Group Number: 2
Group Member Name: Lau Chun Kui (Kelvin), Leung Ka Shek (Eric),
Lui Ching Fung (Brian), Kwok Hei Wut (Charlotte), Siu Wai Na (Aura)

Project title: Smart Investor Profile Explorer (Client Collaboration with Zenith – Capstone Project)

Objectives

The primary objective of this project is to develop a smart, data-driven investor profile exploration tool utilising real operational data provided by Zenith. This tool is designed to help Zenith better understand the characteristics and preferences of different types of investors. Specifically, the project aims to identify patterns and trends in investor behaviour, generate actionable insights for investment decision-making, and uncover potential business opportunities. In addition, we aim to build predictive models that can enhance the accuracy of investor-matching, enabling Zenith to improve the efficiency of its recommendation and matching systems.

Dataset Description

1. Introduction to the Dataset

For this project, Zenith provided us with a dataset containing real-world operational investor data. This dataset includes various types of information such as investor types, geographic locations, and industry tags. The data was collected from internal systems and external sources and reflects actual investor engagement and behaviour.

2. Dataset Characteristics

The original dataset was very large, with more than 2,000,000 records. After resolving coding issues in the raw file, we extracted around 155,000 valid records for further processing. Following a thorough data cleaning and preparation process, we finalised a clean dataset containing approximately 136,277 rows and 9 structured columns.

The dataset includes key fields such as investor types, geographic tags, and industry categories. In total, there are 28 distinct investor types, 204 unique geographic categories, and 16 industry categories. The dataset is highly multi-dimensional and somewhat imbalanced, particularly in the distribution of geographic and industry labels.

3. Data Preprocessing Steps

Before conducting any analysis or modelling, we performed a series of preprocessing steps to improve the quality and usability of the dataset. We began by removing duplicate and invalid records. Then, we handled missing or malformed values, including empty entries, default system values such as “[0],” or incomplete URLs such as “www.”

After that, we standardised the textual fields to ensure consistency across all data entries. This was especially important for the geographic and industry fields, which contained a wide variety of formats and naming conventions. We converted these free-text entries into structured categories that could be used effectively in our machine learning models and visualisations.

At the end of this process, we obtained a clean and structured dataset consisting of 136,277 records and 9 well-defined columns. This dataset provided a strong foundation for the subsequent analysis and modelling stages.

Remove Duplication

1. Remove Duplication

Step 1: remove duplicate iteration
(third line of code) The variable "files" inside pd.read_parquet() of the original file is revised to "file" .

```
num_files = 15
files = [f"files/batch_{i}.parquet" for i in range(num_files)]

dfs = [pd.read_parquet(file) for file in files]
combined_df = pd.concat(dfs, ignore_index=True)
combined_df
```

	Investors	Primary Contact	Description	Geography	Preferred Industry	Preferred Investment Type
0	Techstars	David Cohen	Founded in 26, Techstars is an accelerator bas...	Africa, Americas, Asia, Canada, Middle East, O...	Beverages, Computer Hardware, Education and Tr...	Accelerator/Incubator, Early Stage VC, Later S...

Step 2: checking all other possible duplications (rows with exactly same row values --> found to be 590 number of rows).

```python
# If values in a column conatains unhashable data e.g. list / ndarray, formula will return error --> last three columns
# Thus, only use columns with hashable types --> subset of other columns
# This formula will return a series of boolean values indicating whether each row is a duplicate, True = Duplicate

check_duplicates = combined_df.duplicated(subset=['Investors', 'Primary Contact', 'Description', 'Geography', 'Preferred Industry', 'Preferred Investment Type', 'Primary Investor T
print( check_duplicates )

print( check_duplicates.sum() )          # Count of duplicates --> 590
```

Python

```
0         False
1         False
2         False
3         False
4         False
          ...
155339    False
155340    False
155341    False
155342    False
155343    False
```

Handle Missing Value (e.g. 'www' , '[0])

```python
combined_df.drop(columns=['Primary Contact'], inplace=True, errors='ignore')
```

```python
print(combined_df)
```

```
                       Investors  \
0                      Techstars
1                  Y Combinator
2       Plug and Play Tech Center
3                       Gaingels
4                         Antler
...                          ...
155339                ECS Tuning
155340                  ECSEL JU
155341                    Ecster
155342                ECU Health
155343             ECU Worldwide

                                            Description  \
0       Founded in 26, Techstars is an accelerator bas...
1       Founded in 25, Y Combinator is an accelerator ...
```

```
155339                ECS Tuning
155340                  ECSEL JU
155341                    Ecster
155342                ECU Health
155343             ECU Worldwide

                                            Description  \
0       Founded in 26, Techstars is an accelerator bas...
1       Founded in 25, Y Combinator is an accelerator ...
2       Founded in 26, Plug and Play Tech Center is an...
3       Founded in 214, Gaingels is a venture capital ...
4       Founded in 217, Antler is a venture capital in...
...                                                   ...
155339  Manufacturer and distributor of automotive par...
155340
155341  Operator of payment solutions for both busines...
155342
155343  ECU Worldwide is a provider of logistic servic...
...
155342                                              [0]
155343                                        [logistics]

[139710 rows x 9 columns]
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Remove Duplication and Invalid Records

```
    duplicates_index_number = check_duplicates[check_duplicates == True]
    duplicates_index_number.index
```

```
Index([ 11372,  26404,  26415,  26517,  32353,  32422,  32967,  33122,  33561,
        34570,
       ...
       153768, 153769, 153774, 153861, 153899, 154109, 154162, 154866, 154867,
       154873],
      dtype='int64', length=590)
```

```
    combined_df.drop(index = duplicates_index_number.index, inplace=True)  # Drop duplicates
    combined_df
```

	Investors	Primary Contact	Description	Geography	Preferred Industry	Preferred Investment Type	
0	Techstars	David Cohen	Founded in 26, Techstars is an accelerator bas...	Africa, Americas, Asia, Canada, Middle East, O...	Beverages, Computer Hardware, Education and Tr...	Accelerator/Incubator, Early Stage VC, Later S...	Acc

154754 rows × 10 columns

Step 3: Check duplication for same investor inputs

```
    check_duplicates_2 = combined_df.duplicated(subset=['Investors', 'Description'])
    print(check_duplicates_2)
    print( check_duplicates_2.sum() )                # Count of duplicates --> 15044
```

```
0         False
1         False
2         False
3         False
4         False
          ...
155339    False
155340    False
155341    False
155342    False
155343    False
Length: 154754, dtype: bool
15044
```

Handle Missing Values (Remove Shifted Data)

3. Remove shifted data.

Step 1: Check shift by checking "www." in the column "Description" and remove

```
df_removed_shifted = drop_df[['Description']][drop_df['Description'].str.contains('www.')]
df_removed_shifted
```

	Description
94	www.aceandcompany.com
118	www.hf.com
251	www.nhqv.com
338	www.tdpfund.com
406	www.bdtmsd.com
...	...
155013	www.bebig.com
155014	www.ezag.com
155016	www.eckertseamans.com
155059	www.ecooilandgas.com
155230	www.ecoplastindia.com

3152 rows × 1 columns

```
index_list_removed_shifted = df_removed_shifted.index.tolist()
index_list_removed_shifted
```

```
[94,
 118,
 251,
 338,
```

Drop shifted rows and store in a new dataframe "drop_df_removed_shift"

```
drop_df_removed_shift = drop_df.drop(index= index_list_removed_shifted )
drop_df_removed_shift
```

Python

	Investors	Description	Geography	Preferred Industry	Preferred Investment Type	Primary Investor Type	geography_tags	preferred_investment_type_tags	preferred_industry_tags
0	Techstars	Founded in 26, Techstars is an accelerator bas...	Africa, Americas, Asia, Canada, Middle East, O...	Beverages, Computer Hardware, Education and Tr...	Accelerator/Incubator, Early Stage VC, Later S...	Accelerator/Incubator	[africa, americas, asia, canada, middle east, ...	[accelerator/incubator, early stage vc, later ...	[beverages, computer hardware, education and t...
1	Y Combinator	Founded in 25, Y Combinator is an accelerator ...	Africa, Americas, Asia, Europe, Oceania, Unite...	Biotechnology, Commercial Transportation, Comm...	Accelerator/Incubator, Early Stage VC, Later S...	Accelerator/Incubator	[africa, americas, asia, europe, oceania, unit...	[accelerator/incubator, early stage vc, later ...	[biotechnology, commercial transportation, com...
2	Plug and Play Tech Center	Founded in 26, Plug and Play Tech Center is an...	Africa, Americas, Asia, Canada, Europe, Middle...	Aerospace and Defense, Animal Husbandry, Aquac...	Accelerator/Incubator, Early Stage VC, Later S...	Accelerator/Incubator	[africa, americas, asia, canada, europe, middl...	[accelerator/incubator, early stage vc, later ...	[aerospace and defense, animal husbandry, aqua...
3	Gaingels	Founded in 214, Gaingels is a venture capital ...	Africa, Americas, Asia, Canada, Europe, Middle...	Business Products and Services (B2B), Consumer...	Early Stage VC, Later Stage VC, PE Growth/Expa...	Venture Capital	[africa, americas, asia, canada, europe, middl...	[early stage vc, later stage vc, pe growth/exp...	[business products and services (b2b), consume...
4	Antler	Founded in 217, Antler is a venture capital in...	Australia, Brazil, Canada, China, Denmark, Est...	Agriculture, Business Products and Services (B...	Accelerator/Incubator, Early Stage VC, Seed Round	Venture Capital	[australia, brazil, canada, china, denmark, es...	[accelerator/incubator, early stage vc, seed r...	[agriculture, business products and services (...
...	...	...	...	...	...	...	...	...	...
155339	ECS Tuning	Manufacturer and distributor of automotive par...		Commercial Products, Transportation	Add-on, Buyout/LBO, Merger/Acquisition	0	[]	[add-on, buyout/lbo, merger/acquisition]	[commercial products, transportation]
155340	ECSEL JU			Semiconductors, Software	0	0	[]	[0]	[semiconductors, software]
		Operator of							

Handle Missing Values (Remove '.com' and '.org'

Step 2: Check shift by checking ending with ".com" in the column "Description" , and remove

