CIC STUDENT WORKING GROUP

Analyzing Pandemic Responses Project

- Grow your data analysis skills!
- Build a project & get a certificate for your portfolio
- Join monthly peer check-ins

bit.ly/CICSWG

COVID INFORMATION COMMONS (CIC) STUDENT WORKING GROUP

Spring 2024 Analyzing Pandemic Responses Project

The COVID Information Commons (CIC) Student Working Group has developed a new COVID-19 data science project for our members! Below is a description of the research project we have developed for the Spring 2024 term.

Over the next nine weeks (February 4th - April 5th, 2024), Working Group members may participate in the below data science project, learning how to analyze government policy responses to the COVID-19 pandemic. We will use the extensive Oxford COVID-19 Government Response Tracker (OxCGRT) dataset (details below) to perform our analysis. Together, we will learn new skills, practice advanced quantitative and qualitative methods, and uncover evidence-based insights that illuminate effective strategies for future pandemics. The dataset's complexity offers immense potential to identify and understand the most impactful policies in managing global health crises.

This is a great opportunity for STEM learners and researchers of all academic backgrounds to learn new skills in a friendly and supportive environment! The project was designed for beginner to intermediate data science learners. We expect that participants may spend between 3 to 6 hours per week on the project, depending on their interest and level of experience.

To begin, please review the project description below. Next, begin working through Milestones 1-5. Email CICStudentWorkingGroup@columbia.edu with any questions.

Participation, Deadlines, and Certificates:

To participate, students should register themselves or their team (max 3 members) via this Google Form. There is no deadline to register and students may begin the project at any point in time. Registrants will receive updates from the Working Group about project opportunities and deadlines.

Final Submission: Each participant or team will create a data science project on pandemic policy responses, learning data preprocessing, model selection, and analysis along the way. Final submissions are due to CICStudentWorkingGroup@columbia.edu by 5PM ET on Friday, April 5, 2024.

A final submission consists of a data visualization and a 1-2 page written analysis of research insights (if

citations, use APA format). This submission should showcase your adeptness in data analysis and your ability to effectively share your insights with others. Selected projects may be featured on the website.

Scientists often work together to bring about new ideas and research projects. Please remember to cite any support you receive from other students, scientists, or faculty members in your final submission.

Please send your final submissions in one of the following formats:

- GitHub Repository (Include visualizations and code in the repository)
- Colab Notebook
- Tableau Public Links

If these formats are not suitable for your project, you are welcome to propose alternative formats that align with your chosen analysis tools. Email CICStudentWorkingGroup@columbia.edu with any questions. Ensure that codes and visualizations are clearly presented. For formats emphasizing code, consider submitting a PowerPoint presentation alongside your work for easy analysis.

Certificates and Recognition: Participants will receive a certificate of completion for all work submitted to CICStudentWorkingGroup@columbia.edu by 5pm ET on April 5, 2024. If a group of students works on a team project, each group member will receive a separate certificate.

Participants may receive an additional (optional) certificate for presenting their research findings to the group during the CIC Student Working Group meeting held via Zoom at 11AM ET on Friday, April 5, 2024. Information about this and other upcoming meetings can be found on the Working Group webpage. Students who present their research will receive an additional certificate attesting to their ability to communicate complex scientific concepts to a broad audience.

About the Analyzing Pandemic Responses Project - Problem Statement

We will harness the OxCGRT COVID Policy Dataset (Oxford COVID-19 Government Response Tracker, Blavatnik School of Government, University of Oxford) to pinpoint key policies that significantly shaped the trajectory of the COVID-19 pandemic worldwide from 2020 to 2022. We seek meaningful insights to address critical, unanswered questions regarding policy effectiveness. The challenge lies in comprehending the intricate interplay between diverse government policies, the virus' spread, and COVID's multivariate effects on society, the global economy, and public health.

Dataset Description:

The Oxford COVID-19 Government Response Tracker (OxCGRT) collects and organizes information about countries' policy responses to the COVID-19 pandemic over time. Learn more about OxCGRT on Oxford University's website.

Key details about the OxCGRT Dataset:

1. **Government Response Indices:** OxCGRT reflects how strict government policies were in closing schools, restricting travel, banning large gatherings, etc
2. **Time Series Data:** Tracks how policies changed over the course of the pandemic (between 2020 and 2022)
3. **Global Coverage / Geospatial Data:** Tracks policy effectiveness across different geographic regions and countries
4. **Data Sources:** Sources data from government announcements, official reports, and news articles
5. **Policy Stringency Index:** OxCGRT's Stringency Index summarizes the strictness of policies

Challenges with this Data:

Working with datasets like OxCGRT can be challenging due to their complexity. While these datasets offer valuable information, navigating them might seem overwhelming. To support your learning, you are free to use other datasets for this project. If this one seems too tough, look for different datasets linked to pandemic rules. Alternatively, if you have a broader research question, you may consider supplementing OxCGRT with an additional dataset from a reputable source (U Washington has some good tips on finding datasets; Data to Policy Navigator can help you identify trustworthy datasets; explore the datasets available via the COVID Info Commons). The goal of the project is a sound analysis, so choose a dataset that aligns with your skills and interests. Embrace challenges and explore until you find the right fit!

Project Goals:

With this dataset, participants will develop a research question related to pandemic policy. Example research questions might include:

- Were Containment or Vaccination policies more effective in country X?
- Was Economic Policy Y more effective in 2020 or 2021 and why?
- How did a change in Health Policy Z impact Vaccination Policy A?

Note that your research questions may change or become more specific as you complete your analysis.

You will refine your research question in the five Milestones outlined below. Each Milestone is approximately 2 weeks in length, assuming participants spend ~5 hours per week on the project. You may, of course, work ahead or begin the project at any time.

The early Milestones support data exploration, cleaning, and preprocessing. Midway through the project, we begin to apply data models and analysis tools to the data. Finally, you will complete your analysis and prepare your findings for group sharing.

Note: Milestones 2 and 3 contain an Optional Challenge for students who are interested in exploring the OxCGRT dataset with machine learning techniques. It is not required to complete these Challenges to finish your project. These challenges are recommended for participants who have some pre-existing knowledge of Machine Learning.

Collaboration and Mentored Learning

Remember that this project is not just about the end result; it's also about the learning experience and collaboration within the Working Group! We encourage you to share ideas, seek feedback, and help one another throughout the project. We will encourage collaboration in our monthly Working Group meetings, during Office Hours (more info about these meetings below), and over Slack (join here).

Participants with advanced data science experience may act as peer mentors for the Analyzing Pandemic Responses Project. Please use this form to register your interest in either receiving or providing mentorship support to project participants. There is no deadline to register. Working Group leaders will work to match all mentors and mentees with a partner as quickly as possible. Mentors will receive a certificate of appreciation for their collaborative support.

Note that all project participants are subject to the NEBDHub Code of Conduct. Please review the Code of Conduct for important information about community expectations in Zoom webinars, on the CIC Slack channel, digital communications, and more.

Tools You May Need (optional, use your preferred tools or email the project team with questions):

- Microsoft Excel/Google Sheets
- A free Kaggle Account
- A Tableau Public Account
- A Google Colab Notebook or Jupyter Notebook

Important Links:

- Project Participant Registration (there is no deadline to register)
- Mentor / Mentee Registration Form (there is no deadline to register)
- Slack Channel for Discussion (#studentwg-analyzing-pandemic-responses) for weekly tips and reminders as the APR Project progresses.
- CICStudentWorkingGroup@columbia.edu (email us with any questions, we're here to help!)
- Join the Working Group Listserv to receive updates and be added to upcoming Office Hours and Monthly Meetings Google calendar invites

Mark Your Calendars!

Note that all meetings are optional to join, the only requirement to receive a certificate of completion is a final submission by Friday, April 5th at 5PM (ET):
Zoom link for all calls

Friday, February 2nd, 2024 at 11AM (ET) - Monthly WG Meeting and Project Kickoff

- The Analyzing Pandemic Responses Project will be launched at the February WG Meeting!

Wednesday, February 28th, 2024 at 12PM (ET) - Office Hours Session #1 / Mentoring Session

- During this open meeting, we'll discuss your projects, make recommendations, and troubleshoot problems.
- We will also hear from COVID researcher Courtney Baird about their personal experiences using these data science techniques.

Friday, March 1st, 2024 at 11AM (ET) - Monthly WG Meeting

- We will continue to collaborate on the APR Project during the March WG Meeting.

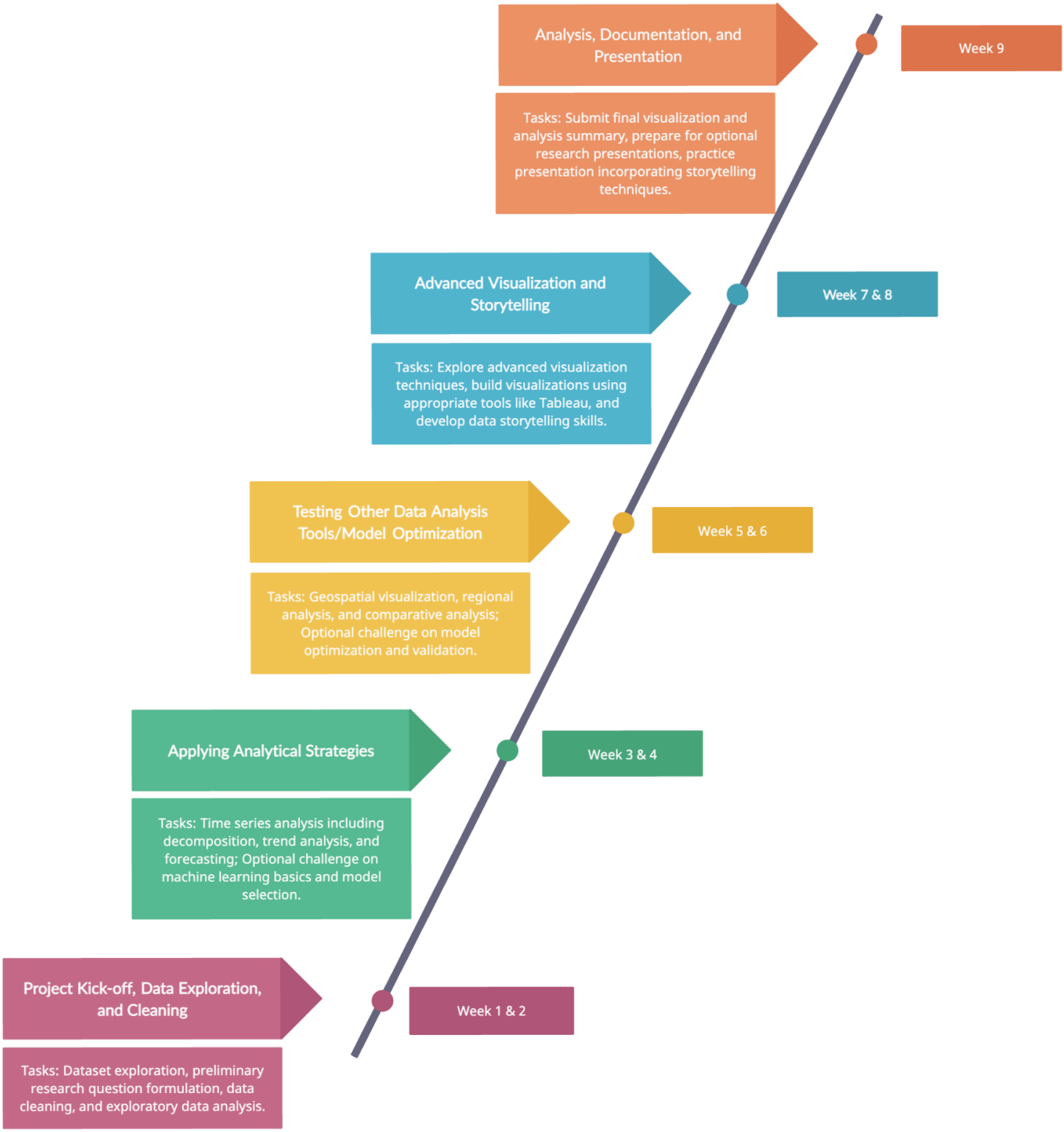
Wednesday, March 27th, 2024 at 12PM (ET) - Office Hours Session #2 / Mentoring Session

- During this open meeting, we'll discuss your projects, make recommendations, and troubleshoot problems.
- We will also hear from COVID researchers Niu Gao and Kathryn Corvey about their personal experiences using these data science techniques.

Friday, April 5th, 2024 at 11AM (ET) - Monthly WG Meeting

- Final project submissions are due to CICStudentWorkingGroup@columbia.edu by 5PM (ET)

- We will wrap up the Analyzing Pandemic Responses Project during the April WG Meeting. Working Group members may present their project to the group for an additional certificate.
- Mentors will receive a certificate of appreciation in the days following the end of the Project.


Project Kick-off, Data Exploration, and Cleaning
Tasks: Dataset exploration, preliminary research question formulation, data cleaning, and exploratory data analysis.
Week 1 & 2
Applying Analytical Strategies
Tasks: Time series analysis including decomposition, trend analysis, and forecasting; Optional challenge on machine learning basics and model selection.
Week 3 & 4
Testing Other Data Analysis Tools/Model Optimization
Tasks: Geospatial visualization, regional analysis, and comparative analysis; Optional challenge on model optimization and validation.
Week 5 & 6
Advanced Visualization and Storytelling
Tasks: Explore advanced visualization techniques, build visualizations using appropriate tools like Tableau, and develop data storytelling skills.
Week 7 & 8
Analysis, Documentation, and Presentation
Tasks: Submit final visualization and analysis summary, prepare for optional research presentations, practice presentation incorporating storytelling techniques.
Week 9

Milestones and Recommended Timeline

Milestone 1: Project Kick-off, Data Exploration, and Cleaning

Weeks of February 4th and February 11th, 2024

Objective: Familiarize yourself with the OxCGRT dataset and define the possible scope of your analysis. Clean the dataset and gain insights from Exploratory Data Analysis (EDA).

Week 1: Introduction, Dataset Overview & Preliminary Research Question

> Introduction & Dataset Overview

1. Project kick-off meeting on Friday, February 2nd.
    a. Project leaders will provide an overview of the project's objectives, expectations, and timeline. We will be available to answer any of your questions and help you get started.
2. Dataset Overview
    a. Go through the above documentation. Read project guidelines and goals.
    b. Get started with the OxCGRT dataset! Give yourself time to explore the dataset documentation.
    c. Refer to the data documentation to understand the data definitions and dataset files.
    d. Identify key variables and indices within the dataset.
    e. Explore the dataset's structure, variables, and challenges. Write down some initial impressions, questions, and thoughts as you go.

> Preliminary Research Question

1. Formulate a broad research question related to pandemic policies.
    a. Review OxCGRT documentation and codebooks. You might like to explore the analysis conducted by Our World in Data for inspiration.
    b. As you explore, consider the research question you want to answer! What interesting ideas emerge as you crawl through these spreadsheets?
2. Consider your research scope.
    a. Start to narrow down your research question. A research question like: 'Was Policy A or Policy B more effective?' is probably too big for you to tackle! Instead, try something like 'Was Policy A or Policy B more effective in Country Y between June 2020 and June 2021?'.
    b. Also, decide who you are going to tell your data story to. Will your final audience be a group of scientists and peers? Is it the general public? How would you frame your

analysis differently for each of these groups? Write down your initial ideas. You will explain your rationale for analysis in Milestone 5.

c. Need some inspiration? Watch this presentation on how one set of NIH-funded researchers at Brown University approached a pandemic policy analysis.

Note: We invite all participants to talk to each other about your ideas and hypotheses throughout the project - this can help spark creativity! Through collaboration, you may come up with innovations neither of you would have thought of alone. Consider signing up to be a Mentor / Mentee (Registration Form) or use the CIC Slack Channel (#studentwg-analyzing-pandemic-responses-spring-2024) to brainstorm with others.

Week 2: Data Exploration and Cleaning

> Initial Data Exploration

1. Begin by exploring the data and perform some basic data cleaning and preprocessing.
    a. Identify missing values and outliers.
2. Learn the basics of data exploration and use the following tools as they best apply to your project
    a. Learn best practices for deduplicating your data, fixing structures, and removing irrelevant data.
    b. Use Pandas Profiling for quick and easy data exploration.
    c. Examine data distributions and summary statistics (try Jupyter Notebooks/Pandas).

> Data Cleaning and Preprocessing

1. Data Cleaning and Integration.
    a. Now that you know how to approach your dataset, you can begin the cleaning process. This will ensure that extraneous outliers and missing values don't skew your analysis.
        i. Watch an introduction to data cleaning video
    b. Learn how bias can impact the data cleaning process and explore basic data science ethics concepts (more ethics material here).
2. Data Transformation:
    a. Learn about data transformation and how it differs from data cleaning.
        i. Convert date columns to a standard date format.
        ii. Aggregate data wherever needed (e.g., daily, weekly or monthly).
    b. Begin the data transformation process, recognizing what variables you will need to answer your initial research question. Consider removing irrelevant data (e.g., if you

want to look at pandemic policies in Germany, you can remove data related to Indonesian policies).

3. Exploratory Data Analysis (EDA)
    a. Create basic visualizations (line plots, bar charts) to understand the data's distribution. This will help you visualize key concepts and refine your research question.
    b. Generate summary statistics for key variables (school policies, lockdown, etc.). You may wish to learn about pivot tables and create a few in your excel workbook.
    c. Brush up on some statistics concepts with this Flashcard video series on Descriptive Statistics from the NSDC.

Milestone 2: Applying Analytical Strategies

Weeks of February 18th and February 25th, 2024

Objective: Explore Time Series Analysis; Choose appropriate analytical methods, start building models, and define the approach.

Weeks 3 & 4: Time Series Analysis (Required) and/or Machine Learning (Optional Challenge)

> Time Series Analysis (Required)

1. Time Series Decomposition:
    a. Let's explore this dataset using statistics. We will deconstruct the data into components that represent categories and patterns. We'll use Time Series Analysis to review trends and seasonal changes. We will consider forecasting techniques, then wrap up by writing down our initial findings from this exercise.
    b. Learn how to decompose the time series into trend, seasonality, and residual components and why it's important for data visualization.
2. Trend Analysis:
    a. Let's start with trend analysis! Learn what trends are in time series analysis.
    b. Plot the trend component in Excel (or a Google Colab notebook if you're comfortable in Python) to identify long-term patterns or trends.
3. Seasonal Analysis:
    a. Learn what seasonality is in a time series and why it matters to your COVID analysis.
    b. Visualize seasonal effects using appropriate plots in Excel or a Google Colab notebook. Identify any recurring patterns (e.g. weekly or monthly effects). (An advanced tutorial for seasonal analysis visualization can be found here.)

4. Time Series Forecasting (not required):
    a. Learn about forecasting techniques like ARIMA or exponential smoothing.

> Machine Learning (Optional Challenge)

Note that this is an optional challenge, recommended for participants who have some previous experience with Machine Learning. Expect that this Optional Challenge will require an additional five hours of project work per week.

1. Understand Machine Learning
    a. Machine learning uses algorithms to learn from data. These algorithms find patterns, develop understanding, make decisions, and evaluate those decisions. In machine learning, datasets are split into two subsets: The first subset is known as the *training data* - it's a portion of our actual dataset that is fed into the machine learning model to discover and learn patterns. In this way, it trains our model. We will explore models in this Milestone.
    b. The other subset is known as the *testing data*. Once your machine learning model is built (with your training data), you need unseen data to test your model. This data is called testing data, and you can use it to evaluate the performance and progress of your algorithms' training and adjust or optimize it for improved results. Testing data has two main criteria. It should represent the actual dataset and be large enough to generate meaningful predictions. We will optimize our models in the next Milestone.
    c. Explore the differences between training data and test data with this tutorial from Google.
2. Machine Learning Basics
    a. We have our dataset, but before we jump into model selection, we need to explore the basics of Machine Learning. This will help us understand how ML can help us explore this particular COVID policy dataset. There is a lot to know about machine learning, so select the resources that support your learning best. Share resources you think your peers would benefit from in the Slack Channel.
    b. Learn more about how and when not to use Machine Learning with this Masterclass video from Matthew Carbone and the NSDC.
    c. Learn the basics of Machine Learning.
    d. Watch lectures on Machine Learning from the OpenDS4All GitHub
    e. Explore the essentials with the NSDC's Machine Learning Videos
    f. Refer to this checklist of best practices while implementing machine learning models.
3. Machine Learning & COVID Research

a. If you need some additional inspiration, here are a few examples of how Machine Learning has been used by scientists to support COVID-19 research.
    i. Combating Covid-19 using machine learning and deep learning: Applications, challenges, and future perspectives
    ii. Comparing machine learning algorithms for predicting COVID-19 mortality
    iii. The role of machine learning in health policies during the COVID-19 pandemic and in long COVID management
    iv. An Easy-to-Use Machine Learning Model to Predict the Prognosis of Patients With COVID-19: Retrospective Cohort Study

4. Create Training, Validation and Testing Datasets
    a. Before you can pick your model, you need to develop your Training, Validation and Testing (TV&T) datasets. There are a number of ways to do this, some tips on how to proceed below.
        i. A Comprehensive Guide to Train-Test-Validation Split.
        ii. Understanding training and testing data in terms of overfitting and underfitting challenges in Machine Learning
    b. Consider the integration of external datasets (see the 'Challenges with this Data' header in the project introduction for recommendations on selecting external data). Learn best practices for merging datasets.
5. Model Selection
    a. Now that we have our TV&T datasets ready, we can start selecting the type of ML model that we want to use for our research. There are *many many* ML models available (some examples here). Here are some of the most frequently used ML models, some of which may be useful for your analysis.
        i. Linear Regression - used to predict the value of a dependent variable based on one or more independent variables. It assumes a linear relationship between these variables. It is best used for trend analysis and forecasting.
        ii. Logistic Regression - used for binary classification problems, where the output is categorical and represents two classes. It estimates the probability that a given instance belongs to a particular category.
        iii. Decision Trees - used for both classification and regression problems. They model decisions and their possible consequences, representing an upside-down tree structure. It's a good choice when interpretability is important, like in strategic planning or medical diagnosis.
        iv. Random Forest - an ensemble method using multiple decision trees for classification, regression, and other tasks. It improves the predictive accuracy

and controls over-fitting. It's versatile and can be used in various domains, including e-commerce, finance, and healthcare.

v. Support Vector Machines (SVM) - used for classification and regression, working well for both linear and non-linear relationships. Effective in high-dimensional spaces, suitable for text classification and image recognition.

vi. K-Nearest Neighbors (KNN) - a simple, instance-based learning algorithm used for classification and regression. It predicts the output based on the majority vote or average of the K nearest instances. It is best suited for small datasets where the relationship between variables is complex.

b. Your choice of ML model is *very important*! We strongly encourage you to message our team in the Slack channel when you've selected your model. Our team and mentor cohort will help guide you at this stage.

6. Initial model development
    a. Now that you've selected your model, you'll want to begin developing that model and performing preliminary analysis. You may begin by training models using the preprocessed OxCGRT dataset and validate the performance at each stage.
    b. There are numerous methods to analyze performance. To begin with, consider exploring popular options such as the confusion matrix, precision, recall, and F1 score.

Your Summary Analysis

As you complete this Milestone, take some time to jot down notes on what you have learned. What key insights about COVID-19 have you learned from this exploration? What pieces of analysis do you want to bring into your final visualization and data story? How does this compare with the EDA you identified in the previous Milestone?

Milestone 3: Testing Other Data Analysis Tools/Model Optimization

Weeks of March 3rd and March 10th, 2024

Objective: Explore Geospatial Analysis; Improve model or analysis performance, fine-tune parameters, and validate results.

Weeks 5 & 6: Geospatial Analysis (Required) and/or Machine Learning Optimization (Optional Challenge)

> Geospatial Analysis (Required)

1. Geospatial Visualization
    a. In the previous Milestone, we analyzed data by examining change over time. Now, let's consider how data represented spatially can also give us useful insights. We'll observe data when plotted at a regional level, then learn how to compare that data. We'll wrap up by writing down our initial findings from this exercise.
    b. Learn about geospatial visualizations and how they can help with data analysis. Consider how this visualization format is (or is not) helpful for analyzing COVID policies.
    c. Explore the concepts of geospatial data with this NSDC Flashcard series.
2. Regional Analysis
    a. Plot the data on maps to visualize the evolution of COVID policies using either Excel, or a Google Colab notebook if you're comfortable in Python.
    b. Consider using libraries like `folium` or `geopandas` for this. You can also create heatmaps, choropleth maps, or other relevant plots.
3. Comparative Analysis
    a. Create a few simple visualizations with your data. Compare the impact of COVID-19 policies across different locations or regions. Does one analysis format in particular support your research goals?

> Machine Learning Optimization (Optional Challenge)

Note that this is an optional challenge, recommended for participants who have some previous experience with Machine Learning. To work on this section, you must have completed the other Optional Challenge in Milestone 2. Expect that this Optional Challenge will require an additional five hours of project work per week.

1. Model Optimization
    a. Once you've started to develop your model, you'll need to optimize it for research success! This is where you will use the Validation data that you developed in the last Milestone.
        i. Start the optimization by fine-tuning your hyperparameters, taking your initial results into account. Setting your hyperparameters will allow you to refine your model and increase its performance.
        ii. 10 Tools for optimization in Python.
        iii. Additional tips on hyperparameter tuning.

        iv. Begin with an introduction to fine-tuning, then expand your knowledge with this detailed guide.

2. Model or Analysis Validation
    a. Next, you'll need to validate your model. Learn about model validation here. Validation lets us confirm that your model can process data it hasn't seen yet and that the model is working the way you want it to.
    b. Perform cross-validation, hypothesis testing to ensure robustness.
        i. Address overfitting/underfitting issues. Overfitting occurs when the model is trained too well and learns the details and noise in the training data set instead of the true underlying patterns.
        ii. Read this guide on how to ensure model robustness.
        iii. Experiment with cross-validation using the popular K-Fold technique.
3. Some suggested optimization topics to explore:
    a. Feature Engineering: This step may be completed by some of you in the previous milestone. But it is recommended to enhance model performance by preprocessing Oxford Policy Response Dataset features effectively.
    b. Hyperparameter Tuning: Experiment with hyperparameters tailored to the dataset to optimize model performance.
    c. Cross-Validation: Validate models using subsets of the dataset to ensure robustness in policy response analysis.
    d. Ensemble Methods: Combine models to predict policy impacts accurately based on the Oxford dataset.
    e. Model Evaluation Metrics: Select metrics aligned with policy impact assessment objectives.
    f. Feature Selection: Identify influential policy features to enhance model efficiency and interpretability.

Your Summary Analysis

As you complete this Milestone, take some time to jot down notes on what you have learned. What key insights about COVID-19 have you learned from this exploration? What pieces of analysis do you want to bring into your final visualization and data story? How does this compare with the EDA you identified in the previous Milestone?

Milestone 4: Advanced Visualization and Storytelling

Weeks of March 17th and March 24th, 2024

Objective: Create advanced visualizations and communicate insights effectively.

Weeks 7 & 8: Explore Advanced Data Visualization Techniques

1. Finalize your Analysis
    a. At this point, we have taken our data and produced information from it. Now, we want to transform that information into insights. It's the insights that we will communicate and share with our audience in the form of a final analysis document and data visualization.
    b. Read up on the Data > Information > Insights transformation process.
    c. Learn about the difference between an analytical research paper, argumentative paper/persuasive paper. (Your final submission should be an analytical paper which takes into account the data from OxCGRT!)
        i. Here is some information about how to develop an academic tone for your analytical paper.
    d. Think about the key insights you want to share with your audience. What is one key insight you want someone to take away from your research? What did you find that made you say "wow!"? We'll explore the best way to communicate that wow factor with visualizations and data storytelling below.
    e. What happens if your data doesn't align with your initial hypothesis? That means you've learned something new! "Scientific failure" is an important part of the process - don't get discouraged.
2. Exploring Advanced Visualizations
    a. Now that we've considered some ways to analyze COVID-19 policy data, we're ready to visualize our ideas.
    b. Consider your target audience and what kind of story we'll need to tell to persuade them that our insights are relevant and informative. In this Milestone, we'll brainstorm and then build a visualization.
    c. Explore advanced chart types (e.g., stacked area charts, bubble charts, etc.).
    d. Decide which visualization type is most appropriate for your research question. Compare types and how they communicate specific ideas or trends.
3. Build your Visualization

a. Now we come to the fun part! Give yourself plenty of time to build your visualization. You can either use your Google Colab notebook (again, if you're comfortable in Python) or learn Tableau.
    i. If you're interested in Tableau, begin by creating a Tableau Public account. There are many good Tableau tutorials out there. We recommend the following to get you started:
        1. Learn Tableau in 15 minutes
        2. Tableau for Data Science and Data Visualization
        3. Create Covid-19 in India Dashboard
b. Accessible Visualizations
    i. Consider how your visualization might appear to people with disabilities who may not be able to distinguish between muted colors or see your chart at all. Are there any changes you can make so that differently-abled scientists can also learn about your research?
    ii. Other tools:
        1. Color Contrast Checker
        2. SAS Graphics Accelerator
        3. TwoTone Data Sonification Tool
        4. Making Visual Studio Accessible

4. Data Storytelling
    a. Write 1-2 pages summarizing your research question, your reasoning for selecting your data visualization choices, and key insights from your data analysis. You may also wish to include your outstanding research questions that could not be answered by the dataset.
    b. Learn to tell a compelling data-driven story.
    c. As you write your research summary, consider what your key insight is. Then, identify a possible policy recommendation you would make as a result of that key insight. Make sure to note what further research might be needed, if any, before your recommendation is adopted.

Milestone 5: Final Analysis, Documentation, and Presentation
Week of March 31st, 2024
Optional Research Presentations: Friday, April 5th

Week 9: Final Analysis & Presentation Preparation

1. Submit your Final Visualization and Analysis
    - You're nearly done! In previous Milestones, we've learned about the importance of data visualization, the different ways we can analyze data, and then found new ways to plot data through data storytelling. In this Milestone, we'll refine our visualization and get ready to share our visualization with others. We'll prepare how we discuss our visualization and request feedback from our peers.
    - Submit a final visualization that you feel best answers your research question from Milestone 1 to CICStudentWorkingGroup@columbia.edu. In your packet (any format you choose), include the 1-2 page summary you developed in Milestone 4.
      Submissions are due by 5pm ET on Friday, April 5, 2024. Students who create and share a research submission may receive a certificate of completion for this project.
    - If you are using a Google Colab notebook, share your final visualizations with lc3460@columbia.edu.
    - If you are using Tableau, publish your Tableau Public visualization so others can view your final workbook. Alternative: Instructions for Desktop to Public publishing
2. Final Presentation Preparation (Optional)
    - In our final CIC Student Working Group meeting of the semester (Friday, April 5, 2024), we will open the floor to students who are interested in presenting their research findings and visualizations. We will spend a few minutes discussing your research question, your storytelling goals, and rationale for selecting a particular visualization format.
    - Practice your presentation by incorporating storytelling techniques.
    - For the presentation, share your discoveries, how you did your research, and what you found in a clear and interesting way. Show why your analysis is important, how it supports scientific discovery, and can support future research. Share your key insight, what you think it means, and what policy recommendations you would make as a result. If you think there are opportunities for further research and exploration in this area, talk about them with the group!

- This is a *friendly* and collaborative working group environment so you should feel comfortable sharing your ideas with your colleagues. Feedback will strengthen your research practices.
- Feel free to request feedback from the CIC team or other Working Group members as you develop your presentation. Accepting feedback and finding new ways to frame your analysis is an important part of the scientific process! Doing this will make your final project much stronger.

3. Future Practice (Optional)
    - The best way to improve your data visualization skills is through continued practice. Keep your skills sharp by taking on additional visualization challenges, like these ones offered through Makeover Monday.