**MTEch KE5106**

**data warehousing for business analytics**

**Project report**

# Singapore Hotel Reviews: Data Collection, Analysis and Insights

**teAM MEMBERS**

BHARAT NAGARaju-A0178258N

Chan Su-Wen Philemon - A0110588E

GUO LEI – A0059762U

Vigneshram selvaraj - A0178215A

# 1.0 introduction

Tourism is one of the major industries in Singapore, attracting 17.4 millions of internal tourists in 2017, more than 3 times of its total population[[1]](#footnote-2). It is interesting to find out the most preferred hotel(s) in Singapore and understand the customers’ profile and sentiments. Relying on reviews on two most popular online review forums, Trip Advisor and Yelp, first, we aim to understand the profile of Singapore hotel reviewers. Secondly, we will recommend most liked hotel(s) based on data analysis. Lastly, we will help the websites identify lead users and establish correlations among their behavioral attributes.

Both Trip Advisor and Yelp were selected because of their popularity and the purpose of cross validation. Following that, we chose six hotels in Singapore, representing the wide range of service offerings from the airport hotel (e.g. Crown Plaza Hotel), family hotels (e.g. Village Hotel) to luxury hotels (e.g. Marina Bay Sands). The rest of the report is structured as follows. We start with data collection methods, and technical approach and processes. Next, data exploration and analysis results are presented. We conclude this report with reflections and recommendations.

# 2.0 Data collection

Data scraped from Trip advisor and Yelp were done with the help of python libraries such as Beautiful soup to obtain data from HTML files. This is done via traversing the HTML parse tree to search for key HTML tags that contain hotel reviewer and hotel information. Textblob python library was also used to process text data for the use of analyzing the sentiments in hotel reviews obtained from Trip Advisor and Yelp. The data was then ingested to document database, MongoDB, which enables ease of retrieval of data in JSON format.

# 3.0 Technical approach

Review and User details from Trip Advisor has been scraped using Python – BeautifulSoup, Request and LXML packages. Scraped data from Trip Advisor and Yelp are stored in NoSQL document database MongoDB using Pymongo into four main collections – User and Reviews for the ease of analysis and maintenance and two each for yelp and trip advisor.. Each website has their own set of user and review collections (group of documents) due to the differences in data fields extracted from the websites. Also user data is kept in a separate table due to the frequent change of data and to avoid the cost of frequent updates.

We used PyMongo, a python library to work with MongoDB, to retrieve data from the respective collections in the database. Aggregation pipeline stages were prototyped using Studio 3T (Integrated Development Environment for MongoDB) and the resultant pipeline was incorporated into the python code. For example, the simple two stage aggregation pipeline to get the list of cities that the TripAdvisor reviewers reside in for the data of hotels in Singapore are as follows:

def cityOrigin():

cities = users\_collection.aggregate( [

{ "$group" :

{ "\_id" : "$user\_loc\_list", "count" : { "$sum" : 1.0 } } },

{ "$sort" : { "count" : -1 } } ], );

return cities

Following which, the aggregated data can be visualized or further processed in Tableau or in Python Pandas with visualization from python libraries such as seaborn and matplotlib.

Also, User Reviews were analyzed using Natural Language Toolkit NLTK & Wordcloud libraries.

Words like “hasn’t” , “lemme” and colloquial words were replaced with proper English words like has not, let me. Using NLTK, Comments were tokenized and stripped of Stopwords(Prepositions, Fillers, etc.).Resultant array of words were stemmed and lemmatized. The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance:  
am, are, is $\Rightarrow$ be   
car, cars, car's, cars' $\Rightarrow$ car  
The result of this mapping of text will be something like:  
the boy's cars are different colors $\Rightarrow$   
the boy car be differ color

A word cloud and frequency graph was plotted to understand the most frequently described comments. This technique was applied on positive and negative feedback separately in order to obtain detailed insights separately. All Ratings 3 and above are classified as positive while anything below three was considered negative.

# 4.0 analytics

**1. Exploratory Analysis: Profile of Singapore Hotel Reviewers**

First, we conducted data cleansing, preparation and exploratory analysis to understand the profile of those reviewers. After extensive effort, we cleansed the raw data sets, recoded variables into the right types (e.g. from categorical to numerical), and created new variables (e.g. grouping numbers of helpful votes and user contributions into categories) for analytical purposes. Also, we removed duplications and missing values for better data quality and validity.

We studied the trend of users joining Trip Advisor. **As indicated in Figure 1, it seems a decreasing trend in recent year with less number of users joining Trip Advisor** **and a sharp fall in 2018**. A possible explanation is that the company has passed the fast growing stage, from 2012 to 2015, and now it is at the mature stage. Figure 1.1 further explains this downward trend, e.g. lesser people joining Trip Advisor in 2018 compared to that of 2017.

After massive data cleansing, we managed to categorize the reviewers according their locations. **Figure 6.1 indicated that most of the reviewers are from UK, Australia, US, and Singapore.** **Most of them from UK, Australia, US, and Singapore. Hotels can win the customers by providing facilities famous among citizens of these countries.**

**Among Singapore users, the distribution from Level 1 to Level 6 varies from 14% to 22%, which indicated the majority of Singapore users were experienced reviewers shown in Figure 6.2. As shown in Figures 2 and 3, more than half, or 53%, of the reviews are active contributors who are at Level 4 and Level 5. Accordingly, the average number of cities visited by them is increasing across different levels with 8 cities by Level 1 but 168 cities by Level 6.** Moreover, the most frequent number of cities visited by Level 1 was 3 cities, but that of Level 3 was 8 cities, shown in Figure. 5.1 and 5.2. Following that, we investigated the relationship between reviewers’ contribution and their respective levels. Figure 3 shows that about one third of Level 6 reviewers contributed more than 100 reviews, whilst none of the users at Level 1, 2 and 3 have contributed 100 reviews. **The finding of helpful votes by level is consistent, as presented in Figure 4, users of the higher levels, 4 and above, tend be more helpful than those who are at the lower levels.**

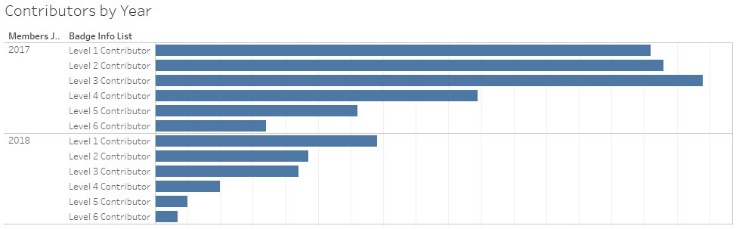
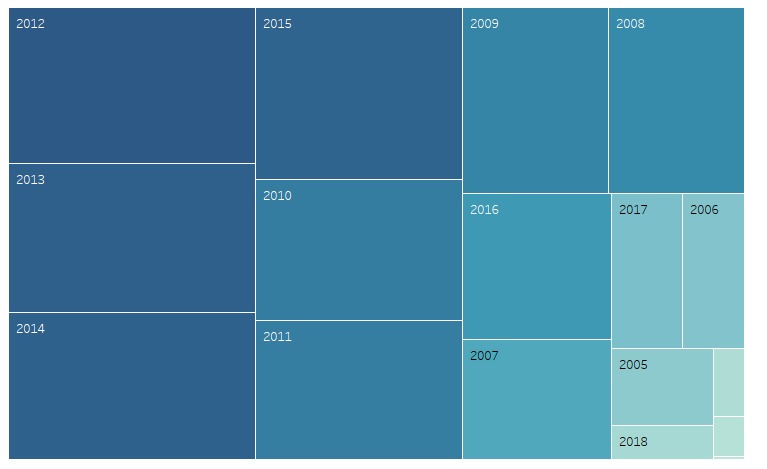


