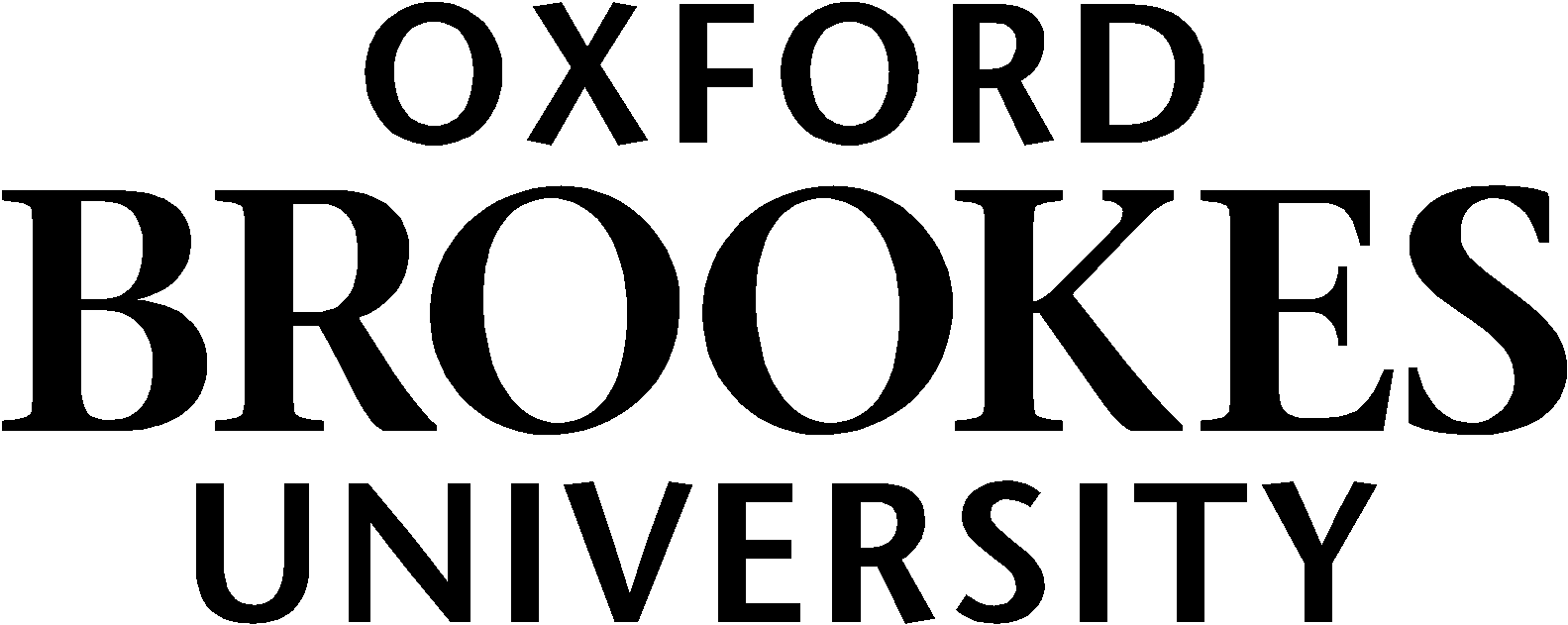
****

**Assessment cover**

**STUDENTS, PLEASE COPY THIS PAGE AND USE AS THE COVER PAGE FOR YOUR SUBMISSION**

|  |  |  |  |
| --- | --- | --- | --- |
| Module No: | DALT7002 | Module title: | Data Science Foundations |

|  |  |
| --- | --- |
| Assessment title : | Data Science Foundations |

|  |  |
| --- | --- |
| Due date and time**:** | **1:00 PM, 28 Mar 2024** |

|  |  |
| --- | --- |
| Estimated total time to be spent on assignment: | 30 hours per student |

**LEARNING OUTCOMES**

|  |
| --- |
| **On successful completion of this assignment, students will be able to achieve the module following learning outcomes (LOs):** *LO numbers and text copied and pasted from the module descriptor* |
| 1. Demonstrate the ability to identify and integrate data of various types from traditional and alternative sources, and make informed judgements about their use in data science research |
| 1. Critically evaluate the methodologies applied in data collection, data processing, data analysis & dissemination of research findings |
| 1. Critically assess methods and data strengths and limitations combined to application of R |

|  |  |  |
| --- | --- | --- |
| **Engineering Council AHEP4 LOs assessed (from S2 2022-23)**  *LOs copied and pasted from the AHEP4 matrix (add rows as required)* | |  |
| **LO number** | **LO text** | **Met? (Y/N)** |
|  |  |  |
|  |  |  |
|  |  |  |

