SCHOOL OF SCIENCE & ENGINEERING

SPRING 2024

CSC 5356 01 Data Engineering and Visualization

Project#3 Viz of 2 Large Datasets

April 15th, 2024

Realized by:

Khadija Salih Alj

Noura Ogbi

Supervised by:

Prof. Tajjeeddine Rachidi

Table of Contents

I. Introduction

The purpose of this report is to outline the methodology for identifying and visualizing two very large datasets, one being Geospatial, and the other being Field-based. This project involves following the "**What/Why/How**" steps/methodology to describe the datasets, identify tasks, and perform visualizations with manipulations.

Here is the link to our Demo: https://www.youtube.com/watch?v=k-rFz0Z0vHg

II. Geospatial Visualization

1. Dataset

The earthquake dataset (https://raw.githubusercontent.com/plotly/datasets/master/earthquakes-23k.csv) spans from January 1965 to December 2016. It includes attributes:

- **Date**: The date of the earthquake event in the format MM/DD/YYYY.
- **Latitude**: The geographical latitude coordinate of the earthquake event.
- **Longitude**: The geographical longitude coordinate of the earthquake event.
- **Magnitude**: The magnitude of the earthquake event on the Richter scale.

This dataset records earthquake events globally, with each entry specifying the date of the earthquake, its geographical location using latitude and longitude coordinates, and the earthquake's magnitude on the Richter scale.

```
Date,Latitude,Longitude,Magnitude
01/02/1965,19.246,145.616,6.0
01/04/1965,1.8630000000000002,127.352,5.8
01/05/1965,-20.579,-173.972,6.2
01/08/1965,-59.076,-23.557,5.8
01/09/1965,11.937999999999999,126.427,5.8
01/10/1965,-13.405,166.62900000000002,6.7
01/12/1965,27.357,87.867,5.9
01/15/1965,-13.309000000000001,166.21200000000002,6.0
01/16/1965,-56.452,-27.043000000000003,6.0
01/17/1965,-24.563000000000002,178.487,5.8
01/17/1965,-6.807,108.988,5.9
```

Figure1: Earthquake Data (sample)

2. Visualization and Manipulation

a. Definition

Geospatial Visualization refers to the process of representing and displaying geospatial data, such as geographic information, spatial relationships, and physical features, in a visual format. This visualization technique often involves the use of maps, geographic coordinates, and spatial layers to convey patterns, trends, and insights related to geographical phenomena. Geospatial visualization enables the comprehension of complex spatial data through graphical representations, aiding in the analysis, interpretation, and communication of geospatial information.

b. What, How, Why?

Idiom	Map Chart
What? Data	The earthquake dataset provides records of earthquake events globally from January 1965 to December 2016. It includes attributes such as date, latitude, longitude, and magnitude.
How? Encode	For the map chart, the data is encoded using latitude and longitude coordinates to plot earthquake locations on a global map. The magnitude of each earthquake is represented by the intensity or color of the data points.
Why? Task	Actions: Locate, explore, and show the density and magnitude of earthquake events across the globe over the specified period. Target: Identify insights of the distribution and intensity of earthquakes, highlighting regions with higher seismic activity.
Scale	The scale of the visualization is global, covering earthquake events from January 1965 to December 2016.

c. Source Code

The following Python code demonstrates how to create a map chart using the earthquake dataset. The code uses the Plotly library to generate a density Mapbox visualization, encoding earthquake data based on latitude, longitude, and magnitude to visualize their global distribution and intensity.

```
earthquakes = 'https://raw.githubusercontent.com/plotly/datasets/master/earthquakes-23k.csv'
df = pd.read_csv(earthquakes)
df.head()
fig = px.density_mapbox(df,
                        lat = 'Latitude',
                        lon = 'Longitude',
                        z = 'Magnitude',
                        radius = 10,
                        center = dict(lat = 9, lon = 9),
                        zoom = 1,
                        hover_name = 'Date',
                        mapbox_style = 'open-street-map',
                        title = 'Earthquake Data 1965 - 2016')
fig.show()
```

Figure2: Code for Generating Geospatial Density Map of Earthquake Data (1965 - 2016)

d. Visualization

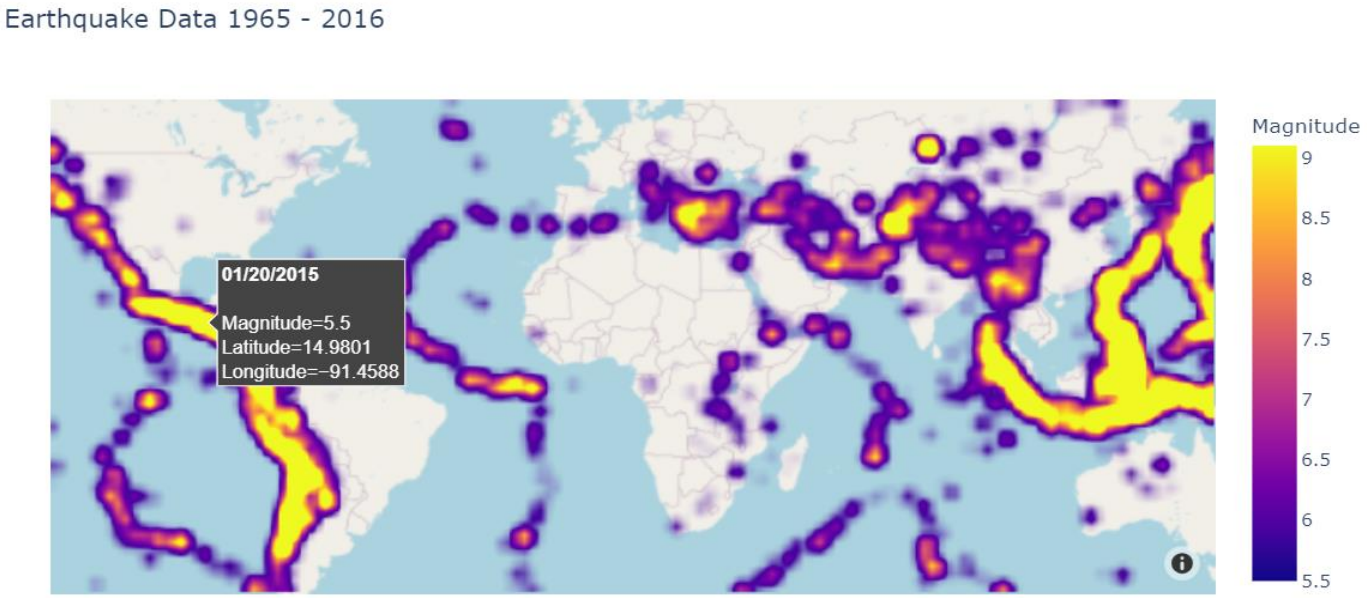
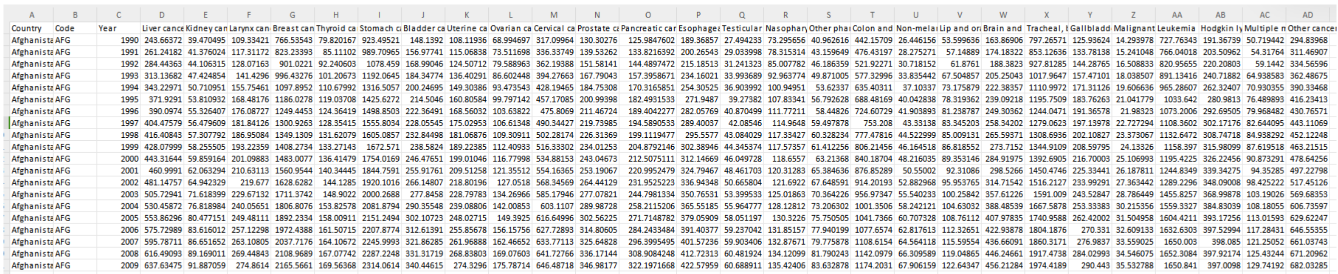


Figure3: Map Chart Visualizing Earthquake Density and Magnitude (1965 - 2016)

III. Field Visualization

1. Dataset

The dataset (cancer.csv) includes cancer death rates for various countries. In our project, we focus on Morocco, spanning from 1990 to 2016. It contains attributes such as **Country**, **Code**, **Year**, and specific **cancer types** including Liver, Kidney, Larynx, Breast, Thyroid, Stomach, Bladder, Uterine, Ovarian, Cervical, Prostate, Pancreatic, Esophageal, Testicular, Nasopharynx, Other pharynx, Colon and rectum, Non-melanoma skin, Lip and oral cavity, Brain, and nervous system, Tracheal, bronchus, and lung, Gallbladder and biliary tract, Malignant skin melanoma, Leukemia, Hodgkin lymphoma, Multiple myeloma, and Other cancers. The dataset provides information on cancer death rates per 100,000 population for Morocco and covers a wide range of cancer types over the specified time period.



Country	Code	Year	Liver cance	Kidney can	Larynx can	Breast can	Thyroid ca	Stomach ca	Bladder ca	Uterine ca	Ovarian ca	Cervical ca	Prostate ca
Afghanista	AFG	1990	243.66372	39.470495	109.33421	766.53543	79.820167	923.49521	148.1392	108.11936	68.994697	317.09964	130.30276
Afghanista	AFG	1991	261.24182	41.376024	117.31172	823.23393	85.11102	989.70965	156.97741	115.06838	73.511698	336.33749	139.53262
Afghanista	AFG	1992	284.44363	44.106315	128.07163	901.0221	92.240603	1078.459	168.99046	124.50712	79.588963	362.19388	151.58141
Afghanista	AFG	1993	313.13682	47.424854	141.4296	996.43276	101.20673	1192.0645	184.34774	136.40291	86.602448	394.27663	167.79043
Afghanista	AFG	1994	343.22971	50.710951	155.75461	1097.8952	110.67992	1316.5057	200.24695	149.30386	93.473543	428.19465	184.75308
Afghanista	AFG	1995	371.9291	53.810932	168.48176	1186.0278	119.03708	1425.6272	214.5046	160.80584	99.797142	457.17085	200.99398
Afghanista	AFG	1996	390.0974	55.326407	176.08727	1249.4453	124.36419	1498.8503	222.36491	168.56032	103.63822	475.8069	211.46724
Afghanista	AFG	1997	404.47579	56.479609	181.84126	1300.9263	128.35415	1555.8034	228.05545	175.02953	106.61348	490.34427	219.73985
Afghanista	AFG	1998	416.40843	57.307792	186.95084	1349.1309	131.62079	1605.0857	232.84498	181.06876	109.30911	502.28174	226.31369
Afghanista	AFG	1999	428.07999	58.255505	193.22359	1408.2734	133.27143	1672.571	238.5824	189.22385	112.40933	516.33302	234.01253
Afghanista	AFG	2000	443.31644	59.859164	201.09883	1483.0077	136.41479	1754.0169	246.47651	199.01046	116.77998	534.88153	243.04673
Afghanista	AFG	2001	460.9991	62.063294	210.63113	1560.9544	140.34445	1844.7591	255.91761	209.51258	121.35512	554.16365	253.19067
Afghanista	AFG	2002	481.14757	64.942329	219.677	1628.6282	144.1285	1920.1016	266.14807	218.80196	127.0518	568.34569	264.44129
Afghanista	AFG	2003	505.72941	71.618399	229.67132	1711.3742	148.9022	2000.2688	277.8458	228.79783	134.26966	585.17946	277.07821
Afghanista	AFG	2004	530.45872	76.818984	240.05651	1806.8076	153.82578	2081.8794	290.35548	239.08806	142.00853	603.1107	289.98728
Afghanista	AFG	2005	553.86296	80.477151	249.48111	1892.2334	158.00911	2151.2494	302.10723	248.02715	149.3925	616.64996	302.56225
Afghanista	AFG	2006	575.72989	83.616012	257.12298	1972.4388	161.50715	2207.8774	312.61391	255.85678	156.15756	627.72893	314.80605
Afghanista	AFG	2007	595.78711	86.651652	263.10805	2037.7176	164.10672	2245.9993	321.86285	261.96888	162.46652	633.77113	325.64828
Afghanista	AFG	2008	616.49093	89.169011	269.44843	2108.9689	167.07742	2287.2248	331.31719	268.83803	169.07603	641.72766	336.17144
Afghanista	AFG	2009	637.63475	91.887059	274.8614	2165.5661	169.56368	2314.0614	340.44615	274.3296	175.78714	646.48718	346.98177

