QF605 Financial Data Science

Amanda Goh Hui Wen, Cheok De Jun, Richard Wang Zhi Chen, Tan Si Yin Jodie Ethelda, Yap Kian Tiong

PROJECT I

An Overview

Task	Data Cleaning	Data Analysis
• An external research firm has developed a proprietary system that forecasts the prices of ETF • To review quality of data and propose corrected values • To analyse signal's effectiveness in forecasting ETF prices	• Ensuring prices in date are in correct formats • Dropping duplicate rows • Adjusting data for rows with negative prices • Handling outliers	• Analysis of adjusted closing prices using Time Series Decomposition and ACF • Analysis of relationship between signal changes and ETF returns • Analysis of signal using MAE

PROJECT I

Understanding the Data

Signal column are predictions generated by the proposed system

Open, High, Low, Close & Adj Close columns are historical prices for a well-known broad market ETF

Date	Signal	Open	High	Low	Close	Adj Close
11/19/2015	13.76854	116.44	116.65	115.74	116.06	108.2816
11/20/2015	13.608819	116.48	117.36	116.38	116.81	108.98132
11/23/2015	12.990589	116.71	117.89	116.68	117.39	109.52245
11/24/2015	12.667435	116.88	118.42	116.56	118.25	110.32484
11/25/2015	13.01991	118.3	119.32	118.11	119.17	111.18316
11/27/2015	12.879819	119.27	119.9	118.88	119.62	111.603
11/30/2015	13.184791	120.02	120.07	119.05	119.1	111.11785
12/01/2015	12.922631	119.61	119.91	118.9	119.89	111.8549
12/02/2015	13.118076	119.73	120.04	118.45	118.6	110.65137
12/03/2015	12.91654	118.93	119.45	116.14	116.6	108.78541
12/04/2015	13.155278	116.6	117.93	116.34	117.78	109.88632
12/07/2015	12.430221	117.64	117.67	115.51	116.01	108.23495
12/08/2015	13.030335	114.96	116.01	114.46	115.37	107.63783
12/09/2015	12.439604	115.03	116.17	113.6	114.08	106.43429
12/10/2015	13.249671	113.89	115.08	113.65	114.46	106.78884
12/11/2015	13.067526	113.27	113.34	111.53	111.91	104.40974
12/14/2015	13.054955	111.87	112.38	110.28	111.11	103.66337
12/15/2015	12.231405	111.73	112.96	111.6	112.71	105.15612
12/16/2015	13.078074	113.49	114.65	112.83	114.43	106.76083

PROJECT I

Data Cleaning

Check Data Formats

- Ensuring prices are in float64 format, and date is in datetime format
- Checked that date format and sequence are correct

Cleaning Data

- Dropping duplicate rows
- Forward filling NaN values

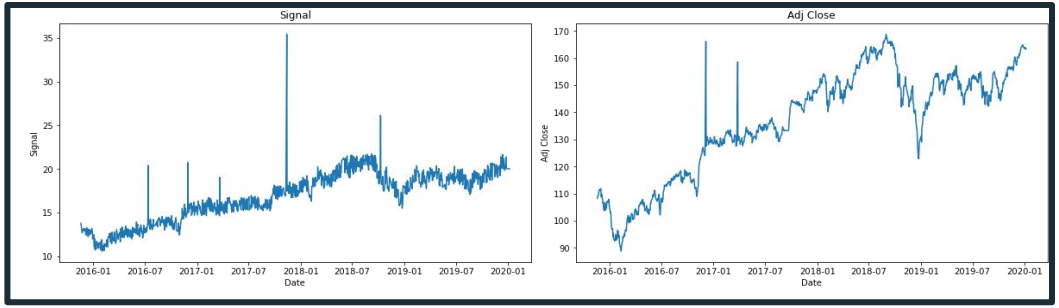
Correcting Negative or Zero Prices

- Set negative prices to NaN
- Forward fill

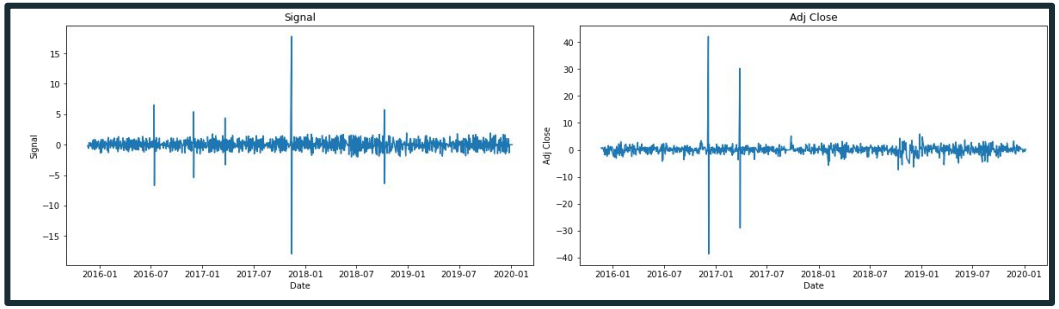
Difference the Data

- Period = 1

Before differencing (1038, 7)



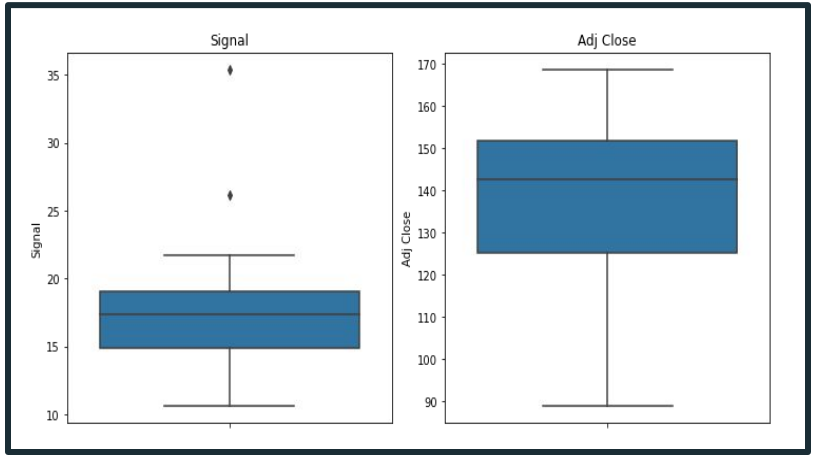
After differencing (1037, 7)



PROJECT I

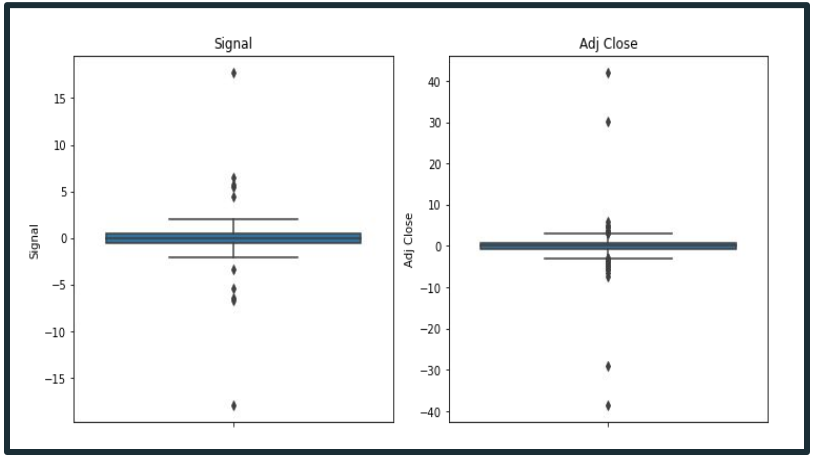
Data Cleaning (Handling Outliers)

Before differencing



- Seemingly no outliers

After differencing



- Returns seem to be normally distributed with a mean of 0
- The kurtosis of signal and adj close returns of ETF are significantly higher

PROJECT I

Data Cleaning (Handling Outliers)

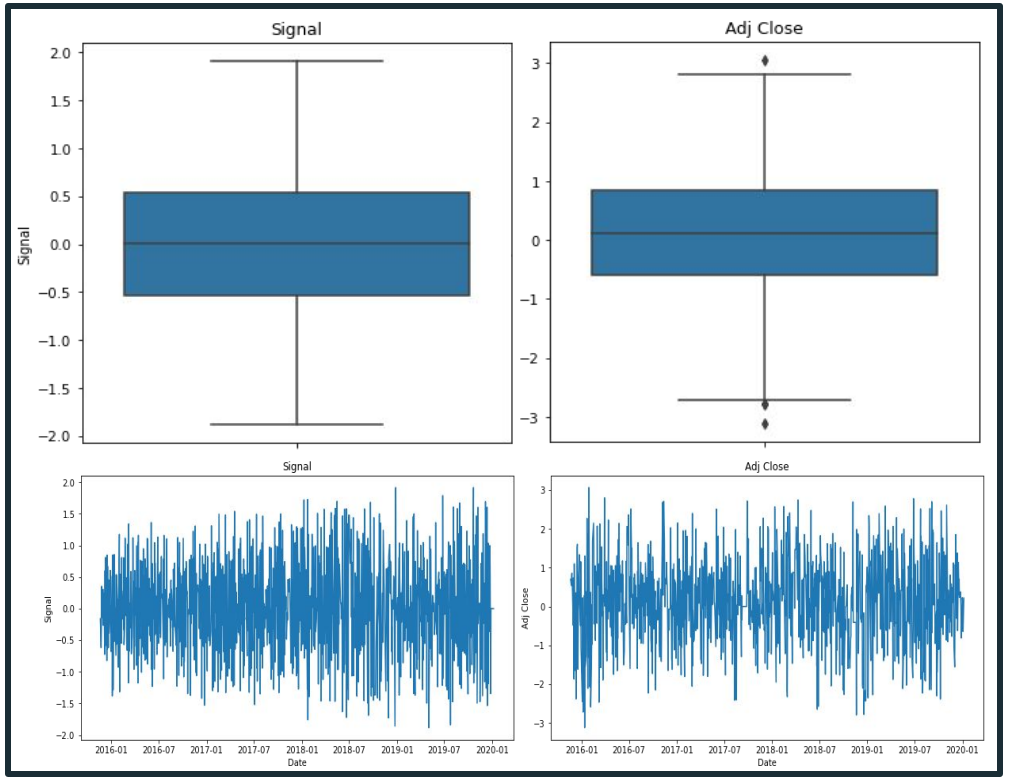
Outlier Detection

- Start at one-tenth point of the data
- Get the rolling mean & std of differenced data
- Check if “today’s” data point is > mean +/- threshold * std
  - If yes, identify as outlier

Outlier Handling

- Assign outliers as NaN values
- Forward fill numbers
  - This would assume the previous day’s trend
  - If fill NA with 0s, we will assume no trend for outlier days

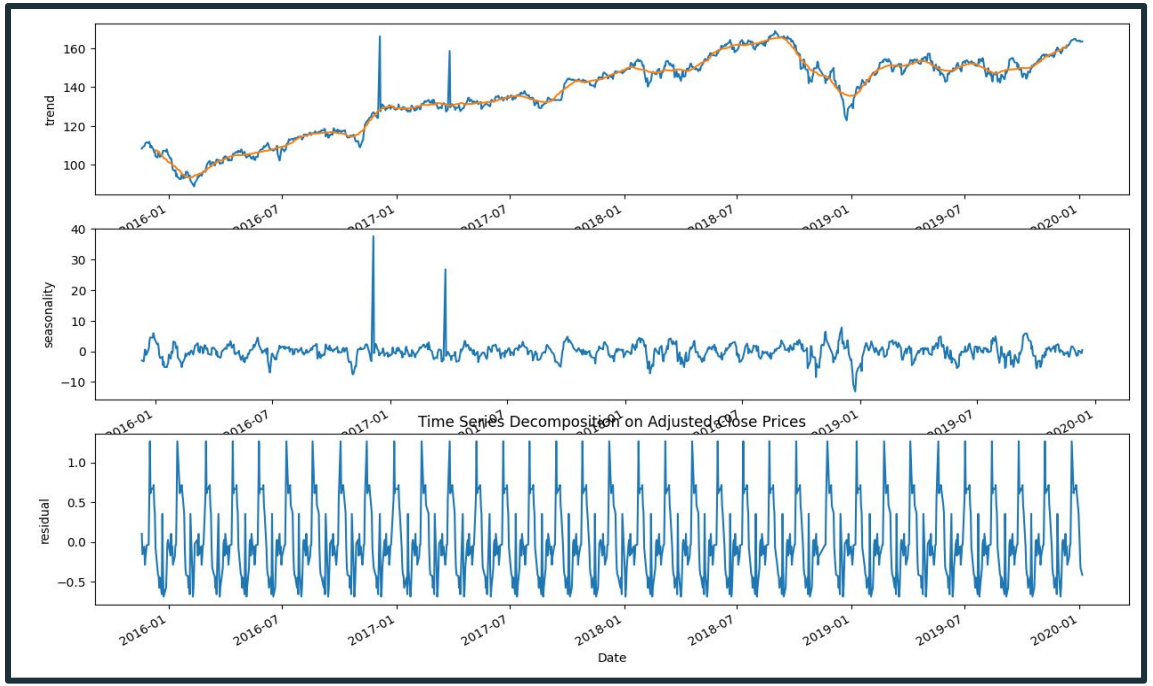
Box Plots and Differenced Data After Outlier Handling



PROJECT I

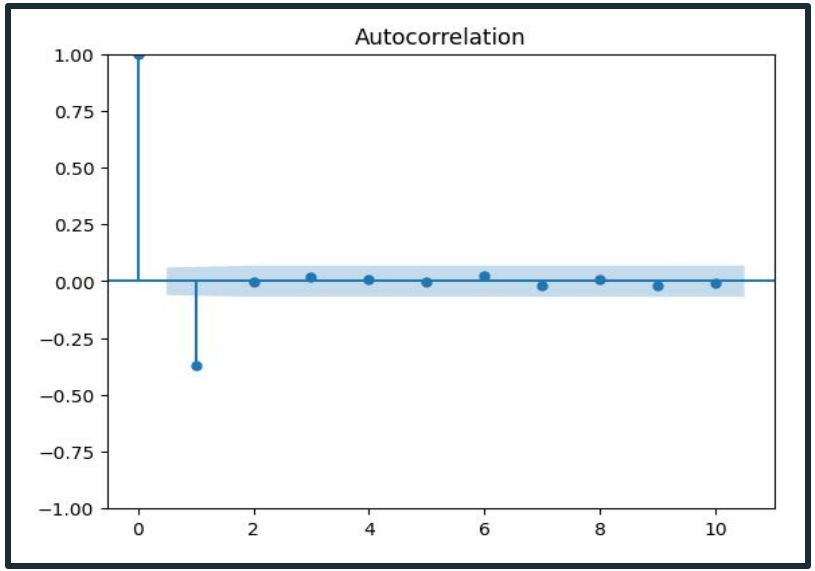
Time Series Decomposition

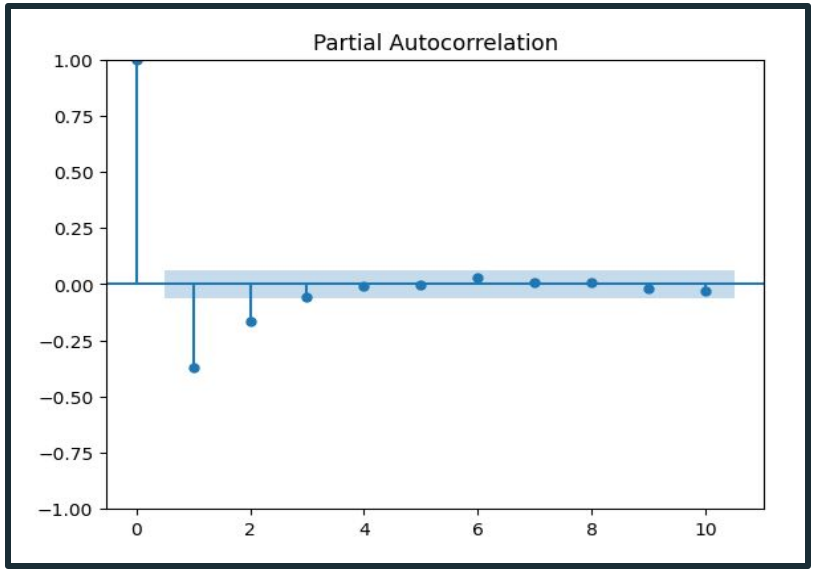
- Using additive method as we see constant variance
- Upward trend till end 2018
- Does not have clear seasonality after decomposition
- Residuals show clear pattern
- There is still time-series information



PROJECT I

ACF/PACF of Adjusted Close





- ACF is significant at lag 1, whereas PACF is shows a geometric decay at each m lag
- This can signify a moving average (MA) component in an ARIMA model if used

PROJECT I

Cross-Correlation

Generally low correlation between Signal changes and ETF returns

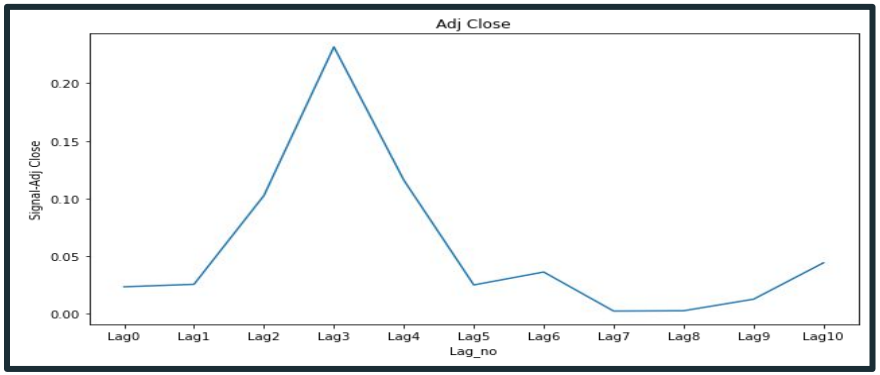
Before Outlier Handling

- Highest correlation between Signal change and Adj Close ETF returns at lag 3

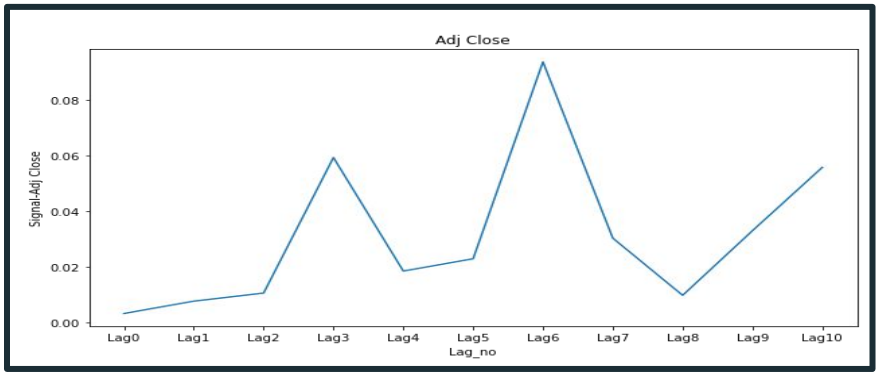
After Outlier Handling

- Highest correlation between Signal change and Adj Close ETF returns at lag 6

Before Outlier Handling



After Outlier Handling



PROJECT I

Analysis of Signal's Effectiveness in Forecasting ETF Price

What is considered "good"?	• Evaluating against Baseline: Naive • Taking the lag of price as next horizon forecast, essentially forecasting change in price to be 0.
Metrics for Comparison	• Mean absolute error • % of time signal beats Naive: No. of times signal beats naive / total no. of rows ○ Outlier prediction from Signal can drive up MAE unfairly
Test Set	• 2016 onwards • 1009 points

PROJECT I

Analysis of Signal's Effectiveness in Forecasting ETF Returns

- MAE of Signal vs Naive (Std)
- Signal unable to beat Naive by a large margin

Prices	Naive	Signal	Signal at lag 3	Signal (Outlier Handling)	Signal at lag 6 (Outlier Handling)
Open	1.07 (0.00)	1.35 (1.28)	1.32 (1.28)	1.27 (0.80)	1.26 (0.81)
High	0.92 (0.00)	1.25 (1.28)	1.16 (1.28)	1.17 (0.80)	1.11 (0.81)
Low	1.03 (0.00)	1.31 (1.28)	1.29 (1.28)	1.25 (0.80)	1.17 (0.81)
Close	1.18 (0.00)	1.45 (1.28)	1.43 (1.28)	1.37 (0.80)	1.32 (0.81)
Adj Close	1.12 (0.00)	1.40 (1.28)	1.37 (1.28)	1.32 (0.80)	1.27 (0.81)

- % of time Signal beats Naive
- Signal unable to beat Naive by a large margin

Prices	Signal	Signal at lag 3	Signal (Outlier Handling)	Signal at lag 6 (Outlier Handling)
Open	0.38	0.41	0.38	0.40
High	0.34	0.42	0.35	0.41
Low	0.36	0.41	0.36	0.43
Close	0.36	0.42	0.36	0.42
Adj Close	0.36	0.41	0.36	0.42

PROJECT I

Analysis of Signal's Effectiveness in Forecasting ETF Adj Close Returns by Year

- MAE of Signal vs Naive (Std) per year
- Signal unable to beat Naive by a large margin

Year	Naive	Signal	Signal at lag 3	Signal (Outlier Handling)	Signal at lag 6 (Outlier Handling)	Sample Count
2016	1.243047	1.496629	1.328046	1.408498	1.327340	252
2017	0.990522	1.425036	1.288485	1.257901	1.209483	252
2018	1.238025	1.475460	1.529969	1.399995	1.416818	250
2019	1.146655	1.322327	1.295795	1.322327	1.172815	252
2020	0.323080	0.323080	0.323080	0.323080	1.199989	3

- % of time Signal beats Naive per year
- Signal unable to beat Naive by a large margin

Year	Naive	Signal	Signal at lag 3	Signal (Outlier Handling)	Signal at lag 6 (Outlier Handling)	Sample Count
2016	0.00	0.39	0.45	0.40	0.46	252
2017	0.00	0.31	0.41	0.31	0.35	252
2018	0.00	0.39	0.41	0.40	0.40	250
2019	0.00	0.38	0.42	0.38	0.50	252
2020	0.00	0.00	0.00	0.00	0.00	3

PROJECT I

Conclusion

Efficacy of Product	Recommendation

• Product does not forecast ETF prices better than naive across all time periods
• Efficacy of the product can be greatly improved in terms of its predictive quality, as the value of its prediction hinges on beating the baseline
	• Accounting for cross correlation during analysis
	• Compare against naive when modelling
	• Identify specific characteristics of the ETF to forecast (e.g., sector-specific ETFs, leveraged ETFs) as these can influence their price dynamics
	• Research and gather more uncorrelated features that can forecast ETF prices
	  ○ i.e. Macroeconomic Indicators: interest rates, GDP growth
	  ○ News Sentiment: Sentiment analysis of news related to the ETF or its underlying assets if applicable

PROJECT II An Overview

Task	Data Cleaning/Scraping	Data Analysis

• Merge given data and data harvested from h1bdata.info
• Run through exploratory data analysis
• Pick out trends and insights from the dataset, such as:
  ○ Salary trends
  ○ Biggest hirer of H1B Visa etc.
	• Used BeautifulSoup to scrape data from h1bdata.info
	• Merged with existing sample data (total 5 sets)
	• Resolved conflicts in rows
	  ○ “L” vs “Large”
	  ○ Salary in different units
	• Standardised the datasets
	  ○ Column types, categorical values for rows
	• Cleaned data
	  ○ Drop rows with missing salary data
		• Analysis of salary over the years by job type and experience
		• Analysis of job demand by state (in US only)
		• Analysis of job demand by employer size
		• Analysis of job demand by experience level

PROJECT II

Data Cleaning/Scraping

For Sample Data

- Organised jobs as TYPES according to job title
- Cleaned data, dropping rows with missing salary value
- Standardised variables: COMPANY & EMPLOYEE LOCATIONS & STATES (for US only), EMPLOYER SIZE

For Scraped Data

- Scraped data using BeautifulSoup based on job TYPES
- Standardise salary into annual salary

Merging Data

- Keeping variables YEAR, EMPLOYER, JOB TITLE, TYPE, COMPANY & EMPLOYEE LOCATIONS & STATES, EMPLOYER SIZE, y (salary in '000s), EXP, EMPLOYMENT TYPE, WORK SETTING, REMOTE RATIO, WORK MODELS
- Replace NaN values with UNKNOWN for relevant columns (i.e., EMPLOYEE_STATE if EMPLOYEE_LOCATION is known)

	YEAR	EMPLOYER	JOB TITLE	TYPE	COMPANY LOCATION	COMPANY STATE	EMPLOYEE LOCATION	EMPLOYEE STATE	EMPLOYER SIZE	y	EXP
0	2023	NaN	DATA DEVOPS ENGINEER	DATA ENGINEER	GERMANY	NaN	GERMANY	NaN	L	95.01	MID
1	2023	NaN	DATA ARCHITECT	DATA ARCHITECT	UNITED STATES	NaN	UNITED STATES	NaN	M	186.00	SENIOR
2	2023	NaN	DATA ARCHITECT	DATA ARCHITECT	UNITED STATES	NaN	UNITED STATES	NaN	M	81.80	SENIOR
3	2023	NaN	DATA SCIENTIST	DATA SCIENTIST	UNITED STATES	NaN	UNITED STATES	NaN	M	212.00	SENIOR
4	2023	NaN	DATA SCIENTIST	DATA SCIENTIST	UNITED STATES	NaN	UNITED STATES	NaN	M	93.30	SENIOR

	YEAR	EMPLOYER	JOB TITLE	TYPE	COMPANY LOCATION	COMPANY STATE	EMPLOYEE LOCATION	EMPLOYEE STATE	EMPLOYER SIZE	y
0	2023	UNKNOWN	DATA DEVOPS ENGINEER	DATA ENGINEER	GERMANY	UNKNOWN	GERMANY	UNKNOWN	L	95.01
1	2023	UNKNOWN	DATA ARCHITECT	DATA ARCHITECT	UNITED STATES	UNKNOWN	UNITED STATES	UNKNOWN	M	186.00
2	2023	UNKNOWN	DATA ARCHITECT	DATA ARCHITECT	UNITED STATES	UNKNOWN	UNITED STATES	UNKNOWN	M	81.80
3	2023	UNKNOWN	DATA SCIENTIST	DATA SCIENTIST	UNITED STATES	UNKNOWN	UNITED STATES	UNKNOWN	M	212.00
4	2023	UNKNOWN	DATA SCIENTIST	DATA SCIENTIST	UNITED STATES	UNKNOWN	UNITED STATES	UNKNOWN	M	93.30
...	...	...	...	...	...	...	...	...	...	...
101667	2018	LDISCOVERY LLC	DATA MANAGEMENT TEAM LEAD	DATA MANAGEMENT	UNITED STATES	VA	UNITED STATES	VA	M	104.40

