**Visualizing the Global Footprint of Historical Disease Outbreaks**

**Objective**

Pandemics have been a recurring challenge throughout human history, significantly affecting global health systems, economies, and societies. The severity of these outbreaks varies, but their consequences are often devastating, with millions of lives lost, economies disrupted, and healthcare systems strained to their limits. From the Bubonic Plague of the 14th century to the more recent COVID-19 pandemic, each of these events has left an indelible mark on the world. The death tolls, ranging from hundreds of thousands to over a hundred million in some cases, highlight the immense scale of these health crises and the urgency of effective preparedness.

In collaboration with a healthcare consulting firm, we developed a simple visualization tool to help clients understand the historical impact of pandemics. The firm needed a straightforward solution to highlight trends and death tolls from major outbreaks, providing a framework for assessing the historical context. We created foundational visualization focusing on key pandemics and their effects on various regions. This analysis offered valuable insights into how past pandemics influenced hospital operations, aiding organizations in better preparing for future crises.

The data was sourced from publicly available datasets, historical records, and research databases that catalog information on pandemics. The original dataset provides information on various historical epidemics and pandemics, listing their names, dates, affected locations, suspected diseases, and estimated death tolls. It spans a wide range of events, from ancient plagues like the Plague of Megiddo (1350 BC) to later outbreaks such as the 1592 Malta plague epidemic. The causes of these diseases vary, including conditions like bubonic plague, smallpox, influenza, and typhoid fever. The dataset highlights the significant impact of these events on populations across Europe, Asia, the Americas, and other regions. To analyze, few changes were made in later stages of the project in the dataset.

**Color Scheme and Visual Design Rationale**

In this project, we have used gradients of red for the charts to emphasize the severity of the death tolls, with darker shades indicating higher mortality rates, making the data more visually intuitive. The dark grey background was selected to ensure that the focus remains on the vibrant red color of the visualizations, enhancing contrast and readability. This color scheme provides clarity and enables viewers to quickly interpret the scale of each pandemic's impact. Additionally, the subdued background reduces distractions, fostering a more professional and polished appearance for the overall visualization.

**KPI**

Death Toll (Estimate): The Death Toll (Estimate) KPI measures the estimated number of deaths attributed to each disease outbreak, providing a quantitative reflection of the human cost of pandemics. We determined this KPI by extracting death toll estimates from historical records, ensuring accuracy by utilizing sources such as academic articles, government health agencies, and historical data sets. The estimate helps gauge the severity of each pandemic, allowing for direct comparisons across outbreaks. This metric is critical in understanding the scale of mortality and serves as a primary indicator for prioritizing public health interventions and resource allocation in response to pandemics. It enables stakeholders to assess the global impact of various diseases, facilitating a better understanding of their long-term societal consequences and aiding future preparedness efforts.

Timeline of Events (Start and End Years): The Timeline of Events KPI tracks the start and end years of each disease outbreak, offering a chronological perspective of pandemics over time. We derived this KPI by reviewing historical records and disease outbreak reports, meticulously recording the onset and conclusion of each event. By analyzing the timeline, we can identify patterns of disease emergence, fluctuations in frequency, and the duration of each pandemic. This KPI is essential for understanding the dynamics of disease spread, revealing insights into how factors like global mobility, medical advancements, and public health responses influence the lifespan of pandemics. The timeline also helps forecast future disease outbreaks, providing valuable information for the timely implementation of containment measures and medical interventions.

Disease Category Distribution: The Disease Category Distribution KPI classifies pandemics into broad disease groups, such as Plague, Pox, Viral Infection, and others, to highlight the types of diseases responsible for the greatest fatalities throughout history. This categorization was determined by reviewing disease classifications and aligning them with established medical frameworks to ensure accurate grouping. The categorization allows for a deeper understanding of the types of pathogens responsible for historical global health crises and provides insight into trends over time. By analyzing the distribution of these categories, health professionals can prioritize research and preventive measures for diseases that have had the most devastating effects historically. This KPI helps guide future public health policies by identifying which disease types pose the greatest risks and require sustained focus for prevention, surveillance, and vaccine development.

**Preparing the Dataset**

**Data Cleaning**:

Utilized the Death toll\_check Python script to clean and preprocess the dataset. Key cleaning steps include:

* For the BC column, if the value is negative, it is reversed (i.e., BC = -BC).
* For the Present column, set values to 2024 where applicable.
* Removed all text within parentheses to ensure clean data.
* Normalized spacing within text entries to maintain consistency.
* Converted whole number entries directly into integers.
* Handled range formats (e.g., "5-10" becomes "10", extracting the upper limit).
* Transformed "million" and "billion" notations into corresponding integer values (e.g., "2-3 million" becomes 3 million).
* Processed values with a '+' symbol (e.g., "10,000+" becomes 10,000).
* If an unrecognized format was encountered, the system prompted the user for manual input or allowed the entry of "Unknown" for inestimable values.
* Stored the processed values in a new column, labeled "G."