Year	Pancreatic car	Esophagea	Testicular	Nasophary	Other phar	Colon and	Non-mela	Lip and ora	Brain and
1990	125.9847602	189.36857	27.494233	73.295656	40.962616	442.15709	26.446156	53.599636	163.86906
1991	133.8216392	200.26543	29.033998	78.315314	43.159649	476.43197	28.275271	57.14889	174.18322
1992	144.4897472	215.18513	31.241323	85.007782	46.186359	521.92271	30.718152	61.8761	188.3823
1993	157.3958671	234.16021	33.993689	92.963774	49.871005	577.32996	33.835442	67.504857	205.25043
1994	170.3165851	254.30525	36.903992	100.94951	53.62337	635.40311	37.10337	73.175879	222.38357
1995	182.4931533	271.9487	39.27382	107.83341	56.792628	688.48169	40.042838	78.319362	239.09218
1996	189.4042277	282.05769	40.870499	111.77211	58.44826	724.60729	41.903893	81.238787	249.30362
1997	194.5890533	289.40037	42.08546	114.9648	59.497878	753.208	43.33138	83.345203	258.34202
1998	199.1119477	295.5577	43.084029	117.33427	60.328234	777.47816	44.522999	85.009131	265.59371
1999	204.8792146	302.38946	44.345374	117.57357	61.412256	806.21456	46.164518	86.818552	273.7152
2000	212.5075111	312.14669	46.049728	118.6557	63.21368	840.18704	48.216035	89.353146	284.91975
2001	220.9952479	324.79467	48.461703	120.31283	65.384636	876.85289	50.55002	92.31086	298.5266
2002	231.9525223	336.94348	50.665804	121.6922	67.648591	914.20193	52.882968	95.953765	314.71542
2003	244.7981334	350.76531	53.399533	125.01863	70.364226	956.97347	55.540233	100.25842	357.61226
2004	258.2115206	365.55185	55.964777	128.12812	73.206302	1001.3506	58.242121	104.63032	388.48539
2005	271.7148782	379.05909	58.051197	130.3226	75.750505	1041.7366	60.707328	108.76112	407.97835
2006	284.2433484	391.40377	59.237042	131.85157	77.940199	1077.6574	62.817613	112.32651	422.93878
2007	296.3995495	401.57236	59.903406	132.87671	79.775878	1108.6154	64.564118	115.59554	436.66091
2008	308.9084248	412.72313	60.481924	134.12099	81.790243	1142.0979	66.309589	119.04865	446.24661
2009	322.1971668	422.57959	60.688911	135.42406	83.632878	1174.2031	67.906159	122.64347	456.21284

Year	Tracheal, b	Gallbladd	Malignant	Leukemia	Hodgkin ly	Multiple m	Other cancer
1990	797.26571	125.93624	14.293978	727.76343	191.36739	50.719442	294.83968
1991	853.12636	133.78138	15.241048	766.04018	203.50962	54.31764	311.46907
1992	927.81285	144.28765	16.508833	820.95655	220.20803	59.1442	334.56596
1993	1017.9647	157.47101	18.038507	891.13416	240.71882	64.938583	362.48675
1994	1110.9972	171.31126	19.606636	965.28607	262.32407	70.930355	390.33468
1995	1195.7509	183.76263	21.041779	1033.642	280.9813	76.489893	416.23413
1996	1244.0471	191.36578	21.98323	1073.2006	292.69505	79.968482	430.76571
1997	1279.0623	197.13978	22.727294	1108.3602	302.17176	82.644095	443.11069
1998	1308.6936	202.10827	23.373067	1132.6472	308.74718	84.938292	452.12248
1999	1344.9109	208.59795	24.13326	1158.397	315.98099	87.619518	463.21515
2000	1392.6905	216.70003	25.106993	1195.4225	326.22456	90.873291	478.64256
2001	1450.4746	225.33441	26.187811	1244.8349	339.34275	94.35285	497.22798
2002	1516.2127	233.99291	27.363442	1289.2296	348.09008	98.425222	517.45126
2003	1591.009	243.52847	28.786449	1455.8257	368.99878	103.19026	569.68353
2004	1667.5878	253.33383	30.215356	1559.3327	384.83039	108.18055	606.73597
2005	1740.9588	262.42002	31.504958	1604.4211	393.17256	113.01593	629.62247
2006	1804.1876	270.331	32.609133	1632.6303	397.52994	117.28431	646.55355
2007	1860.3171	276.9837	33.559025	1650.003	398.085	121.25052	661.03743
2008	1917.4738	284.02993	34.546075	1652.3084	397.92174	125.43244	671.20962
2009	1974.4189	290.443	35.532788	1650.841	397.0098	129.74192	682.03285

Figure4: Cancer Death rates by Country (1990-2016) (sample)

2. Visualization and Manipulation

a. Definition

Field Visualization refers to the process of representing and displaying data and observations within a specific field or domain in a visual format. This visualization technique is used to convey patterns, trends, and relationships inherent in the data, aiding in the analysis, interpretation, and communication of information related to a particular field or domain. Field visualization often involves the use of graphical elements such as charts, graphs, and diagrams to present the data in a comprehensive and understandable manner, facilitating insights and decision-making within the domain.

b. What, How, Why?

Idiom	Map Chart, Horizontal Bar Chart
What? Data	The dataset contains cancer rates per 100,000 population for different countries from 1990 to 2016. The attributes: Country, Code, Year, and specific cancer types like Liver cancer, Kidney cancer, Breast cancer, etc.
How? Encode	**Map Chart**: Displayed the overall cancer rates across all countries using a map chart. **Horizontal Bar Chart**: Specifically for Morocco, a horizontal bar chart was used to compare cancer rates by type.
Why? Task	Please see below table.

