In-Depth Visualization of the COVID-19 Cases Analysis in Colorado by County and Open Research Dataset Reference

Domain Description

COVID-19 is an infectious disease that causes respiratory illness. As it is becoming more widespread around the world in the past months in 2020, we need to understand and treat it from a more holistic perspective than ever before, where collected numbers of tests are merged with current research so that we could explore and discover more in common using a good visualization.

Dataset Overview

World COVID-19 Positive Cases (also Tested/Death Cases) in Time Series:

Novel Corona Virus (COVID-19) epidemiological data since 22 January 2020. The data, a CSV file in 1.2 MB, is compiled by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) from various sources including the World Health Organization (WHO), DXY.cn. Pneumonia. 2020, BNO News, National Health Commission of the People's Republic of China (NHC), China CDC (CCDC), Hong Kong Department of Health, Macau Government, Taiwan CDC, US CDC, Government of Canada, Australia Government Department of Health, European Centre for Disease Prevention and Control (ECDC), Ministry of Health Singapore (MOH).

Colorado Smoking Data by county:

The data, a CSV file 8KB in size, is from The State of Colorado and the Colorado Department of Public Health and Environment (CDPHE) and represents the predicted (modeled) prevalence of adults (age 18+) who currently Smoke Cigarettes for each census tract in Colorado. Currently, smoking is defined as having smoked at least 100 cigarettes (5 packs) in your lifetime and now smoke cigarettes on some days or every day. The estimates represent the average that was derived from multiple years of Colorado Behavioral Risk Factor Surveillance System data (2014-2017) that are based on statistical models and are not direct survey estimates. Using the best available data, CDPHE was able to model census tract estimates based on demographic data and background knowledge about the distribution of specific health conditions and risk behaviors. 9 census tracts are displayed in the map as "No Estimate" because of either with a known population of less than 50 (7) or exclusively containing a federal correctional institution as 100% of their population (2).

Colorado Asthma Data by county:

The data, a CSV file 286KB in size, is from The State of Colorado and the Colorado Department of Public Health and Environment and represents the predicted (modeled) prevalence of Asthma among adults (age 18+) for each census tract in Colorado. Asthma is defined as ever being diagnosed with Asthma by a doctor, nurse, or other health professionals, and still having the condition. The estimates represent the average that was derived from multiple years of Colorado Behavioral Risk Factor Surveillance System data (2014-2017). 9 census tracts are displayed in the map as "No Estimate" because of either with a known population of less than 50 (7) or exclusively containing a federal correctional institution as 100% of their population (2).

Colorado COVID-19 Positive Cases and Rates of Infection by County Data:
This dataset contains the number of COVID-19 Positive Cases by County of Identification and County Rate of Infection Per 100,000 Persons.

COVID-19 Open Research Data:
The metadata provided for papers, a 50MB CSV file, is combined from sources of CZI, PMC, BioRxiv/MedRxiv with 29500 records and 14 columns in total, where 1236 records are coming from CZI, 27337 from PMC, 566 from bioRxiv, and 361 from medRxiv. In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 29,000 scholarly articles, including over 13,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.

Coming with the open research data is a collection of JSON files in 2GB that contain detailed contents of each paper included in the metadata, especially text_body, and has shared part of other columns in the open research data, including ‘title’, ‘authors’, and ‘abstract’. The JSON structure stays the same for any type of source paper.

Data Quality and Cleaning

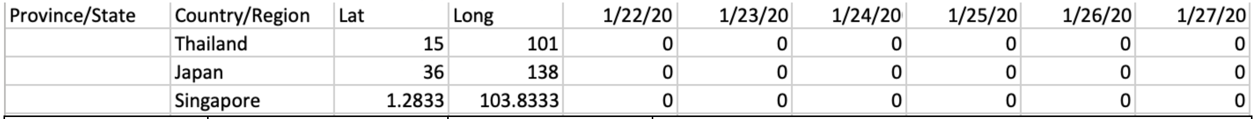
There is no obvious missing data in all 5 datasets (or compiled datasets) after exploratory data analysis, however, there is something special about one dataset, the COVID-19 Open Research Data. In the dataset, the column of ‘sha’ with hash is filled for 17K papers that have PDF records. One thing to notice is that one paper's metadata can be associated with more than one PDFs/shas under the paper for PMC sourced papers. The column ‘has_full_text’ is used to indicate whether PDF papers that were processed with full text or not, and there are 13K out of the total records. The column ‘pmcid’ is populated for PMC sourced papers only, while the column ‘doi’ is populated for those from BioRxiv/MedRxiv and others, and ‘WHO #Covidence’ is populated for all CZI records. 'pubmed_id' and 'Microsoft Academic Paper ID' are the columns for some of the records only.

Deriving Attributes and/or Integrating Multiple Datasets

Positive cases from World COVID-19 Positive Cases (also Tested/Death Cases) in Time Series could derive basic statistics like mean, quartiles, standard deviation, daily increase, and accumulation by country and the generated attributes could compare with each other by using either line charts or box plots, when users are interested in finding trends and fitted models for positive cases, e.g., Task 1. The datasets of Colorado Smoking Data by county, Colorado Asthma Data by county, and Colorado COVID-19 Positive Cases and Rates of Infection by County could be combined through the shared attribute ‘county’ so that smoking, asthma, and COVID-19 positive cases figure are easier to be shown at the same time, if users especially have some preference over checking causality between them for each county in Colorado, e.g., Task 2&3. The COVID-19 Open Research Data could work as a complementary section using NLP techniques to generate new attributes like keywords and topics, which enables the users to explore existing research papers for topics and questions of their interests related to COVID-19, e.g., for Task 4&5.

Data Abstraction

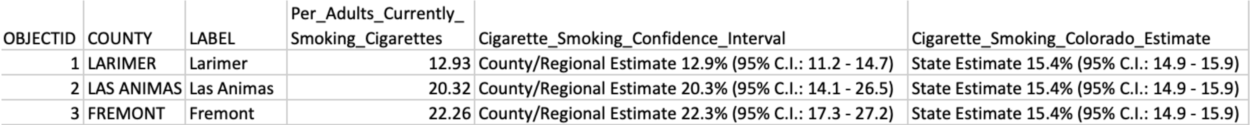
World COVID-19 Positive Cases (also Tested/Death Cases) in Time Series:



Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20
	Thailand	15	101	0	0	0	0	0	0
	Japan	36	138	0	0	0	0	0	0
	Singapore	1.2833	103.8333	0	0	0	0	0	0

	Data/Dataset Type	Attribute Type	Semantic Meaning
Province/State	Position/Geometry	Categorical	PROVINCE
Country/Region	Position/Geometry	Categorical	COUNTRY
Lat	Position/Geometry	Quantitative	LATITUDE
Long	Position/Geometry	Quantitative	LONGITUDE
Date	Attribute/Table	Ordinal	DATE SEQUENCE
Value	Attribute/Table	Quantitative	POSITIVE CASES IN TIME SERIES

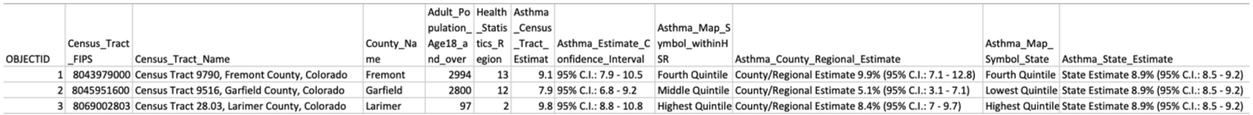
Colorado Smoking Data by county:



OBJECTID	COUNTY	LABEL	Per_Adults_Currently_Smoking_Cigarettes	Cigarette_Smoking_Confidence_Interval	Cigarette_Smoking_Colorado_Estimate
1	LARIMER	Larimer	12.93	County/Regional Estimate 12.9% (95% C.I.: 11.2 - 14.7)	State Estimate 15.4% (95% C.I.: 14.9 - 15.9)
2	LAS ANIMAS	Las Animas	20.32	County/Regional Estimate 20.3% (95% C.I.: 14.1 - 26.5)	State Estimate 15.4% (95% C.I.: 14.9 - 15.9)
3	FREMONT	Fremont	22.26	County/Regional Estimate 22.3% (95% C.I.: 17.3 - 27.2)	State Estimate 15.4% (95% C.I.: 14.9 - 15.9)

	Data/Dataset Type	Attribute Type	Semantic Meaning
OBJECTID	Attribute/Table	Ordinal	OBJECTID
COUNTY	Position/Geometry	Categorical	COUNTY
LABEL	Position/Geometry	Categorical	COUNTY
Per_Adults_Currently_Smoking_Cigarettes	Attribute/Table	Quantitative	CIGARETTES PER ADULT
Cigarette_Smoking_Confidence_Interval	Attribute/Table	Quantitative	COUNTY CIGARETTE ESTIMATE CI
Cigarette_Smoking_Colorado_Estimate	Item/Table		STATE CIGARETTE ESTIMATE

