Project

Group 1

2023-05-04

Let's load the data sets in to the file

Diabetes Data sets

```
diabetes_data <- read.csv("C:/Users/SRI HARSHA S/OneDrive - Indiana University/Documents/R/Project Files
head(diabetes_data)
```

```
##   Age Sex HighChol CholCheck BMI Smoker HeartDiseaseorAttack PhysActivity
## 1   4   1        0         1  26      0                    0            1
## 2  12   1        1         1  26      1                    0            0
## 3  13   1        0         1  26      0                    0            1
## 4  11   1        1         1  28      1                    0            1
## 5   8   0        0         1  29      1                    0            1
## 6   1   0        0         1  18      0                    0            1
##   Fruits Veggies HvyAlcoholConsump GenHlth MentHlth PhysHlth DiffWalk Stroke
## 1      0       1                 0       3        5       30        0      0
## 2      1       0                 0       3        0        0        0      1
## 3      1       1                 0       1        0       10        0      0
## 4      1       1                 0       3        0        3        0      0
## 5      1       1                 0       2        0        0        0      0
## 6      1       1                 0       2        7        0        0      0
##   HighBP Diabetes
## 1      1        0
## 2      1        0
## 3      0        0
## 4      1        0
## 5      0        0
## 6      0        0
```

Description/ Summary of the data:

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.2.3
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
describe(diabetes_data)
```

```
## diabetes_data
##
##  18  Variables      70692  Observations
## ---------------------------------------------------------------------------
## Age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    70692        0       13    0.988    8.584    3.194        3        4
##      .25      .50      .75      .90      .95
##        7        9       11       12       13
##
## lowest :  1  2  3  4  5, highest:  9 10 11 12 13
##
## Value          1     2     3     4     5     6     7     8     9    10    11
## Frequency    979  1396  2049  2793  3520  4648  6872  8603 10112 10856  8044
## Proportion 0.014 0.020 0.029 0.040 0.050 0.066 0.097 0.122 0.143 0.154 0.114
##
## Value         12    13
## Frequency   5394  5426
## Proportion 0.076 0.077
## ---------------------------------------------------------------------------
## Sex
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2    0.744    32306    0.457   0.4963
##
## ---------------------------------------------------------------------------
## HighChol
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2    0.748    37163   0.5257   0.4987
##
## ---------------------------------------------------------------------------
## CholCheck
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2    0.072    68943   0.9753  0.04826
##
## ---------------------------------------------------------------------------
## BMI
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    70692        0       80    0.997    29.86    7.425       21       22
##      .25      .50      .75      .90      .95
##       25       29       33       39       43
##
## lowest : 12 13 14 15 16, highest: 87 89 92 95 98
## ---------------------------------------------------------------------------
## Smoker
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2    0.748    33598   0.4753   0.4988
##
## ---------------------------------------------------------------------------
```

```
## HeartDiseaseorAttack
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2    0.378    10449   0.1478   0.2519
##
## ------------------------------------------------------------------------------
## PhysActivity
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2    0.626    49699    0.703   0.4176
##
## ------------------------------------------------------------------------------
## Fruits
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2    0.713    43249   0.6118    0.475
##
## ------------------------------------------------------------------------------
## Veggies
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2     0.5    55760   0.7888   0.3332
##
## ------------------------------------------------------------------------------
## HvyAlcoholConsump
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2    0.123     3020  0.04272  0.08179
##
## ------------------------------------------------------------------------------
## GenHlth
##        n  missing distinct     Info     Mean      Gmd
##    70692        0        5    0.933    2.837    1.232
##
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
##
## Value          1     2     3     4     5
## Frequency   8282 19872 23427 13303  5808
## Proportion 0.117 0.281 0.331 0.188 0.082
## ------------------------------------------------------------------------------
## MentHlth
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    70692        0       31    0.685    3.752    6.285        0        0
##      .25      .50      .75      .90      .95
##        0        0        2       15       30
##
## lowest :  0  1  2  3  4, highest: 26 27 28 29 30
## ------------------------------------------------------------------------------
## PhysHlth
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    70692        0       31    0.818     5.81    8.949        0        0
##      .25      .50      .75      .90      .95
##        0        0        6       30       30
##
## lowest :  0  1  2  3  4, highest: 26 27 28 29 30
## ------------------------------------------------------------------------------
## DiffWalk
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2    0.567    17866   0.2527   0.3777
```

```
##
## ------------------------------------------------------------------------------
## Stroke
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2    0.175     4395  0.06217   0.1166
##
## ------------------------------------------------------------------------------
## HighBP
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2    0.738    39832   0.5635    0.492
##
## ------------------------------------------------------------------------------
## Diabetes
##        n  missing distinct     Info      Sum     Mean      Gmd
##    70692        0        2     0.75    35346      0.5      0.5
##
## ------------------------------------------------------------------------------
```

We can see here that there are no missing values in the data set and we can see that many of them are categorical.

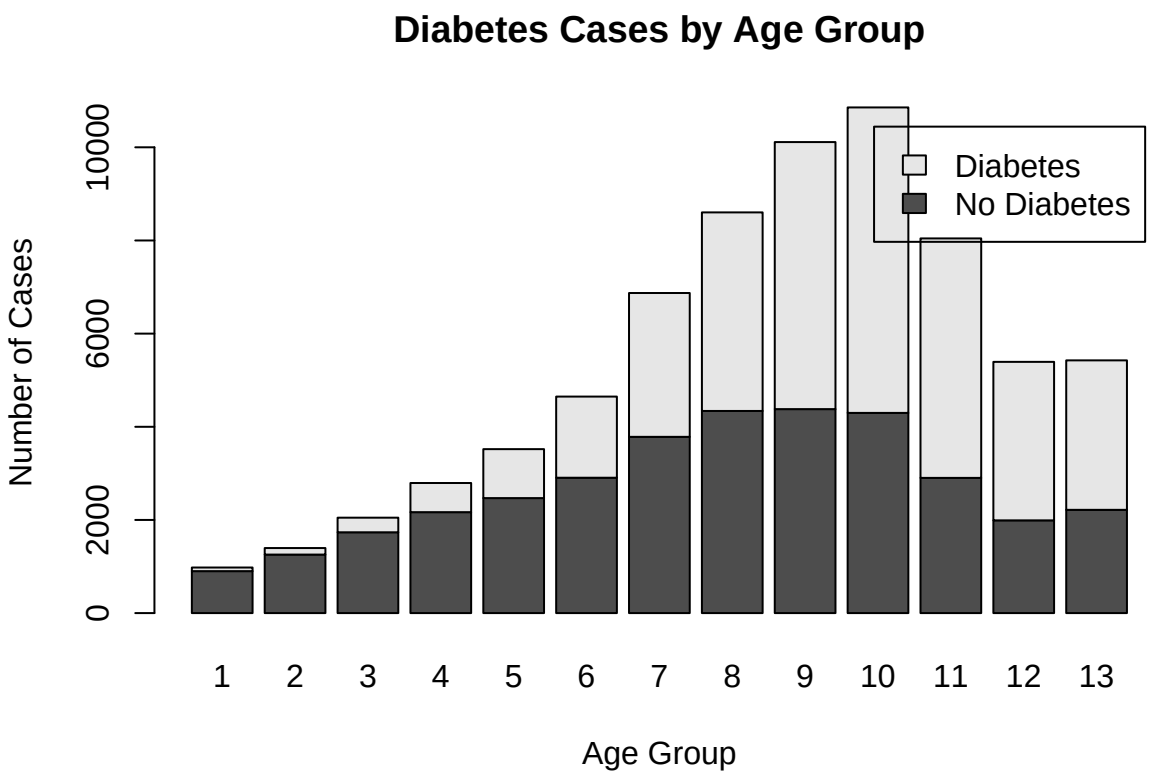
Lets see whether we have enough data of the samples who are diabetic

