R Markdown for Capstone - Data Cleaning

Anna Vaughan

January 16, 2018

# Capstone Project - Data Cleaning

## I wanted to explore the files I had imported and clean them up (take out any blank fields, check for NA values etc).

### First I renamed the files to show their content (and viewed a summary to check them), then tidied them to removed blank columns and checked for NA values

EOC\_data\_resource\_2016\_DR4\_005\_03\_Provider\_Domicile <- EOC\_data\_resource\_2016\_DR4\_005\_03 EOC\_data\_resource\_2016\_DR3\_027\_03\_Subject\_Domicile <- EOC\_data\_resource\_2016\_DR3\_027\_03 EOC\_data\_resource\_2016\_DR3\_018\_03\_Subject\_Age <- EOC\_data\_resource\_2016\_DR3\_018\_03 summary(EOC\_data\_resource\_2016\_DR3\_018\_03\_Subject\_Age) summary(EOC\_data\_resource\_2016\_DR3\_027\_03\_Subject\_Domicile) summary(EOC\_data\_resource\_2016\_DR4\_005\_03\_Provider\_Domicile) Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03 <- select(EOC\_data\_resource\_2016\_DR4\_005\_03\_Provider\_Domicile, -X5) Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03 <- select(EOC\_data\_resource\_2016\_DR3\_018\_03\_Subject\_Age, -X5) Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03 <- select(EOC\_data\_resource\_2016\_DR3\_027\_03\_Subject\_Domicile, -X5) any(is.na(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03)) any(is.na(Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03)) any(is.na(Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03))

### I then reviewed the values in each of the files more closely, first viewing a summary for each of them, then checking for any outlier values in each field

summary(Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03) unique(Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03$`Age Band`) unique(Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03$Subject Group (Detailed Level)) unique(Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03$`Cycle Year`) unique(Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03$Number of Applications) summary(Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03) unique(Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03$`Cycle Year`) unique(Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03$Subject Group (Detailed Level)) unique(Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03$Applicant Domicile (Region))

#### ‘Unknown’ was a category in Applicant Domicile, so I checked how many

summary(Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03$`Applicant Domicile (Region)` == "'Unknown'") summary(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03) unique(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03$Cycle Year) unique(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03$`Provider Name`) unique(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03$Applicant Domicile (Region)) unique(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03$Number of Applications)

#### ‘Unknown’ was a category in Applicant Domicile, so I checked how many

summary(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03$Applicant Domicile (Region) == “‘Unknown’”)

### I then checked the format of the text within the files to see what needed tidying up, and standardising to allow easier manipulation of the data

summary(Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03) #### Single quotes needed removing from some fields Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03$`Subject Group (Detailed Level)` <- gsub(pattern = "'", replacement = "", x = Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03$Subject Group (Detailed Level)) Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03$`Age Band` <- gsub(pattern = "'", replacement = "", x = Subject\_Age\_EOC\_data\_resource\_2016\_DR3\_018\_03$Age Band) Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03$`Subject Group (Detailed Level)` <- gsub(pattern = "'", replacement = "", x = Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03$Subject Group (Detailed Level)) Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03$`Applicant Domicile (Region)` <- gsub(pattern = "'", replacement = "", x = Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03$Applicant Domicile (Region)) Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03$`Provider Name`<- gsub(pattern = "'", replacement = "", x = Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03$Provider Name) Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03$`Applicant Domicile (Region)` <- gsub(pattern = "'", replacement = "", x = Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03$Applicant Domicile (Region))

### The Subject Domicile file contained data from a range of years, but the first way I would want to look at the subject vs domicile data was for one year only to see the relative balance of subjects and domiciles at a single point in time, so I created a new smaller file for 2016 only.

class(Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03$`Cycle Year`) Subject\_Domicile\_2016 <- subset(Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03, Subject\_Domicile\_EOC\_data\_resource\_2016\_DR3\_027\_03$Cycle Year == 2016)

### Finally I renamed columns within the Provider Domicile file to make them more meaningful, ahead of selecting and subsetting

names(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03)[names(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03) == ‘Cycle Year’] <- ‘cycle\_year’ names(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03)[names(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03) == ‘Provider Name’] <- ‘provider\_name’ names(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03)[names(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03) == ‘Applicant Domicile (Region)’] <- ‘applicant\_domicile\_(region)’ names(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03)[names(Provider\_Domicile\_EOC\_data\_resource\_2016\_DR4\_005\_03) == ‘Number of Applications’] <- ‘number\_of\_applications’