Report 1

Vanessa Machuca and Luis Espino

February 4, 2018

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   age = col_integer(),
##   Medu = col_integer(),
##   Fedu = col_integer(),
##   traveltime = col_integer(),
##   studytime = col_integer(),
##   failures = col_integer(),
##   famrel = col_integer(),
##   freetime = col_integer(),
##   goout = col_integer(),
##   Dalc = col_integer(),
##   Walc = col_integer(),
##   health = col_integer(),
##   absences = col_integer(),
##   G1 = col_integer(),
##   G2 = col_integer(),
##   G3 = col_integer()
## )
```

```
## See spec(...) for full column specifications.
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   age = col_integer(),
##   Medu = col_integer(),
##   Fedu = col_integer(),
##   traveltime = col_integer(),
##   studytime = col_integer(),
##   failures = col_integer(),
##   famrel = col_integer(),
##   freetime = col_integer(),
##   goout = col_integer(),
##   Dalc = col_integer(),
##   Walc = col_integer(),
##   health = col_integer(),
##   absences = col_integer(),
##   G1 = col_integer(),
##   G2 = col_integer(),
##   G3 = col_integer()
## )
```

```
## See spec(...) for full column specifications.
```

```
student_both<-merge(student_mat,student_por,by=c("school","sex","age","address","famsize","Pstatus","Me
ncol(student_both)
```

