R Markdown for Capstone - Exploratory Data Analysis

Anna Vaughan

January 16, 2018

# Capstone Project - Exploratory Data Analysis

## I wanted to understand the relative balance of subject areas applied for to MMU and UoS, to see which subjects were more popular at MMU, and so which subjects UoS could seek to draw students away from MMU for.

### I wanted to explore applications by subject area for each institution graphically - I first plotted a scatterplot of application numbers by subject by provider, but realised I needed to look at percentages instead, so calculated what (rounded) percentage of applications each institution received for each subject area and plotted that instead

ggplot(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small, aes(x = subjectgroupsummarylevel, y = Applications, col = providername)) + geom\_point() + geom\_smooth(lwd = 2, se = FALSE) Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small\_Total <- Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small %>% group\_by(providername) %>% mutate(totalapps = sum(Applications)) Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small\_Extra <- Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small\_Total %>% mutate(subjappspercent = Applications/totalapps) Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small\_Extrasubjappspercent \* 100 Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small\_Extra <- Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small\_Extra %>% mutate(subjappspercent = round(subjappspercent, 0)) ggplot(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small\_Extra, aes(x = subjectgroupsummarylevel, y = subjappspercent, col = providername)) + geom\_point() + geom\_smooth(lwd = 2, se = FALSE)

#### The scatterplot didn’t give a clear view of differences, so I tried a column chart instead

ggplot(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small\_Extra, aes(x = subjectgroupsummarylevel, y = subjappspercent, fill = providername)) + geom\_col(position = “dodge”)

### The column chart suggested there are major differences between MMU and UoS’ application numbers (green and red columns), so I created a table to show the differences more clearly

Subject\_Group\_2016\_MMU\_UoS\_Percentage\_Only <- filter(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small\_Extra, providername == c(“The Manchester Metropolitan University”, “The University of Salford”))

#### To make the table easier to read, I removed UCAS total rows, and columns other than provider, subject and percentage, before spreading out the provider columns

Subject\_Group\_2016\_MMU\_UoS\_Percentage\_Only <- filter(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small\_Extra, providername != “Total (UCAS wide)”) Subject\_Group\_2016\_MMU\_UoS\_Percentage\_Only <- select(Subject\_Group\_2016\_MMU\_UoS\_Percentage\_Only, providername, subjectgroupsummarylevel, subjappspercent) Subject\_Group\_2016\_MMU\_UoS\_Percentage\_Only\_Spread <- spread(Subject\_Group\_2016\_MMU\_UoS\_Percentage\_Only, providername, subjappspercent)

#### The percentage differences between MMU and UoS seemed substantial for several subjects, so I calculated the difference in the percentage of applications that each subject area made up between MMU and Salford (so that any subject that Salford had a lower percentage in shows as negative), and reordered the data to show the largest differences first

Subject\_Group\_2016\_MMU\_UoS\_Percentage\_Only\_Spread\_Totals <- mutate(Subject\_Group\_2016\_MMU\_UoS\_Percentage\_Only\_Spread, difference = The University of Salford - The Manchester Metropolitan University) Subject\_Group\_2016\_MMU\_UoS\_Percentage\_Only\_Spread\_Totals <- arrange(Subject\_Group\_2016\_MMU\_UoS\_Percentage\_Only\_Spread\_Totals, difference)