Gautam_Kapila_Unit3AssignmentSubmission

G. Kapila

5/23/2019

Q1 - GitHub Closing

Working directory in Windows is "C:/Users/Gautam/Documents/GMS/Sem 1 - 01 - Doing Data Science/Lecture 03"

Commands are executed in GitBash, shown below:

```
# cd Documents/GMS/Sem\ 1\ -\ 01\ -\ Doing\ Data\ Science/Lecture\ 03/
# git init
# git clone https://github.com/caesar0301/awesome-public-datasets
```

Q2 - Data Summary

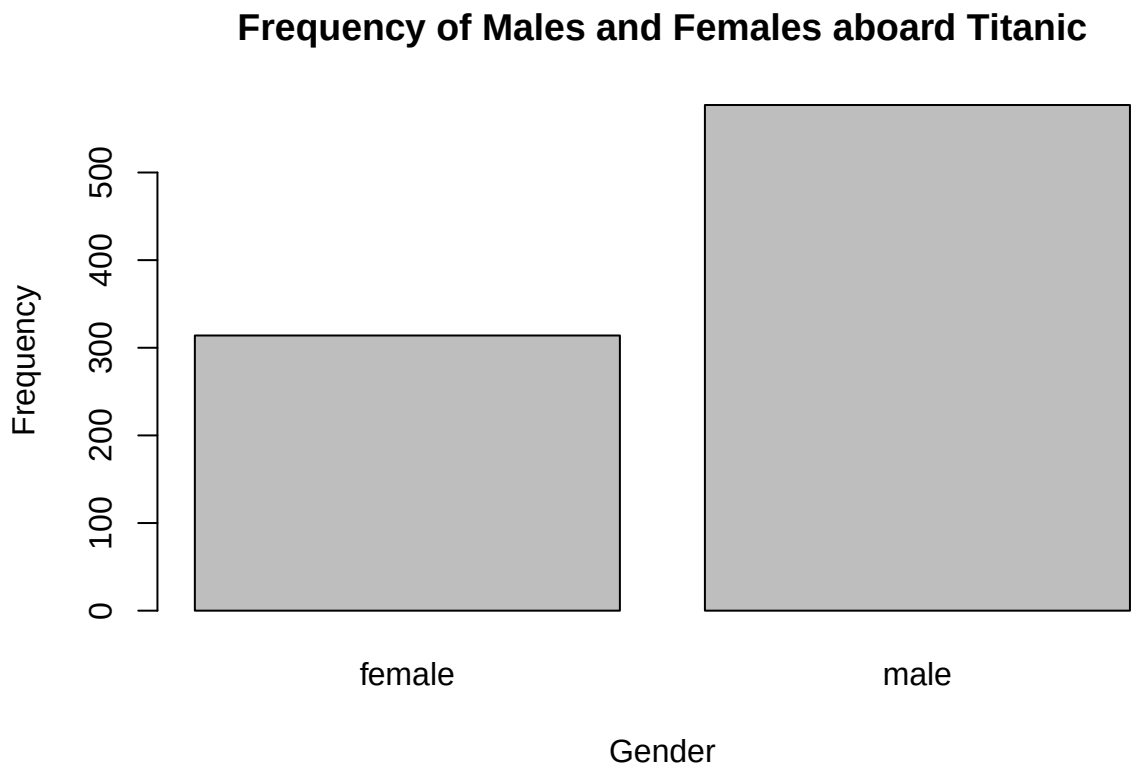
```
library(plyr)

sessionInfo()
```

```
## R version 3.5.3 (2019-03-11)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] plyr_1.8.4
##
## loaded via a namespace (and not attached):
##  [1] compiler_3.5.3  magrittr_1.5    tools_3.5.3     htmltools_0.3.6
##  [5] yaml_2.2.0      Rcpp_1.0.1      stringi_1.4.3   rmarkdown_1.12
##  [9] knitr_1.23      stringr_1.4.0   xfun_0.7        digest_0.6.18
## [13] evaluate_0.13
```

```
tdf <- read.table("titanic.csv/titanic.csv", TRUE, sep = ",")
cnt <- count(tdf$Sex)
cnt
```

```
##        x freq
## 1 female  314
## 2   male  577
```

```
barplot(cnt$freq,names.arg=c(levels(cnt$x)),xlab = 'Gender',ylab = 'Frequency',main = 'Frequency of Mal
```




```
# Extracting columns for age, fare and survival below
afs <- tdf[,c(6,10,2)]
# Calculating mean per column
sapply(afs, mean,na.rm=TRUE)
```

```
##        Age       Fare   Survived
## 29.6991176 32.2042080  0.3838384
```

Q3 - Function Building

```
# Function to take in file name, and create corresponding objects

sleepDataAnalysis <- function(fileName='') {
  sData     <- read.table(fileName, TRUE, sep = ",", na.strings = c(" ","NA"))
  minSleep  <- min(sData$Duration,na.rm = TRUE)
  maxSleep  <- max(sData$Duration,na.rm = TRUE)
  medianAge <- median(sData$Age,na.rm = TRUE)
  mRSES     <- mean(sData$RSES,na.rm = TRUE)
  sdRSES    <- sd(sData$RSES,na.rm = TRUE)

  MedianAge <- medianAge
  SelfEsteem<- mRSES/5
  SE_SD     <- sdRSES/5
  DurationRange <- maxSleep - minSleep

  report    <- data.frame(MedianAge,SelfEsteem,SE_SD,DurationRange)

  return(round(report,2))

}

adf <- sleepDataAnalysis('sleep_data_01.csv')

adf
```

```
##   MedianAge SelfEsteem SE_SD DurationRange
## 1        14       3.62  1.24             7
```

Q4 FiveThirtyEight Data

```
sessionInfo()
```

```
## R version 3.5.3 (2019-03-11)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] plyr_1.8.4
```

```
##
## loaded via a namespace (and not attached):
##  [1] compiler_3.5.3  magrittr_1.5    tools_3.5.3     htmltools_0.3.6
##  [5] yaml_2.2.0      Rcpp_1.0.1      stringi_1.4.3   rmarkdown_1.12
##  [9] knitr_1.23      stringr_1.4.0   xfun_0.7        digest_0.6.18
## [13] evaluate_0.13
```

```
library(fivethirtyeight)

df <- college_recent_grads

dim(df)
```

```
## [1] 173  21
```

```
names(df)
```

```
##  [1] "rank"                        "major_code"
##  [3] "major"                       "major_category"
##  [5] "total"                       "sample_size"
##  [7] "men"                         "women"
##  [9] "sharewomen"                  "employed"
## [11] "employed_fulltime"           "employed_parttime"
## [13] "employed_fulltime_yearround" "unemployed"
## [15] "unemployment_rate"           "p25th"
## [17] "median"                      "p75th"
## [19] "college_jobs"                "non_college_jobs"
## [21] "low_wage_jobs"
```

Q5 Data Summary

```
names(df)
```

```
##  [1] "rank"                        "major_code"
##  [3] "major"                       "major_category"
##  [5] "total"                       "sample_size"
##  [7] "men"                         "women"
##  [9] "sharewomen"                  "employed"
## [11] "employed_fulltime"           "employed_parttime"
## [13] "employed_fulltime_yearround" "unemployed"
## [15] "unemployment_rate"           "p25th"
## [17] "median"                      "p75th"
## [19] "college_jobs"                "non_college_jobs"
## [21] "low_wage_jobs"
```

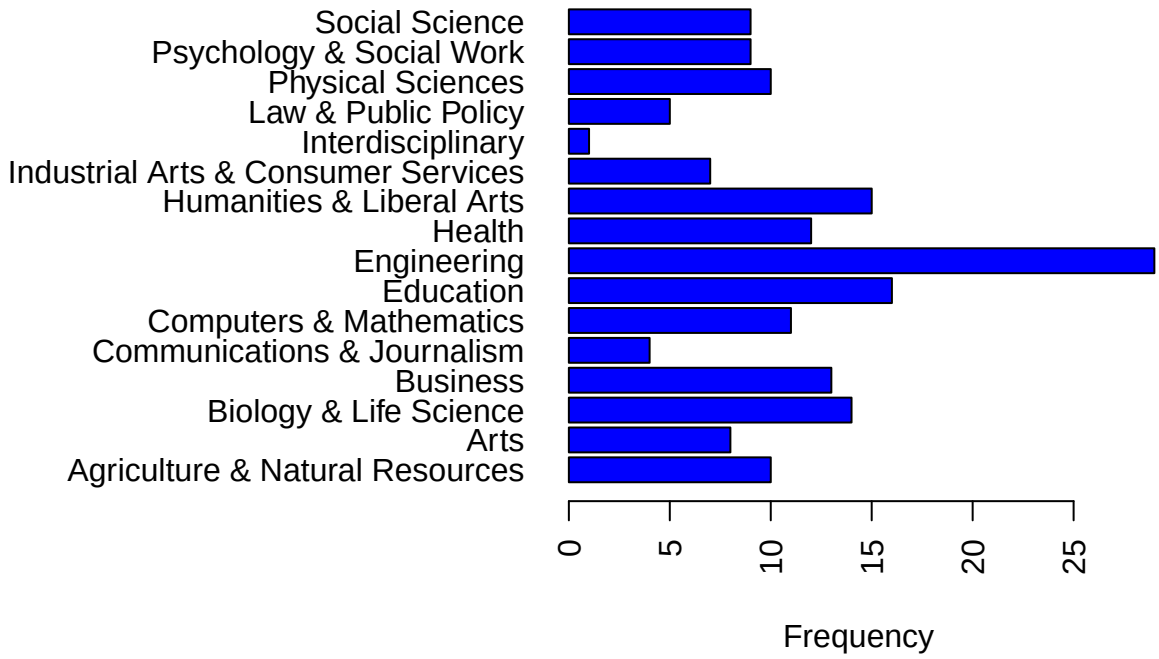
```
length(names(df))
```

```
## [1] 21
```

```r
library(plyr)

cnt.major_category = count(df$major_category)

cnt.major_category
```

```
##                                    x freq
## 1      Agriculture & Natural Resources   10
## 2                                  Arts    8
## 3                Biology & Life Science   14
## 4                              Business   13
## 5           Communications & Journalism    4
## 6               Computers & Mathematics   11
## 7                             Education   16
## 8                           Engineering   29
## 9                                Health   12
## 10           Humanities & Liberal Arts   15
## 11 Industrial Arts & Consumer Services    7
## 12                    Interdisciplinary    1
## 13                 Law & Public Policy    5
## 14                   Physical Sciences   10
## 15            Psychology & Social Work    9
## 16                       Social Science    9
```

```r
par(mar=c(5.1,15,4.1,2.1),las=2)

barplot(cnt.major_category$freq,names.arg=c(levels(cnt.major_category$x)),xlab = 'Frequency',main = 'Di
```




```
write.csv(df,file = 'output_data.csv',row.names = FALSE)
```

Q6 GitHub Repo for HomeWork

```
# https://github.com/gkapila07/msds_homeworks/tree/master/dds
```