Oxford Brookes University

SUBJECT

Data Science Foundations

Coursework Assignment

Authors:

Nikolaos Alexoudis (19178012)

December 8, 2021

1) Data Selection & Cleaning.

The data used for this research came from individual resources. For the median house prices, data from the "Data from Office of National Statistics (ONS): "Median price paid by ward, England and Wales, year ending December 1995 to year ending Dec 2020" (Excel Sheet 1a)" found on:

"https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/medianpricepaidbywardhpssadataset37"

Where data for the years 2000 to 2020 were used, divided in quarterly averages. The data was all ready to use from the start and we chose to use only 20 years worth of data since this amount satisfies are queries.

Broadband data were supplied from the "House of Commons Library" site's, "Broadband connectivity and speeds" table in excel. we only use four attributes called wardcode, average download speed, connections receiving super-fast speeds and connections receiving over 300 mbps speed for the purposes of this table. The data is found on:

"https://commonslibrary.parliament.uk/constituency-data-broadband-coverage-and-speeds/postcode"

Council tax data was supplied from the "Oxfordshire County Council" official website. There five individual links (one for each district) lead to sites where data about the tax on individual tax bands, per different town are supplied. The local authority code was attached manually to each town, since there are only five of them. For the purposes of this research only 64 entries were considered out of the total. The initial link to the council's site is:

"https://www.oxfordshire.gov.uk/council/about-your-council/government-oxfordshire/district-councils"

The data for the ward table was found on "Open Geography portalx", were after the data was copied to an excel spreadsheet, the data about our districts of concern were filtered out. This table's sole purpose is to be used in a referential way to all the other tables. This is because the table's structure easily allows us to identify each ward to a district. The link to this site is:

"https://geoportal.statistics.gov.uk/datasets/e169bb50944747cd83dcfb4dd66555b1/explore"

The initial cleaning process happened in "excel spreadsheet". For each individual table the column names were changed to fit the purposes of this project using "DB Browser (SQLite)". This means that SQL reserved symbols were replaced(i.e. dots for hyphens). For data on broadband, only the columns that could be used for the queries listed in the coursework's description and plus one of my choosing, were used. The data for council tax were manually loaded in "DB Browser", since no collective file could be found. The above reason justifies the comparatively small size of the mentioned table.

2) Legal Issues

The data we used are all publicly available from government maintained websites. The reason this data is available is for the government to provide transparency on the well being of its citizens and on its taxing policy.

The data about the house prices are all maintained by ONS (Office of National Statistics) which itself reports directly to the United Kingdom's Parliament. It is clear in the terms of the mentioned data that all records lie under Open Government Licence which means that they all are available for personal use. The validity of the data is guaranteed by the organisation which collected the data. Since the ONS is the largest statistical producer in the country it is safe to assume that the data is valid.

The data about broadband speed was provided by the UK's Parliament site on the section of constituency data. The Parliament itself got its data from ofCom which in their site they mention that the data is protected from the freedom of information act of 2000. This means that this data can be publicly used. The fact that this data is provided from a public agent that constitutes a legitimate source means that its accuracy is upheld.

Data about council tax is provided by Oxfordshire County Council which is committed to the highest standards of quality which is exactly what is expected from a public provider. The issue that arises with intellectual property rights in this case though, is the fact that we can only use council tax data from this source only and only if we do not make a financial gain out of it. For the purposes of this research we do not plan to use any of the listed information for financial gain.

The information used in the ward table is collected from Open Geography Portal X. Their data can be trusted since they are taken from official data from ONS geography. This set of data is also protected from the open government licence on data sharing.

The vagueness of the data, on the dimension of specificity on the whereabouts of each unit is also detrimental to the legal use of this data. Since no more specific clue than town (i.e. postcode or address) was used, each individual's personal information is protected.

3)Structured and Semi-Structured Data

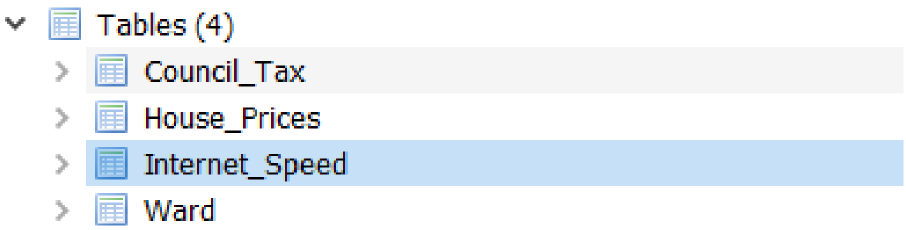
Structured database types like SQL, store data in tables which are usually defined with a proper schema so that helpful relationships and input constraints can be applied. Relational tables allow for functional dependency, meaning that one column or attribute in one table uniquely identifies another column in the same table. This allows for a referential ability between tables that share an attribute. We can use that attribute, usually called a primary key to one table and a foreign key to another, to reference between attributes belonging to different tables.

On the other hand, semi structured database types, like XML, have no need of proper pre-defined schema and are purposefully designed to be understood by humans and computers with relative ease. We can also witness from the name that XML (eXtensible Markup Language) can be extended to allow user defined tags in natural language. This adds to the readability of XML and makes for an ideal choice even for beginners in the world of database building. The way XML files are being structured allows for an easy merger of different sources into one. The above benefit combined with the easily interpreted representation make for a very simplified way for information to be passed around businesses.

However XML has its limitations. Its redundancy makes the final version of the data to have a very large markup on labour, storage and transmission costs. Most crucially related to the purposes of this project the non normalisation of data makes XML unsuitable for this project. Since we need normalised tables that are related through primary keys so we can retrieve information from one table using an attribute from another to satisfy our given queries. We can safely conclude that SQL is the most suitable approach for the creation of the databases used in this project.

4)Data Model and Implementation

The tables are normalized to 3NF meaning no duplicates and appropriate keys are given.



Tables (4)
- Council_Tax
- House_Prices
- Internet_Speed
- Ward

Figure 1: Data Set Tables

Council_Tax contains information about council tax for each town, hence the primary key we arranged is the name of each town in the column “Town”. Bands A to H are all in integer type while L_AuthCode (local authority code) and Town are in text type.

House_Prices contains quarterly averages of house prices for each ward in the span of twenty years (2000-2019). The primary key is hence the Wardcode. Wardcode is in text type, while YearendingMar2000-YearendingDec2019 are all integer.

Internet_Speed contains information about the average download speed, the percentage of connections receiving super fast speeds and the percentage of connections receiving speeds above 300 mb/s. Wardcode is again the column we use as a primary key here. Wardcode is again text while all the other values are in integer type.

The ward table contains the columns Wardcode, WardName, L_AuthCode and L_AuthName. Those column names contain the ward code, ward name, local authority code and local authority name. All the columns are of course in text type in this table.

5)R code and Justifications

The libraries I used for this assignment are haven, dplyr, dbplyr, RSQLite and DBI. I connected DB Browser and R Studio by the use of

```
conn <- dbConnect(SQLite(),"DF3.db")
```

command, where DF3.db is my final database.

0.1 R Code and Database Queries:

Q1: To fetch the eight quarterly averages from the house data we make use of the command:

```
q1 <- dbGetQuery(conn, ’SELECT YearendingDec2015, YearendingMar2015, YearendingJun2015, YearendingSep2015, YearendingDec2016, YearendingMar2016, YearendingJun2016, YearendingSep2016 FROM House_Prices INNER JOIN Ward ON Ward.Wardcode = House_Prices.Wardcode WHERE Ward.WardName = "Thame"’)
```

Where we get eight individual prices for the ward of Thame for years 2015 and 2016. The inbuilt function mean() gives as the mean of the two years.

Q2: To find the increase between the house prices of the ward Thame between 2015 and 2016, we use the above data points and we separate them per year. The initial four go to the mean of 2015 and the latter four go to the mean of 2016 (we make use of the mean() function). By dividing initial average price by final, subtracting one and multiplying by one hundred we get our result. We find out that prices increased 22.971% from the year 2015 to 2016 in the ward of Thame.

Q3: To find out the highest average house price for spring 2004 in the county of Oxfordshire we use the following command:

```
q3 <- dbGetQuery(conn, ’SELECT WardName FROM Ward INNER JOIN House_Prices ON Ward.Wardcode = House_Prices.Wardcode WHERE House_Prices.YearendingMar2004 = (SELECT max(YearendingMar2004) FROM House_Prices)’)
```

If we print the variable we will see the highest priced ward in terms of house prices is Holywell for spring 2004.

Q4: To find the broadband speed for Bicester West we use the following command:

```
q4 <- dbGetQuery(conn, ’SELECT Avg_DownloadSpeedMbs FROM Ward INNER JOIN Internet_Speed ON Ward.Wardcode = Internet_Speed.Wardcode WHERE Ward.WardName = "Bicester West"’)
```

The answer is 62mb/s.

Q5: The query we choose to make, asks “what is the ward with the highest percentage of super fast seeds?”. This question is essential for someone that wants to move somewhere where it is relatively easier to get a high speed connection. To find the answer we use the following command:

```
q5 <- dbGetQuery(conn, ’SELECT WardName FROM Ward INNER JOIN Internet_Speed ON Ward.Wardcode = Internet_Speed.Wardcode WHERE Internet_Speed.ConnectionsReceivingSuperfastSpeeds = (SELECT max(ConnectionsReceivingSuperfastSpeeds) FROM Internet_Speed)’)
```

The answer is Marston.

Q6: To find the average between bands F,G and H in Littlemore we use the following command:

```
q6 <- dbGetQuery(conn, ’SELECT Band_F, Band_G, Band_H FROM Council_Tax WHERE Council_Tax.Town = "Littlemore"’)
```

Identically to query number one we use the combination of mean() and as.numeric() to find the answer. The mean of bands F,G and H for the city of Littlemore is £3695.727.

Q7: To find the difference between the charges in band A council tax between Claydon and Prescote of the district Cherwell we make use of the following commands:

```
q9a <- dbGetQuery(conn, ’SELECT Band_A FROM Council_Tax WHERE Council_Tax.Town = "Claydon"’)

q9b <- dbGetQuery(conn, ’SELECT Band_A FROM Council_Tax WHERE Council_Tax.Town = "Prescote"’)

claydonA <- as.numeric(q9a[1,1])

prescoteA <- as.numeric(q9b[1,1])
```

We simply subtract the variables claydonA and prescoteA to find the answer. Claydon’s Band A is £53.53 more expensive from Prescote’s Band A.

1 6) Testing.

The testing of this system will be conducted in the video file, contained in the ZIP folder this report occupies.