Big data for Business

Big Data Health Analytics Project

Global Cancer, U.S. Health, and Clinical Diagnostics

Md Nahid Rahman

Phase I,II & III (Nahid)

Introduction of Phase I, II, III

In today's data-driven world, understanding the root causes and patterns of cancer is more critical than ever. This project integrates three phases of analysis, using **SQL** and **Python** to uncover mortality patterns and prepare datasets for advanced modeling, followed by visualization and insight generation through **Tableau** and **SAS Viya Visual Analytics**, leveraging two major datasets that provide both a global perspective and clinical-level insights. The first is the **Global Cancer Dataset (2019–2023)** by Ankush Panday, which contains over **900K+ records** from the **50 most populated countries**. It focuses on cancer **incidence**, **mortality**, and key **risk factors** such as **tobacco use**, **obesity**, **UV radiation**, and **healthcare access**. The second dataset is the **CDC BRFSS 2022 Heart Disease Dataset**, a comprehensive public health survey of over **400,000 U.S. adults**, covering data on **mental health**, **physical health**, **BMI**, **sleep patterns**, and **heart disease risks**. A third clinical dataset, the **Breast Cancer Wisconsin (Diagnostic) Dataset** by M. Yasser H, includes detailed tumor-level data collected between 1989 and 1995, with features like **area**, **texture**, **smoothness**, and **compactness**, helping identify **benign** versus **malignant tumors**.

In **Phase 1**, the focus is on exploring **global cancer trends** using data visualizations that make complex patterns easier to understand. A world map using color gradients highlights countries with the **highest cancer incidence**, signaling regions in urgent need of **healthcare intervention**. A pie chart on cancer mortality shows which types—such as **lung**, **colorectal**, and **breast cancer**—contribute the most to global deaths. A comparative bar chart reveals significant **disparities in mortality** between countries with advanced healthcare systems and those with limited access to treatment. The treemap on cancer risk factors illustrates how **tobacco use**, **pollution**, and **unhealthy lifestyles** drive many of these outcomes. Additionally, a scatter plot examining **incidence vs. mortality** by age and cancer type emphasizes that **early detection** and **treatment**

availability are key to improving survival. The final chart in this phase compares **obesity** and **physical activity levels** across countries, showing that regions with **high obesity** often report **low physical activity**, reinforcing the link between **lifestyle choices** and cancer risk.

Phase 2 shifts the analysis toward **U.S. population health**, using the **CDC BRFSS 2022 dataset** alongside the global cancer data. This phase highlights the inverse relationship between **mental and physical health** across different age groups, with mental health issues peaking in the **18–24 age group**, while physical health problems increase with age. A heatmap shows how self-reported health status differs across **racial groups**, with **White adults** generally reporting better health than **Black**, **Hispanic**, and **American Indian/Alaska Native** populations—suggesting gaps in healthcare access. Another visual links **average sleep time** to **heart disease risk**, reinforcing how **insufficient sleep** can affect cardiovascular health. A line chart shows that **heart disease cases sharply rise** after the age of 50, peaking in the 65–74 age range. A dual-axis chart reveals that **White adults** report the highest combination of **heart disease and overweight status**, while **Asian adults** report the lowest. The phase concludes by revisiting global cancer data, showing that countries like **Argentina**, **Afghanistan**, **Russia**, **Ukraine**, and **Venezuela** have some of the **highest cancer mortality rates**. A focused comparison between **Canada**, **USA**, and **Mexico** provides further insights into **regional healthcare performance** in North America.

In **Phase 3**, the project moves into a **clinical perspective** by analyzing the **Breast Cancer Wisconsin Diagnostic dataset**, enriched by visual tools from both **Tableau** and **SAS Viya**. The first visual examines how **tobacco use** relates to cancer mortality, with a bubble chart showing a strong positive trend ($R^2 = 0.999$), confirming tobacco as a dominant risk factor. A cluster analysis then explores how **family history** and **obesity** combine to increase risk in specific country groups. Another cluster analysis connects **UV exposure** to rising cancer rates, particularly in older age groups. A comparison between **Canada and the U.S.** shows that despite similar population densities, **Canada reports slightly higher incidence and mortality**, prompting questions about **healthcare policies or environmental conditions**. A comprehensive dashboard summarizes the

global factors—**tobacco**, **UV**, **obesity**, and **family history**—into one visual narrative. On the clinical side, charts compare tumor **smoothness**, **symmetry**, **area**, and **texture**, showing that **malignant tumors** tend to have more irregular features. A bubble chart plotting **radius**, **area**, and **perimeter** clearly distinguishes **malignant** from **benign** clusters. A final multivariate dashboard helps identify **high-risk tumor characteristics** and patterns in growth, supporting **early diagnosis** and **clinical decision-making**. SAS Viya visualizations of **tumor size distribution** and **age-specific cancer mortality** add another dimension, showing how demographics influence both **risk** and **outcomes**.

Altogether, these three phases create a complete and insightful picture of cancer—from **global distribution and population health factors** to **individual tumor analysis**. By combining **international statistics**, **U.S. health trends**, and **clinical features**, this project provides a powerful example of how **big data** can support **evidence-based healthcare policy**, **targeted prevention**, and **early detection strategies**. The ultimate goal is not just to visualize numbers, but to turn those insights into **actionable solutions** that contribute toward a **healthier, cancer-aware world**.

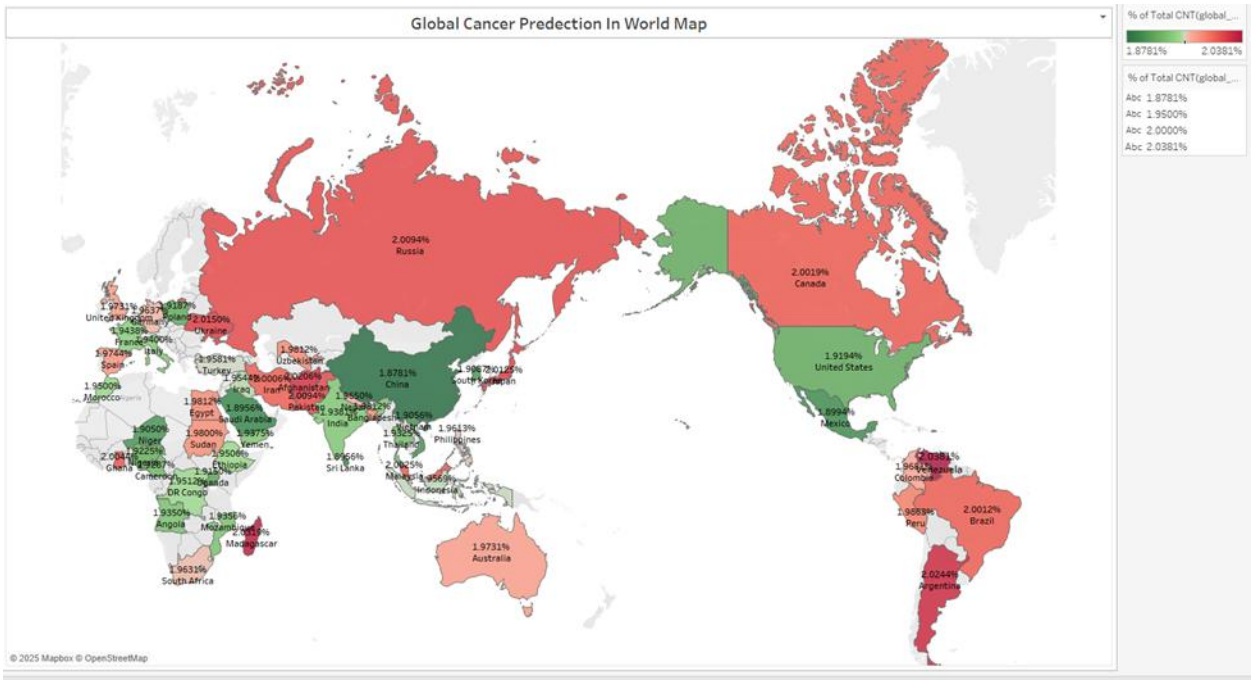

Figure 1.1: Global Cancer Prediction World Map

Explanation:

This Global Cancer Prediction World Map visualizes cancer incidence rates across different countries using a color gradient. Countries in red indicate higher cancer incidence, while green represents lower rates. The percentage values displayed on each country show their contribution to global cancer cases, allowing for comparative analysis. Users can interact with the map by applying filters for specific cancer types, helping to analyze regional variations. The legend on the right provides a reference for interpreting the percentage scale of cancer incidence. Countries shown in gray/white may have missing or unreported data, which could affect accuracy. This visualization helps in identifying high-risk regions, where healthcare policies and cancer prevention strategies need more focus. Additional analysis could explore the relationship between cancer incidence and risk factors like pollution, smoking, and healthcare infrastructure. These insights are crucial for policymakers and researchers in designing targeted interventions to improve global health outcomes.

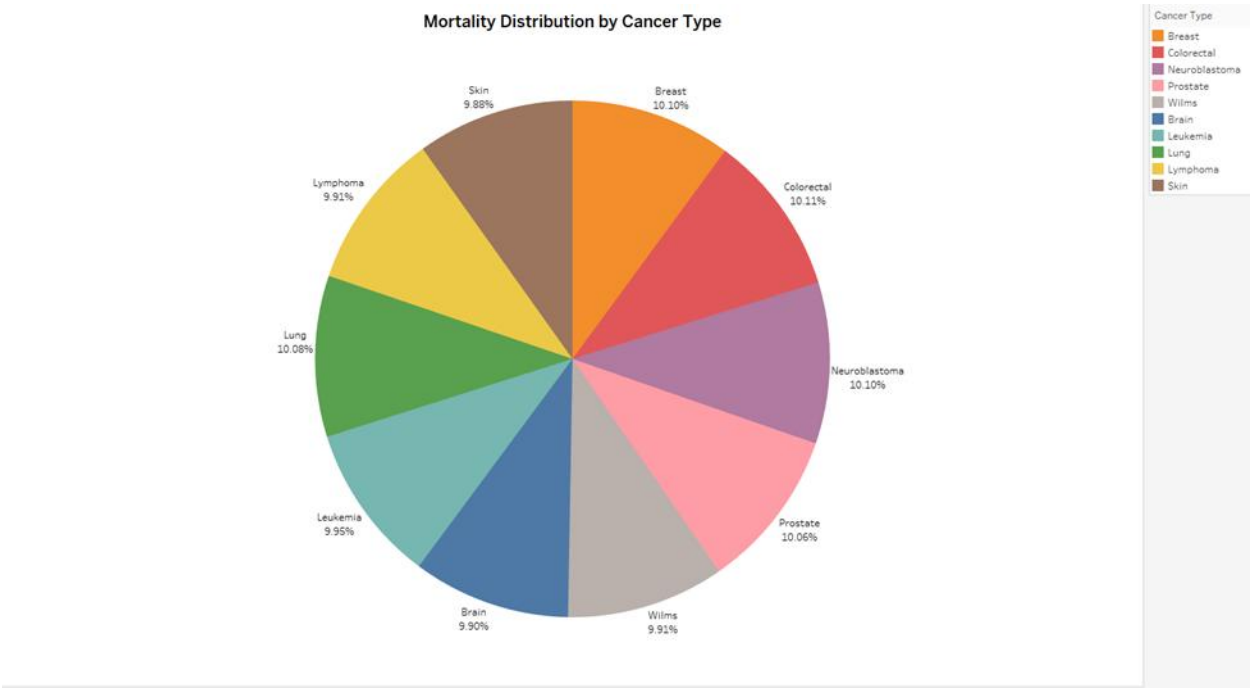

Figure 1.2: Mortality Distribution by Cancer Type

Explanation:

This pie chart illustrates the mortality distribution among various cancer types based on a global cancer prediction dataset. Each slice represents a specific cancer type, with its size proportional to the percentage of total cancer-related deaths. The percentages have been formatted to two decimal places for clarity. Lung, Colorectal, and Breast cancers show the highest mortality rates, while others, such as Lymphoma and Wilms, have comparatively lower percentages. This visualization helps in understanding which cancers contribute the most to overall mortality, providing valuable insights for healthcare research, policy-making, and targeted prevention strategies. By analyzing mortality distribution, medical professionals and researchers can prioritize resources and interventions to combat the most life-threatening cancers more effectively.

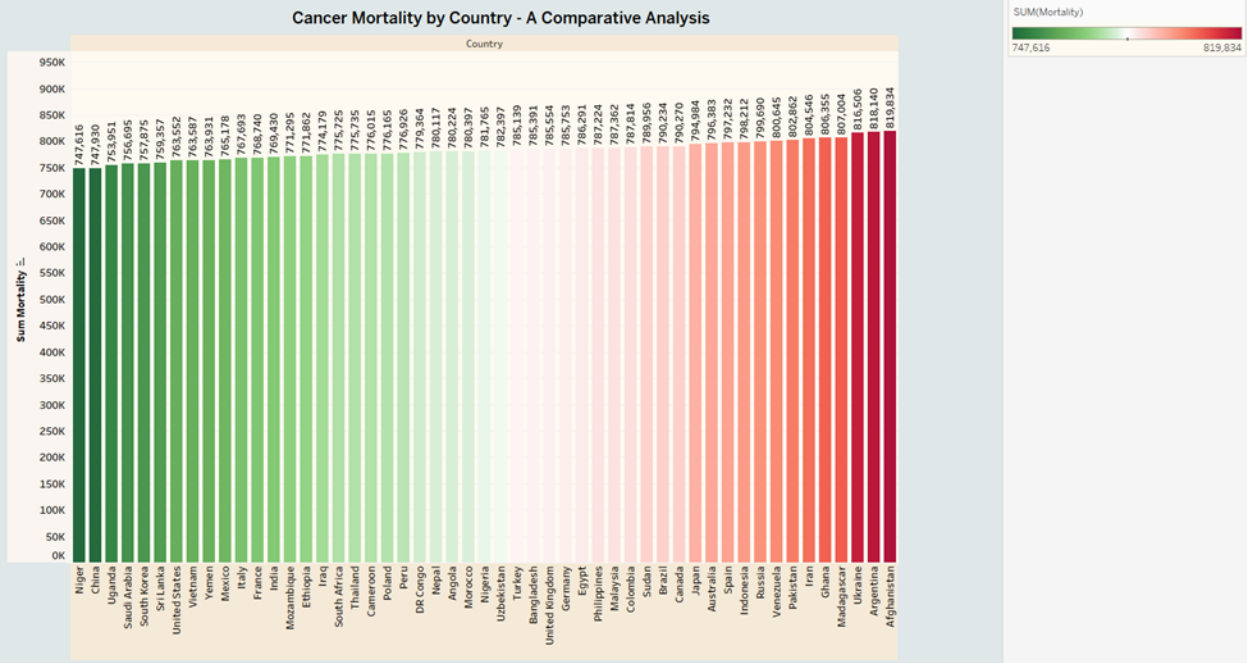

Figure 1.3: Cancer Mortality by Country - A Comparative Analysis

Explanation:

This bar chart visually represents the cancer mortality rates across different countries. Each bar corresponds to a country, with its height indicating the total mortality count due to cancer. The color gradient from green to red helps highlight variations in mortality, where green represents lower mortality rates and red indicates higher mortality rates. Countries on the left, such as Niger and China, have relatively lower cancer-related deaths, whereas countries on the right, such as Afghanistan and Argentina, exhibit the highest mortality rates. This variation can be influenced by multiple factors, including healthcare infrastructure, lifestyle choices, environmental conditions, and access to medical treatment. The chart allows for easy comparison between countries, providing insights into potential risk factors and areas that require improved healthcare interventions. By analyzing such data, policymakers and researchers can work toward better cancer prevention and treatment strategies worldwide.

