Data Analysis for SD

Load Data

```
data_path <- "../../data/original/train.csv"
voice <- read.csv(data_path)
head(voice)
```

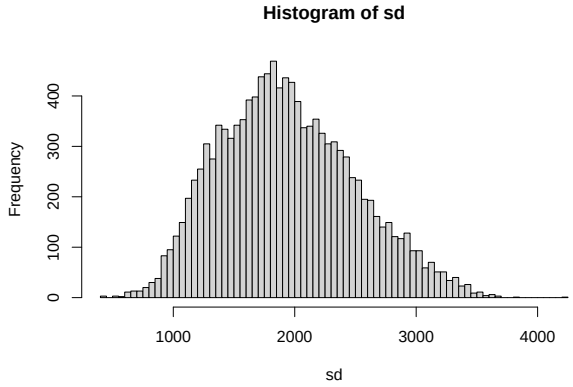
```
##   id meanfreq       sd   median      Q25      Q75      IQR        skew
## 1  0 3521.667 2332.212 2997.294 1660.408 4621.867 2961.459  0.11656897
## 2  1 4189.998 2430.977 4302.741 1832.028 5901.071 4069.043  0.04560770
## 3  2 3154.455 2150.497 2609.968 1460.612 4053.928 2593.316 -0.16147499
## 4  3 4384.338 3029.302 3426.479 1596.072 7283.314 5687.242  0.02416762
## 5  4 4557.150 3158.111 4543.116 1608.165 8074.335 6466.170  0.11711588
## 6  5 4069.004 2983.199 2565.487 1305.284 6961.581 5656.297  0.13049391
##        kurt   sp.ent         sfm     mode centroid  meanfun   minfun   maxfun
## 1 0.9817728 2.308696 0.008450270 1761.333 3521.667 32.33476 153.1934 3995.790
## 2 0.9214181 3.522410 0.022862796 2095.499 4189.998 42.56545 154.0434 3993.462
## 3 0.3882481 2.027891 0.006853276 1577.728 3154.455 26.15712 153.4610 3995.524
## 4 1.4739316 4.823092 0.084471270 2192.669 4384.338 37.56627 153.6399 3994.671
## 5 1.2885699 3.820815 0.100988194 2279.075 4557.150 29.34924 153.8535 3994.646
## 6 0.7668548 3.726702 0.073939204 2035.002 4069.004 29.89368 153.2515 3995.253
##      meandom       mindom    maxdom   dfrange  modindx      age gender accent
## 1 0.06084856 9.842593e-04 194.17128 194.17029 5914.581 twenties female canada
## 2 0.04495757 7.060266e-04 102.27859 102.27788 7693.945 twenties female canada
## 3 0.08144125 2.950821e-04 164.99316 164.99287 5261.606 twenties female canada
## 4 0.01039643 3.165859e-08  29.66787  29.66787 7942.756      nan    nan    nan
## 5 0.01848914 9.267869e-07  85.19259  85.19259 8383.634      nan    nan    nan
## 6 0.01521549 6.052965e-07  32.57839  32.57839 7575.469      nan    nan    nan
```

Visualizing the Data

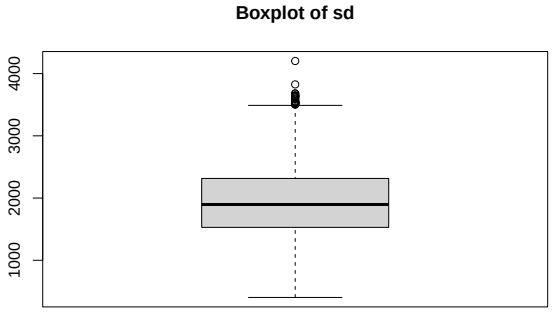
We selected `sd` column to perform the analysis.

First, we load the data and draw a histogram of the `sd` column to get an initial understanding of the data distribution.

```
sd <- voice$sd
hist(sd, breaks = 80, main = "Histogram of sd", xlab = "sd")
```




```
boxplot(sd, main = "Boxplot of sd")
```



Then, we generate the descriptive statistics of the `sd` column.

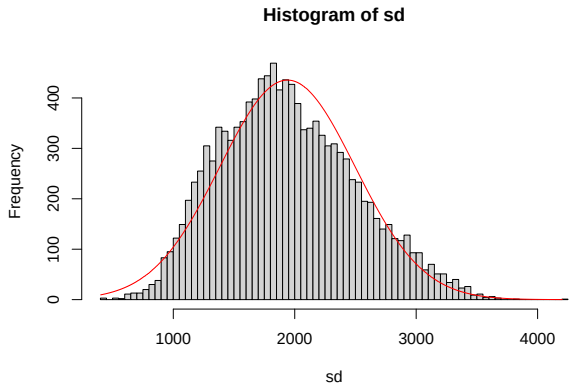
```
library(psych)
describe(sd, type = 1)
```

```
##    vars     n   mean     sd  median trimmed    mad    min     max   range skew
## X1    1 12135 1940.3 555.86 1896.32 1918.18 580.72 402.55 4202.62 3800.07 0.33
##    kurtosis   se
## X1    -0.31 5.05
```

