# Group ID - MSc in Data Analytics

Author: I. Fadieiev

e-mail: sba23139@student.cct.ie

Student ID: sba23139

Abstract

This research delves into a comprehensive analysis of Ireland's population estimates from 1950 to 2023, with a focus on age groups and genders. Utilizing Python, the study conducted data preprocessing, eliminating irrelevant information and addressing missing values. Visualizations of age structures across all years were created, accompanied by a flexible function for generating population pyramids. The study further explored mean age variations over time, employing machine learning to predict changes. Clustering techniques identified baby booms, and predictive models calculated populations for specific ages. Statistical analyses compared the data against normal, binomial, and Poisson distributions, offering insights into the underlying population dynamics.

Introduction

This study focuses on a analysis of Ireland's population estimates from 1950 to 2023, employing various data analytics techniques and methods. The foundation of this research lies in Python programming, where an initial phase of data preprocessing was undertaken to refine the dataset by removing extraneous information and addressing missing values.

Using sophisticated approaches to population estimation, the research progresses into visualizing age structures across the temporal spectrum, offering a nuanced portrayal of Ireland's population dynamics. The development of a dynamic function for generating population pyramids adds versatility to the analytical toolkit. Expanding into predictive modeling, machine learning algorithms were applied to anticipate mean age variations over time. Clustering techniques were employed to identify significant events, such as baby booms, contributing insights into generational patterns.

The study also embraces statistical analyses to characterize the distribution of population data, scrutinizing fits against models like normal, binomial, and Poisson distributions. This work stands at the intersection of data analytics and population trends, offering a comprehensive methodology applicable to a diverse array of datasets. The insights generated not only deepen our understanding of Ireland's data landscape but also hold practical implications for data analysts, policymakers, and researchers navigating the complexities of population analytics and forecasting.

1. Data preparation and Visualization

1.1. Exploratory Data Analysis

Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics often plotting them visually [1]. For my study I decided to make such steps for EDA: Data exploration. Data cleaning and Data filling.

1.1.1. Data Exploration:

The initial phase of the exploratory data analysis involved a visual inspection of the dataset. This facilitated a initial understanding of its structure, allowing for informed decisions in subsequent stages. Specific attention was directed towards such columns as 'UNIT,' 'STATISTIC Label,' and 'Age Group.' The exploration revealed that 'UNIT' and 'STATISTIC Label' had only one unique value each, leading to a judicious decision to drop these columns. Simultaneously, 'Age Group' underwent a meticulous examination to identify overlapping age groups, guiding the selection of only those relevant to the study.

Additionally, a detailed investigation into missing values patterns was conducted. Notably, missing values were consistently observed in the '0 - 4 years' age group, laying the groundwork for a targeted approach in the subsequent data cleaning phase.

1.1.2. Data Cleaning:

With insights gleaned from the exploratory phase, data cleaning focused on enhancing the dataset's coherence. The identified redundant columns, 'UNIT' and 'STATISTIC Label,' were removed to streamline the data. The 'Age Group' column underwent refinement to retain only the essential age groups for the study.

A crucial aspect of the cleaning process was the targeted handling of missing values in the '0 - 4 years' age group. Recognizing that these missing values pertained to 'Under 1 year' and '1 - 4 years,' a systematic approach was employed to calculate and fill these gaps, ensuring the dataset's completeness and accuracy.

1.2.3. Data Filling:

The final step in this exploratory journey involved a strategic approach to data filling. The calculated missing values in the '0 - 4 years' age group were systematically addressed by summing 'Under 1 year' and '1 - 4 years' for each Year and Sex category. This meticulous data filling process aimed to eliminate gaps in crucial age group information, laying the foundation for subsequent analyses.

In summation, the combination of data exploration, cleaning, and filling strategies not only refined the dataset but also provided a methodologically sound basis for the ensuing phases of the analysis. The decisions made were driven by a nuanced understanding of the dataset's intricacies, ensuring a robust foundation for further exploration and interpretation.

1.2. Methods used

1.2.1. Data Cleaning:

The data cleaning phase addressed, redundancies, and missing values in the dataset. This not only improved the overall quality of the data but also laid a solid foundation for subsequent analyses. By ensuring data integrity, the neural network model's training process was set on a robust trajectory.

1.2.2. Data Filling:

Filling missing values, particularly in the '0 - 4 years' age group, was a crucial aspect of data preparation. By calculating and imputing missing values based on the sum of 'Under 1 year' and '1 - 4 years' for each Year and Sex category, the dataset's completeness was ensured. This step was pivotal in preventing information gaps that could potentially hinder the neural network's ability to learn and generalize from the data.

1.2.3. Feature Selection:

Feature selection was a key strategy to enhance the efficiency and interpretability of the neural network model.

In essence, this comprehensive approach to data preparation, encompassing data cleaning, data filling, and feature selection, laid the groundwork for a more efficient and focused neural network training process.

1.3. Visualisations used

Appropriate visualizations must be used to engender insight into the dataset and to illustrate your final insights gained in your analysis. [0-20]

All design and implementation of your visualizations must be justified and detailed in full., making reference to Tufts Principles [0-30]

2. Machine learning for Data Analytics

2.1. Project management framework

For my data science project involving population estimates in Ireland, the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework appears to be the most suitable choice. Here's the rationale behind this selection:

- Comprehensive and Iterative Nature: CRISP-DM is a cyclical and iterative process that accommodates the complexities of data science projects. Given the diverse analyses in my project, from exploring age structures and predicting mean age changes to clustering baby booms, the iterative nature of CRISP-DM allows me to refine objectives, adjust models, and explore new insights as I progress.

- Phased Approach: CRISP-DM consists of distinct phases, including Business (problem) Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This structured approach aligns well with the various components of my project. For instance, the Business Understanding phase allows me to define and refine project objectives based on initial findings, while the Modeling and Evaluation phases cater to the machine learning aspects of predicting mean age changes and assessing model performance.

- Flexibility: CRISP-DM provides the flexibility needed for data science projects where the goals and insights may evolve. As I navigate through the different phases, I can adapt my approach based on emerging patterns and findings. This adaptability is valuable in a project like mine, where the analyses are diverse, ranging from visualizations to machine learning model development.

- Industry Standard: CRISP-DM is a widely recognized and accepted framework in the data science community. Its popularity is a testament to its effectiveness and adaptability across various domains. Using a well-established framework like CRISP-DM enhances the reproducibility and communication of my methodology.

In summary, CRISP-DM offers a well-rounded, flexible, and industry-recognized framework that aligns with the diverse analytical aspects of my study. Its iterative nature, structured phases, and adaptability make it a suitable choice for navigating the complexities of exploring and predicting population trends in Ireland.

2.2. Machine learning techniques

Provide an explanation of why you chose a supervised, unsupervised, or semi-supervised machine learning technique for the dataset you used for ML modeling. [0 - 20]

2.3. Predictions

Machine learning models have a wide range of uses, including prediction, classification, and clustering. It is advised that you assess several approaches (at least two), choose appropriate hyperparameters for the optimal outcomes of Machine Learning models using an approach of hyperparameter tunning, such as GridSearchCV or RandomizedSearchCV. [0 - 30]

2.4. Clustering

Show the results ML modeling comparisons in a table or graph format. Review and critically examine the machine learning models' performance based on the selected metric for supervised, unsupervised, and semi-supervised approaches. [0 - 30]

Demonstrate the similarities and differences between your Machine Learning modelling results using the tables or visualizations.

2.5. Machine Learning effectiveness

Provide a report along with an explanation and interpretation of the relevance and effectiveness of your findings. [0 - 20]

3. Statistics

3.1. Dataset summary

Summarise your dataset clearly, using relevant descriptive statistics and appropriate plots. These should be carefully motivated and justified, and clearly presented. You should critically analyse your findings, in addition to including the necessary Python code, output and plots in the report. You are required to plot at least three graphs. [0-35]

3.2. Discrete distributions

Use two discrete distributions (Binomial and/or Poisson) in order to explain/identify some information about your dataset. You must explain your reasoning and the techniques you have used. Visualise your data and explain what happens with the large samples in these cases. You must work with Python and your mathematical reasoning must be documented in your report. [0-30]

3.3. Normal distribution

Use Normal distribution to explain or identify some information about your dataset. [0-20]

3.4. Distributions summary

Explain the importance of the distributions used in point 3 and 4 in your analysis. Justify the choice of the variables and explain if the variables used for the discrete distributions could be used as normal distribution in this case. [0-15]

4. Programming

4.1. Programmatical exploration

The project must be explored programmatically, this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should also be applied). [0-50]

Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.

4.2. Programming paradigms

Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project.

Conclusions

References

1. Prabhu, T.N. (2019). *Exploratory data analysis in Python.* [online] Medium. Available at: https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce.