Assessment cover

Module No:	DALT7016	**Module title:**	Data Visualisation
Assessment title:	Assignment Report		
Due date and time:		Friday 15th December 13:00	
Estimated total time to be spent on assignment		30 hours	

LEARNING OUTCOMES

On successful completion of this assignment, students will be able to achieve the module following learning outcomes (LOs):
Critically analyse data visualisation approaches with respect to human sensory modalities
Create appropriate visualisations for temporal, dynamic, and high dimensionality data
Devise methodologies for data interaction to facilitate exploratory data analysis

Engineering Council AHEP4 LOs assessed (from S2 2022-23)		**Met? (Y/N)**
M1	Apply a comprehensive knowledge of mathematics, statistics, natural science and engineering principles to the solution of complex problems. Much of the knowledge will be at the forefront of the particular subject of study and informed by a critical awareness of new developments and the wider context of engineering	
M2	Formulate and analyse complex problems to reach substantiated conclusions. This will involve evaluating available data using first principles of mathematics, statistics, natural science and engineering principles, and using engineering judgment to work with information that may be uncertain or incomplete, discussing the limitations of the techniques employed	
M3	Select and apply appropriate computational and analytical techniques to model complex problems, discussing the limitations of the techniques employed	
M17	Communicate effectively on complex engineering matters with technical and non-technical audiences, evaluating the effectiveness of the methods used	

Statement of Compliance *(please tick to sign)*

[X]
I declare that the work submitted is my own and that the work I submit is fully in accordance with the University regulations regarding assessments
(www.brookes.ac.uk/uniregulations/current)

Table of Contents

DALT7016 Data Visualisation Assignment Report

Part 1: Context and EDA

Context and Source Description

In 2022, the Centers for Disease Control and Prevention – CDC –, (2023a) postulates that almost 47% of all Americans have experienced at least a third of the risk factors of heart disease: high cholesterol, high blood pressure and have smoked.

By using findings from CDC's Behavioural Risk Factor Surveillance System (BRFSS) telephone survey of 2022 (the largest annual health survey of the United States),
we developed a profile (Table 2) that highlights the characteristics that the 'typical or mode' individual who suffered a heart-attack has in the US.
The profile has considered and applied the relative population sizes of survey respondents against the total population size of the U.S in 2022, - statistics obtained from the US Census Bureau report of 2022 (US Census Bureau 2022a; US Census Bureau 2022b). This is to ensure a true depiction of the reality among demographics.

The 246,022 observations, both categorical and numeric, from respondents (from across all the 50 US states) and 39 variables, refines our profiles to better match population data.

By focusing on modifiable risk factors (excluding age, race and sex), our data helps readers identify some trends between life-style and heart-disease.

