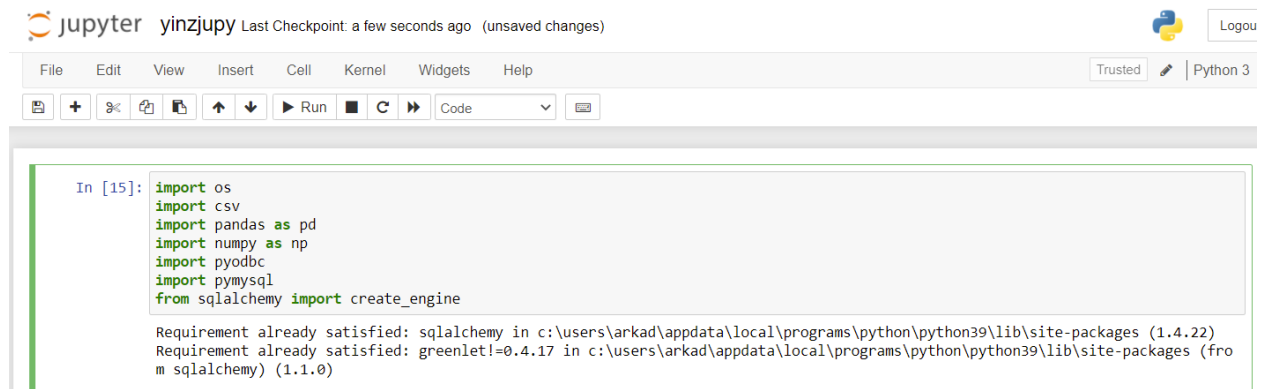


DATA ENGINEERING ASSESMENT

DATA INGESTION –

Libraries



The image shows a JupyterLab interface with a code editor. The code imports several libraries: os, csv, pandas as pd, numpy as np, pyodbc, pymysql, and create_engine from sqlalchemy. Below the code, a message indicates that the requirements are already satisfied for sqlalchemy (1.4.22) and greenlet (0.4.17).

```
In [15]: import os
import csv
import pandas as pd
import numpy as np
import pyodbc
import pymysql
from sqlalchemy import create_engine
```

Requirement already satisfied: sqlalchemy in c:\users\arkad\appdata\local\programs\python\python39\lib\site-packages (1.4.22)
Requirement already satisfied: greenlet!=0.4.17 in c:\users\arkad\appdata\local\programs\python\python39\lib\site-packages (from sqlalchemy) (1.1.0)

Reading TSV File



The image shows a JupyterLab interface with a code editor. The code opens a TSV file, reads it using csv.reader, and then creates a pandas DataFrame. The DataFrame columns are listed, and the final DataFrame is assigned to final_df.

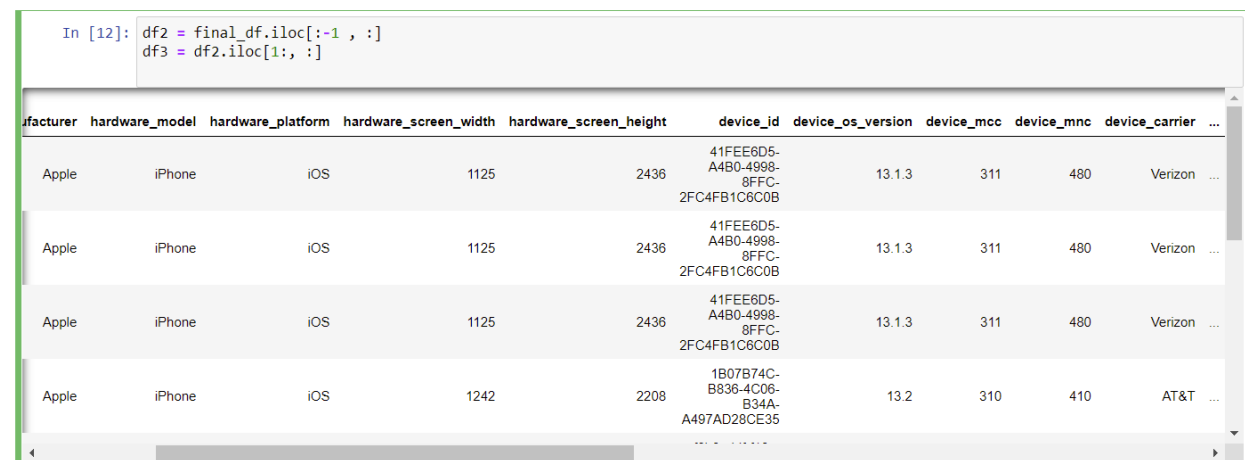
```
tsv_file = open("C:/Users//arkad//OneDrive//Desktop//Assignments//sample_dataset_with_header.tsv")
read_tsv = csv.reader(tsv_file, delimiter="\t", )
read_tsv

<_csv.reader at 0x1d6edc82760>

df = pd.DataFrame(read_tsv)
print(df.head())
df.columns=['hardware_manufacturer', 'hardware_model', 'hardware_platform', 'hardware_screen_width', 'hardware_screen_height',
'device_id', 'device_os_version', 'device_mcc', 'device_mnc', 'device_carrier', 'session_id', 'session_start_date_time',
'session_end_date_time', 'action_mysql_id', 'action_type_major', 'action_type_minor', 'action_resource_major',
'action_resource_minor', 'action_request_time', 'action_invisible_time', 'action_sort_order', 'geopip_country_code', 'geopip_time_zone']

final_df = df
```

Cleaning



The image shows a JupyterLab interface with a code editor. The code filters the DataFrame to keep only the last row and the first row. Below the code, a preview of the DataFrame is shown, displaying columns like hardware_manufacturer, hardware_model, hardware_platform, hardware_screen_width, hardware_screen_height, device_id, device_os_version, device_mcc, device_mnc, and device_carrier.

```
In [12]: df2 = final_df.iloc[:-1, :]
df3 = df2.iloc[1:, :]
```

hardware_manufacturer	hardware_model	hardware_platform	hardware_screen_width	hardware_screen_height	device_id	device_os_version	device_mcc	device_mnc	device_carrier	...
Apple	iPhone	iOS	1125	2436	41FEE6D5-A4B0-4998-8FFC-2FC4FB1C6C0B	13.1.3	311	480	Verizon	...
Apple	iPhone	iOS	1125	2436	41FEE6D5-A4B0-4998-8FFC-2FC4FB1C6C0B	13.1.3	311	480	Verizon	...
Apple	iPhone	iOS	1125	2436	41FEE6D5-A4B0-4998-8FFC-2FC4FB1C6C0B	13.1.3	311	480	Verizon	...
Apple	iPhone	iOS	1242	2208	1B07B74C-B836-4C06-B34A-A497AD28CE35	13.2	310	410	AT&T	...

Inserting into MySQL Database

```
In [17]: engine = create_engine("mysql+pymysql://{user}:{pw}@localhost/{db}"
                                .format(user="root",
                                        pw="Abcd1234",
                                        db="yinzcam_test"))

df3.to_sql('yinzdata4', con = engine, if_exists = 'append')
```

MySQL Database

The screenshot shows the MySQL Workbench interface. The 'Schemas' pane on the left lists databases including 'yinzcam_test'. The 'SQL Files 4*' editor shows a query: `USE yinzcam_test;` followed by `SELECT * FROM yinzdata4;`. The 'Result Grid' displays 15 rows of data with columns: index, hardware_manufacturer, hardware_model, hardware_platform, hardware_screen_width, hardware_screen_height, device_id, device_os_version, device_mcc, and device_n. The 'Table: yinzdata' pane shows columns: yid, hardware_manufacturer, hardware_model, hardware_platform, hardware_screen_width, hardware_screen_height, device_id, device_os_version, device_mcc, and device_n. The 'Output' pane shows a list of actions and their results, including 'Error loading schema content', 'DROP TABLE 'yinzcam_test.yinzdata3'', 'DROP TABLE 'yinzcam_test.yinzdata'', 'DROP TABLE 'yinzcam_test.yinzdata'', 'USE yinzcam_test', and 'SELECT * FROM yinzdata4'.

