Faculty of Technology, Design & Environment

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**:**	5:00 PM, 18 Dec 2023

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
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

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:

Statement of Compliance *(please tick to sign)*

School of Engineering, Computing & Mathematics

☑ 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*)

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.

1 Data Selection and Cleaning

In the initial phase of the data science process, data acquisition played an important role in addressing the specific problem at hand. This endeavour required proficient database management skills, particularly when dealing with unstructured data. Utilizing renowned database management systems such as MySQL, SQLite, and MongoDB proved instrumental in handling unstructured datasets. Collecting and acquiring data can frequently lead to errors such as missing values, typos, inconsistent formats, duplicated entries for the same real-world entity, and breaches of both business and data integrity rules (Hu, J., 2021).

For the crucial task of data cleaning, the raw dataset underwent meticulous scrubbing using the R programming language. This step aimed to ensure the integrity and reliability of the data, preparing it for subsequent exploration and analysis.

The data collection process involved obtaining datasets from reputable sources, namely The United Kingdom Parliament and The Office for National Statistics (ONS). Notably, the broadband speed dataset originated from the House of Commons Library of the United Kingdom Parliament, a trusted repository known for its impartial analyses and statistical research. Similarly, council tax datasets were sourced from District Council, the government website of United Kingdom. The ONS assumes the responsibility of promoting and safeguarding the publication of official statistics for the public good.

This comprehensive approach to data collection and cleaning sets the foundation for subsequent stages in the data science process.

House Median: The data on housing prices obtained from below source.

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

Broadband: Dataset downloaded from House of Common Library

Constituency data: broadband coverage and speeds (Published June, 2022)

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

Council Tax: The dataset of Council tax is downloaded from United Kingdom Government website of District councils - Oxfordshire

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

Next, clean the data to ensure its validity. This involves checking for missing values, identifying invalid entries, and correcting data range issues. In this process need to replace invalid entries with suitable default values to make them usable. Additionally, some columns might be merged or split for better organization. According to the Figure 1, Data needs to get imported, merged, cleaned, standardized, verified, and deduplicated before being exported for analysis or use.

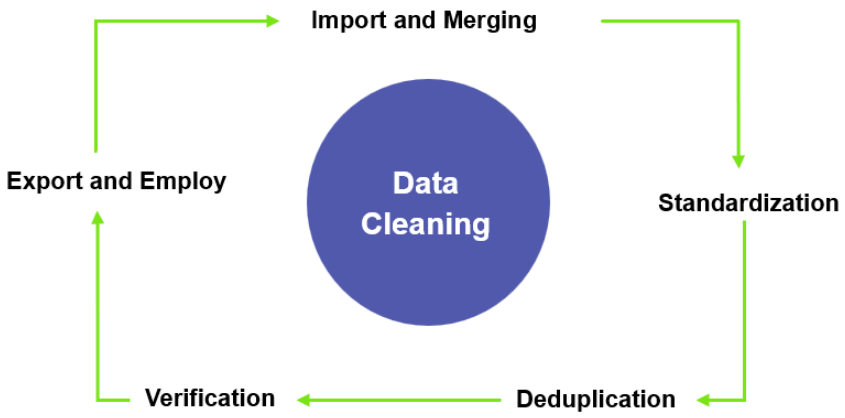

Figure 1: Data Cleaning Process

Furthermore, each file contains redundant information that needs to be removed before importing it into the database. Therefore, the data cleaning process will be split into two distinct stages for this assignment

a) Excel:

In the beginning, we will make use of Excel, a software tool, to choose the specific sheets or pieces of information that are necessary for the task.

House Median: In house median data set, I used Excel Sheet 1a and delete the initial 5 unwanted rows. I selected the data from Year ending Mar 2019 to Year ending Dec 2022.

Broadband: I have processed the broadband dataset by removing unnecessary rows and updating the column names. The revised column names are as follows: 'Ward_Code,' 'Superfast_Availability,' 'Average_Download_Speed,' and 'Receiving_under_10_Mbps.'

Council Tax: I have gathered data pertaining to Council Tax for various districts in Oxfordshire and subsequently renamed the column headers. The updated column names are now Band_A, Band_B, Band_C, Band_D, Band_E, Band_F, Band_G, and Band_H. Additionally, the dataset now includes the extracted ward codes.

District: The district dataset has been derived from the House median dataset, encompassing two columns originally labelled Local Authority Code and Local Authority Name. Subsequently, these column names have been updated to District_Code and District_Name, respectively.

Ward: This dataset is also derived from the House median price dataset includes two columns Ward_Code and Ward_Name. District_Code column has been added to perform the task.

b) R Programming:

The next phase involves utilizing the R programming language to refine and reshape the data, ensuring its accuracy by addressing null values and adjusting the data format as required.

I imported all the necessary files and conducted a process to eliminate duplicate entries, ensuring the dataset's cleanliness and integrity.

```
library(haven)
library(dplyr)
library(odbc)
library(RSQLite)
library(DBI)


#----------------------house_price_data-------------------------
house_price_data <- read.csv("D:/Study Material/assignment data/Daatabase -SQL/SQL_sonam/SQL project/HousePrice.csv")

#----------------------broadband-------------------------------
broadband <- read.csv("D:/Study Material/assignment data/Daatabase -SQL/SQL_sonam/SQL project/BroadBand.csv")

#----------------------council_tax-----------------------------
council_tax <- read.csv("D:/Study Material/assignment data/Daatabase -SQL/SQL_sonam/SQL project/CouncilTax.csv")

#----------------------district--------------------------------
district <- read.csv("D:/Study Material/assignment data/Daatabase -SQL/SQL_sonam/SQL project/District.csv")

#----------------------ward------------------------------------
ward <- read.csv("D:/Study Material/assignment data/Daatabase -SQL/SQL_sonam/SQL project/Ward.csv")
```

```
# Remove duplicate
district <- district %>% distinct(District_Code, .keep_all = TRUE)
```

2 Structured and semi-structured data

In the realm of data analysis, the difference between structured and semi-structured data plays a pivotal role in shaping the efficiency of processes.

Structured Data:

Structured data epitomizes a well-defined format with discernible relationships between data points. This format allows for the easy identification of individual records, fostering efficient analysis. Typically stored in databases like SQL, structured data is organized into tables featuring rows and columns. Each row represents a record, and each column signifies an attribute of that record. The inclusion of identification keys, such as primary keys, facilitates the straightforward mapping of records to specific fields. SQL stands out as the most common example of structured data storage.

Semi-structured Data:

In contrast, semi-structured data deviates from the rigid structure of relational databases. Semi structured data is characterized by variations in attribute presence, occurrence, and types across objects, including missing attributes and diverse representations of related information (Suciu, D., 1998). This characteristic makes semi-structured data more amenable to analysis than its unstructured counterpart. While it does not strictly adhere to the organizational norms of structured data, semi-structured data still maintains a certain degree of order. XML serves as a popular format for semi-structured data, employing tags to represent data elements and attributes for additional information.

Key Differences and Considerations:

Structure: Structured data is characterized by a fixed schema with pre-defined relationships, offering a clear advantage in terms of organization. On the other hand, semi-structured data embraces a more flexible structure featuring tags and hierarchies.

Analysis: Structured data is generally more straightforward to query and analyse using traditional database tools. In contrast, semi-structured data, while still analytically accessible, may require specialized tools or techniques due to its flexible nature.

Storage Efficiency: Structured data, with its rigid structure, often demands more storage space. In contrast, semi-structured data excels in storage efficiency by leveraging nested structures and hierarchies to optimize space utilization.

Understanding the difference of both structured and semi-structured data is essential for making informed decisions during the data analysis process. Based on the information and understanding of the data, I am using SQL for this assignment.

3 Legal and Ethical Issues

According to legal standards and acts is imperative when collecting and utilizing data for the analysis process, with the primary goal of safeguarding personal information. Researchers should observe the following guidelines during data collection for research purposes:

Data containing personal information should only be collected if it is pertinent to the research process. Whenever possible, researchers should avoid including personally identifiable information in the dataset. Prior to using data for research, researchers must obtain consent from the data owner. In cases where personally identifiable information is necessary for the research dataset, prompt de-identification processes should be implemented. It is essential to

encrypt data before transmitting it over the public internet. Devices housing datasets must be securely protected to prevent unauthorized. Dealing with uncertainty in research is important, but putting more ethical rules on the research and limiting how it is done also brings up its own set of ethical issues (Facca, D., Smith, M.J., Shelley, J., Lizotte, D. and Donelle, L., 2020).

The ethical dimensions of data collection and utilization encompass the following considerations:

Informed Consent: Securing approval from individuals whose data is utilized for research purposes is fundamental.

Beneficence: Researchers bear the responsibility of ensuring that their research contributes to societal welfare without causing harm.

Anonymity and Confidentiality: Maintaining the anonymity of involved individuals is preferred; when anonymity is unfeasible, researchers should guarantee the confidentiality of the data.

Privacy: Respecting the privacy of individuals is paramount, particularly when engaging in research involving individual datasets.

Expanding on the researcher's role in obtaining informed consent, it is crucial to provide clear, understandable information to participants, ensuring their voluntary agreement. De-identification involves the removal or anonymization of personally identifiable information to protect individual privacy.

For secure data transmission and storage, encryption methods suitable for public internet use should be highlighted. Additionally, securing devices with encryption, access controls, and regular security updates is essential. While the discussion already touches on informed consent and beneficence, the ethical principles of justice and integrity should also be considered. Justice involves fair distribution of benefits and burdens, while integrity underscores the importance of honesty, accuracy, and avoidance of biases in research reporting.

4 Data model and implementation

Normalisation of Data: A table is deemed to be in the third normal form (3NF) when it successfully eliminates partial dependencies and transitive dependencies among its columns. In the context of database design, it is imperative to comprehend and address these dependencies to enhance the overall structure. Normalization of data is important for decision-making methodologies as the information needs to be expressed in numerical terms and made

comparable for aggregation into a unified score for each alternative. (Vafaei, N., Ribeiro, R.A. and Camarinha-Matos, L.M., 2018)

Partial Dependency occurs when a composite attribute uniquely identifies non-key attributes, creating a dependency on only part of the primary key. Resolving this is vital for data integrity and normalization. Transitive Dependency involves a chain of attribute dependencies, leading to inefficiencies and redundancy. Identifying and addressing transitive dependencies is crucial for optimizing the database design.

In the current dataset, taking proactive measures to eliminate dependencies and adhere to the principles of 3NF involves restructuring the dataset into five distinct tables. This strategic division includes the creation of tables for DISTRICT, WARD, COUNCIL_TAX, BROADBAND, and MEDIAN_HOUSE_PRICE, and. Each table serves a specific purpose, contributing to a more organized and normalized database structure. This meticulous approach not only aligns with normalization principles but also enhances data management, retrieval, and overall system efficiency. I have connected all data tables through Ward_Code.

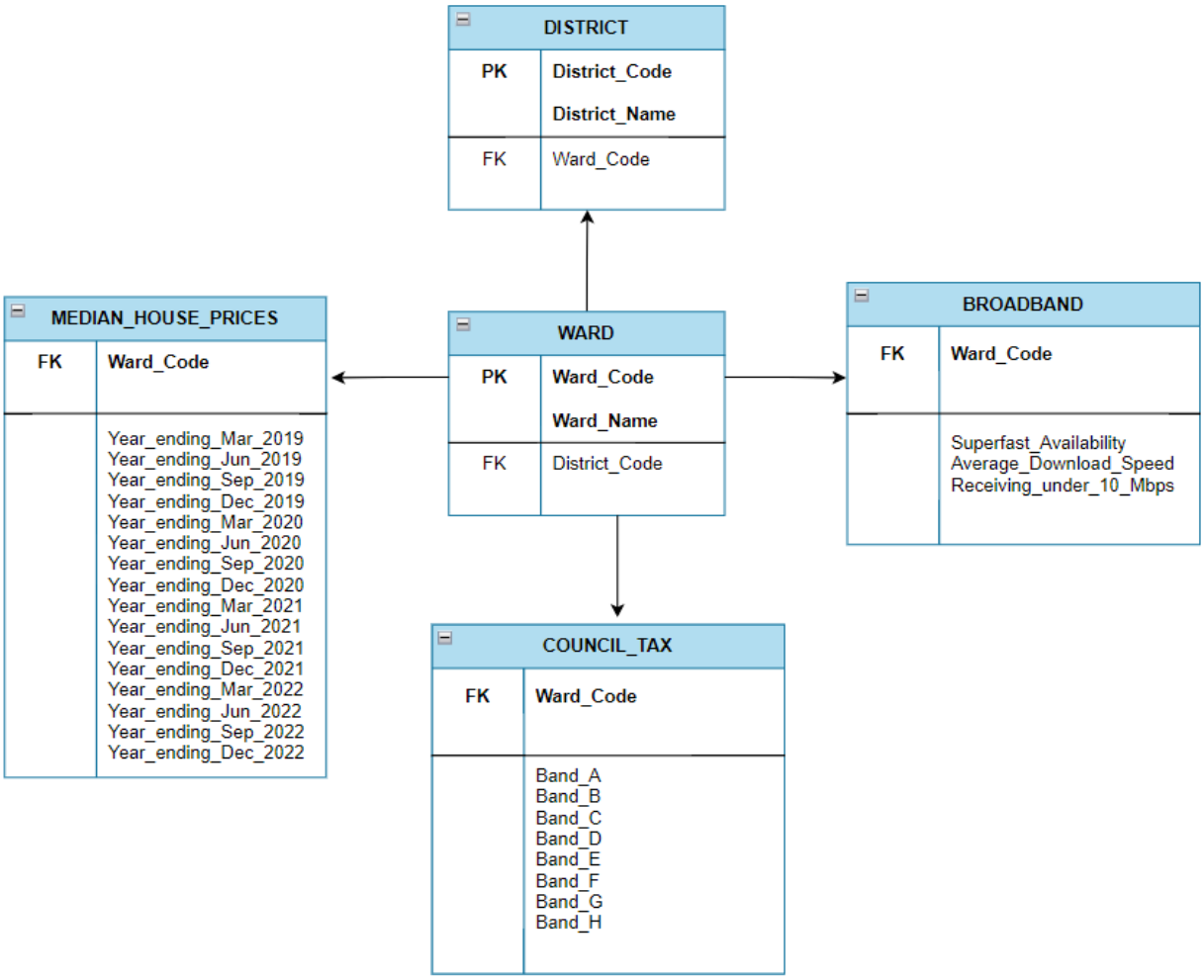

Figure 2: Schema

Appropriate design of SQL database tables:

1. **DISTRICT Table:** District_Code serves as the primary key, serving as the unique identifier for each row within the table.
2. **WARD Table:** Within the Ward table, the District_Code acts as a foreign key, establishing a reference to the DISTRICT table. Simultaneously, the Ward_Code functions as the primary key, guaranteeing the unique identification of each row in this table.
3. **MEDIAN_HOUSE_PRICE Table:** In house price table, the foreign key is Ward_Code, referencing the corresponding key in the WARD table. The primary key is ID, set to auto-increment, ensuring the unique identification of each row in this table.
4. **BROADBAND Table**: In this table, the Ward_Code serves as the foreign key, linking to the Ward table, while the primary key is ID, set to auto-increment for distinct identification within this table.
5. **COUNCIL_TAX table:** In this table, the Ward_Code serves as the foreign key, linking to the Ward table, while the primary key is ID, set to auto-increment for distinct identification within this table.

Connection of Database and R

Here, database connection with R is necessary. As I am using SQLite and R language to perform the tasks.

RSQLite enables seamless interaction with SQLite databases in R, allowing for easy creation, querying, and manipulation of data. Meanwhile, DBI provides a standardized interface for R, ensuring a consistent connection and interaction with diverse database management systems, simplifying the database-related tasks in R.

```
library(haven)
library(dplyr)
library(odbc)
library(RSQLite)
library(DBI)
```

Database Connection: Establishes a connection to an SQLite database located at the specified path.

Insert Ward: Checks the count of records in the 'WARD' table, then inserts the 'ward' data frame into the 'WARD' table.

Insert Council Tax: Checks the count of records in the 'COUNCIL_TAX' table and inserts the 'council_tax' data frame into the 'COUNCIL_TAX' table.

Insert house_price_data: Checks the count of records in the 'MEDIAN_HOUSE_PRICES' table and inserts the 'house_price_data' data frame into the 'MEDIAN_HOUSE_PRICES' table.

Insert District: Checks the count of records in the 'DISTRICT' table, removes duplicate records based on 'District_Code' in the 'district' data frame, and inserts the 'district' data frame into the 'DISTRICT' table.

Insert Broadband: Checks the count of records in the 'BROADBAND' table and inserts the 'broadband' data frame into the 'BROADBAND' table.

```
#------------ Database Connection -------------
connection <- dbConnect(RSQLite::SQLite(), "D:/Study Material/assignment data/Daatabase -SQL/SQL_sonam/SQL project/data_base.db")

# Insert Table
dbGetQuery(connection, "SELECT count(*) FROM (table)")
dbWriteTable(connection, name = "(table)", value = (dataframe), append = TRUE)

# Delete values
dbGetQuery(connection, "delete from (table)")

# Disconnect
dbDisconnect(connection)
```

5 R code to implement the SQL database queries with testing

Question 3: The average price of houses in two years, 2020 and 2021

```
  Ward_Code Ward_Name       District_Name Avg_of_2020 Avg_of_2021
1 E05011719   Wootton Vale of White Horse      445750      502500
>
```

The query provides average house prices for the "Wootton" ward in 2020 and 2021. For the "Vale of White Horse" district, the average house price was £445,750 in 2020 and £502,500 in 2021.

Question 4: The average increase (or decrease) in prices (in percent) between two years, 2019 and 2020 in Deddington ward.

```
  Ward_Code  Ward_Name District_Name Average_of_prices_for_2020 Average_of_prices_for_2021 Price_Change (%)Percentage_Change
1 E05010930 Deddington      Cherwell                   390875                   471250       80375                   20
>
```

For the "Deddington" ward in the "Cherwell" district, the average house price increased from £390,875 in 2020 to £471,250 in 2021, resulting in a £80,375 (20%) rise.

Question 5: Find a ward which has the highest house price in Mar 2021.

```
  Ward_Code   Ward_Name    District_Name Highest house price for Mar 2021
1 E05010781 Chadlington West Oxfordshire                            805908
>
```

The query identifies the ward with the highest house price in March 2021. In this specific result, the ward with Ward Code E05010781, named "Chadlington" in the "West Oxfordshire" district, has the highest house price for March 2021, recorded at £805,908.

Question 6: Find a broadband speed (average downland), or (superfast) broadband availability (%), in a particular ward (or a postcode) of a district

```
  Ward_Code Ward_Name    District_Name Average_Download_Speed Superfast_Availability
1 E05010783  Leafield West Oxfordshire                 117.1                 98.60%
>
```

The query provides broadband statistics for the ward with Ward Code E05010783, named "Leafield," in the "West Oxfordshire" district. The average download speed is reported as 117.1 Mbps, and the superfast broadband availability is noted at 98.60%.

Question 7: Calculate the average of Receiving_under_10_Mbps in West Oxfordshire District.

```
  Average_Receiving_under_10_Mbps
1                            5.51
>
```

The query calculates the average percentage of "Receiving under 10 Mbps" for wards in the 'West Oxfordshire' district, providing a result of approximately 5.51%.

Question 8: Calculate average council tax charge for a Leafield town in a West Oxfordshire district for any three bands A, B and C of properties.

```
 Ward_Code Ward_Name    District_Name  Band_A  Band_B  Band_C Average Council tax Charge
 E05010783  Leafield West Oxfordshire 1522.93 1776.76 2030.58                   1776.757
```

The output shows council tax information for the ward 'Leafield' in the 'West Oxfordshire' district, including individual band charges (Band_A, Band_B, Band_C) and the calculated average council tax charge, which is approximately 1776.76.

Question 9: Calculates the difference between council tax charges of same bands but of two different towns of the same district.

the query measures the disparity in Band C council tax charges between the 'Churchill' and 'North Leigh' wards.

```
dbGetQuery(connection,"
          SELECT  DISTINCT(select c.Band_C from COUNCIL_TAX c, WARD a WHERE c.Ward_Code = a.Ward_Code and a.Ward_Name = 'Churchill') -
(select c.Band_C from COUNCIL_TAX c, WARD a WHERE c.Ward_Code = a.Ward_Code and a.Ward_Name = 'North Leigh') as Difference
FROM COUNCIL_TAX c, WARD a

   ")
Difference
      39.5
```

Question 10: find a town which has the lowest council tax charges for Band B properties

```
+  )
  Ward_Code   Ward_Name    District_Name Minimun Council tax charges
1 E05010781 Chadlington West Oxfordshire                    3118.47
>
```

The query identifies the ward with the minimum council tax charges for properties classified under Band F in the "West Oxfordshire" district. In this specific result, the ward with Ward Code E05010781, named "Chadlington," has the minimum council tax charges recorded at £3,118.47.

Video Recording Link:

https://drive.google.com/drive/folders/1BUL-Ck6MGUlnSzRucnwsmG8vI9QD5R_5?usp=sharing

References

Hu, J., 2021. Data cleaning and feature selection for gravelly soil liquefaction. *Soil Dynamics and Earthquake Engineering*, *145*, p.106711.

Vafaei, N., Ribeiro, R.A. and Camarinha-Matos, L.M., 2018. Data normalisation techniques in decision making: case study with TOPSIS method. *International journal of information and decision sciences*, *10*(1), pp.19-38.

Suciu, D., 1998. An overview of semistructured data. *ACM SIGACT News*, *29*(4), pp.28-38.

Facca, D., Smith, M.J., Shelley, J., Lizotte, D. and Donelle, L., 2020. Exploring the ethical issues in research using digital data collection strategies with minors: A scoping review. *Plos one*, *15*(8), p.e0237875.