

# One Data Science Programme Week 3

Recap - Introduction to Data Wrangling and Data Visualisation

Asma Alfayez

**what is data wrangling?**

# Data wrangling

Data Transformation – calculating the mean

```
Dark.Arts <- data$Defense.Against.the.Dark.Arts  
head(Dark.Arts, 10) #see the first 10 rows/data points
```

```
## [1] -6.889120 -4.536762 -5.440189 -3.675312 -3.542801 -5.999016  4.261754  
## [8] -3.769207  5.077157  5.695134
```

```
mean(Dark.Arts, na.rm = TRUE) #na.rm means removing NA (aka missing data)
```

```
## [1] -0.3878635
```

## Transform

id	time1	time2
1	62	60
2	59	45
3	64	50



“Add Change column”

“Convert time1 to minutes”

id	time1	time2	change	time1_min
1	62	60	-2	1.03
2	59	45	-6	0.98
3	64	50	-14	1.06

## Organise

id	time1	time2
1	62	60
2	59	45
3	64	50



“Convert rows to columns”

“Order rows by id and time”

id	time	x
1	1	62
2	1	59
3	1	64
1	2	60
2	2	45
3	2	50

## Aggregate

id	time	x
1	1	62
2	1	59
3	1	64
1	2	60
2	2	45
3	2	50



“Group by Time”

“Calculate mean and standard deviation”

time	mean	sd
1	61.66	60
2	51.66	45

# Piping

```
library(magrittr) #this is needed for the "%>%" function
library(dplyr) # this is needed for functions such as select() and summarise()

#Select Defence Against the Dark Arts and view the first 10 rows
data %>%
  select(Defense.Against.the.Dark.Arts) %>%
  head(10)
```

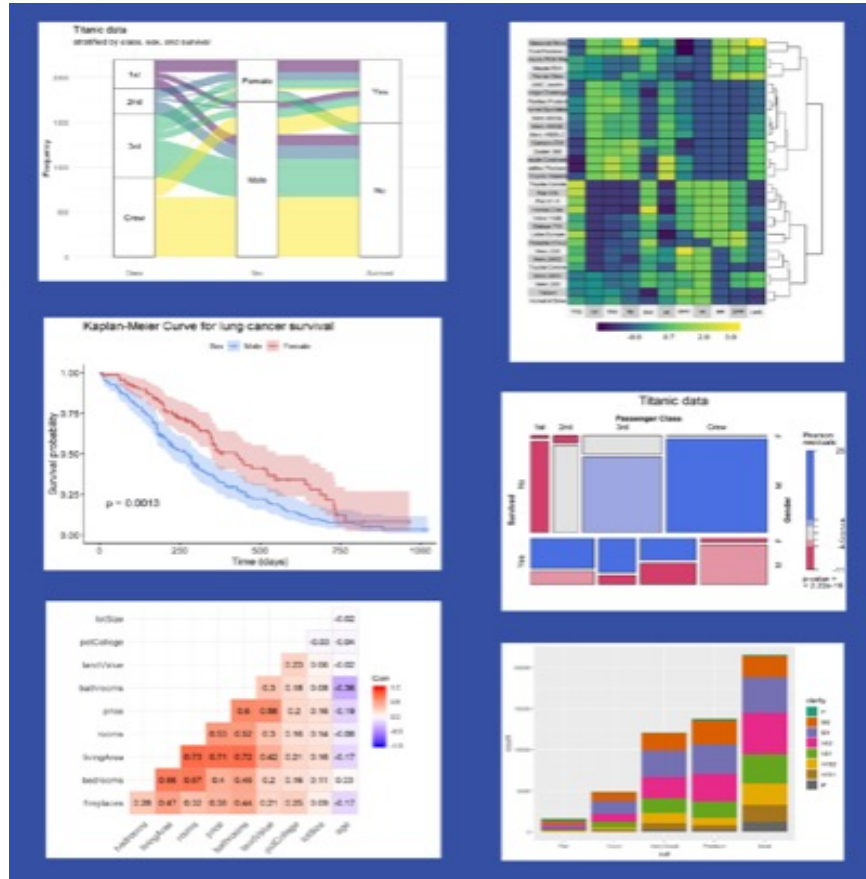
```
##      Defense.Against.the.Dark.Arts
## 1                                -6.889120
## 2                                -4.536762
## 3                                -5.440189
## 4                                -3.675312
## 5                                -3.542801
## 6                                -5.999016
## 7                                 4.261754
## 8                                -3.769207
## 9                                 5.077157
## 10                               5.695134
```

*#Add the `group\_by()`` function to further subset the data, so that we can calculate means for different houses*

```
data %>%  
  group_by(Hogwarts.House) %>%  
  select(Defense.Against.the.Dark.Arts) %>%  
  summarise(n = n(),  
            Dark.Art.Mean = mean(Defense.Against.the.Dark.Arts))
```

```
## # A tibble: 4 × 3  
##   Hogwarts.House      n Dark.Art.Mean  
##   <fct>          <int>         <dbl>  
## 1 Gryffindor      327         -4.86  
## 2 Hufflepuff      529         -4.89  
## 3 Ravenclaw      443          4.72  
## 4 Slytherin       301          4.86
```

# Data visualisation



For inquiries, email me:  
[asma.Alfayez.17@ucl.ac.uk](mailto:asma.Alfayez.17@ucl.ac.uk)

Cc:

[alvina.lai@ucl.ac.uk](mailto:alvina.lai@ucl.ac.uk)  
[albert.henry.16@ucl.ac.uk](mailto:albert.henry.16@ucl.ac.uk)  
[n.fitzpatrick@ucl.ac.uk](mailto:n.fitzpatrick@ucl.ac.uk)  
[rini.veeravalli.18@ucl.ac.uk](mailto:rini.veeravalli.18@ucl.ac.uk)