Report 1

Vanessa Machuca and Luis Espino

February 6, 2018

This data looks at student achievement in a math course and portuguese course at two secondary education Portuguese schools. A dataset is provided for each course and there are approximately 393 students that overlap with both. We found the data through the UCI Machine Learning Repository. The observational units for the dataset are students.

The dataset includes demographic variables like student school name (binary), sex (binary), age (numeric), and whether the student lives in an urban or rural area.

It also includes variables that relate to family, including family size, parents cohabitation status (living apart or together), mother and father education (ordinal), mother and father job (teacher, health care related, civil services (e.g. administrative or police), at home or other), student guardian (mother, father or other), and quality of family relations (1 being very bad to 5 being excellent).

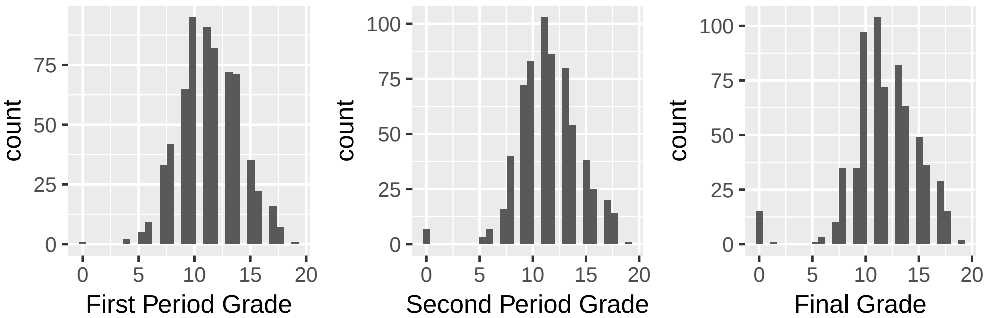
School-related variables include home to school travel time (ordinal), weekly study time (ordinal), and number of past class failures ($1 - 3$ if $n < 3$, 4 otherwise). For each of two courses, students were given a graded after the first and second period as well as a final grade.

Academic routine, habits and performance were also measured. These variables included travel time from home to school (nominal), weekly study time (nominal), number of past class failures (nominal), reason to choose school (categorical), and number of absences (numeric). Additionally, educational resources were assessed through binary measurements of extra educational support, family educational support, extra paid classes within the course subject, extracurricular activities, attended nursery school, college aspiration, and home internet access.

Finally, nominal variables that measured more personal aspect of life included frequency of outings with friends, workday alcohol consumption, weekend alcohol consumption, current health status, and being in a romantic relationship (binary).

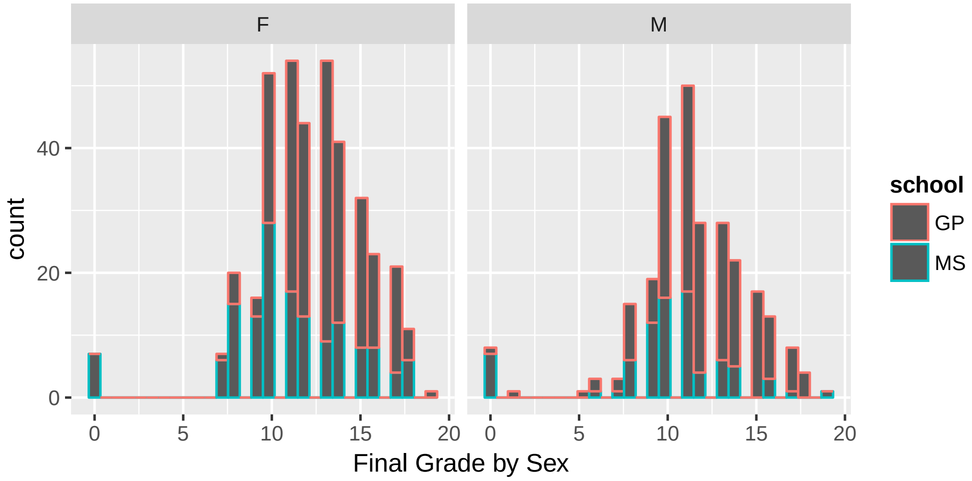
For the purposes of this assignment, we will only look at the observations for students in the portuguese course, since more students took this course.

Let us take a look at the distributions of grades with the following histograms.



