Assignment4-Copy1

August 9, 2020

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 **Ann Arbor, Michigan, United States**, or **United States** more broadly.

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

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., **Ann Arbor, Michigan, United States** 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

```
In [96]: import matplotlib.pyplot as plt
         import seaborn as sns
         import pandas as pd
         import numpy as np
         import os

         %matplotlib notebook
```

```
In [97]: df=pd.read_csv('city_day.csv')
```

```
In [98]: df.head()
```

Out[98]:

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	
0	Ahmedabad	2015-01-01	NaN	NaN	0.92	18.22	17.15	NaN	0.92	27
1	Ahmedabad	2015-01-02	NaN	NaN	0.97	15.69	16.46	NaN	0.97	24
2	Ahmedabad	2015-01-03	NaN	NaN	17.40	19.30	29.70	NaN	17.40	29
3	Ahmedabad	2015-01-04	NaN	NaN	1.70	18.48	17.97	NaN	1.70	18
4	Ahmedabad	2015-01-05	NaN	NaN	22.10	21.42	37.76	NaN	22.10	39

	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	133.36	0.00	0.02	0.00	NaN	NaN
1	34.06	3.68	5.50	3.77	NaN	NaN
2	30.70	6.80	16.40	2.25	NaN	NaN
3	36.08	4.43	10.14	1.00	NaN	NaN
4	39.31	7.01	18.89	2.78	NaN	NaN

```
In [99]: df['Year'] = df['Date'].apply(lambda x: x[:4])

         df=df[(df['City']=='Delhi') | (df['City']=='Mumbai') | (df['Date']=='2020'
         df['City'].unique()
         df=df.drop('Date',axis=1)

In [100]: df=df.groupby(['City']).mean()
          df['CO'] *= 100
          df.reset_index()

Out[100]:      City       PM2.5        PM10         NO        NO2         NOx
          0   Delhi  117.196153  232.809229  38.985595  50.785182  58.567023  41.99
          1  Mumbai   35.198393   96.745059  31.014133  25.555212  55.038795  13.83

                     CO         SO2         O3   Benzene    Toluene    Xylene
          0  197.605276  15.901253  51.323610  3.544480  17.185042  1.438339  259.4
          1   56.836190  15.197516  33.058946  1.098605   0.013455  0.008010  105.3

In [101]: df= (df[['SO2', 'CO', 'NO2']].rename(columns={
                    'SO2':'SO2 [ppb]', 'CO':'CO [10^2ppm]',
                    'NO2':'NO2 [ppb]'}).unstack())

In [102]: df=df.to_frame()
          df=df.reset_index()

In [103]: df=df[['level_0','City', 0]].rename(columns={
                    'level_0':'Pollutant','City':'city',0:'Value'})

In [104]: df

Out[104]:      Pollutant    city       Value
          0    SO2 [ppb]   Delhi   15.901253
          1    SO2 [ppb]  Mumbai   15.197516
          2  CO [10^2ppm]   Delhi  197.605276
          3  CO [10^2ppm]  Mumbai   56.836190
          4    NO2 [ppb]   Delhi   50.785182
          5    NO2 [ppb]  Mumbai   25.555212

In [105]: # Draw a nested violinplot and split the violins for each region in part
          plt.figure(figsize=(10,5))
          sns.barplot(x="Pollutant", y="Value",hue='city',data=df,palette='BuGn_r')
          plt.xlabel('Pollutant')
          plt.ylabel('Emissions')
          plt.rc('xtick', labelsize=15);
          plt.rc('ytick', labelsize=15);
          plt.legend(loc='upper left', fontsize=13);
          ttl = plt.title('Comparison of pollutant emissions in 2020 between Mumbai

<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

In [109]:
```
plt.savefig('Course2-asign4.pdf')
```

In [ ]: