ASSIGNMENT

By Sabuj Mandal

1) Load the data and impute missing values Imputation of missing values:

● Replace the null values (NA) of gender column with its mode or median and explain why mode/median used to replace NA values.

Solution:

Missing Value Treatment

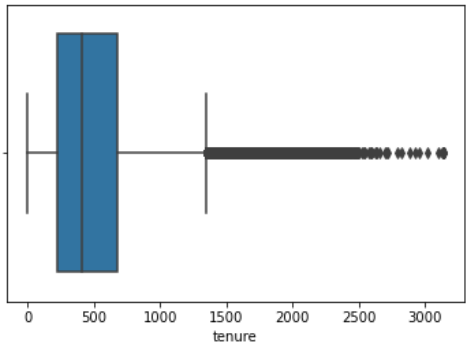
```
In [217]: data.isnull().sum()

Out[217]: userid                    0
          age                       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
          age_group                 0
          tenure_in_years           2
          tenure_year_group        11
          dtype: int64
```

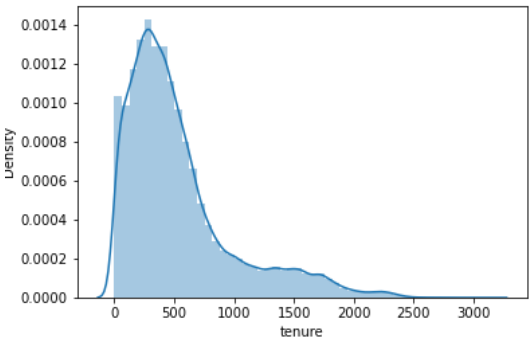
Facing missing values in real life project is very much common. Missing values affect the performance of our model along with its predictive capacity. NA values have that capacity to change all the statistical parameters of our project which will lead to bad performance of our model and we wolud not be able to reach our desired accuracy.

As gender here is a categorical column, there is no mean nor median, so in these type of cases our best option to use the mode (the most frequent value).Mode is the most frequent value in our data set. Mode is thus used to impute missing values in columns which are categorical in nature. Also mode is the values that reflects the central tendency better than mean or median. Mode is also not influenced by outliers.

After mode, it is the median that reflects the central tendency the best. Which implies that for continuous data, the use of the median is better than mean. Median is the middle score of data-points when arranged in order.

● Replace the null values (NA) of tenure column (numerical variable) with its median, and explain why mode/median used to replace NA values.



