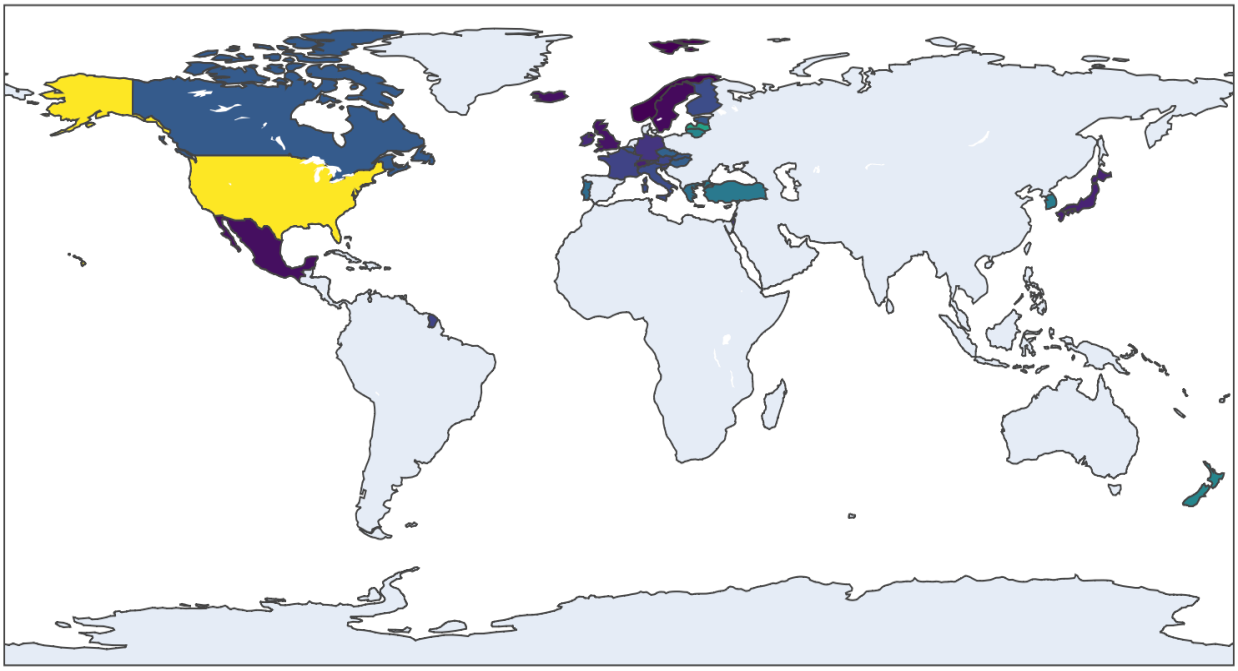

COURSE PROJECT: BIG DATA CONCEPT AND IMPLEMENTATIONS

Road Accident Death Rates in OECD Countries

Zeynep Elabiad

I535-MANAGEMENT, ACCESS,
AND USE OF BIG AND COMPLEX DATA

Spring 2023

INTRODUCTION

Road accidents are a significant and fatal concern across many countries globally. According to the World Health Organization (WHO), road accidents result in approximately 1.3 million deaths and 20-50 million injuries annually. Moreover, road accidents cause property damage, medical expenses, emotional trauma, and other long-lasting damages. Understanding the elements that lead to road accidents and distinguishing effective mitigation strategies is crucial. This project explores the relationship between alcohol consumption, road accident death rates, and the Human Development Index (HDI) in OECD countries. The analysis will focus on the latest available data and emphasize country comparisons, utilizing Big Data Tools in the Google Cloud Platform.

BACKGROUND

Road accidents are a major problem affecting many countries. Because of the complex nature of traffic accidents, numerous factors can impact the rates of fatalities resulting from accidents. By comprehending the underlying causes of road accidents, we can distinguish effective actions to reduce their occurrence and severity.

The objective of the project is to gain insights into the factors responsible for road accidents caused by alcohol consumption and DUI in OECD countries, as well as the impact of HDI on the death rate of road accidents while exploring the environment of a Google Cloud Platform. Investigating the relationship between alcohol consumption, road accident death rates, and HDI and understanding of these factors can provide valuable insights for policymakers to create effective policies and strategies to mitigate road accidents and enhance public safety. By using these insights, policymakers can develop targeted interventions that address the underlying causes of road accidents, resulting in safer roads and healthier communities. Furthermore, a limited amount of literature is available on the connection between alcohol consumption, road accident death rates, and HDI in OECD countries, which encouraged me to select this topic for the project.

DATA & METHODOLOGY

*The Organization for Economic Co-operation and Development (OECD***)** is a significant global organization that collaborates with governments to identify solutions to common challenges and exchange knowledge for improved quality of life. Additionally, the organization provides various resources, including databases, statistics, maps, educational outputs, publications, and visualizations, to individuals interested in multiple topics. For my project, I utilized three datasets made available by the OECD: Alcohol Consumption[1], Road Accident Death Rates [2], and Road Accidents Involving Casualties[3] datasets. The Alcohol Consumption dataset provides information on the annual sales of pure alcohol in liters per person aged 15 years and older for each country in the OECD. The road accident death rates dataset displays the number of fatalities in road accidents per 100,000 individuals in the population annually for each country in the OECD. Meanwhile, the Road Accidents Involving Casualties dataset showcases the yearly total number of persons injured and deaths due to road accidents for each country in the OECD.

Additionally, I utilized HDI dataset from the United Nations Development Program (UNDP) website. The UNDP compiles the Human Development Index data that comes from United Nations Agencies. *The Human Development Index (HDI)* is an average index across key indicators in human development (Health, Education, and Gross National Per Capita Income).[4] According to the United Nations Development Program, this index is a good way to compare countries with similar incomes but different human development outcomes and can be helpful in making policies. The HDI ratio values range from 0 to 1; the higher the score, the better the result.

Due to data limitations and missing values, I narrowed down the scope of my project and focused on countries that are members of the OECD. Currently, there are 38 members of the OECD, and I included 30 of these 38 the OECD countries' data between 2010-2020 in the project. The countries included in my study are Belgium, Canada, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Latvia, Lithuania, Mexico, New Zealand, Norway, Portugal, Slovakia, Slovenia, South Korea, Spain, Sweden, Switzerland, Turkiye, the United Kingdom, and the United States of America.

Using Jupyter Notebook and the Python programming language, I prepared a Python file to clean, merge, and transform the downloaded datasets. Afterward, I uploaded these datasets that were mentioned above and the Python file to the GitHub repository[5] to serve as the **data ingestion step** in the Pipeline process.

---

[1] OECD (2023), Alcohol consumption (indicator). doi: 10.1787/e6895909-en (Accessed on 2 April 2023)
[2] OECD (2023), Road accidents (indicator). doi: 10.1787/2fe1b899-en (Accessed on 2 April 2023)
[3] ITF (2023), "Road accidents", ITF Transport Statistics (database), https://doi.org/10.1787/g2g55585-en (accessed on 2 April 2023).
[4] https://hdr.undp.org/data-center/human-development-index#/indicies/HDI
[5] https://github.com/Zeynepelabiad/road-accidents

Data Lifecycle and Pipeline on Google Cloud Platform

Google Cloud Platform (GCP) is a cloud computing platform by Google that offers public cloud infrastructure and platform services, such as computing, machine learning, storage, data analytics learning, and more. The data lifecycle management on GCP involves using its services to manage and process data at every stage of the lifecycle. Users can create a data pipeline with tools offered by the Google Cloud Platform to manage different data lifecycle stages. For my project, I created a manual data pipeline using GCP services for data lifecycle stages.

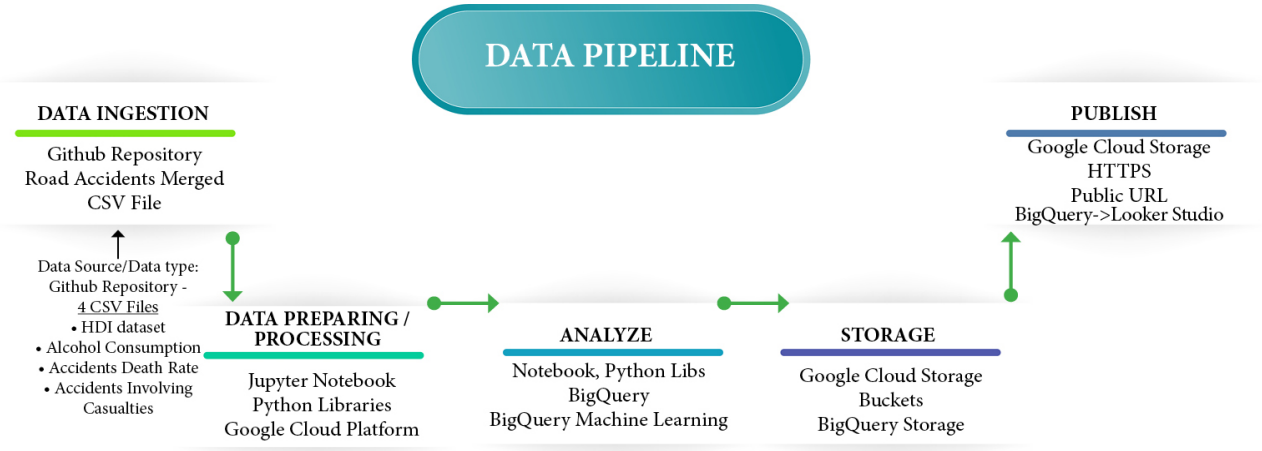

Figure 1. Manual Data Pipeline on Google Cloud Platform

Before ingesting my data, I created a Compute Engine WM instance on GCP, chose the boot disk version as Debian GNU/Linux 10, and gave full access to all Cloud APIs. Next, I remotely accessed the Compute Engine instance using Secure Shell (SSH) and installed the GitHub package. Then I cloned the GitHub repository link that has uploaded datasets and Python Transformation file to the instance folder called “Road Accidents” using the SSH window.

In order to transform the data, I navigated to the Road Accidents instance folder and executed the Python transformation code in the SSH window. I installed several necessary packages, including pandas, seaborn, matplotlib, CSV, and requests, to facilitate **data processing** and transformation using the SSH window. The Python transformation file imported Python libraries, cleaned, and selected 30 out of 38 OECD countries with data spanning from 2010 to 2020, converted data types, merged all the datasets, and exported the resulting CSV file and a png file.

Google Cloud Platform provides various storage options to **store data**, such as Cloud Storage, Big Query, Cloud Data Store, etc. I used Cloud Storage, which stores objects as Buckets for my project. I changed the bucket's location from Multi-Regional to Regional to reduce costs and enhance speed while creating it. I also copied the merged CSV dataset – oecd_df.csv and the png file into the road_accidents folder of the bucket I created in Cloud Storage.

We can **publish**/make accessible the bucket files in the Cloud Storage and change/edit access permissions. I published the TRAFFIC_ACCIDENTS_DEATH_RATE.png on the public URL. The PNG file displays bar plots that represent the road accident death rate by OECD member countries.

According to the graph, the USA, Latvia, and Korea are the top three OECD countries with the highest death rates resulting from road accidents.

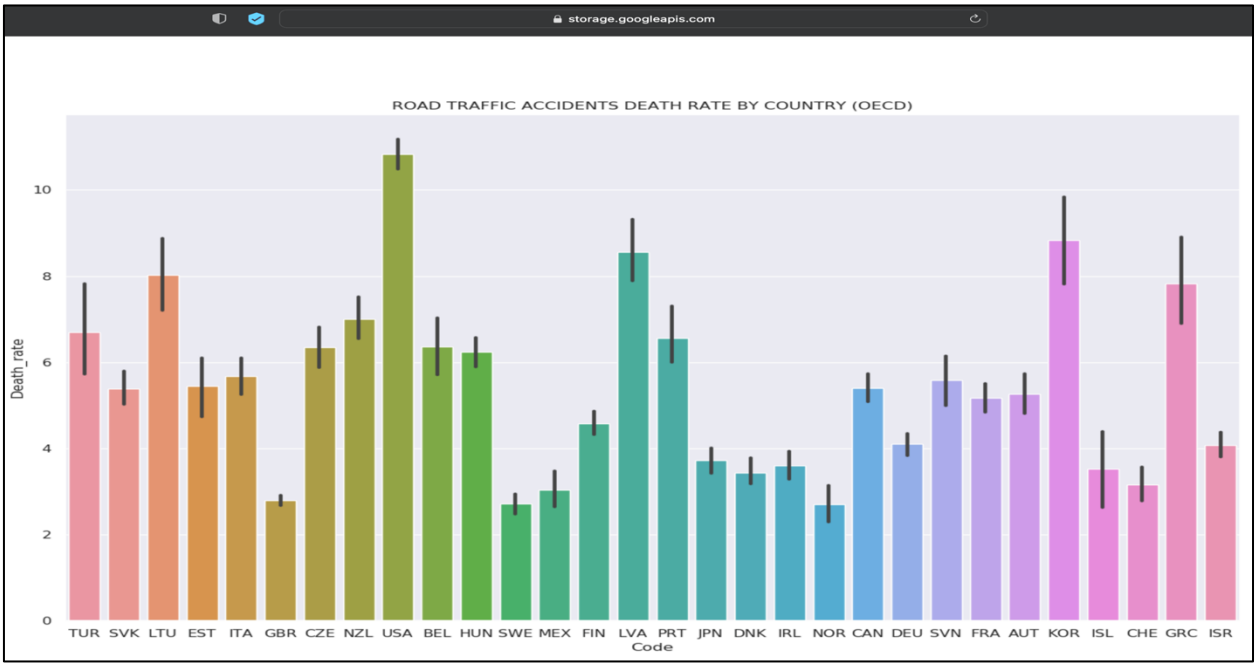

Figure 2. TRAFFIC_ACCIDENTS_DEATH_RATE. png Barplot on the Public URL

BigQuery

BigQuery is a highly scalable, fully managed, serverless data warehouse that runs on the Google Cloud Platform. It has a query engine allows us to run SQL queries on large amounts of data without any maintenance, even data not stored in BigQuery. It can ingest real-time and batch data, perform large-scale machine learning for **data analysis**, and store data in BigQuery. Google Cloud Storage also allows users to export data in buckets directly to BigQuery on the Google Cloud Platform. This means that data stored in Cloud Storage can be analyzed and processed by BigQuery without requiring extra data transfer or manipulation steps. I used the BigQuery web UI to import the transformed dataset - oecd_df from the Cloud Storage Bucket to the BigQuery table to analyze the dataset.

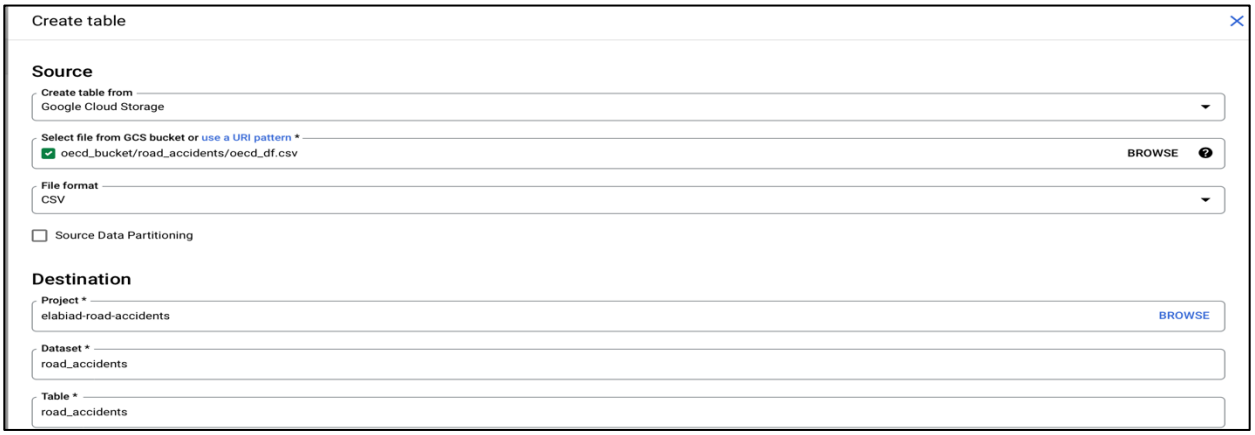

Figure 3. Creating road_accidents table from Google Cloud Storage Bucket on BigQuery

RESULTS

BigQuery offers a variety of tools for analyzing data, such as SQL-like queries, machine learning algorithms, and data visualization. Using BigQuery, we can explore the relationships between different variables in the oecd_df dataset by calculating correlations between the features. Correlations can be used to identify patterns and make predictions. To compute the correlation between features in our dataset, we can use the CORR() function in BigQuery. The function produces a value between -1 and 1, where -1 represents a strong negative correlation, 0 indicates no correlation, and 1 represents a strong positive correlation. By calculating the correlation between alcohol consumption and road accident death rates or calculating the correlation between HDI and road accident death rates in several OECD countries, we can obtain valuable insights regarding the strength of their relationship.

Correlations

Upon analyzing the **correlation between road accident death rates and HDI** in 30 OECD member countries, I have observed a significant negative correlation between the two variables. This indicates that countries with higher HDI tend to have lower rates of traffic accident deaths, providing valuable insight into the relationship between these factors.

Top 5 Countries that have the highest correlation between road accident death rates and HDI:

OECD Country	Correlation
South Korea	-0.975
Switzerland	-0.961
Lithuania	-0.930
Canada	-0.91
Norway	-0.903

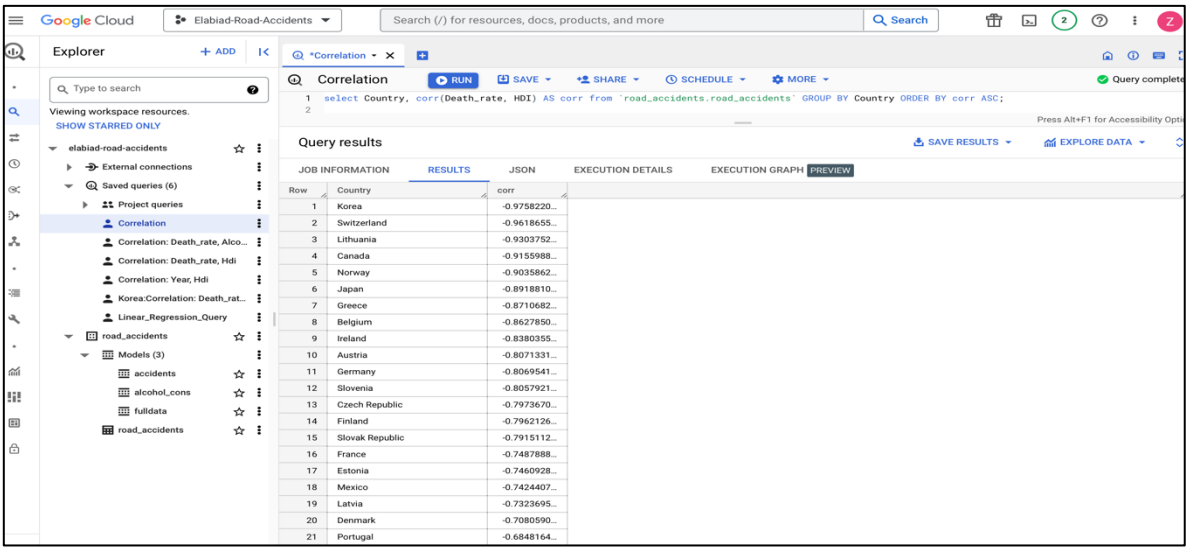

Figure 4. Correlation Between HDI and Road Accident Death Rates by OECD Country

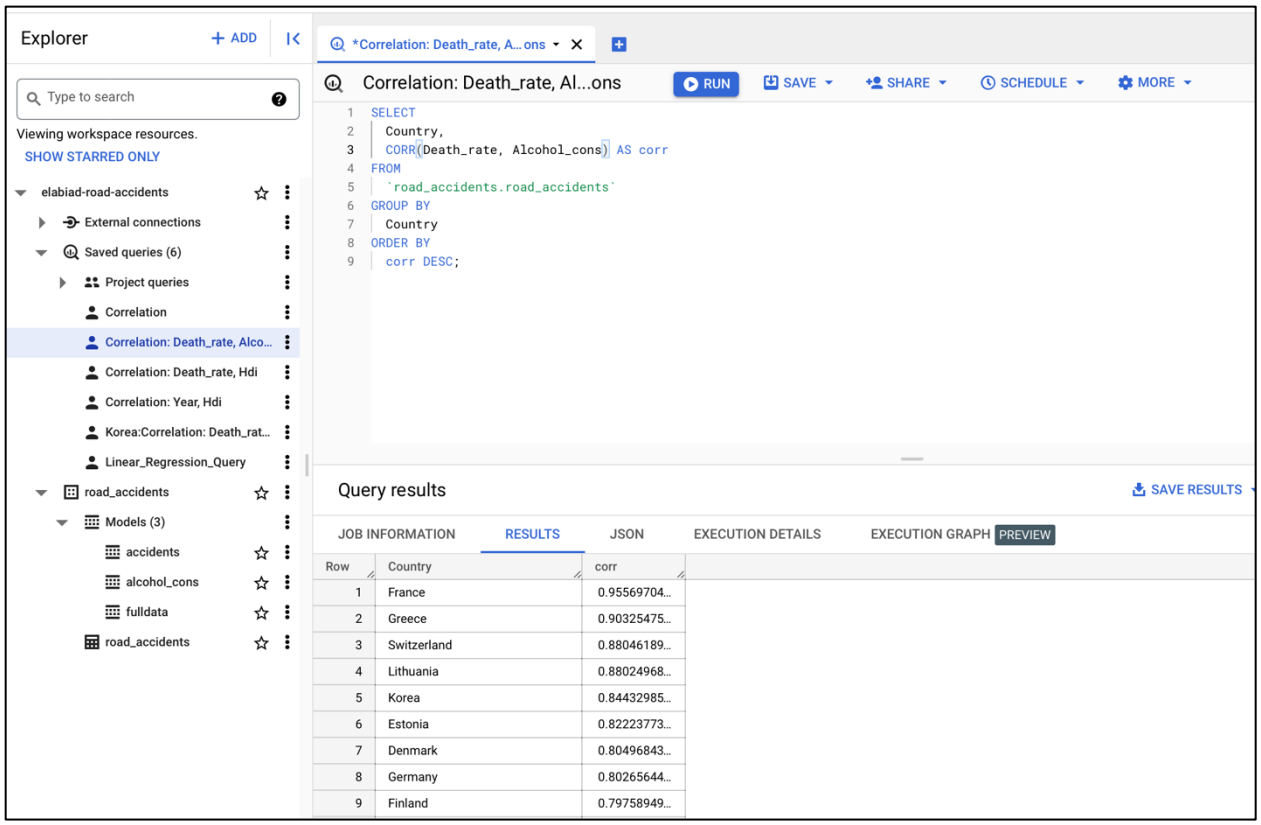

Figure 5. Correlation Between Alcohol Consumption and Road Accident Death Rates

After examining the **correlation between alcohol consumption and road accident death rates** in 30 OECD member countries, I found a moderate positive correlation between the two variables. Countries with higher levels of alcohol consumption tend to experience more fatal accidents of driving under the influence (DUI). In addition to alcohol consumption, many other factors may contribute to road accident death rates, such as weather conditions, road infrastructure, driver stress levels, vehicle safety, etc.

Top 5 Countries that have the highest correlation between road accident death rates and alcohol consumption:

OECD Country	**Correlation**
France	0.955
Greece	0.903
Switzerland	0.880
Lithuania	0.880
Korea	0.844

BigQuery ML

BigQuery Machine learning allows us to analyze and make predictions using machine learning algorithms on the dataset, such as linear regression and logistic regression. I utilized a linear regression algorithm in BigQuery to create a model for predicting road accident death rates using HDI, and alcohol consumption features. The model achieved an R-squared value of 0.83 in predicting OECD countries' road accident death rate.

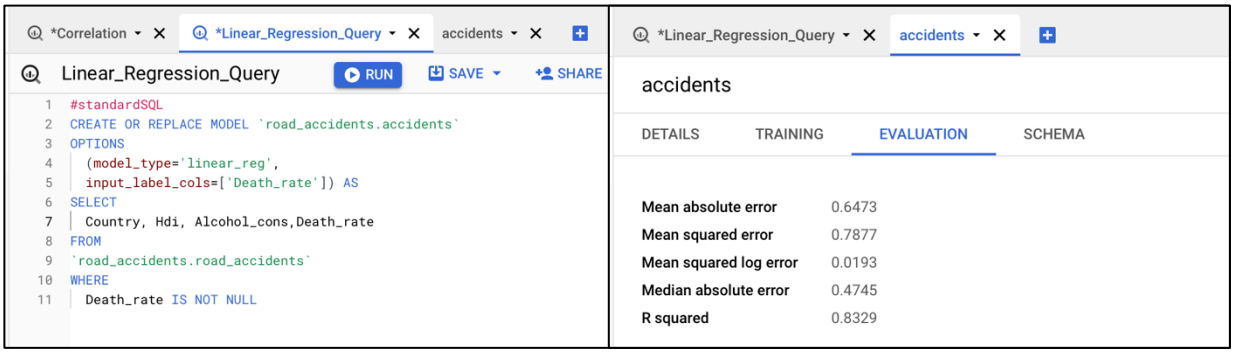

Figure 6. Application of Linear Regression Algorithm on the Road Accident Dataset

BigQuery also offers an integrated powerful visualization platform called BigQuery Looker Studio. It allows users to create custom visualizations or use pre-built visualizations to gain insights into users' data. We also can create reports, interactive visualizations, and dashboards to explore data. I used BigQuery Looker Studio to create interactive visualizations to understand historical data better and identify patterns and trends.

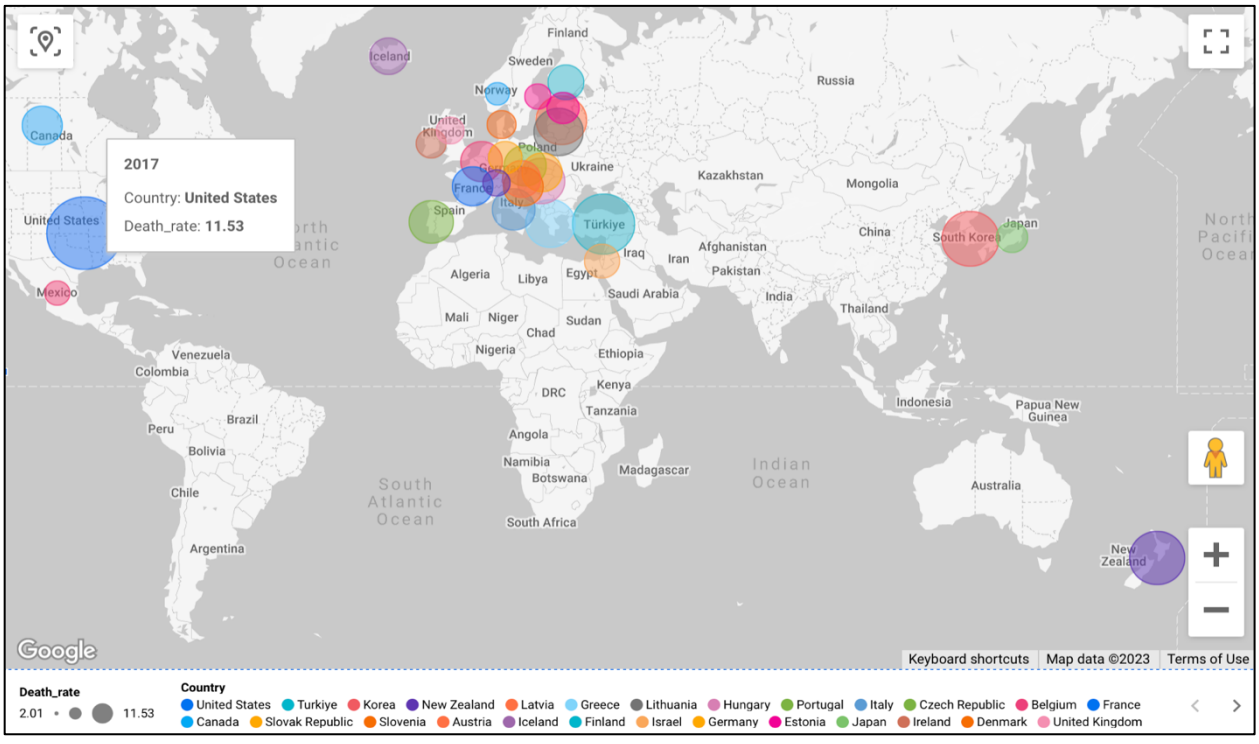

Figure 7. A Bubble Map of Road Accident Death Rates in 2017

Discussion

Road accidents significantly impact individuals, families, and society, resulting in substantial economic costs. Given the complexity of traffic accidents, a range of factors can affect the rate of fatalities resulting from them. Therefore, it is crucial to comprehend the underlying causes of road accidents to identify effective measures for reducing their frequency and severity. Alcohol consumption is widely recognized as a significant risk factor for road accidents, and its impact on road safety is a subject of ongoing research. I observed a moderate positive correlation between alcohol consumption and road accident death rates, indicating that increased alcohol consumption led to higher road accident deaths. So, it is essential for officials to address this issue by implementing strict DUI laws and discouraging extreme alcohol consumption, education starting from primary school on the dangers of drunk driving, and road safety rules. Policymakers and officials can promote road safety for everyone and reduce road accident death rates by addressing the causes of the problem. Another factor that may impact road accident death rates is HDI. HDI measures a country's human development, income, education, and healthcare level. I have observed a strong negative correlation between road accident death rates and HDI. That means the OECD countries with lower HDI tend to have higher rates of traffic accident deaths. Countries with lower levels of HDI may lack road infrastructure, limited access to healthcare and education, and less knowledge of road safety. Analyzing the correlation between HDI and road accident death rates can help policymakers to identify areas for improving education, government policies, social programs, health care, and road safety.

To accomplish the project goals, I chose the Google Cloud Platform, which is a user-friendly and great service for implementing data lifecycle stages and pipelines, even for managing small-scale data in various formats. I applied the knowledge I gained from the course's “Data Lifecycle and Pipelines” module to create a manual pipeline. Additionally, the “Ingesting and Store” module helped me to ingest the dataset into BigQuery and store the data quickly and efficiently. I employed various Python libraries using SSH on a Virtual Machine Instance of the Google Cloud Platform to process and clean the data. For analysis, I utilized the powerful BigQuery and BigQuery Machine Learning. Furthermore, I stored the data in Google Cloud Storage and BigQuery table, which provided scalability and reliability. To create informed visualizations, I used BigQuery Looker Studio's capabilities on the Google Cloud Platform.

During the project, I encountered several challenges. Firstly, I faced difficulty installing the Plotly Express library in the SSH window of the VM instance due to my limited experience with cloud computing platforms, and I had to revise my transformation Python codes. Furthermore, while I initially added OECD country road infrastructure investment as another feature to the dataset, I later had to remove it since most of the correlations between OECD countries and road accident death rates were either strongly positive or strongly negative, indicating the need for a different study to identify the causes for each country.

The GCP has a comprehensive range of features, including storage, analysis, machine learning, and visualization capabilities which guided me to select it for my project. If I hadn't used GCP and had worked with the data locally, I would have had to use a powerful computer equipped with software tools like Tableau, Python IDE, SQL databases, etc., to handle the data and apply machine learning algorithms and visualization. It could have resulted in significant limitations and the need to skip important steps in the project due to the software and expertise required. Overall, the Google Cloud Platform's broad features enabled me to accomplish my project objectives efficiently. The experience showed me valuable data management and analysis skills in a cloud computing environment.

Conclusion

With this project, I studied the relationship between alcohol consumption, HDI, and road accident death rates. The results showed that alcohol consumption positively correlates to higher death rates in road accidents. It also showed that the higher a country's HDI score, the lower its accident death rate is. Given these results, I would conclude that we need a holistic approach that focuses on solving road safety injuries and fatalities. It is not enough to just focus on singular issues such as alcohol consumption and speed. This can be deduced from observing the negative correlation between road accidents and HDI. This means that countries must invest in sectors such as education, health, infrastructure, etc., to address the issues. Understanding these issues can provide valuable insights for policymakers in making efficient policies and plans to reduce road accidents and improve community safety.

REFERENCES:

- OECD (2023), Alcohol consumption (indicator). doi: 10.1787/e6895909-en (Accessed on 2 April 2023)
- OECD (2023), Road accidents (indicator). doi: 10.1787/2fe1b899-en (Accessed on 2 April 2023)
- ITF (2023), "Road accidents", ITF Transport Statistics (database), https://doi.org/10.1787/g2g55585-en (accessed on 2 April 2023)
- United Nations Development Programme (UNDP). Human Development Index (HDI) https://hdr.undp.org/data-center/human-development-index#/indicies/HDI
- The Investopedia Team. Human Development Index (HDI) https://www.investopedia.com/terms/h/human-development-index-hdi.asp
- OECD: https://en.wikipedia.org/wiki/OECD
- Course Materials (INFO-1 535: MANAGEMENT, ACCESS, AND USE OF BIG AND COMPLEX DATA **)**
- Ingesting New Datasets into BigQuery https://www.cloudskillsboost.google
- Rent-a-VM to Process Earthquake Data https://www.cloudskillsboost.google