**COVID-19 Data Exploration**

**Excel**

1. Download dataset from <https://ourworldindata.org/covid-deaths>
   1. Data updated daily (last update around 3pm on Mar. 30, 2022 from time of download)
   2. Data ranges from 1/1/20 – 3/29/22
   3. # Records: **172799**; # Fields: **67**
2. Make copy of raw data onto new workbook
3. Cut & paste Population field into base table in order to avoid joining on every query later
4. Create desired tables (in separate files) from raw table by removing unnecessary fields
   1. coviddeaths – save as .csv which is format supported by MySQL import wizard
   2. covidvaccinations - save as .csv which is format supported by MySQL import wizard
5. Import both files into MySQL using Table Data Import Wizard

**MySQL**

* Load datasets into MySQL

Results:

METHOD 1

* coviddeaths (Import Wizard): 581.308 s; 164617 records
* covidvaccinations (Import Wizard): 743.813 s; 20492 records
  + upon closer evaluation of both datasets in MySQL, it is clear that they weren’t completely imported. There were several missing records in both and there were limitations in setting the data types for every column right in the wizard. Not to mention the extremely long amount of time it took for both datasets to be imported via this method. Its unusual because the final message didn’t mention any data import errors upon completion; you had to keep track of the “show logs” option in order to see that errors were being encountered
    - for reasons mentioned above and after further research of best methods to import datasets into MySQL, the LOAD DATA INFILE command was used. But before that, DROP both tables previously imported

METHOD 2

1st Attempt – command line

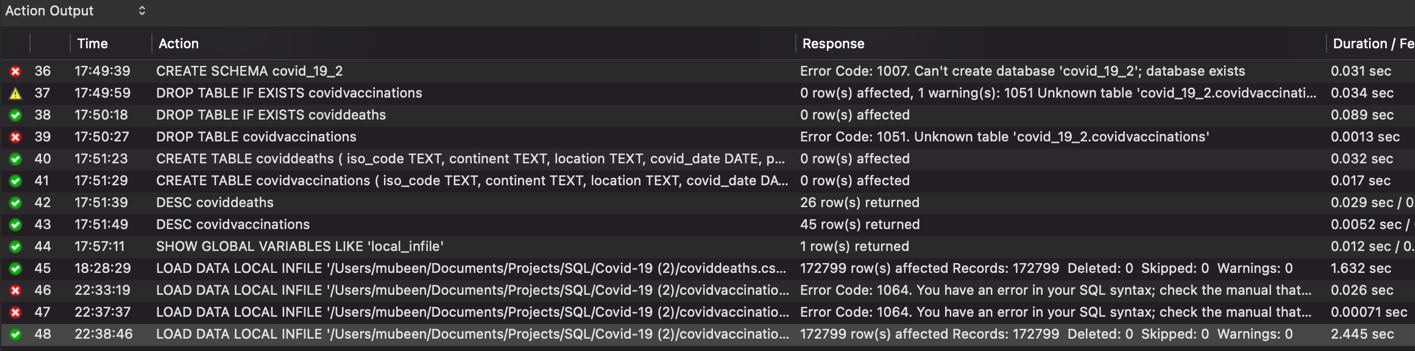
* coviddeaths (LOAD DATA): 1.91s; 172799 records
* covidvaccinations (LOAD DATA):
  + although this method worked really fast, I still noticed that the date columns for both tables weren’t being read properly. I realized that the issue lied in the differences in date formats between the input columns and MySQL’s default date formats. So, I performed some preprocessing transformations within the LOAD DATA command in order to resolve this. This entailed using user defined variables and specifying all columns of the tables in the load command
    - once again, I had to DROP both tables before recreating them

