Data Understanding Report

Ammar Hasan 150454388

20 November 2018

Contents

1 Introduction

This report documents the Data Understanding stage of the CRISP DM cycle. The Data Understanding stage involves the process of collecting insights about the data, which are used to help form hypothesis for later analysis. The process involves data descriptions, explorations and quality verification. In this project this would involve attempting to analyse the relationship between archetypes and progress in the program (steps and questions).

2 Data Background

As described in the Business Understanding stage, the data is provided from an online learning program about Cyber Security hosted by Future Learn. The data is divided into a set datasets that are collected at 7 runs of the program that occur in different time intervals and are anonymised. The data used for this project is derived from the last run for questions responses, archetype survey and step activity tables. Moreover, data was cleaned and merged in some of the Data Preparation steps carried before this step to ensure that the data is clear of anomalies and is aggregated.

3 Data Description

This is described in detail in the Data Preparation report, but to summarise the data from the question responses, archetype survey and step activity is preprocessed (cleaned, aggregated and merged) to form a dataset with the following fields:

- learner-id
  - This is the field used to uniquely identify learners from the online program throughout the data (used to help join the cleaned tables together).
- archetype
  - This field is derived from the archetype field from the archetype survey table and is used to represent the type of learner a user is according to their responses.
- week-completed-tasks
  - There are 3 fields of this variation for each week, and they represent the completed tasks for the users in a given week. This is derived from the step activity table.
- week-total-marks
  - There are 3 fields of this variation for each week, and they represent the total marks gained (1 for each correct answer) for all attempts on questions for the users in a given week. This is derived from the question response table.
- week-total-attempts
  - There are 3 fields of this variation for each week, and they represent the total attempts on all questions for the users in a given a week. This is derived from the question response table.

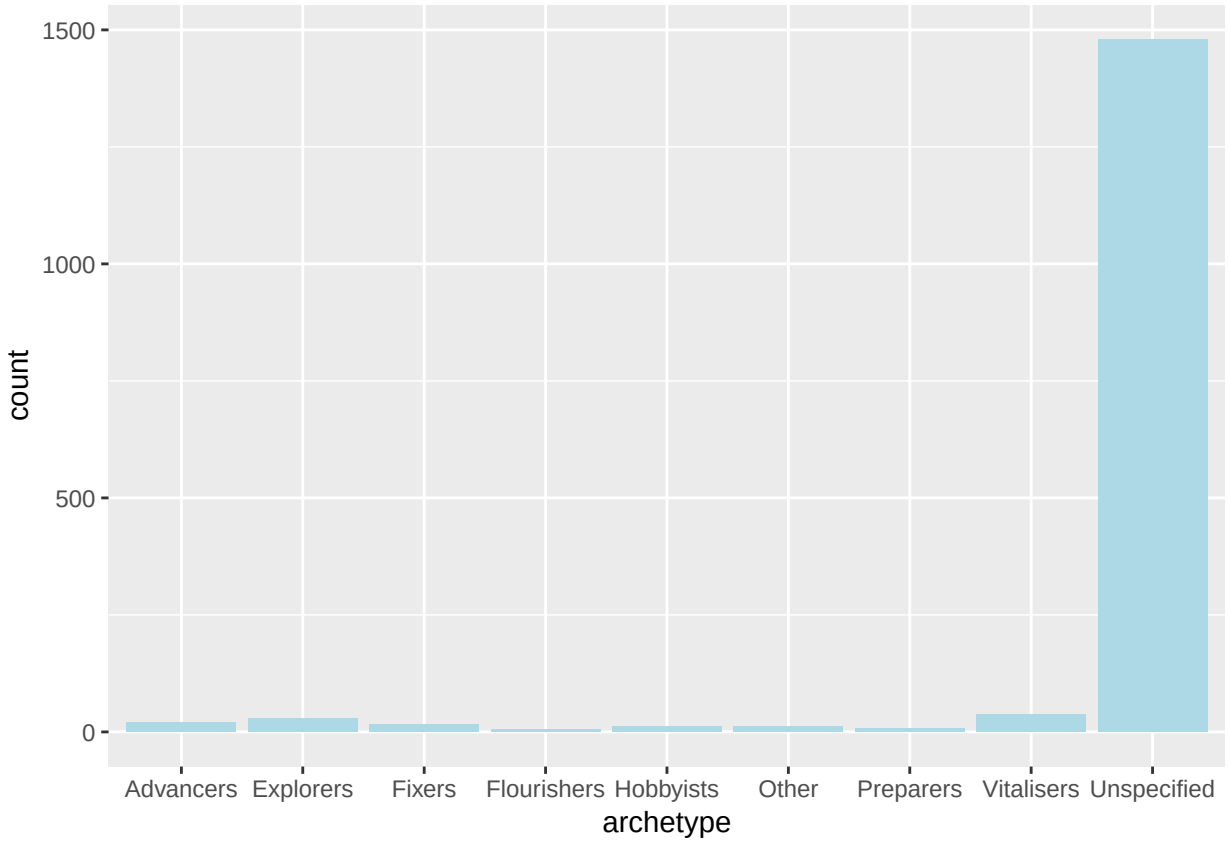

Figure 1: Count of Archetypes

4 Data Exploration and Analysis

This section concentrates on the process of exploring the data for patterns and interesting features using Exploratory Data Analysis and some unsupervised learning techniques (e.g. Cluster Analysis).

4.1 Graphical Summaries (Exploratory Data Analysis)

4.1.1 Frequencies (Histogram)

This section will look into how are the learners divided across the different archetypes.

4.1.1.1 All Archetypes

```
# call ggplot bar to count archetypes
 ggplot(progressTypeDf, aes(archetype)) +
  geom_bar(fill = 'Light Blue')
```

As seen in figure 1 the archetypes are dominated by people that choice to not specify an archetype at all (NA), it seems that a lot of people did not fill out the archetype survey.

