Data Foundations Assignment

This repository contains the code and documentation for a Data Foundations assignment completed during an MSc program. The assignment focused on data selection, cleaning, legal considerations, structured and semi-structured data, data modeling, and implementation. The project utilized R programming language and SQL databases.

Data Selection & Cleaning

The data used for this research came from various public resources. The following sources were used:

Median House Prices: Data obtained from the Office of National Statistics (ONS) website. Specifically, the "Median price paid by ward, England and Wales" dataset for the years 2000 to 2020 was used. The data was divided into quarterly averages.

Data Source: ONS Median House Prices

Prices Broadband Data: Data sourced from the "Broadband connectivity and speeds" table on the House of Commons Library website. Four attributes were considered: ward code, average download speed, connections receiving super-fast speeds, and connections receiving over 300 Mbps.

Data Source: House of Commons Library

Council Tax Data: Data provided by the Oxfordshire County Council. Separate links were used for each district, containing information about tax bands and rates per town.

Data Source: Oxfordshire County Council

Council Tax Data Ward Table: Data obtained from the Open Geography portal. The relevant data was filtered out to include only the districts of interest. This table serves as a reference for the other tables.

Data Source: Open Geography Portal

The initial cleaning process was performed using Excel spreadsheets. Column names were modified to suit the project's requirements, and SQL reserved symbols were replaced. Only the necessary columns were used for broadband data, as specified in the coursework description. The council tax data was manually loaded into the database since no collective file was available.

Legal Issues

All the data used in this project is publicly available from government-maintained websites. The data is provided for transparency and public use by the respective authorities. The validity of the data is ensured by the organizations that collected and maintained it.

House Prices Data: The Office of National Statistics (ONS) maintains and reports the house price data directly to the UK Parliament. The data is licensed under the Open Government Licence, allowing for personal use.

Broadband Data: The broadband speed data is provided by the UK Parliament, sourced from Ofcom. The data is publicly available and can be used for research purposes.

Council Tax Data: The council tax data is provided by the Oxfordshire County Council, a trusted public provider. However, the use of council tax data is limited to this specific source and should not be utilized for financial gain.

Ward Table: The ward table data is collected from the Open Geography Portal, using official data from ONS geography. The data is shared under the Open Government Licence.

The data used in this project does not include any personally identifiable information, ensuring privacy and compliance with legal requirements.

Methodology

In this research project, a methodological approach was employed to analyze and interpret the available data. The data was collected from various publicly accessible sources, including government websites and official repositories. The chosen data sources were carefully selected to ensure reliability and validity.

To organize and store the data effectively, an SQLite database was utilized. The tables within the database were structured in a normalized form, adhering to the third normal form (3NF) to eliminate redundancies and establish appropriate keys. This database design allowed for efficient data management, enabling relationships between tables through primary and foreign keys.

R, along with several relevant libraries such as haven, dplyr, dbplyr, RSQLite, and DBI, was employed for data manipulation, analysis, and querying. The connection between R Studio and the SQLite database was established using the "DBI" and "RSQLite" packages, facilitating seamless interaction with the database.

The R code included various queries to extract specific information from the database. Each query was carefully crafted to retrieve the required data based on the research questions. Functions such as dbGetQuery() were utilized to execute SQL queries and retrieve the desired results.

It is important to acknowledge the limitations of the available data and the modest scope of this research project. The focus was on utilizing the provided data sources and analyzing specific aspects within the designated regions. Nevertheless, this methodology provided valuable insights into the selected variables and their relationships, allowing for informed interpretations and potential further exploration in the future.