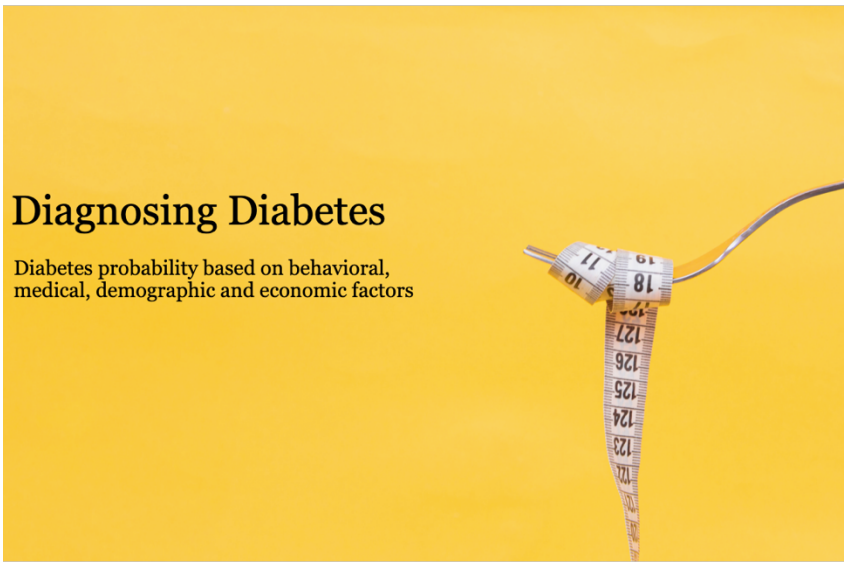
MGT 6203 Team 29 Final Report

Olivia Watson (owatson6),
Selda Kocaman (skocaman7),
Kelly Cristina Ribeiro Yogui (kyogui3),
Thu Thi Diem Le (tdiem6)
Hande Pehlivan (hpehlivan3)



Contents

Background

Nearly 10% of Americans have type II diabetes [1], costing each almost ten years of their life and the U.S. over $325 billion annually. ADA declares that, *"The total estimated 2017 cost of diagnosed diabetes of $327 billion includes $237 billion in direct medical costs and $90 billion in reduced productivity. […] People with diagnosed diabetes incur average medical expenditures of $16,752 per year, of which about $9,601 is attributed to diabetes. On average, people diagnosed with diabetes have medical expenditures approximately 2.3 times higher than what expenditures would be in the absence of diabetes."* [2] The American Diabetes Association (ADA) states that "*Care for people with diagnosed diabetes accounts for one in four health care dollars in the U.S.*" [2] Analyzing diabetes risk factors that impact a person's probability of having diabetes can be useful to improve healthcare programs, lower the U.S.'s healthcare burden, enhance the private sector's productivity, and improve U.S. citizens' wellbeing. The private sector could save $90 billion dollars from improved productivity, reduced absenteeism, and reduced insurance costs. Governments could re-allocate ¼ of all health care dollars to solving other crises or reduce taxes to reflect reduced health care needs. [2]

Objective

Ultimately, we seek to identify behavioral and societal diabetes risk factors to improve health care programs, lower the U.S.'s healthcare burden, improve the private sector's productivity, and improve U.S. citizens' wellbeing. In this project, we set out to identify the factors associated with a diabetes diagnosis and use modeling to accurately predict someone's likelihood of having diabetes. Understanding the main factors which may lead to a diabetes diagnosis is paramount to our society's survival. For citizens, this means extending their life expectancy, avoiding nearly $10k in diabetes associated costs per year, and enhancing their health. To do this, we will use the widely accepted CDC – 2021 Behavioral Risk Factor Surveillance Survey (BRFSS) dataset [3], the Annual Cost of Living dataset [4], and the US Populations by State [5] information.

Initial Hypotheses

We anticipated that the most significant variables would be related to obesity, diet, blood pressure, cholesterol, exercise, and cost of living based on past academic studies and desk research. [8]

We hypothesized states with lower costs of living may suffer from higher diabetes density: lower- and middle-income areas, associated with lower costs of living, often suffer from higher rates of disease. Home environment, unhealthy behaviors, obesity, stress, and medication accessibility are a few risks contributing to a higher incidence of diabetes among low-income households.

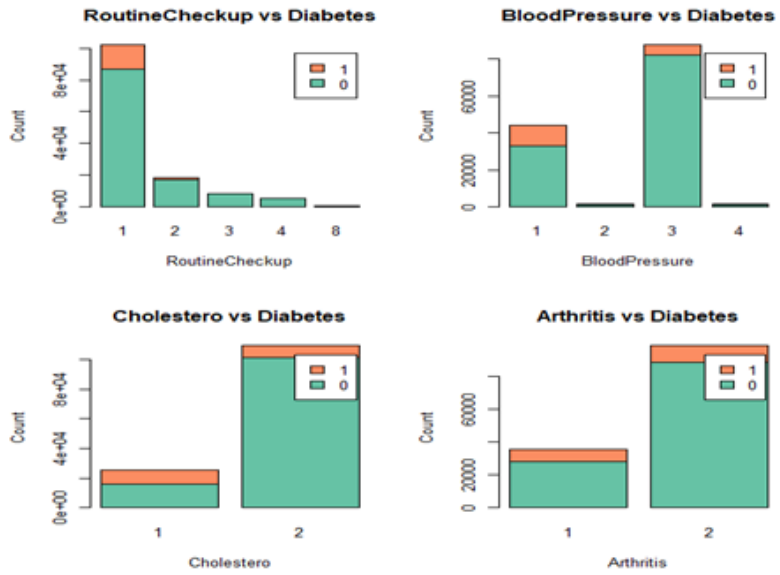


Figure 1: Bar graphs illustrating the occurrence of Diabetes against some other factors.

Academic Research Paper Review

Social Determinants of Health and Diabetes: A Scientific Review [6] explores the impact of socioeconomic status (SES) on diabetes onset and progression, highlighting the importance of social determinants of health in shaping health outcomes. The prevalence of diabetes is higher among individuals with lower income and educational levels. The neighborhood environment and food insecurity also significantly influence diabetes outcomes. The article also emphasizes the importance of health insurance in accessing diabetes screenings/care and the substantial cost burden associated with diabetes care. Investigating the impact of the cost of living on diabetes occurrence is crucial, given the high healthcare costs and the increased risk of undiagnosed diabetes among uninsured individuals. Living in a resource-deprived environment also contributes to disparities in diabetes risk, diagnosis, and outcomes.

Epidemiology of Diabetes and Diabetes-Related Complications [7] provides an overview of the state of Diabetes, the types of diabetes, the non-modifiable and modifiable risk factors, and a review of co-morbidities and associated healthcare costs. Notably, the work underscores the lag in diagnosis and the number of people who may go undiagnosed: for every 10 diagnosed with diabetes, there is an estimated 3-4 who go undiagnosed. Similarly, there is a large proportion of the population described as pre-diabetic. This 25% of the population described as pre-diabetic is critically important to understand and of interest for further research. Much like other articles and papers we've reviewed, modifiable risk factors include diet and physical activity, and non-modifiable risk factors include age, race, family history, and low birth weight.

These research studies and reviews have been considered in our datasets and variable selection for this work.

Methodology

Data Sources

The data sources used in this analysis are:

1. CDC - 2021 BRFSS Survey Data and Documentation [3]
   BRFSS dataset includes behavioral, medical condition, and demographic data. More on the background of the BRFSS survey can be found here.
2. Cost-of-living index [4]
   The 2022 Annual Average Cost of Living table has the Cost-of-Living Index and Grocery, Housing, Utilities, Transportation, Health, and Misc. values for US states.
3. US Populations by State [5]
   The 2023 US Population table contains information about 2023 population, 2022 population, Growth and Density Per Miles.

In the *Datasets* section at the end of this report, images of the first rows and links to download the datasets are provided.

Data Processing & Cleaning

The overall Data Processing & Cleaning is described below. The code is attached to this report and also may be accessed through the links provided in the *Codes* section at the end of this document.

1. Codebook Review: To have a better understanding of the variables in the BRFSS dataset, the Codebook Report [9] was reviewed. The original dataset has 438,693 rows and 303 columns. Diabetes is classified as:
   a. Yes (1)
   b. Yes, but female told only during pregnancy (2)
   c. No, and No, pre-diabetes, or borderline diabetes (3)

   We transformed the diabetes column and created our binary column with the Yes values (1) qualifying as having diabetes (1) in our data set.

2. Handling Missing Values: We did not impute missing values as the data is individual level healthcare inputs. Variables with more than 15 % missing values and non-relevant variables as informed by academic research are excluded. The data was then reduced to 124 columns. After re-plotting the missing values, we removed the rows with NAN value. This reduced our row count to 289,330.

3. Initial Variable Reduction: To select the variables for initial modeling, we used Lasso, PCA, correlation and the risk factors for diabetes according to our literature review. We analyzed multicollinearity and correlation, using 0.70 as threshold. After selecting the variables and the unique values for each one, the response with ‘don’t know’ or refused to answer are excluded from the analysis. Other variables were transformed in this process for efficiency in analysis including children at home and others.

4. Final BRFSS Data: After cleaning, the BRFSS dataset has 22 columns and 134,702 rows. The final variables include state, race, gender, general health, physical activity, employment, children at home, alcohol consumption, health insurance, BMI, age, blood pressure, education, income, among others. 12.8% of respondents in our cleaned dataset have diabetes.

5. Merging Data: BRFSS State column, *_STATE*, is numeric and the mapping from them to their respective labels is available in the BRFSS Codebook [9], as shown in *Figure A1* in the Appendix. To merge on State, we input the corresponding codes in the Cost-of-Living dataset, creating an additional field called *STATE_code*. A snippet is provided in *Figure A2* in the Appendix. The third dataset, US Population, has a string column called *state*, which matches with the *State* field from the Cost-of-Living dataset.

Exploratory Data Analysis

1. Variable Analysis: Before starting the analysis, columns were renamed. Variables' descriptions and values can be found in *Table A1 - Variables Description and values* in the Appendix. We examined how variable distribution varies between those diagnosed with diabetes and those who are not with summary statistics, correlation, bar charts, and box plots.

2. Initial Data Links: We found diabetes can be linked to various factors such as race, body mass index, physical activity, and others. Figure 2 shows the correlations between variables, where no two pairs of predictors are significantly correlated. Figure A5 in the Appendix shows four bar charts between predictors and diabetes variables. The bar charts show that variables have a broad distribution. Some of the variables including gender, children, and health insurance show minimal differences with diabetes. Other bar charts can be found in our GitHub repository in the Visualizations folder.

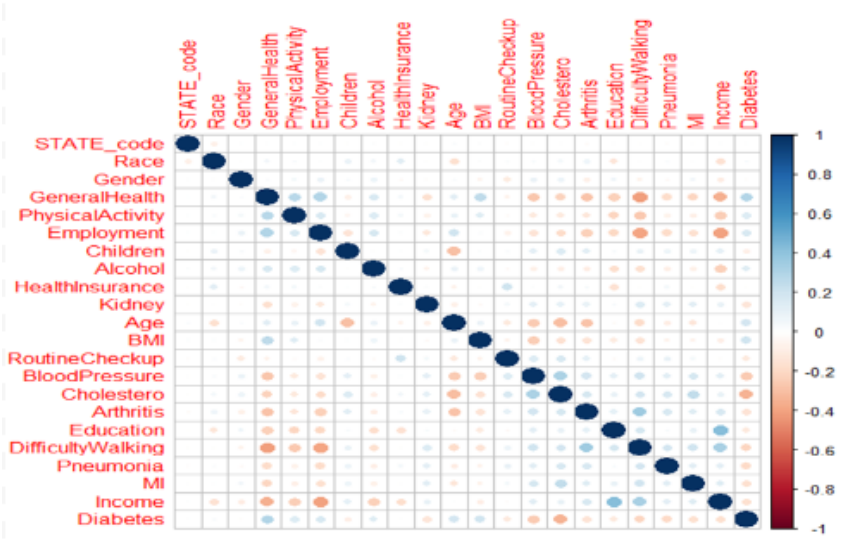


Figure2: Correlations between variables, where no two pairs of predictors are significantly correlated

3. Relationship Between Cost of Living and Diabetes: Figure 3 shows a clear trend that the states with a lower index for Cost of Living, Groceries, and Housing tend to have a higher ratio of diabetes. Groceries and Housing charts can be found in the Appendix (*Figures A3 and A4*). Home environment, unhealthy behaviors, obesity, stress, and medication accessibility are a few risks contributing to a higher incidence of diabetes among low-income households.

   The chart plots the density of diabetes with the cost-of-living index (average at 100). This is represented as the number of people with diabetes divided by the density of the state. The density is calculated by the population divided by size of the state. From the plots, we see that Montana, South Dakota, Nebraska, Kansas, New Mexico, Wyoming, and North Dakota generally have the highest density of diabetes and generally have lower associated costs. Whereas New York and Massachusetts and California typically have a lower density of diabetes with a higher cost of living.

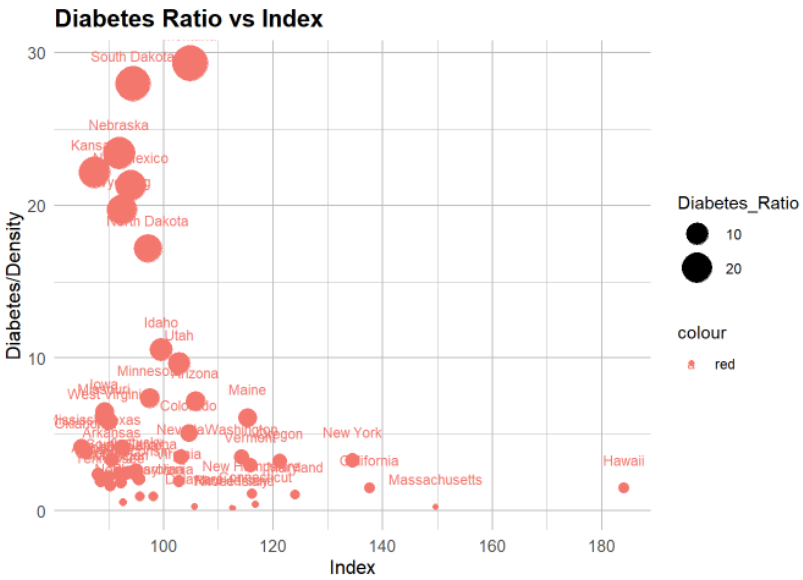


Figure 3: Graphs between Diabetes Ratio vs Cost of Living Index

Modeling

We built four classification models.

