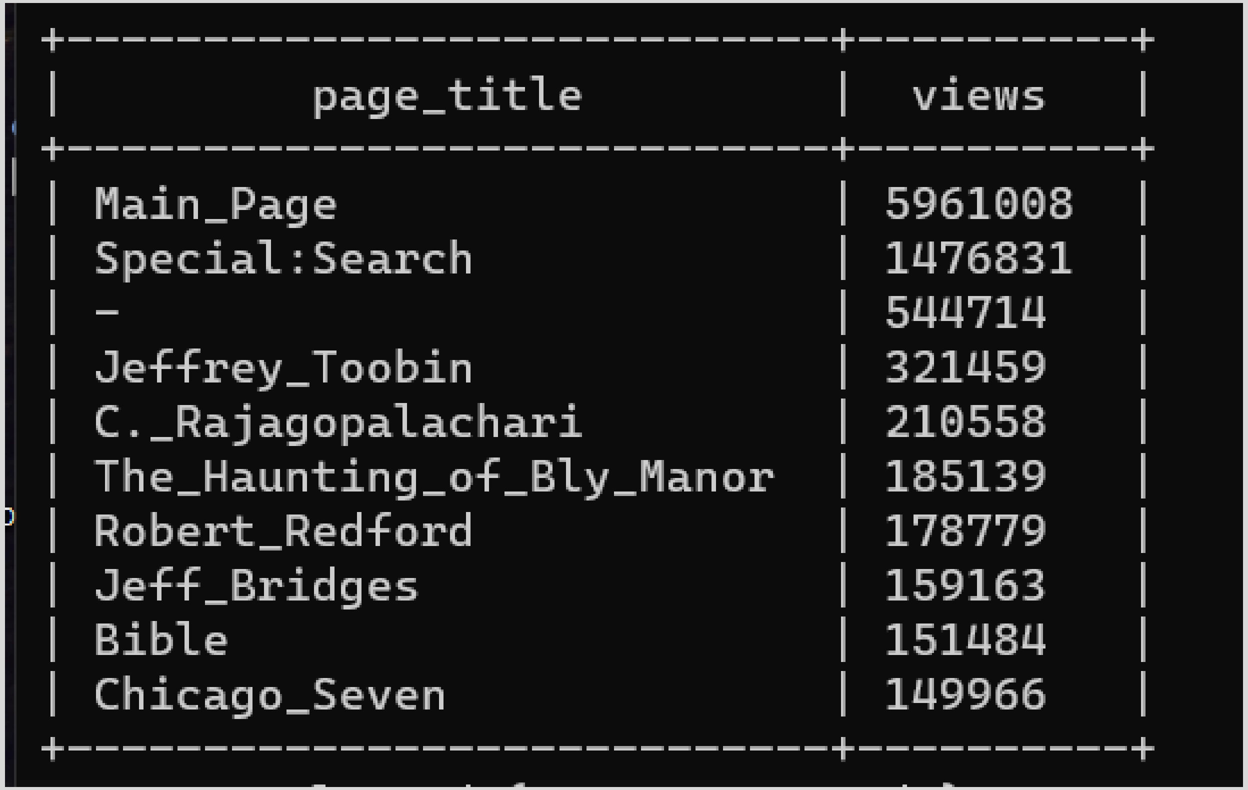
PROJECT_1

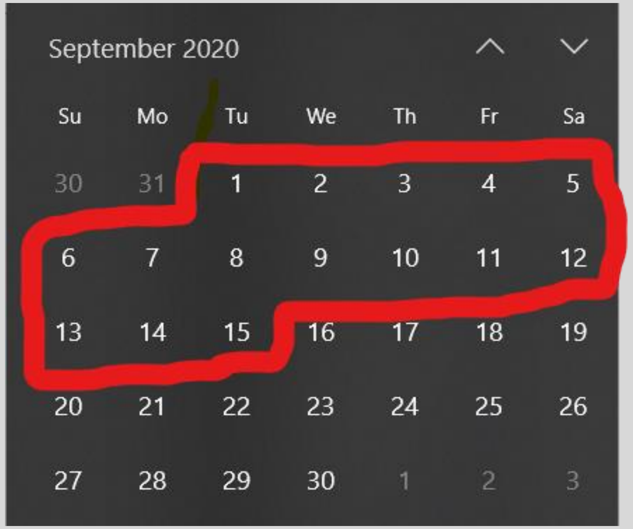
VICTOR PULLAS

1. Which English Wikipedia article got the most traffic on **October 20**?



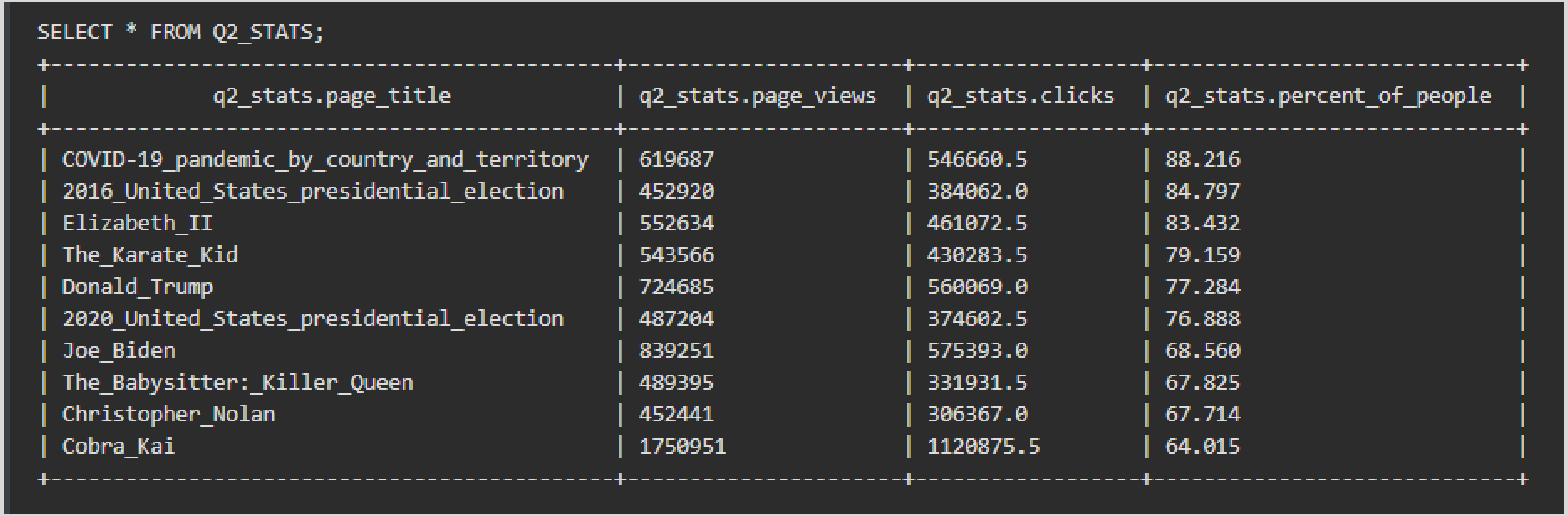
```
+-----------------------------+-----------+
|         page_title          |   views   |
+-----------------------------+-----------+
| Main_Page                   | 5961008   |
| Special:Search              | 1476831   |
| -                           | 544714    |
| Jeffrey_Toobin              | 321459    |
| C._Rajagopalachari          | 210558    |
| The_Haunting_of_Bly_Manor   | 185139    |
| Robert_Redford              | 178779    |
| Jeff_Bridges                | 159163    |
| Bible                       | 151484    |
| Chicago_Seven               | 149966    |
+-----------------------------+-----------+
```

2. What **English** wikipedia article has the largest fraction of its readers follow an **internal link** to another wikipedia article?



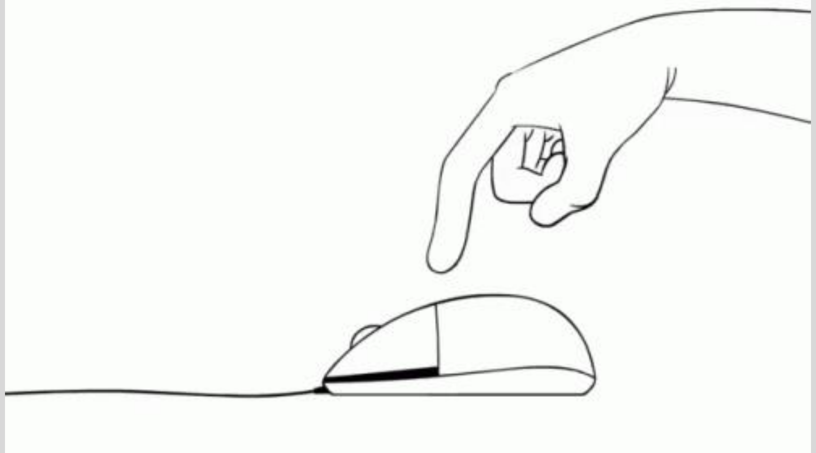
September 2020

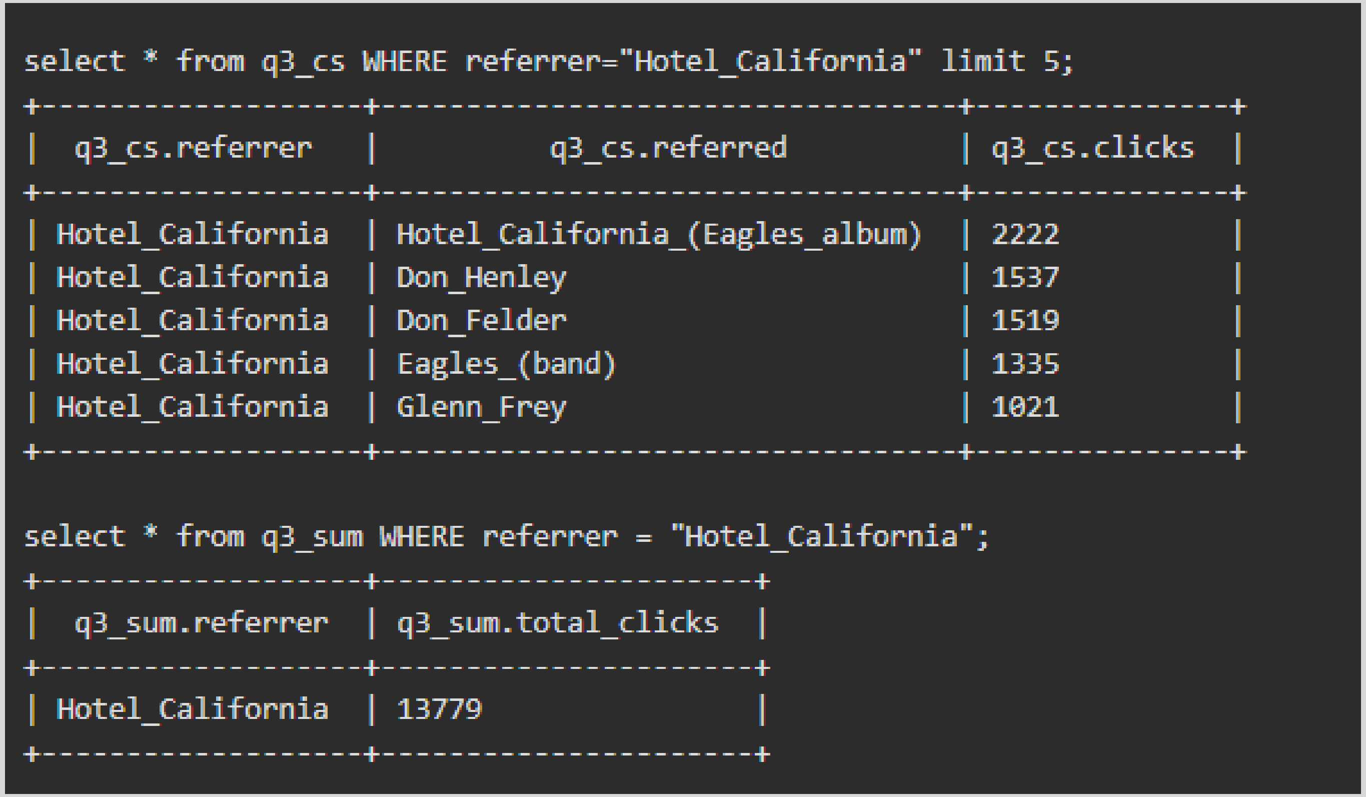
Su	Mo	Tu	We	Th	Fr	Sa
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	1	2	3



```
SELECT * FROM Q2_STATS;
+-------------------------------------------+----------------------+------------------+---------------------------+
|            q2_stats.page_title            | q2_stats.page_views  | q2_stats.clicks  | q2_stats.percent_of_people |
+-------------------------------------------+----------------------+------------------+---------------------------+
| COVID-19_pandemic_by_country_and_territory | 619687               | 546660.5         | 88.216                    |
| 2016_United_States_presidential_election  | 452920               | 384062.0         | 84.797                    |
| Elizabeth_II                              | 552634               | 461072.5         | 83.432                    |
| The_Karate_Kid                            | 543566               | 430283.5         | 79.159                    |
| Donald_Trump                              | 724685               | 560069.0         | 77.284                    |
| 2020_United_States_presidential_election  | 487204               | 374602.5         | 76.888                    |
| Joe_Biden                                 | 839251               | 575393.0         | 68.560                    |
| The_Babysitter:_Killer_Queen              | 489395               | 331931.5         | 67.825                    |
| Christopher_Nolan                         | 452441               | 306367.0         | 67.714                    |
| Cobra_Kai                                 | 1750951              | 1120875.5        | 64.015                    |
+-------------------------------------------+----------------------+------------------+---------------------------+
```

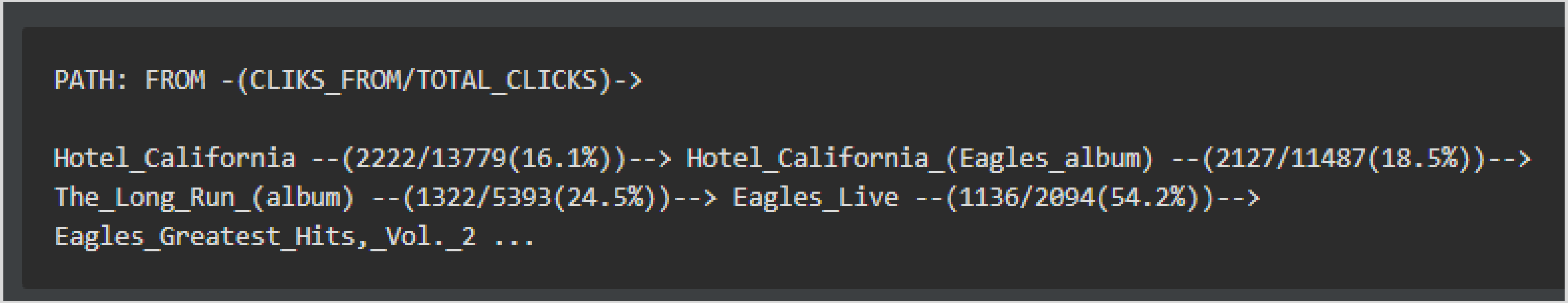
3. What series of Wikipedia articles, starting with Hotel California, keeps the largest fraction of its readers clicking on internal links? This is similar to (2), but you should continue the analysis past the first article.





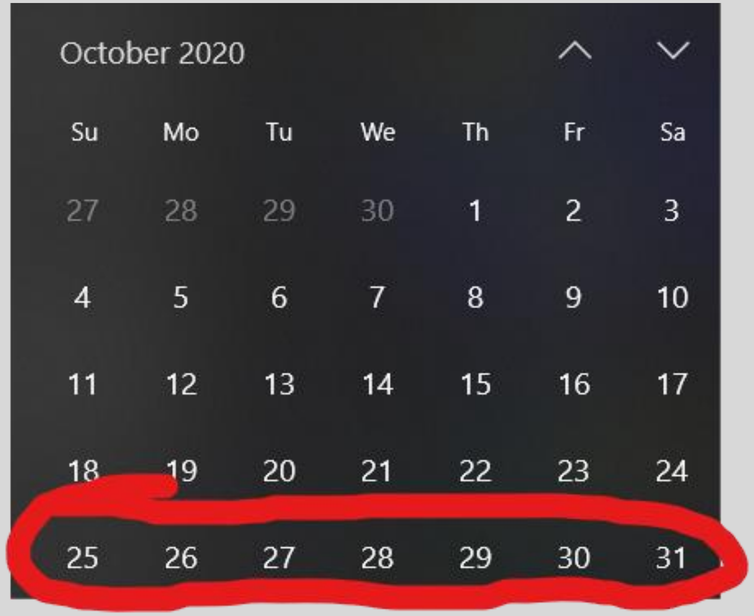
```
select * from q3_cs WHERE referrer="Hotel_California" limit 5;
+-------------------+----------------------------------+---------------+
|  q3_cs.referrer   |          q3_cs.referred          | q3_cs.clicks  |
+-------------------+----------------------------------+---------------+
| Hotel_California  | Hotel_California_(Eagles_album)  | 2222          |
| Hotel_California  | Don_Henley                       | 1537          |
| Hotel_California  | Don_Felder                       | 1519          |
| Hotel_California  | Eagles_(band)                    | 1335          |
| Hotel_California  | Glenn_Frey                       | 1021          |
+-------------------+----------------------------------+---------------+

select * from q3_sum WHERE referrer = "Hotel_California";
+-------------------+----------------------+
|  q3_sum.referrer  | q3_sum.total_clicks  |
+-------------------+----------------------+
| Hotel_California  | 13779                |
+-------------------+----------------------+
```

Question 3 Continued:



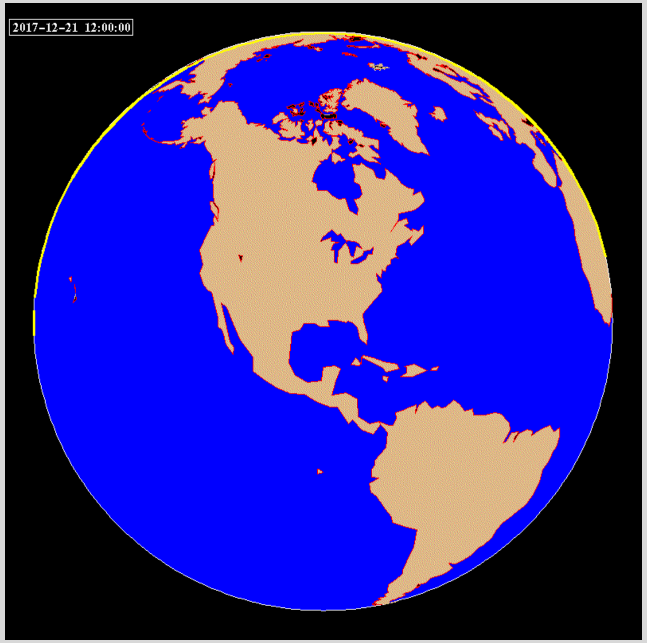
```
PATH: FROM -(CLIKS_FROM/TOTAL_CLICKS)->

Hotel_California --(2222/13779(16.1%))--> Hotel_California_(Eagles_album) --(2127/11487(18.5%))-->
The_Long_Run_(album) --(1322/5393(24.5%))--> Eagles_Live --(1136/2094(54.2%))-->
Eagles_Greatest_Hits,_Vol._2 ...
```

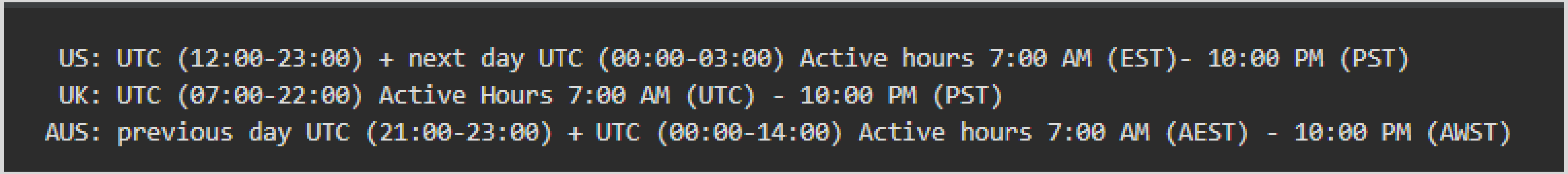
4. Find an example of an English Wikipedia article that is relatively more popular in the UK. Find the same for the US and Australia.



October 2020

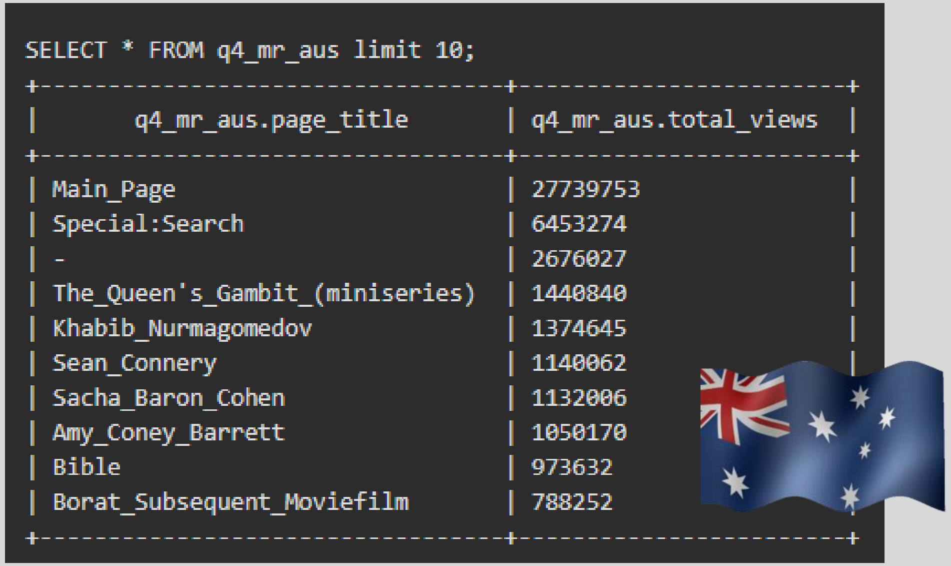
Su	Mo	Tu	We	Th	Fr	Sa
27	28	29	30	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31



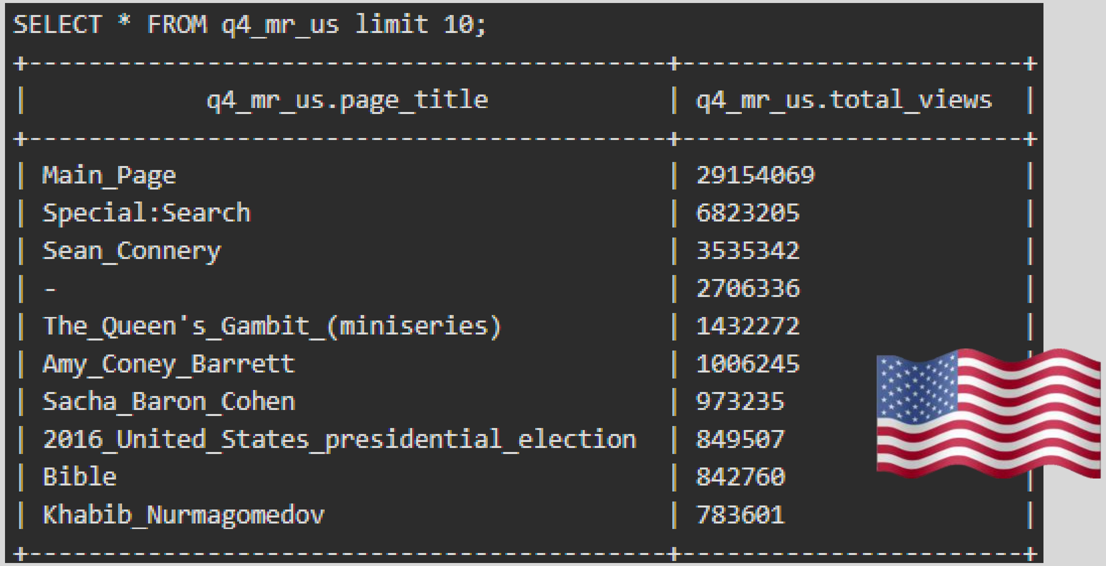


```
 US: UTC (12:00-23:00) + next day UTC (00:00-03:00) Active hours 7:00 AM (EST)- 10:00 PM (PST)
 UK: UTC (07:00-22:00) Active Hours 7:00 AM (UTC) - 10:00 PM (PST)
AUS: previous day UTC (21:00-23:00) + UTC (00:00-14:00) Active hours 7:00 AM (AEST) - 10:00 PM (AWST)
```

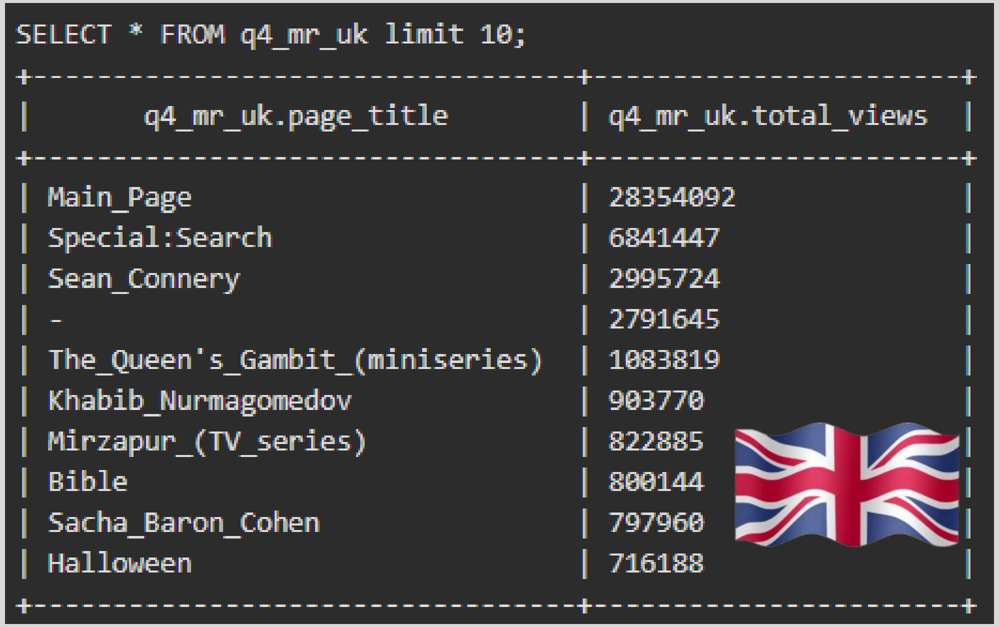
Question 4: Continued



```
SELECT * FROM q4_mr_aus limit 10;
+-----------------------------------+------------------------+
|        q4_mr_aus.page_title       | q4_mr_aus.total_views  |
+-----------------------------------+------------------------+
| Main_Page                         | 27739753               |
| Special:Search                    | 6453274                |
| -                                 | 2676027                |
| The_Queen's_Gambit_(miniseries)   | 1440840                |
| Khabib_Nurmagomedov               | 1374645                |
| Sean_Connery                      | 1140062                |
| Sacha_Baron_Cohen                 | 1132006                |
| Amy_Coney_Barrett                 | 1050170                |
| Bible                             | 973632                 |
| Borat_Subsequent_Moviefilm        | 788252                 |
+-----------------------------------+------------------------+
```

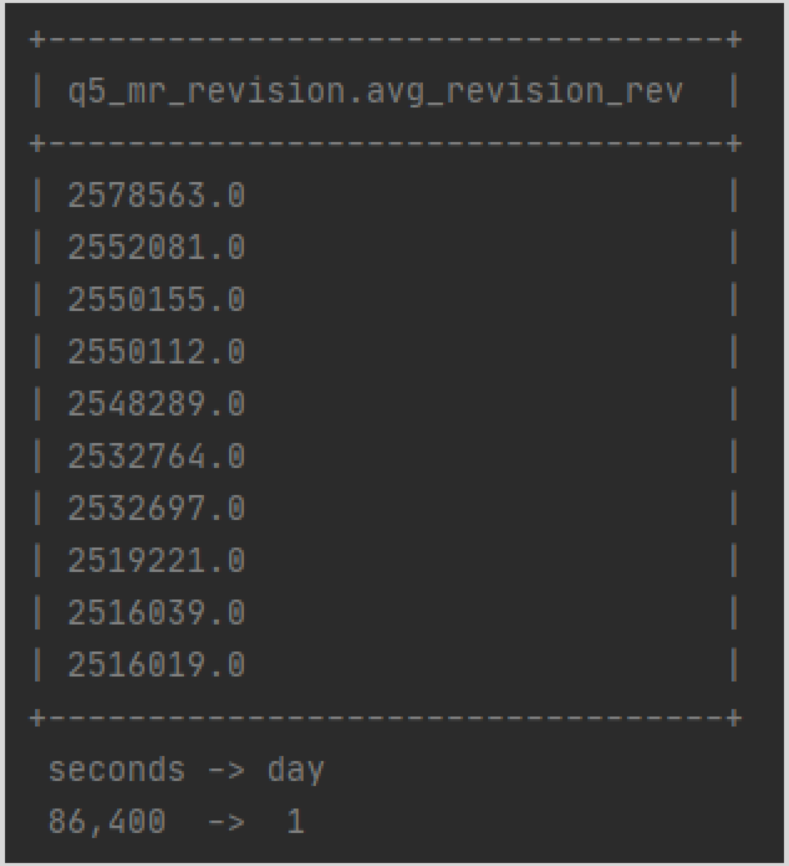


```
SELECT * FROM q4_mr_us limit 10;
+---------------------------------------------+------------------------+
|              q4_mr_us.page_title            | q4_mr_us.total_views   |
+---------------------------------------------+------------------------+
| Main_Page                                   | 29154069               |
| Special:Search                              | 6823205                |
| Sean_Connery                                | 3535342                |
| -                                           | 2706336                |
| The_Queen's_Gambit_(miniseries)             | 1432272                |
| Amy_Coney_Barrett                           | 1006245                |
| Sacha_Baron_Cohen                           | 973235                 |
| 2016_United_States_presidential_election    | 849507                 |
| Bible                                       | 842760                 |
| Khabib_Nurmagomedov                         | 783601                 |
+---------------------------------------------+------------------------+
```

