Class 2 - Data Exploration

Load Libraries

```
library(psych)
library(corrgram)
```

Load Dataset

```
knitr::opts_chunk$set(echo = TRUE)
marketing = read.csv("E:/MSc DSc/Sem 1/Business Analytics/Ch3_marketing.csv", stringsAsFactors = TRUE)
str(marketing)
```

```
## 'data.frame':    172 obs. of  7 variables:
##  $ google_adwords : num  65.7 39.1 174.8 34.4 78.2 ...
##  $ facebook       : num  47.9 55.2 52 62 40.9 ...
##  $ twitter        : num  52.5 77.4 68 86.9 30.4 ...
##  $ marketing_total: num  166 172 295 183 150 ...
##  $ revenues       : num  39.3 38.9 49.5 40.6 40.2 ...
##  $ employees      : int  5 7 11 7 9 3 10 6 6 4 ...
##  $ pop_density    : Factor w/ 3 levels "High","Low","Medium": 1 3 3 1 2 1 2 1 3 2 ...
```

```
marketing$pop_density = factor(marketing$pop_density,
                               ordered = TRUE,
                               levels = c('Low','Medium','High'))
```

Focusing on google_adwords, and pop_density

```
summary(marketing$pop_density)
```

```
##    Low Medium   High
##     68     52     52
```

```
summary(marketing$google_adwords)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.65   97.25  169.47  169.87  243.10  321.00
```

```
sd(marketing$google_adwords)
```

```
## [1] 87.47228
```

```r
var(marketing$google_adwords)
```

```
## [1] 7651.4
```

Tabular Exploration

```r
summary2 = function(x) {
    results = c(summary(x),'StdDev.' = sd(x),'Var.' = var(x), 'IQR' = IQR(x))
    return(results)
}
summary2(marketing$google_adwords)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.    StdDev.
##   23.65000   97.24750  169.47500  169.86849  243.10500  321.00000   87.47228
##       Var.        IQR
## 7651.39954  145.85750
```

```r
summary3 = function(x) {
    results = c('Min' = min(x),
                'Q1' = quantile(x, 0.25),
                'Median' = median(x),
                'Mean' = mean(x),
                'Q3' = quantile(x, 0.75),
                'Max' = max(x),
                'StdDev' = sd(x),
                'Var' = var(x),
                'IQR' = IQR(x))
    results
}
summary3(marketing$google_adwords)
```

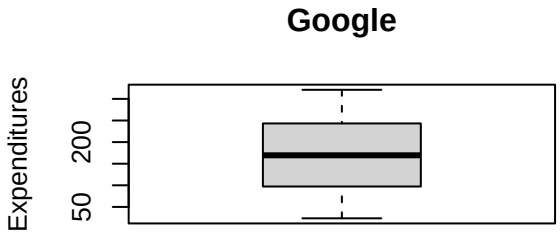
```
##        Min     Q1.25%     Median       Mean     Q3.75%        Max     StdDev
##   23.65000   97.24750  169.47500  169.86849  243.10500  321.00000   87.47228
##        Var        IQR
## 7651.39954  145.85750
```

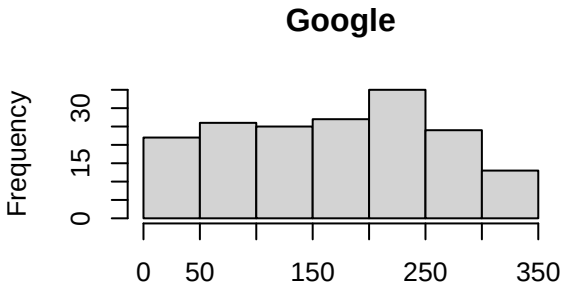
```r
summary(marketing$pop_density)
```

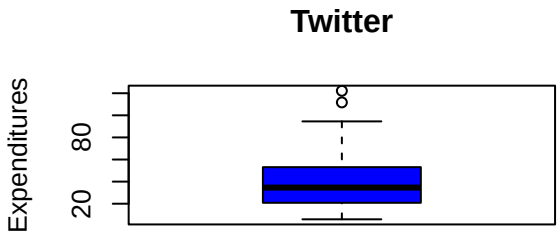
```
##    Low Medium   High
##     68     52     52
```

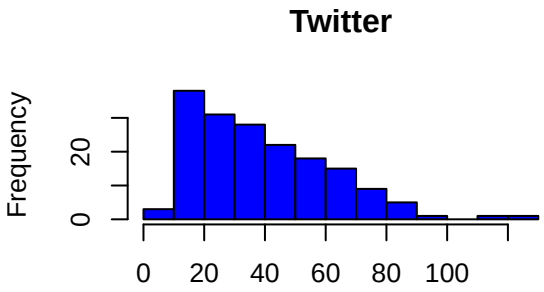
Graphical Exploration

```r
#layout(matrix(1:4,ncol = 2)) # or par(mfrow = c(2,2))
par(mfrow=c(2,2))
boxplot(marketing$google_adwords, ylab = 'Expenditures', main = 'Google')
hist(marketing$google_adwords, main = 'Google', xlab = NULL)
boxplot(marketing$twitter, ylab = 'Expenditures', col = 'blue', main = 'Twitter')
hist(marketing$twitter, col = 'blue', main = 'Twitter', xlab = NULL)
```










```
par(mfrow=c(1,1))
```

Analyzing two variables together

```
summary(marketing)
```

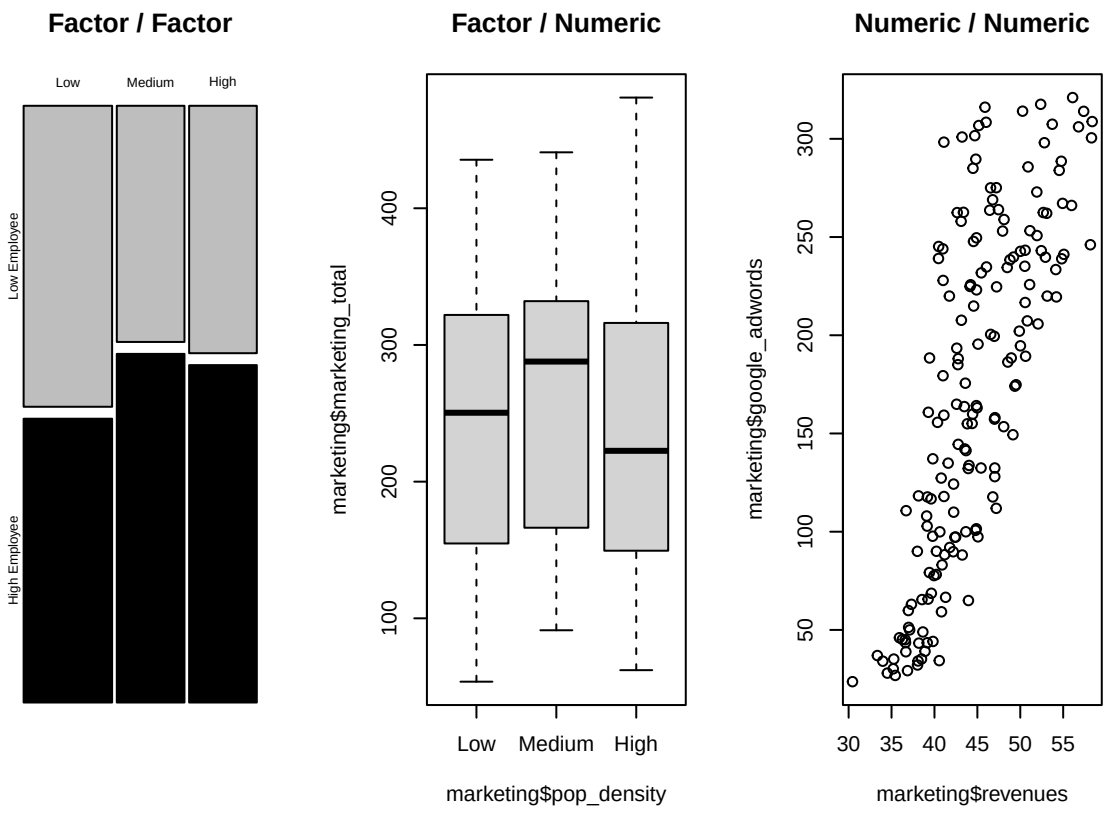
```
##  google_adwords      facebook        twitter       marketing_total
##  Min.   : 23.65   Min.   : 8.00   Min.   :  5.89   Min.   : 53.65
##  1st Qu.: 97.25   1st Qu.:19.37   1st Qu.: 20.94   1st Qu.:158.41
##  Median :169.47   Median :33.66   Median : 34.59   Median :245.56
##  Mean   :169.87   Mean   :33.87   Mean   : 38.98   Mean   :242.72
##  3rd Qu.:243.10   3rd Qu.:47.80   3rd Qu.: 52.94   3rd Qu.:322.62
##  Max.   :321.00   Max.   :62.17   Max.   :122.19   Max.   :481.00
##     revenues       employees      pop_density
##  Min.   :30.45   Min.   : 3.000   Low   :68
##  1st Qu.:40.33   1st Qu.: 6.000   Medium:52
##  Median :43.99   Median : 8.000   High  :52
##  Mean   :44.61   Mean   : 7.866
##  3rd Qu.:48.61   3rd Qu.:10.000
##  Max.   :58.38   Max.   :12.000
```

```
marketing$emp_factor = cut(marketing$employees , 2)
```

```
levels(marketing$emp_factor) = c('Low Employee', 'High Employee')
table1 = table(marketing$pop_density,marketing$emp_factor)
table1
```

```
##
##          Low Employee High Employee
##   Low               35            33
##   Medium            21            31
##   High              22            30
```

```
par(mfrow=c(1,3))
mosaicplot(table1,
           col=c('gray','black'),
           main = 'Factor / Factor')
boxplot(marketing$marketing_total ~ marketing$pop_density,
        main = 'Factor / Numeric')
plot(marketing$revenues, marketing$google_adwords,
    main = 'Numeric / Numeric')
```




```
par(mfrow=c(1,1))
```

Correlation

```
cor(marketing$google_adwords,marketing$revenues)
```

```
## [1] 0.7662461
```

```
#cor(marketing$google_adwords, marketing$facebook)
```

```
cor.test(marketing$google_adwords, marketing$revenues)
```

```
##
##  Pearson's product-moment correlation
##
## data:  marketing$google_adwords and marketing$revenues
## t = 15.548, df = 170, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6964662 0.8216704
## sample estimates:
##       cor
## 0.7662461
```

```
cor_test = function(x,y) {
    results = cor.test(x,y)
    output = c(#'Data' = results$data.name,
                'Correlation Coefficient' = results$estimate,
                'p-value' = results$p.value)
    output
}
```

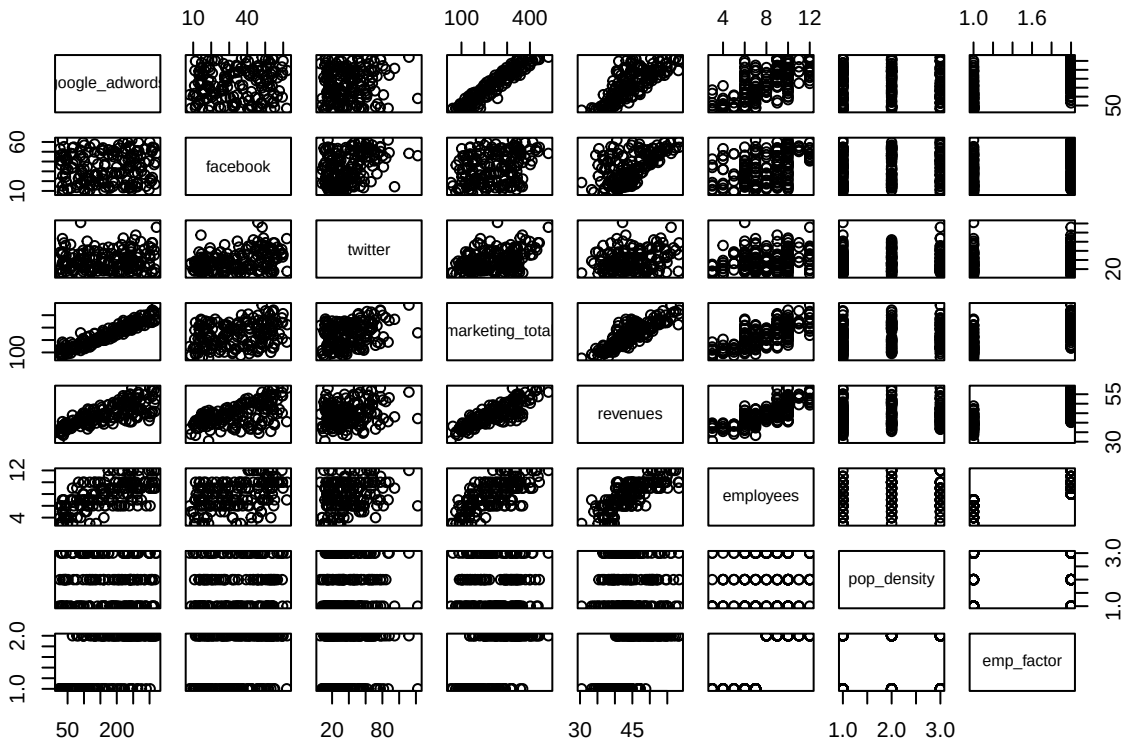
```
cor(marketing[,1:6])
```

```
##                 google_adwords   facebook   twitter marketing_total  revenues
## google_adwords      1.00000000 0.07643216 0.0989750       0.9473566 0.7662461
## facebook            0.07643216 1.00000000 0.3543410       0.3102232 0.5778213
## twitter             0.09897500 0.35434096 1.0000000       0.3758691 0.2696854
## marketing_total     0.94735659 0.31022316 0.3758691       1.0000000 0.8530354
## revenues            0.76624608 0.57782131 0.2696854       0.8530354 1.0000000
## employees           0.66103123 0.41019661 0.2290618       0.7210171 0.7656857
##                 employees
## google_adwords  0.6610312
## facebook        0.4101966
## twitter         0.2290618
## marketing_total 0.7210171
## revenues        0.7656857
## employees       1.0000000
```

```
corr.test(marketing[,1:6])
```

```
## Call:corr.test(x = marketing[, 1:6])
## Correlation matrix
##                 google_adwords facebook twitter marketing_total revenues
## google_adwords            1.00     0.08    0.10            0.95     0.77
## facebook                  0.08     1.00    0.35            0.31     0.58
## twitter                   0.10     0.35    1.00            0.38     0.27
## marketing_total           0.95     0.31    0.38            1.00     0.85
## revenues                  0.77     0.58    0.27            0.85     1.00
## employees                 0.66     0.41    0.23            0.72     0.77
##                 employees
## google_adwords       0.66
## facebook             0.41
## twitter              0.23
## marketing_total      0.72
## revenues             0.77
## employees            1.00
## Sample Size
## [1] 172
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                 google_adwords facebook twitter marketing_total revenues
## google_adwords            0.00     0.39    0.39               0        0
## facebook                  0.32     0.00    0.00               0        0
## twitter                   0.20     0.00    0.00               0        0
## marketing_total           0.00     0.00    0.00               0        0
## revenues                  0.00     0.00    0.00               0        0
## employees                 0.00     0.00    0.00               0        0
##                 employees
## google_adwords       0.00
## facebook             0.00
## twitter              0.01
## marketing_total      0.00
## revenues             0.00
## employees            0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

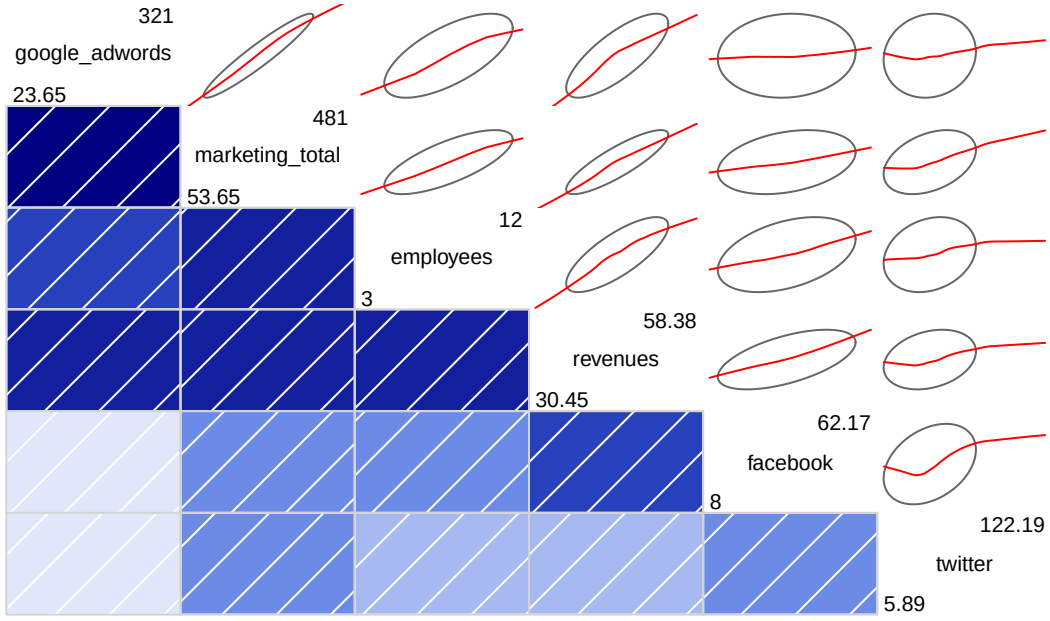
```
pairs(marketing)
```



Corrgram

```
corrgram(marketing,order=TRUE,
         main = "Correlogram of Marketing Data Ordered",
         lower.panel = panel.shade,
         upper.panel = panel.ellipse,
         diag.panel = panel.minmax,
         text.panel = panel.txt)
```

Correlogram of Marketing Data Ordered




```
corrgram(marketing,order=FALSE,
         main = "Correlogram of Marketing Data Unordered",
         lower.panel = panel.conf,
         upper.panel = panel.shade,
         diag.panel = panel.minmax,
         text.panel = panel.txt)
```

Correlogram of Marketing Data Unordered