UNSW DataSoc x Atlassian: Datathon 2020

Question Pack

Table of Contents

Problem Statement

How have countries' characteristics
(social, cultural, economic, and political values and indicators)
impacted their **response** to COVID-19 and its **effectiveness**?

Please specify your target audience when presenting your analysis.

Guidance

The problem statement is quite broad in scope. You may want to narrow it down and study specific aspects for the datathon.

Below are some potential areas to explore. Those in **bold** are suggested or have data provided.

Case statistics

- **Number of cases** (incremental vs cumulative, reported vs expected)
- **Number of reported deaths**
- **Number of deaths** (reported vs excess deaths)

Country specific actions and characteristics

- Policy actions taken, including but not limited to
  - **Control measures** (domestic restrictions, border restrictions)
- Country characteristics, including but not limited to
  - **Health system capabilities** (patient beds, ICU beds, medical expertise, health funding, public vs private insurance)
  - **Economic characteristics** (GDP, GDP per capita, inequality/Gini index, unemployment, levels of infrastructure/physical connectedness)
  - Government characteristics (level of democracy, level of federal vs state control)
  - Societal characteristics (population, level of trust in government, household size, age distribution, life expectancy, education, crime rates)
  - **Social values** (individualistic vs collectivistic values - Article)

News articles for inspiration

- Do low-trust societies do better in a pandemic? - The Economist
- Comparing nation's COVID containment measures - Nature

Judging criteria

Criteria	Considerations	Marks
Answer Quality	(1) Hypothesis • Defined a clear problem statement and hypothesis that is used to frame the analysis	/ 5
	(2) Research - translating the question into a data problem • Extensive detailed research/usage of country-specific characteristics based on problem statement or hypothesis • Usage of multiple reputable sources to support findings • Summarises findings and hypotheses concisely • Clear referencing to data sources	/ 10
	(3) Analysis • Usage of clear quantitative analysis as evidence to support theories • Efficient and hypothesis-driven usage of datasets (provided and/or external) to support answer • Uses compelling visualisations to support thesis/hypotheses	/ 20
	(4) Insights • Conclusions are clearly communicated and interpreted. • The conclusions and insights address the problem statement and answer the problem/question effectively.	/ 15
Presentation Quality	• Effectively conveys the answer through appropriate visualisations • Demonstrates an understanding of the context and impact of the analysis	/ 20
	• Concise visual presentation • Clear delivery and message throughout the presentation • Engaging and easy to follow	/ 10
Creativity of Investigation	• Degree of creativity and original thinking in identifying novel questions to ask about the data • Innovative method of analysis to generate useful insights	/ 10
Code Quality	• Readability and interpretability, effective documentation • Modelling detail and accuracy • Reproducibility of analysis, transparency and accountability	/ 10
Total		/ 100

Datasets

Introduction

To save you some time, we pulled all the main datasets for easy download from: https://drive.google.com/drive/u/3/folders/1Pm4fvAr7OyXTzHfu_XNgtG15vbHB9gKT

We designated the datasets into three categories:

- **Core datasets** are those that capture aspects of COVID-19 response and effectiveness, and are likely to be relevant whatever your scope is
- The **suggested datasets** capture various country characteristics. You may choose to use some, all or none of them, depending on the scope that you set and the questions you'd like to explore
- We also provide pointers to some **other datasets** that may be relevant

Treat the whole list of datasets (including the core datasets) as a *recommendation*. They are intended to provide direction and save you some time. If there are other datasets that you wish to use, by all means go for it. As long as you target the problem statement and properly cite your sources, you're good.

Documentation

Below is extensive documentation on the datasets which we recommend using. They are extremely useful for finding out which files would be useful to your team, without downloading them.

For each dataset, we have included:

- Description
- Source(s)
- Columns (listing all the features)
- Examples of records in tables

Prepared by Arik Friedman, Principal Data Scientist at Atlassian.

Core datasets

Novel Coronavirus 2019 dataset

Description: For 253 locations, and for the date range 22 Jan 2020 - 23 September 2020, this dataset provides the number of confirmed cases, recovered cases, and the number of deaths.
Source: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
(note that the kaggle dataset pulls information from https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series, so it has the same information, but some of the files are in a format that is easier to process since it merges file-per-day CSVs into a single daily-report file.)

The files ***time_series_covid_19_recovered.csv***, ***time_series_covid_19_deaths.csv***, and ***time_series_covid_19_confirmed.csv*** capture for 253 locations, and for the date range 22 January 2020 - 12 September 2020, the cumulative number of recovered cases, deaths and confirmed cases respectively.
Columns:

- Province/State,
- Country/Region,
- Lat,
- Long,
- 1/22/20,
- 1/23/20,
- 1/24/20,
- …. (dates in the format m/d/yy)

Example record: *New South Wales,Australia,-33.8688,151.2093,0,0,0 , …*

The file ***covid_19_data.csv*** contains similar information to the tables above, but in a different format. It's a merge of the file-per-day records from GitHub.
Columns:

- SNo,
- ObservationDate,
- Province/State,
- Country/Region,
- Last Update,
- Confirmed,
- Deaths,
- Recovered.

Example record: *1,01/22/2020,Anhui,Mainland China,1/22/2020 17:00,1.0,0.0,0.0*

The files ***time_series_covid_19_deaths_US.csv*** and ***time_series_covid_19_confirmed_US.csv*** capture for 3340 US locations, and for date range 22 January 2020 - 12 September 2020, the number of deaths and number of confirmed cases respectively.

Columns:

- UID,
- iso2,
- iso3,
- code3,
- FIPS,
- Admin2,
- Province_state,
- Country_Regsion,
- Lat,
- Long_,
- Combined_Key,
- Population,
- 1/22/20,
- 1/23/20,
- 1/24/20,
- …. (dates in the format m/d/yy)

Example record:
84001001,US,USA,840,1001,Autauga,Alabama,US,32.53952745,-86.64408227,"Autauga, Alabama, US",0,0,0, …

See https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data for explanation of fields. There is also a lookup table for the locations on https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/UID_ISO_FIPS_LookUp_Table.csv.

The original dataset includes additional files that were not pulled into the folder:

- The dataset includes also COVID19_open_line_list.csv and COVID19_line_list_data.csv, which have individual level information.
- There are also some detailed datasets for some countries with additional information. For example, https://www.kaggle.com/sudalairajkumar/covid19-in-usa?select=us_states_covid19_daily.csv also has data about tests, hospitalisations, etc.

Oxford Covid-19 Government Response Tracker

Description: OxCGRT collects publicly available information on 17 indicators of government response taken between January 1st and October 2nd 2020. Responses are captured at country level, plus national level for USA.

- Eight of the policy indicators (C1-C8) record information on *containment and closure* policies, such as school closures and restrictions in movement.
- Four of the indicators (E1-E4) record *economic* policies such as income support to citizens or provision of foreign aid.
- Five indicators (H1-H5) record *health system* policies such as the Covid-19 testing regime or emergency investments into healthcare.
- Finally, there is a miscellaneous indicator (M1) for notes that do not fit elsewhere.
- There are also indices derived from the indicators (via simple average), see details on https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/index_methodology.md about the methodology.
- The dataset also includes the number of confirmed cases and the number of deaths for each location, taken from the ECDC data.

See detailed codebook for the dataset on
https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/codebook.md
Pay attention to Data Quality caveats on https://github.com/OxCGRT/covid-policy-tracker

Source:
https://raw.githubusercontent.com/OxCGRT/covid-policy-tracker/master/data/OxCGRT_latest.csv
File: OxCGRT_latest.csv
Note that there is also a version of the data in
https://github.com/OxCGRT/covid-policy-tracker/tree/master/data/timeseries which is organised as time series data per indicator: one file for each indicator, one row for each country, one column for each date.

Columns:

- CountryName,
- CountryCode,
- RegionName,
- RegionCode,
- Date,
- C1_School closing,
- C1_Flag,
- C2_Workplace closing,
- C2_Flag,
- C3_Cancel public events,

- C3_Flag,
- C4_Restrictions on gatherings,
- C4_Flag,
- C5_Close public transport,
- C5_Flag,
- C6_Stay at home requirements,
- C6_Flag,
- C7_Restrictions on internal movement,
- C7_Flag,
- C8_International travel controls,
- E1_Income support,
- E1_Flag,
- E2_Debt/contract relief,
- E3_Fiscal measures,
- E4_International support,
- H1_Public information campaigns,
- H1_Flag,
- H2_Testing policy,
- H3_Contact tracing,
- H4_Emergency investment in healthcare,
- H5_Investment in vaccines,
- M1_Wildcard,
- ConfirmedCases,
- ConfirmedDeaths,
- StringencyIndex,
- StringencyIndexForDisplay,
- StringencyLegacyIndex,
- StringencyLegacyIndexForDisplay,
- GovernmentResponseIndex,
- GovernmentResponseIndexForDisplay,
- ContainmentHealthIndex,
- ContainmentHealthIndexForDisplay,
- EconomicSupportIndex,
- EconomicSupportIndexForDisplay

Example: United States,USA,Colorado,US_CO,20200910,2.00,0,2.00,1,1.00,1,4.00,1,1.00,0,1.00,1,1.00,1,3.00,2.00,0,2.00,,,2.00,1,2.00,2.00,,,,60155,1979,61.11,61.11,71.43,71.43,68.59,68.59,65.15,65.15,87.50,87.50)

Community mobility reports

Description: The data shows how visits to places, such as grocery stores and parks, are changing in each geographic region, for the period between 15th February 2020 and 27th September 2020. The place categories covered are: Grocery & pharmacy, Parks, Transit stations, Retail & Recreation, Residential, Workplaces.