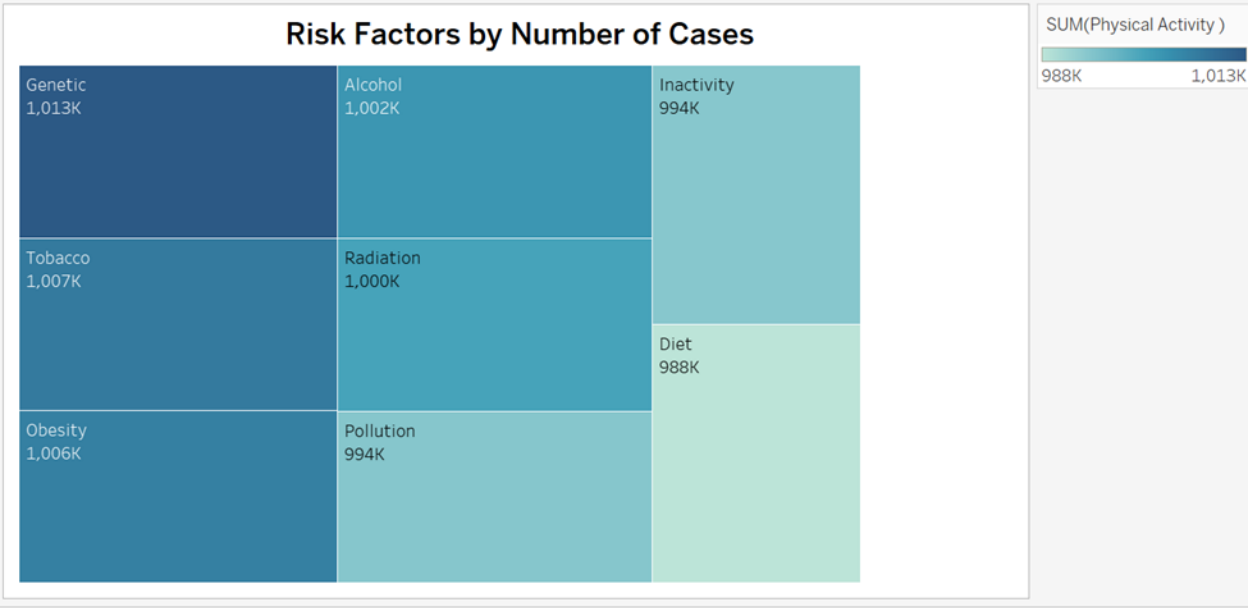

Figure 1.4: Cancer Risk Factors by Number of Cases

Explanation:

“Risk Factors by Number of Cases” presents a treemap visualization that illustrates the contribution of various risk factors to cancer cases. Each block represents a specific risk factor, with its size proportional to the number of cases attributed to that factor. Genetic factors have the highest number of associated cases, followed closely by tobacco as another major contributor. Obesity, alcohol, and radiation also have significant impacts, while inactivity and pollution are among the lower contributors, but still play a notable role. Diet-related factors contribute the least among the displayed risk factors. The color gradient represents variations in case numbers, with darker shades indicating higher values. This visualization allows for quick comparison between risk factors, making it easier to identify major contributors and understand which factors need more focus in cancer prevention strategies.

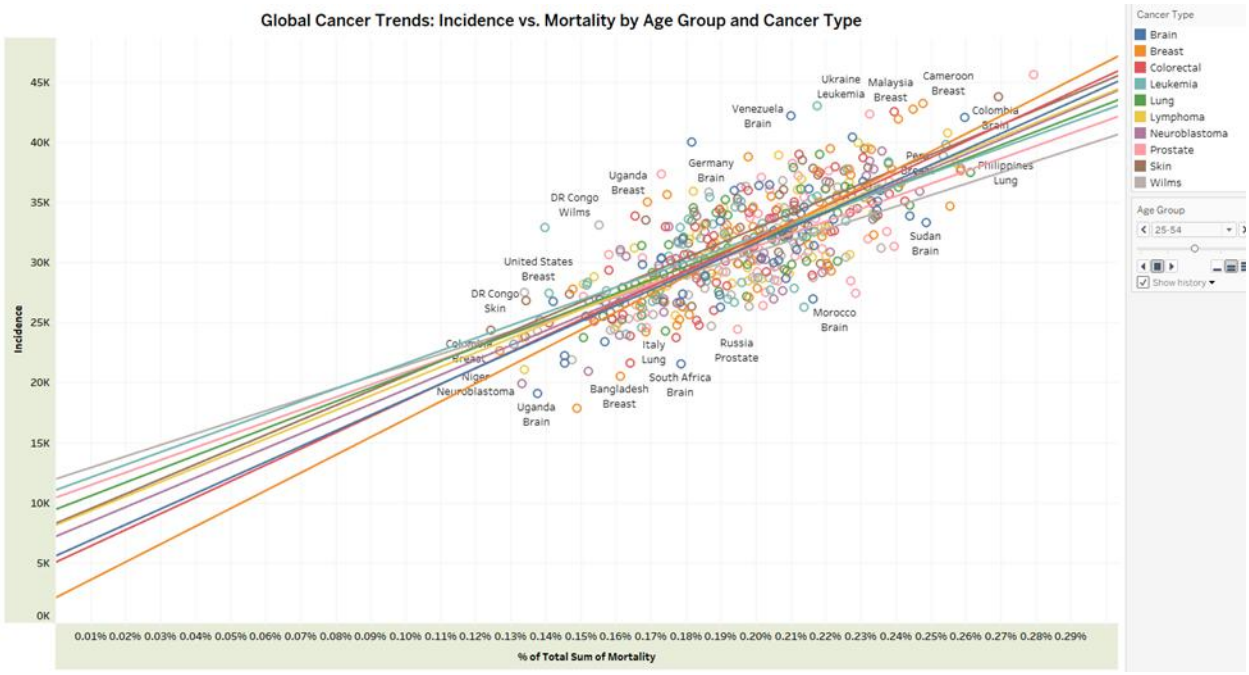

Figure 1.5: Global Cancer Trends: Incidence vs. Mortality by Age Group and Cancer Type.

Explanation:

This visualization illustrates the relationship between cancer incidence and mortality across different countries. The X-axis represents the percentage of total mortality, indicating how deadly cancer is in various regions, while the Y-axis shows the total number of cancer cases reported. Each point on the chart represents a country, with the placement indicating its incidence and mortality rates. Trend lines help show patterns in the data, revealing a correlation between higher incidence and higher mortality. The chart also allows filtering by age group, providing insights into how cancer affects different demographics. Labels for key countries highlight significant trends, making it easier to compare their positions. The overall trend suggests that countries with more cancer cases tend to have higher mortality rates. This analysis can help identify regions that may require better healthcare interventions and resources.

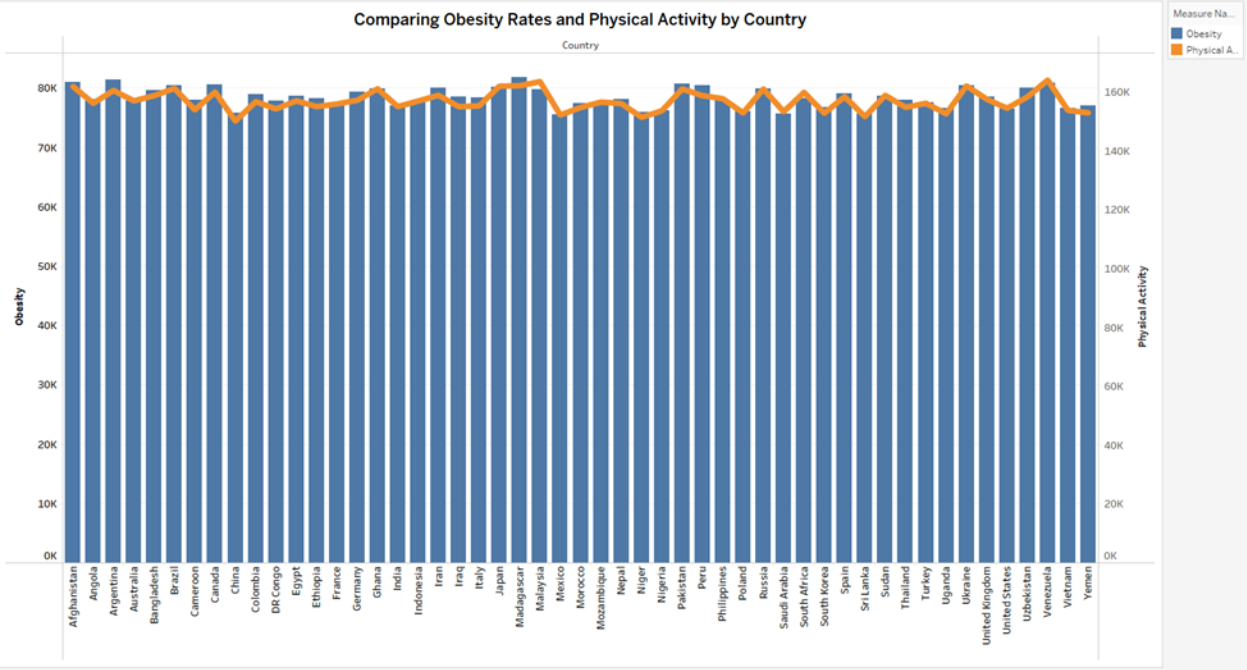

Figure 1.6: Obesity vs. Physical Activity Levels Across Countries

Explanation:

This graph compares obesity rates and physical activity levels across different countries. The blue bars represent obesity percentages, showing how obesity varies by country. The orange line represents physical activity levels, allowing us to see trends. Countries with higher obesity rates generally show lower physical activity levels. Some countries may have high obesity but moderate activity, indicating other health factors. The dual-axis chart helps visualize the inverse relationship between these two metrics. This analysis can help in public health research to address obesity-related risks. Governments and policymakers can use this data to promote health initiatives. Further analysis could include diet, urbanization, or healthcare access for deeper insights. Understanding these trends is key to reducing obesity rates and promoting a healthier society.

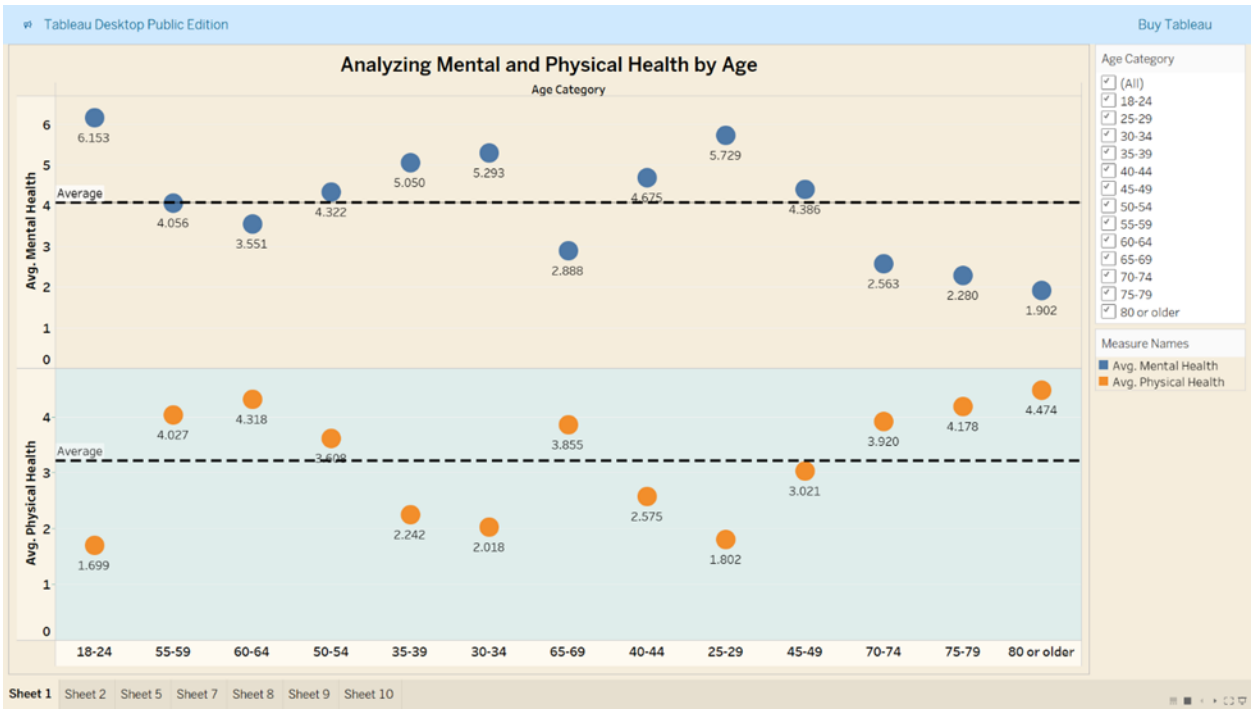

Figure 1.7 Analyzing Mental and Physical Health by Age (2022 CDC BRFSS Heart Disease Data United States)

Explanation:

This graph visualizes average mental and physical health indicators across different age groups from the 2022 CDC BRFSS Heart Disease Data (United States). The Y-axis represents the average number of days per month that respondents reported experiencing poor mental or physical health. The upper light beige section represents average mental health issues, shown using blue bubbles, indicating the number of days per month respondents experienced mental health problems. The lower light blue section represents average physical health issues, shown using orange bubbles, indicating the number of days per month respondents experienced physical health problems. The black dashed lines in each section represent the overall average across all age groups, serving as a reference point for comparison. Mental health issues are more frequent in younger age groups, with the highest average in the 18–24 age group, which gradually declines as age increases. Physical health issues are more prevalent in older age groups, peaking in the 60–64, 70–74, and 80+ categories, aligning with expectations of declining physical

health with age. The lowest mental health issues are reported in the 75–79 age group, possibly due to better coping mechanisms or underreporting by older adults. The lowest physical health issues are reported in the 18–24 age group, reinforcing the trend of better physical well-being in younger populations. This visualization effectively demonstrates an inverse relationship between mental and physical health across age groups, emphasizing the need for age-specific health interventions.