index	hardware_manufacturer	hardware_model	hardware_platform	hardware_screen_width	hardware_screen_height	device_id	device_os_version	device_mcc	device_n
1	Apple	iPhone	iOS	1125	2436	41FEE6D5-A4B0-4998-BFFC-2FC4B1C6C0B	13.1.3	311	480
2	Apple	iPhone	iOS	1125	2436	41FEE6D5-A4B0-4998-BFFC-2FC4B1C6C0B	13.1.3	311	480
3	Apple	iPhone	iOS	1125	2436	41FEE6D5-A4B0-4998-BFFC-2FC4B1C6C0B	13.1.3	311	480
4	Apple	iPhone	iOS	1242	2208	B807B74C-8836-4C06-B34A-A497AD28CE35	13.2	310	410
5	Samsung	SM-G930P	Android	1080	1920	8b2eddf-f16c-4f9a-8156-c100be491c8	8.0.0	310	655
6	Samsung	SM-G930P	Android	1080	1920	8b2eddf-f16c-4f9a-8156-c100be491c8	8.0.0	310	655
7	Lge	LG-V410	Android	800	1219	de840045-7ae4-401f-84ce-3ed18e48b6a	5.0.2	310	410
8	Lge	LM-Q720	Android	1080	2034	3537c85e-8db4-4a66-ae4-d652baf8f5ee	9	311	870
9	Lge	LM-Q720	Android	1080	2034	3537c85e-8db4-4a66-ae4-d652baf8f5ee	9	311	870
10	Lge	LM-Q720	Android	1080	2034	3537c85e-8db4-4a66-ae4-d652baf8f5ee	9	311	870
11	Samsung	SM-G935T	Android	1080	1920	b20919c3-6364-43df-bb08-e1db145d0a7d	8.0.0	310	260
12	Apple	iPhone	iOS	1125	2436	61e93852-2890-4e83-8f9a-CSAAC4C98D0E	13.1.3	310	410
13	Apple	iPhone	iOS	1125	2436	61e93852-2890-4e83-8f9a-CSAAC4C98D0E	13.1.3	310	410
14	Apple	iPhone	iOS	750	1334	9f9f7125-0A89-48E2-4A47-EA7737FE208	13.1.3	310	150
15	Apple	iPhone	iOS	878	1767	F7ABF678F17001441F80B7177F777790	11.1	111	480

SQL Queries –

1.

```
5 • SELECT action_type_major, count(*) as Count
6 FROM yinzdata4
7 GROUP BY action_type_major;
```

The screenshot shows the 'Result Grid' in MySQL Workbench. It displays the results of the SQL query: `SELECT action_type_major, count(*) as Count FROM yinzdata4 GROUP BY action_type_major;`. The table has two columns: 'action_type_major' and 'Count'.

action_type_major	Count
PUSH_VIEW	45511
PUSH_CLICK	43947
V	1175198
AD_BAN_IMP	547409
AD_BAN_CLICK	5400

2.

```
6 • SELECT
7   DATEDIFF(MAX(session_end_date_time),
8           MIN(session_start_date_time)) AS DateDiff,
9   TIMEDIFF(MAX(session_end_date_time),
10          MIN(session_start_date_time)) AS TimeDiff
11 FROM
12   yinzdata4;
```

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

DateDiff	TimeDiff
25	601:01:32

3.

```
12 • SELECT DATE(session_start_date_time) as Day, AVG(c) as Average, MIN(c) as Minimum, MAX(c) as Maximum
13 FROM(
14   SELECT session_start_date_time, count(action_type_major) as c
15   FROM yinzdata4
16   WHERE session_start_date_time BETWEEN CONCAT(DATE(session_start_date_time), ' 00:00:00') AND CONCAT(DATE(session_start_date_time), ' 23:59:59')
17   GROUP BY session_start_date_time
18 ) a
19 GROUP BY DATE(session_start_date_time);
```

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

Day	Average	Minimum	Maximum
2019-07-12	4.4183	1	110
2019-07-11	4.3182	1	107
2019-07-08	6.6485	1	183
2019-07-10	4.7548	1	119
2019-07-09	4.4631	1	58
2019-07-05	4.2556	1	56
2019-07-04	4.2683	1	73
2019-07-02	2.6545	1	23
2019-07-06	4.4060	1	105
2019-07-01	2.8000	2	13
2019-07-03	4.4703	1	110
2019-06-30	2.0000	2	2

Result 1 x

4.

Limit to 1000 rows

```

4  /* What is the average dwell time per page view? */
5
6  • SELECT action_resource_major, AVG(TIMEDIFF(action_invisible_time,action_request_time)) as AverageTimeInSeconds FROM yinzdata4
7  WHERE action_invisible_time != '1970-01-01 00:00:00' AND action_type_major = 'V'
8  GROUP BY action_resource_major;

```

Result Grid

action_resource_major	AverageTimeInSeconds
HOME	37.0969
GT_BOX	39.1231
GT_CONTAINER	51.4409
VOD	674.5865
LOGIN	23.1109
Gameday	24.6812
FANS	24.4582
ONBOARDING	58.8067
TEAM	16.0779
NEWS	98.2710
INTERACTIVE_MAP	42.6252
GT_DRIVE	31.8978
WIDGET	73.9438
GT_PLAYER	39.5870
EVENTS_CALENDAR	38.7355
MPRNTA	10.0275

Result 1 x

5.

