Learning analysis: Exploring the interactions and demographics between different demographics

Adanna V. Obibuaku (180251870)

21/11/2021

1. Indroduction

1.1 Background

The report follow the CRISP-DM methodology to gain an insight of the online courses provided by Newcastle University, specifically the *Cyber Security: Safety at Home, Online, in life* online course hosted *FutureLearn* website. The main business goal is to enhance Newcastle University's online resources. This is done by motioning the performance and engagement of certain users. The *Cyber Security: Safety at Home, Online, in life* has several a total of 7 different runs of the online course between the years September 2016 - September 2019. *FutureLearn* has collected several data sets for each run of the course. Utilizing this data means to gain an insight of pupils engagement and performance and provide appropriate intervention in areas which are limited. In return this will enhance the quality of the online course. To achieve this we examine the engagement and performance of the online course using the collected data. The type of demographics this report focuses on includes:

- Age range
- Employment status
- Region
- Gender
- Highest education level.

The report looks closely the demographics of pupil learning the course, their interactions, engagement and performance Additionally, it explores reasoning behind the performance of each group; trying to identify constraits a demographics may have for engagament or learning the online course.

Additionally the report includes two models:

1. The first model takes uses two data sets; demographics data (age, gender etc). and usuage of data which gives an insight about student engagaments (number of weekly activities/quizes etc), to predict the performance of a user.
2. The second model makes usues of data which provides the average performance of a specific demographics across different runs of the module to predict the expected performance of a specific demographics for the next run of the module

1.2.1 The Data

The intial data is collected by the *FutureLearn* organisation. There are a total of 53 files. This is collected over the 7 runs of the online course. However we only need a few subset of data from the data set:

1. The enrollment of each run of the course. This includes fields such as age, gender, country etc. This will help exploring the performance between different demographics.
2. The step activity for each run of the course. This includes fields specifying the number of activities competed by a pupil. This allow to measure the student engagement.
3. The quizzes for each run of the course. This includes fields specifying the number of quizzes by the user. This allows us to measure the student engagement. Including the performance of a user.

An additional data set is used collect by the Kaggle wesite. This provides country mapping by ISO codes to continent regions. This data set is used in combination with the enrollment course, where an additional row called "region" to map the detected country of a pupil to a continent. This would provides a brief summary of where the students are from.

Furthermore, the data set collected for the 7 runs of the online course is combined into 1 data set. This provides a summary of the course data. For example, figure 1 illustrates how the enrollment data set collect for each run is combined.

2. Methodology

This section is going to be used to explore the data, to identify and relationship and trends within the data.

2.1 Verify data quality

The data sets that is going to be used to analyse: *Exploring the interactions and demographics between different demographics*. These data set mostly had some columns which were empty fields which were unknown. For example in *enrollment* data set collect for each run:

- Some rows of Learner_id were empty. Rows with empty learner_ID were simply removed.
- The "unrolled_at" column hand many missing rows. This column was not used within our analysis, therefore not a problem.
- The "purchased_statement_at" column was completely empty. However, this is not an error it indicates that the online course was free. Therefore, there was no purchases.
- Column describing the type of user (gender, age) rows were labeled "Unknown". This suggest that the pupil did not answer it. These fields were removed from out data set.

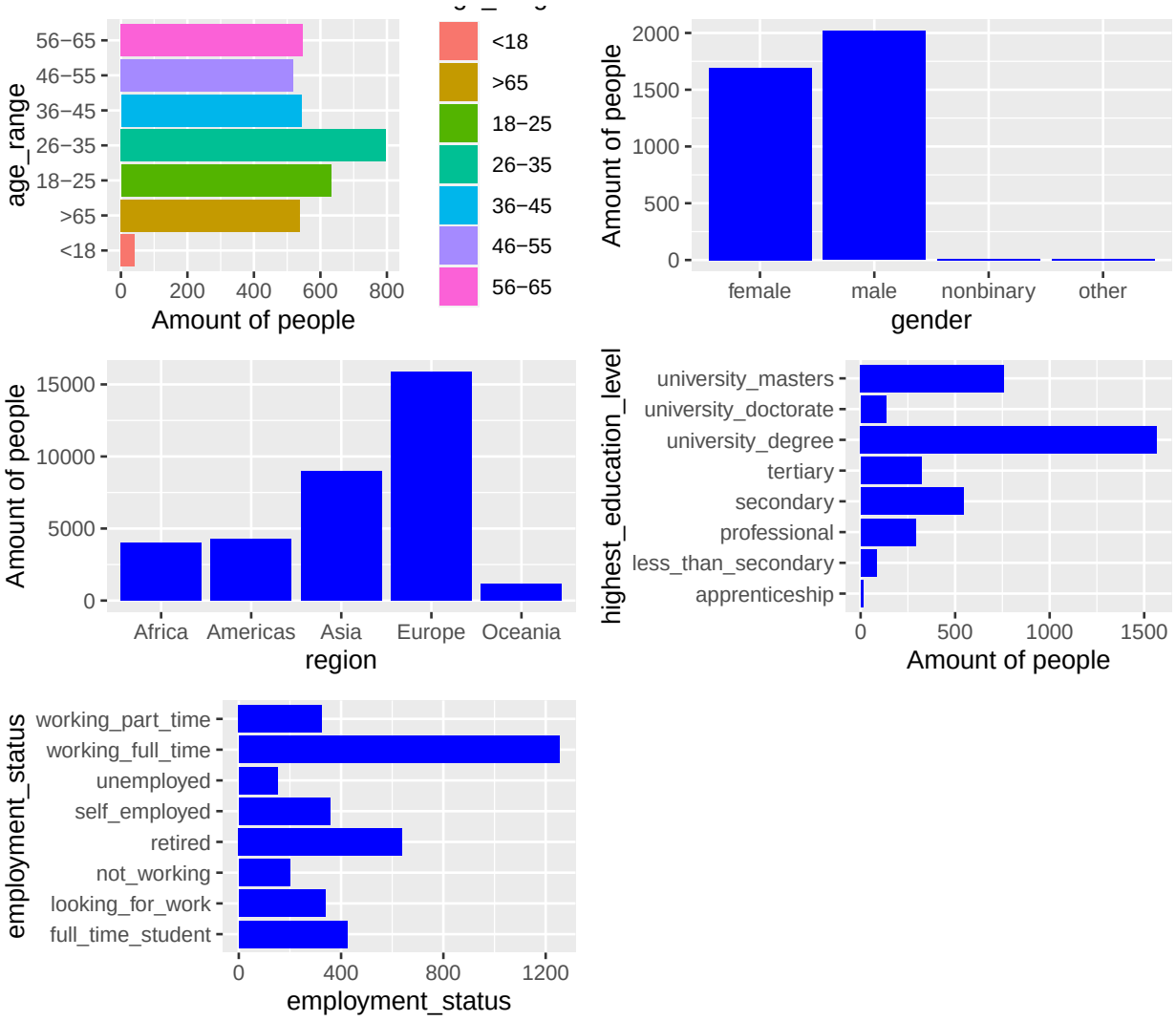

Figure 1: Bar graph illustrating the number of enrollments between different demographics

2.2 Monitoring pupils attendance between different demographics

Here we are looking at the amount of pupils enrolled within the online course, between different demographics.

This graphs is a summary of the total 7 different runs of the course.

From the data it is found the highest number of pupils enrolled in the online course are aged 26-35. While, the lowest people less than 18. More male pupils are enrolled in the course more than female. Approximately 2000 males and 1750 females enrolled. In terms of locations, Europe has the highest number of pupils enrolled in this course, the second highest being pupils from Asia. Lowest, is Oceania. Pupils enrolled in this course, mostly have at least a university degree and are working full time.

Overall, the data illustrates that a significant amount of people who choose to enroll in this course appears to be between the ages 26-35, working full time with a high university degree from Europe.

2.3 Monitoring pupils performance between different demographics

2.3.1 Illustration of the performance across each run of the module

The graph is an illustration of the demographics of pupils performance across different runs of the online course.

