Final Project Report

Haichuan Zhao & Lixin Li

INFT4001-Special Topics II

Aprl 17, 2024

# Introduction

In this project, we selected the UK traffic accident dataset as the research data. The project involves designing an Entity-Relationship Diagram (ERD) based on the dataset, importing the data into a database, performing SQL queries, and conducting optimization comparisons for typical queries. The decision to choose this dataset was based on several considerations: Firstly, the dataset is sufficiently large, allowing us to effectively demonstrate the time differences before and after optimization in SQL query performance comparisons. Secondly, the dataset is comprehensive and authentic, containing accident, vehicle, and casualty data from 2018 to 2022, spanning five years. Lastly, the dataset provides rich dimensional information, enabling us to extract interesting insights.

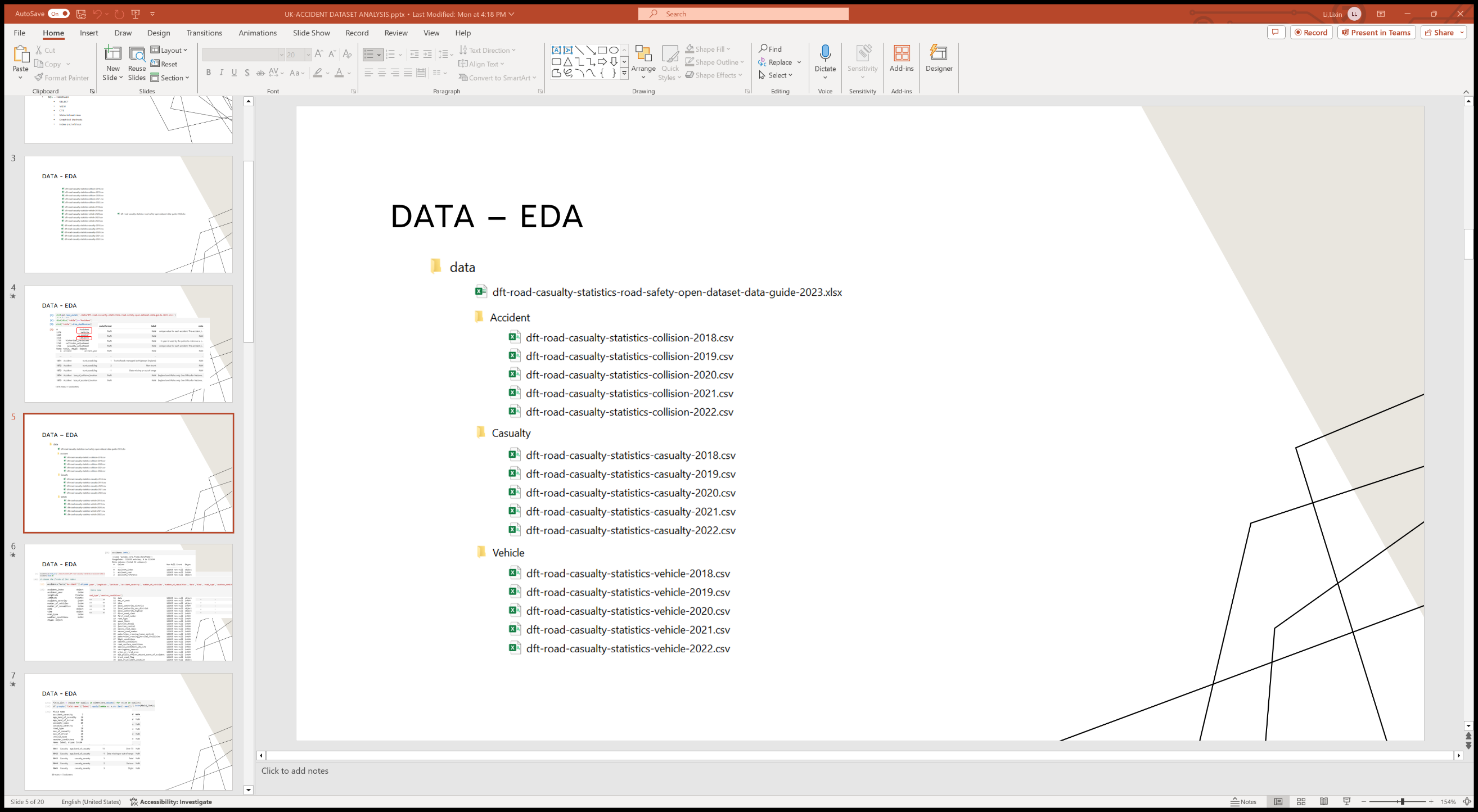
In this project, Lixin is responsible for data processing, including data exploration, Entity-Relationship Diagram (ERD) creation, and data Extract, Transform, Load (ETL) processes. Haichuan is responsible for SQL querying and performance optimization tasks. The final presentation slides and report documents are collaboratively completed by both team members.

# Data Processing

We use Python to process data and import it into the database. Visual Paradigm is used to design the ERD, and SQL scripts are used to create the data structure. Python is also used to execute the SQL scripts.

## Exploratory Data Analysis (EDA)

We explore the data to determine the field scope, data types, and lengths for the fact table. We also determine the method of importing data.



Accidents as an example of fact tables:

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

Dimention tables:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Entity-Relationship Diagram (ERD)

This is the ERD corresponding to the data, including the foreign key constraint relationships between tables.

A diagram of a computer

Description automatically generated

The SQL script to initialize data structure:

A screenshot of a computer program

Description automatically generated

## Extract, Transform, Load (ETL)

Three tasks are accomplished through Python code: initializing data structure, importing dimension data, and importing fact table data.

initializing data structure:

A screenshot of a computer program

Description automatically generated

importing dimension data:

A screenshot of a computer code

Description automatically generated

importing fact table data:

A computer screen shot of a computer code

Description automatically generated

# SQL Queries

# Conclusion

In summary,

# References:

Road Safety Data - data.gov.uk

https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data