An Analysis on Diabetes

Introduction

Diabetes influences 37.3 million Americans. That's 1 in 10 people. Even worse, 1 in 3 adults in America, or 96 million adults have prediabetes. It's the seventh leading cause of death in the US and costs an estimate $327 billion in medical costs ever year.

Scientifically, diabetes is a serious chronic disease where people lose the ability to regulate the glucose levels in their blood. During digestion, food breaks down into sugar in the bloodstream. In a healthy body, this causes the pancreas to release insulin, a hormone that allows sugar to be used for energy. Diabetes is characterized as a problem with this feedback cycle (people are either unable to use the insulin or produce enough insulin). Diabetes has no cure, but early diagnoses and lifestyle changes (ie: losing weight, eating healthier, and being active) can be quite helpful.

Thus, it is quite evident that diabetes is a disease that affects a lot of people a lot. I truly believe that technology should be used to help people live better quality lives, and I think a statistical deep dive into understanding the inner workings of diabetes can help medical professionals find connections and make medications.

Data Source

I opted to study two diabetes datasets from Kaggle. The first is a dataset of 253,680 survey responses to the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS). Every year, the CDC collects health-related responses from 400,000+ Americans regarding both chronic health conditions and preventative services/lifestyle changes. The data used in this analysis contained 21 features and 70,692 participant surveys. I used this dataset for the majority of my analysis. Towards the end of my analysis, I also used a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases that consisted of data from females at least 21 years old of Pima Indian heritage. This was a much smaller database, and had 768 rows and 8 factors.

Both datasets are linked in the appendix.

Data Ethics

Since this is health data, it is integral to think of the data ethics behind the collection and usage of this data.Unlike other sectors, health data is bound to a few legal standards including the Health Insurance Portability and Accountability Act of 1996 (HIPPA). HIPPA requires that sensitive patient information cannot be disclosed without the patient's consent or knowledge.Since this is government data, it it is likely that the appropriate protocols were made in anonymizing the data. We must continue to respect the consent of the participants in this study and not perform analyses that could have a malicious intent. It is also important to consider the downstream effects of this analysis. This dataset currently is being used solely for an exploratory purpose; this paper does not attempt to interpolate extraneous details about each user in a harmful way nor does it state that the data collected her is representative of the whole US population. Both pieces of data were voluntarily filled out, and thus our sample size was not randomly sampled. Thus, we must ensure that the paper does not get misconstrued this way since that could be potentially harmful (ie generalizing the results of this study to a population at large). It is critical to work with data in an ethical fashion; the rows of data are the lives of real people.

Load Data

First, we input the data. The sheets were downloaded from Kaggle as a csv and the imported.

```
us_diabetics <- read_csv("us_diabetics.csv")
```

```
## Rows: 70692 Columns: 22

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (22): Diabetes_binary, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke,...

##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
pima_diabetics <- read_csv("pima_diabetics.csv")
```

```
## Rows: 768 Columns: 9

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (9): Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, D...

##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Data Import, Cleaning and Tidying, and Exploration

This data is already tidy. We want to ensure it is clean as well. First, we ensure that all the data sets are of the correct variable type.

Printing the first 10 rows of the US Diabetic data set we see:

```
us_diabetics
```

```
## # A tibble: 70,692 x 22
##    Diabetes_binary HighBP HighChol CholCheck   BMI Smoker Stroke
##              <dbl>  <dbl>    <dbl>     <dbl> <dbl>  <dbl>  <dbl>
##  1               0      1        0         1    26      0      0
##  2               0      1        1         1    26      1      1
##  3               0      0        0         1    26      0      0
##  4               0      1        1         1    28      1      0
##  5               0      0        0         1    29      1      0
##  6               0      0        0         1    18      0      0
##  7               0      0        1         1    26      1      0
##  8               0      0        0         1    31      1      0
##  9               0      0        0         1    32      0      0
## 10               0      0        0         1    27      1      0
## # ... with 70,682 more rows, and 15 more variables: HeartDiseaseorAttack <dbl>,
## #   PhysActivity <dbl>, Fruits <dbl>, Veggies <dbl>, HvyAlcoholConsump <dbl>,
## #   AnyHealthcare <dbl>, NoDocbcCost <dbl>, GenHlth <dbl>, MentHlth <dbl>,
## #   PhysHlth <dbl>, DiffWalk <dbl>, Sex <dbl>, Age <dbl>, Education <dbl>,
## #   Income <dbl>
```

Here, all the values are integers but they're encoded as doubles. We correct this:

```
us_diabetics <- us_diabetics %>%
  mutate(across(everything(), as.integer))
```

Now, all of the numbers are correctly coded as integers.

```
us_diabetics
```

```
## # A tibble: 70,692 x 22
##    Diabetes_binary HighBP HighChol CholCheck   BMI Smoker Stroke
##              <int>  <int>    <int>     <int> <int>  <int>  <int>
##  1               0      1        0         1    26      0      0
##  2               0      1        1         1    26      1      1
##  3               0      0        0         1    26      0      0
##  4               0      1        1         1    28      1      0
##  5               0      0        0         1    29      1      0
##  6               0      0        0         1    18      0      0
##  7               0      0        1         1    26      1      0
##  8               0      0        0         1    31      1      0
##  9               0      0        0         1    32      0      0
## 10               0      0        0         1    27      1      0
## # ... with 70,682 more rows, and 15 more variables: HeartDiseaseorAttack <int>,
## #   PhysActivity <int>, Fruits <int>, Veggies <int>, HvyAlcoholConsump <int>,
## #   AnyHealthcare <int>, NoDocbcCost <int>, GenHlth <int>, MentHlth <int>,
## #   PhysHlth <int>, DiffWalk <int>, Sex <int>, Age <int>, Education <int>,
## #   Income <int>
```

Similarly, a lot of the health diagnostic data have been encoded in binary digits this can be helpful for modelling, but there might also be a time and place where it would be beneficial to have the individual labels too. Thus, we create an alternate version of this tibble where we replace each number with the written status as a factor.

```
us_diabetics_factored <- us_diabetics %>%
   mutate(Diabetes_binary = factor(Diabetes_binary)) %>%
  mutate(Diabetes_binary = fct_recode(Diabetes_binary, "Not Diabetic" = "0",
                                      "Diabetic" = "1")) %>%
   mutate(HighBP = factor(HighBP)) %>%
  mutate(HighBP = fct_recode(HighBP, "No" = "0", "Yes" = "1")) %>%
   mutate(HighChol = factor(HighChol)) %>%
  mutate(HighChol = fct_recode(HighChol, "No" = "0", "Yes" = "1")) %>%
  mutate(Smoker = factor(Smoker)) %>%
  mutate(Smoker = fct_recode(Smoker, "No" = "0", "Yes" = "1")) %>%
  mutate(Stroke = factor(Stroke)) %>%
  mutate(Stroke = fct_recode(Stroke, "No" = "0", "Yes" = "1")) %>%
  mutate(HeartDiseaseorAttack = factor(HeartDiseaseorAttack)) %>%
  mutate(HeartDiseaseorAttack = fct_recode(HeartDiseaseorAttack,
                                           "No" = "0", "Yes" = "1")) %>%
  mutate(PhysActivity = factor(PhysActivity)) %>%
  mutate(PhysActivity = fct_recode(PhysActivity, "No" = "0", "Yes" = "1")) %>%
  mutate(Fruits = factor(Fruits)) %>%
  mutate(Fruits = fct_recode(Fruits, "No" = "0", "Yes" = "1")) %>%
  mutate(Veggies = factor(Veggies)) %>%
  mutate(Veggies = fct_recode(Veggies, "No" = "0", "Yes" = "1")) %>%
  mutate(HvyAlcoholConsump = factor(HvyAlcoholConsump)) %>%
  mutate(HvyAlcoholConsump = fct_recode(HvyAlcoholConsump, "No" = "0", "Yes" = "1")) %>%
 mutate(AnyHealthcare = factor(AnyHealthcare)) %>%
  mutate(AnyHealthcare = fct_recode(AnyHealthcare, "No" = "0", "Yes" = "1")) %>%
  mutate(NoDocbcCost = factor(NoDocbcCost)) %>%
  mutate(NoDocbcCost = fct_recode(NoDocbcCost, "No" = "0", "Yes" = "1")) %>%
  mutate(GenHlth = factor(GenHlth)) %>%
  mutate(GenHlth = fct_recode(GenHlth, "Excellent" = "1", "Fair" = "4",
                              "Good" = "3", "Poor" = "5", " Very Good" = "2"))  %>%
  mutate(DiffWalk = factor(DiffWalk)) %>%
  mutate(DiffWalk = fct_recode(DiffWalk, "No" = "0", "Yes" = "1")) %>%
  mutate(Sex = factor(Sex)) %>%
  mutate(Sex = fct_recode(Sex, "Female" = "0", "Male" = "1")) %>%
  mutate(Education = factor(Education)) %>%
  mutate(Education = fct_recode(Education, "No School/Only Kindergarden" = "1",
                                "Grades 1-8" = "2", "Grades 9-11" = "3",
                                "Grade 12 / GED" = "4",
                                "1-3 Years of College" = "5",
                                "4+ Years of College" = "6")) %>%
  mutate(MentHlth = factor(MentHlth)) %>%
  mutate(CholCheck = factor(CholCheck)) %>%
  mutate(DiffWalk = fct_recode(CholCheck, "No" = "0", "Yes" = "1")) %>%
  mutate(Income = factor(Income)) %>%
  mutate(Income = fct_recode(Income, " Less than $10,000" = "1",
                             "$10,000 - $14,999" = "2",
                             "$15,000 - $19,999" = "3",
                             "$20,000 - $24,999" = "4",
                             "$25,000 - $34,999" = "5",
                             "$35,000 - $49,999" = "6",
                             "$50,000 - $74,999" = "7", "$75,000+" = "8"))

us_diabetics_factored
```

```
## # A tibble: 70,692 x 22
##    Diabetes_binary HighBP HighChol CholCheck   BMI Smoker Stroke
##    <fct>           <fct>  <fct>    <fct>     <int> <fct>  <fct>
##  1 Not Diabetic    Yes    No       1            26 No     No
##  2 Not Diabetic    Yes    Yes      1            26 Yes    Yes
##  3 Not Diabetic    No     No       1            26 No     No
##  4 Not Diabetic    Yes    Yes      1            28 Yes    No
##  5 Not Diabetic    No     No       1            29 Yes    No
##  6 Not Diabetic    No     No       1            18 No     No
##  7 Not Diabetic    No     Yes      1            26 Yes    No
##  8 Not Diabetic    No     No       1            31 Yes    No
##  9 Not Diabetic    No     No       1            32 No     No
## 10 Not Diabetic    No     No       1            27 Yes    No
## # ... with 70,682 more rows, and 15 more variables: HeartDiseaseorAttack <fct>,
## #   PhysActivity <fct>, Fruits <fct>, Veggies <fct>, HvyAlcoholConsump <fct>,
## #   AnyHealthcare <fct>, NoDocbcCost <fct>, GenHlth <fct>, MentHlth <fct>,
## #   PhysHlth <int>, DiffWalk <fct>, Sex <fct>, Age <int>, Education <fct>,
## #   Income <fct>
```

We repeat this for the other dataset:

```
pima_diabetics <- pima_diabetics %>%
  mutate(Pregnancies = as.integer(Pregnancies),
         BloodPressure = as.integer(BloodPressure),
         SkinThickness = as.integer(SkinThickness),
         Insulin = as.integer(Insulin),
         Age = as.integer(Age),
         Outcome = as.integer(Outcome))

pima_diabetics
```

```
## # A tibble: 768 x 9
##    Pregnancies Glucose BloodPressure SkinThickness Insulin   BMI
##          <int>   <dbl>         <int>         <int>   <int> <dbl>
##  1           6     148            72            35       0  33.6
##  2           1      85            66            29       0  26.6
##  3           8     183            64             0       0  23.3
##  4           1      89            66            23      94  28.1
##  5           0     137            40            35     168  43.1
##  6           5     116            74             0       0  25.6
##  7           3      78            50            32      88  31
##  8          10     115             0             0       0  35.3
##  9           2     197            70            45     543  30.5
## 10           8     125            96             0       0   0
## # ... with 758 more rows, and 3 more variables: DiabetesPedigreeFunction <dbl>,
## #   Age <int>, Outcome <int>
```

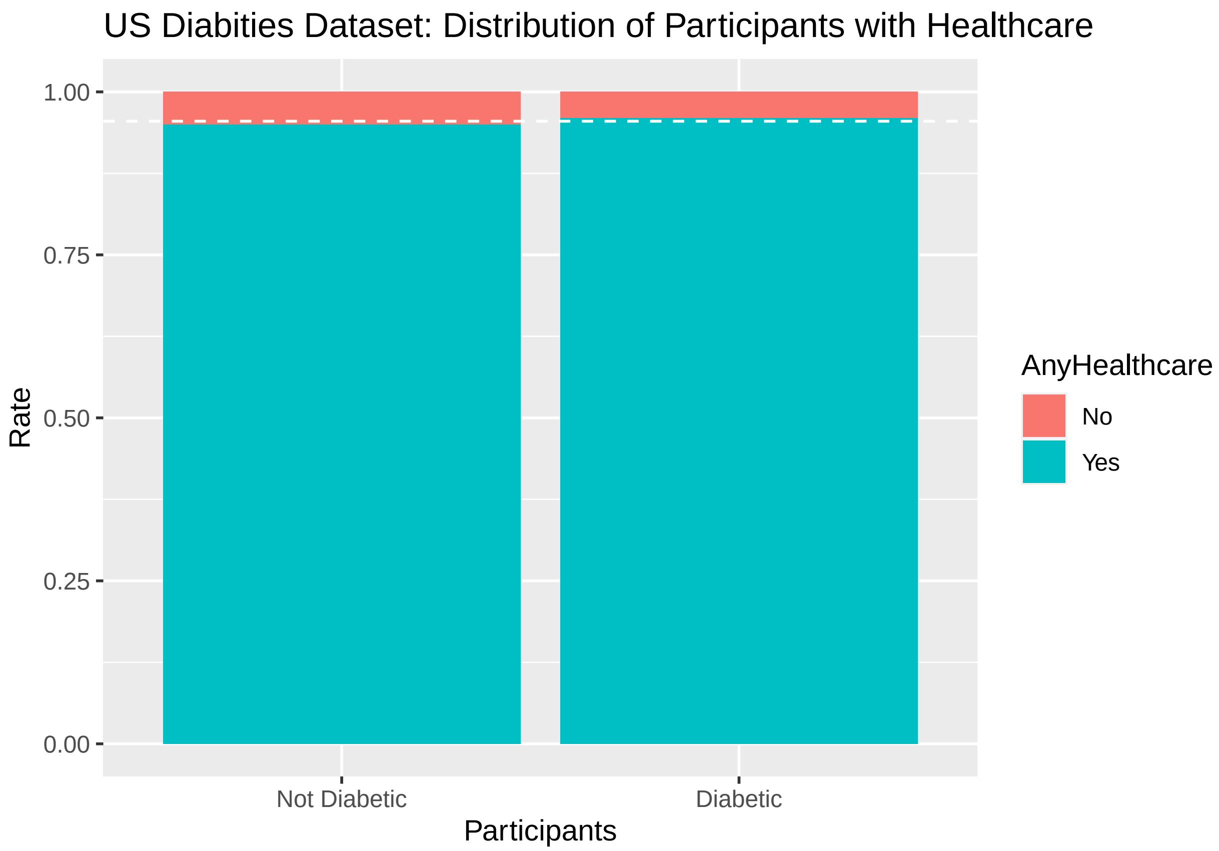
Question 1: What relationships are present between socioeconomic markers and diabetes? (For US Americans)

In the US dataset there are four factors that indicate some relation to socioeconomic status:

- *AnyHealthcare*: Checks to see if you have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc (0 = no 1 = yes).
- *NoDocbcCost*: Checks if there was a time in the past 12 months when a participant needed to see a doctor but could not because of cost (0 = no 1 = yes)?
- *Income*: Income Scale (1-8 Scale based on INCOME2 handbook)
- *Education*: Education level (1-6 Scale based on EDUCA code book)

First, we plot each of them to gain an understanding of the general distribution of the variables. Since they are all categorical, we have four bar plots.

