CS306 Group Project – Group 17 – Natural Disasters

Group Members:

Ege Özgür (30609) / Arda Berk Çetin (31077) / Arman Gökalp (29398) / Doğaç Görgülü (28395) / Zeynep Özgür Gün (29502)

Link to the Github repository: https://github.com/dogacgorgulu/CS306-Group17

Project Description:

In this project, we chose to study the natural disasters dataset, and analyze the various attributes such as death rate, injured rate, economic damage etc. to draw some conclusions and gain a better understanding. The dataset includes all the countries and what damage they got from each disaster, so we can study them locally as well. We aim to see how the damages have changed throughout the years and interpret the data as the project goes on, so that the vulnerability of the countries or their development over the years could be seen clearly and it would be easier to come up with solutions.

Data Cleaning and Tables:

We have five main tables, named “countries-table”, “disasters_table”, “deaths”, “injuries” and “economic_damage”.

1) “countries-table.csv” file includes “country name”, “iso-code”, “year” and “population” attributes. “country_names” represents the name of the countries, “iso_codes” represents iso codes of the countries,”year” represents the years starting from 1950 to 2020 to declare the populations of the country according to specific years. “Population” shows the population of the country in the selected year between 1950 and 2020. The data is collected from “ourworldindata.org/natural-disasters” and “https://ourworldindata.org/world-population-growth”. We merged the countries from natural disasters data and population information from world population growth data. Raw data is uploaded to the Excel file. Data types are determined. Related data are copied to another Excel sheet. With Excel’s built-in functions, duplicates are removed and we get the clean data we wanted to work on.

2) “disasters_table” table has “Country - Year” as the key attribute, representing a certain country and year, in which some disasters has happened. Other attributes holds the

number of people affected by those disasters. These attributes are “Number of people affected from drought”, “Number of people affected by earthquakes”, “Number of people affected by volcanic activity”, “Number of people affected by floods”, “Number of people affected by storms”, “Number of people affected by landslides”, “Number of people affected by wildfires”, “Number of people affected by extreme temperatures”, and finally, “Number of total people affected by disasters”. First, we removed the years before 1950 since there were lots of missing data using the sort & filter option in excel. Later on, again using the filter option, we removed data from contries with large missing values as well as continent/organization data since we are focusing on countries. Finally, we merged the year and country columns to create a key attribute to easily access to all the disasters which occured in a certain year and country. The table can be found on our GitHub repository.

3) “deaths.csv” file includes “ISO Code”, which is the key attribute, “Country name”, ”Year”, “Number of deaths from drought”, “Number of deaths from earthquakes”, “Number of deaths from volcanic activity”, “Number of deaths from floods”, “Number of deaths from storms”, “Number of deaths from landslides”, “Number of deaths from wildfires”, “Number of deaths from extreme temperatures” attributes. Raw data set retrieved from “ourworldindata.org/natural-disasters”. Raw data is uploaded to the Excel file. Data types are determined manually. Related data are copied to another Excel sheet. With Excel’s built-in functions, duplicates are removed, unrelated disasters (includes 8 of the disasters) and years (includes only between 1950-2020) are filtered out. Lastly, excel file saved in “.csv” format and uploaded to GitHub repository.

4) “injuries.csv” table includes the number of people injured from droughts, earthquakes, volcanic activities, floods, storms, landslides, wildfires, and extreme temperatures as attributes. They are each represented by a separate column. The raw dataset is retrieved from the same source as my team members: ourworldindata.org/natural-disasters. This raw data had to be cleaned. This was done through filtering the continents and ex-countries from the countries column using the inbuilt filtering mechanism of Excel. Same process was applied to years in order to delete any date earlier than 1950. This was done to sync our injury data with population data that my other team members are working with. Afterwards, any column that was not about injury statistics was deleted. Lastly, ISO codes were added to distinguish countries. The csv file can be found on the Github repository.

5) “economic_damages.csv” file includes the following attributes: “country_names”, “iso_code”, “year”, “total_reconstruction_costs”, “total_insured_damages” and “total_economic_damages”, with the last three attributes representing the total costs

from all natural disasters. Raw data set retrieved from “ourworldindata.org/natural-disasters”. To create this table, countries and years were taken from the original dataset and iso codes of each country were added. Also, the countries which no longer exist were removed. Each damage was categorized in the original dataset, such as reconstruction costs from drought, reconstruction costs from earthquakes etc. but in this table, they are summed under the same column to represent the total damages from all natural disasters, making it easier to see the total costs. The file can be found in the GitHub repository.

ER Diagram: