R Language Programming

(CSC 522)

Project on

Data analysis of Facebook

By,

Priya Phapale

Student ID- 94662

Guided by,

Prof. Patricia Hoffman

TABLE OF CONTENT

INTRODUCTION

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. Now a days it is easy to get data from various sources but difficult to find useful data and relationship between different attributes in order to predict future. Many companies nowadays are investing money in studying past data in order to understand upcoming trends or predicting future and gaining profit in business. The process of analysis data consists of data gathering, reviewing, and then analyzing to form some sort of finding or conclusion.

Why I chose Facebook for analysis?
First reason is pretty obvious that Facebook is most popular social networking site and Second is data associated with it is huge. Study shows that the average American spends 2 hours each day social networking like Facebook, Instagram, YouTube etc. These two hours' worth of clicks, views, likes, shares and comments that all go into massive databases to be culled for further analysis. This data is used for better understanding of behaviors and improve the user experience, to create marketing profiles, to gather user census data, and a lot more. Not only do people spend a lot of time on social media, they seem to be comfortable revealing a lot of personal information.

OBJECTIVE

Objective of project is to visualize data and discover different relationships between the attributes, analyze information of users, number of pages you like, number of friend count, finding mean and median of it, keeping track on number of male and female friends, what is a age group of a person, likes received, correlation between variables and understand the different patterns by plotting graphs as per data provided. For many companies this analysis is very useful to make a profit in their respective businesses.

FLOWCHART OR STRATEGY

Steps for performing data analysis

1. Define why you need data analysis.
   - What are some ways to increase sales opportunities with our current resources?
   - Which products customer is interested?
   - Where and how to earn profit
2. Begin collecting data from sources.
   - Both structured and unstructured data that can be gathered from internal and external sources.
3. Clean through unnecessary data.
   - To generate accurate results, data scientists must identify and purge duplicate data, anomalous data, and other inconsistencies that could skew the analysis.
4. Begin analyzing the data.
   - Clustering analysis, association rule mining, which could unveil hidden patterns in data that weren't previously visible.
5. Interpret the results and apply them.
   - The final step is interpreting the results from the data analysis

PACKAGES AND FUNCTIONS

I am using different packages for facebook data analysis. My goal is create different plots using datasets like histograms, Scatterplots, Boxplots, Subsetting etc

1. **ggplot2**
   This package offers a powerful graphics language for creating elegant and complex plots. ggplot2 allows you to create graphs that represent both univariate and multivariate numerical and categorical data in a straightforward manner.
   The **qplot()** function can be used to create the most common graph types. While it does not expose **ggplot**'s full power, it can create a very wide range of useful plots.
2. **gridExtra**
   Provides a number of user-level functions to work with "grid" graphics, notably to arrange multiple grid-based plots on a page, and draw tables.
3. **reshape2**
   makes it easy to transform data between wide and long formats.
4. **GGally**
5. It extends 'ggplot2' by adding several functions to reduce the complexity of combining geometric objects with transformed data. Some of these functions include a pairwise plot matrix, a two group pairwise plot matrix, a parallel coordinates plot, a survival plot, and several functions to plot networks.
6. **Plyr**
   plyr is an R package that makes it simple to split data apar, and mash it back together. plyr makes it easy to control the input and output data format with a consistent syntax.
7. **Dplyr**
   dplyr is the next iteration of plyr, focused on tools for working with data frames identify the most important data manipulation tools needed for data analysis Provide fast performance for in-memory data. It allows us to use the same interface to work with data no matter where it's stored, whether in a data frame, a data table or database.

DATA SOURCE

For performing data analysis, I am using “pseudo_facebook” data set. This data is ideal dataset for studding relationships, rations, correlations between different numeric variables as well as categorical variables.

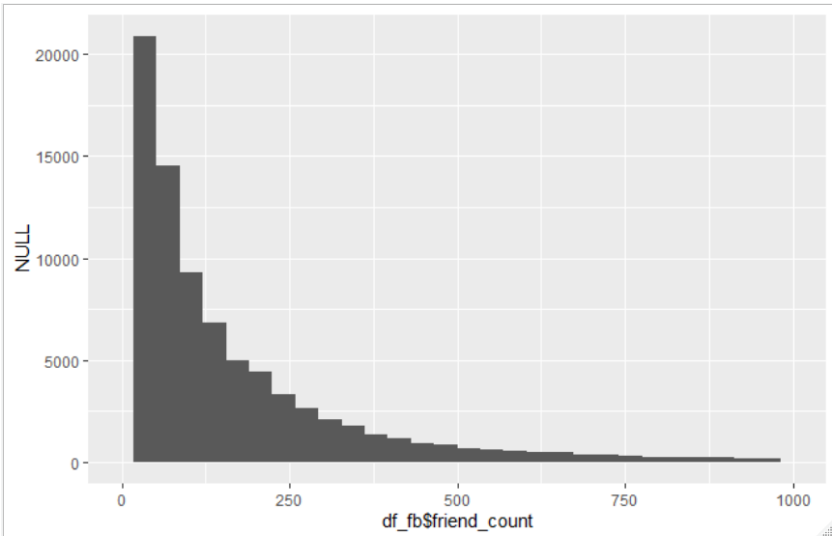
I have downloaded this data from internet. Data set consists of Userid, Age, DOBdate, DOBMonth, DOByear, Gender, Tenure, FriendCount etc.

