Assignment4

August 4, 2022

1 Assignment 4

Before working on this assignment please read these instructions fully. In the submission area, you will notice that you can click the link to **Preview the Grading** for each step of the assignment. This is the criteria that will be used for peer grading. Please familiarize yourself with the criteria before beginning the assignment.

This assignment requires that you to find **at least** two datasets on the web which are related, and that you visualize these datasets to answer a question with the broad topic of **economic activity or measures** (see below) for the region of **Arceburgo, Minas Gerais, Brazil**, or **Brazil** more broadly.

You can merge these datasets with data from different regions if you like! For instance, you might want to compare **Arceburgo, Minas Gerais, Brazil** to Ann Arbor, USA. In that case at least one source file must be about **Arceburgo, Minas Gerais, Brazil**.

You are welcome to choose datasets at your discretion, but keep in mind **they will be shared with your peers**, so choose appropriate datasets. Sensitive, confidential, illicit, and proprietary materials are not good choices for datasets for this assignment. You are welcome to upload datasets of your own as well, and link to them using a third party repository such as github, bitbucket, pastebin, etc. Please be aware of the Coursera terms of service with respect to intellectual property.

Also, you are welcome to preserve data in its original language, but for the purposes of grading you should provide english translations. You are welcome to provide multiple visuals in different languages if you would like!

As this assignment is for the whole course, you must incorporate principles discussed in the first week, such as having as high data-ink ratio (Tufte) and aligning with Cairo's principles of truth, beauty, function, and insight.

Here are the assignment instructions:

- State the region and the domain category that your data sets are about (e.g., **Arceburgo, Minas Gerais, Brazil** and **economic activity or measures**).
- You must state a question about the domain category and region that you identified as being interesting.
- You must provide at least two links to available datasets. These could be links to files such as CSV or Excel files, or links to websites which might have data in tabular form, such as Wikipedia pages.
- You must upload an image which addresses the research question you stated. In addition to addressing the question, this visual should follow Cairo's principles of truthfulness, functionality, beauty, and insightfulness.

- You must contribute a short (1-2 paragraph) written justification of how your visualization addresses your stated research question.

What do we mean by **economic activity or measures**? For this category you might look at the inputs or outputs to the given economy, or major changes in the economy compared to other regions.

1.1 Tips

- Wikipedia is an excellent source of data, and I strongly encourage you to explore it for new data sources.
- Many governments run open data initiatives at the city, region, and country levels, and these are wonderful resources for localized data sources.
- Several international agencies, such as the United Nations, the World Bank, the Global Open Data Index are other great places to look for data.
- This assignment requires you to convert and clean datafiles. Check out the discussion forums for tips on how to do this from various sources, and share your successes with your fellow students!

1.2 Example

Looking for an example? Here's what our course assistant put together for the **Ann Arbor, MI, USA** area using **sports and athletics** as the topic. Example Solution File

2 MY FINAL PROJECT - COVID-19 ANALYSIS

3 1. Region and Domain

State the region and the domain category that your data sets are about. India(IND), United States (USA), Brazil (BRA) and Russia (RUS)

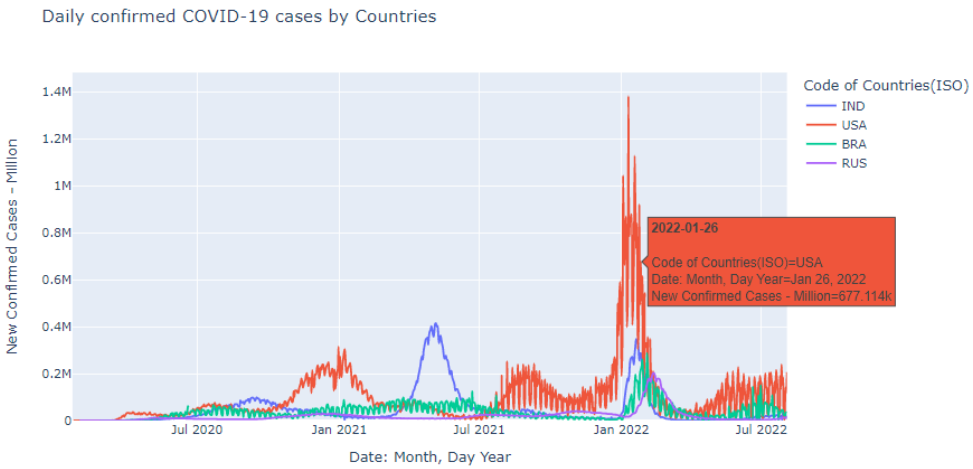
COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University

4 2. Research Question

You must state a question about the domain category and region that you identified as being interesting. Question: What is the daily number of confirmed cases in these countries? Daily cases: where are confirmed cases increasing or falling?

5 3. Links

You must provide at least two links to publicly accessible datasets. These could be links to files such as CSV or Excel files, or links to websites which might have data in tabular form, such as Wikipedia pages. The coursera does not accept files > 25 mb, so, I downloaded the file from : https://covid.ourworldindata.org/data/owid-covid-data.csv and then select the countries that we will analyze.



Graph1

6 4. Image

You must upload an image which addresses the research question you stated. In addition to addressing the question, this visual should follow Cairo's principles of truthfulness, functionality, beauty, and insightfulness.

7 5. Discussion

You must contribute a short (1-2 paragraph) written justification of how your visualization addresses your stated research question Daily new confirmed COVID-19 cases (in million people) compared with some European and non-European countries'. This visualization was concerned with answering the question of What is the daily number of confirmed cases in these countries and where are confirmed cases have increased or decreased over the years.

The plot indicates that the daily cases increased throught the early of 2022, with a doward trend in recent months on these countries.

8 THE CODE:

```
In [83]: #import pandas as pd
         #df = pd.read_csv('Downloads/owid-covid-data.csv')
         #india = df[df['iso_code'] == 'IND']
         #usa = df[df['iso_code'] == 'USA']
         #bra = df[df['iso_code']=='BRA']
         #rus = df[df['iso_code']== 'RUS']
         #df = pd.concat([india,usa,bra,rus], axis=0)
         #df.to_csv('DataFrameAssignment.csv')

In [110]: !pip install chart_studio
          import pandas as pd
          import numpy as np
```

```
import chart_studio.plotly as py
import seaborn as sns
import plotly.express as px
%matplotlib inline

# Make Plotly work in your Jupyter Notebook
from plotly.offline import download_plotlyjs, init_notebook_mode, plot, i
init_notebook_mode(connected=True)
# Use Plotly locally
```

```
Requirement already satisfied: chart_studio in /opt/conda/lib/python3.6/site-packag
Requirement already satisfied: plotly in /opt/conda/lib/python3.6/site-packages (fr
Requirement already satisfied: six in /opt/conda/lib/python3.6/site-packages (from
Requirement already satisfied: retrying>=1.3.3 in /opt/conda/lib/python3.6/site-pac
Requirement already satisfied: requests in /opt/conda/lib/python3.6/site-packages
Requirement already satisfied: tenacity>=6.2.0 in /opt/conda/lib/python3.6/site-pac
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in /opt/conda/lib/python3.6/si
Requirement already satisfied: idna<2.7,>=2.5 in /opt/conda/lib/python3.6/site-pack
Requirement already satisfied: urllib3<1.23,>=1.21.1 in /opt/conda/lib/python3.6/si
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.6/site-
You are using pip version 9.0.1, however version 22.2.2 is available.You should con
```

In [111]:
```
df = pd.read_csv('readonly/DataFrameAssignment.csv')
```

```
df.head()
```

Out[111]:

	Unnamed: 0	iso_code	continent	location	date	total_cases	new_c
0	84981	IND	Asia	India	2020-01-30	1.0	
1	84982	IND	Asia	India	2020-01-31	1.0	
2	84983	IND	Asia	India	2020-02-01	1.0	
3	84984	IND	Asia	India	2020-02-02	2.0	
4	84985	IND	Asia	India	2020-02-03	3.0	

	new_cases_smoothed	total_deaths	new_deaths	\
0	NaN	NaN	NaN	
1	NaN	NaN	NaN	
2	NaN	NaN	NaN	
3	NaN	NaN	NaN	
4	NaN	NaN	NaN	

	...	female_smokers	male_smokers
0	...	1.9	20.6
1	...	1.9	20.6
2	...	1.9	20.6
3	...	1.9	20.6
4	...	1.9	20.6

```
   handwashing_facilities  hospital_beds_per_thousand  life_expectancy  \
0                   59.55                        0.53            69.66
1                   59.55                        0.53            69.66
2                   59.55                        0.53            69.66
3                   59.55                        0.53            69.66
4                   59.55                        0.53            69.66

   human_development_index  excess_mortality_cumulative_absolute  \
0                    0.645                                   NaN
1                    0.645                                   NaN
2                    0.645                                   NaN
3                    0.645                                   NaN
4                    0.645                                   NaN

   excess_mortality_cumulative  excess_mortality  \
0                          NaN               NaN
1                          NaN               NaN
2                          NaN               NaN
3                          NaN               NaN
4                          NaN               NaN

   excess_mortality_cumulative_per_million
0                                      NaN
1                                      NaN
2                                      NaN
3                                      NaN
4                                      NaN

[5 rows x 68 columns]
```

Line plots on Date axes Line plots can be made on using any type of cartesian axis, including linear, logarithmic, categorical or date axes. Line plots on date axes are often called time-series charts.

Plotly auto-sets the axis type to a date format when the corresponding data are either ISO-formatted date strings or if they're a date pandas column or datetime NumPy array.

```
In [113]: fig = px.line(df, x="date", y="new_cases", color="iso_code", title = 'Dai
                  line_shape="spline", render_mode="svg", range_y=[0,df['new_cases'
                        labels={
                               "new_cases": "New Confirmed Cases - Million",
                               "date": "Date: Month, Day Year",
                               "iso_code": "Code of Countries(ISO)"
                           },
                       )
          #fig

In [ ]:
```