Process Documentation of the Visualization of NHANES Dataset on Health Indicators of Lifestyle Habits

Group ID: 1

Group Members: Zachary Trabookis, Acheme Christopher Acheme, Srushti Sameer Tikekar

Overview and Motivation

The Centre for Disease Control (CDC) of the United States Government among other things, through the National Health and Nutrition Examination Survey (NHANES), collects a broad spectrum of health and nutritional information from various US population groups. This data is used by researchers and policymakers to study patterns and proffer solutions for various health concerns affecting the populace. While they generate an expansive dataset, it appears to be too broad for analytical purposes. The authors of this sub-dataset extracted the most relevant features/attributes (including physiological measurements, lifestyle choices, and biochemical markers) which were hypothesized to have strong correlations with age. For our purposes we took respondents' identity and mapped to additional questionnaires data, then created a visualization to demonstrate the relationship between lifestyle habits - specifically physical exercise and health indicators. The goal is that through this visualization, policy makers and individuals can see a correlation between lifestyle habits and the health of individuals in gender and age categories. Hopefully, this allows the government and interested parties including the individual to properly design programs to enhance their health through lifestyle habits.

Related Work

Engaging in physical activity is essential for maintaining good health. The CDC, for example, recommends that adults perform at least 150 minutes of moderate aerobic exercise weekly to significantly lower their chances of early death, heart disease, type 2 diabetes, and some cancers. Beyond physical benefits, regular exercise also boosts mental health, sharpens cognitive abilities, and enhances overall life quality. It is also worthy of note to mention that the CDC also provided recommendations for various age groups and special groups like adults with chronic health conditions and disabilities, as well as pregnant women to further underscore the importance of physical exercise for everyone [7].

The initial report produced by the CDC on *National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset* [1] gave us inspiration to build out a dashboard showing metrics for lifestyle habits affecting participants health metrics based on age, gender, or other categorical measures. This report showed us that we could limit the number of input fields from the vast NHANES dataset and focus on things like lifestyle habits (ie. exercise level, food spending habits) and health metrics (ie. diabetes, blood and cholesterol, weight, age) which could factor into the participants' overall health.

Looking at a CDC published article on *Percentage* of Adults Aged ≥18 Years with Diagnosed Diabetes,† by Urbanization Level§ and Age Group* [8] provided insight into how they went about using collected health metric data to report to the public. This article specifically relates to diabetes and age group which is closely related to what we're trying to accomplish with our dashboard. The CDC further subdivides the bars to show categorization of metropolitan areas which could have been utilized on our dashboard.

Questions

We want to provide relevant information for users of this visualization to answer the following questions: Questions we can answer:

1. Does the Insulin Level and Blood Glucose Level Affect the Body Mass Index?
2. What percentage of women to men engage in vigorous physical exercise?
3. Which age group is the most active in physical activity?
4. How many {female/male} participants (or percentage of participants) aged 60 and above engage in vigorous physical exercise?
5. What are the health indicators for the {female/male} age group 60 and older?

However, we did not just stumble on these questions. We initially started with some ambitious questions in our first iteration even before drawing a line or a point for visualization. Our initial set of questions included:

1. What is the ratio between females and males that are diabetic, borderline or show no sign of diabetes?
2. What percentage of respondents fall into different ethnic groups?
3. Do Females and Male have Different Insulin Levels as per their BMI?
4. Is the metric for someone's body mass index (BMI) directly related to their waist dimension?
5. Does the weight of a respondent increase or decrease their insulin levels?
6. Can gender affect how weight changes insulin levels?
7. Where are (males/females) with diabetes located within the United States?
8. What specific ethnic groups have a higher count of diabetes within the United States?
9. Do men have a higher or lower body mass index (BMI) than women?

10. What is the Glucose trend with respect to BMI?
11. What are normal and abnormal insulin vs glucose levels?
12. Does insulin level increase as glucose levels increase for diabetic, non-diabetic and borderline respondents?
13. Are men more at risk of diabetes than women?
14. What age range is more at risk of diabetes in men vs women?
15. Do biomarker values differ for men at risk vs women at risk of diabetes?
16. Do people who workout have less chances of getting diabetes?
17. How many adult people exercise?
18. Diabetes Status as per age group?

Data

Dataset Description: The items in the NHANES dataset were collected through examination and questionnaires. It is composed of 2278 respondents representing persons captured in the survey.

Data preprocessing: Below are the preprocessing steps for converting the questionnaire dataset with labels. More details on the dataset variables will be found in Table 2.

- This subset of data was used to bin respondents greater than or equal to 65 years old as "senior", while all other individuals under 65 years old as "non-senior."
- Diabetic: Answered "Yes" to "Have you been told by a doctor that you have diabetes?" or had Plasma Glucose ≥ 126 mg/dl.
- Undiagnosed Diabetic: Answered "No" but had Plasma Glucose ≥ 126 mg/dl.
- Pre-diabetic: Had Plasma Glucose between 100 - 125 mg/dl.
- Not Diabetic: Had Plasma Glucose ≤ 100 mg/dl.
- Case I: Includes records of diabetic (1), pre/undiagnosed diabetic (1), and non-diabetic (0) patients.
- Case II: Includes only pre/undiagnosed diabetic (1) and non-diabetic (0) patients. Diabetic patients are excluded.
- Having heart diseases: Answered "Yes" to having had one of the following: congestive heart failure, coronary heart disease, heart attack, or stroke (Label = 1).
- Not having heart diseases: Answered "No" to all conditions (Label = 0).

The dataset for this project was sourced from various questionnaire datasets for the year 2013-2014. After preprocessing, including handling missing values, the final dataset was constructed as follows:

Table 1: Dataset Type, Items, Attributes

Dataset	Items	Attributes
Tabular	Represent individual survey respondents in the United States population range from adults to children. Individual respondent age restriction is not defined. Total item count in the subset of NHANES data is 2278.	Quantitative (6 values) Ordinal (2 values) Categorical (2 values) There are 7 feature attributes that have a strong correlation to prediction age. There is 1 target attribute for age.

Table 2: Dataset Information and Description

Variable Name	Role	Type	Demographic	Item/ Attribute	Description
SEQN	ID	Continuous		Item	Respondent Sequence Number
age_group	Target	Categorical	Age	Attribute	Respondent's Age Group senior - 65 or older non-senior - under 65
RIDAGEYR	Other	Quantitative	Age	Attribute	Respondent's Age
RIAGENDR	Feature	Categorical	Gender	Attribute	Respondent's Gender 1 - male 2 - female
PAQ605	Feature	Ordinal		Attribute	If the respondent engages in moderate or vigorous-intensity sports, fitness, or recreational activities in the typical week

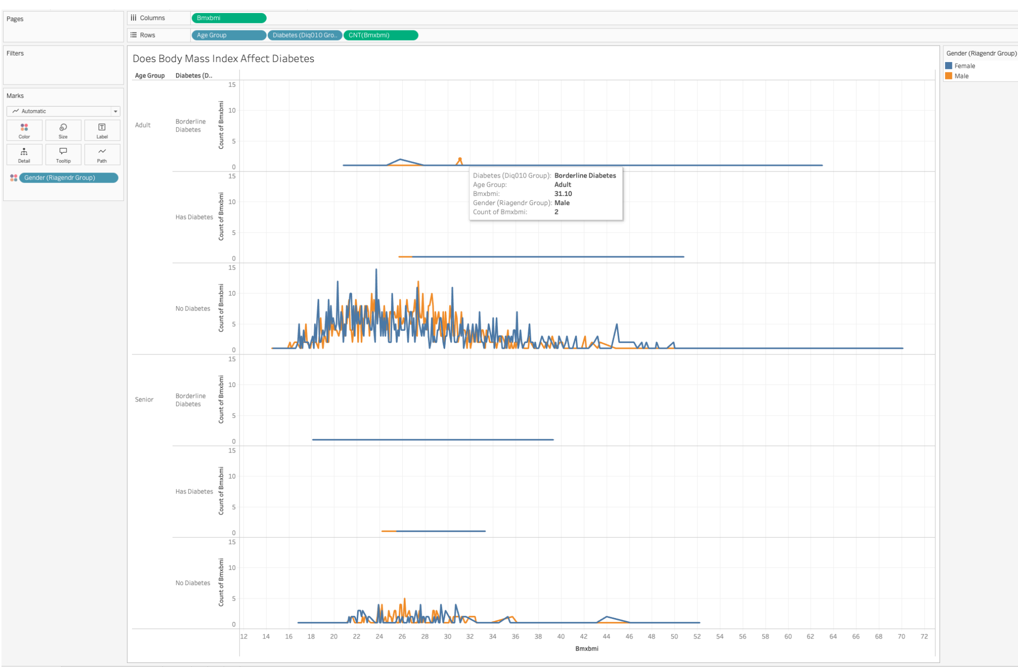
					See additional information below in PAQ605 table: 1 - respondent takes part in weekly moderate or vigorous-intensity physical activity 2 - represents that they do not vigorously work out. 7 - refused
BMXBMI	Feature	Quantitative		Attribute	Respondent's Body Mass Index
BMXWAIST	Feature	Quantitative		Attribute	Respondent’s Waist Circumference (cm)
LBXGLU	Feature	Quantitative		Attribute	Respondent's Blood Glucose after fasting(mg/dL)
DIQ010	Feature	Ordinal		Attribute	If the Respondent is diabetic See additional information below in DIQ010 table: 1 - yes to having diabetes 2 - no to having diabetes 3 - borderline to having diabetes
LBXGLT	Feature	Quantitative		Attribute	Respondent’s Two Hour Glucose(OGTT) (mg/dL)
LBXIN	Feature	Quantitative		Attribute	Respondent's Blood Insulin Levels (uU/mL)
BPQ080	Feature	Categorical		Attribute	Doctor told you - high cholesterol level 1 - Yes 2 - No 7 - Refused 9- Don’t Know