These datasets show how visits and length of stay at different places change compared to a baseline. The baseline is the median value, for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020. The datasets show trends over several months with the most recent data representing approximately 2-3 days ago.

"Location accuracy and the understanding of categorized places varies from region to region, so we don't recommend using this data to compare changes between countries, or between regions with different characteristics (e.g. rural versus urban areas)."

See more detailed explanation on

https://www.google.com/covid19/mobility/data_documentation.html?hl=en

Source: https://www.google.com/covid19/mobility/

File: Global_Mobility_Report.csv (it's also possible to download separate files per country (Region_CSVs) but the global mobility report contains all of that data)

Columns:

- country_region_code,
- country_region,
- sub_region_1,
- sub_region_2,
- metro_area,
- iso_3166_2_code,
- census_fips_code,
- date,
- retail_and_recreation_percent_change_from_baseline,
- grocery_and_pharmacy_percent_change_from_baseline,
- Parks_percent_change_from_baseline,
- transit_stations_percent_change_from_baseline,
- workplaces_percent_change_from_baseline,
- residential_percent_change_from_baseline

Example record: *AE,United Arab Emirates,Abu Dhabi,,,AE-AZ,,2020-03-22,-27,-9,-30,-44,-14,11*

Suggested datasets

World development indicators - health systems

Description: The dataset describes various health spending per capita by Country, as well as doctors, nurses and midwives, and specialist surgical staff per capita. *"The motivation behind the dataset was to explore whether health spending levels (public or private), or hospital staff have any effect on the rate at which Covid-19 spreads in a country? Can we use this data to predict the rate at which Cases or Fatalities will grow?"*

Source: https://www.kaggle.com/danevans/world-bank-wdi-212-health-systems
File: 2.12_Health_systems.csv
Columns: (see https://www.kaggle.com/danevans/world-bank-wdi-212-health-systems for detailed descriptions)

- Country_Region
- Province_State
- World_Bank_Name
- Health_exp_pct_GDP_2016
- Health_exp_public_pct_2016
- Health_exp_out_of_pocket_pct_2016
- Health_exp_per_capita_USD_2016
- per_capita_exp_PPP_2016
- External_health_exp_pct_2016
- Physicians_per_1000_2009-18
- Nurse_midwife_per_1000_2009-18
- Specialist_surgical_per_1000_2008-18
- Completeness_of_birth_reg_2009-18
- Completeness_of_death_reg_2008-16

Example record: Germany,,Germany,11.1,84.7,12.4,4714.3,5463.3,NaN,4.2,13.2,108,100,100

Hofstede national culture dimensions (2015)

Description: Culture indicators for 111 countries. 65 of them have full information (all indices available), the rest have some missing values. Excerpt from https://data.world/adamhelsinger/geerthofstedeculturaldimension: *Geert Hofstede's cultural dimensions theory proposes a method of analyzing cultures based on a handful of continuums*.

- **Power distance index (PDI)**: The power distance index is defined as "the extent to which the less powerful members of organizations and institutions (like the family) accept and expect that power is distributed unequally."
- **Individualism vs. collectivism (IDV)**: This index explores the "degree to which people in a society are integrated into groups."
- **Uncertainty avoidance index (UAI)**: The uncertainty avoidance index is defined as "a society's tolerance for ambiguity," in which people embrace or avert an event of something unexpected, unknown, or away from the status quo.
- **Masculinity vs. femininity (MAS)**: In this dimension, masculinity is defined as "a preference in society for achievement, heroism, assertiveness and material rewards for success."
- **Long-term orientation vs. short-term orientation (LTO)**: This dimension associates the connection of the past with the current and future actions/challenges.
- **Indulgence vs. restraint (IND)**: This dimension is essentially a measure of happiness; whether or not simple joys are fulfilled.

See also https://hi.hofstede-insights.com/national-culture

Source: https://data.world/adamhelsinger/geerthofstedeculturaldimension

File: 6-dimensions-for-website-2015-08-16.csv (note that this is semi-colon delimited rather than comma delimited format. The file is available also in excel format).

Columns:

- ctr
- country
- pdi
- idv
- mas
- uai
- ltowvs
- ivr

Example record: *COS;Costa Rica;35;15;21;86;#NULL!;#NULL!*

Worldwide governance indicators

Description: The governance indicators are:

- Voice and Accountability
- Political Stability and Absence of Violence/Terrorism
- Government Effectiveness
- Regulatory Quality
- Rule of Law
- Control of Corruption

See https://info.worldbank.org/governance/wgi/Home/Documents for detailed documentation.

Source: https://info.worldbank.org/governance/wgi/
File: wgidataset-1.xls

Columns: The data is provided in 6 sheets, once for each indicator. All sheets have a similar format: one row per country (with Country/Territory and code fields to identify the country), and a set of columns for each year (the range of years is 1996-2018, not every year included):

- Estimate: Estimate of governance (ranges from approximately -2.5 (weak) to 2.5 (strong) governance performance)
- StdErr: Standard error reflects variability around the point estimate of governance.
- NumSrc: Number of data sources on which estimate is based
- Rank: Percentile rank among all countries (ranges from 0 (lowest) to 100 (highest) rank)
- Lower: Lower bound of 90% confidence interval for governance, in percentile rank terms
- Upper: Upper bound of 90% confidence interval for governance, in percentile rank terms

Example record: *Andorra, ADO, 1.32, 0.48, 1.00, 87.10, 72.04, 96.77, …*

Other datasets

ECDC daily case numbers

Description: Day-by-day numbers of cases and deaths in each country. Includes also 14-days cumulative cases per 100,000 based on 2019 population data. Data goes from December 31st 2019 to October 1st 2020 and covers 209 countries/territories. See https://www.ecdc.europa.eu/en/interpretation-covid-19-data.

Source:
https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

Columns: dateRep, day, month, year, cases, deaths, countriesAndTerritories, geoId, countryterritoryCode, popData2019, continentExp, Cumulative_number_for_14_days_of_COVID-19_cases_per_100000

ECDC weekly data

Description: Several data files from the Eurpoean Centre for disease prevention and control.

Source: https://www.ecdc.europa.eu/en/covid-19/situation-updates-covid-19/covid-19-data/weekly (follow the links for the specific datasets on that page for detailed data dictionaries).

- **Data on hospital and ICU admission rates**. For 29 countries, and for dates between February and September 2020, there is information for the four indicators *Daily hospital occupance*, *Daily ICU occupancy*, *Weekly new hospital admissions per 100k* and *Weekly new ICU admissions per 100k*. File: hops_icu_all_Data_2020-10-01.xlsx.
- **Data on testing for COVID-19 by week and country**. For 30 countries, and for weeks 1-39 in 2020, there is data about: new cases, tests down, population, testing rate, and positivity rate. File: weekly_testing_data_EUEEAUK_2020-10-01_0.xlsx.
- **Daily subnational 14-day notification rate of new COVID-19 cases**. For 558 regions in 14 European countries, this file reports the 14-day notification rate per 100k on a daily basis (date range 26 February 2020 to 27 September 2020). File: subnational_daily_data_2020-10-01.xlsx.
- **Daily national 14-day notification rate of new COVID-19 cases and deaths**. For 208 countries (+ continental totals) and for each day Between January 2nd and September 27th 2020, provides the 14-day notification rate of newly reported COVID-19 cases per 100k population and the14-day notification rate of reported deaths per million population. File: daily_national_incidence_2020-10-01.xlsx.

- **Weekly subnational 14-day notification rate of new COVID-19 cases**. For 419 regions in 31 European countries, and for weeks 13-39 in 2020, provides the 14-day notification rate of new COVID-19 cases per 100k population. File: subnational_weekly_data_2020-10-01.xlsx.
- **Country response measures to COVID-19**. For 31 European countries, reports response measures taken by the countries (total 25 responses, such as daycare closures, mandatory masks, partial or full bans on mass gatherings), with start and end date for each response. File: response_graphs_2020-19-16.csv.

World development indicators

Description: The main data file (WDIdata.csv) has annual measurements of 1431 indicators over 264 countries.
Source: http://datatopics.worldbank.org/world-development-indicators/
See the list of indicators on https://data.worldbank.org/indicator/
User guide: http://datatopics.worldbank.org/world-development-indicators/user-guide.html

The economist tracker for COVID-19 excess deaths

Description: Number of excessive deaths, from the Economist analysis of COVID-19 deaths across countries.
Source: https://github.com/TheEconomist/covid-19-excess-deaths-tracker
File: all_weekly_excess_deaths.csv (per country raw data and other files available via the Github link).

Columns: country, region, region_code, start_date, end_date, year, week, population, total_deaths, covid_deaths, expected_deaths, excess_deaths, non_covid_deaths, covid_deaths_per_100k, excess_deaths_per_100k, excess_deaths_pct_change

Additional resources

- COVID-19 related datasets from the world bank
- A survey paper summarising COVID-19 open source datasets
- Google open data for EU and US