Sathwik Kesappragada
Analyst, Audience Insights – Vice Media
Part I:

1. Use an SQL query to extract the ==article_id== from the URL. Please paste the query in the box below. You may also link to a separate sheet/document if necessary.

   article_id length is 6 across all values however we are dealing with varying lengths of url

```sql
SELECT SUBSTRING(url, POSITION('icle/' in url) + 5, 6)
FROM case_study;
```

2. Using the ==published_date== field, calculate the unique daily and weekly article output. How does content output trend throughout the week?

The observed time frame of this dataset is from 05/01/2021 to 07/31/2021, however instead of having 92 days we are missing 5 dates, specifically ['2021-05-02', '2021-05-08', '2021-06-13', '2021-06-20','2021-07-03'].

```sql
-- unique daily output, 87 days in total
SELECT published_date,
COUNT(DISTINCT article_id)
FROM case_study
GROUP BY published_date;
-- unique weekly output, 14 weeks
SELECT
COUNT(DISTINCT article_id),
EXTRACT('week' FROM published_date) as "week number"
FROM case_study
GROUP BY 2
ORDER BY 1 DESC;
-- The day of week Sunday (0) to Saturday (6)
-- article trend throughout week
SELECT
COUNT(DISTINCT article_id),
EXTRACT('dow' FROM published_date) as "day of week"
FROM case_study
GROUP BY 2;
```

count bigint	day of week numeric
15	0
219	1
228	2
233	3
258	4
205	5
25	6

- On weekends less content gets published
- Marginally exponential growth in article output throughout the weekdays with the exception of Friday
- During the weekdays on average there are 229 articles produced
- During the weekends on average there are 20 articles produced

3. VICE's editorial department is structured into different sections (e.g. Motherboard, US Culture and Horoscope). How does each <mark>section</mark> contribute to editorial output?

```sql
SELECT
COUNT(DISTINCT article_id) as "total articles",
SUM(page_views) as "total traffic",
vertical
FROM case_study
GROUP BY 3;
```

total articles bigint	total traffic bigint	vertical character varying
144	6611382	Horoscope
558	17288463	Motherboard
484	6791201	US Culture

Motherboard leads the pack with the most traffic and articles generated. US Culture follows second in both categories and Horoscope falls at the bottom with significantly much less content however if we look at (sum of views/ # of unique articles), we get an interesting finding telling us that although there is less supply, on average horoscope articles get more views. US Culture isn't getting as many impressions as we would like. There may need to be an improvement in generated material to attract a lot more users or develop a strategy to increase reach by focusing on search and posting frequently on social media.

- 56.33% of traffic curious about tech-related news
- 22.13% of traffic headed towards culture side
- 21.54% of traffic wanted to engage with astrology
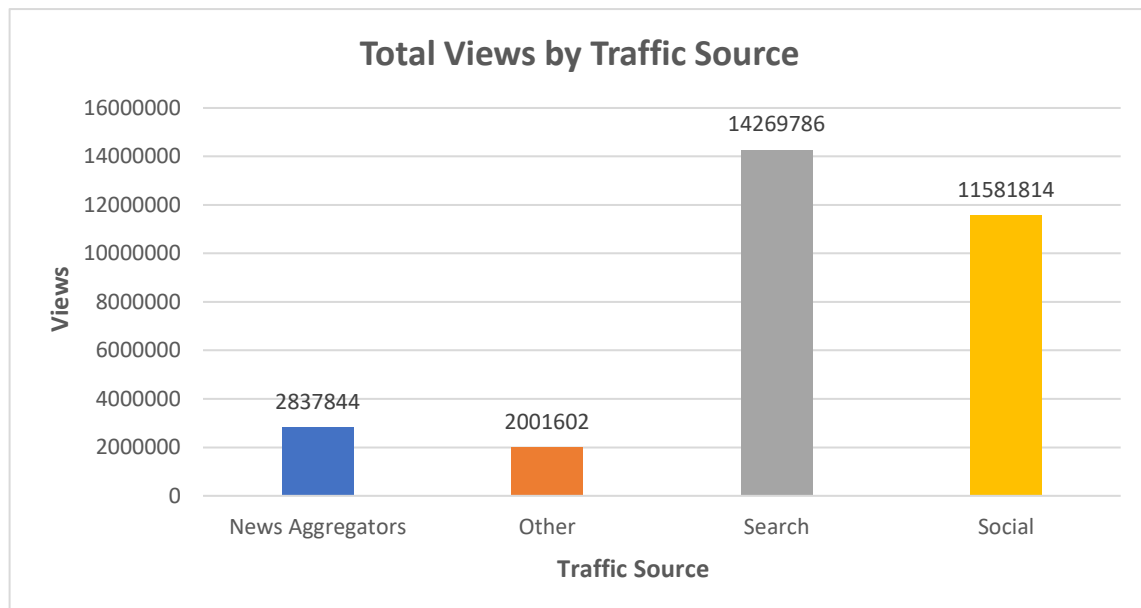
Total traffic/ Total articles