<u>Further preprocessing for project:</u>

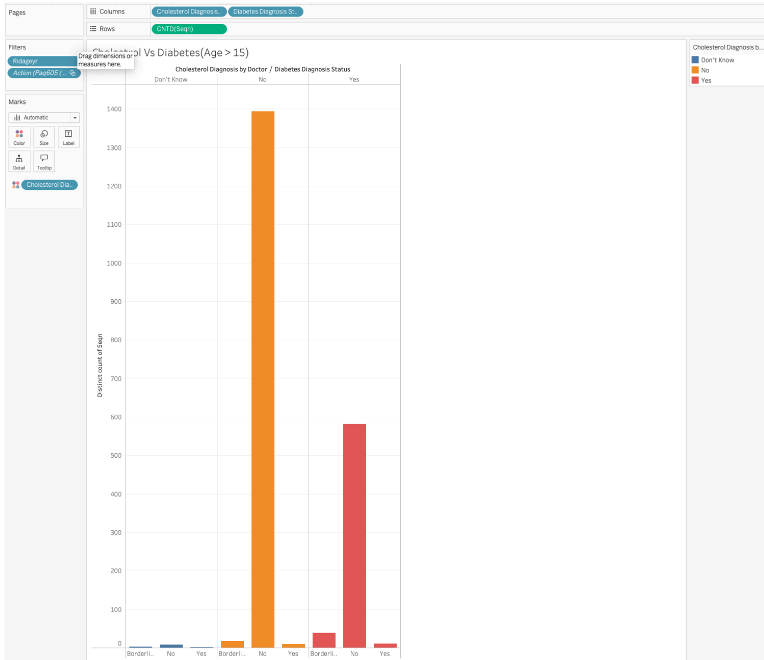
- We started searching for additional questionnaire data which can be mapped to the respondents for [Cardiovascular](), [Consumer Behaviour](), [Blood Measures](), [Blood Pressure & Cholesterol](), [Diet Behavior & Nutrition](), [Demographic Variables and Sample Weights](), [Diabetes](), [Occupation]().
- Preprocessed the data using a Python code and added new column fields like Race, Country, Education, Marital Status, Occupation details, etc and imported the new merged excel into tableau for Visualization. [[Source Code]()]
- We observed the data is now inconsistent where a lot of data was missing for the 2000+ respondents we are making our visualization based on. But we did observe that a field of High Cholesterol level was quite consistent for the respondents age >= 16.

Exploratory Data Analysis

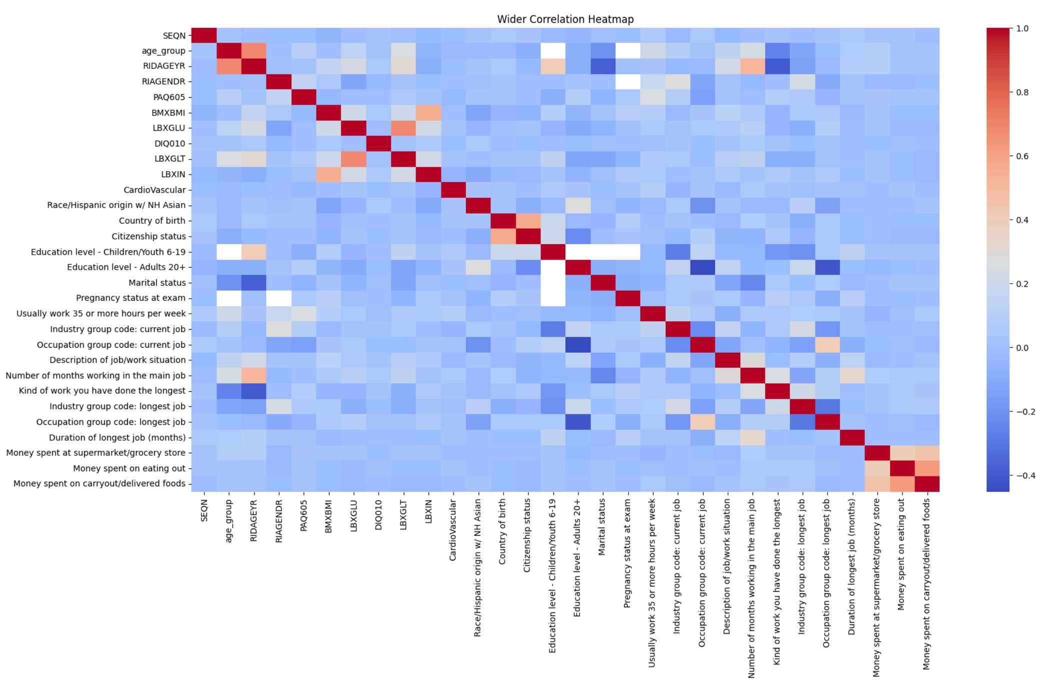
We used Excel and Tableau to conduct exploratory data analysis of the dataset. Starting with summary statistics and visualizations, we examined the relationships between variables such as patient age, BMI, diabetes status etc. Our first attempt was to visualize a trend (line graph) correlating increasing BMI with the diabetic status of respondents where colors differentiated each gender.



We checked the correlation of cholesterol levels and diabetes status of respondents reported with the graph below:



Next, we tried to find a correlation between all the attributes we have now in excel and plotted a heatmap and observed that we do not see much correlation between the other attributes which we integrated in the dataset. This was interesting as it helped tell us about the kind of questions a visualization of the dataset could address. In the heatmap below the correlation color range on the right goes from 0.0 - 1.0, with the closer to 1.0 meaning that the fields that intersect (x and y axes) are strongly correlated.

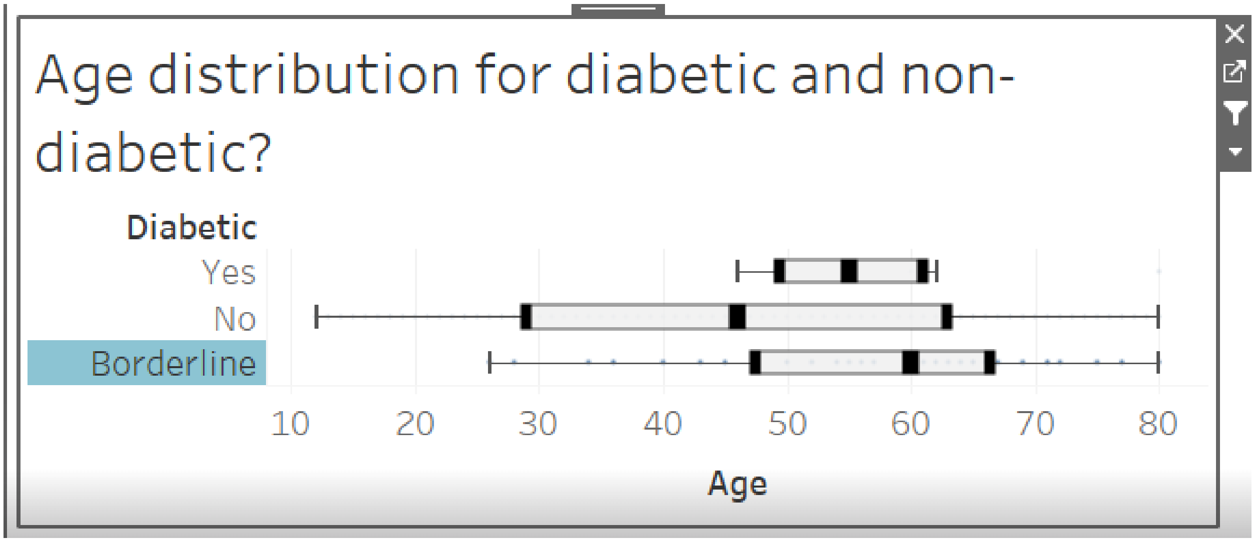


Another health metric that we wanted to show on our dashboard was waist circumference measurement when compared with Body Mass Index (BMI), filtered by gender and diabetic status. Waist measurement indicates fat distribution around the abdomen and is a widely accepted metric indicating potential health risk leading to higher risks of chronic diseases like diabetes and heart disease. BMI considers height and weight attributes, not taking into account fat stored on the body. BMI also doesn't differentiate between muscle and fat, so muscular people might have higher BMI despite having a lower body fat percentage.

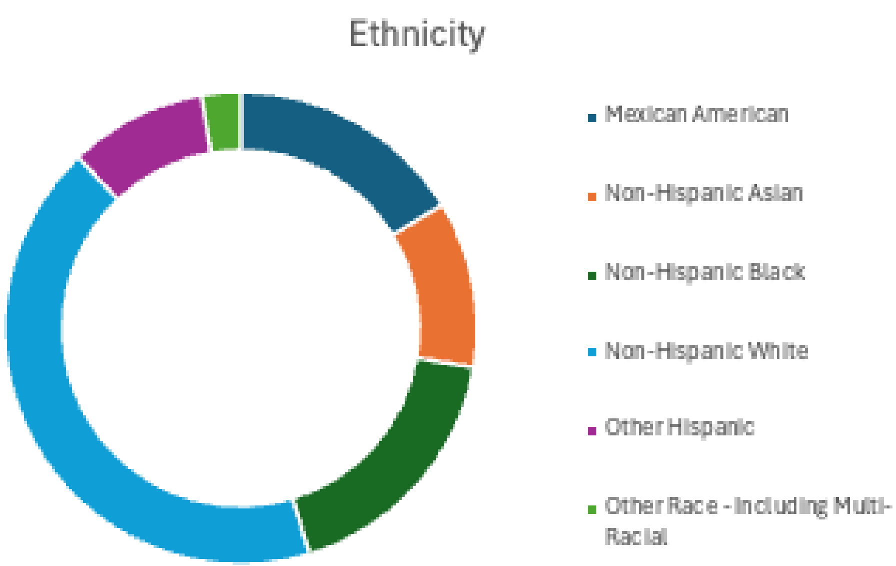
The chart below groups waist circumference (cm) versus BMI metrics using color markers for gender (ie. Male, Female) participants. This allows the user to see the distribution by gender more easily. A shape marker was used to further classify the diabetic indicator where a doctor has classified the participant as previously having been diagnosed with diabetes (ie. Yes ➕, No ▢ , Borderline ○). Because there were so many participants with the No diabetes indication, it was hard to see shapes for the other two diabetes indicators for Yes and Borderline, so in the final dashboard we left off this shape marker.

