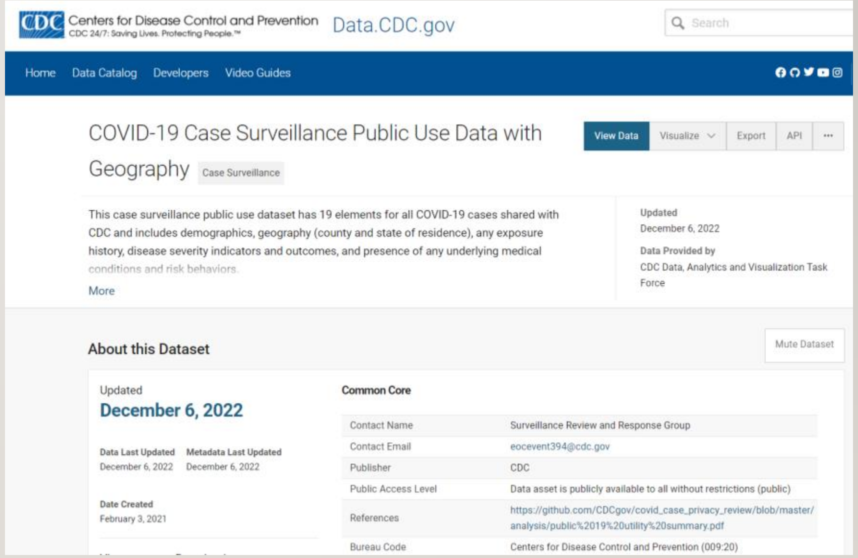
COVID-19 CASE SURVEILLANCE DATA

BY: TONY FONG, VICTOR VERDUZCO, PABLO ILABACA, MICHAEL MEDINA, JOHN MARTINEZ

GROUP 3

BACKGROUND INFORMATION ABOUT THE DATASET

- The dataset was from the Center for Disease and Control Prevention (CDC), and it shows data on Covid-19 cases.
- Link to Data Set: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4
- Data Set size: 11.2 GB



COVID-19 Case Surveillance Public Use Data with Geography

Case Surveillance

This case surveillance public use dataset has 19 elements for all COVID-19 cases shared with CDC and includes demographics, geography (county and state of residence), any exposure history, disease severity indicators and outcomes, and presence of any underlying medical conditions and risk behaviors.

More

Updated
December 6, 2022

Data Provided by
CDC Data, Analytics and Visualization Task Force

About this Dataset

Updated
December 6, 2022

Data Last Updated
December 6, 2022

Metadata Last Updated
December 6, 2022

Date Created
February 3, 2021

Common Core

Contact Name	Surveillance Review and Response Group
Contact Email	eocevent394@cdc.gov
Publisher	CDC
Public Access Level	Data asset is publicly available to all without restrictions (public)
References	https://github.com/CDCgov/covid_case_privacy_review/blob/master/analysis/public%2019%20utility%20summary.pdf
Bureau Code	Centers for Disease Control and Prevention (009:20)

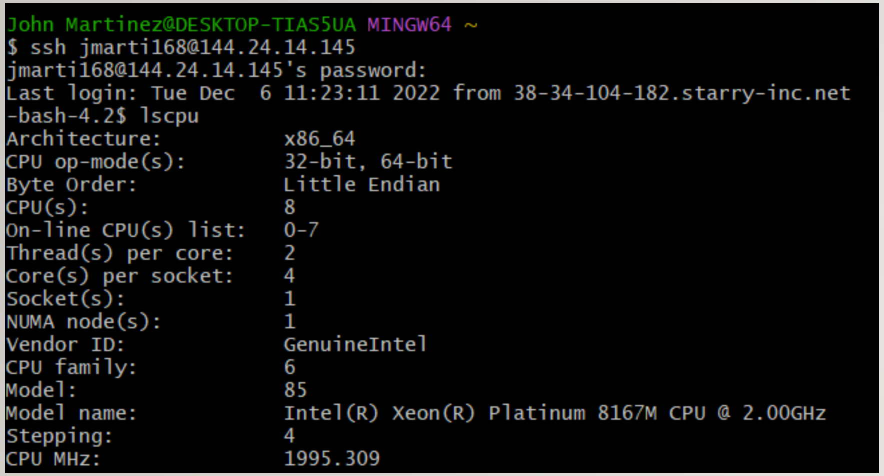
HOW UNIQUE IS OUR DATASET?

