



Post-Facto Analysis of Cardio- Vascular Diseases and their Causes

DATA 511 - C4: Visualization Concept

Nayantara Mohan – 2129865

Rohit Lokwani – 2129904

Shubha Changappa Palachanda – 2129848

Sravan Kumar Reddy Hande – 2129885

Tharun Kumar Reddy Karasani – 2129858

Prof. Nathan Mannheimer

Table of Contents

Executive Summary	3
Introduction	4
Objective and Scope.....	4
Design Process	5
Background Research	6
Existing Survey Review and Design Questions	6
Competitive Analysis	6
Visualization 1: Analysis of Heart Data	6
Visualization 2: Heart Disease Dashboard	8
Summary	9
Dataset Processing and EDA	10
Data set	10
Data Cleaning and Preparation	11
Visual Inspection of the Dataset	11
Transforming Data	12
Generating Synthetic Fields.....	12
Decision Trees.....	14
Design Ideation	16
Prototype Development	17
Prototype Round 1.....	17
Usability Testing.....	21
Prototype Round 2.....	24
Key Takeaways.....	24
Final Dashboard	27
Individual Visualizations	28

Dashboard Properties..... 35

Future Scope 36

Acknowledgements 37

References 38

Appendix 40

1. Executive Summary

This report provides a detailed overview of our process of developing an interactive dashboard on Cardiovascular Health. Starting with the topic introduction, the motivation, and the end goal, the report provides a detailed end-to-end description of the design process adopted from the research and ideation stages to the development and deployment stages. For the brevity of the document, greater emphasis is made on critical tasks like data profiling, data preparation, and design ideation using Five Design Sheets (FDS). With this pretext, the report focuses on the iterative prototype development stage and the summary of user feedback at each of these stages. This is followed by the presentation of the final dashboard, which highlights how it fulfills the design principles taught in class. This dashboard is then evaluated against Shneidermann's Information Seeking and Visualization mantras and Tufte's graphical excellence and integrity principles. Lastly, the report sheds light on the future scope of this project, on how this will enhance the end-user experience.

Link to the final dashboard: [Cardiovascular Health](#)

2. Introduction

Cardiovascular Diseases (CVDs) stand among the leading causes of death globally [1]. CVDs are defined as a condition where the heart's structure and function are affected. It is a lifestyle disease and can be prevented by addressing behavioral risk factors [1]. Given the fast-paced changes to the lifestyle of people in the modern era, we can observe an accelerated growth rate of susceptibility to various health ailments. CVDs are non-communicable diseases that can be prevented if behavioral risk factors are addressed. According to the Centers for Disease Control and Prevention, in the United States, one person dies from cardiovascular disease every 36 seconds [3]. It is also the leading cause of death in the United States [4]. Although the specific cause of CVD is unknown, several factors can raise your chances of developing it. These are referred to as "risk factors." The more risk factors you have, the more likely you will develop a CVD. Some of these factors include but are not limited to high blood pressure, smoking, high cholesterol, diabetes, inactivity, and obesity [5]. The sooner a CVD is diagnosed, the easier it is to treat it.

Keeping the above in mind, we did a literature survey on the risk factors of developing a CVD and looked for datasets with the relevant risk parameters. While conducting our user research about the presence of CVD, we realized that many people were unaware of their blood pressure and hadn't got their glucose and cholesterol levels checked. This raised a concern to us as these health parameters played a vital role in determining CVD presence. We visioned to create a self-awareness dashboard for general folk and medical students that would be used to understand the impact of various risk factors on the likelihood of developing a CVD.

2.1 Objective and Scope

The Objective is to look at the historical data of various individuals with and without a CVD and the presence and levels of their risk factors to come up with a post-facto analysis of the current trend and susceptibility rate of having a CVD. Our goal is to discuss the hidden associations within the predictor variables in the data.

The outcome of this project is to create an interactive dashboard and a report that embeds a self intuitive visual story enabling the users to navigate through the key metrics and findings. We anticipate results like the users identifying the correlation between CVD and synthetic fields like Body Mass Index (BMI), blood pressure range, to name a few, quantifying the risk associated with each of the features in our dataset and understanding the ones which have a significant contribution towards a CVD.

3. Design Process

The design process adopted for this project was inspired by Ben Fry's Model of Information Visualization Design Process [13] and Cooper et al.'s goal-oriented design [14]. The first steps involved coming up with a topic and addressing the questions that the data visualization will be answering. This was followed by conducting a literature survey and research on the subject matter to bolster our hypothesis further. The relevant dataset was acquired and parsed in parallel. The parsing involved performing EDA, data cleaning, and this was followed by visually ideating the design on paper. The Five Design Sheets (FDS) and development methods were adopted from a conventional product design approach. Since those were the significant steps in the process, they are explained in detail in the following sections. We then followed that up by building the low-fi prototype in Tableau. Consequently, we refined the dashboard by iterating it from user feedback inputs in over three rounds of usability testing and finalized the dashboard to host it.

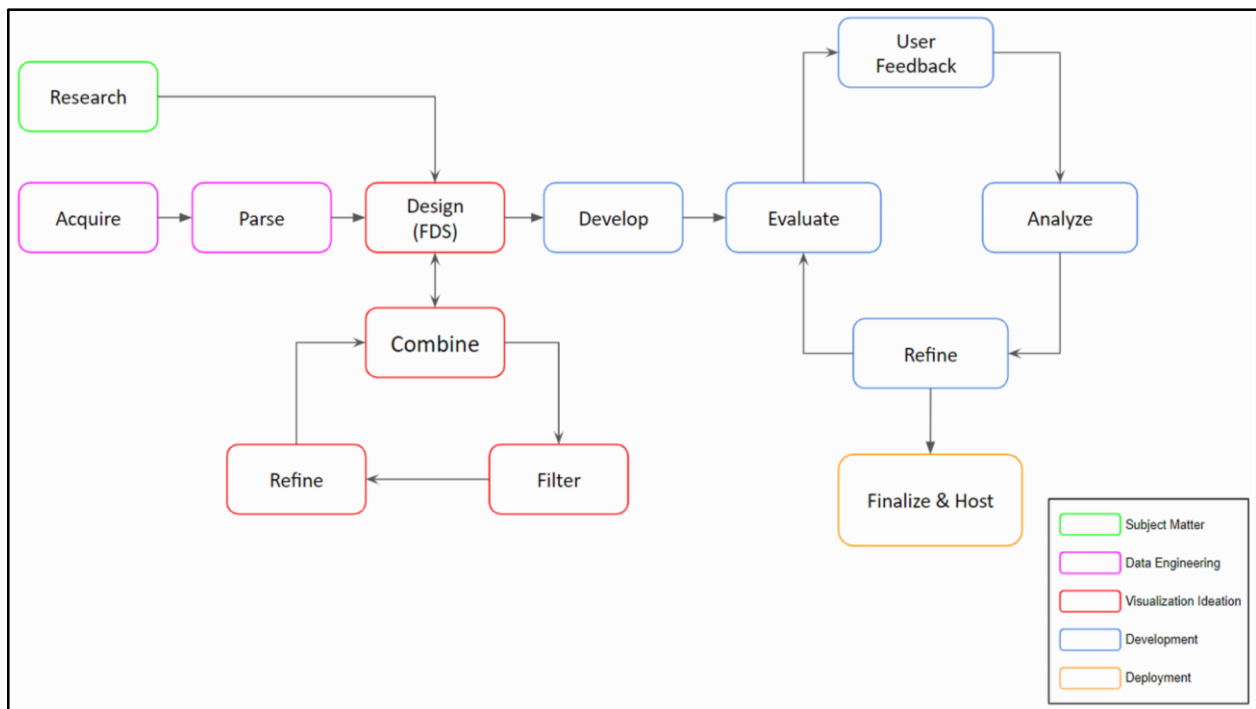


Figure 3.1: Design Process for the Cardiovascular Dashboard

4. Background Research

4.1 Existing Survey Review and Design Questions

During the preliminary research, the team reviewed cardiovascular-related articles by WHO [6], PubMed [7], CDC [8], and a paper by the Journal of American College of Cardiology [9]. Based on the information gathered from these sources, the team came up with the following hypotheses, which served as a baseline for the design process:

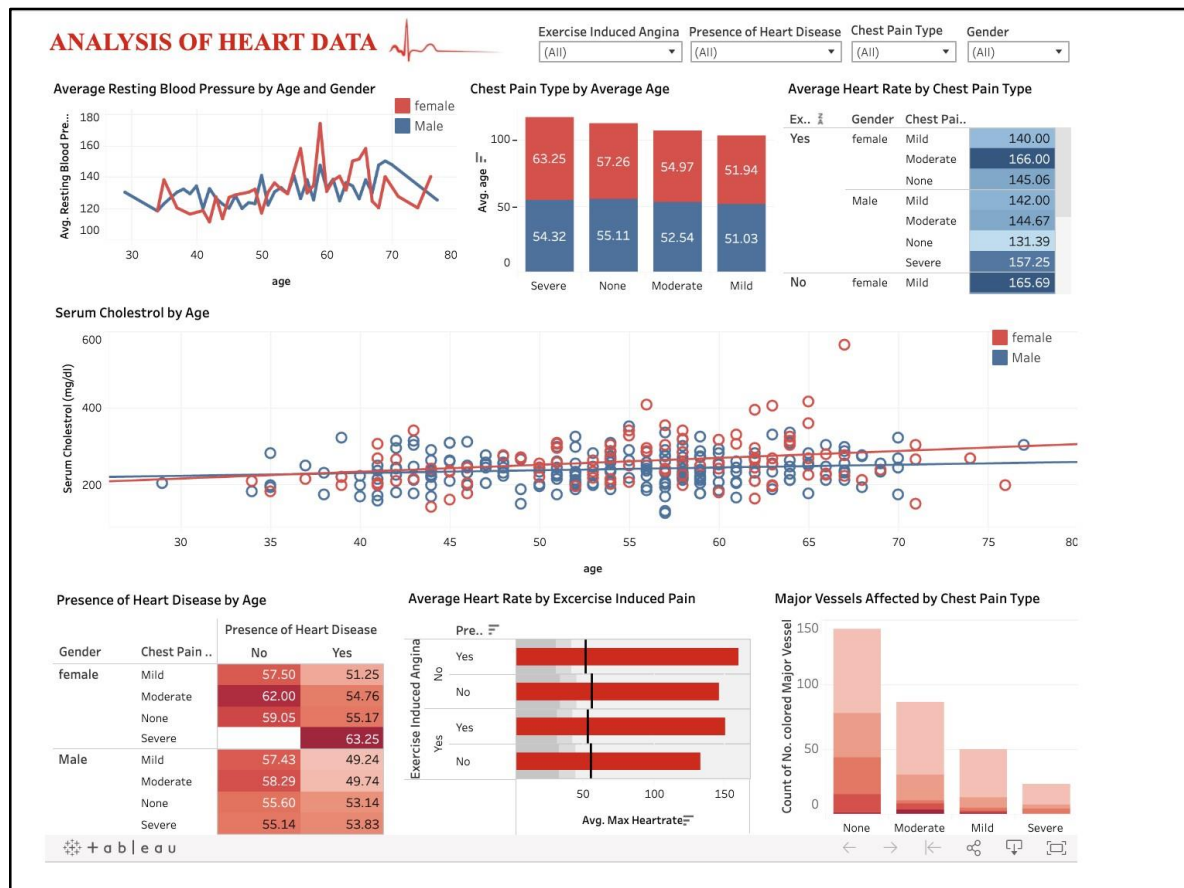
1. Adverse effects of behavioral risk factors (contributing to CVD development)
2. Correlation between obesity and lack of physical exercise (contributing to CVD development)
3. Advantages of managing Hypertension, glucose levels

4.2 Competitive Analysis

As there were no dashboards that used the same dataset, we looked at previous work based on similar behavioral risk factors and parameters related to our dataset. We listed down the pros and cons of these dashboards to better understand what our final dashboard could look like. This evaluation was unbiased and helped us gain significant insights. Some of these visualizations helped us improve the ways of representing population data, addiction themes, selection of color palette, and the placement of the filter options.

