EDA

Holly Probasco

Libraries

```
lapply(c("readr", "dplyr", "tidyverse", "ggplot2", "tidyr", "scales"), library,
  character.only = TRUE)
```

Introduction Section

This data is the results from a yearly survey done by the CDC regarding general health. This survey from 2015 had 253,680 responses where respondents gave answers to different health-related questions. These answers are used by the CDC to help track health objectives as well as better implement disease prevention. CDC Site

The variables we are working with in this file are going to be used to look at diabetes indicators. Since this was a survey, few. This analysis will focus on 4 predictor variables specifically that I think would create an accurate predictive model for Diabetes. The variables I have chosen to look into more in depth are Smoker, GenHlth, and BMI. I thought smoker would be interesting here because we know that smoking negatively affects basically everything. It causes cancer in many forms, increases heart disease and attack risk, damages cells, and that's not even half of them. Therefore, looking at Smoker here makes sense to me as a possible indicator for Diabetes. The GenHlth variable is a scale of 1-5 (1 being excellent) on how healthy in general the respondent thinks they are. I chose this variable because I think generally people think they are healthier than they actually are. So, looking at this variable could show that people who answered this question with a good score could actually have a higher risk based on other factors. Lastly, I chose BMI because generally diabetes is associated with people who are overweight. BMI is a measure of this, so I wanted to see if there was a spike at a certain BMI, or if there is a dependence on BMI with the other variables. Risk Factors according to CDC

The purpose of the EDA done in this file is to get to know the dataset used as well as the variables chosen. We will look at different summaries of the data visually with tables and charts, to look at the distribution of each variable.

The ultimate goal of modeling in this context is to take existing data in order to predict which factors may result in diabetes for people. In this way, we aim to prevent other people with similar traits from getting the disease.

Data

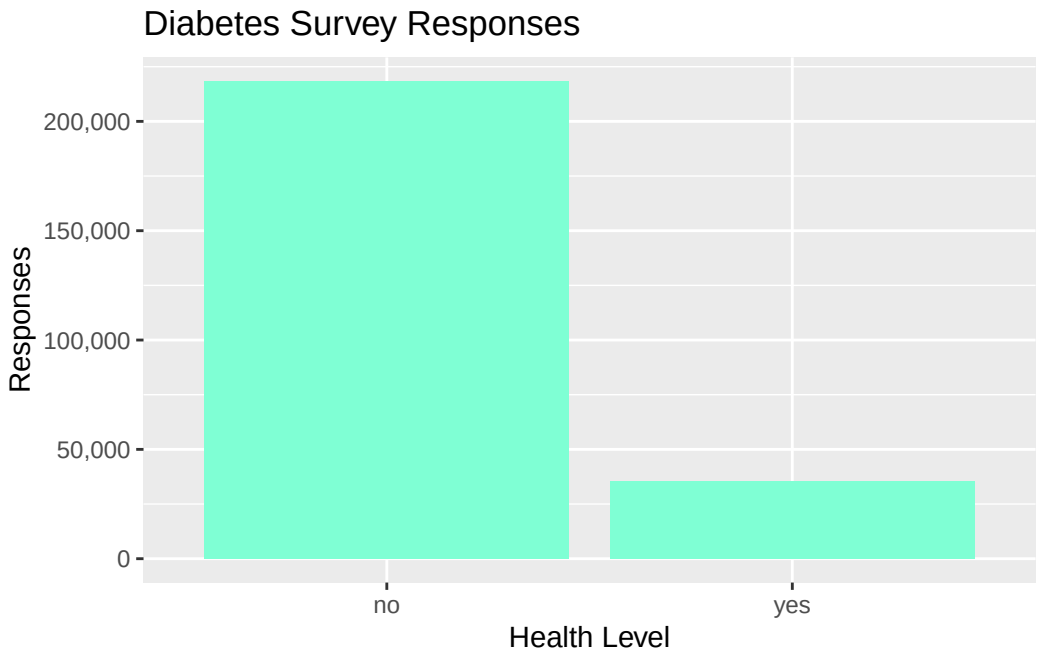
```
diabetes_data <- read_csv("diabetes_binary_health_indicators_BRFSS2015.csv") |>
select(Diabetes_binary, BMI, Smoker, GenHlth) |> drop_na() |>
mutate(Diabetes = factor(if_else(Diabetes_binary == 0, "no", "yes")),
Smoking_status = factor(if_else(Smoker == 0, "no", "yes")), Health_level = case_when(
  GenHlth == "1" ~ "excellent",
  GenHlth == "2" ~ "verygood",
  GenHlth == "3" ~ "good",
  GenHlth == "4" ~ "fair",
  GenHlth == "5" ~ "poor",
  .default = "NA"
)) |>
mutate(Health_level = factor(Health_level,
levels = c("excellent", "verygood", "good", "fair", "poor"),ordered = TRUE))
```

Summarizations

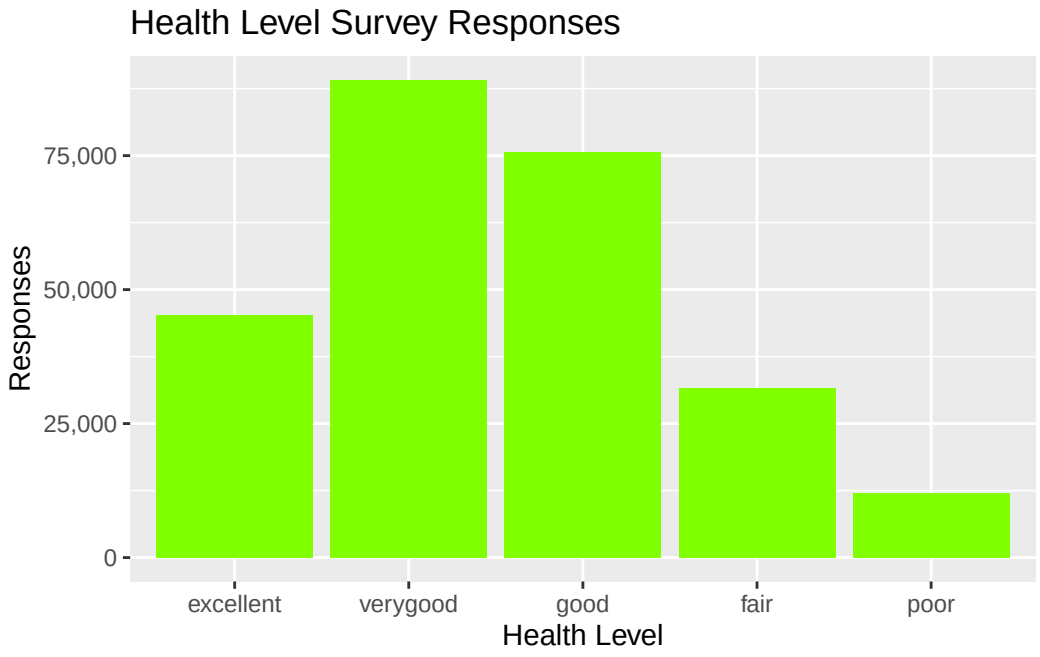
Univariate Explorations

First, we do some univariate exploration on the three variables chosen. This way, we can see the spread of the data we are working with and if there is anything that jumps out visually.