Waist Circumference vs Body Mass Index (BMI)
by Gender and Diabetes Status

Furthermore, we explored the age distribution of respondents captured in the dataset with their diabetes status by using a box and whisker plot. We moved away from the story where we wanted to show groups with diabetes and focused on a story to show the relationship with lifestyle habits and health indicators through some blood measures.

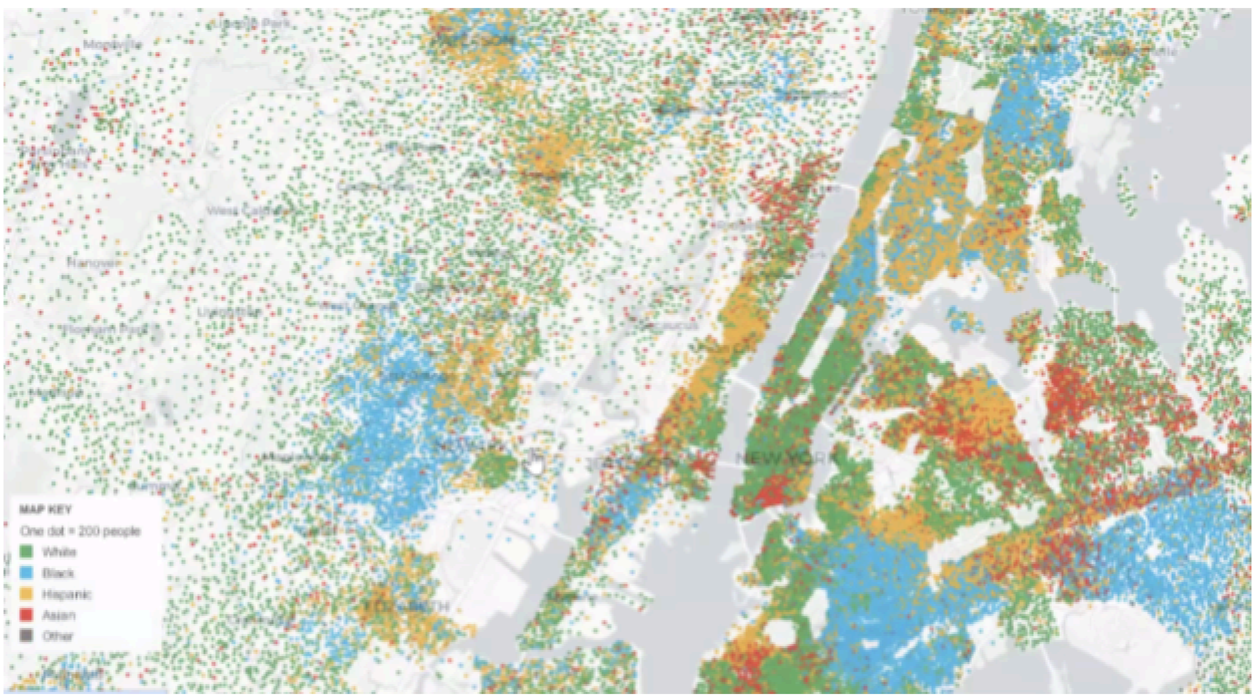


We considered the distribution of ethnicity with a doughnut plot shown below. We chose not to pursue this direction as we sought to answer questions related to age and gender.



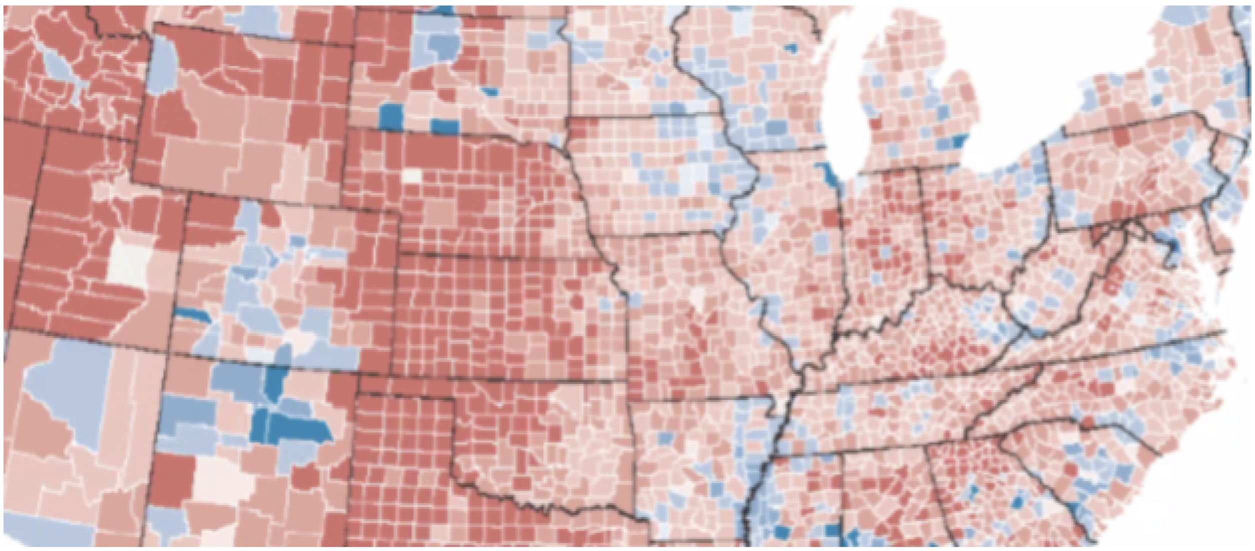
Geographic Participant Information
The data collected by NHANES related to the United States of America (USA) population included geocode and Census related information for each participant. After having reached out to a NHANES Research Data Center (RDC) analyst, the access to use this geospatial Latitude/Longitude and Census information for City/State was denied for use by the public because of classification/usage restrictions only provided to federal government research facility employees. This prevented us from being able to show any of the map mock-ups below on the final dashboard.

Participant Ethnic Group Population Density Map of the United States of America
Latitude and longitude geocode information in conjunction with ethnicity data would allow for a visualization of point data on a USA map representing individuals by ethnic group for a given geographic area. This class example shows points, representing individuals, marked with different color markers by ethnic groups of the New York City area. This point map would allow the user to filter by ethnic group the map points and corresponding charts on the dashboard would adjust showing only the ethnic group data selected. The example below was pulled from a class and does not reflect the mockup of participant ethnic group population density for the USA map.



**New York City Map Representation by Ethnic Group:
Points Represent Individual People and Color Markers for Each Ethnic Group**

Participant Diabetes Status County Map of the United States of America
Census information for state, city, tract are defined in the NHANES dataset per participant. To show a breakdown by county, the data would need to be cleaned up to map a participant to a county. Using a Choropleth map of participant information for county and diabetes status (Yes, No, Borderline) would provide the user with an interface to see what parts of the country have more or less diabetes issues due to their lifestyle habits (ie. Exercise level (Vigorous, No)). The example below was pulled from class and does not reflect the mockup of participant diabetes status per county for the USA map.



Choropleth Map Showing Counties of USA by Diabetes Status

Design Evolution

We started redesigning the dashboards individually for more inputs and lining up the charts to a story and came up with some visualization (Below) . Each dashboard as a different story

Design 1:

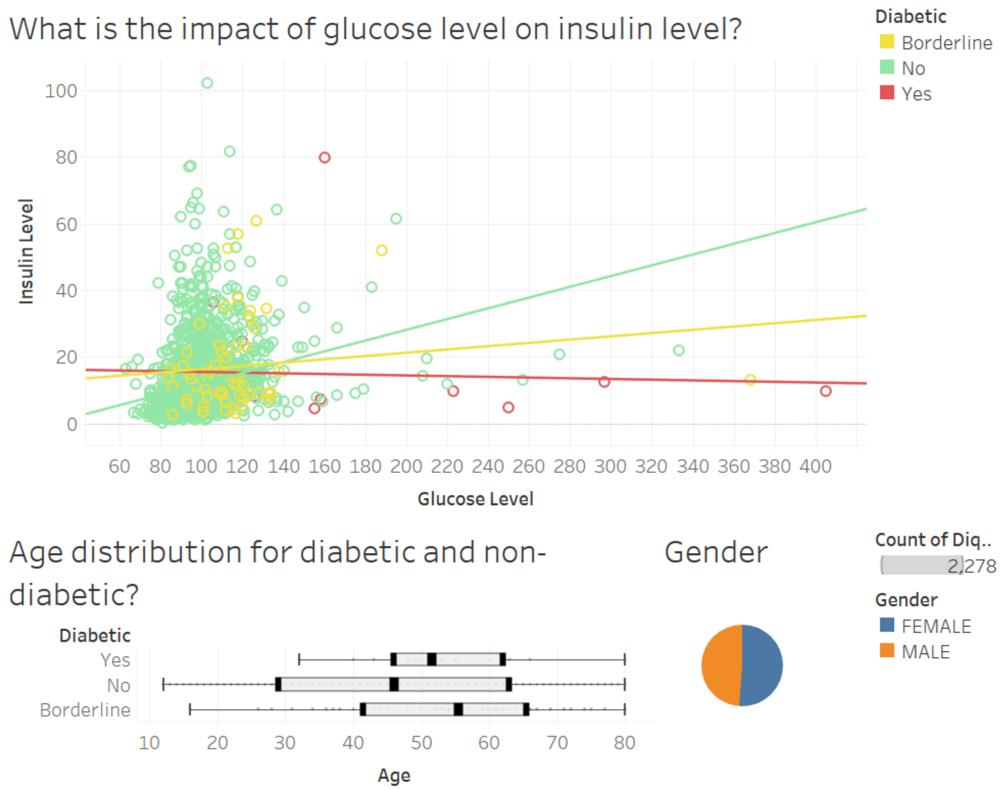
- The dashboard illustrates the connections between lifestyle habits and health indicators. It reveals that individuals with higher BMIs tend to have elevated glucose and insulin levels, which may suggest an increased risk of diabetes. There is also a noticeable correlation between diabetes and cholesterol issues, emphasizing the need for a more integrated approach to health management. The data indicates that a significant portion of participants do not engage in vigorous exercise, reflecting a predominantly sedentary lifestyle. Encouraging regular physical activity could help in weight management, blood sugar regulation, and cholesterol improvement, ultimately reducing the overall risk of diabetes and associated health concerns.

