In [1]:
```
import re
import jovian
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud, STOPWORDS
import emoji
import demoji
from collections import Counter
```

In [2]:
```
def rawToDf(file, key):
    split_formats = {
        '12hr' : '\d{1,2}/\d{1,2}/\d{2,4},\s\d{1,2}:\d{2}\s[APap][mM]\s-\s',
        '24hr' : '\d{1,2}/\d{1,2}/\d{2,4},\s\d{1,2}:\d{2}\s-\s',
        'custom' : ''
    }
    datetime_formats = {
        '12hr' : '%m/%d/%y, %I:%M %p - ',
        '24hr' : '%m/%d/%y, %H:%M - ',
        'custom': ''
    }

    with open('WhatsAppChatwithSeniorsiyyipoyam.txt', 'r', encoding="utf8") as raw
        raw_string = ' '.join(raw_data.read().split('\n')) # converting the list sp
        user_msg = re.split(split_formats[key], raw_string) [1:] # splits at all th
        date_time = re.findall(split_formats[key], raw_string) # finds all the date

        df = pd.DataFrame({'date_time': date_time, 'user_msg': user_msg}) # export

    # converting date-time pattern which is of type String to type datetime,
    # format is to be specified for the whole string where the placeholders are ext
    df['date_time'] = pd.to_datetime(df['date_time'], format=datetime_formats[key]

    # split user and msg
    usernames = []
    msgs = []
    for i in df['user_msg']:
        a = re.split('([\w\W]+?):\s', i) # lazy pattern match to first {user_name}
        if(a[1:]): # user typed messages
            usernames.append(a[1])
            msgs.append(a[2])
        else: # other notifications in the group(eg: someone was added, some left
            usernames.append("grp_notif")
            msgs.append(a[0])

    # creating new columns
    df['user'] = usernames
    df['msg'] = msgs

    # dropping the old user_msg col.
    df.drop('user_msg', axis=1, inplace=True)

    return df
```

In [3]:
```
df = rawToDf('WhatsAppChatwithSeniorsiyyipoyam.txt', '12hr')
```

In [4]:
```
df.tail()
```

Out[4]:

	date_time	user	msg
39994	2023-05-29 09:59:00	grp_notif	Your security code with Leela Sai Gitam change...
39995	2023-06-16 13:10:00	grp_notif	Your security code with Leela Sai Gitam change...
39996	2023-06-20 20:04:00	grp_notif	Your security code with ~ PAVAN changed. Tap t...
39997	2023-06-21 12:27:00	grp_notif	Your security code with Leela Sai Gitam change...
39998	2023-06-21 15:03:00	grp_notif	Your security code with Issaku B19-17 changed....

In [5]:
```
df.shape
```

Out[5]:
```
(39999, 3)
```

Data Cleaning

In [6]:
```
media = df[df['msg']=="<Media omitted> "] #no. of images, images are represented by
media.shape
```

Out[6]:
```
(8508, 3)
```

In [7]:
```
df["user"].unique()
```

Out[7]:
```
array(['Mohith Medisetty😊 ', 'Radhakrishna B31', 'Issaku B19-17',
       'grp_notif', 'Gowtham B19', '+91 99630 84797', 'Vivek B12',
       'Ganesh Raju B19', '+91 94911 46169', '+91 99890 40722',
       'Devesh B19', 'Vishal B11', 'Pavan B12', 'RISHIK B31',
       '+91 70751 20218', 'Harish Bokka', '+91 6300 753 171', 'Eswar B11',
       'siddhu B12', 'Rohit B11', 'Nazeer Syed B12', 'Purna Chandra',
       '+91 93926 47949', 'Harsha B11', 'Gopichand B33', 'Karthik B19',
       'Vivek B11', 'Tarun B11', '+91 77802 81730', 'Leela Sai Gitam',
       'Sahasra B31', 'Prudhvi B11', '+91 6303 069 013',
       '+91 73860 59767', '+91 83092 39722', 'Avinash B11',
       '+91 70931 14342', '+91 98497 12312', 'Sai Chand B31',
       '+91 6302 872 024', 'Teja Chowdary B18', '+91 79896 12034',
       '+91 77805 83590', 'Chandu Narapinni', 'Hemanth Thumati',
       '+91 6303 651 428', 'Subash', 'Adharsh B11', '+91 70329 42422',
       'Shiva B18', 'Abhiram Gorela B11', '+91 83744 82158',
       'Praneeth B12', '+91 6305 554 597', 'Aditya B19',
       '+91 93468 85459', 'Bhargav B31', 'Tejus Varma B11',
       'Srikanth B11', '+91 90307 43589', 'Ankit Cr', 'Sai Prasana',
       '+91 6303 836 949', 'Pranav Sharma B12', '+91 83318 90905',
       '+91 95507 58987', '+91 79891 20147', '+91 86399 43935',
       '+91 93983 28616', 'Suresh B12', '+91 95538 07882',
       '+91 93900 84363', 'Harshidh B11', '+91 93470 09078'], dtype=object)
```