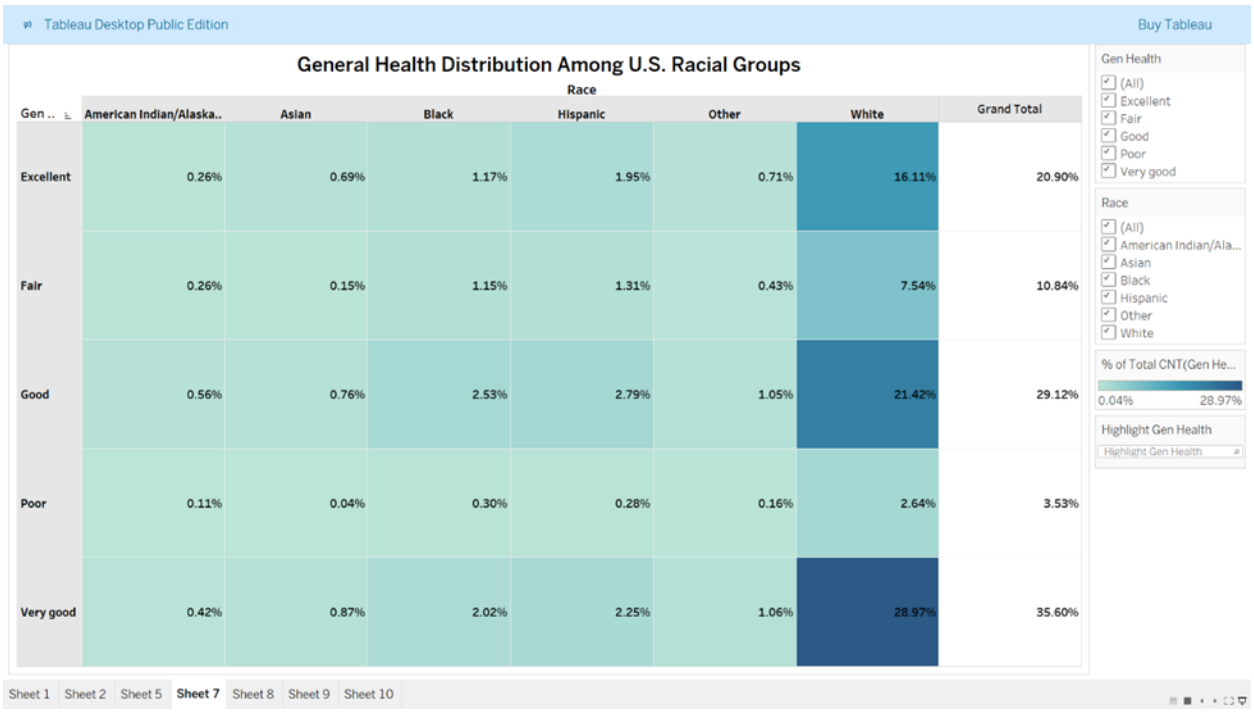

Figure 1.8 General Health Distribution Among U.S. Racial Groups: Insights from the 2022 CDC BRFSS Heart Disease Data (United States)

Explanation:

This heatmap is based on the 2022 CDC BRFSS dataset, a U.S. health survey covering 400,000+ adults to analyze general health distribution across racial groups. The X-axis represents race, while the Y-axis represents general health categories (Excellent, Very Good, Good, Fair, Poor), with percentages displayed in each cell. The heatmap uses a shading scale from light to dark blue, where darker shades indicate higher percentages, making trends more visible. The majority of respondents rated their health as "Very Good" or "Good," meaning over 64% of surveyed adults consider themselves in good health. The White population reports the highest percentage in "Very Good" and "Good" health, suggesting overall better self-reported health. Black, Hispanic, and American Indian/Alaska Native groups show a more even distribution across all health categories, indicating possible health disparities. The highest percentage of "Fair" and "Poor" health also comes from the White population, likely due to their larger sample size. Instead of raw counts, percentages were used to ensure fair comparison across racial groups with different population sizes. The Grand Total column helps compare each racial group's

general health distribution against the entire U.S. adult population in the dataset. This heatmap confirms that self-reported general health varies by race, with White respondents reporting better health on average, while Black, Hispanic, and other groups show a more balanced spread across all categories, possibly reflecting broader socioeconomic and healthcare access disparities.

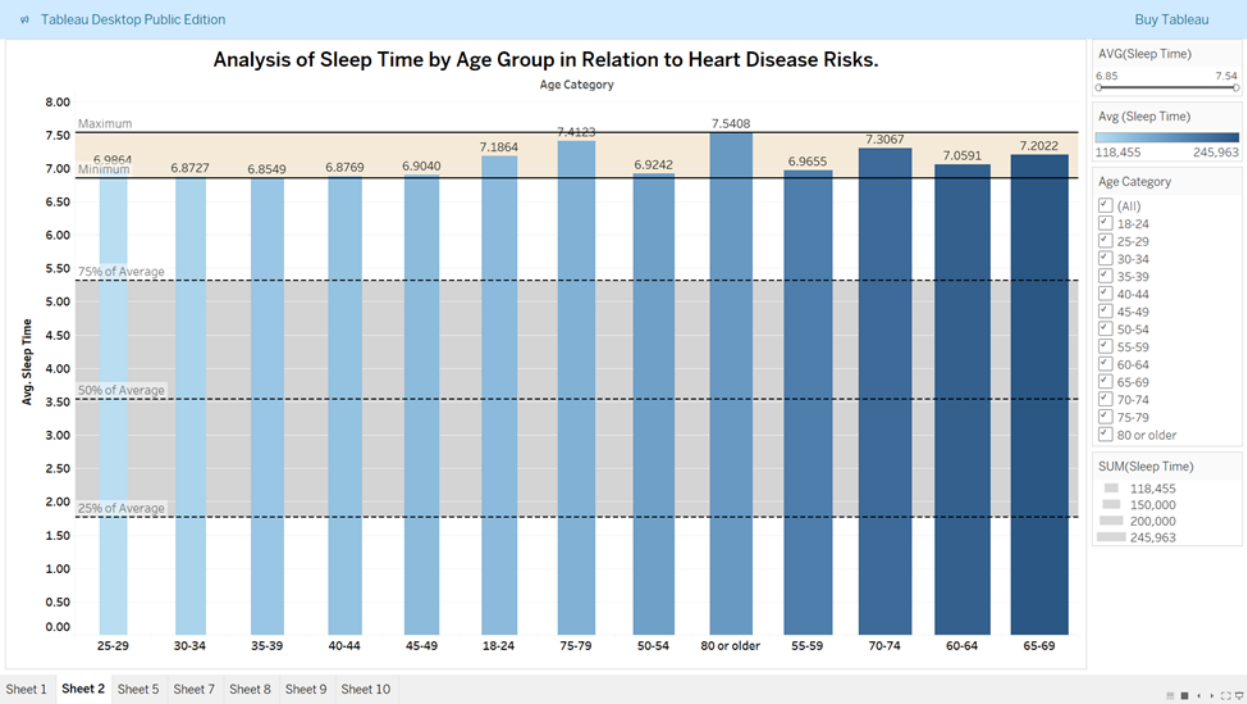

Figure 1.9 Analysis of Sleep Time by Age Group in Relation to Heart Disease Risks (2022 CDC BRFSS Heart Disease Data, United States)

Explanation:

This visualization represents average sleep time (AVG Sleep Time) across different age categories using data from the 2022 CDC Behavioral Risk Factor Surveillance System (BRFSS) survey. The dataset includes responses from over 400,000 U.S. adults, tracking key health indicators, including sleep habits, which play a crucial role in heart disease risk factors. The bar chart shows the average sleep duration for each age group, ranging from 6.85 to 7.54 hours, which falls within the CDC's recommended 7–9 hours of sleep for adults. Minimum and Maximum Reference Lines (shaded in beige) indicate the range of sleep times, helping identify any potential outliers or deviations from normal patterns. Percentile Bands (25%, 50%, 75%) (shown in gray) provide additional insight into the distribution of sleep durations, showing where most of the population's sleep falls. Sleep is a key determinant of heart health, and inadequate sleep has been linked to high blood pressure, obesity, diabetes, and other heart disease risks, which are part of this dataset. The dataset focuses on key risk factors for heart disease, such as high blood pressure, obesity, smoking, and lack of physical activity, making sleep an important variable in

cardiovascular health research. The inclusion of sleep data in this survey highlights its importance and the need for early interventions to promote healthy sleep habits. While this graph focuses on sleep trends by age, further analysis could correlate sleep time with other risk factors such as obesity, smoking, alcohol consumption, and diabetes. Overall, this visualization helps understand how sleep duration varies across age groups and how it could be used to identify high-risk populations for heart disease prevention strategies.

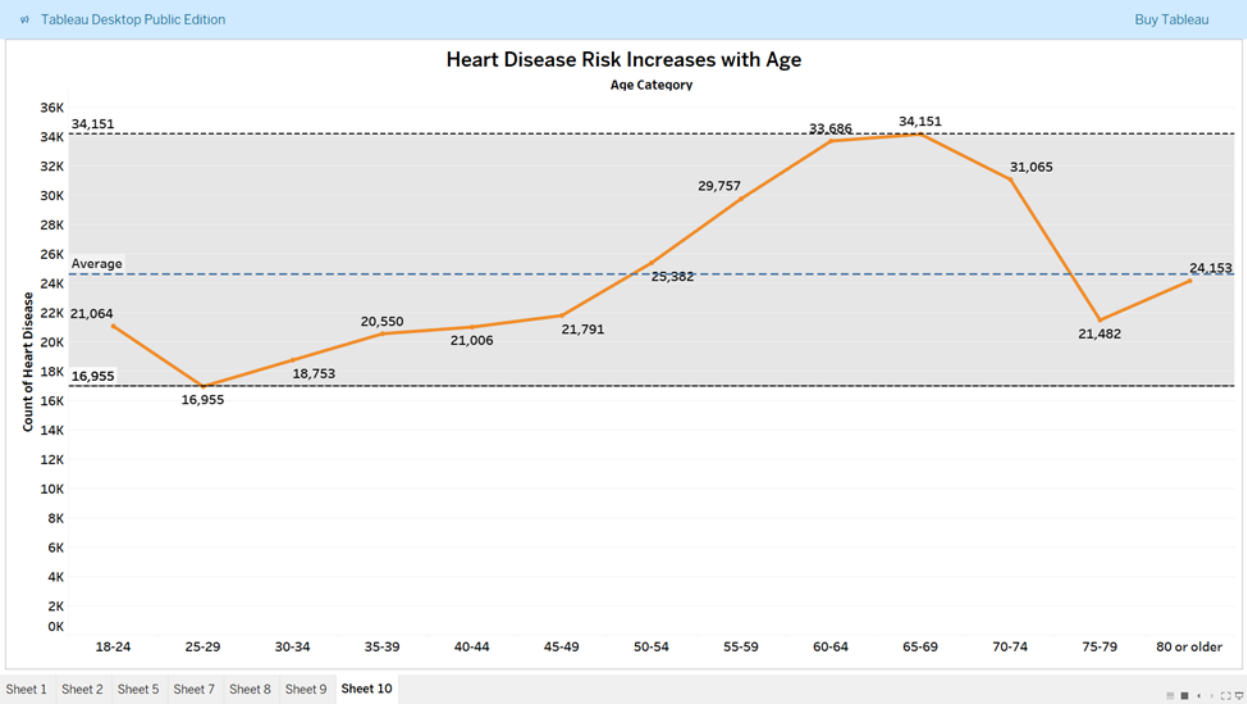

Figure 1.10 Heart Disease Risk Increases with Age: Analysis of 2022 CDC BRFSS Heart Disease Data (United States)

Explanation:

This analysis is based on the 2022 CDC Heart Disease Dataset, collected through the Behavioral Risk Factor Surveillance System (BRFSS). The dataset includes responses from over 400,000 U.S. adults, focusing on heart disease cases across different age groups. Heart disease is a leading cause of death in the United States, and understanding how age affects heart disease risk helps in creating better prevention strategies and improving public health awareness. The line chart shows the count of heart disease cases by age category. The trend starts low in younger adults, rises with age, peaks in the 70–74 age group, and then declines slightly in older populations. The lowest heart disease count appears in the 25–29 age group, which is expected as younger individuals typically have fewer risk factors. Heart disease cases gradually rise from age 30 onward, indicating that lifestyle habits, diet, and physical activity become critical factors in middle age. The highest case count occurs in the 70–74 age group, reflecting natural aging, weakened heart function, and accumulated risk factors. After age 75, case counts begin to decline, possibly due to mortality bias or medical underreporting in older populations. The dashed

blue reference line shows the average case count (~24,000), with younger groups falling below average and older adults, especially from age 50+, exceeding it. Since heart disease risk increases with age, focusing on healthy habits early in life—such as a balanced diet, regular exercise, and routine check-ups—can reduce future risk. This analysis confirms that heart disease risk peaks in the 70s, but preventive actions taken earlier can greatly reduce long-term impact.