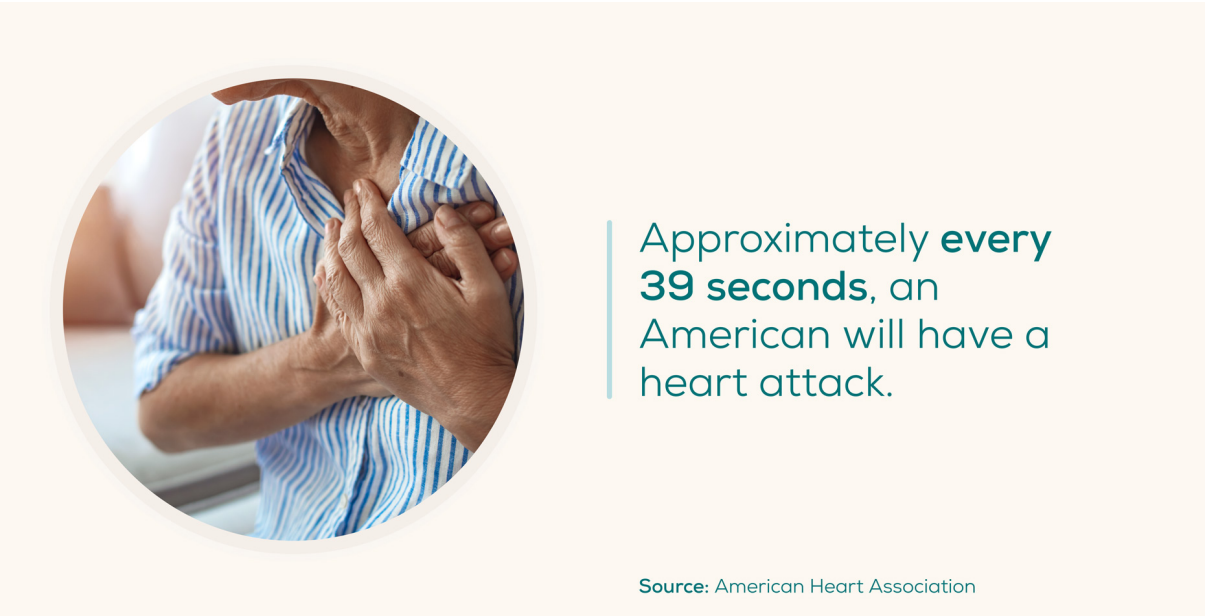


Figure 1 93 Heart Disease Facts and Statistics to Know for 2022: By Everlywell (2022) - Own work by the original uploader, CDC, https://www.everlywell.com/blog/heart-health/heart-disease-facts/

Dataset Description and Summary Statistics

The dataset derived from BRFSS 2022 consisting of 246,022 categorical and numeric observations, across 39 variables, acquires the frequency of (cardiac, respiratory, mental) health conditions and lifestyle choices of adult Americans (18 to +80 years old).

However, we have specifically selected 15 variables (see *Table 1* and *Table 2)* to create a *heart attack profile of the individual prone to experiencing it*.
One limitation of our is dataset is that it does not capture information about respondents' hereditary and non-modifiable factors (except age, race, and sex).
Therefore, we approach this problem by investigating trends in modifiable risk factors that are at the scope of the public, and could inspire positive lifestyle changes.
For example, the average BMI - μ 30.16 -, of persons with confirmed heart attacks, suggests an obese majority. Which may alert readers to consider not falling on the BMI range that commonly holds heart attack sufferers.

The summary of statistics *(Table 2*) demonstrates the total population proportion, total female population proportion and total male population proportion who had heart attacks and share the mode risk group characteristic. The values are transformed to mirror the demographics and population size reported by the US Census Bureau in 2022.

To limit collecting, unreliable information (i.e., self-diagnosis), the participants were requested to answer "Yes" to having a health-condition only if medical staff confirmed the results. However, we must cognise the potential limitations to the survey conducted, that might limit the validity of the data obtained.
This can include the honesty of the responses to the survey as health issues are generally a very private matter. Therefore, some respondents may cease to disclose their private health information, which could potentially skew the data and lower the surveyed rate of cardiac issues. Another limitation could be the possibility that people with a heart issue may have an increased likelihood to respond to the survey due to their familiarity with the subject matter. This, however, would work adversely to the previous issue of sensitivity to revealing private information as it would potentially increase the rate of cardiac issue in an unrepresentative manner. Thus, our profile and visualisations are guides that attempt to approximate our understanding of potential trends between heart attacks and particular demographics.

Table 1: List of Variables

Variable	Description
Heart Attack	The number of adult persons with a heart-attack diagnosis confirmed by a medical staff.
Had Angina	The number of adult persons with an angina diagnosis, confirmed by a medical staff.
Had Stroke	The number of adult persons with a stroke diagnosis, confirmed by a medical staff.
Had Asthma	The number of adult persons with an asthma diagnosis, confirmed by a medical staff.
Had COPD	The number of adult persons with a Chronic Obstructive Pulmonary Disease diagnosis, confirmed by a medical staff.
High Risk Group Characteristics (mode)	The mode characteristics found in persons, who experience or have experienced heart attacks in their life. And which diagnosis was confirmed by medical staff. The uneven spread of responses from different risk groups with distinct demographics is considered in the weighing of percentages. High Risk Group Characteristics captures the relative significance of the relationship between these groups and heart attacks.
%	For each *High-Risk Group Characteristic*, the percentage % approximates us to understand the true number of adult American citizens (both women and men) with said characteristic. The values are proportionate to population data in the USA census of 2022.
Female	Female represents the percentage (%) of adult American women that possess the named *High-Risk Group Characteristic*. The values are proportionate to population data in the USA census of 2022.

Male	Male represents the percentage (%) of adult American men that posses the named *High-Risk Group Characteristic*. The values are proportionate to population data in the USA census of 2022.
Respondents Total	The total number of respondents who had a heart attack and simultaneously have the 'named' *High-Risk Group Characteristic* (row).
Race	The self-identification of race of Americans. The categories reflect a social definition of race, socially recognised by Americans.
Age Group	Age grouped in intervals of 5 (inclusive). This ranges from 18 years of age to 80 years of age or older.
BMI	The average Body Mass Index of Americans. Metrically calculated as: weight(kg) // height(m)
Kidney Disease	The number of adult persons with a kidney disease diagnosis, confirmed by a medical staff.
Diabetes	The number of adult persons with a diabetes diagnosis, confirmed by a medical staff.
Has Smoked	The number of adult persons who have or currently smoke.
Drinks	The number of adult persons who have or currently drink.
Physically Active	The number of adult persons with the corresponding *Risk Factor* diagnosis (confirmed by a medical staff), who **are** physically active.
Unhealthy Days	The number of adult persons with the corresponding *Risk Factor* diagnosis (confirmed by a medical staff), who have had **n days** (mode) of poor health.
Poor Mental Health Days	The number of adult persons with the corresponding *Risk* Factor diagnosis (confirmed by a medical staff), who have had **n days** (mode) of poor mental health days

Table 2: Summary Statistics – The Profile of the Common Adult American Citizen with a Heart Attack Diagnosis in 2022

Risk	High Risk Group Characteristic (mode)		Data Type	Population %	Female Pop	Male Pop	Respondents Total
Heart Attack	**Race**	Multiracial, Non-Hispanic	Category	0.14%	0.87%	1.47%	13,435
	Age Group	65 - 69	Category	0.42%	1%	1.34%	2,155
	BMI	μ 30.16	Numeric	0.14%	μ 30.66	μ 29.86	340
	Poor Mental Health Days	7	Numeric	0.14%	7	6	8,610
	Kidney Disease	Yes	Category	2.07%	0.84%	1.50%	11,540
	Diabetes	No	Category	1.30%	0.75%	1.59%	8,333
	Ever Smoked	Never	Category	0.85%	1.05%	1.29%	5,471
	Drinks	No	Category	1.50%	0.98%	1.36%	8,109
	Physically Active	Yes	Category	1.56%	0.82%	1.52%	8,514
	Unhealthy Days in a Month	0	Numeric	0.98%	0.61%	1.73%	6,018
	Had Angina	Yes	Category	0.07%	0.81%	1.53%	14,953
	Had Stroke	Yes	Category	0.03%	0.83%	1.51%	10,112
	Had Asthma	Yes	Category	0.03%	1.11%	1.22%	36,529
	Had COPD	Yes	Category	0.04%	0.89%	1.45%	994
Total Frequency					127, 811	118,211	

Part 2: Design

The purpose of the static visualisation is to capture the prevalence of heart attacks among the mode race/ethnicity from the BRFSS 2022 survey.
According to the U.S census data, 58.9% of Americans are white only. This aligns with our dataset (Figure 2), where the number of white respondents is higher. Consequently, the visual spread in visualisation_1 is anticipated to exhibit a steadier increase compared to other demographics.

The bar chart should highlight noticeable peaks and troughs in the reported cases within the Hispanic, Multiracial (Non-Hispanic) and Other race only, (Non-Hispanic) groups. This should be noticeable from the 40-45 age range. The comparable bar charts should prompt further inquiry into the specific social or biological factors that may contribute to these fluctuations among minority groups.

The comparative bar charts vividly illustrate the actual counts of reported heart attacks categorised by the U.S definition of race. Additionally, we have confirmed that the ethnic or racial group with the highest incidence of heart-disease, based on the relative sample size, is Multiracial Non-Hispanic.
For this reason, our focus remains on refining our heart attack profile using data solely from respondents in the BRFSS 2022 survey.

The interactive visualisation_2 delves into a more detailed exploration of the continuously increasing number of reported heart attack cases within the mode race/ethnicity group that exhibits the highest record of heart attacks.
Visualisation_2 displays cumulative frequency information when hovering over each plotted dot of the graph, providing a total count of heart attack reports for each age-group.

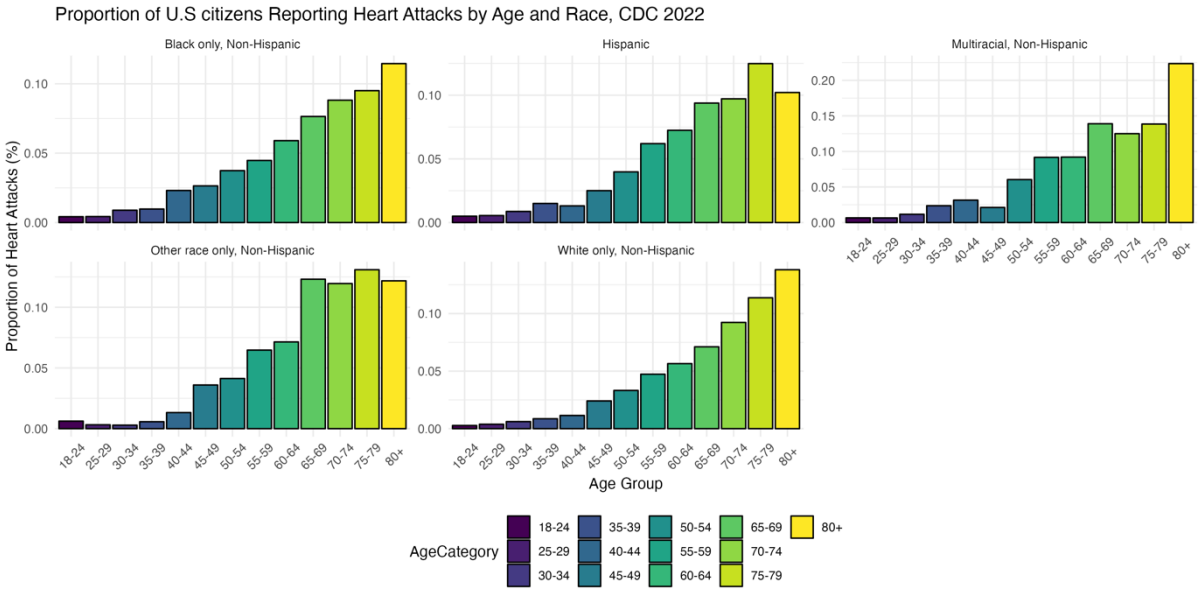
Both graphs will use the *viridis* package “magma” to cater to colour-blind audiences and mature audiences with declining vision. The target audience for the visualisations includes mature 65–69-year-olds and older, who may benefit from straightforward, informative bar chart. The graphs should ignite questions about the factors related to the trend of heart attacks and cardiac complications by age and instil health consciousness among mature viewers.

Originally the dataset combines categorical and numeric values, but to provide effective visualisations, it is required to recode ‘HadHeartAttack’ values, from categorical to numeric.

Part 3: Final Visualisations

Visualisation 1 Commentary

Figure 2: Visualisation 1 – *Proportion of U.S Citizens Reporting Heart Attacks by Age and Race, CDC 2022*



The comparative bar charts were created through the library ggplot2, and use the records of BRFSS 2022 survey respondents who answered 'Yes' to having experienced a heart attack which was confirmed by medical staff.

The static png file size is relatively large considering that the data is extracted from a large dataset. The colour pallete contributes to this. Compressing the file further would deteriorate the quality and limit audiences with visual difficulties to learn from the data.

Furthermore, since it is a set of comparative bar charts very close to each other, it is crucial to ensure that all the items are clearly readable, which is why the borders are outlined in black and the background is white. Also, the *viridis* colour palette is applied to enhance accessibility.

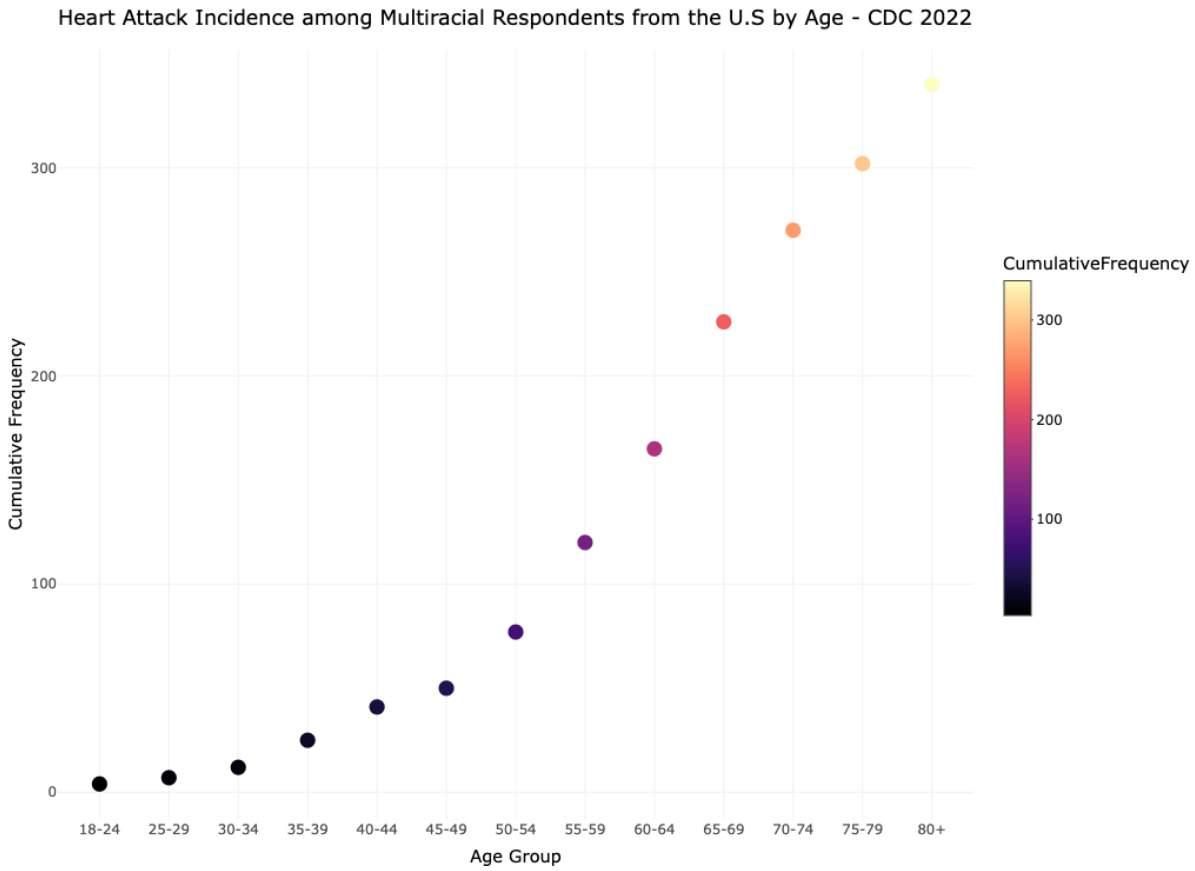
We have attempted to ensure that the *values* on the y and x axis are equally reflected and distributed on every graph, but given that the 'multiracial, non-Hispanic' reported over 0.15% heart attacks, its graph appears smaller in comparison to the rest.

However, the simplicity of the graphs should permit viewers notice that the peak value for 'multiracial, non-Hispanic' is significantly higher (>0.20%) than the other groups.

The complications with the R script limited our ability to apply a *y* axis label of 0.20 for each bar chart.

Visualisation 2 Commentary

Figure 3: Visualisation 2 – *Heart Attack Incidence Among Multiracial Respondents from the U.S by Age – CDC 2022*



The interactive html visualisation was created using the libraries plotly and htmlwidgets. The R script (lines 104-111) outlines the logic for extracting data on Multiracial, Non-Hispanic respondents who had heart-attacks which were confirmed by medical staff.

Notably, over 50% of the respondents who reported heart attacks were aged 64 and above. While, the information does not provide detailed insights into causation or correlation, it communicates that heart disease may be associated with age-related changes and factors. The research requires information on the hereditary and non-modifiable health conditions of the respondents to explore the relationship between age and heart disease in detail for this particular group.

Similarly to visualisation_1, the file size for visualisation_2 is large due to the large dataset used for generating figures.
Additionally, the inclusion of interactive libraries augments the size. However, despite the large size, the visualisation is responsive when interacting with plotted dots, without detectable delays.

Finally, the interactive graph performs as intended as it illustrates the cumulative frequency of the primary ethnic/racial group (with a relatively higher confirmed cases of heart attacks).

Reference list

BRFSS (2022) *2022 BRFSS Questionnaire* , *CDC Gov*. Available at: https://www.vdh.virghttps//www.cdc.gov/brfss/questionnaires/pdf-ques/2022-BRFSS-Questionnaire-508.pdfinia.gov/content/uploads/sites/68/2022/03/BRFFS-Questionnaire-2022NEW.pdf (Accessed: 8 December 2023).

BRFSS (2023) *2022 BRFSS Questionnaire*, *Centers for Disease Control and Prevention* , pp. 18–117. Available at: https://www.cdc.gov/brfss/questionnaires/pdf-ques/2022-BRFSS-Questionnaire-508.pdf (Accessed: 10 November 2023).

CDC (2023a) *2022 BRFSS Survey Data and Documentation*, *Centers for Disease and Control Prevention*. Available at: https://www.cdc.gov/brfss/annual_data/annual_2022.html (Accessed: 7 November 2023).

CDC (2023b) *Heart Disease Risk Factors*, *Centers for Disease Control and Prevention*. Available at: https://www.cdc.gov/heartdisease/risk_factors.htm (Accessed: 8 December 2023).

Everlywell (2022) *93 Heart Disease Facts and Statistics to Know for 2022*. Available at: https://www.everlywell.com/blog/heart-health/heart-disease-facts/ (Accessed: 15 December 2023).

National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health (2023) *Summary Matrix of Calculated Variables (CV) in the 2022 Data File*, *CDC Gov*. Available at: https://www.cdc.gov/brfss/annual_data/2022/summary_matrix_22.html (Accessed: 14 December 2023).

Pytlak, K. (2023) *Personal Key Indicators of Heart Disease*, *Kaggle*. Available at: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data (Accessed: 27 November 2023).

US Census Bureau (2022a) *2022 National and State Population Estimates Press Kit*, *Census Gov*. Available at: https://www.census.gov/newsroom/press-kits/2022/2022-national-state-population-estimates.html (Accessed: 10 December 2023).

US Census Bureau (2022b) *U.S. Census Bureau QuickFacts: United States*, *United States Census Gov*. Available at: https://www.census.gov/quickfacts/fact/table/US/PST045222#PST045222 (Accessed: 5 December 2023).