In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
df = pd.read_csv("facebook user data - facebook user data.csv")
df
```

Out[1]:

	userid	age	dob_day	dob_year	dob_month	gender	tenure	friend_count	friendships_ini
0	2094382	14	19	1999	11	male	266.0	0	
1	1192601	14	2	1999	11	female	6.0	0	
2	2083884	14	16	1999	11	male	13.0	0	
3	1203168	14	25	1999	12	female	93.0	0	
4	1733186	14	4	1999	12	male	82.0	0	
...	...	...	...	...	...	...	...	...	
98998	1268299	68	4	1945	4	female	541.0	2118	
98999	1256153	18	12	1995	3	female	21.0	1968	
99000	1195943	15	10	1998	5	female	111.0	2002	
99001	1468023	23	11	1990	4	female	416.0	2560	
99002	1397896	39	15	1974	5	female	397.0	2049	

99003 rows × 15 columns

In [2]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99003 entries, 0 to 99002
Data columns (total 15 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   userid                 99003 non-null  int64
 1   age                    99003 non-null  int64
 2   dob_day                99003 non-null  int64
 3   dob_year               99003 non-null  int64
 4   dob_month              99003 non-null  int64
 5   gender                 98828 non-null  object
 6   tenure                 99001 non-null  float64
 7   friend_count           99003 non-null  int64
 8   friendships_initiated  99003 non-null  int64
 9   likes                  99003 non-null  int64
 10  likes_received         99003 non-null  int64
 11  mobile_likes           99003 non-null  int64
 12  mobile_likes_received  99003 non-null  int64
 13  www_likes              99003 non-null  int64
 14  www_likes_received     99003 non-null  int64
dtypes: float64(1), int64(13), object(1)
memory usage: 11.3+ MB
```

In [3]:
```
df.describe()
```

Out[3]:

	userid	age	dob_day	dob_year	dob_month	tenure	friend_
count	9.900300e+04	99003.000000	99003.000000	99003.000000	99003.000000	99001.000000	99003.0

	userid	age	dob_day	dob_year	dob_month	tenure	friend_
mean	1.597045e+06	37.280224	14.530408	1975.719776	6.283365	537.887375	196.3
std	3.440592e+05	22.589748	9.015606	22.589748	3.529672	457.649874	387.3
min	1.000008e+06	13.000000	1.000000	1900.000000	1.000000	0.000000	0.0
25%	1.298806e+06	20.000000	7.000000	1963.000000	3.000000	226.000000	31.0
50%	1.596148e+06	28.000000	14.000000	1985.000000	6.000000	412.000000	82.0
75%	1.895744e+06	50.000000	22.000000	1993.000000	9.000000	675.000000	206.0
max	2.193542e+06	113.000000	31.000000	2000.000000	12.000000	3139.000000	4923.0

In [4]:
```
df.shape #How Many rows and column in the data
```

Out[4]: (99003, 15)

Question 1

Load the data and impute missing values

Imputation of missing values:

In [5]:
```
df2 = df.copy() #Make a duplicate data so that our original data remain uneffected
df2
```

Out[5]:

	userid	age	dob_day	dob_year	dob_month	gender	tenure	friend_count	friendships_ini
0	2094382	14	19	1999	11	male	266.0	0	
1	1192601	14	2	1999	11	female	6.0	0	
2	2083884	14	16	1999	11	male	13.0	0	
3	1203168	14	25	1999	12	female	93.0	0	
4	1733186	14	4	1999	12	male	82.0	0	
...	...	...	...	...	...	...	...	...	
98998	1268299	68	4	1945	4	female	541.0	2118	
98999	1256153	18	12	1995	3	female	21.0	1968	
99000	1195943	15	10	1998	5	female	111.0	2002	
99001	1468023	23	11	1990	4	female	416.0	2560	
99002	1397896	39	15	1974	5	female	397.0	2049	

99003 rows × 15 columns

In [6]:
```
df2.isnull().sum()
```

```
userid                   0
```

Out[6]:
```
age                      0
dob_day                  0
dob_year                 0
dob_month                0
gender                 175
tenure                   2
friend_count             0
friendships_initiated    0
likes                    0
likes_received           0
mobile_likes             0
mobile_likes_received    0
www_likes                0
www_likes_received       0
dtype: int64
```

Clearly from above result there is only two columns 'gender' and 'tenure' which have 175 and 2 null values respectively