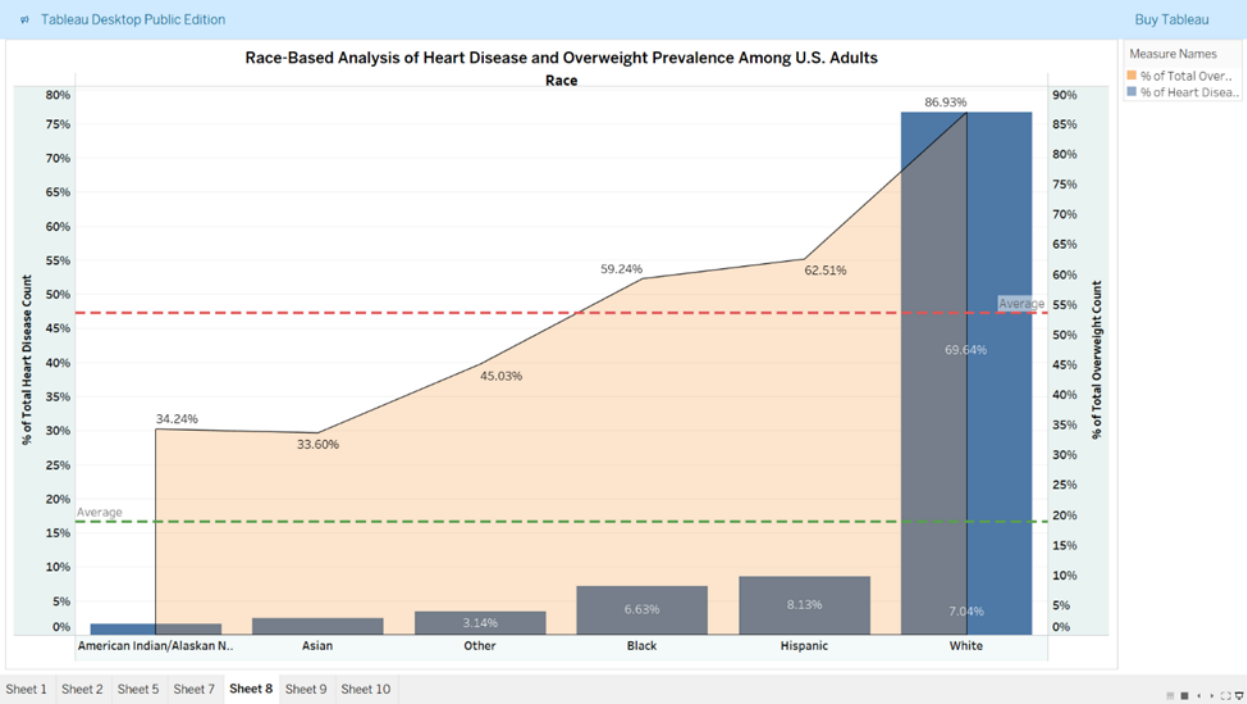

Figure 1.11 Race-Based Analysis of Heart Disease and Overweight Prevalence Among U.S. Adults: Insights from the 2022 CDC BRFSS Heart Disease Data (United States)

Explanation

This visualization presents a race-based analysis of heart disease and overweight prevalence among U.S. adults, using data from the 2022 CDC BRFSS Heart Disease Data. The blue bars represent the percentage of total heart disease cases for each racial group, while the orange area represents the percentage of overweight individuals within those groups. White adults have the highest heart disease prevalence (69.64%) and overweight percentage (86.93%), suggesting a strong correlation between excess weight and cardiovascular risks. Hispanic and American Indian/Alaska Native groups also show high overweight percentages, which may contribute to increased heart disease risk over time. Black respondents report a lower heart disease rate despite moderate overweight levels, indicating that other factors like genetics, healthcare access, or lifestyle may influence cardiovascular health. Asian respondents have the lowest overweight prevalence and heart disease cases, reinforcing the idea that lower BMI may serve as a protective factor. Two reference lines (dashed) are included: a red line marking the

average heart disease rate and a green line marking the average overweight rate, helping identify groups above or below national trends. The clear labeling, color contrast, and dual Y-axes ensure an effective comparison of how excess weight influences heart disease risk across racial groups. These findings suggest that public health initiatives should emphasize obesity prevention, particularly among White, Hispanic, and American Indian/Alaska Native populations, as they exceed national averages in both metrics.

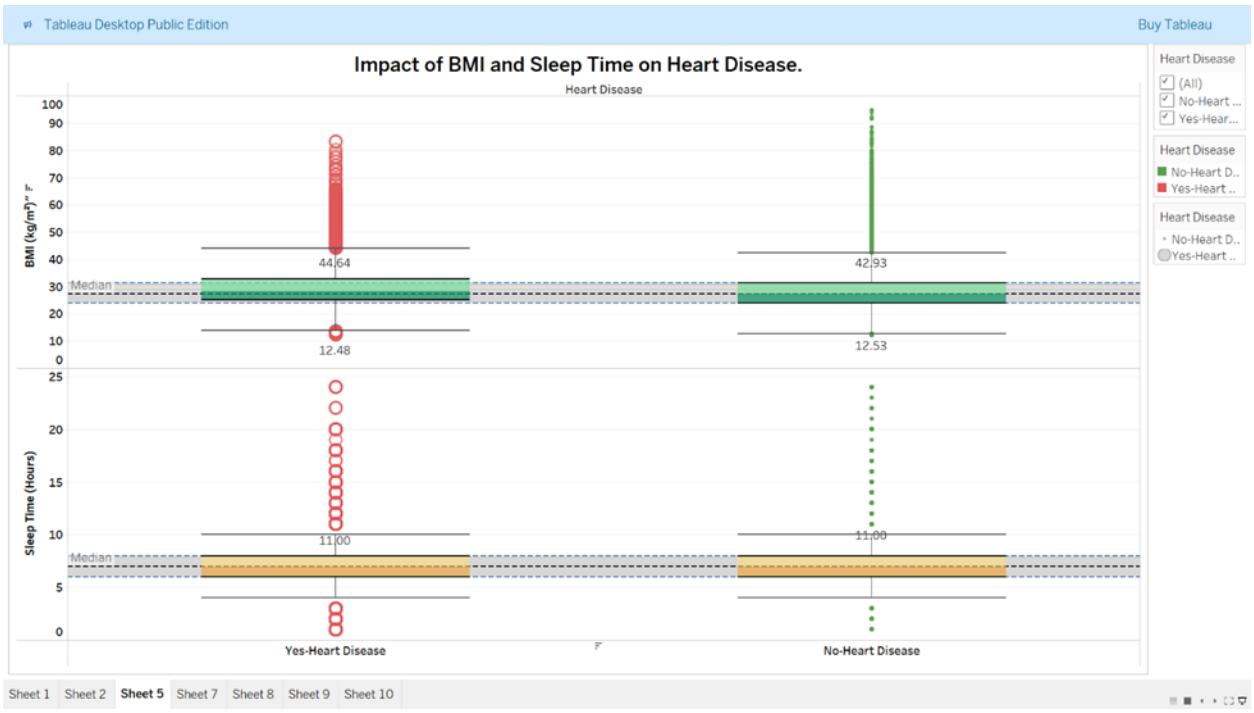

Figure 1.12 Impact of BMI (Body Max Index) and Sleep Time on Heart Disease: Analysis Aof 2022 CDC BRFSS Heart Disease Data (United States)

Explanation:

This visualization is based on the 2022 CDC BRFSS dataset, which includes health survey data from over 400,000 U.S. adults, focusing on key heart disease indicators such as BMI (Body Mass Index), Sleep Time, Smoking, and Physical Activity. The box plot explores the relationship between BMI, Sleep Time, and Heart Disease, categorizing individuals into two groups: those with heart disease ("Yes") and those without ("No"). The X-axis represents Heart Disease status, while the Y-axis shows BMI (kg/m²) and Sleep Time (Hours), allowing for a clear side-by-side comparison of health patterns. Each box plot displays the median (solid line), interquartile range (IQR), whiskers, and outliers, providing insight into the distribution of values. The median BMI for heart disease patients is 44.95, which is higher than 43.10 for those without the condition, and heart disease patients show more outliers, indicating higher obesity prevalence. Both groups have a similar median Sleep Time (11 hours), but heart disease patients show more extreme outliers, possibly pointing to poor sleep patterns. The BMI whiskers for heart disease patients are wider, suggesting greater variability. Red represents "Yes–Heart Disease"

and Green represents "No–Heart Disease", with dashed median lines aiding in visual clarity. The dataset was originally composed of 300+ variables and was reduced to 40 key indicators, with missing values cleaned in a version called "heart_2020_cleaned.csv". Overall, the box plot confirms that higher BMI is associated with heart disease, while sleep time shows less variation, though extreme sleep durations in heart disease patients warrant further investigation.

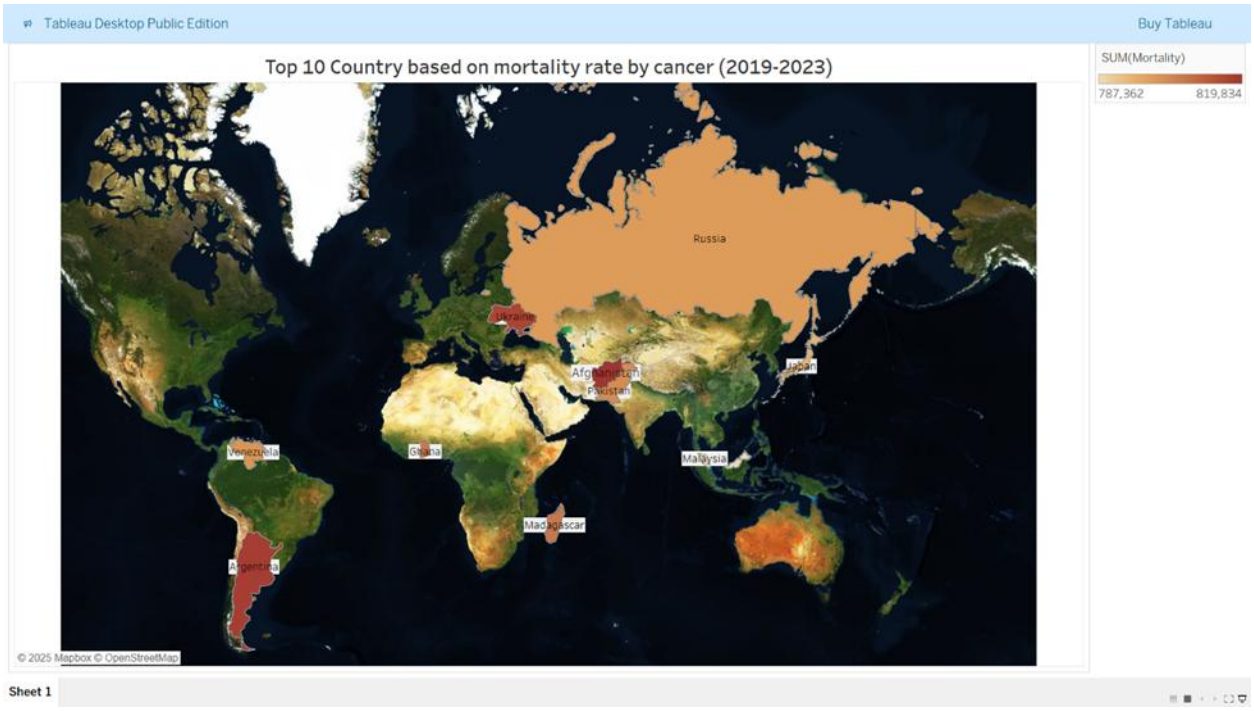

Figure : 1.13 the top 10 countries with the highest cancer mortality rates from 2019 to 2023 (Globally). Ankush 2023; Sophia et al. 2021; Nguyen 2022.

Explanation:

This visualization presents the top 10 countries with the highest cancer mortality rates from 2019 to 2023, using a satellite map view. The dataset includes 160,000 records from the 50 most populated countries, covering cancer incidence, risk factors, and healthcare metrics. The satellite map uses color intensity to show total cancer deaths, with darker shades representing higher mortality rates. Argentina and Afghanistan appear in the darkest colors, indicating they have the highest cancer mortality in this analysis. Russia and Ukraine also show significant mortality, reflecting serious health challenges related to cancer deaths. Venezuela is highlighted as another high-mortality country, possibly due to economic and healthcare system struggles. Countries like Madagascar, Malaysia, Ghana, and Japan appear in lighter shades but still rank in the top 10 for mortality rates. The satellite background helps clearly mark the locations of each country, making it easy to visualize geographic patterns of cancer deaths. Each country's dot represents a hotspot where healthcare systems may face challenges in cancer prevention and

treatment. Overall, this output helps researchers and policymakers identify critical regions needing attention and plan better global cancer control strategies.

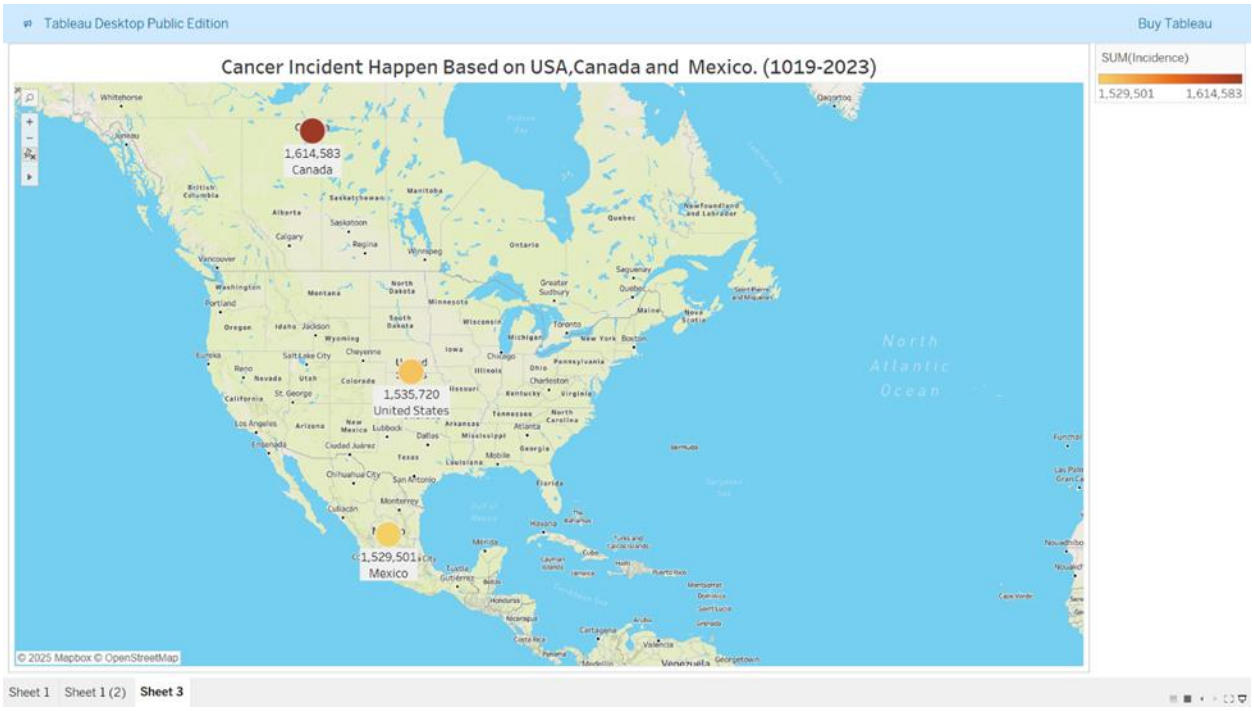

Figure 1.14: Cancer incident cases recorded between 2019 and 2023 in the USA, Canada, and Mexico. Ankush 2023; Sophia et al. 2021; Nguyen 2022.

