What lifestyle or socioeconomic factors can help predict diabetes?

Why we selected this topic?

We chose this topic because we found a robust dataset that should help us achieve our goal.

Description of data.

Our data has over 250,000 records and 22 columns. This data is from the 2015 health survey conducted annually by the CDC.

Questions we hope to answer with the data.

With this large of a dataset we hope to see what factors can predict diabetes. Secondarily, we want to examine how the results are affected by age and gender.

Technologies we will use.

Github for collaboration.

Supervised learning using logistic regression.

Postgres for our database.

AWS for web hosting.

Tableau for our dashboard.

Unsupervised Machine Learning to determine trends among Gender, Age and BMI.

Data Science Capstone 2021 Project

Team Members:
Peter Tsivis, John Murphy
Kyle Norton, Sharda Raj

What lifestyle or socioeconomic factors can help predict diabetes

- Selected topic
  - Diabetes

- Reason they selected the topic
  - To determine what lifestyle and socio-economic attributes indicate the likelihood of developing diabetes.
  - We found a robust data set that could help us achieve our goals.

Technologies used

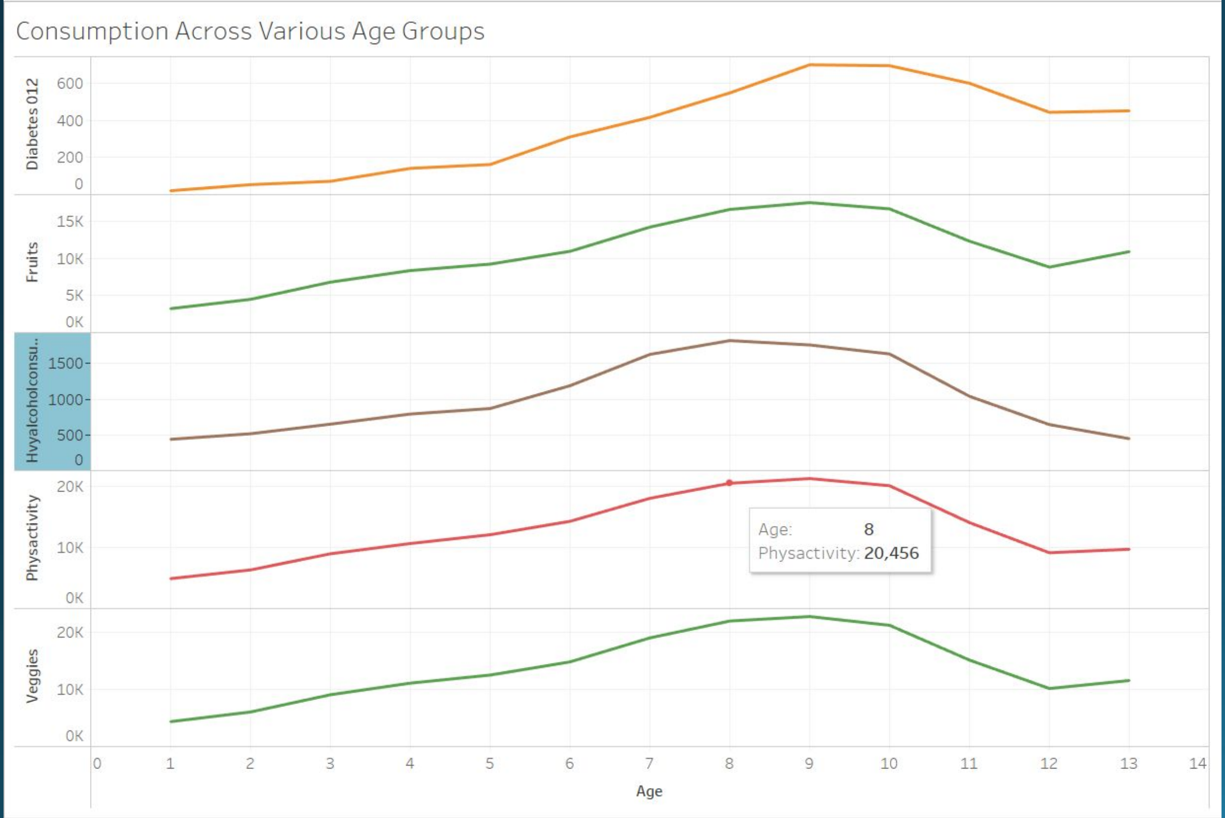
Systems Used

- Postgres SQL, Tableau,AWS
- Description of interactive elements-
  - Gender
  - Age
  - BMI
  - (the above based on linear regression/ logistical regression done)