PROJECT II

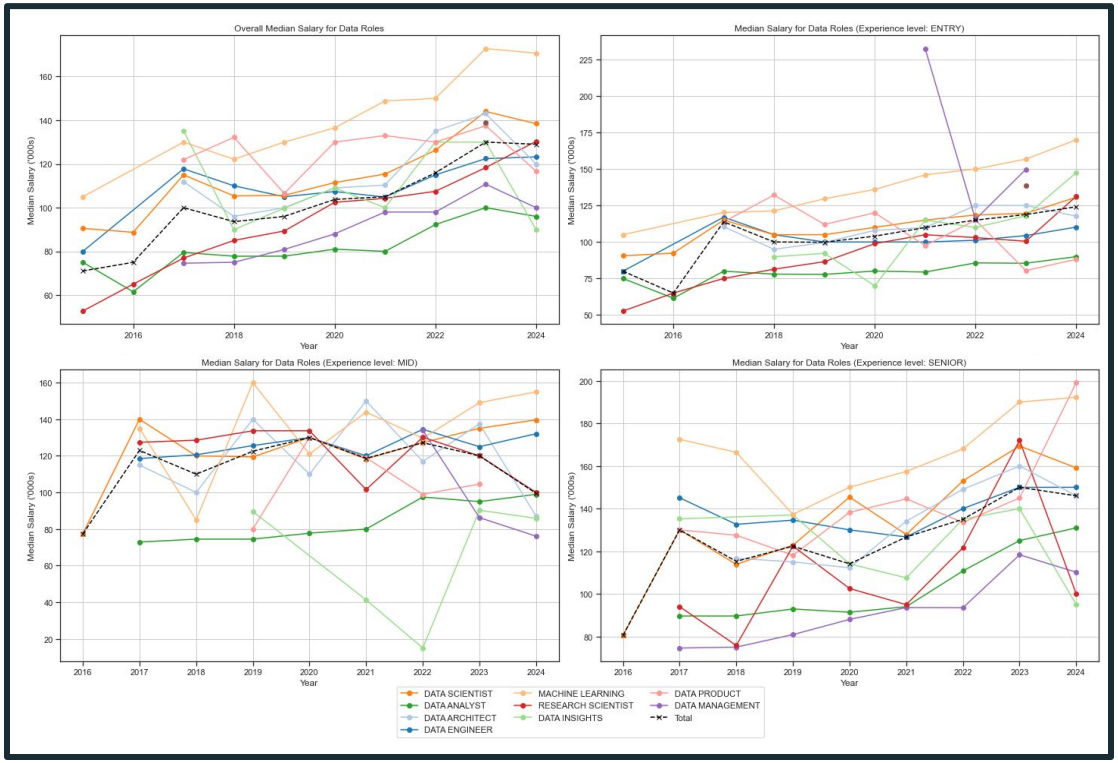
Analysing Median Salary

Overall, median salaries have increased across time for all roles

- Most prominent for machine learning roles, which also has the highest median salaries across most years
- Interestingly, in 2024, median salaries saw a drop for all data roles except research scientists

Increasing salary trend is most observable for entry and senior level roles

- Median salaries tend to fluctuate more across time for higher experience levels
- Machine learning roles generally tend to hold the highest median salaries at every experience level

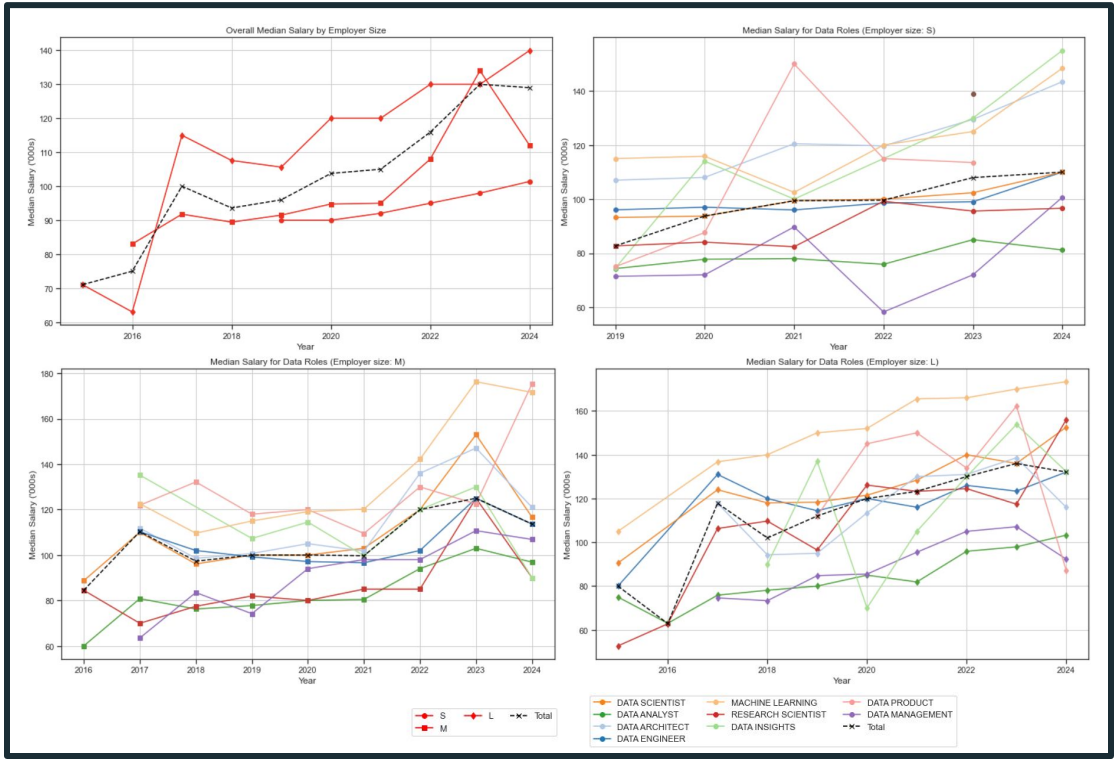


PROJECT II

Analysing Median Salary

As expected, large companies tend to offer the highest salaries, followed by medium and small companies

Across all companies, machine learning roles tend to hold one of the highest median salaries



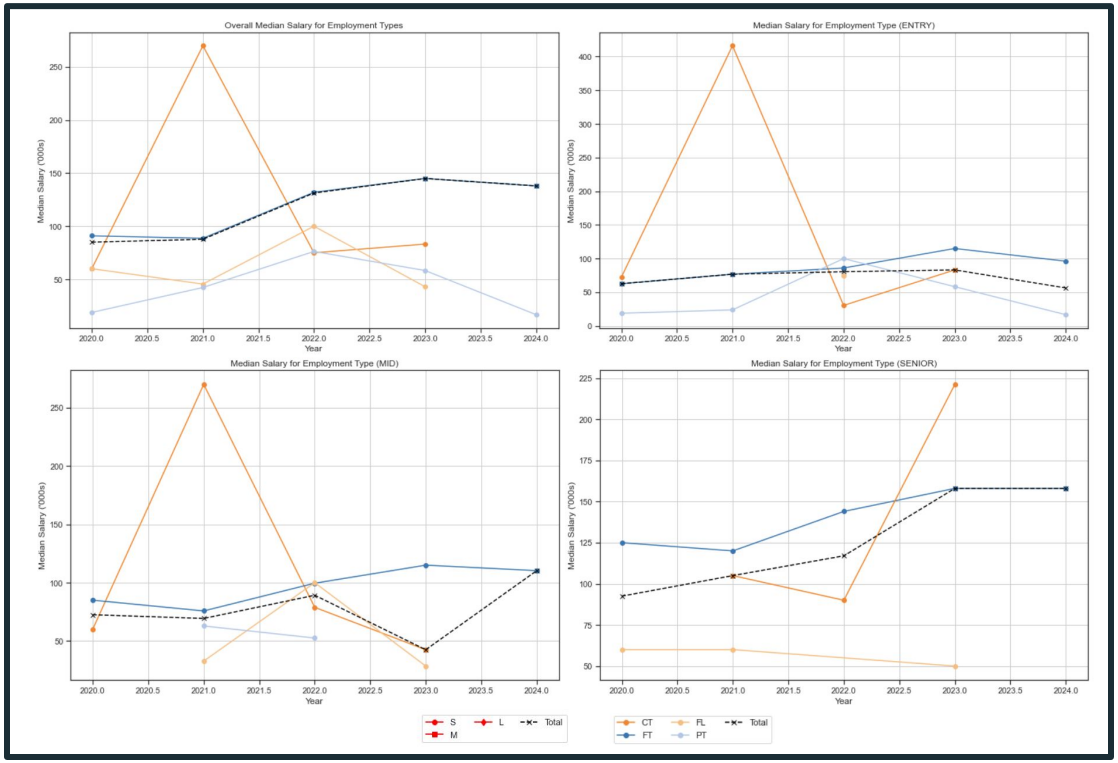
PROJECT II

Analysing Median Salary

Salaries for full-time roles have generally increased throughout the years before dipping slightly in 2024

- Salaries for freelance/consultant roles increased significantly from 2020 to 2021 before dropping sharply in 2022
- This could be due to the limited sample size of freelance/consultant roles

While salaries for entry and mid roles mirror the overall trend, this is not the case for senior roles, where salaries for full-time and consultant roles have maintained high



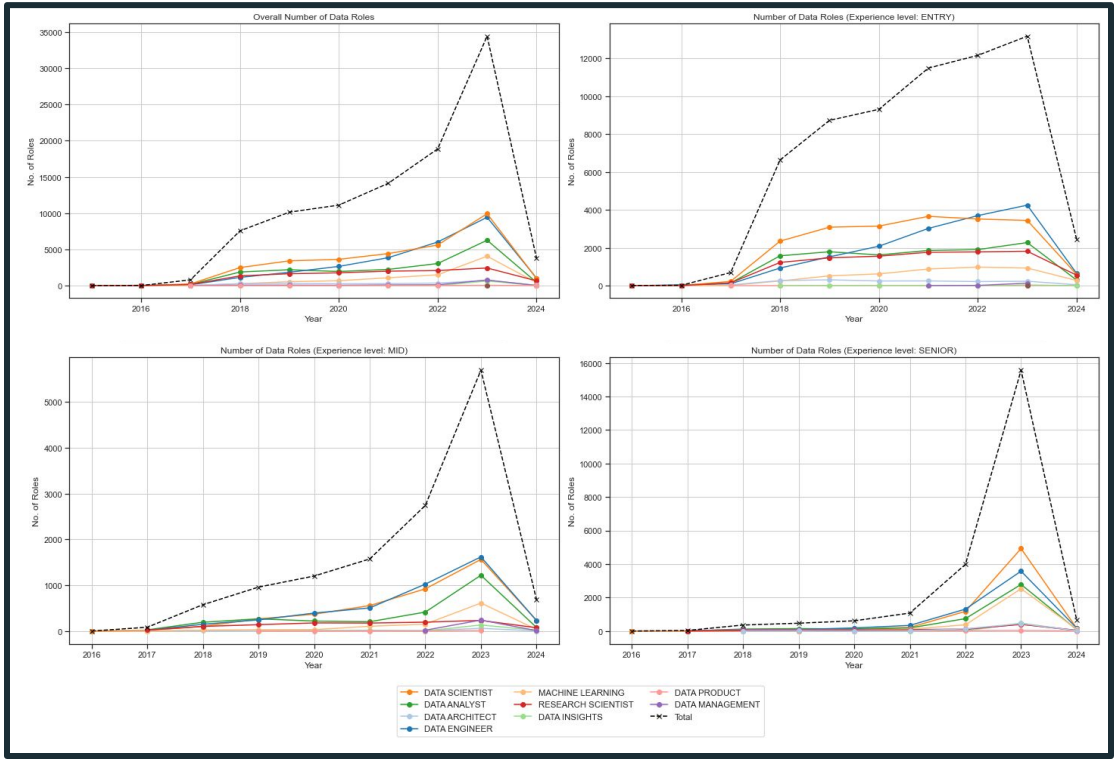
PROJECT II

Analysing Job Demand

Overall, number of data roles have increased across time, with the steepest increase in 2023, followed by a sharp drop in 2024

- This trend is especially observable for data scientist, data engineer and data analyst roles
- Observations are similar across all experience levels
- However, we should note that the data for 2024 is incomplete, and may attribute to the seeming decline in data roles

Demand for data scientists, data engineers and data analysts are the strongest across all experience levels

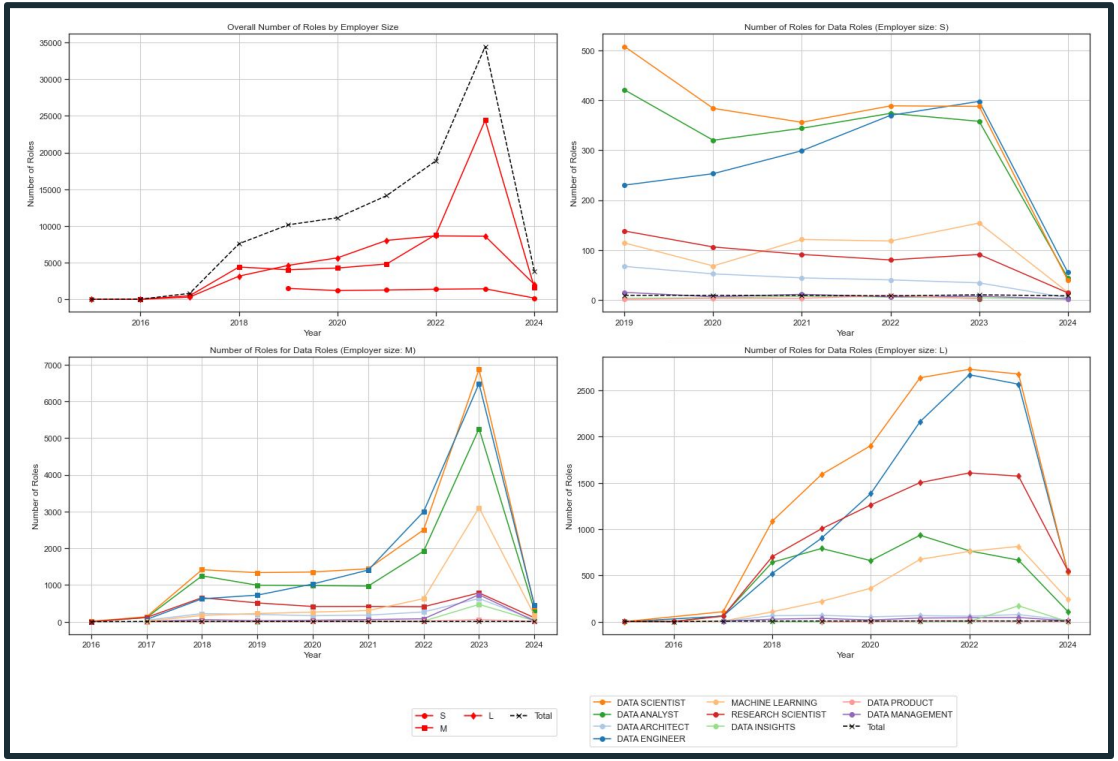


PROJECT II

Analysing Job Demand

For small & medium companies, the top 3 roles in demand are data engineer, data scientist and data analyst roles

For large companies, the top 3 roles in demand are data scientist, data engineer and research scientist roles

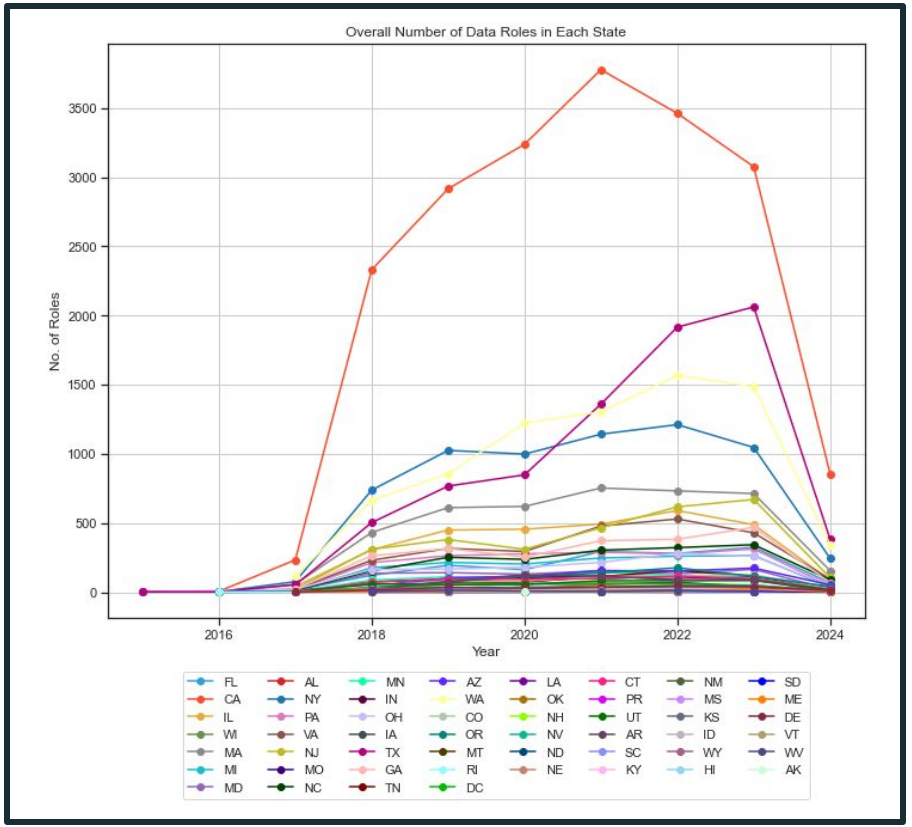


PROJECT II

Analysing Job Demand in US

In US, the number of data roles have increased up until 2021
Following which, demand has been on the decline from 2021 to 2023

- This trend is especially observable for CA, which is also the largest employer of data jobs across all time



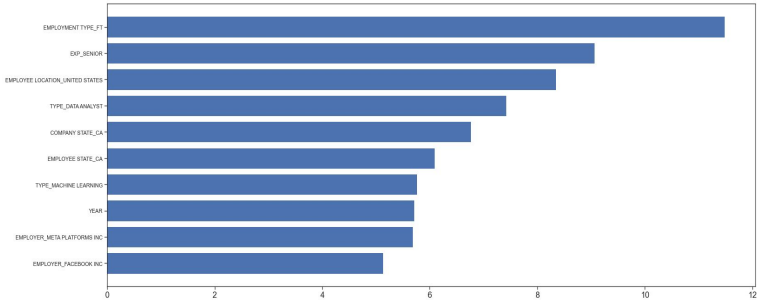
PROJECT II

Feature Importance

Mean Absolute Error (std)

