STA 106 Summary

Sanah Keswani-Santiago and Andrew

2024-03-17

Import Data

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

```
insurance <- read.csv("insurance.csv")

insurance$children <- as.factor(insurance$children)
region <- insurance$region
houseSize <- insurance$children
cost <- insurance$charges
```

Data Summary

```
# Average Cost by Region and House Size
aggregate(cost~region+houseSize, data = insurance, FUN = mean)
```

```
##       region houseSize      cost
## 1  northeast         0 11626.463
## 2  northwest         0 11324.371
## 3  southeast         0 14309.868
## 4  southwest         0 11938.505
## 5  northeast         1 16310.206
## 6  northwest         1 10230.256
## 7  southeast         1 13687.042
## 8  southwest         1 10406.485
## 9  northeast         2 13615.153
## 10 northwest         2 13464.315
## 11 southeast         2 15728.471
## 12 southwest         2 17483.486
## 13 northeast         3 14409.913
## 14 northwest         3 17786.161
## 15 southeast         3 18449.846
## 16 southwest         3 10402.442
## 17 northeast         4 14485.193
## 18 northwest         4 11347.019
```

```
## 19 southeast         4 14451.024
## 20 southwest         4 14933.261
## 21 northeast         5  6978.973
## 22 northwest         5  8965.796
## 23 southeast         5 10115.442
## 24 southwest         5  8444.159
```

```
# Standard Deviation of Cost by Region and House Size
aggregate(cost~region+houseSize, data = insurance, FUN = sd)
```

```
##       region houseSize      cost
## 1  northeast         0 10339.487
## 2  northwest         0 10551.248
## 3  southeast         0 14801.663
## 4  southwest         0 11340.917
## 5  northeast         1 13157.214
## 6  northwest         1  9031.057
## 7  southeast         1 12779.192
## 8  southwest         1 10651.506
## 9  northeast         2  9246.112
## 10 northwest         2 11135.470
## 11 southeast         2 14940.357
## 12 southwest         2 14782.150
## 13 northeast         3 12896.085
## 14 northwest         3 14173.184
## 15 southeast         3 12497.837
## 16 southwest         3  6455.847
## 17 northeast         4  6646.318
## 18 northwest         4  5563.298
## 19 southeast         4 12795.518
## 20 southwest         4 12107.035
## 21 northeast         5  2159.275
## 22 northwest         5        NA
## 23 southeast         5  2895.416
## 24 southwest         5  4985.139
```

```
# Sample Size by Region and House Size
aggregate(cost~region+houseSize, data = insurance, FUN = length)
```

```
##       region houseSize cost
## 1  northeast         0  147
## 2  northwest         0  132
## 3  southeast         0  157
## 4  southwest         0  138
## 5  northeast         1   77
## 6  northwest         1   74
## 7  southeast         1   95
## 8  southwest         1   78
## 9  northeast         2   51
## 10 northwest         2   66
## 11 southeast         2   66
## 12 southwest         2   57
## 13 northeast         3   39
```

```
## 14 northwest         3   46
## 15 southeast         3   35
## 16 southwest         3   37
## 17 northeast         4    7
## 18 northwest         4    6
## 19 southeast         4    5
## 20 southwest         4    7
## 21 northeast         5    3
## 22 northwest         5    1
## 23 southeast         5    6
## 24 southwest         5    8
```

```
# Table for Means, Standard Deviation and Sample Size by Group (Region)
groupMeansInsReg <- by(cost, region, mean)
groupSDsInsReg <- by(cost, region, sd)
groupNisInsReg <- by(cost, region, length)
insRegSummary <- rbind(groupMeansInsReg, groupSDsInsReg, groupNisInsReg)
insRegSummary <- round(insRegSummary, digits = 4)
colnames(insRegSummary) = names(groupMeansInsReg)
rownames(insRegSummary) = c("Means", "Std. Dev", "Sample Size")
insRegSummary
```

```
##             northeast northwest southeast southwest
## Means        13406.38  12417.58  14735.41  12346.94
## Std. Dev     11255.80  11072.28  13971.10  11557.18
## Sample Size    324.00    325.00    364.00    325.00
```

```
# Table for Means, Standard Deviation and Sample Size by Group (House Size)
groupMeansInsHou <- by(cost, houseSize, mean)
groupSDsInsHou <- by(cost, houseSize, sd)
groupNisInsHou <- by(cost, houseSize, length)
insHouSummary <- rbind(groupMeansInsHou, groupSDsInsHou, groupNisInsHou)
insHouSummary <- round(insHouSummary, digits = 4)
colnames(insHouSummary) = names(groupMeansInsHou)
rownames(insHouSummary) = c("Means", "Std. Dev", "Sample Size")
insHouSummary
```

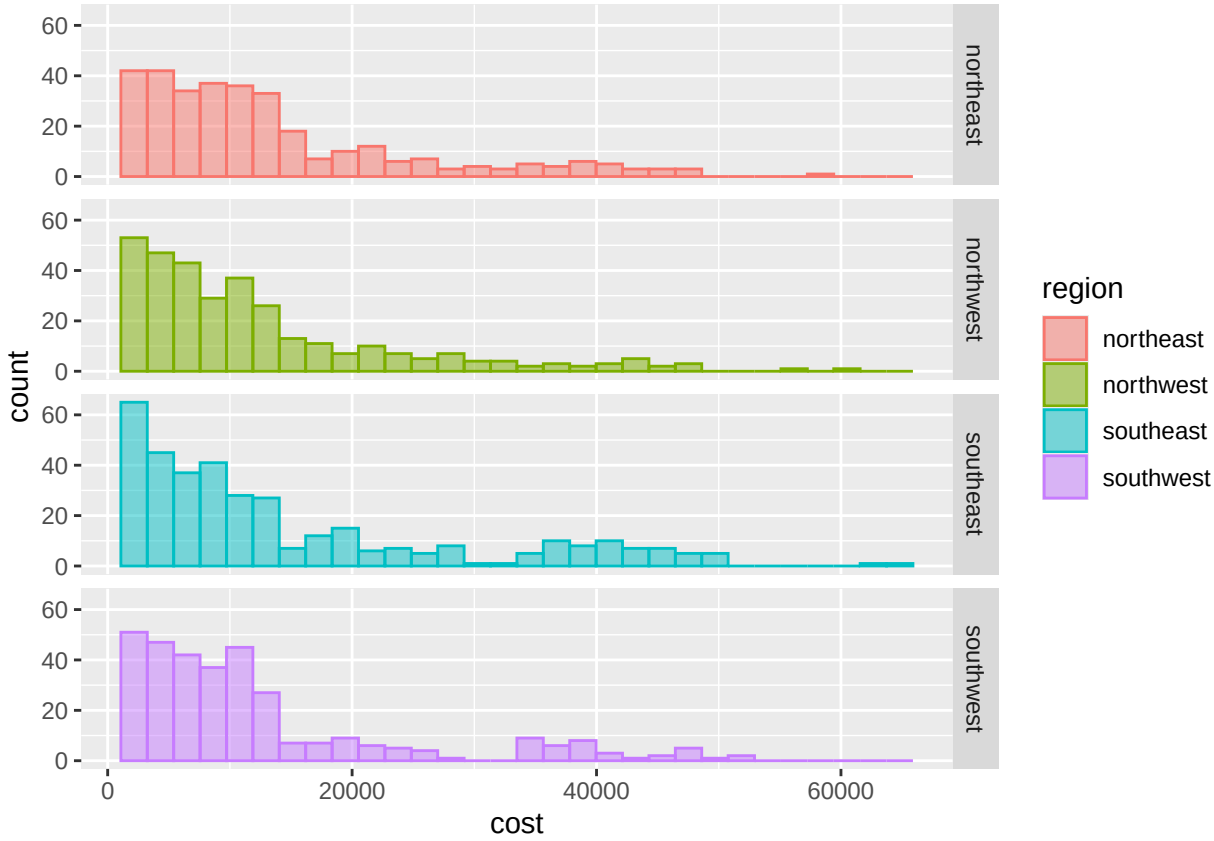
```
##                    0        1        2        3         4        5
## Means       12365.98 12731.17 15073.56 15355.32 13850.656 8786.035
## Std. Dev    12023.29 11823.63 12891.37 