userid	age	dob_day	dob_year	dob_month	gender	tenure	friend_count
2094382	14	19	1999	11	male	266	0
1192601	14	2	1999	11	female	6	0
2083884	14	16	1999	11	male	13	0
1203168	14	25	1999	12	female	93	0
1733186	14	4	1999	12	male	82	0
1524765	14	1	1999	12	male	15	0
1136133	13	14	2000	1	male	12	0
1680361	13	4	2000	1	female	0	0
1365174	13	1	2000	1	male	81	0
1712567	13	2	2000	2	male	171	0
1612453	13	22	2000	2	male	98	0
2104073	13	1	2000	2	male	55	0
1918584	13	5	2000	3	male	106	0
1704433	13	21	2000	3	male	61	0

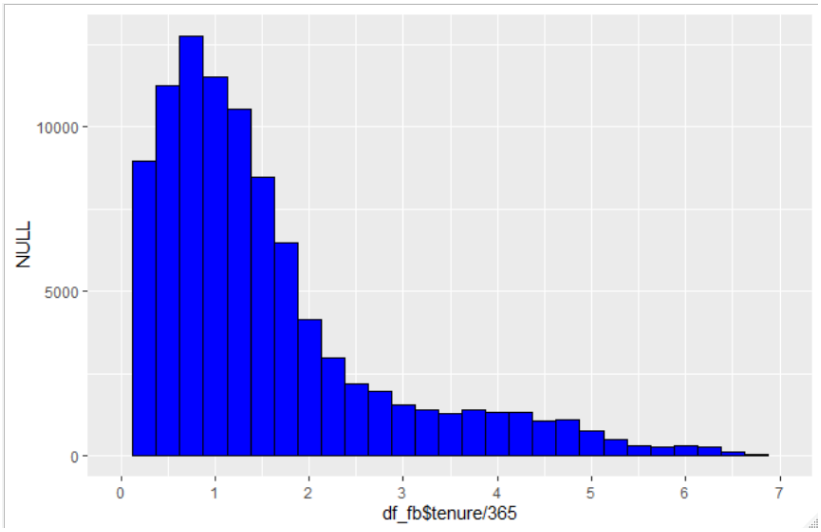
ACTUAL ANALYSIS USING R LANGUAGE

Below are some examples of data analysis of facebook using different plots

1. histogram of friend count



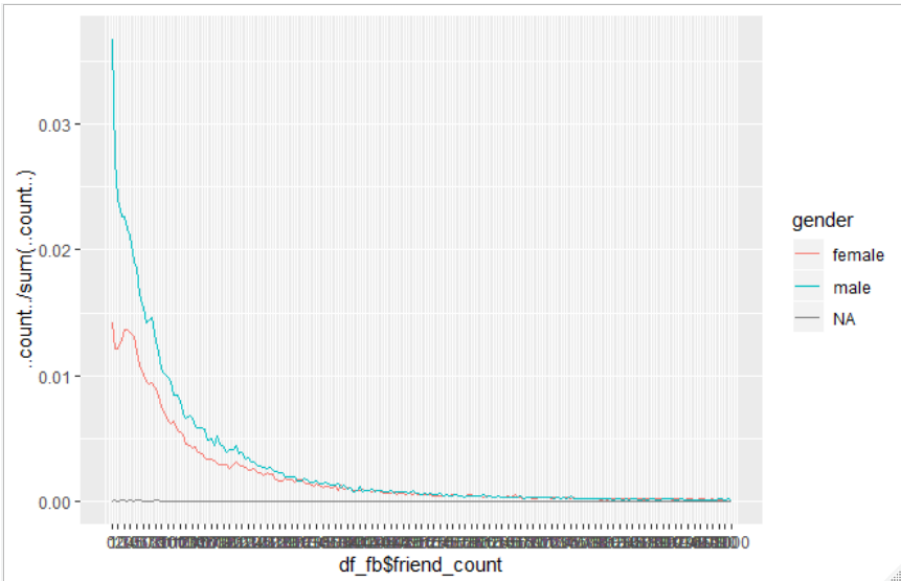
2. Histogram of tenure on Facebook (number of days; tenure/365)