```
sum(diabetes_data$Diabetes)/nrow(diabetes_data)
```

```
## [1] 0.5
```

Exploratory data analysis:

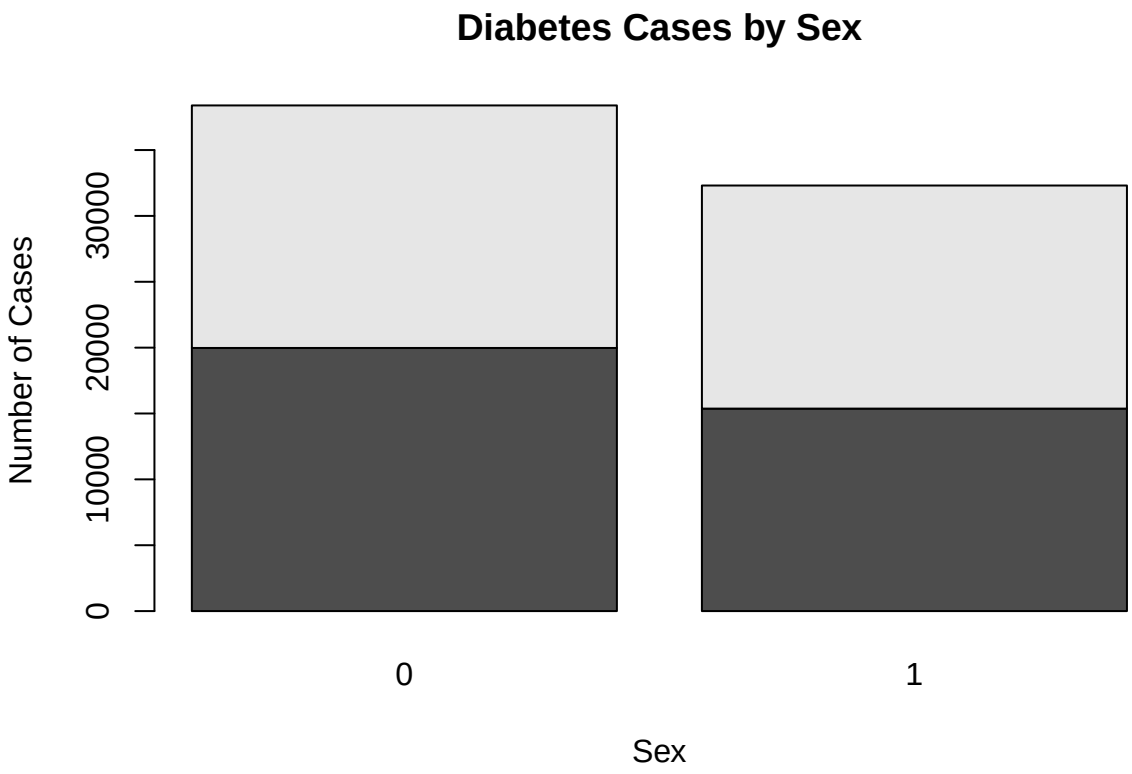
```
# create a table with counts of diabetes cases by age group
age_table <- table(diabetes_data$Diabetes, diabetes_data$Age)

# create a bar chart
barplot(age_table, main = "Diabetes Cases by Age Group", xlab = "Age Group", ylab = "Number of Cases",
```



From the chart, we can see that the data for both diabetes and non-diabetes cases across different age groups. As the age group of the patient increases there is increasing trend of diabetic cases.

Sex vs diabetes:

```
# create a table with counts of diabetes cases by age group
Sex_table <- table(diabetes_data$Diabetes, diabetes_data$Sex)

# create a bar chart
barplot(Sex_table, main = "Diabetes Cases by Sex", xlab = "Sex", ylab = "Number of Cases")
```



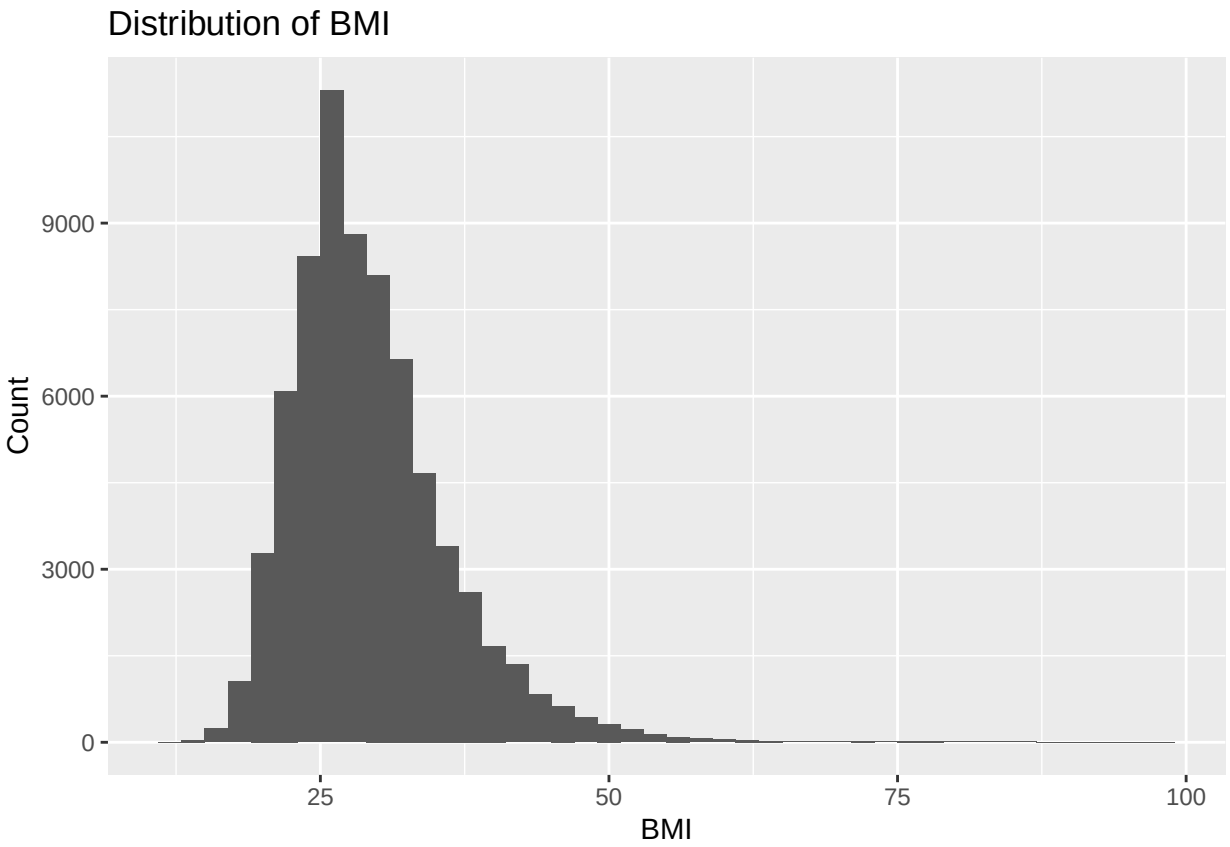
The bar plot shows that the number of diabetes cases is higher in females than in males.

Distribution of BMI:

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

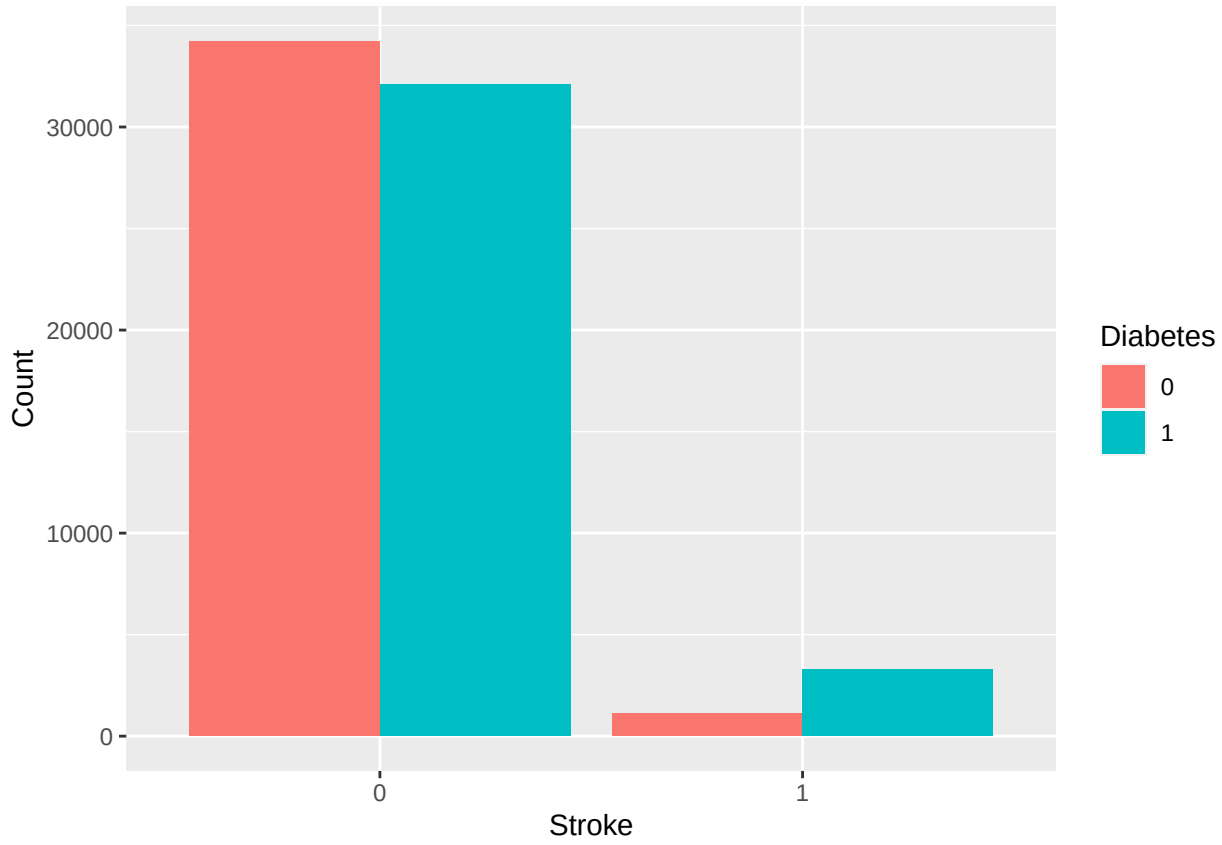
```
ggplot(data = diabetes_data, aes(x = BMI)) + geom_histogram(binwidth = 2) + xlab("BMI") + ylab("Count")
```



The histogram shows a peak at a BMI value of around 25, we can infer that this value is the most common BMI value in the population being studied and it is right skewed.

Stroke vs Diabetes:

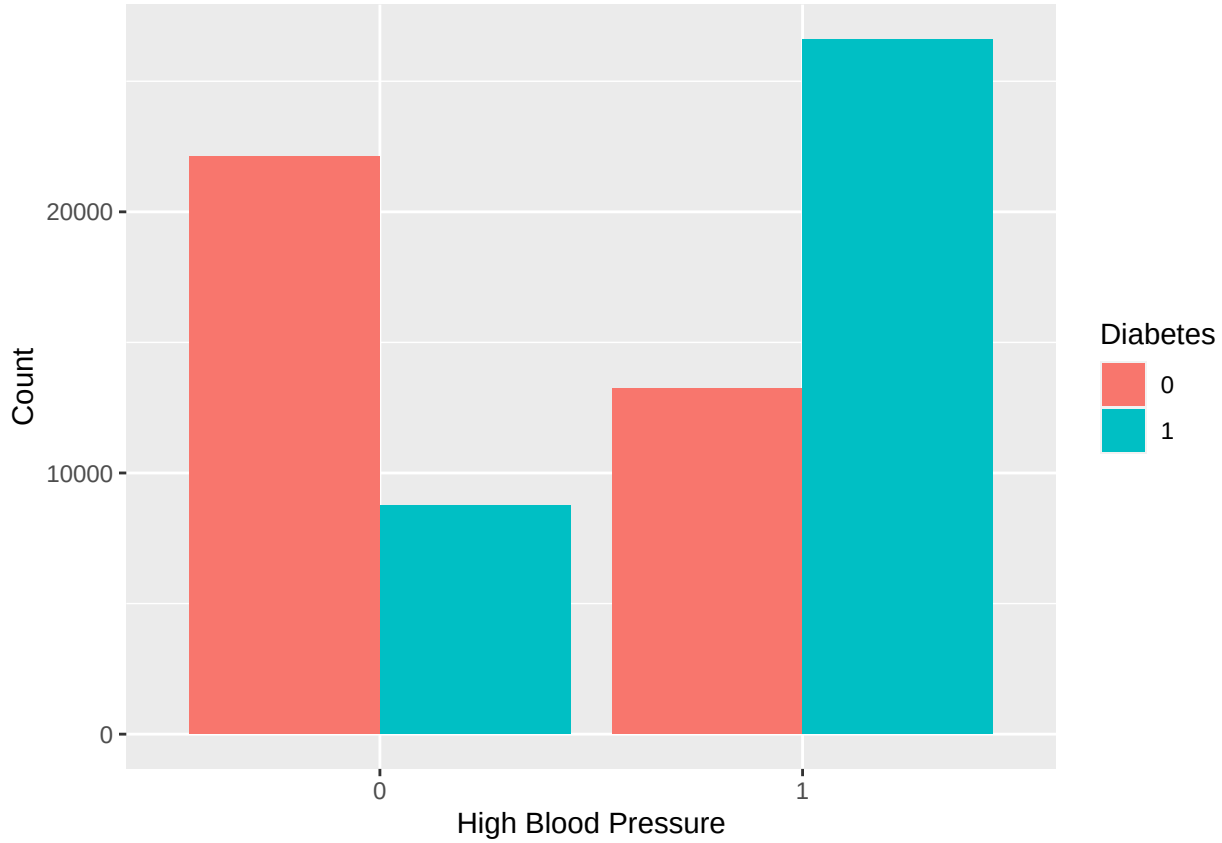
```
diabetes_data$Stroke <- as.factor(diabetes_data$Stroke)
diabetes_data$Diabetes <- as.factor(diabetes_data$Diabetes)
ggplot(data = diabetes_data, aes(x = Stroke, fill = Diabetes)) + geom_bar(position = "dodge") + xlab("S
```



From the bar plot we can say that the number of individuals with diabetes who have had a stroke is higher than the number of individuals without diabetes who have had a stroke.

HighBp vs Diabetes:

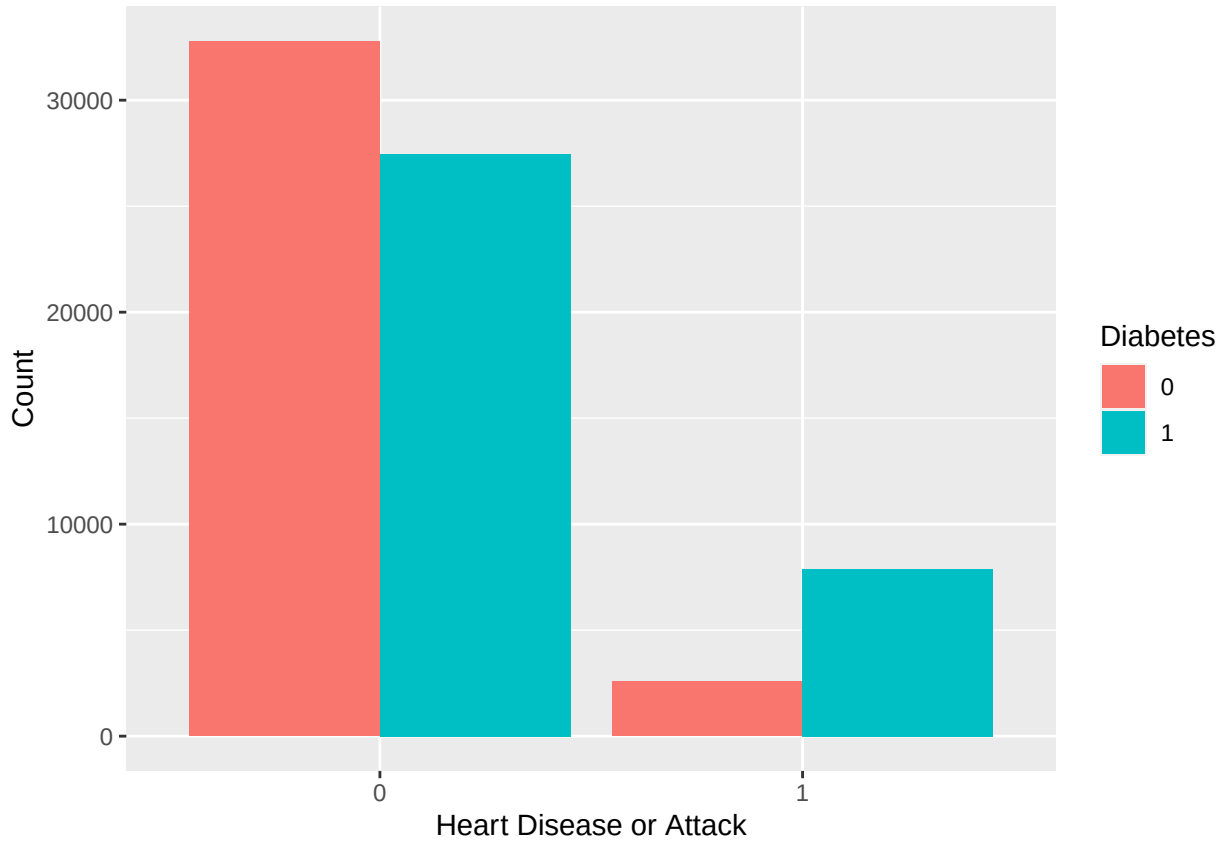
