**SECTIONS (Add links to doc sections)**

### [TEAM,-PROJECT SPECIFICATIONS & NEXT STEPS](#tflqhxz2rbp6)

### [DOCUMENTATION FOR PROFESSOR AND RESULTS](#1w48wa2vkjwb)

### [TECHNICAL DOCUMENTATION w CODE SCREENSHOTS](#hqjgio32je0l)

### [DASHBOARD SPECIFICATIONS](#o4xkh7r5i8kg)

### [TECHNICAL ISSUES & DATA DEPOSITORY](#nourjffze59)

### [APPENDIX WITH ALL RELATED NOTES & EMAILS](#gaslr2j9xpeu)

######################## TEAM, PROJECT SPECIFICATIONS & NEXT STEPS

///// TEAM

[CURRENT]

Ada, Jane Zhijun Liu, Stanley, Fiona

[PREVIOUS]

Fiona

///// NEXT STEPS

////// SPECIFICATIONS: OVERALL STEPS & EXPECTED RESULTS - 1st Deadline June 13

[Initial]

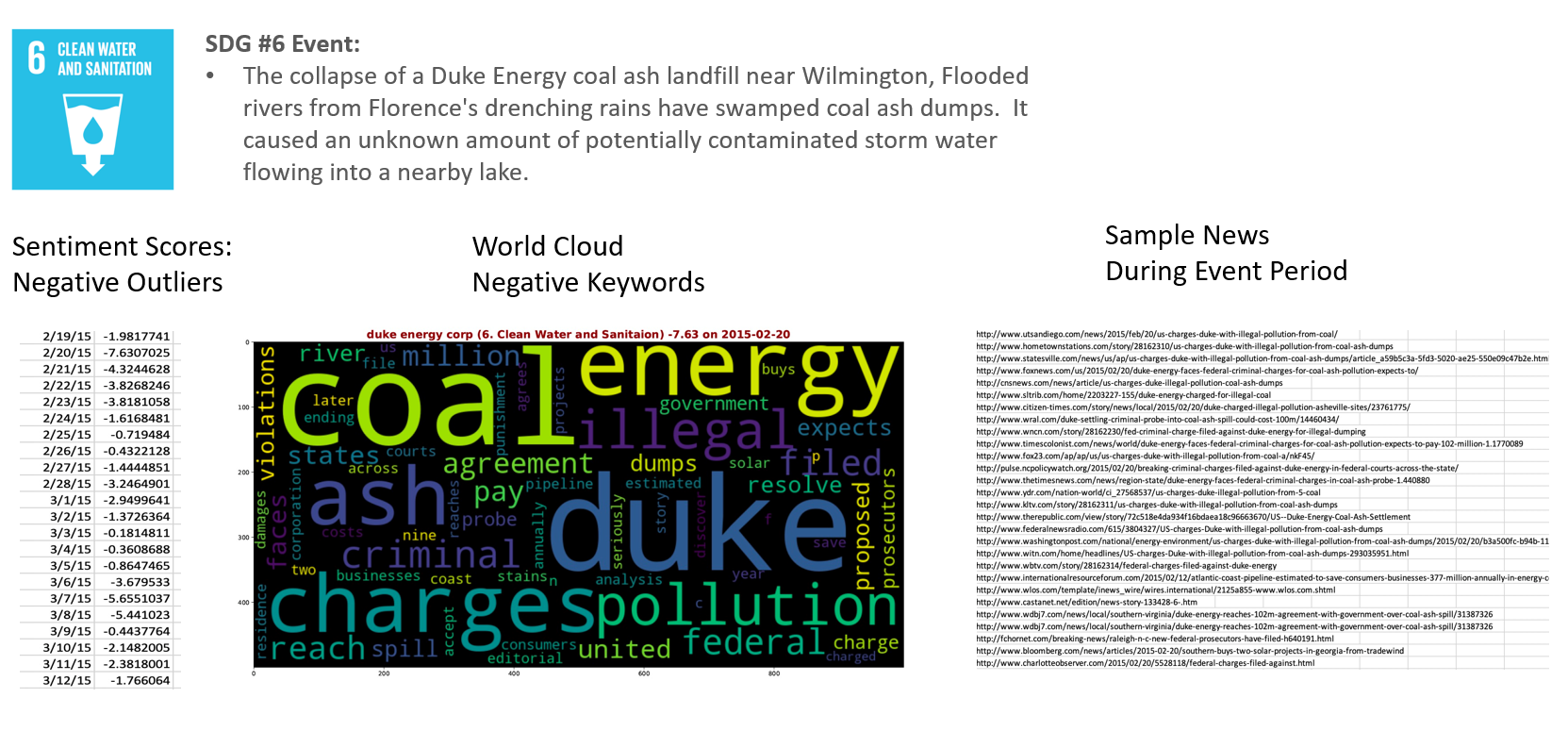
Do daily Dashboard shows word clouds for companies that have outliers on any given SDGs; the main page is a heatmap of companies and SDGs; or a list of SDG-Company combinations that have outliers; then after clicking on one, it shows the Word Clouds

* Jane & Ada: outlier detection and word cloud automation & algorithm improvement
* Stanley: dashboard connection & automation

[PHASE 1]

Heatmap: In the first dashboard (main page), EVERY DAY there will be a heatmap of 17 SDGs X N companies (600+); and if no outlier defined, then cell is white. If 1rst threshold then color is a bit red; and if 2nd threshold hit, then color is more red, and so on

For each outlier in the heatmap, include the following graphs



For outlier time series, include line graph with T+/-10 (10 days before, and 10 days after the outlier event)

For URLS, there is no ranking of relevance yet. For now just show the first 20 or 30 URLS

[PHASE 2]

apply the code to rank the urls by relevance and also filter the keywrods that show up to make sure they are related to the SDG topic

[DASHBOARD]

######################## DOCUMENTATION FOR PROFESSOR AND RESULTS

\*\*\*This applies to only projects that interact with Professor

\*\*Updated required weekly <Wed

[Documentation\_Outlier Detection](https://docs.google.com/document/d/1VHCdOMBJUSVwbClXptYEafVCRp9ym7yc8TF54xGcNls/edit)

/////// Update 07/30

**Outline:**

* **Method 1**: absolute value
* **Method 2**: absolute daily changes
* [**Method 3**: Combine Method 1(absolute STS) and Method 2 (absolute daily change in STS)](#kix.xts5vc485c37)
* [Once a day, once a week, once a month cutoff](#kix.m1svkeejqbau)
* [**Clustering outlier by SDG**](#kix.dnbrd7u7ksw)

**Method 1**: absolute value

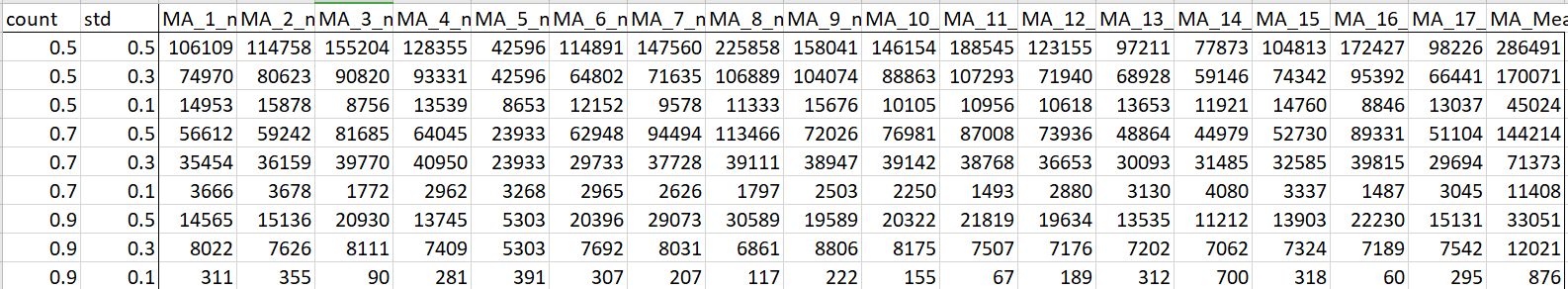
1. Absolute value of SDG STS > 3, 3.5, 4
2. The news volume is larger than certain quantiles and the standard deviation is smaller than certain quantiles. Quantile = 0.5, 0.7, 0.9 for news volumes and quantile = 0.5, 0.3, 0.1 for standard deviation.
3. Calculate the percentage: number of outliers in each quantile combination/total observations of STS by SDG

Total observations of STS by SDG：



1. Calculate the percentage in respective bins : number of outliers in each quantile combination/total observations of STS within respective bins

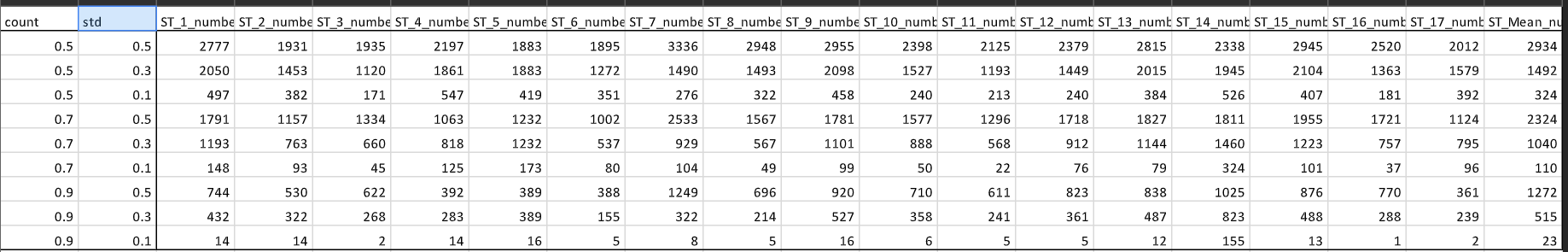
Total observations of STS within respective bins:



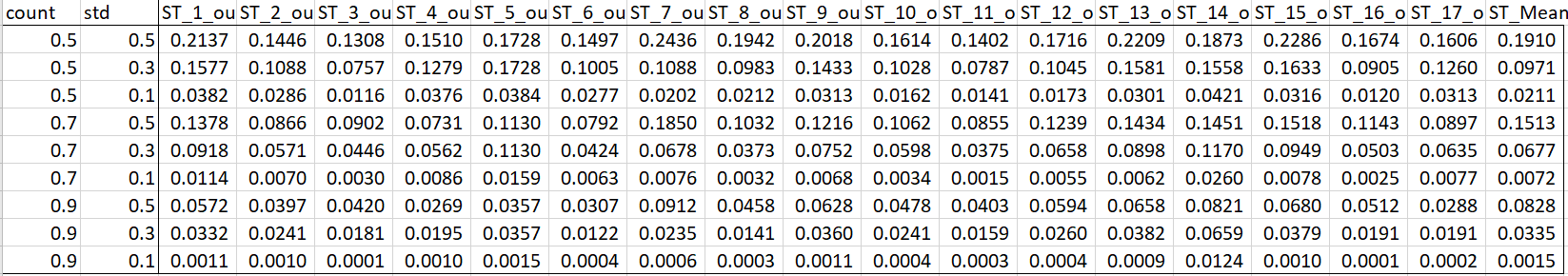
Below are the number of outliers under each quantile and its percentage

|STS| >3:

