Task: infosec trends on stackexchange

General Instructions

The goal of this project is to identify generically "interesting" insights from the data. The tasks below are examples of questions that could lead to interesting insights. You may use these questions as a starting point or feel free to develop your own set of questions.

Assume that your audience is the data science team. Your submission should consist of a reproducible report such as an Rmarkdown or Jupyter notebook.

Please submit your results as a zipped attachment to cds-recruiting@cert.org within one week of receipt of the task.

Data

data.zip contains several tables generated from activity at https://security.meta.stackexchange.com/. We downloaded security.meta.stackexchange.com.7z from https://archive.org/download/stackexchange and did some simple pre-processing to convert from xml to csv format.

Tasks

1. Visualize the distribution of user reputation (Reputation column of Users.csv).
2. Build a simple predictive model for reputation. For example, your model could use 'number of badges' or 'time since first activity' to predict reputation.
3. Characterize the content of comments. Can text analysis be used to group comments in a natural way?
4. Given that the time for this task is limited, briefly state some questions you'd most like to tackle next and outline how you might approach them.