Factor Analysis and Outcome Prediction of QUT Higher Degree Research Students
Hugh Andersen
Kerrie Mengersen
Brendan van Rooyen

Introduction
Significant time, effort and resources are expended by students and universities to ensure student success and high university rankings. This has prompted research [1][2][3][4] using statistical models to analyse student profiles, engagement types and to predict student success or failure.

This research project uses corporate QUT demographic and engagement data on Higher Degree Research students from years 2015 and 2016 for the analysis. Basic data visualisation shows the make up of HDR students, this is followed by an exploratory factor and K means cluster analysis. The factor analysis results are used in a logistic regression model to determine which factors influence a successful student outcome.

Data Visualisation
The plots below illustrate some of the diversity of the data. QUT HDR students are predominantly Australian or Chinese, study Science and Engineering, Medicine or Creative Industries (Figure 1). Students use different forms of engagement, the highest is logged into QUT resources while on campus (Figure 2). Student course loading and engagement hours seems to affect outcome (Figure 3).

Factor Analysis
A factor analysis was conducted on the demographic and engagement data using the R Psych package [5] for variable reduction and to better understand student profiles.

Demographic Factor Analysis
The number of factors was chosen using scree plots which showed that 14 and 8 factors for demographic and engagement profiles should be used.

Factor diagrams show correlation between factors, but is high enough to be problematic. The factor analysis captured 0.43 and 0.80 of the cumulative variance.

K means Cluster Analysis
The factors were used to create clusters to further profile the students using the K means algorithm. By counting number of students in each cluster a linkage diagram could be created.

Only the main links between demographic clusters (top row) and engagement clusters are shown. Medical, science and engineering and students in their 40s tend to work off campus and logon remotely.

HDR Outcome Prediction
A model to predict student outcome (pass or fail) was created using the student factors. A Logistic regression [6] model was chosen due to its interpretive ability. The model result is 68% accurate.

References and Acknowledgements

My thanks to QUT Strategic Intelligence Unit for providing the data, and my supervisors.