Queries used for Visualizations

Query 1: Regioned Based Query:

```
SELECT
    SUM(CASE WHEN abstract LIKE '%New York%' OR abstract LIKE '%America%' THEN 1 ELSE 0 END) AS America_count,
    SUM(CASE WHEN abstract LIKE '%Asia%' OR abstract LIKE '%China%' OR abstract LIKE '%Europe%' OR abstract LIKE '%Africa%'
    OR abstract LIKE '%Australia%' OR abstract LIKE '%Russia%' OR abstract LIKE '%India%' OR abstract LIKE '%London %' OR abstract LIKE '%Ukraine%'
    OR abstract LIKE '%Israel%' THEN 1 ELSE 0 END) AS Other_Continents_count
FROM  {{ source ('NYT_DB', 'ARTICLE') }}
```

Query 2: Number of articles by Month:

```
SELECT
    TO_CHAR(pub_date, 'YYYY-MM') AS publication_month,
    COUNT(*) AS num_articles
FROM
    {{ source ('NYT_DB', 'ARTICLE') }}
GROUP BY
    publication_month
ORDER BY
    publication_month
```

Query 3: Top 5 Keywords by group

```
WITH RankedKeywords AS (
  SELECT keyword_name, keyword_value,
         RANK() OVER (PARTITION BY keyword_name ORDER BY COUNT(*) DESC) AS Rank_keywords
  FROM {{ source ('NYT_DB', 'KEYWORDS')}}
  GROUP BY keyword_name, keyword_value
)
SELECT keyword_name, keyword_value
FROM RankedKeywords
WHERE Rank_keywords <= 5
```

Query 4: Number_of_articles by Type_of_material

```
SELECT
    type_of_material,
    COUNT(*) AS num_articles
FROM
    {{ source ('NYT_DB', 'ARTICLE') }}
GROUP BY
    type_of_material
```

Query 5:Number_of_articles belong to each section

```
select section.section_name as section_name, count(fact_nyt.article_id) as article_count from {{ source('NYT_DB', 'SECTION') }} as section join {{ source('NYT_DB', 'FACT_NYT') }}

as fact_nyt on section.section_id = fact_nyt.section_id group by section.section_name
```

Query 6:Distribution of keywords across different sections of articles

```
select section.section_name, array_agg(distinct keywords.keyword_value) AS
keyword_names from {{ source('NYT_DB', 'SECTION') }}

inner join {{ source ('NYT_DB', 'FACT_NYT') }} on section.section_id =
fact_nyt.section_id inner join {{ source ('NYT_DB', 'ARTICLE_KEYWORD') }}

on fact_nyt.article_id = article_keyword.article_id inner join {{ source ('NYT_DB',
'KEYWORDS')}} on article_keyword.keyword_id = keywords.keyword_id group by
section.section_name
```

Query 7:Number of Articles and Average Word Count of each authors:

```
with author_table as (

    select *

    from {{source('nyt_db', 'author')}}

),

article_author_table as (

    select *

     from {{source('nyt_db', 'article_author')}}

),

article_table as (

    select *

     from {{source('nyt_db', 'article')}}

)

SELECT

    author_table.firstname,

           author_table.lastname,
    COUNT(article_author_table.articleid) AS num_articles,
    AVG(article_table.word_count) AS avg_word_count
FROM
    author_table
JOIN
    article_author_table ON author_table.authorid = article_author_table.authorid
JOIN
    article_table ON article_author_table.articleid = article_table._id
GROUP BY
    author_table.authorid, author_table.firstname, author_table.lastname
```

Query 8: Article Section Classification: Metrics

```
SELECT
        SECTION_NAME AS group_name,
        COUNT(*) AS actual_count,
        SUM(CASE WHEN SECTION_NAME = PREDICTED_SECTION_NAME THEN 1 ELSE 0 END) AS correct_predictions,
        COUNT(*) - SUM(CASE WHEN SECTION_NAME = PREDICTED_SECTION_NAME THEN 1 ELSE 0 END) AS incorrect_predictions,
        ROUND((correct_predictions/actual_count)*100, 2) AS accuracy
    FROM
        classifications_results
    GROUP BY
        SECTION_NAME
```

```
ORDER BY
    correct_predictions DESC;
```