Analysis

Contents

Data Preparation

```r
voice = read.csv("../../data/original/train.csv")
voice = voice[voice$age != "nan", ]
voice = voice[voice$gender == "male", ]
AGE_LEVELS = c("teens", "twenties", "thirties", "fourties", "fifties", "sixties", "seventies")
voice$age <- factor(voice$age, levels = AGE_LEVELS)
head(voice)
```

```
##    id meanfreq       sd   median      Q25      Q75      IQR       skew
## 14 13 2395.584 1425.829 1883.654 1431.296 2935.433 1504.137  0.2293413
## 15 14 5012.665 2963.707 6353.490 1705.367 7848.439 6143.072 -0.3733059
## 19 18 2749.071 1526.183 2482.678 1382.738 4077.986 2695.248 -0.1231436
## 20 19 2687.434 1634.787 2216.735 1153.558 4409.890 3256.332 -0.1546336
## 21 20 2807.396 1640.620 2434.370 1410.506 4392.110 2981.604 -0.3097823
## 22 21 2434.442 1433.957 1887.645 1275.208 3693.556 2418.347 -0.2732455
##         kurt   sp.ent         sfm     mode centroid  meanfun   minfun   maxfun
## 14 0.6316582 1.346211 0.003008815 1198.292 2395.584 21.29166 153.1219 3994.158
## 15 2.7109194 3.961114 0.110746250 2506.832 5012.665 50.10727 152.4209 3995.314
## 19 0.2052341 2.140214 0.009356702 1375.035 2749.071 33.03064 152.9522 3996.042
## 20 0.3491321 2.139938 0.013563351 1344.217 2687.434 34.00013 152.5279 3995.969
## 21 0.5242957 2.087635 0.050972614 1404.198 2807.396 31.79936 152.8019 3994.976
## 22 0.2360166 2.084844 0.009249036 1217.721 2434.442 31.32352 152.7743 3995.663
##       meandom       mindom    maxdom   dfrange  modindx   age gender accent
## 14 0.02383211 7.181296e-11  39.08173  39.08173 4100.888 teens   male     us
## 15 0.01081940 2.593320e-06  28.00432  28.00431 9699.244 teens   male     us
## 19 0.04023470 1.147567e-07  94.49204  94.49204 5262.030 teens   male    nan
## 20 0.02115684 2.354625e-05  51.08478  51.08476 5394.165 teens   male    nan
## 21 0.03580633 6.213739e-07 122.87537 122.87537 5617.723 teens   male    nan
## 22 0.06437463 4.170119e-05 140.29124 140.29120 4863.189 teens   male    nan
```

```
check_normality <- function(column) {
  for (age in AGE_LEVELS) {
    data_age = voice[voice$age == age, ]
    data_age = data_age[[column]]
    result = shapiro.test(data_age)$p.value
    print(paste("Shapiro-Wilk test for", age, ":", result))
  }
}
```

Data Analysis

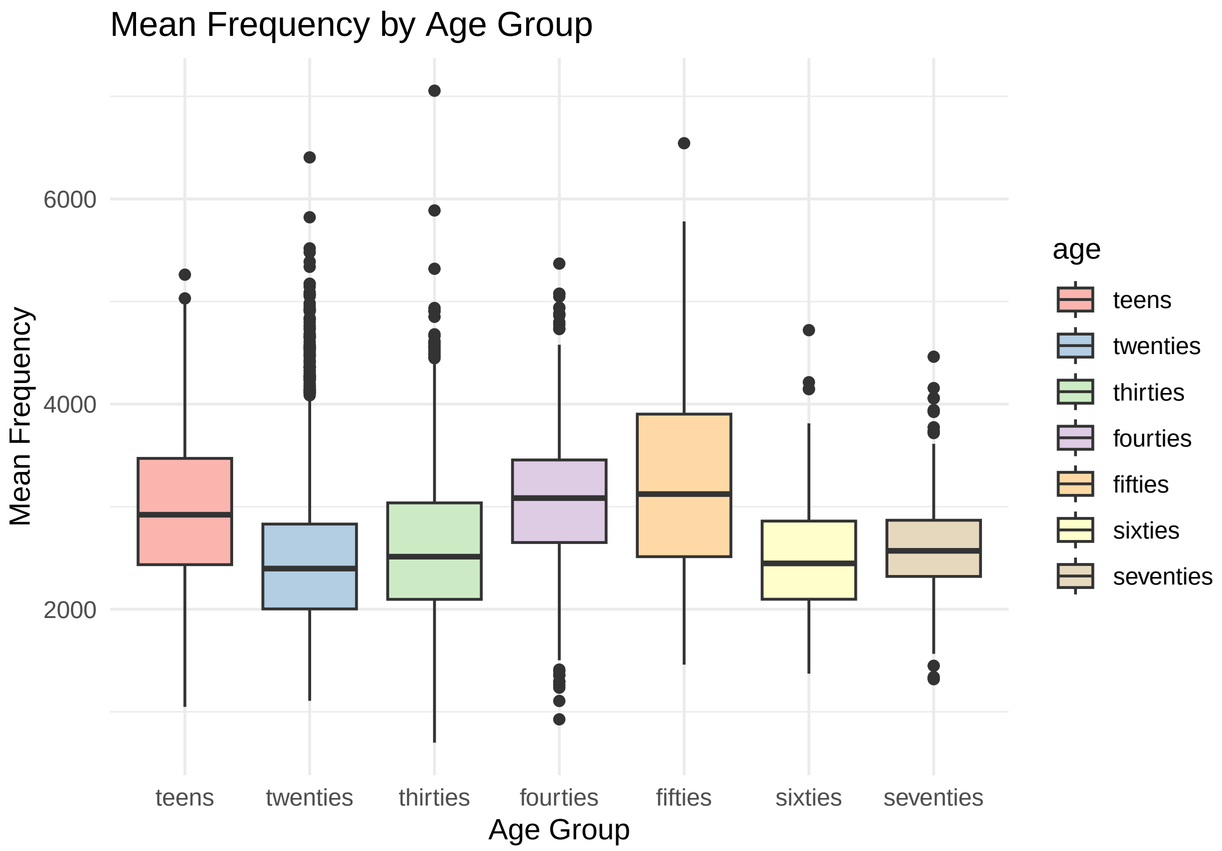
Mean Frequency

Visualizing the data

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
ggplot(voice, aes(x = age, y = meanfreq, fill = age)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "Mean Frequency by Age Group", x = "Age Group", y = "Mean Frequency") +
  theme_minimal()
```



Check for normality

```
check_normality("meanfreq")
```

```
## [1] "Shapiro-Wilk test for teens : 0.000206087522582567"
## [1] "Shapiro-Wilk test for twenties : 2.46132409846886e-33"
## [1] "Shapiro-Wilk test for thirties : 3.43165344350297e-23"
## [1] "Shapiro-Wilk test for fourties : 0.000151879311986604"
## [1] "Shapiro-Wilk test for fifties : 4.59719709834552e-07"
## [1] "Shapiro-Wilk test for sixties : 0.00119644475752247"
## [1] "Shapiro-Wilk test for seventies : 0.000227803277640362"
```

Kruskal-Wallis Test

```
kruskal.test(meanfreq ~ age, data = voice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  meanfreq by age
## Kruskal-Wallis chi-squared = 1123.9, df = 6, p-value < 2.2e-16
```

Pairwise Wilcoxon Test

```
pairwise.wilcox.test(voice$meanfreq, voice$age, p.adjust.method = "BH")
```

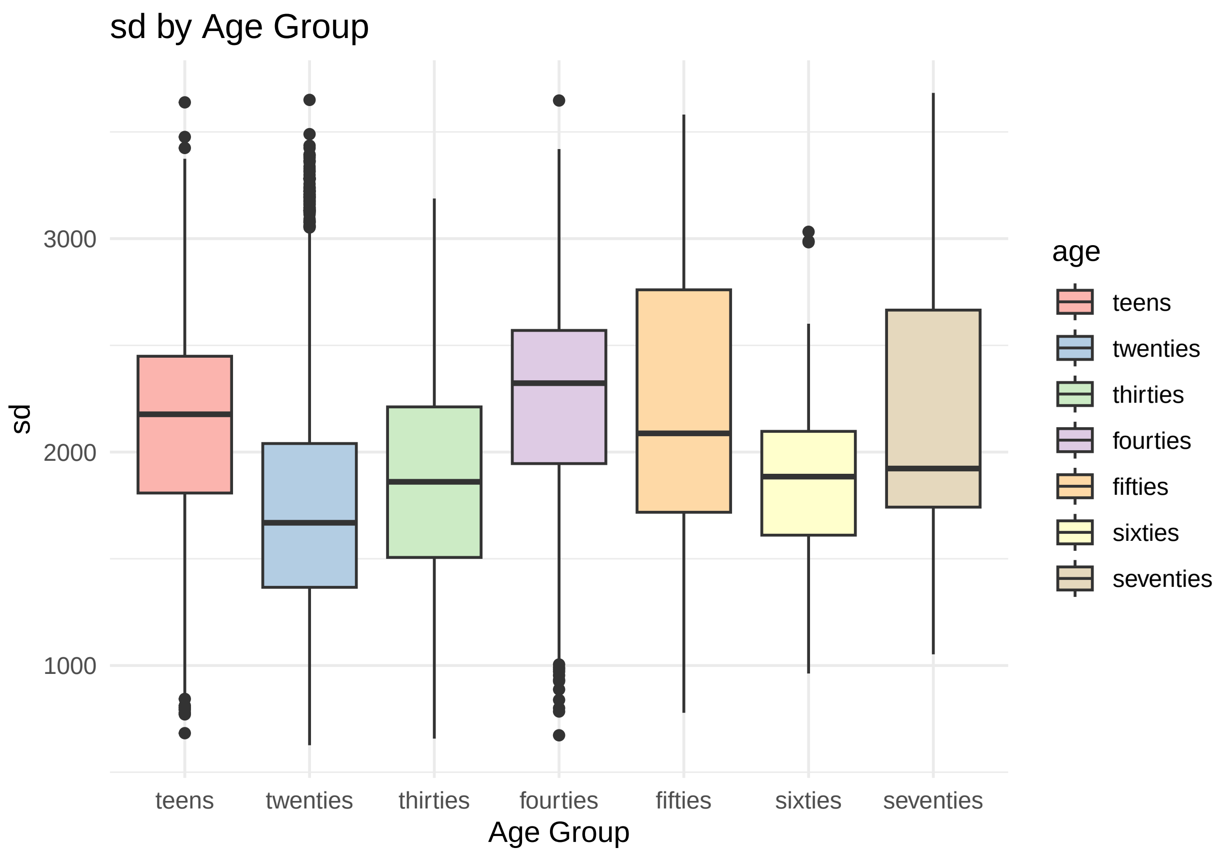
```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  voice$meanfreq and voice$age
##
##           teens   twenties thirties fourties fifties sixties
## twenties  < 2e-16 -        -        -        -       -
## thirties  < 2e-16 1.1e-12  -        -        -       -
## fourties  0.00047 < 2e-16  < 2e-16  -        -       -
## fifties   1.9e-07 < 2e-16  < 2e-16  0.00083  -       -
## sixties   1.7e-14 0.42174  0.04525  < 2e-16  < 2e-16 -
## seventies 3.6e-16 2.9e-08  0.23866  < 2e-16  < 2e-16 0.00664
##
## P value adjustment method: BH
```

Standard Deviation

Visualizing the data