12330.87  9139.223 3808.436
## Sample Size   574.00   324.00   240.00   157.00    25.000   18.000
```

```
# Histograms of Cost by Region
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

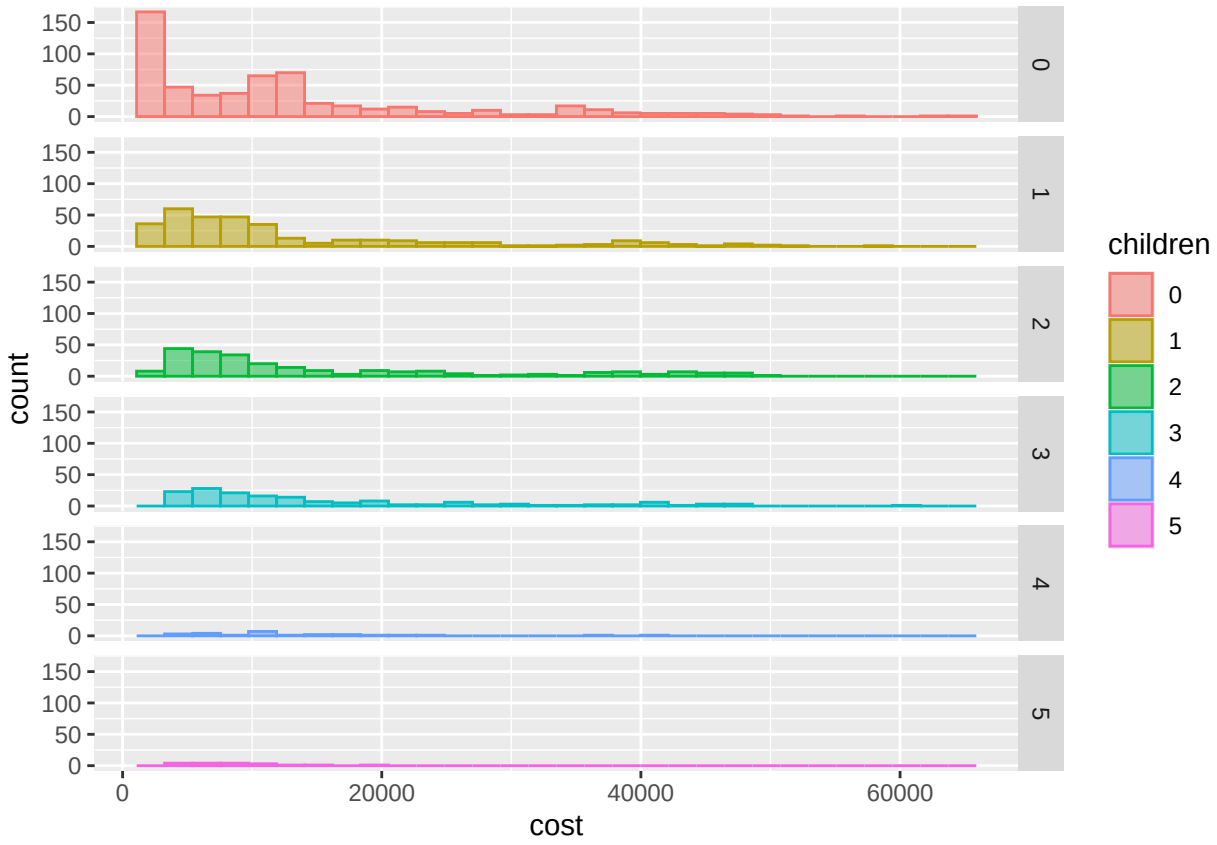
```
ggplot(insurance, aes(x=cost, color=region,fill=region)) + geom_histogram(position="identity", alpha=0.5
```

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

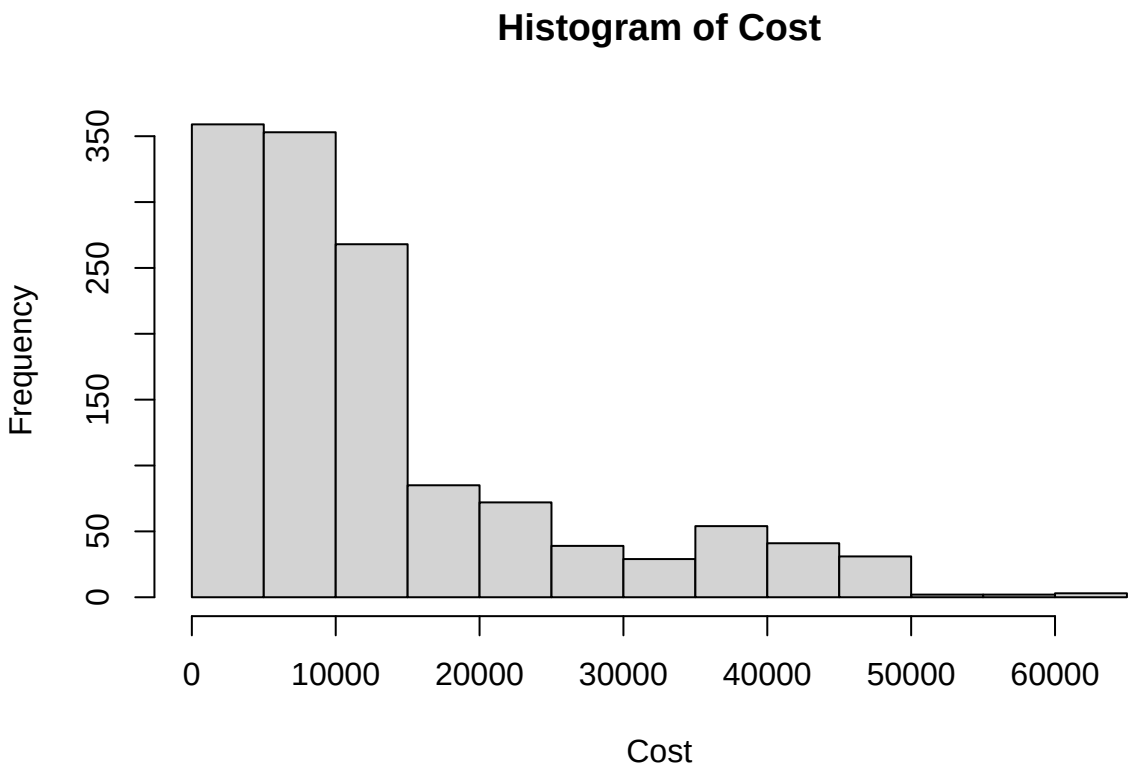



