Summary Report

1.0 Introduction

The ceremonial county of Oxfordshire is located in the far west of the South East England statistical region. It borders Warwickshire to the north-west, Northamptonshire to the north-east, Buckinghamshire to the east, Berkshire to the south, Wiltshire to the south-west, and Gloucestershire to the west. The county of Oxfordshire is divided into five districts: Oxford, Cherwell, South Oxfordshire, Vale of White Horse, and West Oxfordshire. The county has six constituencies: Banbury, Henley, Oxford West and Abingdon, Oxford East, and Wantage.

This report summarises all the processes involved in data gathering, selection, and cleaning for designing an SQLite database system for the house price, council tax, and broadband speed in Oxfordshire. It also provides a report on legal and/or ethical issues of the proposed system, data model and implementation of the proposed system, and a comparison between structured and semi-structured data models. The data set used in implementing the SQLite database designed for the coursework is the house prices for two districts (Cherwell and West Oxfordshire), council tax from thirty-nine towns/parishes from Cherwell and West Oxfordshire, and the broadband speed for all constituencies within Cherwell and West Oxfordshire districts .

1.1 Data gathering

The data sources for the house price, council tax, and broadband speed are secondary data sources. The house price data was collected through the Office of National Statistics (ONS) website. This data provides the “Median price paid by ward, England and Wales, year ending Dec 1995 to year ending Dec 2020” (Excel Sheet 1a). The council tax data for the five districts in Oxfordshire was collected through the Oxfordshire county council website. The data contains the council tax charges by tax band for various parishes and towns in any district. The broadband speed data for the constituency in Oxfordshire was sourced through the UK parliament website. The data contains broadband connectivity and speed for different parts of the United kingdom based on the Ofcom data.

1.1 Data Selection and Data Cleaning Description

The house price data collected from the Office of National Statistics website was stored in a Microsoft Excel file format (xlsx). The data contains house prices from December 1995 to March 2020 for various districts in England and Wales. For this coursework, I utilised the filter options available in the Microsoft Excel package to select house prices from 2018 to 2020 for two districts (Cherwell and West Oxfordshire) out of the five districts in Oxfordshire and stored them in a new Microsoft Excel document. The Local Area Code (LAC) attribute in the house price data was used to identify the wards and house prices belonging to Cherwell and West Oxfordshire districts.

The council tax data for nineteen towns/parishes within Cherwell and twenty towns/parishes within West Oxfordshire, making it a total of thirty-nine towns, were sourced from the Cherwell district website and the West Oxfordshire district website. The data were collated and stored in a Microsoft Excel document. To look up which ward a particular town/parish belongs to, I use the UK postcode website to allocate the selected thirty-nine towns/parishes to their respective wards within Cherwell and West Oxfordshire.
The broadband speed data collected from the UK parliament website was stored in a Microsoft Excel file format (xlsx). The data contains broadband speed for different constituencies in the United Kingdom. I utilised the filter options available in the Microsoft Excel package to select the average download speed and superfast availability data for the constituencies within the Cherwell district and West Oxfordshire districts. To achieve this, I used a lookup dataset from the Office for National Statistics website to help me identify the various consistencies in the Cherwell and West Oxfordshire districts. I also used the lookup dataset to lookup MSOA code and Ward code in Cherwell and West Oxfordshire district. The lookup dataset assisted me in marching each MSOA code to its wardcode which helped me create a link between the broadband speed for each ward with Cherwell and West Oxfordshire.
All data cleaning processes were done manually using Microsoft excel. There was no need to worry about missing data because all the data were complete for the dataset used for this coursework.

2.0 Data Model and Implementation

2.1 Database Normalisation

The dataset was normalised up to the third normal form (3NF) with an appropriate definition of primary keys and foreign keys to store the house price data, council tax data, and broadband data into an SQLite database. To achieve 3NF, the house price table, council tax table, and the broadband table were further separated into six tables (District, BroadBand, HousePrice, CounciltaxData, Townward, and Constituency). Below is a description of the various tables in the database to hold house price data, council tax data, and broadband data.

2.1.1 District Table

This table is created by separating Local Authority Code (LAC) and Local Authority Name (LAN) from the house price data with LAC as the primary key. It is labeled in the database as DistrictTable.