```
diabetes_data$HighBP <- as.factor(diabetes_data$HighBP)
diabetes_data$Diabetes <- as.factor(diabetes_data$Diabetes)
ggplot(data = diabetes_data, aes(x = HighBP, fill = Diabetes)) + geom_bar(position = "dodge") + xlab("H
```



The plot shows that the number of individuals with diabetes who have high blood pressure is higher than the number of individuals without diabetes who have high blood pressure.

HeartAttack vs Diabetes:

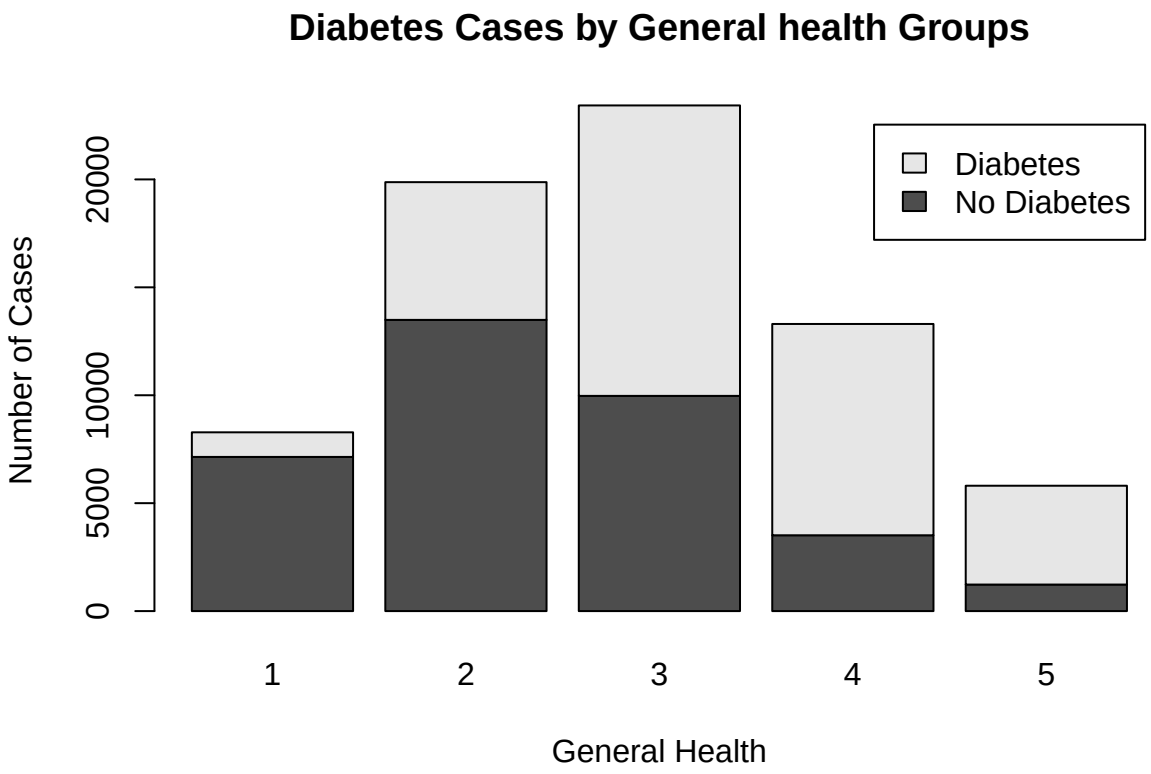
```
diabetes_data$HeartDiseaseorAttack <- as.factor(diabetes_data$HeartDiseaseorAttack)
diabetes_data$Diabetes <- as.factor(diabetes_data$Diabetes)
ggplot(data = diabetes_data, aes(x = HeartDiseaseorAttack, fill = Diabetes)) + geom_bar(position = "dod
```



The chart shows that the number of individuals with diabetes who have heart disease or have had a heart attack is higher than the number of individuals without diabetes who have heart disease or have had a heart attack.

General health vs Diabetes:

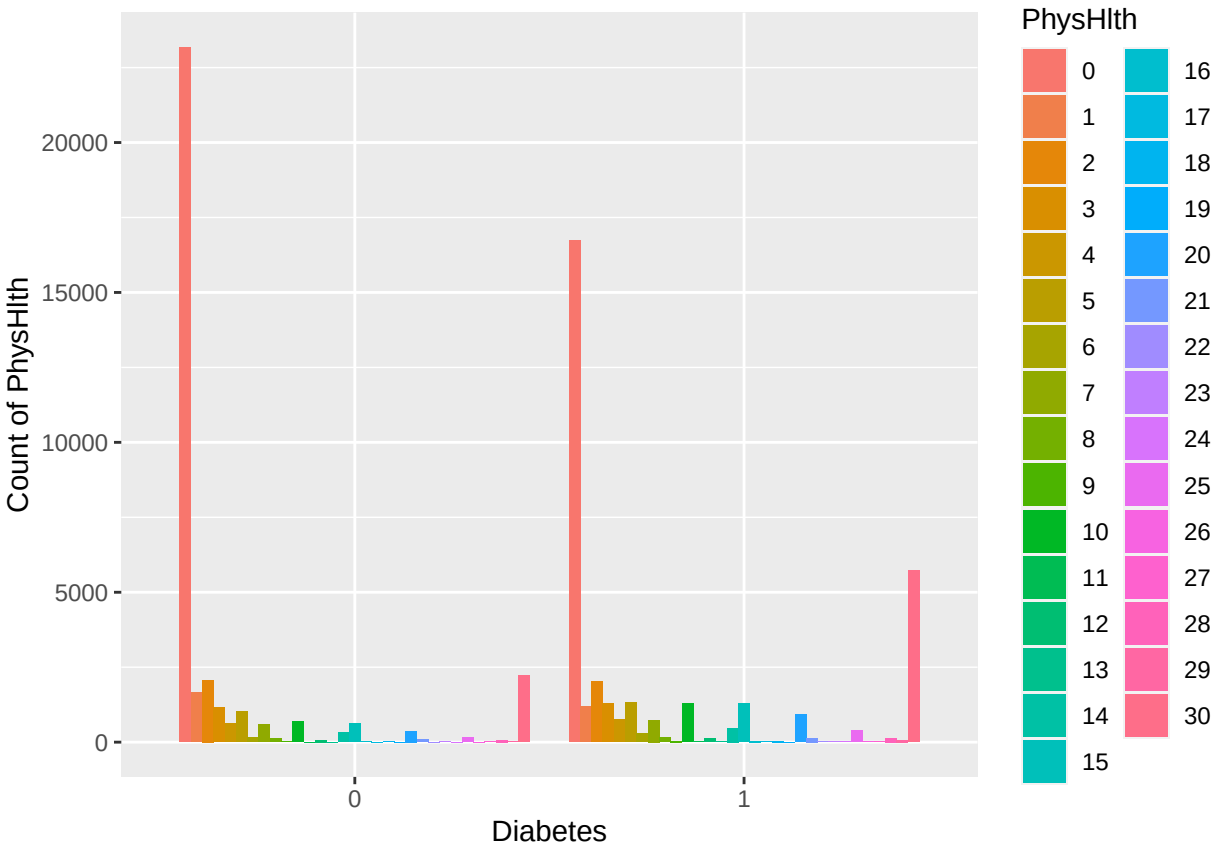
```
# create a table with counts of diabetes cases by General health group
GenHlth_table <- table(diabetes_data$Diabetes, diabetes_data$GenHlth)