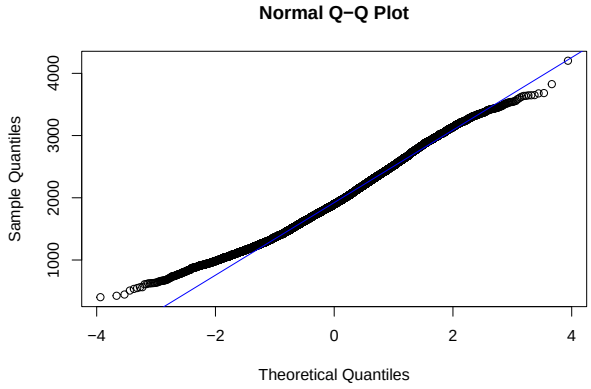
Assessing Data Normality

After that, we assess the normality of the `sd` column by drawing a histogram with a normal curve and a Q-Q plot.

```
hist(sd, breaks = 80, main = "Histogram of sd", xlab = "sd")
# impose a normal curve on the histogram
xpt <- seq(402, 4203, by = 0.1)
n_den <- dnorm(xpt, mean = mean(sd), sd = sd(sd))
ypt <- n_den * length(sd) * 50
lines(xpt, ypt, col = "red")
```



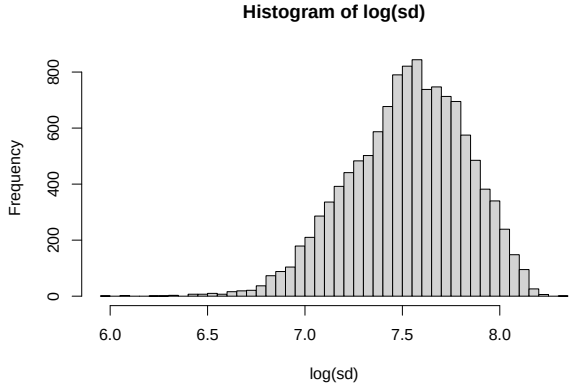

```
qqnorm(sd)
qqline(sd, col = "blue")
```



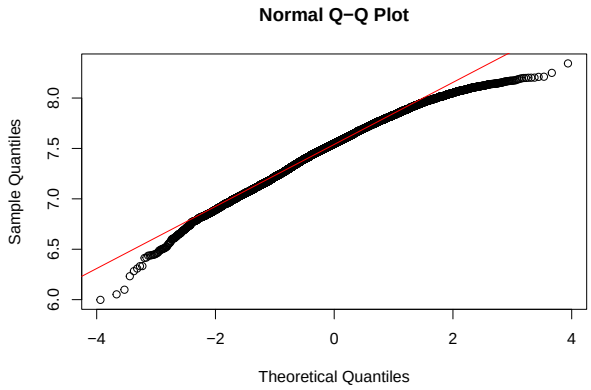
Transformation

We found that the data is almost normally distributed, but not perfect. We tried to log-transform the data to see if it can be improved.

```
sd_trans <- log(sd)
hist(sd_trans, breaks = 80, main = "Histogram of log(sd)", xlab = "log(sd)")
```



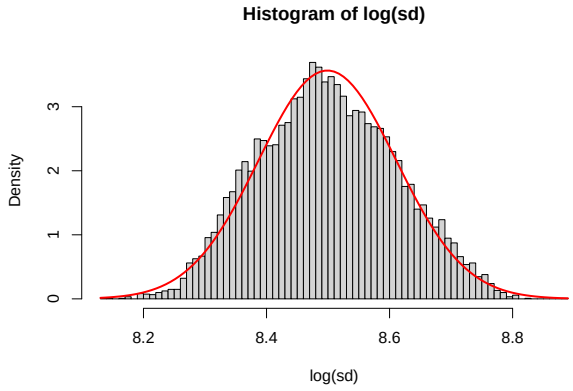
```r
qqnorm(sd_trans)
qqline(sd_trans, col = "red")
```



We observed that after log-transformation, the data's fit to a normal distribution did not improve as expected.

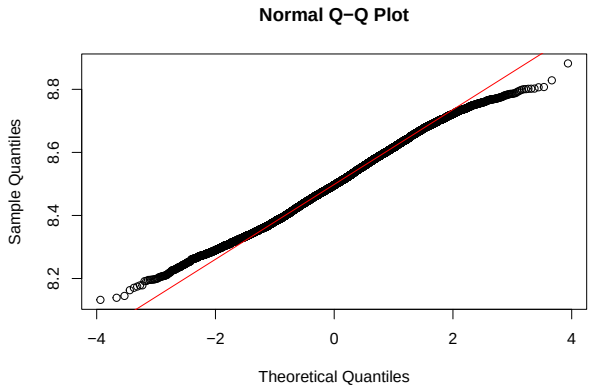
Therefore, we explored an alternative transformation: $y = \log(x + 3000)$

```r
sd_trans <- log(sd + 3000)

hist(
  sd_trans,
  breaks = 80,
  main = "Histogram of log(sd)",
  xlab = "log(sd)",
  probability = TRUE
)

curve(
  dnorm(x, mean = mean(sd_trans), sd = sd(sd_trans)),
  col = "red",
  lwd = 2,
  add = TRUE
)
```



And we checked the Q-Q plot of the transformed data.

```
qqnorm(sd_trans)
qqline(sd_trans, col = "red")
```



Finally, we calculated the descriptive statistics of the transformed data.

```
describe(sd_trans)
```

```
##    vars     n mean   sd median trimmed  mad  min  max range skew kurtosis se
## X1    1 12135  8.5 0.11    8.5     8.5 0.12 8.13 8.88  0.75 0.07    -0.44  0
```