2.1.2 BroadBand Table

This table is created from the broadband data. It holds the Middle Layer Super Output Code (MSOACD11), Middle Layer Super Output name (MSOA_name), Average download speed, Super availability, Constituency code, and Ward code in Cherwell and West Oxfordshire district. The table has the primary key as

MSOACD11 and the foreign key as ward code linked to the house price table. This table is labeled in the database as BroadBand.

2.1.3 Towns Ward Table

This table is created to link the council tax table to the house price table. The table holds the towns/parishes in Cherwell and West Oxfordshire, Ward name, and ward code to which each town/parish belongs. The table has a primary key as Towns_parishes and a foreign key as ward code reference to the house price table. This table is labeled in the database as TownsWard.

2.1.4 Council Tax Table

This table is created to hold the council tax by tax band for the different towns/parishes within Cherwell and West Oxfordshire. The table has the primary key as Towns_parishes and is linked to the townward table. This table is labeled in the database as CouncilTaxData.

2.1.5 House Price Table

This table holds the quarterly house prices for the different wards within Cherwell and West Oxfordshire from 2018 to 2020. The table has the primary key as ward code and the foreign key as local area code reference to the district table. This table is labeled in the database as HOUSEPRICE.

2.1.6 Constituency Table

The constituency table is created to link the broadband table and the district table. The table holds the constituency code, constituency name, region name, and local area code. The table has a primary key as constituency code and foreign key as local area code reference to the district table. This table is labeled in the database as Constituency.

2.2 Database Schema

The figure below shows the database schema for the SQLite database to store the house price data, council tax data, and broadband data. The database schema displays the relationship that exists between the various tables in the database. It also helps identify the data types used for each attribute in any table and the primary and foreign keys in each table in the database.

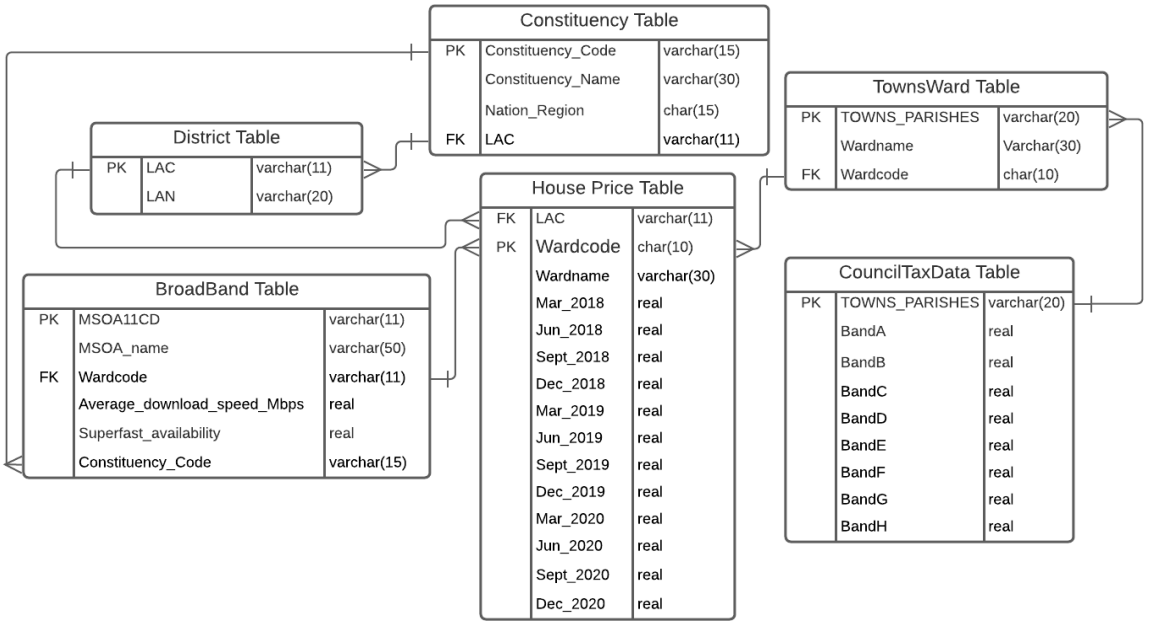

Figure 1: House Price, Council Tax and Broadband Speed Database Schema

2.3 Database creation and population

The database creation which includes creating tables, defining data type, primary and foreieng keys was done using SQLite software (see appendix for SQL code) and population of the database was done using R programming language.

2.4 Main steps of how R code is to be run

The first step is to read the normalised dataset stored in Microsoft Excel format (DataScience FoundationNEW (1).xlsx) into R studio. The data to be stored for each table in the database is stored as a sheet in the Microsoft Excel file and read into R as a data frame. The second step is to connect to the database created using SQLite and populate the database with the normalised data stored as data frames in R. To populate the database, I used the dbwriteTable function with append equals true. The third step is to design the R code to implement the tasks outlined in the coursework brief. All the R codes required to implement that task state in the coursework brief have been tested and implement all the required tasks correctly. To implement the coursework task in R, I used some packages like tidyverse, readxl and RSQLite.

For the first task, I found 2018, 2019 and 2018 to 2019 average house prices for Banbury Cross and Neithrop ward in Cherwell district. The R code to implement this task has been tested to perform the task accurately. In the second task, I found that the average increase in house prices from 2018 to 2019 for Banbury Cross and Neithrop ward in Cherwell district is approximately 2.7%. In task three, I found the maximum house price for the last quarter of 2019 (Dec 2019) and Alvescot and Filkins ward in West Oxfordshire district had the highest house price with 55500

pounds. The first three tasks required me to use the INNER JOIN method on HOUSEPRICE and DistrictTable, which are different tables in my database.

In task four, I found the broadband speed (average download speed) for Banbury Cross and Neithrop ward in Cherwell district to be 42.51833. This task was achieved by joining three different tables: HOUSEPRICE, DistrictTable and BroadBand. The R code has been tested and implemented correctly.

In task five, I designed a code to retrieve the average broadband speed by ward. This is because each ward may have one or more broadband speeds. The result from this query outlines the average broadband speed for any ward in our database. To achieve this, I used the INNER JOIN method of two tables HOUSEPRICE and BroadBand.

In task six, I found the average council task (Band A to Band C) for Langford town in Alvescot and Filkins ward in West Oxfordshire district. I joined three different tables (HOUSEPRICE, TownsWard and DistrictTable) to achieve this task.

In task seven, I found the Band A council tax difference between Alvescot and Bampton in West Oxford District. To achieve this, I created a virtual table to perform this task. The R code to implement this task has been tested and implemented correctly.

3.0 Structured and semi-structured data

In the current era of big data, in which tremendous amounts of data are generated every second, it is still not clear which data model is best suited for handling such massive volumes of data. The term structured data refers to data that has been predefined and formatted to a set structure before it is stored, also known as schema-on-write. The data is easy to enter, store and analyse. Structured data are stored in rows and columns, a format managed by a programming language called Structured Query Language (SQL)(Groff, Weinberg and Oppel, 2002). Structured data is handled in a relational model, which manages it in the form of rows and tables, thus enabling the content of a table to be processed efficiently. Although on-write-schema data definition offers several advantages for structured data, this limits its flexibility and use cases since it can only be used for the purpose for which it is predefined. With structured data, data managers have access to more tools that have been tried and tested. They can choose from more products when using structured data (Praveen and Chandra, 2017).

The term semi-structured data refers to data that does not fit into a relational database. It has no fixed schema but has some structural properties or a loose organisational structure. Emails, XML, and JSON are examples of semi-structured data. Although XML data cannot be restricted by having a fixed database schema, there is difficulty in interpreting the relationships between the data and the schema. Querying XML data is less efficient than querying structured data. XML syntax is

verbose and redundant which leads to higher costs associated with storing and transporting data.

In order to manage the house price, council tax, and broadband data efficiently and effectively, I propose the use of SQL over XML. Although SQL has its major drawback in having a fixed database schema, it is preferred over XML because it allows complex queries and large amounts of data to be retrieved quickly and efficiently. In SQL, the system automatically updates the related database table which is not the same for XML. SQL allows for relationships to be created between the various tabel which XML does not support.

4.0 Legal and/or ethical issues

The house price, broadband speed, and council tax data are all online open-source data that contain public sector information licensed under the Open Government Licence version 3.0. The data do not contain personal information or any exemption that the open government licence does not cover. All data processing with the data is done lawfully without any manipulation to falsify the data.
The house price data is sourced through the Office of National Statistics website. The data contains the Median price paid for residential property in England and Wales by property type and electoral ward. Annual data updated quarterly. The data cannot be sold but can be used for academic purposes, this does not present any ethical or legal issues. Similarly, Council tax data is sourced from the Oxfordshire county council website. The data hold council tax by band for different towns and parishes in Oxfordshire. This data does not pose any legal or ethical issues as it is legally and ethically allowed to be used by educational establishments (click here). The broadband speed data is sourced from the UK parliament house of commons library website. It contains broadband connectivity and speeds for different parts of the UK, based on Ofcom data. It does not contain any personal information and does not pose any ethical or legal issues. Since all the data is open-source data, it does not present any legal or ethical issues.

Reference

Groff, J.R., Weinberg, P.N. and Oppel, A.J., 2002. *SQL: the complete reference* (Vol. 2). McGraw-Hill/Osborne.

Praveen, S. and Chandra, U., 2017. Influence of structured, semi-structured, unstructured data on various data models. *Int. J. Sci. Eng. Res*, *8*, pp.67-69.

Appendix

Link to data sources

Office of National Statistics (ONS) website -https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/medianpricepaidbywardhpssadataset37

Oxfordshire county council website - https://www.oxfordshire.gov.uk/council/about-your-council/government-oxfordshire/district-councils

UK parliament website. - https://commonslibrary.parliament.uk/constituency-data-broadband-coverage-and-speeds/

House price data collected from the Office of National Statistics website - https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/medianpricepaidbywardhpssadataset37

Council tax data for nineteen towns/parishes within Cherwell district - https://www.cherwell.gov.uk/directory/22/council-tax-charges-2021-2022

Council tax data for nineteen towns/parishes within West Oxfordshire district- https://www.westoxon.gov.uk/council-tax-and-benefits/council-tax-bands-charges-and-appeals/

The broadband speed data collected from the UK parliament website- https://commonslibrary.parliament.uk/constituency-data-broadband-coverage-and-speeds/

Lookup dataset from the Office for National Statistics website- https://geoportal.statistics.gov.uk/datasets/middle-layer-super-output-area-2011-to-ward-to-lad-december-2020-lookup-in-england-and-wales-v2/explore

SQL Code

```
-- create the district table

CREATE TABLE DistrictTable(

LAC varchar(11) NOT NULL PRIMARY KEY,

LAN varchar(20));

-- create the house price  table

CREATE TABLE HOUSEPRICE(
```

```
LAC varchar(11) NOT NULL,
Wardcode char(10) NOT NULL PRIMARY KEY,
Wardname varchar(30),
Mar_2018 real,
Jun_2018 real,
Sep_2018 real,
Dec_2018 real,
Mar_2019 real,
Jun_2019 real,
Sep_2019 real,
Dec_2019 real,
Mar_2020 real,
Jun_2020 real,
Sep_2020 real,
Dec_2020 real,
foreign key (LAC)REFERENCES DistrictTable(LAC));
-- create the townsward table
CREATE TABLE TownsWard(
TOWNS_PARISHES varchar(20) NOT NULL PRIMARY KEY,
Wardname varchar(30),
Wardcode char(10),
foreign key (Wardcode)REFERENCES HOUSEPRICE(Wardcode));
-- create the constituency table
CREATE TABLE Constituency(
Constituency_Code varchar(15) NOT NULL PRIMARY KEY,
Constituency_Name varchar(30),
Nation_Region char(15),
LAC varchar(11),
foreign key (LAC)REFERENCES DistrictTable(LAC));
 -- create the broadband table
CREATE TABLE BroadBand(Constituency_Code varchar(15),
```

```
MSOA11CD varchar(11) NOT NULL PRIMARY KEY,
MSOA_name varchar(50),
Wardcode varchar(11),
Average_download_speed_Mbps real,
Superfast_availability real,
foreign key (Wardcode)REFERENCES HOUSEPRICE(Wardcode));
-- create the council tax table
CREATE TABLE CouncilTaxData(
TOWNS_PARISHES varchar(20) NOT NULL PRIMARY KEY,
BandA real,
BandB real,
BandC real,
BandD real,
BandE real,
BandF real,
BandG real,
BandH real,
foreign key (TOWNS_PARISHES)REFERENCES TownsWard(TOWNS_PARISHES));
```