It appears that females pupils performed better than male pupils on runs 2, 4, 5 and 7. Male pupils performed better on runs 1, 3 and 6. Pupils who identify as non-binary tend to perform better than male and female pupils. However, there the population of non binary pupils within the data, is not large enough to support this claim. Pupils from countries in European and Oceania performed better than pupils from countries Asia, Africa and Americas. European pupils appears to perform consistently the same across the 7 runs of online course, while Oceania significantly performed better than all the other regions on run 4 of the course. Those who are under 18 appear to significantly perform the best. However given that the population of pupils under 18 within the data is small, this claim is not supported. Therefore, we disregard this data. Pupils aged 56-65 and 65> tend to perform better.

There is no significant distinction in who performs between pupils with different forms of highest education. However, those with a highest form of education being a university degree appear to perform consistently through each run. Across each run, there does not appear a significant distinction in performance between pupils with different employment status. However, pupils who are retired perform at a consistent rate through each run of the course.

2.3.2 Illustration from the total run

This summaries data from the different runs of the online course, to compare the average and distribution of different demographics performance.

There is more variance in the gender of male students than they are for female students. Furthermore, The mean of male and females students are approximately the same. Pupils from Europe

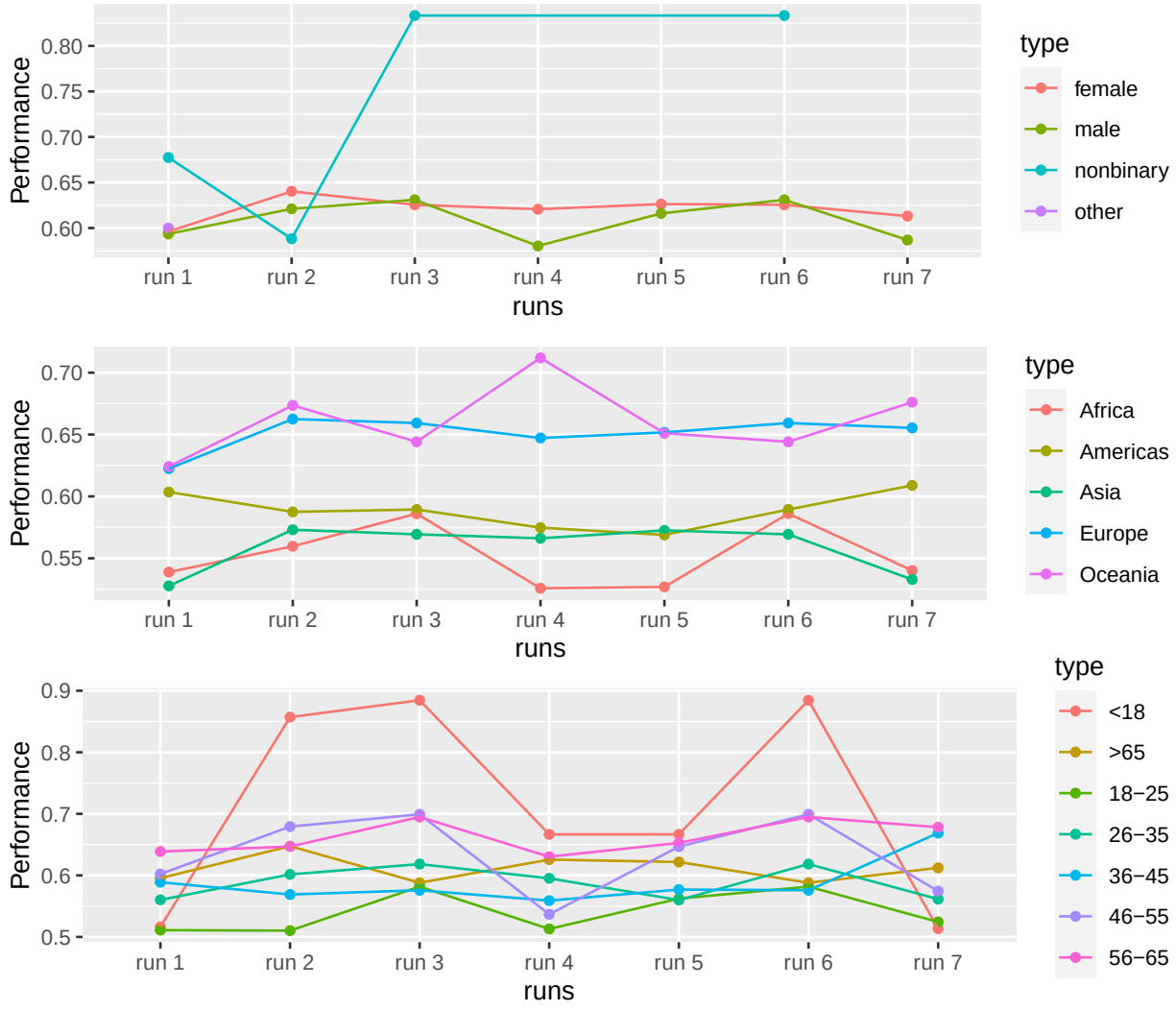

Figure 2: The average performance across different runs of the module compared between different demograpics

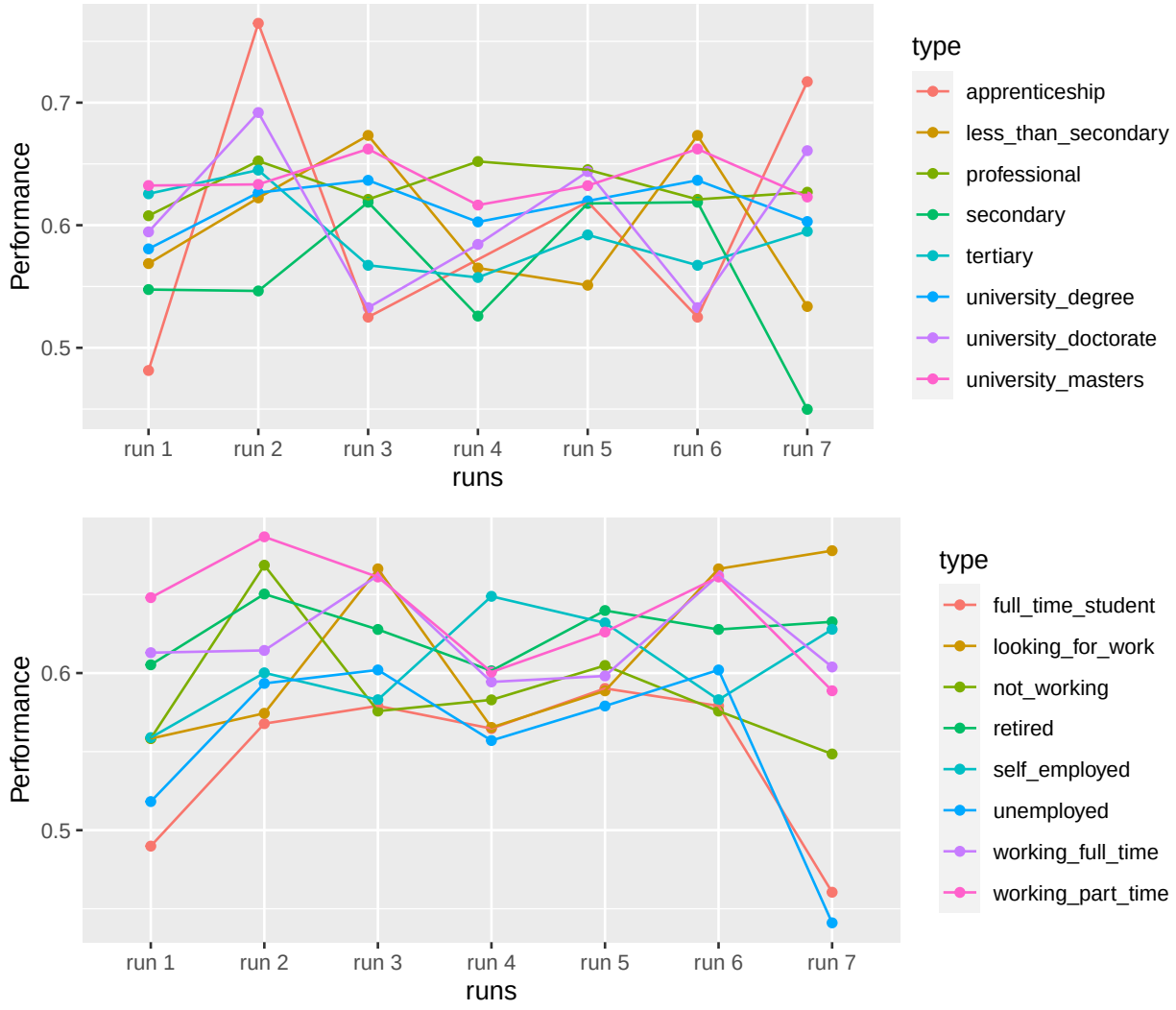

Figure 3: The average performance across different runs of the module compared between different demograpics

