Project Report

# GitHub URL

<https://github.com/RuthAnderson95/UCDPA_RuthAnderson>

# Abstract

For my project I was unable to obtain information directly from my job due to GDPR as it contains sensitive information. I wanted to pick a dataframe that related closely to my current work, I therefore chose to work with Pensions Data which was available on data.gove.ie. I broke down the data, filtered and cleaned it and made visual representations and insights. I compared data from two separate years and merged these files together.

# Introduction

I chose to use this project case as the data relates to an area I work in. Working in a Life and Pension’s Wealth Management brokerage, I deal with DC, DB and other pension queries on a daily basis. It was very interesting to discover that this data set was readily available on the data.gov.ie website. I chose this data set as now when I apply these skills in my job, I will have a project already completed using the same type of data. The data was easy to access, from a trustworthy source and related uniquely to my current job role.

# Dataset

I first explored datasets on Kaggle, as this is an open source for dataframes, whilst exploring Kaggle I found a project on data on American Occupational Pensions, this spiked my interest and I began searching for Irish based pension datasets, I was unable to find any on Kaggle, and so following a google search I located the data.gov.ie website. The datasets were easily accessible on this website. I chose to pull data from this site as it is a government official website. I knew it was a reliable source for data as it was directly related to the Irish Government.

The first data set I downloaded was from 2020 data. It was an even split of male and female participants used. 162 men and 162 women. The data looks at people holding an occupational pension from their current employment, people who hold one only from their previous employment and also self employed people with an occupational pension. It defines the type of pension and the age range of people who hold these schemes. (<https://data.gov.ie/dataset/pmq32-persons-with-occupational-pension-coverage?package_type=dataset>)

The second data frame I downloaded was for a comparison. This data is from 2015 and shows similar information to the first data set. This data set was also downloaded from the data.gov.ie website. I was happy to download my second data source from the same website for the same above reasons. It is a trustworthy source.

This data shows information for both men and women holding different types of pensions, their age range and also includes a bracket of people who have not yet started or were unsure of the type of pension they held. It was interesting to see that there were people paying into a pension scheme but not knowing what type of scheme they were a part of.

(<https://data.gov.ie/dataset/pmq15-persons-with-an-occupational-pension?package_type=dataset>)

I downloaded these datasets as csv before uploading them to Jupyter Notebook using pandas.

# Implementation Process

(Describe your entire process in detail)

* I first ensured to important pandas and numpy to ensure I could evaluate the data accurately.
* Having previously downloaded my data set to a csv, I imported the csv using pd.read() giving my data set the name occ\_pen\_df.
* To ensure it imported I printed the dataframe. After confirming it has imported I used the .head() feature to show the first 20 rows of my dataframe.
* I could see NaN in some areas so I ran .isnull() to confirm what data was missing from my dataframe. This confirmed in a Boolean that the VALUE column had some missing data information.
* I ran occ\_pen\_df.VALUE.isnull() to confirm.
* To get the correct number of missing values I used the .sum() method. (occ\_pen\_df.isnull().sum()). This confirmed all missing data was in the VALUE column.132 values were missing.
* I chose to fill this data using interpolate. I renamed the dataframe with the updates VALUE column as occ\_pension\_updated\_df.
* I used Groupby, List and .mean() functions to further investigate the data.
* I used numpy to create an array called Gender\_Pen\_Type\_np. Which showed the type of pension for example Defined Benefit or Define Contribution and the gender of the participant beside it.
* I used .value\_counts() to define the data by sex.
* After ensuring the data for the 2020 csv was filtered and cleaned I imported data from a 2015 file for comparison.
* I downloaded the csv from data.gov.ie and then imported using pandas. I saved this as fifteen\_occ\_pen\_df
* I printed using .head() also to get a select overview of my dataframe.
* To compare the number of participants I used value\_counts() on the sex and discovered an even nu7mber of men and women but less participant data.
* Using (fifteen\_occ\_pen\_df.isnull().sum()) I was able to determine once again that only the VLAUE column has NaN values.
* I once again used interpolate to fill. This time I used the limit direction as backwards as the very first value was NaN and I wanted this to be filled. This was then saved as twentyfifteen.
* I determined the types of occupational pension using print(twentyfifteen['Type of Occupational Pension'].value\_counts()).
* I then compared this to the 2020 data using print(occ\_pension\_updated\_df['Pension Type'].value\_counts()) #value count to see difference to 2015
* I then began to merge the two data sets.
* I used .concat to merge my files. merged\_df = pd.concat([occ\_pension\_updated\_df,twentyfifteen], axis=1, ignore\_index=False)
* I saved this as merged\_df and used .head(20) for a snapshot of data.
* Happy with the results I then saved this as cleaned\_df.
* I saved the cleaned\_df as a csv.
* cleaned\_df.to\_csv('occupational-pension-merged.csv') and cleaned\_df.to\_csv(r"C:\Users\ruth\Desktop\PROJECT\occupational-pension-merged.csv")
* having cleaned the data set and merged them I then proceed to add visualization tools.
* Importing seaborn as sns and matplotlib.pyplot as plt to do this.
* I made three graphs from the data I had imported.
* The visualizations offer an insight into pension contributions and therefore also an insight into salaries.
* Results of these graphs are below.

# Results

(Include the charts and describe them)

sns.scatterplot(data=occ\_pension\_updated\_df, x='Sex', y = 'Pension Type')

A white rectangular chart with blue dots

Description automatically generated

Using the scatter plot on the first data set we can see the even number of participants.

sns.scatterplot(data=occ\_pension\_updated\_df, x='Pension Type', y = 'VALUE', hue='Sex')

plt.legend(bbox\_to\_anchor=(1.02, 0.15), loc='upper left', borderaxespad=0)

A graph with orange and blue dots

Description automatically generated

This graph shows the values held in each pension type from 2020. The hue is the Sex so we can clearly define which gender holds the greater values.

sns.scatterplot(data=twentyfifteen, x='Type of Occupational Pension', y = 'VALUE', hue='Sex')

plt.legend(bbox\_to\_anchor=(1.02, 0.15), loc='upper left', borderaxespad=0)

plt.xticks(rotation=45)

A graph with numbers and dots

Description automatically generated

This graph is similar to the previous but shows 2015 data. The 2015 data has extra types of occupational pensions. Once again the Hue is used to define Sex to make the graph easier to read.

# Insights

1. The data is limited but as there is an even number of each gender used it shows an insight into which gender holds more value in their pensions and also which pension is preferred by genders. This is interesting as men tend to make more money than women, and the graphs reflect that in the value of pension schemes being topped by males.
2. Women in the 2020 data set were more likely to be members of a hybrid pension scheme - a hybrid scheme is not fully DB (defined benefit) or DC (defined contribution) but contains some characteristics of both
3. DB pension values remained very similar between 2015 and 2020 – why? This is potentially due to the fact that DB scheme are in the majority of workplaces, no longer offered. DB schemes became too expensive for employers to offer as people now liv longer.
4. Men favored a DC pension plan – these are the most common occupational pension schemes in Ireland. The charts show an interesting insight into earnings and therefore into pension contributions and values.
5. The 2015 data set offers an interesting insight into peoples knowledge of their pension schemes. With the addition of the do not know/not started data. It is interesting to see that some people willing pay into a pension scheme without knowing what benefits they are paying into. It also is curious to see that perhaps people have opted not to begin their pensions, this could be for many reasons including not being in the company as defined amount of time, or their financial situation.
6. according to the Central Bank of Ireland there are 74,866 DC pension schemes and 597 DB schemes active.

# References

*Occupational pension funds in Ireland: What do we know* (no date) *Central Bank of Ireland*. Available at: https://www.centralbank.ie/statistics/statistical-publications/behind-the-data/occupational-pension-funds-in-ireland-what-do-we-know#:~:text=There%20are%2074%2C866%20DC%20and,the%20Irish%20pension%20fund%20sector. (Accessed: 16 July 2023).