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 as a written document and a Python notebook separately via **Turnitin on Canvas by Thursday 12th of December 2024 at 2pm at the latest (Late Stream – Monday 13th of January 2025 at 2pm).** The written part of this assignment should be 2,500 words, which can be exceeded by 10% without penalty (i.e., the hard limit is 2,750 words). Tables, figures, (including legends and captions), appendices, references, footnotes and endnotes do not count towards the word limit. Content from after the word limit will not be graded.

Note that the written report should not include any Python code or screenshots of your notebook. Python code must be submitted in a separate Jupyter notebook.

Context

The goal of this assignment is to apply your data science skills to a real-world scenario in the form of a mock census. Some of the workshops in this module will be spent on tasks for this project which will help you to plan, think about, and ultimately execute the assignment, but it is essential for you to dedicate personal study time to this assignment 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 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 that you will work with has been simulated 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 "N/A" 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).

As a first step, you will need to clean this data set. You will find that there are missing entries, and, potentially, some responses from the population are false. Part of the grading for the assignment will assess your rationale in correcting these details.

Note: While the format follows the 1881 census, the data will resemble modern-day demographics.

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 data set 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 set 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 significant amount of unemployment, we should re-train people for new skills.
ii. Old age care. If there is evidence that there will be an increased number 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 into the town), then the schooling spend should increase.
iv. General infrastructure. If the town is expanding, then services (waste collection; road maintenance, etc.) will require more investment.

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 people of retirement age 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 the death rate for the town?

These are merely suggestions, and there are many other analyses that could be undertaken that will be discussed in the lectures and the labs. 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 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 total number of marks available for this assignment is 80. We expect to see fully reasoned answers to the questions which are supported by evidence derived from the supplied data set. For more detail, see the rubric on the next page.

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). You may also find it useful to read around the topic and undertake library/online research, should you wish to achieve the highest grades.

Please upload the following files **separately** (you cannot upload a zip file, for example):

i. Your written report.
ii. The code you wrote to produce the results and/or visualisations used in the assignment (i.e. the jupyter notebook).

Marking Rubric

Criteria	1	2	3	4	5
Coding	No jupyter notebook has been uploaded, what has been uploaded is too short to correspond to the level of analysis required to answer (a) and (b), or is not of sufficient quality to be able to analyse the data adequately.	Example of code has been uploaded. The code might not be complete, or it might have some obvious omissions or errors inside it. Commenting is not to a sufficient standard to inform an unfamiliar reader. No markdown is used, or, it is not used sufficiently to structure the notebook effectively.	The code supplied meets a baseline competency, where it is complete to the minimum standards and is written with appropriate and relevant syntax for the analysis undertaken. There is sufficient commenting to follow the procedures stated in the report and there may be appropriate use of some markdown.	The code is written to a more advanced level and well-structured with an advanced use of markdown and comments that cross-matches with the written report from beginning to end (e.g. Data intro; EDA; Analysis and Visualsation).	The code supplied is fully complete and can generate the important items contained in the report; what is in the report matches the notebook fully. The code has minimal to no errors, is efficient, and contains extensive commenting or potential use of docstrings. It could be suitable for publication in an official repository.
Data Cleaning	The data cleaning has not been attempted or there is no evidence of data cleaning found within the code or the report.	The data cleaning has been attempted, but is either too simplistic, generalises or makes irrational assumptions about the data. (e.g. imputes all missing ages to be the same)	A baseline understanding of the data cleaning required for the dataset has been written into a structured methodology, where the data have been used to inform simple plausible imputation. (e.g. using the median age of a particular demographic to impute missing age).	Advanced understanding of data cleaning for the dataset is outlined and detailed in the report, including advanced methods of imputation and/or additional considerations of errors within the data that are used. (e.g. plausible/appropriate use of random selection methods).	A superior understanding of the data cleaning process for the data is clearly written, detailing the methodology followed, as well as adequate justification why these methods were used. At the highest levels, the cleaning is equivalent to a professional in the field.
Analysis	No attempt has been made at analysing and interpreting the census data. If an analysis has been attempted, it is not of sufficient quantity and quality, and/or the logic surrounding the interpretations of the data are wholly implausible/illogical.	The report contains an analysis along with an attempt to interpret the data. However, the analysis is incomplete, or interpretation is limited.	A basic analysis of the data is detailed in the report and meets the minimum requirements of the briefing document. This is paired with sufficient quantification and may contain baseline attempts to infer implications from the results.	A more advanced analysis of the data is detailed; The results shown are appropriately quantified, with significances stated, along with a more detailed breakdown of the inferences made from the results (e.g. comparing birth and death rates against entire UK population etc).	The analysis is of a superior nature, containing a thorough analysis, as well as a fully detailed discussion of the interpretations and implications of the results. At the highest levels, the analysis is to a professional level of insight and covers many different angles.

Arguments and Data Visualisation	The report makes no effort to formulate any arguments to justify the decisions made on infrastructure projects for points (a) and (b), or the argument in answer to the main tasks (a) and (b) is incoherent.	Both (a) and (b) are addressed in the report, 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 baseline standard in the report; The report uses the data and representative visualisations to help justify these arguments, but may not have sufficient analysis and/or the correct visualisations to form a full cohesive justification.	Both (a) and (b) are argued to an advanced level in the report; The report uses the data and an extensive (but appropriate) array of visualisations to help to justify these arguments. The analysis complements these arguments and there are very limited-to-no gaps in the logic between the arguments and the logic.	High quality arguments are written for (a) and (b) that include balancing the different needs of the population that have been identified from the data. This includes extremely well-presented visualisations, with detailed captions. At the highest levels, and if this were a "real life" exercise, the work would be publishable.