Creating Start and End Year Columns:

* Added columns for Start Year and End Year for better temporal organization of the data.

Disease and Location Cleanup:

* Subtypes of diseases were excluded to maintain focus on main categories.

Locations were split into separate columns using the Textsplit function in Excel for easier analysis

**Continents assigned using py script** ‘location’.

Remaining done manually. Old locations assigned continents acc to today’s time. For Excel file ‘country to continent’, We write a python script to enriches a pandemic dataset by adding continent information for each location in the data. First, it loads two Excel files: one with the Pandemic data and another that maps each country to its continent. The script then defines a function to look at location columns (like Location\_1, Location\_2…7) in each row, checks if each location has a corresponding continent in the mapping file and adds the continents as a comma-separated list in a new 'Continents' column. Continents namely: Africa, Antarctica, Asia, Europe, North America, South America, Oceania. Some locations were

**Final Cleanup:**

* Manually reviewed and filled in any remaining gaps for location data.

**Exporting Cleaned Data:**

* Saved the cleaned dataset as a CSV file to facilitate seamless import into Tableau for visualization.Step 2: Importing Data into Tableau

**Opening Tableau:**

* Created a new workbook in Tableau to begin the visualization process.

**Connecting to the Data:**

* Imported the cleaned CSV file into Tableau by selecting "Text File" from the connection options.

Verifying Data Types: Ensured that Tableau correctly identified and assigned appropriate data types to each column]

**Data Visualization:**

**Calculated Fields (CF)**

* Known and unknown death

KD: IF NOT ISNULL([Death toll (estimate)]) THEN [Death toll (estimate)] END

* Combined locations: To merge the location data across columns (1 to 7)

IFNULL([Location 1], '') +

IFNULL(', ' + [Location 2], '') +

IFNULL(', ' + [Location 3], '') +

IFNULL(', ' + [Location 4], '') +

IFNULL(', ' + [Location 5], '') +

IFNULL(', ' + [Location 6], '') +

IFNULL(', ' + [Location 7], '')

* D2 : to get events having greater than 2000000 deaths

IF [Death toll (estimate)] > 2000000 THEN [Event] END

* D filter: To get only the events of D2 (apply dfilter to filter than do TRUE, to apply it)

IF [Death toll (estimate)] > 2000000 THEN TRUE ELSE FALSE END

* CO: constant 0. A constant value of 0 is used to create a reference line.
* Disease category: all diseases divided into:
  + Plague: If the disease is "Bubonic plague," it is labeled as "Plague."
  + Pox: If the disease is "Smallpox," it is categorized as "Pox."
  + Influenza: Any mention of "Influenza" keeps its name as "Influenza."
  + Viral Hemorrhagic Fever: Diseases like "Yellow fever" and the viruses "Ebola" or "Marburg" fall under this category.
  + Vector-Borne Disease: This includes diseases spread by insects, such as "Malaria" and "Dengue fever."
  + Bacterial Infection: "Cholera" and "Typhus" are both classified as bacterial infections.
  + Viral Infection: Several diseases, including "Measles," "COVID," "Poliomyelitis," "HIV/AIDS," "Hepatitis," and "Zika virus," are grouped as viral infections.
  + Other: If a disease does not fit into any of the categories above, it is labeled as "Other."

**Details table**

Distinct count of continents, countries, known and unknown deaths, with filters based on years. The dataset covers 25 continents and 151 countries. Total deaths across pandemics: 54,948,136 from 430 BC to 2022, with 35 events where death tolls are unknown.

**Map Visualization:**

map visualizes historical disease outbreaks around the world, using shades of red-blue to represent death tolls, shown by side filter, where darker indicate higher death tolls (ranging from 2 to 200,000,000). When hovering over specific countries, like Canada, a tooltip appears showing detailed information about the epidemic in that area (location, event name, disease, start year and its death toll). The map uses geographic data (latitude and longitude) to plot locations, with various data fields like "Disease," "Start Year," and "Event" helping to categorize and detail the outbreaks.

**Tree map**

The treemap visualization in Tableau represents various pandemics by their estimated death tolls, with each rectangle's size and color intensity corresponding to the death toll. Larger and darker red rectangles indicate pandemics with higher death tolls, while smaller and lighter ones represent those with fewer deaths. The Bubonic Plague has the largest and darkest rectangle, showing the highest death toll at 321,990,641. Other significant pandemics include Influenza and HIV/AIDS, displayed in red shades, while COVID, Smallpox, and Salmonella enterica appear smaller and lighter. Tooltips provide additional details, such as the specific death toll for each pandemic with its name. This treemap effectively visualizes the scale of mortality across different pandemics, highlighting the disproportionate impact of certain diseases.

