Data Mining for Student Success and Perseverance

Sameer Bhatnagar
Jonathan Guillemette
Micheal Dugdale
Sahir Bhatnagar
Nathaniel Lasry

2017-05-29

Contents

Chapter 1

Preface

This report summarizes the work done by our team on using college registration records at three diffrenet anglophone CEGEPS in Montreal in order to find predictors of attrition.

Chapter 2

Introduction

Chapter 3

Literature

The most important relevant work for this project is (Jorgensen et al., 2009).

Chapter 4

Descriptive Statistics

Here in we will describe - the data set - the methods by which we label students at risk - the distributions of at-risk students by - demographic indicators - registration record indicators

4.1 Demographics across the colleges and major programs

4.2 Fraction of students who change colleges

Chapter 5

Methods Centered on Determining Predictive Factors

This chapter will be focused on methods which have a sound probabilistic framework, and allow for inference into the statistical importance of predictive factors.

- Logistic Regression
- Mixed Effects Models

Chapter 6

Methods centered on predicting at-risk students

This chapter will explore the use of machine learning algorithms whose primary focus is prediction. This comes at the expense of model interpretatbility, and this tradeoff will be discussed herein as well.

- Decision Trees and Random Forests
- Neural Networks

6.1 Example one

6.2 Example two

Chapter 7

Comparisons

This chapter will compare the effectiveness of the methods developped in the previous two chapters, and compare them to more basic approaches to identifying students at-risk.

For example, we know that some CEGEPs have implemented a policy whereby they identify students as being at risk based on their mid-term assessments: if the student receives a certain number of "at-risk" or "failing" results, they are automatically sent an email referring them to academic support services.

Based on this, we can ask the following research questions : - how effective is this approach at identifying students who drop-out? - how does this approach compare to our models from the previous chapters?

We begin with a basic logistic regression with demographic variables, and as well as the number of each type of results of mid-term assessment, for students in their last term at the college. With these predictors, we try to predict if students are about to graduate, or simply not register again.

	Estimate	Std. Error	z value	Pr(>|z|)
num__pass	0.5877	0.02321	25.32	1.747e-141
num__at__risk	1.455	0.03664	39.7	0
num__failing	2.193	0.04653	47.15	0
num__courses	-0.7442	0.02373	-31.36	6.315e-216
SexeM	0.2111	0.0386	5.469	4.537e-08
birth__placeQuebec	-0.5914	0.04815	-12.28	1.108e-34
LangueMaternelleAU	-0.193	0.05173	-3.731	0.0001907
LangueMaternelleFR	0.3504	0.04914	7.13	1.006e-12
(Intercept)	0.5654	0.0694	8.147	3.741e-16

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	24452 on 18370 degrees of freedom
Residual deviance:	17188 on 18362 degrees of freedom

Chapter 8

Continuing Education at Dawson

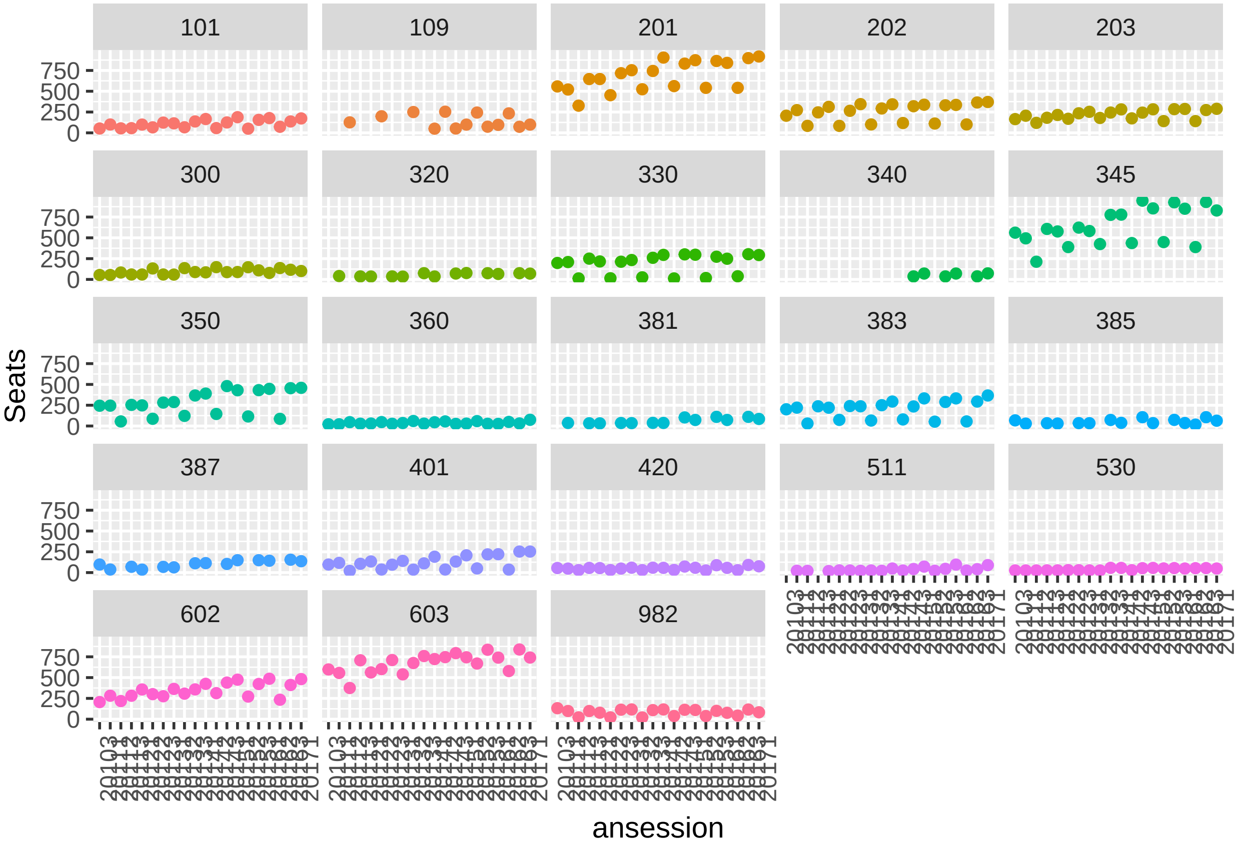
The department of Continuing at Education offers evening courses across many disciplines throughout the year. The purpose of this report is to give a broad overview of the demographics and sector-level metrics for students taking courses within Cont Ed.

8.1 Services

Herein, we look at the services offered by the Division of Continuing Education.

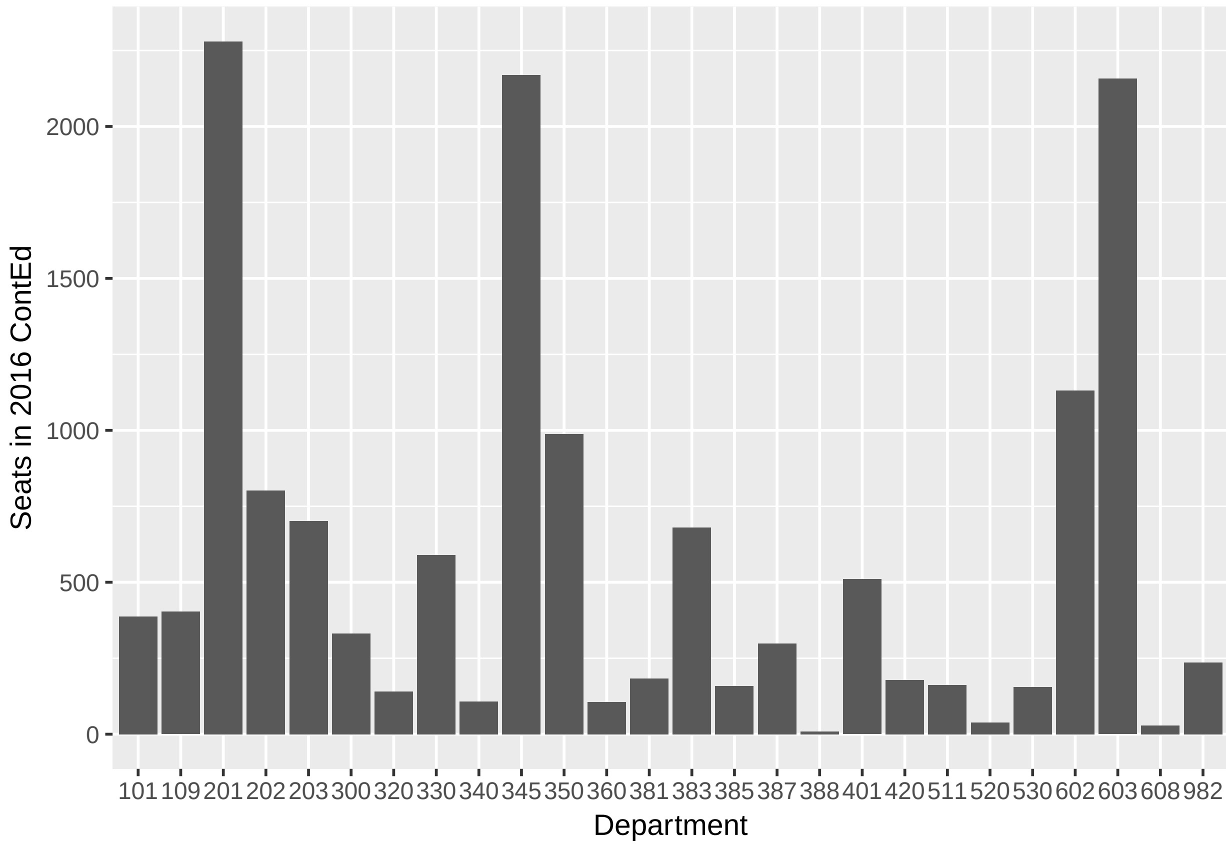
8.1.1 Continuing Education's Growth over Time

How have each of the departments increased their ContEd offerings over time?



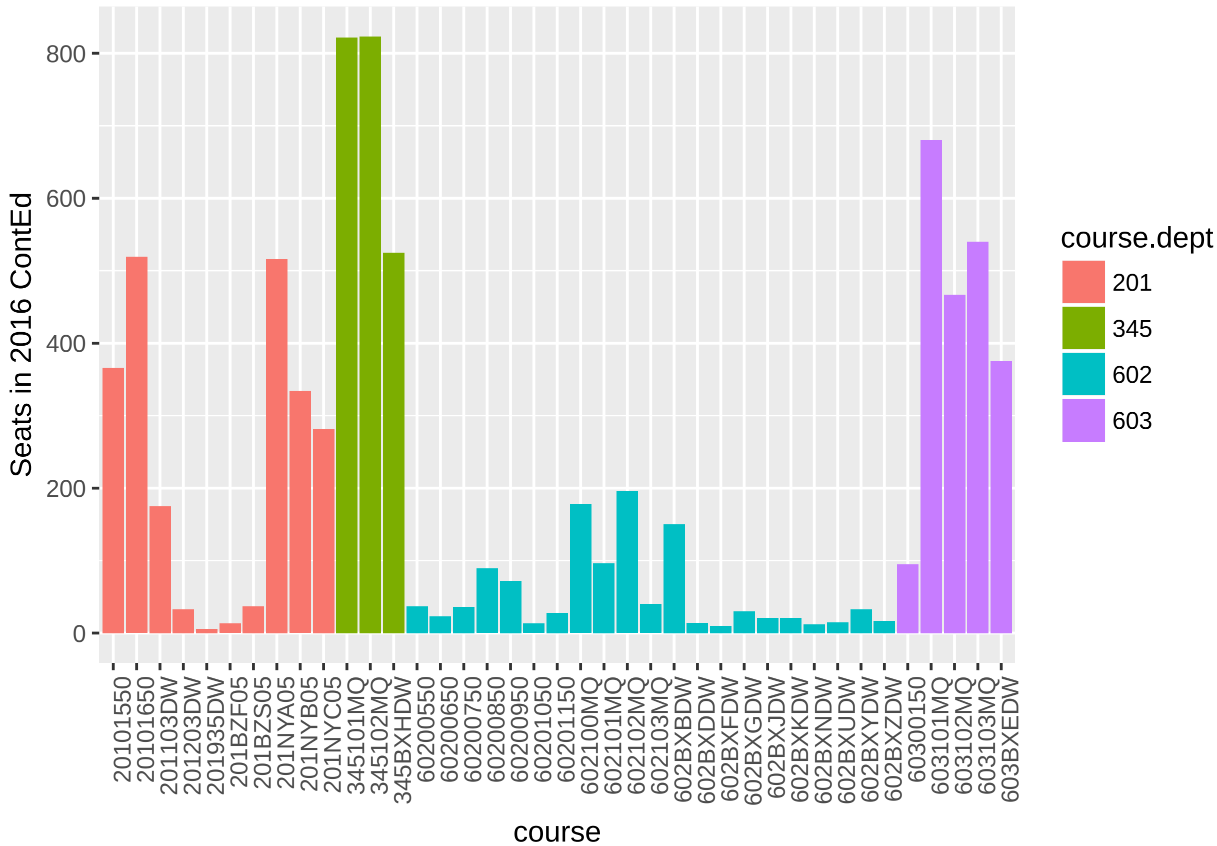
8.1.2 Department Level Seat Distributions

Which parts of Continuing Education are the most important in 2016?



8.1.3 Course Level distributions

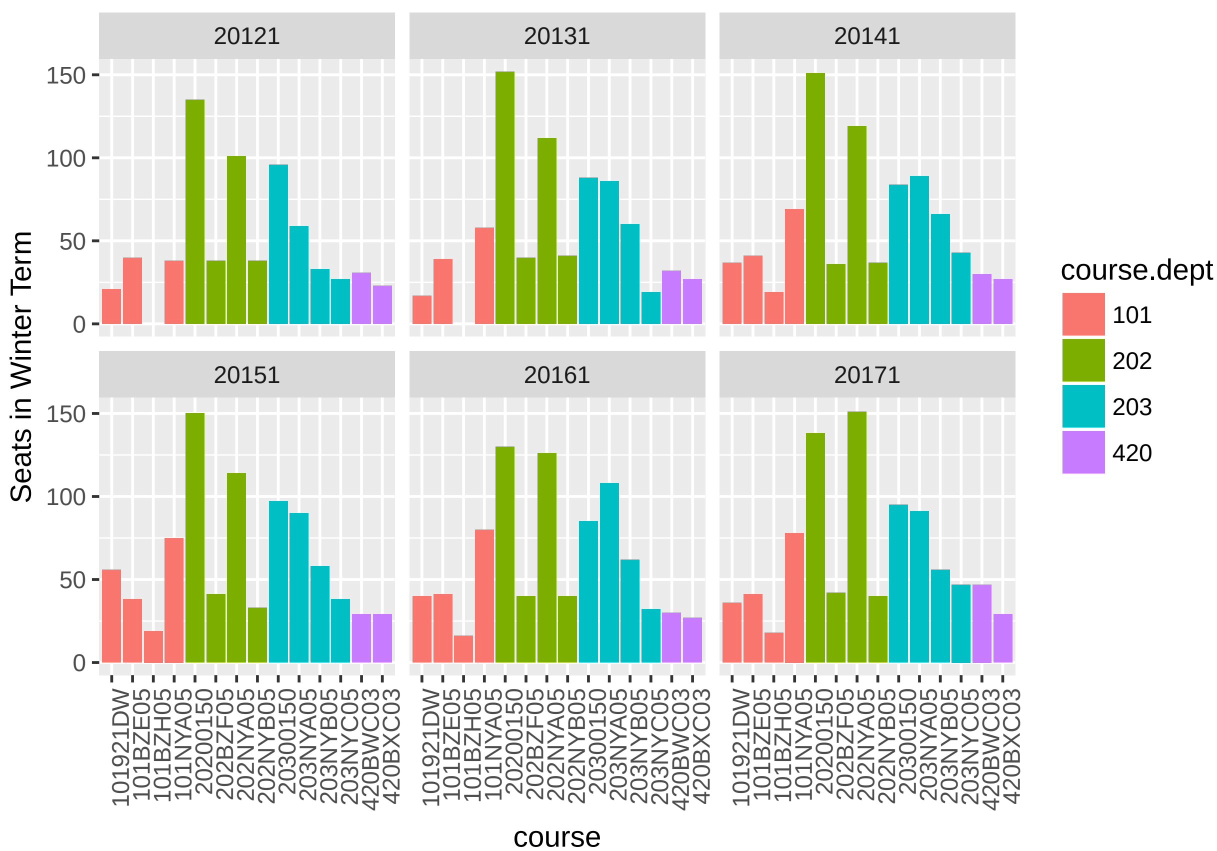
If we look at the three most important departments (Math,Humanities and English) in 2016, how are the seats distrinuted across courses?



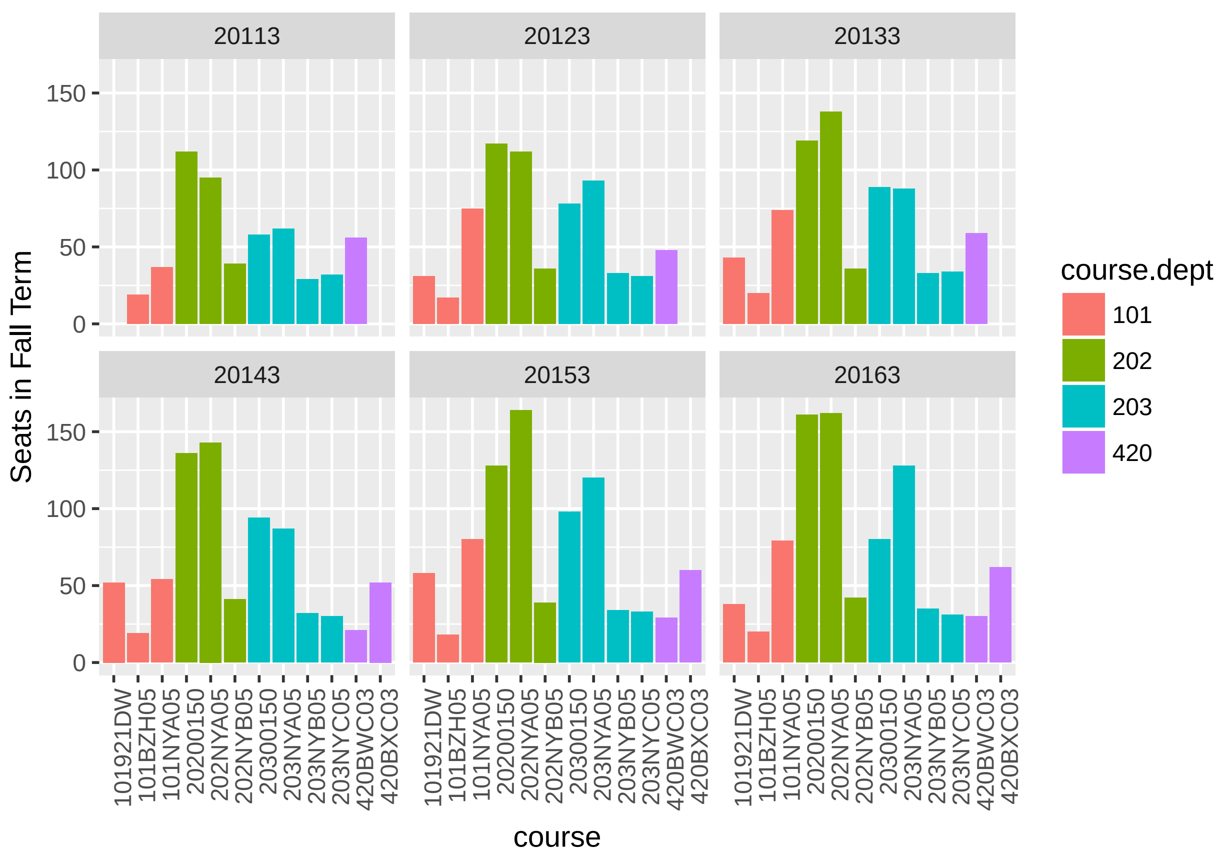
8.1.4 Specialized spaces

If we look at the departments that require specialized spaces (i.e. labs), in 2016, how are the seats distributed by course? - Has this evolved over time? - What are the seasonal variations?

8.1.4.1 Winter


20121
20131
20141
20151
20161
20171
Seats in Winter Term
0
50
100
150
101921DW
101BZE05
101BZH05
101NYA05
20200150
202BZF05
202NYA05
202NYB05
20300150
203NYA05
203NYB05
203NYC05
420BWC03
420BXC03
course
course.dept
101
202
203
420

8.1.4.2 Fall


20113
20123
20133
20143
20153
20163
Seats in Fall Term
150
100
50
0
course
101921DW
101BZH05
101NYA05
20200150
202NYA05
202NYB05
20300150
203NYA05
203NYB05
203NYC05
420BWC03
420BXC03
course.dept
101
202
203
420

8.1.4.3 Summer

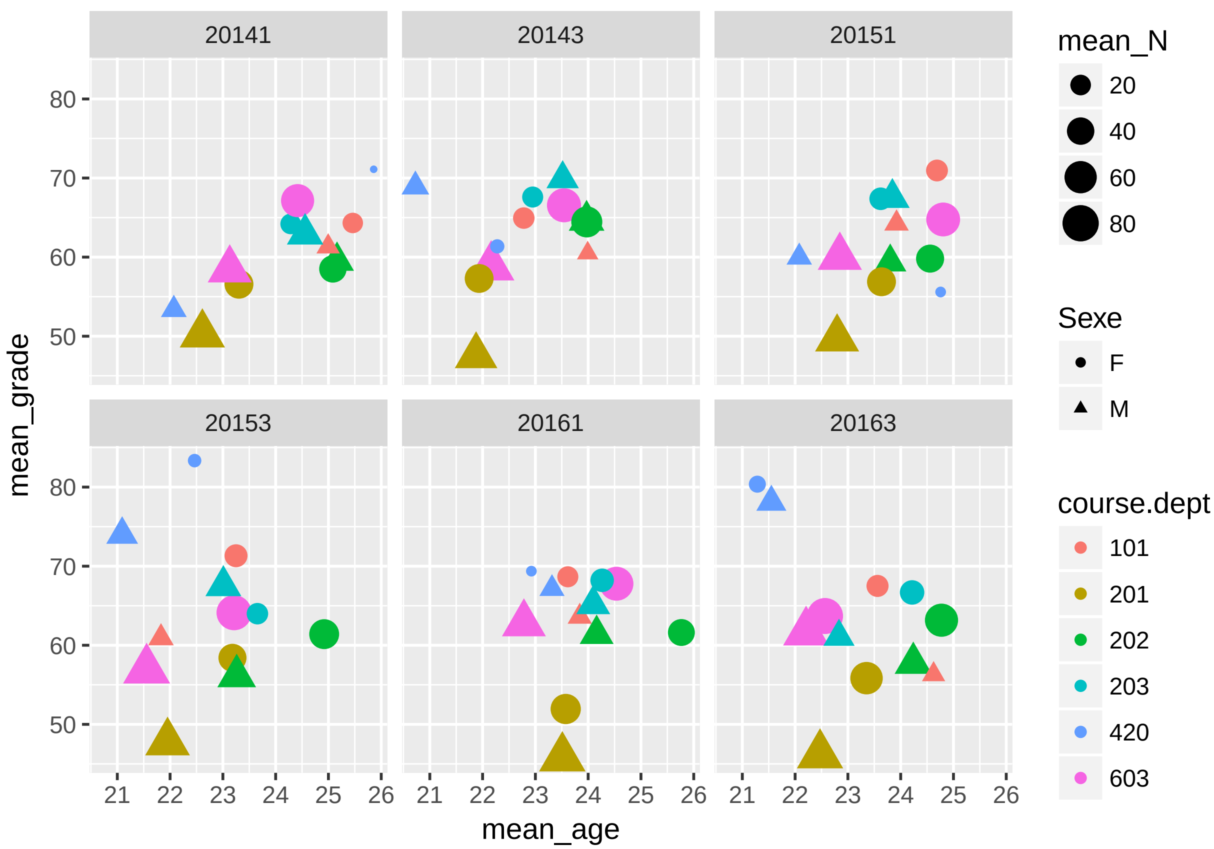
8.2 Students

8.2.1 Demographics

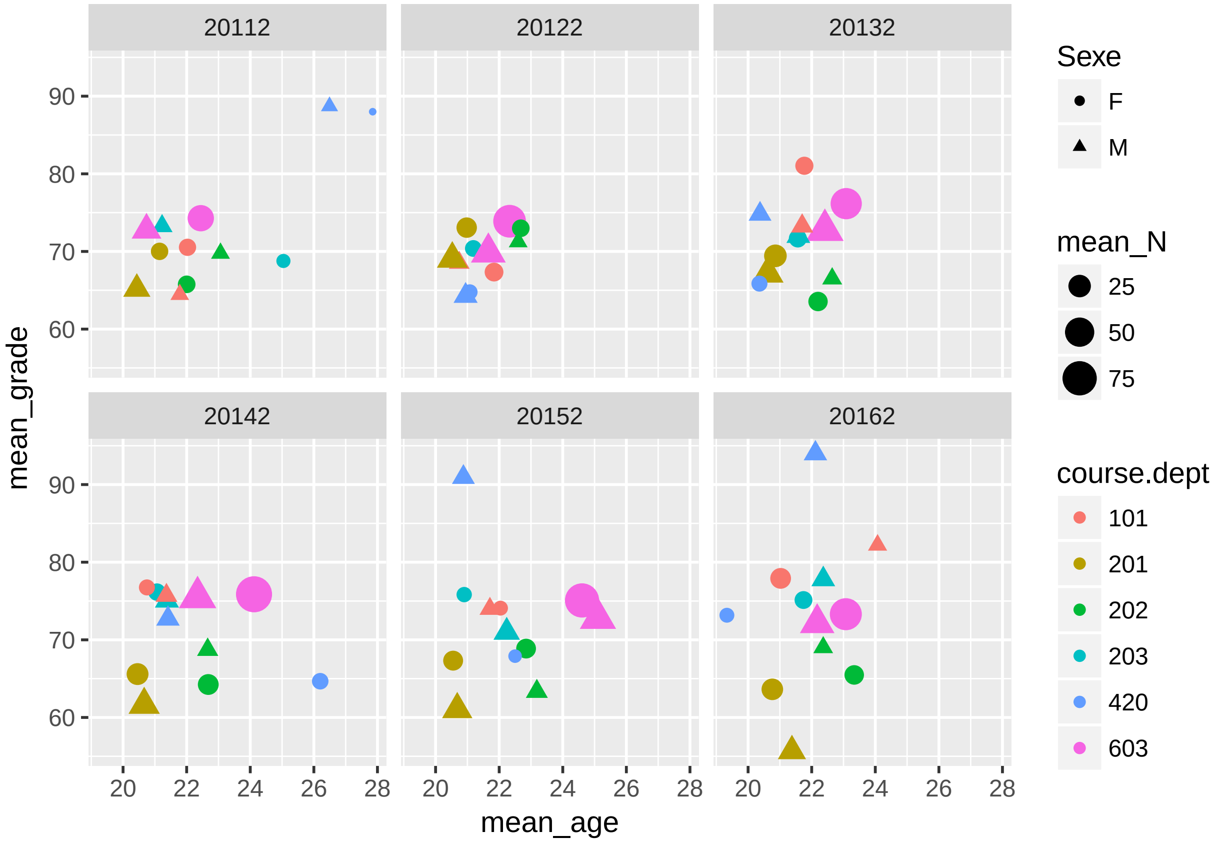
Who are the students using continuing education services?

- are there gender differences?
- are there age differences?
- do these demographics change over time?
- are there seasonal variations?
- are there differences in different departments?

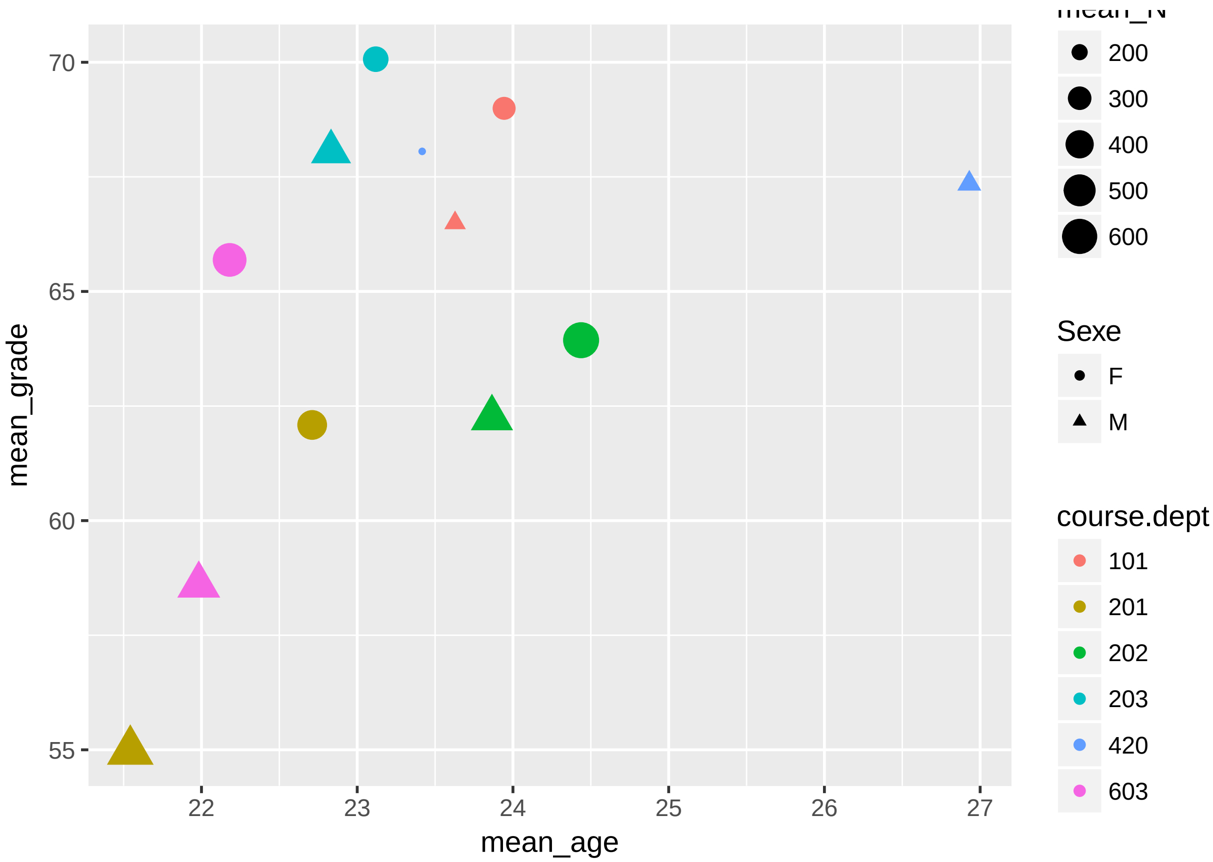
8.2.1.1 Sexe



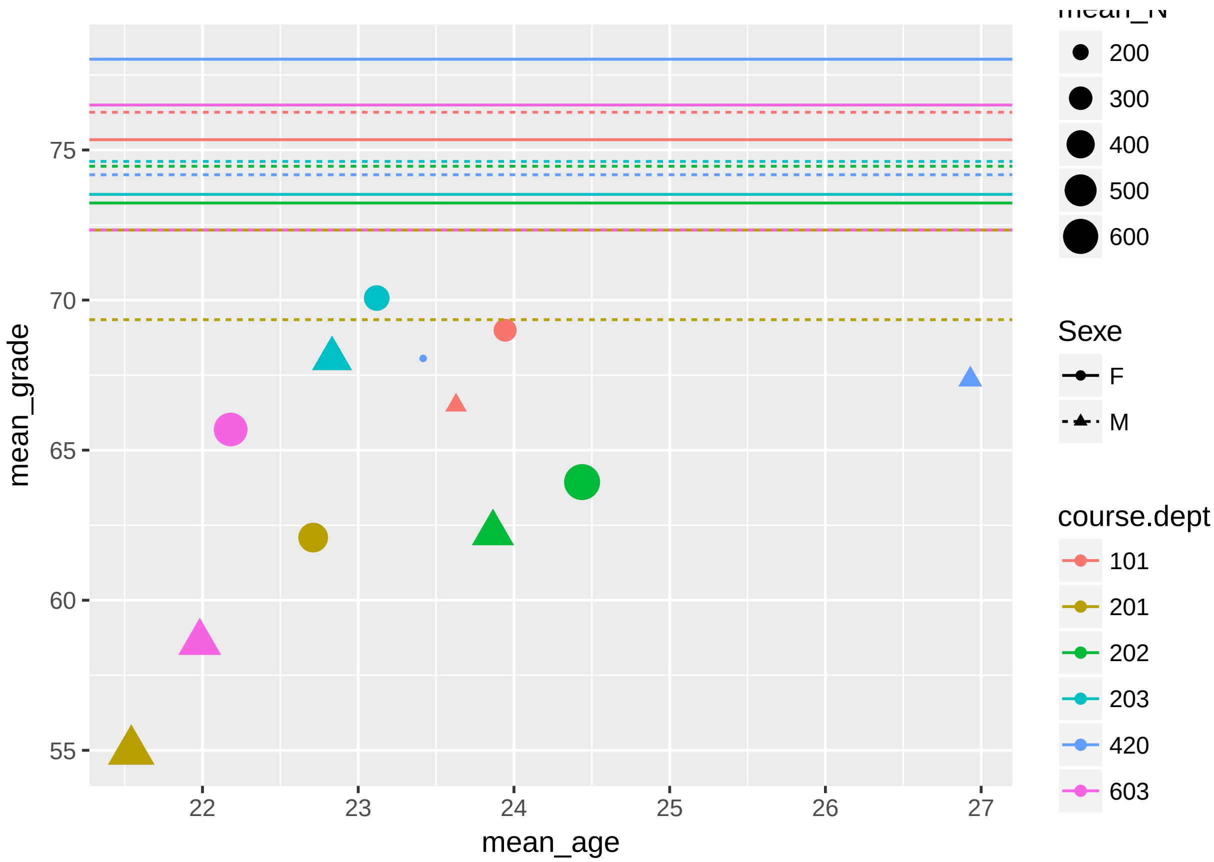
What is impact of condensed summer term?



Finally, as there seems to be relatively little change in these patterns over time, we can collapse all over the past seven years,



and then add on the average grade achieved by students with the same gender, in the same courses from the same disciplines, but from the "regular" day division.

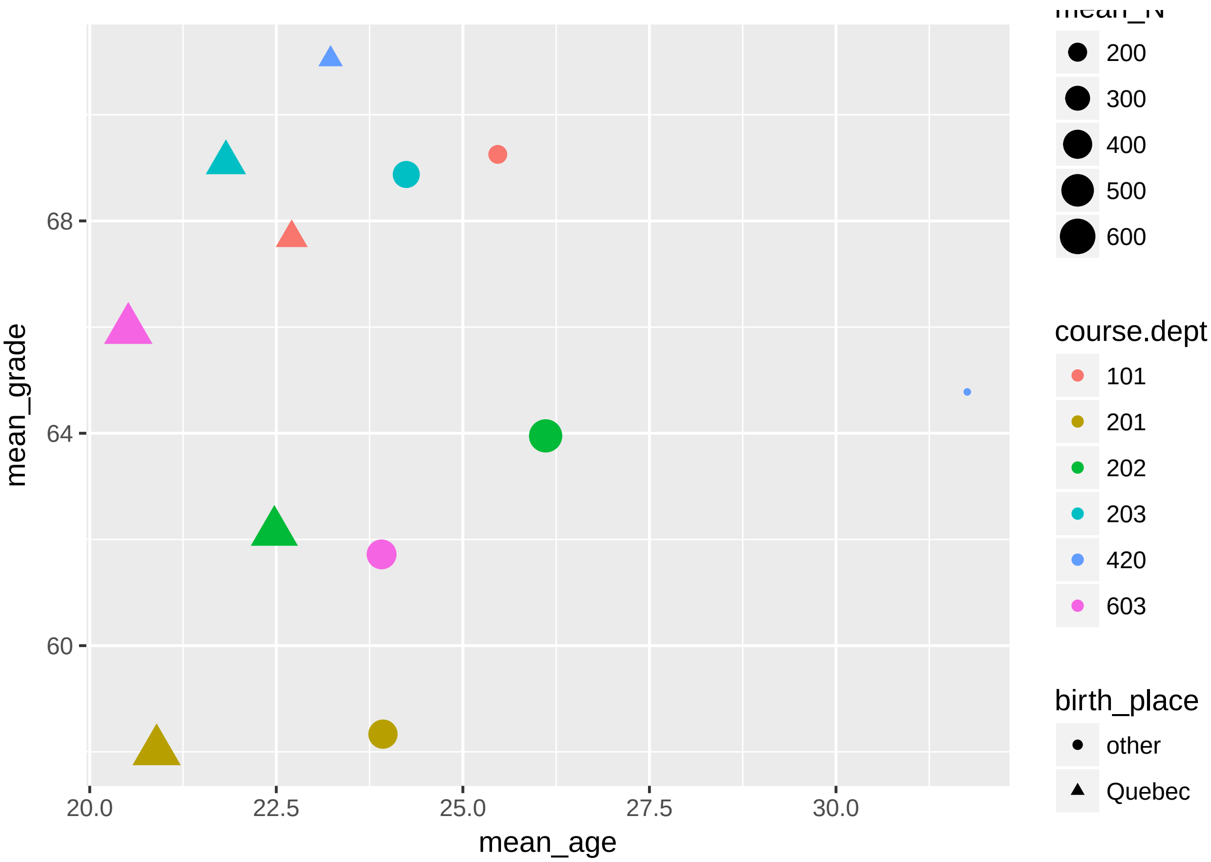


8.2.1.2 Birth Place

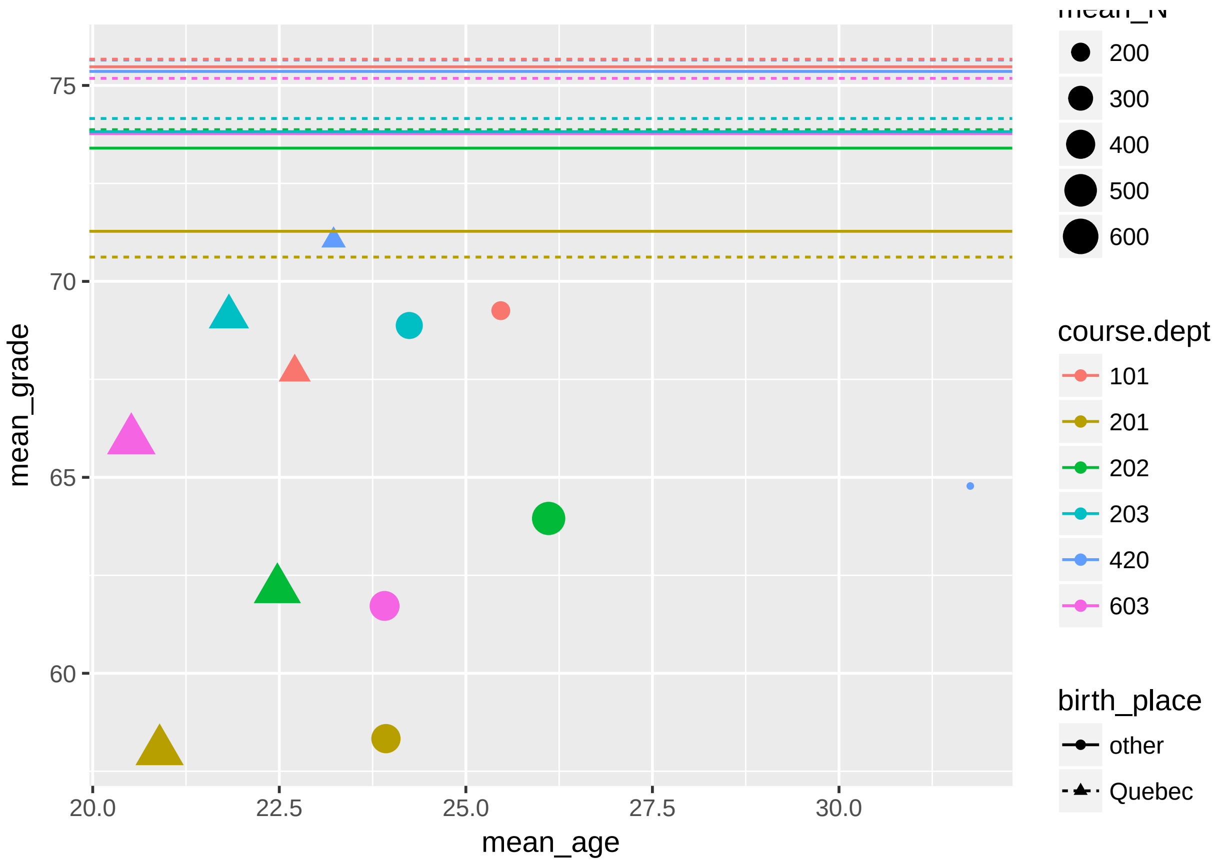
Now, instead of looking at effect of gender, we focus instead on Birth Place. To Quebec residents stand out in any consistent way, as compared to those born elsewhere?

Again, what is impact of condensed summer term?

Finally, as there seems to be relatively little change in these patterns over time, we can collapse all over the past seven years,

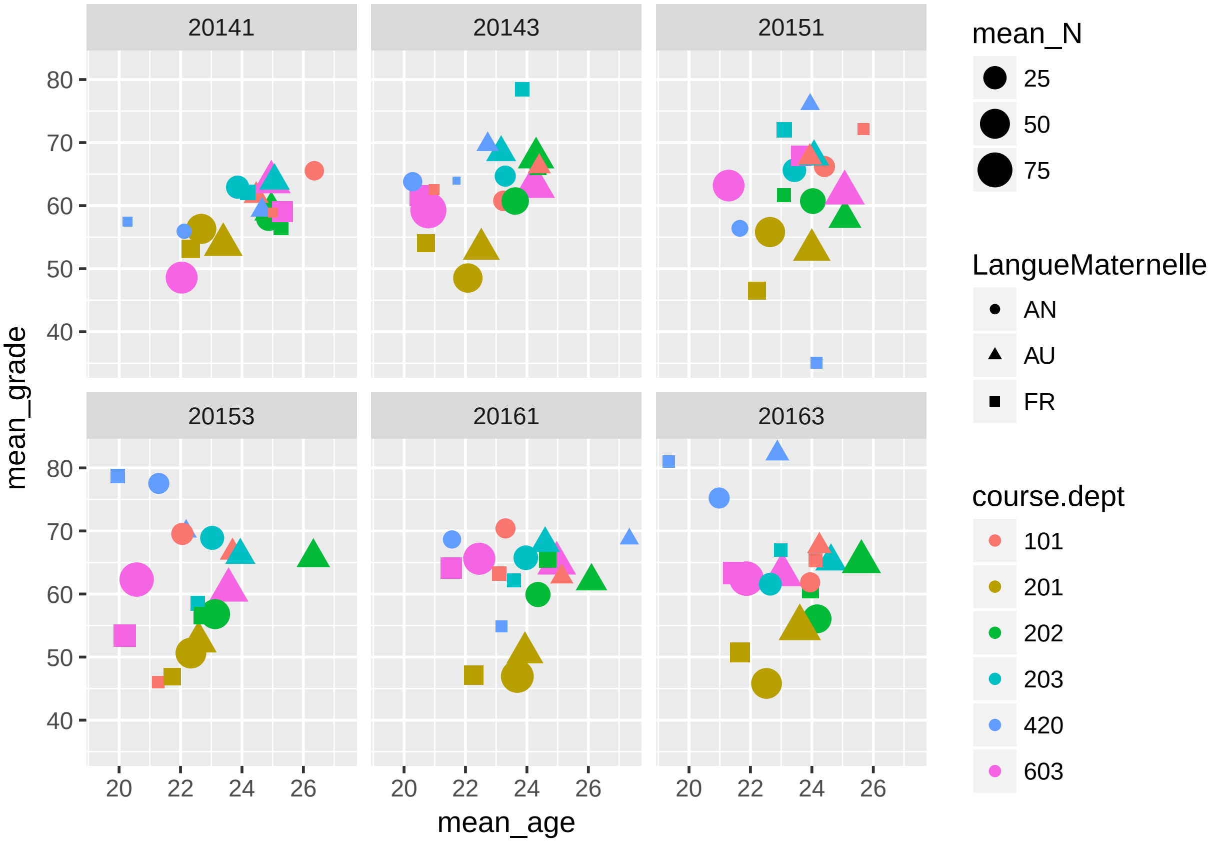


and then add on the average grade achieved by students with the same **birth place** (namely quebec residents vs *other*), in the same courses from the same disciplines, but from the "regular" day division.

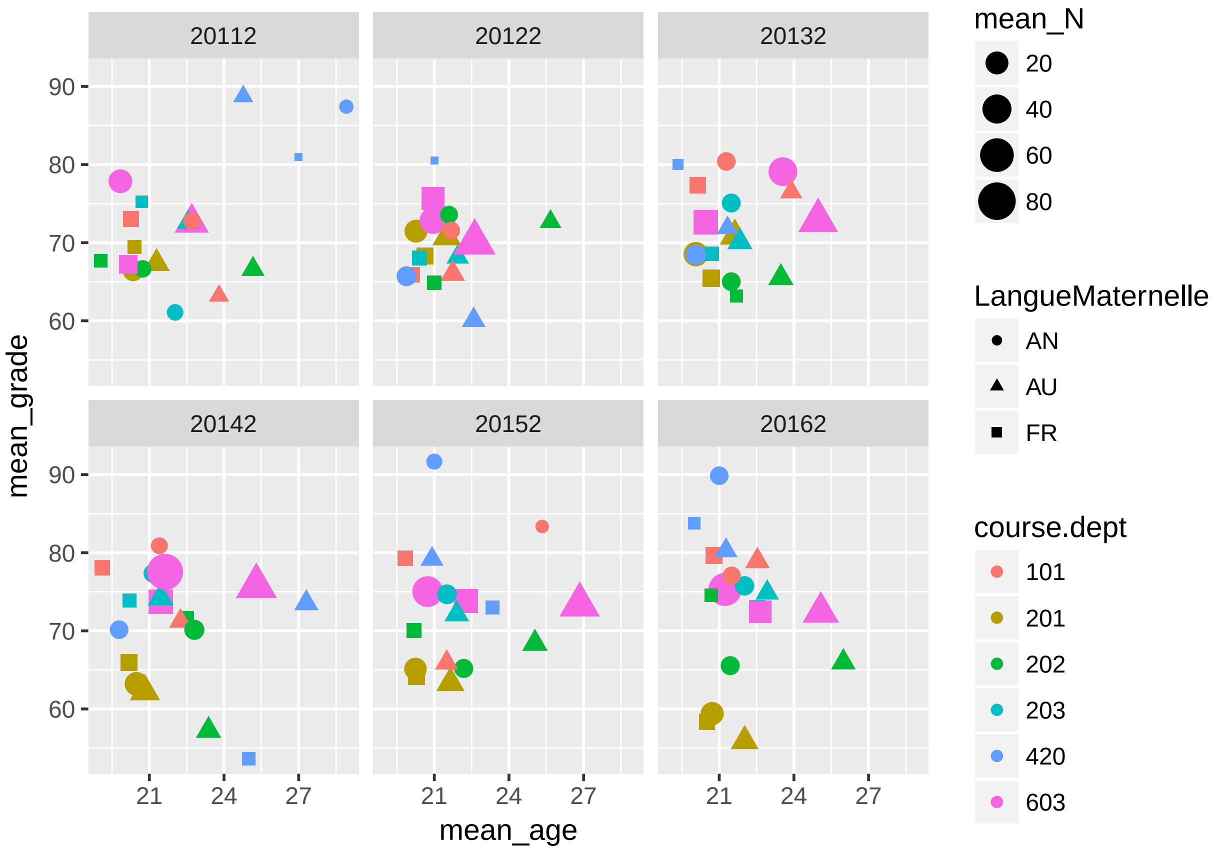


8.2.1.3 Mother Tongue

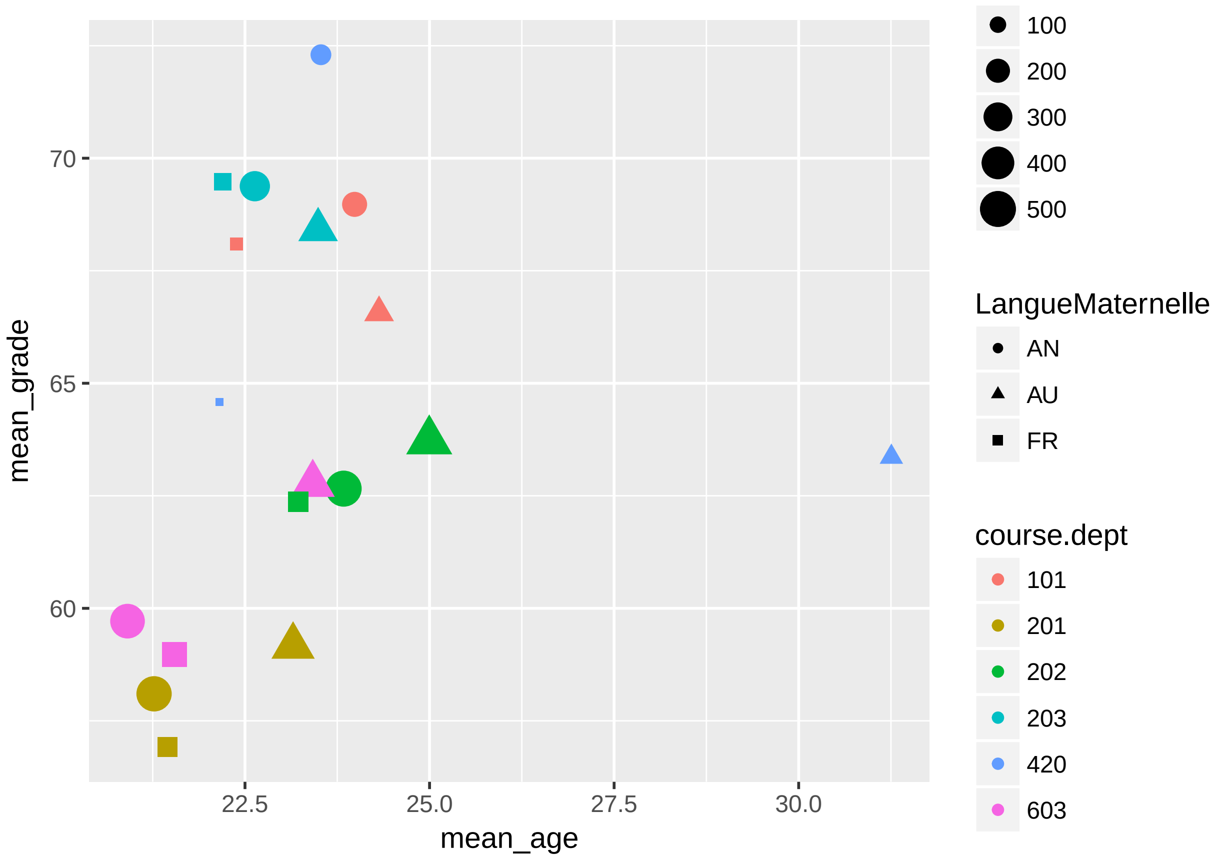
Finally we look at the possible impact of Mother Tongue. How do anglophones, francophones, and allophones compare in Continuing Education?



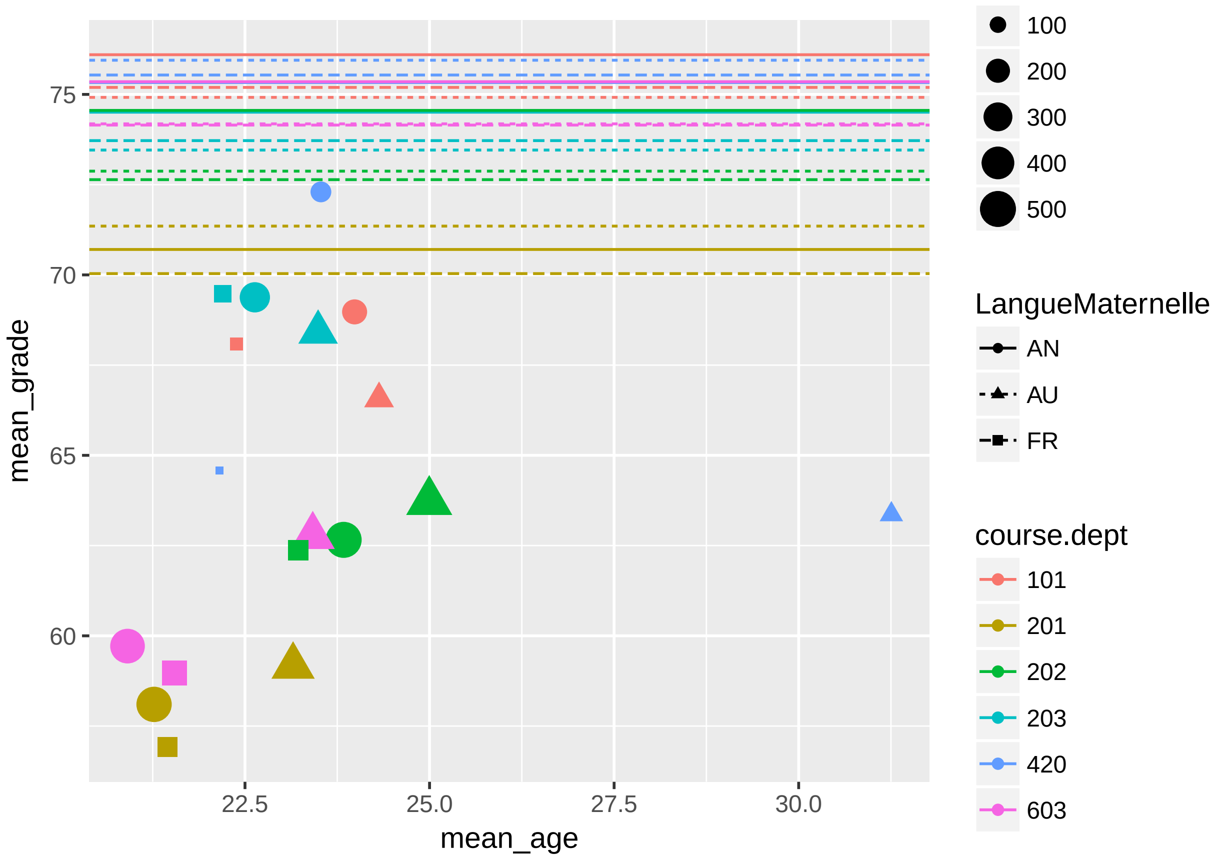
Again, what is impact of condensed summer term?



Finally, as there seems to be relatively little change in these patterns over time, we can collapse all over the past seven years,

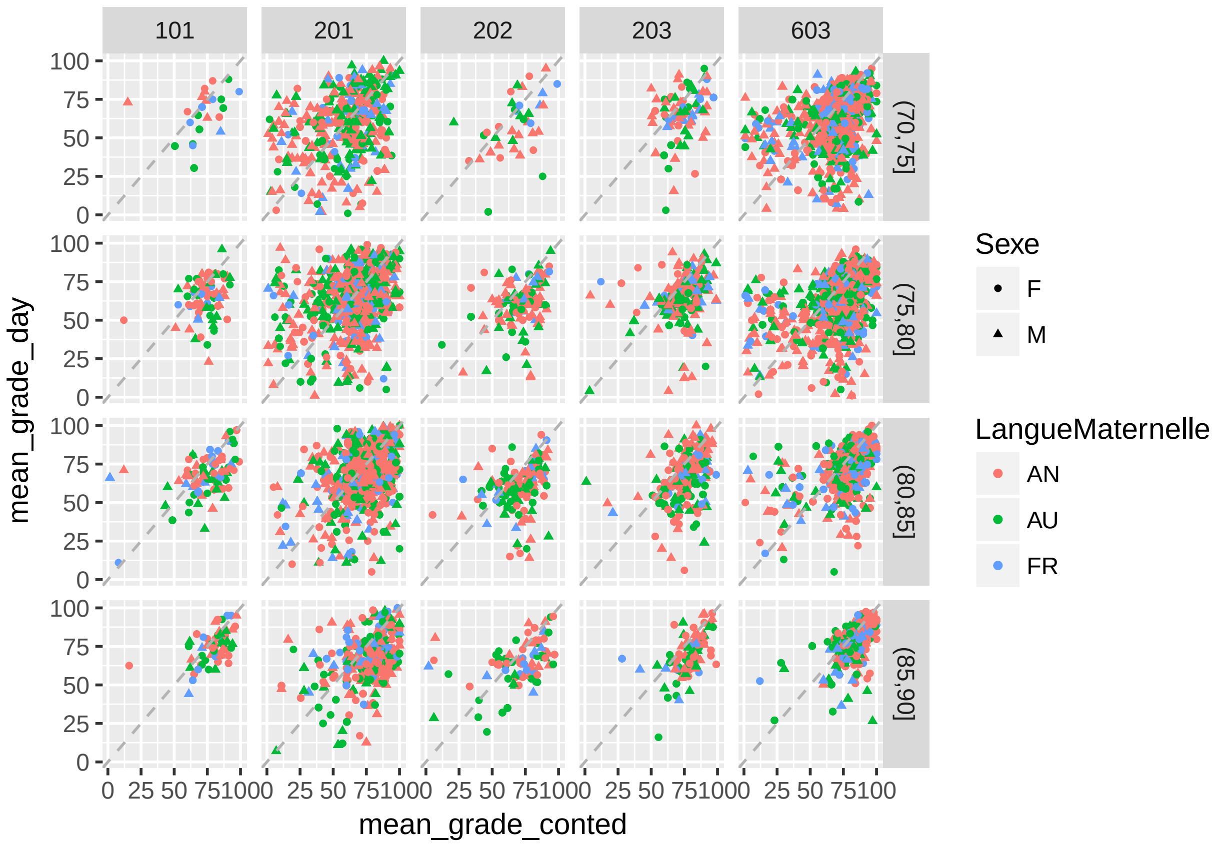


and then add on the average grade achieved by students with the same **mother tongue**, in the same courses from the same disciplines, but from the "regular" day division.



8.2.2 Success Rates

There are many students who take courses through both Continuing Education, and the regular day division over the course of their time at Dawson. One way of measuring the difference in the student experience here would be to look at those students alone, and compare the average of their grades in each division, while controlling for discipline. (the diagonal in each facet is simply meant to provide a reference for a hypothetical 1-to-1 correlation.) In order to control for student strength, we bin by overall average of high school grades.



101
201
202
203
603
(70,75]
(75,80]
(80,85]
(85,90]
mean_grade_day
mean_grade_conted
Sexe
F
M
LangueMaternelle
AN
AU
FR

Chapter 9

Final Words

We have finished a nice book.

Bibliography

Jorgensen, S., Fichten, C., and Havel, A. (2009). *Predicting the At Risk Status of College Students: Males and Students with Disabilities. Final Report Presented to PAREA, Spring 2009.*