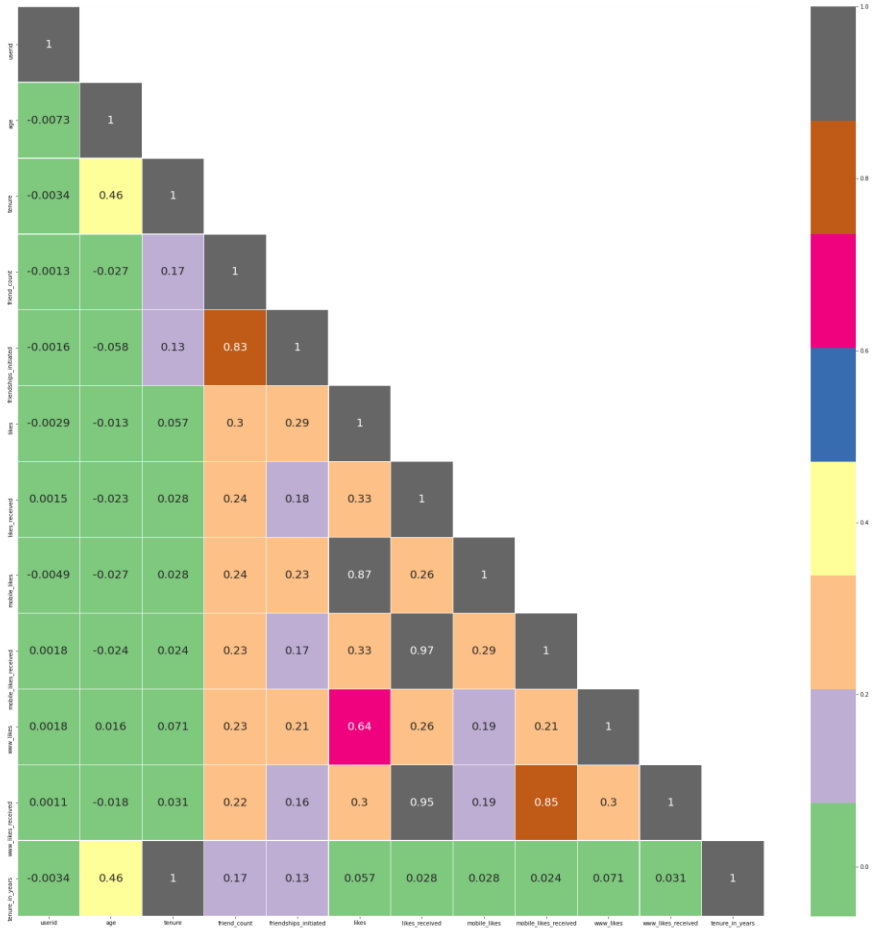


From the above box plot and distribution plot, we may note that the data is skewed. There are several or large number of data points which act as outliers. Outliers data points will have significant impact

on the mean and hence, in such cases, it is not recommended to use mean for replacing the missing values. Using mean value for replacing missing values may not create a great model and hence gets ruled out.

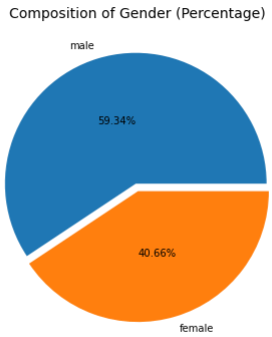
So in this case it's better to use median to replace NA values.

2) Plot heatmap / correlation matrix on all the columns.



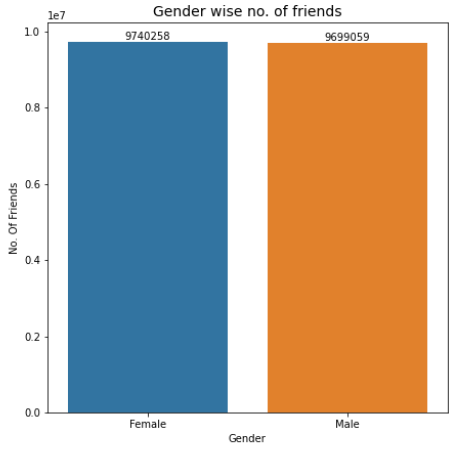
3) Analysis based on gender of the users

- What is composition of male and female users?



In our dataset, there are total 99003 no. of datapoints are present. Among them after filling missing values, 59.34% of total users, 58749 are male and 40254 are female which is 40.66% of total users.

- Which category of gender has more friends?



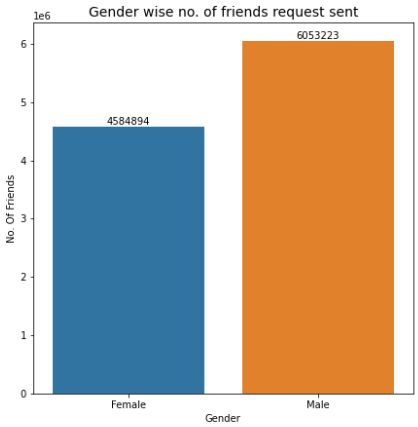

```
In [262]: data.groupby("gender")[["friend_count"]]
              .aggregate(['count','sum','mean'])
```

Out[262]:

	friend_count		
	count	sum	mean
gender			
female	40254	9740258	241.969941
male	58749	9699059	165.093176

From this above bar plot we can say females have a little bit more no of friends than males. In an average if we look after the above table we can state that females have also higher avg. no friends than males. There is a high positive correlation between friendships initiated and no. of friends which is 0.83, which can be noticed from our correlation matrix above. So we can state that having friends is hugely depend on the no. of friend request the person send. It's quite obvious too. But very low positive correlation (0.17) with tenure in Facebook, it states that tenure has very low effect on having no. of friends in Facebook, it means using Facebook for long time does not help you to get more friends.

● **Which category of gender initiated more friendships?**




```
In [264]: data.groupby("gender")[["friendships_initiated"]]
          .aggregate(['count','sum','mean'])
```

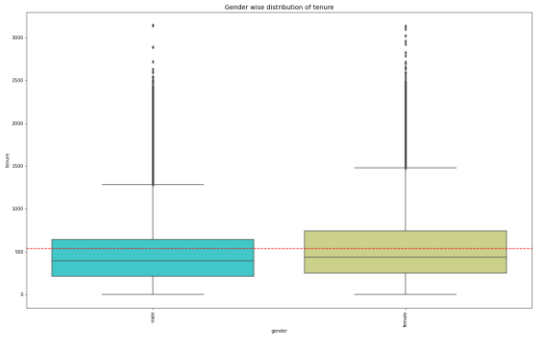
Out[264]:

	friendships_initiated		
	count	sum	mean
gender			
female	40254	4584894	113.899091
male	58749	6053223	103.035337

Unlike the above situation, here we can say that, although females are higher in having more no of friends , from the above bar plot and table we can observe that, males are higher in total no of friend requests sent or friendship initiated than females, but from the table we can observe that no. of friends request sent by each female which is 113.89 is higher than that of each male which is 103.03.

