US Exchange Traded Funds and Mutual Funds

Executive summary

	summary
1. Objective	Analyze and extract insights from US funds data
2. Dataset	Static and time series data on ETFs and Mutual Funds with several attributes
3. Method	• EDA on Python • Visualization on PowerBI

4. Conclusion	• Analysis performed and provision made to extract further insights
	• Further course is proposed in terms of
	  ○ Ancilary data analysis that can be performed
	  ○ Modeling approaches that can be used to perform predictions

Detailed analysis follows next page onwards.

Objective

- I study and analyze a public dataset on US Exchange Traded Funds and Mutual Funds.
- I set up personal objective to answer the following questions through this exercise:
    - How has the **popularity of funds progressed over the years** – overall and certain cohort based?
    - How has the **financial performance of funds progressed over the years** – overall and certain cohort based?
    - How are **different fund performance indicators correlated**, if at all?
    - What **ETF and MF category should I invest in, if I favor a particular performance metric** over others?

Dataset

- 4 datasets available on US Funds dataset from Yahoo Finance (kaggle.com) were collectively analyzed:

datasets	attributes	rows	description
ETFs	142	**2K** *(2310)*	- attributes are **static** - majorly include (amongst others) quantified: ○ general fund aspects (e.g. total_net_assets, fund family, inception date, etc.) ○ portfolio indicators (e.g. cash, stocks, bonds, sectors, etc.) ○ Financial ratios (price/earning, Treynor and Sharpe ratios, alpha, and beta) ○ ESG scores for MFs ○ historical yearly and quarterly returns (e.g. year_to_date, 1-year, 3-years, etc.)
MutualFunds	298	**24K** *(23,783)*	
ETF prices	8	**3.8M** *(3,866,030)*	- attributes are **time series** data ○ ETFs ▪ fund, price date, open-high-low-close prices, volume traded ○ Mutual Funds ▪ fund, price date, nav per share
MutualFund prices	3	**75.6M** *(75,657,739)*	

Method

1) **Exploratory Data Analysis on Python**
   a) **Organizing data:**
      i) CSV files ---> first stored in Databricks Catalog under personal volume ---> imported as Pandas dataframes into notebooks
   b) **Data exploration, cleaning and analysis:**
      i) Unique entries of fund symbols are cross-checked against lengths of the dataframe
      ii) **DATA TYPES** are identified.
         (1) The motivation is to identify categorical attributes that can later be used for cohort-based- visualizations
      iii) Data type for relevant attributes is transformed to 'date-time'
      iv) Noise is removed e.g. undesired class ('Other OTC') within category 'exchange_name' is dropped as we wish to analyze only for major exchanges. Moreover, there were only 9 rows associated with the class.
      v) **MISSING VALUES** are explored; columns with lowest and highest missing values are identified
      vi) Missing values are exclusively identified for certain columns of interest (categorical and performance-based columns) that are to be analyzed later in the study
         (1) Deleting all rows with missing values in at least one of these columns leads to significant data loss as big as 70%
         (2) Instead, I choose to delete rows with null values in 'all' the columns together. This reduces data loss.
         (3) I also delete rows with missing values in categorical columns.
         (4) Thoughts on Data amputation:
            (a) tricky because the fund performances are characteristics of so many factors.
            (b) For modelling purposes, 5year performance metrics can proxy for 10 year performance metrics (and vice versa), if later we see correlation in these.
      vii) Few **CATEGORICAL COLUMNS** are handpicked from columns with data type as 'objects'. Number of classes within each of these are identified. I limit further analysis to the cohorts based on these categorical columns.
         (1) These are ['exchange_name', 'investment_type', 'size_type', 'fund_category', 'fund_family']
         (2) Out of these, 'fund_category' and 'fund_family' have much more classes (84 and 150) than I expected, so later after a point, I further limit **COHORT BASED ANALYSIS** to only 'exchange_name', 'investment_type' and 'size_type' attributes.
      viii) Split of 'Count of funds' in respective categories is visualized using vertical bar plots
      ix) Split of 'Count of funds' in 'fund_category' and 'fund_family' is visualized for top 15 classes only (horizontal plot)

x) Out of the many attributes present in the dataframe, following are handpicked to analyze and compare performance of funds:
    (1) 'total_net_assets', 'fund_mean_annual_return_5years', 'fund_sharpe_ratio_5years', 'fund_treynor_ratio_5years', 'fund_price_earning_ratio'

xi) I visualize following performances (on above attributes) of cohorts of funds based on 'exchange_name', 'investment_type', 'size_type':
    (1) Sum of total net assets (*I could have also presented another metric here - 'average net asset value per fund'*)
    (2) Average of 5 year fund annual return
    (3) Average of 5 year Sharpe ratio
    (4) Average of 5 year Treynor ratio
    (5) Average of Price/Earning ratio
        (a) Small and Value ETFS are best performers

xii) I add 4 more attributes to performance indicators - 10 year counterparts of the above list.

xiii) **CORRELATIONS**. Then I normalize these performance specific columns (using standard scaler) to study Correlations between various fund performance indicators on two levels.
    (1) These levels are
        (a) Overall funds; and
        (b) Funds split on categorical cohorts
    (2) In general, Growth Funds seem to have higher correlation between 5year and 10year performance counterparts as compared to Value and Blend funds.
        (a) One possible explanation can be the focus of Growth ETFs on technology and other high-growth sectors, which have performed well over the past decade, leading to stronger correlations between 5-year and 10-year performance.
        (b) General trend of this correlation is Growth > Blend > Value

xiv) Similar analysis can be performed on other cohorts.

xv) I **TRIM** price dataset to start from 2001 after ensuring that the amount of data that dates back before 2001 is fairly less.

xvi) Necessary attributes from ETF static dataset are merged with ETF prices dataset that has time-series data

xvii) Then I study YoY growth (popularity) of funds in different cohorts
    (1) Most popular and growing categories - 'Large' sized funds with 'blend' investment type linked with 'NYSE'

xviii) Finally, I also visualize YoY overall trading volume in different cohorts using a heatmap.
    (1) ETF trading linked with 'NYSE' peaked around 2008-09, with 'Large' sized 'Growth' and 'Blend' ETF contributing almost entirely towards this peak.

(a) *One possible explanation is that during the global financial crisis, investors may have seen ETFs as a more attractive option to diversify their portfolios and reduce risk.*

(b) *Large sized Growth and Blend ETFs may have been particularly popular during this time as investors may have been looking for exposure to companies with strong growth potential.*

(c) *This increased demand for ETFs may have led to higher trading volumes on NYSE in general.*

(2) MF trading displays a different trend. Mutual Funds have been peaking recently (continuous growth), with 'Large' sized 'Growth' MFs being the hot ones.

(a) *Investors seem to be turning to mutual funds for their ability to potentially generate higher returns through active management and stock picking.*

2) <u>Visualization on PowerBI</u>

a) 4 levels of dynamic visualizations have been presented where ETFs have been compared with Mutual Funds.

b) These are hierarchical in nature i.e. as we progress from level 1 to 4, we can go further granular in analysis.

c) In each level, visualizations can be filtered on custom date ranges.

d) Provisions have also been added to filter visualizations on cohorts present within the following categories (more similar filters can be added):

i) Name of the exchange

ii) Type of investment

iii) Size of the fund

e) Description of these levels:

i) **Funds Outlook**: Here I present an overall outlook of US funds. Following dynamic visualizations are included:

(1) YoY growth of funds

(2) YoY average price/earnings trend

(3) Frequency distribution of funds price/earning

(4) Average 5year fund performance metrics compared against 10year counterparts

(5) YoY average sector-wise composition of ETF portfolios

(6) YoY ESG score trends for Mutual funds

(7) YoY average NAV per share trend for Mutual funds

(8) YoY class wise (stocks, bonds, etc) composition of Mutual funds

(9) Top performing funds based on P/E:

ii) **Categorical Splits**: This is very close to the initial EDA performed over Python. Based on name of exchange, type of investment and size of fund, visualizations include cohort-wise splits of the following:

(1) Count of funds in market

(2) Total net assets

(3) 5year mean annual return

(4) 5year Sharpe and Treynor ratio

iii) **Temporal Trends**: This includes time-wise progress of fund performance indicators. Visualizations include temporal trends of all things discussed under level **'Funds Outlook'**

iv) **Single Fund Level**: This is for visualizing progress of single funds exclusively.

Conclusion

- Summary
  - Following questions (among many others) can now be answered:
    1. How has the popularity of funds progressed over the years – overall and certain cohort based?
    2. How has the financial performance of funds progressed over the years – overall and certain cohort based?
    3. How are different fund performance indicators correlated, if at all?
    4. What ETF and MF category should I invest in, if I favor a particular performance metric over others?
- Key future course (**of interest to me**; amongst **many** other possible) include:
  - Studying correlations between performance indicators for other cohorts. Also, several other visualizations are possible
  - Since this is a time series dataset, we can try and study the stationarity in the data
  - **Modeling** - Scope of modeling future predictions using ARIMA, LSTMs and Transformers