In [7]:
```
null_gender = pd.isnull(df2["gender"]) # Check that which rows containing null value
df2[null_gender]
```

Out[7]:

	userid	age	dob_day	dob_year	dob_month	gender	tenure	friend_count	friendships_ini
3224	1781137	25	3	1988	1	NaN	203.0	1	
4123	1110111	31	12	1982	1	NaN	227.0	2	
5920	2000185	25	15	1988	2	NaN	239.0	4	
6432	1137054	59	18	1954	2	NaN	2076.0	4	
6870	1783336	62	5	1951	9	NaN	2047.0	5	
...	...	...	...	...	...	...	...	...	
94778	1748557	68	27	1945	4	NaN	1862.0	790	
95317	2173780	50	11	1963	9	NaN	2129.0	863	
95479	1442490	108	1	1905	7	NaN	1332.0	967	
97530	2119521	74	26	1939	9	NaN	1998.0	1609	
98216	1966857	102	6	1911	12	NaN	2389.0	2548	

175 rows × 15 columns

The above result shows the only rows which have null values in 'gender' column

In [8]:
```
null_tenure = pd.isnull(df2['tenure']) #check that which rows are conaining null val
df2[null_tenure]
```

Out[8]:

	userid	age	dob_day	dob_year	dob_month	gender	tenure	friend_count	friendships_ini
35057	1414063	88	14	1925	12	female	NaN	48	
63979	2095829	103	1	1910	1	female	NaN	137	

The above result shows the only rows which have null values in 'tenure' column

Question 1(a)

Replace the null values (NA) of gender column with its mode or median and explain

why mode/median used to replace NA values

In [9]:
```
df2['gender'] = df2['gender'].fillna(df2['gender'].mode()[0]) #Filling null values i

df2.isnull().sum() #check that null values in gender column has been removed or not
```

Out[9]:
```
userid                   0
age                      0
dob_day                  0
dob_year                 0
dob_month                0
gender                   0
tenure                   2
friend_count             0
friendships_initiated    0
likes                    0
likes_received           0
mobile_likes             0
mobile_likes_received    0
www_likes                0
www_likes_received       0
dtype: int64
```

Clearly we can see from above result that there is no null value in the gender column now

We use mean, median, mode values for filing in place of null values because:- • When we train our model then it shows error on null values • model acccuracy does not affected • if we drop the rows containing null values then we can loose a useful information during training of our model

Question 1(b)

Replace the null values (NA) of tenure column (numerical variable) with its median,

and explain why mode/median used to replace NA values

In [10]:
```
df2['tenure'] = df2['tenure'].fillna(df2['tenure'].median()) #Filling null values in
df2.isnull().sum() #Checking if null values in tenure column has been removed or not
```

Out[10]:
```
userid                   0
age                      0
dob_day                  0
dob_year                 0
dob_month                0
gender                   0
tenure                   0
friend_count             0
friendships_initiated    0
```

```
likes                   0
likes_received          0
mobile_likes            0
mobile_likes_received   0
www_likes               0
www_likes_received      0
dtype: int64
```

we can clearly see from above result that no null value is present in any column now

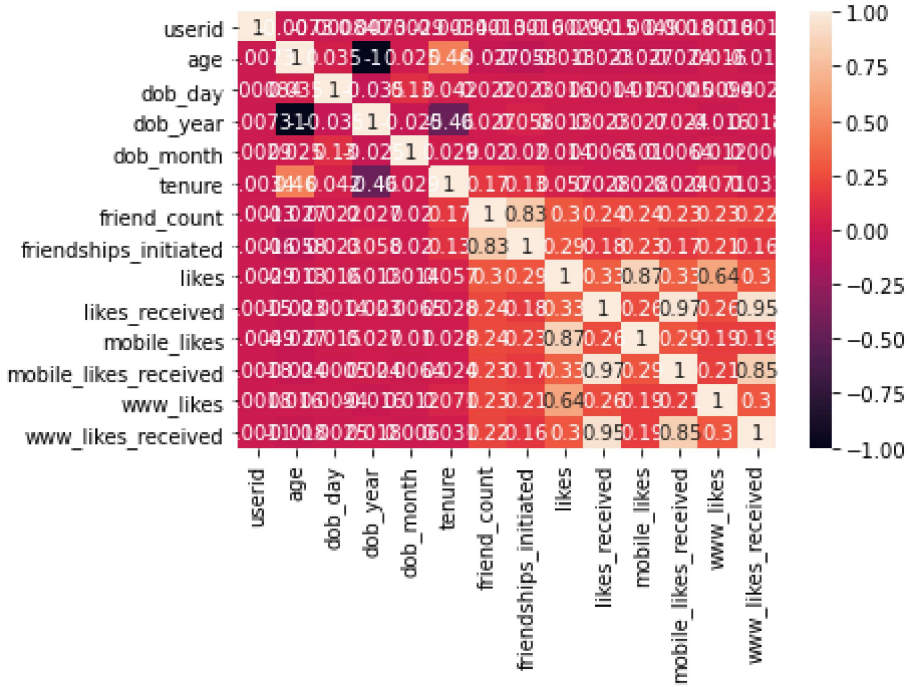
We use mean, median, mode values for filing in place of null values because:- • When we train our model then it shows error on null values • model acccuracy does not affected • if we drop the rows containing null values then we can loose a useful information during training of our model

Question 2

Plot heatmap / correlation matrix on all the columns.

In [11]:
```
sns.heatmap(df2.corr(), annot=True)
```

Out[11]: <AxesSubplot:>



Question 3

Analysis based on gender of the users

Question 3(a)

What is composition of male and female users?

In [12]:
```
df2['gender'].value_counts()
```

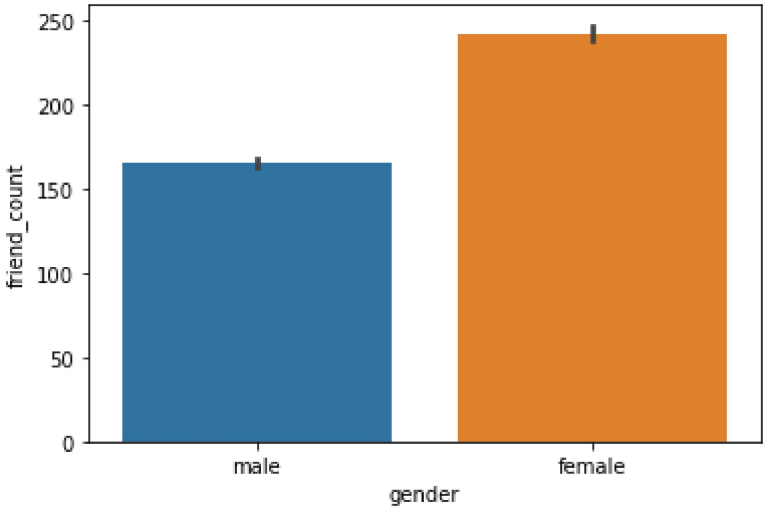
Out[12]:
```
male      58749
female    40254
Name: gender, dtype: int64
```

Clearly there are 56749 male and 40254 female in the gender column

Question 3(b)

Which category of gender has more friends?

In [13]:
```
sns.barplot(data = df2,x = 'gender',y = 'friend_count')
plt.show()
```

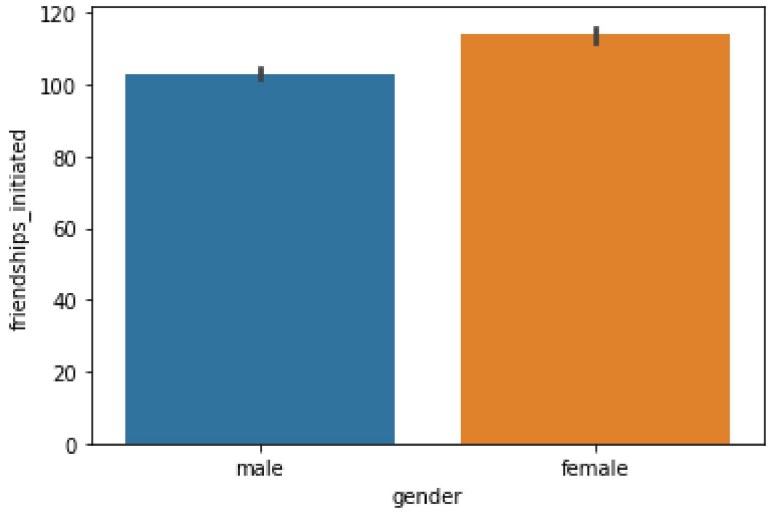


Clearly from the above graph Female have more friends than Male

Question 3(c)

Which category of gender initiated more friendships?

In [14]:
```
sns.barplot(data = df2,x = 'gender',y = 'friendships_initiated')
plt.show()
```



Clearly from the above graph female initiated more friends than Male

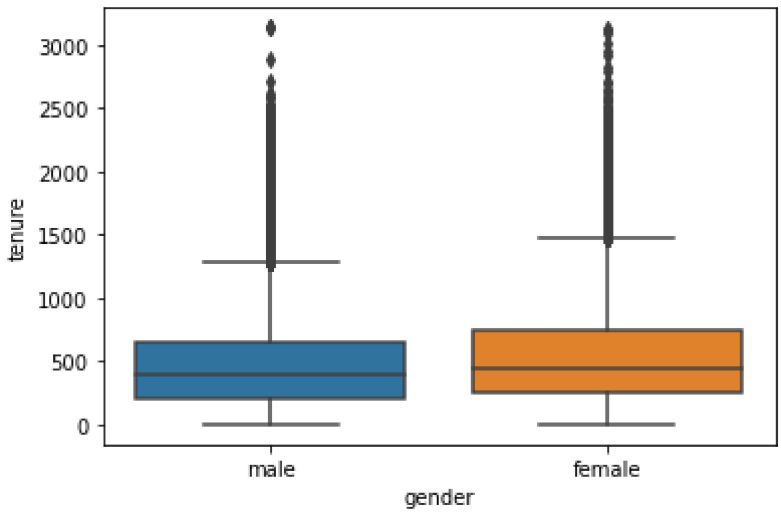
Question 3(d)

What is the distribution of tenure across different categories of gender?

In [15]:
```
sns.boxplot(df2.gender,df2.tenure)
```

```
C:\Users\roxma\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning:
Pass the following variables as keyword args: x, y. From version 0.12, the only vali
d positional argument will be `data`, and passing other arguments without an explici
t keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[15]:
```
<AxesSubplot:xlabel='gender', ylabel='tenure'>
```



The above graph shows the distribution of tenure over categories of gender

Question 4

Analysis based on the least active users on Facebook

Question 4(a)

How many users have no friends?

In [16]:
```
df2.friend_count.isin([0]).sum()
```

Out[16]: 1962

Clearly from above result there are 1962 user who have no friends

Question 4(b)

How many users did not like any posts?

In [17]:
```
df2.likes.isin([0]).sum()
```

Out[17]: 22308

Clearly from the above result there are 22308 user who did not like any post

Question 4(c)

How many users did not receive any likes?

In [18]:
```
df2.likes_received.isin([0]).sum()
```

Out[18]: 24428

Clearly from above result there are 24428 user who did not received any like

Question 5

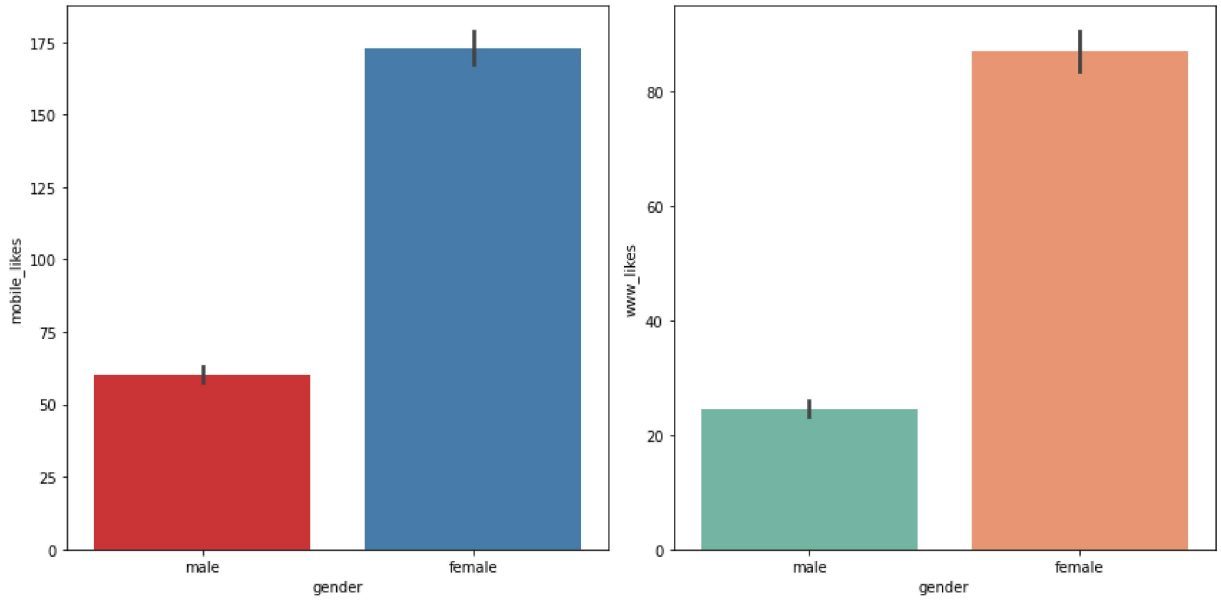
Analysis based on the user accessibility (Mobile Devices vs. Web Devices)

Question 5(a)

What is the average number of posts liked by users (based on gender) through web vs.

mobile devices?

In [19]:
```
plt.figure(figsize=(12,6))
plt.subplot(1,2,1)
sns.barplot(x='gender',y='mobile_likes',data=df2, palette='Set1')
plt.subplot(1,2,2)
sns.barplot(x='gender',y='www_likes',data=df2, palette='Set2')
plt.tight_layout()
```



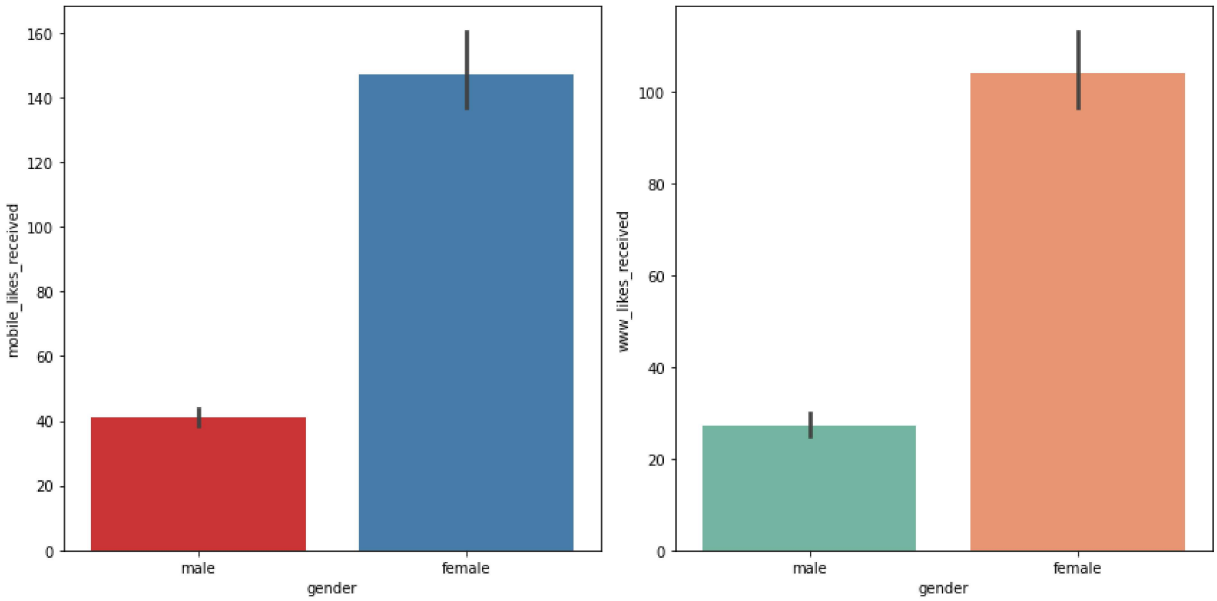
From the above graph it is clear that average no of posts liked by male and female is 60, 170 repectively from mobile and from web is 25, 90 repectively.

Question 5(b)

What is the average number of likes received by users (based on gender) through web

vs. mobile devices?

In [20]:
```
plt.figure(figsize=(12,6))
plt.subplot(1,2,1)
sns.barplot(x='gender',y='mobile_likes_received',data=df2,palette='Set1')
plt.subplot(1,2,2)
sns.barplot(x='gender',y='www_likes_received',data=df2,palette='Set2')
plt.tight_layout()
```



from tha above graph we can say that average number of likes received by male and female is 40, 140 respectively from the mobile and from the web is 25, 105 repectively

Thank You