For this homework I chose a dataset that included the counties in each state with demographic information such as median age, counts of males and females and female percentage. I chose this dataset because I thought it would be interesting to combine this dataset with our COVID dataset to see what interesting correlations or links could be found. I do three different analyses in this assignment which I will go through along with their corresponding question answered. I did not have to clean the data up at all, I simply loaded it in as we had done previously in class with our county population dataset after uploading it.

**Code to load in datasets:**

covid\_df = spark.read.csv("dbfs:/databricks-datasets/COVID/covid-19-data/us-counties.csv", header=True, inferSchema=True)



countydemo\_df = spark.read.format("csv").load("dbfs:/FileStore/shared\_uploads/bmnesfed@utexas.edu/us\_county.csv",header=True, inferSchema=True)

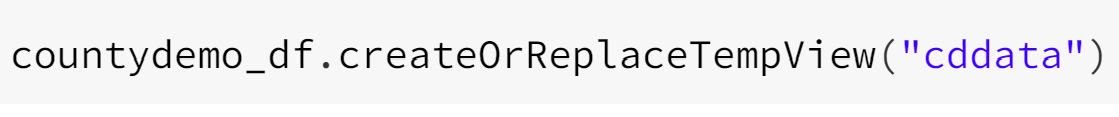


**Code to create views:**

covid\_df.createOrReplaceTempView("cvdata")

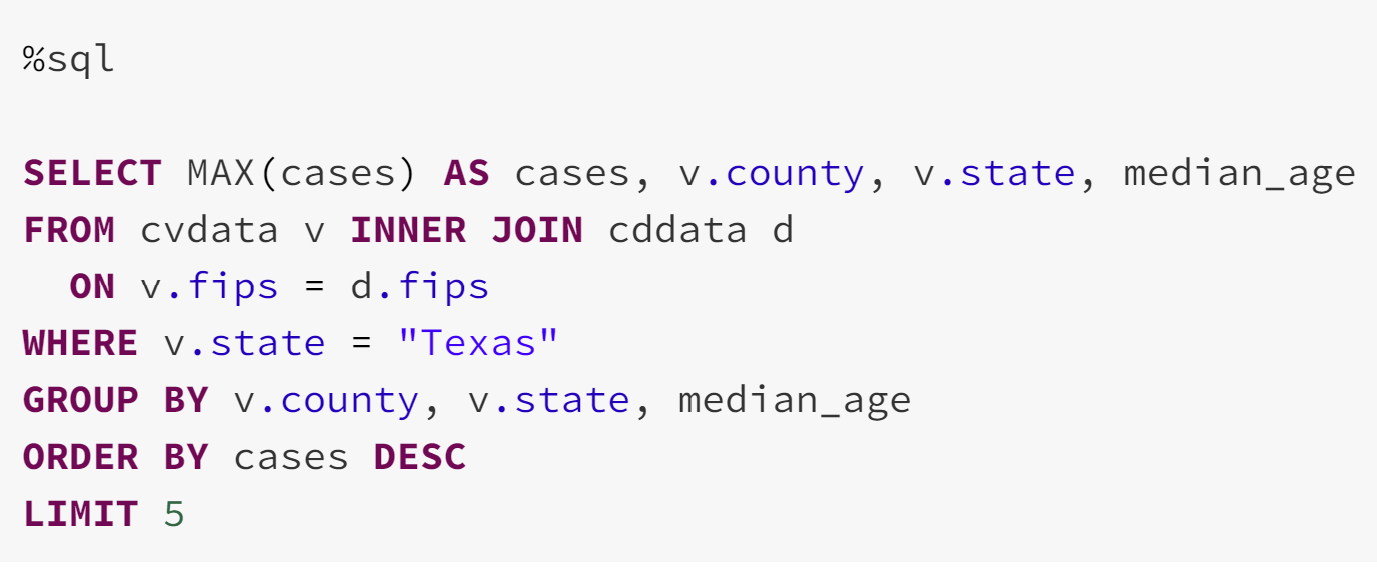


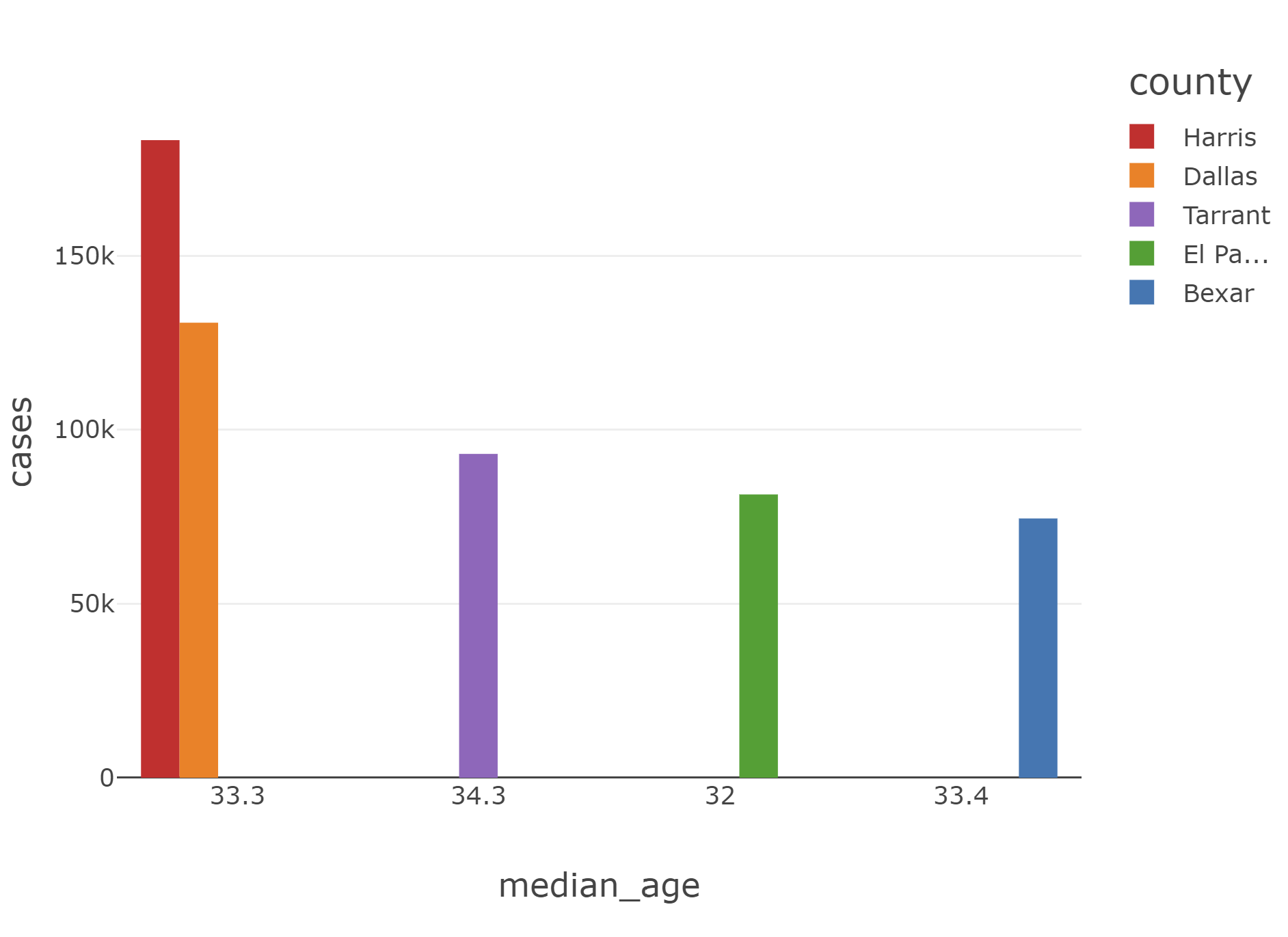
countydemo\_df.createOrReplaceTempView("cddata")

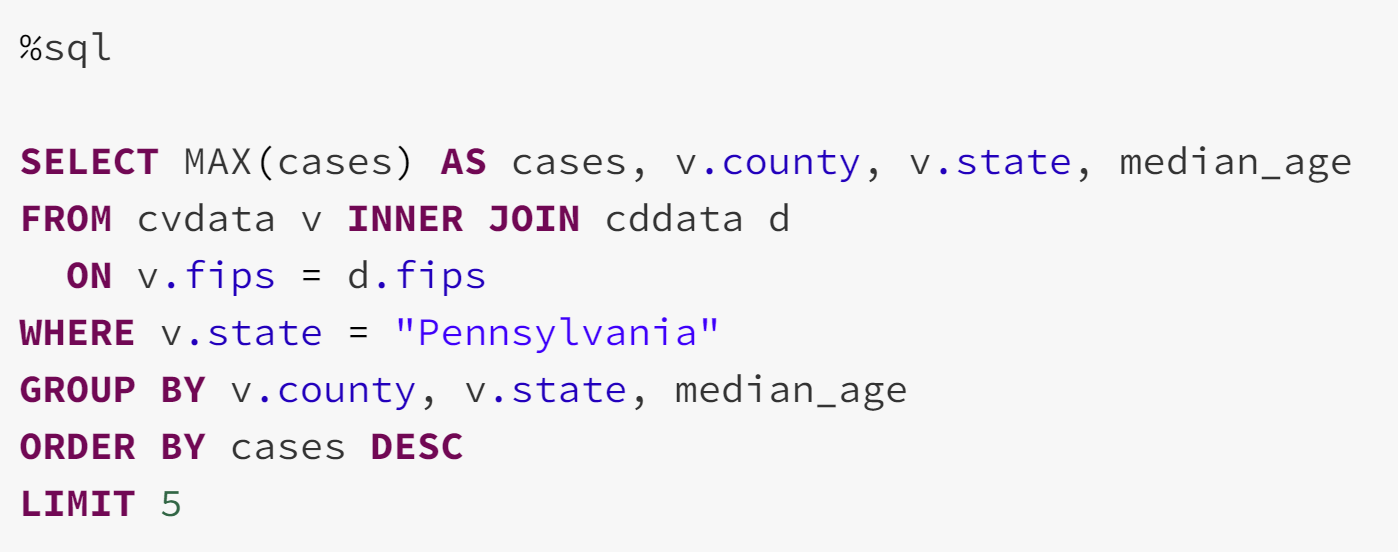


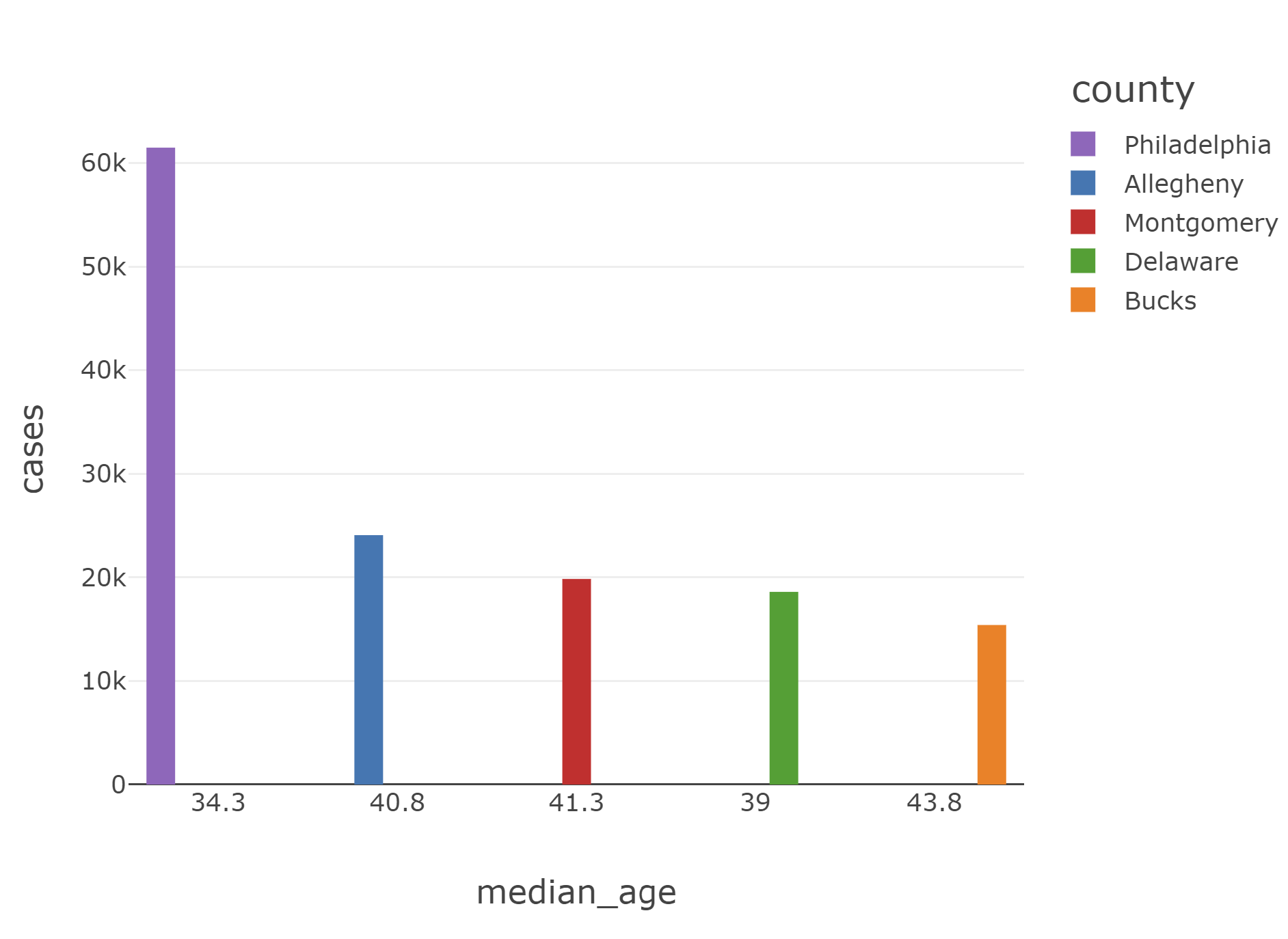
1. **How does the median age of the top 5 counties (in terms of most cases) in Texas (a less restrictive state) compare to the median age of the top 5 counties in Pennsylvania?**

In order to answer this question, I decided to look at the top 5 counties in regards to most COVID cases in Texas and then the same for Pennsylvania. I then compare the findings for each to see if a less restrictive state like Texas compared to a more restrictive state like Pennsylvania has any glaring differences in cases and/or median age trend.





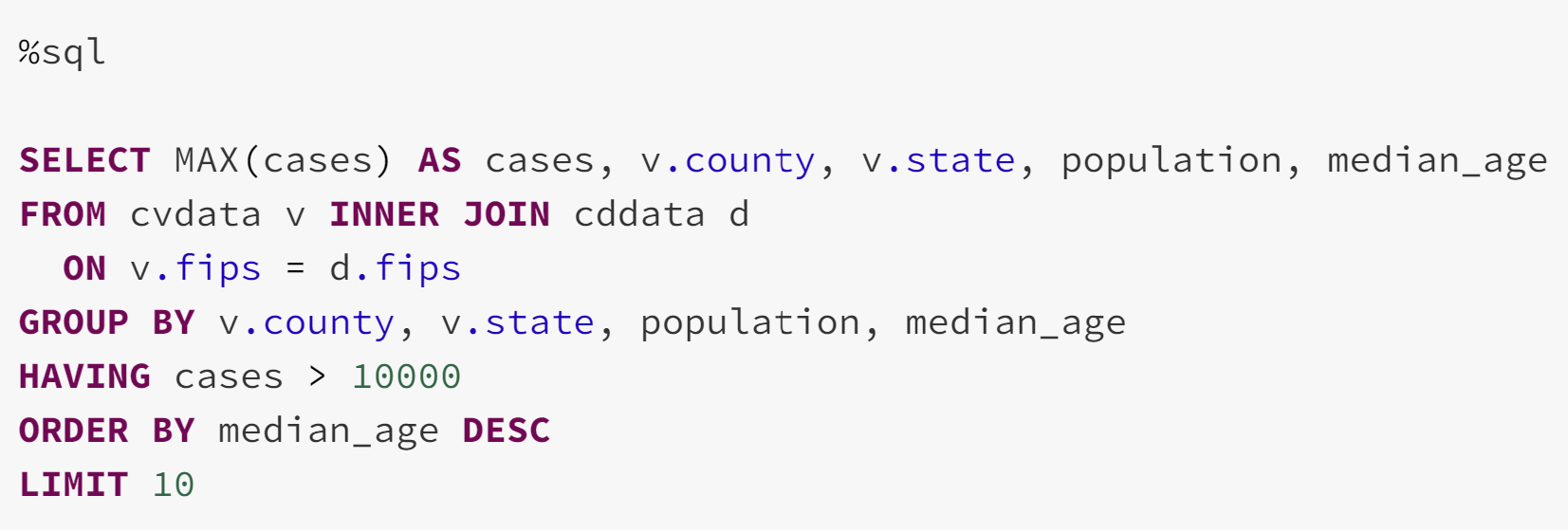


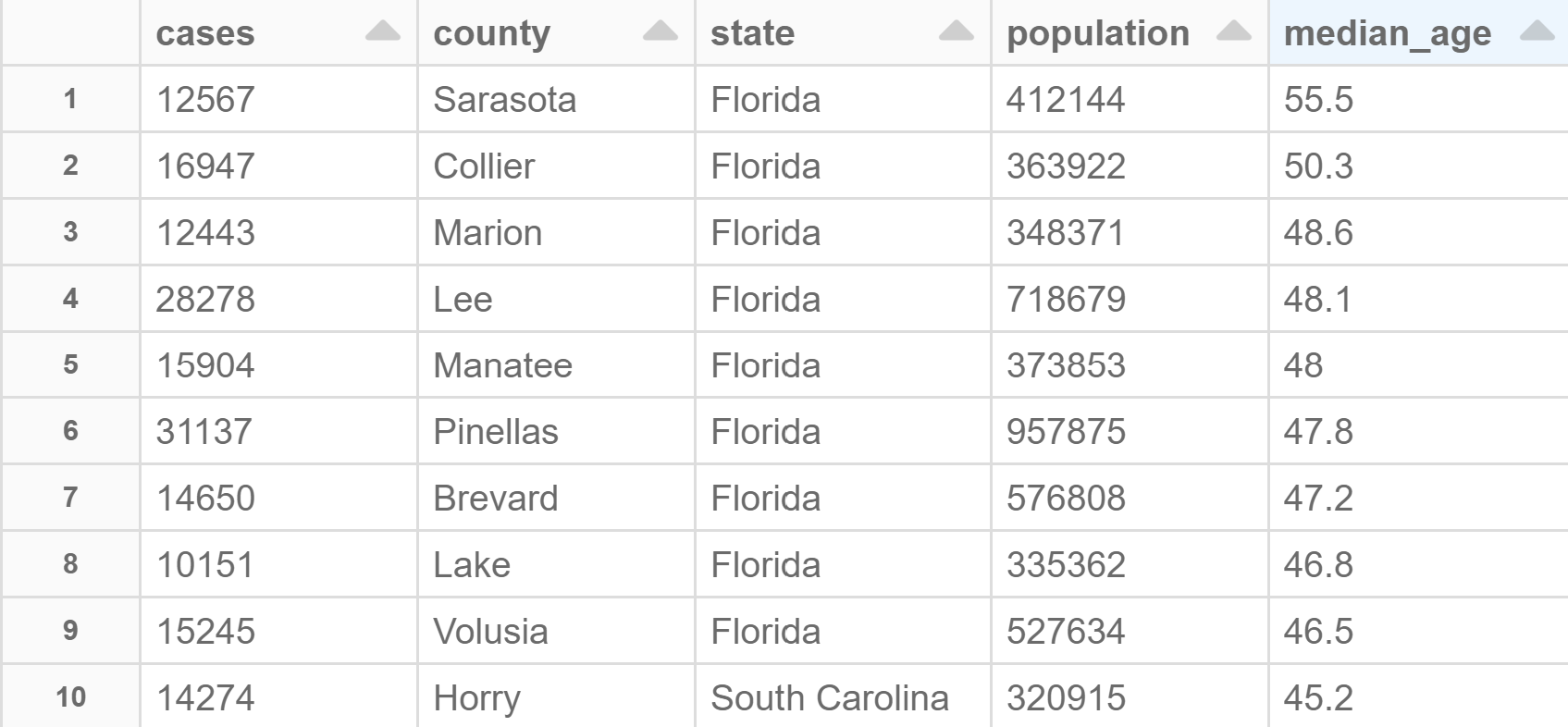


As we can see from the numbers, both overall cases and median age, I found some very interesting things. First off, it is clear Texas has many more cases overall as their #1 county, Harris, has over 100,000 more cases than Pennsylvania’s #1 county. Texas is a much bigger state than Pennsylvania, but the large difference is still pretty shocking. Another interesting note is that each of Pennsylvanias’ counties had a median age of equal to or greater than each of Texas’s counties. So the high number of cases in Texas could be because of younger people going out more often than in Pennsylvania because of the restriction differences. This analysis has shown (from a small sample size) that if there are less COVID restrictions in a state, younger people may be eager to go out, increasing the cases in a younger populated county. Texas generally has a younger population and Pennsylvania seems to have an older population, leading to their various results in handling the situation.

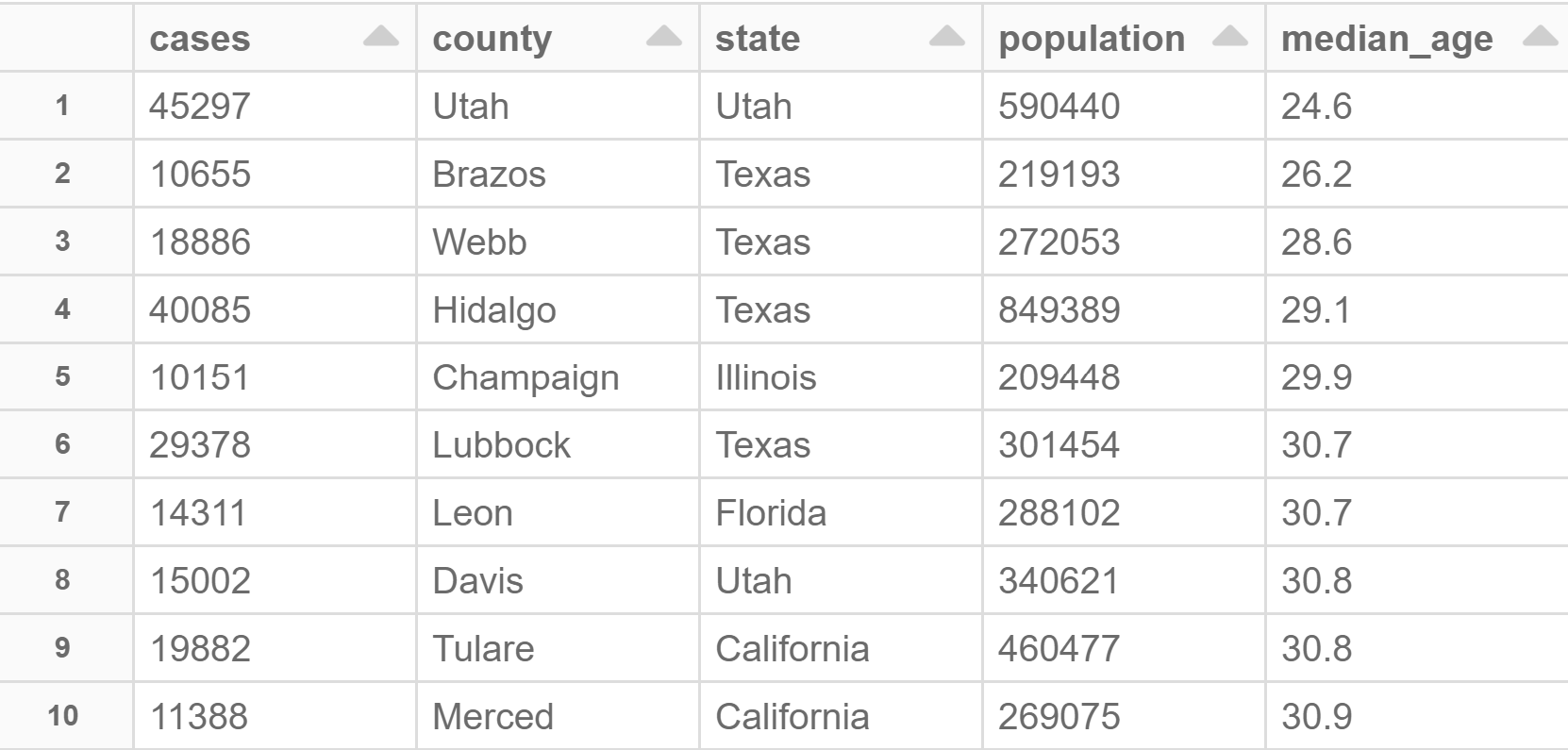
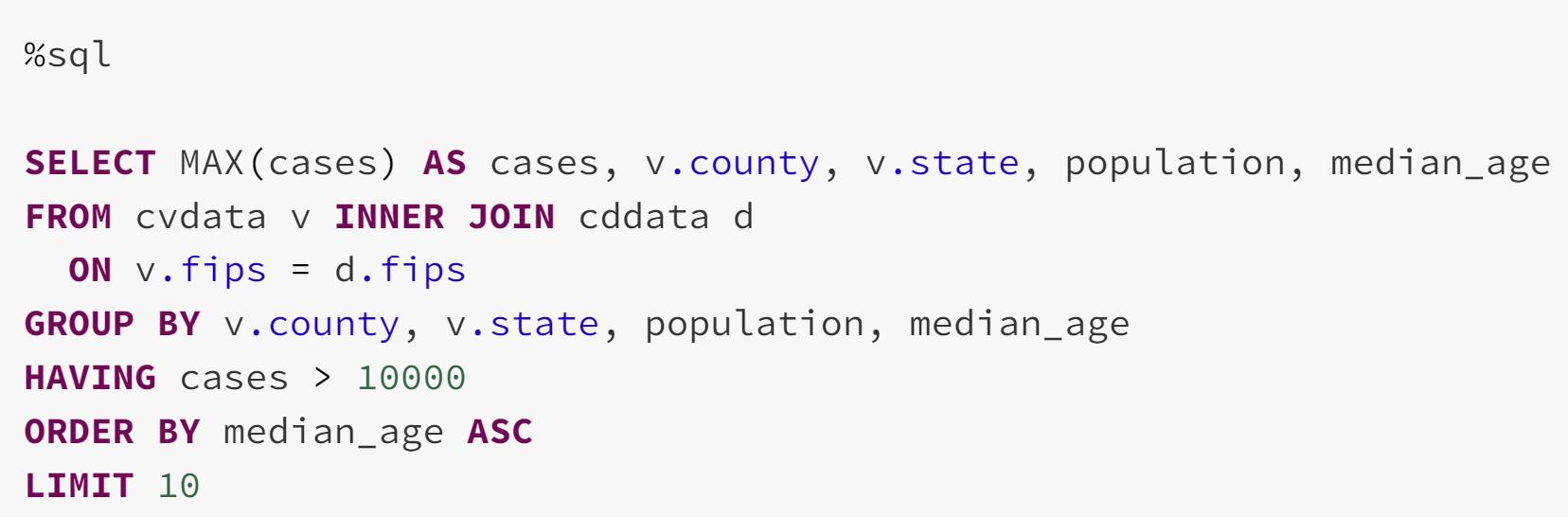
1. **Which states have higher median ages (a high amount of older people having cases) and which ones have lower median ages (a high amount of younger people having cases)?**

For this question, I thought it’d be interesting which states have a high amount of older people having COVID and which states have a high amount of younger people having COVID. After creating my output, I thought simply the tables were enough to be able to see interesting things.





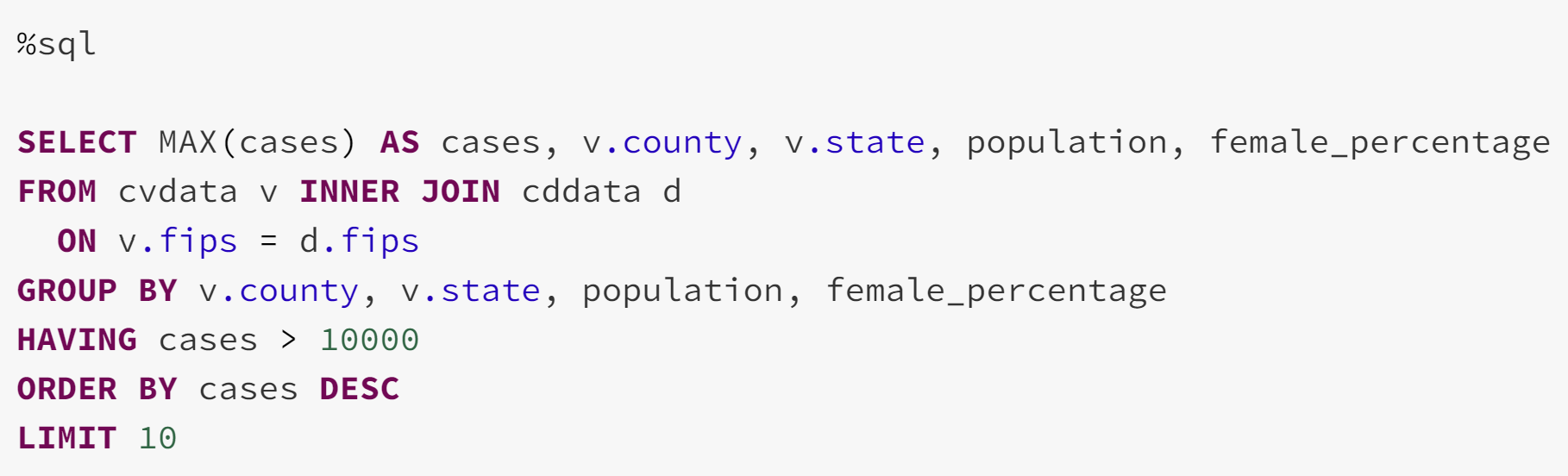
Taking the top 10 counties, it is clear which state has the highest amount of older people having cases, Florida. This finding makes a lot of sense since a lot of older people go to Florida to retire, therefore Florida just has some of the oldest counties in the U.S.

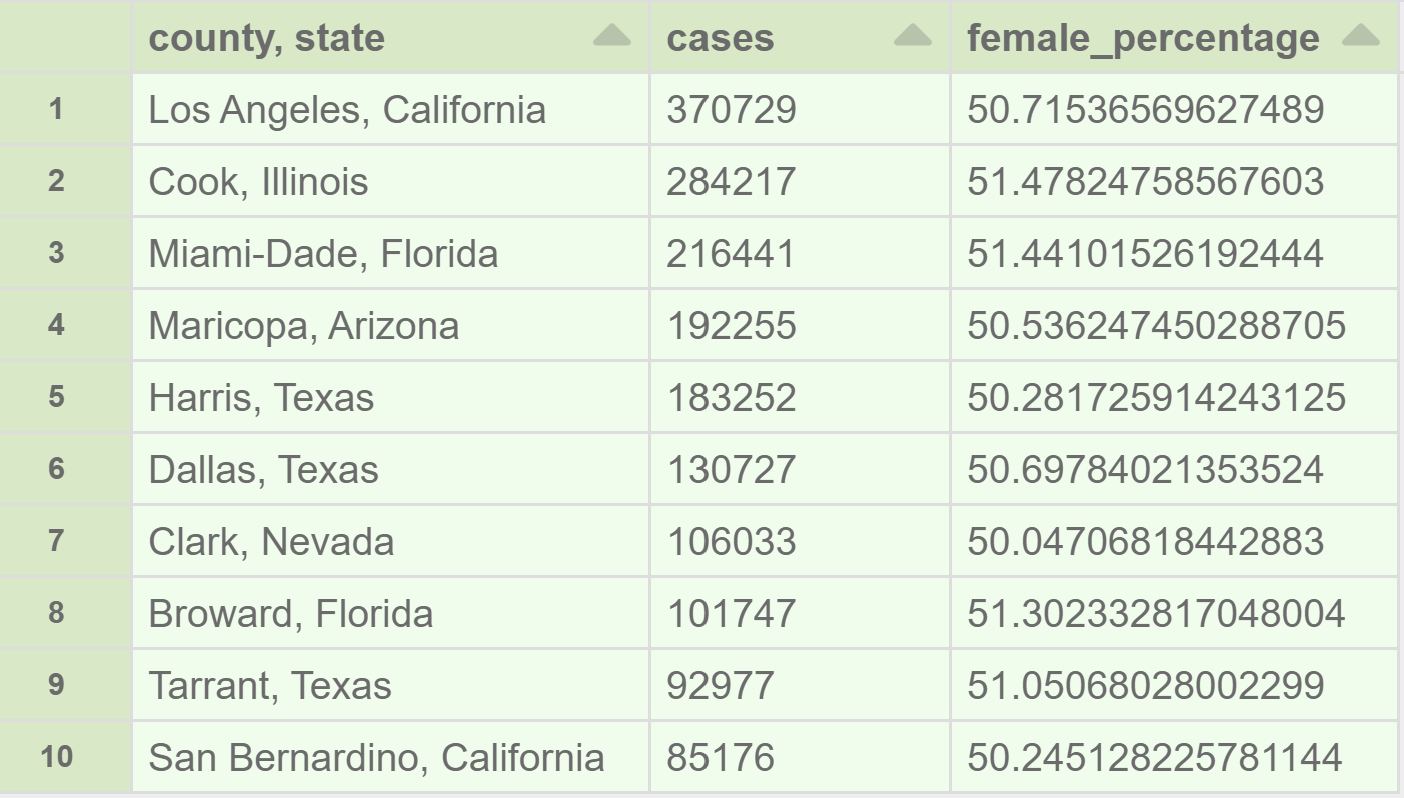


Looking at the top 10 states with a high amount of younger people dying, there is not one super clear trend or common state, although Texas has four of the top six youngest counties. This reinforces information we found in #1 when we discovered a lot of Texas’s counties were younger than other states. It is an interesting mix of states. So in this question, it was seen that Florida is host to many of the older counties getting COVID and while there’s a mix of states with younger counties getting COVID, Texas again is near the top of the list.

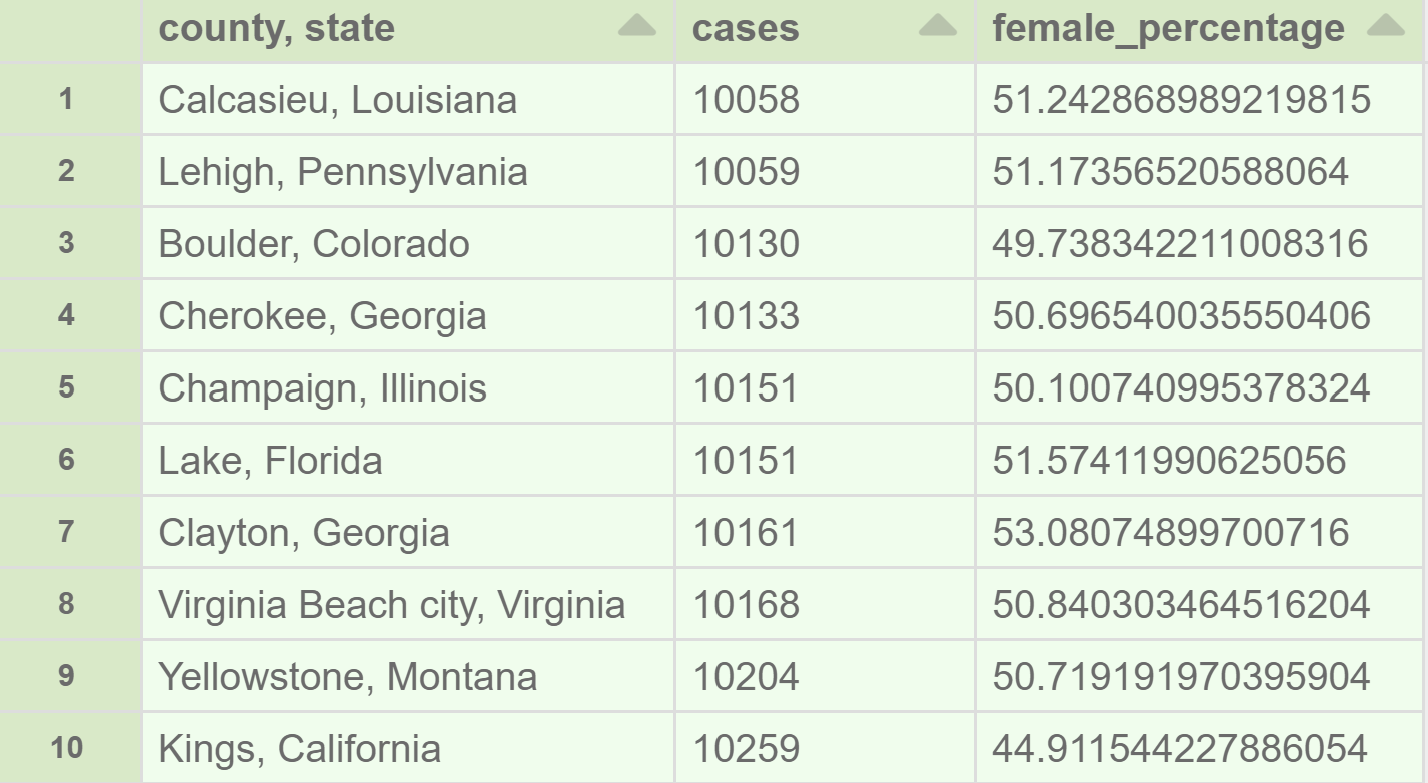
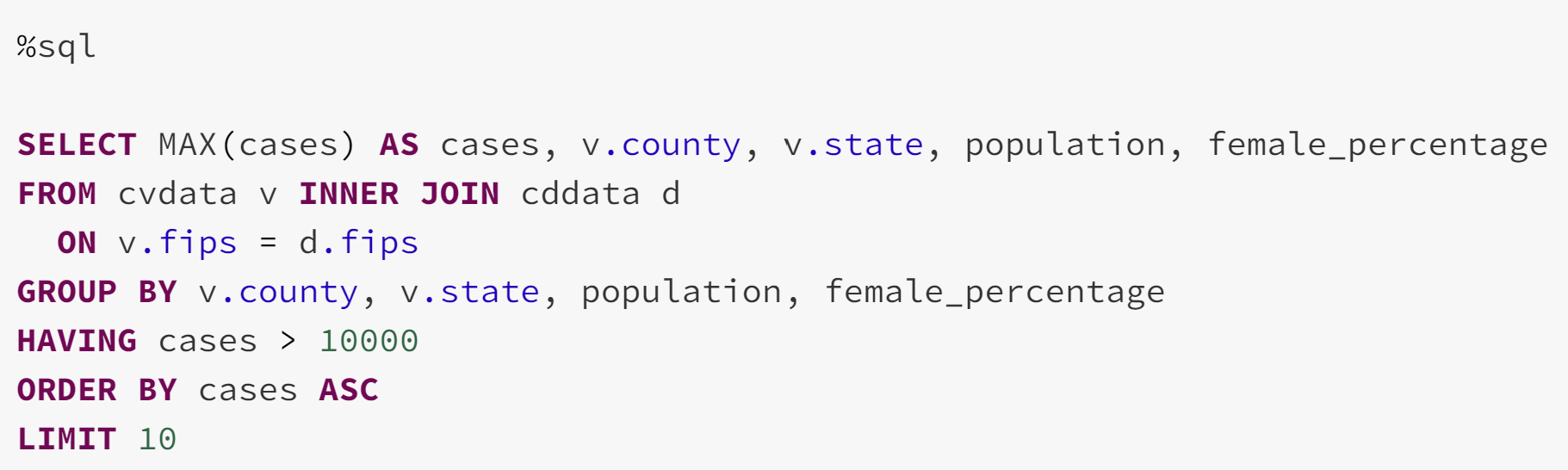
1. **Does female percentage have any correlation with COVID cases?**

For this final question, I simply wanted to analyze if female percentage had any correlation with COVID cases (does a higher female percentage usually mean a higher number of cases?). I just created two pivot tables to compare the highest and lowest amounts of cases in the top counties and their female percentages to see if there would be any interesting trends.





The pivot table shows the counties with the ten highest amount of cases and their corresponding female percentages. They are all around 50%-51%, not showing anything too interesting.



This pivot table shows the lowest ten counties in terms of COVID cases. Again, there is not too much interesting information, but it does show us that Kings, California has a low percentage of females. However, that’s most likely just a coincidence as there doesn’t seem to be too much telling information from either of these tables, an interesting exploration nonetheless.