- We realized that the information perceived through interconnectivity of the charts is missing. The next step was to identify how these graphs can be interconnected through the permutation of the attributes which we have currently identified.

Design 2 (Alternate):

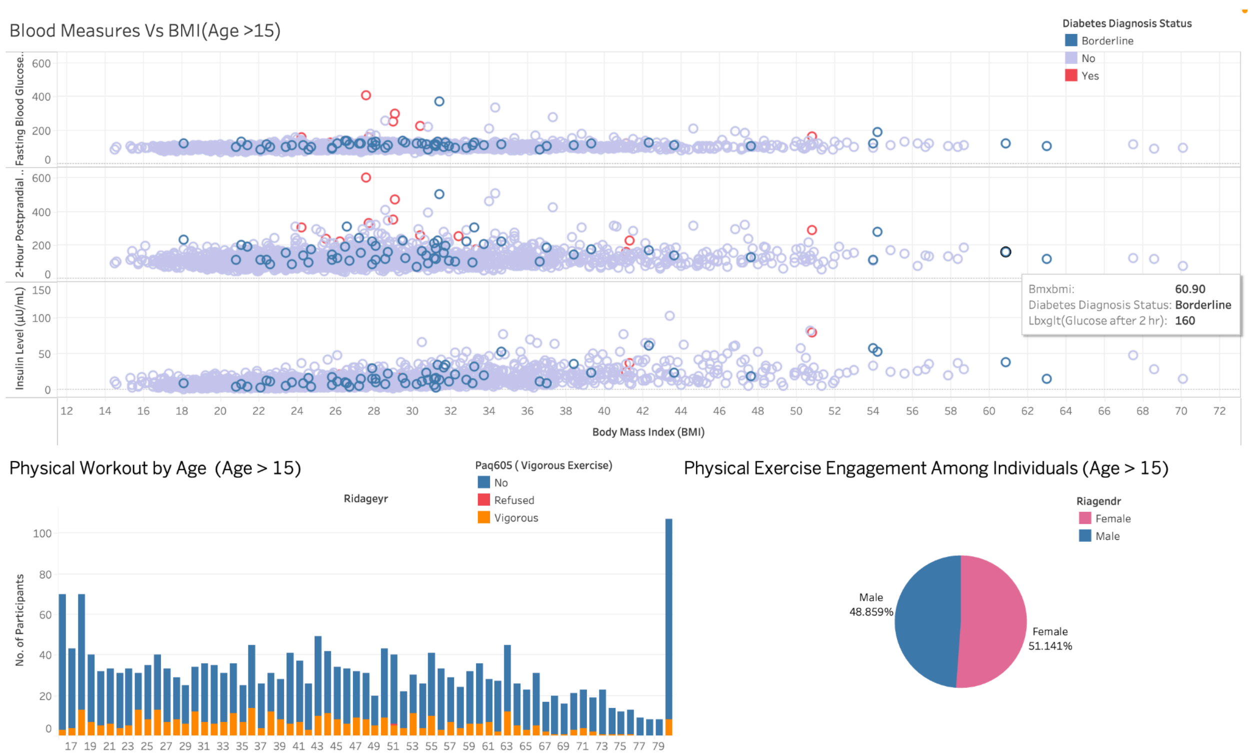
We conceived a second visualization seeking to show the trend of diabetes biomarkers - specifically the relationship between insulin level and glucose level for various age groups and gender.

We realized that our visualization lacked interactivity and the scatter plot (shown below) did not really change with the bottom two aggregate graphs. Hence we decided to focus on our first option which had more potential for interactivity and hence a better visualization.



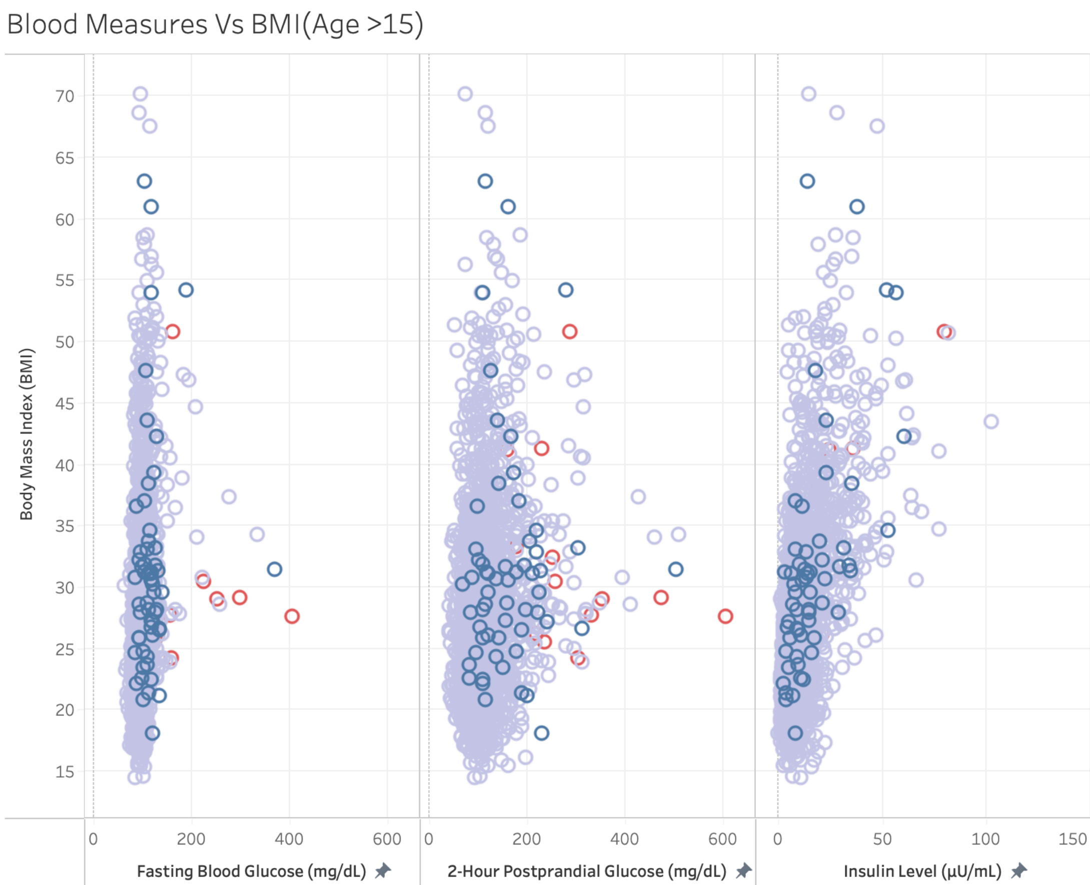
Design 3 (Final Blueprint):

- Based on feedback from the previous designs, we identified three interconnected charts that collectively convey a cohesive story.
- A stacked bar chart for vigorous workouts by age was added, inspired by Design 2.
- We chose a bar chart format over a box plot to provide more detailed information.
- To improve readability, we binned age groups in the final visualization.
- This reduces the number of bars and optimizes the chart size for better selection and viewing.



Design 4 (Changes from Professor Feedback):

- Blood Measures vs. BMI
  - Changing BMI to the Y-axis and Blood Measures to X-axis to see if the data points in the scatterplot would spread out to show a better distribution of points.



- Physical Exercise vs. Age
  - We may want to select our age group categorical by factoring in diabetic changes. This [article](#) by the CDC has specified age groups as follows. Because we wanted the user to compare themselves with other participants in an age group we decided to go from an individual age to a group selection.

    Examples of age grouping:

    ≤17 : Teenagers
    18–34: Younger Adults
    35-59: Older Adults
    60 and older: Seniors

Implementation

Dashboard Development Tools
We used several tools to build out this analytics dashboard, among them being Github (repository, projects, and pages).

- **Github**
  Provides a cloud repository to store our code changes. Code repository for this dashboard can be found at https://github.com/ztraboo/cpsc-6030-course-project**.** Git was used to create commits, track and merge changes, and perform Pull Requests (PR) as needed.

  Branches
  We created several branches to help organize new feature changes, development progress from all team members, a main branch for final release, and a separate branch (ie. gh-pages) to serve the web application from.

  main - Final release changes were merged from `develop` into this branch and is used to track major changes between releases.

