COMP6037: Foundations of Data Analytics

Coursework 1: Semester 1, 2024-25

School of Engineering, Computing and Mathematics
Oxford Brookes University

Introduction

This is individual coursework, so an independent submission is expected from each student.

This coursework is worth 40% of the overall module marks.

In this coursework, you will prepare data models and analyse data using data collected from different sources. You will develop a software system which users can use to find information about house prices and council tax charges in different areas of Oxfordshire.
There are multiple districts (or local authorities) in Oxfordshire which include City of Oxford, Cherwell, South Oxfordshire, Vale of White Horse and West Oxfordshire. Each of the district (local authority) has different areas or wards. For example, City of Oxford has different wards such as Barton and Sandhills, Summertown, Headington, Cowley, etc. Similarly, Vale of White Horse district includes wards such as Abingdon Abbey Northcourt, Cumnor, Faringdon, and so on.

Data sources

You will need to collect data about house prices and council tax charges using different data sources. The data should be cleaned (manually or automatically) and be stored in a SQL database (see the marking scheme).

You should use datasets that are published on by the UK government, either centrally or through a public body that would be available to a member of the UK public. You can find different data sources. Some of the examples data sources are given below.

1) Data from Office of National Statistics (ONS):
“Median price paid by ward, England and Wales, year ending Dec 1995 to year ending Dec 2022” (Excel Sheet 1a)
https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/medianpricepaidbywardhpssadataset37

2) HM Land Registry n.d, HM Land Registry Open Data:
Price Paid Data,
http://landregistry.data.gov.uk/app/ppd

3) Council Tax data

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

Examples:

Cherwell district: Council tax charges by tax band for various parish and Town council areas
https://www.cherwell.gov.uk/directory/146/council-tax-charges-202324

Oxford City Council tax charges:
https://www.oxford.gov.uk/info/20152/council_tax_bands_and_charges/120/council_tax_charges

Tasks

Using the data (collected from data sources), you should produce a unified data set and model that could be used to develop the system. You must ensure that all data used is normalised to 3NF (3rd Normal Form). You should use SQL database system (SQLite) to store and process the data. You must demonstrate that you can query the data set in R. i.e., data stored in SQLite database is queried in R (as specified below).

You are required to write a report and explain all the processes (and stages) that you undertake in order to collect, clean and structure data as well as implement the system (see the marking scheme for details).

In order to implement the system, the following tasks are to be completed.

SQL database:

1. Normalise the data to 3NF and store it in SQL database tables. The data should be stored (represented) in multiple tables.
2. Define appropriate keys, data types and relationships between the tables.

Write R code in order to implement the following SQL database queries:

3. For a given ward in a particular district (e.g., City of Oxford, Cherwell, etc.) calculate the average price of houses in two years, e.g., average of prices for 2021 and 2022.
   Note that each year prices are given as (or divided into) quarters such as Mar 2021, Jun 2021, Sep 2021, Dec 2021. These need to be taken into account when calculating average price for two years.

4. Considering a particular district in Oxfordshire, find a ward which has the highest house price in a particular (quarter of a) year, for example, Mar 2021 or Dec 2219.
5. Based on the data about council tax charges, write a query that calculates an average council tax charge for a particular town in a particular district for any three bands of properties. (e.g., council tax charges for the Banbury town of Cherwell district are as follow: Band A = £1,376.67; Band B = £1,606.11; Band C = £1,835.56 and so on).
6. Different towns may have different council charges for a particular band of properties. For instance, Barford and Bicester (of Cherwell district) have respectively £1451.42 and £1517.34 council tax charges for Band A properties. Write a query that calculates the difference between council tax charges of same bands but of two different towns of the same district.
7. Considering a particular district (e.g. Cherwell), find a town which has the lowest council tax charges for Band B properties.

   [Note: The above council tax charges are used as examples. They may not be up to date]

Marking scheme

Data selection and cleaning

- Describe the stages (steps) you took to identify (or search), obtain, clean, and use the data sets related to house prices, broadband and council tax charges. Give a justification of the approaches (or methods) used in identifying, obtaining, cleaning, and using the datasets. Data selection and cleaning processes should follow appropriate quality criteria. **[10 marks]**

Structured and semi-structured data

- In this coursework you are storing data in a SQL database. However, it is possible to use other data tools (such as XML, etc) to represent and store such data. You should describe the structured data model (SQL) and semi-structured data model (XML) and give your suggestion – whether it should use SQL or XML or both. You should give clear reasons in order to justify your suggestions. **[4 marks]**

Data model and implementation

SQL database

- Appropriate design of SQL database tables – i.e., Normalisation of data to 3NF; Definition of correct definition of keys, relationships, and data types. Explain the normalistion steps, i.e. how un-normalised data is converted to normalized data. **[6 marks]**

R Code

- Design of R code: Good structure, use of comments; explanation of the main steps of how your R code is to be run and used **[5 marks]**
- Execution and testing of R code – Test and explain that queries (using R and SQL) return correct results**. [15 marks** (3 marks for each tasks 3-7) **]**

Total = 40 marks (40% of the module marks)

Submission details

The final submission date is: **1:00 PM, 9 Dec 2024 (Week 12).**

- Submit coursework via Moodle (Turnitin). You must
  - Submit an electronic copy of your R code and SQL database
  - Submit an electronic copy of report as detailed above
  - Submit a recorded video of the demo of the system you've developed. It should demonstrate that the system performs the required tasks. The length of the video must be limited to 3 minutes.

- Report should be limited to 2000 words. R and SQL code are not included in the page count.
- Final marks and feedback will be given after the exam committee.

Hints:

- You have to do the core part in R and SQL.
- The data cleaning can be done in whatever you want (manual or automatic) as long as you say in report what it is and how you've clean the data
- You might want to look up wikipedia for the districts in Oxfordshire - from the districts you can find out the major towns
- The ONS has some nice tools to help with things like converting postcodes to MSOAs, electoral wards etc. at their Open Geography Portal (http://geoportal.statistics.gov.uk/)

Learning Outcomes

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
2. Critically evaluate the methodologies applied in data collection, data processing, data analysis & dissemination of research findings
3. Critically assess methods and data strengths and limitations combined to application of R

Referencing

You are required to cite the work of others used in your solution and include a list of references (use the university recommended referencing style) and must avoid plagiarism and collusion.

https://www.brookes.ac.uk/library/library-services/information-skills/citing-references-in-your-work-and-plagiarism/