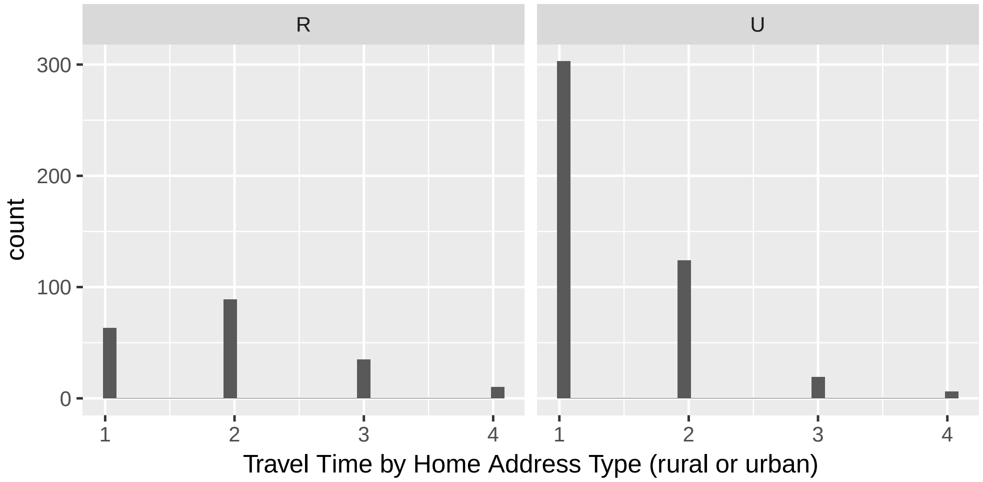
The distributions of each grade seem to be symmetric which is to be expected with grades.

```
> ggplot(student_por, aes(G3)) + geom_histogram(aes(color=school)) + facet_wrap('sex') + theme(legend.ti
```



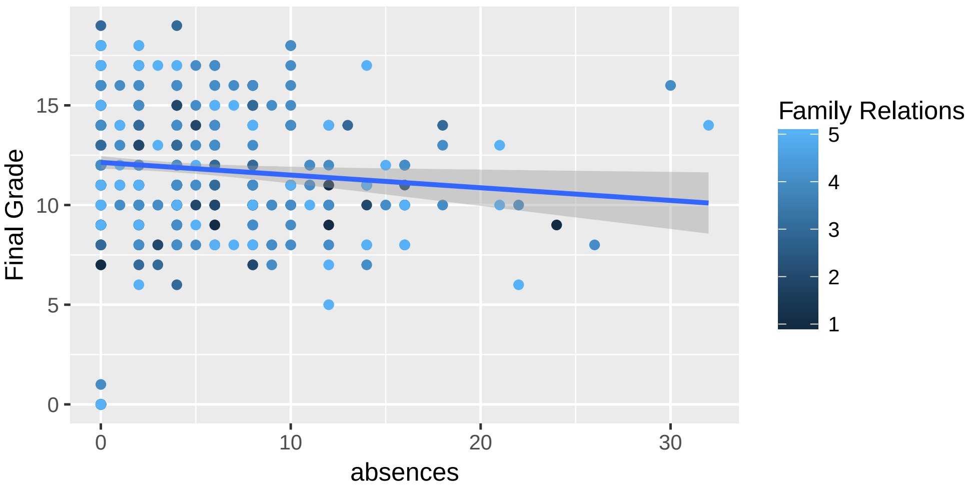
Similarly, the distributions for final grades are fairly symmetric by sex. The range of grades for each is identical and they also both have a few outliers with a final grade of 0. Moreover, notice that students from Gabriel Pereira (GP) make up a larger proportion of the sample than students from Mousinho da Silveira (MS).

```
> ggplot(student_por, aes(traveltime)) + geom_histogram(aes()) + facet_wrap('address') + theme(legend.ti
```



The travel time variable is coded as follows: 1 - travel 15 minutes, 2 - travel 15 to 30 minutes, 3 - travel 30 minutes to 1 hour, or 4 - travel over an hour. Observe that the majority of people who traveled about 15 minutes live in an urban environment. Perhaps this is due to school placement in an urban setting.

```
> ggplot(student_por, aes(absences, G3)) + geom_point(aes(color=famrel)) + labs(col="Family Relations")
```



Intuitevely, we would assume that as number of absences increase, a decrease in student final grade. While this trend exists, it is not very prominent in our data, as shown by the regression line with a slope close to 0 above. Additionally, quality of family relations is pretty randomly spread across the dotplot above, so there are no immediately visible trends as they relate absences and final grades.

```
> table(student_por$famsup, student_por$paid)

       no yes
  no  243   8
  yes 367  31
```

It is interesting to note that people not supported by family in their education varied in terms of having extra paid lessons for the course or not. I would assume that there would be more of a correlation betweeen family support and having paid classes.

```
Skim summary statistics
 n obs: 649
 n variables: 33

Variable type: character
   variable missing complete   n min max empty n_unique
 activities       0      649 649   2   3     0        2
    address       0      649 649   1   1     0        2
    famsize       0      649 649   3   3     0        2
     famsup       0      649 649   2   3     0        2
       Fjob       0      649 649   5   8     0        5
   guardian       0      649 649   5   6     0        3
     higher       0      649 649   2   3     0        2
   internet       0      649 649   2   3     0        2
       Mjob       0      649 649   5   8     0        5
    nursery       0      649 649   2   3     0        2
       paid       0      649 649   2   3     0        2
    Pstatus       0      649 649   1   1     0        2
     reason       0      649 649   4  10     0        4
   romantic       0      649 649   2   3     0        2
     school       0      649 649   2   2     0        2
```

```
  schoolsup       0       649 649    2   3      0          2

Variable type: integer
   variable missing complete   n  mean   sd p0 p25 median p75 p100     hist
   absences       0      649 649  3.66 4.64  0   0      2   6   32 â■Ğâ■Ćâ■Ąâ■Ąâ■Ąâ■Ąâ■Ąâ■Ą
        age       0      649 649 16.74 1.22 15  16     17  18   22 â■Ěâ■Ğâ■Ğâ■Ęâ■Ćâ■Ąâ■Ąâ■Ą
       Dalc       0      649 649  1.5  0.92  1   1      1   2    5 â■Ğâ■Ćâ■Ąâ■Ąâ■Ąâ■Ąâ■Ąâ■Ą
   failures       0      649 649  0.22 0.59  0   0      0   0    3 â■Ğâ■Ąâ■Ąâ■Ąâ■Ąâ■Ąâ■Ąâ■Ą
     famrel       0      649 649  3.93 0.96  1   4      4   5    5 â■Ąâ■Ąâ■Ąâ■Ćâ■Ąâ■Ğâ■Ąâ■Ě
       Fedu       0      649 649  2.31 1.1   0   1      2   3    4 â■Ąâ■Ęâ■Ąâ■Ğâ■Ąâ■Ěâ■Ąâ■Ě
   freetime       0      649 649  3.18 1.05  1   3      3   4    5 â■Ćâ■Čâ■Ąâ■Ğâ■Ąâ■Ęâ■Ąâ■Ć
         G1       0      649 649 11.4  2.75  0  10     11  13   19 â■Ąâ■Ąâ■Ćâ■Čâ■Ęâ■Ğâ■Ćâ■Ą
         G2       0      649 649 11.57 2.91  0  10     11  13   19 â■Ąâ■Ąâ■Ąâ■Ěâ■Ğâ■Ğâ■Ćâ■Ą
         G3       0      649 649 11.91 3.23  0  10     12  14   19 â■Ąâ■Ąâ■Ąâ■Ćâ■Ğâ■Ğâ■Čâ■Ć
      goout       0      649 649  3.18 1.18  1   2      3   4    5 â■Ćâ■Ęâ■Ąâ■Ğâ■Ąâ■Ęâ■Ąâ■Ě
     health       0      649 649  3.54 1.45  1   2      4   5    5 â■Čâ■Ćâ■Ąâ■Čâ■Ąâ■Čâ■Ąâ■Ğ
       Medu       0      649 649  2.51 1.13  0   2      2   4    4 â■Ąâ■Ęâ■Ąâ■Ğâ■Ąâ■Ęâ■Ąâ■Ğ
  studytime       0      649 649  1.93 0.83  1   1      2   2    4 â■Ęâ■Ąâ■Ğâ■Ąâ■Ąâ■Ćâ■Ąâ■Ą
 traveltime       0      649 649  1.57 0.75  1   1      1   2    4 â■Ğâ■Ąâ■Ěâ■Ąâ■Ąâ■Ąâ■Ąâ■Ą
       Walc       0      649 649  2.28 1.28  1   1      2   3    5 â■Ğâ■Ěâ■Ąâ■Čâ■Ąâ■Čâ■Ąâ■Ć
```

Although our compiled PDF does not show it (could not find the appropriate font), the histograms for age, grades 1, 2, and 3, and weekend alcohol consumption all seem relatively symmetric. The distributions of all other variables are either skewed or relatively level across different levels of that variable.

Additionally, note that while the range of student ages is 15-22, most students were on the younger side, with an average age of 16.74. Most students are on the more positive side of health and on the lower extremes of studytime, travel time, weekly alcohol consumption, workday alcohol consumption, number of absences, and failures. The remaining variables we relatively even across the board.

While the sample size is fairly large, we do not know that participants were randomly chosen. Thus, we cannot rule out the possibility of bias in our data. Additionally, this would only be generalizable to the population of students at these two schools taking these courses.