10 ACADEMY

Kifiya AI Mastery Program 2

Week 1, Task 1

Interim Report

by

Jerusalem Fetene Mekonnin

August 30, 2024

Introduction

Python is a simple, clear and intuitive programming language used for many scientific and mathematical, and statistical and financial applications. Owing to this many scientists and data analysts gain valuable insights in the least amount of tine and resource while their statistical data analysis.

Data analysis is simply working with numbers and extracting valuable insights from them using descent skills of mathematics, programming and domain area working on i.e., financial, statistical, descriptive etc.

Financial analysis is the process of evaluating businesses, projects, budgets, and other finance-related transactions to determine their performance and suitability. Typically, financial analysis is used to analyze whether an entity is stable, solvent, liquid, or profitable enough to warrant a monetary investment. Stock market prediction involves trying to determine the future value of a company stock or other financial instrument traded on an exchange.

Objective

The express aim of this report is to analyze the correlations between news sentiment and stock market movements of the financial news data. Here below are some of the objectives of this report:

- To perform descriptive statistics
- To perform sentiment analysis on the 'headline' text to quantify the tone
- To perform time series analysis
- To establish statistical correlations between the sentiment derived from news articles and the corresponding stock price movements
- To perform publisher analysis

About Dataset

Financial News and Stock Price Integration Dataset

FNSPID (Financial News and Stock Price Integration Dataset), is a comprehensive financial dataset designed to enhance stock market predictions by combining quantitative and qualitative data.

Python Libraries

1.Pandas

Pandas is the open-source python library that is widely used for data analysis and data. Its main purpose is to perform data analysis on the structured data and focuses on the fundamental data processing.

2.NumPy

NumPy package comes with a wide collection of numerical functions that makes it an important library in academia and finance industry.

3. SciPy

Scipy is used for financial computation and other numerical integrations in the finance industry4.
4. Polars

Polars is a blazingly fast DataFrame library for manipulating structured data.

5. Pyfolio

It is an open-source library that provides risk analysis reports and performance results of financial portfolios based on the returns.

6. QuantPy

A framework for quantitative finance in python.

7. Statsmodel

Statsmodel is used for statistical tests and statistical data exploration. Some best models of statsmodel includes linear regression model, discrete model, time series analysis, bayesian analysis.

8. Statistics

This module provides functions for calculating mathematical statistics of numeric

9. Pynance

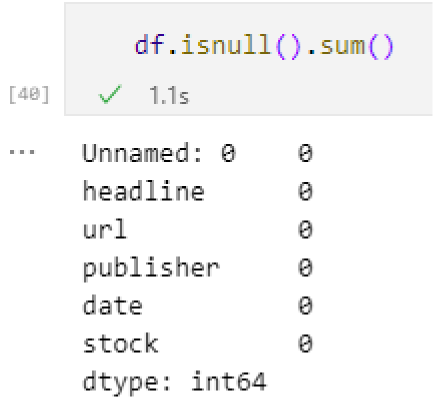
Pynance will work wonders for a stock market trader. It is an open-source python package that retrieves, analyses and visualizes the data from stock market derivatives.

Dataset Features

1) **headline**: Article release headline, the title of the news article, which often includes key financial actions like stocks hitting highs, price target changes, or company earnings.
2) **url**: The direct link to the full news article.
3) **publisher**: Author/creator of article.
4) **date**: The publication date and time, including time zone information (UTC-4 time zone).
5) **stock**: Stock ticker symbol (unique series of letters assigned to a publicly traded company). For example (AAPL: Apple)

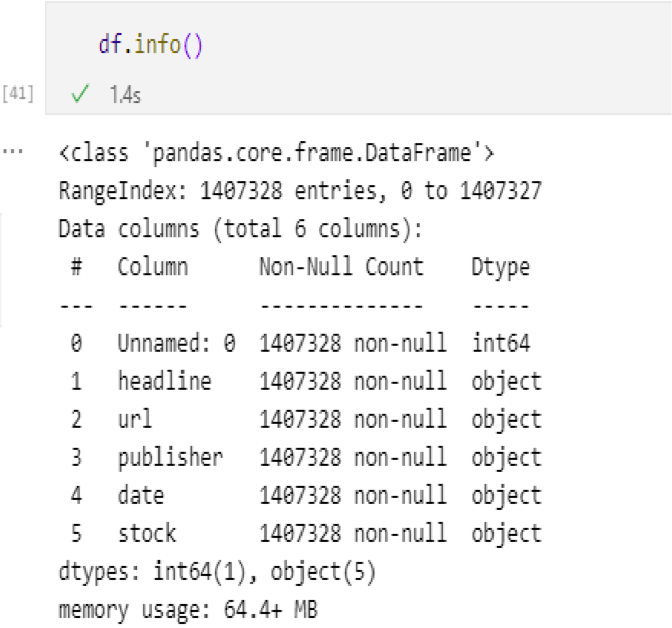
Data Processing

The technique used for data preprocessing is checking only the null values for each columns as shown in the figure below.



```
df.isnull().sum()
[40] ✓ 1.1s

...  Unnamed: 0    0
     headline      0
     url           0
     publisher     0
     date          0
     stock         0
     dtype: int64
```

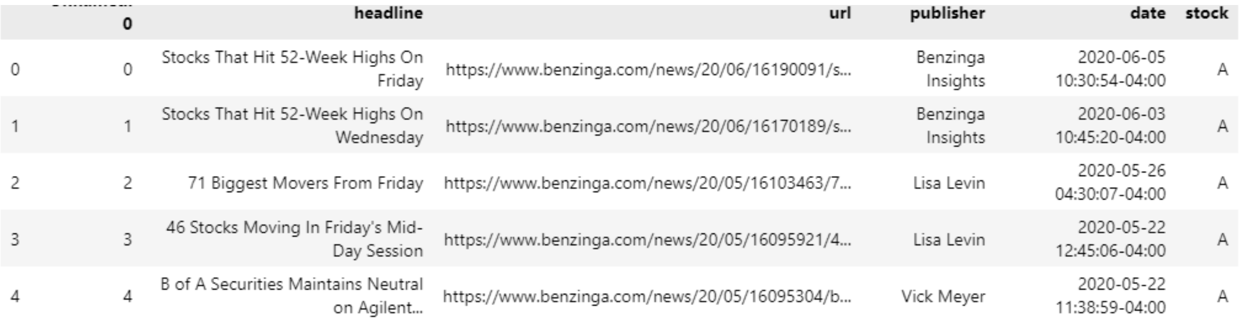


```
df.info()
[41] ✓ 1.4s

...  <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 1407328 entries, 0 to 1407327
     Data columns (total 6 columns):
      #   Column      Non-Null Count    Dtype
     ---  ------      --------------    -----
      0   Unnamed: 0  1407328 non-null  int64
      1   headline    1407328 non-null  object
      2   url         1407328 non-null  object
      3   publisher   1407328 non-null  object
      4   date        1407328 non-null  object
      5   stock       1407328 non-null  object
     dtypes: int64(1), object(5)
     memory usage: 64.4+ MB
```

Data Analysis (EDA) of the data structure of the dataset

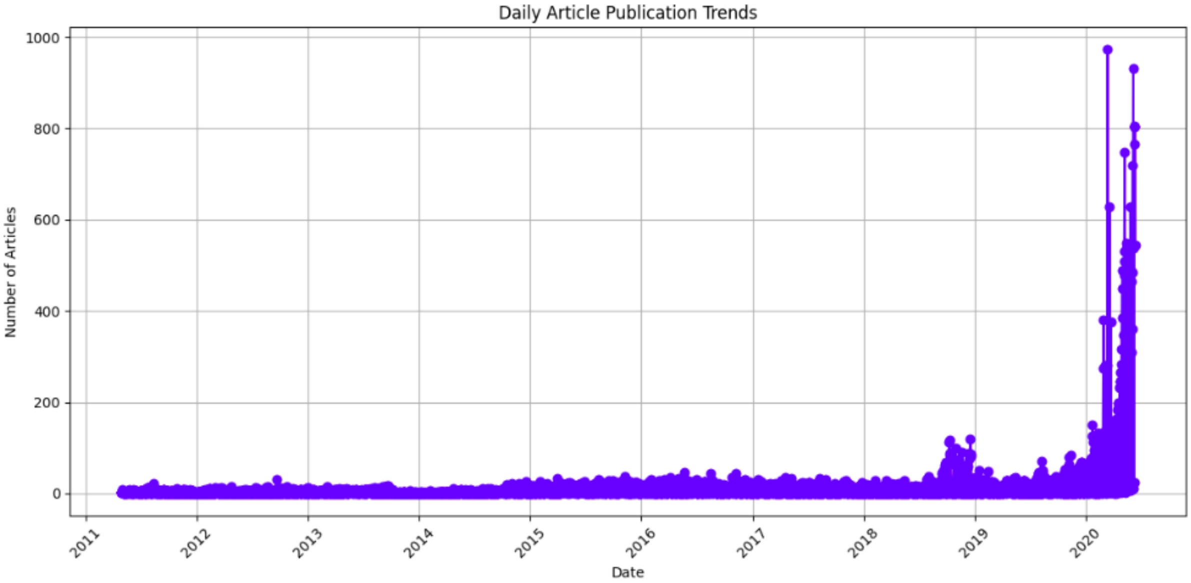
Here below shows us the dataset has columns ‘headline’, ‘url’, ‘publisher’, ‘date’ (publication date), and the ‘stock’.



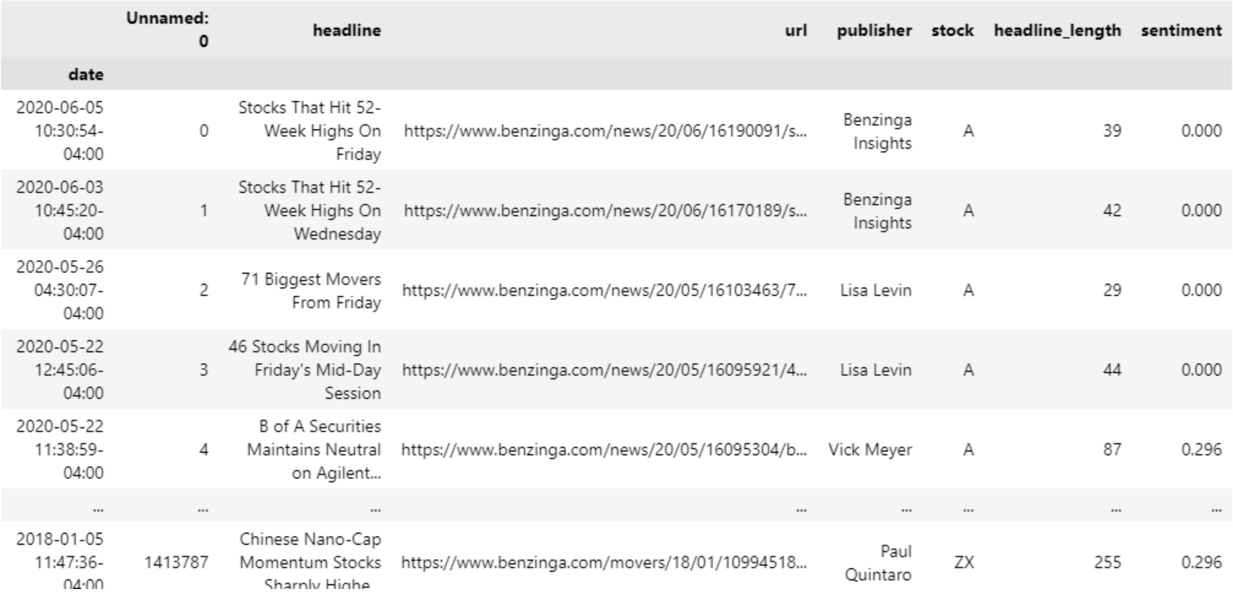
	0	headline	url	publisher	date	stock
0	0	Stocks That Hit 52-Week Highs On Friday	https://www.benzinga.com/news/20/06/16190091/s...	Benzinga Insights	2020-06-05 10:30:54-04:00	A
1	1	Stocks That Hit 52-Week Highs On Wednesday	https://www.benzinga.com/news/20/06/16170189/s...	Benzinga Insights	2020-06-03 10:45:20-04:00	A
2	2	71 Biggest Movers From Friday	https://www.benzinga.com/news/20/05/16103463/7...	Lisa Levin	2020-05-26 04:30:07-04:00	A
3	3	46 Stocks Moving In Friday's Mid-Day Session	https://www.benzinga.com/news/20/05/16095921/4...	Lisa Levin	2020-05-22 12:45:06-04:00	A
4	4	B of A Securities Maintains Neutral on Agilent...	https://www.benzinga.com/news/20/05/16095304/b...	Vick Meyer	2020-05-22 11:38:59-04:00	A

Data Visualization

Here below the figure illustrates us the respective number publications from 2011 to 2020. Here we can conclude that the number of publications has been going little bit in constant number of articles publication from 2011 to mid of 2018 and form the mid of 2018 to the last of the 2019 some how increase and lastly and dramatically from 2020 it is increasing.

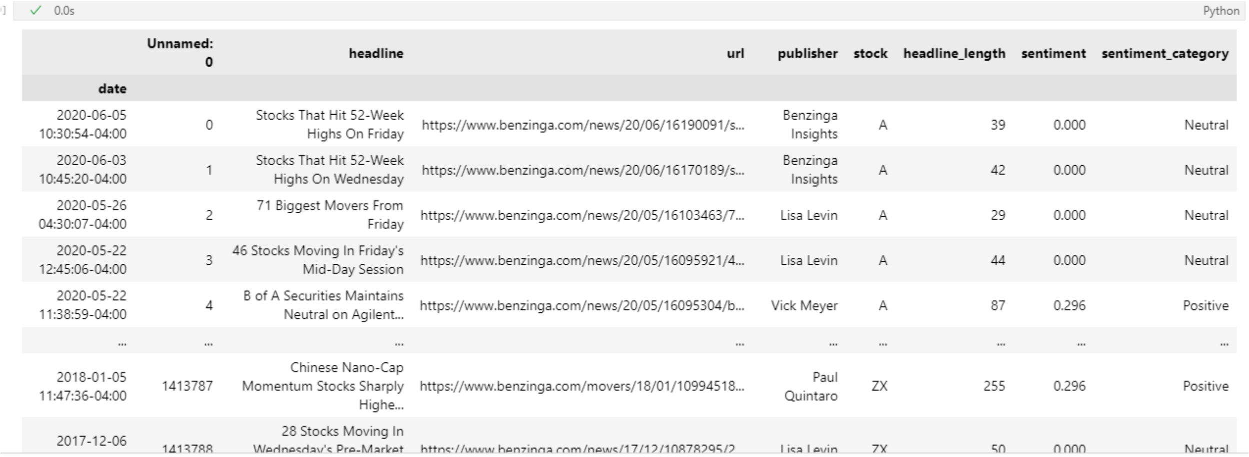


Sentiment analysis



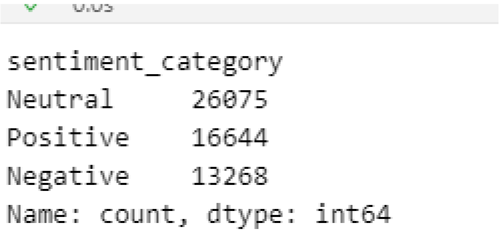
	Unnamed: 0	headline	url	publisher	stock	headline_length	sentiment
date							
2020-06-05 10:30:54-04:00	0	Stocks That Hit 52-Week Highs On Friday	https://www.benzinga.com/news/20/06/16190091/s...	Benzinga Insights	A	39	0.000
2020-06-03 10:45:20-04:00	1	Stocks That Hit 52-Week Highs On Wednesday	https://www.benzinga.com/news/20/06/16170189/s...	Benzinga Insights	A	42	0.000
2020-05-26 04:30:07-04:00	2	71 Biggest Movers From Friday	https://www.benzinga.com/news/20/05/16103463/7...	Lisa Levin	A	29	0.000
2020-05-22 12:45:06-04:00	3	46 Stocks Moving In Friday's Mid-Day Session	https://www.benzinga.com/news/20/05/16095921/4...	Lisa Levin	A	44	0.000
2020-05-22 11:38:59-04:00	4	B of A Securities Maintains Neutral on Agilent...	https://www.benzinga.com/news/20/05/16095304/b...	Vick Meyer	A	87	0.296
...	...	...	...	...	...	...	...
2018-01-05 11:47:36-04:00	1413787	Chinese Nano-Cap Momentum Stocks Sharply Highe	https://www.benzinga.com/movers/18/01/10994518...	Paul Quintaro	ZX	255	0.296

The below figure illustrates us the 'headline_length' and the sentiment value associated with the news headline. The sentiment might be 'negative', 'neutral' and 'positive'



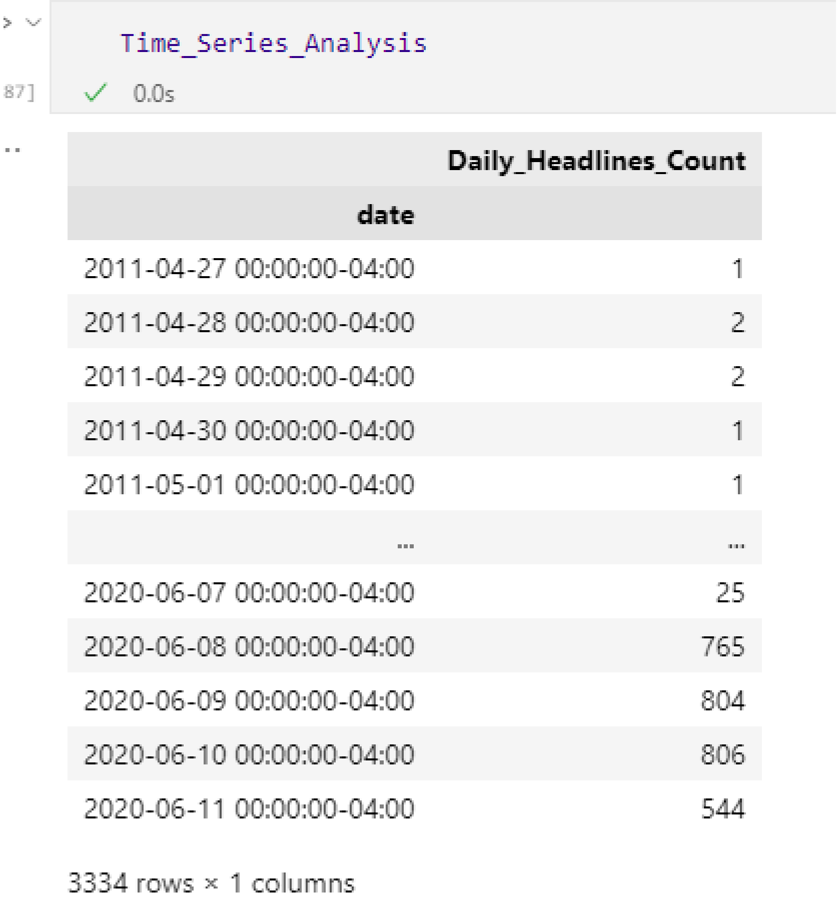
	Unnamed: 0	headline	url	publisher	stock	headline_length	sentiment	sentiment_category
date								
2020-06-05 10:30:54-04:00	0	Stocks That Hit 52-Week Highs On Friday	https://www.benzinga.com/news/20/06/16190091/s...	Benzinga Insights	A	39	0.000	Neutral
2020-06-03 10:45:20-04:00	1	Stocks That Hit 52-Week Highs On Wednesday	https://www.benzinga.com/news/20/06/16170189/s...	Benzinga Insights	A	42	0.000	Neutral
2020-05-26 04:30:07-04:00	2	71 Biggest Movers From Friday	https://www.benzinga.com/news/20/05/16103463/7...	Lisa Levin	A	29	0.000	Neutral
2020-05-22 12:45:06-04:00	3	46 Stocks Moving In Friday's Mid-Day Session	https://www.benzinga.com/news/20/05/16095921/4...	Lisa Levin	A	44	0.000	Neutral
2020-05-22 11:38:59-04:00	4	B of A Securities Maintains Neutral on Agilent...	https://www.benzinga.com/news/20/05/16095304/b...	Vick Meyer	A	87	0.296	Positive
...	...	...	...	...	...	...	...	...
2018-01-05 11:47:36-04:00	1413787	Chinese Nano-Cap Momentum Stocks Sharply Highe...	https://www.benzinga.com/movers/18/01/10994518...	Paul Quintaro	ZX	255	0.296	Positive
2017-12-06	1413788	28 Stocks Moving In Wednesday's Pre-Market	https://www.benzinga.com/news/17/12/10878295/2	Lisa Levin	ZX	50	0.000	Neutral

Here below is the amount of sentiment category ('negative', 'neutral', and 'positive')



```
sentiment_category
Neutral     26075
Positive    16644
Negative    13268
Name: count, dtype: int64
```

Time Series Analysis

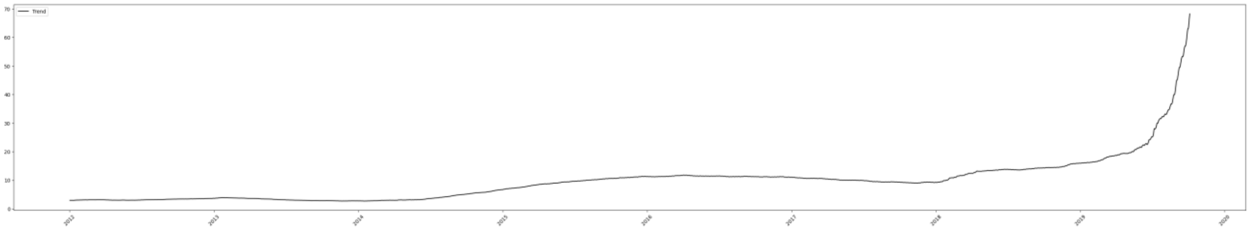


```
Time_Series_Analysis
```

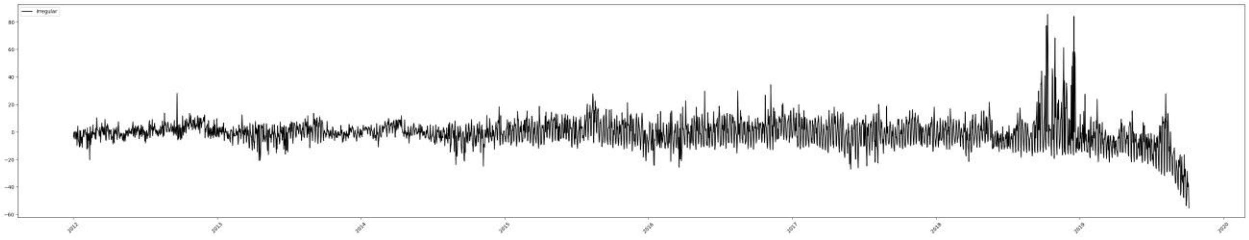
	Daily_Headlines_Count
date	
2011-04-27 00:00:00-04:00	1
2011-04-28 00:00:00-04:00	2
2011-04-29 00:00:00-04:00	2
2011-04-30 00:00:00-04:00	1
2011-05-01 00:00:00-04:00	1
...	...
2020-06-07 00:00:00-04:00	25
2020-06-08 00:00:00-04:00	765
2020-06-09 00:00:00-04:00	804
2020-06-10 00:00:00-04:00	806
2020-06-11 00:00:00-04:00	544

3334 rows × 1 columns

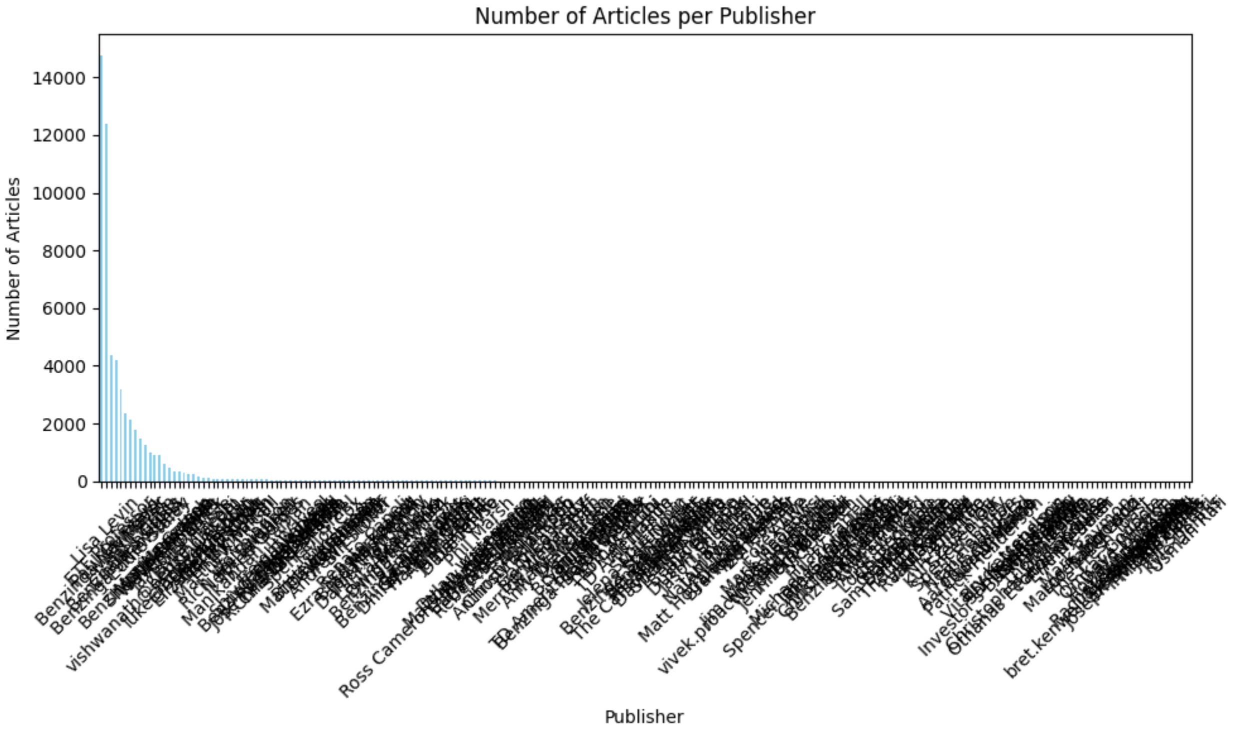
Seasonal variation of the Time Series Analysis



Irregular variation of the Time Series Analysis



Number of articles vs the publishers


Number of Articles per Publisher
Number of Articles
14000
12000
10000
8000
6000
4000
2000
0
Lisa Levin
Ross Cameron
Phil Marsh
Publisher