Overview

Day	2
Name	Practice #1 - Let's explore again
Skills	Literacy, reflexivity, design, create
Time	1h *(finishing in this time line is not required nor graded)*
Submission	No submission

Practice

Preliminary

tools info

Plotly

The plotly Python library is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases.

Built on top of the Plotly JavaScript library (plotly.js), plotly enables Python users to create beautiful interactive web-based visualizations that can be displayed in Jupyter notebooks, saved to standalone HTML files, or served as part of pure Python-built web applications using Dash. The plotly Python library is sometimes referred to as "plotly.py" to differentiate it from the JavaScript library.

→ Plotly Python Open Source Graphing Library

environment setup

It would be easier here to use your local environment :

- conda install -c plotly plotly=4.9.0

Open the notebook template, and test that plotly is working properly running the 2 first cells.

Data exploration

Step 1 : Context and data overview

- Run the following cells and have a look at the data manipulation and merging
- From the dataframe **df_merged**, list files, variables and their type & range (if applicable) in order to have a global picture.

  → Tips : df.head, df.info, df.describe

  Is it the same starting point as last week ? Do we have more / less dimensions in data ?

- Read a bit about the indicators explanations
- Given the variables available, what are the questions that are coming to you ? What kind of relationship do you anticipate or are curious about ?

Step 2 : Exploring each variable

- Get a better look at each individual variable alone. Let's do this for one numerical variable *(what do their distribution look like ?).*

  With this data here, we have more dimensions than last week (*for each country, for each year, we have values that can be investigated)*

- So let's start by a numerical variable that has a meaning even if it's summed across one dimension (years / countries). Let's look at the numerical variable **population** over time summed for all countries.

  → Tips : look for distribution graph in the Plotly Express gallery

  What about gdp per capita ? Can you do the same ?

- Now, let's try to see how we can still see distribution for another variable that doesn't make sense if it's summed : life expectancy. Because the basic options of Plotly Express does not allow us to specify the aggregation function, we are using the **go.Histogram** class from plotly.graph_objects.

  Can you do the same for the fertility rate ?

- Then, let's visualise another variable, the fertility rate, on both dimensions : for each country and for each year. What kind of graph can show an evolution for many entries ?
  → Tips : px.line, line_group

- What if you would like to see the evolution of fertility rate over time for **only one country**, let's say France ? How would you do ?

  → Tips : df.loc[df[x] == **'coucou'**]

Step 3 : Exploring variables relationships

Discovering the parallel coordinate plot

Before we look at the relationship between two variables, let's discover a chart type that is not widespread but very useful to have an overview of many numerical variables distribution : the Parallel Coordinates plot.

The best practice when we have a lot of lines (here 1 line = 1 country) is to have opacity under 100% to see the most overlapped area. This feature is not available in Plotly.

The plot in 1950 is provided. Can you do the same for 2019 ?
Do you learn something from comparing the two graphs ?

Exploring

Let's have a look at the relationship between **fertility rate** and **gdp per capita**.

- What would be your first idea about this ? How would you feel this relationship would be ?

- In order to fix one dimension before charting, let's say we want to see the relationship between **fertility rate** and **gdp per capita** first in **1950** and then in **2019**.

  What types of charts can we make to investigate **correlation** ? (data to viz)

  Make a chart representing the relationship between those 2 variables in **1950**.

  → Tips : df.loc[df[x] == **'coucou'**]
  → Tips : px.scatter

Let's improve a bit this chart by adding the name of the country on hover and log axis for gdp per capita.

→ Tips : in the chart option add a parameter : hover_data=['**variable**']
→ Tips : in the chart option add a parameter : log_x=True

- Now, let's add more context with the country's population. Can you see a way to add this variable to the chart ? What other visual variables are available for us to encode ? From the ones available, which one would be easier to see ?

  You have plotted this for the year 1950. Can you do it again for 2019 ?

  Have a look at the situation in 1950 then 2019. Does that match the idea you had before ? Do you see a lot of difference between 1950 and 2019 ?

  → Tips : size

- It's quite a long time between 1950 and 2019. Let's see if we can have a view every 10 years and plot those charts next to each other. I've created a dataframe for selected years **df_selected_years** for simplicity of use.

  Let's plot a small multiple : the same chart we did before but for each year present in the dataframe. Have a look at the documentation on subploting with Plotly.