Limit to 1000 rows

```

1  /* What is the average number of page views per session and per device?*/
2
3  • SELECT session_id,device_id, AVG(c) FROM(
4  SELECT session_id,device_id,count(action_resource_major)c
5  FROM yinzdata4
6  WHERE action_invisible_time != '1970-01-01 00:00:00' AND action_type_major = 'V'
7  GROUP BY session_id,device_id)a
8  GROUP BY session_id,device_id;

```

Result Grid

session_id	device_id	AVG(c)
B795939B-1554-4061-8FE3-B5B91D457EBD	41FEE6D5-A4B0-4998-8FFC-2FC4FB1C6C0B	1.0000
1FDA8CB6-9091-47F7-8CA9-FCB1A4F6D207	f6b6eddf-f16c-4f9a-8156-c100cbe491c8	6.0000
A7D9DAD9-7927-438C-A83E-B8C64AC7154B	de840045-7ae4-401f-84ce-3ed18e48beea	2.0000
EBD97072-3634-44C3-8C1A-D0B02D4EDF80	3537c85e-8cb4-4a66-ae4-dd62baf8f6ee	1.0000
2E720869-E70D-4F67-93D5-04ABC46871ED	b20919c3-6364-43df-bbd8-e1cb145d0a7d	5.0000
C215C95C-06CE-4622-AE25-F22F2297B1C7	61E93852-2890-4E83-BF9A-C5AAC4C98D0E	1.0000
C0C4BC86-EF49-4AAF-9263-25A31D2D11FE	5FF87125-DA89-49E2-A4A7-EA7737F1E208	1.0000
5CF713B1-6021-49B2-9FD6-A33A1A635012	A2A315C8-642E-4F13-9032-2F790AD8AAFC	5.0000
082D3975-AE9F-4C84-95D6-495DA76D77C3	AD35B2EA-C79B-4C83-8CC8-9AE6F1B7635E	1.0000
2A98CB2F-C598-47EC-B601-643226A27438	069A1154-865C-4A96-ACB6-0F3E680F3CC6	1.0000
A515B043-6371-46E0-A7C0-33A788702533	C8E81B69-C3BE-4E7E-89BA-C2E0974A5BE0	1.0000
1CF6802A-AA82-42A7-BB35-7E7C9FC52EBF	0a35c659-fb5b-4c77-ad1c-f2b51e7cc763	1.0000

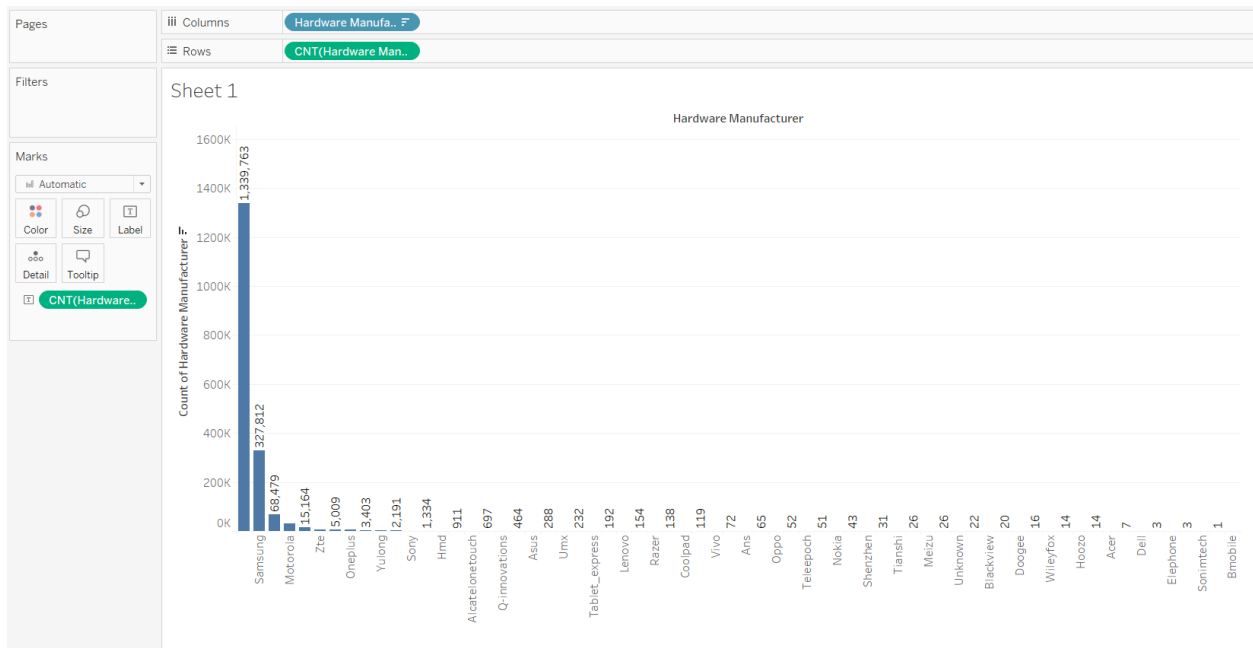
Result 1 x

DATA ANALYSIS AND VIZUALISATION –

Most used smartphone brand

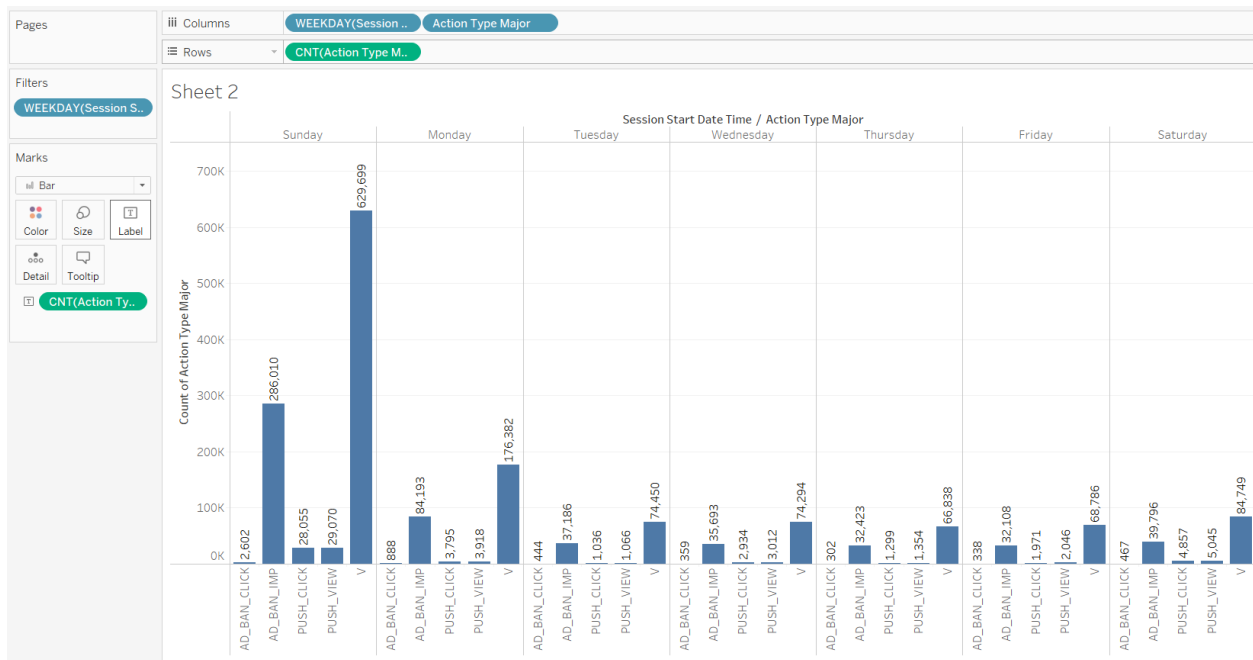
```
4 • SELECT hardware_manufacturer as Company, count(hardware_manufacturer) as Count
5 FROM yinzdata4
6 GROUP BY Company;
```

Company	Count
Apple	1339763
Samsung	327812
Lge	68479
Oneplus	4750
Google	15164
Umx	281
Huawei	5009
Motorola	31823
Htc	1334
Bullittgrouplimited	51
Td	3403
Asus	459
Blu	464



Views

We can see that Sundays have more views



What is the bounce rate or retention rate of features (pages)? Which features perform the best, and which perform the worst?

Are advertisements more effective on certain pages of the app than others? At certain times? Where/when should advertisements be placed to get maximum yield?

