Introduction (data choice, data cleaning)

This study poses two questions:
- Which is the best pension fund in Israel?
- Are the most popular pension funds the best?

The study is based on data store grouping fund pension data from 1999 to 2022, and uses clustering algorithms to find the answer. Each row related a reporting period of a pension fund, giving us information about performance, cost and risks. The first step was to understand the shape of our data, how many different funds we have (300), and remove the lines where the value 'NaN' appeared. After cleaning we stayed with 154 different funds.

```
df = pd.read_csv( filepath_or_buffer: 'pensia-net1999-2022.csv', encoding='utf-8')
print(len(df))
print(df.columns)
# understand how many pension fund we have
print(df['FUND_ID'].nunique())
# understand how many Fund have only NaN value at the columns 'AVG_ANNUAL_YIELD_TRAILING_3YRS'
print(df.groupby('FUND_ID')['AVG_ANNUAL_YIELD_TRAILING_3YRS'].apply(lambda x: x.isna().all()).sum())
# delete the NaN raws
df = df.dropna(subset=['AVG_ANNUAL_YIELD_TRAILING_3YRS', 'AVG_DEPOSIT_FEE', 'SHARPE_RATIO', 'STOCK_MARKET_EXPOSURE'])
print(len(df))
# checking how many funds left
print(df['FUND_ID'].nunique())
```

Preparing the data for the algorithm

After data cleaning, one important step is to choose the right variables that give an indication about what a good pension fund is. The choice was to take the average yield trailing in 3 years, the average deposit fee, the sharpe ratio (the ratio between the gain and the risk), and the stock market exposure. Given each row represented a reporting period, the best choice was to group funds names and funds IDs by mean. This operation gives us unique rows, each row represents a pension fund.

GMM algorithm

GMM algorithm was launched with K=6, we received this :

cluster	AVG_ANNUAL_YIELD_TRAILING_3YRS	AVG_DEPOSIT_FEE	SHARPE_RATIO	STOCK_MARKET_EXPOSURE
0	6.204030	1.130004	0.770077	162.014092
1	5.077736	2.276417	1.104421	292.203970
2	4.539508	2.121896	2.134988	0.878050
3	3.011763	1.375935	0.591787	9.957400
4	6.540453	2.016127	1.267505	1433.607177
5	7.413284	2.021005	1.116220	14461.108854

The choice was to take cluster 2 as the best one. We can see that the average annual yield is not the highest but the risk is very low (stock market exposure low), and the sharpe ratio is the highest.
Within this cluster, the 5 'best' pension funds were the 5 with the highest sharpe ratio.

```
The 5 best pension fund with GMM algo are :
(שקלי) מגדל מקפת אישית כספי                                  39
(שקלי) הפניקס פנסיה מקיפה -כספי                              64
...קי מיטב דש פנסיה מקיפה מסלול בסיסי למקבלי קצבה            66
(שקלי) הראל פנסיה - כספי                                     99
מנורה מבטחים פנסיה - מסלול קצבה לזכאים קיימים                88
```

BGMM algorithm

We ran the BGMM algorithm on the same data store we ran the GMM algorithm on.
The BGMM algorithm proposed us 5 clusters :

cluster_bgmm				
0	4.888707	1.701216	0.859549	175.057415
1	7.233770	1.971371	1.053714	2020.963794
2	4.841886	1.968409	2.507621	89.958870
3	7.147561	2.046717	1.137223	15888.596505
7	0.216193	1.144151	-0.378111	0.000000

with these weights:

```
Weights of clusters :
Cluster 0: 0.5980
Cluster 1: 0.1182
Cluster 2: 0.2113
Cluster 3: 0.0589
Cluster 4: 0.0041
Cluster 5: 0.0029
Cluster 6: 0.0021
Cluster 7: 0.0045
```

Here we choose a new time the cluster 2, and we found that the 5 ‘best’ pension found inside cluster 2 are :

```
the best 5 pension fund with BGMM algo are :
5.285385      (שקלי) מגדל מקפת אישית כספי                      39
5.279747      (שקלי) הפניקס פנסיה מקיפה -כספי                  64
4.680000      ...קי מיטב דש פנסיה מקיפה מסלול בסיסי למקבלי קצבה  66
3.818621      (שקלי) הראל פנסיה - כספי                         99
3.676500      מנורה מבטחים פנסיה - מסלול קצבה לזכאים קיימים    88
```

We can see that the result is the same for both of the algorithms. The same results can confirm the reliability of our results.

Silhouette score

The silhouette score is a metric evaluating how well-grouped data points are in clustering. The silhouette score was calculated on the 2 algorithms. Here the results :

```
Silhouette GMM: 0.11101180726864737
Silhouette BGMM: 0.10729976330468119
```

These results show the proximity between the clusters. There are no ‘very bad’ funds or ‘perfect’ funds and one of the reasons is that every Fund is regulated by Israeli Law. For example if the sharpe ratio is very low, the regulators will verify the risk the fund takes, and there is a possibility of closure if the fund is not profitable at all.

Wisdom of crowds

In the data store, there are no columns ‘NUMBER OF CLIENTS’, but in order to estimate the popularity of a fund, the best solution was to use the 'DEPOSITS' variable.
Each fund was grouped by Fund name and Fund ID to have unique values, and the deposit value of each fund was the mean of each deposit fund. In this way, funds with more data have no advantage over funds with few data.

The result was :

FUND_ID	FUND_NAME	DEPOSITS
566.685316	מנורה מבטחים פנסיה - כללי	2009	15
403.191139	מגדל מקפת אישית כללי	2102	32
363.997089	הראל פנסיה - גילעד כללי	2172	53
363.394800	אלטשולר שחם פנסיה מקיפה מסלול לבני 50 ומטה	9757	116
352.140800	הפניקס פנסיה מקיפה - מסלול לבני 50 ומטה	9974	138

Even though the same companies like Menora, Harel, and Phoenix appear in this ranking, we can see that the most popular funds are general funds.

These funds are mass-market default products, often presented as safe and simple pension funds.

The 'best' funds in our algorithms were chosen for their risk/return ratio, which is a very specific metric and can define an excellent pension fund.

The most popular pension funds are not necessarily the 'best'.