```
## [1] 53
```

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

```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
## Warning: package 'gridExtra' was built under R version 3.3.2
```

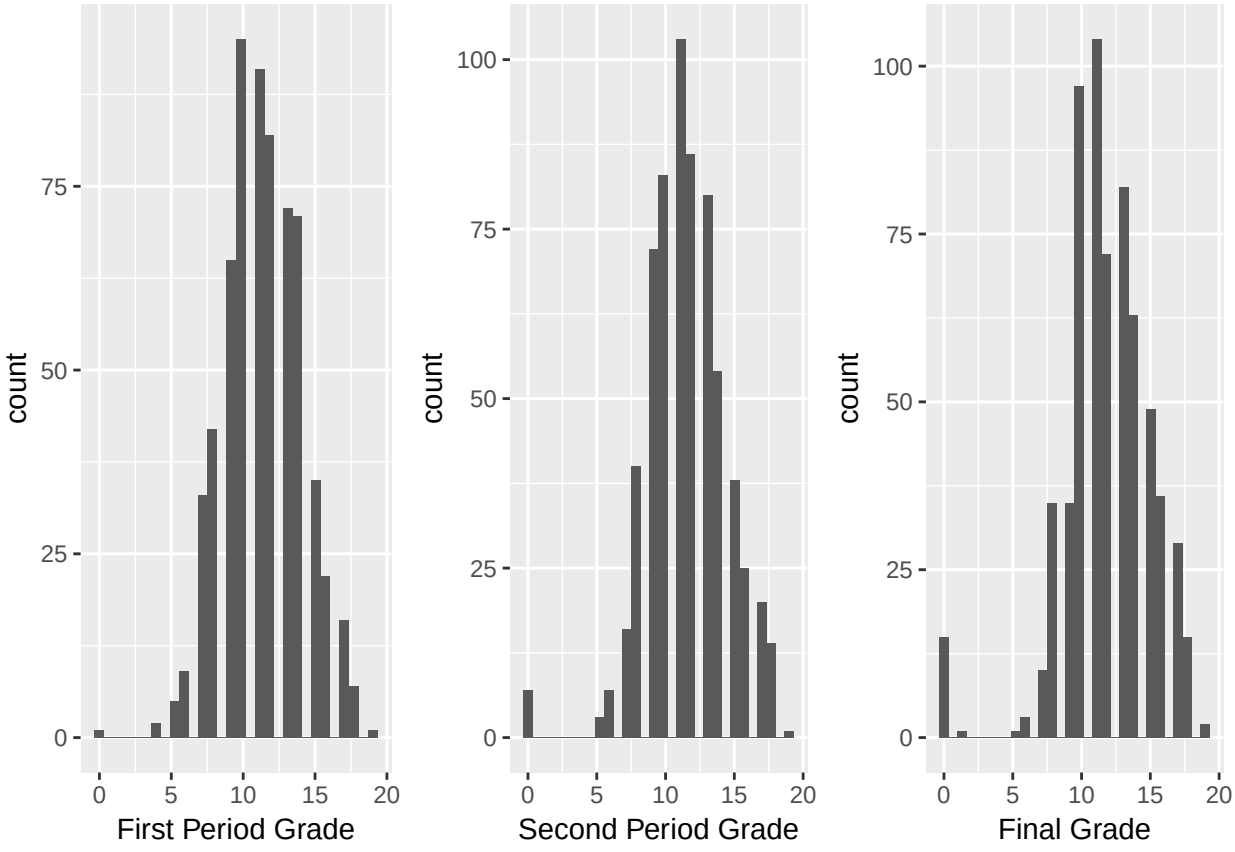
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
first <- ggplot(student_por, aes(x=G1)) + geom_histogram() + labs(x="First Period Grade")

second <- ggplot(student_por, aes(x=G2)) + geom_histogram() + labs(x="Second Period Grade")

final <- ggplot(student_por, aes(x=G3)) + geom_histogram() + labs(x="Final Grade")

grid.arrange(first, second, final, ncol=3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

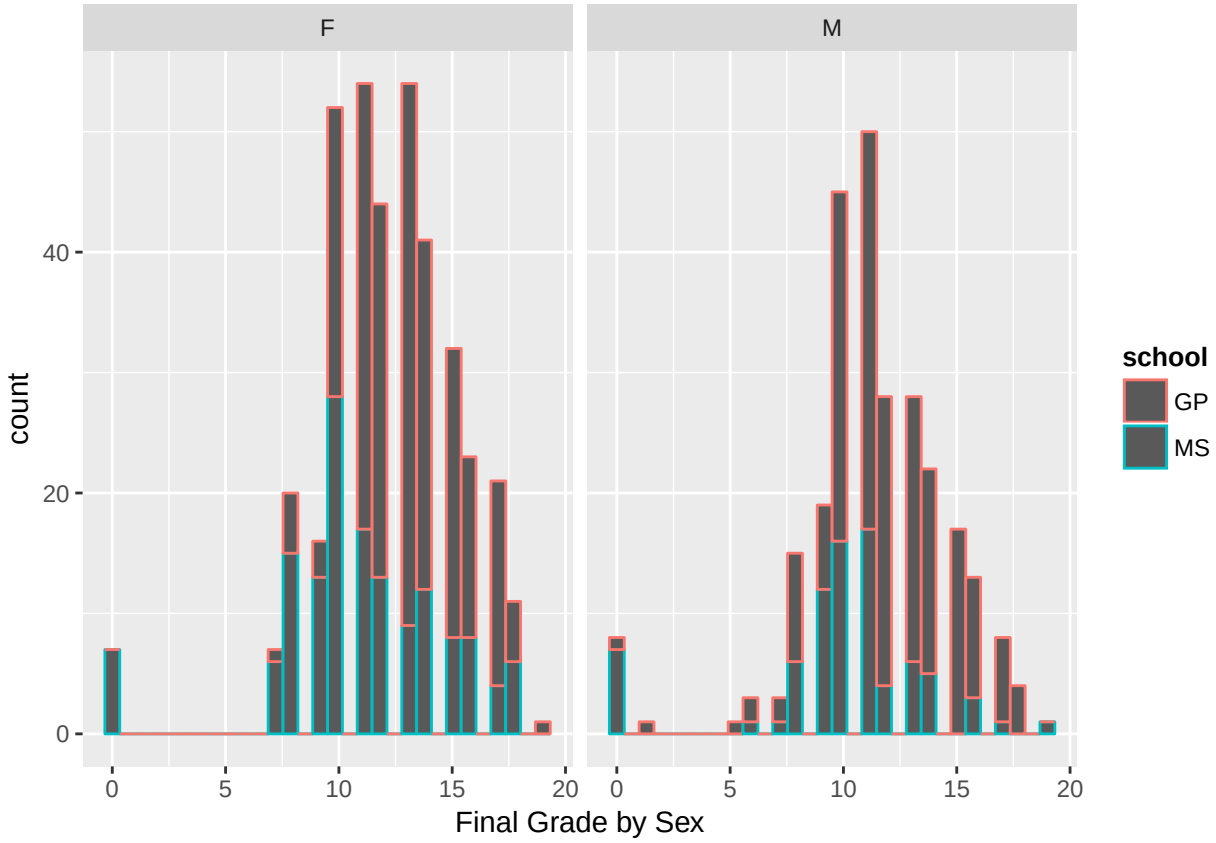
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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.titl
```

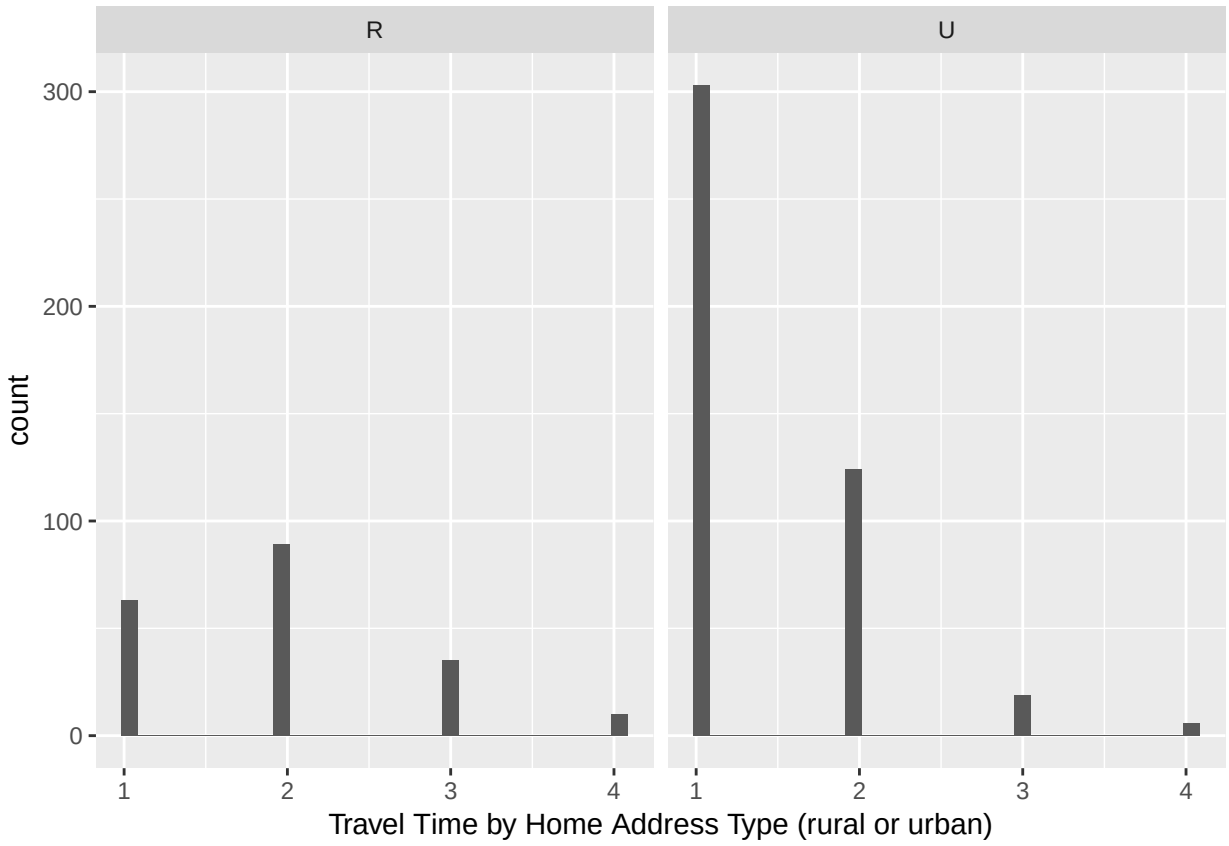
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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).

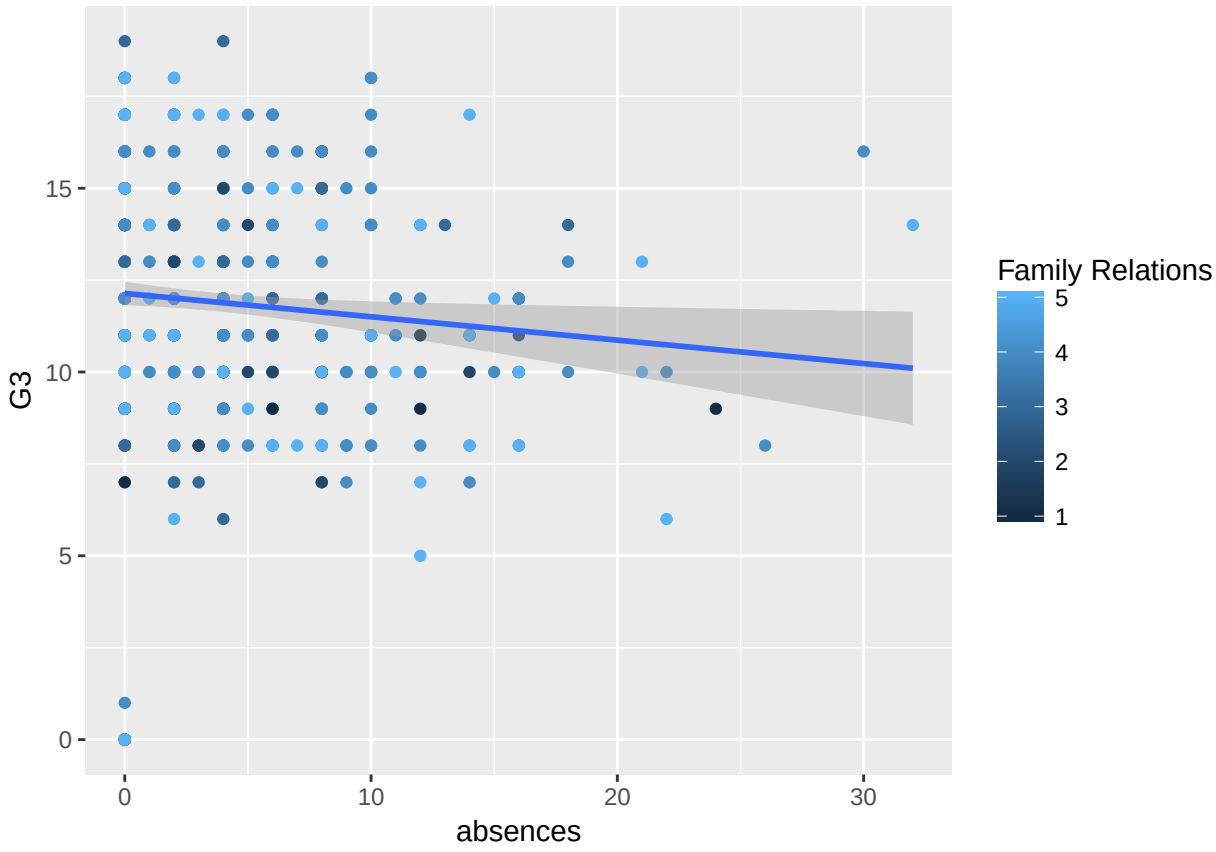
```
ggplot(student_por, aes(traveltime)) + geom_histogram(aes()) + facet_wrap('address') + theme(legend.titl
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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 the school's placement in an urban setting.

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



Intuitively, we would assume that as number of absences increase, a student's final grade would decrease. While this trend exists, it is not very prominent in our data, as shown by the regression line in our dot plot above.