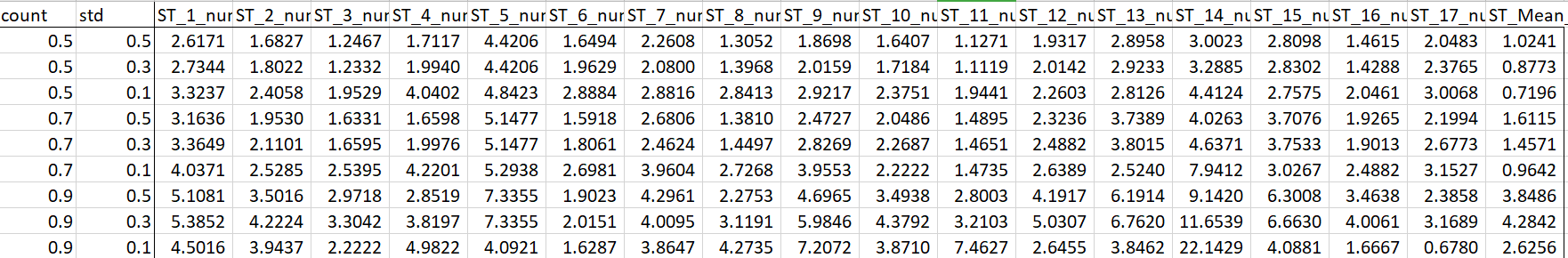
Number of outliers:



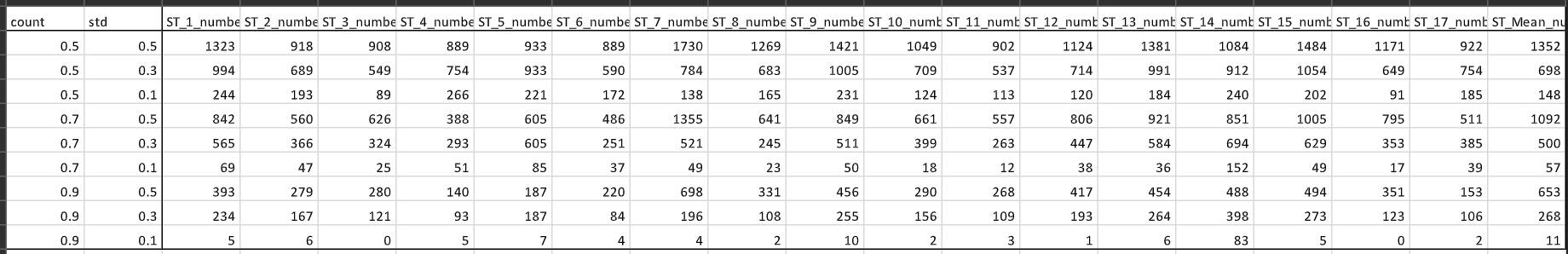
Percentage: in (%)



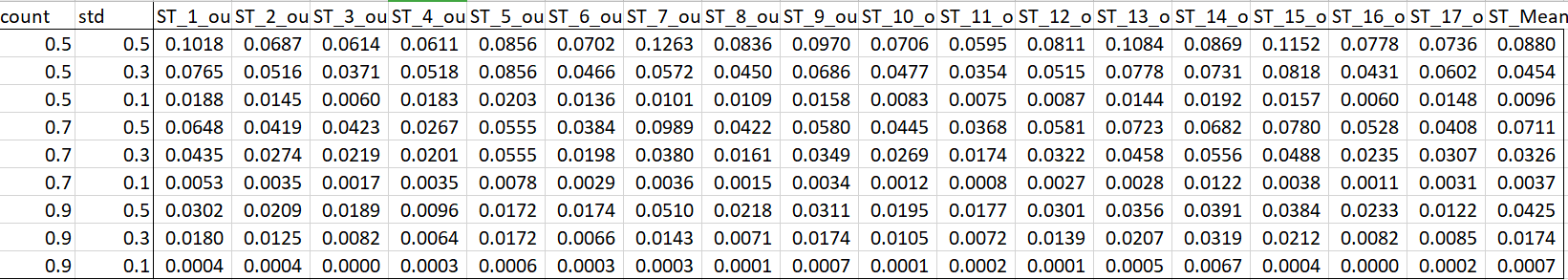
Percentage in respective bins: in (%)



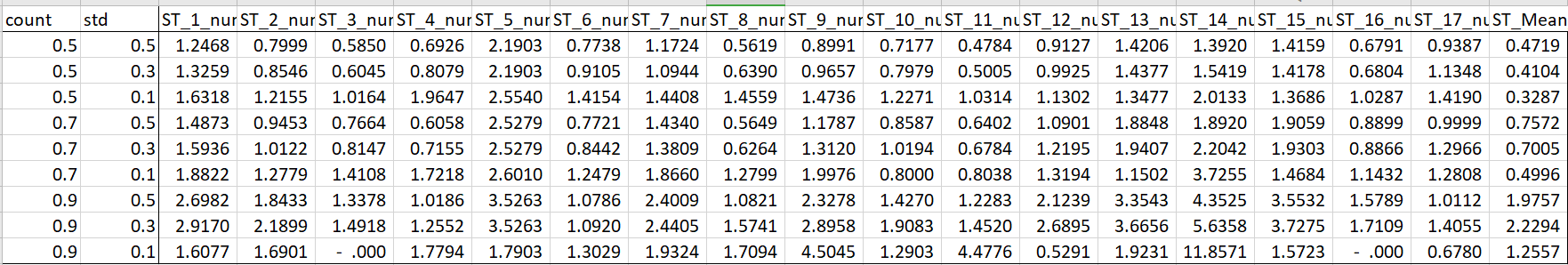
|STS| > 3.5:



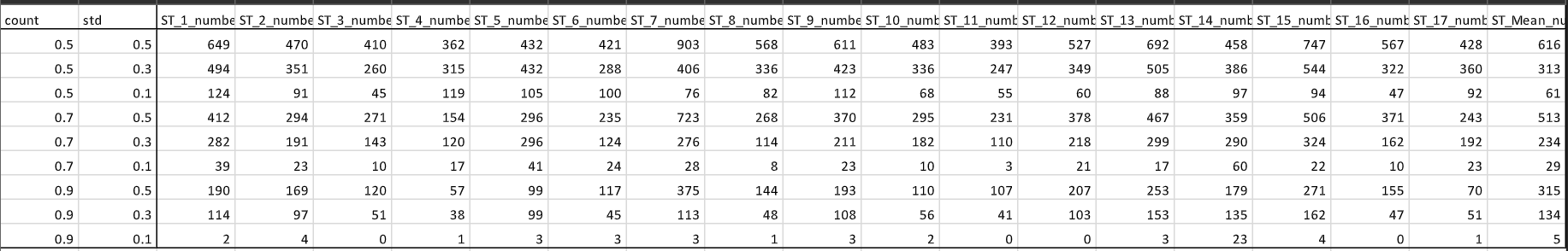
Percentage: in (%)



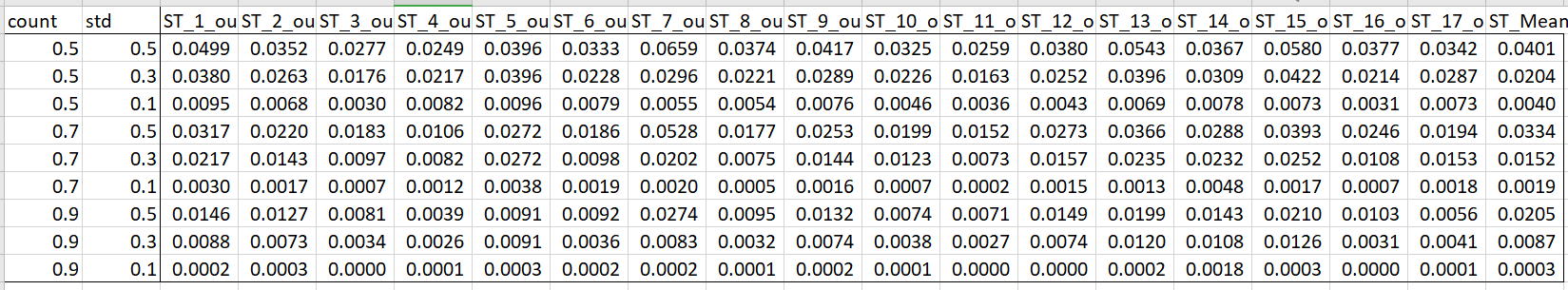
Percentage in respective bins: in (%):



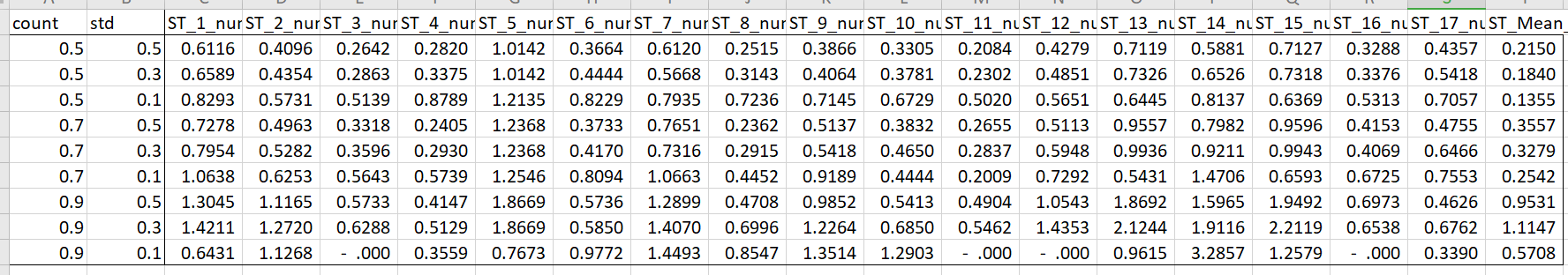
|STS| > 4:

****

Percentage: in (%)



Percentage in respective bins: in (%):



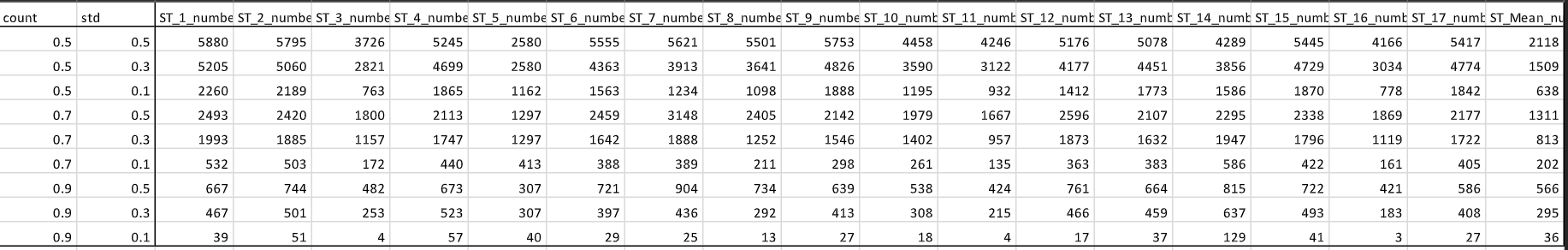
**Method 2**: absolute daily changes

1. Absolute daily change of SDG STS. Diff = | today STS - yesterday STS| > 1, 1.5
2. The news volume is larger than certain quantiles and the standard deviation is smaller than certain quantiles. Quantile = 0.5, 0.7, 0.9 for news volumes and quantile = 0.5, 0.3, 0.1 for standard deviation.

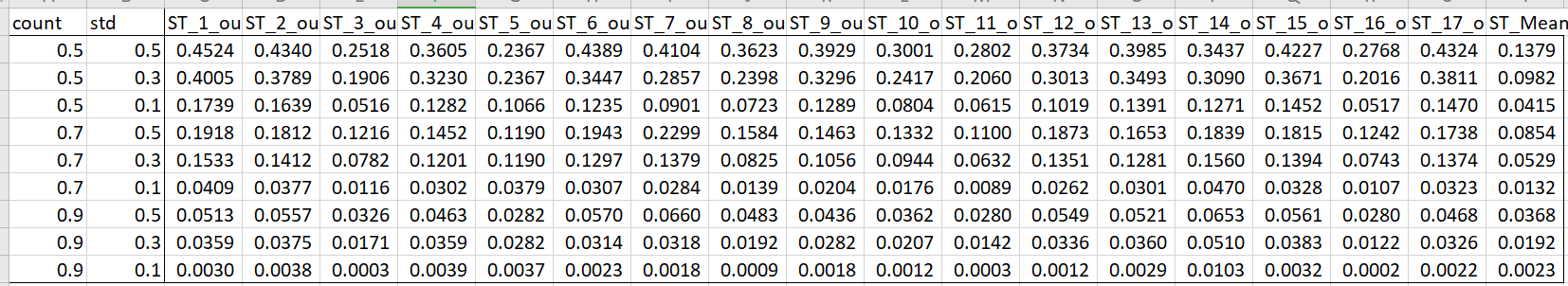
Below are the number of outliers under each quantile and its percentage

|Diff| > 1:

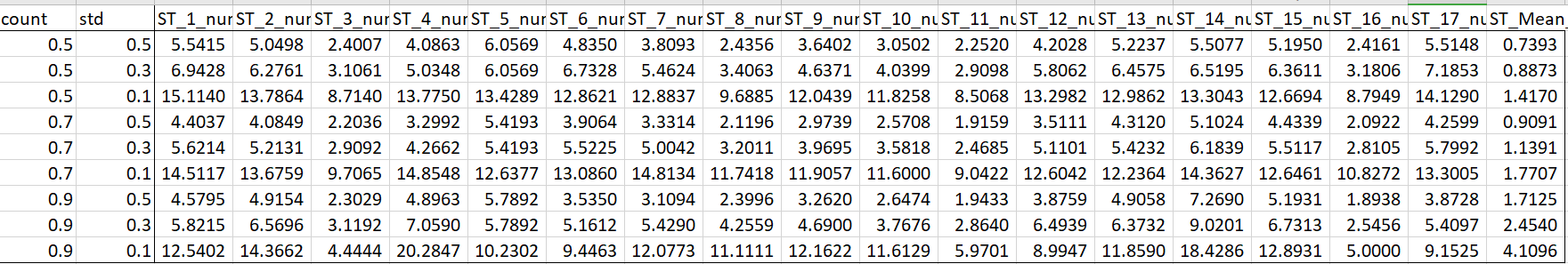
Number of outliers:



Percentage: in (%)

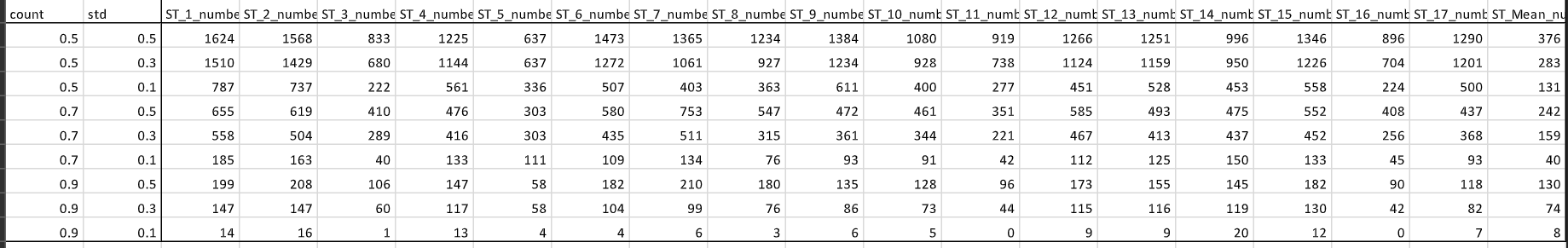


Percentage in respective bins: in (%):



|Diff| > 1.5:

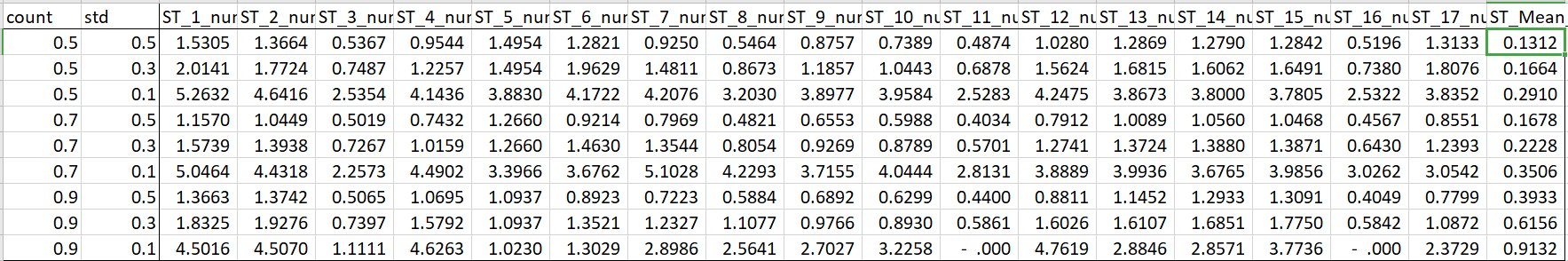
Number of outliers:



Percentage: in (%)



Percentage in respective bins: in (%):

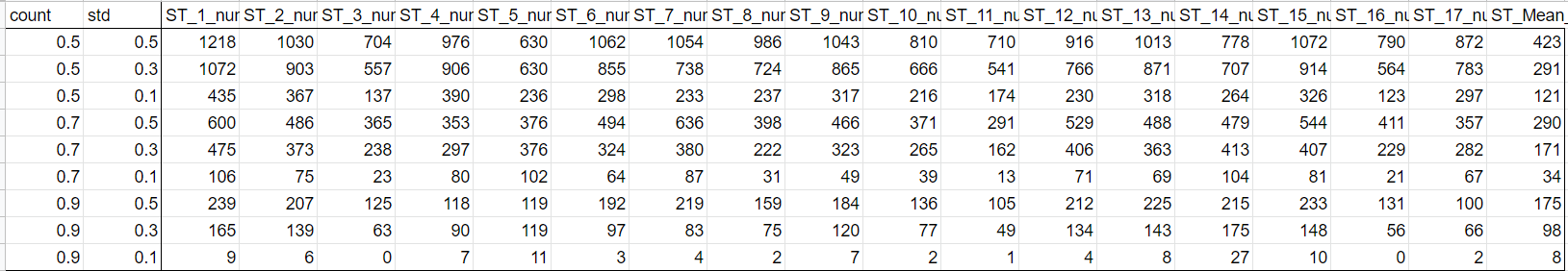


**Method 3**: Combine Method 1(absolute STS) and Method 2 (absolute daily change in STS)

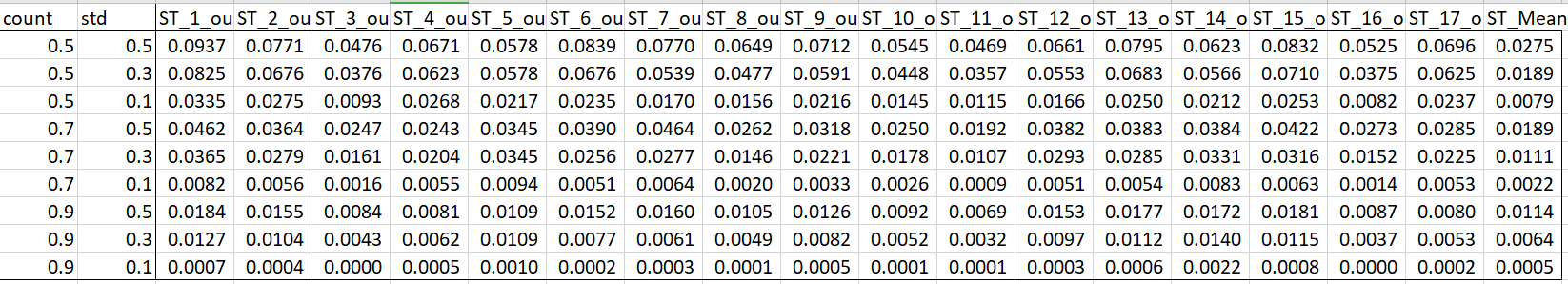
1. Absolute value of SDG STS > 2.5, 3. And, Absolute daily change of SDG STS. Diff = | today STS - yesterday STS| > 1, 1.5
2. The news volume is larger than certain quantiles and the standard deviation is smaller than certain quantiles. Quantile = 0.5, 0.7, 0.9 for news volumes and quantile = 0.5, 0.3, 0.1 for standard deviation.
3. Calculate the percentage: number of outliers in each quantile combination/total observation in each SDG.

Below are the number of outliers under each quantile and its percentage

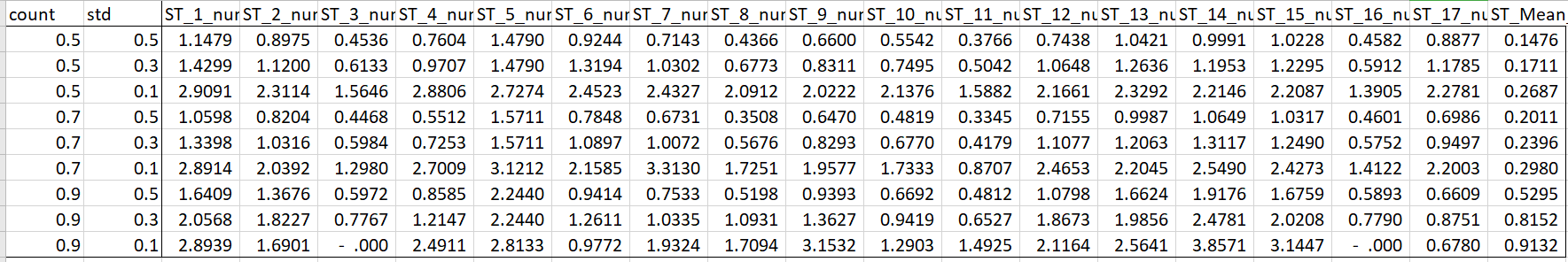
|STS| =2.5 & |Diff| = 1:



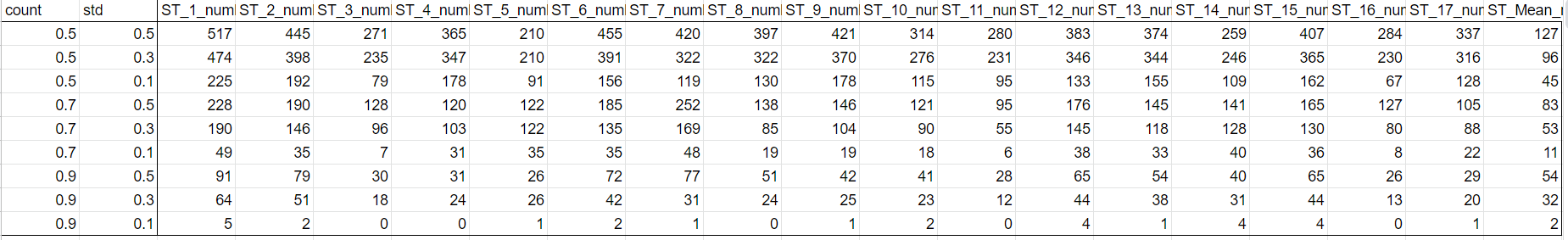
Percentage: in (%)



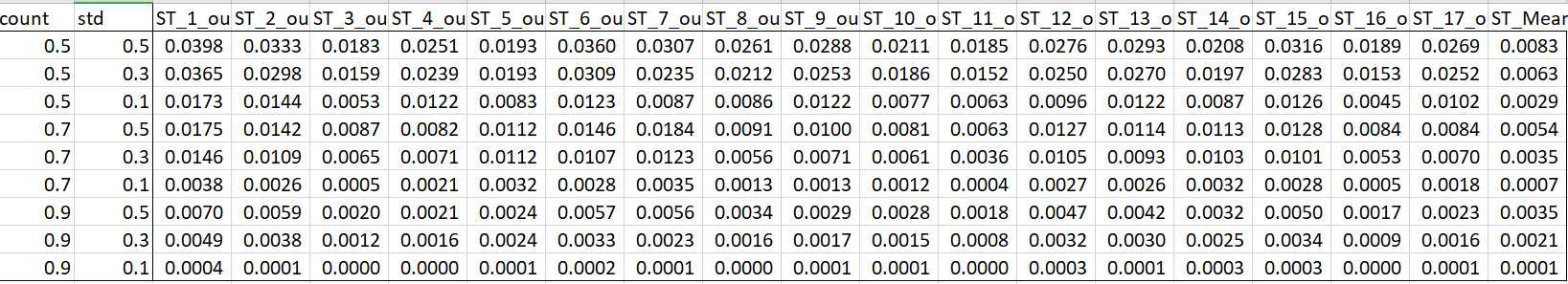
Percentage in respective bins: in (%):



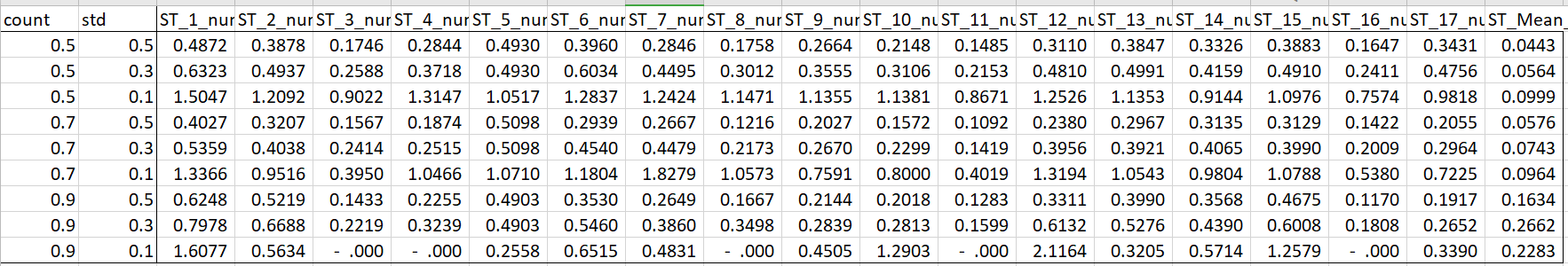
|STS| =2.5 & |Diff| = 1.5:



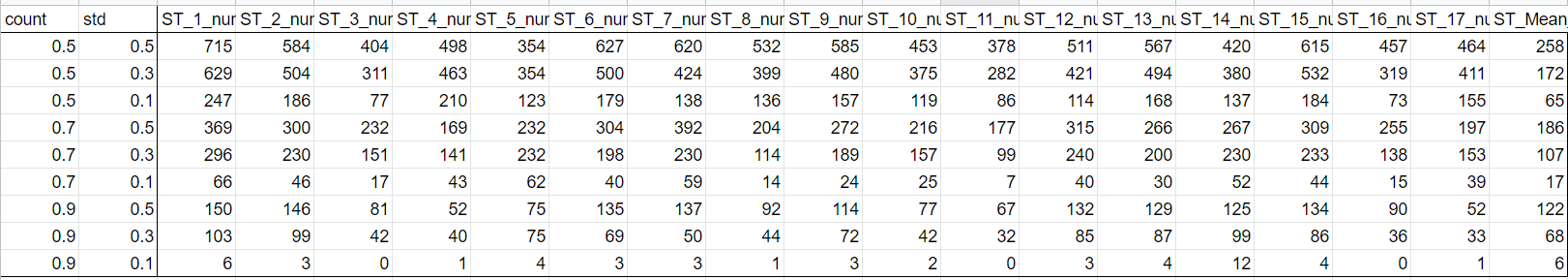
Percentage: in (%)



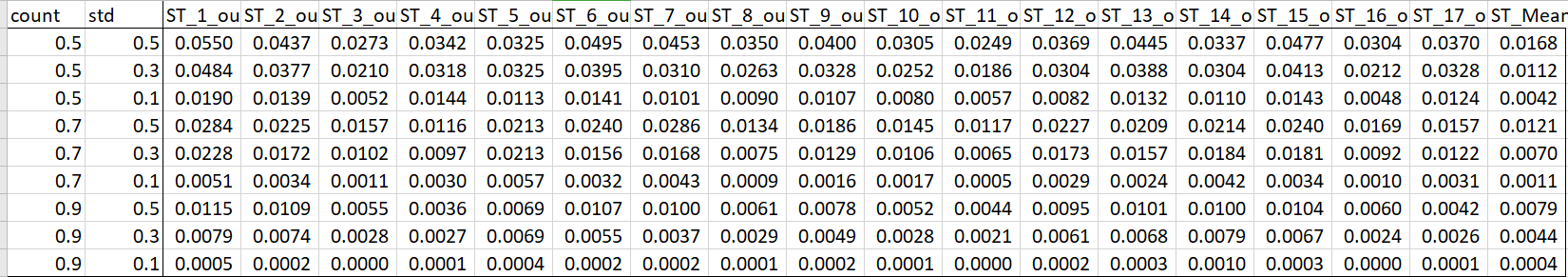
Percentage in respective bins: in (%):



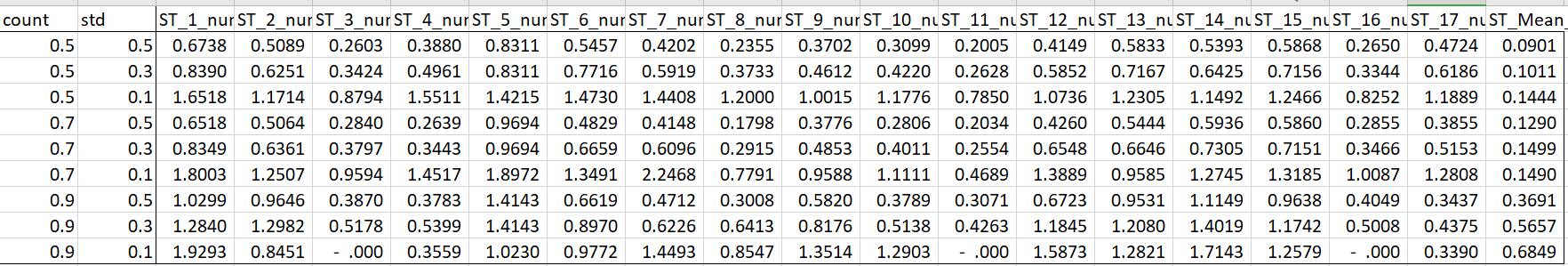
|STS| = 3 & |Diff| = 1:



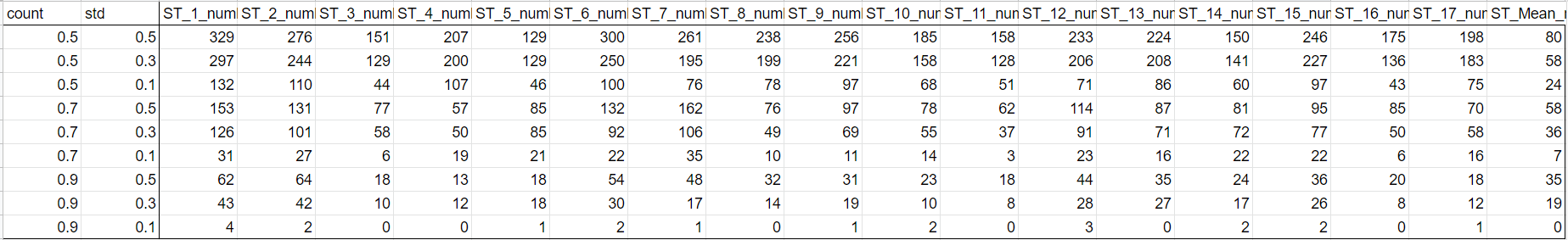
Percentage: in (%):



Percentage in respective bins: in (%):



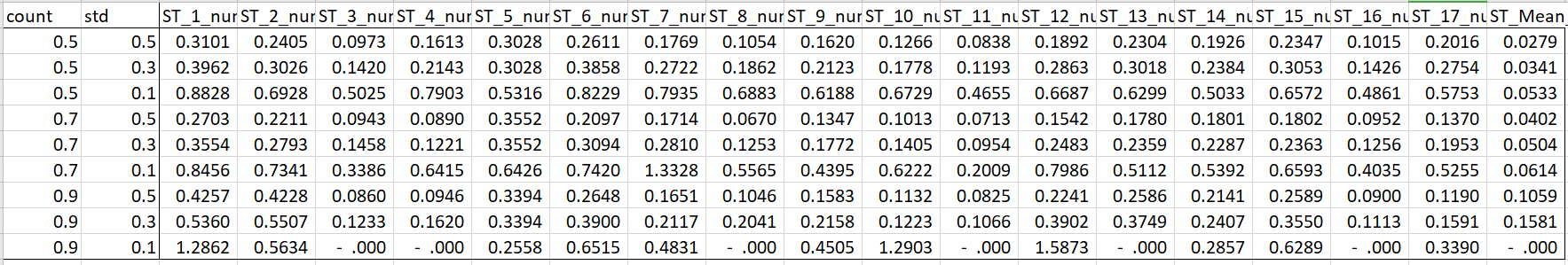
|STS| = 3 & |Diff| = 1.5:



Percentage: in (%):



Percentage in respective bins: in (%):



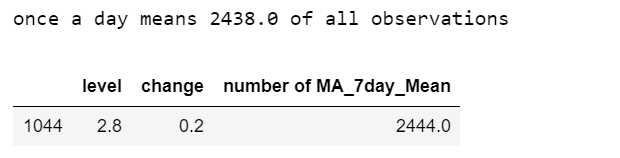
**Once a day, once a week, once a month cutoff**

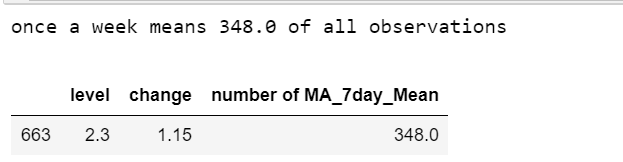
Steps:

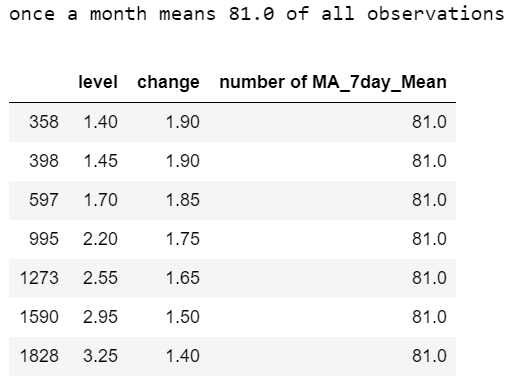
1. Restrict sample to certain ncount and standard deviation quantile, 0.5 for each, or 0.7 and 0.3.
2. Determine level and change cut-off to hit the described percentage.

“once a day” means 1/630 = 0.15% of all observations, “once a week” is 1/(630\*7), “once a month” is 1(630\*30) = 0.0053%. we have about 1,535,992 observations, so the ideal number for “once a day” is 0.15% \* 1,535,992 = 2438, ideal number of “ once a week” is 348, ideal number of “once a month” is 81.

After restricting the sample by ncount > 0.5 quantile and standard deviation < 0.5. I started from a wide range for grid search. I set levels from 1 to 3.5 with 0.05 increment, which includes 1.5, 1.55, 1.6, …, 3.45, 3.5 and set change from 0 to 2 with 0.05 increment. There are 2000 combinations in total. The combinations closest to the desired number are listed below,

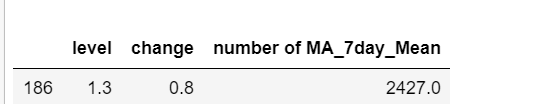




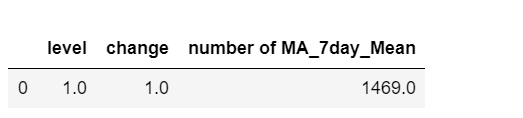


For the “once a day”, daily change = 0.2, might be too small. Below are some other results, after I setting the lower border of change to 0.5 and 1

change from 0.5 to 2 with 0.05 increment:



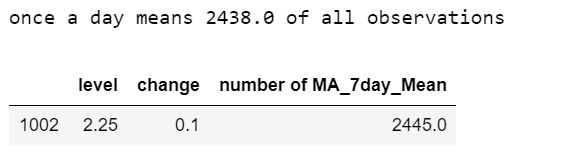
change from 1 to 2 with 0.05 increment

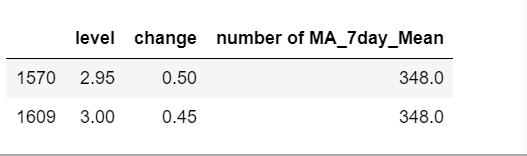


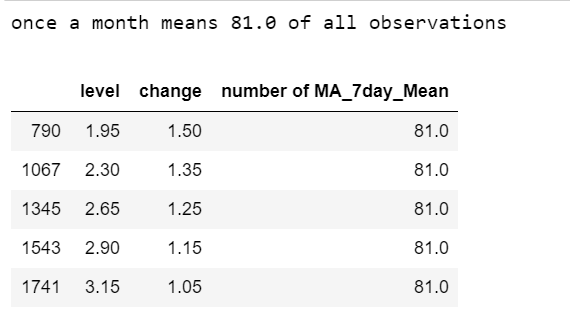
It seems change = 1.0 is too large to hit the “once a day”?

I also repeated the same steps ncount > 0.7 quantile and standard deviation < 0.3.

I started from a wide range for grid search. I set levels from 1 to 3.5 with 0.05 increment, which includes 1.5, 1.55, 1.6, …, 3.45, 3.5 and set change from 0 to 2 with 0.05 increment. There are 2000 combinations in total. The combinations closest to the desired number are listed below,

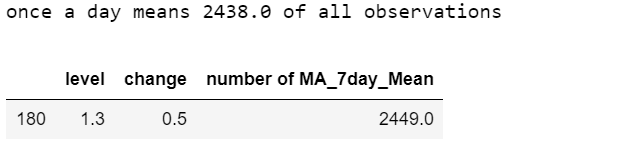


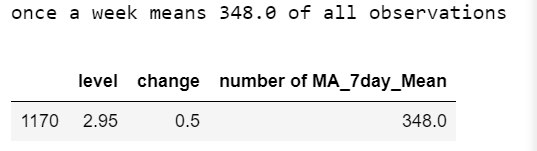




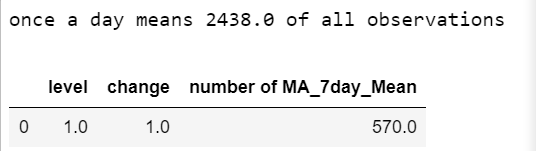
Results after I setting the lower border of change to 0.5 and 1:

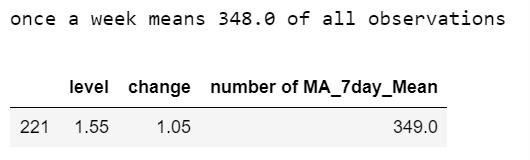
change from 0.5 to 2 with 0.05 increment:





change from 1 to 2 with 0.05 increment:





Also, you can find number of observation of all combinations through the link below,

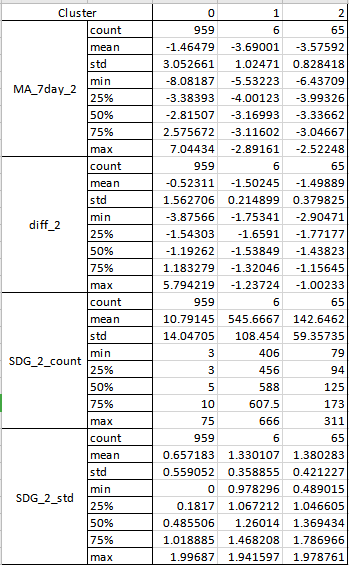
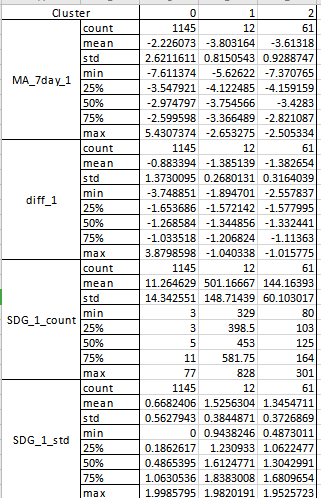
[observation with quantile = (0.7,0.3), level = (1, 3.5, 0.05), change = (0, 2, 0.05)](https://docs.google.com/spreadsheets/d/1n3rtZSgMT9vX5857BC6rwarGDl0keCsWZo8ku2Dr9W4/edit?usp=sharing)

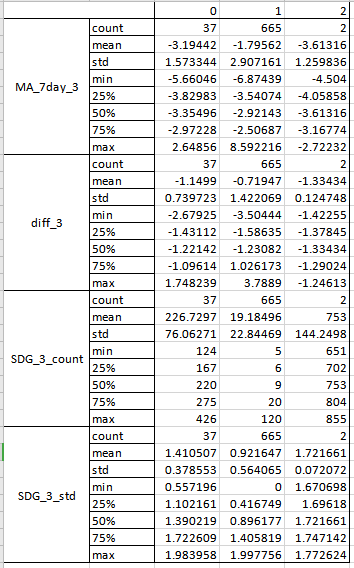
[observation with quantile = (0.5,0.5), level = (1, 3.5, 0.05), change = (0, 2, 0.05)](https://docs.google.com/spreadsheets/d/1aW465HO-KlNaR8LBWNi6puEU6kPFUUyrf_Y1TkTVxmw/edit?usp=sharing)

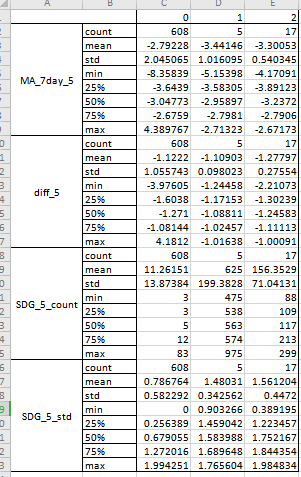
**Clustering outlier by SDG**

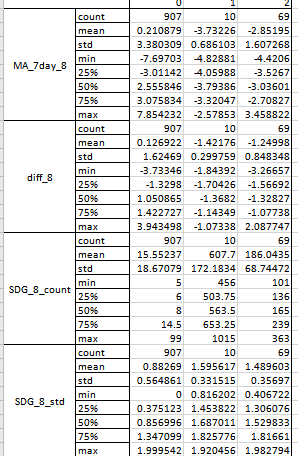
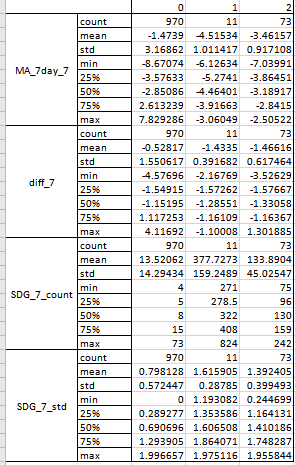
**Without Sector**

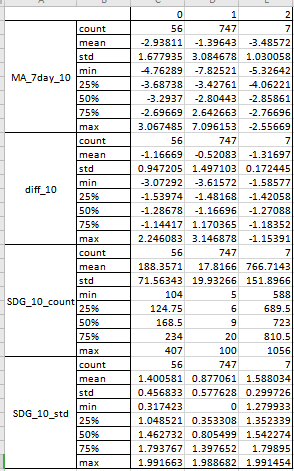
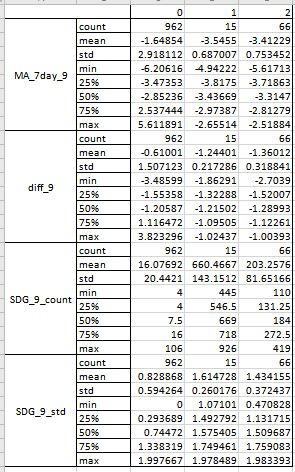
1. Absolute value of SDG STS > 2.5. And, Absolute daily change of SDG STS. Diff = | today STS - yesterday STS| > 1
2. The news volume is larger than 0.5 quantiles and the standard deviation is smaller than 0.5 quantiles.
3. Apply K-Means Clustering on level, diff, news volume, and standard deviation by SDG

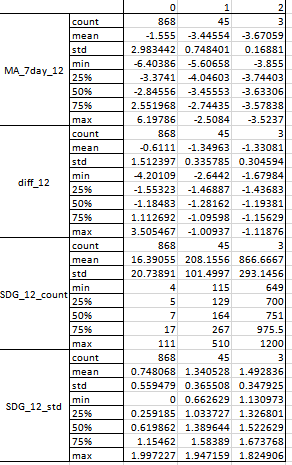
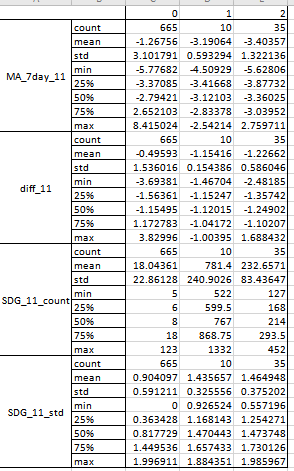


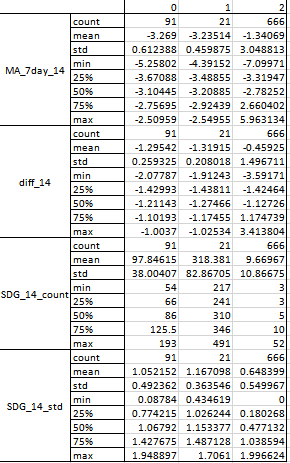
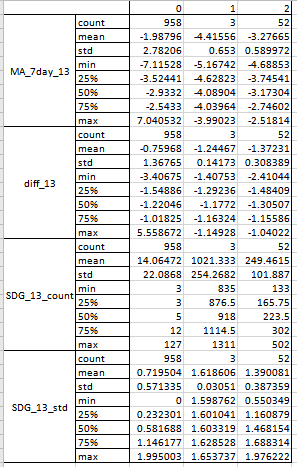


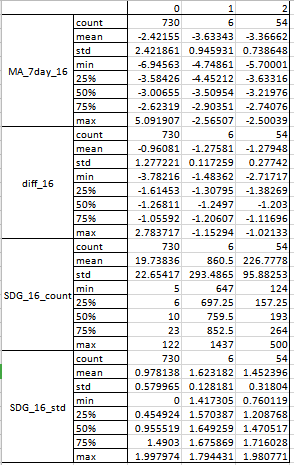
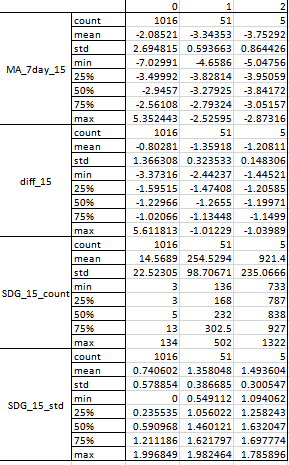


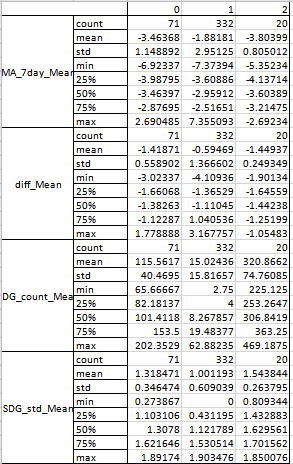
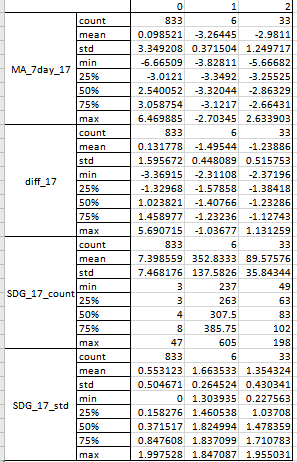










****

**With Sector**

1. Absolute value of SDG STS > 2.5. And, Absolute daily change of SDG STS. Diff = | today STS - yesterday STS| > 1
2. The news volume is larger than 0.5 quantiles and the standard deviation is smaller than 0.5 quantiles.
3. Apply K-Means Clustering on level, diff, news volume, and standard deviation, GICS Sector by SDG

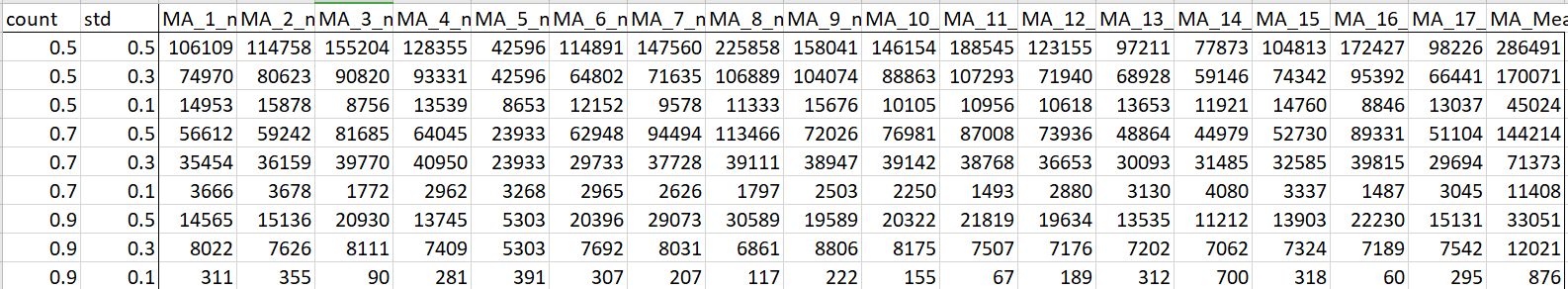
Note: the result don’t have much difference with Without Sector

Other Analysis:

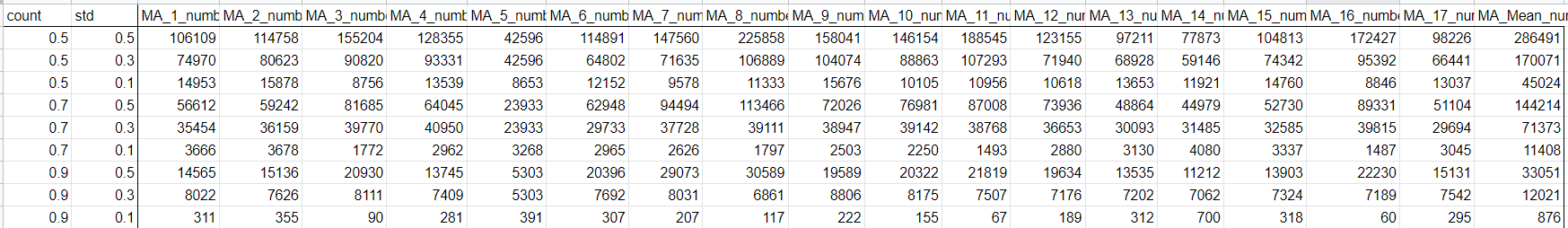
Total observations of STS by SDG：



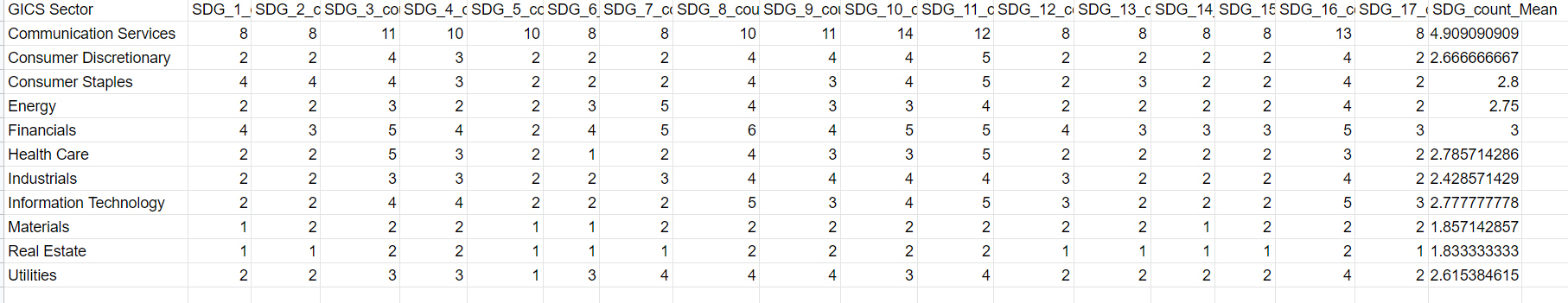
Total observations of STS within respective bins:



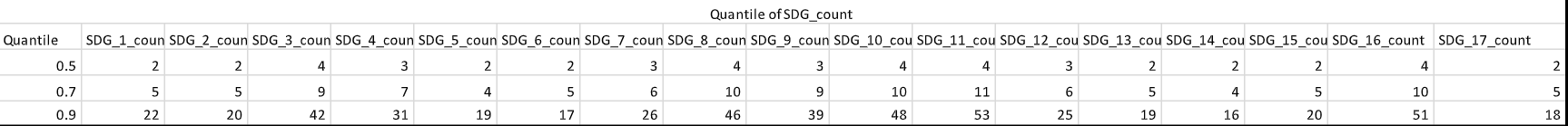
Number of STS under each quantile of news volume and standard deviation:



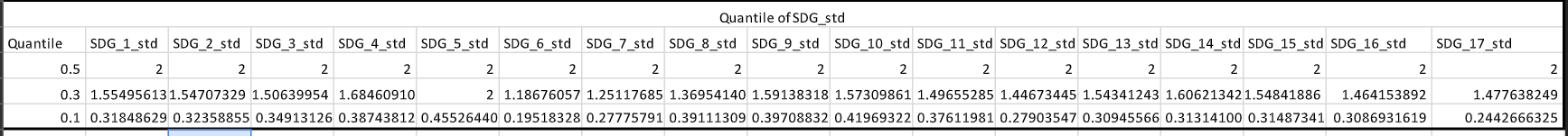
Median of news volume by Sector:



Quartile of News Volume by SDG:



Quantile of standard deviation by SDG:

****

total observations in each SDG.

######################## TECHNICAL DOCUMENTATION w CODE SCREENSHOTS

\*\*\*This needs to include a Workflow DIAGRAM with major key steps

\*\*\*this will be copy/pasted in the GITHUB

**Outline:**

* Outlier detection
* Word cloud
  + Rank
  + unranked

**Subsection - outlier detection**

You can find more information about outlier detection in the below doc.

[Documentation\_Outlier Detection](https://docs.google.com/document/d/1VHCdOMBJUSVwbClXptYEafVCRp9ym7yc8TF54xGcNls/edit)

SDG STS

Level & change by certain thresholds

Counts and std thresholds

Outliers

Path: \Users\mac\Desktop\Dropbox\SDG\outlier\_Fiona\MXUS

*Note: All files also unde the sample folder in google drive*

* Outlier. ipynb

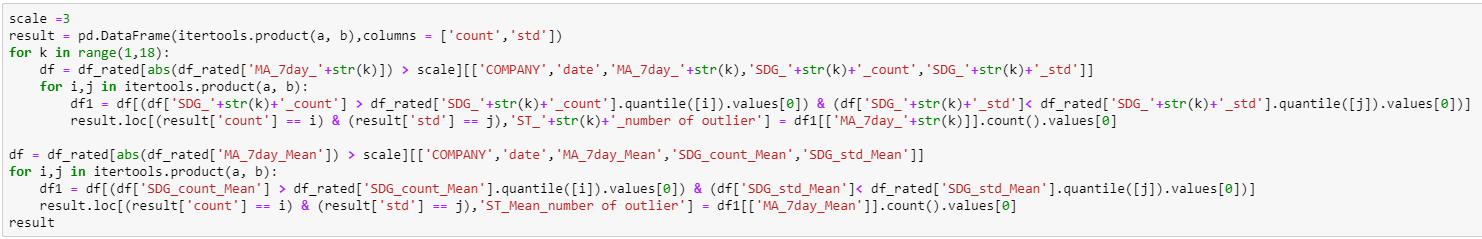
Method 1: absolute value

1. set the quantile of count and standard deviation



* 1. Create a result dataframe
  2. Iterate each SDG and for each SDG filter its absolute STS (MA\_7day) based on the scale, it returns df
  3. For each quantile combination, filter the df based on the quantile, it returns df1, then the number of outliers is the number of non-missing values of the STS, stored the it back to result dataframe
  4. Get the aggregate table in result dataframe

Note: change the scale if using other absolute values.

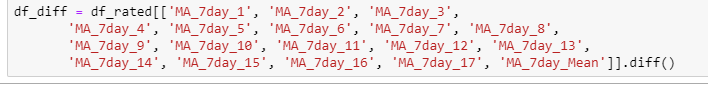


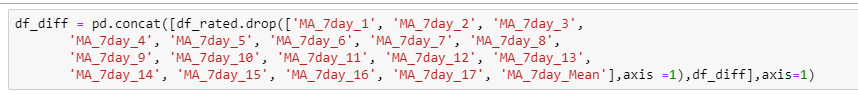
Method 2: absolute daily changes

1. set the quantile count and standard deviation



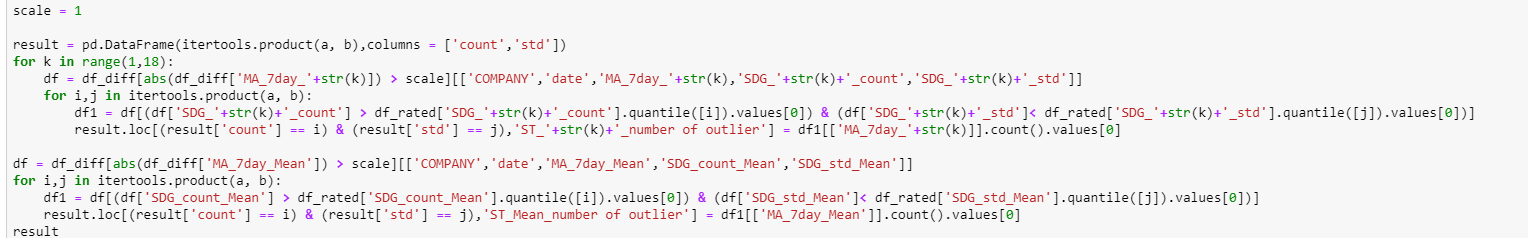
1. Create a df with absolute daily changes in STS and add news volume and std back





1. Create a result dataframe
   1. Iterate each SDG and for each SDG filter its absolute daily changes of STS (MA\_7day) based on the scale, it returns df
   2. For each quantile combination, filter the df based on the quantile, it returns df1, then the number of outliers is the number of non-missing values of the STS, stored the it back to result dataframe
   3. Get the aggregate table in result dataframe

Note: change the scale if using other absolute values

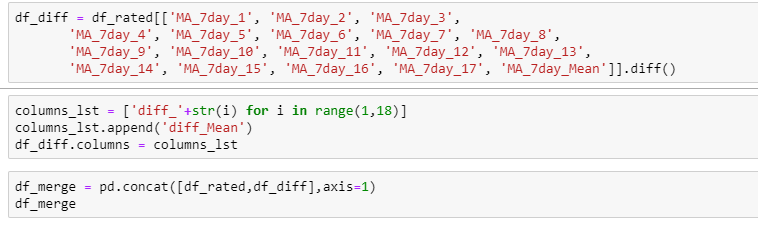


Method 3: Combine Method 1(absolute STS) and Method 2 (absolute daily change in STS)

1. set the quantile count and standard deviation

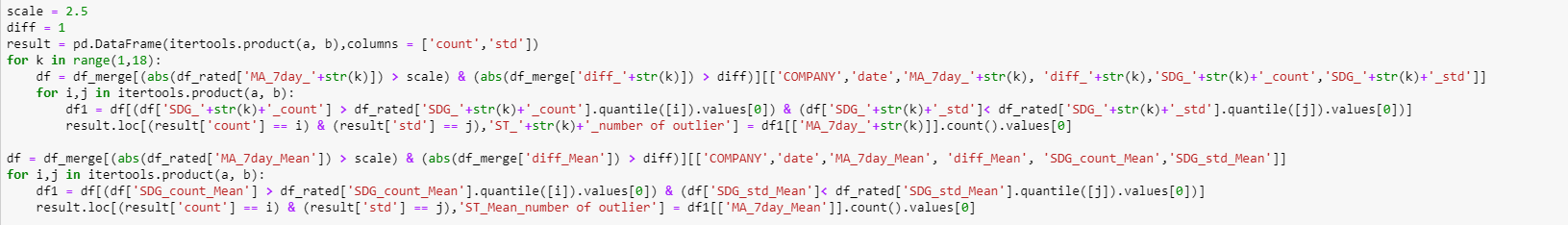


1. Create a merge dataframe with absolute daily change



1. Create a result dataframe
   1. Iterate each SDG and for each SDG filter its absolute STS (MA\_7day) based on the scale and absolute daily changes of STS (MA\_7day) based on the scale , it returns df
   2. For each quantile combination, filter the df based on the quantile, it returns df1, then the number of outliers is the number of non-missing values of the STS, stored the it back to result dataframe
   3. Get the aggregate table in result dataframe

Note: change the scale and diff with respective value



# **subsection - word cloud**

# **Outline**

* Summary
* Workflow
* Model
  + Input
  + Path: on old mac: 1 384 900 708
  + Output
  + Steps & Code Screenshot
* Others
  + Automation the Raw Sentiment Score – using Crontab

**Summary**

This project works on generating word clouds of News title according to SDG score breaking days (dramatic changes in SDG scores for a particular company). Key words of News related to SDG can help us understand what happened in these days that caused SDG changes.

**Workflow**

Find outliers

Ranked (semantic comparison)

World Cloud

& time series plots &

unranked

**Model:**

**Step1: find outliers**

Relevant files & paths:

1. find SDG outliers – all

* filename: find SDG outliers – all. jpyter
* path: /Users/macglobalai/Documents/WorldCloud\_Fiona/find outliers
* output: SDG outliers list – all.csv

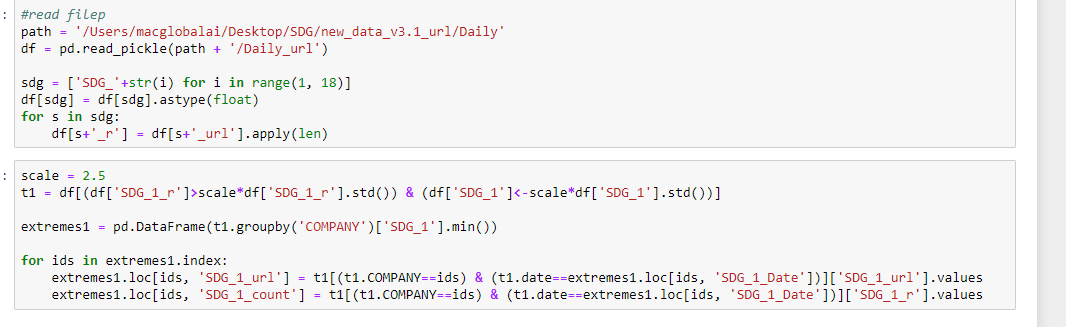
1. find SDG outliers – specific

* filename: find SDG outliers – specific.jpyter
* path: /Users/macglobalai/Documents/WorldCloud\_Fiona/find outliers
* output: all files under ‘SDG outliers list – specific’

Steps & screenshot:

Find outliers based on standard deviation and number of news

If the sdg score is large than scale \* the standard deviation and the title lens large than threshold, we recognize it as outliers.



**Step2: Word Cloud**

**Approach 1: Ranked**

Relevant files & paths: (SDG outliers – all)

*Note: All files also unde the sample folder in google drive*

* path: /Users/macglobalai/Documents/WordCloud\_Fiona/outlier/no rank/ SDG outliers\_all (ranked)/
* filename: SDG outliers\_all (ranked).jupyter

Input and output:

|  |  |  |  |
| --- | --- | --- | --- |
| function | input |  | output |
| similarity\_scores | df\_url | '/Users/macglobalai/Desktop/SDG/new\_data\_v3.1\_url/Daily/Daily\_url' | ranked URLs and word clouds |
| Df\_theme | '/Users/macglobalai/Documents/WordCloud\_Fiona/df\_theme.csv' |
| company | e.g. 'pfizer inc' |
| SDG\_Value | e.g. -56.83 |
| date | e.g. ‘2016-05-14’ |
| SDG | e.g. 1 |
| timeseries\_SDG | df\_url, company, start\_date, end\_date | Same as above | Time series SDG |

Relevant files & paths (SDG outliers – specific)

* path: /Users/macglobalai/Documents/WordCloud\_Fiona/outlier/no rank/ SDG outliers\_specific (ranked)/
* filename: SDG outliers\_specific (ranked).jypter
* output: all results under file called ‘SDG outliers\_specific (ranked)\_results’

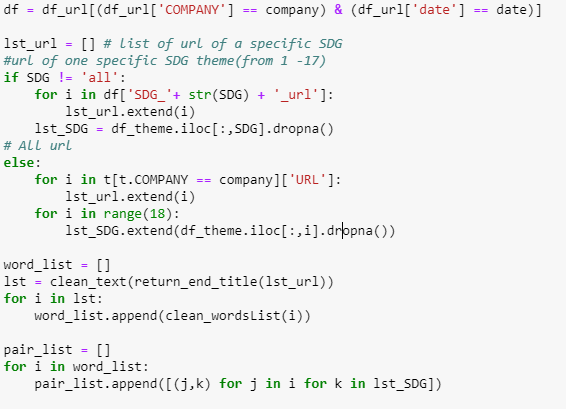
Main function:

|  |  |  |  |
| --- | --- | --- | --- |
| function | input |  | output |
| similarity\_scores | df\_url | '/Users/macglobalai/Desktop/SDG/new\_data\_v3.1\_url/Daily/Daily\_url' | ranked URLs and word clouds |
| Df\_theme | '/Users/macglobalai/Documents/WordCloud\_Fiona/df\_theme.csv' |
| company | e.g. 'pfizer inc' |
| SDG\_Value | e.g. -56.83 |
| date | e.g. ‘2016-05-14’ |
| SDG | e.g. ‘all’ |
| timeseries\_SDG | df\_url, company, start\_date, end\_date | Same as above | Time series SDG |

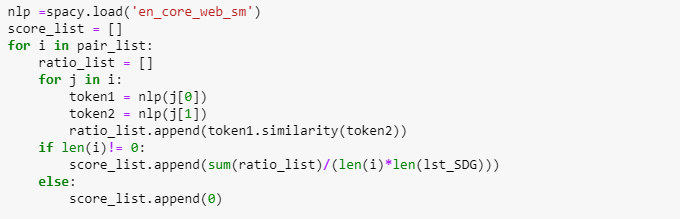
Steps and screenshots:

1. rank URLs by semantic similarities

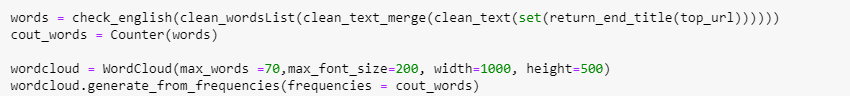
1.1. Combine the one word from key themes on SDG topic and one word from URL.

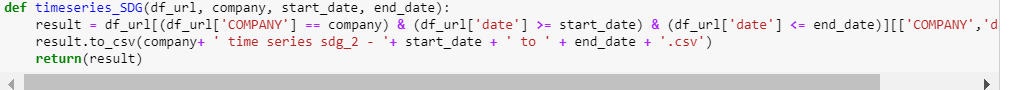


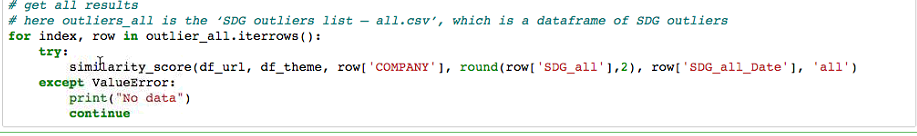
1.2. Calculate word by word similarity score for each combination by spaCy library in python and get an aggregate score for each url.



1.3. Rank URLs by the scores.

2. extract the key words from top ranked URLs

3. generate the 3 deliverables for each SDG outliers, including word cloud, time series, urls list

4. Using the loop to iterate each rows of ‘SDG outliers list – all.csv’ and use the values as the input of similarity\_scores function. <same steps for timeseries\_SDG function>

**Approach 2: No Rank**

Extract the key words from all URLs. generate the 3 deliverables for each SDG outliers, including word cloud, time series, urls list

Relevant files & paths (SDG outliers – all)

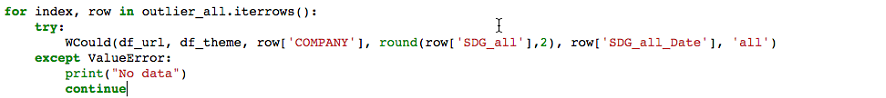
* path: /Users/macglobalai/Documents/WordCloud\_Fiona/outlier/no rank/ SDG outliers\_all (no rank)/
* filename: SDG outliers\_all (no rank).jypter
* output: all results under file called ‘SDG outliers\_all (no rank)\_results’

Input& Output:

|  |  |  |  |
| --- | --- | --- | --- |
| function | input |  | output |
| WCould | df\_url | '/Users/macglobalai/Desktop/SDG/new\_data\_v3.1\_url/Daily/Daily\_url' | ranked URLs and word clouds |
| Df\_theme | '/Users/macglobalai/Documents/WordCloud\_Fiona/df\_theme.csv' |
| company | e.g. 'pfizer inc' |
| SDG\_Value | e.g. -56.83 |
| date | e.g. ‘2016-05-14’ |
| SDG | e.g. ‘all’ |
| timeseries\_SDG | df\_url, company, start\_date, end\_date | Same as above | Time series SDG |

*Using loop to get all results:*

Using the loop to iterate each rows of ‘SDG outliers list – all.csv’ and use the values as the input of WCould function. <same steps for timeseries\_SDG function>



Relevant files & paths (SDG outliers – specific)

* path: /Users/macglobalai/Documents/WordCloud\_Fiona/outlier/no rank/ SDG outliers\_specific (no rank)/
* filename: SDG outliers\_specific (no rank).jypter
* output: all results under file called ‘SDG outliers\_specific (no rank)\_results’

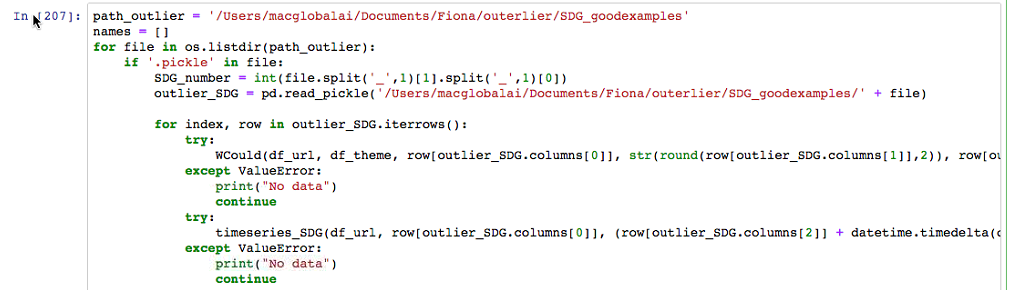
Input & Output:

|  |  |  |  |
| --- | --- | --- | --- |
| function | input |  | output |
| WCould | df\_url | '/Users/macglobalai/Desktop/SDG/new\_data\_v3.1\_url/Daily/Daily\_url' | ranked URLs and word clouds |
| Df\_theme | '/Users/macglobalai/Documents/WordCloud\_Fiona/df\_theme.csv' |
| company | e.g. 'pfizer inc' |
| SDG\_Value | e.g. -56.83 |
| date | e.g. ‘2016-05-14’ |
| SDG | e.g. 1 |
| timeseries\_SDG | df\_url, company, start\_date, end\_date | Same as above | Time series SDG |

Steps & Screenshots:

*Using loop to get all results:*

Using the loop to iterate each rows of each file under ‘SDG outliers list – specific’ and use the values as the input of WCould function and timeseries\_SDG function



######################## DASHBOARD SPECIFICATIONS

//// FOR EACH DASHBOARD:

Path for Plotly Python File:

Path for Input data (specify computer/full path /MongoDB): In ryver

SDG with url 2020 sample data: <https://www.dropbox.com/s/5puf0x5za60d921/urls_2020.csv?dl=0>

Web link: <https://gai-word-cloud.herokuapp.com/>

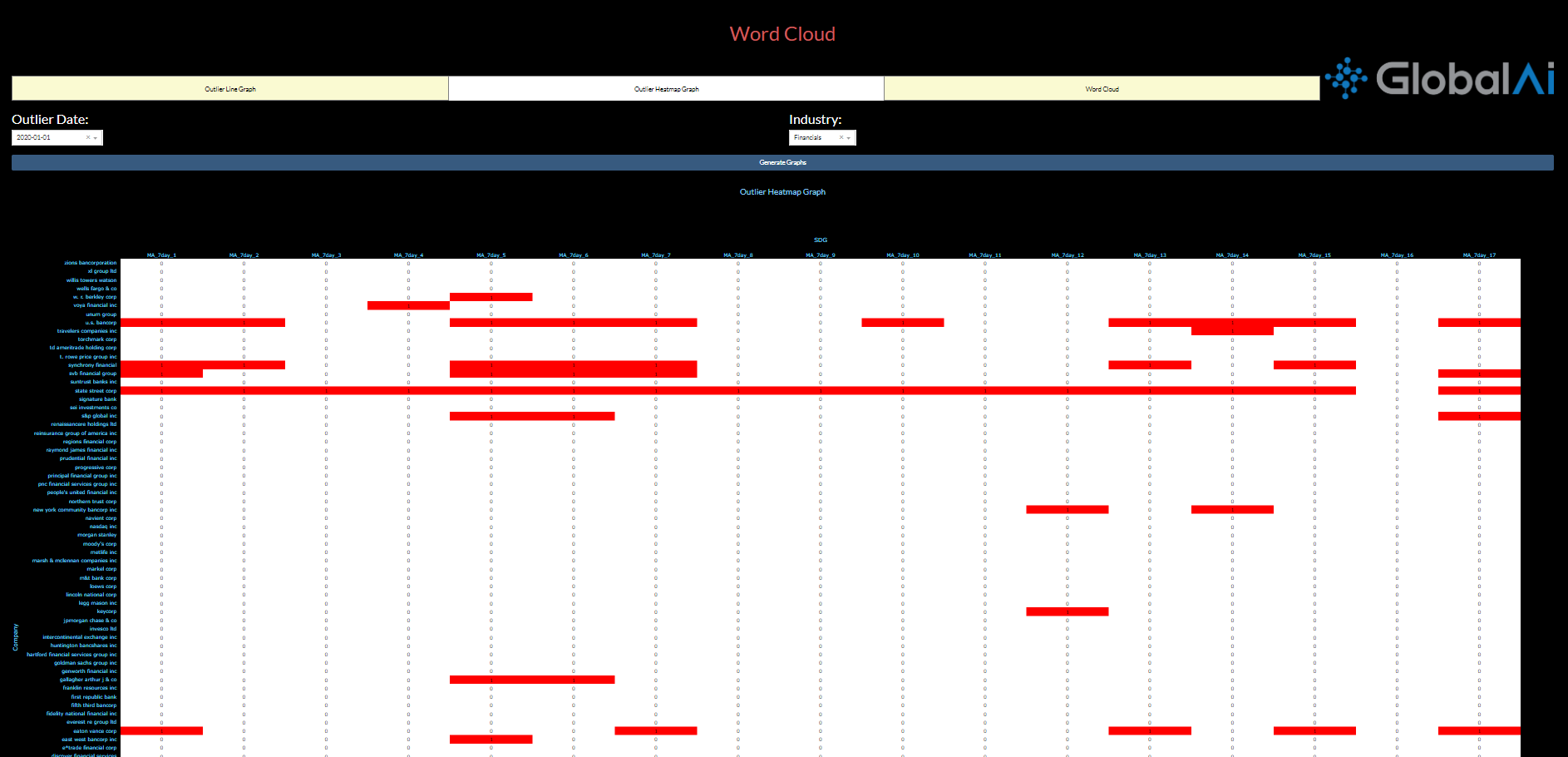
Heroku details: <https://gai-word-cloud.herokuapp.com/>

Task Scheduler (daily frequency):

Template Implementation:

/////SPECIFICATIONS

///// SCREENSHOTS OF DASHBOARDS (add date to each):

07/27

######################## TECHNICAL ISSUES & DATA DEPOSITORY

/// RYVER Channel? Last Update?

nlp-SDG+\_Outliers-Word-clouds

/// GOOGLE FOLDER

nlp-SDG+\_Outliers-Word-clouds

/// GITHUB ID? Last Update?

Word-clouds <https://github.com/globalaiorg/word-clouds>

/// Full Code Path(s)? All relevant file are also under google drive/sample code

old mac: 1 384 900 708

Word cloud: /Users/macglobalai/Documents/WordCloud\_Fiona

Outlier: https://drive.google.com/drive/folders/1bKHeOc-sU8MBPjk\_SjJJPqjpeMOwuGLO?usp=sharing

/// Full Raw Data Path?

SDG with url: <https://www.dropbox.com/s/sgia67meafs5ei4/urls.csv?dl=0>

Explain about the column: <https://www.dropbox.com/s/4aegbrtgz6p85ln/Column_Description.docx?dl=0>

SDG with url 2020 sample data: <https://www.dropbox.com/s/5puf0x5za60d921/urls_2020.csv?dl=0>

SDG with url - daily updated:

<https://www.dropbox.com/s/ro7ucgccrf73u9o/Daily_v1v2_MXUS_URL.pkl?dl=0>

/// Dashboard Link?

<https://gai-word-cloud.herokuapp.com/>

/// PRODuction data Folder

SDG with url 2020 sample data: <https://www.dropbox.com/s/5puf0x5za60d921/urls_2020.csv?dl=0>

/// PRODUCtion Code Folder

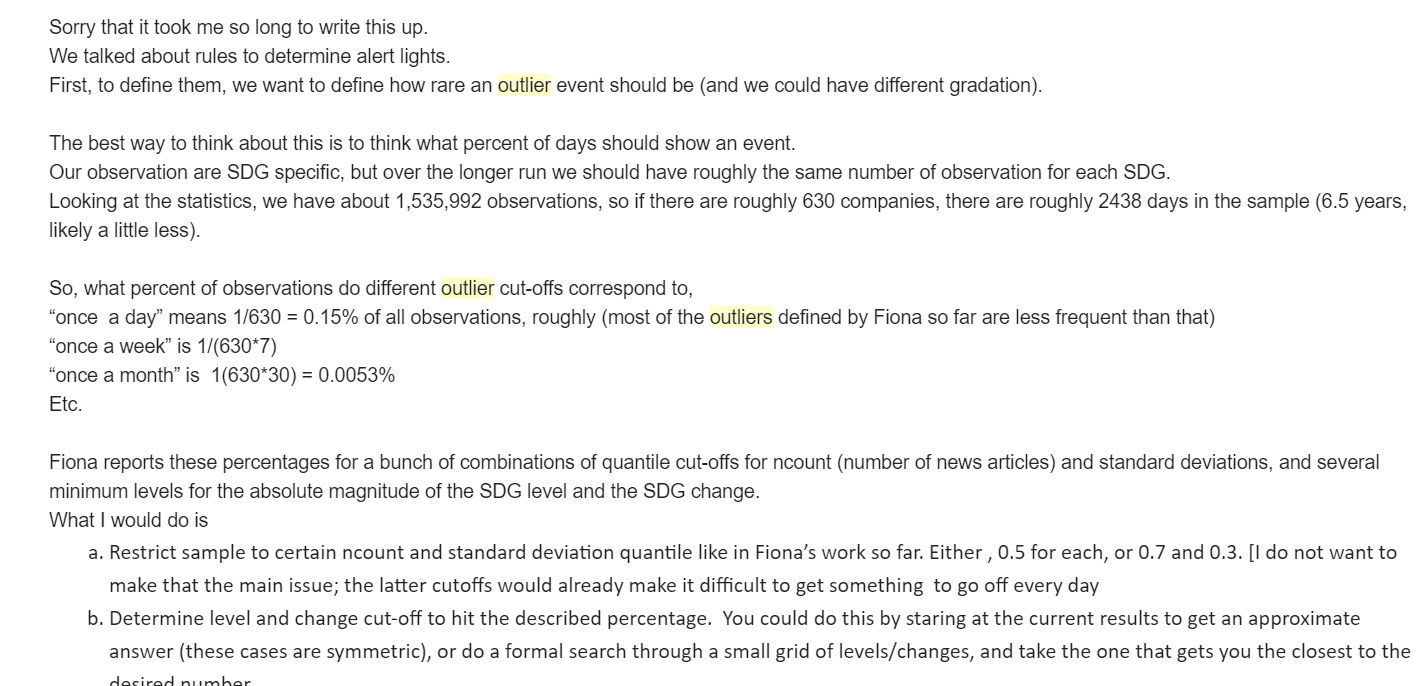
<https://drive.google.com/drive/u/2/folders/14Q-8g6LWO7gaYlUW5Uhxpp5i0Z-fH7yC>

######################## APPENDIX WITH ALL RELATED NOTES & EMAILS

\*\*Here include ALL related information from email threads; notes; etc. – even if its 100 pages long we need everything here; if file too big, then include a downloadable link there

/////// SCREENSHOTS

/////// EMAIL NOTES



//////// OTHER