Machine Learning

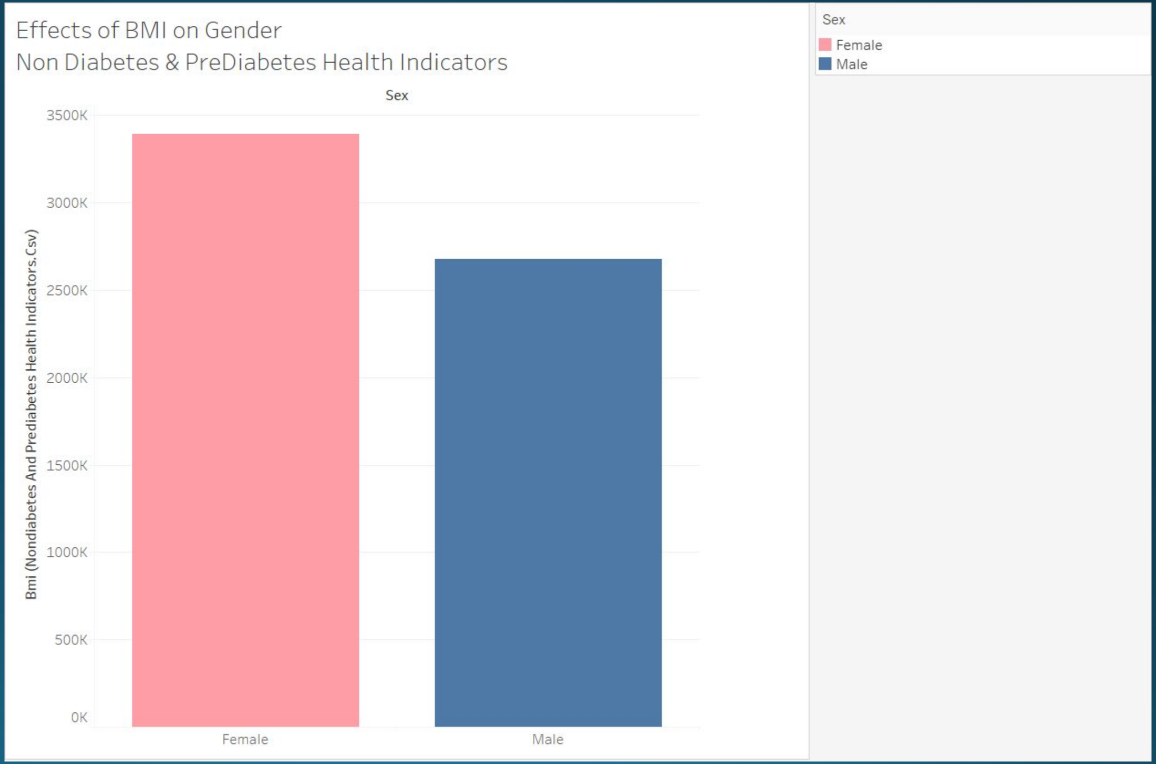
- Unsupervised Machine Learning to determine trends among Gender, Age and BMI.