```
df_removed_shifted_2 = drop_df_removed_shift[drop_df_removed_shift['Description'].str.endswith('.com', na=False)]
df_removed_shifted_2
```

Python

	Investors	Description	Geography	Preferred Industry	Preferred Investment Type	Primary Investor Type	geography_tags	preferred_investment_type_tags	preferred_industry_tags
545	Anderson Lock &amp	andersonlockandsafe.com	Anderson Lock & Safe is a provider of Lock...	0	0	Merger/Acquisition	[anderson lock & safe is a provider of loc...	[0]	[0]
1133	Italian Food &amp	ifbcorp.com	Operator of an investment holding company. The...	0	Consumer Non-Durables, Retail, Software	Merger/Acquisition	[operator of an investment holding company. th...	[consumer non-durables, retail, software]	[0]
1852	J&amp	jnmpartners.com	Founded 221, J&M Partners is a private equ...	0	0	Buyout/LBO, Early Stage VC, Later Stage VC, PE...	[founded 221, j&m partners is a private eq...	[0]	[0]
1854	J&amp	jnpef.com	Founded in 218, J& Private Equity is a pri...	0	0	Buyout/LBO, Later Stage VC, PE Growth/Expansion	[founded in 218, j& private equity is a pr...	[0]	[0]
1856	J&amp	jrcollisioncenters.com		0	Transportation	Merger/Acquisition	[]	[transportation]	[0]
...	...	...	...	...	...	...	...	...	...
152015	Dr. Lutz &amp	lutz-cie.com	Founded in 211, Dr. Lutz & Cie is a ventur...	0	0	Early Stage VC, PE Growth/Expansion, Seed Round	[founded in 211, dr. lutz & cie is a ventu...	[0]	[0]
152125	Drake Real Estate &amp	drake-investments.com	Founded in 212, Drake Real Estate & Invest...	0	0	0	[founded in 212, drake real estate & inves...	[0]	[0]
153734	E &amp	eandiventuresllc.com	Founded in 216, E & I Ventures is a ventur...	0	Software	Debt - General, Early Stage VC, Seed Round	[founded in 216, e & i ventures is a ventu...	[software]	[0]
153881	E.B. Horsman &amp	ebhorsman.com		0	Commercial Products, Computer	Merger/Acquisition	[]	[commercial products, computer hardware]	[0]

Step 3: Check shift by checking ending with ".org" in the column "Description" , and remove

```
df_removed_shifted_3 = drop_df_removed_shift[drop_df_removed_shift['Description'].str.endswith('.org', na=False)]
df_removed_shifted_3
```

Python

	Investors	Description	Geography	Preferred Industry	Preferred Investment Type	Primary Investor Type	geography_tags	preferred_investment_type_tags	preferred_industry_tags
8460	JSSATE - Science &amp	jssstepnoida.org	JSSATE - Science & Technology Entrepreneur...	0	Software	Accelerator/Incubator	[jssate - science & technology entrepreneu...	[software]	[0]
22647	Eli &amp	broadfoundation.org	Eli & Edythe Broad Foundation is an indepe...	0	Services (Non-Financial)	0	[eli & edythe broad foundation is an indep...	[services (non-financial)]	[0]
60934	Hope &amp	makefoodyourbusiness.org	Founded in 214, Hope & Main is an incubato...	0	Consumer Non-Durables	Accelerator/Incubator, Early Stage VC, Seed Round	[founded in 214, hope & main is an incubat...	[consumer non-durables]	[0]
68343	Innovation &amp	accelerator.childrenshospital.org	Innovation & Digital Health Accelerator is...	0	Healthcare Technology Systems, Other Healthcare	Accelerator/Incubator	[innovation & digital health accelerator i...	[healthcare technology systems, other healthcare]	[0]
71960	Waggies By Maggie And Friends	waggies.org	providing meaningful employment to persons wit...	None	legal services	None	[providing meaningful employment to persons wi...	[legal services]	[]
117324	The West Coast Consortium for Technology &amp	westcoastctip.org	Founded in 211, The West Coast Consortium for ...	United States	Healthcare Devices and Supplies, Healthcare Te...	Accelerator/Incubator	[founded in 211, the west coast consortium for...	[healthcare devices and supplies, healthcare t...	[united states]
118319	Boys &amp	bgcmla.org	Boys and Girls Clubs of Metro Los Angeles is a...	0	0	Merger/Acquisition	[boys and girls clubs of metro los angeles is ...	[0]	[0]

Standardize textual tags into structured fields - Work on One Hot Encoding: Geography Tags

4. Work on One Hot Encoding

For each tag column, do the below steps individually to avoid duplicates during explode in step 1
Step 1: extract the unique items in each tag columns, review and categorize in excel
Step 2: import back the excel file to dataframe, do lookup to the existing dataset dataframe, do one hot encoding after that .
Step 3: combine the resulting encoding columns to the existing dataset dataframe.

- Extract Tag Column - Geography Tags

```
geography_tags_list = drop_df_removed_shift['geography_tags'].explode().dropna().unique().tolist()
print("Geography Tags:", geography_tags_list)

df_to_excel = pd.DataFrame(geography_tags_list)
df_to_excel.columns = ['Location']
df_to_excel
```

```
Geography Tags: ['africa', 'americas', 'asia', 'canada', 'middle east', 'oceania', 'united kingdom', 'united states', '
```

	Location
0	africa
1	americas
2	asia
3	canada
4	middle east
...	...
3691	indigenous communities transportation infrastr...
3692	etc
3693	providing services and support to clients to m...
3694	operator of a holding company based in niort
3695	france. the company is involved in fund manage...

3696 rows × 1 columns

```
%pip install pycountry

import pandas as pd
import zipfile
import io
import requests
import pycountry
```

```
Requirement already satisfied: pycountry in c:\users\leung\anaconda3\lib\site-packages (24.6.1)
Note: you may need to restart the kernel to use updated packages.
```

```
# Step 1: 下載並讀取 GeoNames 的城市資料
url = "http://download.geonames.org/export/dump/cities15000.zip"
response = requests.get(url)

if response.status_code == 200:
    with zipfile.ZipFile(io.BytesIO(response.content)) as z:
        with z.open('cities15000.txt') as f:
            columns = [
                'geonameid', 'name', 'asciiname', 'alternatenames', 'latitude', 'longitude',
                'feature_class', 'feature_code', 'country_code', 'cc2', 'admin1_code',
                'admin2_code', 'admin3_code', 'admin4_code', 'population', 'elevation',
                'dem', 'timezone', 'modification_date'
            ]
            df_geo = pd.read_csv(f, sep='\t', header=None, names=columns)
else:
    raise Exception("GeoNames 資料下載失敗，請檢查網路。")

# Step 2: 建立 地名 → 國家名稱 的對照字典
def get_country_name(code):
    try:
        return pycountry.countries.get(alpha_2=code).name
    except:
        return None

df_geo['country_name'] = df_geo['country_code'].apply(get_country_name)
```

Extract Tag Column - Geography Tags

```
                                      Location Standardized_Country
0                                       africa                 N/A
1                                     americas                 N/A
2                                         asia           Philippines
3                                       canada                Canada
4                                  middle east                 N/A
...                                        ...                 ...
3691  indigenous communities transportation infrastr...          N/A
3692                                        etc                N/A
3693  providing services and support to clients to m...          N/A
3694       operator of a holding company based in niort           N/A
3695  france. the company is involved in fund manage...          N/A

[3696 rows x 2 columns]
```

```
geography_categories = pd.read_excel('Tag Columns Classification.xlsx', sheet_name='geography_tags')
geography_categories
```

	Location	Standardized_Country	Geography Categories_2nd	Geography Categories_3rd	Geography Categories_by countries
0	serves as managing director at dr. hettich bet...	NaN	Not Relevant	Not Relevant	Not Relevant
1	board member at dimo.	NaN	Not Relevant	Not Relevant	Not Relevant
2	48	NaN	Not Relevant	Not Relevant	Not Relevant
3	65	NaN	Not Relevant	Not Relevant	Not Relevant
4	president at moladin. he is an angel investor.	NaN	Not Relevant	Not Relevant	Not Relevant
...	...	...	...	...	...
3691	gabon	Gabon	Gabon	Gabon	Gabon
3692	florida.	NaN	Florida	Florida	United States
3693	chad	Chad	Chad	Chad	Chad
3694	austria.	NaN	Austria	Austria	Austria
3695	north carolina.	NaN	North Carolina	North Carolina	United States

3696 rows × 5 columns

Extract Tag Column - Preferred Investment Type Tags

- Extract Tag Column - Preferred Investment Type Tags

Extract all the unique items in the column to excel file and review manually

```
investment_tags = pd.Series( drop_df_removed_shift['preferred_investment_type_tags'].explode().unique().tolist() )

# To output an excel file, remove the # sign in the next line
investment_tags.to_excel('preferred_investment_type_tags.xlsx', index=False, header=['Preferred Investment Type Tags'])
```

After manual review, import the respective excel worksheet

```
investment_categories = pd.read_excel('Tag Columns Classification.xlsx', sheet_name='preferred_investment_type_tags')
investment_categories
```

	Preferred Investment Type Tags	Investment Categories
0	accelerator/incubator	Accelerator/incubator
1	early stage vc	Venture capital
2	later stage vc	Venture capital
3	seed round	Seed round
4	pe growth/expansion	Private equity
...	...	...
401	served as non-executive director at unith. he ...	Not relevant
402	new york	Not relevant
403	aluminum cans	Not relevant
404	building refurbishment	Not relevant
405	serves as board member at rogue signal. he is ...	Not relevant

406 rows × 2 columns

Extract Tag Column - Preferred Industry Tags

- Extract Tag Column - Preferred Industry Tags

Extract all the unique items in the column to excel file and review manually

```
industry_tags = pd.Series( drop_df_removed_shift['preferred_industry_tags'].explode().unique().tolist() )

# To output an excel file, remove the # sign in the next line
industry_tags.to_excel('preferred_industry_tags.xlsx', index=False, header=['preferred_industry_tags'])

industry_tags
```

```
0                                        beverages
1                                 computer hardware
2                education and training services (b2b)
3               educational and training services (b2c)
4                                           energy
                              ...
1002    serves as interim chief executive officer and ...
1003    chief growth officer at mighty (business/produ...
1004                             advisor at foody.vn.
1005    chief operating officer at cheezburger. he has...
1006     safety products for organizations across canada.
Length: 1007, dtype: object
```

After manual review, import the respective excel worksheet

