



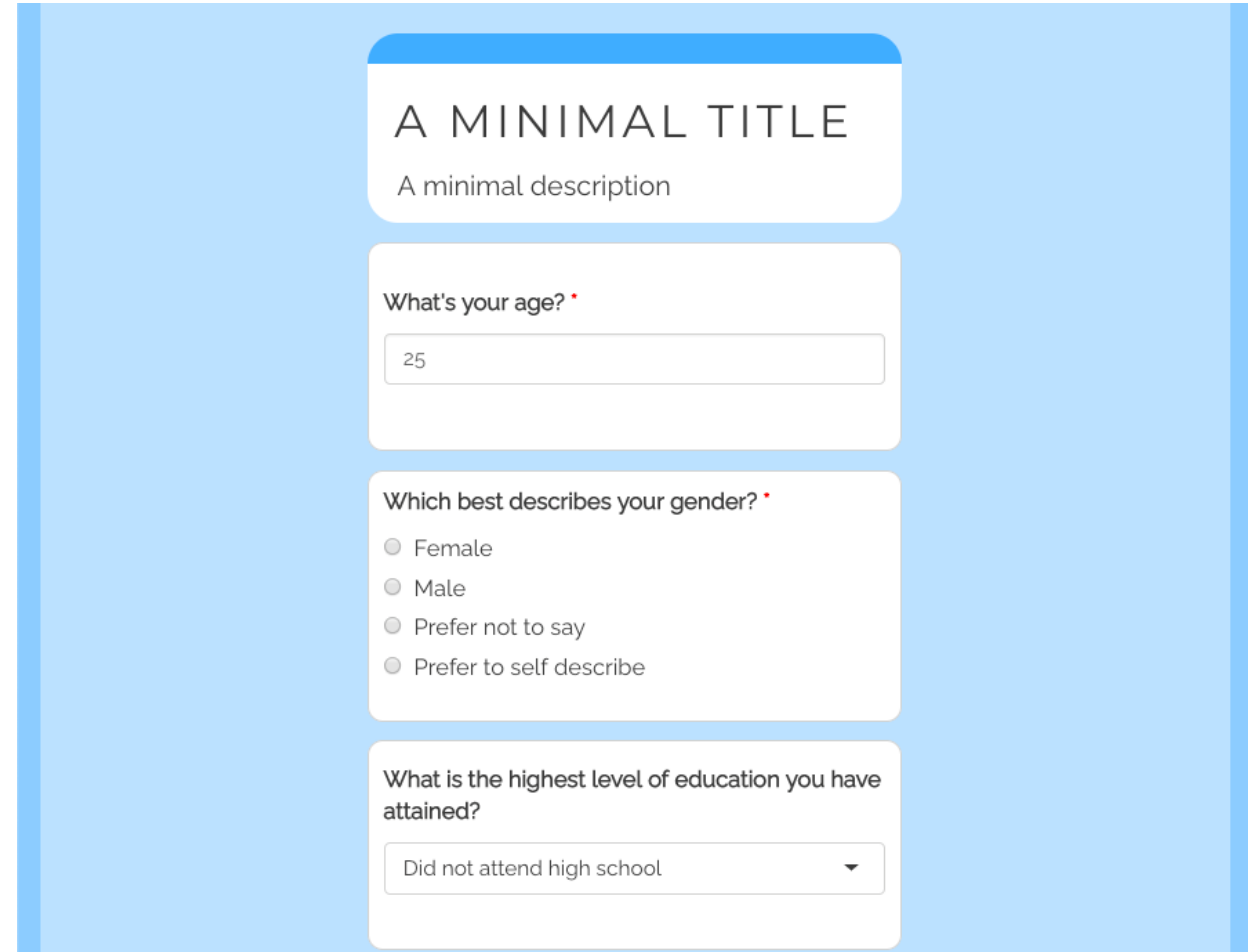
Project Proposal:

Experimentation of Machine Learning Techniques to Create a
Diabetes Risk Prediction Application

Matthew Vu

Overall Project Topic and Objective:

The topic of my capstone project is to experiment with several machine learning techniques, such as logistic regression, naïve bayes, support vector machines, k-nearest neighbor, and decisions trees, with the goal of creating a questionnaire application that can predict the risk of diabetes. Also, this capstone project will experiment with clustering the observations using Gower's distance and Partitioning Around Medoids (PAM) to see if clustering improves model accuracy.



A MINIMAL TITLE

A minimal description

What's your age? *

25

Which best describes your gender? *

☐ Female

☐ Male

☐ Prefer not to say

☐ Prefer to self describe

What is the highest level of education you have attained?

Did not attend high school ▼

Idea of what the questionnaire should look like using ShinyR. Source: <https://cran.r-project.org/web/packages/shinysurveys/vignettes/surveying-shinysurveys.html>

Research Questions:

1. What behaviors and habits are associated with diabetes risk?
2. What machine learning algorithms out of the following is most effective at predicting diabetes risk: logistic regression, naïve bayes, support vector machines, k-nearest neighbor, and decisions trees
3. Will clustering the data using Gower's distance and the Partition Around Medoids (PAM) clustering algorithm improve model accuracy?

Motivation and Importance for Research:

1. There still exists a high prevalence of diabetes, and many individuals do not know that they have diabetes.
 - a) According to the Center for Disease Control and Prevention's (CDC) National Diabetes Statistics Report, over 37 million Americans have diabetes; however, about 1 in 5 of these individuals do not know that they have it (CDC, 2022).
2. There are dangerous effects of prolonged untreated diabetes.
 - a) According to the Center for Disease Control and Prevention's (CDC) National Diabetes Statistics Report, prolonged high blood sugar levels can lead to cardiovascular disease and kidney damage (CDC, 2022).
3. Early intervention for individuals at risk of diabetes can help reduce the chances of becoming diabetic
 - a) According to a study conducted by the Diabetes Prevention Program Research Group, subjects who were administered into lifestyle early intervention programs to address their risk of diabetes contracted diabetes 58 percent less than subjects who did not participate in these programs (Diabetes Prevention Program Research Group, 2002).
4. Living with diabetes comes with financial burden
 - a) According to an academic article published in the *Diabetes Care* journal, "People diagnosed with diabetes, on average, have medical expenditures about 2.3 times higher than what expenditures would be in the absence of diabetes" (Petersen, 2018)

Further Motivation and Literature Review:

- Creating my digital questionnaire application will give individuals yet another resource to easily be notified if they are at-risk of diabetes. This will allow them to take appropriate action, such as lifestyle changes or get a full diabetes screening.
- However, in addition to wanting to create an application to help individuals take appropriate action, I would also like to fill gaps in past research

Academic Literature Review: Source 1

- **Title:** “Likelihood Prediction of Diabetes at Early State Using Data Mining Techniques”
- **Summary:** The book chapter describes the methodology of experimenting with data mining techniques, such as Naïve Bayes, Logistic Regression, and Random Forest, to understand if behavioral risk factors can be used to predict diabetes. They used a dataset containing diabetes behavioral risk factors, such as experiencing polyuria (excess urination), polydipsia (excessive thirst), episode of sudden weight loss, and muscle weakness. They discovered with this dataset that Random Forest performed the best. Also, they proposed the creation of a user-friendly application to assess if a person is at risk of diabetes. Although the application was proposed, it was never created.
- **Ways I can see improvement in research:** The data could consist of other predictors that fall more in line with a person’s habits, such as diet, sleep, smoking, etc. They could have checked for multicollinearity and performed more EDA (all columns present in the dataset were used). They could have removed extraneous collinear predictors. They could have included experimentation with support vector machines (SVM) and k-nearest neighbors (KNN) could have been conducted. They could have experimented with clustering the data to improve model accuracy. The application to assess diabetes risk could have been created.
- **Citation:** Islam, M.M., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2019). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. *In Computer Vision and Machine Intelligence in Medical Image Analysis* (pp. 113–125). Springer Singapore. https://doi.org/10.1007/978-981-13-8798-2_12

