

Group Project Data

0 Points Possible

 Add Comment

▼ Details

Important Dates

- Release of dataset: March 10
- Proposal due date: March 31
- Final report due date: April 30

Note: Group members and data cannot be changed after the proposal due date.

Note: Each group can have 1 - 3 members.

Overview

This is an open-ended exploration of the COVID-19 dataset. The dataset and some example ideas to get started are provided, but the analysis and discovery to be undertaken is up to each group. You may pick and choose which features/attributes to use in conducting your analyses. You may also derive more features/attributes from preexisting features/attributes.

Learning Objectives

1. **Data Literacy:** Understand and interpret complex datasets related to public health.
2. **Analytical Skills:** Develop the ability to conduct thorough data analysis, including statistical and trend analysis.
3. **Critical Thinking:** Engage in hypothesis generation and testing, encouraging a deeper understanding of the data and its implications.
4. **Technical Proficiency:** Gain hands-on experience with data manipulation, visualization, and modeling techniques.
5. **Communication Skills:** Enhance the ability to present findings clearly and effectively.



COVID-19 Dataset - United States - Our World in Data

This COVID-19 dataset is provided by [Our World in Data \(OWID\)](https://ourworldindata.org/coronavirus)  [\(https://ourworldindata.org/coronavirus\)](https://ourworldindata.org/coronavirus) and includes both original and derived features. The dataset has

been curated by Bryan Tuck for in-depth analysis of the pandemic's impact within the United States.

Dataset Overview:

There are two datasets:

- An original COVID-19 dataset including 429K records and raw features for many locations, the features are incomplete for some records and require preprocessing [[Full Dataset \(https://canvas.uh.edu/courses/18140/files/6288897?wrap=1\)](https://canvas.uh.edu/courses/18140/files/6288897?wrap=1)] 
(https://canvas.uh.edu/courses/18140/files/6288897/download?download_frd=1)]
- A subset of the dataset that reflects the situation in the United States, which has been preprocessed to some extent [[Data Subset \(https://canvas.uh.edu/courses/18140/files/6288899?wrap=1\)](https://canvas.uh.edu/courses/18140/files/6288899?wrap=1)]  (https://canvas.uh.edu/courses/18140/files/6288899/download?download_frd=1)]

These datasets provide detailed information on cases, deaths, hospitalizations, and other metrics relevant to the pandemic. Our World in Data has compiled this information, drawing from official sources and making it accessible for public analysis. The dataset is part of a larger collection that tracks the global impact of the COVID-19 pandemic, with daily updates throughout its duration.

Note: The original dataset can be used for creating new tasks than the example tasks below, which is highly preferred.

The Meaning of Some Original Variables:

- date: The date when the data was recorded
- total_cases: Total confirmed cases of COVID-19
- new_cases: New confirmed cases of COVID-19 on the given date
- total_deaths: Total deaths attributed to COVID-19
- new_deaths: New deaths attributed to COVID-19 on the given date
- total_cases_per_million: Total confirmed cases of COVID-19 per 1,000,000 people
- total_deaths_per_million: Total deaths attributed to COVID-19 per 1,000,000 people
- icu_patients: Number of COVID-19 patients in intensive care units (ICUs) on the given date
- hosp_patients: Number of COVID-19 patients in the hospital on the given date
- weekly_hosp_admissions: Number of COVID-19 patients newly admitted to hospitals in the given week

The details about other variables can be found at this [Github Repository](https://github.com/owid/covid-19-data/tree/master/public/data) 
(<https://github.com/owid/covid-19-data/tree/master/public/data>).

The Meaning of The Derived Features (in data subset):

- `daily_case_change_rate`: The percentage change in new cases compared to the total cases on the previous day
- `daily_death_change_rate`: The percentage change in new deaths compared to the total deaths on the previous day
- `hospitalization_rate`: The percentage of total COVID-19 cases that resulted in hospitalization on the given date
- `icu_rate`: The percentage of total COVID-19 cases that required intensive care on the given date
- `case_fatality_rate`: The percentage of total COVID-19 cases that resulted in death
- `7day_avg_new_cases`: The 7-day rolling average of new COVID-19 cases
- `7day_avg_new_deaths`: The 7-day rolling average of new COVID-19 deaths
- `hospitalization_need`: Categorical assessment of hospitalization rates as 'Low', 'Medium', or 'High' based on quantile distribution
- `icu_requirement`: Categorical assessment of ICU rates as 'Low', 'Medium', or 'High' based on quantile distribution

The derived categorical assessments ('hospitalization_need' and 'icu_requirement') are relative to the dataset and are based on the distribution of the data within the United States. These labels are intended to facilitate the analysis and may not represent absolute thresholds for public health action.

The subset presented here includes only the most pertinent variables for a focused analysis on the United States, allowing for a clear understanding of the trends and patterns specific to the U.S. during the COVID-19 pandemic.

Data Preprocessing:

- The "date" column has been converted to a datetime object to facilitate time series analysis.
- Quantile-based discretization function (`pd.cut`) has been used to convert continuous variables into categorical variables for "hospitalization_need" and "icu_requirement".
- Rolling window functions have been used to calculate 7-day averages for new cases and deaths.

Example Tasks:

These tasks are just examples. You are not limited to these areas and are encouraged to explore with creativity and critical thinking.

1. Exploratory Data Analysis (EDA)

- **Objective:** Conduct a thorough exploratory analysis to uncover underlying patterns and insights.
- **Task:** Visualize trends in `total_cases`, `total_deaths`, and `hospitalization_rate`. Perform clustering and analysis to investigate any correlations or surprising patterns in the data and hypothesize their causes.

2. Classification

- **Objective:** Classify days or periods into different risk categories based on COVID-19 data.
- **Task:** Use classification models to categorize days into varying levels of hospitalization_need. Explore which features are most predictive and discuss the implications of your findings.

3. Regression

- **Objective:** Utilize regression models to predict future healthcare requirements based on current trends.
- **Task:** Predict the number of ICU admissions using variables like total_cases and total_deaths. Analyze how well your model performs and discuss its potential real-world applicability.

4. Outlier Detection

- **Objective:** Identify and analyze outliers to uncover unusual patterns or data errors.
- **Task:** Use statistical methods to detect outliers in key variables such as new_cases, new_deaths, icu_patients, and hosp_patients. Techniques like the Interquartile Range (IQR), Z-scores, or visual methods (box plots) can be employed.

Note: Picking the above example task and using the data subset are fine, but the novelty will be limited. Your score will be based on the **Novelty, Comprehensiveness and Depth** of your analysis. For example:



- For classification, comparing multiple models with different parameter settings to justify the choice of the best model is better than randomly picking a model to generate results (the same for other tasks)
- Conducting multiple tasks (e.g., preprocessing, exploratory analysis, classification and clustering) is better than picking only the easiest task
- Choosing challenging tasks (and justify the challenges) is better than choosing easy tasks (without good justification of the challenges)
- Creating new thoughtful tasks and justify them with in-depth analysis is always preferred, **especially on the Full Dataset.**
 - The raw full dataset can be used for more analytic tasks, such as pattern analysis/comparison in different locations/regions
 - Using raw data may need more data preprocessing and new feature/label derivation, which are also considered as new tasks
 - It is welcome to use techniques beyond the lecture contents, such as time series analysis (e.g., forecasting important features), using LSTM/CNN/Transformer models on the dynamic full dataset and analyze the correlation/patterns of the time series between different locations

Also, the **Correctness** of the code and the **Clearness** of the proposal/report are important.

Submission Instructions:

- Please submit the proposal (PDF, 2-3 pages) to Canvas (Please do not submit to this post. Instead, use this [link \(https://canvas.uh.edu/courses/18140/assignments/538098\)](https://canvas.uh.edu/courses/18140/assignments/538098)).
- Please submit all of your python notebooks and a report (PDF, 5-10 pages) in a zip file to Canvas (Please do not submit to this post. Instead, use this [link \(https://canvas.uh.edu/courses/18140/assignments/538099\)](https://canvas.uh.edu/courses/18140/assignments/538099)).
- Please include the generated figures in the report.
- Please include a good README file to describe what's included in the zip file and how to run your code to get your figures/results as presented in your report. Markdown file for README is recommended.
- Please ensure that all required plots/results can be generated by running your code.
- Please make sure that names and UH IDs of all team members are included in the proposal, README, and report.

Note: The code will be run to reproduce the figures and results in the report.

Helpful Links: [VS Code](https://code.visualstudio.com/) , [Markdown](https://www.markdownguide.org/) 
(<https://www.markdownguide.org/>), [Jupyter Notebook](https://jupyter.org/) , (<https://jupyter.org/>).