Explanation:

This street view map shows the cancer incident cases recorded between 2019 and 2023 in the USA, Canada, and Mexico. The data comes from a large dataset containing 160,000 records of cancer incidence, healthcare metrics, and risk factors from the 50 most populated countries. Each country is represented by a colored circle, with its size and color intensity based on the number of reported cancer cases. Canada reports the highest cancer incidence with 1,614,583 cases, shown with the darkest color and largest circle. The United States follows with 1,535,720 cases, and Mexico reports 1,529,501 cases, both displayed with lighter shades. The street view background helps clearly locate the countries and visually compare the cancer burden across North America. This output allows viewers to understand the regional distribution of cancer cases and highlights areas needing healthcare attention.

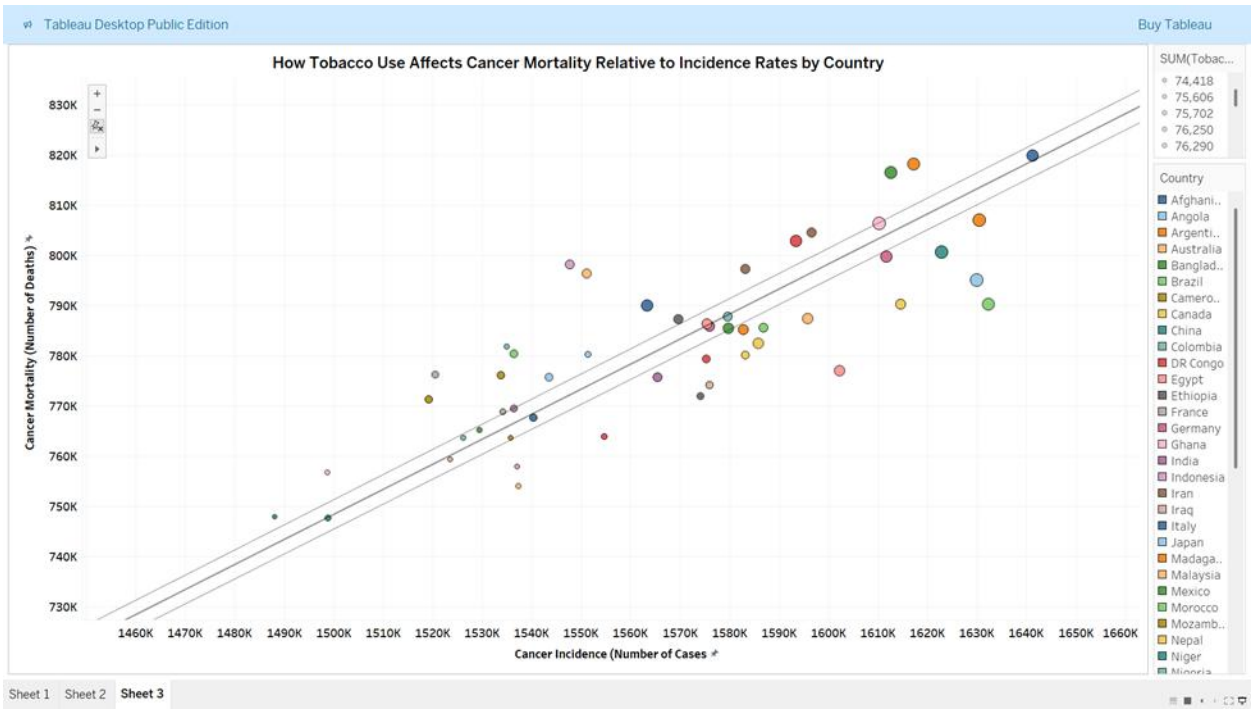

Figure 1.15: Cancer Mortality vs. Incidence with Tobacco Use as a Risk Factor (Top 50 Populated Countries 2019–2023)

Explanation:

This graph analyzes the relationship between cancer incidence (number of cases) and cancer mortality (number of deaths) across 51 countries. The dataset comes from Kaggle's Cancer Dataset (Top 50 Populated Countries), which includes health metrics and risk factors like tobacco use. The X-axis shows the total number of cancer cases per country, and the Y-axis shows the total number of cancer deaths. Each circle (bubble) represents a country, and the size of the bubble shows the percentage of tobacco use in that country. A trend line is added to show the overall pattern — countries with more cases also tend to have more deaths. The R-squared value is 0.999, meaning the linear model explains nearly all variation in mortality based on incidence. The p-value is < 0.0001, which means the relationship is statistically very significant. The slope of the line (coefficient 0.4989) means that for every 1 case increase, about 0.5 deaths are expected. Countries with larger bubbles above the line may have higher mortality due to higher tobacco use.

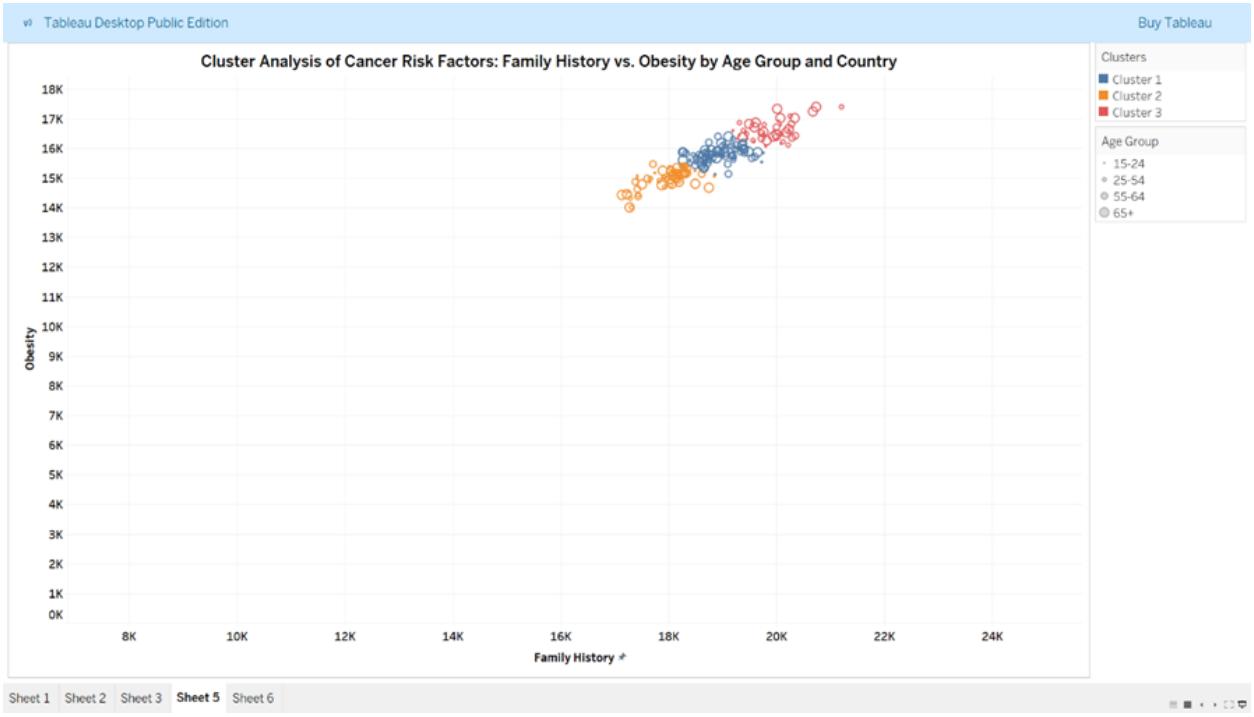

Figure 1.16: Cluster Analysis of Cancer Risk Factors: Family History vs. Obesity by Age Group and Country ("Cancer Dataset (Top 50 Populated Countries 2019–2023)")

Explanation:

This graph shows a cluster analysis of countries based on two key cancer risk factors: Family History (X-axis) and Obesity (Y-axis). The data comes from the Kaggle dataset titled "Cancer Dataset (Top 50 Populated Countries)", which includes global cancer indicators and risk factors. The graph uses three clusters (red, blue, and orange) to group countries with similar values for family history and obesity. Each circle represents a country, and the size of the bubble reflects its Age Group (15–24, 25–54, 55–64, 65+). The color shows cluster membership: Cluster 1 (blue), Cluster 2 (orange), and Cluster 3 (red), each having different risk factor patterns. The filter is set on Age Group to focus only on older populations, which are more prone to cancer-related risks. Most data points are concentrated between 17,000–21,000 Family History and 14,000–17,000 Obesity, indicating this range is common across countries. Cluster 3 (red) represents countries

with both high obesity and high family history, possibly indicating higher cancer vulnerability. The analysis helps identify which age groups and countries share similar risk patterns, useful for targeted health strategies.

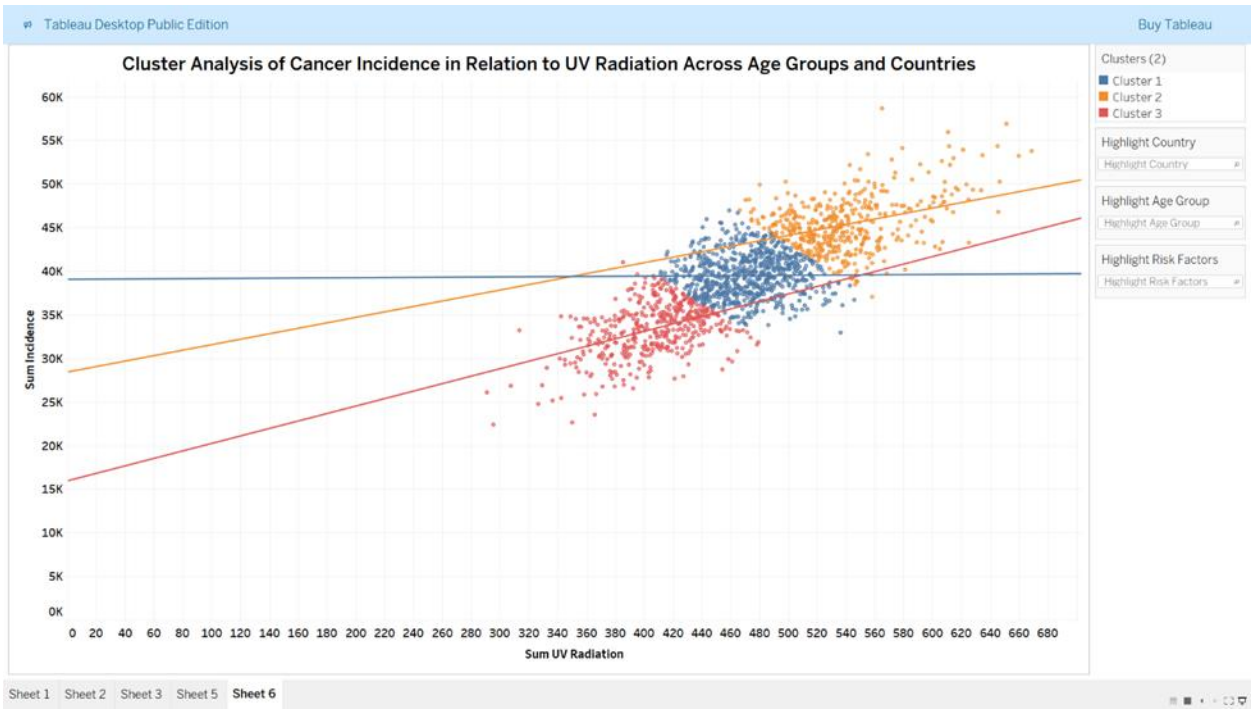

Figure 1.17: Cluster Analysis of Cancer Incidence in Relation to UV Radiation Across Age Groups and Countries. (Cancer Dataset: Top 50 Populated Countries 2019–2023)

Explanation:

The data source is "Cancer Dataset (Top 50 Populated Countries)" from Kaggle by Ankush Panday, featuring 160,000 health records. This chart presents a cluster analysis of cancer incidence in relation to UV radiation, using data filtered by age groups (15–24, 25–54, 55–64, 65+). The X-axis shows total UV Radiation exposure, while the Y-axis represents the total number of cancer incidence cases. The data is divided into three clusters (blue, orange, red), grouped by similar patterns using Tableau's built-in clustering model. Each dot represents a specific country-age group combination, and trend lines show the direction and strength of the relationship within each cluster. According to the trend line model, Cluster 3 and Cluster 2 show a positive correlation between UV exposure and cancer incidence, while Cluster 1 shows no significant trend ($p = 0.789$). The overall R-squared is 0.7468, which means about 75% of the variation in cancer incidence is explained by UV radiation levels. All clusters except Cluster 1 have p-values < 0.0001, confirming that UV radiation significantly affects cancer incidence in those

groups. The dataset includes data from 51 countries, 8 risk factors (like tobacco, obesity, genetic), and 4 age groups, providing rich insights. This graph helps visualize how UV radiation may impact cancer risks differently by region and age, aiding targeted prevention strategies.