**Bar Chart for Death Toll:**

This bar chart compares the death tolls of various disease categories (all diseases reported divided into broad diseases category of 8), visually demonstrating the scale of impact each disease has had. The vertical axis shows the estimated death tolls, while the horizontal axis lists disease categories. The Plague has the highest death toll at 322,050,727, represented by the tallest and dark red, while Typhus has the lowest with 3,402,000 deaths, shown by the smallest, lightest red-pink. The colors, ranging from red-pink (for lower death tolls) to dark red (for higher death tolls), provide a quick visual cue for severity, with a legend on the right indicating the range from 3,402,000 to 322,050,727. This bar chart effectively highlights the significant variance in mortality rates among historical pandemics and disease outbreaks, making it easier to compare their relative impact immediately.

**Timeline Visualization:**

To make this visualization a few CF were created.

CO= constant 0. Added a constant value of 0 to create a baseline for the timeline.

D2+ to get the incidents having deaths more than 200000

Dfilter To get only the events of D2+

Dual Axis: The timeline uses a dual-axis approach with Start Year on the columns and CO (0) on the rows to create a reference line in the middle.

The Timeline Visualization process involves several key steps to effectively display historical pandemics. The Marks Card includes three marks: All, Start Year 1 (y1), and Start Year 2 (y2), with detailed tooltips providing information about the event name, death toll, and whether it meets the D2+ criteria (death toll greater than 200,000). The Dfilter (True) is applied to filter out events with death tolls below 200,000, ensuring that only relevant events remain on the timeline. Annotations are added to each event for added clarity and context. Additionally, the Axis Adjustment modifies the timeline to start from the year 165 (instead of -430 BC), making the axis easier to read and understand. The tick marks are fixed with a 200 interval, providing a clear and consistent scale throughout the visualization.

**Insights**

Disproportionate Mortality Across Pandemics: The disproportionate mortality across pandemics is evident when looking at the historical death tolls of various diseases. The Bubonic Plague, for instance, stands out as the most devastating, with an estimated death toll exceeding 320 million, demonstrating how certain pandemics can have catastrophic global consequences. The impact of HIV/AIDS and Influenza is also significant, as these viral diseases have contributed greatly to global mortality, underscoring their long-term and widespread effects on societies worldwide. On the other hand, diseases like Typhus and Salmonella, which have lower death tolls, may indicate more localized or less severe outbreaks that didn’t spread as widely or result in as high a fatality rate as the major pandemics. This variance in death tolls highlights the differing scales of impact that various diseases can have, with some causing widespread devastation while others have more contained or regional effects. It also reflects the varying nature of infectious diseases—some spreading rapidly across continents, while others remain more limited in their reach and severity.

Geographic and Temporal Spread: The geographic and temporal spread of pandemics is vividly illustrated through the map and timeline visualizations, providing insights into how disease outbreaks have evolved across different regions and time periods. The map visualization clearly shows how some pandemics, such as the Bubonic Plague, were concentrated in specific regions like Europe and Asia, reflecting the limited global reach of certain outbreaks during those times. In contrast, diseases like HIV/AIDS and COVID-19 had a much broader, more global impact, affecting nearly every corner of the world and highlighting the interconnectedness of modern societies. The timeline visualization further enhances this understanding by showing how pandemics often occurred in clusters, corresponding with particular historical events or time periods. For example, the plagues during the Middle Ages, the 1918 influenza pandemic, and the rise of HIV/AIDS in the late 20th century all align with specific moments in history, offering a perspective on how societal conditions, advancements, or challenges influenced the spread and severity of diseases. This dual approach—spatial and temporal—helps to contextualize the historical impact of pandemics, offering a deeper understanding of both their regional spread and their timing in history.

Diseases-by-category impact: The category-specific impact of different disease types reveals important patterns in how pandemics affect global mortality. Viral infections, such as Measles, COVID-19, and HIV/AIDS, collectively contribute to a significantly larger share of global deaths compared to bacterial diseases. This suggests that viral pandemics tend to have a wider and more enduring impact, often spreading across continents and causing long-term effects on public health. These viruses, through rapid transmission and evolving mutations, have a capacity to affect a larger portion of the global population over extended periods, leading to higher mortality rates. In contrast, bacterial diseases like Typhus and Cholera typically have more localized impacts, although they can still cause substantial fatalities. Vector-borne diseases, including Malaria and Dengue fever, also stand out for their significant mortality rates, particularly in tropical regions. These diseases, transmitted by insects such as mosquitoes, highlight the critical role of environmental factors—such as climate, water sources, and population density—in facilitating their spread. Warmer climates and stagnant water provide ideal breeding grounds for mosquitoes, making tropical and subtropical areas more susceptible to these diseases. This pattern emphasizes the importance of environmental and ecological factors in determining the spread and severity of certain diseases, which is crucial for targeting prevention and control measures.