Exercises: Joins and melts

Nathan Green, Imperial College London

07/09/2019

First load packages and read in the data

```
library(dplyr)
library(ggplot2)
library(dataPakistan)

file_location <- system.file("extdata", package = "dataPakistan")

dat1 <- readxl::read_excel(path = paste0(file_location, "/MPQA _March SNID.xlsx"))

dat2 <- readxl::read_excel(path = paste0(file_location, "/LQAS %Pass _ District level_ Sep18-Mar19.xlsx
```

1. Both data sets have Province and District in them. Perform an inner join on these using `merge`.

```
xx = merge(x=dat1, y=dat2)
```

What happens and how can you fix it?

After joining correctly what do you see? Explain.

2. Perform a left join using `merge`.
3. Perform a right join using `merge`
4. Perform an outer join using `merge`.
5. Now using `dplyr` we will do some filter joins.

Load the package

```
suppressPackageStartupMessages(library(dplyr))
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

Perform an anti-join using.

Which Province has the most missing entries?

6. Perform a semi-join.
7. Melt `dat1` so that Province and District are the id columns and the others are the measure vaiables which form the variable and value columns.

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 3.5.3
```

8. Cast this long array so that it sums across District using `dcast()`.