Figure 1: Members Joined by Year Figure 1.1: Members Joined by Year

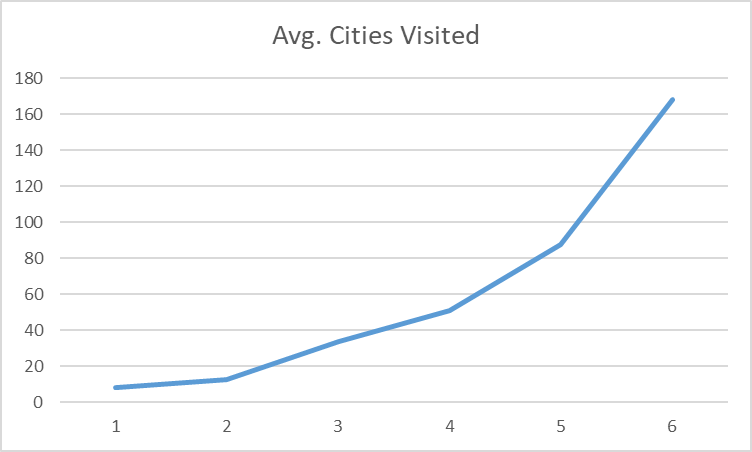
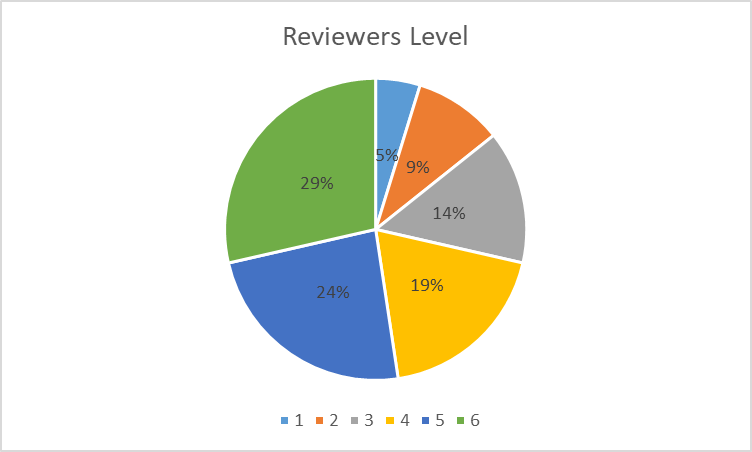


Figure 2: Reviewer by Levels Figure 3: Average Cities Visited by Levels

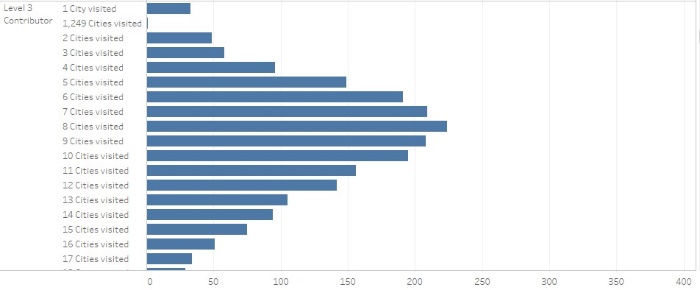
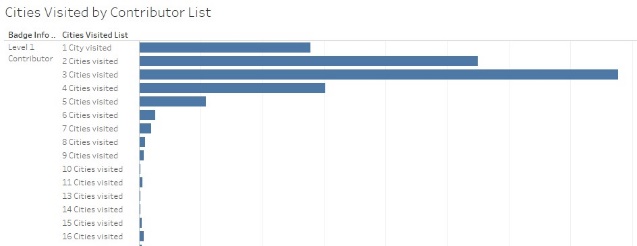


Figure 4.1: Number of Cities Visited by Level 1 Figure 4.2: Number of Cities Visited by Level 3

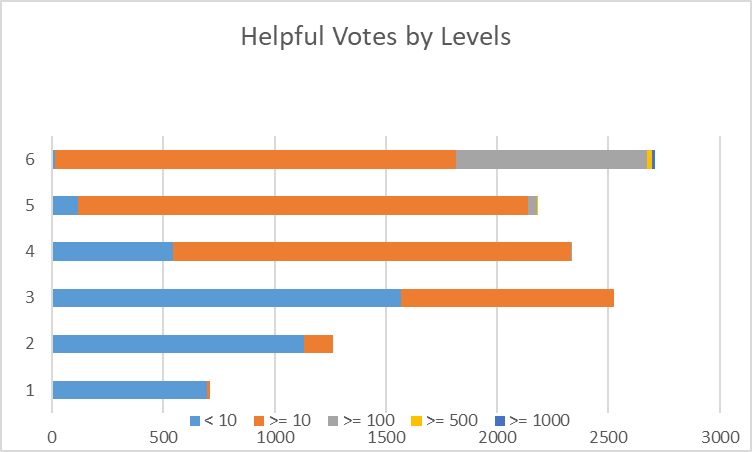
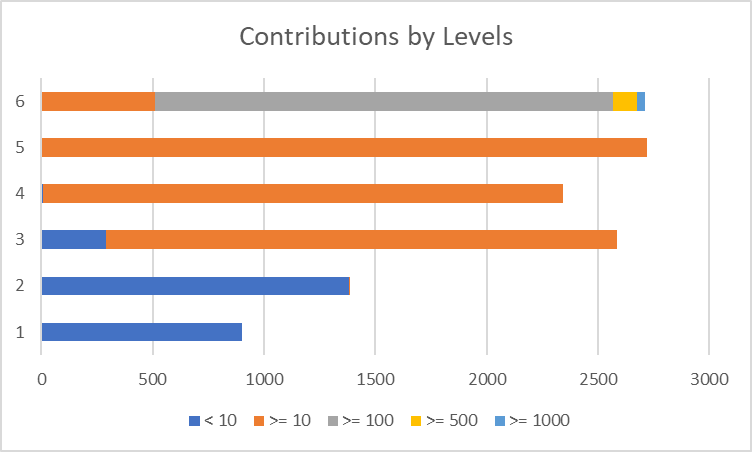


Figure 5.1: Contributions by Levels Figure 5.2: Helpful Votes by Levels

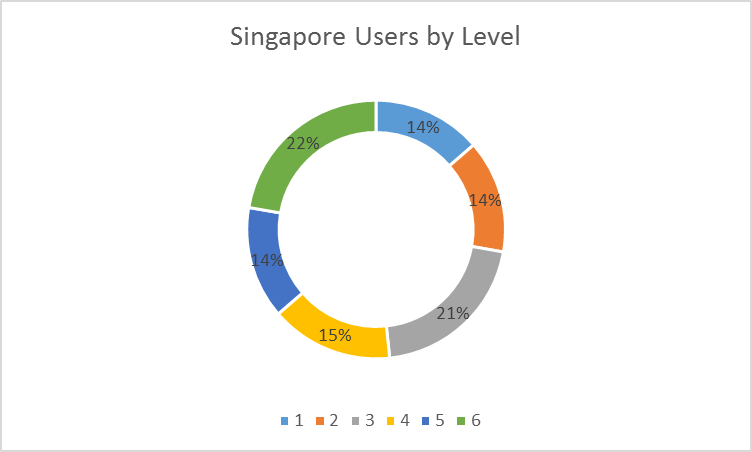
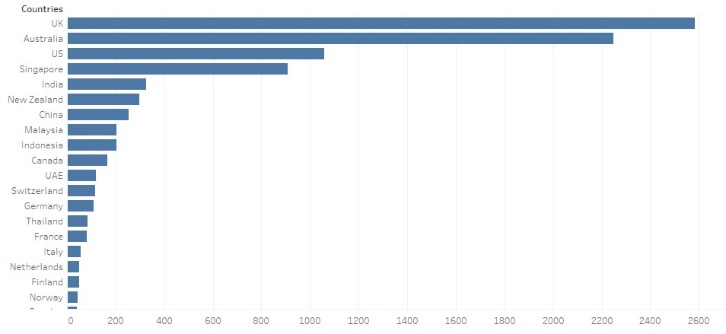
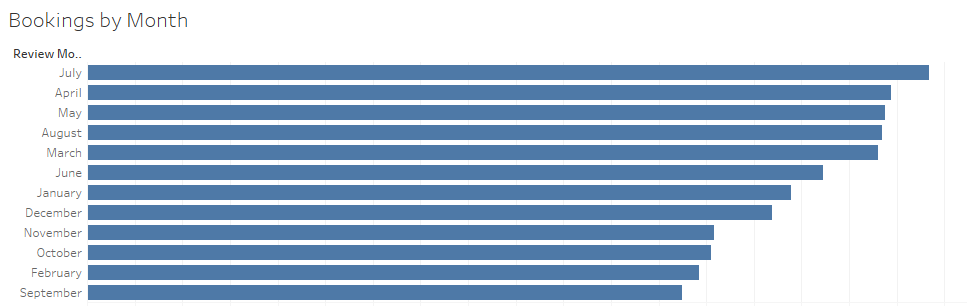


Figure 6.1: Visitor’s Countries Figure 6.2: Singapore Users by Level

**2. Business Insights: The Most Liked Hotel(s) in Singapore**

**Hotel Bookings are in peak in July and surprisingly not in December. In Oct, Nov, Feb and Sep the bookings drop by 30 %. Hotels can provide promotions during this month to attract customers. This is shown in the below figure.**

****

As explained in the Introduction section, we selected 6 hotels for comparison, namely Crowne Plaza Changi Airport, Marina Bay Sands, Oasia Hotel, Park Hotel, Shangri La and Village Hotel. There are a total of more than 18,000 reviews for all six hotels, the number of reviews were distributed across them, from 12% to 21% (see Figure 7). We conducted sentiment analysis of the feedbacks, as shown in Figure 8 & 9, the positive sentiments were centering around ‘location, service and experience’, on the other hand, negative sentiments were about ‘price, service and disappointment’. At the hotel level, Park Hotel received more positive feedback than others.

**However, it is still unclear which hotel was the most preferred one by reviewers. We further looked into numerical rating. Interesting, the average rating of each of the hotels was very similar, 4.38 out of 5. In order to generate business insights, we compared the lowest rating (i.e. 1) and the highest rating (i.e. 5) across six hotels. The result indicated that Shangri la has received a much higher percentage of higher rating (see Figure 11).**

As such, we concluded that Shangri La and Park Hotel were most liked by the reviewers.

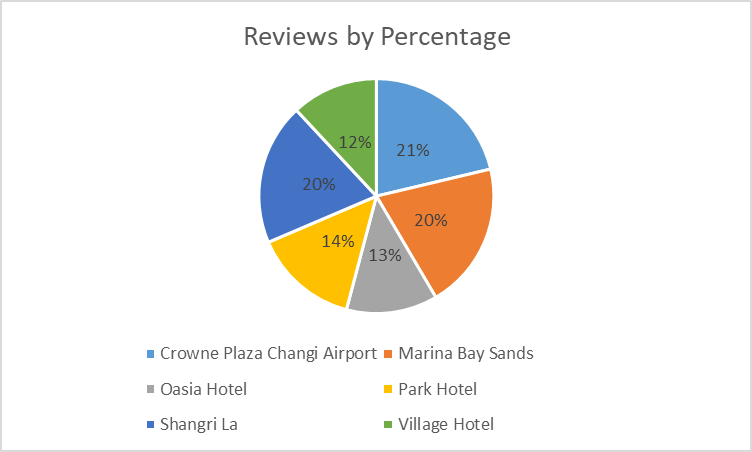


Figure 7: Hotel Reviews Distribution Figure 8: Word Cloud of Positive Sentiment

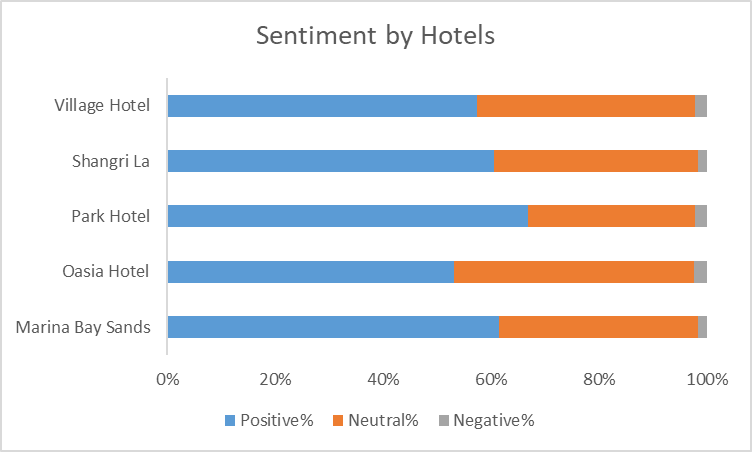


Figure 9: Word Cloud of Negative Sentiment Figure 10: Sentiment by hotels

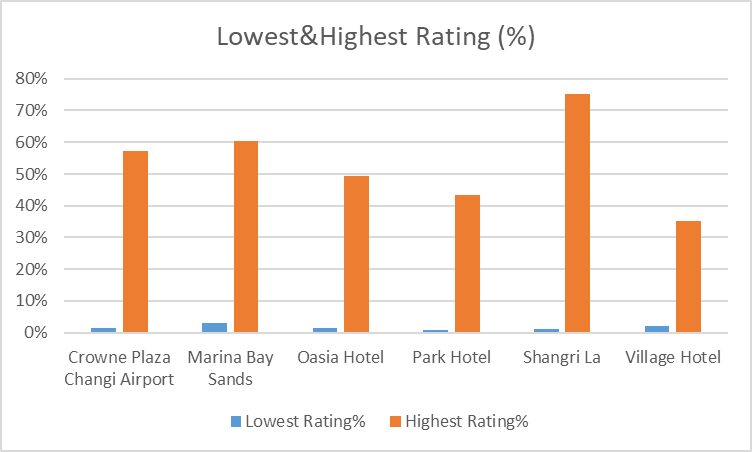
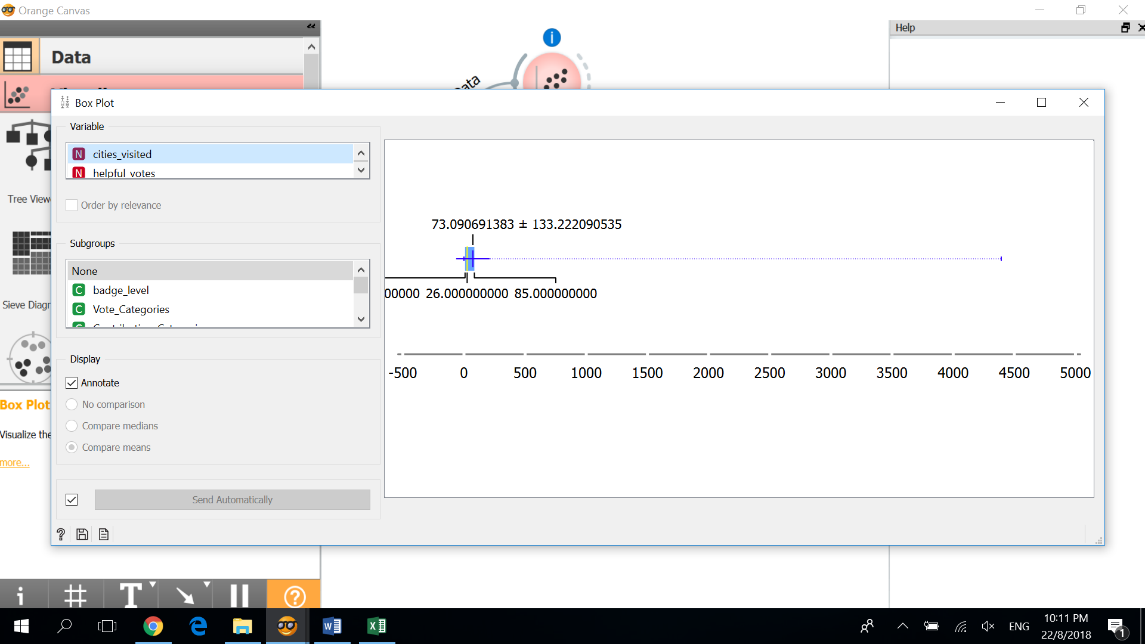
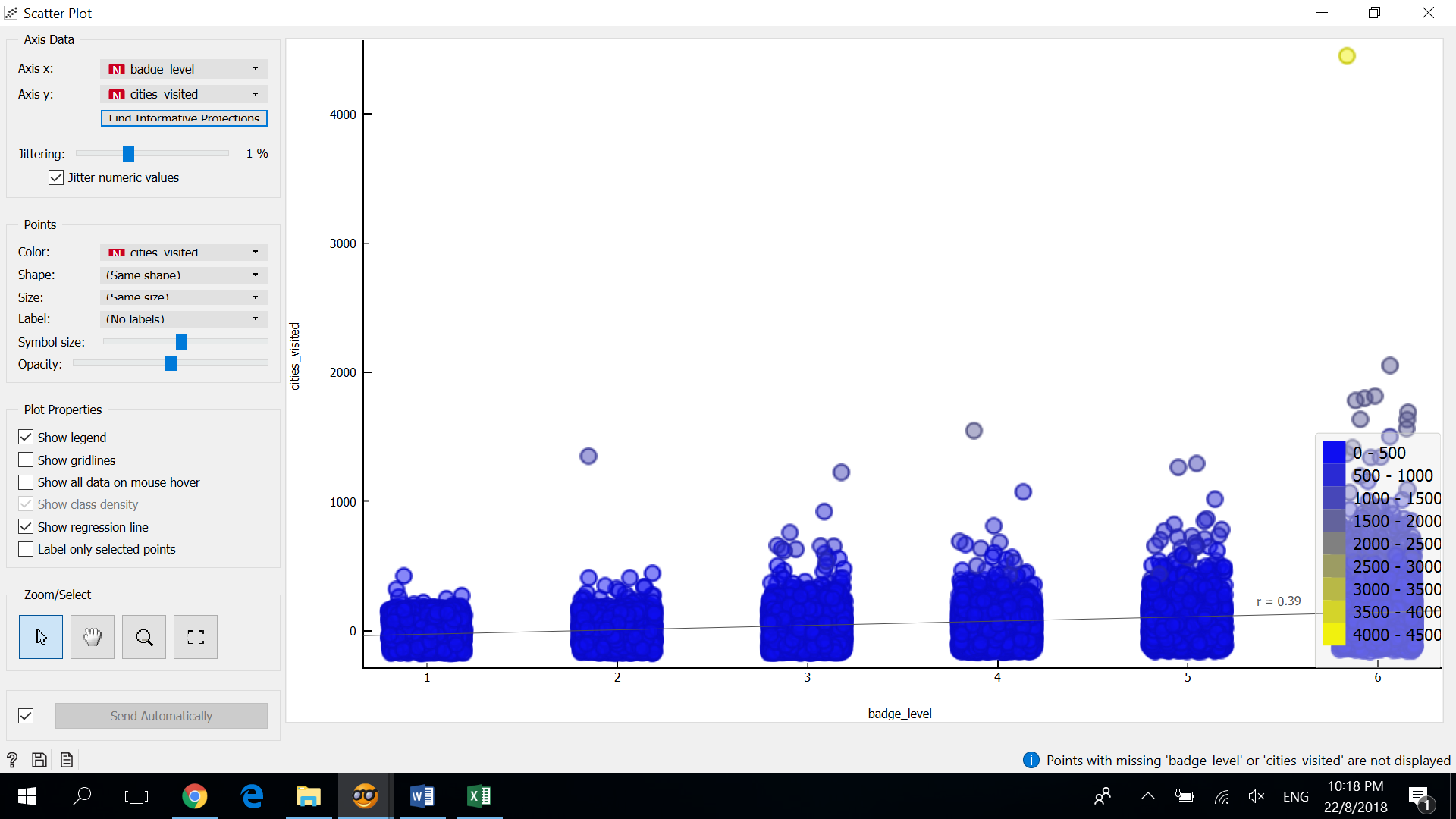


Figure 11: Percentage of Lowest and Highest Ratings

**3. Further Findings: Lead Users and Behavioral Attributes**

We used Boxplot and Scatter Plot to identify lead users whom Trip Advisor may want to reward and motivate their behavior. **We filtered the outliers from variables of ‘cities visited’, ‘contributions’ and ‘helpful votes’ from each of the six user levels, see Figure 12 to Figure 14. By identifying those lead users, Trip Advisor will be able to provide personalized service, rewards or recognitions to them.**

Finally, we examined the correlations between ‘cities visited’, ‘contributions’ and ‘helpful votes’. The results were presented from Figure 15 to Figure 17, it shows that helpful votes were strongly correlated to contributions but no correlation with cities visited. There was a moderate correlation between cities visited and contributions. As such, the results suggested that the increasing number of cities visited by the reviewers is not necessarily leading to their contribution to the Trip Advisor community.

  
Figure 12: Outlier of Cities Visited Figure 12.1: Outlier of Cities Visited by Levels

