SMU
SINGAPORE MANAGEMENT UNIVERSITY

# Financial Data Science Group Project

Group member:

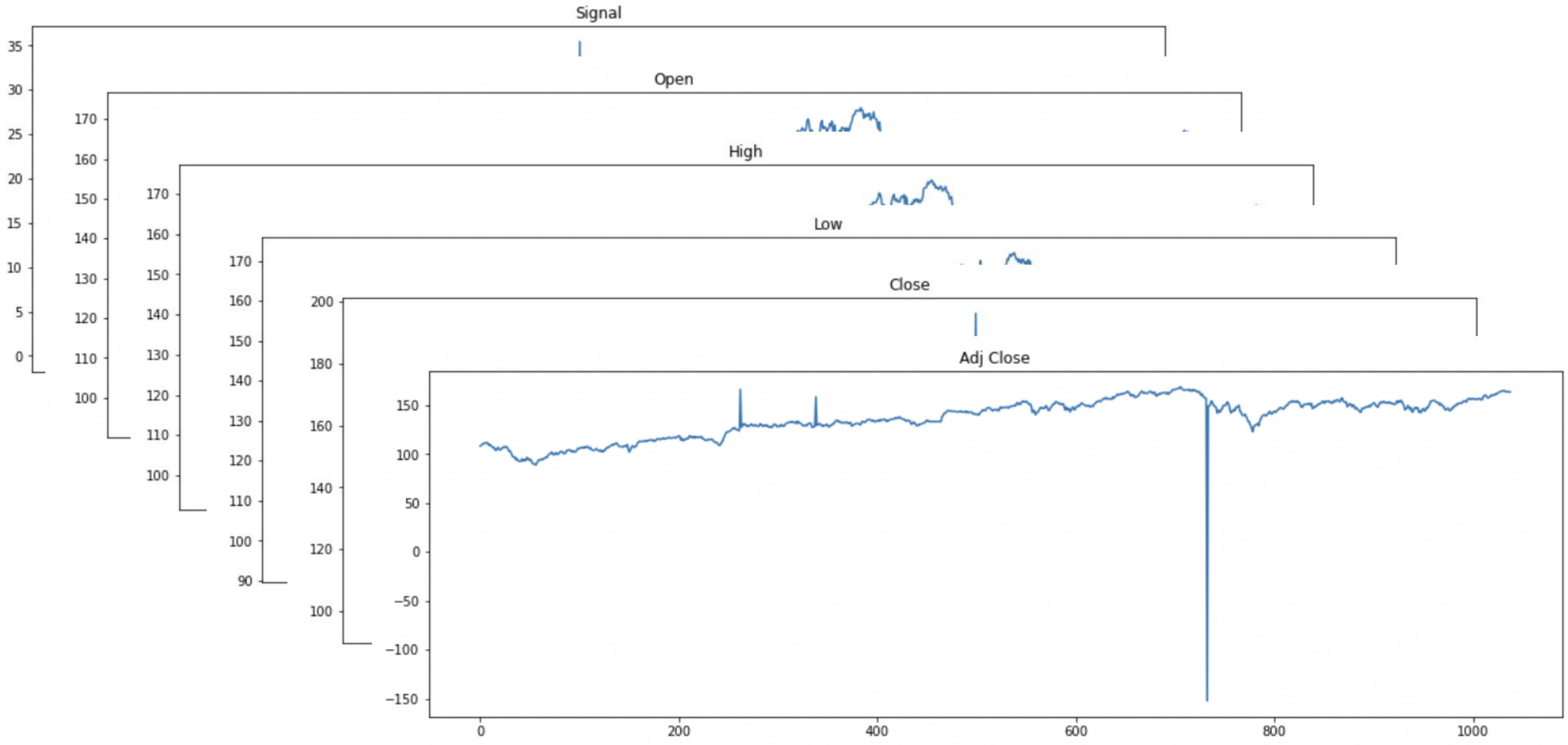
Deng Qiuyun, Jeriel Wong, Ma Qiyuan, Qian Yutao, Zeng Taili

# Project 1

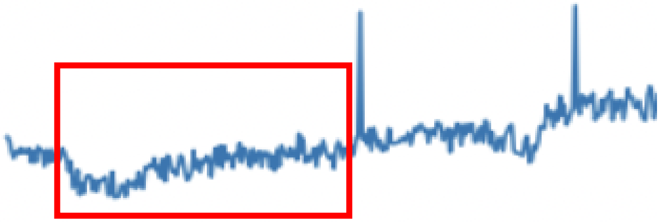
# Data Preview

| | Date | Signal | Open | High | Low | Close | Adj Close |
|---|---|---|---|---|---|---|---|
| **0** | 11/19/2015 | 13.768540 | 116.440002 | 116.650002 | 115.739998 | 116.059998 | 108.281601 |
| **1** | 11/20/2015 | 13.608819 | 116.480003 | 117.360001 | 116.379997 | 116.809998 | 108.981323 |
| **2** | 11/23/2015 | 12.990589 | 116.709999 | 117.889999 | 116.680000 | 117.389999 | 109.522453 |
| **3** | 11/24/2015 | 12.667435 | 116.879997 | 118.419998 | 116.559998 | 118.250000 | 110.324837 |
| **4** | 11/25/2015 | 13.019910 | 118.300003 | 119.320000 | 118.110001 | 119.169998 | 111.183159 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **1033** | 12/30/2019 | 0.000000 | 165.979996 | 166.210007 | 164.570007 | 165.440002 | 163.623688 |
| **1034** | 12/31/2019 | 0.000000 | 165.080002 | 166.350006 | 164.710007 | 165.669998 | 163.851135 |
| **1035** | 1/2/2020 | 0.000000 | 166.740005 | 166.750000 | 164.229996 | 165.779999 | 163.959946 |
| **1036** | 1/3/2020 | 0.000000 | 163.740005 | 165.410004 | 163.699997 | 165.130005 | 163.317093 |
| **1037** | 1/6/2020 | 0.000000 | 163.850006 | 165.539993 | 163.539993 | 165.350006 | 163.534668 |

# Data Preview



# Methodology



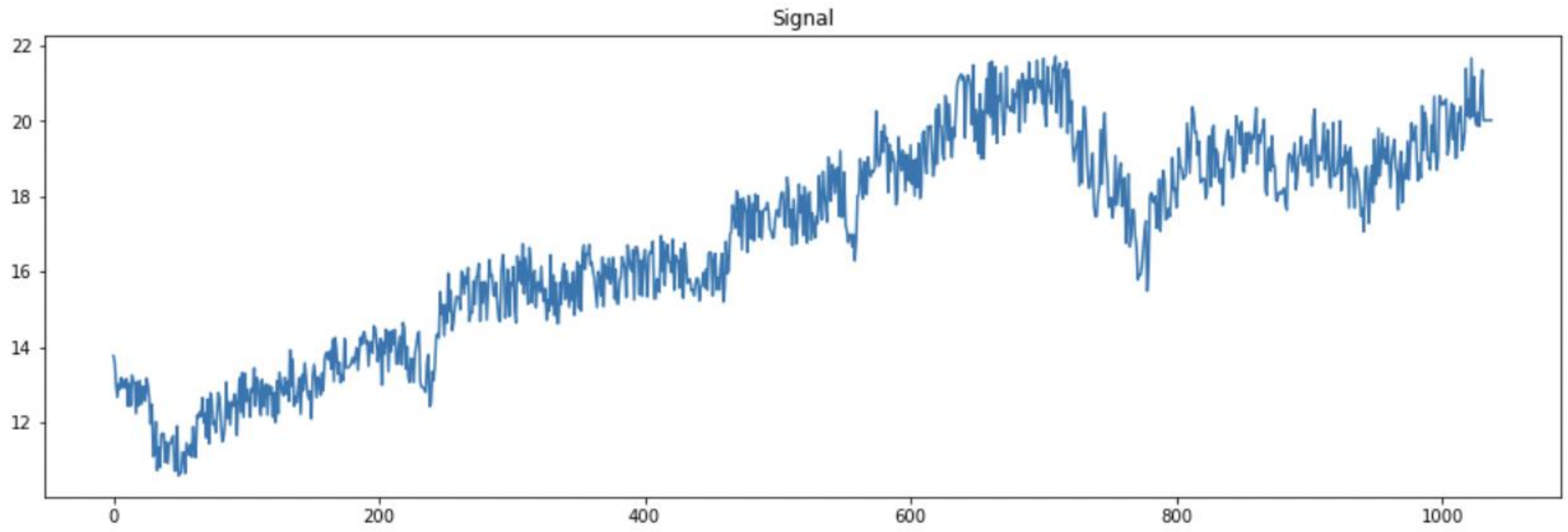
**Rolling Window Sigma Rule**

**1.** Set a window, calculate the mean and standard deviation of the window.

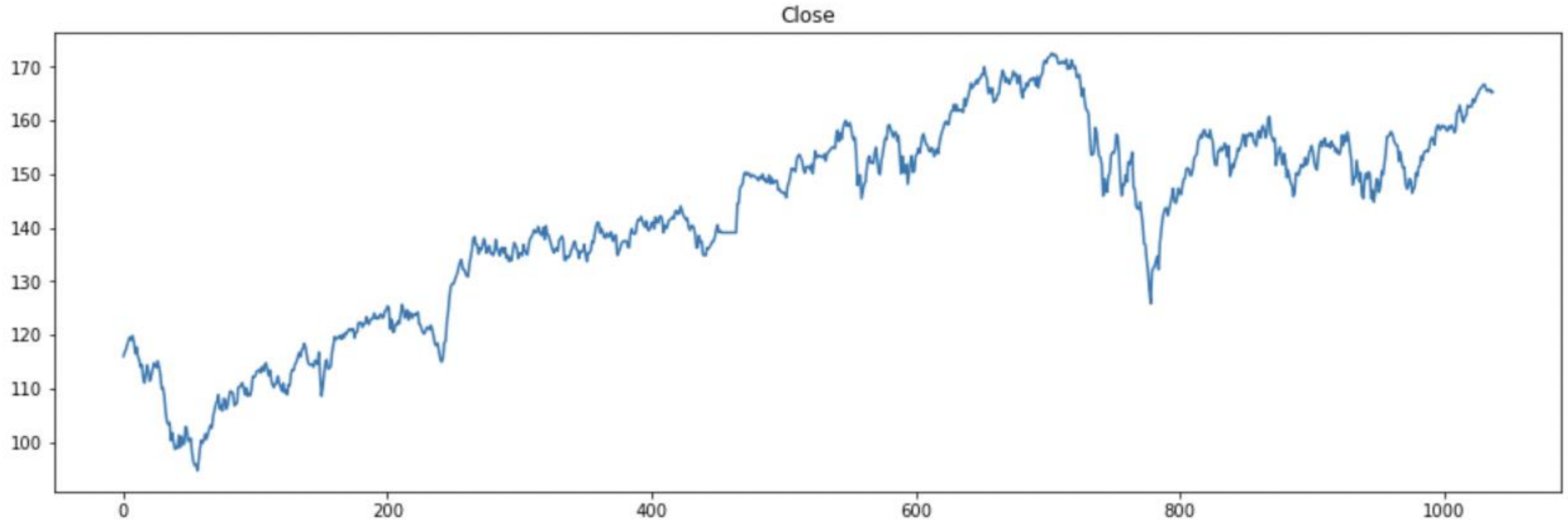
**2.** Determine if the next point is deviated from the mean for a certain times of standard deviation.

**3.** If so, replace the outlier with a proposed value,  for example, use the former value or use linear interpolation.

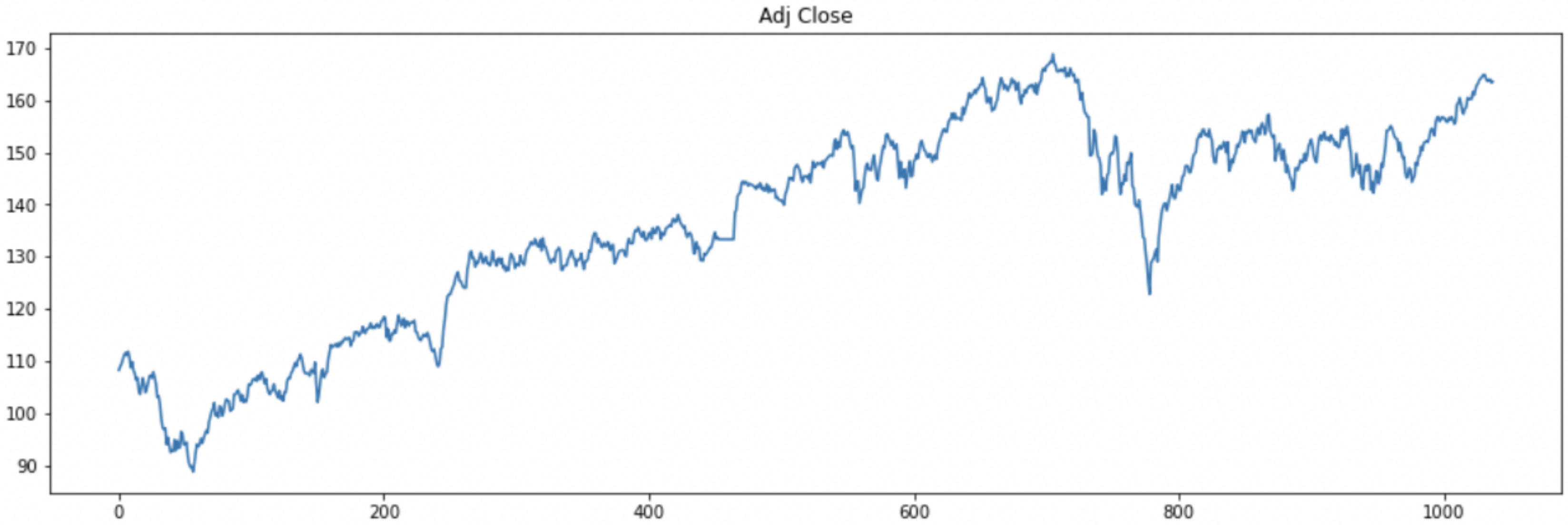
# Signal Result



# Close Result



# Adj Close Result



# Signal-Close OLS and Cointegration Analysis

| | Signal | Close | Adj Close |
|---|---|---|---|
| **Signal** | 1.000000 | 0.958261 | 0.964855 |
| **Close** | 0.958261 | 1.000000 | 0.998006 |
| **Adj Close** | 0.964855 | 0.998006 | 1.000000 |

| | cointegration result | p-value |
|---|---|---|
| **0** | -3.520307 | 0.030637 |

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Close   R-squared:                       0.918
Model:                            OLS   Adj. R-squared:                  0.918
Method:                 Least Squares   F-statistic:                 1.162e+04
Date:                Tue, 21 Jun 2022   Prob (F-statistic):               0.00
Time:                        16:39:36   Log-Likelihood:                -3193.9
No. Observations:                1037   AIC:                             6392.
Df Residuals:                    1035   BIC:                             6402.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.8766      1.015     33.385      0.000      31.885      35.868
Signal         6.4080      0.059    107.795      0.000       6.291       6.525
==============================================================================
Omnibus:                        2.438   Durbin-Watson:                   0.943
Prob(Omnibus):                  0.296   Jarque-Bera (JB):                2.296
Skew:                          -0.100   Prob(JB):                        0.317
Kurtosis:                       3.114   Cond. No.                         106.
==============================================================================
```

# Log Signal-Close OLS and Cointegration Analysis

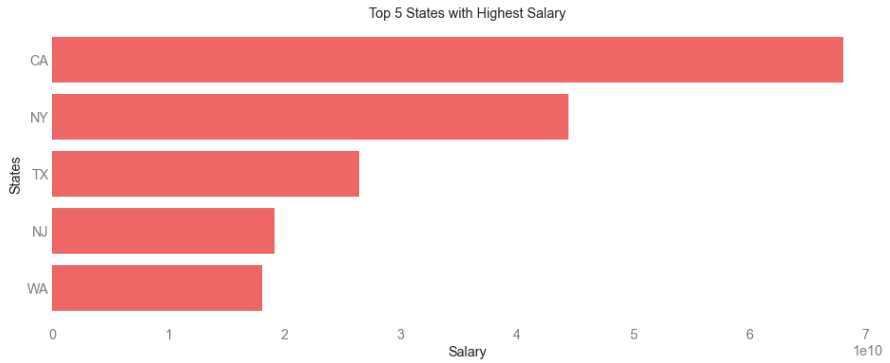
| | Log Signal | Log Close | Log Adj Close |
|---|---|---|---|
| **Log Signal** | 1.000000 | 0.001980 | 0.001732 |
| **Log Close** | 0.001980 | 1.000000 | 0.986661 |
| **Log Adj Close** | 0.001732 | 0.986661 | 1.000000 |

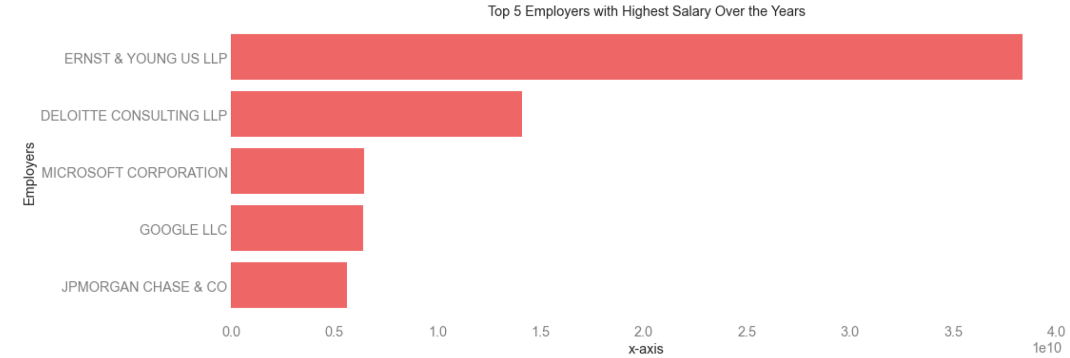
| | cointegration result | p-value |
|---|---|---|
| **0** | -18.965781 | 0.0 |

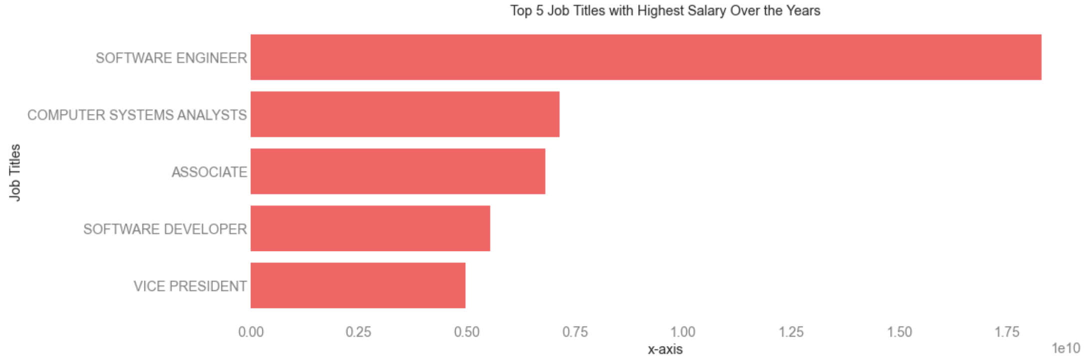
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Log Close   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                  0.004057
Date:                Tue, 21 Jun 2022   Prob (F-statistic):              0.949
Time:                        16:40:50   Log-Likelihood:                 3272.0
No. Observations:                1037   AIC:                            -6540.
Df Residuals:                    1035   BIC:                            -6530.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0003      0.000      1.064      0.288      -0.000       0.001
Log Signal     0.0004      0.007      0.064      0.949      -0.013       0.014
==============================================================================
Omnibus:                       71.712   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              180.788
Skew:                          -0.374   Prob(JB):                     5.52e-40
Kurtosis:                       4.904   Cond. No.                         21.9
==============================================================================
```

# Project 2

# Top 5 Stats



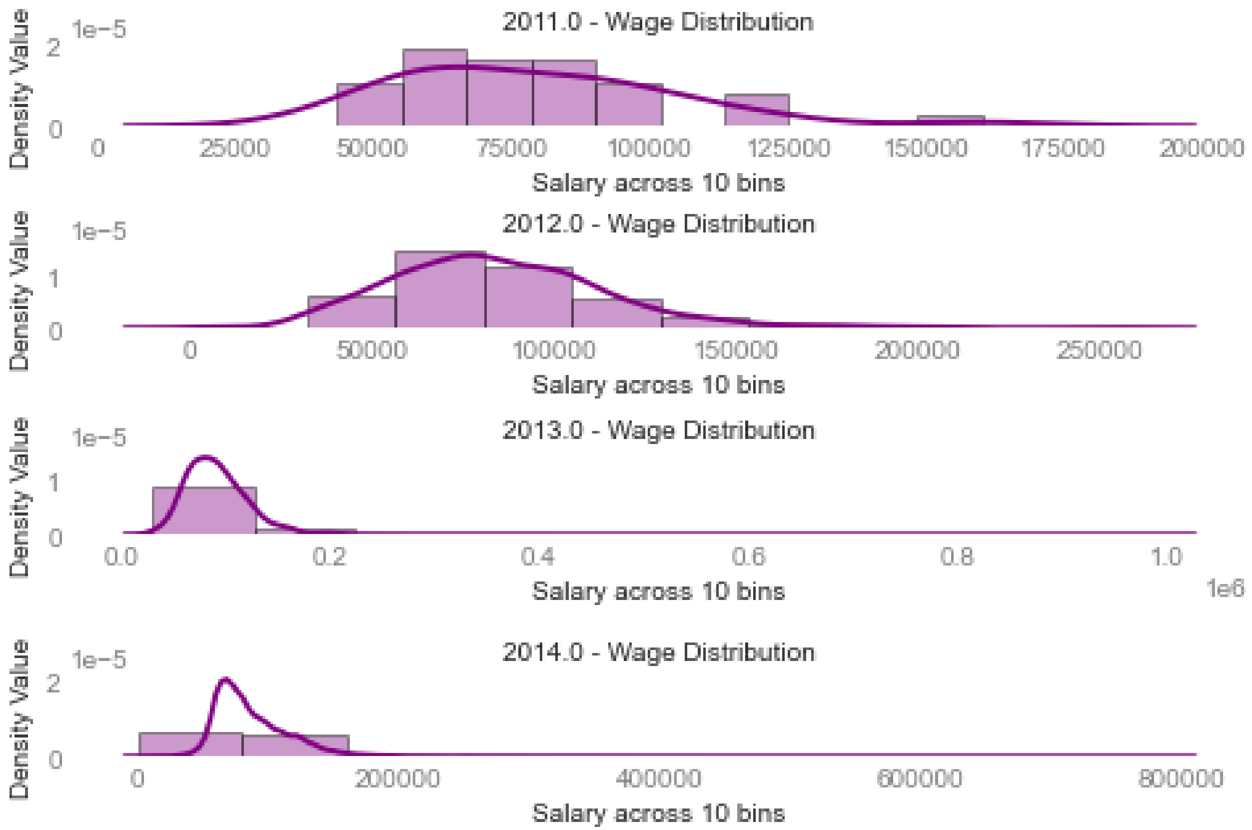

## Jobs with Largest pay increments

```
In [4]: df_key['delta'].nlargest(n=5)

job title
LEAD ARCHITECT                        1.516549e+07
GLOBAL HEAD OF INVESTOR RELATIONS     3.600000e+06
GENERAL COUNSEL                       2.034195e+06
SURVEY RESEARCHER                     9.219400e+05
CHIEF CREATIVE OFFICER                7.800000e+05
Name: delta, dtype: float64
```

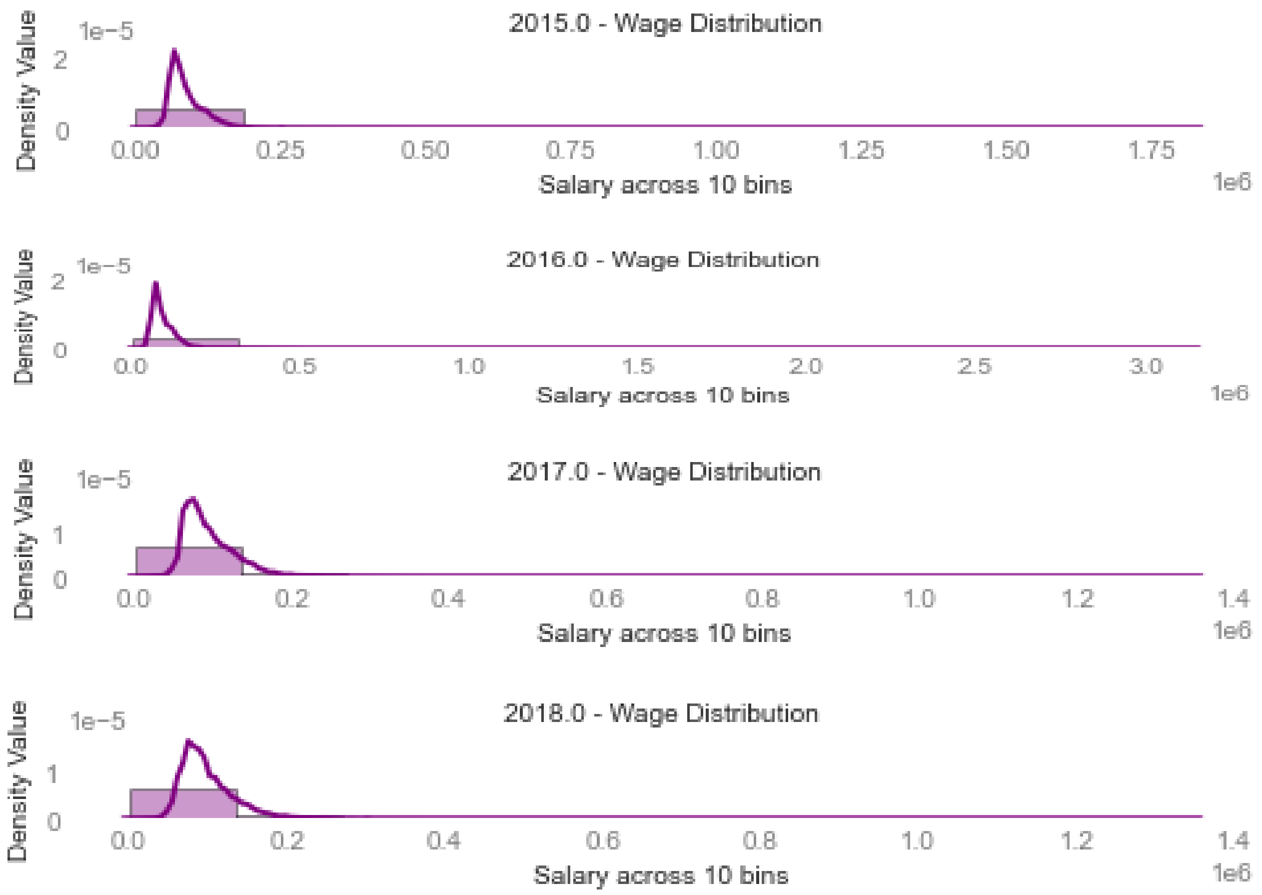
## Jobs with Largest pay decrease ¶

```
[5]: df_key['delta'].nsmallest(n=5)

job title
PATENT ATTORNEY - PROSECUTION                             -1780000.0
ADVERTISING SALES MANAGER                                  -879434.0
DISTINGUISHED RESEARCH SCIENTIST                           -557379.0
COMPUTER GRAPHICS PIPELINE LEAD                            -524912.0
DIRECTOR FOR THE CENTER OF COMPUTATIONAL QUANTUM PHYSICS   -490685.0
Name: delta, dtype: float64
```

# Wage Distribution – Across Years (2011 – 2014)

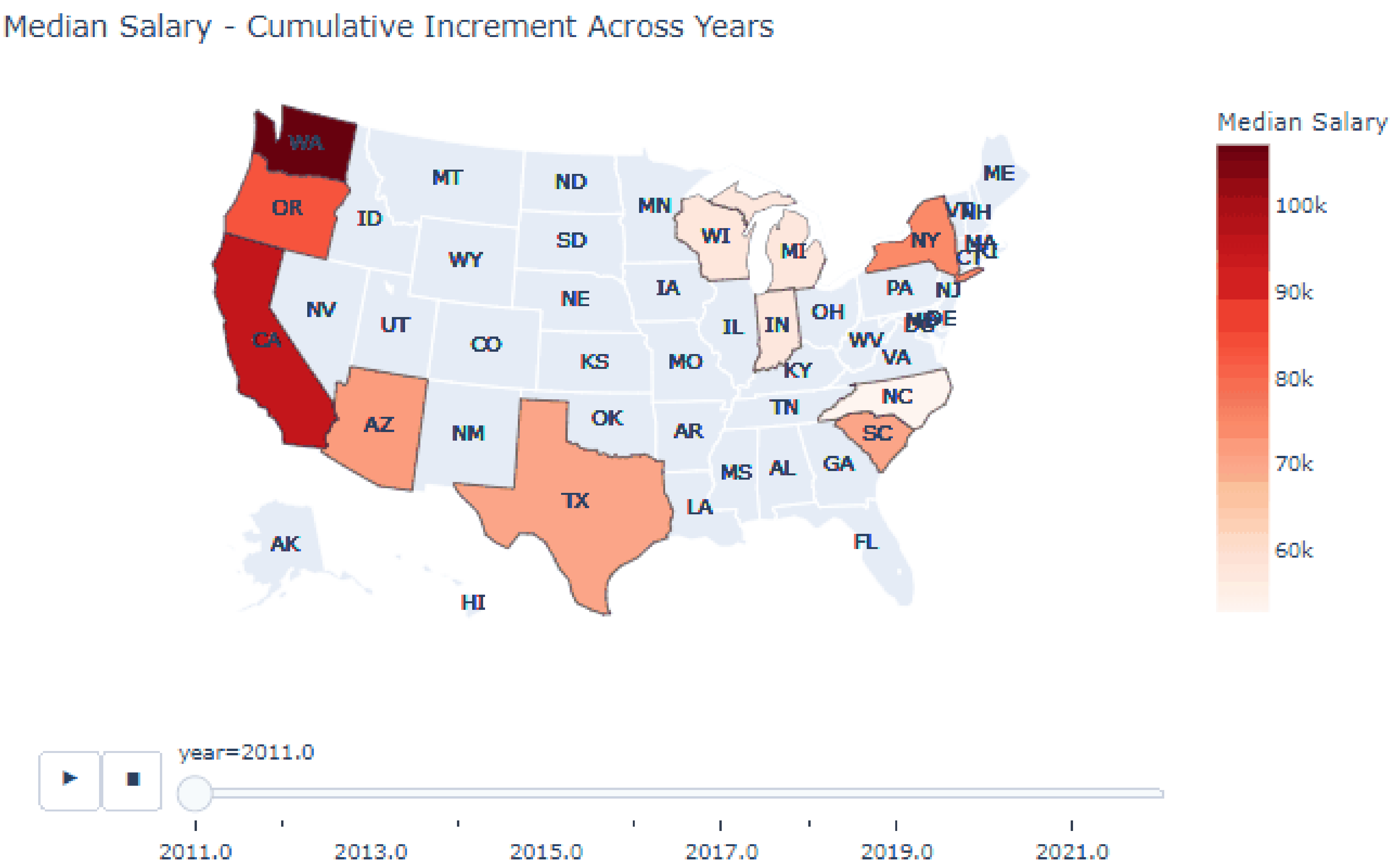


# Wage Distribution – Across Years (2015 – 2018)



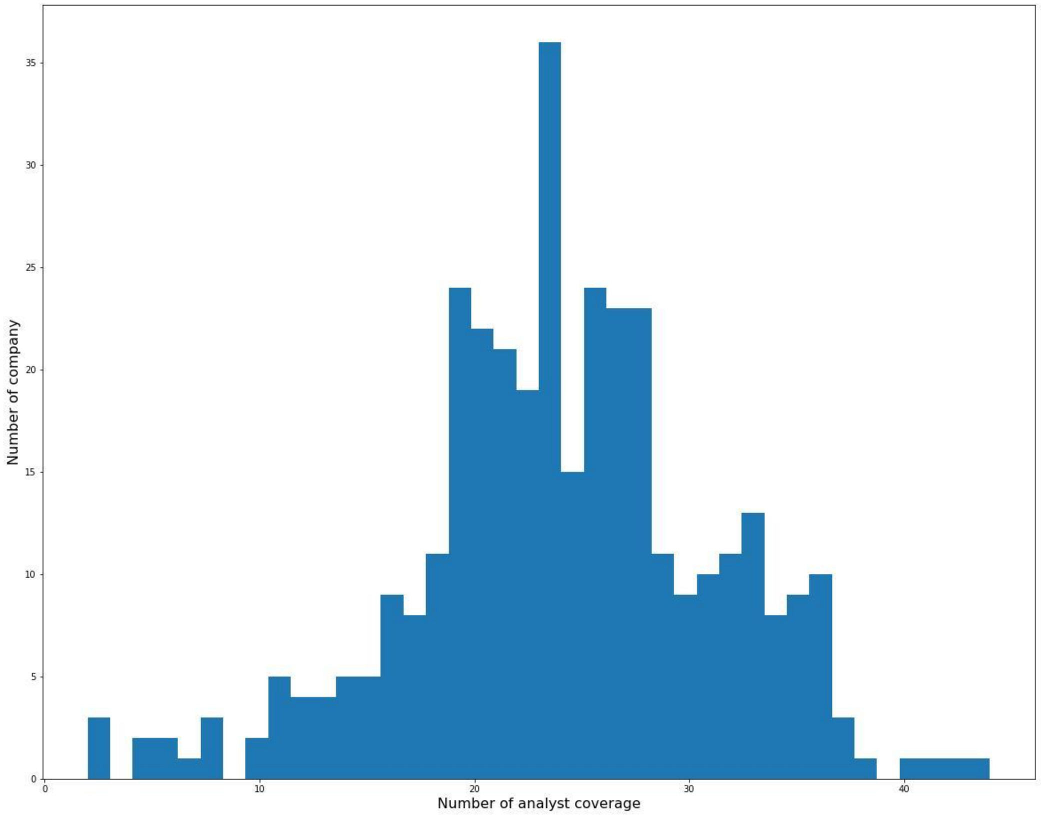
# Wage Distribution – Across Years (2019 – 2022)

# Median Salary – Across Cities – Across Years



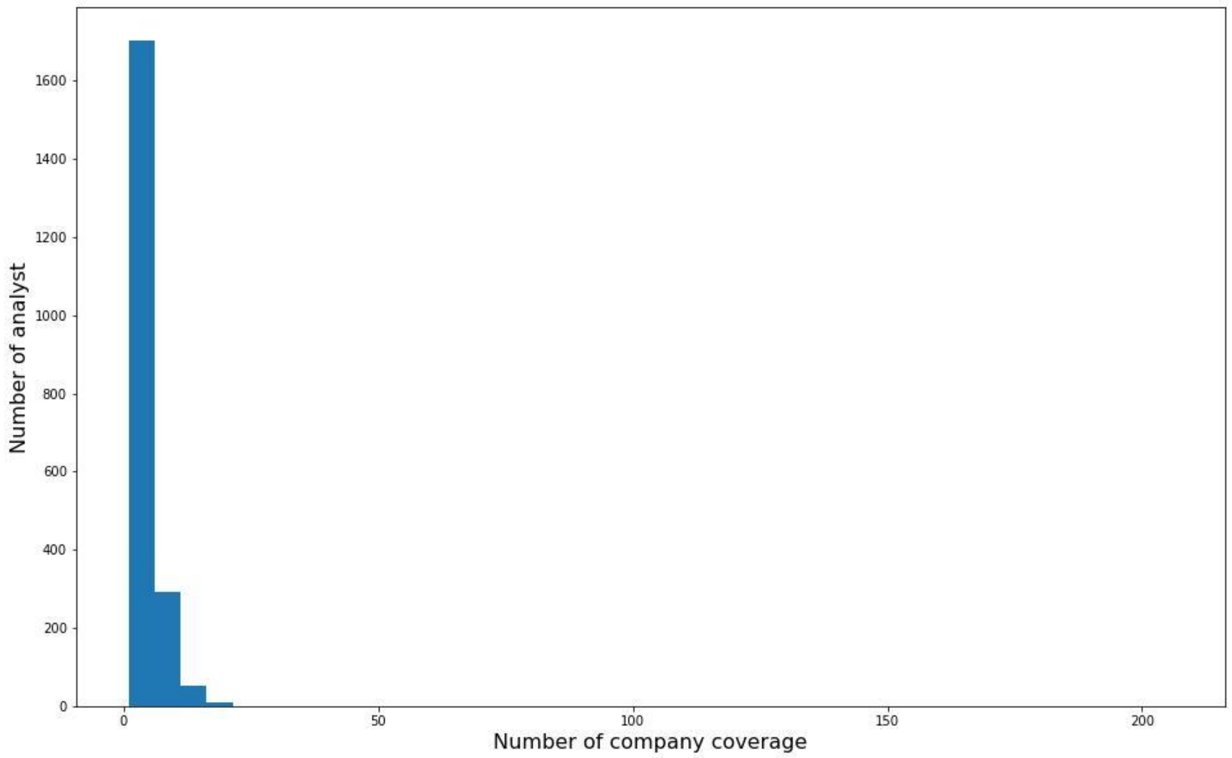
# Project 3

# 1. Which company has the highest analyst coverage?




- The company that has the highest analyst coverage is ADS GR Equity, and the number of analysts is 44.

# 2. Which analyst covers the most companies?




- The analyst that covers the most companies is Antpagna, and the number of companies is 206.

# 3.1 Similarity Matrix

| | NESN SW Equity | ROG SW Equity | NOVN SW Equity | HSBA LN Equity | SAP GR Equity | AZN LN Equity | ASML NA Equity | SAN FP Equity | MC FP Equity | FP FP Equity | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NESN SW Equity | 0 | 0.006173 | 0.006173 | 0.064996 | 0.069851 | 0.006173 | 0.011027 | 0.006173 | 0.119851 | 0.011027 | ... |
| ROG SW Equity | 0.006173 | 0 | 4.172021 | 0.006173 | 0.077601 | 4.033926 | 0.106173 | 2.433926 | 0.006173 | 0.106173 | ... |
| NOVN SW Equity | 0.006173 | 4.172021 | 0 | 0.006173 | 0.077601 | 3.447129 | 0.106173 | 3.176783 | 0.006173 | 0.106173 | ... |
| HSBA LN Equity | 0.064996 | 0.006173 | 0.006173 | 0 | 0.064996 | 0.006173 | 0.006173 | 0.006173 | 0.064996 | 0.006173 | ... |
| SAP GR Equity | 0.069851 | 0.077601 | 0.077601 | 0.064996 | 0 | 0.006173 | 0.378884 | 0.097082 | 0.069851 | 0.011027 | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| TUI LN Equity | 0.004854 | 0 | 0 | 0 | 0.067354 | 0 | 0.067354 | 0 | 0.004854 | 0.004854 | ... |
| GFS LN Equity | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| LHA GR Equity | 0 | 0 | 0 | 0 | 0.0625 | 0 | 0.0625 | 0 | 0 | 0.109649 | ... |
| BMW3 GR Equity | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| UHRN SW Equity | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.111111 | 0 | ... |

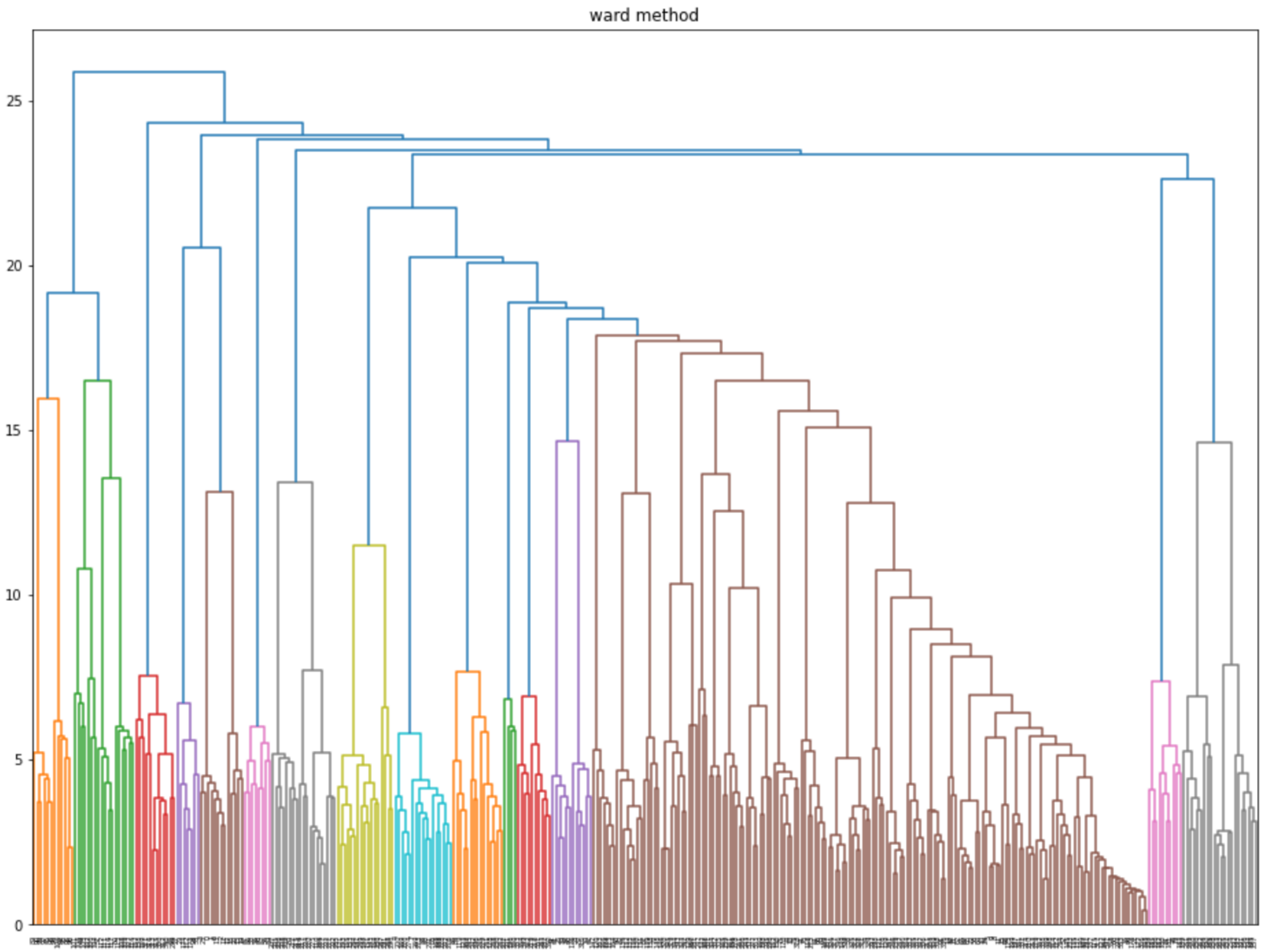
- In this similarity Matrix, we consider all the analysts.
- In the next page, we make some restriction on the analysts.

# 3.2 Restricted to analysts that cover >=3 companies and <=20 companies

| | NESN SW Equity | UNA NA Equity | ULVR LN Equity | BN FP Equity | GIVN SW Equity | KYG ID Equity | HEN3 GR Equity | SY1 GR Equity | LISN SW Equity | LISP SW Equity | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **NESN SW Equity** | 0 | 2.874071 | 1.648674 | 3.42169 | 0.502632 | 0.46765 | 1.381214 | 0.302632 | 0.927356 | 1.026623 | ... |
| **UNA NA Equity** | 2.874071 | 0 | 1.716135 | 2.735182 | 0.135965 | 0.312888 | 1.575659 | 0.135965 | 0.560689 | 0.588528 | ... |
| **ULVR LN Equity** | 1.648674 | 1.716135 | 0 | 1.565341 | 0.135965 | 0.384317 | 1.27169 | 0.135965 | 0.417832 | 0.3171 | ... |
| **BN FP Equity** | 3.42169 | 2.735182 | 1.565341 | 0 | 0.302632 | 0.634317 | 1.422881 | 0.302632 | 0.727356 | 0.826623 | ... |
| **GIVN SW Equity** | 0.502632 | 0.135965 | 0.135965 | 0.302632 | 0 | 0.894298 | 0.185965 | 2.859017 | 0.283333 | 0.45 | ... |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **BLND LN Equity** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| **PRX NA Equity** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| **MNDI LN Equity** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| **SKG ID Equity** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| **RYAAY US Equity** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |

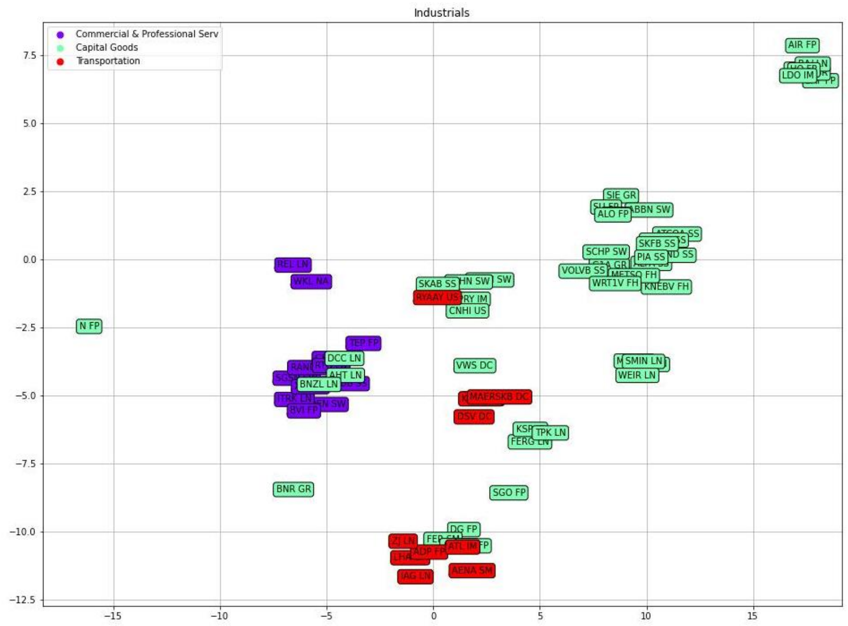
- In this matrix, we can see that some relationships with small similarity are eliminated, since extreme company coverage is avoided, similarity with largest number and smallest number disappear as well.

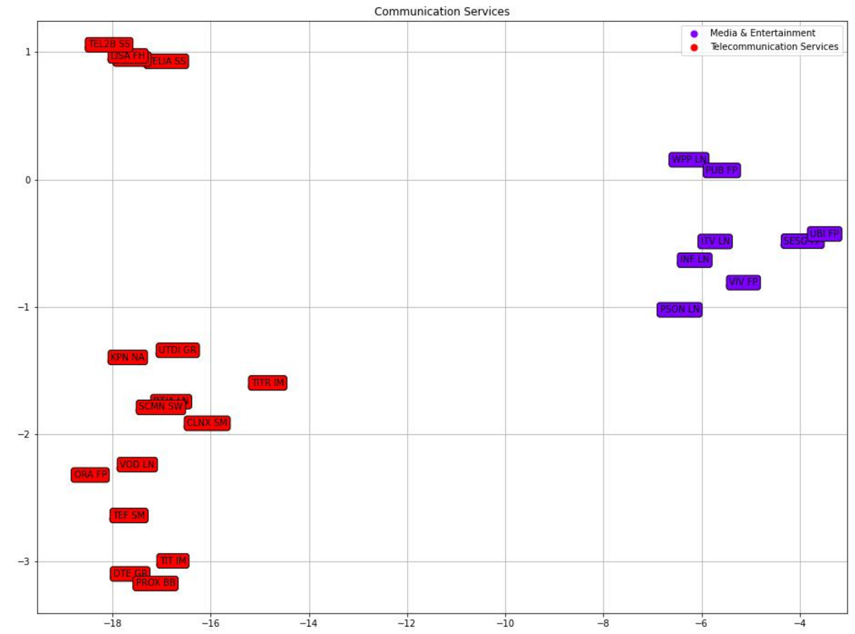
# 4.Hierarchical




- Ward method shows the best result, since it considers each factor in the most balance way.

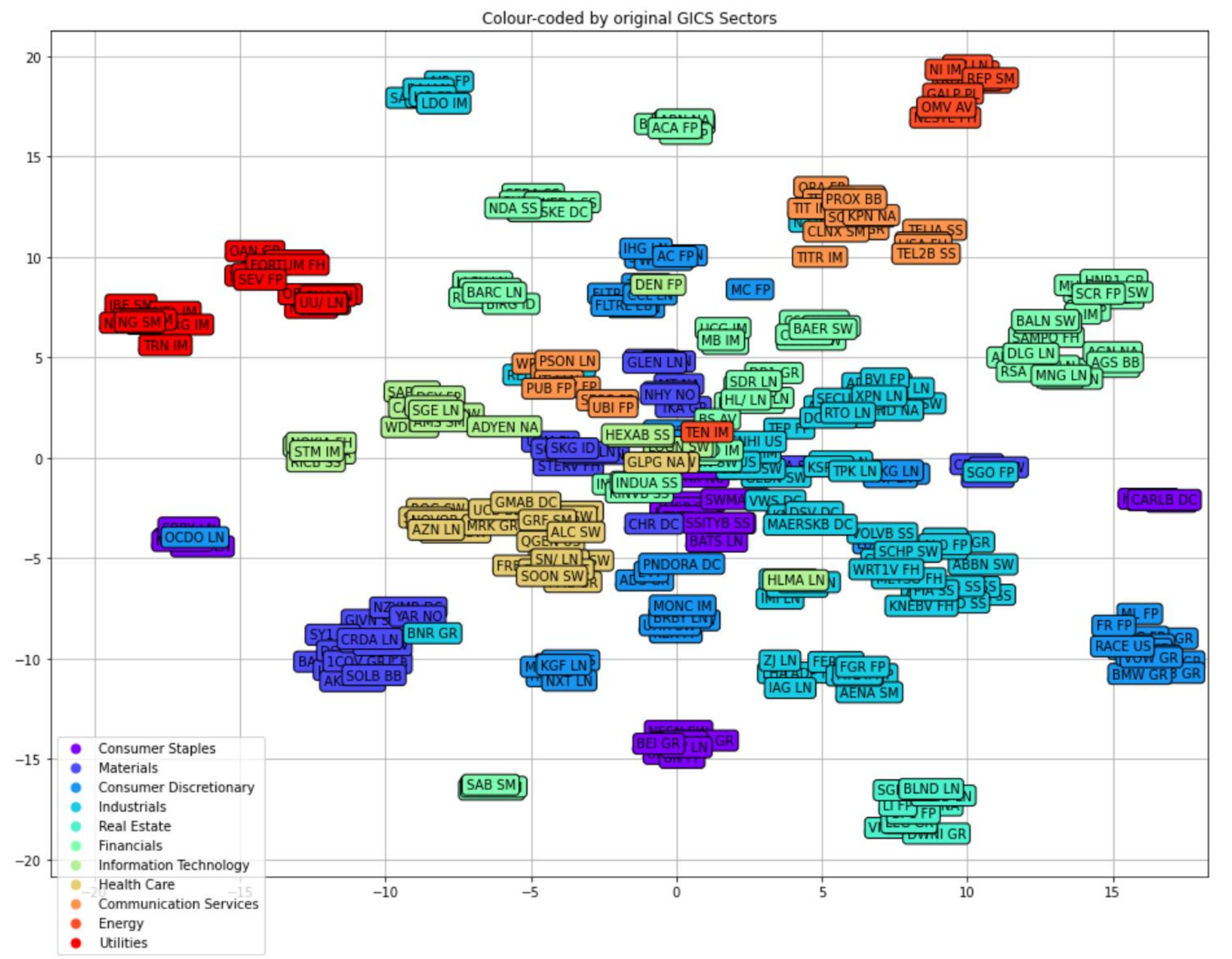
# 5.Homogeneous and Heterogeneous



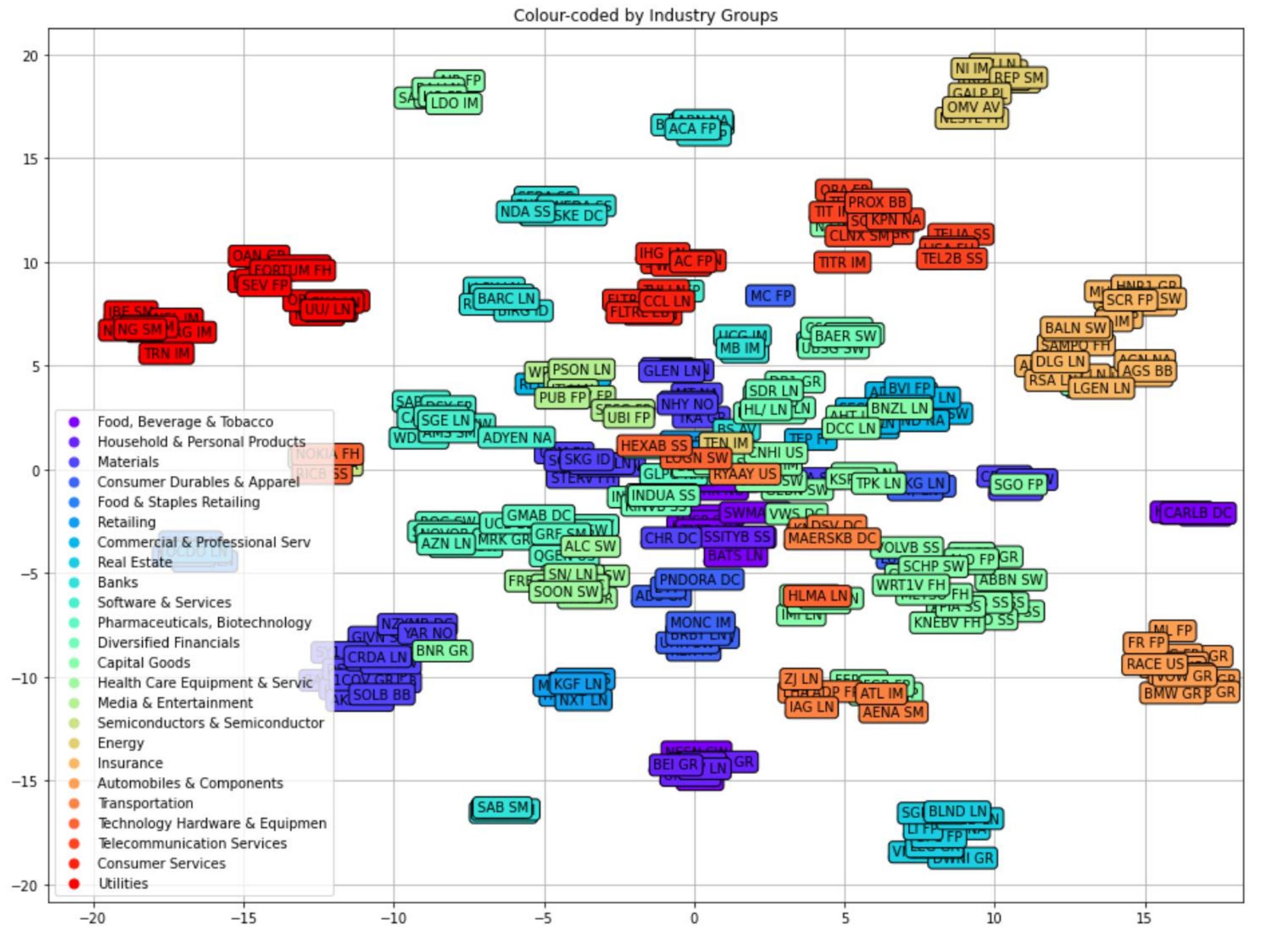




| | GICS_SECTOR_NAME | Distance |
|---|---|---|
| 0 | Health Care | 2.302836 |
| 1 | Utilities | 2.662624 |
| 2 | Energy | 2.708263 |
| 3 | Real Estate | 3.035293 |
| 4 | Information Technology | 4.303414 |
| 5 | Financials | 6.270991 |
| 6 | Industrials | 7.240057 |
| 7 | Communication Services | 7.582258 |
| 8 | Consumer Staples | 7.718867 |
| 9 | Materials | 8.765679 |
| 10 | Consumer Discretionary | 11.228444 |

# Colored by Sectors



# Colored by Industry Groups



Thanks for watching!