student-success-analysis

Data loading and getting the feeling of the dataset

```
students_org <- readxl::read_excel("student_data.xlsx")

# 370 rows, 39 columns
dim(students_org)
```

```
## [1] 370  39
```

```
# Show column names
names(students_org)
```

```
##  [1] "school"       "sex"          "age"          "address"      "famsize"
##  [6] "Pstatus"      "Medu"         "Fedu"         "Mjob"         "Fjob"
## [11] "reason"       "guardian"     "traveltime"   "studytime"    "failures_mat"
## [16] "failures_por" "schoolsup"    "famsup"       "paid_mat"     "paid_por"
## [21] "activities"   "nursery"      "higher"       "internet"     "romantic"
## [26] "famrel"       "freetime"     "goout"        "Dalc"         "Walc"
## [31] "health"       "absences_mat" "absences_por" "G1_mat"       "G2_mat"
## [36] "G3_mat"       "G1_por"       "G2_por"       "G3_por"
```

```
# Show first few rows
head(students_org)
```

```
## # A tibble: 6 x 39
##   school sex     age address famsize Pstatus  Medu  Fedu Mjob     Fjob   reason
##   <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <dbl> <dbl> <chr>    <chr>  <chr>
## 1 GP     F        18 U       GT3     A           4     4 at_home  teach~ course
## 2 GP     F        17 U       GT3     T           1     1 at_home  other  course
## 3 GP     F        15 U       LE3     T           1     1 at_home  other  other
## 4 GP     F        15 U       GT3     T           4     2 health   servi~ home
## 5 GP     F        16 U       GT3     T           3     3 other    other  home
## 6 GP     M        16 U       LE3     T           4     3 services other  reputa~
## # ... with 28 more variables: guardian <chr>, traveltime <dbl>,
## #   studytime <dbl>, failures_mat <dbl>, failures_por <dbl>, schoolsup <chr>,
## #   famsup <chr>, paid_mat <chr>, paid_por <chr>, activities <chr>,
## #   nursery <chr>, higher <chr>, internet <chr>, romantic <chr>, famrel <dbl>,
## #   freetime <dbl>, goout <dbl>, Dalc <dbl>, Walc <dbl>, health <dbl>,
## #   absences_mat <dbl>, absences_por <dbl>, G1_mat <dbl>, G2_mat <dbl>,
## #   G3_mat <dbl>, G1_por <dbl>, G2_por <dbl>, G3_por <dbl>
```

```
# Show details for each column
summary(students_org)
```

```
##     school              sex                 age          address
##  Length:370         Length:370         Min.   :15.00   Length:370
##  Class :character   Class :character   1st Qu.:16.00   Class :character
##  Mode  :character   Mode  :character   Median :17.00   Mode  :character
##                                        Mean   :16.58
```

```
##                                       3rd Qu.:17.00
##                                       Max.   :22.00
##    famsize            Pstatus              Medu          Fedu
##  Length:370         Length:370         Min.   :0.0   Min.   :0.000
##  Class :character   Class :character   1st Qu.:2.0   1st Qu.:2.000
##  Mode  :character   Mode  :character   Median :3.0   Median :3.000
##                                        Mean   :2.8   Mean   :2.557
##                                        3rd Qu.:4.0   3rd Qu.:3.750
##                                        Max.   :4.0   Max.   :4.000
##      Mjob               Fjob              reason            guardian
##  Length:370         Length:370         Length:370         Length:370
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    traveltime      studytime      failures_mat     failures_por
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :2.000   Median :0.0000   Median :0.0000
##  Mean   :1.446   Mean   :2.043   Mean   :0.2784   Mean   :0.1324
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :4.000   Max.   :4.000   Max.   :3.0000   Max.   :3.0000
##   schoolsup            famsup            paid_mat           paid_por
##  Length:370         Length:370         Length:370         Length:370
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   activities          nursery             higher            internet
##  Length:370         Length:370         Length:370         Length:370
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    romantic              famrel         freetime         goout
##  Length:370         Min.   :1.000   Min.   :1.000   Min.   :1.000
##  Class :character   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000
##  Mode  :character   Median :4.000   Median :3.000   Median :3.000
##                     Mean   :3.935   Mean   :3.224   Mean   :3.116
##                     3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000
##                     Max.   :5.000   Max.   :5.000   Max.   :5.000
##       Dalc            Walc           health       absences_mat
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 0.000
##  Median :1.000   Median :2.000   Median :4.000   Median : 4.000
##  Mean   :1.484   Mean   :2.295   Mean   :3.562   Mean   : 5.381
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 8.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :75.000
##   absences_por        G1_mat          G2_mat          G3_mat
##  Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
```

```
##  Median : 2.000   Median :11.00   Median :11.00   Median :11.00
##  Mean   : 3.632   Mean   :10.89   Mean   :10.75   Mean   :10.46
##  3rd Qu.: 6.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :32.000   Max.   :19.00   Max.   :19.00   Max.   :20.00
##      G1_por          G2_por          G3_por
##  Min.   : 0.00   Min.   : 5.00   Min.   : 0.00
##  1st Qu.:10.00   1st Qu.:11.00   1st Qu.:11.00
##  Median :12.00   Median :12.00   Median :13.00
##  Mean   :12.14   Mean   :12.27   Mean   :12.55
##  3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00
##  Max.   :19.00   Max.   :19.00   Max.   :19.00
```

```
# Check the class of the column. "numeric", "character"...
sapply(students_org, class)
```

```
##       school          sex          age      address      famsize      Pstatus
##  "character"  "character"    "numeric"  "character"  "character"  "character"
##         Medu         Fedu         Mjob         Fjob       reason     guardian
##    "numeric"    "numeric"  "character"  "character"  "character"  "character"
##   traveltime    studytime failures_mat failures_por    schoolsup       famsup
##    "numeric"    "numeric"    "numeric"    "numeric"  "character"  "character"
##     paid_mat     paid_por   activities      nursery       higher     internet
##  "character"  "character"  "character"  "character"  "character"  "character"
##     romantic       famrel     freetime        goout         Dalc         Walc
##  "character"    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"
##       health absences_mat absences_por       G1_mat       G2_mat       G3_mat
##    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"
##       G1_por       G2_por       G3_por
##    "numeric"    "numeric"    "numeric"
```

```
# Let's check if any columns exceed the maximum or minumum values specified in the pdf
# This makes sense only for numerical values

colMax <- students_org %>%
  select(where(is.numeric)) %>%
  sapply(., max, na.rm = TRUE)
colMax
```

```
##          age         Medu         Fedu   traveltime    studytime failures_mat
##           22            4            4            4            4            3
## failures_por       famrel     freetime        goout         Dalc         Walc
##            3            5            5            5            5            5
##       health absences_mat absences_por       G1_mat       G2_mat       G3_mat
##            5           75           32           19           19           20
##       G1_por       G2_por       G3_por
##           19           19           19
```

```
# Every column has normal maximum value
```

```
# Are there any na values?
students_org %>% filter(is.na(.))
```

```
## # A tibble: 0 x 39
## # ... with 39 variables: school <chr>, sex <chr>, age <dbl>, address <chr>,
## #   famsize <chr>, Pstatus <chr>, Medu <dbl>, Fedu <dbl>, Mjob <chr>,
## #   Fjob <chr>, reason <chr>, guardian <chr>, traveltime <dbl>,
## #   studytime <dbl>, failures_mat <dbl>, failures_por <dbl>, schoolsup <chr>,
```

```
## #   famsup <chr>, paid_mat <chr>, paid_por <chr>, activities <chr>,
## #   nursery <chr>, higher <chr>, internet <chr>, romantic <chr>, famrel <dbl>,
## #   freetime <dbl>, goout <dbl>, Dalc <dbl>, Walc <dbl>, health <dbl>, ...
```

```
sum(apply(students_org, 2, is.nan))
```

```
## [1] 0
```

```
students_org %>% filter(is.null(.))
```

```
## # A tibble: 0 x 39
## # ... with 39 variables: school <chr>, sex <chr>, age <dbl>, address <chr>,
## #   famsize <chr>, Pstatus <chr>, Medu <dbl>, Fedu <dbl>, Mjob <chr>,
## #   Fjob <chr>, reason <chr>, guardian <chr>, traveltime <dbl>,
## #   studytime <dbl>, failures_mat <dbl>, failures_por <dbl>, schoolsup <chr>,
## #   famsup <chr>, paid_mat <chr>, paid_por <chr>, activities <chr>,
## #   nursery <chr>, higher <chr>, internet <chr>, romantic <chr>, famrel <dbl>,
## #   freetime <dbl>, goout <dbl>, Dalc <dbl>, Walc <dbl>, health <dbl>, ...
```

```
# Drop these values just in case they show up with another dataset
# We will continue using "student" variable
students <- students_org %>% filter_all(all_vars(!is.na(.) & !is.nan(.) & !is.null(.)))
```

```
# Show average grade for all schools

schools <- students %>%
  select("school") %>%
  distinct(.)
schools # [GP, MS]
```

```
## # A tibble: 2 x 1
##   school
##   <chr>
## 1 GP
## 2 MS
```

```
subject_final_grade_names <- names(students)[grepl("G3", names(students))]
subject_final_grade_names
```

```
## [1] "G3_mat" "G3_por"
```

