Exploratory Data Analysis

The datasets on personality type assessment, depression risk assessment, and Facebook information were obtained from the myPersonality project [1].

The depression risk assessment dataset contained the 6,561 responses to the Center for Epidemiologic Studies Depression Scale (CES-D Scale) [2] questionnaire. The CES-D is a 20-item questionnaire that asks respondents to rate how often over the past week they experienced situations related to depression (e.g. restless sleep, poor appetite, feeling lonely). Response options range from 0 to 3 for each item (0 = Rarely or None of the Time, 1 = Some or Little of the Time, 2 = Moderately or Much of the time, 3 = Most or Almost All the Time). Scores range from 0 to 60, with high scores indicating greater depression risk.

The personality type assessment dataset contained the responses of 3,137,694 individuals to the International Personality Item Pool (IPIP) version of the NEO Personality Inventory [3]. This questionnaire measures individuals on the five dimensions (Neuroticism, Extraversion, Openness to Experience, Agreeableness, Conscientiousness) of personality defined by the Five-Factor Model of Personality [4]. Scores for each dimension range from 0 to 5, with high scores indicating stronger associations with that personality dimension.

The Facebook datasets included 22,043,394 Facebook status updates from 153,727 users, and demographics information from 4,282,857 users.

In preparation for exploratory data analysis, the datasets were loaded as dataframes in R. Most of the dataframes were merged to select the information relevant to this research. Table 2 shows a summary of the data from the merged datasets.

Table 2:

	Dataset	# of Rows	# of Unique Users
A	Merged: CES-D and Status Updates	16,496,881	115,873
B	Merged: Five Factor and Status Updates	228,761	1,047
C	Merged: Five Factor, CES-D, and Status Updates	216,395	981
D	Merged: CES-D and Demographics	6,242	5,664
E	Merged: Five Factor and Demographics	2,907,658	2,907,658

A. Users who responded to the CES-D questionnaire and who had more than 1 status update

Figure 1 shows a histogram of the number of status updates for these users. It shows that majority of the users in this dataset have less than 100 status updates. Figure 2 shows a histogram of the CES-D questionnaire results for this set of users. The CES-D scores range from 0-60 and the histogram shows a normal distribution across this range.

Figure 1: Number of status updates

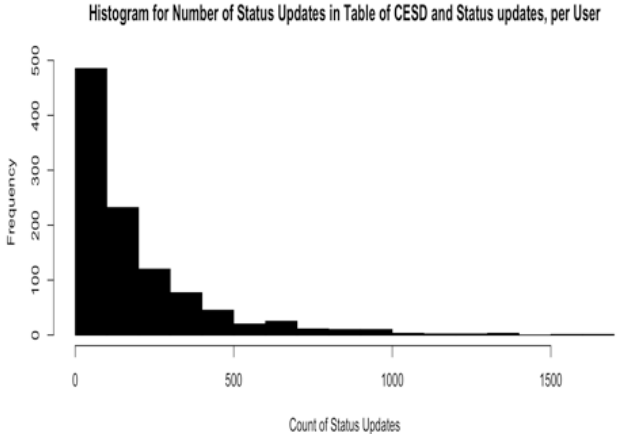
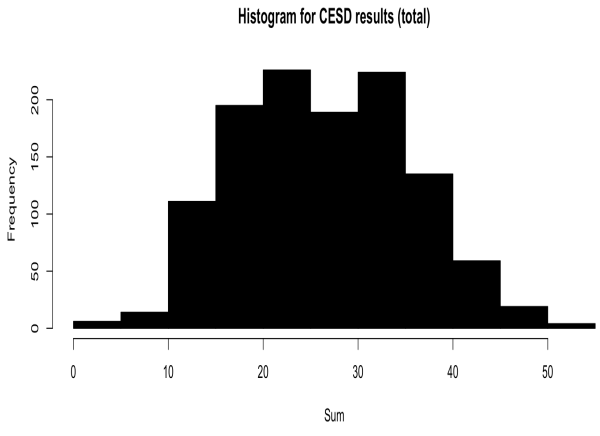


Figure 2: CES-D scores



B. Users who responded to the Five Factor Personality assessment and who had more than 1 status update

Figures 3-7 show the histograms of the Five Factor Personality assessment scores per dimension for these users. All of the histograms show a normal distribution, with most skewing right, on the higher end of the range, with the exception of Neuroticism, which was skewed left. This indicates that most of the users had lower scores on the Neuroticism personality dimension, compared to the other dimensions.

