Exploring Student Records

Paul Murrell

The University of Auckland

August 13 2009

Overview

Motivation:	STATS 220
Problem statement:	What background do the students have? Does background dictate performance?
Preconceptions:	A Comp Sci group Comp Sci perform better
Some answers:	Preparing the data Visualizing the data

The raw data

Originally an Excel file (plus file describing variables).

ID The student ID number.

Name The student name.

Term The term in which the student took the paper (e.g., semester 1, 2008), but as a code (e.g., 1083).

Subject The paper subject, as a letter code, e.g., `STATS`, or for older papers it could be a subject number, e.g., 475.

Catalog The paper number, e.g., 220.

Acad Prog The academic program of the student (at the time the paper was taken), e.g., `BA` or `BSC`.

Grade The grade for the paper, e.g., `A`, or blank if currently enrolled.

Cumulative GPA The student's GPA (per semester).

The raw data

These variables are also in the report, but I did not use them.

Status Whether the student is currently enrolled in STATS 220 (`E`). A couple of values are, worryingly, blank, but I don't know what that means yet.

Points The number of points for the paper.

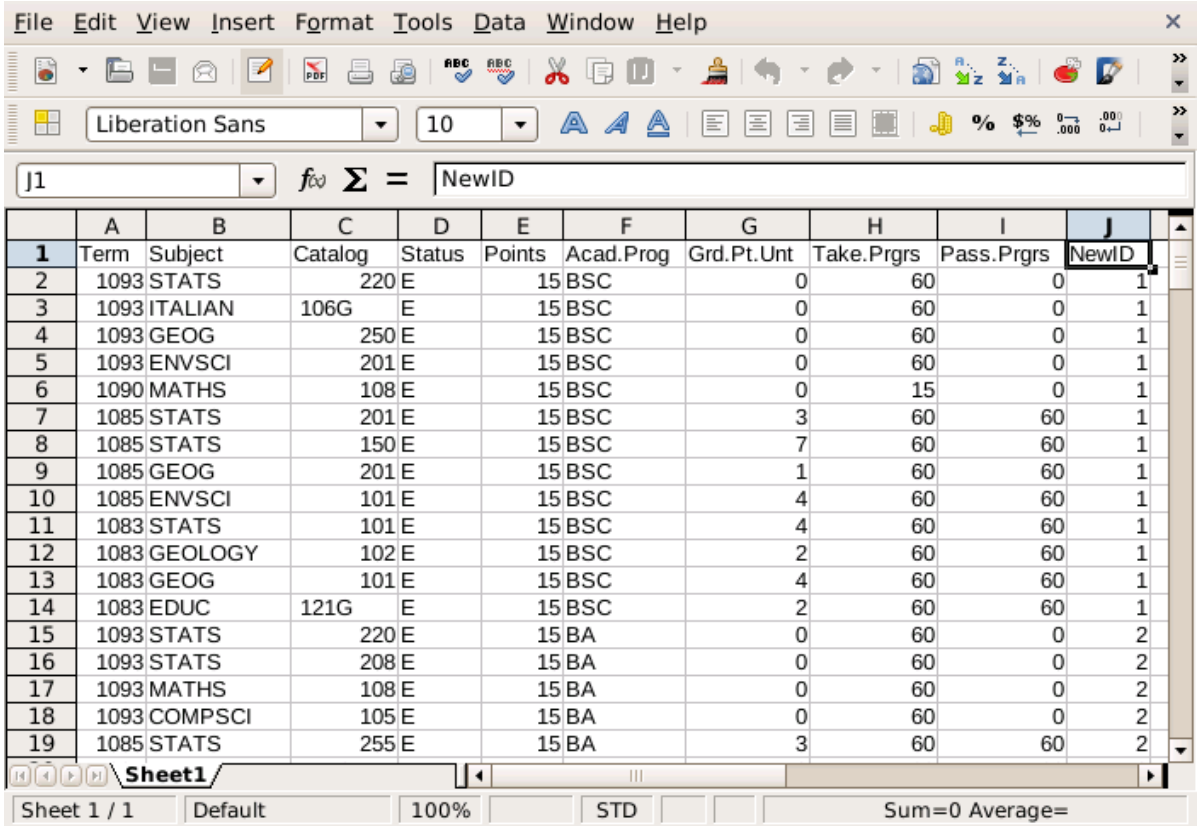
Grd Pt/Unt The grade point contribution of this paper.

Take Prgrs Points achieved in a semester.

Pass Prgrs Points achieved in a semester (not sure how this differs from previous, but I did not use these two anyway).

The raw data

Exported the Excel file as CSV, removed names, grades, and GPA, and replaced `ID` with `NewID` which provides anonymous unique identifier.



File Edit View Insert Format Tools Data Window Help

Liberation Sans 10

J1 NewID

	A	B	C	D	E	F	G	H	I	J
1	Term	Subject	Catalog	Status	Points	Acad.Prog	Grd.Pt.Unt	Take.Prgrs	Pass.Prgrs	NewID
2	1093	STATS	220	E	15	BSC	0	60	0	1
3	1093	ITALIAN	106G	E	15	BSC	0	60	0	1
4	1093	GEOG	250	E	15	BSC	0	60	0	1
5	1093	ENVSCI	201	E	15	BSC	0	60	0	1
6	1090	MATHS	108	E	15	BSC	0	15	0	1
7	1085	STATS	201	E	15	BSC	3	60	60	1
8	1085	STATS	150	E	15	BSC	7	60	60	1
9	1085	GEOG	201	E	15	BSC	1	60	60	1
10	1085	ENVSCI	101	E	15	BSC	4	60	60	1
11	1083	STATS	101	E	15	BSC	4	60	60	1
12	1083	GEOLOGY	102	E	15	BSC	2	60	60	1
13	1083	GEOG	101	E	15	BSC	4	60	60	1
14	1083	EDUC	121G	E	15	BSC	2	60	60	1
15	1093	STATS	220	E	15	BA	0	60	0	2
16	1093	STATS	208	E	15	BA	0	60	0	2
17	1093	MATHS	108	E	15	BA	0	60	0	2
18	1093	COMPSCI	105	E	15	BA	0	60	0	2
19	1085	STATS	255	E	15	BA	3	60	60	2

Sheet1

Sheet 1 / 1 Default 100% STD Sum=0 Average=

Data preparation

Generate new variable `Current` to indicate whether the paper is currently being taken (`!Current` gives papers that the student has taken in the past).

```
> classData <- read.csv("Data/transcripts-blind.csv",
+                       stringsAsFactors=FALSE,
+                       strip.white=TRUE)

> classData$Current <- classData$Term > 1090
```

Data preparation

```
> head(classData[c("NewID", "Term", "Current")], 10)

   NewID Term Current
1      1 1093    TRUE
2      1 1093    TRUE
3      1 1093    TRUE
4      1 1093    TRUE
5      1 1090   FALSE
6      1 1085   FALSE
7      1 1085   FALSE
8      1 1085   FALSE
9      1 1085   FALSE
10     1 1083   FALSE
```

Data preparation

Some older papers have a numeric Subject.

```
> head(classData[!is.na(as.numeric(classData$Subject)),
+                c("NewID", "Term", "Subject")])

   NewID Term Subject
62     5 1005     641
63     5 1005     616
64     5 1005     610
65     5 1005     600
66     5 1003     641
67     5 1003     641
```

Data preparation

The file `subjectNumbers.txt` contains translations from subject numbers to subject names (begun work on more comprehensive file).

```
3         ANTHRO
13         ECON
26         MATHS
29         PHIL
30         POLITICS
285         POLITICS
405         BIOSCI
410         CHEM
445         MATHS
453         PHYSICS
475         STATS
530         HUMANBIO
600         ACCTG
610         COMLAW
616         ECON
641         MGMT
675         ENGSCI
```

Data preparation

```
> subjNumbers <- read.table("Data/subjectNumbers.txt",
+                           sep="\t",
+                           col.names=c("SubjectNumber",
+                             "SubjectName"),
+                           stringsAsFactors=FALSE)
```

Merge this table with `classData`.

```
> classData <- merge(classData,
+                    subjNumbers,
+                    by.x="Subject",
+                    by.y="SubjectNumber",
+                    all.x=TRUE)
> classData$SubjectName[is.na(classData$SubjectName)] <-
+     classData$Subject[is.na(classData$SubjectName)]
```

Data preparation

```
> head(classData[c("NewID", "Subject", "SubjectName")], 10)
```

	NewID	Subject	SubjectName
1	101	13	ECON
2	101	13	ECON
3	101	13	ECON
4	101	26	MATHS
5	101	26	MATHS
6	5	285	POLITICS
7	5	285	POLITICS
8	5	285	POLITICS
9	101	29	PHIL
10	101	29	PHIL

Data preparation

Generate new `School` variable which maps each subject to a school or faculty.

The file `school.txt` contains translations from subject names to schools or faculties.

```
ACADPRAC          Academic Practice          Education
ACCTG          Accounting         Business and Economics
ANCHIST          Ancient History         Arts
ANTHRO          Anthropology         Arts
ARCHDES          Architectural Design         Creative Arts and Industries
ARCHDRC          Architectural Media         Creative Arts and Industries
ARCHGEN          Architecture  -  General         Creative Arts and Industries
ARCHHTC          Architectural History, Theory and Criticism         Creative Arts and Industries
ARCHPRM          Architectural Practice and Management         Creative Arts and Industries
ARCHTECH          Architectural Technology         Creative Arts and Industries
```

Data preparation

```
> schools <- read.table("Data/school.txt",
+                       sep="\t", quote="",
+                       strip.white=TRUE,
+                       stringsAsFactors=FALSE,
+                       col.names=c("Subject", "FullName",
+                         "School", "EMPTY"))
```

Merge this table with `classData`.

```
> classData <- merge(classData, schools[, c(1, 3)],
+                    by.x="Subject", by.y=1)
```

Data preparation

```
> head(classData[c("NewID", "SubjectName", "School")], 10)

   NewID SubjectName                School
1      9       ACCTG Business and Economics
2      5       ACCTG Business and Economics
3     13       ACCTG Business and Economics
4     13       ACCTG Business and Economics
5     17       ACCTG Business and Economics
6     13       ACCTG Business and Economics
7     18       ACCTG Business and Economics
8     25       ACCTG Business and Economics
9     25       ACCTG Business and Economics
10    18       ACCTG Business and Economics
```

Data preparation

Generate new variable `Year` from `Term`.

```
> classData$Year <- 2000 + (classData$Term - 1000) %/% 10

> head(classData[c("NewID", "Term", "Year")], 10)

   NewID Term Year
1      9 1033 2003
2      5 1025 2002
3     13 1043 2004
4     13 1035 2003
5     17 1063 2006
6     13 1045 2004
7     18 1043 2004
8     25 1093 2009
9     25 1085 2008
10    18 1045 2004
```

Data preparation

All of the data are per-paper.

Now want to generate per-student data (102 students).

Focus on each student's **history** by dropping all papers that are currently being taken.

```
> pastPapers <- subset(classData, !Current)
```

```
> dim(classData)
[1] 2722   14
> dim(pastPapers)
[1] 2334   14
```

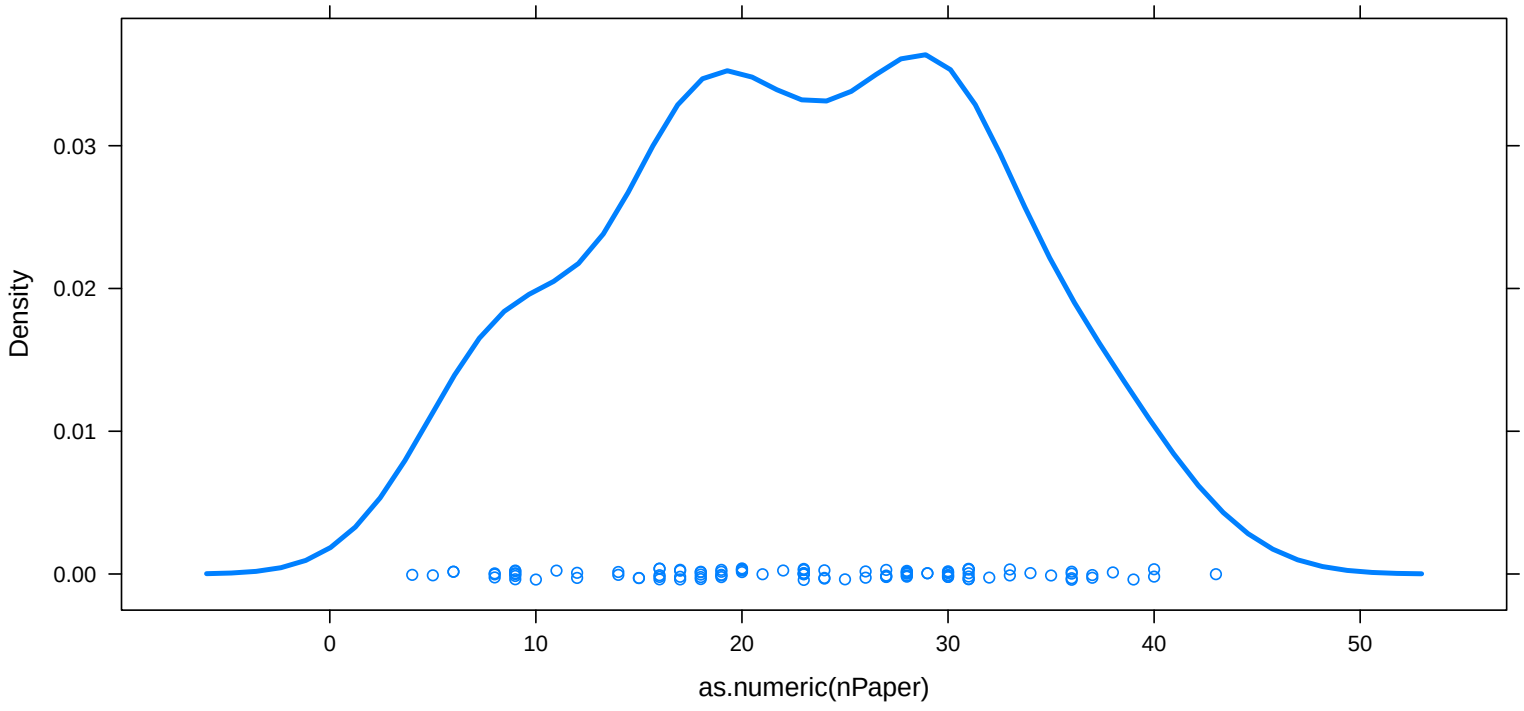
Student history

How many papers has each student taken in the past?

```
> nPaper <- table(pastPapers$NewID)
```

```
> library(lattice)
```