```
library(ggplot2)

ggplot(voice, aes(x = age, y = sd, fill = age)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "sd by Age Group", x = "Age Group", y = "sd") +
  theme_minimal()
```



Check for normality

```
check_normality("sd")
```

```
## [1] "Shapiro-Wilk test for teens : 3.12540191275697e-07"
## [1] "Shapiro-Wilk test for twenties : 1.12013417763867e-28"
## [1] "Shapiro-Wilk test for thirties : 1.5441243466243e-11"
## [1] "Shapiro-Wilk test for fourties : 8.77439432933489e-23"
## [1] "Shapiro-Wilk test for fifties : 5.49672435341787e-12"
## [1] "Shapiro-Wilk test for sixties : 0.00619069396044784"
## [1] "Shapiro-Wilk test for seventies : 6.99563637915044e-18"
```

Kruskal-Wallis Test

```
kruskal.test(sd ~ age, data = voice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  sd by age
## Kruskal-Wallis chi-squared = 1254, df = 6, p-value < 2.2e-16
```

Pairwise Wilcoxon Test

```
pairwise.wilcox.test(voice$sd, voice$age, p.adjust.method = "BH")
```

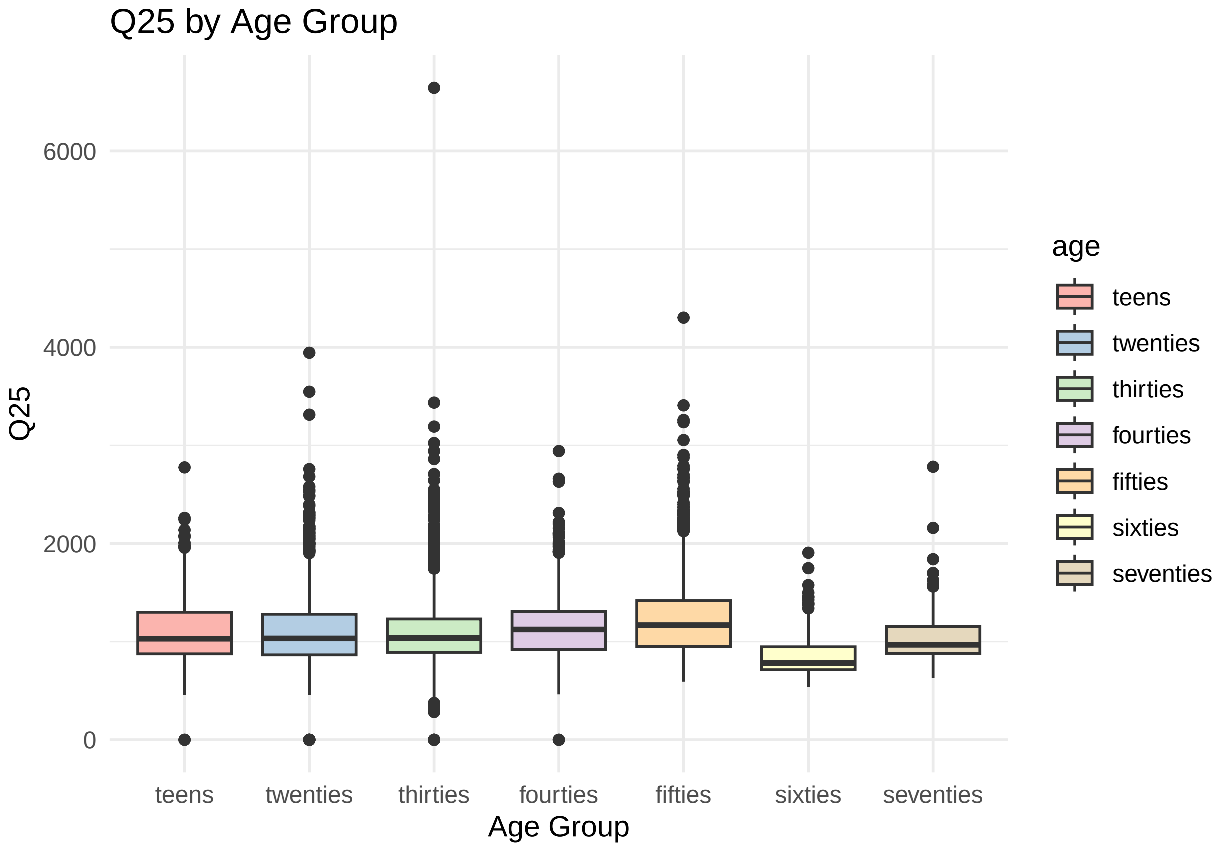
```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  voice$sd and voice$age
##
##           teens   twenties thirties fourties fifties sixties
## twenties  < 2e-16 -        -        -        -       -
## thirties  < 2e-16 < 2e-16  -        -        -       -
## fourties  5.6e-06 < 2e-16  < 2e-16  -        -       -
## fifties   0.8433  < 2e-16  < 2e-16  0.0291   -       -
## sixties   6.2e-15 6.8e-06  0.6443   < 2e-16  1.1e-10 -
## seventies 0.0053  < 2e-16  3.5e-12  7.3e-08  0.0725  4.4e-05
##
## P value adjustment method: BH
```

Q25

Visualizing the data