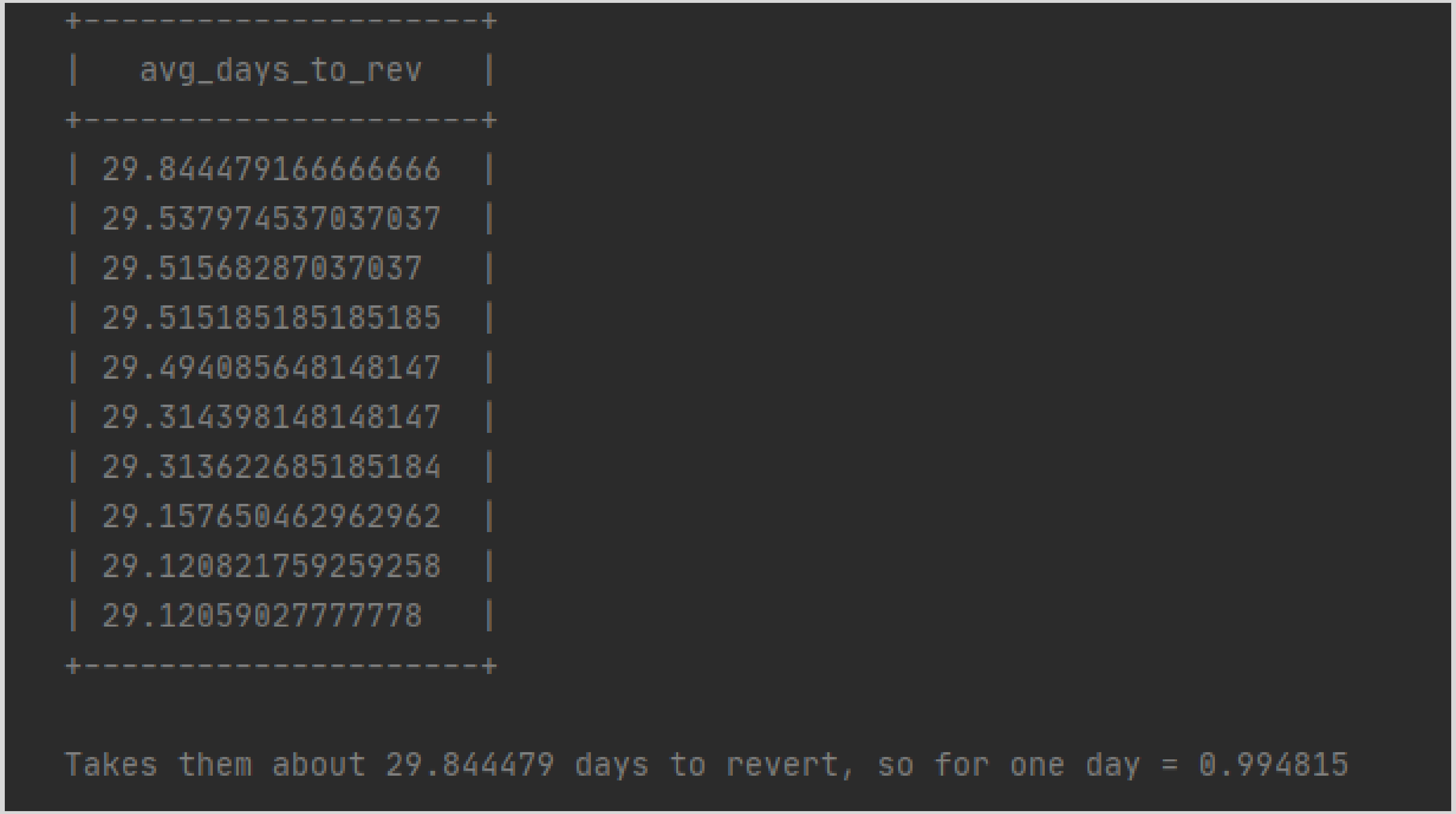


```
SELECT * FROM q4_mr_uk limit 10;
+-----------------------------------+------------------------+
|        q4_mr_uk.page_title        | q4_mr_uk.total_views   |
+-----------------------------------+------------------------+
| Main_Page                         | 28354092               |
| Special:Search                    | 6841447                |
| Sean_Connery                      | 2995724                |
| -                                 | 2791645                |
| The_Queen's_Gambit_(miniseries)   | 1083819                |
| Khabib_Nurmagomedov               | 903770                 |
| Mirzapur_(TV_series)              | 822885                 |
| Bible                             | 800144                 |
| Sacha_Baron_Cohen                 | 797960                 |
| Halloween                         | 716188                 |
+-----------------------------------+------------------------+
```

5. Analyze how many users will see the average vandalized wikipedia page before the offending edit is reversed.

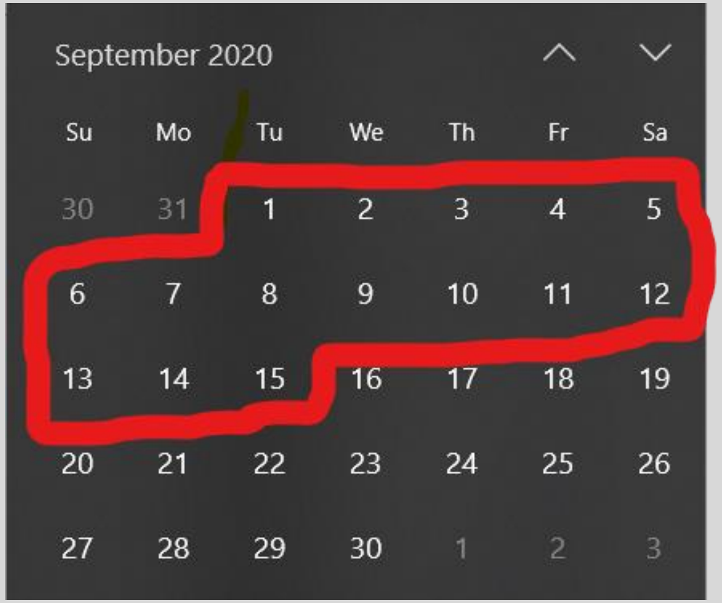


```
+-----------------------------------+
| q5_mr_revision.avg_revision_rev   |
+-----------------------------------+
| 2578563.0                         |
| 2552081.0                         |
| 2550155.0                         |
| 2550112.0                         |
| 2548289.0                         |
| 2532764.0                         |
| 2532697.0                         |
| 2519221.0                         |
| 2516039.0                         |
| 2516019.0                         |
+-----------------------------------+
 seconds -> day
 86,400  ->  1
```



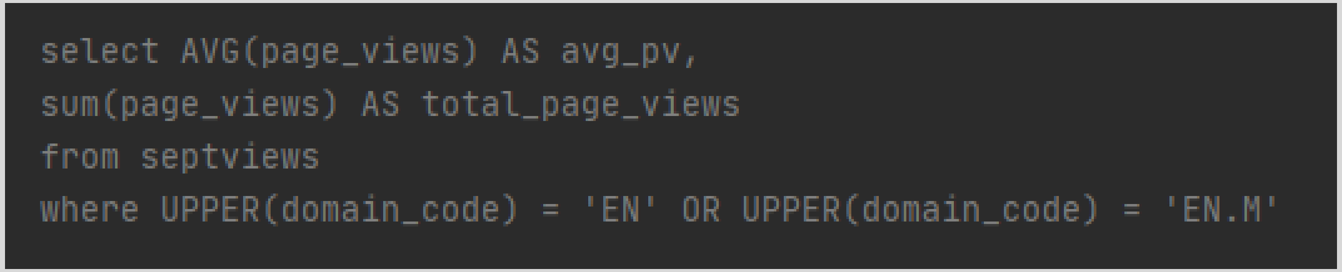
```
+---------------------+
|   avg_days_to_rev   |
+---------------------+
| 29.844479166666666  |
| 29.537974537037037  |
| 29.51568287037037   |
| 29.515185185185185  |
| 29.494085648148147  |
| 29.314398148148147  |
| 29.313622685185184  |
| 29.157650462962962  |
| 29.120821759259258  |
| 29.12059027777778   |
+---------------------+


Takes them about 29.844479 days to revert, so for one day = 0.994815
```

Question 5 Continued:

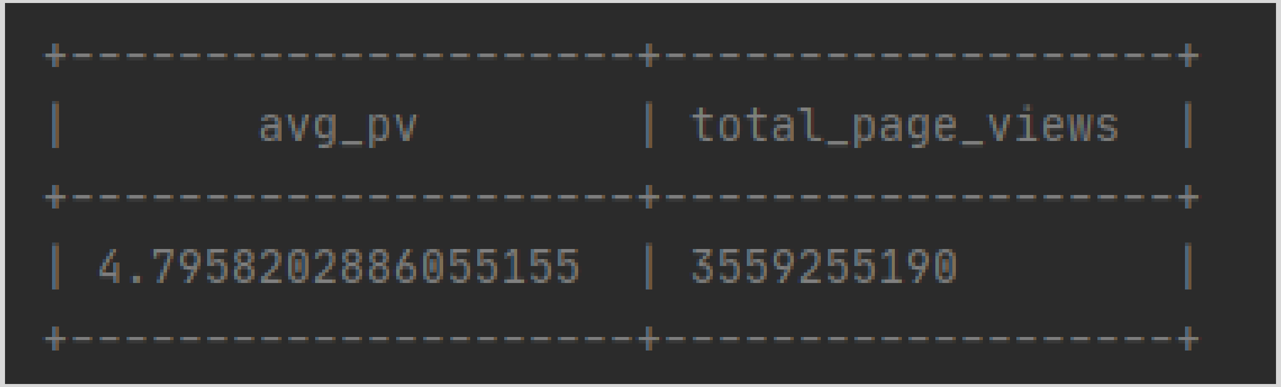


September 2020

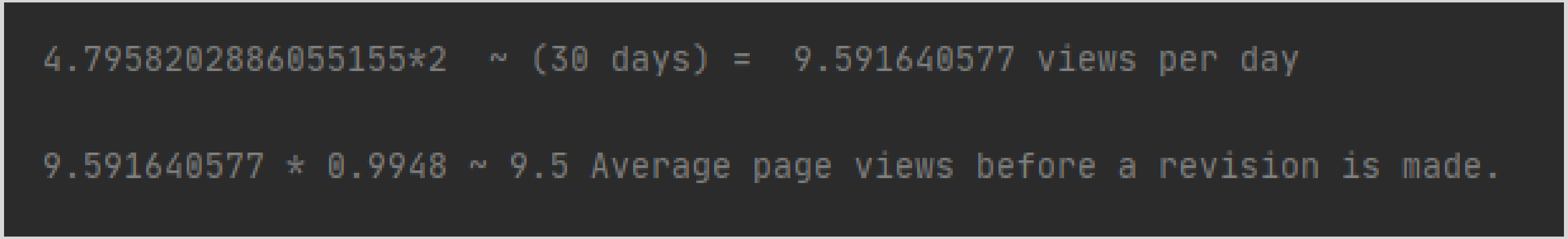
Su	Mo	Tu	We	Th	Fr	Sa
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	1	2	3



```
select AVG(page_views) AS avg_pv,
sum(page_views) AS total_page_views
from septviews
where UPPER(domain_code) = 'EN' OR UPPER(domain_code) = 'EN.M'
```

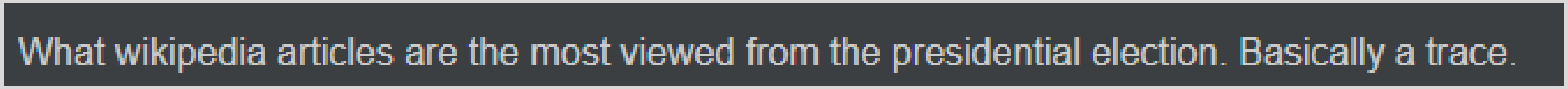


```
+----------------------+--------------------+
|        avg_pv        | total_page_views   |
+----------------------+--------------------+
| 4.7958202886055155   | 3559255190         |
+----------------------+--------------------+
```

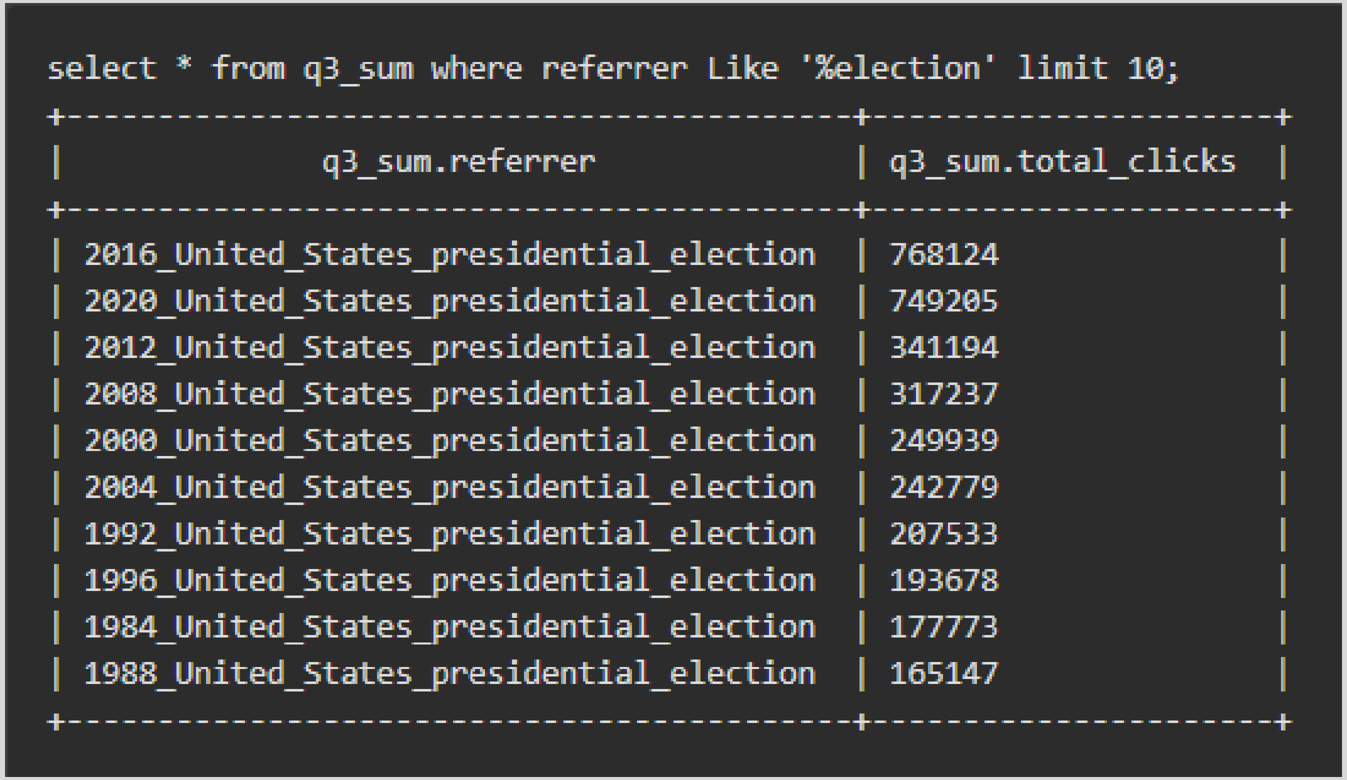


```
4.7958202886055155*2  ~ (30 days) =  9.591640577 views per day

9.591640577 * 0.9948 ~ 9.5 Average page views before a revision is made.
```

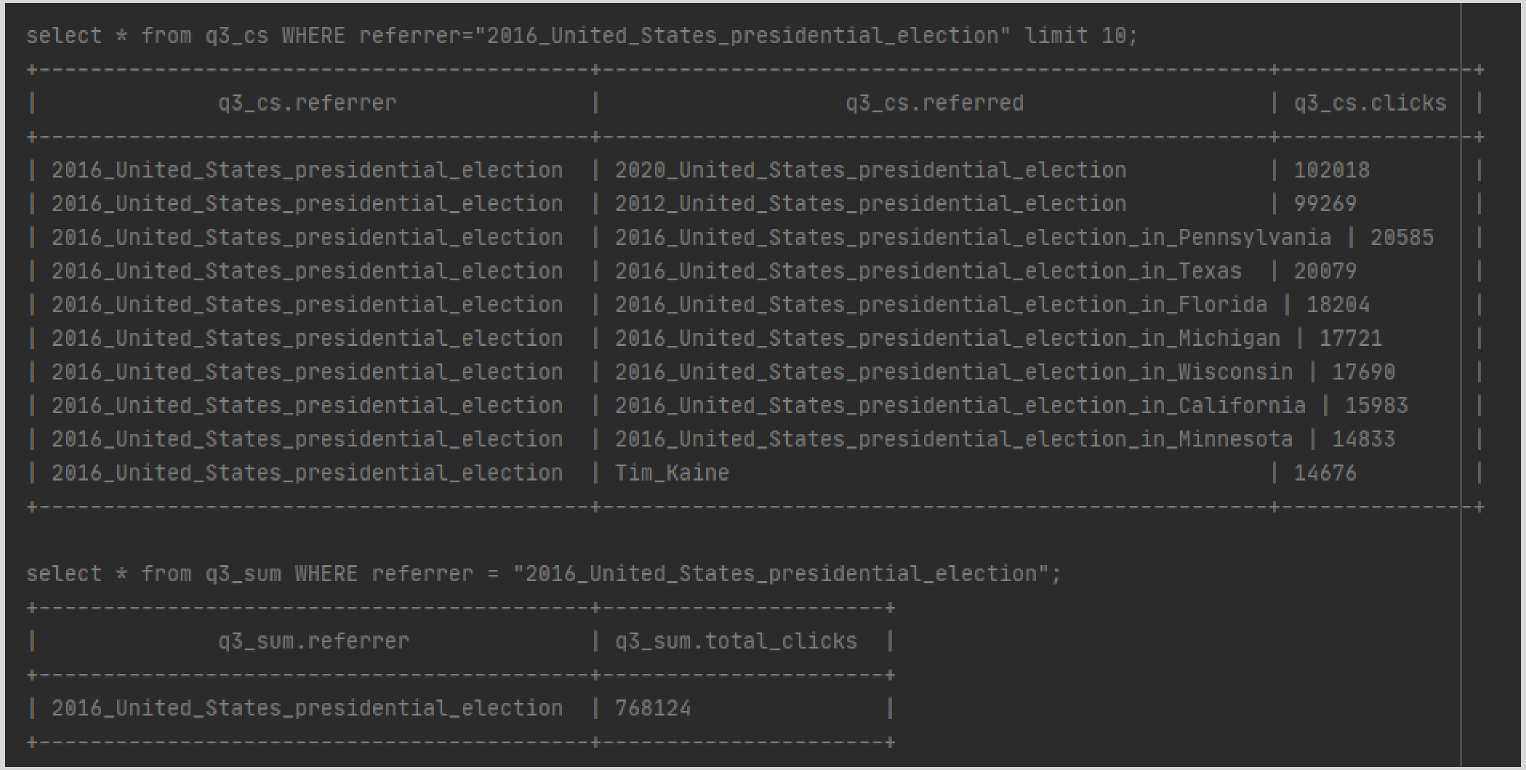
6. Run an analysis you find interesting on the wikipedia datasets we're using.



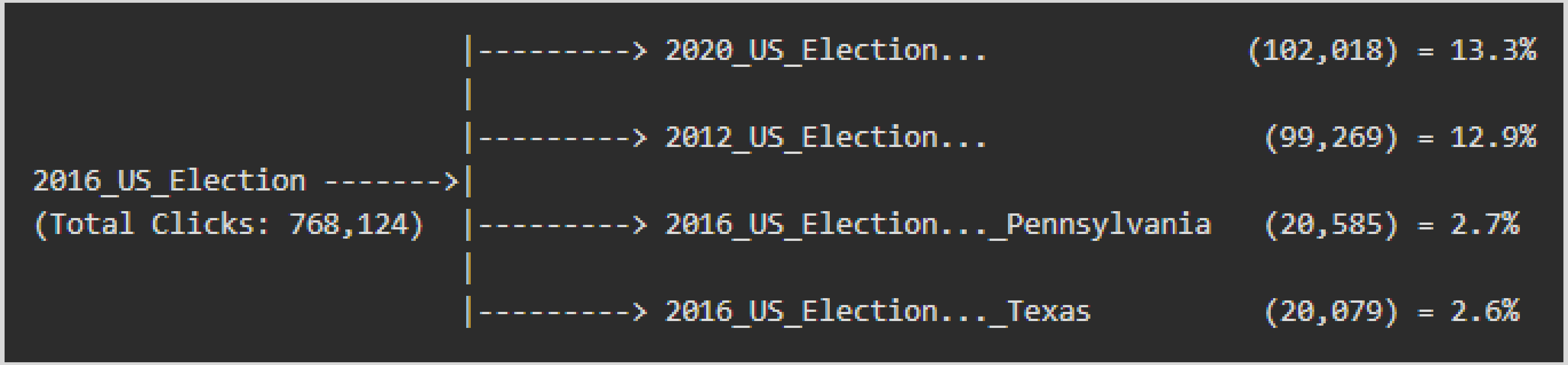
What wikipedia articles are the most viewed from the presidential election. Basically a trace.