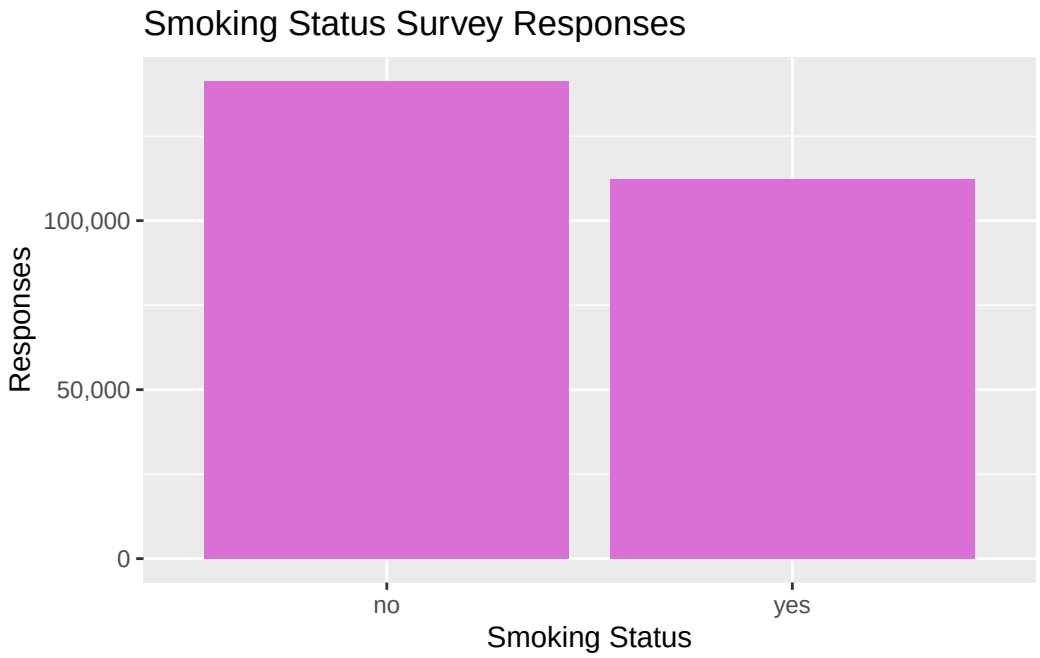
```
ggplot(data = diabetes_data, aes(x = Diabetes)) + geom_bar(fill = "aquamarine") +
labs(x = "Health Level", y = "Responses", title = "Diabetes Survey Responses") +
scale_y_continuous(labels = scales::label_comma())
```



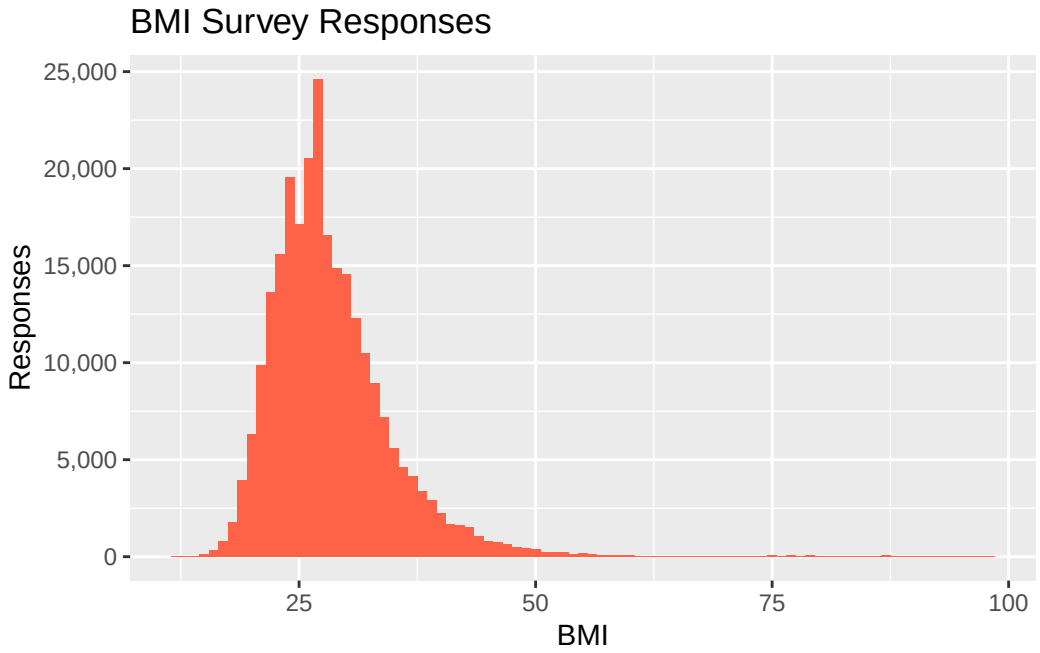

```
ggplot(data = diabetes_data, aes(x = Health_level)) +
geom_bar(fill = "chartreuse") +
labs(x = "Health Level", y = "Responses", title = "Health Level Survey Responses") +
scale_y_continuous(labels = scales::label_comma())
```



```
ggplot(data = diabetes_data, aes(x = Smoking_status)) + geom_bar(fill = "orchid") +
labs(x = "Smoking Status", y = "Responses", title = "Smoking Status Survey Responses") +
scale_y_continuous(labels = scales::label_comma())
```




```
ggplot(data = diabetes_data, aes(x = BMI)) +
geom_histogram(fill = "tomato", binwidth = 1) + labs(y = "Responses",
title = "BMI Survey Responses") +
scale_y_continuous(labels = scales::label_comma())
```



Based on these graphs, we can see the following for the variables: Diabetes responses show that there is a lot more people in this survey that do not have diabetes than those that do Health Level is right skewed, implying that respondents do consider themselves to be generally on the healthy side Smoking Status is close to even, which will be helpful in modeling as we have similar group sizes for each BMI peaks at just under 30, showing that many people in this survey are overweight. However, there are also many responses over 50, and BMI does not really even go that high, implying that respondents many not actually know what BMI is

Bivariate Exploration

Next, we can look to see how each variable distributes with/without the presence of Diabetes, because since this is what we will be doing our analysis on later, it would be helpful to see if there is a visual change when Diabetes is present vs. not.

```
diabetes_data |> group_by(Diabetes, Smoking_status) |>
summarize(count = n(), .groups = "drop")
```

```
# A tibble: 4 x 3
  Diabetes Smoking_status  count
  <fct>    <fct>           <int>
1 no       no             124228
2 no       yes             94106
```

```
3 yes      no               17029
4 yes      yes              18317
```

Based on this table, we can see that

nonsmoker/no diabetes = 124228 / (124228 + 94106) = 56.9%

smoker/no diabetes = 94106 / (124228 + 94106) = 43.1%

nonsmoker/diabetes = 17029 / (17029 + 18317) = 48.2%

smoker/diabetes = 18317 / (17029 + 18317) = 51.8%

Therefore, it does not look like, based on this table, smoking has a drastic effect on whether or not someone may have diabetes

Next, let's look at BMI

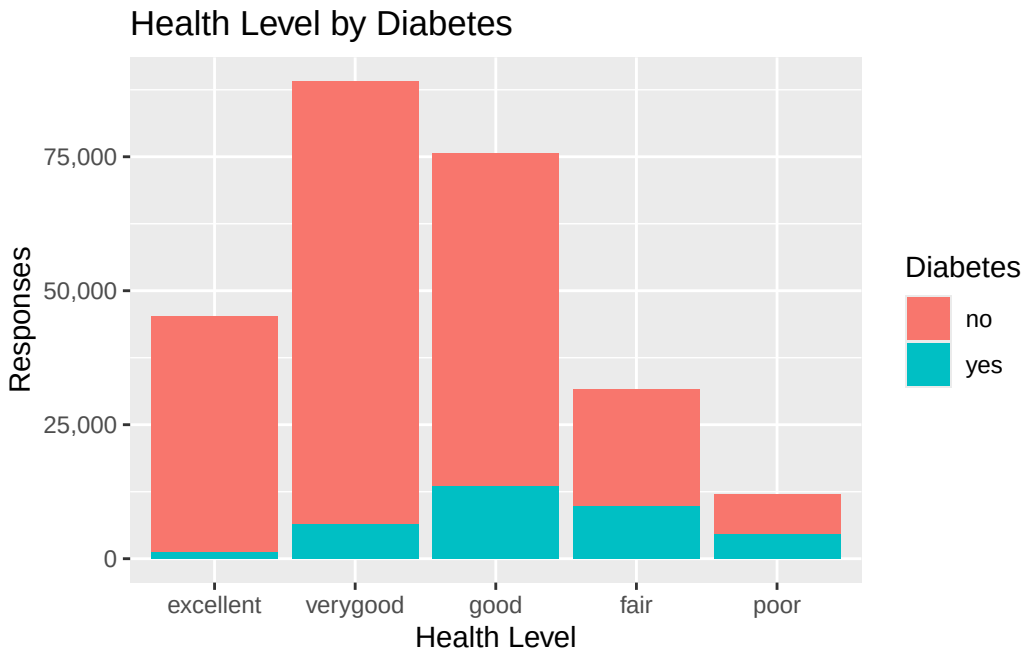
```
diabetes_data |>
  group_by(Diabetes) |>
  summarize(mean_BMI = mean(BMI), .groups = "drop")
```

```
# A tibble: 2 x 2
  Diabetes mean_BMI
  <fct>       <dbl>
1 no           27.8
2 yes          31.9
```

Looking at the mean here, we do see that BMI for those with Diabetes is 4 points higher than those without. Obesity is considered to be a BMI of 30 or higher, which could point to why this is the case.

Plot of Health Level vs Diabetes

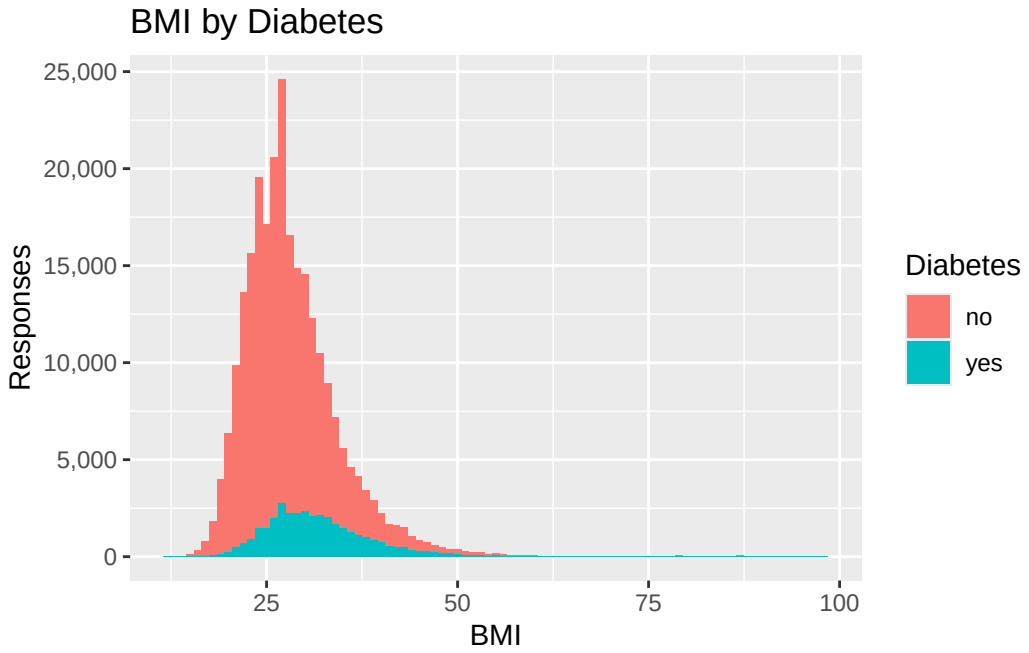
```
ggplot(data = diabetes_data, aes(x = Health_level, fill = Diabetes)) + geom_bar() +
labs(x = "Health Level", y = "Responses", title = "Health Level by Diabetes") +
scale_y_continuous(labels = scales::label_comma())
```



There are a handful of people who gave themselves an excellent health score who also have diabetes. Not all people who have Diabetes are unhealthy, but an interesting spread nonetheless.

Plot of BMI by Diabetes

```
ggplot(diabetes_data, aes(x = BMI, fill = Diabetes)) + geom_histogram(binwidth = 1) +
labs(y = "Responses", title = "BMI by Diabetes") + scale_y_continuous(labels = scales::label
```



There seems to be a normal curve for both no diabetes as well as those with diabetes. It looks like the Diabetes curve peaks at a higher BMI than no Diabetes

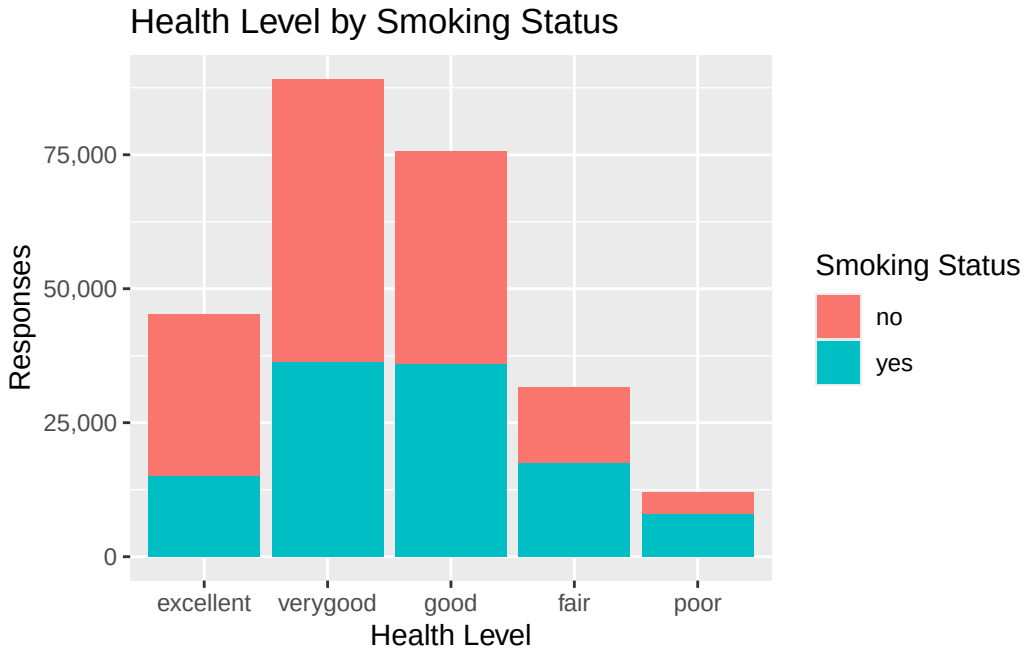
Now, we should also look to see how the variables relate to each other.

```
diabetes_data |>
  group_by(Health_level) |>
  summarize(mean_BMI = mean(BMI), .groups = "drop") |> arrange(mean_BMI)
```

```
# A tibble: 5 x 2
  Health_level mean_BMI
  <ord>           <dbl>
1 excellent        25.8
2 verygood         27.6
3 good             29.5
4 poor             30.6
5 fair             30.7
```

Unsurprisingly, BMI gets larger as Health_Level decreases. Though poor and fair have switched in order, they are extremely similar

```
ggplot(data = diabetes_data, aes(x = Health_level, fill = Smoking_status)) + geom_bar() +
labs(x = "Health Level", y = "Responses", title = "Health Level by Smoking Status", fill = "S
scale_y_continuous(labels = scales::label_comma())
```



Interestingly, more people that gave themselves an excellent health rating smoke than those who gave themselves a poor rating

Click here for the Modeling Page