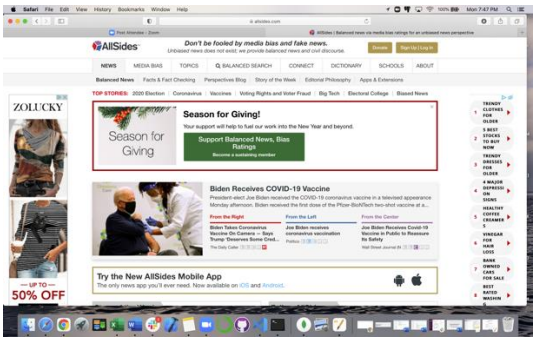
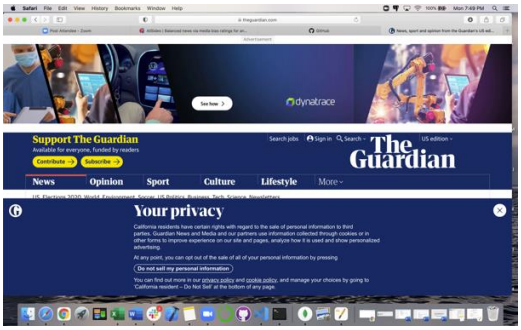
ETL Project: News Headlines
Team Members: Abby Herrup, Brett Thompson, Tim Schurmann and Clarence Robinson
The Sources: Allsides.com, Guardian IP
Tools: Google Cloud, Python, Juypter Notebook, SQL, Postgres

The initial process was to identify the sources we would be using to collect our data.
Once the data was obtained through web scrapping API's, we cleaned and organized our data using Pandas in Juypter notebooks. Once the data was organized in an easy read format the last step was to transfer our final output into a Data Base.

Allsides.com Data Source



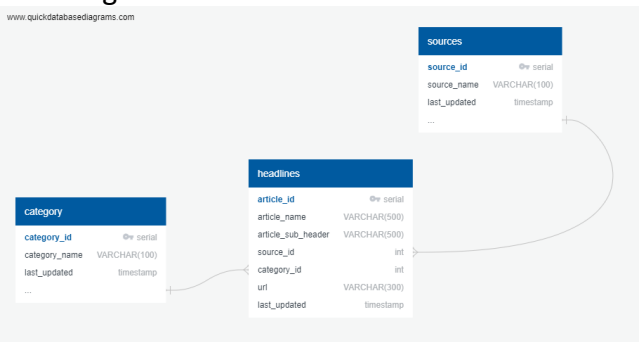
The Guardian Data Source



Extraction: We web scrapped data from Allsides.com and then collected data using pandas from The Guardian API, used a loop to save off the information from the API in a list, used lists to create base data frame and then took date in data frame and converted it to standard format using datetime.

Transformation: We pulled in various data sources and loaded them into the data frames, we reviewed the files and transformed them into data frames into Article, Source, Category, and Sub Header. We did an initial connection to the Postgres database using PG admin to store our original clean data sets. We used the quick database website to create the initial table schema that got loaded into the Postgres database that generated the first set of tables. After running the queries and created the new tables with only the relevant information we reconnected to the database and generated additional tables for the data frames.

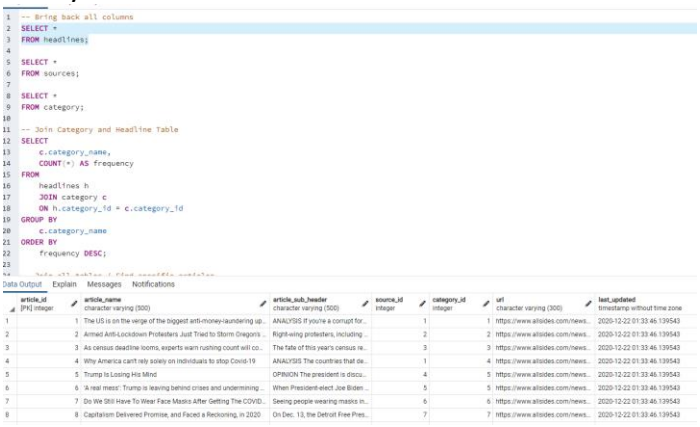
ERD Diagram



Upload: The last step was to transfer our final output into a Database. We created a database and respective tables to match the columns from the final Panda's Data Frame using SQL and then connected to the database using SQL Alchemy and loaded the result to Postgres. Here we were able to perform multiple queries to suit a desired criterion.

