Melissa Viator

BU MET College Computer Science Department

CS689 Term Project Report

Fall 2022

Natural Disasters Data Warehouse Design & Implementation

Introduction

Natural disasters include all geophysical, meteorological and climate events which pose significant threats to human health and safety. The number and cost of weather and climate disasters is rising due to a combination of population growth and development along with the influence of human-caused climate change (climate.gov). For this reason, it is even more pertinent to understand the historical nature of natural disasters. The purpose of this project is to design and implement a dimensional data warehouse from various sources on natural disasters, climate, weather, and economic indicators to understand the statistics and associations of natural disaster events. Specifically, this project will 1) design a dimensional data warehouse based on relevant information from four data sources, 2) iteratively extract, transform, and load (ETL) the data into the data warehouse, accounting for slowly changing dimensions, and 3) visualize analysis results of the dimensional model, answering the predefined business questions.

Business Questions

The following business questions were developed to define the scope of the project and guide the design of the dimensional data warehouse. The six questions are central to the projects aims in understanding the statistics, associations, and implications of natural disaster events.

1. What is the average magnitude and duration by natural disaster type? What were the natural disaster events with the highest magnitude and longest duration?
2. What locations are most prone to natural disaster occurrences?
3. How have the number of natural disaster occurrences evolved over time? Is there a seasonality for particular types of natural disasters?
4. What is the average damage by natural disaster type? Which natural disaster type has the highest number of average deaths and people affected?
5. Is there a difference in a country's population/GDP on years where a natural disaster did or did not occur?
6. What is the relationship between temperatures and greenhouse gas emissions and what effects do they have on natural disaster occurrences?

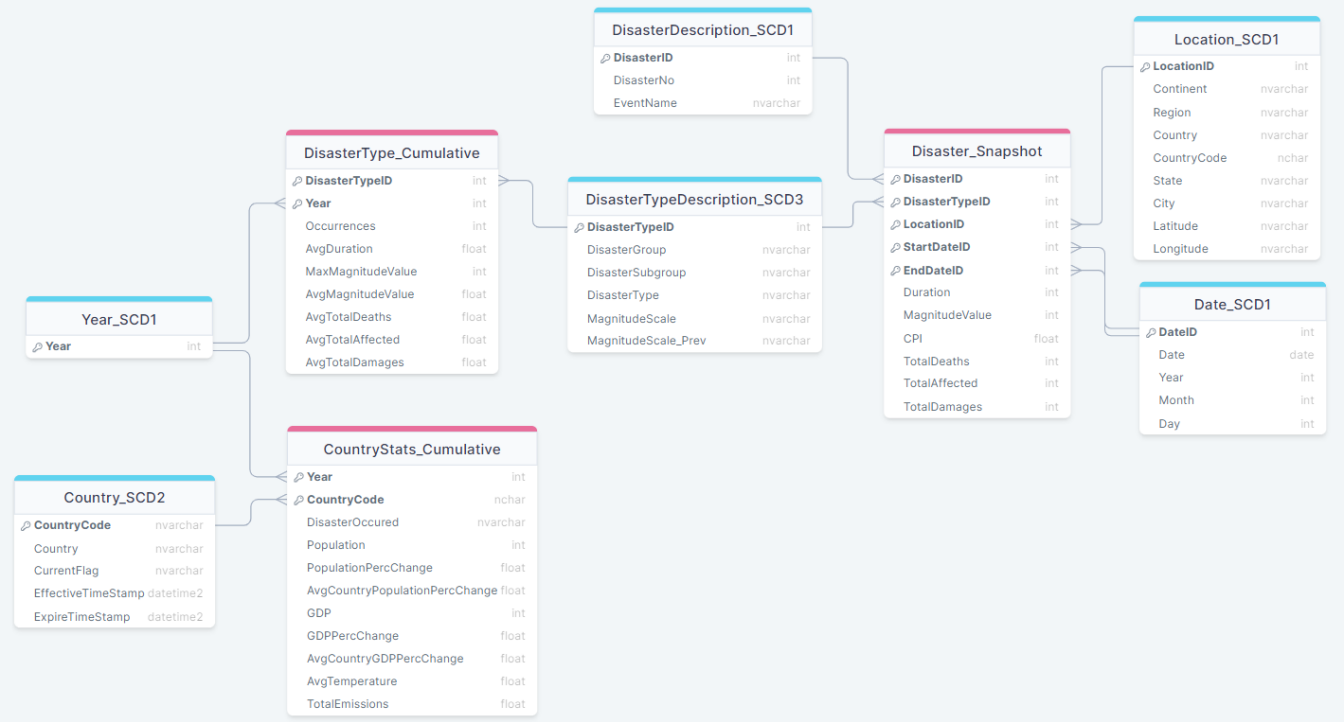
Datasets

The following list provides an overview of the four datasets utilized to build the dimensional data warehouse. In combination, the datasets contribute vital information on natural disasters, climate, weather, and economic indicators. With the array of datasets from various sources, the ETL process must account for disparities among the datasets.

- Natural Disasters dataset
  - Provides information on natural disaster events such as event name, grouping, magnitude, damages, etc.
  - Date Range: 1900 - 2022
  - Countries: 228
  - Dimensions: 50 columns x 16,399 rows
- Temperature dataset
  - Provides average daily temperature by city
  - Date Range: 1995 - 2020
  - Countries: 124
  - Dimensions: 8 columns x 1,134,925 rows
- Greenhouse Gas Emissions dataset
  - Provides greenhouse gas emissions (all, CO2, CH4, N2O, and F-Gas) by country and year
  - Date Range: 1990 - 2019
  - Countries: 195
  - Dimensions: 35 columns x 973 rows
- GDP dataset
  - Provides population and GDP per capita by country and year
  - Countries: 169
  - Date Range: 1900 - 2020
  - Dimensions: 5 columns x 21,682rows

Dimensional Design

The following diagram represents the dimensional design of the natural disasters data warehouse. The design follows a constellation schema with three fact tables (in red) and six dimensions (in blue). See below for further information on the dimensional tables and fact tables included in the data warehouse.



Link: https://drawsql.app/teams/melissa-viator/diagrams/cs689-project

Dimensional Tables

- DisasterDesc
  - Captures the disaster number and event name of the natural disaster
  - SCD Type 1 to overwrite errors in recording the natural disaster event
- DiasterTypeDesc
  - Captures the group, subgroup, type, and magnitude scale of the natural disaster
  - SCD Type 3 to capture changes in the magnitude scale of the natural disaster with additional columns; this is important to understand if the measurement has changed (i.e., Celsius to Fahrenheit)
  - Includes disaster type hierarchy (group → subgroup → type)
- Location
  - Captures continent, region, country, state, city, latitude, and longitude of the natural disaster
  - SCD Type 1 to overwrite changes in a location's country name, code, etc.; changes in the location does not affect the snapshot of the natural disaster
  - Includes location hierarchy (continent → region → country → state → city → latitude/longitude)
- Country
  - Captures country code and country name on a wider grain than the location dim
  - SCD Type 2 to capture changes in the changes in the country's name/existence with additional rows (i.e., if a country separates and no longer exists)

- Date
  - Captures date, year, month, and day of the natural disaster's start date and end date
  - SCD Type 1 to overwrite errors when capturing the date of natural disaster event
  - Includes date hierarchy (year → month → day)
  - Role playing dimension for the start and end date of the natural disaster
- Year
  - Captures year on a wider grain than the date dim
  - SCD Type 1 to overwrite errors when capturing the year

Fact Tables

- Disaster
  - Snapshot: Instance from one natural disaster event
  - Measures: Duration, magnitude value, CPI, total deaths, total affected, total damages
- DisasterType
  - Cumulative: aggregation of disaster statistics grouped by natural disaster type and year
  - Measures: occurrences, average duration, maximum magnitude value, average magnitude value, average total deaths, average total affected, average total damages
- CountryStats
  - Cumulative: aggregation of economic and climate statistics grouped by country and year
  - Measures: disaster occurred, population, population percent change, average country population percent change, GDP, GDP percent change, average country GDP percent change, average temperature, total emissions

Extract, Transform, and Load

The ETL process was iteratively performed in a combination of SQL and Python. See the documentation file attached for instructions on how to run the ETL process. The list below details the steps of one ETL iteration.

- **Extract**
  - Read csv files of the four datasets into Python dataframes as staging tables
- **Transform**
  - Data cleaning procedure in Python
    - Remove irrelevant data
      - Remove irrelevant variables
      - Filter data
      - Remove duplicate observations
    - Fix structural errors
      - Transpose dataframes
      - Clean latitude and longitude variables (built function, clean_lat_lon, to assist)
      - Compute city and state based on latitude and longitude (built function, get_location and add_location, to assist)

        - Resolve incongruent naming conventions for country code and country variables (built function, get_countrycode and get_country_name, to act like a lookup table)
        - Adjust datatypes
    - Handle missing data
        - Null values were removed when applicable, however, null values remain in the dataframes for some variables
  - Build dimensional tables in Python
    - Merge datasets
    - Unique observations based on composite of variables
- **Load**
  - Create dimensional tables and fact staging tables in SQL to prepare the loading of the data from Python
  - Insert extracted and transformed dimensional tables and fact staging tables into data warehouse

ELT Challenges

The transformation phase of the ELT process was the most challenging and laborious step in the project. There were a number of disparities among the four datasets that attributed to data inconsistencies. The country name and country code variables were particularly taxing to resolve the incongruent naming conventions. Methods were employed to handle structural errors and unify the datasets.

Slowly Changing Dimensions

To show the implementation with slowly changing dimensions, the ETL process was performed in two waves of data: wave 1 included years 1900 - 2009 and wave 2 included years 2010 - 2022. The ETL process above was performed separately for each wave of data (note that the second wave of data was loaded into temporary tables). After the load of the second wave of data, the dimensional tables from the two waves were merged, upholding the slowly changing dimensions. The list below summarizes how the three types of SCDs were implemented when merging the new data wave.

- SCD Type 1:
  - Check if the record from the source exists in the dimension
    - If the matching record exists, overwrite the dimensional record
    - If the record does not exist, insert the record
- SCD Type 2:
  - Check if the record from the source exists in the dimension
    - If the matching record exists, check if the column(s) of interest have updated values
      - If the column(s) have updates, add new record and update the previous record's flag and timestamps
    - If the record does not exist, insert the record with an active flag and expire timestamp to future date
- SCD Type 3:

- Check if the record from the source exists in the dimension
  - If the matching record exists, check if the column(s) of interest have updated values
    - If the column(s) have updates, set the previous value from the current field and the new value from the source
  - If the record does not exist, insert the record with a null as the previous value

Recall that Country_dim is a SCD Type 2. The screenshot below shows the delta after merging the data waves and upholding the requirements of SCD2. Notice that the country, Yugoslavia, is expired and the timestamp of the expiration.

CountryCode	Country	CurrentFlag	EffectiveTimeStamp	ExpireTimestamp
SRB	Yugoslavia	Expired	2022-11-16 19:00:48.057	2022-11-16 19:03:12.340
SRB	Serbia	Active	2022-11-16 19:02:12.167	9999-12-31 00:00:00.000

Recall that DisasterTypeDesc_dim is a SCD Type 3. The screenshot below shows the delta after merging the data waves and upholding the requirements of SCD3. Notice the current and previous magnitude scale.

DisasterTypeID	DisasterGroup	DisasterSubgroup	DisasterType	MagnitudeScale	MagnitudeScale_Prev
10	Natural	Meteorological	Extreme temperature	Celius	°C

Create Fact Tables

Recall that the fact staging tables have been extracted, transformed, and loaded along with the dimensional tables. Once the dimensional tables are loaded and merged, the fact tables are created in SQL. First, the staging tables from the two data waves are merged. Next the dimensions composed in the fact tables are joined back to the source by the natural keys. Finally, the measures are calculated and the record is interested into the fact table. While some of the fact tables are created with simple joins and aggregate function, the CountryStats_fact employed window function to calculate its measures. The SQL code below demonstrates inserting records into the CountryStats_fact with methods such as nested queries, case statements, casting, windowing and partitioning, etc.

```
INSERT INTO CountryStats_fact (CountryCode, Year, DisasterOccurred, Population,
PopulationPercChange, AvgCountryPopulationPercChange, GDP, GDPPercChange,
AvgCountryGDPPercChange, AvgTemperature, TotalEmissions)
SELECT CountryCode, Year,
CASE WHEN CountryCode+CAST(t.Year AS nvarchar(50)) IN (
             SELECT DISTINCT CountryCode+CAST(Year AS nvarchar(50)) AS CountryCodeYear
             FROM Disaster_fact
             LEFT JOIN Location_dim ON Disaster_fact.LocationID =
Location_dim.LocationID
             LEFT JOIN Date_dim ON Disaster_fact.StartDateID = Date_dim.DateID)
       THEN 'Yes'
       ELSE 'No'
       END AS DisasterOccurred,
Population, CAST(CAST((Population - PrevPopulation) AS FLOAT)/nullif(PrevPopulation, 0) *
100 AS DECIMAL (18,2)) AS PopulationPercChange,
AVG(CAST(CAST((Population - PrevPopulation) AS FLOAT)/nullif(PrevPopulation,0) * 100 AS
DECIMAL (18,2))) OVER (PARTITION BY CountryCode) AS AvgCountryPopulationPercChange,
GDP, CAST(CAST((GDP - PrevGDP) AS FLOAT)/nullif(PrevGDP, 0) * 100 AS DECIMAL (18,2)) AS
GDPPercChange,
AVG(CAST(CAST((GDP - PrevGDP) AS FLOAT)/nullif(PrevGDP, 0) * 100 AS DECIMAL (18,2))) OVER
(PARTITION BY CountryCode) AS AvgCountryGDPPercChange,
AvgTemperature, TotalEmissions
FROM (
SELECT CountryCode, Country, Year,
Population, LAG(Population) OVER (PARTITION BY CountryCode ORDER BY Year) AS
PrevPopulation,
GDP, LAG(GDP) OVER (PARTITION BY CountryCode ORDER BY Year) AS PrevGDP,
AvgTemperature, TotalEmissions
FROM CountryStats_fact_stage ) t
ORDER BY Country, Year DESC;
```

Final Tables

The final dimensional tables and fact tables have been integrated into the data warehouse. The screenshots below illustrate the first ten rows of the tables.

DisasterDesc_dim:

DisasterID	DisasterNo	EventName
4	1902-0003-GTM	Santa Maria
5	1902-0010-GTM	Santa Maria
7	1903-0012-COM	Mount Karthala
17	1907-0001-IND	Bubonic
21	1909-0001-CHN	Bubonic
24	1909-0002-IDN	Bubonic
27	1910-0001-CHN	Pneumonic
48	1918-0001-BGD	Influenza
49	1918-0015-CAN	Spanish Influenza
53	1919-0001-IDN	Mt. Kelud

DisasterType_dim:

DisasterTypeID	DisasterGroup	DisasterSubgroup	DisasterType	MagnitudeScale	MagnitudeScale_Prev
13	Natural	Biological	Animal accident	Not Applicable	NULL
1	Natural	Biological	Epidemic	Vaccinated	NULL
2	Natural	Biological	Insect infestation	Not Applicable	NULL
3	Natural	Climatological	Drought	Km2	NULL
14	Natural	Climatological	Glacial lake outburst	Not Applicable	NULL
4	Natural	Climatological	Wildfire	Km2	NULL
15	Natural	Extra-terrestrial	Impact	Not Applicable	NULL
5	Natural	Geophysical	Earthquake	Richter	NULL
6	Natural	Geophysical	Mass movement (dry)	Not Applicable	NULL
7	Natural	Geophysical	Volcanic activity	Not Applicable	NULL

Location_dim:

LocationID	Continent	Region	Country	CountryCode	State	City	Latitude	Longitude
1	Africa	Eastern Africa	Burundi	BDI	Bururi	Bururi	-3.86	29.67
2	Africa	Eastern Africa	Burundi	BDI	Cankuzo	Cankuzo	-3.14	30.4
2237	Africa	Eastern Africa	Burundi	BDI	Gitega	Gitega	-3.47	29.99
2374	Africa	Eastern Africa	Comoros	COM	Anjouan	Sima	-12.23	44.3
15	Africa	Eastern Africa	Ethiopia	ETH	Dire Dawa	Dire Dawa	9.67	42.19
17	Africa	Eastern Africa	Ethiopia	ETH	Gedarif State	Al Qadarif	13.21	34.76
27	Africa	Eastern Africa	Kenya	KEN	Manafwa	Bugobero	0.88	34.28
47	Africa	Eastern Africa	Malawi	MWI	Sennar State	Ad Dinder	13.45	34.29
2747	Africa	Eastern Africa	Mauritius	MUS	Moka	Moka	-20.23	57.52
2095	Africa	Eastern Africa	Rwanda	RWA	Kigali City	Nyarugenge District	-2.06	30.01

Date_dim:

DateID	Date	Year	Month	Day
9817	2022-01-01	2022	1	1
10547	2022-01-02	2022	1	2
8134	2022-01-03	2022	1	3
8781	2022-01-04	2022	1	4
7854	2022-01-05	2022	1	5
9537	2022-01-07	2022	1	7
8213	2022-01-08	2022	1	8
10528	2022-01-09	2022	1	9
7791	2022-01-10	2022	1	10
8901	2022-01-11	2022	1	11

Country_dim:

CountryCode	Country	CurrentFlag	EffectiveTimeStamp	ExpireTimestamp
AFG	Afghanistan	Active	2022-11-16 19:00:47.810	9999-12-31 00:00:00.000
ALB	Albania	Active	2022-11-16 19:00:47.813	9999-12-31 00:00:00.000
DZA	Algeria	Active	2022-11-16 19:00:47.817	9999-12-31 00:00:00.000
AND	Andorra	Active	2022-11-16 19:00:47.817	9999-12-31 00:00:00.000
AGO	Angola	Active	2022-11-16 19:00:47.820	9999-12-31 00:00:00.000
ATG	Antigua and Barbuda	Active	2022-11-16 19:00:47.823	9999-12-31 00:00:00.000
ARG	Argentina	Active	2022-11-16 19:00:47.827	9999-12-31 00:00:00.000
ARM	Armenia	Active	2022-11-16 19:00:47.830	9999-12-31 00:00:00.000
AUS	Australia	Active	2022-11-16 19:00:47.830	9999-12-31 00:00:00.000
AUT	Austria	Active	2022-11-16 19:00:47.830	9999-12-31 00:00:00.000

Year_dim:

Year
2022
2021
2020
2019
2018
2017
2016
2015
2014
2013

Disaster_fact:

DisasterID	DisasterTypeID	LocationID	StartDateID	EndDateID	Duration	MagnitudeValue	CPI	TotalDeaths	TotalAffected	TotalDamages
394	5	152	632	632	1	6	11.62799038	2	49350	2000
470	5	1847	787	787	1	7	12.83661491	19	7618	600
558	5	631	916	916	1	7	14.32816969	29	88112	4000
629	5	319	1010	1010	1	7	16.3855937	21	3745	200
733	5	350	1116	1116	1	8	21.0017319	23000	4993000	1000000
929	5	940	150	150	1	7	6.424965817	3022	79540	40000
1355	5	335	644	644	1	6	11.62799038	125	139720	35000
1363	5	648	718	718	1	8	11.97858736	120	108356	15000
1429	5	1639	747	747	1	7	12.31072819	183	326073	3000
1581	5	1620	935	935	1	7	14.94324951	878	88665	5000

DisasterType_fact:

DisasterTypeID	Year	Occurances	AvgDuration	MaxMagnitudeValue	AvgMagnitudeValue	AvgTotalDeaths	AvgTotalAffected	AvgTotalDamages
4	2019	14	5	169967	169967	23	4743	932750
4	2017	15	28	59000	7568	25	3883	2417428
4	2016	10	31	6000	6000	6	1361	785875
4	2007	18	9	1600	1600	11	5392	919490
4	2000	30	1	680000	30823	5	8077	426485
4	1994	13	1	10000	2732	14	15464	76000
4	1987	7	1	25000	7170	191	152752	105000
5	2022	22	1	8	6	91	12960	1900774
5	2021	28	1	7	6	182	10699	807563
5	2020	17	1	7	6	19	29666	1622877

CountryStats_fact:

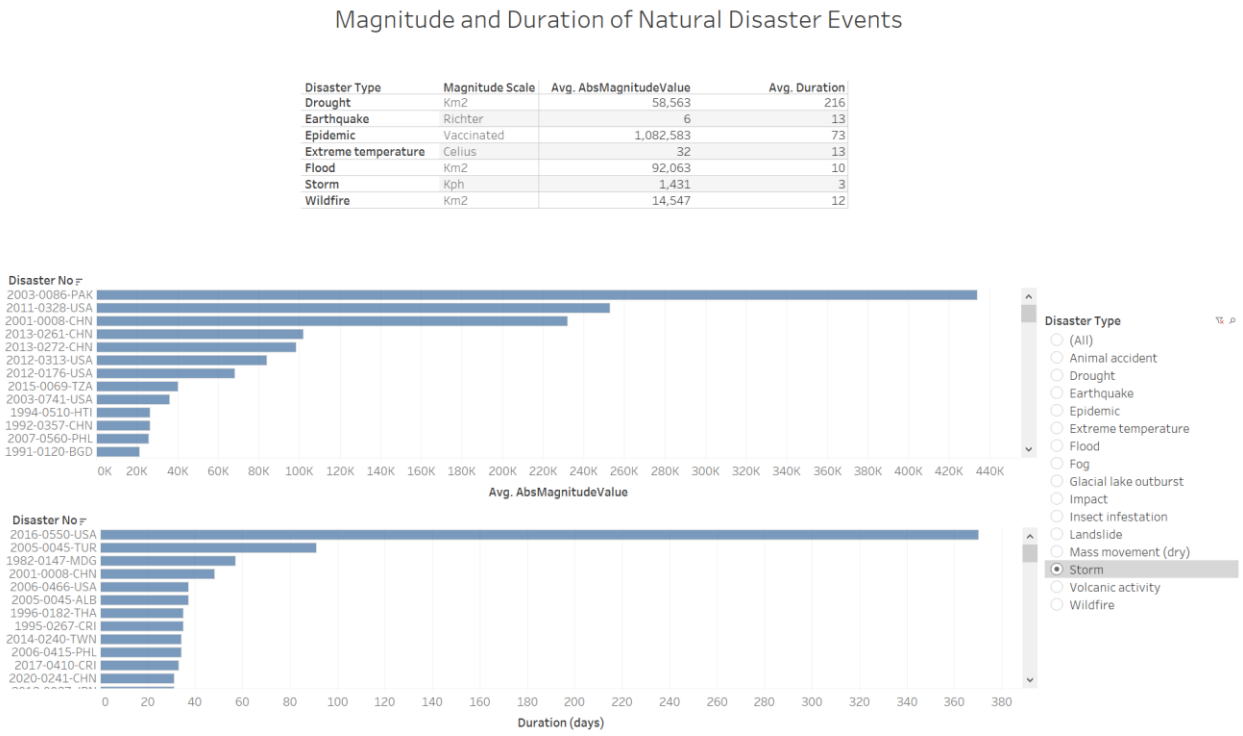
CountryCode	Year	DisasterOccurred	Population	PopulationPercChange	AvgCountryPopulationPercChange	GDP	GDPPercChange	AvgCountryGDPPercChange	AvgTemperature	TotalEmissions
ALB	2018	Yes	3063	0.33	1.934305	11104	3.76	3.4225	59.8893150684932	9.22
ALB	2017	Yes	3053	0.3	1.934305	10702	3.48	3.4225	60.3846575342466	9.42
ALB	2016	Yes	3044	0.3	1.934305	10342	3.09	3.4225	58.9868852459017	8.85
ALB	2015	Yes	3035	0.3	1.934305	10032	2.28	3.4225	60.0901639344263	8.38
ALB	2014	No	3026	0.3	1.934305	9808	1.53	3.4225	60.2156164383562	8.52
ALB	2013	No	3017	0.3	1.934305	9660	0.71	3.4225	60.4131506849315	8.09
ALB	2012	Yes	3008	0.27	1.934305	9592	1.14	3.4225	59.8844262295082	8.04
ALB	2011	No	3000	0.27	1.934305	9484	2.83	3.4225	59.9186301369863	8.49
ALB	2010	Yes	2992	0.13	1.934305	9223	4.11	3.4225	60.5687671232876	7.78
ALB	2009	Yes	2988	-0.07	1.934305	8859	3.95	3.4225	59.4010958904109	7.52

Visualization

Tableau was employed to analyze the Natural Disaster data warehouse and answer the business questions with visualizations. The following dashboards detail the results of their corresponding business questions. See the attached Tableau workbook for a dynamic interaction with the visualizations.

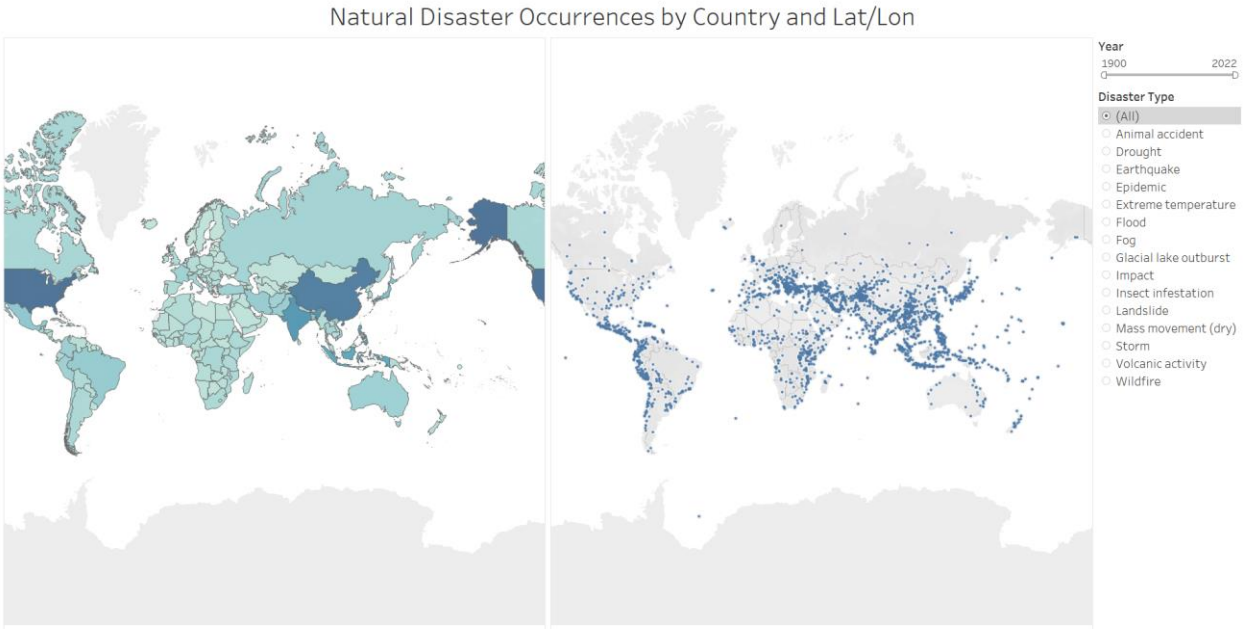
1. *What is the average magnitude and duration by natural disaster type? What were the natural disaster events with the highest magnitude and longest duration?*

The following dashboard highlights the magnitude and duration of natural disaster events. The table summarizes the average magnitude value and average duration by disaster type. Note the different magnitude scales for each natural disaster type. The bar charts rank the natural disaster event with the highest magnitude value and longest duration by natural disaster type.



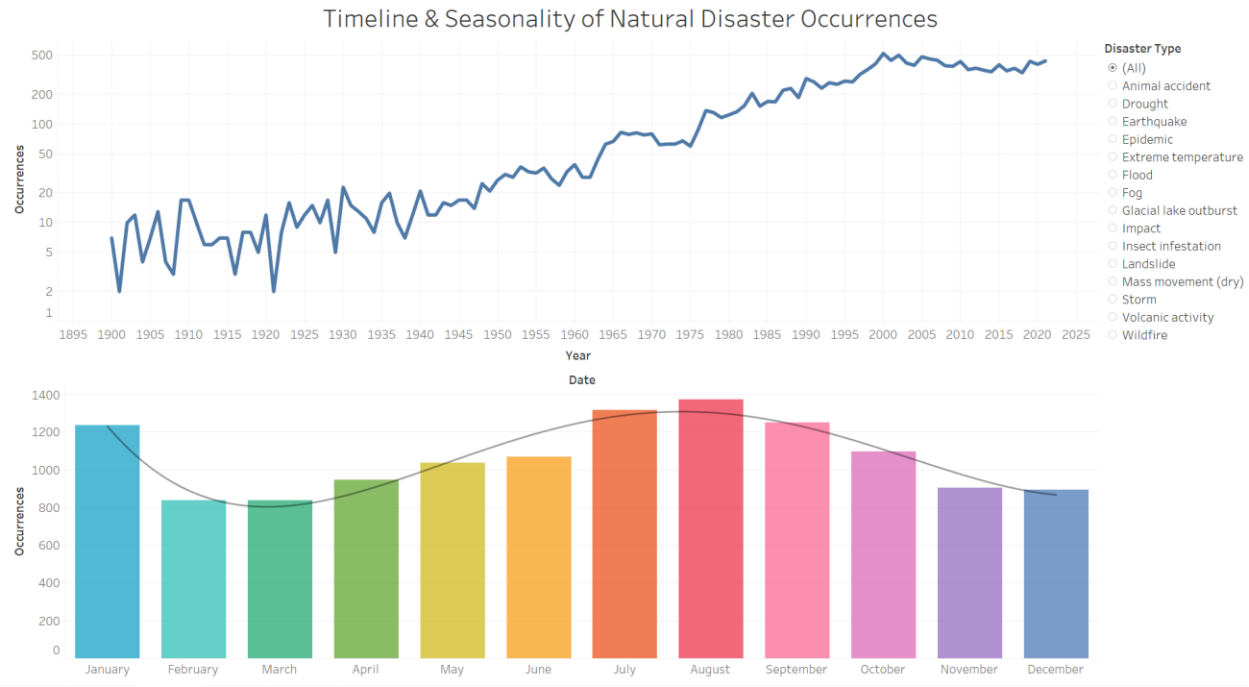
2. *What locations are most prone to natural disaster occurrences?*

The following dashboard visualizes the location of natural disaster occurrences by country (left) and latitude/longitude (right). Filtering options are available for year and disaster type. Notice the countries with a high number of natural disaster occurrences such as the United States and China.



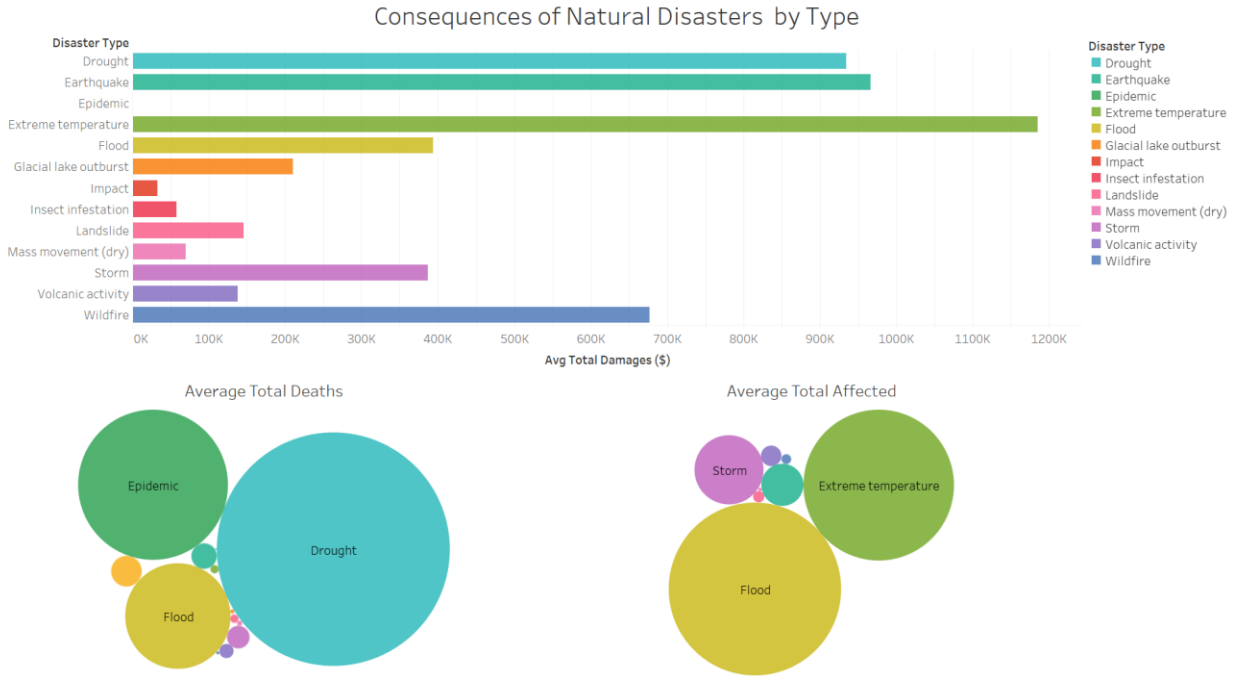
3. *How have the number of natural disaster occurrences evolved over time? Is there a seasonality for particular types of natural disasters?*

   The following dashboard seeks to understand the time dimension of natural disaster occurrences on different grains. The timeline reveals a rising number of natural disaster occurrences over the last 122 years. The bar chart illustrates a trend of seasonality of natural disaster occurrences with peaks in July through September. The seasonality becomes even more distinct when filtering on the natural disaster type.



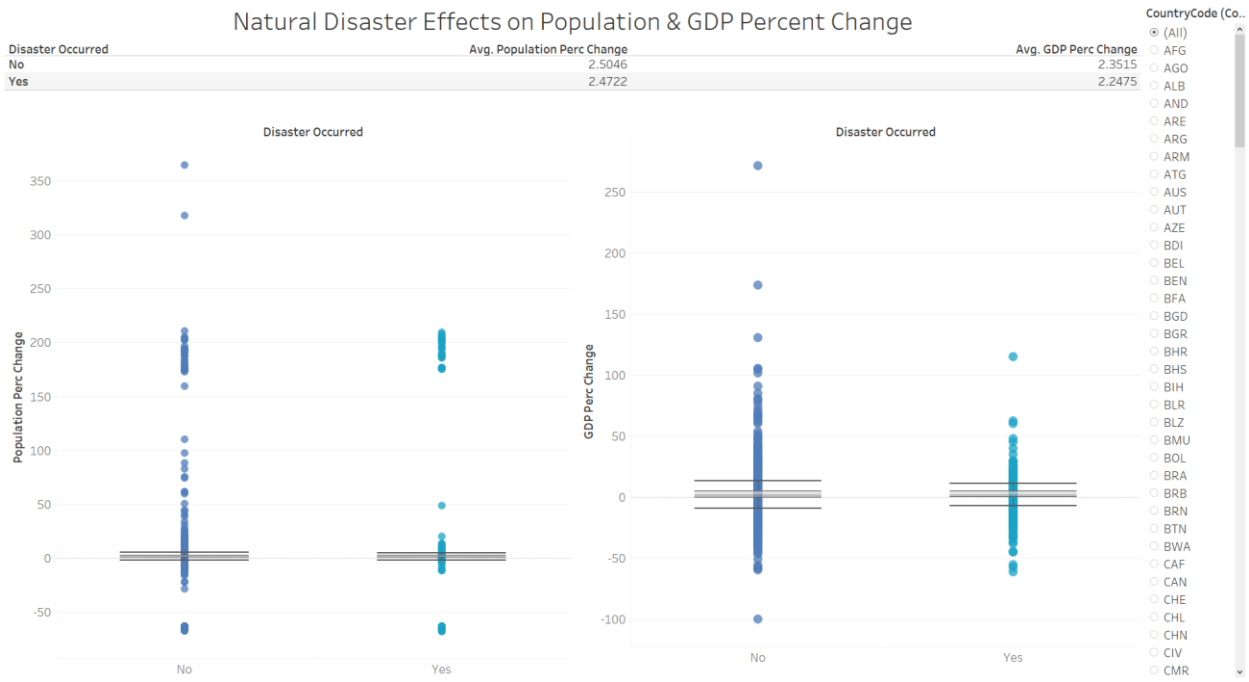
4. *What is the average damage by natural disaster type? Which natural disaster type has the highest number of average deaths and people affected?*

   The following dashboard exams the consequences of natural disasters, including the total damages, total deaths, and total number of people affected. The bar chart visualizes the average total damages by natural disaster type; on average, extreme temperatures is the most expensive natural disaster in terms of damages with $1.2 million damages on average. The bubble charts visualize the natural disaster types that result in the highest number of deaths and highest number of people affected, on average; while droughts result in the highest number of deaths, floods result in the highest number of people affected.



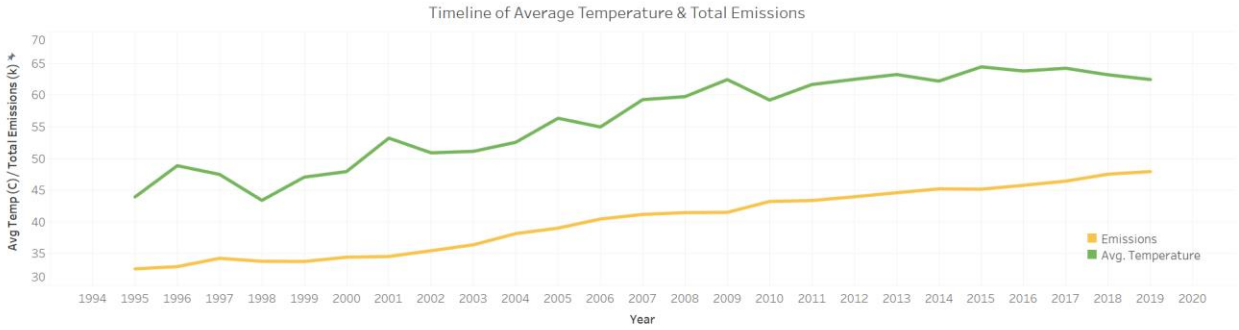
5. *Is there a difference in a country's population/GDP on years where a natural disaster did or did not occur?*

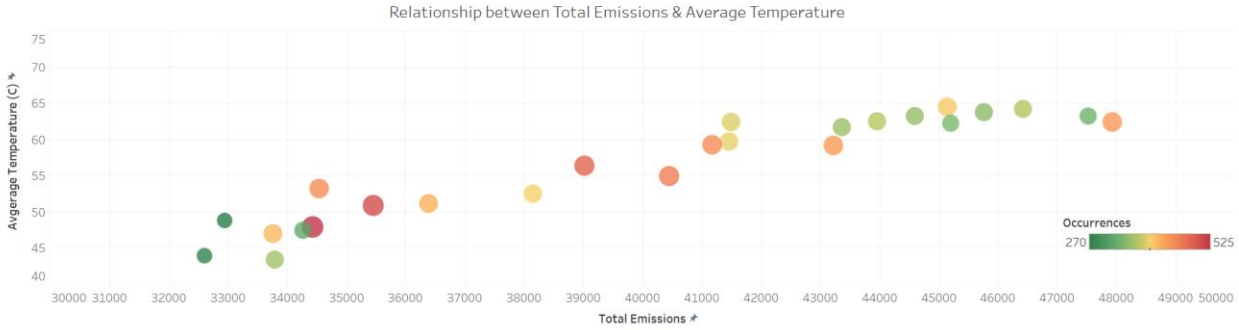
   The following dashboard seeks to understand if there is a disparity between the growth of a country's population and GDP on years when a natural disaster occurred. On average, the population and GDP percent change is approximately 0.1 percentage point higher on years where a natural disaster did not occur. The box plots examine the distribution of the population and GDP percent change stratified by years in which a natural disaster did or did not occur. Filtering options by country are available.



6. *What is the relationship between temperatures and greenhouse gas emissions and what effects do they have on natural disaster occurrences?*

   The following dashboard examines the association between climate and natural disaster occurrences. First, the line chart plots the timeline of average temperature (in Celsius) and total emissions (is thousands) . Notice the steady rise in both average temperature and total emissions over the last 25 years. Next, the bubble chart examine the relationship between total emissions and average temperature (strong positive correlation) in conjunction with the number of natural disaster occurrences. We are seeing an increase in natural disaster occurrences as both total emission and average temperature increase.





Conclusion

The goal of this project was to design and implement a dimensional data warehouse to capture the historical nature of natural disasters and their associations/impacts on climate, weather, and economic indicators. By determining the project scope with business questions, carefully designing a constellation schema, and iteratively performing the ETL process I was able to implement a functional data warehouse ready for analysis. The visualization section uncovers the successes and potential improvements of the data warehouse all related back to the initial step of defining the purpose of the data warehouse.

This project was a practical and challenging experience implementing a full stack data warehouse. The ETL process was by far the most difficult and iterative aspect of the implementation process given the variety of data sources and practicality of slowly changing dimensions. I enjoyed working with both Python and SQL to implement the data warehouse as well as learning Tableau to visualize the answers to the business questions.

Sources

Natural Disasters dataset: https://public.emdat.be/

Temperature dataset: https://www.kaggle.com/datasets/subhamjain/temperature-of-all-countries-19952020

Greenhouse Gas Emissions dataset: https://www.climatewatchdata.org/data-explorer/

GDP dataset: https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2020

References

Climate.gov: https://www.climate.gov/disasters2020