Notes for One Data Science Week 2 Slides
Data Wrangling and Visualisation

Prepared by: Kai Xiang Lim

17/06/2022

Welcome!

Welcome to Week 2! This week, we will learn about data wrangling and data visualisation. We will go through the tutorial step-by-step, the codes here have been presented in a presentation where you can find the link here.

Now, let's continue our data science journey with Hogwart's latest cohorts of students (and their enrolment test results).

Loading the data

We will first load the data from Github using the `read.csv` function. The data is also available through this link.

The csv file is loaded into R environment and stored as an object called "data".

```
data <- read.csv("Hogwarts_enrolment_data.csv")
```

Without piping

We can use the dollar sign ($) to select a column (Defense Against the Dark Arts) from `data` and see the first 10 rows of the column using the `head()` function, and then calculate the mean of this subject.

```
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
```

With 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
```

```
#Select Defence Against the Dark Arts and calculate its mean
data %>%
  select(Defense.Against.the.Dark.Arts) %>%
  summarise(Dark.Art.Mean = mean(Defense.Against.the.Dark.Arts))
```

```
##   Dark.Art.Mean
## 1    -0.3878635
```

```
#Select Defence Against the Dark Arts and calculate its mean,
# whilst showing number of students

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

```
##      n Dark.Art.Mean
## 1 1600    -0.3878635
```

The group_by function

```
#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 x 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
```

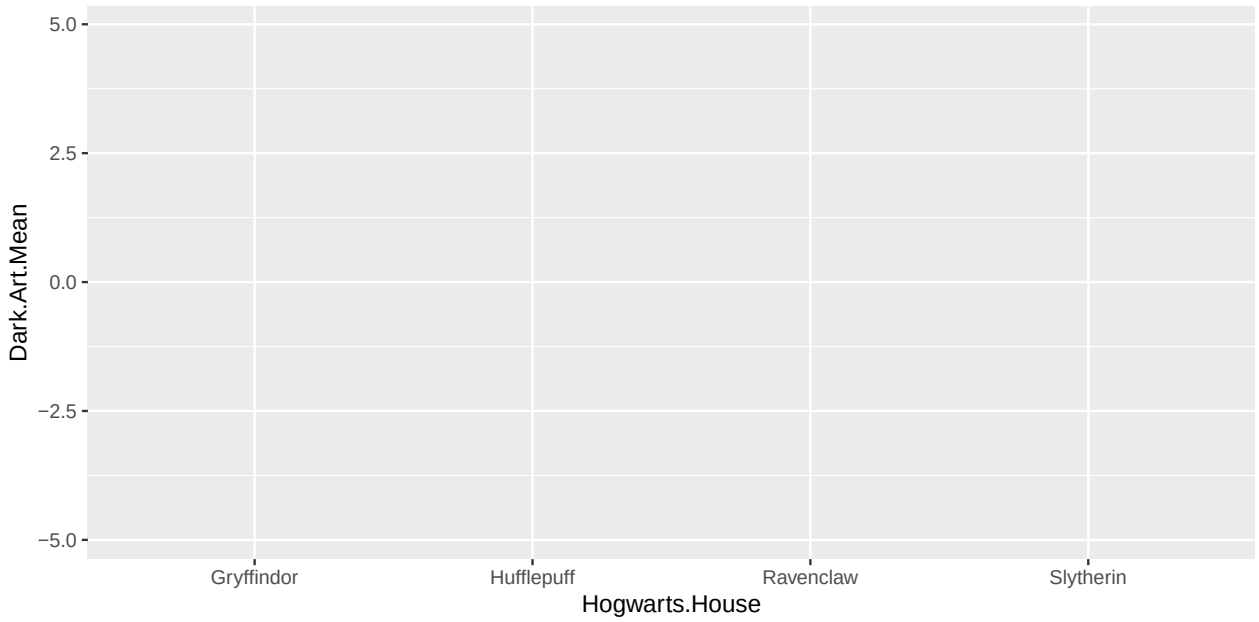
Preparing the data for data visualisation

```
# Save the output as a new object - data.for.plotting
  data %>%
  group_by(Hogwarts.House) %>%
  select(Defense.Against.the.Dark.Arts) %>%
  summarise(n = n(),
            Dark.Art.Mean = mean(Defense.Against.the.Dark.Arts)) ->
  data.for.plotting

# show data.for.plotting
data.for.plotting
```

```
## # A tibble: 4 x 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
```

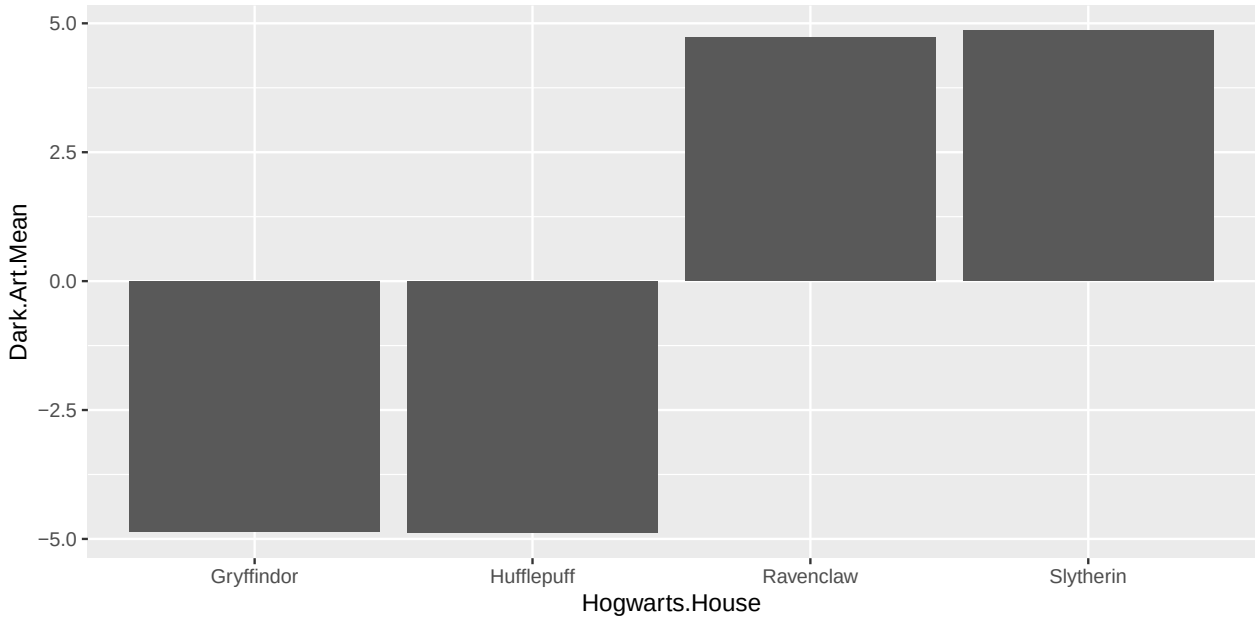
Using ggplot2 for data visualisation

```
# load another library called ggplot2, which is used for data visualisation.
library(ggplot2)

# This will only show an empty plot with the x and y axes.
data.for.plotting %>%
  ggplot(aes(x = Hogwarts.House, y = Dark.Art.Mean))
```