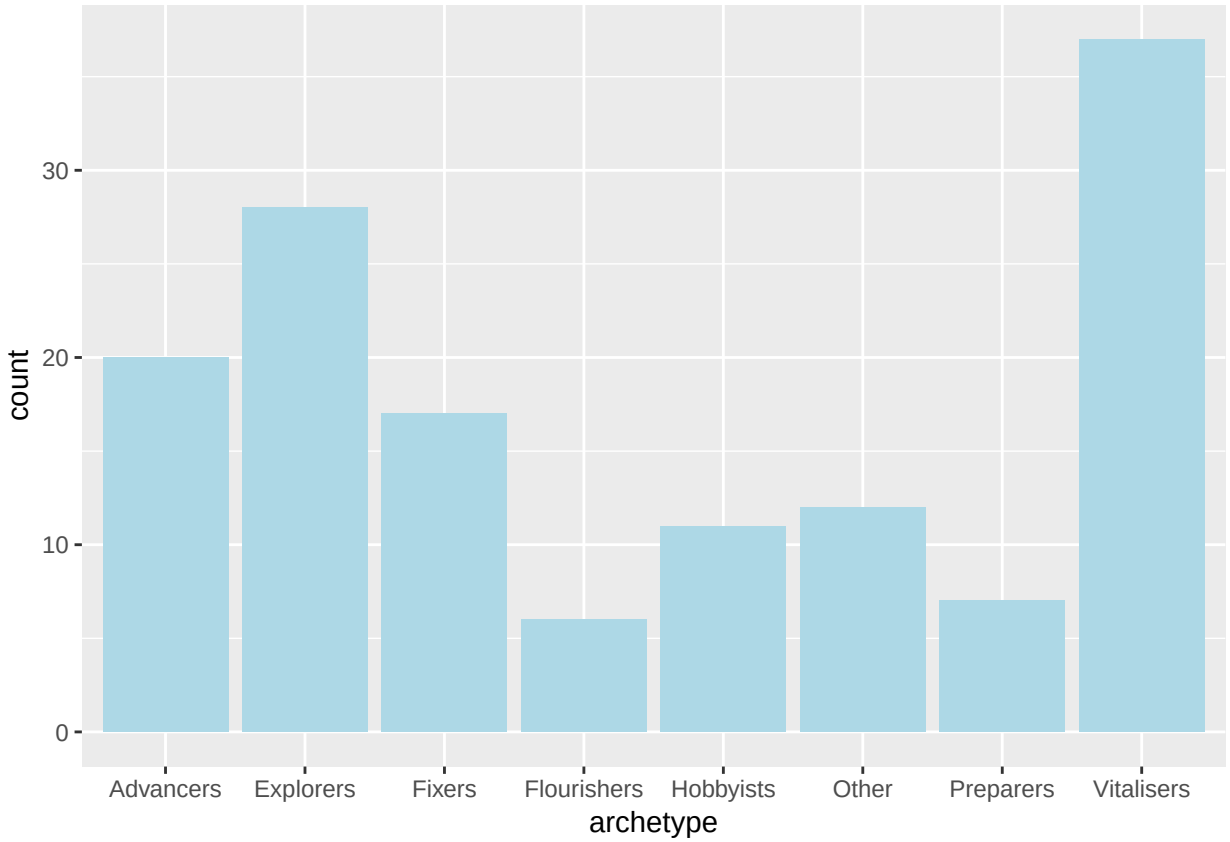

Figure 2: Histogram of Archetypes (No Unspecified)

4.1.1.2 Archetypes Without Unspecified

```
# Subset unspecified out this time
ggplot(subset(progressTypeDf, archetype !='Unspecified'), aes(archetype)) +
  geom_histogram(stat='count', fill = 'Light Blue')
```

Figure 2 shows that archetypes without the unspecified archetypes seem more balanced between learners, with vitalisers dominating, and flourishers and preparers being comparatively the smallest.

4.1.2 Location and Spread Measurements

This section is where interesting observations on the structure of that ranges and averages of the data are recorded. This would be done by using quantile based Box and Whisker plots for individual variables and variables by each archetype.

4.1.2.1 Individual Variable Boxplots

```
# gather data to plot each variable individually, function
# for later use
gatherProgress = function(df) {
  return(gather(df, Variable, Value, -c(learner_id, archetype)))
```

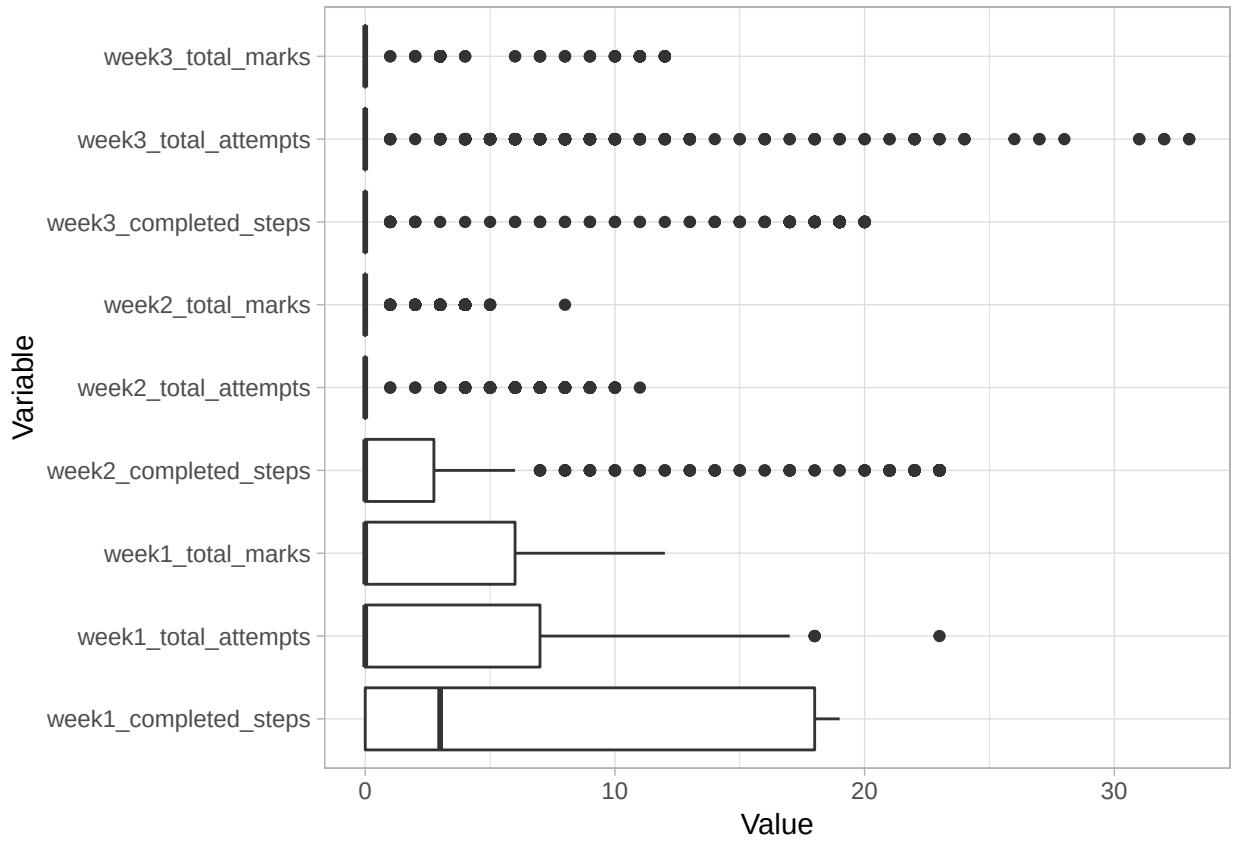

Figure 3: Boxplots for all Numeric Variables

```
}

# plotting with ggplot, use function for later use
boxplotGathered = function(df){
  # plot with ggplot, flip to fit variable names
  ggplot(df, aes(x = Variable, y = Value)) +
    geom_boxplot() +
    coord_flip() +
    theme_light()
}

boxplotGathered(gatherProgress(progressTypeDf))
```

