Data Science Final Project

Reflection

Our project was about understanding the relationship between weather and climate disasters and state populations. We used two main datasets: one about natural disaster declarations in the U.S., and the other about U.S. population by state. The goal was to explore how different natural disasters affect the populations in each state and find any patterns or trends over time.

At the beginning, it was hard to find the right datasets that would work for our project. We needed data that showed both the types of natural disasters and the populations of states over time. We chose two datasets that we found on Kaggle. The US Natural Disaster Declarations dataset had categories that needed to be cleaned up, and the US Population by State dataset had missing data for some states in certain years, so we had to figure out how to deal with that.

We ended up deleting columns from the Natural Disaster dataset and keeping the US population the same. We coded the necessary portions to clear data that we did not need like Incident_begin_date, Incident_end_date, Disaster_closeout_date, Fips, etc which were not used in our graphs or needed for the data we were examining. We then created a new CSV with the filtered data and stored that file in our google cloud project for use.

Designing the ETL pipeline was one of the biggest challenges. The transformation part was difficult. For example, aligning the two datasets by year and state took a lot of work, and the disaster dataset had inconsistent date formats. We used Python and MySQL to make the pipeline work, and while it was complicated, it ended up being flexible and easy to update when needed.

Analyzing the data and figuring out what insights we could get from it wasn’t always easy. We had to balance between statistical methods and clear visualizations. A big challenge was figuring out how to graph the relationship between population size and disaster types. Some of the graphs were tricky to make look clear and meaningful, so we spent time improving them until they made sense.

Setting up Google Cloud Storage for our transformed data was another important task. The main challenge here was setting up the service accounts and managing permissions. It wasn't easy to make sure the data was secure but still accessible to everyone in our group. After a while, we were able to set up the cloud storage properly and store our data for future use. It was also difficult to upload the existing file into code we were using mainly because it had to be a certain file type for the code to run.

One of the most important lessons we learned was how crucial data cleaning and transformation are. Without clean data, it's hard to analyze anything. We spent a lot of time making sure the data was accurate and aligned, which paid off when we started the analysis. This showed us how important it is to clean and prepare data properly before trying to analyze it. It led to a smoother process for our group to find the data we wanted to use.

Working together as a team was a big part of the project's success. We had to communicate clearly and divide the tasks so that everyone knew what they were responsible for. At first, it was hard to set up the cloud storage because not everyone was familiar with Google Cloud and the ETL setup, but we met over the period of weeks to learn and use both.

Creating meaningful graphs and charts was another area where we learned a lot. Making sure that the visualizations were clear and easy to understand took several attempts. We used libraries like matplotlib and learned that it's important to not just show data, but to be specific in what was important in it. The clearer the graph, the easier it is to understand it.

This project helped us improve our Python skills, especially for data manipulation using pandas. We also learned how to work with cloud storage solutions like Google Cloud and how to use MySQL to store data. We got better at using matplotlib to create visualizations.

We gained hands-on experience with cloud storage, particularly with Google Cloud. We now understand how to manage service accounts, control permissions, and make sure our data is secure. This is a skill we can use in future data projects where we need to store and access large datasets.

Building and managing the ETL pipeline was a valuable learning experience. We now understand how to automate the data processing steps and ensure that the pipeline can handle

updates. We also learned how to handle issues like missing data and data inconsistencies.
We learned that good communication is essential for a successful project. By splitting the work, collaborating, and having regular check-ins, we were able to overcome the challenges we faced. This project taught us how to work better as a team and how to make the most of each person's strengths.

Overall, this project was a great learning experience. It gave us hands-on experience with the entire data science process, from data selection and exploration to ETL implementation and cloud storage setup. We faced several challenges, but by working together and learning new skills along the way, we were able to gain valuable insights about how natural disasters affect state populations. These skills and lessons will help us in future data science projects.