Retrieve User Activity Data on an Online Forum Using SQL

Data Analyst
Lenara Sitshayeva

ChatData

A popular network of question-and-answer (Q&A) websites in the fields of data analytics, data science and artificial intelligence. They help budding data analysts stay up-to-date with current innovations, find answers to burning questions, and stay active in the data community

Project Goal: to learn how ChatData sites are being used in the real world to understand which features are useful to the users and what additional features might be worth introducing.

Datasets: three separate CSV files attached: *posts*, *comments* and *users*.

Task 1: Create the ERD and Database and Load the Data

Entities:

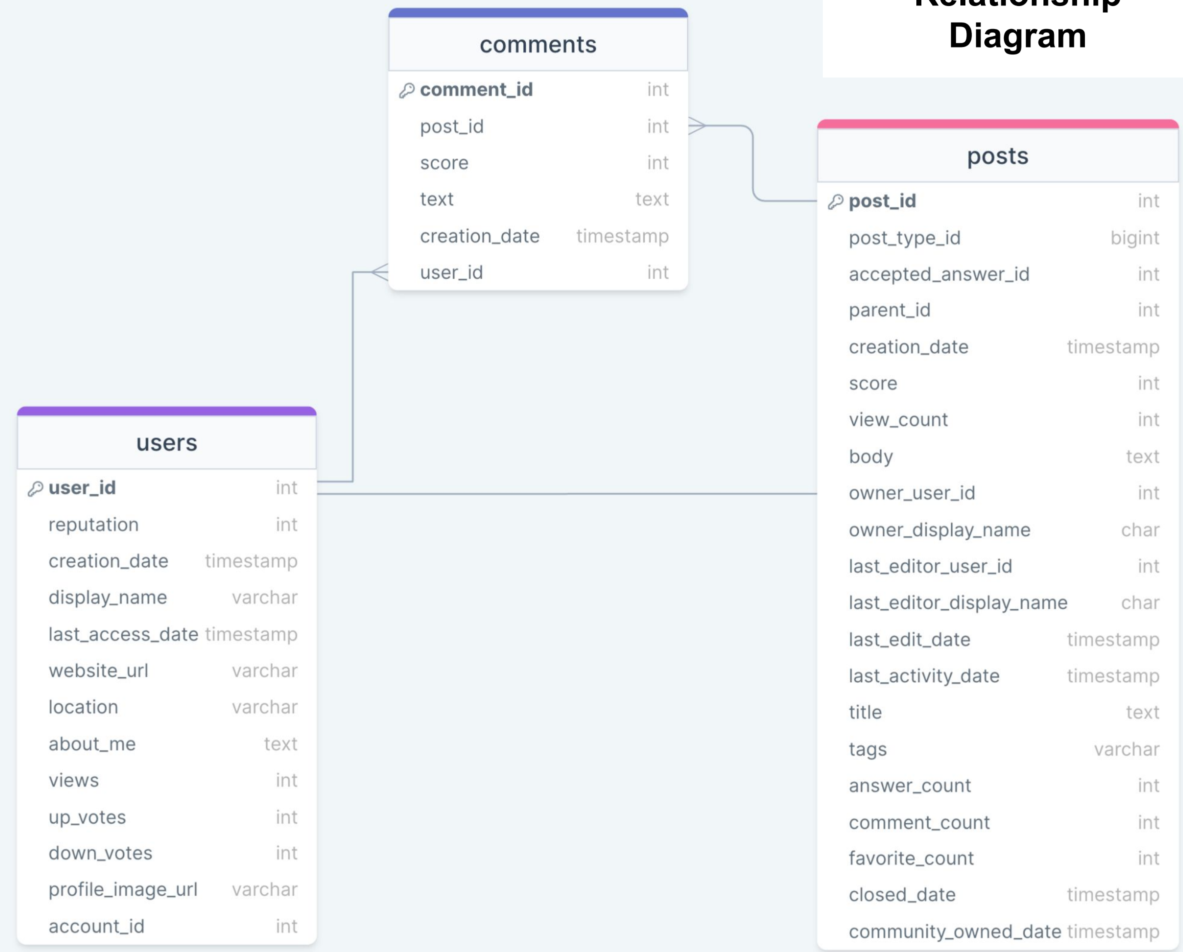
- Posts
- Comments
- Users

Relationship:
Entities **Users** and **Comments** have one to many relationship:
One user can write many comments.
One comment can be written by user.

Relationship:
Entities **Posts** and **Comments** have one to many relationship:
One post can have many comments.
One comment can be written for only one post.

Relationship:
Entities **Users** and **Posts** have one to one relationship:
One user can be identified as one post owner.
One post can have one owner user.

ChatData Entity Relationship Diagram



comments	
comment_id	int
post_id	int
score	int
text	text
creation_date	timestamp
user_id	int

users	
user_id	int
reputation	int
creation_date	timestamp
display_name	varchar
last_access_date	timestamp
website_url	varchar
location	varchar
about_me	text
views	int
up_votes	int
down_votes	int
profile_image_url	varchar
account_id	int

posts	
post_id	int
post_type_id	bigint
accepted_answer_id	int
parent_id	int
creation_date	timestamp
score	int
view_count	int
body	text
owner_user_id	int
owner_display_name	char
last_editor_user_id	int
last_editor_display_name	char
last_edit_date	timestamp
last_activity_date	timestamp
title	text
tags	varchar
answer_count	int
comment_count	int
favorite_count	int
closed_date	timestamp
community_owned_date	timestamp

- **Is the data organised in a way that would lend itself to being managed in a relational database?**

Data is structured and can be managed in a relational database.

- **How would the different tables be connected?**

Table posts, comments and users connected by foreign keys.

- **What are the primary and foreign keys?**

Table **Users** has **Primary Key** *User_Id*.

Table **Comments** has **Primary Key** *Comment_Id*.

Table **Posts** has **Primary Key** *Post_Id*.

Table **Comments** has **Foreign Key** *Post_Id* that relates to table **Posts**.

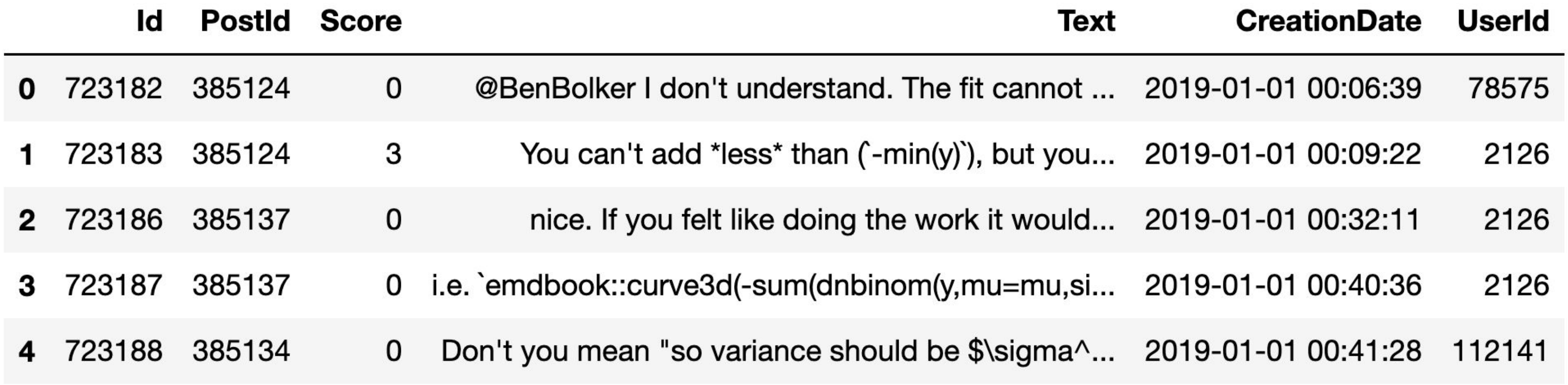
Table **Comments** has **Foreign Key** *User_Id* that relates to table **Users**.

Table **Posts** has **Foreign Key** *Owner_User_Id* that relates to table **Users**.

- **Would this give us a 3NF model?**

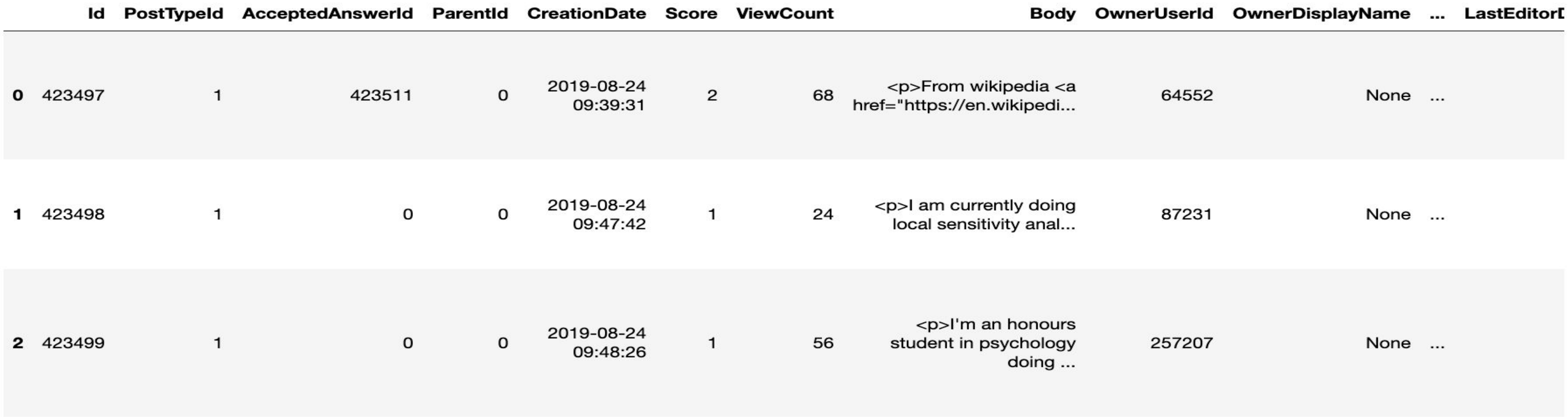
  Database model is in the 3NF model.

Table Comments



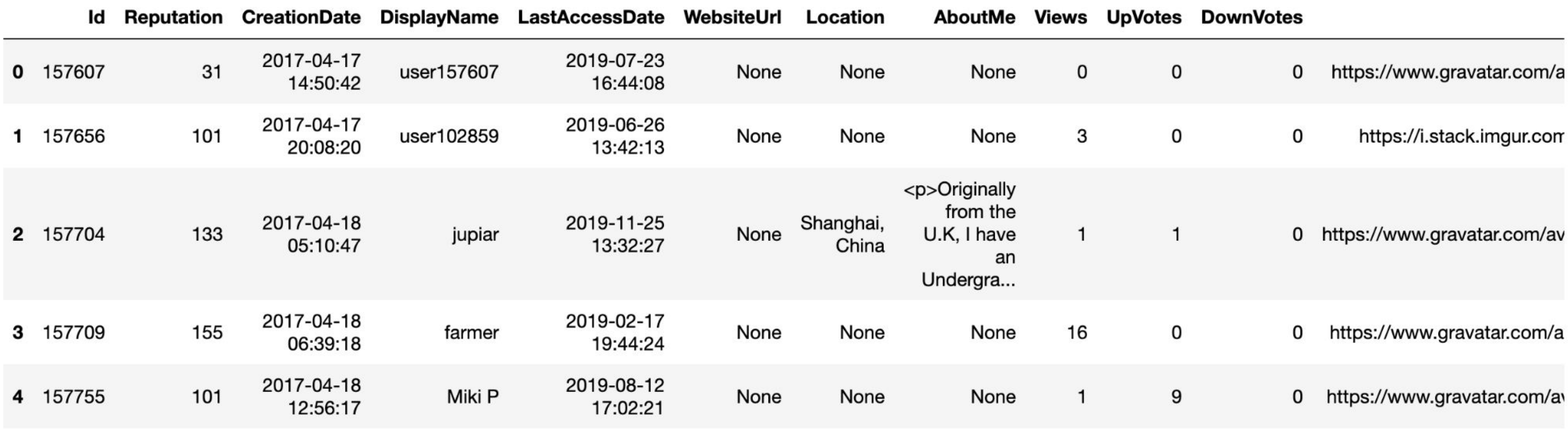
	Id	PostId	Score	Text	CreationDate	UserId
0	723182	385124	0	@BenBolker I don't understand. The fit cannot ...	2019-01-01 00:06:39	78575
1	723183	385124	3	You can't add *less* than (`-min(y)`), but you...	2019-01-01 00:09:22	2126
2	723186	385137	0	nice. If you felt like doing the work it would...	2019-01-01 00:32:11	2126
3	723187	385137	0	i.e. `emdbook::curve3d(-sum(dnbinom(y,mu=mu,si...	2019-01-01 00:40:36	2126
4	723188	385134	0	Don't you mean "so variance should be $\sigma^...	2019-01-01 00:41:28	112141

Table Posts



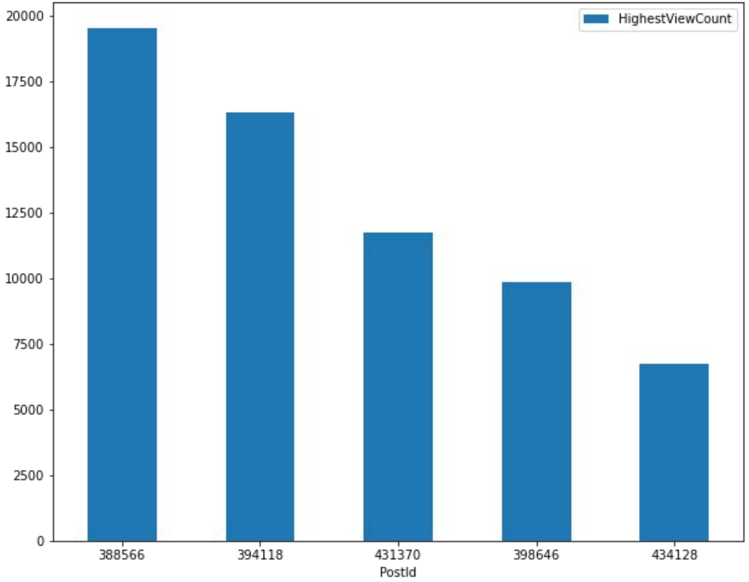
	Id	PostTypeId	AcceptedAnswerId	ParentId	CreationDate	Score	ViewCount	Body	OwnerUserId	OwnerDisplayName	...	LastEditorD
0	423497	1	423511	0	2019-08-24 09:39:31	2	68	<p>From wikipedia <a href="https://en.wikipedi...	64552	None	...	
1	423498	1	0	0	2019-08-24 09:47:42	1	24	<p>I am currently doing local sensitivity anal...	87231	None	...	
2	423499	1	0	0	2019-08-24 09:48:26	1	56	<p>I'm an honours student in psychology doing ...	257207	None	...	

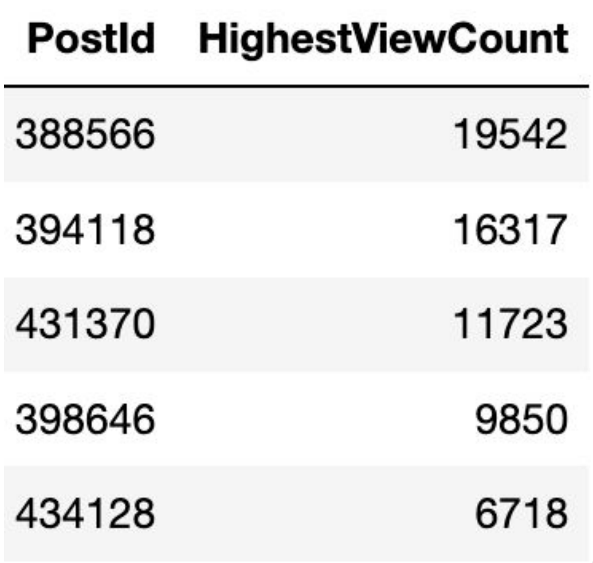
Table Users



	Id	Reputation	CreationDate	DisplayName	LastAccessDate	WebsiteUrl	Location	AboutMe	Views	UpVotes	DownVotes	
0	157607	31	2017-04-17 14:50:42	user157607	2019-07-23 16:44:08	None	None	None	0	0	0	https://www.gravatar.com/a
1	157656	101	2017-04-17 20:08:20	user102859	2019-06-26 13:42:13	None	None	None	3	0	0	https://i.stack.imgur.com
2	157704	133	2017-04-18 05:10:47	jupiar	2019-11-25 13:32:27	None	Shanghai, China	<p>Originally from the U.K, I have an Undergra...	1	1	0	https://www.gravatar.com/av
3	157709	155	2017-04-18 06:39:18	farmer	2019-02-17 19:44:24	None	None	None	16	0	0	https://www.gravatar.com/a
4	157755	101	2017-04-18 12:56:17	Miki P	2019-08-12 17:02:21	None	None	None	1	9	0	https://www.gravatar.com/av

Task 2: Create Single Table Queries to Analyze Engagement

<table>
<tr><td><b>How many posts have 0 comments?</b></td><td><b>21713</b></td><td><b>51,41% of all posts</b></td></tr>
<tr><td><b>How many posts have 1 comment?</b></td><td><b>6460</b></td><td><b>15,30% of all posts</b></td></tr>
<tr><td><b>How many posts have 2 comments or more?</b></td><td><b>14061</b></td><td><b>33,29% of all posts</b></td></tr>
<tr><td><b>Find the 5 posts with the highest View Count</b></td><td>


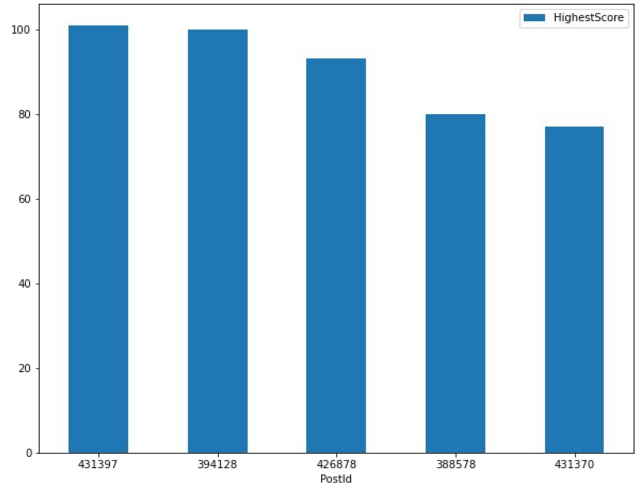
</td><td>


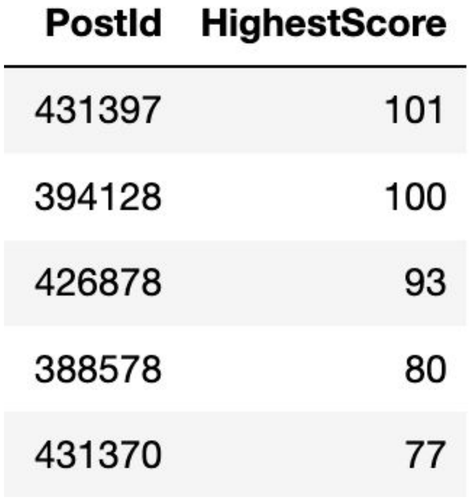
PostId	HighestViewCount
388566	19542
394118	16317
431370	11723
398646	9850
434128	6718

</td></tr>
</table>

Task 2: Create Single Table Queries to Analyze Engagement

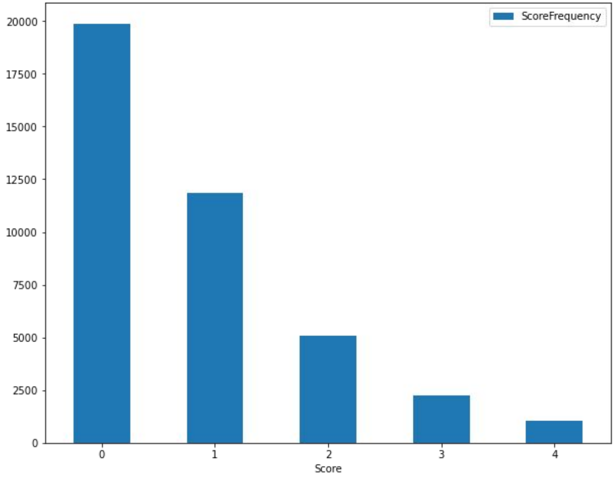
Find the top 5 posts with the highest scores

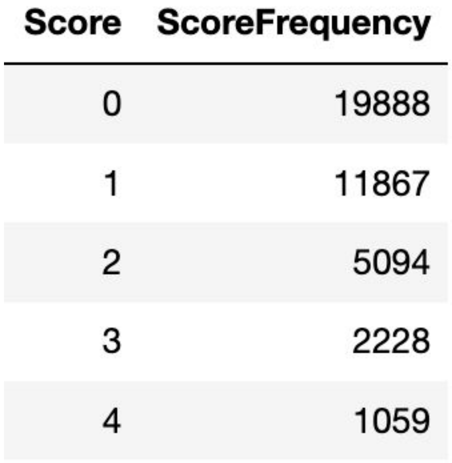




PostId	HighestScore
431397	101
394128	100
426878	93
388578	80
431370	77

What are the 5 most frequent scores on posts?





Score	ScoreFrequency
0	19888
1	11867
2	5094
3	2228
4	1059

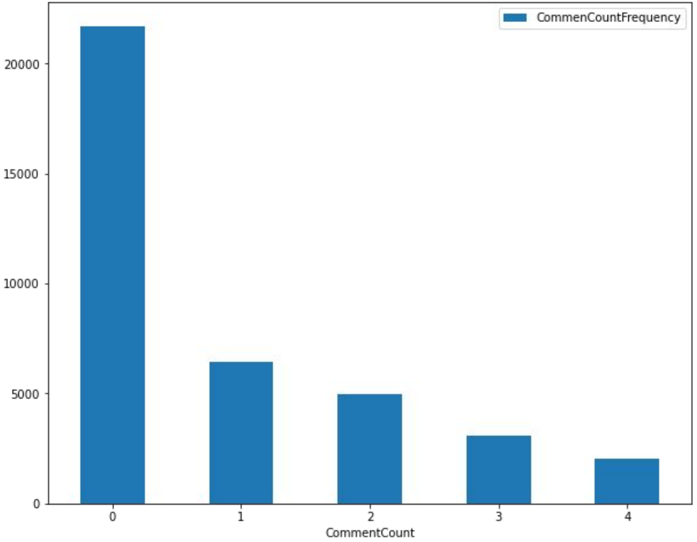
Task 2: Create Single Table Queries to Analyze Engagement

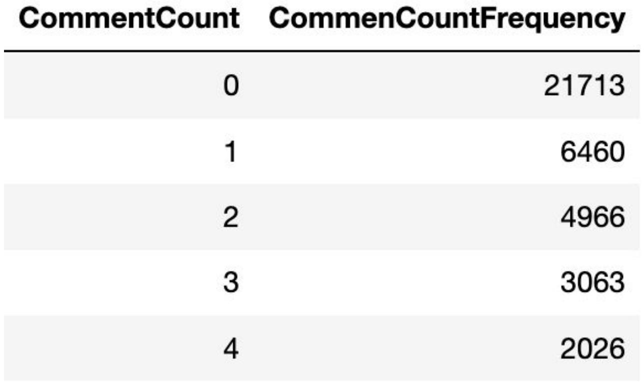
How many posts have the keyword "data" in their tags?

2242

5,3% of all posts

What are the 5 most frequent Comment Count for posts?





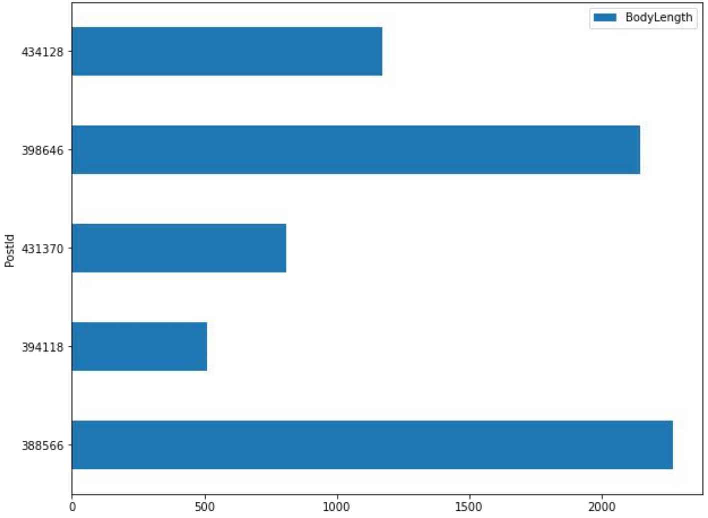
CommentCount	CommenCountFrequency
0	21713
1	6460
2	4966
3	3063
4	2026

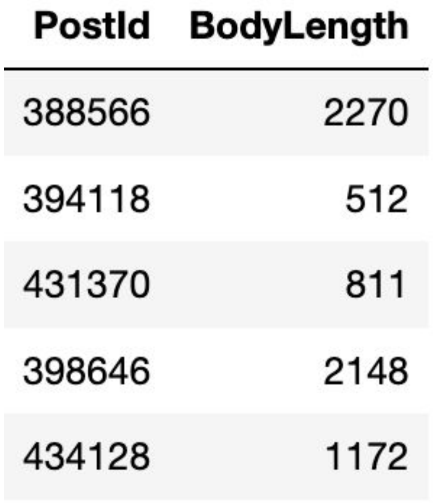
How many posts have an accepted answer?

5341

12,65% of all posts

Task 2: Create Single Table Queries to Analyze Engagement

<table>
<tr><td><b>What is the average reputation of table users?</b></td><td><b>312.35</b></td><td></td></tr>
<tr><td><b>What is the min reputation of users?</b></td><td><b>1</b></td><td></td></tr>
<tr><td><b>What is the max reputation of users?</b></td><td><b>228662</b></td><td><b>732 times higher than average users reputation</b></td></tr>
<tr><td><b>What is the length of the body of 5 most viewed posts?</b></td><td>


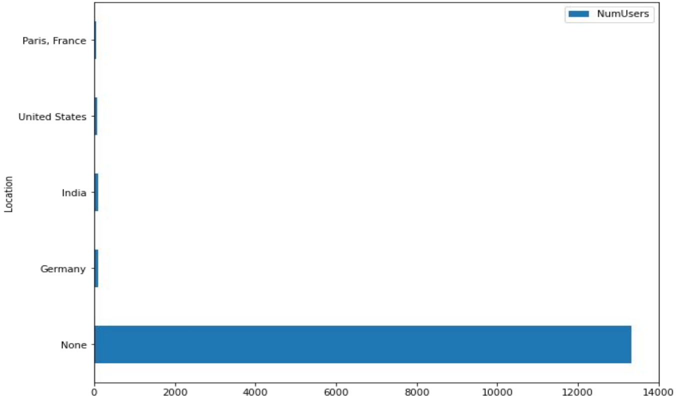
</td><td>

<table>
<tr><th>PostId</th><th>BodyLength</th></tr>
<tr><td>388566</td><td>2270</td></tr>
<tr><td>394118</td><td>512</td></tr>
<tr><td>431370</td><td>811</td></tr>
<tr><td>398646</td><td>2148</td></tr>
<tr><td>434128</td><td>1172</td></tr>
</table>
</td></tr>
</table>

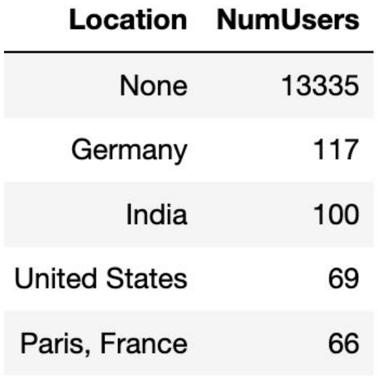
Task 2: Create Single Table Queries to Analyze Engagement

How many different locations are there in the users table?

1900

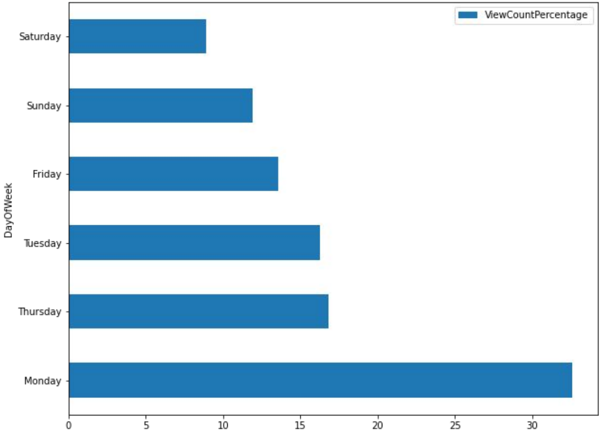
What are the top 5 locations of users?

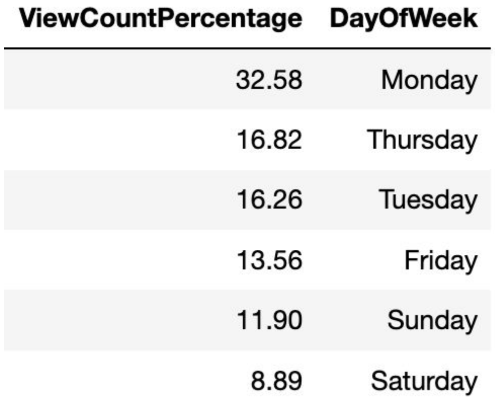




Location	NumUsers
None	13335
Germany	117
India	100
United States	69
Paris, France	66

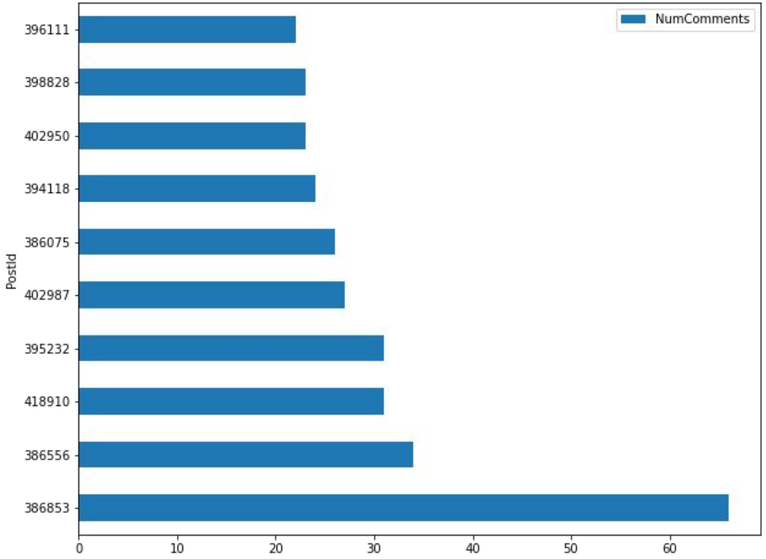
Rank the days of the week from highest to lowest in terms of the volume of View Count as a percentage.

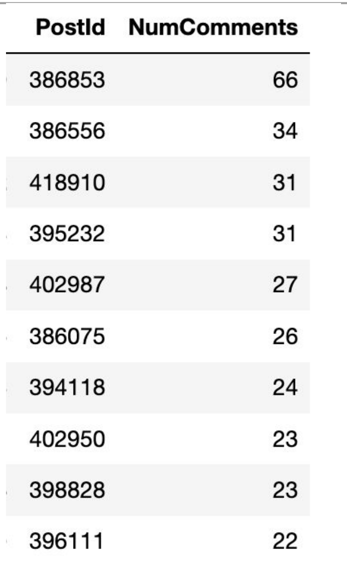




ViewCountPercentage	DayOfWeek
32.58	Monday
16.82	Thursday
16.26	Tuesday
13.56	Friday
11.90	Sunday
8.89	Saturday

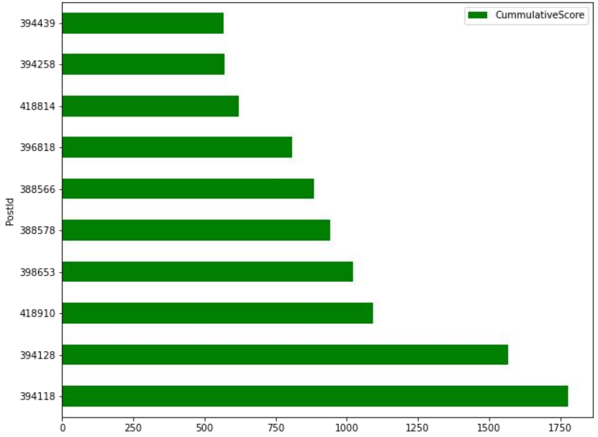
Task 3: Create Cross table queries to Further Analyze Engagement

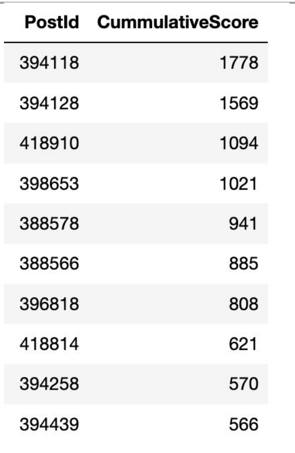
<table>
<tr>
<td><b>How many posts have been created by a user that has a filled out the "About Me" section?</b></td>
<td><b>17189</b></td>
<td><b>40,7% of all posts</b></td>
</tr>
<tr>
<td><b>Considering only the users with an "About Me," how many posts are there per user?</b></td>
<td><b>1,0</b></td>
<td></td>
</tr>
<tr>
<td><b>Not taking into account the Comment Count field in the table posts, what are the Top 10 posts in terms of number of comments?</b></td>
<td>


</td>
<td>

<table>
<tr><th>PostId</th><th>NumComments</th></tr>
<tr><td>386853</td><td>66</td></tr>
<tr><td>386556</td><td>34</td></tr>
<tr><td>418910</td><td>31</td></tr>
<tr><td>395232</td><td>31</td></tr>
<tr><td>402987</td><td>27</td></tr>
<tr><td>386075</td><td>26</td></tr>
<tr><td>394118</td><td>24</td></tr>
<tr><td>402950</td><td>23</td></tr>
<tr><td>398828</td><td>23</td></tr>
<tr><td>396111</td><td>22</td></tr>
</table>
</td>
</tr>
</table>

Task 3: Create Cross table queries to Further Analyze Engagement

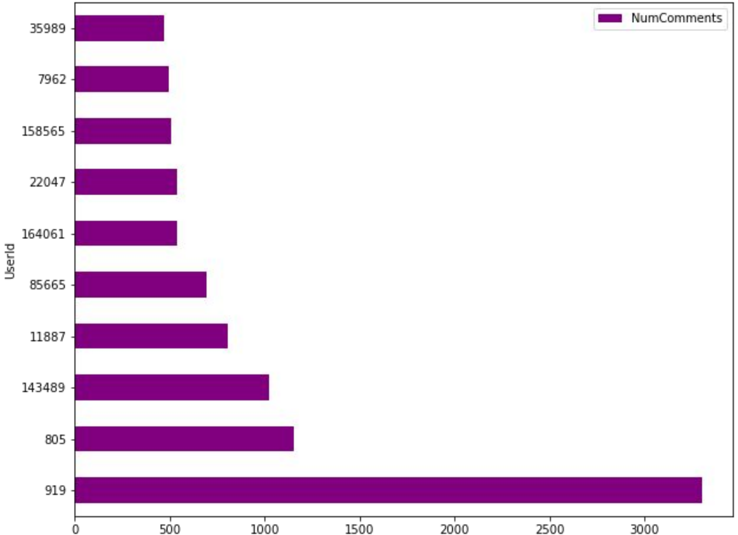
What are the Top 10 posts which have the highest cummulative (post score + comment score) score?

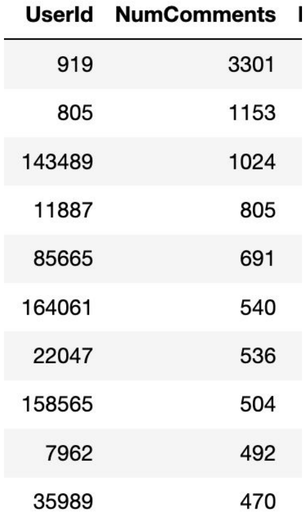




PostId	CummulativeScore
394118	1778
394128	1569
418910	1094
398653	1021
388578	941
388566	885
396818	808
418814	621
394258	570
394439	566

Who are the top 10 users who comment the most?

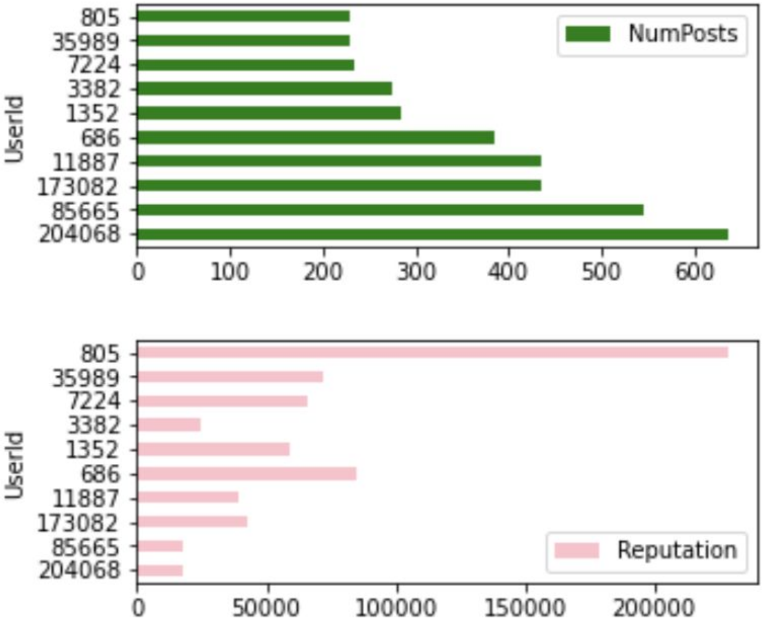


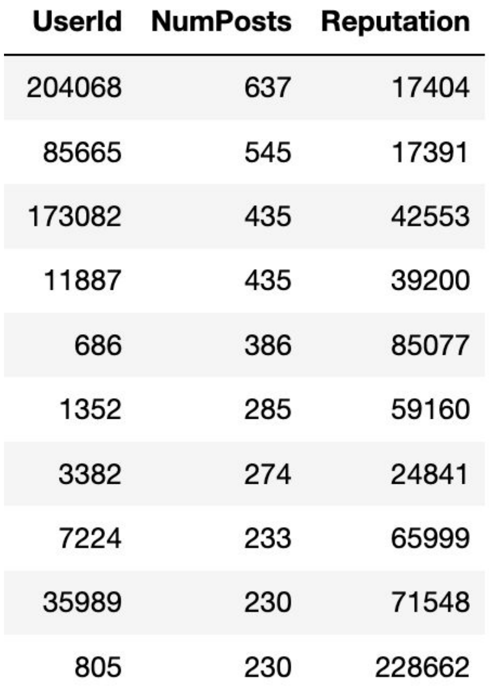


UserId	NumComments	I
919	3301	
805	1153	
143489	1024	
11887	805	
85665	691	
164061	540	
22047	536	
158565	504	
7962	492	
35989	470	

Task 3: Create Cross table queries to Further Analyze Engagement

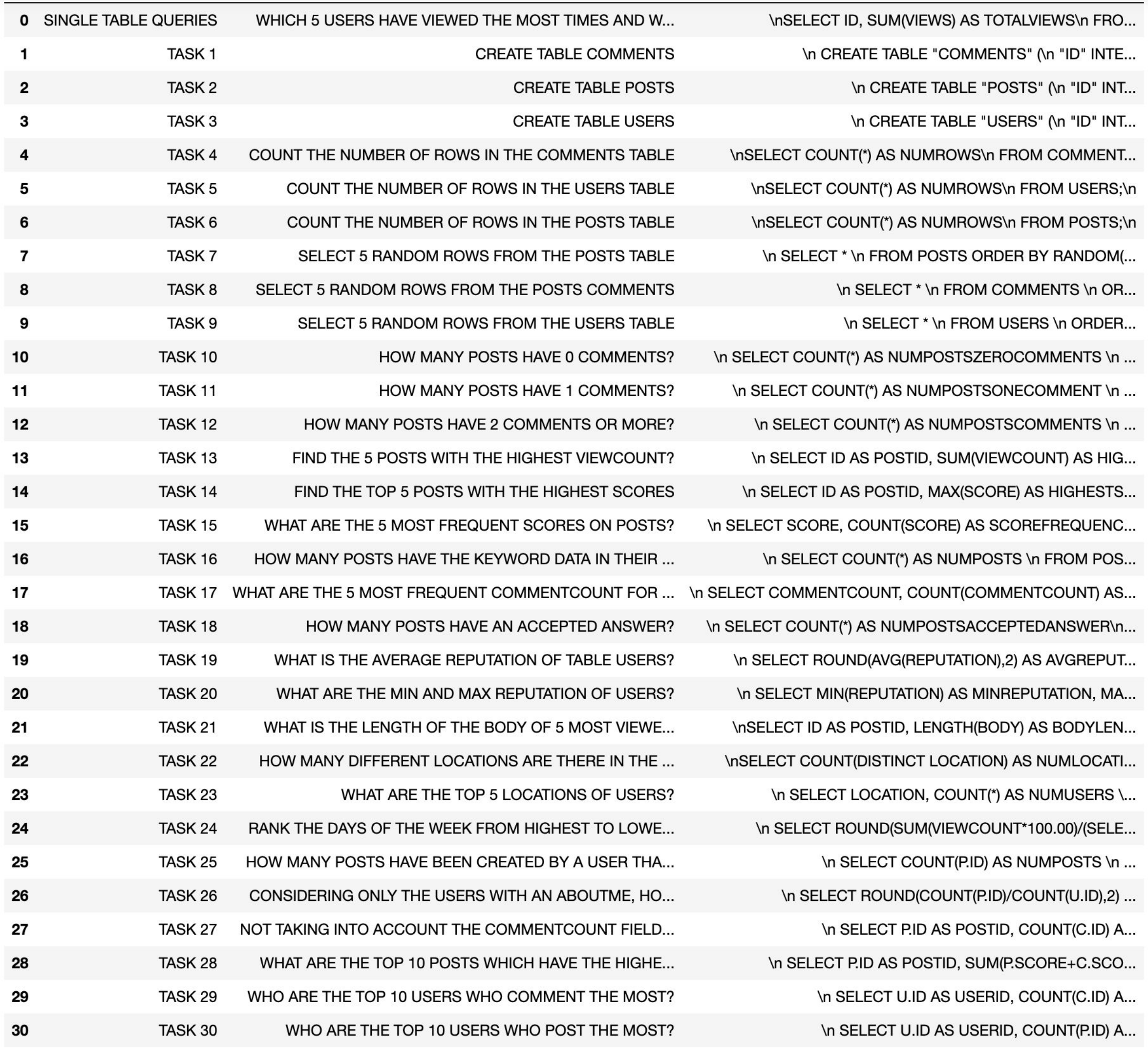
Who are the top 10 users who post the most?





UserId	NumPosts	Reputation
204068	637	17404
85665	545	17391
173082	435	42553
11887	435	39200
686	386	85077
1352	285	59160
3382	274	24841
7224	233	65999
35989	230	71548
805	230	228662

Task 4: Check the Queries table



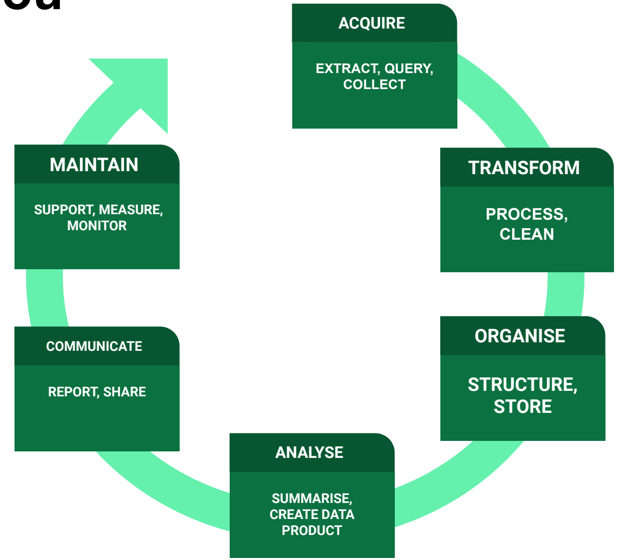
0	SINGLE TABLE QUERIES	WHICH 5 USERS HAVE VIEWED THE MOST TIMES AND W...	\nSELECT ID, SUM(VIEWS) AS TOTALVIEWS\n FRO...
1	TASK 1	CREATE TABLE COMMENTS	\n CREATE TABLE "COMMENTS" (\n "ID" INTE...
2	TASK 2	CREATE TABLE POSTS	\n CREATE TABLE "POSTS" (\n "ID" INT...
3	TASK 3	CREATE TABLE USERS	\n CREATE TABLE "USERS" (\n "ID" INT...
4	TASK 4	COUNT THE NUMBER OF ROWS IN THE COMMENTS TABLE	\nSELECT COUNT(*) AS NUMROWS\n FROM COMMENT...
5	TASK 5	COUNT THE NUMBER OF ROWS IN THE USERS TABLE	\nSELECT COUNT(*) AS NUMROWS\n FROM USERS;\n
6	TASK 6	COUNT THE NUMBER OF ROWS IN THE POSTS TABLE	\nSELECT COUNT(*) AS NUMROWS\n FROM POSTS;\n
7	TASK 7	SELECT 5 RANDOM ROWS FROM THE POSTS TABLE	\n SELECT * \n FROM POSTS ORDER BY RANDOM(...
8	TASK 8	SELECT 5 RANDOM ROWS FROM THE POSTS COMMENTS	\n SELECT * \n FROM COMMENTS \n OR...
9	TASK 9	SELECT 5 RANDOM ROWS FROM THE USERS TABLE	\n SELECT * \n FROM USERS \n ORDER...
10	TASK 10	HOW MANY POSTS HAVE 0 COMMENTS?	\n SELECT COUNT(*) AS NUMPOSTSZEROCOMMENTS \n ...
11	TASK 11	HOW MANY POSTS HAVE 1 COMMENTS?	\n SELECT COUNT(*) AS NUMPOSTSONECOMMENT \n ...
12	TASK 12	HOW MANY POSTS HAVE 2 COMMENTS OR MORE?	\n SELECT COUNT(*) AS NUMPOSTSCOMMENTS \n ...
13	TASK 13	FIND THE 5 POSTS WITH THE HIGHEST VIEWCOUNT?	\n SELECT ID AS POSTID, SUM(VIEWCOUNT) AS HIG...
14	TASK 14	FIND THE TOP 5 POSTS WITH THE HIGHEST SCORES	\n SELECT ID AS POSTID, MAX(SCORE) AS HIGHESTS...
15	TASK 15	WHAT ARE THE 5 MOST FREQUENT SCORES ON POSTS?	\n SELECT SCORE, COUNT(SCORE) AS SCOREFREQUENC...
16	TASK 16	HOW MANY POSTS HAVE THE KEYWORD DATA IN THEIR ...	\n SELECT COUNT(*) AS NUMPOSTS \n FROM POS...
17	TASK 17	WHAT ARE THE 5 MOST FREQUENT COMMENTCOUNT FOR ...	\n SELECT COMMENTCOUNT, COUNT(COMMENTCOUNT) AS...
18	TASK 18	HOW MANY POSTS HAVE AN ACCEPTED ANSWER?	\n SELECT COUNT(*) AS NUMPOSTSACCEPTEDANSWER\n...
19	TASK 19	WHAT IS THE AVERAGE REPUTATION OF TABLE USERS?	\n SELECT ROUND(AVG(REPUTATION),2) AS AVGREPUT...
20	TASK 20	WHAT ARE THE MIN AND MAX REPUTATION OF USERS?	\n SELECT MIN(REPUTATION) AS MINREPUTATION, MA...
21	TASK 21	WHAT IS THE LENGTH OF THE BODY OF 5 MOST VIEWE...	\nSELECT ID AS POSTID, LENGTH(BODY) AS BODYLEN...
22	TASK 22	HOW MANY DIFFERENT LOCATIONS ARE THERE IN THE ...	\nSELECT COUNT(DISTINCT LOCATION) AS NUMLOCATI...
23	TASK 23	WHAT ARE THE TOP 5 LOCATIONS OF USERS?	\n SELECT LOCATION, COUNT(*) AS NUMUSERS \...
24	TASK 24	RANK THE DAYS OF THE WEEK FROM HIGHEST TO LOWE...	\n SELECT ROUND(SUM(VIEWCOUNT*100.00)/(SELE...
25	TASK 25	HOW MANY POSTS HAVE BEEN CREATED BY A USER THA...	\n SELECT COUNT(P.ID) AS NUMPOSTS \n ...
26	TASK 26	CONSIDERING ONLY THE USERS WITH AN ABOUTME, HO...	\n SELECT ROUND(COUNT(P.ID)/COUNT(U.ID),2) ...
27	TASK 27	NOT TAKING INTO ACCOUNT THE COMMENTCOUNT FIELD...	\n SELECT P.ID AS POSTID, COUNT(C.ID) A...
28	TASK 28	WHAT ARE THE TOP 10 POSTS WHICH HAVE THE HIGHE...	\n SELECT P.ID AS POSTID, SUM(P.SCORE+C.SCO...
29	TASK 29	WHO ARE THE TOP 10 USERS WHO COMMENT THE MOST?	\n SELECT U.ID AS USERID, COUNT(C.ID) A...
30	TASK 30	WHO ARE THE TOP 10 USERS WHO POST THE MOST?	\n SELECT U.ID AS USERID, COUNT(P.ID) A...