Figure 3 shows some concerning results, as most averages lie around 0 and most other values are treated as outliers (in particular when it comes to marks and question attempts). This problem occurs probably because many learners eventually drop out perhaps, but further analysis is required. In the next boxplot all users who did not try a single question and did not finish a single step in the program are subsetted out.

```
# subset anyone that didn't use the program (no steps & questions)
progressTypeDfSub = subset(progressTypeDf, !(week1_completed_steps == 0 &
                                               week2_completed_steps == 0 &
                                               week3_completed_steps == 0 &
```

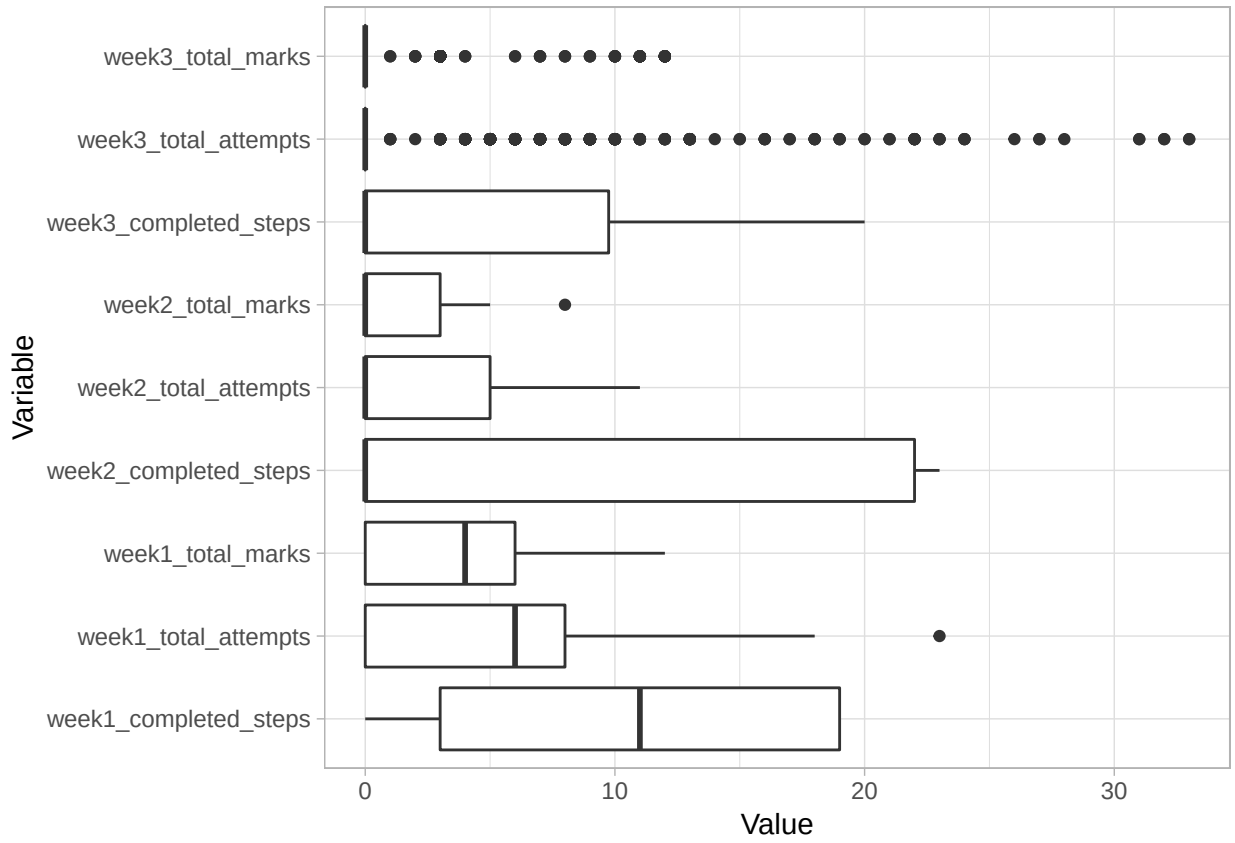

Figure 4: Boxplots for all Numeric Variables (Only who Tried the Program)

```
                                                week1_total_marks == 0 &
                                                week2_total_marks == 0 &
                                                week3_total_marks == 0 ))

# plot like before by using functions
boxplotGathered(gatherProgress(progressTypeDfSub))
```

Shockingly, figure 4 shows that the averages and ranges climb up significantly and that outliers are mostly gone (except in the final 3rd week question attempts and marks), this is probably because many people that are signing up, but are never attempting to use the service at all. Nonetheless, it can also be noticed that:

- The averages are still generally low (many users are still not attempting most questions and steps).
- That generally the questions get less attempts and marks compared to steps judging by the lower ranges and averages (people more inclined to do steps compared questions).
- And lastly, the final weeks have lower averages and ranges (a lot of people stop or attempt less questions and quizzes by the end of the course as people drop out).

4.1.2.2 Boxplots by Archetype

Now that the progress spread and location was analysed on its own, in this section it will be analysed for different given archetypes.