- 45912 views – Horoscope
- 30983 views – Motherboard
- 14,031 views – US Culture

4. Write a query creating a table that shows:
   a. each section
   b. its unique content output
   c. its traffic contribution
   d. section's share of the total output and traffic cover the full data period.

```sql
CREATE TABLE vice_data(
    sector VARCHAR NOT NULL,
    content_ouput VARCHAR NOT NULL,
    traffic_contribution VARCHAR NOT NULL,
    total_output INT NOT NULL,
)
```

Part II:

1. Create a chart(s) showing the traffic channel distribution split by section. What do you think is the value of each traffic source, and how would you use them to assess performance?

Total Views by Traffic Source

Traffic Source	Views
News Aggregators	2837844
Other	2001602
Search	14269786
Social	11581814

Our two biggest traffic channels are search and social. Successful search engine optimization and social media marketing resulted in a significant number of users arriving at our platform. In a world where nearly 68% of online experiences begin with a search this discovery checks out. Additionally having a strong presence in numerous platforms (16.6 M YouTube, 4.9 M Instagram, 2 M Twitter, 10.5 M Facebook, 7.8 M Snapchat) increases visibility. It is good to see about 84% of our viewership comes from two sources that we have some control over. According to Chartbeat traffic driven by news aggregators have been trending downward but it is great for diversification purposes. Continued efforts towards optimizing search strategy and social marketing will drive viewership.

- ~46.49% Search
- ~37.74% Social
- ~9.25% News Aggregators
- ~6.52% Other

2. Using your preferred mix of tables and charts, give a sense of how content output and traffic is distributed across the locales and languages. Remember that despite sharing different locales, en_us and en_uk belong to the same language.

a. Is there a least / most effective language for content to be published in? Is there a difference in traffic source mix depending on language? Does the number of translations for an article appear to be related to its total traffic?
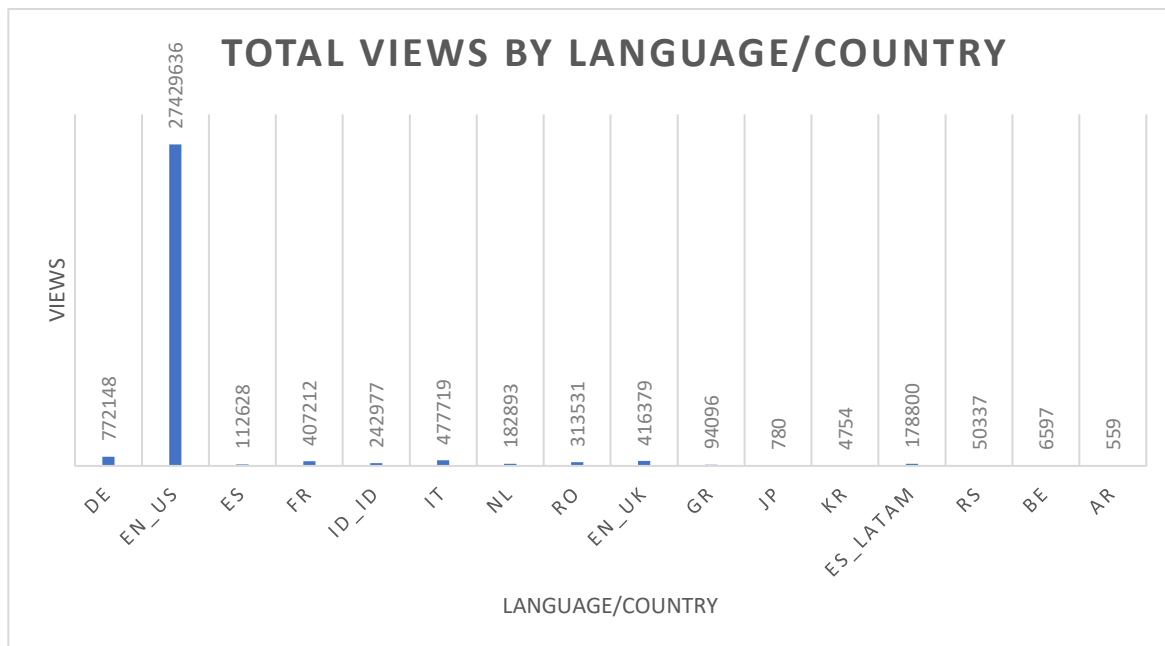
US English has the most views by a landslide and Denmark placing at second.