4.2.1 Visualization 1: Analysis of Heart Data ([Link](#))

This visualization was built using Tableau. The dashboard shows various parameters like average resting blood pressure, chest pain, and exercise-induced pain and maps these variables against cardiac consequences. Although the dashboard gave us an insight into how consistent color schemes make the dashboard aesthetically pleasing, it helped us critically assess certain design decisions that would be better for the end-user.

Figure 4.1: [Analysis of Heart Data](#)**Pros:**

- The dashboard has a consistent color scheme making it visually appealing, one of the primary aims of any visual design [21].
- It incorporates Shneiderman's mantra of information visualization [18] to a certain extent supporting features like filtering, zooming, panning, and refining.
- Since line graphs rank high for continuous data on Mackinley's effectiveness ranking [22], the visualization seems to obey them.
- The scatterplot is fitted with a line to convey the trend to improve the understanding of the end-user.

Cons:

- Adding a comprehensive title to the dashboard could help better comprehend the story.
- A lot of information is shown at a single glance, without annotations to guide the user.
- If the goal of the 'major vessels affected by chest pain type' visualization is to aid comparison, stacked bars would probably not be the best representation.

- Representation of information in tabular format requires excessive cognitive effort and might not be suitable for end-users who want to see the trends and patterns instead of the underlying numbers.
- Unordered labels in the table and the X-axis graphs do not follow the principle of importance ordering [23].
- There is an inconsistency in using color to denote variables like male and female, thus making it hard for the end-user to interpret.
- The alignment of visualizations does not convey a good story.

4.2.2 Visualization 2: Heart Disease Dashboard ([Link](#))

This dashboard displays the primary cause of heart diseases and depicts them as per age and gender. This chart helped us develop the idea of feature importance to be one of the visualizations we would like to show. It also helped us reduce our color scheme as multiple colors can be confusing to digest from an end-user perspective.

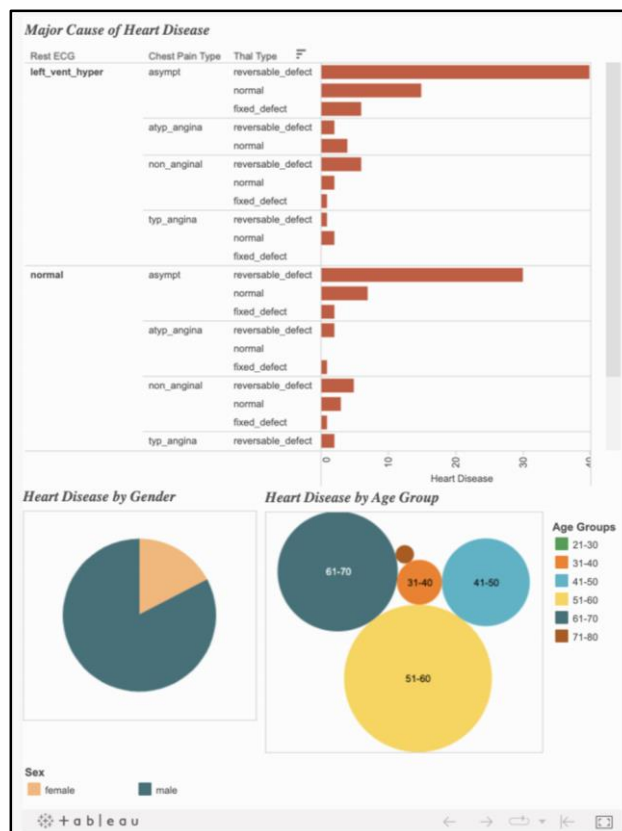


Figure 4.2: [Heart Disease Dashboard](#)

Pros:

- Conveys the information succinctly with simple visualizations to aid the user's understanding.
- The use of horizontal bar charts aids in human cognition

Cons:

- The variables have not been renamed to make it intuitive for the user to understand.
- The conveyed story is hard to comprehend and leaves the end user guessing.
- The use of multiple colors can be confusing to the end-user.
- The "heart disease by age group" visualization could have been a bar chart as the current form makes it hard to compare as the human cognition finds it hard to compare areas.
- This dashboard does not obey the Shneiderman mantra [18], as the zooming and filtering options are missing.
- The red color is not the best choice for inclusivity [17].
- The pie chart used to denote the female and male distribution isn't normalized or labeled well on the chart.

4.2.3 Summary

With this task of background research and competitive analysis, we laid out the pros and cons for the preceding dashboards, which helped us better understand the different subcategories in the vast topic of heart diseases that were already in discussion among the general public. Counterintuitively this also helped us identify the key concepts/hypothesis that is yet to be analyzed, like contributions of an individual's blood pressure, cholesterol, glucose level, their physicality, their lifestyle choices, and their activity quotient in developing a CVD and understanding the different types of these feature combinations that corresponds to lower/higher development probability. This hypothesis formed a good baseline for our design ideation process and helped us zero in on the data source that would best support the pertinent analysis.

5. Dataset Processing and EDA

5.1 Data set

To fulfill the pre-defined goals, our visual story is built upon the "Cardiovascular Disease" dataset from Kaggle [11], a subsidiary of Google LLC that publishes reliable datasets for user exploration and model building. This "Cardiovascular Disease" dataset consists of medical observations of 70,000 subjects and 12 features that list down the key indicators of each subject's lifestyle. These features are further classified into three types, namely, Objective (representing factual information), Examination (describing results of medical examinations), and Subjective (representing information provided by the subjects)

Sl No.	Feature Name	Feature Type	Data Type
1	Age	Objective	Quantitative
2	Height	Objective	Quantitative
3	Weight	Objective	Quantitative
4	Gender	Objective	Ordinal
5	Systolic Blood Pressure	Examination	Quantitative
6	Diastolic Blood Pressure	Examination	Quantitative
7	Cholesterol	Examination	Ordinal
8	Glucose	Examination	Ordinal
9	Smoking	Subjective	Ordinal
10	Alcohol Intake	Subjective	Ordinal
11	Physical Activity	Subjective	Ordinal
12	Presence of CVD	Examination	Ordinal

Table 5.1: Dataset Parameters and their Types

5.2 Data Cleaning and Preparation process

The cleaning process and data preparation were executed using Tableau Prep and Alteryx. Following modifications were carried out to aid the analysis and derive more insights:

- How susceptible is the person to have a CVD?
- What's the addiction theme of the data?
- What are the CVD rates for different glucose and cholesterol ranges?
- Using Systolic blood pressure and Diastolic blood pressure to determine if a person is hypotensive/hypertensive or not

5.2.1: Visual Inspection of the dataset

To understand the spread of the data corresponding to features in the original data set, multiple box plots were plotted using Tableau. This helped identify apparent outliers in the data sets, which were removed using Tableau Prep.

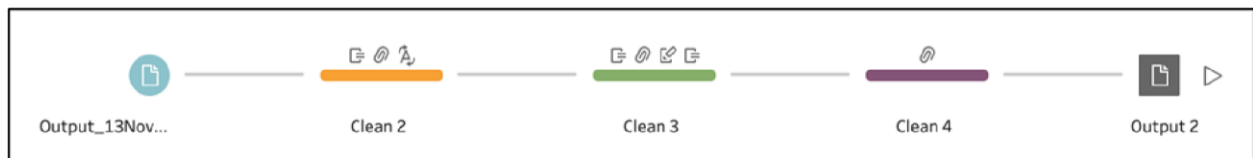


Figure 5.1: Snapshot of Tableau Prep Cleaning Workflow

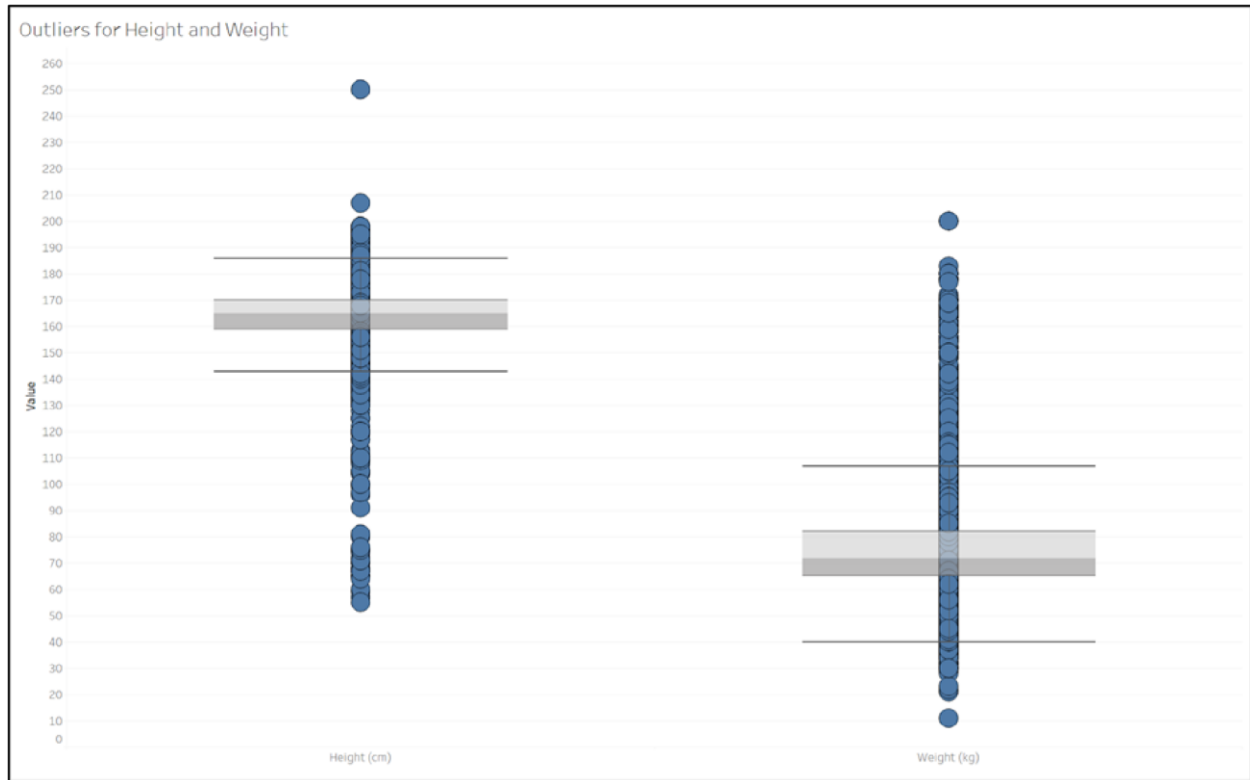


Figure 5.2: Box plot created using Tableau indicating outliers in height and weight. Some data points indicate patients with height and weight as low as 64.6cm and 11kgs, respectively.

