**Observations on Facebook\_profiling\_before\_preprocessing.html:-**

1. Dataset Statistics

* Number of variables - 15
* Number of observations - 99003
* Missing cells - 177
* Missing cells (%) - < 0.1%
* Duplicate rows - 0
* Duplicate rows (%) – 0

1. Variable Types
   1. Numerical – 14
   2. Categorical – 1
2. Variables (columns)
   1. No missing values in “User Ids”, “Age”, “dob\_day”, “dob\_year”, “dob\_month”,
   2. “gender” is categorical and has 175 missing records and “tenure” has 2 missing records, which is 0.2% of overall data
   3. “friend\_count”, friendships\_initiated”, "likes", "likes\_received", "mobile\_likes", "mobile\_likes\_recieved" & "www\_likes\_recieved" have Zeros
   4. “friend\_count” & friendships\_initiated " have zeros with counts as 1,962 and 2,997 respectively, which is 3% of the overall data
   5. “likes” have 22308 zeros, which is 22.5% of the overall data
   6. “likes\_recieved” has skewed values & 24,428 zeros, which is 24.7% of overall data. IQR is 58.
   7. “mobile\_likes” has 35,056 zeros, which is 35.4% of overall data
   8. "mobile\_likes\_recieved" has values & 30,003 zeros, which is 30.3% of overall data. IQR is 33.
   9. “www\_likes” has 60,999 zeros which is 61.6% of overall data
   10. "www\_likes\_recieved" has skewed values and 36,864 zeros, which is 37.2% of overall data. IQR is 20.
3. Warnings

* likes\_received is highly skewed (γ1 = 112.0745682) - Skewed
* mobile\_likes\_received is highly skewed (γ1 = 107.5312999) - Skewed
* www\_likes\_received is highly skewed (γ1 = 126.257317) - Skewed
* userid has unique values - Unique
* friend\_count has 1962 (2.0%) zeros - Zeros
* friendships\_initiated has 2997 (3.0%) zeros - Zeros
* likes has 22308 (22.5%) zeros - Zeros
* likes\_received has 24428 (24.7%) zeros - Zeros
* mobile\_likes has 35056 (35.4%) zeros - Zeros
* mobile\_likes\_received has 30003 (30.3%) zeros - Zeros
* www\_likes has 60999 (61.6%) zeros - Zeros
* www\_likes\_received has 36864 (37.2%) zeros – Zeros

Problem Statement :-

To increase the utilization of FB across all ages

What can be done as part of preprocessing of data:-

1. **drop rows which have less than 1% of zeros, which belong to “gender” & “tenure”. Total no. of records will be 98,758. - NA**
2. **Minimum and maximum ages are 13 & 113 respectively, can we come up with an range to fill the zeros of other columns. Eg: “age\_group” as 13-17, 18-24, 25-35, 36-50, 51-80, 81-113 etc.,**
3. **Bucketing the age into ranges will help to analyze the “friend\_count” – Multivariate- Gender, Age group & Count**
4. **Bucketing the age into ranges will help to analyze “friendship\_initiated” along with gender – Multivariate- Gender, Age group & Count**
5. “like” & “mobile\_likes” can be combined to “total\_likes” and can be analyzed based on age group, -  **– Multivariate- Gender,Age group & Count**
6. **likes\_received (γ1 = 112.0745682), mobile\_likes\_received (γ1 = 107.5312999) & www\_likes\_received (γ1 = 126.257317) are highly Skewed, can we combine them together as “total\_likes\_received” – Multivariate- Gender,Age group & Count**
7. **Similar to age group, tenure can also be grouped by range of months “tenure\_range” as …..**

questions:-

* 1. what’s the percentage of people using FB mobile app & www

no. of people using mobile app based on “mobile\_likes” & “mobile\_likes\_received” can be used to understand the usage of FB mobile app

* 1. Who has more no. of friends? Male/Female?
  2. Who are the most active users based on age group, combining the likes and likes received
  3. Who have been using FB since long based on the tenure, combined with mobile and www
  4. Are people who are using FB since long, really active? Based on the likes
  5. Can something be done based on the user Ids, like count of user IDs with highest likes, friend requests etc.,
  6. What can be done on skewed values?
  7. Can anything be correlated? Like age and friend\_count
  8. Male and likes, female & likes?
* Maximum no. of users belong to age-groups between **18 to 24 years** & **25 to 35 years (around 50%).**
* Female users have more than friends and are interested in friendships (**more than 20%**) than Male and Other gender users.
* On an average users belonging to **Above 80 years** age-group have more no. of friends (**more than 50%**). But this needs to be thoroughly validated.
* On an average users belonging to **18 to 24 years** & **25 to 35 years** age-groups have more no. of friends compared (**more than 50%**). This figure seems more appropriate based on the no. of users.
* Since likes received are skewed, median of total likes received across age-groups shows that users **Above 80** years have more no. of likes received. But this needs to be thoroughly validated.
* On an average users belonging to **13 to 17** & **18 to 24 years** age-groups have more no. of likes given.
* Female users have more no. likes received and given.
* There are users with longest tenure more than 2000 days who seem to be still active.
* Comparison between usage of mobile and web applications based on categories, friend counts and friendships initiated, likes and likes received shows **around 80%** users using Facebook Mobile application.
* likes\_received is highly skewed (γ1 = 111.1536706)
* mobile\_likes\_received is highly skewed (γ1 = 106.6461839)
* www\_likes\_received is highly skewed (γ1 = 125.2173279)
* total\_likes\_received is highly skewed (γ1 = 111.1536771)
* df\_index has unique values
* userid has unique values
* friendships\_initiated has 1333 (1.4%) zeros
* likes has 20644 (21.2%) zeros
* likes\_received has 22764 (23.4%) zeros
* mobile\_likes has 33392 (34.3%) zeros
* mobile\_likes\_received has 28339 (29.1%) zeros
* www\_likes has 59335 (61.0%) zeros
* www\_likes\_received has 35200 (36.2%) zeros
* total\_likes has 20644 (21.2%) zeros
* total\_likes\_received has 22764 (23.4%) zeros

**10.1 Conclusion**

* **Home Teams** has won *more matches* **(46.2%)**, *scored more goals* **(57.6%)** than Away Teams
* **2nd half (59.9%)** has witnessed *more goals* than 1st half.
* Each season witnessed *more goals* in the **Home season** than Away Season.
* **3.51%** of matches has been won by Teams inspite of being behind during Half time
* **8.51%** of matches has been *Goalless*
* Only **6 teams** has participated in all 25 seasons from 1993-94 till 2017-18
* **Top 10 teams** has been identified, who can put themselves in the future seasons Top 10

**Reality Check**

* + **80%** of the Top 10 mentioned above has made to 2018-19 season
  + **60%** of the Top 10 mentioned above has made to 2019-20 season
  + **90%** of the Top 10 mentioned above has made to 2020-21 season (as on 05/Apr/2021)