Session 5 – Data

Management using R

Assignment – 1

1. How many vowels are there in the names of USA States?
2. > names(USArrests) # to know what is the column name for State names
3. [1] "Murder" "Assault" "UrbanPop" "Rape"
4. > USA\_States <- rownames(USArrests) # names of states
5. > USA\_States
6. [1] "Alabama" "Alaska" "Arizona" "Arkansas" "California"
7. [6] "Colorado" "Connecticut" "Delaware" "Florida" "Georgia"
8. [11] "Hawaii" "Idaho" "Illinois" "Indiana" "Iowa"
9. [16] "Kansas" "Kentucky" "Louisiana" "Maine" "Maryland"
10. [21] "Massachusetts" "Michigan" "Minnesota" "Mississippi" "Missouri"
11. [26] "Montana" "Nebraska" "Nevada" "New Hampshire" "New Jersey"
12. [31] "New Mexico" "New York" "North Carolina" "North Dakota" "Ohio"
13. [36] "Oklahoma" "Oregon" "Pennsylvania" "Rhode Island" "South Carolina"
14. [41] "South Dakota" "Tennessee" "Texas" "Utah" "Vermont"
15. [46] "Virginia" "Washington" "West Virginia" "Wisconsin" "Wyoming"
16. > USA\_States <- paste(USA\_States, collapse = "") # converting the names to a string
17. > USA\_States <- tolower(strsplit(USA\_States, "")[[1]]) # converting to lower case and spliting each letter
18. > USA\_States <- USA\_States[USA\_States %in% letters]
19. > (table(USA\_States)) # Frequency of each letter
20. USA\_States
21. a b c d e f g h i j k l m n o p r s t u v w x y z
22. 61 2 12 11 28 2 8 15 44 1 10 15 14 43 36 4 22 32 19 8 5 11 2 6 1
23. > distribution <- as.data.frame(table(USA\_States)) # converted to data frame
24. > names(distribution)
25. [1] "USA\_States" "Freq"
26. > colnames(distribution) <- c("letters", "Freq") # changed column names
27. > names(distribution)
28. [1] "letters" "Freq"
29. > library(dplyr)
30. filter(distribution, letters %in% c("a","e","i","o","u")) # finding number of vowels in the names of USA States
31. letters Freq
32. 1 a 61
33. 2 e 28
34. 3 i 44
35. 4 o 36
36. 5 u 8

2. Visualize the vowels distribution.

vowel\_dist <- filter(distribution, letters %in% c("a","e","i","o","u"))

> vowel\_dist

letters Freq

1 a 61

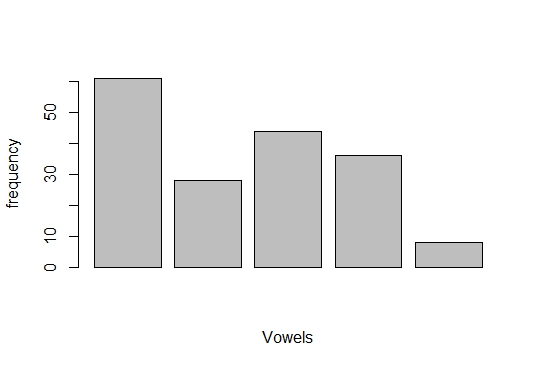
2 e 28

3 i 44

4 o 36

5 u 8

> barplot(vowel\_dist$Freq, axes = TRUE, axisnames = TRUE, xlab = "Vowels", ylab = "frequency")



library(ggplot2)

> # since we want vowel distribution, assiging colour to position of vowel

> area.color <- c("withcolour",NA,NA,NA,"withcolour",NA,NA,NA,"withcolour",

+ NA,NA,NA,NA,NA,"withcolour",NA,NA,NA,NA,"withcolour",

+ NA,NA,NA,NA,NA)

> area.color

[1] "withcolour" NA NA NA "withcolour" NA

[7] NA NA "withcolour" NA NA NA

[13] NA NA "withcolour" NA NA NA

[19] NA "withcolour" NA NA NA NA

[25] NA

plot.vowel <- ggplot(data = distribution, aes(x=letters, y=Freq, fill=area.color))+

+ geom\_bar(stat = "identity") +

+ xlab(colnames(distribution)[1]) +

+ ylab(colnames(distribution)[2])

> plot.vowel # vowel distribution is visualized with different colour