Future work: Add dates for when the articles were published and for when the articles were scrapped.

Query Result Screenshots



```sql
-- Bring back all columns
SELECT *
FROM headlines;

SELECT *
FROM sources;

SELECT *
FROM category;

-- Join Category and Headline Table
SELECT
    c.category_name,
    COUNT(*) AS frequency
FROM
    headlines h
    JOIN category c
    ON h.category_id = c.category_id
GROUP BY
    c.category_name
ORDER BY
    frequency DESC;
```

Data Output | Explain | Messages | Notifications

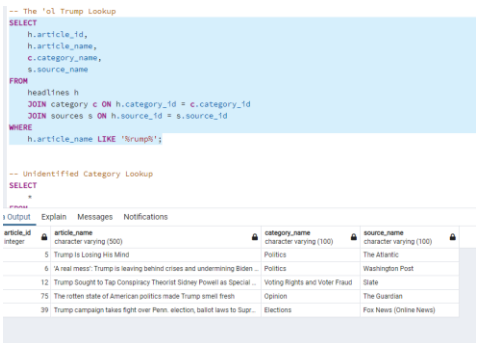
article_id [PK] integer	article_name character varying (500)	article_sub_header character varying (500)	source_id integer	category_id integer	url character varying (300)	last_updated timestamp without time zone
1	1 The US is on the verge of the biggest anti-money-laundering up...	ANALYSIS If you're a corrupt for...	1	1	https://www.allsides.com/news...	2020-12-22 01:33:46.139543
2	2 Armed Anti-Lockdown Protesters Just Tried to Storm Oregon's ...	Right-wing protesters, including ...	2	2	https://www.allsides.com/news...	2020-12-22 01:33:46.139543
3	3 As census deadline looms, experts warn rushing count will co...	The fate of this year's census re...	3	3	https://www.allsides.com/news...	2020-12-22 01:33:46.139543
4	4 Why America can't rely solely on individuals to stop Covid-19	ANALYSIS The countries that de...	1	4	https://www.allsides.com/news...	2020-12-22 01:33:46.139543
5	5 Trump Is Losing His Mind	OPINION The president is discu...	4	5	https://www.allsides.com/news...	2020-12-22 01:33:46.139543
6	6 'A real mess': Trump is leaving behind crises and undermining ...	When President-elect Joe Biden ...	5	5	https://www.allsides.com/news...	2020-12-22 01:33:46.139543
7	7 Do We Still Have To Wear Face Masks After Getting The COVID...	Seeing people wearing masks in...	6	6	https://www.allsides.com/news...	2020-12-22 01:33:46.139543
8	8 Capitalism Delivered Promise, and Faced a Reckoning, in 2020	On Dec. 13, the Detroit Free Pres...	7	7	https://www.allsides.com/news...	2020-12-22 01:33:46.139543

```sql
-- Join Category and Headline Table
SELECT
    c.category_name,
    COUNT(*) AS frequency
FROM
    headlines h
    JOIN category c
    ON h.category_id = c.category_id
GROUP BY
    c.category_name
ORDER BY
    frequency DESC;

-- Join all tables / Find specific articles
SELECT
    h.article_id,
    h.article_name,
    s.source_name,
```

ta Output | Explain | Messages | Notifications

category_name character varying (100)	frequency bigint
Politics	8
Business	6
World news	6
Coronavirus	5
Joe Biden	5
Elections	4
Opinion	4
Voting Rights and Voter Fraud	4



```sql
-- The 'ol Trump Lookup
SELECT
    h.article_id,
    h.article_name,
    c.category_name,
    s.source_name
FROM
    headlines h
    JOIN category c ON h.category_id = c.category_id
    JOIN sources s ON h.source_id = s.source_id
WHERE
    h.article_name LIKE '%rump%';


-- Unidentified Category Lookup
SELECT
```

Output | Explain | Messages | Notifications

article_id integer	article_name character varying (500)	category_name character varying (100)	source_name character varying (100)
5	Trump Is Losing His Mind	Politics	The Atlantic
6	'A real mess': Trump is leaving behind crises and undermining Biden ...	Politics	Washington Post
12	Trump Sought to Tap Conspiracy Theorist Sidney Powell as Special ...	Voting Rights and Voter Fraud	Slate
75	The rotten state of American politics made Trump smell fresh	Opinion	The Guardian
39	Trump campaign takes fight over Penn. election, ballot laws to Supr...	Elections	Fox News (Online News)