

HONG KONG

Theme: Big Data Analytics and Sustainable and Green Finance in Hong Kong
a green and sustainable finance hub in Asia



Team: The Greenist

CHIU Seen Yung

LEUNG Sum Nam

LIU Hoi Ying

WONG Cheuk Wah



Objective

Generate insights and inform responsible investment decisions in Hong Kong with big data research.



- Analyze datasets to provide insights about ESG risks or opportunities and how it affect investment choices.



- Find out the attitudes (negative, positive or neutral) of Hong Kong people about sustainability to roll out better green finance products.



- Help investors to create an investment portfolio that takes sustainability and profitability into account.

Utilize big data analytics and comprehend how different stakeholders respond to climate risks and green financing investment

Significance

Climate change is not just a local or investment issue. It is a human issue that affects us all.



Significance 1

The global risk caused by “**extreme weather events**” is **higher** than that caused by “**weapons of mass destruction**”.

Nearly **200** countries have signed the **Paris Agreement** at the 2015 United Nations Climate Change Conference to **combat climate change**.



Significance 2

The **financial sector** plays an **important** role in promoting **sustainable economic development**.

HK Financial Secretary **Paul Chan** has remarked the target of HK in pursuing **green finance hub** as well as growing **international interest** in green finance.



City / Climate Level

Task 1 – Manage Physical Climate Risk

This task aims to find out the correlation between the frequency of sustainability related keywords and the physical climate risks metrics. The government can then utilize the findings to promote sustainability with the right keywords.

Task 2 – Facilitate Green Finance

This task aims to find out the sentiment and attitude of Hong Kong citizens on social media towards posts about sustainability. Provide the government or business a picture and revise their effort in promoting ESG to the public.



Investment Level

Task 3 – Stock Cluster Analysis

This task will group stocks in S&P 500 index and Hang Seng Index according to their financial and ESG performance. Allow investors to build an investment portfolio considering both profitability and sustainability.







Task 4 – Stock Anomaly Analysis

This task will find anomalies in Hang Seng Index from 2017 to 2022 and summarize the events happened right before based on the provided news headlines. Help investors devise investment strategies and seize good market timing.



1

Task 1 – Manage Physical Climate Risk

 co2	 Annual Surface Temperature	 發展	 可持續	 ESG	 綠色
The amount of co2 in the unit of millions of metric tons from 2017-2021	Temperature change with respect to a baseline climatology, corresponding to the period 1951-1980, from 2017-2021, in terms of degree Celsius	A normalized number of the frequency of the word '發展' in the social big data across 2017-2021	A normalized number of the frequency of the word '可持續' in the social big data across 2017-2021	A normalized number of the frequency of the word 'ESG' in the social big data across 2017-2021	A normalized number of the frequency of the word '綠色' in the social big data across 2017-2021

2

Task 2 – Facilitate Green Finance

*'co2' and 'annual surface temperature' are two chosen *physical climate risk metrics*
 *2022 figure is **dropped** as the text are only until 30th September, which leads to *inaccurate comparison*



Sentiment
(binary variable)

- 1: when the text has the number of wow, haha, or love reaction larger than the mean → **Positive Sentiment**
- 0: other situations → **Negative Sentiment**



Cluster

Cluster divided through **k-means clustering**

3

Task 3 – Stock Cluster Analysis

- **494** stock tickers in total are extracted from **S&P 500** and **Hang Seng Index**.



Variables for Financial Performance

Beta

EV/Ebita Ratio

PE-Ratio

Earning Per Share (EPS)

Profit Margin



Variables for ESG Performance

Total ESG Score (total_esg)

Environment Risk Score (e_esg)

Governance Risk Score (g_esg)

Social Risk Score (s_esg)

Significant Controversy Level (highestcontroversy)

Outcome Variable → Cluster Category (Cluster)

**ESG Risk Ratings assess the degree to which a company's enterprise business value is at risk driven by environmental, social and governance issues. The rating employs a two-dimensional framework that combines an assessment of a company's exposure to industry-specific material ESG issues with an assessment of how well the company is managing those issues. The final ESG Risk Ratings scores are a measure of unmanaged risk on an absolute scale of 0-100(range), with a lower score signaling less unmanaged ESG Risk.*