Figure 3: Neuroticism

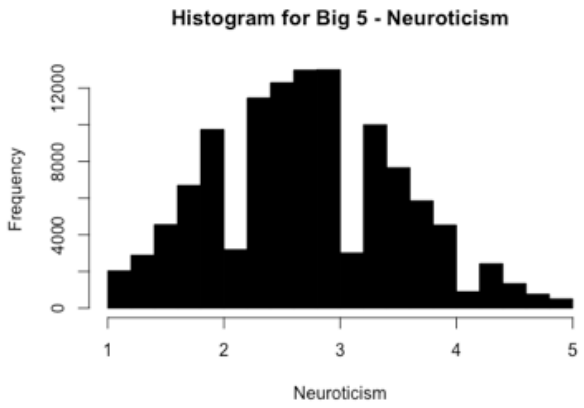


Figure 4: Extraversion

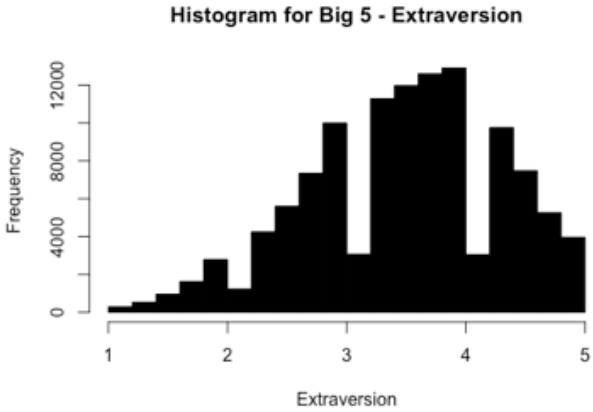


Figure 5: Openness to experience

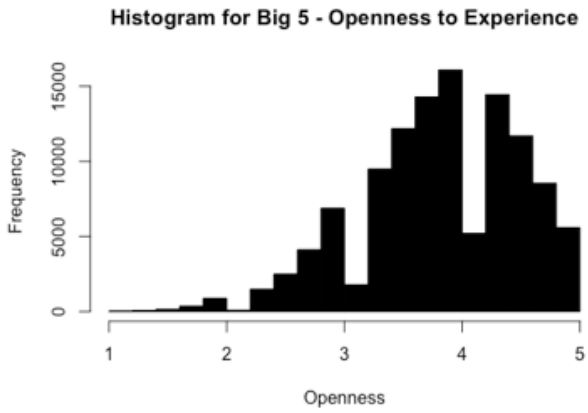


Figure 6: Agreeableness

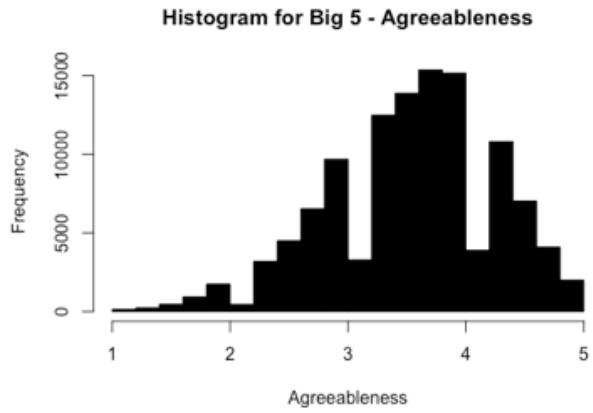
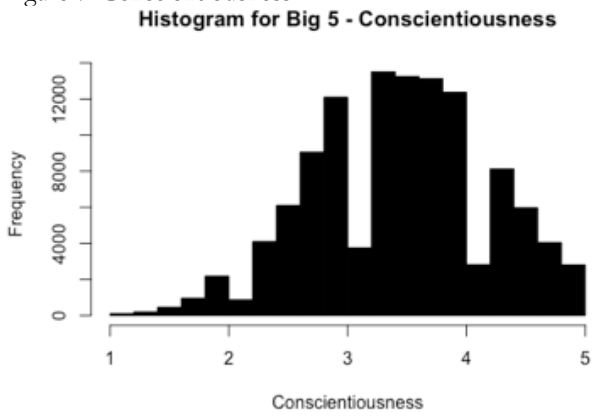


Figure 7: Conscientiousness



C. Users who responded to the Five Factor Personality assessment and the CES-D questionnaire and who had more than 1 status update

Figure 8 shows the correlations between the scores for each of the personality dimensions and the CES-D scores for the users in this dataset. This indicates that Neuroticism is negatively correlated with the other personality dimensions, but positively correlated with depression risk. Figure 9 shows that majority of the users in this dataset are at risk of depression.

