Project Milestone 1: Impact of COVID-19 on Stress Level

DSC540 Data Wrangling

Jyoti Dave

2024-12-20

Project Subject Area This project aims to analyze how various demographic factors, COVID-19 exposure, and personal circumstances influenced stress levels during the pandemic by integrating data from multiple sources to identify patterns in different regions and population groups. We aim to uncover patterns and correlations that explain the rise in stress, contributing to public health efforts in mental health management.

Data Sources:

Flat File:

- **Description:** A dataset containing stress level, health issues, increased work hours, hours worked per day, stress level, technology and many more.
- **Source:** Kaggle.
- **Link/File:** https://www.kaggle.com/code/tanechklangburam/impact-of-covid-19-on-stress-level/notebook6

API:

- **Description:** Covid tracking API, contains datasets such as state, hospitalized, positive, negative case details by state wise in USA.
- **Source:** Covid tracking API.
- **Link:** https://api.covidtracking.com/v1/states/current.json

Website:

- **Description:** A publicly accessible Wikipedia website providing COVID-19 pandemic statistics in a tabular format.
- **Source:** Wikipedia.
- **Link:** https://en.wikipedia.org/wiki/Statistics_of_the_COVID19_pandemic_in_the_United_States

Relationships: The data sources are connected as follows:

- **CSV File:** Contains a list of mental stress level, health issues and other metadata about working pattern during COVID-19. I need to add the location column as it's missing.
- **API:** Contains the location("state") of each record along with other health related metadata**.**
- **Website:** Contains a list of the number of deaths happening for each state**.**
- All 3 of these data sources are related by location**.**

Approach/Plan:

1. **Data Extraction and Cleaning:**
   - Download the flat file dataset and clean it for duplicates, missing values, and formatting inconsistencies.
   - Query the API to fetch COVID-19 data, ensuring compliance with API rate limits.
   - Scrape or manually retrieve tabular data from the Wikipedia website.
2. **Integration and Transformation:**
   - Standardize location identifiers (e.g., state) across datasets.
   - Merge datasets based on shared attributes such as state.
   - Add missing column in the flat file.
3. **Analysis:**
   - Explore relationships between stress level distribution by state, sector, technology, work hours on productivity etc.
   - Visualize insights using tools like python.
4. **Reporting:**
   - Summarize the findings in a report and provide actionable recommendations for stakeholders.

Challenges:

- **Data Quality:** Inconsistencies in location identifiers across datasets may hinder integration.
- **API Rate Limits:** Managing requests within the limitations of the API might slow the data collection process.
- **Web Scraping Constraints:** Some websites may restrict automated data extraction, requiring manual data collection.
- **Data Integration:** Aligning data with different formats and structures could be complex and time-consuming.
- **Missing and Incomplete Data:** Some records contain missing values, particularly in critical fields like "Hours Worked Per Day." While addressing this through data cleaning and imputation, these techniques may introduce bias or affect the accuracy of the results. Missing data for other key variables may also lead to an incomplete picture, especially if missing values are not random but instead reflect certain employee groups or industries disproportionately.
- **Reliance on Self-Reported Data:**

Much of the data is likely self-reported, meaning it depends on the individual perceptions and interpretations of the employees. Self-reported data can introduce biases, such as over-reporting or under-reporting of stress levels, productivity changes, or health issues. This reliance on subjective reporting limits the precision of measures like “Stress Level” or “Productivity Change” and may not capture objective work outcomes accurately.

Ethical Implications:

- **Privacy Concerns:** Ensuring that no personally identifiable information (PII) from the datasets is exposed or misused is critical.
- **Bias in Data:** Increased work hours and stress level can be influenced by affected by covid attribute, and analysis must account for and communicate these biases.