Group 05 – Project 3 Write Up

This analysis was conducted by Sabina Kamalova, Amar Patil, Andrew Montemayor, and Earl Lewis. In it we analyze 2019 Coronavirus data from the Kaggle Dataset, Covid-19 – Analysis, Visualization & Comparisons (January – February 2020). Our goal with this dataset was to highlight the nations that have most impacted by the pandemic and to comment on their recovery levels. By examining the disparities in how countries were affected, we can gain more clear understanding of the global response to COVID-19 and the ongoing efforts to rebuild in the aftermath of this crisis. We will also consider extenuating factors outside our dataset, though these were less individually quantifiable due to the limitations of the dataset.

We chose this dataset to show after pandemic effect on the world. Our desire to contribute to the understanding and management of the pandemic. To make global impact easier to understand. We developed higher level questions to help provide more detailed answers.
We chose 3 high level questions that we would use to try and address this issue.

1. Which are the top 10 countries with the highest number of COVID-19 cases?
2. Does latitude influence the number of Covid-19 cases?
3. Does the continent affect the number of COVID-19 cases?

To proceed with answering these questions, we first needed to look at the dataset we are working with and modify it to better fit the questions that we are posing. Below is our raw data set, it has 8 columns and provides a good working amount of data for each country within the data. Reviewing the columns and their titles, we considered which columns would have the most impact in answering our questions, which columns need to be manipulated, and which can be dropped all together.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 231 entries, 0 to 230
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Serial Number    231 non-null    int64
 1   Country          231 non-null    object
 2   Total Cases      231 non-null    object
 3   Total Deaths     225 non-null    object
 4   Total Recovered  211 non-null    object
 5   Active Cases     213 non-null    object
 6   Total Test       213 non-null    object
 7   Population       229 non-null    object
dtypes: int64(1), object(7)
memory usage: 14.6+ KB
```

Firstly, we dropped all the null value in the table, as the overall population or the total cases were insignificant as compared to other countries in that continent. Hence, the overall impact on the analysis was minimized.

```
# Dropping all the null value
clean_covid_df = covid_df.dropna().reset_index()
```

```
clean_covid_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 197 entries, 0 to 196
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   index            197 non-null    int64 
 1   Serial Number    197 non-null    int64 
 2   Country          197 non-null    object
 3   Total Cases      197 non-null    object
 4   Total Deaths     197 non-null    object
 5   Total Recovered  197 non-null    object
 6   Active Cases     197 non-null    object
 7   Total Test       197 non-null    object
 8   Population       197 non-null    object
dtypes: int64(2), object(7)
memory usage: 14.0+ KB
```

We did analysis on a regional basis and to do that, made a new data frame to consolidate the countries by continent for analysis. This feature engineering helped look at a high-level overview of the dataset and how geography can play a factor in total COVID cases, death cases and recovery.

.

```
# Change the column name to help in merge
lat_long_df.rename(columns={'country':'Country'}, inplace=True)
lat_long_df.head()
```

	country_code	latitude	longitude	Country
0	AD	42.546245	1.601554	Andorra
1	AE	23.424076	53.847818	UAE
2	AF	33.93911	67.709953	Afghanistan
3	AG	17.060816	-61.796428	Antigua and Barbuda
4	AI	18.220554	-63.068615	Anguilla

```
# merging to get continent to help groupby continent
covid_fin = pd.merge(clean_covid_df,continent_df,how="left",on=["Country"])
covid_fin.head()
```

	index	Serial Number	Country	Total Cases	Total Deaths	Total Recovered	Active Cases	Total Test	Population	Continent
0	0	1	United States	104,196,861	1,132,935	101,322,779	1,741,147	1,159,832,679	334,805,269	North America
1	1	2	India	44,682,784	530,740	44,150,289	1,755	915,265,788	1,406,631,776	Asia
2	2	3	France	39,524,311	164,233	39,264,546	95,532	271,490,188	65,584,518	Europe
3	3	4	Germany	37,779,833	165,711	37,398,100	216,022	122,332,384	83,883,596	Europe
4	4	5	Brazil	36,824,580	697,074	35,919,372	208,134	63,776,166	215,353,593	South America

Additionally, to show the the overall impact of COVID cases on the world map, a data for latitude and longitude for each country were added to the above DataFrame.