```
### graph ###
ggplot(us_diabetics_factored) +
  geom_bar(aes(x = Diabetes_binary, fill = AnyHealthcare), position = "fill") +
  ggtitle("US Diabities Dataset: Distribution of Participants with Healthcare") +
  xlab("Participants") +
  ylab("Rate") +
  geom_hline(yintercept = mean(us_diabetics$AnyHealthcare), color = "white", linetype = "dashed")
```



```
### stats ###
us_diabetics %>%  summarise(mean_value_ALL = mean(AnyHealthcare))
```

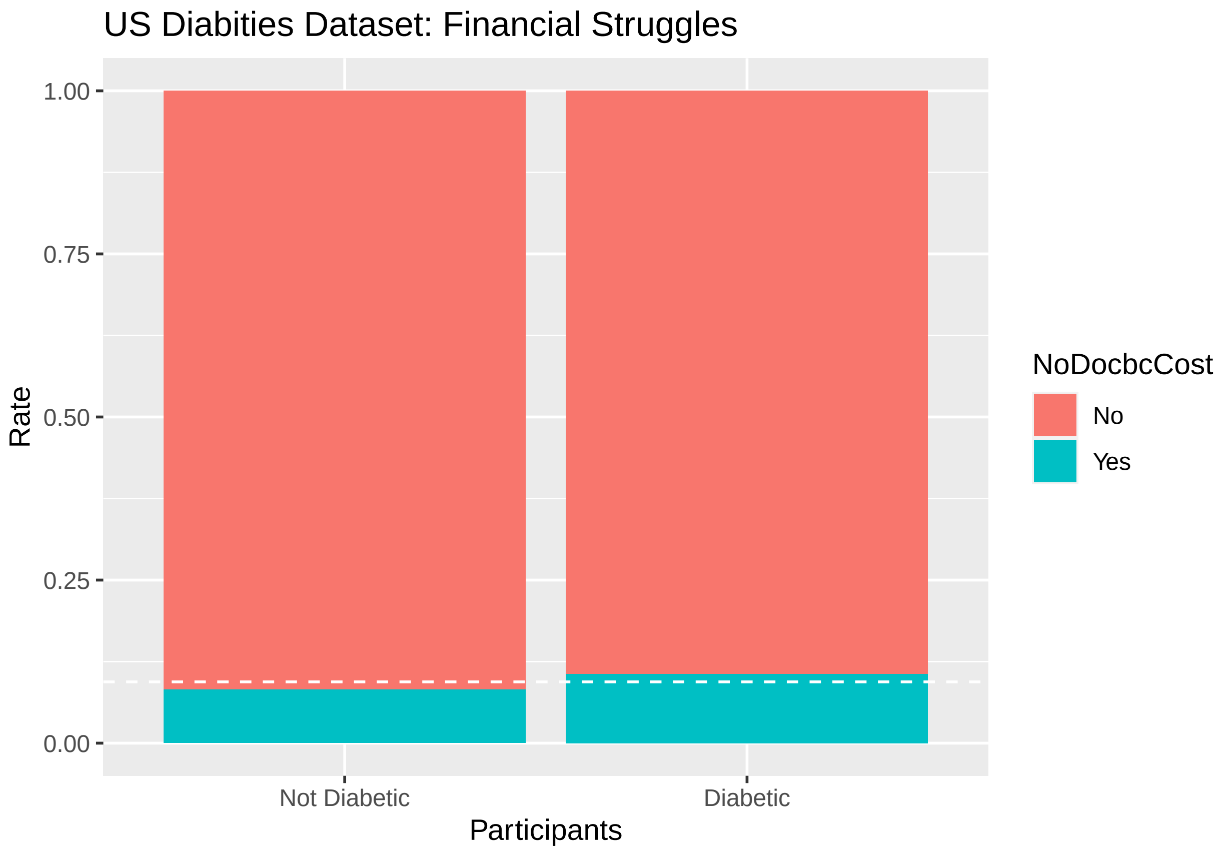
```
## # A tibble: 1 x 1
##   mean_value_ALL
##            <dbl>
## 1          0.955
```

```
us_diabetics %>% group_by(Diabetes_binary) %>% summarise(mean_value = mean(AnyHealthcare))
```

```
## # A tibble: 2 x 2
##   Diabetes_binary mean_value
##             <int>      <dbl>
## 1               0      0.950
## 2               1      0.960
```

95.49% of all people in the dataset have healthcare – 95.01% of non-diabetics have health insurance and 95.97% of diabetics have health insurance. As we can see from the graph above, there does not seem to be a significant difference in regards to non-diabetics and diabetics with regards to having health insurance.

```
### graph ###
ggplot(us_diabetics_factored) +
  geom_bar(aes(x = Diabetes_binary, fill = NoDocbcCost), position = "fill") +
  ggtitle("US Diabities Dataset: Financial Struggles") +
  xlab("Participants") +
  ylab("Rate")  +
  geom_hline(yintercept = mean(us_diabetics$NoDocbcCost), color = "white", linetype = "dashed")
```



```
### stats ###
us_diabetics %>%  summarise(mean_value_ALL = mean(1- NoDocbcCost))
```

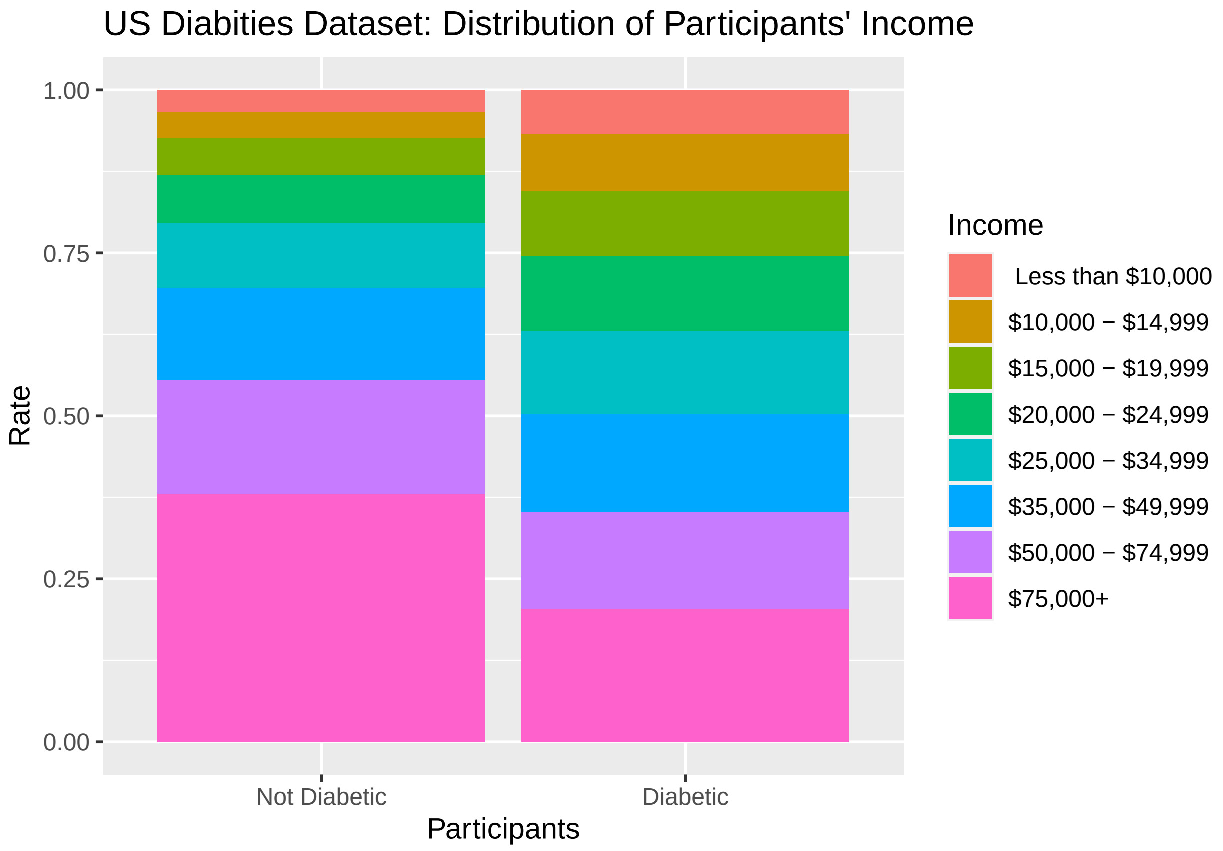
```
## # A tibble: 1 x 1
##   mean_value_ALL
##            <dbl>
## 1          0.906
```

```
us_diabetics %>% group_by(Diabetes_binary) %>% summarise(mean_value = 1- mean(NoDocbcCost))
```

```
## # A tibble: 2 x 2
##   Diabetes_binary mean_value
##               <int>      <dbl>
## 1                 0      0.918
## 2                 1      0.894
```

Here, we see while for both parties going to the doctor was viable, diabetics had a slightly higher reported frequency (by 2.39%) of going to the doctor not be viable.

```
### graph ###
ggplot(us_diabetics_factored) +
  geom_bar(aes(x = Diabetes_binary, fill = Income), position = "fill") +
  ggtitle("US Diabities Dataset: Distribution of Participants' Income") +
  xlab("Participants") +
  ylab("Rate")
```



```
### stats ###
us_diabetics %>%  summarise(mean_value_ALL = mean(Income))
```

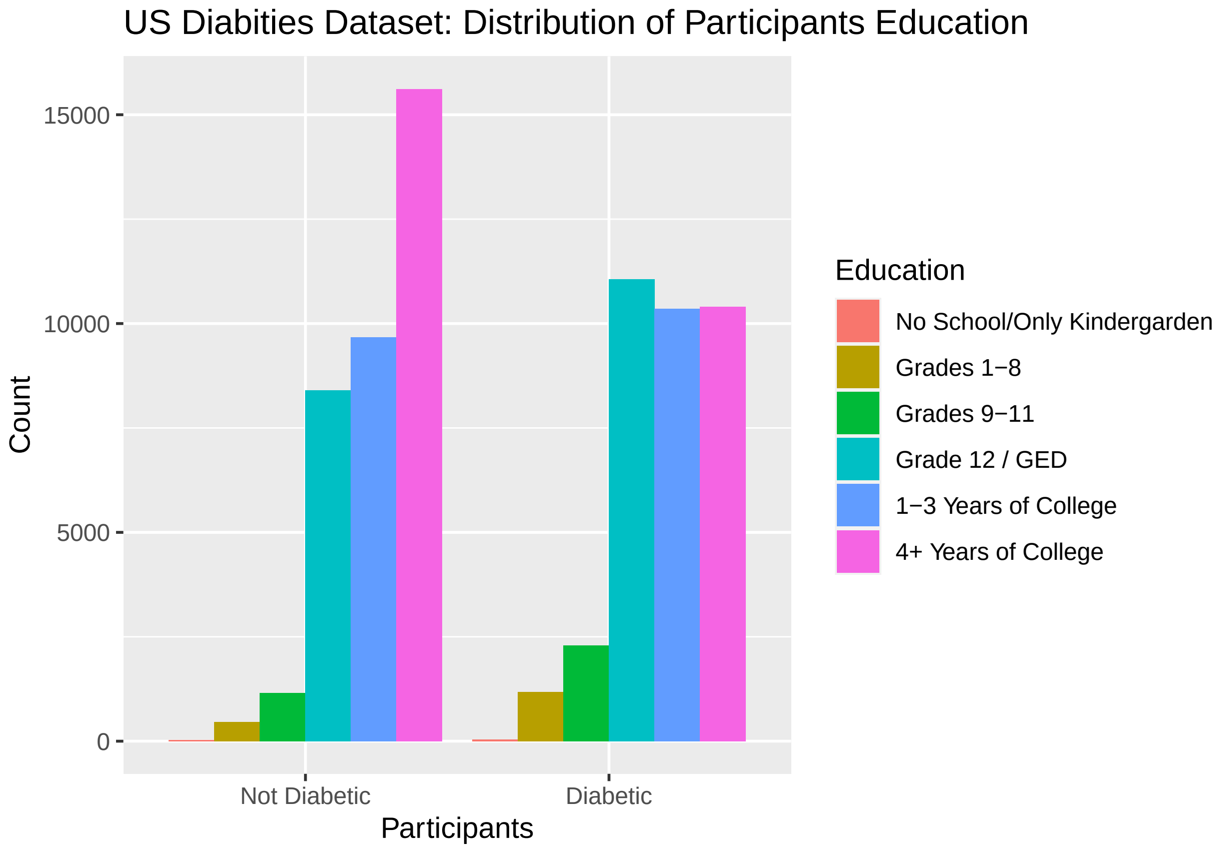
```
## # A tibble: 1 x 1
##   mean_value_ALL
##            <dbl>
## 1           5.70
```

```
us_diabetics %>% group_by(Diabetes_binary) %>% summarise(mean_value = mean(Income))
```

```
## # A tibble: 2 x 2
##   Diabetes_binary mean_value
##             <int>      <dbl>
## 1               0       6.19
## 2               1       5.21
```

This is interesting: non-diabetics tend to be slightly wealthier than diabetics; we see this visually in the graph above, but also when we look at the mean values for each group. The average code for non-diabetics was 6.18 ($35,000 - $49,999) and the average code for diabetics was 5.21 ($25,000 - $34,999).

```
### graph ###
ggplot(us_diabetics_factored) +
  geom_bar(aes(x = Diabetes_binary, fill = Education), position = "dodge") +
  ggtitle("US Diabities Dataset: Distribution of Participants Education") +
  xlab("Participants") +
  ylab("Count")
```



```
### stats ###
us_diabetics %>%  summarise(mean_value_ALL = mean(Education))
```

```
## # A tibble: 1 x 1
##   mean_value_ALL
##            <dbl>
## 1           4.92
```

```
us_diabetics %>% group_by(Diabetes_binary) %>% summarise(mean_value = mean(Education))
```

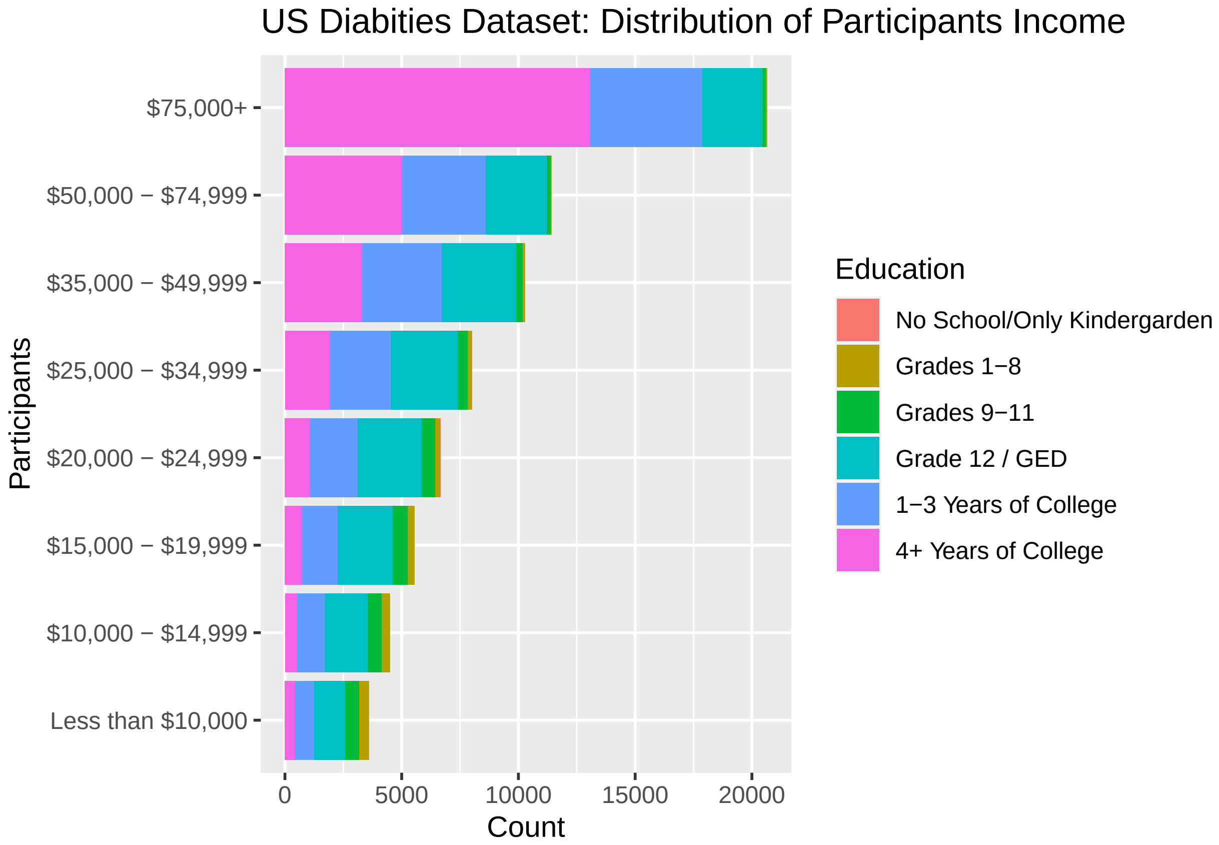
```
## # A tibble: 2 x 2
##   Diabetes_binary mean_value
##             <int>      <dbl>
## 1               0       5.10
## 2               1       4.75
```

There is a noticeable difference between education levels and diabetic standing; non diabetics tend to be more educated than diabetics. We see this above in the graph and also when we look at the average education level for each group; non diabetics average a level of 5.09 (1-3 Years of College) and diabetics average 4.75 level (Grade 12/ GED).

Thus, from our exploratory analysis we see that education level and income seem to be the most significant when we consider a participant's diabetic status.

Education level and income tend to be reflective of each other; let's create a model that takes in income and outputs education level.

```
ggplot(us_diabetics_factored) + geom_bar(aes(x = Income, fill = Education)) + coord_flip() +  ggtitle("
  xlab("Participants") +
  ylab("Count")
```



The relationship between education and income is evident: higher income levels tend to correlate with a higher level of education.

Next, we make a model that uses income and education as a proxy to determine a person's diabetic standing.

In statistics, when there are many features which makes it hard to pick the appropriate features to construct a model or to determine feature importance, one theory called forward model selection, opts to iteratively add features based on their correlation values. I utilized this idea and opted to create multiple models and compare their relative R^2 values.

```
socioeconomic_mod <- lm(us_diabetics$Diabetes_binary ~ Income + Education, data = us_diabetics_factored)
socioeconomic_mod_2 <- lm(us_diabetics$Diabetes_binary ~ Education, data = us_diabetics_factored)
socioeconomic_mod_3 <- lm(us_diabetics$Diabetes_binary ~ Income, data = us_diabetics_factored)
socioeconomic_mod_4 <- lm(us_diabetics$Diabetes_binary ~ Income + Education + NoDocbcCost + AnyHealthca



summary(socioeconomic_mod)
```

```
##
## Call:
## lm(formula = us_diabetics$Diabetes_binary ~ Income + Education,
##     data = us_diabetics_factored)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77463 -0.46592 -0.04932  0.45708  0.67599
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.665545   0.056390  11.803  < 2e-16 ***
## Income$10,000 - $14,999       0.032151   0.010850   2.963 0.003046 **
## Income$15,000 - $19,999      -0.007698   0.010407  -0.740 0.459461
## Income$20,000 - $24,999      -0.034599   0.010105  -3.424 0.000617 ***
## Income$25,000 - $34,999      -0.072131   0.009861  -7.315  2.6e-13 ***
## Income$35,000 - $49,999      -0.111319   0.009599 -11.597  < 2e-16 ***
## Income$50,000 - $74,999      -0.154989   0.009562 -16.208  < 2e-16 ***
## Income$75,000+               -0.253230   0.009226 -27.448  < 2e-16 ***
## EducationGrades 1-8           0.076937   0.057245   1.344 0.178954
## EducationGrades 9-11          0.039028   0.056591   0.690 0.490418
## EducationGrade 12 / GED      -0.011310   0.056116  -0.202 0.840270
## Education1-3 Years of College -0.029736   0.056133  -0.530 0.596291
## Education4+ Years of College  -0.088307   0.056154  -1.573 0.115816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4848 on 70679 degrees of freedom
## Multiple R-squared:  0.05999,    Adjusted R-squared:  0.05983
## F-statistic: 375.9 on 12 and 70679 DF,  p-value: < 2.2e-16
```

```
summary(socioeconomic_mod_2)
```

```
##
## Call:
## lm(formula = us_diabetics$Diabetes_binary ~ Education, data = us_diabetics_factored)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71828 -0.51692 -0.05898  0.48308  0.60031
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.62667    0.05686  11.021  < 2e-16 ***
## EducationGrades 1-8             0.09161    0.05814   1.576   0.1151
## EducationGrades 9-11            0.03942    0.05748   0.686   0.4928
## EducationGrade 12 / GED        -0.05839    0.05697  -1.025   0.3054
## Education1-3 Years of College -0.10974    0.05697  -1.926   0.0541 .
## Education4+ Years of College  -0.22697    0.05694  -3.986 6.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4924 on 70686 degrees of freedom
## Multiple R-squared:  0.03016,    Adjusted R-squared:  0.03009
## F-statistic: 439.7 on 5 and 70686 DF,  p-value: < 2.2e-16
```

```
summary(socioeconomic_mod_3)
```

```
##
## Call:
## lm(formula = us_diabetics$Diabetes_binary ~ Income, data = us_diabetics_factored)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68608 -0.46083 -0.01729  0.48566  0.65151
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.659928   0.008091  81.563  < 2e-16 ***
## Income$10,000 - $14,999  0.026155   0.010864   2.408   0.0161 *
## Income$15,000 - $19,999 -0.017855   0.010392  -1.718   0.0858 .
## Income$20,000 - $24,999 -0.051036   0.010048  -5.079  3.8e-07 ***
## Income$25,000 - $34,999 -0.097631   0.009746 -10.018  < 2e-16 ***
## Income$35,000 - $49,999 -0.145590   0.009404 -15.481  < 2e-16 ***
## Income$50,000 - $74,999 -0.199096   0.009282 -21.450  < 2e-16 ***
## Income$75,000+          -0.311434   0.008770 -35.511  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4862 on 70684 degrees of freedom
## Multiple R-squared:  0.05454,    Adjusted R-squared:  0.05445
## F-statistic: 582.5 on 7 and 70684 DF,  p-value: < 2.2e-16
```

```
summary(socioeconomic_mod_4)
```

```
##
## Call:
## lm(formula = us_diabetics$Diabetes_binary ~ Income + Education +
##     NoDocbcCost + AnyHealthcare, data = us_diabetics_factored)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80097 -0.46973  0.00822  0.45072  0.81742
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.545195   0.056850   9.590  < 2e-16 ***
## Income$10,000 - $14,999       0.027617   0.010835   2.549 0.010810 *
## Income$15,000 - $19,999      -0.009877   0.010391  -0.951 0.341832
## Income$20,000 - $24,999      -0.037643   0.010094  -3.729 0.000192 ***
## Income$25,000 - $34,999      -0.077047   0.009862  -7.812  5.7e-15 ***
## Income$35,000 - $49,999      -0.117217   0.009616 -12.190  < 2e-16 ***
## Income$50,000 - $74,999      -0.162566   0.009596 -16.941  < 2e-16 ***
## Income$75,000+               -0.261888   0.009283 -28.211  < 2e-16 ***
## EducationGrades 1-8           0.074130   0.057145   1.297 0.194557
## EducationGrades 9-11          0.030975   0.056494   0.548 0.583496
## EducationGrade 12 / GED      -0.021174   0.056022  -0.378 0.705463
## Education1-3 Years of College -0.040795   0.056039  -0.728 0.466630
## Education4+ Years of College  -0.100724   0.056062  -1.797 0.072394 .
## NoDocbcCostYes                0.011550   0.006503   1.776 0.075733 .
## AnyHealthcareYes              0.142477   0.009056  15.733  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.484 on 70677 degrees of freedom
## Multiple R-squared:  0.0633, Adjusted R-squared:  0.06311
## F-statistic: 341.1 on 14 and 70677 DF,  p-value: < 2.2e-16
```

Here, from the summary we see that the R^2 value for our first model (income and education) is 0.059 or roughly 6%. Thus, only 6% of the variation in diabetic standing can be attributed to education and income. Similarly, the model with just income had an R^2 value of 0.4862, and the model with just education had an R^2 value of 0.03. For fun, I added a model with all the socioeconomic markers in this data set; it had an R^2 value of 0.063; this was not much higher than what we saw with just education and income, which makes sense – the other two factors didn't seem to differentiate between the two groups in a visible way. Thus, income seemed to explain slightly more variation than education with regards to diabetic standings which makes sense; healthcare, especially in America, is expensive and so with a higher income you are more likely to get the resources you need to either catch diabetic symptoms early or manage your diabetes well.

I did some literature readings to see how much evidence backs up this claim. According to WHO, income is extremely linked to diabetic status; in high-income countries, diabetic mortality rates decreased between 2000 and 2010, but in low-income countries the rates increased. Similarly, according to the American Diabetes Association, the "…prevalence of diabetes disproportionately impacted lower-income populations."

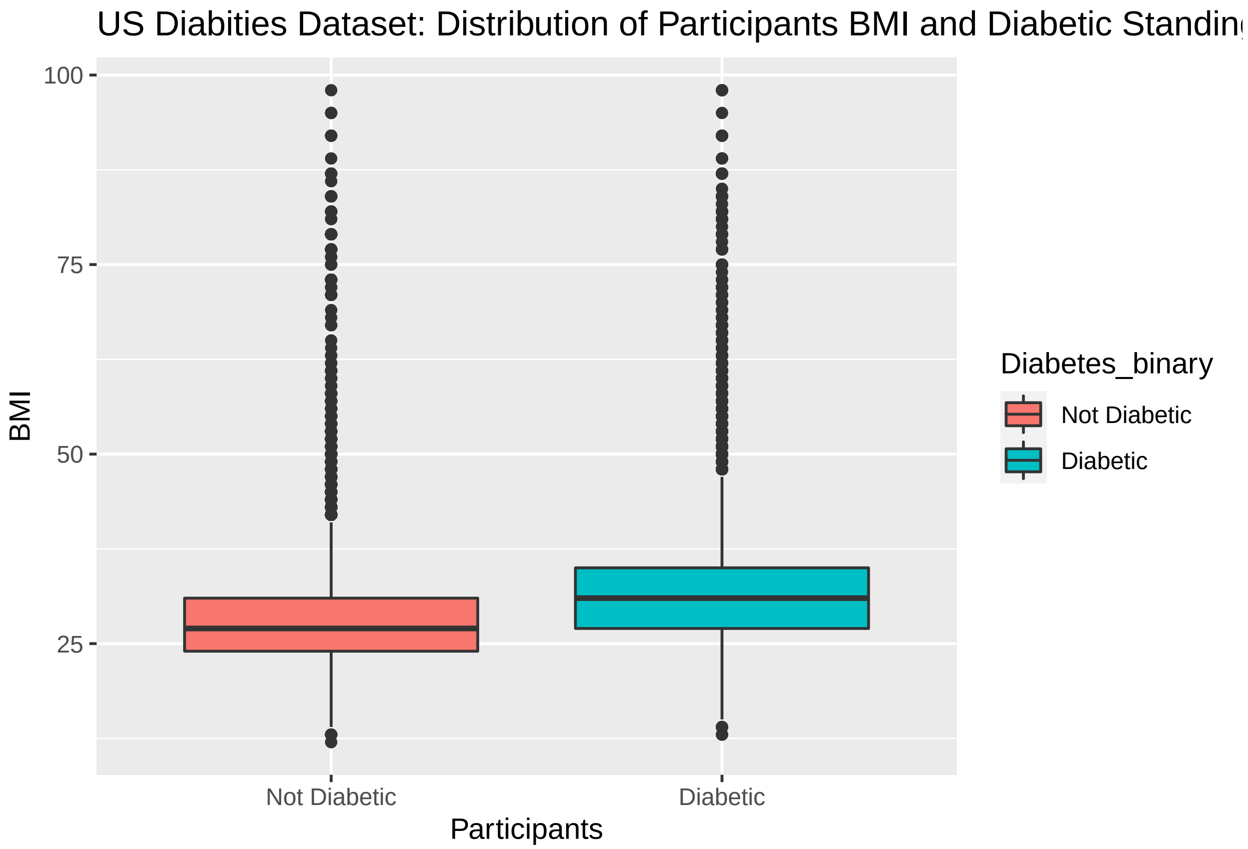
The articles above also discuss the intersect between race and income, an avenue I think would be interesting to consider in future data sets that include ethnicity data.

Thus, we see that diabetic standing is correlated with income.

Question 2: Is there a relationship between BMI and diabities? If so, is there a relationship between BMI and blood pressure and/or BMI and cholesterol that could perhaps compound the relationship?

Let's begin by graphing the relationship between BMI and diabetes:

```
ggplot(us_diabetics_factored) + geom_boxplot(aes(x = Diabetes_binary, y = BMI, fill = Diabetes_binary))
  ggtitle("US Diabities Dataset: Distribution of Participants BMI and Diabetic Standing") +
  xlab("Participants") +
  ylab("BMI")
```



```
us_diabetics %>% group_by(Diabetes_binary) %>%  summarize(median = median(BMI))
```

```
## # A tibble: 2 x 2
##   Diabetes_binary median
##             <int>  <dbl>
## 1               0     27
## 2               1     31
```

Thus, we can see that diabetic patients have a slightly higher BMI (31) than non diabetics (27). Note that a healthy BMI is 18.5 - 24.9,27 falls into overweight, and 31 falls into obese.

To check if the difference between these two are actually significant, we will perform a T-test. A T-test is a statistical test that allows us to compare the means of two different groups (for us, this is the group with diabetes and the group without diabetes). It allows us to determine if the BMI has a significant effect on the population.

```
t.test(BMI ~ Diabetes_binary, var.equal = TRUE, data = us_diabetics)
```

```
##
```

```
##  Two Sample t-test
##
## data:  BMI by Diabetes_binary
## t = -81.591, df = 70690, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -4.274321 -4.073781
## sample estimates:
## mean in group 0 mean in group 1
##        27.76996        31.94401
```

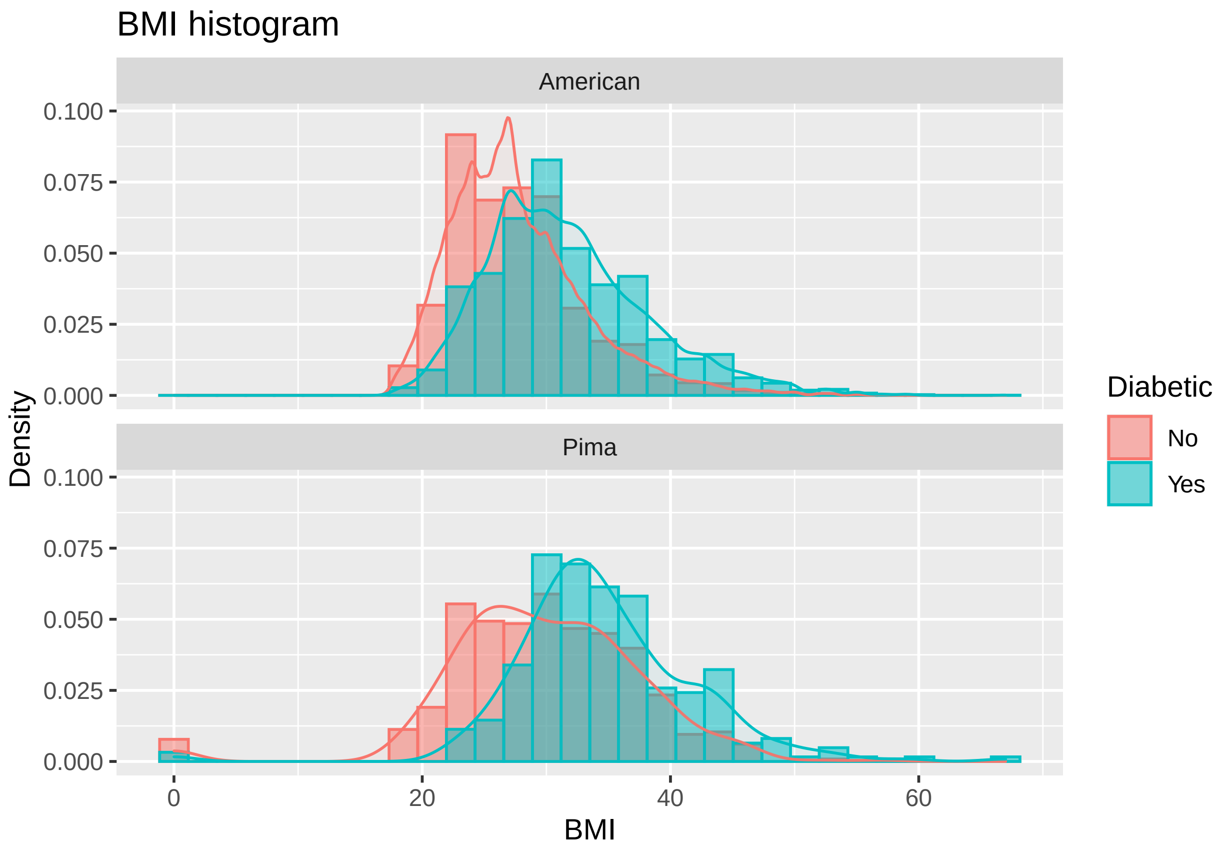
Here, we see that the p-value is quite small (2.2e-16). This tells us that it is quite likely that the difference in means between the two populations (27.7 for non diabetics and 31.9 for diabetics) has a very minuscule probability of the difference between them being by chance and is thus statistically significant. Thus, we can conclude that BMI is indeed an important factor.

I was also interested to see if the Pima Indian dataset had similar BMI and diabetes distributions as the American dataset. I used a semi-join to filter the American diabetes dataset to include the BMI ranges in the Pima Indian dataset so that we could see the comparable differences in an easier way.

```r
pima_diabetics2 <- pima_diabetics %>% mutate(BMI = as.integer(BMI))
us_pima_bmi_match <- semi_join(us_diabetics, pima_diabetics2, by = c("BMI"))

us_bmi_match <- us_pima_bmi_match %>%
  select(Diabetes_binary, BMI) %>%
  mutate(Data = "American") %>%
  rename(Diabetic = Diabetes_binary) %>%
  mutate(Diabetic = ifelse(Diabetic == 1, "Yes", "No"))


pima_bmi_match <- pima_diabetics2 %>%
  select(Outcome, BMI) %>%
  mutate(Data = "Pima") %>%
  rename(Diabetic = Outcome) %>%
  mutate(Diabetic = ifelse(Diabetic == 1, "Yes", "No"))

combined_ds <- rbind(us_bmi_match, pima_bmi_match)


combined_ds %>%
  mutate(Diabetic2 = ifelse(Diabetic == "Yes", 1, 0))  %>%
  filter(Data == "American" & Diabetic2 == 0)
```

```
## # A tibble: 35,005 x 4
##    Diabetic   BMI Data     Diabetic2
##    <chr>    <int> <chr>        <dbl>
##  1 No          26 American         0
##  2 No          26 American         0
##  3 No          26 American         0
##  4 No          28 American         0
##  5 No          29 American         0
##  6 No          18 American         0
##  7 No          26 American         0
##  8 No          31 American         0
##  9 No          32 American        0
## 10 No          27 American        0
## # ... with 34,995 more rows
```

```
ggplot(combined_ds, aes(x=BMI, color=Diabetic, fill=Diabetic)) +
geom_histogram(aes(y=..density..), position="identity", alpha=0.5)+
geom_density(alpha=0.05)+
labs(title="BMI histogram",x="BMI", y = "Density")+
  facet_wrap(~ Data,  ncol = 1)
```

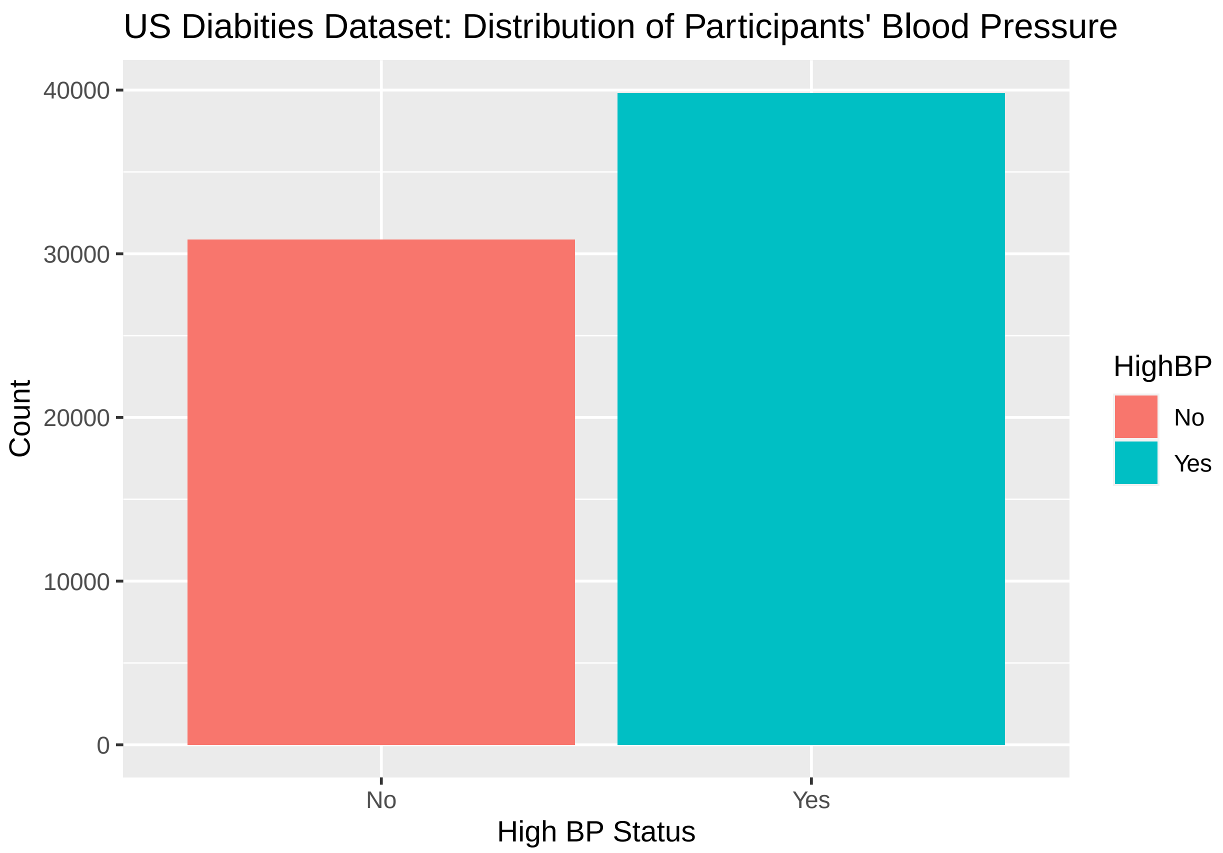


It is interesting that both BMI ranges seem to be normally distributed, and for both groups non diabetics tend to have a lower BMI than diabetics. This analysis was performed to ensure that the trend found in the American data set was mirrored in other populations.For the rest of the analysis only the US dataset will be used.

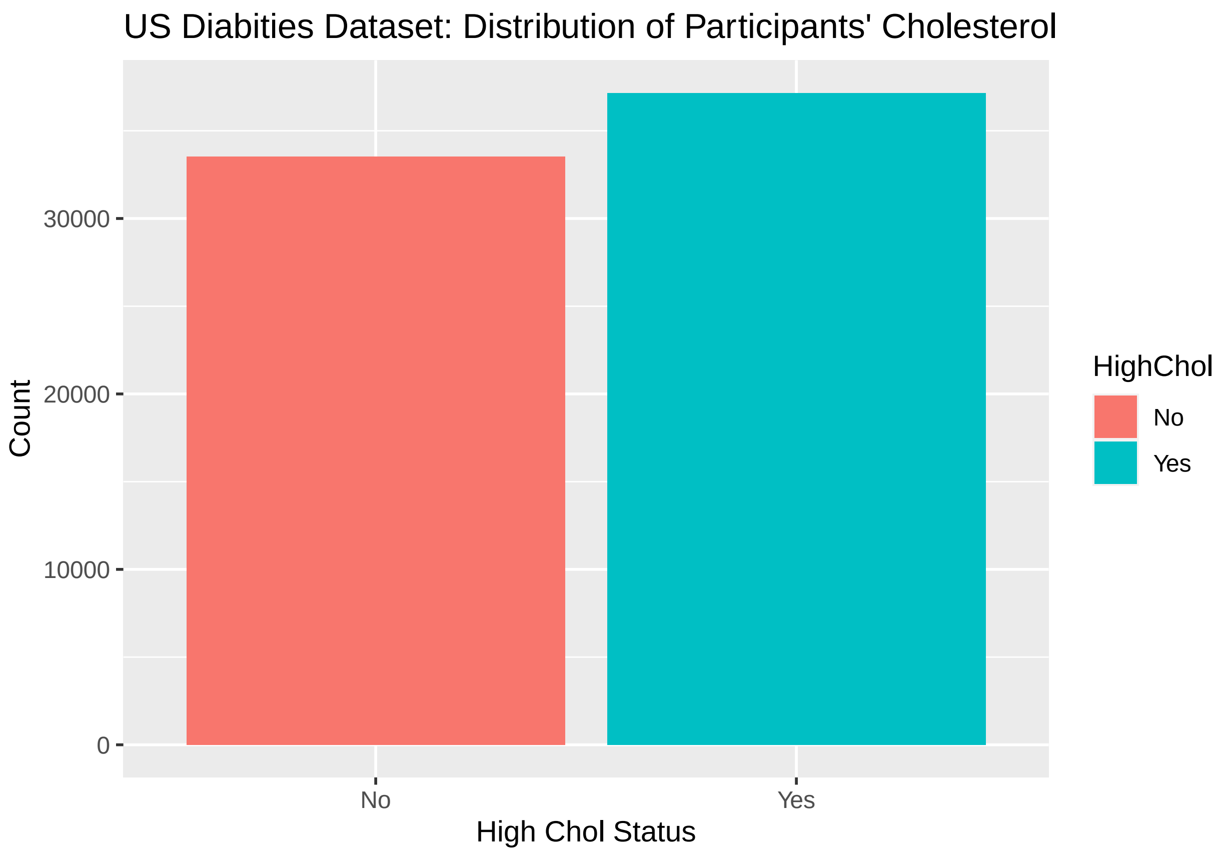
Evidently, BMI is an important factor. However,what if BMI's relationship with diabetes is confounded by blood pressure or cholesterol?

First, let's graph the distribution of blood pressure and cholesterol in the data:

```
ggplot(us_diabetics_factored) + geom_bar(aes(x = HighBP, fill = HighBP)) +
  ggtitle("US Diabities Dataset: Distribution of Participants' Blood Pressure") +
  xlab("High BP Status") +
  ylab("Count")
```



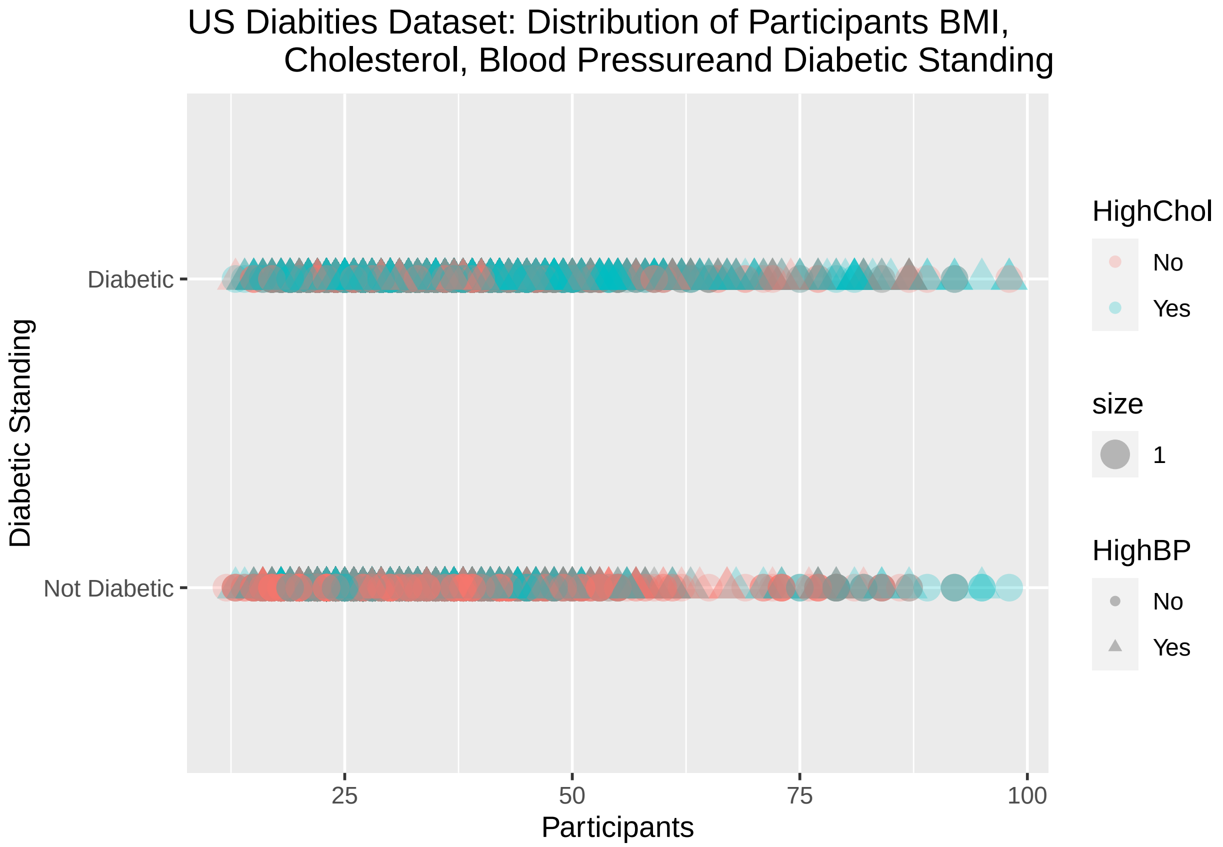
```
ggplot(us_diabetics_factored) + geom_bar(aes(x = HighChol, fill = HighChol)) +
  ggtitle("US Diabities Dataset: Distribution of Participants' Cholesterol") +
  xlab("High Chol Status") +
  ylab("Count")
```



Thus, we see that in the data set there are more people with high cholesterol than not (35700+ participants), but not by a lot (32500 participants). Similarly, there are more people with high blood pressure than not; roughly 30000 participants had normal blood pressure and 40000 participants had high blood pressure.

Let's visualize this data in conjunction with diabetes and BMI.

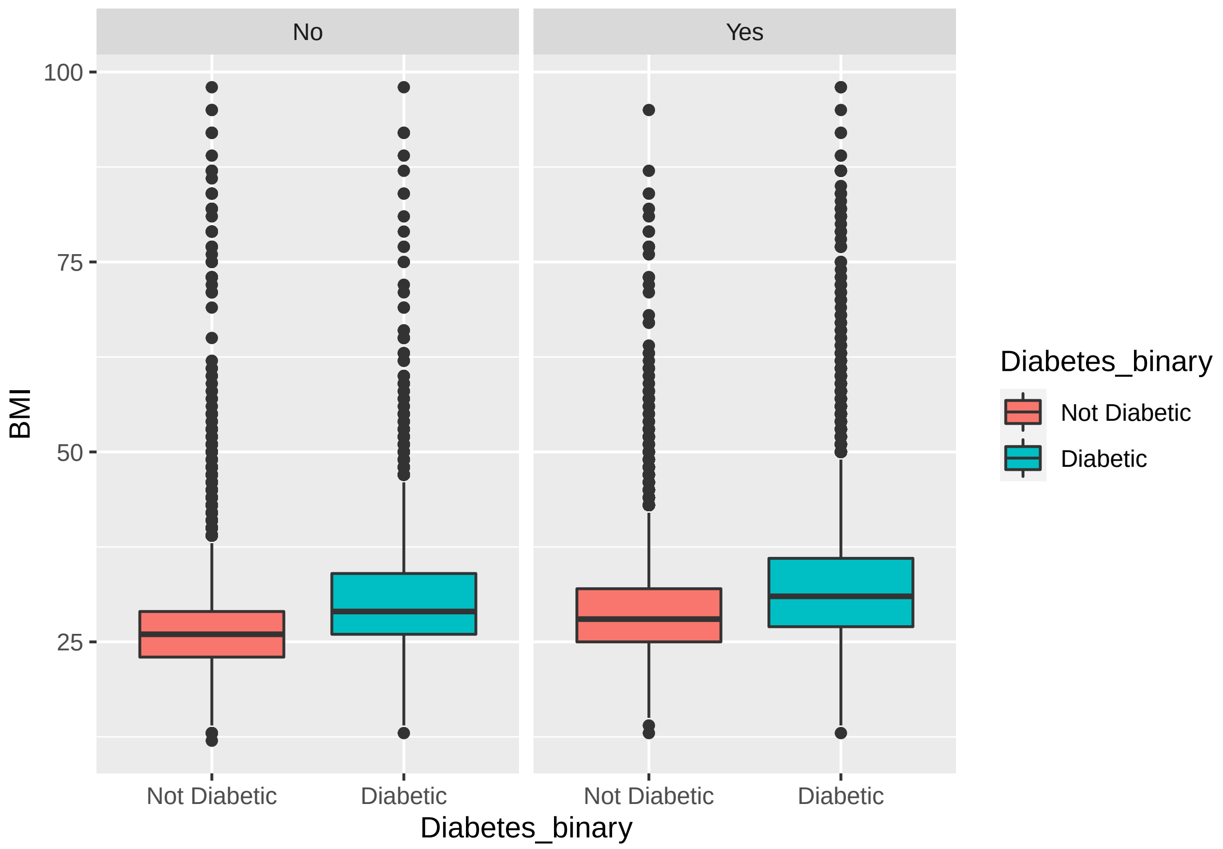
```
ggplot(us_diabetics_factored) +
  geom_point(aes(x = BMI, y = Diabetes_binary,
                 shape = HighBP,color = HighChol, size = 1),
             alpha = 0.25) +
  ggtitle("US Diabities Dataset: Distribution of Participants BMI,
          Cholesterol, Blood Pressureand Diabetic Standing") +
  xlab("Participants") +
  ylab("Diabetic Standing")
```



Here we see that there does seem to be a tendency between having normal cholesterol and no diabetes. Similarly, there seems to be a relationship between having high blood pressure and being diabetic. However, a lot of the data points are on top of each other, and even with adjusted alpha levels and sizes it is a little hard to visualize or quantify. It is difficult to understand if the relationship between BMI and diabetes is not affected by cholesterol and/or BP. Let's perform more specific explorations.

We graph the BMI distribution with respect to blood pressure:

```
ggplot(us_diabetics_factored) +
  geom_boxplot(aes(x = Diabetes_binary, y = BMI, fill = Diabetes_binary)) +
  facet_wrap(~ HighBP)
```



```
us_diabetics %>%
  group_by(HighBP, Diabetes_binary) %>%
  summarize(medianBMI = median(BMI), meanBMI = mean(BMI))
```

```
## `summarise()` has grouped output by 'HighBP'. You can override using the `.groups` argument.
```

```
## # A tibble: 4 x 4
## # Groups:   HighBP [2]
##   HighBP Diabetes_binary medianBMI meanBMI
##    <int>           <int>     <dbl>   <dbl>
## 1      0               0        26    26.9
## 2      0               1        29    30.4
## 3      1               0        28    29.2
## 4      1               1        31    32.4
```

Here we see that the BMI measurements tend to mirror each other for participants with high blood pressure and participants without high blood pressure (non diabetics tend to have a lower BMI regardless). However, non diabetics and diabetics that have normal blood pressure tend to have a median and mean blood BMI lower than non diabetics and diabetics with a high blood pressure.

Here, we perform a T-test for each blood pressure group for each diabetic standing to see if blood pressure affects the BMI in a sizable way.

```
us_diabetics_dib <- us_diabetics %>% filter(Diabetes_binary == 0)
us_diabetics_nodib <- us_diabetics %>% filter(Diabetes_binary == 1)
```

```
t.test(BMI ~ HighBP, var.equal = TRUE, data = us_diabetics_dib)
```

```
##
##  Two Sample t-test
##
## data:  BMI by HighBP
## t = -34.356, df = 35344, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.429646 -2.167381
## sample estimates:
## mean in group 0 mean in group 1
##        26.90976        29.20827
```

```
t.test(BMI ~ HighBP, var.equal = TRUE, data = us_diabetics_nodib)
```

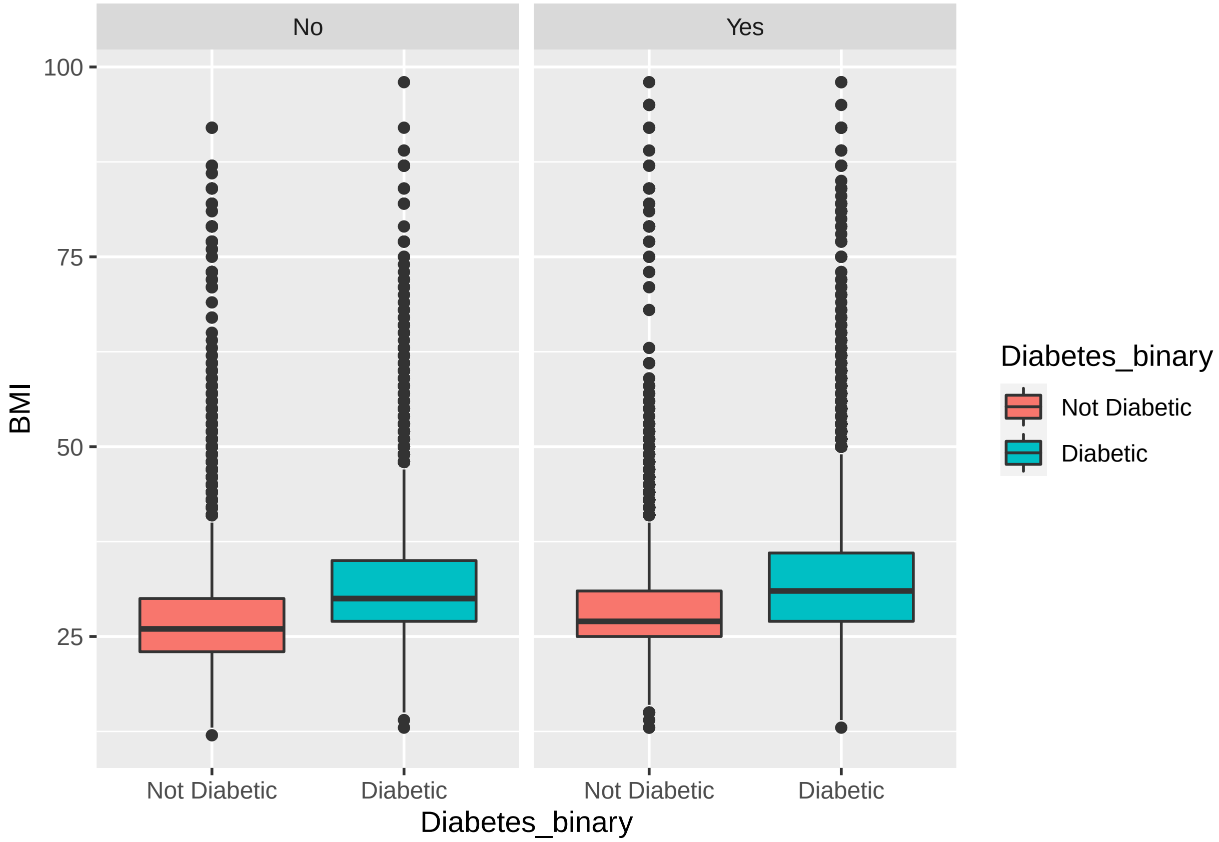
```
##
##  Two Sample t-test
##
## data:  BMI by HighBP
## t = -22.205, df = 35344, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.178490 -1.825095
## sample estimates:
## mean in group 0 mean in group 1
##        30.43731        32.43911
```

Here, for both groups, the p-value is quite small (< 2.2e-16), and thus we can stipulate that there is a statistically significant difference between the BMI groups that have diabetes and high blood pressure and groups that have no diabetes and high blood pressure as well as a statistically significant difference between the BMI of groups that have diabetes and normal blood pressure and groups that don't have diabetes and have normal blood pressure.

Thus, there exists a relationship between BMI, diabetes, and blood pressure.

We repeat this study for cholesterol as well:

```
ggplot(us_diabetics_factored) +
  geom_boxplot(aes(x = Diabetes_binary, y = BMI, fill = Diabetes_binary)) +
  facet_wrap(~ HighChol)
```



```
us_diabetics %>%
  group_by(HighChol, Diabetes_binary) %>%
  summarize(medianBMI = median(BMI), meanBMI = mean(BMI))
```

```
## ‘summarise()‘ has grouped output by ’HighChol’. You can override using the ‘.groups‘ argument.
```

```
## # A tibble: 4 x 4
## # Groups:   HighChol [2]
##   HighChol Diabetes_binary medianBMI meanBMI
##      <int>           <int>     <dbl>   <dbl>
## 1        0               0        26    27.4
## 2        0               1        30    31.7
## 3        1               0        27    28.4
## 4        1               1        31    32.1
```

Once again, the relative distributions between high and low blood pressure mirror each other; diabetics consistently have a higher BMI than non diabetics. However, as we can see from the table, having high cholesterol tends to make the BMI slightly higher. To check if this difference is statistically significant we perform a T-test.

```
t.test(BMI ~ HighChol, var.equal = TRUE, data = us_diabetics_dib)
```

```
##
##  Two Sample t-test
```

```
##
## data:  BMI by HighChol
## t = -15.777, df = 35344, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.197696 -0.932993
## sample estimates:
## mean in group 0 mean in group 1
##        27.36376        28.42910
```

```
t.test(BMI ~ HighChol, var.equal = TRUE, data = us_diabetics_nodib)
```

```
##
##  Two Sample t-test
##
## data:  BMI by HighChol
## t = -4.2786, df = 35344, p-value = 1.886e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.5195576 -0.1930900
## sample estimates:
## mean in group 0 mean in group 1
##        31.70523        32.06156
```

The p-values for both groups were quite small (< 2.2e-16, and 1.886e-05) respectively. So, there is a statistically significant difference between the BMI groups that have diabetes and high cholesterol and groups that have no diabetes and high cholesterol as well as a statistically significant difference between the BMI of groups that have diabetes and normal cholesterol and groups that don't have diabetes and have normal cholesterol.

To validate the findings from this exploration, I did some research. According to a John Hopkins Medicine paper, "If you have diabetes, you are twice as likely to have high blood pressure." Thus, our visual titled "US Diabetes Dataset: Distribution of Participants BMI, Cholesterol, Blood Pressure and Diabetic Standing" makes sense. Similarly, the National Library of Medicine (NLM) released a paper that stated "Body Mass Index is Strongly Associated with Hypertension." This matches what we found in our box plots and statistical tests. Similarly, there is literature from NLM that states that diabetics tend to develop cholesterol problems due to both the medication and nature of diabetes. According to the CDC, "69% had high blood pressure, and 44% had high cholesterol" for adults diagnosed with diabetes.

Thus, our findings that these three variables are indeed related match the scientific community's consensus as well!

Question 3: Which health diagnositcs are good predictors for diabeties? (For US Americans) & Which health diagnositcs are good predictors for diabeties? (For Pima Indians) How do they differ?

First, we split our data into training and testing data. We will adopt the 75/25 practice.

```
set.seed(3476)
trainIndex <- createDataPartition(us_diabetics$Diabetes_binary, p = .75,
                                  list = FALSE,
                                  times = 1)
us_train <- us_diabetics[ trainIndex,]
us_test <- us_diabetics[-trainIndex,]
```

Now, we have our training and test datasets.Let's create our preliminary model first.

```
us_train %>% group_by(Diabetes_binary) %>% summarize(count = n())
```

```
## # A tibble: 2 x 2
##   Diabetes_binary count
##             <int> <int>
## 1               0 26510
## 2               1 26510
```

Here, we see that there is an even number of people with diabetes and without diabetes. Thus, our baseline model should assume that either everyone has diabetes or nobody does; our accuracy regardless will be 50%.

Theoretically, the best model occurs when we use all the features.

```
us_diabetics_mod_ALL <- lm(Diabetes_binary ~ GenHlth + HighBP + HighChol + CholCheck + BMI + Smoker + S
  HeartDiseaseorAttack + PhysActivity + Fruits + Veggies + HvyAlcoholConsump +
  AnyHealthcare + NoDocbcCost + NoDocbcCost + MentHlth + PhysHlth + DiffWalk +
  Sex + Age + Education + Income, data = us_diabetics)

us_train %>%
  add_predictions(us_diabetics_mod_ALL)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Diabetes_binary == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.747
```

Thus, the best model with all the features has an accuracy of 74.7% on the training data. Let's test the testing data next:

```
us_test %>%
  add_predictions(us_diabetics_mod_ALL)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Diabetes_binary == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.747
```

On the testing data, this model had an efficacy of 74.7%. This is exactly the same (with 3 significant figures) to the other training data, and so we do not need to worry about overfitting.

However, we prioritize the simplest model, and so let's do an exploration to see what the most important factors are; printing out the summary of the model will provide us with more information about the p-values of the coefficients.

```
summary(us_diabetics_mod_ALL)
```

```
##
## Call:
## lm(formula = Diabetes_binary ~ GenHlth + HighBP + HighChol +
##     CholCheck + BMI + Smoker + Stroke + HeartDiseaseorAttack +
##     PhysActivity + Fruits + Veggies + HvyAlcoholConsump + AnyHealthcare +
##     NoDocbcCost + NoDocbcCost + MentHlth + PhysHlth + DiffWalk +
##     Sex + Age + Education + Income, data = us_diabetics)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55380 -0.33167  0.03057  0.33686  1.25612
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.5860745  0.0176034 -33.293  < 2e-16 ***
## GenHlth               0.1048536  0.0019300  54.328  < 2e-16 ***
## HighBP                0.1560050  0.0036716  42.489  < 2e-16 ***
## HighChol              0.1092921  0.0033986  32.158  < 2e-16 ***
## CholCheck             0.1638210  0.0102258  16.020  < 2e-16 ***
## BMI                   0.0118392  0.0002387  49.593  < 2e-16 ***
## Smoker               -0.0018721  0.0032557  -0.575 0.565285
## Stroke                0.0252435  0.0067701   3.729 0.000193 ***
## HeartDiseaseorAttack  0.0488488  0.0048118  10.152  < 2e-16 ***
## PhysActivity         -0.0069207  0.0036958  -1.873 0.061132 .
## Fruits               -0.0051995  0.0033656  -1.545 0.122375
## Veggies              -0.0113480  0.0040221  -2.821 0.004782 **
## HvyAlcoholConsump    -0.1228700  0.0078163 -15.720  < 2e-16 ***
## AnyHealthcare         0.0133504  0.0078881   1.692 0.090559 .
## NoDocbcCost          -0.0013439  0.0057291  -0.235 0.814534
## MentHlth             -0.0008718  0.0002157  -4.041 5.32e-05 ***
## PhysHlth             -0.0015972  0.0002025  -7.886 3.17e-15 ***
## DiffWalk              0.0281748  0.0045367   6.210 5.31e-10 ***
## Sex                   0.0447107  0.0032740  13.656  < 2e-16 ***
## Age                   0.0246927  0.0006337  38.968  < 2e-16 ***
## Education            -0.0066329  0.0017527  -3.784 0.000154 ***
## Income               -0.0109883  0.0008930 -12.305  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4155 on 70670 degrees of freedom
## Multiple R-squared:  0.3095, Adjusted R-squared:  0.3093
## F-statistic:  1508 on 21 and 70670 DF,  p-value: < 2.2e-16
```

We see that HighBP, HighChol, CholCheck, BMI, GenHlth, Stroke, HeartDiseaseorAttack,PhysActivity, Veggies, HvyAlcoholConsump,MentHlth, PhysHlth, DiffWalk,Sex,Age, Education, and Income all have three asterisks (***). This tells us that these are potentially relevant to our model. In our earlier questions, we determined that education and income are probably correlated, and income had a greater correlation with diabetes. Thus, in our first model we will include income, not education. Next, we found out that BMI, HighBP, High Chol have statistically significant relations with each other and diabetes. Thus, we include them as well.

```
us_diabetics_mod_2 <- lm(Diabetes_binary ~  HighBP + HighChol + BMI + Income, data = us_diabetics)

us_train %>%
  add_predictions(us_diabetics_mod_2)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Diabetes_binary == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.712
```

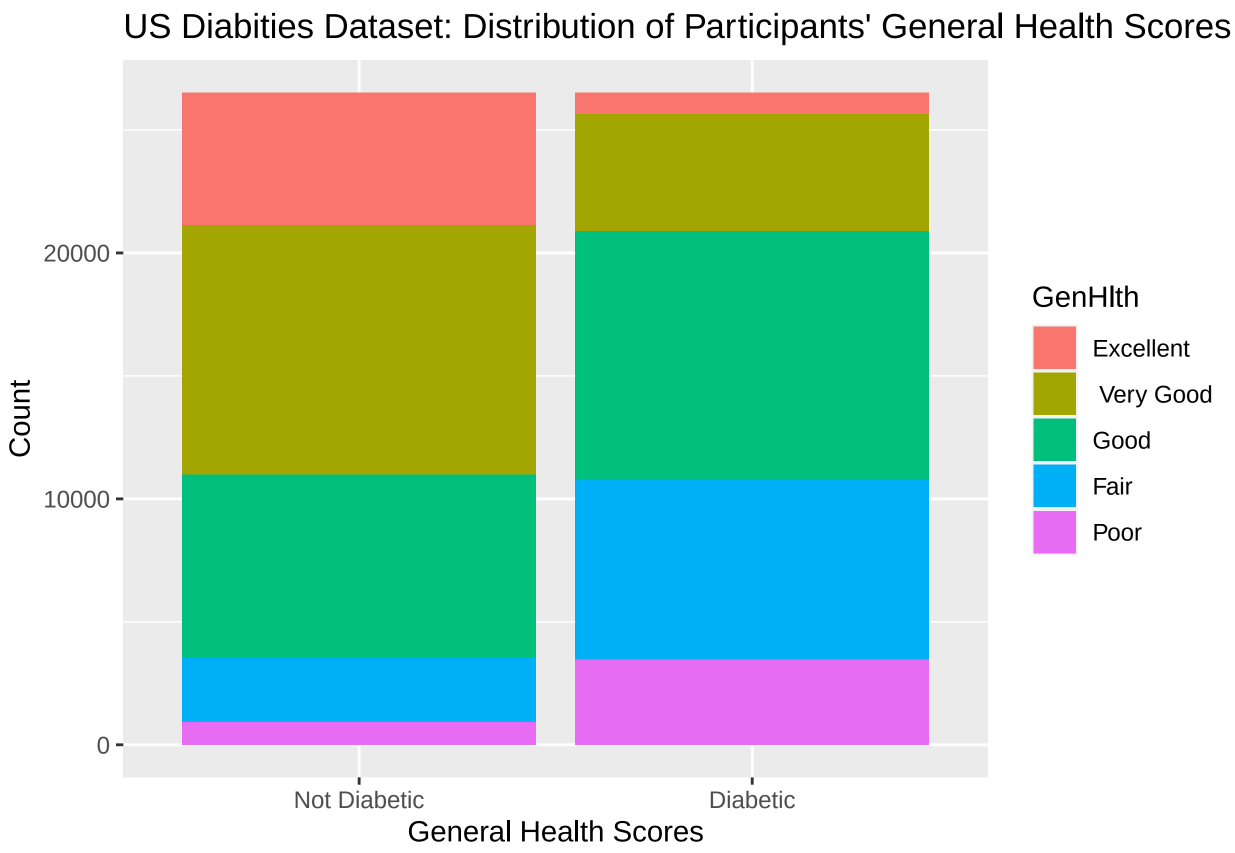
This is pretty decent – 71%. The literature I read suggested that blood pressure is more related with diabetes than cholesterol, so let's remove cholesterol from the model.

```
us_diabetics_mod_3 <- lm(Diabetes_binary ~  HighBP + BMI + Income, data = us_diabetics)

us_train %>%
  add_predictions(us_diabetics_mod_3)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Diabetes_binary == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

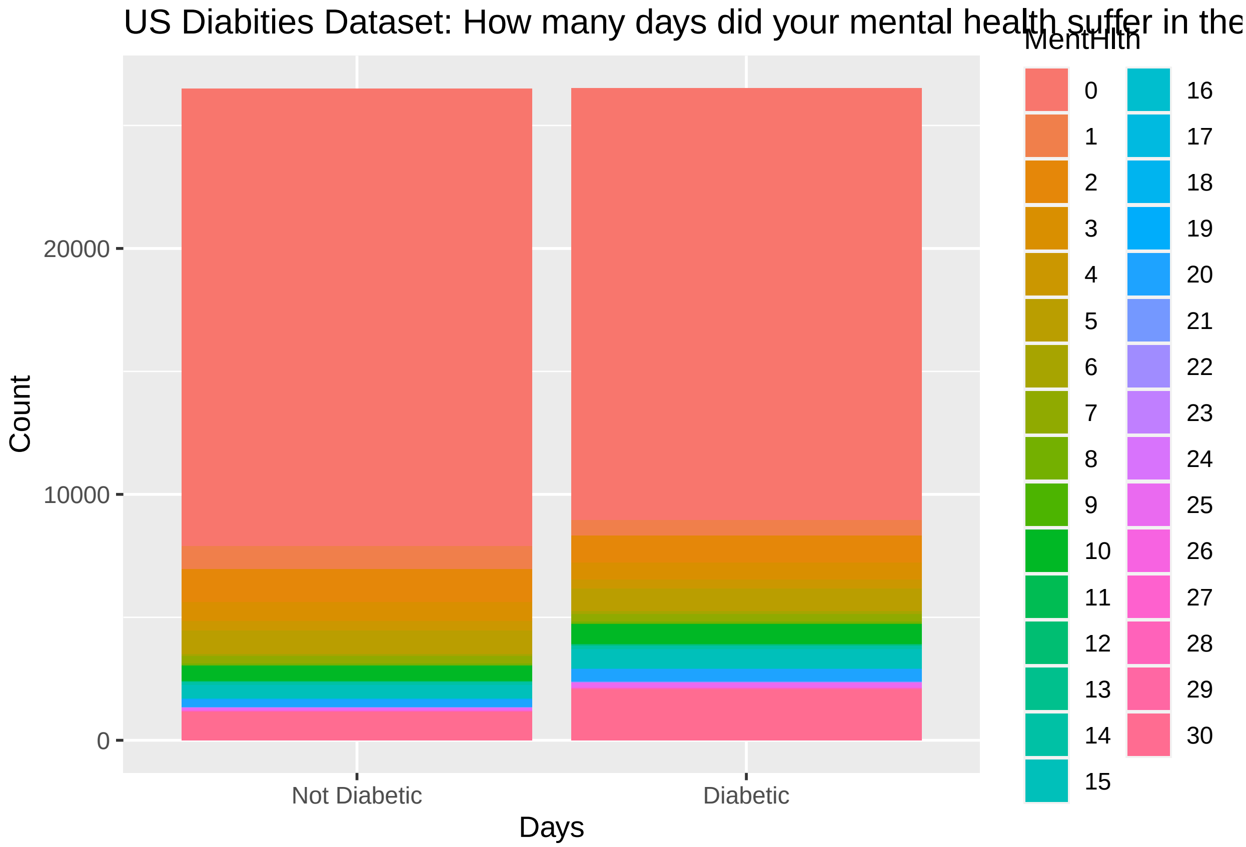
```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.702
```

Here, we see the efficacy only went down by 1%. Thus, we will abstain from using cholesterol data in this model. I want my model to be agnostic of potentially subjective features. Thus, I will abstain from using all three features (GenHlth, MentHlth and PhysHlth (measurement of one's general health (scale of 1-5), mental health (number of days in the last month they felt there mental health was poor), and physical health (answer to the question: "Did you exercise in the last 30 days?"))), since they are features that a participant fills out by themselves. Since all three features were statistically significant, I will graph their relationship with diabetes and select the one that has the most visible impact for the next iteration of the model.

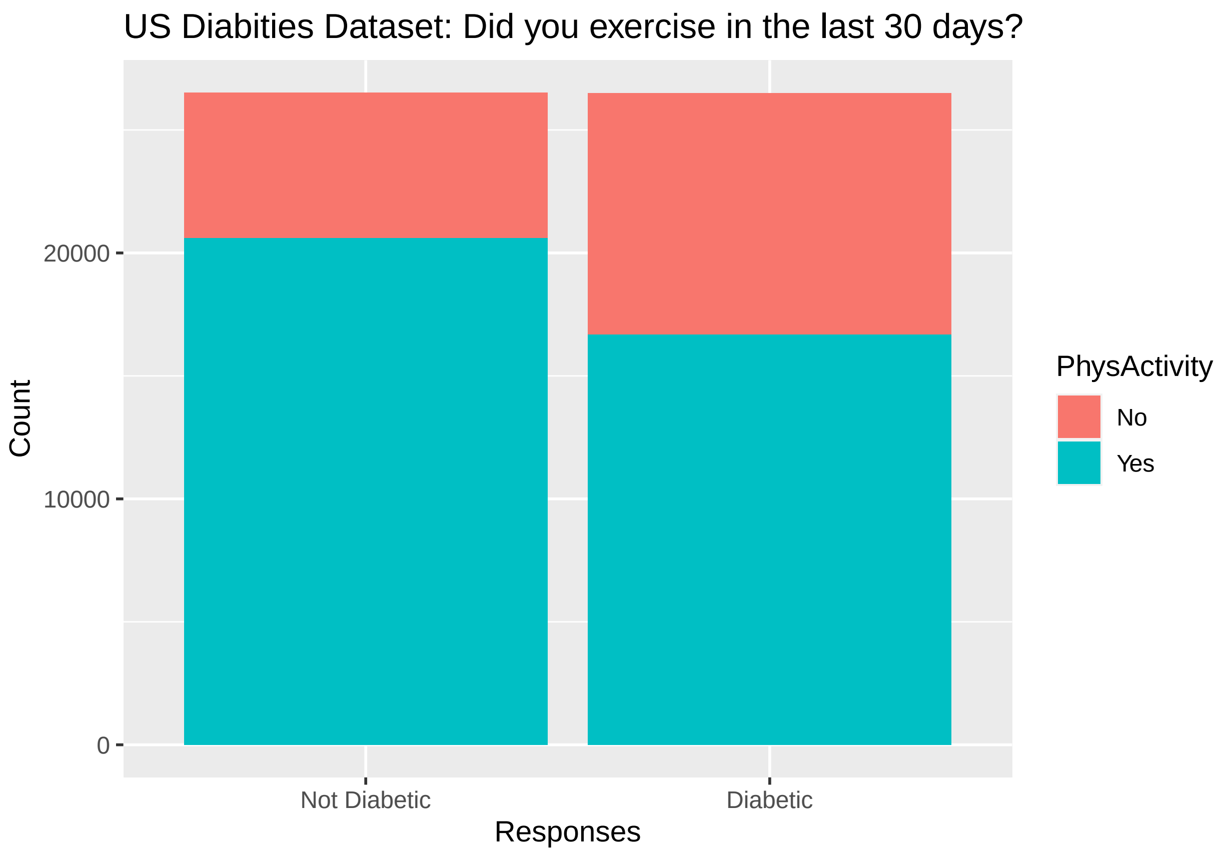
```
us_train %>%
  mutate(Diabetes_binary = factor(Diabetes_binary)) %>%
  mutate(Diabetes_binary = fct_recode(Diabetes_binary, "Not Diabetic" = "0",
                                      "Diabetic" = "1")) %>%
  mutate(GenHlth = factor(GenHlth)) %>%
  mutate(GenHlth = fct_recode(GenHlth, "Excellent" = "1", "Fair" = "4",
                              "Good" = "3", "Poor" = "5", " Very Good" = "2")) %>%
  ggplot() +
  geom_bar(aes(x = Diabetes_binary, fill = GenHlth)) +
  ggtitle("US Diabities Dataset: Distribution of Participants' General Health Scores") +
  xlab("General Health Scores") +
  ylab("Count")
```



```
us_train %>%
  mutate(Diabetes_binary = factor(Diabetes_binary)) %>%
  mutate(Diabetes_binary = fct_recode(Diabetes_binary, "Not Diabetic" = "0",
                                      "Diabetic" = "1")) %>%
  mutate(MentHlth = factor(MentHlth)) %>%
  ggplot() +
  geom_bar(aes(x = Diabetes_binary, fill = MentHlth)) +
  ggtitle("US Diabities Dataset: How many days did your mental health suffer in the last 30 days?") +
  xlab("Days") +
  ylab("Count")
```



```
us_train %>%
  mutate(Diabetes_binary = factor(Diabetes_binary)) %>%
  mutate(Diabetes_binary = fct_recode(Diabetes_binary, "Not Diabetic" = "0",
                                      "Diabetic" = "1")) %>%
  mutate(PhysActivity = factor(PhysActivity)) %>%
  mutate(PhysActivity = fct_recode(PhysActivity, "No" = "0", "Yes" = "1")) %>%
  ggplot() +
  geom_bar(aes(x = Diabetes_binary, fill = PhysActivity)) +
  ggtitle("US Diabities Dataset: Did you exercise in the last 30 days?") +
  xlab("Responses") +
  ylab("Count")
```



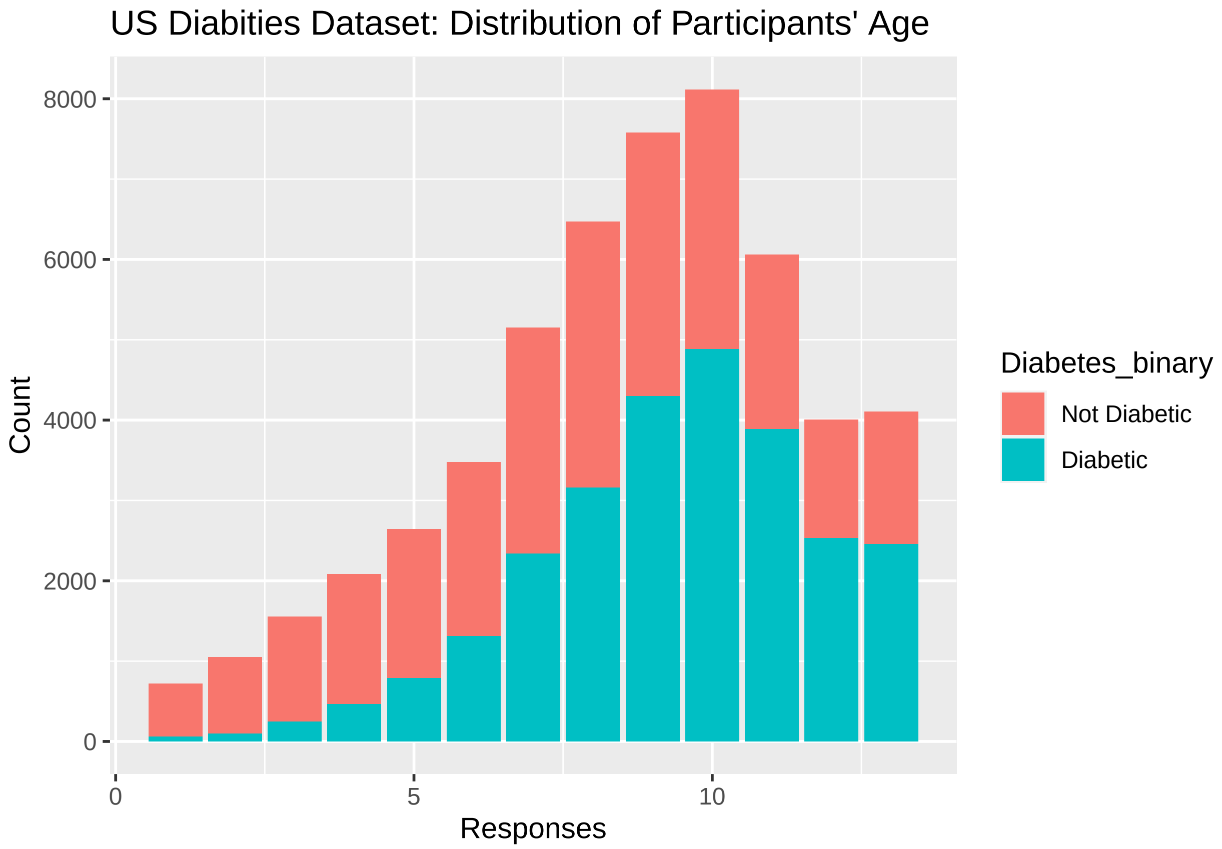
We see the most variation in the general health scores; non diabetics report more “Very Good” and “Excellent” scores than diabetics (and similarly, fewer “Poor” and “Fair”) scores. The mental health data was roughly the same, as was the physical activity data (here, non diabetics indicated exercising a little bit more than diabetics, but nothing very exaggerated). Thus, I opted to incorporate the general health feature.

```
us_diabetics_mod_4 <- lm(Diabetes_binary ~  HighBP + BMI + Income + GenHlth, data = us_diabetics)

us_train %>%
  add_predictions(us_diabetics_mod_4)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Diabetes_binary == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

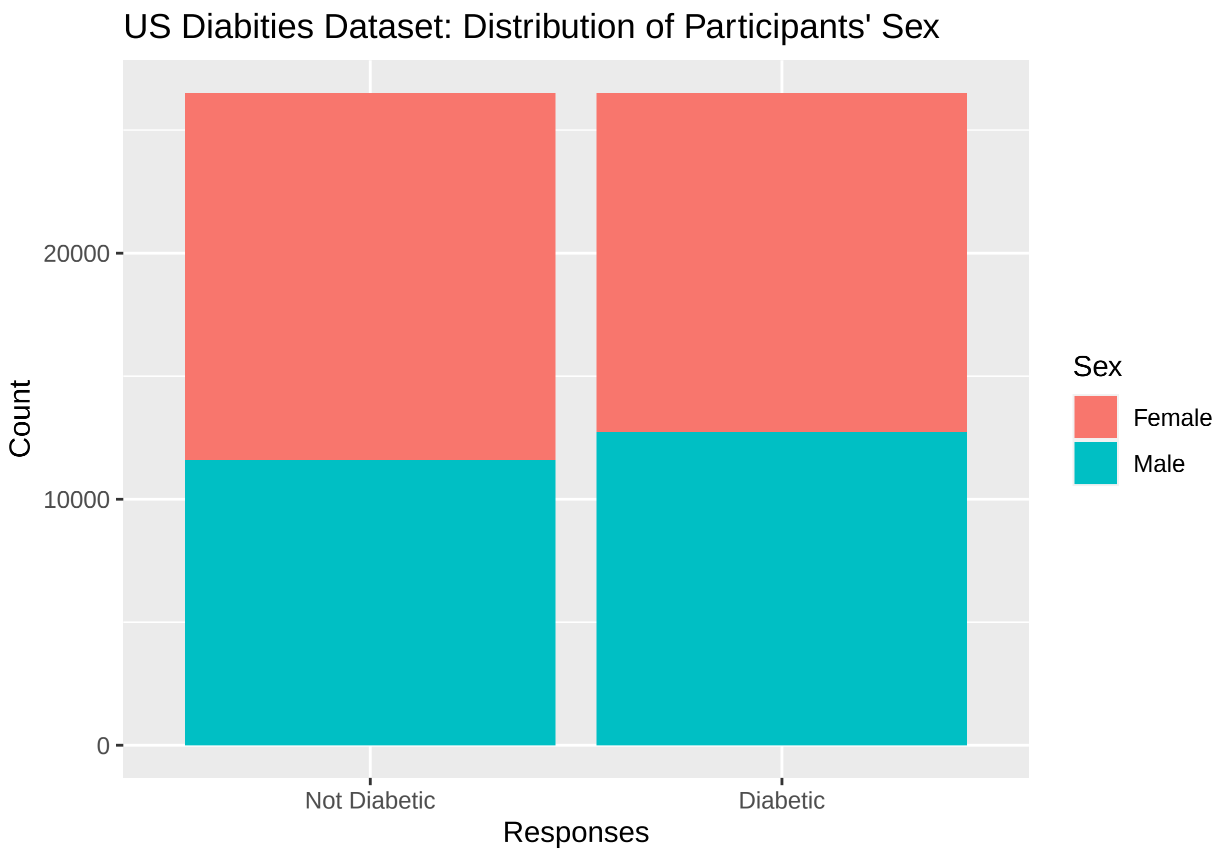
```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.726
```

This accuracy is not bad given the model with all the features had roughly a 74% accuracy. Let’s consider the distribution of gender and sex and see if adding both/one of these makes sense or increases the accuracy model by a considerable amount.

```
us_train %>%
  mutate(Diabetes_binary = factor(Diabetes_binary)) %>%
  mutate(Diabetes_binary = fct_recode(Diabetes_binary, "Not Diabetic" = "0",
                                      "Diabetic" = "1")) %>%
  ggplot() +
  geom_bar(aes(x = Age, fill = Diabetes_binary)) +
  ggtitle("US Diabities Dataset: Distribution of Participants' Age") +
  xlab("Responses") +
  ylab("Count")
```



```
us_train %>%
  mutate(Diabetes_binary = factor(Diabetes_binary)) %>%
  mutate(Diabetes_binary = fct_recode(Diabetes_binary, "Not Diabetic" = "0",
                                      "Diabetic" = "1")) %>%
  mutate(Sex = factor(Sex)) %>%
  mutate(Sex = fct_recode(Sex, "Female" = "0",
                                      "Male" = "1")) %>%
  ggplot() +
  geom_bar(aes(x = Diabetes_binary, fill = Sex)) +
  ggtitle("US Diabities Dataset: Distribution of Participants' Sex") +
  xlab("Responses") +
  ylab("Count")
```



```
us_train
```

```
## # A tibble: 53,020 x 22
##    Diabetes_binary HighBP HighChol CholCheck   BMI Smoker Stroke
##              <int>  <int>    <int>     <int> <int>  <int>  <int>
##  1               0      1        0         1    26      0      0
##  2               0      1        1         1    26      1      1
##  3               0      0        0         1    26      0      0
##  4               0      0        0         1    29      1      0
##  5               0      0        0         1    18      0      0
##  6               0      0        1         1    26      1      0
##  7               0      0        0         1    31      1      0
##  8               0      0        0         1    32      0      0
##  9               0      0        0         1    27      1      0
## 10               0      0        0         1    21      0      0
## # ... with 53,010 more rows, and 15 more variables: HeartDiseaseorAttack <int>,
## #   PhysActivity <int>, Fruits <int>, Veggies <int>, HvyAlcoholConsump <int>,
## #   AnyHealthcare <int>, NoDocbcCost <int>, GenHlth <int>, MentHlth <int>,
## #   PhysHlth <int>, DiffWalk <int>, Sex <int>, Age <int>, Education <int>,
## #   Income <int>
```

There seems to be a trend in which as age increases the proportion of diabetics increased; intuitively, this makes sense. Similarly, participants' sex does not seem to affect their likelihood of getting diabetes either. Let's add age to the model and see if it increases the model efficacy by a reasonable amount.

```
us_diabetics_mod_5 <- lm(Diabetes_binary ~  HighBP + BMI + Income + GenHlth + Age, data = us_diabetics)

us_train %>%
  add_predictions(us_diabetics_mod_5)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Diabetes_binary == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.739
```

Adding age increases our efficacy to roughly 74% which almost matches the accuracy of our model accuracy when we have all features. Similarly, the majority of these features are accessible to most people and are (generally) not subjective measurement. Let's test our 5th model on the testing data and see how it performs.

```
us_test %>%
  add_predictions(us_diabetics_mod_5)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Diabetes_binary == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.733
```

This model seems good – the accuracy is 73.3% which is pretty good considering the test model with all the features was 74.7%. It is a little lower, but this is to be expected for a model that is much simpler. I think that the decrease in accuracy is worth the accessibility and simplicity of the model.

Next, let's explore a model for the Pima Indian dataset. First, we split our data into training and testing:

```
set.seed(3476)
trainIndex <- createDataPartition(pima_diabetics$Outcome, p = .75,
                                  list = FALSE,
                                  times = 1)
pima_train <- pima_diabetics[ trainIndex,]
pima_test <- pima_diabetics[-trainIndex,]
```

Now, we have our training and test datasets.Let's create our preliminary model first.

```
pima_train %>% group_by(Outcome) %>% summarize(count = n())
```

```
## # A tibble: 2 x 2
##   Outcome count
##     <int> <int>
## 1       0   382
## 2       1   194
```

Thus, our baseline model should predict that nobody has diabetes; this gives us an accuracy of 66%.

Once again, our first model will use all the features to give us a target value.

```
names(pima_train)
```

```
## [1] "Pregnancies"              "Glucose"
## [3] "BloodPressure"            "SkinThickness"
## [5] "Insulin"                  "BMI"
## [7] "DiabetesPedigreeFunction" "Age"
## [9] "Outcome"
```

```
pima_diabetics_mod_ALL <- lm(Outcome ~  Pregnancies + Glucose + BloodPressure + SkinThickness + Insulin
                               + BMI+
                                 Age,
                             data = pima_diabetics)

pima_train %>%
  add_predictions(pima_diabetics_mod_ALL)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Outcome == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.795
```

The efficacy with all the features is 79.5%.

On the testing data, the efficacy is 75%. There is a little bit of discrepancy in the test and training data, but I don't think this is due to overfitting but instead a lack of data; pima_test only has 192 rows. Making the testing data larger makes the training data smaller, and through testing I discovered that the percentages between the training and testing data have discrepancies but stay between 70% to 80% regardless of how I split them up. Thus, this is just

```
pima_test
```

```
## # A tibble: 192 x 9
##    Pregnancies Glucose BloodPressure SkinThickness Insulin   BMI
##          <int>   <dbl>         <int>         <int>   <int> <dbl>
##  1           1      85            66            29       0  26.6
##  2           1      89            66            23      94  28.1
##  3           7     100             0             0       0  30
##  4           1     103            30            38      83  43.3
##  5           1     115            70            30      96  34.6
##  6           3     126            88            41     235  39.3
##  7           7     196            90             0       0  39.8
##  8          10     122            78            31       0  27.6
##  9          11     138            76             0       0  33.2
## 10           9     102            76            37       0  32.9
## # ... with 182 more rows, and 3 more variables: DiabetesPedigreeFunction <dbl>,
## #   Age <int>, Outcome <int>
```

```
pima_test %>%
  add_predictions(pima_diabetics_mod_ALL)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Outcome == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1  0.75
```

From out previous exploration, we saw that BloodPressure, BMI, and Age were important factors. Let's see the efficacy of our model with just these features.

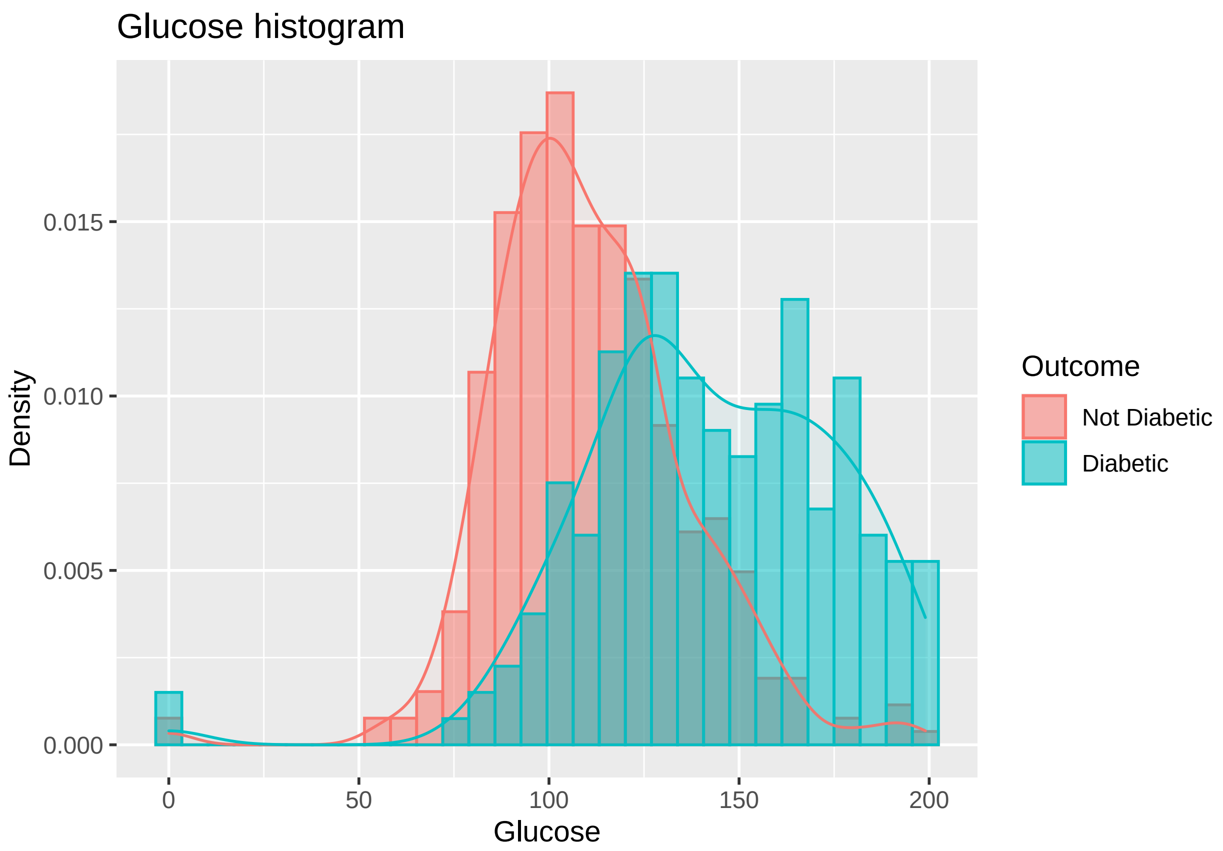
```
pima_diabetics_mod_2 <- lm(Outcome ~ BloodPressure +
                                BMI+
                                Age,
                             data = pima_diabetics)

pima_train %>%
  add_predictions(pima_diabetics_mod_2)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Outcome == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.684
```

This gets us to an efficacy of 68.4%! Intuitively, glucose seems as though it would be an important indicator of diabetes. Let's graph the distribution of glucose.

```
pima_train2 <- pima_train %>%
  mutate(Outcome = factor(Outcome)) %>%
  mutate(Outcome = fct_recode(Outcome, "Not Diabetic" = "0",
                                      "Diabetic" = "1"))

ggplot(pima_train2, aes(x=Glucose, color=Outcome, fill=Outcome)) +
geom_histogram(aes(y=..density..), position="identity", alpha=0.5)+
geom_density(alpha=0.05)+
labs(title="Glucose histogram",x="Glucose", y = "Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Here, we see that the distributions of glucose for both non diabetics and diabetics are roughly normally distributed, they have different means; non diabetics have a mean that is smaller than diabetics. Given the nature of diabetes, this makes sense. Let's see how much our model's accuracy increases if we add glucose to our other features, and let's also see what the accuracy is if we just use glucose to make our predictions.

```
pima_diabetics_mod_3 <- lm(Outcome ~ BloodPressure +
                                BMI+
                                Age +
                             Glucose,
                             data = pima_diabetics)


pima_diabetics_mod_4 <- lm(Outcome ~  Glucose,
                             data = pima_diabetics)

pima_train %>%
  add_predictions(pima_diabetics_mod_3)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Outcome == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.785
```

```r
pima_train %>%
  add_predictions(pima_diabetics_mod_4)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Outcome == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.759
```

We see that the model with just glucose levels gives us a 75.8% chance of predicting diabetes, and adding Glucose to our previous feature increases the efficacy to roughly 78.4%. Let's see if we can simplify our model more by getting rid of potential features that are not contributing a lot.

```r
pima_diabetics_mod_5 <- lm(Outcome ~ BloodPressure +
                            Glucose,
                             data = pima_diabetics)


pima_diabetics_mod_6 <- lm(Outcome ~  BMI+
                             Glucose,
                             data = pima_diabetics)


pima_train %>%
  add_predictions(pima_diabetics_mod_5)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Outcome == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.757
```

```r
pima_train %>%
  add_predictions(pima_diabetics_mod_6)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Outcome == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.783
```

Here, we see that a model with just BMI and Glucose gets us to an efficacy of 78.2%. I think this is a good modeling terms of accuracy and simplicity. Let's test it on the testing data now.

```
pima_test %>%
  add_predictions(pima_diabetics_mod_3)  %>%
  mutate(pred = ifelse(pred > 0.5, 1, 0)) %>%
  mutate(correct = ifelse(Outcome == pred, 1, 0)) %>%
  summarize(score = mean(correct, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   score
##   <dbl>
## 1 0.734
```

Considering the efficacy of our testing model was originally 75%, this model with only 2 predictors is pretty solid. Once again, the discrepancy between the testing and training efficacies is notable, but I think this is due to a lack of data more than a brittle model.

Conclusion

This analysis into diabetes dataset resulted in some interesting (but still intuitive) outcomes.

Our first question discussed the socioeconomic factors that could potentially contribute to diabetes. This avenue was considered as I wanted to remind the readers that healthcare and health problems tends to be a function of inequality not just general health. Here, we found that income (and similarly, education) were important variables and helped explain some of the variation seen in diabetes. The importance of income was also later explored and utilized in the third question where we constructed a more fleshed-out model to predict diabetes. This analysis was only done on the US American dataset.

In the second question we considered the relationship between BMI and Diabetes, as well as the relationship between BMI and high levels of cholesterol as well as BMI and high blood pressure to see if there were possible relations between them.The analyses showed us that diabetics have a higher BMI on average than non-diabetics. We found using bar charts and T-tests that there were statistically significant differences in the BMI's for non diabetics and diabetics based on if they had high blood pressure/high cholesterol. This makes intutive sense and research articles from the internet (linked below in the appendix) backed up this claim. The relationships between these models were important in deciding what features were important to graph in the third section, where we made a model. We also analyzed the BMI distribution of the Pima Indians using a semi-join such that the range of the BMI's for both US and Pima Indians was the same in order to allow us to view their differences more clearly. We found that for both populations the BMI was roughly normally distributed and that the means and general distributions mirrored each other.

In the final question, the aim was to build models for each population to predict diabetes. In our earlier explorations we discovered for the US population that income, blood pressure, BMI, and cholesterol were important features. Through trial and error we landed on a model that had five features (HighBP + BMI + Income + GenHlth + Age) and had an efficacy of roughly 73-74% on the training and testing data. For the Pima Indians, we performed a similar analysis and produced a model that had two features (BMI and Glucose) and had an efficacy between 73-78% for the testing and training models. The reason for this discrepancy between the two models is likely due to the small size of the data set. Similarly, glucose, understandably, feels like the most important feature in diagnosing diabetes, which is why the Pima Indians model has considerably fewer features with a higher level of accuracy. The process in which this was accomplished was not statistically backed-up and was based on intuition. Thus, there is probably room for improvement in these models, but given the upper bounds we determined in the analysis I think they are pretty adequate.

Please note that the results of this analysis was brought on by US American data in the 2015, and that it was a voluntary survey. Similarly, the results of this dataset that utilized the Pima Indians dataset was pruned from a larger dataset, and there is no information provided on how the data was collected (and thus I assume voluntary as well). Thus,we cannot make any notions of causation only correlation from this analysis. We can treat this as an exploration that can provide the foundation for possible interesting variables to explore/model in a more controlled laboratory setting. I think it would be interesting to play with pre-diabetes data and use models to predict the likelihood of developing diabetes based on a patient's the health data. If we are able to create a robust enough model, maybe we can figure out what features are important and fix those in an effort to prevent diabetes. All the literature I read stated that the best way to control diabetes was to go to the doctor as early as possible and that preventative care was the best recommended path. I think that considering the scale of this disease, more R&D into this field is important.

Appendix

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset https://www.kaggle.com/datasets/mathchi/diabetes-data-set https://www.who.int/news-room/fact-sheets/detail/diabetes https://www.diabetes.org/newsroom/press-releases/2021/income-related-inequalities-in-diabetes-have-widened-over-past-decade-cdc-study-finds https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-high-blood-pressure https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6316192/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4457375/ https://www.cdc.gov/diabetes/library/spotlights/diabetes-facts-stats.html#:~:text=37.3%20million%20Americans%E2%80%94about%201,t%20know%20they%20h
https://stackoverflow.com/questions/50782005/r-geom-bar-facet-grid-vertical-instead-of-horizontal https://www.programmingr.com/examples/r-dataframe/merge-data-frames/#:~:text=Using%20rbind()%20to%20merge%20two%