Conclusions

1. Only **49%** of all posts have at least one comment or 2 and more comments. And **51%** of all posts don’t have any comment. This shows us that we should analyse posts subjects and find out what subjects are interesting and popular for ChatData users.
2. Based on analysing posts *View Counts* and *Scores* variables we can define the 5 most popular and interesting subjects of posts that are the most viewed and received the highest scores.
3. Only **12%** of all posts have accepted answer.
4. The most viewed posts are the posts with body length that varies from **512** to **2270 characters**.
5. ChatData users from **1900** locations. **72%** of all users location isn’t defined. Among known locations are **Germany**, **India** and **United States**.
6. The highest *View Count* is on **Monday** and **Thursday**. On this days of week posts on ChatData are most viewed.
7. Based on analysing users ids with the highest post and comment counts we can identify the most active and productive ChatData audience.

Task 5: Report back on process

Data Analysis LifeCycle



- **Which aspects of the data analysis lifecycle are you primarily involved with on this project?**

This project fit into five stages of Data Analysis LifeCycle: **Acquire, Transform, Organise, Analyse, Communicate, Maintain data.**

- **What activities would you need to do before undertaking this project? Think about where the data came from.**

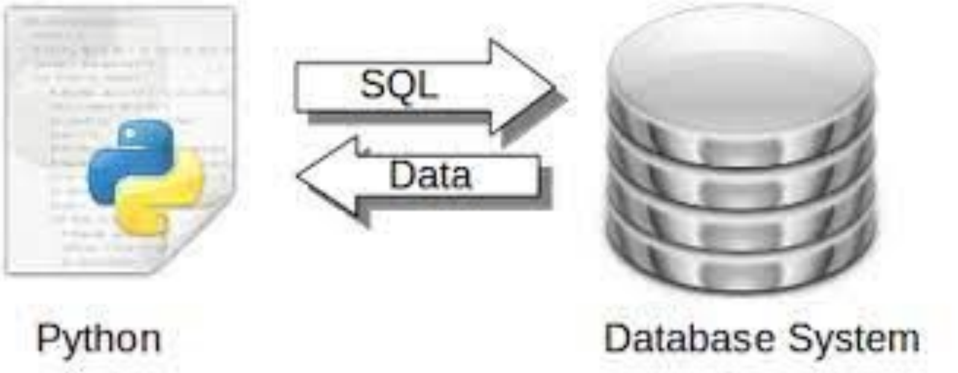
Data classification – the first step to protecting companies and users sensitive data.

Data classification provides one of the most basic ways for organisations to determine and assign relative values to the data they possess. The process of *data classification* allows Data Analyst to categorise stored data by sensitivity and business impact so Data Analyst can understand associated risks with the data.

Task 5: Report back on process

- **What would you need to do to allow this analysis to be repeated with updated data, and how would this solution be maintained?**

Analysis can be repeated with updated data by replacing data sources. We can edit an existing database sources by updating the database connection.

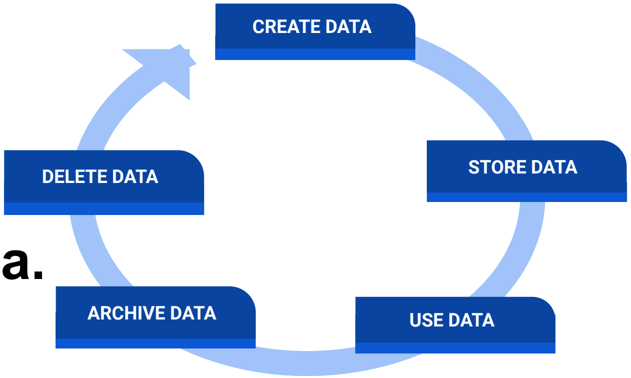


- **How does this project fit into a broader data lifecycle?**

This project fit into three stages of Data LifeCycle:

Create, Store, Use data, Archive data, Delete Data.

Data LifeCycle



Task 5: Report back on process

Requirements

- **What aspects of the requirements were not 100% clear to you?**

In order to use variables from datasets for analysis we need to know detailed information how they were measured and counted, definitions of variables, data types.

- **Would it have been easier if you could talk directly to Oliver? If so, what sorts of questions would you want to ask him?**
- What exactly do you want to find out?
- How will analysis results be used?
- What data visualizations will be needed?
- Who should be able to access the information?
- Will reports be developed and maintained?
- What information will be on report?
- What reports currently exist in another format?
- What changes might be made to existing reports?
- What ETLs or stored procedures need to be developed, if any?

Task 5: Report back on process

● How did you analyse the requirements?

Step 1: *Categorize Requirements*

- Entity Relationship Diagram
- Single Table Queries
- Cross table queries
- Queries table

Step 2: *Interpret Requirements*

- Identify Entities and Relationships, create ERD
- Writing Single Table Queries
- Writing Cross table queries
- Writing Queries table

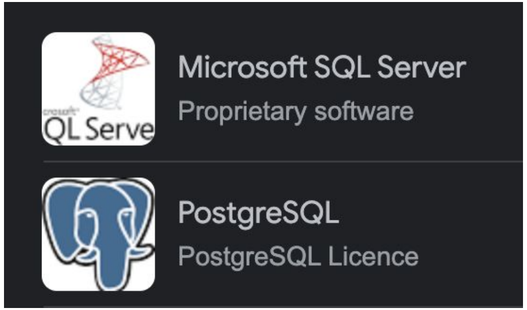
Step 3: *Record Requirements*

- Entity Relationship Diagram is exported in image file
- Retrieving data for Single Table Queries
- Retrieving data for Cross table queries
- Queries table is written in file

Task 5: Report back on process

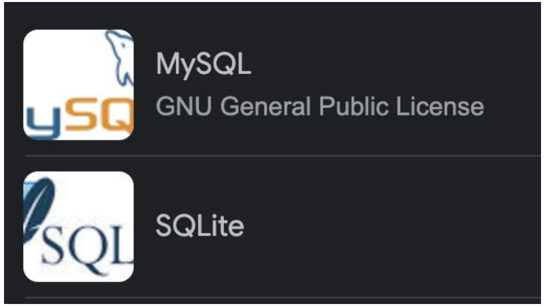
Tools

- **Were the tools you used appropriate for the job?**



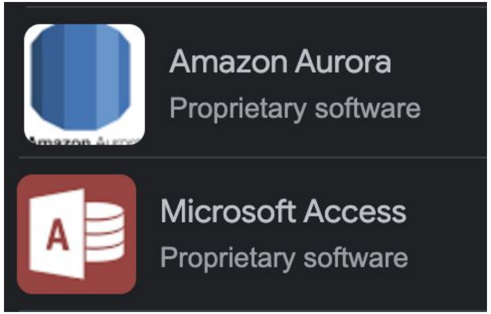
Microsoft SQL Server
Proprietary software

PostgreSQL
PostgreSQL Licence



MySQL
GNU General Public License

SQLite



Amazon Aurora
Proprietary software

Microsoft Access
Proprietary software

- **Why was SQLite used over PostgreSQL for this project?**

SQLite and *PostgreSQL* are among the most widely used relational database management systems (RDMS). They are both open-source and free, but they have some major differences that should be considered when choosing a database to use for your business.

SQLite is highly useful for:

- Standalone apps
- Small apps that don't require expansion.
- Apps need to read or write files to disk directly.

PostgreSQL is recommended when:

- Data integrity and reliability is highly concerned.
- Custom Procedures which is extensible to run the complex task.
- Complexity with ease.

Task 5: Report back on process

Tools

- **What are the benefits of adopting relational database technologies for an IT organisation?**

Adopting relational database technologies helps improve organizational security, integration, compliance, and performance.

- Improved data sharing and data security
- Effective data integration
- Consistent, reliable data
- Data that complies with privacy regulations
- Increased productivity
- Better decision-making

Task 5: Report back on process

Data Quality

- **What did you think of the organisation of the source data and how that mapped to the structure of your database? Was there a natural mapping from the CSV files to the database tables?**

Data mapping is the process of connecting a data field from one source to a data field in another source. This reduces the potential for errors, helps standardize data and makes it easier to understand data.

In Relational Databases data gets stored in a table format so by using CSV File, the database can be created.

- **Was the data consistent? In other words, were there any issues with the data that prevented you from producing good quality results?**

Data is considered consistent if two or more values in different locations are identical.

Poor quality data can seriously harm business. It can lead to inaccurate analysis, poor customer relations and poor business decisions.

Task 5: Report back on process

- **Did you find the definitions of the data were detailed enough to assist you in the tasks?**

A *data dictionary* is a collection of metadata such as object name, data type, size, classification, and relationships with other data assets. A data dictionary acts as a reference guide on a dataset.

- **What steps would you take to ensure that any concerns about data quality are dealt with appropriately? Who would you talk to about this?**

- **Fix data in the source system.** Fixing data in the source system is often the best way to ensure effective customer experiences and analysis on the other end of the process.
- **Fix the source system to correct data issues.** The source system that collects data can be set up to automatically cleanse data before it enters the database.
- **Accept bad source data and fix issues during the ETL phase.** Before customer data can be analyzed, it’s frequently put through an extract, transform, and load (ETL) process.
- **Apply precision identity/entity resolution.** One of the most significant issues with many customer/users databases is that they have multiple records for the same customer/user, and no way to tell that these pieces of information are interrelated.

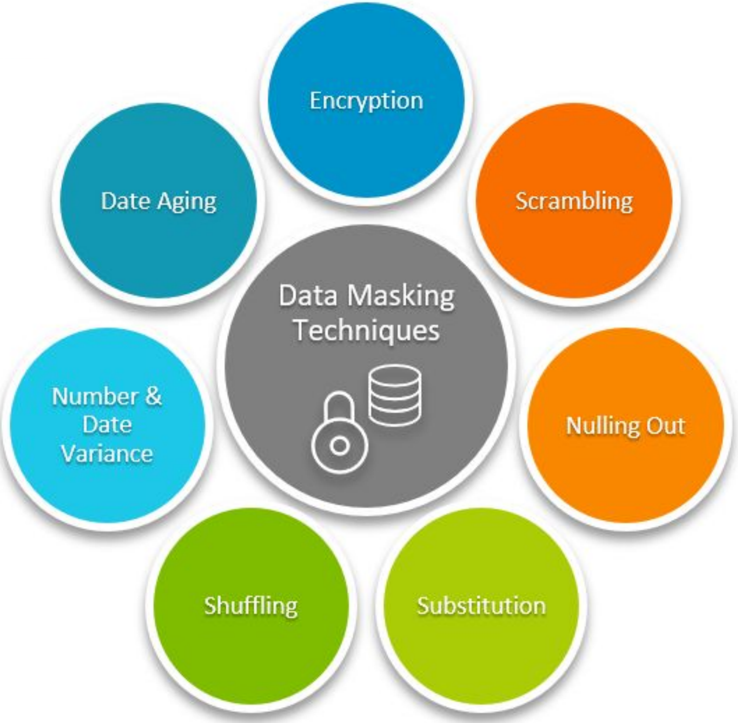
Task 5: Report back on process

Security, ethics and legislation

- **Was the data sufficiently masked, or are there personal details present in the data?**

Personal data is information that relates to an identified or identifiable individual.What identifies an individual could be as simple as a name or a number or could include other identifiers such as an IP address or a cookie identifier, or other factors.
Users dataset contain Location and ProfileImageUrl. These columns contain information that relates to identifiable individual.

- **Can you think of possible techniques (e.g. statistics) that could unmask the participants in this data?**



Task 5: Report back on process

Security, ethics and legislation

- **Are there any ethical considerations with our intended use of this data?**

Ethical considerations are the ethical practices that govern how data is gathered, stored, and exchanged. These can include obtaining unambiguous and informed consent, storing data securely, and obtaining permissions to use or share data.

- **Is the use of this data covered by any legislation, and if so, what is that legislation?**

The Data Protection Act (DPA) controls how personal information can be used and person rights to ask for information about himself.

The Data Protection Act 2018 controls how your personal information is used by organisations, businesses or the government. The Data Protection Act 2018 is the UK's implementation of the General Data Protection Regulation (GDPR).

- **What did you learn about the data analysis lifecycle?**

The **Data Analytics Lifecycle** outlines how data is created, gathered, processed, used, and analyzed to meet business objectives. It provides a structured method of handling data so that it may be transformed into knowledge that can be applied to achieve business project objectives.

- **Which stages are predominant in this project?**

Organise and Analyse Data stages are predominant in this project. These stages involves *organizing*, *storing*, and *retrieving* data as necessary over the life of a data project.

- **What issues arose around data privacy and which types of data were in this project?**

Datasets contain ***for internal use only (sensitive)*** and ***public (unrestricted) data***. For internal use only that is classified as sensitive, would not have a severe impact if lost or destroyed (email, location data). Public data that is classified as public includes data and files that are not critical to business needs or operations.

- **Can you explain the business drivers for relational database technology and some of the design issues?**

The primary benefit of the relational database approach is the **ability to create meaningful information by joining the tables**. Joining tables allows to understand the relations between the data, or how the tables connect. SQL includes the ability to count, add, group, and also combine queries.

Ten Common Database Design Issues:

1. Poor design/planning
2. Ignoring normalization
3. Poor naming standards
4. Lack of documentation
5. One table to hold all domain values
6. Using identity/guid columns as your only key
7. Not using SQL facilities to protect data integrity
8. Not using stored procedures to access data
9. Trying to build generic objects
10. Lack of testing