CSDA1050 Advanced Analytics Capstone Course

Project Sprint 2

Improving student's graduation in Education

By Sylvain Kamto

Student ID: 11060

1- Introduction/ Background

Currently less than 65% of the students complete their studies as planned. Part of the students will move to work without graduation or change the branch of studies to another institute, but too many have either delayed in their studies (12.3%) or will completely discontinue (8.5%)

The delayed and dropout students pose significant direct costs to cities and schools due to reduced funding from government. Dropouts especially have challenges in finding a job and this problem is causing serious impacts on society in the long run.

To alleviate this problem, we are here by initiating a concept project on how to apply analytics to improve graduation in schools. The core of the idea is the following: utilize advanced analytics and machine learning to identify students who have elevated risk to dropout or delay in studies, so that interventions and support actions can be initiated early enough.

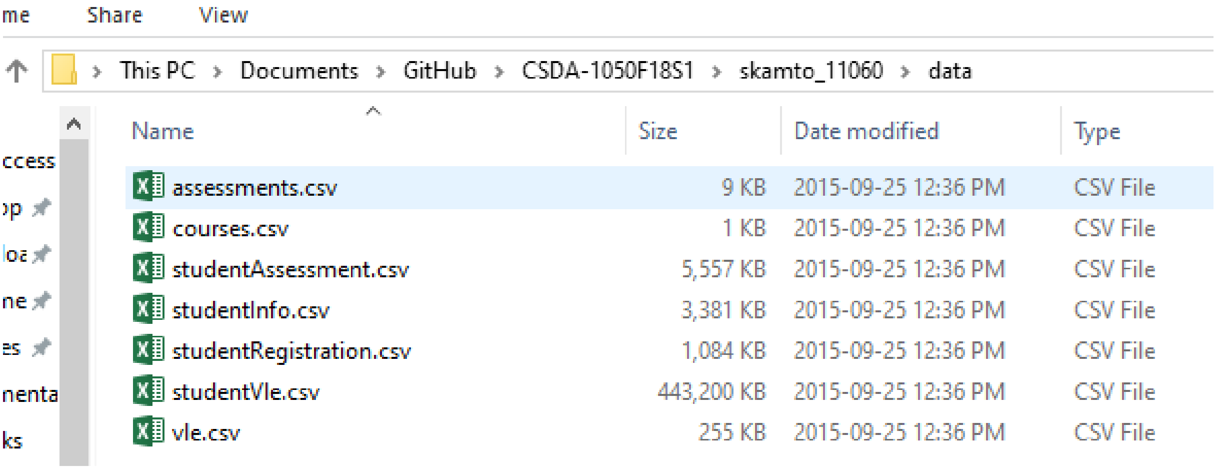
2- Research Question

2.1. Predicts which students have elevated risk of delayed studies or even dropping out

2.2. Predict student academic outcomes to better guidance and support

3- Dataset

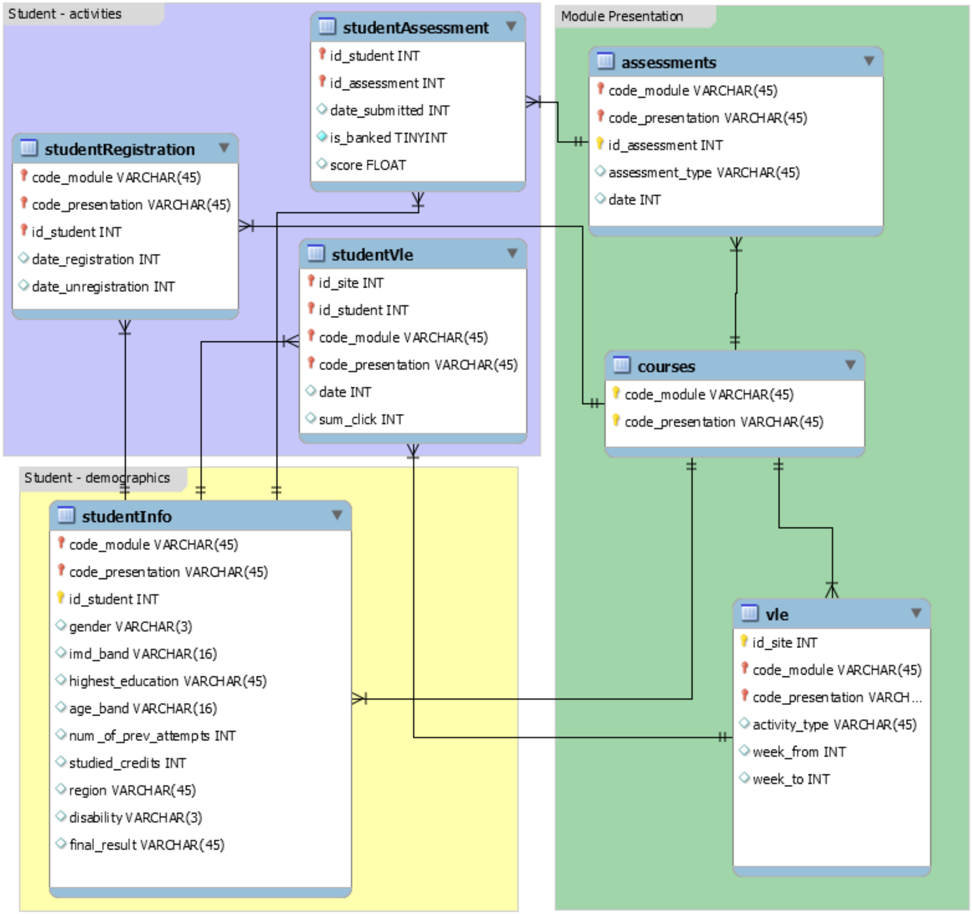
Data were collected from the anonymised Open University Learning Analytics Dataset (OULAD). It contains data about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules). Presentations of courses start in February and October - they are marked by “B” and “J” respectively. The dataset consists of tables connected using unique identifiers. All tables are stored in the csv format.


