Exploratory Data Analysis

of

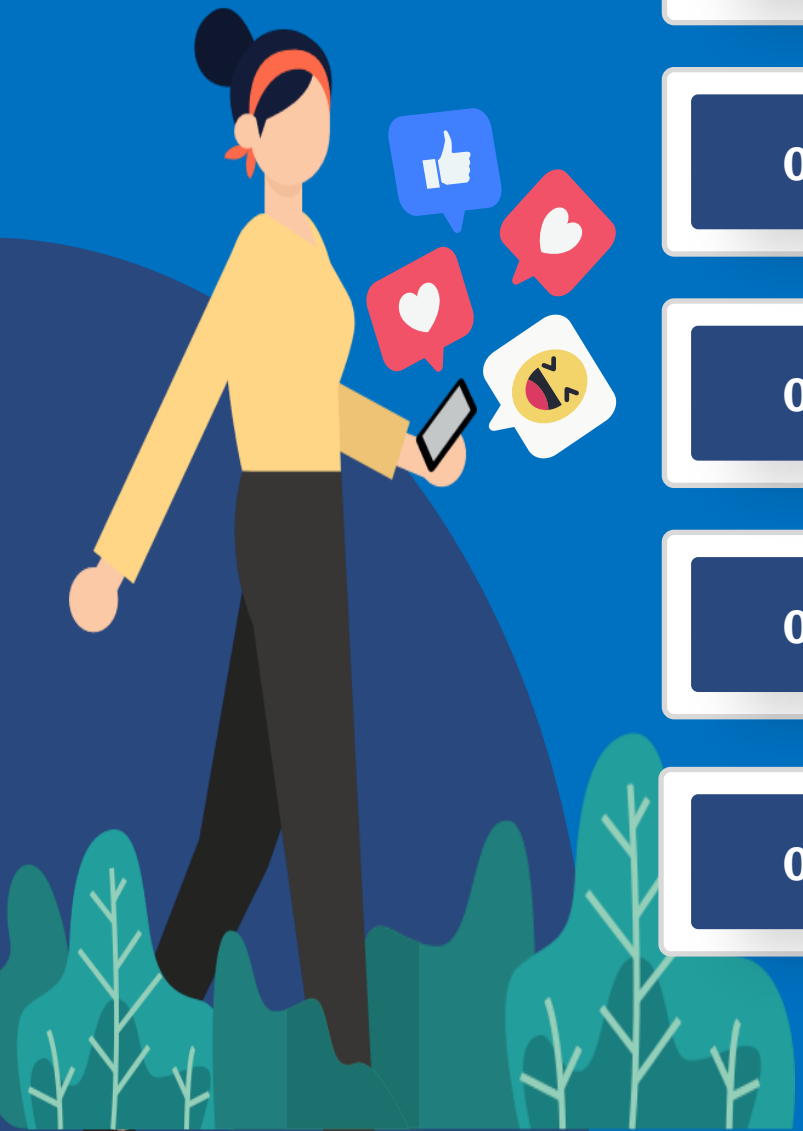
FACEBOOK DATASET

Submitted By

Aditi Goyal

(CDF Program)

Agenda



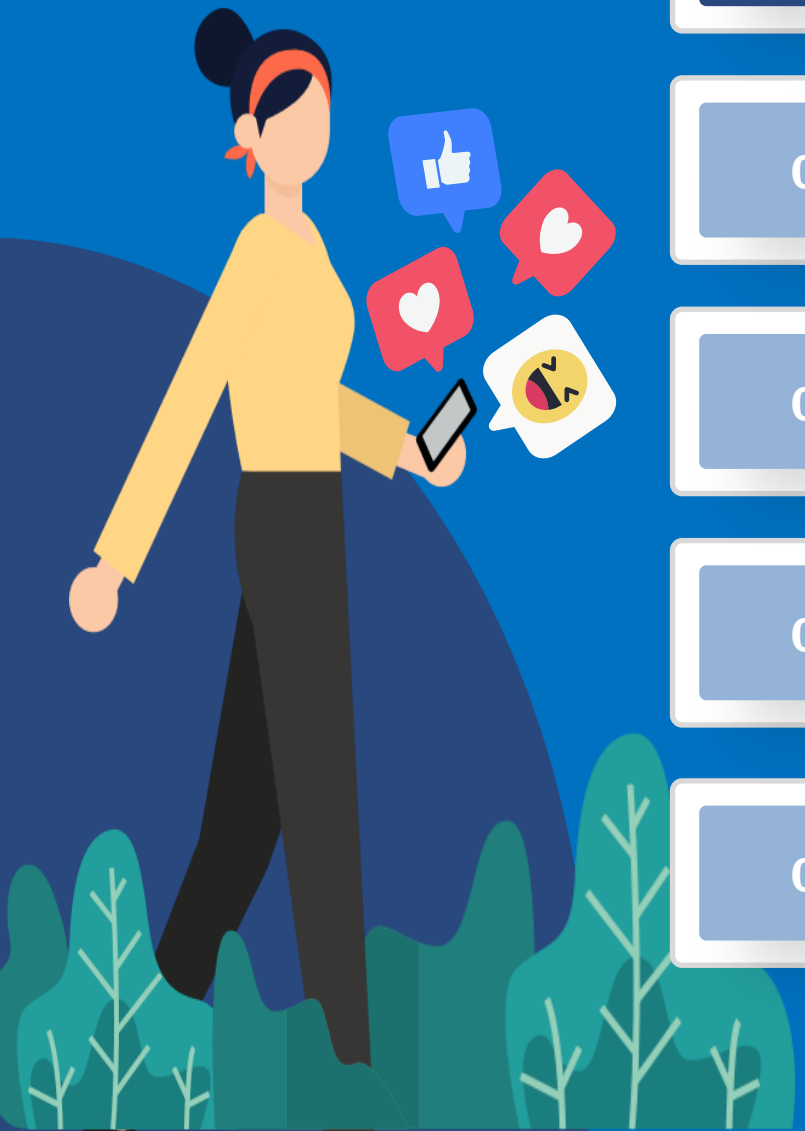
01 Problem Statement

02 Data Collection and Visualization

03 Analysis of Dataset

04 Summary

05 Reference



Analyze important parameters for using social media, FACEBOOK to derive insights on various trends such as age group of users, friend counts, platform and how these parameters are interlinked.

problem statement

- **Dataset Information**: The data set has 15 types of attributes
- There are certain differences in the way Facebook is used by different age group and gender
- These result will help the company to utilize the patterns in the next set of iteration development and improving their application and user experience

Attributes

The following attributes affect the Facebook usage and popularity

Column Name	Description
userid	A numeric value uniquely identifying the user
age	Age of the user in years
dob_year	Year part of the user's date of birth
gender	Gender of the user
tenure	Number of days since the user has been on FB
friend_count	Number of friends the user has
friendships_initiated	Number of friendships initiated by the user
likes	Total number of posts liked by the user
likes_received	Total Number of likes received by user's posts
mobile_likes	Number of posts liked by the user through mobile app
mobile_likes_received	Number of likes received by user through mobile app
www_likes	Number of posts liked by the user through web
www_likes_received	Number of likes received by user through web

Data Collection and Visualization

Data Acquisition

```
fb_data=pd.read_csv('https://github.com/insaid2018/Term-1/raw/master/Data/Projects/facebook_data.csv', na_values='NAN')
```

```
fb_data.head()                    # Displaying first five rows of the data from imported dataset.
```

	userid	age	dob_day	dob_year	dob_month	gender	tenure	friend_count	friendships_initiated	likes	likes_received	mobile_likes	mobile_likes_received
0	2094382	14	19	1999	11	male	266.0	0	0	0	0	0	0
1	1192601	14	2	1999	11	female	6.0	0	0	0	0	0	0
2	2083884	14	16	1999	11	male	13.0	0	0	0	0	0	0
3	1203168	14	25	1999	12	female	93.0	0	0	0	0	0	0
4	1733186	14	4	1999	12	male	82.0	0	0	0	0	0	0

```
print('Data Shape:', fb_data.shape)       #Printing shape of data
```

```
Data Shape: (99003, 15)
```

This shows that our data has **99,003 values** and **15 variables**.

Understanding of Data

```
fb_data.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
```

```
# To check which column is unique, null and not-null values.
fb_data.agg(['count', 'size', 'nunique'])
```