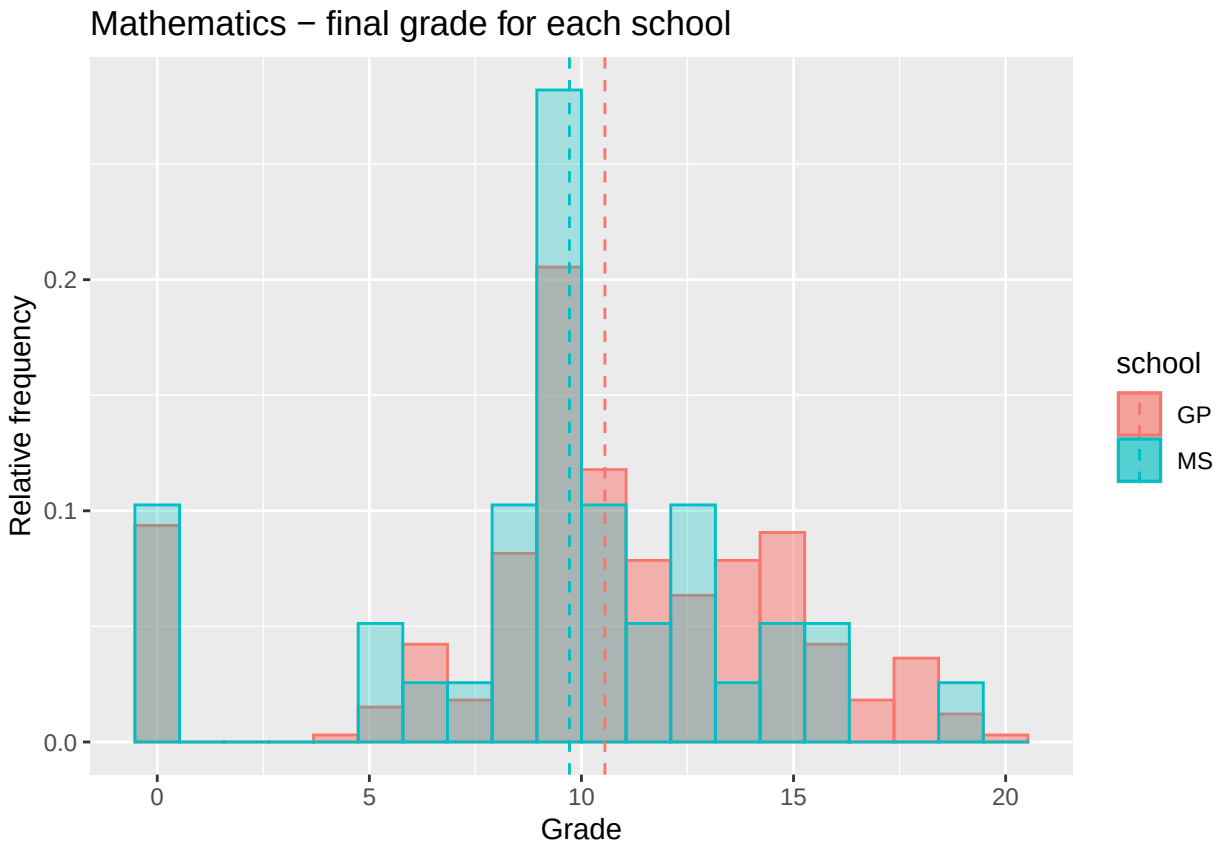
```
students_final_grade <- students %>% select("school", subject_final_grade_names)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(subject_final_grade_names)` instead of `subject_final_grade_names` to silence this mess
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

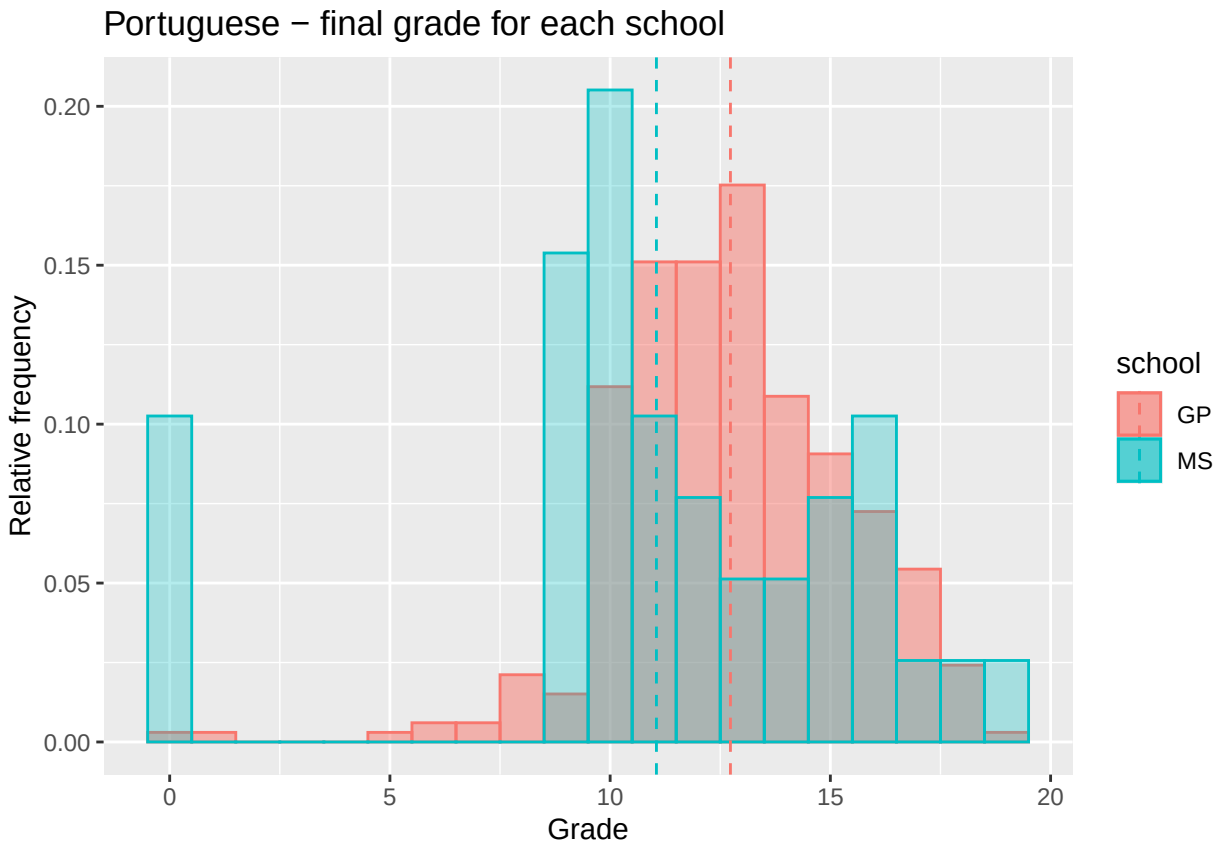
```
# Select only the subject grade and school
gp_mat <- students_final_grade %>%
  filter(school == "GP") %>%
  select(G3_mat, school)
gp_por <- students_final_grade %>%
  filter(school == "GP") %>%
  select(G3_por, school)
ms_mat <- students_final_grade %>%
  filter(school == "MS") %>%
  select(G3_mat, school)
ms_por <- students_final_grade %>%
```

