R Markdown for Capstone - Data Wrangling

Anna Vaughan

January 16, 2018

# Capstone Project - Data Wrangling

## I wanted to further explore the files I had imported, and rearrange their contents into new files that showed the information in a format that gave me access to the combinations of variables that I needed to better understand the data.

### First I created a reduced data frame to compare overall student numbers for MMU and Salford, by removing columns, grouping by and filtering

MMU\_Salf\_App\_Numbers\_1 <- select(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03, “cycle\_year”, “provider\_name”, “number\_of\_applications”) MMU\_Salf\_App\_Numbers\_2 <- MMU\_Salf\_App\_Numbers\_1 %>% group\_by(cycle\_year, provider\_name) %>% summarise(number\_of\_applications = sum(number\_of\_applications)) MMU\_Salf\_App\_Numbers\_3 <- filter(MMU\_Salf\_App\_Numbers\_2, provider\_name %in% c(“M40 The Manchester Metropolitan University”, “S03 The University of Salford”))

### I then created data frames in provider order, and that spread the year values and provider names into columns, and generated the total sum of applications by institution, to understand the changes in overall application numbers by institution

MMU\_Salf\_App\_Numbers\_4 <- arrange(MMU\_Salf\_App\_Numbers\_3, provider\_name, cycle\_year) MMU\_Salf\_App\_Numbers\_5 <- spread(MMU\_Salf\_App\_Numbers\_4, cycle\_year, number\_of\_applications) MMU\_Salf\_App\_Numbers\_6 <- spread(MMU\_Salf\_App\_Numbers\_4, provider\_name, number\_of\_applications) sum(MMU\_Salf\_App\_Numbers\_6$`M40 The Manchester Metropolitan University`) sum(MMU\_Salf\_App\_Numbers\_6$S03 The University of Salford)

### I used the data frame in provider order to plot applications to each institution over time, and saw an upward trend for both institutions, with MMU larger than Salford

ggplot(MMU\_Salf\_App\_Numbers\_4, aes(x = cycle\_year, y = number\_of\_applications, col = provider\_name)) + geom\_point() + geom\_smooth(lwd = 2, se = FALSE)

### I realised I could do with tidying up the column names in MMU\_Salf\_App\_Numbers\_6 by adding underscore

names(MMU\_Salf\_App\_Numbers\_6) <- gsub(" “,”.“, names(MMU\_Salf\_App\_Numbers\_6)) str(MMU\_Salf\_App\_Numbers\_6)

### I could see that MMU got more applications than Salford each year, but wanted to know how the balance each received changed over the years, by looking at the relative percentages of applications received by each institution (creating a total column and calculating the percentage, gathering institution name columns into one, before plotting a chart to show it)

MMU\_Salf\_App\_Numbers\_7 <- MMU\_Salf\_App\_Numbers\_6 %>% mutate(total = M40.The.Manchester.Metropolitan.University + S03.The.University.of.Salford) MMU\_Salf\_App\_Numbers\_8 <- MMU\_Salf\_App\_Numbers\_7 %>% mutate(MMU\_percentage = M40.The.Manchester.Metropolitan.University/total, UoS\_percentage = S03.The.University.of.Salford/total)

#### when trying to gather institution name columns into one, cycle.year wasn’t recognised, so lookedup up fix on stackoverflow

MMU\_Salf\_App\_Numbers\_8 <- as.data.frame(MMU\_Salf\_App\_Numbers\_8) MMU\_Salf\_App\_Numbers\_9 <- gather(MMU\_Salf\_App\_Numbers\_8, “HE\_institution”, “n”, 2:3)

MMU\_Salf\_App\_Numbers\_10 <- select(MMU\_Salf\_App\_Numbers\_9, cycle\_year, MMU\_percentage, UoS\_percentage, HE\_institution)

#### I created two tables, one for each institution, to separate out percentages and then combine again (removing redundant columns and creating new columns)

MMU\_only\_percentage <- filter(MMU\_Salf\_App\_Numbers\_10, HE\_institution %in% “M40.The.Manchester.Metropolitan.University”) UoS\_only\_percentage <- filter(MMU\_Salf\_App\_Numbers\_10, HE\_institution %in% “S03.The.University.of.Salford”) MMU\_only\_percentage <- select(MMU\_only\_percentage, cycle\_year, MMU\_percentage, HE\_institution) UoS\_only\_percentage <- select(UoS\_only\_percentage, cycle\_year, UoS\_percentage, HE\_institution) MMU\_only\_percentage <- mutate(MMU\_only\_percentage, percentage = MMU\_percentage) UoS\_only\_percentage <- mutate(UoS\_only\_percentage, percentage = UoS\_percentage) MMU\_only\_percentage <- select(MMU\_only\_percentage, cycle\_year, HE\_institution, percentage) UoS\_only\_percentage <- select(UoS\_only\_percentage, cycle\_year, HE\_institution, percentage) MMU\_Salf\_App\_Numbers\_11 <- bind\_rows(UoS\_only\_percentage, MMU\_only\_percentage)

#### I could then plot a bar chart to see how the percentage of applications varied over time

ggplot(MMU\_Salf\_App\_Numbers\_11, aes(x = cycle\_year, y = percentage, fill = HE\_institution)) + geom\_bar(stat = “identity”) + geom\_smooth()

### I could see that the relative application numbers to each institution varied only slightly, so decided to explore the subject mix applied for at each institution further

#### First I imported and cleaned the Provider by Subject data

EOC\_data\_resource\_2016\_DR4\_016\_03 <- read\_csv(“UCASData/EOC\_data\_resource\_2016-DR4\_016\_03.csv”, col\_names = TRUE, skip = 5) EOC\_data\_resource\_2016\_DR4\_016\_03\_Provider\_Subject <- EOC\_data\_resource\_2016\_DR4\_016\_03 summary(EOC\_data\_resource\_2016\_DR4\_016\_03\_Provider\_Subject) Provider\_Subject\_EOC\_data\_resource\_2016\_DR4\_016\_03 <- select(EOC\_data\_resource\_2016\_DR4\_016\_03\_Provider\_Subject, -X5)

#### I created and cleaned a data frame only containing applications by subject for 2016 overall, to compare with the individual providers if useful

Subject\_Group\_By\_Year\_Summary <- Provider\_Subject\_EOC\_data\_resource\_2016\_DR4\_016\_03 %>% group\_by(Cycle Year, Subject Group (Summary Level) ) %>% summarise(sum = sum(Number of Applications)) Subject\_Group\_2016\_Summary <- filter(Subject\_Group\_By\_Year\_Summary, Cycle Year == 2016) Provider\_Subject\_EOC\_data\_resource\_2016\_DR4\_016\_03$`Provider Name` <- gsub(pattern = "'", replacement = "", x = Provider\_Subject\_EOC\_data\_resource\_2016\_DR4\_016\_03$Provider Name) Provider\_Subject\_EOC\_data\_resource\_2016\_DR4\_016\_03$`Subject Group (Summary Level)` <- gsub(pattern = "'", replacement = "", x = Provider\_Subject\_EOC\_data\_resource\_2016\_DR4\_016\_03$Subject Group (Summary Level))

### I then created and cleaned a data frame summarising subject and year just for MMU and UoS, and filtered to only include 2016 data

Subject\_Group\_By\_Year\_MMU\_UoS\_Summary <- Provider\_Subject\_EOC\_data\_resource\_2016\_DR4\_016\_03 %>% filter(Provider Name %in% c(“M40 The Manchester Metropolitan University”, “S03 The University of Salford”)) Subject\_Group\_2016\_MMU\_UoS\_Summary <- filter(Subject\_Group\_By\_Year\_MMU\_UoS\_Summary, Cycle Year == 2016) Subject\_Group\_2016\_Summary$`Subject Group (Summary Level)` <- gsub(pattern = "'", replacement = "", x = Subject\_Group\_2016\_Summary$Subject Group (Summary Level))

#### I then combined the data frames to create one showing 2016 subject application numbers for MMU, UoS and Summary (Total), and cleaned up the result

Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary <- bind\_rows(Subject\_Group\_2016\_MMU\_UoS\_Summary, Subject\_Group\_2016\_Summary) Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary[is.na(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary)] <- 0 names(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary)[names(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary) == ‘Number of Applications’] <- ‘numapps’ names(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary)[names(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary) == ‘Cycle Year’] <- ‘cycleyear’ names(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary)[names(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary) == ‘Provider Name’] <- ‘providername’ names(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary)[names(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary) == ‘Subject Group (Summary Level)’] <- ‘subjectgroupsummarylevel’ names(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary)[names(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary) == ‘Applications’] <- ‘applications’

#### Within the new data frame I created a combined Applications column to show total applications for MMU, UoS and Summary, removed unnecessary columns and cleaned up the data

Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary <- mutate(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary, Applications = numapps + sum) Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small <- select(Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary, providername, subjectgroupsummarylevel, Applications) Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small$providername <- gsub(pattern = "S03 ", replacement = "", x = Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small$providername) Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small$providername <- gsub(pattern = "M40 ", replacement = "", x = Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small$providername) Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small$providername <- gsub(pattern = "0", replacement = "Total (UCAS wide)", x = Subject\_Group\_2016\_MMU\_UoS\_Plus\_Summary\_Small$providername)