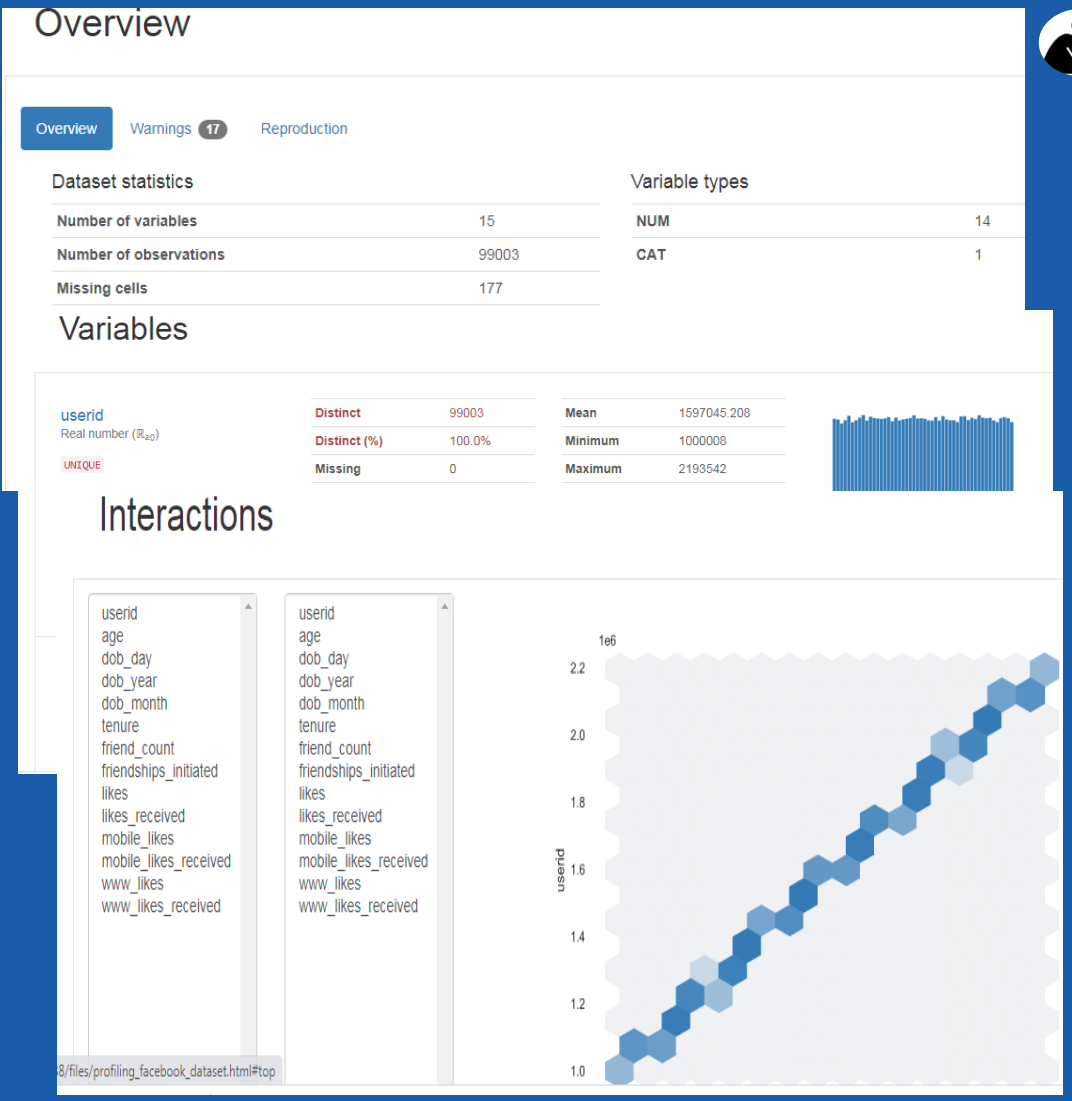
	userid	age	dob_day	dob_year	dob_month	gender	tenure	friend_count	friendships_initiated	likes	likes_received	mobile_likes
count	99003	99003	99003	99003	99003	98828	99001	99003	99003	99003	99003	99003
size	99003	99003	99003	99003	99003	99003	99003	99003	99003	99003	99003	99003
nunique	99003	101	31	101	12	2	2426	2562	1519	2924	2681	2396

```
fb_data.describe()
```

	userid	age	dob_day	dob_year	dob_month	tenure	friend_count	friendships_initiated	likes	likes_received
count	9.900300e+04	99003.000000	99003.000000	99003.000000	99003.000000	99001.000000	99003.000000	99003.000000	99003.000000	99003.000000
mean	1.597045e+06	37.280224	14.530408	1975.719776	6.283365	537.887375	196.350787	107.452471	156.078785	142.689363
std	3.440592e+05	22.589748	9.015606	22.589748	3.529672	457.649874	387.304229	188.786951	572.280681	1387.919613
min	1.000008e+06	13.000000	1.000000	1900.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.298806e+06	20.000000	7.000000	1963.000000	3.000000	226.000000	31.000000	17.000000	1.000000	1.000000
50%	1.596148e+06	28.000000	14.000000	1985.000000	6.000000	412.000000	82.000000	46.000000	11.000000	8.000000
75%	1.895744e+06	50.000000	22.000000	1993.000000	9.000000	675.000000	206.000000	117.000000	81.000000	59.000000
max	2.193542e+06	113.000000	31.000000	2000.000000	12.000000	3139.000000	4923.000000	4144.000000	25111.000000	261197.000000

- Data is clean. There are **no negative values**
- There are only **two genders** in our dataset
- Mean is approximately equal to median which is at 50th percentile for first 6 columns (till tenure) and is considerably more for remaining 9 columns
- We also see that there is significant difference between 75th percentile and max value for columns like, like_received, mobile_likes, mobile_likes_received, www_likes, www_likes_received
- Mimimum **age** of facebook user is **13 years and maximum is 113 years** and **average age** of facebook users is **37**

Pandas Profiling



Observations from Pandas Profiling before Data Processing

Dataset info:
Number of variables: 15
Number of observations: 99003
Missing cells: 177 (0.1%)

Variables types:
Numeric = 14
Categorical = 1

- **www_likes** has **60999 (67%) zero values**.
- **age** is **highly correlated** with **dob_year**.
- **mobile_likes_received** and **www_likes_received** are **highly correlated** with **likes_received**.
- **likes_received**, **mobile_likes_received** and **www_likes_received** are **highly skewed** with **skewness** of **112.0745**, **107.5312** and **126.2573** respectively.
- The distribution is **positive (right) skewed** as the **mean (142.689363)** is to the **right** of the **median (8)** for **likes_received**.
- **mobile_likes_received** and **www_likes_received** are also **positive (right) skewed** for same reason .
- **Gender** has data divided into **2** distinct values.
- All the **remaining columns** can be studied in the same way.

Missing Data

```
fb_data.isnull().sum()
```

```
userid                   0
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
```

- We have replaced null value of column gender by **mode of the column**
- Additionally we have removed two rows where tenure value was missing as it was not having any impact on our dataset.
- Similarly we have also remove column of users' day and month of birth as it will not fetch any analysis.
- Now are data set has total of **99001 rows** and **13 columns**

We saw that **0.176%** of entries are missing for gender and **0.002%** of rows doesn't have tenure

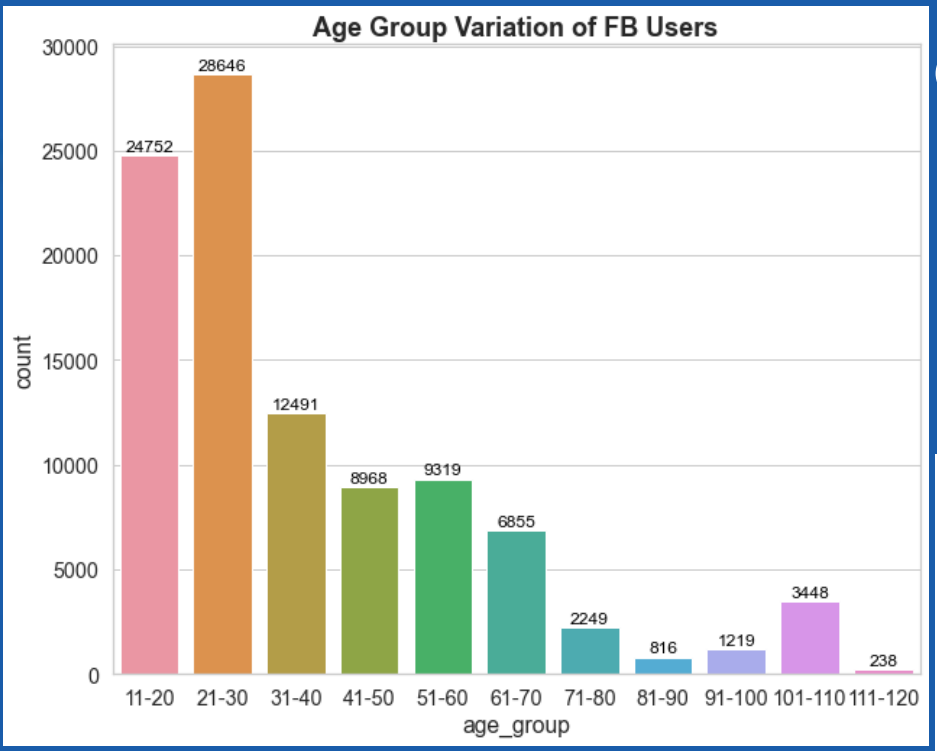
```
#Adding missing value in gender column which is equal to mode of the column
fb_data['gender']=fb_data['gender'].fillna(fb_data['gender'].mode()[0])
print(fb_data['gender'].unique())
```

```
['male' 'female']
```

```
#dropping the two rows where value is missing for column 'tenure'
fb_data.dropna(inplace=True)
fb_data.info()
```