In [8]:
```
grp_notif = df[df['user']=="grp_notif"] #no. of grp notifications
grp_notif.shape
```

Out[8]:
```
(542, 3)
```

In [9]:
```
df.drop(media.index, inplace=True) #removing images
df.drop(grp_notif.index, inplace=True)
```

In [10]:
```
df.tail()
```

Out[10]:

	date_time	user	msg
39943	2022-12-24 13:34:00	+91 93926 47949	Kindly share this both messages in the class ...
39944	2022-12-24 13:44:00	+91 93926 47949	
39946	2023-01-08 08:07:00	+91 93926 47949	https://youtu.be/krGEHHy9zd8
39964	2023-03-15 17:22:00	+91 93926 47949	*_Greetings from Devtown_* ```We are glad to...
39975	2023-04-23 17:37:00	+91 93926 47949	Basic 7 days long *A to Z Learn Everything abo...

In [11]:
```
df.reset_index(inplace=True, drop=True)
df.shape
```

Out[11]:
```
(30949, 3)
```

Q. Who are the least active and most active persons in the group?

In [12]:
```
df.groupby("user")["msg"].count().sort_values(ascending=False)
```

Out[12]:
```
user
Harsha B11          4905
Eswar B11           4765
RISHIK B31          3107
+91 93926 47949     2213
Harish Bokka        1837
                    ...
+91 73860 59767        2
+91 93900 84363        1
+91 93983 28616        1
+91 95507 58987        1
+91 6303 651 428       1
Name: msg, Length: 73, dtype: int64
```

In [13]:
```
df
```

Out[13]:

	date_time	user	msg
0	2021-01-08 09:10:00	Mohith Medisetty😊	ha
1	2021-01-08 10:02:00	Radhakrishna B31	Happy Pongal ki matram perfect ga reply estharu
2	2021-01-08 10:03:00	Mohith Medisetty😊	avadu ichadu ra niku
3	2021-01-08 10:03:00	Mohith Medisetty😊	inka pongal ee ralle
4	2021-01-08 10:03:00	Radhakrishna B31	Sir ki ra
...	...	...	...
30944	2022-12-24 13:34:00	+91 93926 47949	Kindly share this both messages in the class ...
30945	2022-12-24 13:44:00	+91 93926 47949	
30946	2023-01-08 08:07:00	+91 93926 47949	https://youtu.be/krGEHHy9zd8
30947	2023-03-15 17:22:00	+91 93926 47949	*_Greetings from Devtown_* ```We are glad to...
30948	2023-04-23 17:37:00	+91 93926 47949	Basic 7 days long *A to Z Learn Everything abo...

30949 rows × 3 columns

Q.How many emojis I have used?

In [14]:
```python
# Download and update demoji library's emoji database
demoji.download_codes()

# Function to extract emojis from a text
def extract_emojis(text):
    emojis = demoji.findall(text)
    return list(emojis.keys())

# Extract emojis from messages
df['emojis'] = df['msg'].apply(extract_emojis)

# Count the number of emojis used by each participant
emoji_counts = df.explode('emojis').groupby('user')['emojis'].value_counts()

# Print the most used emojis by each participant
for user, emojis in emoji_counts.groupby(level=0):
    most_used_emoji = emojis.index[0][1]
    count = emojis[0]
    print(f"Most used emoji by {user}: {most_used_emoji} (Count: {count})")
```

```
C:\Users\gmass\AppData\Local\Temp\ipykernel_20064\3918567757.py:2: FutureWarning:
The demoji.download_codes attribute is deprecated and will be removed from demoji
in a future version. It is an unused attribute as emoji codes are now distributed
directly with the demoji package.
  demoji.download_codes()
```