```
# Histograms of Beck Score by House Size
library(ggplot2)
ggplot(insurance, aes(x=cost, color=children,fill=children)) + geom_histogram(position="identity", alpha
```

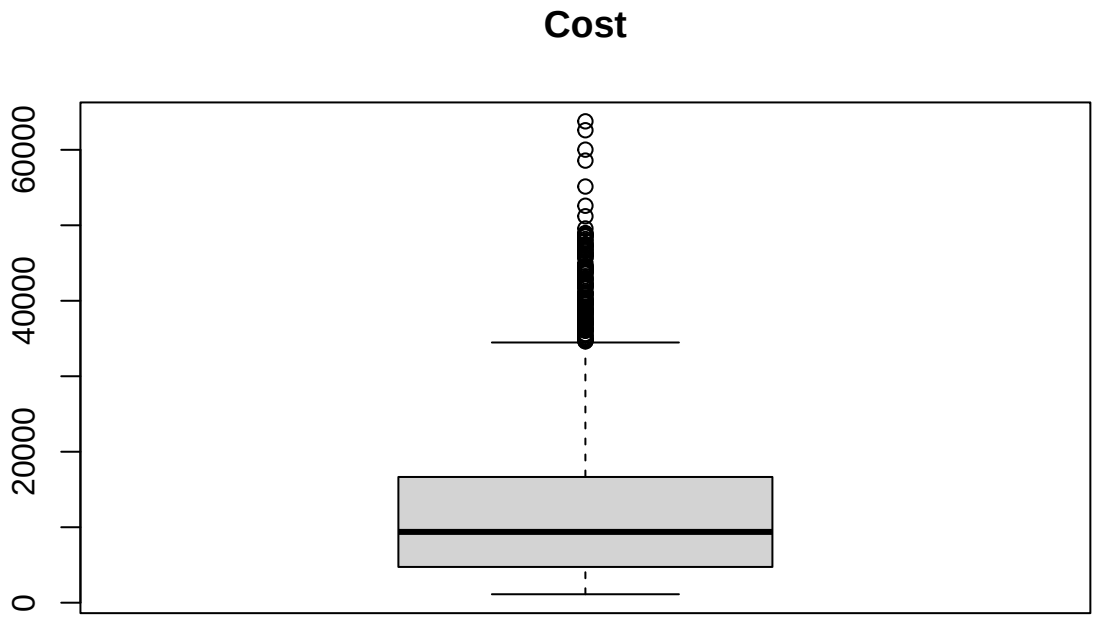
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



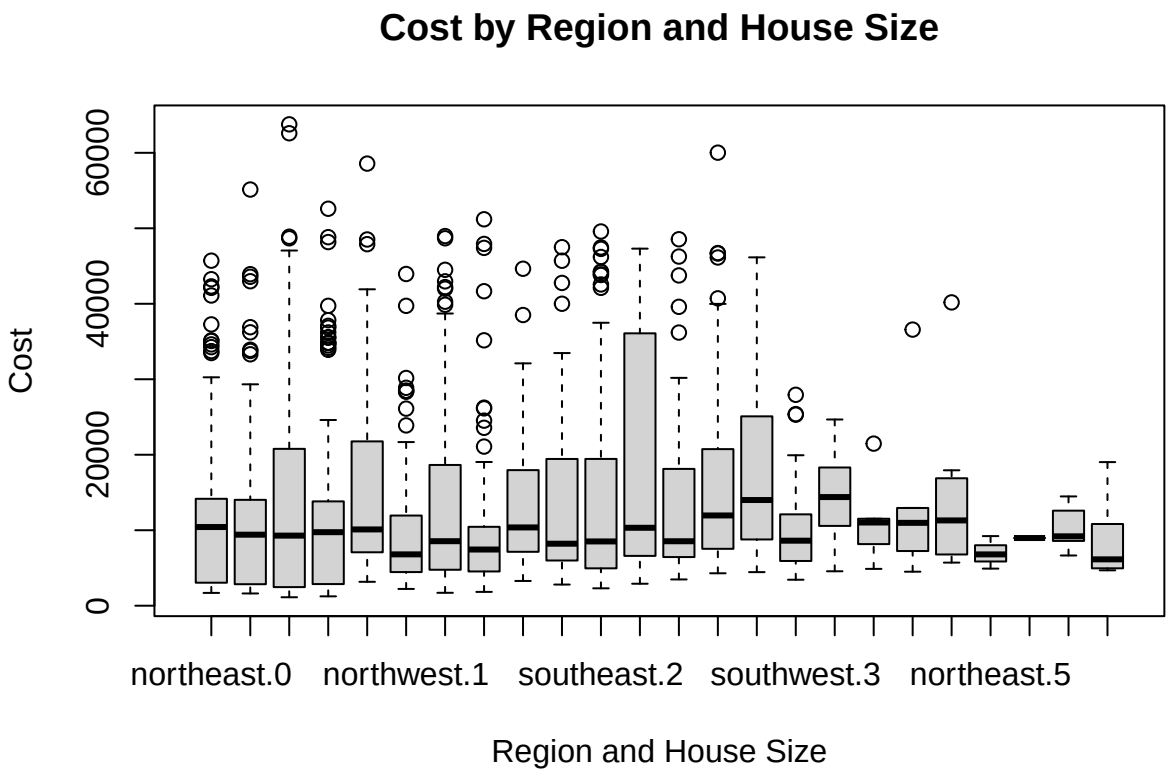

```
# Overall Histogram of Cost
hist(cost, xlab = "Cost", ylab = "Frequency", main = "Histogram of Cost")
```



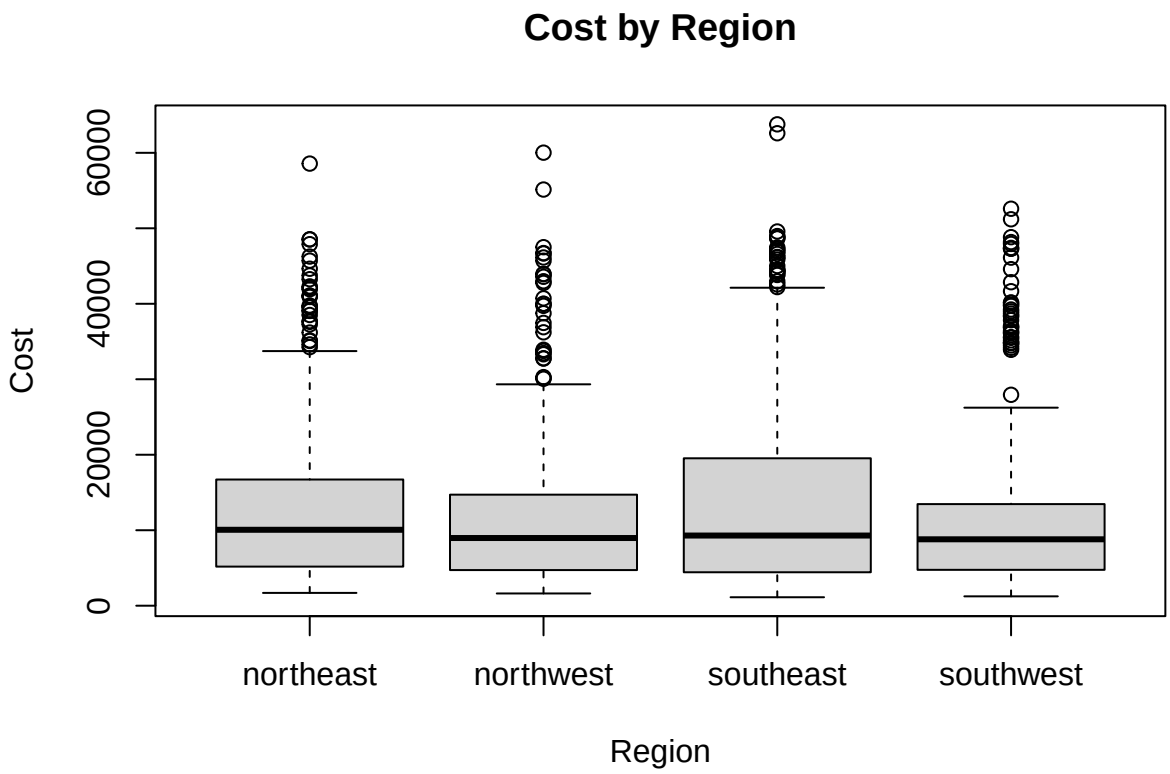

```r
# Overall Boxplot of Cost
boxplot(cost, main = "Cost")
```



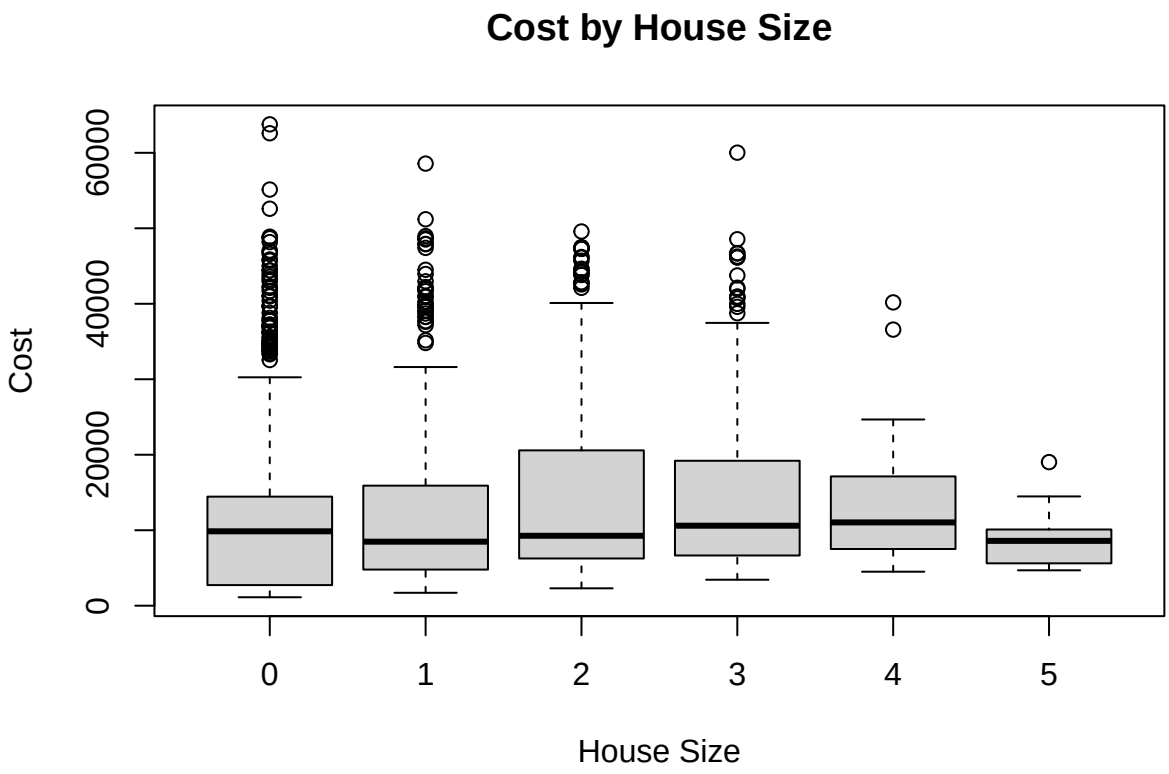

```
# Boxplots of Cost by Region and House Size
boxplot(cost~region+houseSize, data = insurance, xlab = "Region and House Size", ylab = "Cost", main =
```




```
# Boxplots of Beck Score by Region
boxplot(cost~region, data = insurance, xlab = "Region", ylab = "Cost", main = "Cost by Region")
```




```
# Boxplots of Beck Score by House Size
boxplot(cost~houseSize, data = insurance, xlab = "House Size", ylab = "Cost", main = "Cost by House Size
```



Standard Deviations by group

```
# Standard Deviation of Cost by Region
aggregate(cost~houseSize, data = insurance, FUN = sd)
```

```
##   houseSize      cost
## 1         0 12023.294
## 2         1 11823.631
## 3         2 12891.368
## 4         3 12330.869
## 5         4  9139.223
## 6         5  3808.436
```

```
# Sd by House Size
aggregate(cost~region, data = insurance, FUN = sd)
```

```
##      region     cost
## 1 northeast 11255.80
## 2 northwest 11072.28
## 3 southeast 13971.10
## 4 southwest 11557.18
```

```r
# Overall sd
aggregate(cost~1, data = insurance, FUN = sd)
```

```
##       cost
## 1 12110.01
```