```
  filter(school == "MS") %>%
  select(G3_por, school)

# Rename all columns to "grade"
gp_mat <- gp_mat %>% rename(grade = G3_mat)
gp_por <- gp_por %>% rename(grade = G3_por)
ms_mat <- ms_mat %>% rename(grade = G3_mat)
ms_por <- ms_por %>% rename(grade = G3_por)


# TODO: can this data be grouped and used dynamically? (support multiple schools and mutliple subjects)

# scale_this <- function(x){
#   (x - mean(x, na.rm=TRUE)) / sd(x, na.rm=TRUE)
# }

# gp_mat_scaled <- gp_mat %>% mutate(G3_mat = scale_this(G3_mat))
# gp_por_scaled <- gp_por %>% mutate(G3_por = scale_this(G3_por))
# ms_mat_scaled <- ms_mat %>% mutate(G3_mat = scale_this(G3_mat))
# ms_por_scaled <- ms_por %>% mutate(G3_por = scale_this(G3_por))
```

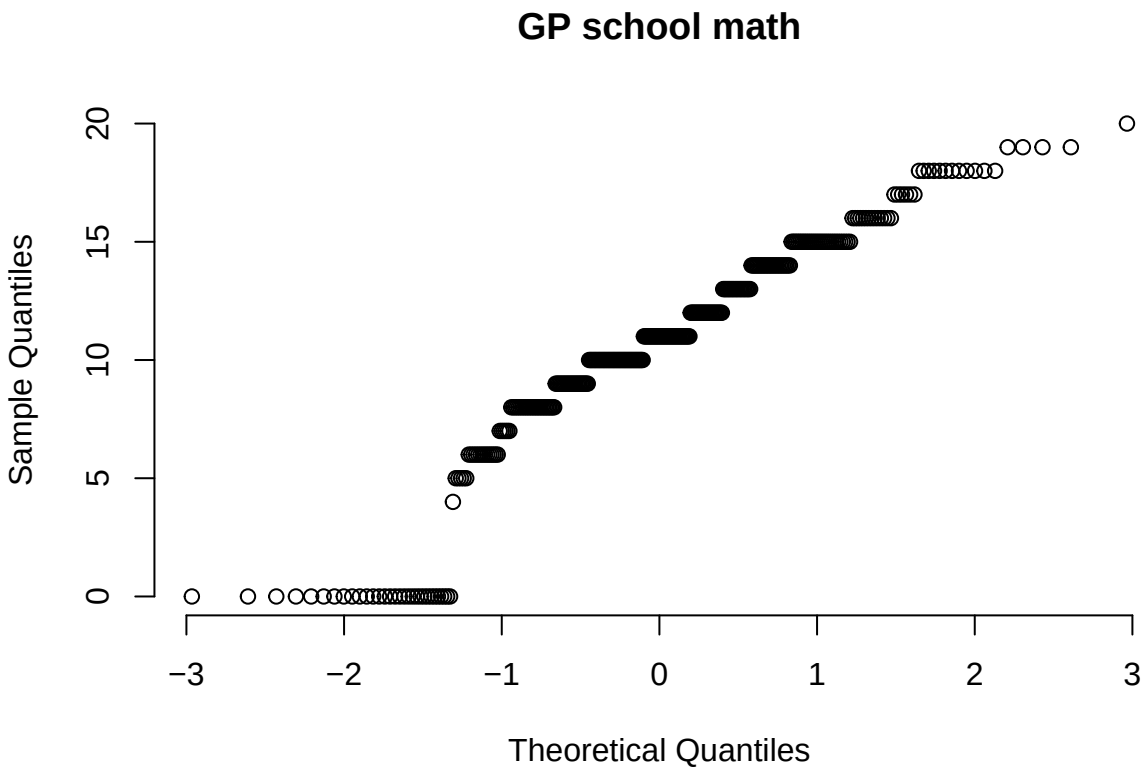
```
# Plot math -- final grade
ggplot(gp_mat, aes(x = grade, y = (..count.. / sum(..count..)))) +
  geom_histogram(bins = 20, aes(color = school, fill = school), alpha = 0.5) +
  geom_histogram(data = ms_mat, bins = 20, aes(color = school, fill = school), alpha = 0.3) +
  geom_vline(data = gp_mat, aes(xintercept = mean(grade), color = school), linetype = "dashed") +
  geom_vline(data = ms_mat, aes(xintercept = mean(grade), color = school), linetype = "dashed") +
  xlab("Grade") +
  ylab("Relative frequency") +
  labs(title = "Mathematics - final grade for each school")
```



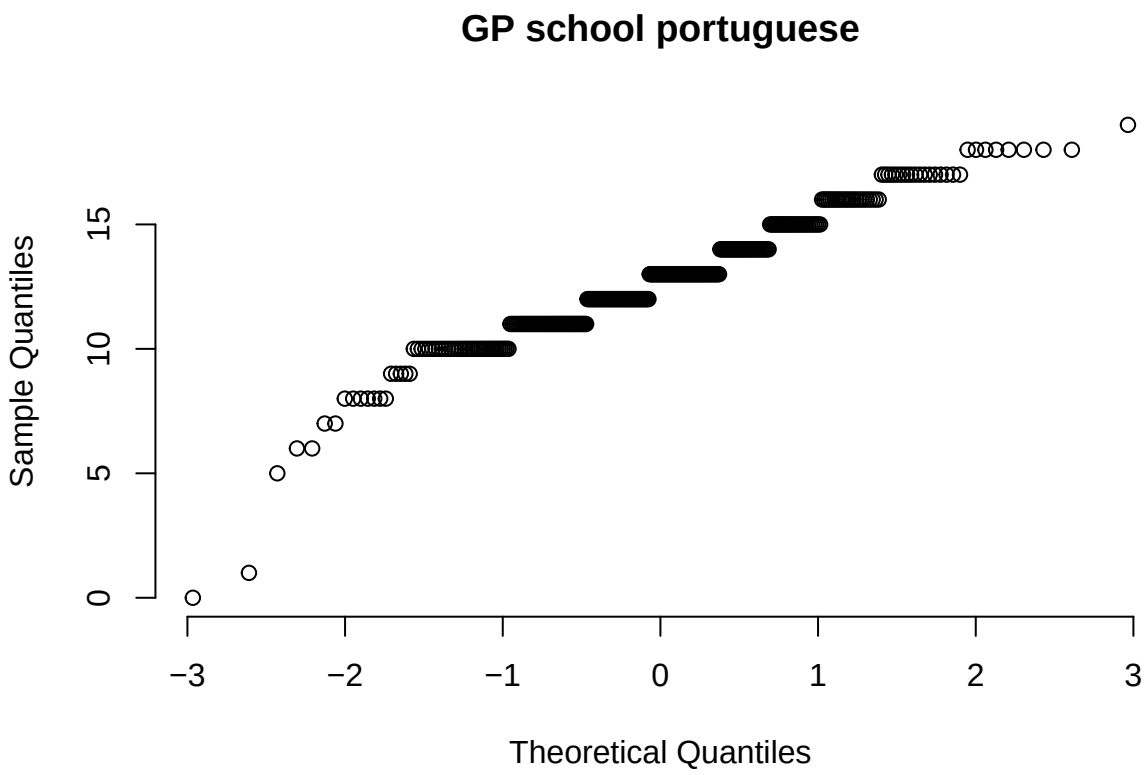

