Chad Huntebrinker’s Midterm

Chad Huntebrinker

2024-10-19

#Chad Huntebrinker  
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

#Read Data  
cell.data <- read.csv("cell\_clinical\_data.csv")  
  
#Part 1: Clean the data  
#Problem 1  
colnames(cell.data)

## [1] "Study.ID" "Patient.ID"   
## [3] "Sample.ID" "Age"   
## [5] "Annotation.Source" "Cancer.Type"   
## [7] "Cancer.Type.Detailed" "Cell.Line.Source"   
## [9] "DepMap.ID" "Ethnicity.Category"   
## [11] "Fraction.Genome.Altered" "Growth.Medium"   
## [13] "Mutation.Count" "Mutation.Rate"   
## [15] "Oncotree.Code" "Ploidy"   
## [17] "Primary.Tumor.Site" "Race.Category"   
## [19] "Sample.Type" "Sex"   
## [21] "TMB..nonsynonymous." "Tumor.Type"

#Problem 2  
Patient.ID\_duplicates <- cell.data[duplicated(cell.data$Patient.ID), ]  
Sample.ID\_duplicates <- cell.data[duplicated(cell.data$Sample.ID), ]  
  
#Problem 3  
cell.data$Pt.id <- paste0("Pt\_", seq(1, nrow(cell.data)))  
  
#Problem 4  
cell.data$Pt.id <- paste0(cell.data$Pt.id, "\_", sub("\_.\*", "", cell.data$Sample.ID))  
  
#Problem 5  
cell.data <- cell.data %>%  
 relocate(Pt.id, .before = Study.ID)  
  
#Problem 6  
#Change all the unwanted columns to NULL to remove them  
cell.data$Patient.ID <- NULL  
cell.data$Sample.ID <- NULL  
cell.data$Annotation.Source <- NULL  
cell.data$Cell.Line.Source <- NULL  
cell.data$Cancer.Type.Detailed <- NULL  
cell.data$DepMap.ID <- NULL  
cell.data$Race.Category <- NULL  
cell.data$Sample.Type <- NULL  
cell.data$TMB..nonsynonymous. <- NULL  
cell.data$Tumor.Type <- NULL  
  
#Problem 7  
cell.data$Age[is.na(cell.data$Age)] <- sample(1:100, sum(is.na(cell.data$Age)), replace = TRUE)  
  
#Problem 8  
cell.data <- cell.data %>%  
 rename(Ethnicity = Ethnicity.Category)  
  
#Problem 9  
ethnicities <- c("White", "Hispanic", "Black")  
cell.data$Ethnicity[is.na(cell.data$Ethnicity)] <- sample(ethnicities, sum(is.na(cell.data$Ethnicity)), replace = TRUE)  
  
#Problem 10  
min\_value <- min(cell.data$Fraction.Genome.Altered, na.rm = TRUE)  
max\_value <- max(cell.data$Fraction.Genome.Altered, na.rm = TRUE)  
  
cell.data$Fraction.Genome.Altered[is.na(cell.data$Fraction.Genome.Altered)] <-   
 runif(sum(is.na(cell.data$Fraction.Genome.Altered)), min = min\_value, max = max\_value)  
  
#Problem 11  
media\_types <- c("DMEM", "RPMI-1640", "MEM", "EMEM", "FBS")  
  
cell.data$Growth.Medium[is.na(cell.data$Growth.Medium)] <-   
 sample(media\_types, sum(is.na(cell.data$Growth.Medium)), replace = TRUE)  
  
#Problem 12  
cell.data$Mutation.Count[is.na(cell.data$Mutation.Count)] <-   
 sample(1:20000, sum(is.na(cell.data$Mutation.Count)), replace = TRUE)  
  
#Problem 13  
cell.data$Mutation.Rate[is.na(cell.data$Mutation.Rate)] <-   
 with(cell.data, Mutation.Count[is.na(Mutation.Rate)] / (Fraction.Genome.Altered[is.na(Mutation.Rate)] \* 3000000))  
  
#Problem 14  
cell.data$Ploidy[nchar(cell.data$Ploidy) > 4] <- NA  
  
#Problem 15  
min\_value <- as.numeric(min(cell.data$Ploidy, na.rm = TRUE))  
max\_value <- as.numeric(max(cell.data$Ploidy, na.rm = TRUE))  
  
cell.data$Ploidy[is.na(cell.data$Ploidy)] <- round(runif(sum(is.na(cell.data$Ploidy)), min = min\_value, max = max\_value), 2)  
  
#Problem 16  
cell.data$Primary.Tumor.Site <- sub("\_.\*", "", cell.data$Primary.Tumor.Site)  
  
#Problem 17  
#Rename all the values using ifelse statements  
cell.data$Primary.Tumor.Site <- ifelse(cell.data$Primary.Tumor.Site == "Large", "Large Intestine",   
 cell.data$Primary.Tumor.Site)  
cell.data$Primary.Tumor.Site <- ifelse(cell.data$Primary.Tumor.Site == "Biliary", "Biliary Tract",   
 cell.data$Primary.Tumor.Site)  
cell.data$Primary.Tumor.Site <- ifelse(cell.data$Primary.Tumor.Site == "Central", "Central Nervous System",   
 cell.data$Primary.Tumor.Site)  
cell.data$Primary.Tumor.Site <- ifelse(cell.data$Primary.Tumor.Site == "Haematopoietic And Lymphoid Tissue", "Haematopoietic",   
 cell.data$Primary.Tumor.Site)  
cell.data$Primary.Tumor.Site <- ifelse(cell.data$Primary.Tumor.Site == "Salivary", "Salivary Gland",   
 cell.data$Primary.Tumor.Site)  
cell.data$Primary.Tumor.Site <- ifelse(cell.data$Primary.Tumor.Site == "Small", "Small Intestine",   
 cell.data$Primary.Tumor.Site)  
cell.data$Primary.Tumor.Site <- ifelse(cell.data$Primary.Tumor.Site == "Soft", "Soft Tissue",   
 cell.data$Primary.Tumor.Site)  
cell.data$Primary.Tumor.Site <- ifelse(cell.data$Primary.Tumor.Site == "Upper", "Upper Aerodigestive Tract",   
 cell.data$Primary.Tumor.Site)  
cell.data$Primary.Tumor.Site <- ifelse(cell.data$Primary.Tumor.Site == "Urinary", "Urinary Tract",   
 cell.data$Primary.Tumor.Site)  
  
#Problem 18  
cell.data$Primary.Tumor.Site[is.na(cell.data$Primary.Tumor.Site)] <-   
 sub(" .\*", "", cell.data$Cancer.Type[is.na(cell.data$Primary.Tumor.Site)])  
  
#Problem 19  
sex\_type <- c("Male", "Female")  
  
cell.data$Sex[is.na(cell.data$Sex)] <- sample(sex\_type, sum(is.na(cell.data$Sex)), replace = TRUE)  
  
cell.data$Sex[cell.data$Cancer.Type == "Breast Cancer"] <- "Female"  
cell.data$Sex[cell.data$Cancer.Type == "Prostate Cancer"] <- "Male"  
  
#Problem 20  
anyNA(cell.data)

## [1] FALSE

#Part 2: Analysis and interpretation  
#Problem 1  
summary(cell.data)

## Pt.id Study.ID Age Cancer.Type   
## Length:2826 Length:2826 Min. : 0.25 Length:2826   
## Class :character Class :character 1st Qu.: 29.00 Class :character   
## Mode :character Mode :character Median : 52.00 Mode :character   
## Mean : 49.89   
## 3rd Qu.: 70.00   
## Max. :100.00   
## Ethnicity Fraction.Genome.Altered Growth.Medium   
## Length:2826 Min. :0.0117 Length:2826   
## Class :character 1st Qu.:0.2318 Class :character   
## Mode :character Median :0.4757 Mode :character   
## Mean :0.4540   
## 3rd Qu.:0.6458   
## Max. :0.9771   
## Mutation.Count Mutation.Rate Oncotree.Code Ploidy   
## Min. : 8.0 Min. : 0.0000 Length:2826 Length:2826   
## 1st Qu.: 60.0 1st Qu.: 0.0001 Class :character Class :character   
## Median : 218.0 Median : 0.0006 Mode :character Mode :character   
## Mean : 1401.0 Mean : 65.8118   
## 3rd Qu.: 425.8 3rd Qu.: 104.9339   
## Max. :19713.0 Max. :3119.6184   
## Primary.Tumor.Site Sex   
## Length:2826 Length:2826   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##

str(cell.data)

## 'data.frame': 2826 obs. of 13 variables:  
## $ Pt.id : chr "Pt\_1\_127399" "Pt\_2\_1321N1" "Pt\_3\_1321N1" "Pt\_4\_143B" ...  
## $ Study.ID : chr "ccle\_broad\_2019" "ccle\_broad\_2019" "cellline\_ccle\_broad" "ccle\_broad\_2019" ...  
## $ Age : num 41 21 33 13 41 28 9 97 72 73 ...  
## $ Cancer.Type : chr "Soft Tissue Sarcoma" "Glioma" "Glioma" "Bone Cancer" ...  
## $ Ethnicity : chr "Black" "Caucasian" "Black" "Caucasian" ...  
## $ Fraction.Genome.Altered: num 0.66 0.565 0.564 0.501 0.5 ...  
## $ Growth.Medium : chr "EMEM" "EMEM" "DMEM" "MEM" ...  
## $ Mutation.Count : int 163 13784 950 8609 15333 290 2186 260 2272 248 ...  
## $ Mutation.Rate : num 8.23e-05 8.14e-03 5.61e-04 5.72e-03 1.02e-02 ...  
## $ Oncotree.Code : chr "SYNS" "ASTR" "GBM" "OS" ...  
## $ Ploidy : chr "8.09" "3.45" "8.25" "3.61" ...  
## $ Primary.Tumor.Site : chr "Soft" "Central Nervous System" "Central Nervous System" "Bone" ...  
## $ Sex : chr "Female" "Male" "Female" "Female" ...

dim(cell.data)

## [1] 2826 13

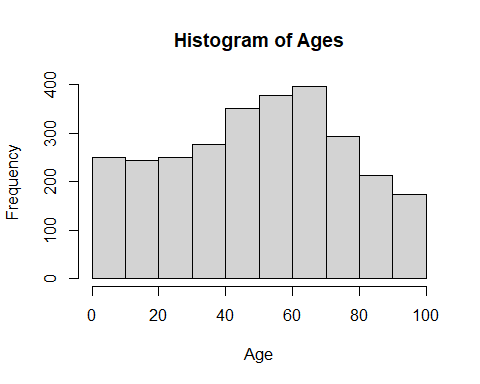
names(cell.data)

## [1] "Pt.id" "Study.ID"   
## [3] "Age" "Cancer.Type"   
## [5] "Ethnicity" "Fraction.Genome.Altered"  
## [7] "Growth.Medium" "Mutation.Count"   
## [9] "Mutation.Rate" "Oncotree.Code"   
## [11] "Ploidy" "Primary.Tumor.Site"   
## [13] "Sex"

#Problem 2  
table(cell.data$Study.ID)

##   
## ccle\_broad\_2019 cellline\_ccle\_broad cellline\_nci60   
## 1739 1020 67

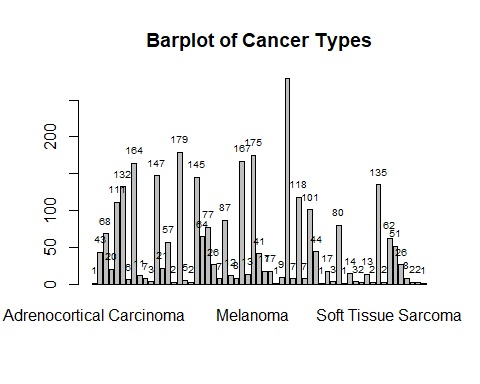
#Problem 3  
#I prefer to use a histogram because it shows all the ages and the number of times they are used  
hist(cell.data$Age, main = "Histogram of Ages", xlab = "Age")



#Problem 4  
cancer\_table <- table(cell.data$Cancer.Type)  
names(which.max(table(cell.data$Cancer.Type)))

## [1] "Non-Small Cell Lung Cancer"

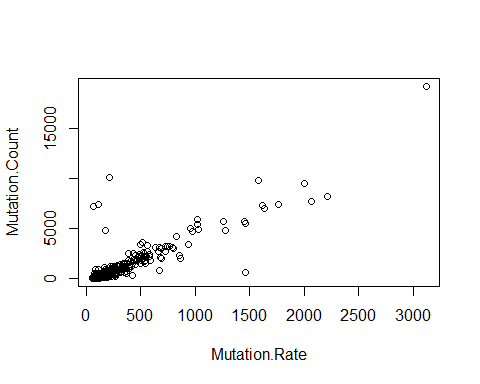
barplot\_cancer <- barplot(cancer\_table, main = "Barplot of Cancer Types",)  
text(barplot\_cancer, cancer\_table, labels = cancer\_table, cex = 0.6, pos = 3)



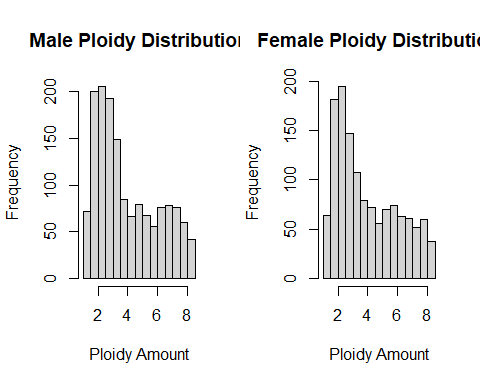
#Problem 5  
table(cell.data$Growth.Medium[cell.data$Cancer.Type == "Bladder Cancer"])

##   
## 2/3 DMEM:1/3 RPMI+10%FBS DMEM DMEM + 10% FBS   
## 1 9 2   
## DMEM + 10%FBS DMEM + 15% FBS DMEM +10 %FBS   
## 1 1 2   
## DMEM +10%FBS DMEM +15%FBS Eagle   
## 1 1 1   
## EMEM EMEM +10%FBS EMEM+10%FBS   
## 8 3 1   
## FBS Ham's F12 +10% FBS L15 + 10% FBS   
## 13 1 1   
## McCoy's 5A +10% McCoy's 5A +10% FBS MEM   
## 2 1 6   
## MEM + 10 % FBS RPMI-1640 RPMI +10%FBS   
## 1 8 2   
## RPMI 1640 + 10%FBS RPMI1640+10%FBS   
## 1 1

#Problem 6  
temp.cell.data <- cell.data[cell.data$Mutation.Rate > 1, ]  
plot(Mutation.Count~Mutation.Rate, data = temp.cell.data)



#Problem 7  
par(mfrow = c(1, 2))  
  
temp.cell.data <- cell.data[cell.data$Sex == "Male", ]  
hist(as.numeric(temp.cell.data$Ploidy), main = "Male Ploidy Distribution",  
 xlab = "Ploidy Amount")  
  
temp.cell.data <- cell.data[cell.data$Sex == "Female", ]  
hist(as.numeric(temp.cell.data$Ploidy), main = "Female Ploidy Distribution",  
 xlab = "Ploidy Amount")



#Problem 8  
cell.data.obj <- list(  
 c(cell.data$Cancer.Type),  
 c(cell.data$Growth.Medium),  
 c(cell.data$Primary.Tumor.Site),  
 c(cell.data$Oncotree.Code)  
)  
  
#Problem 9  
# Initialize a new column for age group  
cell.data$Age.Group <- NA  
  
# Categorize ages using a for loop  
for (i in 1:nrow(cell.data)) {  
 if (cell.data$Age[i] <= 20) {  
 cell.data$Age.Group[i] <- "Child"  
 }  
 else if (cell.data$Age[i] <= 60) {  
 cell.data$Age.Group[i] <- "Adult"  
 }  
 else {  
 cell.data$Age.Group[i] <- "Senior"  
 }  
}  
mean(cell.data$Mutation.Count[cell.data$Age.Group == "Child"])

## [1] 1550.923

mean(cell.data$Mutation.Count[cell.data$Age.Group == "Adult"])

## [1] 1412.203

mean(cell.data$Mutation.Count[cell.data$Age.Group == "Senior"])

## [1] 1319.28

con\_table <- table(cell.data$Study.ID, cell.data$Ethnicity)  
par(mfrow = c(1, 1))  
#con\_table <- as.data.frame(con\_table)  
barplot(con\_table, col = c("Red", "Blue", "Black"), main = "Distribution of Ethnicity")  
legend("topleft", legend = c("Red = ccle\_broad\_2019", "Blue = cellline\_ccle\_broad", "Black = cellline\_nci60"),  
 title = "Legend", cex = 0.7)

