# Aim

The Group Project provides you with a chance to analyse the Social Web using knowledge obtained from this unit and a computer based-statistical package. For this project, we will focus on identifying a chosen company’s Twitter image.

# Method

To complete this project:

1. Read through this specification.
2. Form a group and register the members in your group using th[e Project Groups](https://vuws.westernsydney.edu.au/webapps/bb-group-mgmt-LEARN/execute/groupInventoryList?course_id=_41520_1) section of vUWS.
3. Choose a company that is active on Twitter, check that it is not already on the list of [Group Project Twitter Handles.](https://vuws.westernsydney.edu.au/webapps/blackboard/content/launchLink.jsp?course_id=_41520_1&content_id=_8003343_1&mode=cpview) Then submit the Twitter handle of the company using the same link. Note that a given company cannot be allocated to more than one group. If duplicate company names are found on the list, the group with the later time stamp will be asked to (and or should) find a new company.
4. Complete the data analysis required by the specification.
5. Write up your analysis using your favourite word processing/typesetting program, making sure that all of the working is shown and presented well.
6. Include the student declaration text on the front page of your report. Please make sure that the names and student numbers of each group member are clearly displayed on the front page. **If a group member did not contribute to any part of the project, state their contribution as 0% (no contribution means 0 mark).**
7. Submit the report as a **PDF** file by the due date using th[e Submit Group Project](https://vuws.westernsydney.edu.au/webapps/assignment/uploadAssignment?content_id=_7913437_1&course_id=_41520_1&group_id=&mode=cpview) link

All code and the outputs must be shown in the project, also include comments in the code to explain what you tried to do. Put the code at appropriate locations in document (not in an Appendix). Any submissions other than a PDF file will not be marked.

# Report Format

Once the required analysis is performed by the group, the members of the group are to write up the analysis as a report. Remember that the assessor will only see the group's report and will be marking the group's analysis based on your report. Therefore, the report should contain a clear and concise description of the procedures carried out, comments on the code, explanations of what you tried to do, the analysis of results and any conclusions reached from the analysis.

The required analysis in this specification covers the material presented in lectures and labs. Students should use R software to carry out the required analysis and then present the results from the analysis in the report. DO NOT include or use R packages that are NOT used in the workshop sessions or in the lecture sessions. Failure to adhere to this policy will result in zero marks for the corresponding question where it is used.

# Project Description

Due Week 13, Friday 11:59 pm

The company "Progressive Business Private Ltd", also known as PBP wants to start using social media to promote its business. They have approached your team with a request to find what other businesses have done successfully using social media. PBP are particularly interested in using Twitter and so have asked your group to perform the following analysis on Twitter. To begin, find a company (say X) that has a Twitter handle with over 10,000 followers and 1500 tweets, then perform the following tasks using the chosen Twitter handle. Note, no two groups should use the Twitter handle for the same company X.

## Analysing relationship between month in which the tweeter account was created and followers’ count

Tweeter user accounts have attributes such as *followers\_count* and *created\_at,* where *followers\_count* represents the number of followers an account has and *created\_at* represents the time and date of the creation of the tweeter user account.An example of *created\_at* is "created\_at": "Mon Nov 29 21:18:15 +0000 2010" and an example of *followers\_count* is "followers\_count": 21. You may want to frequently refer to the rtweet api documentation available a[t rtweet.pdf (r-project.org)](https://cran.r-project.org/web/packages/rtweet/rtweet.pdf) while answering questions in this document.

Company X wants to know if kind of relationship exists between follower count and the month in which the user account was created. Hence complete the following activities.

1. Use the *get\_followers()* API in rtweet library to get 5000 user IDs for the accounts following Company X’s twitter handle. Then use the *lookup\_users()* API to get the Twitter User object details of those followers. Reference these followers by a variable, **users** **[1 mark]**
2. Filter out those followers that do not contain a creation date. Reference these followers by a variable, **users.filtered**. You may refer to [User object | Docs | Twitter Developer Platform](https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/user) to learn about Twitter User Object attributes **[1 mark]**
3. Find the creation date, and follower count of each twitter user in **users.filtered**. Reference them by variables, **users.filtered.acc.date** and **users.filtered.follower.count** respectively. Use the account creation date to find the name of **month** in which the user account was created. You may use the *months()* function on **users.filtered.acc.date** to get the month names. Reference the months using a variable, **months**. **[1 mark]**.
4. Use the data found in step 3 (previous step) to create a table in the following format, where followers count represents the follower count of users in each month of the account creation. **[2 marks]**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **“January”** | **…** | **“December”** |
| **Followers’ count** |  |  |  |

1. The Company wants to analyse whether the proportion of followers’ count in each month are equal or not. Suggest a suitable test. Clearly outline the null hypothesis and the alternate hypothesis **[2 marks]**
2. What is the conclusion of your test? Interpret your results **[1 mark]**.

## Finding Themes in Tweets

Company X wants to analyse similar tweets of those users with high follower count. Assume a follower count of greater than 100 is considered as high. Complete the following activities

1. Search and download 1000 tweets in English from users using Company X name in the query. Filter the tweets from those users having more than 100 followers count and save them in a variable **tweets [1 mark]**
2. Clean and pre-process the tweet text data in tweets. **C**reate a Document **T**erm **M**atrix. Use term frequency weights in your analysis and ensure empty documents are not considered. Store the result in the variable, **tweets.dtm.nz** **[1 marks]**
3. Clustering is one way to find themes. Find the most appropriate number of clusters using the elbow method by using cosine distance. **[2 marks]**
4. Find the number of tweets in each cluster **[0.5 mark].**
5. Visualize your clustering in 2-dimensional vector space. Show each cluster in a different colour and symbol. Add a legend to your plot. **[2 marks]**
6. Comment on your visualization. **[0.5 mark]** 13. Find the top 20 words in each cluster. **[0.5 mark]**
7. Display a word cloud of one cluster **[0.5 mark].**
8. Find and plot the hierarchical clustering of the top 20 words in the largest cluster. Use complete linkage. **[2 marks]**

## Building Networks

Company X wants you to find some of the influential users in the network of friends. Hence you are required to create a network to find them.

1. Find the 10 most popular friends of the chosen Twitter handle. Assume a popular friend is one who has most followers. **[1 mark]**
2. Obtain a 2.0**-degree egocentric** graph centred at the chosen Twitter handle and plot the graph. The egocentric graph should be the most popular 10 friends of the chosen Twitter handle. Plot the network. Ensure the vertex names are short so that the network is not cluttered. You may need to rename the usernames to satisfy this criterion. **[3 marks]**
3. Compute the **closeness** centrality score for each Twitter handle in your graph. **[1 mark]**
4. List the top 3 most central people in your graph according to the **closeness** centrality. **[1 mark]** 20. Comment on your results. **[1 mark]** Important notes:

Note that in Section 8.3, depending on the friends of the chosen twitter handle, you possibly will reach the rate limit of the Twitter API.

See this https://developer.twitter.com/en/docs/basics/rate-limiting.html for more information on the rate limit.

I strongly recommend saving your objects as an RData file once the data is downloaded from Twitter. Share your data file with your group mates so that everyone has the same dataset. This will make your group results consistent. For more information on how to save your objects see: https://stackoverflow.com/questions/19967478/how-to-save-data-file-into-rdata

The company wants the above three-part analysis to be written up as a professional report. Each part should have its own section of the report and all questions should have thoughtful answers. Include all the R code along with its output in your assignment. Do not simply dump the output of data without considering its length and its value. Output without the code, or code without the output will result zero marks for the relevant section.