Academic Literature Review: Source 2

- **Title:** "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques"
- **Summary:** The academic article describes using multiple machine learning algorithms, such as support vector machine, decision tree, logistic regression, random forest, neural network, and Gaussian Naive Bayes classifiers, to predict the risk of type 2 diabetes. They used a dataset that consisted of predictors describing behavioral and habitual risk factors associated with diabetes, such as difficulty seeing, depression, exercise, smoking, vaccines (flu), and history of heart disease. The dataset was sourced from the CDC's 2014 Behavioral Risk Factor Surveillance System (BRFSS), which is the system responsible for conducting health-related telephone surveys asking about risky behaviors, chronic diseases, and the use of preventative services. They discovered that neural networks performed the best with highest the AUC.
- **Ways I can see improvement:** The research could experiment with k-nearest neighbors (KNN). Also, they could experiment and analyze predictors describing diet, high blood pressure, high cholesterol, and difficulty walking. They could have checked for multicollinearity and performed more EDA. They could have removed extraneous collinear predictors. They could have experimented with clustering the data to improve model accuracy. Furthermore, an application that can deploy their research was not proposed or created.
- **Citation:** Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Preventing Chronic Disease*, 16.
<https://doi.org/10.5888/pcd16.190109>

Why Researching These Questions are Important to Me:

1. Add to the conversation and contribute to the use of data science concepts within the medical field
2. Use the concepts I learned through my previous courses to make a useful resource that can help individuals make data-driven decisions on the actions they need to take (i.e. make lifestyle changes or get a full diabetes screening)
3. By sharing this project on GitHub, I hope to get others involved in the use of data science in medicine. (They can improve on my methods or use the project as inspiration for their own methods)

Things I Hope to Review and Learn:

(other than answering the research questions)

1. Things I hope to review from previous courses:

- Data cleaning
- Data Preparation
- Exploratory Data Analysis (i.e. univariate data exploration, multivariate data exploration, chi-square hypothesis testing, ANOVA hypothesis testing, Cramer's V metric of association, Variance Inflation Factor (VIF) to measure multicollinearity)
- Checking for multicollinearity
- Supervised ML algorithms (i.e. logistic regression, naïve bayes, support vector machines, k-nearest neighbor, and decisions trees)

2. Things I hope to learn more about through experience:

- The concept of clustering data observations to improve model accuracy
- Gower's distance and Partition Around Medoids (PAM) clustering

Data Part 1: Metadata

- **Name:** Behavioral Risk Factor Surveillance System: Public Health Surveys of 400k people in 2015
- **Distributor:** Centers for Disease Control and Prevention
- **Description:** It comes directly from the CDC and contains 441, 456 survey responses to the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS) Health Survey. Also, the 330 variables are the answers to these health survey questions. Moreover, it comes in the form of a 516 MB .csv file with 441, 457 rows and 330 columns
- **File Type and Size:** The dataset comes in the form of a 516 MB .csv file with 441, 457 rows and 330 columns.
- **Reason for use:** The data consist of several useful behavioral and habitual risk factors and predictors associated with diabetes, such as diet, high blood pressure, high cholesterol, and difficulty walking. Also, the dataset has plenty of variables to choose from in case I need any more variables. Also, the CDC is a reliable source, and this survey is the CDC's main method of getting public health data.
- **Initial Data Processing (a starting point):** Selected predictor variables that are known to be associated with diabetes, such as age, difficulty walking, high blood pressure, high cholesterol, BMI, etc inspired from another dataset from another Kaggle user [Alex Teboul](#) and domain knowledge
- **Citation:** Centers for Disease Control and Prevention (2018). *Behavioral Risk Factor Surveillance System: Public Health Surveys of 400k people in 2015*. [Data set]. kaggle. <https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system>

Data Part 2: Data Dictionary of the BFRSS Heath Survey Dataset

Behavioral Risk Factor Surveillance System

2015 Codebook Report

Land-Line and Cell-Phone data

August 23, 2016



Link to Codebook:

https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Data Part 3: Initial Cleaned and More Manageable Dataset

- Diabetes_binary = Has diabetes?
- HighBP = Has been told they have high blood pressure by a doctor, nurse, or other health professional
- HighChol = Has EVER been told by a doctor, nurse or other health professional that their blood cholesterol is high
- CholCheck = Has had their cholesterol checked within past five years
- BMI = Body mass index
- Smoker = Has smoked at least 100 cigarettes in your entire life
- Stroke = Had a stroke
- HeartDiseaseorAttack = Had a coronary heart disease or myocardial infarction
- PhysActivity = Reported doing physical activity or exercise during the past 30 days other than their regular job
- Fruits = Consumes fruit one or more times per day
- Veggies = Consumes vegetables one or more times per day
- HvyAlcoholConsump = Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
- AnyHealthcare = Has any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service
- NoDocbcCost = There was a time in the past 12 months when they needed to see a doctor but could not because of cost
- GenHlth = General health (Excellent, very good, good, fair, Poor; 1, 2, 3, 4, 5)
- MentHlth = How many days during the past 30 days was their mental health not good
- PhysHlth = How many days during the past 30 days was their physical health not good
- DiffWalk = Has serious difficulty walking or climbing stairs
- Sex = Sex
- Age = Age category (must be of 18 years or older and categories are organized in increments of 5 years)
- Education = Highest grade or year of school completed (never, elementary, some high school, high school, some college, college; 1, 2, 3, 4, 5, 6)
- Income = income level category (<10,000, <15,000, <20,000, <25,000, <35,000, <50,000, <75,000, >75,000; 1, 2, 3, 4, 5, 6, 7, 8)

Diabetes_bin	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1	0	1	26	0	0	0	1	0	1	0	1	0	3	5	30	0	1	4	6	8
0	1	1	1	26	1	1	0	0	1	0	0	1	0	3	0	0	0	1	12	6	8
0	0	0	1	26	0	0	0	1	1	1	0	1	0	1	0	10	0	1	13	6	8
0	1	1	1	28	1	0	0	1	1	1	0	1	0	3	0	3	0	1	11	6	8
0	0	0	1	29	1	0	0	1	1	1	0	1	0	2	0	0	0	0	8	5	8

Table 1: The first 5 rows of the “Behavioral Risk Factor Surveillance System: Public Health Surveys of 400k people in 2015” dataset . Data source:

<https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system>

Data Part 4: Cleaned Data Structure and Summary

```
{  
# look at the structure of the dataset  
str(data_balanced)  
  
'data.frame': 80258 obs. of 22 variables:  
 $ diabetes      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
 $ high_bp       : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 1 2 ...  
 $ high_chol     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...  
 $ chol_check    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 1 2 ...  
 $ BMI           : num 31.8 25.6 33.3 23.1 34.3 ...  
 $ smoker        : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 2 1 1 ...  
 $ stroke        : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...  
 $ heart_disease : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
 $ physical_activity: Factor w/ 2 levels "0","1": 1 2 1 2 2 2 2 2 2 1 ...  
 $ fruits        : Factor w/ 2 levels "0","1": 2 1 2 2 1 2 2 1 1 1 ...  
 $ veggies       : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 1 2 2 ...  
 $ heavy_drinker : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
 $ health_coverage : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...  
 $ NoDocbcCost   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
 $ general_health : Factor w/ 5 levels "1","2","3","4",...: 2 2 2 3 1 3 3 4 3 3 ...  
 $ mental_health : num 0 0 0 0 0 0 4 0 0 ...  
 $ physical_health : num 0 0 0 0 0 0 30 0 0 ...  
 $ diff_walking   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
 $ sex           : Factor w/ 2 levels "F","M": 2 2 1 2 1 1 2 1 1 2 ...  
 $ age_cat       : Factor w/ 13 levels "1","2","3","4",...: 11 13 4 9 5 11 9 6 3 7 ...  
 $ education_level : Factor w/ 6 levels "1","2","3","4",...: 4 6 6 6 5 6 6 6 6 4 ...  
 $ income_level   : Factor w/ 8 levels "1","2","3","4",...: 8 7 8 8 8 3 8 1 8 8 ...
```

```

# summary statistics of the variables
summary(data_balanced)

