Scan II

To scan for outliers, we decided to focus only on postgraduate (including graduate diploma and graduate certificate level) and bachelor degree level observations.

We looked at two forms of the data, the first being the total count of postgraduate and undergraduate for each SA1 subsection. The second was the proportion of these counts against the totals found with their respective SA1.

We visualized the distribution of the data using the base R boxplot function to find the extent of the outliers. We then reassigned this function to a vector and observed the out attribute to find. What we deemed noteworthy from the total count compared to the proportion outliers is the dramatic decrease when comparing the totals by SA1 region as opposed to the percentages.

Once the outliers were identified for each post and undergraduate education level, we used a capping function from Dr Anil Dolgun, to winsorise and reassign to vectors in order to visualize using the base R boxplot function.

***Replace entire R chunk for Scan II with:***

*#let's mutate post and grad into one (total counts)*

*edu\_uni\_tot <- education %>% mutate(Post = (education$`Postgraduate Degree Level` + education$`Graduate Diploma and Graduate Certificate Level`),*

*UnderGrad = education$`Bachelor Degree Level`) %>%*

*select(SA1, Post, UnderGrad)*

*head(edu\_uni\_tot)*

*edu\_uni\_tot$SA1 <- as.numeric(edu\_uni\_tot$SA1)*

*edu\_uni\_tot2 <- edu\_uni\_tot %>% left\_join(geo\_lookup[, c('SA1\_7DIG16','LGA\_NAME17')], by = c('SA1' = 'SA1\_7DIG16'))*

*head(edu\_uni\_tot2)*

*glimpse(edu\_uni\_tot2)*

*# visualize the totals*

*box\_post\_tot <- boxplot(edu\_uni\_tot2[,2], main = 'Post Grad Students per SA1)*

*length(box\_post\_tot$out) # 532 outliers for post grad*

*box\_under\_tot <- boxplot(edu\_uni\_tot2[,3], main = 'UnderGrad Students per SA1)*

*length(box\_under\_tot$out) # 454 outliers for under grad*

*# mutate post and grad into one (but as a proportion of the total)*

*edu\_uni\_prop <- education %>% mutate(Post = (education$`Postgraduate Degree Level` + education$`Graduate Diploma and Graduate Certificate Level`)/education$Total,*

*UnderGrad = education$`Bachelor Degree Level`/education$Total) %>%*

*select(SA1, Post, UnderGrad)*

*head(edu\_uni\_prop)*

*glimpse(edu\_uni\_prop)*

*# visualize the proportions*

*box\_post\_prop <- boxplot(edu\_uni\_prop2[,2], main = 'Proportion of Post Grad Students per SA1')*

*length(box\_post\_prop$out) # 203 outliers, significant drop in outliers when you take into account the proportion*

*box\_under\_prop <- boxplot(edu\_uni\_prop2[,3], main = 'Proportion of Under Grad Students per SA1')*

*length(box\_under\_prop$out) # 122 outliers*

*boxplot(edu\_uni\_tot2[,2:3])*

*boxplot(edu\_uni\_prop2[,2:3]) # both visuals*

*# finding proportions drastically reduces the count of outliers*

*################################*

*# Handling Outliers*

*################################*

*# Replace NAs with 0 values in the proportions (NA's occured due to 0/0 mutate)*

*edu\_uni\_prop2[which(is.na(edu\_uni\_prop2$Post)),] <- 0*

*edu\_uni\_prop2[which(is.na(edu\_uni\_prop2$UnderGrad)),] <- 0*

*cap <- function(x){*

*quantiles <- quantile( x, c(.05, 0.25, 0.75, .95 ) )*

*x[ x < quantiles[2] - 1.5\*IQR(x) ] <- quantiles[1]*

*x[ x > quantiles[3] + 1.5\*IQR(x) ] <- quantiles[4]*

*x*

*} # from Module 6 of MATH 2349, credit to Dr Anil Dolgun*

*post\_cap <- cap(edu\_uni\_prop2$Post)*

*boxplot(post\_cap, main = 'Post Grad Proportions by LGA Code (Outliers Winsorised)')*

*summary(post\_cap)*

*under\_cap <- cap(edu\_uni\_prop2$UnderGrad)*

*boxplot(under\_cap, main = 'UnderGrad Proportions by LGA Code (Outliers Winsorised)')*

*summary(under\_cap)*

*boxplot(post\_cap, under\_cap)*

Transform

We further investigated the distributions of both the Post Graduate and Under Graduate education levels found within the tidied education dataframe. After filtering new dataframes and running histogram functions on the Post Graduate and Under Graduate education levels, we found that both distributions were heavily positively skewed (right skewed). In order to visualize a normal distribution for better understanding of the data spread, both Post and UnderGrad undertook a log transformation and reassigned to their own vectors. Running a histogram function over these new vectors visualises a clearer normal distribution.

Additional Step: Simple Linear Regression

Here we wanted to investigate the linear relationship between undergraduate and post graduate education levels. In order to achieve a post graduate level of education, obviously an undergraduate level would have to be undertaken. The purpose of this analysis is to find if a consistent proportion of university attendees continue their studies to a post graduate level.

For the machine learning aspect, we trained the model on the numeric data from the undergraduate variable to predict the postgraduate variable.