# create a bar chart
barplot(GenHlth_table, main = "Diabetes Cases by General health Groups", xlab = "General Health", ylab =
```



Here, as the scale increases, the diabetic cases are increasing, and non-diabetic cases are decreasing.

Physical Health Vs Diabetes:

```
ggplot(diabetes_data, aes(x = Diabetes, fill = factor(PhysHlth))) + geom_bar(position = "dodge") + xlab
```



This bar plot represents the number of individuals with diabetes who have had physical illness/injury in the past 30 days is higher than the individuals without diabetes.

Statistical analysis:

1 Research Question:

Which health and demographic factors are most strongly associated to the risk of developing diabetes?

Null Hypothesis: There is no significant association between demographic and health factors and the risk of developing diabetes.

Alternate Hypothesis: There is a significant association between demographic and health factors and the risk of developing diabetes.

```
table(diabetes_data$Age, diabetes_data$Diabetes)
```

```
##
##          0    1
##    1   901   78
##    2  1256  140
##    3  1735  314
##    4  2167  626
##    5  2469 1051
##    6  2906 1742
##    7  3784 3088
```

```
##   8  4340 4263
##   9  4379 5733
##   10 4298 6558
##   11 2903 5141
##   12 1991 3403
##   13 2217 3209
```

```
chisq.test(diabetes_data$Age, diabetes_data$Diabetes)
```

```
##
##  Pearson's Chi-squared test
##
## data:  diabetes_data$Age and diabetes_data$Diabetes
## X-squared = 6179.1, df = 12, p-value < 2.2e-16
```

```
table(diabetes_data$Sex, diabetes_data$Diabetes)
```

```
##
##         0     1
##   0 19975 18411
##   1 15371 16935
```

```
chisq.test(diabetes_data$Sex, diabetes_data$Diabetes)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  diabetes_data$Sex and diabetes_data$Diabetes
## X-squared = 139.26, df = 1, p-value < 2.2e-16
```

```
table(diabetes_data$HighChol, diabetes_data$Diabetes)
```

```
##
##         0     1
##   0 21869 11660
##   1 13477 23686
```

```
chisq.test(diabetes_data$HighChol, diabetes_data$Diabetes)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  diabetes_data$HighChol and diabetes_data$Diabetes
## X-squared = 5911.8, df = 1, p-value < 2.2e-16
```

```
table(diabetes_data$CholCheck, diabetes_data$Diabetes)
```

```
##
##         0     1
##   0  1508   241
##   1 33838 35105
```

```
chisq.test(diabetes_data$CholCheck, diabetes_data$Diabetes)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  diabetes_data$CholCheck and diabetes_data$Diabetes
## X-squared = 939.63, df = 1, p-value < 2.2e-16
```

```
table(diabetes_data$Smoker, diabetes_data$Diabetes)
```

```
##
##         0     1
##   0 20065 17029
##   1 15281 18317
```

```
chisq.test(diabetes_data$Smoker, diabetes_data$Diabetes)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  diabetes_data$Smoker and diabetes_data$Diabetes
## X-squared = 522.48, df = 1, p-value < 2.2e-16
```

```
table(diabetes_data$PhysActivity, diabetes_data$Diabetes)
```

```
##
##         0     1
##   0  7934 13059
##   1 27412 22287
```

```
chisq.test(diabetes_data$PhysActivity, diabetes_data$Diabetes)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  diabetes_data$PhysActivity and diabetes_data$Diabetes
## X-squared = 1779, df = 1, p-value < 2.2e-16
```

```
table(diabetes_data$HvyAlcoholConsump, diabetes_data$Diabetes)
```

```
##
##         0     1
##   0 33158 34514
##   1  2188   832
```

```
chisq.test(diabetes_data$HvyAlcoholConsump, diabetes_data$Diabetes)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  diabetes_data$HvyAlcoholConsump and diabetes_data$Diabetes
## X-squared = 635.09, df = 1, p-value < 2.2e-16
```

```
table(diabetes_data$GenHlth, diabetes_data$Diabetes)
```

```
##
##         0     1
##   1  7142  1140
##   2 13491  6381
##   3  9970 13457
##   4  3513  9790
##   5  1230  4578
```

```
chisq.test(diabetes_data$GenHlth, diabetes_data$Diabetes)
```

```
##
##  Pearson's Chi-squared test
##
## data:  diabetes_data$GenHlth and diabetes_data$Diabetes
## X-squared = 12304, df = 4, p-value < 2.2e-16
```

```
#Welch-T test for the numerical data
t.test(diabetes_data$MentHlth ~ diabetes_data$Diabetes)
```

```
##
##  Welch Two Sample t-test
##
## data:  diabetes_data$MentHlth by diabetes_data$Diabetes
## t = -23.227, df = 67626, 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:
##  -1.539325 -1.299751
## sample estimates:
## mean in group 0 mean in group 1
##        3.042268        4.461806
```

```
t.test(diabetes_data$PhysHlth ~ diabetes_data$Diabetes)
```

```
##
##  Welch Two Sample t-test
##
## data:  diabetes_data$PhysHlth by diabetes_data$Diabetes
## t = -57.985, df = 64069, 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:
##  -4.433070 -4.143176
## sample estimates:
## mean in group 0 mean in group 1
##        3.666355        7.954479
```

```
t.test(diabetes_data$BMI ~ diabetes_data$Diabetes)
```

```
##
##  Welch Two Sample t-test
##
## data:  diabetes_data$BMI by diabetes_data$Diabetes
## t = -81.591, df = 68653, 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
```

Interpretation:

The results of the Chi-squared tests indicate a significant association between all the predictor variables (Age, Sex, HighChol, CholCheck, Smoker, PhysActivity, HvyAlcoholConsump, and GenHlth) and the response variable (Diabetes), with p-values less than 0.05. Hence we can reject the null hypothesis. This suggests that each of these predictor variables is associated with the risk of developing diabetes.

The Welch Two Sample t-test also indicates a significant difference in the mean values of the Mental Health variable between those with diabetes and those without diabetes, with a p-value less than 0.05 and can reject null hypothesis. This suggests that there is a significant difference in mental health between those with diabetes and those without diabetes.

The Welch Two Sample t-test also indicates a significant difference in the mean values of the Physical Health variable between those with diabetes and those without diabetes, with a p-value less than 0.05 and can reject null hypothesis. This suggests that there is a significant difference in mental health between those with diabetes and those without diabetes.

Lets do the logistic regression to know the association.

```
# assuming your data frame is named "data" with columns "age", "sex", and "diabetes"
logit_model <- glm(Diabetes ~ Age + Sex + HighChol + CholCheck + Smoker + PhysActivity +
                   HvyAlcoholConsump + GenHlth + MentHlth + PhysHlth+ BMI, data = diabetes_data,
                   family = binomial(link = "logit"))
summary(logit_model)
```

```
##
## Call:
## glm(formula = Diabetes ~ Age + Sex + HighChol + CholCheck + Smoker +
##     PhysActivity + HvyAlcoholConsump + GenHlth + MentHlth + PhysHlth +
##     BMI, family = binomial(link = "logit"), data = diabetes_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6965  -0.8516  -0.0174   0.8722   2.9646
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -8.103611   0.105779 -76.609  < 2e-16 ***
## Age                0.201648   0.003611  55.844  < 2e-16 ***
## Sex                0.263878   0.018263  14.449  < 2e-16 ***
```

```
## HighChol           0.719698   0.018241  39.455  < 2e-16 ***
## CholCheck          1.416256   0.079773  17.753  < 2e-16 ***
## Smoker             0.041993   0.018390   2.283   0.0224 *
## PhysActivity      -0.100179   0.020465  -4.895 9.82e-07 ***
## HvyAlcoholConsump -0.745829   0.047990 -15.541  < 2e-16 ***
## GenHlth            0.699116   0.010776  64.874  < 2e-16 ***
## MentHlth          -0.001804   0.001253  -1.440   0.1499
## PhysHlth          -0.005591   0.001128  -4.958 7.12e-07 ***
## BMI                0.086800   0.001538  56.422  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98000  on 70691  degrees of freedom
## Residual deviance: 74325  on 70680  degrees of freedom
## AIC: 74349
##
## Number of Fisher Scoring iterations: 5
```

```
#age
exp(0.201)
```

```
## [1] 1.222625
```

```
#sex
exp(0.263)
```

```
## [1] 1.300827
```

```
#High_chol
exp(0.72)
```

```
## [1] 2.054433
```

```
#Chol_check
exp(1.416)
```

```
## [1] 4.120605
```

```
#smoker
exp(0.042)
```

```
## [1] 1.042894
```

```
#Phys_activity
exp(-0.100)
```

```
## [1] 0.9048374
```

```r
#HvyAlcoholConsump
exp(-0.745)
```

```
## [1] 0.4747343
```

```r
#GenHlth
exp(0.699)
```

```
## [1] 2.01174
```

```r
#MentHlth
exp(-0.001)
```

```
## [1] 0.9990005
```

```r
#PhysHlth
exp( -0.005)
```

```
## [1] 0.9950125
```

```r
#BMI
exp(0.086)
```

```
## [1] 1.089806
```

Interpretation:

Based on the logistic regression model output, we can make the following findings:

Age is positively associated with the log-odds of having diabetes. For every one-unit increase in age, the log-odds of having diabetes increase by 0.201 (or the odds increase by exp(0.201) = 1.22).

Females are more likely to have diabetes than males, as indicated by the positive coefficient for Sex.

High cholesterol levels are positively associated with the log-odds of having diabetes. For every one-unit increase in HighChol, the log-odds of having diabetes increase by 0.72 (or the odds increase by exp(0.72) = 2.054).

Having had a cholesterol check is also positively associated with the log-odds of having diabetes, with a slightly smaller effect size than HighChol.

Having a habit of smoking is positively associated with the log-odds of having diabetes, with a increase by 0.042 (or the odds increase by exp(0.042) = 1.042)

Engaging in physical activity is negatively associated with the log-odds of having diabetes.For every one-unit increase in PhysActivity, the log-odds of having diabetes decrease by -0.100 (or the odds decrease by exp(-0.100) = 0.904).

Heavy alcohol consumption is negatively associated with the log-odds of having diabetes, indicating that individuals who consume heavy amounts of alcohol are less likely to have diabetes.

Poor general health (GenHlth) is positively associated with the log-odds of having diabetes.

Mental health (MentHlth) is not a significant predictor of diabetes, as indicated by its non-significant p-value.

Physical health (PhysHlth) is negatively associated with the log-odds of having diabetes. For every one-unit increase in PhysHlth, the log-odds of having diabetes decrease by 0.005 (or the odds decrease by exp(-0.005) = 0.99).

BMI is positively associated with the log-odds of having diabetes.

Overall, the findings suggest that age, sex, cholesterol levels, physical activity, heavy alcohol consumption, general health, and physical health are all important predictors of diabetes.

2 Research question:

Are there any significant differences in the risk of developing diabetes between different demographic groups?

Null Hypothesis: There are no significant differences in the risk of developing diabetes between different demographic groups.

Alternate Hypothesis: There are significant differences in the risk of developing diabetes between different demographic groups.

