Exploratory Data Analysis: Diabetes and Fracture Prevalences

To begin the data cleaning process, it was important to first identify the target population. Among the 26,510 population, all members were male and from the same market (San Antonio); however, each had a unique ‘anon\_id.’ This information allowed me to drop the columns for ‘member\_gender’ and ‘market’ and begin merging the sheets on ‘anon\_id’. The first page I merged was the ‘Fracture\_pop’.

df\_merged = pd.merge(left=df\_combined\_short, right=dfs['Fracture\_pop'], on= 'anon\_id', how= 'outer')

df\_merged

The following sheets were merged following the same code pattern: ‘Diabetic\_Pop’, 'Smoking\_pop', 'A1c\_lab\_pop', 'Vit\_D\_Pop','testost\_pop', 'calcium\_pop', and 'Insulin\_pop'. When deciding what sheets to merge into the data frame, I considered their relevance to diabetes and the possible correlations that could be drawn. For example, A1c value would be a direct indicator of diabetic status. Similarly, vitamin D values would be interesting to observe across different age ranges and how it correlates to diabetes.

As I merged each new sheet, I dropped irrelevant columns that would affect the tidiness of the data frame. The dropped columns included dates, descriptions, lab result names, and value measures that would not work well in matplotlib and seaborn graphs. However, I created new true/false columns from the information provided in diagnosis dates before dropping the dates.

#new columns

# dfs['new column']= dfs['old column'].notna()

# creates a true/false column based on the information from the old column

df\_merged['diabetic\_status']= df\_merged['earliest\_diabetes\_dx'].notna()

df\_merged

I used the same code to create ‘Fracture\_status’ and ‘smoker\_status.’ Next, to ensure that all the data would be compatible with matplotlib and seaborn graphing tools, I collected all the columns with values and converted them to numeric using the following code:

col\_list =[]

for col in df\_all\_columns:

print(col)

if 'val' in col or 'value' in col:

col\_list.append(col)

col\_list

for item in col\_list:

df\_final[item]= pd.to\_numeric(df\_final[item], errors='coerce')

Finally, I replaced and combined duplicated races to consolidate the ‘member\_race’ column. As a result, I was left with a final data frame of both qualitative and quantitative values that could be expressed in various graphs to find correlations.

In my EDA and advanced visualizations, I created an s group based on age to better support my graphs.

def s\_group(age):

if age < 70:

gp = 'u70'

else:

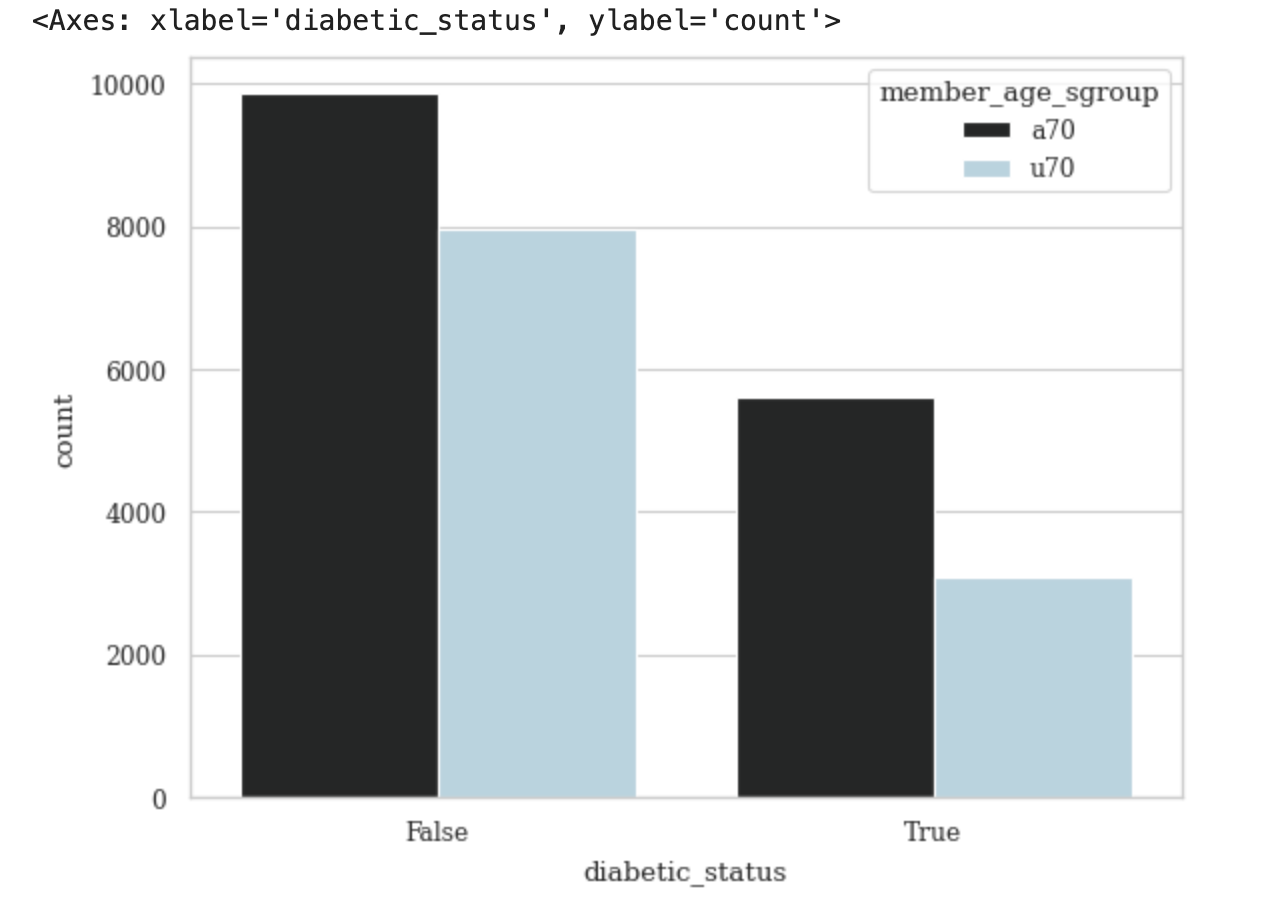
gp = 'a70'

return gp

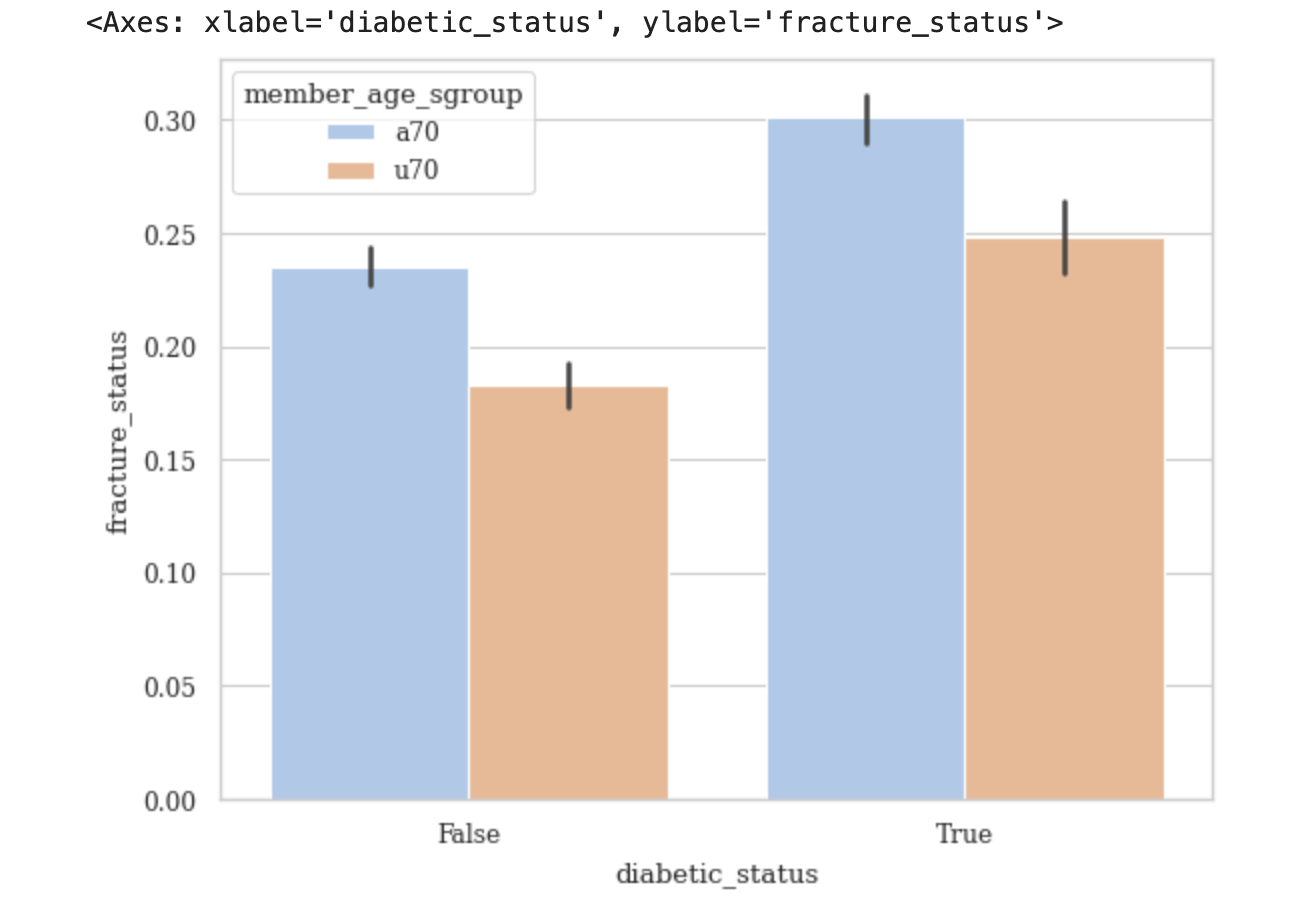
df\_final['member\_age\_sgroup']= df\_final['member\_age'].apply(lambda x: s\_group(x))

df\_final['member\_age\_sgroup']

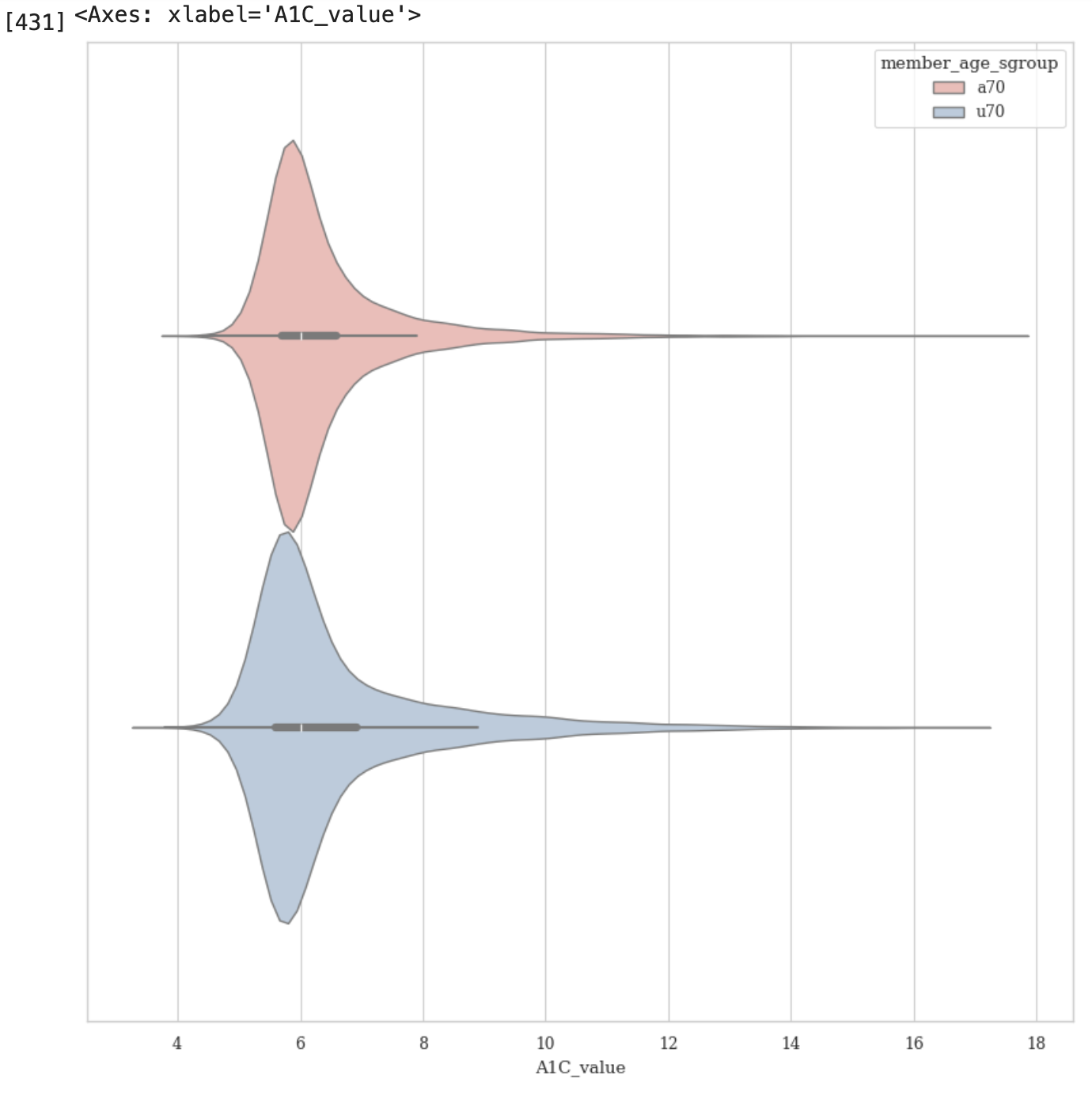
Using the information from the final data frame, I created general graphs of age and race distribution among the target population. Using a seaborn countplot, I found that the majority of the target population is above 70 and non-diabetic, and the minority is under 70 and diabetic.



Next, I explored the relationship between fractures and diabetics in both age groups (above 70 and below 70). Through the analysis of a barplot, I found that diabetics above 70 were most likely to suffer from bone fractures, and non-diabetics under 70 were least likely to suffer fractures.

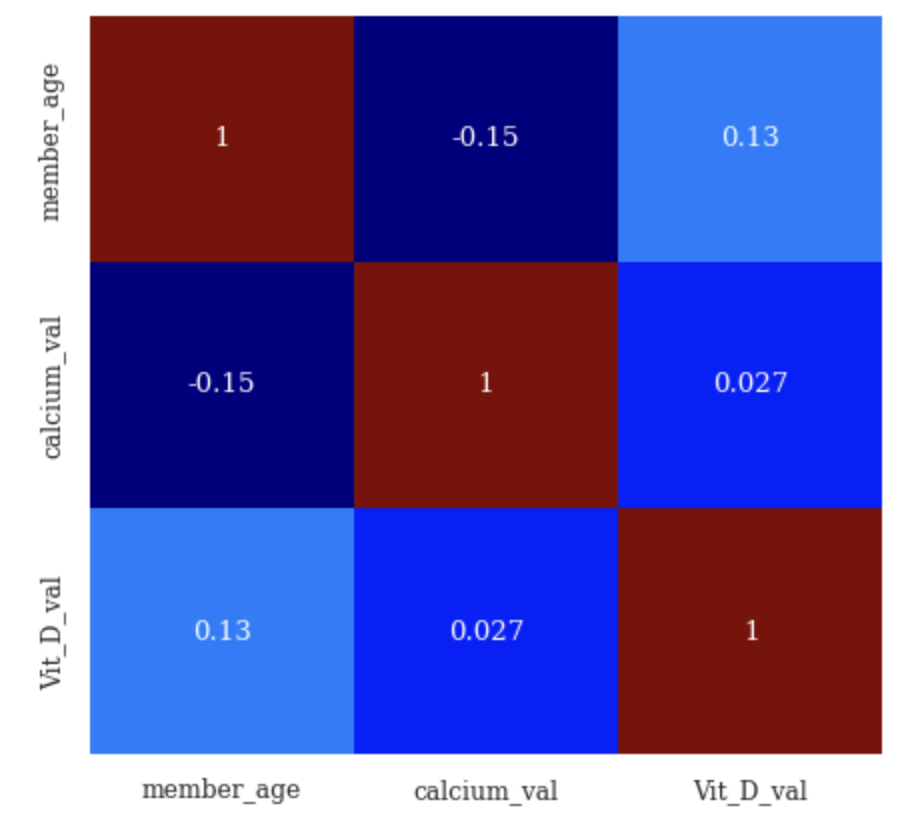


Once I found the correlation between diabetic status, age, and fractures, I wanted to know if A1c values would be consistent with this finding. I hypothesized that higher A1c values would be in the minority since most of the population is comprised of non-diabetic above 70 year old men. To support my hypothesis, I used a violin plot.



The violin plot proved that the average A1C value among the target population is between 5.0 and 7.5. A healthy A1C value is below 5.7%, a level of 5.7% to 6.4% indicates prediabetes, and a level of 6.5% or more indicates diabetes.[[1]](#footnote-0) The A1C value distribution is consistent with the distribution of diabetic v. non-diabetics within the target population. The majority of members have healthy to pre-diabetic A1C values, while the minority of members have diabetic A1C values.

For the advanced data visualization, I used a heatmap to find the correlation between member age, calcium value, and vitamin D value.



The heatmap proved that as a member's age increases, vitamin D increases-- most likely due to additional supplements. Yet, as a member’s age increases, calcium value decreases. Additionally, the heatmap displays that there is almost no correlation between vitamin D value and calcium value. To further analyze the correlation between calcium value and age, I crafted a jointplot using member age, calcium value, and set the hue to ‘diabetic\_status’. In the resulting graph, I found that while calcium values are mostly consistent among diabetics and non-diabetics, lower calcium values belong mainly to diabetics between 75-88. This finding strengthens the fact that diabetes leads to a "deterioration of calcium and bone metabolism."[[2]](#footnote-1) Other graphs I used to delve deeper into the analysis of the target population include: prevalence of fractures among diabetics, distribution of diabetes across race and age, and the effects of smoking on calcium and testosterone values.

In conclusion, the key insights from the data analysis suggests that age and diabetic status adversely affect calcium and vitamin D levels. Across the target population, the majority of men are in good health across all racial demographics. However, diabetes seems to affect mostly Hispanic and White men. The relatively small amounts of Asian and Black men affected by diabetes indicates a disparity that could be due in part to lack of information. A practical implication for the data analyzed would be the suggestion that more Asian and Black men get tested for diabetes. Similarly, a deeper investigation into cultural practices would be interesting to observe in both Hispanic and White demographics and their resulting effect on men’s health.

1. <https://www.cdc.gov/diabetes/managing/managing-blood-sugar/a1c.html#:~:text=Your%20A1C%20Result&text=A%20normal%20A1C%20level%20is,for%20developing%20type%202%20diabetes>. [↑](#footnote-ref-0)
2. <https://rdcu.be/dHBpf> [↑](#footnote-ref-1)