GA CAPSTONE PROJECT:

Influence of Climate Change in the Media on iShares Global Clean Energy ETF Activity

- Building a Predictive Model -

Author: Theresa Waters
Course: General Assembly Data Science (Part-Time)
Date: Spring 2022

Agenda

- Objectives
- Data Extraction
- Data Exploration
- Modeling
- Next Steps
- Key Takeaways

Objectives

Question: Is a clean energy ETF's volume influenced by media coverage of climate change related topics?

- Use text from The New York Times Article Archive to gain insight into climate-change related topic trends over time
- Incorporate article trends into a time series model of ICLN ETF volume data
- Determine if climate change discussion in the media is a predictor for ICLN volume activity

Objectives

New York Times: Article Archive API

Filter via Topic Modelling for Climate Change related articles from the last 10 years

Climate Change: daily articles (2012-2022)

+

Alpha Vantage: API

Filter for iShares Global Energy Clean ETF (ICLN) daily results from the last 10 years

ICLN: daily volume (2012-2022)

ARIMAX

Time Series Forecasting Model

Agenda

- Objectives
- Data Extraction
- Data Exploration
- Data Modeling
- Next Steps
- Key Takeaways

Data Extraction – The New York Times

	headline	date	doc_type	material_type	section	keywords	abstract
0	Trump's Book Club: A President Who Doesn't Rea...	2018-12-01	article	News	NaN	['Books and Literature', 'United States Politi...	President Trump, who is not a reader, has used...
1	Kareem Hunt Is Cut by the Chiefs After a Video...	2018-12-01	article	News	NaN	['Football', 'Domestic Violence']	Hunt, a star running back, was also suspended ...
2	Agency Pulls Back on Its Warning Against Talk ...	2018-12-01	article	News	NaN	['Hatch Act (1939)']	Casual conversations about impeachment and inv...
3	A China Hawk Gains Prominence as Trump Confron...	2018-12-01	article	News	NaN	['United States International Relations', 'Uni...	Michael Pillsbury, the president's top outside...
4	Wilmer Flores Is a Met No More	2018-12-01	article	News	NaN	['Baseball']	A fan favorite, Flores did not receive a contr...
5	Will Trump Speak Up Against China's Oppression?	2018-12-01	article	Editorial	NaN	['United States International Relations', 'Hum...	In Argentina, President Trump has a chance to ...

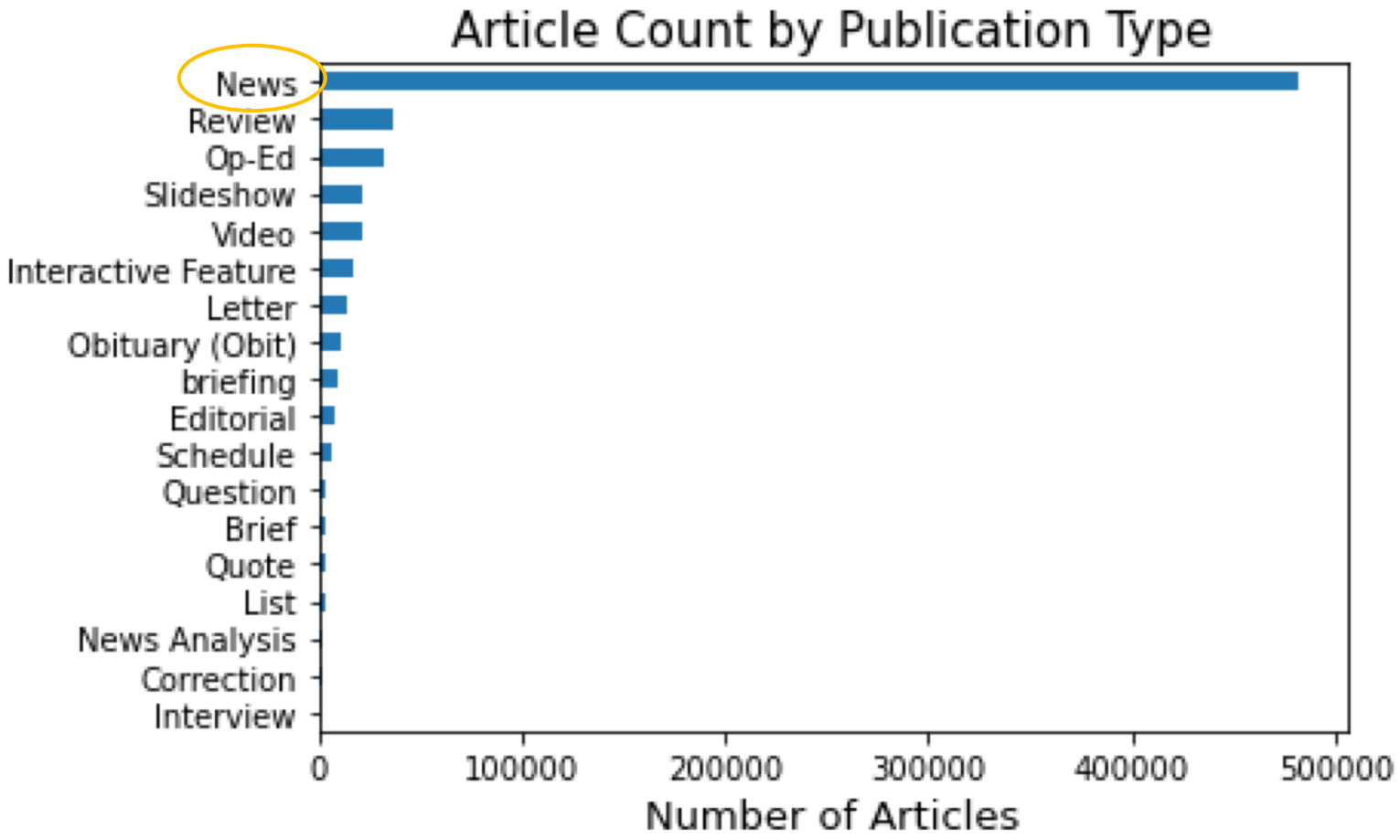
Notes:

- NYT API provides text data: headline, keywords, abstract

Agenda

- Objectives
- Data Extraction
- Data Exploration
- Modeling
- Next Steps
- Key Takeaways

