

Supporting Documentation: Task 2 and 3

Student: Úna McGinn

CA675 Assignment 1

Following trial and error in an attempt to obtain the top close to 50,000 records, the following query was tried.

```
select top 50000 * from posts where posts.ViewCount > 129462
ORDER BY posts.ViewCount
```

The screenshot shows the Stack Overflow query editor interface. The query entered is: `select top 50000 * from posts where posts.ViewCount > 129462 ORDER BY posts.ViewCount`. The results table shows columns: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, Body, OwnerUserId, OwnerDisplayName, LastEditorUserId, LastEditorDisplayName, and LastEditDate. The first row has Id 166914, ViewCount 129464, and Body starting with 'I have one Android Device running Jelly...'. The table indicates 49029 rows returned in 30312 ms.

Id	PostTypeId	AcceptedAnswerId	ParentId	CreationDate	DeletionDate	Score	ViewCount	Body	OwnerUserId	OwnerDisplayName	LastEditorUserId	LastEditorDisplayName	LastEditDate
166914...	1	16698105		2013-05-22 12:15:05		57	129464	<p>I have one Android Device running Jelly ...	560932		560932		2013-...
1677121	1	1677198		2009-11-04 22:30:03		130	129466	<p>I am a new user to <code>git</code> and...	185119		1067326		2017-...
565026	1			2019-06-12 12:48:35		8	129468	<p>Windows 10 is displaying a <code>vbscript ...	7086300				
3184883	1	3184937		2010-07-06 09:05:26		87	129468	<p>I have the following piece of code <code>ip ...	328681		503413		2014-...
3786022	1	3786023		2010-09-21 12:24:56		83	129468	<p>How To Include CSS and jQuery in my W...	423903		771496		2019-...
271820...	1	27184261		2014-11-28 04:43:09		238	129468	<p>I am trying to get the difference between t...	4263409		2303865		2020-...
338427...	1			2015-11-21 11:24:16		29	129471	<p>I want to delete a News from database an...	user4578154		1252759		2015-...
5181096	1	5181180		2011-03-01 22:18:20		72	129471	<p>I change the position of a UIView with full ...	174622		5176709		2020-...
9952000	1	9952100		2012-03-31 00:40:39		79	129472	<p>I'm having problems getting my <code>cv ...	1304360		490018		2020-...
9723912	1	9724069		2012-03-15 16:25:24		49	129472	<p>Here is some Java code to reverse a strin...	1241677		445131		2014-...
7364	1	7455		2008-08-10 21:58:24		84	129475	<p>Does anyone know of a good method for ...	277		1420625		2021-...
693788	1	693794		2009-03-28 23:44:29		237	129478	<p>What is better: <code>void foo()</code> ...	83871	Zlre	895245	Zlre	2016-...
5157809	1	5157928		2011-03-01 17:16:18		12	129482	<p>Is there any way to store mycol result in p...	589744		227884		2011-...
770474	1	770509		2009-04-26 23:36:53		606	129483	<p>I know that in terms of several distributed ...	86751		-1		2020-...
158190...	1	15822811		2013-04-04 18:37:25		83	129483	<p>I have a list of 4 pandas dataframes cont...	2246074		1007939		2017-...
174809...	1	17482796		2013-07-05 04:16:14		140	129484	<p>I am developing an application using Ql. I...	2243767		872616		2017-...

From the above, I can see that the lowest ViewCount in the dataset is 129464. Therefore, for the 1st dataset I will select the records with Viewcount > 129464 to remove the need to account for the possibility that there may be multiple videos with 129464 views.

Dataset Part 1:

```
select top 50000 * from posts where posts.ViewCount > 129464
ORDER BY posts.ViewCount
```

Result: 49028 rows

Dataset Part 2:

```
select top 50000 * from posts where posts.ViewCount <= 129464 and posts.ViewCount > 77000
ORDER BY posts.ViewCount
```

Result: 47385 rows

From the above, I can see that the lowest ViewCount in the dataset is 77001. There are less than 50,000 records in the dataset so all values in this range are in the dataset.

Dataset Part 3:

```
select top 50000 * from posts where posts.ViewCount <= 77000 and posts.ViewCount > 55500
ORDER BY posts.ViewCount
```

Result: 46963 rows

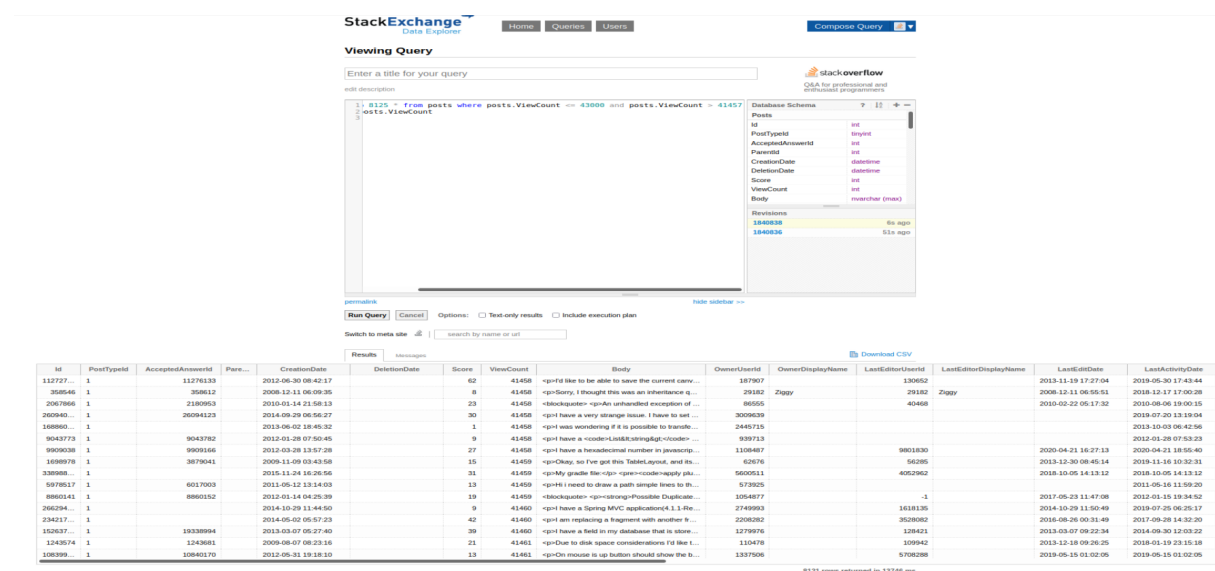
Dataset Part 4:

select top 50000 * from posts where posts.ViewCount <= 55500 and posts.ViewCount > 43000
ORDER BY posts.ViewCount

Result: 48499 rows

Dataset Part 5:

Tried the following: select top 8125 * from posts where posts.ViewCount <= 43000 and
posts.ViewCount > 35000
ORDER BY posts.ViewCount



The screenshot shows the StackExchange Data Explorer interface. A query is entered in the top text area: `select top 8125 * from posts where posts.ViewCount <= 43000 and posts.ViewCount > 41457 ORDER BY posts.ViewCount`. The results are displayed in a table with columns: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, Body, OwnerUserId, OwnerDisplayName, LastEditorUserId, LastEditorDisplayName, LastEditDate, and LastActivityDate. The table shows 8125 rows of data, with the first few rows having ViewCounts ranging from 41458 to 41460. The interface also includes a 'Database Schema' panel on the right showing the structure of the 'Posts' table.

Id	PostTypeId	AcceptedAnswerId	ParentId	CreationDate	DeletionDate	Score	ViewCount	Body	OwnerUserId	OwnerDisplayName	LastEditorUserId	LastEditorDisplayName	LastEditDate	LastActivityDate
112777	1	11276133		2012-06-30 08:42:17		62	41458	<p>I'd like to be able to save the current cam...	187987	Ziggy	130652		2013-11-19 17:27:04	2019-06-30 17:43:44
358648	1	358612		2008-12-11 06:09:35		8	41458	<p>Sorry, I thought this was an inheritance q...	29182	Ziggy	29182	Ziggy	2008-12-11 06:55:51	2018-12-17 17:00:28
2067866	1	2180953		2010-01-14 21:58:13		23	41458	<blockquote><p>An unhandled exception of ...	86555		40468		2010-02-22 05:17:32	2010-08-06 19:00:15
200480	1	26094123		2014-09-26 06:56:27		30	41458	<p>I have a very strange issue. I have to set ...	300638				2018-07-26 13:19:04	2019-07-26 13:19:04
168960	1			2013-06-02 15:46:32		1	41458	<p>I was wondering if it is possible to resu...	2446715				2013-10-03 06:42:56	
9043773	1	9043782		2012-01-28 07:50:45		9	41458	<p>I have a <code>LinkedList<string>glt</code>...	939713				2012-01-28 07:53:23	
9909038	1	9909166		2012-03-28 13:57:28		27	41458	<p>I have a hexadecimal number in my script...	1109487		9901830		2020-04-21 16:27:13	2020-04-21 18:55:40
1608978	1	3879041		2009-11-09 03:43:58		15	41458	<p>Okay, so I've got this <code>Table</code> and it...	62676		56295		2013-12-30 08:40:14	2019-11-16 10:32:31
338988	1			2015-11-24 16:26:56		31	41459	<p>My gradle file <code>-cp <code>=</code>apply pl...	5600511		4052962		2018-10-05 14:13:12	2018-10-05 14:13:12
5978517	1	6017003		2011-05-12 13:14:03		13	41459	<p>Hi I need to draw a path simple lines to th...	573825				2011-05-16 11:59:20	
8860141	1	8860152		2012-01-14 04:25:39		19	41459	<blockquote><p>strong>Possible Duplicate...	1054577		-1		2017-05-23 11:47:08	2012-01-15 19:34:52
262294	1			2014-10-29 11:44:50		9	41460	<p>I have a Spring MVC application 5.1.9Re...	2749983		1418136		2014-10-29 11:50:49	2019-07-25 06:25:17
234217	1			2014-05-02 05:57:23		42	41460	<p>I am replacing a fragment with another fr...	2208282		3528082		2016-08-26 00:31:49	2017-08-28 14:32:20
152637	1	19338994		2013-03-07 05:27:40		39	41460	<p>I have a field in my database that is store...	1279976		128421		2013-03-07 09:22:34	2014-09-30 12:03:22
1243574	1	1243601		2009-08-07 06:23:16		21	41461	<p>Due to disk space considerations I'd like t...	115478		109942		2013-12-18 09:26:25	2018-01-10 23:15:18
198399	1	10840170		2012-05-31 19:18:10		13	41461	<p>On mouse to up button should show the b...	1337506		5706288		2019-05-15 01:02:05	2019-05-15 01:02:05

By trial and error, used the following interval to obtain the next 8125 records.

select top 8125 * from posts where posts.ViewCount <= 43000 and posts.ViewCount > 41457
ORDER BY posts.ViewCount

Result: 8125 rows

For the final subset I take the top 8125 records i.e.(the number of records required to make up 200,000 records). These 5 batches have 49028, 47385, 46963, 48499 and 8125 records respectively, making up the total dataset of 200,000.