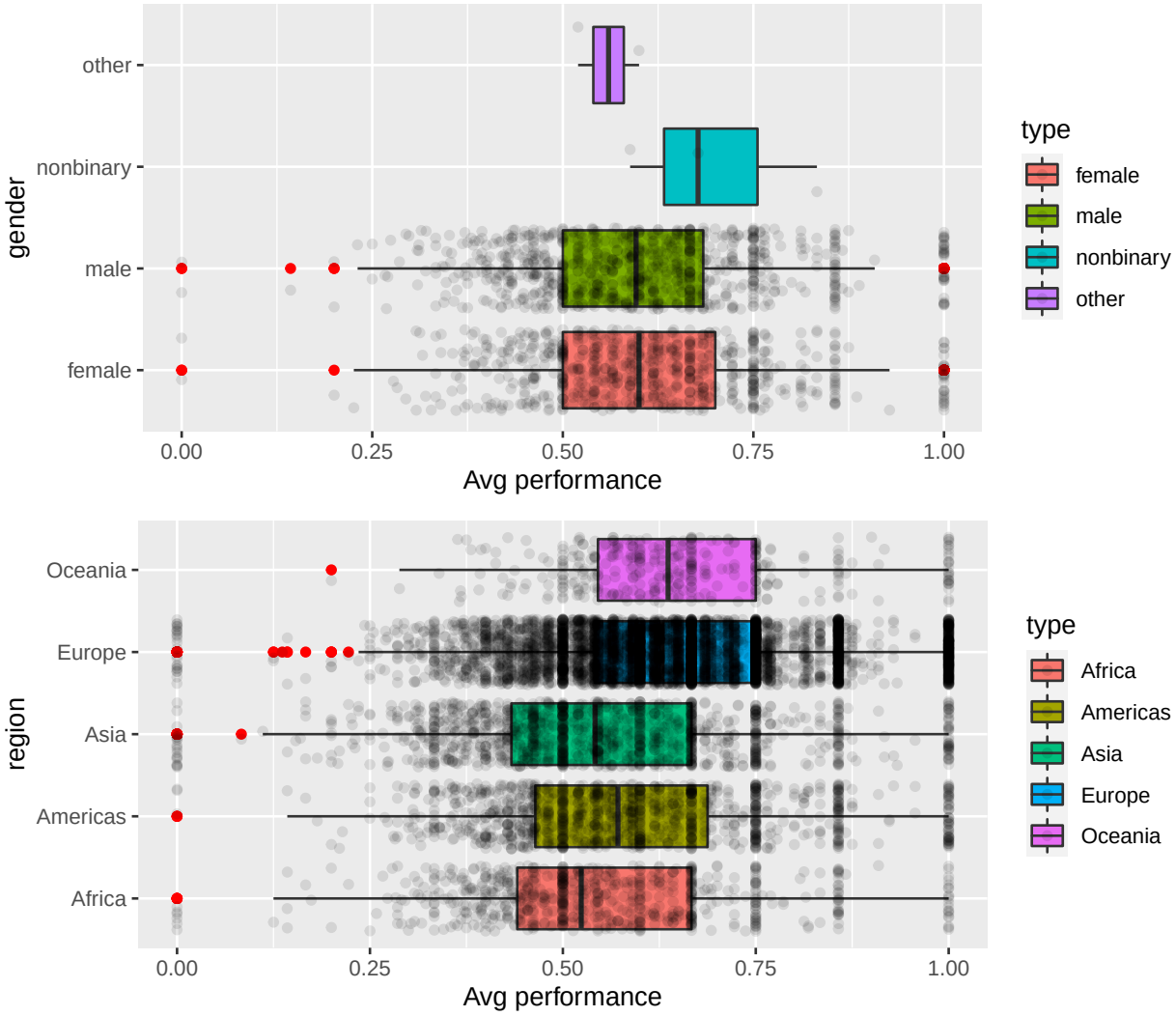

Figure 4: Boxplots illustrating the distribution of the average performance between different demographics

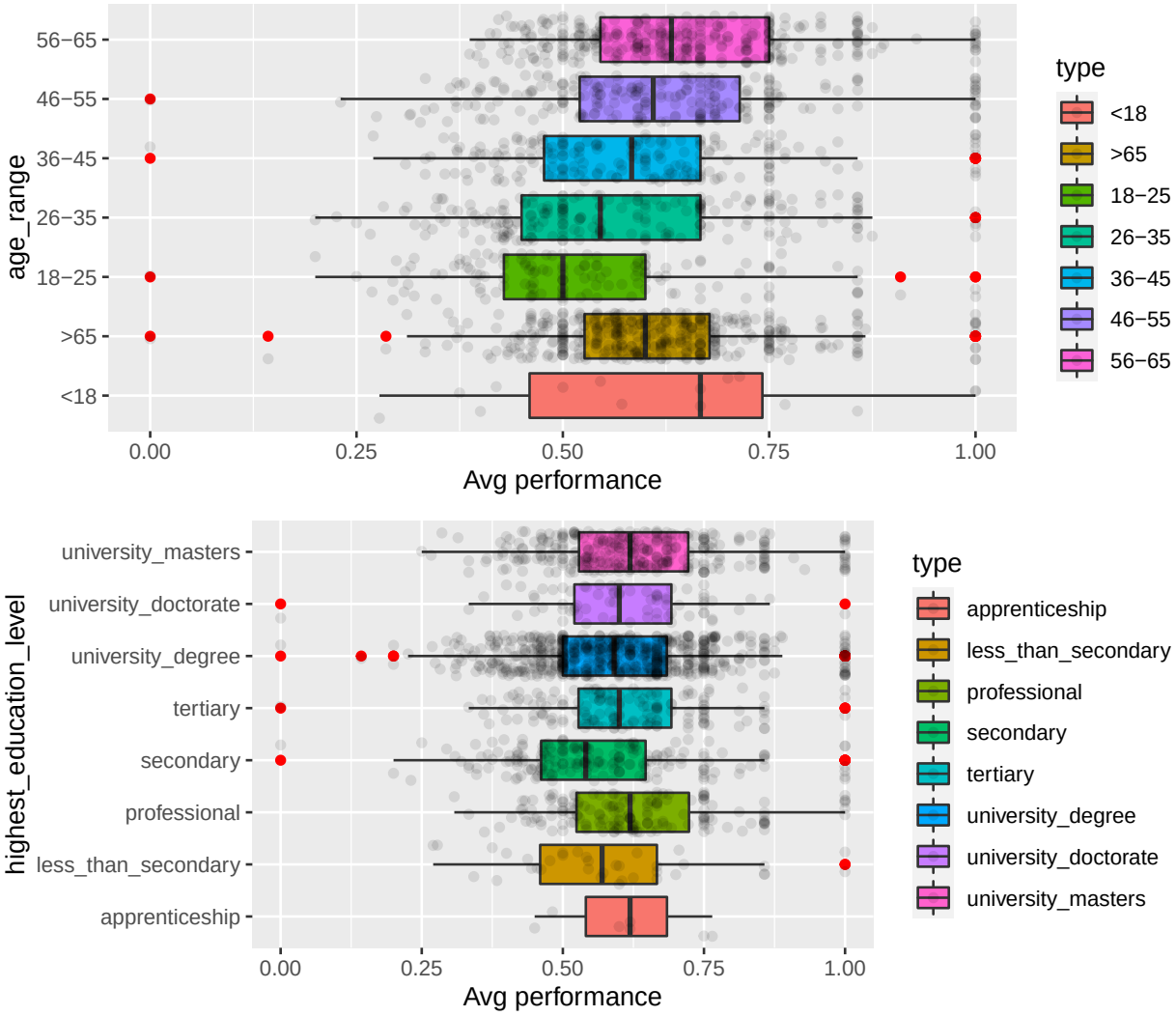

Figure 5: Boxplots illustrating the distribution of the average performance between different demographics

have a higher mean than pupils from other regions. Additionally the data suggest that pupils from Europe are suited at the higher end of performance.

As mentioned before, there is not enough pupils that are less than 18 to precisely define them. Therefore, this information is disregarded. However, from the data it shows that pupils who are more than 65 have a average higher performance. The second highest average performance are those who are between ages (56-65). Furthermore, pupils who are more than 65 have a low IQR, suggesting that the scores between each individual is similar/consistent. Pupils who are ages between 18 and 25 have a lower performance average. The graph, illustrates that the distribution of pupils aged 18-25 are situated at the lower end the average performance. The pupils with the highest education level being an apprenticeship have the highest average performance. However, as the data from those who have an apprenticeship is not enough to accurately represent the population, the data is disregarded. Pupils with a university have the highest average performance, while the second highest average performance comes from pupils with a tertiary background. Pupils who have obtained a secondary schooling at a highest level education have an low average performance.

Pupils working full time have the highest average performance. This is followed by pupils who have retired. The distribution of people who have retired have a small IQR. This suggest that the performance obtained by each individual is fairly similar. Pupils who are full time students have a low average performance. Additionally the distribution of full time students is situated at the lower end of the average performance.

2.4 Monitoring pupils engagement between different demographics

There is no clear strong pattern between the number of activities completed. However there it shows a slight negative correlation between the number of activities and the performance of each student. This is looking at student engagement across different demographics. Student engagement is measured by the average number of quizzes completed in a week including the average number of activities completed in a week.

It appears that pupils ages 65 have the highest number of activities completed. The second highest is those who are aged 56-65. The lowest number of activities is completed is by pupils aged 18-25. Second lowest being pupils from age 26-35. Those who perform the highest number of quizzes per week are those aged 26-35. Those who perform the lowest number of quizzes per week are those aged 56-65.

Those who have retired have a higher number of number of activities completed on average. The second highest is those who are not working. Those who are retired complete the lowest amount of quizzes per week, while full time student complete the highest number of quizes per week.

Pupils who are either non-binary complete the highest number of activities and lowest number of activities. However since the sample size from this population is small, the results are inconclusive. This hold true for pupils who identify as “other”. Female pupils tend to have complete more activities in a week than Male pupils. The data further illustrates the female pupils complete less quizzes in a week than male pupils.

Pupils who have an apprenticeship tend to complete more activities and less quizzes than all the different demographics of higher education. However, given that the sample of the population collected for pupils in apprenticeships is small. This data is disregarded. Therefore, pupils who

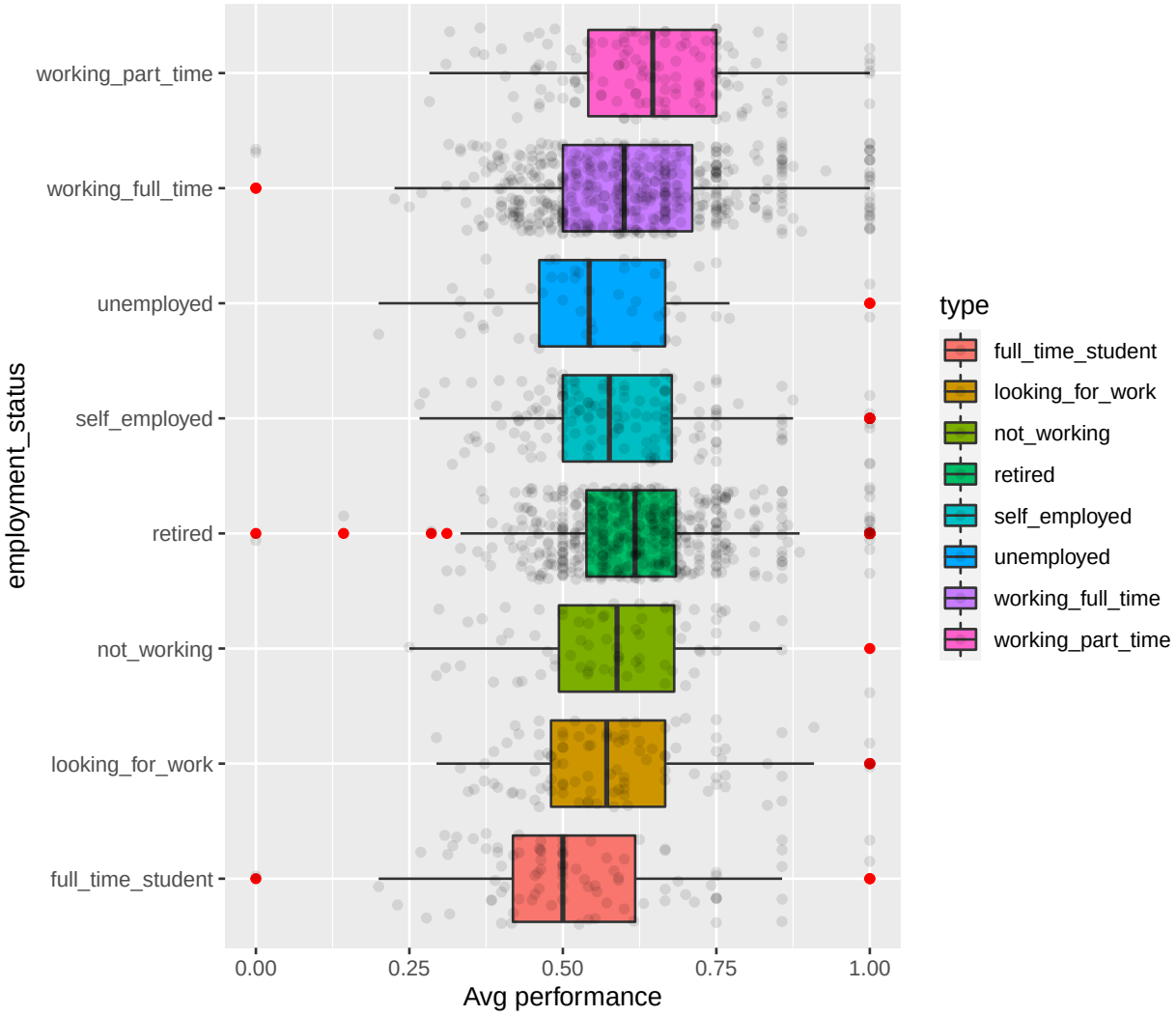

Figure 6: Boxplots illustrating the distribution of the average performance between different demographics

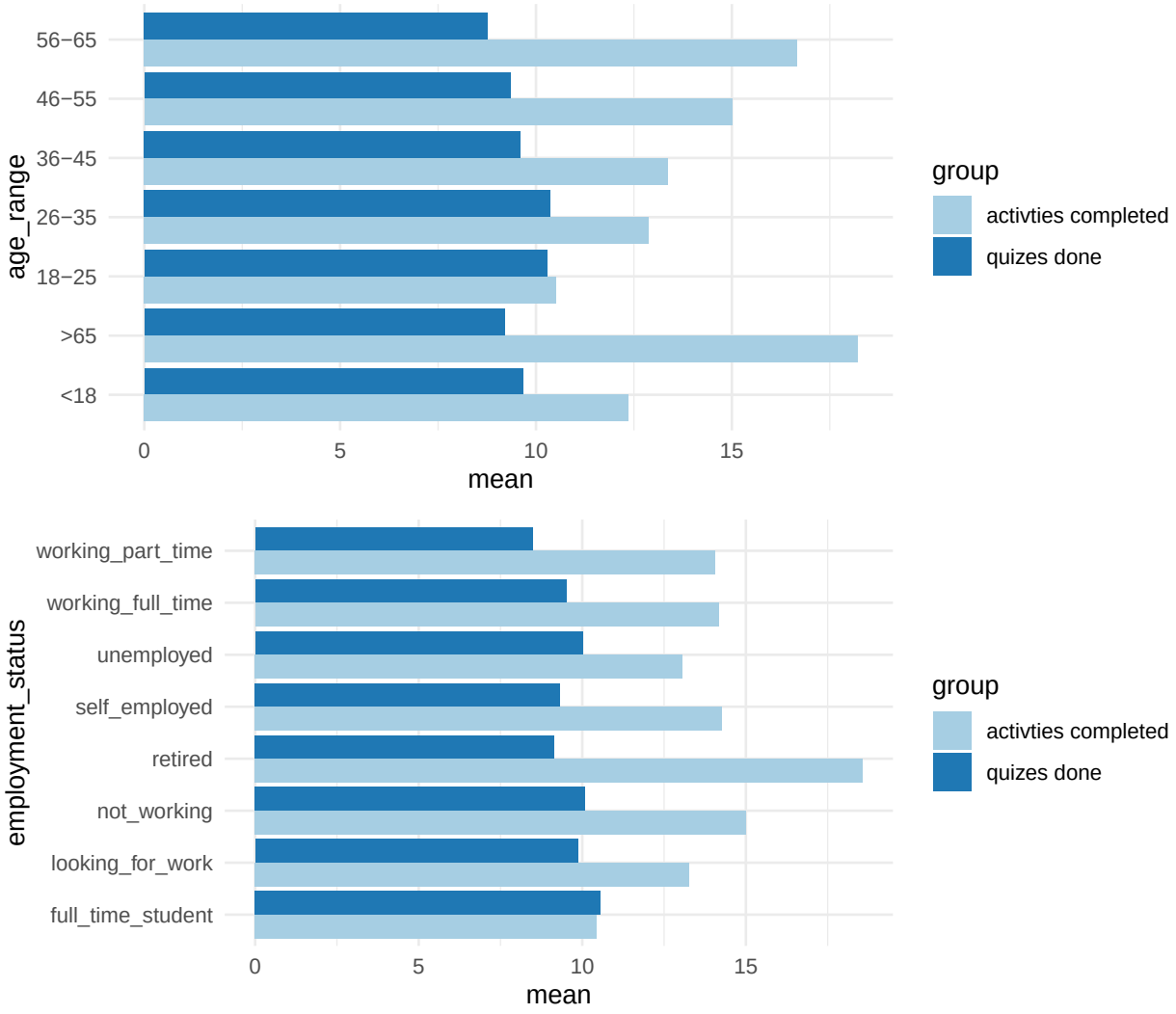

Figure 7: Bar graphs illustrating student engagement measured using number of quizes and number of activites completed.

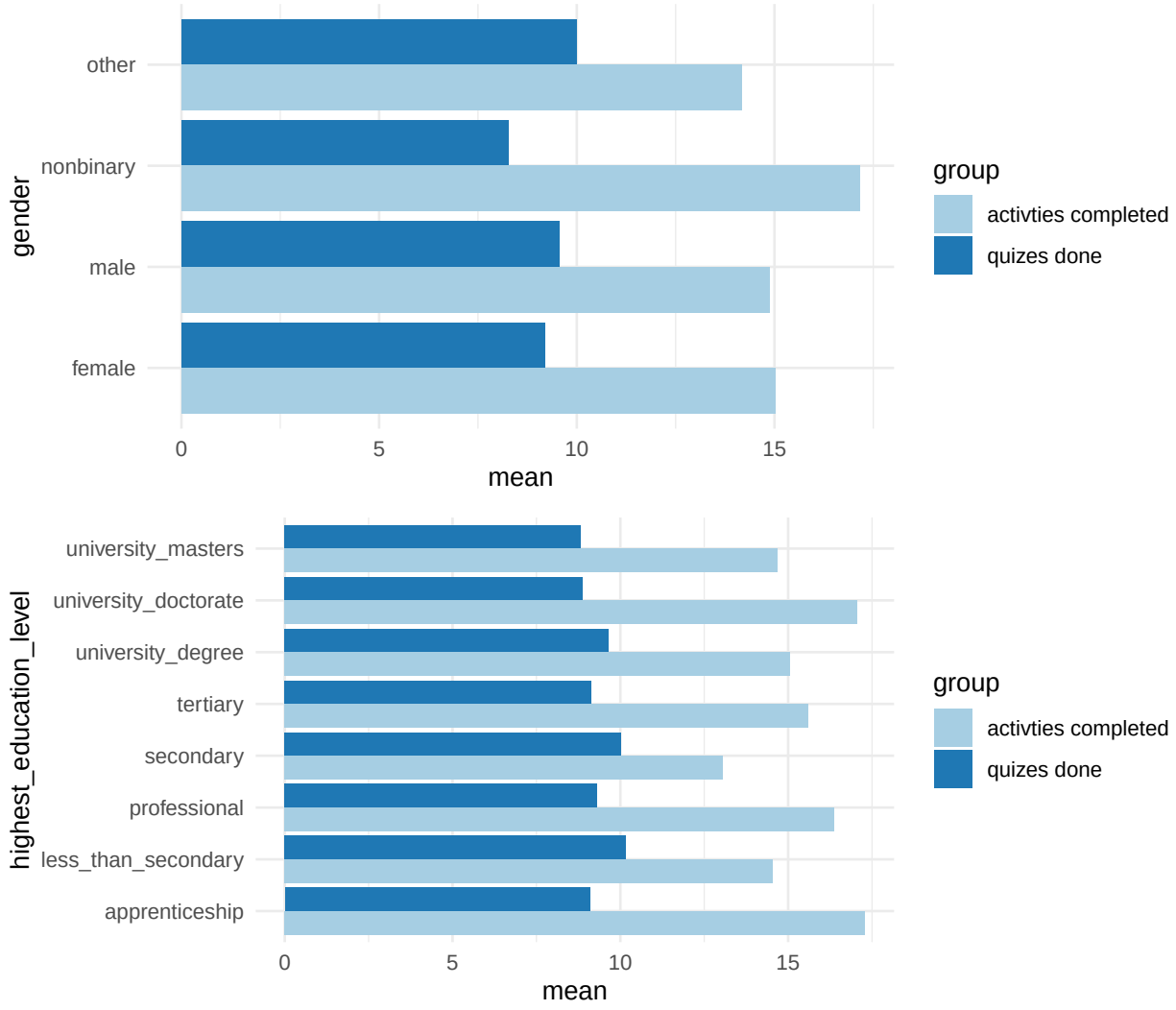

Figure 8: Bar graphs illustrating student engagement measured using number of quizes and number of activites completed.

have a university doctorate degree have the highest number of activities completed in a week The second highest are Pupils who are professionals. However, professional pupils complete a lowest number of quizzes in a week.