```
library(ggplot2)

ggplot(voice, aes(x = age, y = Q25, fill = age)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "Q25 by Age Group", x = "Age Group", y = "Q25") +
  theme_minimal()
```



Check for normality

```r
check_normality("Q25")
```

```
## [1] "Shapiro-Wilk test for teens : 1.21472348440465e-13"
## [1] "Shapiro-Wilk test for twenties : 9.33195036404546e-36"
## [1] "Shapiro-Wilk test for thirties : 3.20374994943887e-43"
## [1] "Shapiro-Wilk test for fourties : 2.74194675313038e-17"
## [1] "Shapiro-Wilk test for fifties : 6.31465727775898e-22"
## [1] "Shapiro-Wilk test for sixties : 9.92813083401321e-13"
## [1] "Shapiro-Wilk test for seventies : 1.68633323756025e-17"
```

Kruskal-Wallis Test

```r
kruskal.test(Q25 ~ age, data = voice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Q25 by age
## Kruskal-Wallis chi-squared = 306.77, df = 6, p-value < 2.2e-16
```

Pairwise Wilcoxon Test

```r
pairwise.wilcox.test(voice$Q25, voice$age, p.adjust.method = "BH")
```

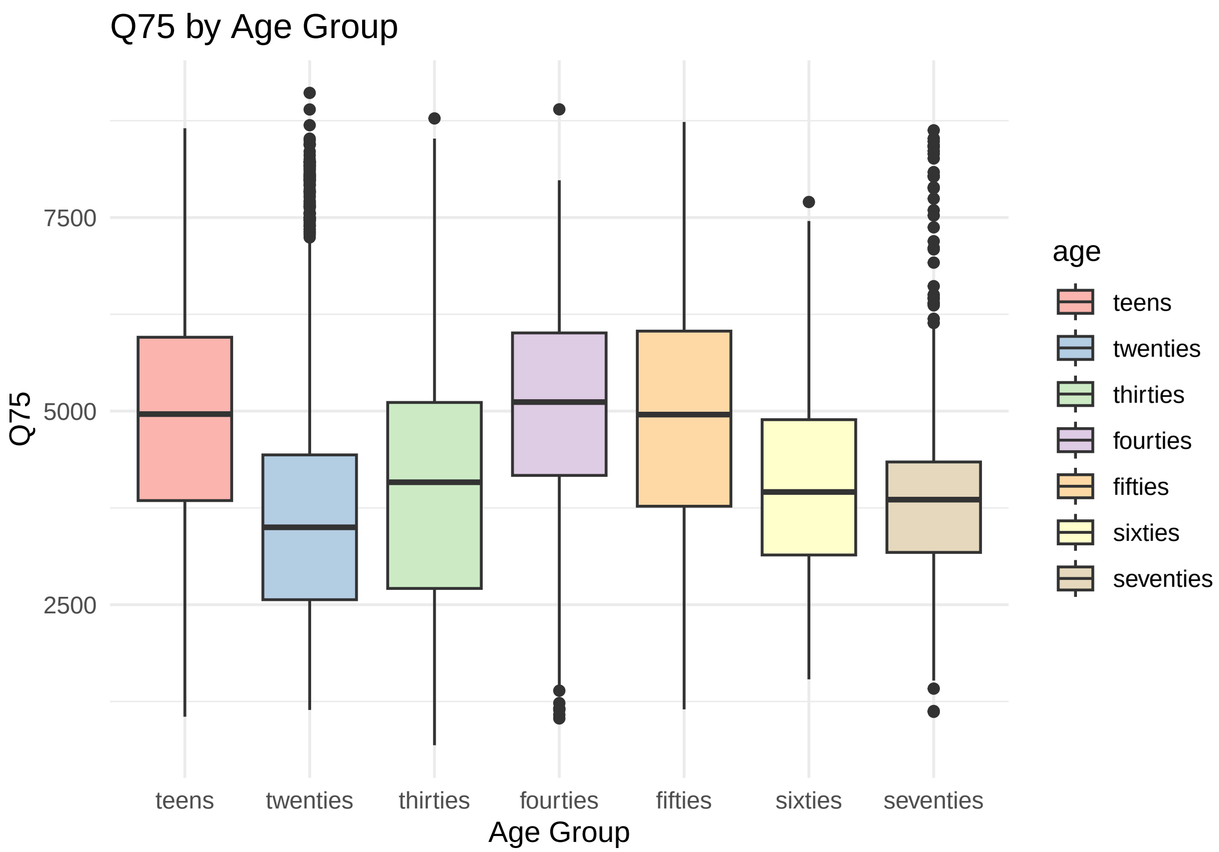
```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  voice$Q25 and voice$age
##
##           teens   twenties thirties fourties fifties sixties
## twenties  0.68134 -        -        -        -       -
## thirties  0.81332 0.76215  -        -        -       -
## fourties  0.00044 1.3e-10  3.4e-11  -        -       -
## fifties   5.8e-08 5.5e-14  2.0e-14  0.00040  -       -
## sixties   < 2e-16 < 2e-16  < 2e-16  < 2e-16  < 2e-16 -
## seventies 0.00094 0.00050  8.7e-05  2.7e-15  < 2e-16 < 2e-16
##
## P value adjustment method: BH
```

Q75

Visualizing the data

```r
library(ggplot2)

ggplot(voice, aes(x = age, y = Q75, fill = age)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "Q75 by Age Group", x = "Age Group", y = "Q75") +
  theme_minimal()
```



Check for normality