4

Task 4 – Stock Anomaly Analysis

**Controversies Research identifies companies involved in incidents and events that may negatively impact stakeholders, the environment or the company's operations. Controversies are rated on a scale from one to five with five denoting the most serious controversies with the largest potential impact.*



From Social Big Data (Total 321946 records)

- Pubtype == U → Only extract news data
- Date - Headline
- Keywords (Derived from headlines)

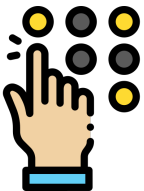
From Hang Seng Index Historical Monthly Data (Total 69 monthly records)

- Date
- Close Price (Close)
- **Outcome Variable**: Anomaly (1 = Yes; 0 = No)

Task 1 – Manage Physical Climate Risk

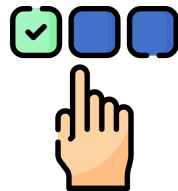
Counting

- Count keywords frequency in the social big data by year



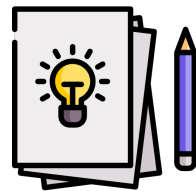
Select

- Select top sustainability related words through visualization



Create

- Create data frame with the normalized word frequency and the two physical climate risk attributes across 2017-2021

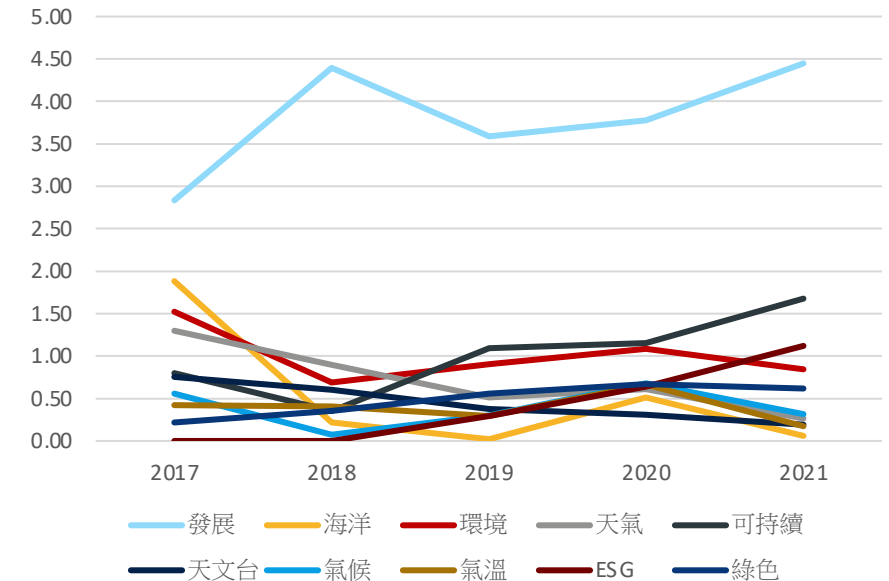


Correlation

- Find the correlation between the frequency of sustainability related keywords and the two physical climate risk metrics



% of the words appeared in total of the year



Top four sustainability related words – “發展, 可持續, ESG, 綠色” were chosen

Task 2 – Facilitate Green Finance



1. Select

Select texts from social platform only



2. Create

- Create new binary variable for all text → sentiment
- 1: when the text has the number of wow, haha, or love reaction larger than the mean → **positive sentiment**
 - 0: other situation → **neutral** or **negative sentiment**



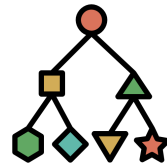
3. Tokenize

Tokenize the texts and fit the Count Vectorize model



4. Fit

Fit in the K-means clustering model with $K = 4$



5. Divide

Text is divided into 4 clusters



6. Random

Randomly select 10-20 texts from each cluster to define the topic of the cluster



7. Compare

Compare % of sentiment = 1 and sentiment = 0 across clusters

Task 3 – Stock Cluster Analysis

1. Data Pre-processing

1. Collect stock symbols from Wikipedia by web scraping.
2. Collect financial and ESG info for each ticker using yahoo finance library.
3. Explore the relationships among attributes with Correlation Matrix.
4. For the attributes with low correlation → plot a scatterplot and identify the outlier that is deviated from the general pattern.
5. Standardize the variables using **StandardScaler** → remove the mean and scales each feature/variable to unit variance.

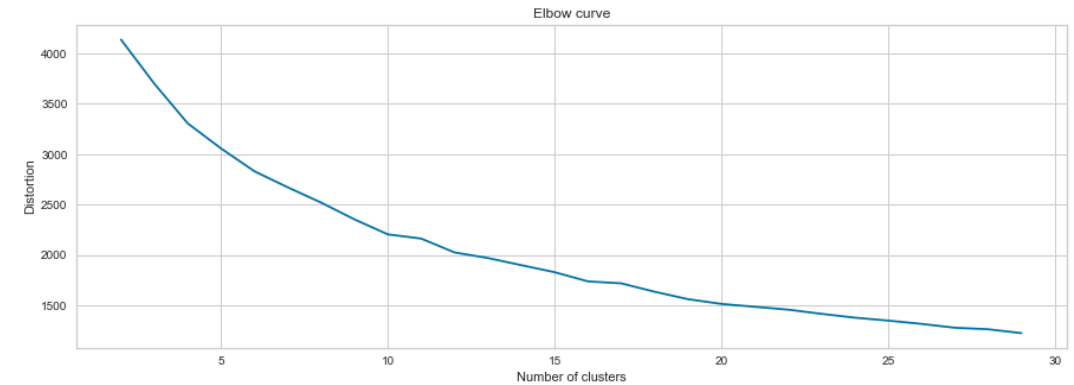
2. Hyperparameter Tuning

1. Plot an **Elbow Curve** to choose the possible range of the number of clusters (2-10).
2. Since the 'elbow' point is not very significant → Confirm + determine the optimal number of clusters with **Silhouette Visualizer** using the silhouette metric for cluster evaluation.

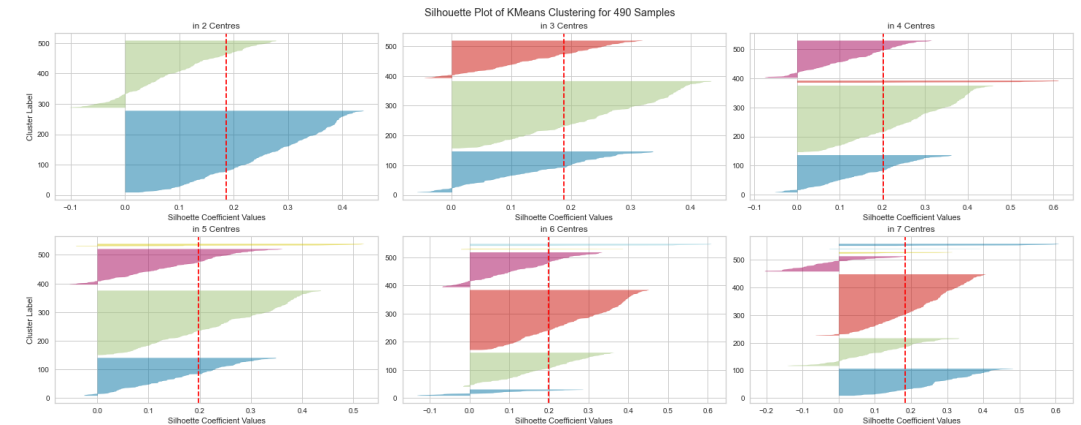
3. Model Fitting

1. Fit in the K-means clustering model with the chosen optimal K = XXXX (depend on how many clusters are chosen).

Elbow curve to locate cluster range



Silhouette plot to select optimal cluster



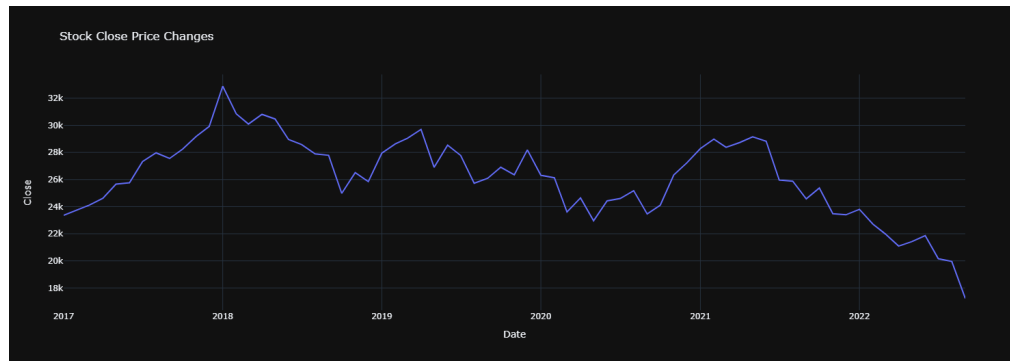
Task 4 – Stock Anomaly Analysis

1. Text Pre-processing

1. Remove special characters and English characters.
2. Extract keywords based on the **TF-IDF score** using Jieba Chinese text segmentation so to evaluate how relevant a word to a document in a collection of document.

3. Anomaly Detection Using ARIMA Model

1. Select the optimal order for the ARIMA model using Akaike's Information Criterion.
2. The minimum AIC score determines the optimal order while d (the number of non-seasonal differences needed for stationarity) has been determined previously.
3. Fit the ARIMA model to the preprocessed data.
4. Compute the Squared Error for each observation.
5. Compute the threshold for the errors in the data → Formula: $\text{threshold} = \text{mean}(\text{squared_errors}) + (z * \text{standard_deviation}(\text{squared_errors}))$.
6. If the errors > the threshold → the observation is flagged as an anomaly (= 1).

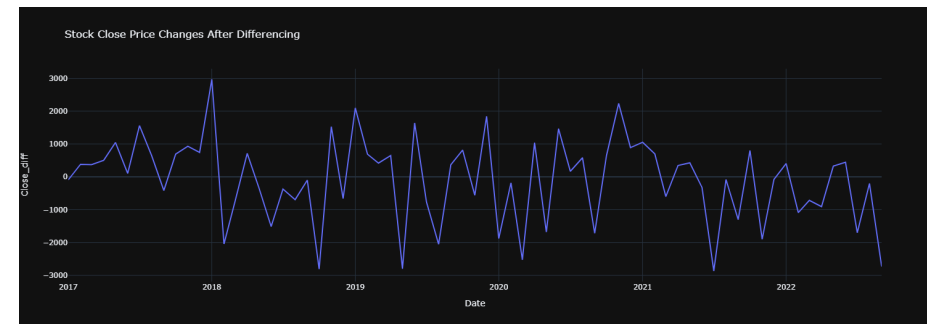


2. Time Series Data Pre-processing

1. Augmented Dickey-Fuller unit root test is conducted to test whether the close price over time is stationary or not.
2. Use first order differencing method to transform the data to stationary.

4. Anomaly and Associated Events Analysis using Topic Modelling

1. Among the news headlines, partition the headlines within the month before each anomaly into new datasets:
 - If anomaly happened in Jan 2018, a new dataset within 2017-12-01 and 2017-12-31 will be created.
2. For each newly created dataset, use TF-IDF Vectorizer to put the extracted keywords into a vector by considering the importance of each word in a headline, which is scaled by its importance across all the headlines in the corpus.
3. Decide the optimal number of component with GridSearchCV → which is 2.
4. Perform Topic Modelling: For each anomaly dataset, divide the headlines into two topics (weather % social-economics focus) using Latent Dirichlet Allocation.
5. Use pyLDAvis to interpret the topics → extract the top 30 most relevant terms for each topic (which are considered the events associated with a particular anomaly).



Task 1 – Manage Physical Climate Risk



Results – Correlation Matrix Table

Original table with raw data	co2	Annual Surface Temperature Change	發展	可持續	ESG	綠色
2017	5.85	1.389	2.83	0.8	0	0.22
2018	5.79	1.234	4.4	0.35	0	0.35
2019	5.68	1.827	3.59	1.09	0.29	0.56
2020	4.59	1.901	3.78	1.15	0.64	0.67
2021	5.4775	2.038	4.45	1.68	1.12	0.62

*co2 means the co2 emission over the year

*annual surface temperature change means its change over the years

*others are based on the percentage of the keyword frequency over the total word frequency over the year

Correlation Matrix Table

	發展	可持續	ESG	綠色
co2	-0.14475	-0.3893	-0.51798	-0.73385
Annual Surface Temperature Change	0.229853	0.940526	0.891166	0.89373

From the correlation matrix table, we can observe a **negative correlation** between the climate risk metric “co2” with the chosen four sustainability-related words, while a **positive correlation** for “annual surface temperature change”.



Implications

The increase in the frequency of sustainability-related keywords between 2017 and 2021 has shown that there is an increase in awareness of sustainability among the public and an increase in promotion of ESG from the government. And this can be one explanation of the decrease in physical climate risk - "CO2 emissions".



Suggested solutions – Promoting with the right keywords

Although "co2 emission" has decreased, "annual surface temperature change" has increased over the years. Therefore, the Hong Kong government can try to find if "annual surface temperature change" as well as other physical climate risks have any negative correlation with other sustainability-related words. Hence, increase the awareness of the relative physical climate risk with the right keywords.

Task 2 – Facilitate Green Finance



Results – Text Clustering

Cluster	Sentiment	# of texts	Total # of texts	% of the cluster
0 (ESG related)	0	130,552	154,985	84%
	1	24,433		16%
1 (entertainment)	0	450	450	100%
	1 -			0
2 (political/ other news)	0	3,178	3,952	80%
	1	774		20%
3 (online tutorials)	0	607	614	99%
	1	7		1%



From the text clustering table besides, we can observe that only a small portion of Hong Kong people is talking about sustainability in a positive way (16%).



Implications

*Sentiment = 1 – When the # of Wow, Love, or Haha reaction is larger than mean (positive)

*Sentiment = 0 – Otherwise (neutral or negative)

This result may imply that the general public believes the current effort devoted by the government or corporations to sustainability is not enough. As a result, they may not have sufficient knowledge on ESG or have suspicions about green finance, causing hesitation when support is needed.



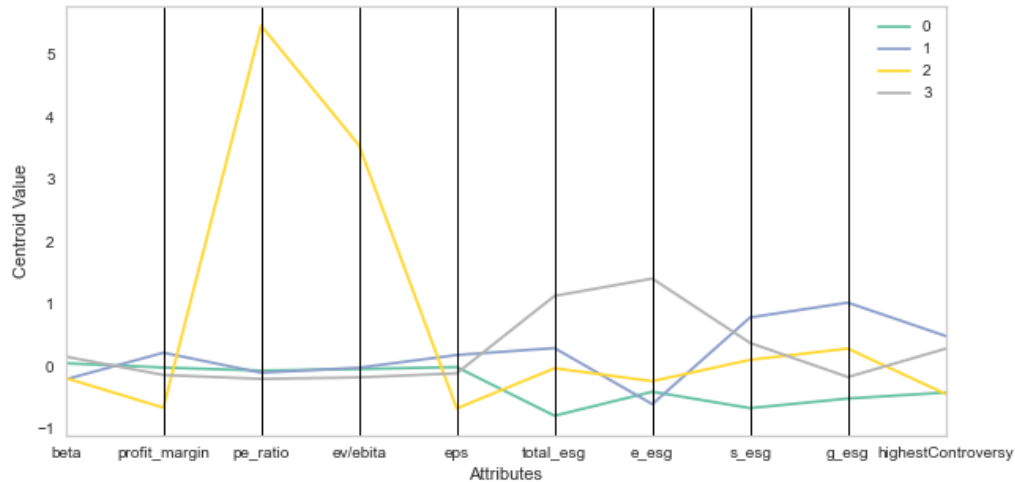
Suggested solutions – Revise Effort

The government or corporations can regularly carry out testing on how Hong Kong residents feel and behave towards messages regarding sustainability on social media. Since the current majority of Hong Kong residents' feelings are relatively negative. The government can try to improve and expand on the current public education campaign on ESG so that the public can have more knowledge and possibly a more positive attitude toward different green finance schemes.

Task 3 – Stock Cluster Analysis



Results – Parallel Coordinates for 4 Clusters



Except cluster 2 which is relatively not attractive due to its expensiveness in terms of valuation (highest P/E ratio), all clusters have similar financial performance but various risk scores in terms of ESG.

Cluster 0 is the most attractive due to its lowest total ESG risk score overall and second highest profit margin.



Implications

Without a doubt, a company's objective is to maximize profit, and traditionally, it will focus on achieving "beautiful" financial performance. However, with rising environmental and global concerns, it is critical for businesses to take responsibility for their global footprint, which can indeed help attract more investors.



Suggested solutions – Portfolio Design

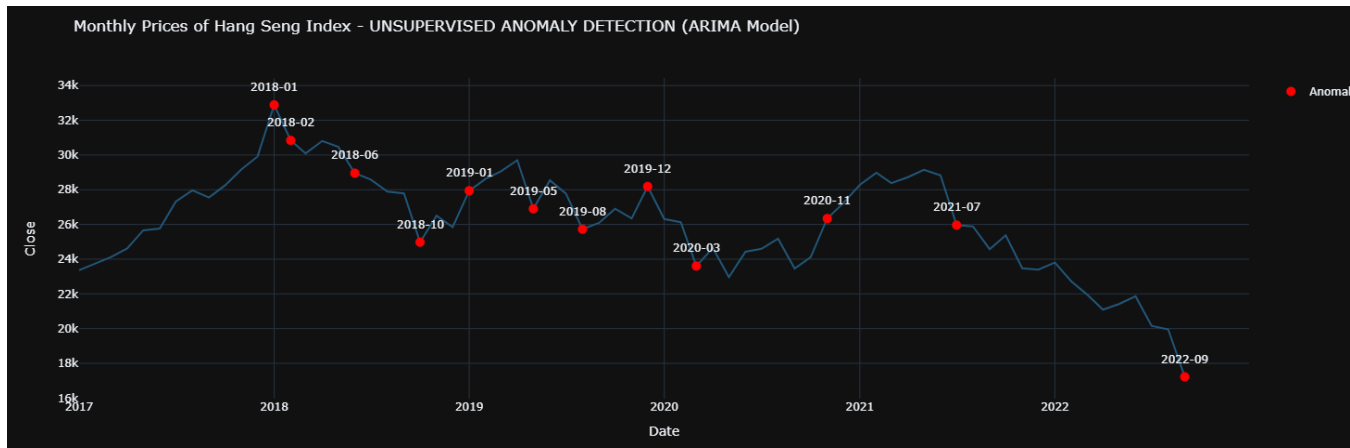
A well-diversified portfolio is vital to any investor's success. Traditionally, sector, market cap, and profitability are the major factors to consider when picking stocks that satisfy the corresponding level of risk. With rising ESG concerns among investors, it is necessary for investors to take sustainability into account when designing their portfolio. Hence, investors can refer to our cluster analysis and pick the group with the lowest risk score.

Task 4 – Stock Anomaly Analysis



Results – Anomaly Detection

	Positive Anomaly		Negative Anomaly
Topic 1 - Socio Economics	['發展', '中國', '香港', '合作', '經濟', '持續', '習近平', '國際', '企業', '印尼', '全球', '海嘯', '創新', '綠色', '環保', '增至', '世界', '論壇', '投資', '深圳', '金融', '科技', '國家', '項目', '舉行', '推動', '市場', '美國', '社會', '地震', '傾瀉', '風暴']	Topic 1 - Socio Economics	['山竹', '颱風', '日本', '中國', '香港', '機場', '航班', '取消', '關西', '政府', '影響', '風災', '襲港', '服務', '市民', '暴雨', '居民', '北海道', '吹襲', '強颱風', '經濟', '停課', '美國', '中心', '地震', '交通', '大阪', '恢復', '傾瀉', '風暴']
Topic 2 - Weather	['天文台', '颱風', '風球', '天氣', '香港', '信號', '沙德爾', '三號', '浪卡', '熱帶', '年度', '日本', '風暴', '氣旋', '澳門', '強風', '改發', '考慮', '失蹤', '明日', '新聞', '天晴', '生效', '八號', '本港', '維持', '乾燥', '威尼斯', '暴雨', '氣溫']	Topic 2 - Weather	['天文台', '天氣', '山竹', '颱風', '驟雨', '警告', '雷暴', '酷熱', '暴雨', '風球', '信號', '香港', '最高', '持續', '生效', '發展', '本港', '今年', '氣溫', '未來', '熱帶', '死亡', '澳門', '下午', '今日', '明日', '狂風', '改發', '多雲', '襲港']



Six negative anomalies and six positive anomalies are detected.



Implications

The keywords are the top 30 most relevant terms associated with the anomalies among the Hang Seng Index stock prices.

Positive words such as '合作'、'綠色'、'環保'、'持續' are shown to boost the stock price and climate-related terms such as '颱風'、'暴雨'、'機場' are more likely to affect the socio-economic condition in society and lower the price.

However, the result also proves there is no significant relationship between weather and the monthly stock price.



Suggested solutions – Investment Timing

Timing the market is often a key component of actively managed investment strategies, and it is always a basic strategy for traders.

In this case, investors and traders may use this stock anomaly detection as a reference point when deciding when to invest, possibly enabling more accurate market timing.

What are the Limitations and How can be Done Better?

Limitations and Improvement



Limitations

1 City / Climate Level Task 1 – Manage Physical Climate Risk

- Lacked physical climate risk data across 2017-2022. Our team have found other datasets online, but they only covered other years before 2017 while the social text dataset covered 2017-2022.

2 City / Climate Level Task 2 – Facilitate Green Finance

- Limited testing on different number of clusters to figure out which cluster is the best, so the result may not be fully comparable and comprehensive.

3 Investment Level Task 3 – Stock Cluster Analysis

- Lacked a comprehensive overview on ESG scores, the ESG scores are only based on one main ESG rating organization which is Sustainalytics.

4 Investment Level Task 4 – Stock Anomaly Analysis

- Lacked an all-inclusive prediction on future stock price or anomalies, we only identified unexpected or unusual stock price in the past.



Improvement

1 City / Climate Level Task 1 – Manage Physical Climate Risk

- If we could have more physical climate risk data covering 2017-2022, we could have insights and increase accuracy in finding the correlation between the physical climate risk metrics and the word frequency.

2 City / Climate Level Task 2 – Facilitate Green Finance

- If we could repeat with different number of clusters, i.e. less than or more than 4 clusters, we could give an accurate result and decide the most ideal cluster.

3 Investment Level Task 3 – Stock Cluster Analysis

- If we could be granted with permission in the future, we could adopt other credible ESG rating organizations, like MSCI ESG Ratings.

4 Investment Level Task 4 – Stock Anomaly Analysis

- If we could give a prediction to stock price or anomalies in the future, we could inform investors more accurate investment strategies and take advantage of favorable market conditions.

What are the key takeaways?

Conclusions



City / Climate Level

1

Task 1 – Manage Physical Climate Risk

- Analyzed CO₂ and Annual Surface Temperature datasets
- Determined the relationship between the frequency of sustainability related keywords and the physical climate risk metrics.
- Enabled the government to release more accurate information about physical climate risks in public with the suitable keywords.

2

Task 2 – Facilitate Green Finance

- Concluded that how Hong Kong residents feel and behave towards messages regarding sustainability on social media.
- Reminded the government or corporations to reevaluate their public education campaign on ESG before launching green finance schemes.



Investment Level

3

Task 3 – Stock Cluster Analysis

- Carried out a comprehensive stock cluster analysis, in which stocks in S&P 500 index and Hang Seng Index will be grouped based on their ESG and financial performance.
- Permitted investors to create an investment portfolio that takes sustainability and profitability into account.

4

Task 4 – Stock Anomaly Analysis

- Investigated anomalies in Hang Seng Index between 2017 and 2022, followed by a summary of recent occurrences based on the specified news stories.
- Assisted investors in creating investment plans and identifying optimal market timing.