

Supporting Documentation: Task 1

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 Data Explorer 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>Vista...</code> | 7086300 | | | | |
| 3184883 | 1 | 3184937 | | 2010-07-06 09:05:26 | | 87 | 129468 | <p>I have the following piece of code <code>ip...</code> | 328681 | | 503413 | | 2014-... |
| 3786022 | 1 | 3786073 | | 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>cmr...</code> | 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> or ... | 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... | 583744 | | 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 Qt. 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. At the top, there's a navigation bar with 'Home', 'Queries', and 'Users'. Below it, a 'Viewing Query' section contains a text input for a query title and a query editor. The query entered is: `select top 8125 * from posts where posts.ViewCount <= 43000 and posts.ViewCount > 41457 ORDER BY posts.ViewCount`. To the right of the query editor is a 'Database Schema' panel showing the structure of the 'Posts' table, including columns like 'Id', 'PostTypeId', 'AcceptedAnswerId', 'ParentId', 'CreationDate', 'OwnerUserId', 'OwnerDisplayName', 'LastEditorUserId', 'LastEditorDisplayName', 'LastEditDate', and 'LastActivityDate'. Below the query editor are buttons for 'Run Query', 'Cancel', and 'Options'. At the bottom, there's a 'Results' section displaying a table of query results. The table has columns corresponding to the 'Posts' table schema. The first few rows of the results are visible, showing various post details. A status bar at the bottom right indicates '8125 rows returned in 13746 ms'.

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.