**CDC PLACSE DATA**

**PROJECT OVERVIEW:**

This project will utilize the CDC PLACES website health database to access Indiana and Kentucky state datasets for analysis and visualization. PLACES is a collaboration between CDC, the Robert Wood Johnson Foundation, and the CDC Foundation. PLACES provides health data for small areas across the country. This allows local health departments and jurisdictions, regardless of population size and rurality, to better understand the burden and geographic distribution of health measures in their areas and assist them in planning public health interventions. Goal is to provides model-based, population-level analysis and community estimates of health measures to all counties, places (incorporated and census designated places), census tracts, and ZIP Code Tabulation Areas (ZCTAs) across the United States.

Project Details:

* Data will be queried from CDC PLACES using the online interface and tools. Key datasets to be analyzed factors like Prevalence, Prevention, Health risk behaviors, Health outcomes of various diseases, insurance, medicines which further segregated in small geographical parts like:
  + County
  + Place
  + Census Tract
  + ZIP Code Tabulation Area (ZCTA)
* Data will be queried with parameters like time period, location, demographics, health outcomes, and other variables relevant to research questions.
* Analysis will focus on visualizing trends over time and geographic patterns, aided by charts, graphs, and maps.
* All analyses and visualizations produced will be publicly shared to highlight the accessibility and value of Places data.
* If feasible, the project may also explore using automated methods like API access or web scraping to efficiently query and extract data from the PLACES site.

Done understanding of project and gone through the data.

**DELIVERABLES:**

* Published public health data stories and visualizations based on CDC PLACES data
* Code and technical documentation for accessing, analyzing, and visualizing PLACES data
* Use case descriptions and examples demonstrating applied uses of CDC PLACES data
* Need to collect links manually from the data set as no input provided by requister, Development of Python Script and formating of downloaded data files
* Need to discover data set of site, Combined list of links and API of data either releases or updated in 2023.
* Developed Python Script and downloaded files.

**PROJECT BACKGROUND**

* PLACES reports county, place, census tract, and ZCTA data and uses small area estimation methods to obtain 36 (29 in the 2022 and 2021 releases; 27 in 2020 release) chronic disease measures for the entire United States.
* The data are published through a public, interactive “PLACES” website that allows users to view, explore, and download data by county, place, census tract, and ZCTA.
* Although limited data are available at the county and metropolitan levels, PLACES represents a first-of-its-kind data analysis to release information for all US counties, places, census tracts, and ZCTAs. This system complements existing surveillance data by providing estimates necessary to understand the health issues affecting the residents of local areas of all sizes and regardless of urban or rural status; develop and implement effective and targeted prevention activities; identify health problems; and establish key health objectives.

**INITIAL TECHNICAL INVESTIGATION**

The team initially collected all 58 links from where data need to download and then explored using the CDC PLACES API to directly access the required datasets. Data links collected with the filter for 2 states Indiana and Kentucky as per suggestion from project owner.

Web scrapping is possible for direct download data from these links and working on it.

Web scraping tools like Selenium and Python could be leveraged to query and extract the data needed from the website.

**PROJECT PLAN**

If web scraping of CDC PLACES is viable, this method will be used to obtain filtered, two states data which are released or updated in 2023. The scraped data will be stored in a database for further analysis.

Data visualizations and interactive dashboards will then be developed to bring the data to life. Use cases and applications of the data will be showcased to promote open public health data.

All code and processes documented will be open source to allow reproducibility and collaboration. Ethics and responsible data use will be ensured throughout the project.

**PROGRESS**:

Our project aims to provide a comprehensive analysis of the "500 Cities" dataset, with a particular focus on health behaviors, outcomes, and preventive services.

* As there was no links provided to us so we manually collected all links for the data either released or updated in 2023. Further need to use Web scraping programs to scrape data off of websites, even if they don't have an API. We've used web scraping successfully before to get data from other sites like CDC Wonder.
* We'll need to research more to see if web scraping CDC PLACES and automate whole download process. After downloading need to update the downloaded files in required format for further analysis and visualizations.
* The team also started working on a Python script that could eventually be used to automate downloading the data if the team does get the green light to scrape the site. The script will help grab the data quickly and efficiently.
* Python script for automatic downloading of files is now fully functional
* Additional Python code for further file formatting is in progress (about half completed)

**CHALLENGES:**

* Getting to Know the Data Sources: At the start, understanding where the CDC data came from was little challenging. We had to dive deep into websites and APIs, which, were pretty complex to navigate.

* Hunting for API Links and Task Details: Getting those specific API links and clear task details are challenging as it slowed down our progress in building the scripts we needed.

* Dealing with Data Discrepancies: The data formatting and differences between the old and new indicator names were a bit difficult.

* Scripting Woes: Even though we made headway in scripting for data retrieval, there were moments when we had to make the scripts more adaptable to changes in the data structure.

**Project Plan and Progress Summary:**

**September:**

**Plan**: Initiate project discussions, explore CDC data sources, and conduct web/API exploration for data understanding.

**Progress**: Started project discussions and ongoing exploration of CDC data sources, focusing on understanding data through web and API exploration.

**Octobe**r:

**Plan**: Download sample data, analyze it, and continue Python script development for data collection. Seek API source links and task-specific details.

**Progress**: Successfully downloaded sample data, analyzed and summarized findings. Continued Python script development, seeking API source links and specific task information.

**November:**

**Plan**: Progress with Python script development for automated data retrieval, address formatting challenges, and resolve differences in indicator names.

**Progress**: Continued developing Python scripts for automated data retrieval; successfully downloaded data in CSV format. Worked on file formatting, partially completed code. Encountered challenges with differing indicator names and worked on establishing relationships between old and new indicators.

Overall Summary:

The project started off with discussions and exploration of CDC data sources. It progressed as planned, focusing on data understanding, sample analysis, and Python script development. However, challenges emerged with formatting and indicator names, requiring ongoing efforts to address these issues while ensuring successful automation of data retrieval.