8/5/2020

Alfred D. Hull: adh8mb, Farhan Kanani, fk3ak;

Niraja Bohidar, nb4dg, Francisco Estrada, fge8tj, Dara Maguire, dkm2bd

CS 5010: Project 1

The Effect of State Humidity  
Covid-19 Cases

# **Phase I. Business Understanding: “Introduction”**

**Project Charter:**

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), also known as COVID-19, has manifested into a pandemic that has spread to 188 countries and territories with more than 17.5 million cases as of August 2020. In the United States alone, there have 4.68 million confirmed cases and rising.

Though the complete means of the method of transmission is still unclear, several scientists postulate that transmission of COVID-19 occurs primarily in droplets through the air. To combat this, health officials have recommended that people wear masks and face shields and maintain social distancing of at least 6 feet. However, many researchers believe that to help prevent transmission; indeed, we need to protect the air.

One overlooked aspect of air quality is humidity. Infection control specialist Stephanie Taylor identified that midrange humidity levels have been able to help prevent the transmission of airborne pathogens. When air is too dry, large air droplets dry out to become smaller droplets. In moister air, airborne viruses like COVID-19, do not seem to be as virulent, though the mechanism is not fully understood. Though creating a more humid environment is not a complete solution to prevent transmission, it seems to be another step in creating a safer environment.

**Who might care:**

This project may be useful for Healthcare professionals, businesses, household owners, and the general public. It would be in their best interest to consider implementing these precautions to deter the spread of COVID-19.

**Set-Objective:**

The objective is to corroborate the evidence that humidity may have a role in reducing transmission by looking at the humidity in each State and the number of cases in those states. Should our data corroborate these findings, further justification for businesses to maintain a midrange (40%-60%) humidity level in their establishments would be optimal.

**Cost and benefits:**

Though businesses may invest more in quality humidifiers to maintain this level of humidity, it may be a necessity to create a safer environment for people.

**Constraints, Limitations, and Assumptions (CLAs):**

Should we find a trend between humidity and number of COVID cases, it is essential to remember that this is evidence of correlation, not causation.

# **Phase II. Data Understanding: “The Data”**

***Initial Data Collection***: For the Data Collection, we found two websites to scrape. One website was data regarding humidity levels across all states in The United States of America. The other website was COVID-19 Cases by State, which included Total, New Cases, and Deaths. The procedure included web scraping these websites, cleaning the data, and merging them into one CSV file. This section helped lay the foundation for the data description, analysis, and testing/evaluation aspects of this project. </li>

***Data Description***: The Data Description describes the data acquired, highlighting the format and source of data. Additionally, it highlights the number of cases/variables and their descriptions. These attributes help evaluate whether the data is suitable and sufficient for the analysis.

***Exploratory Data Analysis (EDA)***: EDA examines the data more closely. It highlights the first findings and initial hypothesis of the study of interest: aggregates, Statistical Analysis, Relationships between variables, and Visualizations (graphs and plots).

***Data Testing***: The Data Testing portion utilizes the data analysis to come up with queries to reveal interesting and useful information about our data and topic of interest.

***Data Quality***: Data Quality lists our conclusions. Here we summarize our findings, explain how these results are relevant to organizations like the government and health officials.

## **Initial Data Collection Report**

The first dataset used was put together by the WorldOMeters information group. The dataset is updated daily to reflect the update count of each of the variables of interest. These include the Total Cases, Deaths, Tests, New Deaths, Cases, and Active Deaths of the COVID 19 Pandemic. The source data set is located at <https://www.worldometers.info/coronavirus/country/us/>.

The second dataset used was put together by the USA World Media Group. The dataset rank locations with Average Humidity Data by State. The dataset showcases the State, its population, and Average Humidity level in percentage. The source data set is located at [www.usa.com/rank/us--average-humidity--state-rank.htm](http://www.usa.com/rank/us--average-humidity--state-rank.htm).

The following Data Collection Report is a simple listing of the data sources acquired along with their locations, the procedures used to procure them, and any difficulties encountered—this section aides both with future replication of this project and with the execution of similar future projects.