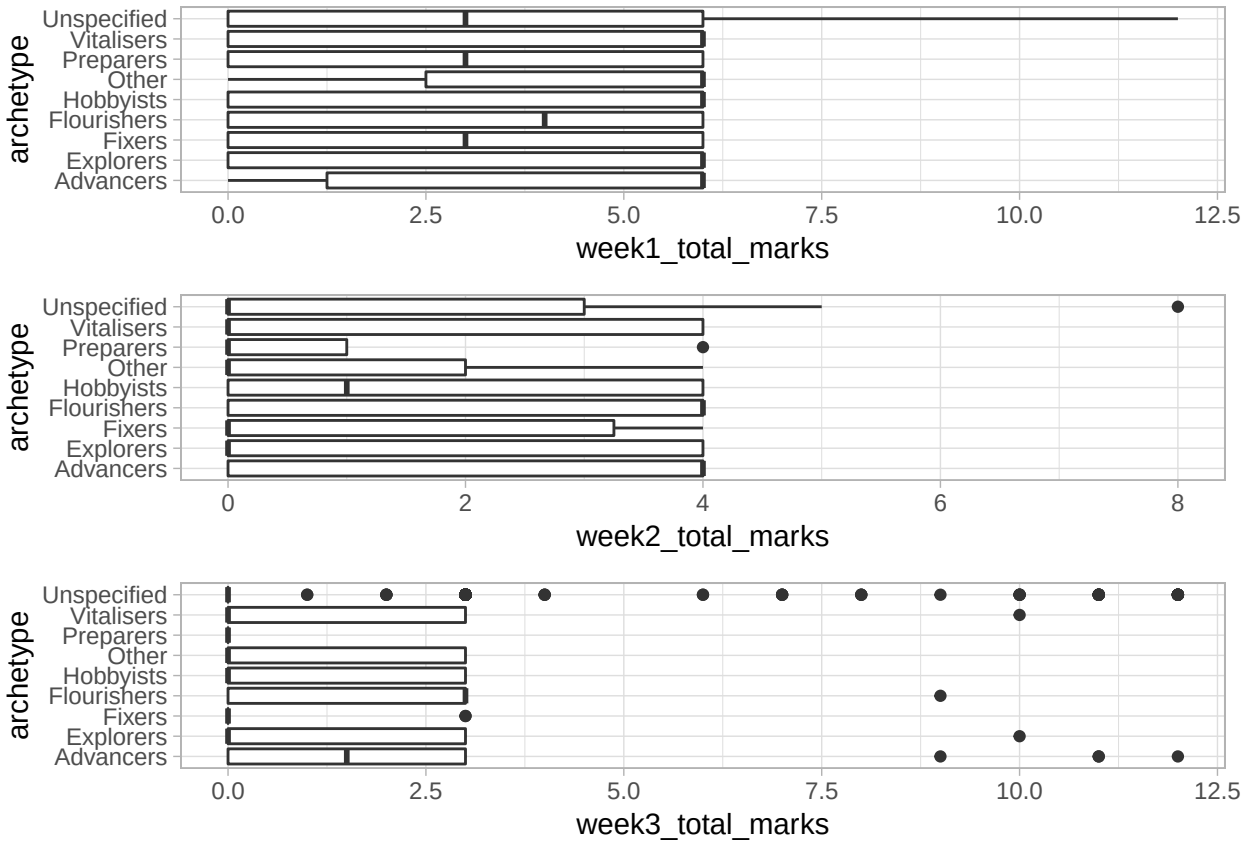

Figure 5: Total Marks by Archetype

```
# main ggplot object
g = ggplot(progressTypeDfSub) + theme_light()

# create three plots, one for each week
p1 = g +
  geom_boxplot(aes(archetype, week1_total_marks)) +
  coord_flip()
p2 = g +
  geom_boxplot(aes(archetype, week2_total_marks)) +
  coord_flip()
p3 = g +
  geom_boxplot(aes(archetype, week3_total_marks)) +
  coord_flip()

# combine using gridExtra grid arrange
grid.arrange(p1, p2, p3)
```

Interestingly in figure 5, it seems that fixers, prepares and learners that didn't specify an archetype dropped their averages first (struggles with tests or outright gave up on them) compared to other archetypes. Moreover, advancers and flourishers seem to generally perform best, in particular by the final week.

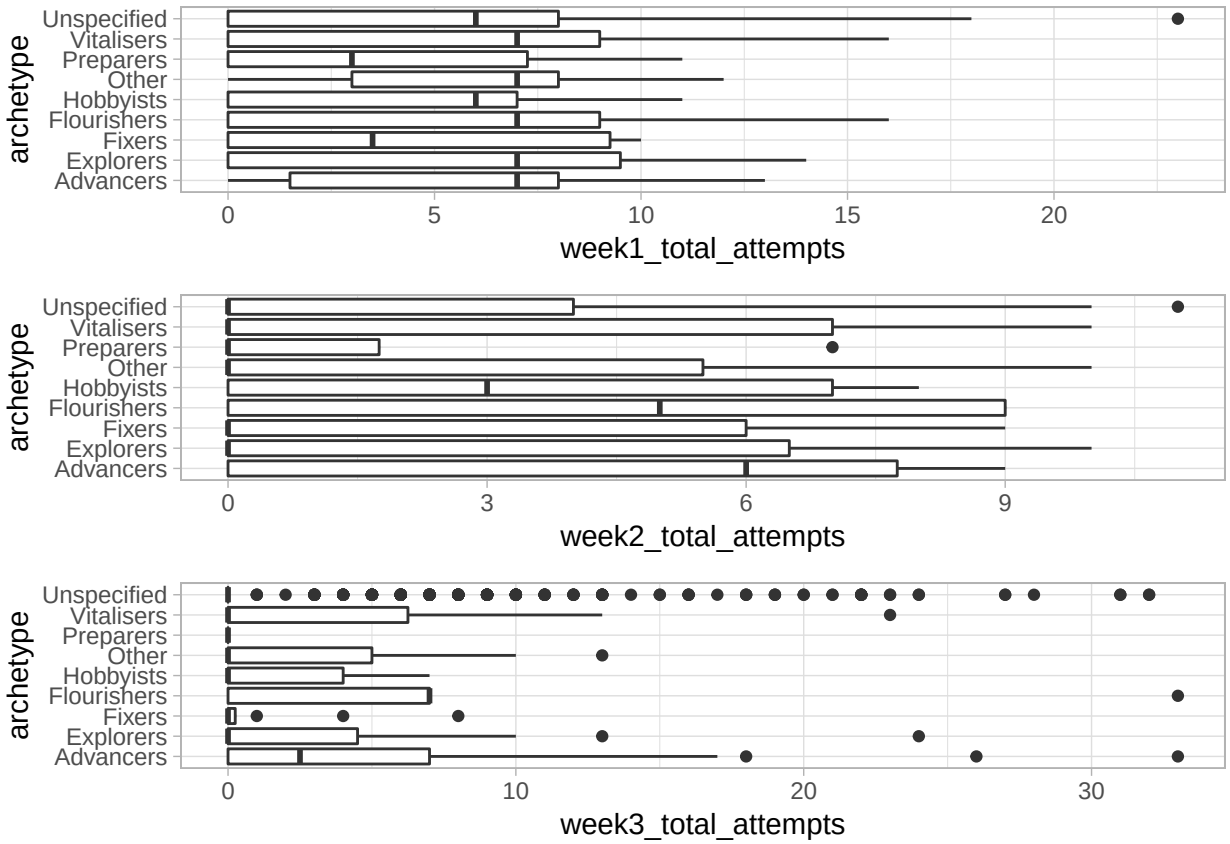

Figure 6: Total Question Attempts by Archetype

```
# main ggplot object
g = ggplot(progressTypeDfSub) + theme_light()

# create three plots, one for each week
p1 = g +
  geom_boxplot(aes(archetype, week1_total_attempts)) +
  coord_flip()
p2 = g +
  geom_boxplot(aes(archetype, week2_total_attempts)) +
  coord_flip()
p3 = g +
  geom_boxplot(aes(archetype, week3_total_attempts)) +
  coord_flip()

# combine using gridExtra grid arrange
grid.arrange(p1, p2, p3)
```

The same patterns can be observed in figure 6, as was in the question marks totals the unspecified, preparers and fixers seem to attempt the tests less than the other archetypes groups, also suggesting that the difference in total marks is brought by a lack of attempts rather than poor performance (after all you will get a greater total if you attempt more questions). It could be that the program is too boring, hard or simply unappealing to these types of learner causing them to drop off (especially by the third week). On the other, advancers and flourishers again perform best, and do not seem to be badly affected by the final week.

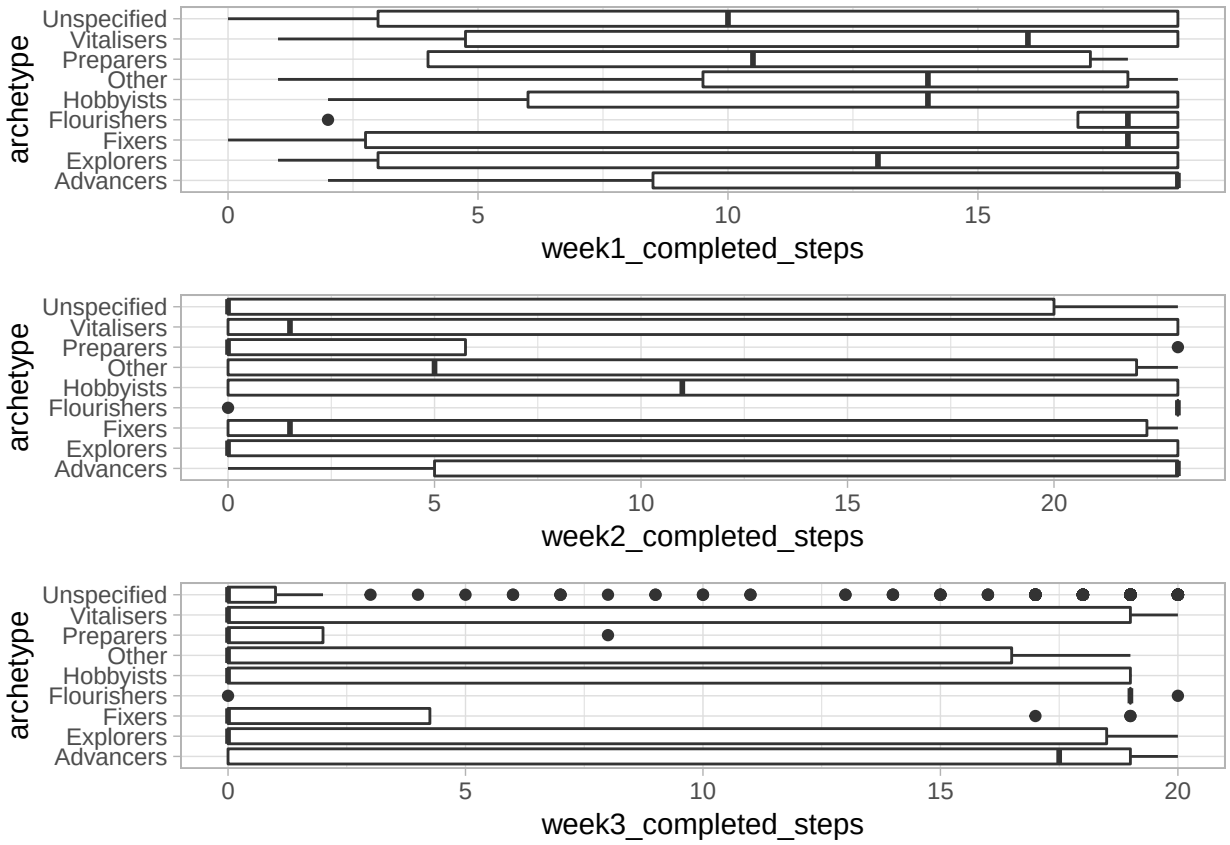

Figure 7: Completed Steps by Archetype

```
# main ggplot object
g = ggplot(progressTypeDfSub) + theme_light()

# create three plots, one for each week
p1 = g +
  geom_boxplot(aes(archetype, week1_completed_steps)) +
  coord_flip()
p2 = g +
  geom_boxplot(aes(archetype, week2_completed_steps)) +
  coord_flip()

p3 = g +
  geom_boxplot(aes(archetype, week3_completed_steps)) +
  coord_flip()

# combine using gridExtra grid arrange
grid.arrange(p1, p2, p3)
```

Again, similar observations are seen in figure 7 to the previous sections, the preparers, fixers and unspecified archetypes are the more likely to attempt less as weeks go on. However, as noticed with the variables box plots, the learners generally seem to be more keep to try out the steps rather than the questions judging by the higher averages and ranges. Also, flourishers seem to perform consistently here, but as seen with the histograms this is the smallest group of learners so this should be taken with a grain of salt, nonetheless

they and advancers still seem to be the best performers.

