NHANES Exploratory Data Analysis Project: Visualization

Tiffany Taylor and Dr. Kasthuri Kannan

6/24/2020

The following is a project completed using the NHANES data set for years 2013 and 2014. It provides a visual exploratory analysis of the relationships between age, gender, diabetes, and glucose levels.

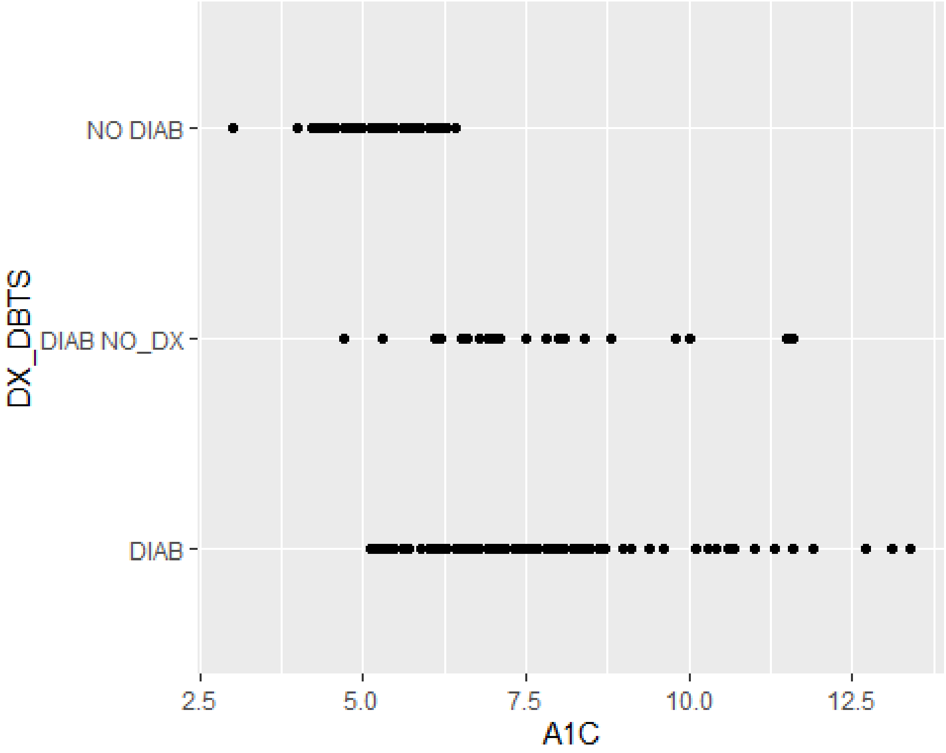
```
library(RCurl)
URL_text_1 <- "https://raw.githubusercontent.com/kannan-kasthuri/kannan-kasthuri.github.i
o"
URL_text_2 <- "/master/Datasets/HANES/NYC_HANES_DIAB.csv"
URL <- paste(URL_text_1,URL_text_2, sep="")
HANES <- read.csv(text=getURL(URL))
HANES$GENDER <- factor(HANES$GENDER, labels=c("M","F"))
HANES$AGEGROUP <- factor(HANES$AGEGROUP, labels=c("20-39","40-59","60+"))
HANES$HSQ_1 <- factor(HANES$HSQ_1, labels=c("Excellent","Very Good","Good", "Fair", "Poor
"))
HANES$DX_DBTS <- factor(HANES$DX_DBTS, labels=c("DIAB","DIAB NO_DX","NO DIAB"))
HANES <- na.omit(HANES)
str(HANES)
```

```
## 'data.frame':    1112 obs. of  23 variables:
##  $ KEY              : chr  "134040A" "134460B" "134490A" "134620A" ...
##  $ GENDER           : Factor w/ 2 levels "M","F": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SPAGE            : int  29 28 27 24 30 26 31 32 34 32 ...
##  $ AGEGROUP         : Factor w/ 3 levels "20-39","40-59",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ HSQ_1            : Factor w/ 5 levels "Excellent","Very Good",..: 2 2 2 1 1 3 1 2 1
3 ...
##  $ UCREATININE      : int  105 53 314 105 163 150 46 36 177 156 ...
##  $ UALBUMIN         : num  0.707 1 8 4 3 2 2 0.707 4 3 ...
##  $ UACR             : num  0.00673 2 3 4 2 ...
##  $ MERCURYU         : num  0.37 0.106 0.487 2.205 0.979 ...
##  $ DX_DBTS          : Factor w/ 3 levels "DIAB","DIAB NO_DX",..: 3 3 3 3 3 3 3 3 3 3 .
..
##  $ A1C              : num  5 5.2 4.8 5.1 4.3 5.2 4.8 5.2 4.8 5.2 ...
##  $ CADMIUM          : num  0.2412 0.1732 0.0644 0.0929 0.1202 ...
##  $ LEAD             : num  1.454 1.019 0.863 1.243 0.612 ...
##  $ MERCURYTOTALBLOOD: num  2.34 2.57 1.32 14.66 2.13 ...
##  $ HDL              : int  42 51 42 61 52 50 57 56 42 44 ...
##  $ CHOLESTEROLTOTAL : int  184 157 145 206 120 155 156 235 156 120 ...
##  $ GLUCOSESI        : num  4.61 4.77 5.16 5 5.11 ...
##  $ CREATININESI     : num  74.3 73 80 84.9 66 ...
##  $ CREATININE       : num  0.84 0.83 0.91 0.96 0.75 0.99 0.9 0.84 0.93 1.09 ...
##  $ TRIGLYCERIDE     : int  156 43 108 65 51 29 31 220 82 35 ...
##  $ GLUCOSE          : int  83 86 93 90 92 85 72 87 96 92 ...
##  $ COTININE         : num  31.5918 0.0635 0.035 0.0514 0.035 ...
##  $ LDLESTIMATE      : int  111 97 81 132 58 99 93 135 98 69 ...
##  - attr(*, "na.action")= 'omit' Named int [1:415] 2 15 16 24 26 28 33 34 35 39 ...
##   ..- attr(*, "names")= chr [1:415] "2" "15" "16" "24" ...
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0
--

## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts()
--
## x tidyr::complete() masks RCurl::complete()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
```

```
ggplot(data=HANES)+
  geom_point(mapping = aes(x=A1C, y=DX_DBTS))
```



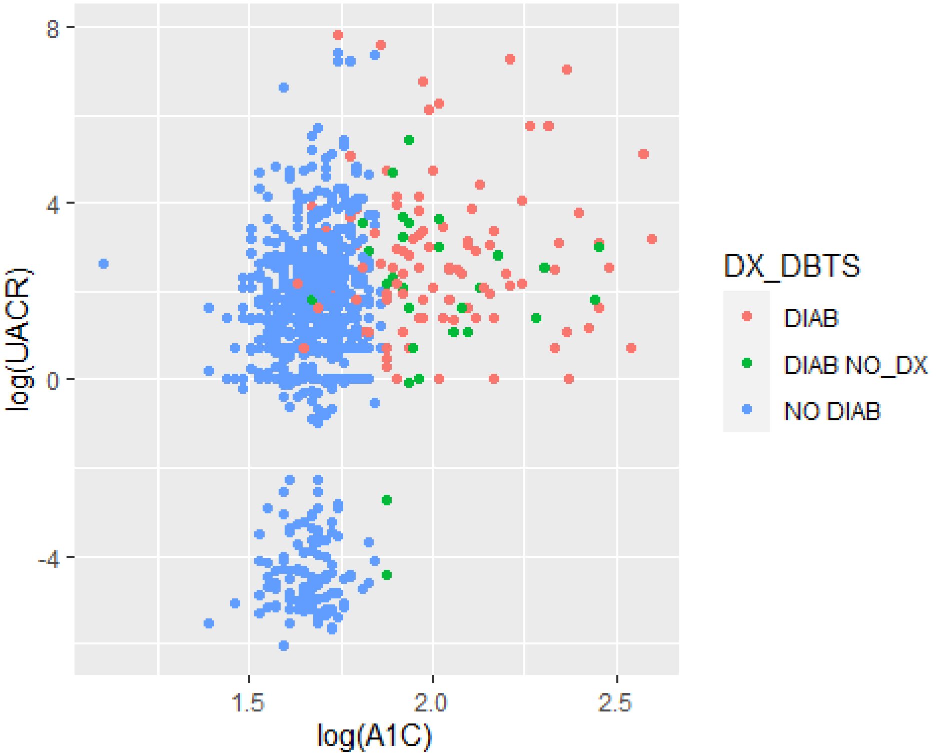
```
nrow=(HANES)
```

There are 1112 observations and 23 variables.

```
ncol=(HANES)
```

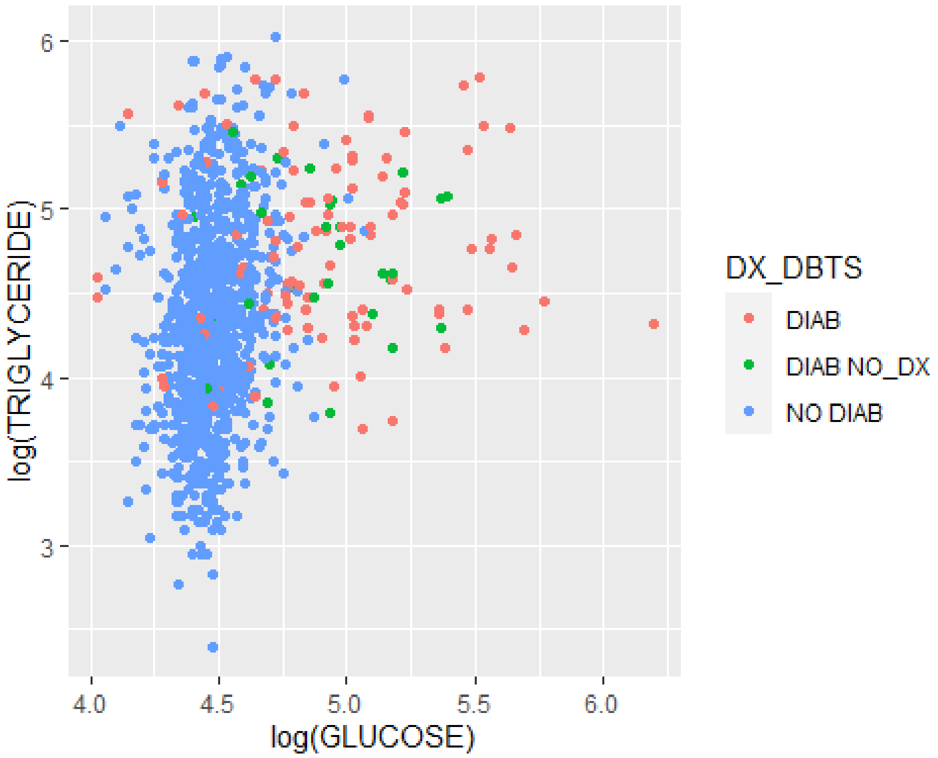
In the image above, we can see that most of the people with diabetes have an A1C level of at least 5.0. There is a lot of density around the 7.5 test level area. This makes sense because the consensus is that people get diabetes when their test levels start reaching near 6%. This plot shows a positive relationship between A1C levels and having diabetes. In other words, the higher someone's A1C levels, the more likely they are to have been diagnosed with diabetes. This confirms my hypothesis.

```
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(A1C), y = log(UACR), color=DX_DBTS))
```



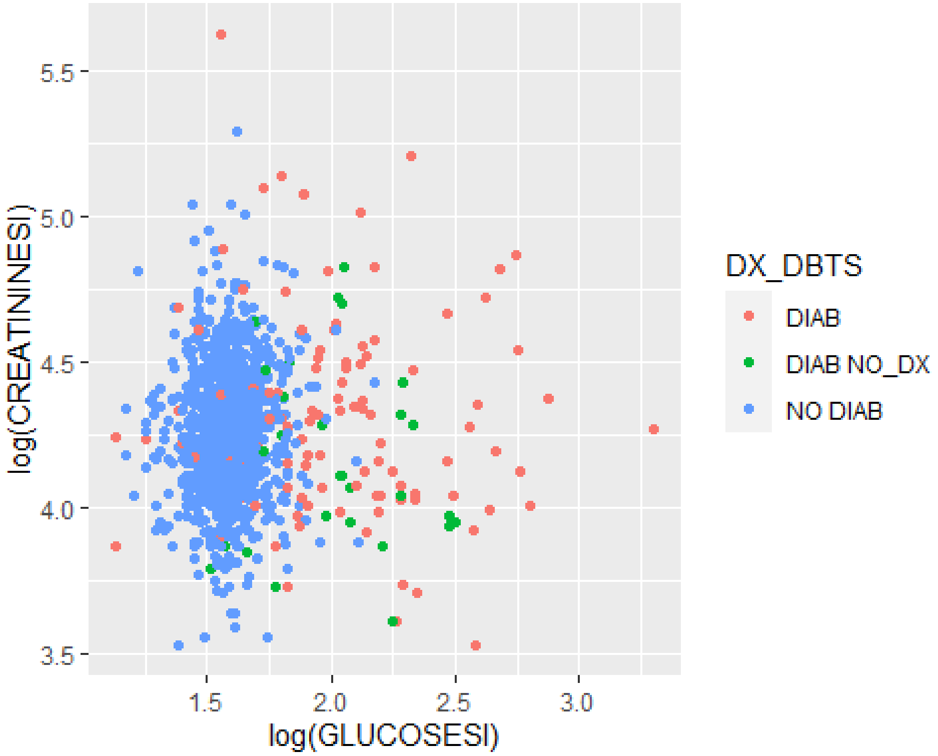
Above is the creation of a ggplot with aesthetic color for the variable DX_DBTS. This color visualization/graph reveals many things. Here it can be seen that most of the people in the clusters (each dot represents a person) does not have diabetes. Thus, it appears that UACR and A1C may not be risk factors for diabetes. Just to recap, A1C is a blood test level. Research shows that A1C test results of 6.5% or above indicate diabetes and prediabetes is from an A1C test of 5.7% to 6.4% so these results are consistent with mainstream data.
Source: https://www.virginiamason.org/whatarenormalbloodglucoselevels

```
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(GLUCOSE), y = log(TRIGLYCERIDE), color=DX_DBTS))
```



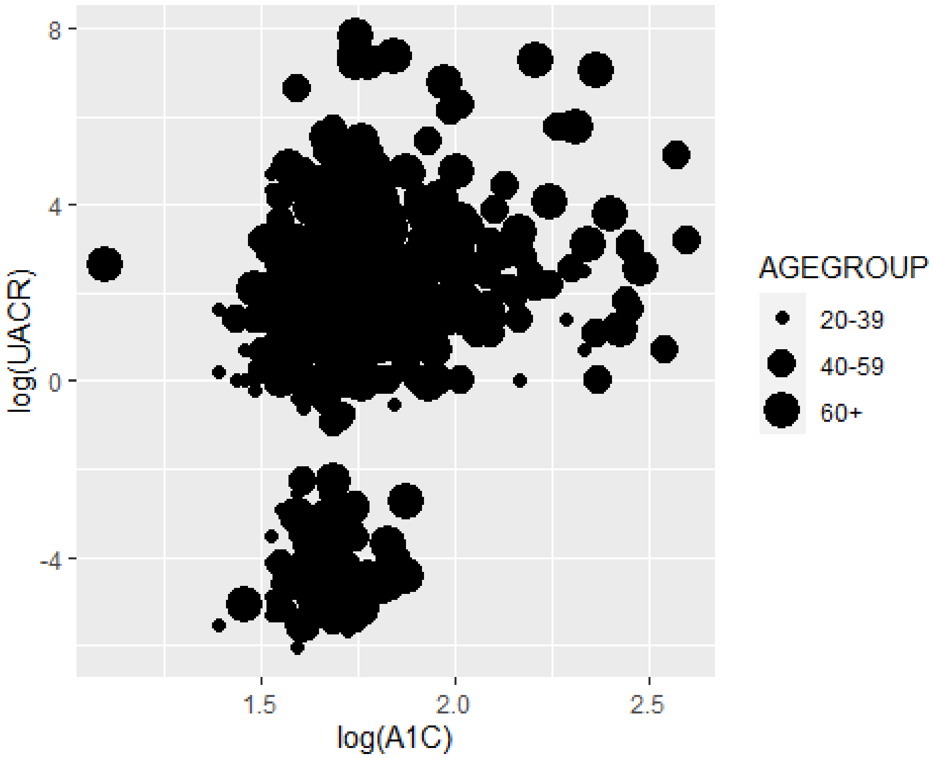
Above is a creation of a ggplot with aesthetic color for the variable DX_DBTS and x log Glucose with y log Triglyceride. This color visualization/graph reveals many things. Here it can be seen that most of the people have no diabetes. The people that do have diabetes when considering the log(triglyceride) vs log(Glucose) are seldom and sparse. This may indicate that triglyceride and glucose levels are not sole risk factors for diabetes.

```
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(GLUCOSESI), y = log(CREATININESI), color=DX_DBTS))
```



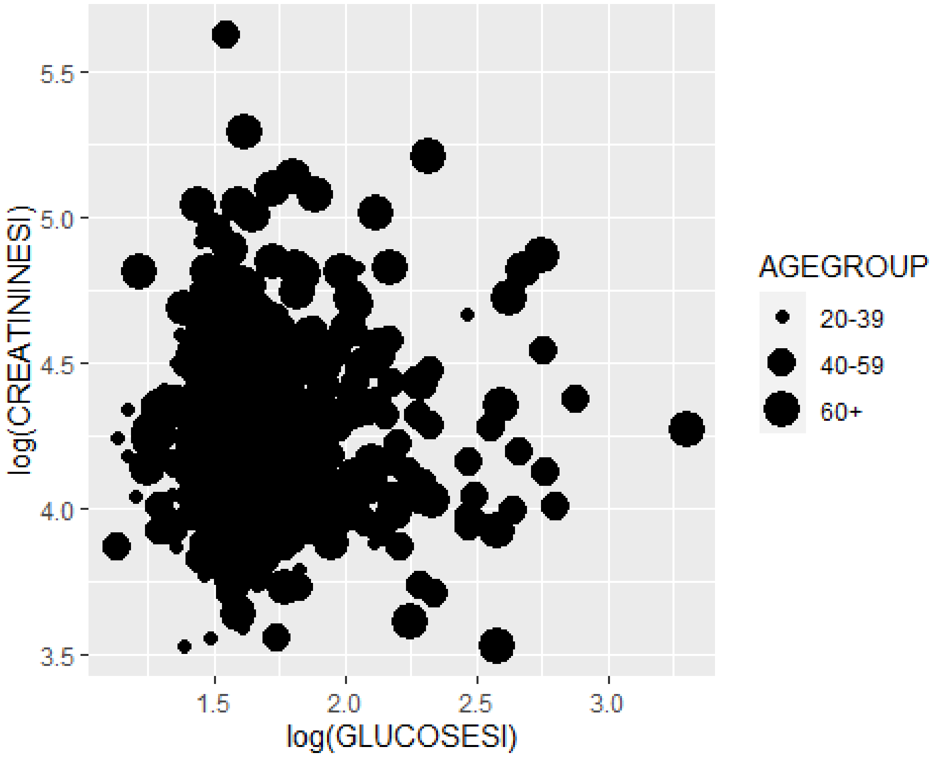
```r
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(A1C), y = log(UACR), size=AGEGROUP))
```

```
## Warning: Using size for a discrete variable is not advised.
```

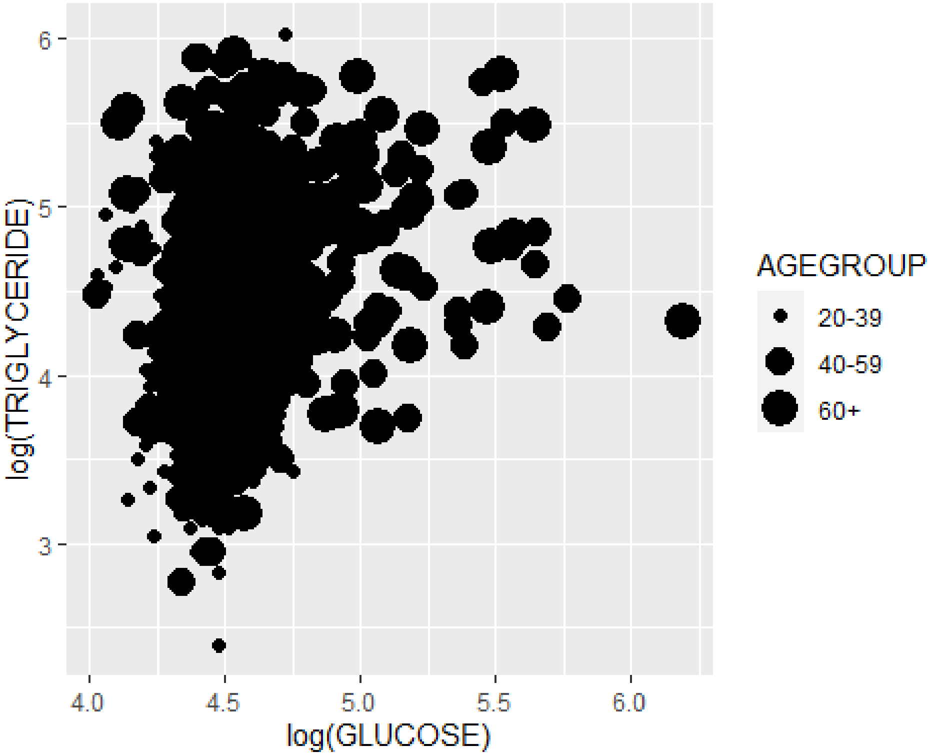


```r
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(GLUCOSESI), y = log(CREATININESI), size=AGEGROUP))
```

```
## Warning: Using size for a discrete variable is not advised.
```

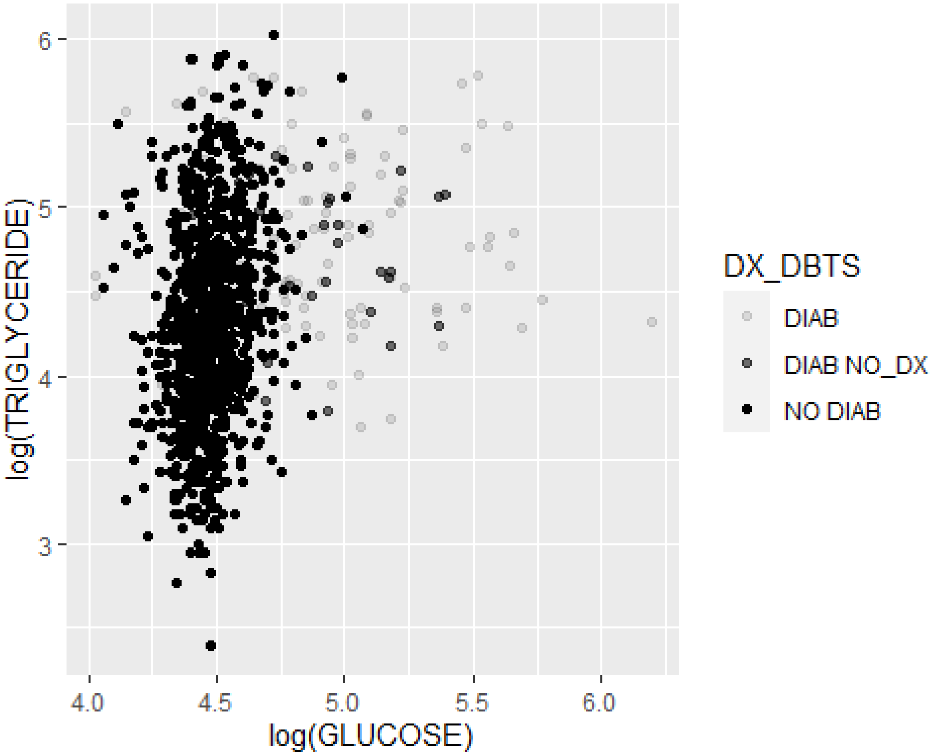


```
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(GLUCOSE), y = log(TRIGLYCERIDE), size=AGEGROUP))

## Warning: Using size for a discrete variable is not advised.
```



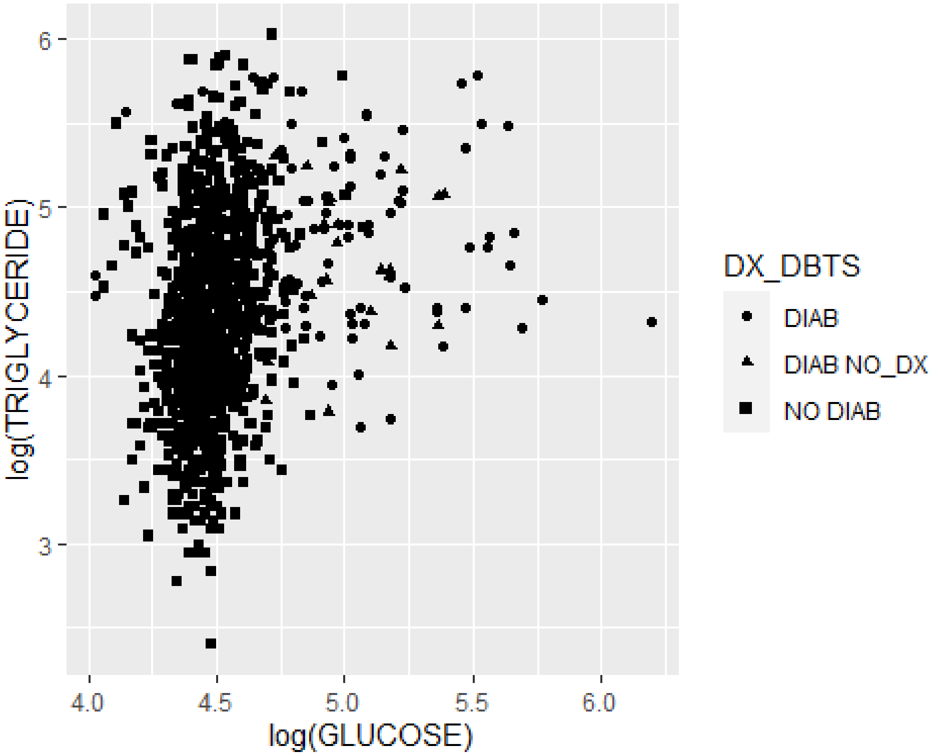
For the above graph, we see that the dots are congregated in one general areal between values of 3 and 5.5 on the log(Triglyceride) scale and from 4.25 and 4.75 on the log(glucose) scale. It is difficult to see the shape of the dots to indicate which dot is large or small so another type of graph would be best to look at indicating the age group. I need to retry this graph adjusting for transparency and shape of the dots so I am better able to view what exactly is going on and analyze sufficiently what is going on.

```
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(GLUCOSE), y = log(TRIGLYCERIDE), alpha=DX_DBTS))

## Warning: Using alpha for a discrete variable is not advised.
```



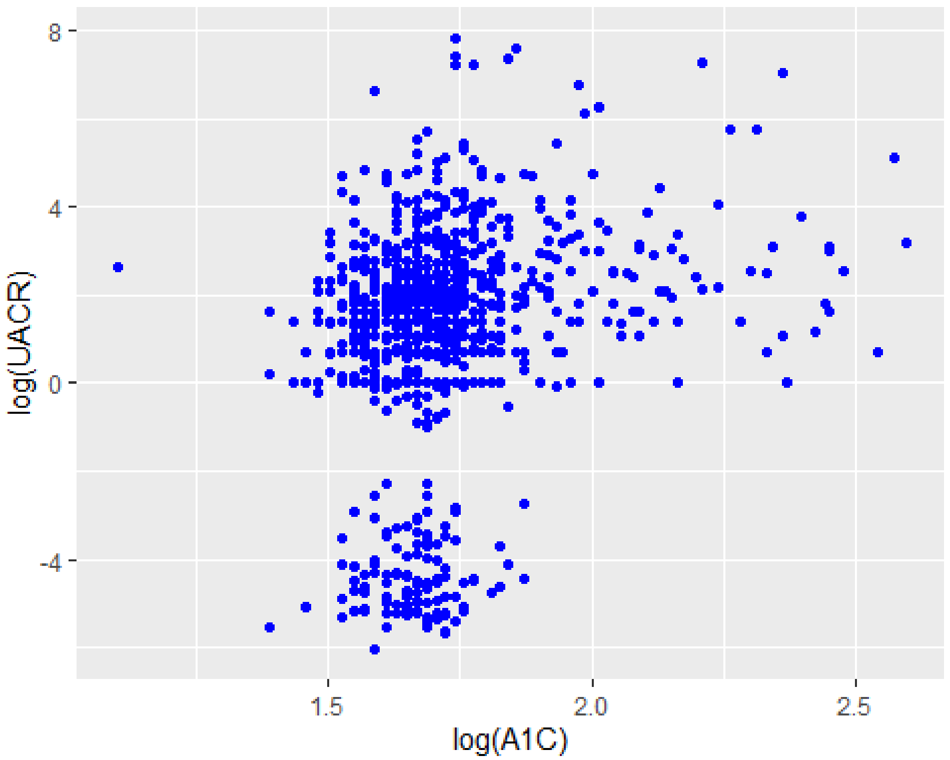
Now, after mapping DX_DBTS to the alpha aesthetic, the control over the transparency of the points is significantly improved. Here we see that most of the people have no diabetes.

```
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(GLUCOSE), y = log(TRIGLYCERIDE), shape=DX_DBTS))
```

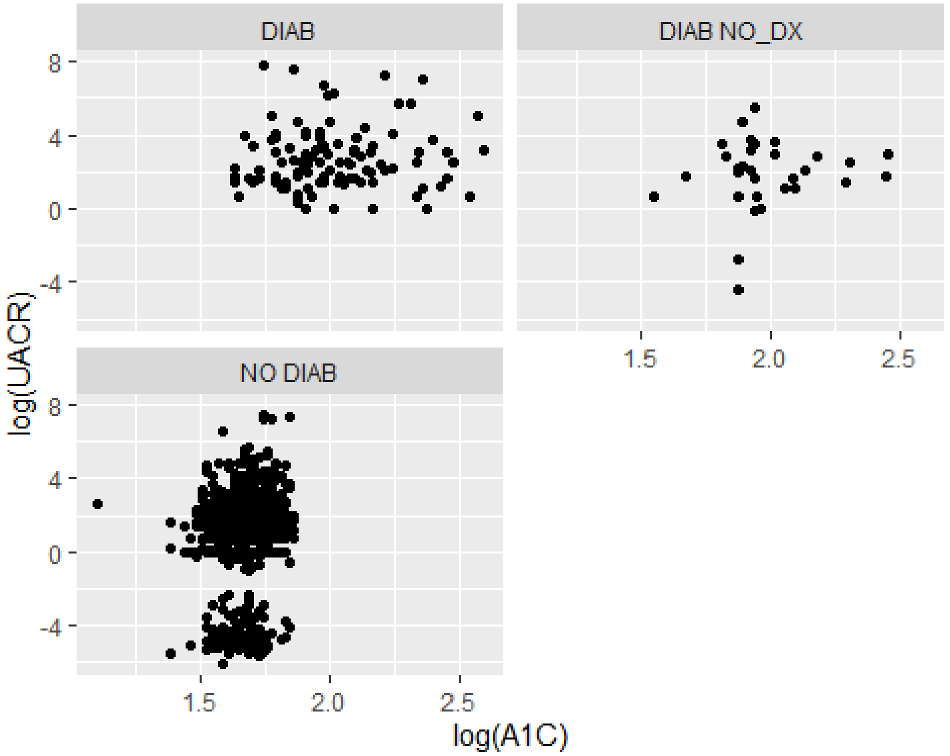


Now the shapes of the dots have altered and produce the same information as above. The above graph is provided to present an alternative viewing using visualization with shapes methods in R.

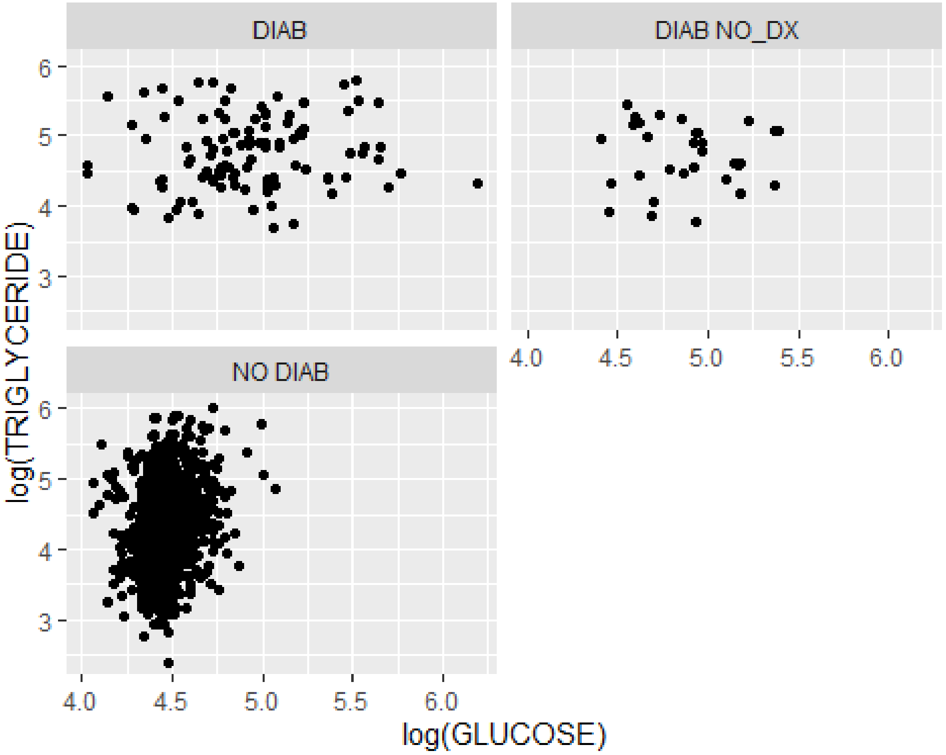
```r
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(A1C), y = log(UACR)), color="blue")
```



```r
ggplot(data=HANES) +
  geom_point(mapping = aes (x = log (A1C), y = log(UACR))) +
  facet_wrap(~ DX_DBTS, nrow = 2)
```

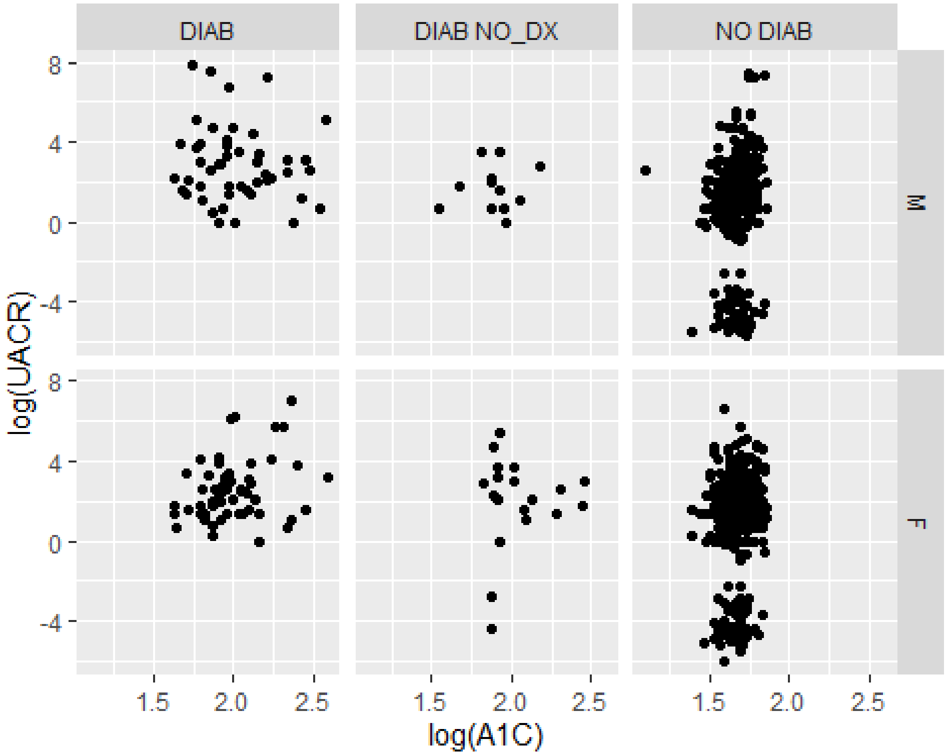


```
ggplot(data=HANES) +
  geom_point(mapping = aes (x = log (GLUCOSE), y = log(TRIGLYCERIDE))) +
  facet_wrap(~ DX_DBTS, nrow = 2)
```

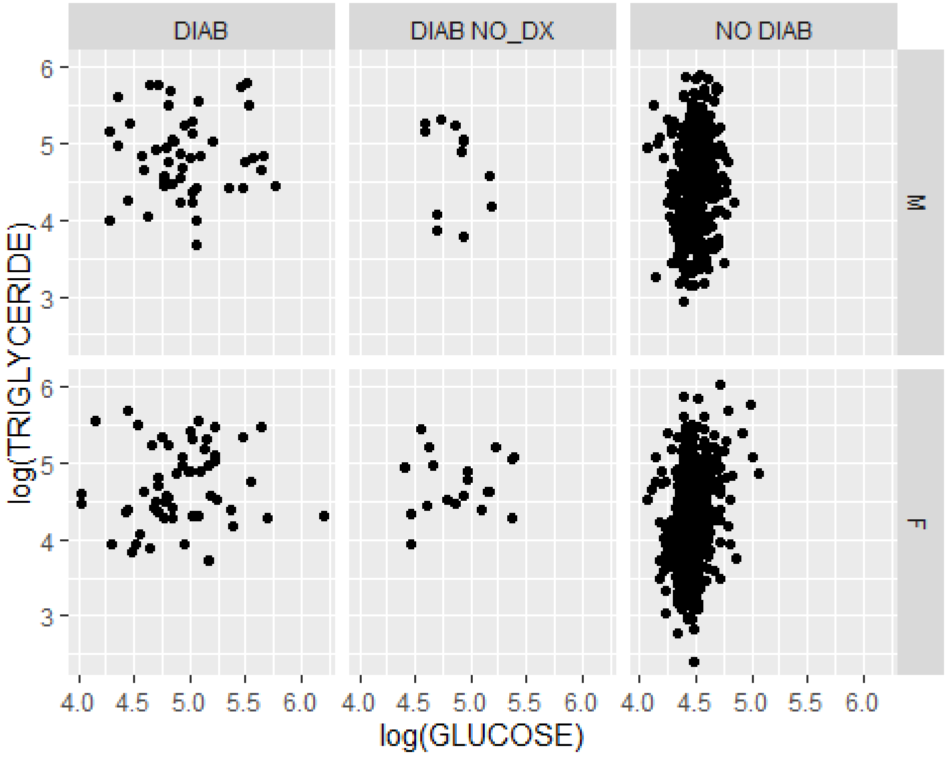


This visualization is great in enabling us to view DIAB, DIAB No_DX, and NO DIAB all based on log triglyceride (x-axis) and log glucose (y-axis).

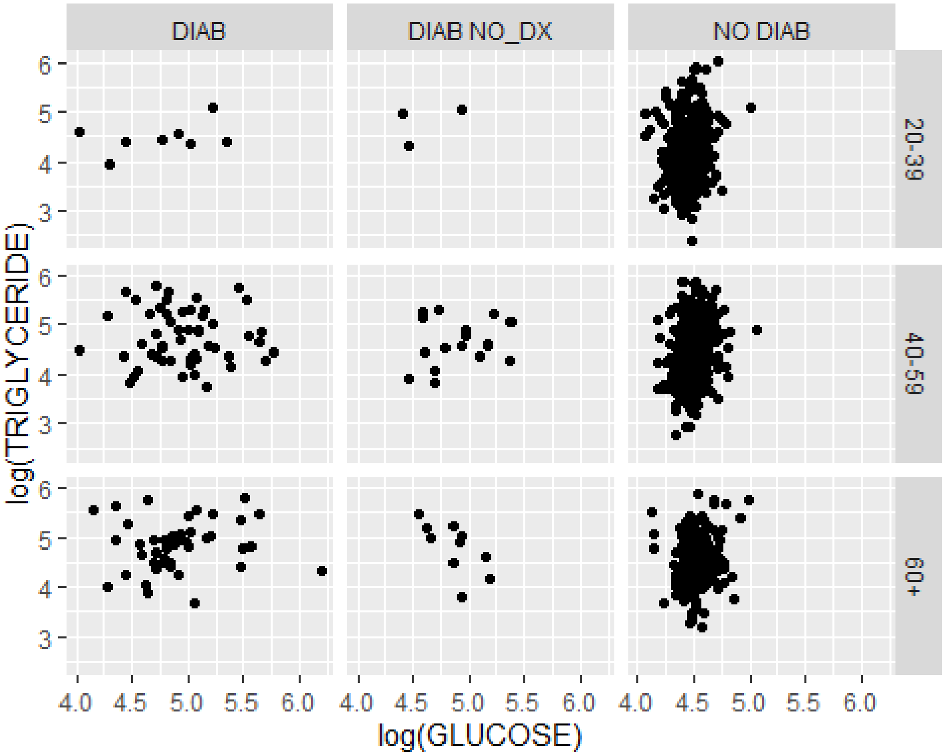
```
# Make a ggplot with facet grid - GENDER vs DX_DBTS
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(A1C), y = log(UACR))) +
  facet_grid(GENDER ~ DX_DBTS)
```



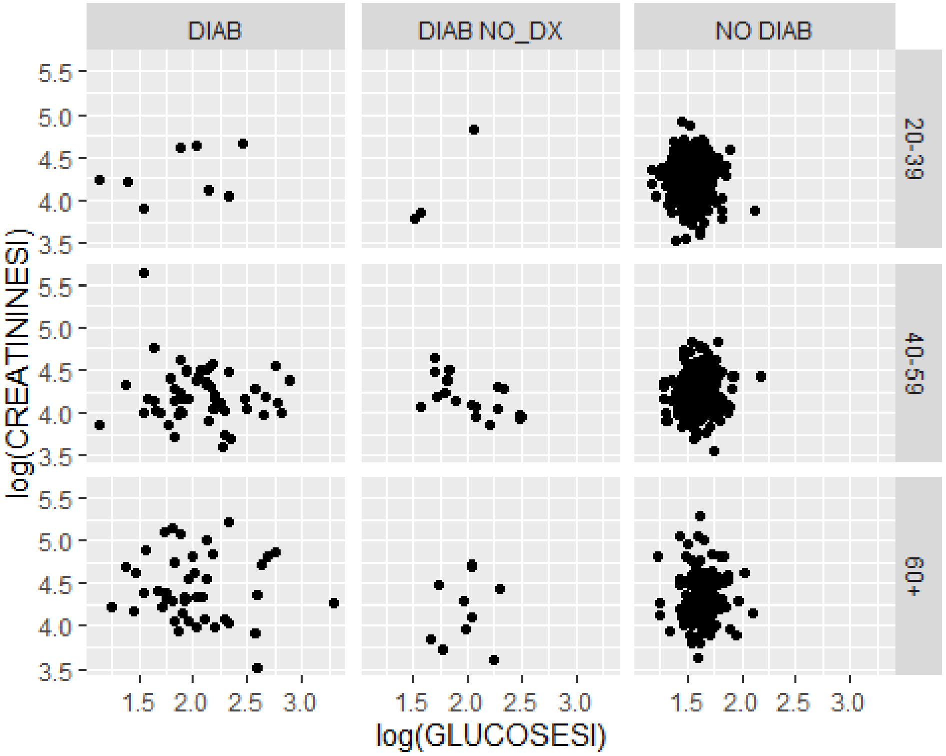
```r
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(GLUCOSE), y = log(TRIGLYCERIDE))) +
  facet_grid(GENDER ~ DX_DBTS)
```



```r
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(GLUCOSE), y = log(TRIGLYCERIDE))) +
  facet_grid(AGEGROUP ~ DX_DBTS)
```

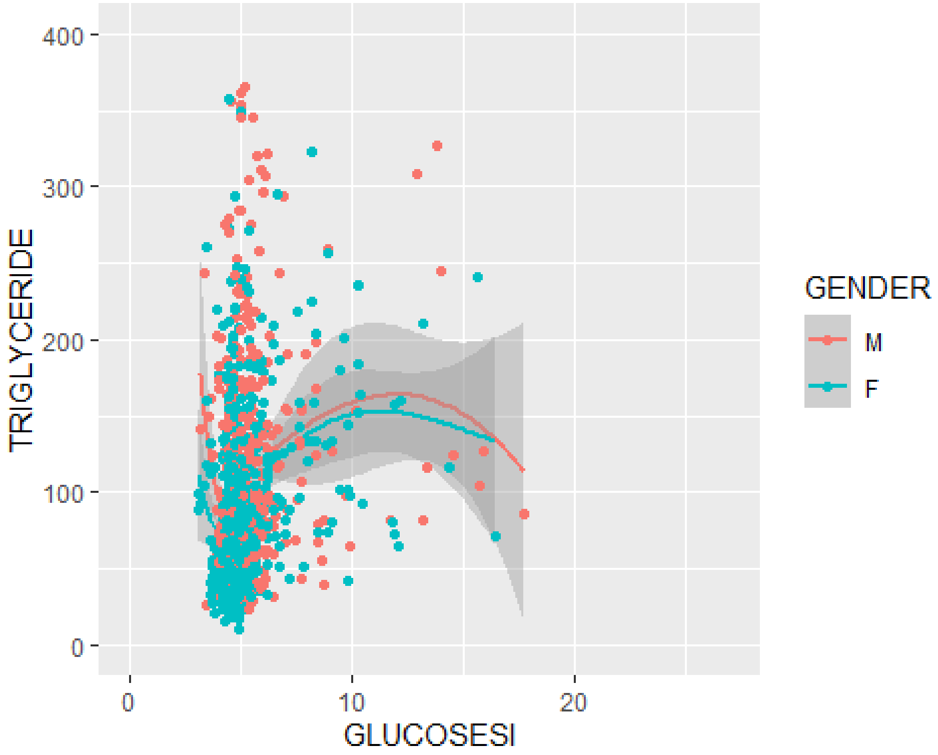


```r
ggplot(data = HANES) +
  geom_point(mapping = aes(x = log(GLUCOSESI), y = log(CREATININESI))) +
  facet_grid(AGEGROUP ~ DX_DBTS)
```



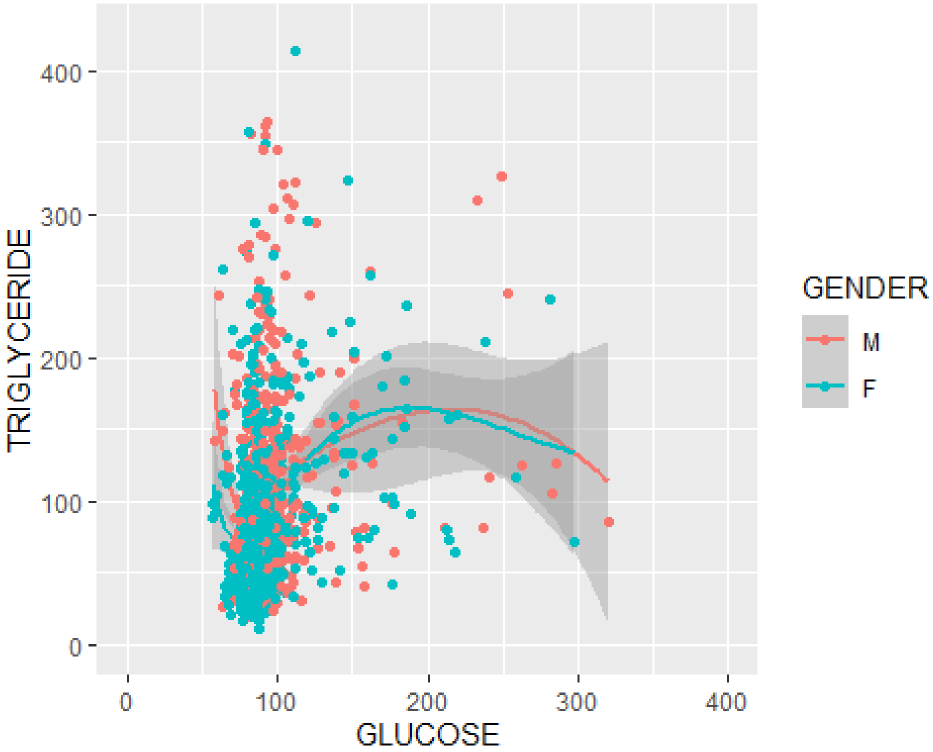
```r
ggplot(data = HANES) +
  geom_smooth(mapping = aes(x = GLUCOSESI, y = TRIGLYCERIDE, color=GENDER)) +
  xlim(c(0,27)) + ylim(c(0,400)) +
  geom_point(mapping = aes(x = GLUCOSESI, y = TRIGLYCERIDE, color=GENDER))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## Warning: Removed 2 rows containing missing values (geom_point).
```



Above are ggplot graphs with the facet grid for (1) Gender vs. DX_DBTS, (2) Gender vs DX_DBTS, and (3)Glucose and Triglyceride instead of A1c and uarc (age group vs. dx_dbts). There is also a plot of a smooth and point geom in the same plot. For the graph above I first graphed it with coordinates showing ylim(c(0,425)) for GlucoseSI and then altered it to c(0,100) to better view the graph. Once I adjusted for this, I saw that this graph is indeed an appropriate graph for gaining more knowledge concerning this material.

```
ggplot(data = HANES, mapping = aes(x = GLUCOSE, y = TRIGLYCERIDE, color=GENDER)) +
  xlim(c(0,400)) + ylim(c(0,425)) +
  geom_smooth() +
  geom_point()

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning: Removed 1 rows containing missing values (geom_point).
```



Above is the visualization of a smooth and point geom in the same plot. In the above graph, we can more clearly see where the congregation of points is located based on one's gender. We can see that most of the female dots are around the 80 to 100 mg/l glucose measurement and between the 20 and 100 Triglyceride mark. Men on the other hand show more dots extending past the 100 glucose mark. This may indicate that men - who have higher levels of glucose - may be more susceptible to diseases with high glucose levels as a risk factor.

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

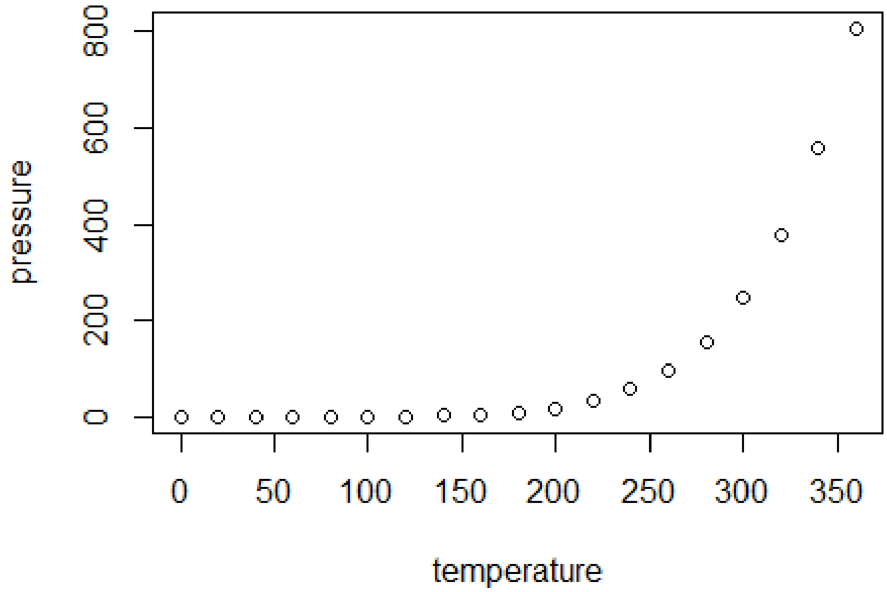
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.