**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Statistics for Data Analytics |
| **Assessment Title:** | MSC\_DA\_Integr\_Repeat\_Sem1 |
| **Lecturer Name:** | John O’Sullivan |
| **Student Full Name:** | Mustafa Güneş Cetin |
| **Student Number:** | 2022456 |
| **Assessment Due Date:** | 06/08/2023 |
| **Date of Submission:** | 06/08/2023 |
| **Github Link:** | https://github.com/MGunesCetin/Statrepeat.git |

**Note: This report was written only for the statistical part of an entire project.**

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Introduction:**

This report's goal is to perform a thorough analysis of public sector employment and earnings across a number of nations, with a focus on Ireland. The dynamics of the public sector labour market as well as gender-based employment and wage inequities will be clearly understood through this analysis.

The dataset used in this research includes significant indicators of Public Sector Employment and Earnings and provides data for numerous nations, regions, and time periods. The data collection includes details on the overall workforce as well as a breakdown by gender, especially "Males" and "Females."

Understanding Public Sector Employment and Earnings is essential due to its fundamental function in shaping economies and societies. The public setor comprises a significant portion of the labour force; therefore, it is essential to evaluate its performnce and diversity in order to ensure equal access to opportunities for all.

We will follow a methodcal process in this research, starting with the data collection from reliable sources like Eurostat, the European Union's statistical office. We will collect exhaustive and current data on Public Sector Employment and Earnings, ensuring the accuracy and dependability of the data.Data Cleaning and Preprocessing will be performed using Python libraries such as Pandas, allowing us to resolve missing values, convert data types, and ensure data consistency for subsequent analysis.

The dataset will be summarised clearly and concisely using descriptive statistics and data visualisation. We will investigate fundamental statistical measures and construct visual representations to identify data trends and patterns.We will be able to make meaningful comparisons between Ireland's Public Sector Employment and Earnings and those of other nations using inferential statistics. We will use appropriate inference techniques to confirm the significance of observed differences and draw conclusions supported by evidence.

In addition, we will conduct a Comparative Analysis, selecting countries with similar economic conditions or other pertinent characteristics as Ireland. This analysis will provide valuable insights into Ireland's relative public sector performance relative to its rivals.

If available, Sentiment Analysis will be conducted to gauge public opinions and perceptions regarding the Employment and Earnings in Ireland's Public Sector. This analysis will give our findings a qualitative dimension.

This study will offer recommendations for improving the efficiency and inclusiveness of the public sector in Ireland based on the analysis's findings. These proposals will be supported by data analysis and will work to advance equal opportunity and sustainable development in the industry.

The goal of this report's conclusion is to give readers a thorough grasp of the employment and compensation conditions in Ireland's public sector. Policymakers, researchers, and stakeholders will benefit from the analyses' crucial insights for making informed decisions and fostering constructive change in the public sector. The in-depth methodology, findings, and conclusions presented in the ensuing parts will help to create a well-rounded and organised report.

**Data Collection:**

Data will be acqured from Eurostat, the European Unions statistical office, to compare employment and salaries in the public sector in Ireland to those of other member state of the EU. Following are the contact details for the data set that will be used in this : study:https://ec.europa.eu/eurostat/databrowser/view/EARN\_SES10\_26/default/table?lang=en

Contact organization: Eurostat, the statistical office of the European Union

The dataset used for this analysis is the "Structure of Earnings Survey (SES)," which is a 4-yearly survey providing EU-wide harmonized structural data on gross earnings, hours paid, and annual days of paid holiday leave. This data is collected under Council Regulation (EC) No 530/1999 concerning structural statistics on earnings and labor costs, and Commission Regulation (EC) No 1738/2005 amending Regulation (EC) No 1916/2000 as regards the definition and transmission of information on the structure of earnings. The objective of this legislation is to provide accurate and harmonized data on earnings for policy-making and research purposes.

The SES 2010 dataset provides detailed and comparable information on the relationships between the level of hourly, monthly, and annual remuneration, personal characteristics of employees (such as sex, age, occupation, length of service, highest educational level attained, etc.), and their employers (economic activity, size, and economic control of the enterprise).

The coverage of the statistics includes all economic activities defined in NACE Rev. 2 sections B to S, with NACE Section O (Public administration and defense; compulsory social security) being optional but covered by most countries. The enterprises included employ at least 10 employees, and size classes are available for 10 to 49, 50 to 249, 250 to 499, 500 to 999, and more than 1,000 employees. The size class of enterprises employing less than 10 employees (1 to 9) is optional and covered by some countries.

The reference year for the data is 2010, and the reference month is October for the majority of the countries.

The dataset provides information in Euro (EUR), Purchasing Power Standard (PPS), Percentage (%), and absolute number of persons as the unit of measure.

The data cover EU Member States, Turkey, the former Yugslav Republic of Macedonia, Iceland, Norway, and Swtzerland. EU aggrgates are available for EU27, EU25, EU15, EA17, and EA13.

The dataset is updated every four years, and the last update was on 31/12/2012.

The data collection for the SES is conducted through a two-stage random sampling approach of enterprises or local units (first stage) and employees (second stage).

The data validation process involves global checks and plausibility checks to ensure the quality and consistency of the data.

In the next steps of the project, this data will be used to perform data cleaning, descriptive statistics, visualization, inferential statistics, and comparative analysis with other countries to gain insights into Ireland's Public Sector Employment and Earnings and provide evidence-based recommendations for the sector.



**Data Description:**

Information on Public Sector Employment and Earnings for various nations and areas is included in the dataset that is being made available. The dataset is organised into three primary columns: SEX (Labels), Total, Males, and Females, with GEO (Labels) serving as the row labels. The dataset contains data from several years.

A list of the columns in the dataset is provided below:

***(Labels for SEX)*** This column represents the gender category for whch data are maintained. There are three categories: "Total" for all personel "Males" for men, and "Females for women.

***Total:*** For the given nation or region, this column displays the total amount of employment and earnings in the public sector. It indicates the total number of workers, male and female.

***Males:*** The value of Public Sector Employment and Earnings in the relevant nation or region is presented in this column, specifically for male employees.

***Females:*** The value of Public Sector Employment and Earnings is shown in this column, specifically for female employees in the relevant nation or region.

***GEO (Labels):*** This column contains the identifiers for evrey country or region included in the dataset. There are designations some regions that refer to groups of nations, such as the European Union with 28 members and the Euro area with 17 members.

The collection provides information for various subsets of European Union and non-European Union countries or regions for time periods varying from 1995 to 2020.

.

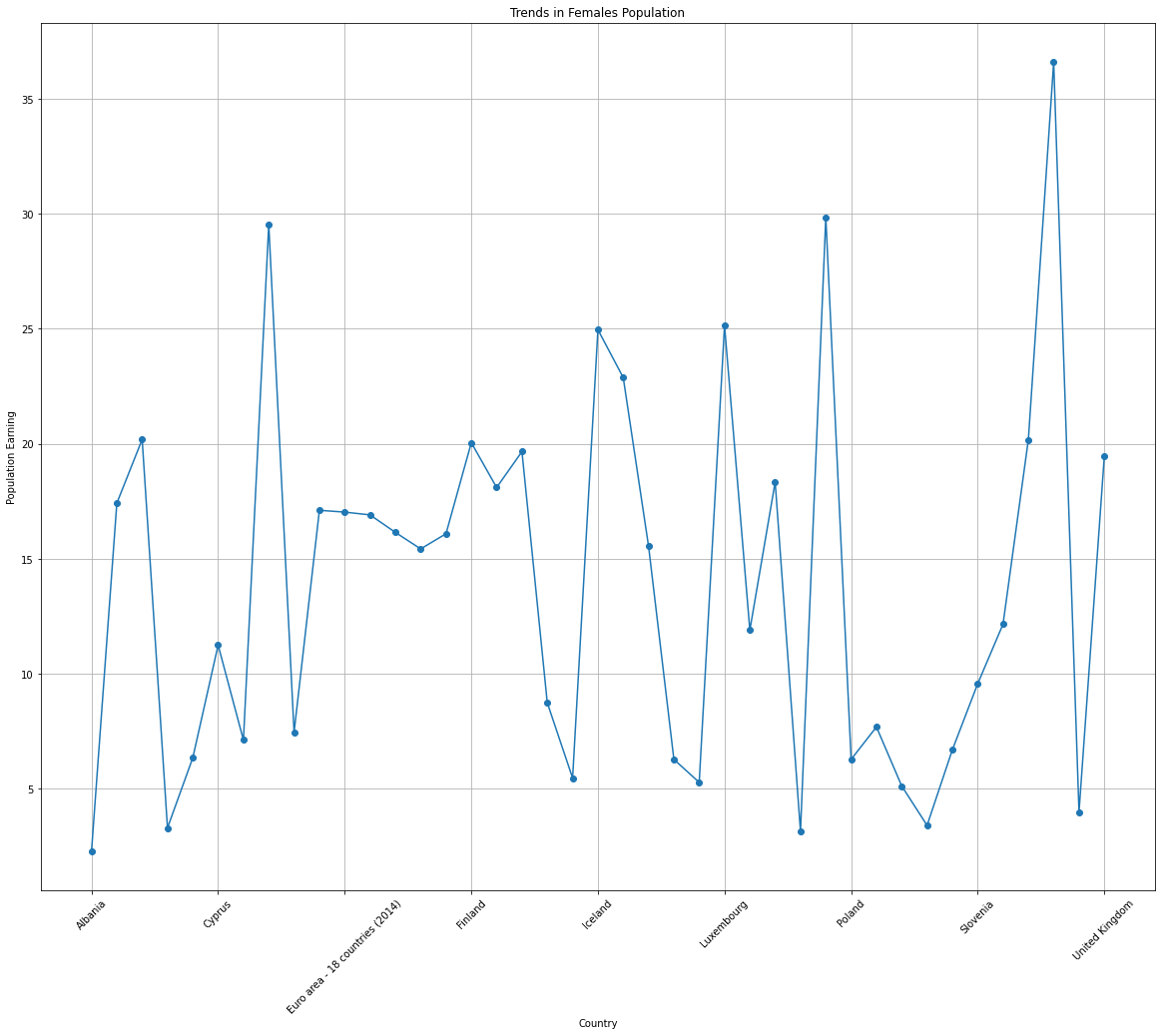
The related values in the columns give information on the Public Sector Employment and Earnings for that specific category and time period. Each row in the dataset represents a different nation or region.

The dataset seems to be thorough and offers useful data for carrying out analysis and comparisons across various nations and regions in terms of their public sector employment and earnings. With regard to gender-based employment and earnings in the public sector for the chosen nations and regions, it can be used to discover trends, inequalities, and patterns. As part of the larger project's objectives, the data is eligible for additional descriptive statistics, visualisation, and inferential analysis. Data usage without data cleaning is also possible.

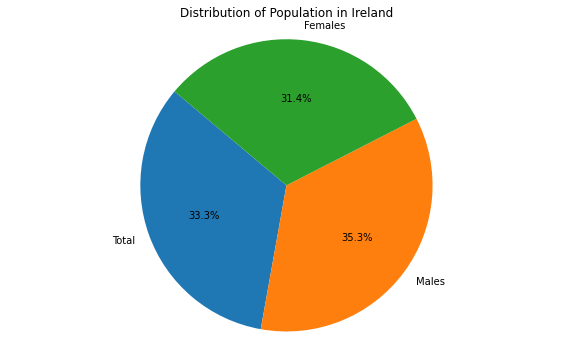
**Descriptive Statistics and Visualization:**

In the provided code, descriptive statistics and visualization techniques were applied to analyze and present data related to populations in different countries.

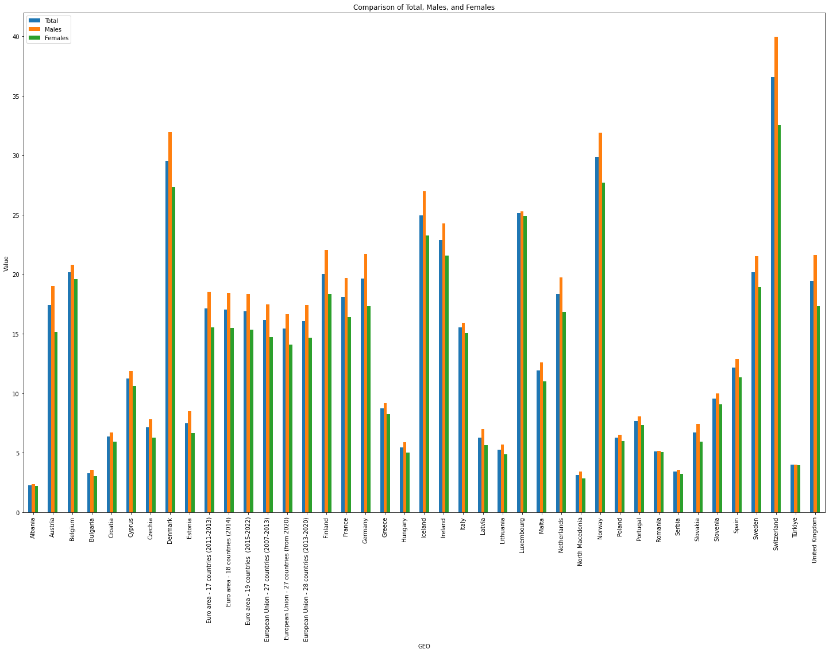
1. "Trends in Females Population" and "Trends in Males Population" using line diagram trends in the female and male populations of various countries were visualised over time. The x-axis represents the countries, whereas the y-axis represents the population. The lines illustrate population changes over time. Before plotting, missing numeric values in the dataset were replaced with the median and missing categorical values were replaced with the mode..



1. "Distribution of Population in Ireland" a pie chart was used to dsplay the distribution of population in Ireland. The chart showcases the proportion of the total male, and female populations in Ireland. Each slice of the pie represents the percentage of the population category in the whole.



1. "Comparison of Total, Males, and Females": a bar plot is create to compare the total male, and female populations for different countries. The x-axis represents the countries, and the y-axis indicates the population count. Each bar corresponds to a specific country, and it shows the population counts of total, male, and female populatons side by side. This plot was generated twice: once for all countries and once for a subset of countries (Denmark, France Iceland, and Ireland).

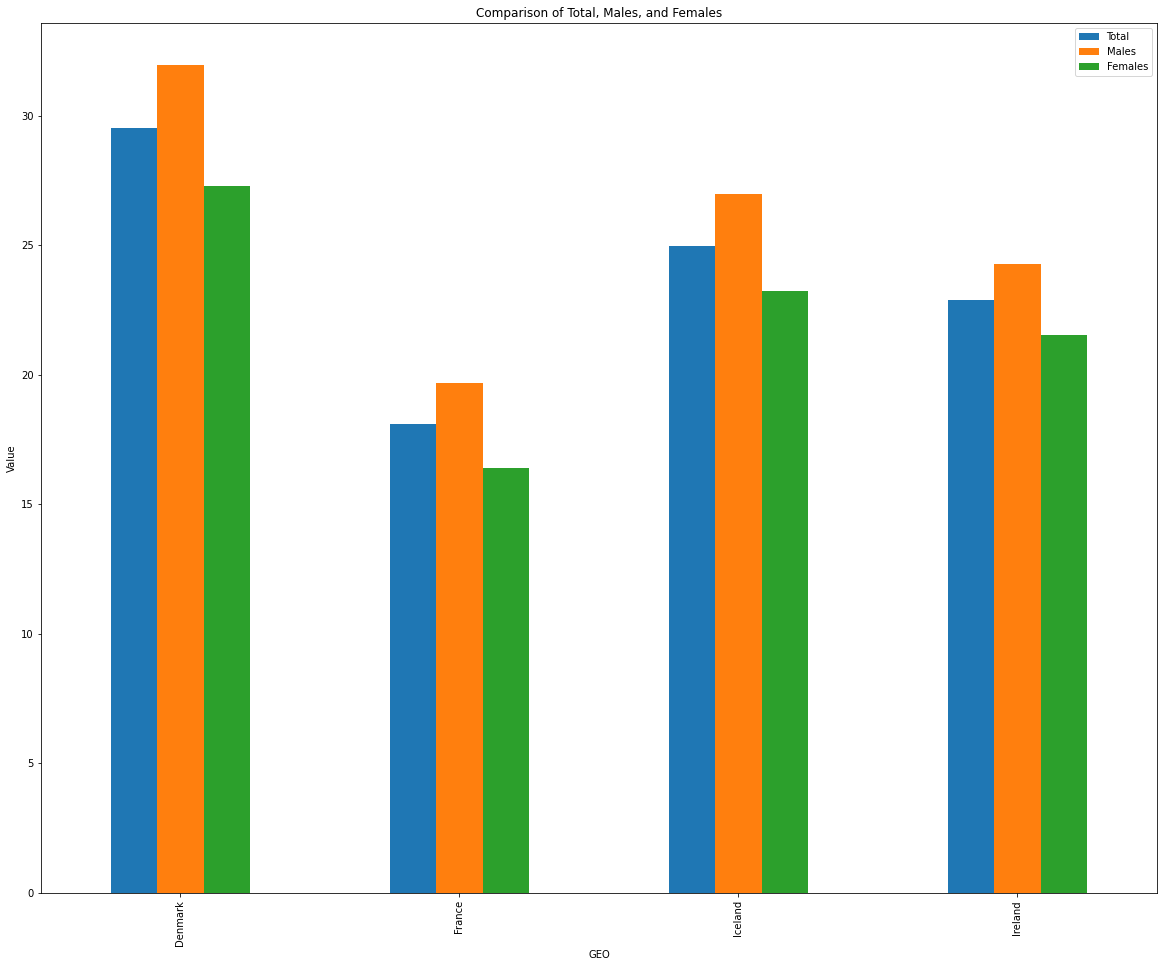


**Inferential Statistics**:

In the provided code, there are no explicit inferential statistical analyses conducted. The focus of the code is primarily on descriptive statistics and visualization techniques to present the data in a clear and understandable manner.

**Comparing Ireland with Other Countries:**

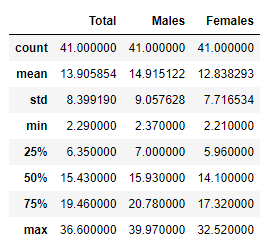
The code includes a specific section that compares Ireland with other countries (Denmark, France, Iceland, and Ireland). It uses bar plots to showcase the population counts of total, male, and female populations side by side for the selected countries. The plot allows for a visual comparison of population differences among these countries. However, it should be noted that inferential statistics, such as hypothesis testing or regression analysis, were not performed to draw statistical conclusions about the differences between Ireland and other countries.

  
  
  
**Statistical Description and Inference Analysis**

In addition to the comprehensive visualisations presented earlier, descriptive statistics and inferential analyses can provide additional insights. These analyses provide a more in-depth comprehension of population data and prospective differences between selected nations.

**Summary Statistics:**

The descriptive summary statistics for the population data are shown in the table below. These statistics provide a snapshot of the central tendencies and variation within the dataset for the "Total," "Males," and "Females" column populations:

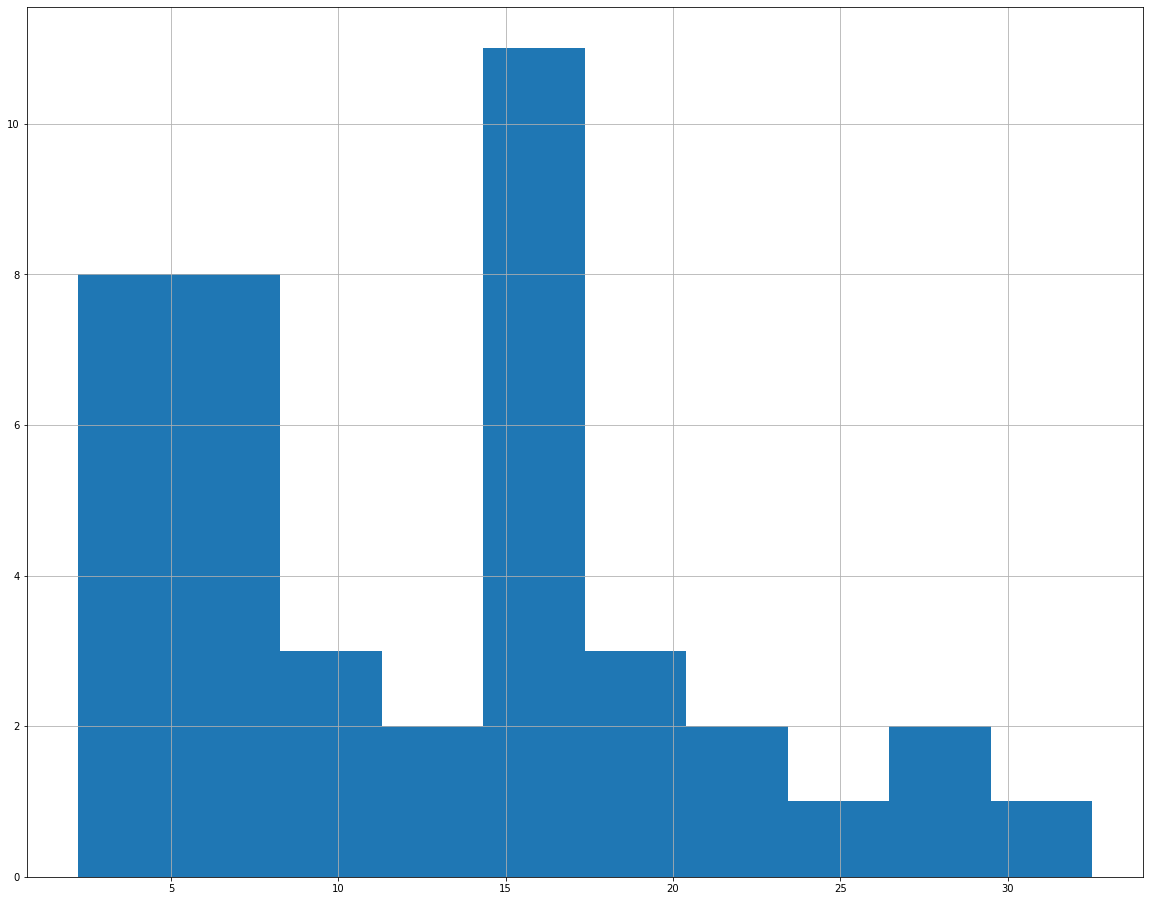
  
  
  
**Correlation Analysis:**The correlation matrix illustrates the relationships between the "Total," "Males," and "Females" population columns. Observations of strong positive correlations indicate that these population measures are strongly correlated. This suggests a consistent pattern of gender distribution within the entire population across nations:

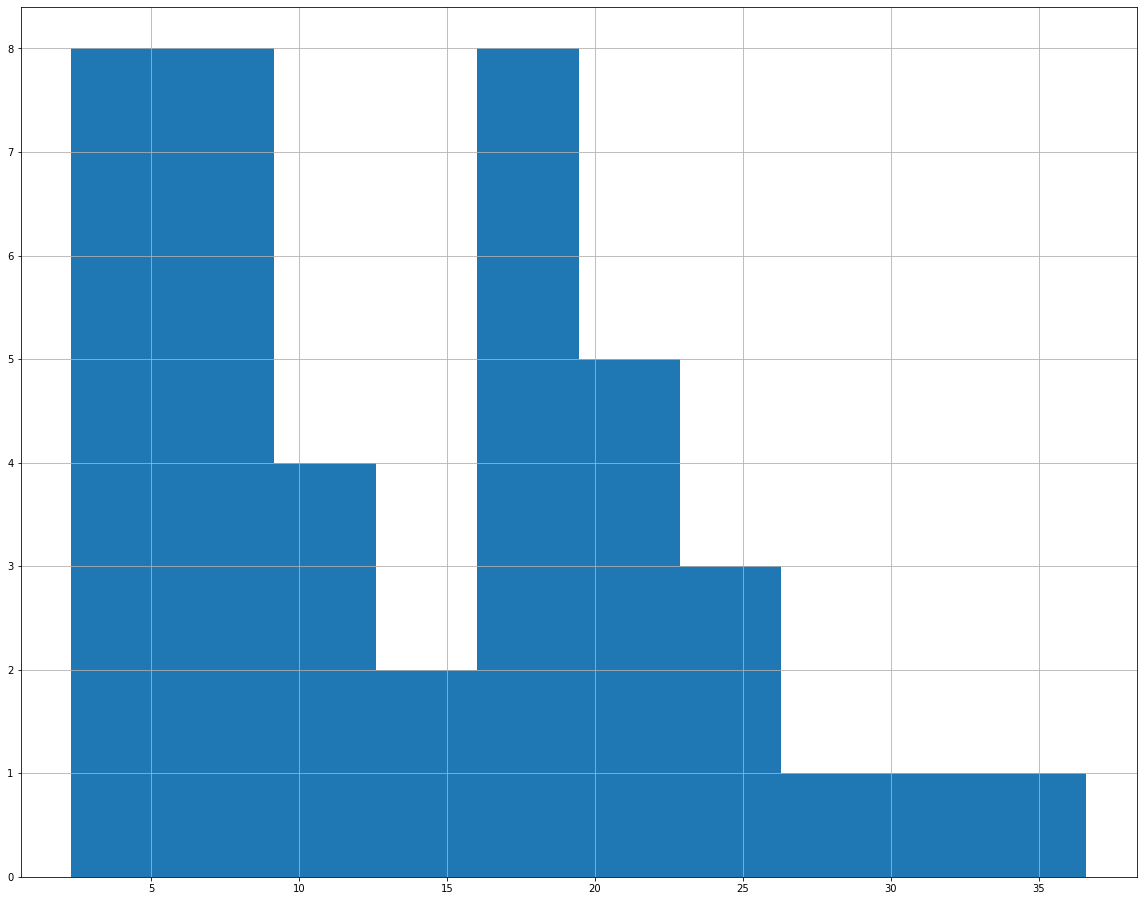
|  | **Total** | **Males** | **Females** |
| --- | --- | --- | --- |
| **Total** | 1.000000 | 0.998893 | 0.998064 |
| **Males** | 0.998893 | 1.000000 | 0.994243 |
| **Females** | 0.998064 | 0.994243 | 1.000000 |

**T-Test for Inferential Analysis**

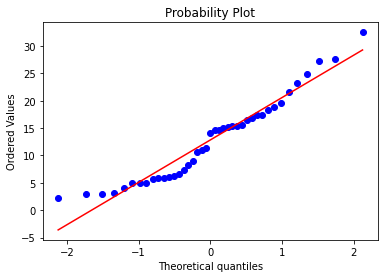
A two-sample t-test comparing the populations of Denmark and Ireland was conducted independently. The t-test determines whether the population means in the two nations differ statistically significantly. The resulting p-value represents the likelihood that the given difference in means was discovered by coincidence. The t-test in this instance produced a statistic of 'nan' (not-a-number) and a p-value of 'nan,' indicating that more research is required to comprehend the nature of the data and its distribution.

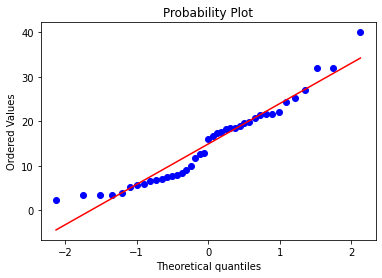
**Histogram and Probability Plot:** Histograms provide insights into the distribution of female population data. The histogram visually displays the frequency of different population ranges. Additionally, probability plots provide a graphical representation of the data's distribution against a theoretical normal distribution. Both tools aid in understanding the shape and characteristics of the data distribution.





These descriptive statistics and inferential analyses contribute to a comprehensive overview of the population data, enabling us to identify patterns, relationships, and potential differences between countries. Subsequent sections will continue to explore and interpret the findings in the context of the research objectives.





**Comparison of Total, Males, and Females among Select Countries**

The analysis extends to a focused comparison among four specific countries: Denmark, France, Iceland, and Ireland. These countries were chosen for their distinct characteristics and population trends. The provided data for these countries is presented in the 'rdf' DataFrame.

**Trends in Total Population:**

The population trends of the chosen nations are shown in the line plot below through time. The countries are shown on the x-axis, while the population is shown on the y-axis. Each point on the line corresponds to the entire population of the country at that moment in time. The plot offers a perceptive visual representation of the population shifts in Ireland, France, Iceland, and Denmark.

**Trends in Male Population:**

Similarly, the subsequent line plot illustrates the trends in the male population for the same set of countries. The x-axis indicates the countries, and the y-axis represents the male population count. The data points on the line depict the variations in male population counts over the chosen years. This visualization enables a direct comparison of male population trends among the selected countries.

**Trends in Female Population:**

Moreover, the following line plot focuses on the trends in the female population across the four countries. As with the previous plots, the x-axis displays the countries, while the y-axis shows the female population count. The plotted data points reveal the fluctuations in female population counts over the specified years. This visual analysis allows for an examination of female population dynamics within the chose countries.