Figure 8: Correlations between the Five Factor personality dimension scores and the CES-D scores

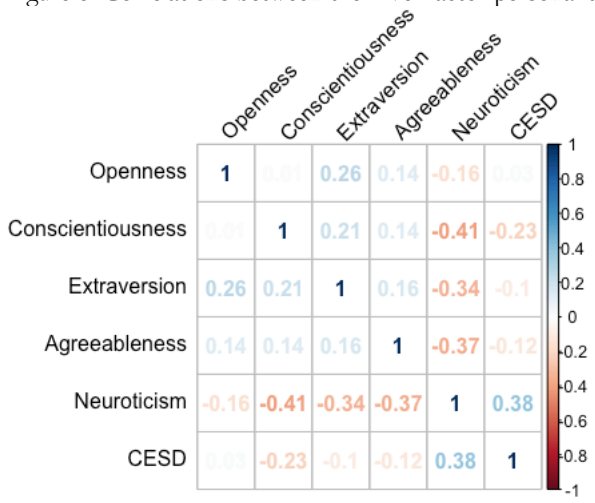
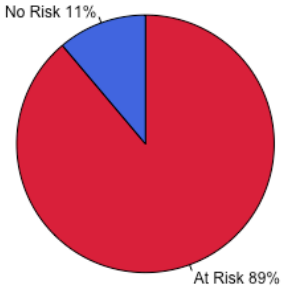
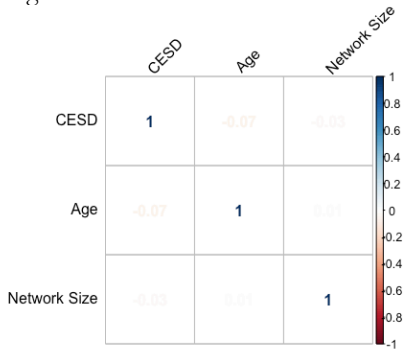


Figure 9: The CES-D score label distribution for the users in the dataset



D. Users who responded to the CES-D questionnaire and who provided demographic information via Facebook
Figure 10 shows the correlations between the CES-D questionnaire scores and Facebook demographic information: age and network size, for the users in this dataset. This indicates that there are no significant correlations between the CES-D scores and age and network size.

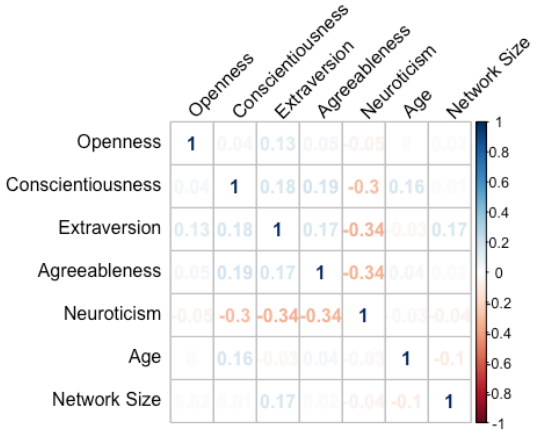
Figure 10: Correlations between CES-D scores and Age, Network Size



E. Users who responded to the Five Factor Personality assessment and who provided demographic information via Facebook

Figure 11 shows the correlations between the Five Factor personality dimension scores and Facebook demographic information: age and network size, for the users in this dataset. This indicates that there are no significant correlations between the Five Factor personality scores and age and network size.

Figure 11: Correlations between Five Factor personality scores and Age, Network Size



References

[1] Stillwell, David & Kosinski, Michal. (2012). myPersonality project: Example of successful utilization of online social networks for large-scale social research.

[2] Radloff, L. S. (1977). The CES-D scale: A self report depression scale for research in the general population. Applied Psychological Measurements, 1(3), 385-401.

[3] Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. Journal of Research in Personality, 40, 84-96.

[4] Costa, P. T., Jr., & McCrae, R. R. (1992). The five-factor model of personality and its relevance to personality disorders. Journal of Personality Disorders, 6(4), 343-359. doi:http://dx.doi.org.ezproxy.lib.ryerson.ca/10.1521/pedi.1992.6.4.343