

Final Project Report

Group members:

Vikramjeet Singh Kundu (vkundu2)

Samarth Jain (sjain224)

Siddhant Evre (sevre2)

Project Summary: Life Expectancy and World Indicators Analysis

Objective:

The aim of this project was to design and implement a data warehouse and business intelligence solution for the Life Expectancy (WHO) dataset.

We have picked this dataset from Kaggle which is essentially a platform for data scientists and machine learning practitioners and has a lot of datasets to choose from.

Link to the dataset: <https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated>

Couple of important points to note for our dataset:

- It contained important information for 179 countries from 2000-2015. The information consisted of demographic, economic and health information along with Life Expectancy.
- It consists of 21 variables along with 2864 rows.

Project Objectives and Methodology:

1. Designing Data Warehouse Schema: Design a robust data warehouse schema that aligns with business intelligence goals.

For designing our data warehouse schema, we used Lucid Chart which is a very convenient web-based diagramming solution and can be used to make a comprehensive database schema quickly and with ease.

We made one fact table along with two-dimension tables as it can be seen below.

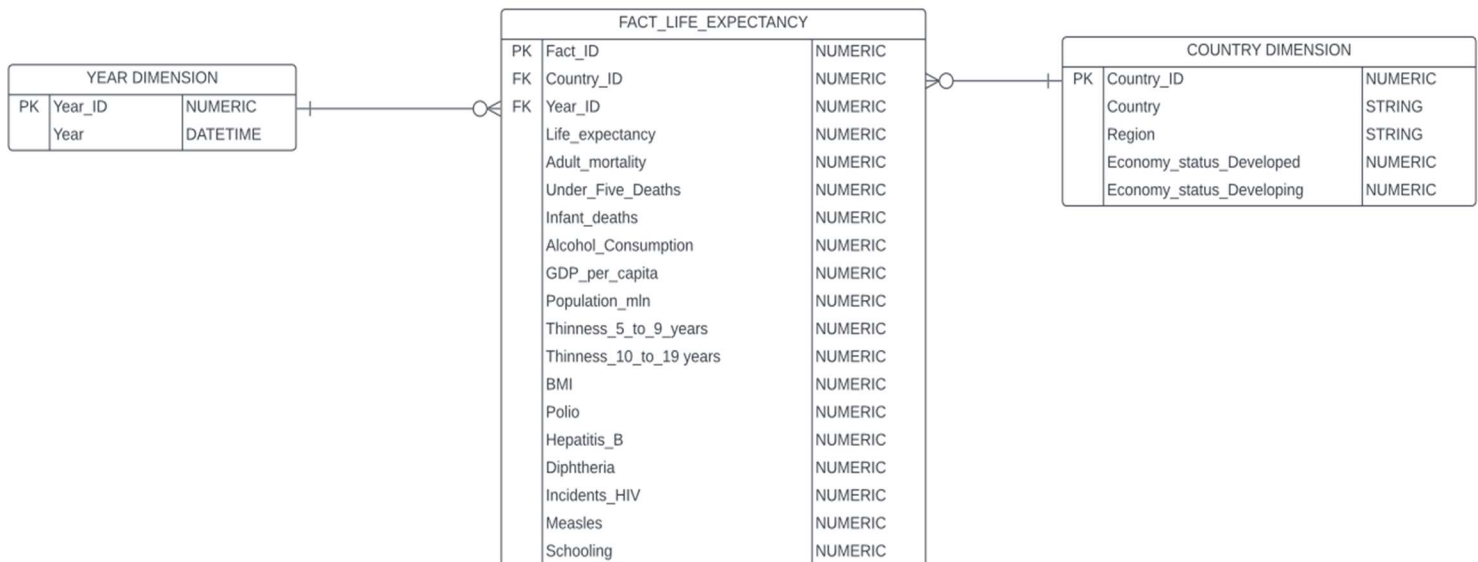
Primary key for the fact table was the Fact_ID and Country_ID along with Year_ID were foreign keys within the fact table which were referenced in a separate Country and Year dimension respectively.

The granularity of the fact table is the life expectancy data per country per year. The schema essentially employs normalization to reduce data redundancy and the dimensions are separated from the fact table in order to allow for more efficient querying. This schema allows us to analyze the life expectancy and health indicators as well as trends and correlations between a variety of factors across multiple countries and years.

Database schema along with a link to access it on Lucid Chart.

https://lucid.app/lucidchart/545c892a-3bce-4603-a0c0-0074def13efb/edit?viewport_loc=-216%2C-649%2C3391%2C1665%2C0_0&invitationId=inv_3bcca794-4af9-4a7d-a6cc-93fe140cc2d0

Fact Table Grain : Life Expectancy data per year per country



2. ETL Development: Instead of using ETL tools such as Informatica/PowerCenter/Talend Open Studio/AWS Glue to develop an automated ETL process to regularly extract data from the Life Expectancy dataset, we performed ETL in the PyCharm CE IDE using Python scripts. We then loaded the extracted data seamlessly in Tableau for analysis with the help of a cloud-based data warehouse Amazon Redshift and cloud-based storage solution, S3 buckets.

Since our data was static and was not being constantly updated, we didn't feel the need to perform ETL using external tools. Simple Python scripts along with SQL queries sufficed the needs of our particular use case and database schema.

- **Data Extraction**

The initial step involved sourcing a comprehensive dataset containing various health and economic indicators from around the world.

The raw data was provided in CSV format and included metrics on life expectancy, alcohol consumption, and other health-related statistics.

- **Database Creation and Schema Definition:**

An SQLite database, life_expectancy.db, was set up to store and manage the data.

The database schema was designed to include separate tables for country attributes, yearly data, and fact table, as shown above in the schema/flowchart to record life expectancy and other metrics.

- **Data Transformation:**

Python scripts were written to transform the raw data into a structured format, suitable for analysis. Mappings for countries and years were established to convert textual information into numerical identifiers that could be more efficiently stored and queried.

- **Data Loading:**

The transformed data was loaded into the SQLite database, populating the Country_Dimension, Year_Dimension, and Fact_Life_Expectancy tables.

Validations were performed post-loading to ensure data integrity and to verify that the transformation process had not introduced any errors.

The entire Python script 'DWBI Project ETL.py' along with the database 'life_expectancy.db' that was created have been attached in the files submitted on Canvas.

AWS Configuration and Data Warehousing

- AWS S3 Bucket Configuration:

An AWS S3 bucket was created as a centralized and secure storage solution for the transformed datasets, which were uploaded in preparation for Redshift ingestion.

- IAM Role and Policy Setup:

An IAM role was established with AmazonS3ReadOnlyAccess to authorize the Redshift cluster so that we could access the S3 bucket data securely without exposing broader permissions than necessary.

- Redshift Cluster Initialization:

The Redshift cluster was provisioned with the appropriate node type and quantity to balance performance with cost, considering the expected query load and data volume.

Tables within Redshift were created, mirroring the schema defined in SQLite but adjusted for Redshift's columnar storage and optimized for query performance.



```
+ load-data-fact_life_expectancy-edb1 x load-data-year_dimension-a1b6 x load-data-country_dimension-4a0c x Fact table x 2 dimension tables
▶ Run Limit 100 Explain Isolated session dwbi-project dev
1 CREATE TABLE Fact_Life_Expectancy (
2   Fact_ID INTEGER IDENTITY(1,1) PRIMARY KEY,
3   Country_ID INTEGER REFERENCES Country_Dimension(Country_ID),
4   Year_ID INTEGER REFERENCES Year_Dimension(Year_ID),
5   Life_expectancy REAL,
6   Adult_mortality REAL,
7   Under_five_deaths REAL,
8   Infant_deaths REAL,
9   Alcohol_consumption REAL,
10  GDP_per_capita INTEGER,
11  Population_mln REAL,
12  Thinness_ten_nineteen_years REAL,
13  Thinness_five_nine_Years REAL,
14  BMI REAL,
15  Hepatitis_B INTEGER,
16  Polio INTEGER,
17  Diphtheria INTEGER,
18  Incidents_HIV REAL,
19  Measles INTEGER,
20  Schooling REAL
21 );
```

```
+ load-data-fact_life_expectancy-edb1 x load-data-year_dimension-a1b6 x load-data-country_dimension-4a0c x Fact table x 2 dimension tables x
Run Limit 100 Explain Isolated session dwbi-project dev
1 CREATE TABLE Country_Dimension (
2     Country_ID INTEGER PRIMARY KEY,
3     Country VARCHAR(255) NOT NULL,
4     Region VARCHAR(255) NOT NULL,
5     Economy_status_Developed INTEGER NOT NULL,
6     Economy_status_Developing INTEGER NOT NULL
7 );
8
9 CREATE TABLE Year_Dimension (
10     Year_ID INTEGER PRIMARY KEY,
11     Year INTEGER NOT NULL
12 );
```

- Data Ingestion into Redshift:

The COPY command was used to bulk load the data from the S3 bucket into the corresponding Redshift tables efficiently.

```
+ load-data-fact_life_expectancy-edb1 x load-data-year_dimension-a1b6 x load-data-country_dimension-4a0c x Fact table x 2 dimension tables x
Run Limit 100 Explain Isolated session dwbi-project dev
1 COPY dev.public.fact_life_expectancy (country_id, year_id, life_expectancy, adult_mortality,
2 under_five_deaths, infant_deaths, alcohol_consumption, gdp_per_capita, population_mln,
3 thinness_ten_nineteen_years, thinness_five_nine_years, bmi, hepatitis_b, polio, diphtheria, incidents_hiv, measles, schooling)
4 FROM 's3://dwbi-project-etl-dataset/Fact_table.csv' IAM_ROLE 'arn:aws:iam::394451548552:role/AmazonS3ReadOnlyAccess-DWBI'
5 FORMAT AS CSV DELIMITER ',' QUOTE ''
6 IGNOREHEADER 1
7 REGION AS 'us-east-2'
```

```
+ load-data-fact_life_expectancy-edb1 x load-data-year_dimension-a1b6 x load-data-country_dimension-4a0c x Fact table x 2 dimension tables x
Run Limit 100 Explain Isolated session dwbi-project dev
1 COPY dev.public.year_dimension (year_id, year)
2 FROM 's3://dwbi-project-etl-dataset/Year_Dimension.csv' IAM_ROLE 'arn:aws:iam::394451548552:role/AmazonS3ReadOnlyAccess-DWBI'
3 FORMAT AS CSV DELIMITER ',' QUOTE '' IGNOREHEADER 1
4 REGION AS 'us-east-2'
```

```
+ load-data-fact_life_expectancy-edb1 x load-data-year_dimension-a1b6 x load-data-country_dimension-4a0c x Fact table x 2 dimension tables x
Run Limit 100 Explain Isolated session dwbi-project dev
1 COPY dev.public.country_dimension (country_id, country, region, economy_status_developed, economy_status_developing)
2 FROM 's3://dwbi-project-etl-dataset/Country_Dimension.csv' IAM_ROLE 'arn:aws:iam::394451548552:role/AmazonS3ReadOnlyAccess-DWBI'
3 FORMAT AS CSV DELIMITER ',' QUOTE '' IGNOREHEADER 1
4 REGION AS 'us-east-2'
```

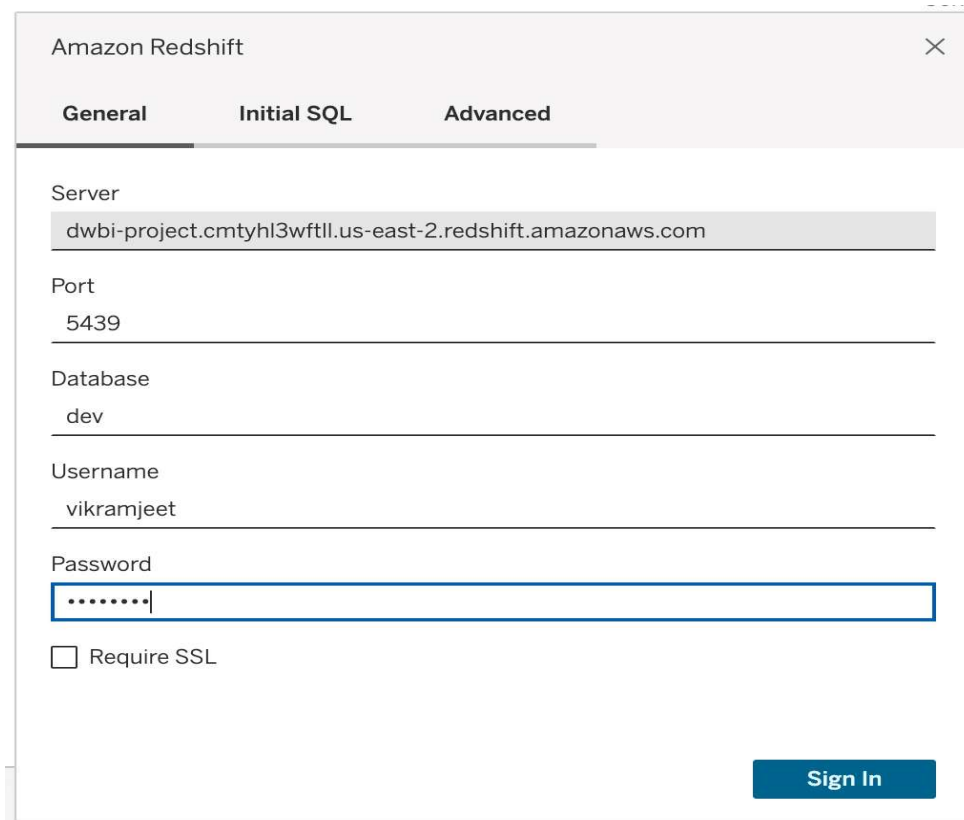
Post-load, a series of SQL queries were executed to validate record counts, data types, and the presence of any null values in critical fields.

3. Report and Dashboard Development:

We utilized Tableau to develop a comprehensive set of reports and dashboards for facilitating data analysis and trend identification. A live connection was established between Tableau and the AWS Redshift data warehouse.

Please find the below details to connect to our Redshift data warehouse:

- **Server** - dwbi-project.cmtyh13wftll.us-east-2.redshift.amazonaws.com
- **Port** - 5439
- **Database** - dev
- **Username** - vikramjeet
- **Password** - Anurag99



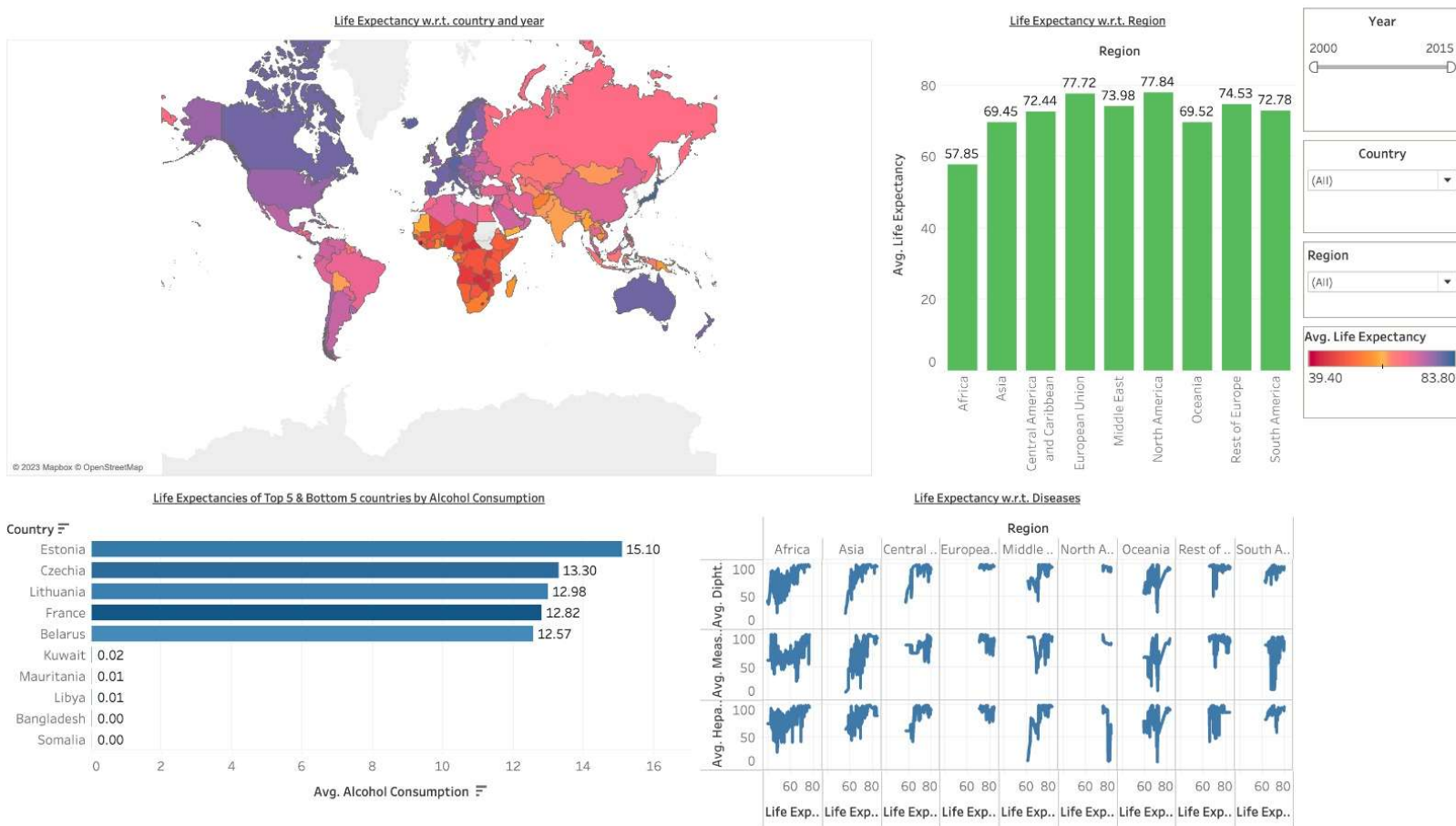
The screenshot shows the 'Amazon Redshift' connection window in Tableau. It has three tabs: 'General', 'Initial SQL', and 'Advanced'. The 'General' tab is selected. The fields are filled with the following information:

- Server: dwbi-project.cmtyh13wftll.us-east-2.redshift.amazonaws.com
- Port: 5439
- Database: dev
- Username: vikramjeet
- Password: [masked with dots]
- ☐ Require SSL

A 'Sign In' button is located at the bottom right of the window.

Please note that you will need to download specific drivers related to AWS Redshift based on your operating system.

We connected the country and year dimension tables to the fact life expectancy table using their respective primary and foreign keys, and then proceeded to create the below dashboard with 4 visualizations.



- **Visualization Number 1:**

Life Expectancy by Country and Year (Map Visualization) -

Countries are color-coded based on average life expectancy; wherein different hues represent different ranges of life expectancy. Developed nations, exhibit higher life expectancy, visible through the concentration of similar hues. The map allows for year selection, showing how life expectancy changes over time, and therefore indicating overall global health improvements or highlighting regions where lower life expectancy is observed.

- **Visualization Number 2:**

Life Expectancy by Region (Vertical Bar Chart) -

The bar chart shows variations in life expectancy across different regions.

Regions such as North America and European Union has higher life expectancy compared to regions such as Africa, possibly due to differences in economic development, healthcare systems, and social factors. The bar chart also displays the change in life expectancy over selected years when the year filter is manipulated.

- **Visualization Number 3:**

Top & Bottom 5 Countries by Alcohol Consumption (Horizontal Bar Chart) -

This visualization exhibits a comparison of average alcohol consumption in top 5 and bottom 5 countries. Life expectancy is displayed in the detail of the bar graph. It may reveal cultural or regional patterns in alcohol consumption, with certain countries or regions displaying significantly higher or lower consumption.

- **Visualization Number 4:**

Life Expectancy with Respect to Diseases (Line Graph) -

The line graph expresses the relationship between life expectancy and the prevalence or incidence of various diseases across different regions. Each small section corresponds to a region and a particular disease, and the trend lines indicate whether a higher incidence of certain diseases correlates with life expectancy within that region. This visualization can be particularly useful to identify regions where certain diseases have a significant impact on life expectancy, highlighting areas where healthcare interventions might be needed.

- **Filters and Interactivity:**

The dashboard includes interactive elements such as filters for countries, year and regions, which allow to refine the data displayed and focus on specific aspects of the data.

Selecting a country on the map will update visualization number 4 to reflect data for that country, providing a detailed view of that country's health indicators.

- **Overall Insights:**

The dashboard provides a comprehensive overview of global health trends, emphasizing the disparities in life expectancy across different regions and countries, year-wise. This suggests potential areas of concern where public health efforts could be targeted to improve life outcomes.

By analyzing the data, stakeholders can identify patterns and correlations that could inform policy decisions, healthcare resource allocation, and other strategic initiatives aimed at improving global health.sss