3. Summary of friend count

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -Inf   3.434   4.407    -Inf   5.328   8.502
```

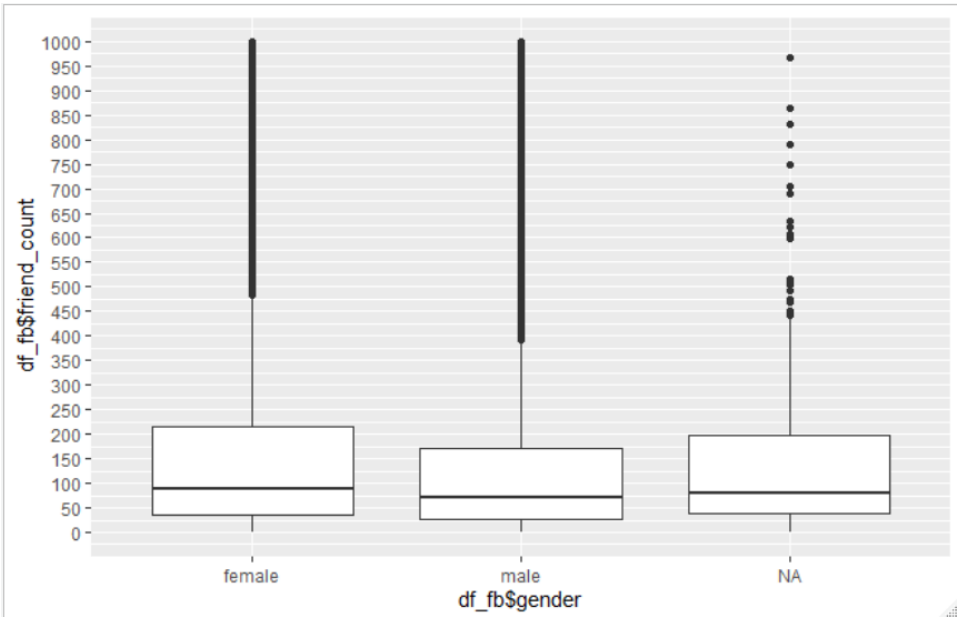
4. Plot for friend count vs relative count, colored by gender



5. Subsetting

