DSCI561 - Final Project - Figures for Paper

Team 2

12/2/2021

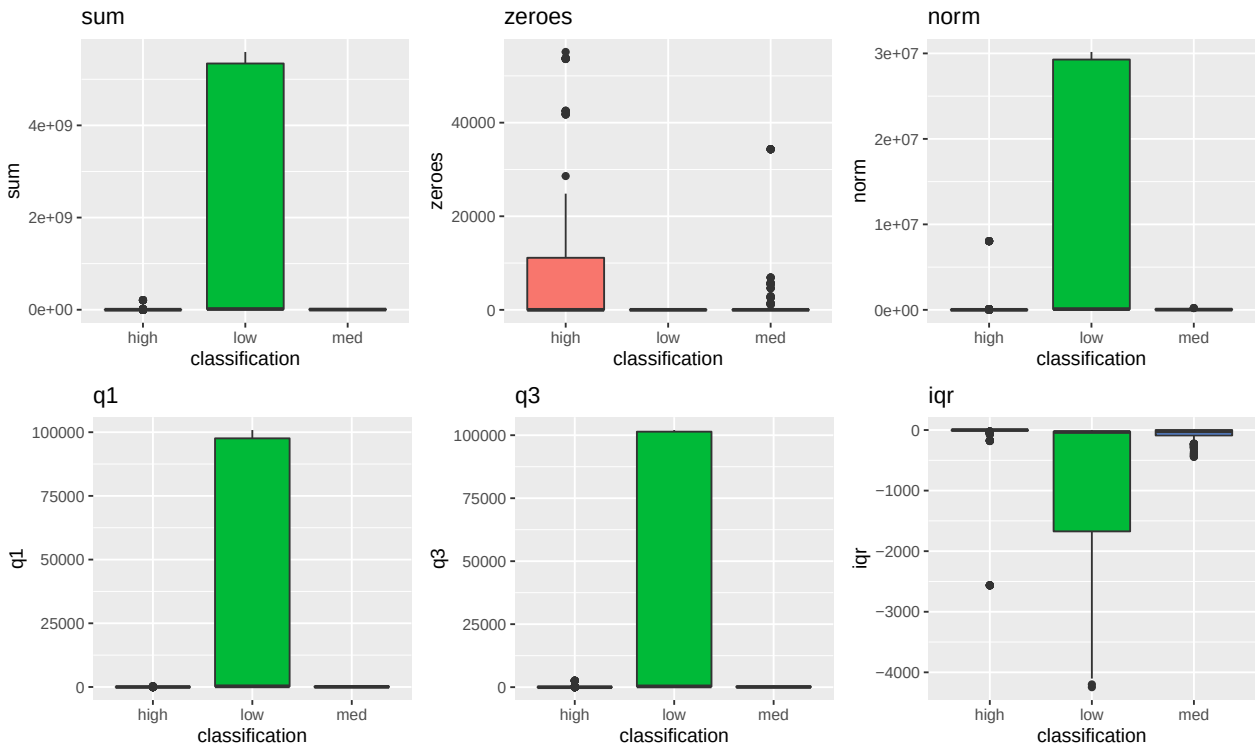
```
df = read_csv("../data/df.csv")
```

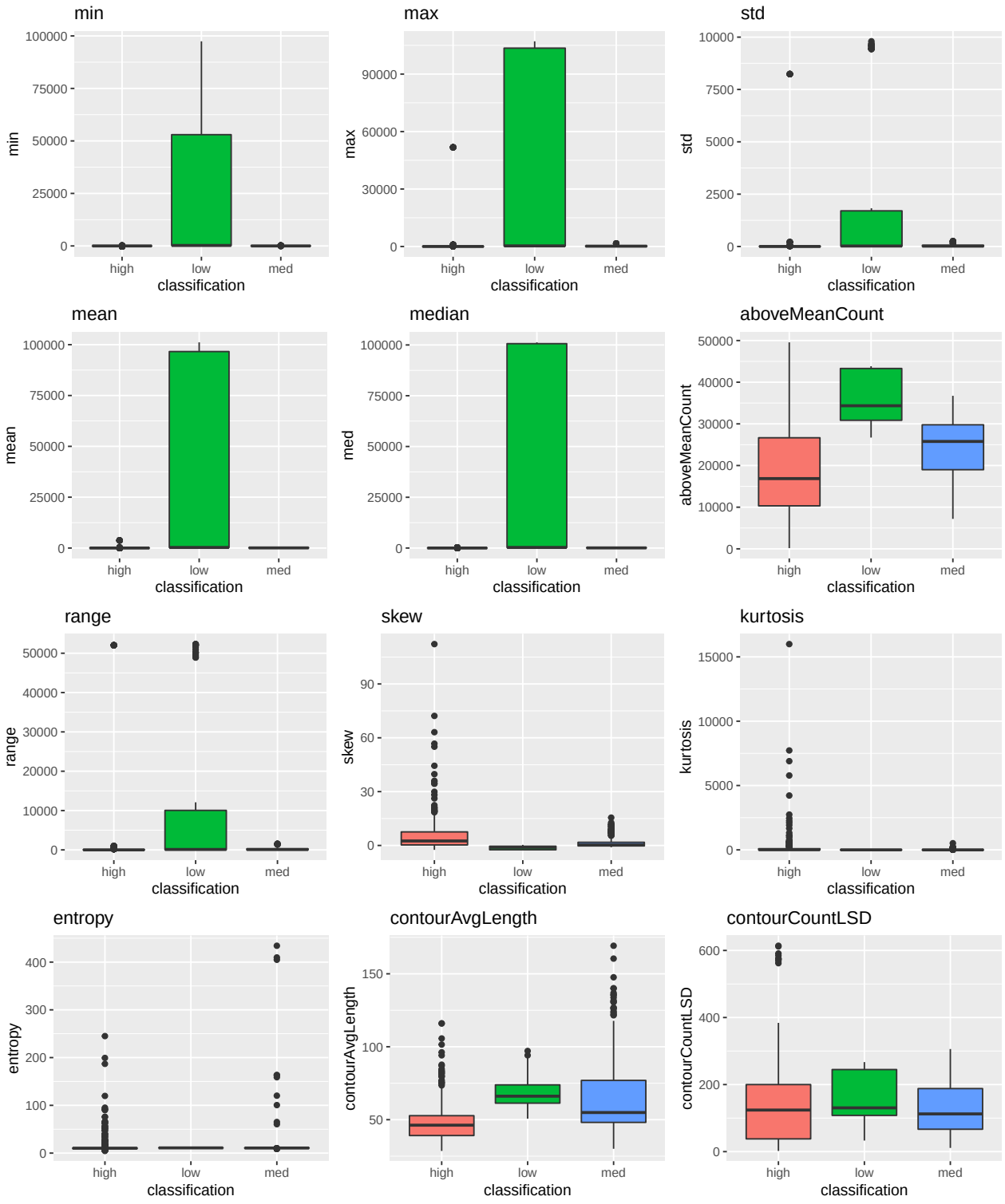
```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   classification = col_character()
## )
```

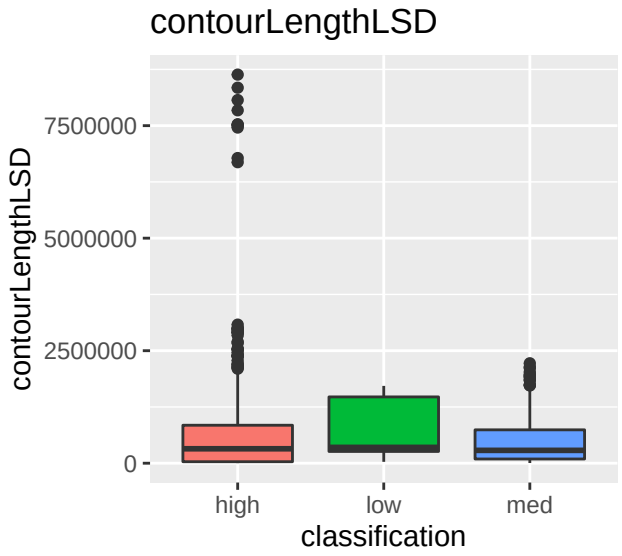
```
## See spec(...) for full column specifications.
```

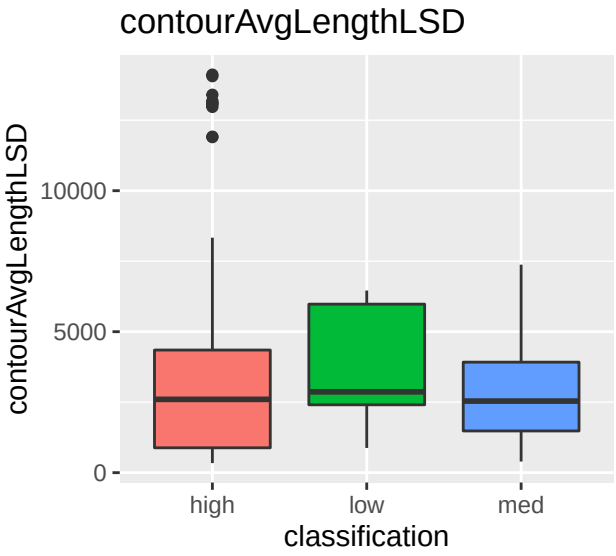
Creating boxplots of our initial variables.

```
dfTrain = df[df$validate==0,]
```




min
max
std
mean
median
aboveMeanCount
range
skew
kurtosis
entropy
contourAvgLength
contourCountLSD
classification
high
low
med






```
s.df = read_csv("../data/standardizeddf.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   classification = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

Comparing medium and high compression means:

```
print(paste("sum:", t.test(s.df$sum[s.df$classification=='high'], s.df$sum[s.df$classification=='med'])$
```

```
## [1] "sum: 0.000655984872869012"
```

```
print(paste("norm:", t.test(s.df$norm[s.df$classification=='high'], s.df$norm[s.df$classification=='med
```

```
## [1] "norm: 0.182177130341782"
```

```
print(paste("q1:", t.test(s.df$q1[s.df$classification=='high'], s.df$q1[s.df$classification=='med'])$p.
```

```
## [1] "q1: 5.32943316507823e-05"
```

```
print(paste("q3:", t.test(s.df$q3[s.df$classification=='high'], s.df$q3[s.df$classification=='med'])$p.
```

```
## [1] "q3: 0.000282878384081688"
```

```
print(paste("iqr:", t.test(s.df$iqr[s.df$classification=='high'], s.df$iqr[s.df$classification=='med'])$
```

```
## [1] "iqr: 0.00112849670979531"
```

```
print(paste("min:", t.test(s.df$min[s.df$classification=='high'], s.df$min[s.df$classification=='med'])$
```

```
## [1] "min: 0.000115091212302161"
```

```
print(paste("max:", t.test(s.df$max[s.df$classification=='high'], s.df$max[s.df$classification=='med'])$
```

```
## [1] "max: 0.41537106038783"
```

```
print(paste("std:", t.test(s.df$std[s.df$classification=='high'], s.df$std[s.df$classification=='med'])$
```

```
## [1] "std: 0.415144399299369"
```

```
print(paste("mean:", t.test(s.df$mean[s.df$classification=='high'], s.df$mean[s.df$classification=='med
```

```
## [1] "mean: 0.000655984872869017"
```

```
print(paste("med:", t.test(s.df$med[s.df$classification=='high'], s.df$med[s.df$classification=='med'])$
```

```
## [1] "med: 0.000105202530292408"
```

```
print(paste("range:", t.test(s.df$range[s.df$classification=='high'], s.df$range[s.df$classification=='
```

```
## [1] "range: 0.790078692439351"
```

Comparing medium and high compression means:

```
print(paste("zeroes:", t.test(s.df$zeroes[s.df$classification=='high'], s.df$zeroes[s.df$classification=
```

```
## [1] "zeroes: 1.14420037678045e-08"
```

```
print(paste("aboveMeanCount, low/medium:", t.test(s.df$aboveMeanCount[s.df$classification=='low'], s.df$
```

```
## [1] "aboveMeanCount, low/medium: 1.15499198297181e-42"
```

```
print(paste("aboveMeanCount, low/high:", t.test(s.df$aboveMeanCount[s.df$classification=='high'], s.df$a
```

```
## [1] "aboveMeanCount, low/high: 1.65161889282011e-83"
```

```
print(paste("aboveMeanCount, medium/high:", t.test(s.df$aboveMeanCount[s.df$classification=='high'], s.d
```

```
## [1] "aboveMeanCount, medium/high: 2.90441728067163e-26"
```

```
print(paste("skew, low/medium:", t.test(s.df$skew[s.df$classification=='low'], s.df$skew[s.df$classifica
```

```
## [1] "skew, low/medium: 1.56619962615283e-33"
```

```
print(paste("skew, low/high:", t.test(s.df$skew[s.df$classification=='high'], s.df$skew[s.df$classificat
```

```
## [1] "skew, low/high: 2.1013598882982e-38"
```

```
print(paste("skew, medium/high:", t.test(s.df$skew[s.df$classification=='high'], s.df$skew[s.df$classifi
```

```
## [1] "skew, medium/high: 6.04786919927293e-17"
```

```
print(paste("kurtosis, low/medium:", t.test(s.df$kurtosis[s.df$classification=='low'], s.df$kurtosis[s.d
```

```
## [1] "kurtosis, low/medium: 2.08705956478844e-05"
```

```
print(paste("kurtosis, low/high:", t.test(s.df$kurtosis[s.df$classification=='high'], s.df$kurtosis[s.di
```

```
## [1] "kurtosis, low/high: 1.69827404958106e-05"
```

```
print(paste("kurtosis, medium/high:", t.test(s.df$kurtosis[s.df$classification=='high'], s.df$kurtosis[s
```

```
## [1] "kurtosis, medium/high: 5.10033542294577e-05"
```

```
print(paste("entropy, low/medium:", t.test(s.df$entropy[s.df$classification=='low'], s.df$entropy[s.df$c
```

```
## [1] "entropy, low/medium: 0.0151052654990131"
```

```
print(paste("entropy, low/high:", t.test(s.df$entropy[s.df$classification=='high'], s.df$entropy[s.df$cl
```

```
## [1] "entropy, low/high: 7.40743690716382e-06"
```

```
print(paste("entropy, medium/high:", t.test(s.df$entropy[s.df$classification=='high'], s.df$entropy[s.di
```

```
## [1] "entropy, medium/high: 0.130338898915938"
```

```
print(paste("contourAvgLength, low/medium:", t.test(s.df$contourAvgLength[s.df$classification=='low'], s
```

```
## [1] "contourAvgLength, low/medium: 0.0161823630974807"
```

```
print(paste("contourAvgLength, low/high:", t.test(s.df$contourAvgLength[s.df$classification=='high'], s
```

```
## [1] "contourAvgLength, low/high: 1.38611456955862e-33"
```

```
print(paste("contourAvgLength, medium/high:", t.test(s.df$contourAvgLength[s.df$classification=='high']
```

```
## [1] "contourAvgLength, medium/high: 2.23068020501849e-27"
```

```
print(paste("contourCountLSD, low/medium:", t.test(s.df$contourCountLSD[s.df$classification=='low'], s.c
```

```
## [1] "contourCountLSD, low/medium: 0.00551183925630991"
```

```
print(paste("contourCountLSD, low/high:", t.test(s.df$contourCountLSD[s.df$classification=='high'], s.di
```

```
## [1] "contourCountLSD, low/high: 0.533465380700969"
```

```
print(paste("contourCountLSD, medium/high:", t.test(s.df$contourCountLSD[s.df$classification=='high'], s
```

```
## [1] "contourCountLSD, medium/high: 0.000712593080793888"
```

```
print(paste("contourLengthLSD, low/medium:", t.test(s.df$contourLengthLSD[s.df$classification=='low'], s
```

```
## [1] "contourLengthLSD, low/medium: 0.0108357365477674"
```

```
print(paste("contourLengthLSD, low/high:", t.test(s.df$contourLengthLSD[s.df$classification=='high'], s
```

```
## [1] "contourLengthLSD, low/high: 0.00121250616821494"
```

```
print(paste("contourLengthLSD, medium/high:", t.test(s.df$contourLengthLSD[s.df$classification=='high']
```

```
## [1] "contourLengthLSD, medium/high: 1.34095717405509e-07"
```

```
print(paste("contourAvgLengthLSD, low/medium:", t.test(s.df$contourAvgLengthLSD[s.df$classification=='lo
```

```
## [1] "contourAvgLengthLSD, low/medium: 0.000239188068650828"
```

```
print(paste("contourAvgLengthLSD, low/high:", t.test(s.df$contourAvgLengthLSD[s.df$classification=='high
```

```
## [1] "contourAvgLengthLSD, low/high: 0.526099699762322"
```

```
print(paste("contourAvgLengthLSD, medium/high:", t.test(s.df$contourAvgLengthLSD[s.df$classification=='h
```

```
## [1] "contourAvgLengthLSD, medium/high: 0.00188787502047971"
```

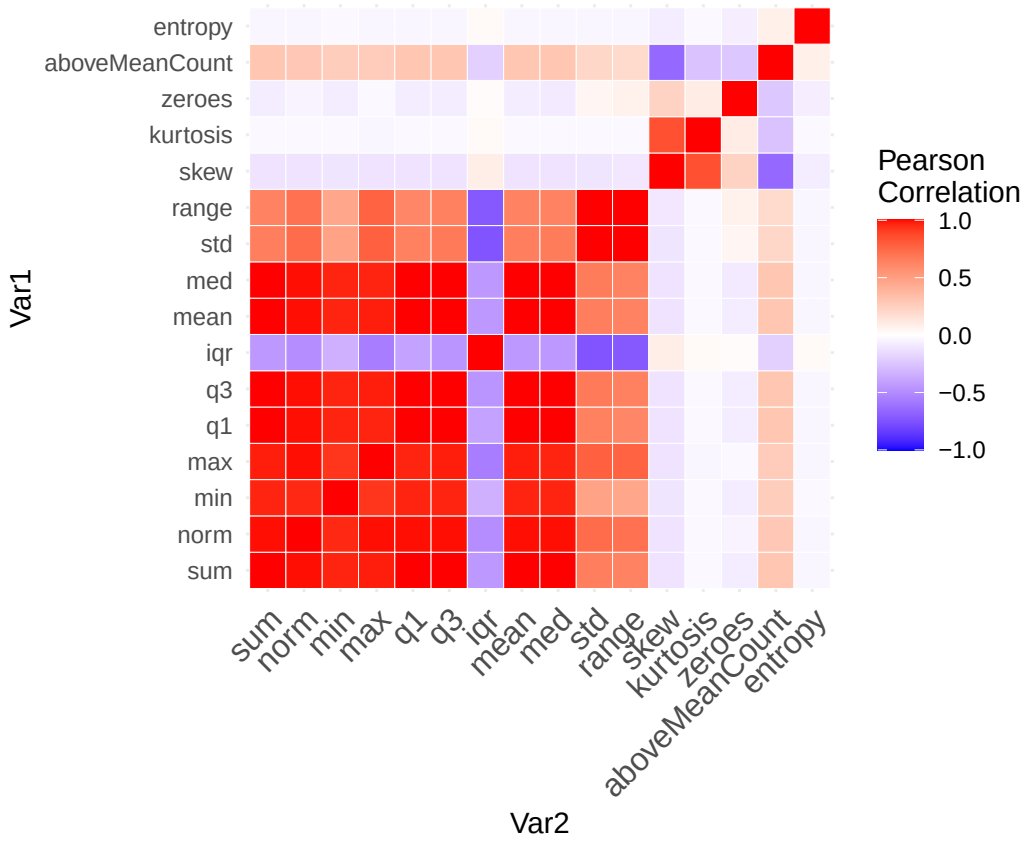
Correlation

```
c.df = df %>% dplyr::select(sum,norm,min,max,q1,q3,iqr,mean,med,std,range,skew,kurtosis,zeroes,aboveMean
cormat <- round(cor(c.df),2)

melted_cormat <- melt(cormat)

ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Pearson\nCorrelation") +
  theme_minimal()+
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
```

```
    size = 12, hjust = 1))+
 coord_fixed()
```




```
c.contour.df = df %>% dplyr::select(contourCountLSD,contourAvgLength,contourLengthLSD,contourAvgLengthLS
cor(c.contour.df)
```

```
##                     contourCountLSD contourAvgLength contourLengthLSD
## contourCountLSD           1.0000000      -0.10945736        0.9192746
## contourAvgLength         -0.1094574       1.00000000       -0.1291245
## contourLengthLSD          0.9192746      -0.12912446        1.0000000
## contourAvgLengthLSD       0.9849047      -0.07755503        0.9281226
##                     contourAvgLengthLSD
## contourCountLSD              0.98490469
## contourAvgLength            -0.07755503
## contourLengthLSD             0.92812256
## contourAvgLengthLSD          1.00000000
```

```
cormat <- round(cor(c.contour.df),2)

melted_cormat <- melt(cormat)

ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Pearson\nCorrelation") +
  theme_minimal()+
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
```

```
    size = 12, hjust = 1))+
coord_fixed()
```