- Our dataset is unique because we aren’t just looking at the number of Covid-19 cases all around the country, but we are looking at if those with underlying conditions were more affected and had harsher outcomes.
- We are also looking for the average age group that was affected the most by Covid-19, and draw up conclusions as to why this happened.

CLUSTER SPECIFICATIONS

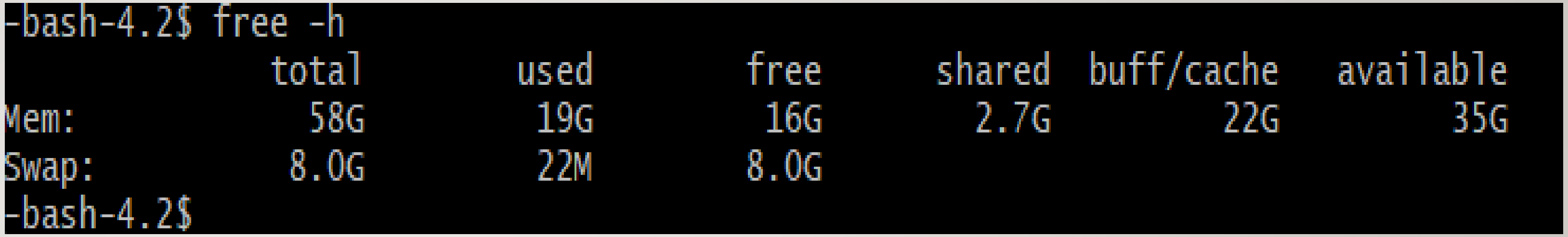
- Run the command **Hadoop version** to show the version of your Hadoop (Our version is 3.1.2)
- Run the command **lscpu** to get the CPU specifications (The number of CPUs used is 8, and the CPU speed is 1995.309 MHZ)
- Run the command **yarn node –list –all** to get the total number of nodes used (Total number of nodes is 3)
- Run the command **free –h** to get the memory specifications

# CPUs and CPU Speed



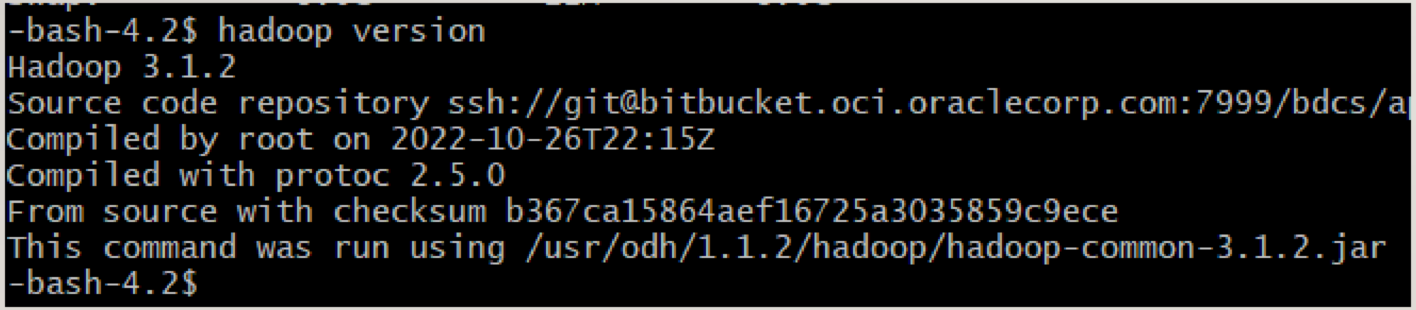
```
John Martinez@DESKTOP-TIAS5UA MINGW64 ~
$ ssh jmarti168@144.24.14.145
jmarti168@144.24.14.145's password:
Last login: Tue Dec  6 11:23:11 2022 from 38-34-104-182.starry-inc.net
-bash-4.2$ lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                8
On-line CPU(s) list:   0-7
Thread(s) per core:    2
Core(s) per socket:    4
Socket(s):             1
NUMA node(s):          1
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 85
Model name:            Intel(R) Xeon(R) Platinum 8167M CPU @ 2.00GHz
Stepping:              4
CPU MHz:               1995.309
```

Memory Size



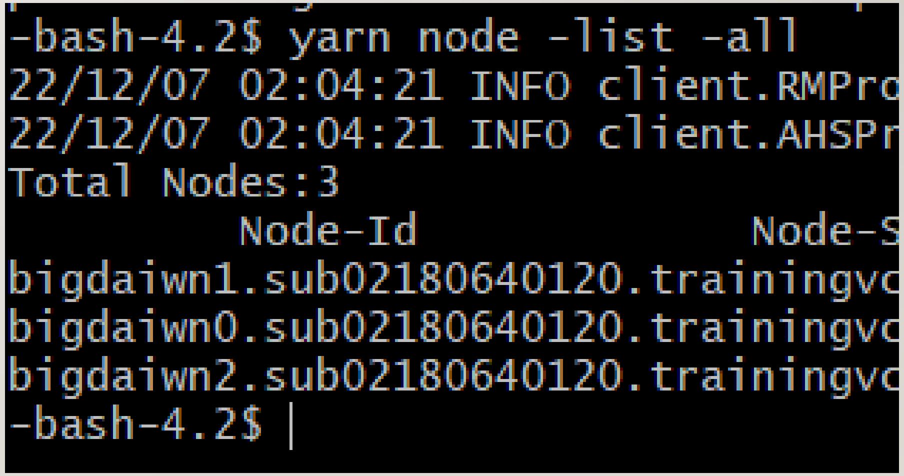
```
-bash-4.2$ free -h
              total        used        free      shared  buff/cache   available
Mem:            58G         19G         16G        2.7G         22G         35G
Swap:          8.0G         22M        8.0G
-bash-4.2$
```

Hadoop Version



```
-bash-4.2$ hadoop version
Hadoop 3.1.2
Source code repository ssh://git@bitbucket.oci.oraclecorp.com:7999/bdcs/a
Compiled by root on 2022-10-26T22:15Z
Compiled with protoc 2.5.0
From source with checksum b367ca15864aef16725a3035859c9ece
This command was run using /usr/odh/1.1.2/hadoop/hadoop-common-3.1.2.jar
-bash-4.2$
```

# of Nodes



```
-bash-4.2$ yarn node -list -all
22/12/07 02:04:21 INFO client.RMPro
22/12/07 02:04:21 INFO client.AHSPr
Total Nodes:3
         Node-Id                Node-S
bigdaiwn1.sub02180640120.trainingvc
bigdaiwn0.sub02180640120.trainingvc
bigdaiwn2.sub02180640120.trainingvc
-bash-4.2$ |
```

