# Reflection on the Data Science Project

By: Hewan Kasie, Paige O'Neil, Naina Sharma, and Kaleigh West

**Data Selection and Exploration**

Our project focused on analyzing societal trends by looking at two datasets:

1. **Mental Health Dataset**: This dataset focused on the prevalence of various mental health disorders across different countries, with a specific focus on age-standardized rates for conditions like schizophrenia, depression, and anxiety.
2. **Causes of Death Dataset**: This dataset provides detailed information on global mortality causes, filtered for the United States, to understand the relationship between health trends and preventable causes of death.

We selected these datasets based on their relevance to society; the serious effects of mental health disorders are becoming increasingly apparent and have been closely influencing public health policies. Causes of Death provided the perfect link to mental health with respect to mortality data. Both were obtained from Kaggle and contained enough historical data for a decent trend analysis.

The process of data selection was not devoid of its own challenges. Initial explorations showed inconsistent data formats and a number of values that were missing, especially when the two different sets were merged. For instance, one set used "Country/Territory" while the other used "Entity." Such discrepancies first needed to be standardized and restructured carefully before they could actually be merged. Another challenge was dealing with inconsistent data units between the two datasets. In particular, the Mental Health Dataset reported prevalence of diseases in terms of a percentage of the population, whereas the Causes of Death Dataset reported absolute numbers of cases. However, we felt that we would be able to work through these challenges, which we did by narrowing down our investigation to the US only and by using published data on the US yearly population to convert the prevalence rates to numbers of cases. The general understanding derived from this phase is the importance of the cleanliness and consistency of data that is required for any further analysis. Given more time, an additional line of exploration would involve other social determinants of health datasets to increase our current level of information for more insightful analysis.

**ETL Process (Setup & Implementation)**

The **ETL (Extract, Transform, Load)** process was a critical component of the project. Our objective was to clean, transform, and merge the datasets so that we could draw meaningful insights. The process consisted of several steps:

1. **Extraction**: We initially extracted the datasets from Kaggle in CSV format. Both datasets were comprehensive but contained significant noise, especially with missing or inconsistent data entries.
2. **Transformation**:
    - **Cleaning**: The cleaning process involved renaming columns for clarity and removing irrelevant columns. For example, unrelated variables from the mental health dataset were removed. An example is renaming "Entity" to "Country/Territory" to match the Causes of Death dataset. We also multiplied the Mental Health Dataset data by the population of the United States (for that given year, as reported by the US Census Bureau) so that the mental health data could be more easily compared to the Causes of Death data.
    - **Filtering**: Next, after cleaning, we filtered this data to focus only on data related to the U.S. Because both datasets had an extensive geographical scope, limiting our analysis to only the U.S. made our data more manageable for us to draw conclusions from.
    - **Merging**: Merging the data sets proved to be the most intricate part of the transformation process. The two datasets were combined using the shared keys "Country/ Territory" and "Year", giving us one single dataset which contained both mental health and mortality data. United States population data for every year were also added so that we could contextualize the rates of the disorders about the population.
3. **Loading**: Once the data was cleaned and merged, it was loaded into **Google Cloud Storage** and **MongoDB** for centralized access. Google Cloud allowed us to store large datasets while ensuring accessibility for team members and collaborators. Storing the data in a MongoDB cluster will allow for ease of querying and finding particular data values in the future.

The ETL implementation process had a number of challenges. For example, in one dataset, country names were spelled out in full, while the other used country codes. It took us a great deal of time to debug these issues since we needed to manually match the rows or use string-matching techniques. We also had to overcome missing data. This was particularly problematic in the Causes of Death dataset, which had several missing years for many countries. Dropping those rows was the practical thing to do, but it made the dataset incomplete. Python, especially the Pandas, and NumPy libraries, were used to manipulate and clean the data. These allowed us to efficiently manipulate very large datasets. We utilized MongoDB for the management and querying of the database once the data was loaded. Finally, the cleaned data was hosted in Google Cloud Storage for ease of access and sharing among collaborators.

**Data Analysis**

After cleaning and merging the data, analysis of trends and relationships in the data that could provide insight into public health policies was performed. We analyzed the correlation between mental health disorders and preventable deaths since the results could have informed discussions around public health interventions.

Our analysis revealed the following key insights:

1. **Mental health disorders have been on the rise in the U.S.** We saw remarkable year-on-year increases in mental health disorders, especially anxiety, depression, and schizophrenia. These findings point out the dire need for better mental health support and resources.
2. **Anxiety disorders had the highest prevalence,** as a big proportion of the population was suffering from them, while in usual conditions, anxiety is not presented in actuality in the public health data.
3. **A noticeable correlation exists between mental health trends and causes of preventable deaths.** Specifically, there was an association between higher rates of mental health disorders and increased deaths from preventable causes such as suicide and substance abuse. This finding strengthens the notion that mental health is a priority in public health. Additionally, deaths related to alcohol consumption had a positive correlation with depressive disorders and a negative correlation with anxiety disorders. We also found a strong positive correlation between alcohol use disorders and cirrhosis and other chronic liver diseases, a finding consistent with our hypotheses.

We used visualization techniques to display these trends: line graphs showing the increase in trend for mental health disorders, bar charts comparing different disorders, and scatter plots to visualize the correlation between mental health data and preventable deaths.

**Cloud Storage and Documentation**

For managing the large datasets, we used Google Cloud Storage, which provided a secure and scalable environment. This allowed us to store both the raw and cleaned datasets in one place for easy access by the team and for future reference. Cloud storage also made real-time collaboration possible, with team members able to access and work on the data at the same time. We set up access control by making the bucket itself public to all users so that anyone with a link can view the data, and each team member was made a bucket owner so that everyone on our team would have equal access to the dataset. This allowed us to control who could view, edit, and upload the data, keeping sensitive information secure. Version control was also implemented in order to track changes and allow us to roll back to previous versions if necessary. This process really underscored data security and accessibility in collaborative work. Cloud-based storage made not only the data more accessible but also ensured that all members of the team were using the latest version of the dataset.

**Challenges Faced**

Throughout the project, we encountered several technical and collaborative challenges:

1. **Data Quality Issues**: Both datasets had inconsistent formats, missing values, and irrelevant columns. Resolving these issues required extensive cleaning and transformation.
2. **Debugging ETL Errors**: The ETL pipeline often failed when merging datasets due to discrepancies in the shared keys. These errors were challenging to debug, requiring us to explore the datasets manually and apply string-matching techniques.
3. **Cloud Integration**: Setting up **Google Cloud Storage** was initially tricky, particularly with regard to configuring access permissions and ensuring that all data was stored securely. Additionally, getting the JSON key so that our script has the proper credentials to upload to our bucket proved to be challenging.
4. **Database Creation**: We had difficulty in deciding if we should use MySQL or MongoDB for the database portion of the project. For both SQLite and Mongo, we felt that we had experience querying/working with data that was already loaded into MySQL or Mongo, but not much with uploading our own data into a database in MySQL or Mongo. Ultimately, we decided to create a cluster in Mongo for which to put our data. We learned a lot in the process of working through taking what we did in class and expanding our knowledge through further research and experimentation to have a more well-rounded understanding of working with databases, not only for querying but also for data storage before querying.
5. **Team Coordination**: With different team members working on different aspects of the project, coordination was key. Although we encountered some delays due to conflicting schedules, these challenges ultimately helped improve our communication and delegation skills.

**Lessons Learned and Skills Gained**

Important here were the lessons of how communicative members in a team should be, as well as how technical issues can pose issues in a project. It was vital to clearly explain ideas and report progress, especially when there was some problem at certain moments. Also, learning to use the strengths of every member within the team helped in getting through some challenges, allowing us to get the job on time.

In terms of **technical skills**, we gained proficiency in:

- **Python**: We used **Pandas** for data manipulation, **Matplotlib** for data visualization, and **NumPy** for numerical analysis.
- **Cloud Storage**: We became familiar with using **Google Cloud** for data storage, which involved setting up cloud buckets and managing access controls.

- **Data Visualization**: Creating clear and compelling graphs helped us communicate the results effectively to our audience.

In the future, more advanced data visualization techniques can be used, such as an interactive dashboard with Plotly or Tableau. Some aspects of the ETL pipeline can also be automated. For example, we could schedule checkpoints for tasks using Airflow, which will further streamline the process.

**Future Improvements**

Looking ahead, we would focus on the following improvements:

1. **Automating ETL Processes**: We could automate data extraction, transformation, and loading using Python scripts or workflow orchestration tools like **Apache Airflow**.
2. **Incorporating More Datasets**: Adding more datasets, such as information on healthcare access, could deepen the analysis and provide more context for the findings.
3. **Interactive Visualizations**: We plan to explore **interactive visualizations** that allow users to explore the data and trends on their own, providing a more engaging experience.

This project was a valuable learning experience, providing us with the skills necessary for future data science projects. It deepened our understanding of societal trends and enhanced our technical expertise in data transformation, analysis, and cloud storage.