1. Project Overview

The goal of this project was to extend Homework 1 by fully developing a cloud-based data pipeline and analytics workflow using a real-world dataset. I continued working with the **Texas Workers' Compensation Non-Subscriber Employer dataset**, it contains information about companies that opted out of workers' compensation coverage in Texas.

2. Data Sourcing

- **Data Source**:
  Texas Open Data Portal:
  https://data.texas.gov/dataset/Workers-compensation-non-subscriber-employer-infor/azae-8krr
- **Steps Taken**:
  - Reviewed and understood the structure of the dataset
  - Manually created a **data dictionary** with fields, descriptions, types, and constraints
  - Uploaded the raw dataset to **Snowflake** using the web interface

3. Data Storage

- Chose **Snowflake** as the cloud data platform
- Created a new database and schema to store raw and cleaned data
- Loaded the CSV file into a raw table in Snowflake
- Executed SQL queries to explore and validate the data

4. Data Transformation

- Used **DBT (Data Build Tool)** to write and manage SQL transformations
- Created a cleaned version of the dataset with the following modifications:
  - Removed duplicates and null values
  - Standardized date formats to `YYYY-MM-DD`
  - Added calculated columns:
    - `duration_days` (difference between start and end date)
    - `year`, `month`, `quarter` (from start date)

- DBT compiled and executed SQL code, creating transformed tables directly in Snowflake

5. Data Modeling

- Modeled the dataset with a basic dimensional schema in Snowflake:
    - **Fact Table**: `fact_duration` containing company ID, duration, and dates
    - **Dimension Tables**:
        - `dim_company` with company info
        - `dim_date` with calendar breakdown
- Used DBT to generate and document these models in a modular and reusable way

6. Data Visualization

- Used **Tableau** to connect directly to Snowflake and create interactive dashboards
- Visuals included:
    - **Filter** by year and state
    - **Pie Chart**: Company distribution of non-subscribers
    - **Column Chart**: Frequency by ZIP code
    - **Line Chart**: Opt-out trends over time
    - **Heat Map**: Company vs. duration

7. GitHub Repository

- Uploaded:
    - DBT project scripts (`.sql` files for transformations and models)
    - SQL scripts for Snowflake setup
    - Data dictionary and data mapping in Markdown format
    - README file explaining project steps

8. Conclusion

This project gave me the chance to build a full cloud data pipeline using Snowflake and DBT. I cleaned and organized real-world data, and then used Tableau to create clear, useful visuals. Working through each step helped me get more comfortable with cloud storage, SQL, and ETL best practices.