VAST 2010 MINI CHALLENGE - 2

MC2 - Characterization of Pandemic Spread

Sindhu Sree Aita
Master of Science in Computer Science
Arizona State University
Tempe, Arizona, USA
saita1@asu.edu

Avinash Senthil Kumaran
Master of Science in Computer Science
Arizona State University
Tempe, Arizona, USA
asenth10@asu.edu

Vineethkrishna Vemireddy
Master of Science in Computer Science
Arizona State University
Tempe, Arizona, USA
vvemired@asu.edu

Sucharith Reddy Desireddy
Master of Science in Computer Science
Arizona State University
Tempe, Arizona, USA
sdesired@asu.edu

Aishwarya Reddy Dwaram
Master of Science in Computer Science
Arizona State University
Tempe, Arizona, USA
adwaram1@asu.edu

Madhura Ganga
Master of Science in Computer Science
Arizona State University
Tempe, Arizona, USA
mganga@asu.edu

I. INTRODUCTION

The Mini Challenge-2 is about the pandemic that broke across the world in 2009. The health officials gave specifics regarding the epidemic, including hospital admittance and mortality statistics for cities affected by the outbreak, which will aid us in describing the disease's spread across countries and handle them from any further occurrence. Inspecting the given data, pre-processing it and plotting different visualizations will help us in future study. We have taken attributes like gender, age, mortality rate, frequency of symptoms etc., into consideration to build these visualizations.

A. Key Words

Multi-line chart, Stacked bar chart, , Tree map chart, Bubble chart

B. Tools/Languages used

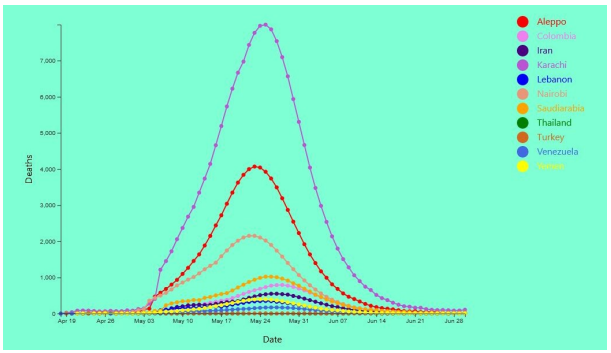
D3.js, CSS, HTML, Bootstrap, python tools - Numpy, pandas and mlxtend

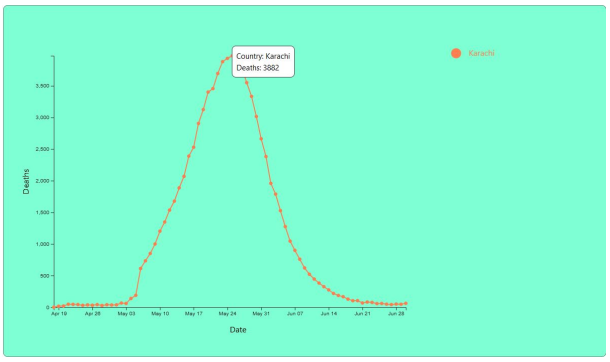
II. VISUALIZATION DESIGN

To display the data and respond to the questions/challenges provided in the mini challenge, we used five distinct visualizations.

A. Visualization 1 : Multi-Line Chart analysing the death toll of patients in each country/city

In this visualization, we estimated the number of patients that died on a given day and managed to represent it using a line chart. We used data pre-processing to acquire the death count in different cities because the given data was so large. Initially, we computed the death toll for all patients who died as a result of the epidemic. For better analysis, this method is applied to all countries. This is a multi-line chart that is used to determine trends and anomalies across countries. The lines for each country/city are represented by different color hues. We may filter the graph for each country/city or gender for better analysis.





Interactions involved:

When you hover the mouse over a specific path, respective country/city gets illuminated which assists us with distinguishing the path of a specific country/city from the drop-down choice.

Used drop-down interaction for country/city.

Used on-click Gender selection on navigation bar.

Anomalies Found:

As the death toll for Thailand and Turkey is too low, the graph for these two countries is not consistent and distorted.

