“The Chatting Traders” Project

CMPSC 310: An Introduction to Data Science - Spring, 2019

By Thao Phuong, Bakhytzhan Altynbayev

Abstract

Nowadays, communication is a crucial component of everyday life. It is one of those skill sets that might possibly bring people closer to success. While pursuing daily goals to success, traders often utilize their communication skills in winning the daily economical battles. This project analyses four files that describe various communication models used by traders in ForEx trading system. It analyzes scenarios of what possible insights the data can provide us. What are some types of discussions that the most popular post chatters are talking about? Could we find the hidden success formula of Warren Buffett? Let’s dive in and see.

Introduction

The following four files describe the communication between traders in ForEx trading system:

1. users.tsv - contains unique user ids and account creation dates;
2. messages.tsv - contains unique message ids, send dates, sender ids (consistent with those in 'users.tsv'), and message types;
3. discussions.tsv - contains unique discussion ids, creation dates, creator ids (consistent with those in 'users.tsv'), and discussion categories;
4. discussion_posts.tsv - contains unique post ids, discussion ids (consistent with those in 'discussions.tsv'), and creator ids (consistent with those in 'users.tsv')

By combining, merging, concatenating, grouping, ordering, sorting and ranking mentioned above files we will be on the way to answers on the following interesting questions:

1. Simple descriptive statistics:
    a. How many users are in the database?
    b. What is the time span of the database?
    c. How many messages of each type have been sent?
    d. How many discussions of each type have been started?
    e. How many discussion posts have been posted?
2. What is the distribution of activity ranges?
3. What is the distribution of message activity delays in EACH category?
4. What is the distribution of discussion categories by the number of posts? What is the most popular category?
5. What is the distribution of post activity delays in the most popular category?

6. A box plot with whiskers that shows all appropriate statistics for message activity delays in EACH category, post activity delays, and activity ranges.

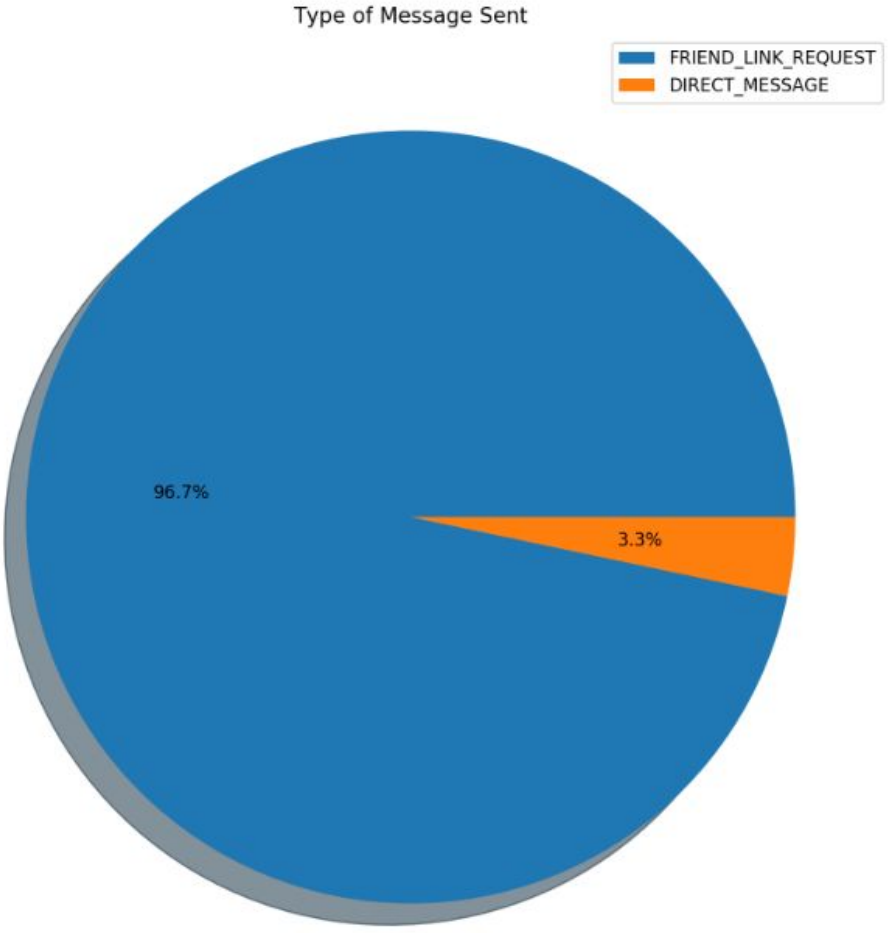
Methods for data acquisition and processing

- Obtain the needed raw data, namely four files;
- Use pandas module to import files and turn them into DataFrame for further analysis;
- Processing the raw data by: using pandas software library and techniques including merging, concatenating, sorting (ordering), grouping, aggregating.
- Present the results in form of a plain number, pie chart, histogram, box plot (the form depends on each question' delivery instruction).

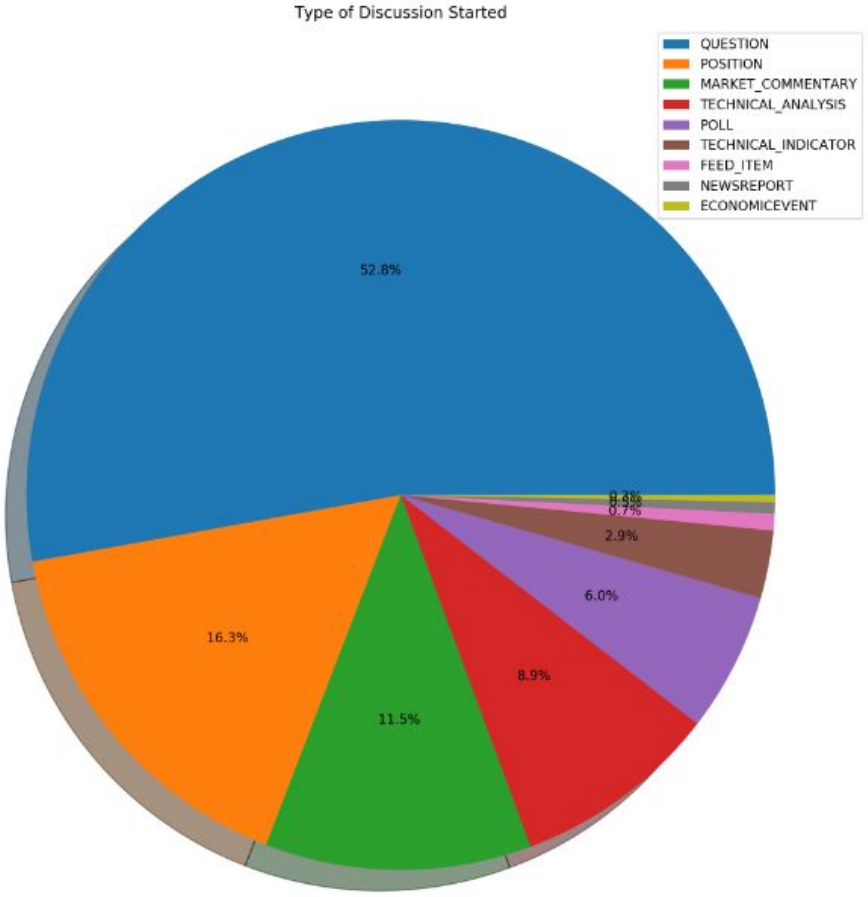
Results

Part I: Simple descriptive statistics

1. How many users are in the database? 53103 users.
2. What is the time span of the database? 1778 hours.
3. How many messages of each type have been sent?

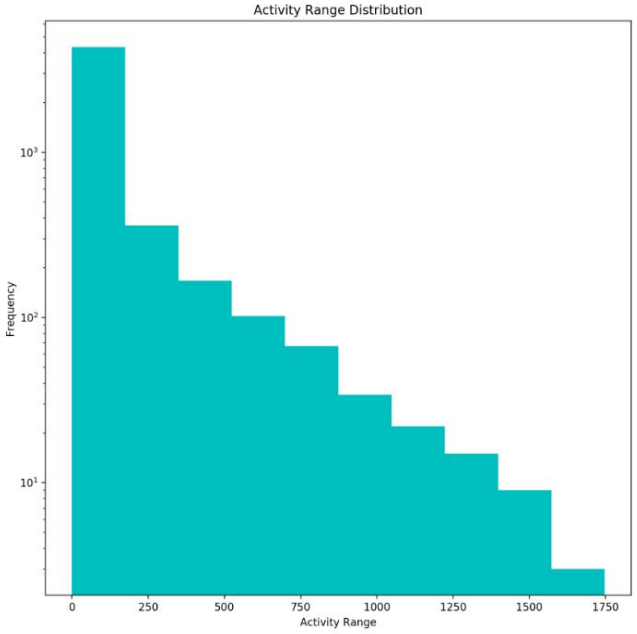


4. How many discussions of each type have been started?

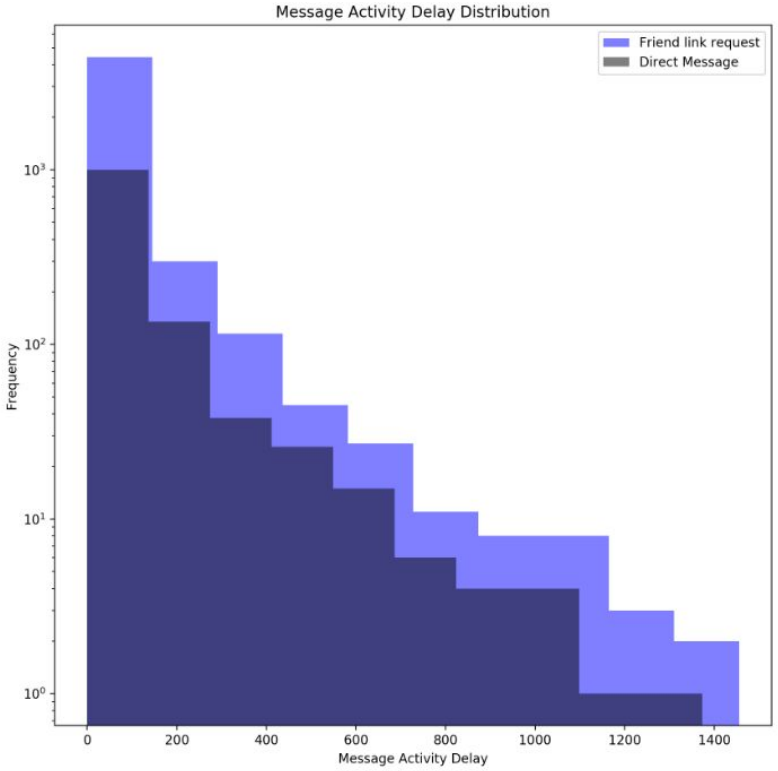


5. How many discussion posts have been posted? 1980 discussions.

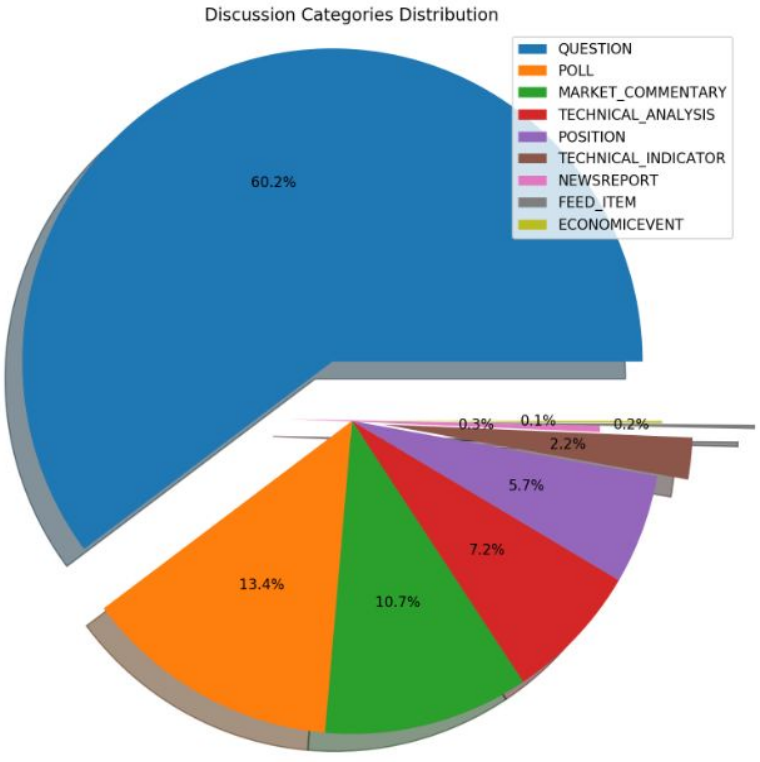
Part II: Activity range



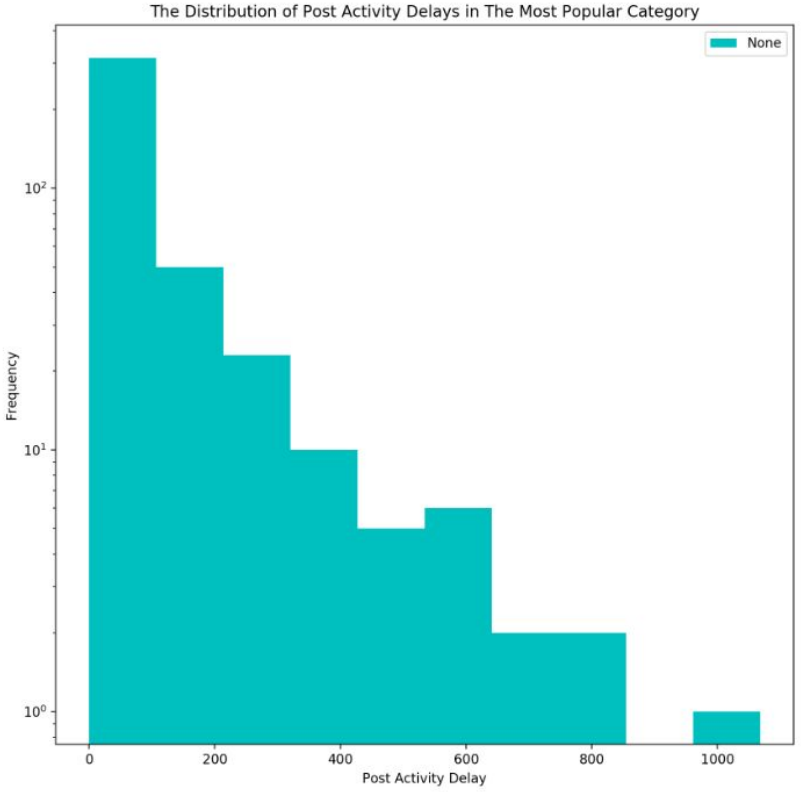
Part III: Message activity delay



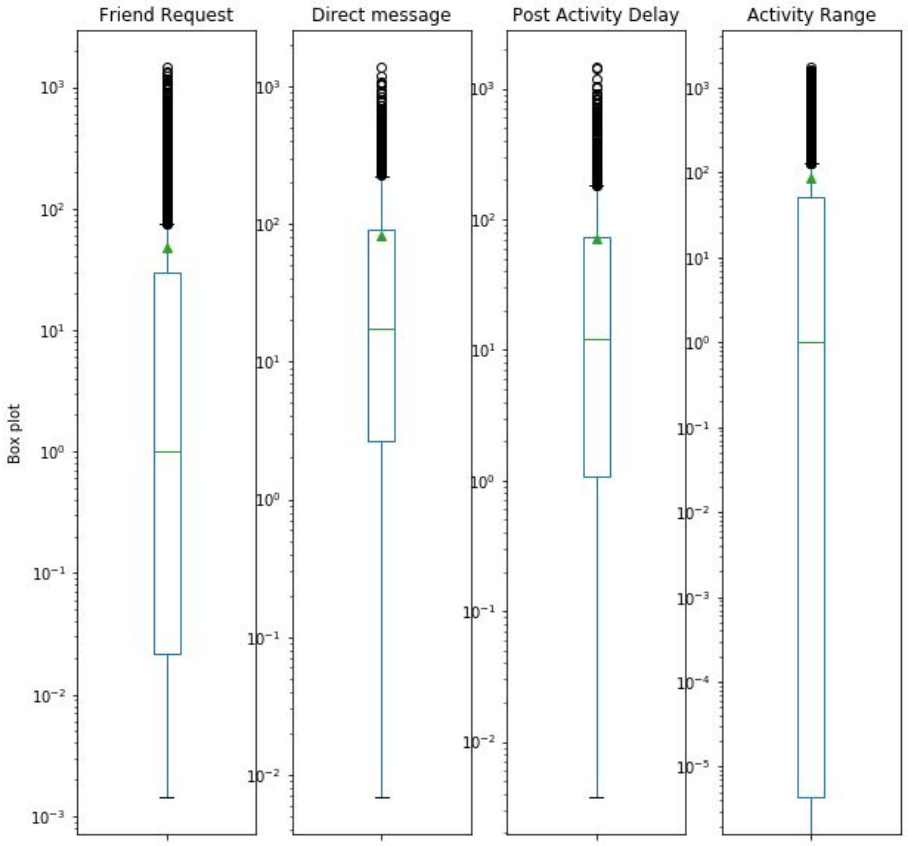
Part IV: Discussion categories by number of posts



Part V: Post activity delay



Part VI: Box plot



Conclusion

By using build-in Pandas function, we were able to extract and tell that the DataFrame contains a significant number of 53103 users.

Being able to combine all four files into a Single DataFrame, we were able to find the most maximums and minimums of time related numeral in it. Thus, it seems that the time span of database is 1778 hours.

By accessing the elements of message DataFrame we've located and counted the amount of messages sent (divided by their type). Traders seem to be a very friendly community, showing 96.7% of message types being a friend request. The remaining 3.3% stood for direct messaging.

In a similar way to finding and counting the types of messages, we've analysed a discussion DataFrame to see how many discussions of each type were sent by traders. The project shows that more than half of the discussion topics were raised in form of "question" type. On the second most-ranked position we could see "polls" raised by traders. The third most discussed topic was "market commentaries" provided by traders.

By accessing the post DataFrame, we've searched and counted the most unique discussion posts that we could find. Overall, there are 1980 discussion posts opened on various topics of trading industry.

While measuring the activity range (between the first and last messages in any category), the results have shown that the most frequent range was between 0 and 250 days.

Similar to measuring the activity range, the results have shown that the most frequent message activity delay was between 0 and 200 days (for both: friend link request and direct message).

Similar to the research question in Section 1.4 (from Simple Descriptive Statistics) the most discussed topic (in discussion categories distribution) was "question" with 60.2%. Second leading position was upon "poll" with 13.4%. The third ranked topic was "market commentary" with 7.2% from the total.

By looking at the distribution of post activity delay (in the most popular category) we could see that the most frequent delay was between 0 and 100 days.

Lastly, a box plot with whiskers lets us see all appropriate statistics for message activity delays in each category.