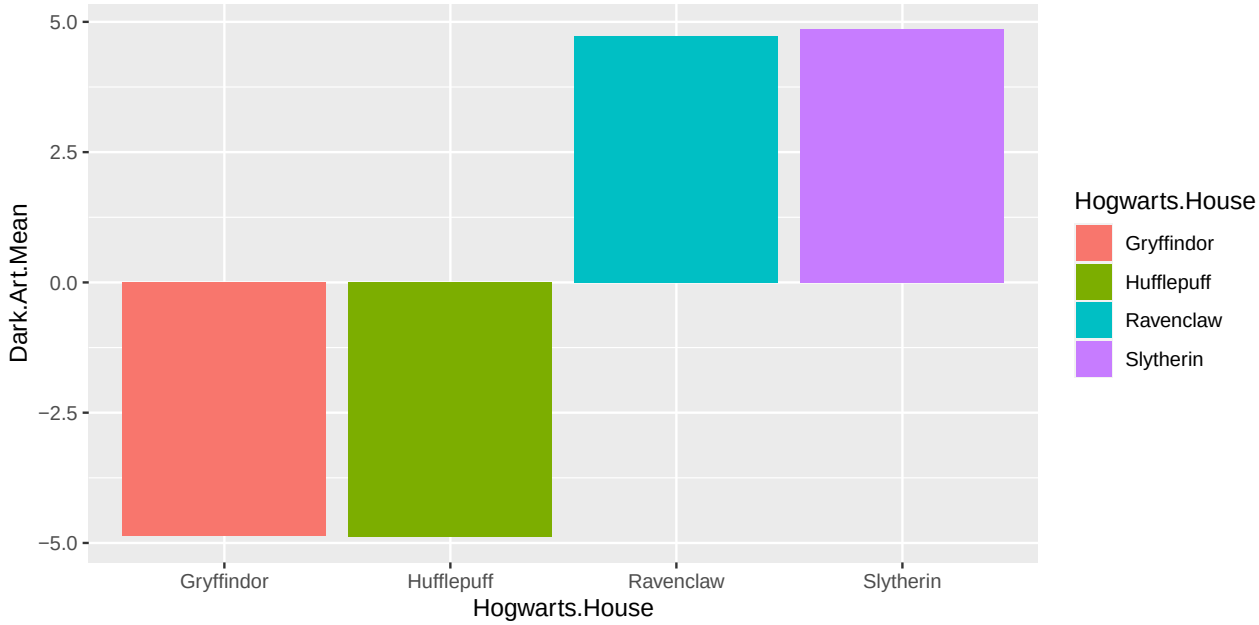



```
# To add the barplots, we need to add the geom_bar() function.
data.for.plotting %>%
  ggplot(aes(x = Hogwarts.House, y = Dark.Art.Mean)) +
```

```
  geom_bar(stat = "identity")
```

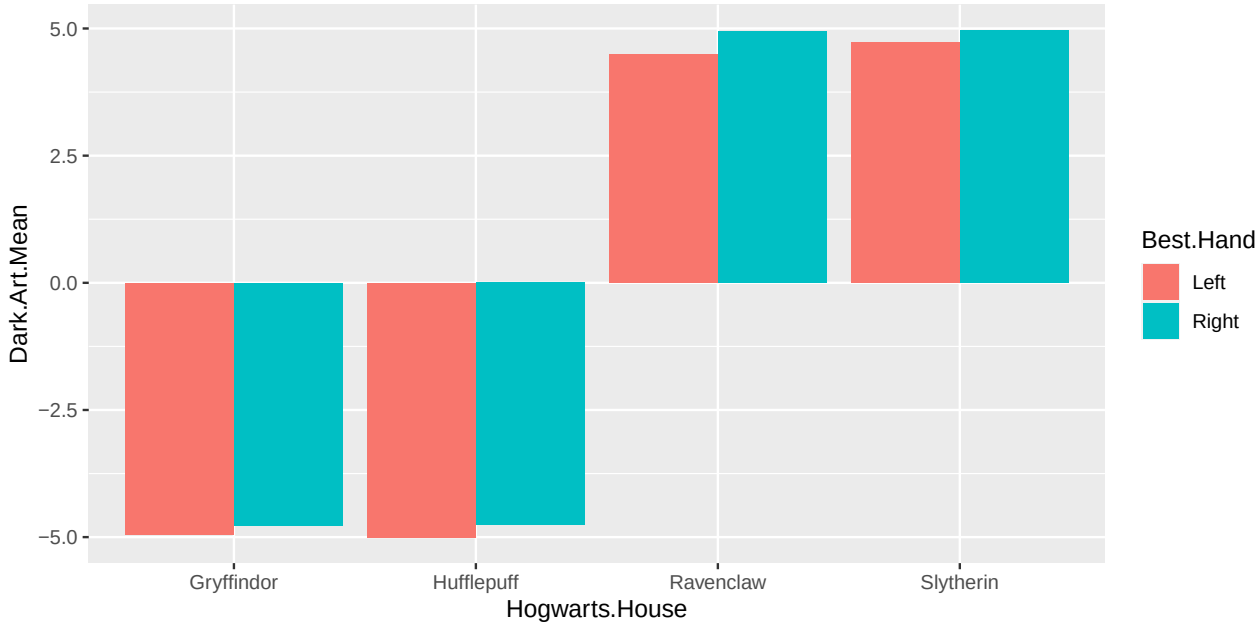



```
# Add fill colour by using 'fill=Hogwarts.House'
# such that each House has a different colour.
data.for.plotting %>%
  ggplot(aes(x = Hogwarts.House, y = Dark.Art.Mean, fill = Hogwarts.House)) +
  geom_bar(stat = "identity")
```



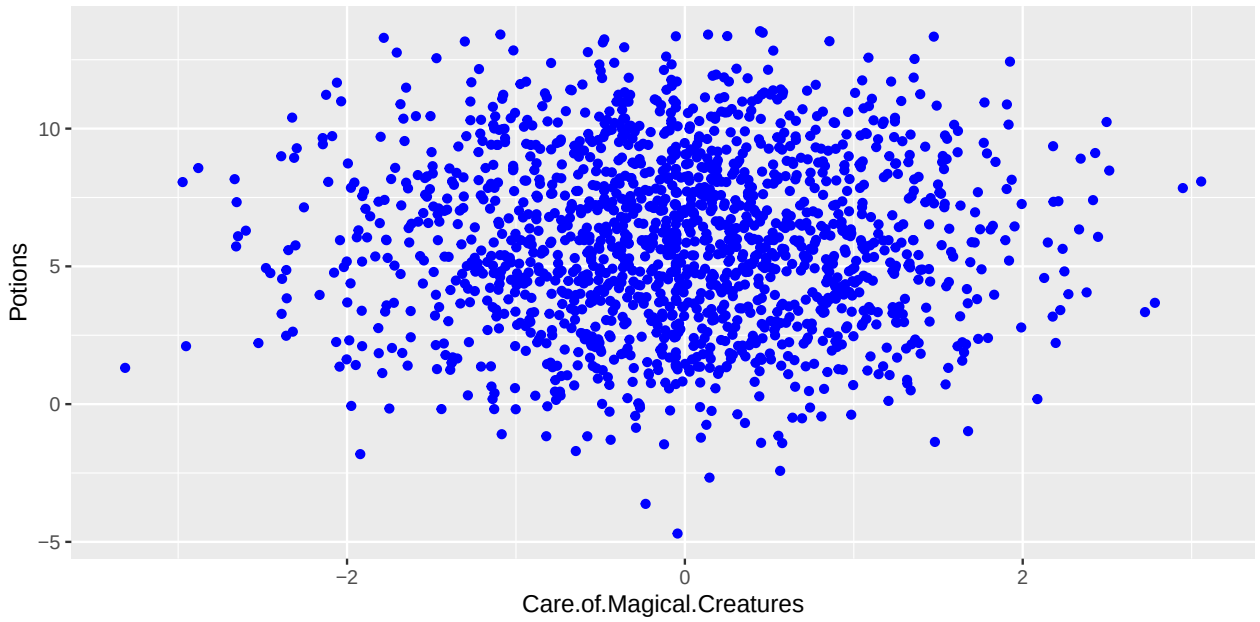

```
# We add another "group_by" variable: Best.Hand together with Hogwarts.House.
# This will give us means for 8 groups
# (4 Houses and within each house there are two means for the best hand,
# one for left and one for right)
  data %>%
  group_by(Hogwarts.House, Best.Hand) %>%
```

```
  select(Defense.Against.the.Dark.Arts) %>%
  summarise(n = n(),
            Dark.Art.Mean = mean(Defense.Against.the.Dark.Arts)) ->
  data.for.plotting.2

# Plot the second graph
data.for.plotting.2 %>%
  ggplot(aes(x = Hogwarts.House, y = Dark.Art.Mean, fill = Best.Hand)) +
  geom_bar(stat = "identity", position = position_dodge())
```



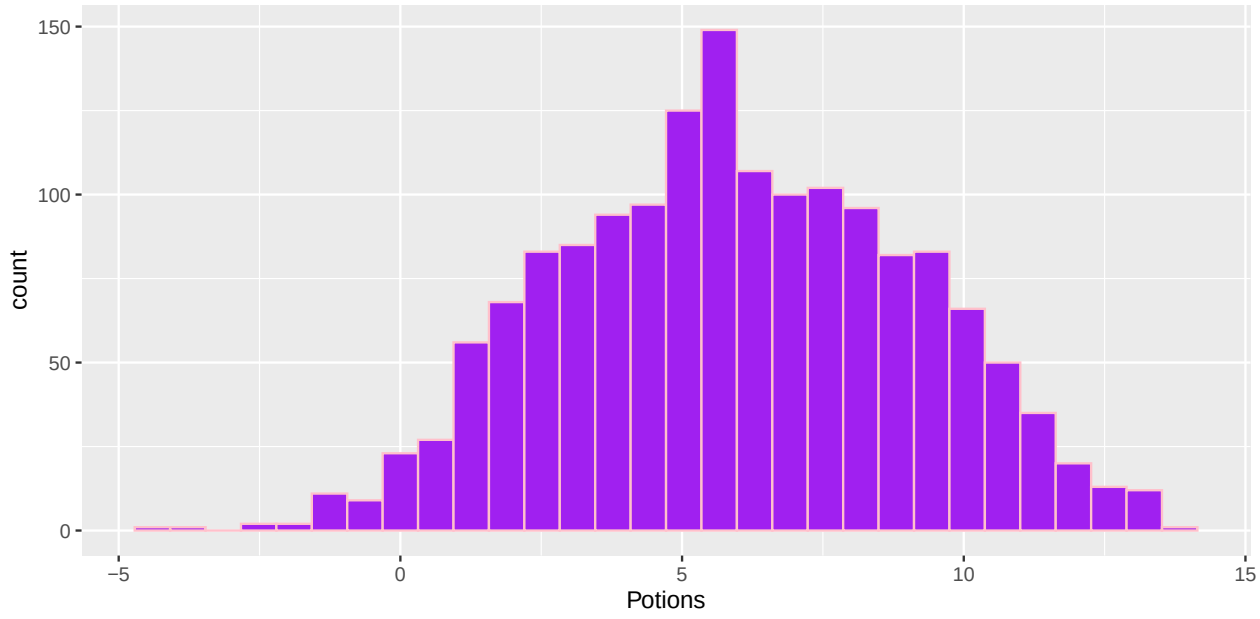
Scatterplot

```
ggplot(data,aes(x=Care.of.Magical.Creatures,y=Potions))+
  geom_point(color="blue")
```



Histogram

```
ggplot(data,aes(x=Potions))+
  geom_histogram(fill="purple", color="pink")
```



Wide to long format to plot histogram with facets

```
long_data <- tidyr::pivot_longer(data, cols=colnames(data)[7:18],
                                 names_to="subject",values_to="score")

ggplot(long_data,aes(x=score,fill=subject))+
```

```
  geom_histogram()+
  facet_wrap(~subject, scale="free_x")+
  theme(legend.position = "none")
```