So from the variables 'friend_count' & 'friendships_initiated' we can simply state that females have higher rate of acceptance of their friend requests than that of males.

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




```
In [266]: data.groupby("gender")[["tenure"]]
          .aggregate(['count','sum','mean'])
```

Out[266]:

	tenure		
	count	sum	mean
gender			
female	40254	23637975.0	587.220525
male	58749	29614237.0	504.080699

In our dataset, our male users have spent more time than females did. Male users have spent 29614237 days in Facebook where total females users have spent 23637975 days in Facebook. But on an average a female user have spent more time (587.22 days) than that of a male user(504.08 days). So, we can simply conclude it that female users have spent more time than male users in Facebook.

4) Analysis based on the least active users on Facebook

● How many users have no friends?

```
data1= data[data["friend_count"]==0]
data1.shape[0]
```

```
1962
```

In our dataset there are 1962 users don't have any no. of friends.

```
data1.gender.value_counts()
```

```
male      1459
female     503
Name: gender, dtype: int64
```

```
data1.groupby("gender")[["friendships_initiated"]].aggregate(['count','sum','mean'])
```

	friendships_initiated		
	count	sum	mean
gender			
female	503	0	0
male	1459	0	0

From our dataset we can sate that total 1962 users don't have any friends. Among them 1459 users are male and 503 users are female. And the obvious reason we can see from the table that they have not initiated any friendship to anyone. Also their average day of spending time in Facebook are about to half than total average.

● How many users did not like any posts?

There are total 22308 users did not have liked any post. Among them 16719 are males and 5589 are females. We can observe that these users have very less no of friends in their account. Average no of friend of a male and female are 241 and 165 respectively, here for theses users this average no is approximately 1/3 of the total users. Although they spend a quite good amount of time in Facebook comparing to the all users.

```
data2=data[data["likes"]==0]
data2.shape[0]
```

```
22308
```

```
data2.gender.value_counts()
```

```
male      16719
female     5589
Name: gender, dtype: int64
```

```
data2.groupby("gender")[["friend_count"]]
.aggregate(['count','sum','mean'])
```

	friend_count		
	count	sum	mean
gender			
female	5589	461541	82.580247
male	16719	1288556	77.071356

```
data2.groupby("gender")[["tenure"]]
.aggregate(['count','sum','mean'])
```

	tenure		
	count	sum	mean
gender			
female	5589	2495572.0	446.51494
male	16719	7494531.0	448.26431

- How many users did not receive any likes?

```
data3=data[data["likes_received"]==0]
data3.shape[0]
```

```
24428
```

In our dataset there are 24428 users did not receive any like for their post.

```
data3.gender.value_counts()
```

```
male      18188
female     6240
Name: gender, dtype: int64
```

```
data3.groupby("gender")[["friend_count"]]
.aggregate(['count','sum','mean'])
```

	friend_count		
	count	sum	mean
gender			
female	6240	377656	60.521795
male	18188	1233997	67.846767

```
data3.groupby("gender")[["friendships_initiated"]]
.aggregate(['count','sum','mean'])
```

	friendships_initiated		
	count	sum	mean
gender			
female	6240	205030	32.857372
male	18188	820270	45.099516

```
data3.groupby("gender")[["tenure"]]
.aggregate(['count','sum','mean'])
```

	tenure		
	count	sum	mean
gender			
female	6240	2728119.0	437.198558
male	18188	7751579.0	426.191940

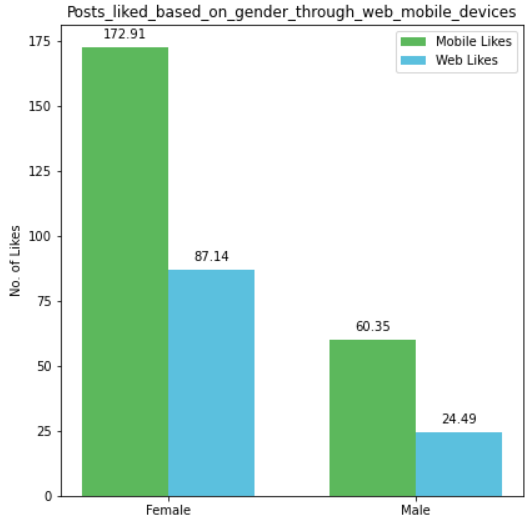
```
data3.age_group.value_counts()
```

```
20-30      9182
13-19      5031
31-40      3615
51-65      2632
41-50      2233
Over 65    1735
```

24428 users did not receive any likes. Among then 18188 are males and 6240 are females. There could be one thing clearly observed from here that because of the very less no of friends these users did not receive any like, these users are very less no of friends, with comparison to total users they have approximately 1/9th no of friends. Also they have sent very no of friend request, although their tenure in Facebook on an average is good.