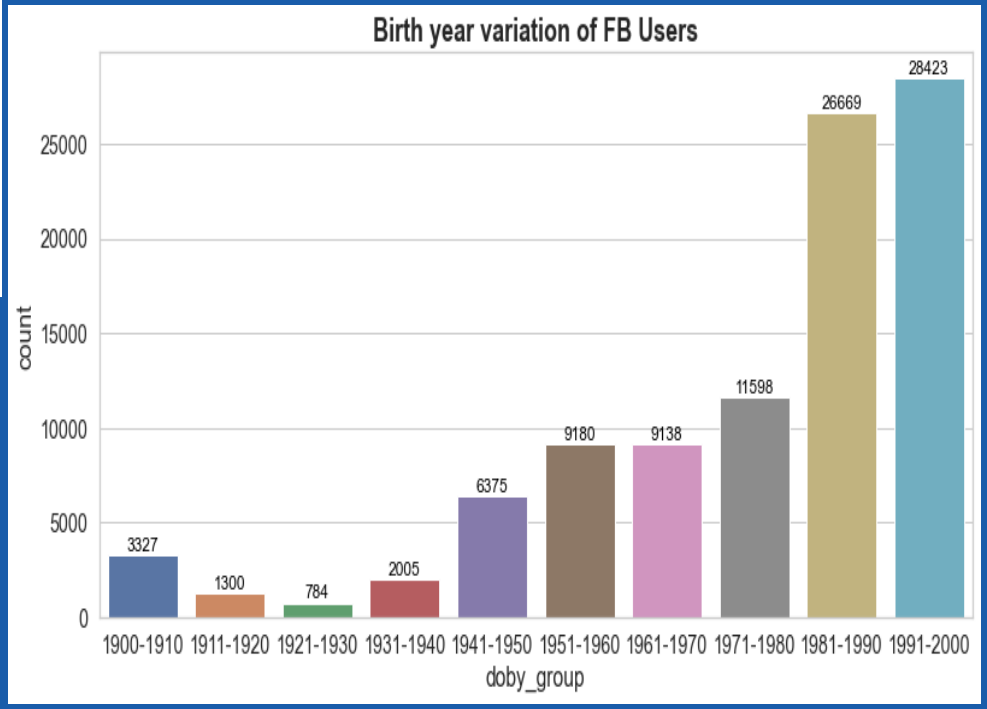
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 98826 entries, 0 to 99002
Data columns (total 15 columns):
```

Grouping

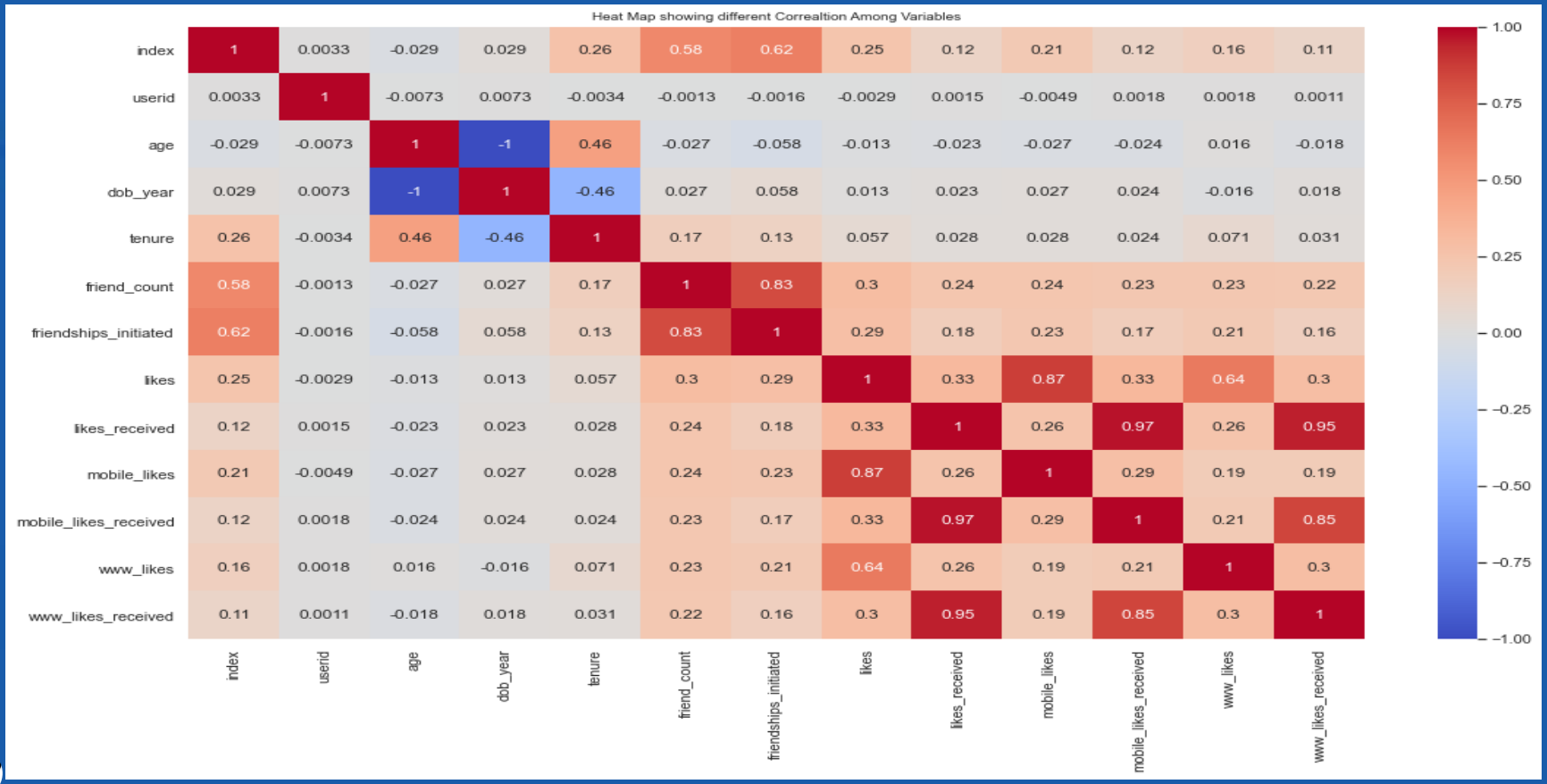


- Facebook users count increases gradually for people born in year 1921 and onwards
- **Maximum users** given in our dataset are born between **1981-2000**, however we see more number of users in year range 1900-1910 compared to 1911

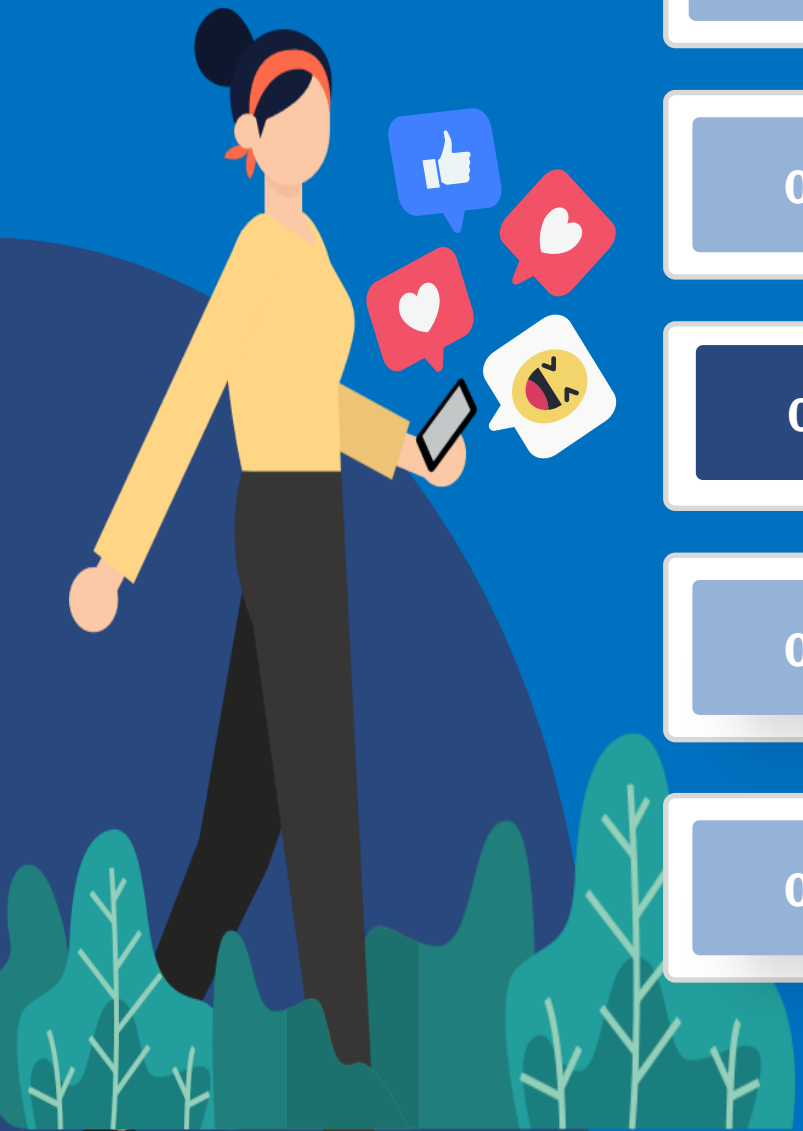
- People of **age group 11-30** are the **heaviest users** of Facebook and then usage gradually declines
- One thing to note is that **sudden rise** of users in **age group 110-110** which can be linked to people who were **born in 1990-1910**



Correlation Matrix

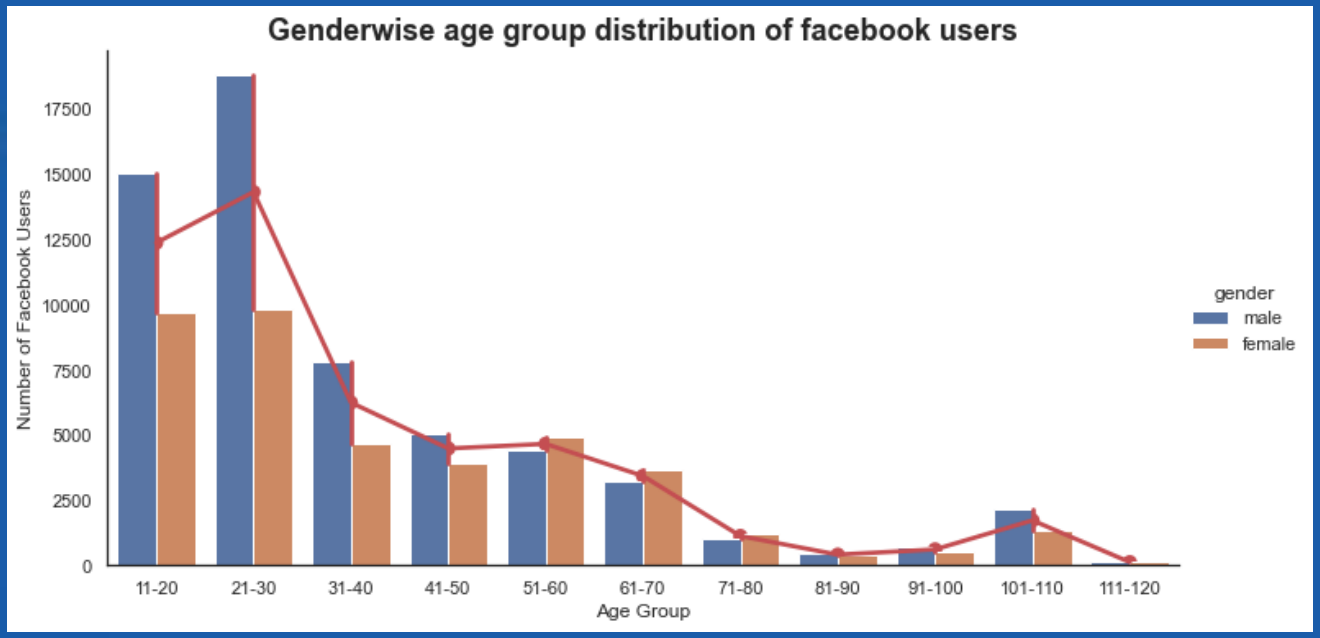


- Most of the likes received are mobile_likes_received **(postively correlated)**
- Most of the likes given are also through mobile_likes **(positively correlated)**
- 'dob_year' is **negatively correlated** with 'age' and 'tenure' also 'age' & 'tenure' strongly correlated.
- The correlation among variables (likes' & 'mobile_likes', 'likes' & 'www_likes', 'likes_received' & 'mobile_likes_received', 'likes_received' & 'www_likes_received', 'www_likes_received' & 'mobile_likes_received', tenure and 'friend_count' & 'friendships_initiated') is so strong that it can indicate a **situation of multicollinearity**.



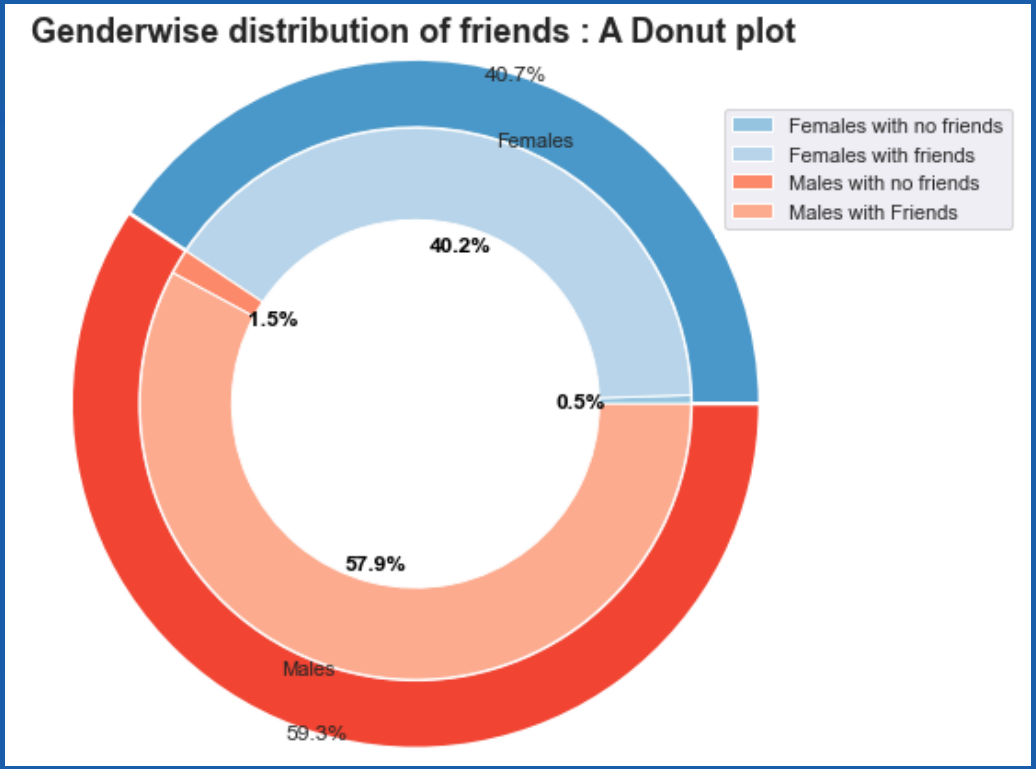
Analysis of Dataset

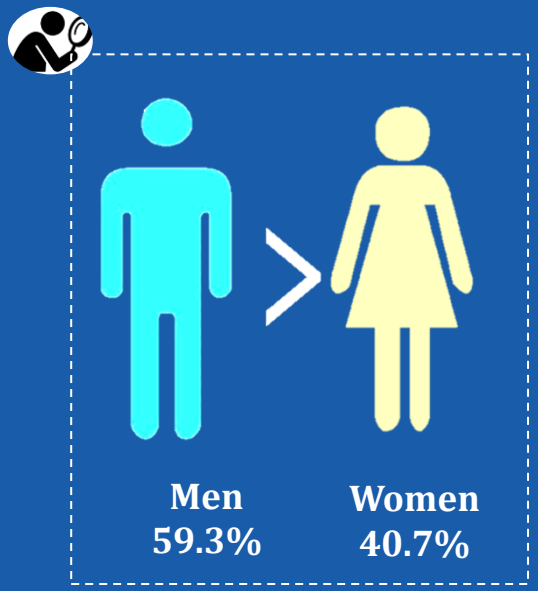
Which gender and age group use Facebook more?



- This shows that males between age of **11 to 30 years are twice more** than males of age group 31 onwards.