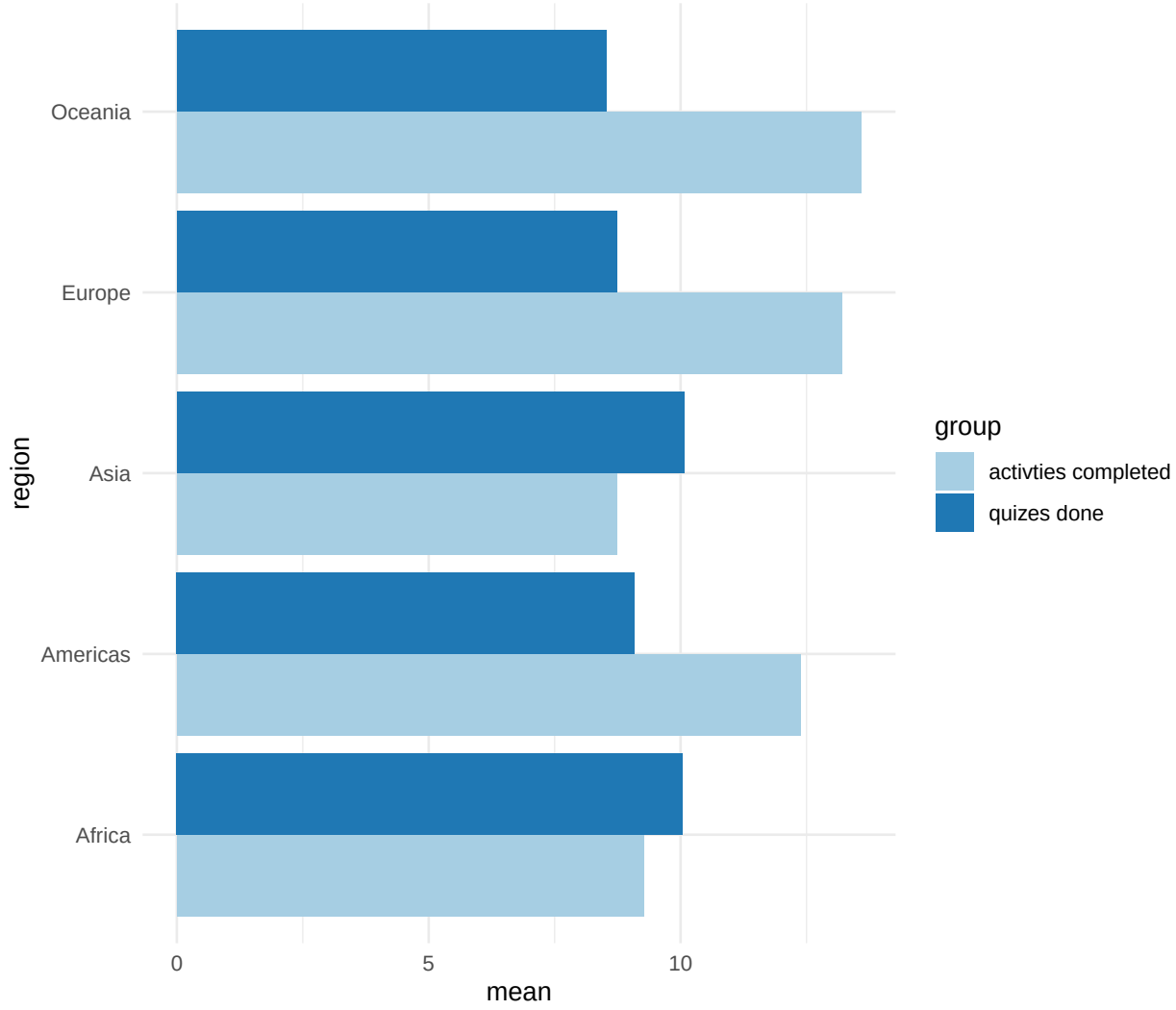

Figure 9: Bar graphs illustrating student engagement measured using number of quizes and number of activites completed.

Pupils from countries in Oceania complete the highest number of activities a week. The lowest number of activities completed are pupils from countries in Africa. Pupils with the lowest number of quizzes completed are from countries in Europe. Pupils with the highest number of quizzes are from countries in Africa

2.5 Findings and Reasioning

The people who had a better higher performance included:

- Pupils aged 56-65 and 65 tend to perform better.

- Those who are retired and or working full time perform better
- Full time students perform the lowest
- Pupils with university degree perform better
- Pupils from countries in Oceania

Generally it was observed that a pupil will complete more quizzes per week have a lower performance. This is because pupil with a higher number of quizzes tend to make attempts in multiple combination until an answer is correct. This suggest that they may have not gained the right understanding and are most likely guessing. Additionally, it's observed that the number of activities completed and performance has less of a negative relationship than number of quizzes completed per week and performance. Perhaps this is indication why some demographics (>65) had a higher performing (as they took much less attempts to answer questions) and activities. Activities take more effort than quizzes, as involves self research to find an answer. That indicates those who are more engaged take more activities, thus gaining a better understand and a higher performance in quizzes (less reattempts).

Full time students had a lower performance and completed more quizzes than any pupils with different employment status. This holds true for pupils with the age range 18-25. Retired people tend to perform better among est those with different employment status. The data also suggest that pupils >65 perform the best. Therefore, it is reasonable to suggest that full time students tends to be those 18-25 are less engaged because as they are completing other studies.

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

2.6 Models

We are going to be looking at implementing two models

1. Predicts the number performance by each student
2. Uses data to estimate the performance of a student

2.6.1 Model 1: Linear Model to predict pupil performance

This model take in two main data sets - demographic data, including the gender, age, highest level of education, employment status, region and employment area. Additionally makes use of the data specifying the number of engagements. This is measured by the number of quizzes and the number of activities completed by a pupils week by week. The data sets are split into 2 parts: One to be used as test data and the other for training data.

The demographic and engagement data is put into a linear regression model. The output is the prediction of a pupil performance.

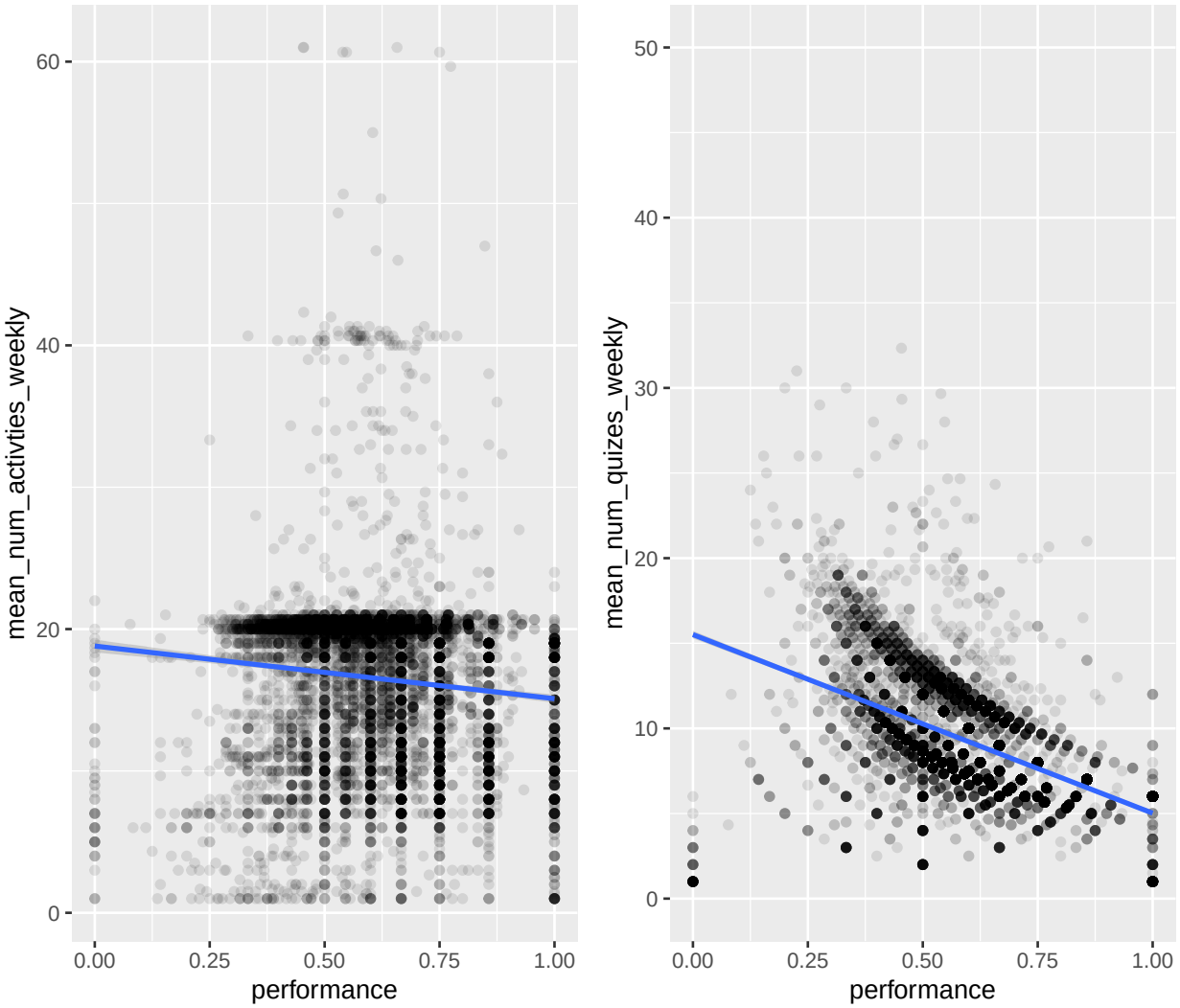

Figure 10: Scatter graphs illustrating the realationships between the performance and the average number of quizes completed and the realationship between the performance and the average number of activities completed weekly

2.6.1.1 Preparing our data The code below demonstrates the coding steps that were taken to prepare our data for the model.

```
# Select rows we need
total_enrollements_model_data = total_enrollments[-c(2:6,13,8)]
total_enrollements_model_data = total_enrollements_model_data %>%
  filter_all(all_vars(.!="Unknown")) # remove all rows that contain "unknown"

# Remove NA values
total_enrollements_model_data =
  na.omit(performanceVsDf(total_enrollements_model_data))

# Calculates average quizzes performed by an individual
avg_quizes_weekly = avgNumberOfQuizCompletedWeekly(total_quizes)
avg_quizes_weekly = data.frame(row.names =
                                 avg_quizes_weekly$learner_id, vals =
                                 avg_quizes_weekly$mean_num_quizes_weekly)
total_enrollements_model_data$mean_num_quizes_weekly =
  avg_quizes_weekly[total_enrollements_model_data$learner_id,]

# Calculates average activities performed by an individual
avg_activties_weekly = avgNumberOfActivtiesCompletedWeekly(total_activties)
avg_activties_weekly = data.frame(row.names
                                  = avg_activties_weekly$learner_id, vals
                                  = avg_activties_weekly$mean_num_activties_weekly)
total_enrollements_model_data$mean_num_activties_weekly =
  avg_activties_weekly[total_enrollements_model_data$learner_id,]

# Removing learner id column
total_enrollements_model_data = total_enrollements_model_data[-1]

# Translate columns into numbers
performance = total_enrollements_model_data$performance
avg_quizes_weekly = total_enrollements_model_data$mean_num_quizes_weekly
avg_activties_weekly = total_enrollements_model_data$mean_num_activties_weekly
total_enrollements_model_data =
  apply(total_enrollements_model_data[-c(7:9)], 2, asDoubleFactor)

# Combine dataframe again
total_enrollements_model_data = data.frame(total_enrollements_model_data,
                                           avg_quizes_weekly,
                                           avg_activties_weekly, performance)

# Our finial defined dataframe used for the model
total_enrollements_model_data = na.omit(total_enrollements_model_data)
```

2.6.1.2 Model This illustrates the model coefficients including the train and test error.

```
## [1] "The coefficients of the model"
```

```
##                              Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept)              0.6765835738 0.0347316152  19.4803371 1.735459e-70
## gender                  -0.0015965316 0.0087297626  -0.1828837 8.549316e-01
## age_range                0.0072371936 0.0025583080   2.8288984 4.777634e-03
## highest_education_level  0.0023834214 0.0027083278   0.8800343 3.790824e-01
## employment_status        0.0058465793 0.0022726461   2.5725868 1.025799e-02
## employment_area         -0.0017601316 0.0007667036  -2.2957133 2.192798e-02
## region                   0.0215858825 0.0042068916   5.1310765 3.549714e-07
## avg_quizes_weekly       -0.0227612559 0.0013258477 -17.1673228 3.731376e-57
## avg_activties_weekly     0.0006901234 0.0007445473   0.9269033 3.542323e-01
```

Error	Results
Test Error	0.0142717
Train Error	0.0165305

The test error and train error are small. This shows the model produces an accurate estimation of the performance of a pupil. This is further shown on figure 10, as all our residuals (the error measured by taking away the actual results from the expected result in the test data) is small and relatively close to the fitted line. This model can be used to provide more insight on how a pupil is performing, and identify any students not gaining the correct understanding. This information can then be used to decide on the suitable intervention for a pupil which will further help enhance the Newcastle online learning course.

2.6.2 Model 2: Time series model to forcast future performance.

The idea of this model is a predictive analytic model that makes use of the data of the average performance of each a specific demographic from each run of the module. The example on figure 12, is predicting average performance the next run based on different regions. This predictive model can be also used for other demographics (age, gender etc).

2.6.2.1 Preparing our data The model uses two data set: One containing the student performance and one containing the demographic performance that we want to predict. As an example we focus on regions of pupils from: Africa, Oceania, Americas Asia and Europe. Therefore 5 models are built to forecast their future performance. For each model, the data is:

1. Filtered to select only pupils from a specific region
2. Maps the pupils to their performance on the second data set.

The predictive model can be used to find out more on other demographics (age, gender etc)

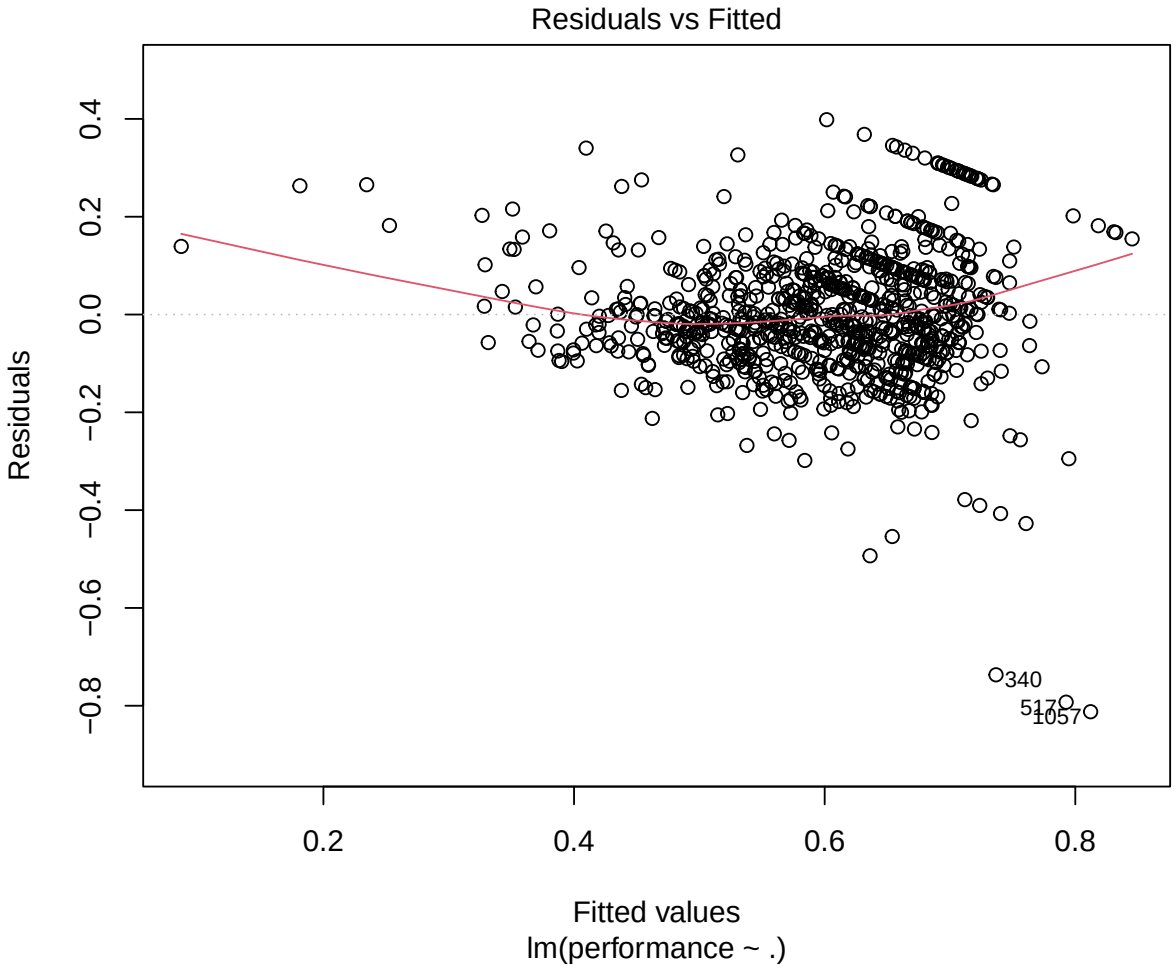

Figure 11: Model 1: illustrating where the residuals lies on the fitted line.