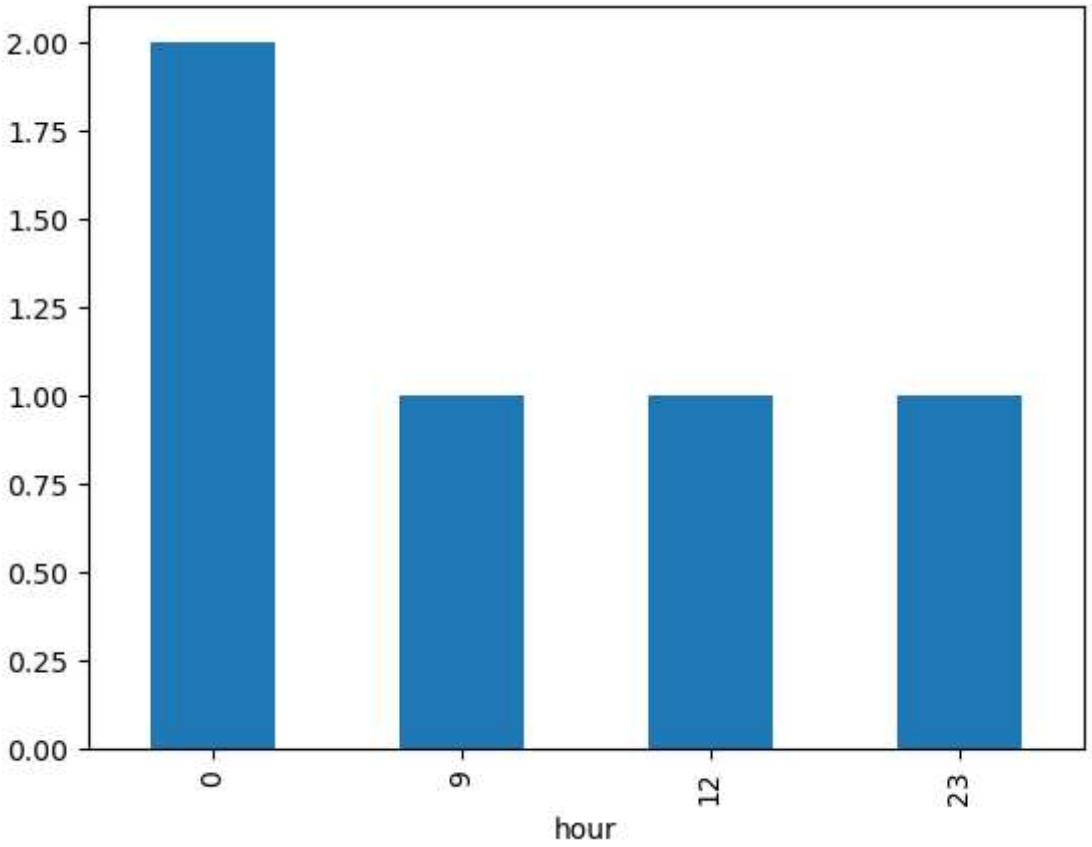
Most used emoji by +91 6300 753 171: 😂 (Count: 6)
Most used emoji by +91 6302 872 024: 😂 (Count: 5)
Most used emoji by +91 6303 069 013: 😂 (Count: 3)
Most used emoji by +91 6305 554 597: 🍔 (Count: 1)
Most used emoji by +91 70329 42422: 😂 (Count: 1)
Most used emoji by +91 70751 20218: 😂 (Count: 12)
Most used emoji by +91 70931 14342: ✨ (Count: 2)
Most used emoji by +91 77802 81730: 👍 (Count: 1)
Most used emoji by +91 77805 83590: 😂 (Count: 1)
Most used emoji by +91 79891 20147: 🍾 (Count: 1)
Most used emoji by +91 79896 12034: 😂 (Count: 5)
Most used emoji by +91 83092 39722: 🎉 (Count: 2)
Most used emoji by +91 83744 82158: 😂 (Count: 1)
Most used emoji by +91 86399 43935: 😳 (Count: 3)
Most used emoji by +91 93468 85459: 😂 (Count: 4)
Most used emoji by +91 93926 47949: 😂 (Count: 46)
Most used emoji by +91 94911 46169: 😂 (Count: 38)
Most used emoji by +91 95538 07882: 😅 (Count: 1)
Most used emoji by +91 98497 12312: 😁 (Count: 6)
Most used emoji by +91 99630 84797: 😂 (Count: 84)
Most used emoji by +91 99890 40722: 😂 (Count: 16)
Most used emoji by Abhiram Gorela B11: ☝️ (Count: 1)
Most used emoji by Adharsh B11: 😂 (Count: 59)
Most used emoji by Aditya B19: 😂 (Count: 12)
Most used emoji by Ankit Cr: 😂 (Count: 5)
Most used emoji by Avinash B11: 😂 (Count: 6)
Most used emoji by Bhargav B31: 😂 (Count: 2)
Most used emoji by Chandu Narapinni: 🍼 (Count: 1)
Most used emoji by Devesh B19: 😂 (Count: 36)
Most used emoji by Eswar B11: 😂 (Count: 227)
Most used emoji by Ganesh Raju B19: 😂 (Count: 253)
Most used emoji by Gopichand B33: 🎂 (Count: 1)
Most used emoji by Gowtham B19: 😂 (Count: 43)
Most used emoji by Harish Bokka: 🤣 (Count: 292)
Most used emoji by Harsha B11: 😂 (Count: 35)
Most used emoji by Hemanth Thumati: 🙄 (Count: 4)
Most used emoji by Issaku B19-17: 🎂 (Count: 7)
Most used emoji by Karthik B19: 😁 (Count: 14)
Most used emoji by Leela Sai Gitam: 😂 (Count: 2)
Most used emoji by Mohith Medisetty😁: 😂 (Count: 220)
Most used emoji by Nazeer Syed B12: 😂 (Count: 1)
Most used emoji by Pavan B12: 🎂 (Count: 1)
Most used emoji by Pranav Sharma B12: 😂 (Count: 3)
Most used emoji by Praneeth B12: ✋ (Count: 1)
Most used emoji by Prudhvi B11: 😂 (Count: 129)
Most used emoji by Purna Chandra: 😂 (Count: 90)
Most used emoji by RISHIK B31: 🤣 (Count: 39)
Most used emoji by Radhakrishna B31: 😂 (Count: 63)
Most used emoji by Rohit B11: 😂 (Count: 40)
Most used emoji by Sahasra B31: 😳 (Count: 3)
Most used emoji by Sai Chand B31: 🤣 (Count: 5)
Most used emoji by Shiva B18: 😂 (Count: 87)
Most used emoji by Srikanth B11: 😂 (Count: 5)
Most used emoji by Subash: 🔥 (Count: 7)
Most used emoji by Tarun B11: ❤️ (Count: 11)
Most used emoji by Teja Chowdary B18: 😂 (Count: 11)
Most used emoji by Tejus Varma B11: 😂 (Count: 5)
Most used emoji by Vishal B11: 🤣 (Count: 68)
Most used emoji by Vivek B11: 🥳 (Count: 5)
Most used emoji by Vivek B12: 😂 (Count: 231)
Most used emoji by 𝕤𝕚𝕕𝕕𝕙𝕦 B12: 😂 (Count: 2)

Q. What does my WhatsApp activity tell about my sleep cycle?

In [15]:
```
me=input()
df['hour'] = df['date_time'].apply(lambda x: x.hour)
df[df['user']==me].groupby(['hour']).size().sort_index().plot(x="hour", kind='bar'
```

```
siddhu B12
```

Out[15]:
```
<Axes: xlabel='hour'>
```



How many words do I type on average on weekday vs weekend?

In [16]:
```
df['weekday'] = df['date_time'].apply(lambda x: x.day_name())
```

In [17]:
```
df['is_weekend'] = df.weekday.isin(['Sunday', 'Saturday'])
```

In [18]:
```
msgs_per_user = df['user'].value_counts(sort=True)
msgs_per_user
```

Out[18]:
```
Harsha B11          4905
Eswar B11           4765
RISHIK B31          3107
+91 93926 47949     2213
Harish Bokka        1837
                    ...
+91 73860 59767        2
+91 93983 28616        1
+91 95507 58987        1
+91 93900 84363        1
+91 6303 651 428       1
Name: user, Length: 73, dtype: int64
```

In [19]:
```
top5_users = msgs_per_user.index.tolist()[:5]
top5_users
```

Out[19]:
```
['Harsha B11', 'Eswar B11', 'RISHIK B31', '+91 93926 47949', 'Harish Bokka']
```

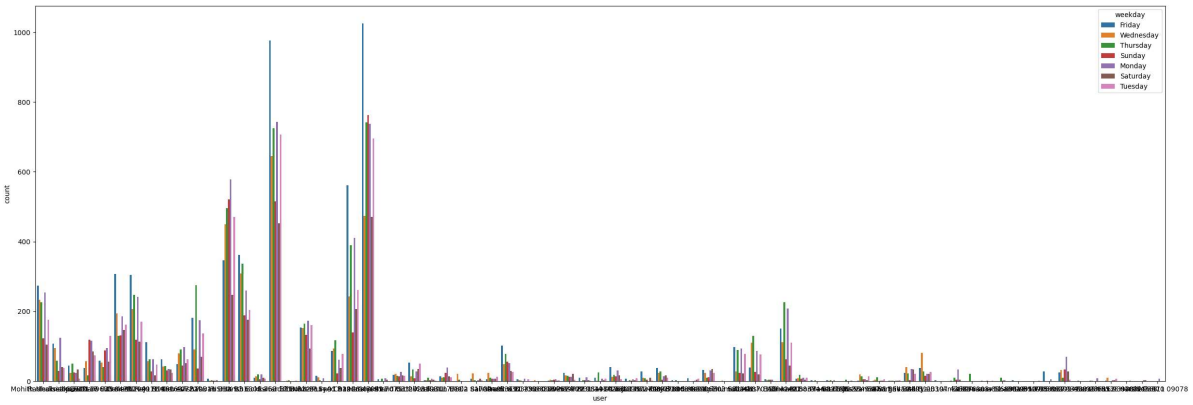
In [20]:
```
df_top5 = df.copy()
df_top5 = df_top5[df_top5.user.isin(top5_users)]
df_top5.head()
```