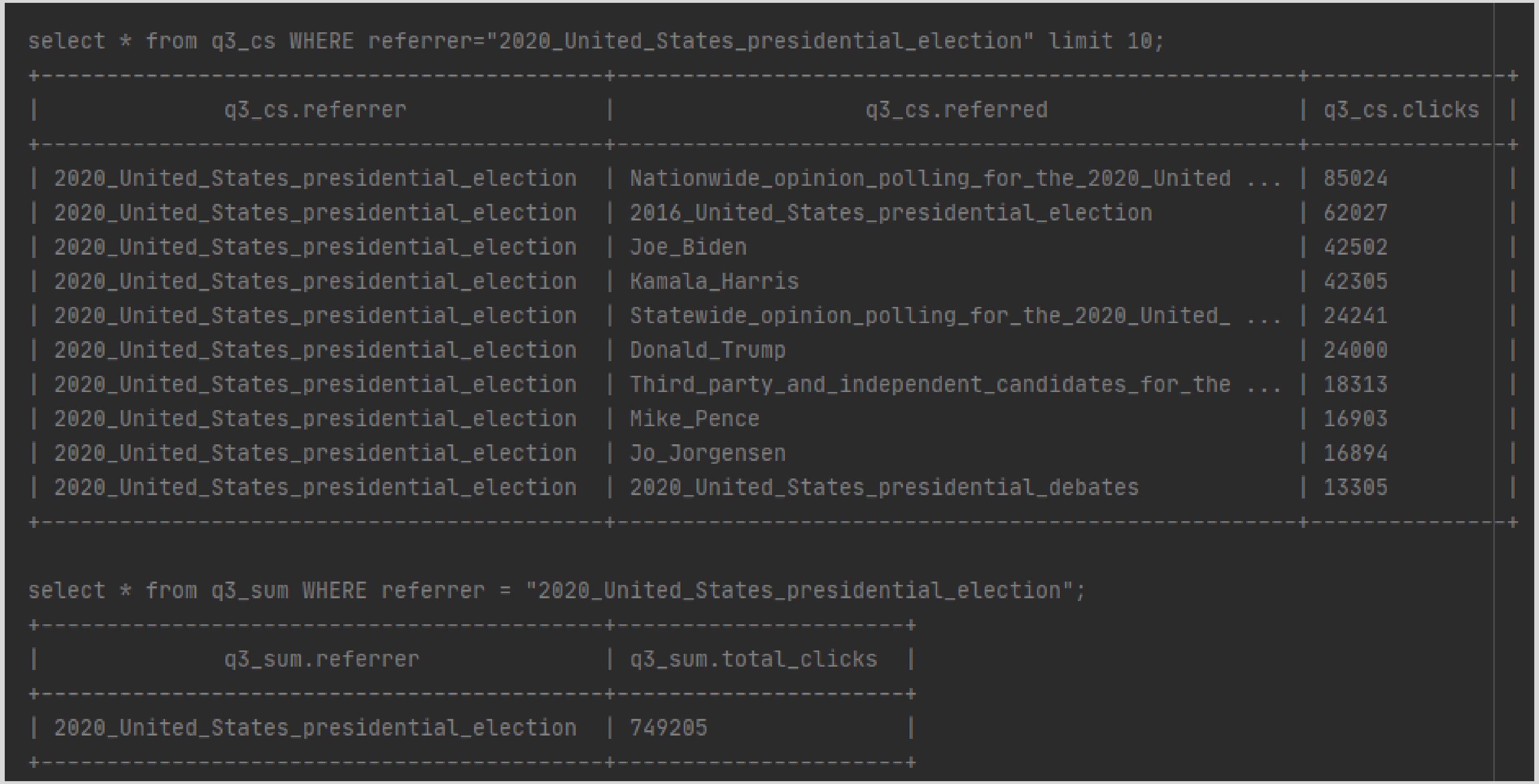
```
select * from q3_sum where referrer Like '%election' limit 10;
+---------------------------------------------+-----------------------+
|              q3_sum.referrer                | q3_sum.total_clicks   |
+---------------------------------------------+-----------------------+
| 2016_United_States_presidential_election    | 768124                |
| 2020_United_States_presidential_election    | 749205                |
| 2012_United_States_presidential_election    | 341194                |
| 2008_United_States_presidential_election    | 317237                |
| 2000_United_States_presidential_election    | 249939                |
| 2004_United_States_presidential_election    | 242779                |
| 1992_United_States_presidential_election    | 207533                |
| 1996_United_States_presidential_election    | 193678                |
| 1984_United_States_presidential_election    | 177773                |
| 1988_United_States_presidential_election    | 165147                |
+---------------------------------------------+-----------------------+
```



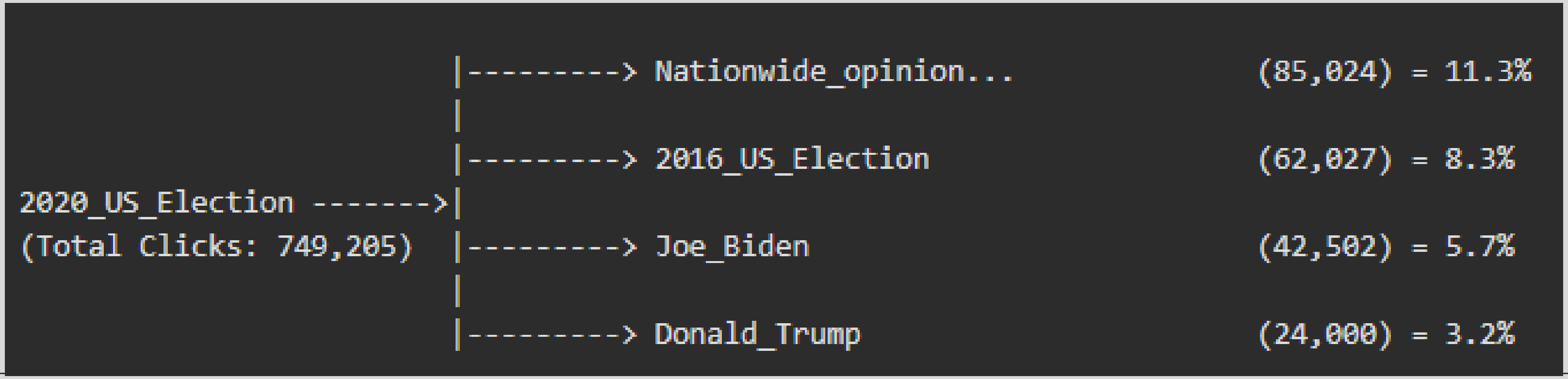
```
select * from q3_cs WHERE referrer="2016_United_States_presidential_election" limit 10;
+-----------------------------------------+----------------------------------------------------+--------------+
|             q3_cs.referrer              |                  q3_cs.referred                    | q3_cs.clicks |
+-----------------------------------------+----------------------------------------------------+--------------+
| 2016_United_States_presidential_election | 2020_United_States_presidential_election          | 102018       |
| 2016_United_States_presidential_election | 2012_United_States_presidential_election          | 99269        |
| 2016_United_States_presidential_election | 2016_United_States_presidential_election_in_Pennsylvania | 20585  |
| 2016_United_States_presidential_election | 2016_United_States_presidential_election_in_Texas  | 20079       |
| 2016_United_States_presidential_election | 2016_United_States_presidential_election_in_Florida | 18204      |
| 2016_United_States_presidential_election | 2016_United_States_presidential_election_in_Michigan | 17721     |
| 2016_United_States_presidential_election | 2016_United_States_presidential_election_in_Wisconsin | 17690    |
| 2016_United_States_presidential_election | 2016_United_States_presidential_election_in_California | 15983   |
| 2016_United_States_presidential_election | 2016_United_States_presidential_election_in_Minnesota | 14833    |
| 2016_United_States_presidential_election | Tim_Kaine                                          | 14676       |
+-----------------------------------------+----------------------------------------------------+--------------+

select * from q3_sum WHERE referrer = "2016_United_States_presidential_election";
+-----------------------------------------+---------------------+
|             q3_sum.referrer             | q3_sum.total_clicks |
+-----------------------------------------+---------------------+
| 2016_United_States_presidential_election | 768124              |
+-----------------------------------------+---------------------+
```



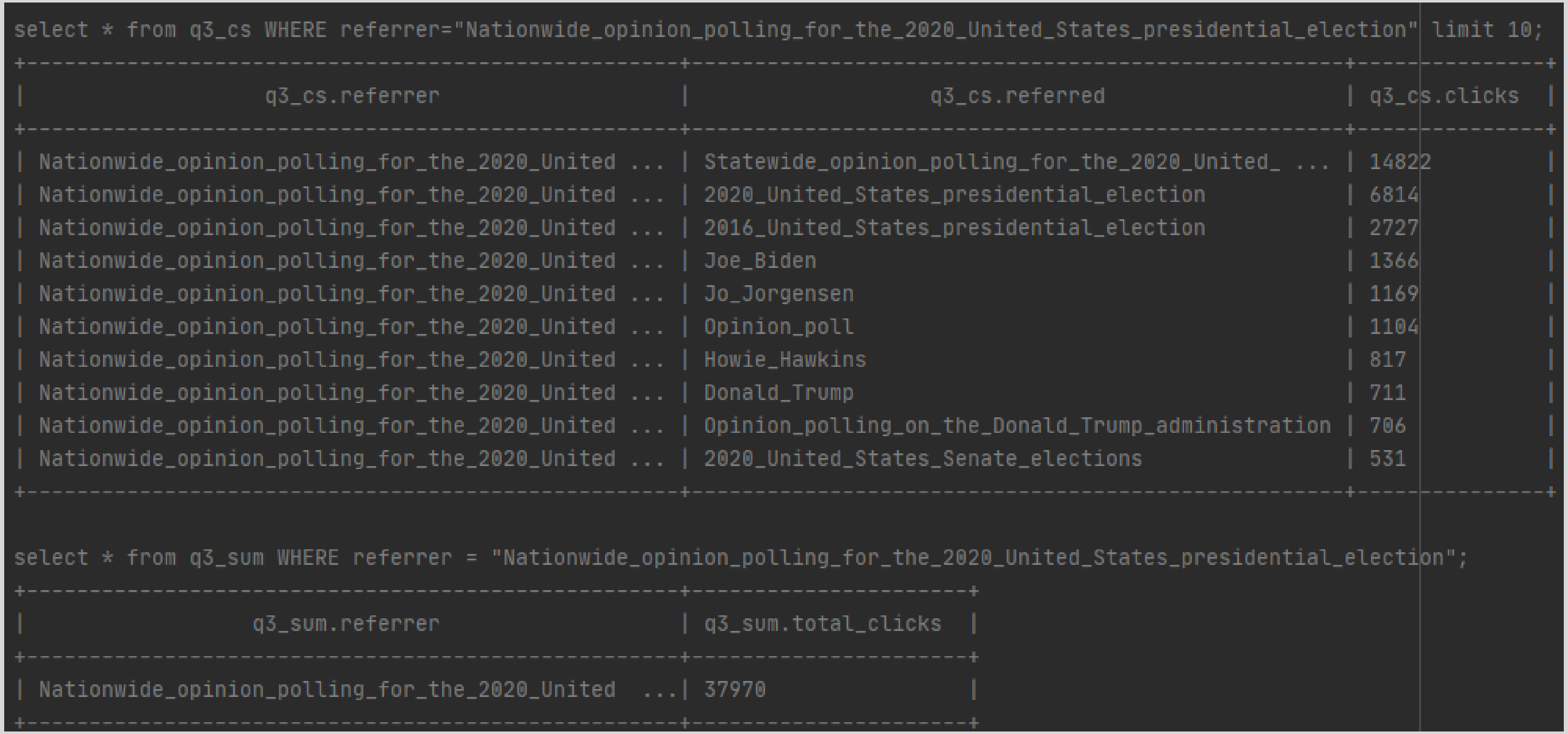
```
                            |---------> 2020_US_Election...                 (102,018) = 13.3%
                            |
                            |---------> 2012_US_Election...                  (99,269) = 12.9%
2016_US_Election ------->|
(Total Clicks: 768,124)     |---------> 2016_US_Election..._Pennsylvania    (20,585) = 2.7%
                            |
                            |---------> 2016_US_Election..._Texas           (20,079) = 2.6%
```



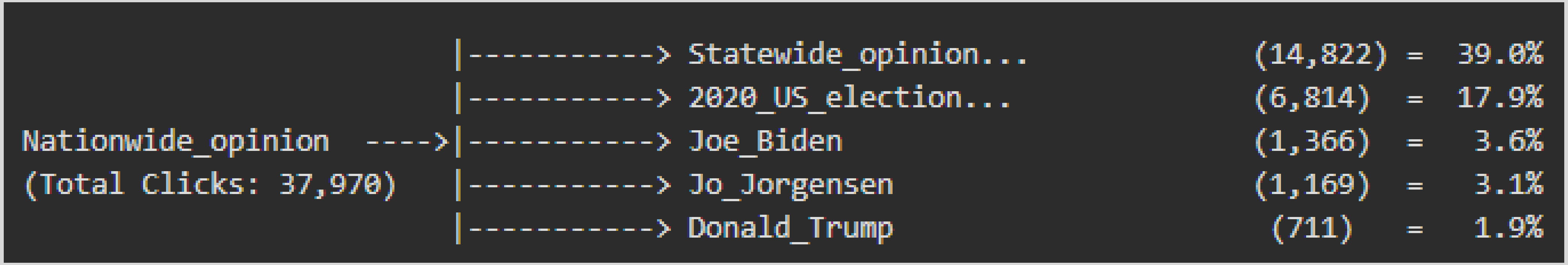
```
select * from q3_cs WHERE referrer="2020_United_States_presidential_election" limit 10;
+-------------------------------------------+----------------------------------------------------+---------------+
|              q3_cs.referrer               |                   q3_cs.referred                   | q3_cs.clicks  |
+-------------------------------------------+----------------------------------------------------+---------------+
| 2020_United_States_presidential_election  | Nationwide_opinion_polling_for_the_2020_United ... | 85024         |
| 2020_United_States_presidential_election  | 2016_United_States_presidential_election           | 62027         |
| 2020_United_States_presidential_election  | Joe_Biden                                          | 42502         |
| 2020_United_States_presidential_election  | Kamala_Harris                                      | 42305         |
| 2020_United_States_presidential_election  | Statewide_opinion_polling_for_the_2020_United_ ... | 24241         |
| 2020_United_States_presidential_election  | Donald_Trump                                       | 24000         |
| 2020_United_States_presidential_election  | Third_party_and_independent_candidates_for_the ... | 18313         |
| 2020_United_States_presidential_election  | Mike_Pence                                         | 16903         |
| 2020_United_States_presidential_election  | Jo_Jorgensen                                       | 16894         |
| 2020_United_States_presidential_election  | 2020_United_States_presidential_debates            | 13305         |
+-------------------------------------------+----------------------------------------------------+---------------+

select * from q3_sum WHERE referrer = "2020_United_States_presidential_election";
+-------------------------------------------+-----------------------+
|              q3_sum.referrer              | q3_sum.total_clicks  |
+-------------------------------------------+-----------------------+
| 2020_United_States_presidential_election  | 749205                |
+-------------------------------------------+-----------------------+
```



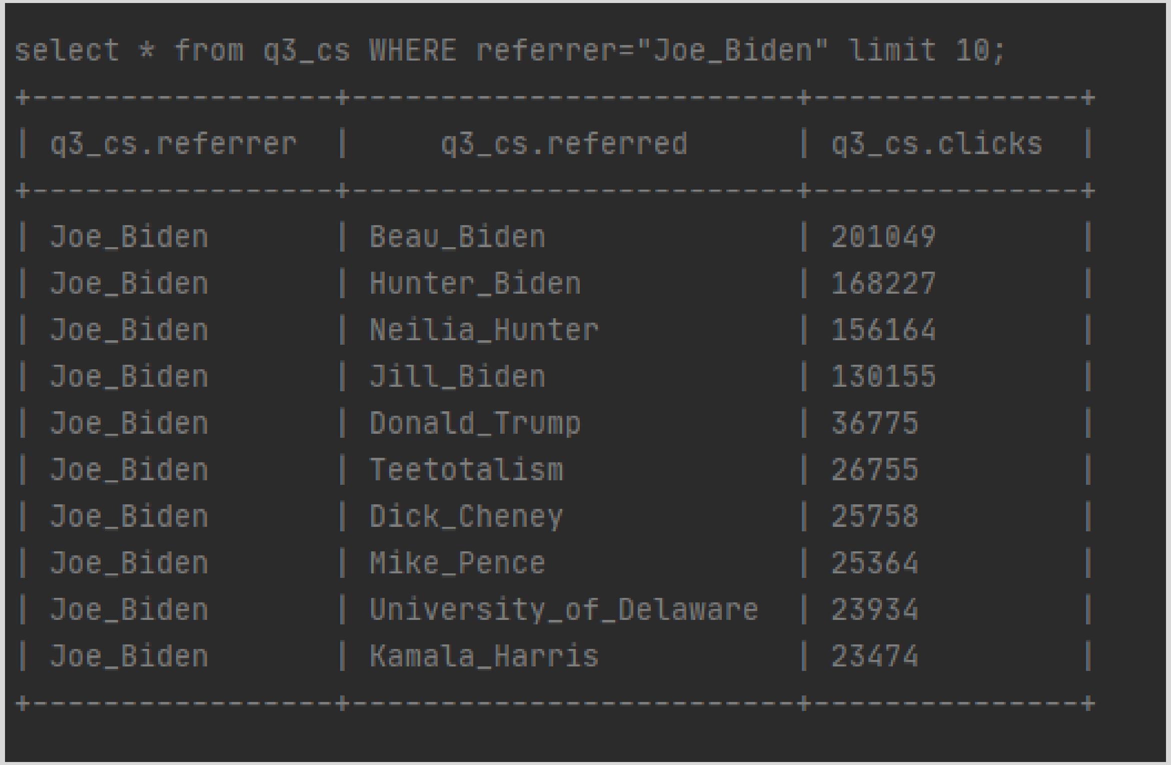
```
                              |---------> Nationwide_opinion...            (85,024) = 11.3%
                              |
                              |---------> 2016_US_Election                 (62,027) = 8.3%
2020_US_Election ------->|
(Total Clicks: 749,205)       |---------> Joe_Biden                        (42,502) = 5.7%
                              |
                              |---------> Donald_Trump                     (24,000) = 3.2%
```



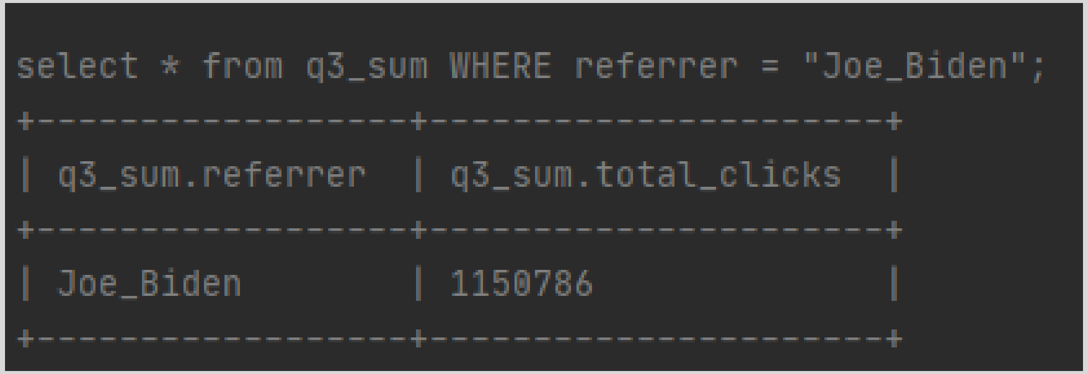
```
select * from q3_cs WHERE referrer="Nationwide_opinion_polling_for_the_2020_United_States_presidential_election" limit 10;
+-----------------------------------------------------+-----------------------------------------------------+---------------+
|                  q3_cs.referrer                     |                   q3_cs.referred                    | q3_cs.clicks  |
+-----------------------------------------------------+-----------------------------------------------------+---------------+
| Nationwide_opinion_polling_for_the_2020_United ...  | Statewide_opinion_polling_for_the_2020_United_ ...  | 14822         |
| Nationwide_opinion_polling_for_the_2020_United ...  | 2020_United_States_presidential_election            | 6814          |
| Nationwide_opinion_polling_for_the_2020_United ...  | 2016_United_States_presidential_election            | 2727          |
| Nationwide_opinion_polling_for_the_2020_United ...  | Joe_Biden                                           | 1366          |
| Nationwide_opinion_polling_for_the_2020_United ...  | Jo_Jorgensen                                        | 1169          |
| Nationwide_opinion_polling_for_the_2020_United ...  | Opinion_poll                                        | 1104          |
| Nationwide_opinion_polling_for_the_2020_United ...  | Howie_Hawkins                                       | 817           |
| Nationwide_opinion_polling_for_the_2020_United ...  | Donald_Trump                                        | 711           |
| Nationwide_opinion_polling_for_the_2020_United ...  | Opinion_polling_on_the_Donald_Trump_administration  | 706           |
| Nationwide_opinion_polling_for_the_2020_United ...  | 2020_United_States_Senate_elections                 | 531           |
+-----------------------------------------------------+-----------------------------------------------------+---------------+

select * from q3_sum WHERE referrer = "Nationwide_opinion_polling_for_the_2020_United_States_presidential_election";
+-----------------------------------------------------+-----------------------+
|                 q3_sum.referrer                     | q3_sum.total_clicks   |
+-----------------------------------------------------+-----------------------+
| Nationwide_opinion_polling_for_the_2020_United  ...| 37970                 |
+-----------------------------------------------------+-----------------------+
```



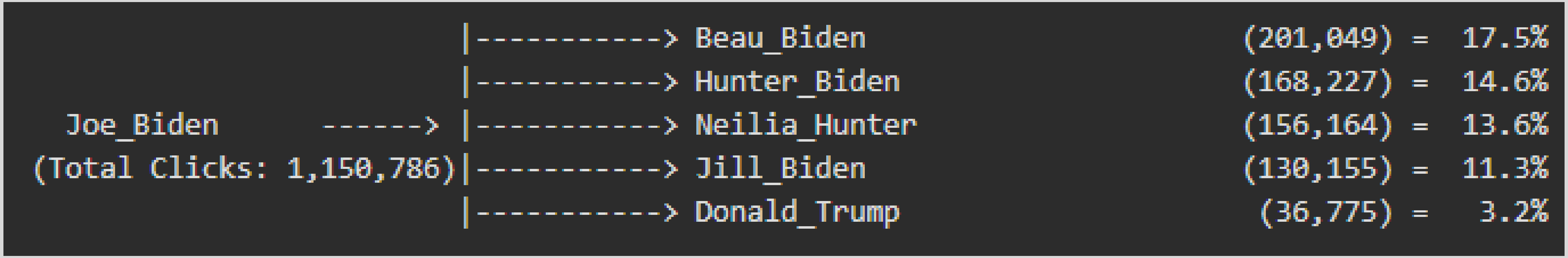
```
                             |-----------> Statewide_opinion...            (14,822) =  39.0%
                             |-----------> 2020_US_election...             (6,814)  =  17.9%
Nationwide_opinion    ---->  |-----------> Joe_Biden                       (1,366)  =   3.6%
(Total Clicks: 37,970)       |-----------> Jo_Jorgensen                    (1,169)  =   3.1%
                             |-----------> Donald_Trump                     (711)   =   1.9%
```



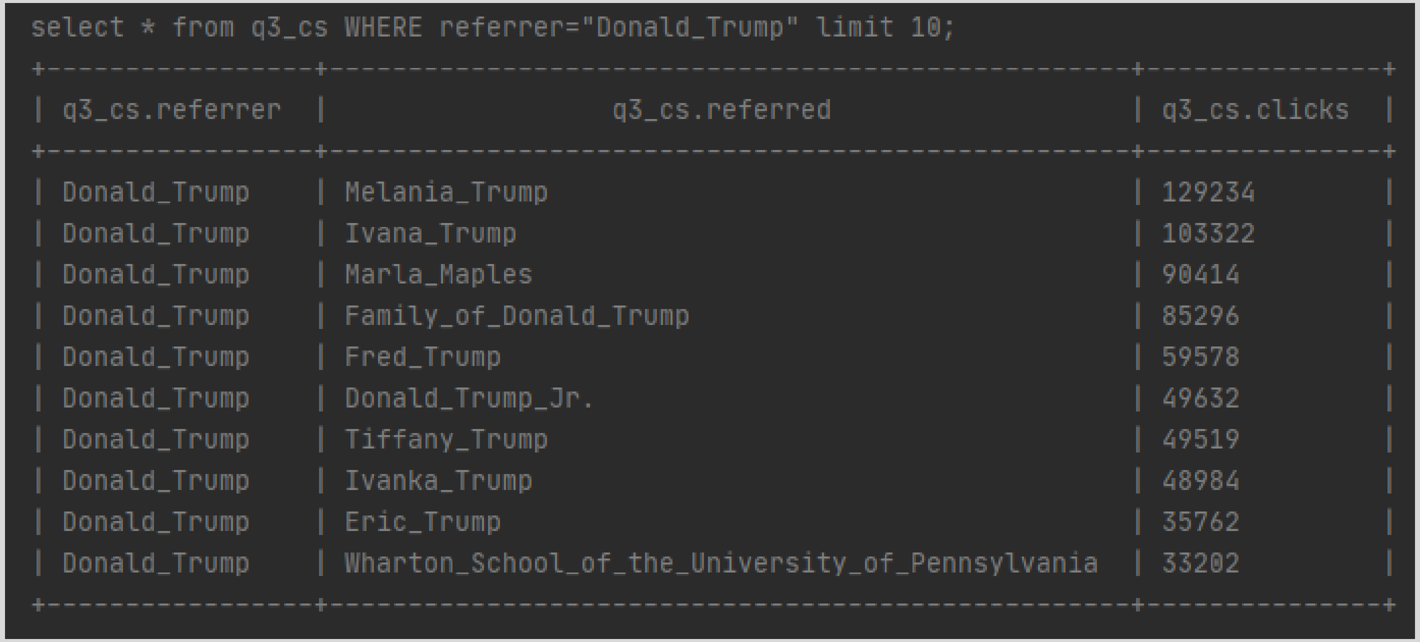
```
select * from q3_cs WHERE referrer="Joe_Biden" limit 10;
+-----------------+-------------------------+---------------+
| q3_cs.referrer  |     q3_cs.referred      | q3_cs.clicks  |
+-----------------+-------------------------+---------------+
| Joe_Biden       | Beau_Biden              | 201049        |
| Joe_Biden       | Hunter_Biden            | 168227        |
| Joe_Biden       | Neilia_Hunter           | 156164        |
| Joe_Biden       | Jill_Biden              | 130155        |
| Joe_Biden       | Donald_Trump            | 36775         |
| Joe_Biden       | Teetotalism             | 26755         |
| Joe_Biden       | Dick_Cheney             | 25758         |
| Joe_Biden       | Mike_Pence              | 25364         |
| Joe_Biden       | University_of_Delaware  | 23934         |
| Joe_Biden       | Kamala_Harris           | 23474         |
+-----------------+-------------------------+---------------+
```



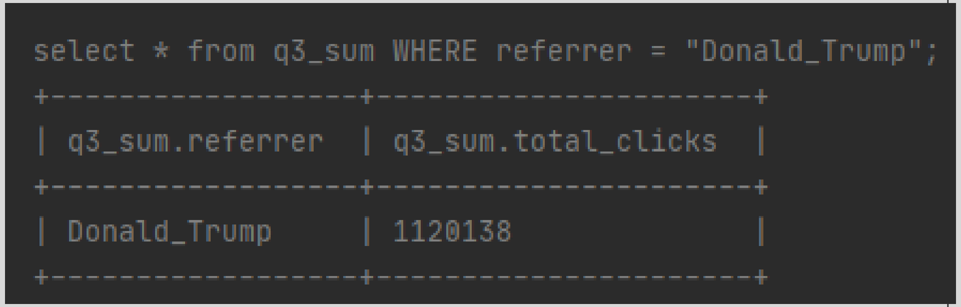
```
select * from q3_sum WHERE referrer = "Joe_Biden";
+-------------------+-----------------------+
| q3_sum.referrer   | q3_sum.total_clicks   |
+-------------------+-----------------------+
| Joe_Biden         | 1150786               |
+-------------------+-----------------------+
```



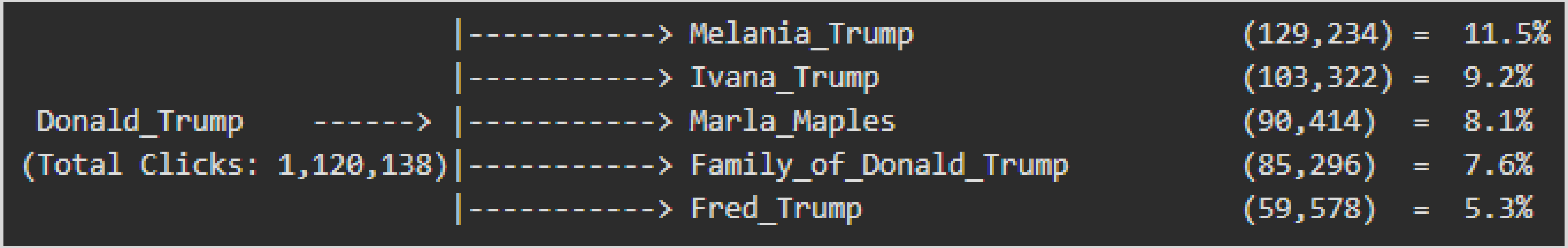
```
                               |-----------> Beau_Biden                   (201,049) =  17.5%
                               |-----------> Hunter_Biden                 (168,227) =  14.6%
   Joe_Biden            ------> |-----------> Neilia_Hunter                (156,164) =  13.6%
 (Total Clicks: 1,150,786)|-----------> Jill_Biden                   (130,155) =  11.3%
                               |-----------> Donald_Trump                  (36,775) =   3.2%
```



```
select * from q3_cs WHERE referrer="Donald_Trump" limit 10;
+---------------+------------------------------------------------+--------------+
| q3_cs.referrer |                q3_cs.referred                 | q3_cs.clicks |
+---------------+------------------------------------------------+--------------+
| Donald_Trump  | Melania_Trump                                  | 129234       |
| Donald_Trump  | Ivana_Trump                                    | 103322       |
| Donald_Trump  | Marla_Maples                                   | 90414        |
| Donald_Trump  | Family_of_Donald_Trump                         | 85296        |
| Donald_Trump  | Fred_Trump                                     | 59578        |
| Donald_Trump  | Donald_Trump_Jr.                               | 49632        |
| Donald_Trump  | Tiffany_Trump                                  | 49519        |
| Donald_Trump  | Ivanka_Trump                                   | 48984        |
| Donald_Trump  | Eric_Trump                                     | 35762        |
| Donald_Trump  | Wharton_School_of_the_University_of_Pennsylvania | 33202      |
+---------------+------------------------------------------------+--------------+
```



```
select * from q3_sum WHERE referrer = "Donald_Trump";
+-----------------+---------------------+
| q3_sum.referrer | q3_sum.total_clicks |
+-----------------+---------------------+
| Donald_Trump    | 1120138             |
+-----------------+---------------------+
```



```
                            |-----------> Melania_Trump             (129,234) =  11.5%
                            |-----------> Ivana_Trump               (103,322) =  9.2%
 Donald_Trump     ------->  |-----------> Marla_Maples              (90,414)  =  8.1%
(Total Clicks: 1,120,138)   |-----------> Family_of_Donald_Trump    (85,296)  =  7.6%
                            |-----------> Fred_Trump                (59,578)  =  5.3%
```

Questions?

GitHub link

https://github.com/victorpullas/Project1