**Data Science on House Prices & Broadband**

**Data selection and cleaning**

Obtaining and utilizing datasets related to house prices, broadband speeds, and council tax charges for specific areas such as Oxfordshire involves a systematic approach. In this process, the key stages include identification of data, data acquisition, cleaning, and utilization. The links provided in the coursework served as valuable sources for obtaining the necessary datasets.

**1. Identification of data:** The first stage involved identifying reliable sources for the datasets related to house prices, broadband speeds, and council tax charges in Oxfordshire. The provided links were crucial in this regard, leading to official sources such as the Office for National Statistics (ONS) for house prices, the Commons Library for broadband coverage and speeds, and the Oxfordshire County Council for council tax charges.

**2. Data Acquisition:** Once the sources were identified, the next step was to access and download the relevant datasets. For house prices, the ONS link provided access to the "Median Price Paid by Ward" dataset, which includes comprehensive information on house prices in different areas. The Commons Library link offered datasets on broadband coverage and speeds for constituencies, while the Oxfordshire County Council link led to datasets on council tax charges. Obtaining data from official government and parliamentary sources, such as the ONS and the Commons Library, is a robust method. These organizations are dedicated to collecting and maintaining high-quality datasets, ensuring that the information retrieved is comprehensive, up-to-date, and reflective of the actual situation. Relying on such authoritative sources enhances the credibility of the data.

**3. Manual Filtering:** After obtaining the datasets, manual filtering was necessary to focus on the specific areas of interest within Oxfordshire. For house prices, the ONS dataset contained information for multiple regions, and filtering was performed to isolate data relevant to Oxfordshire. Similarly, the broadband speeds dataset from the Commons Library was filtered to include only the constituencies within Oxfordshire. The council tax charges dataset from the Oxfordshire County Council was filtered to include information on districts or specific areas within Oxfordshire. The decision to manually filter datasets to specific areas within Oxfordshire was driven by the need for precision and relevance in the analysis. Manual filtering allows for a targeted approach, ensuring that the datasets are tailored to the requirements. Automation may introduce errors, especially when dealing with geographically specific data, making manual intervention a prudent choice for accuracy.

**4. Data Cleaning:** With the filtered datasets in hand, the next crucial step was data cleaning. This involved addressing missing values, removing duplicates, and standardizing data formats. Inconsistencies in naming conventions or data entry errors were corrected to ensure the datasets were accurate and ready for analysis. The emphasis on data cleaning as a part of the preparation process is essential for maintaining data quality. Addressing missing values, removing duplicates, and standardizing formats contribute to the reliability of the datasets. By adhering to these practices, the datasets become more robust and suitable for analysis, minimizing the risk of inaccurate or biased results.

The methods used to find, acquire, clean, and use datasets for broadband speeds, council tax rates, and home prices in Oxfordshire were carefully planned to satisfy high-quality standards, guaranteeing the stability and dependability of the data for further analysis. Authoritative sources, careful data cleaning, and manual curation all work together to maintain the dataset's overall integrity and the validity of any conclusions that may be made from it.

**Structured and semi-structured data**

**Structured Data Model (SQL):** A Structured Query Language (SQL) database is a robust and widely used choice for storing structured data, providing a clear and organized format for representing relationships among various entities. SQL database offers a structured and efficient solution for this project. Here are the reasons why:

* **Tabular Representation:** SQL databases organize data in tables, with each table representing a specific entity (e.g., houses, broadband data, council tax information). This tabular structure simplifies data retrieval and supports the relational nature of the data.
* **Data Integrity:** SQL databases enforce data integrity constraints through relationships and foreign key constraints, ensuring that the data is consistent and accurate. This is crucial when dealing with multiple datasets that may have interconnected information.
* **Query Capabilities:** SQL provides powerful query capabilities, allowing for complex data retrieval, aggregation, and analysis. This is beneficial when exploring relationships between house prices, broadband speeds, and council tax charges in Oxfordshire.

**Semi-Structured Data Model (XML):** XML (eXtensible Markup Language) is a semi-structured data model that represents data in a hierarchical and self-descriptive format. While XML is flexible, it may not be the most efficient choice for storing the structured data in this project. Herer are the reasons why:

* **Hierarchical Structure:** XML organizes data hierarchically, which is suitable for representing complex relationships. However, for the relatively structured nature of this project, a hierarchical model might introduce unnecessary complexity.
* **Flexibility:** XML allows for flexibility in data representation, accommodating variations in data structures. This flexibility may be advantageous in certain scenarios, but it could lead to increased complexity and potential challenges in maintaining data integrity.

Considering the structured nature of the data in this project and the need for efficient relational queries, a SQL database is the primary recommendation. SQL's tabular structure, data integrity enforcement, and powerful query capabilities align well with this project's objectives.

However, if there is a specific requirement for a more flexible and self-descriptive representation of the data, XML could be used in conjunction with SQL. XML can be employed to store metadata or supplementary information, providing an additional layer of flexibility while the core structured data remains in the SQL database. While SQL is the primary recommendation, the use of XML can be considered if there are specific requirements for a more flexible and extensible data representation alongside the structured SQL database.

**Legal and/or ethical issues**

It is crucial to ensure compliance with data protection regulations, respect individual privacy, and maintain transparency in handling sensitive information. The proposed system for collecting and utilizing data on house prices, broadband speeds, and council tax charges in Oxfordshire necessitates careful attention to legal and ethical considerations to ensure compliance and safeguard individual rights.

**Data Privacy and Protection:**

* **Legal Perspective:** Adherence to data protection laws, such as GDPR, is mandatory, requiring explicit consent for collecting and processing personal information.
* **Ethical Perspective:** Respecting privacy rights is paramount, necessitating transparent communication and ethical practices like anonymization.

**Accuracy and Fairness:**

* **Legal Perspective:** Ensuring data accuracy is crucial to avoid legal consequences, particularly in areas like house prices and council tax charges.
* **Ethical Perspective:** Ethical responsibility lies in preventing biases and ensuring fairness in data representation to avoid perpetuating inequalities.

**Security of Data:**

* **Legal Perspective:** Robust security measures are required by law to protect against unauthorized access or breaches that could lead to legal consequences.
* **Ethical Perspective:** Ethical obligations include investing in cybersecurity to uphold trust and protect individuals' data.

**Transparency and Informed Consent:**

* **Legal Perspective:** Legal requirements dictate transparent communication about data collection practices, purposes, and obtaining informed consent.
* **Ethical Perspective:** Informed consent is an ethical imperative, providing individuals control over their data and fostering trust.

The proposed system should align with legal requirements and ethical principles to protect individuals' rights, ensure data accuracy, and promote fairness.

**Data model and implementation**

**SQL**

The process of normalizing a relational database involves organizing and structuring tables to minimize data redundancy and enhance data integrity. The normalization process follows several normal forms, with the goal often being the achievement of the third normal form (3NF).

In the initial stage, known as the First Normal Form (1NF), the focus is on eliminating duplicate columns and ensuring that each column contains atomic values. All the provided tables, including "AreaTable," "AvailabilityTable," "LocalAuthority," "MedianPricesPaid," "SpeedTable," "Ward," and "CouncilTax," meet the criteria for 1NF. Each table exhibits simple atomic attributes without any repeating groups.

Moving on to the Second Normal Form (2NF), the objective is to remove partial dependencies by ensuring that non-prime attributes are fully functionally dependent on the primary key. In the "MedianPricesPaid" table, this involves creating a composite primary key consisting of "LocalAuthorityCode" and "WardCode" to address partial dependencies and ensure a more normalized structure.

Finally, in the Third Normal Form (3NF), the goal is to eliminate transitive dependencies by ensuring that non-prime attributes are not dependent on other non-prime attributes. In the "CouncilTax" table, for instance, the band taxes are directly dependent on the primary key "WardCode" and are not influenced by other non-prime attributes like "District" or "Area," thus adhering to the principles of 3NF.

By successfully achieving 3NF, the tables are effectively organized with minimal data redundancy. Each table has a well-defined primary key, and relationships among attributes are appropriately handled. This normalized database schema contributes to better data management, integrity, and overall efficiency.

I have created a relationship diagram for my database for visualization and clarity.  
A screenshot of a computer

Description automatically generated

**R Code**

I've organized the queries specified in the coursework into distinct R files, numbered 1 through 8, for clarity. These files are now compressed into a single folder for ease of access. It's important to be aware that each R file contains code that references a specific location for the database. Before executing the code, make sure to modify the location to match the actual storage location of the database on your system.

Query 1:

Code

A screenshot of a computer program

Description automatically generated  
Output

A screenshot of a computer

Description automatically generated

My code produces the average price calculated over a span of two years for properties in the Carfax ward within the Oxford district for the years of 2018 and 2019, yielding a result of 803.

Query 2:

Code

A close-up of a computer screen

Description automatically generated

Output



The output of my code indicates a calculated average percentage increase in prices, amounting to 2.14%, for the Summertown ward in the Oxford district during the years 2018 and 2021.

Query 3:

Code   
A screenshot of a computer code

Description automatically generated

Output

A screenshot of a computer

Description automatically generated

The output of my code shows that Benson & Crowmarsh in the district of South Oxfordshire has the highest house price in the year of 2021 in the quarter of March out of all the Wards in South Oxfordshire.

Query 4:

Code

A screenshot of a computer program

Description automatically generated

Output

A screenshot of a computer

Description automatically generated

The output of my code shows that 99.9% of areas in Holywell have available superfast broadband speed.

Query 5:

Code

A computer screen shot of a code

Description automatically generated

Output

A screenshot of a computer code

Description automatically generated

My code output gives useful information, namely the average download speed for a given area—Churchill, in this case—as well as the area code for that area. These specifics include important and practical information for users.

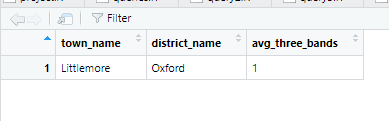
Query 6:

Code

A screenshot of a computer

Description automatically generated

Output



The code output for this query is inaccurate. Despite investing considerable time in troubleshooting, I have been unable to rectify the issue, specifically in calculating the average of the three bands. While the code appears that it should work, the practical outcome shows an average value of 1. This could be due to problems with the data stored or it could be a small issue in my code which has been overlooked.

Query 7:

Code  
A screenshot of a computer code

Description automatically generated

Output

A screenshot of a computer

Description automatically generated

Once again, the output for this is incorrect. The problem seems to appear when trying to calculate between the integers of each Band even though they are stored properly. Again, the code appears that it should work however in practice it does not.

Query 8:

Code  
A screenshot of a computer program

Description automatically generated

Output

A screenshot of a computer

Description automatically generated

The code successfully identifies Adderbury in the Cherwell district as the town with the lowest council tax charges for Band B properties, amounting to £1,695.