**[Michael Leggett](https://app.slack.com/team/U01MS4EHYMB" \t "_blank)**  [11:54 PM](https://gwuvirtdatapt-okg9923.slack.com/archives/C026HF4TM0V/p1626494082022500)

If anyone can meet tomorrow that would be great. I'm not sure meeting on Sunday is going to give us enough time to get going on what we need to. I met with Udit again today. He pointed out a lot of fundamental problems with our main question and data. I want to go to office hours to talk to Zeb at noon and possibly Gabe in the morning if anyone else would like to join. I'll just skip to the meat of what I propose we need to work on:

1. Please look at the files I have in my branch: 'cdc\_db\_cleaned\_part1.csv', 'cdc\_db\_cleaned\_part2.csv', and 'cdc\_main\_database.ipynb'. The CSV files are ALL the states split into two separate files as to avoid one massive file which I wasn't able to export from pyspark anyway. You will notice there are only 4 columns: res\_state, age\_group, sex, and race. I am happy to explain why at some other point but basically most of the other columns fell into two categories: not helpful for analysis and/or there were way too many missing values and therefore the column was required to be dropped. If we can get this run through the SQL script and formatted the way we need to by state, that would be awesome!
2. You will also notice two other files that we will need to use: 'age\_sex.csv' and 'race.csv'. These came directly from the Census and the links to the tables can be found in the readme in my branch under Census Data. This will need to be merged with the final data table by state. Why do we need this?  We need to have the age groups with covid AND the age groups without covid; we need to have the genders with covid AND the genders without covid; we need the races with covid AND the races without covid in order for the ML model to do any sort of analysis. The data from the CDC gives us the factors with covid, the Census data gives us the factors without covid. For example, with the CDC data we know how many males have covid in each state - but we also need to know how many males didn't have covid in each state. That is what the Census data will provide us (with some subtraction of how many males had covid). The only work that needs to be done on the census data (as far as I am currently aware) is to reorganize the ages given into the same age groups that the CDC data has.
3. The NA data question which Nic hammered us on might be easier to handle than expected. Udit thinks that using one hot encoding will be enough to handle the NA values and we won't have to do any fancy SMOTE oversampling or filling in values with medians or whatever. This is mainly the point I want to talk to Zeb about tomorrow.
4. Udit had some good ideas about how we can import the state mask mandate policy so remind me to talk about that at some point.
5. Finally, about our target variable. We really don't have one. We kept saying "whether or not a person has covid" but the only way that would be a viable target variable would be with data on every individual person in the country with a variable that says 'covid\_yn' and the values being Yes or No on whether or not the person has covid. We can't do that since that data just doesn't exist and would be 300mil+ rows. We need a very specific column name that we will have in our data. Udit suggested that our target variable be calculating the average # of covid cases per day per state or the # of covid patients. I'm not sure I 100% understand what the model is predicting in this case, which is another discussion point I will have with Zeb/Gabe in office hours.

**Today**

**New**



[**David Watson**](https://app.slack.com/team/U01MN48NC22)  [10:19 AM](https://gwuvirtdatapt-okg9923.slack.com/archives/C026HF4TM0V/p1626531545023300)

I'll be on and off the computer between now and 5:30 p.m. today and then online after 11 a.m. tomorrow



[**David Watson**](https://app.slack.com/team/U01MN48NC22)  [11:12 AM](https://gwuvirtdatapt-okg9923.slack.com/archives/C026HF4TM0V/p1626534759025200)

So immediate action items: 1. run through the SQL script (both CDC CSVs) and formatted by state. 2. reorganize Census data ages given into the same age groups that the CDC data has. 3. Need to decide on target variable and then populate a column with data on whatever we decided.

**import** **os**

*# Find the latest version of spark 3.0 from http://www-us.apache.org/dist/spark/ and enter as the spark version*

*# For example:*

*# spark\_version = 'spark-3.1.2'*

spark\_version = 'spark-3.1.2'

os.environ['SPARK\_VERSION']=spark\_version

*# Install Spark and Java*

!apt-get update

!apt-get install openjdk-11-jdk-headless -qq > /dev/null

!wget -q http://www-us.apache.org/dist/spark/$SPARK\_VERSION/$SPARK\_VERSION-bin-hadoop2.7.tgz

!tar xf $SPARK\_VERSION-bin-hadoop2.7.tgz

!pip install -q findspark

*# Set Environment Variables*

**import** **os**

os.environ["JAVA\_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"

os.environ["SPARK\_HOME"] = f"/content/**{**spark\_version**}**-bin-hadoop2.7"

*# Start a SparkSession*

**import** **findspark**

findspark.init()

Hit:1 https://cloud.r-project.org/bin/linux/ubuntu bionic-cran40/ InRelease

Ign:2 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86\_64 InRelease

Get:3 http://security.ubuntu.com/ubuntu bionic-security InRelease [88.7 kB]

Ign:4 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86\_64 InRelease

Hit:5 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic InRelease

Hit:6 http://archive.ubuntu.com/ubuntu bionic InRelease

Hit:7 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86\_64 Release

Hit:8 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86\_64 Release

Get:9 http://archive.ubuntu.com/ubuntu bionic-updates InRelease [88.7 kB]

Hit:10 http://ppa.launchpad.net/cran/libgit2/ubuntu bionic InRelease

Hit:13 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu bionic InRelease

Get:14 http://archive.ubuntu.com/ubuntu bionic-backports InRelease [74.6 kB]

Hit:15 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic InRelease

Fetched 252 kB in 2s (102 kB/s)

Reading package lists... Done

In [29]:

*# Start Spark session*

**from** **pyspark.sql** **import** SparkSession

spark = SparkSession.builder.appName("DataFrameBasics").getOrCreate()

In [30]:

*# Read in data from S3 Buckets*

**from** **pyspark** **import** SparkFiles

url = "https://cdc-main-database.s3.amazonaws.com/COVID-19\_Case\_Surveillance\_Public\_Use\_Data\_with\_Geography.csv"

spark.sparkContext.addFile(url)

df = spark.read.csv(SparkFiles.get("COVID-19\_Case\_Surveillance\_Public\_Use\_Data\_with\_Geography.csv"), header=**True**)

In [31]:

df.show()

+----------+---------+---------------+----------+----------------+--------------+------+-------+---------+-------------------------------+-------------------+--------------------+-----------+--------------------+--------------+-------+-------+--------+------------------------+

|case\_month|res\_state|state\_fips\_code|res\_county|county\_fips\_code| age\_group| sex| race|ethnicity|case\_positive\_specimen\_interval|case\_onset\_interval| process|exposure\_yn| current\_status|symptom\_status|hosp\_yn| icu\_yn|death\_yn|underlying\_conditions\_yn|

+----------+---------+---------------+----------+----------------+--------------+------+-------+---------+-------------------------------+-------------------+--------------------+-----------+--------------------+--------------+-------+-------+--------+------------------------+

| 2020-08| KY| 21| GRAYSON| 21085| 0 - 17 years| NA| NA| NA| null| 0| Missing| Missing|Laboratory-confir...| Symptomatic| No|Missing| No| null|

| 2021-01| NC| 37| MARTIN| 37117| 0 - 17 years| NA| NA| NA| 0| null| Missing| Unknown|Laboratory-confir...| Asymptomatic| No|Unknown| No| null|

| 2021-03| VT| 50| FRANKLIN| 50011|18 to 49 years| NA| NA| NA| 0| 0| Missing| Yes|Laboratory-confir...| Symptomatic| No|Missing| No| Yes|

| 2020-09| GA| 13| CATOOSA| 13047| 0 - 17 years|Female|Unknown| Missing| null| null| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| 2021-04| PA| 42| CLINTON| 42035| 0 - 17 years|Female| NA| NA| 0| 0| Missing| Yes|Laboratory-confir...| Symptomatic| No| No| NA| null|

| 2020-10| PA| 42| CRAWFORD| 42039| 0 - 17 years|Female| NA| NA| 0| 0| Missing| Yes|Laboratory-confir...| Symptomatic| No| No| NA| null|

| 2021-04| OR| 41| CROOK| 41013| 0 - 17 years|Female| NA| NA| null| 0| Missing| Missing|Laboratory-confir...| Symptomatic| No|Missing| No| null|

| 2020-08| AR| 05| LONOKE| 05085| 0 - 17 years|Female| NA| NA| 0| null| Missing| Missing|Laboratory-confir...| Symptomatic| No|Missing| NA| Yes|

| 2021-03| OH| 39| MEDINA| 39103| 0 - 17 years|Female| NA| NA| null| 0|Contact tracing o...| Yes| Probable Case| Symptomatic|Missing|Missing| NA| null|

| 2020-07| OK| 40| MUSKOGEE| 40101| 0 - 17 years|Female| NA| NA| null| null| Missing| Missing|Laboratory-confir...| Unknown| No|Missing| No| null|

| 2020-12| ME| 23| OXFORD| 23017| 0 - 17 years|Female| NA| NA| null| null| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| No| null|

| 2020-12| OR| 41| POLK| 41053| 0 - 17 years|Female| NA| NA| null| 0| Missing| Missing| Probable Case| Symptomatic| No|Missing| No| null|

| 2020-08| NC| 37| RICHMOND| 37153| 0 - 17 years|Female| NA| NA| 0| null| Missing| Unknown|Laboratory-confir...| Unknown| No|Unknown| No| null|

| 2020-12| WI| 55| ST. CROIX| 55109| 0 - 17 years|Female| NA| NA| null| 0| Missing| Missing|Laboratory-confir...| Symptomatic|Unknown|Missing| Unknown| null|

| 2020-12| NC| 37| WILKES| 37193| 0 - 17 years|Female| NA| NA| 0| 0| Missing| Unknown|Laboratory-confir...| Symptomatic|Unknown|Unknown| No| null|

| 2020-07| SC| 45| ABBEVILLE| 45001|18 to 49 years|Female| NA| NA| null| null| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| 2020-11| OK| 40| BECKHAM| 40009|18 to 49 years|Female| NA| NA| null| null| Missing| Missing|Laboratory-confir...| Unknown| No|Missing| No| null|

| 2020-04| NC| 37| BLADEN| 37017|18 to 49 years|Female| NA| NA| 1| 0| Missing| Unknown|Laboratory-confir...| Symptomatic|Unknown|Unknown| No| null|

| 2020-11| TN| 47| CARROLL| 47017|18 to 49 years|Female|Missing| Missing| 0| null| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| 2021-05| PA| 42| GREENE| 42059|18 to 49 years|Female| NA| NA| 0| null| Missing| Missing|Laboratory-confir...| Unknown|Unknown|Unknown| Unknown| null|

+----------+---------+---------------+----------+----------------+--------------+------+-------+---------+-------------------------------+-------------------+--------------------+-----------+--------------------+--------------+-------+-------+--------+------------------------+

only showing top 20 rows

In [32]:

*# Filter dates between March 2020 and December 2020*

df = df.filter((df.case\_month >= '2020-03') & (df.case\_month <= '2020-12'))

In [33]:

df.show()

+----------+---------+---------------+----------+----------------+--------------+------+-------+---------+-------------------------------+-------------------+-------------------+-----------+--------------------+--------------+-------+-------+--------+------------------------+

|case\_month|res\_state|state\_fips\_code|res\_county|county\_fips\_code| age\_group| sex| race|ethnicity|case\_positive\_specimen\_interval|case\_onset\_interval| process|exposure\_yn| current\_status|symptom\_status|hosp\_yn| icu\_yn|death\_yn|underlying\_conditions\_yn|

+----------+---------+---------------+----------+----------------+--------------+------+-------+---------+-------------------------------+-------------------+-------------------+-----------+--------------------+--------------+-------+-------+--------+------------------------+

| 2020-08| KY| 21| GRAYSON| 21085| 0 - 17 years| NA| NA| NA| null| 0| Missing| Missing|Laboratory-confir...| Symptomatic| No|Missing| No| null|

| 2020-09| GA| 13| CATOOSA| 13047| 0 - 17 years|Female|Unknown| Missing| null| null| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| 2020-10| PA| 42| CRAWFORD| 42039| 0 - 17 years|Female| NA| NA| 0| 0| Missing| Yes|Laboratory-confir...| Symptomatic| No| No| NA| null|

| 2020-08| AR| 05| LONOKE| 05085| 0 - 17 years|Female| NA| NA| 0| null| Missing| Missing|Laboratory-confir...| Symptomatic| No|Missing| NA| Yes|

| 2020-07| OK| 40| MUSKOGEE| 40101| 0 - 17 years|Female| NA| NA| null| null| Missing| Missing|Laboratory-confir...| Unknown| No|Missing| No| null|

| 2020-12| ME| 23| OXFORD| 23017| 0 - 17 years|Female| NA| NA| null| null| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| No| null|

| 2020-12| OR| 41| POLK| 41053| 0 - 17 years|Female| NA| NA| null| 0| Missing| Missing| Probable Case| Symptomatic| No|Missing| No| null|

| 2020-08| NC| 37| RICHMOND| 37153| 0 - 17 years|Female| NA| NA| 0| null| Missing| Unknown|Laboratory-confir...| Unknown| No|Unknown| No| null|

| 2020-12| WI| 55| ST. CROIX| 55109| 0 - 17 years|Female| NA| NA| null| 0| Missing| Missing|Laboratory-confir...| Symptomatic|Unknown|Missing| Unknown| null|

| 2020-12| NC| 37| WILKES| 37193| 0 - 17 years|Female| NA| NA| 0| 0| Missing| Unknown|Laboratory-confir...| Symptomatic|Unknown|Unknown| No| null|

| 2020-07| SC| 45| ABBEVILLE| 45001|18 to 49 years|Female| NA| NA| null| null| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| 2020-11| OK| 40| BECKHAM| 40009|18 to 49 years|Female| NA| NA| null| null| Missing| Missing|Laboratory-confir...| Unknown| No|Missing| No| null|

| 2020-04| NC| 37| BLADEN| 37017|18 to 49 years|Female| NA| NA| 1| 0| Missing| Unknown|Laboratory-confir...| Symptomatic|Unknown|Unknown| No| null|

| 2020-11| TN| 47| CARROLL| 47017|18 to 49 years|Female|Missing| Missing| 0| null| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| 2020-12| PA| 42| INDIANA| 42063|18 to 49 years|Female| NA| NA| 0| 0| Missing| Yes|Laboratory-confir...| Symptomatic| No|Unknown| No| null|

| 2020-03| MS| 28| MADISON| 28089|18 to 49 years|Female| NA| NA| 0| 0|Clinical evaluation| Missing|Laboratory-confir...| Symptomatic| No| No| No| Yes|

| 2020-12| PA| 42| VENANGO| 42121|18 to 49 years|Female| NA| NA| 0| null| Missing| Missing|Laboratory-confir...| Unknown|Unknown|Unknown| Unknown| null|

| 2020-11| GA| 13|WASHINGTON| 13303|18 to 49 years|Female|Unknown| Missing| null| null| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| 2020-12| TN| 47|WASHINGTON| 47179|18 to 49 years|Female|Missing| Missing| 0| null| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| 2020-06| IA| 19| BOONE| 19015|50 to 64 years|Female| NA| NA| null| 47| Missing| Missing|Laboratory-confir...| Symptomatic|Missing|Missing| NA| null|

+----------+---------+---------------+----------+----------------+--------------+------+-------+---------+-------------------------------+-------------------+-------------------+-----------+--------------------+--------------+-------+-------+--------+------------------------+

only showing top 20 rows

In [34]:

*# Drop unnecessary columns*

df = df.drop('case\_month', 'state\_fips\_code', 'res\_county', 'county\_fips\_code', 'case\_positive\_specimen\_interval', 'case\_onset\_interval')

In [35]:

df.show()

+---------+--------------+------+-------+---------+-------------------+-----------+--------------------+--------------+-------+-------+--------+------------------------+

|res\_state| age\_group| sex| race|ethnicity| process|exposure\_yn| current\_status|symptom\_status|hosp\_yn| icu\_yn|death\_yn|underlying\_conditions\_yn|

+---------+--------------+------+-------+---------+-------------------+-----------+--------------------+--------------+-------+-------+--------+------------------------+

| KY| 0 - 17 years| NA| NA| NA| Missing| Missing|Laboratory-confir...| Symptomatic| No|Missing| No| null|

| GA| 0 - 17 years|Female|Unknown| Missing| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| PA| 0 - 17 years|Female| NA| NA| Missing| Yes|Laboratory-confir...| Symptomatic| No| No| NA| null|

| AR| 0 - 17 years|Female| NA| NA| Missing| Missing|Laboratory-confir...| Symptomatic| No|Missing| NA| Yes|

| OK| 0 - 17 years|Female| NA| NA| Missing| Missing|Laboratory-confir...| Unknown| No|Missing| No| null|

| ME| 0 - 17 years|Female| NA| NA| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| No| null|

| OR| 0 - 17 years|Female| NA| NA| Missing| Missing| Probable Case| Symptomatic| No|Missing| No| null|

| NC| 0 - 17 years|Female| NA| NA| Missing| Unknown|Laboratory-confir...| Unknown| No|Unknown| No| null|

| WI| 0 - 17 years|Female| NA| NA| Missing| Missing|Laboratory-confir...| Symptomatic|Unknown|Missing| Unknown| null|

| NC| 0 - 17 years|Female| NA| NA| Missing| Unknown|Laboratory-confir...| Symptomatic|Unknown|Unknown| No| null|

| SC|18 to 49 years|Female| NA| NA| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| OK|18 to 49 years|Female| NA| NA| Missing| Missing|Laboratory-confir...| Unknown| No|Missing| No| null|

| NC|18 to 49 years|Female| NA| NA| Missing| Unknown|Laboratory-confir...| Symptomatic|Unknown|Unknown| No| null|

| TN|18 to 49 years|Female|Missing| Missing| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| PA|18 to 49 years|Female| NA| NA| Missing| Yes|Laboratory-confir...| Symptomatic| No|Unknown| No| null|

| MS|18 to 49 years|Female| NA| NA|Clinical evaluation| Missing|Laboratory-confir...| Symptomatic| No| No| No| Yes|

| PA|18 to 49 years|Female| NA| NA| Missing| Missing|Laboratory-confir...| Unknown|Unknown|Unknown| Unknown| null|

| GA|18 to 49 years|Female|Unknown| Missing| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| TN|18 to 49 years|Female|Missing| Missing| Missing| Missing|Laboratory-confir...| Missing|Missing|Missing| Missing| null|

| IA|50 to 64 years|Female| NA| NA| Missing| Missing|Laboratory-confir...| Symptomatic|Missing|Missing| NA| null|

+---------+--------------+------+-------+---------+-------------------+-----------+--------------------+--------------+-------+-------+--------+------------------------+

only showing top 20 rows

In [36]:

df.groupby('res\_state').count().orderBy('count', ascending=**False**).show()

+---------+-------+

|res\_state| count|

+---------+-------+

| CA|2517956|

| NY|1242822|

| IL|1030427|

| FL| 858632|

| OH| 714699|

| PA| 666585|

| TN| 596611|

| NC| 589914|

| AZ| 579023|

| NJ| 552458|

| GA| 516835|

| IN| 515233|

| WI| 440126|

| MN| 429299|

| MA| 386352|

| VA| 369707|

| AL| 355128|

| SC| 331485|

| IA| 327246|

| MI| 326148|

+---------+-------+

only showing top 20 rows

In [37]:

df.groupby('age\_group').count().orderBy('count', ascending=**False**).show()

+--------------+-------+

| age\_group| count|

+--------------+-------+

|18 to 49 years|9066479|

|50 to 64 years|3422181|

| 65+ years|2401830|

| 0 - 17 years|1754297|

| NA| 188666|

| Missing| 96772|

+--------------+-------+

In [38]:

df.groupby('sex').count().orderBy('count', ascending=**False**).show()

+-------+-------+

| sex| count|

+-------+-------+

| Female|8550138|

| Male|7734161|

| NA| 514000|

|Unknown| 112611|

|Missing| 19315|

+-------+-------+

In [39]:

df.groupby('race').count().orderBy('count', ascending=**False**).show()

+--------------------+-------+

| race| count|

+--------------------+-------+

| White|7250574|

| NA|2886927|

| Unknown|2772223|

| Missing|1509396|

| Black|1252891|

| Multiple/Other| 806242|

| Asian| 303508|

|American Indian/A...| 124836|

|Native Hawaiian/O...| 23628|

+--------------------+-------+

In [40]:

df.groupby('ethnicity').count().orderBy('count', ascending=**False**).show()

+-------------------+-------+

| ethnicity| count|

+-------------------+-------+

|Non-Hispanic/Latino|6922096|

| Unknown|3751967|

| NA|3471207|

| Hispanic/Latino|1574833|

| Missing|1210122|

+-------------------+-------+

In [41]:

df.groupby('process').count().orderBy('count', ascending=**False**).show()

+--------------------+--------+

| process| count|

+--------------------+--------+

| Missing|15901328|

| Clinical evaluation| 513880|

|Routine surveillance| 189659|

| Multiple| 128827|

| Laboratory reported| 82065|

|Contact tracing o...| 70802|

| Unknown| 24163|

| Other| 8776|

|Other detection m...| 8674|

| Provider reported| 1892|

|Routine physical ...| 152|

| Autopsy| 7|

+--------------------+--------+

In [42]:

df.groupby('exposure\_yn').count().orderBy('count', ascending=**False**).show()

+-----------+--------+

|exposure\_yn| count|

+-----------+--------+

| Missing|15034468|

| Yes| 1438353|

| Unknown| 457404|

+-----------+--------+

In [43]:

df.groupby('current\_status').count().orderBy('count', ascending=**False**).show()

+--------------------+--------+

| current\_status| count|

+--------------------+--------+

|Laboratory-confir...|15699194|

| Probable Case| 1231031|

+--------------------+--------+

In [44]:

df.groupby('symptom\_status').count().orderBy('count', ascending=**False**).show()

+--------------+-------+

|symptom\_status| count|

+--------------+-------+

| Symptomatic|7519108|

| Missing|6319041|

| Unknown|2763959|

| Asymptomatic| 328117|

+--------------+-------+

In [45]:

df.groupby('hosp\_yn').count().orderBy('count', ascending=**False**).show()

+-------+-------+

|hosp\_yn| count|

+-------+-------+

|Missing|7044878|

| No|6341619|

|Unknown|2622532|

| Yes| 921196|

+-------+-------+

In [46]:

df.groupby('icu\_yn').count().orderBy('count', ascending=**False**).show()

+-------+--------+

| icu\_yn| count|

+-------+--------+

|Missing|13606834|

|Unknown| 2076167|

| No| 1154405|

| Yes| 92819|

+-------+--------+

In [47]:

df.groupby('death\_yn').count().orderBy('count', ascending=**False**).show()

+--------+-------+

|death\_yn| count|

+--------+-------+

| No|8345229|

| Missing|5609568|

| Unknown|1879180|

| NA| 865827|

| Yes| 230421|

+--------+-------+

In [48]:

df.groupby('underlying\_conditions\_yn').count().orderBy('count', ascending=**False**).show()

+------------------------+--------+

|underlying\_conditions\_yn| count|

+------------------------+--------+

| null|15642277|

| Yes| 1266461|

| No| 21487|

+------------------------+--------+

In [49]:

*# Drop unnecessary columns based on number of values missing*

df = df.drop('ethnicity', 'process', 'exposure\_yn', 'current\_status', 'symptom\_status', 'hosp\_yn', 'icu\_yn', 'death\_yn', 'underlying\_conditions\_yn')

In [50]:

df.show()

+---------+--------------+------+-------+

|res\_state| age\_group| sex| race|

+---------+--------------+------+-------+

| KY| 0 - 17 years| NA| NA|

| GA| 0 - 17 years|Female|Unknown|

| PA| 0 - 17 years|Female| NA|

| AR| 0 - 17 years|Female| NA|

| OK| 0 - 17 years|Female| NA|

| ME| 0 - 17 years|Female| NA|

| OR| 0 - 17 years|Female| NA|

| NC| 0 - 17 years|Female| NA|

| WI| 0 - 17 years|Female| NA|

| NC| 0 - 17 years|Female| NA|

| SC|18 to 49 years|Female| NA|

| OK|18 to 49 years|Female| NA|

| NC|18 to 49 years|Female| NA|

| TN|18 to 49 years|Female|Missing|

| PA|18 to 49 years|Female| NA|

| MS|18 to 49 years|Female| NA|

| PA|18 to 49 years|Female| NA|

| GA|18 to 49 years|Female|Unknown|

| TN|18 to 49 years|Female|Missing|

| IA|50 to 64 years|Female| NA|

+---------+--------------+------+-------+

only showing top 20 rows

In [51]:

*# Replace "Missing" and "Unknown" with "NA"*

df = df.replace(['Missing', 'Unknown'], 'NA')

In [52]:

df.show()

+---------+--------------+------+----+

|res\_state| age\_group| sex|race|

+---------+--------------+------+----+

| KY| 0 - 17 years| NA| NA|

| GA| 0 - 17 years|Female| NA|

| PA| 0 - 17 years|Female| NA|

| AR| 0 - 17 years|Female| NA|

| OK| 0 - 17 years|Female| NA|

| ME| 0 - 17 years|Female| NA|

| OR| 0 - 17 years|Female| NA|

| NC| 0 - 17 years|Female| NA|

| WI| 0 - 17 years|Female| NA|

| NC| 0 - 17 years|Female| NA|

| SC|18 to 49 years|Female| NA|

| OK|18 to 49 years|Female| NA|

| NC|18 to 49 years|Female| NA|

| TN|18 to 49 years|Female| NA|

| PA|18 to 49 years|Female| NA|

| MS|18 to 49 years|Female| NA|

| PA|18 to 49 years|Female| NA|

| GA|18 to 49 years|Female| NA|

| TN|18 to 49 years|Female| NA|

| IA|50 to 64 years|Female| NA|

+---------+--------------+------+----+

only showing top 20 rows

In [53]:

*# Filter for 'AL', 'AK', 'AZ', 'AR', 'CA', 'CO', 'CT', 'DE', 'DC', 'FL', 'GA', 'HI', 'ID', 'IL', 'IN', 'IA', 'KS', 'KY', 'LA', 'ME', 'MD', 'MA', 'MI', 'MN', 'MS', 'MO'*

df\_part1 = df.filter((df.res\_state == ('AL' **or** 'AK' **or** 'AZ' **or** 'AR' **or** 'CA' **or** 'CO' **or** 'CT' **or** 'DE' **or** 'DC' **or** 'FL' **or** 'GA' **or** 'HI' **or** 'ID' **or** 'IL' **or** 'IN' **or** 'IA' **or** 'KS' **or** 'KY' **or** 'LA' **or** 'ME' **or** 'MD' **or** 'MA' **or** 'MI' **or** 'MN' **or** 'MS' **or** 'MO')))

In [54]:

*# Filter for 'MT', 'NE', 'NV', 'NH', 'NJ', 'NM', 'NY', 'NC', 'ND', 'OH', 'OK', 'OR', 'PA', 'RI', 'SC', 'SD', 'TN', 'TX', 'UT', 'VT', 'VA', 'WA', 'WV', 'WI', 'WY'*

df\_part2 = df.filter((df.res\_state == ('MT' **or** 'NE' **or** 'NV' **or** 'NH' **or** 'NJ' **or** 'NM' **or** 'NY' **or** 'NC' **or** 'ND' **or** 'OH' **or** 'OK' **or** 'OR' **or** 'PA' **or** 'RI' **or** 'SC' **or** 'SD' **or** 'TN' **or** 'TX' **or** 'UT' **or** 'VT' **or** 'VA' **or** 'WA' **or** 'WV' **or** 'WI' **or** 'WY')))

In [55]:

df\_part1.toPandas().to\_csv('cdc\_db\_cleaned\_part1.csv')

In [56]:

df\_part2.toPandas().to\_csv('cdc\_db\_cleaned\_part2.csv')