```
> # subset of facebook dataframe, with only data about males
> df_males <- df_fb[df_fb$gender == "male",]
> str(df_males)
'data.frame':   58749 obs. of  15 variables:
 $ userid               : int  2094382 2083884 1733186 1524765 1136133 1365174 1712567 1612453
 2104073 1918584 ...
 $ age                  : int  14 14 14 14 13 13 13 13 13 13 ...
 $ dob_day              : int  19 16 4 1 14 1 2 22 1 5 ...
 $ dob_year             : int  1999 1999 1999 1999 2000 2000 2000 2000 2000 2000 ...
 $ dob_month            : Factor w/ 12 levels "1","2","3","4",..: 11 11 12 12 1 1 2 2 2 3 ...
 $ gender               : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
 $ tenure               : int  266 13 82 15 12 81 171 98 55 106 ...
 $ friend_count         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ friendships_initiated: int  0 0 0 0 0 0 0 0 0 0 ...
 $ likes                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ likes_received       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ mobile_likes         : int  0 0 0 0 0 0 0 0 0 0 ...
```

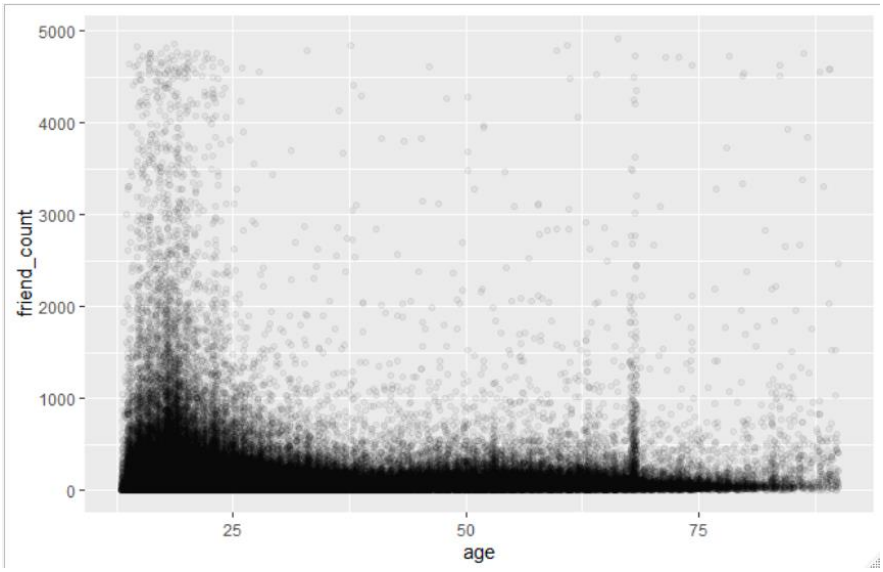
6. Boxplot of friend count vs gender



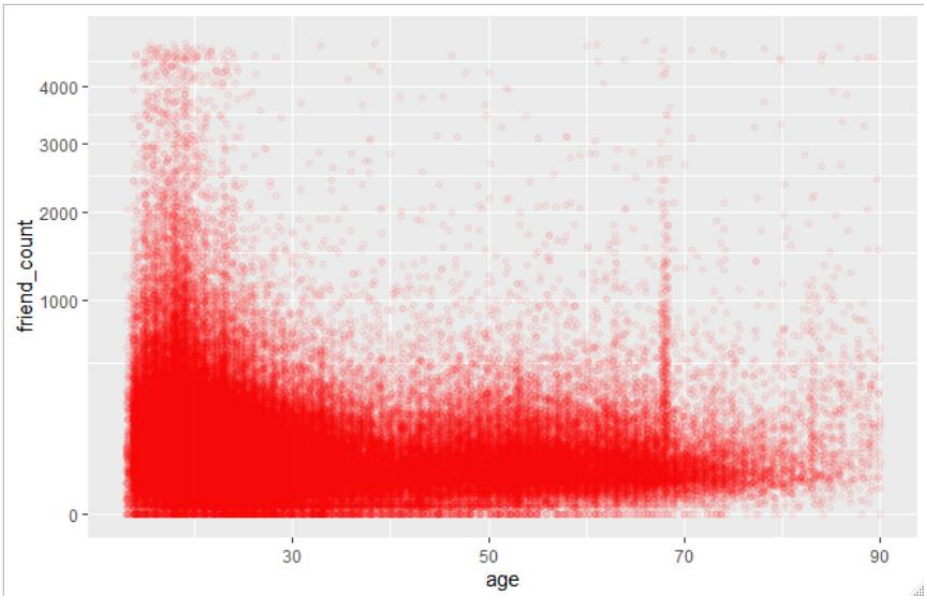
7. Creating column for mobile check ins based on mobile likes

```
> # creating column for mobile check ins based on mobile likes
> df_fb$mobile_checkin <- 0
> df_fb$mobile_checkin[df_fb$mobile_likes > 0] <- 1
> df_fb$mobile_checkin <- as.numeric(df_fb$mobile_checkin)
> summary(df_fb$mobile_checkin)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.6459  1.0000  1.0000
```

8. Using geom_jitter- histogram of age vs friend count



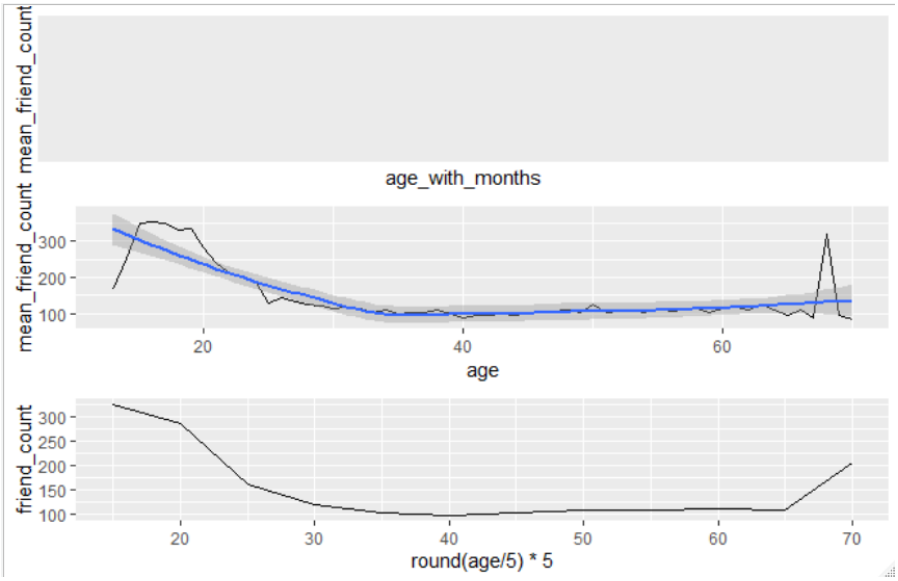
9. Scatterplot of age vs friend count



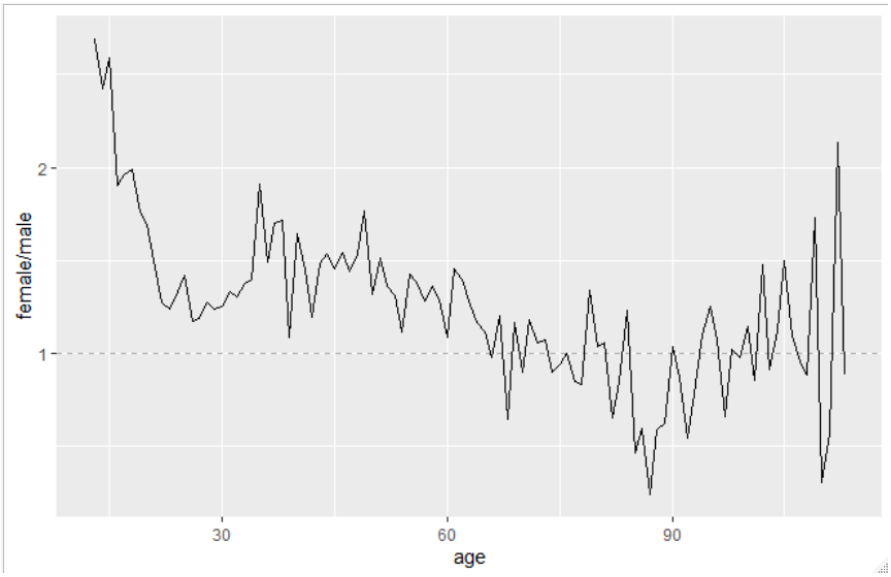
10. Correlation co-efficient between age and friend count

