Project 1:

•••

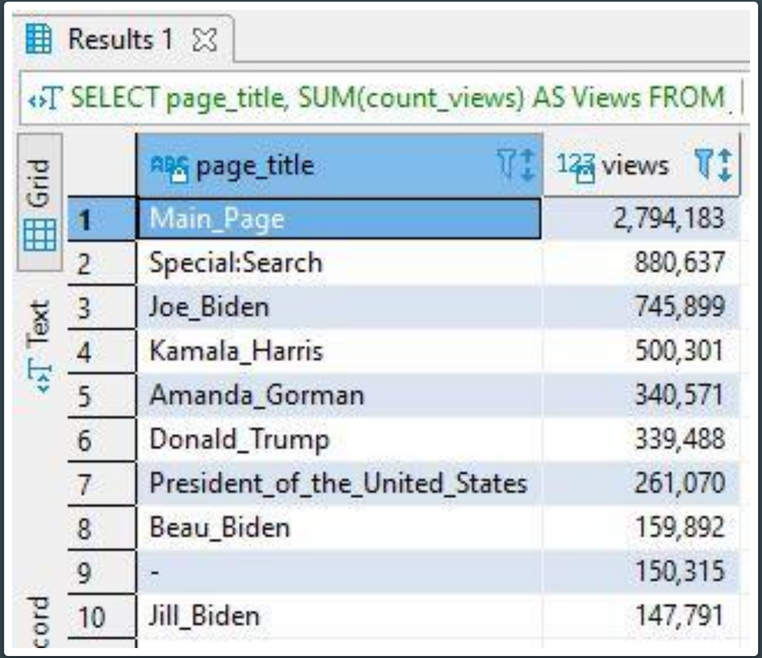
Delaney Lekien

Questions 1:

Which English wikipedia article got the most traffic on January 20, 2021?

What we needed to solve this:

- All pageviews data from January 20th, 2021.
- Filtering the domain code to English articles only.
- Query for the total Sum of the all the views for each individual page.
  - Grouping by Page Title in order to avoid duplicates.
- Lastly, summing all the page views associated with each page title.



```
SELECT page_title, SUM(count_views) AS Views FROM
```

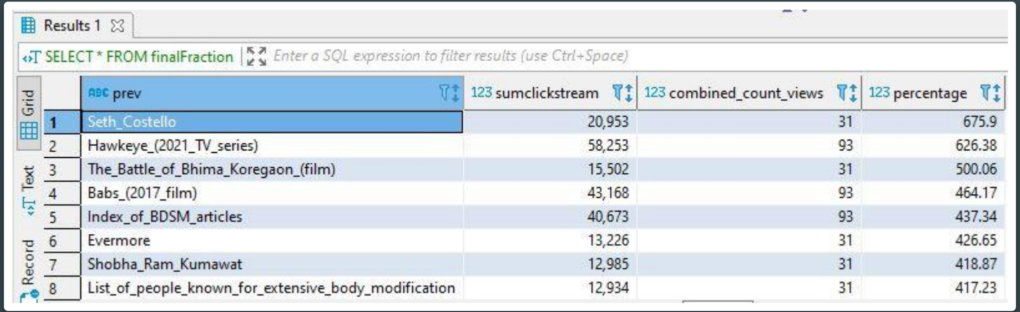
	page_title	views
1	Main_Page	2,794,183
2	Special:Search	880,637
3	Joe_Biden	745,899
4	Kamala_Harris	500,301
5	Amanda_Gorman	340,571
6	Donald_Trump	339,488
7	President_of_the_United_States	261,070
8	Beau_Biden	159,892
9	-	150,315
10	Jill_Biden	147,791

Questions 2:

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

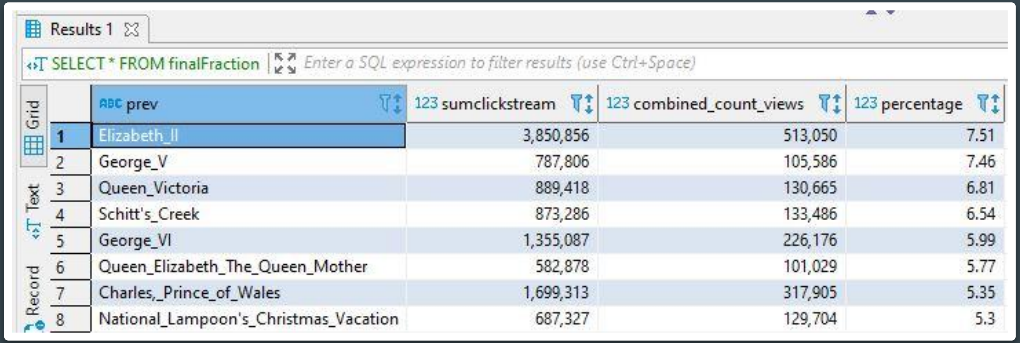
What we needed to solve this:

- December clickstream data filtered to only show internal links.
- December page views data
  - First assumption, since all December page views data cannot be downloaded.
- Percentage of internal clicks for a page over the amount of traffic that page got.



```
SELECT * FROM finalFraction
```

	prev	sumclickstream	combined_count_views	percentage
1	Seth_Costello	20,953	31	675.9
2	Hawkeye_(2021_TV_series)	58,253	93	626.38
3	The_Battle_of_Bhima_Koregaon_(film)	15,502	31	500.06
4	Babs_(2017_film)	43,168	93	464.17
5	Index_of_BDSM_articles	40,673	93	437.34
6	Evermore	13,226	31	426.65
7	Shobha_Ram_Kumawat	12,985	31	418.87
8	List_of_people_known_for_extensive_body_modification	12,934	31	417.23



```
SELECT * FROM finalFraction
```

	prev	sumclickstream	combined_count_views	percentage
1	Elizabeth_II	3,850,856	513,050	7.51
2	George_V	787,806	105,586	7.46
3	Queen_Victoria	889,418	130,665	6.81
4	Schitt's_Creek	873,286	133,486	6.54
5	George_VI	1,355,087	226,176	5.99
6	Queen_Elizabeth_The_Queen_Mother	582,878	101,029	5.77
7	Charles,_Prince_of_Wales	1,699,313	317,905	5.35
8	National_Lampoon's_Christmas_Vacation	687,327	129,704	5.3

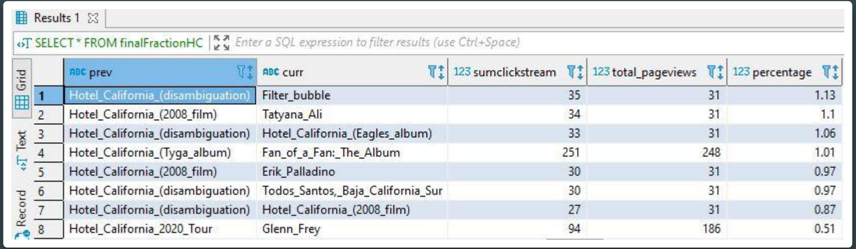
Questions 3:

What series of wikipedia articles, starting with Hotel California keeps the largest fraction of its readers clicking on internal links?

What we needed to solve this:

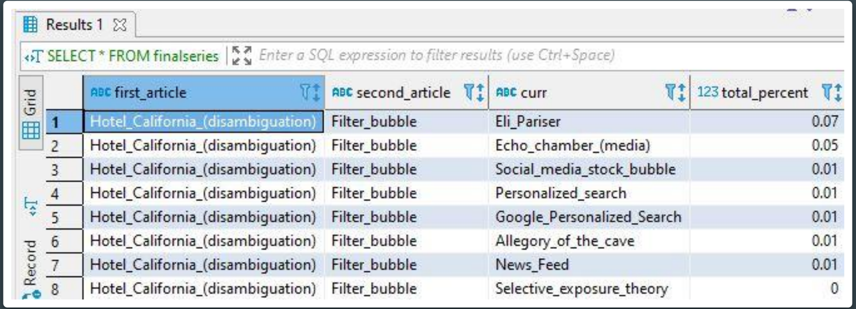
- Using the same clickstream data as question two, and filtering it to start with Hotel California.
- December page views assumption data.
- Percentage of internal links starting with Hotel California
  - Then following that top chain to it's next highest top chain.

Hotel California (disambiguation) -> Filter_bubble ("Hotel California effect, related to media filter bubbles") -> Eli Pariser



```
SELECT * FROM finalFractionHC
```

	prev	curr	sumclickstream	total_pageviews	percentage
1	Hotel_California_(disambiguation)	Filter_bubble	35	31	1.13
2	Hotel_California_(2008_film)	Tatyana_Ali	34	31	1.1
3	Hotel_California_(disambiguation)	Hotel_California_(Eagles_album)	33	31	1.06
4	Hotel_California_(Tyga_album)	Fan_of_a_Fan:_The_Album	251	248	1.01
5	Hotel_California_(2008_film)	Erik_Palladino	30	31	0.97
6	Hotel_California_(disambiguation)	Todos_Santos,_Baja_California_Sur	30	31	0.97
7	Hotel_California_(disambiguation)	Hotel_California_(2008_film)	27	31	0.87
8	Hotel_California_2020_Tour	Glenn_Frey	94	186	0.51



```
SELECT * FROM finalseries
```

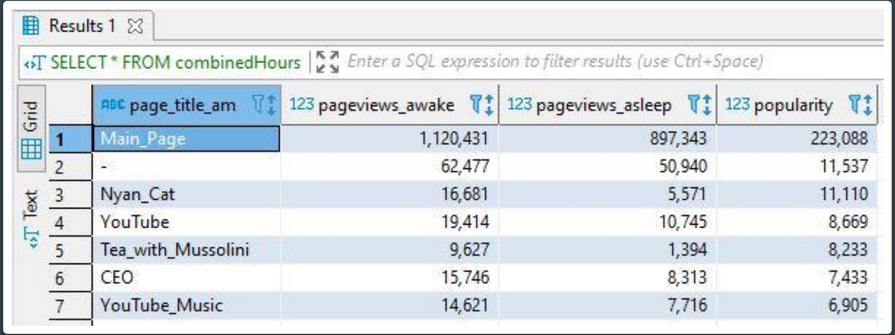
	first_article	second_article	curr	total_percent
1	Hotel_California_(disambiguation)	Filter_bubble	Eli_Pariser	0.07
2	Hotel_California_(disambiguation)	Filter_bubble	Echo_chamber_(media)	0.05
3	Hotel_California_(disambiguation)	Filter_bubble	Social_media_stock_bubble	0.01
4	Hotel_California_(disambiguation)	Filter_bubble	Personalized_search	0.01
5	Hotel_California_(disambiguation)	Filter_bubble	Google_Personalized_Search	0.01
6	Hotel_California_(disambiguation)	Filter_bubble	Allegory_of_the_cave	0.01
7	Hotel_California_(disambiguation)	Filter_bubble	News_Feed	0.01
8	Hotel_California_(disambiguation)	Filter_bubble	Selective_exposure_theory	0

Questions 4:

Find an example of an English wikipedia article that is relatively more popular in the Americas than elsewhere.

What we needed to solve this:

- Data from page views during popular hours in the Americas (7pm - 11pm EST or 14:00 - 18:00 UTC)
- Data from page views during resting hours in the Americas (1am-5am EST or 20:00 - 1:00 UTC)
- The difference between pageviews of awake hours versus pageviews of hours asleep.



```
SELECT * FROM combinedHours
```

	page_title_am	pageviews_awake	pageviews_asleep	popularity
1	Main_Page	1,120,431	897,343	223,088
2	-	62,477	50,940	11,537
3	Nyan_Cat	16,681	5,571	11,110
4	YouTube	19,414	10,745	8,669
5	Tea_with_Mussolini	9,627	1,394	8,233
6	CEO	15,746	8,313	7,433
7	YouTube_Music	14,621	7,716	6,905

Questions 5:

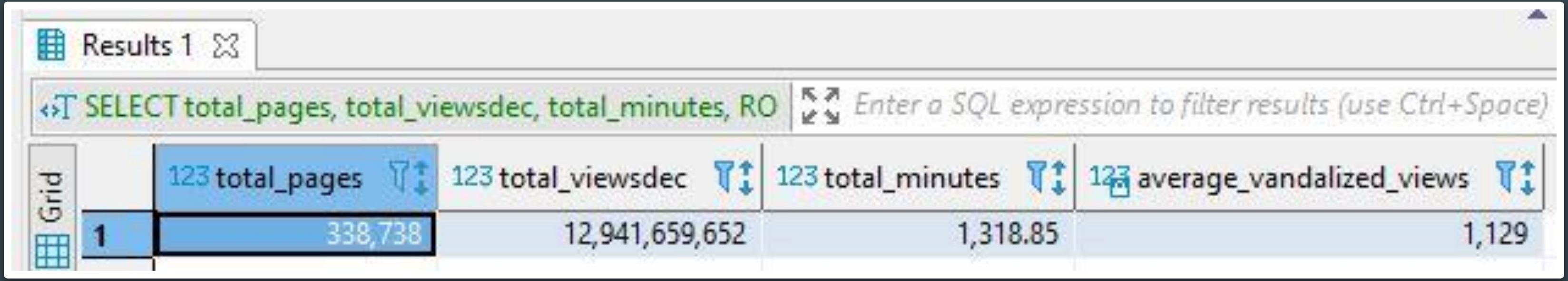
Analyze how many users will see the average vandalized wikipedia page before the offending edit is reversed.

What we needed to solve this:

- Average minutes a vandalized webpage stays up in the month of December.
- Pageviews in December
- Total pages visited in December

The data I used to answer this question:

- Event_entity = revision
- Event_type = create
- Event_timestamp
- Revision_is_identity_reverted = true
- Revision_seconds_to_identify_revert



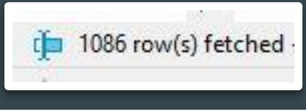
```
SELECT total_pages, total_viewsdec, total_minutes, RO
```

	total_pages	total_viewsdec	total_minutes	average_vandalized_views
1	338,738	12,941,659,652	1,318.85	1,129

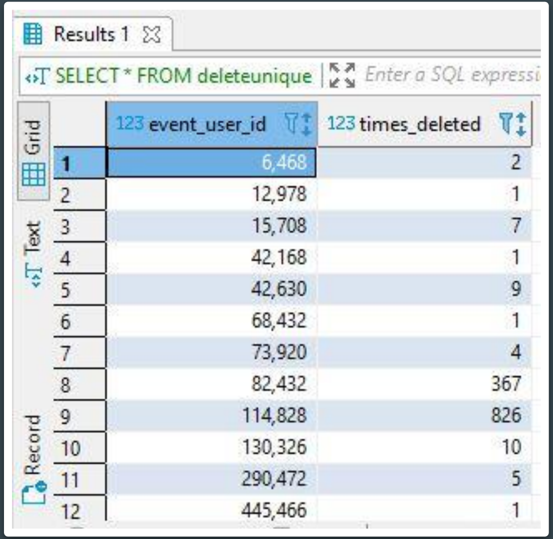
Questions 6:

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

I wanted to know how many unique users deleted a page on Wikipedia, and how many times did they do?

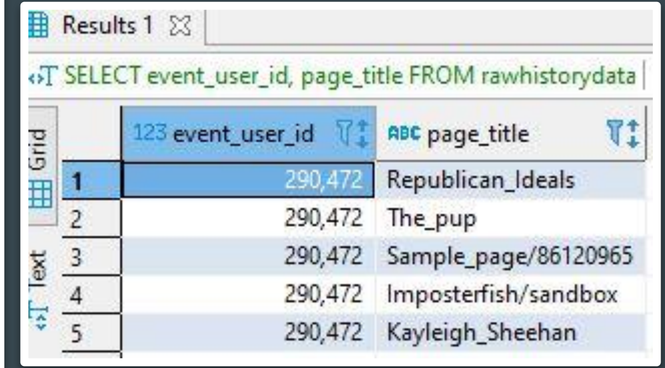


1086 row(s) fetched



SELECT * FROM deleteunique

	event_user_id	times_deleted
1	6,468	2
2	12,978	1
3	15,708	7
4	42,168	1
5	42,630	9
6	68,432	1
7	73,920	4
8	82,432	367
9	114,828	826
10	130,326	10
11	290,472	5
12	445,466	1



SELECT event_user_id, page_title FROM rawhistorydata

	event_user_id	page_title
1	290,472	Republican_Ideals
2	290,472	The_pup
3	290,472	Sample_page/86120965
4	290,472	Imposterfish/sandbox
5	290,472	Kayleigh_Sheehan

Git Repo:

https://github.com/delaney-lekien/Project1_DL