Big Data Analysis on American Community Census Blocks

GOUTHAM RAJ GANESAN
MONISH PHATARPEKAR
NANJESH RAMESH
SOWRABH SHANTHANAND
MSIS Graduate Students
California State University, Los Angeles

Abstract: This project will demonstrate the usage of Hadoop, MapReduce, and Hive on big data. We will apply the knowledge learned during the lecture, extensive researches and development of HiveQL in order to generate data and visualize it on Power BI, Tableau and 3D maps. We are using the Survey Data of the Census Blocks of American Communities as the foundation to generate results. Fundamentals of this project include a report paper, a tutorial on the queries, and one group presentation.

URL: https://www.kaggle.com/safegraph/census-block-group-american-community-survey-data
Dataset size: 8GB
Cluster version: Oracle Cloud 4.2
No of nodes: 6
Memory size: 180GB
CPU Speed: 2.2 GHz

1. Introduction

Based on the list of data provided by our instructor, we have done some researches and exclusively decided which data we are using for this project. We are going to manipulate and filter the datasets below following with step:

- American Community Census Blocks; data size is 2GB.
- Cleaning down the information to have a detailed comparison between the years.
- From the dataset, sorted out the economic characteristics such as Family Income, Poverty Status, Per-Capita Income and so on.
- The tools we are using is HiveQL, Putty, Terminal, Oracle Cloud, Tableau, Excel 3D maps and Power BI.

2. Manipulating Datasets

2.1. Tools and Data Preprocessing

- We extracted our Community Census data from the corresponding website in .csv format: Census Block Group American Community Data from Kaggle.com.

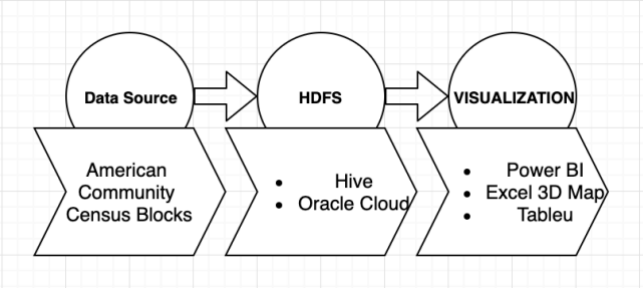

Figure 1: Data Processing

- We used basic commands to connect to the data source such as pscp, mkdir, and also uploaded our data to HDFS and used Hive queries to create external tables on the .csv data.
- Then, we used Hive queries to select the desired data from the external table and filtered out unimportant data. (cleaned the data)
- We used hive queries to analyse the data.
- Lastly, we used Power BI, Tableau and Excel 3D maps to reproduce the selected data in the form of information by generating the appropriate graphs and maps.

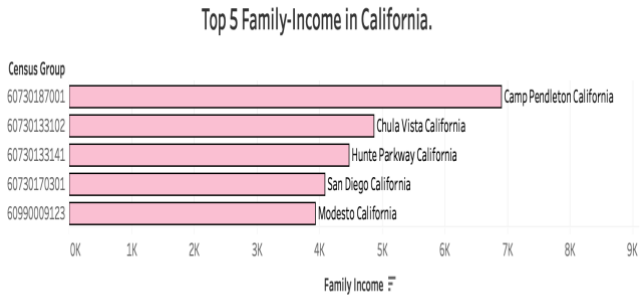

Figure 2: Top 5 Family-Income in California

The above graph shows the Census Blocks in California with highest Family-Income. The census blocks across Camp Pendleton have recorded the family-income of $6.9k.

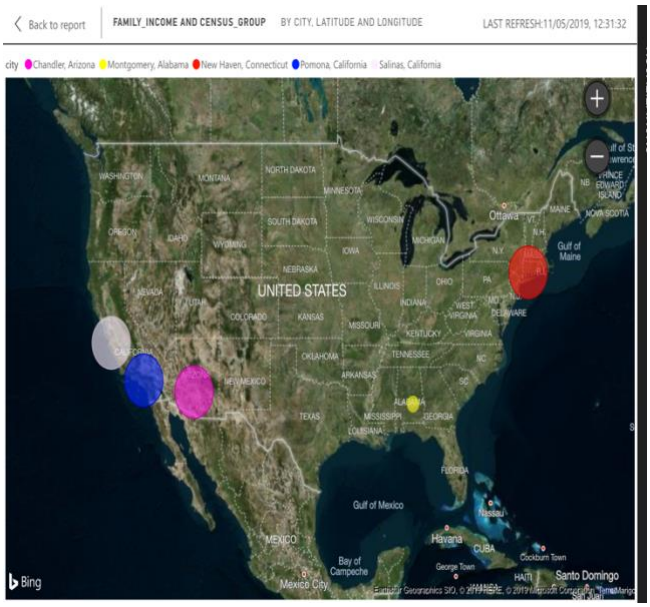

Figure 3: Bottom 5 Family-Income in California

In the above Map Visualization we have represented the Census Blocks with low family-income. The Census Blocks in Montgomery City, Alabama State (yellow point on the map) has recorded low family-income.