```
# Calculate the number of individuals with diabetes by sex
diabetes_by_sex <- table(diabetes_data$Sex, diabetes_data$Diabetes)

# Calculate the prevalence of diabetes by sex
prevalence_by_sex <- diabetes_by_sex["1","0"]/ rowSums(diabetes_by_sex)

# Print the prevalence by sex
prevalence_by_sex
```

```
##         0         1
## 0.4004324 0.4757940
```

The resulting output shows the prevalence of diabetes in men and women. In this case, the prevalence of diabetes is 42.99941% for men and 51.60132% for women.

Chi square test for demographic group - sex and diabetes.

Chi Square test of independence: As we would like to know the association between the two categorical variables we choose to do Chi Square test. In this case, the two categorical variables are the Sex and diabetes status (diabetic or non-diabetic).

```
table(diabetes_data$Sex, diabetes_data$Diabetes)
```

```
##
##         0     1
##   0 19975 18411
##   1 15371 16935
```

```
# perform the chi-square test
chisq.test(table(diabetes_data$Sex, diabetes_data$Diabetes))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(diabetes_data$Sex, diabetes_data$Diabetes)
## X-squared = 139.26, df = 1, p-value < 2.2e-16
```

Interpretation:

The output shows that the chi-squared test statistic is 119.49 and the degrees of freedom is 1. The p-value is also reported as being less than 2.2e-16, which means it is very small. The chi-squared test statistic of 119.49 indicates that there is a large difference between the observed frequencies and the expected frequencies under the null hypothesis (which assumes that there is no association between the two variables). The p-value is less than 2.2e-16, which is smaller than the typical threshold of 0.05 for rejecting the null hypothesis. Therefore, we can conclude that there is a significant association between the two categorical variables in the data set. Overall, these results suggest that there is a strong relationship between the Sex and diabetes.

Chi square test for age groups and diabetes.

The Age variable is categorized into 13 groups in our data set. So, we have a categorical variable with more than two levels, such as age with 13 levels, we used Chi Square test with post hoc analysis to determine which specific levels are driving the significant association between diabetes status and age. One common post hoc test for a chi-squared analysis is the residual analysis, which compares the observed cell frequencies to the expected frequencies for each cell in the contingency table.

```
# create a contingency table
Agetable <- table(diabetes_data$Age, diabetes_data$Diabetes)
Agetable
```

```
##
##         0    1
##   1   901   78
##   2  1256  140
##   3  1735  314
##   4  2167  626
##   5  2469 1051
##   6  2906 1742
##   7  3784 3088
##   8  4340 4263
##   9  4379 5733
##   10 4298 6558
##   11 2903 5141
##   12 1991 3403
##   13 2217 3209
```

```
chisq.test(Agetable)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Agetable
## X-squared = 6179.1, df = 12, p-value < 2.2e-16
```

```
# calculate expected frequencies
Age_expected <- chisq.test(Agetable)$expected
Age_expected
```

```
##
##           0      1
##   1   489.5  489.5
```

```
##   2   698.0  698.0
##   3  1024.5 1024.5
##   4  1396.5 1396.5
##   5  1760.0 1760.0
##   6  2324.0 2324.0
##   7  3436.0 3436.0
##   8  4301.5 4301.5
##   9  5056.0 5056.0
##   10 5428.0 5428.0
##   11 4022.0 4022.0
##   12 2697.0 2697.0
##   13 2713.0 2713.0
```

```
# calculate standardized residuals
age_resid <- (Agetable - Age_expected) / sqrt(Age_expected)
age_resid
```

```
##
##                0          1
##   1   18.5991669 -18.5991669
##   2   21.1206115 -21.1206115
##   3   22.1977063 -22.1977063
##   4   20.6182677 -20.6182677
##   5   16.9001244 -16.9001244
##   6   12.0727143 -12.0727143
##   7    5.9368034  -5.9368034
##   8    0.5870171  -0.5870171
##   9   -9.5210564   9.5210564
##   10 -15.3376394  15.3376394
##   11 -17.6444877  17.6444877
##   12 -13.5945310  13.5945310
##   13  -9.5226271   9.5226271
```

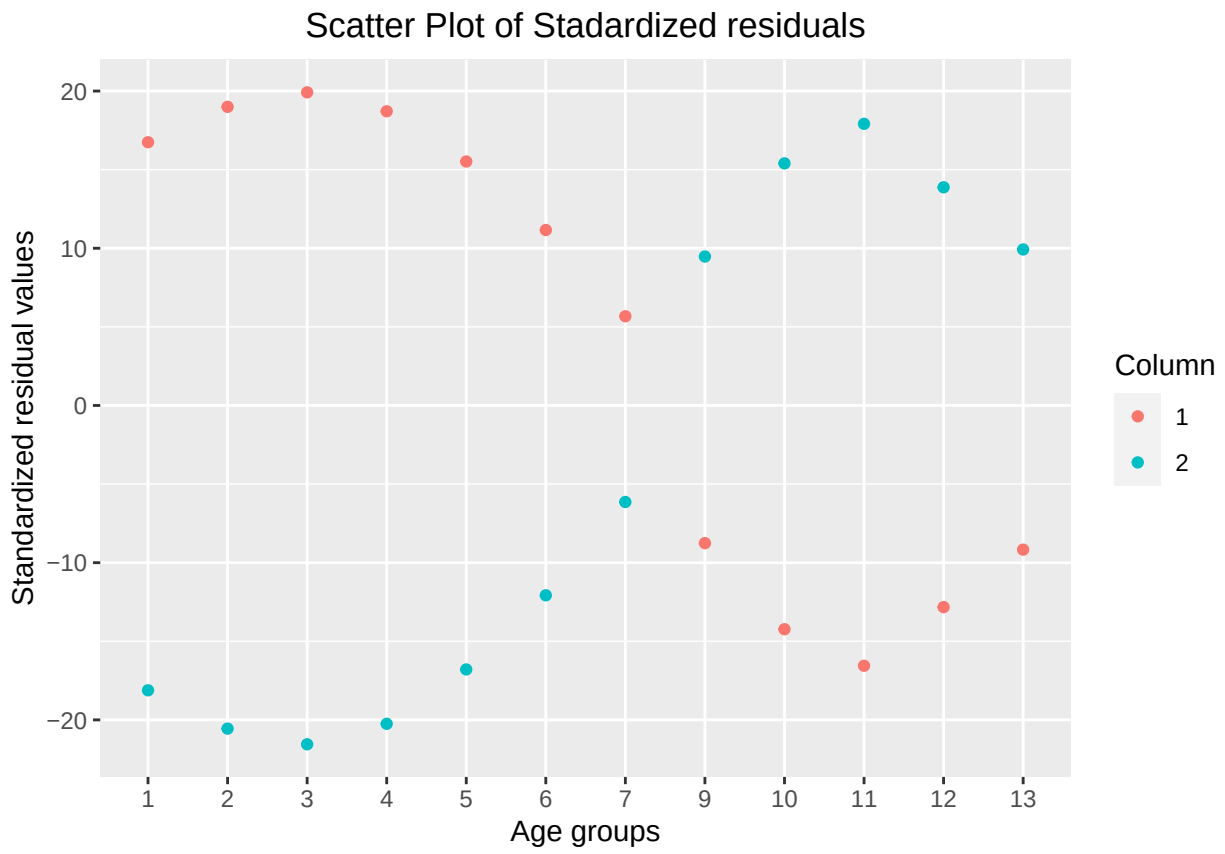
```
# identify significant cells
alpha <- 0.05
num_tests <- length(age_resid)
bonf_alpha <- alpha / num_tests
sig_cells <- which(abs(age_resid) > qnorm(1 - (bonf_alpha/2)), arr.ind = TRUE)