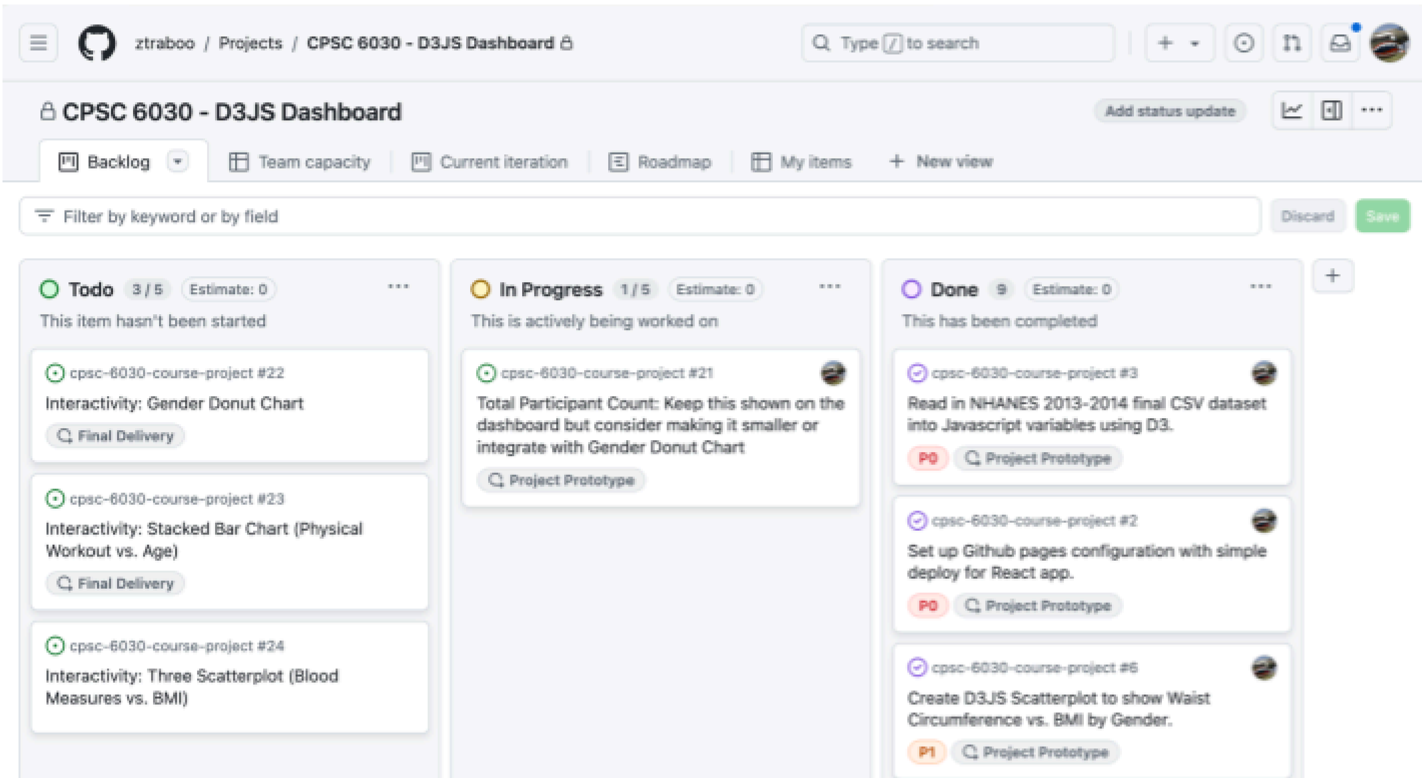
  gh-pages - Used to deploy to Github Pages for our final static site delivery. Issuing a `npm run deploy -- -m "New Commit Message"` command to merge changes into this branch and kick off a Github action to deploy the static site.

  develop - Contains the latest in-progress work from multiple feature branches from individual team contributors.

  feature/<username>/<change> - New feature branches were created from the `develop` branch to allow modification to the code without affecting other team members' changes. Pull Requests (PR) were used to merge changes into the `develop` branch. PR allowed us to track changes and have discussions prior to merging changes.

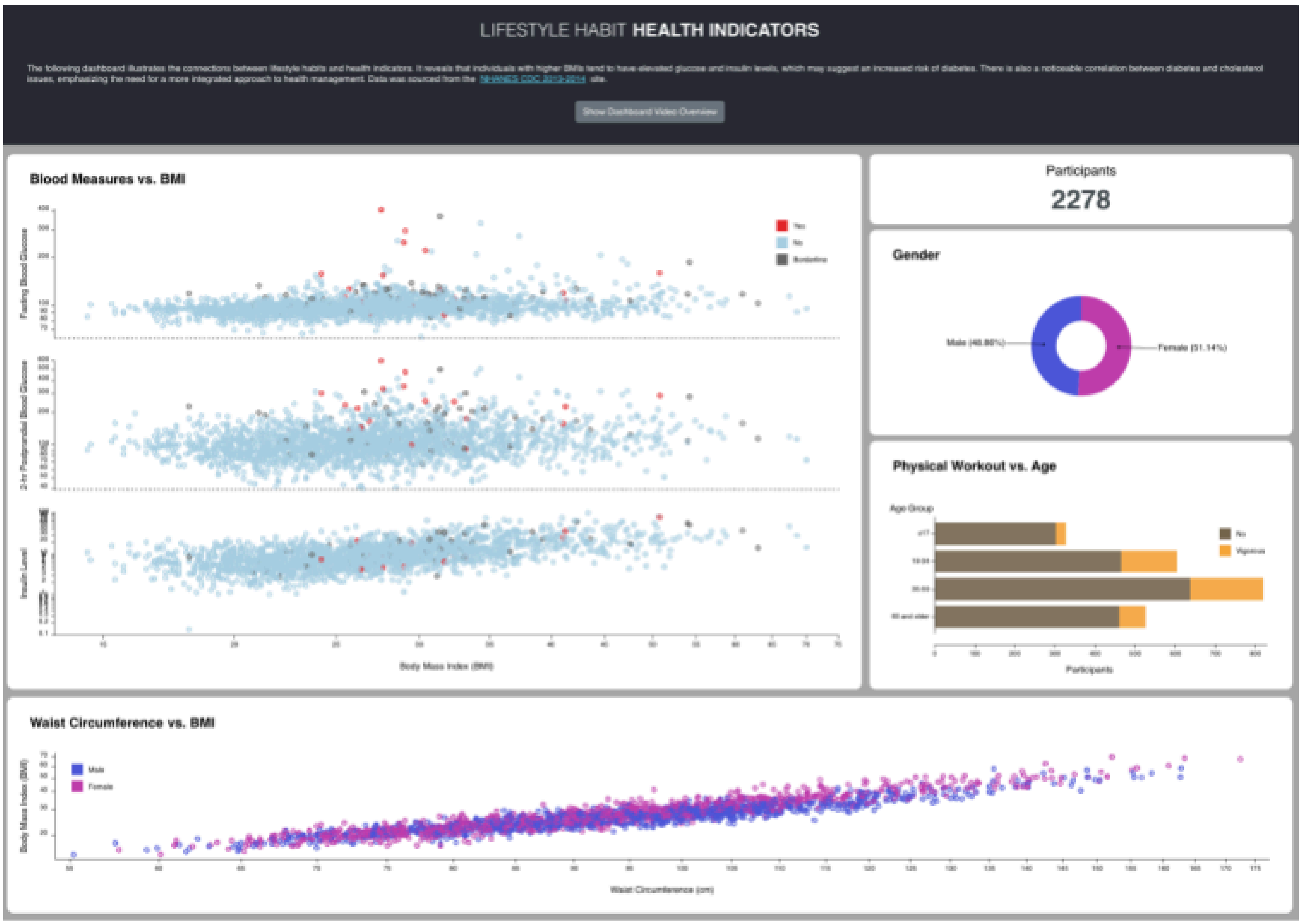
- **Github Project**
  A tool within the repository that allowed our team to divide up the tasking needed to complete the dashboard work. This project tool allowed us to create tasks, define a specified interval work period for the task to limit scope, and assign the task to one or more team members. The Kanban style dashboard shows all tasks and their task progress states (Todo, In Progress, and Done) to see at a glance where we are with

completing the work for the dashboard project.



A Kanban Style Dashboard for Project Tasking

- **Github Pages**
  The final delivery website for our dashboard project can be found at https://ztraboo.github.io/cpsc-6030-course-project/. The dashboard was developed with React frontend framework and npm publishing tools. When using npm to build out the site for final delivery, it generates static website files that can be hosted on Github Pages.



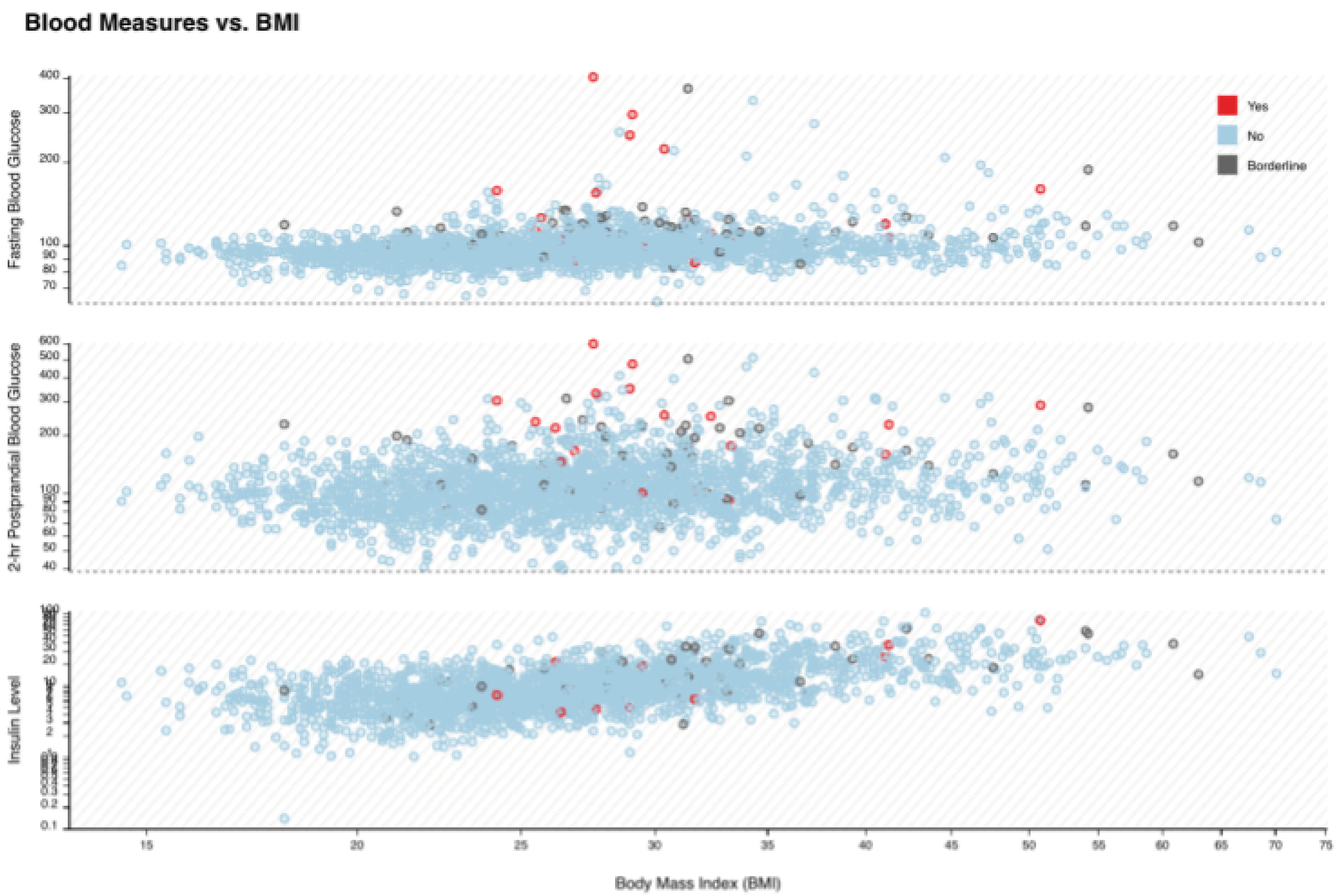
Github Pages: Final Delivery for Dashboard Project