```
industry_categories = pd.read_excel('Tag Columns Classification.xlsx', sheet_name='preferred_industry_tags')
industry_categories = industry_categories.loc[:,'preferred_industry_tags':'industries categories']
industry_categories
```

	preferred_industry_tags	industries categories
0	animal husbandry	Agriculture

Output the Final Dataframe after One Hot Encoding for further study.
Combine the encoded columns to the original dataset dataframe.

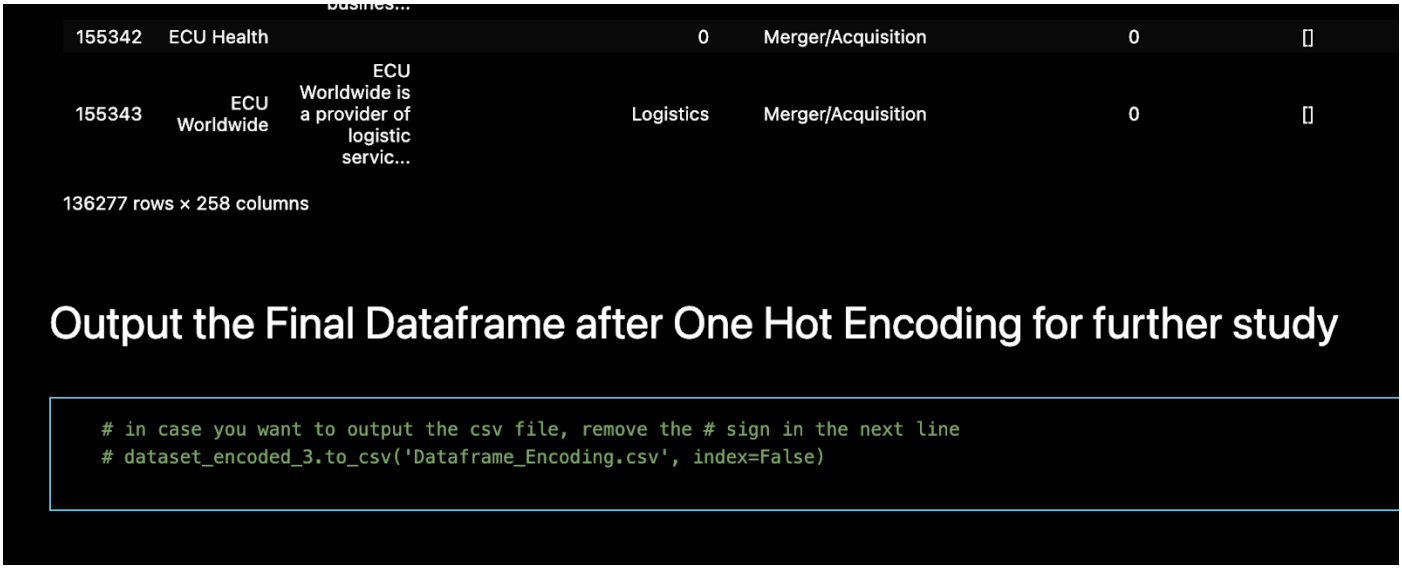
Combine the encoded columns to the original dataset dataframe

```
dataset_encoded_3 = dataset_encoded_2.merge(
    industry_encode_columns,
    left_index=True,
    right_index=True,
    how='inner'
)

dataset_encoded_3

# To output an excel file, remove the # sign in the next line
# dataset_encoded_2.to_excel('dataset_encoded_2.xlsx', index=False)
```

	Investors	Description	Geography	Preferred Industry	Preferred Investment Type	Primary In
0	Techstars	Founded in 26, Techstars is an accelerator bas...	Africa, Americas, Asia, Canada, Middle East, O...	Beverages, Computer Hardware, Education and Tr...	Accelerator/Incubator, Early Stage VC, Later S...	Accelerator/Inc
1	Y Combinator	Founded in 25, Y Combinator is an accelerator ...	Africa, Americas, Asia, Europe, Oceania, Unite...	Biotechnology, Commercial Transportation, Comm...	Accelerator/Incubator, Early Stage VC, Later S...	Accelerator/Inc
2	Plug and Play Tech Center	Founded in 26, Plug and Play Tech Center is an...	Africa, Americas, Asia, Canada, Europe, Middle...	Aerospace and Defense, Animal Husbandry, Aquac...	Accelerator/Incubator, Early Stage VC, Later S...	Accelerator/Inc
		Founded in 214, Gaingels	Africa, Americas, Asia,	Business Products and	Early Stage VC, Later	



Methodology & Analysis

To carry out the analysis, we used Python to clean the data and prepare it for modeling. After preprocessing, we used data visualization to explore patterns and gain insights. Then, we applied machine learning models to make predictions and support investor classification.

This approach was chosen because the original dataset was unstructured and messy. By cleaning and structuring the data first, we could ensure reliable and meaningful analysis results.

Pre-Processing

To ensure high data quality and model readiness, we performed a series of rigorous preprocessing steps on the raw dataset, which initially contained over 2 million records.

First, we removed duplicate records. After resolving issues with file loading and filtering out repeated entries and investors with duplicated inputs, the dataset was reduced from 155,344 rows to 139,710 unique records across 10 columns.

Next, we addressed missing values. Although some records had incomplete entries in fields such as “Preferred Industry,” “Preferred Investment Type,” and “Primary Investor Type,” we retained these rows due to the client’s preference for maximizing available information. This resulted in a final dataset of 139,710 rows with missing values preserved.

We also detected and removed shifted or malformed entries in the “Description” column. These included rows containing improperly placed URLs (e.g., entries starting with “www.” or ending with “.com” or “.org”), which accounted for over 3,400 rows. After filtering these cases, we obtained a cleaner dataset with 136,277 rows and 9 well-structured columns.

Lastly, we conducted one-hot encoding on three key categorical tag columns: “Geography,” “Preferred Investment Type,” and “Preferred Industry.” We first extracted and reviewed unique values from each column manually, then standardized and mapped them to curated category lists. These tag categories were matched using external databases such as GeoNames and PyCountry for geographic standardization. The encoded tag features were merged back into the main dataset, expanding the feature space from 9 to 258 columns.

The resulting dataset was now fully cleaned, structured, and enriched with encoded features, providing a solid foundation for machine learning modeling and visualization.

Exploratory Data Analysis (EDA)

1. Preferred Geography Categories

According to Figure 1 and Figure 2, which rank the regions based on the information filled by the respondents, the US (21.60%), Europe (16.80%), and Canada (12.40%) scored the top three in terms of the investors’ preferred geography categories.

Dashboard 1 summarises the investors’ preferred country based on the country after filtering out the entries that are not on the expected country level such as Europe and Los Angeles. The map reflected that the investors’ vote predominantly goes to North America, which is in line with Figure 1 and 2.

Interestingly, discordant with Figure 1 and 2, Dashboard 1 highlighted that the Philippines are notably one of the most popular countries among investors.

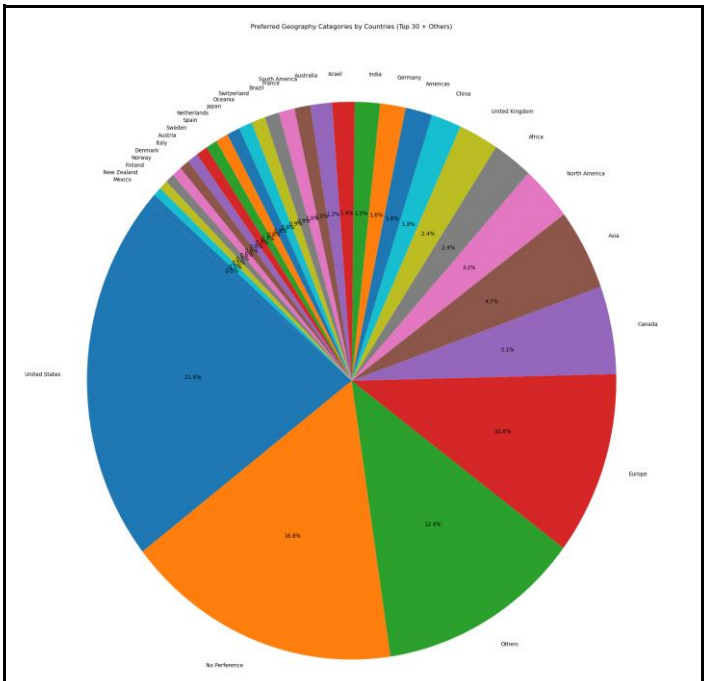


Figure 1 : Preferred Geography Categories by Countries (Top 30) - Pie Chart

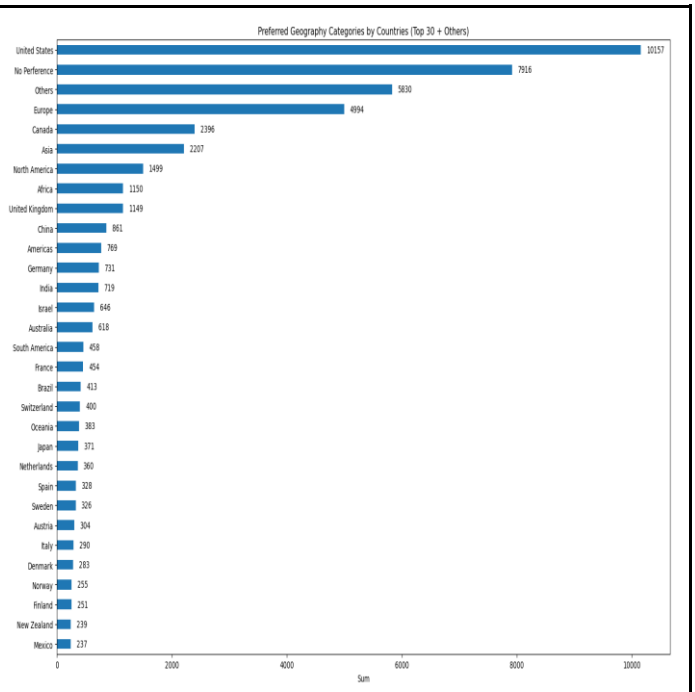
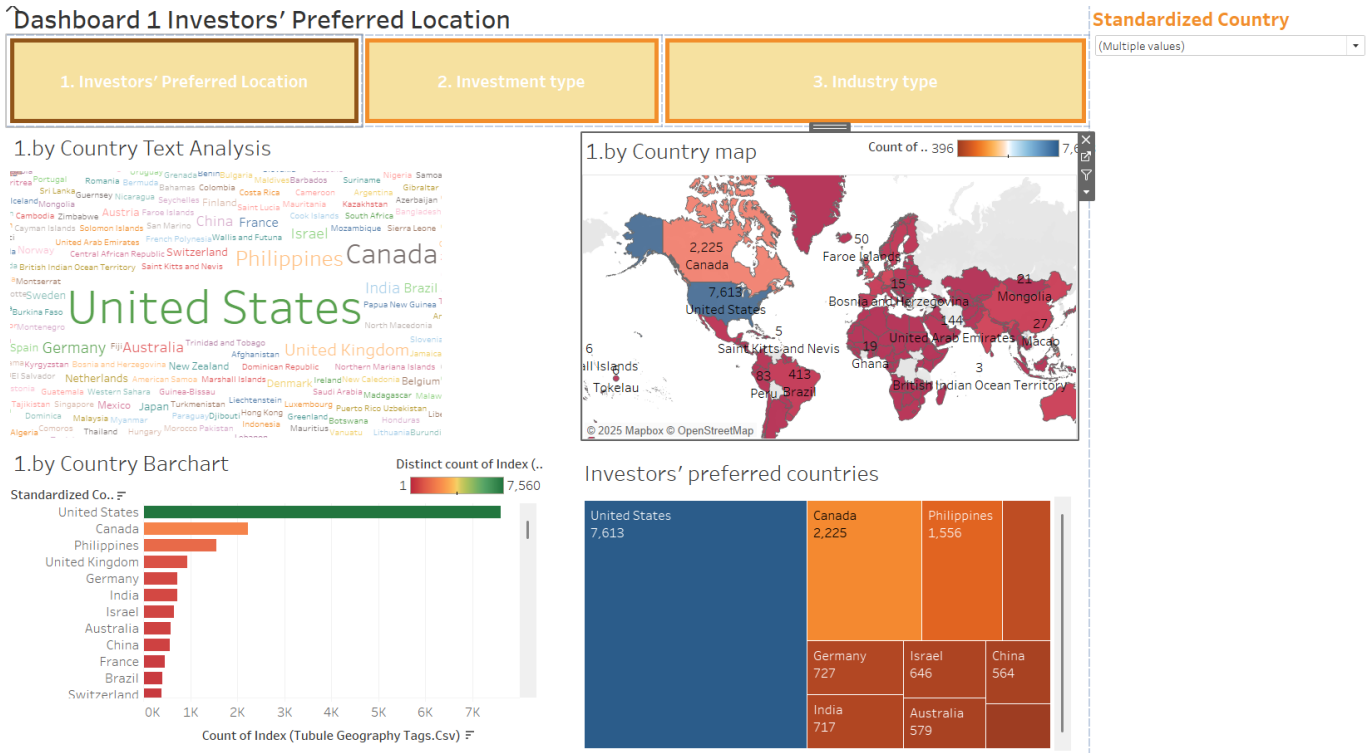


Figure 2 : Preferred Investment Type - Barchart



Dashboard 1 : Investors' Preferred Location

2.Investors' Preferred Investment Type

Figure 3, Figure 4, along with Dashboard 2 indicated that Mergers and Acquisitions gained the greatest support from the investors when it comes to the investment type, with 75,183 investors (37.6%), over one-third of the respondents, picking such investment type as the most preferred. This might be explained by the fact that mergers often involve more than one investor in the investment while acquisitions denote takeover of the existing business, thus enabling investors to equally share the risks with their partners or further develop the original business by utilising the existing resources. This might potentially diminish the risk.

In addition, 40,586 investors (20.30%) and 21,233 investors (10.6%) chose Venture capital and Buyout respectively, thus ranking the second and third most popular investment types.

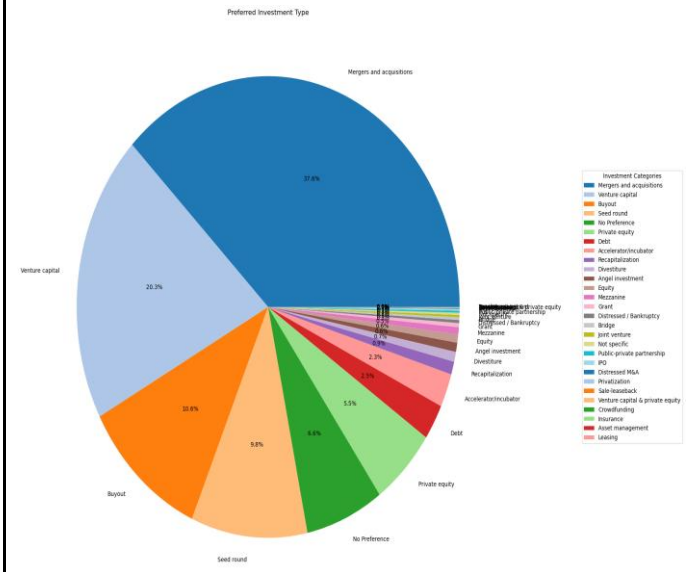


Figure 3 : Preferred Investment Type - Pie Chart

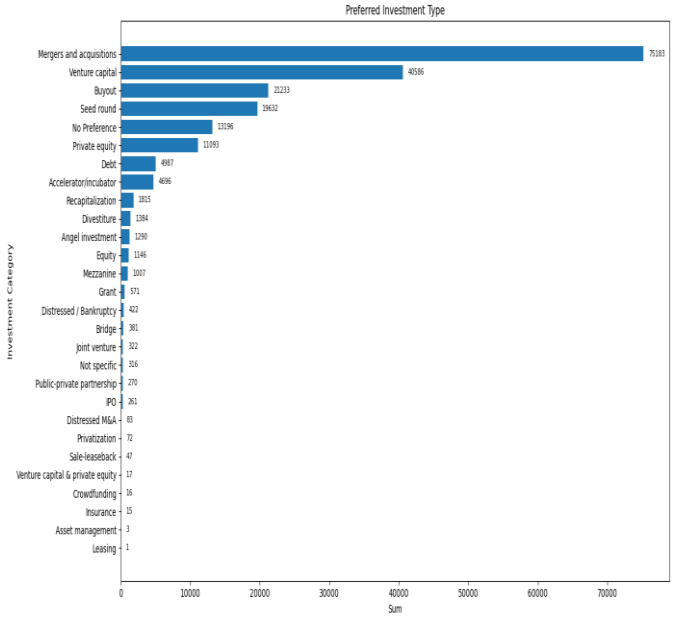
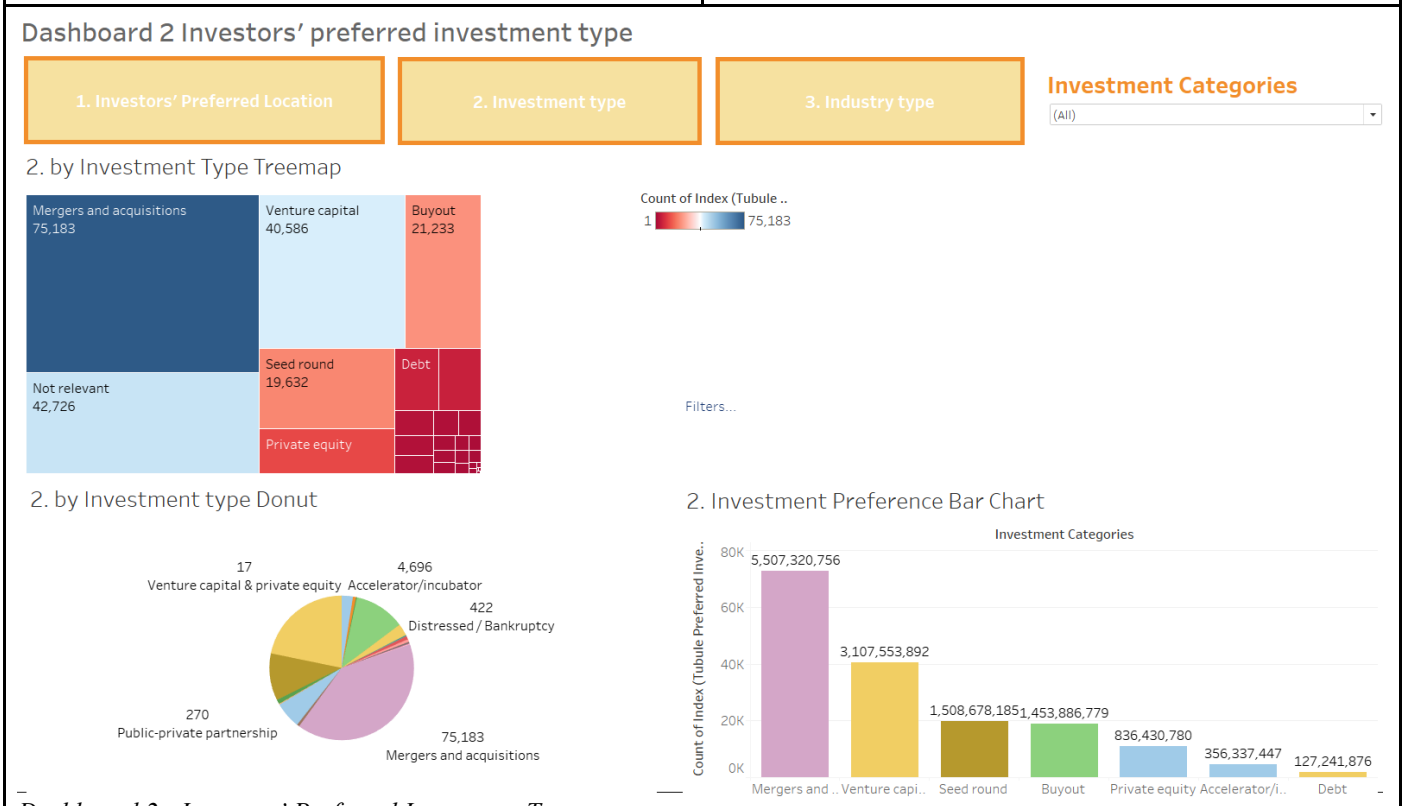


Figure 4 : Preferred Investment Type - BarChart



Dashboard 2 : Investors' Preferred Investment Type

3.Investors' Preferred Industry Type

With reference to Figure 5, 6, and Dashboard 3, it is reported that STEM, with 46,515 votes and accounting for 17.2% of the total count of votes, shows the highest ranking overall in respect of the investors' preferred industry type. Such a high number of votes might be explained by the rapid advancement and growth in the technological field, especially the Artificial Development (AI), which are generally considered as lucrative and profitable from the perspective of investors and thus worth the investments, in recent years.

Another noteworthy observation is that the manufacturing, industrialising-related industry that gains 12.6% of the votes, is the second most popular following the STEM industry, which can possibly attributed to the fact that such industry often goes hand in hand with the top 1 STEM industry, with the manufacturing of silicon chip and semi-conductor that are indispensable for the production of electronic gadgets serving as a prominent example.

Not surprising, the healthcare industry is the third most popular, which can undoubtedly be ascribed to the aging population that is a pressing global phenomenon and drives up the demand for the talents, equipment, and medicines in relation to the healthcare field.

As for the correlation between the preferred industry type, correlation between tourism and catering industry is 0.46 - moderate in strength - which implies that the investors who prefer tourism are more inclined to the catering industry.

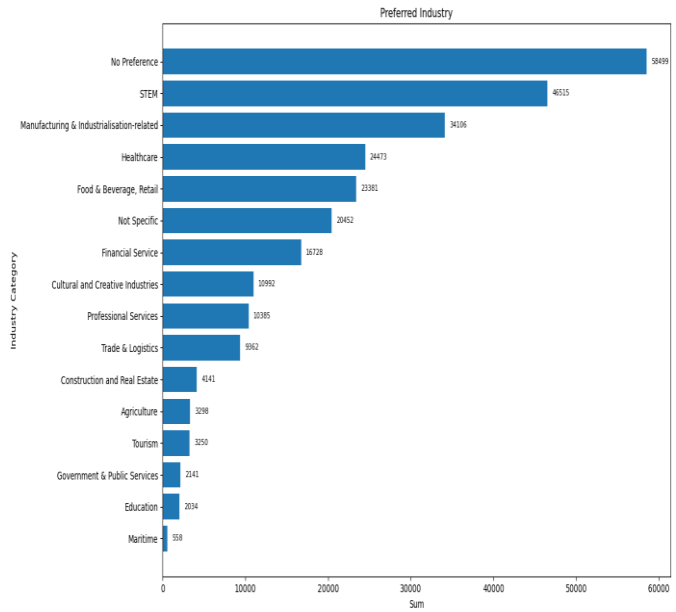


Figure 5 : Preferred Industry Type - BarChart

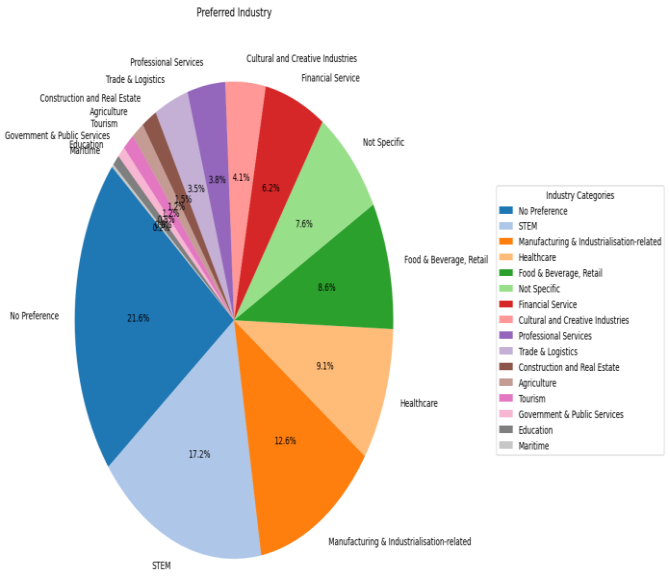
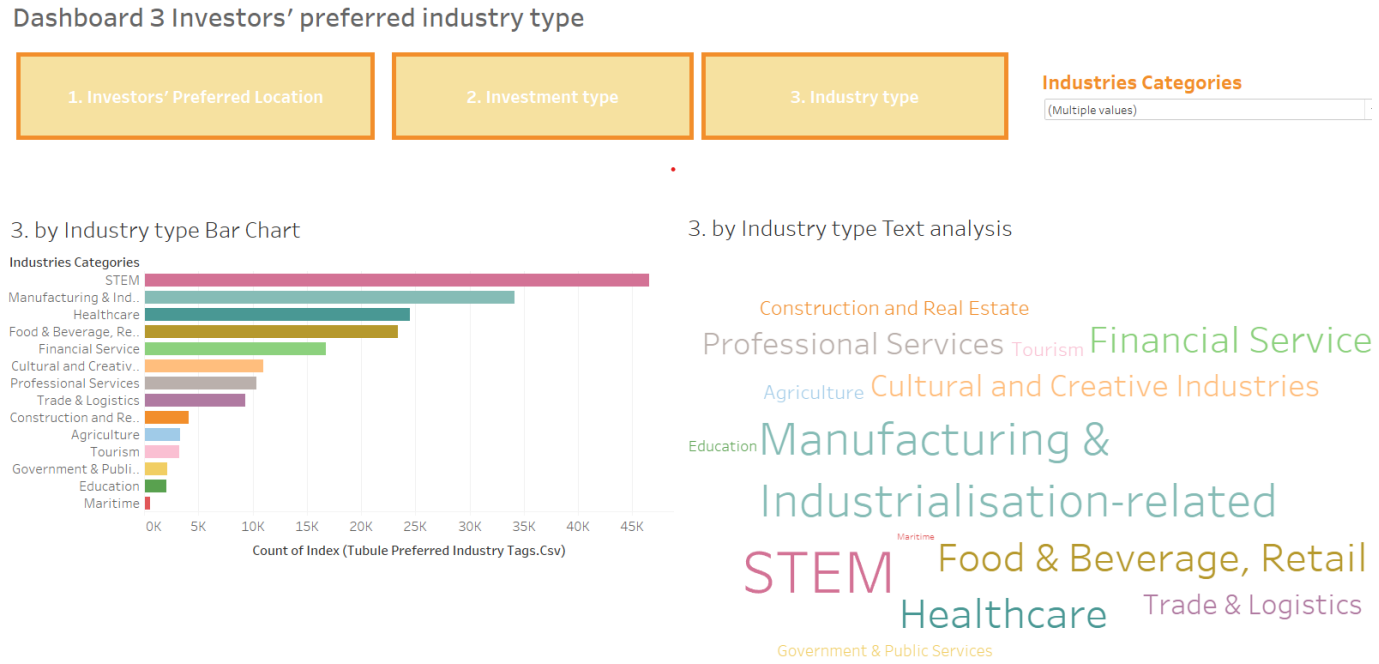


Figure 6 : Preferred Industry Type - Pie Chart



Dashboard 3 : Investors' Preferred Industry Type

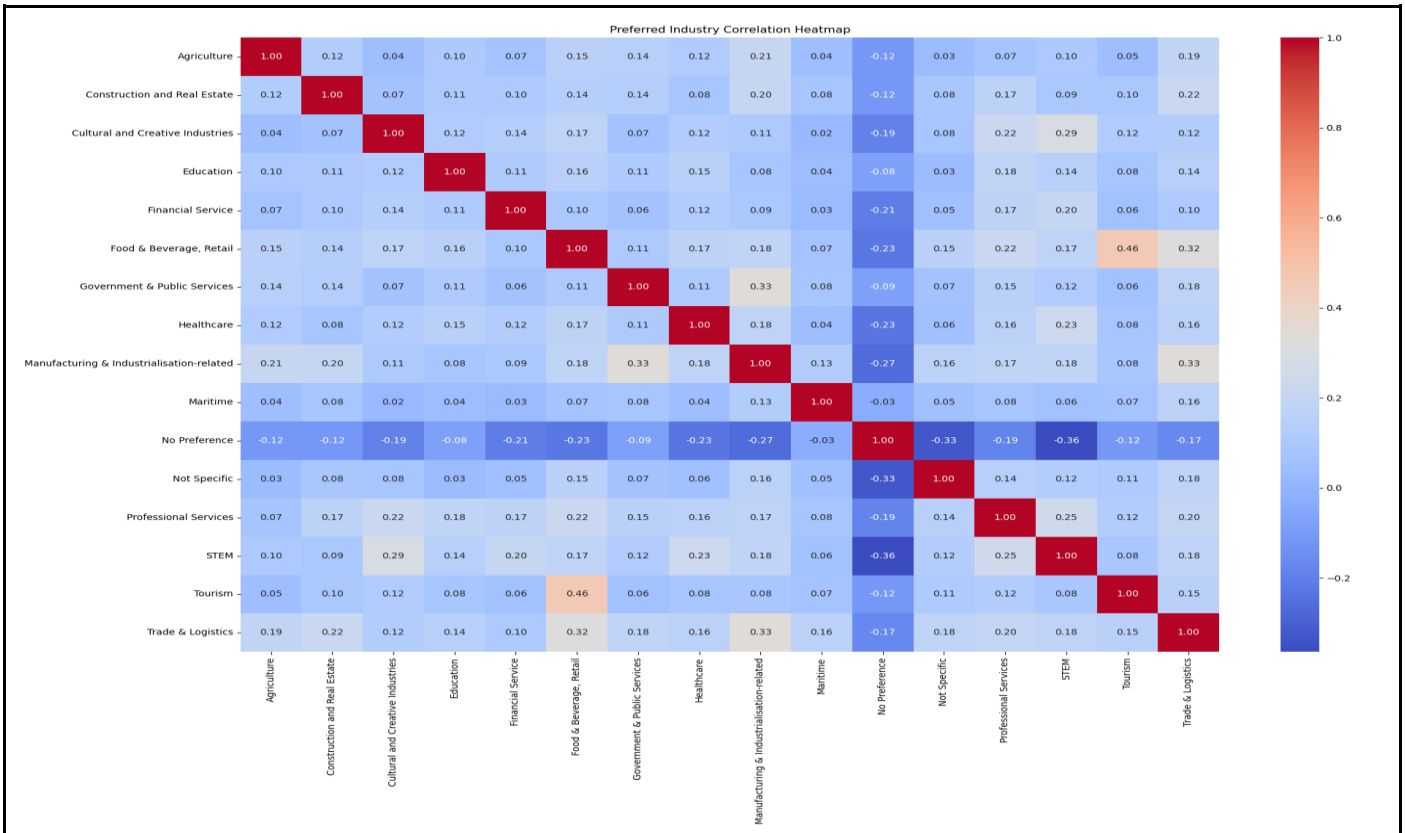


Figure 7 : Correlation between Preferred Industry - Heatmap

Machine Learning

In the machine learning session, 3 domains were covered – Clustering, Association Rule, Supervised Learning.

Clustering	- Investor Cluster (investor + their main preference of sub-tags) - Tag Cluster (sub-tags) → Recommendation of relevant sub-tags
Association Rule	- Among the 3 main tags - Between "Primary Investor Type" and "Investment Type"
Supervised Learning	- Use 2 main tag columns to forecast the Remaining 1 main tag Column

Primary Investor Type: It is the investor's background.
Geography, Investment Type and Industry: They are the investment preference from investors.
For easy identification, column names "Geography, Investment Type and Industry" were renamed as "main tags" and its row elements as "sub tags".

Clustering: Sub-tags were used to find investor clusters and their main preference. They were also used to find tag clusters within each main tag, which can be used for recommendation to investors with similar interest.

Association Rule: The association among the 3 main tags and the association between "Primary Investor Type" and "Investment Type" were studied.

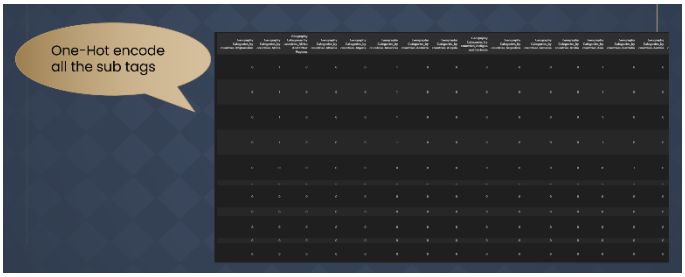
Supervised Learning: 2 main tag columns were used to forecast the remaining 1 main tag column.

The detail of each domain is going to be outlined in the following one by one.

Clustering

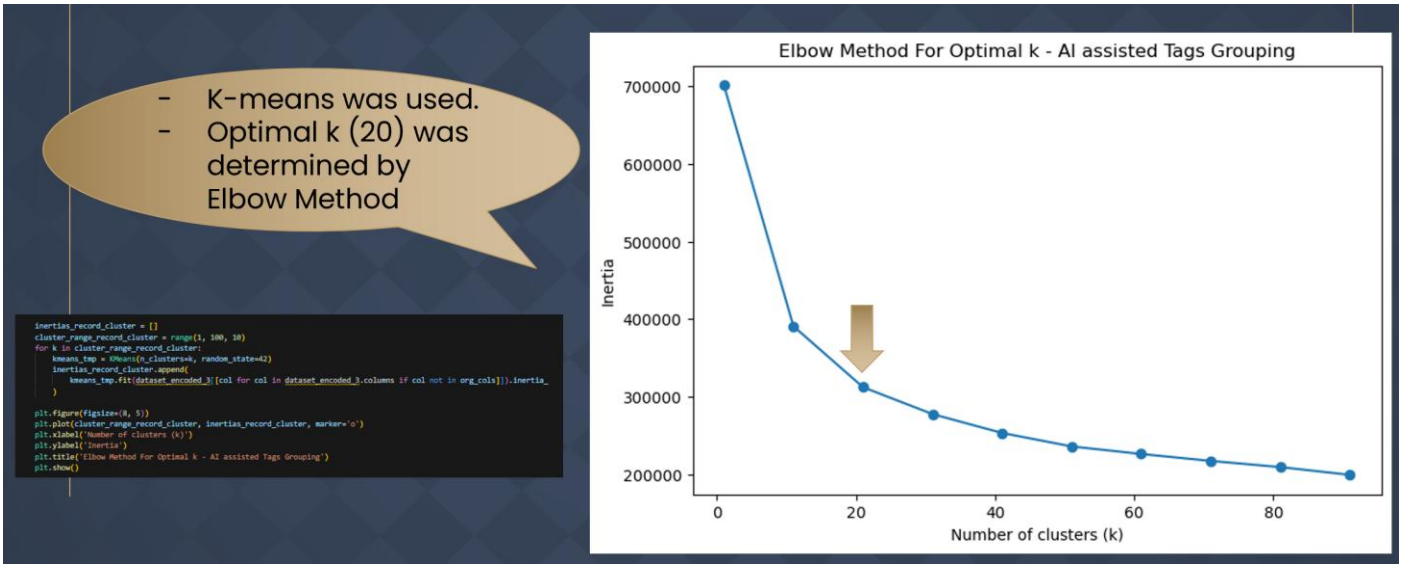
The process of finding the two clusters is demonstrated.

(Investor Cluster)



Step 1 - Encoding

One-Hot Encoding was performed to all the sub tags.



Step 2 – Apply Algorithm

K-means clustering was used to calculate and determine the optimal number of k using the Elbow Method, which was found to be 20 clusters.

Step 3 – Outcome Interpretation

Investors sharing the same interest were successfully divided into clusters and could be extracted for other purposes.

Step 4 – Find the main preference of sub tags in each cluster

Cluster_row	Geography Categories_by countries_Afghanistan	Geography Categories_by countries_Africa	Geography Categories_by countries_Africa And Other Regions	Geography Categories_by countries_Albania	Geography Categories_by countries_Algeria	Geography Categories_by countries_Americas	Geography Categories_by countries_Andorra	Geography Categories_by countries_Angola	Geography Categories_by countries_Antigua and Barbuda	Geography Categories_by countries_Argentina
0	0.000000	0.030303	0.000000	0.001036	0.000000	0.017612	0.000518	0.000259	0.000259	0.001036
1	0.000000	0.000310	0.000000	0.000000	0.000000	0.000225	0.000028	0.000000	0.000000	0.000056
2	0.000000	0.006729	0.000000	0.000000	0.000000	0.004527	0.000367	0.000122	0.000000	0.000367
3	0.000000	0.020958	0.000000	0.000000	0.000000	0.019162	0.001198	0.000000	0.000599	0.002994
4	0.000000	0.027030	0.000000	0.000620	0.000248	0.017979	0.000248	0.000000	0.000124	0.001860
5	0.000000	0.003715	0.000000	0.000149	0.000000	0.001337	0.000000	0.000000	0.000000	0.000446
6	0.000000	0.083601	0.000000	0.006431	0.000000	0.083601	0.006431	0.000000	0.000000	0.006431
7	0.000000	0.001962	0.000000	0.000000	0.000000	0.001744	0.000000	0.000000	0.000000	0.000000
8	0.000000	0.020330	0.000000	0.000000	0.000000	0.013977	0.000000	0.000000	0.000000	0.002541
9	0.000000	0.041026	0.000000	0.002564	0.000000	0.043590	0.000000	0.000000	0.000000	0.002564
10	0.000000	0.012638	0.000000	0.000527	0.000000	0.009479	0.000527	0.000000	0.000000	0.002106
11	0.000000	0.019868	0.002208	0.000000	0.000000	0.011038	0.002208	0.000000	0.000000	0.004415
12	0.000000	0.040204	0.000000	0.000638	0.000000	0.030632	0.001914	0.000638	0.000000	0.000000
13	0.000000	0.003564	0.000000	0.000000	0.000086	0.001503	0.000000	0.000086	0.000086	0.000301
14	0.000000	0.003281	0.000000	0.000000	0.000080	0.002080	0.000000	0.000000	0.000000	0.000320
15	0.000000	0.072539	0.000000	0.000000	0.000000	0.062176	0.000000	0.000000	0.005181	0.000000
16	0.000000	0.001034	0.000000	0.000000	0.000000	0.001264	0.000230	0.000000	0.000000	0.000115
17	0.000000	0.044498	0.000000	0.000809	0.000000	0.025890	0.000809	0.000000	0.000000	0.001618
18	0.000000	0.013770	0.000000	0.000459	0.000153	0.005814	0.000000	0.000000	0.000000	0.000918
19	0.000447	0.046673	0.000000	0.000893	0.000447	0.036623	0.000223	0.000223	0.000000	0.002903

Firstly, the average number of occurrence of sub tags was calculated in each cluster.

```
def top_3_values(row):
    return row.nlargest(3)

top_3_cluster_geography = record_cluster_geography.apply(top_3_values, axis=1)
top_3_cluster_investment = record_cluster_investment.apply(top_3_values, axis=1)
top_3_cluster_industry = record_cluster_industry.apply(top_3_values, axis=1)

top_3_cluster_industry
```

Cluster_row	Industries categories_Cultural and Creative Industries	Industries categories_Financial Service	Industries categories_Food & Beverage, Retail	Industries categories_Healthcare	Industries categories_Manufacturing & Industrialisation-related	Industries categories_Not Relevant	Industries categories_Not Specific
0	NaN	NaN	NaN	0.360336	0.419052	NaN	NaN
1	NaN	0.568362	NaN	0.960769	NaN	0.819944	NaN
2	NaN	NaN	NaN	NaN	1.628166	NaN	0.145218
3	NaN	NaN	NaN	0.403764	0.382036	NaN	NaN
4	0.188145	0.159863	NaN	NaN	NaN	NaN	NaN
5	NaN	NaN	1.408143	NaN	0.134827	NaN	NaN
6	NaN	NaN	NaN	0.827358	1.158077	NaN	NaN
7	NaN	NaN	NaN	NaN	0.131190	NaN	1.050282
8	NaN	NaN	NaN	4.191233	0.415532	NaN	NaN
9	NaN	NaN	1.320513	2.182564	NaN	NaN	NaN
10	NaN	2.733097	NaN	NaN	NaN	NaN	0.140600
11	NaN	NaN	NaN	NaN	0.295806	0.320971	NaN
12	NaN	NaN	NaN	NaN	4.246511	NaN	NaN
13	NaN	0.021299	NaN	0.057142	NaN	0.800581	NaN
14	0.123830	NaN	NaN	NaN	NaN	NaN	0.111628
15	NaN	NaN	3.266839	NaN	7.029016	NaN	NaN
16	NaN	NaN	NaN	NaN	0.187179	0.292954	0.239535
17	NaN	NaN	4.137179	NaN	NaN	NaN	NaN
18	NaN	NaN	0.275133	NaN	0.278794	0.669510	NaN
19	NaN	NaN	NaN	1.134212	0.405315	NaN	NaN

Secondly, the top 3 sub tags of each main tag were filtered in each cluster

```
Cluster 0:
['Geography Categories_by countries_United States', 'Geography Categories_by countries_Europe', 'Geography Categories_by countries_Canada']
['Investment Categories_Buyout', 'Investment Categories_Private equity', 'Investment Categories_Mergers and acquisitions']
['industries categories_STEM', 'industries categories_Manufacturing & Industrialisation-related', 'industries categories_Healthcare']


Cluster 1:
['Geography Categories_by countries_United States', 'Geography Categories_by countries_Not Relevant', 'Geography Categories_by countries_Europe']
['Investment Categories_Mergers and acquisitions', 'Investment Categories_Private equity', 'Investment Categories_Buyout']
['industries categories_Not Relevant', 'industries categories_Healthcare', 'industries categories_Financial Service']
```

For all the 20 investor clusters, we successfully extracted their main preference in “Geography, Investment Type, Industry”.

To illustrate, investors in “Cluster 0” preferred to invest in the United States, Europe and Canada. Given the geography options, they were interested in investing in the STEM, Manufacturing & Industrialization-related and Healthcare industries of those regions. With those preferences, investors would like to finance their investment projects in the form Buyout, Private Equity and Mergers and Acquisitions. The same applies for all the remaining clusters which empowers us to group investors into clusters for subsequent treatment.

(Tag Cluster)

```
model = SentenceTransformer('all-MiniLM-L6-v2')
```

Embedding

For each main tag, all its sub-tags were embedded by sentence transformers and arrays of vectors were returned.

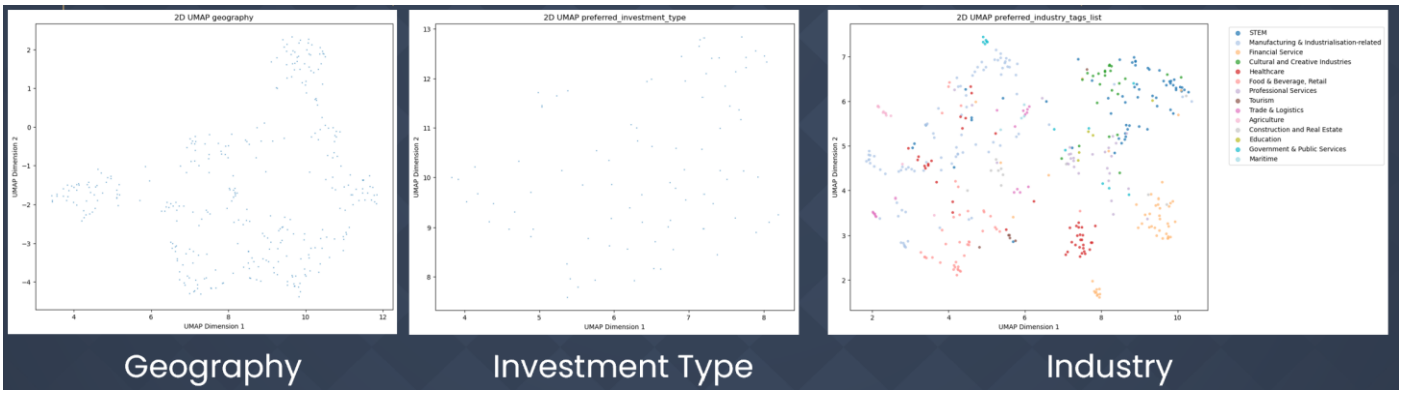
```
import umap.umap_ as umap

#Reduce dimensions using umap (good for large dataset)
umap_model = umap.UMAP(n_components=2, random_state=42)
geography_embeddings_umap = umap_model.fit_transform(geography_embeddings)
geography_embeddings_umap
```

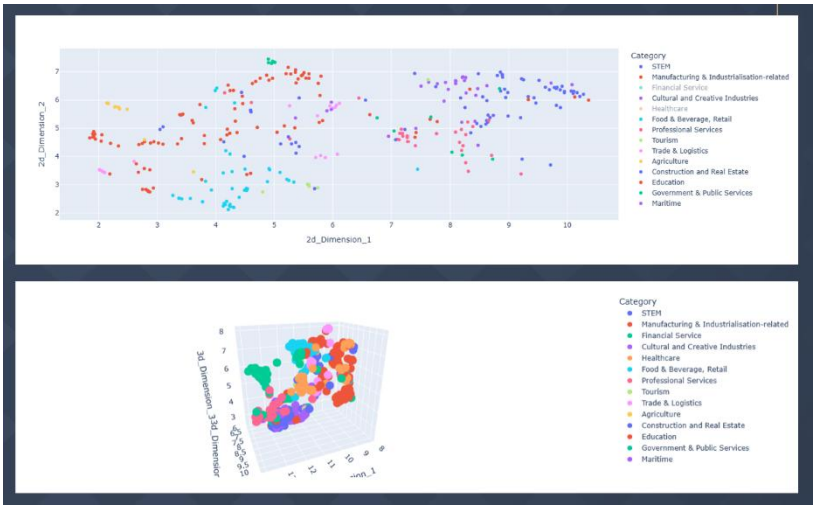
	Geography	Dimension_1	Dimension_2
1	united states	7.426336	-2.142599
2	europe	9.057121	-3.145132
3	canada	6.835590	-2.037454
4	north america	7.304031	-2.366180
5	asia	11.533536	-1.876656
...	...	...	...
3369	saudi arabia.	10.697309	-0.800912
3397	lesotho	10.021734	1.998306
3552	india.	10.518393	-1.364715
3629	africa and other regions.	10.465020	1.893452
3680	guam	7.031823	-0.894585

Dimensionality Reduction

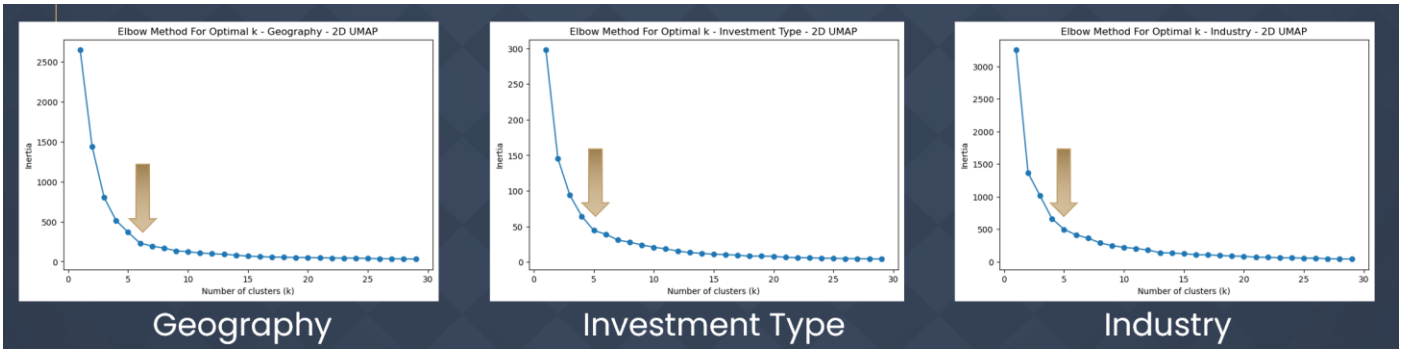
UMAP was used to reduce the dimensions of vectors down to only 2 dimensions while preserving as much of the dataset's structure and characteristics as possible.



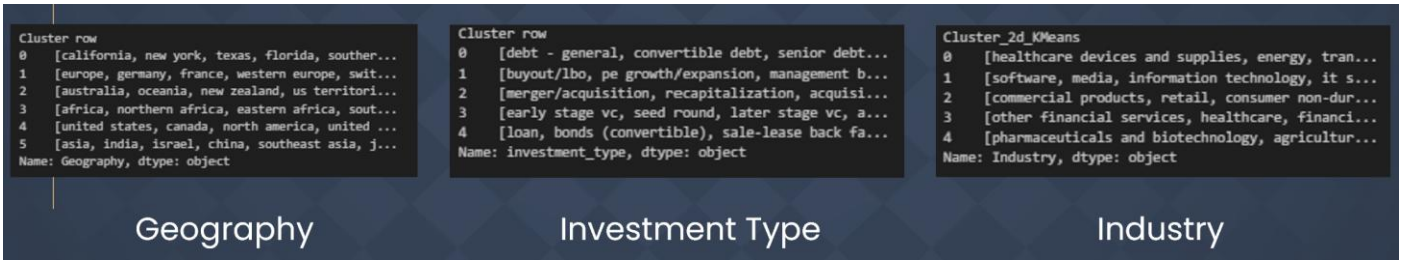
After dimensionality reduction, the outcome seemed generally scattered, like images in Geography and Investment Type. However, when a brief categorization was assisted by LLM (indicated by the color legend of Industry) in the image of industry, the scattered points were indeed grouped.



To get a better visualization, 'Plotly' module was further used to interpret the image of industry. It was observed that the scattered points were intuitively categorized into industries according to their proximity of vector value after dimensionality reduction.

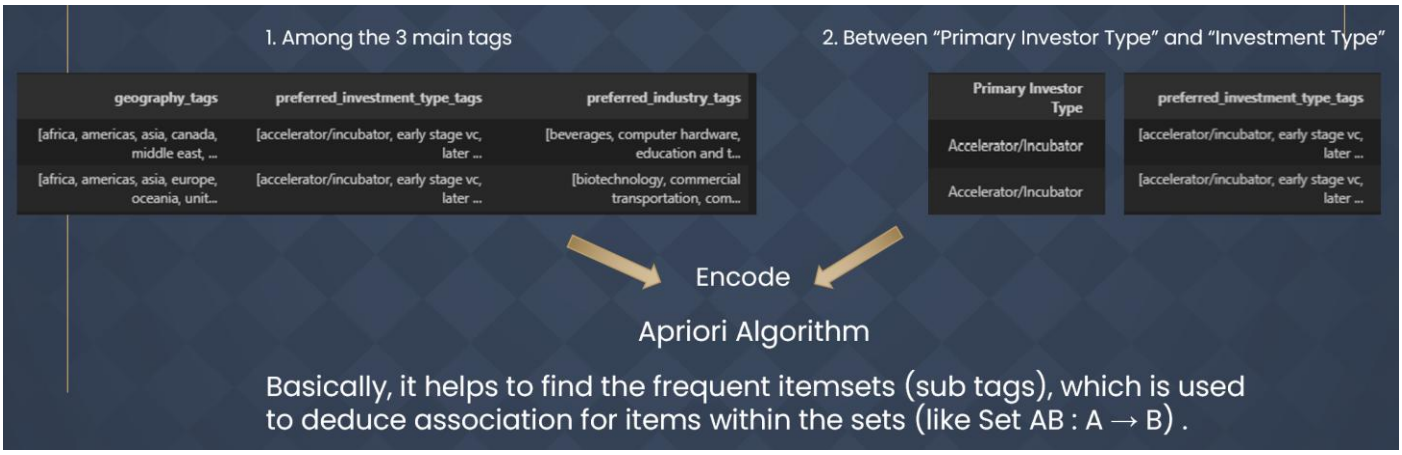


After that, K-means and Elbow methods were again used. The optimal numbers of K were found to be 6, 5 and 5 for Geography, Investment Type and Industry respectively.

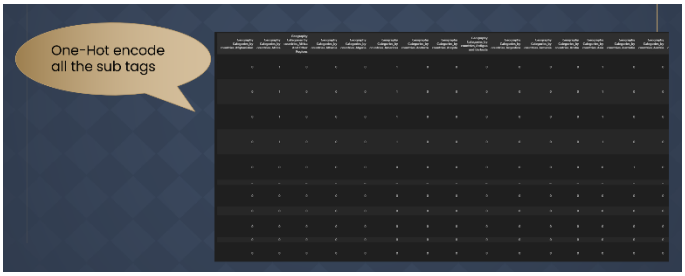


Sub Tags were clustered with result output, which can be used for recommendation to investors of similar interests.
For example, when one investor prefers investment in California, this clustering data could be used to recommend the investor to similar regions (like New York, Texas, Florida, …) that he/she might be interested to look for.

Association Rule



In this study, two associations were studied, one of which was among the 3 main tags and another was between "Primary Investor Type" and "Investment Type". The steps of finding the two associations were the same.



Step 1 - Encoding
One-Hot Encoding was performed to all the sub tags.

```
from mlxtend.frequent_patterns import apriori, association_rules

# Select the columns for association rule mining
data_for_association = dataset_encoded_3_cluster[[col for col in dataset_encoded_3_cluster.columns if col not in org_cols]].drop(columns=['Cluster_row']).applymap(lambda x: 1 if x > (

# Find frequent itemsets
frequent_itemsets = apriori(data_for_association, min_support=0.05, use_colnames=True)

# Generate association rules
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1.0)

rules_sorted = rules.sort_values('lift', ascending=False)
rules_sorted
```

Step 2 – Apply Algorithm
Apriori Algorithm was applied to the encoded sub tags.

There were adjusted parameters
min_support : can be adjusted to a higher level (in this case 0.05) to get frequent itemsets with higher occurrence threshold and thus a higher association for items within sets.

min_threshold : set as 1.0 , which indicates only positive association between items within sets was accepted in the study.

Step 3 – Outcome Interpretation

1. Among the 3 main tags

antecedents	consequents
industries categories_STEM', 'Investment Categories_Venture capital'	Investment Categories_Seed round'
Investment Categories_Seed round'	industries categories_STEM', 'Investment Categories_Venture capital'
Investment Categories_Venture capital'	Investment Categories_Seed round'
Investment Categories_Seed round'	Investment Categories_Venture capital'
Investment Categories_Venture capital'	industries categories_STEM', 'Investment Categories_Seed round'
industries categories_STEM', 'Investment Categories_Seed round'	Investment Categories_Venture capital'
Investment Categories_Buyout'	Investment Categories_Private equity'
Investment Categories_Private equity'	Investment Categories_Buyout'
industries categories_STEM'	Investment Categories_Seed round'
Investment Categories_Seed round'	industries categories_STEM'
industries categories_STEM'	Investment Categories_Venture capital', 'Investment Categories_Seed round'
Investment Categories_Venture capital', 'Investment Categories_Seed round'	industries categories_STEM'
industries categories_Healthcare'	Investment Categories_Venture capital'
Investment Categories_Venture capital'	industries categories_Healthcare'
industries categories_STEM'	Investment Categories_Venture capital'
Investment Categories_Venture capital'	industries categories_STEM'
industries categories_STEM'	industries categories_Healthcare'
industries categories_Healthcare'	industries categories_STEM'
Investment Categories_Mergers and acquisitions'	industries categories_Manufacturing & Industrialisation-related'
industries categories_Manufacturing & Industrialisation-related'	Investment Categories_Mergers and acquisitions'

A → B

20 associations were extracted & sorted in descending order. Those associations could be divided into 3 main groups.

(Green Color) Association between Industry & Investment Type	• Investors who preferred "STEM industry" were likely to invest in the form of "Venture Capital / Seed Round" • Investors who preferred "Manufacturing & Industrialization-related" were likely to invest in the form of "Mergers and Acquisitions"
(Orange Color) Association within Investment Type	• Investors who preferred investment type "Venture Capital" were also interested in investing in the form of "Seed Round" • Investors who preferred investment type "Private Equity" were also interested in investing in the form of "Buyout"
(Yellow Color) Association within Industry	• Investors who preferred "STEM industry" were also interested in investing in "Healthcare industry" as well

2. Between "Primary Investor Type" and "Investment Type"

antecedents	consequents
Investment Categories_Venture capital', 'Investment Categories_Seed round'	Primary Investor Type_Venture Capital'
Primary Investor Type_Venture Capital'	Investment Categories_Venture capital', 'Investment Categories_Seed round'
Investment Categories_Seed round'	Primary Investor Type_Venture Capital', 'Investment Categories_Venture capital'
Primary Investor Type_Venture Capital', 'Investment Categories_Venture capital'	Investment Categories_Seed round'
Primary Investor Type_Venture Capital'	Investment Categories_Seed round'
Investment Categories_Seed round'	Primary Investor Type_Venture Capital'
Primary Investor Type_Venture Capital', 'Investment Categories_Seed round'	Investment Categories_Venture capital'
Investment Categories_Venture capital'	Primary Investor Type_Venture Capital', 'Investment Categories_Seed round'
Investment Categories_Venture capital'	Primary Investor Type_Venture Capital'
Primary Investor Type_Venture Capital'	Investment Categories_Venture capital'
Investment Categories_Mergers and acquisitions'	Primary Investor Type_Corporation'
Primary Investor Type_Corporation'	Investment Categories_Mergers and acquisitions'

A → B

12 associations were extracted & sorted in descending order. Those associations could be divided into 2 main groups.

Green Color	• Investors who had "Venture Capital" background itself were likely to invest in the form of "Venture Capital / Seed Round"
Orange Color	• Investors who had "Corporation" background itself were likely to invest in the form of "Mergers and Acquisition"

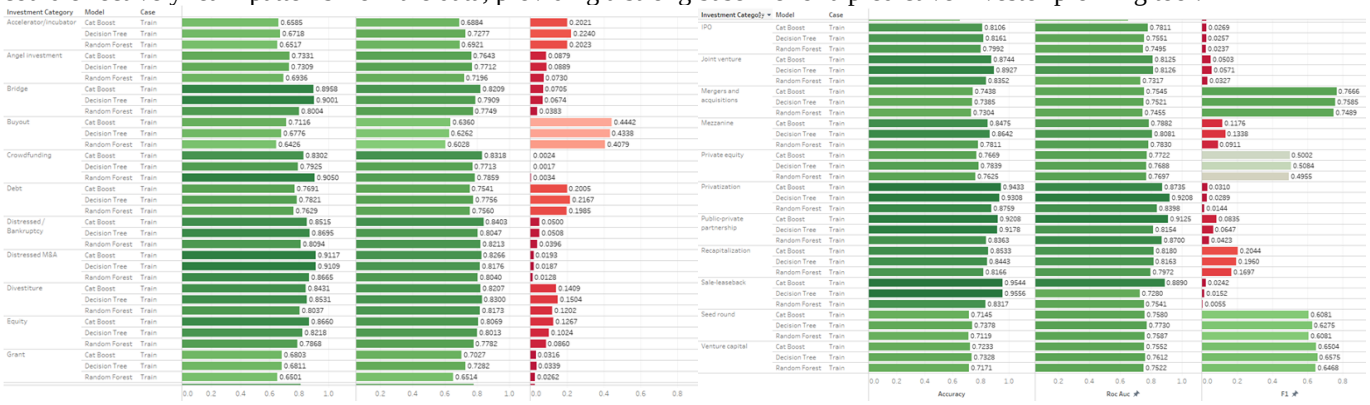
Supervised Learning

Objective:
A multi-label classification task was designed to predict an investor's preferred Investment Category (22 possible types) based on their preferred Geography and Industry.

Challenge:
The target variable exhibited extreme class imbalance (a "long-tail" distribution), where some categories like "Mergers and acquisitions" were very common and others like "Crowdfunding" were rare.

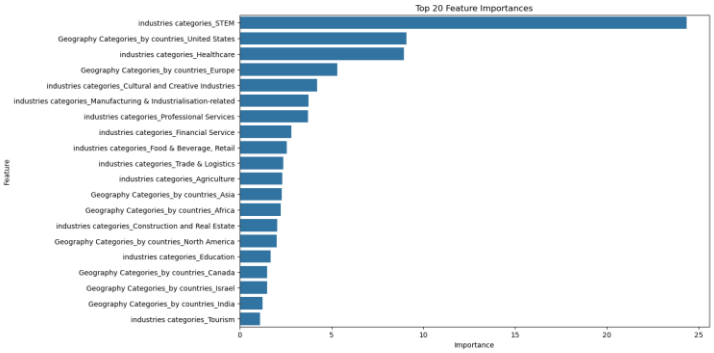
Methodology:
1. Decision Tree: Highly interpretable.
2. Random Forest: An ensemble model for high accuracy and stability.
3. CatBoost: A state-of-the-art gradient boosting model that excels with categorical data.
* Validation: Stratified K-Fold Cross-Validation was used to ensure each fold maintained the original dataset's class distribution, which is crucial for imbalanced data.
* Optimization: The Optuna library with the Tree-structured Parzen Estimator (TPE) was used for efficient hyperparameter optimization based on Bayesian optimization.

Results:
The models were evaluated on Accuracy, ROC AUC, and F1-Score. The results showed that all three models, especially CatBoost, could effectively learn patterns from the data, providing a strong baseline for a predictive investor profiling tool.



Model Deep Dive: Predicting Interest in "Mergers and Acquisitions"

To provide actionable insights for Zenith, we performed a deep dive into our best-performing model: the one predicting investor interest in Mergers and Acquisitions. This analysis focuses on two key areas: understanding which factors drive the model's predictions (Feature Importance) and evaluating its real-world performance (Detailed Metrics).



The feature importance chart reveals the top 20 most influential factors the model uses to identify an investor's interest in M&A.

Key Observations:

Dominance of Industry Preference: An investor's preferred industry is the most powerful predictor. "Industries categories_STEM" is overwhelmingly the most important feature, with an importance score nearly double that of the next feature. This strongly suggests that investors focused on STEM are significantly more likely to be involved in M&A activities, likely reflecting the high volume of technology and biotech acquisitions in the market.

Key Geographic and Industry Signals: The top four features confirm the patterns seen in our EDA. After STEM, an investor's preference for the "United States" and "Healthcare" are the next most significant predictors, followed by a preference for "Europe". This creates a clear profile of a typical M&A-focused investor: one interested in high-value, mature sectors located in major Western markets.

A Blend of Factors: While industry dominates, the model effectively uses a combination of both geography and industry tags to refine its predictions. The presence of features like "Cultural and Creative Industries" and "Manufacturing" in the top 10 shows the model is capturing nuanced patterns beyond just the most obvious categories.

A closer look at the classification report confirms the model is not only effective but also robust and reliable.

Train vs. Test Performance:

train

Accuracy = 0.7438079870915691
F1 Score = 0.7666274711545528
ROC Area under Curve = 0.7544545967514704
Cohen's Kappa = 0.4880829498077005

	precision	recall	f1-score	support
0	0.64310	0.80766	0.71604	19829
1	0.84544	0.70125	0.76663	29751
accuracy			0.74381	49580
macro avg	0.74427	0.75445	0.74133	49580
weighted avg	0.76452	0.74381	0.74640	49580

Accuracy = 0.7371732817037754
F1 Score = 0.7601590106007067
ROC Area under Curve = 0.747799986922402
Cohen's Kappa = 0.47515712471824223

	precision	recall	f1-score	support
0	0.63641	0.80109	0.70931	4962
1	0.83951	0.69451	0.76016	7434
accuracy			0.73717	12396
macro avg	0.7379	0.74780	0.73474	12396
weighted avg	0.75821	0.73717	0.73981	12396

The minimal drop in performance between the training and test sets demonstrates that the model has generalized well and is not overfitting. This gives us high confidence that it will perform reliably on new, unseen investor data.

Challenges and Limitations

This project is hindered by several challenges. First, we experienced significant data quality issues such as incomplete or missing information. These issues limit effective assessments. Additionally, the complexity of the data makes preprocessing and cleaning difficult due to problems like incorrect quotations and misplaced entries. Scalability and performance are further concerns, as the dataset requires substantial computational resources. Moreover, the lack of proper context can lead to misleading conclusions, undermining the reliability of insights. Finally, time constraints prevent the development of predictive models and deeper analyses, compounding the difficulties faced in this area.

On the other hand, this project also faces several limitations. First, data consistency issues arise from the need for further standardization, as inconsistent data positions and formats complicate interpretation. Additionally, data granularity is insufficient, with categories like preferred geography (countries, continents and cities) grouped together, limiting detailed analysis. Ongoing handling of missing data further complicates the situation. The complex interrelationships among features make it challenging to capture meaningful connections in the analysis. Furthermore, the reliance on historical data may not accurately reflect future trends. This necessitates consistent monitoring. Lastly, computational constraints, including resource and time limitations, restrict the scale and depth of analysis.

Future Work

Several improvements are planned to further enhance the analysis. One of the key priorities is to complete and refine the dataset by addressing missing values and eliminating irrelevant data, ensuring that it is both comprehensive and reliable. Second, new features such as investment budget, risk aversion levels, and expected potential growth will be introduced to broaden the analytical scope and provide deeper insights into investor behavior.

Moreover, advanced machine learning techniques, including the implementation of models like XGBoost, will be further utilized to predict investment trends, preferences, and behaviors with greater precision. In addition, we would like to develop a personalized recommendation system to provide tailored insights and preferred investment options that align with individual investor profiles.

Furthermore, Natural Language Processing (NLP) techniques, such as Large Language Models (LLMs), will be employed to analyze investor descriptions and extract valuable context and insights. To ensure sustained performance, continuous optimization of the current machine learning models will remain a focus, enabling the system to adapt and improve over time.

Conclusion

We conducted an investment profile analysis using a 150,000-row dataset provided by Zenith. The analysis involved Exploratory Data Analysis (EDA), the development of an interactive Tableau dashboard, and the application of both supervised and unsupervised machine learning models.

The EDA revealed several significant insights into investment preferences:

- The United States emerged as the most preferred territory for investment, followed by Europe and Canada.
- Merger & Acquisitions was identified as the most popular investment type, surpassing venture capital.
- STEM industries were observed to be the top choice for investment, followed by manufacturing and healthcare.

To enhance data accessibility and visualization, an interactive dashboard was created using Tableau. This dashboard provided detailed insights into investment preferences by country, investment type, and industry type, offering a user-friendly interface for colleagues to explore the data.

Both unsupervised and supervised machine learning models were employed to extract deeper insights and make predictions:

- **Unsupervised Models**:
    - **Clustering** and **Association Rule Mining** were used to uncover hidden patterns within the data, focusing on preferred geography, investment types, industry, and investor profiles.
- **Supervised Models**:
    - Three supervised learning models were developed to predict investment preferences, such as predicting the preferred industry based on geography and investment type.
        - **Decision Tree**
        - **Random Forest**
        - **CatBoost**

These models provided actionable insights and helped in understanding the patterns and preferences of investors, enabling more informed decision-making.

Work Distribution

Even distribution among team members.