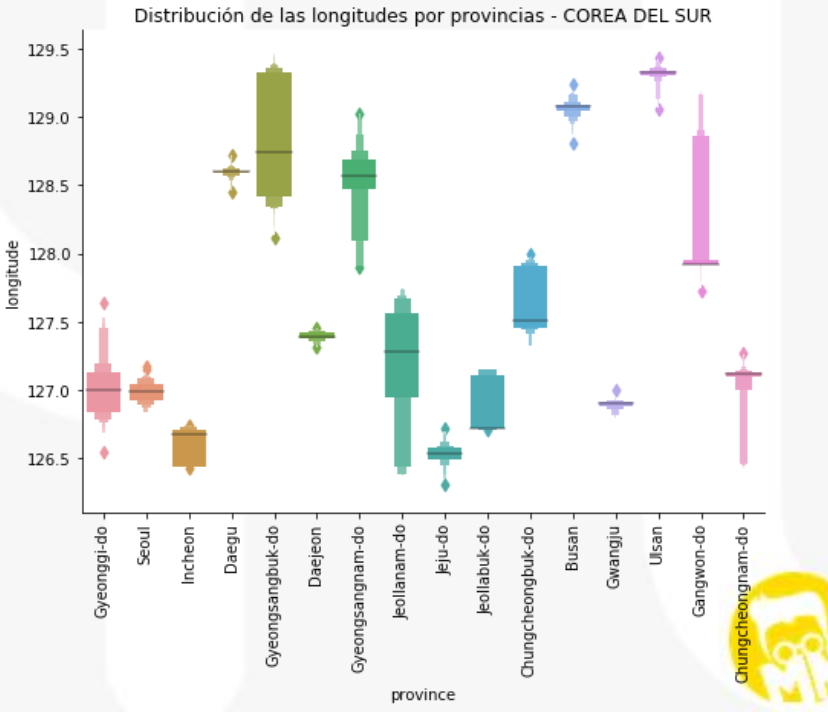
LETTER VALUE PLOT

- Es una extensión de los gráficos de cajas comunes. Utilizan múltiples cajas para encerrar cada vez más proporciones de los datos.

- La primera caja cubre el 50% central y la segunda caja se extiende para cubrir la mitad restante (75% en general, 12.5% sobrante en cada extremo). La tercer caja cubre la otra mitad del área restante (87.5% en total, 6.25% restante en cada extremo), y así sucesivamente hasta que finalice el procedimiento y los puntos restantes se marquen como valores atípicos.

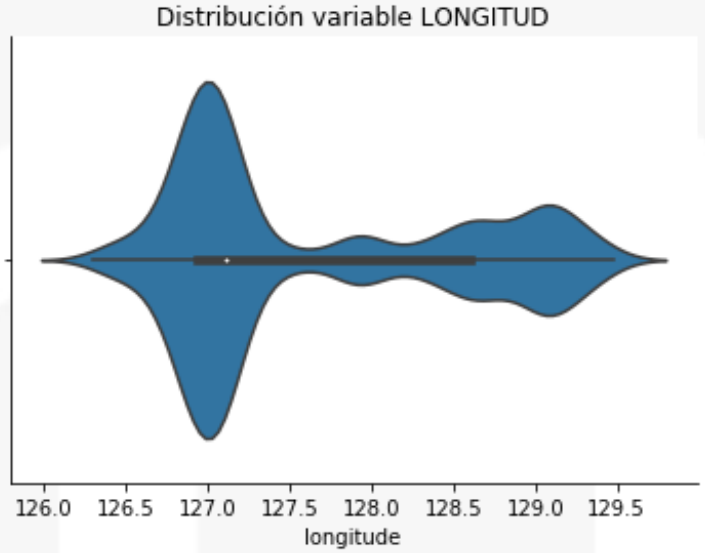
- Cuanto más datos son recogidos, más estable es la estimación de la cola del gráfico.





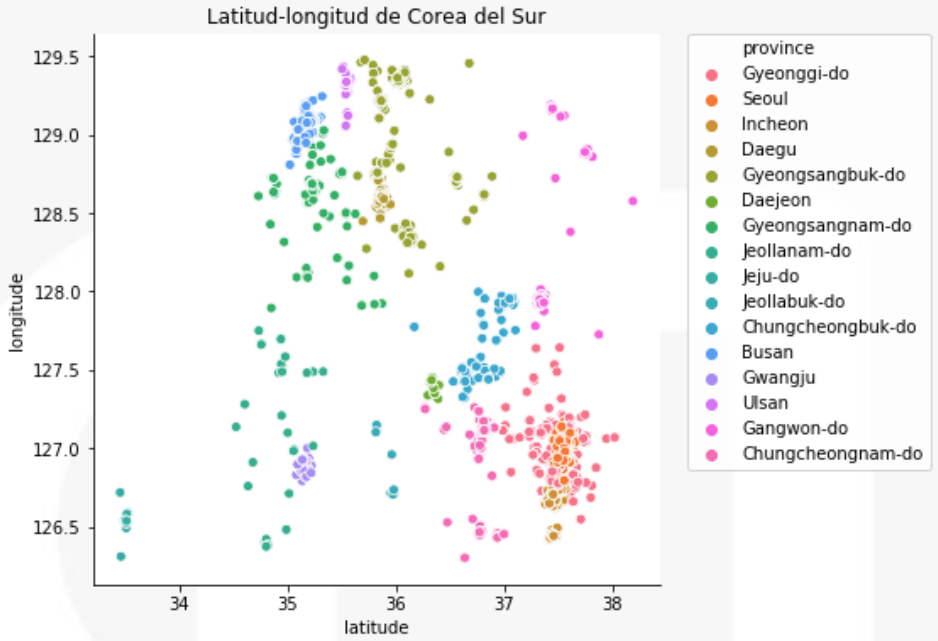
VIOLINPLOT (Gráfico de violín)

- Muestra las distribuciones de datos numéricos para uno o más grupos usando *curvas de densidad*.
- El ancho de cada curva corresponde con la frecuencia aproximada de los puntos de datos en cada región. Las densidades suelen ir acompañadas de un tipo de gráfico superpuesto, como un diagrama de caja, para proporcionar información adicional.
- Es muy útil para comparar grupos en cuanto a diferencia y similitudes (picos, valles, colas, etc).



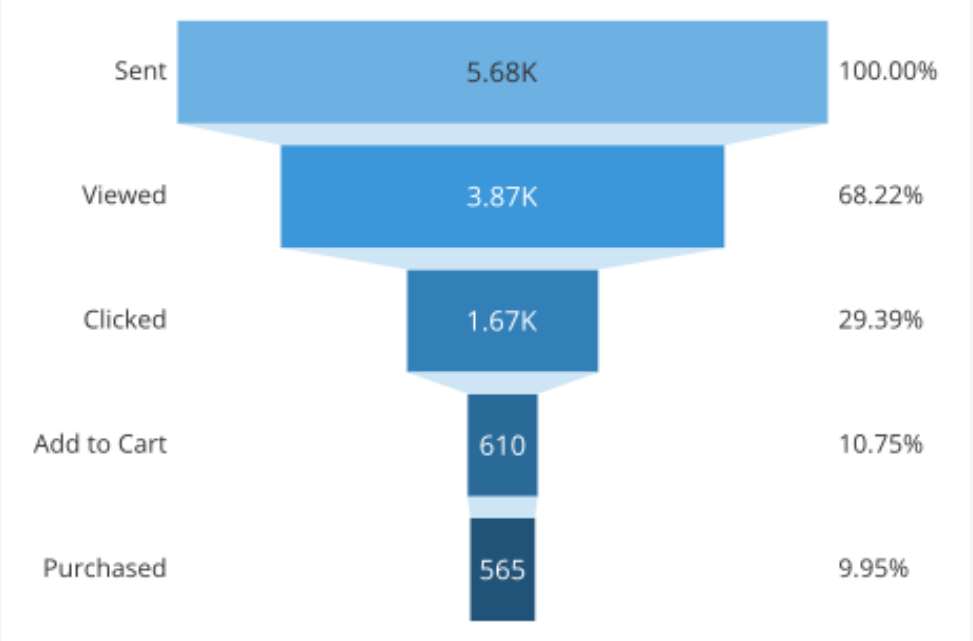
SCATTERPLOT (Gráfico de dispersión)

- Es usado para representar valores entre dos variables numéricas diferentes y analizar la relación entre ellas.
- Cada punto dentro del gráfico de dispersión tiene coordenadas x e y que representan dichas variables.
- Son útiles para encontrar patrones en los datos (agrupamientos especiales, outliers, etc.)
- Adicionalmente es posible añadir una tercera variable identificando los puntos con colores o formas.



FUNNEL CHART (Gráfico de embudo)

- Es un gráfico especial que muestra el flujo de usuarios a través de un negocio o proceso de ventas.
- Es usado cuando se necesita trackear cómo un conjunto inicial de usuarios abandonan un proceso o flujo.
- Permite visualizar la cantidad de usuarios por etapas lo cual ayuda a la toma de decisiones.



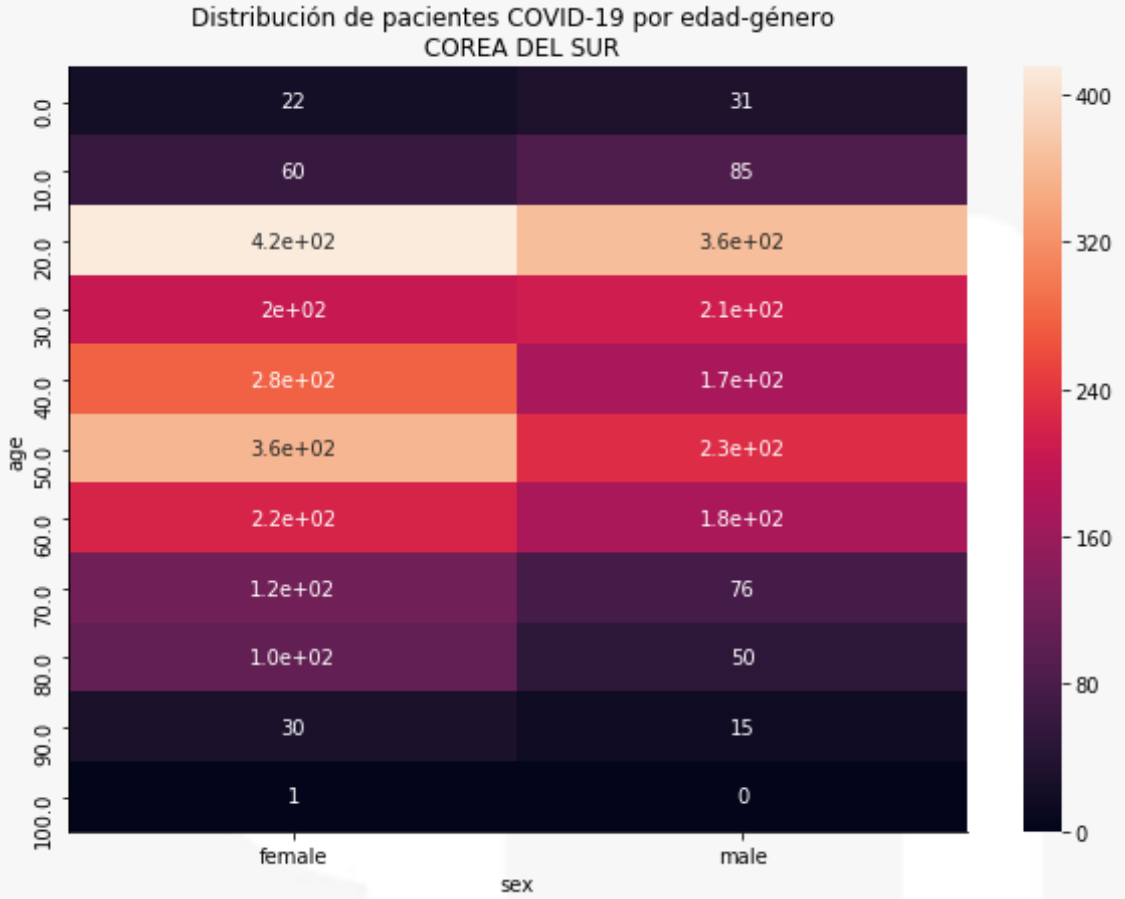
https://chartio.com/learn/charts/funnel-chart-complete-guide/

HEATMAP (Mapa de calor)

- Son usados para mostrar las relaciones entre dos variables (tanto categóricas o numéricas), cada una de ellas representada sobre un eje.

- Puede verse como una matriz coloreada, donde el cambio de tonalidades o colores dentro de las celdas permite percibir la presencia de patrones en caso que existan.

- Los colores de las celdas pueden corresponderse a diferentes tipos de métricas (porcentajes, frecuencias, etc).



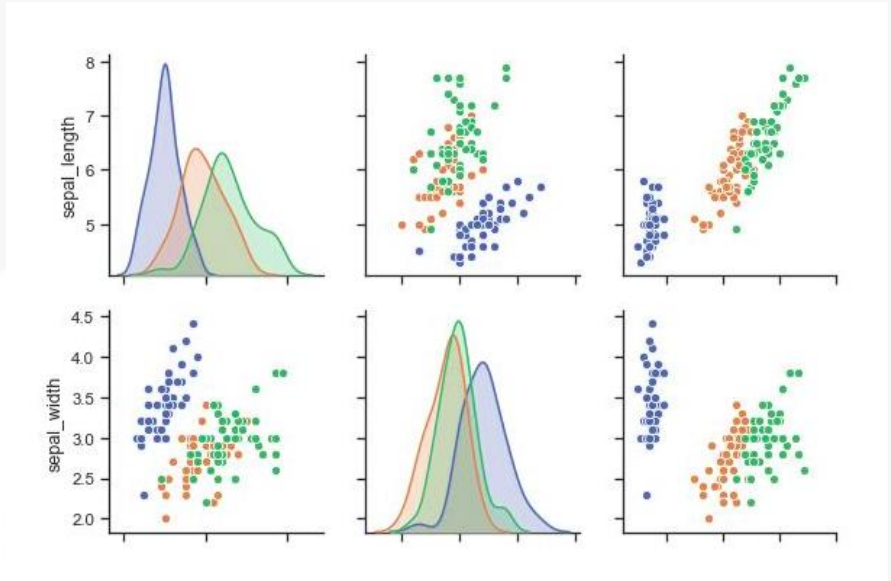
LIBRERÍAS

MATPLOTLIB

- Es quizás la librería más popular para realizar visualización en python. Es de bajo nivel y proporciona mucha libertad (muchísima!).
- Es una herramienta de propósito general (gráficos de barras, de torta, de línea, etc).
- Comenzó a ser desarrollada en 2003 por John Hunter, neurobiólogo, quien quería emular los comandos de MATLAB de Mathworks.
- John murió a los 44 años en 2012 pero matplotlib se mantiene y es desarrollada gracias al aporte de toda la comunidad.
- Link a charla de youtube de la conferencia Scipy 2012 donde John habló sobre la evolución de matplotlib.

SEABORN

- Está construida sobre la base de Matplotlib.
- Interfaz de alto nivel fácil de usar.
- Permite realizar gráficos estadísticos atractivos e informativos y algunos más complejos (como heatmaps o series de tiempo) de manera relativamente simple.
- Ayuda a resolver dos desventajas de Matplotlib:
  - Parámetros predeterminados de Matplotlib
  - Trabajo con dataframes.
- Algunas características destacadas:
  - Visualización de datos univariados y bivariados.
  - Montaje y visualización de modelos de regresión lineal.
  - Trazado de datos estadísticos de series de tiempo.
  - Buen funcionamiento con estructuras de datos NumPy y Pandas.
  - Viene con temas integrados para el diseño de gráficos Matplotlib
  - .Es bastante práctico para mostrar posible correlaciones entre los datos.



Seaborn

PANDAS

- Es una librería open source de alto rendimiento y fácil de usar que proporciona estructuras de datos, como dataframes y herramientas de análisis de datos como las herramientas de visualización.
- La visualización hace que sea realmente fácil crear trazados a partir de un dataframe o series de pandas.
- En muchos casos, en comparación con otras librerías, necesitamos menos código para obtener los mismos resultados.

ALTAIR

- Es una librería estadística declarativa de visualización construida sobre la gramática de visualización popular Vega-Lite.

- Permite crear visualizaciones con una mínima cantidad de código.

- Es bastante amigable, consistente y viene con una simple API.

PLOTLY

- Es un kit de herramientas de visualización basado en la web que incluye funcionalidades como dendrogramas, gráficos 3D y gráficos de contorno.

- Tiene una API que ofrece gráficos de dispersión, de líneas, barras y barras de error, etc.

- Se puede usar desde una jupyter notebook.

GGPLOT

- Es un sistema para crear gráficos declarativamente. Está basado en la gramática de gráficos de R y se relaciona estrechamente con Pandas.

- No se recomienda su uso para crear gráficos altamente personalizados.

MANOS A LA OBRA!