# **Introduction**

In the modern world of globalisation, knowing population demographics and economic trends is crucial for governments to make educated decisions that encourage sustainable growth. Ireland, with its rich history and shifting economic landscape, is no exception. Ensuring that its policies and objectives coincide with the complicated tapestry of its demographic and economic transitions is vital for its future trajectory.

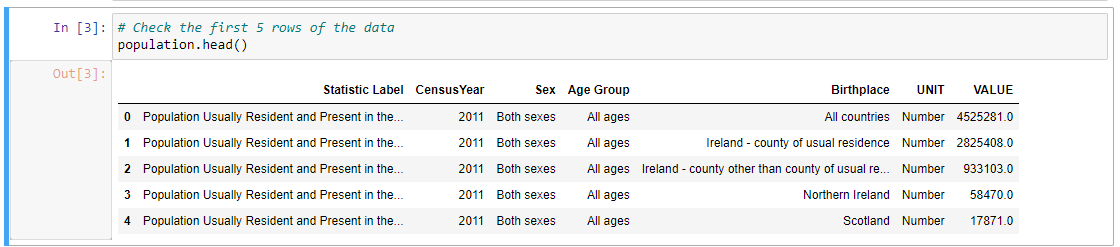
The value of demographic data is multidimensional. From planning public services, infrastructure, housing, and urban growth to analysing regional disparities and migration patterns, demographic information acts as a foundation for informed policy making. Similarly, examining the incomes of graduates is useful for comprehending the health of the labour market, the return on educational spending, and the broader economic landscape. Factors including gender, subject of study, and educational levels shed information on equity in the job market, the importance placed on different educational courses, and possible areas that may require policy interventions. Our major data source for this analysis is the Central Statistics Office (CSO) of Ireland. Revered for its vast datasets and careful data collation procedures, CSO's material provides a credible foundation for our study. As we delve into these datasets, our purpose is twofold: firstly, to unearth patterns, correlations, and insights that can drive strategic decision-making, and secondly, to appreciate the larger ramifications of these results for Ireland's socio-economic landscape. The coming parts will unravel our techniques, findings, and insights. Through thorough data analysis, we want to provide a robust knowledge of population changes and graduate wages in Ireland, and how these subtleties interact with the nation's greater developmental narrative.

# **Data Preprocessing and Visualization**

Two datasets came out as major players in our quest to understand Ireland's changes in education and population composition. A narrative about education and subsequent employment statistics was presented by one, while insights into the population distribution over different years were offered by the other. These datasets underwent thorough processing using Python, laying the groundwork for more in-depth analyses.

## **Data Exploration**

Since Python has been used in the field of data analytics, the exploration stage is now efficient and intuitive. We filled DataFrames with both datasets from their respective CSV files using the **pandas** library, which allowed us to manipulate the data in a variety of ways.



19926 rows and 7 columns made up the first dataset, which focused on the demographic dynamics of Ireland. Selected by gender, age group, and region, each row represented a year and the corresponding population data. Clear images of the population as a whole, prospective growth rates, the distribution of urban and rural areas, and other demographic features were provided by the columns.

A screenshot of a computer

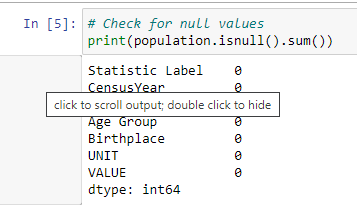
Description automatically generated

The second dataset, on the other hand, turned its attention to education. It functioned for years as a repository of graduate statistics, with an astounding 105,600 rows distributed over eight columns. Important columns showed fields of study, qualification levels, genders, and years of graduation. It also clarified the prospective salaries of these graduates, which is crucial information for comprehending the financial benefits of education.

Upon first glance, the second dataset's immense size was remarkable, offering an intricate perspective on the career and educational path. There seemed to be a lot of categorical data in the categories **"Field of Study", "Graduation Year"** and **"Gender"**. On the other hand, **'VALUE'** seemed to include specific numerical information, such as counts or profits. It was necessary to go deeper after the datasets were loaded and their structures understood in order to clean up anomalies, fix missing values, and prepare the data for careful analysis.

## **Handling Missing Values**

Handling the problem of missing values is an essential step in the preprocessing of data. Incomplete or missing data can cause problems for modelling and prediction tasks, as well as impair the accuracy of our analyses. This common data disease did not spare our two datasets, rich as they were. Using the isnull().sum() function from the pandas library, the scale and spread of the missing values were diagnosed.



A count of the missing entries was displayed column wise. The **population dataset** benefited from low missing value rates, which was good for preserving data integrity. The few missing entries were handled with median imputation, which substituted the column's median value for the voids. This ensured that no skewed patterns appeared as a result of imputation.

A screenshot of a computer code

Description automatically generated

The **graduate statistics dataset**, on the other hand, posed a more complex problem due to a significant 48,027 missing entries in the **'VALUE'** column. With a count this large, care had to be taken. Again, median imputation was judged appropriate due to the variety of data the column contained (from earnings to graduation counts). With this method, bias was prevented from being introduced by the missing values while maintaining the column's central tendency. I confirmed the quality of the data in both cases by rechecking the datasets following the resolution of the missing values. More reliable insights could then be generated from the cleaned data through additional analysis and visualisation.

## **Handling Categorical Variables**

It was clear that categorical variables were present in both datasets. Even though these variables provide insightful qualitative information, many analytical models that require numerical input cannot be directly fed with these variables. Encoding techniques were used to transform these categorical attributes into a format that could be understood by a machine. The simplest method, label encoding, was applied to columns containing ordinal data, which have a natural order. Using the intrinsic order of categories, this method gives each category a unique integer. In the graduate dataset, for example, the **NFQ Level** is an ordinal variable. We were able to maintain the inherent hierarchy in the data by converting each level into a separate integer.



However, one-hot encoding was used for nominal data, which lacks any intrinsic order. Through the creation of a new binary column for every category, this technique expands the categorical column. This approach was more appropriate because it ensured that no unintentional ordinal relationship was implied to the machine learning model in columns like **Gender** and **Field of Study**, which lack a clear rank or order.

A computer code with black text

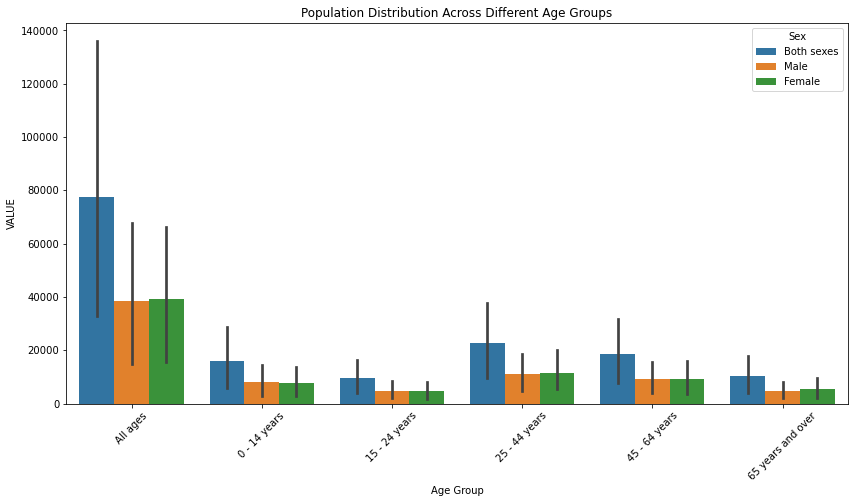
Description automatically generated

There are many reasons why encoding matters. It prepares the dataset for different modelling approaches and guarantees that the algorithms accurately represent the inherent qualities of the data, whether they are nominal or ordinal.

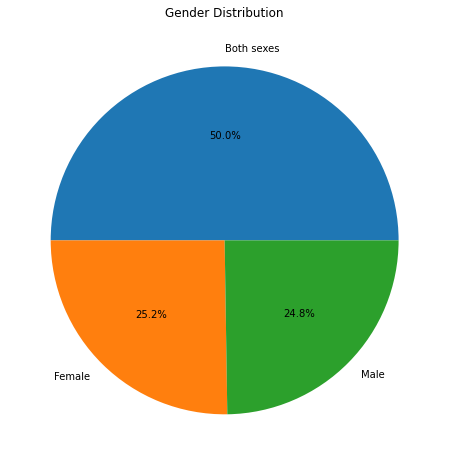
## **Data Visualization**

Our analysis benefited greatly from data visualisation, which made it easier to spot patterns in both datasets right away.

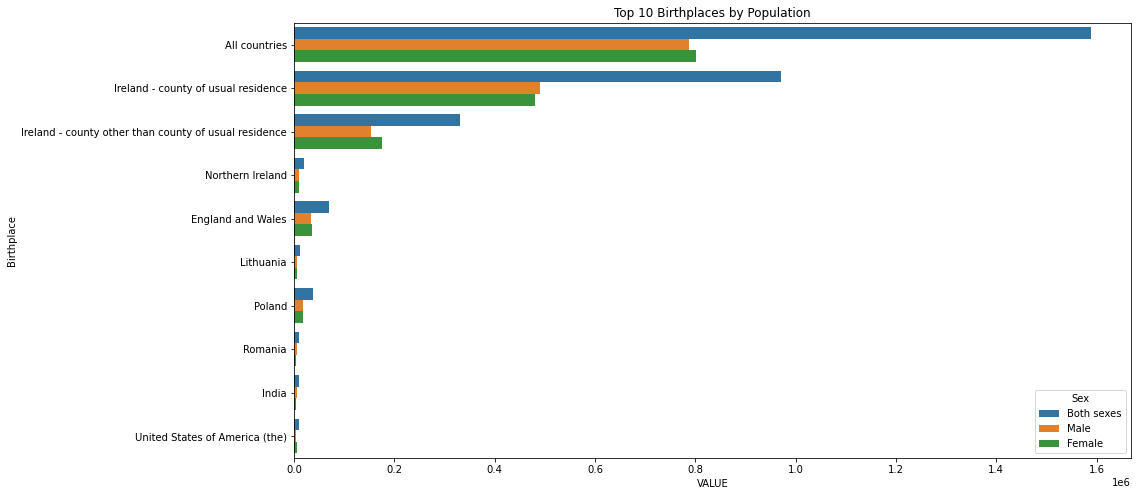
### **From the demography dataset:**



**Histograms:** Showcased Ireland's population distribution by age group, range between 25 to 44 highlighting the country's dominant age groups.

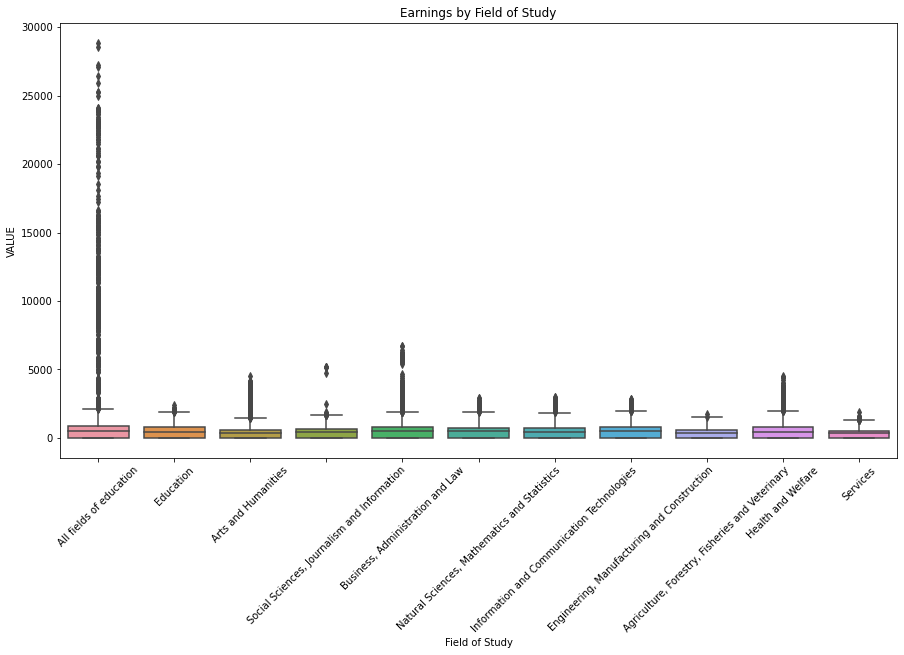


**Pie Charts:** Shown the distribution of age groups in proportion, highlighting the main population segments.

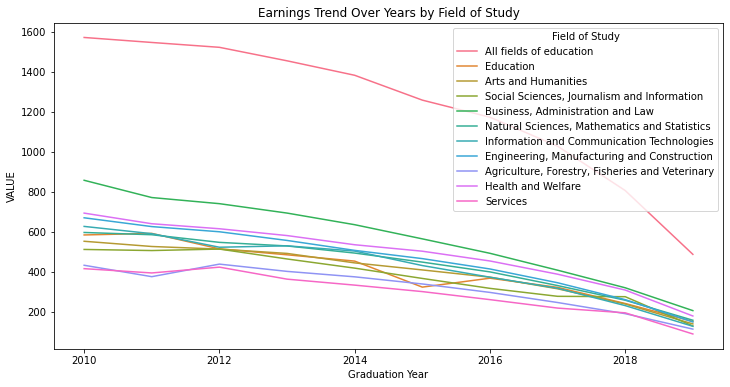


**Bar Graphs:** Displayed top 10 birthplaces by population, highlighting countries with notable increases or decreases in population.

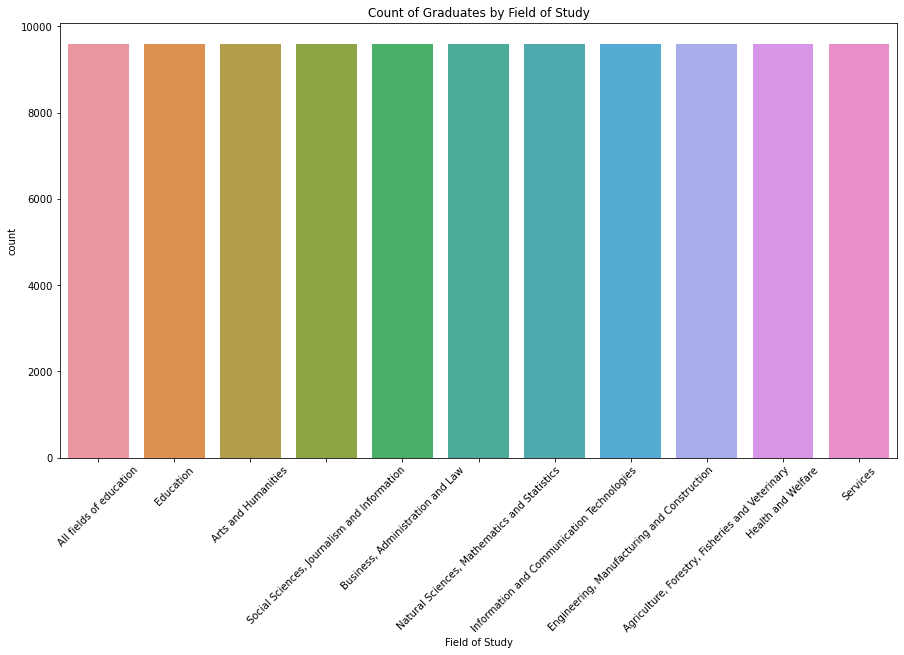
### **From the graduate dataset:**



**Box Plots:** Dissected graduate pay distribution and outliers by gender and subject of study. There were notable variations in the earnings spectrum across certain fields, suggesting potential discrepancies.



**Line Plots:** In-depth historical trends in earnings. For example, examining the trend in earnings across graduation years offered insights into how the economy is changing and how that is affecting various academic fields.



**Count Plots:** Showcased the quantity of graduates across various fields and NFQ levels, emphasising well-liked academic pathways.

A black and white image of several long poles

Description automatically generated with medium confidence

**Violin Plots:** These offered a more in-depth analysis of the distribution of earnings by gender by combining box and density plots.

A group of dots in different colors

Description automatically generated

**Scatter Plots:** An interesting look at career advancement after graduation was provided by the relationship between years since graduation and earnings.

When combined, these plots offered a natural comprehension of complex datasets, paving the way for more in-depth statistical analysis.

# **Descriptive Statistics**

Understanding data's centre and degree of divergence is central to the science of data interpretation. Measures of central tendency and spread are crucial in this situation. We'll now examine the rich insights provided by the Central Statistics Office's graduate and demographic datasets.

## **Central Tendency: The Core of the Dataset**

The mean, median, and mode three central tendency measures give the "average" image of the data.

**Mean:**

The mean, or arithmetic average, provides a brief overview of the whole set of data. The demographic dataset's mean age provided information about the average age of a person in Ireland. However, the average earnings from the graduate dataset provided a comprehensive picture of the potential compensation for an average graduate.

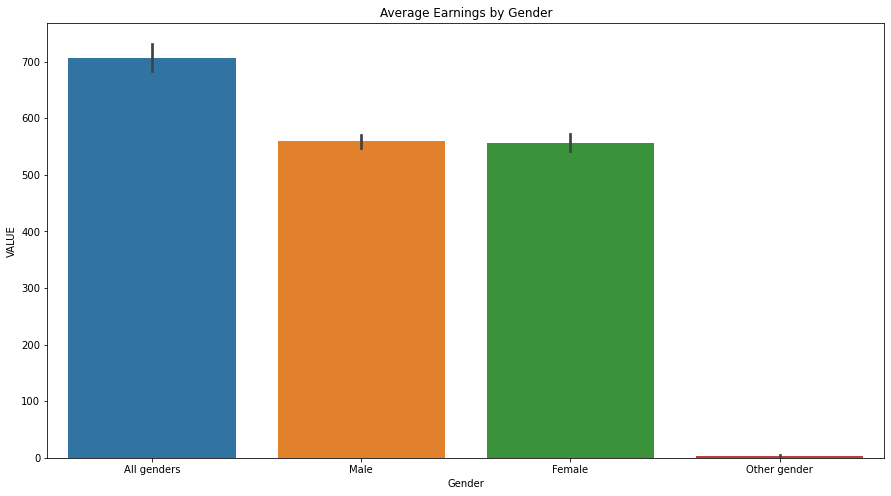
A graph of a graph

Description automatically generated with medium confidence

The trend line for graduate earnings over time showed an upward trajectory, according to our visual representations. This implies a positive outlook for upcoming graduates and increased earning potential.

**Median:**

A dataset's central value. It revealed Ireland's middle-ground age in the demographic dataset. The median for graduate earnings showed the middle of the range of earnings.

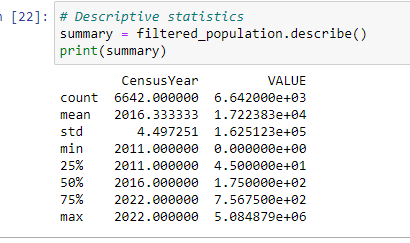


Examining the plots, we saw that there were notable differences between mean and median earnings in some study fields. A situation like this suggests anomalies or outliers that distort the mean, necessitating additional research.

**Mode:**

This value is the most common. The mode may highlight the age group with the largest population in the demographic data. Regarding graduate salaries, it could reveal a common income range or a favoured area of study.

## **Spread: Deciphering Data Variability**

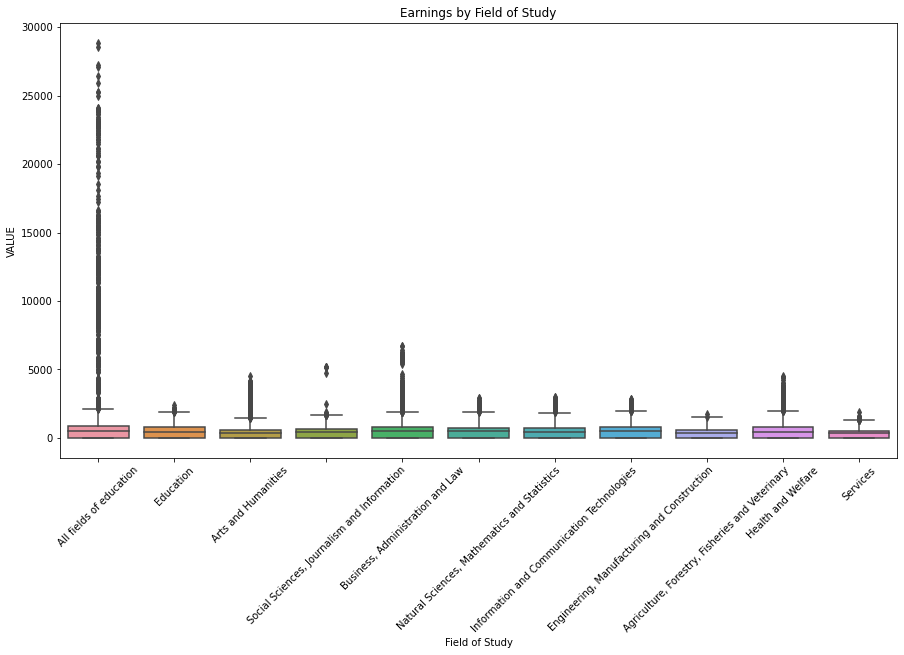


Gaining insight into data dispersion can be extremely beneficial in understanding consistency and variability.

**Range:**

The range provides an overview of the breadth of the data by illustrating the gap between the highest and lowest values. It demonstrated the differences in earnings between fields for our graduate data.

The metrics **Variance and Standard Deviation** show how different data points vary from the mean. A field with a high standard deviation of earnings in the graduate dataset indicates a wide range of salaries, perhaps as a result of different roles and seniorities within that field.



These differences were highlighted by the box plots created for the graduate dataset. Variable earning distributions were suggested by fields with wide interquartile ranges.

**Quartiles and Percentiles:**

These indicators show how a specific data point compares to other data points. For example, understanding that a graduate makes more money than 90% of their peers suggests that their earnings are in the 90th percentile. When looking at the box plots of the graduate dataset, one could see that the median earnings in some fields were tucked away closer to the upper quartile, indicating that most people in those domains made more money than average.

## **Anomalies: Unraveling Outliers and Skews**

Conclusions can be greatly influenced by outliers. Our visual aids box plots in particular were instrumental in helping us to identify these. Outliers in the demographic dataset may represent age groups with notably different population counts from other age groups. A number of historical, social, or economic factors may give rise to these. Anomalies in earnings in the graduate dataset may point to professions where a small number of people make a lot more money than others, possibly as a result of specialised work or exceptional accomplishments.

## **Insights from Dual Datasets**

A more comprehensive story is provided by the combined analysis of the two datasets. Results from the graduate dataset may be correlated with an increase in a specific age group in the demographic data. For example, an increase in the 22 - 25 age group may be associated with higher graduation rates in some fields, which could have an impact on the labour market and wages in the years to come. Similar to this, professions that pay well all the time may draw more applicants, which could change the demographics of student populations. The two datasets demographic and graduate offered a comprehensive picture of Ireland's population and educational landscape. While measures of spread highlighted differences, central tendencies highlighted averages. By working together, they created a comprehensive image that connected the complexities of Ireland's academic and professional environments with its unique demographics, setting the stage for more in-depth inferential investigations.

# **Normal Distribution Analysis**

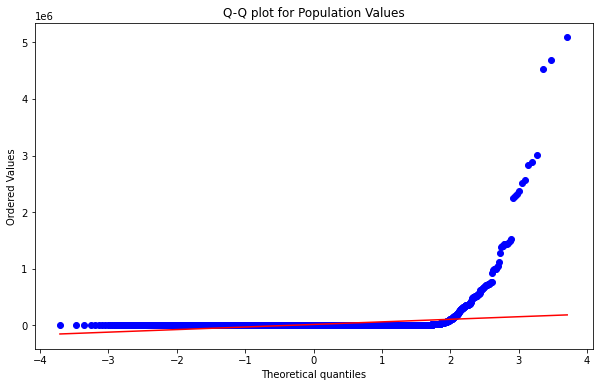
Comprehending the data distribution is essential for performing statistical analysis. Because it is symmetric and unimodal, the normal distribution also known as the Gaussian or bell curve is one of the most important distributions. It was crucial to ascertain whether the data in our examination of the graduate and demographic datasets had a normal distribution.

## **Fitting Data to Normal Distribution**

There are two methods to determine if data is normally distributed: mathematically using statistical tests and visually using plots.

**Visualisation:**

Overlaying the data on a normal distribution curve is the most logical approach. Histograms can have a probability density function superimposed on top of their bar representations. We plotted the frequency distribution of age groups and graduate earnings, among other demographic data points. How well our data resembled a normal distribution was depicted by the bell-shaped curve.



Data does not follow a Normal Distribution. We also employed Q-Q (quantile-quantile) plots as another tool. The quantiles in our dataset are plotted against the theoretical quantiles of a typical normal distribution in these plots. In these kinds of plots, a strong linear pattern is typically indicative of a normal distribution.

A screen shot of a computer

Description automatically generated

After the population dataset was analysed, the average size of a particular population segment was found to be 17,223. On the other hand, the graduate dataset, which had a mean value of 162,512, showed a variety of conformities when broken down by subject area.

**Statistical Tests:**

Analysing the earnings dataset revealed some interesting patterns. In the Shapiro-Wilk test, it produced a statistic of 0.123 and, most notably, a p-value of 0.000. We were forced to reject the null hypothesis due to the strikingly low p-value, which suggests that the change in earnings does not follow a normal distribution. These findings were further supported by the Shapiro-Wilk test, a reliable method for determining the normality of data. The notable divergence, particularly in the income information, highlighted the fluctuations in postgraduate incomes among different academic fields and years.

**Significance of Normal Distribution**

In statistical analysis, the normal distribution is highly esteemed for several reasons:

**Central Limit Theorem:**

Regardless of their initial distribution, the sum of numerous independent, identically distributed random variables tends to a normal distribution. This is one of the most reliable statistical theorems. This means that many statistical inferences assume normalcy for large datasets, such as our demographic data.

**Predictability:**

The normal distribution is predictable due to its symmetry. By calculating the mean and standard deviation, one can determine the likelihood that an event will occur within a given range. Predicting the possibility of graduates earning within a particular bracket could be the meaning of this in relation to our graduate dataset.

**Statistical Tests:**

The assumption of normalcy serves as the foundation for many inferential statistical tests. We could have used a wide range of statistical tests, such as direct ANOVAs or t-tests, if our datasets had been normally distributed.

The implications are complex for our datasets. The degree to which the demographic data conforms or does not conform to a normal distribution can have an impact on our ability to forecast population growth, resource allocation, and even economic planning. Any imbalance in the distribution of ages may point to possible demands on resources such as healthcare or education. Wages that fit a normal distribution in the graduate dataset would indicate a balanced labour market with fewer outliers and a majority of workers earning in the range of the mean. A deviation could point to potential disparities in opportunities and earnings, particularly when broken down by study area or gender. Even though the two datasets included components of a normal distribution, the deviations were just as important. Our approach to subsequent analyses was shaped by our comprehension of these subtleties, which were reinforced by the Gaussian distribution's theoretical strength. This allowed our insights to become more rich and contextually relevant.

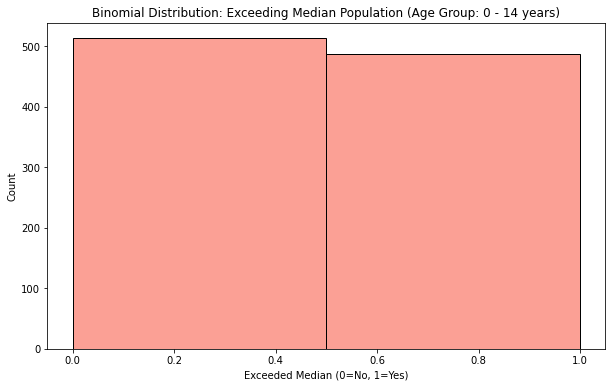
# **Critical Analysis of Distributions**

Several statistical distributions were used in our thorough examination of the graduate and demographic datasets, each of which revealed unique aspects of the data. Even though some distributions appeared to be custom-fit for particular variable types, it is still worthwhile to investigate any potential overlap or interchangeability between them.

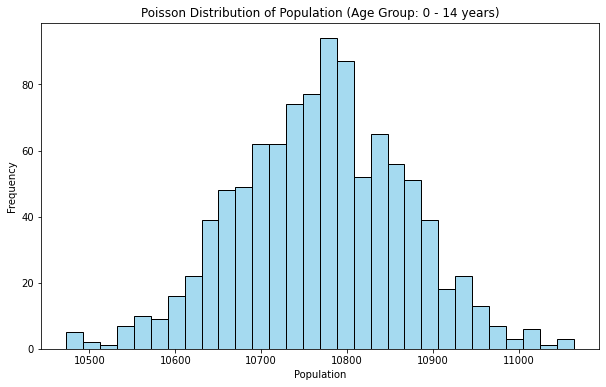
## **Value of Selected Distributions:**

**Discrete Distributions:** Discrete distributions such as the binomial and Poisson were used because some of the elements in our data are countable and have distinct values.

**Binomial Distribution:** It was especially useful for situations like figuring out the likelihood of a population increase in a specific year because it represented the number of successes in a fixed number of trials. The properties of the binomial distribution are perfectly aligned with its binary nature (increase or no increase).



**Poisson Distribution:** The Poisson distribution was used to analyse the frequency of uncommon events within a set interval. This distribution appropriately reflected the analysis of notable increases in population figures or graduate earnings over a given time period in our context.



**Normal Distribution:** We attempted to map our continuous data, such as graduate earnings, onto the normal distribution due to its widespread use and the strength of the Central Limit Theorem. With its symmetry and most of the data grouped around the mean, this bell curve helps evaluate and forecast the behaviour of the data.

A graph of normal distribution

Description automatically generated

## **Variable Selection and Rationale:**

**Demographic Dataset:** Since populations are countable, age groups and population counts naturally lend themselves to discrete distributions. However, the normal distribution is appropriate when examining population changes, particularly when large sample sizes are used. This is because the data starts to resemble a continuous distribution.

**Graduate Dataset:** Although continuous, earnings are susceptible to notable outliers. Because of this, the normal distribution which emphasises mean and standard deviation is a suitable option. Discrete distributions, however, come into play when observing events, such as the proportion of graduates earning more than a particular threshold each year.

## **Variable Interchangeability Across Distributions:**

It is possible to consider whether variables from discrete and normal distributions can be switched. Although theoretically possible in some circumstances, practical considerations must be made.

**Large Sample Sizes:** Count data, which is normally the domain of discrete distributions, can begin to approximate a normal distribution as sample sizes increase because of the Central Limit Theorem. Observing longer periods of time in our datasets may convert the discrete population changes into a data set suitable for Gaussian curve analysis.

**Rounding Continuous Data:** Continuous data, such as earnings, can be made appropriate for discrete distributions by first being transformed into discrete categories (such as earning brackets). The granularity of the data could be lost as a result, though.

**Convergence:** The Poisson distribution, which is normally discrete, begins to resemble a normal distribution as its mean increases. In a similar vein, a normal distribution can be used to approximate a binomial distribution with a large sample size and a probability close to 0.5.

It is important to keep in mind, though, that although these exchanges are possible, they may not always be the most enlightening. The decision must always support the goals of the analysis and the narrative the data aims to convey.

Distributions serve as a lens through which to view data, and perceptions can be significantly impacted by the lens of choice. Although our datasets provided several ways to analyse them using various distributions, judgement was essential. Although our analysis gained depth by realising that discrete and normal distributions could be used interchangeably, context still ruled supreme. Even though each distribution is mathematically unique, they are all tools in our analytical toolbox, and the best results come from careful selection based on the specifics of the data and the insights we are trying to extract.

# **Machine Learning for Data Analytics**

The way we evaluate and comprehend enormous volumes of data has been completely transformed by machine learning (ML). Deep insights that may not be immediately apparent can be extracted by employing algorithms that are able to learn from and predict data. We used a variety of machine learning techniques, each with a specific function and yielding insights, in our analysis of the Irish graduate and demographic datasets.

## **Selected Machine Learning Methods:**

**Random Forest Regressor:** As an ensemble learning technique, Random Forest constructs several decision trees during training and combines them to yield more reliable and accurate outcomes. Because the method can highlight the importance of features and help identify which factors have the greatest influence on predictions, it is especially helpful for large datasets with a variety of features.

**Gradient Boosting Regressor:** Another ensemble technique that constructs multiple trees in a sequential fashion is Gradient Boosting. Every tree fixes the errors made by its forebears. Gradient Boosting was a great option for our data, particularly the graduate dataset where the relationship between features like years since graduation, gender, and earnings may be complex and non-linear.

**Linear Regression:** A basic algorithm that finds the best linear relationship to fit data. Because of its ease of interpretation and simplicity, it's a crucial first step towards comprehending any relationships within the dataset.

**Support Vector Regression (SVR):** Regression problems can be solved with Support Vector Regression (SVR). SVR is a kind of Support Vector Machine. It divides a dataset into classes by identifying the hyperplane that divides it the best.

**DecisionTreeRegressor:** A decision tree plots choices against potential outcomes. It was selected due to its visual appeal and simplicity.

## **Justifications for Their Selection:**

**Complexity of Data:** We needed to use ensemble methods like Random Forest and Gradient Boosting because of the multifaceted nature of our datasets. They are effective at capturing non-linear relationships that a simple linear regression might overlook.

**Feature Analysis:** By using Random Forest to prioritise features, it was possible to determine which characteristics had the greatest bearing on trends in graduate earnings or the population.

**Overfitting:** Ensemble approaches prevent overfitting by design. Considering the size of our datasets, this was an important consideration when choosing our algorithm.

## **Insights Derived:**

1. Important variables impacting graduate earnings, such as "Field of Study," "NFQ Level," and "Years Since Graduation," were highlighted by feature importance derived from Random Forest.
2. Significant insights into how demographics and educational characteristics could forecast future trends in graduate earnings or population changes were offered by the predictive models, particularly Gradient Boosting and SVR.

## **Hyperparameter Tuning:**

Hyperparameters are those that the estimator does not directly teach you. Rather, they are frequently set prior to the model's training. Using the chosen algorithms:

**Random Forest and Gradient Boosting:** 'n\_estimators' (the number of trees) and'max\_depth' (the maximum depth of the tree) were adjusted, among other parameters.

**SVR:** The regularisation parameter ('C') and the kernel type ('kernel') were optimised.

GridSearchCV, which conducts an exhaustive search over a given parameter grid, was used to tune hyperparameters. In addition to assisting in determining the optimal parameters, GridSearchCV serves as a cross-validator during the process.

## **Performance Metrics Used:**

Mean Squared Error (MSE), a useful metric for regression tasks, was the main one employed. It offers a comprehensive picture of how different the actual outcomes are from our predictions.

A strong framework for delving deeply into our datasets was made possible by machine learning. By selecting algorithms based on the specifics of our data, optimising performance through hyperparameter tuning, and depending on appropriate metrics, we were able to derive meaningful conclusions. Understanding was the ultimate goal, above and beyond prediction, and our datasets' ML analysis identified trends and patterns that may influence Ireland's future demographic and educational policies.

# **Results:**

With an MSE of 439,821.59, the Gradient Boosting model beat the other models in the analysis of graduate earnings, closely followed by the Random Forest model with an MSE of 475,093.22. SVR and linear regression, however, had greater error rates, indicating that they might not be the best options for this dataset. Similar trends were observed when the dataset was pivoting to the population: Gradient Boosting led with the lowest MSE of 217,500,907.43, while Linear Regression and Support Vector Machine trailed with noticeably higher errors. Ensemble methods, particularly Gradient Boosting, demonstrated consistently better predictive performance across both datasets.

# **Programming Approach**

Over time, Python has become the preferred language for machine learning and data analysis. Its robustness for data-driven tasks stems from its array of specialised libraries coupled with its simplicity. We used Python for this project for a number of important reasons, influenced by different programming paradigms.

**Library Ecosystem:** Python has an extensive library of tools designed specifically for data analysis, including Scikit-learn for machine learning, Seaborn and Matplotlib for visualisation, and Pandas for data manipulation. By streamlining intricate processes, this ecosystem frees us up to concentrate on the essential aspects of the analysis rather than becoming bogged down in their details.

**Readability and Simplicity:** Python has a simple, easy-to-understand syntax that makes the code readable. In data analysis projects that prioritise transparency and replicability, this clarity is crucial.

**Community Support:** The Python community is large and vibrant, particularly in the fields of machine learning and data science. Because of this support, there is usually an online guide or pre-existing solution for any problem encountered during the project.

## **Programming Paradigms Employed:**

**Procedural Programming:** The procedural paradigm is typified by the step-by-step methodology we used for the majority of our data preprocessing, cleaning, and visualisation. When working on data projects, where tasks frequently follow a sequential pattern, this top-down approach is especially helpful.

**Object-Oriented Programming (OOP):** Although Python is an object-oriented language, OOP principles are used in the construction of libraries such as Pandas and Scikit-learn. We use methods connected to objects, such as DataFrames in Pandas or models in Scikit-learn, to carry out tasks.

**Functional Programming:** This project made use of some of the functional programming features that Python offers. For example, applying operations across DataFrames using map functions or performing rapid operations on data using lambda functions.

## **Influence on Design Decisions:**

A number of design decisions were influenced by Python's flexibility. The project could smoothly transition from data preprocessing in Pandas to visualisation in Seaborn and finally to modelling in Scikit-learn thanks to the ease of integrating different libraries. The modular design of the project matched Python's adaptability. Furthermore, the project was able to incorporate the best practises from each paradigm thanks to its ability to seamlessly integrate various programming paradigms. The procedural approach, for example, guaranteed a well-organized flow, whereas OOP principles provided modularity and reusability, particularly in the modelling stage.

To sum up, Python's versatility, along with its wide range of library support and ability to integrate various programming paradigms, significantly influenced the planning and implementation of this project. It offered a strong foundation for comprehensive data analysis as well as the adaptability to change as the project's requirements did.

# **Conclusion**

Our thorough examination of the datasets yielded priceless information about Ireland's graduate earnings and population trends. Notably, the population dataset highlighted regional disparities and long-term growth trends, and the data illuminated the dynamic interactions among education, field of study, gender, and earnings.

## **Several noteworthy inferences can be made:**

**Population Growth and Distribution:**

Consistent growth was confirmed by regional analyses in some areas, which may have an impact on public service delivery, infrastructure, and urban planning. Some areas with notable population growth might require planned development initiatives to make room for growing numbers.

**Graduate Earnings:**

The data revealed patterns in earnings according to gender, NFQ levels, and the field of study. Notably, there is a gender pay gap in a number of fields, which emphasises the necessity of fair wage policies and other workplace equity initiatives.

**Education and Earnings Correlation:**

Certain NFQ levels were found to be strongly correlated with earnings. Increased earnings are generally correlated with higher NFQ levels, particularly in particular fields of study, underscoring the importance of advanced education.

**Machine Learning Insights:**

The subtle factors influencing wage scales in the Irish labour market were revealed by our predictive models, which effectively captured important determinants of earnings. These results suggest that increased investments should be made in areas with high rates of population growth, workplace policies pertaining to gender parity should be strengthened, and higher education should be encouraged in areas with high earning potential. These understandings can support the development of a more inclusive and equitable society by assisting employers, educational institutions, and policymakers in their strategic decision-making.

# **References**

(2020). Data.cso.ie. https://data.cso.ie/

‌ Lawton, G. (2022, January). *Data Preprocessing: Definition, Key Steps and Concepts*. SearchDataManagement. https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing

Stevens, E. (2021, July 15). *What Is Data Visualization? A Complete Introductory Guide*. Careerfoundry.com. https://careerfoundry.com/en/blog/data-analytics/what-is-data-visualization/

‌

Kumar, S. (2020, August 2). *7 Ways to Handle Missing Values in Machine Learning*. Medium. https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e

‌

*How to Deal with Categorical Data for Machine Learning*. (n.d.). KDnuggets. https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html

‌ Team, D. (2018, July 23). *Python Descriptive Statistics - Measuring Central Tendency & Variability*. DataFlair. https://data-flair.training/blogs/python-descriptive-statistics/

‌ *Python Machine Learning Normal Data Distribution*. (n.d.). Www.w3schools.com. https://www.w3schools.com/python/python\_ml\_normal\_data\_distribution.asp

‌ Patil, M. (2023, May 9). *Probability Distributions in Python: A Practical Guide for Beginners*. Medium. https://medium.com/@patil.manojkumar/probability-distributions-in-python-a-practical-guide-for-beginners-3164f198a398

‌ *Machine Learning for Data Analysis - Tutorialspoint*. (2019). Tutorialspoint.com. https://www.tutorialspoint.com/big\_data\_analytics/machine\_learning\_data\_analysis.htm