When does engagement with push notifications peak, and when is it at its lowest point?
When does push-notification engagement fall off after a game?

```

1 • SELECT DISTINCT(DATE(session_start_date_time)) FROM yinzdata4 ORDER BY DATE(session_start_date_time);
2
3 • SELECT device_id FROM yinzdata4 WHERE (DATE(session_start_date_time)) = 2019-07-18;
4
5 • SELECT session_start_date_time FROM yinzdata4 WHERE session_end_date_time = MAX(session_end_date_time);
6
7 • SELECT action_type_major FROM yinzdata4 WHERE date(session_start_date_time) = 2019-07-28;

```

```

4 • SELECT hardware_manufacturer as Company, count(hardware_manufacturer) as Count
5 FROM yinzdata4
6 GROUP BY Company;
7
8

```

Result Grid |   Filter Rows: | Export:  | Wrap Cell Content: 

	Company	Count
▶	Apple	1339763
	Samsung	327812
	Lge	68479
	Oneplus	4750
	Google	15164
	Umx	281
	Huawei	5009
	Motorola	31823

```

1 • SELECT COUNT(distinct device_id)
2 FROM yinzdata4
3 WHERE date(session_start_date_time) BETWEEN "2019-07-11" AND "2019-07-17"
4 AND
5 action_type_major = 'V';
6
7 • SELECT COUNT(distinct device_id)
8 FROM yinzdata4
9 WHERE date(session_start_date_time) BETWEEN "2019-07-04" AND "2019-07-10"
10 AND
11 action_type_major = 'V';
12

```

<

Result Grid |   Filter Rows: | Export:  | Wrap Cell Content: 

	COUNT(distinct device_id)
▶	57696


```

1  USE yinzcam_test;
2  ● SELECT * FROM yinzdata4;
3
4  ● SELECT action_type_major, count(DISTINCT action_type_major)
5  FROM yinzdata4
6  WHERE action_type_major = 'AD_BAN_CLICK';
7
8  ● SELECT action_resource_major, count(action_resource_major)
9  FROM yinzdata4
10 WHERE action_type_major = 'AD_BAN_CLICK'
11 GROUP BY action_resource_major
12 ORDER BY 2 DESC
13 LIMIT 5;

```

<

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch

	action_resource_major	count(action_resource_major)
▶	HOME	4730
	NEWS	371
	MEDIA	70
	INJ	69
	GALLERY	60

```

1  ● SELECT t as Date, dayname(t) as Day, c as Count FROM
2  (SELECT DATE(session_start_date_time)t, count(DATE(session_start_date_time))c
3  FROM yinzdata4
4  WHERE action_type_major = 'AD_BAN_CLICK'
5  GROUP BY DATE(session_start_date_time)
6  ORDER BY 2 DESC
7  LIMIT 5)x;

```

<

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch

	Date	Day	Count
	2019-07-14	Sunday	1304
	2019-07-07	Sunday	1298
▶	2019-07-08	Monday	495
	2019-07-15	Monday	393
	2019-07-13	Saturday	245

```

1 • SELECT t as Day, c as Count FROM
2 (SELECT DAYNAME(session_start_date_time)t, count(DISTINCT DAYNAME(session_start_date_time))c
3 FROM yinzdata4
4 WHERE action_type_major = 'AD_BAN_CLICK'
5 GROUP BY DAYNAME(session_start_date_time)
6 ORDER BY 2 DESC
7 LIMIT 5)x;
8
9 • SELECT MAX(session_end_date_time) FROM yinzdata4;
10
11 • SELECT action_type_major, session_start_date_time
12 from yinzdata4
13 WHERE session_start_date_time > '2019-07-22 00:00:00';

```

<

Result Grid   Filter Rows: | Export:  | Wrap Cell Content: 






action_type_major	session_start_date_time
-------------------	-------------------------

```

1 • SELECT action_resource_major, count(action_resource_major) from yinzdata4
2 WHERE DAYNAME(session_start_date_time) = 'Saturday' AND action_type_major = 'AD_BAN_CLICK'
3 GROUP BY action_resource_major
4 ORDER BY 2 DESC
5 LIMIT 5;

```

<

Result Grid   Filter Rows: | Export:  | Wrap Cell Content:  | Fetch rows: 

action_resource_major	count(action_resource_major)
HOME	413
NEWS	29
GALLERY	10
MEDIA	8
INJ	5