```
> # correlation co-efficient between age and friend count using pearson correlation coefficient
> cor.test(df_fb$age, df_fb$friend_count, method="pearson")

        Pearson's product-moment correlation

data:  df_fb$age and df_fb$friend_count
t = -8.6268, df = 99001, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03363072 -0.02118189
sample estimates:
        cor
-0.02740737
```

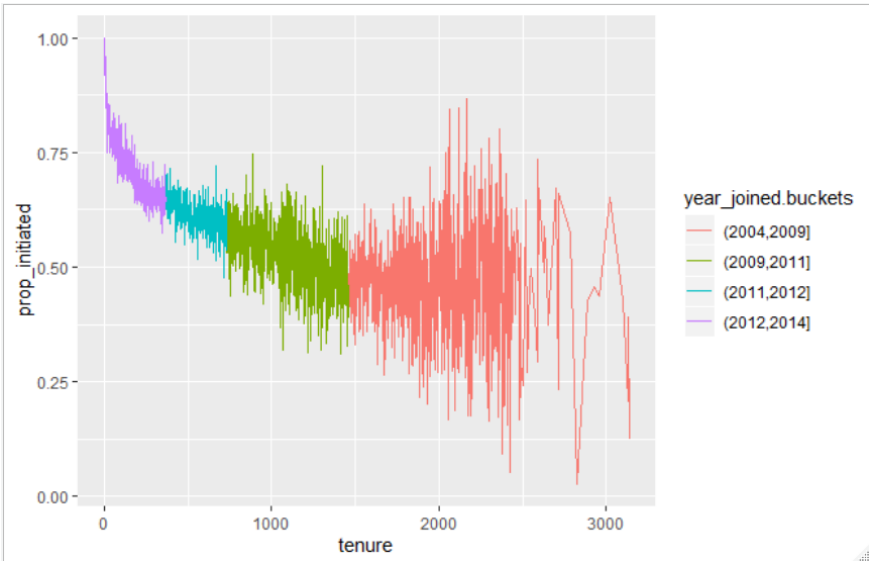
11. Create new dataframe that has mean and median friend counts for each age in months



12. Ratio of friend_counts of both genders



13. Friendship proportion vs tenure on facebook



CONCLUSION

There are many more other packages, libraries and functions we can use to make deep data analysis. R language provides a hug set of packages, built in functionality which can be used for creating different patterns that will help to predict future a by analyzing past data. R language is considered to be the best programming language for any statistician as it possesses an extensive catalog of statistical and graphical methods.

REFERANCES

1. https://www.r-project.org/about.html
2. http://www.businessdictionary.com/definition/data-analysis.html
3. https://www.makeuseof.com/tag/what-is-data-analysis/
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3553267/