# CIS5200 Term Project Tutorial

**Authors:** **Nanjesh Mandya Ramesh, Monish Phatarpekar, Goutham Raj Ganesan and Sowrabh Shanthanand**

**Instructor: Jongwook Woo**

**Date: 05/05/2019**

# Lab Tutorial

nmandya@calstatela.edu

mphatar@calstatela.edu

gganesa@calstatela.edu

sshanth@calstatela.edu

05/05/2019

## Data Analysis of Census Block Group American Community Survey Data

## Objectives

In this hands-on lab, you will learn how to:

- Uploaded the dataset for US Census Group Survey
- HiveQL queries to perform the analysis
- Visualization in Power BI, Tableau and 3D map

## Platform Spec

- Oracle Big Data Compute Edition
- CPU Speed: 2.195 GHz
- # of CPU cores: 4 cores, 1 Socket
- # of nodes: 6 nodes
- Total Memory Size: 32GB

## INTRODUCTION

A Census Block Group (CBG) is the most granular level the US Census Bureau reports data on, and covers ~1500 households.

We include all demographic data from the American Community Survey (2016) 5-year estimate on the Census Block Group level.

The data includes

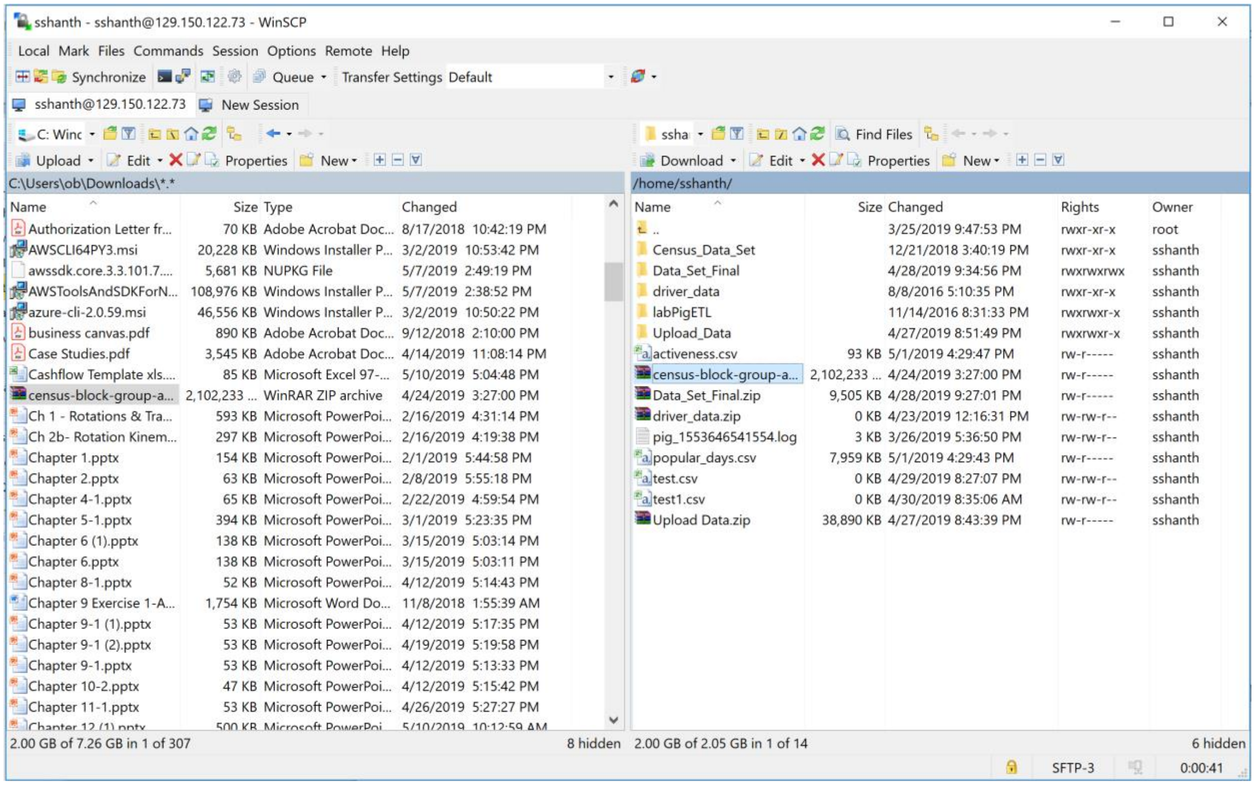
- All census block group boundaries formatted as a GeoJSON file.
- Census attribute tables identified by their Census table ID
- Metadata mapping attribute names to a table ID, census block groups to cities and counties, and census block groups to geographic statistics such as percentage land and water.

# Step 1: Upload Dataset to server

**Server IP:** 129.150.122.73

- Download the dataset from the url: [https://www.kaggle.com/safegraph/census-block-group-american-community-survey-data](https://www.kaggle.com/safegraph/census-block-group-american-community-survey-data)

- Use the software WinSCP to upload the dataset safegraph_open_census_data.zip of size 2GB from local system to oracle server.


sshanth - sshanth@129.150.122.73 - WinSCP
Local Mark Files Commands Session Options Remote Help
Synchronize
Queue
Transfer Settings Default
sshanth@129.150.122.73
New Session
C: Winc
Upload
Edit
Properties
New
ssha
Find Files
Download
SFTP-3
0:00:41

C:\Users\ob\Downloads\*.*

| Name | Size | Type | Changed |
|---|---|---|---|
| Authorization Letter fr... | 70 KB | Adobe Acrobat Doc... | 8/17/2018 10:42:19 PM |
| AWSCLI64PY3.msi | 20,228 KB | Windows Installer P... | 3/2/2019 10:53:42 PM |
| awssdk.core.3.3.101.7.... | 5,681 KB | NUPKG File | 5/7/2019 2:49:19 PM |
| AWSToolsAndSDKForN... | 108,976 KB | Windows Installer P... | 5/7/2019 2:38:52 PM |
| azure-cli-2.0.59.msi | 46,556 KB | Windows Installer P... | 3/2/2019 10:50:22 PM |
| business canvas.pdf | 890 KB | Adobe Acrobat Doc... | 9/12/2018 2:10:00 PM |
| Case Studies.pdf | 3,545 KB | Adobe Acrobat Doc... | 4/14/2019 11:08:14 PM |
| Cashflow Template xls.... | 85 KB | Microsoft Excel 97-... | 5/10/2019 5:04:48 PM |
| census-block-group-a... | 2,102,233 ... | WinRAR ZIP archive | 4/24/2019 3:27:00 PM |
| Ch 1 - Rotations & Tra... | 593 KB | Microsoft PowerPoi... | 2/16/2019 4:31:14 PM |
| Ch 2b- Rotation Kinem... | 297 KB | Microsoft PowerPoi... | 2/16/2019 4:19:38 PM |
| Chapter 1.pptx | 154 KB | Microsoft PowerPoi... | 2/1/2019 5:44:58 PM |
| Chapter 2.pptx | 63 KB | Microsoft PowerPoi... | 2/8/2019 5:55:18 PM |
| Chapter 4-1.pptx | 65 KB | Microsoft PowerPoi... | 2/22/2019 4:59:54 PM |
| Chapter 5-1.pptx | 394 KB | Microsoft PowerPoi... | 3/1/2019 5:23:35 PM |
| Chapter 6 (1).pptx | 138 KB | Microsoft PowerPoi... | 3/15/2019 5:03:14 PM |
| Chapter 6.pptx | 138 KB | Microsoft PowerPoi... | 3/15/2019 5:03:11 PM |
| Chapter 8-1.pptx | 52 KB | Microsoft PowerPoi... | 4/12/2019 5:14:43 PM |
| Chapter 9 Exercise 1-A... | 1,754 KB | Microsoft Word Do... | 11/8/2018 1:55:39 AM |
| Chapter 9-1 (1).pptx | 53 KB | Microsoft PowerPoi... | 4/12/2019 5:17:35 PM |
| Chapter 9-1 (2).pptx | 53 KB | Microsoft PowerPoi... | 4/19/2019 5:19:58 PM |
| Chapter 9-1.pptx | 53 KB | Microsoft PowerPoi... | 4/12/2019 5:13:33 PM |
| Chapter 10-2.pptx | 47 KB | Microsoft PowerPoi... | 4/12/2019 5:15:42 PM |
| Chapter 11-1.pptx | 53 KB | Microsoft PowerPoi... | 4/26/2019 5:27:27 PM |
| Chapter 12 (1).pptx | 500 KB | Microsoft PowerPoi... | 5/10/2019 10:12:59 AM |

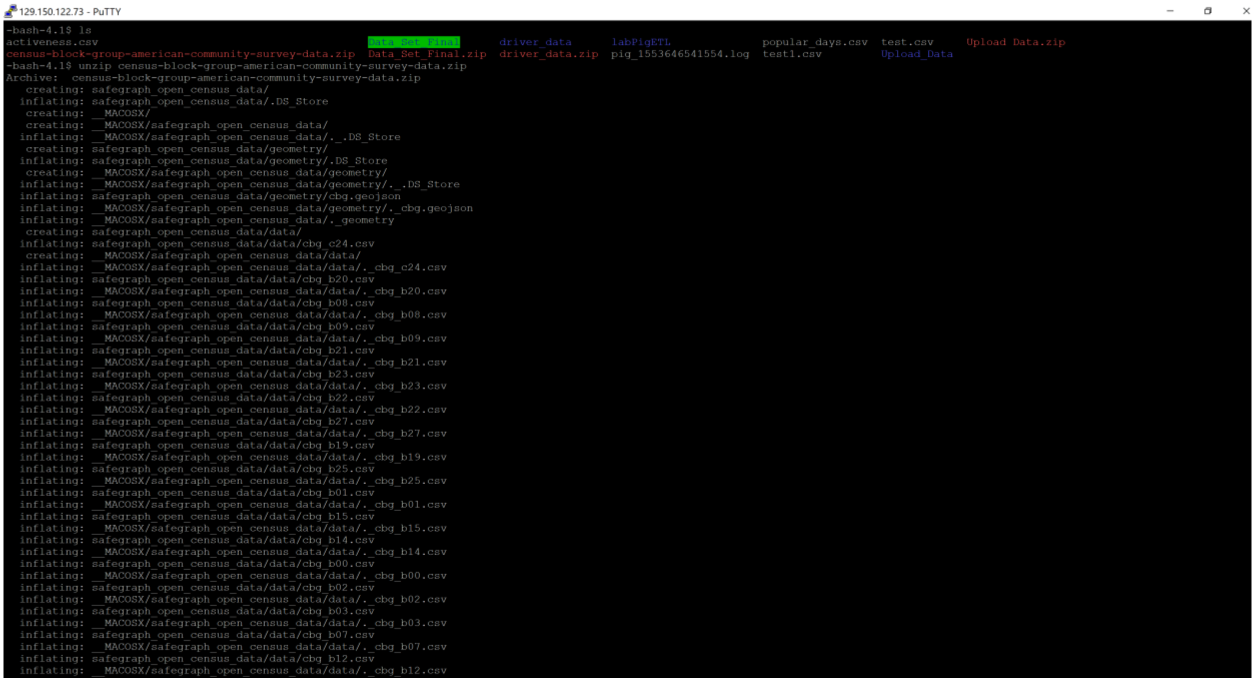
2.00 GB of 7.26 GB in 1 of 307 8 hidden

/home/sshanth/

| Name | Size | Changed | Rights | Owner |
|---|---|---|---|---|
| .. | | 3/25/2019 9:47:53 PM | rwxr-xr-x | root |
| Census_Data_Set | | 12/21/2018 3:40:19 PM | rwxr-xr-x | sshanth |
| Data_Set_Final | | 4/28/2019 9:34:56 PM | rwxrwxrwx | sshanth |
| driver_data | | 8/8/2016 5:10:35 PM | rwxr-xr-x | sshanth |
| labPigETL | | 11/14/2016 8:31:33 PM | rwxrwxr-x | sshanth |
| Upload_Data | | 4/27/2019 8:51:49 PM | rwxrwxr-x | sshanth |
| activeness.csv | 93 KB | 5/1/2019 4:29:47 PM | rw-r----- | sshanth |
| census-block-group-a... | 2,102,233 ... | 4/24/2019 3:27:00 PM | rw-r----- | sshanth |
| Data_Set_Final.zip | 9,505 KB | 4/28/2019 9:27:01 PM | rw-r----- | sshanth |
| driver_data.zip | 0 KB | 4/23/2019 12:16:31 PM | rw-rw-r-- | sshanth |
| pig_1553646541554.log | 3 KB | 3/26/2019 5:36:50 PM | rw-rw-r-- | sshanth |
| popular_days.csv | 7,959 KB | 5/1/2019 4:29:43 PM | rw-r----- | sshanth |
| test.csv | 0 KB | 4/29/2019 8:27:07 PM | rw-rw-r-- | sshanth |
| test1.csv | 0 KB | 4/30/2019 8:35:06 AM | rw-rw-r-- | sshanth |
| Upload Data.zip | 38,890 KB | 4/27/2019 8:43:39 PM | rw-r----- | sshanth |

