# 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.

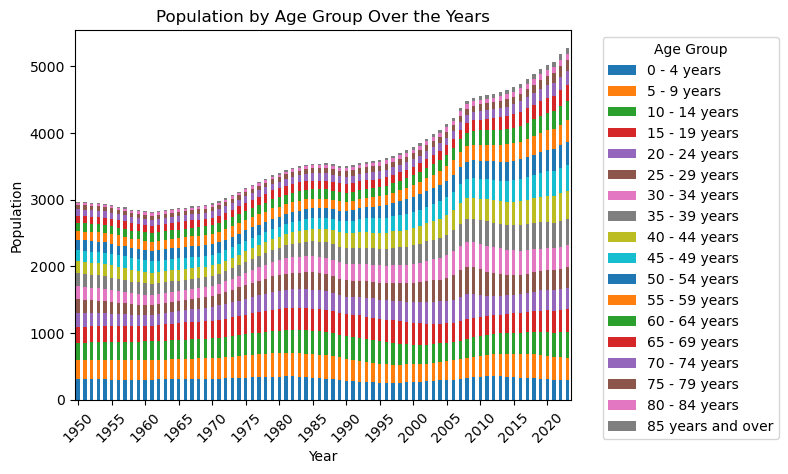
1.2.3. Feature Selection:

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.

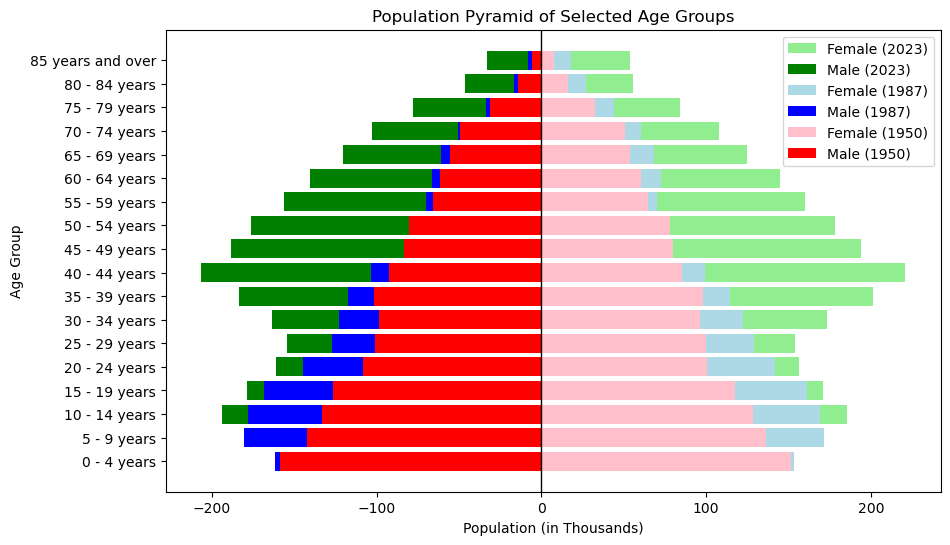
1.3. Visualisations used

“Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.” - Edward R. Tufte [2]. In this section, I detail the visualizations employed to gain insights into Ireland's population estimates from 1950 to 2023. These visualizations align with Tufts Principles, emphasizing clarity, credibility, and the ability to reveal details on demand.



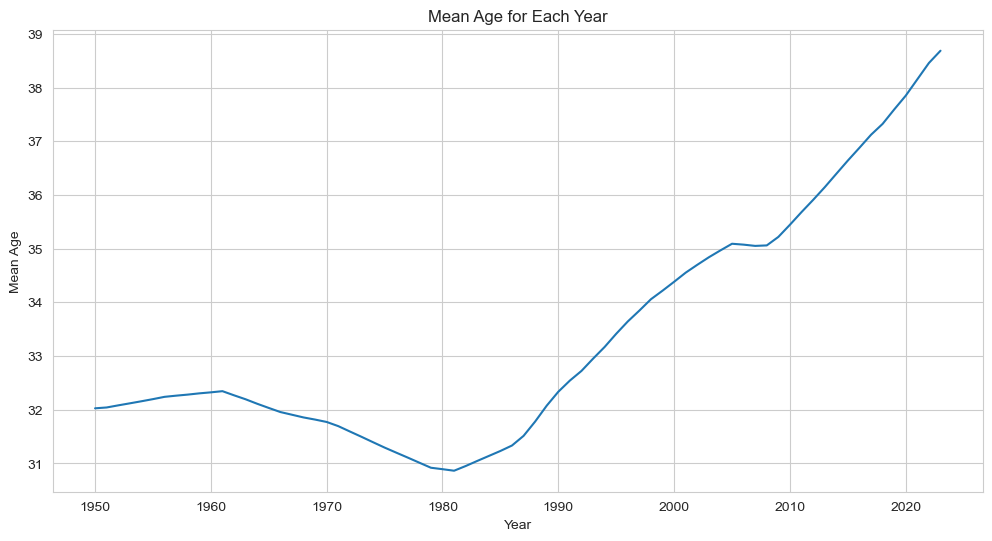
Population by Age Group Over the Years.

This visualization provides a comprehensive overview of population distribution across age groups over time, allowing for the identification of trends and patterns.



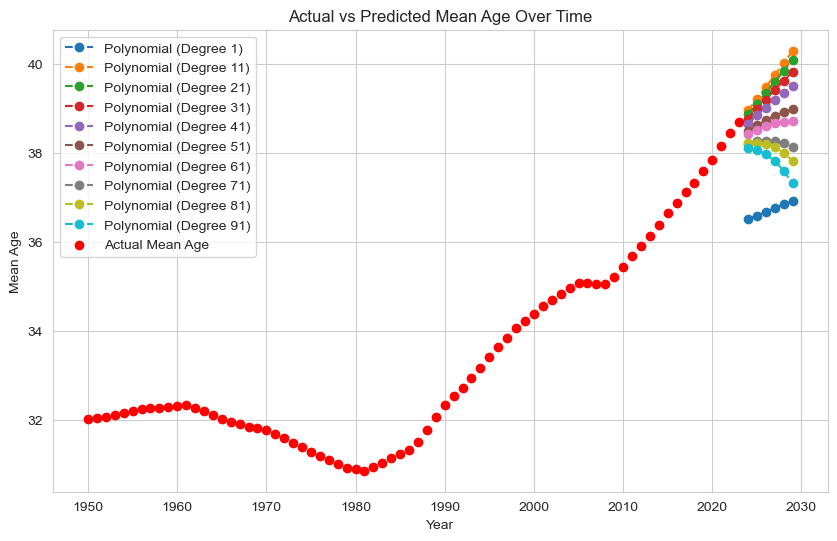
Population Pyramids for Selected Years

Population pyramids offer a visual representation of age and gender distribution simultaneously, aiding in the observation of demographic trends and imbalances.



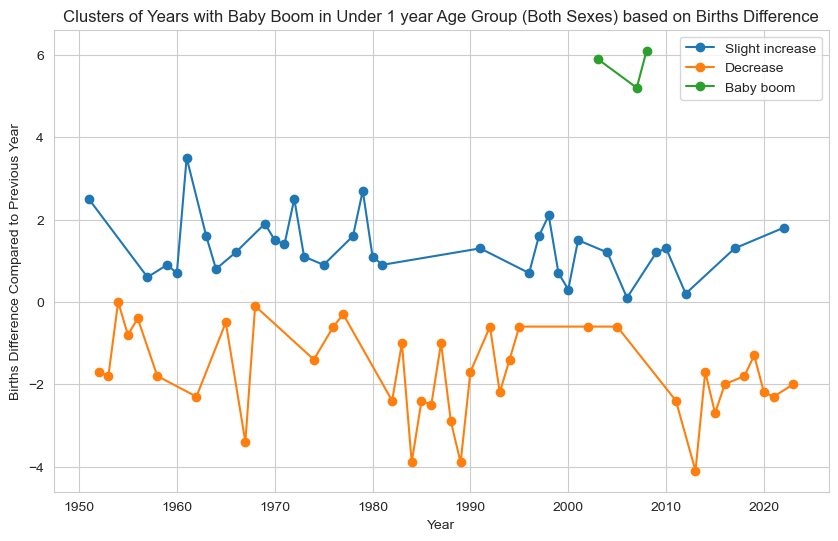
Mean Age by Each Year

Displaying mean age trends over time provides insights into the aging or rejuvenation of the population.



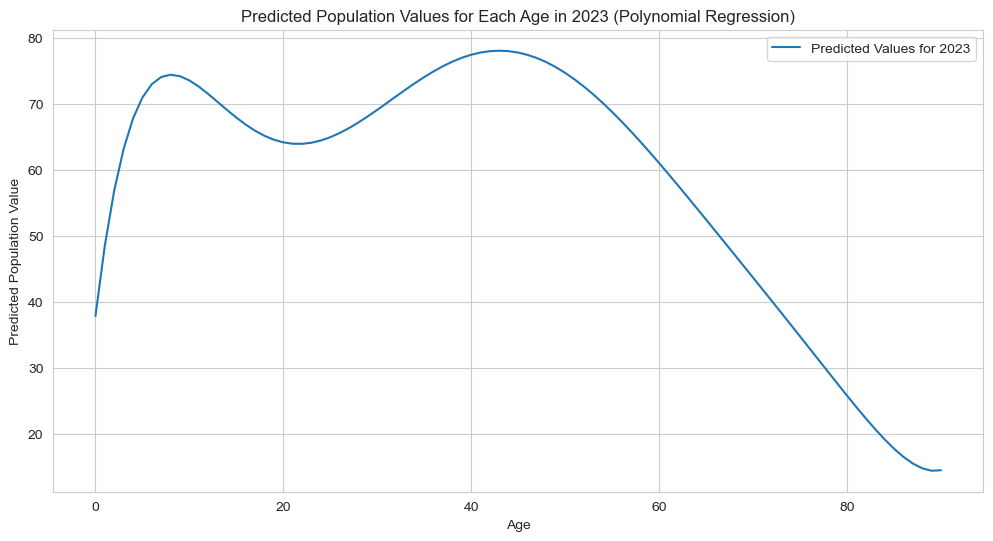
Actual and Predicted Mean Age Over Time

This graph was used to find the best degree for polynomial regression and compare self-made results with GridSearchCV and RandomizedSearchCV results.



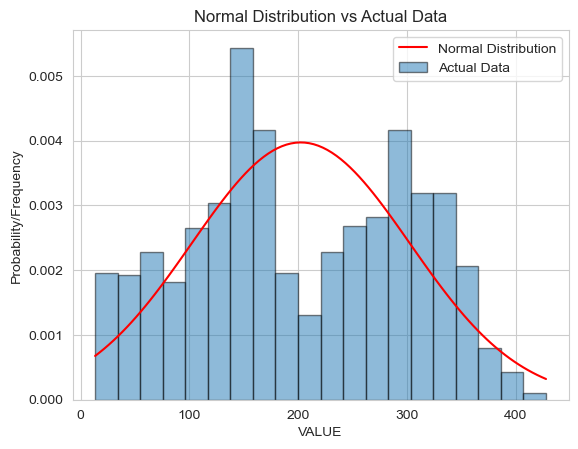
Clusters of Years with Baby Boom.

Identifying clusters of years with baby booms highlights demographic events and their impact on the population structure.



Predicted Population Values for Each Age in Selected Year (Polynomial Regression)

Polynomial regression visualizations predict population values for each age, offering a detailed perspective on age-specific trends.



Normal Distribution vs Actual Data

Binomial Distribution vs Actual Data, Poisson Distribution vs Actual Data, Normal Distribution vs Actual Data are made on the same principle. Comparing actual data with different theoretical distributions assesses the fit and characteristics of the population data.

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

In this section, I elaborate on the rationale behind choosing specific machine learning techniques employed in these analyses.

2.2.1. Linear Regression and Polynomial Regression for Prediction of Mean Age

Explanation: Linear regression and polynomial regression were chosen for predicting mean age over time due to their effectiveness in capturing trends in continuous variables. Linear regression provides a baseline understanding of mean age changes, while polynomial regression allows for capturing potential non-linear relationships.

Justification: The choice of these regression techniques aligns with the goal of predicting a continuous target variable (mean age) based on the temporal aspect of the dataset. Polynomial regression introduces flexibility to capture potential curvature in age trends, ensuring a more nuanced model for predicting mean age changes over the study period.

2.2.2. DBSCAN for Clustering

Explanation: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was selected for clustering due to its capability to identify clusters of varying shapes and densities in the data. In the context of these analyses, the objective was to identify distinct groups of years with similar population structures.

Justification: DBSCAN is well-suited for identifying clusters without making assumptions about their shape, making it appropriate for capturing the diverse patterns in population structures over different years. The ability to handle noise in the dataset ensures robust clustering, particularly when dealing with demographic shifts. And, to be honest, DBSCAN is the only clustering method worked with my data as I expected.

2.2.3. Polynomial Regression to Predict Numbers of People for Each Age

Explanation: Polynomial regression was employed for predicting the numbers of people for each age group based on the values of age group populations. This choice was made to capture potential non-linear relationships between age groups.

Justification: Polynomial regression allows for flexibility in modeling complex relationships between variables. In this case, predicting the numbers of people for each age group may involve non-linear patterns, and polynomial regression provides a versatile approach to capture such nuances in the data.

In conclusion, the choices of machine learning techniques in these analyses are driven by the specific characteristics of the data and the objectives of the analysis. Linear and polynomial regressions suit the prediction of mean age trends, DBSCAN is apt for clustering population structures, and polynomial regression is effective in capturing non-linear relationships when predicting the numbers of people for each age group. These choices align with the underlying structure of the population data and the goals of the study, ensuring that the selected techniques are well-suited to extract meaningful insights from the dataset.

2.3. Predictions

In this study, the application of machine learning models primarily focuses on prediction, with the central goal of forecasting mean age changes and estimating the number of individuals in each age group. To achieve optimal outcomes, a detailed assessment of two key approaches—linear regression and polynomial regression—was conducted. Additionally, hyperparameter tuning was employed through the RandomizedSearchCV and GridSearchCV methods, which were compared to hand-tuning of model parameters.

2.3.1. Linear Regression for Mean Age Prediction:

Approach: Linear regression serves as a fundamental model for predicting mean age changes over the study period.

2.3.2. Polynomial Regression for Mean Age Prediction:

Approach: Recognizing potential non-linear relationships in mean age trends, polynomial regression was employed as a more flexible model.

Hyperparameter Tuning: RandomizedSearchCV and GridSearchCV techniques were applied to explore the optimal degree of the polynomial, allowing for the identification of the most suitable level of complexity in capturing non-linear patterns in mean age changes.

2.3.3. Polynomial Regression for each age population Prediction:

Approach: To predict the number of individuals of each age, polynomial regression was chosen for its versatility in capturing non-linear relationships between variables.

Hyperparameter Tuning: Tuning was made by myself in few iterations, comparing the results and using visual representation to compare results.

2.4. Clustering

Clustering, a fundamental technique for uncovering patterns in data, was employed in this study to discern distinct birth rate patterns over different years. It was the hardest part of the study for me. I tried to use different methods, but only DBSCAN worked as I expected. Despite the challenges encountered, a comprehensive review and critical examination of the machine learning models' performance with a particular focus on the DBSCAN algorithm.

2.4.1. General Information

**DBSCAN for Clustering**

Approach: DBSCAN, a density-based clustering algorithm, was chosen for its ability to identify clusters of varying shapes and sizes, making it suitable for the diverse patterns in population structures over different years.

Performance Evaluation: The clustering performance was assessed using metrics like silhouette score, which measures the compactness and separation of clusters. The primary focus was on achieving a clear distinction between the 'Baby Boom' and 'Not Baby Boom' clusters.

**Alternative Non-Machine Learning Clustering**

Approach: An alternative approach, devoid of machine learning techniques, was also explored for clustering based on domain knowledge and demographic trends.

Performance Evaluation: The non-machine learning clustering approach was evaluated against the same metrics to critically examine its accuracy and efficiency compared to the DBSCAN algorithm.

2.4.2. Performance Examination

**DBSCAN**

Strengths: DBSCAN effectively identified the 'Baby Boom' cluster, providing valuable insights into periods characterized by significant demographic shifts.

Challenges: The algorithm encountered difficulties in creating more nuanced clusters beyond 'Baby Boom' and 'Not Baby Boom.' This limitation could be attributed to the specific nature of population structures, which may not conform to traditional clustering patterns.

**Alternative Clustering Approach:**

Strengths: The non-machine learning clustering approach, driven by domain knowledge, demonstrated higher accuracy in creating more nuanced clusters that aligned with historical events.

Challenges: This approach, while accurate, might be more dependent on the researcher's understanding of demographic trends and could be less adaptable to unexpected patterns in the data.

2.4.3. Critical Examination

**DBSCAN**

Advantages: DBSCAN is robust and capable of handling irregularly shaped clusters.

Limitations: The challenge in creating more than two clusters suggests a limitation in DBSCAN's adaptability to nuanced demographic shifts that may not conform to distinct density-based patterns.

**Alternative Clustering Approach**

Advantages: The alternative approach, grounded in domain knowledge, allows for a more tailored and nuanced clustering based on specific demographic events.

Limitations: The dependence on domain knowledge could potentially lead to biases or overlook unexpected patterns that machine learning algorithms may uncover.

2.5. Machine Learning effectiveness

This study on Ireland's population estimates from 1950 to 2023 underscores the critical role of model selection, data quality, and hyperparameter tuning in achieving meaningful outcomes.

2.5.1. Machine Learning Models

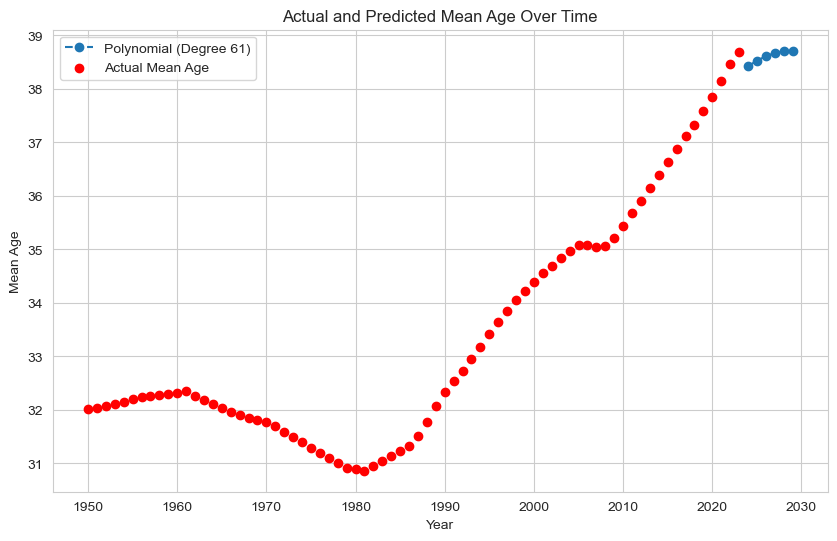
Linear and polynomial regression models were used for predicting mean age changes, capturing both linear and non-linear trends. For my data polynomial regression models worked much better. DBSCAN excelled in identifying 'Baby Boom' clusters, though limitations were acknowledged in creating more nuanced clusters.

2.5.2. Data Quality

Thorough data exploration and cleaning were crucial for ensuring the reliability of analyses and models. Rigorous data quality assessment laid the foundation for meaningful insights and accurate predictions.

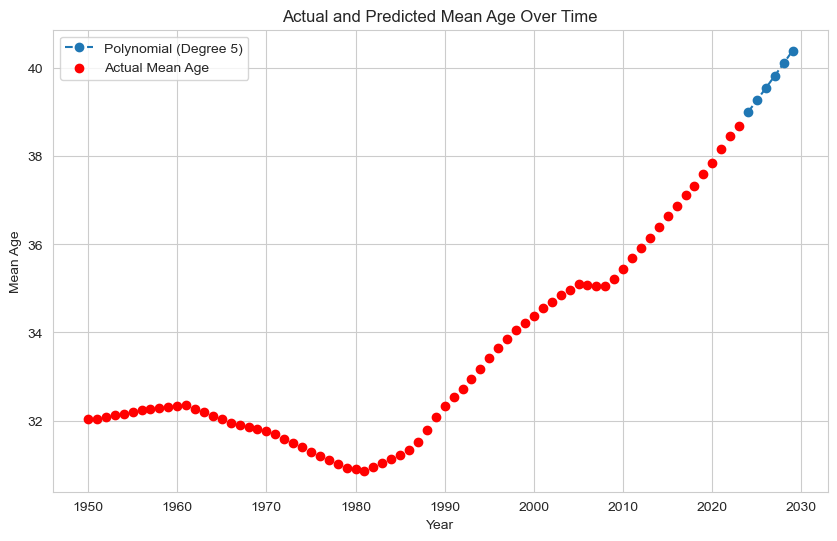
2.5.3. Hyperparameter Tuning

RandomizedSearchCV and GridSearchCV techniques tried to optimize model performance, contributing to improved accuracy and generalizability. But for my specific situation hand tuning worked better.



Fine-tuning by RandomizedSearchCV and GridSearchCV

Fine tuning results by RandomizedSearchCV and GridSearchCV were similar, but hand tuning results seems to be more accurate

Handmade fine-tuning

2.5.4. Relevance of Findings

Mean age predictions provide valuable insights into population dynamics over time. 'Baby Boom' cluster identification contributes historical context, though limitations in nuanced clustering are noted.

2.5.5. Factors for Effectiveness

Model selection, data quality, and hyperparameter tuning emerged as pivotal factors for study effectiveness. The study's recognition of limitations and exploration of alternative approaches enhance its robustness.

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>.
2. München, L.-M.-U., Medieninformatik, L., Butz, A. and Schmidt, A. (n.d.). Vorlesung Advanced Topics in HCI (Mensch-Maschine-Interaktion 2). [online] Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=93c58619b93028d09c71f4a345a2bc19f7cb75c7 [Accessed 12 Nov. 2023].