NaN Replacing Method	Encoding	Lasso	Decision Tree Regressor
Most frequent value	One Hot Only	27.127 (29.336)	28.532 (31.200)
	One Hot + Ordinal	27.240 (29.138)	28.530 (31.184)
New "Unknown" Class	One Hot Only	26.326 (30.618)	28.089 (32.014)
	One Hot + Ordinal	27.231 (29.164)	28.513 (31.194)

Feature Selection by Lasso Regression



Key features being

- EMPLOYMENT_FT
- EXP_SENIOR
- EMPLOYEE_LOCATION_UNITED_STATES
- TYPE_DATA_ANALYST
- COMPANY_STATE_CA
- EMPLOYEE_STATE_CA
- TYPE_MACHINE_LEARNING
- YEAR
- EMPLOYER_META PLATFORMS_INC
- EMPLOYER_FACEBOOK_INC

PROJECT III

An Overview

Task	Data Cleaning	Data Analysis

• GICS can lead to crowded risk profiles within portfolios as many investors rely on the same data
• Develop a new stock classification approach to better isolate idiosyncratic risk based on coverage by sell-side analysts
	• Ensuring prices in date are in correct formats
	• Dropping duplicate rows
	• Imputing NaN numerical values with mean values
		• Analysis of Top 10 companies with highest analyst coverage
		• Analysis of Top 10 analysts covering the most companies
		• Similarity and distance matrix of
		  ○ Companies covered by top analyst
		  ○ Companies covered within 1 SD of analyst coverage
		  ○ Companies covered by analysts with median coverage
		  ○ Hierarchical clustering based on each matrix

PROJECT III

Data Cleaning

Check Data Formats

- Ensuring prices are in float64 format, and date is in datetime format

Cleaning Data

- Dropping duplicate rows

Fix Missing Ratings Values

- Impute NaN values with mean values

Before Imputing

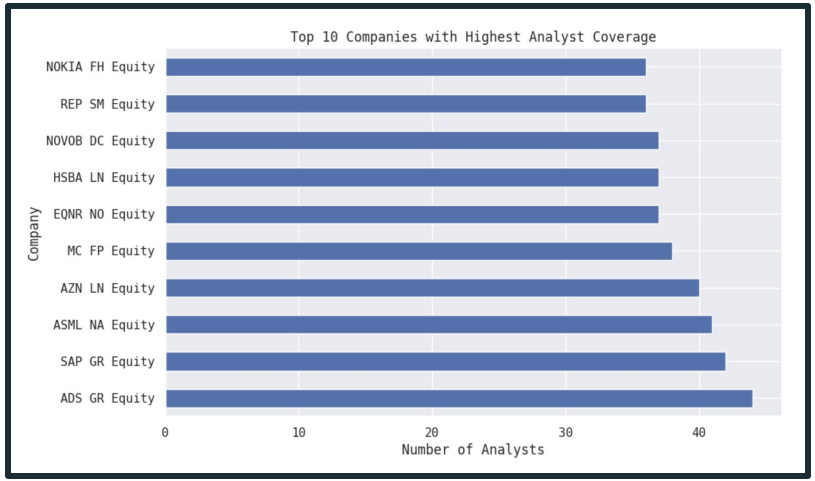
	ANA…	…	BR…	RATING	RECOMMEND…	TARGET_P…	BBTIC…
69	Tarndran	8/9/2017	Shers	NaN	not rated	-2.420000e-14	ROG SW Equity
103	Tarndran	6/19/2017	Shers	NaN	not rated	-2.420000e-14	NOVN SW Equity
297	Tarndran	6/13/2017	Shers	NaN	not rated	-2.420000e-14	SAN FP Equity
435	Tarndran	6/13/2017	Shers	NaN	not rated	-2.420000e-14	NOVOB DC Equi…
777	Tarndran	3/14/2017	Shers	NaN	corporate	-2.420000e-14	BAYN GR Equity

After Imputing

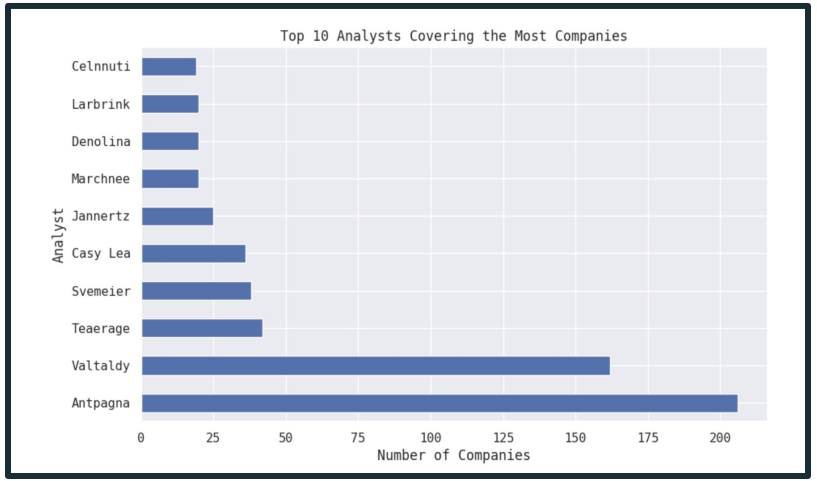
	ANA…	…	BR…	RA…	RECOMMENDA…	TARGET_P…	BBTIC…
0	Tarndran	8/9/2017	Shers	3.796474	not rated	-2.420000e-14	ROG SW Equity
1	Tarndran	6/19/2017	Shers	3.796474	not rated	-2.420000e-14	NOVN SW Equity
8	Tarndran	6/13/2017	Shers	3.796474	not rated	-2.420000e-14	SAN FP Equity
11	Tarndran	6/13/2017	Shers	3.796474	not rated	-2.420000e-14	NOVOB DC Equi…
20	Tarndran	3/14/2017	Shers	3.796474	corporate	-2.420000e-14	BAYN GR Equity

PROJECT III

Analysis of Analyst Coverage



- ADS GR Equity has the highest analyst coverage



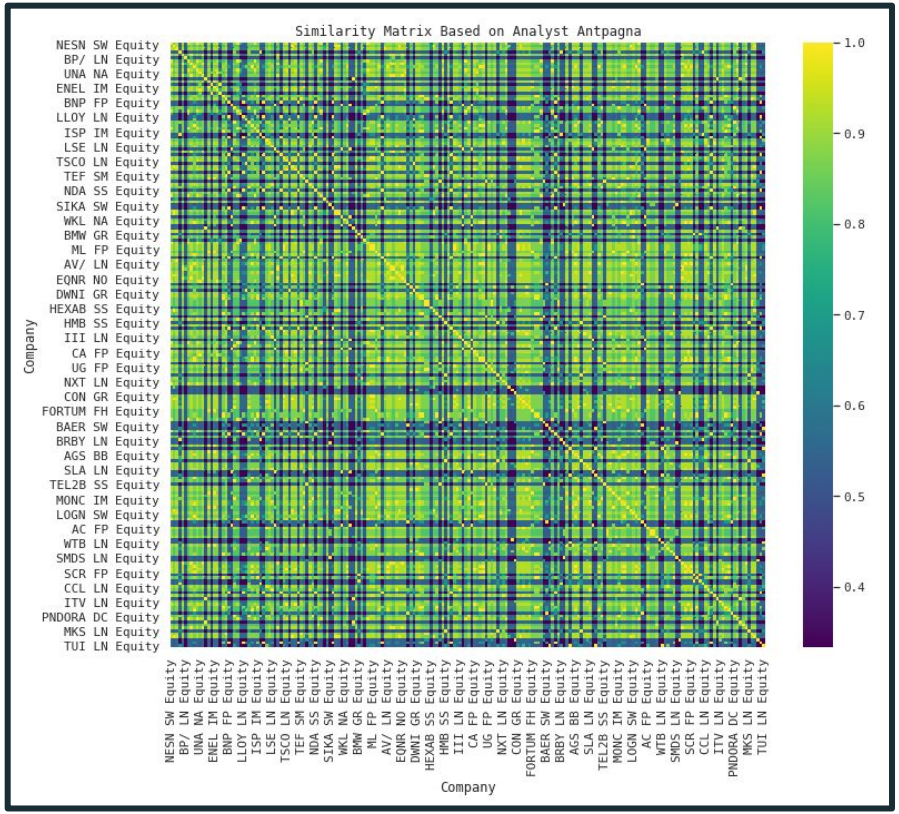
- Antpagna analyst covers the most companies

PROJECT III

Similarity Matrix Based on Companies for Top Analyst (3a)

Antpagna covers almost ⅔ of all unique companies in the dataset

- One-Hot encoding of categorical variables
- Companies covered by Anpagna seem to be similar to each other based on cosine similarity (>85% score), with a few exceptions
  - MC FP Equity is an example of a company different from all others
  - In general, companies that differ from the rest have low cosine similarity with all other companies

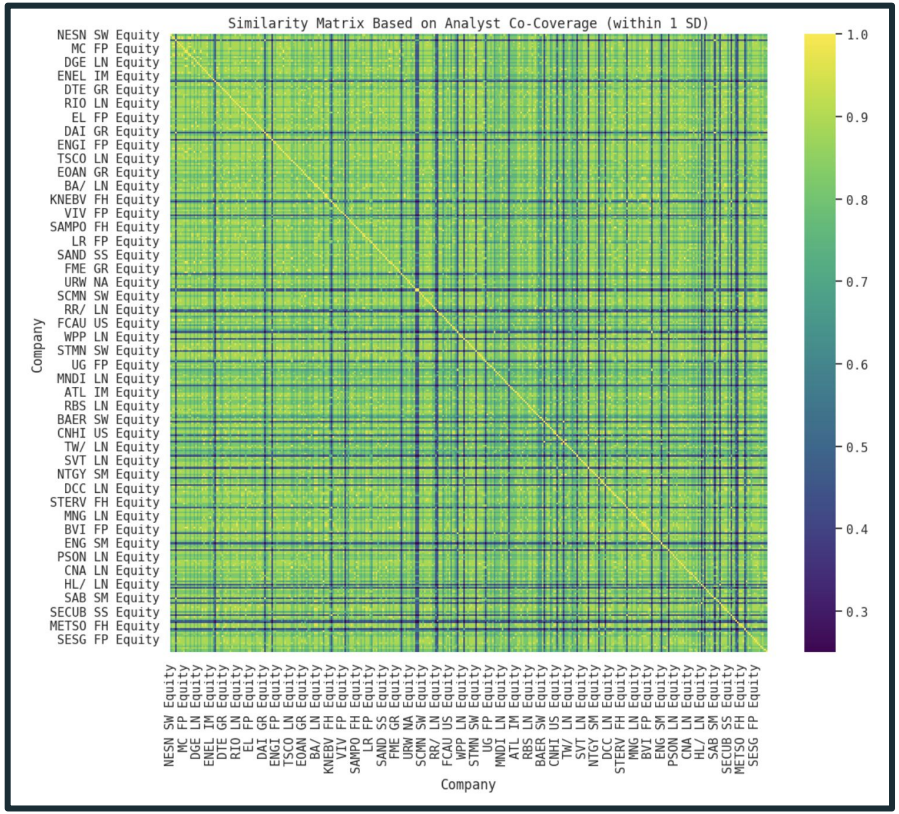


PROJECT III

Similarity Matrix Based on Companies within Analyst Co-Coverage (within 1 SD) (3b)

Companies covered seem to be generally similar to each other based on cosine similarity (>85% score), to a larger extent than analyst Antpagna

- 1974 analysts remain after the filter
- From this, we can understand that analyst Antpagna generally covers a wider variety of companies than most analysts

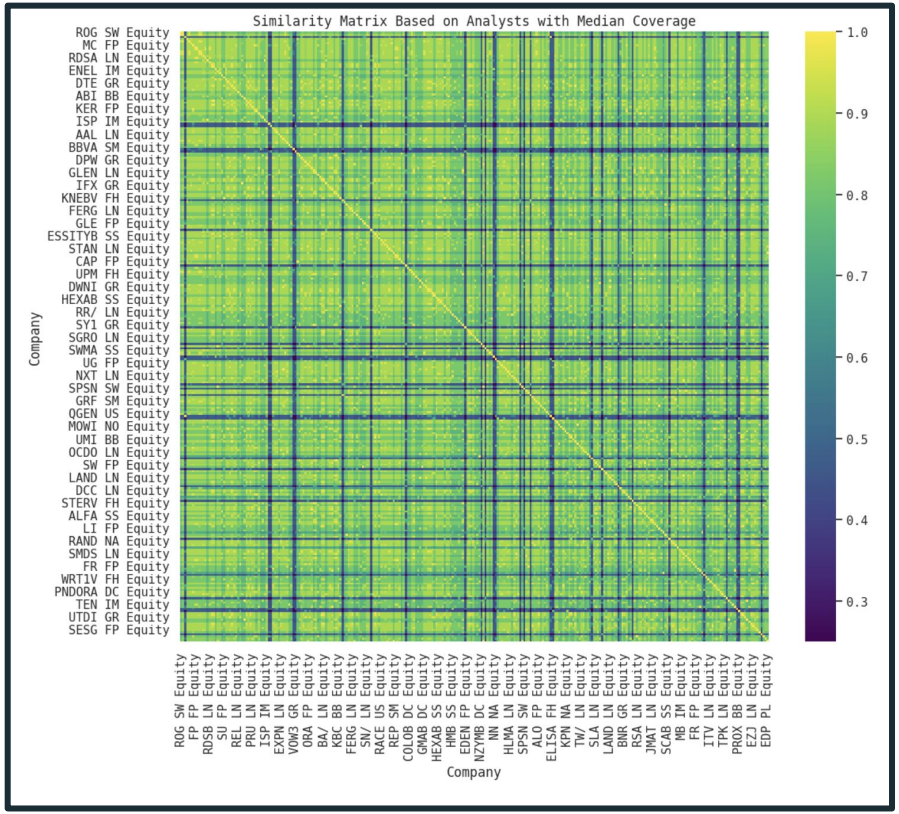


PROJECT III

Similarity Matrix Based on Companies within Analyst Co-Coverage (Median) (3c)

About 10% of the original dataset are analysts covering the median number of companies

- Most companies they cover are similar, with the exception of a small handful
- Those that are different in terms of cosine similarity are different from all other companies.



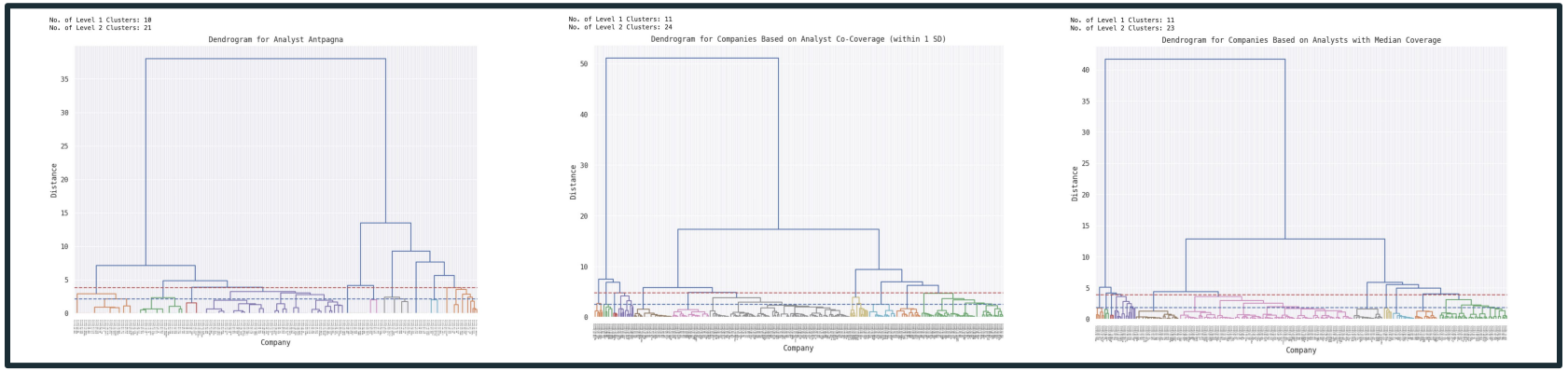
PROJECT III

Hierarchical Clustering

Hierarchical Clustering will be done using distance matrices of the 3 datasets

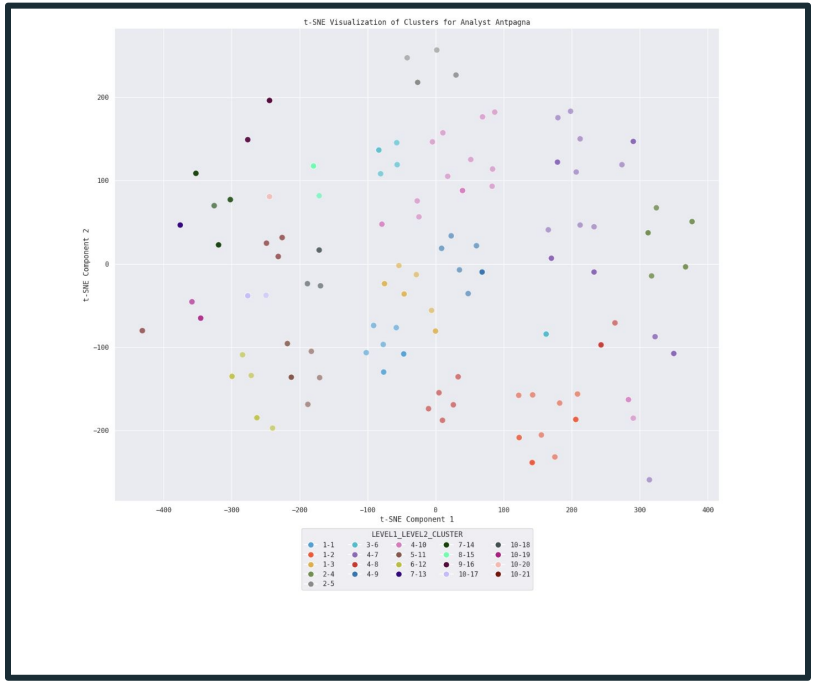
- We use dendrograms to cut each of the 3 datasets into the same number of Level 1 and Level 2 clusters based on original GICS clusters

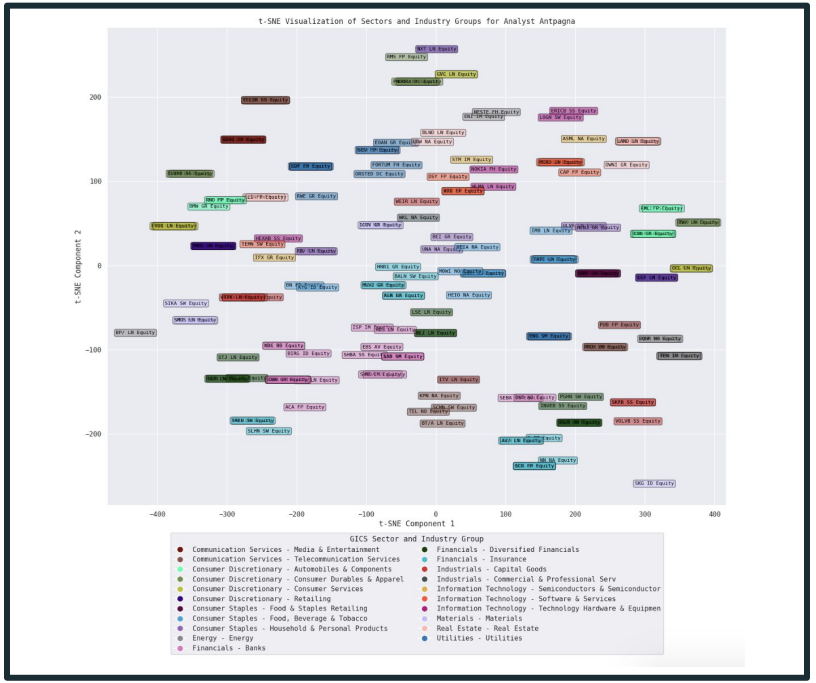
	Level 1			Level 2		
Dataset	Number of Clusters	GICS Silhouette Score	Clustering Silhouette Score	Number of Clusters	GICS Silhouette Score	Clustering Silhouette Score
3a	10	0.17	0.29	21	0.19	0.49
3b	11	0.16	0.14	24	0.16	0.23
3c	11	0.15	0.13	23	0.17	0.23



PROJECT III

t-SNE Cluster Visualisation (for 3a)

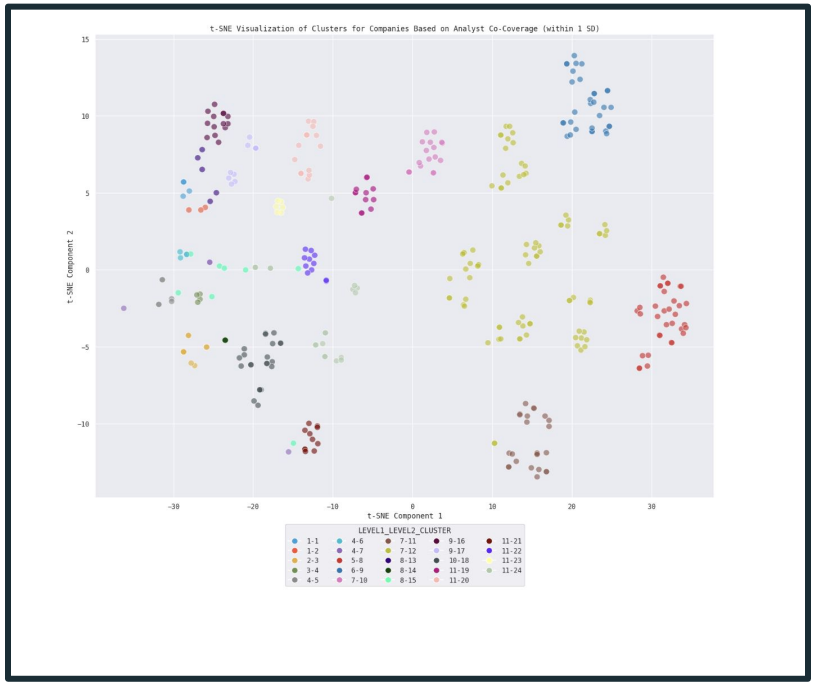


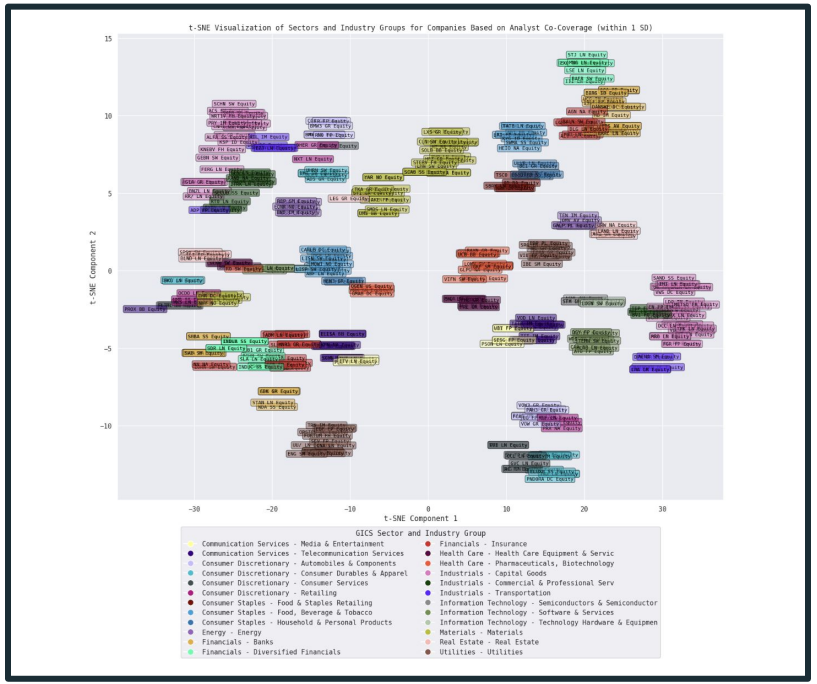


- For companies analysed by Analyst Antpagna, sectors Communication Services, and Financials are most heterogeneous (sectors spread across multiple clusters), and sectors Consumer Staples and Utilities are most homogenous (sectors more clustered together)
- Companies from sector Materials, such as SKG ID Equity and SIKA SW Equity tend to be outliers

PROJECT III

t-SNE Cluster Visualisation (for 3b)

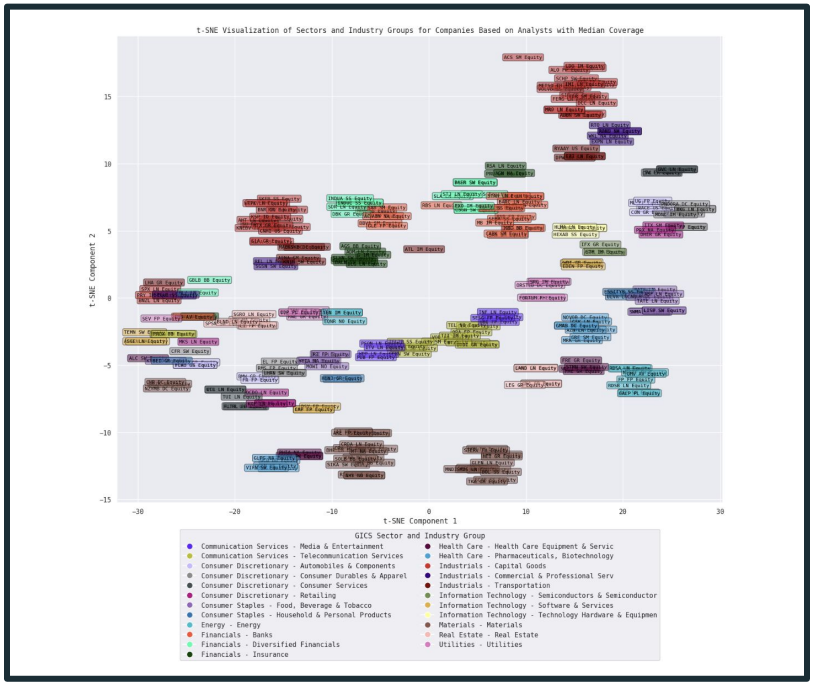




- For companies based on analyst co-coverage (within 1 SD), sectors Financials, and Industrials are most heterogeneous (sectors spread across multiple clusters), and sectors Consumer Discretionary and Materials are most homogenous (sectors more clustered together)
- While there is no particular sector that tends to have outliers, the company PROX BB Equity from sector Communication appears to be an outlier

PROJECT III

t-SNE Cluster Visualisation (for 3c)



- For companies based on analysts with median coverage, sector Information Technology is most heterogeneous (sector spread across multiple clusters), and sectors Healthcare and Financials are most homogenous (sectors more clustered together)
- Companies from sector Industrials and group Transportation, such as LHA GR Equity and ATL IM Equity tend to be outliers