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

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



So, through mobile devices and female users have liked a post more than a male user using mobiles, same way the average posts like by female users through web services are higher than that of male users.

```
data_mob=data[data["mobile_likes"]!=0]
data_mob.shape[0]
```

```
63947
```

```
data_www=data[data["www_likes"]!=0]
data_www.shape[0]
```

```
38004
```

```
data_mob.gender.value_counts()
```

```
male      34472
female    29475
Name: gender, dtype: int64
```

```
data_www.gender.value_counts()
```

```
female    19074
male      18930
Name: gender, dtype: int64
```

```
data_mob.groupby("gender")[["tenure"]]
.aggregate(['count','sum','mean'])
```

	tenure		
	count	sum	mean
gender			
female	29475	17736790.0	601.757082
male	34472	17775738.0	515.657287

```
data_www.groupby("gender")[["friend_count"]]
.aggregate(['count','sum','mean'])
```

	friend_count		
	count	sum	mean
gender			
female	19074	5968956	312.936773
male	18930	4474609	236.376598

```
data_mob.groupby("gender")[["friend_count"]]
.aggregate(['count','sum','mean'])
```

	friend_count		
	count	sum	mean
gender			
female	29475	8693190	294.934351
male	34472	7403564	214.770364

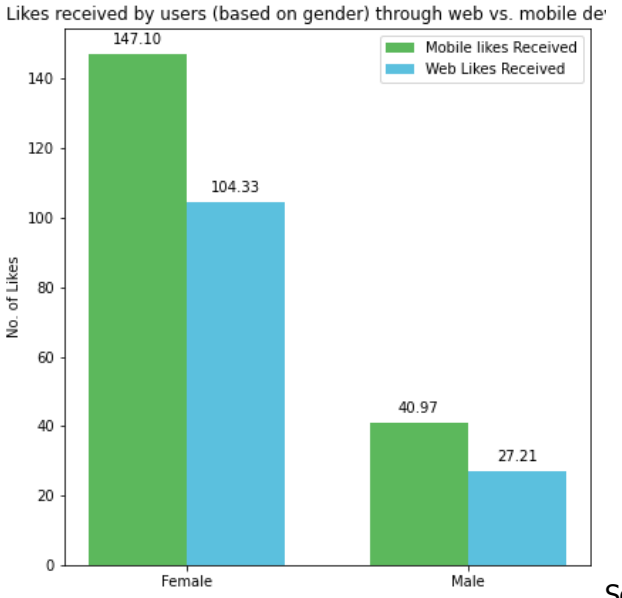
```
data_www.groupby("gender")[["tenure"]]
.aggregate(['count','sum','mean'])
```

	tenure		
	count	sum	mean
gender			
female	19074	13340321.0	699.398186
male	18930	11501788.0	607.595774

63947 posts have been liked by users through mobiles, where 38004 posts have been liked while using web devices. Higher no. of posts are liked by males than females in both the cases. And there 25256 posts are there that have been liked through both mobile and web devices.

But. Interestingly, the posts have been liked through web devices has higher average tenure than other, so when users like post through web devices they spend more time in Facebook.

● What is the average number of likes received by users (based on gender) through web vs. mobile devices?



So, average no of likes (147.10) received by female user is higher than that of male users (40.97), same way female users have received more likes through web services.

```
data_mob.age_group.value_counts(normalize=True), data_www.age_group.value_counts(normalize=True)
```

```
(20-30      0.335653
 13-19      0.227032
 31-40      0.126824
 51-65      0.118598
 Over 65    0.101021
 41-50      0.090872
 Name: age_group, dtype: float64,
 20-30      0.254315
 13-19      0.228134
 51-65      0.177850
 Over 65    0.153852
 31-40      0.095358
 41-50      0.090490
 Name: age_group, dtype: float64)
```

If we try to analyse the comparison of likes between through age_groups, we can observe that people who are in the group of 'Over 65' have liked any posts through web are more in percentage than the users through mobile.