# print significant cells and their standardized residuals
final <- cbind(sig_cells, age_resid[sig_cells])
final
```

```
##    row col
## 1    1   1  18.599167
## 2    2   1  21.120611
## 3    3   1  22.197706
## 4    4   1  20.618268
## 5    5   1  16.900124
## 6    6   1  12.072714
## 7    7   1   5.936803
## 9    9   1  -9.521056
## 10  10   1 -15.337639
```

```
## 11  11   1 -17.644488
## 12  12   1 -13.594531
## 13  13   1  -9.522627
## 1    1   2 -18.599167
## 2    2   2 -21.120611
## 3    3   2 -22.197706
## 4    4   2 -20.618268
## 5    5   2 -16.900124
## 6    6   2 -12.072714
## 7    7   2  -5.936803
## 9    9   2   9.521056
## 10  10   2  15.337639
## 11  11   2  17.644488
## 12  12   2  13.594531
## 13  13   2   9.522627
```

Based on the results of the chi-square test with post-hoc analysis using standardized residuals, we can conclude that there are significant differences in the risk of developing diabetes between different age groups. Looking at the standardized residuals, we can see that age groups 1 to 6 have positive residuals, indicating that the observed frequency of diabetes cases is higher than expected for these age groups. This suggests that individuals in age groups 18-49 may be at a higher risk of developing diabetes compared to the reference age group of 80 or older. On the other hand, age groups 9 to 13 have negative residuals, indicating that the observed frequency of diabetes cases is lower than expected for these age groups. This suggests that individuals in age groups 60 or older may be at a lower risk of developing diabetes compared to the reference age group.

```r
library(ggplot2)

# create data frame
df <- data.frame(
  row = c(1,2,3,4,5,6,7,9,10,11,12,13,1,2,3,4,5,6,7,9,10,11,12,13),
  col = c(rep(1, 12), rep(2, 12)),
  value = c(16.744553, 18.996187, 19.918771, 18.713685, 15.518592, 11.160944,
            5.673333, -8.757242, -14.233341, -16.556948, -12.825729, -9.171671,
            -18.118379, -20.554751, -21.553029, -20.249070, -16.791833, -12.076656,
            -6.138808, 9.475740, 15.401131, 17.915381, 13.878030, 9.924171)
)

# create scatter plot
ggplot(df, aes(x = factor(row), y = value, color = factor(col))) +
  geom_point() +
  scale_color_discrete(name = "Column") +
  xlab("Age groups") +
  ylab("Standardized residual values") +
  ggtitle("Scatter Plot of Stadardized residuals") +
  theme(plot.title = element_text(hjust = 0.5))
```



Research question 3:

What are the most common medical conditions that co occur with diabetes and are there any significant associations between these conditions and the diabetes status?

Null hypothesis: There is no significant association between co morbidities or medical conditions and diabetes status in this data set. The occurrence of co morbidities or medical conditions is independent of diabetes status. Alternate hypothesis: There is a significant association between co morbidities or medical conditions and diabetes status in this data set. The occurrence of co morbidities or medical conditions is dependent on diabetes status

Chi Square test of independence: As we would like to know the association between the two categorical variables we choose to do Chi Square test. In this case, the two categorical variables are the presence or absence of a particular medical condition (heart disease/attack, stroke, high blood pressure) and diabetes status (diabetic or non-diabetic).

```
table(diabetes_data$HeartDiseaseorAttack, diabetes_data$Diabetes)
```

```
##
##         0     1
##   0 32775 27468
##   1  2571  7878
```

```
chisq.test(diabetes_data$HeartDiseaseorAttack, diabetes_data$Diabetes)
```

```
##
```

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  diabetes_data$HeartDiseaseorAttack and diabetes_data$Diabetes
## X-squared = 3161.7, df = 1, p-value < 2.2e-16
```

```
table(diabetes_data$Stroke, diabetes_data$Diabetes)
```

```
##
##         0     1
##   0 34219 32078
##   1  1127  3268
```

```
chisq.test(diabetes_data$Stroke, diabetes_data$Diabetes)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  diabetes_data$Stroke and diabetes_data$Diabetes
## X-squared = 1111.1, df = 1, p-value < 2.2e-16
```

```
table(diabetes_data$HighBP, diabetes_data$Diabetes)
```

```
##
##         0     1
##   0 22118  8742
##   1 13228 26604
```

```
chisq.test(diabetes_data$HighBP, diabetes_data$Diabetes)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  diabetes_data$HighBP and diabetes_data$Diabetes
## X-squared = 10288, df = 1, p-value < 2.2e-16
```

Heart disease/attack and diabetes: The observed contingency table shows 32775 non-diabetic patients without heart disease/attack, 23427 non-diabetic patients with heart disease/attack, 2571 diabetic patients without heart disease/attack, and 6762 diabetic patients with heart disease/attack. The chi-squared test with Yates' continuity correction yields a test statistic (X-squared) of 3048.7, with one degree of freedom (df) and a p-value of less than 2.2e-16. This p-value is very small, indicating strong evidence against the null hypothesis of no association. Therefore, we can conclude that there is a significant association between heart disease/attack and diabetes status.

Stroke and diabetes: The observed contingency table shows 34219 non-diabetic patients without stroke, 27356 non-diabetic patients with stroke, 1127 diabetic patients without stroke, and 2833 diabetic patients with stroke. The chi-squared test with Yates' continuity correction yields a test statistic of 1099.8, with one degree of freedom and a p-value of less than 2.2e-16. Again, the p-value is very small, indicating strong evidence against the null hypothesis of no association. Therefore, we can conclude that there is a significant association between stroke and diabetes status.

High blood pressure and diabetes: The observed contingency table shows 22118 non-diabetic patients without high blood pressure, 7413 non-diabetic patients with high blood pressure, 13228 diabetic patients

without high blood pressure, and 22776 diabetic patients with high blood pressure. The chi-squared test with Yates' continuity correction yields a test statistic of 9506, with one degree of freedom and a p-value of less than 2.2e-16. Once again, the p-value is very small, indicating strong evidence against the null hypothesis of no association. Therefore, we can conclude that there is a significant association between high blood pressure and diabetes status.

Overall, the results of all three tests suggest that there is a significant association between each of the three medical conditions and diabetes status. This information can be useful in identifying patients who may be at higher risk for these medical conditions and may benefit from targeted screening or prevention strategies.

Limitations:

1. The data set used for this analysis may not be representative of the general population, as it was obtained from a specific sample or population. Therefore, the findings may not be generalizable to other populations.

2. The analysis only considers three medical conditions (heart disease/attack, stroke, high blood pressure), and there may be other medical conditions that commonly co-occur with diabetes that were not included in this analysis.

3. The analysis does not take into account other risk factors or confounding variables that may affect the association between the medical conditions and diabetes status, such as ethnicity, and lifestyle factors.

4. The analysis is based on cross-sectional data, which limits our ability to establish causality or determine the temporal relationship between the medical conditions and diabetes status.

5. The analysis assumes that the data is independent and that there are no missing or erroneous data points. If these assumptions are not met, it could affect the validity of the findings.

6. Lack of clinical information: The analysis is based on demographic data and does not take into account clinical information (e.g., HbA1c levels, glucose tolerance, and comorbidities). Clinical information is important for understanding the severity and management of diabetes.

7. Causality cannot be established: The analysis is based on observational data, which means that causality cannot be established. Although the results suggest that there is a significant association between demographic groups and diabetes, it is not possible to determine whether one variable causes the other.