Learning analysis: Exploring the interactions and demographics between different demographics

Adanna V. Obibuaku (180251870)

21/11/2021

1. Introduction

1.1 Background

This report follows the CRISP-DM methodology to gain an insight into the online courses provided by Newcastle University, specifically the Cyber Security: Safety at Home, Online, in life online course hosted by the FutureLearn website. The main business goal is to enhance Newcastle University's online resources. This is done by monitoring the performance and engagement of certain users. Cyber Security: Safety at Home, Online, in life has 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 will provide an insight into pupils engagement and performance and provide appropriate intervention in limited areas. 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 at the demographics of pupils learning the course, their interactions, engagement and performance Additionally, it explores the reasoning behind the performance of each group; trying to identify constraints unique to a specific demographics may have for engagement 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 usage of data which gives an insight about student engagements (number of weekly activities/quizzes etc), to predict the performance of a user.
2. The second model makes use of data that provides the average performance of specific demographics across different runs of the module. This will be used to predict the expected performance of specific demographics for the next run of the module

1.2.1 The Data

Furthermore, the needed data set are specific to 7 runs of the online course. These data sets are combined into 1 data set. For example there are 7 enrollment data sets for each run of the online course. This is combined into 1 data set. This provides a total summary of the course data. Below, illustrates how the enrollment data set collect for each run is combined.

The initial 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 subsets of data from the data set including:

1. The enrollment of each run of the course. This includes fields such as age, gender, country etc. This will help when exploring the performance between different demographics.
2. The step activity for each run of the course. This includes fields specifying the number of activities completed by a pupil. This will be used to measure student engagement.
3. The quizzes for each run of the course. This includes fields specifying the number of quizzes by the user. This will also be used to measure student engagement. Additionally, this data will be used to measure the performance. This is provided with the ratio of answers the pupils got correct.

An additional data set. This data set is used to 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 "detected_country" is used to map the country of each pupil to a continent.

Furthermore, the needed data set are specific to 7 runs of the online course. These data sets are combined into 1 data set. For example, there are 7 enrollment data sets for each run of the online course. This is combined into 1 data set. This provides a total summary of the course data. Below, illustrates how the enrollment data set collected for each run is combined.

```
total_enrollments = translateCodeToCnt(distinct(rbind(cyber.security.1_enrolments, cyber.security.2_enr
                          cyber.security.3_enrolments, cyber.security.4_enrolments,
                          cyber.security.5_enrolments, cyber.security.6_enrolments,
                          cyber.security.7_enrolments), learner_id, .keep_all = TRUE))
```

2. Methodology

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

2.1 Verify data quality

The data sets that are going to be used to analyse our question: *How do different demographics interact and perform with the online course?*, have some empty fields or labelled as "unknown".

For example in the *enrollment* data set collected 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 were no purchases.
- A column describing the type of user (gender, age) rows were labelled "Unknown". This suggests that the pupil did not answer it. These fields were removed from the data set.

2.2 Monitoring pupils attendance between different demographics

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

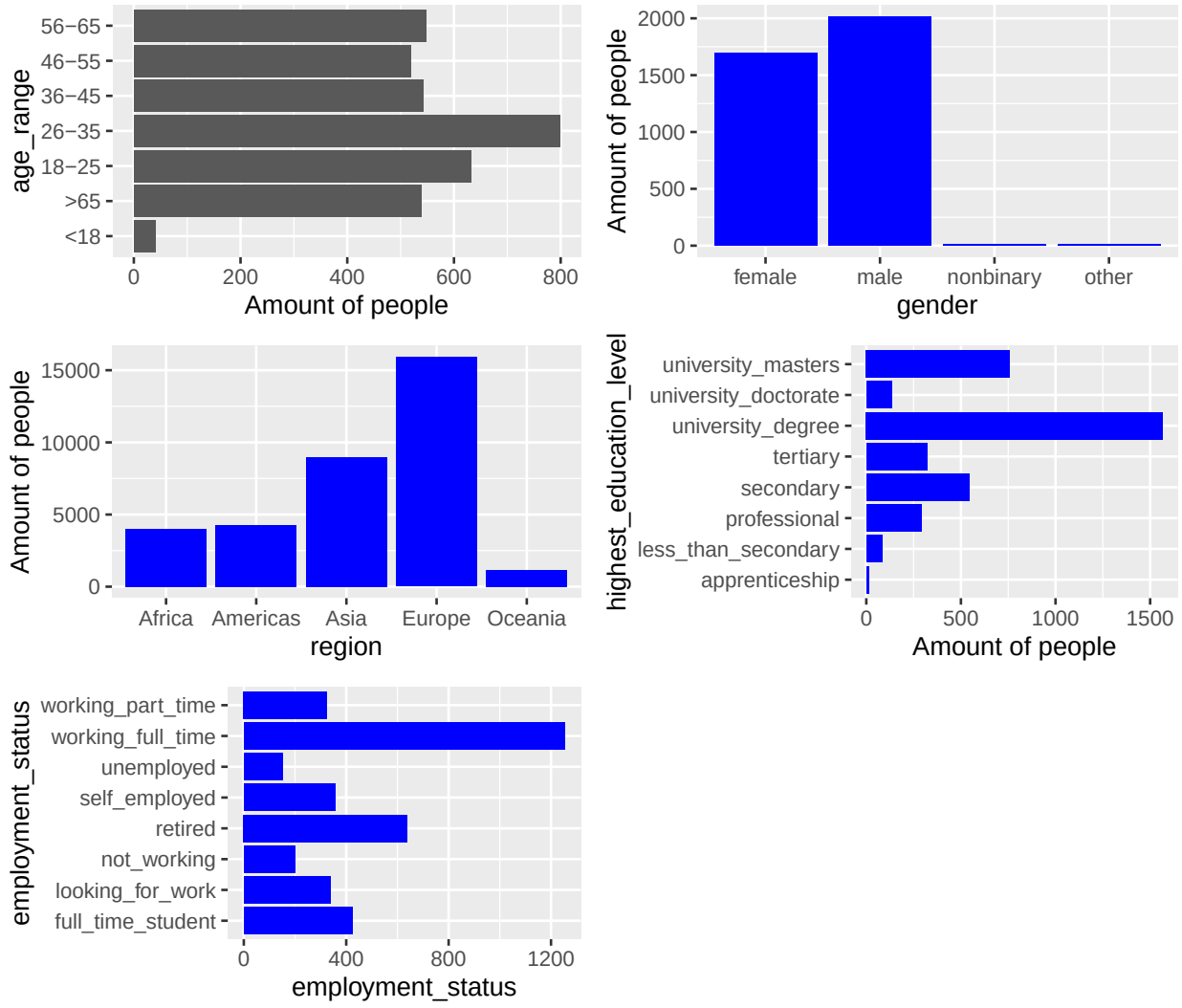

Figure 1: Bar graph illustrating the number of enrollments between different demographics

This graph is a summary of the total 7 different runs of the course. The data illustrated in figure 1, shows 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 than female pupils. There are 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. The lowest, being pupils from Oceania. Most pupils enrolled in this course, have at least a university degree and/or work full time.

Overall, the data illustrates that a significant amount of people who choose to enrol in this course appears pupils from Europe between the ages 26-35, working full time and have at least a university degree.

2.3 Monitoring pupils performance between different demographics

2.3.1 Illustration of the performance across each run of the module

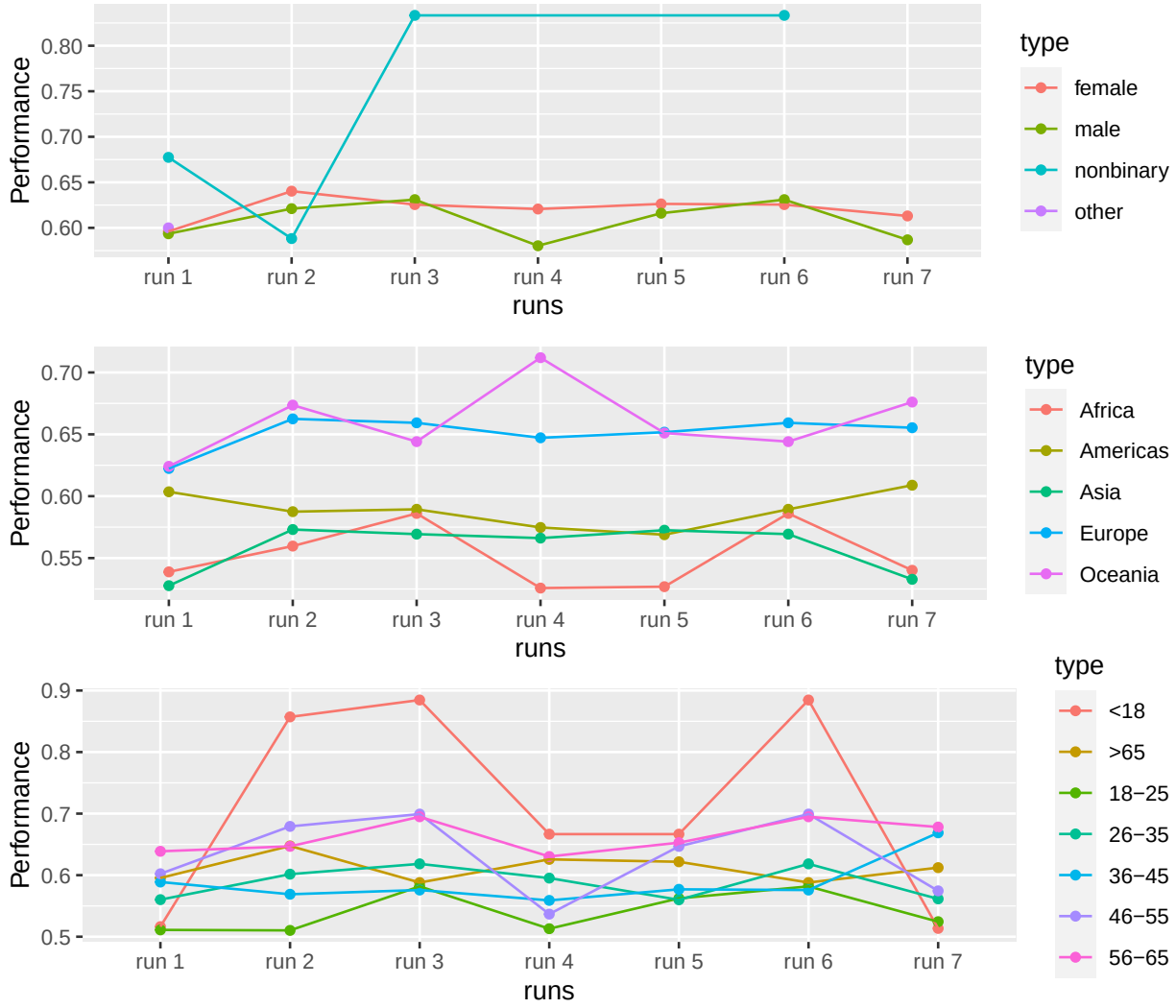

Figure 2: The average performance across different runs of the online course between different demograpics

In Figure 2, the data shows that female 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 Europe and Oceania performed better than pupils from countries Asia, Africa and the Americas. European pupils appear to perform consistently the same across the 7 runs of an online course, while in run 4 Oceania significantly performed better than all the other regions.

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.

Figure 2, shows there is no significant distinction in who performs higher between pupils with different

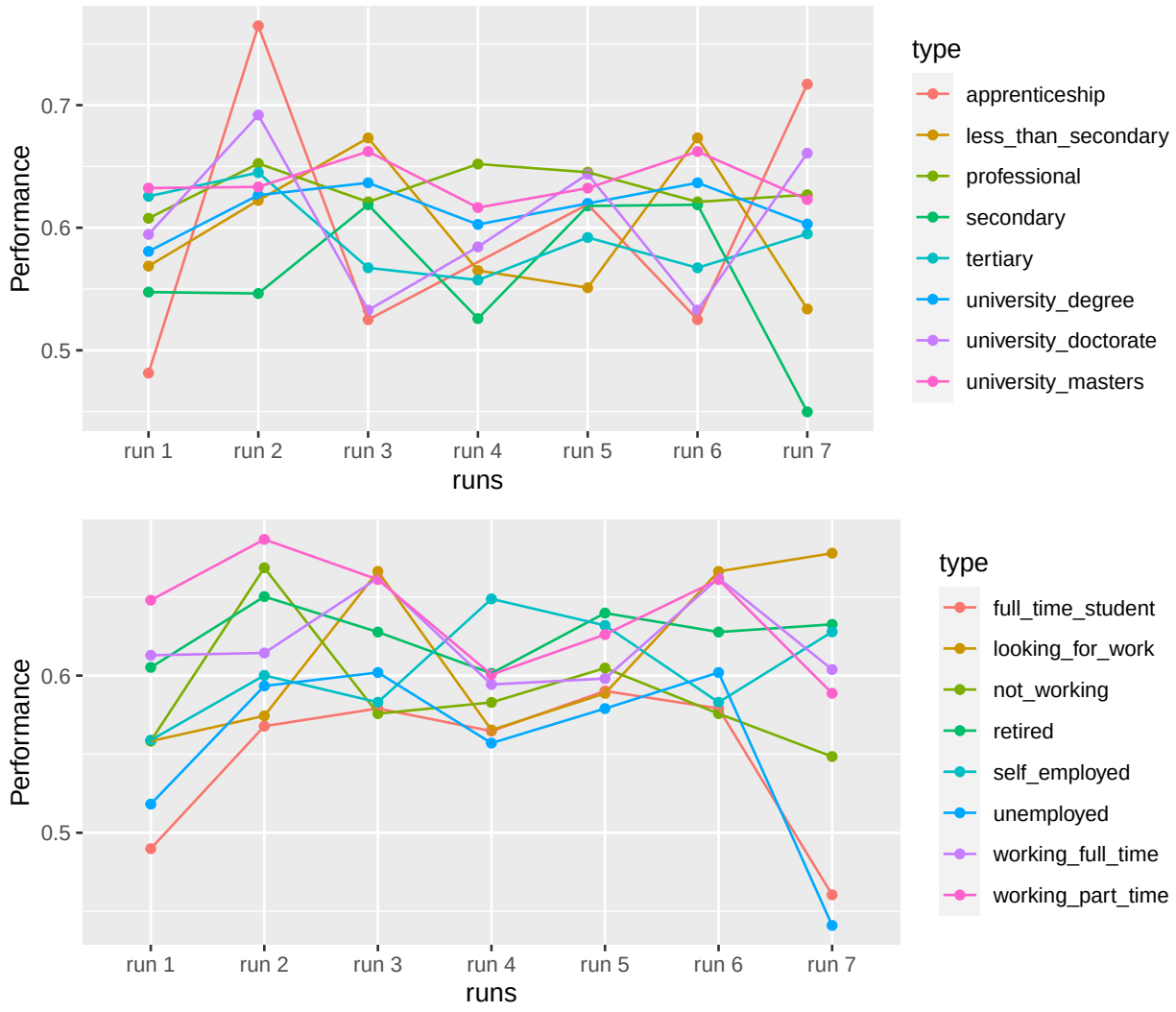

Figure 3: The average performance across different runs of the module compared between different demograpics

levels of the highest education. However, those with the highest education level of a university degree appear to perform consistently through each run.

Additionally, figure 2, illustrates that across each run, there does not appear a significant distinction in performance between pupils with different employment statuses. However, pupils who are retired perform at a consistent rate through each run of the course.

2.3.2 Illustration from the total run

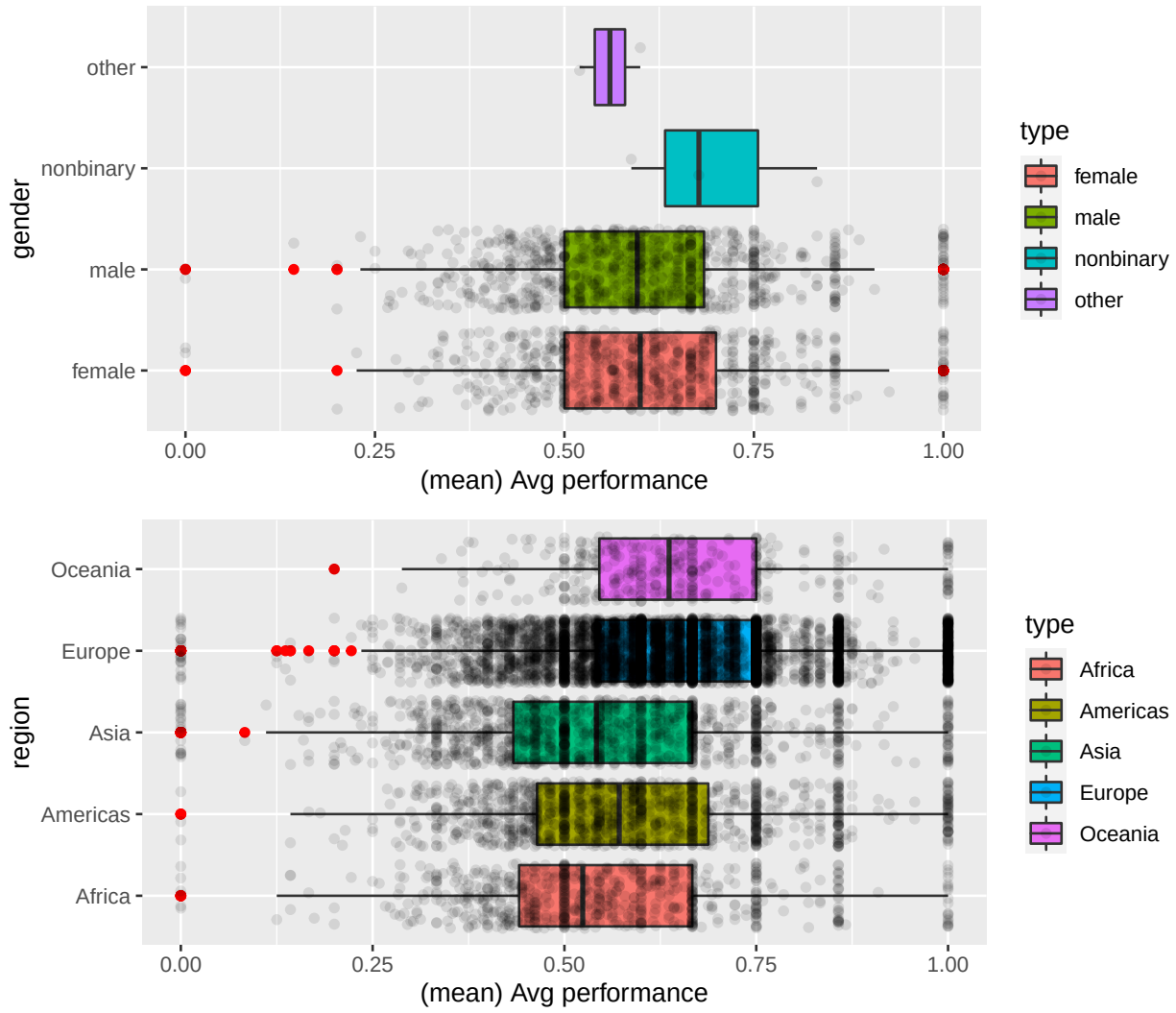

Figure 4: Boxplots illustrating the distribution of the average performance between different demographics

Figure 4, suggest there is more variance in the average performance achieved by male students than the variance for female students. Furthermore, The average performance of male and female students are approximately the same.

Pupils from Europe have a higher average performance than pupils from other regions. Additionally, the data suggest that pupils the average performance from pupils in Europe are mostly distributed at the higher end of the performance.

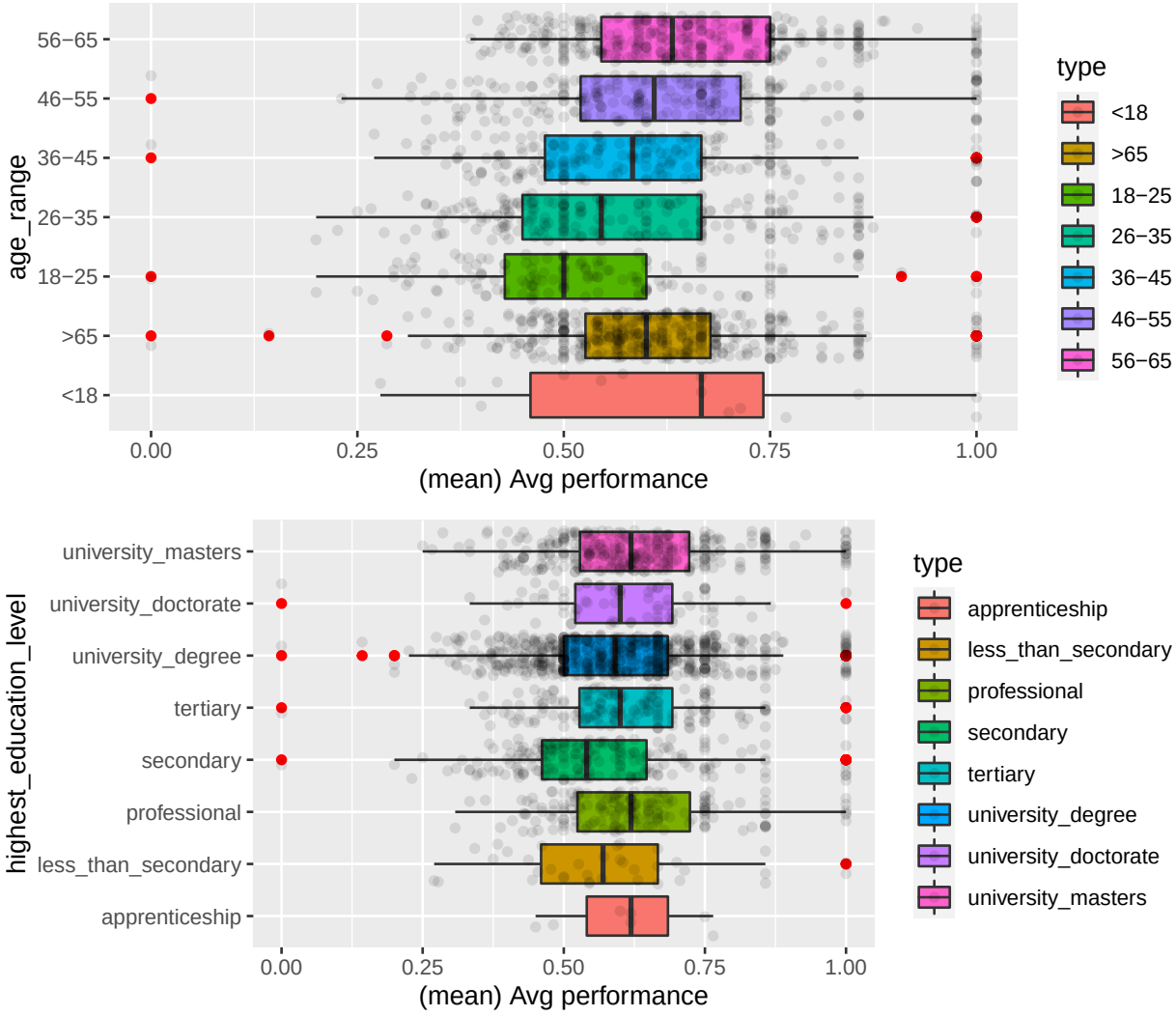

Figure 5: Boxplots illustrating the distribution of the average performance between different demographics

Figure 5, suggest that <18 perform have a higher average performance. However, there are not enough pupils that are less than 18 to precisely define them. Therefore, this information is disregarded. The data shows that pupils who are more than 65 have a higher performance. The second-highest performance is those who are between ages (56-65). Furthermore, pupils who are more than 65 have a low Interquartile range (IQR). This shows that the scores between each individual are similar/consistent. Pupils who are ages between 18 and 25 have a lower performance average. The graph illustrates that the distribution of the average performance of pupils aged 18-25 is situated at the lower end.

Additionally, figure 5 shows 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 too small to accurately represent the population, the data is disregarded. Pupils with a university have the highest average performance. Pupils who have obtained secondary schooling at the highest level of education have a low average performance.

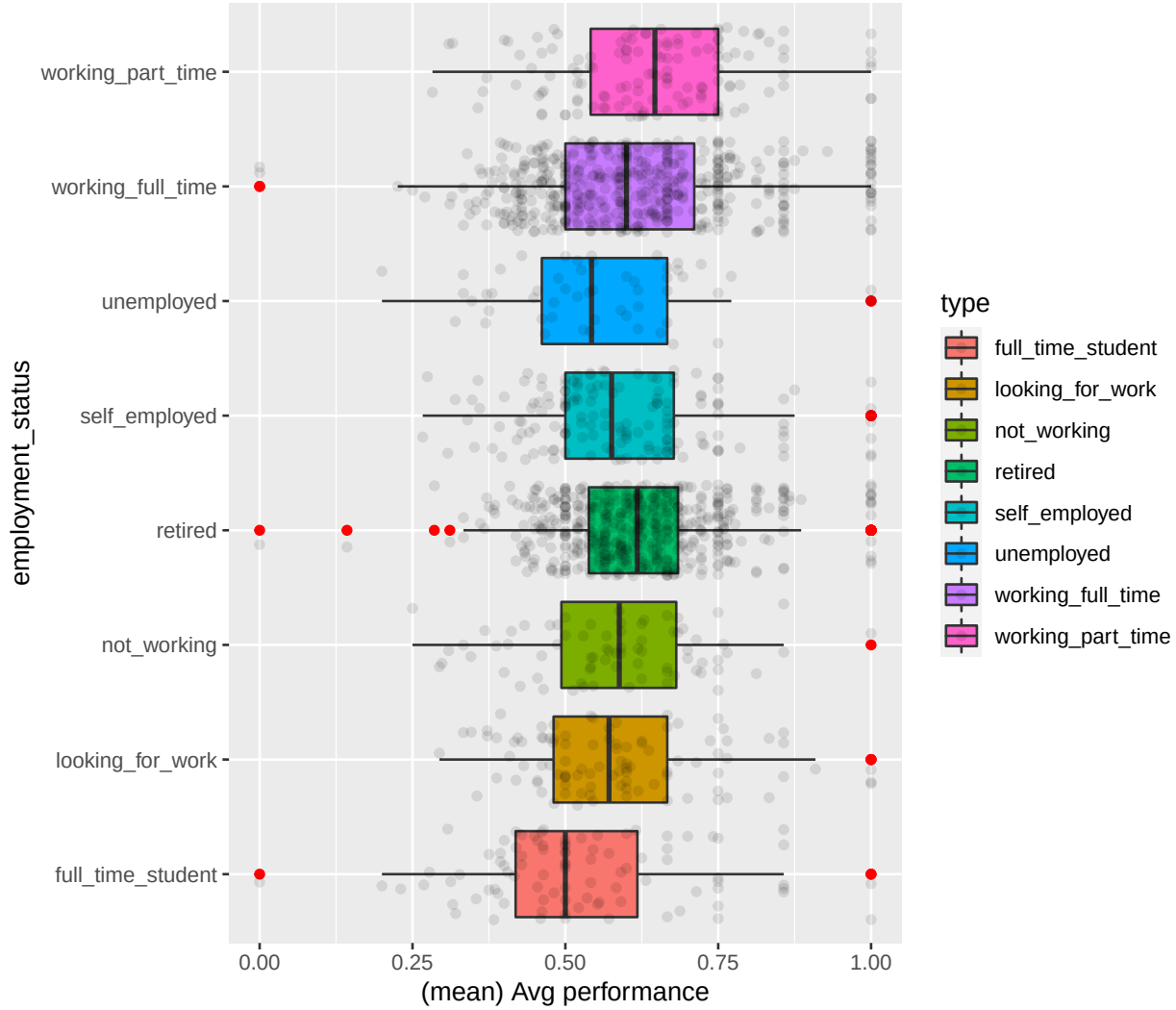

Figure 6: Boxplots illustrating the distribution of the average performance between different demographics

Figure 6, shows that pupils working full time have the highest average performance. This is followed by pupils who have retired. The distribution of people who have retired has a small IQR. This suggests that the performance obtained by each individual is fairly similar. Pupils who are full-time students have a low average performance.

2.4 Monitoring pupils engagement between different demographics

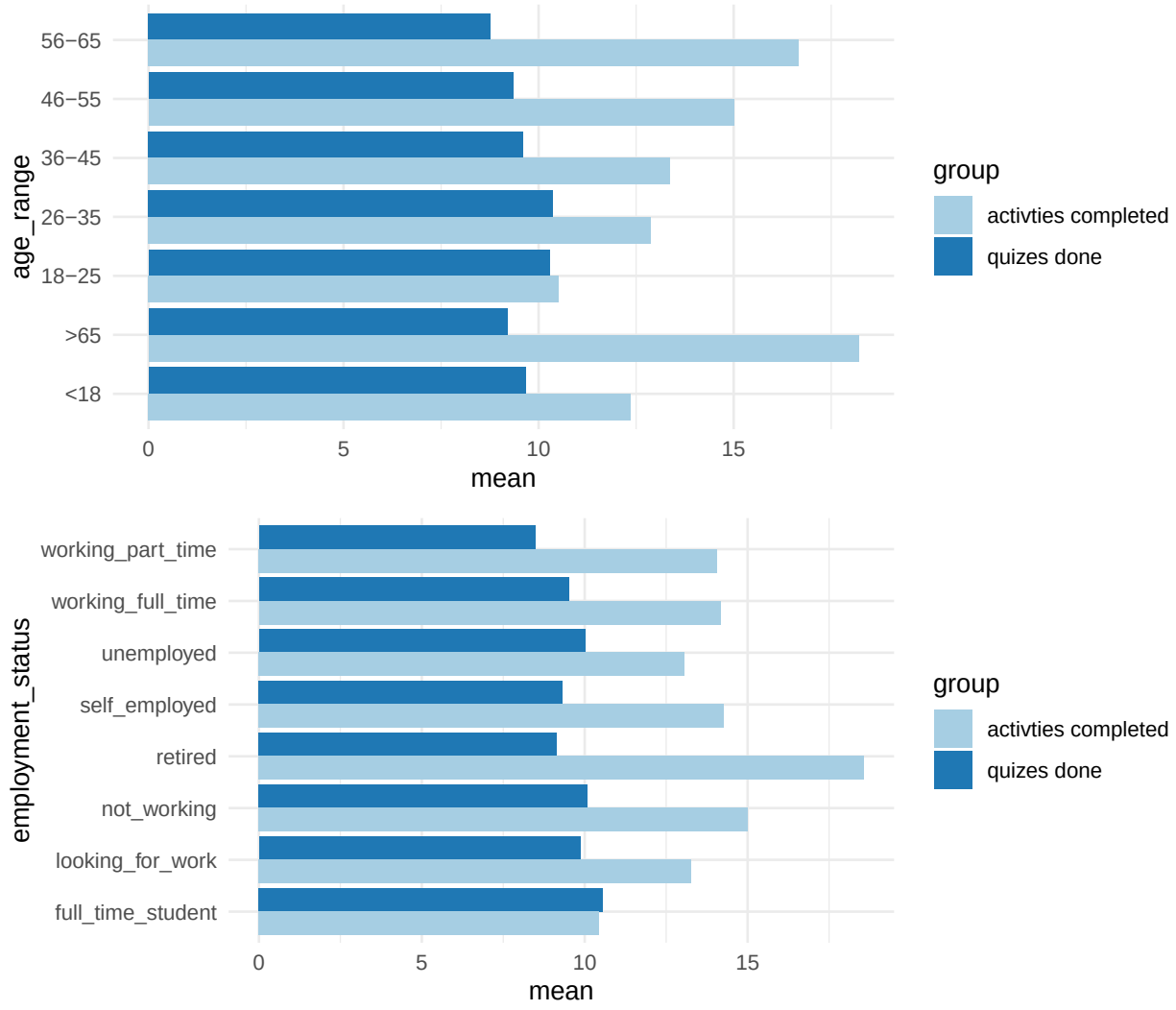

Figure 7: Bar graphs illustrating student engagement measured the mean number of quizzes and number of activities weekly completed.

In figure 7, 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. The second-lowest are 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.

Furthermore, it is shown that those who have retired have a higher number of activities completed on average. The second highest is those who are not working. Those who are retired completely the lowest amount of quizzes per week, while full-time students complete the highest number of quizzes per week.

Figure 8, shows pupils who are non-binary or identify as other 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. Female pupils tend to complete more activities in a week than male pupils. The data further illustrates that female pupil complete fewer quizzes in a week than male pupils.

Pupils who have an apprenticeship tend to complete more activities and fewer quizzes than all the different demographics of higher education. However, given that the sample of the population collected for pupils in

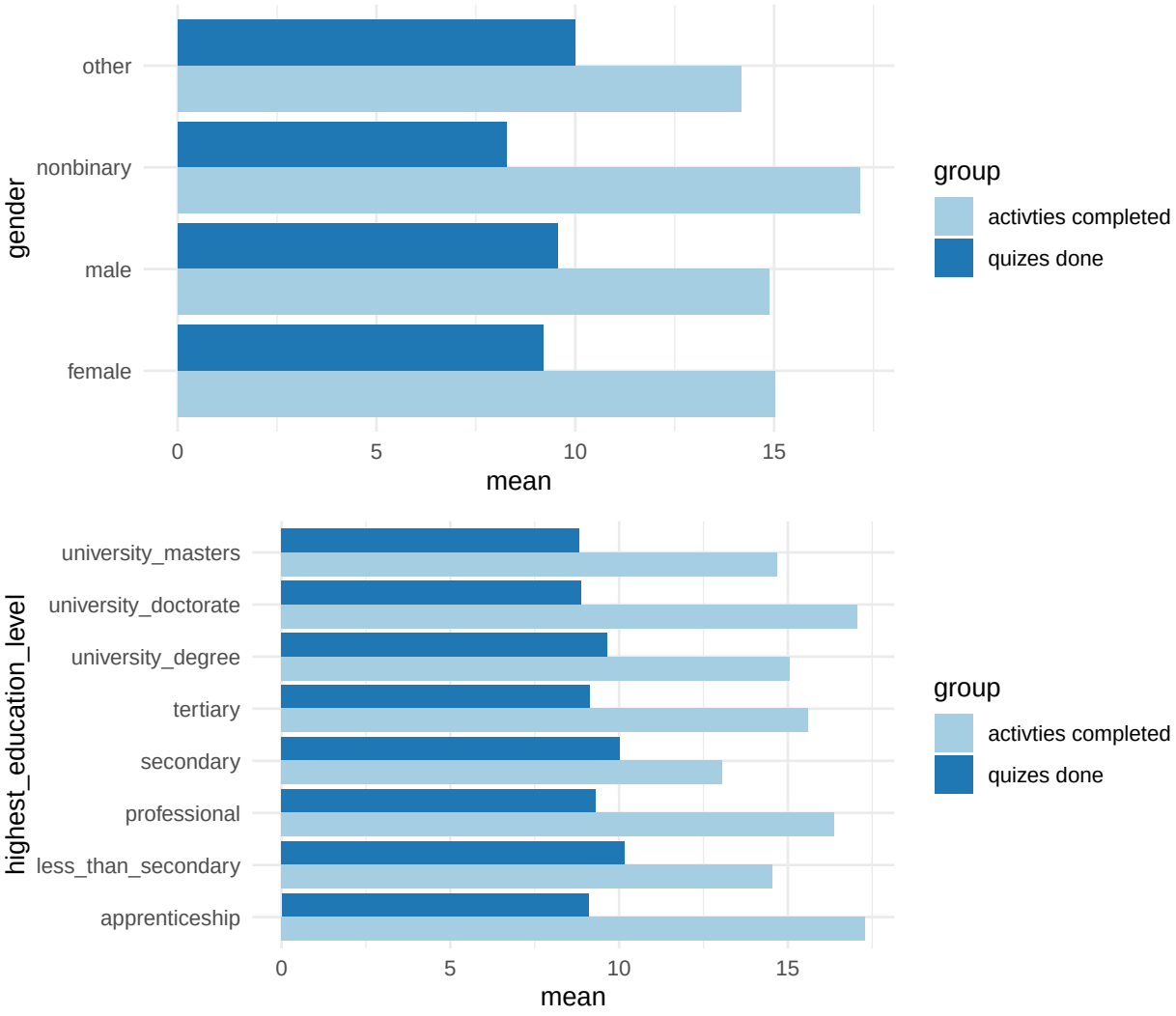

Figure 8: Bar graphs illustrating student engagement measured the mean number of quizzes and number of activities weekly completed.

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

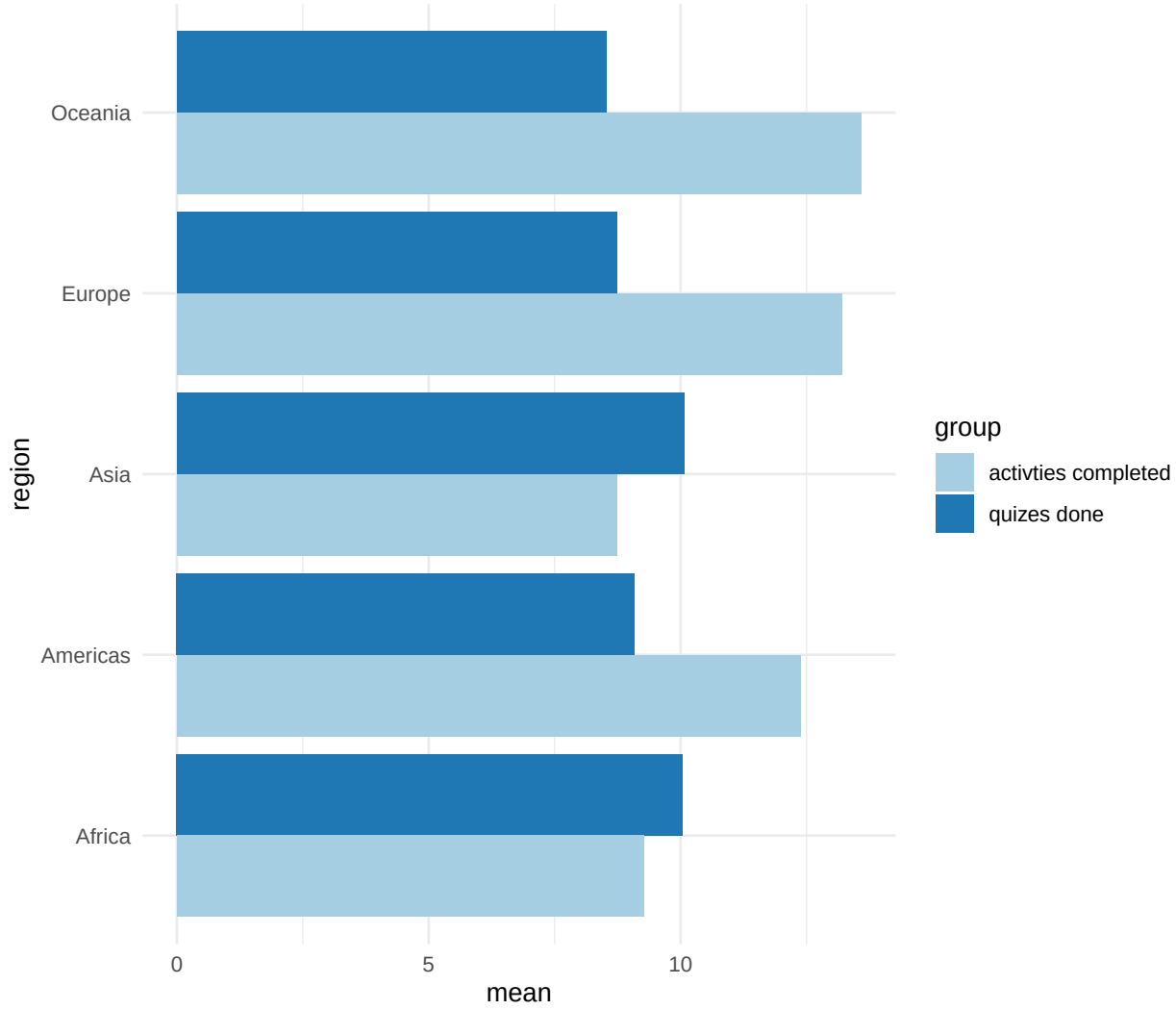

Figure 9: Bar graphs illustrating student engagement measured the mean number of quizzes and number of activities weekly completed.

Figure 9 illustrates that 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. However, pupils with the highest number of weekly quizzes completed are people from countries in Africa. Pupils with the lowest number of quizzes completed are from countries in Europe.

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 degrees perform better

Pupils from countries in Oceania Generally, it was observed that a pupil who completes more quizzes per week have a lower performance. This is because pupils with multiple attempts of quizzes did so with different answer combinations until the answer was correct. This suggests that they may have the right understanding and might be guessing. Additionally, figure 10, shows that the relationships between the number of weekly **quizzes** completed and performance has more of a negative relationship than the relationship between the performance and the number of weekly **activities**. This indicates the reason why some demographics (>65) have a higher-performing (as they took much fewer attempts to answer questions) and more weekly activities completed. Activities take more effort than quizzes, as involves self-research to find an answer. That indicates those who are engaged take more activities, thus gaining a better understanding and higher performance in quizzes (fewer reattempts).

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

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

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 takes in two main data sets - demographic data, including 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 pupils per 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 pupil performance.

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 =
```

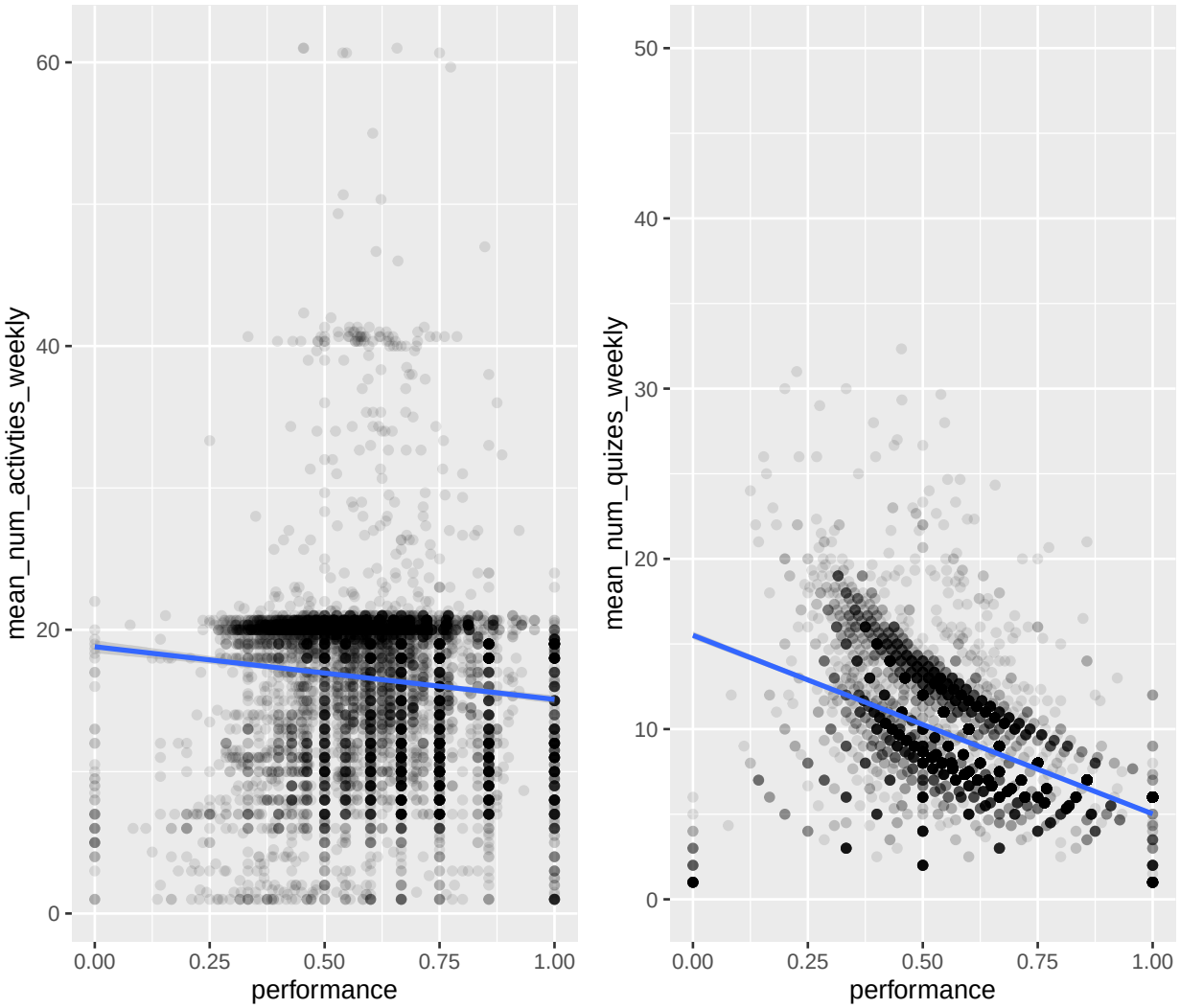

Figure 10: Scatter graphs illustrating the relationships between the performance and the average number of quizzes completed and the relationship between the performance and the average number of activities completed weekly

```
  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 Below illustrates the model coefficients for each exploratory variable 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

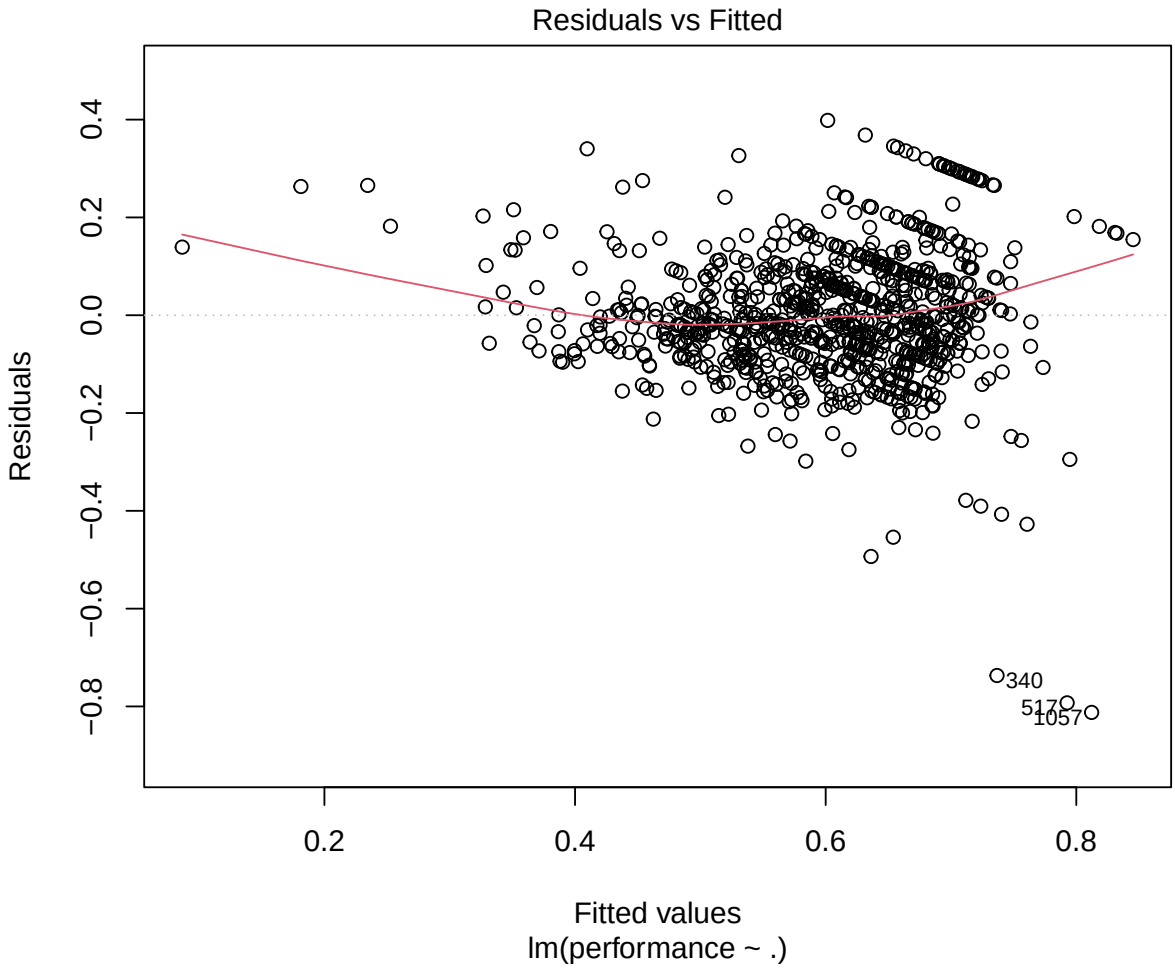

Figure 11: Model 1: illustrating where the residuals lie on the fitted line.

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 in figure 11, as all our residuals (the error measured by taking away the actual results from the expected result) 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 forecast future performance.

The idea of this model is a predictive analytic model that makes use of the data of the average performance of each specific demographic from each run of the module. The example in figure 12, is predicting the

average performance of the next run based on different regions. This predictive model can be also used to predict the performance of 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 used to forecast the average future performance for each region. For each model, the data prepared by

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

2.6.2.2 Model Figure 11, shows the results of each model. This provides the future prediction (run 8) of the average performance. Each model can only forecast 1 average run. This is because there is only 7 observation (obtained within the 7 runs). 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. Therefore to predict more, more data (more runs of the course) would need to be done.

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

3. Evaluation

The business purpose was to look 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, with the intent to identify limitations unique to specific demographics. 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 is used to predict future models. They're made with the intent of finding/forecasting which students have a low performance to allow enough time for Newcastle University to provide suitable intervention.

After review, an additional step that could have been added is monitoring pupils that did not complete any quiz and/or complete any activities. The 2 data set that shows the number of activities and quizzes completed by each pupil only contained the learner_id of pupils who completed at least 1 activity and/or quiz. Therefore, those who completed no quizzes or activities were not included within the data set. Additional analysis is to:

1. 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:

- The percentage of students who take part in complete quizzes and/or activities.
- The percentage of different demographics of pupils who don't complete quizzes and/or activities

2. Validate the performance of those who did not complete any activities by different demographics. This is done by finding the pupils in the *enrollments* data set and do not appear in the *activity* data set then measuring the performance using the *quizzes* data set. This data could then be added to the predictive models. Furthermore, this would answer questions such as:

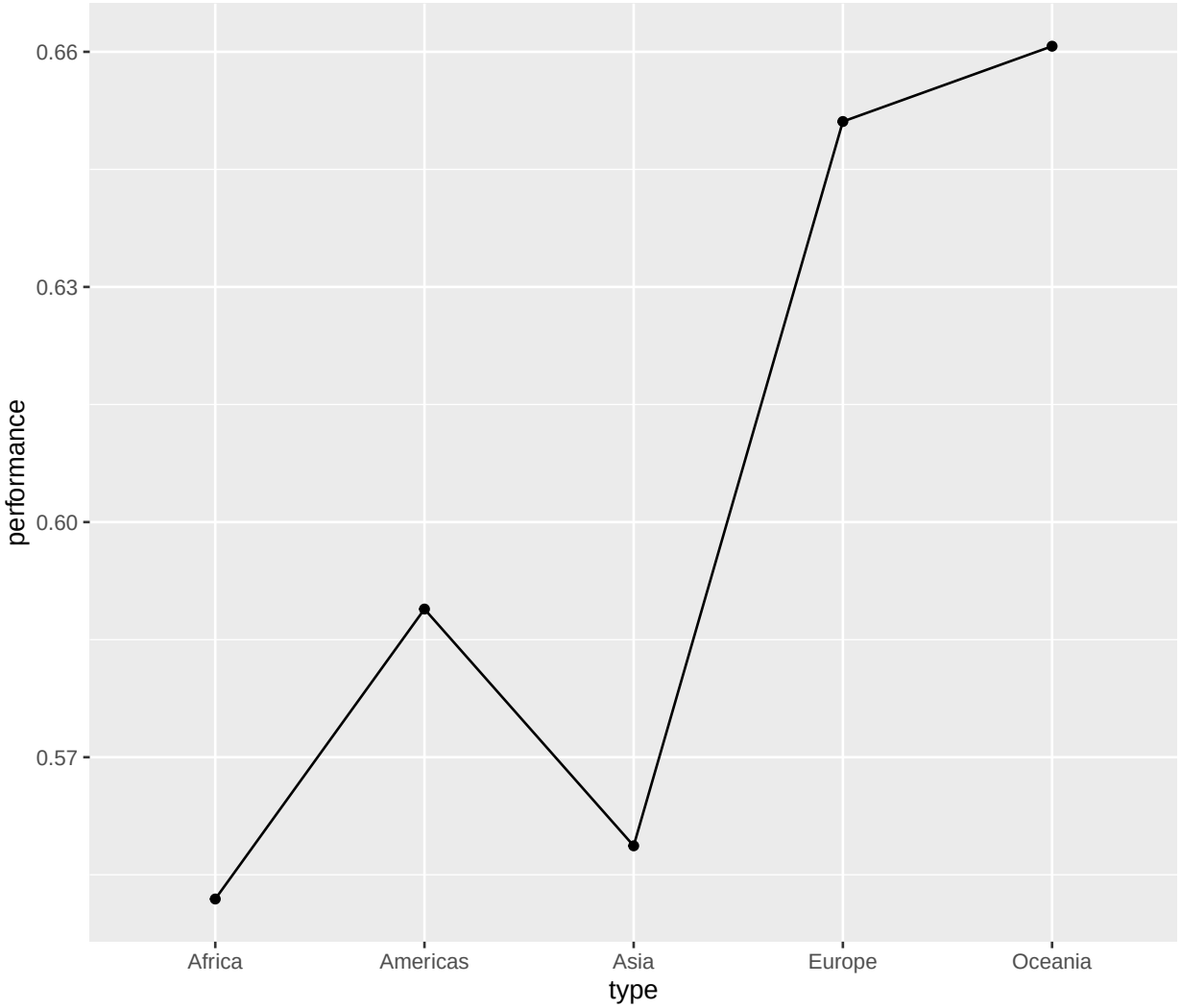

Figure 12: Model 2: Forecasting performance

- Do pupils who take no activity perform better?
- How do different demographics perform within the pupils who don't take activities?

An additional problem 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 for practice and understanding of the content, not a way of assessment. It was identified that pupils tried multiple combinations of answers before getting a question correct. 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 a suitable measurement of performance. This is because quizzes is a way of practice and not assessments. Therefore, the measured "performance" may contain bias. A further helpful method is to introduce an online assessment for the course. Setting a borderline would determine whether a student has passed or not passed.

Another step is to identify reasons for performance for a particular student and look at methods to retain pupils between different demographics. More trial runs could be done using different structures/methods of learning. For example introducing more videos, more engagement with the organisation admin, more content, more or less weekly content. From the data, verify which had more engagement and performance across different demographics. This would allow further identification of the constraints that affect different pupils including the suitable methods for different pupils.

3.1 Deployment

This project is deployment is achieved using projectTemplate, as much of the code is reproducible. The functions within the helper file are used to make to apply the same data transformation which is used within the analysis.