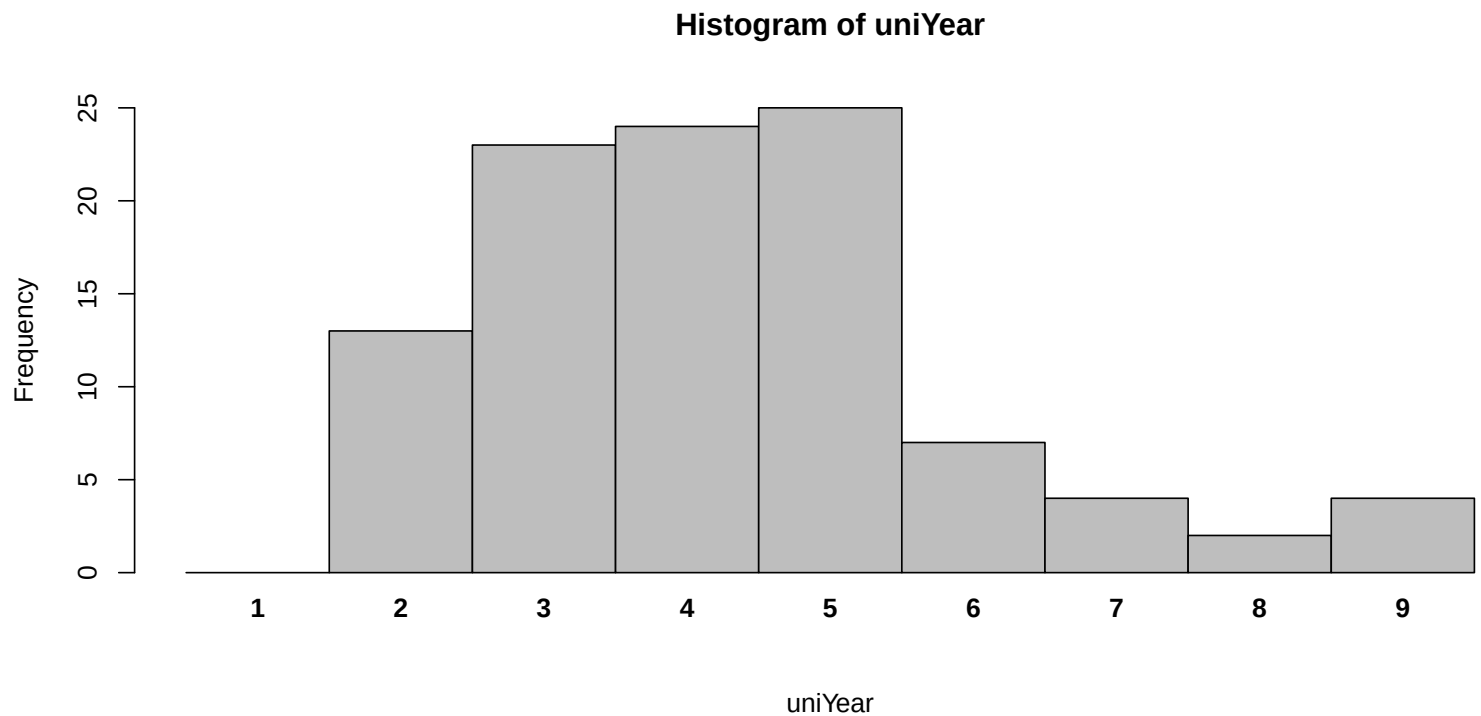
```
> densityplot(as.numeric(nPaper), lwd=3)
```



Student history

Most students are NOT in their second year at university.

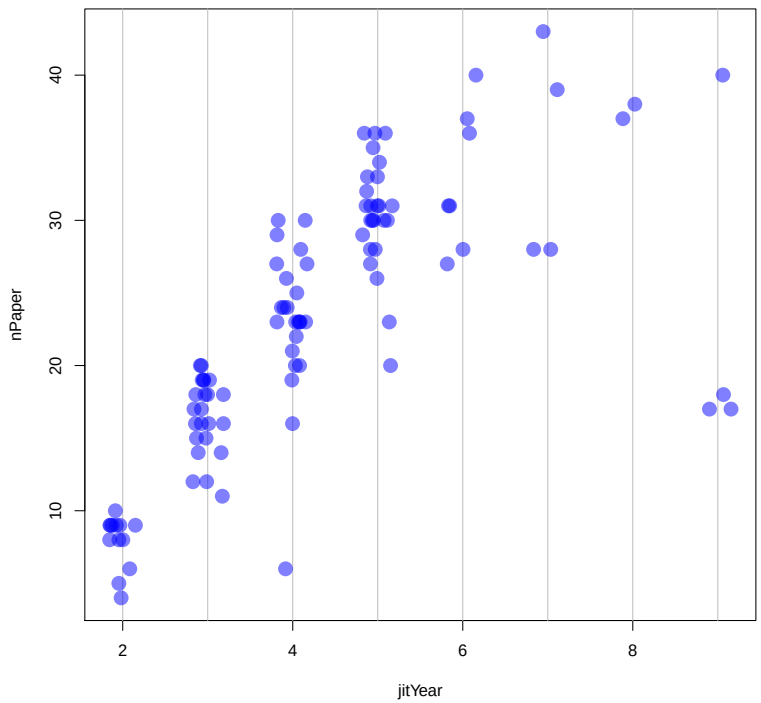
```
> uniYear <- 2009 -
+     tapply(pastPapers$Year, list(pastPapers$NewID), min) + 1
> hist(uniYear, breaks=seq(0.5, 9.5), axes=FALSE, col="grey")
> axis(2)
> mtext(1:9, at=1:9, side=1, font=2)
```



Student history

Time at university mostly corresponds to number of papers taken.

```
> jitYear <- jitter(uniYear)
> plot(jitYear, nPaper, type="n")
> abline(v=2:9, col="grey")
> points(jitYear, nPaper, pch=16,
+        cex=2, col=rgb(0, 0, 1, .5))
```



Student history

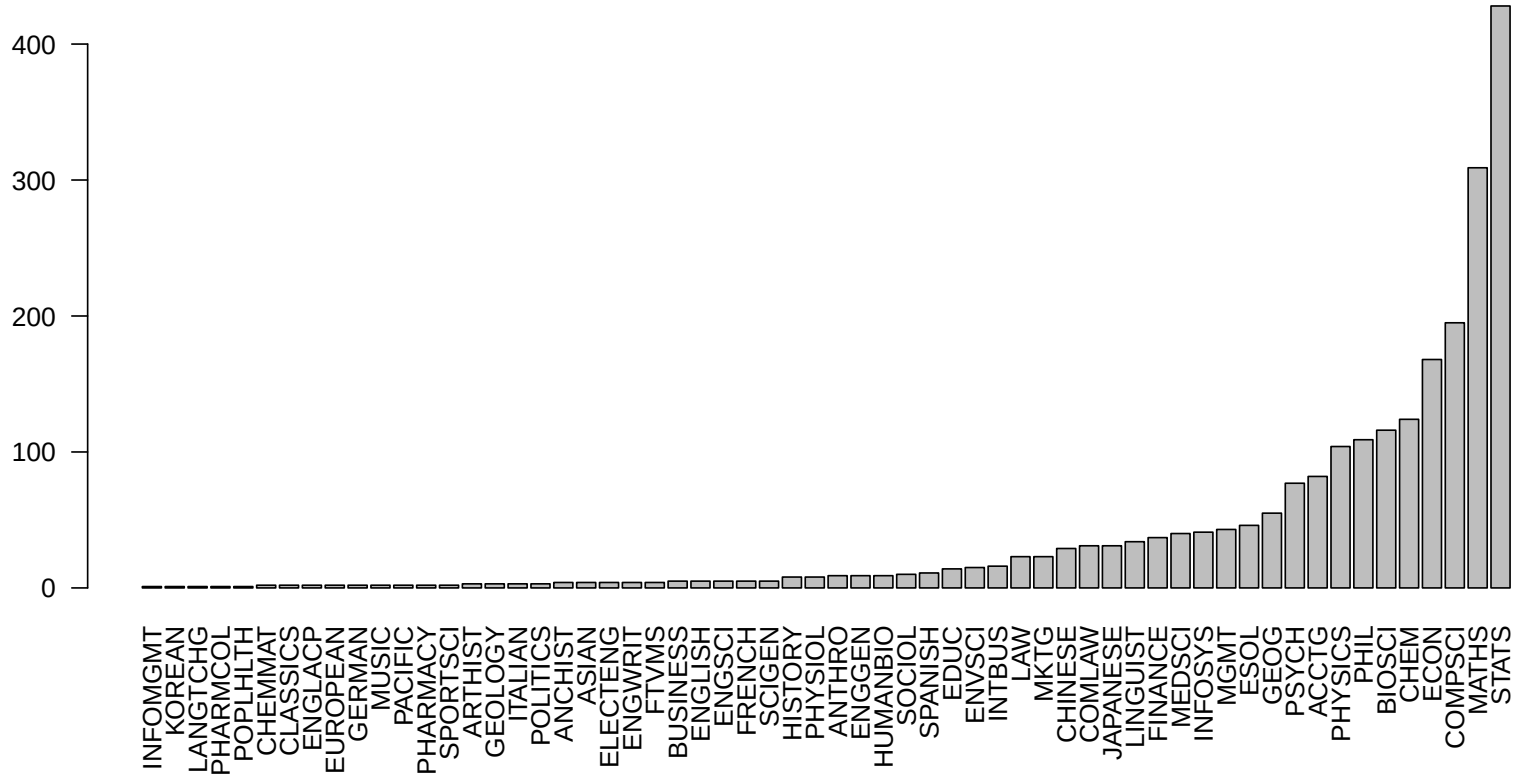
Which subjects have the students taken in the past?

Answer this by counting how many papers each student has taken in each subject.

Student history

The students have taken papers in LOTS of different subjects.

```
> tab <- table(pastPapers$SubjectName)
> ord <- order(tab)
> par(las=2, mar=c(6, 3, 0.5, 0.5))
> barplot(tab[ord])
```



Student history

Too many different subjects to have a count per subject, so only consider the most common subjects (this will also give larger totals in each count).

Generate new variable `Dept` which is based on `Subject`, but only has categories `STATS`, `MATHS`, `COMPSCI`, `ECON`, and `OTHER`.

```
> pastPapers$Dept <- pastPapers$SubjectName
> pastPapers$Dept[!(pastPapers$Dept %in%
+                   c("STATS", "MATHS",
+                     "COMPSCI", "ECON"))] <- "OTHER"
```

Data preparation

```
> head(pastPapers[c("NewID", "SubjectName", "Dept")], 10)
```

	NewID	SubjectName	Dept
1	9	ACCTG	OTHER
2	5	ACCTG	OTHER
3	13	ACCTG	OTHER
4	13	ACCTG	OTHER
5	17	ACCTG	OTHER
6	13	ACCTG	OTHER
7	18	ACCTG	OTHER
9	25	ACCTG	OTHER
10	18	ACCTG	OTHER
11	25	ACCTG	OTHER

Student history

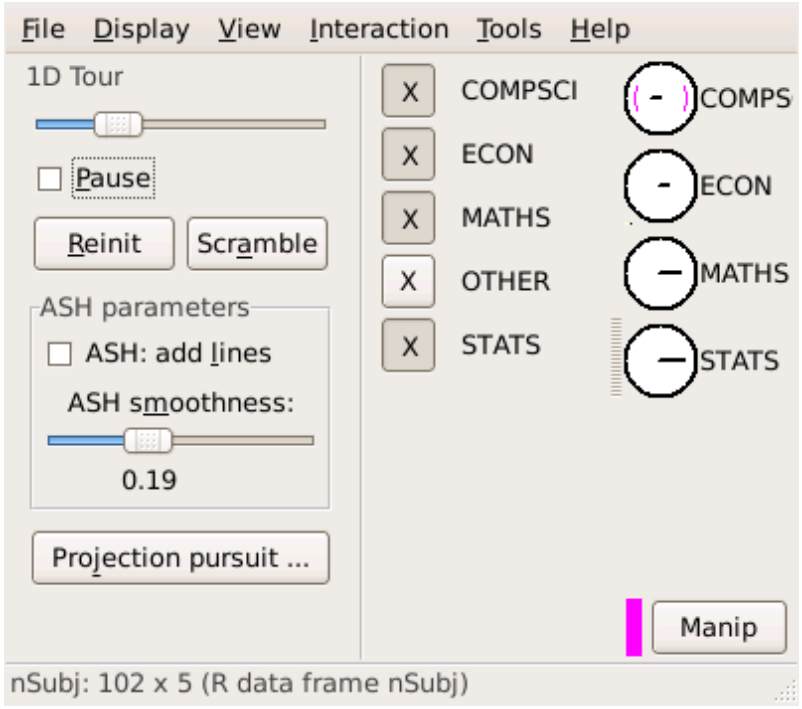
```
> nSubj <- do.call("rbind",
+                  tapply(factor(pastPapers$Dept),
+                         list(ID=pastPapers$NewID),
+                         table,
+                         simplify=FALSE))

> head(nSubj)

  COMPSCI ECON MATHS OTHER STATS
1       0    0     1     5     3
2       2    0     1     3     2
3       0    0     0     0     5
4       2    0     1     0     1
5       0    2     0    14     2
6       0    0     1    15     1
```

Student history

```
> nSubj[1, ]

COMPSCI    ECON   MATHS   OTHER   STATS
      0       0       1       5       3

> pastPapers[pastPapers$NewID == 1,
+            c("NewID", "Dept", "Subject", "SubjectName")]

     NewID  Dept Subject SubjectName
854      1 OTHER    EDUC        EDUC
915      1 OTHER  ENVSCI      ENVSCI
1084     1 OTHER    GEOG        GEOG
1085     1 OTHER    GEOG        GEOG
1090     1 OTHER GEOLOGY     GEOLOGY
1278     1 MATHS   MATHS       MATHS
2106     1 STATS   STATS       STATS
2107     1 STATS   STATS       STATS
2108     1 STATS   STATS       STATS
```

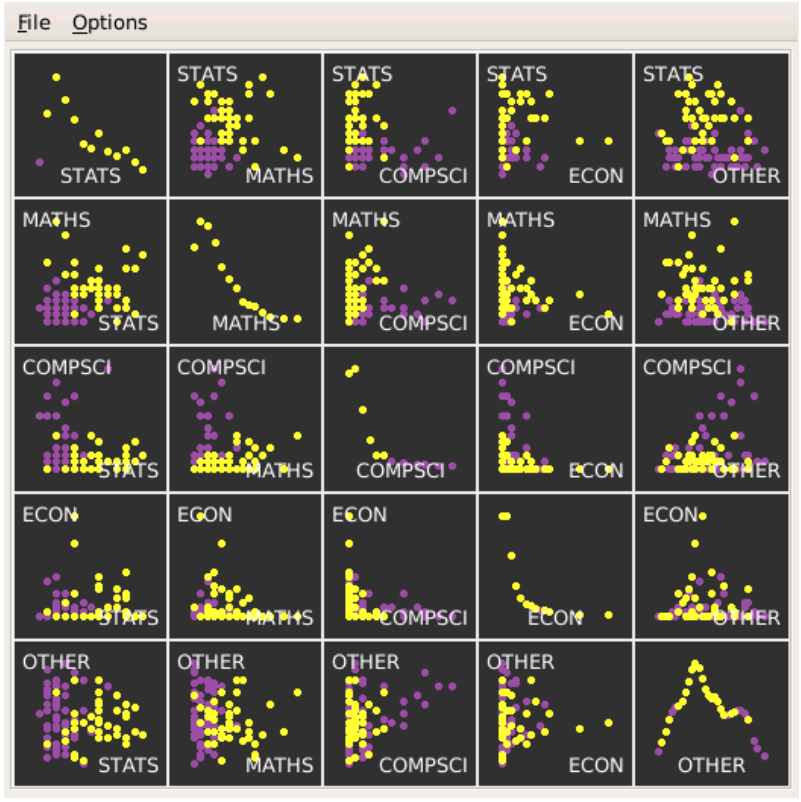
Student history

Two groups : students with several MATHS and/or STATS and students with few (**and** few COMPSCI).



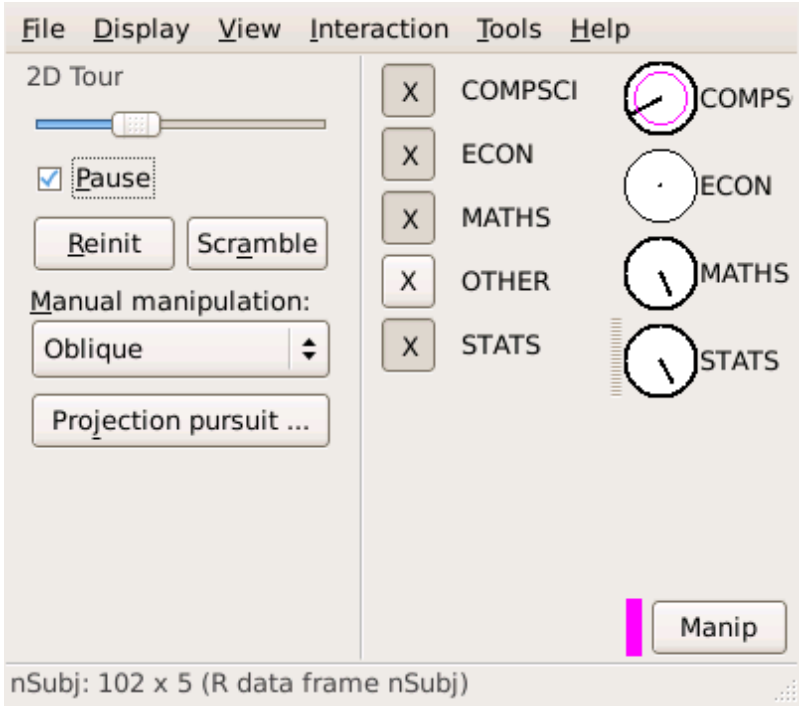
Student history

Two groups : students with several MATHS and/or STATS and students with few (**and** few COMPSCI).



Student history

Three groups : those with several COMPSCI are a separate group.



Student history

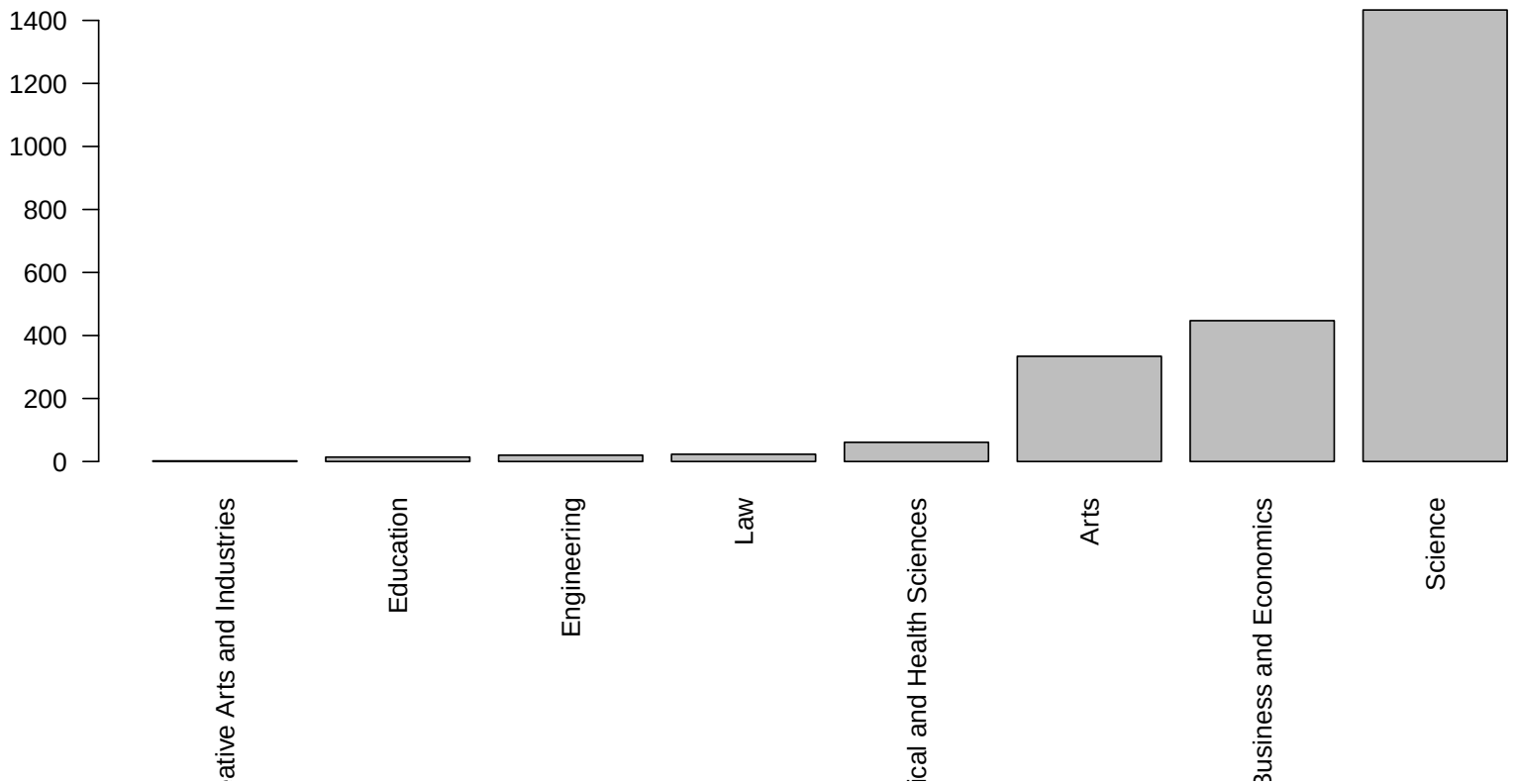
Which schools or faculties have the students taken papers in?

Answer this by counting how many papers each student has taken in each school or faculty.

Student history

The students have taken papers in several different schools/faculties.

```
> stab <- table(pastPapers$School)
> sord <- order(stab)
> par(las=2, mar=c(10, 3, 0.5, 0.5))
> barplot(stab[sord])
```



Student history

Too many different schools/faculties to have a count per subject, so only consider the most common schools/faculties.

Generate new variable `Schl` which is based on `School`, but only has categories `Science`, `Business and Economics`, `Arts`, and `OTHER`.

```
> pastPapers$Schl <- pastPapers$School
> pastPapers$Schl[!(pastPapers$School %in%
+                   c("Science", "Arts",
+                     "Business and Economics"))] <- "OTHER"
```

Data preparation

```
> head(pastPapers[pastPapers$Schl == "OTHER",
+                 c("NewID", "School", "Schl")], 10)
```

	NewID	School	Schl
378	35	Engineering	OTHER
379	20	Engineering	OTHER
854	1	Education	OTHER
855	55	Education	OTHER
856	55	Education	OTHER
857	55	Education	OTHER
858	77	Education	OTHER
859	65	Education	OTHER
861	77	Education	OTHER
862	77	Education	OTHER

Student history

```
> nSchool <- do.call("rbind",
+                    tapply(factor(pastPapers$Schl),
+                           list(ID=pastPapers$NewID),
+                           table,
+                           simplify=FALSE))

> head(nSchool)

  Arts Business and Economics OTHER Science
1    0                      0     1       8
2    0                      2     0       6
3    0                      0     0       5
4    0                      0     0       4
5    5                     10     0       3
6    1                      0    11       5
```

Student history

```
> nSchool[1, ]
```

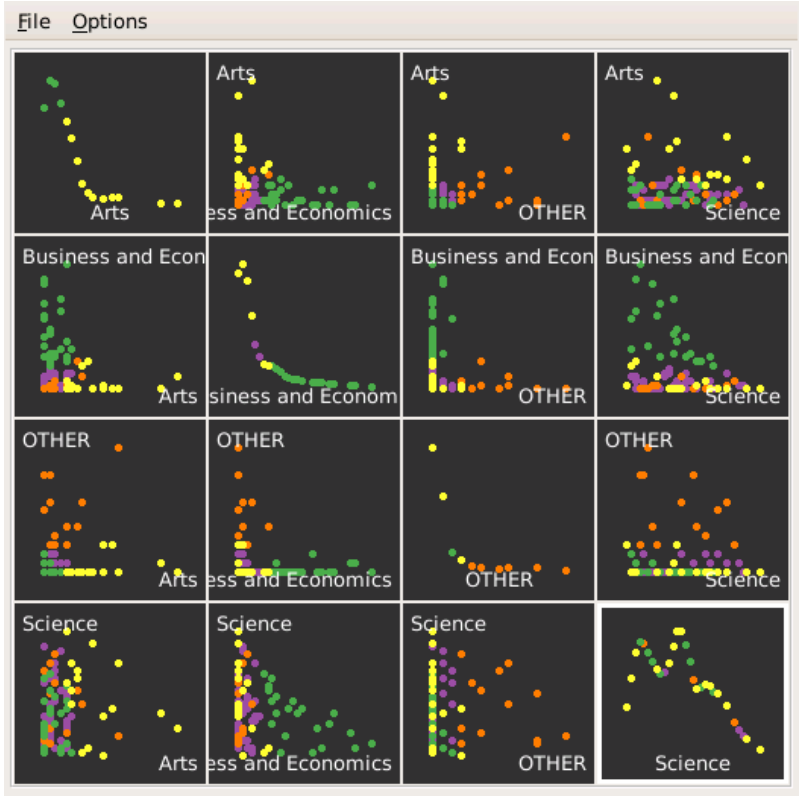
```
                  Arts Business and Economics
                     0                      0
                 OTHER                Science
                     1                      8
```

```
> pastPapers[pastPapers$NewID == 1,
+            c("NewID", "SubjectName", "School", "Schl")]
```

```
     NewID SubjectName    School    Schl
854      1        EDUC Education   OTHER
915      1      ENVSCI   Science Science
1084     1        GEOG   Science Science
1085     1        GEOG   Science Science
1090     1     GEOLOGY   Science Science
1278     1       MATHS   Science Science
2106     1       STATS   Science Science
2107     1       STATS   Science Science
2108     1       STATS   Science Science
```

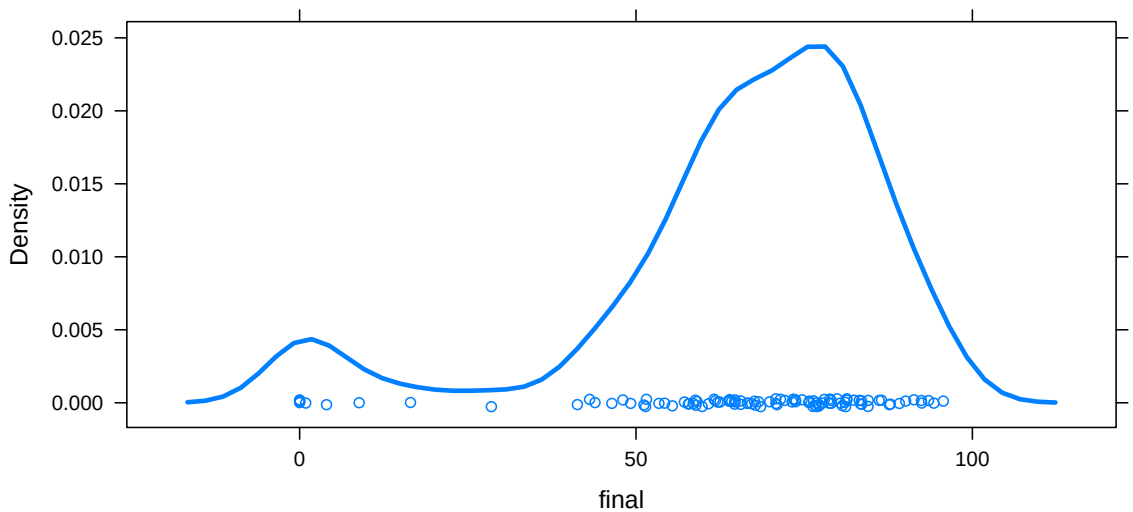
Student history

Four groups : Arts vets, BandE vets, Other vets, and Science.



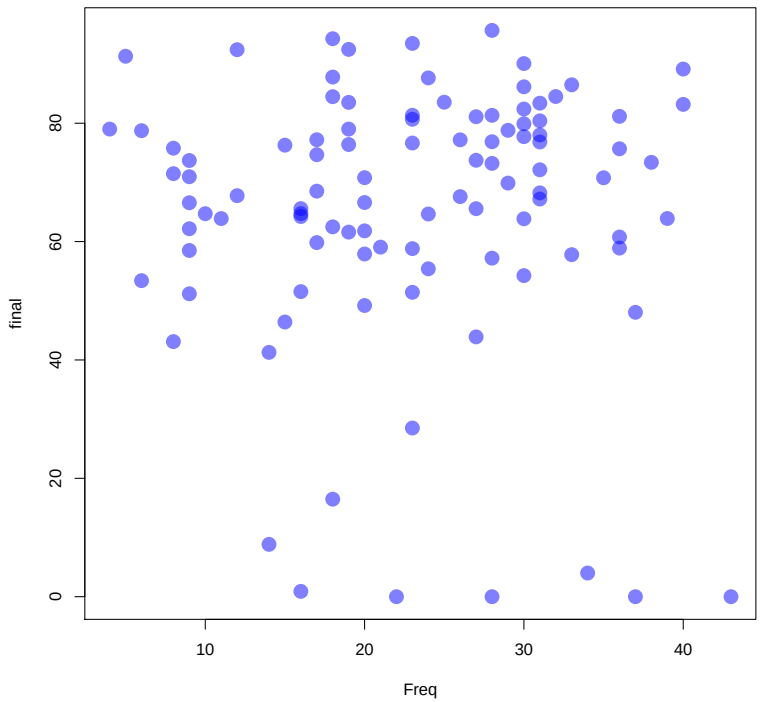
Student performance

```
> exam <- read.csv("PrivData/exam-blind.csv")

> densityplot(~ final, data=exam, lwd=3)
```



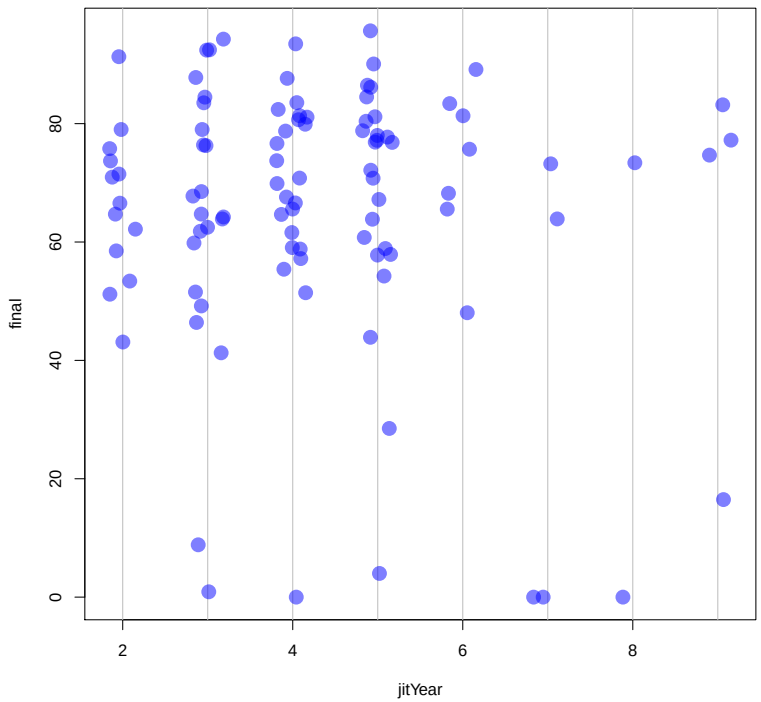
Student performance

```
> examNpaper <- merge(exam, as.data.frame(nPaper),
+                     by.x="NewID", by.y="Var1")

> plot(final ~ Freq, data=examNpaper,
+      pch=16, cex=2, col=rgb(0, 0, 1, .5))
```



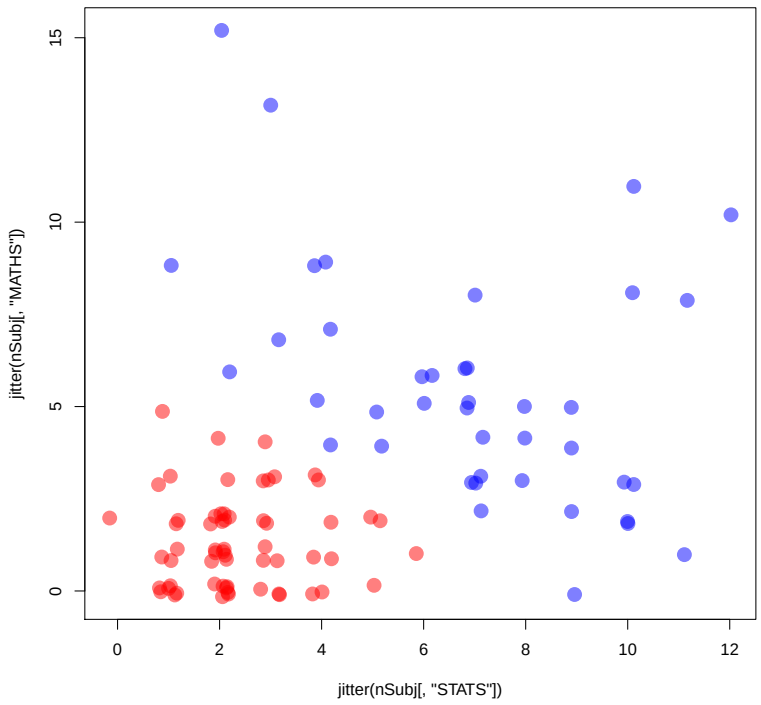
Student performance

```
> examUniYear <- merge(exam, as.data.frame(jitYear),
+                      by.x="NewID", by.y=0)
> plot(final ~ jitYear, data=examUniYear,
+      type="n")
> abline(v=2:9, col="grey")
> points(final ~ jitYear, data=examUniYear,
+        pch=16, cex=2, col=rgb(0, 0, 1, .5))
```



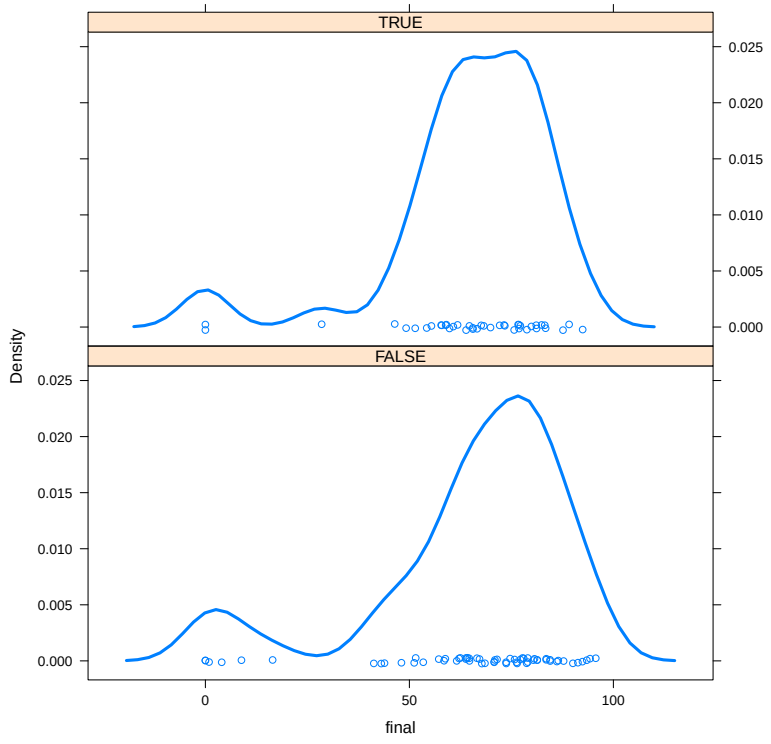
Student performance

```
> grad <- nSubj[, "STATS"] + nSubj[, "MATHS"] >= 8

> plot(jitter(nSubj[, "STATS"]),
+      jitter(nSubj[, "MATHS"]),
+      col=rgb(1:0, 0, 0:1, .5)[grad + 1], pch=16, cex=2)
```



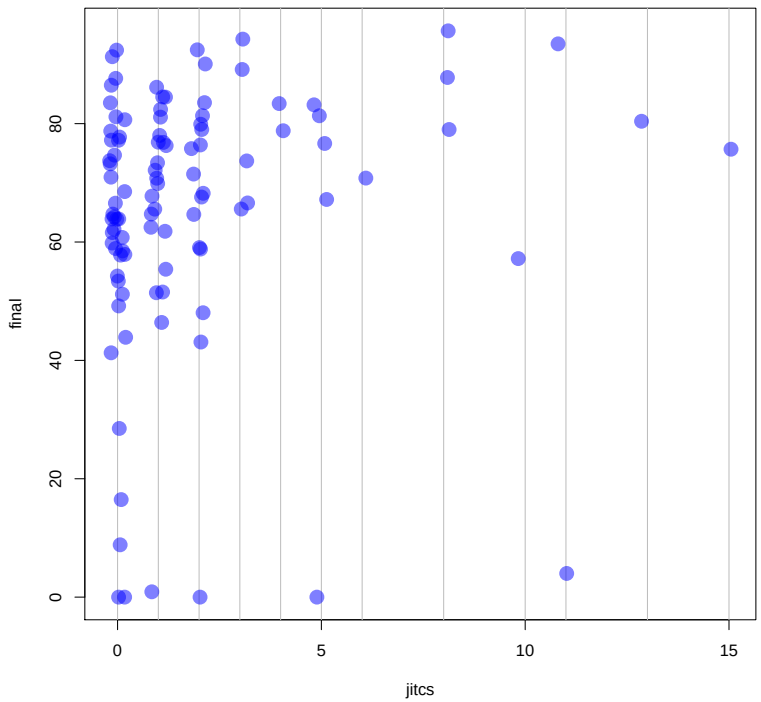
Student performance

```
> examGrad <- merge(exam, grad, by.x="NewID", by.y=0)

> densityplot(~ final | grad, data=examGrad, layout=c(1, 2),
+             lwd=3)
```



Student performance

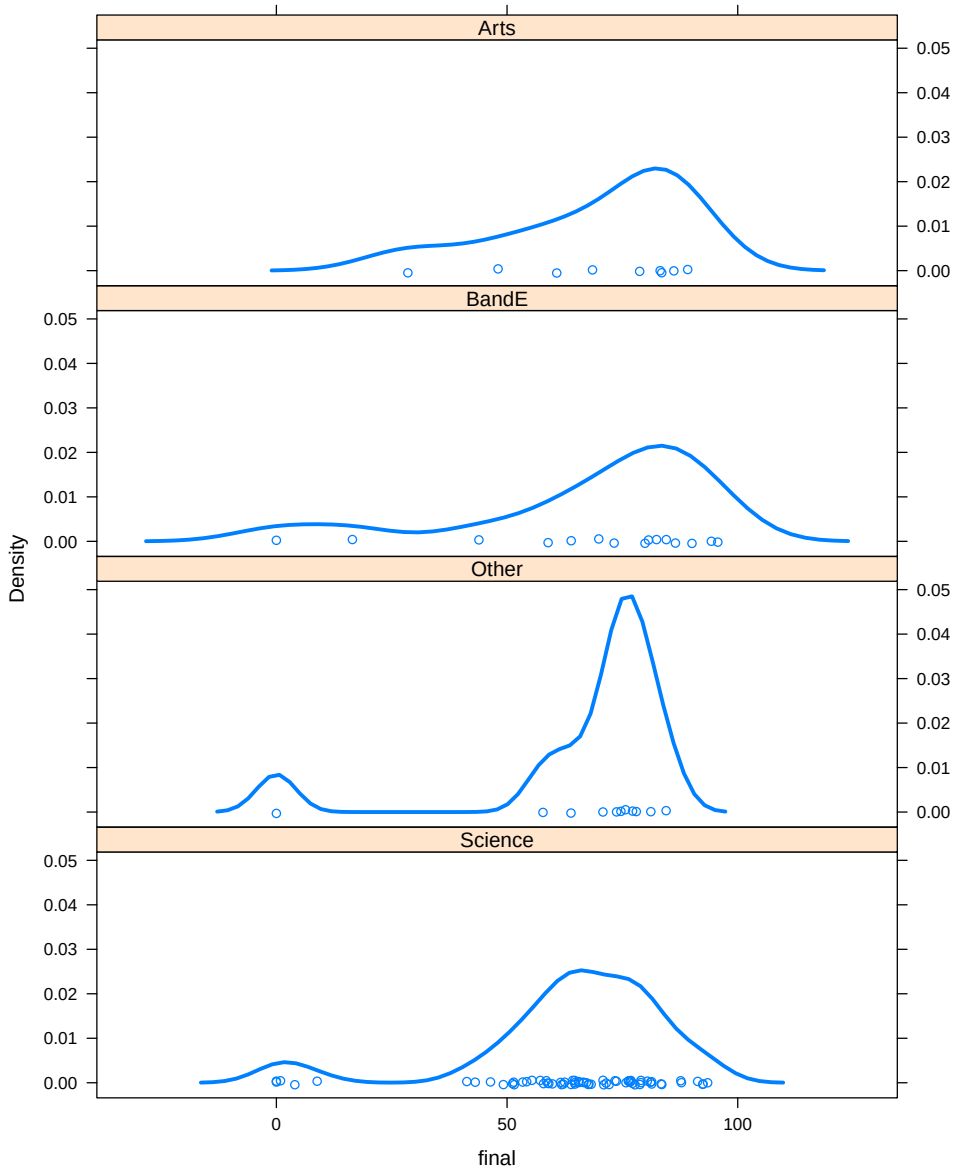
```
> examCompSci <- merge(exam, nSubj[, "COMPSCI", drop=FALSE],
+                      by.x="NewID", by.y=0)
> jitcs <- jitter(examCompSci$COMPSCI)
> plot(final ~ jitcs, data=examCompSci, type="n")
> abline(v=unique(examCompSci$COMPSCI), col="grey")
> points(final ~ jitcs, data=examCompSci,
+        pch=16, cex=2, col=rgb(0, 0, 1, .5))
```



Student performance

```
> SchoolFactor <- read.csv("Data/SchoolFactor-blind.csv")
> SchoolFactor$school <- factor(SchoolFactor$school)
> levels(SchoolFactor$school) <- c("Science", "Other", "BandE", "Arts")
> examSchool <- merge(exam, SchoolFactor)

> head(examSchool)

  NewID final  school
1     1 51.18 Science
2     2 75.78 Science
3     3 91.31 Science
4     4 79.01 Science
5     5 16.48   BandE
6     6 74.68   Other

> densityplot(~ final | school, data=examSchool, layout=c(1, 4),
+             lwd=3)
```

Student performance

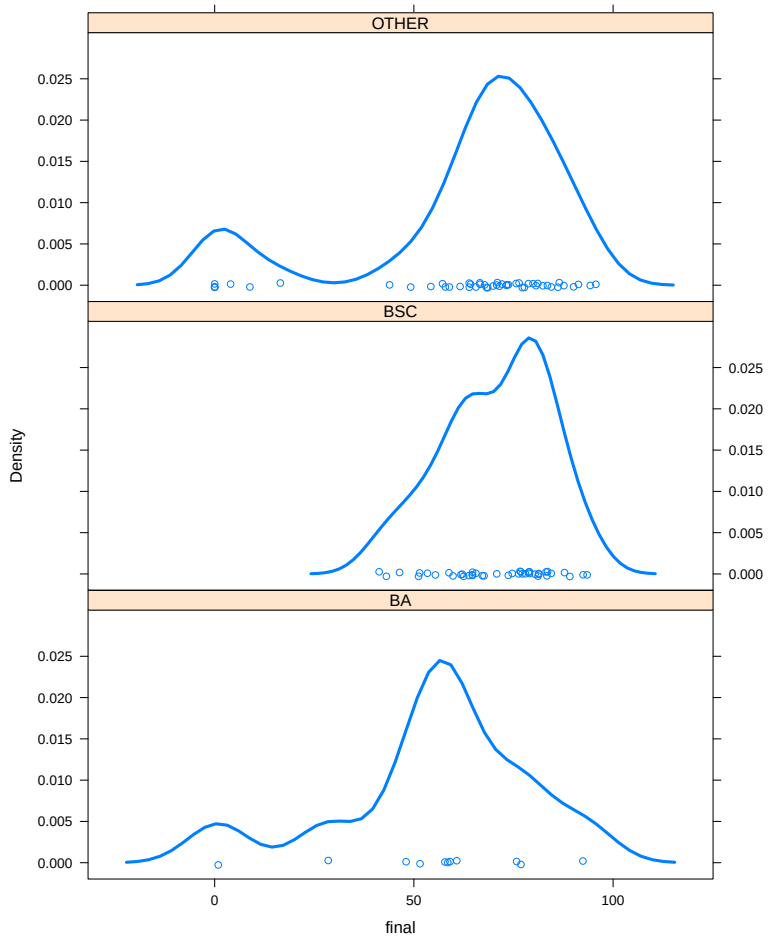


Student performance

```
> program <- aggregate(pastPapers["Acad.Prog"],
+                      list(NewID=pastPapers$NewID),
+                      function(program) {
+                          prog <- paste(sort(unique(program)),
+                                        collapse="-")
+                          switch(prog,
+                                 BA="BA",
+                                 BSC="BSC",
+                                 "OTHER")
+                      })

> head(program)

  NewID Acad.Prog
1     1       BSC
2     2        BA
3     3     OTHER
4     4       BSC
5     5     OTHER
6     6       BSC
```

Student performance

```
> examProgram <- merge(exam, program)
> densityplot(~ final | Acad.Prog, data=examProgram, layout=c(1, 3),
+             lwd=3)
```

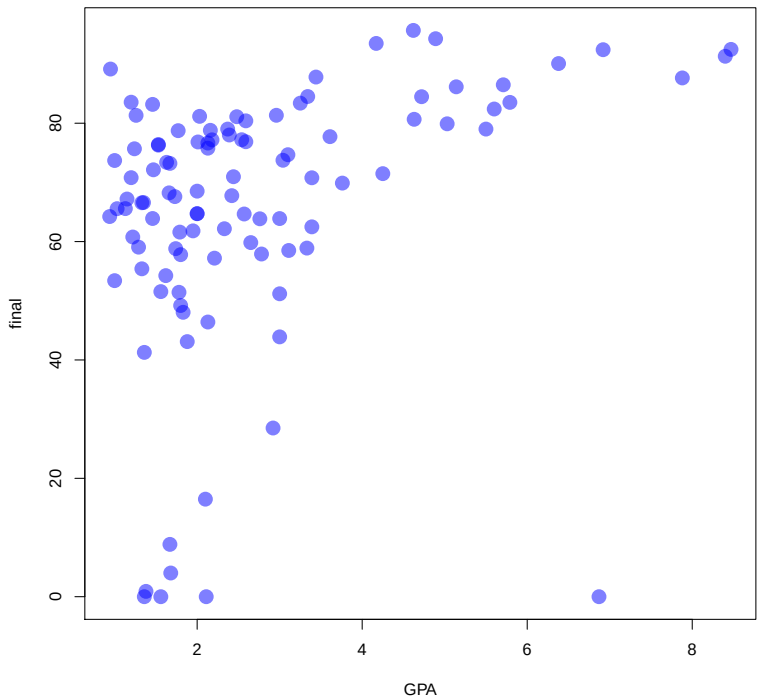


Student performance

```
> table(SchoolFactor$school, program$Acad.Prog)

         BA BSC OTHER
  Science  7  35    25
  Other    1   5     5
  BandE    0   0    15
  Arts     3   2     4
```

Student performance

```
> gpa <- read.csv("PrivData/gpa-blind.csv")
> examGPA <- merge(exam, gpa)

> plot(final ~ GPA, data=examGPA,
+      pch=16, cex=2, col=rgb(0, 0, 1, .5))
```



Summary

- Many students in third, fourth, or fifth year at uni.
- Two student groups: Maths/Stats newbies versus Maths/Stats vets (neither has much Comp Sci)
- More Maths/Stats does not help.
- NOT a separate Comp Sci group, BUT more Comp Sci helps (BUT zero Comp Sci does not doom).
- Four student groups: Arts, BandE, Science, and OTHER.
- Science group worst (BUT BA worse than BSC).
- NO clear evidence found of distinct groups with markedly different performance.
- Best predictor of final mark is GPA.