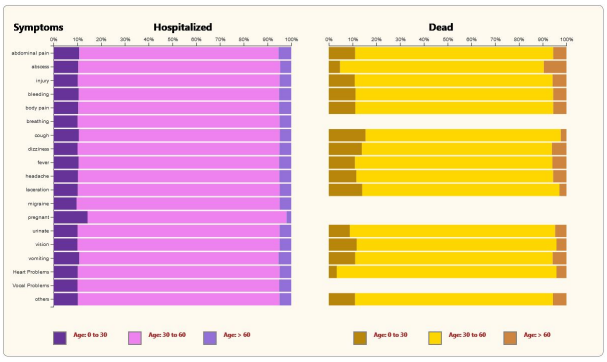
Observations:

We can segregate the graphs based on the patients gender and country/city. Upon breaking down the data based on gender, we can clearly get the similar patterns with no abnormalities. From the visualization it is evident enough that majority of the deaths are happening on May 24th.

B. Visualization 2 : Stacked bar graph showing effect of symptoms on different age groups

In this visualization, we used the stacked bar graph to analyze the effect of symptoms in different age groups. We did the data pre-processing to obtain the count and percentage for men, women and both who are hospitalized with a specific symptom and who died after hospitalizing.

Age is categorized into three categories 0 - 30 years, 30 to 60 years and greater than 60 years. We calculated the count of the number of people for every group for every symptom for men, women and both. From this, we can compare different symptoms from a particular group for a particular gender. Two stacked bar graphs are presented one for hospitalized patients and the other for the dead patients.



Interactions involved:

When you hover the mouse over the graph it gives the count and the percentage of the patients that fall under a particular symptom.

Used drop-down interaction for country/city.

Used on-click Gender selection on navigation bar.

Anomalies Found:

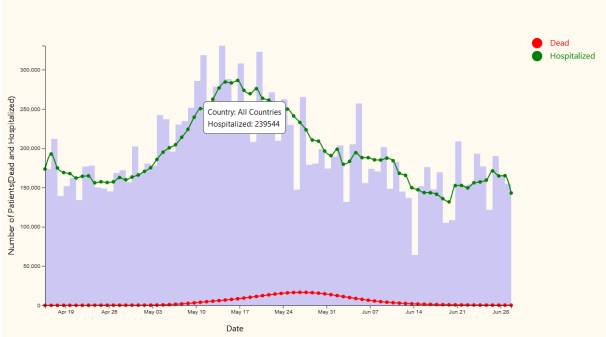
We found that the data set is not helping much to interpret the visualization, as we can see for a particular age group all the symptoms have a similar percentage. If the data could've been more precise the visualization would be more appealing.

For countries like Thailand and Turkey they are very few people who died after hospitalizing, hence the blank data for some symptoms.

C. Visualization 3 : Analysis of pandemic spread (hospitalized and dead) in different cities

In this visualization, we used density estimation to represent the number of people hospitalized every day. We obtained the dead patients' data from the pre-processing and also obtained the mean of the count of the number of people who died and were hospitalized in the previous 6 days for each country/city and each gender.

Each bar in the graph represents the count of patients hospitalized on a particular day in a particular country/city for each gender. Here the green line represents the mean of the hospitalized patients over the previous 6 days. And the red line represents the mean of the patients who died after being hospitalized over the previous 6 days.



Interactions involved:

When you hover the mouse over the green line it displays the number of patients hospitalized and when you hover over the red line it displays the number of patients who died after being hospitalized.

Used drop-down interaction for country/city.

Used on-click Gender selection on navigation bar.

Anomalies Found:

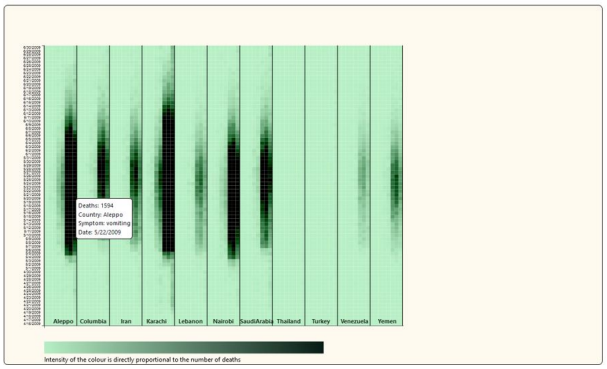
For countries like Thailand and Turkey the death toll is too low every day, so taking the 6 days mean for that low values is resulting in much smaller values, hence the flat red line.