## **Data Collection:**

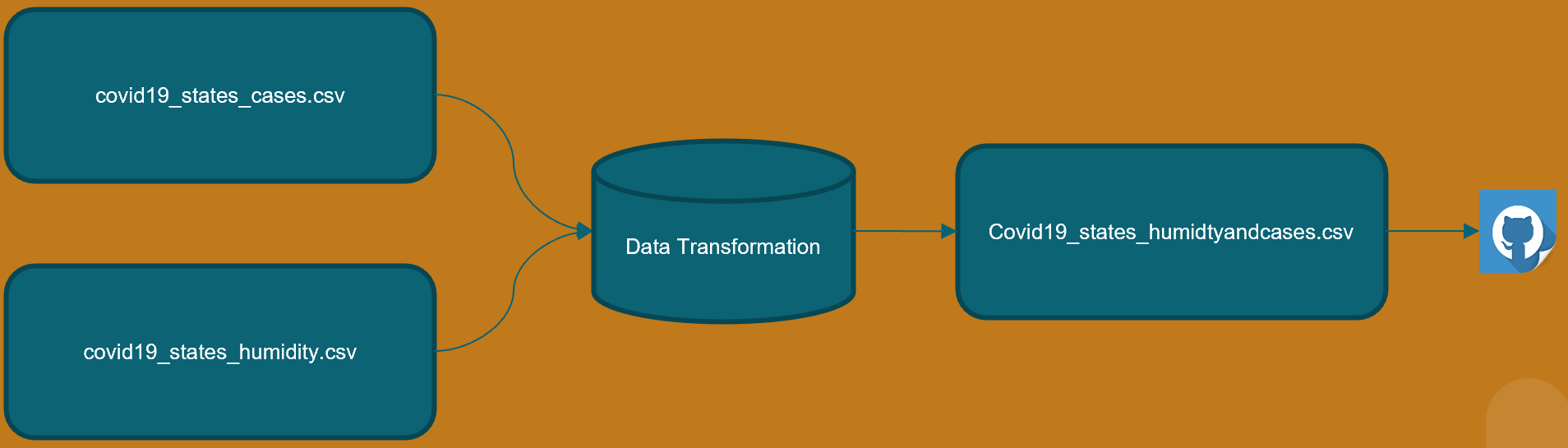
1. Data Source 1: COVID 19 Cases by State in the United States
2. Location: <https://www.worldometers.info/coronavirus/country/us/>
3. COVID 19 Cases by State: Total Cases, Deaths, New Cases, Deaths, and Active Cases from all States in The USA extracted for analysis. Dataset updates daily.
4. Method: This dataset shows the Total COVID 19 Cases by State daily. New cases are updated daily, together with the total deaths, new deaths, and active cases. It contains data from all 50 States showing 4,638,326 total cases with new cases, total deaths, new deaths, and active cases being the nested data. a. Cases report to Health Officials in every State and county. Data is collected and available for the public to view. The data contains deaths and lists no demographic or personal information. Each case is represented as is with no other information is provided. b. The data are contained in the files covid19\_states.csv. More details about the contents and use of the file follows.
5. Obstacles: This is a dataset that is updated daily. As such, due to the daily change, it may not be an appropriate dataset to share as data may have been updated from the time the webpage was scraped. However, from the daily monitoring of COVID cases, this project data analysis would not be affected.
6. Data Source 2: United States Average Humidity State Rank
7. Location: www.usa.com/rank/us--average-humidity--state-rank.htm.
8. The United States, Average Humidity Percentage by State: States and Ranked based on their average humidity percentage level in descending order.
9. Method: This dataset shows the average humidity level across all states. The population of each State is included as well.  
   a. The data are contained in the covid19\_states\_humidity.csv. More details about the contents and use of the file follows.
10. Obstacles: This is a dataset that appears to be updated once a year, gathering humidity levels from states across different periods within a year and finding the average. As such, gathering average levels of humidity could mean the data is measured over time in the same sample group. Hence, we are unsure if the data is from one county or several.

**covid19 states** and **covid19 states humidity** CSV file containing covid19 cases across states and humidity levels are included in the dataset. Humidity levels are consistent with those from the website. COVID cases were obtained and are consistent with data from the first week of July.

**USA States Humidity Levels and COVID19 Cases Merged Data File Structure (covid19\_states\_humidityandcases.csv)** All humidity percentage levels by State, and Covid19 cases, deaths are merged and contained in the file covid19\_states\_humidityandcases.csv. Each line of this file after the header row represents each State, its humidity level, the total number of COVID-19 cases, total deaths, new/active cases, and new deaths. The following format was applied: USAState, TotalCases, NewCases, TotalDeaths, NewDeaths, ActiveCases, AverageHumidity

# **Phase III. Data Preparation: “Beyond the Original Specification:”**

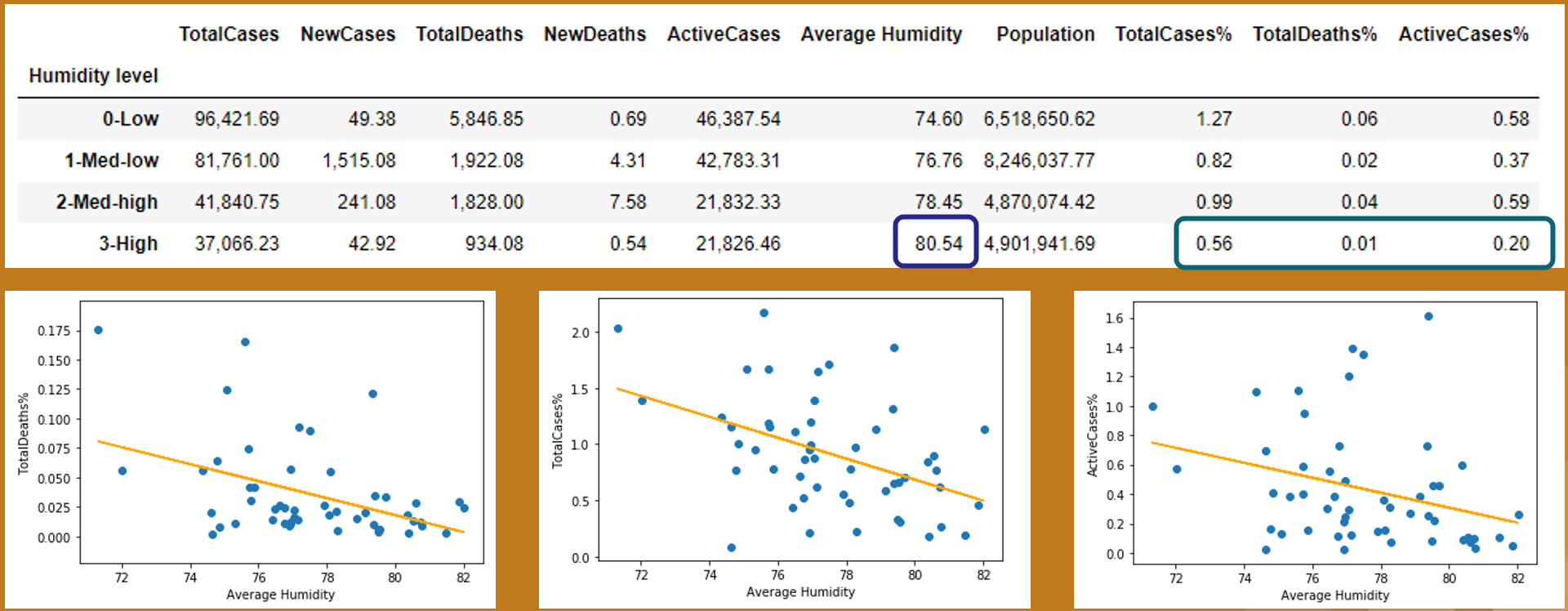
We scraped data from two websites, thus obtaining two separate tables in two separate CSV files. We then joined the files together by using the merge function and pandas to aggregate the data. Once aggregated, we wrote this data into a separate CSV file and then uploaded the file to Github for analysis.



# **Phase IV. Modeling & Analysis: “Results”**

With these additions, we then used panda’s group and aggregate functions to analyze our data. We grouped by humidity levels and looked at both the median and mean values across the various metrics available to us. Of particular interest to us was how the cases and deaths varied among high humidity states vs. medium and low humidity states. A look at the mean grouped data indicated lower cases and death rates within the high humidity grouping. The medians of the grouped data showed similar results.

The next and final step for our analysis was to display our results graphically. We graphed the three most relevant measures – Total Deaths, Total Cases, and Active Cases – all as a percentage of the population. We note that each of them showed results consistent with our hypothesis. States with higher average humidity had lower deaths, total cases, and active cases.



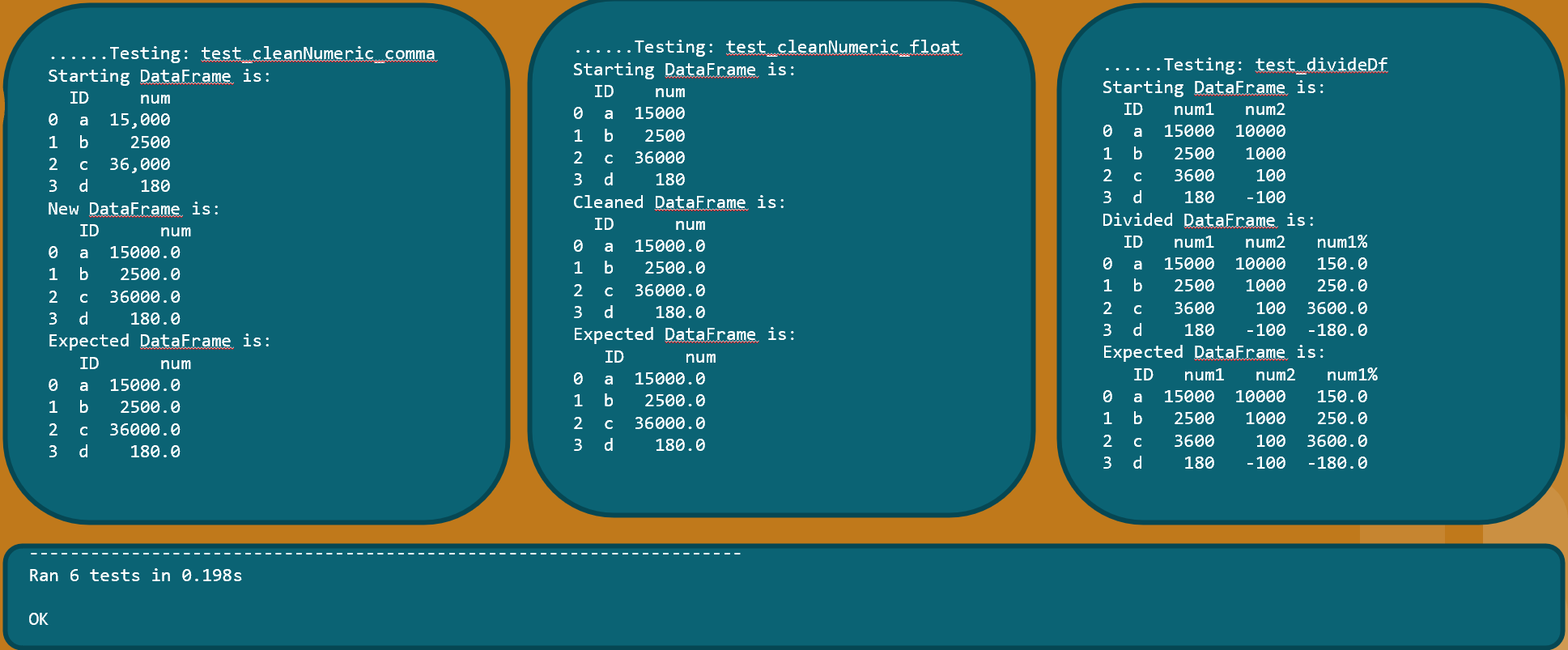
**Phase V. Evaluation: “Testing”**

As we performed our analysis, there were a couple of functions that we wrote to help us. We performed unit testing as we coded each of these.

The first function was the one that cleaned the symbols from the data and transformed them into type float. We wrote unit tests that checked that each symbol was removed correctly, as well as a test to check that the type transformation was successful.

The second function was used to divide 2 columns of our dataframe and add the results as a new column in our dataframe. We wrote unit tests to check that the results were divided correctly and that the new column was named correctly.

Failures of our unit tests helped point to errors in our code as we went, and in the end, all of our unit tests passed.



# **Phase VI. Deployment: “Conclusion”**

In this study to examine the relationship between humidity and the spread of COVID-19 in the fifty continental United States and the Washington District of Columbia, we developed the python code needed to scrape, clean and merge data from two different websites with the information of interest; total COVID-19 cases and deaths, and average overall humidity data for each State. As the web scraping code was developing, we produced python code to test modules of the code to validate its functionality. We saved the data to a .csv file, read into the python programming created, and queried for analysis.

We generarted statistical models in the form of comparison tables and scatterplots to compare the mean, median, and relationship between the number of total COVID-19 cases and deaths for states in different humidity buckets. Results of the data analysis indicated that the humidity band, which ranged from ~74 to ~80 percent, between the states was approximately six percent and that the number of cases and deaths began to drop significantly at humidity levels above 78%.

We recommend that further studies to validate the correlation between humidity and the spread of COVID-19 be conducted. The results of our initial efforts examining the relationship between humidity and COVID-19 highlighted that high humidity negatively impacts the spread. However this for humidity to become a forcing factor on sustaining the virus this in itself is not sustainable as high humidity levels required are not tolerable of humans. Health officials should consider using humidity modeling to identify regions most likely to be at higher risk of infection and estimate the potential of spread.