- Logistic Regression
- KNN (K=317)
- Naive Bayes
- Tree classification

For each, training data was created with 80% of our final data set and testing data with the remaining. Data is scaled for logistic regression and KNN. All variables were used to predict the

diabetes diagnosis. For each model to evaluate performance, confusion matrices and AUC scores were used. Since the data is imbalanced, we up-sampled the diabetes data to see changes on the metrics. Having diabetes is the minority class. We evaluated recall and precision as well, given they are not impacted by the imbalance. As we can see from Table 1 and 2, model accuracies decrease after up-sampling but sensitivity, F1 scores and almost all AUC scores have increased. With up-sampling, False Negatives are decreased for all models. Though accuracy is slightly decreased with up-sampling, up-sampling provides more better and more realistic read of the data.

Table 1: Models Result for Before Up-sampling

Model	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**AUC**	**F1**
Logistic Regression	0.882	0.263	0.974	0.596	0.263	0.849	0.365
KNN	0.898	0.263	0.992	0.834	0.263	0.628	0.4
Decision-Tree	0.884	0.26	0.976	0.617	0.26	0.618	0.366
Naïve Bayes	0.818	0.526	0.861	0.359	0.526	0.693	0.426

Table 2: Models Result for After Up-sampling

Model	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**AUC**	**F1**
Logistic Regression	0.716	0.834	0.698	0.29	0.834	0.848	0.431
KNN	0.817	0.897	0.805	0.405	0.897	0.851	0.558
Decision-Tree	0.856	0.328	0.934	0.424	0.328	0.63	0.37
Naïve Bayes	0.787	0.658	0.806	0.334	0.658	0.732	0.443

Model Interpretation

After comparing the model results, KNN is identified as the best model. Its sensitivity-(recall), F1, and AUC scores are better than other models. Importantly, it also has fewer False Negatives or Type I errors. Meaning, the model correctly identifies those with diabetes at a high rate. The model correctly predicts 89.7 % of those who have diabetes. However, it's precision score is low; 40.5 % of the predicted cases are positive. This highlights room to improve and lower Type II errors.

Discussion

Although KNN performed well, it does not provide a clear notion about each factor's effect on diabetes probability. To be actionable for businesses, healthcare professionals, and governments aimed at curing or reducing diabetes, it's important to understand which and how the factors are associated with diabetes. Logistic regression was the second-best model and can predict correctly 83.4 % of actual diabetes. However, only 29% of the correctly predicted cases actually turned out to be positive. While there's more work to be done to improve the precision, we analyzed the initial results to provide a preliminary look at variable impact. Physical Activity, Health Insurance, Education and Difficulty Walking were not significant in the Logistic Regression model. All remaining variables are significant as can be seen in Figure 4. General Health, Alcohol, Age, BMI, Routine Checkup, Blood Pressure, Cholesterol; Pneumonia, Income and Race have more weight than other variables. The Decision Tree model also identifies the same variables as important.

```
## Call:
## glm(formula = Diabetes ~ ., family = "binomial", data = diabetes_train_up)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9185  -0.7052  -0.2963   0.7244   3.4665
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.3495206  0.0065255 -53.562  < 2e-16 ***
## STATE_code         -0.0130953  0.0062971  -2.080   0.0376 *
## Race                0.1692141  0.0064670  26.166  < 2e-16 ***
## Gender             -0.0623620  0.0065380  -9.538  < 2e-16 ***
## GeneralHealth       0.5917156  0.0082054  72.113  < 2e-16 ***
## PhysicalActivity    0.0062463  0.0067038   0.932   0.3515
## Employment         -0.1059945  0.0077823 -13.620  < 2e-16 ***
## Children           -0.0504012  0.0069199  -7.284 3.25e-13 ***
## Alcohol             0.2019168  0.0065565  30.796  < 2e-16 ***
## HealthInsurance     0.0024629  0.0067053   0.367   0.7134
## Kidney             -0.0788467  0.0070235 -11.226  < 2e-16 ***
## Age                 0.3804814  0.0080784  47.099  < 2e-16 ***
## BMI                 0.4534080  0.0070029  64.746  < 2e-16 ***
## RoutineCheckup     -0.3296324  0.0083493 -39.480  < 2e-16 ***
## BloodPressure      -0.2450564  0.0066721 -36.728  < 2e-16 ***
## Cholestero         -0.6015478  0.0068599 -87.691  < 2e-16 ***
## Arthritis           0.0797508  0.0070236  11.355  < 2e-16 ***
## Education          -0.0081064  0.0071864  -1.128   0.2593
## DifficultyWalking  -0.0007267  0.0079666  -0.091   0.9273
## Pneumonia          -0.2899215  0.0065306 -44.394  < 2e-16 ***
## MI                  0.0290478  0.0069332   4.190 2.79e-05 ***
## Income             -0.1497492  0.0083310 -17.975  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 232157  on 168973  degrees of freedom
## Residual deviance: 157295  on 168952  degrees of freedom
## AIC: 157339
##
## Number of Fisher Scoring iterations: 5
```

