MSc in Artificial Intelligence and Data Science

Assignment – 771766 Fundamentals of Data Science PROJECT

(This assignment is worth 70% of the total marks for this module)

This assignment should be submitted **via turnitin on CANVAS** using the specific coversheet for this assignment, which can also be found on CANVAS by **Friday 28th April 2023 at 2pm *at the latest.***

Maximum Word Count for the written part of this assignment = 2500 words. [*Note that this is a maximum word count, not a goal.*]

Context.

This assignment is intended to stretch your data science skills and develop your programming through the use of a mock census. Some of the contact time on the workshop will be spent on this project to help plan, think about, and ultimately execute the assignment, but further personal study time will be essential to dedicate to this as well.

Project Background Information.

Every ten years, the United Kingdom undertakes a census of the population, with the most recent one having been conducted in 2021. The purpose of such a census is to compare different people across the nation and to provide the government with accurate statistics of the population to enable better planning, to develop policies, and to allocate certain funding.

In the project, you will be provided with a mock census of an imaginary modest town. I would like you to consider yourselves to be part of a local government team who will be making decisions on what to do with an unoccupied plot of land and what to invest in. To address these questions, you will need to clean and analyse the mock census data provided.

About this Mock Census.

The mock census you will be given contains randomly generate data using the Faker package in Python. It has been generated in a similar manner to (and designed to directly emulate the format of) the 1881 census of the UK wherein only a few questions were asked of the population. The fields recorded are as follows:

1. Street Number (this is set to "1" if it is a unique dwelling);
2. Street Name;
3. First Name of occupant;
4. Surname of occupant;
5. Age of occupant;
6. Relationship to the "Head" of the household (anyone aged over 18 can be a "Head" – they are simply the person who had the responsibility to fill in the census details);
7. Marital status (one of: Single, Married, Divorced, Widowed, or "NA" in the case of minors);
8. Gender (one of: Male, Female; note that other responses were not implemented in 1881);
9. Occupation (this field was implemented in a modern style, rather than typical 1881 occupations);
10. Infirmity (we have implemented a limited set of infirmities following the style of 1881);
11. Religion (we have implemented a set of real-world religions).

The first task you will have to do is to clean this dataset. As you will rapidly discover, there are missing entries, and, candidly, some responses from the population are outright lies. Part of the grading for the assignment will assess these details.

The Task. The town from the census is a modestly sized one sandwiched between two much larger cities that it is connected to by motorways. The town does not have a university, but students do live in the town and commute to the nearby cities. Once you have a cleaned dataset to analyse, your task is to decide the following:

(a) What should be built on an unoccupied plot of land that the local government wishes to develop? Your choices are:

(i) High-density housing. This should be built if the population is significantly expanding.

(ii) Low-density housing. This should be built if the population is "affluent" and there is demand for large family housing.

(iii) Train station. There are potentially a lot of commuters in the town and building a train station could take pressure off the roads. But how will you identify commuters?

(iv) Religious building. There is already one place of worship for Catholics in the town. Is there demand for a second Church (if so, which denomination?), or for a different religious building?

(v) Emergency medical building. Not a full hospital, but a minor injuries centre. This should be built if there are many injuries or future pregnancies likely in the population.

(vi) Something else?

Whichever you choose, you must justify it from the data provided to you and argue it is a priority against other choices.

(b) Which one of the following options should be invested in?

(i) Employment and training. If there is evidence for a lot of unemployment, we should re-train people for new skills.

(ii) Old age care. If there is evidence for increasing numbers of retired people in future years, the town will need to allocate more funding for end of life care.

(iii) Increase spending for schooling. If there is evidence of a growing population of school-aged children (new births, or families moving in to the town), then schooling spend should increase.

(iv) General infrastructure. If the town is expanding, then services (waste collection; road maintenance, etc.) will require more investment.

In order to address these two questions, it is suggested that some of the analysis you undertake is:

- Examine the age distribution (age pyramid) of the population. Is it growing or shrinking? Will there be more retired aged people in the future, more school-aged children, more young people, etc.
- Examine unemployment trends. Are certain ages more likely to be unemployed than others.
- Examine religious affiliations. Are any religions growing, or shrinking? Are there any newer religions that are increasing in numbers?
- Examine the divorce and marriage rate. This might impact how you think about housing.
- Examine the occupancy level (how many people per house) and determine if existing housing is being under or over-used.
- Examine the number of university students. All of these are commuters since there are no universities in the town. Are there any other professions that are likely to be commuters?
- What is the birth rate and death rate for the town?

These are merely suggestions, there are plentiful other analyses that could be undertaken that will be discussed in the videos and in class. Ultimately, your answers to (a) and (b) must be justified from the census data, and argued by balancing the different needs of the population and supported through statistics and where appropriate, hypothesis testing. As such, this is a "real" exercise but based on artificial data. [As a disclaimer: any reference to real people or places, living or dead, is purely coincidental and a product of the random generators that have been used.]

Grading.

The following grading rubric will be applied to your supplied answers. The total number of marks available for this assignment is 80. Please note that submitting lots of data is unlikely to attract many marks. Instead, we want to see fully reasoned cases supported by evidence derived from the data supplied.

Given the word count, it is essential to be concise in your answers. It is strongly suggested that you illustrate your answers with appropriate diagrams (i.e. visualisations) or appendices of example calculations. Further, you might need to read around the topic and undertake library/online research to help with this assignment to achieve the highest grades.

Please upload:

(i) Your cover sheet.

(ii) Your written address to the assignment.

(iii) The code you wrote to produce the results and/or visualisations used in the assignment.

Criteria	0 Marks	Up to 6 marks	Up to 12 marks	Up to 20 marks
Coding	No example of code has been uploaded, or what has been uploaded is of very low quality.	Example of code has been uploaded. The code might not be complete, or it might have some obvious omissions or errors inside it. Commenting may be poor to mediocre.	The code supplied is fairly complete, is mostly correct, and is appropriate for the task. Commenting is of a reasonable quality.	The code supplied is complete and can generate the important items contained in the report with ease. The code is also fully correct, efficient, and contains extensive commenting. At the highest levels, it would be suitable for publication in an official repository.
Cleaning	No attempt has been made to clean the data.	Data cleaning has been attempted and the steps taken noted in the write up.	The data has been cleaned to a high standard and blanks filled into a reasonable level. This has been documented in the write up.	The data has been extensively cleaned with a range of appropriate actions undertaken and justified in the write up. At the highest levels, the cleaning is equivalent to a professional in the field.
Analysis	Little to no attempt has been made at analysing the census data. If an attempt has been made, it is very poor quality.	An attempt has been made to produce most of the analysis noted in the briefing, above. Most of it meets a satisfactory level.	All analysis required is present, justified and almost all of it is correct. Key statistics are reported where required.	At the highest levels, the analysis performed is similar to a professional level of insight and covers many different angles.
Arguments and data visualisation	The write up fails to make a coherent argument in answer to the main tasks (a) and (b), above.	Both (a) and (b) are addressed, but the logic used is weak, or incorrect, or fails to adequately reference the data. Visualisations might be basic, or of poor quality.	Both (a) and (b) are argued to a good standard using the data and representative visualisations of an acceptable standard.	High quality arguments are presented for (a) and (b) that include balancing the different needs of the population that have been identified from the data. This includes well-presented visualisations. At the highest levels, and if this were a “real life” exercise, the work would be publishable.