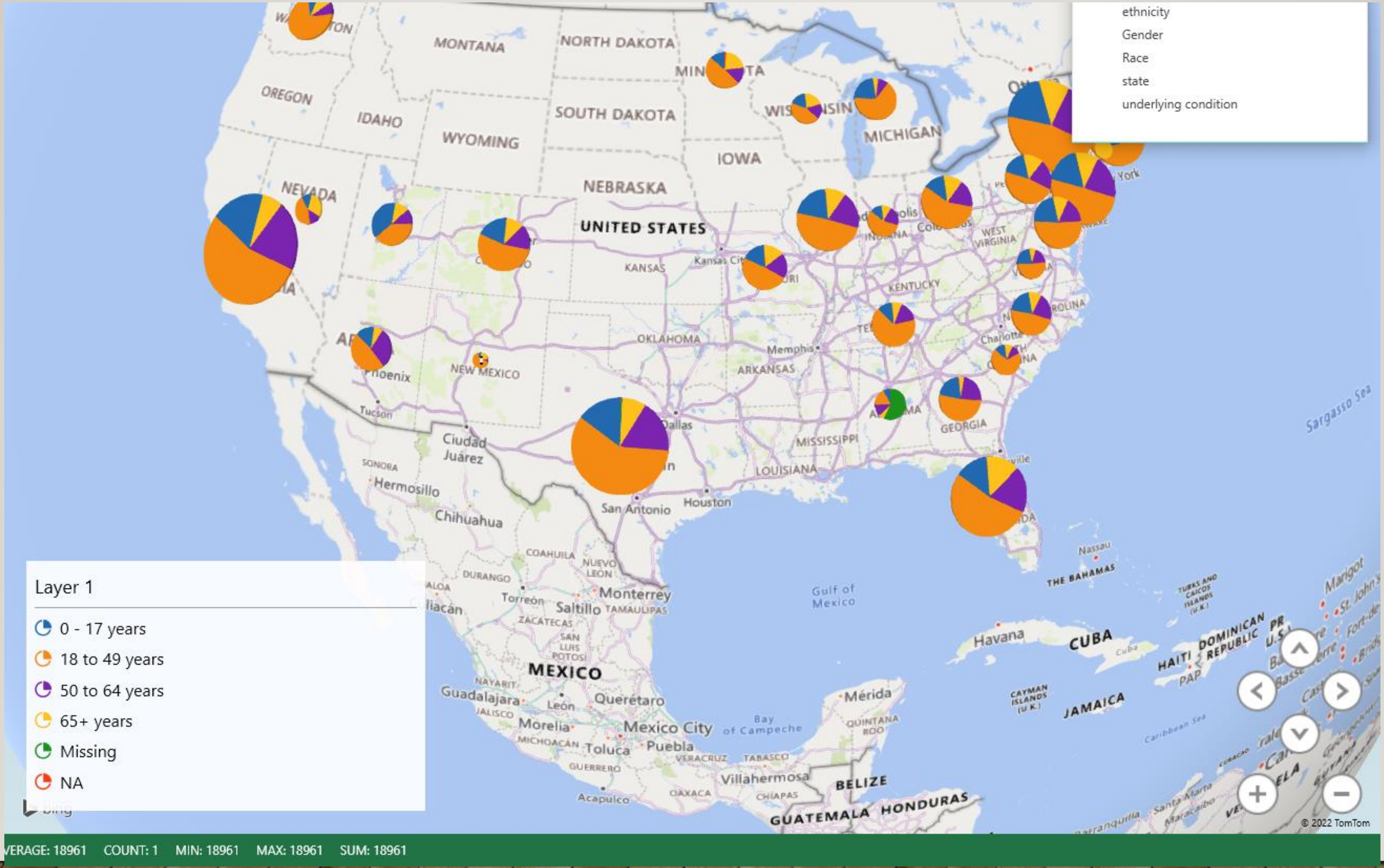
UPLOADING THE DATASET TO HDFS

- We found two ways to upload the dataset to HDFS:

1. Using the command **split -l [number] [filename]** and upload to Google Drive
2. Uploading the dataset tsv or csv file to Google Drive (since they provide 15Gb of storage data) and running the command below to upload it onto HDFS:

```
wget --load-cookies /tmp/cookies.txt
"https://docs.google.com/uc?export=download&confirm=$(wget --
quiet --save-cookies /tmp/cookies.txt --keep-session-cookies --no-
check-certificate
'https://docs.google.com/uc?export=download&id=FILEID' -O- |
sed -rn 's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')&id=1s-
9aKPqcQq8id8oGgW6HBCQzuXKBR1xO"  -O covid19data.csv &&
rm -rf /tmp/cookies.txt
```

