ANALYSIS ON US CENSUS DATA (2000 – 2010)

by

Fadi Dagher

ID: G37186843

12/15/2020

SEAS 6401 – Data Analytics Found Practicum

Fall 2020

Contents

I. Abstract

The US Census is performed every decade, where US residents fill in and mail back a census form, reporting on information such as their age, gender, race, etc., to be used to by the government to better allocate resources. The data generated during the census is very large and can hold interesting statistics within it. The purpose of this paper is to analyze and visualize that data, extracting and displaying interesting information observed in it, while learning a new distributed computing framework.

II. Introduction

The US Census data covers every resident of the United States' 300 million population and includes different attributes about the individual. As a result, analyzing such data is both resource intensive as well as potentially revealing of interesting trends across the US population. The databricks web-based platform for Apache Spark is a powerful platform for processing and analyzing such large amounts of data, providing automated cluster management, and was utilized throughout this project.

III. Methodology

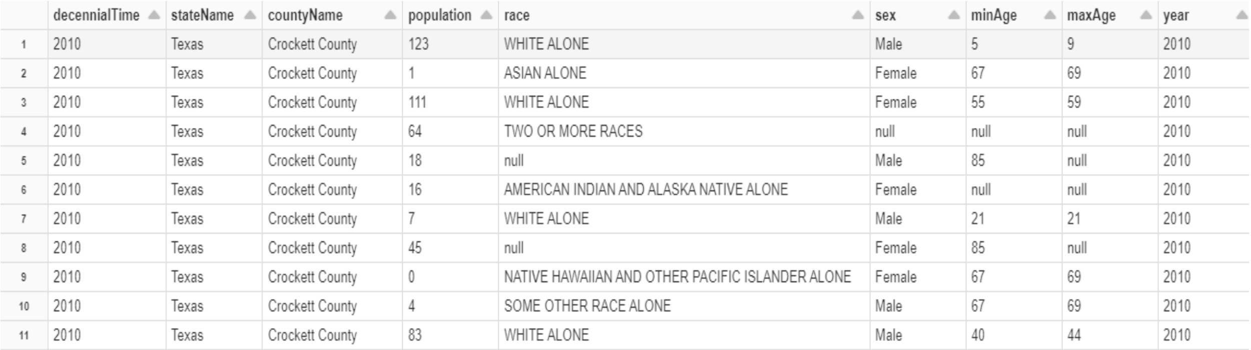
a) Data Identification, Acquisition & Filtering

While the data produced by the US Census is confidential, the US Census Bureau does make available such data without any personal identifiers. The data I retrieved is from Microsoft's Azure Open Datasets Catalog (Microsoft 1). According to Microsoft, "curated open public datasets in Azure Open Datasets are optimized for consumption in machine learning workflows." (Gonlund 2). They provide 2 datasets, by County or by Zip codes, and I chose the latter for this project, for my familiarity with County names more than Zip codes. The dataset is sourced from the United States Census Bureau's Decennial Census Dataset APIs, and is stored in Parquet format, in the East US Azure region. The data is made available through the *azureml.opendataset* package, and is 3.6M rows long. Given its large size, I decided to use Databricks to analyze it, given its distributed computing capabilities, as well as my interest in learning it. Databricks supports several languages (as of this writing): Python, R, Scala, and SQL. I chose Python for data cleansing and aggregation, and SQL for data queries and analysis. Lastly, and for data visualization, I used the native Databricks visualization tools. After collecting the data into a PySpark dataframe, I stored it in the Databricks' DBFS file system for ease of access in the future, as I work on the project. I also complemented the data with a dictionary table of State Name – State Abbreviations, in order to utilize Databricks' map plot function during data visualization phase, which only accepts State name abbreviations (rather than State names). I obtained that table from the USPS website (USPS 3).

Table 1: Code snippet for importing and storing the dataset from Microsoft's AzureML Open Datasets

```
from azureml.opendatasets import UsPopulationCounty
population = UsPopulationCounty()
census_df = population.to_spark_dataframe()
census_df.coalesce(1).write.format("csv").mode('overwrite').option("header",
"true").save("dbfs:/FileStore/Tables/census.csv")
```

Table 2: Sample data from the US Census dataset



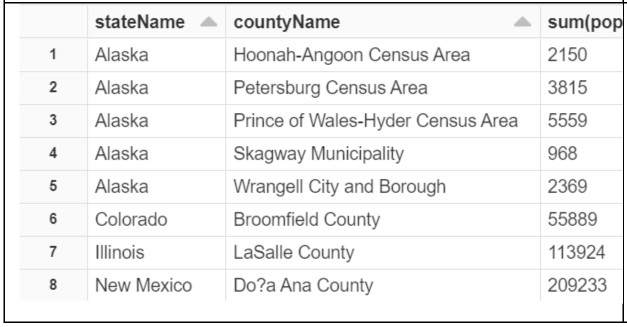
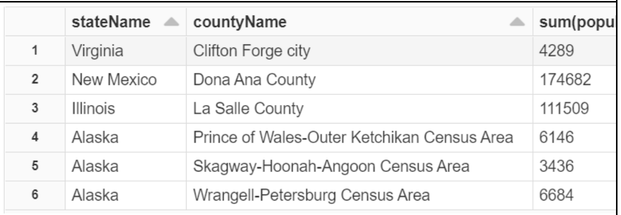
	decennialTime	stateName	countyName	population	race	sex	minAge	maxAge	year
1	2010	Texas	Crockett County	123	WHITE ALONE	Male	5	9	2010
2	2010	Texas	Crockett County	1	ASIAN ALONE	Female	67	69	2010
3	2010	Texas	Crockett County	111	WHITE ALONE	Female	55	59	2010
4	2010	Texas	Crockett County	64	TWO OR MORE RACES	null	null	null	2010
5	2010	Texas	Crockett County	18	null	Male	85	null	2010
6	2010	Texas	Crockett County	16	AMERICAN INDIAN AND ALASKA NATIVE ALONE	Female	null	null	2010
7	2010	Texas	Crockett County	7	WHITE ALONE	Male	21	21	2010
8	2010	Texas	Crockett County	45	null	Female	85	null	2010
9	2010	Texas	Crockett County	0	NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE	Female	67	69	2010
10	2010	Texas	Crockett County	4	SOME OTHER RACE ALONE	Male	67	69	2010
11	2010	Texas	Crockett County	83	WHITE ALONE	Male	40	44	2010