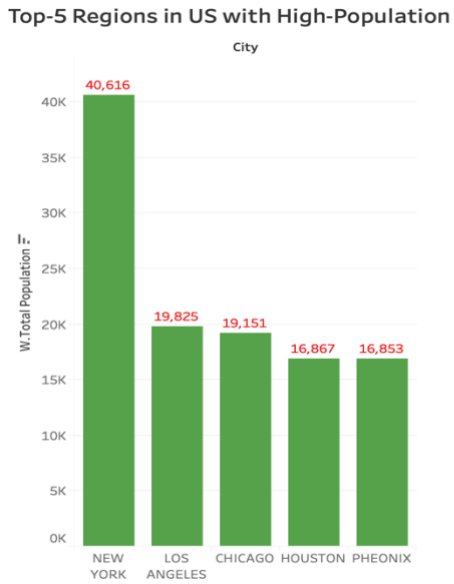

Figure 4: Highly Populated Regions

The above bar graph depicts the highly populated census blocks in the United States. And it is observed that the blocks in the City of New York is densely populated with 40,616 people, followed by Los Angeles which has a population of 19,825.

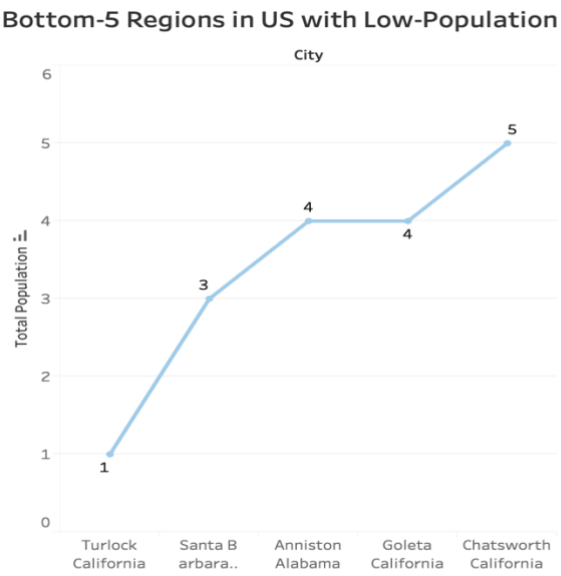

Figure 5: Low Populated Regions

The above line graph represents the low populated census blocks in the United States. And it is observed that the blocks in the City of Turlock has less people per square area which is further followed by Goleta and Chatsworth.

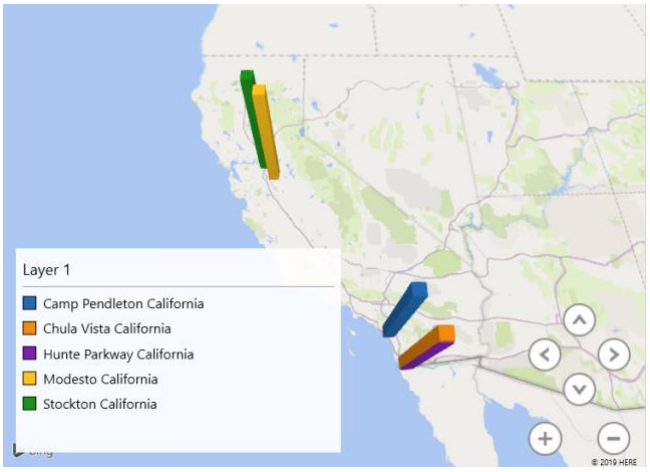

Figure 6: Regions having higher Health-Insurance Coverage

In the above 3D-Map representation we have plotted the regions in California with greater Health insurance Coverage. It is observed that Modesto, Stockton, Hunte Park, Chula Vista and Camp Pendleton have people whose health insurance coverage is significantly high.

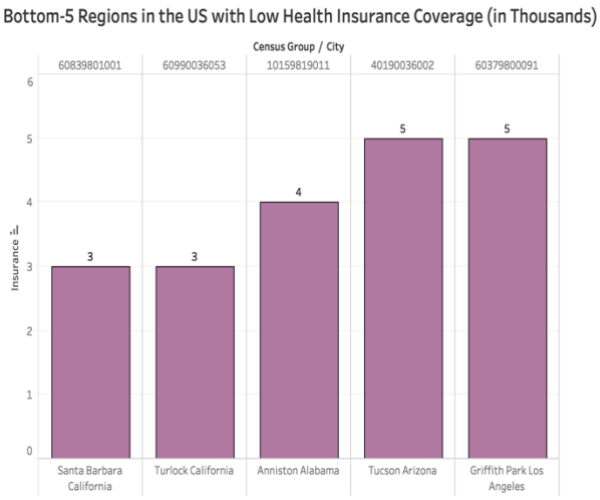

Figure 7: Regions having lower Health-Insurance Coverage

The above bar graphs represents the regions in which the blocks have lower health insurance coverage. We noticed that the city of Santa Barbara and Turlock had blocks with lower insurance coverage followed by Anniston and Tucson.

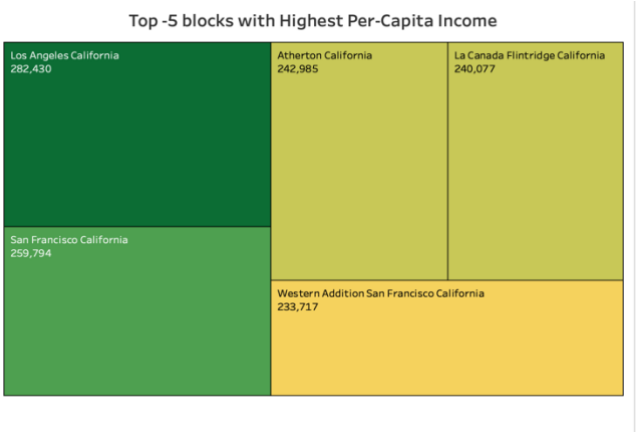

Figure 8: Regions having Highest-Per Capita Income

This is a tree-map representation wherein the larger the size and darker the color signifies a greater value and vice versa for the smaller value. As per the above visualization the City of Los Angeles has the highest per-capita income of $282,430 followed by San Francisco $259,794.

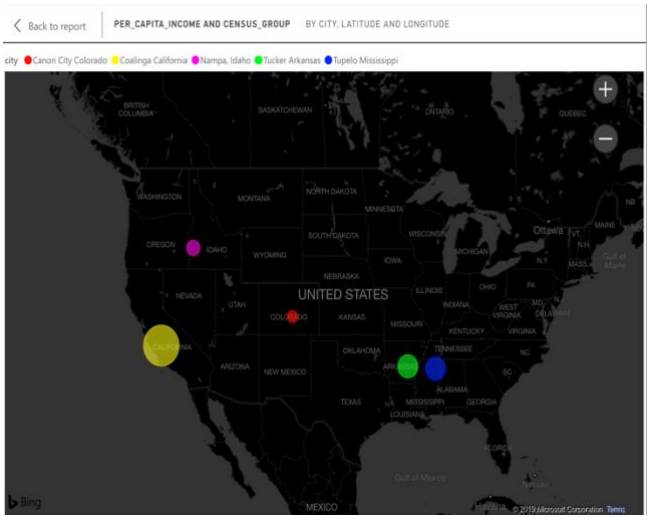

Figure 9: Regions having Lowest-Per Capita Income

In the map representation we can observe the regions with low Per-Capita income. Canon City in Colorado State (red color) has the lowest average earnings per square area followed by Nampa city in Idaho and Tucker city in Arkansas state.

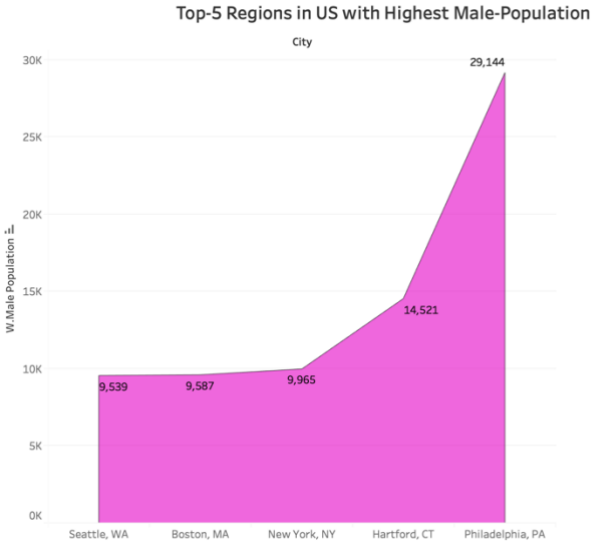

Figure 10: Regions having Highest Male-Population

The above representation is called area under the curve which is giving us a clear insight about the regions with greater male population. It is observed that the city of Philadelphia has blocks with 29,144 males followed by Hartford (14,521 males), New York (9,965 males) and so on.

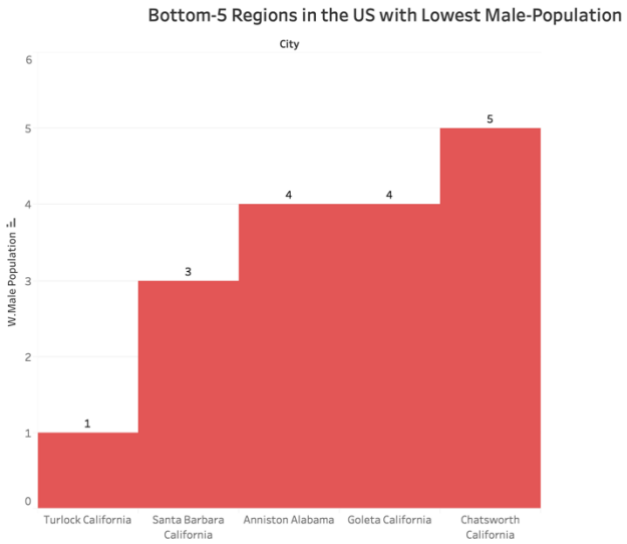

Figure 11: Regions having Lowest Male-Population

The above representation is called step representation which is giving us a clear picture about the regions with lower male population. It is observed that the city of Turlock in California has blocks with fewer male population followed by Santa Barbara, Anniston in Alabama and so on.

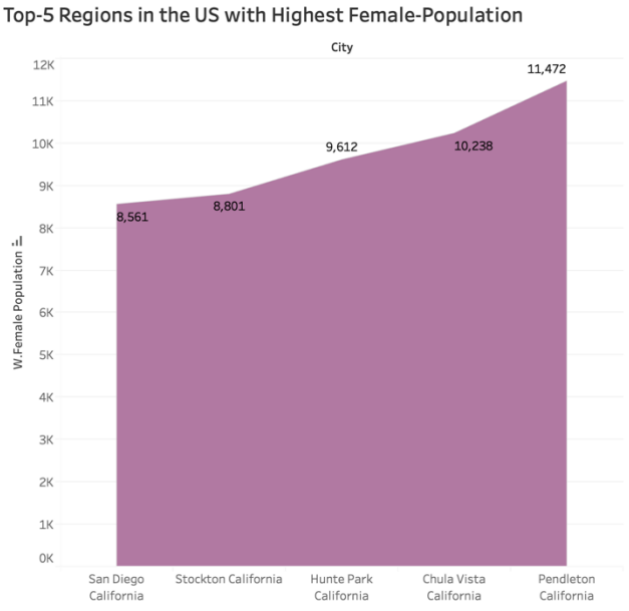

Figure 12: Regions having Highest Female-Population

The above area under the curve representation shows the region-wise count for highest female population. It is evident that Pendleton city is having the highest female population followed by other cities such as Chula Vista, Hunte Park, Stockton and San Diego.

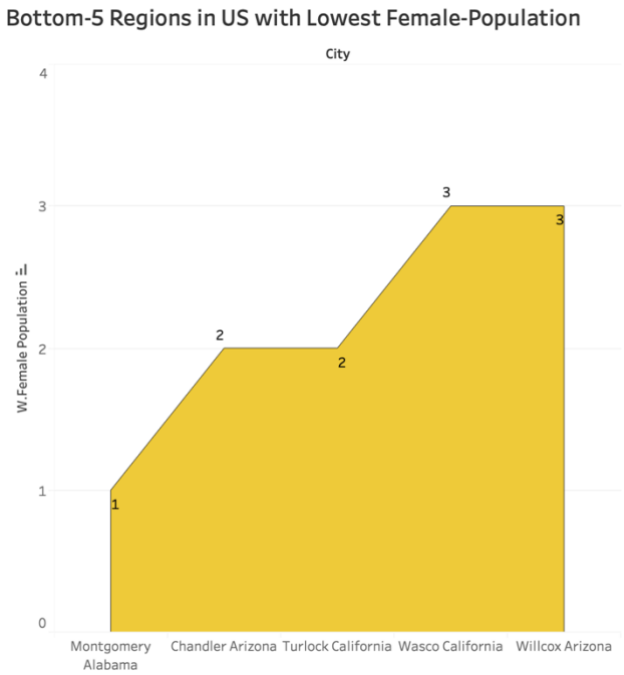

Figure 13: Regions having Lowest Female-Population

The above area under the curve representation shows the region-wise count for lowest female population. It is evident that Montgomery city in Alabama is having the lowest female population followed by other cities such as Chandler in Arizona, Turlock in California and so on.

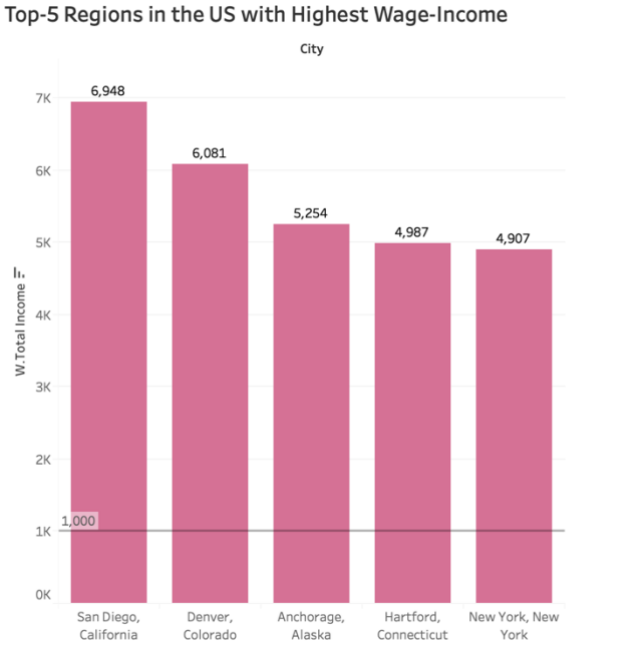

Figure 14: Regions having Highest Wage-Income

The above bar graph represents the Wage-Income corresponding to different regions. It is observed that blocks in San Diego have the highest wage income of $6948 followed by Denver, Colorado with a wage income of $6081.

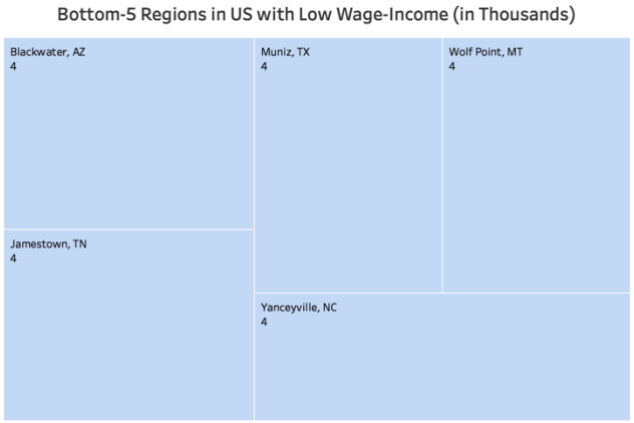

Figure 15: Regions having Lowest Wage-Income

The above tree-map representation depicts the regions with lower wage incomes. It is observed that the cities such as Blackwater(AZ), Muniz(TX), Wolf Point(MT), Jamestown(TN) and Yanceyville(NC) have the lowest wage income.

3. Summary

- We successfully used many tools learned in class such as HiveQL, Oracle Cloud, and Tableau to use and manipulate data.
- Los Angeles has blocks with highest income.
- We can relate the regions with low health insurance coverage and low per-capita income to economic characteristics such as poverty status.
- The city of New York has blocks which are densely populated thereby we could predict the blocks in this city would have more traffic.

4. Github URL

https://github.com/nanjeshgowda/Big-Data-Analysis-on-American-Census-Community-Blocks

5. Reference

https://www.kaggle.com/safegraph/census-block-group-american-community-survey-data

http://proximityone.com/geo_blockgroups.htm

https://en.wikipedia.org/wiki/Census_block_group