Exploratory data analysis

- **Project goal:**
- The goal of this project was to clean, analyze and visualize data from the "*MEPS HC-228 2021 Full Year Population Characteristics*" dataset. The dataset can be found here. I am very passionate about healthcare and how demographic characteristics and socioeconomic status impact medical care and diagnoses. So, I searched a few recent journals and studies on google scholar and PubMed related to my topic of interest and explored their citations, eventually I stumbled upon this massive dataset that I thought would be perfect. The dataset was relatively clean but had many variables that were uniquely coded to identify the variables. I am experienced with survey data and survey codes, so I was able to interpret the data set using the codebook easily and explore the dataset. I liked this dataset for this reason, I spent more time analyzing trends than I did cleaning because I really wanted to explore the overwhelming number of variables in full. I also wanted to keep the integrity of the data as the codebook was attached to the dataset on the original site.
    - The dataset's documentation can be found here with more information about the validity and sample size for the population, along with key codes and other pertinent information regarding the data collected in Rounds 7, 8, and 9 of Panel 23; Rounds 5, 6, and 7 of Panel 24; Rounds 3, 4, and 5 of Panel 25; and Rounds 1, 2, and 3 of Panel 26 of the Medical Expenditure Panel Survey.
    - **Background on the dataset:** Released as an ASCII file (with related R, SAS, SPSS, and Stata programming statements and data user information) and a SAS dataset, SAS transport dataset, Stata dataset, and Excel file, this public use file provides information collected on a nationally representative sample of the civilian noninstitutionalized population of the United States for calendar year 2021. The file contains data collected in Rounds 7, 8, and 9 of Panel 23; Rounds 5, 6, and 7 of Panel 24; Rounds 3, 4, and 5 of Panel 25; and Rounds 1, 2, and 3 of Panel 26 of the Medical Expenditure Panel Survey (i.e., the rounds for the MEPS panels covering calendar year 2021). This year, 2021, is the first data year to include four panels of data; Panel 23 was extended to include Rounds 7, 8, and 9 and Panel 24 was extended to include Rounds 6, and 7. In addition, the Panel 24 Round 5 reference period was extended into 2021 instead of ending on 12/31/2020. This file contains variables pertaining to survey administration, demographics, person-level conditions, health status, disability days, quality of care, health care delays due to COVID-19, COVID-19 vaccinations, Social Determinants of Health (SDOH), employment, and health insurance. The 2021 full-year expenditure, medical care use counts, and income data will be forthcoming. For more details on loading MEPS PUFs into R, SAS, and Stata, please visit the MEPS GitHub page. This was a very large dataset with a lot of data that wasn't pertinent to my study question or study goals, so it was important to isolate the key variables that I needed and select those variables in R. All of the variables in the data set can be found here.
        - **The variables that I isolated and used for this project are:**
            - 263 265 BORNUSA PERSON BORN IN THE US
            - 359 360 DIABAGED AGE OF DIAGNOSIS-DIABETES
            - 357 358 DIABDX_M18 DIABETES DIAGNOSIS
            - 182 185 DOBYY DATE OF BIRTH: YEAR
            - 1 7 DUID PANEL # + ENCRYPTED DU IDENTIFIER
            - 11 20 DUPERSID PERSON ID (DUID + PID)
            - 228 230 EDUCYR YEARS OF EDUC WHEN FIRST ENTERED MEPS

        - 923 925 EVRWRK EVER WRKD FOR PAY IN LIFE AS OF 12/31/21
        - 231 233 HIDEG HIGHEST DEGREE WHEN FIRST ENTERED MEPS
        - 926 931 HRWG31X HOURLY WAGE RD 3/1 CMJ (IMP)
        - 932 937 HRWG42X HOURLY WAGE RD 4/2 CMJ (IMP)
        - 938 943 HRWG53X HOURLY WAGE RD 5/3 CMJ (IMP)
        - 2217 2228 PERWT21P USE FILE PERSON WEIGHT , 2021
        - 190 190 RACEAX ASIAN AMONG RACES RPTD (EDITED/IMPUTED)
        - 191 191 RACEBX BLACK AMONG RACES RPTD (EDITED/IMPUTED)
        - 193 193 RACETHX RACE/ETHNICITY (EDITED/IMPUTED)
        - 187 187 RACEV1X RACE (EDITED/IMPUTED)
        - 188 189 RACEV2X RACE (EDITED/IMPUTED)
        - 192 192 RACEWX WHITE AMONG RACES RPTD (EDITED/IMPUTED)
- Study goals:
    - Identify if there is a correlation between demographic characteristics (such as, age, sex, race, ethnicity, employment status, education, country of birth), and diabetes. The goal is to investigate which demographic factors (mainly race), if any, impact diabetes occurrence rates.
        - So, to summarize:
            - Determine if there is a correlation between race and diabetes.
                - Hypothesis: There is a positive correlation between certain racial groups and the prevalence of diabetes.
            - Determine what role key demographic variables have on instances of diabetes.
                - Hypothesis: The likelihood of developing diabetes is influenced by a combination of factors, including education level, employment status, race, and country of origin.
        - I will then visualize any findings in 4-5 charts and summarize the findings.
    - Questions that I will also keep in mind:
        - What type of variation occurs within my variables?
        - What type of covariation occurs between my variables?
        - Although I have key study goals in mind, I still want to explore each variable and may generate new questions as I investigate every idea that arises throughout the project.
- **Resume skills practiced:** R, data cleaning, data visualization
- **Packages used:** dplyr, tidyverse, ggplot2, psych, sqldf, utils

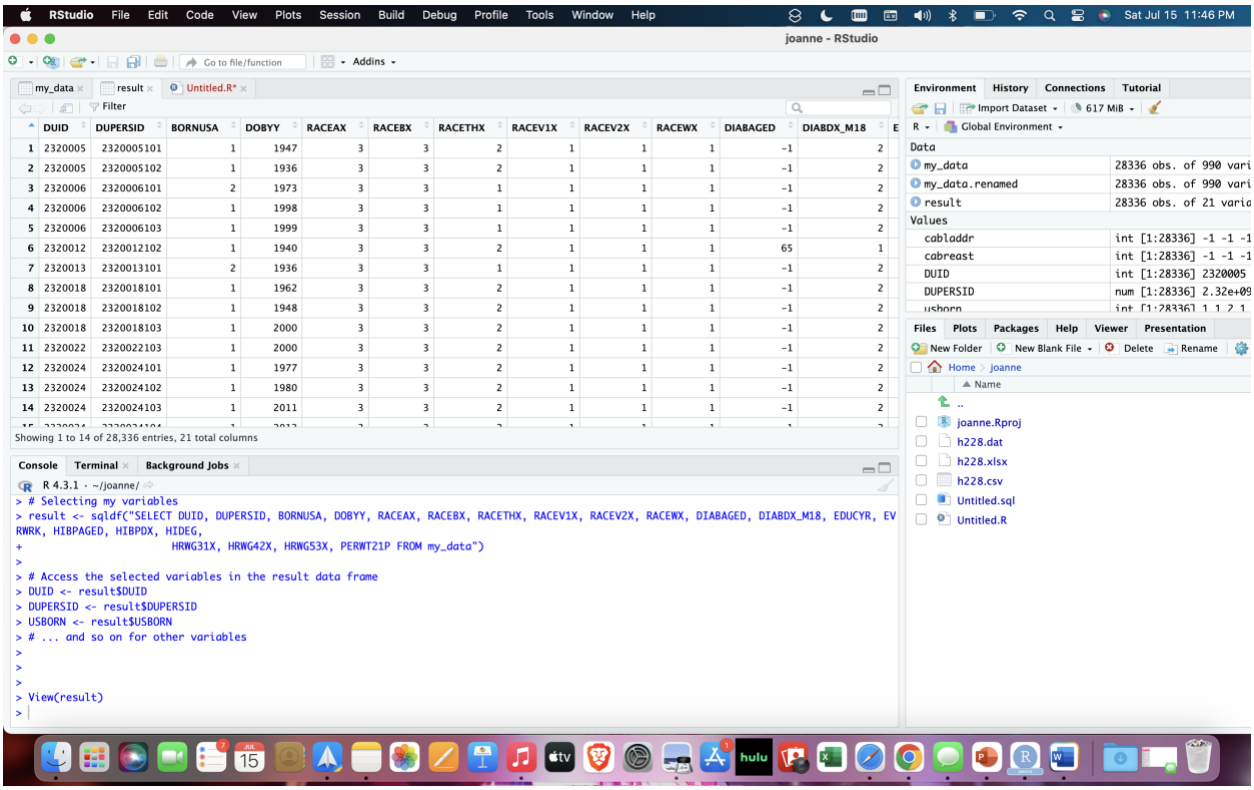
R cleaning and analyzing steps:

1. Selecting the following variables to the following, also referred to this sheet to help identify the variables:

```
library(sqldf)

# Selecting my variables
result <- sqldf("SELECT DUID, DUPERSID, BORNUSA, DOBYY, RACEAX, RACEBX, RACETHX, RACEV1X,
RACEV2X, RACEWX, DIABAGED, DIABDX_M18, EDUCYR, EVRWRK, HIBPAGED, HIBPDX, HIDEG,
                HRWG31X, HRWG42X, HRWG53X, PERWT21P FROM my_data")
```

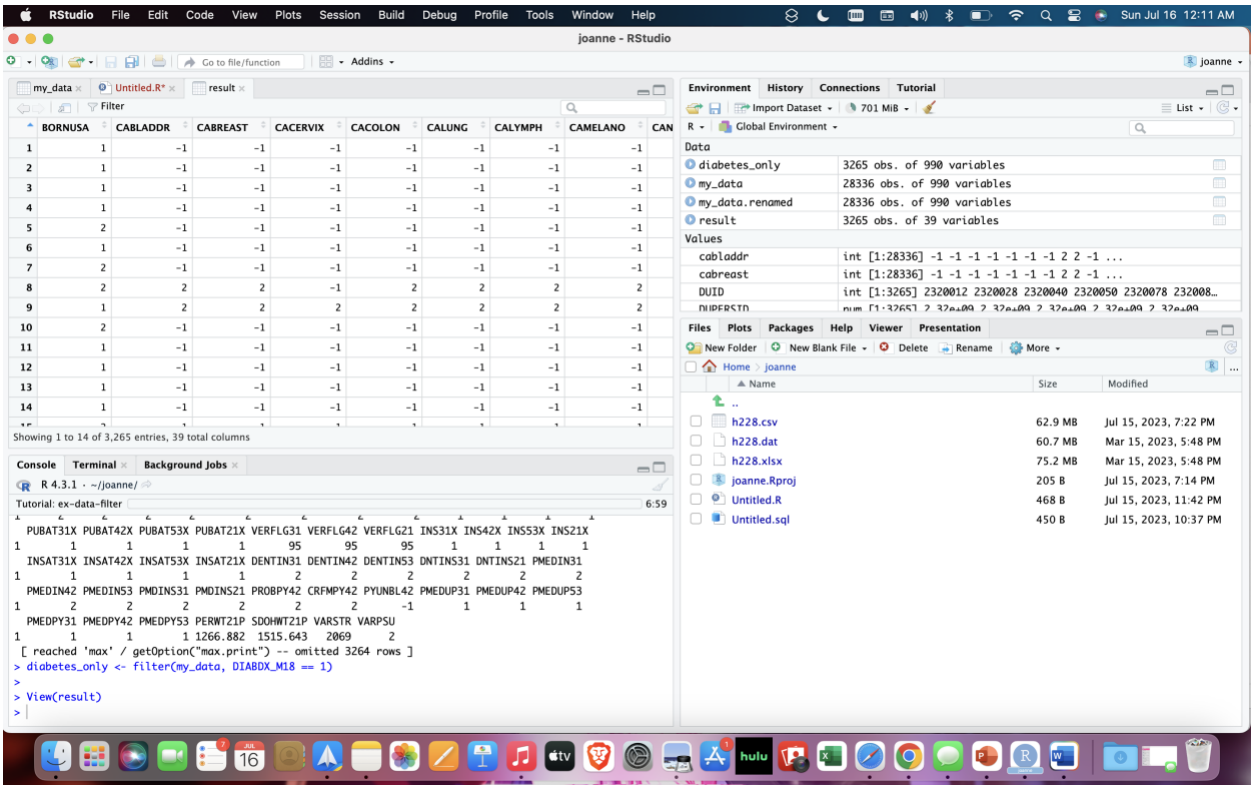
```
# Access the selected variables in the result data frame
DUID <- result$DUID
DUPERSID <- result$DUPERSID
# ... and so on for other variables
```

Result:

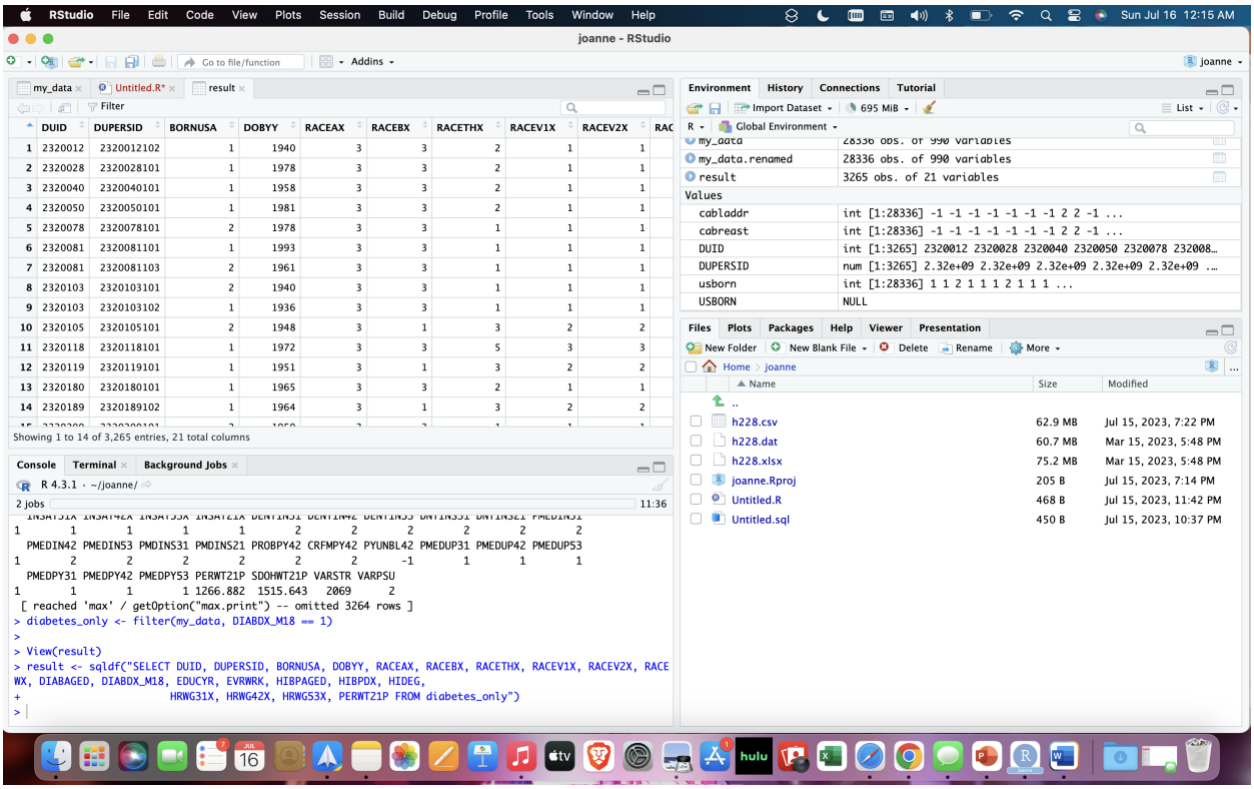


The result was 21 columns and 28,336 entries/rows.
Next, I selected all of the individuals in the data set with diabetes, this is denoted by a 1 in that column so I will pull all of the rows where DIABDX_M18 is 1.

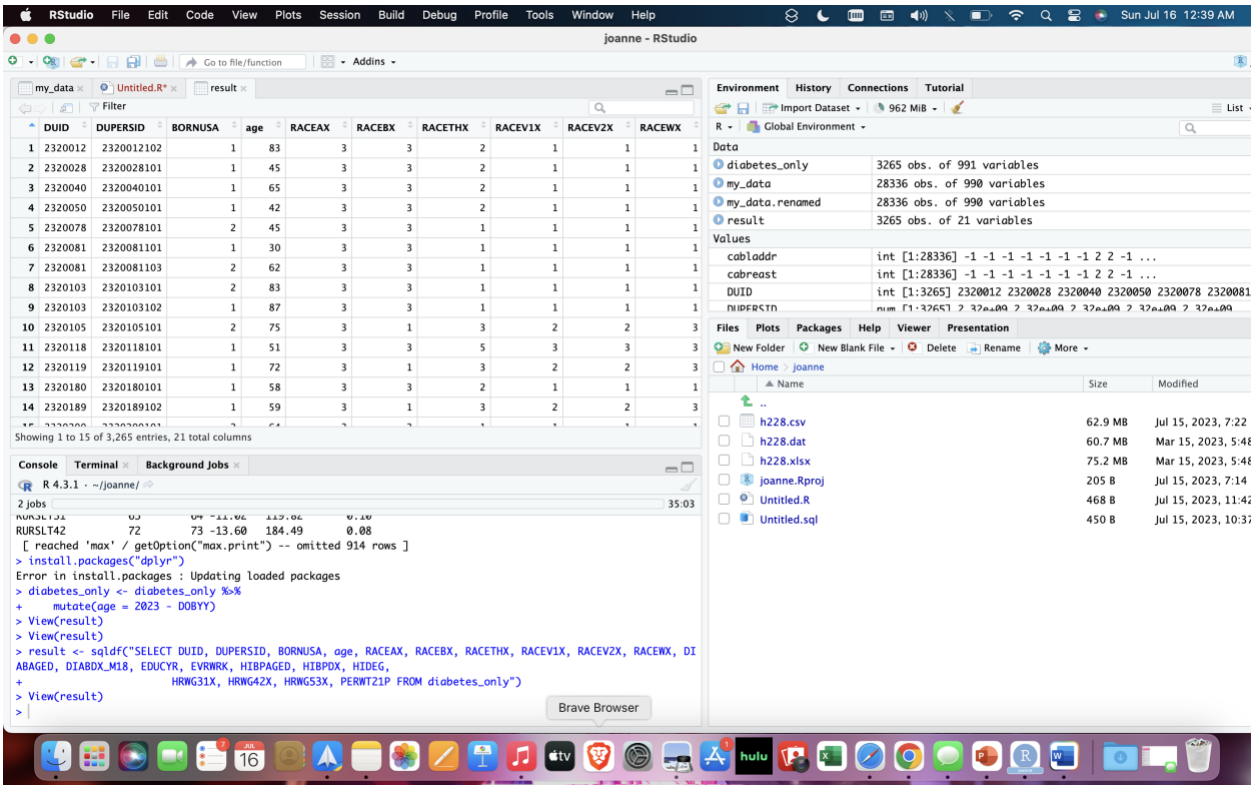
```
# filtering rows where DIABDX_M18 is 1
diabetes_only <- filter(my_data, DIABDX_M18 == 1)
```




```
# Selecting my variables from the new filtered result
result <- sqldf("SELECT DUID, DUPERSID, BORNUSA, DOBYY, RACEAX, RACEBX, RACETHX,
RACEV1X, RACEV2X, RACEWX, DIABAGED, DIABDX_M18, EDUCYR, EVRWRK, HIBPAGED, HIBPDX,
HIDEG,
                HRWG31X, HRWG42X, HRWG53X, PERWT21P FROM diabetes_only")
```

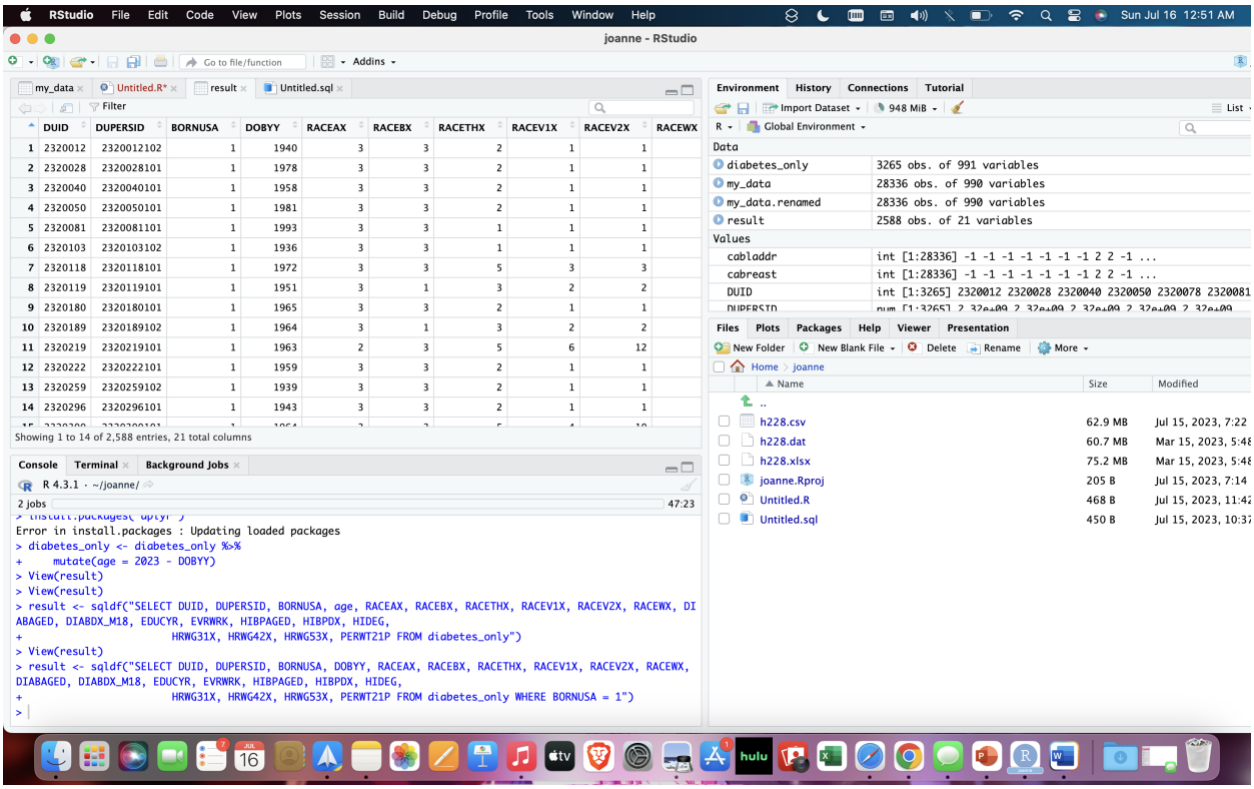



```
# Calculating age using year of birth - current year (2023)
diabetes_only <- diabetes_only %>%
  +     mutate(age = 2023 - DOBYY)
View(result)

result <- sqldf("SELECT DUID, DUPERSID, BORNUSA, age, RACEAX, RACEBX, RACETHX, RACEV1X,
RACEV2X, RACEWX, DIABAGED, DIABDX_M18, EDUCYR, EVRWRK, HIBPAGED, HIBPDX, HIDEG,
                HRWG31X, HRWG42X, HRWG53X, PERWT21P FROM diabetes_only")
```



Next, I wanted to see if there was any interesting data regarding country of birth and presence of diabetes, so I will look at the BORNUSA column and count how many people in the survey with diabetes were born in the US compared to those who weren't. 1=YES, 2=NO. Let's count how many "yes's" there were.

View(result)
result <- sqldf("SELECT DUID, DUPERSID, BORNUSA, DOBYY, RACEAX, RACEBX, RACETHX, RACEV1X, RACEV2X, RACEWX, DIABAGED, DIABDX_M18, EDUCYR, EVRWRK, HIBPAGED, HIBPDX, HIDEG,
+ HRWG31X, HRWG42X, HRWG53X, PERWT21P FROM diabetes_only WHERE BORNUSA = 1")

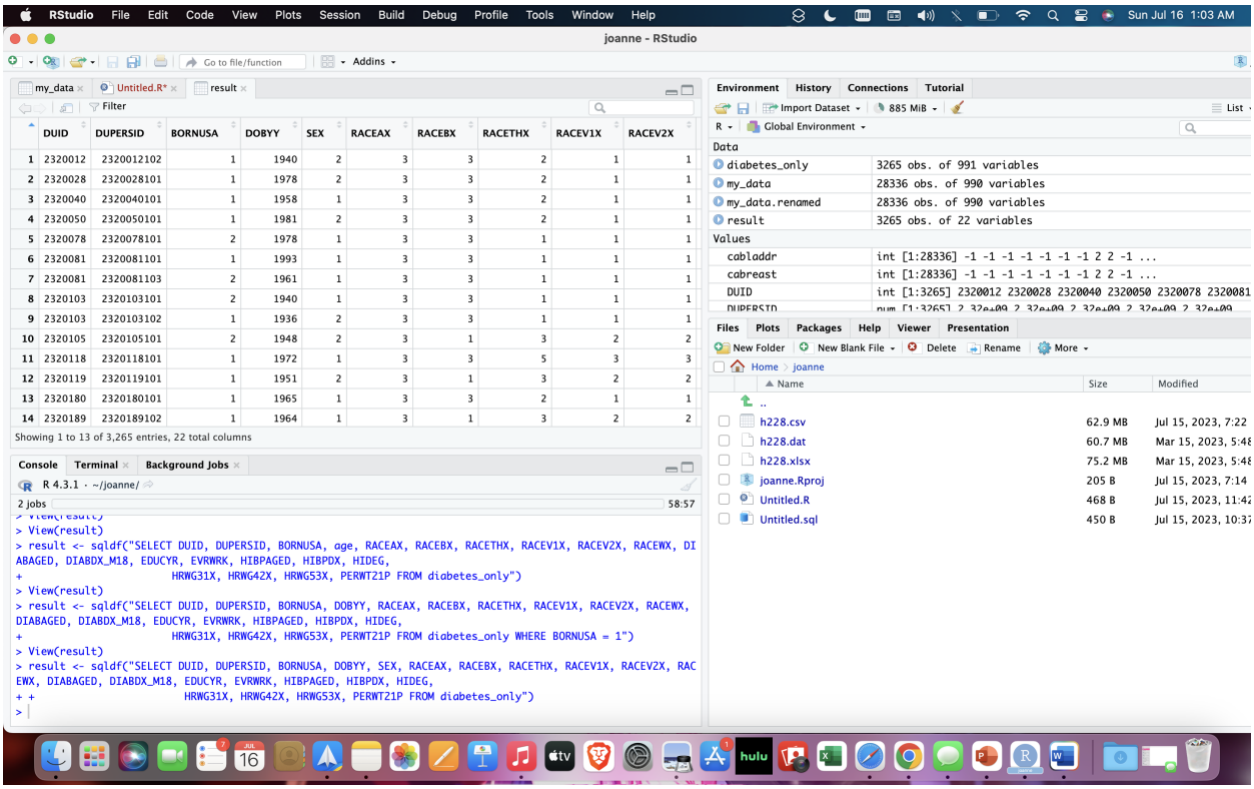


Let's summarize our data so far:

- Out of our original data set of 28336 rows:
  - 3,265 people reported having diabetes.
    - 2,588 of that 3,265 people were born in the USA – That is nearly 80% (79.26% to be exact). So, one trend was identified thus far. US born individuals have a higher chance of being diagnosed with diabetes.

Originally, I didn't extract the sex variable to look at gender differences, I'll do that now.
The SEX variable is coded as the following: 1=MALE, 2=FEMALE

View(result)
result <- sqldf("SELECT DUID, DUPERSID, BORNUSA, DOBYY, SEX, RACEAX, RACEBX, RACETHX, RACEV1X, RACEV2X, RACEWX, DIABAGED, DIABDX_M18, EDUCYR, EVRWRK, HIBPAGED, HIBPDX, HIDEG,
+ HRWG31X, HRWG42X, HRWG53X, PERWT21P FROM diabetes_only")



#Let's see how many men in the sample reported having diabetes.
View(result)
result <- sqldf("SELECT DUID, DUPERSID, BORNUSA, DOBYY, SEX, RACEAX, RACEBX, RACETHX, RACEV1X, RACEV2X, RACEWX, DIABAGED, DIABDX_M18, EDUCYR, EVRWRK, HIBPAGED, HIBPDX, HIDEG,
+ + + HRWG31X, HRWG42X, HRWG53X, PERWT21P FROM diabetes_only WHERE SEX = 1")
>

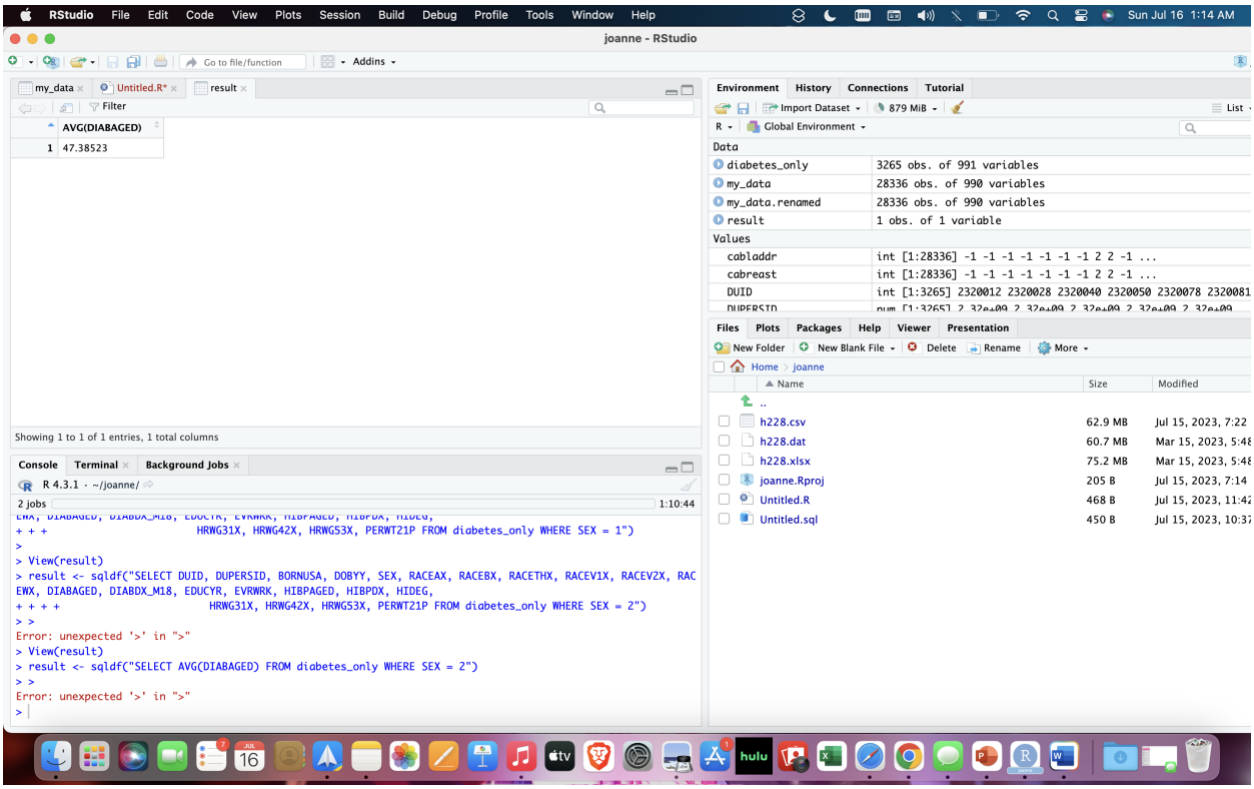
1,518 men reported having diabetes compared to 1,747 women.

# Let's look at the women but let's also pull the average diabetes diagnosis age for men vs. women.
View(result)
result <- sqldf("SELECT AVG(DIABAGED) FROM diabetes_only WHERE SEX = 2")

- Women Average Diabetes Diagnosis Age: 47 (47.38)

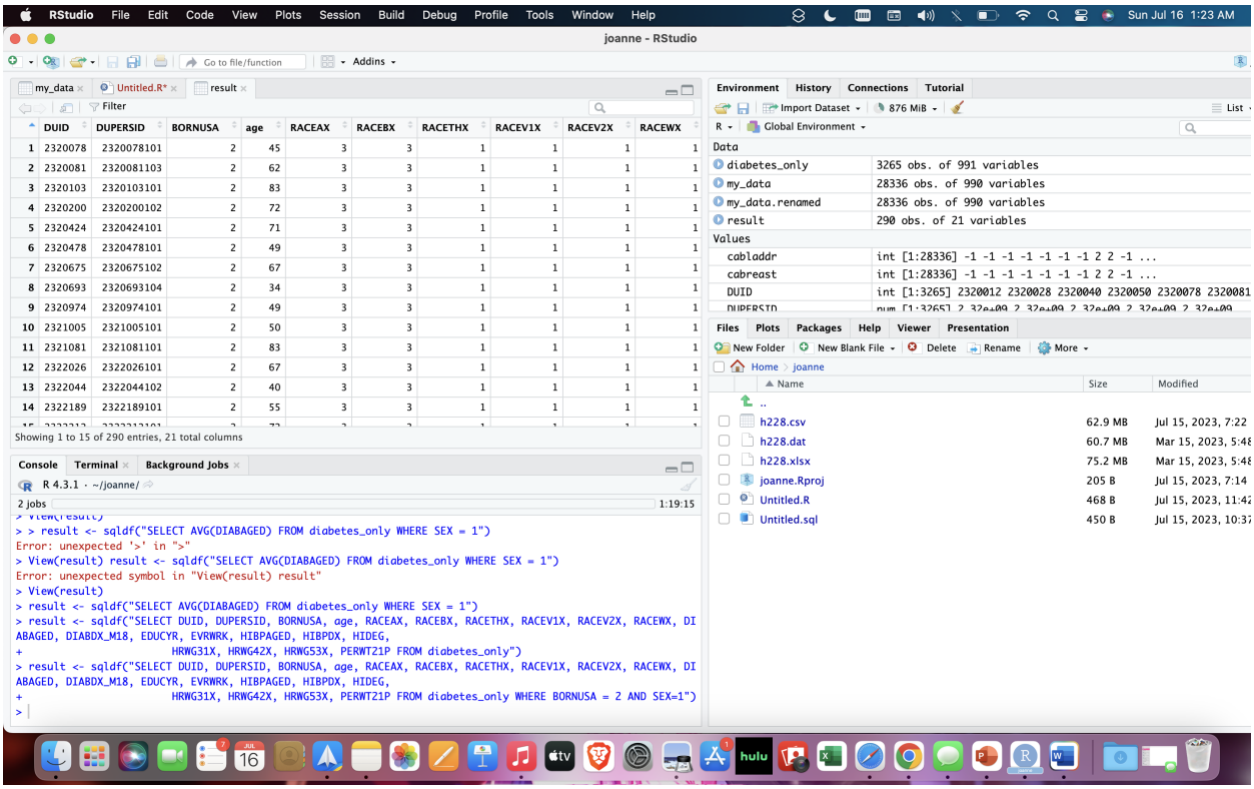
View(result)
result <- sqldf("SELECT AVG(DIABAGED) FROM diabetes_only WHERE SEX = 1")

- Men Average Diabetes Diagnosis Age: 48 (48.07)



#Let's look at differences between both gender and country of birth now.

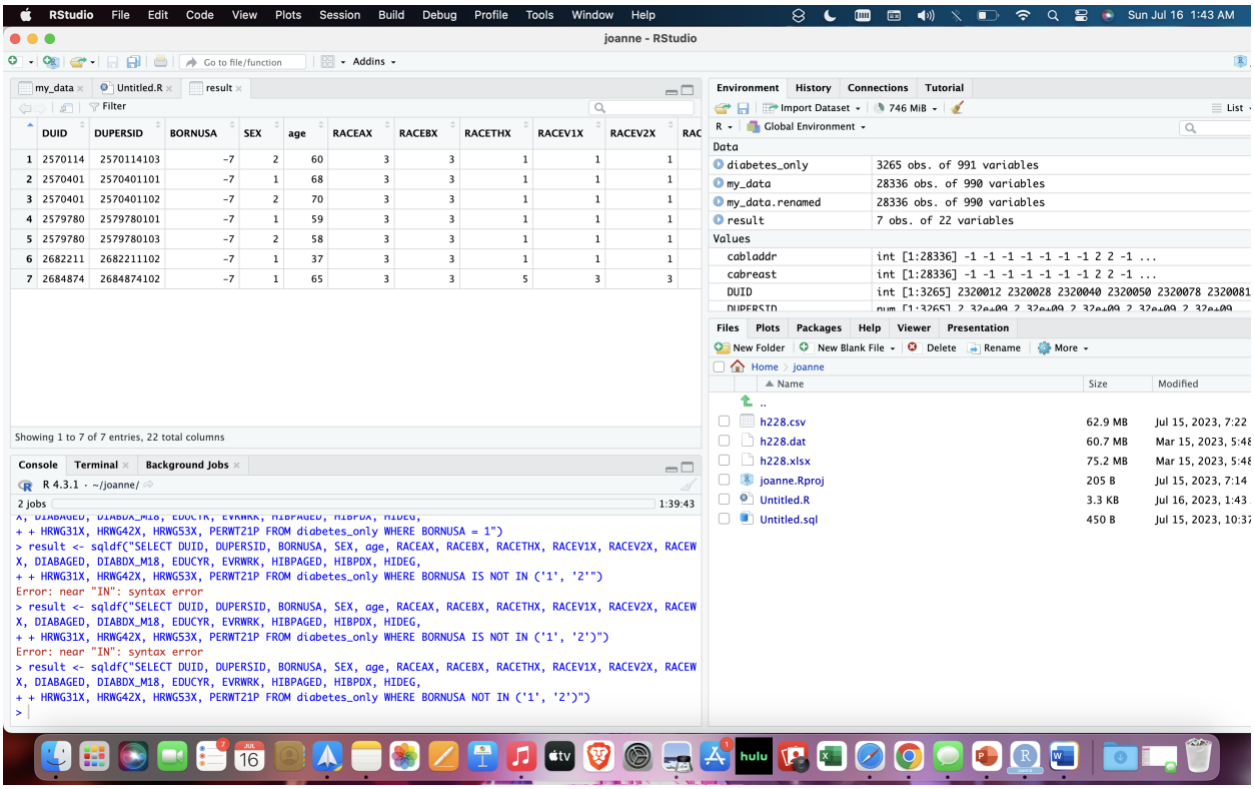
result <- sqldf("SELECT DUID, DUPERSID, BORNUSA, age, RACEAX, RACEBX, RACETHX, RACEV1X, RACEV2X, RACEWX, DIABAGED, DIABDX_M18, EDUCYR, EVRWRK, HIBPAGED, HIBPDX, HIDEG,
+ HRWG31X, HRWG42X, HRWG53X, PERWT21P FROM diabetes_only WHERE BORNUSA = 2 AND SEX=1")
>



290 men who were born in another country other than the US reported having diabetes while, 380 women not born in the US reported having diabetes.

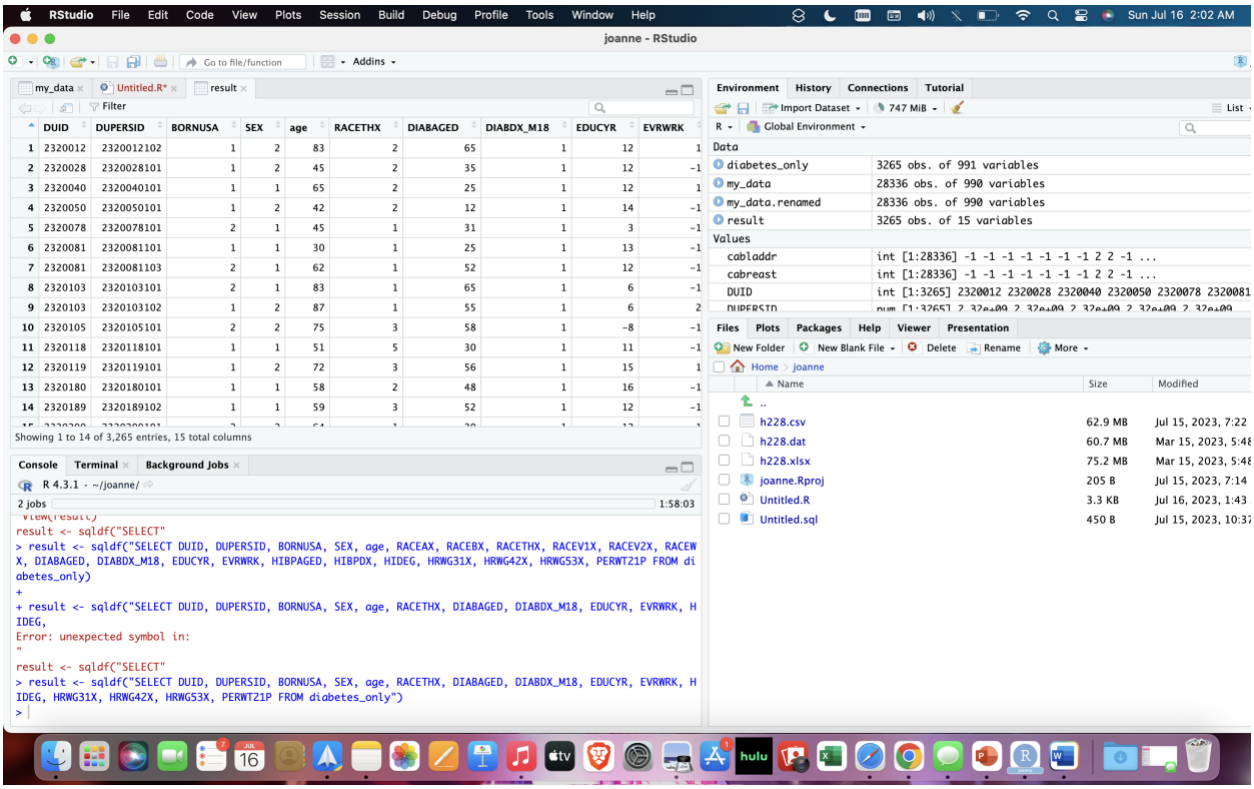
I realized 7 people were not accounted for in the sample and searched other codes for BORNUSA to see why.

result <- sqldf("SELECT DUID, DUPERSID, BORNUSA, SEX, age, RACEAX, RACEBX, RACETHX, RACEV1X, RACEV2X, RACEWX, DIABAGED, DIABDX_M18, EDUCYR, EVRWRK, HIBPAGED, HIBPDX, HIDEG,
+ HRWG31X, HRWG42X, HRWG53X, PERWT21P FROM diabetes_only WHERE BORNUSA NOT IN ('1', '2')")



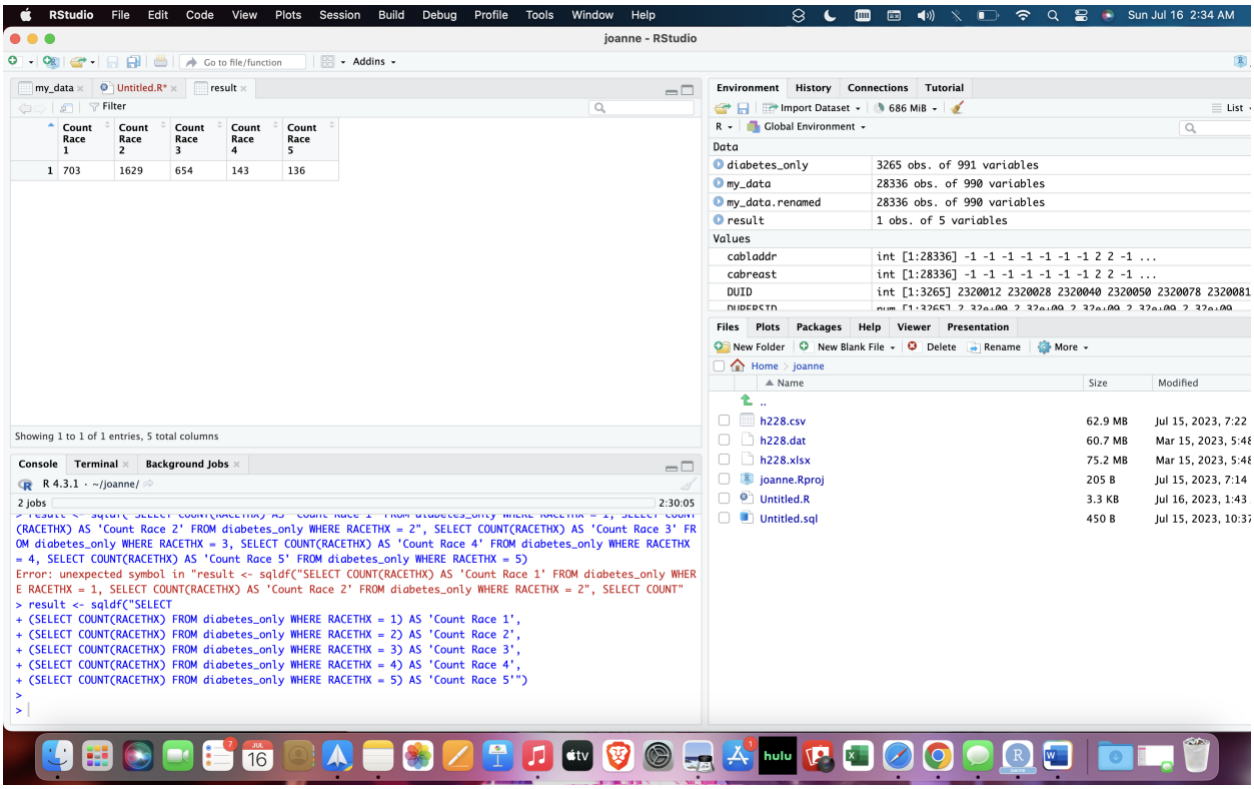
7 patients were coded as -7 for the BORNUSA variable, which means that they refused to answer the survey question.

#Let's select only the key columns now and start to explore race.
result <- sqldf("SELECT DUID, DUPERSID, BORNUSA, SEX, age, RACETHX, DIABAGED, DIABDX_M18, EDUCYR, EVRWRK, HIDEG, HRWG31X, HRWG42X, HRWG53X, PERWT21P FROM diabetes_only")



#Let's select only the key columns now and start to explore race.
result <- sqldf("SELECT DUID, DUPERSID, BORNUSA, SEX, age, RACETHX, DIABAGED, DIABDX_M18, EDUCYR, EVRWRK, HIDEG, HRWG31X, HRWG42X, HRWG53X, PERWT21P FROM diabetes_only")

#Let's start to prepare our dataset to count how many times each race was reported in our diabetes database.
# Count the occurrences of each unique value in the 'RACETHX' column
result <- sqldf("SELECT
(SELECT COUNT(RACETHX) FROM diabetes_only WHERE RACETHX = 1) AS 'Count Race 1',
(SELECT COUNT(RACETHX) FROM diabetes_only WHERE RACETHX = 2) AS 'Count Race 2',
(SELECT COUNT(RACETHX) FROM diabetes_only WHERE RACETHX = 3) AS 'Count Race 3',
(SELECT COUNT(RACETHX) FROM diabetes_only WHERE RACETHX = 4) AS 'Count Race 4',
(SELECT COUNT(RACETHX) FROM diabetes_only WHERE RACETHX = 5) AS 'Count Race 5'")

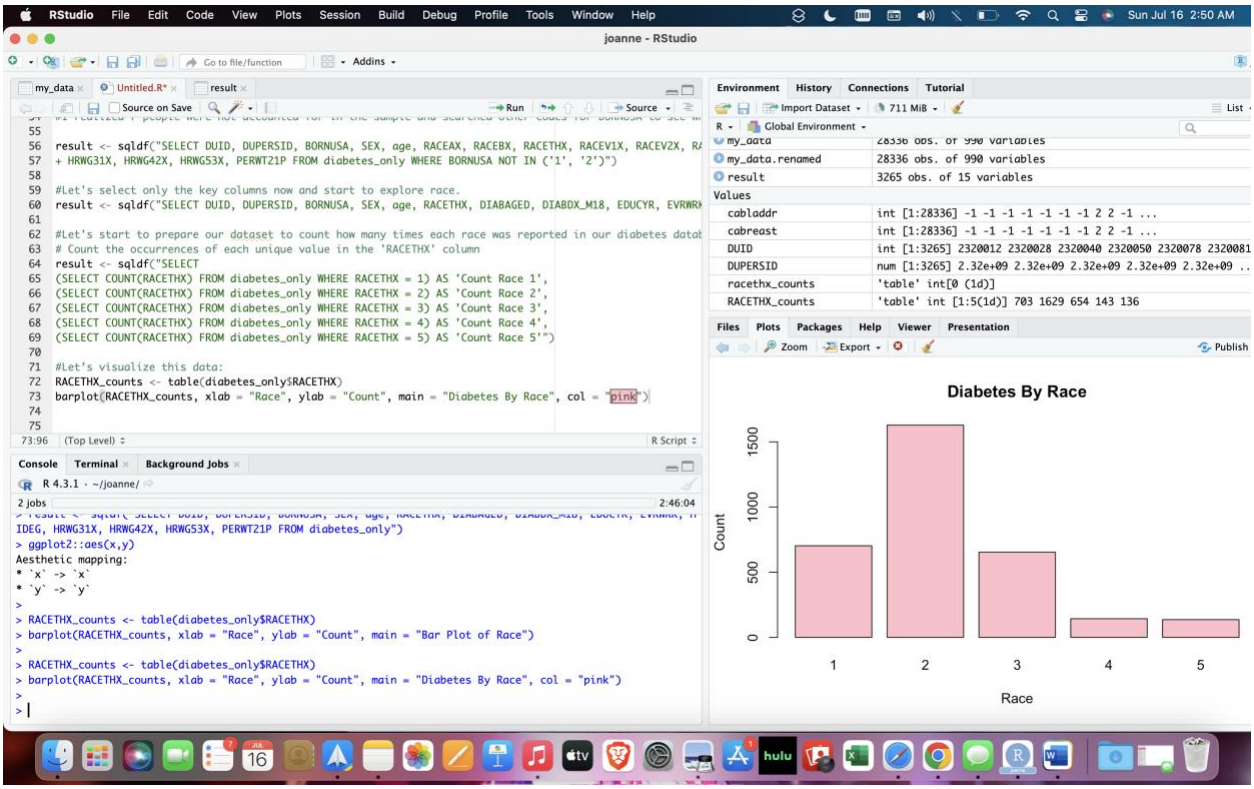


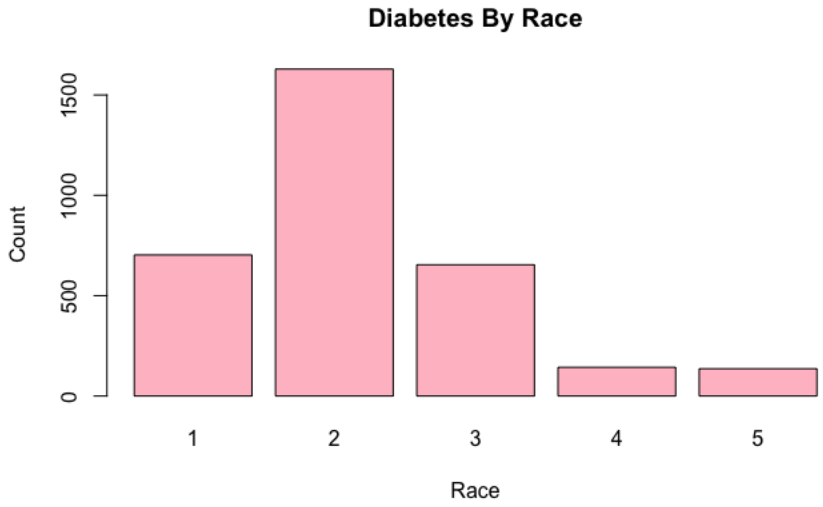
Let's summarize our race counts:

- Race 1 (Hispanic): 703
- Race 2 (Non-Hispanic White Only): 1629
- Race 3 (Non-Hispanic Black Only): 654
- Race 4 (Non-Hispanic Asian Only): 143
- Race 5 (Non-Hispanic Other Race or Multiple Race): 136

Let's visualize this data:

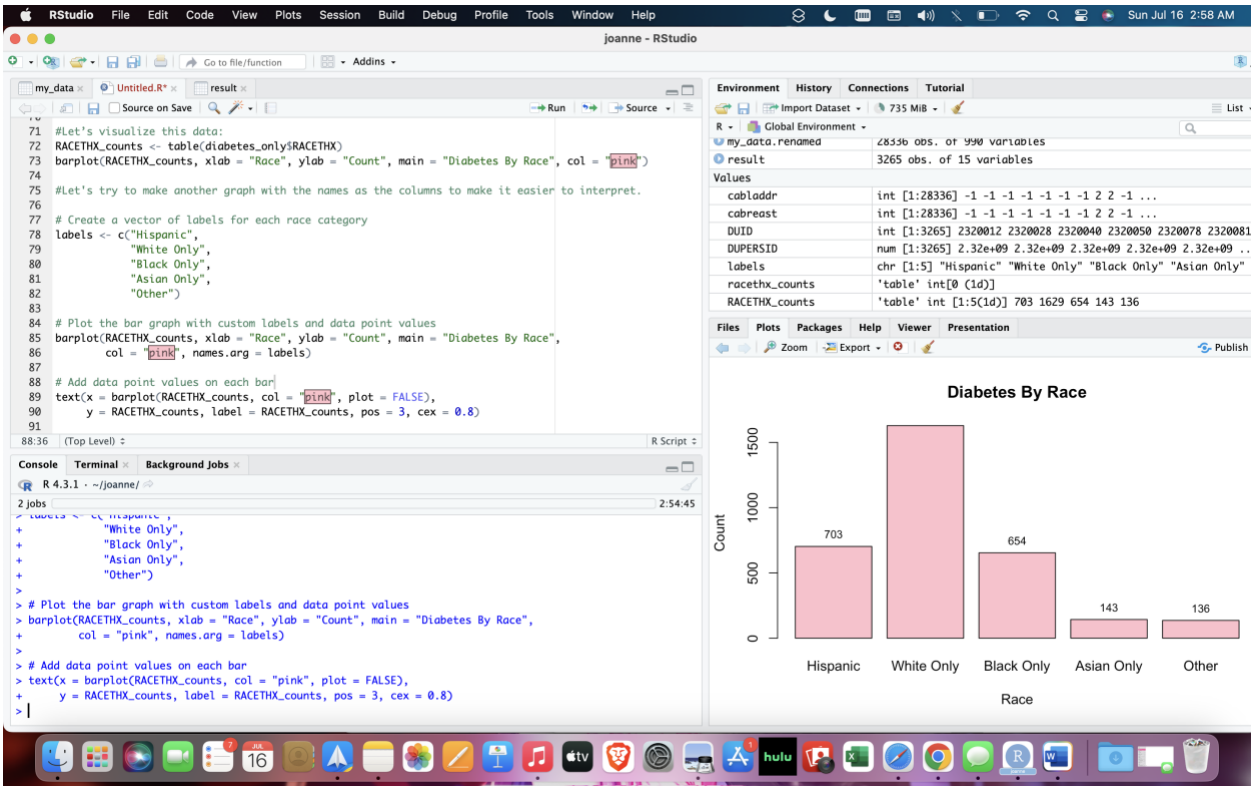
```
RACETHX_counts <- table(diabetes_only$RACETHX)
barplot(RACETHX_counts, xlab = "Race", ylab = "Count", main = "Diabetes By Race", col = "pink")
```

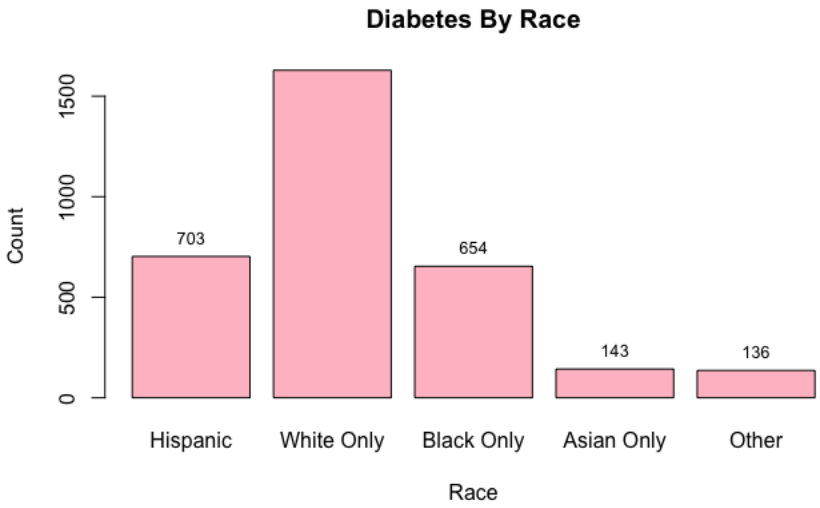





```
# Let's try to make another graph with the names as the columns to make it easier to interpret.
# Create a vector of labels for each race category
labels <- c("Hispanic",
            "White Only",
            "Black Only",
            "Asian Only",
            "Other")

# Plot the bar graph with custom labels and data point values
```

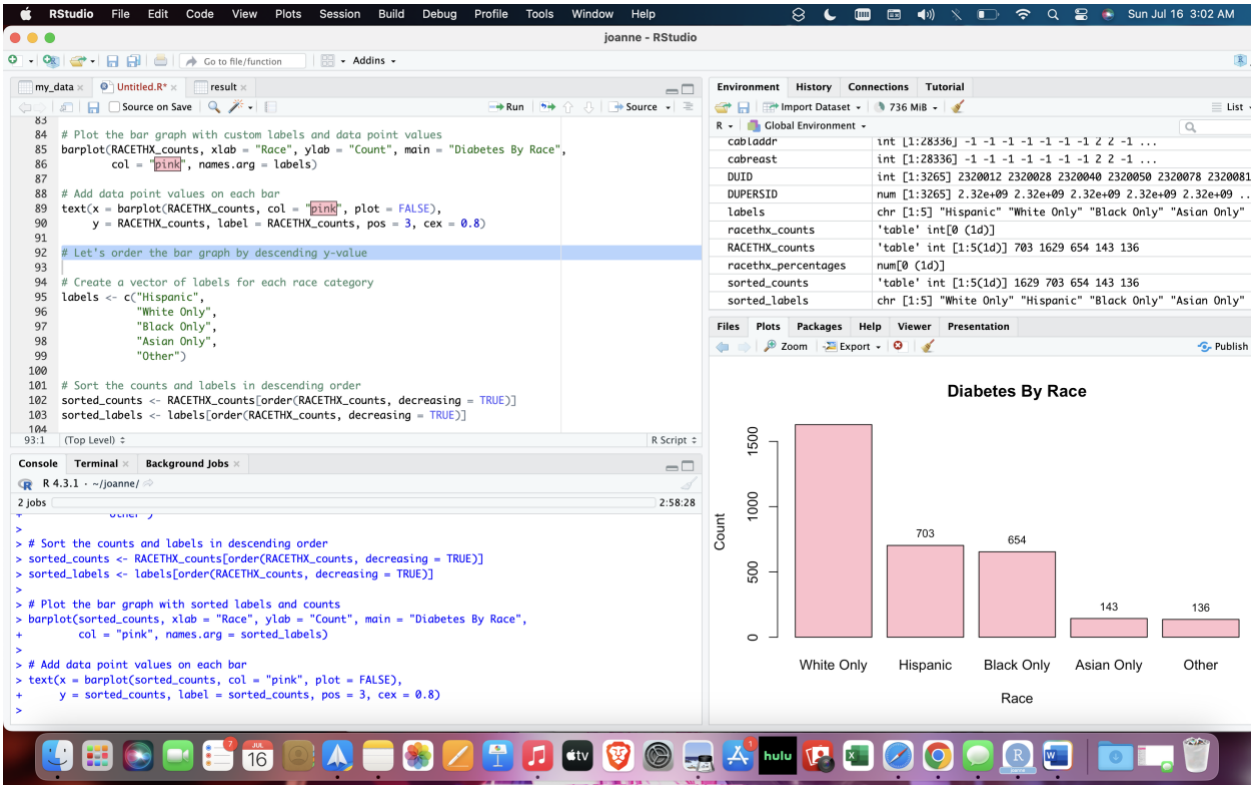
```
barplot(RACETHX_counts, xlab = "Race", ylab = "Count", main = "Diabetes By Race",
        col = "pink", names.arg = labels)

# Add data point values on each bar
text(x = barplot(RACETHX_counts, col = "pink", plot = FALSE),
     y = RACETHX_counts, label = RACETHX_counts, pos = 3, cex = 0.8)
```

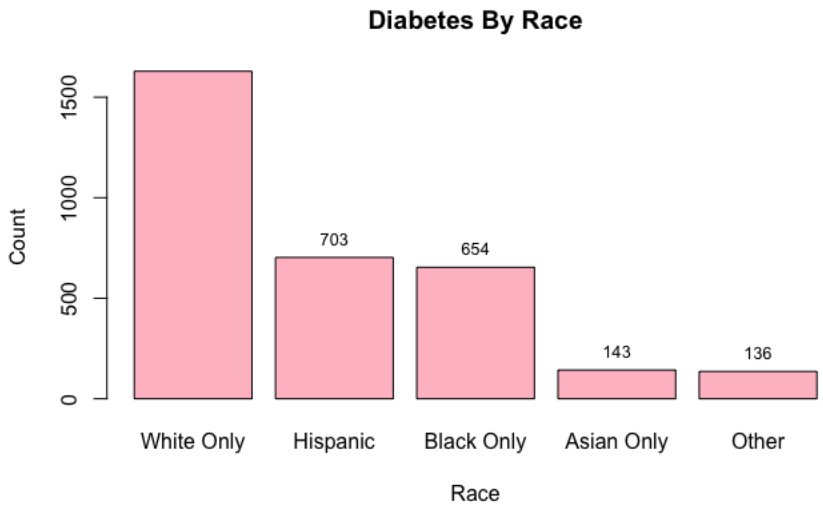




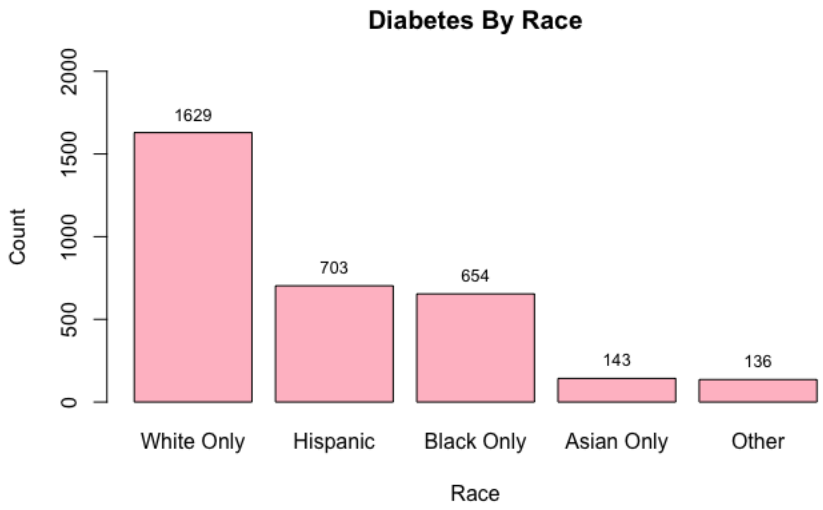

```
# Let's order the bar graph by descending y-value
```

```r
# Create a vector of labels for each race category
labels <- c("Hispanic",
            "White Only",
            "Black Only",
            "Asian Only",
            "Other")

# Sort the counts and labels in descending order
sorted_counts <- RACETHX_counts[order(RACETHX_counts, decreasing = TRUE)]
sorted_labels <- labels[order(RACETHX_counts, decreasing = TRUE)]

# Plot the bar graph with sorted labels and counts
barplot(sorted_counts, xlab = "Race", ylab = "Count", main = "Diabetes By Race",
        col = "pink", names.arg = sorted_labels)

# Add data point values on each bar
text(x = barplot(sorted_counts, col = "pink", plot = FALSE),
     y = sorted_counts, label = sorted_counts, pos = 3, cex = 0.8)
```



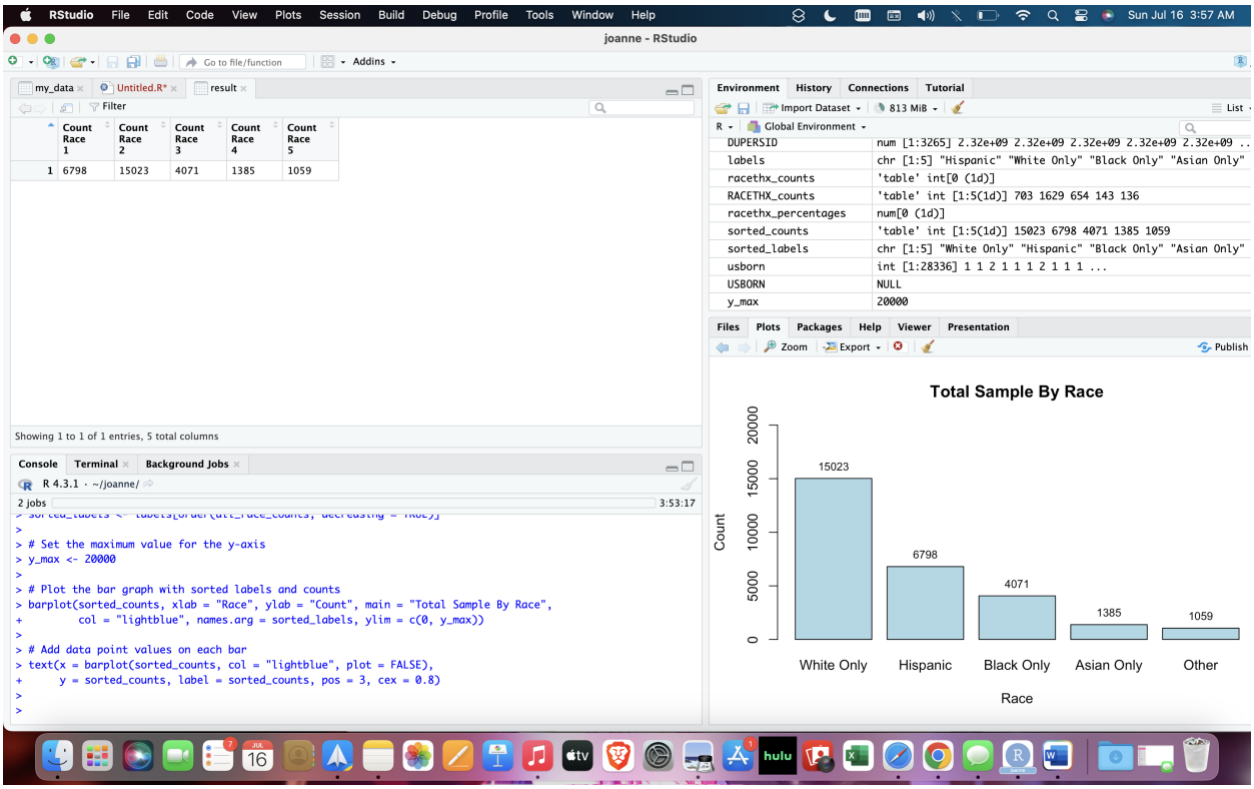


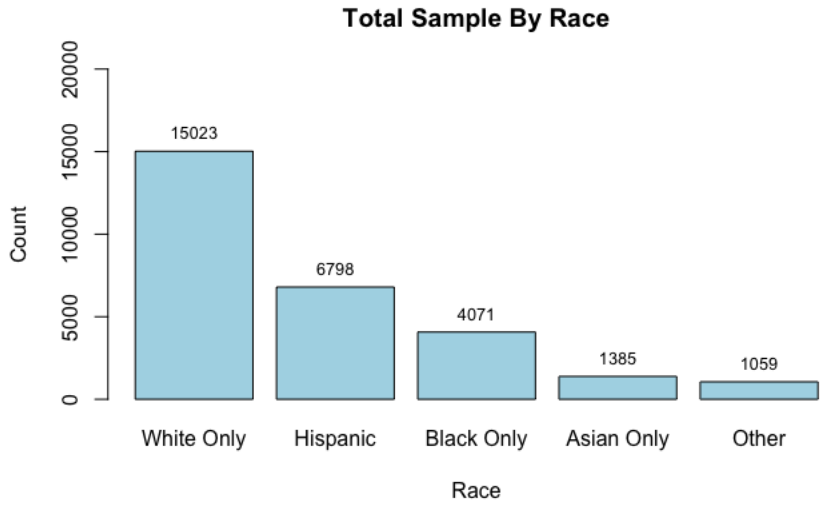

```
# The y-axis scale and maximum value is making it hard to interpret the "white only" column. Let's
update our query to fix this by setting the y-axis's maximum value to 2000.

# Create a vector of labels for each race category
labels <- c("Hispanic",
            "White Only",
            "Black Only",
            "Asian Only",
            "Other")

# Sort the counts and labels in descending order
sorted_counts <- RACETHX_counts[order(RACETHX_counts, decreasing = TRUE)]
sorted_labels <- labels[order(RACETHX_counts, decreasing = TRUE)]

# Set the maximum value for the y-axis
y_max <- 2000

# Plot the bar graph with sorted labels and counts
barplot(sorted_counts, xlab = "Race", ylab = "Count", main = "Diabetes By Race",
        col = "pink", names.arg = sorted_labels, ylim = c(0, y_max))

# Add data point values on each bar
text(x = barplot(sorted_counts, col = "pink", plot = FALSE),
     y = sorted_counts, label = sorted_counts, pos = 3, cex = 0.8)
```



This graph suggests that individuals that identified as “white only” have higher diabetes rates but the sample as whole contains more people from this demographic so it may be helpful to calculate each of these counts for each race over the number of individuals from that race in the survey, for example: ‘total white only with diabetes/total white only in survey’

```
all_race_counts <- table(my_data$RACETHX)
labels <- c("Hispanic",
            "White Only",
            "Black Only",
            "Asian Only",
            "Other")

# Sort the counts and labels in descending order
sorted_counts <- all_race_counts[order(all_race_counts, decreasing = TRUE)]
sorted_labels <- labels[order(all_race_counts, decreasing = TRUE)]

# Set the maximum value for the y-axis
y_max <- 20000

# Plot the bar graph with sorted labels and counts
barplot(sorted_counts, xlab = "Race", ylab = "Count", main = "Total Sample By Race",
        col = "lightblue", names.arg = sorted_labels, ylim = c(0, y_max))

# Add data point values on each bar
text(x = barplot(sorted_counts, col = "lightblue", plot = FALSE),
     y = sorted_counts, label = sorted_counts, pos = 3, cex = 0.8)
```





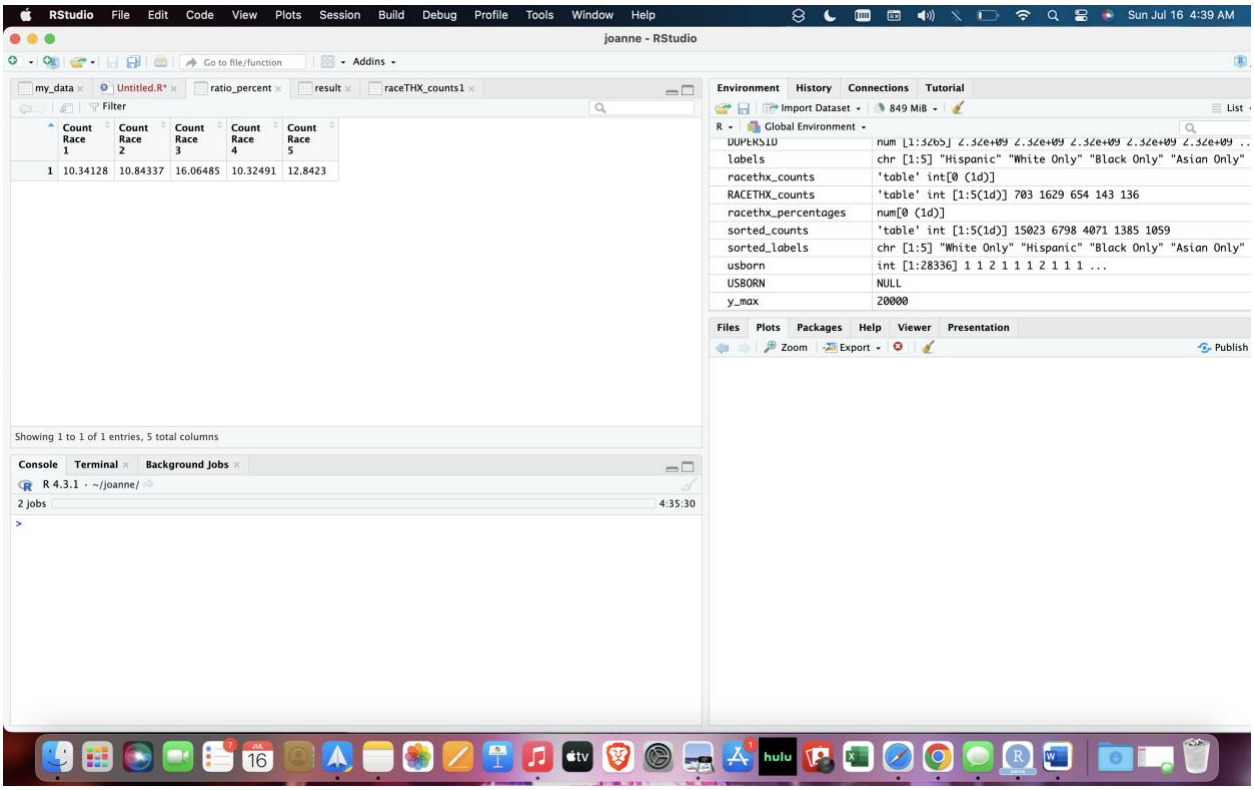
Let's create a new graph with the count variables and data from the "Diabetes by Race" graph divided by the count variables and data from the "Total Sample By Race" graph.

```
raceTHX_counts1 <- sqldf("SELECT
+ (SELECT COUNT(RACETHX) FROM diabetes_only WHERE RACETHX = 1) AS 'Count Race 1',
+ (SELECT COUNT(RACETHX) FROM diabetes_only WHERE RACETHX = 2) AS 'Count Race 2',
+ (SELECT COUNT(RACETHX) FROM diabetes_only WHERE RACETHX = 3) AS 'Count Race 3',
+ (SELECT COUNT(RACETHX) FROM diabetes_only WHERE RACETHX = 4) AS 'Count Race 4',
```

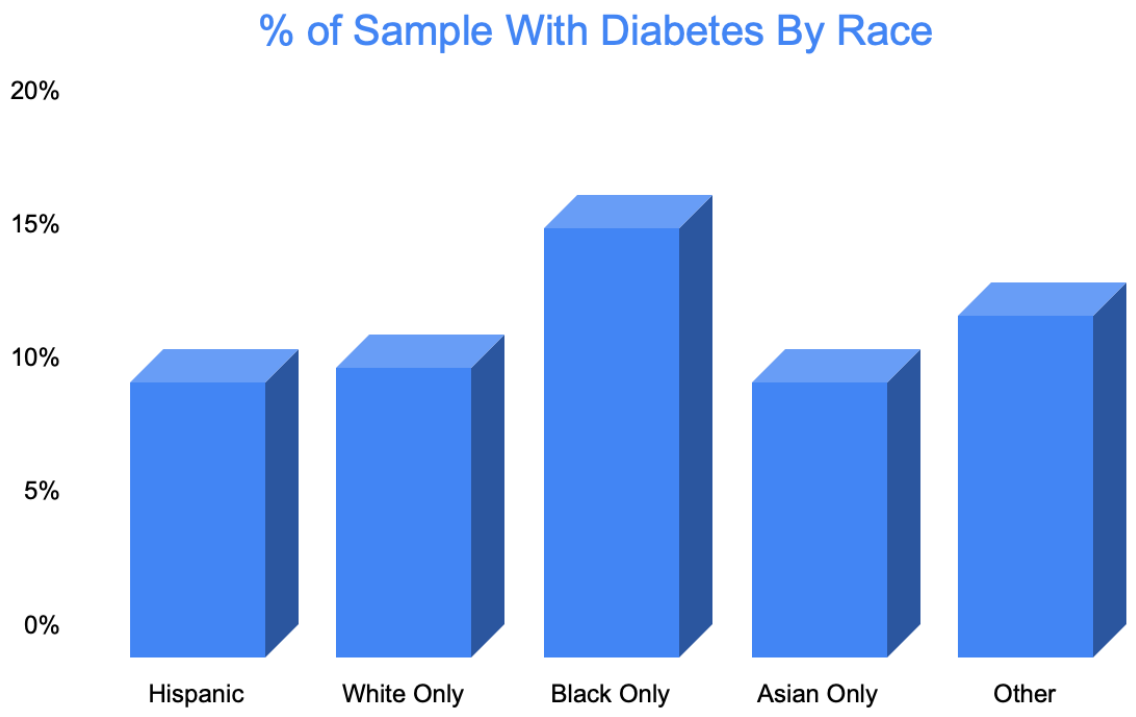
+ (SELECT COUNT(RACETHX) FROM diabetes_only WHERE RACETHX = 5) AS 'Count Race 5'")

View(raceTHX_counts1)

# Calculate the ratio of corresponding values and convert to percentage
ratio_percent <- (raceTHX_counts1 / all_race_counts) * 100



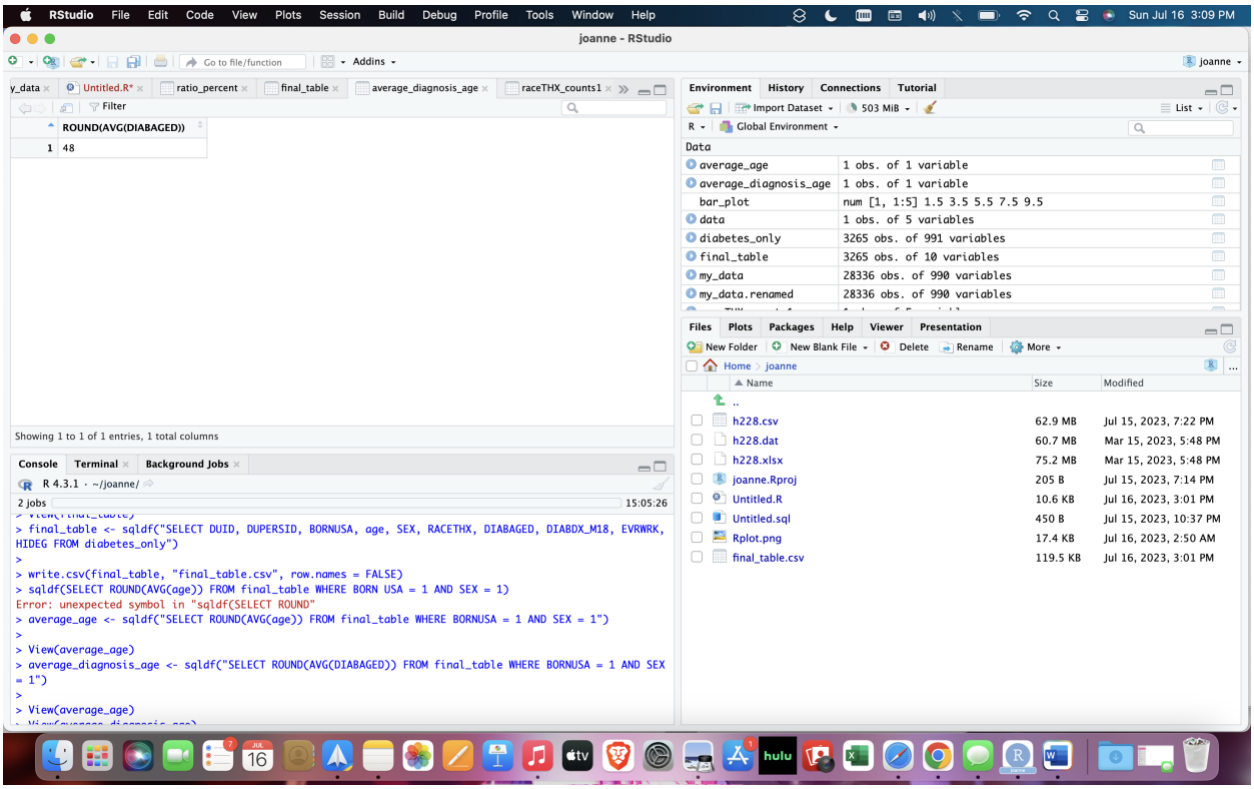
Visualize the data. I used google sheets to visualize this time to showcase some variety. But the dashboard will look much more cohesive.



Hispanic	White Only	Black Only	Asian Only	Other
10%	11%	**16%**	10%	**13%**

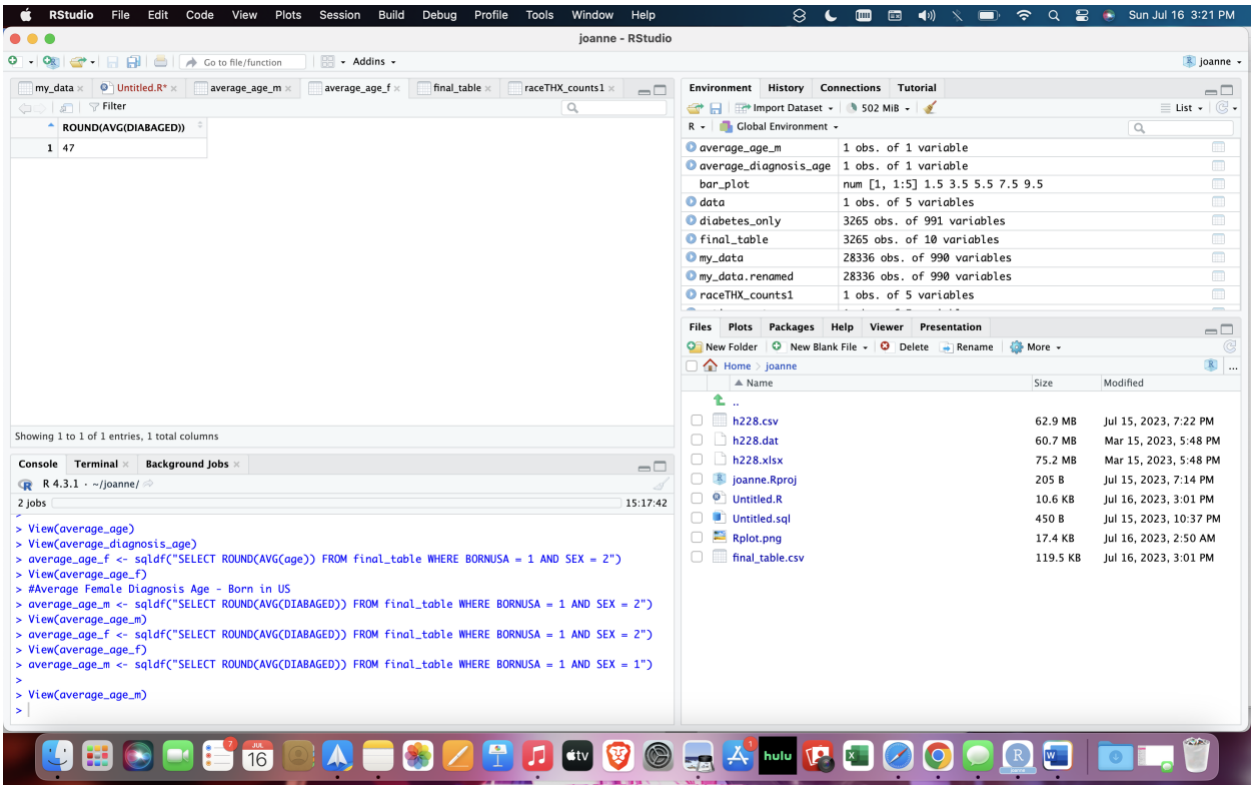
Calculating average diabetes diagnosis age for men and women born in the US versus men and women not born in the US.

#Average Male Diagnosis Age - Born in US
average_age_m <- sqldf("SELECT ROUND(AVG(DIABAGED)) FROM final_table WHERE BORNUSA = 1 AND SEX = 1")



Value: 48

#Average Female Diagnosis Age - Born in US
average_age_f <- sqldf("SELECT ROUND(AVG(DIABAGED)) FROM final_table WHERE BORNUSA = 1 AND SEX = 2")
average_age_f <- sqldf("SELECT ROUND(AVG(DIABAGED)) FROM final_table WHERE BORNUSA = 1 AND SEX = 2")
View(average_age_f)

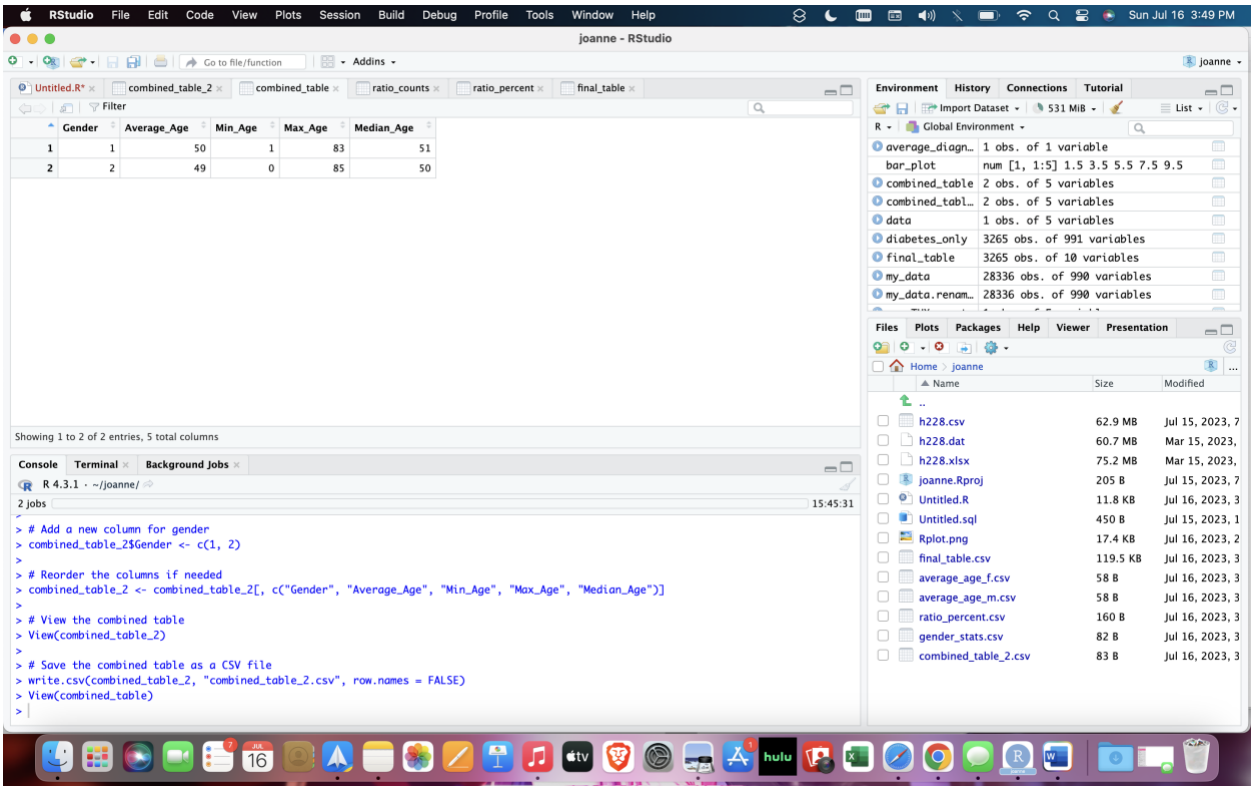


Value: 47

There appears to be no disparity between the average diabetes diagnosis age of men (48) in the US versus female (47). Let's look at the range for our values to see if there are any values skewing our data or if there is anything interesting there.

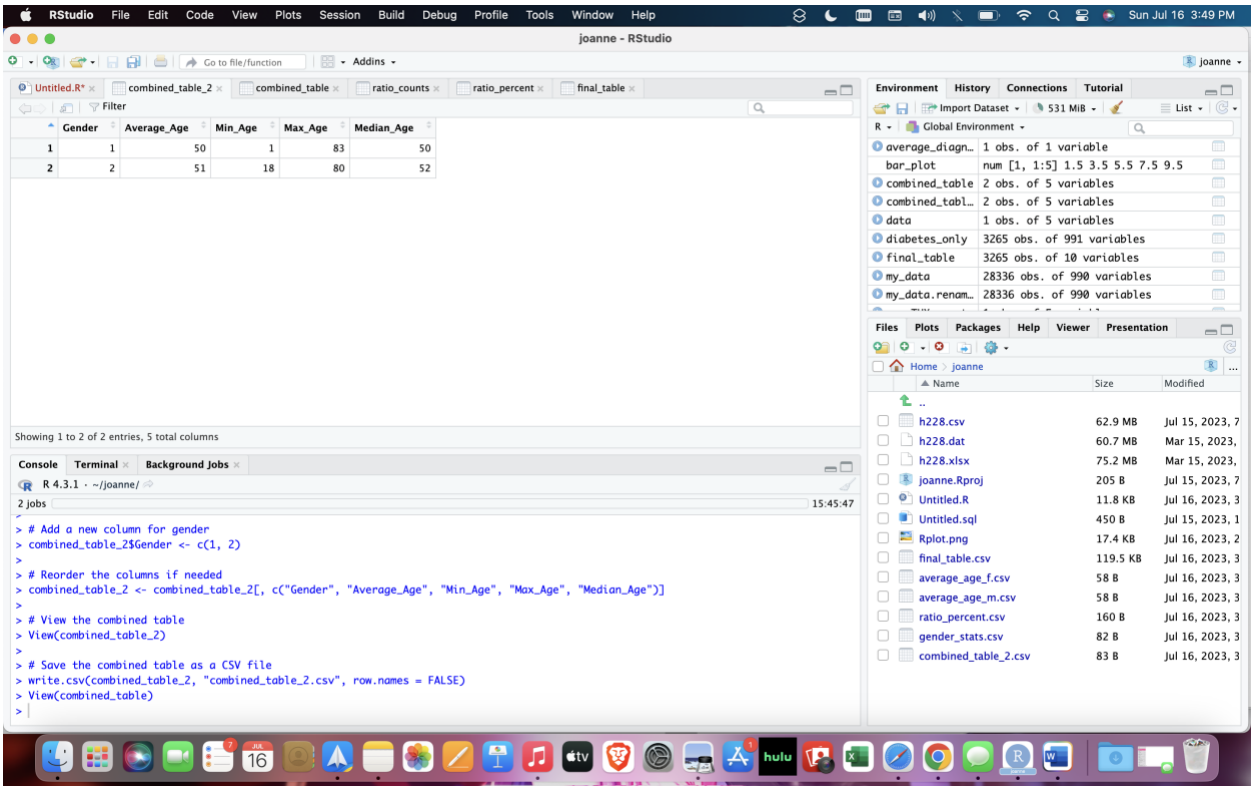
```
#Let's save these values as tables, along with our race ratio table from before.
# Save 'average_age_m' table as a CSV file
write.csv(average_age_m, "average_age_m.csv", row.names = FALSE)

# Save 'average_age_f' table as a CSV file
write.csv(average_age_f, "average_age_f.csv", row.names = FALSE)

# Save 'average_age_f' table as a CSV file
write.csv(ratio_percent, "ratio_percent.csv", row.names = FALSE)

# Let's combine the male and female gender table we created earlier and save it as a csv file.

# Create a new table combining 'average_age_m' and 'average_age_f'
combined_table <- rbind(average_age_m, average_age_f)
# Add a new column for gender
combined_table$Gender <- c(1, 2)

# Reorder the columns if needed
```

```
combined_table <- combined_table[, c("Gender", "Average_Age", "Min_Age", "Max_Age",
"Median_Age")]

# View the combined table
View(combined_table)

write.csv(combined_table, "gender_stats.csv", row.names = FALSE)
```




```
#Let's explore this same query but this time focusing on individuals not born in the US.

# Calculate values for NON US BORN males (SEX = 1) within the age range of 0 to 85
average_age_m_NONUS <- sqldf("SELECT ROUND(AVG(DIABAGED)) AS Average_Age, MIN(DIABAGED)
AS Min_Age, MAX(DIABAGED) AS Max_Age, ROUND(MEDIAN(DIABAGED)) AS Median_Age FROM
final_table WHERE BORNUSA = 2 AND SEX = 1 AND DIABAGED BETWEEN 0 AND 85")
View(average_age_m)

# Calculate values for NON US BORN females (SEX = 2) within the age range of 0 to 85
average_age_f_NONUS <- sqldf("SELECT ROUND(AVG(DIABAGED)) AS Average_Age, MIN(DIABAGED) AS
Min_Age, MAX(DIABAGED) AS Max_Age, ROUND(MEDIAN(DIABAGED)) AS Median_Age FROM
final_table WHERE BORNUSA = 2 AND SEX = 2 AND DIABAGED BETWEEN 0 AND 85")
View(average_age_f)

# Create a new table combining 'average_age_m_NONUS' and 'average_age_f_NONUS'
combined_table_2 <- rbind(average_age_m_NONUS, average_age_f_NONUS)
```

```
# Add a new column for gender
combined_table_2$Gender <- c(1, 2)

# Reorder the columns if needed
combined_table_2 <- combined_table_2[, c("Gender", "Average_Age", "Min_Age", "Max_Age",
"Median_Age")]

# View the combined table
View(combined_table_2)

# Save the combined table as a CSV file
write.csv(combined_table_2, "combined_table_2.csv", row.names = FALSE)
```

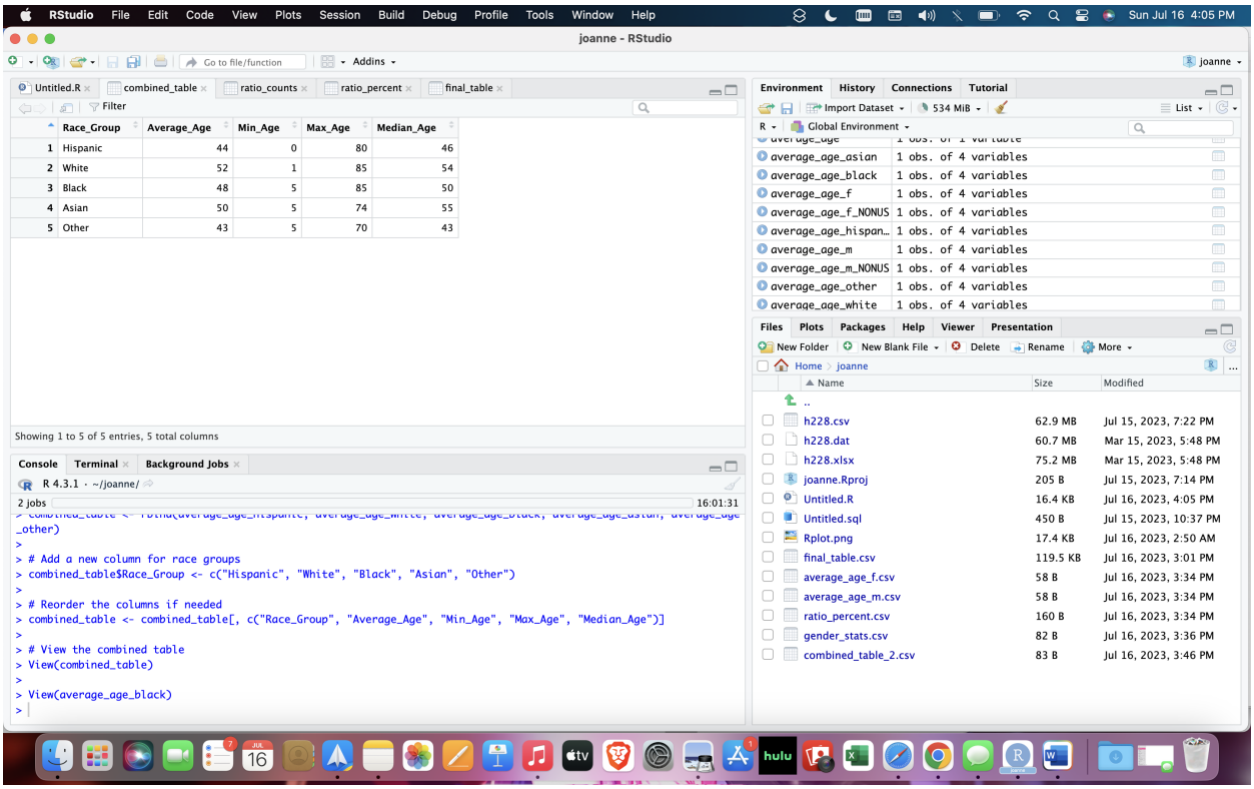


There is nothing too interesting here worth plotting between country of origin. Although the minimum age for diabetes diagnosis for NON-US BORN women (18) is significantly higher than men not born in the USA and both men and women born in the USA. But the '0' values reported for the other groups may be outliers, so I don't think it's too insightful.

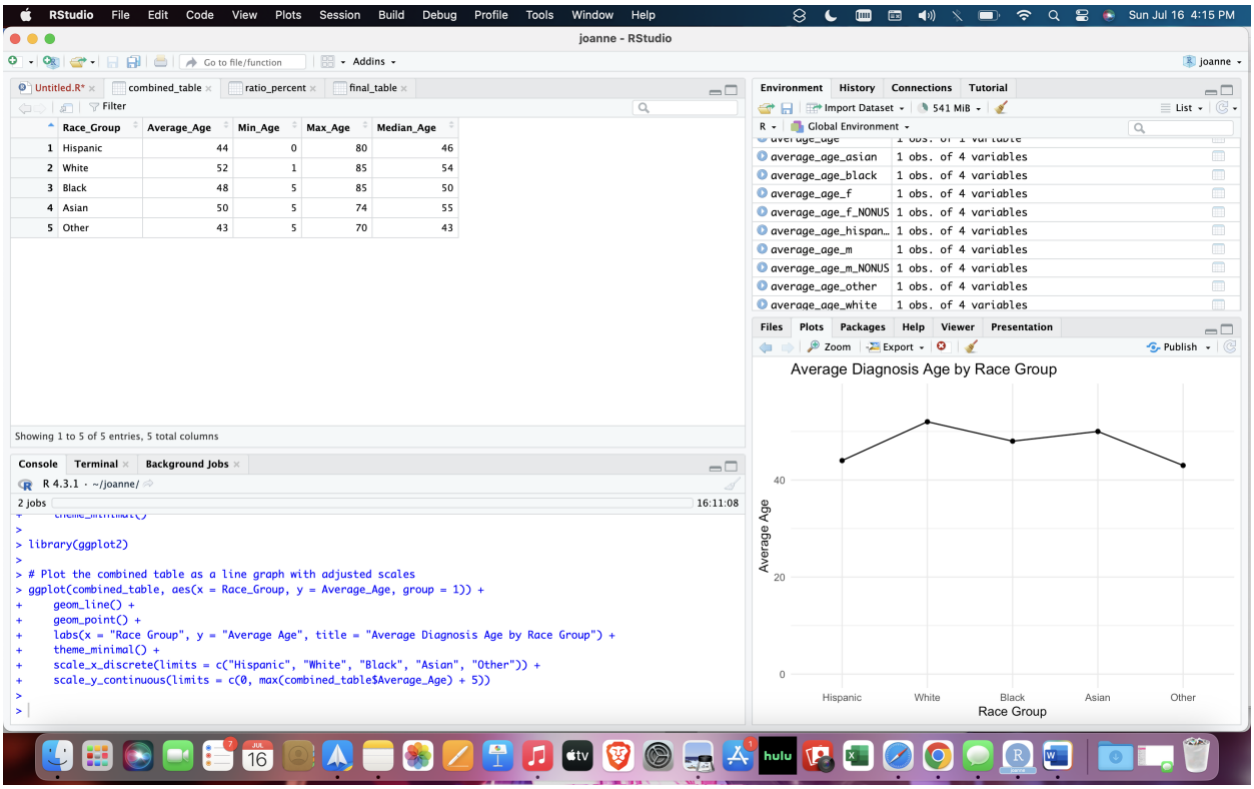
#Let's focus on US BORN individuals now and see whether we identify differences between race and age of diabetes diagnosis.

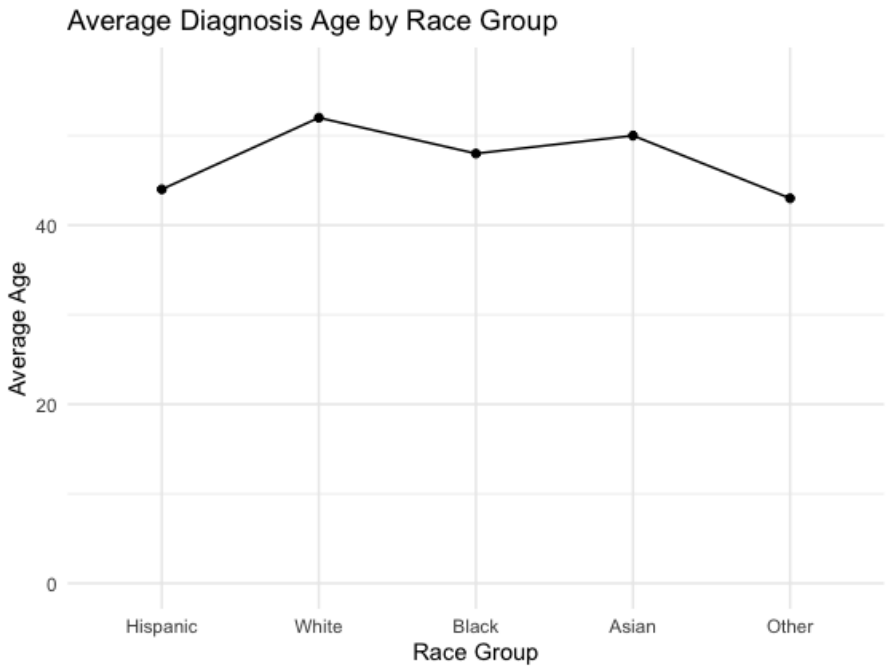
Calculate values for Hispanic race (RACETHX = 1) among U.S. born individuals within the age range of 0 to 85

```
average_age_hispanic <- sqldf("SELECT ROUND(AVG(DIABAGED)) AS Average_Age, MIN(DIABAGED) AS
Min_Age, MAX(DIABAGED) AS Max_Age, ROUND(MEDIAN(DIABAGED)) AS Median_Age FROM
final_table WHERE BORNUSA = 1 AND RACETHX = 1 AND DIABAGED BETWEEN 0 AND 85")

# Calculate values for White race (RACETHX = 2) among U.S. born individuals within the age range of 0 to
85
average_age_white <- sqldf("SELECT ROUND(AVG(DIABAGED)) AS Average_Age, MIN(DIABAGED) AS
Min_Age, MAX(DIABAGED) AS Max_Age, ROUND(MEDIAN(DIABAGED)) AS Median_Age FROM
final_table WHERE BORNUSA = 1 AND RACETHX = 2 AND DIABAGED BETWEEN 0 AND 85")

# Calculate values for Black race (RACETHX = 3) among U.S. born individuals within the age range of 0 to
85
average_age_black <- sqldf("SELECT ROUND(AVG(DIABAGED)) AS Average_Age, MIN(DIABAGED) AS
Min_Age, MAX(DIABAGED) AS Max_Age, ROUND(MEDIAN(DIABAGED)) AS Median_Age FROM
final_table WHERE BORNUSA = 1 AND RACETHX = 3 AND DIABAGED BETWEEN 0 AND 85")

# Calculate values for Asian race (RACETHX = 4) among U.S. born individuals within the age range of 0 to
85
average_age_asian <- sqldf("SELECT ROUND(AVG(DIABAGED)) AS Average_Age, MIN(DIABAGED) AS
Min_Age, MAX(DIABAGED) AS Max_Age, ROUND(MEDIAN(DIABAGED)) AS Median_Age FROM
final_table WHERE BORNUSA = 1 AND RACETHX = 4 AND DIABAGED BETWEEN 0 AND 85")

# Calculate values for Other race (RACETHX = 5) among U.S. born individuals within the age range of 0 to
85
average_age_other <- sqldf("SELECT ROUND(AVG(DIABAGED)) AS Average_Age, MIN(DIABAGED) AS
Min_Age, MAX(DIABAGED) AS Max_Age, ROUND(MEDIAN(DIABAGED)) AS Median_Age FROM
final_table WHERE BORNUSA = 1 AND RACETHX = 5 AND DIABAGED BETWEEN 0 AND 85")

#Let's combine these tables

# Combine the tables for different racial groups
combined_table <- rbind(average_age_hispanic, average_age_white, average_age_black,
average_age_asian, average_age_other)

# Add a new column for race groups
combined_table$Race_Group <- c("Hispanic", "White", "Black", "Asian", "Other")

# Reorder the columns if needed
combined_table <- combined_table[, c("Race_Group", "Average_Age", "Min_Age", "Max_Age",
"Median_Age")]

# View the combined table
View(combined_table)
```




```
# Plot the combined table as a line graph with adjusted scales
library(ggplot2)

ggplot(combined_table, aes(x = Race_Group, y = Average_Age, group = 1)) +
  geom_line() +
  geom_point() +
  labs(x = "Race Group", y = "Average Age", title = "Average Diagnosis Age by Race Group") +
  theme_minimal() +
  scale_x_discrete(limits = c("Hispanic", "White", "Black", "Asian", "Other")) +
  scale_y_continuous(limits = c(0, max(combined_table$Average_Age) + 5))
```

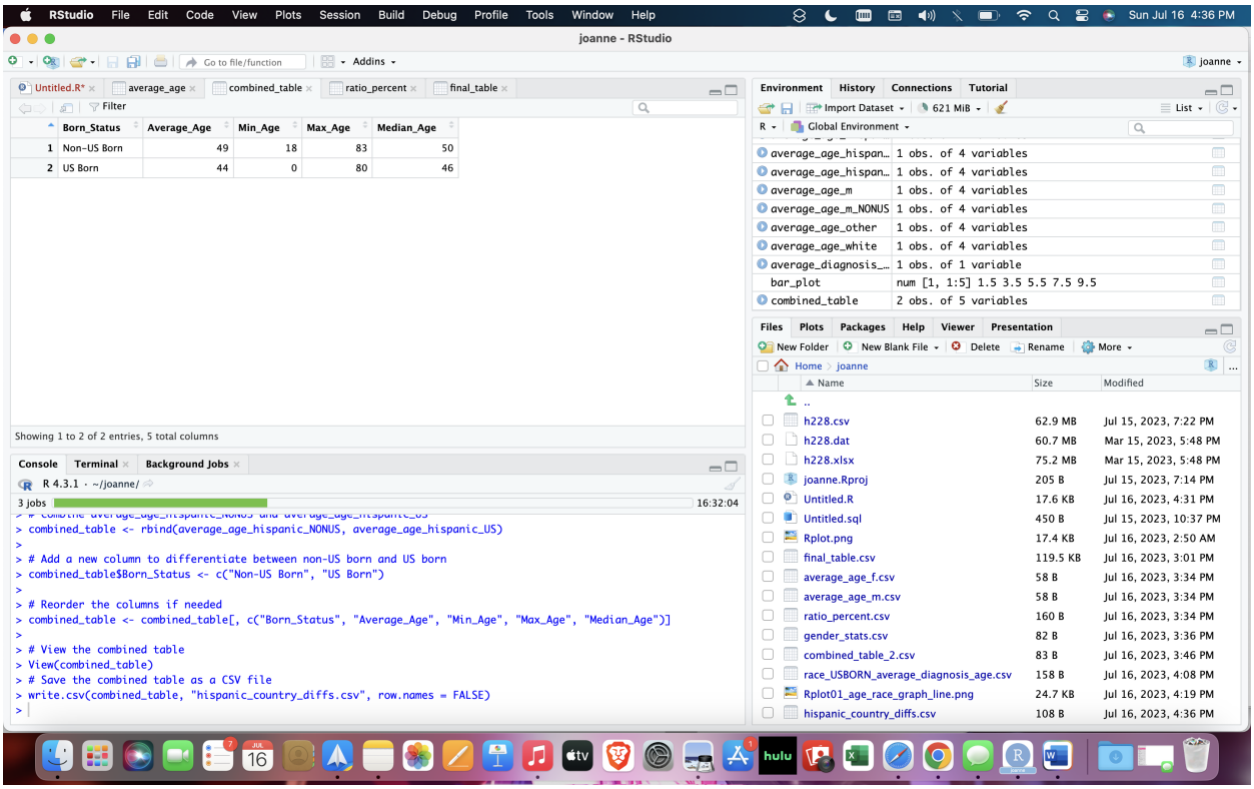




Let's summarize our findings here: White and Asian race groups tend to receive a diabetes diagnosis later than Black, Hispanic, and other groups. Hispanic and other (those that reported another race or multiple races) tend to receive a diabetes diagnosis around the age of 43-44 which suggests that perhaps unhealthy eating habits are developed earlier in these groups compared to other groups.

Please note that these are US BORN Hispanics, let’s see if NON US BORN Hispanics reflect these trends as well.

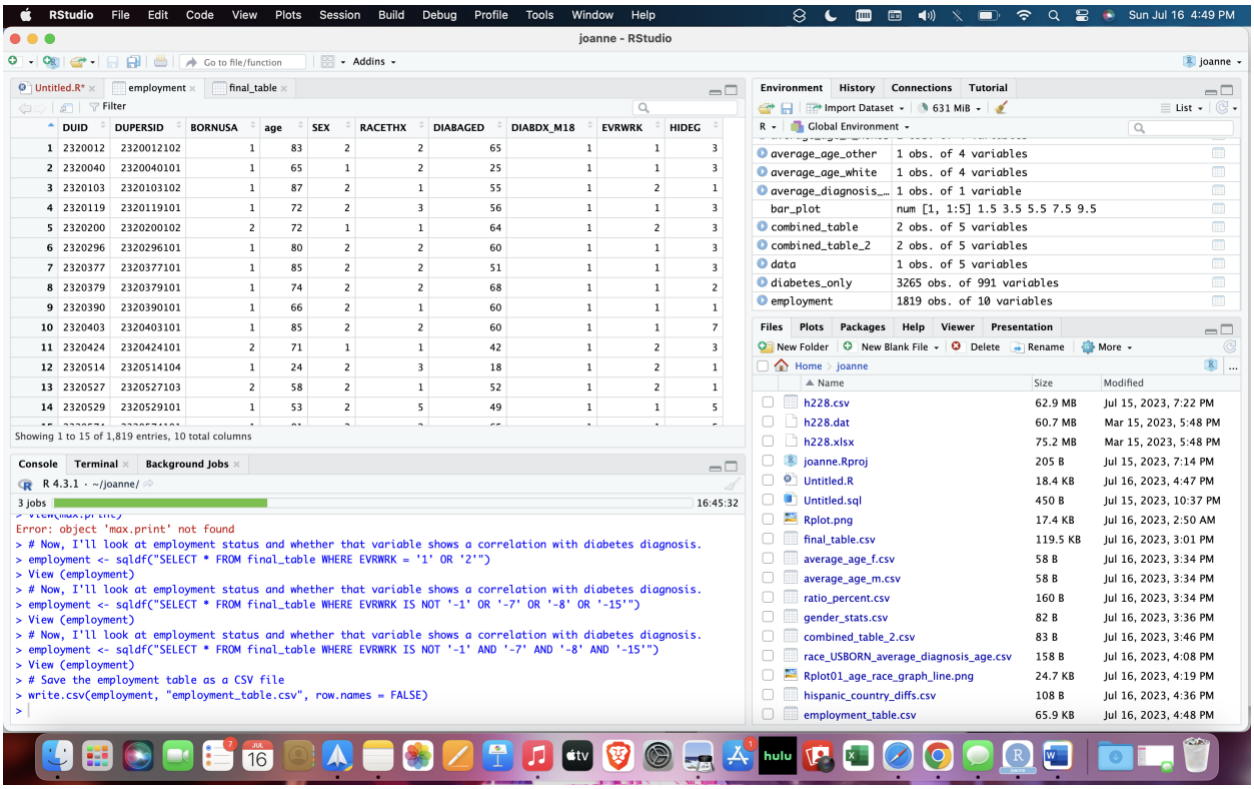
```
# Let's explore NON-US BORN Hispanic diagnosis age compared to US BORN Hispanic ages.

average_age_hispanic_NONUS <- sqldf("SELECT ROUND(AVG(DIABAGED)) AS Average_Age, MIN(DIABAGED) AS Min_Age, MAX(DIABAGED) AS Max_Age, ROUND(MEDIAN(DIABAGED)) AS Median_Age FROM final_table WHERE BORNUSA = 2 AND RACETHX = 1 AND DIABAGED BETWEEN 0 AND 85")
average_age_hispanic_US <- sqldf("SELECT ROUND(AVG(DIABAGED)) AS Average_Age, MIN(DIABAGED) AS Min_Age, MAX(DIABAGED) AS Max_Age, ROUND(MEDIAN(DIABAGED)) AS Median_Age FROM final_table WHERE BORNUSA = 1 AND RACETHX = 1 AND DIABAGED BETWEEN 0 AND 85")

# Combine tables
# Combine average_age_hispanic_NONUS and average_age_hispanic_US
combined_table <- rbind(average_age_hispanic_NONUS, average_age_hispanic_US)

# Add a new column to differentiate between non-US born and US born
combined_table$Born_Status <- c("Non-US Born", "US Born")

# Reorder the columns if needed
combined_table <- combined_table[, c("Born_Status", "Average_Age", "Min_Age", "Max_Age", "Median_Age")]

# View the combined table
View(combined_table)

# Save the combined table as a CSV file
write.csv(combined_table, "hispanic_country_diffs.csv", row.names = FALSE)
```



On average, US born Hispanics receive a diabetes diagnosis much earlier than those not born in the USA (44 vs. 49).

Now, I'll look at employment status and whether that variable shows a correlation with diabetes diagnosis. I have to make sure to remove the other variables in this column as I only need the '1=YES and 2=NO' values.

```
employment <- sqldf("SELECT * FROM final_table WHERE EVRWRK IS NOT '-1' AND '-7' AND '-8' AND '-15'")
View (employment)

# Save the employment table as a CSV file
write.csv(employment, "employment_table.csv", row.names = FALSE)
```



Now, I'll look at employment status and whether that variable shows a correlation with diabetes diagnosis.

employment <- sqldf("SELECT * FROM final_table WHERE EVRWRK IS NOT '-1' AND '-7' AND '-8' AND '-15'")
View (employment)

# Save the employment table as a CSV file
write.csv(employment, "employment_table.csv", row.names = FALSE)

# Select count of values where individuals with diabetes reported being employed vs. being unemployed
emp_count <- sqldf("SELECT COUNT(EVRWRK) FROM employment WHERE EVRWRK = 1")
not_emp_count <- sqldf ("SELECT COUNT(EVRWRK) FROM employment WHERE EVRWRK = 2")

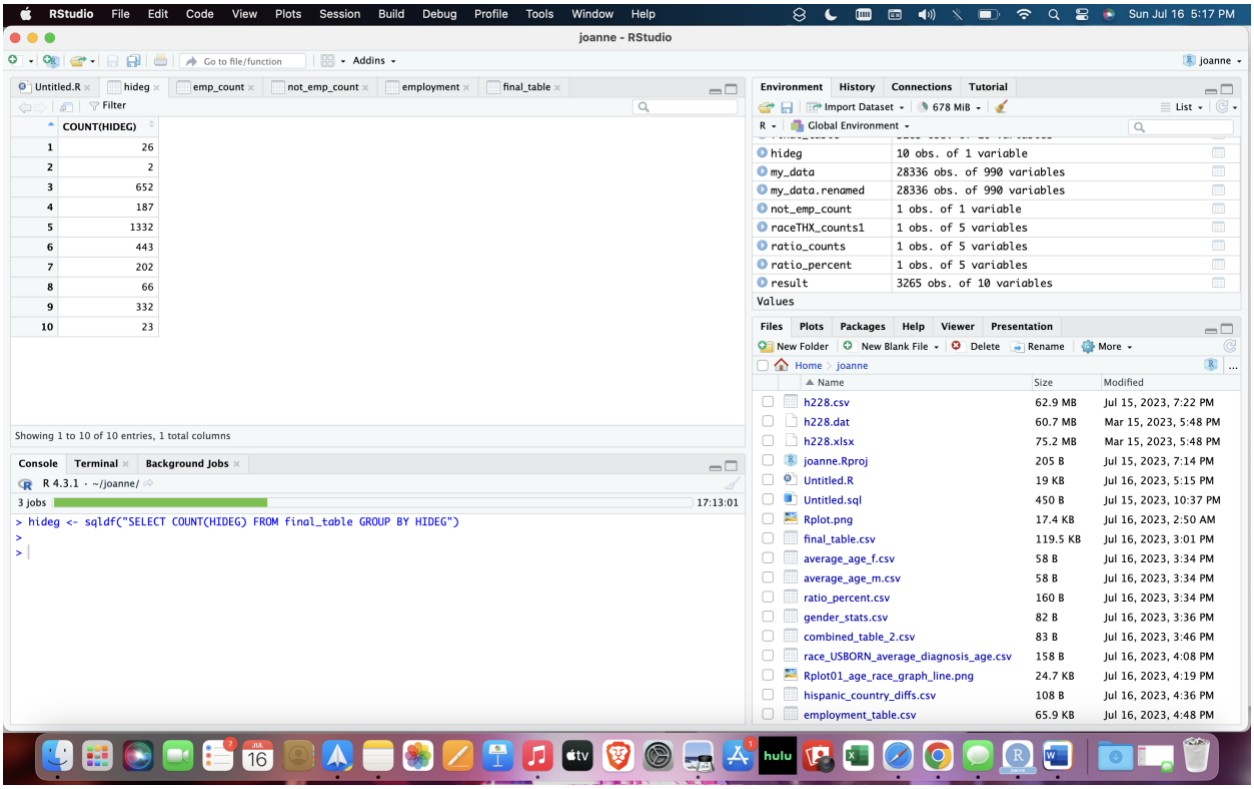
Values:

- Individuals who have worked that reported having diabetes: 1321
- Individuals who have not ever worked that reported having diabetes: 457

These results may be caused by other variables such as age as the survey contains data from children who are unable to work. So, this may only suggest that working adults are more likely to have a diabetes diagnosis which isn't particularly surprising. If we wanted to explore this, we could isolate variables such as age and further explore this. However, we aren't too particularly interested in this area and instead want to focus on the next, and last study aim: the impact of education on diabetes.

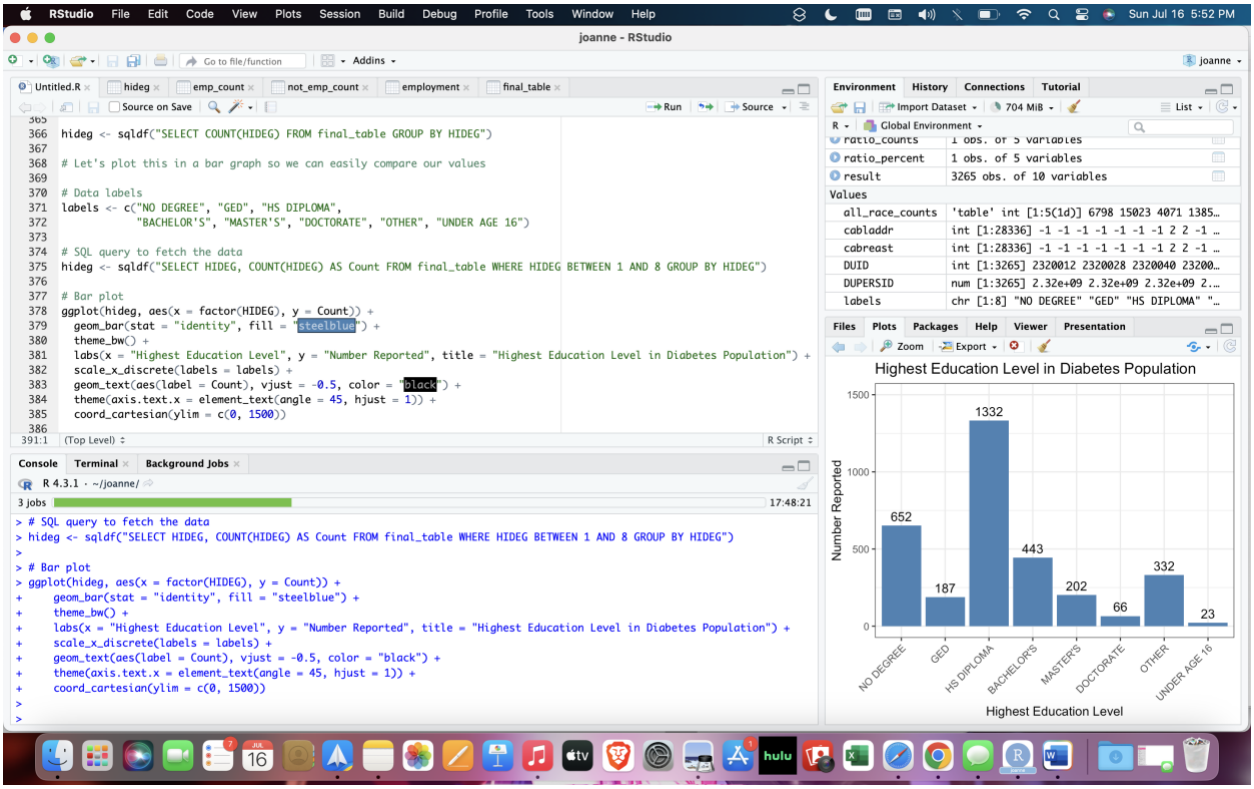
I'll start by counting how many individuals there are for each distinct education level in our final diabetes table.

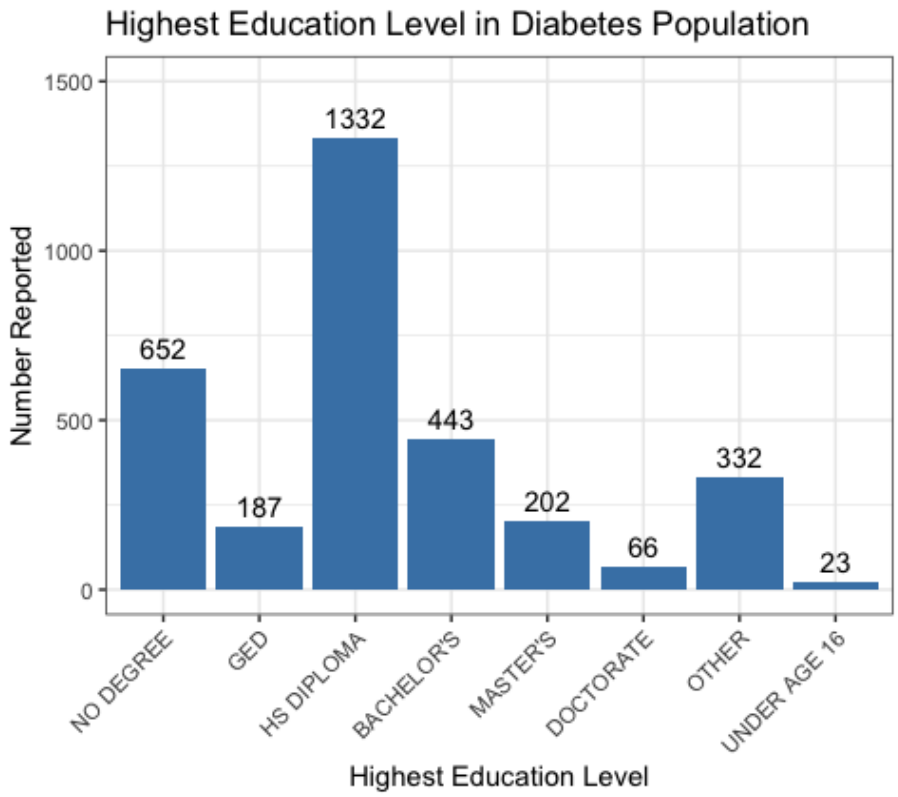
```
hideg <- sqldf("SELECT COUNT(HIDEG) FROM final_table GROUP BY HIDEG")
```



Let's plot this in a bar graph so we can easily compare our values

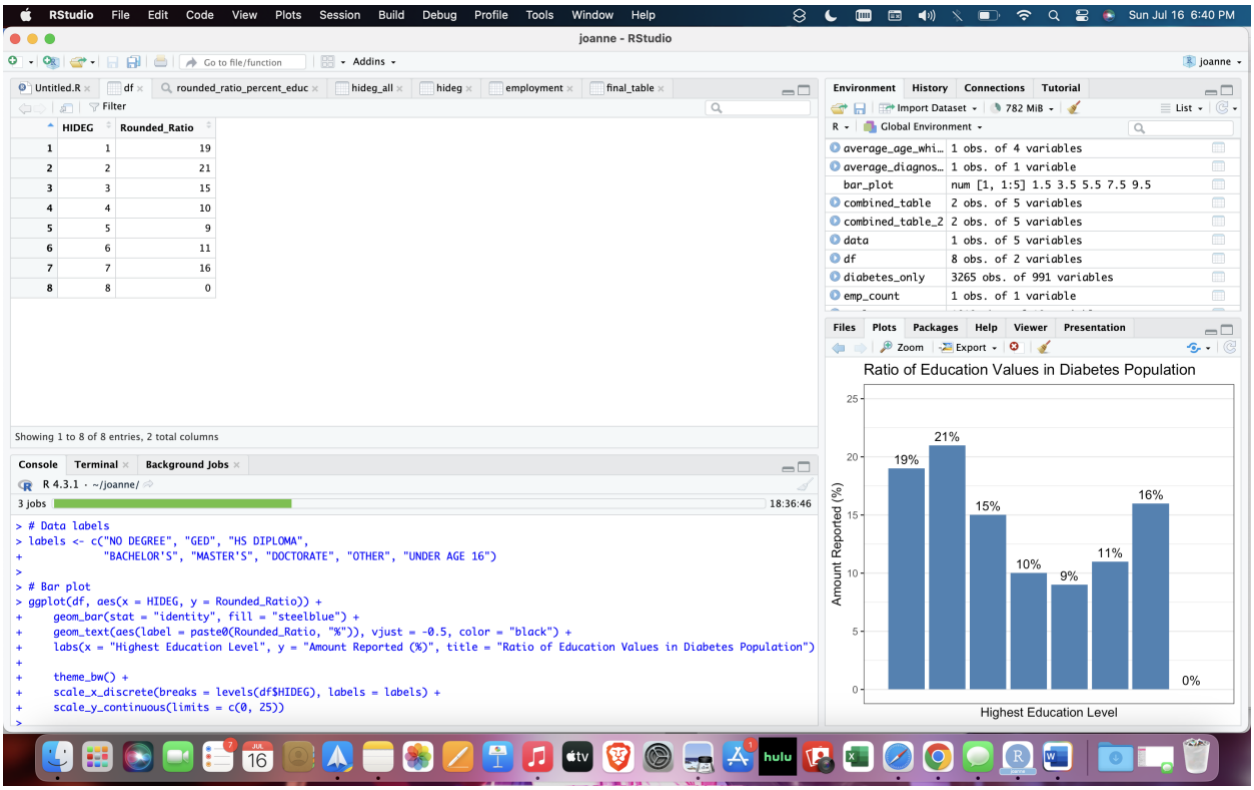
```
# Data labels
labels <- c("NO DEGREE", "GED", "HS DIPLOMA",
            "BACHELOR'S", "MASTER'S", "DOCTORATE", "OTHER", "UNDER AGE 16")

# SQL query to fetch the data
hideg <- sqldf("SELECT HIDEG, COUNT(HIDEG) AS Count FROM final_table WHERE HIDEG BETWEEN 1
AND 8 GROUP BY HIDEG")

# Bar plot
ggplot(hideg, aes(x = factor(HIDEG), y = Count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  labs(x = "Highest Education Level", y = "Number Reported", title = "Highest Education Level in Diabetes
Population") +
  scale_x_discrete(labels = labels) +
  geom_text(aes(label = Count), vjust = -0.5, color = "black") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_cartesian(ylim = c(0, 1500))
```

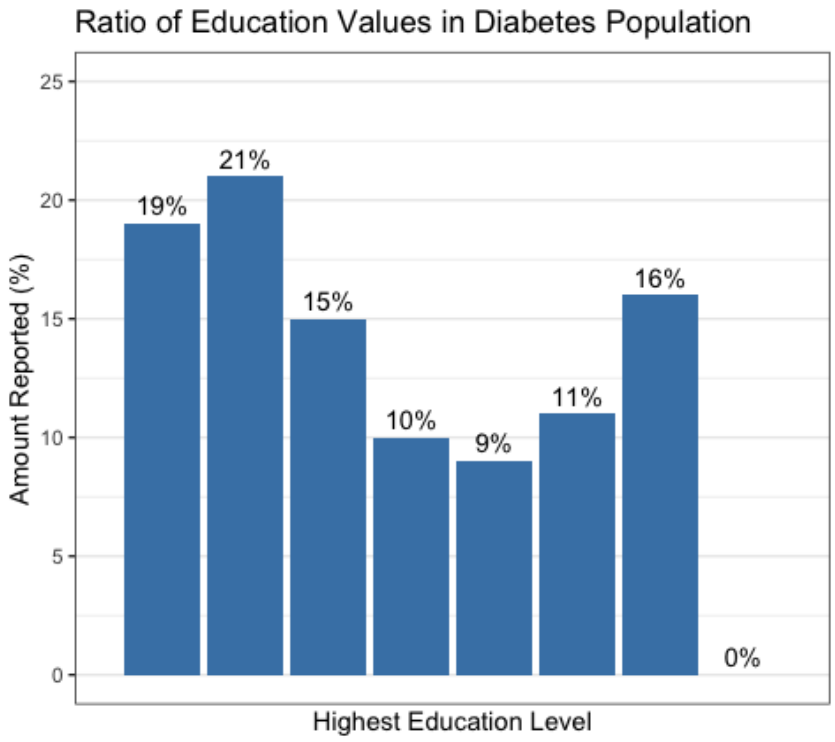

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Sun Jul 16 5:52 PM
joanne - RStudio
Untitled.R
hideg
emp_count
not_emp_count
employment
final_table
Source on Save
Run
Source
hideg <- sqldf("SELECT COUNT(HIDEG) FROM final_table GROUP BY HIDEG")
# Let's plot this in a bar graph so we can easily compare our values
# Data labels
labels <- c("NO DEGREE", "GED", "HS DIPLOMA",
"BACHELOR'S", "MASTER'S", "DOCTORATE", "OTHER", "UNDER AGE 16")
# SQL query to fetch the data
hideg <- sqldf("SELECT HIDEG, COUNT(HIDEG) AS Count FROM final_table WHERE HIDEG BETWEEN 1 AND 8 GROUP BY HIDEG")
# Bar plot
ggplot(hideg, aes(x = factor(HIDEG), y = Count)) +
geom_bar(stat = "identity", fill = "steelblue") +
theme_bw() +
labs(x = "Highest Education Level", y = "Number Reported", title = "Highest Education Level in Diabetes Population") +
scale_x_discrete(labels = labels) +
geom_text(aes(label = Count), vjust = -0.5, color = "black") +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
coord_cartesian(ylim = c(0, 1500))
391:1 (Top Level)
R Script
Console
Terminal
Background Jobs
R 4.3.1 · ~/joanne/
3 jobs
17:48:21
> # SQL query to fetch the data
> hideg <- sqldf("SELECT HIDEG, COUNT(HIDEG) AS Count FROM final_table WHERE HIDEG BETWEEN 1 AND 8 GROUP BY HIDEG")
> # Bar plot
> ggplot(hideg, aes(x = factor(HIDEG), y = Count)) +
+ geom_bar(stat = "identity", fill = "steelblue") +
+ theme_bw() +
+ labs(x = "Highest Education Level", y = "Number Reported", title = "Highest Education Level in Diabetes Population") +
+ scale_x_discrete(labels = labels) +
+ geom_text(aes(label = Count), vjust = -0.5, color = "black") +
+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
+ coord_cartesian(ylim = c(0, 1500))
Environment
History
Connections
Tutorial
Import Dataset
704 MiB
List
R
Global Environment
ratio_counts 1 obs. of 5 variables
ratio_percent 1 obs. of 5 variables
result 3265 obs. of 10 variables
Values
all_race_counts 'table' int [1:5(1d)] 6798 15023 4071 1385…
cabladdr int [1:28336] -1 -1 -1 -1 -1 -1 2 2 -1 …
cabreast int [1:28336] -1 -1 -1 -1 -1 -1 -1 2 2 -1 …
DUID int [1:3265] 2320012 2320028 2320040 23200…
DUPERSID num [1:3265] 2.32e+09 2.32e+09 2.32e+09 2.…
labels chr [1:8] "NO DEGREE" "GED" "HS DIPLOMA" "…
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Highest Education Level in Diabetes Population
Number Reported
1500
1000
500
0
652
187
1332
443
202
66
332
23
NO DEGREE
GED
HS DIPLOMA
BACHELOR'S
MASTER'S
DOCTORATE
OTHER
UNDER AGE 16
Highest Education Level




```
# Next, I'll find the ratio between the education values in our diabetes sample and the education values in our total sample and chart that. This is important because our data may be skewed due to more of the study population only having a HS diploma which could suggest that individuals in the diabetes population are more likely to have a HS diploma only, when this is instead the case for the entire sample (meaning that most people in the sample have a HS diploma). Converting our values into ratios instead could eliminate this issue and make our data more valid.

# SQL query to fetch the data
hideg_all <- sqldf("SELECT HIDEG, COUNT(HIDEG) AS Count FROM my_data WHERE HIDEG BETWEEN 1 AND 8 GROUP BY HIDEG")

# Calculate the ratio of corresponding values and convert to percentage
ratio_percent_educ <- (hideg$Count / hideg_all$Count) * 100

# Round the ratio to the nearest whole number
rounded_ratio_percent_educ <- round(ratio_percent_educ)

# Data labels
labels <- c("NO DEGREE", "GED", "HS DIPLOMA",
        "BACHELOR'S", "MASTER'S", "DOCTORATE", "OTHER", "UNDER AGE 16")

# Create a data frame with the rounded ratios and original HIDEG values
df <- data.frame(HIDEG = hideg$HIDEG, Rounded_Ratio = rounded_ratio_percent_educ)

# Data labels
```

```
labels <- c("NO DEGREE", "GED", "HS DIPLOMA",
            "BACHELOR'S", "MASTER'S", "DOCTORATE", "OTHER", "UNDER AGE 16")

# Bar plot
ggplot(df, aes(x = HIDEG, y = Rounded_Ratio)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = paste0(Rounded_Ratio, "%")), vjust = -0.5, color = "black") +
  labs(x = "Highest Education Level", y = "Amount Reported (%)", title = "Ratio of Education Values in
Diabetes Population") +
  theme_bw() +
  scale_x_discrete(breaks = levels(df$HIDEG), labels = labels) +
  scale_y_continuous(limits = c(0, 25))
```





No Degree	GED	HS Diploma	Bachelor's	Master's	Doctorate	Other	Under Age 16
19%	**21%**	15%	10%	9%	11%	16%	0%