Figure 4: Logistic Regression Summary. All the factors are statistically significant, except Physical Activity, Health Insurance, Education, and Difficulty Walking.

Logistic Regression modeling indicates that being obese, having poor health, and taking cholesterol meds are all associated with an increased likelihood of being diagnosed with diabetes. While non-modifiable, males and people of color have an increased probability of being diagnosed. These results have important implications for both individuals and organizations, healthcare institutions, and governments. With enhanced modeling, the variable results can help segment and identify who is at risk for diabetes and activities to partake in or avoid to prevent an ultimate diagnosis. The following are the odds ratios associated with a few key variables from the logistic regression:

	Odds Ratio from one scale point to the next	Interpretation
Cholesterol Medication	.55	Those taking cholesterol medication have a greater likelihood
General Health	1.8	Those who self describe with worse general health have a greater likelihood
BMI	1.6	Those with higher BMI have a greater likelihood
Pneumonia	.75	Those with or that have had pneumonia have a greater likelihood
Income	.86	Lower income participants have a greater likelihood
Gender	.94	Males have a greater likelihood

Conclusion

We set out to accurately identify individuals with Diabetes. Our modeling focus was on limiting Type II Error (False Negatives) and improving True Positives. The negative impact of settling for a model with higher Type II Error is great. Delaying diagnosis and treatment can result in significant challenges, impacting a person's quality of life and burden on the tax and healthcare systems. KNN and Logistic regression have the fewest False Negatives and a higher Recall percentage compared to the Decision Tree and Naive Bayes. Therefore, we chose KNN as the best model. We wanted to ensure those with Diabetes are diagnosed, and our selected model, KNN, adequately predicts those with the disease as having it.

Our model does indicate that a high number of people would be falsely identified as having diabetes when they in fact do not (Type I Error). Additional data transformation and modeling is required to lower this error.

Logistic regression also provides more information about the relationship between the variables and a positive diabetes diagnosis: general health, age, BMI, blood pressure, pneumonia, routine checkup, race, alcohol consumption and income are significant in predicting diabetes.

Moreover, the data supported our initial hypothesis that lower cost of living is associated with a higher density of diabetes ratio. To understand the real factors behind this effect, a deeper analysis is necessary, focusing on peoples' behaviors in different economic scenarios.

APPENDIX

Label: State FIPS Code
Section Name: Record Identification
Section Number: 0
Question Number: 1
Column: 1-2
Type of Variable: Num
SAS Variable Name: _STATE
Question Prologue:
Question: State FIPS Code

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Alabama	4,586	1.05	1.60
2	Alaska	5,493	1.25	0.23
4	Arizona	10,654	2.43	2.27
5	Arkansas	5,372	1.22	0.95
6	California	6,735	1.54	12.55
8	Colorado	10,476	2.39	1.88
9	Connecticut	8,341	1.90	1.17
10	Delaware	3,640	0.83	0.32

Figure A1: BRFSS Survey Codebook snippet showing the mapping between field values and their corresponding labels.

Rank	State	Index	Grocery	Housing	Utilities	Transport	Health	Misc.	STATE_code
1	Mississippi	85	92.4	67.4	89	91.9	97.7	91.6	28
2	Oklahoma	85.8	93.7	70.2	95.1	90.9	91.2	90.4	40
3	Kansas	87.5	93.7	71.1	98	95.6	100.4	91.6	20
4	Alabama	88.1	97.6	69.6	100.7	89.9	89.6	95	1
5	Georgia	88.6	94.6	75.6	90.3	89.8	94.6	95.1	13
6	Missouri	89.1	95.4	79.9	94.8	91.8	92.3	91.3	29
7	Iowa	89.2	99.5	71.5	93.7	95.8	100	94.8	19
8	Indiana	89.9	93.7	77.4	104	94	95.5	92.9	18
9	West Virginia	90	98.7	68.8	94.4	111.2	101.8	95.3	54
10	Tennessee	90.2	94.4	81.7	93.8	90.5	89.9	94.2	47

Figure A2: Cost of Living Dataset with the *STATE_code* column, created based on the Codebook shown in Figure A1.

