# 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

You must perform appropriate EDA on your dataset, rationalizing and detailing why you chose the specific methods and what insight you gained. [0-20]

1.2. Methods used

You must also rationalise justify and detail all the methods used to prepare the data for ML. [0-30]

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

Explain which project management framework (CRISP-DM, KDD or SEMMA) is required for a data science project. Discuss and justify with real-life scenarios.

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