Data Visualization – The New York Times

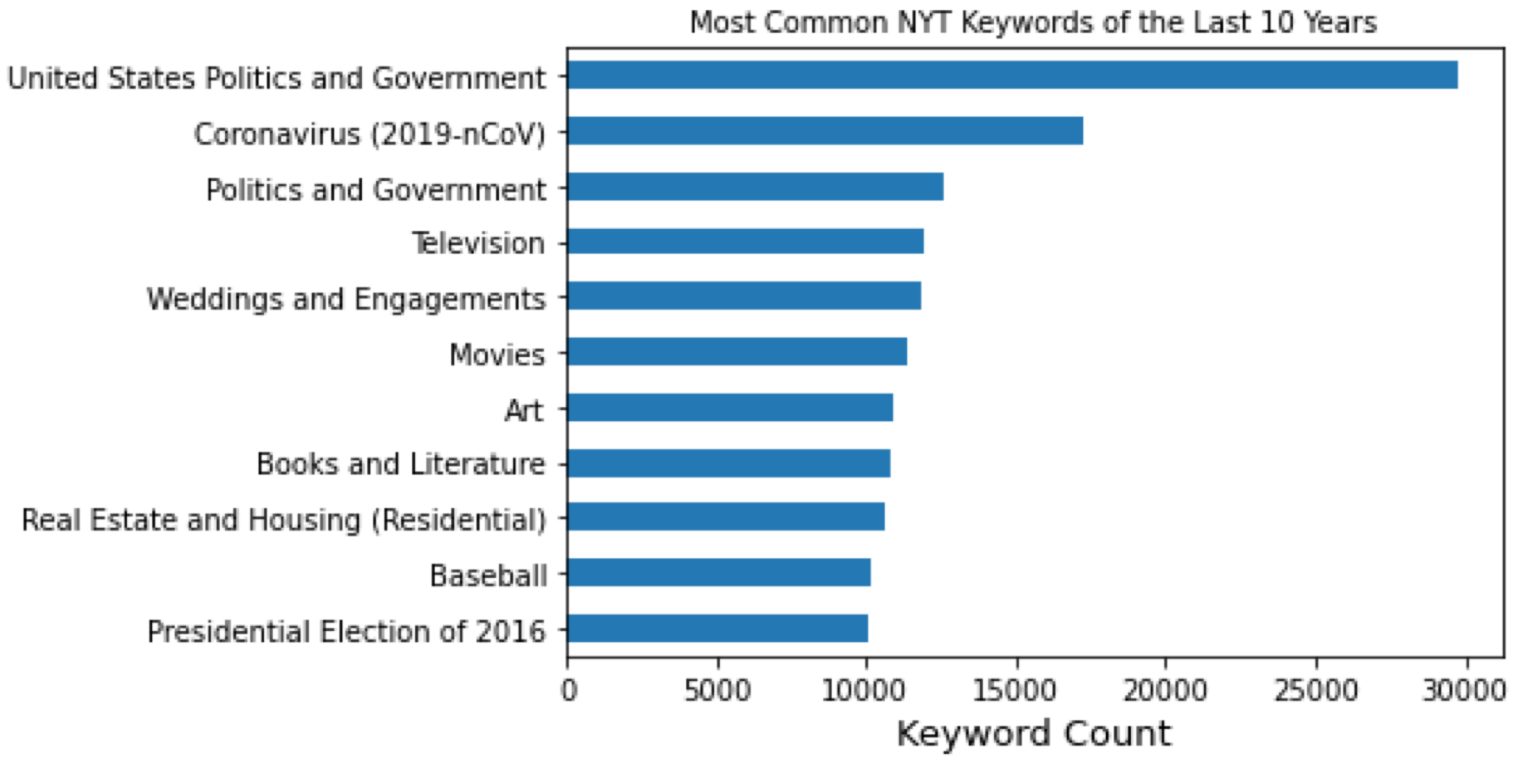


Notes:

- Most of the articles fall under News – this is the publication type I focused on

Data Visualization – The New York Times

Most Common Keywords in the News



Notes:

- Top keywords in the news from the last 10 years do not relate to climate change

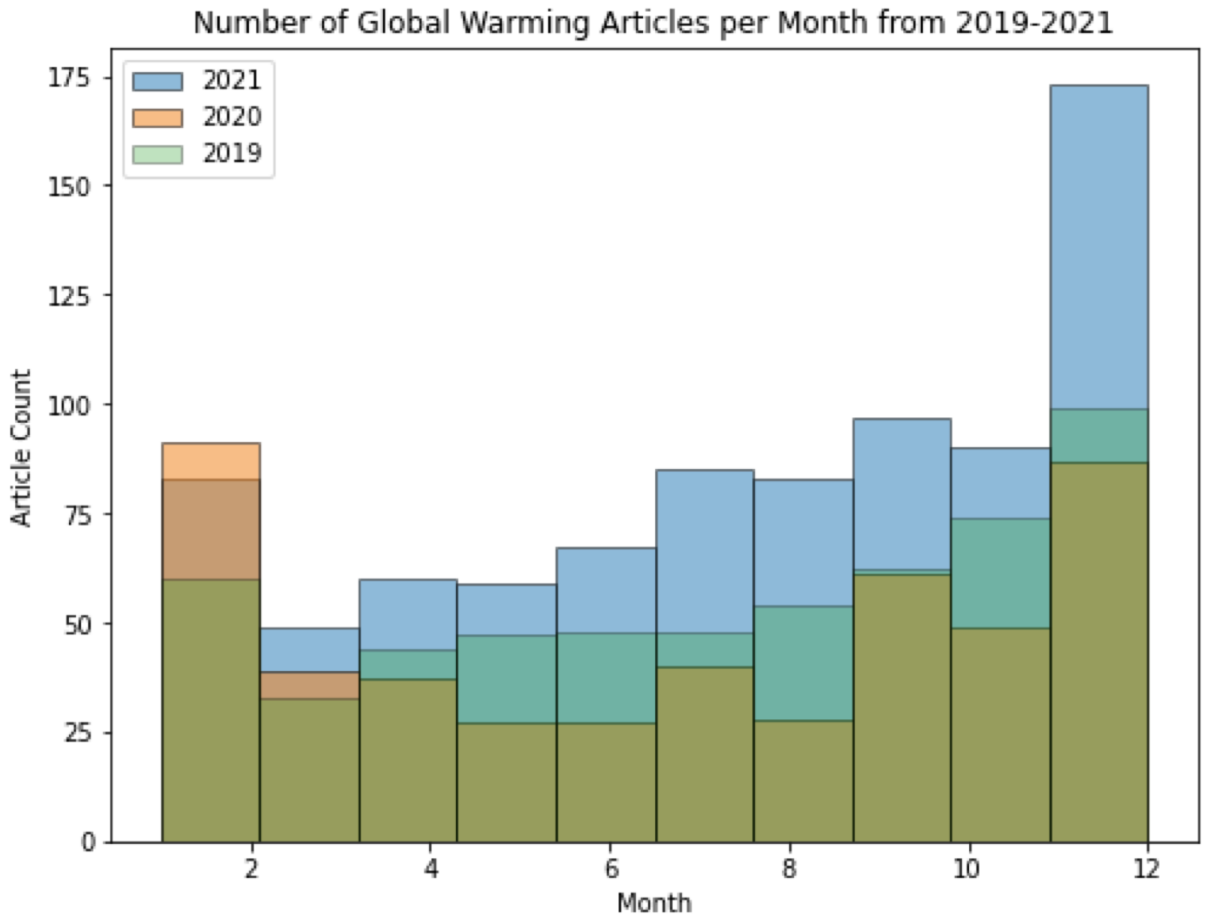
Data Visualization – The New York Times

Most Common Words Across All Article Text

Data Visualization – The New York Times

- Filtering for climate change related articles using only the keywords section: fall under ‘Global Warming’

- Total GW articles from the last 10 years: **4,536**

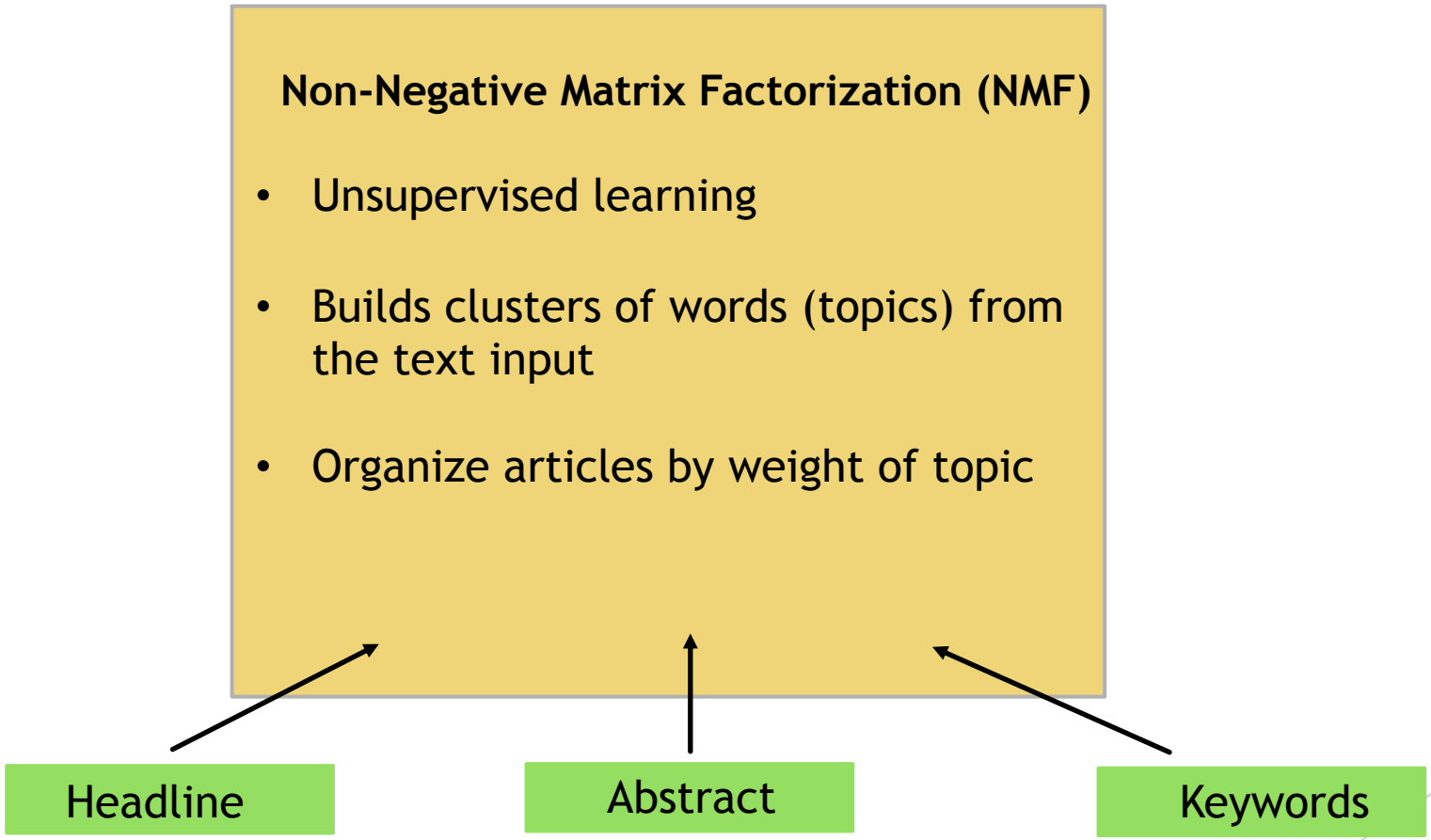


Agenda

- Objectives
- Data Extraction
- Data Exploration
- Modeling
- Next Steps
- Key Takeaways

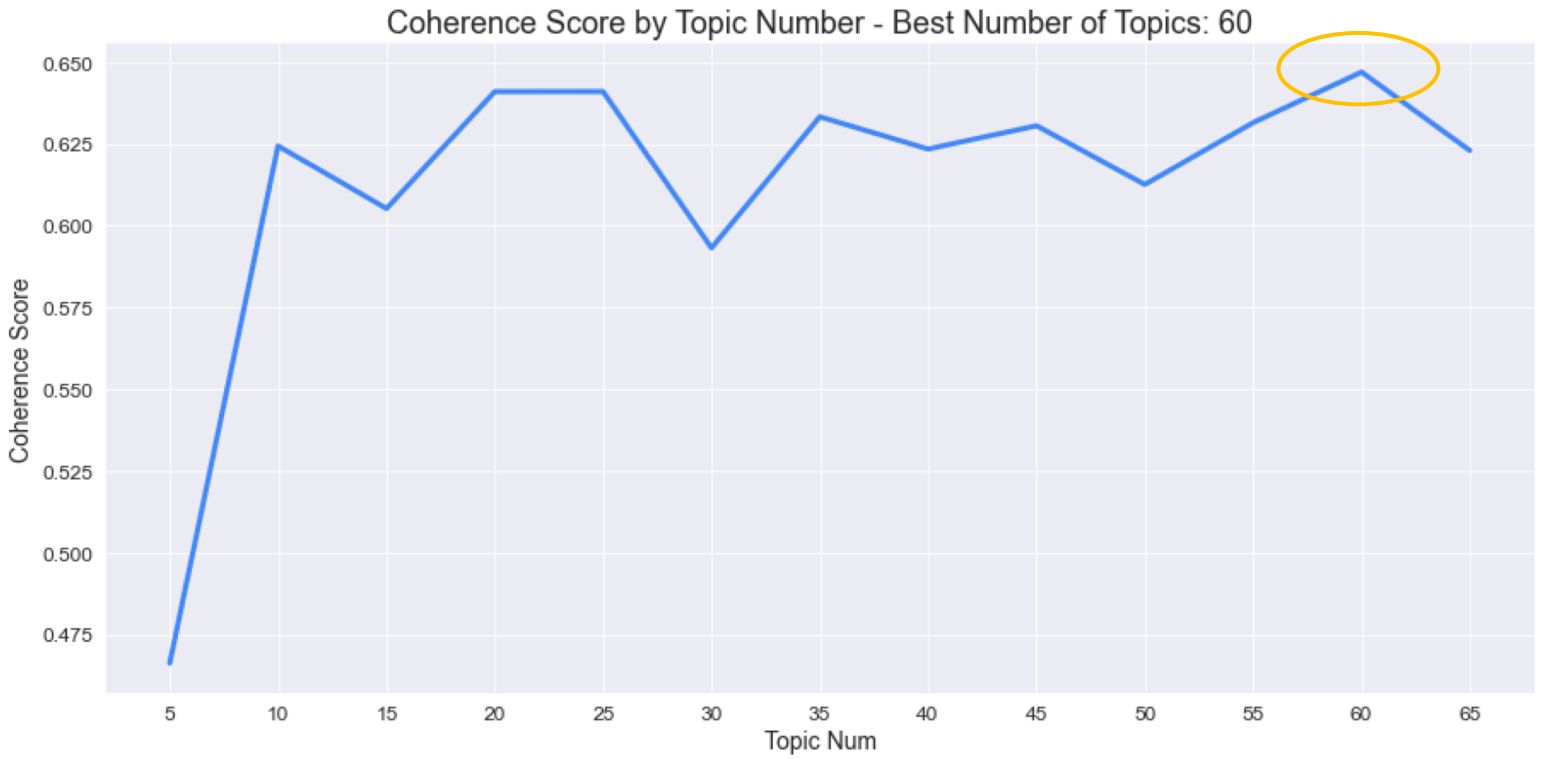
Topic Model – New York Times

- Is just using the 'Global Warming' keyword the best way to group articles related to topics important to climate change? E.g., Renewable energy, emissions, etc.



Topic Model - New York Times

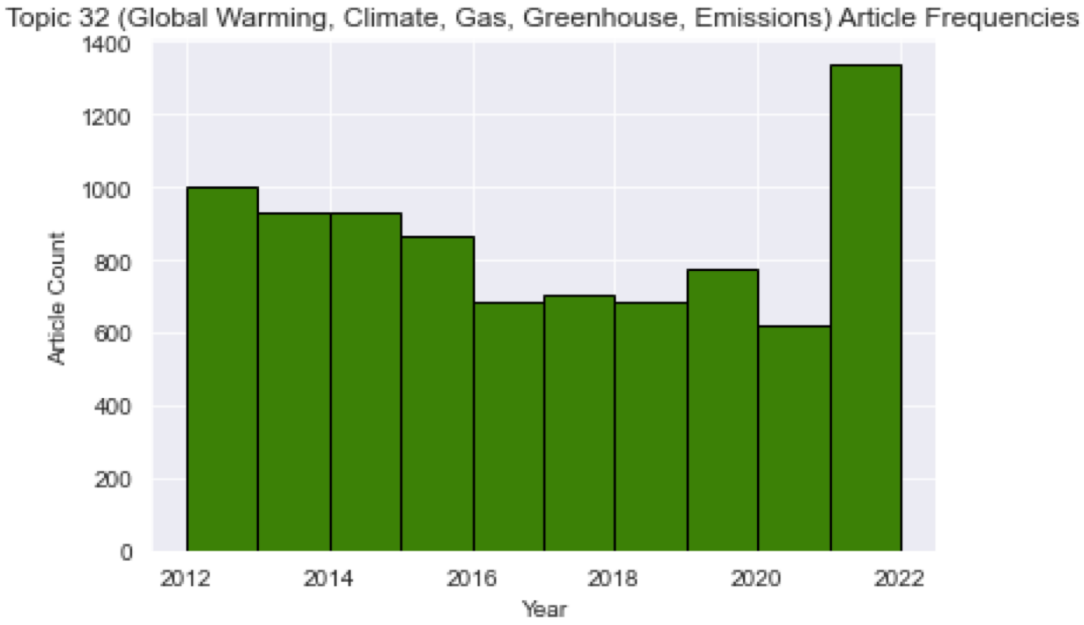
- Determining the best number of topics to input into model using coherence scores (how similar the words in each topic are to one another) - **60 topics**



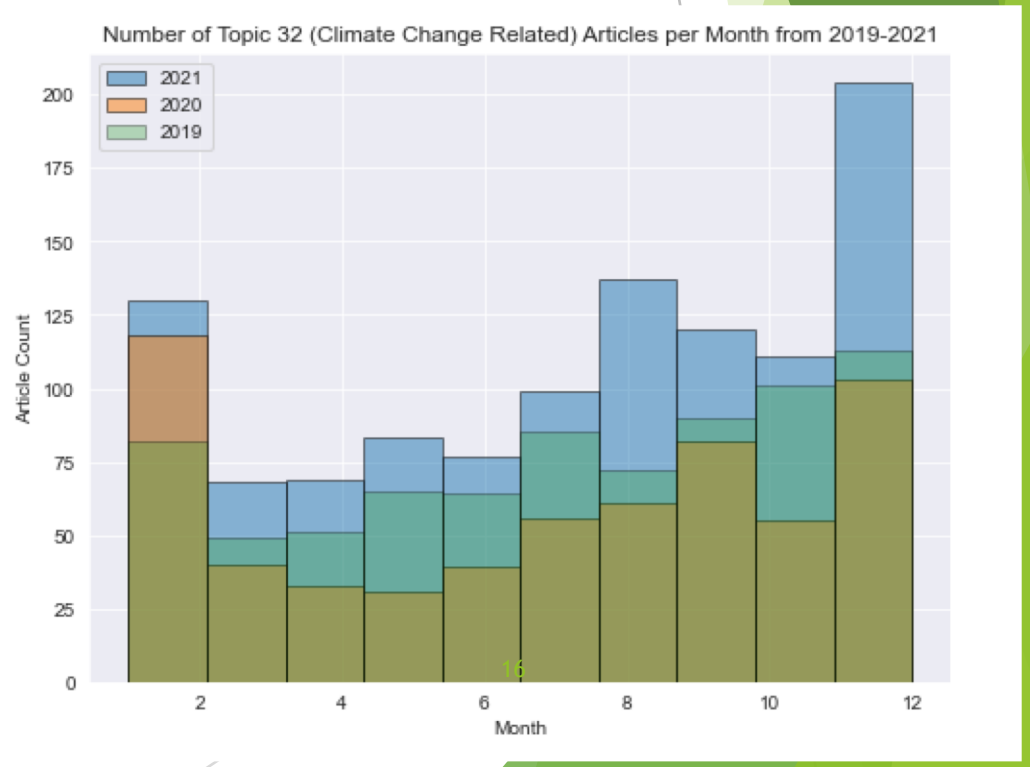
Topic Model - New York Times

27	27	movi film televis festiv documentari oscar award
28	28	right human violat prison group
29	29	merger acquisit divestitur deal billion compani
30	30	execut appoint chang chief director
31	31	immigr emigr illeg border migrant deport
32	32	warm global climat gas emiss greenhous
33	33	tenni open state french wimbledon tournament
34	34	soccer cup world team leagu
35	35	polic shoot offic brutal misconduct
36	36	job labor economi unemploy wage state econom
37	37	school educ student teacher high employe
38	38	trump presid impeach donald russian interfer

Topic Model - New York Times

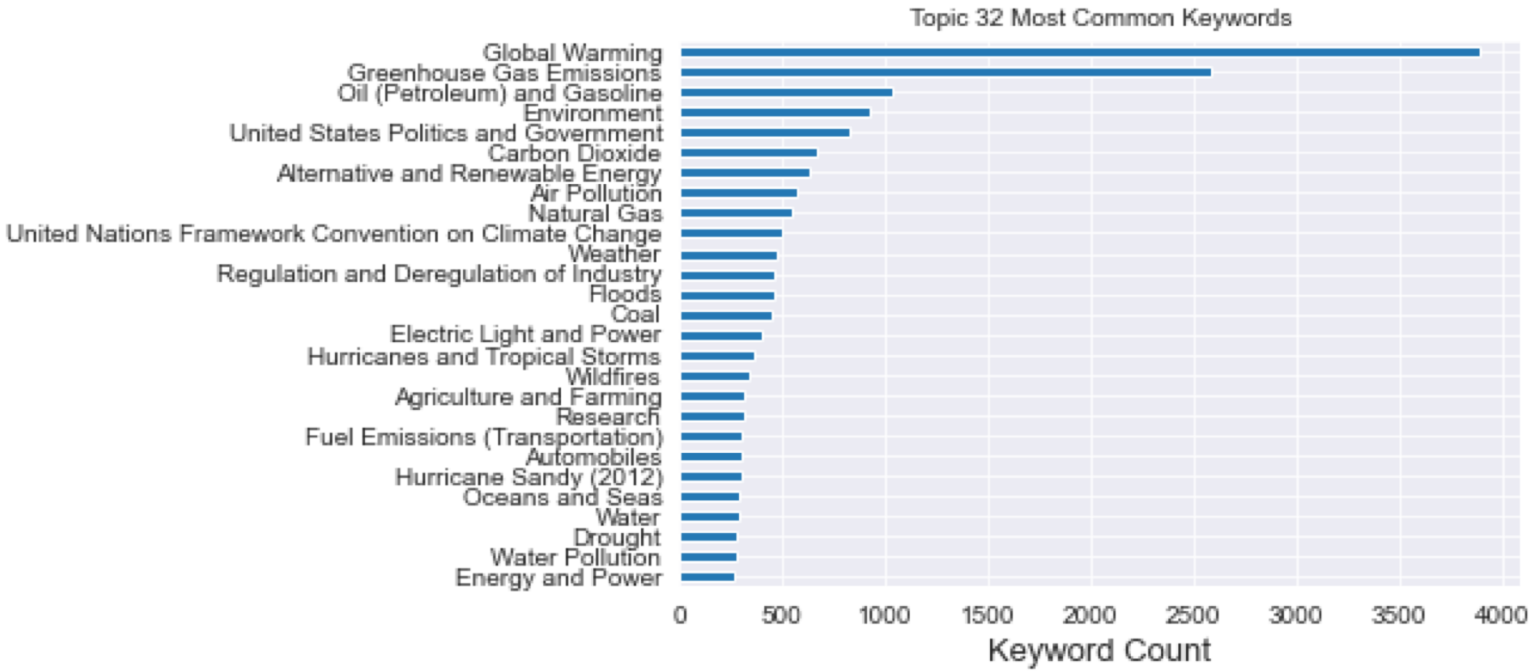


Total number of articles: **8534**



Topic Model – New York Times

- Model is even more diverse than solely using the ‘Global Warming’ keyword as a filter



Agenda

- Objectives
- Data Extraction
- Data Exploration
- Modeling
- Next Steps
- Key Takeaways

Data Extraction – Alpha Vantage API

- Extracted data from 2012-2022 for the iShares Global Clean Energy ETF (ICLN)

	1. open	2. high	3. low	4. close	5. volume
2022-04-22	19.8100	19.8400	19.3100	19.3500	10320130
2022-04-21	20.8500	20.8500	19.6200	19.6800	10410516
2022-04-20	21.2000	21.2500	20.6800	20.6900	5040229
2022-04-19	20.7200	21.1000	20.6000	21.0600	6233716
2022-04-18	20.6900	20.8400	20.5403	20.6300	3723329
...	...	...	...	...	...
2008-07-01	50.0000	50.0000	48.0600	48.8000	14900
2008-06-30	50.7700	50.9600	50.2500	50.2500	17100
2008-06-27	50.7900	50.7900	50.0900	50.1600	7700
2008-06-26	59.9900	59.9999	50.9900	51.0600	9800
2008-06-25	52.2500	52.9800	52.2500	52.7700	2100

3482 rows × 5 columns

Notes:

- AV API gives numerical daily data for price and volume.

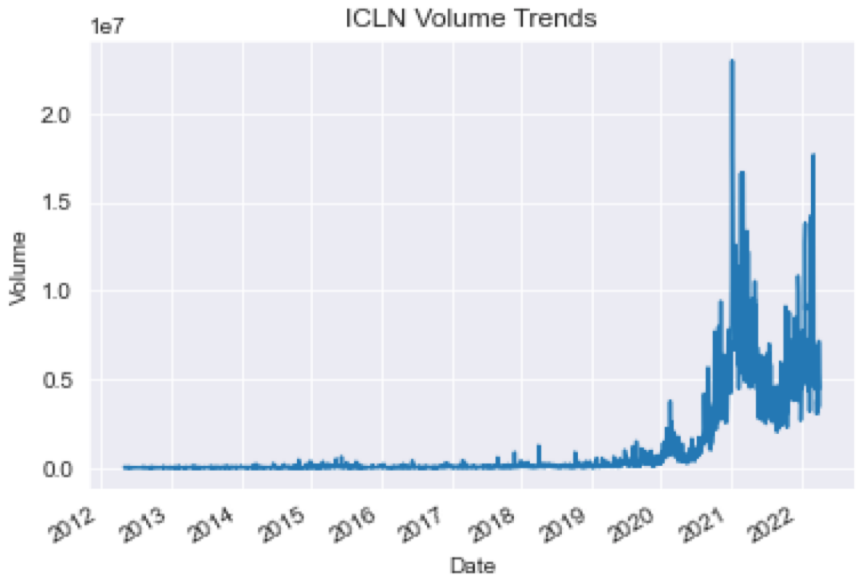
Agenda

- Objectives
- Data Extraction
- Data Exploration
- Modeling
- Next Steps
- Key Takeaways

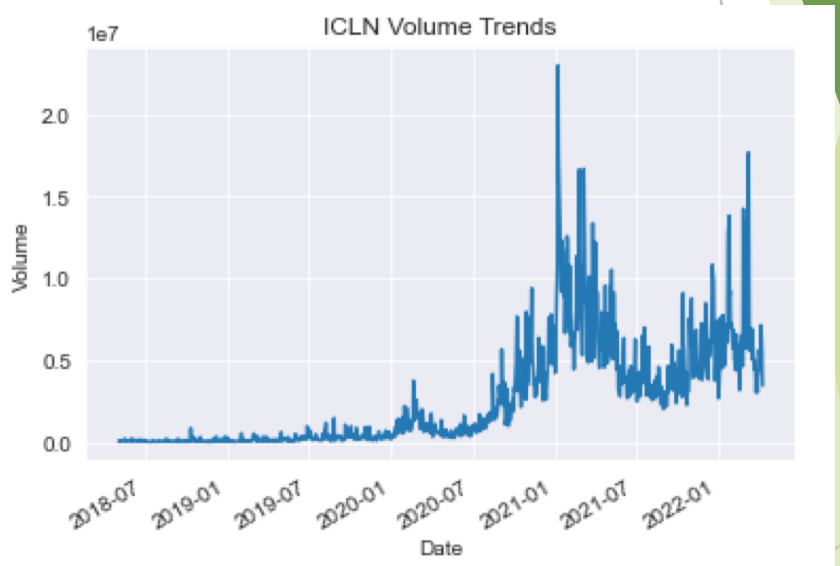
Data Exploration - ICLN

- Most volume activity for the ICLN ETF has occurred in the last 4 years: filtered to use only 2018-2022 data for the time series model

Original: 2012-2022



New: 2018-2022



Merging NYT and ICLN Data

date	1. open	2. high	3. low	4. close	5. volume	article_count
2022-03-31	21.4600	21.7450	21.4600	21.5200	4641040	2.0
2022-03-30	21.6300	21.7700	21.3400	21.3900	4184370	3.0
2022-03-29	21.1900	21.4600	21.0150	21.4400	4000912	4.0
2022-03-28	21.0100	21.2050	20.8050	21.1300	3066803	2.0
2022-03-25	21.2800	21.2800	20.7150	21.0400	3794886	5.0
...	...	...	...	...	...	...
2018-05-08	9.8500	9.8500	9.7200	9.7400	44512	0.0
2018-05-07	9.8500	9.9000	9.8500	9.8600	85477	0.0
2018-05-04	9.7800	9.8500	9.7400	9.8100	63677	1.0
2018-05-03	9.8300	9.8300	9.7000	9.7900	42297	4.0
2018-05-02	9.9000	9.9000	9.7800	9.8100	87889	1.0

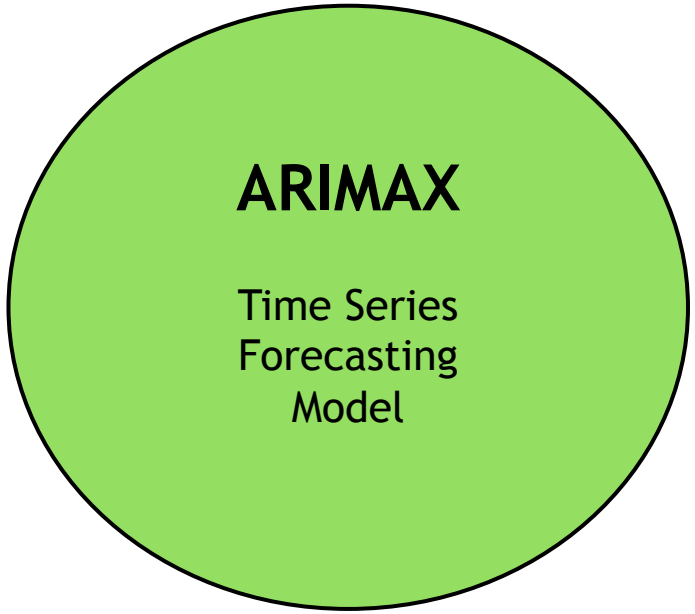
987 rows × 8 columns

Agenda

- Objectives
- Data Extraction
- Data Exploration
- Modeling
- Next Steps
- Key Takeaways

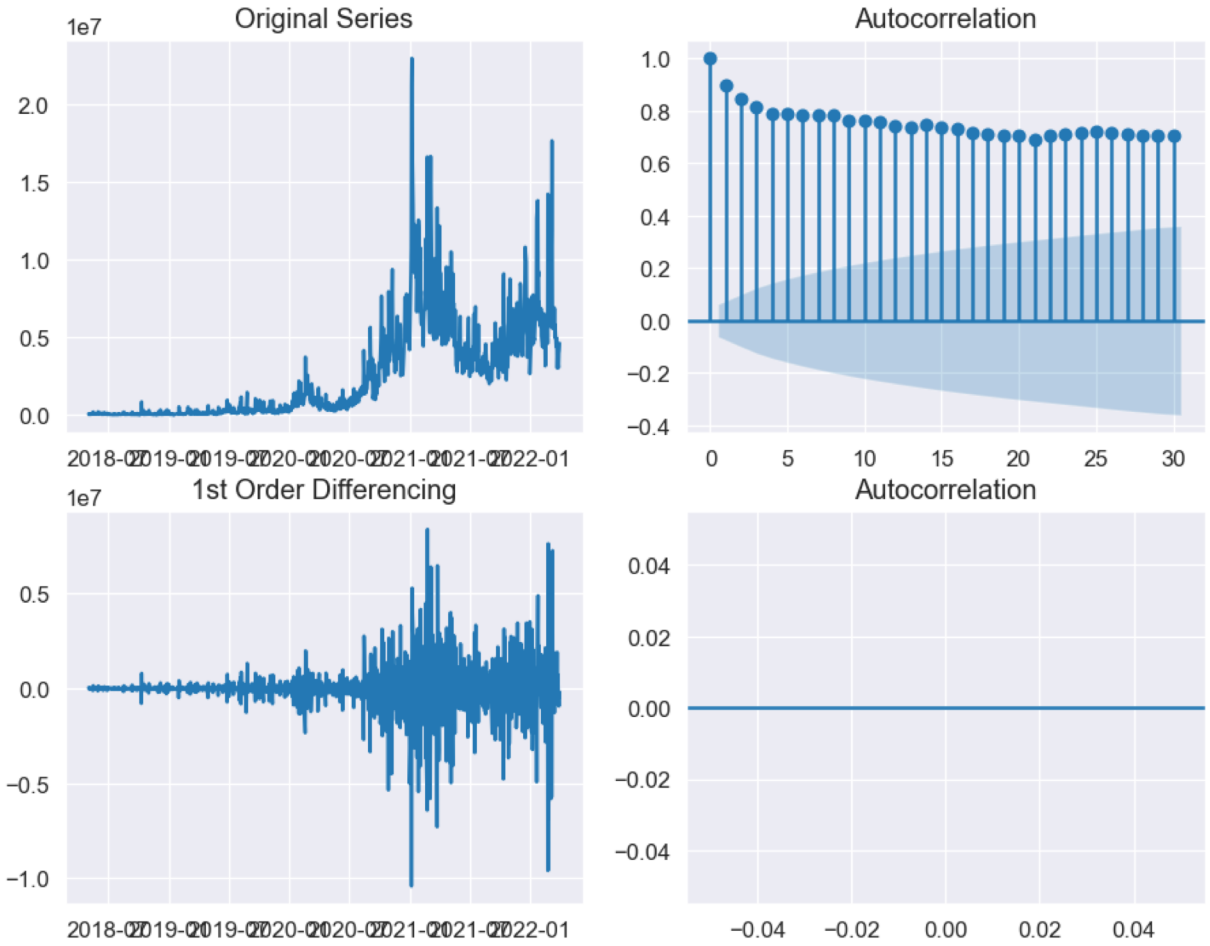
Modeling - ARIMAX

- Implemented a SARIMAX model (without seasonality ‘S’) to predict ICLN ETF volume activity
- Incorporated climate change related NYT article counts into the model as an exogenous variable



Modeling - ARIMAX

- Determining parameters: Augmented Dickey Fuller (ADF) Test = **p-value: 0.418176**



Notes:

- ADF p-value > 0.05
- Volume data are non-stationary and need to be differenced

Modeling - ARIMAX

```
Best model:  ARIMA(1,1,3)(0,0,0)[0]
Total fit time: 5.214 seconds
```

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	740
Model:	SARIMAX(1, 1, 3)	**Log Likelihood**	-11570.196
Date:	Mon, 25 Apr 2022	**AIC**	23152.392
Time:	12:50:31	**BIC**	23180.023
Sample:	0	**HQIC**	23163.046
	- 740		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
article_count	8.881e+04	2.58e+04	3.438	0.001	3.82e+04	1.39e+05
ar.L1	0.4490	0.103	4.371	0.000	0.248	0.650
ma.L1	-0.8623	0.103	-8.386	0.000	-1.064	-0.661
ma.L2	0.0579	0.047	1.235	0.217	-0.034	0.150
ma.L3	-0.0910	0.040	-2.268	0.023	-0.170	-0.012
sigma2	2.405e+12	0.002	1.19e+15	0.000	2.41e+12	2.41e+12

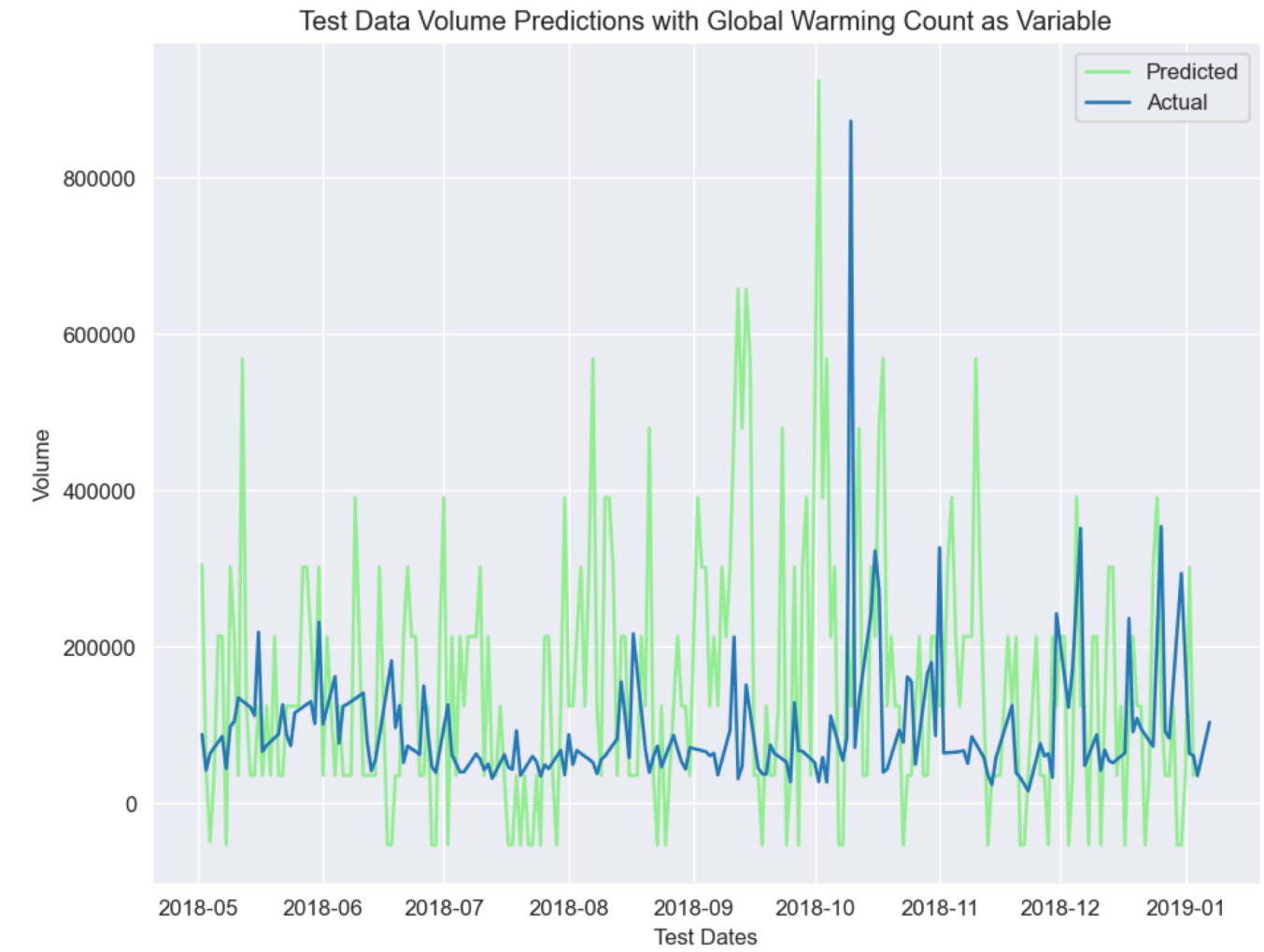
Ljung-Box (L1) (Q):	0.01	**Jarque-Bera (JB):**	1584.40
Prob(Q):	0.94	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.05	**Skew:**	0.97
Prob(H) (two-sided):	0.00	**Kurtosis:**	9.90

Notes:

- Best fitting model (lowest AIC score)
- Article count has significant impact on volume trends
- Article count coefficient is positive – as article counts increase so does ICLN ETF volume

Modeling - ARIMAX

- Split into training and test data for the model
- Test predictions plotted against actual volume data



Notes:

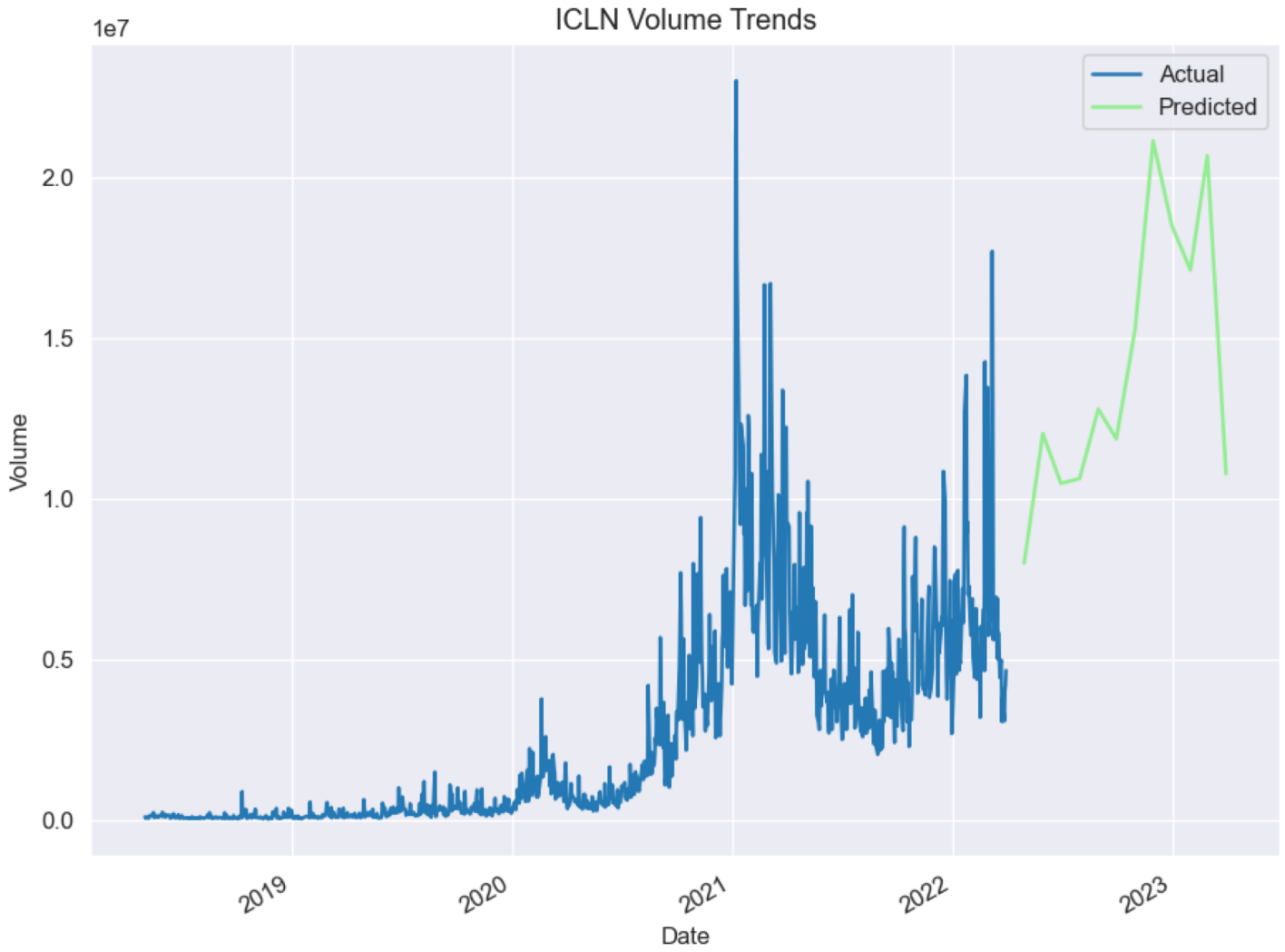
- Although the daily trends do not align perfectly, overall trends align

Modeling - ARIMAX - Forecast

- Between 2020-2021 climate change related article counts increased by 74%
- As climate change related events continue to become more prevalent could we see article counts increase by at least 74% from 2022-2023? What will this mean for the ICLN ETF?

date	month	article_count	article_count_pseudo
2021-05-31	1	52	90.48
2021-06-30	2	78	135.72
2021-07-31	3	68	118.32
2021-08-31	4	69	120.06
2021-09-30	5	83	144.42
2021-10-31	6	77	133.98
2021-11-30	7	99	172.26
2021-12-31	8	137	238.38
2022-01-31	9	120	208.80
2022-02-28	10	111	193.14
2022-03-31	11	134	233.16
2022-04-30	12	70	121.80

Modeling - ARIMAX - Forecast



Notes:

- As article counts increase so will ICLN ETF volume!

Agenda

- Objectives
- Data Extraction
- Data Exploration
- Modeling
- Next Steps
- Key Takeaways

Next Steps

- Is the noise of the daily volume data impacting the ARIMAX model?
- Was first order differencing really the best parameter?
- Incorporate more media sources into the model beyond The New York Times
- Causality: does the model fit volume data for non-clean energy related ETFs? If not, then this is evidence the model is specific for ICLN and potentially other clean energy ETFs

Agenda

- Objectives
- Data Extraction
- Data Exploration
- Modeling
- Next Steps
- Key Takeaways

Key Takeaways

- Topic modeling combined with Natural Language Processing is a successful tool for organizing large collection of texts
- Time series forecasting provides evidence that the prevalence of climate change related news in the media has a positive influence on the activity of a clean energy ETF

Questions?