4.2 Cluster Analysis

To attempt to support the findings in the data exploration section, a cluster analysis will be performed to see how the learner cluster into the archetypes to see if the user selected archetypes for the 7 levels (except others and unspecified) actually divide into 7 groups. Also, the users that didn't attempt any questions and steps are removed since they don't have a score to be clustered with.

```
# subset to only include specfied and not other
clusterDfArch = subset(progressTypeDfSub, archetype != 'Other'& archetype != 'Unspecified')

# remove id and archetypes for k means
# (can't do PCA on them)
clusterDf = clusterDfArch[-11]
clusterDf = clusterDf[-1]

# find cluster using k means with k = 7 (others and unspecified not included)
km = kmeans(clusterDf, 7, iter.max=50, nstart=20)

# Perform the PCA:
pca = prcomp(x=clusterDf)

# Create a dataframe with pca results, cluster and archetype
pcDf = data.frame(firstPC = pca$x[,1], secondPC = pca$x[,2],
                  cluster = km$cluster, archetype = clusterDfArch[11])

# plot, be sure to use manual shape, jitter and alpha to help
# distinguish points
ggplot(pcDf, aes(firstPC, secondPC, color = factor(cluster), shape = factor(archetype))) +
  scale_shape_manual(values=1:nlevels(clusterDfArch$archetype)) +
  geom_jitter(alpha=0.5, size = 2.0)  +
  xlab("First Principle Component") +
  ylab("Second Principle Component") +
  theme_light()
```

Unfortunately figure 8 shows that the data does not divide into 7 clusters with the appropriate users given the current variables. However, there does seem to be some clustering (vitalisers seem to generally cluster next to one another), so perhaps the current set of variables or types of learners need to be changed to find better clusters.

4.3 Conclusions and Data Quality

To conclude, it seems that generally the best performing archetypes are the advancers and flourishers when it comes to steps and questions. On the other hand, the worst performers seem to be the unspecified, fixers and perparers, as they are more likely to drop off by the third week. This could be because advancers and flourishers are more experienced or because the program bores the poorly performing learner types. Moreover, since people that answered the archetype survey are more likely to be interested in the program, it is no surprise that they performed better than who did not answer it (Unspecified).

Moreover, there seems to be a significant number of people that drop off by the third week, and people are general not very keen on doing questions compared to steps. Perhaps questions and third week material are too hard or uninteresting.

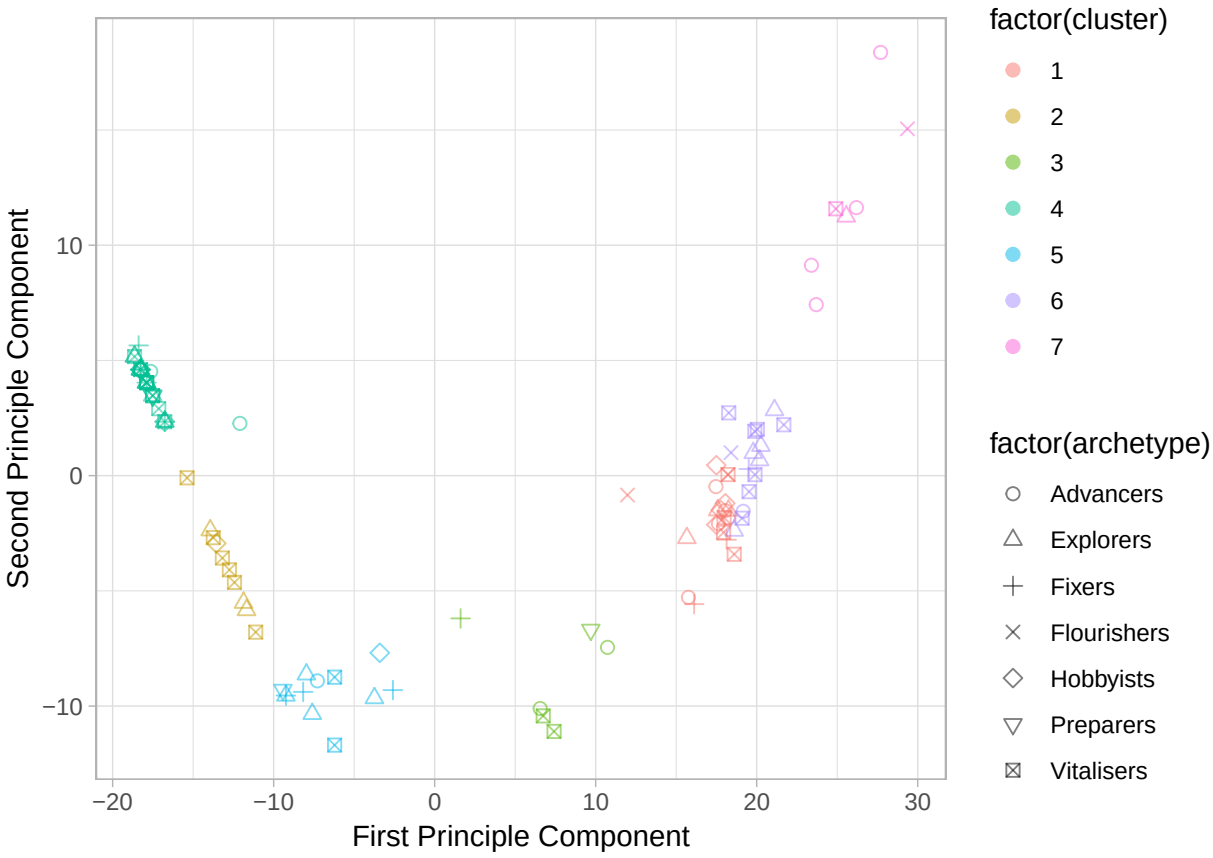

Figure 8: Clusters for the 7 Selectable Archetypes Separated by PCA

However, some problems exists with the data, in that there is not enough evidence to prove that the learners divide into 7 groups given current data and variables. Moreover, there is a really significant number of people dropping off by the third week causing a lot values to be pegged back. In addition, not many people recorded they archetypes by doing the survey, perhaps because the survey was not advertised properly or might be better taken during sign up. Finally, it is important to note that the data had some issues with NA values, in particular with questions dataset which had NA learner ids (as seen in the Data Preparation report).

Suggestions for further research including looking into other variables like video stats, comparing archetypes to locale and investigating the drop off point at the 3rd week.

5 Second Cycle

5.1 Introduction

As suggested in the conclusions, the next cycle looked into further research into comparing archetypes to locale and the third week drop off point. Video stats however fell in favor because no individual video stats were found. Nonetheless, using the 7th run enrollment data, the detected_country was added as a field. For the 3rd week steps analysis, the previously collected data can be used.

5.1.1 Country Analysis

5.1.1.1 User Countries Frequencies

```
# Tabulate
tab =  table(countryProgressDf$detected_country)

# sort table
tabSorted = sort(tab)

# Find top 10 results
top10 = tail(names(tabSorted), 10)

# subset
topCountryProgressDf =
  subset(countryProgressDf, detected_country %in% top10)

# order factor levels
topCountryProgressDf$detected_country =
  factor(topCountryProgressDf$detected_country, levels = rev(top10))

ggplot(topCountryProgressDf) +
  theme_light() +
  geom_bar(aes(detected_country))
```

Figure 9, shows that most users come from the Great Britain. The users from Great Britain seem to dwarf users from other countries (about 3 times as much as the second country!). Nonetheless, the other two top countries are India and South Africa.

5.1.1.2 Top 6 Countries Archetypes

```
# Find top 6 results
top4 = tail(names(tabSorted), 6)

# Remove people who didn't respond
top4Archetypes =
  subset(topCountryProgressDf, detected_country %in% top4 & archetype != 'Unspecified')

# plot country against count with archetype bars
ggplot(top4Archetypes) +
  theme_light() +
  geom_bar(position=position_dodge(), aes(x = detected_country, fill = archetype))
```

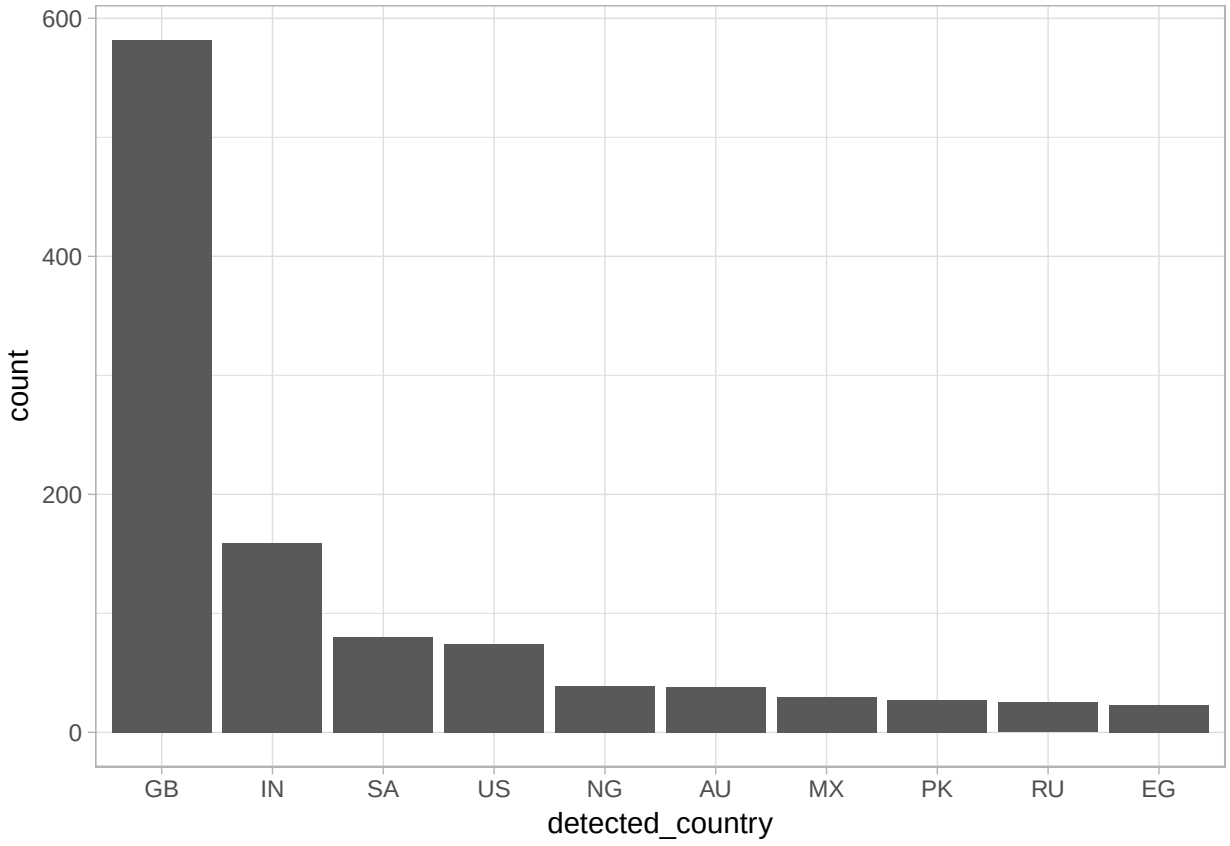

Figure 9: Number of Users by Country (Top 10)

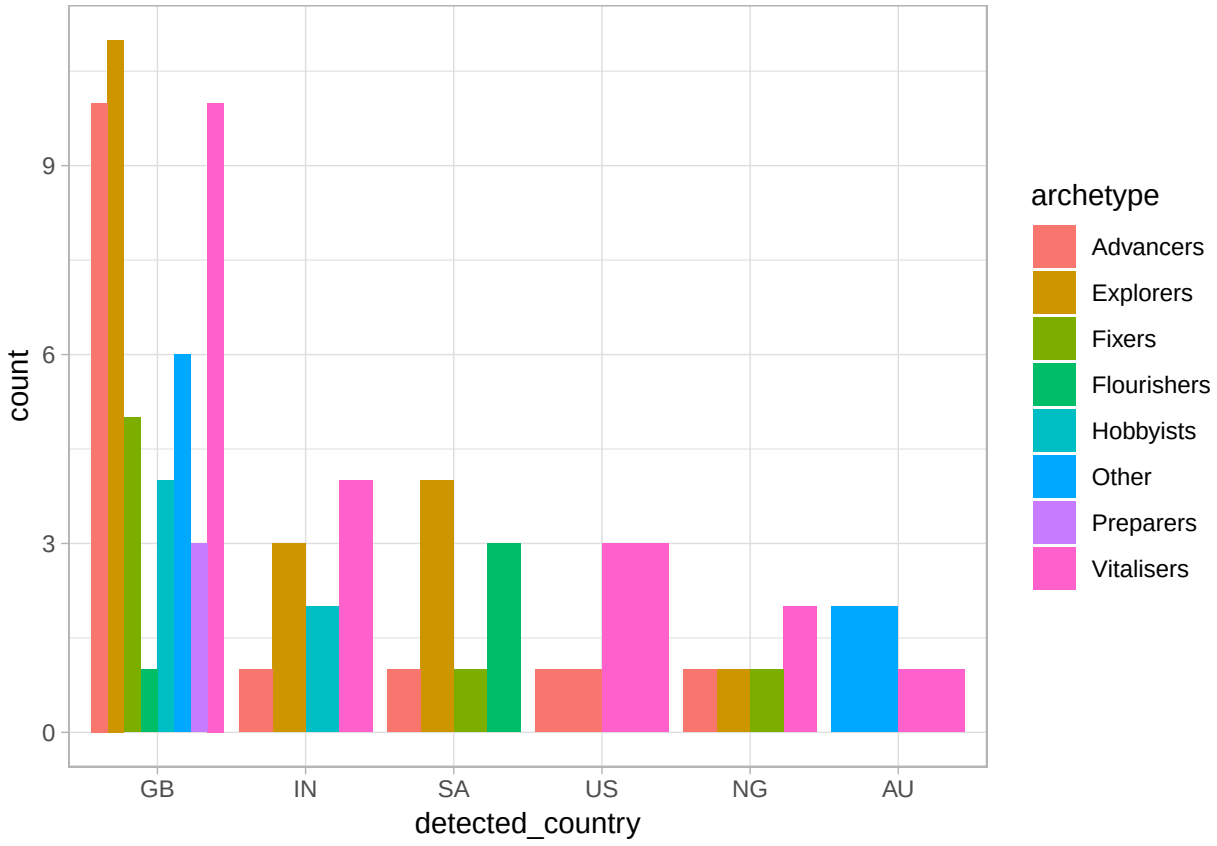

Figure 10: Archetypes by Country (Top 6)

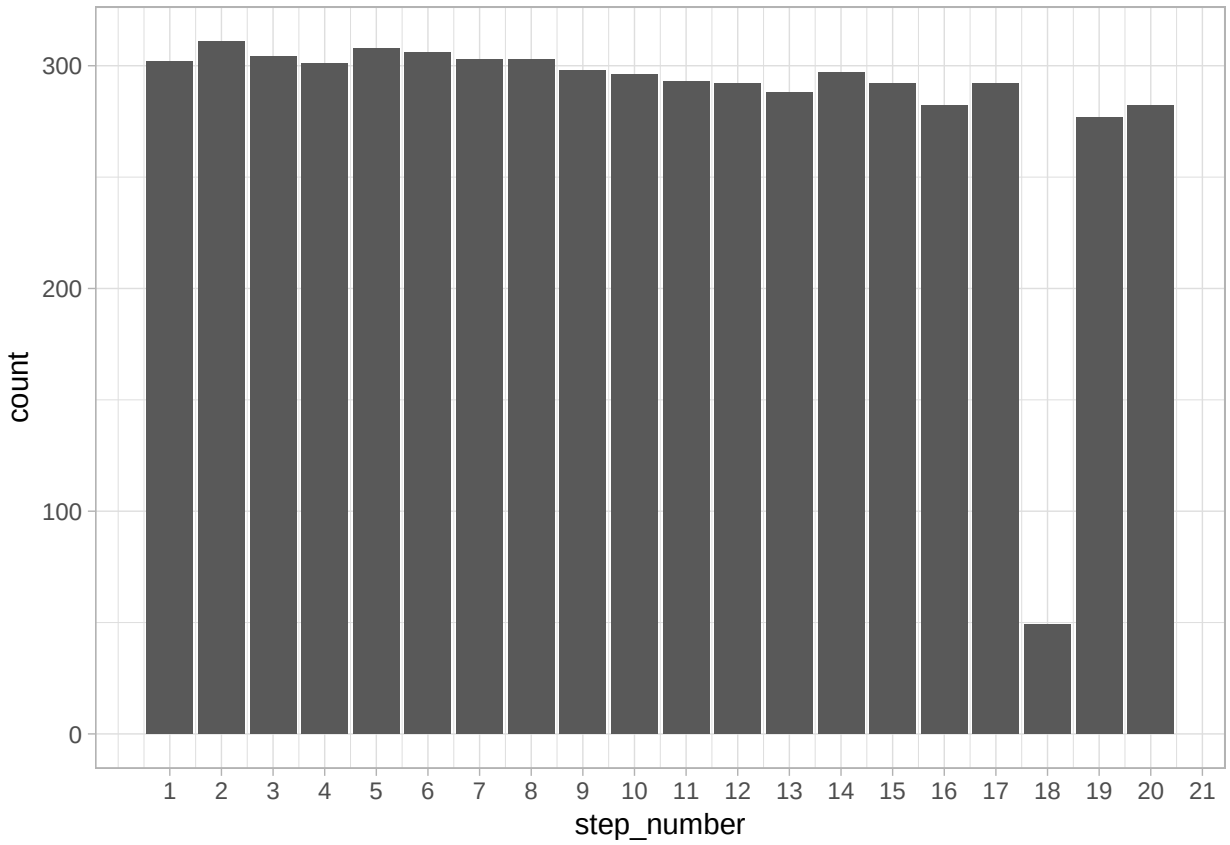

Figure 11: Number of Users in Steps in week 3 (completed/started)

Figure 10 shows that Great Britain seems to have the most varied number of users, which is not surprising considering that most users come from there. Nonetheless, most users from Great Britain are explorers, advancers and vitalisers (in order). When it comes to the other countries, it can be seen that South Africa has a significant number of flourishers compared to other countries (where flourishers are usually not even in the top 4). Moreover, only users from Australia and Great Britain chose 'Other' in the survey.

5.1.2 Steps (3rd week)

```
# plot week 3 steps (cleaned not aggregated) compl/start
ggplot(subset(stepsDf, week_number == 3)) +
  theme_light() +
  geom_bar(aes(step_number)) +
  scale_x_continuous(breaks = seq(
    min(stepsDf$step_number), max(stepsDf$step_number), by = 1))
```

Figure 11 shows that it seems most users skipped the 18th step while the other steps seemed to be done by nearly all users who did week 3 steps. The 18th step is a test.

5.2 Second Cycle Conclusions and Data Quality

The conclude, it seems that the absolute majority of users are from Great Britain, making the analysis difficult. Nonetheless, the most interesting findings were that the most varied archetypes can from Britain and that South Africa had the most flourishers (perhaps there is a booming interest there in cyber security within non experts?). Nonetheless, as stated before this should be taken with a grain of salt, as not only Great Britain dominates the counts, but the counts from other quite small as well. A small issue of note is that there was one learner with a lot of NAs (progress and countries) after the last join, but it is unlikely that their removal has a massive impact (see Data Preparation).

When it comes to the 3rd week steps, it seems that the test in the 18th step in the third week had weak interest. A lot of learners were too intimated from the test in the 18th step and could not complete it or even bother starting it. Perhaps it is too difficult or uninteresting, but either way it seems like it needs revision.