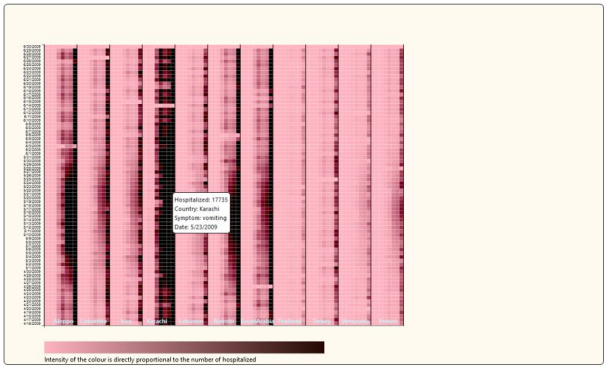
D. Visualization 4 : Treemap showing the death and hospitalized count in the form of color luminance

In this visualization, we used a treemap to represent the number of patients affected by the symptoms in each country/city. We pre-processed the data to obtain the number of people who died and the number of people hospitalized on

each date for each country/city. First chart shows data for the dead patients, whereas the second chart shows data for the hospitalized patients.

Hear each section represents each country/city, 11 countries hence 11 sections. We picked the top 8 symptoms to show in each country/city. Each pixel in a section represents each symptom. The colour gradient is for the count of the dead or hospitalized patients, dark for the higher count, and light for the lower count.





Interactions involved:

When you hover the mouse on a pixel it displays the death or hospitalized count, the country/city selected, symptom selected, and the date.

There is a drop-down to select either dead or hospitalized patients.

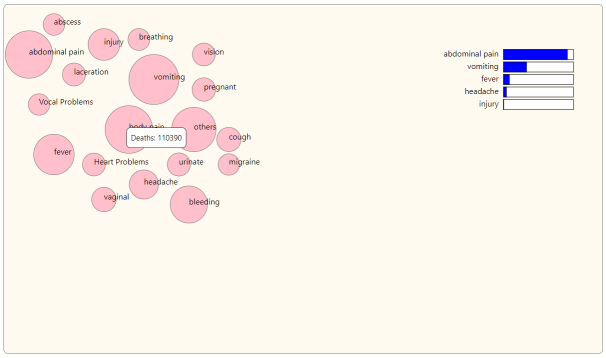
Anomalies Found:

For countries like Thailand and Turkey, the death toll is too low every day. Hence, the low luminance for these two countries.

E. Visualization 5 : Bubble Chart showing the commonality between predominant symptoms in dead patients - Innovative

In this visualization, we used a bubble chart to show the predominant symptoms in patients who died. We pre-processed the data to the dead patients' data from the two datasets. We grouped the symptoms and obtained the commonality of the symptoms by matching each symptom with the count of the other predominant symptom of the dead patient. This is done for dead patients only.

Each bubble in the chart represents each symptom. The size of the bubble changes with respect to the count of the patients who died with that symptom. When you click on each symptom it displays 5 bars on the top right corner of the page which shows the predominant symptoms which are in common with the selected symptom in dead patients.



Interactions involved:

When we hover the mouse over the bubble it shows the count of patients who died and has that particular symptom.

When we click on a bubble it opens 5 bars showing the count of dead patients who has the common symptoms.

When you hover on the bar chart it shows the death count who has the symptom common with the selected symptom.

Anomalies Found:

For some symptoms like vision, cough, pregnancy, etc., the common symptom count shows zero. This is due to the very less count in the dead patients with that particular symptom.

III. ABOUT DATASET

For every country/city, we are given two datasets. The first dataset gives data on hospitalized patients, this dataset gives attributes like patient ID, age, symptoms, Date of hospitalization and gender. Here the symptoms are many-to-many (i.e., one person can have multiple symptoms and one symptom can be possessed for multiple persons). The second dataset gives data on the deaths of the hospitalized patients, this dataset gives two attributes, patient ID and Death date.

Here patient ID acts as a unique key and helps us to club the two datasets. By clubbing, we can assess the patient status whether he is dead or alive after hospitalized. Investigating these datasets will assist us to analyze and understand the intensity of the outbreak by drawing important analytics and this investigation can likewise be utilized for future study in case a new pandemic arrives.

IV. PRE-PROCESSING

To match the requirements, we did data pre-processing to get a better understating of the data. There were many duplicate symptoms in the dataset such as vomit, vomting, vomitting, vomiting for vomiting and ear, ache, ankle, hurts, leg, pain, pn, neck, stomach for body pain, these symptoms are supposed to be the same but were given with different duplicate names, we clubbed such symptoms to be one symptom which helped us to accurate the visualizations. Also, there were unnecessary characters involved in the data, which will lead us to wrong grouping and eventually to bad conclusions. We handled such characters and grouped them appropriately. In some instances, there is no space after the ',' delimiter, in that case, we split the symptom into two different symptoms and assigned them to the respective patients. After grouping, we observed that some symptoms have a significant effect and some don't, so we considered the top symptoms which covered the majority of the dataset.

A. Visualization 1:

Here we clubbed both the datasets using a unique key - patient ID and obtained the data for patients who are dead. From that dataset, we obtained the death toll on each day for each country/city. We got the pattern from the death toll.

B. Visualization 2:

Here we grouped the data in terms of the age groups (0 – 30, 30 – 60, greater than 60). We obtained the data for a number of patients carried out by every symptom and analyzed every symptom based on the age groups in each country/city.

C. Visualization 3:

Here we grouped the datasets to get the data for dead patients and hospitalized patients. Also, calculate the mean of the number of patients who died on a particular date for the corresponding previous 7 days inclusive of the selected date.

D. Visualization 4:

Here we obtained the data of the number of people who died and the number of people hospitalized on each date for each country/city to get the pattern of dead and hospitalized patients in each country/city.

E. Visualization 5:

Here we obtained the dead patients' data from the two datasets. We grouped the symptoms and obtained the commonality of the symptoms by matching each symptom with the count of the other predominant symptom of the dead patient. This is done for dead patients only.

V. LESSONS LEARNT

1) We learnt how to manage teamwork and finish the project deliverables on time.
2) Learnt how to implement the feature extraction on the given dataset and get interesting and desired trends to use them in the data visualization.
3) Since the given dataset is too large, it took us hours to extract the pre-processed data file from the code. We used Google Colab and used multiple cores in the local environment for faster implementation using python.
4) Learnt how to integrate the code which we worked individually to form a single and fully functional project.
5) Learnt how to select up-on charts which is best suitable for the given data to draw meaningful conclusions

VI. ANOMALIES DETECTED

1) In visualization - 1 as the death toll for Thailand and Turkey is too low, the graph for these two countries is not consistent and distorted.
2) In visualization - 2 for countries like Thailand and Turkey they are very few people who died after hospitalizing, hence the blank data for some symptoms.
3) In visualization - 3 for countries like Thailand and Turkey the death toll is too low every day, so taking the 6 days mean for that low values is resulting in much smaller values, hence the flat red line.
4) In visualization - 4 for countries like Thailand and Turkey, the death toll is too low every day. Hence, the blank map for these two countries.
5) In visualization - 5 for symptoms like vision, cough, pregnancy, etc., the common symptom count shows zero. This is due to the very less count in the dead patients with that particular symptom.

REFERENCES

[1] Whiting, M., Haack, J., and Varley, C. Creating realistic, scenariobased synthetic data for test and evaluation of information analytics software. Proc. of BELIV'08, ACM (2008)

[2] G. Grinstein, S. Konecni, C. Plaisant, J. Scholtz and M. Whiting, "VAST 2010 Challenge: Arms dealings and pandemics,"2010 IEEE Symposium on Visual Analytics Science and Technology, Salt Lake City, UT, 2010, pp. 263-264, doi: 10.1109/VAST.2010.5649054.Conference Location:El Paso, Texas USA

[3] Fernanda, B., Martin, W., Jesse, K.,"ManyEyes: a Site for Visualization at Internet Scale", IEEE Transactions on Visualization and Computer Graphics, Irvine, 10.1109/TVCG.2007.70577, Nov. 2007.

[4] Enamul, H., Maneesh, A., "Visual Style and Structure of D3 Visualizations," IEEE Trans. Vis. Comput. Graphics, vol. 26, no. 1, pp. 1236–1245, Jan. 2020.

[5] https://observablehq.com/@d3/bubble-chart