D3 Implementation of Initial Design Mockup

This project visualizes health and lifestyle data using two distinct chart types in D3.js: a scatter plot showing blood measures versus BMI, and a horizontal stacked bar chart that represents physical workout engagement across different age groups. [Source Code]

<u>1. Scatter Plot: Blood Measures vs. BMI</u>

The plot-scatter.js file implements a multi-layered scatter plot showing the relationship between Body Mass Index (BMI) and three blood measures: fasting blood glucose, 2-hour postprandial glucose, and insulin levels. Each measure is plotted on a separate horizontal panel, sharing a common x-axis (BMI), while the y-axis represents each respective measure's values.



Scatterplot charts for Blood Measures vs BMI

- Data Loading and Parsing: The data is loaded from a CSV file, and relevant fields (BMI, glucose levels, insulin, and diabetes diagnosis status) are extracted and converted into numerical values where needed.
- Axes and Scales:

  A single x-axis is shared across the plots using BMI values.

  Individual y-axes are generated for each blood measure to adapt the range to each measure's data.

  Both axes are scaled logarithmically to ensure a greater spread in clustered values, allowing the user to see more of the data points. Our professor recommended this improvement, as we initially had a linear scale, and the values were too close together.

- Plotting: Circles represent data points, with color encoding diabetes diagnosis status. A color scale is used to differentiate between categories: "Yes," "No," and "Borderline."
- Legends:

  A legend provides a clear mapping between color and diabetes status.

2. <u>Horizontal Stacked Bar Chart: Physical Workout vs. Age Group</u>

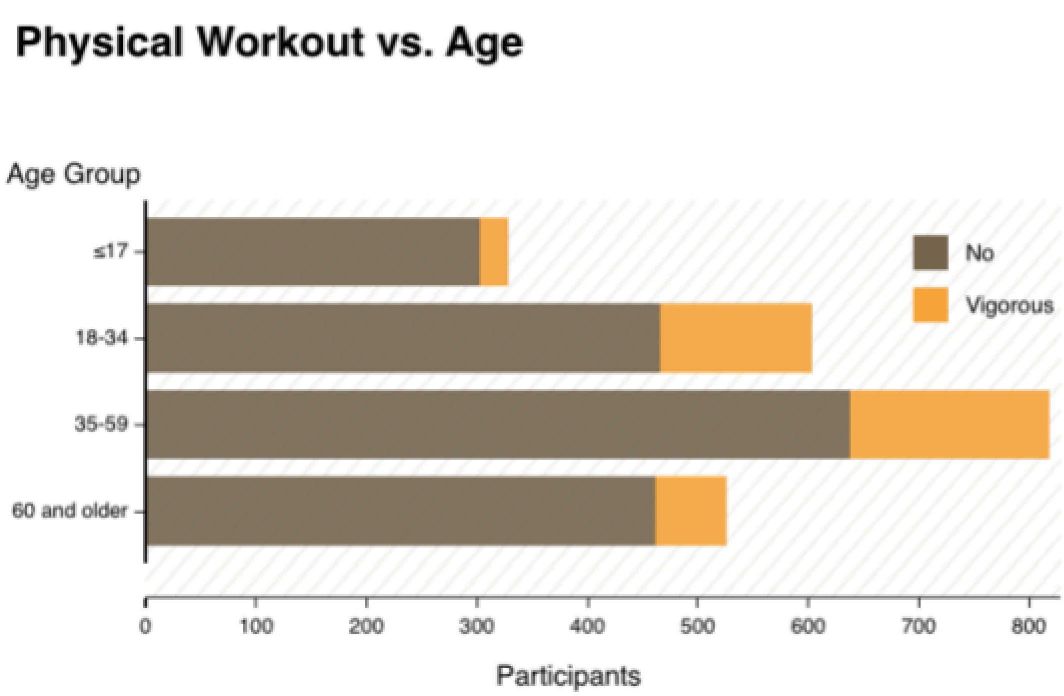
The horizontal-bar.js file visualizes the relationship between age groups and physical workout engagement, categorized into "No" and "Vigorous." Each bar represents an age group, segmented by the count of participants in each workout category.

- Data Aggregation: The data is categorized into four age groups: "≤17," "18-34," "35-59," and "60 and older." Each group counts participants who either perform vigorous workouts or do not, using D3's rollup function to aggregate data.
- Axes and Scales:

  The y-axis uses categorical age groups, while the x-axis represents the number of participants.

  Each age group is represented as a bar, with the total count of participants stacked horizontally.

- Stacked Bar Segments: The bars are divided into two segments per age group (No and Vigorous), with colors differentiating the workout levels.
- Legend and Labeling: A legend identifies the workout levels. Labels on the x-axis represent participant counts, while the y-axis categorizes the data by age group.



Stacked bar chart interface for controlling physical workout (No, Vigorous) and Age Group selection.

HTML Integration

The index.html file organizes the visualizations on a single page, linking each JavaScript file. The structure is simple:

- Blood Measures vs. BMI section includes the scatter plot generated by plot-scatter.js.
- Physical Workout vs. Age section contains the horizontal stacked bar chart from horizontal-bar.js.

References

D3.js Documentation. Available at: https://d3js.org/

React Integration of D3 Charts
Utilizing the React frontend framework and Create React App, our application development environment allowed hot reloading for quick development iteration, built in web server and publishing capability with npm tooling, objects for code reuse between display components made of axis, charts, legends and markers.

We used the following sites as reference for building out this React analytics dashboard:

- SitePoint
  https://www.sitepoint.com/d3-js-react-interactive-data-visualizations/#realworldexampleworldpopulationdashboard
  Provided an overview of how to visualize a complete example driven React dashboard using D3JS from start to finish. Utilized a grid system mechanism for quick chart layout and indicated ways to make the charts responsive by changing container widths and D3 scaler width and range changes for individual charts.

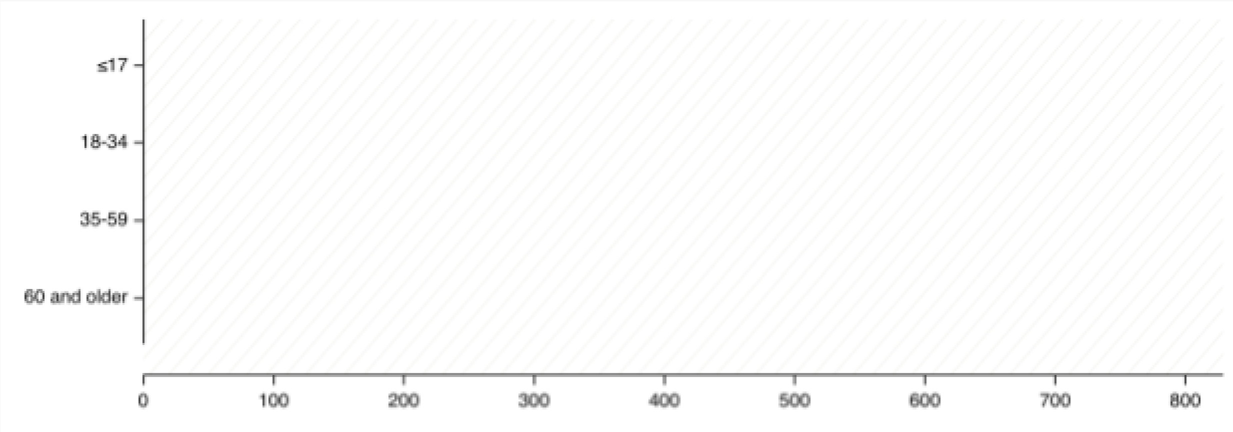
- React Graph Gallery
  https://www.react-graph-gallery.com/
  Extended the D3JS chart example library by providing practical examples for Donut, Bar and Scatterplot charts in React. Limited use of interactivity was included as well.

Display Components

The following outlines reusable React components that were utilized when constructing D3JS built visualizations from the initial design mockup of the standalone D3JS implementation of plain JavaScript and HTML.

Axis
Components for bottom and left numerical and categorical objects were created.
- Numerical objects (AxisBottom, AxisLeft) had attributes for D3JS scaler and pixels per tick mark that were used to construct `<svg>` elements of `<line>` and `<text>` elements uniformly.
- Categorical objects (AxisBottomCategorical, AxisLeftCategorical) had attributes for D3JS scaler that used the scaler domain() call to find group values.

The example below shows both numerical and categorical axes. An additional component (StripGenerator) was used to show the bound width and height of the data and give a nice render to the chart helping it stand out from the background.

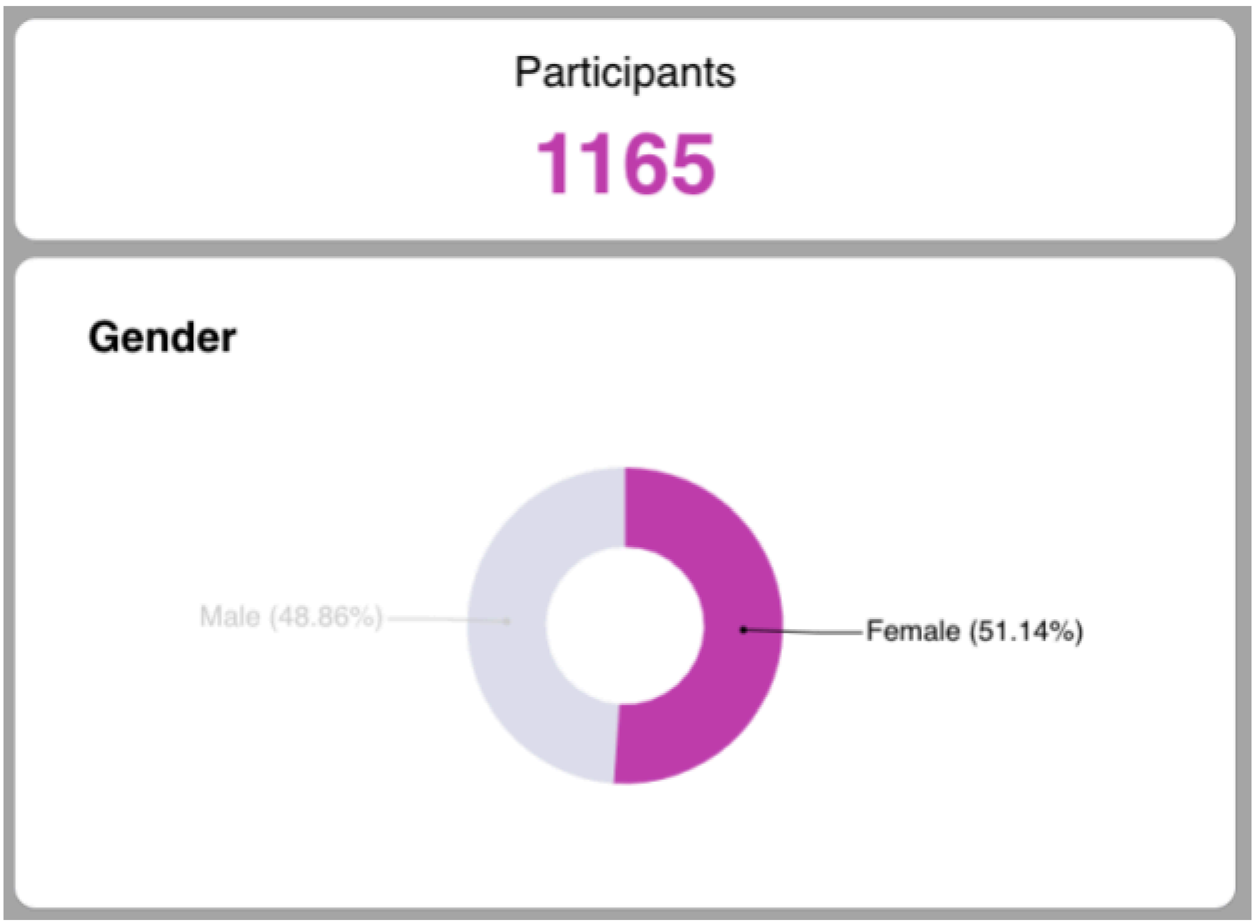


Axis React Display Components Final Rendering for Bottom Numerical and Left Categorical

Charts
Components for generic and dashboard specific donut, bar and scatterplot D3JS inspired chart objects were created.
- Generic Charts (D3DonutChart, D3ScatterplotChart, D3StackedBarChart) provided the main code to build out D3JS logic for these types of charts. Attributes were provided to support mark color, legend names, displaying axes and legends. These objects are defined for reuse for a variety of dashboard charts and vary rendered HTML and SVG element output based on parameters passed.
- Dashboard Specific Charts (DonutChartGender, ScatterplotChartBloodMeasuresVsBMI, ScatterplotChartWaistCircumferenceVsBMI, StackedBarChartAgeVsExercise) were used to customize the data and attributes passed into the generic chart components and these were also the final components used on the dashboard.

Gender Donut Chart - Slice Interaction
A pie-like chart for displaying gender (Male, Female) slices with overall participant percentages is shown. This chart provides an overall metric to see how big the dataset participant numbers were broken down by gender demographic. The user can mouse click a gender slice and see the participant count value decrease to show unique participants for that demographic, and corresponding scatter plot chart dots and stacked bar chart subgroups render only the selected pie slice type values filtering out non-selected types in the chart.

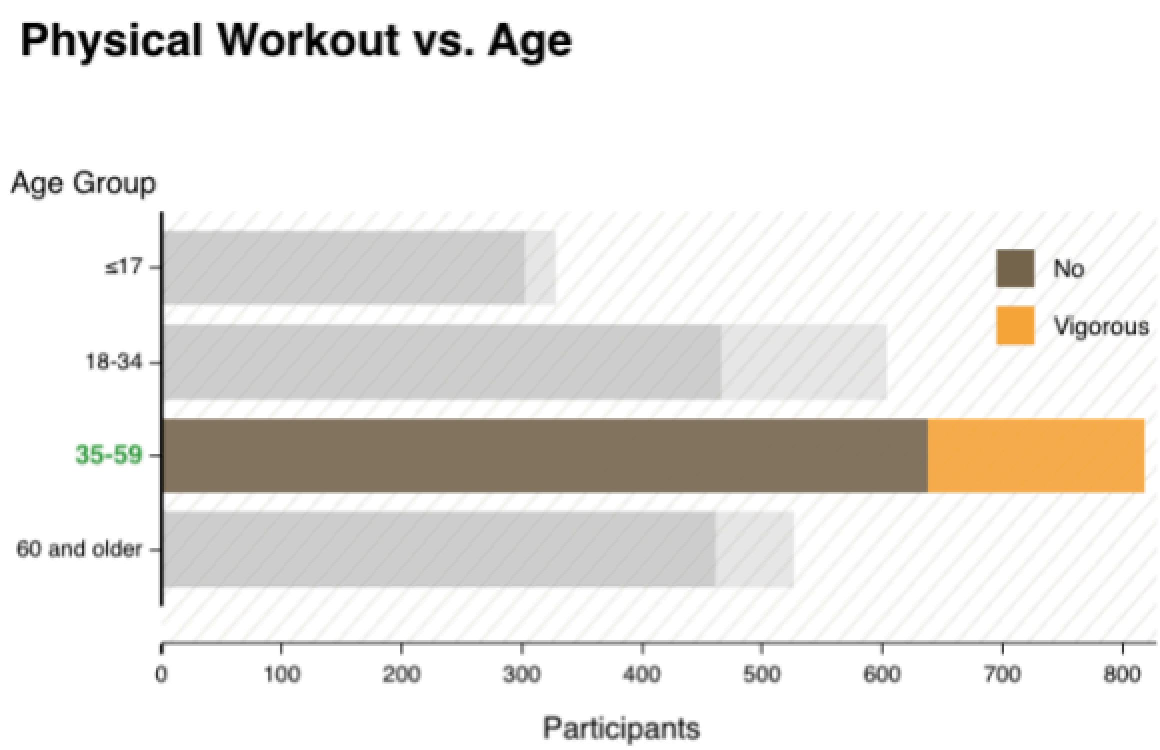
In this example the combined participant count for both male and female is 2278, and when the user clicks the female slice the participant count changes to reflect only the count of 1165 participants for the slice group.



Selecting Donut Chart Slice Updates Participant Count Interactively

Physical Workout Level vs Age Group Stacked Bar Chart - Age Group Categorical Interaction

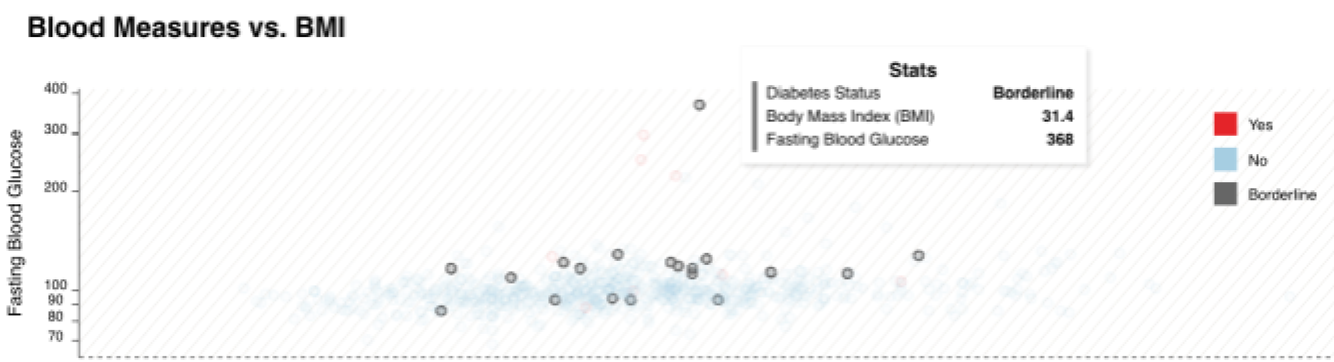
The dataset provides information about a single participant, and when we initially built this chart, we displayed aggregated amounts of distinct participant count per individual age year. To provide a better user dashboard experience and help categorize the participants' ages, we decided to group ages defined for specific ranges based on diabetic changes over time. The user can select an age group and see percentage changes in the donut chart (Participants by Gender) slice percentages and filtering on the joining scatter plot charts (Blood measures vs. BMI and Waist Circumference vs. BMI) for that age group. The legend swatches (No, Vigorous) provide an additional way to filter the data for all charts.