- The spike in males and females of age group 101 to 110 could be that these users have provided default date as 1900 year while setting up the Facebook account so that age can be unknown/ignored. Or these could be fake profiles
- Facebook used by **Male is higher than Female** user between 11 to 50 year age but after 51 age onwards female usage is gradually increased compared to males.
- Possible reason for less females on Facebook :
  - **Privacy**: a lot of women are worried about their privacy online as their pictures, profiles etc. have been misused.
  - **Rural penetration**: While the Internet and smartphones have reached tier 2,3 cities and villages, women still don't get their hands onto smartphones and internet as easily as men do.

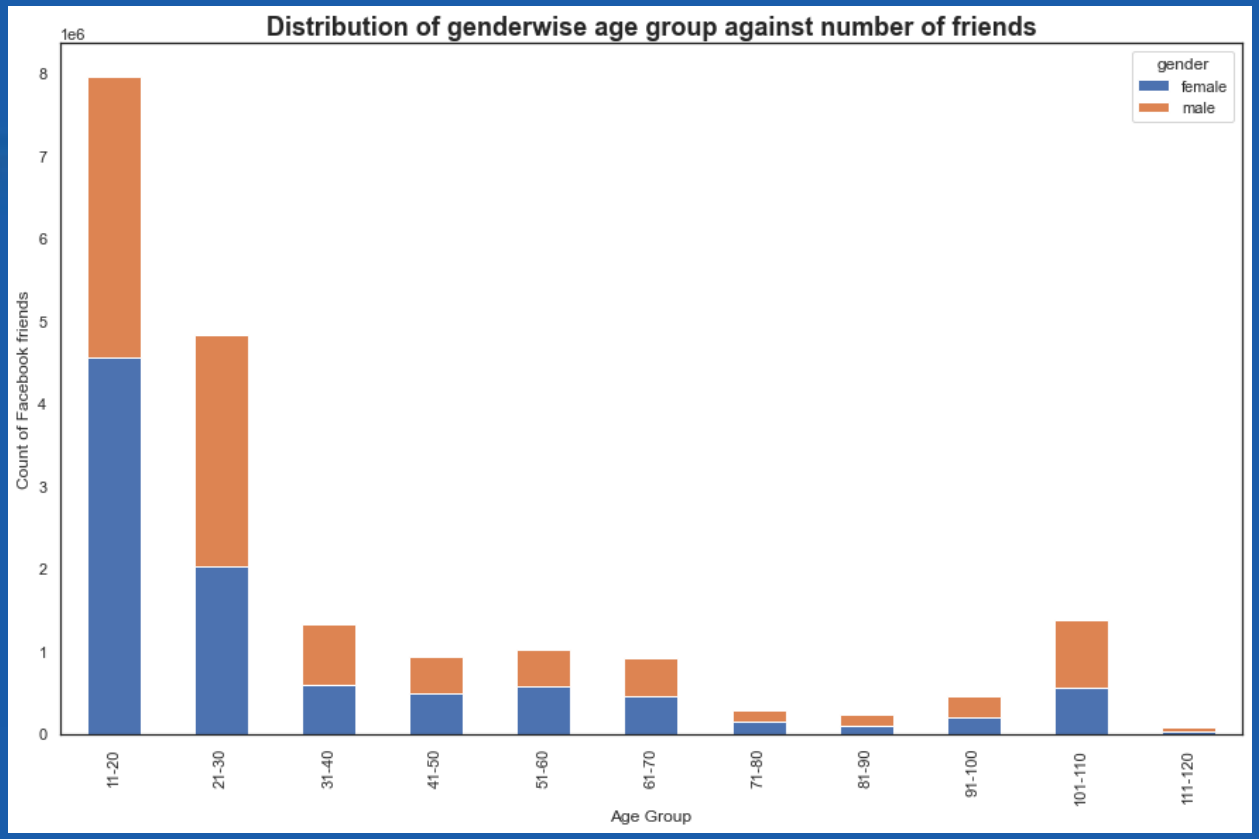
Which gender has more friends on Facebook?





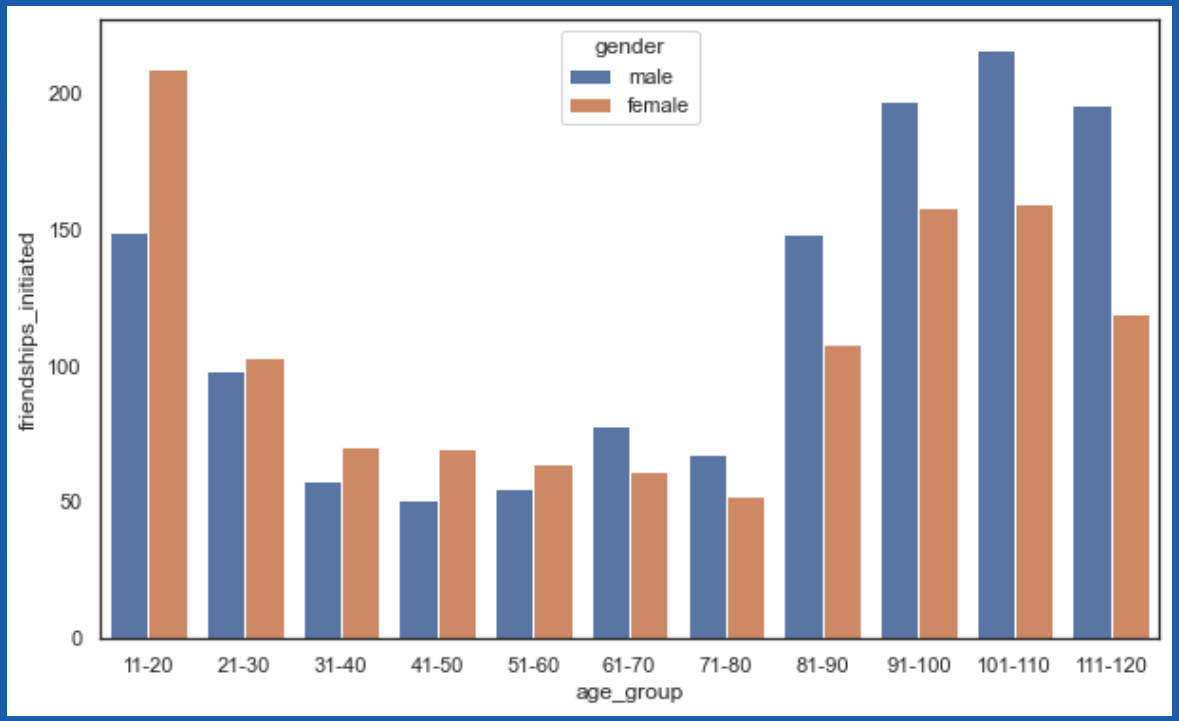
- There are only 0.5% females who doesn't have friends on Facebook.
- Males are heavy users of Facebook but many of them don't have friends in comparison to females
- This shows that females are more interested in building their social network and stay connected

Which age-group has more friends on Facebook?



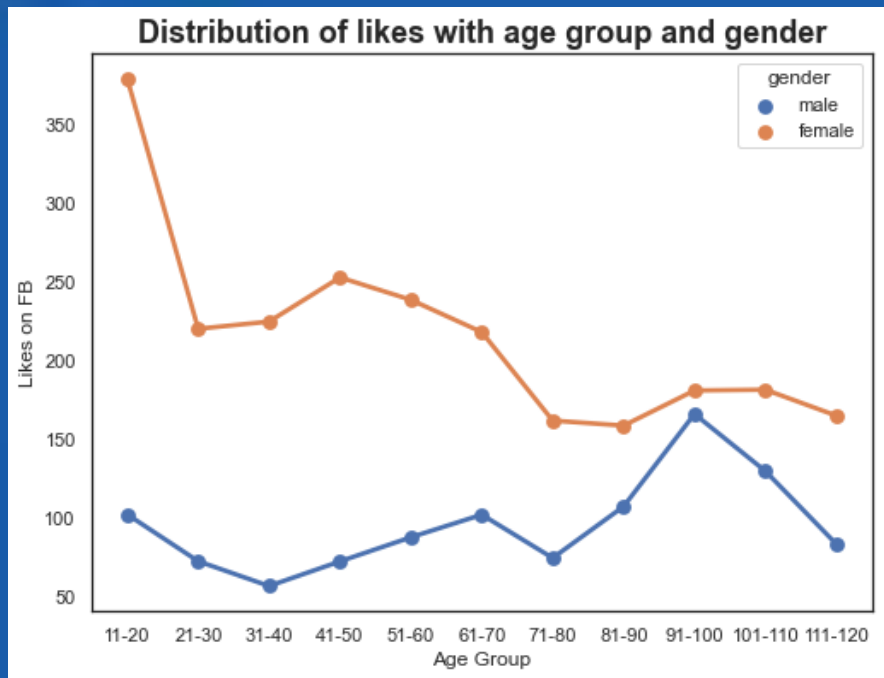
- This shows that users between age of **11 to 30 years have maximum number of friend**s when compared to users of other age group. Youngster tend to have bigger friend circle because of their school/college friends.
- Trend starts declining for subsequent age group but we see friends count for age group **101 to 110 is almost twice** of age group 71-90. Here we see that **males have more friends than females**
- Generally, **females use** social networking sites **to make connections** and stay in touch with family or friends. **Men**, by contrast, **use** social media **to gather the information they need to build influence**.

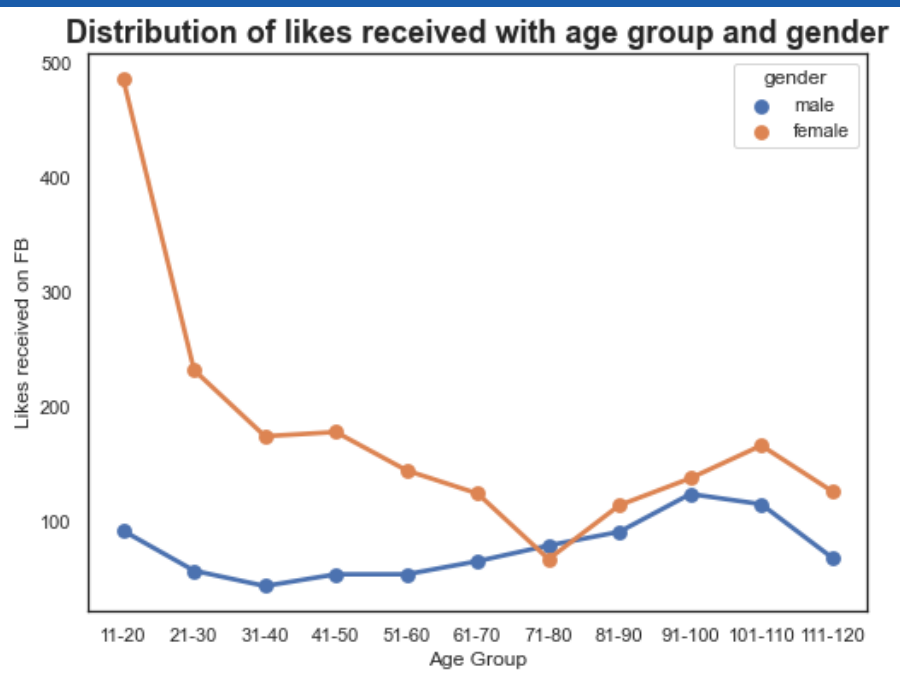