- 90% of total views are on English content
- 70% of Denmark views come from Search
- Most locales/language viewership arrive from social media
- Similar to US English and Denmark, most views for countries Japan, Korea, and France are from Search

It is not fair to expect all readers to act as a monolith. English-speaking material is getting consumed the most which is acceptable however more effort needs to be put into translations to serve other countries. There is an imbalance in the dataset so not all languages/countries are getting fairly captured.

locale	traffic_channel	page_views
ar	News Aggregators	4
	Other	4
	Search	261
	Social	290
be	News Aggregators	14
	Other	220
	Search	890
	Social	5473
de	News Aggregators	51792
	Other	10680
	Search	543078
	Social	166598
en_uk	News Aggregators	82021
	Other	7993
	Search	95977
	Social	230388
en_us	News Aggregators	2544475
	Other	1880944
	Search	13020441
	Social	9983776
es	News Aggregators	3311
	Other	4396
	Search	27913
	Social	77008
es_latam	News Aggregators	4267
	Other	20644
	Search	62928
	Social	90961
fr	News Aggregators	108020
	Other	7914
	Search	209239
	Social	82039
gr	News Aggregators	223
	Other	47865
	Search	12995
	Social	33013
id_id	News Aggregators	495
	Other	1989
	Search	22285
	Social	218208
it	News Aggregators	36755
	Other	9996
	Search	143869
	Social	287099
jp	News Aggregators	182
	Other	50
	Search	278
	Social	270
kr	News Aggregators	17
	Other	3242
	Search	1225
	Social	270
nl	News Aggregators	4073
	Other	3282
	Search	24099
	Social	151439
ro	News Aggregators	1718
	Other	1902
	Search	91783
	Social	218128
rs	News Aggregators	477
	Other	481
	Search	12525
	Social	36854

3. For the last part, we want to issue a recommendation on future content investment. Using the sample data provided, group the content by section and provide quantitative *and* qualitative recommendations about how VICE should increase output most effectively from a traffic standpoint.

   a. This is an open-ended question that can build on the insights obtained in the prior steps. While the data as given is split by section, you should feel free to further segment the content based on language or topic. For example, there are a lot of articles within the US Culture section that contain recipes, and these can be segmented out.

   b. In this part, you do not need to create a data framework explaining all 1,183 articles. However, if you are making the case that a topic, such as recipes, should adjust output up or down, then try to include a metric to support your point. In this simplified scenario, the goal is to maximize pageviews and minimize inefficient content output.

   c. Don't spend too much time on this last part, the goal here is to get a sense of how you approach thinking about and justifying content recommendations.

Vice's mission is to connect the world's curious individuals through meaningful storytelling by crusading against the convention and uncovering compelling facts. People mostly use the internet to keep up to date with news and events, so it is vital to increase user retention and reduce attrition rate.
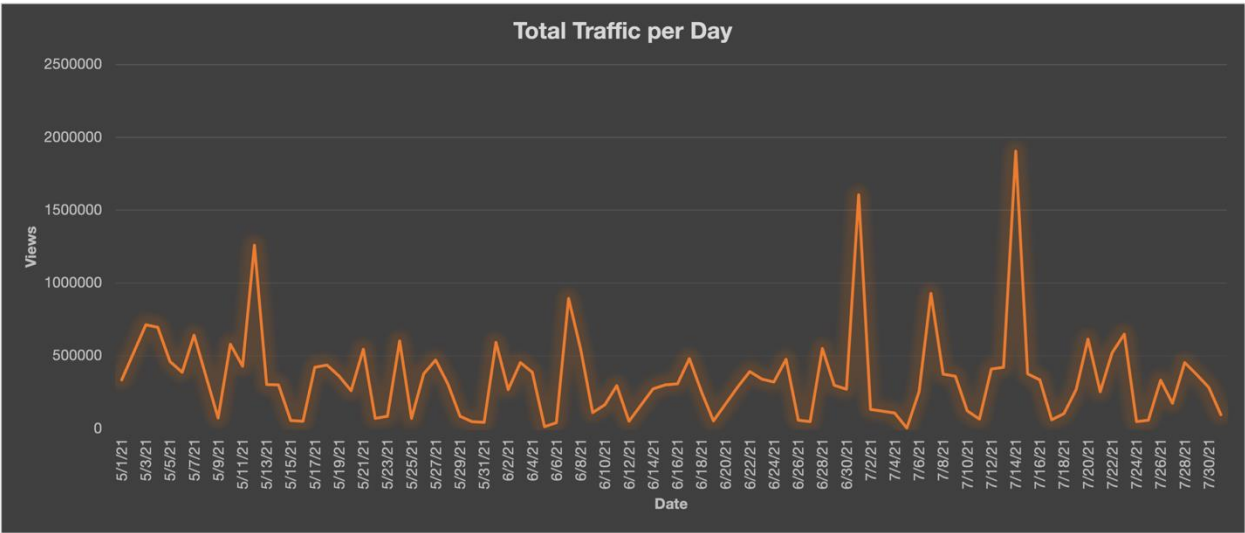

Looking at unique daily and weekly article output vs total traffic per day, we can notice a trend in spike during the middle of every week and weekends tend to be more mellow. It is in our best interest to publishing most works between Monday-Thursday. Additionally recalling back to each vertical's contribution to editorial output, we see that horoscope has a commanding lead in average views per article and this is possibly because people want to increase self-awareness especially during times of stress. Despite having 3 times as less articles, horoscope viewership was only off by 2.65% compared to that of US Culture.

When performing text analysis on the unique article titles segmented by vertical to get a glance at what topics are being discussed most frequently, we can detect patterns and recurring words.
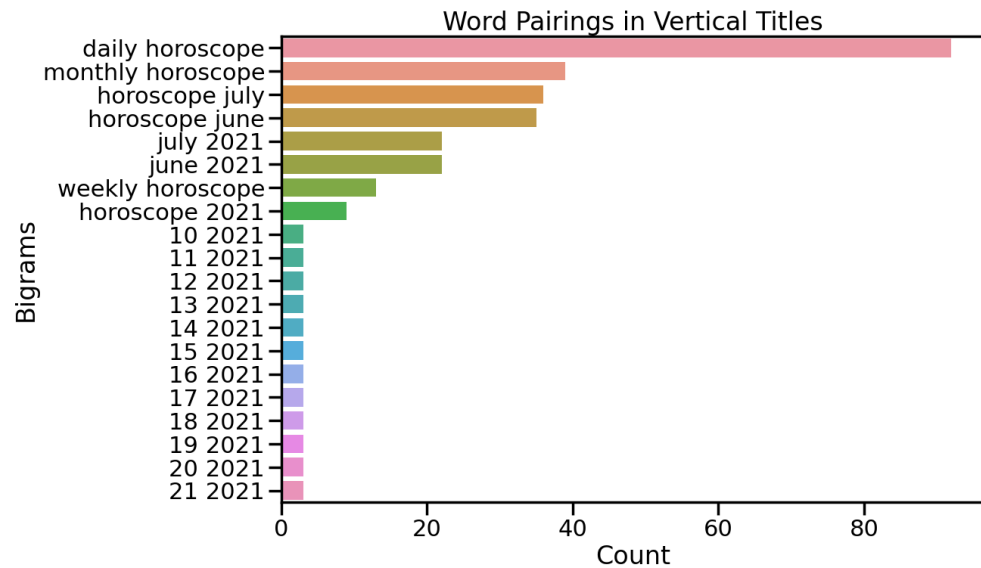
For Motherboard:

For US Culture:

Word Pairings in Vertical Titles

For Horoscope:

Word Pairings in Vertical Titles
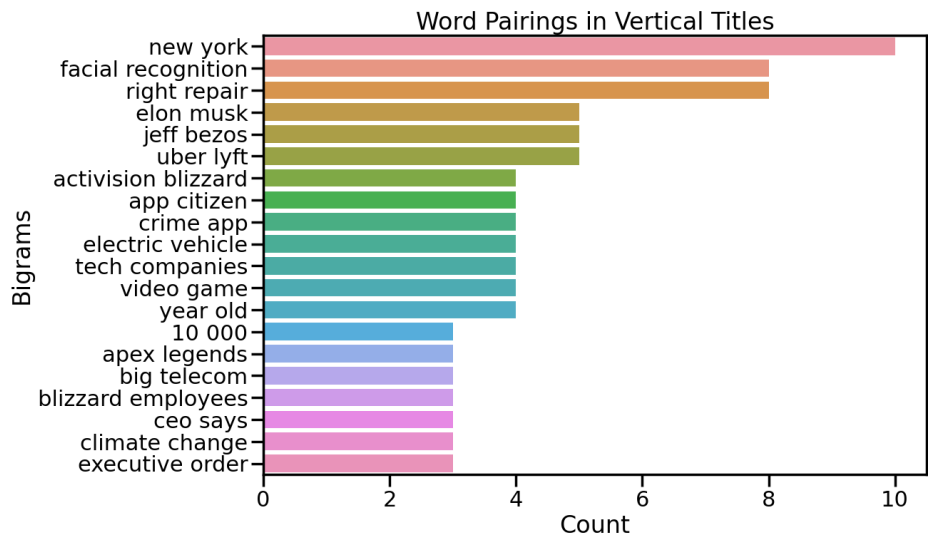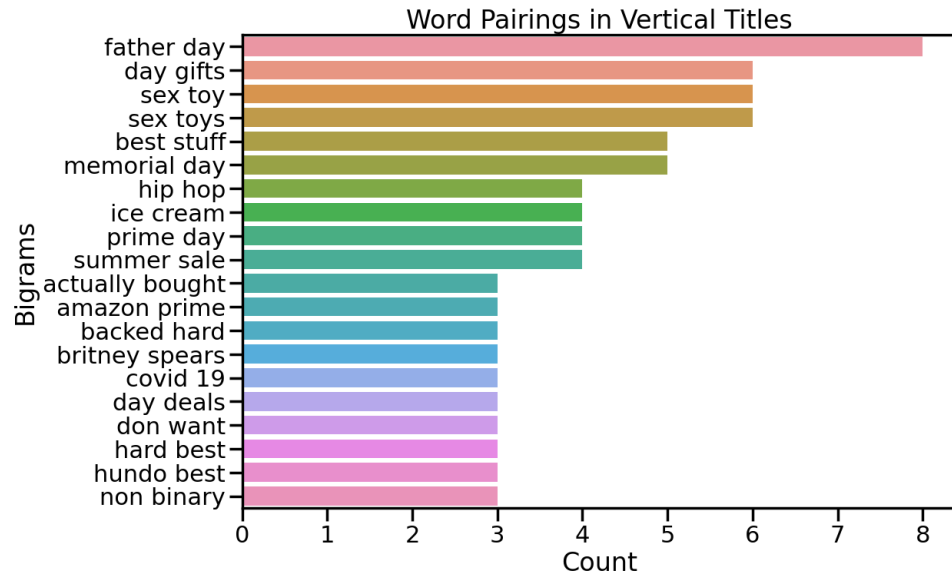
Viewers surf the web to seek information and stay up to date so covering trendy topics yields to a lot of readers. We can spot that US Culture has quite a few articles covering sex toys and even though this sector generates a lot of material focusing on the quality should be emphasized. On the other hand, if we look at vertical views from each respective traffic channel, we can make a case that to get more eyes on US culture articles, we may need to improve our search engine strategy whether it is to use keywords appropriately, adding internal links, pinpointing target audience, or optimizing titles/descriptions. Catchy headlines and relevant content increases clicks.