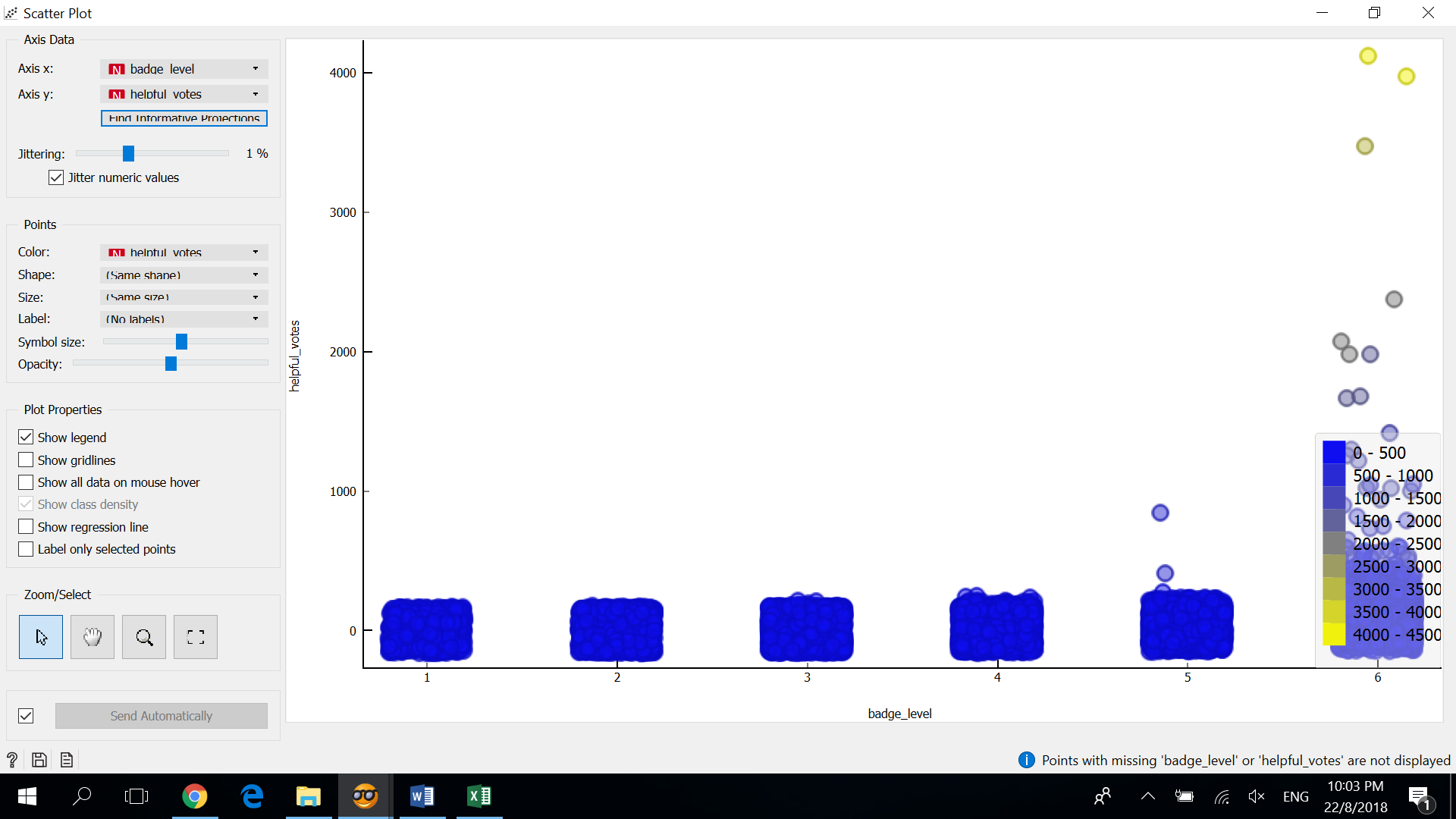
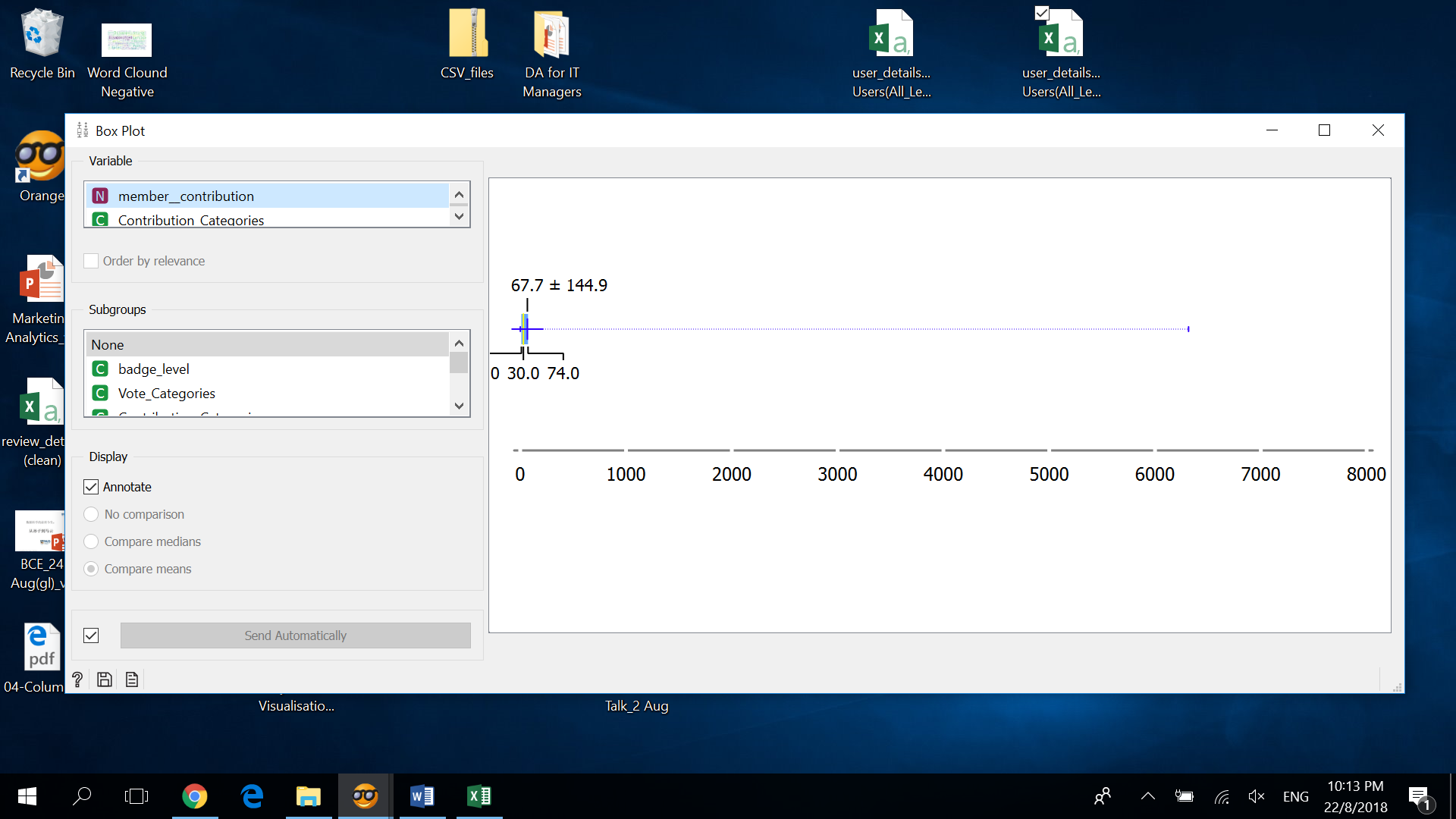


Figure 13: Outlier of Contributions Figure 13.1: Outlier of Contributions by Levels

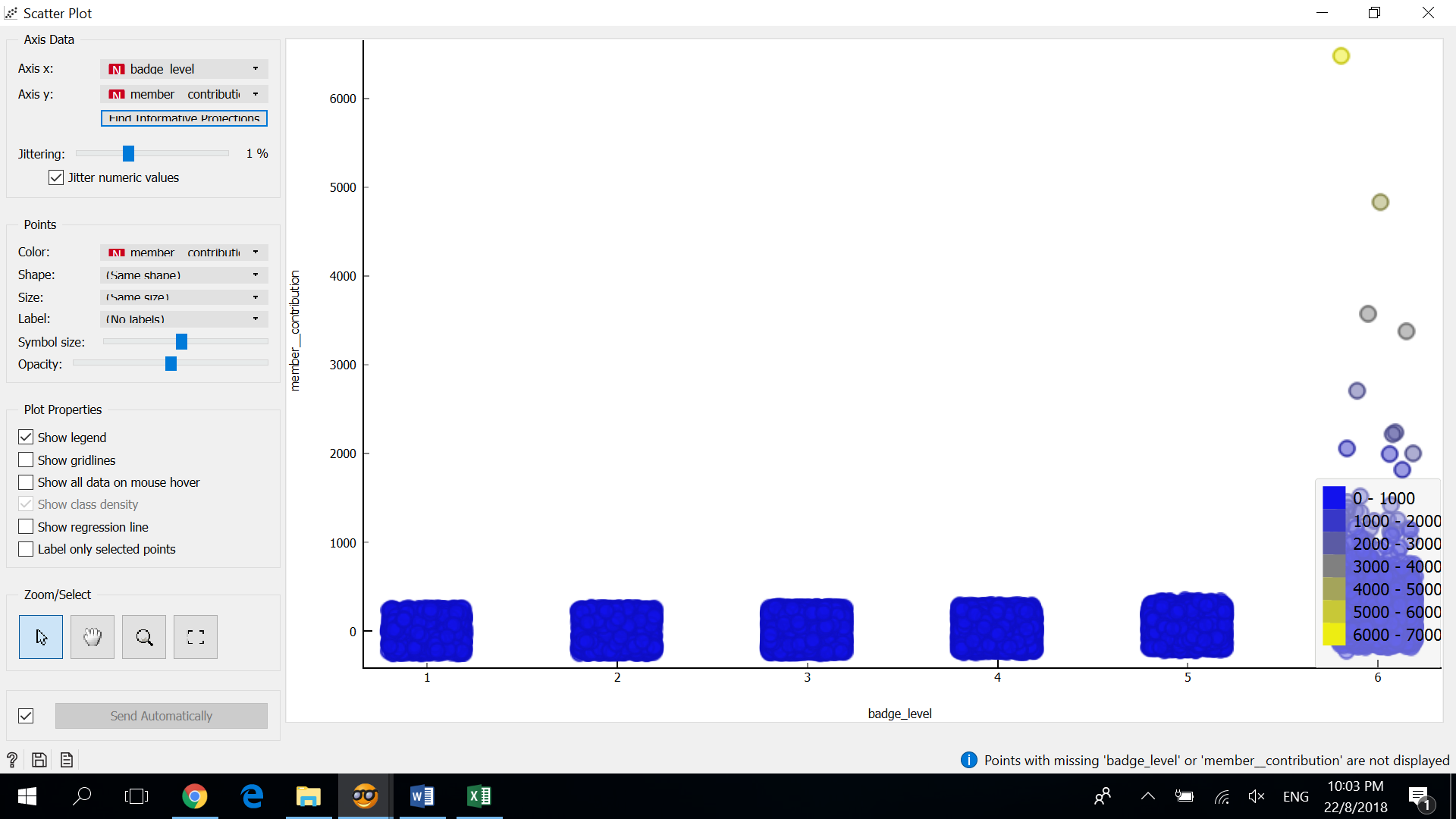
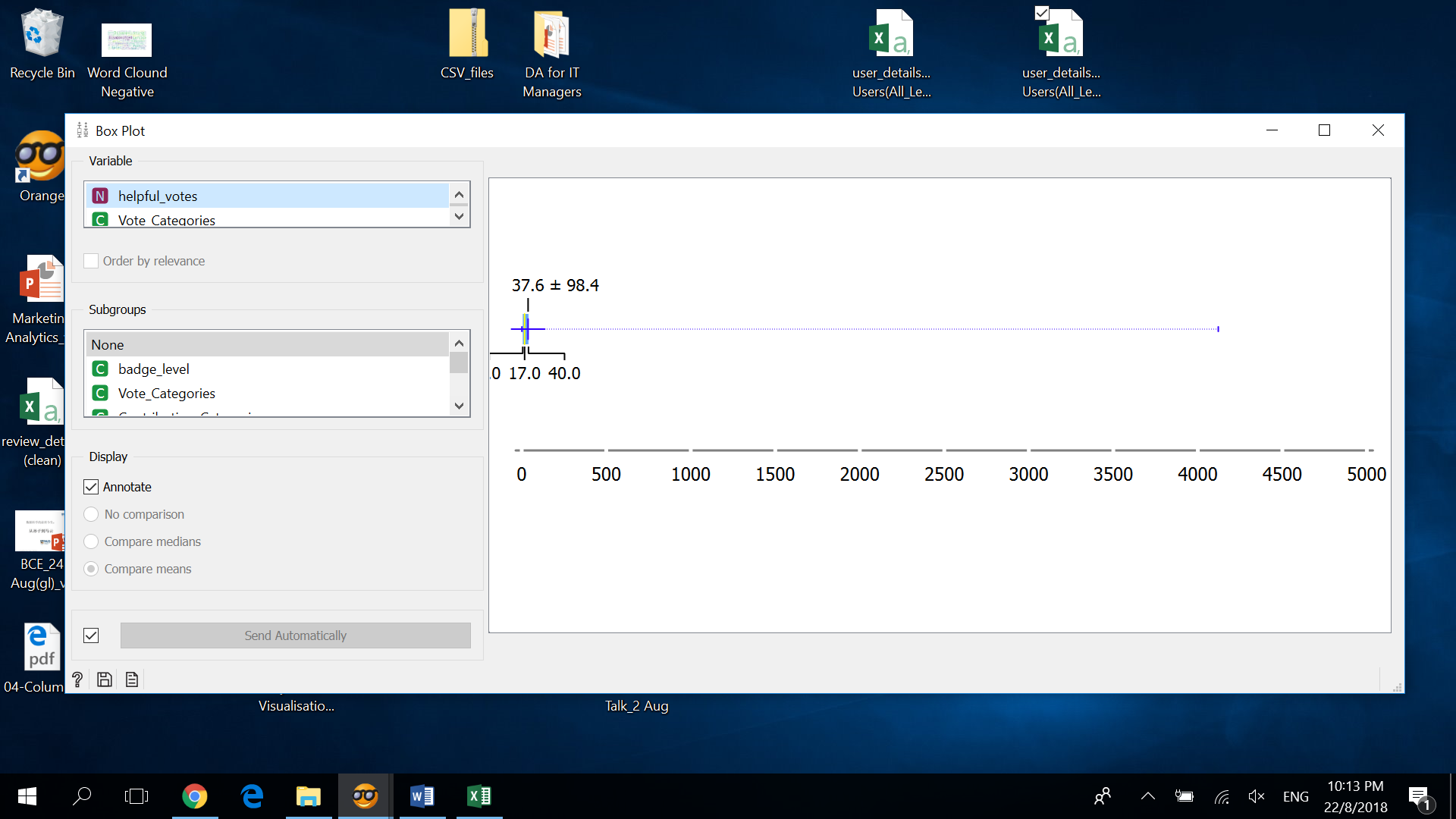


Figure 14: Outlier of Helpful Votes Figure 14.1: Helpful Votes by Levels

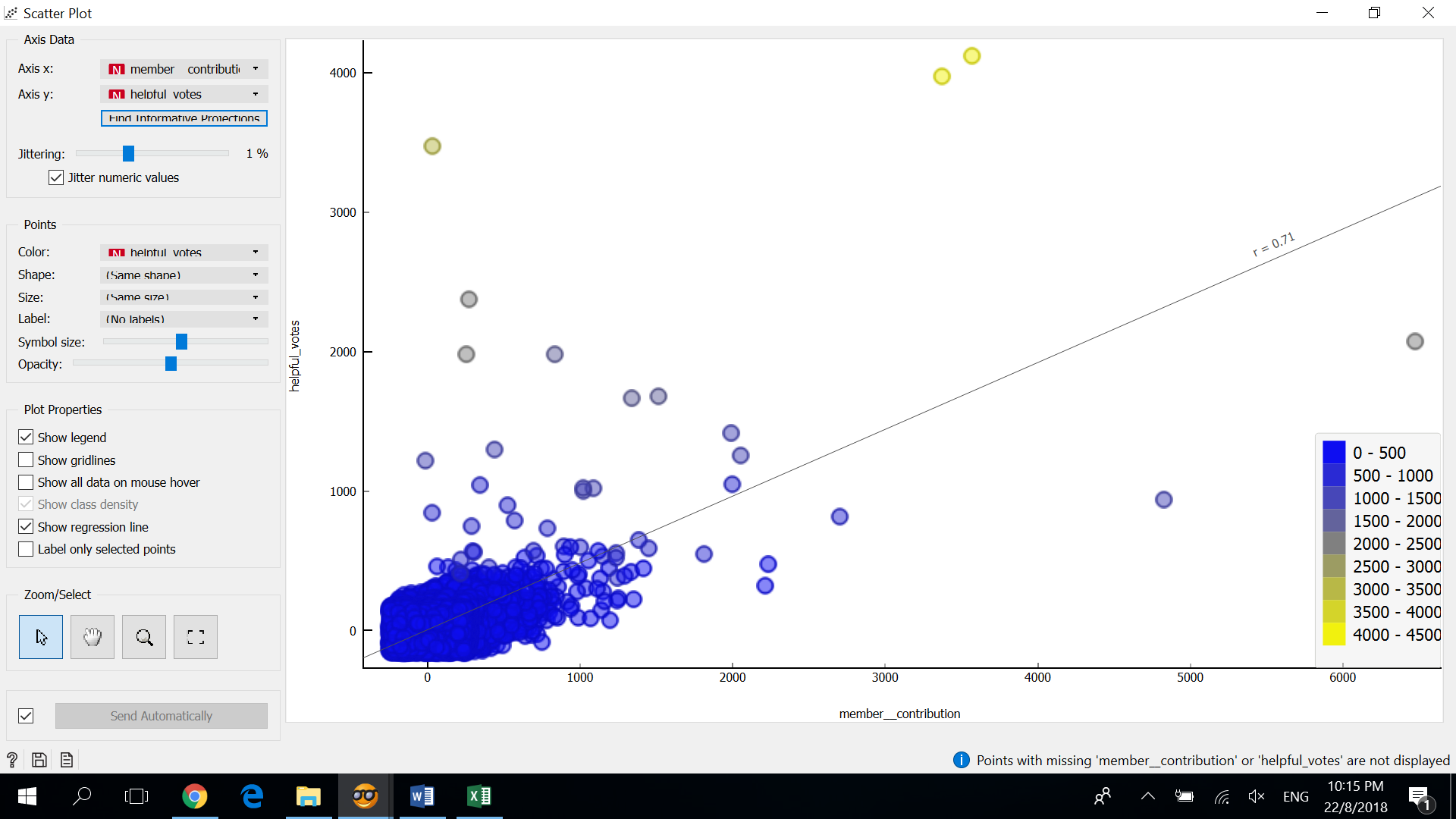
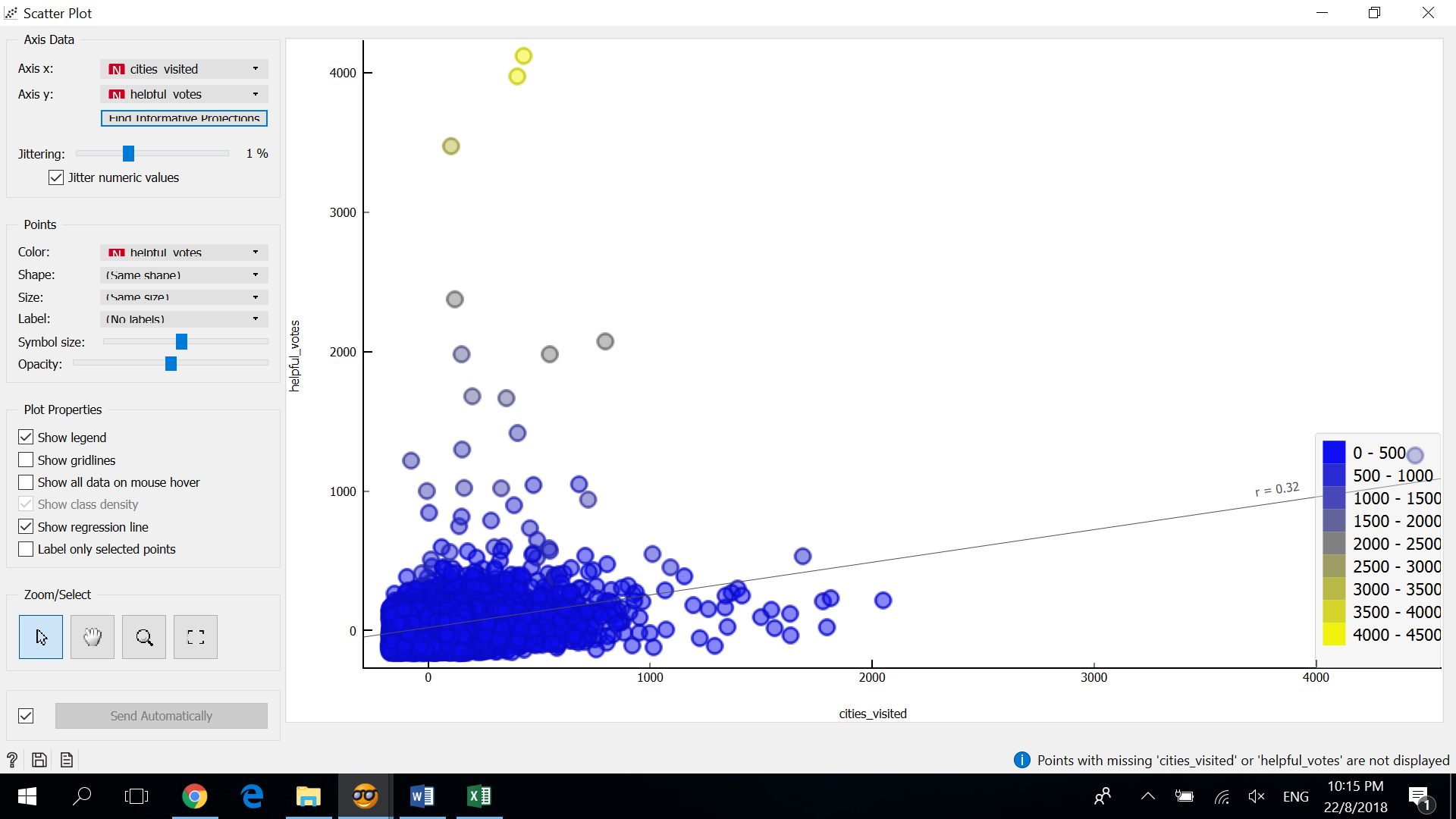


Figure 15: Helpful Votes by Cities Visited Figure 16: Helpful Votes by Contributions

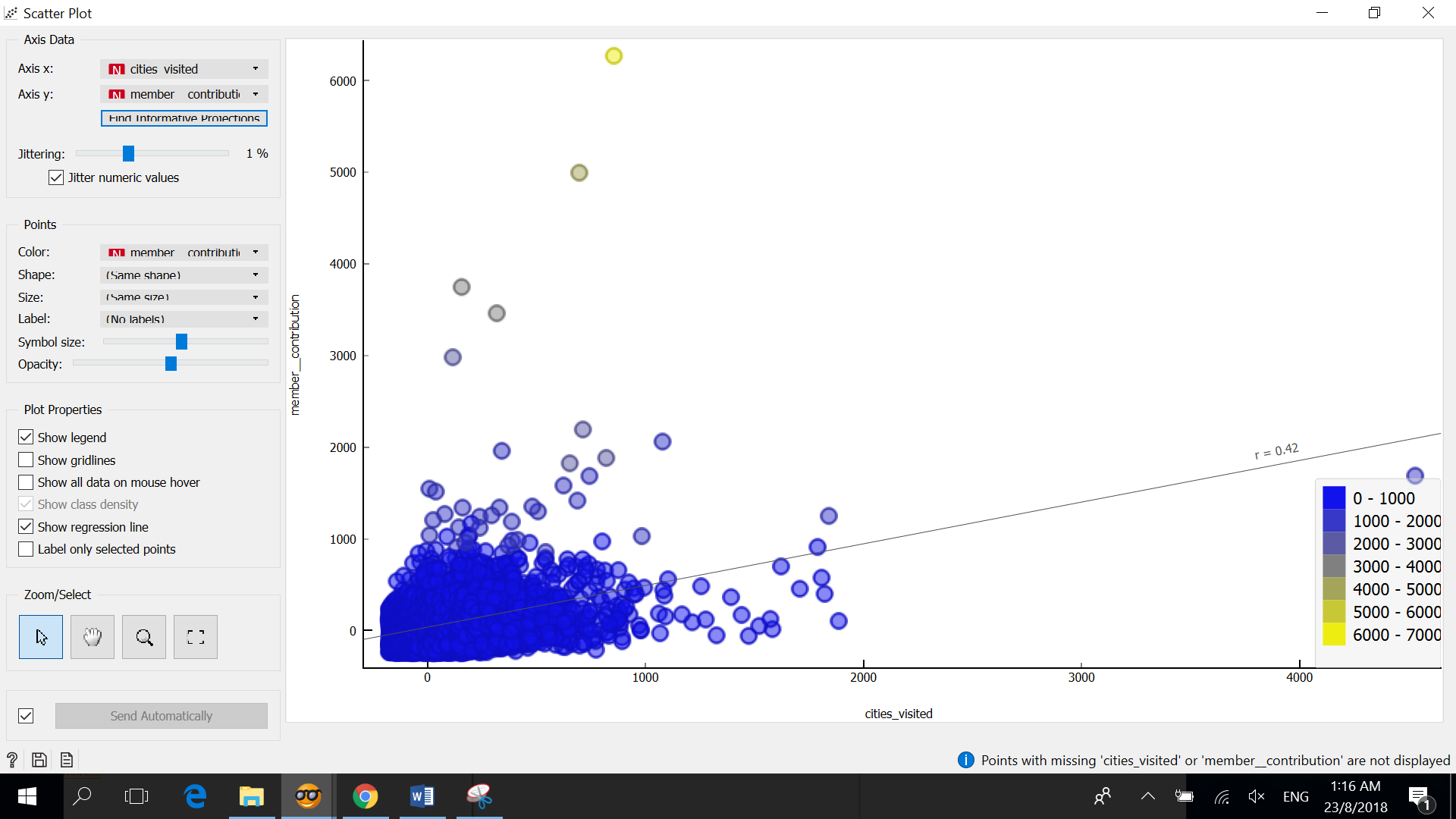


Figure 17: Contributions by Cities Visited

**4. Insights from Yelp**

Yelp is a crowd-sourced review forum as well as the online reservation service. Unlike Trip Advisor, Yelp does not have such a huge and rich user base. But – Yelp makes it up by training small businesses in how to respond to reviews, hosts social events for reviewers (Tagged as Elite Users), and provides data about businesses, including [health inspection](https://en.wikipedia.org/wiki/Health_inspection) scores.

Also, we can compare how the user sentiments vary on a lesser predominant platform.

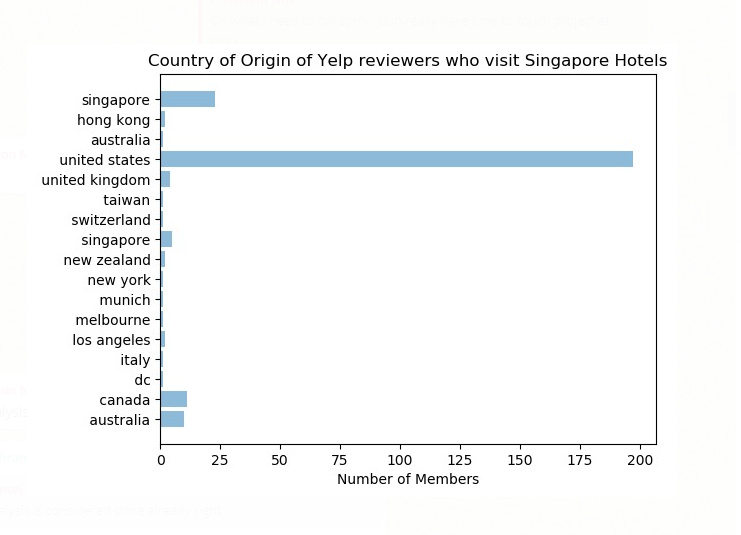
 

Fig 1 – User Location / Country of Domicile Fig 2- Positive Sentiment Cloud

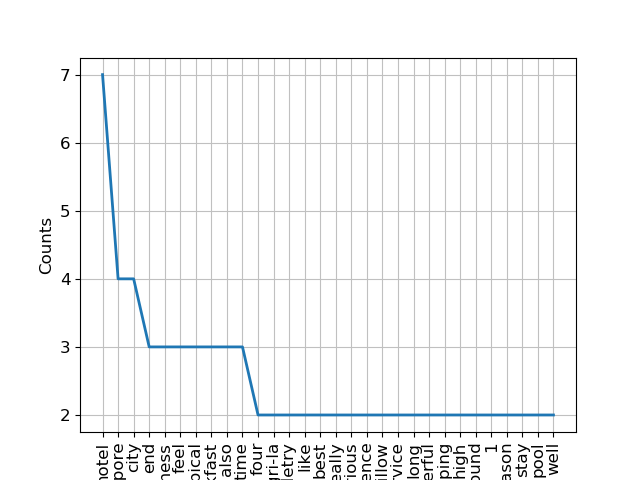
 

Fig 3- Negative Sentiment Cloud Fig 4- Positive Word Frequency

**Comparison of Yelp and Trip Advisor**

1. Similar to Trip Advisor, Yelp data shows that most of the comment writers are based in US
2. Shangri-La has been the top hotel and has been most spoken among top hotels.

# 5.0 Conclusion

Summarizing our discussion, we recommend the websites to:

1. Strengthen their marketing outreach effort to attract more new users to join the review forums, because the market growth rate has been slowing down since the year of 2015.
2. Most of the visitors are from US, UK and Australia. Hence these top 6 hotels can impress the customers by providing facilities which are appealing and liked by people from these countries.
3. Motivate existing users, to write more reviews by providing promotions or gift coupons. It will in turn contribute to the online community by making even inactive users to become active and hence promoting the website too.
4. Also there are a few months where the hotel bookings have dropped considerably. During these months the hotel can provide promotions to attract more customers.

# 6.0 References

1. Wikipedia: Tourism\_in\_Singapore [↑](#footnote-ref-2)