Scale	The cancer rates are per 100,000 population, providing a standardized measure across countries and types.

Action and Target

- Bar Chart
  - Action: Compare and summarize the cancer death rates and find the most spread cancer type in Morocco specifically.
  - Target: Identify the most spread cancer type for future preventive measures.
- Map Chart
  - Action: Locate, compare, and find trends.
  - Target: Show a global overview of cancer rates and highlight geographical patterns.

Comparison between Bar Chart and Map Chart

- Bar Chart
  - Provides a detailed comparison of cancer rates for a specific country.
  - Easier to compare rates for individual cancer types.
  - Suitable for visualizing data for a specific country like Morocco.
- Map Chart
  - Offers a global overview of cancer rates across countries.
  - Highlights geographical patterns and variations in cancer rates.
  - Less detailed compared to the bar chart but provides a broader perspective.

Advantages and Limitations

- Bar Chart
  - Advantages: Detailed, easier comparison for specific data points.
  - Limitations: Limited to a specific country or region, less global perspective.
- Map Chart
  - Advantages: Global overview highlights geographical patterns.
  - Limitations: Less detailed, may not show all data for all countries.

c. Source Code

The visualizations for this project were developed using Python as the programming language. Two primary libraries were employed to create these visualizations: Matplotlib and Geopandas/Plotly. Since the data is not Geospatial (does not include latitude and longitude attributes), we had to merge it with another dataset that is Geospatial, which is 'gapminder'.

```
# Load world data for mapping
world = px.data.gapminder()

# Merge world data with cancer data
merged_data = world.merge(cancer_death_data, left_on='iso_alpha', right_on='Code')
```

Libraries Used:

- **Matplotlib**: Used for creating basic visualizations like bar charts.
- **Geopandas**: Utilized for geospatial data manipulation and visualization, especially for the map chart.
- **Plotly**: Similar to Geopandas, but it is interactive and can be slow for large datasets.

```
import geopandas as gpd
import pandas as pd
import matplotlib.pyplot as plt

# Load the cancer death data
cancer_death_data = pd.read_csv("cancer.csv")
# Calculate the dominant cancer type for each country
cancer_type_cols = [col for col in cancer_death_data.columns if col.endswith(' cancer')]
cancer_death_data['dominant_cancer'] = cancer_death_data[cancer_type_cols].idxmax(axis=1)
# Load world data for mapping
world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
# Merge world data with cancer data
merged_data = world.merge(cancer_death_data, left_on='iso_a3', right_on='Code')
# Plot the map
fig, ax = plt.subplots(figsize=(15, 10))
merged_data.plot(column='dominant_cancer', ax=ax, legend=True,
                 legend_kwds={'title': 'Dominant Cancer', 'loc': 'lower right'},
                 missing_kwds={'color': 'lightgray', 'label': 'Missing data'})
ax.set_title('Distribution of Dominant Cancer Types by Country')
plt.show()
```

Figure5: Code for Visualizing Dominant Cancer Types by Country using Geopandas

```
import pandas as pd
import plotly.express as px

# Load the cancer death data
cancer_death_data = pd.read_csv("cancer.csv")

# Calculate the dominant cancer type for each country
cancer_type_cols = [col for col in cancer_death_data.columns if col.endswith(' cancer')]
cancer_death_data['dominant_cancer'] = cancer_death_data[cancer_type_cols].idxmax(axis=1)

# Load world data for mapping
world = px.data.gapminder()

# Merge world data with cancer data
merged_data = world.merge(cancer_death_data, left_on='iso_alpha', right_on='Code')

# Plot the map
fig = px.choropleth(merged_data,
                    locations='iso_alpha',
                    color='dominant_cancer',
                    hover_name='Country',
                    projection='natural earth',
                    title='Distribution of Dominant Cancer Types by Country',
                    labels={'dominant_cancer': 'Dominant Cancer'})

fig.update_geos(showcountries=True, countrycolor="lightgray")

fig.show()
```

Figure6: Code for Visualizing Dominant Cancer Types by Country using Plotly in Natural Earth Projection

```python
import pandas as pd
import plotly.express as px

# Load the cancer death data
cancer_death_data = pd.read_csv("cancer.csv")

# Calculate the total deaths for each country
cancer_type_cols = [col for col in cancer_death_data.columns if col.endswith(' cancer')]
cancer_death_data['total_deaths'] = cancer_death_data[cancer_type_cols].sum(axis=1)

# Calculate the dominant cancer type for each country
cancer_death_data['dominant_cancer'] = cancer_death_data[cancer_type_cols].idxmax(axis=1)

# Load world data for mapping
world = px.data.gapminder()

# Merge world data with cancer data
merged_data = world.merge(cancer_death_data, left_on='iso_alpha', right_on='Code')

# Create a bubble map
fig = px.scatter_geo(merged_data,
                     locations='iso_alpha',
                     color='dominant_cancer',
                     hover_name='Country',
                     hover_data=['lifeExp', 'dominant_cancer', 'total_deaths'],
                     projection='orthographic',
                     title='Bubble Map of Cancer Deaths by Country',
                     labels={'dominant_cancer': 'Dominant Cancer'})

fig.update_geos(showcountries=True, countrycolor="lightgray")

fig.show()
```

Figure7: Code for Visualizing Dominant Cancer Types by Country using Plotly in Orthographic Projection

```python
import pandas as pd
import matplotlib.pyplot as plt

# Load the cancer death data
cancer_death_data = pd.read_csv("cancer.csv")

# Filter data
country_data = cancer_death_data[cancer_death_data['Country'] == 'Morocco']

# Select columns for cancer types and corresponding deaths
cancer_types = country_data.columns[3:]
deaths = country_data.iloc[0, 3:].values

# Create a horizontal bar chart
plt.figure(figsize=(10, 8))
plt.barh(cancer_types, deaths, color='skyblue')
plt.xlabel('Number of Deaths')
plt.ylabel('Cancer Type')
plt.title('Number of Cancer Deaths in Morocco by Type')
plt.gca().invert_yaxis()  # Invert y-axis to display the highest value at the top
plt.show()
```

Figure8: Code for Generating Horizontal Bar Chart of Cancer Deaths in Morocco

Note: Plotly is used for its interactive capabilities, but it runs slower.

d. Visualization

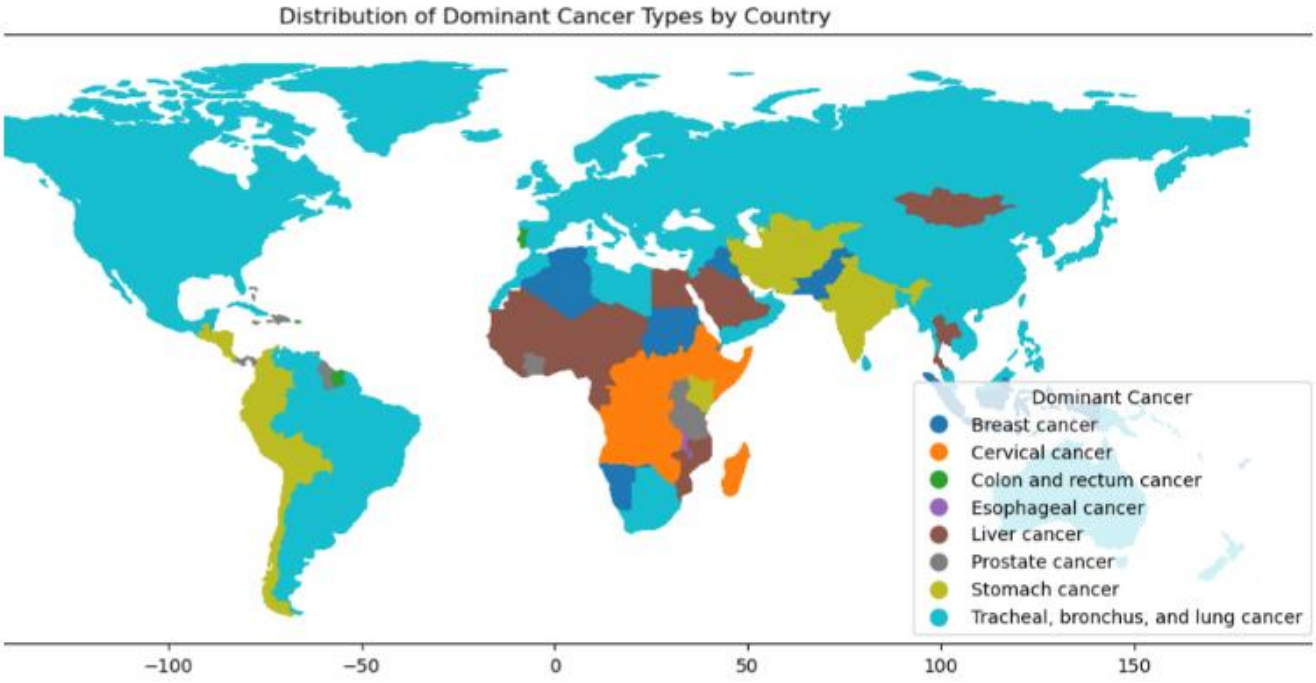


Figure9: Map Visualization of Dominant Cancer Types Worldwide

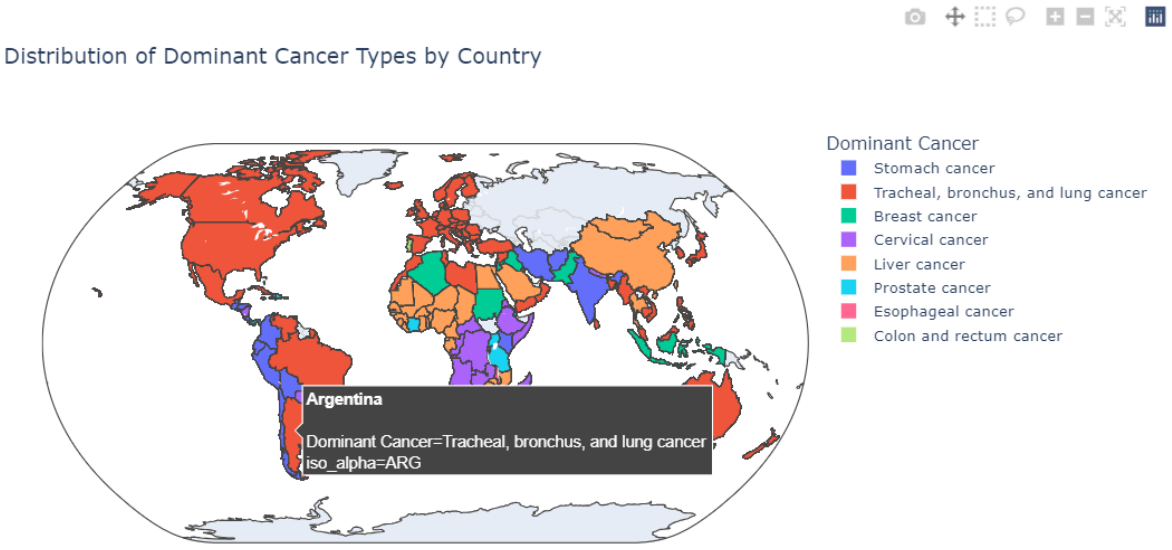


Figure10: Map Visualization of Dominant Cancer Types Worldwide in Natural Earth Projection

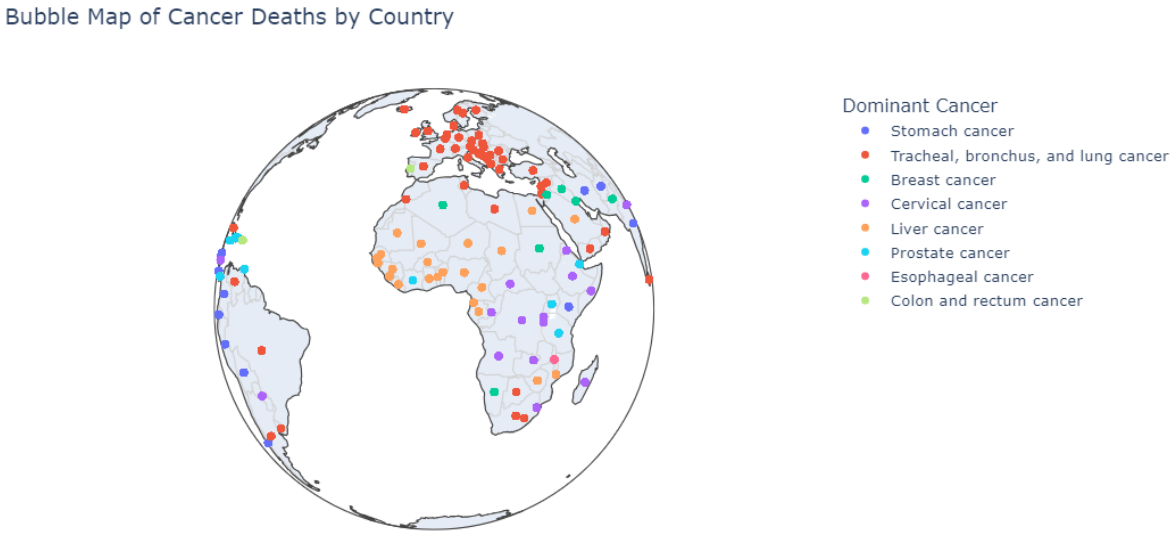


Figure11: Map Visualization of Dominant Cancer Types Worldwide in Orthographic Projection

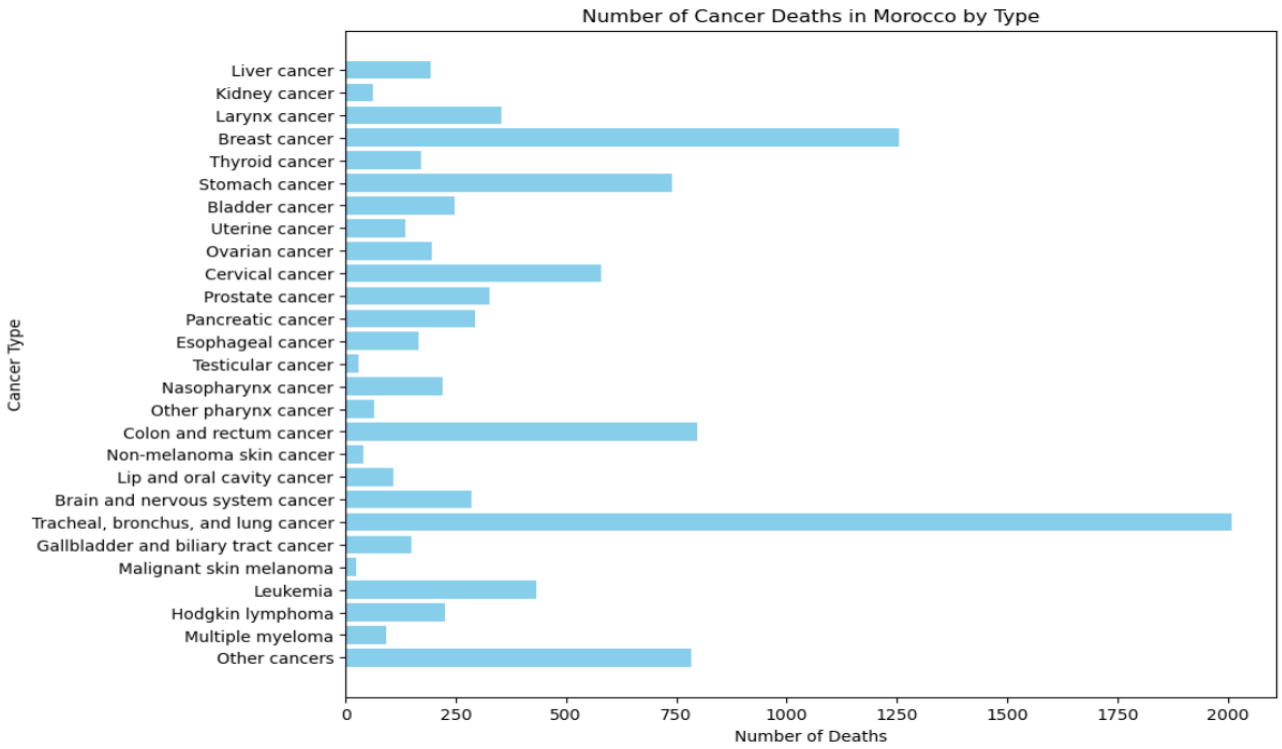


Figure12: Horizontal Bar Chart of Cancer Deaths in Morocco by Type.

IV. Conclusion

In this project, we embarked on a journey to visualize two extensive datasets: geospatial data focusing on earthquake occurrences and field data centered on cancer rates across countries (we did not work with Network since we had the chance to do so in the last project of Neo4j). Adhering to the What/Why/How methodology learned in class, we meticulously described each dataset, pinpointed specific tasks, and employed various visual encodings to effectively communicate the data.

For the geospatial dataset, our objective was to illustrate the global distribution of earthquake events. We aimed to identify areas with high seismic activity as our target. Using a map chart, we encoded earthquake locations using latitude and longitude coordinates, and their magnitudes were represented by color intensity. This visualization offered an insightful view of seismic hotspots and their magnitudes over the years.

For the field dataset, our task was to compare cancer rates across different countries. We focused on comparing cancer rates across Morocco. Two visualizations were employed: a map chart for a global overview and a horizontal bar chart specifically for Morocco. These visualizations utilized encoding through colors and bar lengths, respectively, to represent cancer rates across various types and regions.

In conclusion, this project underscored the significance of data visualization in making sense of large and intricate datasets. These visualizations not only highlighted patterns and trends but also facilitated a deeper understanding of the underlying data, reinforcing the pivotal role of visualization in data analysis and interpretation.