**STUDENT NAMES (ONLY IF GROUP ASSIGNMENT, OTHERWISE ANONYMOUS)**

|  |  |  |
| --- | --- | --- |
| **Student No:** | **Student Name:** | **Group Name and Number:** |
| **19141230** |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Statement of Compliance *(please tick to sign)***

I declare that the work submitted is my own and that the work I submit is fully in accordance with the University regulations regarding assessments *(*[*www.brookes.ac.uk/uniregulations/current*](http://www.brookes.ac.uk/uniregulations/current)*)*

**x**

**RUBRIC OR EQUIVALENT (BELOW)**

**FORMATIVE FEEDBACK OPPORTUNITIES**

|  |
| --- |
|  |

**SUMMATIVE FEEDBACK DELIVERABLES**

|  |  |
| --- | --- |
| **Deliverable content and standard description and criteria** | **Weighting out of 100%** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| **Individual reflection section:**   * Propose further work that would offer improvements and enhancements. * Evaluate personal learning and development in terms of technology/hardware/software/group work. | **10%** |

**Marking grid and peer marking form are attached at the end of this assignment.**

Table of Contents

[Introduction 5](#_Toc162505412)

[Data Selection and Cleaning 5](#_Toc162505413)

[Stages (Steps) and Justifications 5](#_Toc162505414)

[House Prices 5](#_Toc162505415)

[Broadband 6](#_Toc162505416)

[Council Tax Charges 8](#_Toc162505417)

[Legal and/or ethical issues 9](#_Toc162505418)

[Structured and Semi-structured Data 10](#_Toc162505419)

[Data Model and Implementation 10](#_Toc162505420)

[1NF 11](#_Toc162505421)

[2NF 11](#_Toc162505422)

[3NF 12](#_Toc162505423)

[R Code 13](#_Toc162505424)

[Instructions 13](#_Toc162505425)

[Code 14](#_Toc162505426)

[SQL Codes 16](#_Toc162505427)

[1. Create Tables Code 16](#_Toc162505428)

[Location Information (Parent) 16](#_Toc162505429)

[House Price Information (Child) 16](#_Toc162505430)

[Council Tax Information (Child) 17](#_Toc162505431)

[Broadband Information (Child) 17](#_Toc162505432)

[2. Code to Insert Unormalised CSV Data into Normalised Tables 17](#_Toc162505433)

[Location Information 17](#_Toc162505434)

[House Price Information 18](#_Toc162505435)

[Council Tax Information 18](#_Toc162505436)

[Broadband Information 18](#_Toc162505437)

[References: 18](#_Toc162505438)

[Appendixes: 20](#_Toc162505439)

**ASSIGNMENT IN DETAIL**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Introduction This report considers the house price, broadband and council tax charges data for the county of Oxfordshire. The mixed data (qualitative and quantitative) has been extracted from data sources like the Office for National Statistics (ONS), the House of Commons Library (HOCL) and Oxfordshire County Council data. The report is particularly interested in understanding the volatility in property price changes, council tax charges and broadband accessibility in Oxfordshire, particularly in individual wards to gain a general simplified representative insight. Data Selection and CleaningStages (Steps) and JustificationsHouse Prices The “Median House Prices by Ward: HPSSA dataset 37: Year Ending March 2023 Edition of this Dataset” from the Office of National Statistics (ONS) provides the latest version of house prices for a range of national and subnational geographies, since March 1995 to March 2023, for all types of builds/properties in the United Kingdom.  I only used, cleaned, and exported sheet number *1a* into a CSV file from the entire workbook as it contains all the necessary house price data (from all types of properties in the UK) to perform queries from task 3 and 4.  Through manual search I identified then isolated Oxfordshire County’s districts with its corresponding wards’ data. As Holywell of Oxford district showed null values, it was removed from the dataset. This is due to there being none to fewer than five house sales in the given year for the selected geography, which is deemed as inadequate to produce a robust average, therefore it is not reported to avoid presenting misleading values for its house price. Also, as annual price data for 2023 is incomplete, I could not provide a contemporary dataset.  Moreover, the annual data is originally divided into quartiles; therefore, I selected quartiles between 2021 and 2022 data. Firstly, because the queries for *House Prices data* only require two years’ price data to perform the average price comparisons. And secondly, because I will identify which ward in Oxfordshire County has got the highest house price for the quartile of Dec 2022, which can also be interpreted as which ward has the highest house price for the year 2022. With this logic, the values of Dec 2021 and Dec 2022 are useful to investigate the percentage increase or decrease of house prices between two years. Lastly, I supplemented the ‘Band’ attribute, thus enabling users to autonomously identify how much council tax the property alone incurs. The majority of council tax charges of 2021-22 for Oxfordshire’s wards was not shown in the website’s reports, hence why it is not reported in our tables. Broadband The ‘Constituency Data: Broadband Coverage and Speeds’ workbook from the House of Common’s Library (HOCL) provides a simplified version of The Office of Communication (OFCOM) Connected Nations 2023 ‘annual report on progress in the availability of broadband in the UK’. The benefits of HOCL dataset are the user-friendliness and comparably lower file size to OFCOM’s dataset. Selecting data directly from OFCOM, which organises data by postcode, would have augmented the complexities to our data management process and processing time. Given the time constraint, opting for the HOCL Constituency Data proved to be the more efficient and time-saving choice that allowed time to test our code for errors and improvements.  Since the queries (tasks 6 - 7) primarily focus on ward-level information, each tuple in my database represents information specific to one individual ward at a time (no repetitions) within Oxfordshire.  Since the data source organises data based on sub-constituency (areas), - middle layer super output areas (MSOAs) -, rather than wards. To reconcile the data organisation discrepancy between the source and my database, I manually searched for areas that pertain exclusively to Oxfordshire’s wards. I grouped these areas into wards, then calculated the average for attributes ‘Gigabit Availability’, ‘Superfast Availability’ and ‘Average Download Speed’. These values subsequently became the ward’s average values.  In other instances, in the data source file, wards like Kennington and Radley (a singular ward) were recorded as separate entities. I mitigated the problem by calculating the average value between the two entities, to represent the entire ward. Lastly, for transparency, I incorporated an attribute (‘Calculated as’) which reveals whether the value of the ward was derived from a group average, labelled as ‘Grouped’, or if it is its own ‘Individual’ and exact value.  Finally, using averages reduces file size by condensing data and providing a single representative value which reflects trends or status for each attribute, based on the data available for wards in Oxfordshire. Consequently, this simplification allows for easier comparison and analysis across wards. Council Tax Charges The council tax charges for Bands A to H - period 2023-24, are sourced from reports and related meeting discussions from the official website of district councils. I manually inserted these values into my database.  Tuples in the Council Tax Information table, represent the council tax rates for each *ward* in Oxfordshire. This is to ensure that the database is consistent with its aim to focus on charge differences among wards and districts.  The ‘Calculated as’ attribute, provides clarity about how values were derived. By categorising values as derived from either: ‘Town’ rate – where the council tax rate representing a town is applied to all its respective wards. For example, in our table wards Kidlington East and Kidlington West wards (Location ID - 7 and 10) report the same council tax charges because they are considered a single town (Kidlington), in the Cherwell District Council website.  Or, ‘Grouped’ rate – where the ward’s council tax rate is calculated as an average of multiple areas/towns within the ward. For example, Banbury Hardwick ward encompasses two towns with differing rates in each band; later, using these rates I calculated their average and reflected these values into the applicable wards’ bands. Consequently, the grouped average approximates the overall charges for the entire ward in the database.  In contrast, ‘Unparished’ refers to a special standard council tax rate (precept), only applicable for wards with no parish within the City of Oxford district.  In conclusion, this form of categorising and presenting values exemplifies to both current and prospective residents of Oxfordshire about, discrepancies in internet accessibility, house prices, and the related council tax charges across different electoral districts (wards).  As residents wield influence in selecting councillors (who are involved in the fiscal decisions/campaigns of the ward). The database structure raises awareness in the three areas (prices, internet, and tax), which can motivate some to potentially play a more active role in shaping their communities by communicating with their councillors. Legal and/or ethical issues The database aims to align with the Data Protection Act 2018, by providing transparency regarding the source, methods, and calculations used for all data. Table 1 provides a description of the database’s variables and describes parameters in more detail. The purpose of the project’s aim has been explicitly explained in the introduction, and the use of data is only limited to what is adequate and necessary for our research. For transparency, I ensured to state the year/period that the data covers.  While the ONS, Ofcom, House of Commons Library and district councils’ websites have published data and granted permission (Open Government License) for public use, particularly for research purposes. It’s crucial to recognise the risk of inaccuracies due to manual processes. Manual calculations and reporting may lead to errors, resulting in misleading conclusions or decisions which may cause harm to individuals or organisations; Therefore, readers are advised to view this database as a supplementary resource and conduct their own research for validation. Also, because the database lacks security measures such as encryption and regular security audits, our data could be manipulated by sophisticated hackers. Structured and Semi-structured Data While XML offers flexibility in representing hierarchical data structures, and provides additional information about data values, elements, and attributes, through self-generated (natural language) descriptive tags, the information is intermixed in the schema. Conversely, SQL provides structured data, with pre-defined schemas, that enables setting constraints through primary keys, foreign keys and through its defined data types. These constraints also enable creating relations among the tables, in this way enforcing data integrity (consistency), which is not strictly required in XML semi-structured data. The lack of constraints augments the risk of creating inconsistencies and errors in large databases using XML. As our dataset consists of large and diverse data, XML is not appropriate for our project. Moreover, XQuery’s and XPath’s less expressive commands limit XML from performing relational operations efficiently. Conversely, the range of commands such as JOINs, GROUP BY, CASE WHEN make SQL a more effective tool for data analysis. Additionally, XML structures do not require normalisation, consequently, this may increase the user’s required computational resources, augment the file size and may potentially incur high transmission costs. Because of this SQL is recognised as a more efficient alternative to parse in R. Therefore, it is redundant to use semi-structured data for our research as we require immediate responses when querying data. SQL is adequate as it optimises performance and retrieval of our data through its lower memory storage files. Data Model and Implementation The following tables have been normalised to first, second and third normal form. These tables have been simplified for better readability and identification of variables and related values for the user/analyst. The definition of primary keys can be read on [Appendix 1](#_Appendixes:). 1NF To normalise my data into first normal form, I noted all of the atomic attributes required for a complete database, into one structure, and I identified a primary key (Location Id) which can uniquely identify tuples. I also ensured to remove repeating groups.  Year | Location Id (PK) | District code | District Name | Ward code | Ward Name | House Price Id | Quartile Mar 21 | Quartile Jun 21 | Quartile Sep 21 | Quartile Dec 21| Quartile Mar 22 | Quartile Jun 22 | Quartile Sep 22 | Quartile Dec 22 | Band 2021 | Band 2022 | Council Tax Record Id | Period | District code | District Name | Ward code | Ward Name | Calculated as | Band A Charge | Band B Charge | Band C Charge | Band D Charge | Band E Charge | Band F Charge | Band G Charge | Band H Charge | Broadband Record Id | Gigabit Availability | Average Download Speed | Superfast | 2NF Next, to achieve the second normal form, I distributed the attributes from 1NF into the three tables below. The attributes are completely dependent on the primary key, which means that the PK uniquely identifies each tuple.  **House price**  Location Id | District code | District Name | Ward code | Ward Name | House Price Record Id (PK) | Quartile Mar 21 | Quartile Jun 21 | Quartile Sep 21 | Quartile Dec 21| Quartile Mar 22 | Quartile Jun 22 | Quartile Sep 22 | Quartile Dec 22 | Band 2021 | Band 2022 |  **Council Tax**  Council Tax Record Id (PK) | Period | Location Id | District code | District Name | Ward code | Ward Name | Calculated as | Band A Charge | Band B Charge | Band C Charge | Band D Charge | Band E Charge | Band F Charge | Band G Charge | Band H Charge |  **Broadband**  Broadband Record Id (PK) | Year | Location Id | District code | District Name | Ward code | Ward Name | Calculated as | Gigabit Availability Average Download Speed | Superfast |  **Location**  Location Id (PK) | District code | District Name | Ward code | Ward Name | 3NF Finally, to obtain the third normal form from 2NF tables, the transitive dependencies like ‘Year’ and ‘Period’ are removed from the Broadband and Council Tax Information tables and instead are stated in the title of the tables for transparency.  **Parent – Location Information Table:**  Location Id (PK) | District Code | District Name | Ward Code | Ward Name |   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Location Id (PK) | District Code | District Name | Ward Code | Ward Name |   **Child – House Price Information Table, Years 2021-2022:**  House Price Id (PK) | Location Id (FK) | Quartile Mar 21 | Quartile Jun 21 | Quartile Sep 21 | Quartile Dec 21| Quartile Mar 22 | Quartile Jun 22 | Quartile Sep 22 | Quartile Dec 22 | Band 2021 | Band 2022 | Council Tax Record Id (FK) |   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | House Price Id (PK) | Location Id (FK) | Quartile Mar 21 | … | Quartile Dec 22 | Band 2021 | Band 2022 | Council Tax Record Id (FK) |   **Child – Broadband Information Table, Year 2023:**  Broadband Record Id | Location Id | Calculated as | Gigabit Availability | Superfast | Average Download Speed |   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Broadband Record Id (PK) | Location Id (FK) | Calculated as | Gigabit Availability | Superfast | Average Download Speed |   **Child – Council Tax Information Table, Period 2023-23:**  Council Tax Record Id (PK) | Location Id | Calculated as | Band A Charge | Band B Charge | Band C Charge | Band D Charge | Band E Charge | Band F Charge | Band G Charge | Band H Charge |   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Council Tax Record Id (PK) | Location Id | Calculated as | Band A Charge | … | Band H Charge |  R CodeInstructions Download my\_db.db file which contains the database. Next, open the Rscript.R file and run the code from line 1 all the way to line 144. This ensures that the necessary libraries (RSQLite and DBI) are installed and implemented to run the code successfully. The queries are carried out mostly by using SQL language, which are passed through into R via the ‘dbGetQuery()’ command. SQL language is preferred to run the queries as it enables building a coherent and readable structure while querying data, that can be easily understood.  The structure of the code consists of blocks of code separated by ‘#’ comments, which describe each task for readability purposes. Other, comments within the SQL code body ‘- -' describe the function of the slice of code. There is a moderate amount of essential commenting, but this was done to ensure that the reader was able to follow and comprehend the logic throughout.  The testing system involved manually selecting the necessary values to perform the operations using calculators as well as using R and SQLite software to perform the same operations. All tests provided the same results for every task.  *Query 7* was designed to enable readers compare the internet accessibility statistics among the different districts in Oxfordshire. I incorporated the attribute [average gigabit availability](#_Appendixes:) as it is often associated with the presence of fibre optic connections (FTTP/FTTH) in towns, given that fibre provides gigabit-level internet speeds. However, to verify this information, users should carry out further independent research on the matter. Therefore this query serves as an indication of the internet capabilities reported within Oxfordshire. Code #Packages for db syncing and pass SQL code  install.packages("DBI")  install.packages("RSQLite")  library(DBI)  library(RSQLite)  #Connecting DB path to RStudio  my\_db <- "/Users/genesisvega/Desktop/my\_db.db"  mydb <- dbConnect(RSQLite::SQLite(), dbname = my\_db)  #Task 3 - Average House Price for Deddington of Cherwell district.  #Logic for years 21-22 averages  Ded\_Avg\_Hou\_Price <-dbGetQuery(mydb, "SELECT 'Deddington' AS Ward,  (End\_Mar\_2021 + End\_Jun\_2021 + End\_Sep\_2021 + End\_Dec\_2021) /4 AS 'Avg House Price 21',  (End\_Mar\_2022 + End\_Jun\_2022 + End\_Sep\_2022 + End\_Dec\_2022) /4 AS 'Avg House Price 22'  FROM House\_Price\_Information AS HPI  -- Connects tables via foreign key  JOIN Location\_Information AS LI  ON HPI.Location\_Id = LI.Location\_Id  -- filter result by ward & district code  WHERE LI.Ward\_Name = 'Deddington'  AND LI.District\_Code= 'E07000177';")  print(Ded\_Avg\_Hou\_Price)  #Task 4 - 2021 to 2022 Average increase/decrease in house prices in Blackbird Leys of City of Oxford district.  Bbleys\_Avg\_Percent\_Diff <- dbGetQuery(mydb, "SELECT  -- Operation  ROUND(AVG(End\_Dec\_2022 - End\_Dec\_2021) / AVG(End\_Dec\_2021) \* 100, 2) AS 'Avg % Change'  FROM House\_Price\_Information AS HPI  -- Connects tables via foreign key  JOIN Location\_Information AS LI  ON HPI.Location\_Id = LI.Location\_Id  -- filter result by ward & district names  WHERE LI.Ward\_Name = 'Blackbird Leys'  AND LI.District\_Name = 'Oxford';")  print(Bbleys\_Avg\_Percent\_Diff)  #Task 5 - Find ward within Oxfordshire with the highest house price in a quarter of a year.  # Reveal ward with max price  Max\_Houpri\_Ward <- dbGetQuery(mydb, "SELECT LI.Ward\_name AS Ward,  MAX(End\_Dec\_2022) AS 'Highest House Price in Dec 22'  FROM House\_Price\_Information AS HPI  -- Join to retrieve matching ward name  JOIN Location\_Information AS LI  ON HPI.Location\_Id = LI.Location\_Id;")  print(Max\_Houpri\_Ward)  #Task 6 - Superfast availability % in Cholsey of South Oxfordshire district  Chol\_Sfast <- dbGetQuery(mydb, "SELECT BI.Superfast\_Availability  FROM Location\_Information AS LI  -- Join to retrieve matching ward details  JOIN Broadband\_Information\_2023 AS BI  ON BI.Location\_Id = LI.Location\_Id  WHERE LI.Ward\_Name = 'Cholsey'  AND LI.District\_Code = 'E07000179';")  print(Chol\_Sfast)  # Task 6 - Alternative - find through specifying a % figure  particular\_Sfast\_Avail <- dbGetQuery(mydb, "SELECT BI.Superfast\_Availability, LI.Ward\_Code, LI.Ward\_Name  FROM Broadband\_Information\_2023 AS BI  -- Join to retrieve ward details  JOIN Location\_Information AS LI  ON BI.Location\_Id = LI.Location\_Id  -- Filter by specifying a value %  WHERE BI.Superfast\_Availability = '100.00%';")  print(particular\_Sfast\_Avail)  #Task 7 - Comparing All Districts' Averages: Gigabit Availability, Superfast & Download Speed.  Avg\_Broadb\_perf <- dbGetQuery(mydb, "SELECT LI.District\_Name AS District,  -- Calc Averages to 2 decimal points  ROUND(AVG(Gigabit\_Availability), 2) AS 'Avg Gigabit Availability',  ROUND(AVG(Superfast\_Availability), 2) AS 'Avg Superfast',  ROUND(AVG(Average\_Download\_Speed), 2) AS 'Avg Download Speed'  FROM Broadband\_Information\_2023 AS BI  JOIN Location\_Information AS LI  ON BI.Location\_Id = LI.Location\_Id  -- group averages as districts  GROUP BY LI.District\_Name;")  print(Avg\_Broadb\_perf)  #Task 8 - Average council tax for Bicester town by band  Avg\_Bic\_CTax <- dbGetQuery(mydb, "SELECT 'Bicester' AS Town,  --Calc avg for Bicester bands  AVG(Band\_A) AS Avg\_Band\_A,  AVG(Band\_B)AS Avg\_Band\_B,  AVG(Band\_C) AS Avg\_Band\_C  FROM Council\_Tax\_Information\_2023\_24 AS CI  -- Match tables to retrieve observations starting with 'Bicester'  JOIN Location\_Information AS LI  ON CI.Location\_Id = LI.Location\_Id  WHERE Ward\_Name LIKE 'Bicester%' AND District\_Code = 'E07000177';")  print(Avg\_Bic\_CTax)  # Task 9 - Difference btwn C.Tax charges of same bands, same district, two diff towns  # Define operation to workout & show C.Tax differences  Avg\_CTax\_Towns <- dbGetQuery(mydb, "SELECT  ROUND (Band\_A\_1 - Band\_A\_2, 2) AS 'Band A Difference',  ROUND(Band\_B\_1 - Band\_B\_2, 2) AS 'Band B Difference',  ROUND(Band\_C\_1 - Band\_C\_2, 2) AS 'Band C Difference'  --Figures above are derived from...  FROM(  SELECT  -- Calc average charge of each town per band  AVG(CASE WHEN LI.Ward\_Name LIKE 'Abingdon%' THEN Band\_A END) AS Band\_A\_1,  AVG(CASE WHEN LI.Ward\_Name LIKE 'Wantage%' THEN Band\_A END) AS Band\_A\_2,  AVG(CASE WHEN LI.Ward\_Name LIKE 'Abingdon%' THEN Band\_B END) AS Band\_B\_1,  AVG(CASE WHEN LI.Ward\_Name LIKE 'Wantage%' THEN Band\_B END) AS Band\_B\_2,  AVG(CASE WHEN LI.Ward\_Name LIKE 'Abingdon%' THEN Band\_C END) AS Band\_C\_1,  AVG(CASE WHEN LI.Ward\_Name LIKE 'Wantage%' THEN Band\_C END) AS Band\_C\_2  FROM Council\_Tax\_Information\_2023\_24 AS CI  JOIN Location\_Information AS LI  ON CI.Location\_Id = LI.Location\_Id  -- Define district constraint  WHERE LI.District\_Code = 'E07000180'  );")  print(Avg\_CTax\_Towns)  dbDisconnect(mydb) SQL CodesCreate Tables CodeLocation Information (Parent) CREATE TABLE "Location\_Information" (  "Location\_Id" INTEGER UNIQUE,  "District\_Name" TINYTEXT,  "District\_Code" VARCHAR(9),  "Ward\_Name" TINYTEXT,  "Ward\_Code" VARCHAR(9),  PRIMARY KEY("Location\_Id"),  UNIQUE("Location\_Id")  ); House Price Information (Child) CREATE TABLE House\_Price\_Information (  House\_Price\_Record\_Id INTEGER PRIMARY KEY AUTOINCREMENT,  Location\_Id INTEGER,  End\_Mar\_2021 INTEGER,  End\_Jun\_2021 INTEGER,  End\_Sep\_2021 INTEGER,  End\_Dec\_2021 INTEGER,  End\_Mar\_2022 INTEGER,  End\_Jun\_2022 INTEGER,  End\_Sep\_2022 INTEGER,  End\_Dec\_2022 INTEGER,  Band\_2021 CHAR(1),  Band\_2022 CHAR(1),  Council\_Tax\_Record\_Id INTEGER,  FOREIGN KEY(Council\_Tax\_Record\_Id) REFERENCES "Council\_Tax\_Information\_2023\_24" (Council\_Tax\_Record\_Id),  FOREIGN KEY(Location\_Id) REFERENCES Location\_Information (Location\_Id)  ); Council Tax Information (Child) CREATE TABLE "Council\_Tax\_Information\_2023\_24" (  "Council\_Tax\_Record\_Id" INTEGER UNIQUE,  "Location\_Id" INTEGER,  "Calculated\_as" CHAR(9),  "Band\_A" NUMERIC(6, 2),  "Band\_B" NUMERIC(6, 2),  "Band\_C" NUMERIC(6, 2),  "Band\_D" NUMERIC(6, 2),  "Band\_E" NUMERIC(6, 2),  "Band\_F" NUMERIC(6, 2),  "Band\_G" NUMERIC(6, 2),  "Band\_H" NUMERIC(6, 2),  FOREIGN KEY("Location\_Id") REFERENCES "Location\_Information"("Location\_Id"),  PRIMARY KEY("Council\_Tax\_Record\_Id")  ); Broadband Information (Child) CREATE TABLE "Broadband\_Information\_2023" (  "Broadband\_Record\_Id" INTEGER,  "Location\_Id" INTEGER,  "Calculated\_as" CHAR(10),  "Gigabit\_Availability" NUMERIC(4, 2),  "Superfast\_Availability" NUMERIC(4, 2),  "Average\_Download\_Speed" NUMERIC(5, 2),  PRIMARY KEY("Broadband\_Record\_Id" AUTOINCREMENT),  FOREIGN KEY("Location\_Id") REFERENCES "Location\_Information"("Location\_Id")  ); Code to Insert Unormalised CSV Data into Normalised TablesLocation Information INSERT INTO "Location\_Information" ("District\_Name", "District\_Code", "Ward\_Name", "Ward\_Code")  SELECT "Local\_authority\_name", "Local\_authority\_code", "Ward\_name", "Ward\_code"  FROM Unnormalised\_Clean\_Data; House Price Information INSERT INTO "House\_Price\_Information" ("End\_Mar\_2021", "End\_Jun\_2021", "End\_Sep\_2021", "End\_Dec\_2021", "End\_Mar\_2022", "End\_Jun\_2022", "End\_Sep\_2022", "End\_Dec\_2022", "Band\_2021", "Band\_2022")  SELECT "Mar\_2021", "Jun\_2021", "Sep\_2021", "Dec\_2021", "Mar\_2022", "Jun\_2022", "Sep\_2022", "Dec\_2022", "Band\_year\_2021", "Band\_year\_2022"  FROM Unnormalised\_Clean\_Data; Council Tax Information INSERT INTO Council\_Tax\_Information\_2023\_24 (Calculated\_as, Band\_A, Band\_B, Band\_C, Band\_D, Band\_E, Band\_F, Band\_G, Band\_H)  SELECT "Calculated\_as(ct)", B\_A, B\_B, B\_C, B\_D, B\_E, B\_F, B\_G, B\_H  FROM Unnormalised\_Clean\_Data; Broadband Information INSERT INTO Broadband\_Information\_2023 (Calculated\_as, Gigabit\_Availability, Superfast\_Availability, Average\_Download\_Speed)  SELECT "Calculated\_as(bb)", Gigabit\_avail, Superfast\_avail, "Avg\_download\_speed(Mbps)"  FROM Unnormalised\_Clean\_Data;  **Word Count:** 2,453 References: Cherwell District Council (no date) *Council Tax Charges 2023-2024*, *Cherwell District Council*. Available at: https://www.cherwell.gov.uk/directory/146/council-tax-charges-202324 (Accessed: 19 February 2023).  Hewings, S. (2023) *Council Report: Council Tax 2023/24*, *South Oxfordshire District Council*, pp. 13–14. Available at: https://democratic.southoxon.gov.uk/documents/s28767/South%20Council%20Tax%20Report%202023-24.pdf (Accessed: 9 March 2023).  Office for National Statistics (2023) *Median House Prices by Ward: HPSSA Dataset 37 - Year Ending March 2023 Edition of this Dataset*, *ONS*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/medianpricepaidbywardhpssadataset37 (Accessed: 20 February 2023).  Oxford City Council (2022) *Appendix 2 - Council Tax Charges per Band 2023-24, Item 95*, *Oxford City Council*. Available at: https://mycouncil.oxford.gov.uk/documents/s71940/Appendix%202%20-%20Council%20Tax%20Charges%20per%20Band%202023-24.pdf.  UK Parliament (2024) ‘Constituency Data: Broadband Coverage and Speeds’, *commonslibrary.parliament.uk* [Preprint]. Available at: https://commonslibrary.parliament.uk/constituency-data-broadband-coverage-and-speeds/ (Accessed: 11 March 2023).  Vale of White Horse District Council (no date) *Council Tax Calculator*, *Vale of White Horse District Council*. Available at: https://data.whitehorsedc.gov.uk/java/support/Main.jsp?MODULE=Calculator (Accessed: 12 March 2024).  West Oxfordshire District Council (2023) *Council Tax Charges 2023 to 2024*, *West Oxfordshire District Council*, pp. 1–3. Available at: https://www.westoxon.gov.uk/media/blmmubc5/wodc-council-tax-charges-2023-to-2024.pdf (Accessed: 9 March 2024). Appendixes: Appendix 1: Table of Variables   |  |  | | --- | --- | | **Variable** | **Description** | | Location Id (PK) | Autoincrementing value which associates tuples with their corresponding district and ward information | | District Name | Name of district (local administrative unit) | | District Code | UK district’s unique identity code | | Ward Name | Name of ward (electoral district at sub-national level) | | Ward Code | UK ward’s unique identity code | | House Price Record Id (PK) | Autoincrementing value which associates tuples with their corresponding house price information, location information, and council tax charge for 2023-24 | | End March 2021 – End Dec 2022 | Median house price of all types of properties, paid by ward in pounds sterling (£), as quartiles between March 2021 to December 2022 | | Broadband Record Id (PK) | Autoincrementing value which associates tuples with their corresponding broadband information and location | | Calculated as | Method used to calculate records’ values. Categorised into ‘Individual’, ‘Grouped’ or ‘Town’ labels. See, [*Data Selection and Cleaning*](#_Broadband) *(Broadband and Council Tax)* sections for a more detailed explanation | | Gigabit Availability | The percentage of premises capable of receiving speeds of 1 gigabit per second (House of Commons Library, 2023). | | Superfast Availability | The percentage of premises capable of receiving a minimum of 30Mbps download speeds. It does not assume that all lines receive superfast speeds as this is dependent on the package consumers independently subscribe to (House of Commons Library, 2023) | | Average Download Speed (ADS) | Data contains median download speeds for constituencies from Ofcom.  ADSs received based on the mean average. Reflecting both consumer choice and line quality. Users may opt for packages offering higher speeds than what they actually receive. Since this is a mean average, lines with faster speeds have a disproportionate impact on the overall average. (House of Commons Library, 2023) | | Council Tax Record Id (PK) | Autoincrementing value which associates tuples with their corresponding location information and council tax charge for 2023-24 | | Band A - H | A fixed tax fee in pounds sterling (£) assigned to dwellings, based on their property value | |
|  |