Age

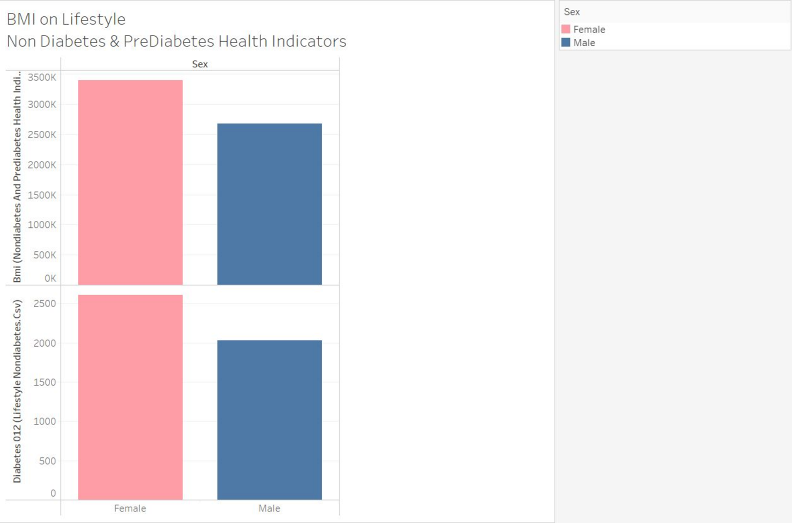


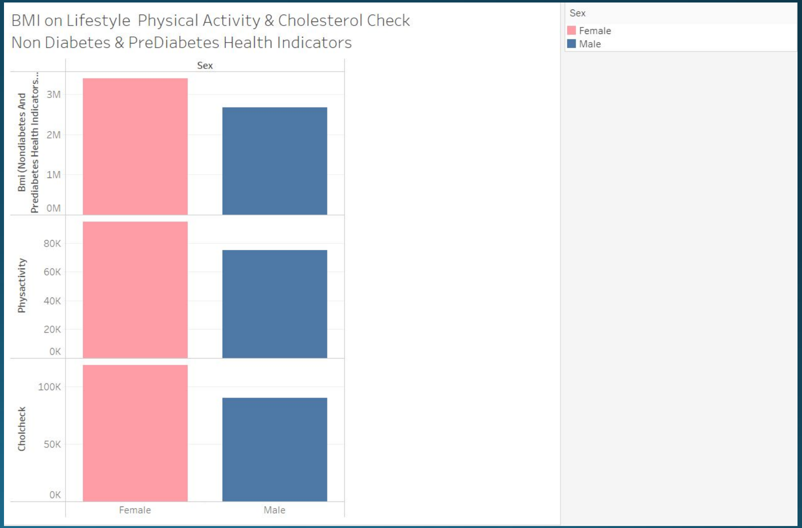
- Between ranges 7-10 (Ages 50-54, 55-59, 60-64 & 65-69) the highest amt of alcohol is consumed.
- Diabetes is the highest amongst age group ranges 9-10 (60-64, & 65-69)
- Fruit consumption decreases among ages 65-74 and increases among ages 75-80+.