5.2.2: Transforming fields to numerics and removing unused features

- Conversion of age in days to years
- Conversion of height in centimeters to meters
- Conversion from numerical to string value indicators for features representing alcohol consumption, physical activity, and smoking activity
- Removal of non-essential columns like Row ID and renaming the indicators according to relevance and their units of measurement

Note: Since clinical data is sensitive, we did not enter missing values or modify outliers mathematically to preserve integrity.

5.2.3: Generating synthetic fields

We leveraged Alteryx to generate the below synthetic fields and bring a tabular structure to the association rules for our final dataset.

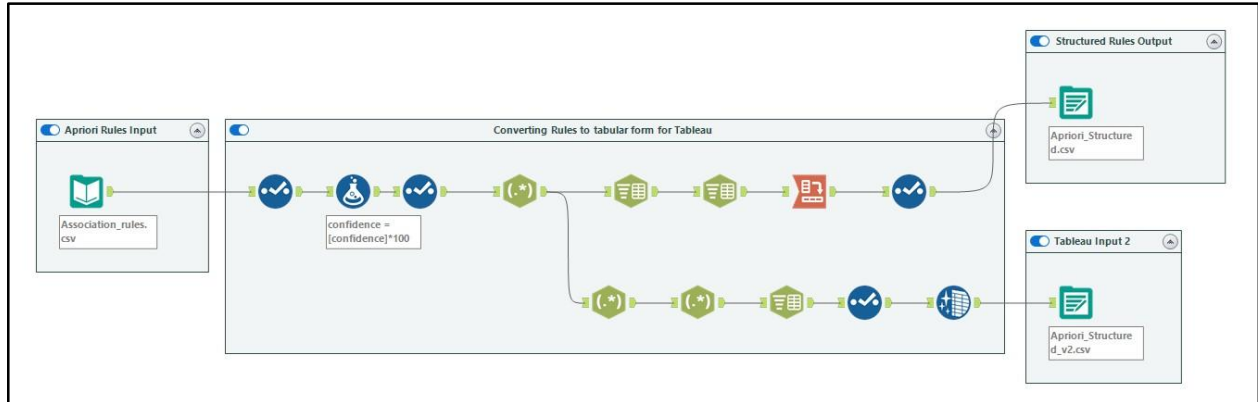


Figure 5.3: Use of Alteryx to manipulate data and generate synthetic fields for the final data set

The following synthetic fields were generated:

1) Body Mass Index (BMI) = Weight / Height² and BMI Category

Category	BMI Range (kg/sq.m)	Reference value assigned
Underweight	< 18.5	-1
Normal	18.5 - 24.9	0
Overweight	25.0 - 29.9	1
Obese	>= 30	2

Table 5.2: BMI Feature Binning [14]

2) Blood Pressure Category

Category	Systolic BP (mmHg)	Condition	Diastolic BP (mmHg)
Normal	< 120	and	< 80
Elevated	120 - 129	and	< 80
Hypertension- Stage 1	130 - 139	or	80 - 89
Hypertension- Stage 2	140 - 180	or	90 - 120
Hypertension- Stage 3	> 180	and/or	> 120

Table 5.3: BP Category Binning [9]

3) Glucose Category

Category	Range (mg/dL)	Reference value assigned
Normal	< 117	1
Above Normal	117 - 137	2
High	> 137	3

Table 5.4: Glucose Feature Binning [9]

4) Cholestrol Category

Category	Range (mg/dL)	Reference value assigned
Normal	< 200	1
Above Normal	200 - 239	2
High	>= 240	3

Table 5.5: Cholesterol Feature Binning [9]

5.2.4 Decision Trees

Along with the traditional EDA techniques, we incorporated specific machine learning algorithms to mine patterns in the data. Decision trees [24] and the Apriori [25] are popular algorithms widely adopted in various pattern mining and machine learning fields. A modified version of the

decision tree, where we have altered the model's hyperparameters to maximize the probability of finding meaningful and interpretable rules. The modifications include limiting the tree depth to 10, limiting the maximum features in a decision cell to 1, and thresholding the samples at each cell to be at least 0.2 factors of the complete dataset. We are not building a predictive model, so we have used the entire dataset to train the model. The `plot_tree` feature available in Matplotlib is used to visualize the decision tree and the rules and sample counts at each cell. From Fig. 5.4, we can see that the desired rules and sample sizes are color-coded and plotted. Orange represents class 0 (Absence of CVD), and Blue represents class 1 (Presence of CVD). Higher the intensity, the greater the confidence of the respective class. The age feature at the root cell seems to be the high impact rule, where subjects above 55 are more prone to CVD. Similarly, we parsed the tree and collected the rules and the sample count for each rule, and used these to provide users with the most common patterns that cause CVD based on the current dataset.

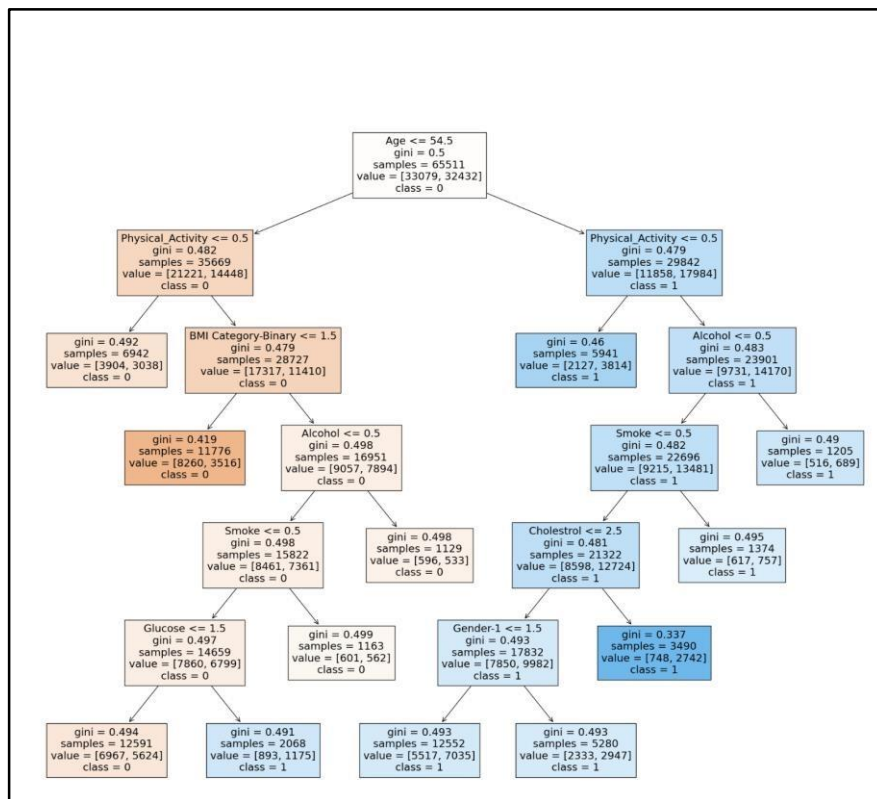


Figure 5.4: Decision Tree output for Feature Importance

Note: This data cleaning process resulted in the removal of 6.41% of data from the original dataset, i.e., the final version has data for 65,513 subjects as opposed to 70,000 subjects in the original.

6. Design Ideation

6.1 Five Design Sheets

We used the five design sheets method to develop the initial design for our first prototype. All the team members came up with multiple ideas for the dashboard during the initial brainstorming session. We discussed each of these ideas and came up with viable options among them. From each of these, we identified common findings and transformations like combining weight and height to derive a synthetic field, BMI. This process enabled us to get a holistic view and a detailed understanding of the data, which further helped with a plan for our final design. Combining and filtering these ideas, we tailored three main concepts to describe our visual story. The first idea was to create an intuitive infographic representing physical activity and addiction themes. The second one was to develop summary visuals that allow users to understand the rest of the data profile better. The third idea was to depict visualizations representing different associations between high-impact features and CVD. We did a deep dive into the visual representations and potential interactions for each of these ideas. These illustrations were further refined for better visual communication. We explored the possibilities of incorporating various filters and interactions and came up with a substantial layout for our visualization. These were evaluated further concerning technical feasibility and keeping the end-user in mind. This entire process enabled us to structure our initial individual introspections into a conclusive dashboard which could eventually lead to a story (Figure 6.1 below).

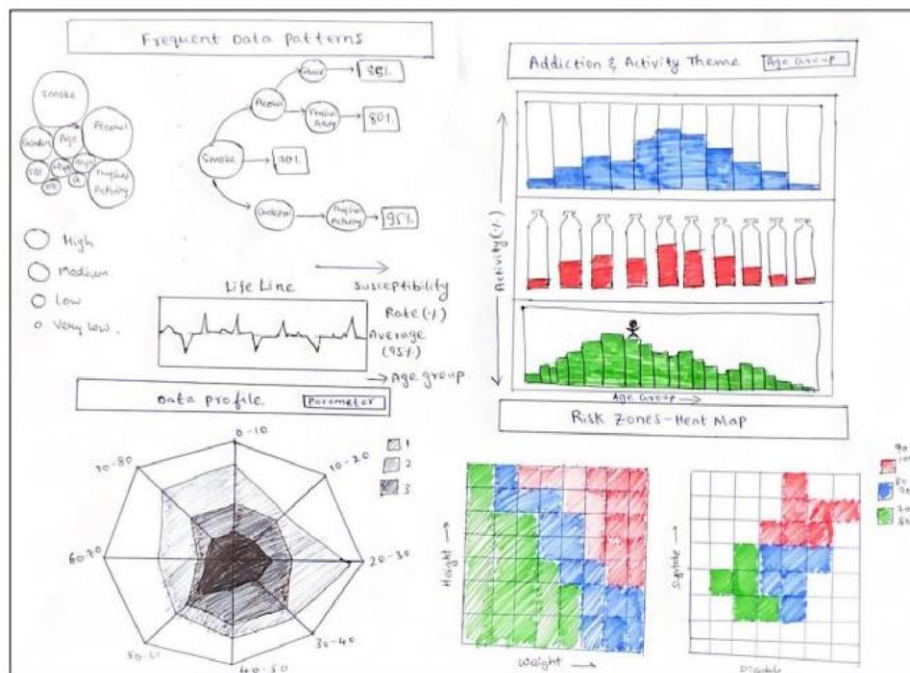


Figure 6.1: Final Design from Five Design Sheets

7. Prototype Development

Our initial prototype revolved around creating graphs similar to what we envisioned in our FDS (Figure 6.1). Post this, we collected feedback from users and refined our dashboard. Our first low-fidelity prototype was created using Tableau Desktop. We followed an iterative process to arrive at our final dashboard. The iterations were multiple, but for the brevity of the document, we stuck to focusing on the significant ones who helped reach the final dashboard.

7.1 Prototype Round 1

Our first prototype was created using Tableau Desktop and hosted on Tableau Public (Fig[7.6]). We primarily focused on showing five layouts: data profile, smoking activity, alcohol activity, physical activity, BMI vs. CVD rate, and data profile.

1) Smoking Activity

This bar chart displayed under the "Activity theme" depicts the percentage demographic who had the behavioral trait of smoking and CVD. The white block in the bar chart showed the defined percentage smoking rate. The orange block was added as a dummy value to resemble a cigarette for aesthetic purposes. Referring to Bertin's vocabulary of retinal variables [23], we thought the bar chart would be easy to interpret and went ahead with the same. During the first round of user interviews, some users liked the design element as it was intuitive, while others confused it as a stacked bar chart. A user also pointed out that of the subjects who have CVD in the age group of 26-30, the bar graph showed a 0% of the smoking rate, which was a data cleaning issue. We resolved this by inspecting the data again and removing this particular age bucket.

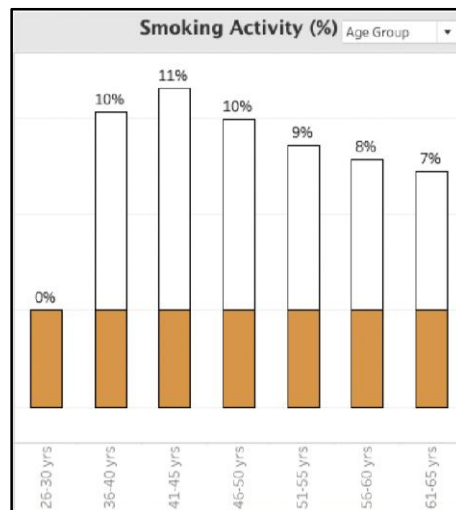


Figure 7.1: Bar graph representing smoking activity across values of a feature

2) Alcohol Activity

This bar chart displayed under the "Activity theme" depicts the percentage demographic who had the behavioral trait of drinking and CVD. We chose a bar chart again to help us with the aesthetic of having a bottle and showing the rise and fall in alcohol levels.

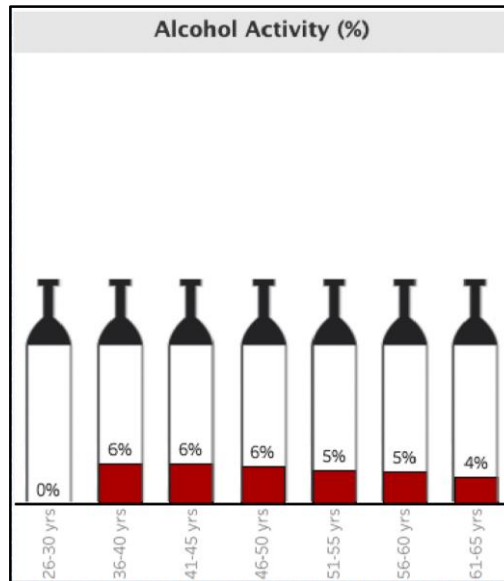


Figure 7.2: Bar graph representing alcohol consumption across values of a feature

3) Physical Activity

The bar chart depicted the rate of physically active people, and the line graph was introduced to see any trends in the same data for different filtered options. However, all the users gave collective feedback during our user feedback that this data seemed spurious, and no real insight was depicted. The change was negligible, and we decided to remove this data field because it gave no additional insight.

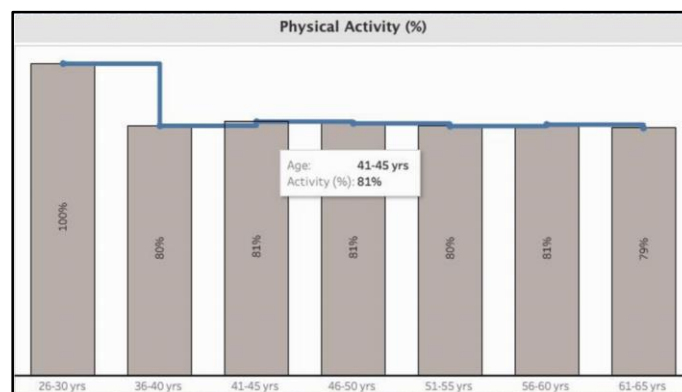


Figure 7.3: Bar graph representing physical activity across values of a feature

4) BMI vs. CVD rate(%)

The heatmap depicts the percentage of the subjects who had CVD for a given age group and blood pressure category. Since this involved multivariate visualization, we used heat maps to show the trend. During user feedback, it was pointed out that there was barely any data trending for the age group of 26-30. We inspected our data and realized there was a lack of data points for this age group for various blood pressure ranges and decided to drop this age group.

BMI Vs CVD Rate (%)											
Overweight/Obesity (BMI > 24.9 kg/m2)						Underweight (BMI < 18.5 kg/m2)					
Age	Low Blood Pressure	Prehypertension	Hyper Stage 1	Hyper Stage 2	Hypertensive Crisis	Age	Low Blood Pressure	Prehypertension	Hyper Stage 1	Hyper Stage 2	Hypertensive Crisis
61-65 yrs	66%	67%	85%	85%	84%	61-65 yrs	37%	100%	83%		
56-60 yrs	50%	62%	83%	86%	85%	56-60 yrs	36%	67%	89%	100%	100%
51-55 yrs	42%	58%	83%	90%	85%	51-55 yrs	19%		38%	100%	
46-50 yrs	35%	57%	84%	91%	89%	46-50 yrs	23%	13%	80%		100%
41-45 yrs	27%	55%	84%	83%	84%	41-45 yrs	10%	100%	33%	100%	50%
36-40 yrs	18%	53%	79%	91%	92%	36-40 yrs	5%	0%	0%	100%	
26-30 yrs	0%										

Figure 7.4: Heat map representing CVD rate for each BMI category

5) Data Profile

The data profile shows the number of people with a CVD for various features chosen in the filter for a given age group. Since we wanted the age group to remain static and wanted the user to play with other parameters to see how they are spread across age groups, we decided to go with a radar chart, as radar charts are a good choice when comparing two or more groups (in our case Age groups) against different features like Glucose or cholesterol. As the patterns for each category of a particular feature (for example, glucose levels) slightly overlapped, the transparent shading of the category and sorting of the CVD rates for each age group aided in improving the visual display for the end-user. We were also curious to see if non-technical users could decipher the same. During our user research, we were told that the scale of the chart and the size of the polygon were not in line and that the age groups of 26-30 and 31-35 were flat for several categories. Hence, we decided to drop all data points across categories for these age buckets and focus on data from the 36-40 bucket.



Figure 7.5: Radar chart showing the split of subjects on multiple features and their CVD rates

The final prototype when put together looked as follows.

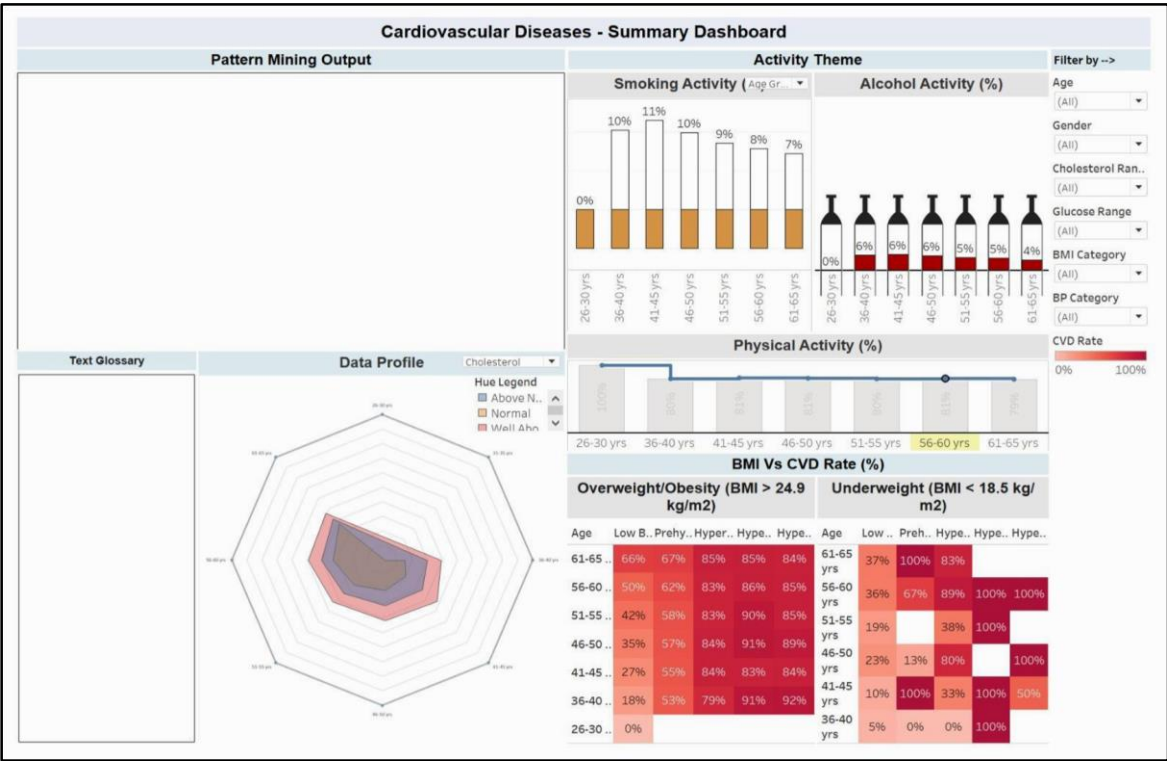


Figure 7.6: First prototype of the interactive visualization

7.1.1 Prototype Usability Testing:

After creating our first prototype, we conducted our first structured usability test. The users were classified into two groups: In-class and real-world potential users.

The Guerrilla testing was first conducted with our in-class users on November 24, 2021, with Andy Wang, Hunter Thompson, and Hriday Baghar as participants. The second usability testing was conducted in a formal setting, and this was kicked off virtually by our three real-world users - Dr. Sanskriti Jaiswal, a Graduate Medical Student at Bangalore Medical College and Research Institute, India; Mr. Sanjeev Sharma – Life Insurance Corporation Employee, and Ms. Karishma Mehta, a Biomedical Research Student at Stony Brook University, Long Island on November 29, 2021.

In both testing sets, the team identified an observer, facilitator, and note-taker. The facilitator was available to answer any questions, but they abstained from any unsolicited prompting. All users were asked to complete a series of tasks that we generated from our understanding of the use cases of the personas. The list of tasks was:

1. Does the user understand the overall story of the visualization?
2. Does the visualization clearly express the effect of addiction and physical activity on CVD in different age groups and other parameters given in the filter options?
3. Navigate through the heat maps represented and give feedback on their understanding.
4. Navigate through the data profile to understand how intuitive it is.

To complete these tasks, the in-class and real-world users were each given 10 mins and 20 mins respectively. After completing these tasks, we asked the users about their thoughts on each visual. The following are some of the questions posed to the users:

1. Were you able to clearly understand what each visual is trying to convey?
2. Is any of the information presented too overwhelming? If yes, do you have ideas for alternatives to the existing visuals?

Refer to tables 7.1 and 7.2 below for user feedback

In-Class User	Positive Feedback	Suggestions/Highlighted Issues
Andy Wang	<ol style="list-style-type: none"> 1. The user liked the data profile feature with the radar chart 2. Found the muted colors in the radar to be aesthetically pleasing 	<ol style="list-style-type: none"> 1. Requested on-click filtering for all the visualizations to make it more interactive 2. Similar filter placement in the smoking theme visual drop down and at the top right corner created confusion while selecting user personas
Hunter Thompson	<ol style="list-style-type: none"> 1. The user liked the overall view and placement of the visuals 2. Appreciated the smoking and alcohol consumption activity visuals represented as cigarette buds and alcohol bottles. 	<ol style="list-style-type: none"> 1. Got confused with the addiction theme visuals and assumed the indicated percentages to be CVD rate instead of demographic percentage 2. Pointed out the necessity of color consistency throughout all visuals 3. Requested improvements to the radar charts to get an easier read on the CVD rates for any highlighted category
Hriday Baghar	<ol style="list-style-type: none"> 1. Liked the addiction pattern visuals created and found the selection of heatmaps for representing BMI graphs to be apt 	<ol style="list-style-type: none"> 1. Requested descriptive titles for all visuals to understand the dashboard better. 2. Requested for more information when hovering over the radar maps. 3. Suggested the addition of Normal BMI range data in the heatmaps to have an ideal frame of reference when perceiving the data

Table 7.1: In-class Guerrilla Testing Feedback

Real-World User	Positive Feedback	Suggestions/Highlighted Issues
Dr. Sanskriti Jaiswal	1. Once familiar with the Tableau public interface, the user was able to navigate and appreciated the filter option.	1. Required assistance to navigate through the dashboard and was not able to easily comprehend the radar map. 2. Requested the use of more legends and intuitive titles for better understanding
Mr. Sanjeev Sharma	1. Found the visuals to be helpful to quote relative premiums to his customers' based on addiction data and cardiovascular health background	1. Requested for dedicated space in the dashboard for data summary of highlighted parameters for better understanding of the scope by the folks in his domain 2. Requested reset button for filter reset to improve user experience
Ms. Karishma Mehta	1. The user was appreciative of the overall concept and the prototype presented was well received 2. They especially liked the concept of using heat maps to represent BP and BMI graphs	1. Questioned the purpose of the line graph used over the physical activity bar graph 2. Found the labels in the legends of the radar graph to be confusing and requested more details on the same 3. Due to the depiction of the cigarette buds in the smoking activity chart, the visual was incorrectly perceived as a stacked bar graph

Table 7.2: Real-World Usability Testing Feedback

7.2 Prototype Round 2

This was created using Tableau Desktop and hosted on Tableau Public (Figure 7.3). We incorporated several changes post our usability testing feedback. We dropped the physical activity chart and finished working on our decision tree visualization. We incorporated all the changes we got from our users in round 1. We also merged alcohol activity and smoking activity under activity themes for CVD.

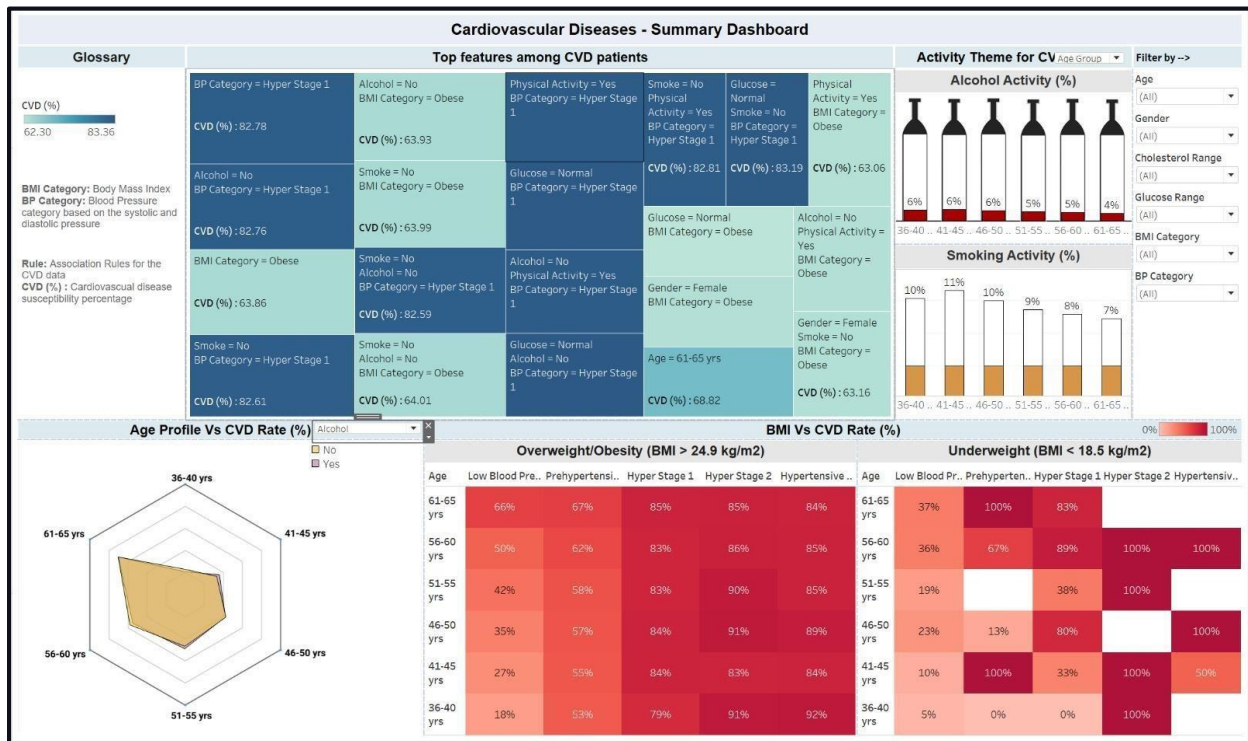


Figure 7.7: Second prototype of the interactive visualization

1) Activity Theme

Based on the feedback received, we realized that the design element could misinterpret the true values. Hence, we decided to drop the aesthetics and go with simple bar charts for our final visualization.

2) BMI vs. CVD rate (%)

We got several constructive feedback for this heatmap. We hadn't shown an option of normal BP range, which we later realized is necessary to be used as a standard for comparing against BMIs for overweight and underweight subjects. A couple of users found the red color to be bright. To also be inclusive of color

vision deficient individuals, we decided to change the color palette to a softer shade and dropped the legend as it was self-intuitive.

3) Age Profile vs. CVD rate (%)

We noticed that only the cholesterol and glucose level variation gave significant information on inspecting the data during this iteration. Thus we decided to drop other parameters, namely gender, an activity that corresponds to 2 categories (with binary indicators), and sticks with only Glucose and cholesterol (corresponding to 3 categories each) in the radar chart depiction.

4) Top features among CVD patients

We chose a decision tree to represent the top features as we used the apriori algorithm to understand the top trends of CVD-positive subjects. However, the users were overwhelmed looking at this chart. There was a lot of text, and all were of similar font and size, which didn't help guide the user. We also received feedback to maintain a similar color scheme throughout the dashboard. We showed this chart to technical and non-technical users to understand how we can improvise. In our final visualization (Figure 8.1), we played with the fonts and hues to guide the users. We highlighted the percentage of CVD positives by increasing the font and lightening the color of the text to make it less verbose at first glance. We also added an extra filter by adding the top 3 risk factors as a starting point for ease of user navigation (Figure 8.1).

7.3 Key Takeaways:

The team reviewed the feedback received from all users and identified the design changes that had to be made to address the voiced usability issues.

Data Visualization	Design recommendations
Overall Dashboard	<ol style="list-style-type: none"> 1. Create intuitive titles and utilize color/font change to indicate different parameters in each visual 2. Select a consistent color palette for the dashboard 3. Improve user navigation by dedicating a glossary space to display information necessary to understand any visual 4. Improve alignment of all visualization to resolve any spacing issues
Addiction theme visuals	<ol style="list-style-type: none"> 1. Remove the redundant filter and utilize a common filter space for all visualizations in the dashboard 2. Remove the design elements used in the graphs as it can cause inaccurate interpretation of the data
BMI vs CVD rate	<ol style="list-style-type: none"> 1. Using a single heat map (instead of 2) to represent the CVD rates for different age groups and BP categories and encoding all the BMI categories as a parameter 2. Add an intuitive title to display the BMI category selected for the respective heat map 3. Add a comprehensive report feature (in the visual/glossary) for the users who wish to detailed data of CVD rates for any/all filtered personas
Age Profile vs CVD rate	<ol style="list-style-type: none"> 1. Using two heatmaps to split the data profile representation w.r.t. cholesterol range and glucose range to increase the legibility of the data represented 2. Increase the axis scale in the radar maps (from 10% to 20%) to create concise visuals where the percentage rates can be easily comprehended with the lowest movement over the graph 3. Selection of uniform color palette across all radar maps
Top risk factors (CVD susceptibility rate%)	<ol style="list-style-type: none"> 1. Addition of a consolidated treemap in the dashboard to display the risk trend among the patient demographic for scenarios derived from the rule mining process 2. Filtering on the treemap based on the top 3 risk factors using bubble graphs to improve readability

Table 7.3: Key Takeaways from Usability Testing

8. Final Dashboard

The final dashboard incorporated a couple more changes, including changing the background to a dark theme, thus making the whole dashboard appear visually appealing. After another round of user interviews, we also changed to a more consistent theme. The design elements could potentially mislead the users; hence they were removed. We tried to keep the data-ink ratio minimal [21] and still convey the story. We utilized the top-left corner of the dashboard and moved the filters there to make the navigation more intuitive for the end-users. Some of the other touch-ups included intuitive headers, images of arrows, consistent fonts, and colors. The color scheme was made more inclusive as well [11]. Since this document cannot be interactive, the plots are vertically stacked for reading appropriately and panned and zoomed to understand the intricate details [18].



Figure 8.1: Final Dashboard Thumbnail

Cardiovascular Diseases - Detailed Report									
Filters for Report ->									
Age	Gender String	BMI Category	BP Category	Cholesterol Range	Glucose Range	Smoking	Alcohol Consumption		
61-65 yrs	Female	(All)	(Multiple values)	Normal	(All)	(All)	(All)		
Report Parameter I	Report Parameter II	Report Parameter III	Report Parameter IV	Report Parameter V	CVD Rate (%)				
Age	BMI Category	Blood Pressure	Cholesterol	Glucose					
61-65 yrs	Normal Weight	Hypertension	Normal	Above Normal	72.2%				
				Normal	83.1%				
		Prehypertension	Normal	Well Above Normal	100.0%				
				Above Normal	63.0%				
				Normal	59.3%				
	Obese	Hypertension	Normal	Well Above Normal	58.8%				
				Above Normal	73.3%				
		Prehypertension	Normal	Normal	81.8%				
				Well Above Normal	74.1%				
				Above Normal	76.5%				
	OverWeight	Hypertension	Normal	Normal	61.8%				
				Well Above Normal	73.3%				
		Prehypertension	Normal	Above Normal	68.0%				
				Normal	86.1%				
				Well Above Normal	81.3%				
	UnderWeight	Hypertension	Normal	Above Normal	64.9%				
				Normal	62.2%				
		Prehypertension	Normal	Well Above Normal	51.1%				
				Normal	100.0%				
				Normal	68.8%				

Figure 8.2: Detailed Report of the Dashboard for selected filters

8.1 Individual Visualizations

The below section talks about the individual visualizations and the information it tries to convey

1) Visual Filters and Addiction Theme Visuals:

To guide the users through the visual narrative, we provided the visual filters at the top left corner of the dashboard. The selections made in these filters reflect the changes in other visualizations. Moving to the right, the users can view the addiction theme visuals. The main purpose of this visual is to talk about the addiction theme. We depicted the smoking and drinking activity among the population who have a CVD. The bar charts are used as they are known to be rated high in Mackinlay's effectiveness ranking for plotting discrete quantitative variables [22]. We used the bar chart with cigarettes as the theme visual for smoking thus reducing the cognitive effort required by the end-users to memorize the graph [23]. The length of the cigarette bud encodes the smoking population (%)

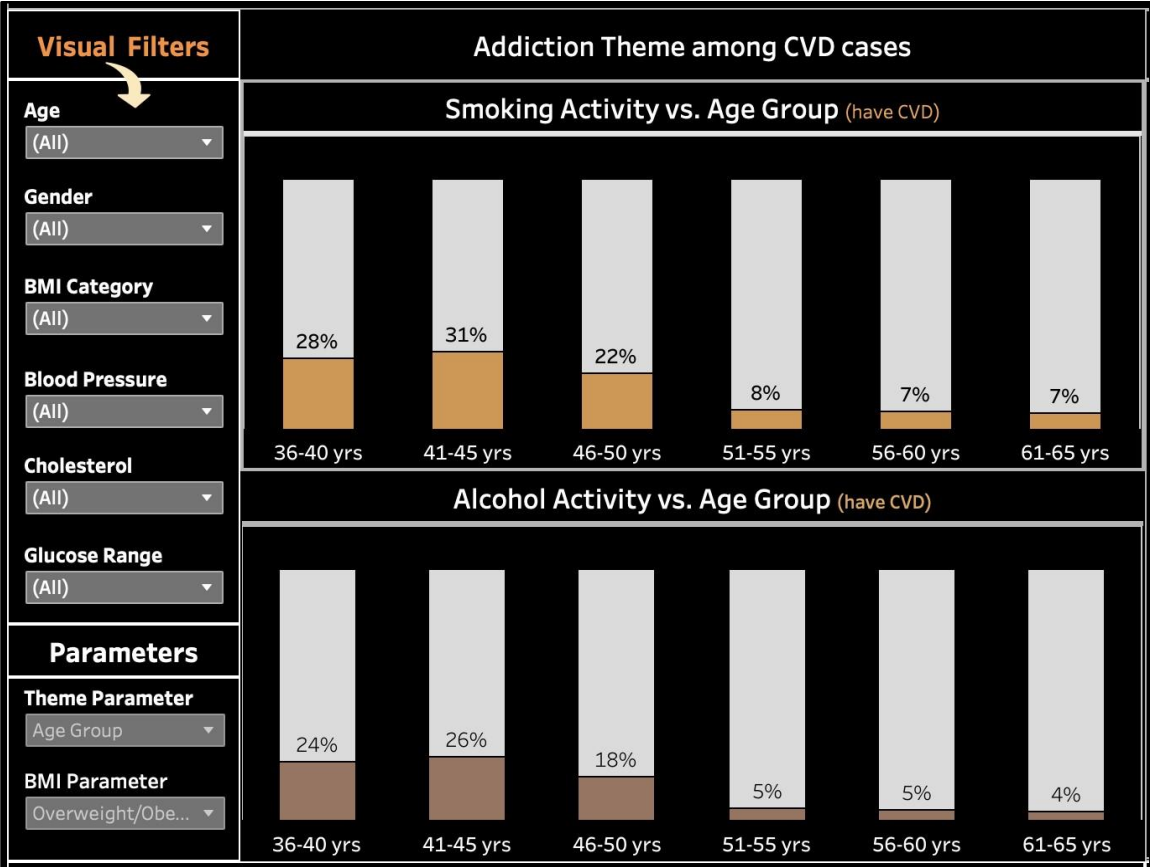


Figure 8.3: Addiction Theme Visualization

2) Summary to Detail Transition

As the prime focus for any visualization, we tried to present the crux of our analysis as a summary sheet and enabled a drill-down of information for the users. The transition can be followed by selecting the summary metrics/factors and then moving to the respective details on the right.

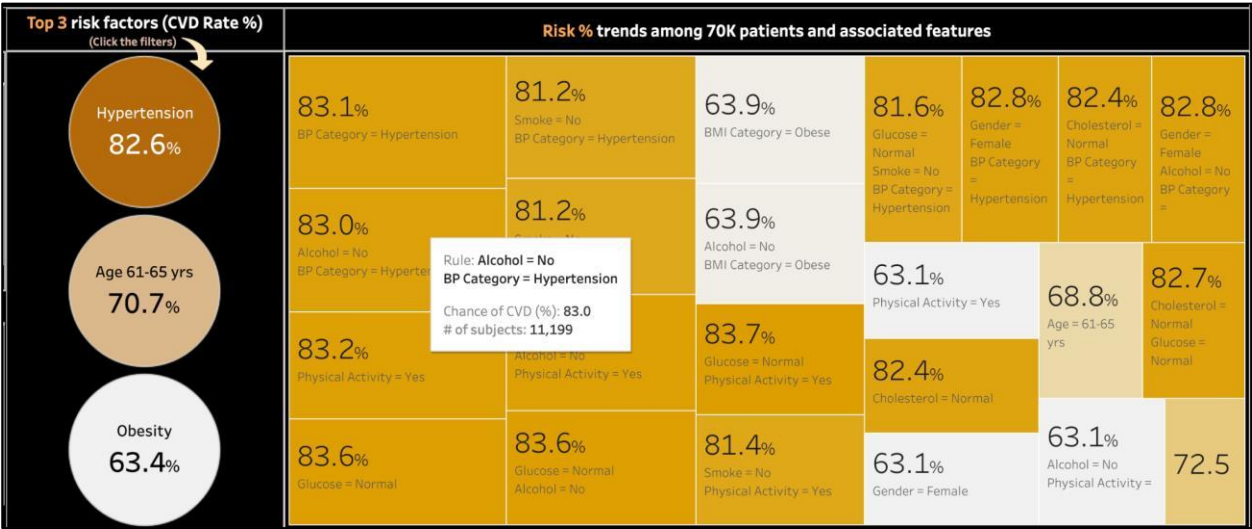


Figure 8.4: Top CVD Features - Default View

The summary visual highlights the Top 3 risk factors among CVD patients. These are the prime features observed in most of the CVD cases. The detailed visual to the right exposes other features associated with the top factors among CVD observations. The dashboard actions enable the users to select any one of the top risk factors and see the respective details to the right. We leveraged the treemap visual to effectively encode two key data properties - # of CVD cases and CVD Rate(%)

For example, subjects identified as females having Hypertension have a higher risk (82.8%) of CVD

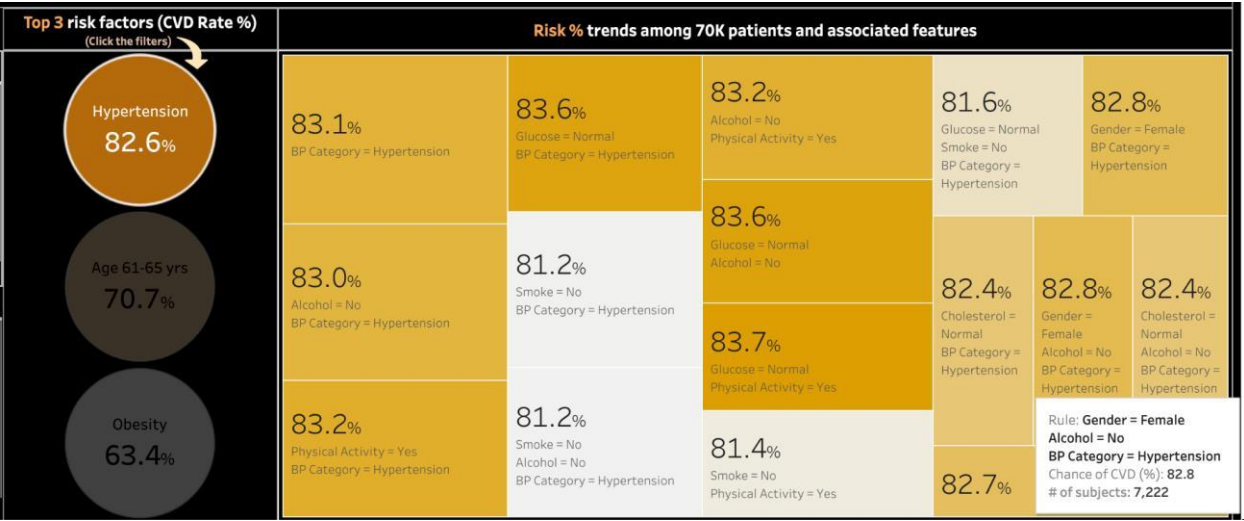


Figure 8.5: Top CVD Features - Specific Feature View

3) Age Profile Radar Chart:

Though the summary and detailed associations in the dashboard provide valuable information, it is essential to understand the general trend/spread of key health metrics like cholesterol and Glucose among various age groups. From Bertin's vocabulary for retinal variables [23], color/hue is used as a preattentive effect to differentiate between nominal variables. We leveraged the Radar chart that encodes age buckets as factors and uses hue to represent the feature classes. The size and scales are consistent for the quantitative variables with minimum lie-factor [21]. Titles are self-explainable. They also help detect patterns, fulfilling Tufte's graphical excellence and integrity principles [21].

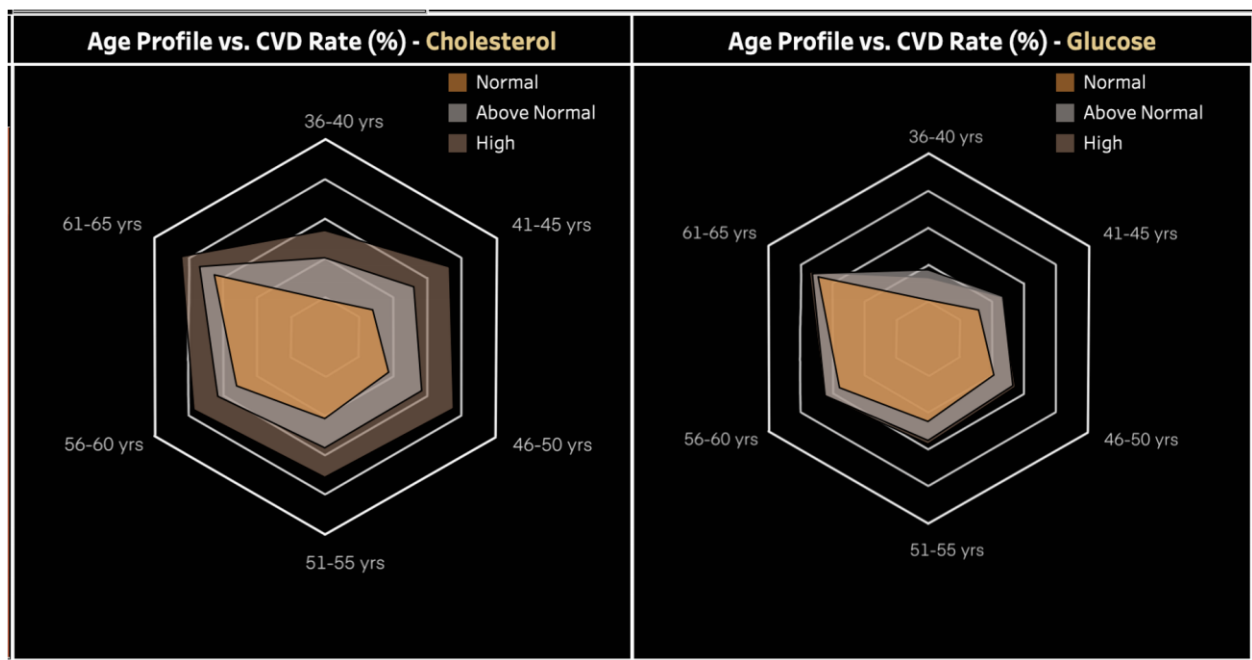


Figure 8.6: Age Profile Visualizations for Different Parameters

4) BMI Heat map

In general use, BMI charts are heat maps that highlight the respective weight categories with all combinations of height and weight. We used a similar heat map concept in this visual to show the relation of Age and Blood Pressure with CVD for each weight category of the BMI. The visual is enabled with a parameter that toggles the views for all three categories of BMI.


BMI Vs CVD Rate (%) - Overweight/Obesity				
Age 	Low Blood Pressure	Normal Blood Pressure	Prehypertension	Hypertension
36-40 yrs	25%	11%	25%	79%
41-45 yrs	5%	17%	33%	85%
46-50 yrs	35%	21%	40%	85%
51-55 yrs	44%	27%	45%	83%
56-60 yrs	60%	36%	53%	83%
61-65 yrs	62%	52%	67%	84%

Figure 8.6: Age Profile Visualizations for Different Parameters

5) Glossary

Given that the dashboard has some derived information like Blood Pressure, Glucose, and Cholesterol ranges, BMI categories, we tried to provide a glossary of information to the users and hidden details like the detailed report. We expect this to be a reference point to the users who want to know the ranges for each category.

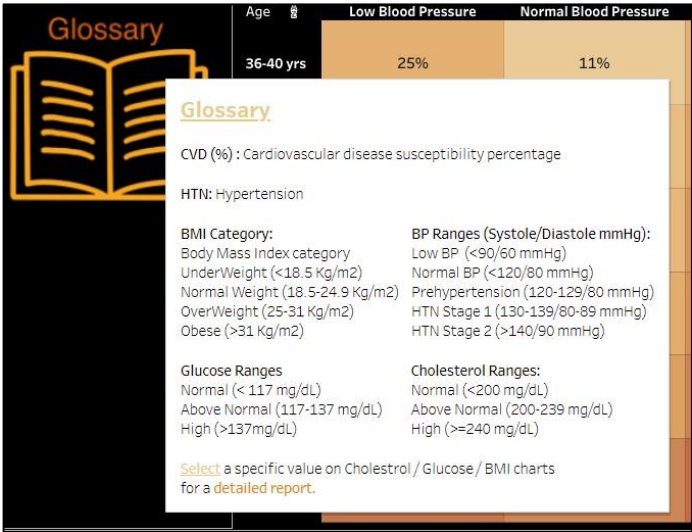


Figure 8.7: Glossary Tooltip

Table 8.1 summarizes the data encodings for all the visualizations

Visual Element	Visual Property	Data Encoded	Information
Summary	Hue	CVD Rate (%)	Visually exposes risk factors that have higher confidence
Details Treemap	Size	# of CVD cases	Highlights the most significant rules in terms of the population over the smaller ones
Details Treemap	Hue	CVD Rate (%)	Visually exposes rules that have higher confidence. The details reveal the combination of features that are tagged to the top risk factor
Age Profile Radar Chart	Hue	Cholesterol/Glucose category	It gives a comparative view between each class category of the feature
BMI Heat Map	Hue	CVD Rate (%)	Highlights risk zones with varying hue in the heat map

Table 8.1 Data Encodings for the Visualizations

Overall Key Findings:

- Hypertension is the key risk factor that is observed in 83% of people having CVD
- The age group of 61-65 yrs with Above normal cholesterol levels has a higher risk of CVD
- Obesity and Hypertension among female individuals are primary associations observed among the CVD cases
- Cholesterol and Glucose levels are other primary health concerns among people having CVD

8.2 Dashboard Properties

Overall, we keep the visualizations simple, aesthetic, yet multivariate simultaneously, thus enhancing easy pattern detection and inference operations and representing data in ways that support cognitive tasks. Hence, amplifying cognition. Since the visualization deals with a lot of information in the form of multivariate data, it is worth noting that the patterns and trends are memorable, but remembering quantitative data in a tabular format would require immense cognitive effort. We highlight principles of importance ordering when placing these visualizations in the final dashboard most effectively, thus gaining brownie points on Tufte's expressiveness and consistency definitions [21].

To better understand the efficacy of our tool, we have evaluated it in terms of Shneiderman's Information Visualization [18] and Information Seeking Mantras [18]:

- **Overview First:** We think our visualization effectively provides an overview of our entire data set. We have added the filters to the top left corner as that is the first place a user's eye would go. We have also added annotations to guide the user on filter options.
- **Zoom:** The top features visualization effectively allows users to zoom in on a particular feature they are interested in by clicking on them. The zoom is effective because the selected feature is highlighted, thereby giving the user feedback they clicked it and keeping the rest of the visualizations visible to maintain a sense of context for the user.
- **Filter:** Our addition, BMI heatmap, and data profile charts provide many options for users to filter for the most relevant data. Filters include Age group, Gender, BP category, amongst others.
- **Details-on-demand:** We provide details on demand in the first top-features graph. On top of it, if the user clicks on any of the blocks on the BMI heatmap, we give it an option to check all the values in a tabular format.
- **History:** Our visualization minimally meets Schneiderman's criteria of history. Users can "undo" the previous action by clicking the undo button, but replay and progressive refinement are not supported. We attribute this to the limitations of using Tableau rather than an intentional design decision on our part.
- **Extract:** Our visualization supports users extracting data using Tableau's "share" function. This allows users to save or share a view they made within the visualization with other users. Users can also easily embed the visualizations to external web pages through either its corresponding URL or automatically generated code.
- **Hover and click:** We allow users to hover and click on the charts, get further information, and pan and zoom on the specific required details.

9. Future Scope

We built a dashboard that conveys the top CVD risk factors and guides the user about how specific lifestyle and physical changes could potentially increase the risk of developing a CVD. Building on the current version, we would like to develop a robust prognosis model further using more features on top of the current dataset. This would ideally identify the risk of developing a CVD at an early stage. We would like to overcome any biases induced by this dataset and test how our dashboard fairs against data from diverse sources. Ultimately, combining this dashboard with a prognosis model would make it more efficient. It would help the general populace, and the healthcare workers get a holistic and patient-specific view to suggest preemptive measures to avoid further complications. Although the current dashboard tries to incorporate most of the visualization principles learned in class, from Bertin's vocabulary to Tufte principles of excellence and integrity, we would still like to include other design principles that we inadvertently missed out on.

10. Acknowledgments

We learned a lot about the different techniques and design principles to create accurate and relevant visualizations throughout this quarter. We want to acknowledge Prof. Nathan Mannheimer for his teachings and consistent guidance throughout the quarter. The topics and concepts covered in our readings have bolstered our conceptual understanding of visualization and human-centered design. The constructive feedback from our Teaching Assistants Andrea and Apoorv has also helped us rethink some of our design decisions that further guided our iterations. Lastly, we'd like to thank our classmates and our real-world users for their patience and inputs with the initial dashboard during the guerilla testing, without whom we could not have managed to come up with a comprehensive and useful visualization.

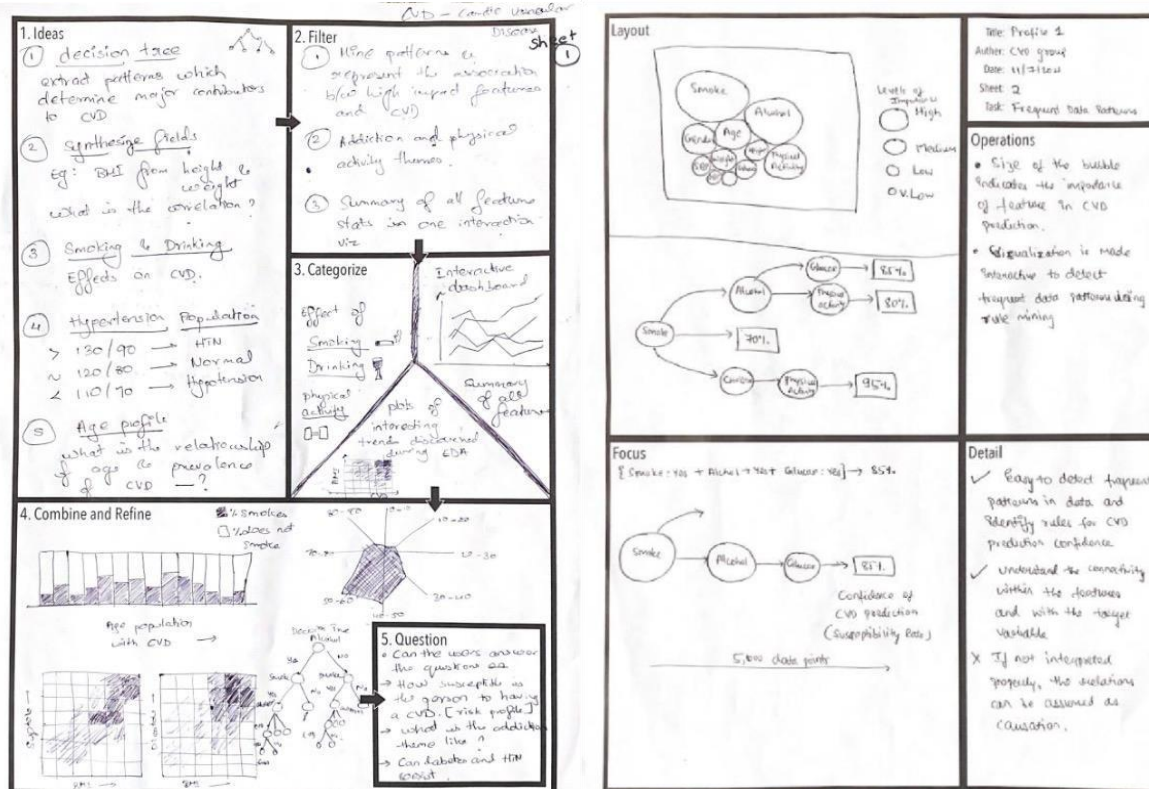
11. References

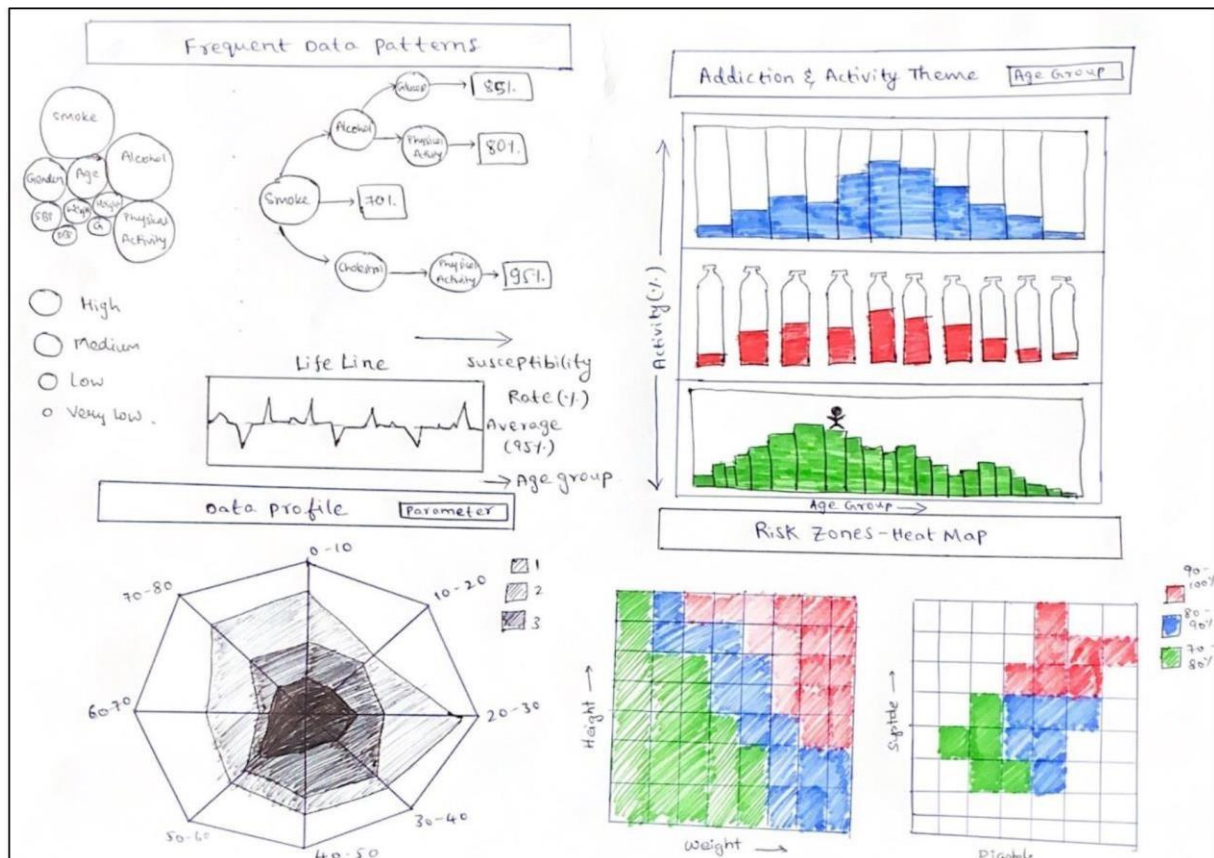
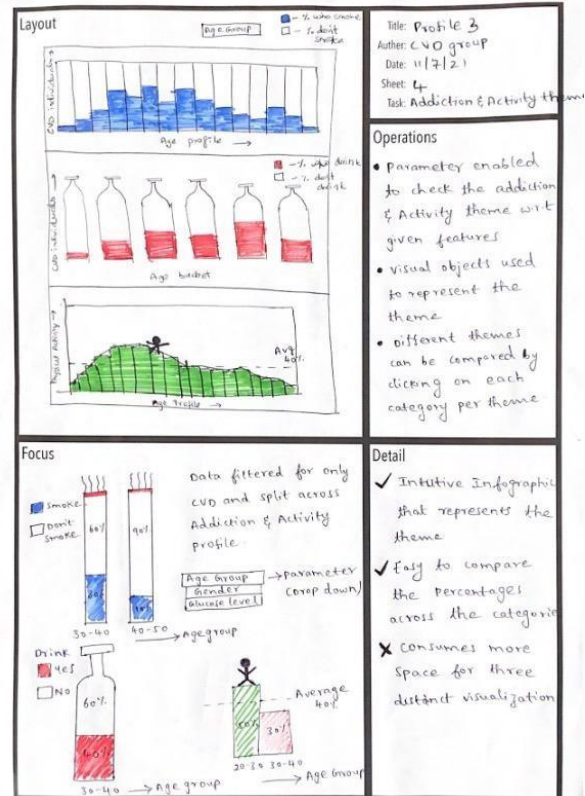
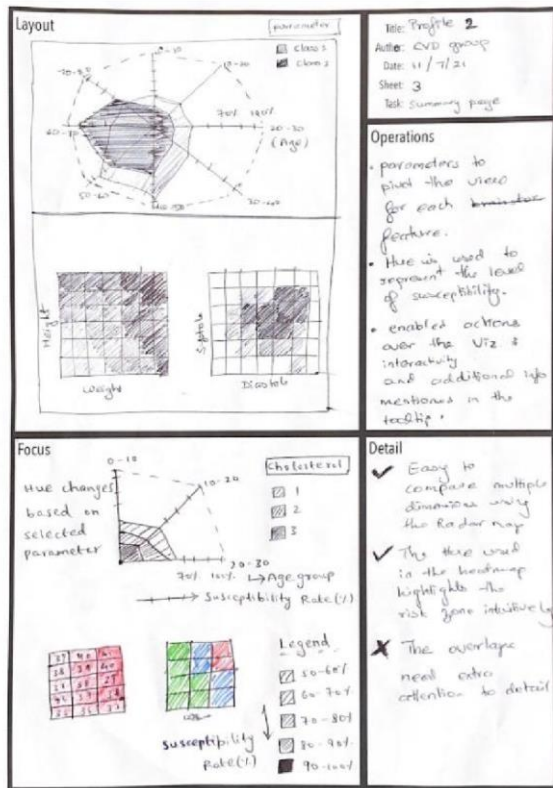
- [1] *Cardiovascular diseases (CVDs)*. (2021, June 11). World Health Organization (WHO). [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
- [2] *The top 10 causes of death*. (2020, December 9). World Health Organization (WHO). <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>
- [3] *Heart Disease Facts | cdc.gov*. (2021, September 27). Centers for Disease Control and Prevention. <https://www.cdc.gov/heartdisease/facts.htm>
- [4] *Leading Causes of Death*. (2021). Centers for Disease Control and Prevention. <https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm>
- [5] NHS website. (2021, November 29). *Cardiovascular disease*. Nhs.Uk. <https://www.nhs.uk/conditions/cardiovascular-disease/>
- [6] Ben F. *Visualizing Data*.
- [7] Flora, G. D., & Nayak, M. K. (2019). A Brief Review of Cardiovascular Diseases, Associated Risk Factors and Current Treatment Regimes. *Current pharmaceutical design*, 25(38), 4063–4084. <https://doi.org/10.2174/1381612825666190925163827>
- [8] *Cardiovascular Health Examination Survey/State Resources/DHDSP/CDC*. (2019). Center for Disease Control and Prevention. https://www.cdc.gov/dhdsp/programs/spha/examination_survey/index.htm
- [9] A. Roth, George A. Mensah, Catherine O. Johnson, Giovanni Addolorato & Others. (2020). Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study, *Journal of the American College of Cardiology*, Volume 76, Issue 25, 2020, Pages 2982-3021, ISSN 0735-1097, <https://doi.org/10.1016/j.jacc.2020.11.010>.
- [10] Stewart J, Manmathan G, Wilkinson P. Primary prevention of cardiovascular disease: A review of contemporary guidance and literature. *JRSM Cardiovasc Dis*. 2017;6:2048004016687211. Published 2017 January 1. doi:10.1177/2048004016687211
- [11] *Cardiovascular Disease dataset*. (2019, January 20). Kaggle. <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>
- [12] *Heart Disease UCI*. (2018, June 25). Kaggle. <https://www.kaggle.com/ronitf/heart-disease-uci>
- [13] M. G. Tsipouras et al., "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling," in *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 4, pp. 447-458, July 2008, doi: 10.1109/TITB.2007.907985.

- [14] *Know Your Risk for Heart Disease* | *cdc.gov*. (2019, December 9). Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/risk_factors.htm
- [15] *Analysis of Heart Disease*. (2020). [Dashboard]. <https://public.tableau.com/app/profile/kadambari.kutre/viz/AnalysisofHeartDiseases/AnalysisofHeartDiseases>
- [16] *Heart Disease Dashboard*. (2019). [Dashboard]. https://public.tableau.com/views/Kartik-HeartDiseaseDashboard/Dashboard1?%3Aembed=y&%3AshowVizHome=no&%3Adisplay_count=y&%3Adisplay_static_image=y&%3AbootstrapWhenNotified=true
- [17] *Types of Color Blindness* | *National Eye Institute*. (2019). National Eye Institute. <https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/color-blindness/types-color-blindness>
- [18] Heer & Shneiderman (2012). Interactive dynamics for visual analysis.
- [19] Cooper, et al. Goal-Directed Design process. *About Face 4*
- [20] Stephen Few's Now You See It. Analytics Press; 1st edition (April 1, 2009). <http://perceptualedge.com/>
- [21] Edward Tufte. (2001). The Visual Display of Quantitative Information. http://www.edwardtufte.com/tufte/books_vdqi
- [22] Mackinlay, J. (1986). Automating the Design of Graphical Presentations of Relational Information. <https://research.tableau.com/sites/default/files/p110-mackinlay.pdf>
- [23] Data Visualization - Basic Principles of Information Visualization by Sunny Solanki. (2020). CoderzColumn. <https://coderzcolumn.com/blogs/data-science/basic-principles-of-information-visualization>
- [24] Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2), 130–135. <https://doi.org/10.11919/j.issn.1002-0829.215044>
- [25] Rakesh A. & Ramakrishnan S. (2015). Fast Algorithms for Mining Association Rules. <http://www.cse.msu.edu/~cse960/Papers/MiningAssoc-AgrawalAS-VLDB94.pdf>

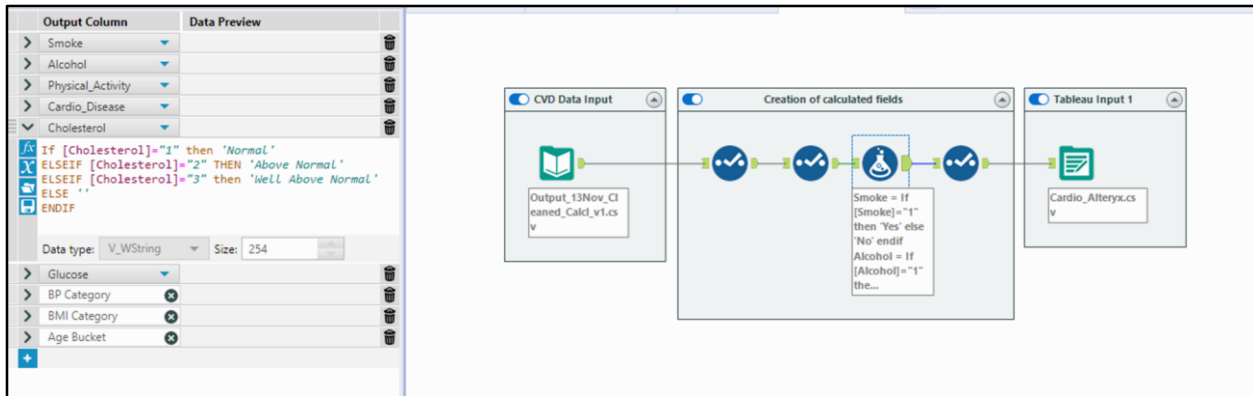
12. Appendix

1) Five Design Sheets





2) Alteryx Data Cleaning



3) Apriori Association Rule Mining

```

rules <- apriori (data=trans, parameter=list (supp=0.10,conf = 0.63,maxlen=4), appearance = list
(default="lhs",rhs="Cardio_Disease=Yes"), control = list (verbose=F)) # get rules that lead to CVD
rules_conf <- sort (rules, by="confidence", decreasing=TRUE) # 'high-confidence' rules.
length(rules_conf)
inspect(head(rules_conf))

```

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{Glucose=Normal, Physical_Activity=Yes, BP.Category=Hypertension}	=> {Cardio_Disease=Yes}	0.1190734	0.8374115	0.1421922	1.691524	7803
[2]	{Glucose=Normal, Alcohol=No, BP.Category=Hypertension}	=> {Cardio_Disease=Yes}	0.1377211	0.8361901	0.1647007	1.689057	9025
[3]	{Glucose=Normal, BP.Category=Hypertension}	=> {Cardio_Disease=Yes}	0.1467092	0.8359273	0.1755047	1.688526	9614
[4]	{Physical_Activity=Yes, BP.Category=Hypertension}	=> {Cardio_Disease=Yes}	0.1483725	0.8323774	0.1782515	1.681355	9723
[5]	{Alcohol=No, Physical_Activity=Yes, BP.Category=Hypertension}	=> {Cardio_Disease=Yes}	0.1383620	0.8317586	0.1663488	1.680105	9067
[6]	{BP.Category=Hypertension}	=> {Cardio_Disease=Yes}	0.1829363	0.8309420	0.2201553	1.678456	11988