Out[20]:

	date_time	user	msg	emojis	hour	weekday	is_weekend
34	2021-01-17 19:20:00	RISHIK B31	Chusa	[]	19	Sunday	True
56	2021-01-18 11:10:00	Harish Bokka	haa	[]	11	Monday	False
69	2021-01-18 11:59:00	Eswar B11	Inka walk out chedama	[]	11	Monday	False
71	2021-01-18 12:01:00	Eswar B11	Kill kill kill ani mana moodu , intrest ani ki...	[🤦]	12	Monday	False
113	2021-01-20 23:47:00	Eswar B11	😊	[😊]	23	Wednesday	False

In [21]:
```
plt.figure(figsize=(30,10))
sns.countplot(x="user", hue="weekday", data=df)
```

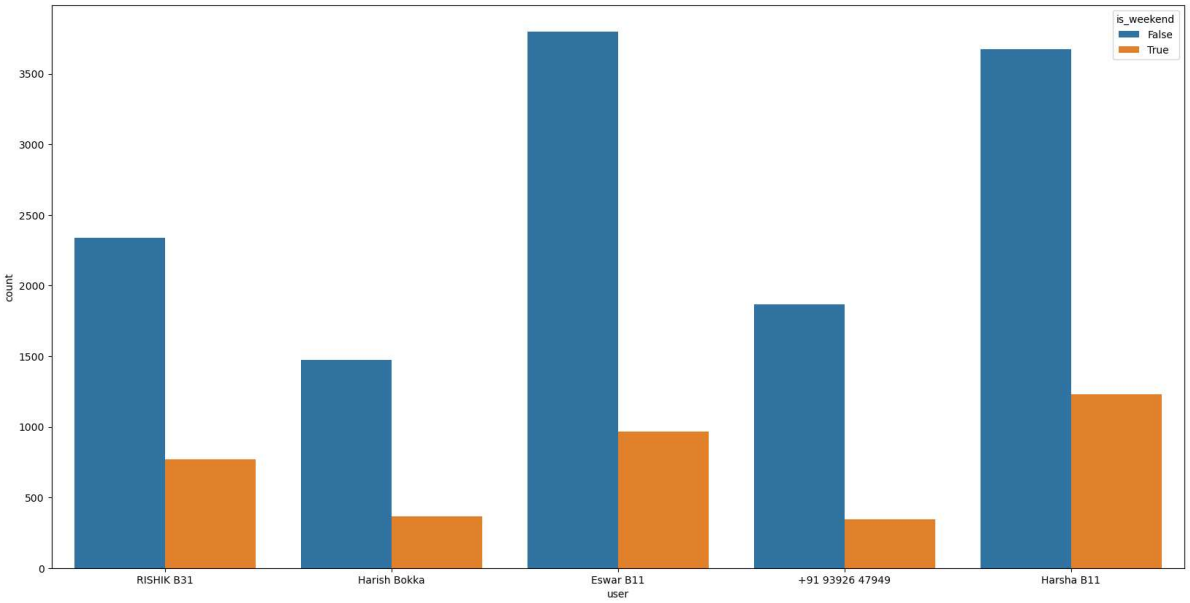
Out[21]:
```
<Axes: xlabel='user', ylabel='count'>
```



In [22]:
```
df_top5['is_weekend'] = df_top5.weekday.isin(['Sunday', 'Saturday'])
```

In [23]:
```
plt.figure(figsize=(20,10))
sns.countplot(x="user", hue="is_weekend", data=df_top5)
```

Out[23]:
```
<Axes: xlabel='user', ylabel='count'>
```



In [24]:
```
def word_count(val):
    return len(val.split())
```

In [25]:
```
def word_count(val):
    return len(val.split())
```

In [26]:
```
df_top5['no_of_words'] = df_top5['msg'].apply(word_count)
```

Q. At what time of day do I use WhatsApp most?

In [27]:
```
x = df.groupby(['hour', 'weekday'])['msg'].size().reset_index()
x2 = x.pivot("hour", 'weekday', 'msg')
x2.head()
```

