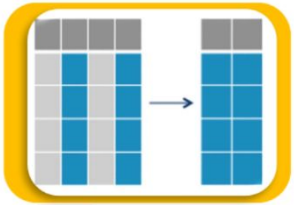
One Data Science Programme Week 3

Revisit Week 2 homework and competition

Kai Xiang Lim

- using your dataset

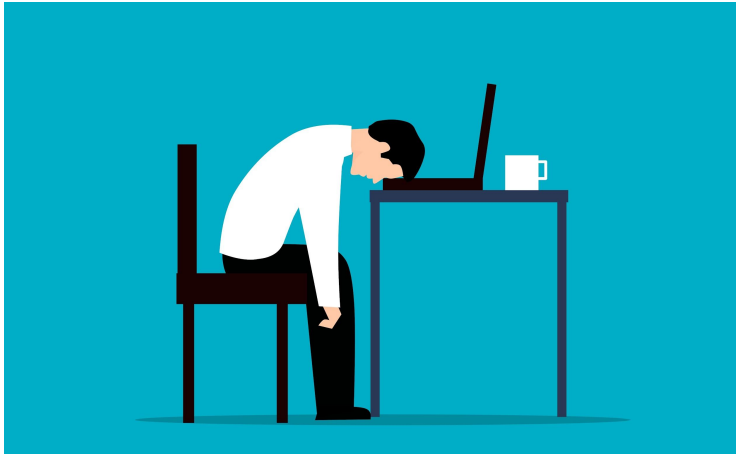
- create a visualisation using ggplot2.



- choose the columns you would like to visualise

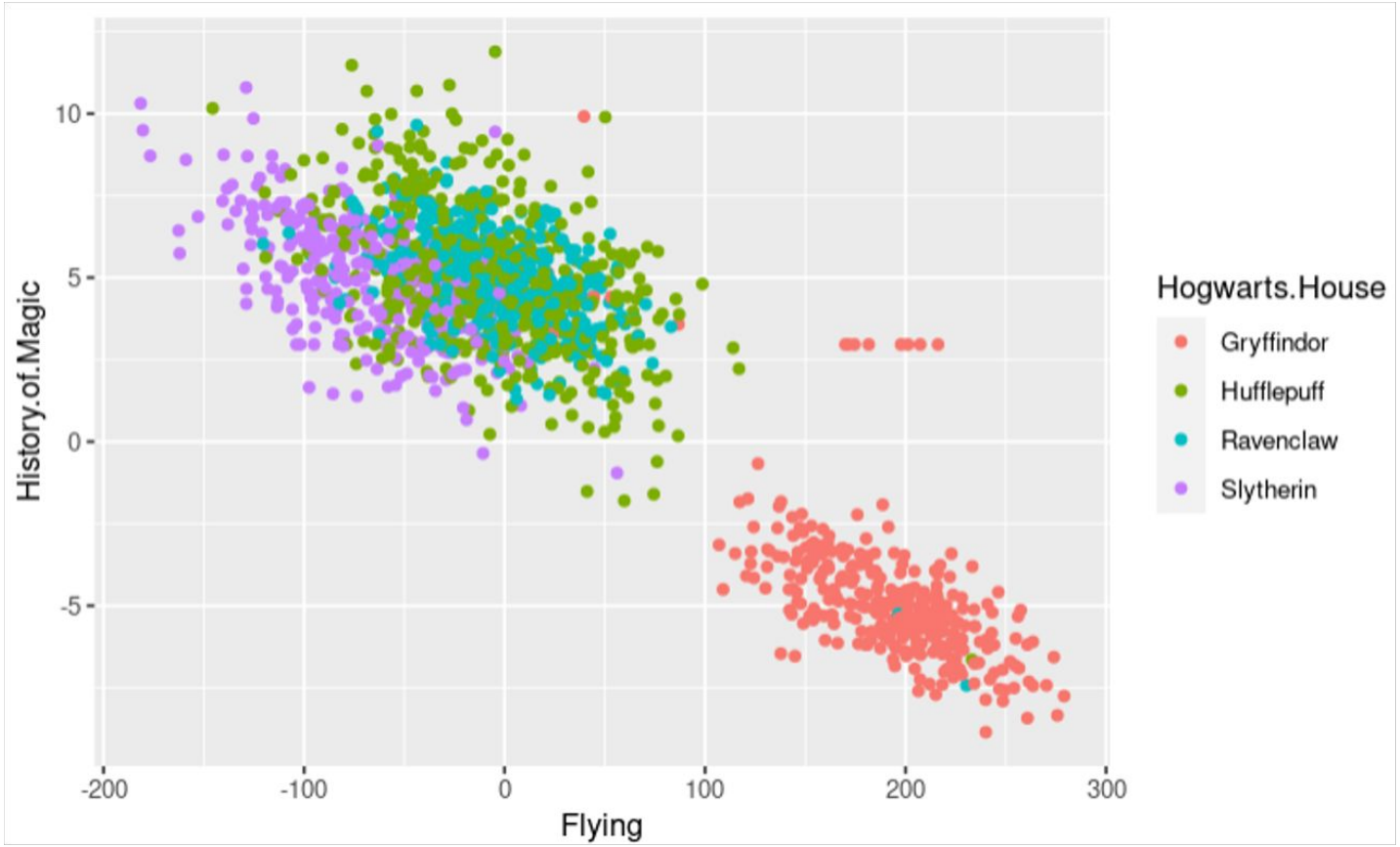
- Prepare a short report showing your codes and your interpretation of the results.

How did you feel about the homework?



Visualisation #1

By Ali Elsayed




```
data %>%
  ggplot(aes(x = Flying, y = History.of.Magic, color = Hogwarts.House)) + geom_point()
```

Analysis

- There is a clear negative correlation between Hogwarts students' scores on Flying and History of Magic, with very few outliers to this conclusion.
- The colours show us that on average Gryffindor students have much greater Flying scores than students from Hufflepuff, Slytherin and Ravenclaw, who tended to score much lower in Flying and higher in History of Magic, which the majority of Gryffindor students scored low in.

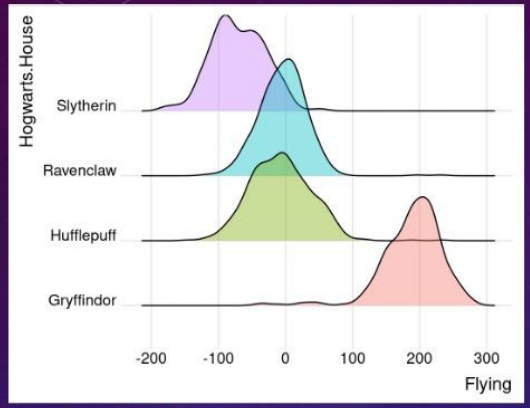
Visualisation #2

By Dylan Caddick

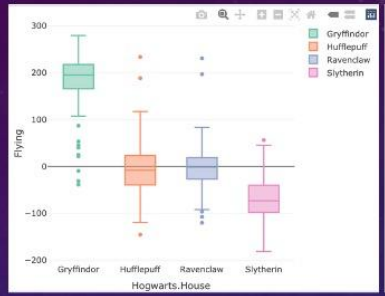
Potter People - An analysis of Hogwart's students

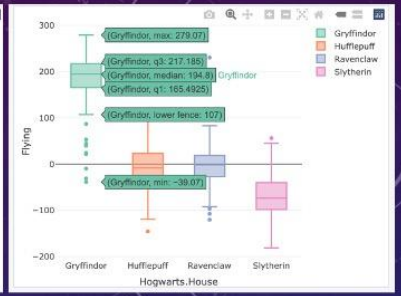
A for loop comparing all Total Scores of each variable for each House (press play or hover over to begin video)

1.
```
for (i in 2:ncol(data_grp_region)) {
  print(ggplot(data_grp_region,
aes(x=Hogwarts.House,
y=data_grp_region[,i], fill =
Hogwarts.House)) +
    geom_bar(stat = "identity", width=0.7,
alpha  = 0.7) +
      ggtitle("Total scores vs Hogwart
Houses") +
       labs(x = "Hogwart House", y =
colnames(data_grp_region[i])))
  Sys.sleep(4)
}
```



2.
```
data %>%
  plot_ly(x = ~Hogwarts.House, y =
~Flying, color = ~Hogwarts.House)
%>%
  add_boxplot()
```





3.
```
ggplot(data, aes(x = Flying, y = Hogwarts.House, fill =
Hogwarts.House)) +
  geom_density_ridges(alpha = 0.4) +
  theme_ridges() +
  theme(legend.position = "none")
```

Code for creating a new table:

```
#Creating a table with total scores for each House
data_grp_region = data.frame(data %>% group_by(Hogwarts.House)  %>%
                             summarise(
                               TotalFlying = sum(Flying),
                               TotalArithmancy = sum(Arithmancy),
                               TotalHerbology = sum(Herbology),
                               TotalDarkArts = sum(Defense.Against.the.Dark.Arts),
                               TotalDivination = sum(Divination),
                               TotalMuggleStudies = sum(Muggle.Studies),
                               TotalAncientRunes = sum(Ancient.Runes),
                               TotalHistory = sum(History.of.Magic),
                               TotalTransfiguration = sum(Transfiguration),
                               TotalPotions = sum(Potions),
                               TotalCare = sum(Care.of.Magical.Creatures),
                               TotalCharms = sum(Charms),
                               .groups = 'drop'))
```

About my graphs

Step 1. Creating the table
To start with, I decided to create a new table that has each 'House' linked to each of their total scores for each variable. I did this by creating a data frame which used 'group_by()' - this groups the sum of the variables according to the 'House' they are in.

Step 2. Creating a series of graphs
I then implemented a FOR loop to go through the each column in the new data frame, where I then plotted the variable total against the 'House' in bar chart form. I added a pause after each graph to give me time to analyze it (or I could have exported it to a file). This allowed me to clearly see a range of graphs where I can then depict one which has an interesting correlation. This can be seen looking Graph 1

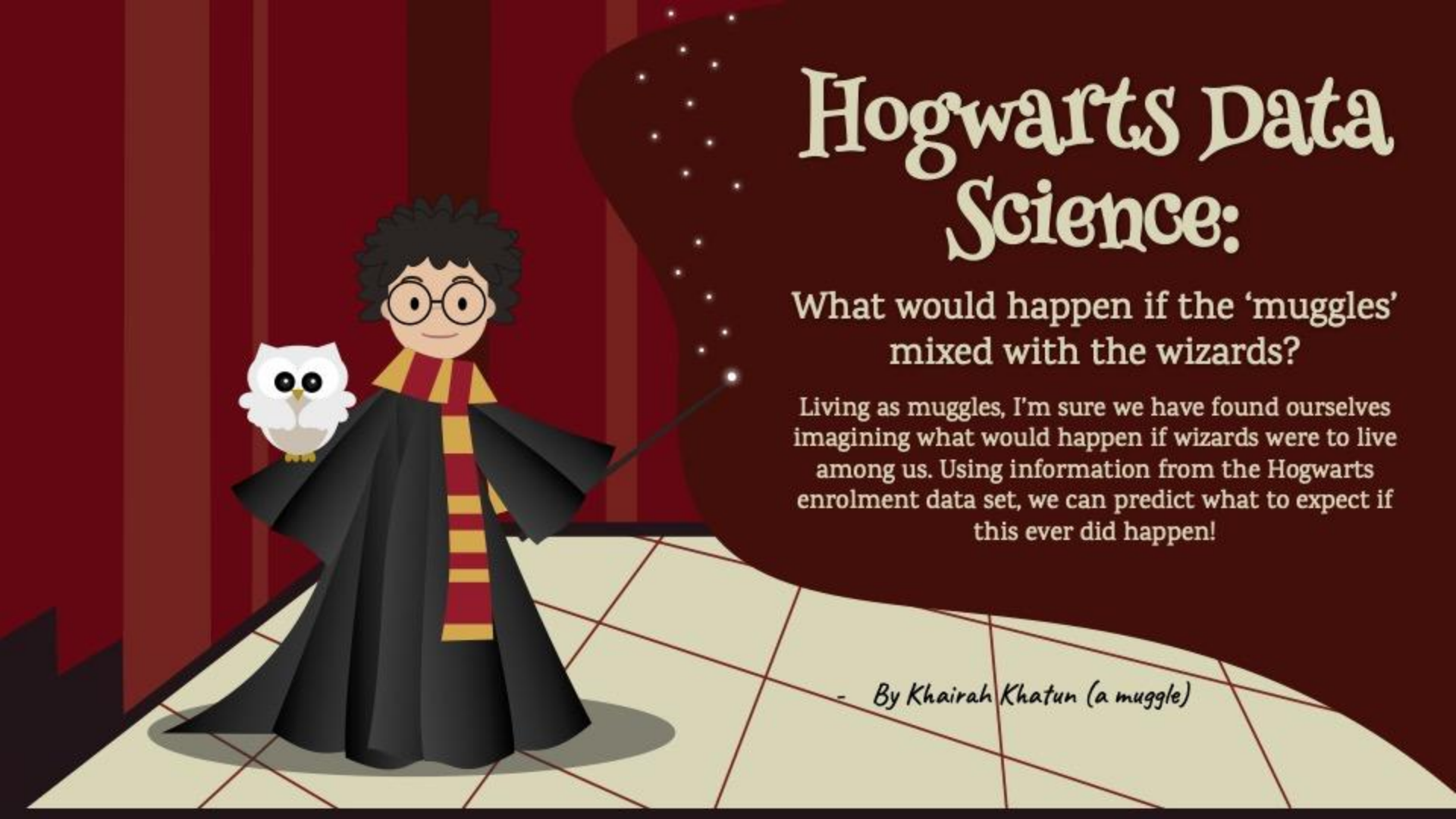
Step 3. Picking the graph and looking further
I decided to choose the Flying score and the 'House' as it showed a dominant House, which was Gryffindor. I implemented a new density graph to see the distributions of the students in each house and their score. It agreed with the totals as Gryffindor had a highest average and others had lower averages, for this I used the original data set to pick out each person – talking about Graph 2. It did show close distributions between the other 'Houses' and without the total score it may be difficult to see which were the weakest ones - Slytherin had the lowest score. However, I wanted to produce a more statistically accurate graph to allow for more observational information.

Step 4. Increasing accuracy by using box plots
I Created a box plot for each 'House'. This was created using 'plotly' to allow for user interactivity to see precise values. You can see how spread out the values are, allowing consistency for each 'House' to be measured. For example, Slytherin is more consistent than the others but achieved lower overall score and also has the lowest min and max . Additionally, it was seen that Gryffindor did have some anomalies, such as a low of -39.07. While average total was high, it did achieve some low results compared to the other Houses.

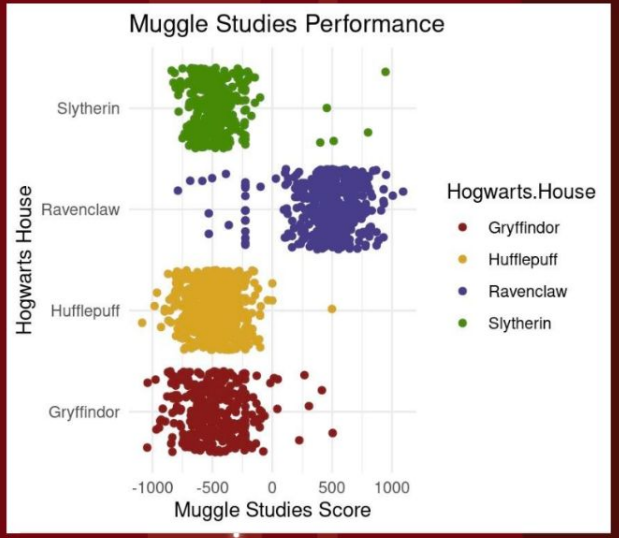
Conclusion
I wanted to simulate a more Data Science approach. I started with a general overall view, then picking out interesting topics by choosing a more specific analysis area (e.g. Which House has the best Flying?). Then I drilled down onto the specific topic, seeing what more information can be uncovered (e.g.How are the scores Distributed?). For example, another question could be – What correlations are there between variables?

Visualisation #3

By Khairah Khatun

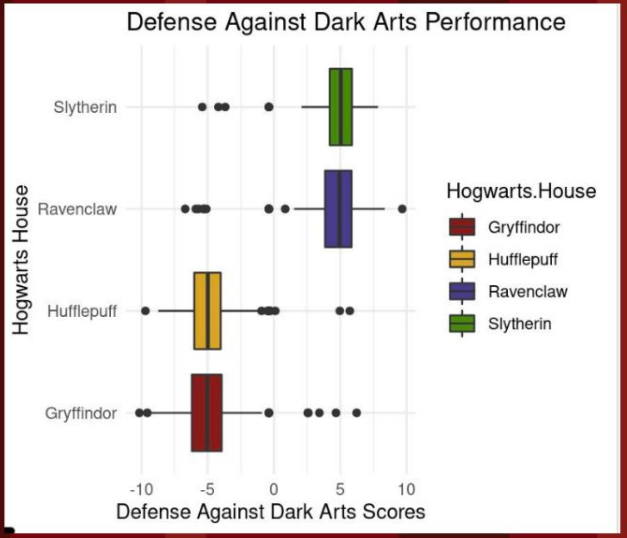

Hogwarts Data Science:
What would happen if the 'muggles' mixed with the wizards?
Living as muggles, I'm sure we have found ourselves imagining what would happen if wizards were to live among us. Using information from the Hogwarts enrolment data set, we can predict what to expect if this ever did happen!
- By Khairah Khatun (a muggle)

- Ravenclaw has the best performance in muggle studies (mean score~466)



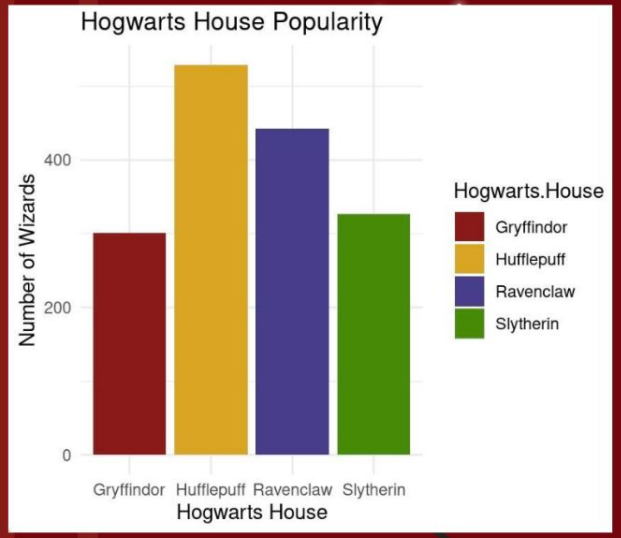

```
ggplot(data, aes(x=Muggle.Studies, y=Hogwarts.House, color=Hogwarts.House)) +  geom_jitter() + ggtitle("Muggle Studies Performance") + xlab("Muggle Studies Score") + ylab("Hogwarts House") + theme_minimal()
```

- Ravenclaw and Slytherin have the highest median dark arts scores




```
ggplot(data, aes(x=Defense.Against.the.Dark.Arts, y=Hogwarts.House, fill=Hogwarts.House)) + geom_boxplot() +  ggtitle("Defense Against Dark Arts Performance") + xlab("Defense Against Dark Arts Scores") + ylab("Hogwarts House") + theme_minimal()
```

- Hufflepuff and Ravenclaw are the most common Hogwarts houses




```
ggplot(data=ODI, aes(x=Hogwarts.House, y=X, fill=Hogwarts.House, xlab="Number of Wizards", ylab="Hogwarts House")) + geom_bar(stat="identity") + ggtitle("Hogwarts House Popularity") + xlab("Hogwarts House") + ylab("Number of Wizards") + theme_minimal()
```

- This shows us that wizards in Ravenclaw understand us muggles the best, and if we ever needed protection from the dark arts they would be helpful - Ravenclaw is also the 2nd most common Hogwarts house (making up 27.7% of wizards) so befriending them certainly won't be difficult
- Contrastingly, Slytherin and Gryffindor have weak muggle studies scores and aren't as good at defending against the dark arts, so perhaps we should keep our distance from them as they could get us into some trouble...
- To conclude, if the wizards ever invaded it would be best to stick close to those in the Ravenclaw house!

Who are the winners?

Ali Elsayed

Dylan Caddick

Khairah Khatun