BMI- Nondiabetic & Prediabetic



BMI- Lifestyle



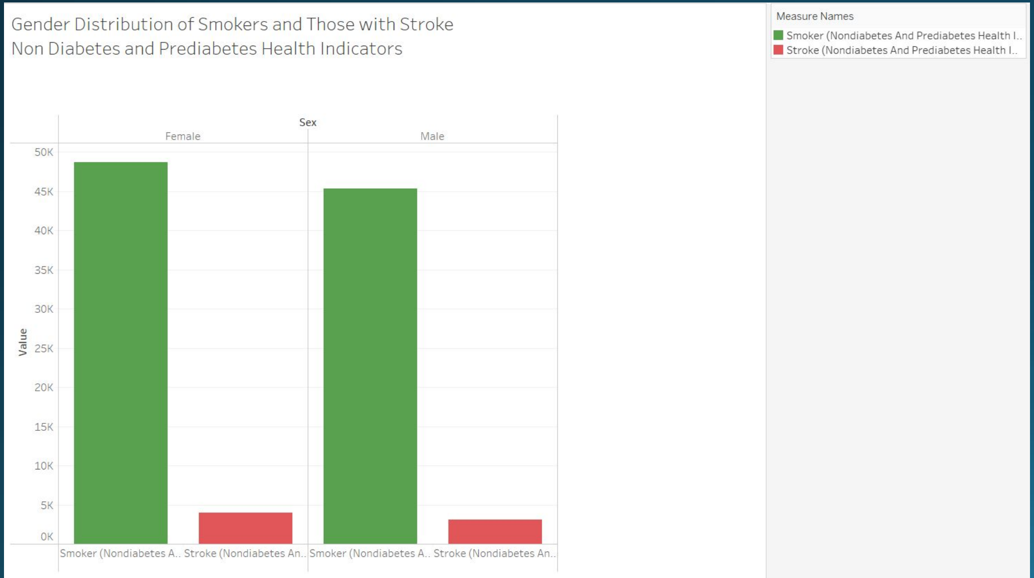


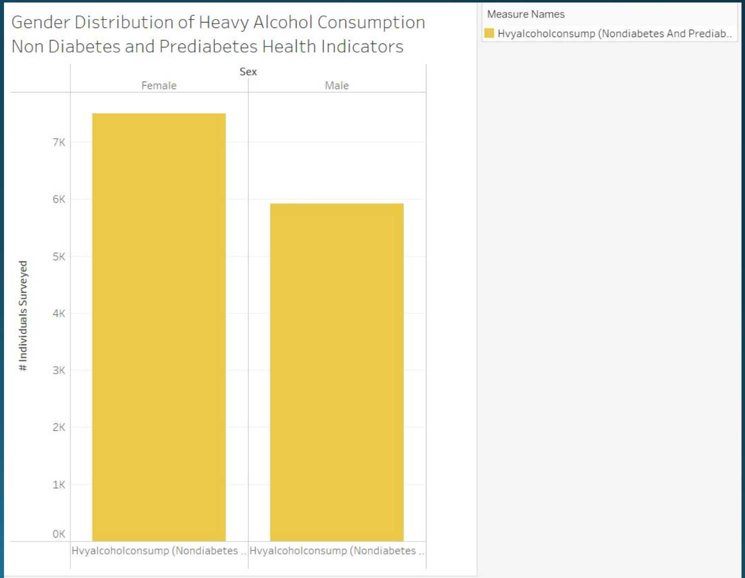
BMI- Socioeconomic

Gender- NonDiabetes and Pre-Diabetes

- More Females than Males are Smokers.
- About the same # of Males and Females have Strokes.

More Females than Males consume Heavy Alcohol.



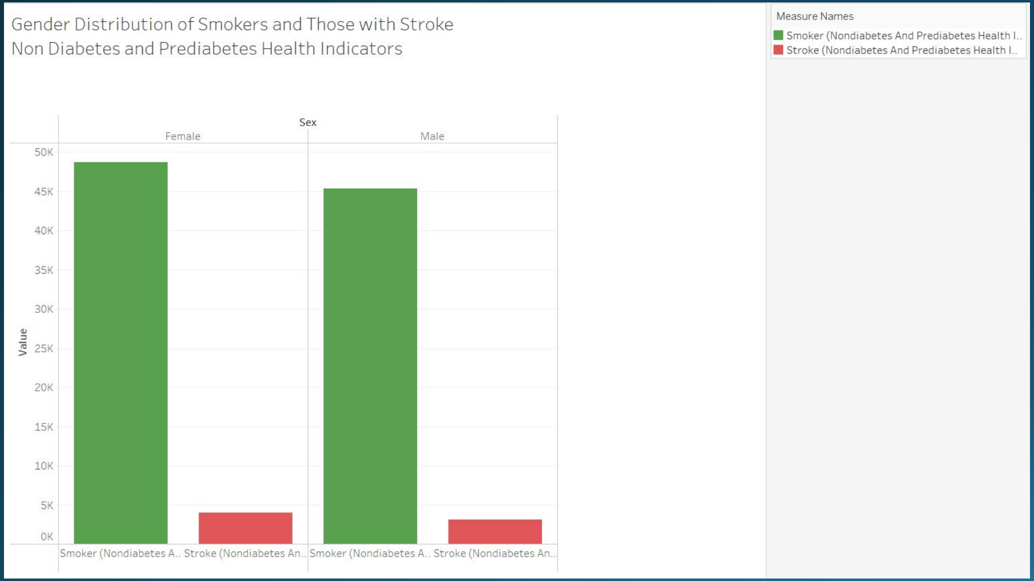


Gender- NonDiabetes and Pre-Diabetes

- More Females than Males are Smokers.
- About the same # of Males and Females have Strokes.

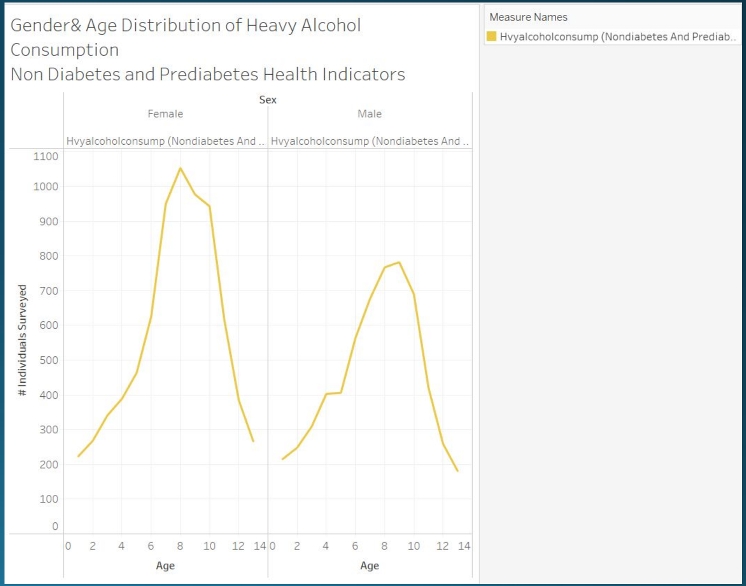
More Females than Males consume Heavy Alcohol.