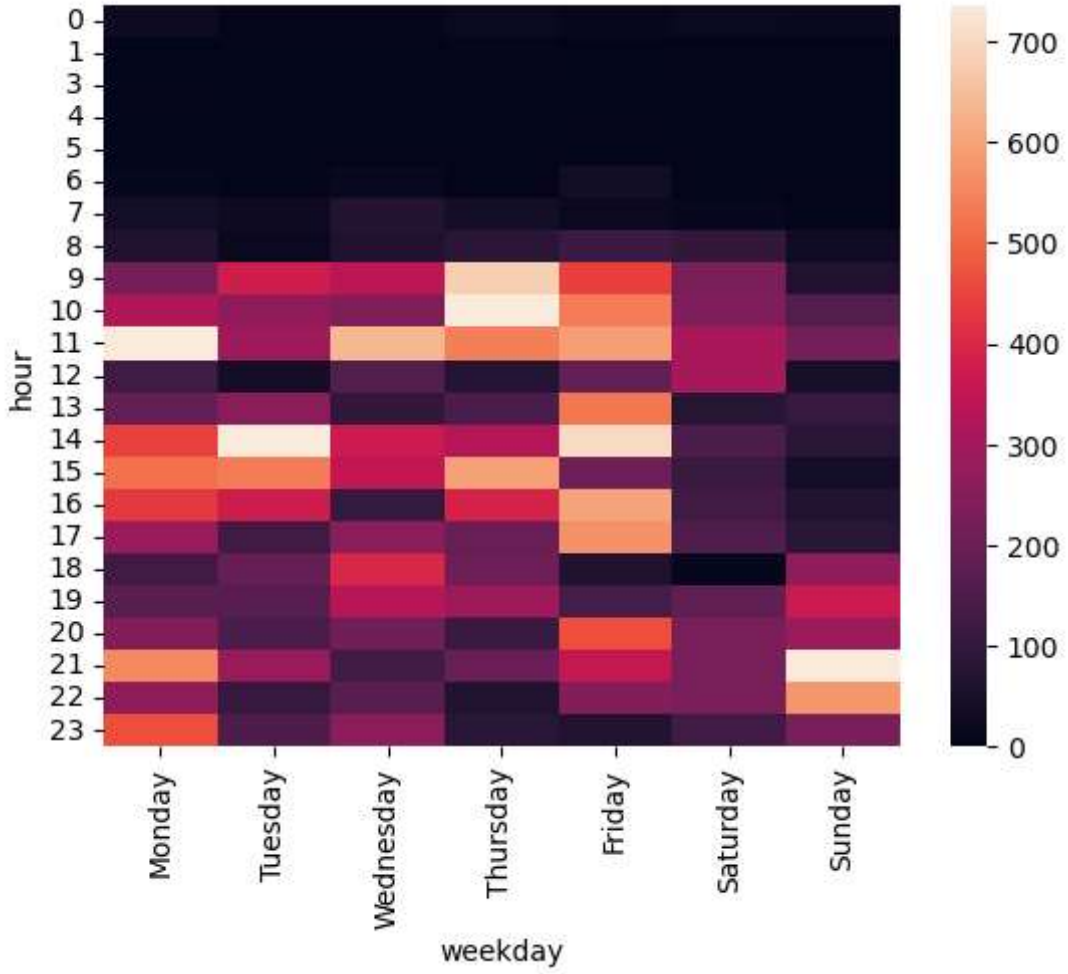
```
C:\Users\gmass\AppData\Local\Temp\ipykernel_20064\1368439187.py:2: FutureWarning:
In a future version of pandas all arguments of DataFrame.pivot will be keyword-onl
y.
  x2 = x.pivot("hour", 'weekday', 'msg')
```

Out[27]:

weekday	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
hour							
0	11.0	27.0	18.0	16.0	22.0	8.0	7.0
1	NaN	NaN	5.0	5.0	8.0	1.0	1.0
3	1.0	3.0	NaN	NaN	NaN	NaN	NaN
4	3.0	NaN	NaN	NaN	NaN	NaN	NaN
5	3.0	1.0	NaN	NaN	NaN	NaN	3.0

In [28]:
```
days = ["Monday", 'Tuesday', "Wednesday", "Thursday", "Friday", "Saturday", "Sunda
sns.heatmap(x2[days].fillna(0), robust=True)
```

Out[28]:
```
<Axes: xlabel='weekday', ylabel='hour'>
```



Let's know whom did I respond the most in the group?

In [29]:
```
my_msgs_index = np.array(df[df['user']==me].index)
print(my_msgs_index, my_msgs_index.shape)
```

```
[ 112  116  118  122 6002] (5,)
```

In [30]:
```
prev_msgs_index = my_msgs_index - 1
print(prev_msgs_index, prev_msgs_index.shape)
```

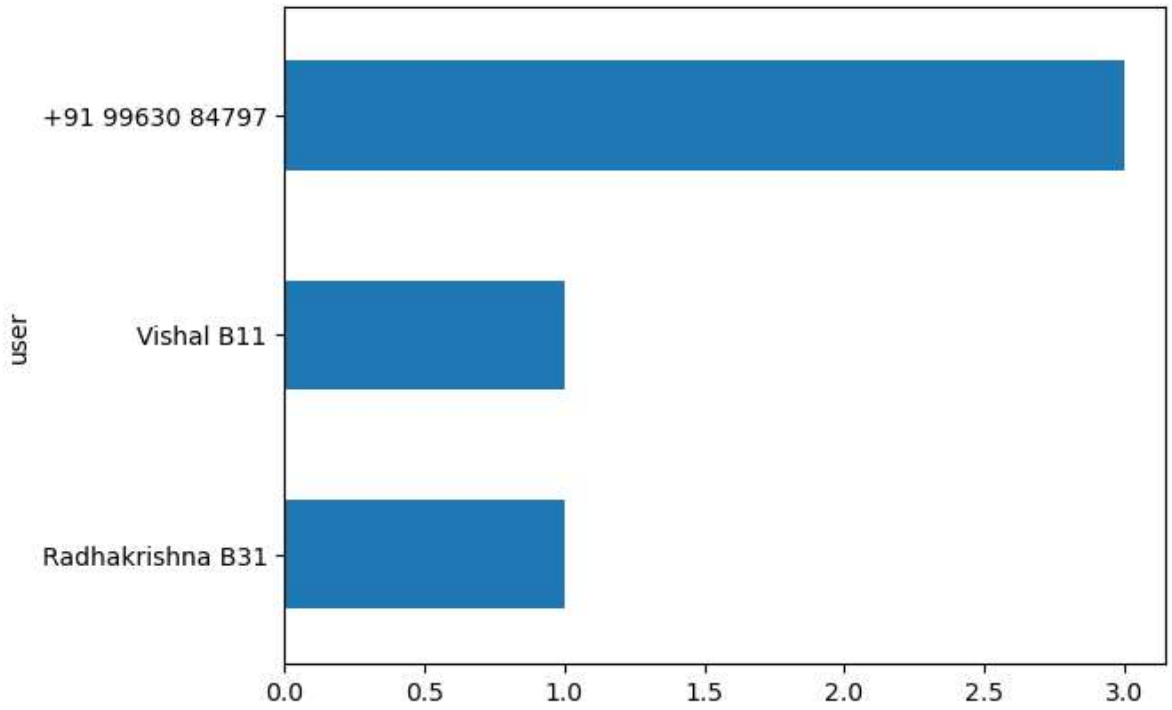
```
[ 111  115  117  121 6001] (5,)
```

In [31]:
```
df_replies = df.iloc[prev_msgs_index].copy()
df_replies.shape
```

Out[31]:
```
(5, 7)
```

In [32]:
```
df_replies.groupby(["user"])["msg"].size().sort_values().plot(kind='barh')
```