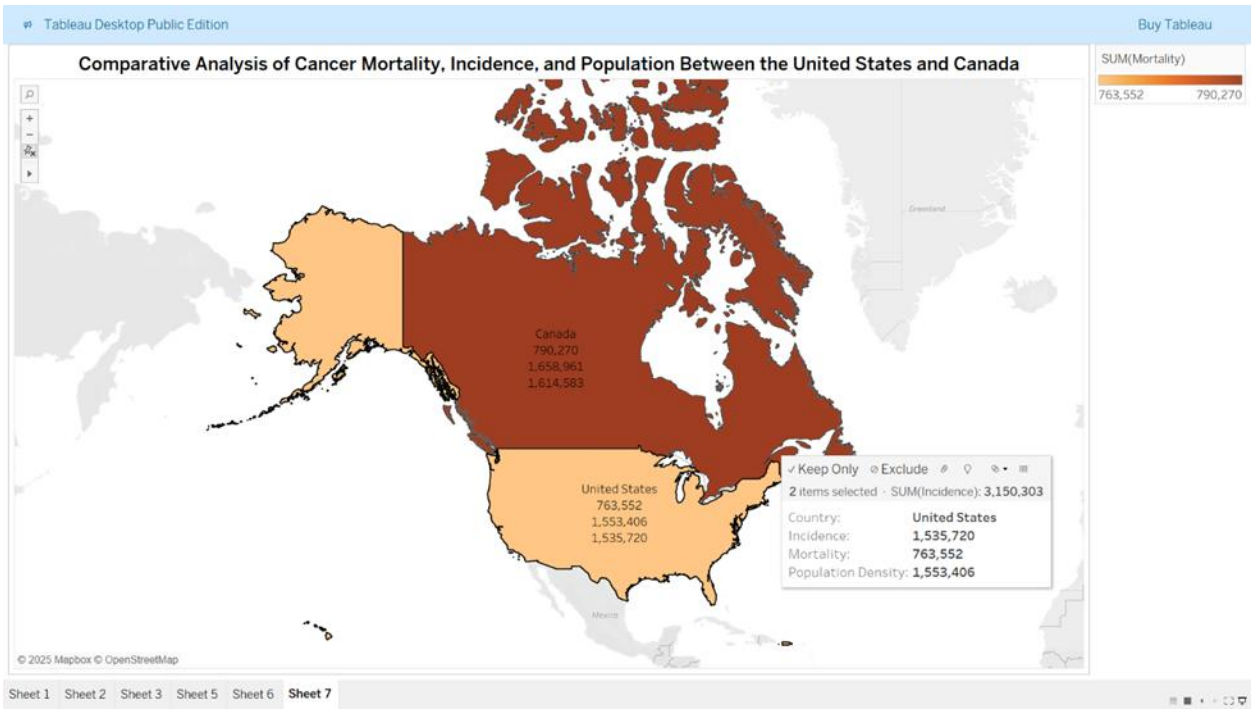

Figure 1.18: Comparative Map of Cancer Mortality, Incidence, and Population Density Between the United States and Canada. Cancer Dataset (Top 50 Populated Countries 2019–2023)

Explanation:

This map compares cancer mortality, incidence, and population density between Canada and the United States using color intensity and text labels. The color shade represents the sum of cancer mortality, with darker shades indicating higher mortality. The text label inside each country shows three values: mortality, population density, and incidence — in that order. Canada reports 790,270 deaths, a population density of 1,658,961, and 1,614,583 incidence cases. The U.S. shows 763,552 deaths, a population density of 1,553,406, and 1,535,720 incidence cases. Despite similar population densities, Canada has slightly higher mortality and incidence than the U.S. This view is filtered to show only two countries for focused analysis and clearer regional comparison. The data source is the Kaggle Cancer Dataset, featuring health metrics from over 160,000 records. This comparison allows us to explore how population factors relate to national-level cancer burden.

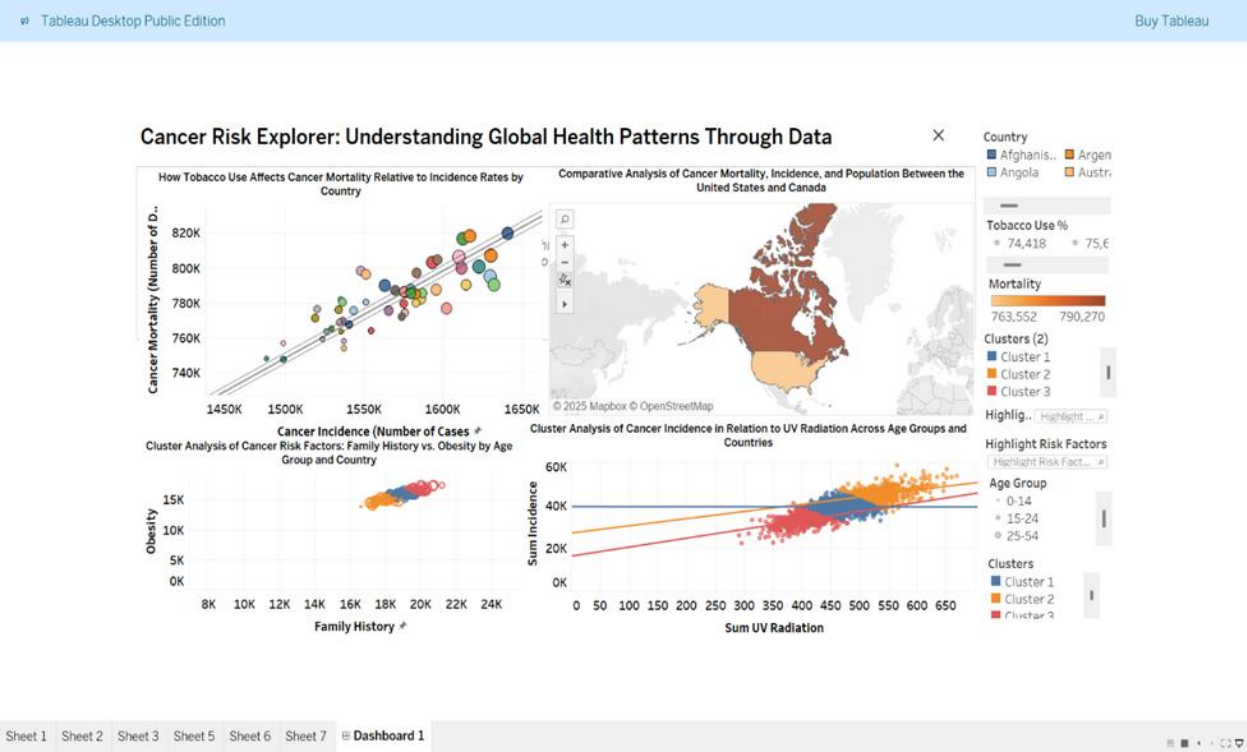

Figure 1.19: Cancer Incidence and Risk Factors Across Countries and Geographical Comparison Between U.S. and Canada. Cancer Dataset (Top 50 Populated Countries 2019–2023)

Explanation:

This dashboard explores cancer incidence, mortality, and key risk factors across countries using multiple visualizations. The top-left scatter plot shows how tobacco use affects cancer mortality relative to incidence rates, with each bubble representing a country. A strong linear trend indicates that countries with higher cancer incidence often face higher mortality, especially where tobacco use is high. The bottom-left bubble chart clusters countries based on Family History and Obesity, revealing age-related health patterns in cancer risks. Each cluster represents similar risk factor profiles, colored for clear segmentation across age groups. The bottom-right scatter plot displays the relationship between UV radiation and cancer incidence, split by clusters and age groups. It shows that higher UV exposure is associated with higher incidence rates in some clusters, highlighting environmental impacts. The geographical map (top-right) compares Canada and the United States in terms of mortality, incidence, and population density.

Although both countries show high values, Canada has slightly higher mortality and incidence, prompting potential investigation into healthcare or environmental differences. All visuals use data from the Kaggle "Cancer Dataset (Top 50 Populated Countries)", offering a global lens on cancer dynamics and public health priorities.

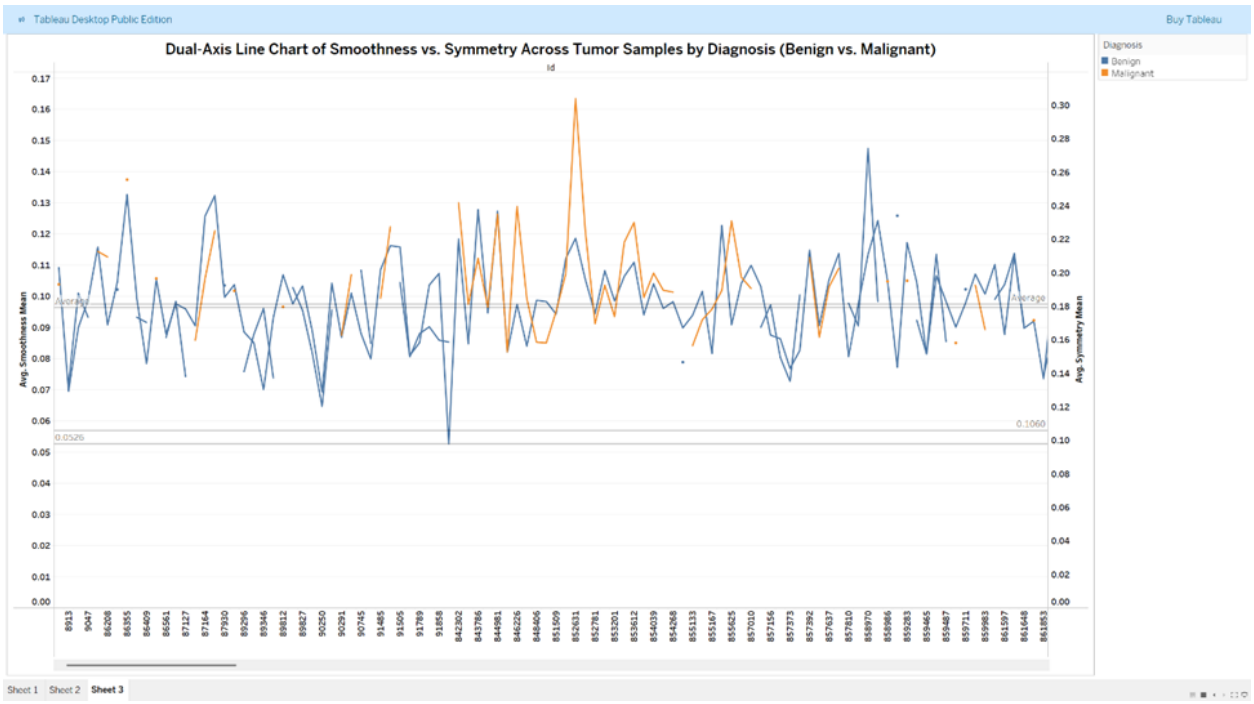

Figure 1.20: Dual-Axis Line Chart of Average Smoothness and Symmetry Across Tumor Samples by Diagnosis (Benign vs. Malignant). Breast Cancer Dataset (Wisconsin, 1989–1995)

Explanation:

This chart uses data from the Breast Cancer Wisconsin (Diagnostic) Dataset, sourced via Kaggle, and processed in Tableau. It displays two continuous variables: Average Smoothness Mean and Average Symmetry Mean. The X-axis represents unique patient IDs (569 total), while the Y-axes show the value ranges of the two metrics. The left Y-axis (0.0526 to 0.1634) is for Avg. Smoothness Mean; the right Y-axis (0.1060 to 0.3040) is for Avg. Symmetry Mean. The lines are color-coded by Diagnosis: blue for Benign (B) and orange for Malignant (M). Each dot/line point indicates the average score of smoothness or symmetry for that specific tumor sample. Malignant tumors tend to show greater variance and higher peaks, especially in symmetry. This dual-axis setup helps compare how tumor texture and shape characteristics behave between diagnosis types. Reference lines mark the average for each metric, aiding in visual comparison across the dataset. The chart enables easy detection of trends and anomalies in tumor features that could support classification modeling in cancer diagnosis.

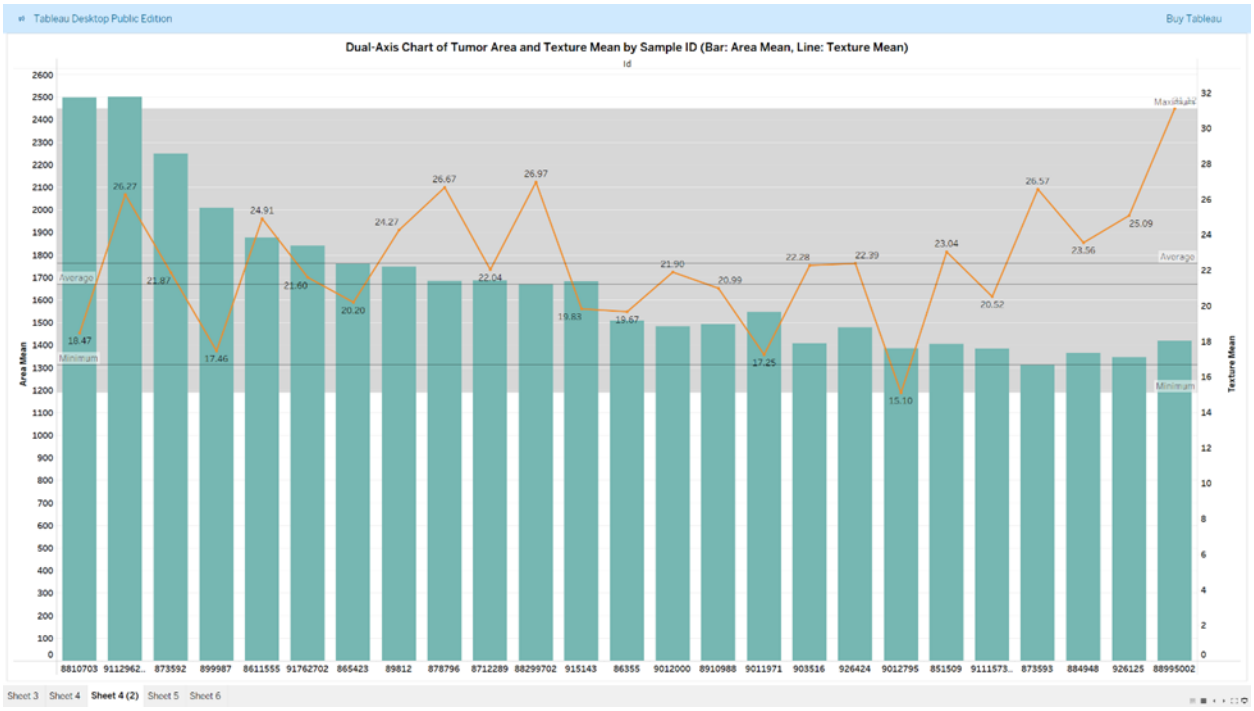

Figure 1.21: Dual-Axis Chart of Tumor Area Mean and Texture Mean by Sample ID (Bar: Area Mean, Line: Texture Mean). Breast Cancer Dataset (Wisconsin, 1989–1995)

Explanation:

This dual-axis chart visualizes Area Mean (as bars) and Texture Mean (as a line) for tumor samples based on their Sample IDs. The dataset is sourced from breast-cancer.csv on Kaggle and includes diagnostic metrics for breast cancer classification. Each sample ID on the X-axis represents an individual tumor record. The Area Mean ranges from 1311 to 2501, while the Texture Mean ranges from 15.10 to 31.12. Area Mean is plotted on the left Y-axis and reflects tumor size, while Texture Mean is on the right Y-axis indicating texture irregularity. Reference lines indicate average, minimum, and maximum values to support quick comparisons. Samples with higher Area Mean often correlate with higher Texture Mean, suggesting more irregular and possibly malignant features. Only 25 out of 569 samples were filtered for this view to ensure clarity. The orange line (Texture) intersects the teal bars (Area), highlighting differences across samples. This chart helps in identifying tumors with significantly higher area or texture, which may warrant closer examination.

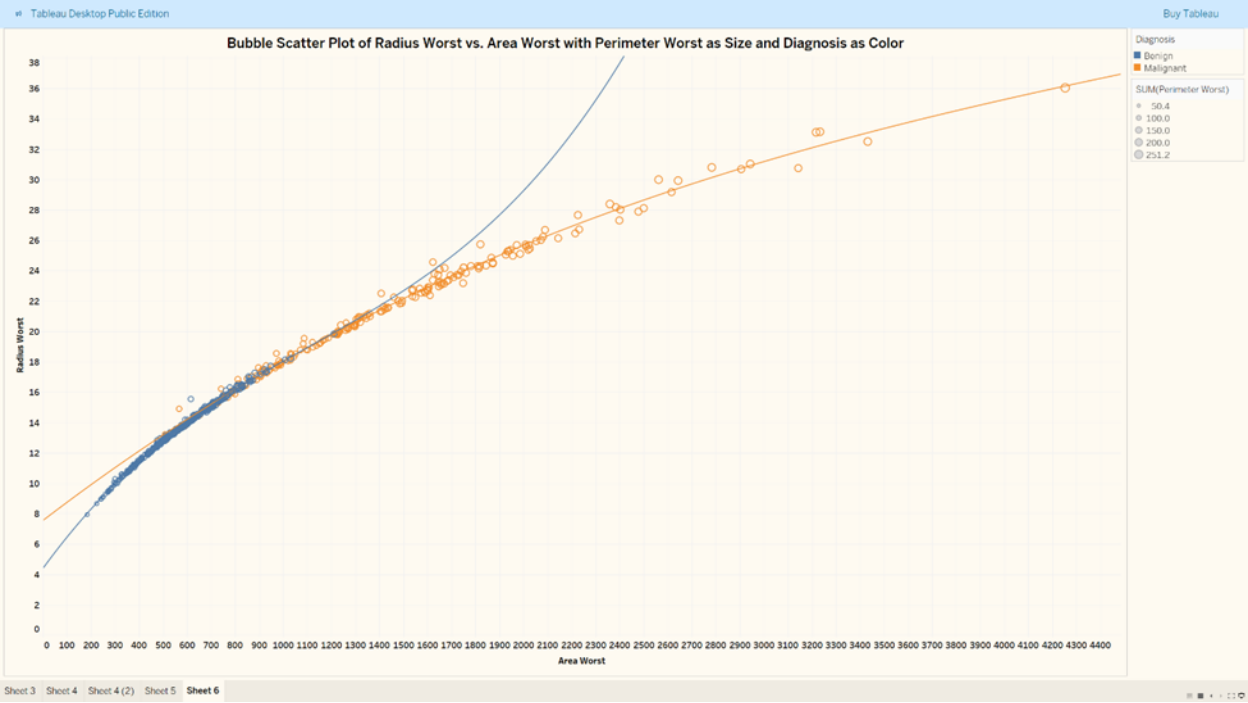

Figure 1.22: Bubble Scatter Plot of Radius Worst vs. Area Worst with Perimeter Worst as Size and Diagnosis as Color (Benign vs. Malignant). (Breast Cancer Dataset, Wisconsin, 1989–1995)

Explanation:

This visualization uses a bubble chart to display the relationship between Area Worst (X-axis) and Radius Worst (Y-axis) for tumor samples. Each bubble represents one tumor ID from the dataset sourced from Kaggle's Breast Cancer Wisconsin (Diagnostic) dataset. The bubble size reflects the Perimeter Worst, helping us visually distinguish tumor severity and size. The color indicates the Diagnosis, with blue for Benign and orange for Malignant tumors. A polynomial trend line of degree 3 has been added to observe the trend of Radius Worst against Area Worst across different diagnoses. The trend lines show that malignant tumors tend to have larger areas and radii than benign ones. R-squared values are very high (0.997 and above), suggesting a strong correlation between Area Worst and Radius Worst. The analysis includes 569 tumor samples, with Area Worst ranging from 185 to 4254 and Radius Worst from 7.93 to 36.04. The trend lines are statistically significant ($p < 0.0001$), confirming that diagnosis type impacts tumor shape

dimensions. This chart helps identify key physical characteristics of tumors that may aid in early breast cancer detection and classification.

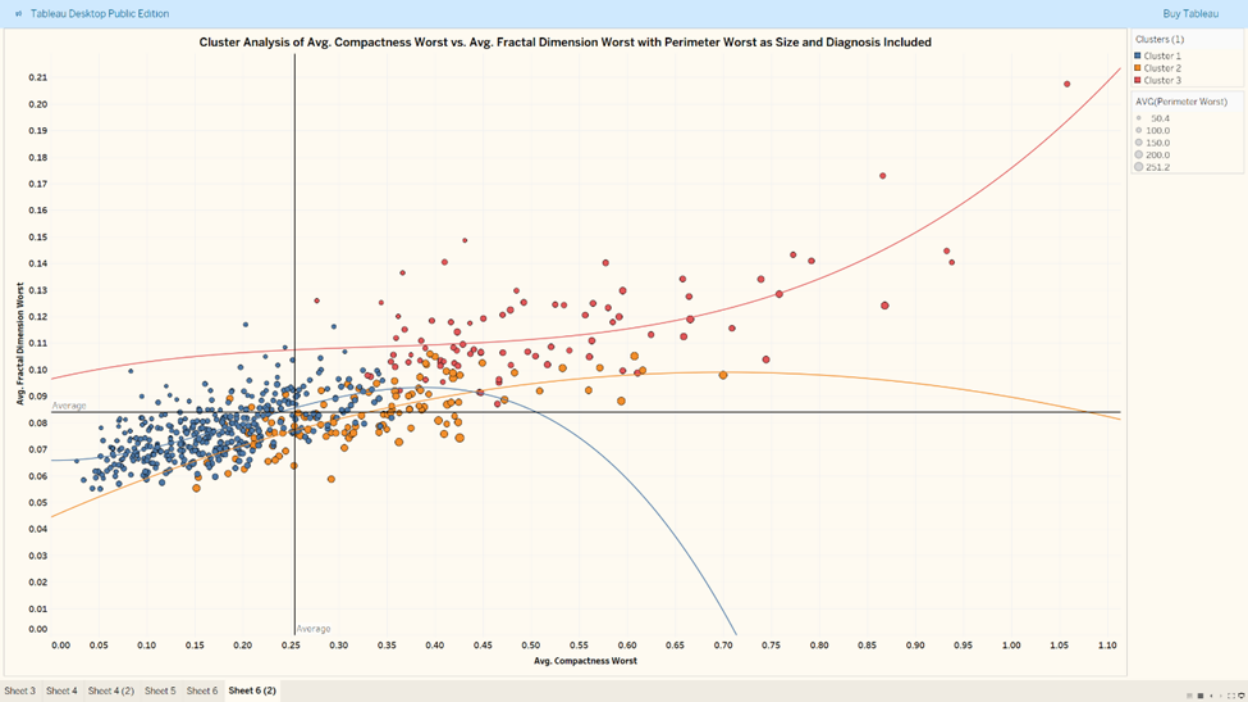

Figure 1.23: Cluster Analysis of Avg. Compactness Worst vs. Avg. Fractal Dimension Worst with Perimeter Worst as Size and Diagnosis Included (Wisconsin, 1989–1995).

Explanation:

This visualization presents a Cluster Analysis of Avg. Compactness Worst vs. Avg. Fractal Dimension Worst using a polynomial trend model. The dataset contains 569 tumor samples from the breast-cancer.csv file, classified into three clusters based on similarities in tumor shape features. Compactness Worst is plotted on the X-axis, and Fractal Dimension Worst is plotted on the Y-axis, indicating the irregularity and structure of tumor edges. The size of each bubble reflects the Average of Perimeter Worst, helping to visualize tumor size differences. Color represents cluster groupings: Cluster 1 (blue), Cluster 2 (orange), and Cluster 3 (red), with added diagnosis detail (Benign or Malignant). The model uses a degree 3 polynomial to fit trend lines per cluster, which helps capture the non-linear relationships between the features. The R-squared value is 0.7528, indicating a reasonably strong fit of the model across all clusters. Individual coefficients for each cluster (Cluster 1, 2, and 3) highlight varied relationships, with some showing higher t-values and lower p-values (significant results). The chart also includes vertical

and horizontal reference lines at average values, offering a baseline for interpretation. This analysis helps identify how shape-related tumor features cluster and correlate, offering insights into potential malignancy groupings for diagnostic prediction.

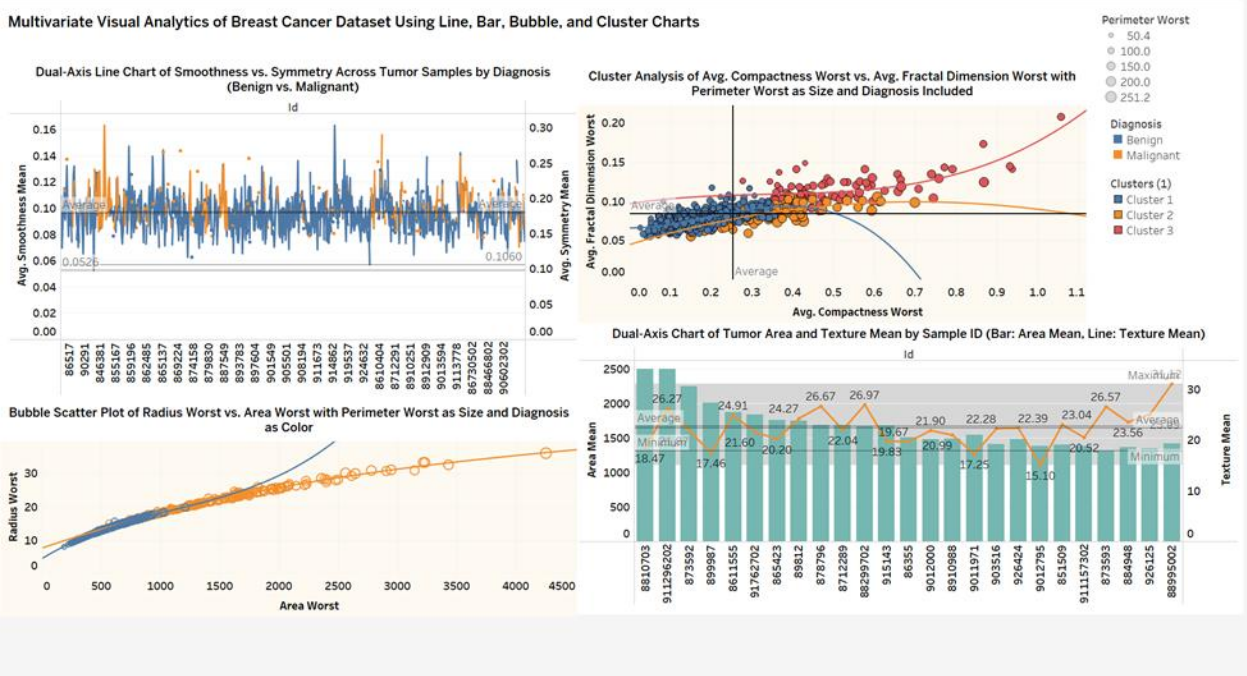

Figure 1.24: Multivariate Visual Analytics of Breast Cancer Dataset Using Line, Bar, Bubble, and Cluster Charts. (Breast Cancer Dataset, Wisconsin, 1989–1995).

Explanation:

This dashboard presents a multivariate analysis of the Breast Cancer Wisconsin (Diagnostic) dataset using four distinct visualization types. The top-left dual-axis line chart compares Avg. Smoothness Mean and Avg. Symmetry Mean across sample IDs, color-coded by diagnosis (Benign or Malignant). The top-right cluster chart shows how Avg. Compactness Worst relates to Avg. Fractal Dimension Worst, grouped into three clusters, with Perimeter Worst as size and diagnosis included. The bottom-left bubble scatter plot maps Radius Worst vs. Area Worst, with bubble size representing Perimeter Worst and color indicating diagnosis. Polynomial trend lines highlight different diagnostic patterns, showing a clear upward trend for malignant tumors in all charts. The bottom-right dual-axis bar and line chart compares Area Mean (bar) and Texture Mean (line) across selected samples. Filters and color legends on the right side help viewers interpret data clusters and diagnostic categories easily. Each graph includes reference lines for minimum, maximum, and average values, improving comparison across measures. The data source is Kaggle's breast-cancer.csv, containing 569 tumor records and more than

30 clinical attributes. Overall, this dashboard aids in understanding how tumor features relate to malignancy, supporting both visual pattern recognition and data-driven insights.

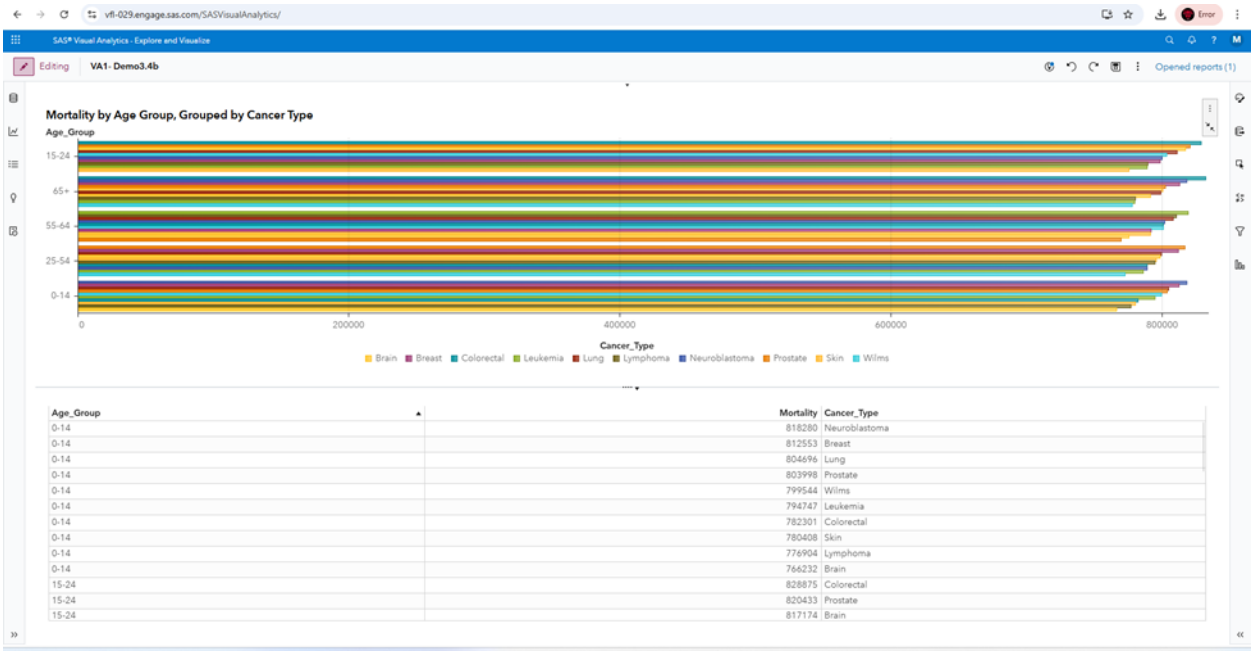

Figure 1.25: Cancer Mortality by Age Group and Cancer Type: A Multivariate Analysis Using SAS Viya. Cancer Dataset (Top 50 Populated Countries 2019–2023).

Explanation:

This graph analyzes cancer mortality by age group and cancer type using the Cancer Dataset (Top 50 Populated Countries) from Kaggle. The dataset includes mortality data across 51 countries and cancer types like breast, lung, prostate, leukemia, and more. As a SAS Viya Learners student, I used SAS® Visual Analytics to build this multivariate chart for exploration and pattern discovery. The bar chart displays mortality counts by age group, grouped and color-coded by cancer type for clear comparison. Age groups range from 0–14 to 65+, showing how cancer mortality varies significantly with age. The 65+ age group shows the highest mortality across nearly all cancer types, especially for colorectal, breast, and prostate cancer. In contrast, Neuroblastoma and Wilms tumors dominate in the youngest age group (0–14). Each colored bar represents a distinct cancer type, helping visualize which cancers are most fatal in each age group. The table below the chart lists exact mortality values to support data transparency and detail. This visual analysis supports public health planning, early detection efforts, and age-focused prevention strategies.

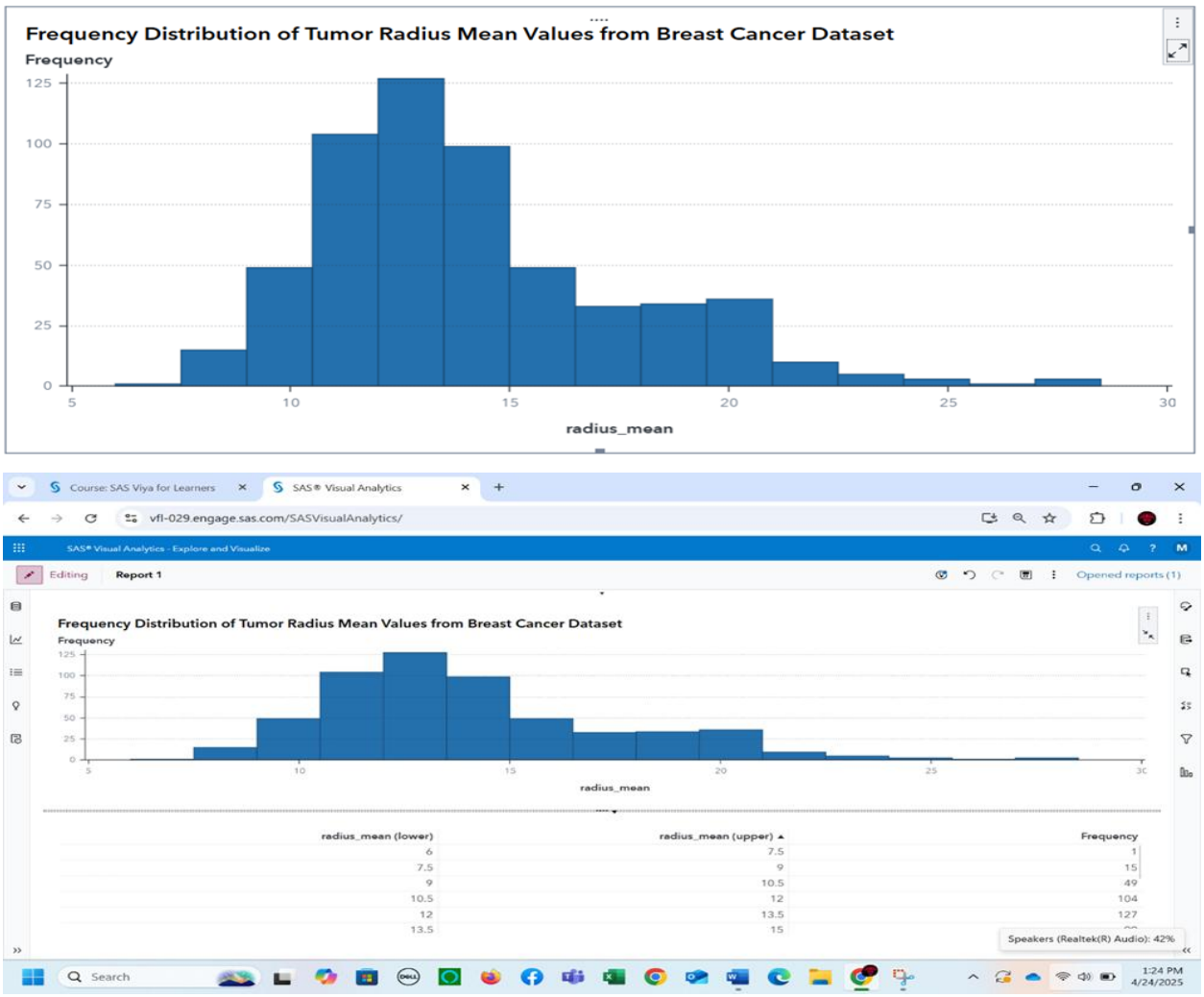

Figure 1.26: Histogram Showing Frequency Distribution of Tumor Radius Mean Values in (Breast Cancer Dataset, Wisconsin, 1989–1995)

Explanation:

This visualization is based on the Breast Cancer Wisconsin dataset and shows the frequency distribution of tumor radius mean values. The dataset contains the average radius of tumors measured for multiple cases. The x-axis of the histogram represents radius_mean, binned into intervals (e.g., 6–7.5, 7.5–9, etc.). The y-axis represents frequency, or how many cases fall into each radius range. The most common tumor size range is between 12 and 13.5, with a frequency of 127 cases. The second most frequent range is 10.5–12, showing 104 cases. As tumor radius increases beyond 15, the frequency starts to drop significantly. Only a few tumors fall in the larger size ranges (e.g.,

27–28.5 has just 3 cases). The histogram is right-skewed, indicating most tumors have a smaller radius. This chart helps in understanding tumor size distribution, useful in diagnostic and treatment decisions.

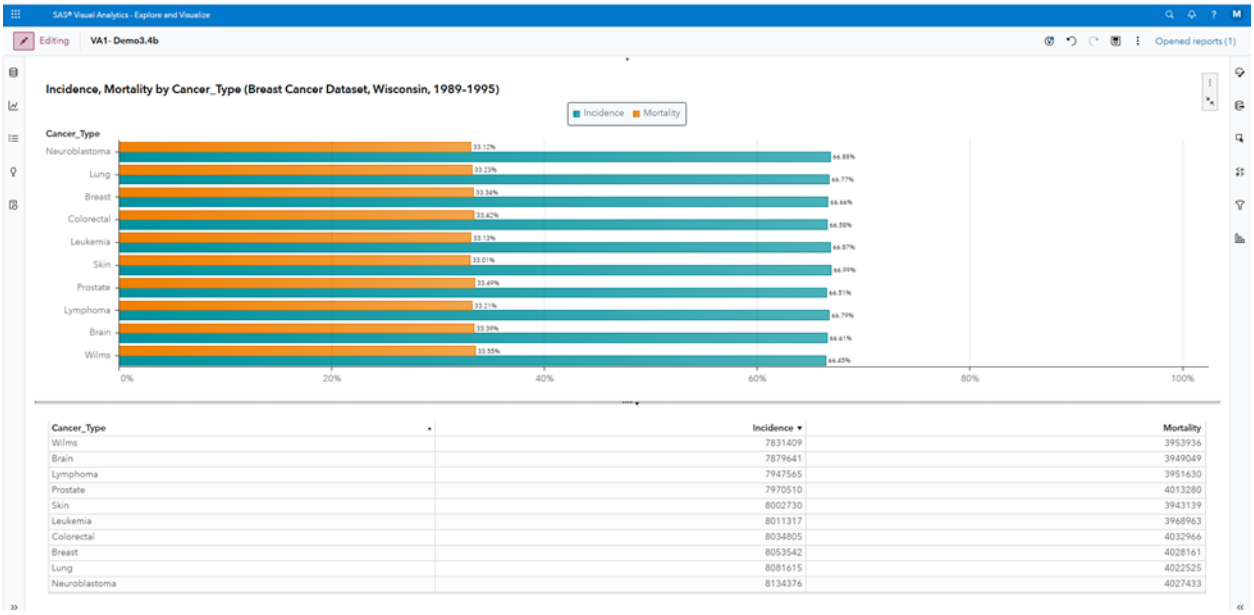

Figure 1.27: Comparison of Cancer Incidence and Mortality by Cancer Type (Breast Cancer Dataset, Wisconsin, 1989–1995)

Explanation:

This visualization uses the Breast Cancer Dataset (Wisconsin, 1989–1995) sourced from Kaggle. It compares cancer incidence and mortality rates across 11 different cancer types. The bar chart displays percentage values for both incidence (blue) and mortality (orange) side by side. Each cancer type shows its share of total cases and total deaths in the dataset. Neuroblastoma has the highest number of cases (~8.13 million) and a relatively close mortality (~4.02 million). Wilms cancer shows the largest mortality percentage (33.55%), with the lowest incidence count (~7.83 million). The horizontal format makes it easy to compare proportional risk for each cancer type visually. The dataset reveals that mortality rates remain around one-third of the total incidence across all types. The table below the chart provides exact incidence and mortality counts, enhancing transparency. This SAS Viya dashboard allows learners and analysts to explore relative risk and survival patterns using grouped bar visuals.

Conclusion: Phase I, II, III

Cancer is a **global health crisis** that touches every part of society—affecting individuals, families, and healthcare systems in unique ways. Through this three-phase project, powered by **data visualization**, we have uncovered both the **macro-level patterns** and **micro-level diagnostics** that shape cancer outcomes worldwide. By analyzing data from the **Global Cancer Dataset (2019–2023)**, the **CDC BRFSS 2022 Heart Disease Dataset**, and the **Breast Cancer Wisconsin Diagnostic Dataset**, this study reveals how **lifestyle**, **demographics**, and **clinical features** intersect in the fight against cancer and heart disease.

In the first phase, global cancer trends were mapped to highlight regions with **high incidence and mortality**, uncovering urgent needs for **prevention programs**, **healthcare investments**, and **lifestyle changes**. Visualizations such as the **Global Cancer Prediction World Map**, **risk factor treemaps**, and **scatter plots** linking incidence to mortality revealed a powerful connection between **behavioral risk factors**—like **tobacco use**, **obesity**, and **physical inactivity**—and cancer outcomes. These findings stress the need for **early detection**, **awareness campaigns**, and **equitable healthcare access**, particularly in underserved nations.

Phase two expanded the analysis to include **U.S. public health trends**, illustrating how **age**, **race**, **sleep patterns**, and **BMI** influence the prevalence of **heart disease** and **general health**. Young adults showed higher mental health challenges, while older groups faced worsening physical health and increasing heart disease risks. The connection between **sleep quality**, **obesity**, and **disease prevalence** was made clearer through detailed visualizations, emphasizing the importance of **personal health behaviors** and **targeted intervention** strategies. Additionally, by comparing cancer mortality across countries, this phase highlighted **regional disparities** and the urgent need for **policy-level improvements** in cancer care.

In the final phase, the focus narrowed to the **clinical level**, using **tumor-specific features** such as **area**, **texture**, **smoothness**, and **symmetry** to distinguish between

benign and malignant tumors. Through advanced visualization techniques in **Tableau** and **SAS Viya**, we identified diagnostic patterns that can aid **early detection** and **treatment planning**. The multivariate dashboards and cluster analyses offered valuable insights into how **tumor characteristics** and **age-based mortality trends** vary across populations. This approach not only complements public health data but also empowers medical professionals with tools to make better, faster decisions.

Altogether, this project demonstrates the **transformative power of big data and visual analytics** in addressing one of humanity's most complex health challenges. By combining **global patterns**, **population-specific risks**, and **individual tumor behavior**, we have created a rich, multidimensional understanding of cancer and related diseases. The knowledge gained through this study can guide **policymakers**, **healthcare providers**, and **communities** toward smarter interventions, stronger prevention programs, and more **personalized care strategies**. Data alone doesn't save lives—but **data-driven action** does. This project is a call to turn insight into impact, paving the way toward a **healthier, more informed, and cancer-resilient world**.

References:

Phase I:

Ankush, Panday. 2023. “*Cancer Dataset (Top 50 Populated Countries)”.* Kaggle, https://www.kaggle.com/datasets/ankushpanday1/cancerdatasettop-50-populated-countries.

Garcia, Sophia, et al. *"National Cancer Registries: A Critical Tool for Tracking Cancer Trends." Public Health Journal*, vol. 42, 2021, pp. 58–72. www.who.int/data.

Nguyen, David. *"Government Health Organizations and Cancer Research: A Global Perspective." World Health Organization*, 2022. www.who.int/publications.

Phase II:

Centers for Disease Control and Prevention (CDC).2022. *Behavioral Risk Factor Surveillance System (BRFSS) Annual Data 2022,* https://www.cdc.gov/brfss/annual_data/annual_2022.html.

Nguyen, David. 2022. *"Government Health Organizations and Cancer Research: A Global Perspective." World Health Organization*. www.who.int/publications.

Pytlak, Kamil. 2022a. *Indicators of Heart Disease (2022 Update).* Kaggle, https://kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

Pytlak, Kamil.2022b. *Heart Disease Prediction Data Processing Notebook.* GitHub, https://github.com/kamilpytlak/data-science-projects/blob/main/heart-disease-prediction/2022/notebooks/data_processing.ipynb.

Ankush, Panday. 2023. “*Cancer Dataset (Top 50 Populated Countries)”.* Kaggle, , https://www.kaggle.com/datasets/ankushpanday1/cancer-datasettop-50-populated-countries.

Sophia, Garcia, et al. 2021. *"National Cancer Registries: A Critical Tool for Tracking Cancer Trends." Public Health Journal*, vol. 42, pp. 58-72. www.who.int/data.

Phase III:

Ankush, Panday. 2023. "*Cancer Dataset (Top 50 Populated Countries)".* Kaggle, , https://www.kaggle.com/datasets/ankushpanday1/cancer-datasettop-50-populated-countries.

Sophia, Garcia, et al. 2021. *"National Cancer Registries: A Critical Tool for Tracking Cancer Trends." Public Health Journal*, vol. 42, pp. 58-72. www.who.int/data.

Nguyen, David. 2022. *"Government Health Organizations and Cancer Research: A Global Perspective." World Health Organization*. www.who.int/publications.

Yasser, M. 2022.*"Breast Cancer Dataset."* Kaggle, https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset.