VAST 2010 MINI CHALLENGE - 2

MC2 - Characterization of Pandemic Spread

Chandrahaas Chintalaboguda
Computer Engineering and Computer Systems
Arizona State University
Tempe Arizona USA
Cchintal@asu.edu

Arun Raj Deva
Computer Science
Arizona State University
Tempe Arizona USA
adeva@asu.edu

Sai Teja Vishal Jangala
Computer Science
Arizona State University
Tempe Arizona USA
sjangala@asu.edu

Vamsi Krishna Kanagala
Computer Science
Arizona State University
Tempe Arizona USA
vkanaga1@asu.edu

Pavan Kalyan Reddy Thota
Computer Science
Arizona State University
Tempe Arizona USA
pthota3@asu.edu

Manikanta Vankayala
Computer Science
Arizona State University
Tempe Arizona USA
mvankaya@asu.edu

1. INTRODUCTION

VAST 2010 Mini challenge-2 is about a major outbreak that occurred in 2009 across the world. We were provided with the information across 11 countries about the pandemic, by the Health officials which will help them to characterize the spread of the disease in those countries. Analyzing this information will help us understand the seriousness of this or any future pandemic and extract statistics. Here we are given spatio-temporal data, and studying this data after its preprocessing, we visually create data flows that will make the realization of data easy for future study. In this project we have taken the attributes like age, gender, mortality rate, symptoms and temporal patterns of the disease into consideration.

Keywords

Multiline chart, Heat Map, Stacked bar chart, Bar Chart, Apriori Principle, Bubble chart, Linked chart, Histogram, Kernel Density Estimation

TOOLS/LANGUAGES USED

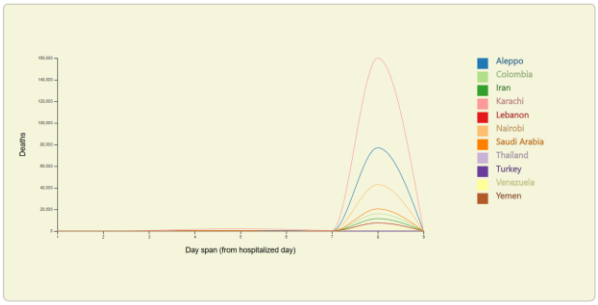
D3 Js, HTML , CSS, Bootstrap, Python libraries such as pandas, NumPy and mlxtend (for data preprocessing)

2. VISUALIZATION DESIGN

We have used five different visualizations to the display the data and answer the questions/ challenges put forth in the mini challenge.

Visualization 1 : Multi-Line Chart effectively
Analyzing the death span of patients in each country

In this visualization we have calculated the life expectancy of a patient who is hospitalized. As the preprocessed data is huge a further preprocessing is done. Initially, we have calculated the death span (Date of death - Date of hospitalized) for all the patients who are expired because of pandemic. This process is applied to all countries for better analysis. This visualization is a multi-line chart used to derive the trends and anomalies across the countries.



Color hues are used to represent the lines for each country. For the better analysis, we can filter the graph for each country.

Interactions involved:

Dropdown interaction – Country, Gender

Hover interaction – On mouse over a particular path, corresponding country will be highlighted which help us to identify the path of a particular country among the dropdown selection.

The graphs can be filtered based on the country and gender of patients. After analyzing the data across the genders, we can find the same patterns without any anomalies.

From the above graph we can clearly conclude that the most of deaths are occurring on the 8th day after the hospitalization of patient.

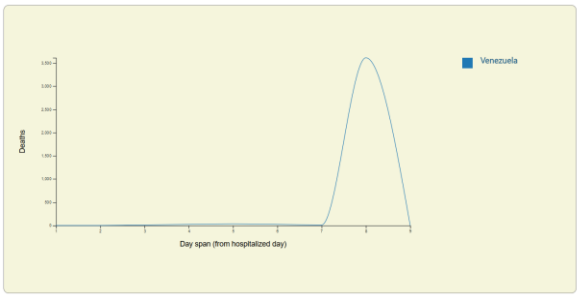


Figure 1

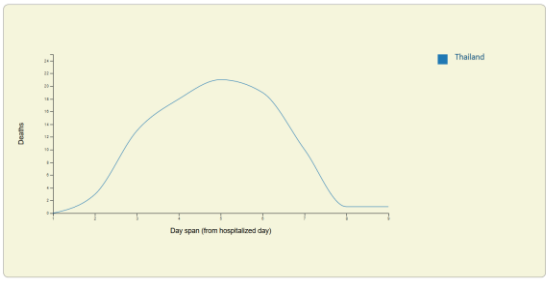


Figure 2

But, when we filter the data based on the country we can find the anomalies in Thailand and Turkey where we can observe the max deaths are on 5th day after the hospitalization of patient. (figure 2)

Whereas all other countries show the similar pattern as of figure 1.

Visualization 2 : Stacked Bar Chart

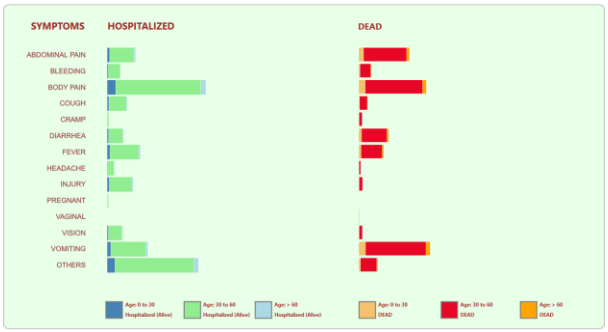
Symptom analysis among the dead and hospitalized

In this visualization, we are analyzing how each symptom is affecting the patients of particular age. For this we have partitioned the patient ages into three groups R1, R2 and R3. R1 is for the patients who are aged between 0-30 years. R2 is for those aged between 30-60 years and R3 corresponds to the patients who are older than 60 years. This trends can be easily studied with the help of a stacked bar chart.

Further preprocessing of Data: From the initial preprocessed data, we have partitioned the age groups into three groups R1, R2 and R3. And calculated the count of each age group for each symptom and for each gender.

Further we also visualized how these age trends varies between the patients who died because of the infection and those who just got hospitalized and recovered from the infection. This also explains the effect of each symptom or which symptom can be highly dangerous and can lead to death, if we consider all the age ranges .

The below figure is a snippet from our visualization.



Interactions involved:

Dropdown/Selection interaction – Gender

Hover interaction – On mouse over a particular block in the graph, the block will be highlighted. It pops out the corresponding symptom and the number of people affected with that symptom in the particular age range.

We used the color hues to differentiate age ranges and also to differentiate the dead and hospitalized patients.

Visualization 3 : Heat Map

Analyzing the death and hospitalized rates across different countries.

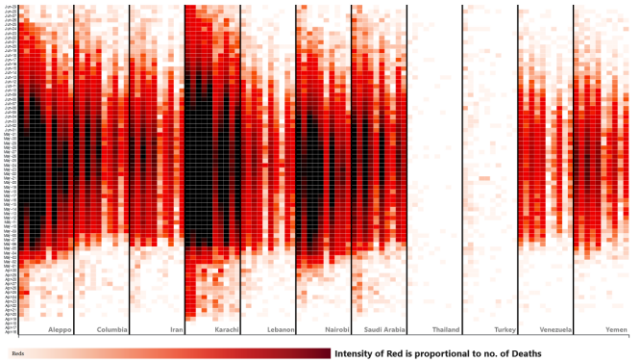
This visualization is a heat map gives the information about the death and hospitalization of the patients over time (April 20-June 13) for the top 10 symptoms of each country during the Pandemic. Using this graph, we can identify which part of the pandemic has largest number of deaths and hospitalized patients. The wave of the pandemic can be analyzed with the help of this visualization. The entire graph is divided into 11 sections and each section represents a country. Within each country, dates are mapped from top to bottom throughout the graph and top 10 syndromes are mapped from left to right on that date for each country. The intensity of the colors utilized in the visualization is proportional to number of deaths/hospitalized patients. In short, as the death/hospitalized rate increases, the color luminance increases. Suppose if we observe column one in the above visualization, the Middle part of visualization is much darker when compared to the other part of the column indicating the part of pandemic that is heavily hit having high rates of deaths/hospitalizations during that period. We have considered top 10 syndromes every country because those top 10 symptoms will have most of the information regarding the deaths and hospitalizations of the patients, which will be helpful in analyzing deaths and hospitalized rates.

Further preprocessing of Data: Initially, we have identified top 10 symptoms globally and calculated the number of deaths/hospitalized of those 10 symptoms in each country for every date of pandemic. We then visualized this data in the form of a heat map.

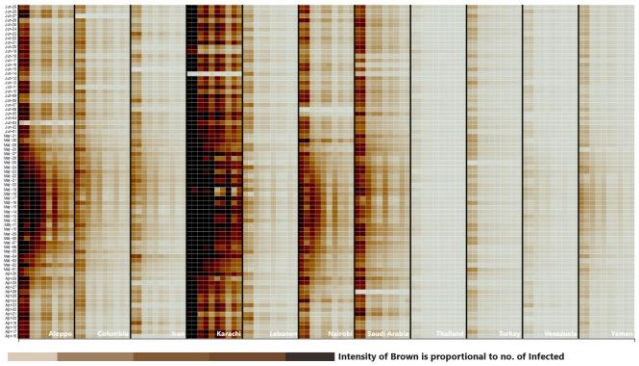
The below images are the snippets of our visualization:

1. **Dropdowns select:** Dead

9



2. **Dropdowns select:** Hospitalized(Alive)



Interactions involved:

Dropdown interaction – Death/Hospitalized

Hover interaction – Country name, Symptom name, Date and Death/Hospitalized count.

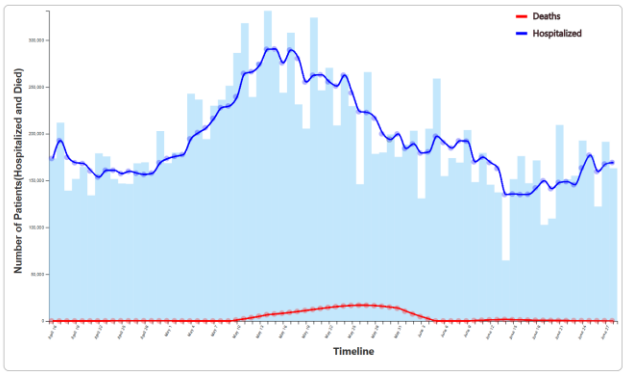
Visualization 4 : Kernel Density Estimation

Country-wise analysis of the pandemic spread

The objective of this visualization is to analyze the country-wide spread of the pandemic. Each country visualization contains bar chart and two smooth curve lines. Bar chart gives the information about the number of people hospitalized on a particular date, and blue curve line represents the windowed mean of the previous 7 death count of patients on a particular date while the red curve line represents the 7-windowed mean of the death count on a particular date.

From this visualization, we can identify the days at which the spread is high for each country and we can interpret the severity of the pandemic using the death count curve. Further, If we observe the visualization, we see that the graph shows the trends of deaths and hospitalized for each country.

The below image is a graph for one country from our visualization.



Further preprocessing:

Initially there were more variations in the data, for which we have taken windowed average of one-week data for deaths as well as hospitalized with which the difference can be analyzed better.

Interactions involved:

Dropdown interaction – Country

Hover interaction – On hovering over the line a tooltip appears which shows the count of death on red line and count of hospitalized on blue line.

Visualization 5 : Bubble Chart (Innovative)

Finding the frequent set of symptoms using Apriori Algorithm

In this visualization, we have found out the frequent set of symptoms which happened to appear together. For finding these symptoms we have used Apriori algorithm. The apriori principle can help us to reduce the number of symptom sets we need to examine. Put simply, the apriori principle states that "*If an itemset is infrequent, then all its* ***supersets*** *must also be infrequent.*"

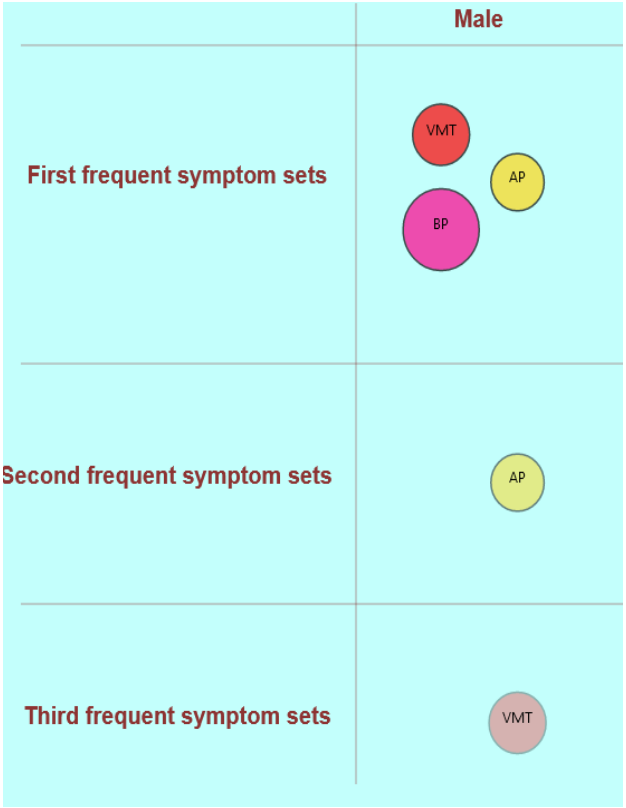
The important steps in the generation of the Apriori algorithm are as follows.

Apriori algorithm:

Step 0. Start with symptom sets containing just a single symptom.
Step 1. Determine the support for symptom sets. Keep the symptom sets that meet your minimum support threshold and remove symptom sets that do not.
Step 2. Using the symptom sets you have kept from Step 1, generate all the possible symptom sets configurations.
Step 3. Repeat Steps 1 & 2 until there are no newer symptom sets.

Further preprocessing of Data: From the initial preprocessed data we have filtered the data according to the gender and patient status. We have used the python library "*mlxtend*" to run the apriori algorithm on our datasets to calculate the support for each symptom set, then plotted the top 3 symptom sets that are

occurring together. We drew a comparison between male patients, female patients and both.



Below is a snippet of our visualization for a particular country.
We used both the color hues and luminance to differentiate top 3 frequent symptom sets.
Color hues – to differentiate the symptoms.
Luminance – to differentiate the symptom frequency.
For example, in the above figure, the color of the body pain symptom circle (BP) is more intense in the most frequent symptom sets part than in the latter. Also, the radius of the symptom is proportional to the number of the patients(hospitalized or dead) having that symptom in each country.

Interactions involved:

Dropdown/Selection interaction – Death/ Hospitalized, Country

Hover interaction – On mouse hover a particular circle/bubble in the graph, the block will be highlighted. It pops out the corresponding symptom and the number of people affected with that symptom.

3. DESCRIPTION OF THE DATASET

In this mini challenge, we are provided with the two different datasets for each country. One dataset provides information about the patients who are hospitalized, and it includes the details such as patient ID, date of hospitalization, age, gender and the symptoms. It may be possible that a particular patient can have multiple symptoms. And the second dataset contains the information about the patients who were expired after hospitalization. It includes the attributes patient ID and the date of death. We used the patient's ID as a key and joined both the datasets to identify if the patient has survived the pandemic. Analyzing all this information will help us to understand the severity of the pandemic by extracting the useful statistics and this analysis can also be used for better analysis in future pandemics(if any).

4. USING THE SYSTEM TO ANSWER THE PROBLEMS IN THE MINI CHALLENGE

In this project, we have analyzed the records that were given, to characterize the spread of the pandemic by considering symptoms of the disease, mortality rates, temporal patterns. And also, we have compared the outbreak across all the countries to find the anomalies and drew conclusions based on the facts generated.

In order to server the requirements, we have preprocessed the data for better understanding. We used the python libraries such as pandas and NumPy for preprocessing. As a first step, we have analyzed the different syndromes given in the dataset and assigned a common alias for the same type of syndrome. For example, abdomen, abd, ab pain are all mapped to the same symptom abdomen pain. In few other cases, we had multiple syndromes without any spaces or extra characters(like -, _ etc.,) and we also mapped them to a common symptom. For example, coughPain, cough-pain, cough_pain are all mapped to Cough Pain. We observed that among all the symptoms present in the given raw data, we have considered the top 26 symptoms(after re-labelling) as we found that these symptoms alone covered 94% of the total dataset.

Task Abstraction -

With this processed data, we answered the questions put forth in the mini challenge in the following way.

In visualization 1, we took the information about the dead patients into consideration and we have derived the death span (Date of death – Date of Hospitalized) of each patient. By plotting these information, we have found the trends and outliers.

{Action, Target} – {(Derive, Annotate),(Trends, Outliers)}

In visualization 2, we have searched and queried the data to find out the age range of the patients effected by each symptom to find out the severity of the symptom in each age group.

{Action, Target} – {(Search, Query),(Features, Distribution)}

In visualization 3, we have summarized the death/hospitalized count for each of the top 10 symptoms in each country and we found out the trends in the impact(death/hospitalized) of each symptom in each country.

{Action, Target} – {(Summary, Compare),(Trends, Outliers)}

In visualization 4, we have calculated the average count of deaths/hospitalized over a week to study the trends in the pandemic for each country.

{Action, Target} – {(Derive, Compare),(Trends, Features)}

In visualization 5, we have consumed the data to find the correlation between the symptoms of both hospitalized and dead patients separately and found the top 3 most frequently occurring symptom sets in each country.

{Action, Target} – {(Present, Compare, Summarize),(Features, Correlation)}

5. DISCUSSION

A) LESSONS LEARNT

- As the datasets we are using is about 15 million records, it took a lot of time while loading the data directly into the front-end and we also encountered issues while preprocessing the data directly. So, we implemented an alternative way to filter the data by writing python scripts. And thus, we used the individual filtered data while loading the web pages.

- For proposing creative visualizations, we went through several blogs on data visualizations especially in D3 and understood how we can use simple D3 graphs to represent complicated analysis or complex data and we succeeded in implementing them in our project.

- Successfully incorporated Data Mining principles for feature extraction and finding unique patterns in the data and further we used these findings to implement them in our visualization.

- We have learnt different methods to link the charts/visualizations and succeeded in implementing one such model in our dashboard and made our application user friendly and easy to navigate.

- With proper planning and execution, we have implemented our visualizations as per our initial proposal.

B) FUTURE WORKS

- The unequal distribution of the data across the countries might have resulted in undesired anomalies and outliers. For example, we can observe in our visualization 1 that the peaks exists on day 5 at 90 deaths. Whereas for the remaining countries, the average deaths reported is in thousands. The same can be observed in the case of Turkey because of this unequal distribution of data. This might have not been the case with good amount of data.

- Similarly, we can observe unequal distribution of data across the age groups of all the patients. This can be observed in visualization-2 where we have huge amount of data for the patients who are aged between 30-60 years.

- In the future, interactions between the visualizations could have been improved with some advanced techniques like linked highlighting, overview + detail etc.,

6. REFERENCES

https://observablehq.com/@d3

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

J. S. Yi, Y. a. Kang, J. Stasko and J. A. Jacko, "Toward a Deeper Understanding of the Role of Interaction in Information Visualization," in IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 6, pp. 1224-1231, Nov.-Dec. 2007, doi: 10.1109/TVCG.2007.70515.

B. Wu, D. Zhang, Q. Lan and J. Zheng, "An Efficient Frequent Patterns Mining Algorithm Based on Apriori Algorithm and the FP-Tree Structure," 2008 Third International Conference on Convergence and Hybrid Information Technology, Busan, 2008, pp. 1099-1102, doi: 10.1109/ICCIT.2008.109.