Visualization of Covid-19 Cases Per Age Group



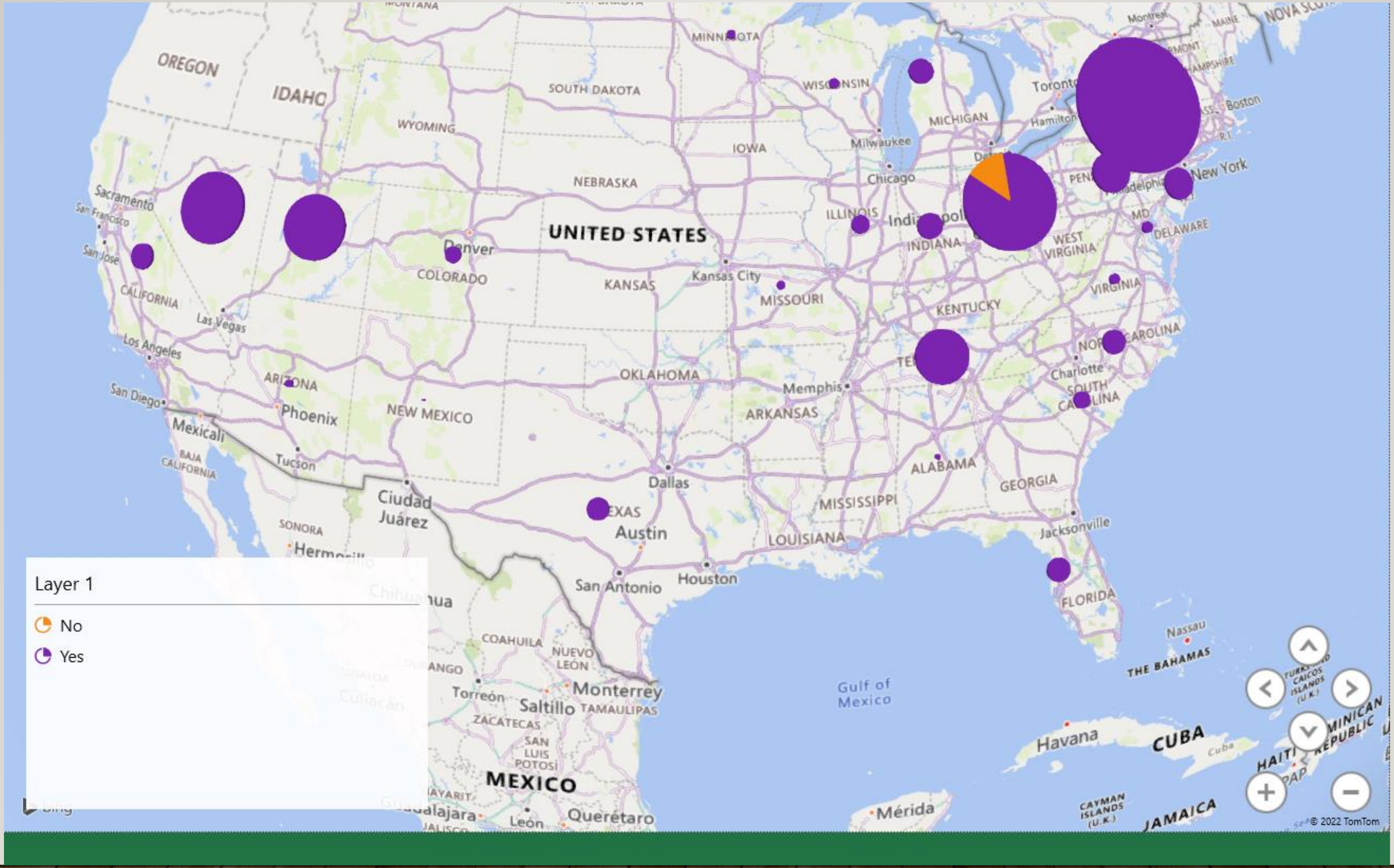
ethnicity
Gender
Race
state
underlying condition

Layer 1

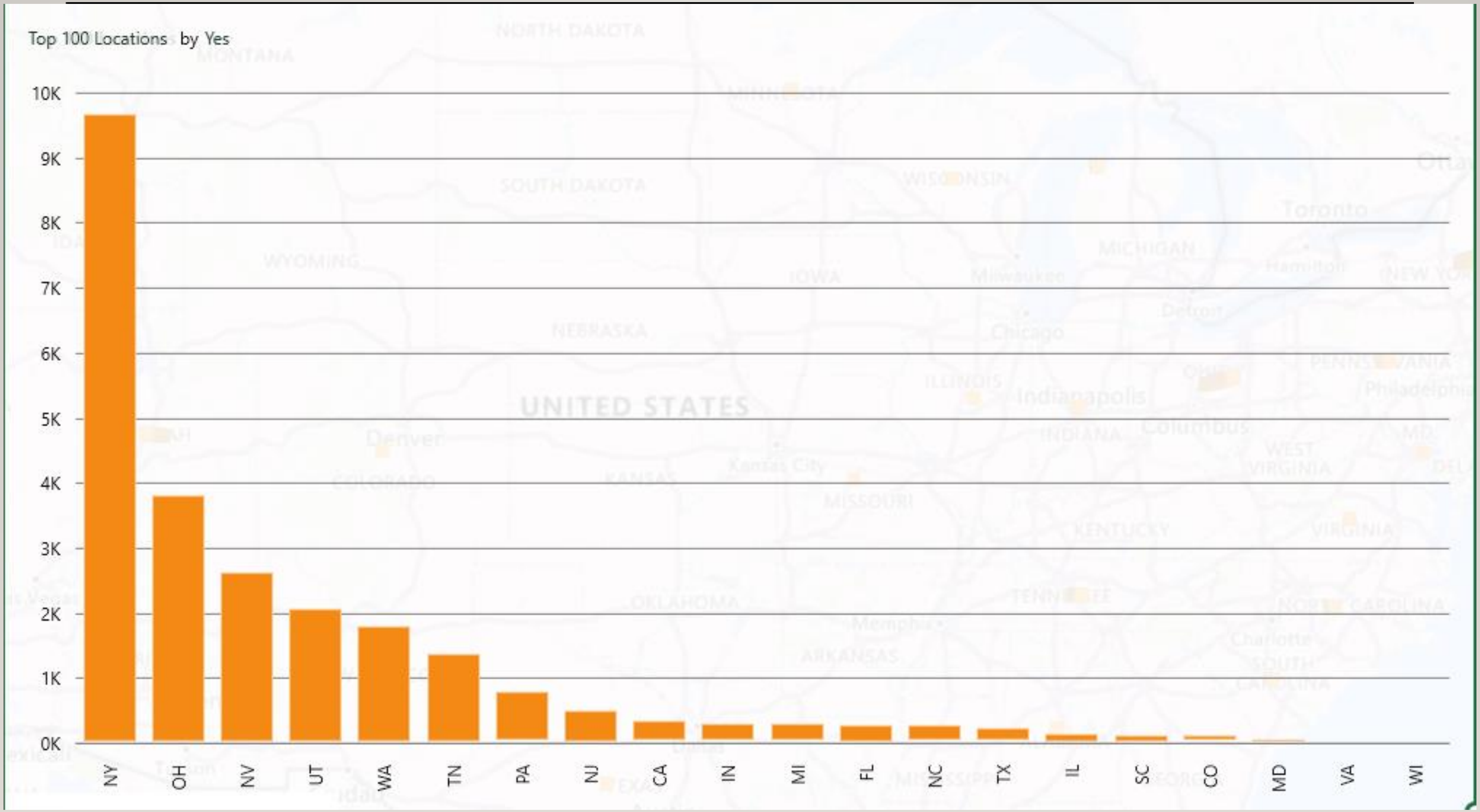
- ◔ 0 - 17 years
- ◔ 18 to 49 years
- ◔ 50 to 64 years
- ◔ 65+ years
- ◔ Missing
- ◔ NA

VERAGE: 18961 COUNT: 1 MIN: 18961 MAX: 18961 SUM: 18961

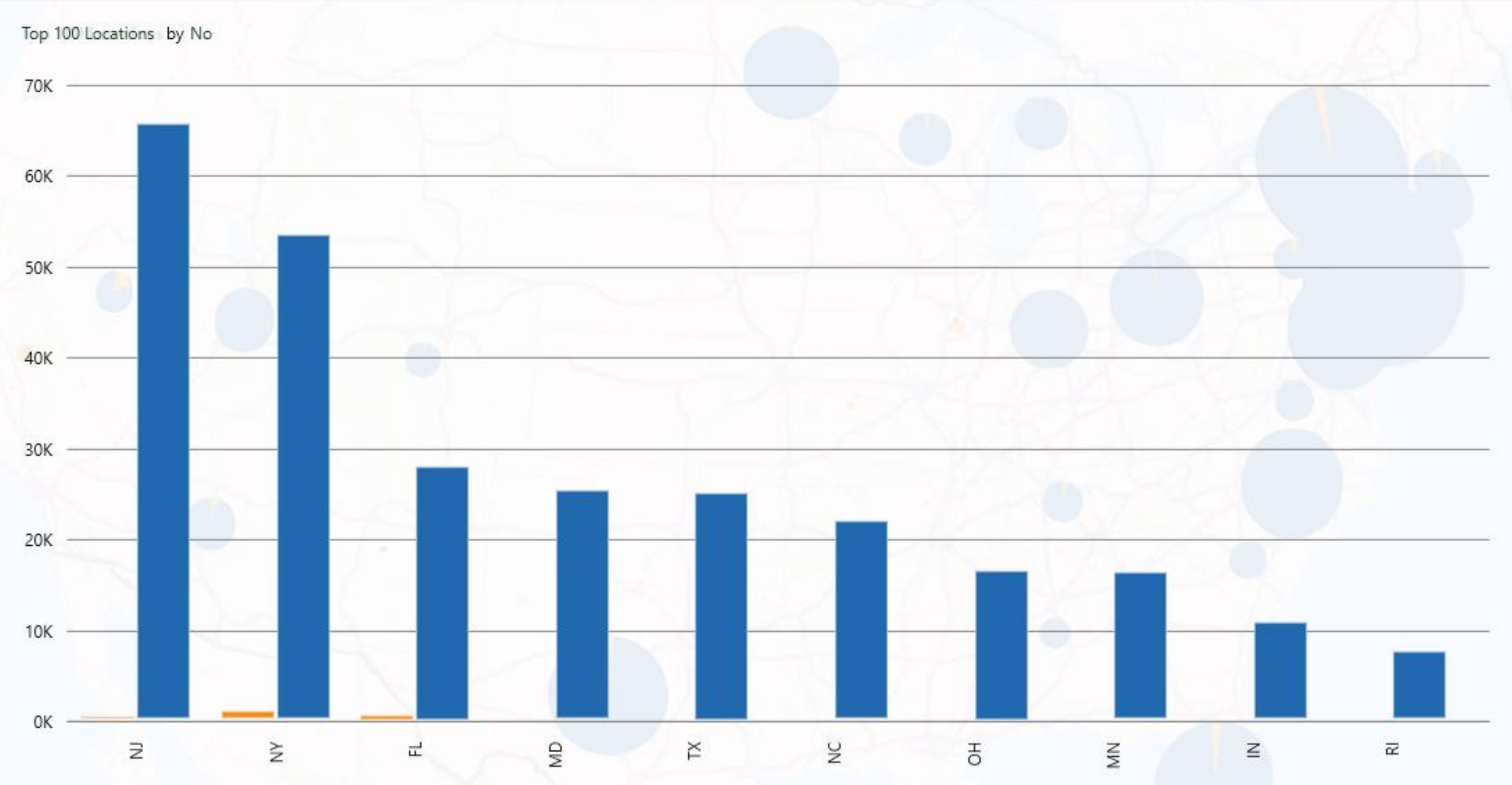
Visualization of Covid-19 Cases With Reported Underlying Conditions



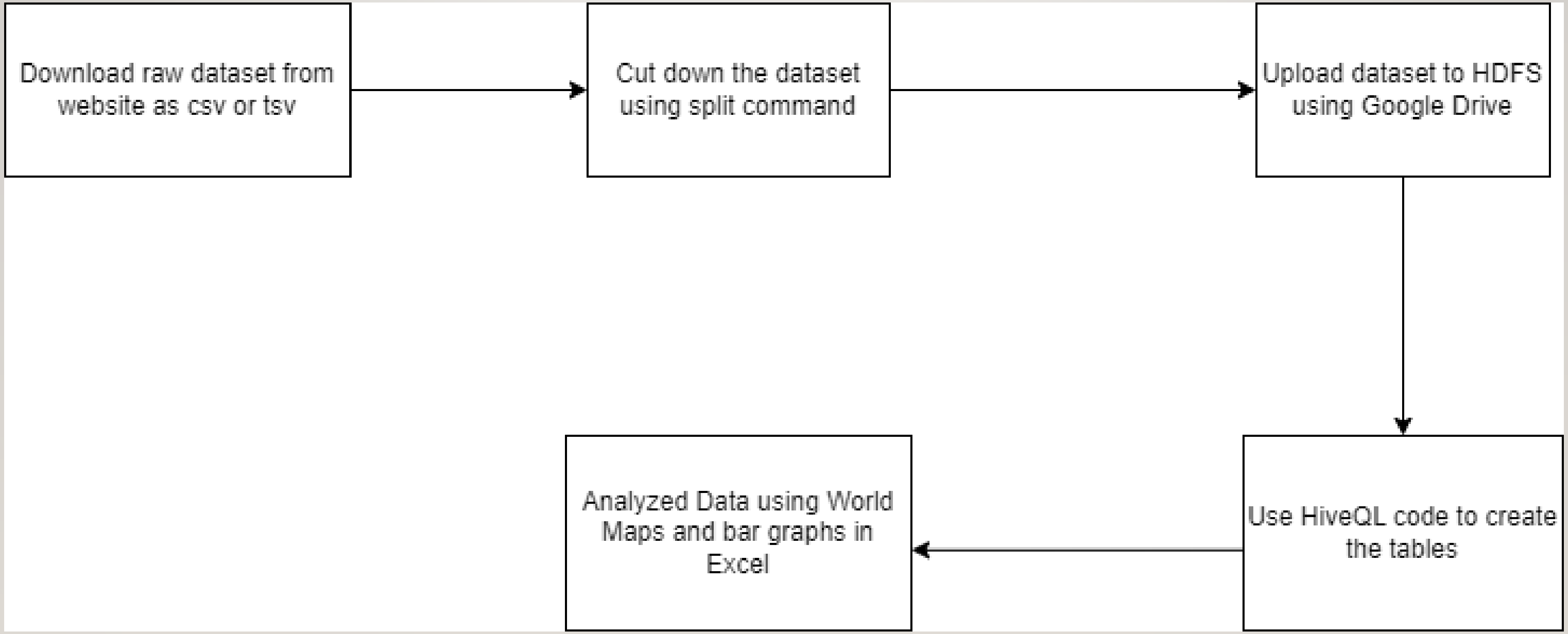
Visualization of Covid-19 Cases With Reported Underlying Conditions



Visualization of Covid-19 Cases Per State and Death Count



WORKFLOW CHART



Download raw dataset from website as csv or tsv

Cut down the dataset using split command

Upload dataset to HDFS using Google Drive

Use HiveQL code to create the tables

Analyzed Data using World Maps and bar graphs in Excel

UNDERSTANDING EPIDEMIC DATA AND STATISTICS: A CASE STUDY OF COVID-19

- Tries to interpret the number of people being infected in each country
- Evaluate how effective the policies each country implements in lowering cases
- Examines covid confirmed cases, recovered, and deaths
- Each country's policies and reactions to the outbreak determine how fast it spread.

DATA MINING AND ANALYSIS OF SCIENTIFIC RESEARCH DATA RECORDS ON COVID-19 MORTALITY, IMMUNITY, AND VACCINE DEVELOPMENT

- Purpose of the research was to investigate and determine early warning systems developed in previous epidemic responses to contain the virus from spreading.
- Examined Covid-19 scientific literature regarding Covid-19 mortality, vaccines, and immunity via data mining.
- Bibliometric analysis was done using the Web of Science Analysis Results tool to search the most dominant keywords and related concepts with Covid-19.
- Factorial analysis was done using R Studio to examine the correlation between different concepts (mortality, immunity, & vaccine development) as well as generate visualizations such as tree maps and conceptual structure maps.

GITHUB LINK

- GitHub Link: https://github.com/mike0nthemic/G5_Big_Data_4560

WORK CITED

- Hoseinpour Dehkordi, A., Alizadeh, M., Derakhshan, P., Babazadeh, P., & Jahandideh, A. (2020). Understanding epidemic data and statistics: A case study of COVID-19. Journal of medical virology, 92(7), 868–882. https://doi.org/10.1002/jmv.25885
- Radanliev, P., De Roure, D., & Walton, R. (2020). Data mining and analysis of scientific research data records on Covid-19 mortality, immunity, and vaccine development-In the first wave of the Covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(5), 1121-1132.
- Phucharoen, C., Sangkaew, N., & Stosic, K. (2020). The characteristics of