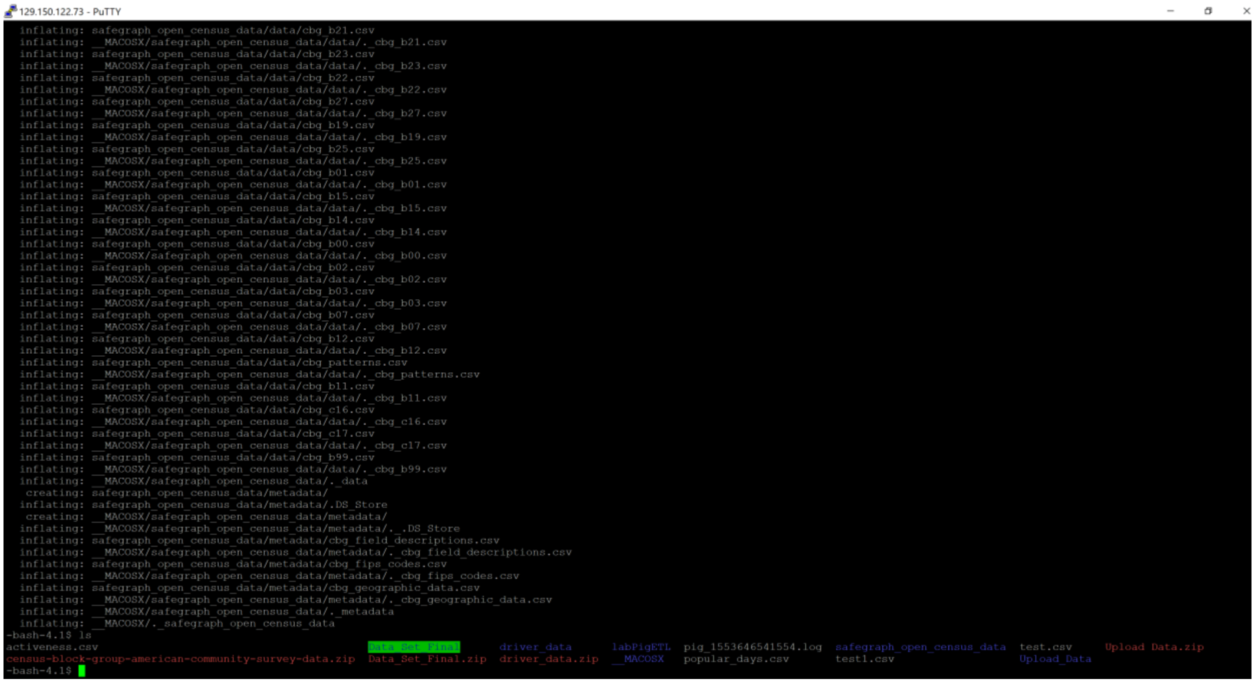
2.00 GB of 2.05 GB in 1 of 14 6 hidden

- Use the command **unzip census-block-group-american-community-survey-data.**zip to extract the files as a csv format.



```
-bash-4.1$ ls
activeness.csv                                         Data_Set_Final      driver_data      labPigETL              popular_days.csv  test.csv      Upload Data.zip
census-block-group-american-community-survey-data.zip  Data_Set_Final.zip  driver_data.zip  pig_1553646541554.log  test1.csv         Upload_Data
-bash-4.1$ unzip census-block-group-american-community-survey-data.zip
Archive:  census-block-group-american-community-survey-data.zip
   creating: safegraph_open_census_data/
  inflating: safegraph_open_census_data/.DS_Store
   creating: __MACOSX/
   creating: __MACOSX/safegraph_open_census_data/
  inflating: __MACOSX/safegraph_open_census_data/._.DS_Store
   creating: safegraph_open_census_data/geometry/
  inflating: safegraph_open_census_data/geometry/.DS_Store
   creating: __MACOSX/safegraph_open_census_data/geometry/
  inflating: __MACOSX/safegraph_open_census_data/geometry/._.DS_Store
  inflating: safegraph_open_census_data/geometry/cbg.geojson
  inflating: __MACOSX/safegraph_open_census_data/geometry/._cbg.geojson
  inflating: __MACOSX/safegraph_open_census_data/._geometry
   creating: safegraph_open_census_data/data/
  inflating: safegraph_open_census_data/data/cbg_c24.csv
   creating: __MACOSX/safegraph_open_census_data/data/
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_c24.csv
  inflating: safegraph_open_census_data/data/cbg_b20.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b20.csv
  inflating: safegraph_open_census_data/data/cbg_b08.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b08.csv
  inflating: safegraph_open_census_data/data/cbg_b09.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b09.csv
  inflating: safegraph_open_census_data/data/cbg_b21.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b21.csv
  inflating: safegraph_open_census_data/data/cbg_b23.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b23.csv
  inflating: safegraph_open_census_data/data/cbg_b22.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b22.csv
  inflating: safegraph_open_census_data/data/cbg_b27.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b27.csv
  inflating: safegraph_open_census_data/data/cbg_b19.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b19.csv
  inflating: safegraph_open_census_data/data/cbg_b25.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b25.csv
  inflating: safegraph_open_census_data/data/cbg_b01.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b01.csv
  inflating: safegraph_open_census_data/data/cbg_b15.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b15.csv
  inflating: safegraph_open_census_data/data/cbg_b14.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b14.csv
  inflating: safegraph_open_census_data/data/cbg_b00.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b00.csv
  inflating: safegraph_open_census_data/data/cbg_b02.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b02.csv
  inflating: safegraph_open_census_data/data/cbg_b03.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b03.csv
  inflating: safegraph_open_census_data/data/cbg_b07.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b07.csv
  inflating: safegraph_open_census_data/data/cbg_b12.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b12.csv
```

- All the data is extract into folder **safegraph_open_census_data**

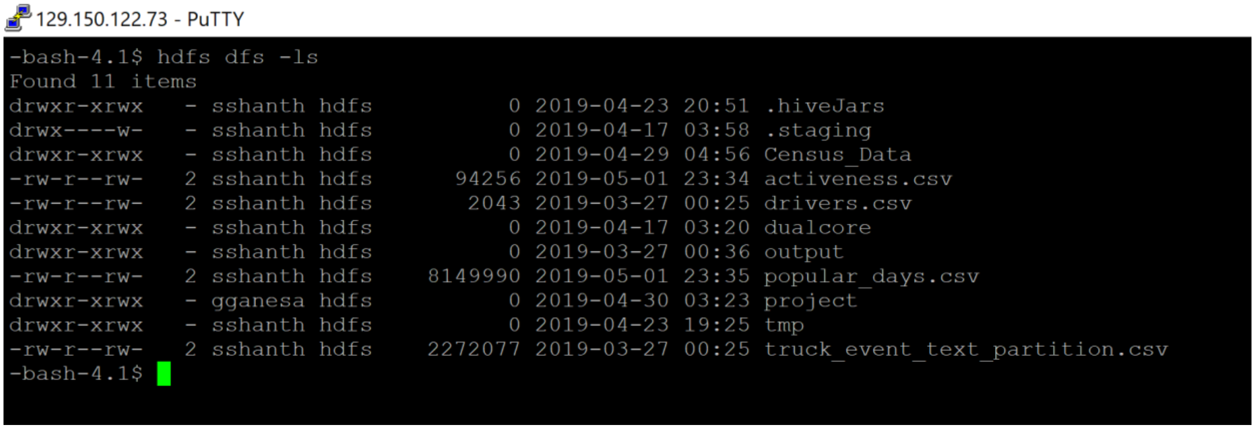


```
  inflating: safegraph_open_census_data/data/cbg_b21.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b21.csv
  inflating: safegraph_open_census_data/data/cbg_b23.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b23.csv
  inflating: safegraph_open_census_data/data/cbg_b22.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b22.csv
  inflating: safegraph_open_census_data/data/cbg_b27.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b27.csv
  inflating: safegraph_open_census_data/data/cbg_b19.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b19.csv
  inflating: safegraph_open_census_data/data/cbg_b25.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b25.csv
  inflating: safegraph_open_census_data/data/cbg_b01.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b01.csv
  inflating: safegraph_open_census_data/data/cbg_b15.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b15.csv
  inflating: safegraph_open_census_data/data/cbg_b14.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b14.csv
  inflating: safegraph_open_census_data/data/cbg_b00.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b00.csv
  inflating: safegraph_open_census_data/data/cbg_b02.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b02.csv
  inflating: safegraph_open_census_data/data/cbg_b03.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b03.csv
  inflating: safegraph_open_census_data/data/cbg_b07.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b07.csv
  inflating: safegraph_open_census_data/data/cbg_b12.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b12.csv
  inflating: safegraph_open_census_data/data/cbg_patterns.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_patterns.csv
  inflating: safegraph_open_census_data/data/cbg_b11.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b11.csv
  inflating: safegraph_open_census_data/data/cbg_c16.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_c16.csv
  inflating: safegraph_open_census_data/data/cbg_c17.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_c17.csv
  inflating: safegraph_open_census_data/data/cbg_b99.csv
  inflating: __MACOSX/safegraph_open_census_data/data/._cbg_b99.csv
  inflating: __MACOSX/safegraph_open_census_data/._data
   creating: safegraph_open_census_data/metadata/
  inflating: safegraph_open_census_data/metadata/.DS_Store
   creating: __MACOSX/safegraph_open_census_data/metadata/
  inflating: __MACOSX/safegraph_open_census_data/metadata/._.DS_Store
  inflating: safegraph_open_census_data/metadata/cbg_field_descriptions.csv
  inflating: __MACOSX/safegraph_open_census_data/metadata/._cbg_field_descriptions.csv
  inflating: safegraph_open_census_data/metadata/cbg_fips_codes.csv
  inflating: __MACOSX/safegraph_open_census_data/metadata/._cbg_fips_codes.csv
  inflating: safegraph_open_census_data/metadata/cbg_geographic_data.csv
  inflating: __MACOSX/safegraph_open_census_data/metadata/._cbg_geographic_data.csv
  inflating: __MACOSX/safegraph_open_census_data/._metadata
  inflating: __MACOSX/._safegraph_open_census_data
-bash-4.1$ ls
activeness.csv                                         Data_Set_Final      driver_data      labPigETL  pig_1553646541554.log  safegraph_open_census_data  test.csv     Upload Data.zip
census-block-group-american-community-survey-data.zip  Data_Set_Final.zip  driver_data.zip  __MACOSX   popular_days.csv       test1.csv                   Upload_Data
-bash-4.1$ 
```

## Step 2: Uploading data to HDFS

Create folders using **mkdir** command, Change the permission for the file using below command to get full access to the file.

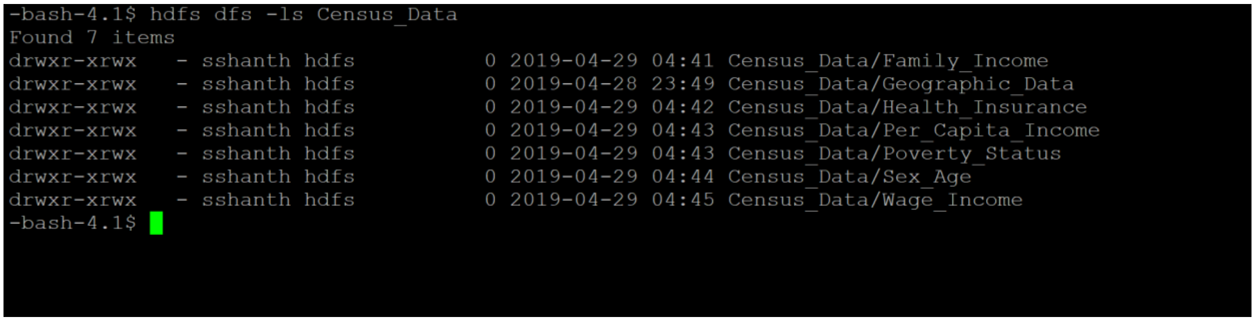
- **hdfs dfs -mkdir Census_Data**
- **hdfs dfs -chmod -R o+w .**
- **hdfs dfs -chmod -R o+w Census_Data**
- **hdfs dfs -mkdir Census_Data/Family_Income**
- **hdfs dfs -chmod -R o+w Census_Data/Family_Income**
- **hdfs dfs -mkdir Census_Data/Geographic_Data**
- **hdfs dfs -chmod -R o+w Census_Data/Geographic_Data**
- **hdfs dfs -mkdir Census_Data/Health_Insurance**
- **hdfs dfs -chmod -R o+w Census_Data/Health_Insurance**
- **hdfs dfs -mkdir Census_Data/Per_Capita_Income**
- **hdfs dfs -chmod -R o+w Census_Data/Per_Capita_Income**
- **hdfs dfs -mkdir Census_Data/Poverty_Status**
- **hdfs dfs -chmod -R o+w Census_Data/Poverty_Status**
- **hdfs dfs -mkdir Census_Data/Sex_Age**
- **hdfs dfs -chmod -R o+w Census_Data/Sex_Age**
- **hdfs dfs -mkdir Census_Data/Wage_Income**
- **hdfs dfs -chmod -R o+w Census_Data/Wage_Income**



```
-bash-4.1$ hdfs dfs -ls
Found 11 items
drwxr-xrwx   - sshanth hdfs          0 2019-04-23 20:51 .hiveJars
drwx----w-   - sshanth hdfs          0 2019-04-17 03:58 .staging
drwxr-xrwx   - sshanth hdfs          0 2019-04-29 04:56 Census_Data
-rw-r--rw-   2 sshanth hdfs      94256 2019-05-01 23:34 activeness.csv
-rw-r--rw-   2 sshanth hdfs       2043 2019-03-27 00:25 drivers.csv
drwxr-xrwx   - sshanth hdfs          0 2019-04-17 03:20 dualcore
drwxr-xrwx   - sshanth hdfs          0 2019-03-27 00:36 output
-rw-r--rw-   2 sshanth hdfs    8149990 2019-05-01 23:35 popular_days.csv
drwxr-xrwx   - gganesa hdfs          0 2019-04-30 03:23 project
drwxr-xrwx   - sshanth hdfs          0 2019-04-23 19:25 tmp
-rw-r--rw-   2 sshanth hdfs    2272077 2019-03-27 00:25 truck_event_text_partition.csv
-bash-4.1$
```

**hdfs dfs –ls**

**hdfs dfs –ls Census_Data**



```
-bash-4.1$ hdfs dfs -ls Census_Data
Found 7 items
drwxr-xrwx   - sshanth hdfs          0 2019-04-29 04:41 Census_Data/Family_Income
drwxr-xrwx   - sshanth hdfs          0 2019-04-28 23:49 Census_Data/Geographic_Data
drwxr-xrwx   - sshanth hdfs          0 2019-04-29 04:42 Census_Data/Health_Insurance
drwxr-xrwx   - sshanth hdfs          0 2019-04-29 04:43 Census_Data/Per_Capita_Income
drwxr-xrwx   - sshanth hdfs          0 2019-04-29 04:43 Census_Data/Poverty_Status
drwxr-xrwx   - sshanth hdfs          0 2019-04-29 04:44 Census_Data/Sex_Age
drwxr-xrwx   - sshanth hdfs          0 2019-04-29 04:45 Census_Data/Wage_Income
-bash-4.1$ 
```

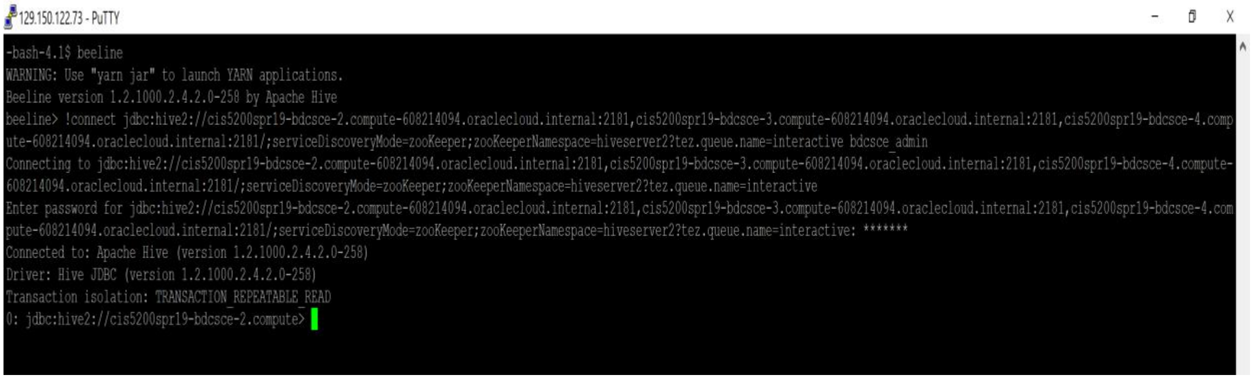
Use hdfs dfs -put command to upload the data from the Linux server to HDFS

- **hdfs dfs -put Family_Income.xlsx Census_Data/Family_Income**
- **hdfs dfs -put Geographic_Data.xlsx Census_Data/ Geographic_Data**
- **hdfs dfs -put Health_Insurance.xlsx Census_Data/Health_Insurance**
- **hdfs dfs -put Per_Capita_Income.xlsx Census_Data/Per_Capita_Income**
- **hdfs dfs -put Poverty_Status.xlsx Census_Data/Poverty_Status**
- **hdfs dfs -put Sex_Age.xlsx Census_Data/Sex_Age**
- **hdfs dfs -put Wage_Income.xlsx Census_Data/Wage_Income**

## Step 3: Connecting server to HIVE

- Use the command **beeline** to connect with to Hive

**!connect jdbc:hive2://cis5200spr19-bdcsce-2.compute-608214094.oraclecloud.internal:2181,cis5200spr19-bdcsce-3.compute-608214094.oraclecloud.internal:2181,cis5200spr19-bdcsce-4.compute-608214094.oraclecloud.internal:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive bdcsce_admin**



```
-bash-4.1$ beeline
WARNING: Use "yarn jar" to launch YARN applications.
Beeline version 1.2.1000.2.4.2.0-258 by Apache Hive
beeline> !connect jdbc:hive2://cis5200spr19-bdcsce-2.compute-608214094.oraclecloud.internal:2181,cis5200spr19-bdcsce-3.compute-608214094.oraclecloud.internal:2181,cis5200spr19-bdcsce-4.comp
ute-608214094.oraclecloud.internal:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive bdcsce_admin
Connecting to jdbc:hive2://cis5200spr19-bdcsce-2.compute-608214094.oraclecloud.internal:2181,cis5200spr19-bdcsce-3.compute-608214094.oraclecloud.internal:2181,cis5200spr19-bdcsce-4.compute-
608214094.oraclecloud.internal:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive
Enter password for jdbc:hive2://cis5200spr19-bdcsce-2.compute-608214094.oraclecloud.internal:2181,cis5200spr19-bdcsce-3.compute-608214094.oraclecloud.internal:2181,cis5200spr19-bdcsce-4.com
pute-608214094.oraclecloud.internal:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive: *******
Connected to: Apache Hive (version 1.2.1000.2.4.2.0-258)
Driver: Hive JDBC (version 1.2.1000.2.4.2.0-258)
Transaction isolation: TRANSACTION_REPEATABLE_READ
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute>
```

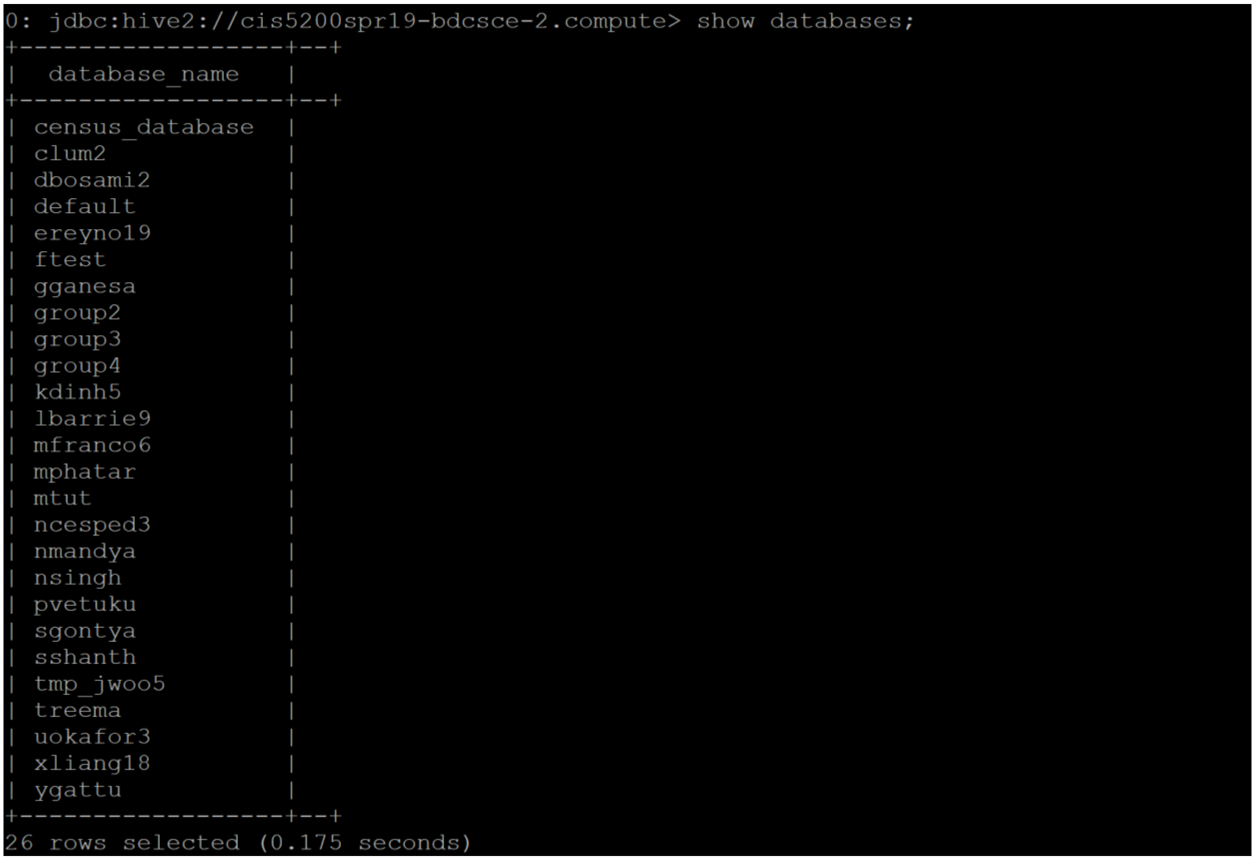
## Step 4: Creating Database

- Use **create database** to create a database named group3.

  **create database group3;**

- To view the newly created database use.

  **Show databases;**



```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> show databases;
+-------------------+--+
|  database_name    |
+-------------------+--+
| census_database   |
| clum2             |
| dbosami2          |
| default           |
| ereyno19          |
| ftest             |
| gganesa           |
| group2            |
| group3            |
| group4            |
| kdinh5            |
| lbarrie9          |
| mfranco6          |
| mphatar           |
| mtut              |
| ncesped3          |
| nmandya           |
| nsingh            |
| pvetuku           |
| sgontya           |
| sshanth           |
| tmp_jwoo5         |
| treema            |
| uokafor3          |
| xliang18          |
| ygattu            |
+-------------------+--+
26 rows selected (0.175 seconds)
```

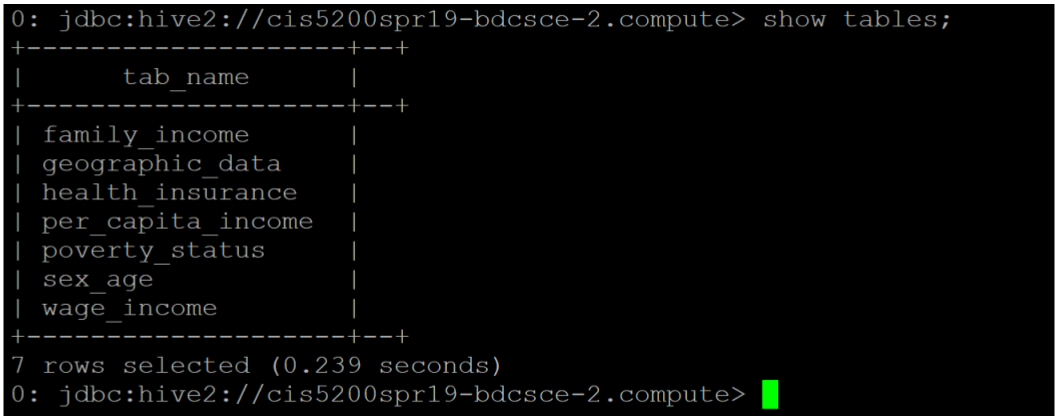
## Step 5: Creating Tables in Database

CREATE EXTERNAL TABLE IF NOT EXISTS Family_Income (Census_Group BIGINT, Family_Income BIGINT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/sshanth/Census_Data/Family_Income' TBLPROPERTIES ('skip.header.line.count'='1');

CREATE EXTERNAL TABLE IF NOT EXISTS Health_Insurance (Census_Group BIGINT, Insurance BIGINT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/sshanth/Census_Data/Health_Insurance' TBLPROPERTIES ('skip.header.line.count'='1');

CREATE EXTERNAL TABLE IF NOT EXISTS Per_Capita_Income (Census_Group BIGINT, Per_Capita_Income BIGINT)

```
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/sshanth/Census_Data/Per_Capita_Income'
TBLPROPERTIES ('skip.header.line.count'='1');

CREATE EXTERNAL TABLE IF NOT EXISTS Poverty_Status (Census_Group BIGINT,
Poverty_Status BIGINT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/sshanth/Census_Data/Poverty_Status'
TBLPROPERTIES ('skip.header.line.count'='1');

CREATE EXTERNAL TABLE IF NOT EXISTS Sex_Age (Census_Group BIGINT,
Total_Population BIGINT, Male_Population BIGINT,Female_Population BIGINT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/sshanth/Census_Data/Sex_Age'
TBLPROPERTIES ('skip.header.line.count'='1');

CREATE EXTERNAL TABLE IF NOT EXISTS Wage_Income (Census_Group BIGINT,
Total_Income BIGINT, Salaried_Wages BIGINT,Non_Salaried_Wages BIGINT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/sshanth/Census_Data/Wage_Income'
TBLPROPERTIES ('skip.header.line.count'='1');

CREATE EXTERNAL TABLE IF NOT EXISTS Geographic_Data ( Census_Group BIGINT,
Amount_Land BIGINT, Amount_Water BIGINT, Latitude BIGINT, Longitude BIGINT )
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/sshanth/Census_Data/Geographic_Data'
TBLPROPERTIES ('skip.header.line.count'='1');
```



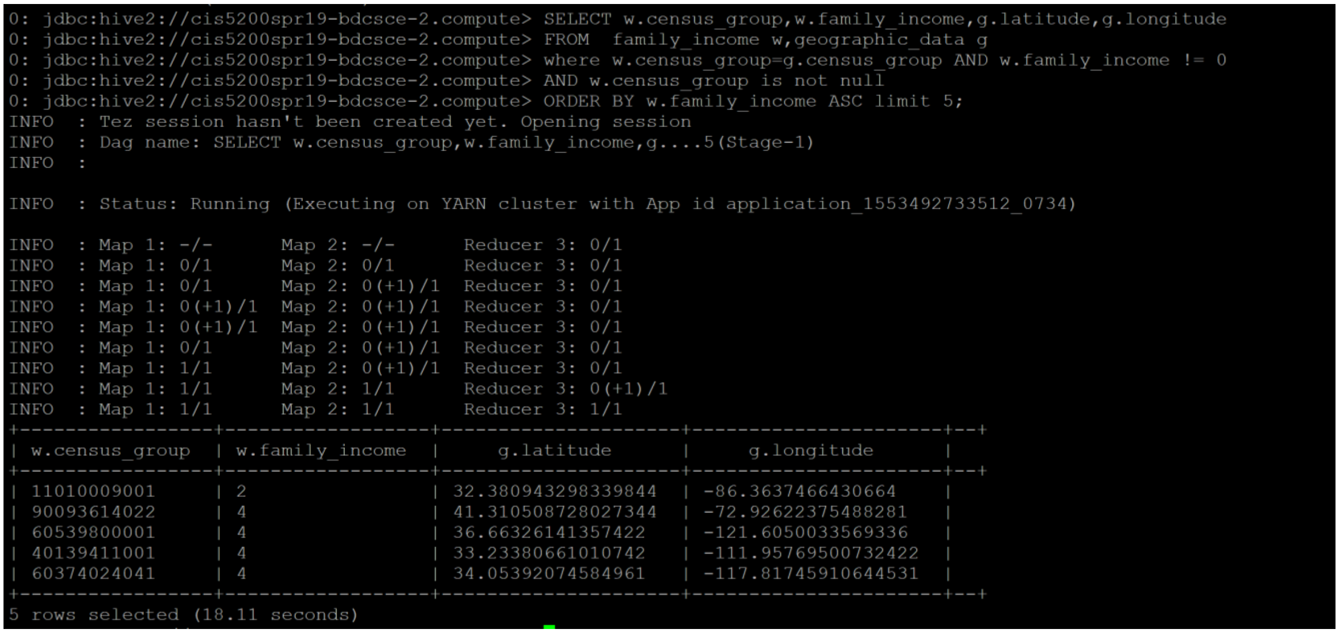
```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> show tables;
+---------------------+--+
|      tab_name       |
+---------------------+--+
| family_income       |
| geographic_data     |
| health_insurance    |
| per_capita_income   |
| poverty_status      |
| sex_age             |
| wage_income         |
+---------------------+--+
7 rows selected (0.239 seconds)
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute>
```

# Step 6: Hive Queries to Analyze Data and the Visualization

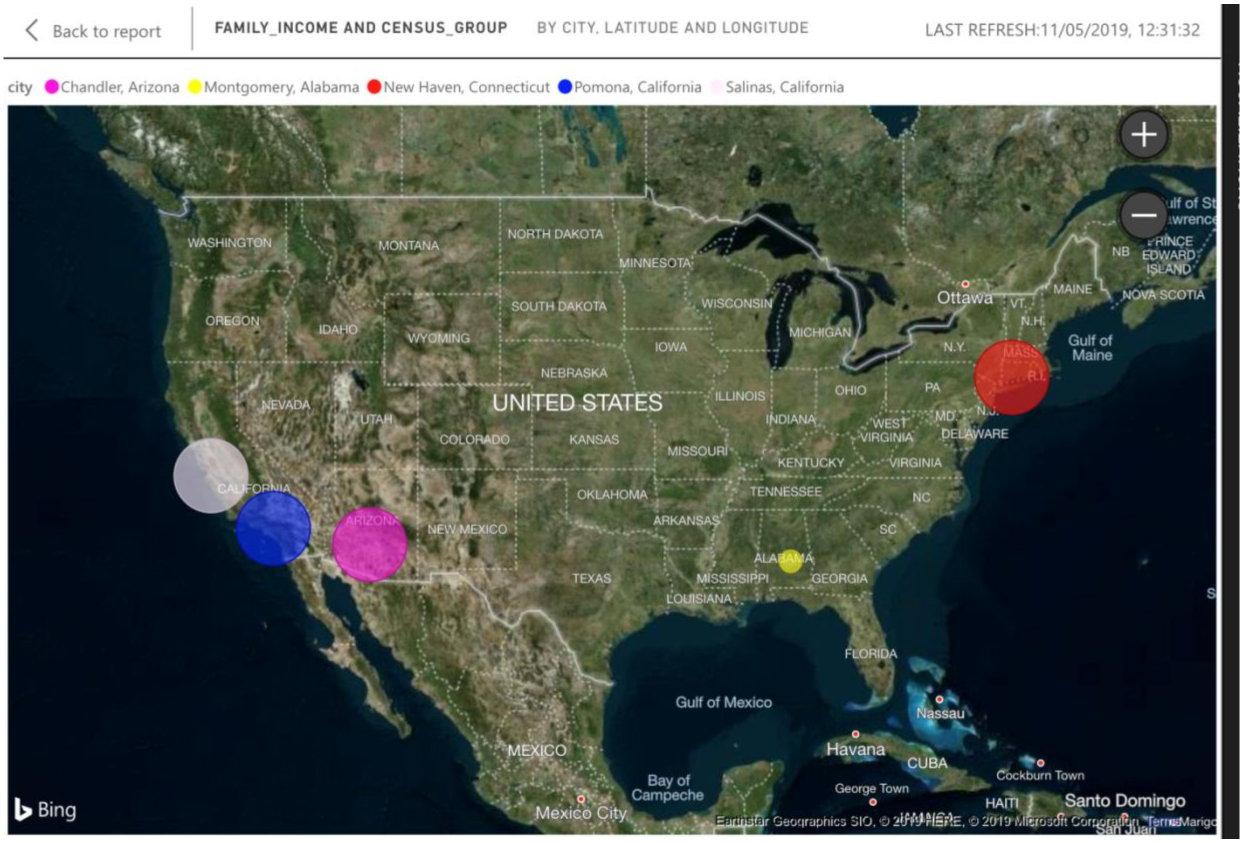
The following Hive queries analyses the Census Block Group American Community Survey Data

1. **Census groups with lowest Family Income**

SELECT w.census_group,w.family_income,g.latitude,g.longitude
FROM  family_income w,geographic_data g
where w.census_group=g.census_group AND w.family_income != 0 AND w.census_group is not null
ORDER BY w.family_income ASC limit 5;



```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> SELECT w.census_group,w.family_income,g.latitude,g.longitude
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> FROM  family_income w,geographic_data g
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> where w.census_group=g.census_group AND w.family_income != 0
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> AND w.census_group is not null
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> ORDER BY w.family_income ASC limit 5;
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: SELECT w.census_group,w.family_income,g....5(Stage-1)
INFO  :

INFO  : Status: Running (Executing on YARN cluster with App id application_1553492733512_0734)

INFO  : Map 1: -/-      Map 2: -/-      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1/1
+-----------------+-------------------+----------------------+----------------------+--+
| w.census_group  | w.family_income   |      g.latitude      |     g.longitude      |
+-----------------+-------------------+----------------------+----------------------+--+
| 11010009001     | 2                 | 32.380943298339844   | -86.3637466430664    |
| 90093614022     | 4                 | 41.310508728027344   | -72.92622375488281   |
| 60539800001     | 4                 | 36.66326141357422    | -121.6050033569336   |
| 40139411001     | 4                 | 33.23380661010742    | -111.95769500732422  |
| 60374024041     | 4                 | 34.05392074584961    | -117.81745910644531  |
+-----------------+-------------------+----------------------+----------------------+--+
5 rows selected (18.11 seconds)
```



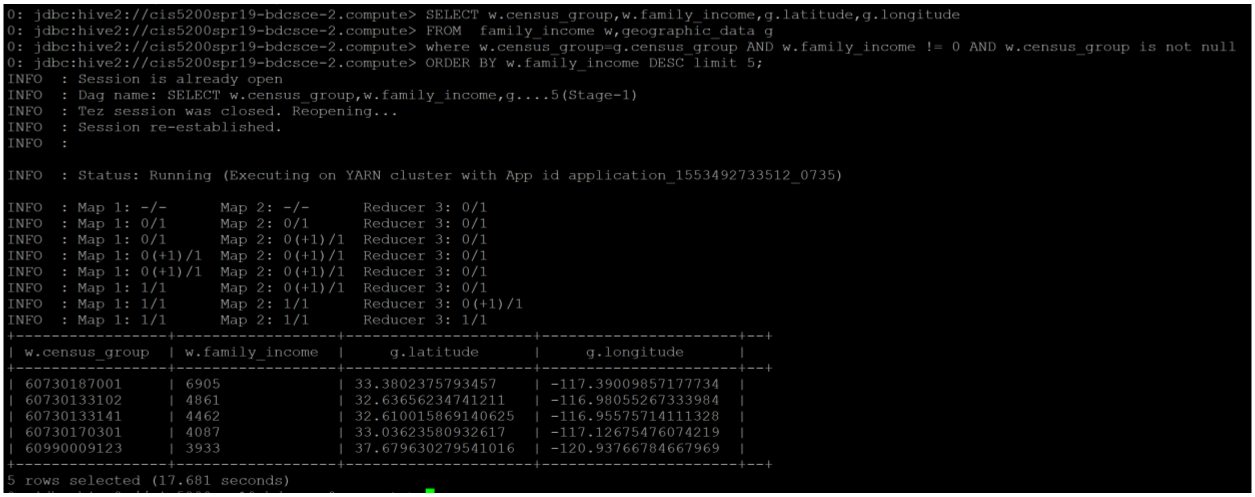
< Back to report | FAMILY_INCOME AND CENSUS_GROUP BY CITY, LATITUDE AND LONGITUDE LAST REFRESH:11/05/2019, 12:31:32

city ● Chandler, Arizona ● Montgomery, Alabama ● New Haven, Connecticut ● Pomona, California ● Salinas, California

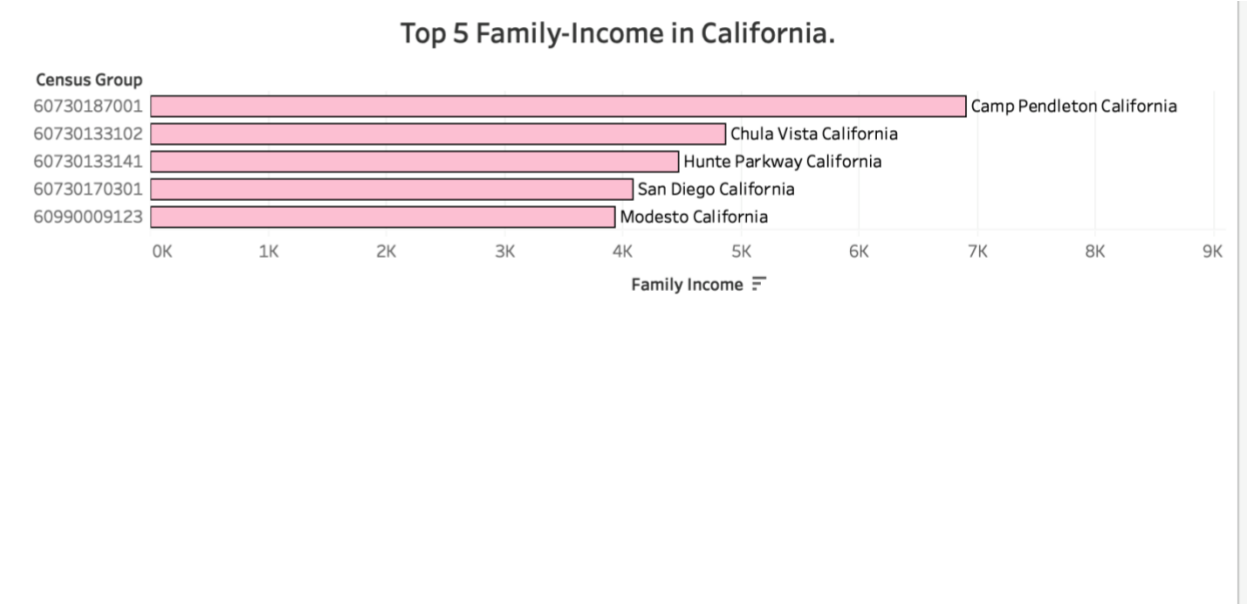
In the above Map Visualization we have represented the Census Blocks with low family-income. The Census Blocks in Montgomery City, Alabama State (yellow point on the map) has recorded low family-income.

2. **Query to find the top 5 Census Blocks with Highest Family Income**

```
SELECT w.census_group,w.family_income,g.latitude,g.longitude
FROM  family_income w,geographic_data g
where w.census_group=g.census_group AND w.family_income != 0 AND w.census_group is not null
ORDER BY w.family_income DESC limit 5;
```



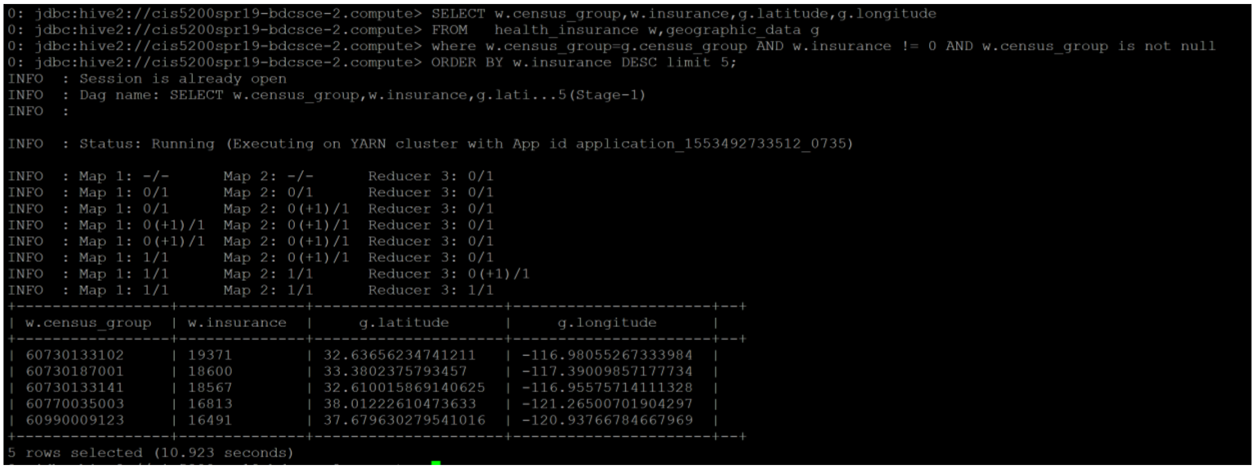
```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> SELECT w.census_group,w.family_income,g.latitude,g.longitude
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> FROM  family_income w,geographic_data g
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> where w.census_group=g.census_group AND w.family_income != 0 AND w.census_group is not null
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> ORDER BY w.family_income DESC limit 5;
INFO  : Session is already open
INFO  : Dag name: SELECT w.census_group,w.family_income,g....5(Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  :

INFO  : Status: Running (Executing on YARN cluster with App id application_1553492733512_0735)

INFO  : Map 1: -/-      Map 2: -/-      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1/1
+-----------------+-------------------+---------------------+----------------------+--+
| w.census_group  | w.family_income   |     g.latitude      |     g.longitude      |
+-----------------+-------------------+---------------------+----------------------+--+
| 60730187001     | 6905              | 33.3802375793457    | -117.39009857177734  |
| 60730133102     | 4861              | 32.63656234741211   | -116.98055267333984  |
| 60730133141     | 4462              | 32.610015869140625  | -116.95575714111328  |
| 60730170301     | 4087              | 33.03623580932617   | -117.12675476074219  |
| 60990009123     | 3933              | 37.679630279541016  | -120.93766784667969  |
+-----------------+-------------------+---------------------+----------------------+--+
5 rows selected (17.681 seconds)
```



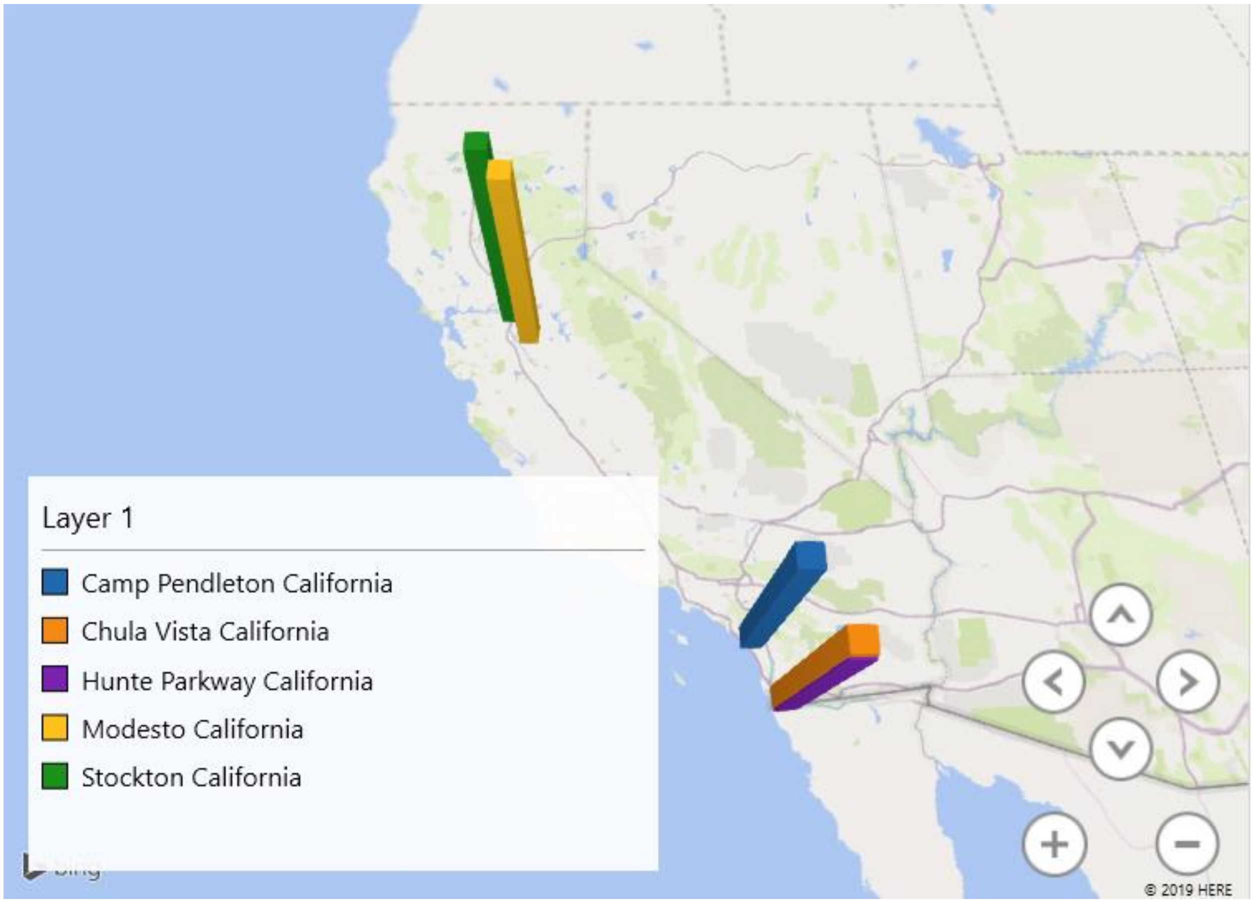
The above graph shows the Census Blocks in California with highest Family-Income. The census blocks across Camp Pendleton have recorded the family-income of $6.9k.

### 3. Query to find the top 5 Census Blocks with Health Insurance

SELECT w.census_group,w.insurance,g.latitude,g.longitude
FROM   health_insurance w,geographic_data g
where w.census_group=g.census_group AND w.insurance != 0 AND w.census_group is not null
ORDER BY w.insurance DESC limit 5;



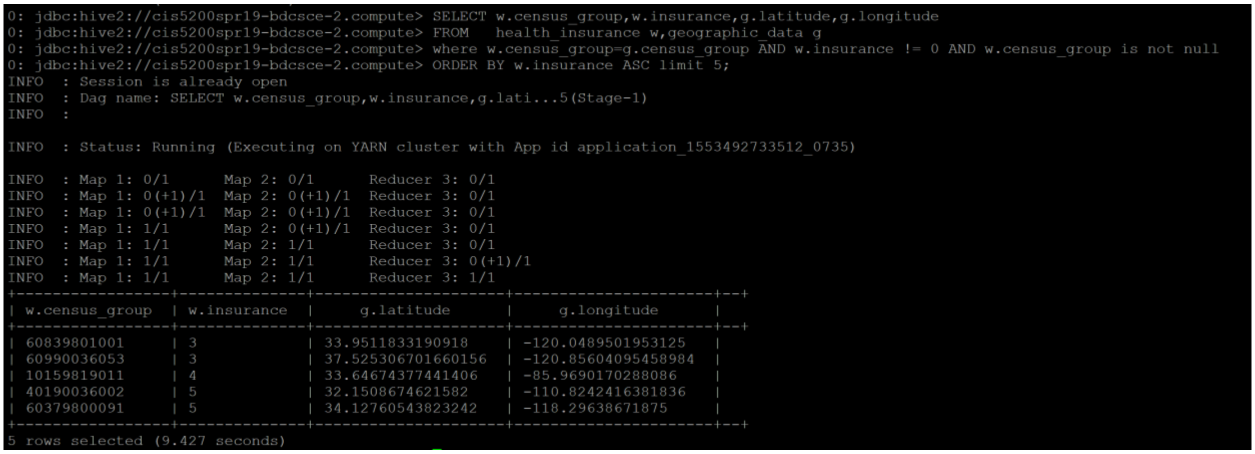
```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> SELECT w.census_group,w.insurance,g.latitude,g.longitude
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> FROM   health_insurance w,geographic_data g
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> where w.census_group=g.census_group AND w.insurance != 0 AND w.census_group is not null
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> ORDER BY w.insurance DESC limit 5;
INFO  : Session is already open
INFO  : Dag name: SELECT w.census_group,w.insurance,g.lati...5(Stage-1)
INFO  :

INFO  : Status: Running (Executing on YARN cluster with App id application_1553492733512_0735)

INFO  : Map 1: -/-      Map 2: -/-      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1/1
+-----------------+-------------+---------------------+----------------------+--+
| w.census_group  | w.insurance  |      g.latitude      |      g.longitude      |
+-----------------+-------------+---------------------+----------------------+--+
| 60730133102     | 19371        | 32.63656234741211   | -116.98055267333984  |
| 60730187001     | 18600        | 33.3802375793457    | -117.39009857177734  |
| 60730133141     | 18567        | 32.610015869140625  | -116.95575714111328  |
| 60770035003     | 16813        | 38.01222610473633   | -121.26500701904297  |
| 60990009123     | 16491        | 37.679630279541016  | -120.93766784667969  |
+-----------------+-------------+---------------------+----------------------+--+
5 rows selected (10.923 seconds)
```



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.

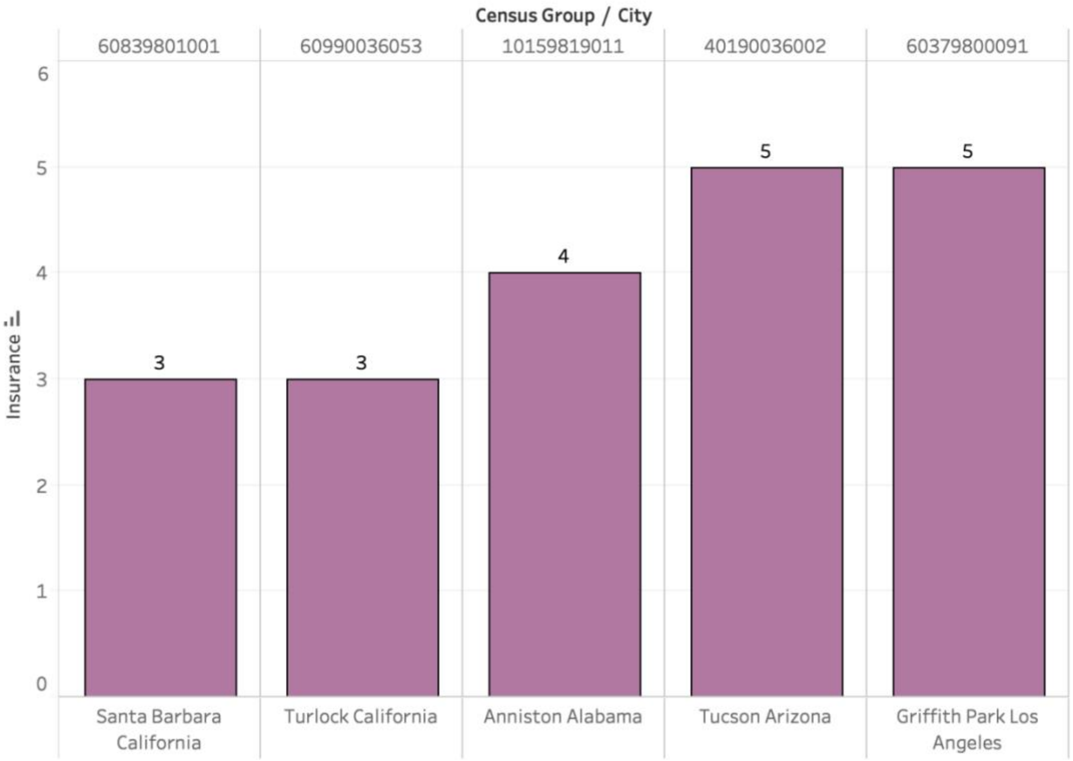
### 4. Query to find the bottom 5 Census Blocks with Health Insurance

```
SELECT w.census_group,w.insurance,g.latitude,g.longitude
FROM   health_insurance w,geographic_data g
where w.census_group=g.census_group AND w.insurance != 0 AND w.census_group is not null
ORDER BY w.insurance ASC limit 5;
```



```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> SELECT w.census_group,w.insurance,g.latitude,g.longitude
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> FROM   health_insurance w,geographic_data g
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> where w.census_group=g.census_group AND w.insurance != 0 AND w.census_group is not null
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> ORDER BY w.insurance ASC limit 5;
INFO  : Session is already open
INFO  : Dag name: SELECT w.census_group,w.insurance,g.lati...5(Stage-1)
INFO  :

INFO  : Status: Running (Executing on YARN cluster with App id application_1553492733512_0735)

INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1/1
+-----------------+--------------+---------------------+----------------------+--+
| w.census_group  | w.insurance  |     g.latitude      |     g.longitude      |
+-----------------+--------------+---------------------+----------------------+--+
| 60839801001     | 3            | 33.9511833190918    | -120.0489501953125   |
| 60990036053     | 3            | 37.525306701660156  | -120.85604095458984  |
| 10159819011     | 4            | 33.64674377441406   | -85.9690170288086    |
| 40190036002     | 5            | 32.1508674621582    | -110.8242416381836   |
| 60379800091     | 5            | 34.12760543823242   | -118.29638671875     |
+-----------------+--------------+---------------------+----------------------+--+
5 rows selected (9.427 seconds)
```

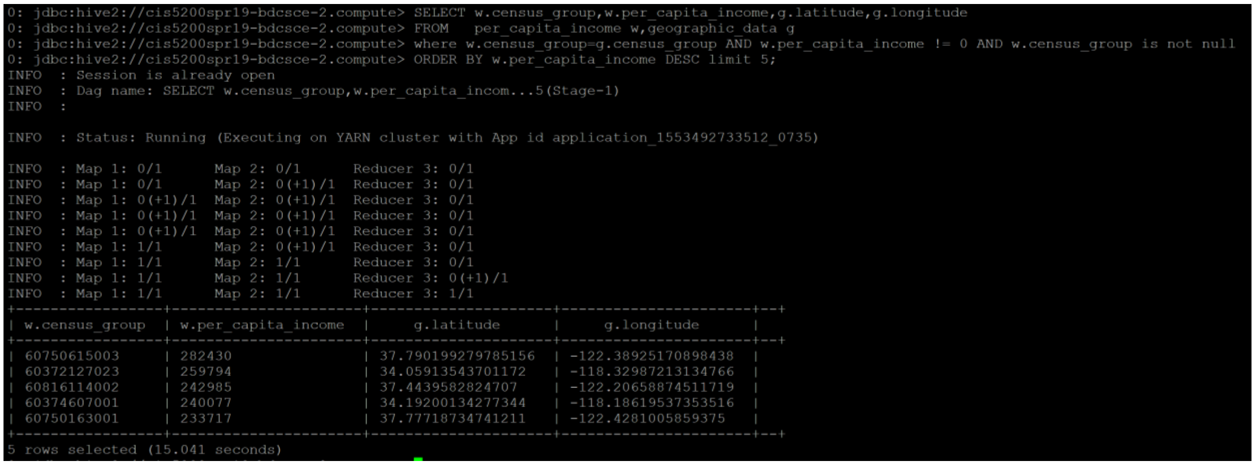
Bottom-5 Regions in the US with Low Health Insurance Coverage (in Thousands)



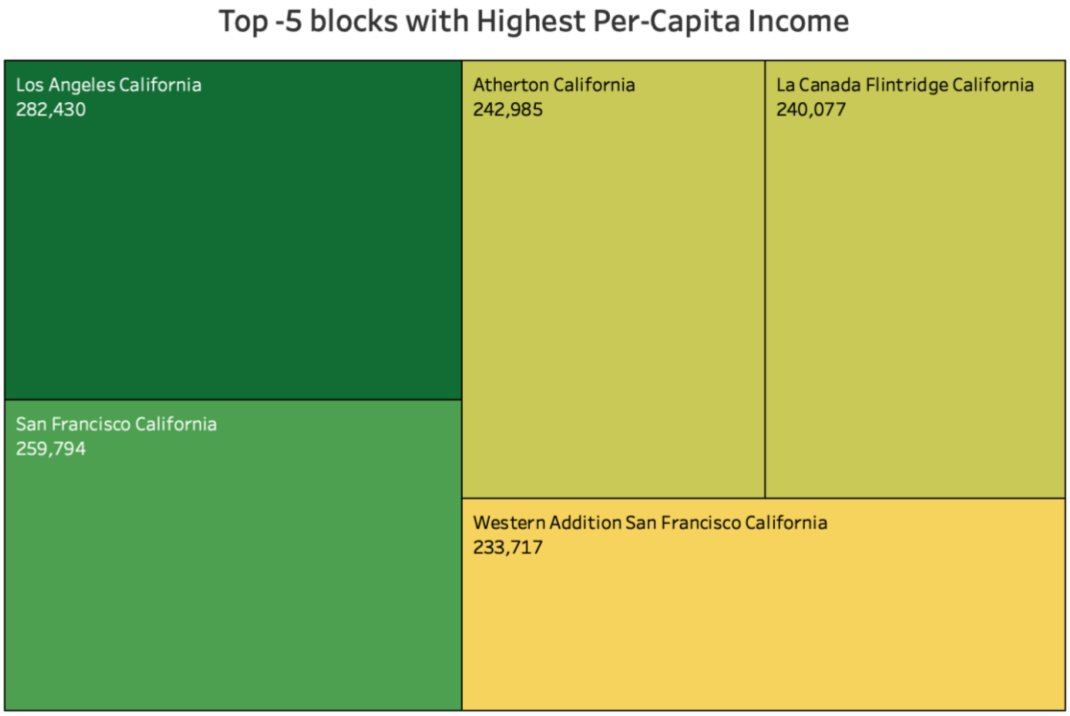
The above bar graphs represents the regions in which the blocks have lower health insurance coverage. We noticed that the city of Santa Barbara and Turlock had blocks with lower insurance coverage followed by Anniston and Tucson.

## 5. Query to find the top 5 Census Blocks with Highest Per-Capita-Income

```
SELECT w.census_group,w.per_capita_income,g.latitude,g.longitude
FROM   per_capita_income w,geographic_data g
where w.census_group=g.census_group AND w.per_capita_income != 0 AND w.census_group is not null
ORDER BY w.per_capita_income DESC limit 5;
```



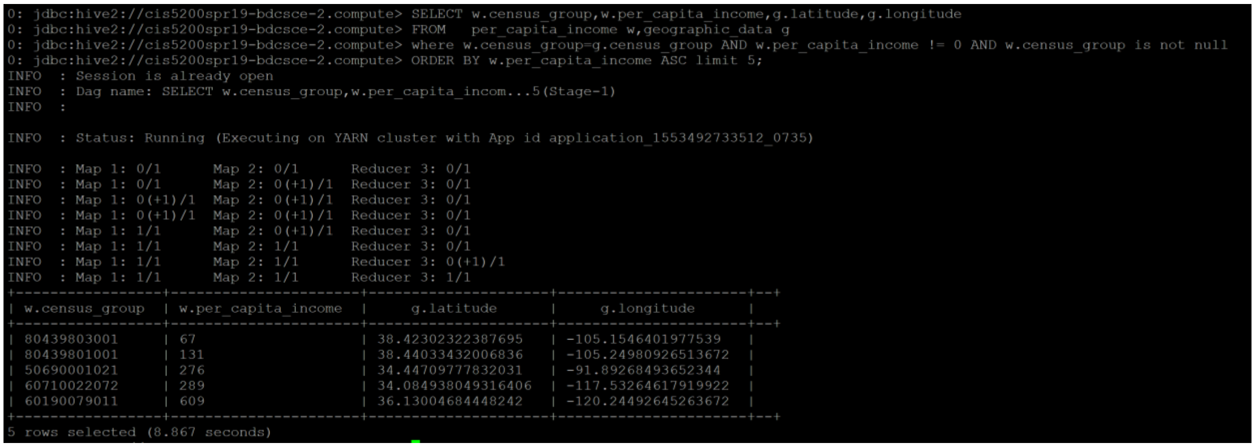
```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> SELECT w.census_group,w.per_capita_income,g.latitude,g.longitude
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> FROM   per_capita_income w,geographic_data g
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> where w.census_group=g.census_group AND w.per_capita_income != 0 AND w.census_group is not null
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> ORDER BY w.per_capita_income DESC limit 5;
INFO  : Session is already open
INFO  : Dag name: SELECT w.census_group,w.per_capita_incom...5(Stage-1)
INFO  :

INFO  : Status: Running (Executing on YARN cluster with App id application_1553492733512_0735)

INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1/1
+-----------------+----------------------+----------------------+----------------------+--+
| w.census_group  | w.per_capita_income  |      g.latitude      |      g.longitude     |
+-----------------+----------------------+----------------------+----------------------+--+
| 60750615003     | 282430               | 37.790199279785156   | -122.38925170898438  |
| 60372127023     | 259794               | 34.05913543701172    | -118.32987213134766  |
| 60816114002     | 242985               | 37.4439582824707     | -122.20658874511719  |
| 60374607001     | 240077               | 34.19200134277344    | -118.18619537353516  |
| 60750163001     | 233717               | 37.77718734741211    | -122.4281005859375   |
+-----------------+----------------------+----------------------+----------------------+--+
5 rows selected (15.041 seconds)
```



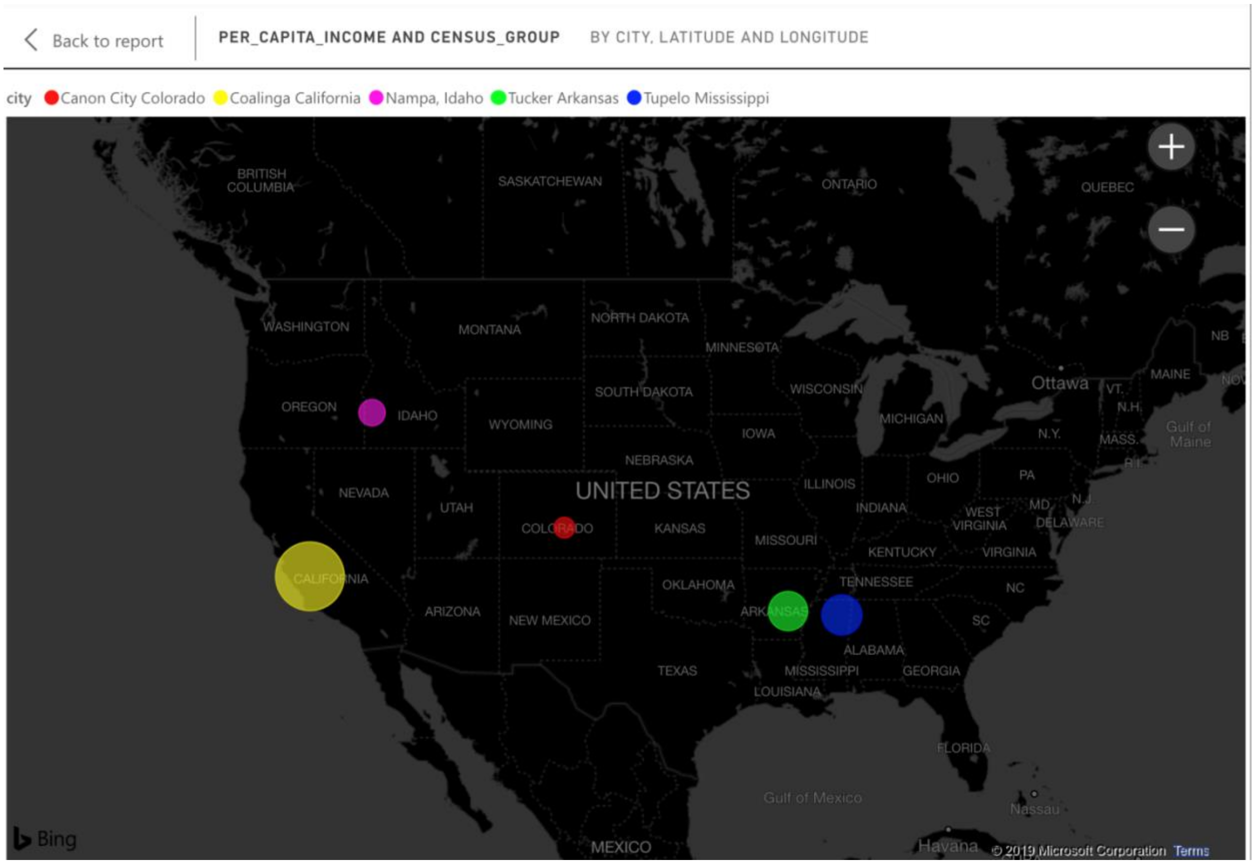
This is a tree-map representation wherein the larger the size and darker the color signifies a greater value and vice versa for the smaller value. As per the above visualization the City of Los Angeles has the highest per-capita income of $282,430 followed by San Francisco $259,794.

## 6. Query to find the bottom 5 Census Blocks with Lowest Per-Capita-Income

```
SELECT w.census_group,w.per_capita_income,g.latitude,g.longitude
FROM   per_capita_income w,geographic_data g
where w.census_group=g.census_group AND w.per_capita_income != 0 AND w.census_group is not null
ORDER BY w.per_capita_income ASC limit 5;
```



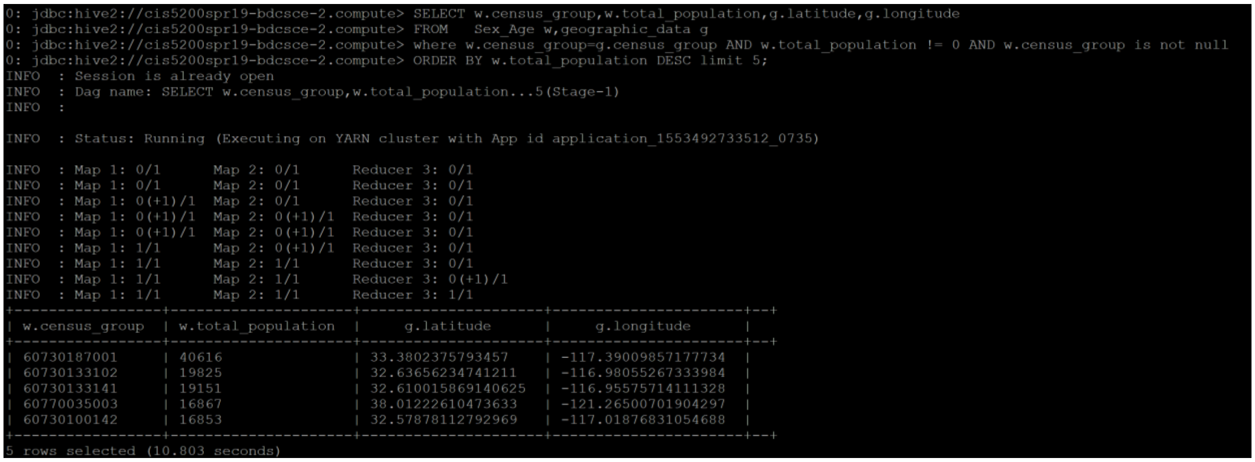
```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> SELECT w.census_group,w.per_capita_income,g.latitude,g.longitude
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> FROM   per_capita_income w,geographic_data g
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> where w.census_group=g.census_group AND w.per_capita_income != 0 AND w.census_group is not null
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> ORDER BY w.per_capita_income ASC limit 5;
INFO  : Session is already open
INFO  : Dag name: SELECT w.census_group,w.per_capita_incom...5(Stage-1)
INFO  :

INFO  : Status: Running (Executing on YARN cluster with App id application_1553492733512_0735)

INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1/1
+-----------------+----------------------+---------------------+----------------------+--+
| w.census_group  | w.per_capita_income  |     g.latitude      |     g.longitude      |
+-----------------+----------------------+---------------------+----------------------+--+
| 80439803001     | 67                   | 38.42302322387695   | -105.1546401977539   |
| 80439801001     | 131                  | 38.44033432006836   | -105.24980926513672  |
| 50690001021     | 276                  | 34.44709777832031   | -91.89268493652344   |
| 60710022072     | 289                  | 34.084938049316406  | -117.53264617919922  |
| 60190079011     | 609                  | 36.13004684448242   | -120.24492645263672  |
+-----------------+----------------------+---------------------+----------------------+--+
5 rows selected (8.867 seconds)
```



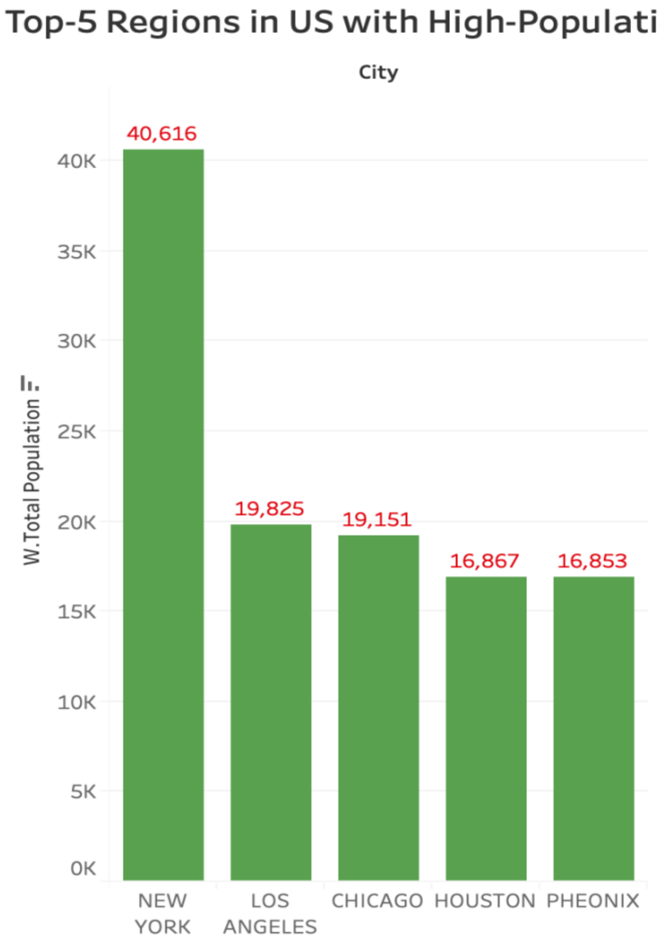
In the map representation we can observe the regions with low Per-Capita income. Canon City in Colorado State (red color) has the lowest average earnings per square area followed by Nampa city in Idaho and Tucker city in Arkansas state.

### 7. Query to find the top 5 Census Blocks with Highest Population

```
SELECT w.census_group,w.total_population,g.latitude,g.longitude
FROM   Sex_Age w,geographic_data g
where w.census_group=g.census_group AND w.total_population != 0 AND w.census_group is not null
ORDER BY w.total_population DESC limit 5;
```



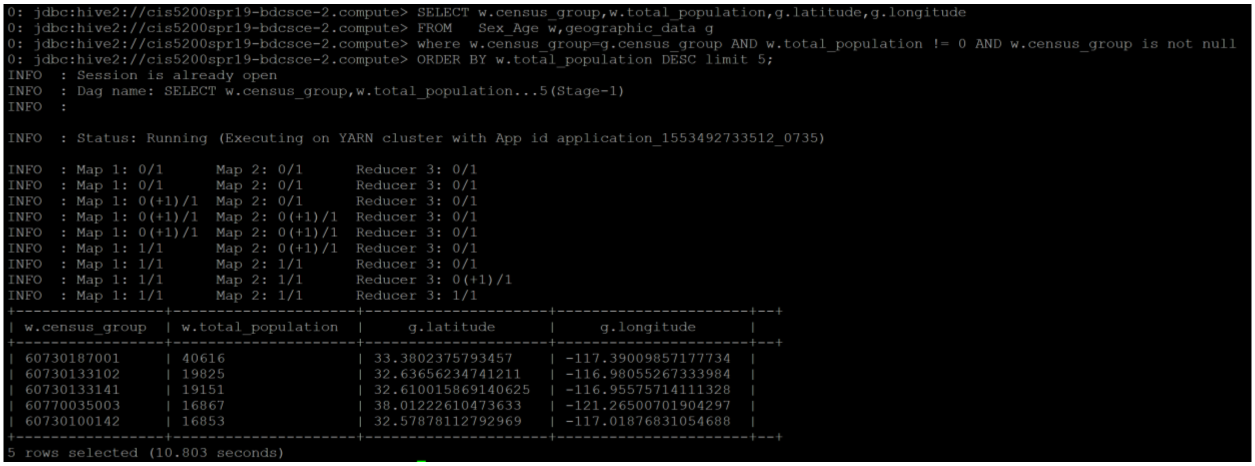
```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> SELECT w.census_group,w.total_population,g.latitude,g.longitude
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> FROM   Sex_Age w,geographic_data g
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> where w.census_group=g.census_group AND w.total_population != 0 AND w.census_group is not null
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> ORDER BY w.total_population DESC limit 5;
INFO  : Session is already open
INFO  : Dag name: SELECT w.census_group,w.total_population...5(Stage-1)
INFO  :

INFO  : Status: Running (Executing on YARN cluster with App id application_1553492733512_0735)

INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1/1
+-----------------+---------------------+----------------------+-----------------------+--+
| w.census_group  | w.total_population  |      g.latitude      |      g.longitude      |
+-----------------+---------------------+----------------------+-----------------------+--+
| 60730187001     | 40616               | 33.3802375793457     | -117.39009857177734   |
| 60730133102     | 19825               | 32.63656234741211    | -116.98055267333984   |
| 60730133141     | 19151               | 32.610015869140625   | -116.95575714111328   |
| 60770035003     | 16867               | 38.01222610473633    | -121.26500701904297   |
| 60730100142     | 16853               | 32.57878112792969    | -117.01876831054688   |
+-----------------+---------------------+----------------------+-----------------------+--+
5 rows selected (10.803 seconds)
```



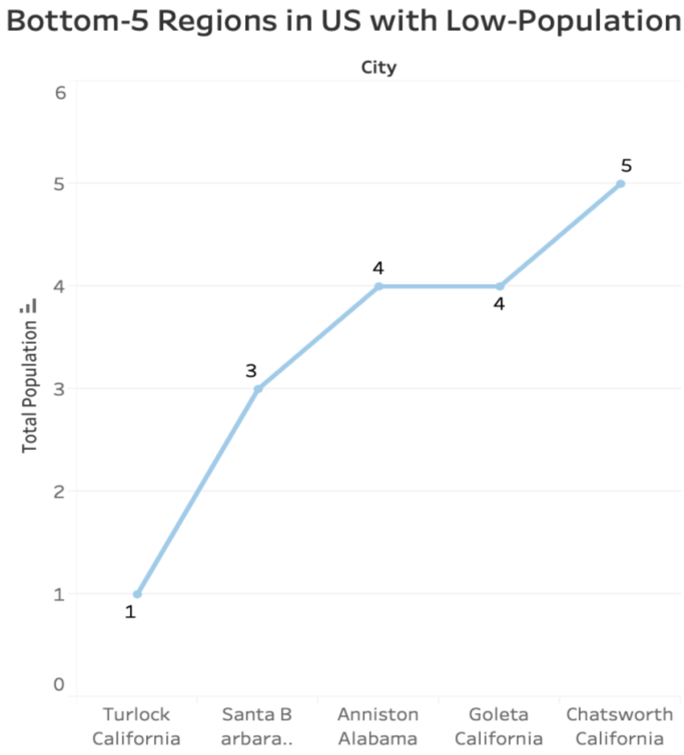
The above bar graph depicts the highly populated census blocks in the United States. And it is observed that the blocks in the City of New York is densely populated with 40,616 people, followed by Los Angeles which has a population of 19,825.

### 8. Query to find the bottom 5 Census Blocks with Lowest Population

SELECT w.census_group,w.total_population,g.latitude,g.longitude
FROM   Sex_Age w,geographic_data g
where w.census_group=g.census_group AND w.total_population != 0 AND w.census_group is not null
ORDER BY w.total_population ASC limit 5;



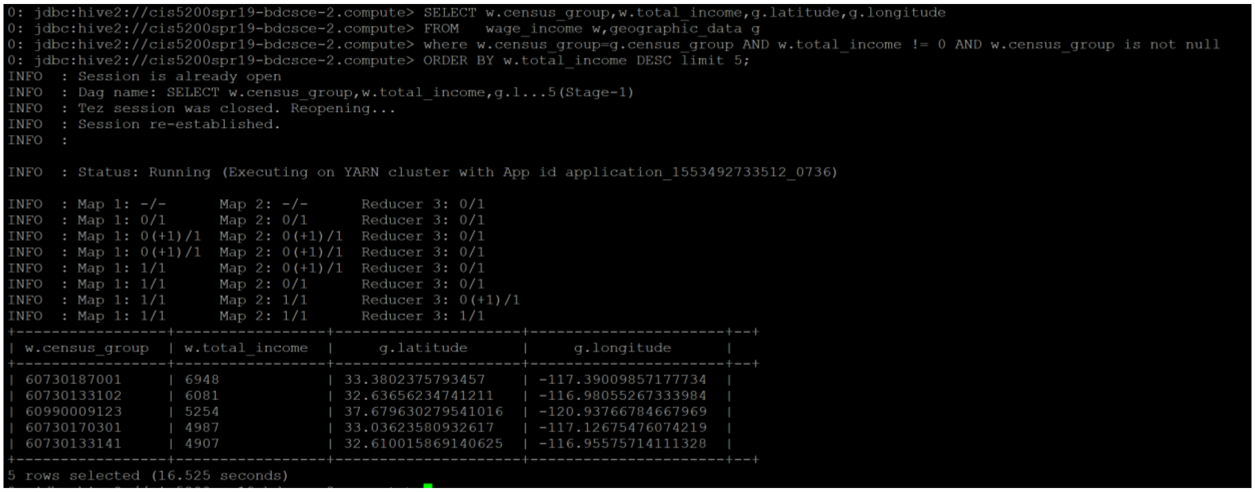
```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> SELECT w.census_group,w.total_population,g.latitude,g.longitude
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> FROM   Sex_Age w,geographic_data g
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> where w.census_group=g.census_group AND w.total_population != 0 AND w.census_group is not null
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> ORDER BY w.total_population DESC limit 5;
INFO  : Session is already open
INFO  : Dag name: SELECT w.census_group,w.total_population...5(Stage-1)
INFO  :

INFO  : Status: Running (Executing on YARN cluster with App id application_1553492733512_0735)

INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1/1
+-----------------+---------------------+---------------------+----------------------+--+
| w.census_group  | w.total_population  |     g.latitude      |     g.longitude      |
+-----------------+---------------------+---------------------+----------------------+--+
| 60730187001     | 40616               | 33.3802375793457    | -117.39009857177734  |
| 60730133102     | 19825               | 32.63656234741211   | -116.98055267333984  |
| 60730133141     | 19151               | 32.610015869140625  | -116.95575714111328  |
| 60770035003     | 16867               | 38.01222610473633   | -121.26500701904297  |
| 60730100142     | 16853               | 32.57878112792969   | -117.01876831054688  |
+-----------------+---------------------+---------------------+----------------------+--+
5 rows selected (10.803 seconds)
```



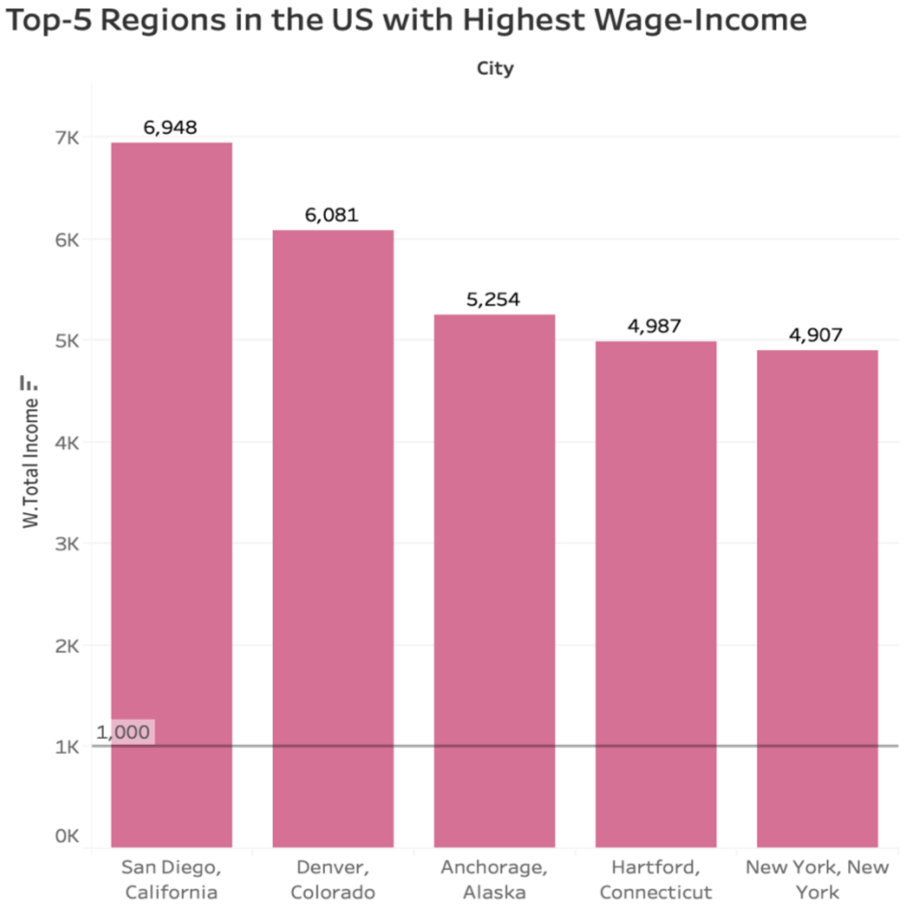
The above line graph represents the low populated census blocks in the United States. And it is observed that the blocks in the City of Turlock has less people per square area which is further followed by Goleta and Chatsworth.

### 9. Query to find the top 5 Census Blocks with Highest Wage-Income

```
SELECT w.census_group,w.total_income,g.latitude,g.longitude
FROM   wage_income w,geographic_data g
where w.census_group=g.census_group AND w.total_income != 0 AND w.census_group is not null
ORDER BY w.total_income DESC limit 5;
```



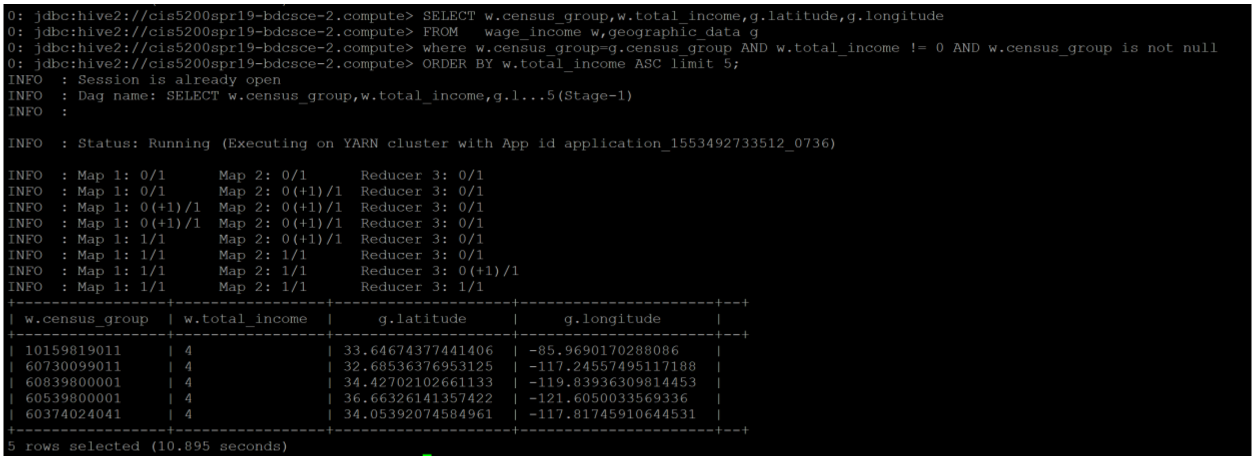
```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> SELECT w.census_group,w.total_income,g.latitude,g.longitude
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> FROM   wage_income w,geographic_data g
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> where w.census_group=g.census_group AND w.total_income != 0 AND w.census_group is not null
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> ORDER BY w.total_income DESC limit 5;
INFO  : Session is already open
INFO  : Dag name: SELECT w.census_group,w.total_income,g.l...5(Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  :

INFO  : Status: Running (Executing on YARN cluster with App id application_1553492733512_0736)

INFO  : Map 1: -/-      Map 2: -/-      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1/1
+-----------------+-----------------+---------------------+----------------------+--+
| w.census_group  | w.total_income  |      g.latitude     |      g.longitude     |
+-----------------+-----------------+---------------------+----------------------+--+
| 60730187001     | 6948            | 33.3802375793457    | -117.39009857177734  |
| 60730133102     | 6081            | 32.63656234741211   | -116.98055267333984  |
| 60990009123     | 5254            | 37.679630279541016  | -120.93766784667969  |
| 60730170301     | 4987            | 33.03623580932617   | -117.12675476074219  |
| 60730133141     | 4907            | 32.610015869140625  | -116.95575714111328  |
+-----------------+-----------------+---------------------+----------------------+--+
5 rows selected (16.525 seconds)
```



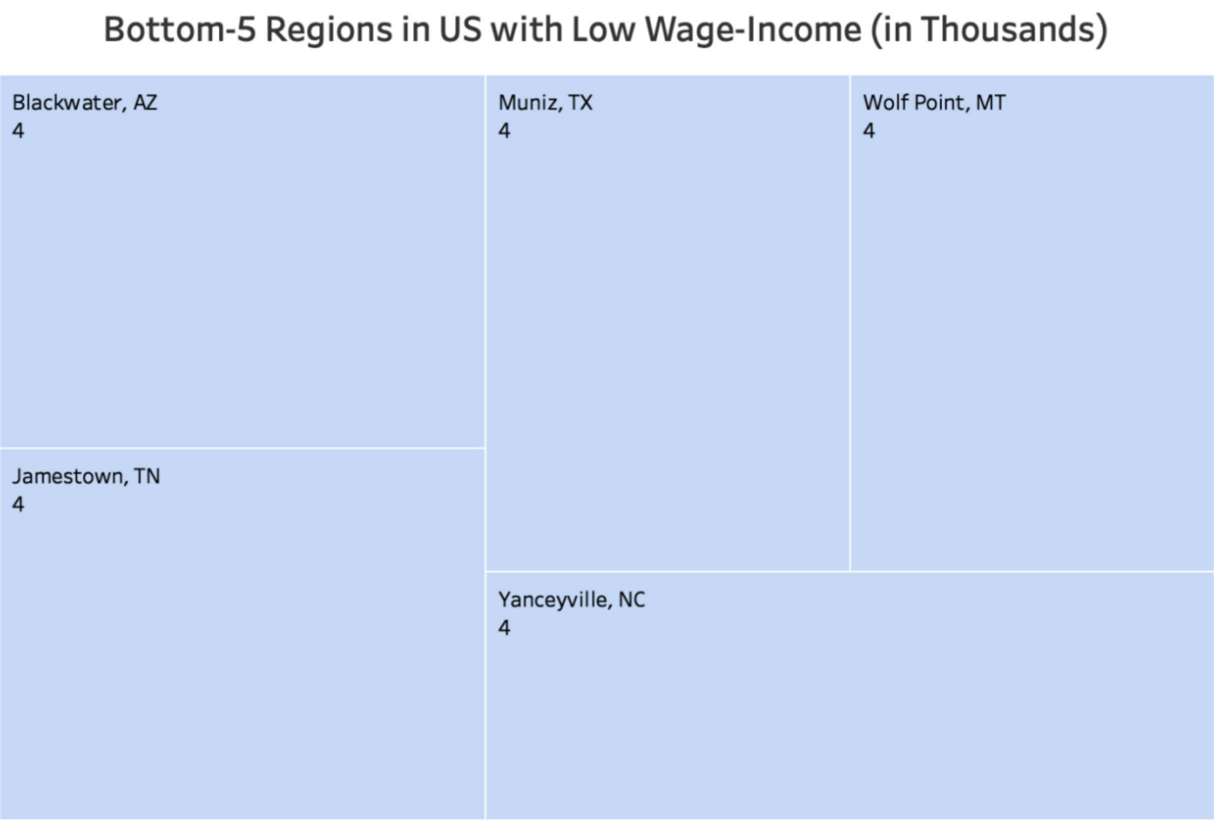
The above bar graph represents the Wage-Income corresponding to different regions. It is observed that blocks in San Diego have the highest wage income of $6948 followed by Denver, Colorado with a wage income of $6081.

### 10. Query to find the bottom 5 Census Blocks with Lowest Population

```
SELECT w.census_group,w.total_income,g.latitude,g.longitude
FROM   wage_income w,geographic_data g
where w.census_group=g.census_group AND w.total_income != 0 AND w.census_group is not null
ORDER BY w.total_income ASC limit 5;
```



```
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> SELECT w.census_group,w.total_income,g.latitude,g.longitude
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> FROM   wage_income w,geographic_data g
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> where w.census_group=g.census_group AND w.total_income != 0 AND w.census_group is not null
0: jdbc:hive2://cis5200spr19-bdcsce-2.compute> ORDER BY w.total_income ASC limit 5;
INFO  : Session is already open
INFO  : Dag name: SELECT w.census_group,w.total_income,g.l...5(Stage-1)
INFO  :

INFO  : Status: Running (Executing on YARN cluster with App id application_1553492733512_0736)

INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/1
INFO  : Map 1: 0/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1/1
+-----------------+-----------------+---------------------+-----------------------+--+
| w.census_group  | w.total_income  |     g.latitude      |      g.longitude      |
+-----------------+-----------------+---------------------+-----------------------+--+
| 10159819011     | 4               | 33.64674377441406   | -85.9690170288086     |
| 60730099011     | 4               | 32.68536376953125   | -117.24557495117188   |
| 60839800001     | 4               | 34.42702102661133   | -119.83936309814453   |
| 60539800001     | 4               | 36.66326141357422   | -121.6050033569336    |
| 60374024041     | 4               | 34.05392074584961   | -117.81745910644531   |
+-----------------+-----------------+---------------------+-----------------------+--+
5 rows selected (10.895 seconds)
```



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.

# References

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

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