Which age-group initiated more friendships?



- As seen above, females up to 60 years of age initiates more friend request compared to males of same age. This strengthens the point that females use social media more to stay connected with families and friends compared to men.
- Yet again we see sudden rise in males of age group 91-110 years, in initiating friend requests. These people have mentioned their birth year between 1900-1920. This shows that people haven't entered all details correctly while creating account on Facebook.
- Additionally males with incorrect date of birth are the ones who have initiated maximum number of friend requests.

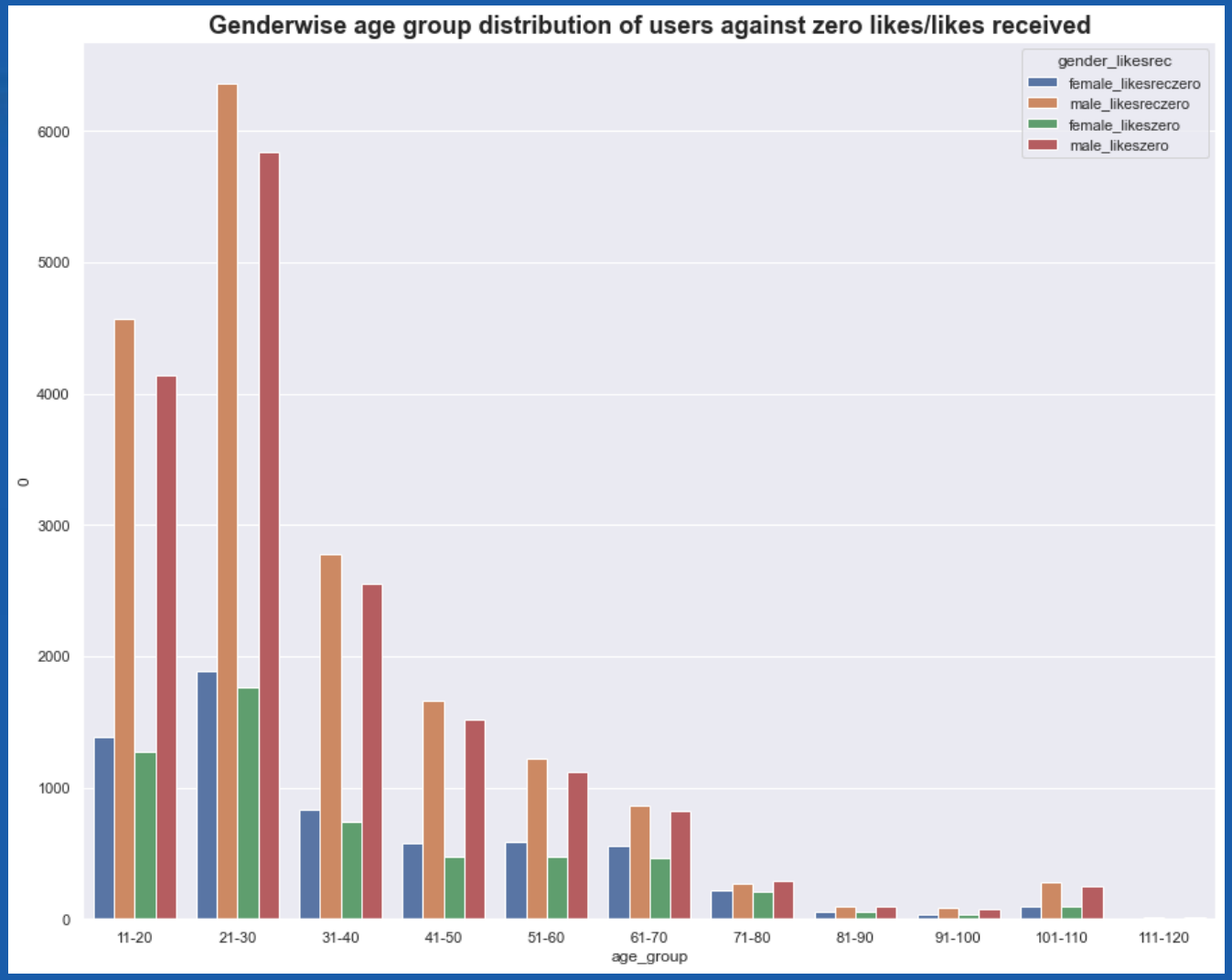
Which age-group received highest likes?





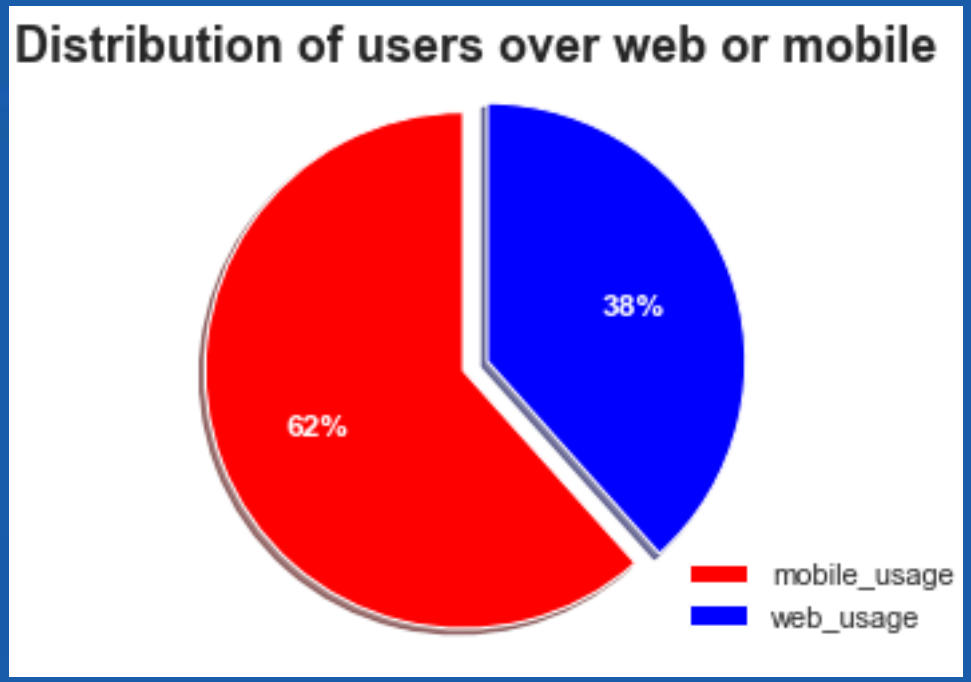
- For both likes and likes_received females outnumber males by significant value.
- Additionally **likes received by males upto age 80 are less than 100**.
- **Females are more active on social media and show their support by liking each others' posts. Also females tend to post more pictures to have better impressions when compared to males thus fetching more likes.**

Which age-group didn't receive any likes?



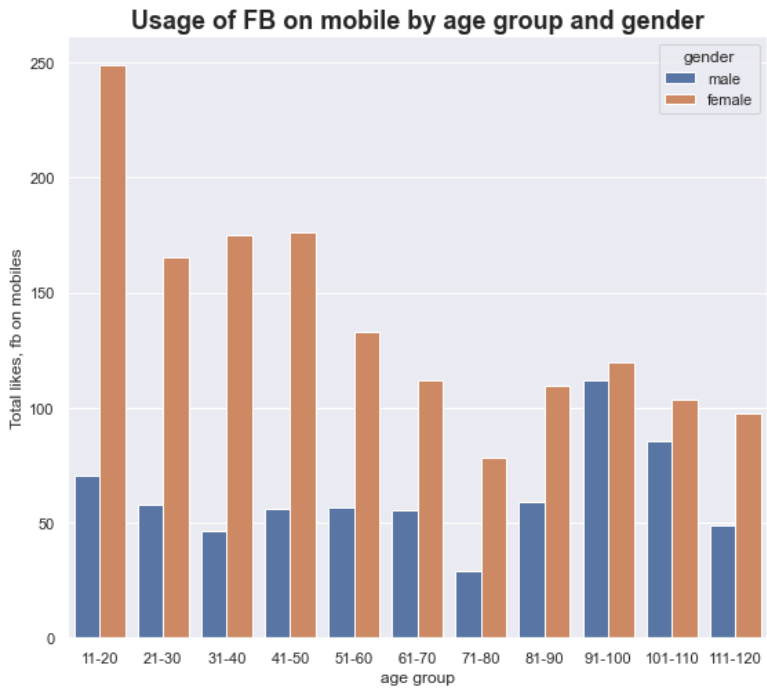
- We see that male population who doesn't have any likes is higher than females in every age group
- A huge spike is noticed for zero likes/likes received in males of age group 21-30. This could be that this age group has more male users who have recently join Facebook and haven't posted much content whereas on the other hand females joined recently but started posting more content and hence got more likes.

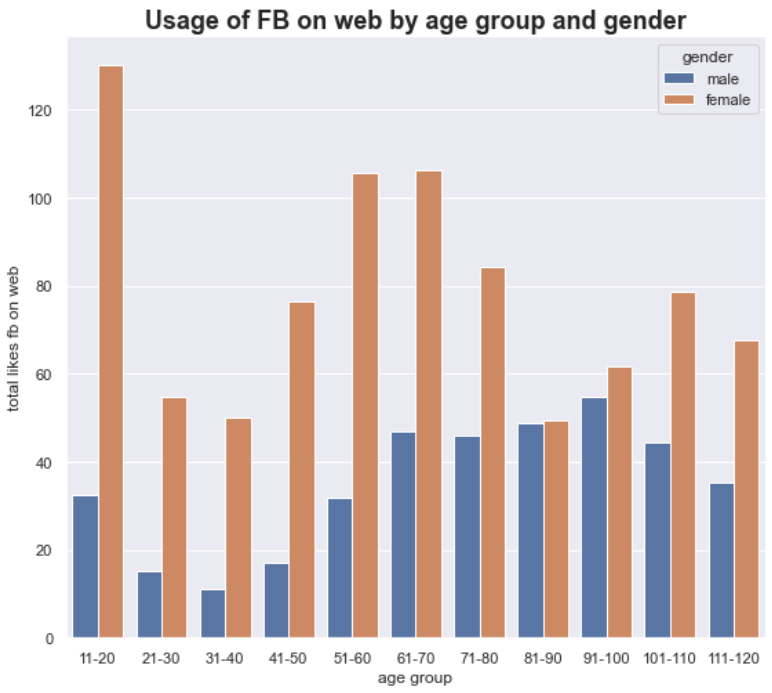
Which platform users are using to access Facebook?



