2021 South Africa Deceased Data Analytics Project

Objective

The aim of this project is to analyse deceased data from South Africa (2021) to uncover meaningful insights about mortality patterns across the country. Specifically, the project seeks to:

- Identify the most common causes of death.
- Determine which provinces record the highest number of deceased cases.
- Explore how causes of death vary by population group and gender.
- Compare the average age of death across genders.
- Assess whether the type of industry in which a person worked contributes to the cause of death.
- Investigate the location of deaths (hospital, home, or other settings).

This project is designed to transform raw data into actionable insights that can inform policymakers, healthcare professionals, and researchers.

Tools and Technologies

- **Python**: Used to preprocess and restructure the dataset. The original dataset contained ~270,000 rows, which was reduced to 50,000 rows for performance considerations. Python was also used to merge scattered data frames into a clean, revised dataset.
- **SQL**: Utilized to aggregate and model the data, enabling structured queries and optimized analysis for visualization.
- **Power BI**: Employed for additional data cleaning (e.g., handling missing values) and the creation of interactive visualizations and dashboards.
- **ChatGPT**: Assisted with formalizing project documentation, generating dashboard themes, and clarifying dataset terminology.
- **GitHub**: Serves as the version control and project hosting platform, making the project publicly accessible.

Analytical Questions

This project will address the following key questions:

1. What are the leading causes of death in South Africa?
2. How do causes of death differ between males and females?
3. Which provinces report the highest number of deaths?
4. Does a person's industry of employment contribute to their cause of death?
5. What is the average age of death, and how does it vary by gender?
6. Where do most deaths occur (e.g., hospitals, homes, other locations)?
7. How do causes of death differ across population groups, and what are the average death rates for each group?

Deliverables

- A cleaned dataset of 50,000 records suitable for analysis.
- An SQL data model optimized for queries and reporting.
- A Power BI dashboard providing visual insights into mortality patterns.
- A GitHub repository documenting the workflow, methodology, and findings.