The purpose of this report is to demonstrate the concepts and techniques used to preprocess data, preparing it for analysis and modelling. We extracted two data sets: a 2016 Australian Census file containing information on education levels in different geographical areas in Victoria and a geographic concordance that matches Census statistical areas to Local Government Areas, which are commonly used. By merging these sets, the data becomes more usable and understandable to a wider range of people. We imported the data into R using appropriate import functions for each data file. We examined both data sets to understand the data and inform how we would preprocess it. This inspection led us to make some adjustments to the data sets, such as trimming rows. We compared the data sets to Hadley Wickham’s Tidy Data Principles (2016) and found that the education data set was not tidy, so we converted the data from wide to long format. We then set one of the variables to an ordered factor. The next step was joining the two datasets. We inspected the results and generated a new variable to group and summarise the data. We inspected the joined data set for missing values and determined it would be appropriate to exclude them. We scanned for outliers, focusing on postgraduate and undergraduate data subsets. This was achieved by Tukey’s Method of identifying outliers, using boxplots, on both total counts and proportional counts for each statistical area. These outliers were addressed through Windsorisation. We investigated the postgraduate and undergraduate data further by examining the distribution of each set on a histogram. The data was heavily positively skewed, so we applied a natural logarithm transformation to normalise the distribution. Finally, we tested how well machine learning could predict the relationship between undergraduate and postgraduate education.