```

diabetes	high_bp	high_chol	chol_check	BMI	smoker	stroke	heart_disease	physical_activity	fruits	veggies
0:40129	0:35658	0:38393	0: 2026	Min. :12.55	0:42237	0:75463	0:68795	0:23437	0:30870	0:16885
1:40129	1:44600	1:41865	1:78232	1st Qu.:25.07	1:38021	1: 4795	1:11463	1:56821	1:49388	1:63373
				Median :28.52						
				Mean :29.80						
				3rd Qu.:33.11						
				Max. :97.65						

heavy_drinker	health_coverage	NoDocbcCost	general_health	mental_health	physical_health	diff_walking	sex	age_cat
0:76752	0: 3727	0:72687	1: 9551	Min. : 0.000	Min. : 0.000	0:72687	F:43628	10 :11941
1: 3506	1:76531	1: 7571	2:22945	1st Qu.: 0.000	1st Qu.: 0.000	1: 7571	M:36630	9 :11511
			3:26638	Median : 0.000	Median : 0.000			8 : 9711
			4:14841	Mean : 3.694	Mean : 5.679			11 : 9137
			5: 6283	3rd Qu.: 2.000	3rd Qu.: 5.000			7 : 7786
				Max. :30.000	Max. :30.000			13 : 6317
								(Other):23855

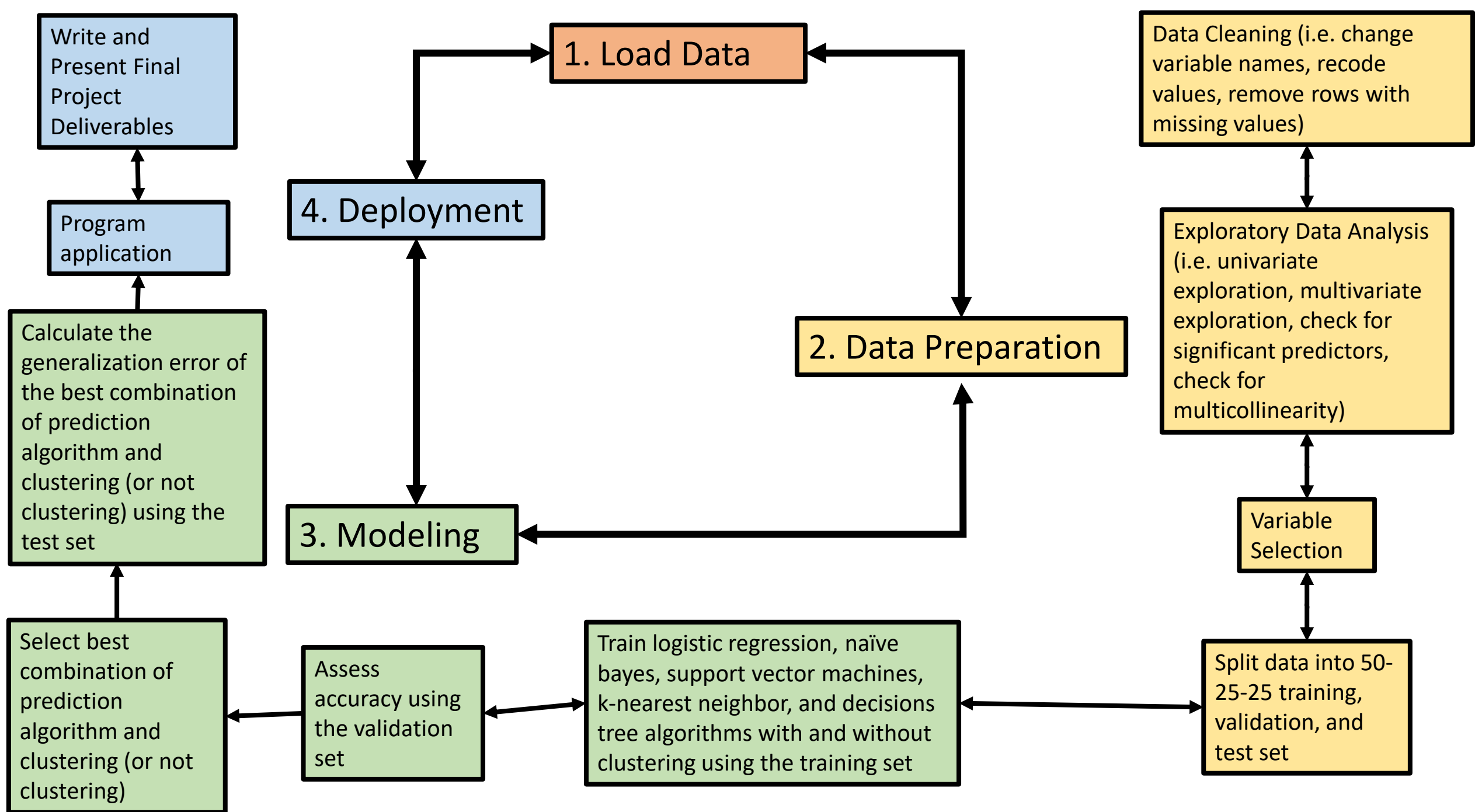
education_level	income_level
1: 74	8 :23539
2: 1837	7 :13067
3: 3944	6 :11785
4:21884	5 : 8988
5:22729	4 : 7547
6:29790	3 : 6232
	(Other): 9100

Methods of Analysis:

1. Exploratory Data Analysis will be conducted in a traditional manner (i.e. univariate exploration, multivariate data exploration, chi-square hypothesis testing, ANOVA hypothesis testing, Cramer's V metric of association, Variance Inflation Factor (VIF) to measure multicollinearity)
2. Select Variables that are not highly associated amongst each other and are significantly associated with diabetes
3. The data will be split into a 50-25-25 training, validation, and test set. Note, we can do this because even the cleaned dataset still has over 200,000 observations. The training set will be used to train the algorithms; the validation set will be used to assess the model's accuracy, choose which ML algorithm is the best, and whether to cluster or not to cluster; and the test set will only be used to finally calculate the generalization error of the best method and ML algorithm
4. Logistic regression, naïve bayes, support vector machines, k-nearest neighbor, and decisions trees will be trained without PAM clustering and evaluated.
5. Logistic regression, naïve bayes, support vector machines, k-nearest neighbor, and decisions trees will be trained with PAM clustering and evaluated by using average error. (note, I use Gower's distance and PAM clustering because I have a mixture of nominal and numerical variables)
6. The best combination of algorithm and clustering (or not clustering) will be used in the application
7. Calculate generalization error on the test set

Software and Packages:

- R
- RStudio
- Tidyverse (for general use)
- ISLR2 (for ML algorithms)
- Daisy (for PAM clustering)
- ShinyR (for application creation)



Progress to Date

On the next few slides...

Cleaned Data Structure and Summary

```
##{r}
# look at the structure of the dataset
str(data_balanced)
##{r}
```

'data.frame': 80258 obs. of 22 variables:

- \$ diabetes : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
- \$ high_bp : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 1 2 ...
- \$ high_chol : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
- \$ chol_check : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 1 2 2 ...
- \$ BMI : num 31.8 25.6 33.3 23.1 34.3 ...
- \$ smoker : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 2 1 1 2 ...
- \$ stroke : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
- \$ heart_disease : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
- \$ physical_activity : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 2 2 2 1 ...
- \$ fruits : Factor w/ 2 levels "0","1": 2 1 2 2 1 2 2 1 1 1 ...
- \$ veggies : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 1 2 ...
- \$ heavy_drinker : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
- \$ health_coverage : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
- \$ NoDocbcCost : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
- \$ general_health : Factor w/ 5 levels "1","2","3","4",...: 2 2 2 3 1 3 3 4 3 3 ...
- \$ mental_health : num 0 0 10 0 0 0 4 0 0 ...
- \$ physical_health : num 0 0 0 0 0 0 30 0 0 ...
- \$ diff_walking : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
- \$ sex : Factor w/ 2 levels "F","M": 2 2 1 2 1 1 2 1 1 2 ...
- \$ age_cat : Factor w/ 13 levels "1","2","3","4",...: 11 13 4 9 5 11 9 6 3 7 ...
- \$ education_level : Factor w/ 6 levels "1","2","3","4",...: 4 6 6 6 5 6 6 6 6 4 ...
- \$ income_level : Factor w/ 8 levels "1","2","3","4",...: 8 7 8 8 8 3 8 1 8 8 ...

```
##{r}
# summary statistics of the variables
summary(data_balanced)
##{r}
```

diabetes	high_bp	high_chol	chol_check	BMI	smoker	stroke	heart_disease	physical_activity	fruits	veggies
0:40129	0:35658	0:38393	0: 2026	Min. :12.55	0:42237	0:75463	0:68795	0:23437	0:30870	0:16885
1:40129	1:44600	1:41865	1:78232	1st Qu.:25.07	1:38021	1: 4795	1:11463	1:56821	1:49388	1:63373
				Median :28.52						
				Mean :29.80						
				3rd Qu.:33.11						
				Max. :97.65						

heavy_drinker	health_coverage	NoDocbcCost	general_health	mental_health	physical_health	diff_walking	sex	age_cat
0:76752	0: 3727	0:72687	1: 9551	Min. : 0.000	Min. : 0.000	0:72687	F:43628	10 :11941
1: 3506	1:76531	1: 7571	2:22945	1st Qu.: 0.000	1st Qu.: 0.000	1: 7571	M:36630	9 :11511
			3:26638	Median : 0.000	Median : 0.000			8 : 9711
			4:14841	Mean : 3.694	Mean : 5.679			11 : 9137
			5: 6283	3rd Qu.: 2.000	3rd Qu.: 5.000			7 : 7786
				Max. :30.000	Max. :30.000			13 : 6317
								(Other):23855

education_level	income_level
1: 74	8 :23539
2: 1837	7 :13067
3: 3944	6 :11785
4:21884	5 : 8988
5:22729	4 : 7547
6:29790	3 : 6232
	(Other): 9100