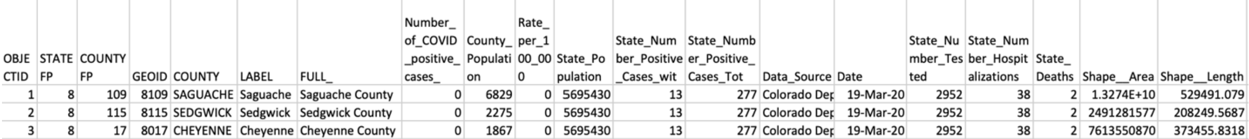
Colorado Asthma Data by county:



OBJECTID	Census_Tract_FIPS	Census_Tract_Name	County_Name	Adult_Population_Age18_and_over	Health_Statistics_Region	Asthma_Census_Tract_Estimat	Asthma_Estimate_Confidence_Interval	Asthma_Map_Symbol_withinHSR	Asthma_County_Regional_Estimate	Asthma_Map_Symbol_State	Asthma_State_Estimate
1	8043979000	Census Tract 9790, Fremont County, Colorado	Fremont	2994	13	9.1	95% C.I.: 7.9 - 10.5	Fourth Quintile	County/Regional Estimate 9.9% (95% C.I.: 7.1 - 12.8)	Fourth Quintile	State Estimate 8.9% (95% C.I.: 8.5 - 9.2)
2	8045951600	Census Tract 9516, Garfield County, Colorado	Garfield	2800	12	7.9	95% C.I.: 6.8 - 9.2	Middle Quintile	County/Regional Estimate 5.1% (95% C.I.: 3.1 - 7.1)	Lowest Quintile	State Estimate 8.9% (95% C.I.: 8.5 - 9.2)
3	8069002803	Census Tract 28.03, Larimer County, Colorado	Larimer	97	2	9.8	95% C.I.: 8.8 - 10.8	Highest Quintile	County/Regional Estimate 8.4% (95% C.I.: 7 - 9.7)	Highest Quintile	State Estimate 8.9% (95% C.I.: 8.5 - 9.2)

	Data/Dataset Type	Attribute Type	Semantic Meaning
OBJECTID	Attribute/Table	Ordinal	OBJECTID
Census_Tract_FIPS	Attribute/Table	Categorical	TRACT FIPS
Census_Tract_Name	Attribute/Table	Categorical	TRACT NAME
County_Name	Position/Geometry	Categorical	COUNTY
Adult_Population_Age18_and_over	Attribute/Table	Quantitative	ADULT POPULATION
Health_Statistics_Region	Attribute/Table	Categorical	HEALTH STAT REGION
Asthma_Census_Tract_Estimate	Attribute/Table	Quantitative	ASTHMA ESTIMATE
Asthma_Estimate_Confidence_Interval	Attribute/Table	Quantitative	ASTEMA ESTIMATE CI
Asthma_Map_Symbol_withinHSR	Attribute/Table	Categorical	ASTHMA MAP SYMBOL
Asthma_County_Regional_Estimate	Attribute/Table	Quantitative	ASTHMA COUNTY ESTIMATE
Asthma_Map_Symbol_State	Attribute/Table	Categorical	ASTHMA STATE SYMBOL
Asthma_State_Estimate	Item/Table		ASTHMA STATE ESTIMATE

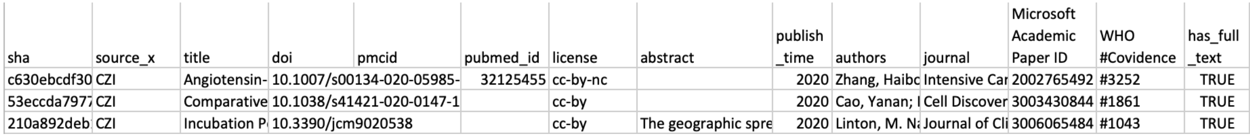
Colorado COVID-19 Positive Cases and Rates of Infection by County:



OBJECTID	STATEFP	COUNTYFP	GEOID	COUNTY	LABEL	FULL_	Number_of_COVID_positive_cases_	County_Population	Rate_per_100_000	State_Population	State_Number_Positive_Cases_wit	State_Number_Positive_Cases_Tot	Data_Source	Date	State_Number_Tested	State_Number_Hospitalizations	State_Deaths	Shape__Area	Shape__Length
1	8	109	8109	SAGUACHE	Saguache	Saguache County	0	6829	0	5695430	13	277	Colorado Dep	19-Mar-20	2952	38	2	1.3274E+10	529491.079
2	8	115	8115	SEDGWICK	Sedgwick	Sedgwick County	0	2275	0	5695430	13	277	Colorado Dep	19-Mar-20	2952	38	2	2491281577	208249.5687
3	8	17	8017	CHEYENNE	Cheyenne	Cheyenne County	0	1867	0	5695430	13	277	Colorado Dep	19-Mar-20	2952	38	2	7613550870	373455.8318

	Data/Dataset Type	Attribute Type	Semantic Meaning
OBJECTID	Attribute/Table	Ordinal	OBJECTID
STATEFP	Attribute/Table	Categorical	STATEFP
COUNTYFP	Attribute /Table	Categorical	COUNTYFP
GEOID	Attribute/Table	Categorical	GEOID
COUNTY	Position/Geometry	Categorical	COUNTY
LABEL	Position/Geometry	Categorical	COUNTY
FULL_	Position/Geometry	Categorical	COUNTY (FULL TEXT)
Number_of_COVID_positive_cases_	Attribute/Table	Quantitative	POSITIVE CASES
County_Population	Attribute/Table	Quantitative	COUNTY POPULATION
Rate_per_100_000	Attribute/Table	Quantitative	POSITIVE CASES RATE
State_Population	Item /Table		STATE POPULATION
State_Number_Positive_Cases_wit	Item /Table		STATE POSITIVE CASES
State_Number_Positive_Cases_Tot	Item /Table		STATE POSITIVE CASES
Data_Source	Attribute/Table	Categorical	DATA SOURCE (CDPHE)
Date	Item /Table		LATEST DATE
State_Number_Tested	Item/Table		STATE TESTED CASES
State_Number_Hospitalizations	Item /Table		STATED HOSPITALIZATION CASES
State_Deaths	Item /Table		STATE DEATHS
Shape__Area	Attribute/Table	Quantitative	COUNTY SHAPE AREA
Shape__Length	Attribute/Table	Quantitative	COUNTY SHAPE LENGTH

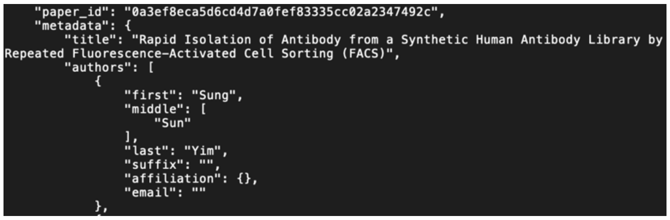
COVID-19 Open Research Data:



sha	source_x	title	doi	pmcid	pubmed_id	license	abstract	publish_time	authors	journal	Microsoft Academic Paper ID	WHO #Covidence	has_full_text
c630ebcdf30	CZI	Angiotensin-	10.1007/s00134-020-05985-		32125455	cc-by-nc		2020	Zhang, Haibc	Intensive Car	2002765492	#3252	TRUE
53eccda7977	CZI	Comparative	10.1038/s41421-020-0147-1			cc-by		2020	Cao, Yanan; I	Cell Discover	3003430844	#1861	TRUE
210a892deb	CZI	Incubation P	10.3390/jcm9020538			cc-by	The geographic spre	2020	Linton, M. Na	Journal of Cli	3006065484	#1043	TRUE

	Data/Dataset Type	Attribute Type	Semantic Meaning
sha	Attribute/Table	Categorical	PDF
source_x	Attribute/Table	Categorical	4 SOURCES
title	Attribute/Table	Categorical	TITLE
doi	Attribute/Table	Categorical	ID
pmcid	Attribute/Table	Categorical	PMC ID
pubmed_id	Attribute/Table	Categorical	PUBMED ID
license	Attribute/Table	Categorical	LICENSE
abstract	Attribute/Table	Categorical	PAPER ABSTRACT
publish_time	Attribute/Table	Ordinal	PAPER PUBLISH TIME
authors	Attribute/Table	Categorical	AUTHORS
journal	Attribute/Table	Categorical	JOURNAL
Microsoft Academic Paper ID	Attribute/Table	Categorical	MICROSOFT ID
WHO #Covidence	Attribute/Table	Categorical	COVIDENCE
has_full_text	Attribute/Table	Categorical (T/F)	FULL TEXT OR NOT

JSON File:



```
    "paper_id": "0a3ef8eca5d6cd4d7a0fef83335cc02a2347492c",
    "metadata": {
        "title": "Rapid Isolation of Antibody from a Synthetic Human Antibody Library by
Repeated Fluorescence-Activated Cell Sorting (FACS)",
        "authors": [
            {
                "first": "Sung",
                "middle": [
                    "Sun"
                ],
                "last": "Yim",
                "suffix": "",
                "affiliation": {},
                "email": ""
            },
“paper_id": <str>,              40-character sha1 of the PDF
"metadata":
  "title": <str>,
  "authors":                    list of author dicts, in order
      "first": <str>,
      "middle": <list of str>,
      "last": <str>,
      "suffix": <str>,
      "affiliation": <dict>,
      "email": <str>,
  "abstract":                   list of paragraphs in the abstract
      "text": <str>,
      "cite_spans":             list of character indices of inline citations
          "start": 151,
          "end": 154,
          "text": "[7]",
          "ref_id": "BIBREF3",
      "ref_spans":              list of dicts similar to cite_spans>,
      "section": "Abstract",
  "body_text":                  list of paragraphs in full body paragraph dicts look the same as above
      "text": <str>,
      "cite_spans": [],
      "ref_spans": [],
      "eq_spans": [],
      "section": "Introduction",
  "bib_entries":
    "BIBREF0":
      "ref_id": <str>,
      "title": <str>,
      "authors": <list of dict>   same structure as earlier but without `affiliation` or `email`
      "year": <int>,
      "venue": <str>,
      "volume": <str>,
      "issn": <str>,
      "pages": <str>,
      "other_ids":
        "DOI": <str>,
  "ref_entries":
    "FIGREF0":
      "text": <str>,              figure caption text
      "type": "figure"
     ...
    "TABREF13":
      "text": <str>,              table caption text
      "type": "table"
  "back_matter": <list of dict>         same structure as body_text
```

Task Descriptions

1. What is the current situation of COVID-19 worldwide and how long does it take to double the positive cases in each country? What are the positive cases trajectories difference?

2. How to define COVID-19 test efficiency so far?

3. Is there a deep relationship between lung-related chronic diseases and positive cases of COVID-19?

4. What do we know about COVID-19 risk factors, according to existing research papers? Are people smoking or having chronic lung diseases are prone to get infected?

5. What are good suggestions for people with lung diseases or smoking to be less likely to get infected than they are supposed to be, if the answer for task 4 is yes, based on current research of COVID-19?

Initial Visualization Design Ideas & Usage Scenario

To manifest the analysis result for Task 1 efficiently, based on side-by-side views from 'Rules of Thumb', we need multiple views that have a main view place holder of a geographic map, one view for a pie chart with basic COVID-19 positive cases composition around the world**,** one view for multi-filtering countries (functioning at the same time with selection on the map), and the rest are juxtaposed views for showing cases trajectories and statistical figures based on users' selection of countries, one of which is a sparklines collection panel similar to the system of LiveRAC we've learned from the textbook. A data table download toolkit will be provided for analysis results, especially for selection over a dozen countries, while the result will be shown in the sparkline view as well. Hues will work for country categories and the point size on the map could be a potential indicator for a certain statistic. Users could get a feeling what is the current COVID-19 situation worldwide from the pie chart at first, then they could use the filtering tool to select multiple countries for which they would like to compare and make predictions from modeling.

It takes a little bit of creativity for Task 2, where we need to have another panel, beside the 1st one and won't show at the same time with it, that aims to utilize 3 datasets with model fittings and diagnostics, e.g., we could create a new index, 'panic index', that is basically dividing positive cases over total tested cases, whose result will be shown in dynamic line charts. For Task 4 and 5, we are supposed to apply NLP techniques such as topic segmentation and recognition to the text_body from the combined dataset of COVID-19 Open Research Data and those JSON files based on paper_id and doc_id, whose result will be shown either through interactive clusters or networks, since the generate attributes could be treated as nodes in a network. There will be a text input placeholder for users to type in keywords or topics of their interests, e.g., "asthma, pneumonia, coronavirus, covid-19" and the result will be provided through a data table instead of plots because users are more likely to use the paper id information from the table to search the entire content online instead of a simple glancing at the result.