2nd Attempt – command line

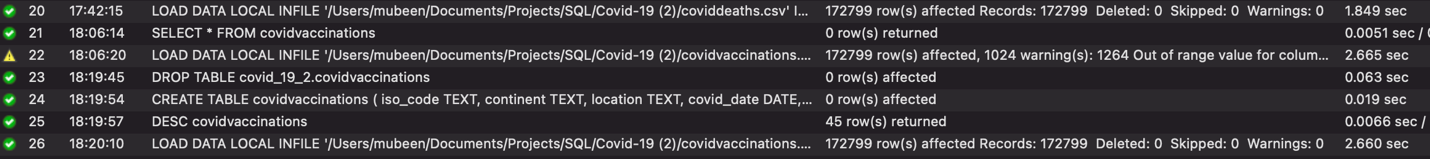
* coviddeaths (LOAD DATA): 1.83s; 172799 records; Deleted: 0 Skipped: 0 Warnings: 1480384
* covidvaccinations (LOAD DATA): 2.89s; 172799 records; Deleted: 0 Skipped: 0 Warnings: 3604158
  + although the previous issue got fixed, the new one I wanted to look deeper into was why I kept getting so many warnings even though the entire datasets seemed to completely get imported. After a deeper look, I noticed that every cell in the input file which was empty resulted in a warning message (incorrect integer value OR data truncated) for the corresponding cell in its MySQL table. This meant that NULL data wasn’t being read properly by the data types (INT & DOUBLE) that defined these cells. MySQL would represent empty cells with 0 for these fields upon mapping. So, I had to perform preprocessing transformations once again by subjecting all INT & DOUBLE columns to user defined variables and replacing every empty (‘ ‘) cell with NULL. [Help from <https://www.youtube.com/watch?v=N-ACTiCAQVM&ab_channel=MichaelKaufmann>]
  + in the midst of doing this, I was also able to configure MySQL workbench to allow LOAD DATA LOCAL INFILE commands (instead of using command line tool from before) by reading several forums. Seems that it was a bug issue within workbench. Just needed to edit the connection: add an option into the Option Box of the Advanced tab of Edit Connection [Solution: <https://bugs.mysql.com/bug.php?id=91891>]

3rd Attempt – workbench

* coviddeaths (LOAD DATA): 1.63s; 172799 records; Deleted: 0 Skipped: 0 Warnings: 0
* covidvaccinations (LOAD DATA): 2.45s; 172799 records; Deleted: 0 Skipped: 0 Warnings: 0



* + I tried fixing aforementioned issue with the REPLACE string function, but noticed something strange. Instead of replacing all empty cells only with NULL values, this function replaced all non-empty cells with NULLS as well. So, **all** INT & DOUBLE defined columns were NULL. After practicing with the function, I noticed that if from\_str wasn’t found in string and anything other than NULL was used for the to\_str, the original str kept its value. BUT, if NULL was used as replacement to\_str under same conditions, the new value always became NULL. So, CASE statements needed to be used for each of these columns: if cell empty, THEN NULL ELSE original value.
  + Also, I noticed I needed to edit the population data type from INT to BIGINT
* 4th Attempt – workbench
* coviddeaths (LOAD DATA): 1.908s; 172799 records; Deleted: 0 Skipped: 0 Warnings: 172798
* covidvaccinations (LOAD DATA): 2.762s; 172799 records; Deleted: 0 Skipped: 0 Warnings: 174801
  + the CASE statement preprocessing worked, **EXCEPT** for the last column for some reason. I was getting warnings only for the last column, and the same message for every single row “data truncated...”. After further research, I found that the issue lied in LINES TERMINATED BY section of load data command. Files that are created on Windows systems usually use two characters as line terminators, so \r\n needed to be used instead for proper reading of the files [<https://dev.mysql.com/doc/refman/8.0/en/load-data.html>, under NOTE]
  + also, I noticed I needed to edit data types for the following: total\_vaccinations, people\_vaccinated, people\_fully\_vaccinated
* 5th Attempt – workbench
* coviddeaths (LOAD DATA): 1.849s; 172799 records; Deleted: 0 Skipped: 0 Warnings: 0
* covidvaccinations (LOAD DATA): 2.660s; 172799 records; Deleted: 0 Skipped: 0 Warnings: 0



* + All data properly & completely loaded

**Quering Data**

* query data of interest

1. for U.S., look at death percentage of total positive cases

side note: numbers match results found on google (as of last day of datasets date range)

quick insights:

* first case reported on 2020-01-22
* first death reported on 2020-02-29
* death % gradually started increasing from 2020-03-24 to 2020-06-01. Then, it started dropping off until the end of 2020 and remained relatively constant for much of 2021 between 1.5-1.7%. by far, the lowest numbers have been from the start of 2022 around 1.2%.

1. total cases vs U.S. population

* as of 2022-03-29, around 24% of U.S. population has been infected by covid

1. countries with highest infection rate compared to population

* most of the countries with the highest infection rates - Andorra, Iceland, Cyprus - don’t have as big populations as some of the bigger countries like U.S. as it stands by the last date of this dataset, U.S. ranks in the lower end of top 50 highest infection rates by country
  + An observation I found in working on this query is that the location field contains both countries and continents (in addition to other miscellaneous roles like international, world, low income, etc.), although there is a separate field for continents. By looking back at the original data, I noticed that continents (in addition to the other miscellaneous roles) were listed in the location field whenever the continent field was empty (Although I took care of NULLs for numeric fields, I forgot to do so for text fields like these. So, I used an UPDATE statement to replace all empty continent cells with NULLs; had to set sql\_safe\_updates mode OFF; 10268 records)



Then, I filtered every query with a WHERE clause that excluded NULLs in the continent field so that continents were no longer appearing in the location field and data was properly being represented. Conversely, in order to get data for continents as a whole, I filtered the opposite WHERE continent IS NULL because when continent did appear in location field, it was represented as an aggregate of it appearing in the continent field.

1. Countries with highest total death count compared to population
   * U.S. has the highest death count at almost 1M, even though its not even in the top 10 for most infected countries
2. Continents with highest total death count compared to population
   * Europe has the highest death count with around 1.8M, followed by N. America with around 1.4M. Africa has the lowest with around 250K

* Added filter to exclude miscellaneous roles

1. Look at global numbers across the world

* Group by date and summing new cases and new deaths will give total numbers across all countries
* World death percentage is around 1.26%
* Total deaths around 6.1M
* Total cases around 485M

1. Total # of people in the world that have been vaccinated

* Performed a JOIN on both location & covid\_date. If only joined on one of the two, the dates or locations wouldn’t correspond for each row, giving inaccurate results
* U.S. started seeing first vaccinations around 2020-12-14. This actually corresponds with the date on The Washingpost Post news article that described the first vaccine shots being given outside trials in U.S. [[https://www.washingtonpost.com/nation/2020/12/14/first-covid-vaccines-new-york/]](https://www.washingtonpost.com/nation/2020/12/14/first-covid-vaccines-new-york/) To get the total vaccinations in each country first, I used the new\_vaccinations column so that a window function could be incorporated to keep a running total of vaccinations for each day. By doing so, the running total value for the last day in each country will represent the total number of vaccinations. I could’ve easily used the total vaccinations column to obtain the same results, but this method takes it a step further to display additional skills (to be sure, running total sum was compared to total vaccinations value for last day). NOTE: The total\_vaccinations and new\_vaccinations columns represent the total number of **doses** given, as some vaccines require more than one dose. Then, I nested the previous query within a subquery in order to obtain the total number of vaccine doses given for each country.
* U.S. has administered the 3rd most vaccine doses in the world (560M), right below China & India
* After exploring the vaccinations data a little more closely, inaccurate results seem to be apparent whenever vaccinations weren’t being reported consistently and consecutively every day for each country. That is why the number of total doses given in the world is off by almost 2B while world population is fairly accurate at around 7.85B

1. Create views to be used for visualizations

* Total deaths by continent
* Covid numbers in the world
* Percent population infected by country

**Tableau**

* Tableau Public doesn’t allow connections to MySQL databases. Ideally, the views previously created in MySQL would have been used to create the visualizations in here. Instead, each views query result set will be copied to individual Excel workbooks and then loaded into Tableau

1. Obtain result set of each MySQL view and copy to separate workbooks in Excel
   1. DeathsByContinent – no NULLs
   2. CovidNumbersWorld – no NULLs
   3. PercentPopulationInfected\_Country
      1. NULLs replaced by 0’s represents missing data
   4. PercentPopulationInfected\_Day
      1. NULLs replaced by 0’s represents days that covid did not affect country

* Since Tableau doesn’t know how to handle NULLs properly, I pretreated each data source by replacing NULLs with 0’s before loading into Tableau. See each field for meaning of NULLs replaced by 0’s

1. Add all data sources into Tableau
2. Sheet1: Global Numbers visual
3. Sheet2: Total Death Count by Continent bar chart
4. Sheet3: Map displaying percent population infected by country

* Changed location data type from string to geographic location. Dragging & dropping onto view resulted in 2 unknowns:

Timor 🡪 Timor Leste

Micronesia 🡪 Federated States of Micronesia

* Added Red-Gold color gradient because covid is linked to something dangerous

1. Sheet4: Time Series visual of percent population infected by country

* Added countries I was interested in seeing as a filter
* Instead of looking at percent population infected, chose the AVG per month
* Added a forecast to predict approximately how percentage of countries infected will look 1 year from Mar. 2022

1. Create Dashboard

* Added a filter Action where if you select a country from the map, its AVG population infected will be displayed
* Wanted to see if I could add another filter where if you selected the continent in Death Count visual, that continent will be highlighted on map