```
# Plot portug -- final grade
ggplot(gp_por, aes(x = grade, y = (..count.. / sum(..count..)))) +
  geom_histogram(bins = 20, aes(color = school, fill = school), alpha = 0.5) +
  geom_histogram(data = ms_por, bins = 20, aes(color = school, fill = school), alpha = 0.3) +
  geom_vline(data = gp_por, aes(xintercept = mean(grade), color = school), linetype = "dashed") +
  geom_vline(data = ms_por, aes(xintercept = mean(grade), color = school), linetype = "dashed") +
  xlab("Grade") +
  ylab("Relative frequency") +
  labs(title = "Portuguese - final grade for each school")
```



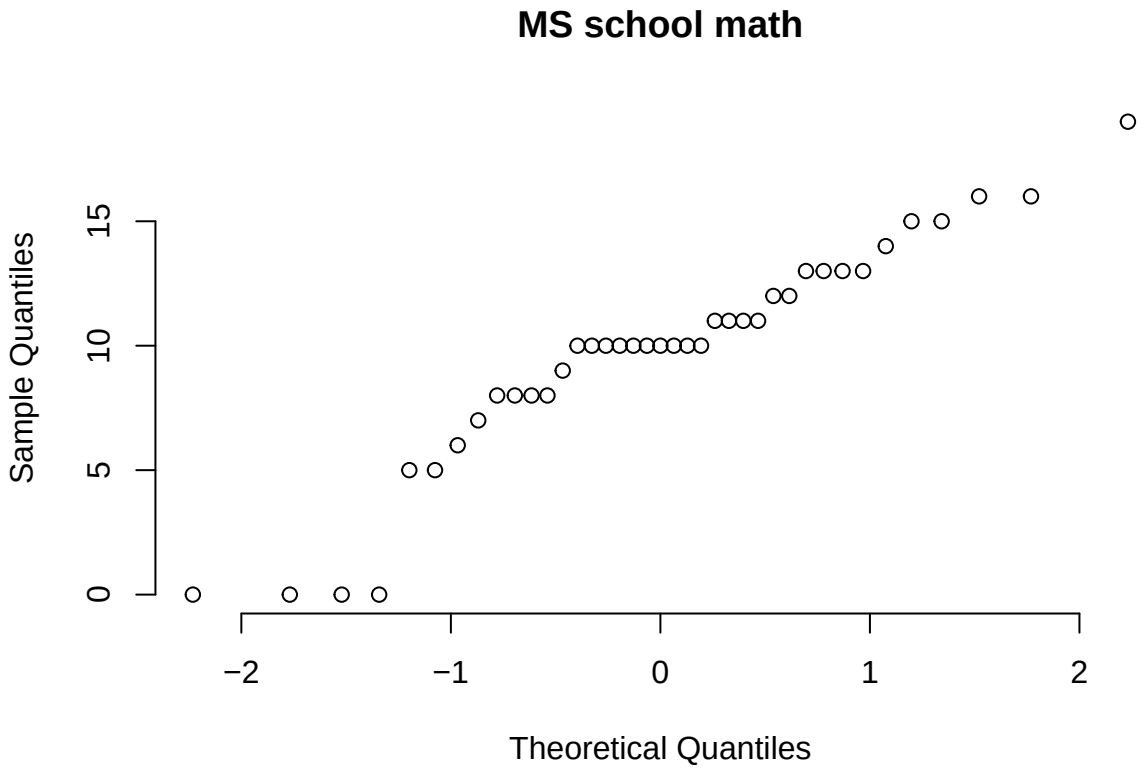

```
## Additional check for normailty is done with qqnorm
qqnorm(gp_mat$grade, pch = 1, frame = FALSE, main = "GP school math")
```



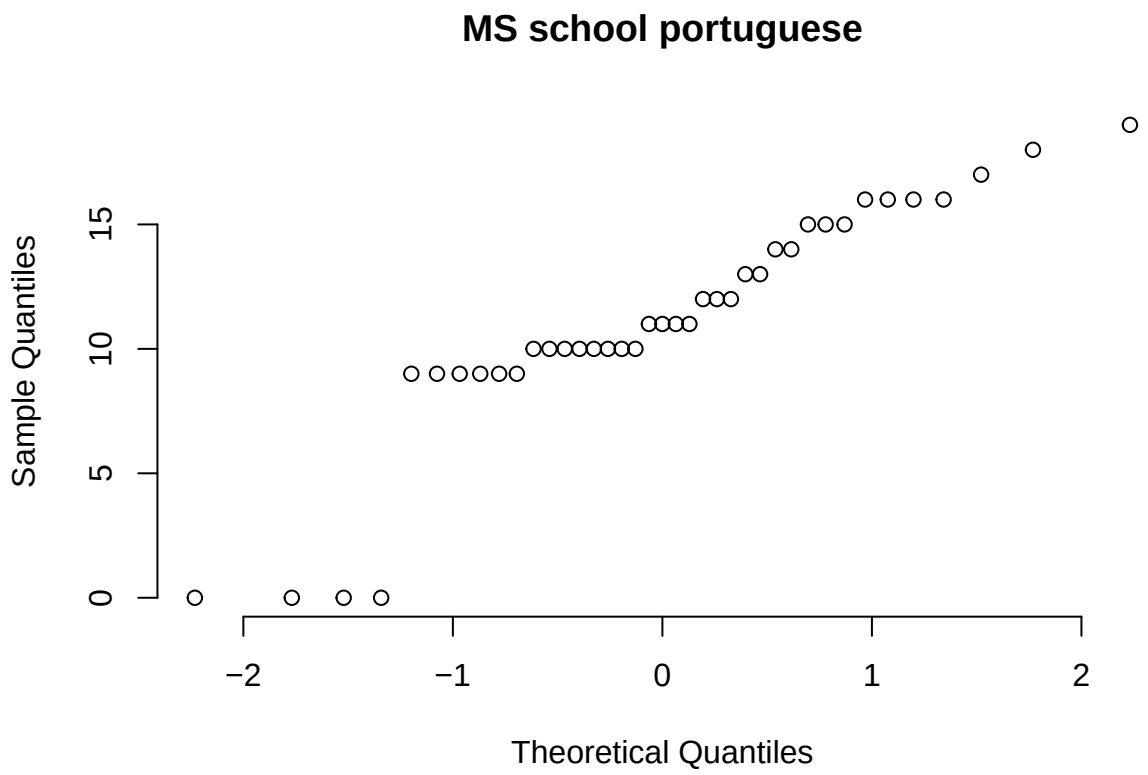
```r
qqnorm(gp_por$grade, pch = 1, frame = FALSE, main = "GP school portuguese")
```




