FINTECH BOOTCAMP

PROJECT 1

Group Members

Roshan Paudyal
Asha Devi

Project Presentation Date : 17 January 2023

DISCLAIMER

Analysis conducted and the outcomes of this project are purely for learning and development purposes. We do not accept any responsibility for the accuracy, completeness or currency of the material included in this project and will not be liable for any loss or damage arising out of any use of, or reliance on, this project outcomes.

MOTIVATION & SUMMARY

Analysis on Superannuation products

- Superannuation – Money saved for retirement
- There is not much insight available regarding the superannuation products comparisons in the market.
- Australian annual superannuation data collection is still relatively new.
- APRA is still in its early phase of a new data collection, systems and processes for collecting and reporting data in accordance with the new reporting requirements and are not fully embedded across the industry.
- We thought to give a try and see what we find with the available data on APRA website.

Our project is about analyzing the Australian Superannuation Funds for period 30 June 2014 to 30 June 2022. Analysis is around demography, fund types, member benefit flows and financial performance.

QUESTIONS & DATA

Data source: Australian Prudential Regulation Authority (APRA) website

File type: Excel format

Questions Posed:

1. Which fund has the maximum number of members accounts?
2. Which gender group holds maximum number of fund accounts?
3. What is the trend in the member accounts by age group?
4. Which sector has the maximum and minimum number of funds?
5. Which sector received the maximum contribution and from whom?
6. Which fund has the highest investment income?
7. Which fund has the lowest investment income?

DATA CLEANUP & EXPLORATION

Data Exploration:

- Data file with multiple sheets and header rows – which we did not anticipated.
  - Coding challenges - multiple sheets and header rows

Data Clean-up:

- Some clean-up at the file level itself:
  - Removed the unwanted header rows and kept only one header row.
  - fixing blank fields and * in the cells, renaming the headers before attempting coding in Jupiter lab / visual studio.
- Clean-up in the Jupiter lab:
  - Retrieve DataFrame shape to view the total number of rows and columns.
  - Identify series count, check for null values, check duplicates, view the DataFrame.
  - Reordering the columns by creating a new DataFrame
  - Set index, sort values, rename

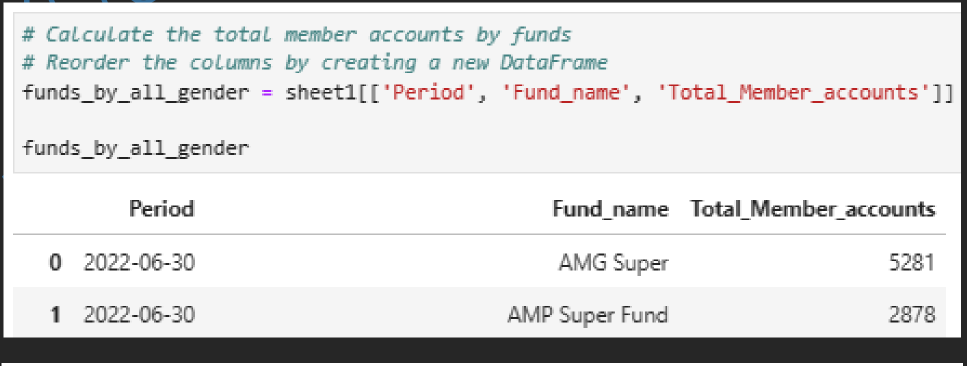
DATA ANALYSIS

Section 1 – Demographics analysis (gender and age)

Section 2 – Fund types, member benefit flows and financial performance analysis

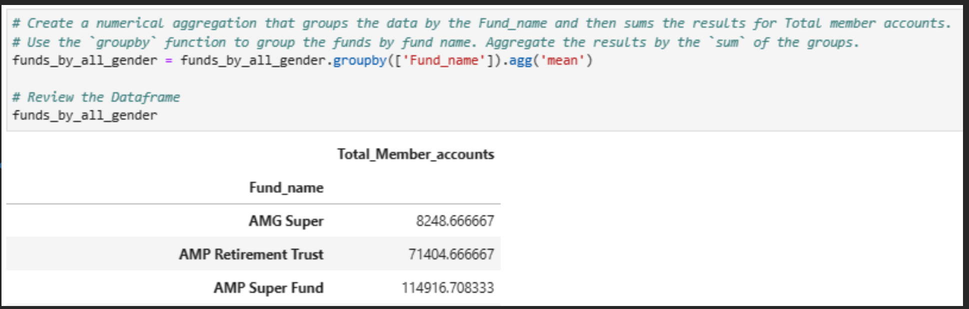
Data analysis continues…

DATA ANALYSIS



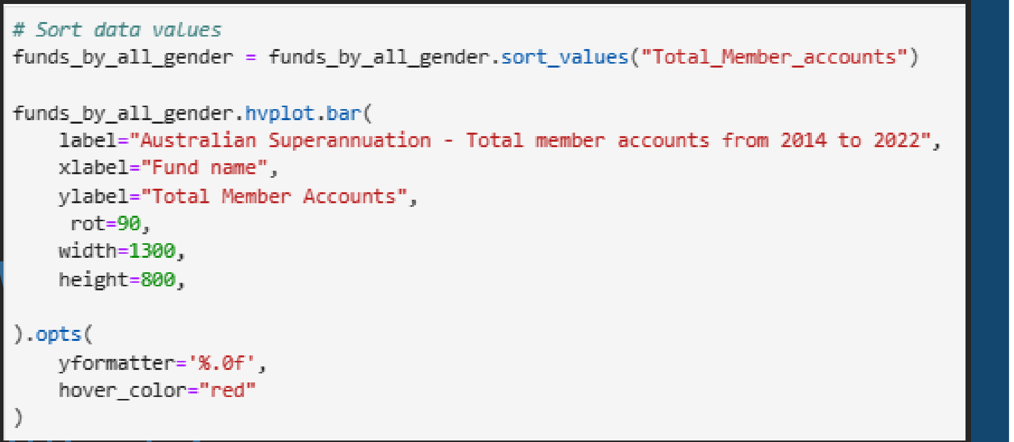
```
# Calculate the total member accounts by funds
# Reorder the columns by creating a new DataFrame
funds_by_all_gender = sheet1[['Period', 'Fund_name', 'Total_Member_accounts']]

funds_by_all_gender
```

```
       Period       Fund_name  Total_Member_accounts
0  2022-06-30       AMG Super                   5281
1  2022-06-30  AMP Super Fund                   2878
```



```
# Create a numerical aggregation that groups the data by the Fund_name and then sums the results for Total member accounts.
# Use the `groupby` function to group the funds by fund name. Aggregate the results by the `sum` of the groups.
funds_by_all_gender = funds_by_all_gender.groupby(['Fund_name']).agg('mean')

# Review the Dataframe
funds_by_all_gender
```

```
                     Total_Member_accounts
Fund_name
AMG Super                      8248.666667
AMP Retirement Trust          71404.666667
AMP Super Fund               114916.708333
```



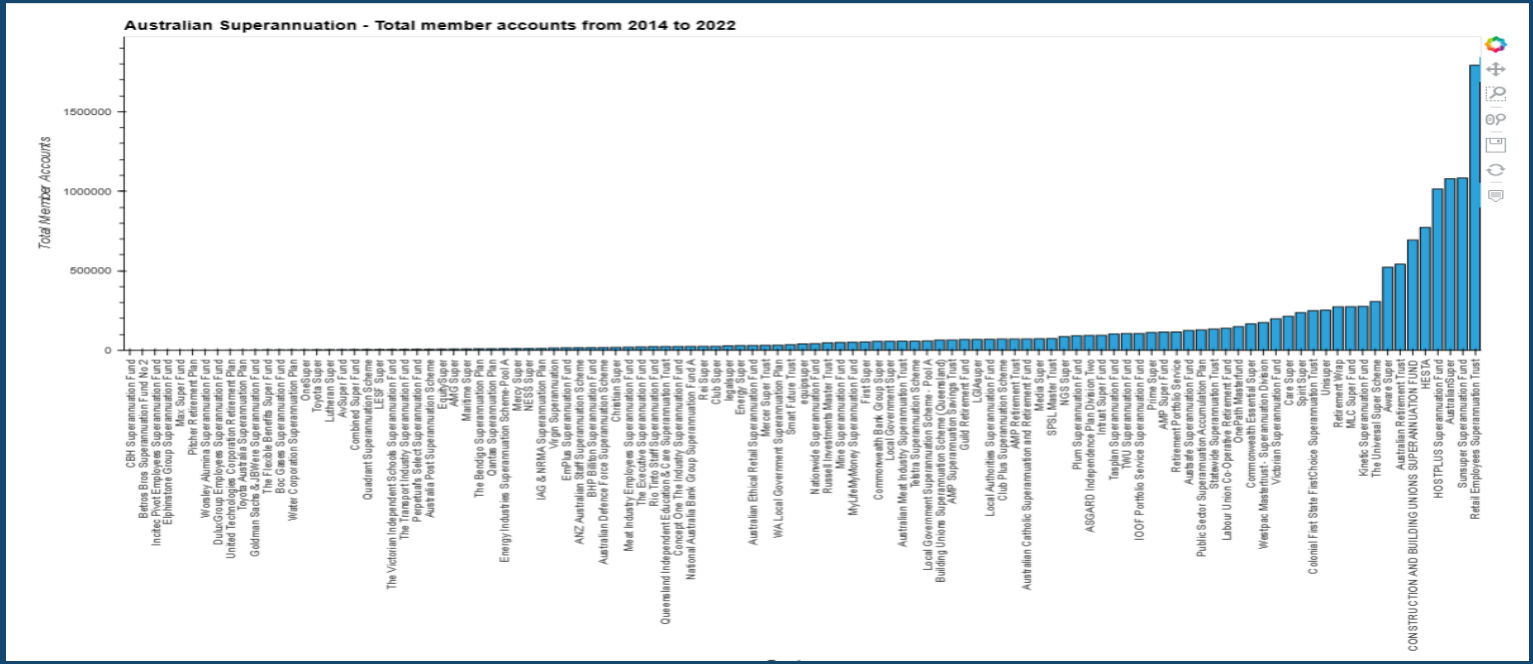
```
# Sort data values
funds_by_all_gender = funds_by_all_gender.sort_values("Total_Member_accounts")

funds_by_all_gender.hvplot.bar(
    label="Australian Superannuation - Total member accounts from 2014 to 2022",
    xlabel="Fund name",
    ylabel="Total Member Accounts",
     rot=90,
    width=1300,
    height=800,

).opts(
    yformatter='%.0f',
    hover_color="red"
)
```

Q1 - Which fund has the maximum number of members accounts?

Answer - Retail Employees Superannuation Trust

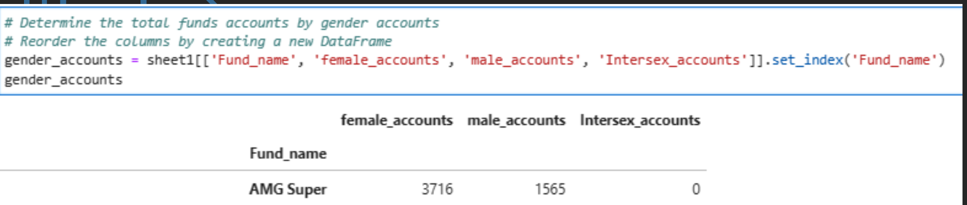
Steps taken to analyse the data:

- View the full DataFrame first
- Reorder the columns by creating a new Dataframe and select the columns that is relevant to this question.
- Create a numerical aggregation that groups the data by the Fund_name and then sums the results for Total member accounts.
- Sort data values and plot the visualisation