b) Data Validation & Cleansing

As table 2 shows, the dataset has 9 columns in total, with the key being the combination of stateName and countyName. The population column holds the total population count for the given stateName/countyName pair, aggregated based on the remaining columns, such as race, sex, or minAge, and maxAge. I will refer to those last 4 columns as "optional" for ease of reference throughout this paper. The values in those 4 optional columns are either filled, or left blank (NULL), allowing the data analyst to choose the rows of interest, based on the goals of their analysis. For example, for the same stateName/countyName pair, we have multiple rows (i.e. population count) where any of those 4 optional columns can be null, indicating that the population count in that row covers all possible values of that column (i.e. the column/criteria is not broken down during the count). This had to be considered when analyzing the data, to avoid over calculating the population count, and unfortunately was not made clear in the dataset description in Azure.

After obtaining and understanding the data scheme, I moved on to data validation. I started by verifying if the same counties and states are covered in both 2000 and 2010. The results of that check were that the 2000 Census data is missing Puerto Rico state, as well as 8 total counties in other states. On the other hand, the 2010 Census data is missing 6 other counties, as showing in table 3 below. Additional checks were later performed on the data to verity that the decennialTime and year columns were identical (in order to remove redundant columns), and that the state names match between the Microsoft and USPS datasets (in order to join correctly).

Table 3: Discrepancies between the data in 2000 and 2010 census

<table>
<tr><th>Counties missing from 2000 Census (in addition to Puerto Rico)</th><th>Counties missing from 2010 Census</th></tr>
<tr><td>
<table>
<tr><th></th><th>stateName</th><th>countyName</th><th>sum(pop</th></tr>
<tr><td>1</td><td>Alaska</td><td>Hoonah-Angoon Census Area</td><td>2150</td></tr>
<tr><td>2</td><td>Alaska</td><td>Petersburg Census Area</td><td>3815</td></tr>
<tr><td>3</td><td>Alaska</td><td>Prince of Wales-Hyder Census Area</td><td>5559</td></tr>
<tr><td>4</td><td>Alaska</td><td>Skagway Municipality</td><td>968</td></tr>
<tr><td>5</td><td>Alaska</td><td>Wrangell City and Borough</td><td>2369</td></tr>
<tr><td>6</td><td>Colorado</td><td>Broomfield County</td><td>55889</td></tr>
<tr><td>7</td><td>Illinois</td><td>LaSalle County</td><td>113924</td></tr>
<tr><td>8</td><td>New Mexico</td><td>Do?a Ana County</td><td>209233</td></tr>
</table></td><td>
<table>
<tr><th></th><th>stateName</th><th>countyName</th><th>sum(popu</th></tr>
<tr><td>1</td><td>Virginia</td><td>Clifton Forge city</td><td>4289</td></tr>
<tr><td>2</td><td>New Mexico</td><td>Dona Ana County</td><td>174682</td></tr>
<tr><td>3</td><td>Illinois</td><td>La Salle County</td><td>111509</td></tr>
<tr><td>4</td><td>Alaska</td><td>Prince of Wales-Outer Ketchikan Census Area</td><td>6146</td></tr>
<tr><td>5</td><td>Alaska</td><td>Skagway-Hoonah-Angoon Census Area</td><td>3436</td></tr>
<tr><td>6</td><td>Alaska</td><td>Wrangell-Petersburg Census Area</td><td>6684</td></tr>
</table></td></tr>
</table>

After the checks were completed, the Azure dataset and USPS table were merged, the value types for population, minAge, maxAge, and year, were converted from type “string” to type “integer”, and the missing state was removed from the dataset. Lastly, the decennialTime column was dropped to avoid data duplication in the dataset. In addition, the race attribute was removed, by filtering in the rows with race=null, then removing that column.

c) Data Extraction, Aggregation & Representation

To help simplify the analysis of the data between the population count in 2000 and 2010, over any of the other attributes (age, gender, location, etc.), four additional columns were added to the dataset, through the use of User Defined Dunctions (UDFs) as listed below:

1- New columns “population at 2000” and “population at 2010” were added to the table, by aggregating the population count over all other columns in the dataset, separating by the year column, and finally dropping the year column.

2- New columns “countIncrease” and “percentIncrease” were added by using the values from the recently created columns in point 1 above, to calculate the total as well as percent difference between the populations in 2000 and 2010, again for ease of analyzing and visualizing the data in future steps.

3- New column ‘Age Bracket’ was added, by converting the values of minAge and maxAge columns into an age bracket of 10 years width. It’s worth noting that the initial minAge and maxAge brackets were not consistent, where in some cases the brackets would be 5 years of width (ex: 10-14), 3 years width (ex: 22-24), 2, 1, and in other cases it would be open (85+).

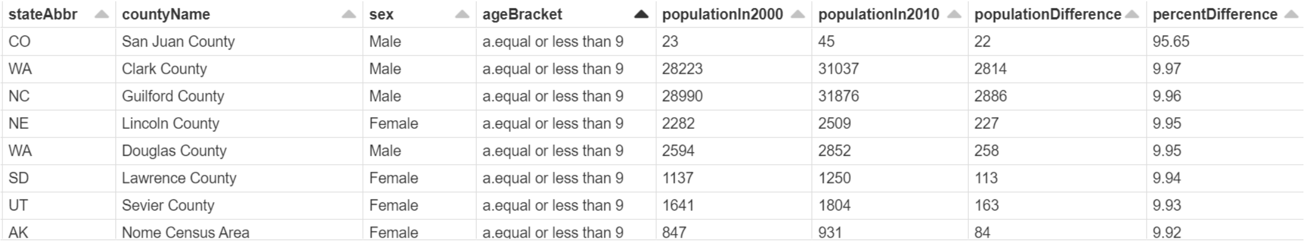
Once that was done, 2 new PySpark dataframes were created, to be directly used during the analysis and visualization phase.

1- The ‘cdfcounty’ dataframe was created from the above described dataframe, by aggregating the data over the county and state fields, then removing the gender and age columns.

2- The ‘cdfstate’ dataframe was created from the cdfcounty dataframe, by aggregating the data over the state field, then removing the county column.

Lastly, three SQL tables were created based on the 3 dataframes above, using the *createOrReplaceTempView()* dataframe function, and were then cached using the *sqlContext.cacheTable function*.

Table 4: Structure and sample data from the three newly create dataframes, to be used in the analysis phase



stateAbbr ▲	countyName ▲	sex ▲	ageBracket ▲	populationIn2000 ▲	populationIn2010 ▲	populationDifference ▲	percentDifference ▲
CO	San Juan County	Male	a.equal or less than 9	23	45	22	95.65
WA	Clark County	Male	a.equal or less than 9	28223	31037	2814	9.97
NC	Guilford County	Male	a.equal or less than 9	28990	31876	2886	9.96
NE	Lincoln County	Female	a.equal or less than 9	2282	2509	227	9.95
WA	Douglas County	Male	a.equal or less than 9	2594	2852	258	9.95
SD	Lawrence County	Female	a.equal or less than 9	1137	1250	113	9.94
UT	Sevier County	Female	a.equal or less than 9	1641	1804	163	9.93
AK	Nome Census Area	Female	a.equal or less than 9	847	931	84	9.92

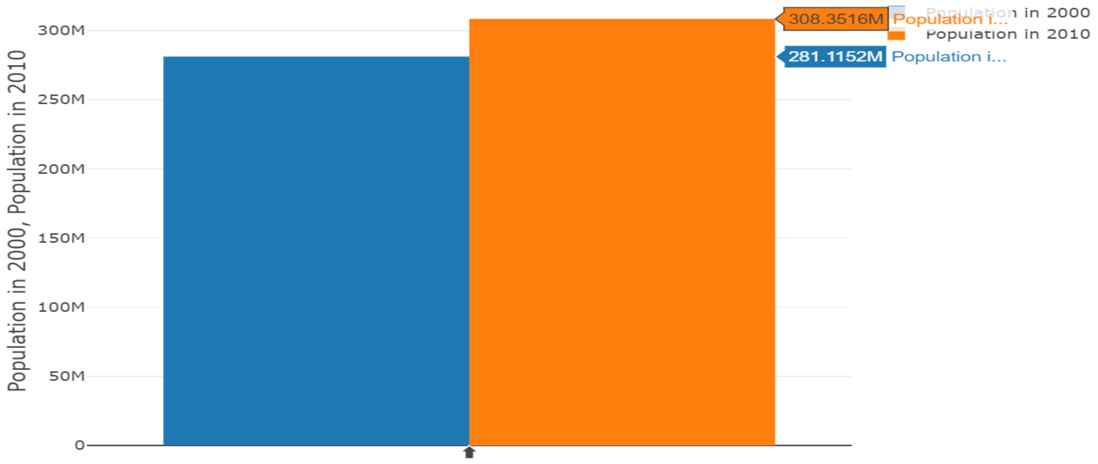
IV. Analysis and Visualization

The analysis and visualization were performed within Databricks, using SQL language to query the data, and the Databricks built-in “display” function to plot the data. The plots were as follows:

a) Population count by Total, Gender, and Age Brackets

The first plot was for the total population in 2000 and 2010, where the analysis revealed that the count increased by roughly 27 million (again this excludes Puerto Rico which was missing from the 2000 dataset provided by MS).

Figure 1: Total population count

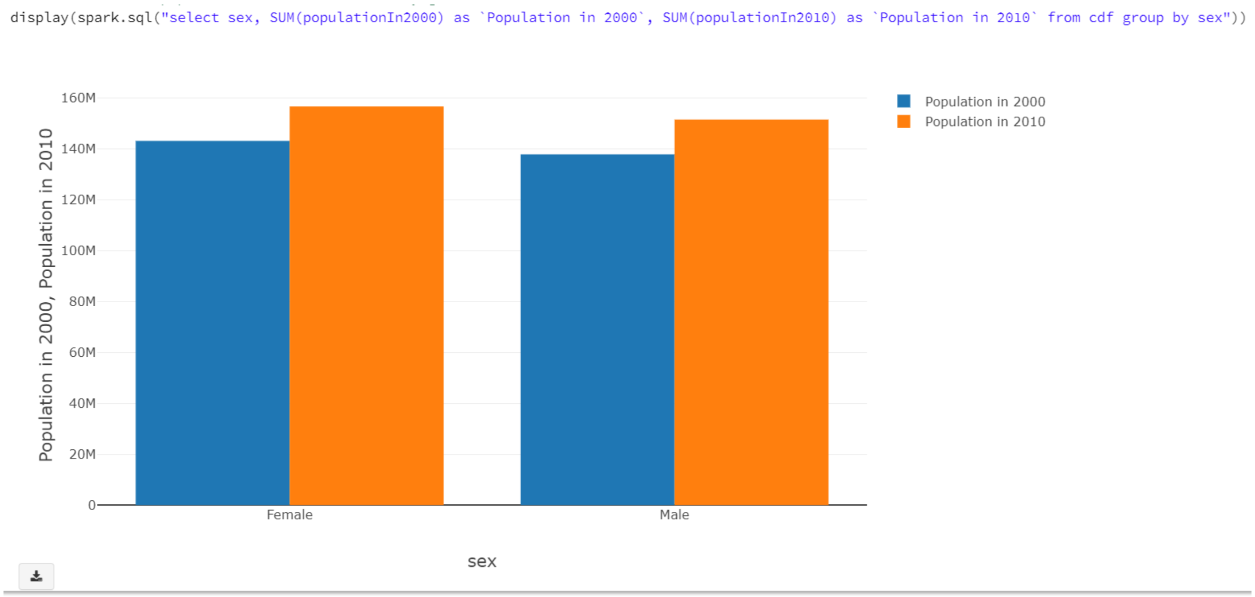


Breaking down the population by gender, we can observe two interesting points:

1- Women count more than men in both census years by roughly 5 million

2- The gap between the count of women and men shrank from around 5.3 million to around 5.2 million

Figure 2: Population count by gender



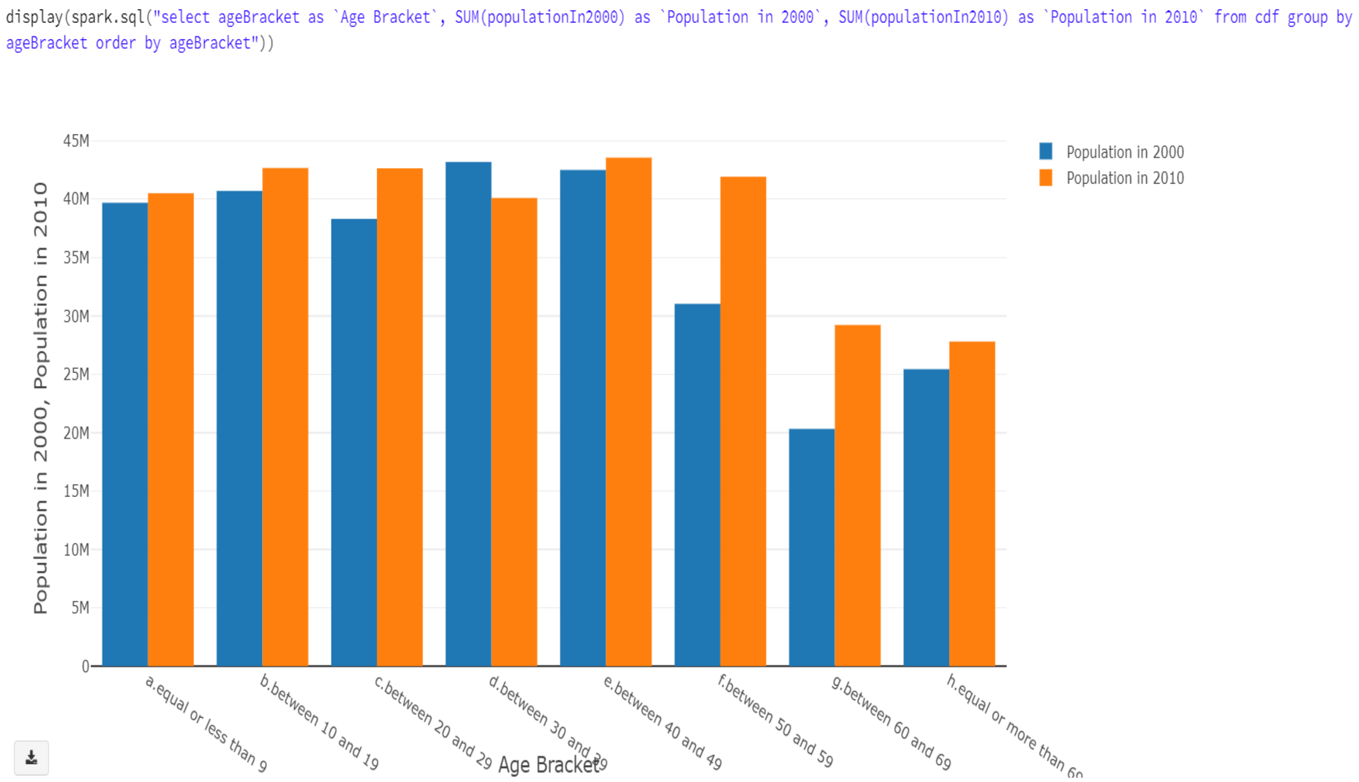
```
display(spark.sql("select sex, SUM(populationIn2000) as `Population in 2000`, SUM(populationIn2010) as `Population in 2010` from cdf group by sex"))
```

The next plot for the population count was broken down by age brackets created in the data aggregation phase (i.e. by decades: 0-10, 11-20, etc.). The plot reveals interesting findings:

1- All age brackets saw a population increase except for one. More on this as follows:
   a. The 50-59 age bracket had the highest total count increase
   b. The 60-69 age bracket had the highest percent increase
   c. The age bracket 70+ saw an increase of over 2 million people
   d. Age bracket 30-39 declined in population count, by around 3 million.

2- There were 31 million people in the 50-59 age bracket in 2000. Ten years later, and as those people move to the 60-69 age bracket, their count decreased to 29.2 million. Population decrease can usually be explained by either immigration out of the country, or death.

Figure 3: Population count by age brackets

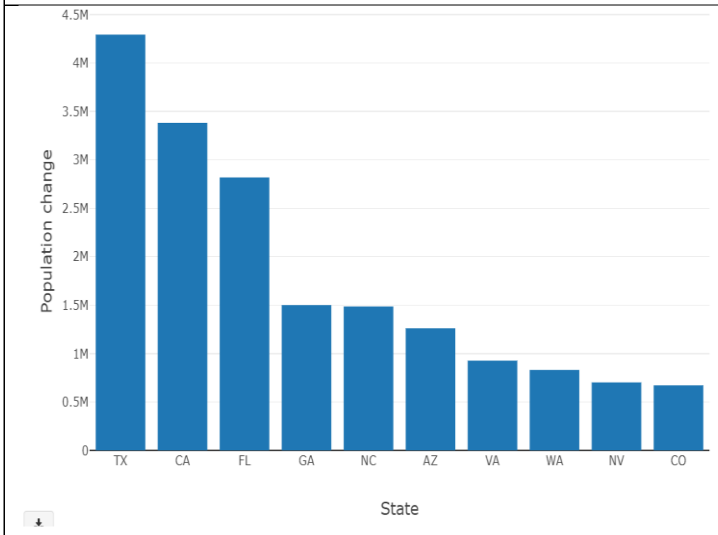


```
display(spark.sql("select ageBracket as `Age Bracket`, SUM(populationIn2000) as `Population in 2000`, SUM(populationIn2010) as `Population in 2010` from cdf group by
ageBracket order by ageBracket"))
```

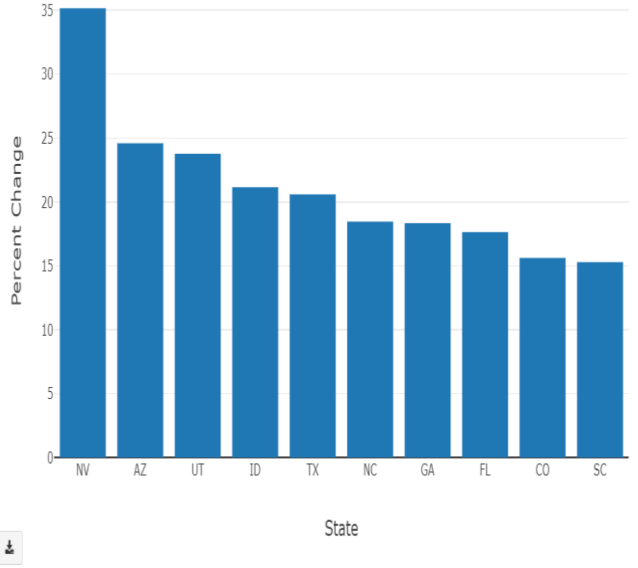
b) Population count by County and State

Texas, California, and Florida had the highest population count increase among the US states. Performing the breakdown by percent (rather than total count) increase, however, shows that Nevada, Arizona, and Utah, are the top 3.

States with highest population count increase

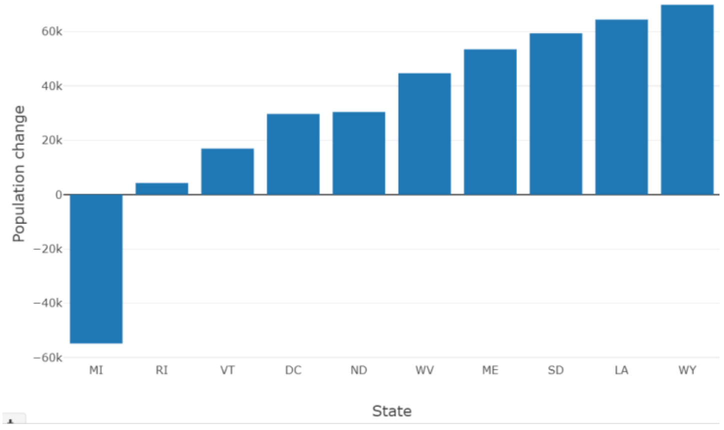


States with highest population percent increase

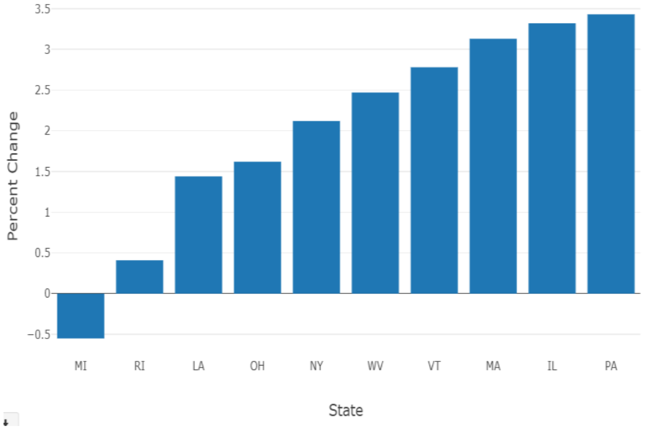


Michigan, Rhode Island, Vermont, and Louisiana were the lowest states in terms of population increase, by count or percent. Michigan is the only state with a negative population increase.

States with lowest population count increase

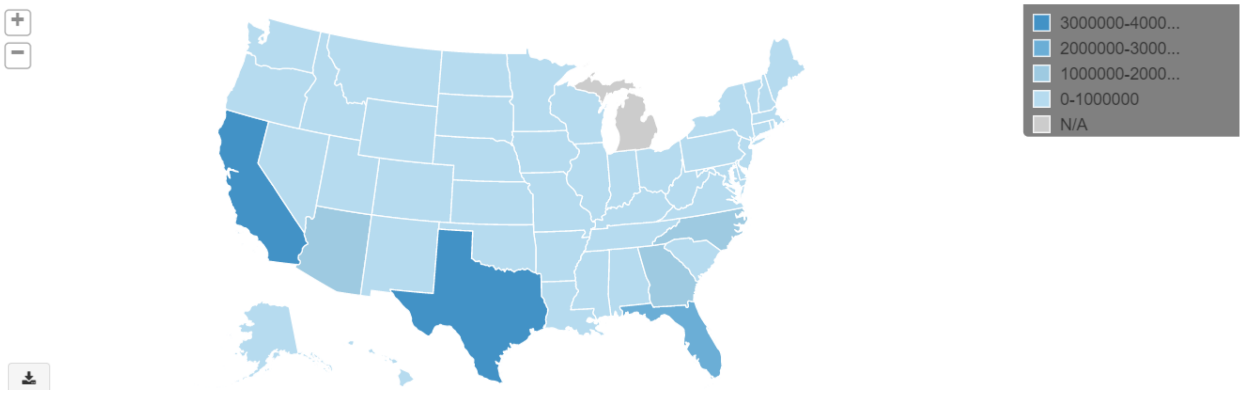


States with lowest population percent increase



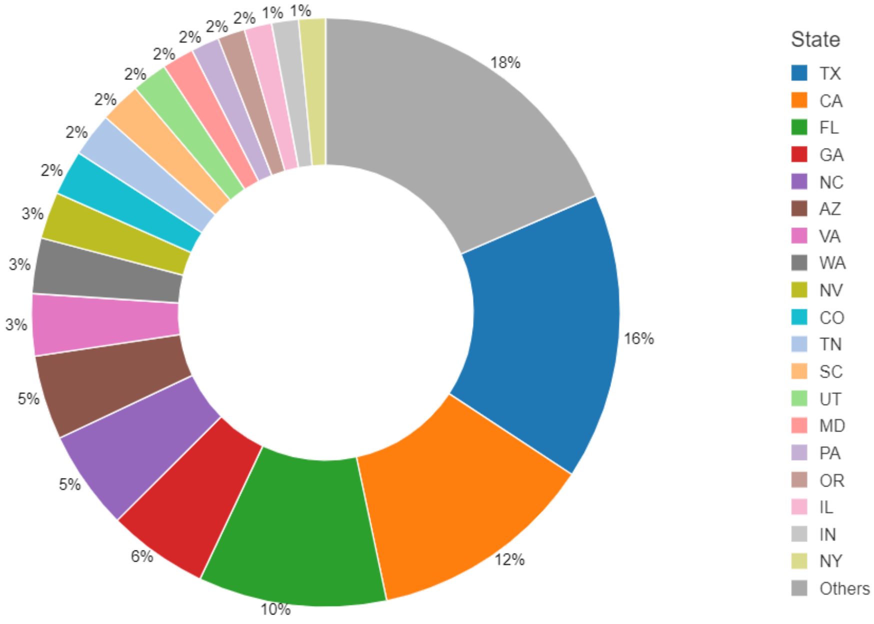
The following is a visualization of population count increase on the US map (the darker the higher the increase)

Figure 4: US map plot for population count increase per state



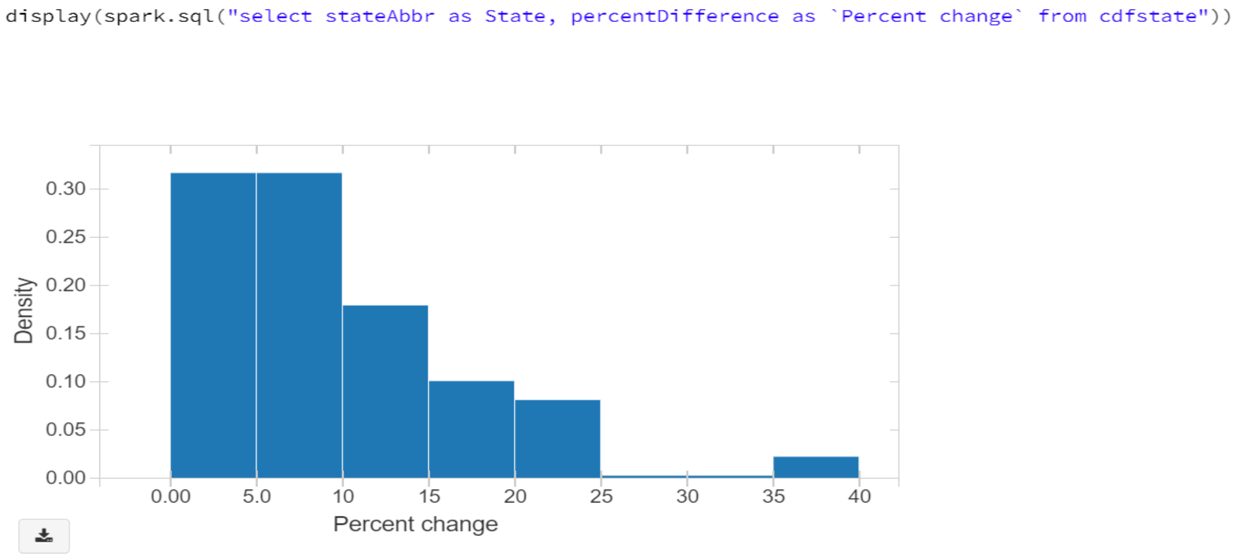
Plotted in a pie chart, we can see that 5 states account for 49% of the total population increase in the US: Texas, California, Florida, Georgia, and North Carolina.

Figure 5: Pie chart of population count per state (percent of total population)



Plotting a histogram of population count change rate per state, we see that the 0-5% and 5-10% brackets are the most common, occurring on around 30% of states, and the highest is the 35-40% rate increase, occurring on 2% of the states (i.e. in 1 state). State names will be shared in following plots.

Figure 6: Histogram for population percent increase per state



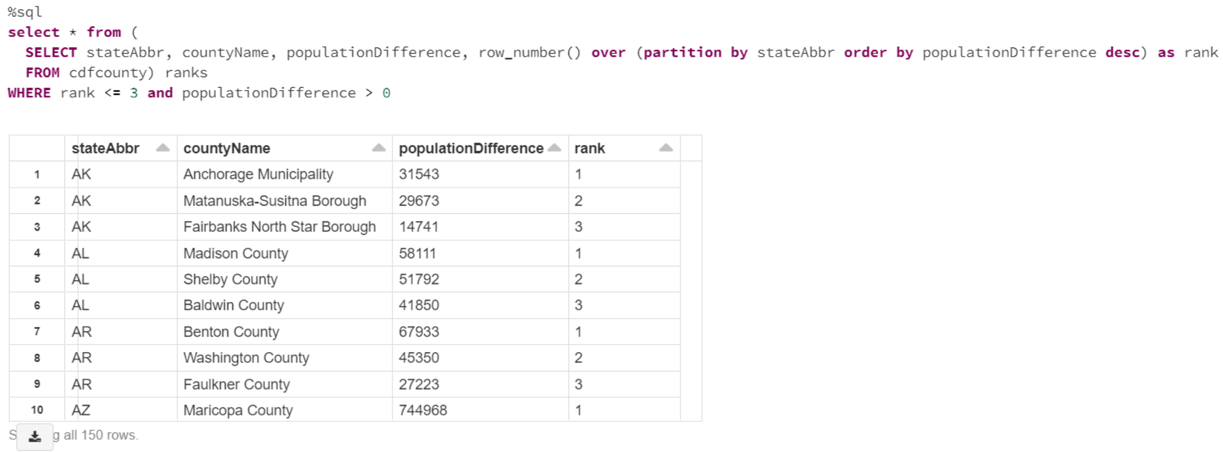
```
display(spark.sql("select stateAbbr as State, percentDifference as `Percent change` from cdfstate"))
```

c) Counties contributing to population increase or decrease

The last part of the analysis was to determine which counties in each state are contributing the most to the population increase of the state, and which ones are contributing to a decrease in population.

The following query shows the top 3 counties in population gain in each state, while the one beneath it shows the counties with most population losses.

Figure 7: Top 3 counties in population gain in each state

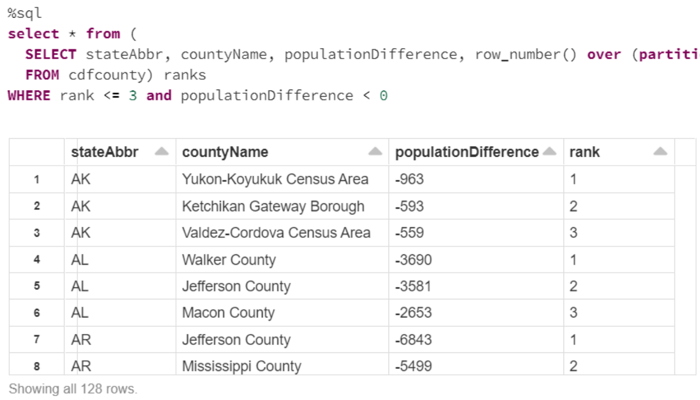


```
%sql
select * from (
  SELECT stateAbbr, countyName, populationDifference, row_number() over (partition by stateAbbr order by populationDifference desc) as rank
  FROM cdfcounty) ranks
WHERE rank <= 3 and populationDifference > 0
```

	stateAbbr	countyName	populationDifference	rank
1	AK	Anchorage Municipality	31543	1
2	AK	Matanuska-Susitna Borough	29673	2
3	AK	Fairbanks North Star Borough	14741	3
4	AL	Madison County	58111	1
5	AL	Shelby County	51792	2
6	AL	Baldwin County	41850	3
7	AR	Benton County	67933	1
8	AR	Washington County	45350	2
9	AR	Faulkner County	27223	3
10	AZ	Maricopa County	744968	1

S g all 150 rows.

Figure 8: Top 3 counties in population decrease in each state



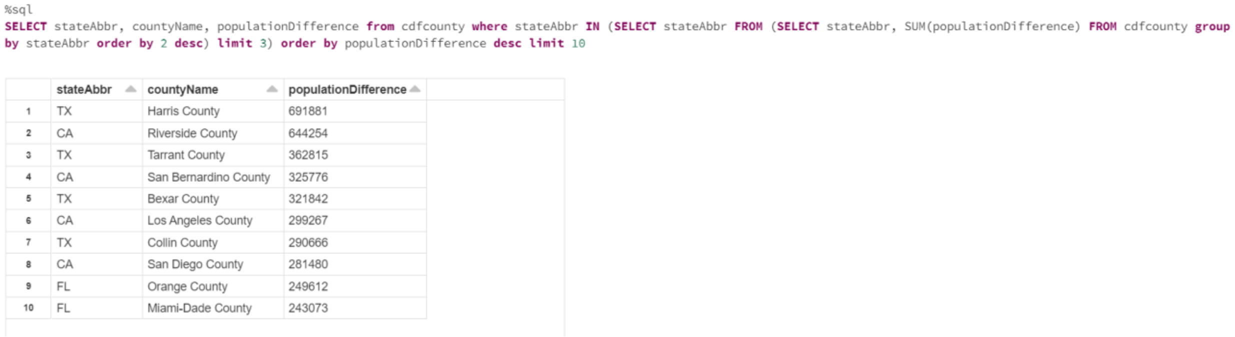
```
%sql
select * from (
  SELECT stateAbbr, countyName, populationDifference, row_number() over (partition by stateAbbr order by populationDifference asc) as rank
  FROM cdfcounty) ranks
WHERE rank <= 3 and populationDifference < 0
```

	stateAbbr	countyName	populationDifference	rank
1	AK	Yukon-Koyukuk Census Area	-963	1
2	AK	Ketchikan Gateway Borough	-593	2
3	AK	Valdez-Cordova Census Area	-559	3
4	AL	Walker County	-3690	1
5	AL	Jefferson County	-3581	2
6	AL	Macon County	-2653	3
7	AR	Jefferson County	-6843	1
8	AR	Mississippi County	-5499	2

Showing all 128 rows.

Below are the top 10 county contributors to population gain in the 3 states with highest population gain, and the figure below it is showing the top 10 county contributors to population loss, in the states with the population loss (i.e. Michigan).

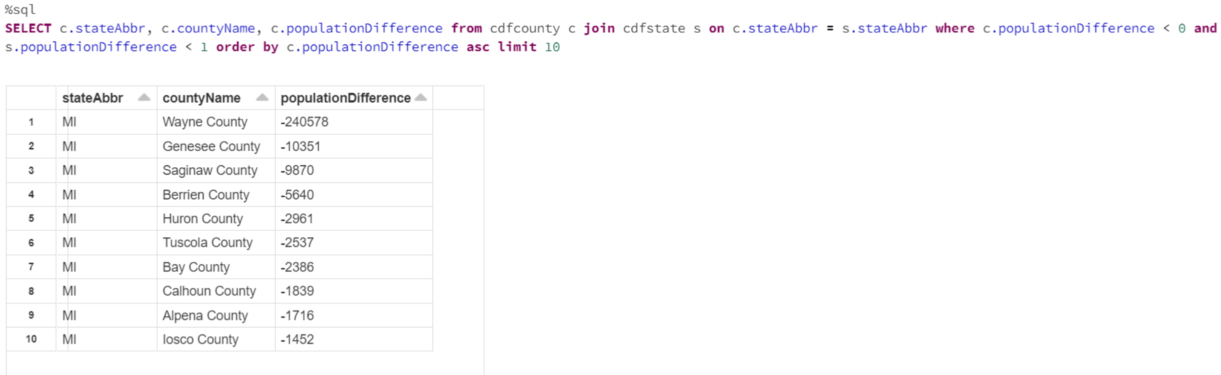
Figure 9: Top 10 county contributors to population gain in the 3 states with highest population gain



```
%sql
SELECT stateAbbr, countyName, populationDifference from cdfcounty where stateAbbr IN (SELECT stateAbbr FROM (SELECT stateAbbr, SUM(populationDifference) FROM cdfcounty group
by stateAbbr order by 2 desc) limit 3) order by populationDifference desc limit 10
```

	stateAbbr	countyName	populationDifference
1	TX	Harris County	691881
2	CA	Riverside County	644254
3	TX	Tarrant County	362815
4	CA	San Bernardino County	325776
5	TX	Bexar County	321842
6	CA	Los Angeles County	299267
7	TX	Collin County	290666
8	CA	San Diego County	281480
9	FL	Orange County	249612
10	FL	Miami-Dade County	243073

Figure 10: Top 10 county contributors to population loss, in the states with the population loss (i.e. Michigan)



```
%sql
SELECT c.stateAbbr, c.countyName, c.populationDifference from cdfcounty c join cdfstate s on c.stateAbbr = s.stateAbbr where c.populationDifference < 0 and
s.populationDifference < 1 order by c.populationDifference asc limit 10
```

	stateAbbr	countyName	populationDifference
1	MI	Wayne County	-240578
2	MI	Genesee County	-10351
3	MI	Saginaw County	-9870
4	MI	Berrien County	-5640
5	MI	Huron County	-2961
6	MI	Tuscola County	-2537
7	MI	Bay County	-2386
8	MI	Calhoun County	-1839
9	MI	Alpena County	-1716
10	MI	Iosco County	-1452

V. Conclusion

In conclusion, the United States has enjoyed a population increase of roughly 10% from 2000 to 2010, with the population increase split between men and women, and distributed across all the age brackets, except for 30-39 which saw a decrease, as well as all states, except for Michigan which saw a population decrease. Wayne County was the biggest contributor to the population decrease in Michigan, while Harris County, Riverside County, and Orange County were the biggest contributors to the population gains in Texas, California, and Florida, respectively (those 3 states being the top 3 in population increase country wide). Immigration/naturalization contributed to the population increase, as the data shows that some of the age brackets' population in 2000, increased 10 years later (when looking at the corresponding age bracket).

Databricks is a great tool for ingesting and analyzing large amount of data, given its distributed architecture. Utilizing spark dataframes and user defined functions helps take advantage of this architecture. One should be mindful that transformations such as filtering a dataset are lazy. As a result, and while it is very quick to build dataframes on top of one another or to utilize such transformation functions, the performance hit of using those functions is sensed when it comes time to taking an action on the data in those dataframes (such as to count or display the tuples). Caching helps reduce the performance hit, however.

VI. References

[1] Microsoft, “US Population by County” source: https://azure.microsoft.com/en-us/services/open-datasets/catalog/us-decennial-census-county/

[2] Gronlund, C. et al “What are Azure Open Datasets and how can you use them?” May 2020, source: https://docs.microsoft.com/en-us/azure/open-datasets/overview-what-are-open-datasets

[3] USPS, “State Abbreviations” source: https://about.usps.com/who-we-are/postal-history/state-abbreviations.htm

[4] Grant Suneson, “how many people died the year you were born”, source: https://eu.usatoday.com/story/money/2020/06/11/how-many-people-died-the-year-you-were-born/111928450/