Data Mining for Student Success and Perseverance

Sameer Bhatnagar
Jonathan Guillemette
Micheal Dugdale
Sahir Bhatnagar
Nathaniel Lasry

2017-06-01

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

This report outlines the results from a three year intercollegiate research project funded by the**PAREA** agency (*Programmme d'Aide à la Recherche en Enseignment et Apprentissage*) from the *Ministère de l'Education* of the provincial Government of Quebec.

In the province of Quebec, students finish their secondary education at what is the equivalent of Grade 11 in other parts of North America. Students are then able to attend **CEGEP** (*Collège d'enseignment général et professionel*) for either

- two years, as part of pre-university program, e.g Science, Social Science, Liberal Arts
- three years, as part of technical program, meant specifically to lead directly to the job market, e.g. Nursing, Civil Engineering Technology, Diagnotsic Imaging Technology

There are 48 CEGEPs in the Quebec network, and public or private, they all fall under the purview of the *Ministère de l'Education et Enseignment Superieur*. Over the past twenty years, there has been significant work(Jorgensen et al., 2003, 2005, 2009; Rivière, 1995; Shaienks et al., 2008) and media (Breton, 2016; Dion-Viens, 2015; Duchaine, 2017) on the topic of student attrition in CEGEP. The scholarly work done has often focused on determing predictors of attrition through surveys, or focused on specific vulnerable sub populations. The media has often reported on government figures, which rely on data that looks at information at a very coarse level of granularity (of students graduated from high schoolhow many obtain diplomas from CEGEP)

Almost all of the CEGEP's use the same database system, known as **CLARA** in order to manage the data related to student admission, registration and graduation. Our research team's main objective is to leverage this uniformity of how data is automatically generated and stored, in order to determine if, in this wealth of data, there might be predictors of student attrition. This effort stands apart from previous work and reports in that - the data analyzed is much finer-grained: the unit of analysis is down to the semester registration records for each student - we look at the general population of students - with every additional college that participates in our study, we can track students who change CEGEPs, and be more accurate in our determinations of which students drop out

Chapter 3

Literature

3.1 Modeling success and attrition in CEGEP

The most important relevant work for this project is (**?**).

3.2 Predictive Modelling in Learning Analytics and Educational Data Mining

(Lang et al., 2017)

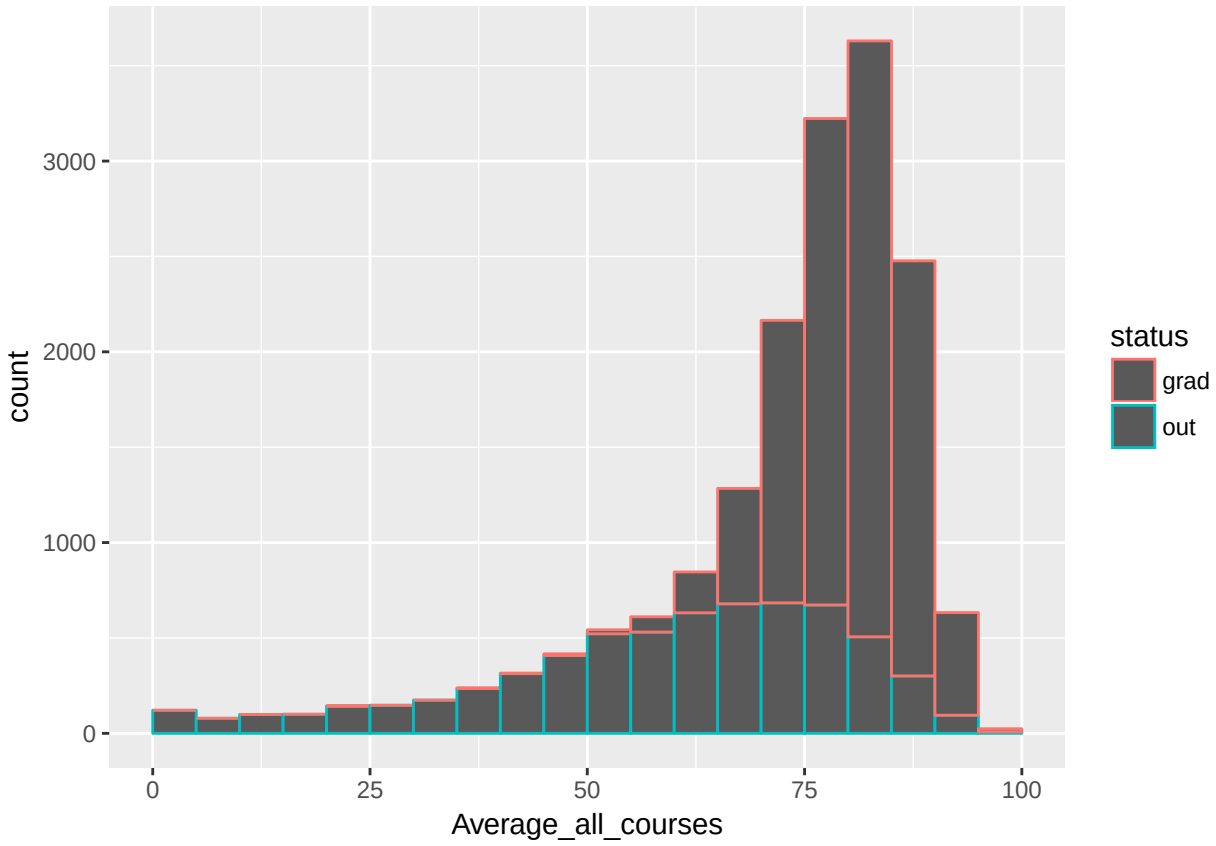
Chapter 4

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

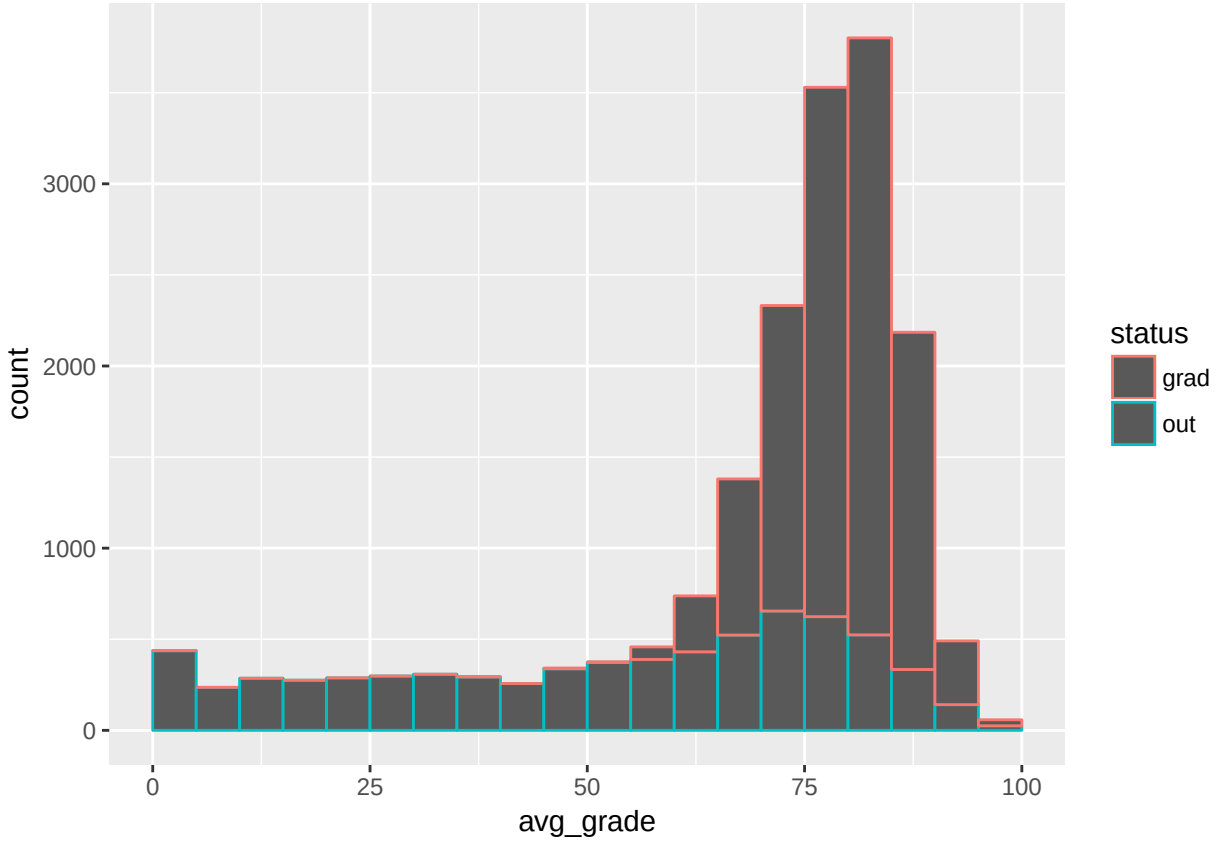
Do students who drop out do so because of poor grades? What fraction of students are counted year after year as drop-outs and labeled as problems to be solved by the system while being exemplary students in terms of academic performance. Armed with this dataset, we can get the answer to that question. Let us begin by looking at the average grades of students who eventually dropped out compared to grades of students who haven't. The following set of graphs will look at that comparison for 3 different semesters: the semester in which they dropped out (or graduated), the semester before that and the one before that. Let us see what the data says.




```
##
##  Welch Two Sample t-test
##
## data:  Average_all_courses by status
## t = 78.251, df = 7349.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  20.36923 21.41600
## sample estimates:
## mean in group grad  mean in group out
##           79.97370           59.08109
```

From the average total grade between the drop outs and the graduates, we can clearly see that the distributions are significantly different, but what is surprising is that 22% of students have an average grade above 75% and that more than 52% of the students who dropped out had an average grade above 60%. In other words, most students who drop out have passing averages. One hypothesis for this effect is that students who end up dropping out start with good semesters and have their performance decline closer to the final semester. Let us verify this hypothesis.

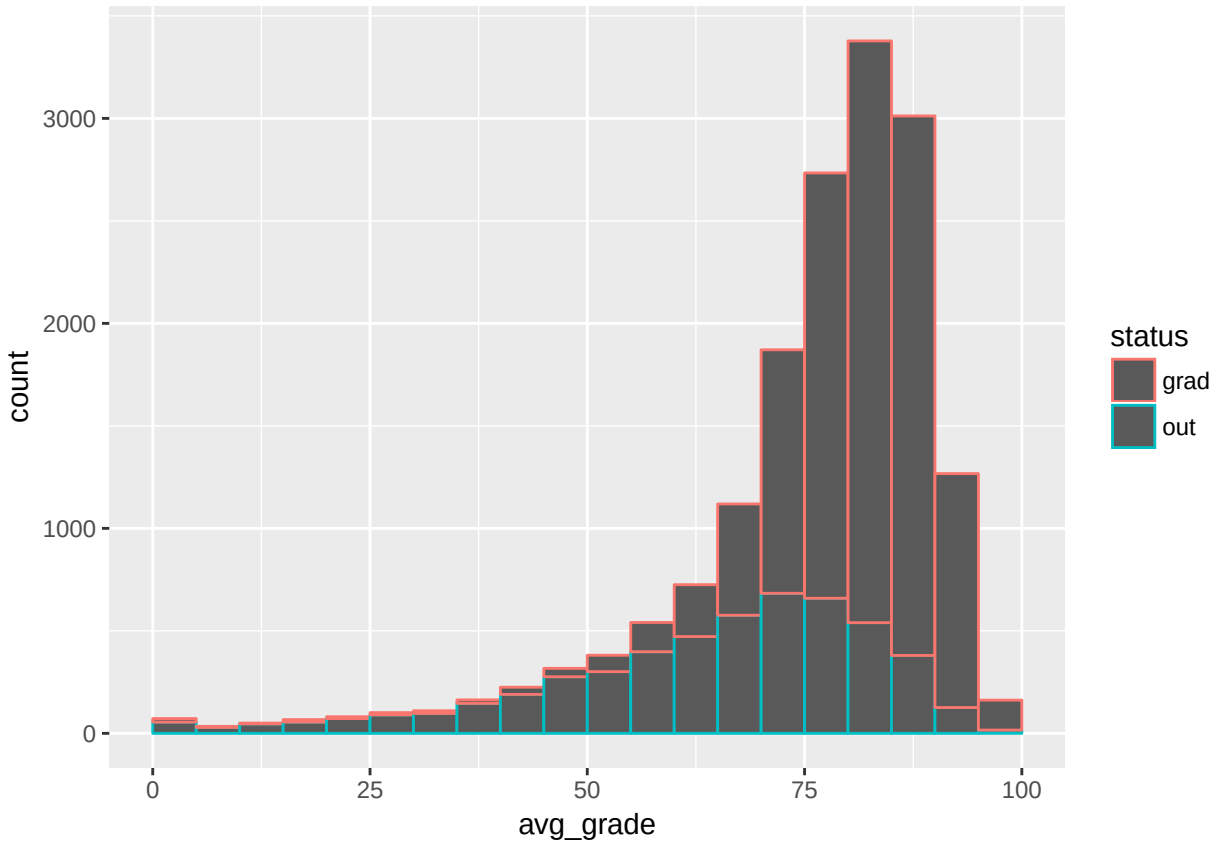
Let us now look at the performance of drop outs vs graduates on a semester by semester basis. Let us start by looking at the average grades of students in their last semester during which they are either graduating or dropping-out.




```
##
##  Welch Two Sample t-test
##
## data:  avg_grade by status
## t = 84.649, df = 7619.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  27.00250 28.28279
## sample estimates:
## mean in group grad  mean in group out
##           79.17656           51.53391
```

In this data, we can clearly observe a stark difference between the graduates and the drop outs. First of all, note that 23% of students still have a term average over 75% in the semester in which they drop out. To push it further 15% of students who drop out have an average grade over 80%. The data clearly suggests that some of the drop outs aren't dropping out because of academic performance. Furthermore, the long tail of the data on the low end of performance suggests that some of the students stopped coming to class prior to the end of the semester resulting in very low grades that serve to drive their cegep average grades from the graph above even lower. 34% of students have a grade below 40 suggesting that they have indeed stopped coming to school some time during the semester. What if these students hadn't stopped coming, would their performance be similar to that of graduates?

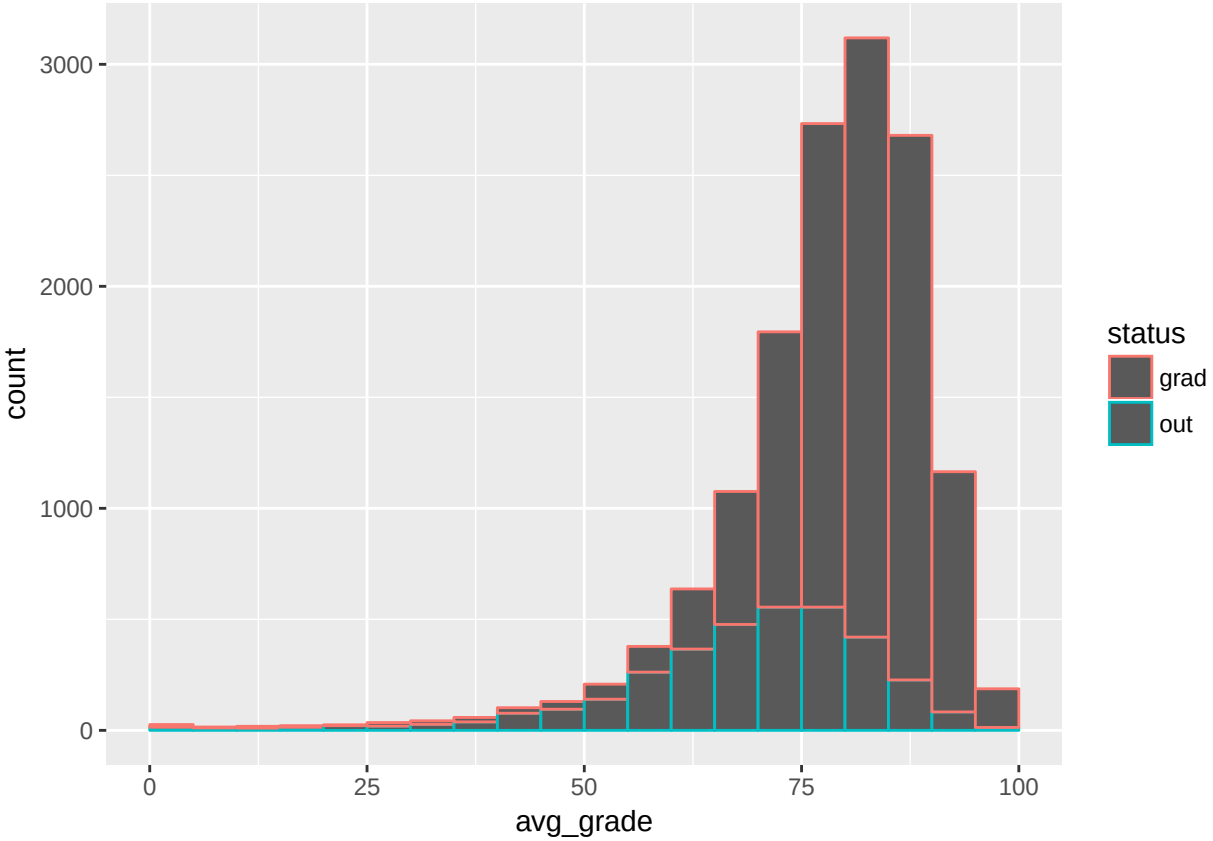
Let us now turn our attention to the semester before the one where they graduate or drop out.




```
##
##  Welch Two Sample t-test
##
## data:  avg_grade by status
## t = 57.702, df = 6604.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  15.65690 16.75814
## sample estimates:
## mean in group grad  mean in group out
##           80.56981           64.36230
```

The data clearly shows that even if there are significant differences between the groups, the drop out student population is getting closer to the graduate population. A section of 33% is observed to have an average grade above 75%.

Finally, let us look at 2 semesters before they graduate or drop-out. We are again expecting the same trend.




```
##
##  Welch Two Sample t-test
##
## data:  avg_grade by status
## t = 42.166, df = 4390.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  10.95430 12.02262
## sample estimates:
## mean in group grad  mean in group out
##           80.42574           68.93729
```

In conclusion, the data strongly suggests that there are approximately 20% of students who consistently have averages above 75%, but still drop out. Therefore, that section of the drop out population is a section that will always go undetected if only traditional academic failure metrics are used to assess which students are likely to drop out. Students who move out to the US or the rest of Canada after their second semester and those who drop out by lack of interest are two potential student types that will always drop out no matter what remedial solutions are offered for them.

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

Over the last twenty years, there has an increasing amount of work in the applied social sciences that explore the use of what (Breiman et al., 2001) refers to as "algorithmic modelling"", as opposed to"data modelling". He describes these as two cultures, the former being made up of mostly computer scientists, and the latter being made up statisticians (the methods therein are the ones explored in the previous chapter of this report). The key metric in such classical methods are **goodness of fit**, and such"explanatory"" modelling aims to find associative and causal relationships between predictors. Meanwhile, "algorithmic"", or"predictive"" methods emphasize determining any function that maps input variables to output responses, with less regard for a probabilitic framework that allows for causation, focusing solely on emprirical precision(Shmueli et al., 2010). In the recently published **Handbok of Learning Analytics**(Lang et al., 2017), published by the Society of Learning Analytics Research, (Bergner, 2017) asserts that the researchers looking into educational data stand to gain from understanding the nuances of both methodologies, as previous work has shown the strengths and weakeness of either in this domain.

The previous chapter explored how classical statistical models can be built and used to determine what are the factors that influence dropout. This is useful for policy makers and admninstrators who want to dedicate resources in the most strategic places. However this chapter will explore models whose inner workings are less interpretable, but whose primary objective is prediction/identification of at-risk students. This is useful in the context where college administration has some blanket intervention that it would like to apply, and we just want to ensure that the students most in need are reached. Despite the less clear interpratability of factors in these *predictive* models (as compared to the *explanatory* models in the previous chapter), we will stil explore methods to "open up the black box", and determine which features are most important in achieving both accrate and sensitive prediction.

6.1 Decision Trees and Random Forests

6.2 Neural Networks

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

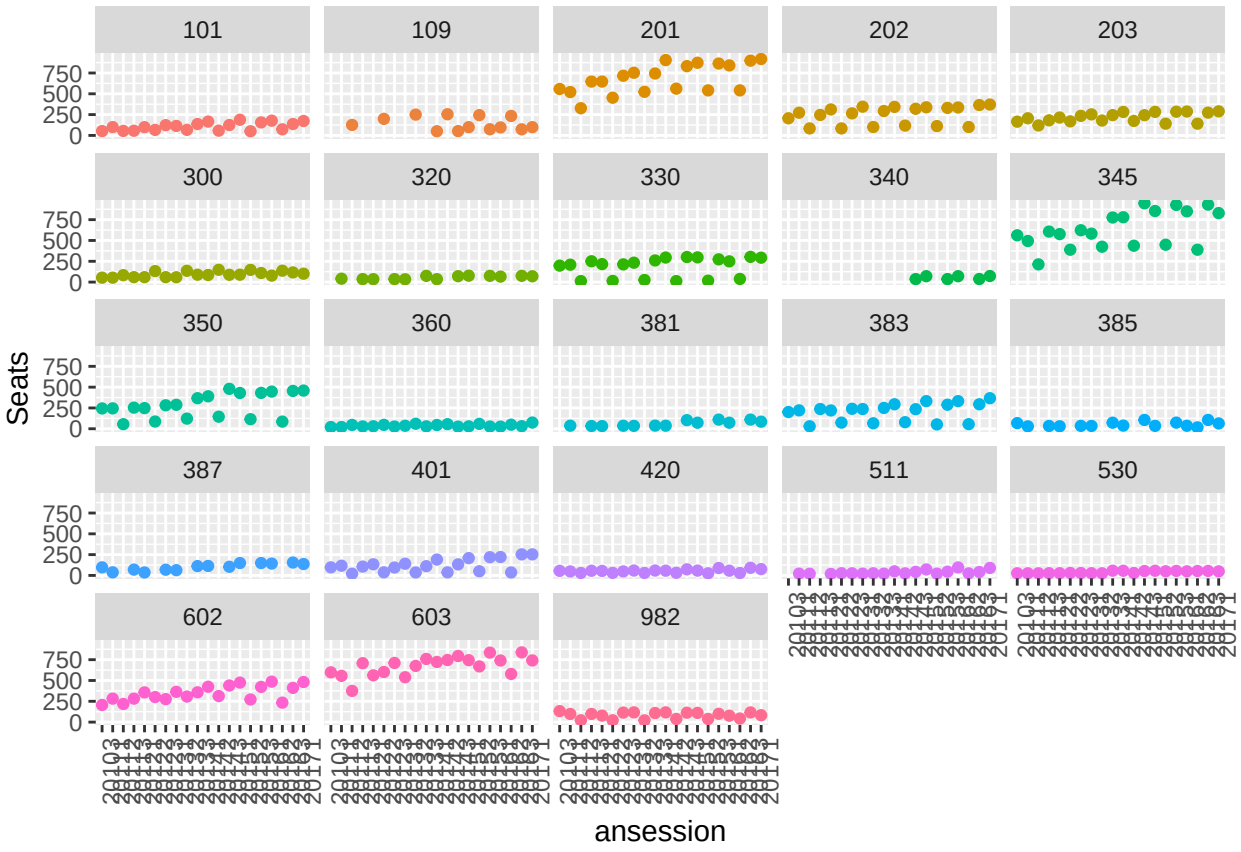
The department of Continuing at Education offers evening courses across many disciplines throughout the year. The purpose of this report is to give a broad overview of the demographics and sector-level metrics for students taking courses within Cont Ed.

8.1 Services

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

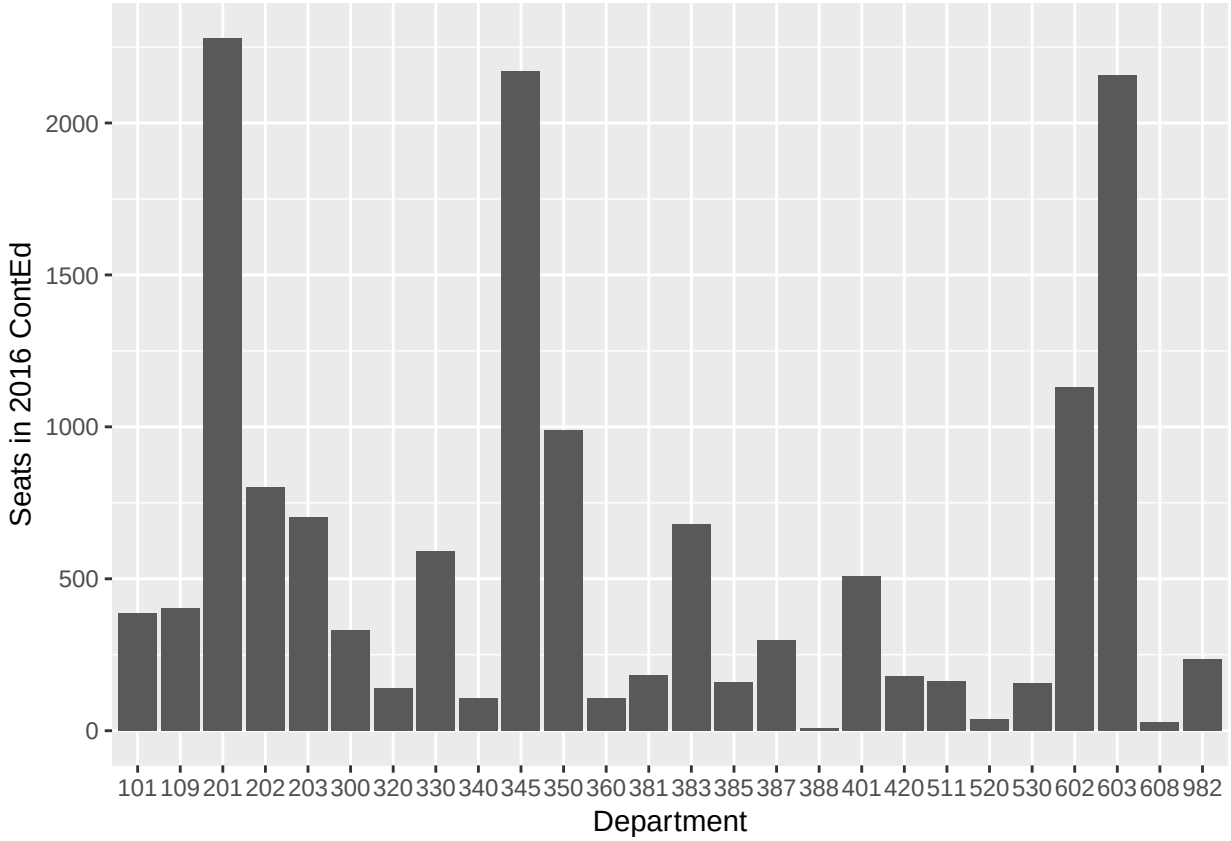
8.1.1 Continuing Education's Growth over Time

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



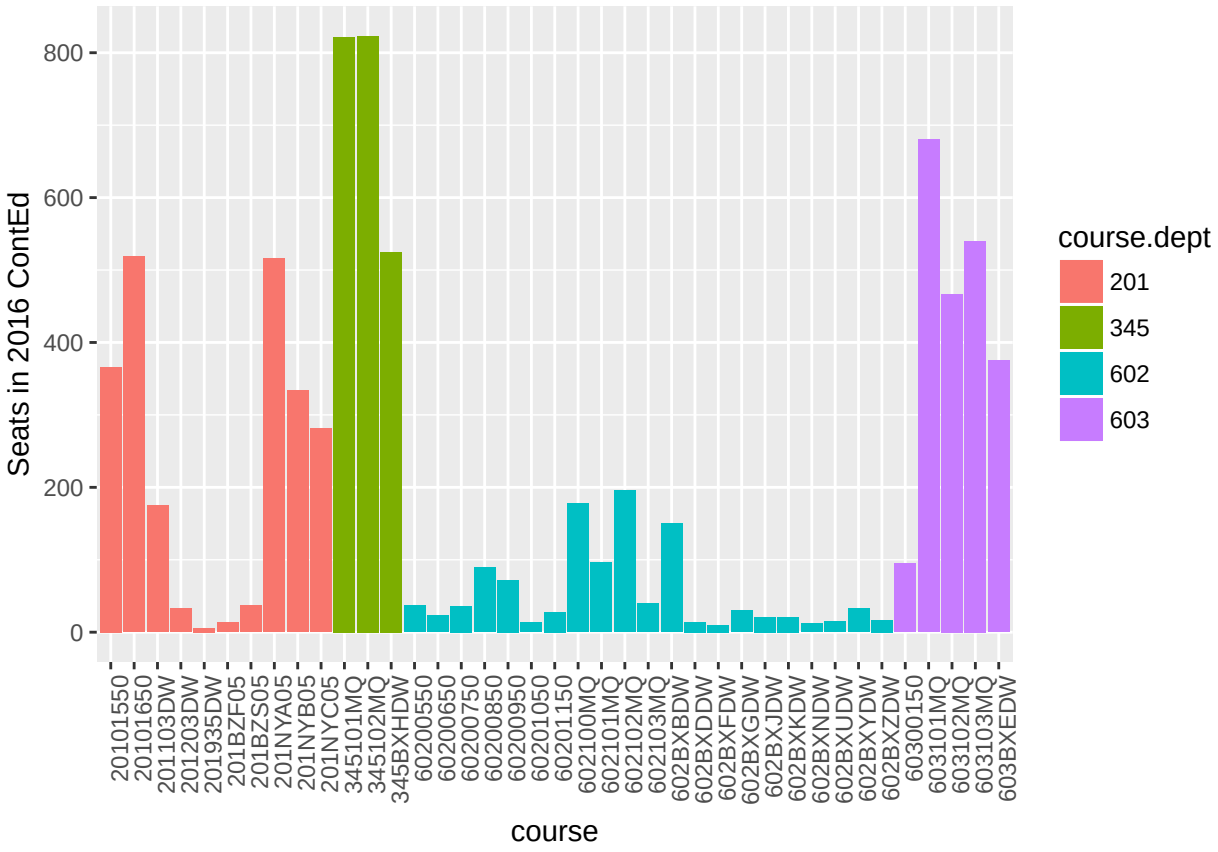
8.1.2 Department Level Seat Distributions

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?

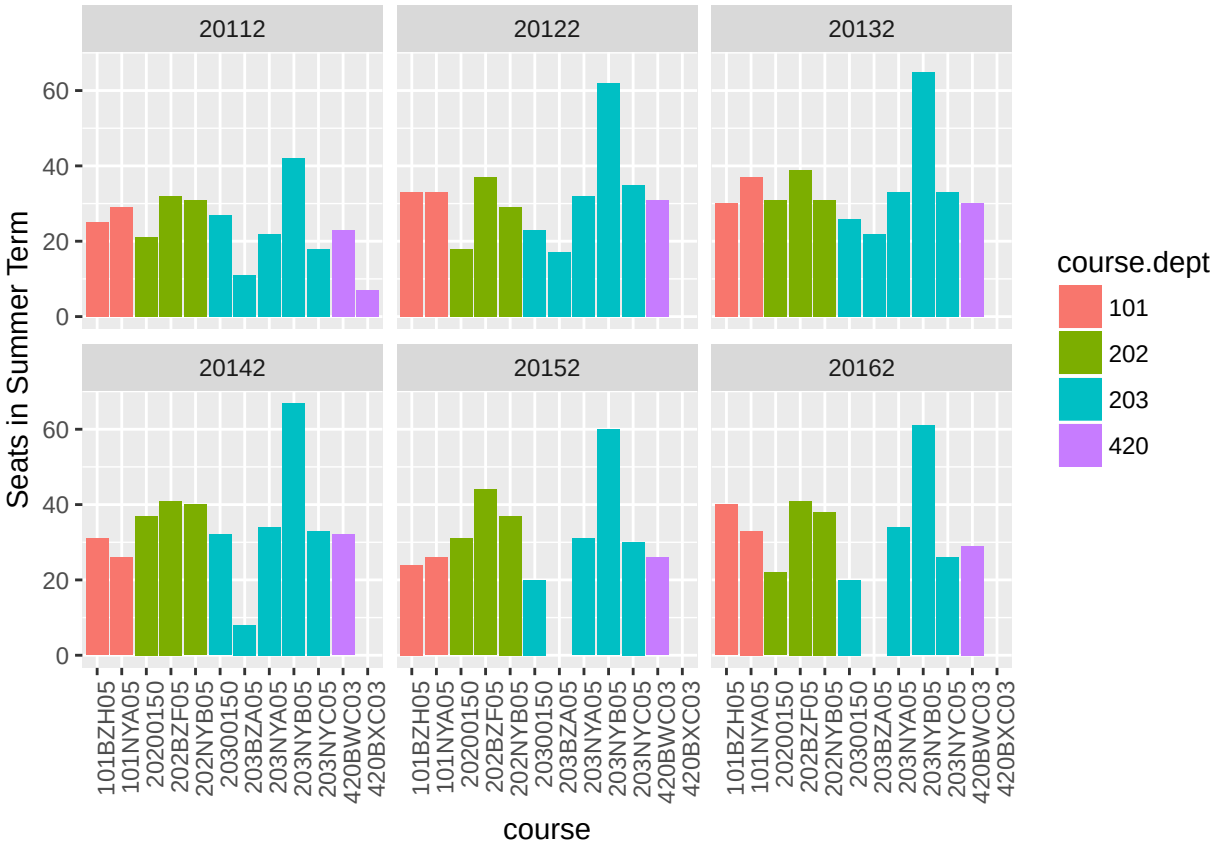
8.1.4.1 Winter


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

8.1.4.2 Fall


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

8.1.4.3 Summer



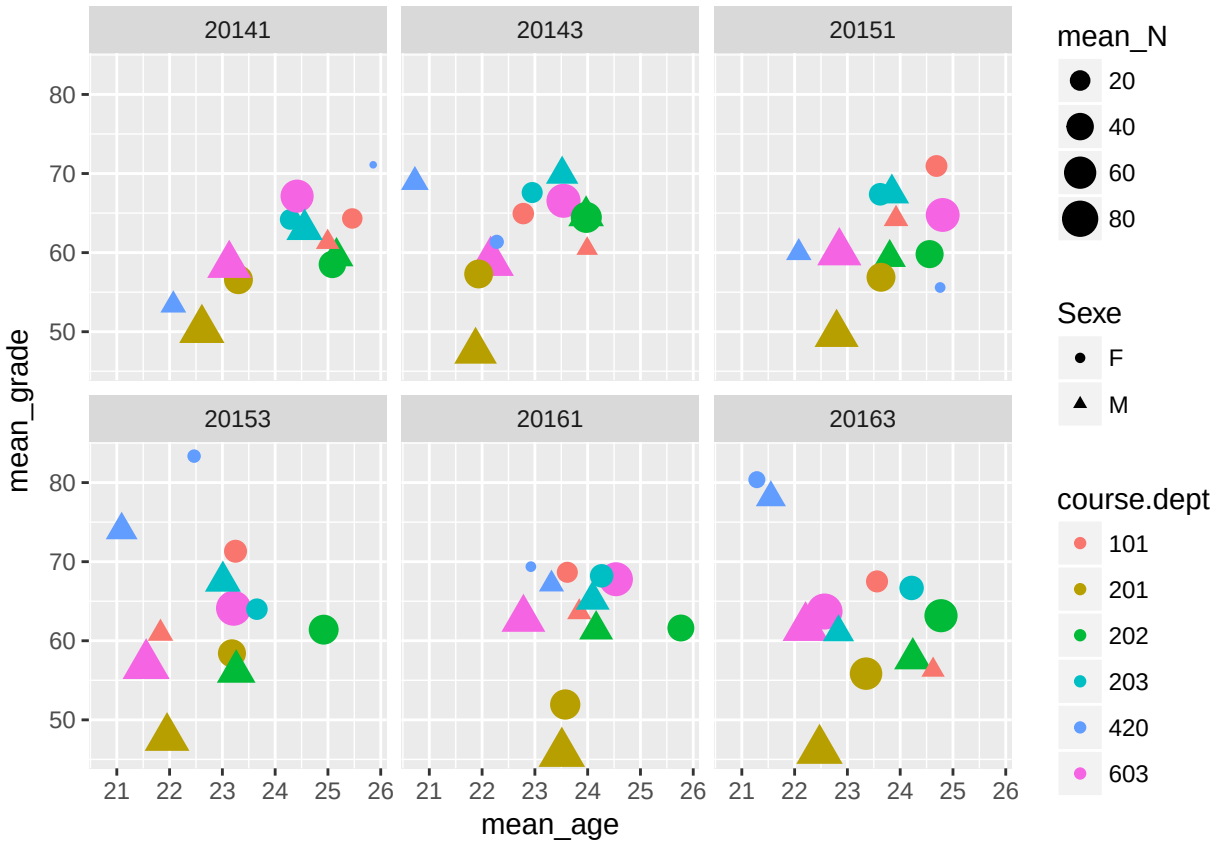
8.2 Students

8.2.1 Demographics

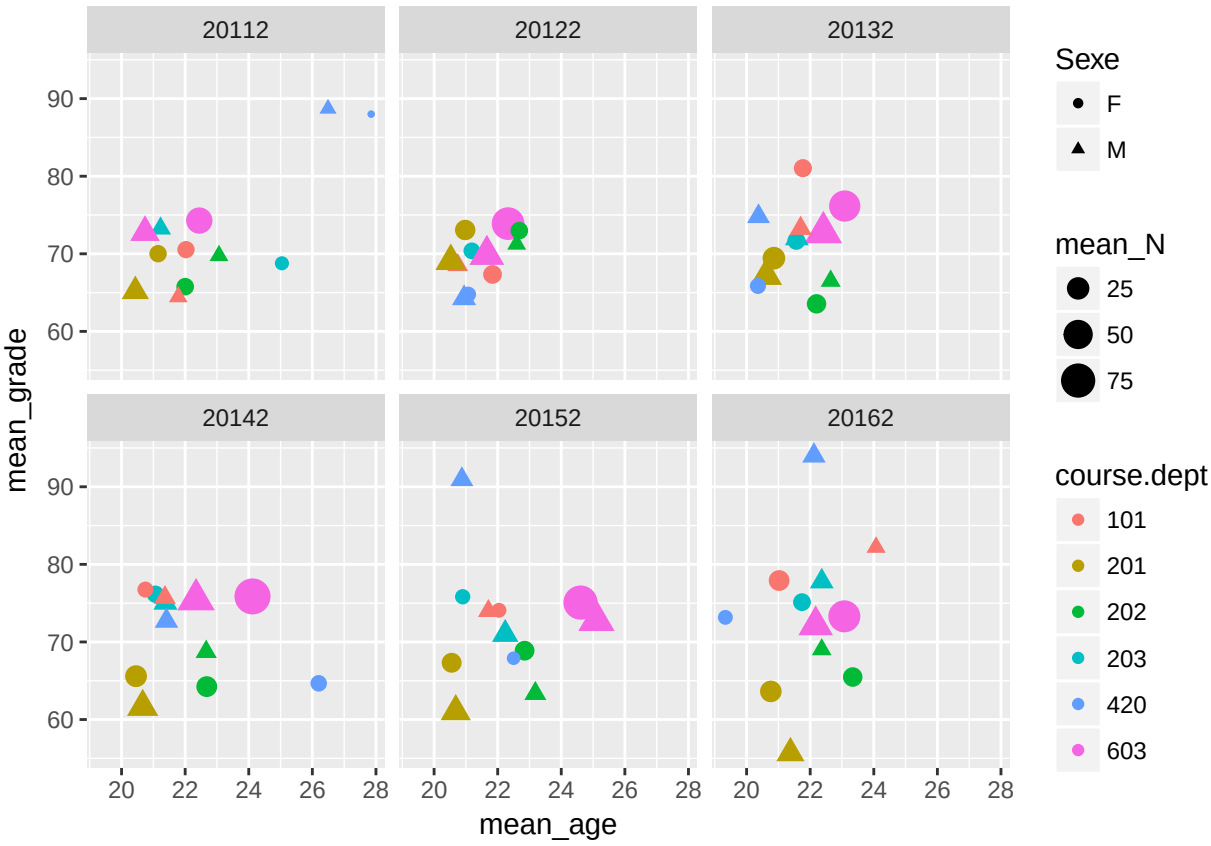
Who are the students using continuing education services?

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

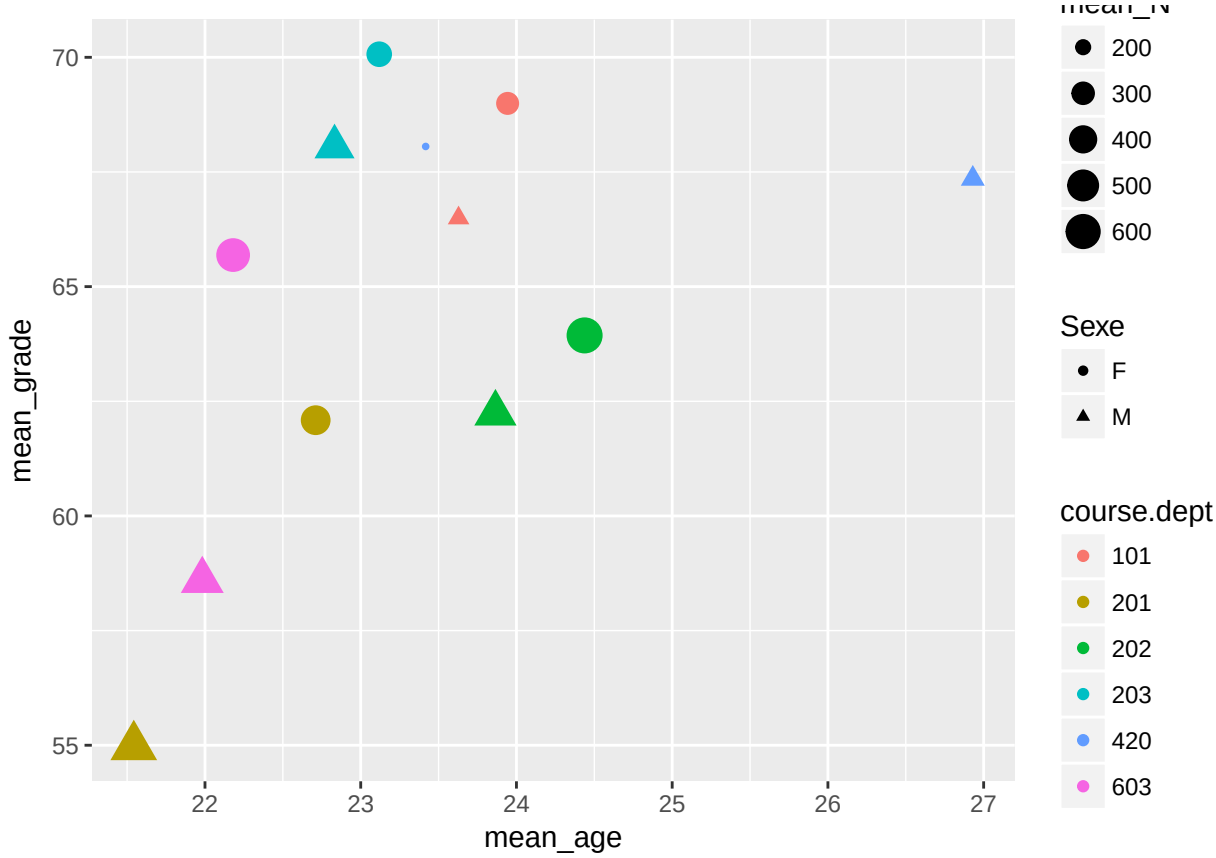
8.2.1.1 Sexe



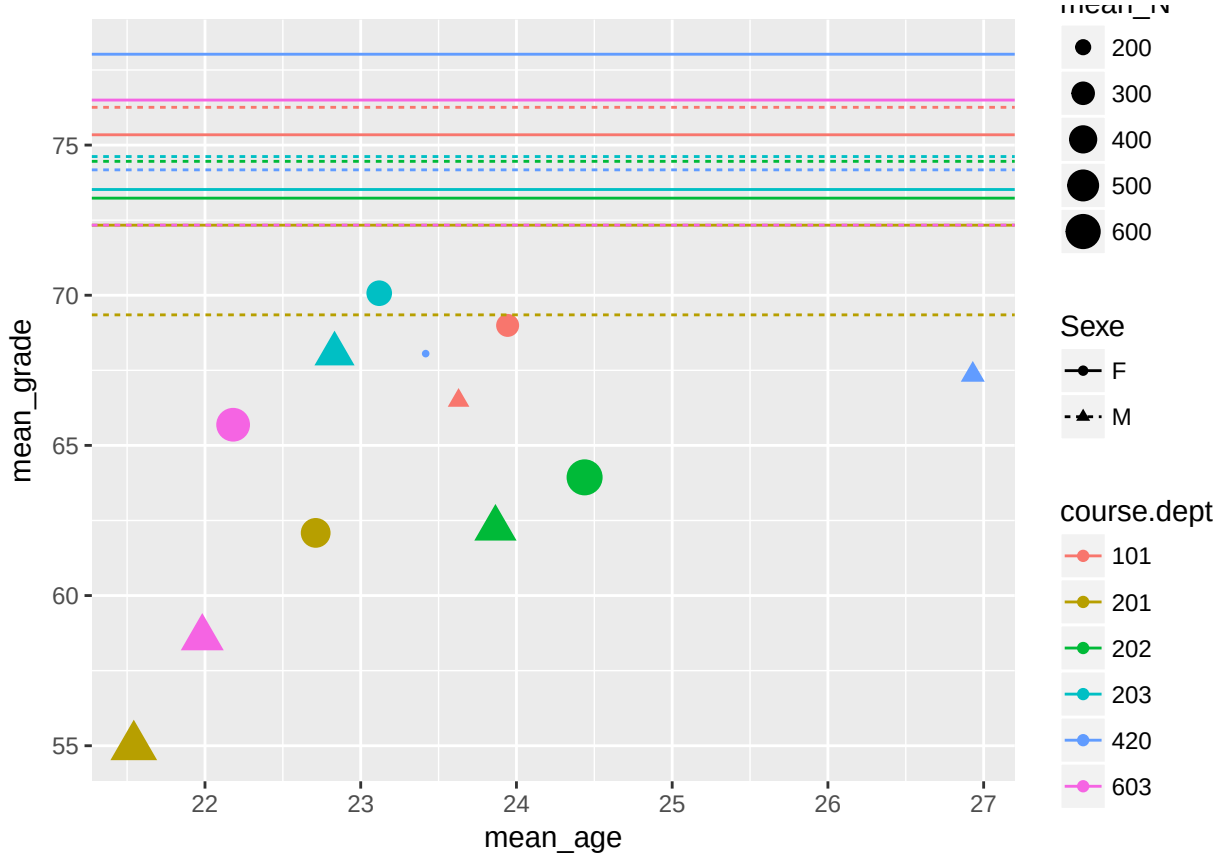
What is impact of condensed summer term?



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



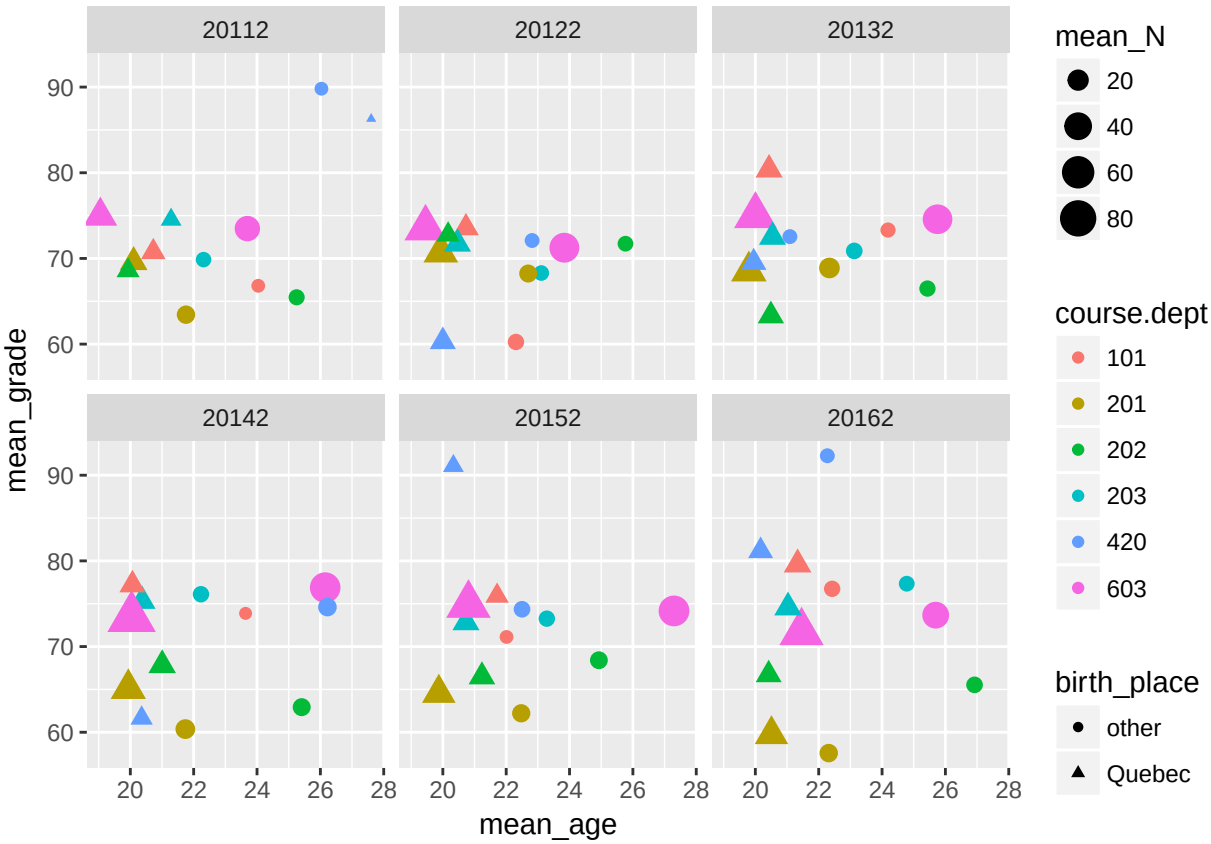
and then add on the average grade achieved by students with the same gender, in the same courses from the same disciplines, but from the "regular" day division.



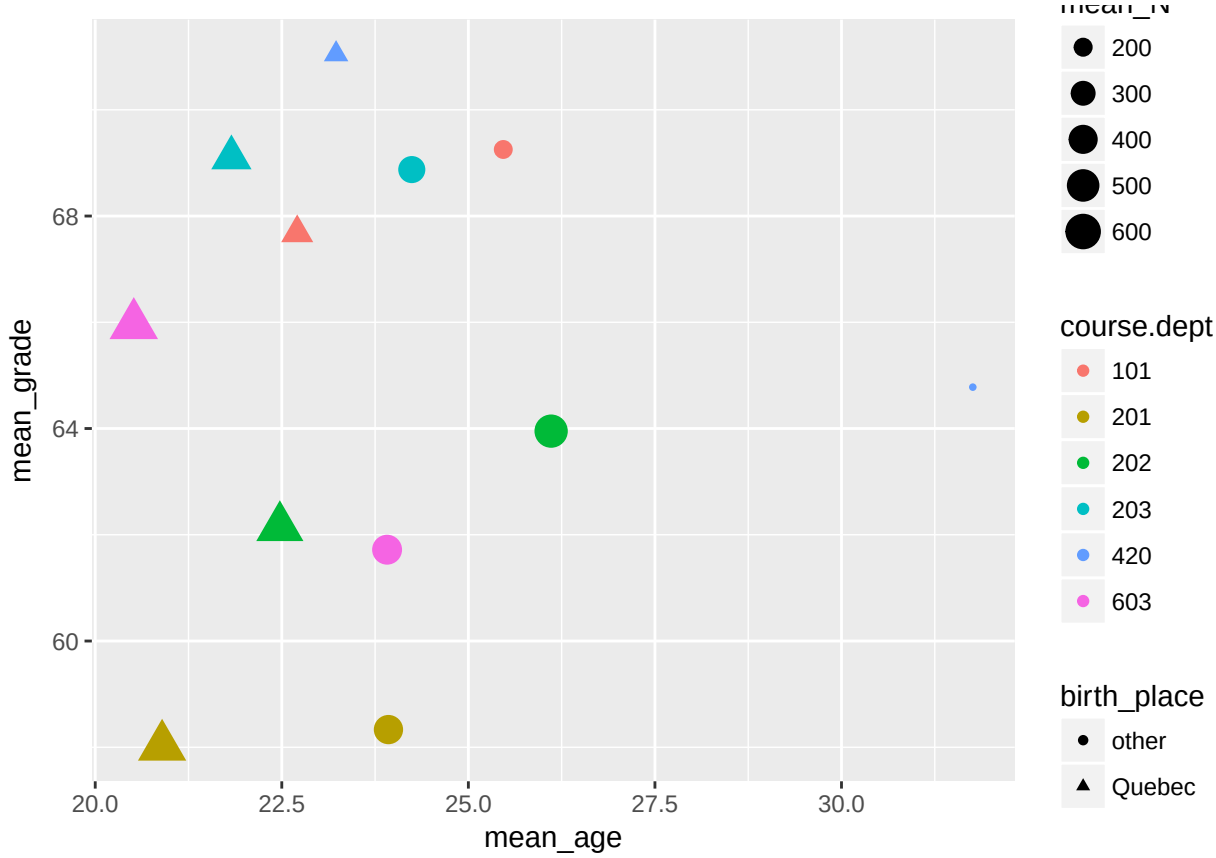
8.2.1.2 Birth Place

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?

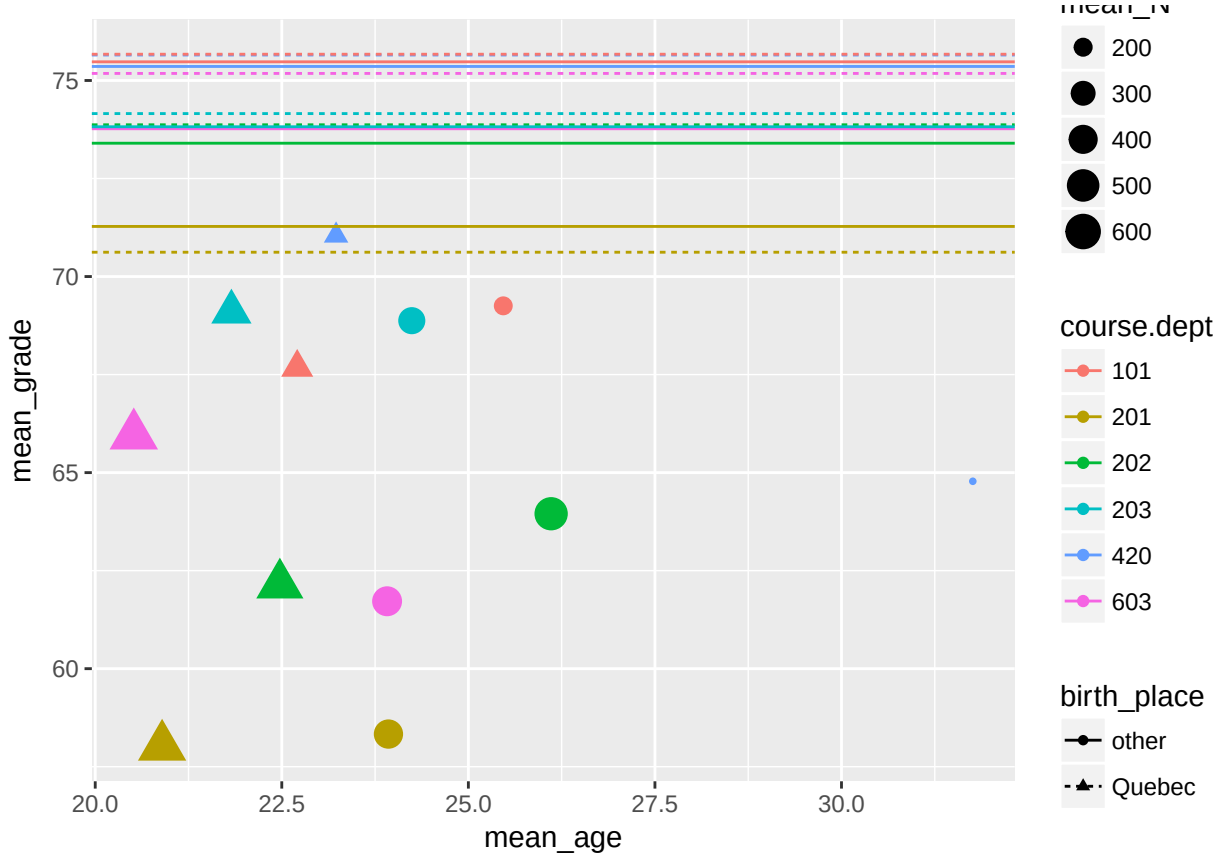
Again, what is impact of condensed summer term?



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

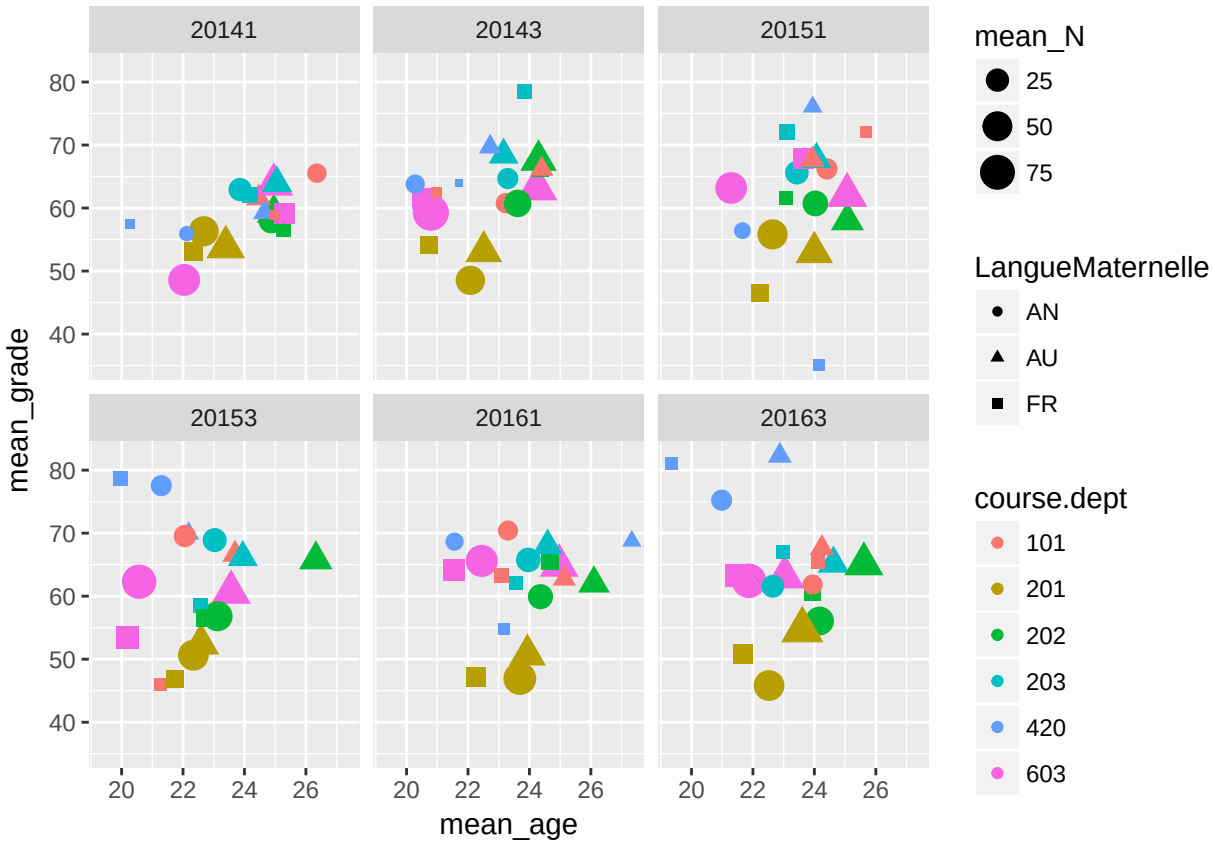


and then add on the average grade achieved by students with the same **birth place** (namely quebec residents vs *other*), in the same courses from the same disciplines, but from the "regular" day division.

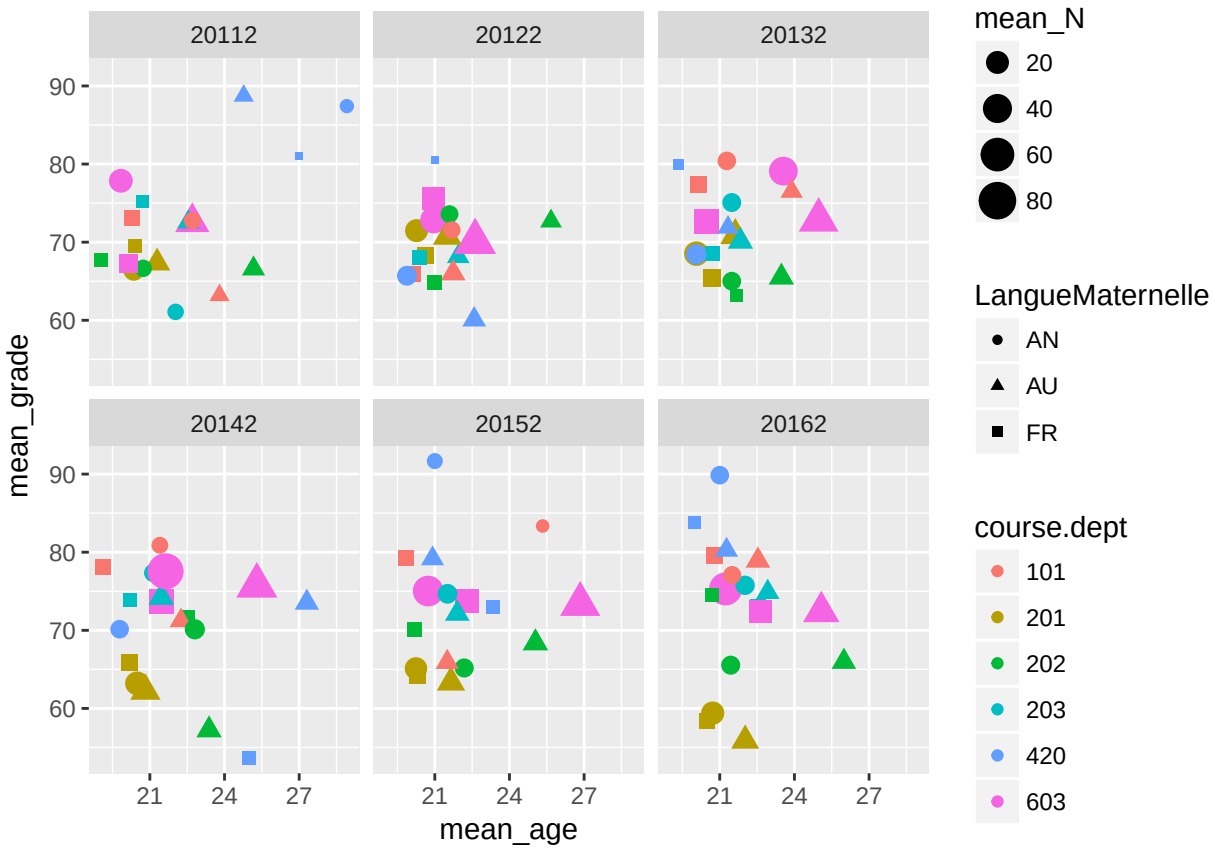


8.2.1.3 Mother Tongue

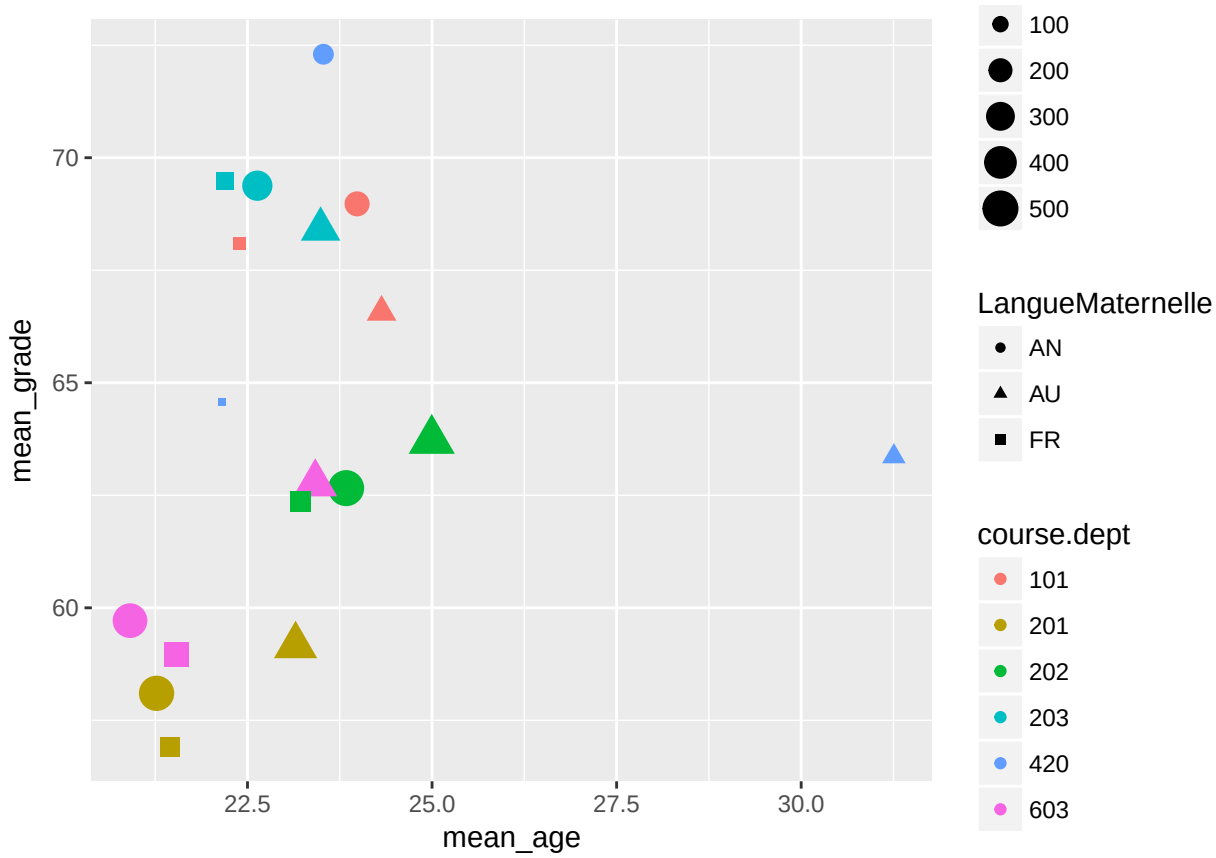
Finally we look at the possible impact of Mother Tongue. How do anglophones, francophones, and allophones compare in Continuing Education?



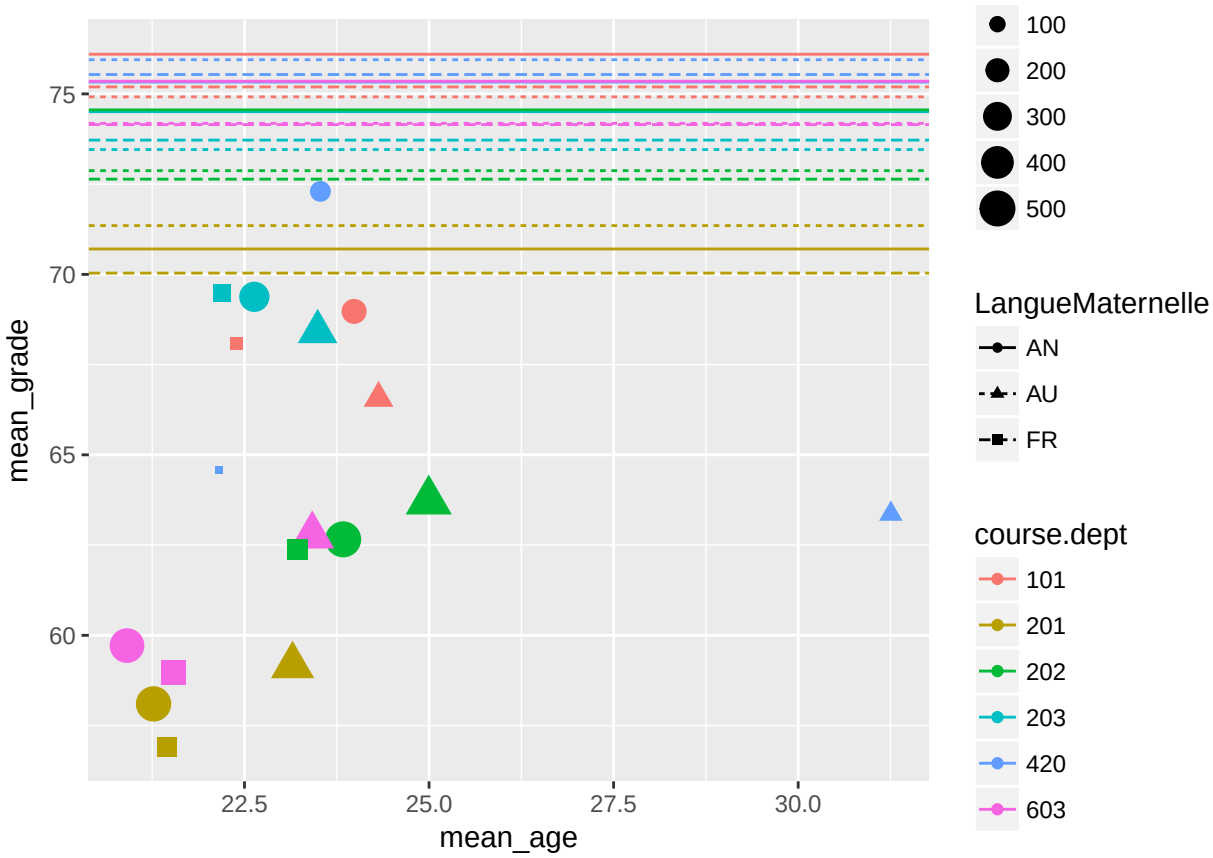
Again, what is impact of condensed summer term?



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

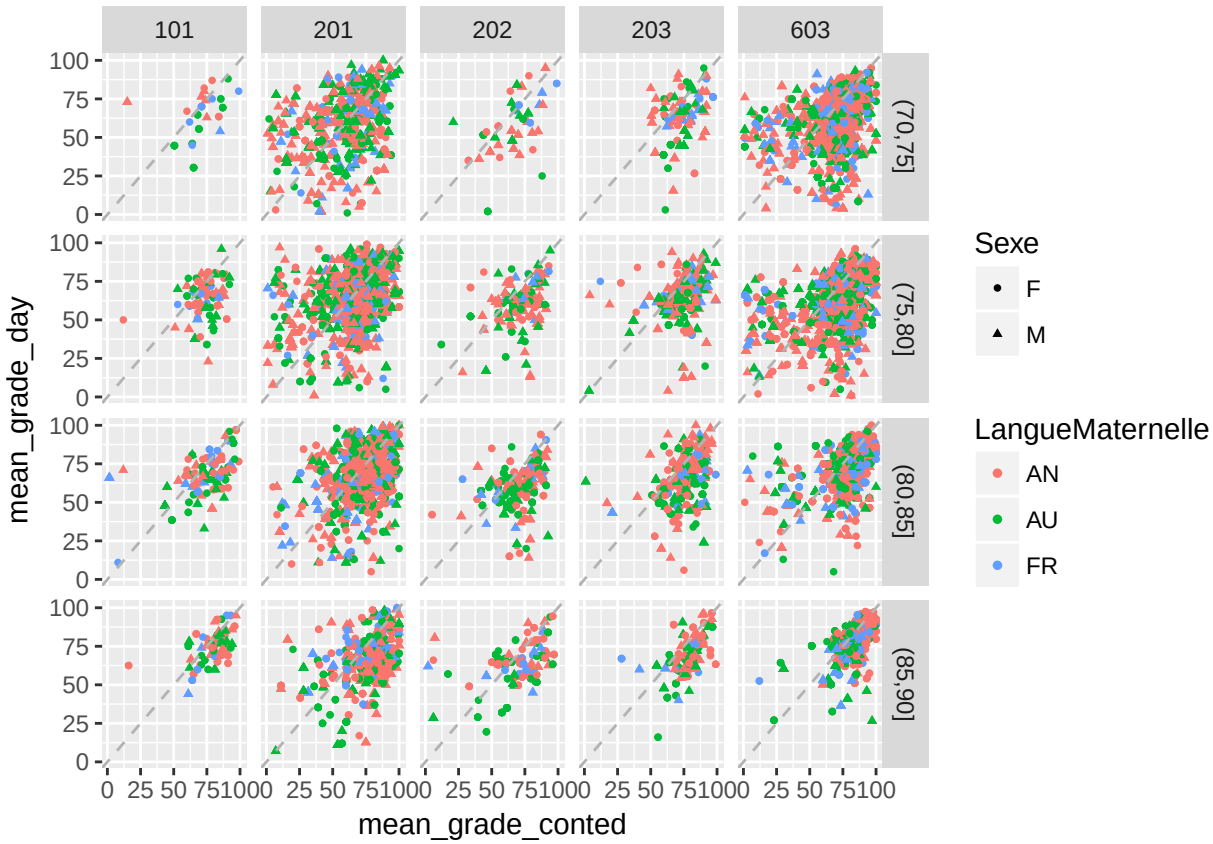


and then add on the average grade achieved by students with the same **mother tongue**, in the same courses from the same disciplines, but from the "regular" day division.



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

Bergner, Y. (2017). Measurement and its Uses in Learning Analytics. In Lang, C., Siemens, G., Wise, A. F., and Gaševic, D., editors, *The Handbook of Learning Analytics*, pages 34–48. Society for Learning Analytics Research (SoLAR), Alberta, Canada, 1 edition.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Breton, B. (2016). Décrocher du cégep et de l'université. *La Presse*.

Dion-Viens, D. (2015). Cégeps: à peine 31% des étudiants obtiennent un diplôme dans les délais.

Duchaine, G. (2017). Qu'est-ce qui cloche au cégep ? - La Presse+.

Jorgensen, S., Ferraro, V., Fichten, C., and Havel, A. (2009). Predicting college retention and dropout: Sex and disability. *Online Submission*.

Jorgensen, S., Fichten, C., Havel, A., Lamb, D., James, C., and Barile, M. (2003). *Students with Disabilities at Dawson College: Success and Outcomes. Final Report Presented to PAREA, Spring 2003.* ERIC.

Jorgensen, S., Fichten, C. S., Havel, A., Lamb, D., James, C., and Barile, M. (2005). Academic performance of college students with and without disabilities: An archival study. *Canadian Journal of Counselling*, 39(2):101.

Lang, C., Siemens, G., Wise, A., and Gasevic, D. (2017). *Handbook of learning analytics*. SOLAR.

Rivière, B. (1995). Comprendre les décrocheurs afin de mieux les aider. *Pédagogie collégiale*, 9(2):11–15.

Shaienks, D., Gluszynski, T., and Bayard, J. (2008). *Les études postsecondaires, participation et décrochage: différences entre l'université, le collège et les autres types d'établissements postsecondaires*. Statistique Canada.

Shmueli, G. et al. (2010). To explain or to predict? *Statistical science*, 25(3):289–310.