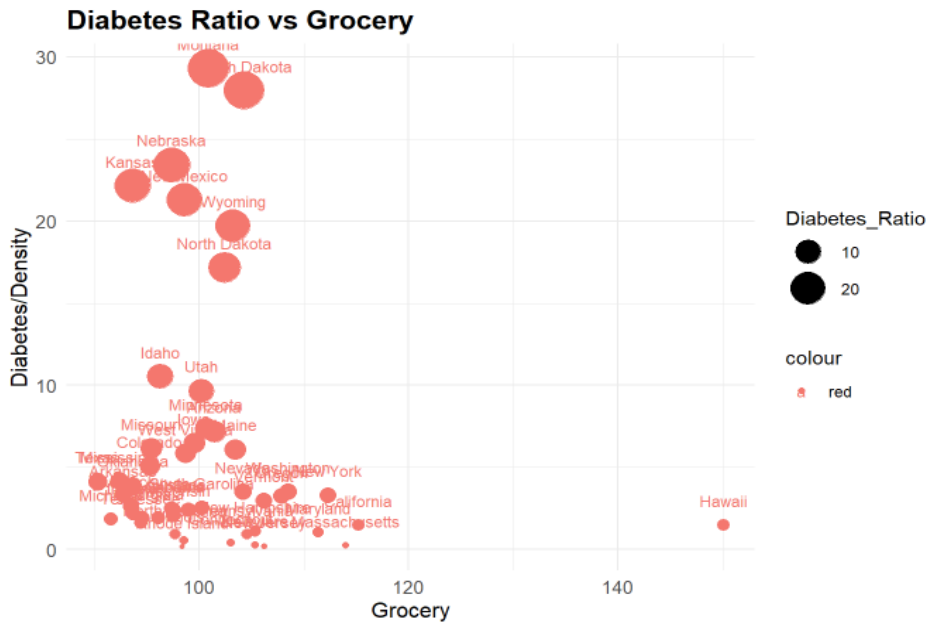


Figure A3: Diabetes Ratio (Diabetes/Density) vs Cost of Grocery Index per state. Like for the General Cost of Living, we can see low Diabetes Ratio for all values of Cost. However, the High Diabetes Ratio States concentrated on lower General Cost of Living and Grocery.

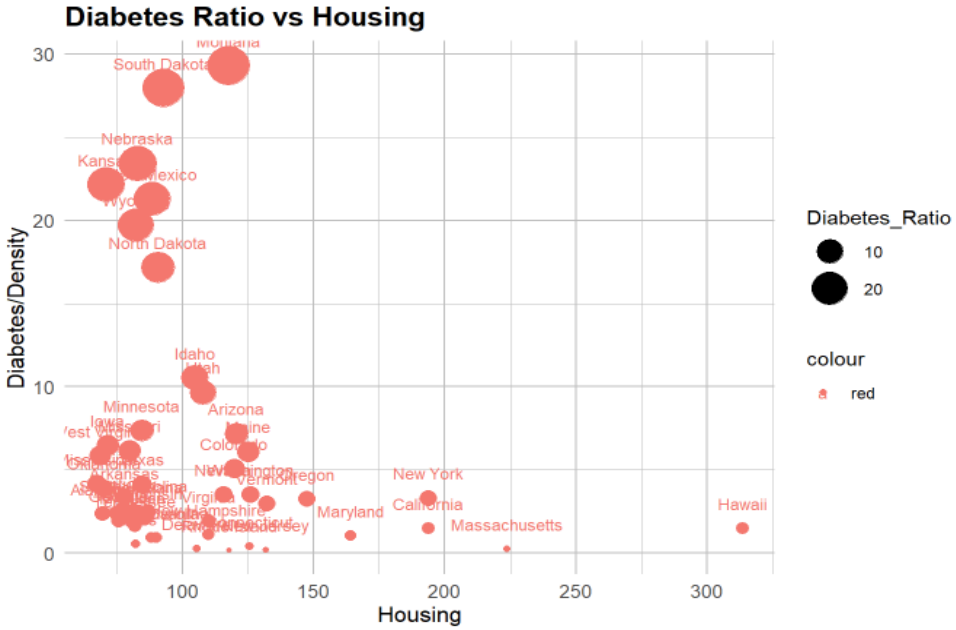


Figure A4: Diabetes Ratio (Diabetes/Density) vs Cost of Housing Index per state. Although some States had their positions in Cost ranking changed if compared to Figure A3, there is no strong effect, and the general trend remains the same.

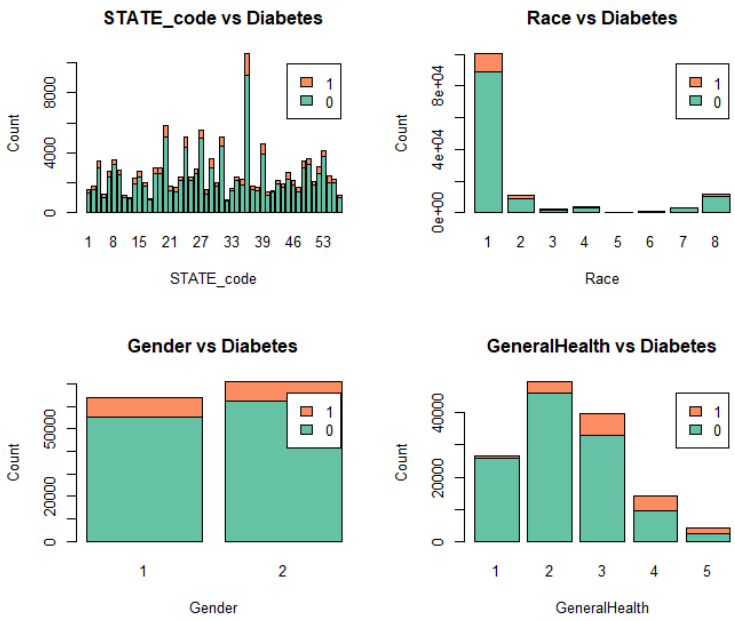


Figure A5: Histogram charts between state_code, Race, Gender, General Health vs Diabetes variables

Table A1: Variables Description and values

General Health	Would you say that in general your health is	1: Excellent, 2: Very good, 3: Good, 4: Fair, 5: Poor
Physical Activity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job	1: Had physical activity or exercise 2: No physical activity or exercise in last 30 days
Employment	Employment	1: Employed, 2: Self-employed, 3: No work >1 y, 4: No work <1 y, 5: Homemaker, 6: Student, 7: Retired, 8: Unable to work
CHILDREN	How many children less than 18 years of age live in your household?	1 - 87 Number of children, 88 None
Alcohol	Adults who reported having had at least one drink of alcohol in the past 30 days	1: Yes, 2: No
Health Insurance	Respondents aged 18-64 who have any form of health insurance	1: Yes, 2: No
Kidney	Ever told you have kidney disease?	1: Yes, 2: No
Age	Six-level imputed age category	1: Age 18 to 24 ,2: Age 25 to 34,3: Age 35 to 44 4 Age 45 to 54, 5: Age 55 to 64, 6: Age 65 >=

BMI	Computed body mass index categories	1: Underweight, 2: Normal Weight ,3: Overweight, 4: Obese
Routine Checkup	Length of time since last routine checkup	1: Within past year (anytime < 12 months ago) 2: Within past 2 years (1 year but < 2 years ago) 3: Within past 5 years (2 years but < 5 years ago) 4: 5 or more years ago, 18,490 4.21 4.83 8 Never
Blood Pressure	Ever Told Blood Pressure High	1 Yes 2 Yes, but female told only during pregnancy 3 No 4 Told borderline high or pre-hypertensive or elevated blood pressure
Cholesterol	Currently taking medicine for high cholesterol	1: Yes, 2: No
Arthritis	Told Had Arthritis	1: Yes, 2: No
Education	Computed level of education completed categories	1: Did not graduate high school 2: Graduated high school 3: Attended college 4: Graduated college
Difficulty Walking	Difficulty Walking or Climbing Stairs	1: Yes, 2: No
Pneumonia	Pneumonia shot ever	1: Yes, 2: No
MI	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)	1 Reported having MI or CHD 2 Did not report having MI or CHD
Income	Computed income categories	1 Less than $15,000 2 $15,000 to < $25,000 3 $25,000 to < $35,000 4 $35,000 to < $50,000 5 $50,000 to < $100,000 6 $100,000 to < $200,000 7 $200,000 or more
State Code		
Race	Computed Race-Ethnicity grouping	1 White only, non-Hispanic 2 Black only, non-Hispanic 3 American Indian or Alaskan Native only, non-Hispanic 4 Asian only, non-Hispanic 5 Native Hawaiian or other Pacific Islander only, Non-Hispanic 6 Other race only, non-Hispanic 7 Multiracial, non-Hispanic 8 Hispanic

Diabetes	(Ever told) you had diabetes	1 Yes 2 Yes, but female told only during pregnancy 3 No 4 No, pre-diabetes or borderline diabetes
Gender	Are you male or female?	1 Male 2 Female

References

[1] Type 2 Diabetes | CDC
https://www.cdc.gov/diabetes/basics/type2.html#:~:text=Healthy%20eating%20is%20your%20recipe,adults%20are%20also%20developing%20it.