```
check_normality("Q75")
```

```
## [1] "Shapiro-Wilk test for teens : 6.40176613621252e-07"
## [1] "Shapiro-Wilk test for twenties : 1.91053310839091e-34"
## [1] "Shapiro-Wilk test for thirties : 4.35642283195017e-19"
## [1] "Shapiro-Wilk test for fourties : 3.02808814200532e-15"
## [1] "Shapiro-Wilk test for fifties : 1.4792237261905e-10"
## [1] "Shapiro-Wilk test for sixties : 5.44525158426606e-05"
## [1] "Shapiro-Wilk test for seventies : 7.37690270028397e-17"
```

Kruskal-Wallis Test

```
kruskal.test(Q75 ~ age, data = voice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Q75 by age
## Kruskal-Wallis chi-squared = 1271.6, df = 6, p-value < 2.2e-16
```

Pairwise Wilcoxon Test

```
pairwise.wilcox.test(voice$Q75, voice$age, p.adjust.method = "BH")
```

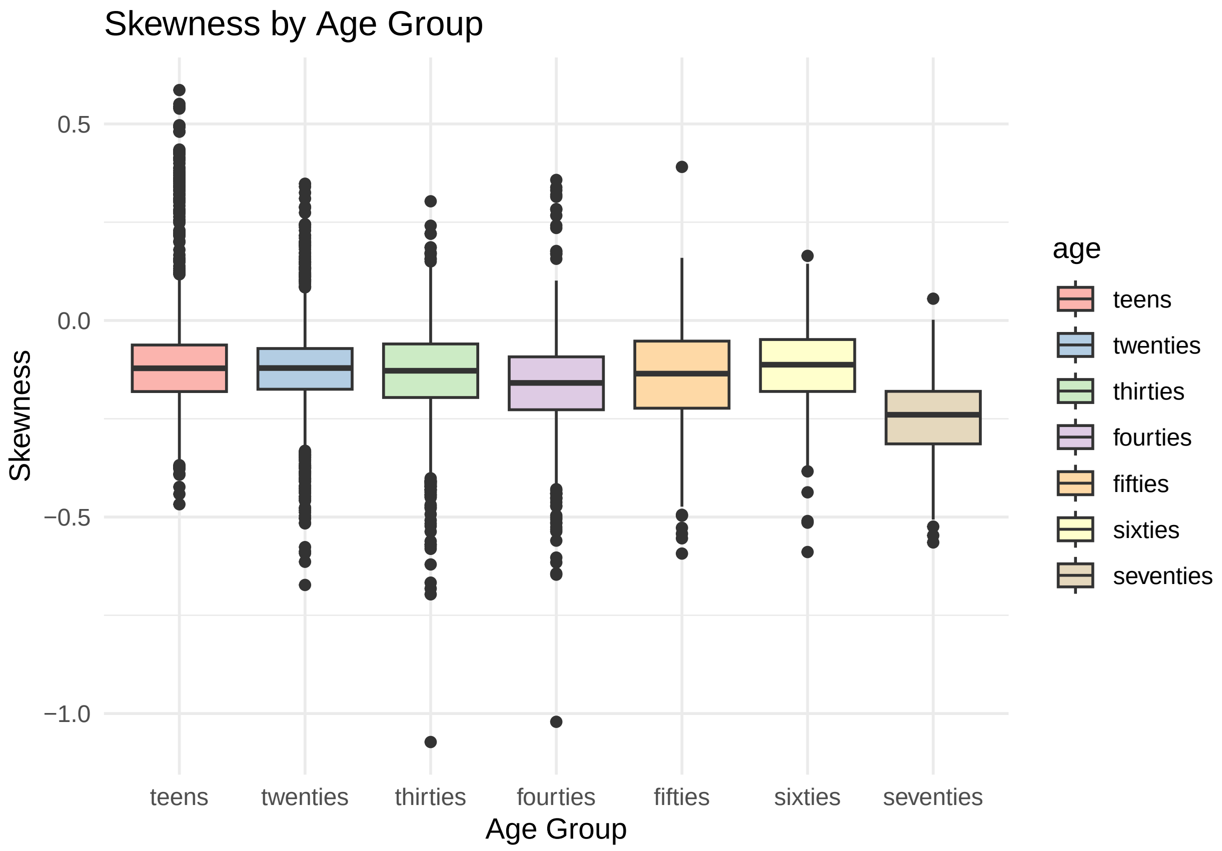
```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  voice$Q75 and voice$age
##
##           teens   twenties thirties fourties fifties sixties
## twenties  < 2e-16 -        -        -        -       -
## thirties  < 2e-16 < 2e-16  -        -        -       -
## fourties  0.025   < 2e-16  < 2e-16  -        -       -
## fifties   0.249   < 2e-16  < 2e-16  0.522    -       -
## sixties   1.9e-14 2.2e-06  0.784    < 2e-16  1.8e-15 -
## seventies < 2e-16 5.4e-07  0.049    < 2e-16  < 2e-16 0.109
##
## P value adjustment method: BH
```

Skewness

Visualizing the data

```
library(ggplot2)

ggplot(voice, aes(x = age, y = skew, fill = age)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "Skewness by Age Group", x = "Age Group", y = "Skewness") +
  theme_minimal()
```



Check for normality

```r
check_normality("skew")
```

```
## [1] "Shapiro-Wilk test for teens : 2.31185240438444e-24"
## [1] "Shapiro-Wilk test for twenties : 8.87168337857448e-30"
## [1] "Shapiro-Wilk test for thirties : 7.29709289949265e-21"
## [1] "Shapiro-Wilk test for fourties : 8.37504982779769e-22"
## [1] "Shapiro-Wilk test for fifties : 2.79238444807594e-06"
## [1] "Shapiro-Wilk test for sixties : 4.10588904555974e-06"
## [1] "Shapiro-Wilk test for seventies : 0.433277489767555"
```

Kruskal-Wallis Test

```r
kruskal.test(skew ~ age, data = voice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  skew by age
## Kruskal-Wallis chi-squared = 610.35, df = 6, p-value < 2.2e-16
```

Pairwise Wilcoxon Test

```r
pairwise.wilcox.test(voice$skew, voice$age, p.adjust.method = "BH")
```

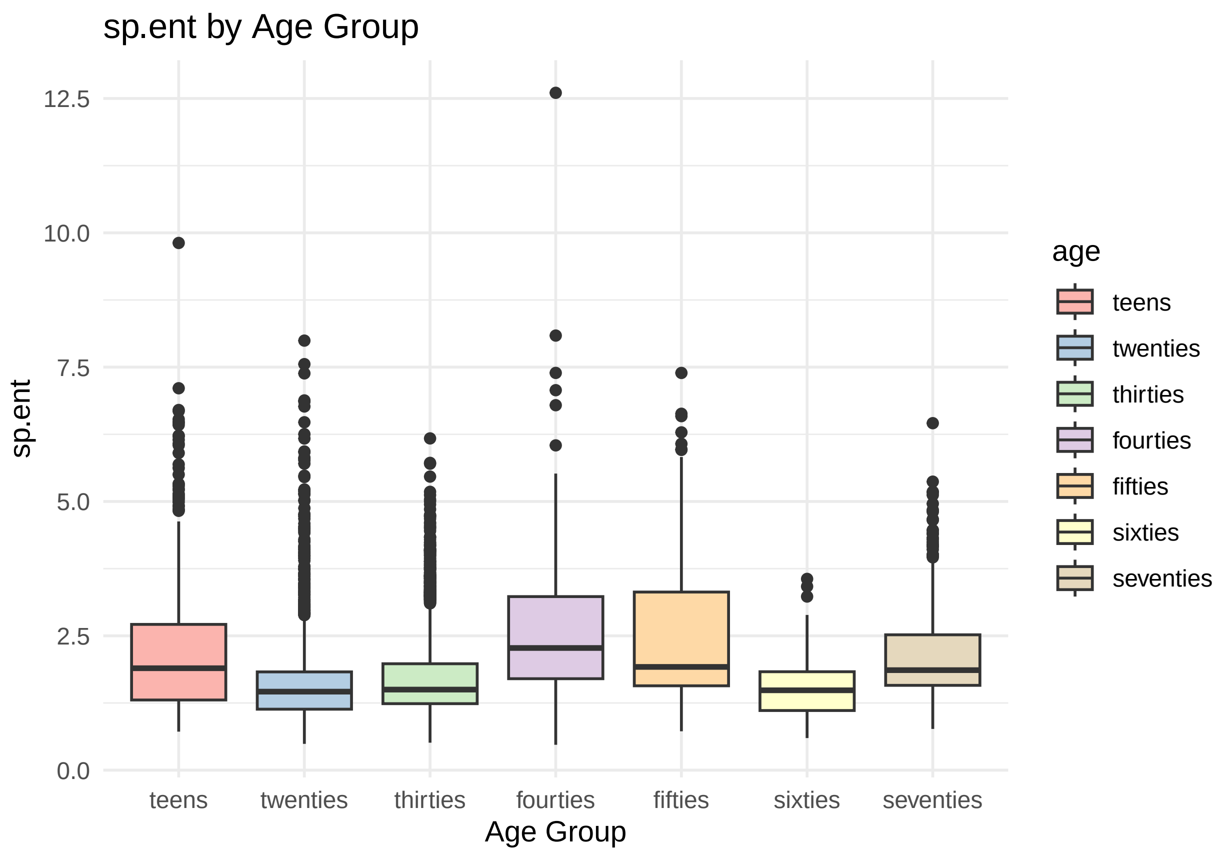
```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  voice$skew and voice$age
##
##           teens   twenties thirties fourties fifties sixties
## twenties  0.2082  -        -        -        -       -
## thirties  0.0061  0.0390   -        -        -       -
## fourties  < 2e-16 < 2e-16  < 2e-16  -        -       -
## fifties   0.0012  0.0116   0.1451   1.4e-05  -       -
## sixties   0.9883  0.2082   0.1056   4.3e-09  0.0390  -
## seventies < 2e-16 < 2e-16  < 2e-16  < 2e-16  < 2e-16 < 2e-16
##
## P value adjustment method: BH
```

sp.ent

Visualizing the data

```r
library(ggplot2)

ggplot(voice, aes(x = age, y = sp.ent, fill = age)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "sp.ent by Age Group", x = "Age Group", y = "sp.ent") +
  theme_minimal()
```



Check for normality

```
check_normality("sp.ent")
```

```
## [1] "Shapiro-Wilk test for teens : 8.88667769534423e-24"
## [1] "Shapiro-Wilk test for twenties : 4.66764665333775e-54"
## [1] "Shapiro-Wilk test for thirties : 5.53907896228274e-44"
## [1] "Shapiro-Wilk test for fourties : 1.02526163050789e-27"
## [1] "Shapiro-Wilk test for fifties : 3.22477664255166e-19"
## [1] "Shapiro-Wilk test for sixties : 7.20902708092716e-06"
## [1] "Shapiro-Wilk test for seventies : 1.02076535287444e-19"
```

Kruskal-Wallis Test

```
kruskal.test(sp.ent ~ age, data = voice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  sp.ent by age
## Kruskal-Wallis chi-squared = 1542.7, df = 6, p-value < 2.2e-16
```

Pairwise Wilcoxon Test

```
pairwise.wilcox.test(voice$sp.ent, voice$age, p.adjust.method = "BH")
```

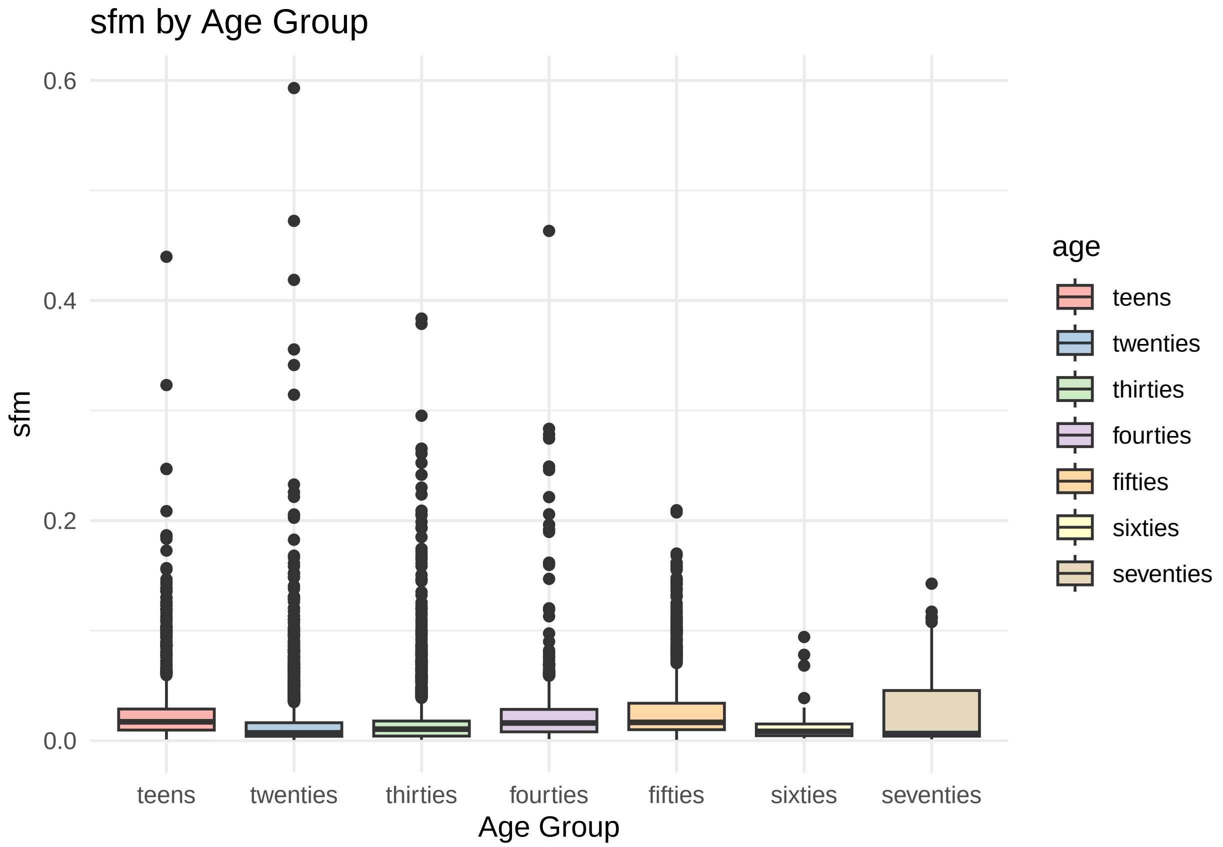
```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  voice$sp.ent and voice$age
##
##           teens   twenties thirties fourties fifties sixties
## twenties  < 2e-16 -        -        -        -       -
## thirties  < 2e-16 8.0e-12  -        -        -       -
## fourties  4.6e-16 < 2e-16  < 2e-16  -        -       -
## fifties   8.9e-06 < 2e-16  < 2e-16  0.0029   -       -
## sixties   1.9e-14 0.7548   0.0045   < 2e-16  < 2e-16 -
## seventies 0.0460  < 2e-16  < 2e-16  6.7e-11  0.0371  < 2e-16
##
## P value adjustment method: BH
```

sfm

Visualizing the data

```
library(ggplot2)

ggplot(voice, aes(x = age, y = sfm, fill = age)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "sfm by Age Group", x = "Age Group", y = "sfm") +
  theme_minimal()
```



Check for normality

```r
check_normality("sfm")
```

```
## [1] "Shapiro-Wilk test for teens : 5.22382547184172e-35"
## [1] "Shapiro-Wilk test for twenties : 1.05517167167322e-73"
## [1] "Shapiro-Wilk test for thirties : 2.68519213405542e-65"
## [1] "Shapiro-Wilk test for fourties : 7.43701410396359e-56"
## [1] "Shapiro-Wilk test for fifties : 1.27045223675796e-29"
## [1] "Shapiro-Wilk test for sixties : 3.26248063856963e-21"
## [1] "Shapiro-Wilk test for seventies : 5.57807551364813e-26"
```

Kruskal-Wallis Test

```r
kruskal.test(sfm ~ age, data = voice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  sfm by age
## Kruskal-Wallis chi-squared = 806.25, df = 6, p-value < 2.2e-16
```

Pairwise Wilcoxon Test

```r
pairwise.wilcox.test(voice$sfm, voice$age, p.adjust.method = "BH")
```

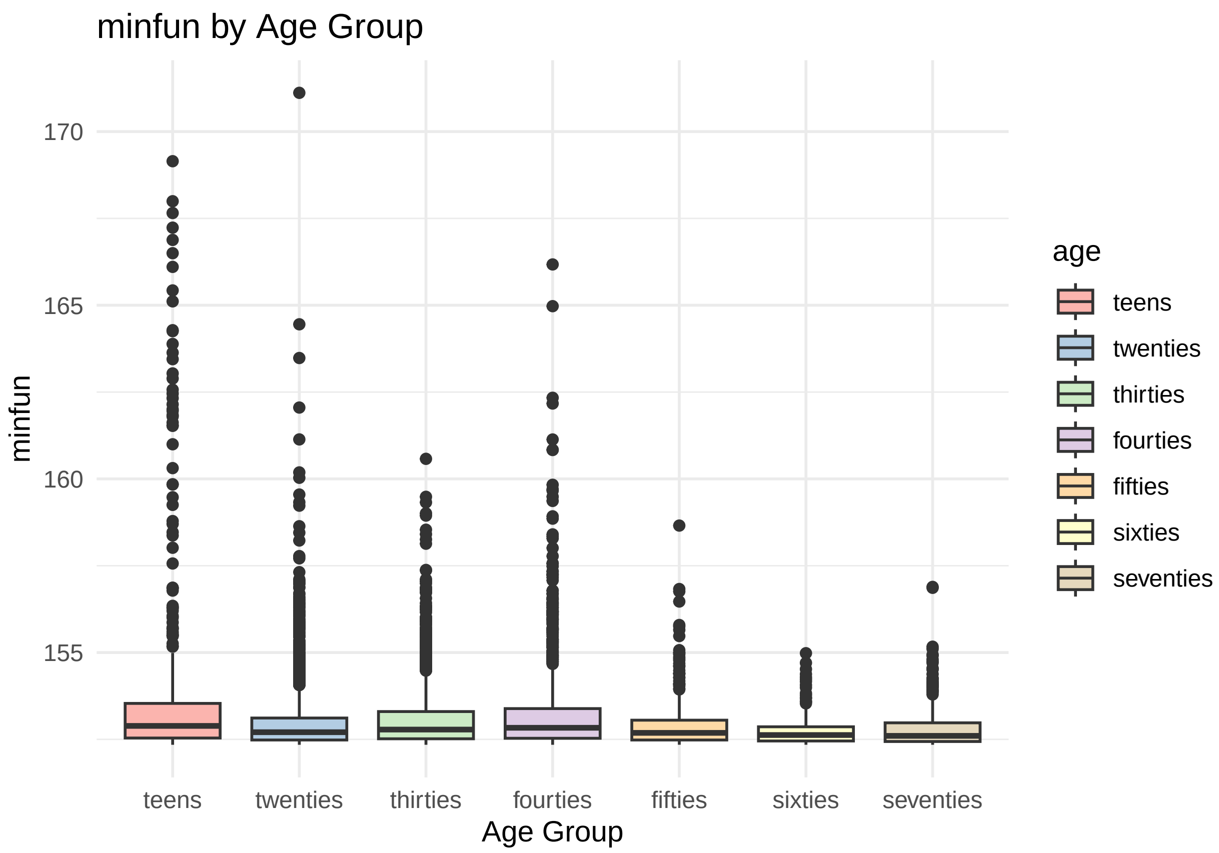
```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  voice$sfm and voice$age
##
##           teens   twenties thirties fourties fifties sixties
## twenties  < 2e-16 -        -        -        -       -
## thirties  < 2e-16 2.9e-10  -        -        -       -
## fourties  0.2501  < 2e-16  < 2e-16  -        -       -
## fifties   0.3911  < 2e-16  < 2e-16  0.0037   -       -
## sixties   < 2e-16 0.5121   0.0363   < 2e-16  < 2e-16 -
## seventies 4.9e-10 0.0020   0.8370   3.3e-12  < 2e-16 0.3644
##
## P value adjustment method: BH
```

minfun

Visualizing the data

```r
library(ggplot2)

ggplot(voice, aes(x = age, y = minfun, fill = age)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "minfun by Age Group", x = "Age Group", y = "minfun") +
  theme_minimal()
```



Check for normality

```r
check_normality("minfun")
```

```
## [1] "Shapiro-Wilk test for teens : 2.78073936026093e-38"
## [1] "Shapiro-Wilk test for twenties : 5.59994750123101e-66"
## [1] "Shapiro-Wilk test for thirties : 1.62325451304873e-52"
## [1] "Shapiro-Wilk test for fourties : 2.75873115318404e-51"
## [1] "Shapiro-Wilk test for fifties : 3.21529108357457e-30"
## [1] "Shapiro-Wilk test for sixties : 8.82837949326311e-17"
## [1] "Shapiro-Wilk test for seventies : 8.11188692941599e-27"
```

Kruskal-Wallis Test

```r
kruskal.test(minfun ~ age, data = voice)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  minfun by age
## Kruskal-Wallis chi-squared = 168.98, df = 6, p-value < 2.2e-16
```

Pairwise Wilcoxon Test

```r
pairwise.wilcox.test(voice$minfun, voice$age, p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  voice$minfun and voice$age
##
##           teens   twenties thirties fourties fifties sixties
## twenties  5.2e-12 -        -        -        -       -
## thirties  0.0001  3.9e-07  -        -        -       -
## fourties  0.0440  1.1e-12  0.0085   -        -       -
## fifties   1.9e-09 0.3278   7.3e-05  6.2e-08  -       -
## sixties   1.4e-10 0.0033   9.0e-07  4.9e-09  0.0327  -
## seventies 2.5e-16 4.9e-06  1.9e-12  5.4e-16  0.0028  0.6161
##
## P value adjustment method: BH
```