

## **Creating Dataframe**

The indiana raw covid data is taken from two links, one is for totals cases/deaths/tests and one is for demographics breakdown.

## **Data Cleaning**

Cleans the raw data in order to create

- Probable deaths (Indiana does include probable deaths in their, only on their dashboard so I am taking this number from CRDT and adding it in later when calculate total\_deaths)
- Indiana\_today (shows total cases, total deaths, and total tests with the date)
- Indiana\_race
- Indiana\_ethnicity

## **Indiana Race Dataframe**

This dataframe is to see cases, tests, and deaths by race

## **Indiana Ethnicity Dataframe**

This dataframe is to see cases, deaths and tests by ethnicity

## **Data Validation**

This was to simulate the data validation check in the CRDT data entry

Does four checks (if not zero will highlight red):

- Difference in sum of cases by race v. total confirmed cases
- Difference in sum of deaths by race v. Total confirmed deaths
- Difference in sum of cases by ethnicity v. Total confirmed cases
- Difference in sum of deaths by ethnicity v. Total confirmed deaths

## **Historical Indiana Data from CRDT Website**

Takes historical data from CRDT data because Indiana does not post their historical data. This is used for time series data quality checks. I made sure to only filter so it was Indiana Only.

## **Converting New Scraped Covid Data into Same Format as CRDT Table**

In order to upload the data to CRDT and for easy comparison between old data and newly scraped data, these cells are transforming new data to the same format as the CRDT table. One thing to note is under total deaths we add confirmed deaths and probable deaths together.

## **Appending the New Scraped Data to CRDT Indiana Historical Data**

In order to do the time series checks which compare the data from the last shift with the data from the newly scraped data, we append the old data with new data.

## **Check for Changes in % of Reporting for Race Data**

On the CRDT dashboard, they report the cases % reported and deaths % reported. This is so you can monitor that change. If there is greater than 3% change since the last shift, the cells will become red.

### **Time Series Diffs Check (amount by which each category increased)**

<https://docs.google.com/spreadsheets/d/1ODWitOgt02LjRWkZQWW6llbPfoAyldxIARmmO32s1ds/edit?pli=1#gid=559698418>

The first tab of the Time Series Checks done by CRDT. This tab is to do the same check to see the difference in cases between newly scraped data and last shift. Orange cells means the category has decreased. If cells have decreased by more than 25 then goes red.

### **Time Series %ofSelf Check (percent by which each category has increased)**

In the second tab of the TimeSeries check is %ofSelf check which shows the percent by which each category has increased. The cells will light up if yellow if greater than 5% change per day, orange if greater than 10% change per day, red if greater than 20% change per day. This is calculated by using the datediff function between this new shift and last shift.

### **Time Series %ofTotal (percent by which percent-total that category increased)**

The third tab of the time series check is %ofTotal. This is calculated dividing each race case by the total cases to see what percent of total cases is by each race. This is repeated for race deaths, ethnicity cases, and ethnicity deaths. Then looking at the percentage change of each. The cells will turn yellow if they have changed by more than 2% per day since last shift. Orange if cells have changed by more than 5% per day since last shift. Red cells have changed by 10% per day since last shift.

### **CRDT Dashboard Changes**

<https://covidtracking.com/race/dashboard>

On the CRDT Dashboard there is a breakdown by cases and deaths by race/ethnicity % compared with the population. The population total is taken from 2018 ACS population data. If there is a 33% higher percentage than population percentage for each case, it will highlight red to show racial disparity. Highlight for deaths\_other is light yellow because should not be compared with the population because so small. This is for both cases and deaths.

### **Per Capita Bar Chart: Cases per 100k people**

On the CRDT dashboard, there is a bar chart for cases per 100,000 people. There is one bar to show the last shift and one bar to show the new shift. This is to anticipate changes in the dashboard.

### **Per Capita Bar Chart: Deaths per 100k people**

On the CRDT dashboard, there is a bar chart for deaths per 100,000 people. There is one bar to show the last shift and one bar to show the new shift. This is to anticipate changes in the dashboard.

### **Testing Data Visualization for Percent of Change**

This visualization is a time series graph to compare white cases vs black cases but can be customized to be any race. It is plotted one line color per white and for cases.

The next graph is to compare black cases vs asian cases in a time series graph.

### **Summary Statistics for Percentage Change for Each Race and Ethnicity**

This is a table for summary statistics for all the numbers including cases for each race and ethnicity. Then for deaths for each race and ethnicity. This data can be used for benchmarks for our data quality checks.

### **Create a csv for today's scraper data**

On your computer if you would like to create a new file csv with today's newly scraped data as the first row. If you already have a file on your computer with historical indiana data, you can avoid this cell. I added a line of code which puts a column which says the state is IN, I did not add it earlier because it interferes with functions such as percent change when there are string values.

### **Add today's scraper data to historical data**

If created a csv in the past and would just like to append new data to it. Use this cell. I added a line of code which puts a column which says the state is IN, I did not add it earlier because it interferes with functions such as percent change when there are string values.