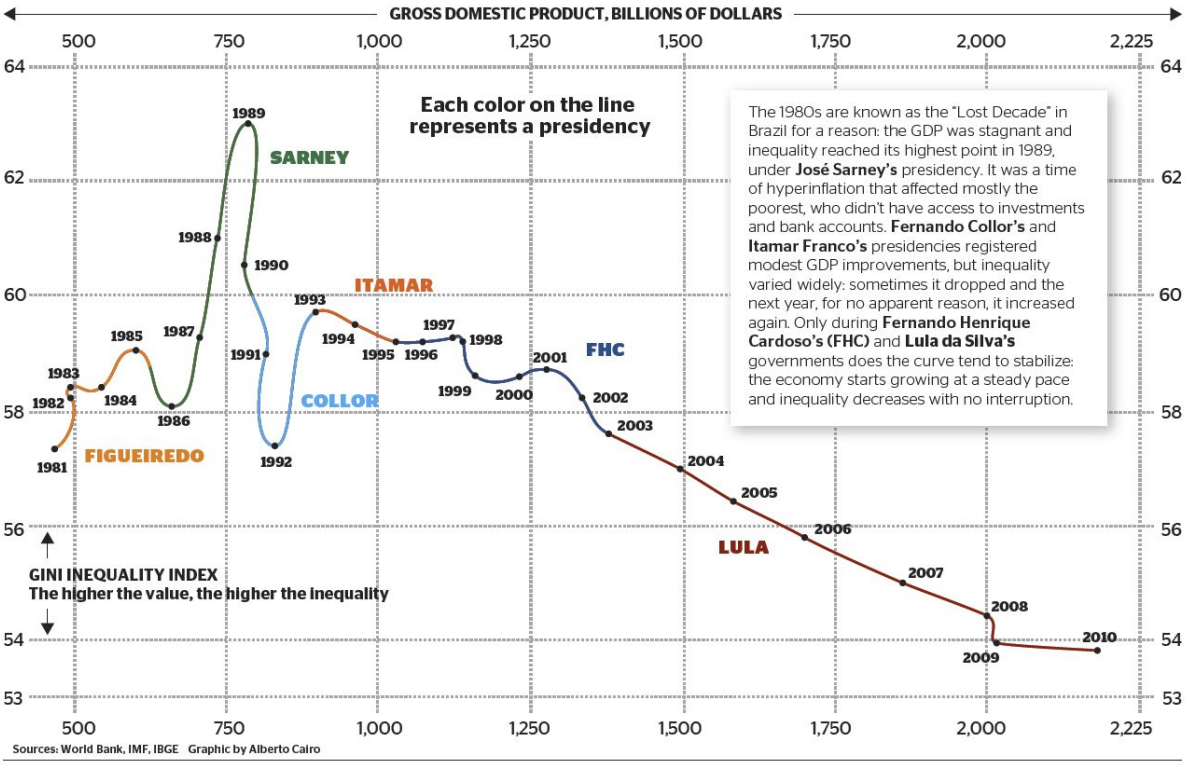
- That's cool. But there is an even cooler feature of Plotly that could do that even better : animation. Use the animation feature of plotly express to see the evolution over time in a single chart

- [Bonus] Let's build a chart that you might see a lot, but which is quite interesting at times : **a connected scatter plot over time**. In this case we keep 2 numerical variables (that we want to investigate) encoded as x and y position of a scatterplot. And we plot a point for each year, which is linked to the previous and the next one. It helps to see the evolution of 2 variables over time.

  Look at this example (source : In praise of connected scatter plots)

When the Brazilian Economy Improves, Inequality Doesn't Drop

The graphic below shows the correlation between Brazilian GDP (horizontal axis) and inequality (vertical axis) between 1981 and 2010. The position of the points, each representing a year, depends on how high GDP and inequality were. You can notice, for instance, that the economy grew between 1986 and 1989 because the line tends to move to the right, but inequality also grew, as the point representing 1989 is much higher than the ones before. You can also see that, during Lula da Silva's government, the economy expanded almost as much as during the terms of the other presidents who preceded him combined.



Let's make our own !

Let's say we want to plot it for France (you can choose the country if you want).

Because it's kind of a weird chart, we will use the plotly.graph_objects module to build a connected scatterplot that we can customize.

From the basic chart provide, let's add more information so that it's more lisible (see documentation) :

- → Make bigger dot (called marker)
- → Encode the dot color as per the year to better see the evolution of time and its direction
- → Show this color scale
- → Reduce the line weight and choose a less flashy color (not to interfere with the colors used for the years)
- → Add x and y axis legend (see documentation)

Search how you can use marker=dict() and line=dict=() options to style this connected scatter plot

Make it again for another country.

Do you feel it's easy to read ? Does it show you more information ?

- [Bonus] We looked here at a specific variable : gdp per capita and fertility rate. What would you like to explore ? What other charts would you like to make ?

Reflection

Meet with one person and discuss :

Data exploration, Chart uses and findings :

- Any trouble in this practice ?
- What did you learn ? What did you find and not find ?
- Are there charts you wanted to make and did not have the time ? Which one ? Can you draw them ?
- Are there questions left where you did not investigate ? Can you imagine/draw chart types & related variables that could be done ?