# Shanghai 2018

```
setwd("~/NCEE")
library(ggplot2)
library(wesanderson)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

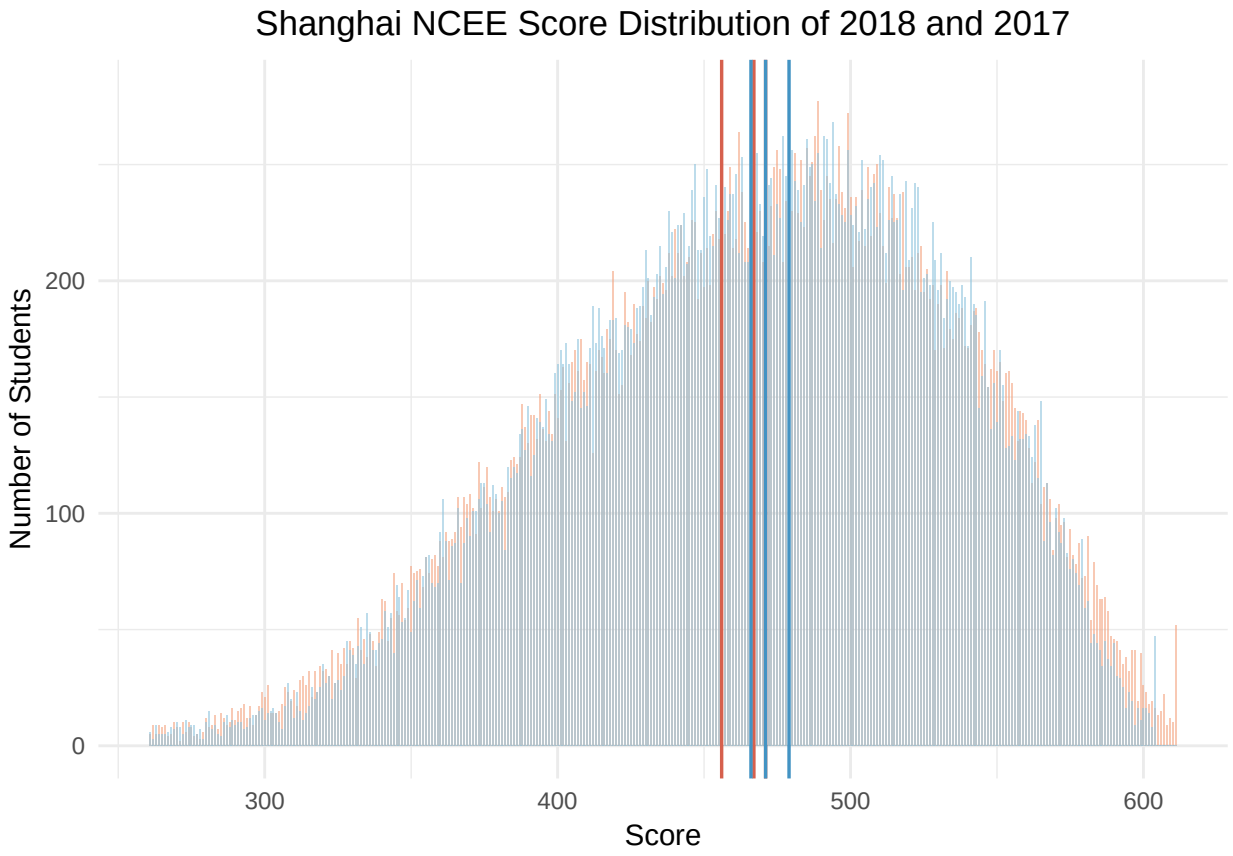
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Shanghai_2018 <- read.csv("Shanghai 2018.csv")[, 1:2]
SH18 <- NA
for (i in 1:nrow(Shanghai_2018)) {
    SH18 <- c(SH18, rep(Shanghai_2018[i, 1], Shanghai_2018[i, 2]))
}
Shanghai_2017 <- read.csv("Shanghai 2017.csv")[, 1:2]
SH17 <- NA
for (i in 1:nrow(Shanghai_2017)) {
    SH17 <- c(SH17, rep(Shanghai_2017[i, 1], Shanghai_2017[i, 2]))
}

score <- c(SH18[-1], SH17[-1])
year <- c(rep(2018, length(SH18) - 1), rep(2017, length(SH17) - 1))
city <- c(rep("Shanghai", length(SH18) - 1), rep("Shanghai", length(SH17) -
    1))

score_table <- data.frame(score, year, city)


mean_2018 <- round(mean(SH18[-1]))
median_2018 <- median(SH18[-1])
lowQ_2018 <- quantile(SH18[-1])[2]
highQ_2018 <- quantile(SH18[-1])[4]
mode_2018 <- Shanghai_2018[which.max(Shanghai_2018$number), ]$grade
mean_2017 <- round(mean(SH17[-1]))
median_2017 <- median(SH17[-1])
lowQ_2017 <- quantile(SH17[-1])[2]
highQ_2017 <- quantile(SH17[-1])[4]
mode_2017 <- Shanghai_2017[which.max(Shanghai_2017$Number), ]$Score
range_18 <- range(SH17[-1])
range_17 <- range(SH17[-1])

statistics <- c("lower range", "25% Q", "mean", "median", "mode", "75% Q", "higher range")
Year_2017 <- c(range_17[1], lowQ_2017, mean_2017, median_2017, mode_2017, highQ_2017,
    range_17[2])
Year_2018 <- c(range_18[1], lowQ_2018, mean_2018, median_2018, mode_2018, highQ_2018,
```

```
    range_18[2])
data.frame(statistics, Year_2017, Year_2018)
```

```
##     statistics Year_2017 Year_2018
## 1  lower range       261       261
## 2         25% Q       421       420
## 3          mean       466       467
## 4        median       471       471
## 5          mode       479       456
## 6         75% Q       516       517
## 7 higher range       604       604
```

```
score_table %>% ggplot(aes(x = score)) + geom_histogram(data = subset(score_table,
    year == "2018"), fill = "#F4A582", binwidth = 0.5, alpha = 0.6) + geom_histogram(data = subset(scor
    year == "2017"), fill = "#92C5DE", binwidth = 0.5, alpha = 0.6) + scale_colour_manual("Density",
    values = c("#F4A582", "#92C5DE")) + geom_vline(xintercept = mean_2018, color = "#D6604D",
    size = 0.6) + geom_vline(xintercept = median_2018, color = "#D6604D", size = 0.6) +
    geom_vline(xintercept = mode_2018, color = "#D6604D", size = 0.6) + geom_vline(xintercept = mean_20
    color = "#4393C3", size = 0.6) + geom_vline(xintercept = median_2017, color = "#4393C3",
    size = 0.6) + geom_vline(xintercept = mode_2017, color = "#4393C3", size = 0.6) +
    theme_minimal() + ggtitle("Shanghai NCEE Score Distribution of 2018 and 2017") +
    theme(plot.title = element_text(hjust = 0.5)) + xlab("Score") + ylab("Number of Students")
```

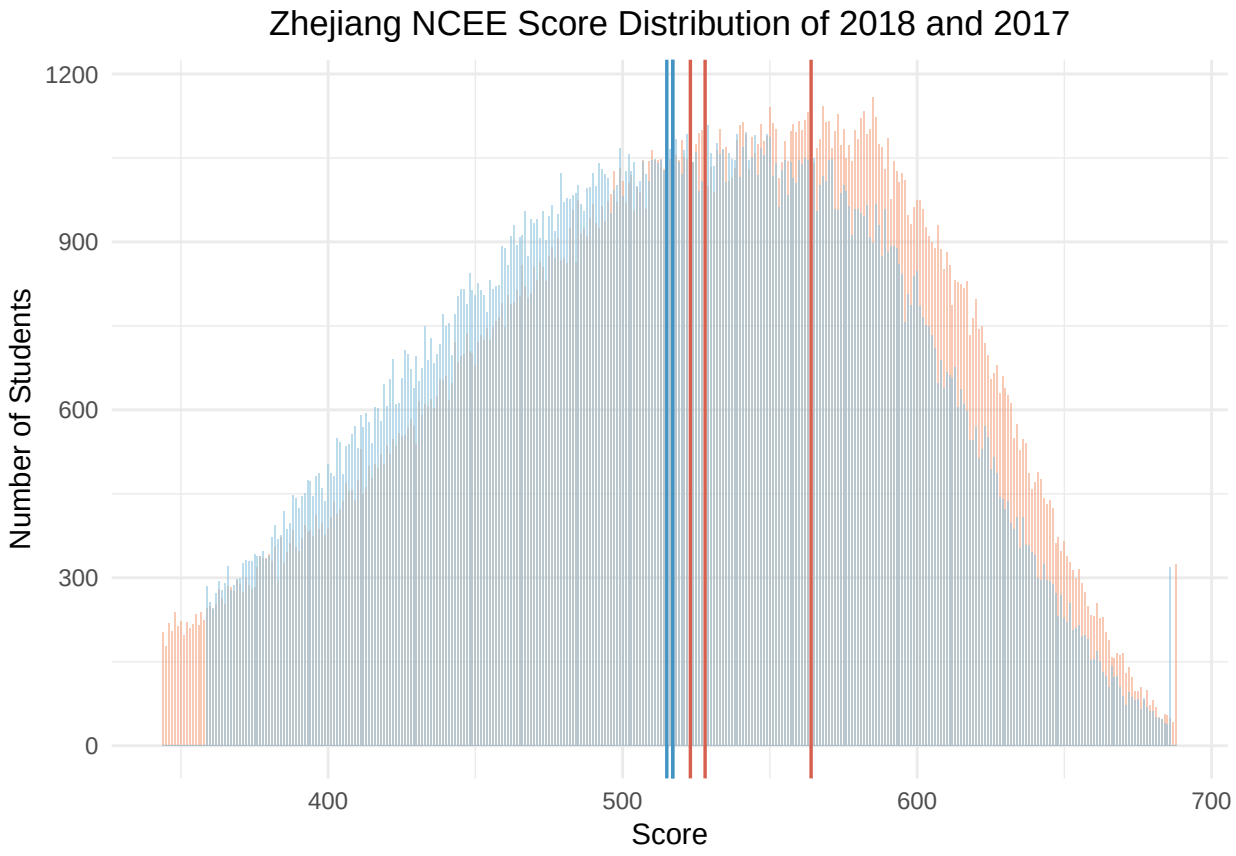



```
Zhejiang_2018 <- read.csv("Zhejiang 2018.csv")[, 1:2]
ZJ18 <- NA
for (i in 1:nrow(Zhejiang_2018)) {
    ZJ18 <- c(ZJ18, rep(Zhejiang_2018[i, 1], Zhejiang_2018[i, 2]))
```

```
}
Zhejiang_2017 <- read.csv("Zhejiang 2017.csv")[, 1:2]
ZJ17 <- NA
for (i in 1:nrow(Zhejiang_2017)) {
    ZJ17 <- c(ZJ17, rep(Zhejiang_2017[i, 1], Zhejiang_2017[i, 2]))
}

score <- c(ZJ18[-1], ZJ17[-1])
year <- c(rep(2018, length(ZJ18) - 1), rep(2017, length(ZJ17) - 1))
city <- c(rep("Zhejiang", length(ZJ18) - 1), rep("Zhejiang", length(ZJ17) -
    1))

score_table <- data.frame(score, year, city)


mean_2018 <- round(mean(ZJ18[-1]))
median_2018 <- median(ZJ18[-1])
lowQ_2018 <- quantile(ZJ18[-1])[2]
highQ_2018 <- quantile(ZJ18[-1])[4]
mode_2018 <- Zhejiang_2018[which.max(Zhejiang_2018$number), ]$score
mean_2017 <- round(mean(ZJ17[-1]))
median_2017 <- median(ZJ17[-1])
lowQ_2017 <- quantile(ZJ17[-1])[2]
highQ_2017 <- quantile(ZJ17[-1])[4]
mode_2017 <- Zhejiang_2017[which.max(Zhejiang_2017$number), ]$score
range_18 <- range(ZJ17[-1])
range_17 <- range(ZJ17[-1])

statistics <- c("lower range", "25% Q", "mean", "median", "mode", "75% Q", "higher range")
Year_2017 <- c(range_17[1], lowQ_2017, mean_2017, median_2017, mode_2017, highQ_2017,
    range_17[2])
Year_2018 <- c(range_18[1], lowQ_2018, mean_2018, median_2018, mode_2018, highQ_2018,
    range_18[2])
data.frame(statistics, Year_2017, Year_2018)
```

```
##     statistics Year_2017 Year_2018
## 1  lower range       359       359
## 2        25% Q       460       468
## 3         mean       515       523
## 4       median       517       528
## 5         mode       517       564
## 6        75% Q       571       583
## 7 higher range       686       686
```

```
score_table %>% ggplot(aes(x = score)) + geom_histogram(data = subset(score_table,
    year == "2018"), fill = "#F4A582", binwidth = 0.5, alpha = 0.6) + geom_histogram(data = subset(scor
    year == "2017"), fill = "#92C5DE", binwidth = 0.5, alpha = 0.6) + scale_colour_manual("Density",
    values = c("#F4A582", "#92C5DE")) + geom_vline(xintercept = mean_2018, color = "#D6604D",
    size = 0.6) + geom_vline(xintercept = median_2018, color = "#D6604D", size = 0.6) +
    geom_vline(xintercept = mode_2018, color = "#D6604D", size = 0.6) + geom_vline(xintercept = mean_20
    color = "#4393C3", size = 0.6) + geom_vline(xintercept = median_2017, color = "#4393C3",
    size = 0.6) + geom_vline(xintercept = mode_2017, color = "#4393C3", size = 0.6) +
    theme_minimal() + ggtitle("Zhejiang NCEE Score Distribution of 2018 and 2017") +
    theme(plot.title = element_text(hjust = 0.5)) + xlab("Score") + ylab("Number of Students")
```


Zhejiang NCEE Score Distribution of 2018 and 2017
Number of Students
1200
900
600
300
0
400
500
600
700
Score