STAT 40001/ STAT 50001 Statistical Computing Fall 2024

Lab-4

1) Go to http://jse.amstat.org/jse_data_archive.htm

a) Import the ***babyboom.dat.txt*** data which has Time of Birth, Sex, and Birth Weight of Babies and number of minutes after midnight.

b) How many observations are recorded?

c) Print first 5 observations

d) Print last 5 observations

```
> babyboom_data <- read.csv("https://jse.amstat.org/datasets/babyboom.dat.txt")
> num_observations <- nrow(babyboom_data)
> print(paste("Number of observations:", num_observations))
[1] "Number of observations: 43"
> print("First 5 observations:")
[1] "First 5 observations:"
> print(head(babyboom_data, 5))
    X0005.......1....3837.......5
1     0104       1    3334      64
2     0118       2    3554      78
3     0155       2    3838     115
4     0257       2    3625     177
5     0405       1    2208     245
> print("Last 5 observations:")
[1] "Last 5 observations:"
> print(tail(babyboom_data, 5))
     X0005.......1....3837.......5
39     2104       2    2121    1264
40     2123       2    3150    1283
41     2217       1    3866    1337
42     2327       1    3542    1407
43     2355       1    3278    1435
```

2) Consider a data set where the columns are separated by $

Col1$Col2$Col3
1$2$3
4$5$6
7$8$9
abc

a) Save the data in a local drive and import it in R by removing $ sign.
b) Export this data to create .csv dataset using write.csv() function

```
> data <- read.table("Q2.txt", sep = "$", header = TRUE)
> print("Imported Data:")
[1] "Imported Data:"
> print(data)
  Col1 Col2 Col3
1    1    2    3
2    4    5    6
3    7    8    9
4    a    b    c
> write.csv(data, file = "Q2.csv", row.names = FALSE)
> data <- read.table("Q2.txt", sep = "$", header = TRUE)
```

```
> print("Imported Data:")
[1] "Imported Data:"
> print(data)
  Col1 Col2 Col3
1    1    2    3
2    4    5    6
3    7    8    9
4    a    b    c
```

3) Weekly SST (sea surface temperature) data starting week of January 3,1990 are provided in the link below

https://www.cpc.ncep.noaa.gov/data/indices/wksst8110.for

a) Import the data in R and determine its dimension. (Note it is fwf data)
b) Print first five observations of the dataset

```
> col_widths <- c(11, 7, 7, 7, 7, 7, 7, 7, 7)
> SST <- read.fwf("https://www.cpc.ncep.noaa.gov/data/indices/wksst8110.for",
+                 widths = col_widths, skip = 4, header = FALSE)
> head(SST, 5)
          V1 V2      V3   V4   V5      V6  V7      V8  V9
1  03JAN1990  23 4-0.4   25.1 -0.3   26.6- 0.0  28.6 0 0.3
2  10JAN1990  23 4-0.8   25.2 -0.3   26.6  0.1  28.6 0 0.3
3  17JAN1990  24 2-0.3   25.3 -0.3   26.5- 0.1  28.6 0 0.3
4  24JAN1990  24 4-0.5   25.5 -0.4   26.5- 0.1  28.4 0 0.2
5  31JAN1990  25 1-0.2   25.8 -0.2   26.7  0.1  28.4 0 0.2
```

4) *bweight* data provided in the Brightspace are Stata data (.dta) sets describing the babyweight along with several variables. Import the data in R and identify the dimension of the data and extract the variables included in the datasets.

```
> library(haven)
> bweight <- read_dta("bweight.dta")
> dataset_dimensions <- dim(bweight)
> print(paste("Dataset Dimensions: ", dataset_dimensions[1], "rows and", dataset_dimen
sions[2], "columns"))
[1] "Dataset Dimensions:  4642 rows and 23 columns"
> variable_names <- colnames(bweight)
> print("Variables in the dataset:")
[1] "Variables in the dataset:"
> print(variable_names)
 [1] "bweight"    "mmarried"   "mhisp"      "fhisp"      "foreign"    "alcohol"    "de
adkids"
 [8] "mage"       "medu"       "fage"       "fedu"       "nprenatal"  "monthslb"   "or
der"
[15] "msmoke"     "mbsmoke"    "mrace"      "frace"      "prenatal"   "birthmonth" "lb
weight"
[22] "fbaby"      "prenatal1"
```

5) URL below contains data related to the number of publications produced by Ph.D. biochemists.

http://www.stata-press.com/data/lf2/couart2.dta

a) Import the data in R. (Note that this is Stata data)
b) List the variables included in the data
c) State the dimension of the data

```
> library(haven)
> Q5<- read_dta("http://www.stata-press.com/data/lf2/couart2.dta")
! curl package not installed, falling back to using `url()`
> head(Q5)
# A tibble: 6 × 6
    art   fem   mar  kid5   phd  ment
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1     0     0     1     0  2.52     7
2     0     1     0     0  2.05     6
3     0     1     0     0  3.75     6
4     0     0     1     1  1.18     3
5     0     1     0     0  3.75    26
6     0     1     1     2  3.59     2
> variable_names <- colnames(bweight_data)
> print("Variables included in the dataset:")
[1] "Variables included in the dataset:"
> print(variable_names)
[1] "art"  "fem"  "mar"  "kid5" "phd"  "ment"
> dataset_dimensions <- dim(bweight_data)
> print(paste("Dataset Dimensions: ", dataset_dimensions[1], "rows and", dataset_dimen
sions[2], "columns"))
[1] "Dataset Dimensions:  915 rows and 6 columns"
```