Females in Age Groups 45-64 consume more Heavy Alcohol than Men.



Gender Distribution of Smokers and Those with Stroke
Non Diabetes and Prediabetes Health Indicators

Measure Names
🟩 Smoker (Nondiabetes And Prediabetes Health I..
🟥 Stroke (Nondiabetes And Prediabetes Health I..

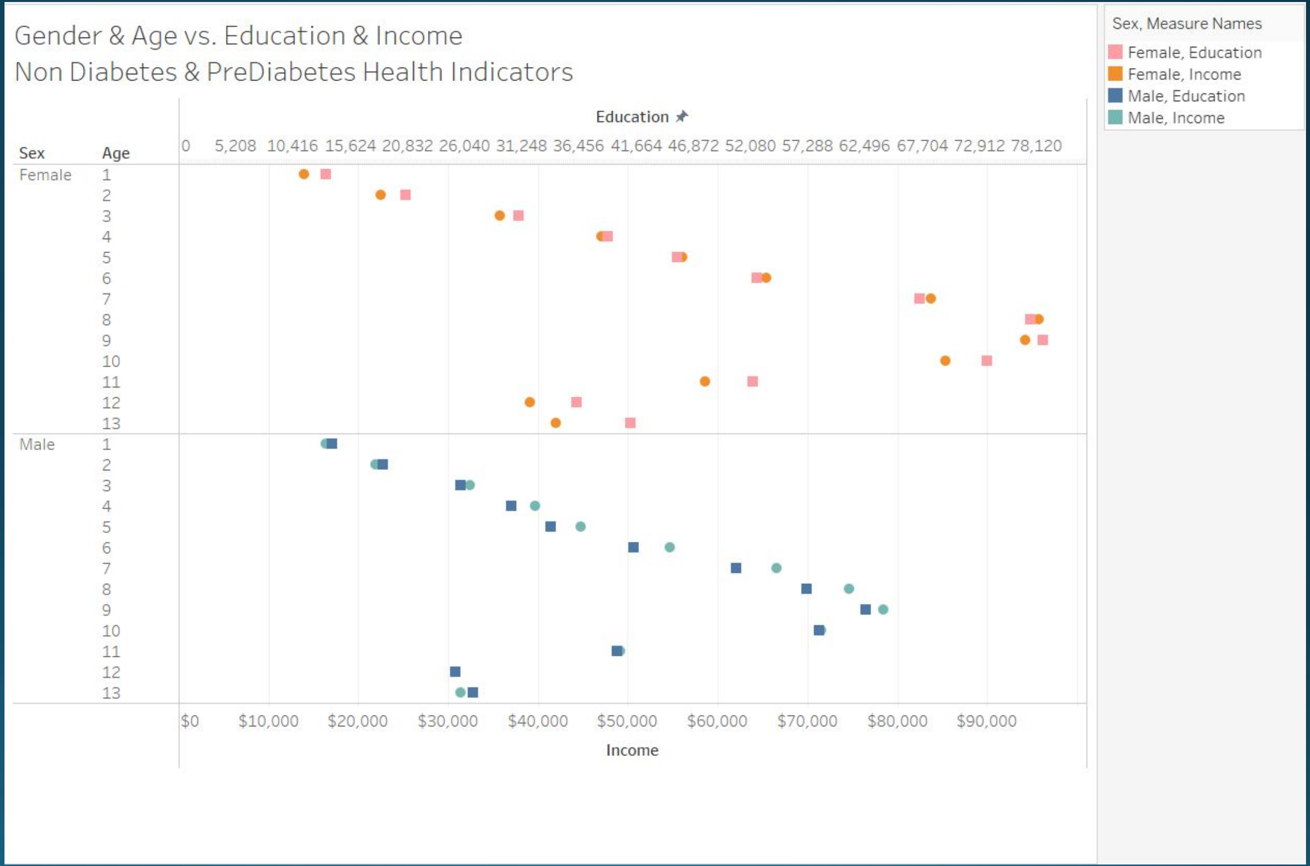


Gender& Age Distribution of Heavy Alcohol Consumption
Non Diabetes and Prediabetes Health Indicators

Measure Names
🟨 Hvyalcoholconsump (Nondiabetes And Prediab..

Gender- NonDiabetes and Pre-Diabetes

- More Females than Males are Smokers.
- About the same # of Males and Females have Strokes.

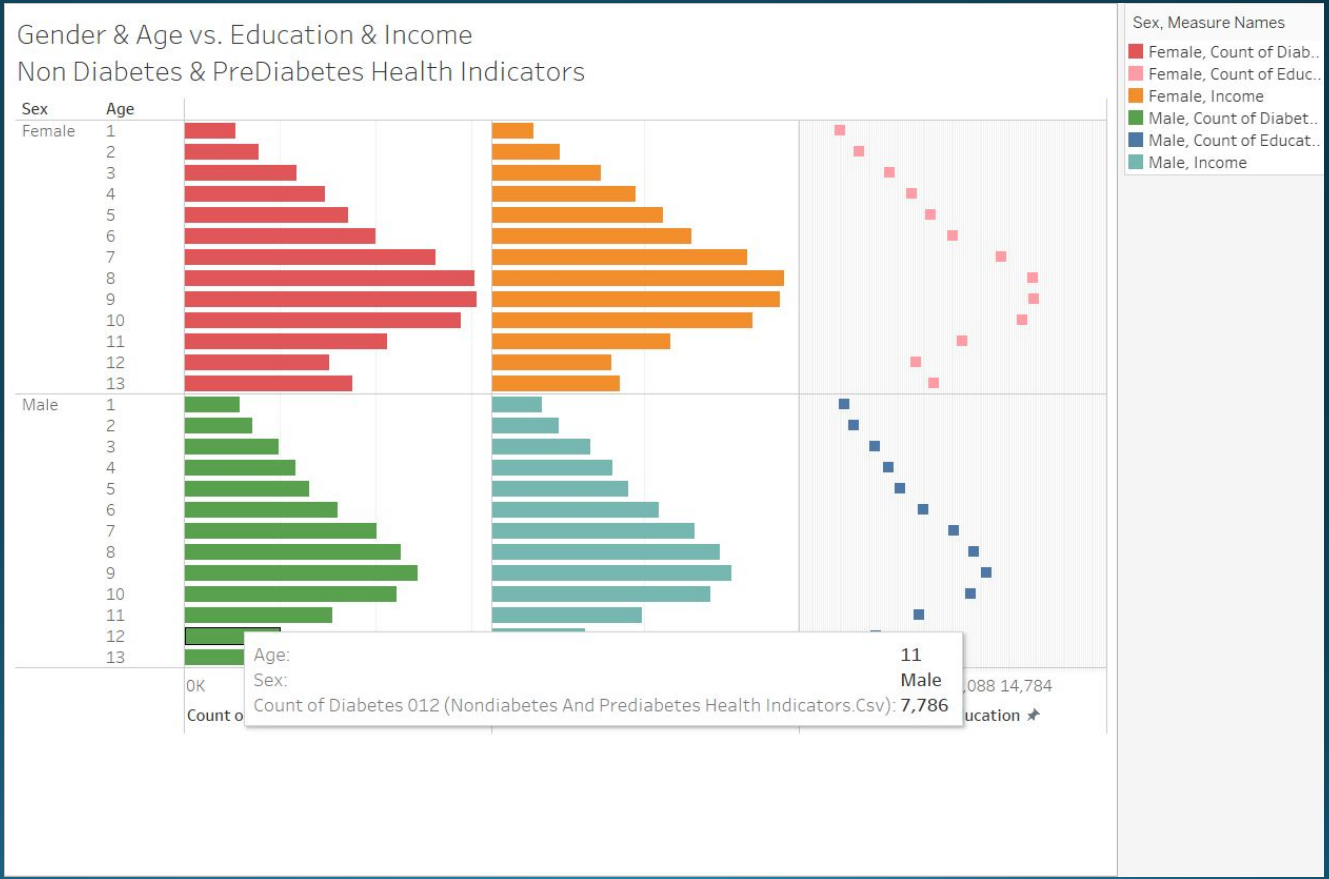


Gender & Age vs. Education & Income

Non Diabetes & PreDiabetes Health Indicators

Gender- NonDiabetes and Pre-Diabetes

- More Females than Males are Smokers.
- About the same # of Males and Females have Strokes.



Gender & Age vs. Education & Income

Non Diabetes & PreDiabetes Health Indicators

Age: 11

Sex: Male

Count of Diabetes 012 (Nondiabetes And Prediabetes Health Indicators.Csv): 7,786

Gender- Lifestyle

Gender- Socioeconomic

Income- NonDiabetes & Prediabetes

Income- Lifestyle

Income- Socioeconomic

Machine Learning

- Insert Machine Learning Screenshots here

Conclusions

Presentation- Mastery

- ✓ Selected topic
- ✓ Reason why they selected their topic
- ✓ Description of their source of data
- ✓ Questions they hope to answer with the data
- ✓ Description of the data exploration phase of the project
- ✓ Description of the analysis phase of the project
- ✓ Technologies, languages, tools, and algorithms used throughout the project
- ✓ Result of analysis
- ✓ Recommendation for future analysis
- ✓ Anything the team would have done differently

Machine Learning Model- Mastery

- ✓ Description of data preprocessing
- ✓ Description of feature engineering and the feature selection, including the
- team's decision-making process
- ✓ Description of how data was split into training and testing sets
- ✓ Explanation of model choice, including limitations and benefits
- ✓ Explanation of changes in model choice (if changes occurred between the
- Segment 2 and Segment 3 deliverables)
- ✓ Description of how model was trained (or retrained, if they are using an
- existing model)
- ✓ Description and explanation of model's confusion matrix, including final
- accuracy score
- Additionally, the model obviously addresses the question or problem the team
- is solving.
- Note: If statistical analysis is not included as part of the current analysis,
- include a description of how it would be included in the next phases of the
- project.

Database

- ✓ Database stores static data for use during the project
- ✓ Database interfaces with the project in some format (e.g., scraping updates
- the database, or database connects to the model)
- ✓ Includes at least two tables (or collections, if using MongoDB)
- ✓ Includes at least one join using the database language (not including any
- joins in Pandas)
- ✓ Includes at least one connection string (using SQLAlchemy or PyMongo)
- Note: If you use a SQL database, you must provide your ERD with
- relationships.

Dashboard

- The dashboard presents a data story that is logical and easy to follow for
- someone unfamiliar with the topic. It includes all of the following:
- ✓ Images from the initial analysis
- ✓ Data (images or report) from the machine learning task
- ✓ At least one interactive element
- Either the dashboard is published or the submission includes a screen capture
- video of it in action.

GitHub

- Main Branch
- All code in the main branch is production-ready.
- All code is clean, commented, easy to read, and adheres to a coding standard
- (e.g., PEP8)
- Main branch should include:
- ✓ All code necessary to perform exploratory analysis
- ✓ All code necessary to complete machine learning portion of project
- ✓ Any images that have been created (at least three)
- ✓ Requirements.txt file
- README.md
- README.md must include:
- ✓ Cohesive, structured outline of the project (this may include images, but
- should be easy to follow and digest)
- ✓ Link to dashboard (or link to video of dashboard demo)
- ✓ Link to Google Slides presentation
- Note: The descriptions and explanations required in all other project
- deliverables should also be in your README.md as part of your outline, unless
- otherwise noted.
- Individual Branches
- ✓ At least one branch for each team member
- ✓ Each team member has at least four commits for the duration of the final
- segment (16 total commits per person)

Description of the data exploration phase of the project.

Description of the analysis phase of the project.

We created 4 logistic regression models.

One of these was for the full dataset and the other three were divided up into health indicators, socioeconomic factors, and lifestyle factors.

We used oversampling, undersampling and combination methods.

Results for the three subsets were not impressive. Our best results came from using the full dataset.

The classification method worked better than the logistical regression on the full dataset.

Recommendations for future analysis

A larger dataset would be preferred for future analysis.

It may be interesting to have two datasets with the same individuals maybe 5 years apart.

The effects of geography on diabetes may also be an area of future analysis.

Things we would have done differently.

Begun the work on our dashboard earlier.

Do a better job of integrating Tableau into our work.