```
# check why all continent are not showing up
cols = ['Continent']
# mask = pd.isna(covid_data[cols])
mask = covid_fin[cols].isnull().any(axis=1)
covid_fin.loc[mask,]
# new_data.info()
# rows_with_missing_values = covid_data[mask]
```

index	Serial Number	Country	Total Cases	Total Deaths	Total Recovered	Active Cases	Total Test	Population	Continent

```
# merge to get latitude and longitude
covid_fin = pd.merge(covid_fin,lat_long_df,how="left",on=["Country"])
covid_fin.head()
covid_fin.tail(10)
```

	index	Serial Number	Country	Total Cases	Total Deaths	Total Recovered	Active Cases	Total Test	Population	Continent	country_code	latitude	longitude
187	210	211	Saint Kitts and Nevis	6,592	46	6,537	9	126,903	53,871	North America	KN	17.357822	-62.782998
188	211	212	Turks and Caicos	6,522	38	6,451	33	611,527	39,741	North America	TC	21.694025	-71.797928
189	212	213	Sao Tome and Principe	6,280	77	6,202	1	29,036	227,679	Africa	NaN	0.255436	6.602781
190	213	214	Palau	5,986	9	5,976	1	68,820	18,233	Oceania	PW	7.51498	134.58252
191	216	217	Nauru	4,621	1	4,609	11	20,509	10,903	Oceania	NR	-0.522778	166.931503
192	217	218	Anguilla	3,904	12	3,879	13	51,382	15,230	North America	AI	18.220554	-63.068615
193	218	219	Macao	3,488	120	3,357	11	7,850	667,490	Asia	MO	22.198745	113.543873

Finally, to help in data analysis datatype for some of the field was converted from string to Int or Float. The conversion must be done twice, first one was to convert number separated with comma using astype and other was decimal number using to_numeric function.

```
col_list = ["Total Cases","Total Deaths","Total Recovered","Active Cases","Total Test","Population"]
for col in col_list:
    covid_fin[col] = covid_fin[col].str.replace(',', '').astype(int, errors='ignore')
covid_fin.info()
```

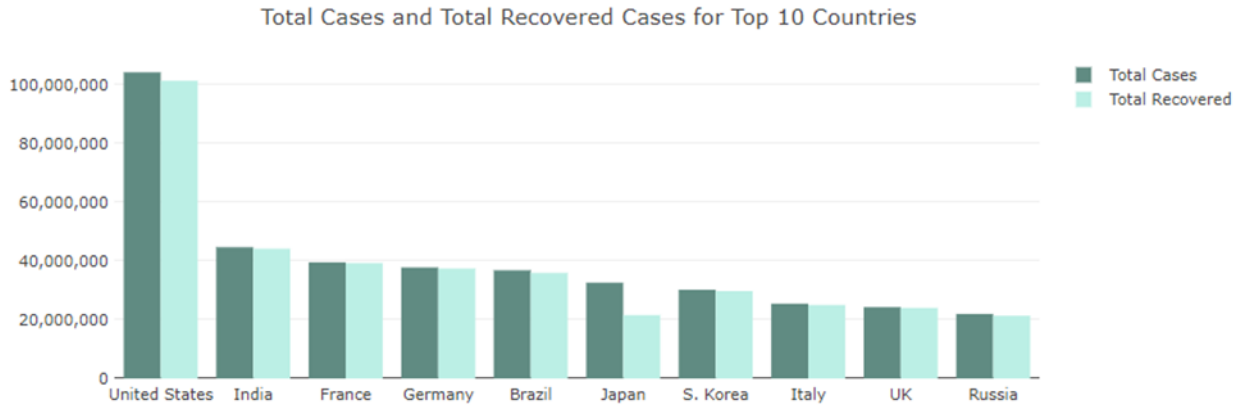
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 197 entries, 0 to 196
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   index            197 non-null    int64
 1   Serial Number    197 non-null    int64
 2   Country          197 non-null    object
 3   Total Cases      197 non-null    int32
 4   Total Deaths     197 non-null    int32
 5   Total Recovered  197 non-null    int32
 6   Active Cases     197 non-null    int32
 7   Total Test       197 non-null    int32
 8   Population       197 non-null    int32
 9   Continent        197 non-null    object
 10  country_code     185 non-null    object
 11  latitude         197 non-null    object
 12  longitude        197 non-null    object
dtypes: int32(6), int64(2), object(5)
memory usage: 15.5+ KB
```

```
col_list = ["latitude","longitude"]
for col in col_list:
    covid_fin[col] = pd.to_numeric(covid_fin[col], errors='coerce')
covid_fin.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 197 entries, 0 to 196
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   index            197 non-null    int64
 1   Serial Number    197 non-null    int64
 2   Country          197 non-null    object
 3   Total Cases      197 non-null    int32
 4   Total Deaths     197 non-null    int32
 5   Total Recovered  197 non-null    int32
 6   Active Cases     197 non-null    int32
 7   Total Test       197 non-null    int32
 8   Population       197 non-null    int32
 9   Continent        197 non-null    object
 10  country_code     185 non-null    object
 11  latitude         195 non-null    float64
 12  longitude        196 non-null    float64
dtypes: float64(2), int32(6), int64(2), object(3)
memory usage: 15.5+ KB
```

Question 1: Which are the top 10 countries with the highest number of COVID-19 cases?

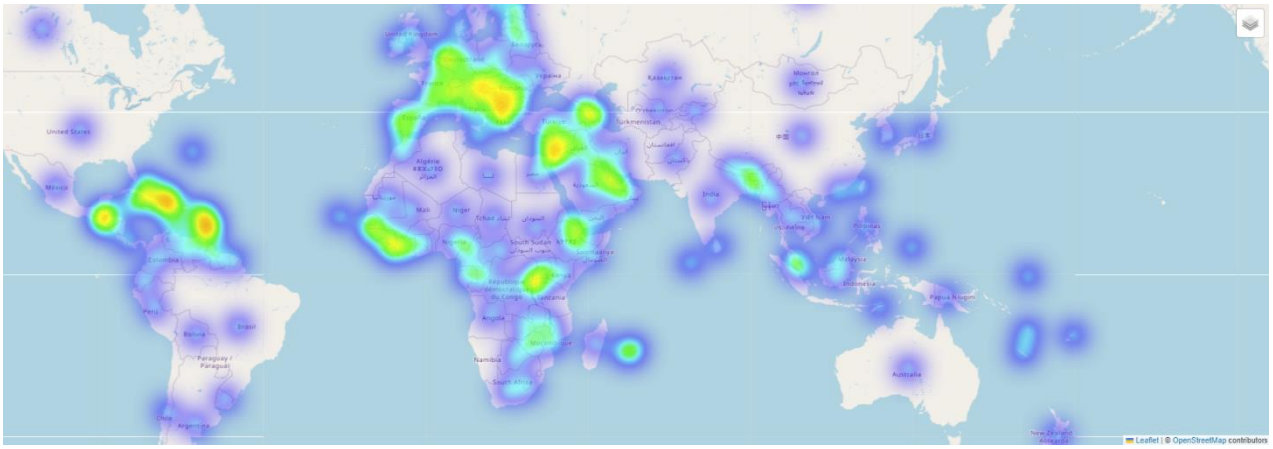
The top 10 countries that have experienced significant outbreaks due to various factors including large populations, high population density, and extensive international travel. The numbers have been influenced by the availability and accuracy of testing, reporting practices, and public health responses. When comparing total COVID-19 cases with recoveries, countries



with the highest case numbers, like the United States (100M), India(44M), and France(40M), also report substantial recoveries. This may be because of strong quarantine process, healthcare systems, extensive testing, and access to vaccines, but these needs to be investigated for each countries. The criteria for recovery used and documented needs further investigation.

Question 2: Does latitude influence the number of COVID-19 cases?

Total Cases and Latitude were considered to study this impact. It can be observed that the Total Case density is increased in areas closer to the equator, which needs to be further investigated. The increase in density near the equator may be due to factors like higher population density warmer climates that encourage social gatherings, and increased travel in these regions. But this cannot be concluded until further analysis is performed on the countries and cities closed to the equator versus moving away from the equator.



Question 3: Does the continent affect the number of COVID-19 cases?

For this analysis total population of each country was not considered as countries with higher population might skew the analysis. To offset this impact and to analyze the Covid cases only we decided to use total test, total cases and recovery.