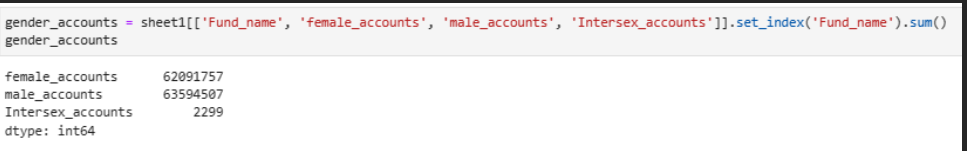
Data analysis continues…

DATA ANALYSIS



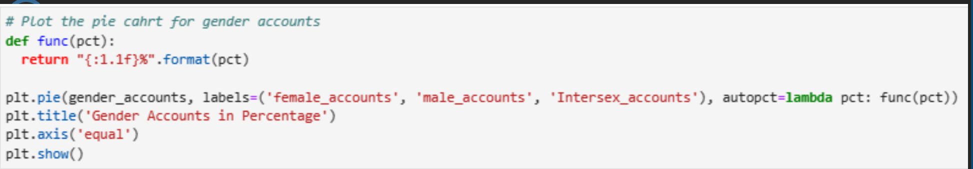
```
# Determine the total funds accounts by gender accounts
# Reorder the columns by creating a new DataFrame
gender_accounts = sheet1[['Fund_name', 'female_accounts', 'male_accounts', 'Intersex_accounts']].set_index('Fund_name')
gender_accounts
```

	female_accounts	male_accounts	Intersex_accounts
Fund_name			
AMG Super	3716	1565	0



```
gender_accounts = sheet1[['Fund_name', 'female_accounts', 'male_accounts', 'Intersex_accounts']].set_index('Fund_name').sum()
gender_accounts
```

```
female_accounts      62091757
male_accounts        63594507
Intersex_accounts        2299
dtype: int64
```



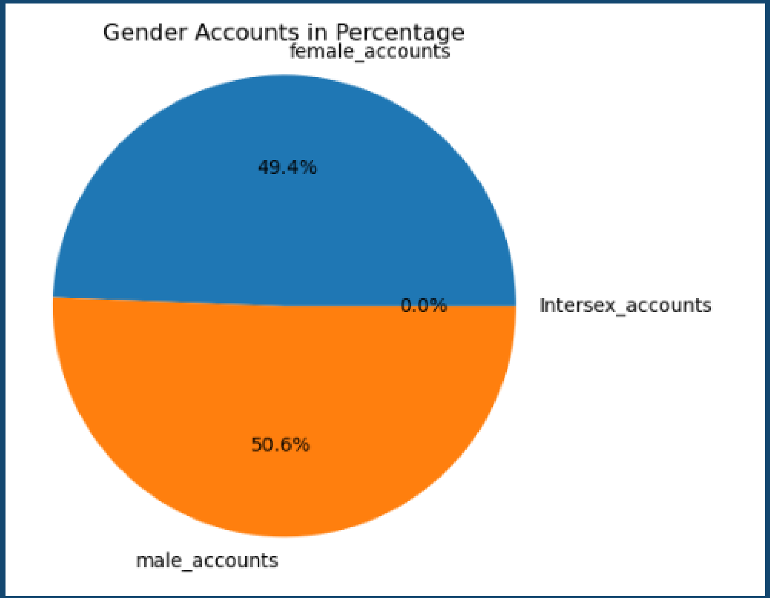
```
# Plot the pie cahrt for gender accounts
def func(pct):
  return "{:1.1f}%".format(pct)

plt.pie(gender_accounts, labels=('female_accounts', 'male_accounts', 'Intersex_accounts'), autopct=lambda pct: func(pct))
plt.title('Gender Accounts in Percentage')
plt.axis('equal')
plt.show()
```

Q2 - Which gender group holds maximum number of fund accounts?

Answer - Plot indicates that there is not much difference between the male and female total accounts whereas the intersex total accounts numbers are almost negligible.

Steps taken to analyse the data:

- Determine the total funds accounts by gender accounts
- Reorder the columns by creating a new DataFrame
- Plot the Pie chart for gender accounts



Data analysis continues…

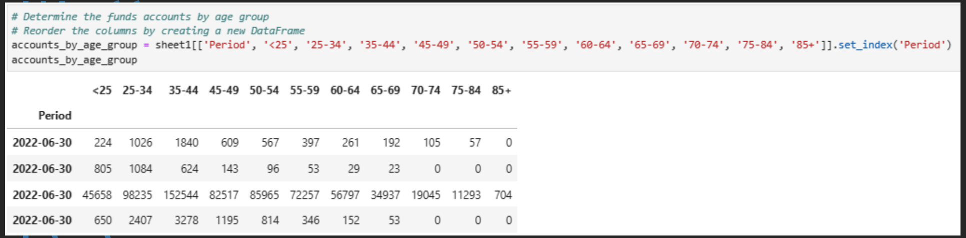
DATA ANALYSIS

Q3 - What is the trend in the member accounts by age group?

Answer - Plot shows that the funds for most of the age groups are trending upwards with period.

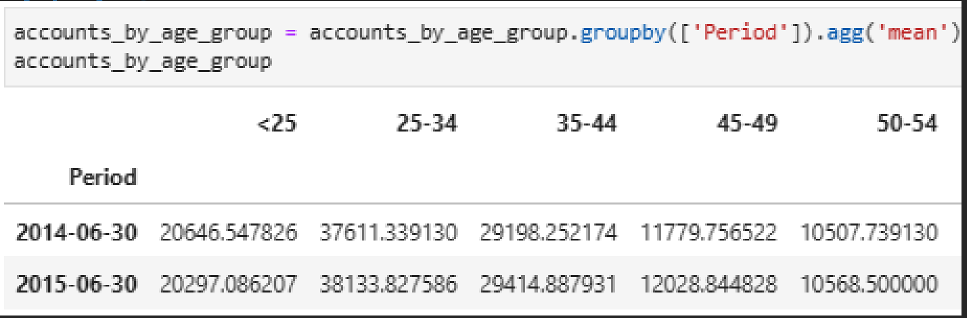
Steps taken to analyse the data:

- Reorder the columns by creating a new DataFrame
- Group the DataFrame by Period and aggregate the data by mean
- Use the hvplot function to plot the visualization to see the trend



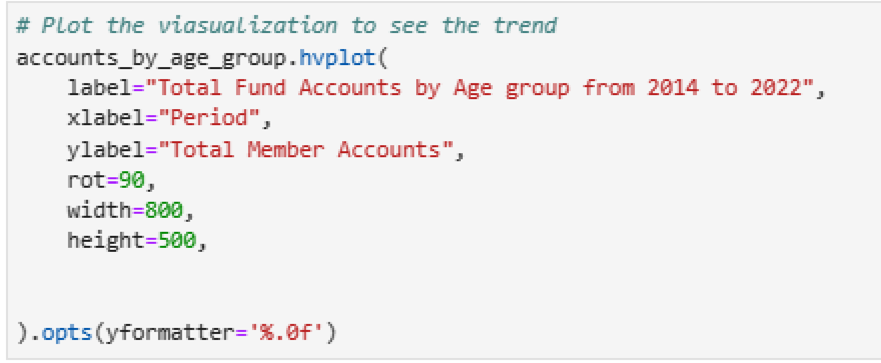
```
# Determine the funds accounts by age group
# Reorder the columns by creating a new DataFrame
accounts_by_age_group = sheet1[['Period', '<25', '25-34', '35-44', '45-49', '50-54', '55-59', '60-64', '65-69', '70-74', '75-84', '85+']].set_index('Period')
accounts_by_age_group
```

	<25	25-34	35-44	45-49	50-54	55-59	60-64	65-69	70-74	75-84	85+
Period											
2022-06-30	224	1026	1840	609	567	397	261	192	105	57	0
2022-06-30	805	1084	624	143	96	53	29	23	0	0	0
2022-06-30	45658	98235	152544	82517	85965	72257	56797	34937	19045	11293	704
2022-06-30	650	2407	3278	1195	814	346	152	53	0	0	0

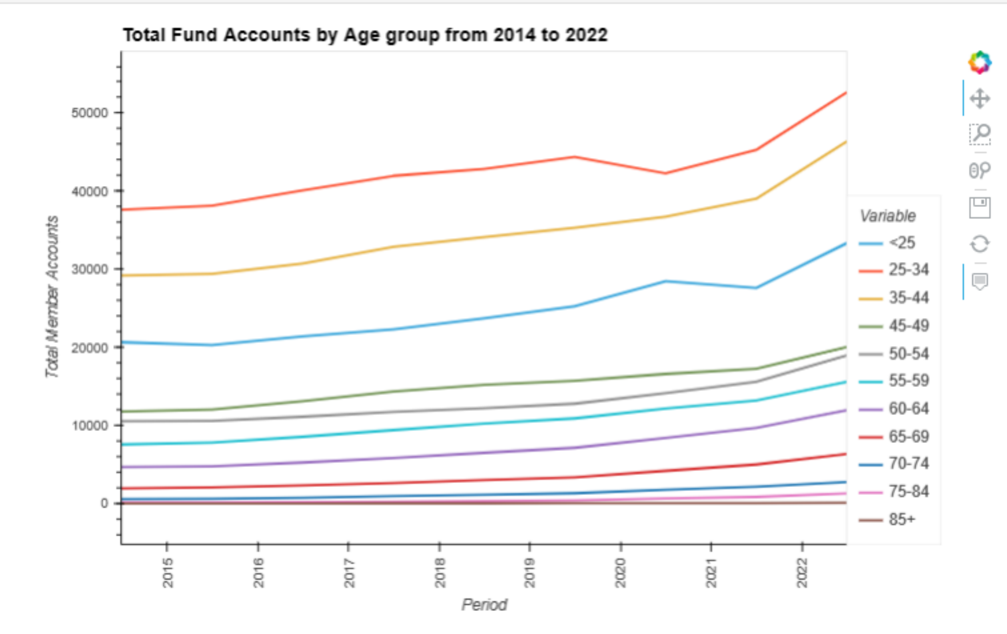


```
accounts_by_age_group = accounts_by_age_group.groupby(['Period']).agg('mean')
accounts_by_age_group
```

	<25	25-34	35-44	45-49	50-54
Period					
2014-06-30	20646.547826	37611.339130	29198.252174	11779.756522	10507.739130
2015-06-30	20297.086207	38133.827586	29414.887931	12028.844828	10568.500000

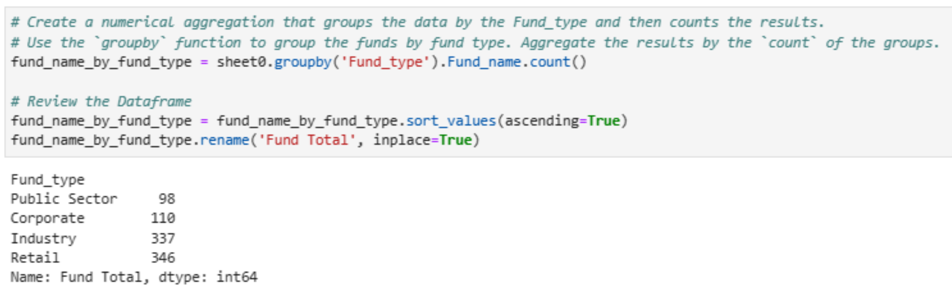


```
# Plot the viasualization to see the trend
accounts_by_age_group.hvplot(
    label="Total Fund Accounts by Age group from 2014 to 2022",
    xlabel="Period",
    ylabel="Total Member Accounts",
    rot=90,
    width=800,
    height=500,


).opts(yformatter='%.0f')
```



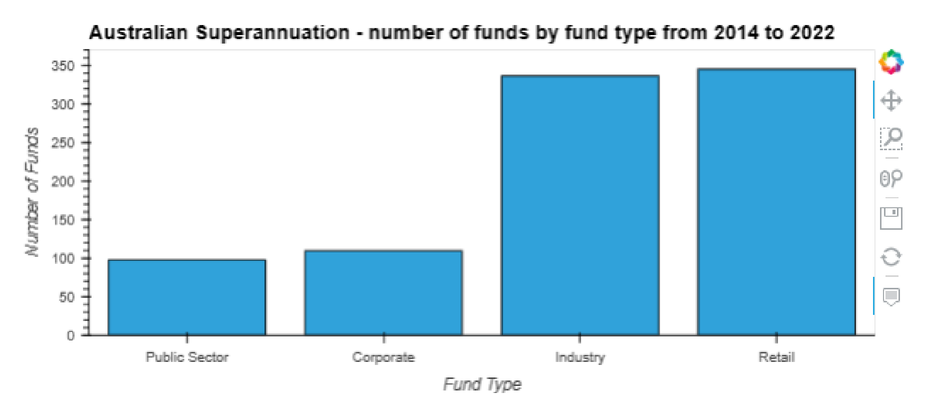
Data analysis continues…

DATA ANALYSIS



```
# Create a numerical aggregation that groups the data by the Fund_type and then counts the results.
# Use the `groupby` function to group the funds by fund type. Aggregate the results by the `count` of the groups.
fund_name_by_fund_type = sheet0.groupby('Fund_type').Fund_name.count()

# Review the Dataframe
fund_name_by_fund_type = fund_name_by_fund_type.sort_values(ascending=True)
fund_name_by_fund_type.rename('Fund Total', inplace=True)
```

```
Fund_type
Public Sector     98
Corporate        110
Industry         337
Retail           346
Name: Fund Total, dtype: int64
```

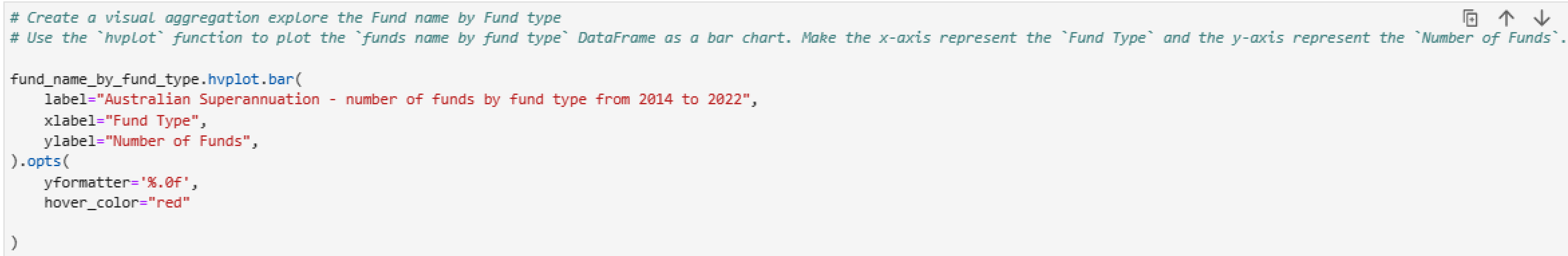


Q4 - Which sector has the maximum and minimum number of funds?

Answer - Retail sector has the maximum number of funds followed by Industry sector whereas, Public sector has the minimum number of funds followed by corporate sector.

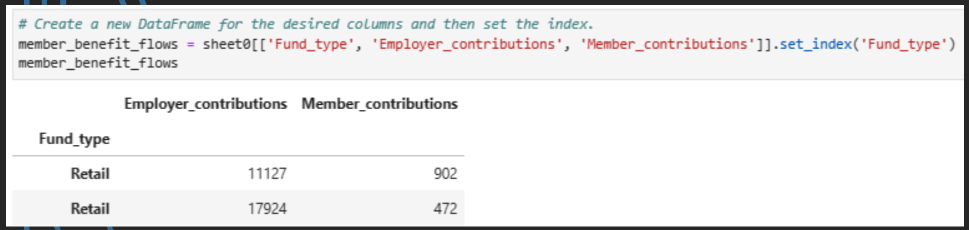
Steps taken to analyse the data:

- Use the `groupby` function to group the funds by fund type. Aggregate the results by the `count` of the groups.
- Use the `hvplot` function to plot the `funds name by fund type` DataFrame as a bar chart. Make the x-axis represent the `Fund Type` and the y-axis represent the `Number of Funds`.



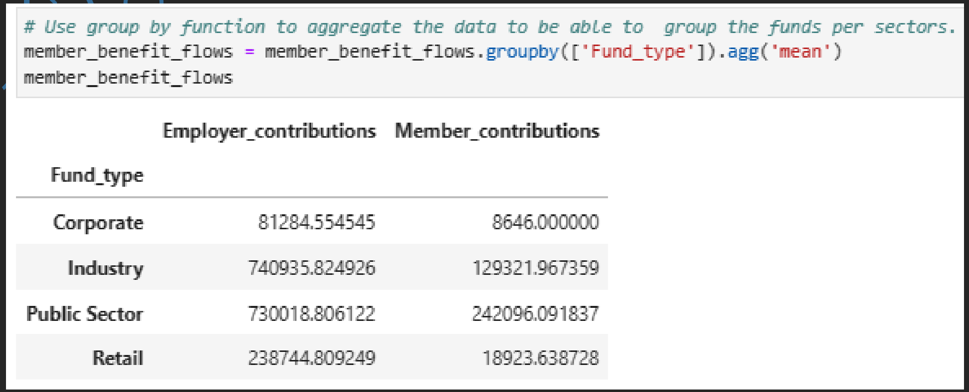
```
# Create a visual aggregation explore the Fund name by Fund type
# Use the `hvplot` function to plot the `funds name by fund type` DataFrame as a bar chart. Make the x-axis represent the `Fund Type` and the y-axis represent the `Number of Funds`.

fund_name_by_fund_type.hvplot.bar(
    label="Australian Superannuation - number of funds by fund type from 2014 to 2022",
    xlabel="Fund Type",
    ylabel="Number of Funds",
).opts(
    yformatter='%.0f',
    hover_color="red"

)
```

Data analysis continues…



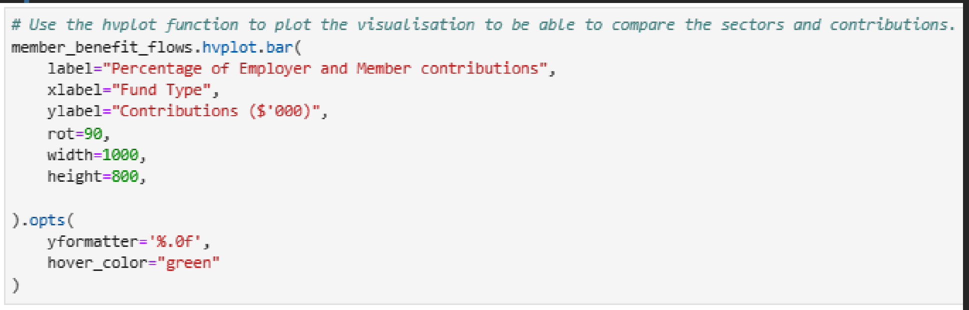
```
# Create a new DataFrame for the desired columns and then set the index.
member_benefit_flows = sheet0[['Fund_type', 'Employer_contributions', 'Member_contributions']].set_index('Fund_type')
member_benefit_flows
```

	Employer_contributions	Member_contributions
Fund_type		
Retail	11127	902
Retail	17924	472



```
# Use group by function to aggregate the data to be able to  group the funds per sectors.
member_benefit_flows = member_benefit_flows.groupby(['Fund_type']).agg('mean')
member_benefit_flows
```

	Employer_contributions	Member_contributions
Fund_type		
Corporate	81284.554545	8646.000000
Industry	740935.824926	129321.967359
Public Sector	730018.806122	242096.091837
Retail	238744.809249	18923.638728



```
# Use the hvplot function to plot the visualisation to be able to compare the sectors and contributions.
member_benefit_flows.hvplot.bar(
    label="Percentage of Employer and Member contributions",
    xlabel="Fund Type",
    ylabel="Contributions ($'000)",
    rot=90,
    width=1000,
    height=800,

).opts(
    yformatter='%.0f',
    hover_color="green"
)
```

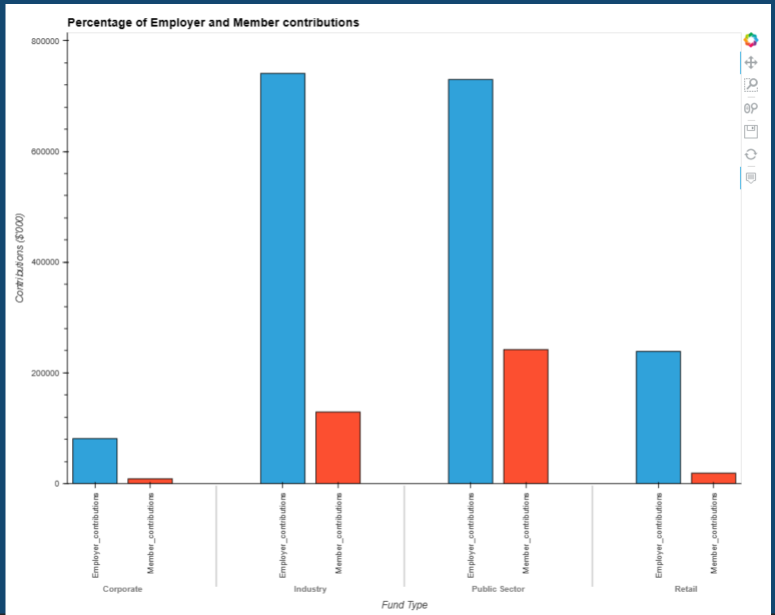
DATA ANALYSIS

Q5 - Which sector received the maximum contribution and from whom?

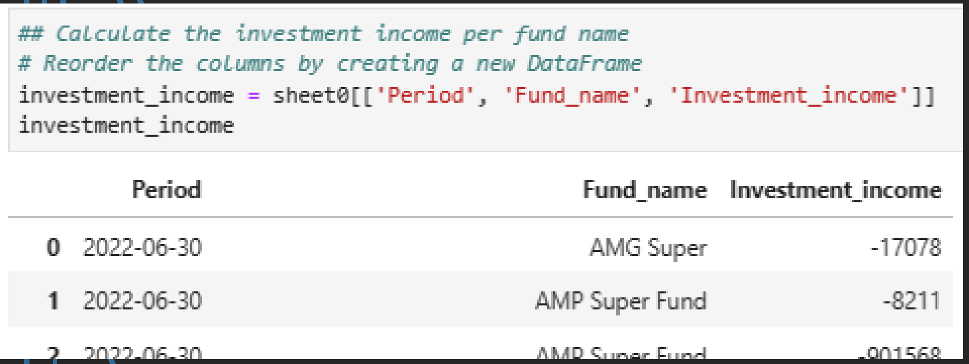
Answer - Industry sector received the maximum contribution from employer.

Steps taken to analyse the data:

- Create a new DataFrame for the desired columns and then set the index.
- Use group by function to aggregate the data to be able to group the funds per sectors.
- Use the hvplot function to plot the visualisation to be able to compare the sectors and contributions.

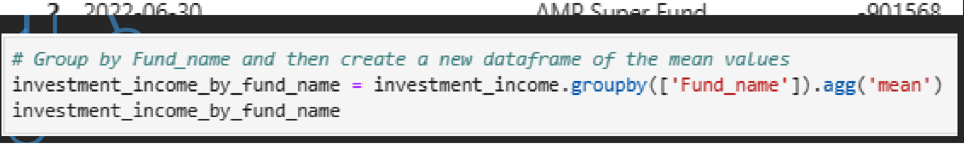


Data analysis continues…

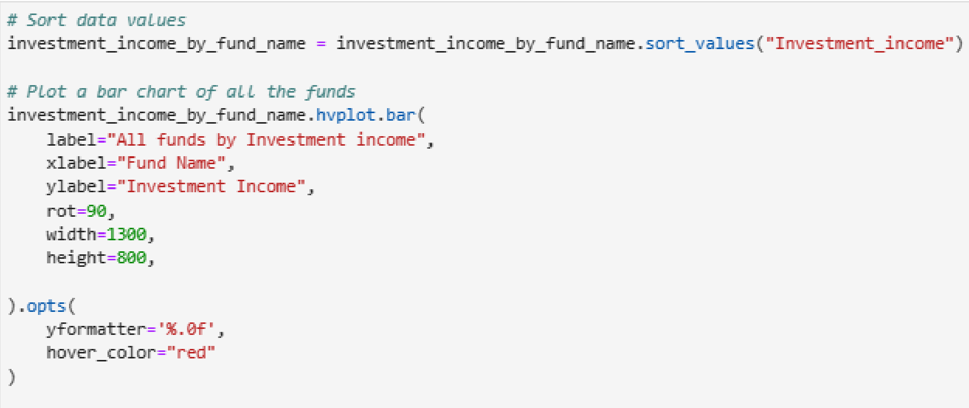


```
## Calculate the investment income per fund name
# Reorder the columns by creating a new DataFrame
investment_income = sheet0[['Period', 'Fund_name', 'Investment_income']]
investment_income
```

	Period	Fund_name	Investment_income
0	2022-06-30	AMG Super	-17078
1	2022-06-30	AMP Super Fund	-8211
2	2022-06-30	AMP Super Fund	-901568



```
# Group by Fund_name and then create a new dataframe of the mean values
investment_income_by_fund_name = investment_income.groupby(['Fund_name']).agg('mean')
investment_income_by_fund_name
```



```
# Sort data values
investment_income_by_fund_name = investment_income_by_fund_name.sort_values("Investment_income")

# Plot a bar chart of all the funds
investment_income_by_fund_name.hvplot.bar(
    label="All funds by Investment income",
    xlabel="Fund Name",
    ylabel="Investment Income",
    rot=90,
    width=1300,
    height=800,

).opts(
    yformatter='%.0f',
    hover_color="red"
)
```

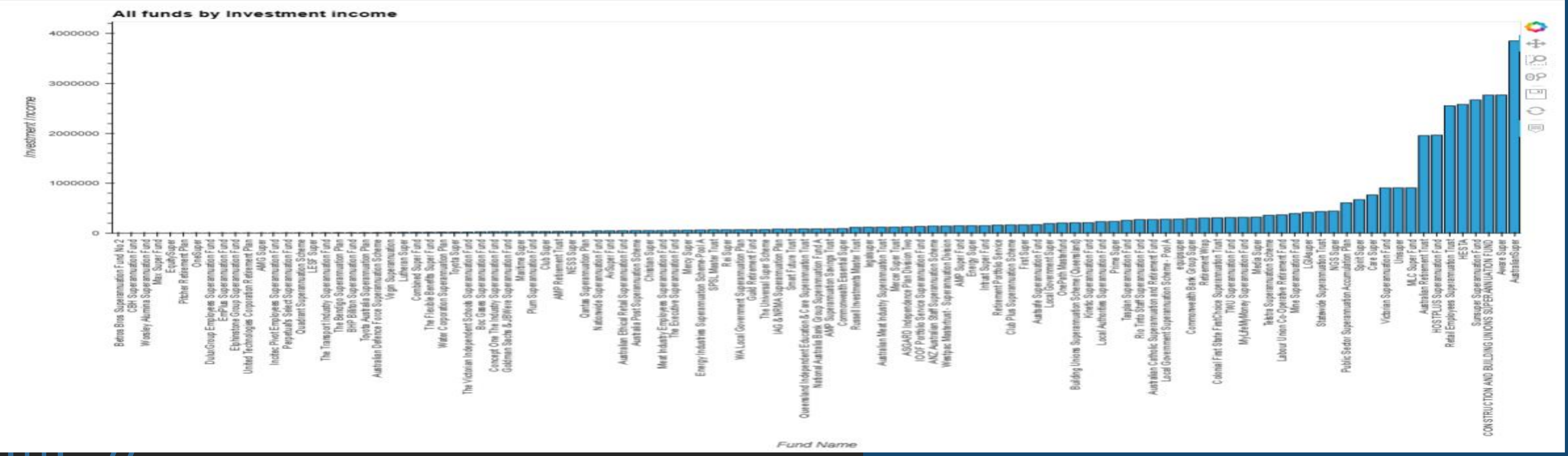
DATA ANALYSIS

Q6 - Which fund has the highest investment income?

Answer - Australian Super

Steps taken to analyse the data:

- Create a new DataFrame for the desired columns and then set the index.
- Reorder the columns by creating a new DataFrame
- Group by Period and Fund_name and then create a new dataframe of the mean values
- Use the hvplot function to plot the visualisation to be able to compare the sectors and contributions.



Data analysis continues…

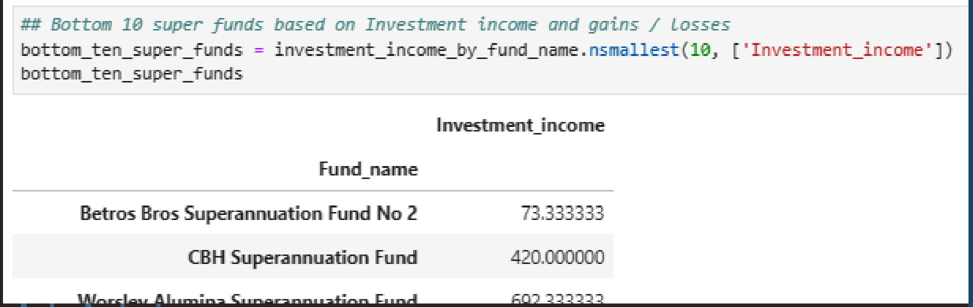
DATA ANALYSIS

Q7 - Which fund has the lowest investment income?

Answer - Betros Bros Superannuation om Fund No 2

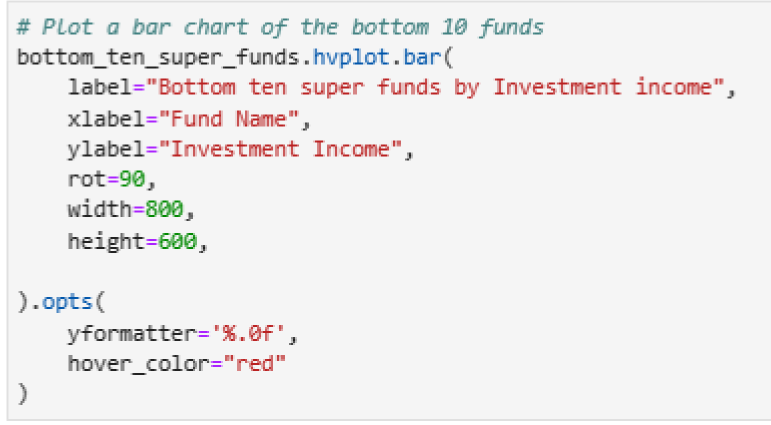
Steps taken to analyse the data:

- Determine the bottom ten funds by using nsmallest function.
- Use hvplot to visualize the outcome.

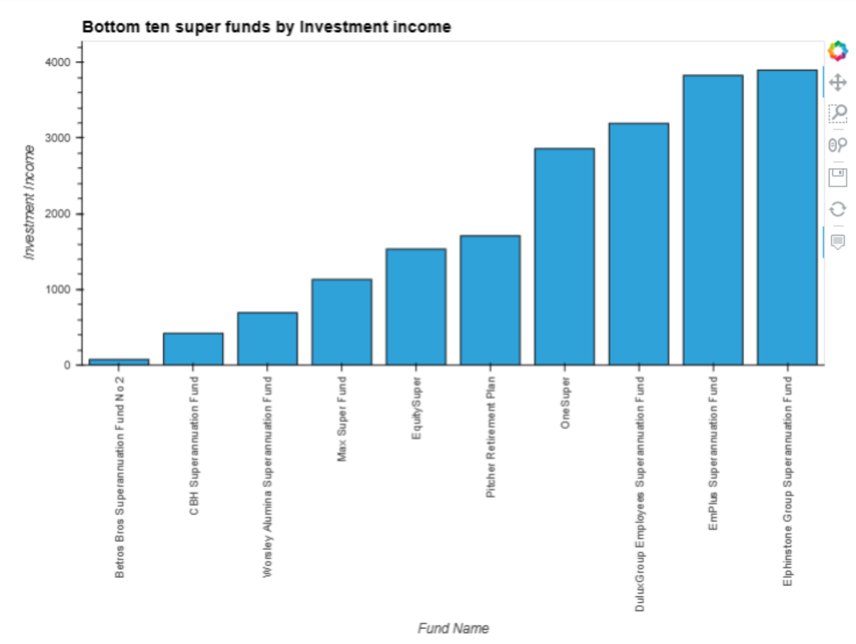


```
## Bottom 10 super funds based on Investment income and gains / losses
bottom_ten_super_funds = investment_income_by_fund_name.nsmallest(10, ['Investment_income'])
bottom_ten_super_funds
```

Fund_name	Investment_income
Betros Bros Superannuation Fund No 2	73.333333
CBH Superannuation Fund	420.000000
Worsley Alumina Superannuation Fund	692.333333



```
# Plot a bar chart of the bottom 10 funds
bottom_ten_super_funds.hvplot.bar(
    label="Bottom ten super funds by Investment income",
    xlabel="Fund Name",
    ylabel="Investment Income",
    rot=90,
    width=800,
    height=600,

).opts(
    yformatter='%.0f',
    hover_color="red"
)
```



Data analysis continues…

DISCUSSION

FINDINGS

- Plot indicates that there is not much difference between the male and female total accounts whereas the intersex total accounts numbers are almost negligible
- Age group 25-34 has the maximum number of superannuation accounts . This indicates that this age group has more working head counts.
- Retails sector holds the maximum number of funds whereas Public sector holds the least.
- Top ten funds by investment income there is a big gap between top performer and rest of the nine funds.

Data analysis continues…

POSTMORTEM

CHALLENGES & RESOLUTION

- Data with multiple sheets and header rows. There was a Coding challenges for multiple sheets and header rows. So, we normalised the data file by removing the unwanted header rows and finding the file path code for multiple sheets.

ADDITIONAL QUESTIONS AROSE BUT UNABLE TO ATTEMPT DUE TO TIME CONSTRAINT.

Compare the performance of investment income for all the funds but due to the complexity of the file structure and time constraint we were unbale to attempt.

WHAT WOULD YOU RESEARCH NEXT IF YOU HAD TWO MORE WEEKS?

Data analysis continues…

QUESTIONS?

Data analysis continues…