Hypothesis Testing for the Relationship between Factor Type Predictors and Diabetes

```
[1] "Chi-squared test of independence between diabetes and high_bp ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 10886, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and high_chol ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 6573.8, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and chol_check ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 1013.9, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and smoker ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 557.53, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and stroke ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 1181.5, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and heart_disease ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 3302, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and physical_activity ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 1998.1, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and fruits ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 261.08, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and veggies ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 484.66, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and heavy_drinker ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 595.49, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and health_coverage ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 36.467, df = 1, p-value = 1.553e-09
```

```
[1] "Chi-squared test of independence between diabetes and NoDocbcCost ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 192.2, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and general_health ->"
```

```
      Pearson's Chi-squared test
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 13058, df = 4, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and diff_walking ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 192.2, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and sex ->"
```

```
      Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 98.856, df = 1, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and age_cat ->"
```

```
      Pearson's Chi-squared test
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 6636.4, df = 12, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and education_level ->"
```

```
      Pearson's Chi-squared test
```

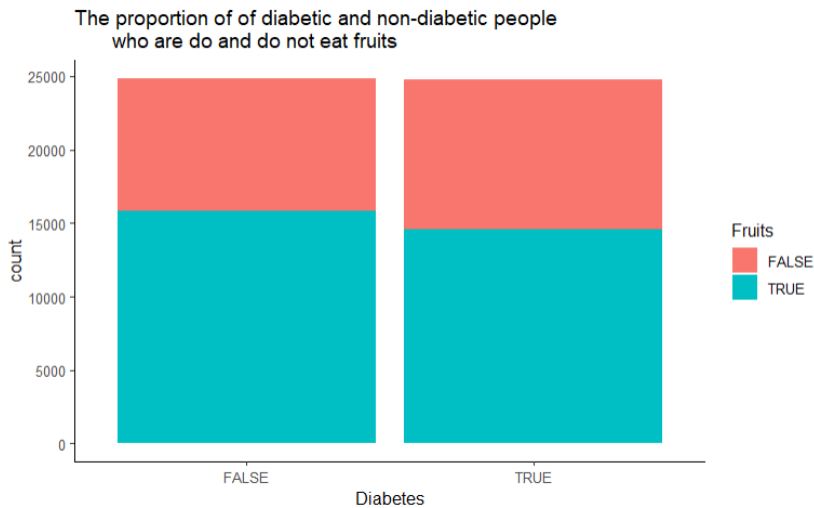
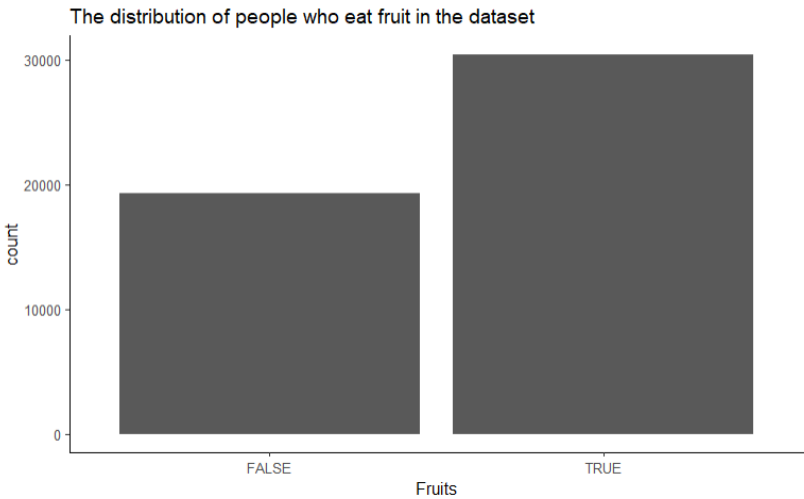
```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 2469.2, df = 5, p-value < 2.2e-16
```

```
[1] "Chi-squared test of independence between diabetes and income_level ->"
```

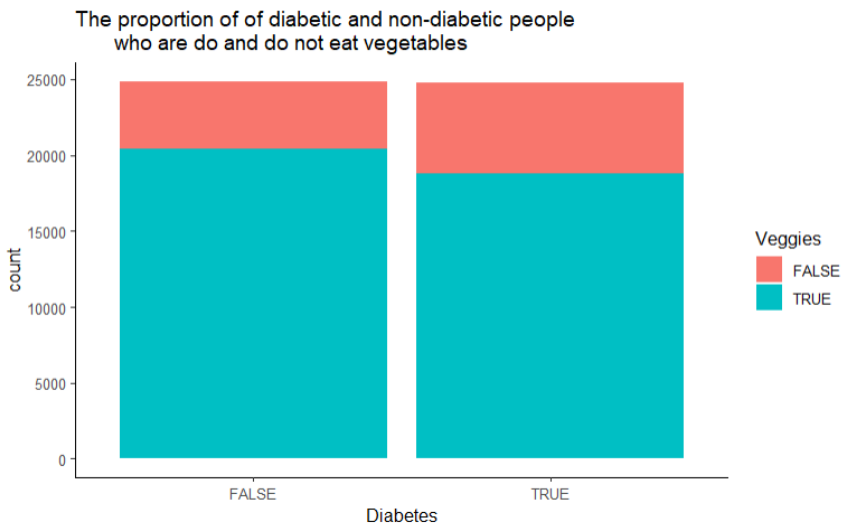
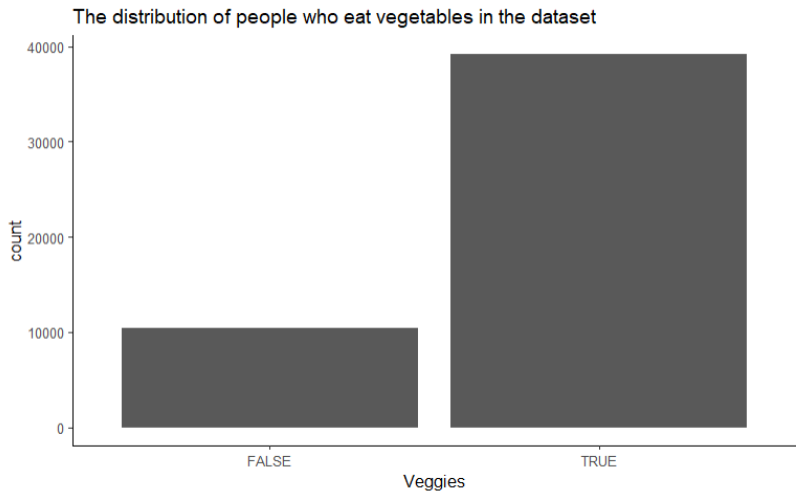
```
      Pearson's Chi-squared test
```

```
data:  table(data_balanced[, 1], data_balanced[, i])  
X-squared = 4365.8, df = 7, p-value < 2.2e-16
```

Exploratory Data Analysis Part 1: Diet

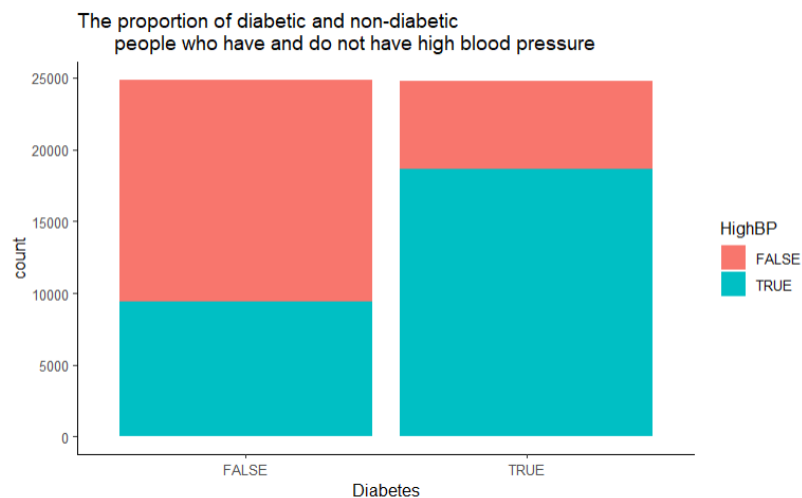
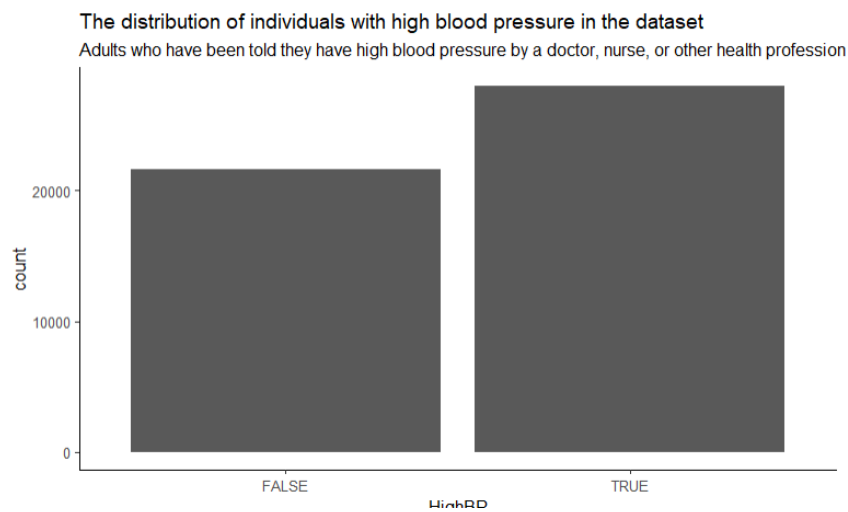


Pearson's Chi-squared test with Yates' continuity correction
data: table(train_data\$Fruits, train_data\$Diabetes_binary)
X-squared = 124.38, df = 1, p-value < 2.2e-16

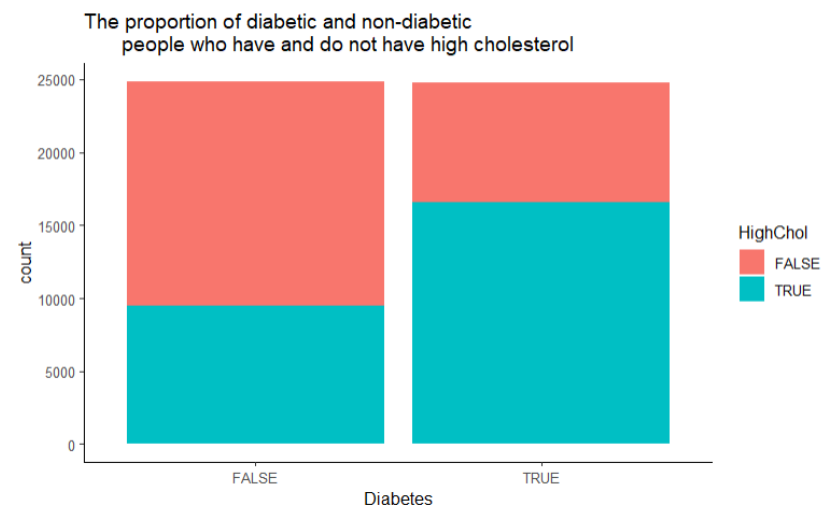
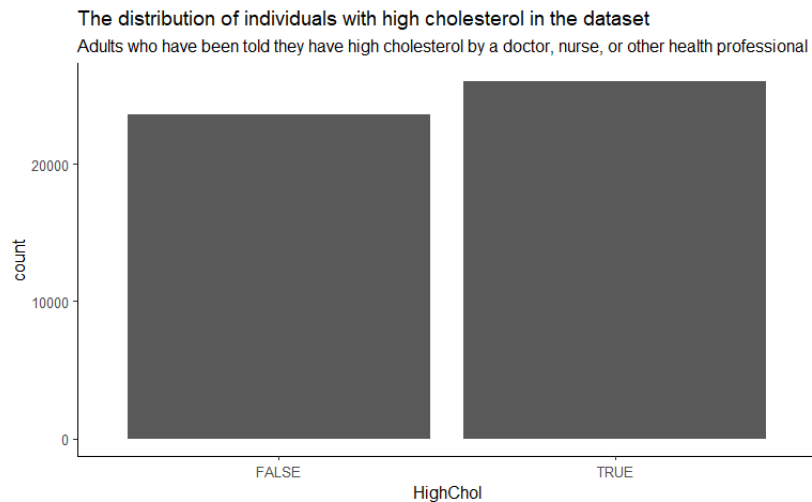


Pearson's Chi-squared test with Yates' continuity correction
data: table(train_data\$Veggies, train_data\$Diabetes_binary)
X-squared = 295.5, df = 1, p-value < 2.2e-16

Exploratory Data Analysis Part 2: High Blood Pressure and High Cholesterol

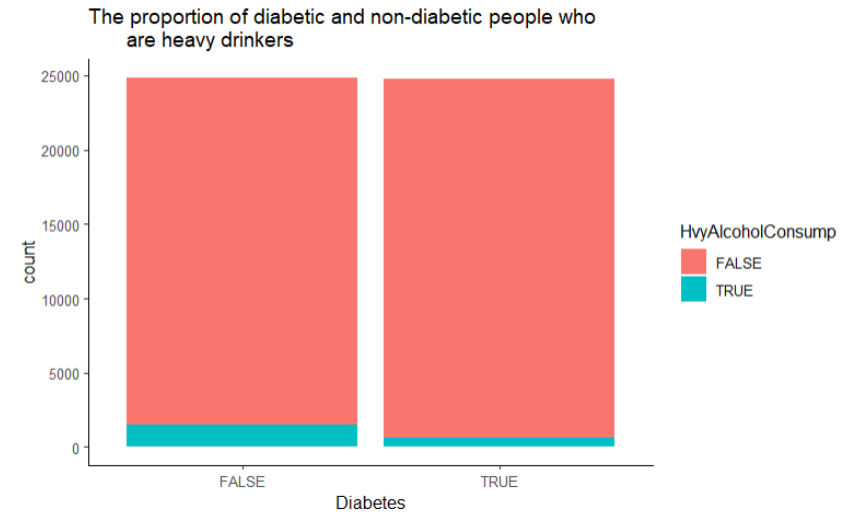
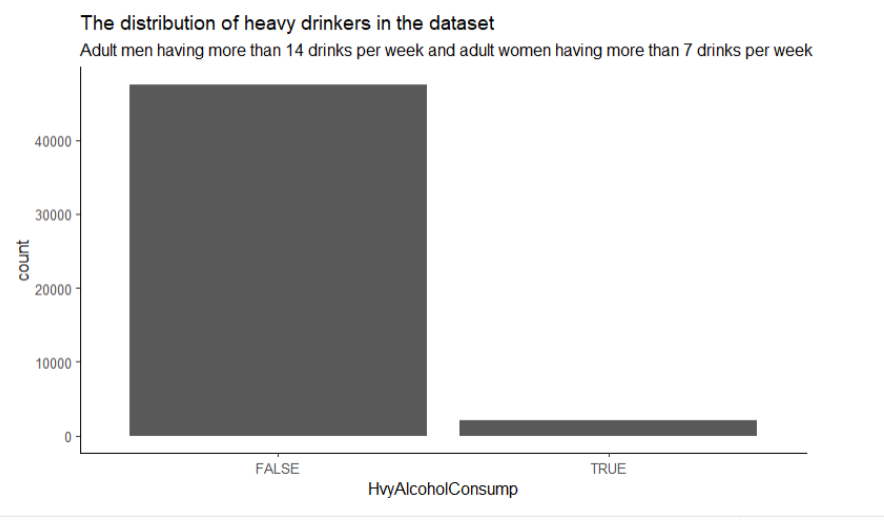


Pearson's Chi-squared test with Yates' continuity correction
data: table(train_data\$HighBP, train_data\$Diabetes_binary)
X-squared = 7102.7, df = 1, p-value < 2.2e-16

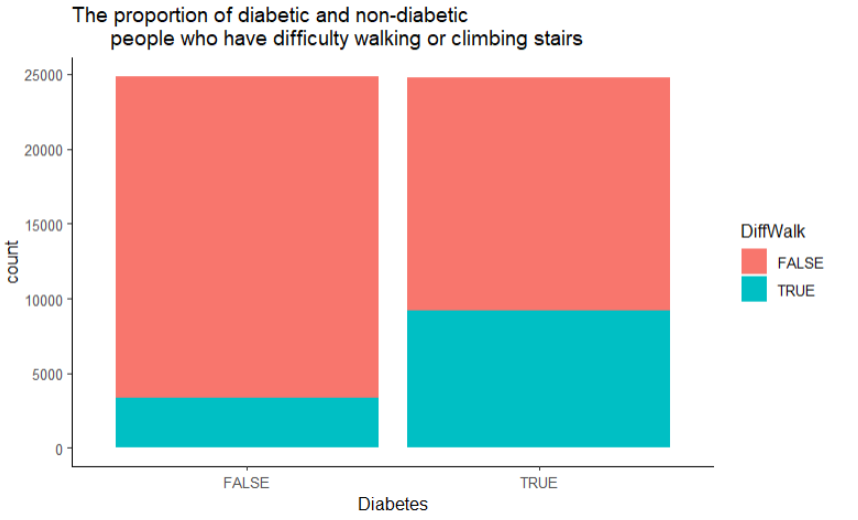
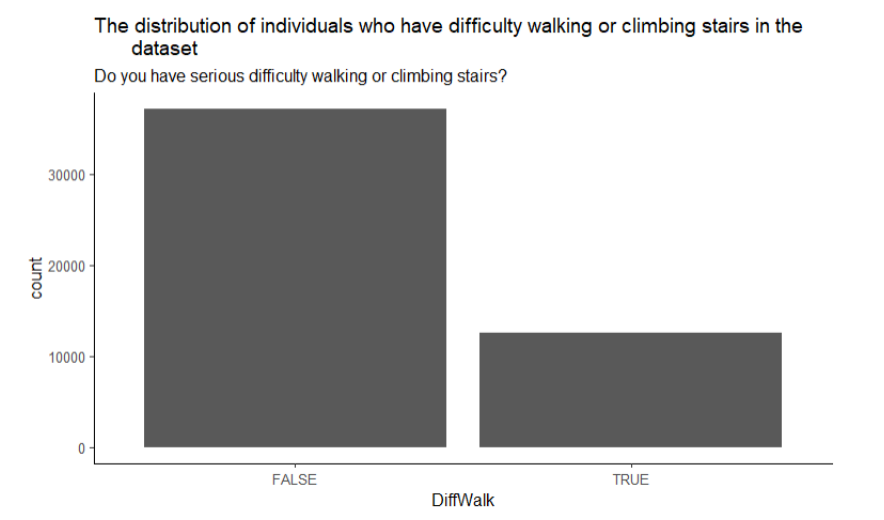


Pearson's Chi-squared test with Yates' continuity correction
data: table(train_data\$HighChol, train_data\$Diabetes_binary)
X-squared = 4097, df = 1, p-value < 2.2e-16

Exploratory Data Analysis Part 3: Heavy Alcohol Consumption and Difficulty Walking

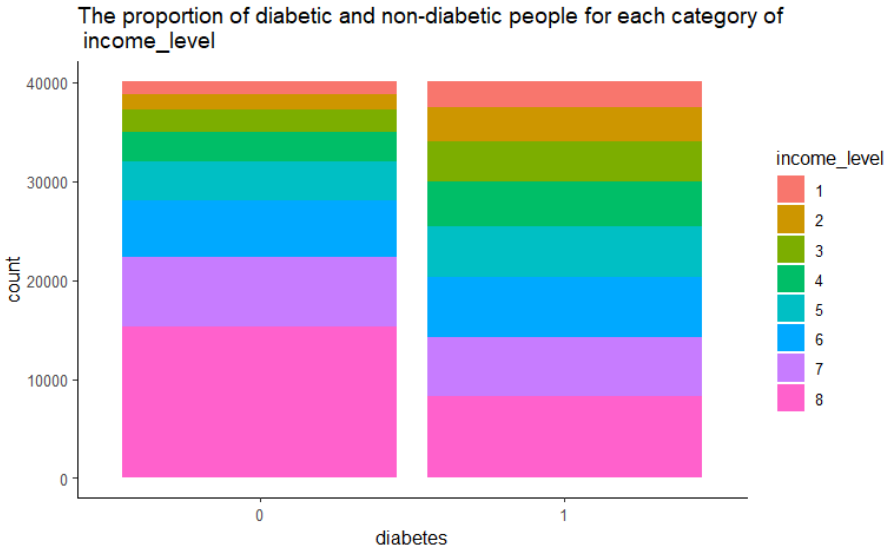
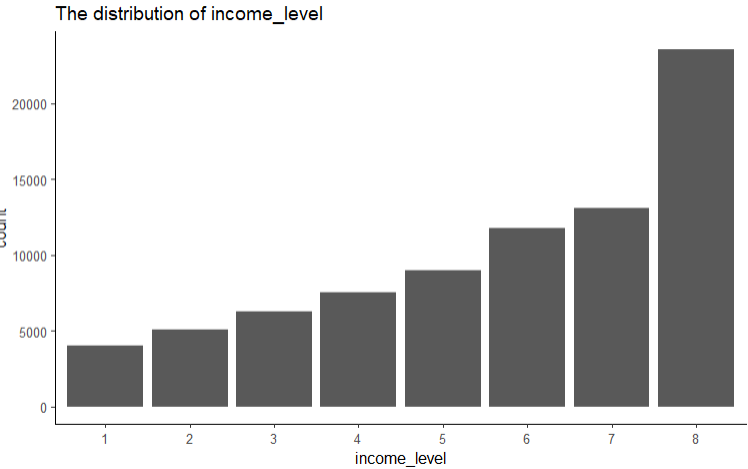


Pearson's Chi-squared test with Yates' continuity correction
data: table(train_data\$HvyAlcoholConsump, train_data\$Diabetes_binary)
X-squared = 414.5, df = 1, p-value < 2.2e-16

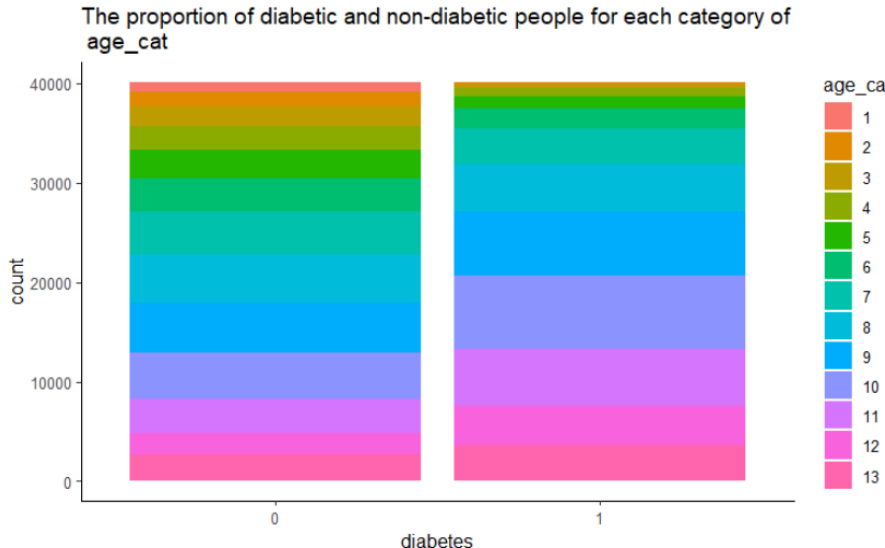
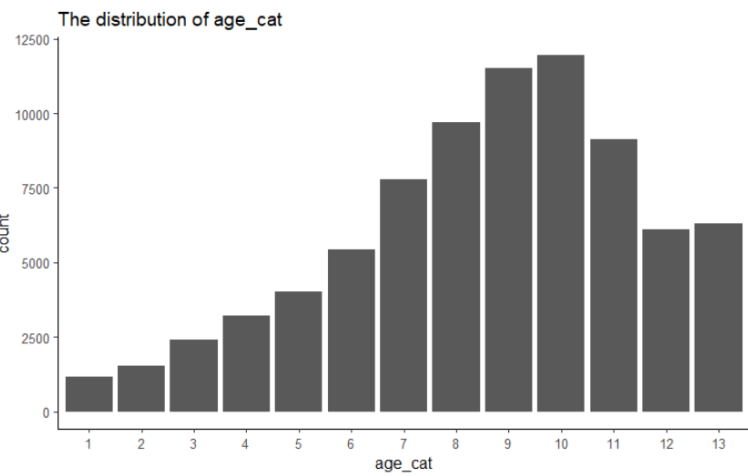


Pearson's Chi-squared test with Yates' continuity correction
data: table(train_data\$Diffwalk, train_data\$Diabetes_binary)
X-squared = 3672.3, df = 1, p-value < 2.2e-16

Exploratory Data Analysis Part 4: Income Level and Age Category



Value	Value Label
1	Less than \$10,000 Notes: If "no," code 02
2	Less than \$15,000 (\$10,000 to less than \$15,000) Notes: If "no," code 03; if "yes," ask 01
3	Less than \$20,000 (\$15,000 to less than \$20,000) Notes: If "no," code 04; if "yes," ask 02
4	Less than \$25,000 (\$20,000 to less than \$25,000) Notes: If "no," ask 05; if "yes," ask 03
5	Less than \$35,000 (\$25,000 to less than \$35,000) Notes: If "no," ask 06
6	Less than \$50,000 (\$35,000 to less than \$50,000) Notes: If "no," ask 07
7	Less than \$75,000 (\$50,000 to less than \$75,000) Notes: If "no," code 08
8	\$75,000 or more



Reported age in five-year age categories calculated variable

CalculatedVariables: 7.11

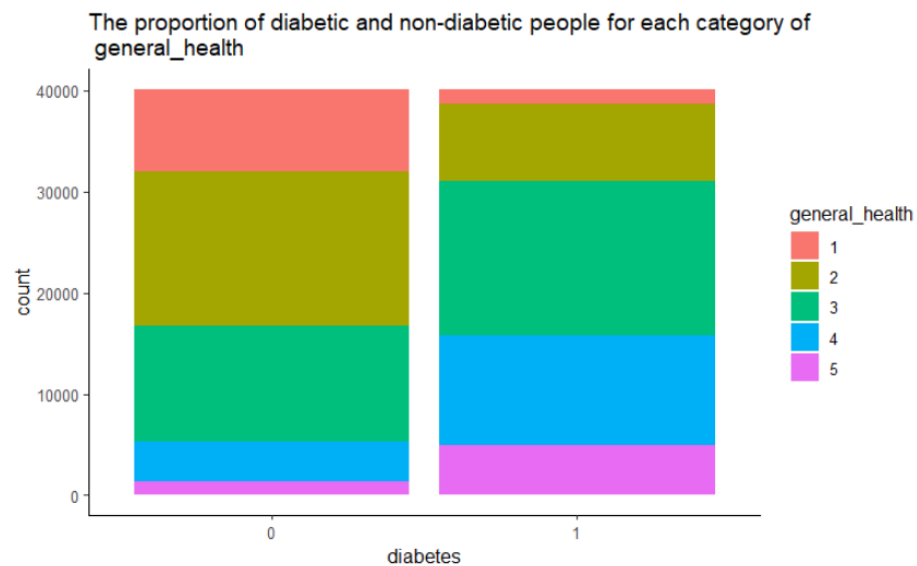
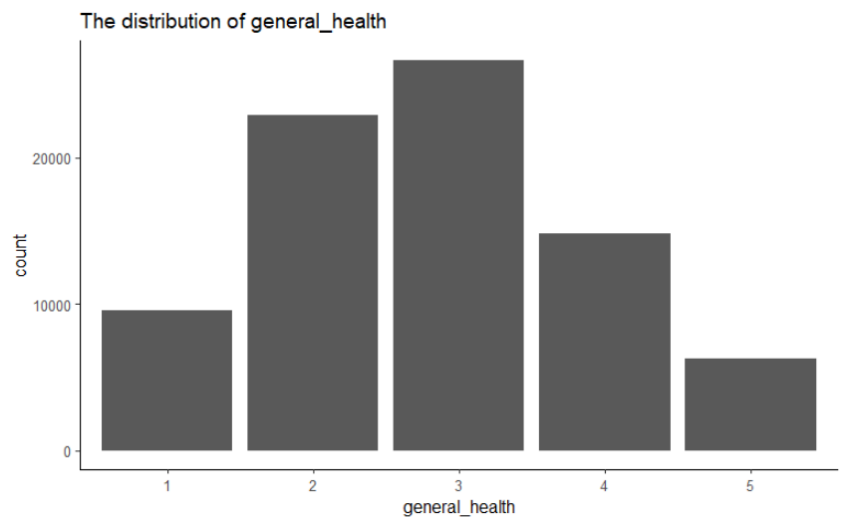
Column: 1971-1972

Prologue:

Description: Fourteen-level age category

Value	Value Label
1	Age 18 to 24 Notes: 18 <= AGE <= 24
2	Age 25 to 29 Notes: 25 <= AGE <= 29
3	Age 30 to 34 Notes: 30 <= AGE <= 34
4	Age 35 to 39 Notes: 35 <= AGE <= 39
5	Age 40 to 44 Notes: 40 <= AGE <= 44
6	Age 45 to 49 Notes: 45 <= AGE <= 49
7	Age 50 to 54 Notes: 50 <= AGE <= 54
8	Age 55 to 59 Notes: 55 <= AGE <= 59
9	Age 60 to 64 Notes: 60 <= AGE <= 64
10	Age 65 to 69 Notes: 65 <= AGE <= 69
11	Age 70 to 74 Notes: 70 <= AGE <= 74
12	Age 75 to 79 Notes: 75 <= AGE <= 79
13	Age 80 or older ..

Exploratory Data Analysis Part 5: General Health and Gender



General Health

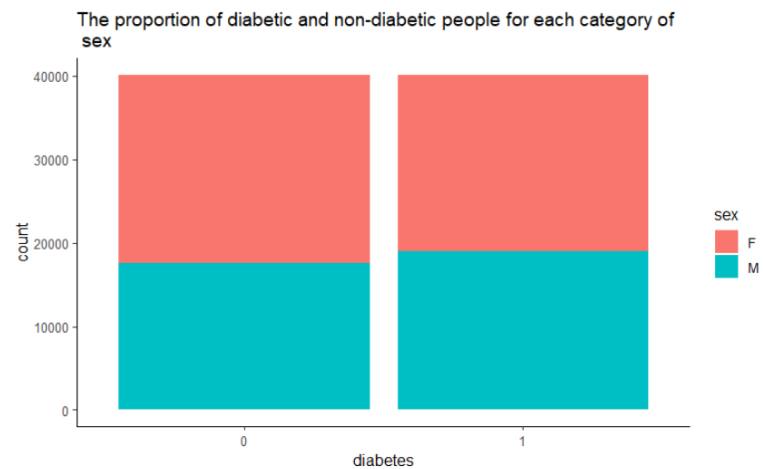
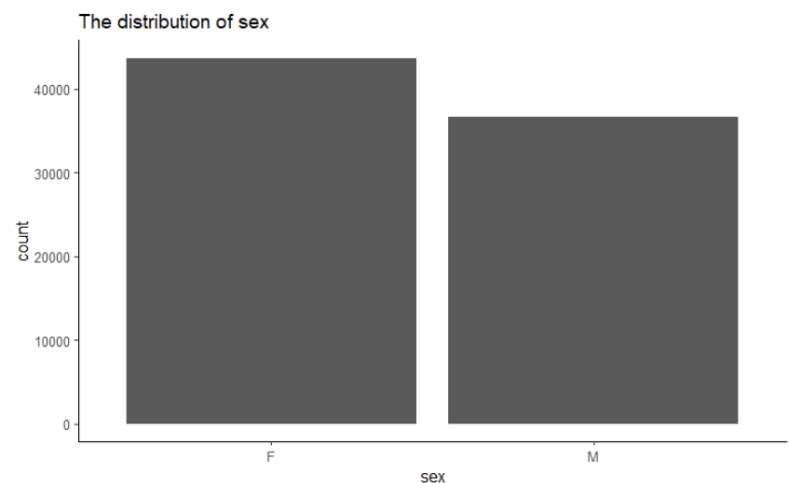
Section: 1.1 Health Status

Column: 90

Prologue:

Description: Would you say that in general your health is:

Value	Value Label
1	Excellent
2	Very good
3	Good
4	Fair
5	Poor



[1] "Chi-squared test of independence between diabetes and sex ->"

Pearson's Chi-squared test with Yates' continuity correction

```
data: table(data_balanced[, 1], data_balanced[, i])
X-squared = 98.856, df = 1, p-value < 2.2e-16
```


We will use Cramer's V to measure the strength of the between factor variables

Table 1. Interpretation of effect size

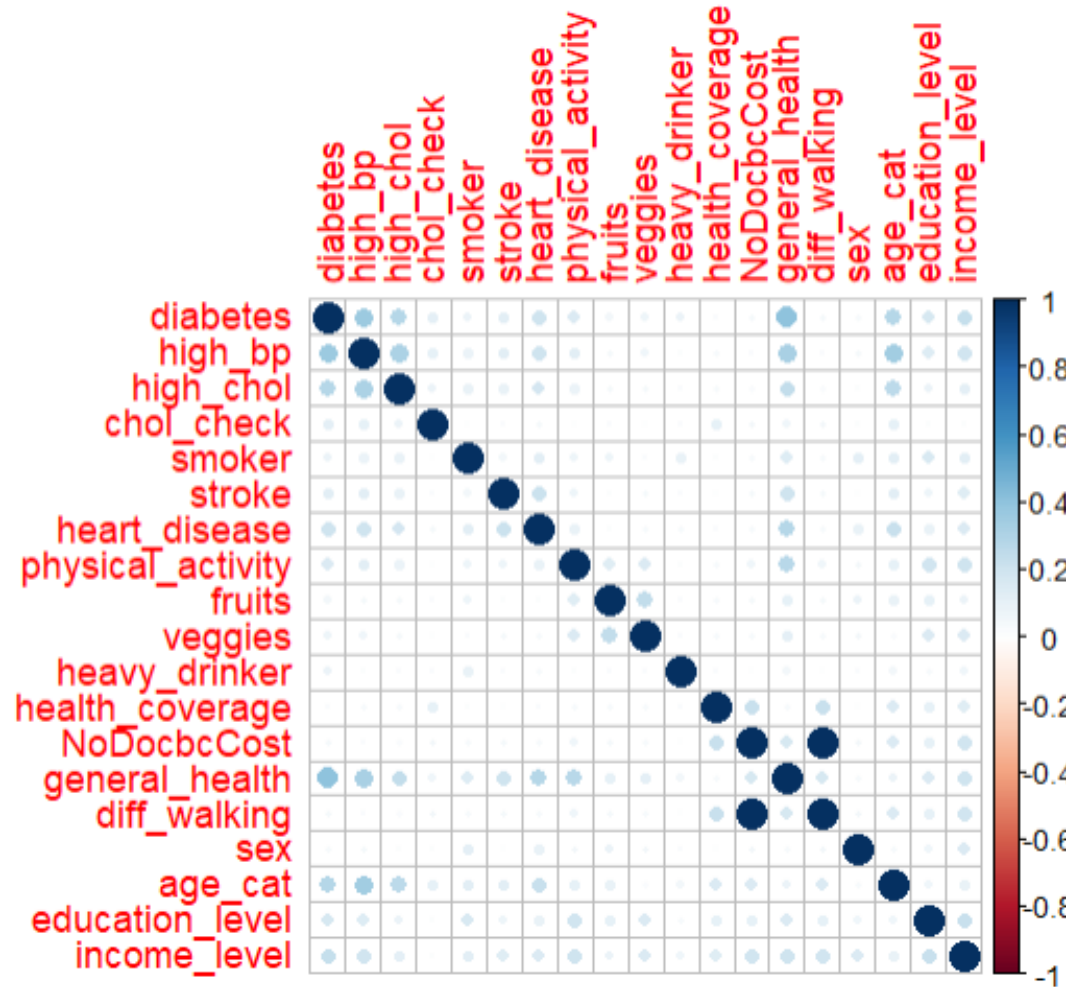
Effect size (ES)	Interpretation
$ES \leq 0.2$	The result is weak. Although the result is statistically significant, the fields are only weakly associated.
$0.2 < ES \leq 0.6$	The result is moderate. The fields are moderately associated.
$ES > 0.6$	The result is strong. The fields are strongly associated.

We will use this general guideline to measure strength between factor variables. Provided by IBM. [Cramér's V - IBM Documentation](#)

Relationships between the Factor Type Predictors and Diabetes.

Also, Multicollinearity Among Factor Type Predictors:

Note: used Cramer's V to measure how related the factor variables



Analysis - Relationship between Factor Type Predictors and Diabetes:

- The factor type predictors that are most strongly related to diabetes with Cramer's V > 0.2 are high_bp (0.36831274), high_chol (0.28622075), heart_disease (0.20287038), general health (0.40335672), age_cat(0.28755565), income_level(0.23323228)

Analysis - Multicollinearity Among Factor Type Predictors :

- High_chol -> high_bp (0.31360472)
- Heart_disease -> high_bp (0.20295323), Stroke (0.21907870)
- Veggies -> Fruits (0.245024490)
- NoDocbcCost -> health_coverage (0.229804631)
- General_health -> high_bp (0.32292487), high_chol (0.24000009), stroke (0.20108227), heart_disease (0.28431391), physical_activity (0.27886606)
- Diff_walking -> health_coverage (0.229804631), NoDocbcCost (1.000000000)

Conclusion:

- Remove high_chol because high_chol closely related to high_bp, but high_bp has a stronger relationship with diabetes
- Remove stroke because stroke is closely related to heart_disease and high_bp, but heart_disease and high_bp have a stronger relationship with diabetes
- Remove fruits because fruits are closely related to veggies, but veggies have a stronger relationship with diabetes (0.077739849)
- Remove health_coverage and NoDocbcCost because health_coverage and NoDocbcCost are closely related to Diff_walking, but Diff_walking is most closely related to diabetes
- Remove general health because it is strongly related to many other variables. Although, general health is most closely related to diabetes, general health can simply be implied with other variables

Tentative Deadlines and Schedule:

- Week 7: all variables selected, data cleaned, visualizations made, and logistic regression trained and analyze
- Week 8 to 9: train and evaluate the rest of the machine learning models (naïve bayes, support vector machines, k-nearest neighbor, and decisions trees). Also, experiment with clustering and not clustering.
- Week 10 to 11: create the application and get started on the final deliverables
- Week 12 to 15: finish the final deliverables.

References

CDC. (2018). *Behavioral Risk Factor Surveillance System* [Data set]. kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv

CDC. (2022, June 29). *National Diabetes Statistics Report*. The Centers for Disease Control and Prevention. <https://www.cdc.gov/diabetes/data/statistics-report/index.html>

Islam, M.M., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2019). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis* (pp. 113–125). Springer Singapore. https://doi.org/10.1007/978-981-13-8798-2_12

Knowler, W., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., & Nathan, D. M. (2002). Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. *The New England Journal of Medicine*, 346(6), 393–403. <https://doi.org/10.1056/NEJMoa012512>

Petersen, M. (2018). Economic Costs of Diabetes in the U.S. in 2017. *Diabetes Care*, 41(5), 917–928. <https://doi.org/10.2337/dci18-0007>

Teboul, Alex (2019). *Diabetes Health Indicators Dataset* [Data set]. kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Preventing Chronic Disease*, 16. <https://doi.org/10.5888/pcd16.190109>