[2] The Cost of Diabetes | ADA
https://diabetes.org/about-us/statistics/cost-diabetes#:~:text=People%20with%20diagnosed%20diabetes%20incur,in%20the%20absence%20of%20diabetes.

[3] CDC - 2021 BRFSS Survey Data and Documentation
https://www.cdc.gov/brfss/annual_data/annual_2021.html

[4] Cost-of-living index
https://meric.mo.gov/data/cost-living-data-series

[5] UP Population by State
https://worldpopulationreview.com/states

[6] Social Determinants of Health and Diabetes: A Scientific Review
Hill-Briggs, Felicia et al. “Social Determinants of Health and Diabetes: A Scientific Review.” *Diabetes care*, vol. 44,1 258–279. 2 Nov. 2020, doi:10.2337/dci20-0053, https://pubmed.ncbi.nlm.nih.gov/33139407/

[7] Epidemiology of Diabetes and Diabetes-Related Complications
Deshpande, Anjali D. " Epidemiology of Diabetes and Diabetes-Related Complications." *Physical Therapy*, Volume 88, Issue 11, 1 November 2008, Pages 1254–1264, https://doi.org/10.2522/ptj.20080020

[8] Preventing Chronic Disease
Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques
https://www.cdc.gov/pcd/issues/2019/19_0109.htm

[9] BRFSS Survey Codebook
BRFSS Codebook 2021 – Page 2
https://www.cdc.gov/brfss/annual_data/2021/pdf/codebook21_llcp-v2-508.pdf

Datasets

The datasets are available for download through the links below:

CDC-2021 BRFSS Survey Data
https://www.dropbox.com/s/t4e5cypfbe63jxf/LLCP2021.XPT%20.zip?dl=0

	X_STATE	FMONTH	IDATE	IMONTH	IDAY	IYEAR	DISPCODE	SEQNO	X_PSU	CTELENM1	PVTRESD1	COLGHOU
1	1	1	01192021	01	19	2021	1100	2021000001	2.021e+09	1	1	
2	1	1	01212021	01	21	2021	1100	2021000002	2.021e+09	1	1	
3	1	1	01212021	01	21	2021	1100	2021000003	2.021e+09	1	1	
4	1	1	01172021	01	17	2021	1100	2021000004	2.021e+09	1	1	
5	1	1	01152021	01	15	2021	1100	2021000005	2.021e+09	1	1	

Figure 3: BRFSS dataset first rows

Cost of Living Data
https://www.dropbox.com/scl/fi/3sm9u74ijs6001b94v2sl/Cost_of_living.xlsx?dl=0&rlkey=nim2iy5x30qwyjiuqfow4l9fe

Rank	State	Index	Grocery	Housing	Utilities	Transportation	Health	Misc.
1	Mississippi	85.0	92.4	67.4	89.0	91.9	97.7	91.6
2	Oklahoma	85.8	93.7	70.2	95.1	90.9	91.2	90.4
3	Kansas	87.5	93.7	71.1	98.0	95.6	100.4	91.6
4	Alabama	88.1	97.6	69.6	100.7	89.9	89.6	95.0
5	Georgia	88.6	94.6	75.6	90.3	89.8	94.6	95.1

Figure 4: Cost of Living dataset first rows

US Population by State
https://www.dropbox.com/s/3ux4nyuqj4tuj30/Population.csv?dl=0

Rank	State	2023 Population	Growth Rate	2022 Population	2010 Population	Growth Since 2010	% of US	Density (/mi²)
1	California	38,915,693	-0.29%	39,029,342	37,253,956	4.46%	11.66%	250
2	Texas	30,500,280	1.57%	30,029,572	25,145,561	21.3%	9.14%	117
3	Florida	22,661,577	1.87%	22,244,823	18,801,310	20.53%	6.79%	423
4	New York	19,496,810	-0.92%	19,677,151	19,378,102	0.61%	5.84%	414
5	Pennsylvania	12,931,957	-0.31%	12,972,008	12,702,379	1.81%	3.87%	289

Figure 5: US Populations dataset first rows

Codes

The README.md and Project Code files are attached to this report. Moreover, they can be accessed through the links:

README file
https://github.gatech.edu/MGT-6203-Spring-2023-Canvas/Team-29/blob/main/README.md

Project Codes
final_code.Rmd
https://github.gatech.edu/MGT-6203-Spring-2023-Canvas/Team-29/blob/main/Final%20Code/final_code.Rmd
final_code.html
https://github.gatech.edu/MGT-6203-Spring-2023-Canvas/Team-29/blob/main/Final%20Code/final_code.html