- The possible reason for using Facebook app over mobile :
  - With increasing demand of smart phone, it is convenient for users to use Facebook over app as mobiles are much more handy compared to laptops.
  - Apps can send push messages that keep users engaged with Facebook.
  - App icon is visible when a user explores their phone, keeping awareness high.
  - App can retain user data, enabling one-click checkouts and minimizing load times.

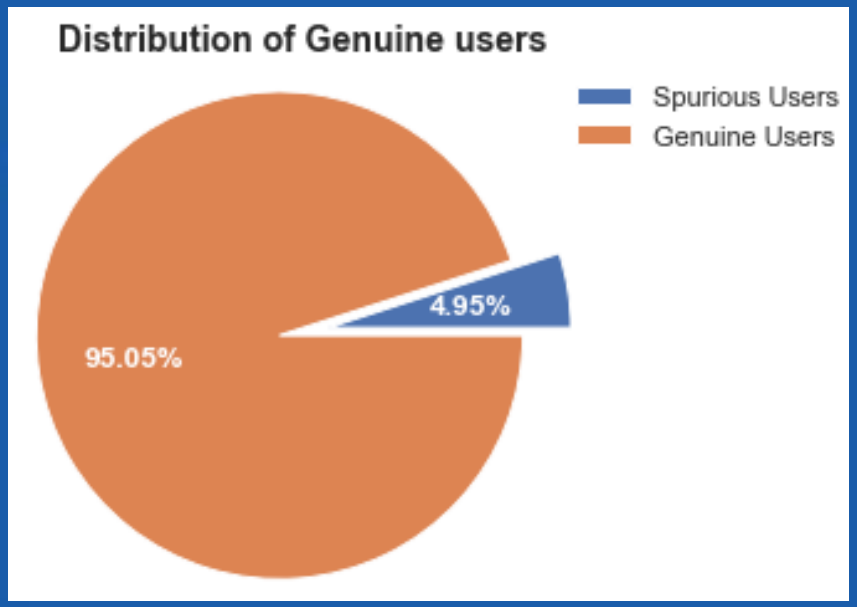
Which gender and age group use mobile to access Facebook?



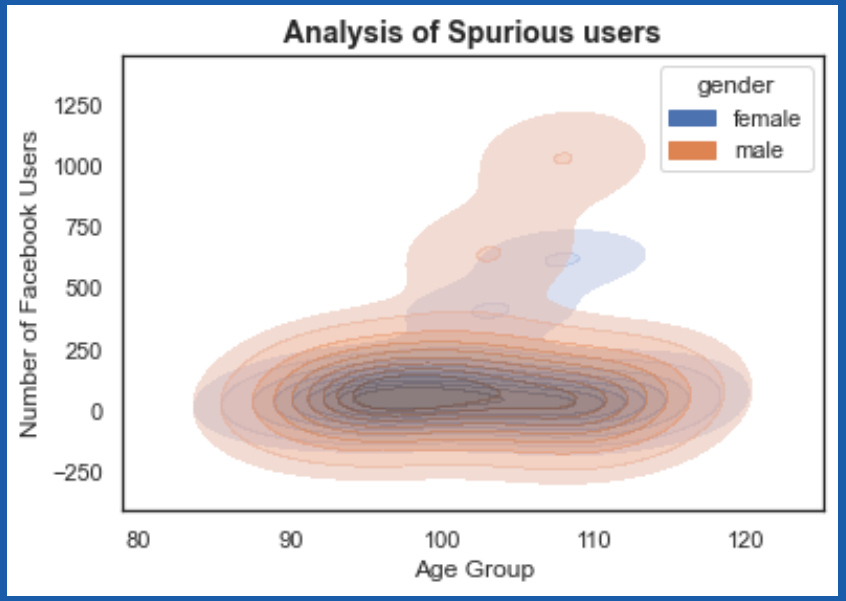


- Based on these graphs we can say that Facebook is mostly used on mobiles. With growing tech era use of mobile phones has become quite necessary and people are using it more to stay connected hence the company should focus on enhancing mobile app to have more user base.
- Also advertisement can be introduces on mobile app which will add income to advertisers
- Another point to note here is that males have less count of likes up to age 80 and then we see sharp increase for age 91-100. The unusual way could be because Facebook has default birth date as 1 Jan 1990. It may be the case when user just skip over detail, or keep maintaining their privacy. Whatever the case, it's important to look over these anomaly and understand the impact.

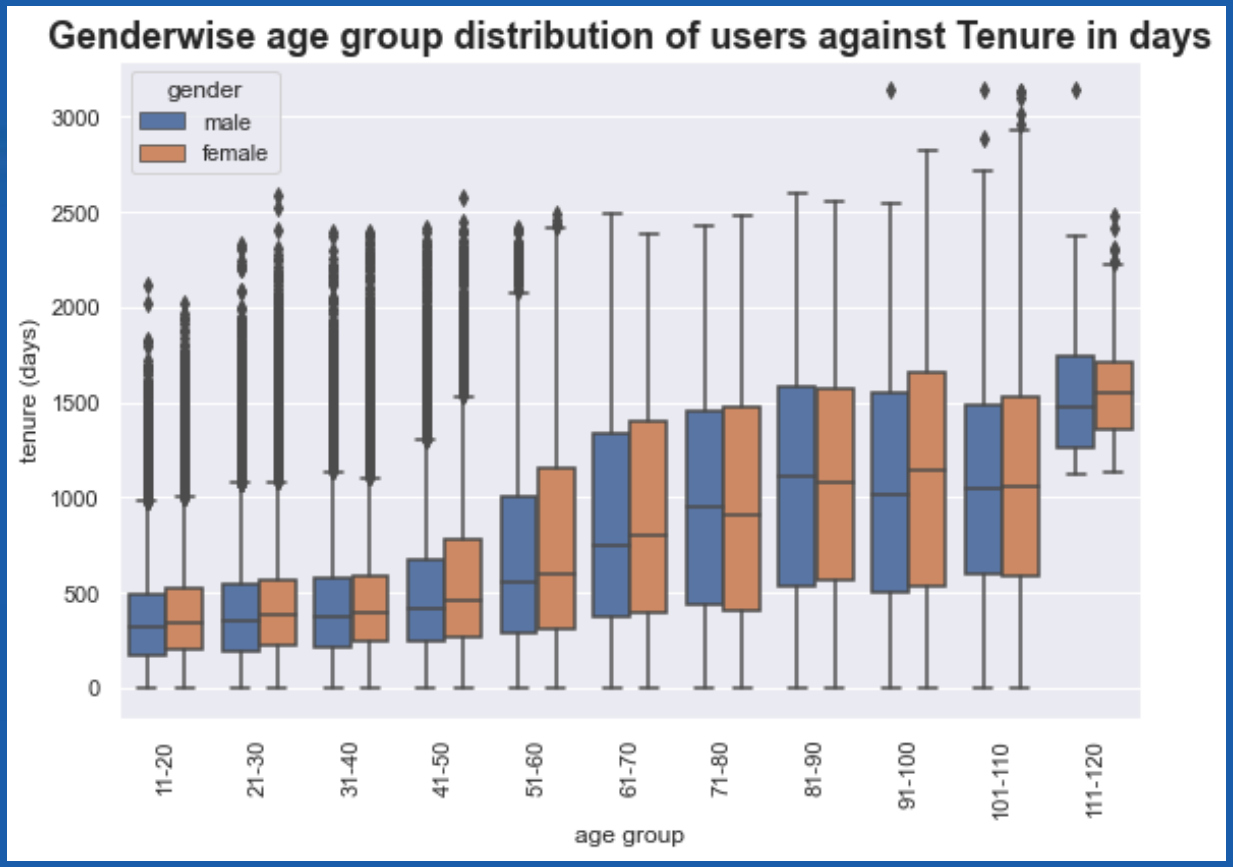
How many users have entered incorrect date of birth ?



- As we can see that graph is densely colored between 90-100, hence users falling under age group 90-110 are the ones who have entered their incorrect date of birth; this could be intentionally or unintentionally

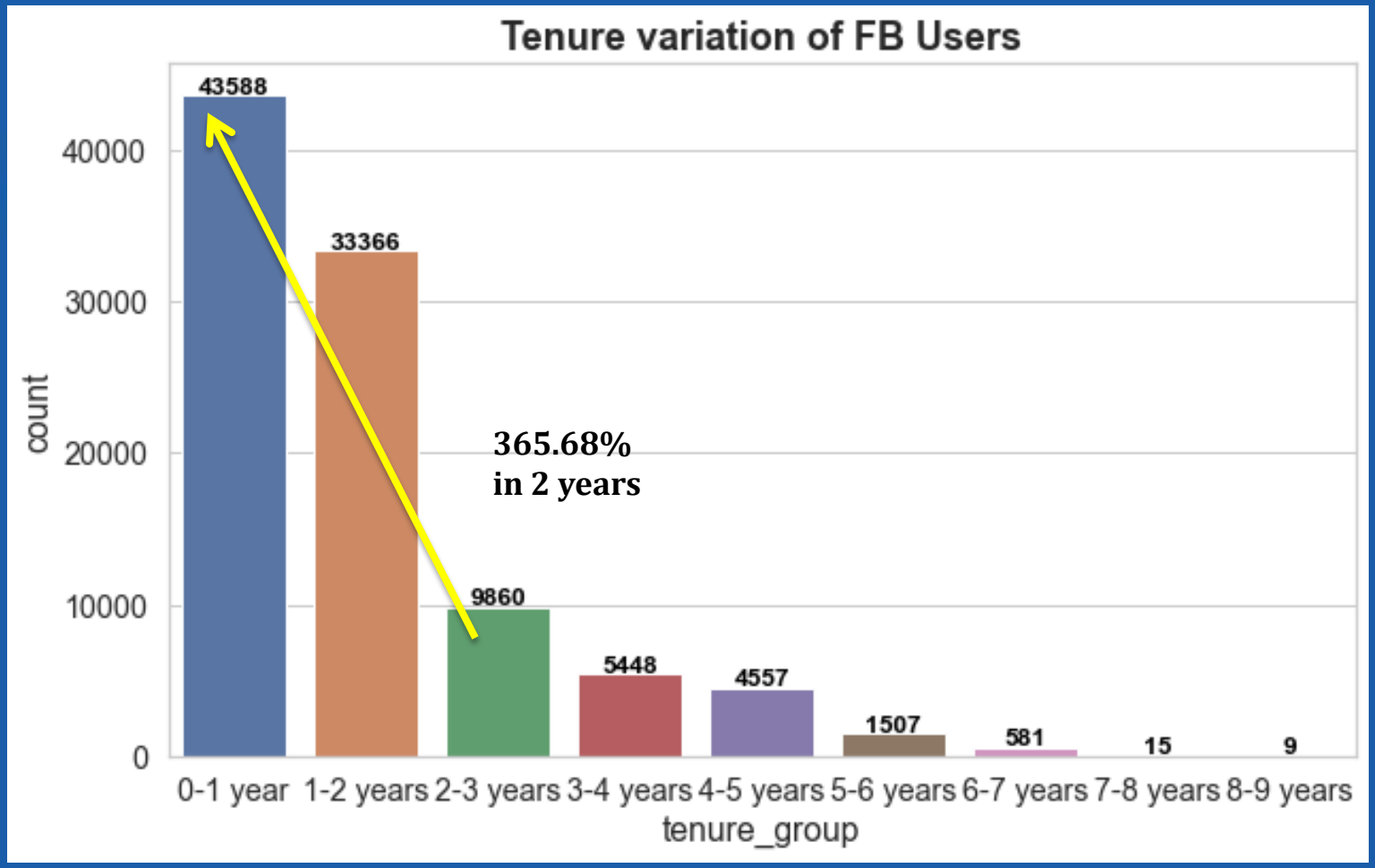


How Facebook is used over years?



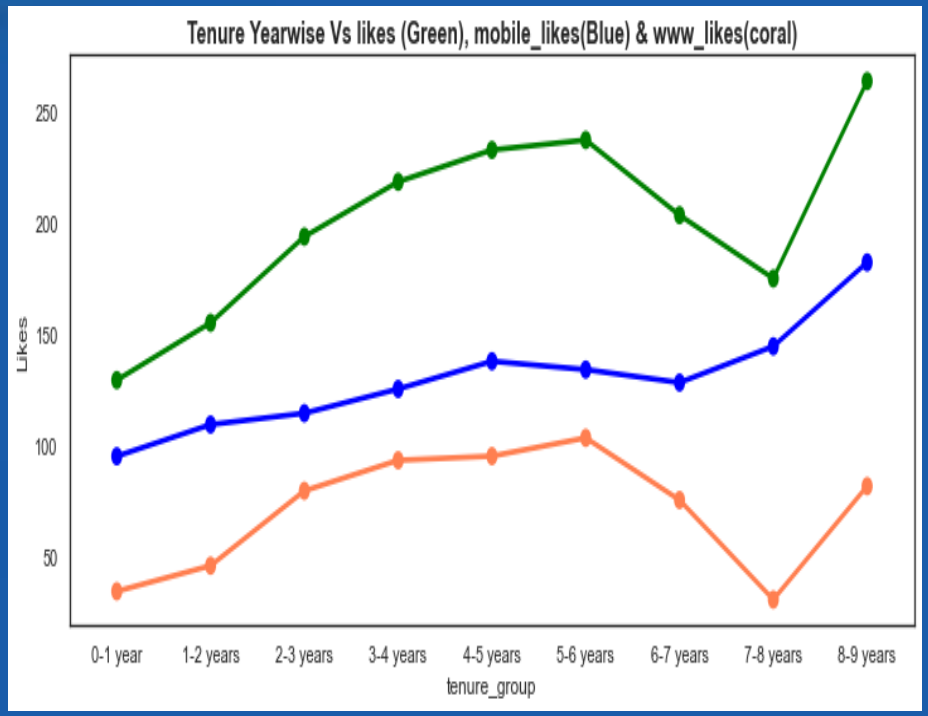
- It shows age group 91-100, 10 year ago started using Facebook for socializing. Additionally it is surprising to see age group 20-29 using Facebook almost 6-7 years ago i.e. 2006-2007.
- Even though female user count is less yet **female tenure is higher** compare to male across all age group
- The above tenure graph shows that Facebook was founded in 2004 as usage is between 2005 to 2013 which is almost 8 to 9 years ago. Also from age and date of birth we can confirm that dataset created around 2013

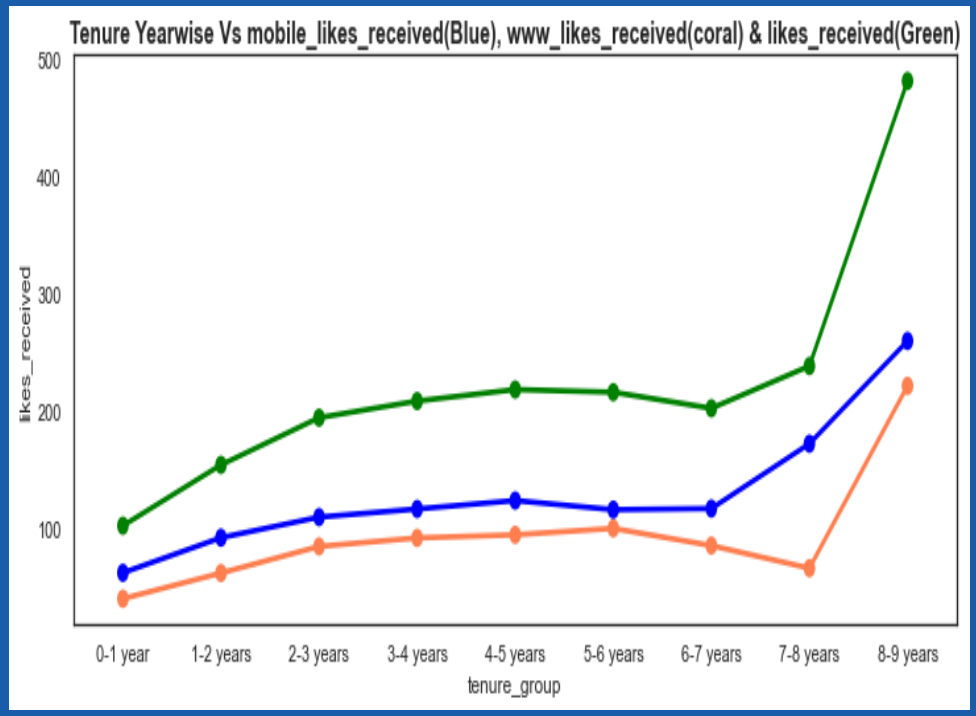
How Facebook is used over years?



- Highest tenure for usage of Facebook is 1 year and avg is 1.5 years.
- This helps in defining the model to engage users for more than a year
- We see a sharp increase in user base in last 2 years (~365.68%).
- Since our data set is of 2013 it shows more people joined from 2009 onwards and that possibly could be because of mobile app as facebook app was launched in 2008.

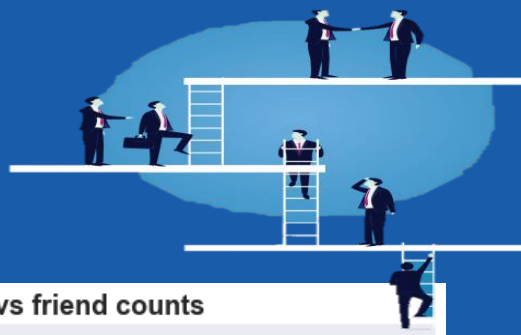
How Facebook is used over years?

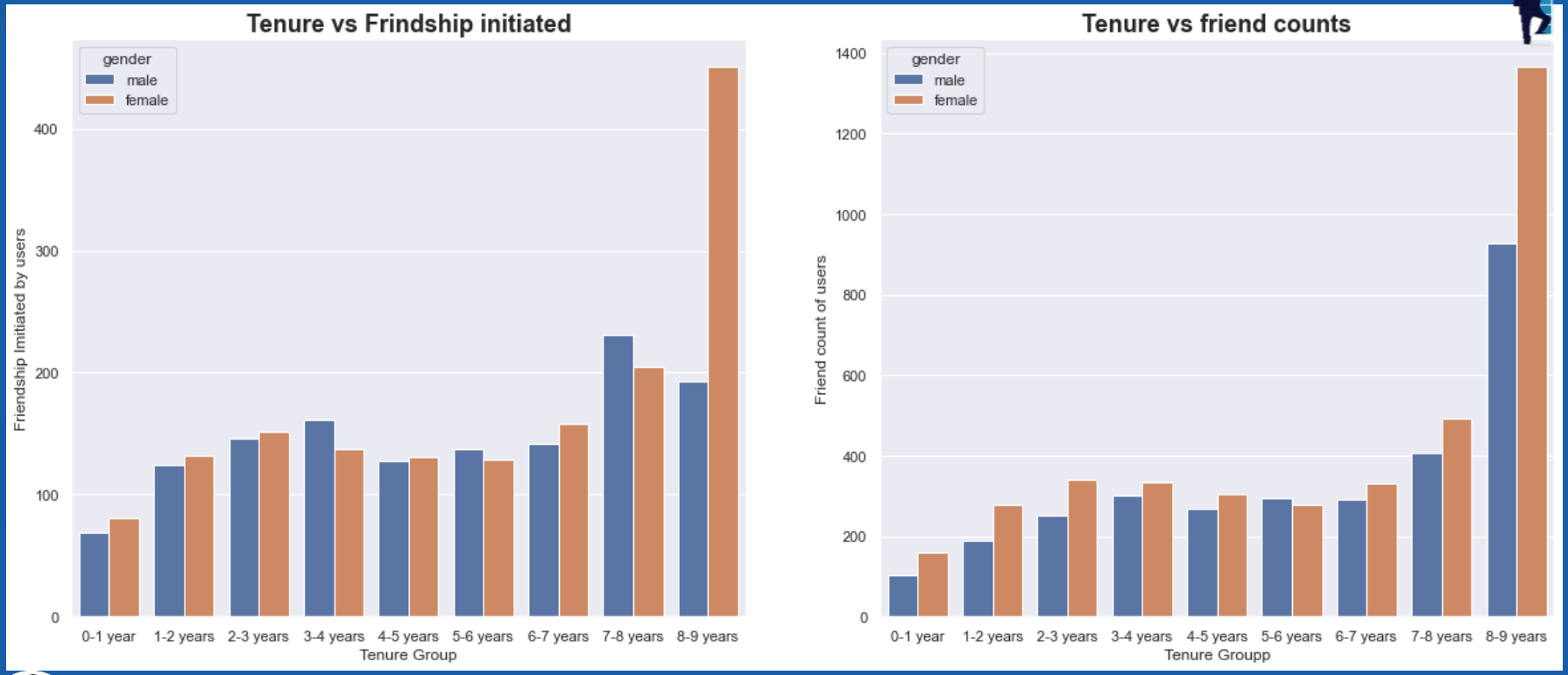




- Mobile likes/Mobile_likes_received are higher in each group compared to www likes/www_likes_received
- For group **7-8 years** we see sharp decrease in www_likes/www_likes_received. **Over the years users have switched over to mobile app instead of web.**

How Facebook is used over years?





- Over the years females have initiated more friendships and have more friends.
- This shows that friendship initiated is proportional to friend counts over the years which we discussed in correlation matrix as well
- Almost across all tenure groups males initiated less friendships except for (3-4, 5-6 and 7-8) years and have less friends in all tenure groups compared to females

Summary

Age Group: Young population in age group 10-30 years is heavy user of Facebook. Hence marketing insight should focus on youngster as they have time and interest.

Likes: Females tend to get more likes than their male counterparts, ~ 2.5 times more. Likes can be very helpful in promoting small or large scale business by introducing advertisements.

Spurious Users: With 4.95% of users entering their incorrect details, raise concerns on genuine accounts on Facebook. Hence developers should come up with strict checks to avoid misleading information such as age.

Gender: With 59.3% of males, they are using Facebook more . Hence Facebook should focus on promoting activities which will increase female user base and at the same time continue working on male activities to keep the count intact

Platform: With the launch of Facebook mobile app, We see that 66% of users access Facebook via mobile. Hence team should work on enhancing their mobile app features and ensure data security at the same time.

Tenure: We see a huge % increase (~ 365.6%) in Facebook users between 2009 and 2013. Facebook mobile app was launched in 2008 and hence we see this huge spike in active users in following years. Marketing team should work to enhance tenure longevity by introducing new features.

Friends: Youngster have wider friend circle . Additionally though male users are more and they initiate more friend requests yet females have more friends. Youngster can get more user base by inviting their friends on facebook. Hence marketing team can come up with attractive strategies here

References

- Dataset : Taken from Github repository
  Reference link to repository: https://github.com/insaid2018/Term-2/blob/master/Term-I-II%20EDA%20Project%20Datasets.ipynb

- Images and Gif: (Google) www.google.com

- Presentation template: https://24slides.com

Thank You!