Grouped by Traffic Channel & Vertical Views
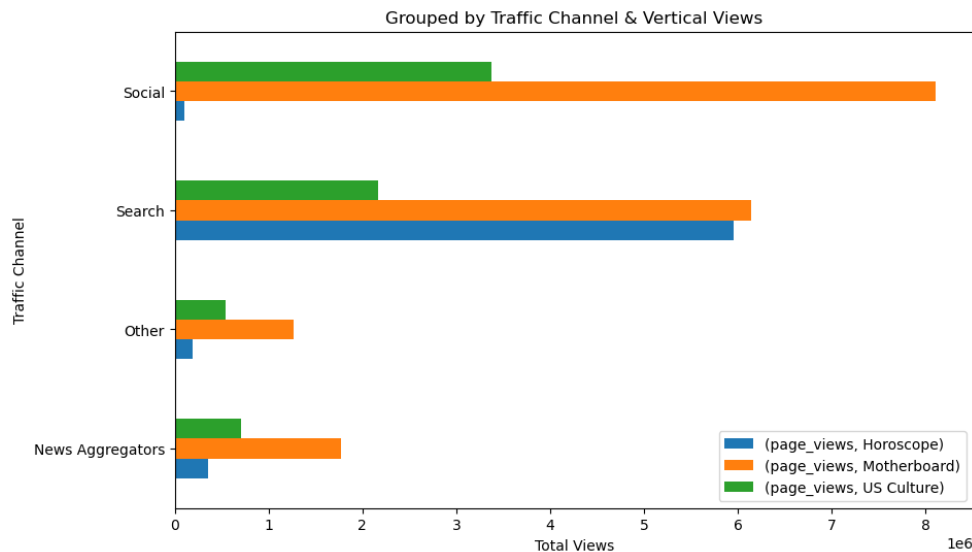
Surprising to see most views for Horoscope came from Search and not social media. An increase in posts on those platforms can potentially drive-up impressions. Noticeably a great amount of traffic like to get informed on tech related news and we would want to replicate its demand for the other verticals.

- Increase horoscope content output, market on social media more
- Realign focus on US Culture by identifying blockers and staying relevant
- Continue staying up to date with tech news

Afterthoughts

To add a notion of duration to these visits/views, we can track the average user session time spent on article/platform, and average daily, weekly, and monthly time spent on platform. Knowing the total number of article views isn't enough information for Vice's product goal of maximizing pageviews and minimizing inefficient content output. To measure the depth and quality of engagement, we can track important user impressions per monthly basis such as article likes, comments, shares, dwell time, and bounce/click-through rate (related section). By weighing the value of an article views vs article likes vs article comments vs article shares differently, we can more accurately capture the engagement activity in one metric. Metrics like ad impressions and ad clicks could be bundled under the umbrella of engagement.

Sentiment analysis on comments from all platforms can be used as a metric to better understand how to produce much more captivating work.

To improve metrics, we can try personalizing the articles to the user that boost engagement and encourage reporters to post more to keep the feed fresh. Having different post types, like live streams or stories, can aid in keeping the content inventory engaging.

Another approach to improving metrics would be to build a model using features that are believed to affect the metric. Usually this would involve, user data (demographics) and event data (browsing behaviors and session events). We can determine factors contributing to target metric and decide on path forward.

Lastly another way to tackle this is to turn our attention towards recommendations. Discovery may be broken so people may not be aware of US culture articles. We can look outside of Vice for consumer surveys to get a benchmark engagement for US culture articles to see if the demand is generally less.