```r
qqnorm(ms_mat$grade, pch = 1, frame = FALSE, main = "MS school math")
```



```
qqnorm(ms_por$grade, pch = 1, frame = FALSE, main = "MS school portuguese")
```



F-test of variance equality

Argument order for F-test of variance equality doesn't matter but in general:

$$\frac{\sigma_1^2}{\sigma_2^2}, \quad \sigma_1^2 > \sigma_2^2$$

p^- probability that under the null hypothesis of obtaining the value (of the test statistic) that's as extreme (or more extreme) than the value we got computed from the sample we have

If $p < \alpha$ we are rejecting the hypothesis $H0$ in favor of $H1$ - falls under right tail => rejection

```
# Let's check variances just as a sanity check:
cat("Mathematics variances", var(gp_mat$grade), var(ms_mat$grade))
```

```
## Mathematics variances 21.38735 19.89204
```

```
cat("Portugeuse variances", var(gp_por$grade), var(ms_por$grade))
```

```
## Portugeuse variances 6.839605 22.1552
```

```
# At first glance, it seems that we will probably reject H0 hypothesis for F-test in the case of Portug
alpha <- 0.05

# H0 - Variance of GP_MAT and MS_MAT are equal
# H1 - not H0
mat_f_test <- var.test(gp_mat$grade, ms_mat$grade, alternative = "two.sided") # F = 1.0752, p = 0.817
mat_f_test
```

```
##
##  F test to compare two variances
##
## data:  gp_mat$grade and ms_mat$grade
## F = 1.0752, num df = 330, denom df = 38, p-value = 0.817
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6349798 1.6605741
## sample estimates:
## ratio of variances
##           1.075171
```

```
# H0 - Variance of GP_POR and MS_MAT are equal
# H1 - not H0
por_f_test <- var.test(gp_por$grade, ms_por$grade, alternative = "two.sided") # F = 1.0752, p = 0.817
por_f_test
```

```
##
##  F test to compare two variances
##
## data:  gp_por$grade and ms_por$grade
## F = 0.30871, num df = 330, denom df = 38, p-value = 1.217e-08
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1823214 0.4767997
## sample estimates:
## ratio of variances
##          0.3087133
```

```
# This part won't be outputed as code in PDF

cat_reject_h0 <- function(prefix_message, is_h0_rejected) {
  cat(prefix_message, "\n")
  if (is_h0_rejected) cat("\tWe are rejecting the H0 hypothesis in favor of H1\n") else cat("\tWe are n
}

var_equal_mat <- if (mat_f_test$p.value < alpha) FALSE else TRUE
cat_reject_h0("For mathemathics variance test:", !var_equal_mat)
```

```
## For mathemathics variance test:
##  We are not rejecting the H0 hypothesis
```

```
var_equal_por <- if (por_f_test$p.value < alpha) FALSE else TRUE
cat_reject_h0("For Portuguese variance test:", !var_equal_por)
```

```
## For Portuguese variance test:
##  We are rejecting the H0 hypothesis in favor of H1
```

```
var_equal_mat
```

```
## [1] TRUE
```

```
var_equal_por
```

```
## [1] FALSE
```

```
# H0 - GP school has equal grades to in mathematics to MS (GP=MS)
# H1 - GP>MS
```

```r
mat_t_test <- t.test(gp_mat$grade, ms_mat$grade, alt = "greater", var.equal = var_equal_mat)
is_gp_mat_better <- if (mat_t_test$p.value < alpha) TRUE else FALSE
cat_reject_h0("For mathemathics T_test test:", is_gp_mat_better)
```

```
## For mathemathics T_test test:
##  We are not rejecting the H0 hypothesis
```

```r
# H0 - GP school has equal grades to in Portuguese to MS (GP=MS)
# H1 - GP>MS
por_t_test <- t.test(gp_por$grade, ms_por$grade, alt = "greater", var.equal = var_equal_por)
is_gp_por_better <- if (por_t_test$p.value < alpha) TRUE else FALSE
cat_reject_h0("For Portuguese T_test test:", is_gp_por_better)
```

```
## For Portuguese T_test test:
##  We are rejecting the H0 hypothesis in favor of H1
```