**Name: Niranjan Vinod Patil.**

**Batch: DA Lab-A2.**

**Roll No: SCETTYDSA15.**

**Course: Descriptive Analytics .**

**Data Preprocessing exercise using R**

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data

tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the *mean* of the data?What is the *median*?

(b) What is the *mode* of the data? Comment on the data’s modality (i.e., bimodal,

trimodal, etc.).

(c) What is the *midrange* of the data?

(d) Can you find (roughly) the first quartile (*Q*1) and the third quartile (*Q*3) of the data?

(e) Give the *five-number summary* of the data.

(f) Show a *boxplot* of the data.

(g) How is a *quantile-quantile plot* different from a *quantile plot*?

(a) Use *smoothing by bin means* to smooth the data, using a bin depth of 3. Illustrate

your steps. Comment on the effect of this technique for the given data.

(b) How might you determine *outliers* in the data?

(c) What other methods are there for *data smoothing*?

Plot an equal-width histogram of width 10.

(b) Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR,

cluster sampling, stratified sampling. Use samples of size 5 and the strata “youth,”

“middle-aged,” and “senior.”

Suppose a group of 12 *sales price* records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into three bins by each of the following methods:

(a) equal-frequency (equidepth) partitioning

(b) equal-width partitioning

(c) clustering

install.packages("dplyr")

library(dplyr)

age = c(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,

        33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)

mean(age)

median(age)

Mode = function(x){

  ta = table(x)

  tam = max(ta)

  if (all(ta == tam))

    mod = NA

  else

    if(is.numeric(x))

      mod = as.numeric(names(ta)[ta == tam])

  else

    mod = names(ta)[ta == tam]

  return(mod)

}

Mode(age)

midrange = (max(age) + min(age))/2

glimpse(midrange)

quantile(age, c(0.25, 0.5, 0.75), type = 1)

summary(age)

boxplot(age,

        main = "AGE",

        col = "blue")

OUTPUT:  
R version 3.5.1 (2018-07-02) -- "Feather Spray"

Copyright (C) 2018 The R Foundation for Statistical Computing

Platform: x86\_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.

You are welcome to redistribute it under certain conditions.

Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.

Type 'contributors()' for more information and

'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or

'help.start()' for an HTML browser interface to help.

Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

WARNING: You are configured to use the CRAN mirror at <https://cran.rstudio.com/>. This mirror supports secure (HTTPS) downloads however your system is unable to communicate securely with the server (possibly due to out of date certificate files on your system). Falling back to using insecure URL for this mirror.

To learn more and/or disable this warning message see the "Use secure download method for HTTP" option in Tools -> Global Options -> Packages.

> install.packages("dplyr")

Installing package into ‘C:/Users/Mehul/Documents/R/win-library/3.5’

(as ‘lib’ is unspecified)

trying URL '<http://cran.rstudio.com/bin/windows/contrib/3.5/dplyr_0.7.6.zip>'

Content type 'application/zip' length 3057193 bytes (2.9 MB)

downloaded 2.9 MB

package ‘dplyr’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\Mehul\AppData\Local\Temp\RtmpyighHa\downloaded\_packages

There were 50 or more warnings (use warnings() to see the first 50)

> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> age = c(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,

+         33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)

> mean(age)

[1] 29.96296

> median(age)

[1] 25

> Mode = function(x){

+   ta = table(x)

+   tam = max(ta)

+   if (all(ta == tam))

+     mod = NA

+   else

+     if(is.numeric(x))

+       mod = as.numeric(names(ta)[ta == tam])

+   else

+     mod = names(ta)[ta == tam]

+   return(mod)

+ }

> Mode(age)

[1] 25 35

> midrange = (max(age) + min(age))/2

> glimpse(midrange)

 num 41.5

> quantile(age, c(0.25, 0.5, 0.75), type = 1)

25% 50% 75%

 20  25  35

> summary(age)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  13.00   20.50   25.00   29.96   35.00   70.00

> boxplot(age,

+         main = "AGE",

+         col = "blue")