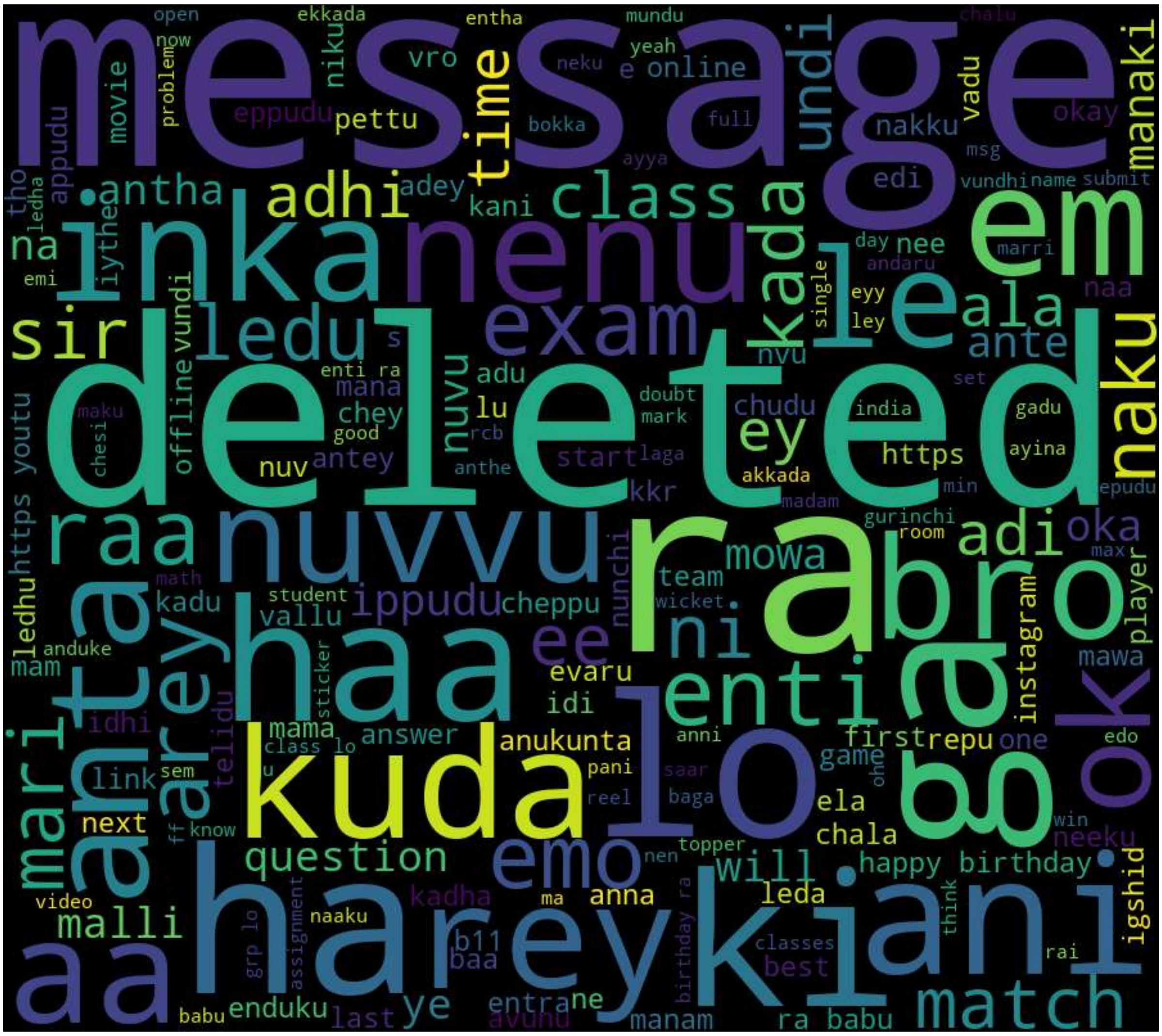
Out[32]:
```
<Axes: ylabel='user'>
```



In [33]:
```
comment_words = ' '
# stopwords = STOPWORDS.update([])

for val in df.msg.values:
    val = str(val)
    tokens = val.split()

    for i in range(len(tokens)):
        tokens[i] = tokens[i].lower()

    for words in tokens:
        comment_words = comment_words + words + ' '


wordcloud = WordCloud(width = 900, height = 800,
                background_color ='black',
#                 stopwords = stopwords,
                min_font_size = 10).generate(comment_words)
```

In [34]:
```
wordcloud.to_image()
```

Out[34]:



message
deleted
ra
inka
nenu
em
haa
hareykiani
kuda
lo
aa
anta
arey
nuvvu
raa
bro
ok
enti
ledu
exam
sir
kada
ala
ante
naku
undi
class
adhi
antha
time
manaki
mari
emo
question
malli
match
ippudu
cheppu
mowa
oka
adi
ee
ni
ey
https
chudu
team
evaru
idi
anukunta
answer
mama
link
next
firstrepu
happy birthday
will
chala
ela
leda
anna
best
ye
enduku
last
manam
ra babu
entra
ne
b11
baa
igshid
instagram
player
mawa
nakku
edi
okay
vadu
nee
naa
eyy
ley
india
gadu
ayina
akkada
kkr
start
antey
chey
good
nuv
adu
lu
nuvu
mana
enti ra
vundi
offline
youtu
iythe
appudu
movie
problem
niku
vro
eppudu
pettu
bokka
full
neku
e online
yeah
mundu
entha
ekkada
open
now
adey
kani
tho
ledha
na
emi
ledhu
mam
anduke
kadu
vallu
student
telidu
idhi
sem
video
naaku
babu
enduku
nunchi
wicket
reel
baga
pani
saar
anni
topper
nen
ma
kadha
game
one
edo
neeku
win
classes
rai
think
vundhiname
submit
msg
ayya
day
andaru
marri
single
set
gurinchi
room
madam
min
epudu
max
doubt
mark
laga
anthe
rcb

In [ ]: