INFS_SP5_2023
Predictive Analytics
PRACTICAL 1

Enna H

Contents

1. load data

```
# load data
census.data <- read.csv(url("http://bit.ly/infs5100_camden_data"))
```

2. data exploration

```
# check data
head(census.data)
```

```
##          OA White_British Low_Occupancy Unemployed Qualification
## 1 E00004120      42.35669     6.2937063   1.893939      73.62637
## 2 E00004121      47.20000     5.9322034   2.688172      69.90291
## 3 E00004122      40.67797     2.9126214   1.212121      67.58242
## 4 E00004123      49.66216     0.9259259   2.803738      60.77586
## 5 E00004124      51.13636     2.0000000   3.816794      65.98639
## 6 E00004125      41.41791     3.9325843   3.846154      74.20635
```

```
# get data structure
str(census.data)
```

```
## 'data.frame':    749 obs. of  5 variables:
##  $ OA           : chr  "E00004120" "E00004121" "E00004122" "E00004123" ...
##  $ White_British: num  42.4 47.2 40.7 49.7 51.1 ...
##  $ Low_Occupancy: num  6.294 5.932 2.913 0.926 2 ...
##  $ Unemployed   : num  1.89 2.69 1.21 2.8 3.82 ...
##  $ Qualification: num  73.6 69.9 67.6 60.8 66 ...
```

- Challenge 1.

```
# rename column OA to Output_Area
census.data <- rename(census.data, Output_Area = OA)
# check data
names(census.data)
```

```
## [1] "Output_Area"   "White_British" "Low_Occupancy" "Unemployed"
## [5] "Qualification"
```

3. descriptive statistics

```
# mean, median, sd, range, quartiles
summary(census.data)
```

```
##  Output_Area        White_British    Low_Occupancy      Unemployed
##  Length:749         Min.   : 7.882   Min.   : 0.000   Min.   : 0.000
##  Class :character   1st Qu.:35.915   1st Qu.: 6.015   1st Qu.: 2.500
##  Mode  :character   Median :44.541   Median :10.000   Median : 4.186
##                     Mean   :44.832   Mean   :11.597   Mean   : 4.510
##                     3rd Qu.:54.472   3rd Qu.:16.107   3rd Qu.: 6.158
##                     Max.   :78.035   Max.   :64.286   Max.   :18.623
##  Qualification
##  Min.   :11.64
##  1st Qu.:36.32
##  Median :55.10
##  Mean   :51.43
##  3rd Qu.:66.23
##  Max.   :88.07
```

```
# for unemployment, range
range(census.data$Unemployed)
```

```
## [1]  0.00000 18.62348
```

- Challenge 2.

```
# use the doBy() package
pacman::p_load(doBy)
```

```
# using the summaryBy() function with grouping by the variable Qualification,
# gives the mean and median for each unique combination of all variables
result <- summaryBy(White_British+Low_Occupancy+Unemployed+Qualification ~ .,
                    data = census.data,
                    FUN = function(x) {c(Mean = mean(x, na.rm = TRUE),
                                         Median = median(x, na.rm = TRUE))})
head(result)
```

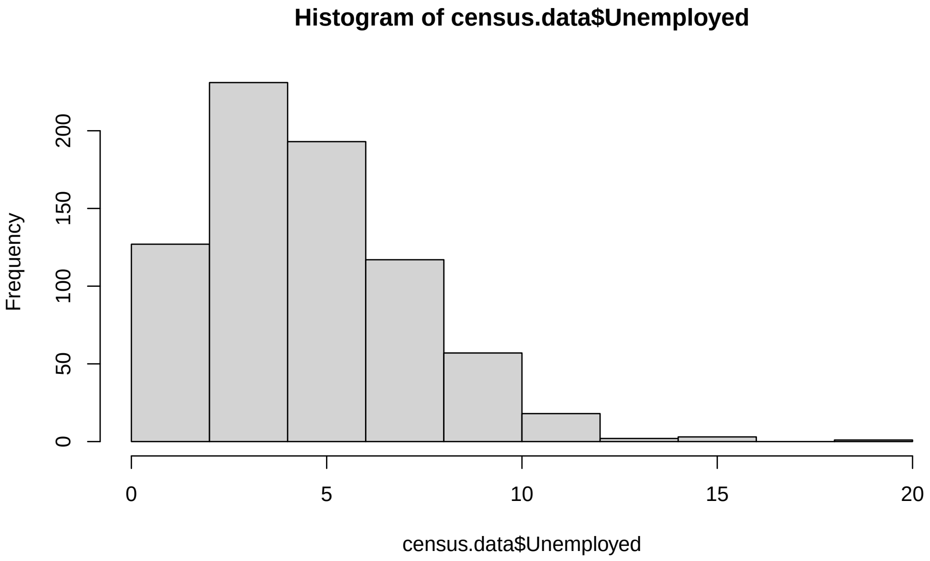
```
##   Output_Area White_British.Mean White_British.Median Low_Occupancy.Mean
## 1   E00004120           42.35669             42.35669          6.2937063
## 2   E00004121           47.20000             47.20000          5.9322034
## 3   E00004122           40.67797             40.67797          2.9126214
## 4   E00004123           49.66216             49.66216          0.9259259
## 5   E00004124           51.13636             51.13636          2.0000000
## 6   E00004125           41.41791             41.41791          3.9325843
##   Low_Occupancy.Median Unemployed.Mean Unemployed.Median Qualification.Mean
## 1            6.2937063        1.893939          1.893939           73.62637
## 2            5.9322034        2.688172          2.688172           69.90291
## 3            2.9126214        1.212121          1.212121           67.58242
## 4            0.9259259        2.803738          2.803738           60.77586
## 5            2.0000000        3.816794          3.816794           65.98639
## 6            3.9325843        3.846154          3.846154           74.20635
##   Qualification.Median
## 1             73.62637
## 2             69.90291
## 3             67.58242
## 4             60.77586
## 5             65.98639
## 6             74.20635
```

```
# using the summaryBy() function without grouping by other variables,
# gives the overall mean and median
result <- summaryBy(White_British+Low_Occupancy+Unemployed+Qualification ~ 1,
                    data = census.data,
                    FUN = function(x) {c(Mean = mean(x, na.rm = TRUE),
                                         Median = median(x, na.rm = TRUE))})
print(result)
```

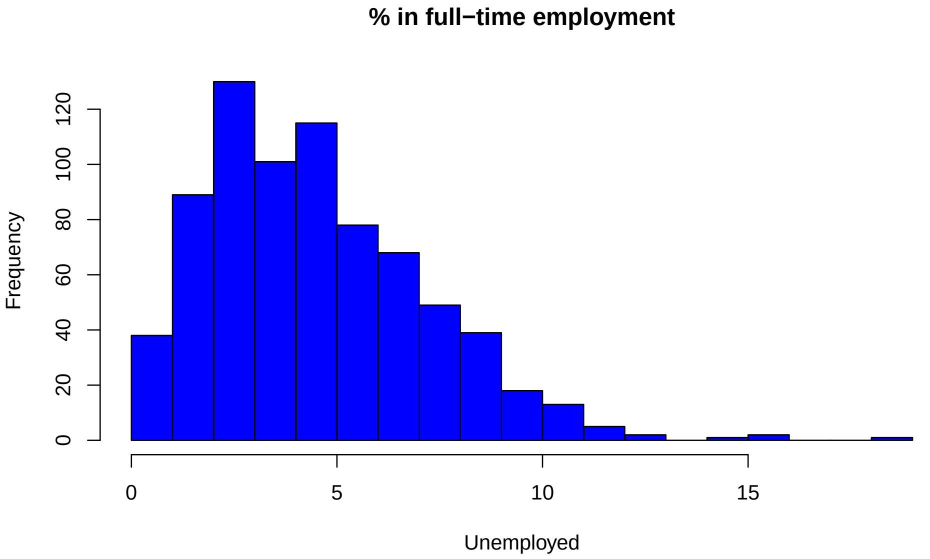
```
##   White_British.Mean White_British.Median Low_Occupancy.Mean
## 1           44.83223             44.54148            11.5972
##   Low_Occupancy.Median Unemployed.Mean Unemployed.Median Qualification.Mean
## 1                   10        4.510309          4.186047           51.42978
##   Qualification.Median
## 1             55.10204
```

4. Univariate plots

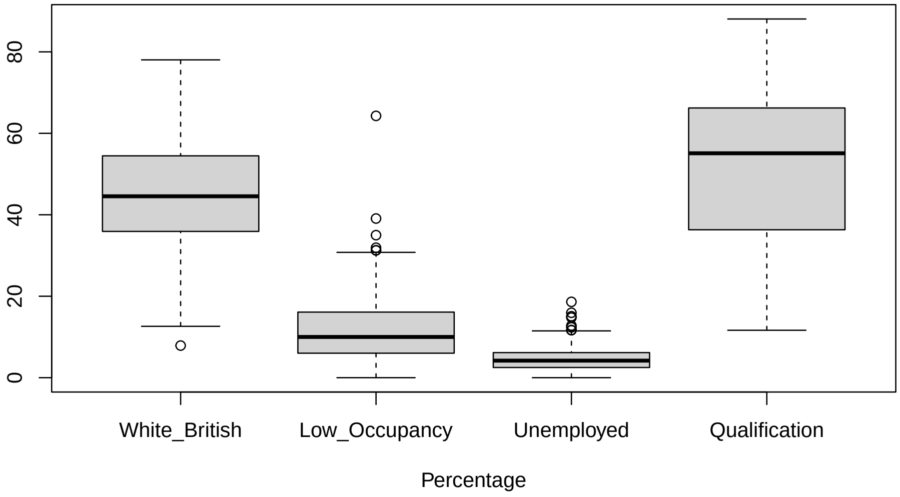
```
# Creates a histogram
hist(census.data$Unemployed)
```



```
# Creates a histogram, enters more commands about the visualisation
hist(census.data$Unemployed, breaks=20, col= "blue", main="% in full-time employment", xlab="Unemployed
```

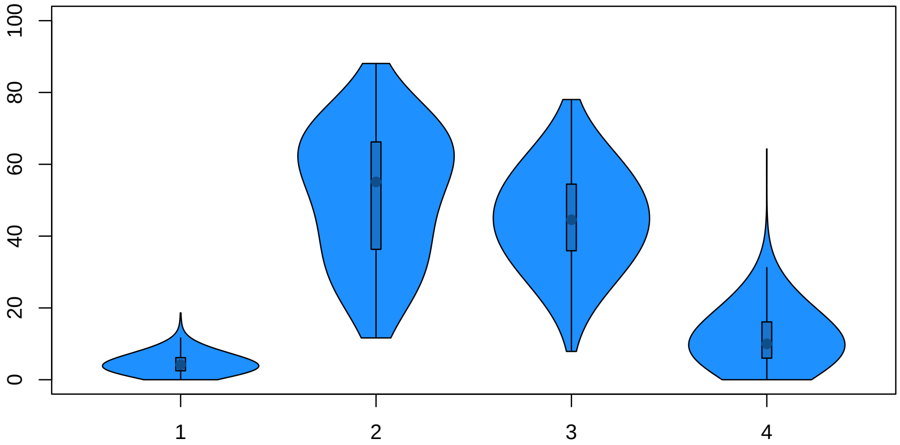


```
# box and whisker plots
boxplot(census.data[,2:5], xlab="Percentage")
```

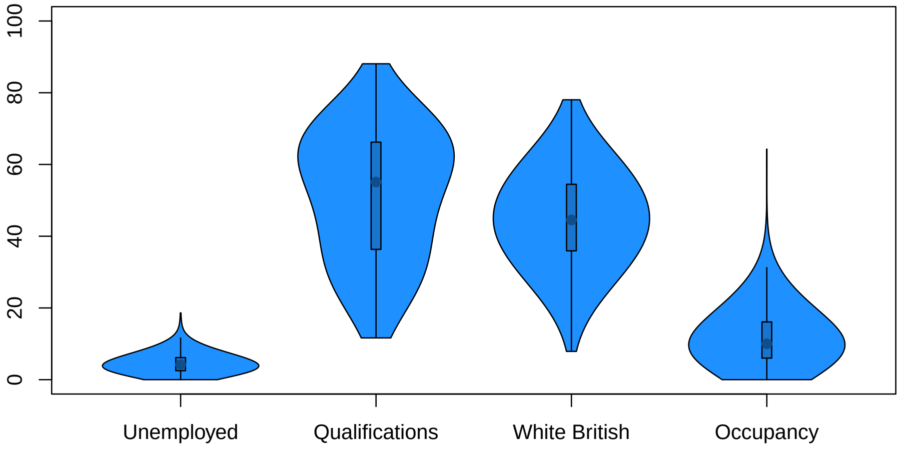


```
pacman::p_load(vioplot)
```

```
# add names to the plot
vioplot(census.data$Unemployed, census.data$Qualification,
census.data$White_British, census.data$Low_Occupancy, ylim=c(0,100), col =
"dodgerblue", rectCol="dodgerblue3", colMed="dodgerblue4")
```



```
# add names to the plot
vioplot(census.data$Unemployed, census.data$Qualification,
census.data$White_British, census.data$Low_Occupancy, ylim=c(0,100), col =
"dodgerblue", rectCol="dodgerblue3", colMed="dodgerblue4",
names=c("Unemployed", "Qualifications", "White British", "Occupancy"))
```



```
# Open a new png device
png("vioplot.png")

# Create the plot
vioplot(census.data$Unemployed, census.data$Qualification,
        census.data$White_British, census.data$Low_Occupancy,
        ylim=c(0,100),
        col = "dodgerblue",
        rectCol="dodgerblue3",
        colMed="dodgerblue4",
        names=c("Unemployed", "Qualifications", "White British", "Occupancy"))

# Close the png device
dev.off()
```

```
## pdf
##   2
```