me
Share
View
This PC › Documents › GitHub › CSDA-1050F18S1 › skamto_11060 › data
ccess
op
loa
ne
es
nenta
ks

Name	Size	Date modified	Type
assessments.csv	9 KB	2015-09-25 12:36 PM	CSV File
courses.csv	1 KB	2015-09-25 12:36 PM	CSV File
studentAssessment.csv	5,557 KB	2015-09-25 12:36 PM	CSV File
studentInfo.csv	3,381 KB	2015-09-25 12:36 PM	CSV File
studentRegistration.csv	1,084 KB	2015-09-25 12:36 PM	CSV File
studentVle.csv	443,200 KB	2015-09-25 12:36 PM	CSV File
vle.csv	255 KB	2015-09-25 12:36 PM	CSV File

4- Dataset Description

This dataset offers two of the elements in the framework: behavior and performance. It contains information about 22 courses, 32,593 students, their assessment results, and logs of their interactions with the VLE represented by daily summaries of student clicks (10,655,280 entries).



courses.csv

File contains the list of all available modules and their presentations. The columns are:

- code_module – code name of the module, which serves as the identifier.
- code_presentation – code name of the presentation. It consists of the year and "B" for the presentation starting in February and "J" for the presentation starting in October.
- length - length of the module-presentation in days.

The structure of B and J presentations may differ and therefore it is good practice to analyse the B and J presentations separately. Nevertheless, for some presentations the corresponding previous B/J presentation do not exist and therefore the J presentation must be used to inform the B presentation or vice versa. In the dataset this is the case of CCC, EEE and GGG modules.

assessments.csv

This file contains information about assessments in module-presentations. Usually, every presentation has a number of assessments followed by the final exam. CSV contains columns:

- code_module – identification code of the module, to which the assessment belongs.
- code_presentation - identification code of the presentation, to which the assessment belongs.
- id_assessment – identification number of the assessment.
- assessment_type – type of assessment. Three types of assessments exist: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).
- date – information about the final submission date of the assessment calculated as the number of days since the start of the module-presentation. The starting date of the presentation has number 0 (zero).
- weight - weight of the assessment in %. Typically, Exams are treated separately and have the weight 100%; the sum of all other assessments is 100%.

If the information about the final exam date is missing, it is at the end of the last presentation week.

vle.csv

The csv file contains information about the available materials in the VLE. Typically these are html pages, pdf files, etc. Students have access to these materials online and their interactions with the materials are recorded. The vle.csv file contains the following columns:

- id_site – an identification number of the material.
- code_module – an identification code for module.
- code_presentation - the identification code of presentation.
- activity_type – the role associated with the module material.
- week_from – the week from which the material is planned to be used.
- week_to – week until which the material is planned to be used.

studentInfo.csv

This file contains demographic information about the students together with their results. File contains the following columns:

- code_module – an identification code for a module on which the student is registered.
- code_presentation - the identification code of the presentation during which the student is registered on the module.
- id_student – a unique identification number for the student.
- gender – the student's gender.
- region – identifies the geographic region, where the student lived while taking the module-presentation.
- highest_education – highest student education level on entry to the module presentation.
- imd_band – specifies the Index of Multiple Depravation band of the place where the student lived during the module-presentation.
- age_band – band of the student's age.
- num_of_prev_attempts – the number times the student has attempted this module.
- studied_credits – the total number of credits for the modules the student is currently studying.
- disability – indicates whether the student has declared a disability.
- final_result – student's final result in the module-presentation.

studentRegistration.csv

This file contains information about the time when the student registered for the module presentation. For students who unregistered the date of unregistration is also recorded. File contains five columns:

- code_module – an identification code for a module.
- code_presentation - the identification code of the presentation.
- id_student – a unique identification number for the student.

- date_registration – the date of student's registration on the module presentation, this is the number of days measured relative to the start of the module-presentation (e.g. the negative value -30 means that the student registered to module presentation 30 days before it started).
- date_unregistration – date of student unregistration from the module presentation, this is the number of days measured relative to the start of the module-presentation. Students, who completed the course have this field empty. Students who unregistered have Withdrawal as the value of the final_result column in the studentInfo.csv file.

studentAssessment.csv

This file contains the results of students' assessments. If the student does not submit the assessment, no result is recorded. The final exam submissions is missing, if the result of the assessments is not stored in the system. This file contains the following columns:

- id_assessment – the identification number of the assessment.
- id_student – a unique identification number for the student.
- date_submitted – the date of student submission, measured as the number of days since the start of the module presentation.
- is_banked – a status flag indicating that the assessment result has been transferred from a previous presentation.
- score – the student's score in this assessment. The range is from 0 to 100. The score lower than 40 is interpreted as Fail. The marks are in the range from 0 to 100.

studentVle.csv

The studentVle.csv file contains information about each student's interactions with the materials in the VLE. This file contains the following columns:

- code_module – an identification code for a module.
- code_presentation - the identification code of the module presentation.
- id_student – a unique identification number for the student.
- id_site - an identification number for the VLE material.
- date – the date of student's interaction with the material measured as the number of days since the start of the module-presentation.
- sum_click – the number of times a student interacts with the material in that day.

Environment setup

First of all you need to download and install R version 3.2.2 and RStudio. After installing required the software, we need to install package data.table, which provides enhanced functionality for the data.frame data type in R, by executing this command:

```
#intall packages
install.packages("data.table")
install.packages("dplyr")
install.packages("tidyr")
install.packages("ggplot2")
install.packages("stringr")
install.packages("DT")
install.packages("knitr")
install.packages("grid")
install.packages("gridExtra")
install.packages("corrplot")
install.packages("methods")
#install.packages("Matrix")
install.packages("reshape2")

install.packages("Rcampdf")
install.packages("ggthemes")
install.packages("qdap")
install.packages("dplyr")
install.packages("tm")
install.packages("wordcloud")
install.packages("plotrix")
install.packages("dendextend")
install.packages("ggplot2")
install.packages("ggthemes")
install.packages("RWeka")
install.packages("reshape2")
install.packages("caret")
```

After installing, we need to load library data.table into the environment. This can be done by executing this command:

```
#start by loading some libraries
```

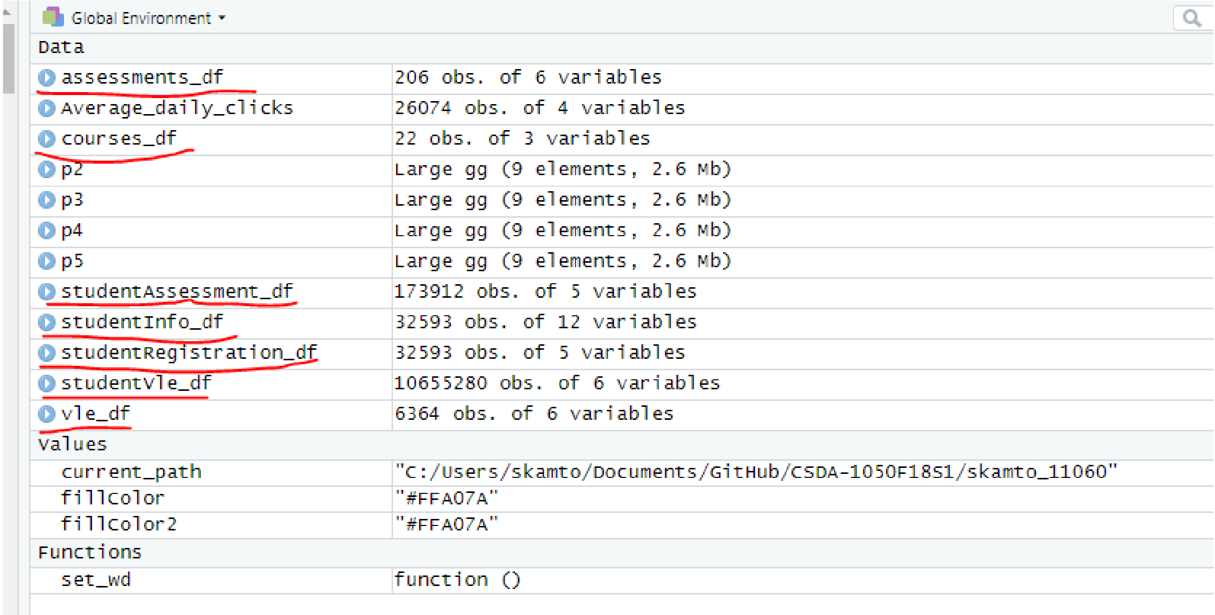
```
library(data.table)
library(dplyr)
library(tidyr)
library(ggplot2)
library(stringr)
library(DT)
library(knitr)
library(grid)
library(gridExtra)
library(corrplot)
library(methods)
library(Matrix)
library(reshape2)
```

Data preparation

First of all we need to download the data by executing this command:

download.file("http://kmi-web29.open.ac.uk:8080/resources/documents/mashupData.RData", destfile = "./mashupData.RData", mode = "wb",quiet = TRUE)

In the next step we will load data into the R environment using: load("mashupData.RData")

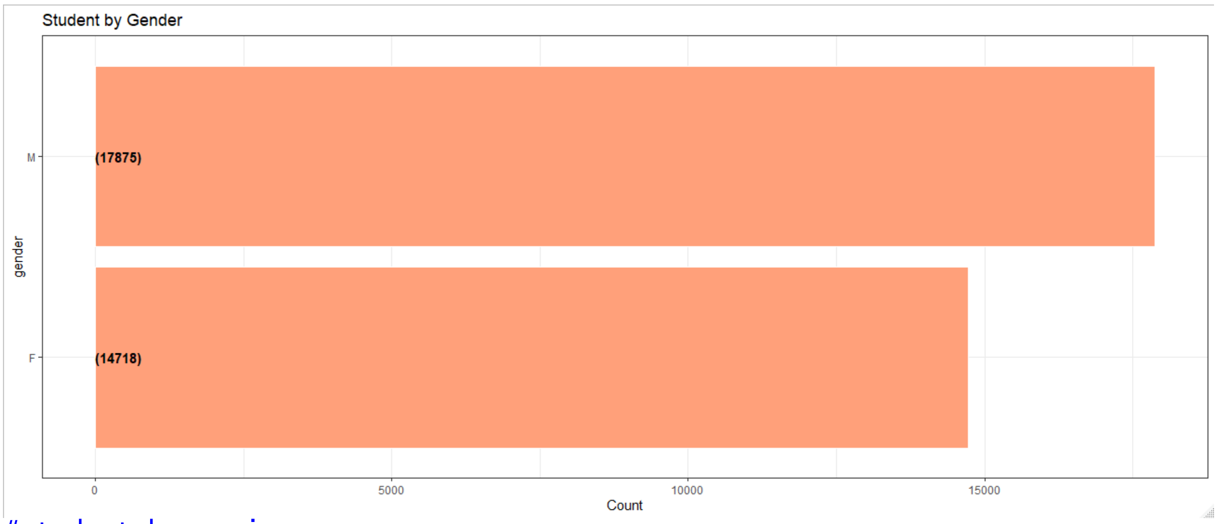


```
Data
  assessments_df             206 obs. of 6 variables
  Average_daily_clicks       26074 obs. of 4 variables
  courses_df                 22 obs. of 3 variables
  p2                         Large gg (9 elements, 2.6 Mb)
  p3                         Large gg (9 elements, 2.6 Mb)
  p4                         Large gg (9 elements, 2.6 Mb)
  p5                         Large gg (9 elements, 2.6 Mb)
  studentAssessment_df       173912 obs. of 5 variables
  studentInfo_df             32593 obs. of 12 variables
  studentRegistration_df     32593 obs. of 5 variables
  studentVle_df              10655280 obs. of 6 variables
  vle_df                     6364 obs. of 6 variables
Values
  current_path               "C:/Users/skamto/Documents/GitHub/CSDA-1050F18S1/skamto_11060"
  fillColor                  "#FFA07A"
  fillColor2                 "#FFA07A"
Functions
  set_wd                     function ()
```

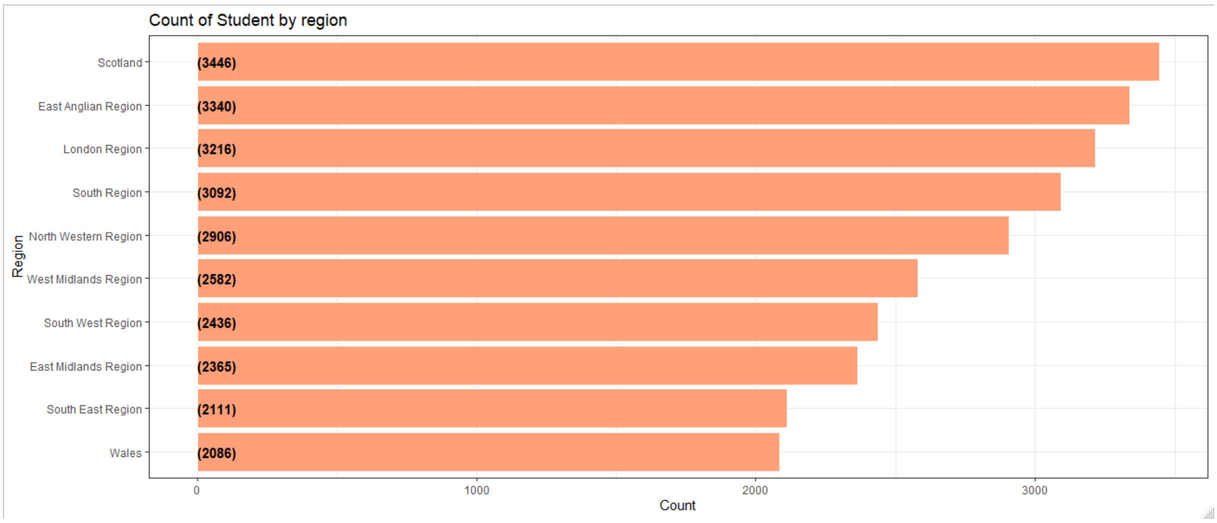
Data Exploration

```
> summary(courses_df)
 code_module          code_presentation   module_presentation_length
 Length:22            Length:22           Min.   :234.0
 Class :character   Class :character   1st Qu.:241.0
 Mode  :character   Mode  :character   Median :261.5
                                       Mean   :255.5
                                       3rd Qu.:268.0
                                       Max.   :269.0
> summary(assessments_df)
code_module        code_presentation  id_assessment   assessment_type         date          weight
 Length:206         Length:206         Min.   : 1752   Length:206         Min.   : 12   Min.   :  0.00
 Class :character   Class :character   1st Qu.:15023   Class :character   1st Qu.: 71   1st Qu.:  0.00
 Mode  :character   Mode  :character   Median :25365   Mode  :character   Median :152   Median : 12.50
                                       Mean   :26474                      Mean   :145   Mean   : 20.87
                                       3rd Qu.:34892                      3rd Qu.:222   3rd Qu.: 24.25
                                       Max.   :40088                      Max.   :261   Max.   :100.00
                                                                          NA's   :11
> summary(vle_df)
    id_site          code_module        code_presentation  activity_type          week_from        week_to
 Min.   : 526721  Length:6364        Length:6364        Length:6364        Min.   : 0.0   Min.   : 0.00
 1st Qu.: 661593  Class :character   Class :character   Class :character   1st Qu.: 8.0   1st Qu.: 8.00
 Median : 730097  Mode  :character   Mode  :character   Mode  :character   Median :15.0   Median :15.00
 Mean   : 726099                                                           Mean   :15.2   Mean   :15.21
 3rd Qu.: 814016                                                           3rd Qu.:22.0   3rd Qu.:22.00
 Max.   :1077905                                                           Max.   :29.0   Max.   :29.00
                                                                           NA's   :5243   NA's   :5243
> summary(studentInfo_df)
 code_module        code_presentation    id_student         gender              region             highest_educ
ation     imd_band
 Length:32593       Length:32593       Min.   :   3733   Length:32593       Length:32593       Length:32593
Length:32593
 Class :character   Class :character   1st Qu.: 508573   Class :character   Class :character   Class :chara
cter   Class :character
 Mode  :character   Mode  :character   Median : 590310   Mode  :character   Mode  :character   Mode  :chara
cter   Mode  :character
                                       Mean   : 706688
                                       3rd Qu.: 644453
                                       Max.   :2716795
   age_band         num_of_prev_attempts studied_credits   disability         final_result
 Length:32593       Min.   :0.0000       Min.   : 30.00   Length:32593       Length:32593
 Class :character   1st Qu.:0.0000       1st Qu.: 60.00   Class :character   Class :character
 Mode  :character   Median :0.0000       Median : 60.00   Mode  :character   Mode  :character
                    Mean   :0.1632       Mean   : 79.76
                    3rd Qu.:0.0000       3rd Qu.:120.00
                    Max.   :6.0000       Max.   :655.00
> summary(studentRegistration_df)
 code_module        code_presentation    id_student      date_registration date_unregistration
 Length:32593       Length:32593       Min.   :   3733   Min.   :-322.00   Min.   :-365.00
 Class :character   Class :character   1st Qu.: 508573   1st Qu.:-100.00   1st Qu.:  -2.00
 Mode  :character   Mode  :character   Median : 590310   Median : -57.00   Median :  27.00
                                       Mean   : 706688   Mean   : -69.41   Mean   :  49.76
                                       3rd Qu.: 644453   3rd Qu.: -29.00   3rd Qu.: 109.00
                                       Max.   :2716795   Max.   : 167.00   Max.   : 444.00
                                                         NA's   :45        NA's   :22521
> summary(studentAssessment_df)
  id_assessment     id_student      date_submitted    is_banked           score
 Min.   : 1752   Min.   :   6516   Min.   :-11    Min.   :0.00000   Min.   :  0.0
 1st Qu.:15022   1st Qu.: 504429   1st Qu.: 51    1st Qu.:0.00000   1st Qu.: 65.0
 Median :25359   Median : 585208   Median :116    Median :0.00000   Median : 80.0
 Mean   :26554   Mean   : 705151   Mean   :116    Mean   :0.01098   Mean   : 75.8
 3rd Qu.:34883   3rd Qu.: 634498   3rd Qu.:173    3rd Qu.:0.00000   3rd Qu.: 90.0
 Max.   :37443   Max.   :2698588   Max.   :608    Max.   :1.00000   Max.   :100.0
                                                                    NA's   :173
> summary(studentVle_df)
code_module        code_presentation    id_student          id_site             date            sum_click
 Length:10655280    Length:10655280    Min.   :   6516   Min.   : 526721   Min.   :-25.00   Min.   :   1.00
0
 Class :character   Class :character   1st Qu.: 507743   1st Qu.: 673519   1st Qu.: 25.00   1st Qu.:   1.00
0
 Mode  :character   Mode  :character   Median : 588236   Median : 730069   Median : 86.00   Median :   2.00
0
                                       Mean   : 733334   Mean   : 738323   Mean   : 95.17   Mean   :   3.71
7
                                       3rd Qu.: 646484   3rd Qu.: 877030   3rd Qu.:156.00   3rd Qu.:   3.00
0
                                       Max.   :2698588   Max.   :1049562   Max.   :269.00   Max.   :6977.00
0
```

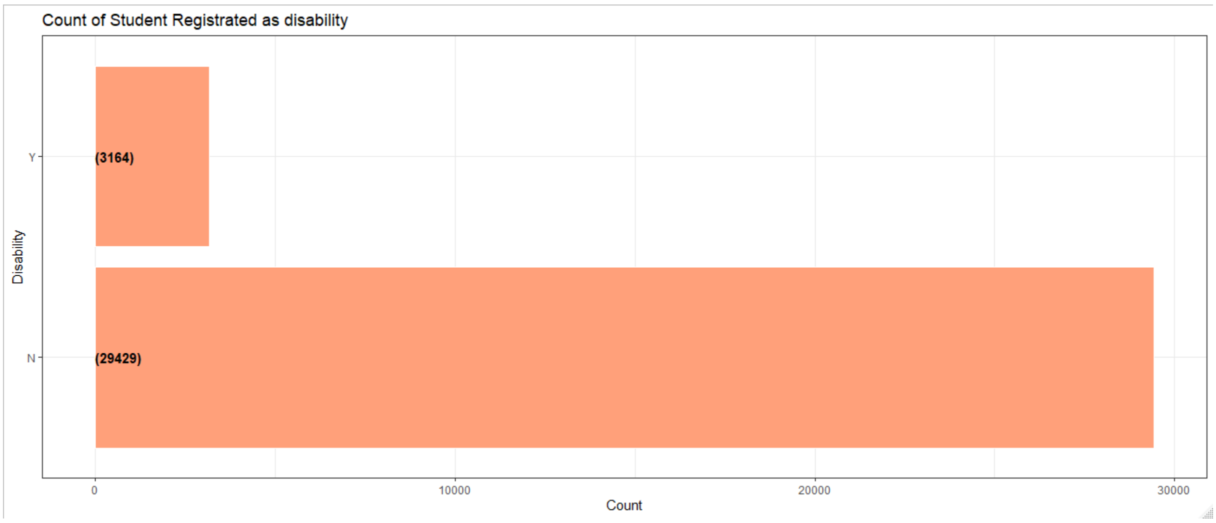
```
fillColor = "#FFA07A"
fillColor2 = "#FFA07A"
#student by gender
studentInfo_df %>%
  group_by(gender) %>%
  filter(!is.na(gender)) %>%
  summarise(Count = n()) %>%
  ungroup() %>%
  mutate(gender = reorder(gender,Count)) %>%
  arrange(desc(Count)) %>%
  head(10) %>%

  ggplot(aes(x = gender,y = Count)) +
  geom_bar(stat='identity',colour="white", fill = fillColor2) +
  geom_text(aes(x = gender, y = 1, label = paste0("(",Count,")",sep="")),
            hjust=0, vjust=.5, size = 4, colour = 'black',
            fontface = 'bold') +
  labs(x = 'gender',
       y = 'Count',
       title = 'Student by Gender') +
  coord_flip() +
  theme_bw()
```



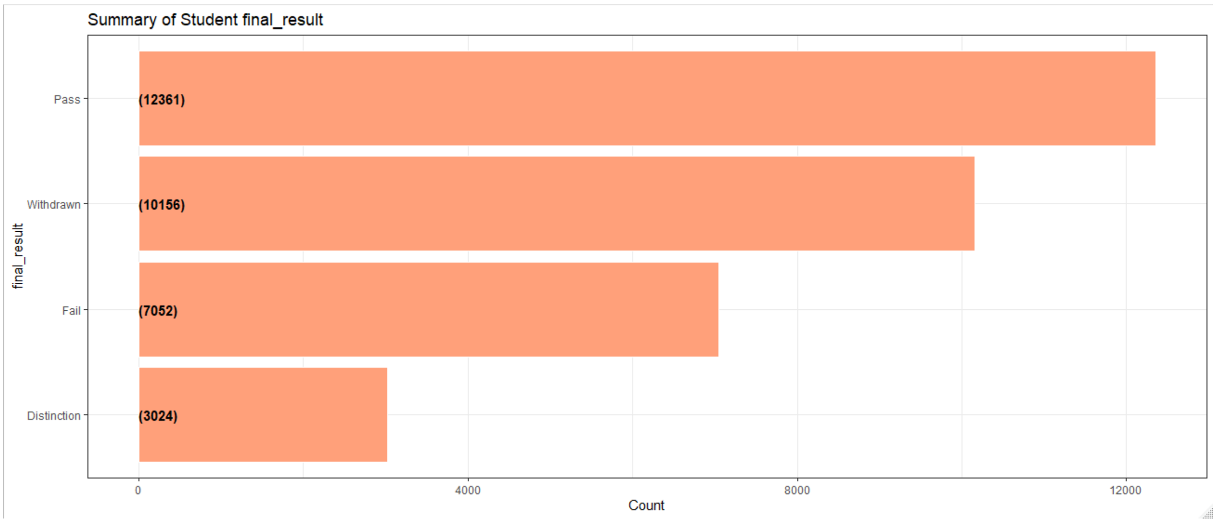
```
#student by region
studentInfo_df %>%
  group_by(region) %>%
  filter(!is.na(region)) %>%
  summarise(Count = n()) %>%
  ungroup() %>%
  mutate(region = reorder(region,Count)) %>%
  arrange(desc(Count)) %>%
  head(10) %>%

  ggplot(aes(x = region,y = Count)) +
  geom_bar(stat='identity',colour="white", fill = fillColor2) +
  geom_text(aes(x = region, y = 1, label = paste0("(",Count,")",sep="")),
            hjust=0, vjust=.5, size = 4, colour = 'black',
            fontface = 'bold') +
  labs(x = 'Region',
       y = 'Count',
       title = 'Count of Student by region') +
  coord_flip() +
  theme_bw()
```



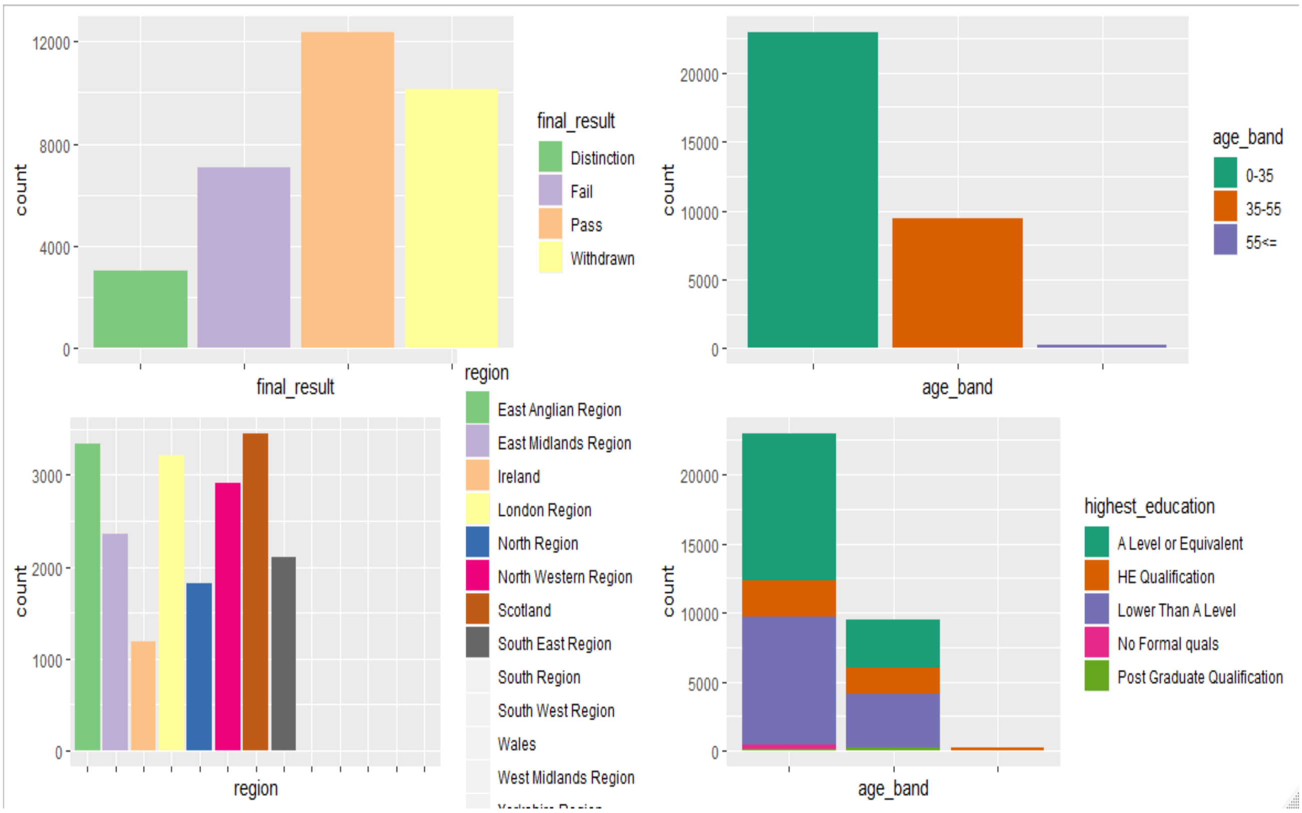
```
#Count of Student by ages
studentInfo_df %>%
  group_by(age_band) %>%
  filter(!is.na(age_band)) %>%
  summarise(Count = n()) %>%
  ungroup() %>%
  mutate(age_band = reorder(age_band,Count)) %>%
  arrange(desc(Count)) %>%
  head(10) %>%

  ggplot(aes(x = age_band,y = Count)) +
  geom_bar(stat='identity',colour="white", fill = fillColor2) +
  geom_text(aes(x = age_band, y = 1, label = paste0("(",Count,")",sep="")),
            hjust=0, vjust=.5, size = 4, colour = 'black',
            fontface = 'bold') +
  labs(x = 'Ages',
       y = 'Count',
       title = 'Count of Student by age_band') +
  coord_flip() +
  theme_bw()
```



```
> #Count of Student by final_result
> studentInfo_df %>%
+   group_by(final_result) %>%
+   filter(!is.na(final_result)) %>%
+   summarise(Count = n()) %>%
+   ungroup() %>%
+   mutate(final_result = reorder(final_result,Count)) %>%
+   arrange(desc(Count)) %>%
+   head(10) %>%
+
+   ggplot(aes(x = final_result,y = Count)) +
+   geom_bar(stat='identity',colour="white", fill = fillColor2) +
+   geom_text(aes(x = final_result, y = 1, label = paste0("(",Count,")",sep
="")),
+             hjust=0, vjust=.5, size = 4, colour = 'black',
+             fontface = 'bold') +
+   labs(x = 'final_result',
+        y = 'Count',
+        title = 'Summary of Student final_result') +
+   coord_flip() +
+   theme_bw()
```



```
> p2 <- ggplot(studentInfo_df, aes(x = final_result)) + geom_bar(aes(fill =
final_result)) +
+   theme(axis.text.x = element_blank()) + scale_fill_brewer(palette="Accen
t")
> p3 <- ggplot(studentInfo_df, aes(x = age_band)) + geom_bar(aes(fill = age
_band)) +
+   theme(axis.text.x = element_blank()) + scale_fill_brewer(palette="Dark2
")
> p4 <- ggplot(studentInfo_df, aes(x = final_result)) + geom_bar(aes(fill =
region)) +
+   theme(axis.text.x = element_blank()) + scale_fill_brewer(palette="Accen
t")
> p5 <- ggplot(studentInfo_df, aes(x = age_band)) + geom_bar(aes(fill = hig
hest_education)) +
+   theme(axis.text.x = element_blank()) + scale_fill_brewer(palette="Dark2
")
> grid.arrange(p2, p3, p4, p5, nrow=2, ncol=2)
```



Prediction Activity

Wrangling

- Calculate the average daily number of clicks (site interactions) for each student from the `studentVle` dataset

- Calculate the average assessment score for each student from the `studentAssessment` dataset

- Merge your click and assessment score average values into the the `studentInfo` dataset

Create a Validation Set

- Split your data into two new datasets, `TRAINING` and `TEST`, by **randomly** selecting 25% of the students for the `TEST`set

Explore

- Generate summary statistics for the variable `final_result`
- Ensure that the final_result variable is binary (Remove all students who withdrew from a courses and convert all students who recieved distinctions to pass)
- Visualize the distributions of each of the variables for insight
- Visualize relationships between variables for insight

Model Training

- We will be allocated one of the following models to test:

  CART (`RPART`), Conditional Inference Trees (`party`), Naive Bayes (`naivebayes`), Logistic Regression (`gpls`)
- Using the `trainControl` command in the `caret` package we will create a 10-fold cross-validation harness:
  ```
  control <- trainControl(method="cv", number=10)
  ```
- Using the standard caret syntax fit our model and measure accuracy:
  ```
  fit <- train(final_result~., data=TRAINING, method=YOUR MODEL,
  metric="accuracy", trControl=control)
  ```

- A summary of our results will be generated and a visualization of the accuracy scores for our ten trials will be created

- Make any tweaks to our model to try to improve its performance

Model Testing

- Use the `predict` function to test our model
  ```
  predictions <- predict(fit, TEST)
  ```
- Generate a confusion matrix for our model test
  ```
  confusionMatrix(predictions, TEST$final_result)
  ```