Stack Bar Chart Age Group Categorical Interaction

Blood Measures vs BMI - Single Point Mouse Over Interaction

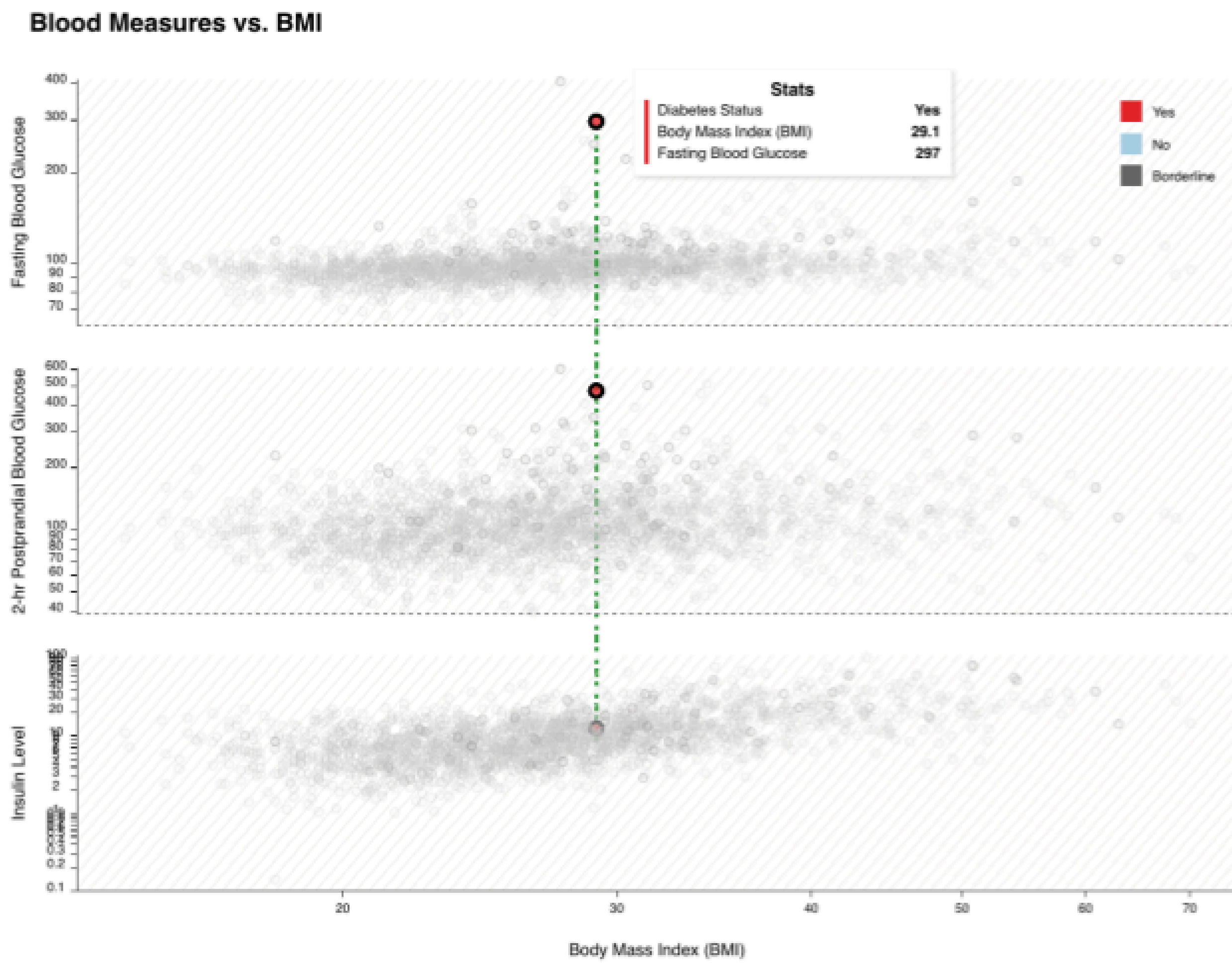
When the user’s mouse hovers over an individual point in the scatter plot two interactions occur. The other similar legend key grouped values (ie. Yes, No, or Borderline diabetic status) are displayed while the additional points are dimmed (opacity set to a low percentage value). This helps the user see these grouped values easier based on diabetic status. The second interaction is that a tooltip is displayed to reflect metrics for the individual point, providing additional insight to the user without having them reference the axes markers or legend color information.



Scatterplot Point Mouse Over Interaction for Single Point: Tooltip Shown and Filtering of Points

Blood Measures vs BMI - Single Point Click Interaction

When the user's mouse clicks a single point all other non-related points are dimmed (opacity set to a low percentage value) and the radial for the dot changes in color (ie. green) and size (ie. scaled up from normal size) to help the user focus on the selected point. Additionally, the related scatter plot chart points for the same participant are selected and these points are linked using a dashed `<line>` element and all chart tooltips are displayed. This helps tie related scatter plot values for the same participant between the three charts to help show a relationship among the points.



Single Point Click Interaction With Additional Related Scatterplot Chart Values

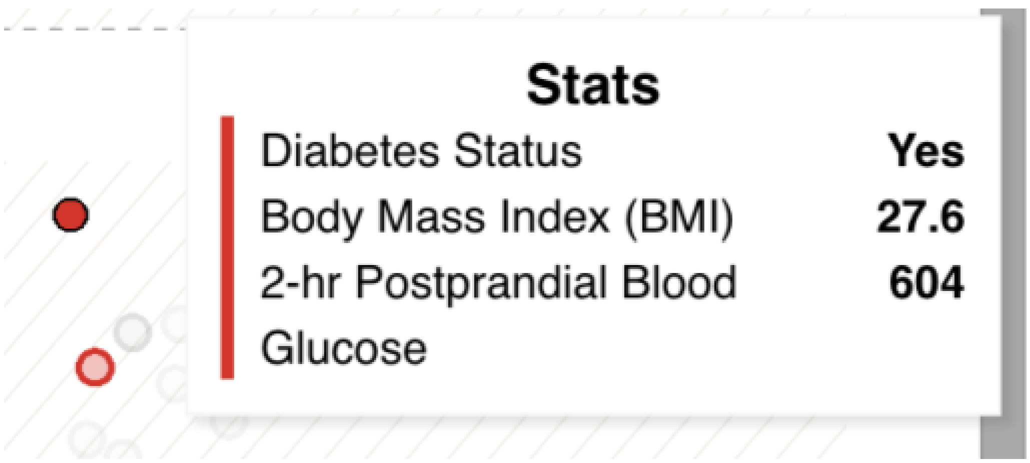
Legend

In order to generate a standard marker color key with names, several constant ordinal D3JS scalers ([Swatches]) were defined in a single file and used in multiple charts, legends, and shapes within the dashboard idiom. SVG Elements of `<rect>` and `<text>` were used to construct a legend. Clicking a color swatch filters the chart dots or bar chart SVG shapes to only display the selected legend key value; all other legend types not selected are dimmed (opacity set to a low percentage) to help the user focus on the selected legend key value.

Legend Color Marker Examples Used on Dashboard

Marks
Ordinal D3JS color scales (Colors) were defined globally and used throughout all charts for consistency between legend swatches, data shape indicators like points and bars, and the tooltip left vertical bar line. A tooltip component (Tooltip) was created and used when the user performs a mouse hover over for a specific scatter plot point or bar chart subgroup SVG shape element.

The example below shows a tooltip showing statistical metrics for the scatterplot point for the Blood Measures vs. BMI chart. These values help the user gather additional information about the specific chart point rather than looking at the surrounding legend or axes tick indicators for that information. Also, in this example the mouse hovered point has the circle shape fill color at full opacity while the other points show dimmed (low opacity value).



Tooltip Example Displaying Stats for Scatterplot Point

Evaluation

Lessons Learned From Data Visualizations
There are so many directions we could have gone with creating a visualization using this same dataset. After looking at the NHANES dataset, we started building charts that could show if participants had diabetes. Since we are not health professionals, we researched on the information health metrics signifies and further explored different datasets for fields like waist circumference,waist-to-hip ratio, high cholesterol level, Blood pressure and A1C metric.

Dashboard Questions Answered Through Graphs
As our video demonstrates, we are able to answer the following questions through interactions of the charts in our visualization.

- Which age group has more indicators of diabetes?
- At 35 and older, do participants who perform vigorous exercise have less indicators of diabetes?
- How many women 60 and above who do vigorous exercise have diabetes status “Yes”?
- How many men 60 and above who do vigorous exercise have diabetes status “Yes”?
- Which gender engages in physical exercise more at different age groups?
- Does waist circumference factor into diabetes risk for {women/men} participants at age group {60 and older} who do not exercise?

Future Improvements for Dashboard
Here are some updates that we would like to improve upon in future changes to the dashboard to include interaction and additional charts.

Additional Interactions
Here are some improvements for interaction for the existing charts on our dashboard. This will allow us to answer the questions proposed in our initial dashboard design.

- Physical Workout vs. Age
    - **Is there any difference in blood measures vs. BMI among the participants who exercise vigorously or not across all age groups?**
      Select the exercise level legend (No, Vigorous) to filter data on other dashboard charts. This helps see diabetes indicators within the Blood Measures vs. BMI chart for all age groups. The user can further filter by age group or gender if needed.

Additional Data and Geographic Visualizations
The data that we are using on the dashboard reflects NHANES 2013-2014, however, there are additional years that we could bring to use in addition to the geospatial information for finding a trend in blood measure metrics which could result in diabetes indications. Also, the data shows information regarding food spending habits (ie. grocery store, eating out) and the dashboard could show how that affects participants' health.

References

1. N. NA. "National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset," UCI Machine Learning Repository, 2019. [Online]. Available: https://doi.org/10.24432/C5BS66.
2. Dinh, A., Miertschin, S., Young, A., & Mohanty, S.D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making, 19.
3. Questionnaire Data for Specific Columns in Dataset
   2013-2014 Questionnaire Data - Continuous NHANES (cdc.gov)
4. Diabetes DIQ_H (cdc.gov)
5. Physical Activity PAQ_H (cdc.gov)
6. Oral Glucise Tolerence Test (OGTT_H): OGTT_H(cdc.gov)
7. How much physical activity do I need? CDC: https://www.cdc.gov/physical-activity-basics/benefits/index.html
8. QuickStats: Percentage of Adults Aged ≥18 Years with Diagnosed Diabetes, by Urbanization Level and Age Group — National Health Interview Survey, United States, 2022. MMWR Morb Mortal Wkly Rep 2024;73:49. DOI: http://dx.doi.org/10.15585/mmwr.mm7302a5