It can be seen from table below (last column) Continent does not have any impact on recovery, though Asia shows relatively higher recovery compared to other continents. If this is compared with total test with population ration Asia has very less test percentage as compared to Europe, Oceania and North America.

The death cases were highest in Africa and lowest in Europe based on total test ratio. The higher death rates in Africa may be attributed to weaker healthcare infrastructure, delayed vaccine access, or higher prevalence of pre-existing conditions, limited data reporting and crowded living conditions, while Europe's stricter public health measures helped reduce the virus's impact.

Continent	TotalTest_TotalCases_Ratio	TotalTest_TotalDeath_Ratio	TotalTest_Population_Ratio	TotalCases_Recovered_Ratio
Oceania	6.44	3633.76	2.10	1.01
Europe	11.39	1567.33	4.16	1.02
Asia	11.50	1377.90	0.55	1.08
North America	10.56	815.66	2.23	1.04
Africa	9.39	454.64	0.08	1.05

COVID-19 Website

Our group developed a comprehensive COVID-19 website designed to provide detailed information and visualizations related to the global impact of the pandemic. The website is structured to offer insights into COVID-19 through various sections, including a home page, a dashboard with interactive data, a map for geospatial analysis, an "About Us" page that introduces the team, and a citation page that credits data sources used.

Home Page:

The home page sets the tone for the website by discussing the broad impacts of COVID-19 on global society. It emphasizes the economic, social, and health-related challenges brought about by the pandemic and introduces the website's purpose: to provide a global perspective on these issues ("home.html").

Dashboard:

The dashboard is a central feature of the website, allowing users to interact with COVID-19 data. Users can filter data by continent and minimum cases, and view visualizations such as sunburst charts and bar charts. These visualizations display data on COVID-19 cases, deaths, and recoveries, helping users explore the pandemic's impact across different regions and understand trends over time ("dashboard.html").

Map:

The map feature provides a geospatial representation of COVID-19 cases worldwide. This section likely uses tools like Leaflet to render maps and create heat maps that show the distribution of cases based on latitude and longitude. This visual approach helps users grasp the geographical spread of the virus more intuitively ("map.html").

The code to create the interactive map was executed through the use of leaflet.js, and upheld a primary focus on visualizing geographical data through a heatmap. The programming process began with the initialization of two base layers: a street map obtained through OpenStreetMap and a topographical map obtained through OpenTopoMap. These two layers are the visual foundation for the map, offering different perspectives to the users such as street-level details or terrain features.

The code then creates a heatmap overlay by processing data points, including “latitude”, “longitude”, and “Total Cases” (the intensity value). The Leaflet.heat plugin is used to generate this heatmap, providing configurable option for radius, blur, and intensity scaling. Which can be set to ensure the effectiveness of the visualization of these data hotspots

Next, the layer controls are created to allow the base layer to be switched and to allow the heatmap to be toggled on or off. The interactive map’s view is set to be centered where the

equator and prime meridian meet and is defaulted to show the street map and have the heatmap turned on.

Finally, the “doWork” function then fetches the data from the API, populating the map with real-time information.

The implementation of this interactive map came to use when addressing the first question in the project hypothesis set, “Does latitude influence the number of COVID-19 cases?”. “Total Cases” and “Latitude” were analyzed to study this impact; and though the study was inconclusive we were able to observe one certain trend from the map visualization. We were able see that the “Total Case” density increased in regions that were closets to the equator. Though, we were not able to draw any conclusive evidence from this trend; we quickly came to the realization that the topographical base layer would be a great foundation to begin an additional study as to the reason for the increase in cases in these regions nearest the equator.

About Us Page:

This page introduces the team members, providing a personal touch to the project. Each member's background is outlined, highlighting their expertise and contributions. This helps build credibility and showcases the diverse skill set within our team ("about_us.html").

Citation Page:

The citation page is essential for maintaining academic and professional integrity. It lists all the data sources used in the project, ensuring transparency and giving credit to original authors and datasets. This section shows that the project is well-researched and relies on credible sources ("citation.html").

Technical Aspects:

The website likely utilizes a backend system, possibly built with Flask (as indicated using Python scripts like app.py and sqlHelper.py), to manage data retrieval and processing. The SQLite database (covid. SQLite) serves as the data storage, housing information on COVID-19 cases, which is then used to generate the visualizations and data tables on the site.

The website is styled using Bootstrap, with the "Minty" theme providing a clean and modern look. The consistent use of this theme across all pages ensures a cohesive user experience. The use of JavaScript libraries like D3.js and Plotly.js for visualizations, along with DataTables for data management, adds a high level of interactivity to the site, allowing users to engage with the data meaningfully.

Conclusion:

The COVID-19 website created by our group is a well-rounded project that effectively combines data analysis, visualization, and user interaction to provide a comprehensive understanding of the pandemic's global impact. The website is not only informative but also accessible, thanks to its user-friendly design and interactive features. The inclusion of various

sections, such as the map and dashboard, ensures that users can explore the data from multiple perspectives, making the project a valuable tool for anyone interested in the effects of COVID-19 worldwide.

Bias and Limitations

When considering our dataset are several bias and limitations that are found within that need to be considered when drawing conclusions. It is limited by the fact that there are a high number of socio-economic factors that would impact datapoints across this analysis that aren't necessarily quantifiable with this type of dataset. A countries individual politics and polices would play into this point as well. We encountered a few null values, which were dropped from the dataset to minimize errors during our cleaning process. Unfortunately, some countries still lack the necessary technology, infrastructure and resources to accurately report and calculate cases. The dataset should include the criteria or process used to determine a positive COVID-19 case, as total cases are the sum of active cases, total deaths, and total recoveries. While deaths can be recorded with some certainty, the criteria used to attribute a death to COVID-19 must be clarified. Additionally, the total cases and deaths reported in some countries appear significantly lower compared to other developed nations, which raises concerns about the consistency and accuracy of reporting and testing practices across different regions, as illustrated in the snapshot below.

Serial Number	Country	Total Cases	Total Deaths	Total Recovered	Active Cases	Total Test	Population
91	China	503,302	5,272	379,053	118,977	160,000,000	1,412,000,000

Several sources have raised concerns about the efficacy of COVID-19 test kits, citing issues with accuracy that resulted in both false negatives and false positives. Hence either total cases, total recovered, or total deaths are questionable.

Conclusion

Reviewing our analysis, we find that COVID-19 case density is higher in regions closer to the equator, but continent or location does not significantly impact recovery rates. The top 10 countries with the most significant outbreaks, including the U.S., India, and Brazil, also report high recovery rates. The reasoning of this needs to be investigated further. Africa's higher COVID-19 death rates are attributed to weaker healthcare systems, delayed vaccine access, and more prevalent pre-existing conditions, while Europe's stricter public health measures led to lower mortality. Though this data gives a very high-level analysis of the COVID impact, a detail analysis is required to make a concrete conclusion. For this, criteria used for recording the data must be set, with vaccine availability, pre-existing conditions if any, restrictions and quarantine process followed.

Bibliography

https://www.kaggle.com/code/imdevskp/covid-19-analysis-visualization-comparisons

https://www.kaggle.com/datasets/paultimothymooney/latitude-and-longitude-for-every-country-and-state

https://www.healthdata.org/news-events/newsroom/news-releases/covid-19-had-greater-impact-life-expectancy-previously-known#:~:text=Key%20takeaways%3A,of%20the%20past%2072%20years

https://www.theguardian.com/world/2022/dec/23/how-accurate-are-chinas-covid-death-numbers

https://www.kff.org/mental-health/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9052715/

6. OpenAI. "ChatGPT Response to Question about MLA Citation." Accessed 6 June 2024, chat.openai.com.

7. Sustainable Development Goals https://www.un.org/sustainabledevelopment/energy/#:~:text=Goal%207%20is%20about%20ensuring,targets%20%E2%80%93%20but%20not%20fast%20enough.

8.Recession drove 2009 energy consumption lower. https://www.bp.com/en/global/corporate/news-andinsights/press-releases/recession-drove-2009-energy-consumption-lower.html

9.Effects of the 2008–2010 automotive industry crisis on the United States. https://en.wikipedia.org/wiki/Effects_of_the_2008%E2%80%932010_automotive_industry_crisis_on_the_United_States

10. Global Energy Review 2021 https://www.iea.org/reports/global-energy-review-2021/renewables Assessing the effects of economic recoveries on global energy demand and CO2 emissions in 2021.