

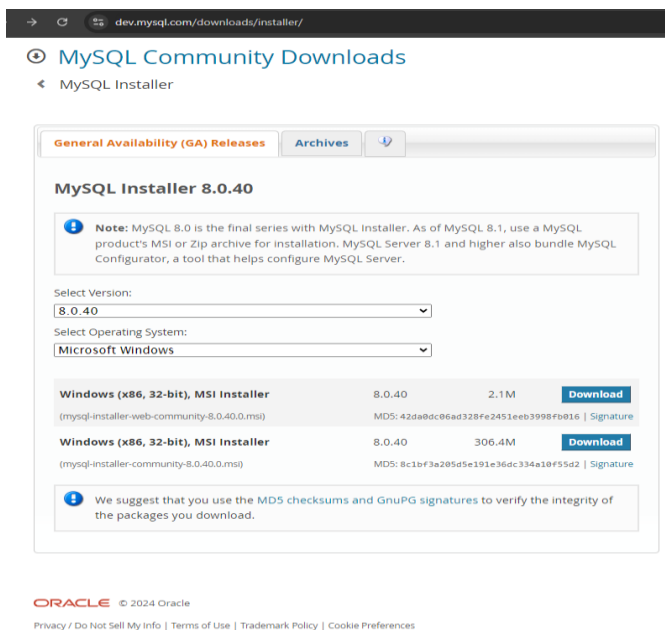
Document highlight:

- Tool Installation
- Create a database (create tables and load the dataset)
- Query run to ensure that (i) tables were created and (ii) that dataset were imported
- Data Cleansing with SQL queries (query and results):
 - Remove special characters
 - Updating and grouping undetermined record values
 - Leveraging SQL clauses (ie., Select, Create, Update, Join), aggregate functions (ie., Count, Sum), and virtual view.

1. Tool setup and installation

Download and install (follow installation guide):

<https://dev.mysql.com/downloads/installer/>







MySQL home screen:

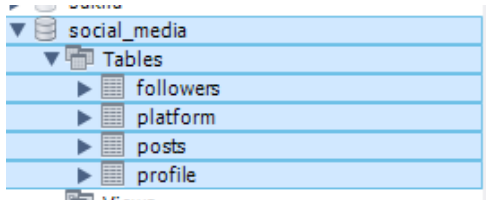


2. Create a database (create tables and load the dataset)

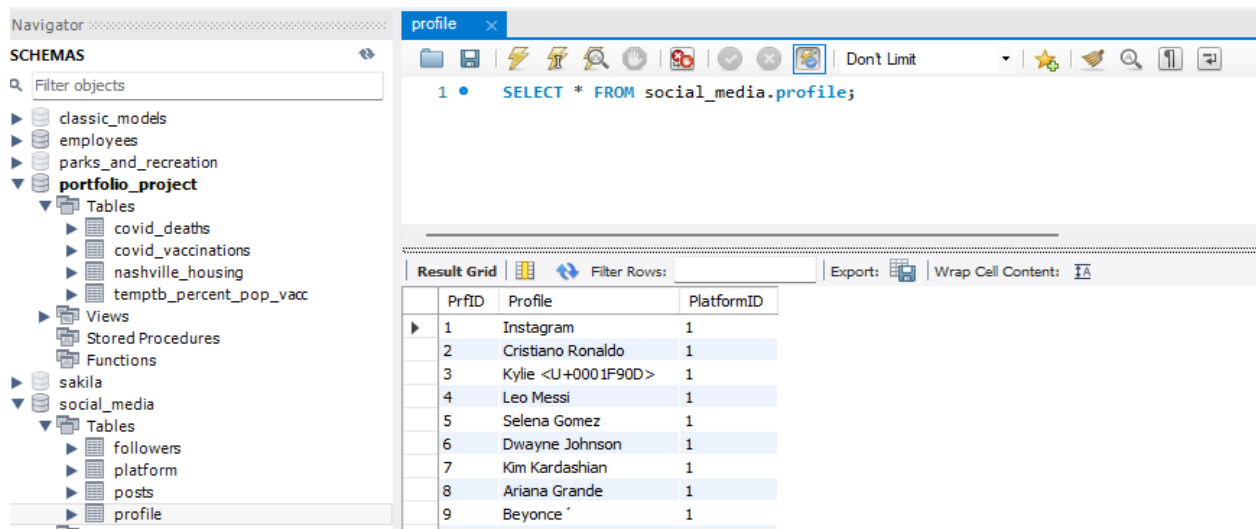
Sampling dataset:

 Followers	2024-11-24 5:15 PM	Microsoft Office Excel Comma Sep
 Platform	2024-11-24 5:17 PM	Microsoft Office Excel Comma Sep
 Posts	2024-11-24 5:16 PM	Microsoft Office Excel Comma Sep
 Profile	2024-11-24 5:14 PM	Microsoft Office Excel Comma Sep

Create a database and import the file (going into the table) in MySQL Workbench:



3. Query run to ensure that (i) tables were created and (ii) that dataset were imported



The screenshot shows the MySQL Workbench interface. The 'Navigator' pane on the left shows the 'social_media' database with tables 'followers', 'platform', 'posts', and 'profile'. The 'profile' table is selected. The 'Query Editor' pane shows the query: `SELECT * FROM social_media.profile;`. The 'Result Grid' pane shows the following data:

PrfID	Profile	PlatformID
1	Instagram	1
2	Cristiano Ronaldo	1
3	Kylie <U+0001F90D>	1
4	Leo Messi	1
5	Selena Gomez	1
6	Dwayne Johnson	1
7	Kim Kardashian	1
8	Ariana Grande	1
9	Beyonce *	1

4. Data cleansing with SQL queries

Remove special characters:

profile x

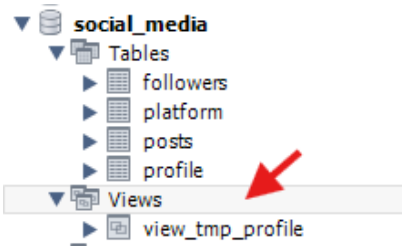
Don't Limit

1 • SELECT * FROM social_media.profile;

Result Grid Filter Rows: Export: Wrap Cell Content: [FA](#)

	PrfID	Profile	PlatformID
1		Instagram	1
2		Cristiano Ronaldo	1
3		Kyrie <U+0001F90D>	1
4		Leo Messi	1
5		Selena Gomez	1
6		Dwayne Johnson	1
7		Kim Kardashian	1
8		Arian Grande	1
9		Beyonce	1
10		Khloe Kardashian	1
11		Justin Bieber	1
12		Kendall	1
13		National Geographic	1
14		Nike	1
15		Taylor Swift	1
16		Jennifer Lopez	1
17		Virat Kohli	1
18		Barbie	1
19		Kourtney Kardashian ...	1
20		Miley Cyrus	1
21		N <U+0001F1E7> <...>	1
22		KARY PERDI	1
23		Kevin Hart	1
24		Zendaya	1


Create a view with a column of cleaned value...



```

4 • DROP VIEW IF EXISTS view_tmp_Profile;
5
6 • CREATE VIEW view_tmp_Profile AS
7     SELECT
8         p1.PrfID,
9         p1.PlatformID,
10        p1.Profile,
11    CASE
12        WHEN SUBSTRING(p1.Profile, 1, locate('<', p1.Profile) - 1) <> '' THEN SUBSTRING(p1.Profile, 1, locate('<', p1.Profile) - 1)
13        END AS Cleaned
14    FROM
15        social_media.profile as p1
16    ;
17

```



19	1	Kourtney Kardashian Barker	<U+2764><U+FE0F><U+200D><U+0001F525>	Kourtney Kardashian Barker
20	1	Miley Cyrus		NULL
21	1	NJ	<U+0001F1E7><U+0001F1F7>	NJ
22	1	KATT PERKINS		NULL
23	1	Kevin Hart		NULL

From the view, find the record that requires an update...

```

18 • SELECT
19     *
20 FROM
21     social_media.view_tmp_Profile p1
22 JOIN
23     social_media.profile as p2
24 ON
25     p1.PrfID = p2.PrfID
26 AND
27     p1.PlatformID = p2.PlatformID
28 AND
29     p1.Cleaned IS NOT NULL
30 ;

```

Result Grid Filter Rows: Export: Wrap Cell Content:						
PrfID	PlatformID	Profile	Cleaned	PrfID	Profile	
3	1	Kylie <U+0001F90D>	Kylie	3	Kylie <U+0001F90D>	
19	1	Kourtney Kardashian Barker <U+2764><U+FE0F><U+200D><U+0001F525>	Kourtney Kardashian Barker	19	Kourtney Kardashian Barker <U+2764><U+FE0F><U+200D><U+0001F525>	
21	1	NJ <U+0001F1E7><U+0001F1F7>	NJ	21	NJ <U+0001F1E7><U+0001F1F7>	
46	1	Shraddha <U+2736>	Shraddha	46	Shraddha <U+2736>	
54	1	Alia Bhatt <U+0001F90D><U+2600><U+FE0F>	Alia Bhatt	54	Alia Bhatt <U+0001F90D><U+2600><U+FE0F>	
66	1	Anitta <U+0001F3A4>	Anitta	66	Anitta <U+0001F3A4>	
69	1	JISOO <U+0001F90D>	JISOO	69	JISOO <U+0001F90D>	
88	1	Bella <U+0001F98B>	Bella	88	Bella <U+0001F98B>	
95	1	disha patani (paatni) <U+0001F98B>	disha patani (paatni)	95	disha patani (paatni) <U+0001F98B>	
65	5	shfa2 - <U+0634><U+0641><U+0627>	shfa2 -	65	shfa2 - <U+0634><U+0641><U+0627>	
22	2	Junya/ <U+3058><U+3085><U+3093><U+3084>	Junya/	22	Junya/ <U+3058><U+3085><U+3093><U+3084>	
28	2	ROD <U+0001FAE0>	ROD	28	ROD <U+0001FAE0>	
41	2	kyle thomas <U+270C><U+FE0F>	kyle thomas	41	kyle thomas <U+270C><U+FE0F>	
44	2	BRIANDA <U+0001F496>	BRIANDA	44	BRIANDA <U+0001F496>	
56	2	Arishfa Khan <U+0001F981>	Arishfa Khan	56	Arishfa Khan <U+0001F981>	
61	2	Ignacia Antonia <U+0001F451>	Ignacia Antonia	61	Ignacia Antonia <U+0001F451>	
65	2	Naim Darrechi <U+0001F3C6>	Naim Darrechi	65	Naim Darrechi <U+0001F3C6>	
90	2	DonaldDucc <U+0001F986>	DonaldDucc	90	DonaldDucc <U+0001F986>	
96	2	JAY CROES <U+0001F4A0>	JAY CROES	96	JAY CROES <U+0001F4A0>	

Update the main table with cleaned record value from the temp (view) table...

```
31
32 • UPDATE
33     social_media.profile as p2
34     JOIN
35         social_media.view_tmp_Profile p1
36     ON
37         p1.PrfID = p2.PrfID
38     AND
39         p1.PlatformID = p2.PlatformID
40     AND
41         p1.Cleaned IS NOT NULL
42     SET
43         p2.Profile = p1.Cleaned
44     ;
```

129 19:11:08 UPDATE social_media.profile as p2 JOIN social_media.view_tmp_Profile p1 ON p1.PrfID = p2.PrfID AND p1.PlatformID = p2.PlatformID AND p1.Cleaned IS NOT NULL ... 1.. 0.281 sec

Check the record post update...

Don't

```
• Select * from social_media.profile;
```

Result Grid			Filter Rows:	Export:	Wrap Cell Content:
PrfID	Profile	PlatformID			
1	Instagram	1			
2	Cristiano Ronaldo	1			
3	Kylie	1			
4	Leo Messi	1			
5	Selena Gomez	1			
6	Dwayne Johnson	1			
7	Kim Kardashian	1			
8	Ariana Grande	1			
9	Beyonce	1			
10	Khloe Kardashian	1			
11	Justin Bieber	1			
12	Kendall	1			
13	National Geographic	1			
14	Nike	1			
15	Taylor Swift	1			
16	Jennifer Lopez	1			
17	Virat Kohli	1			
18	Barbie	1			
19	Kourtney Kardashian Barker	1			
20	Miley Cyrus	1			
21	NJ	1			
22	KATY PERRY	1			

Result Grid			Filter Rows:	Export:	Wrap Cell Contents:	Fetch rows:		
UniqueID	ParcelID	LandUse	PropertyAddress	SaleDate	SalePrice	LegalReference	SoldAsVacant	OwnerName
2045	007 00 0 125.00	SINGLE FAMILY	1808 FOX CHASE DR, GOODLETTSVILLE	2013-04-09 0:00	240000	20130412-0036474	No	FRAZIER, CYRE
16918	007 00 0 130.00	SINGLE FAMILY	1832 FOX CHASE DR, GOODLETTSVILLE	2014-06-10 0:00	366000	20140619-0053768	No	BONER, CHARLI
54582	007 00 0 138.00	SINGLE FAMILY	1864 FOX CHASE DR, GOODLETTSVILLE	2016-09-26 0:00	435000	20160927-0101718	No	WILSON, JAMES
43070	007 00 0 143.00	SINGLE FAMILY	1853 FOX CHASE DR, GOODLETTSVILLE	2016-01-29 0:00	255000	20160129-0008913	No	BAKER, JAY K.
22714	007 00 0 149.00	SINGLE FAMILY	1829 FOX CHASE DR, GOODLETTSVILLE	2014-10-10 0:00	278000	20141015-0095255	No	POST, CHRISTOPHER
18367	007 00 0 151.00	SINGLE FAMILY	1821 FOX CHASE DR, GOODLETTSVILLE	2014-07-16 0:00	267000	20140718-0063802	No	FIELDS, KAREN

25 09:04:29 SELECT * FROM portfolio_project.nashville_housing 24007 row(s) returned 0.000 sec / 0.079 sec

Updating and grouping undetermined record values:

Clean up the remaining record (update the Profile as “Others”)

```
53 • SELECT
54     p1.PrfID,
55     p1.Profile
56 FROM
57     social_media.view_tmp_Profile p1
58 WHERE p1.Profile LIKE '<%'
59 ;
```

Result Grid		Filter Rows:	Export:	Wrap Cell Content:
PrfID	Profile			
27	<U+0001F451>			
30	<U+BC29><U+D0C4><U+C18C><U+B144><U+B2E8>			
6	<U+273F> Kids Diana Show			
55	<U+041C><U+0430><U+0448><U+0430><U+0438><U+041C><U+043...			
19	<U+C6D0><U+C815><U+B9E8> WonJeong			
25	<U+0425><U+041E><U+041C><U+042F><U+041A>			
32	<U+0410>nokhina Liza			
100	<U+0001F49B>julia menu garcia<U+0001F49B>			

Update the main table with cleaned record value from the temp (view) table...

```
2  -- update the main table with cleaned record value from the temp (view) table
3 • UPDATE
4     social_media.profile as p2
5     JOIN
6         social_media.view_tmp_Profile p1
7     ON
8         p1.PrfID = p2.PrfID
9     AND
10        p1.PlatformID = p2.PlatformID
11    AND
12        p1.Profile LIKE '<%'
13 SET
14     p2.Profile = 'Others'
15 ;
141 19:44:26 UPDATE social_media.profile as p2 JOIN social_media.view_tmp_Profile p1 ON p1.PrfID = p2.PrfID AND p1.PlatformID = p2.PlatformID AND p1.Profile LIKE '<%' SET p...
```

Check the record post update...

```
4 • Select * from social_media.profile where Profile like '<%';
5
```

Result Grid		Filter Rows:	Export:	Wrap Cell Content:
PrfID	Profile	PlatformID		

Before...

30	<U+BC29><U+D0C4><U+C18C><U+B144><U+B2E8>
----	--

After...

30	Others
31	Miley Cyrus
32	Akshay Kumar

*****END*****