2.6.2.1 Preparing our data Figure 11, shows the results of each model forecasting the results of pupils based on region. Each model can only predict 1 future run of the module. This is because there is only 7 observation for all runs. Therefore to predict more, more data (more runs of the course) would need to done. There is a low signal-to-noise ratio, therefore it was difficult for the predictive model to extract any possible trend, persistence, lagged errors etc.

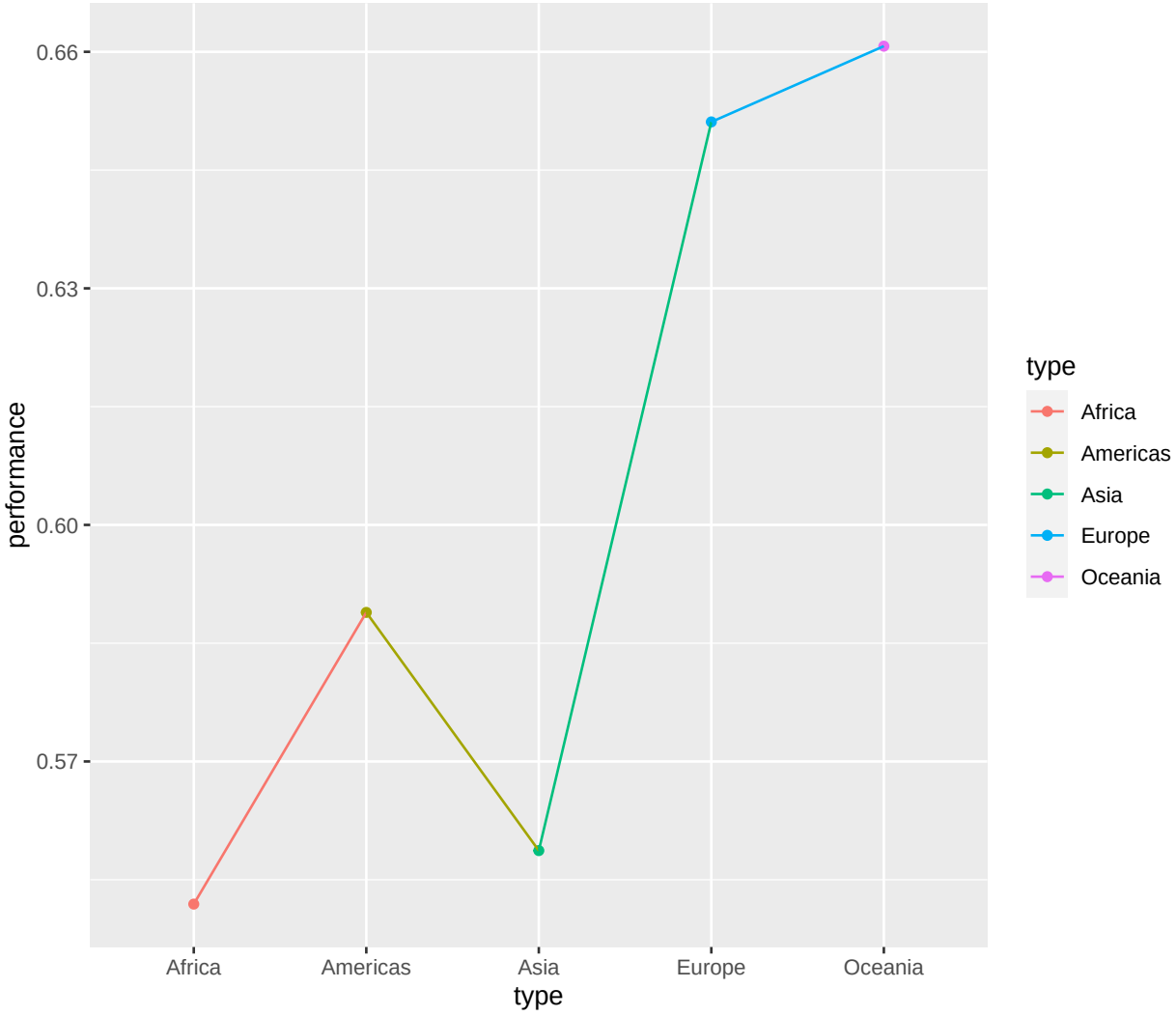

Figure 12: Model 2: Forcasting performance

The purpose of this model, is to allow Newcastle university identify pupils which may have a low performance in the future. This allows enough time to provide any suitable intervention to prevent the low performance. This predictive model could have performed better with more data added. More data added means the model can predict forecast up 2 more years and provide better accuracy.

3. Evaluation

The business purpose was looking at ways to enhance Newcastle online content. This included looking specifically at the online course *Cyber Security: Safety at Home, Online, in life*. This was done by monitoring the interactions and performance between different demographics. This was done with the intent to identify which demographics may face when doing online courses and limitations, ensuring all pupils perform to their best.

The report identifies the performance and interactions between different demographics which are illustrated using graphs. These graphs identified the demographics which perform the best which are usually pupils from countries Oceania and Europe, over the age of 65 and retired. The graphs identified the demographics which perform the worst: This includes full time students, aged between 18-25 and countries from Africa and Asia. Additionally, it includes 2 models. 1 used to predict a pupils performance and the 2nd used forecast pupils performance. This model are used to predict future models. There made with the intent of finding/forecasting which students have low performance and providing suitable interventions before the pupils.

After review an additional step which could have been added. This is the monitoring pupils that did not not complete any quiz and/or complete any activities. The 2 data set that shows the number of activities and quizzes complete by each pupil only contained the learner_id of a pupils who completed at least 1 activity and/or quiz. Therefore, those who completed none quizzes or activities where not included within the data set. An additional analysis is to:

- monitor those who did not complete quizzes or activities by demographics. This is done by finding the pupils in data sets enrollments but not within quizzes or activities data set. This will illustrate:
  - number of students who font complete quizzes and/or activities.
  - illustrate the types of demographics of pupils who don't complete quizzes and/or activities
- Validate the performance of those who did not complete any activities by different demographics. This is done by finding the pupils in data set *enrollments* and not in *activity* data set then measuring the performance using the *quizzes* data set. This data could have been further added towards the predictive models. Furthermore, this would answer questions such as:
  - Do pupils who taken no activity perform better?
  - How do different demographics perform within the pupils who don't take activities perform?

Additional problems with this investigation is the way performance is measured. This is done by the ratio of questions a pupil obtained correctly on quizzes. These quizzes are usually for practice and understanding of the content, not a way of assessment. It was identified that pupils tried multiple combination of answers before getting a question before the got in right. This perhaps indicates the is guessing (therefore not gained the correct understanding), or the pupil is using the quiz to learn. Overall, the number of correct quizzes is not suitable to measure performance as are uses for practice, not assessment. Therefore, this may measured way of "performance" could

contain bias. A further helpful method is to introduce an online assessment for the overall course. Setting a borderline would determine whether a student has passed or not passed.

Another step to explore is identifying reasons of performance for a particular student, and looking at methods to retain pupils between different demographics. A trial could be collected measure different structure methods applied to the run of the online content. For example introducing more videos, more engagement with the organisation admin, more content, more weeks less content. More weeks with less content and verifying which had more engagement and performance across different demographics. This would allow further identification what type of elements affect different pupils are more suitable to different pupils.

3.1 Deployment