Categorical Data

* Cancer type – Nominal Data
* Race – Nominal Data
* State – Nominal Data

Numerical Data

* Percent of the Population below Poverty – Continuous Numerical Data
* Percent of the Population Insured – Continuous Numerical Data
* Population – Discrete Numerical Data

Encoding

Table

Description automatically generated

Table

Description automatically generated

Summary

A screenshot of a computer

Description automatically generated with medium confidence

Glimpse

Text

Description automatically generated

Text

Description automatically generated with low confidence

# Load package

library(tidyverse)

glimpse(CancerCombinedFile3)

# Rename works for Columns

names(CancerCombinedFile3)[names(CancerCombinedFile3) == "State\_name"] <- "State"

names(CancerCombinedFile3)[names(CancerCombinedFile3) == "Race\_name"] <- "Race"

names(CancerCombinedFile3)[names(CancerCombinedFile3) == "Percentage population below poverty"] <- "Poverty%"

names(CancerCombinedFile3)[names(CancerCombinedFile3) == "Percentage population insured"] <- "Insured"

glimpse(CancerCombinedFile3)

Cancer <- CancerCombinedFile3[, 2:9]

head(Cancer, n=10)

#Re-coding

library(dplyr)

library(gapminder)

Cancer$NameR <- NA

Cancer$NameR[Cancer$Name=='Cervix Uteri'] <- 0

Cancer$NameR[Cancer$Name=='Corpus Uteri'] <- 1

Cancer$NameR[Cancer$Name=='Uterus, NOS'] <- 2

Cancer$NameR[Cancer$Name=='Ovary'] <- 3

Cancer$NameR[Cancer$Name=='Vagina'] <- 4

Cancer$NameR[Cancer$Name=='Vulva'] <- 5

Cancer$NameR[Cancer$Name=='Other Female Genital Organs'] <- 6

library(plyr)

# This rename code works for data in a column

Cancer$Race[Cancer$Race == "American Indian or Alaska Native"] <- "NativeAmer"

Cancer$Race[Cancer$Race == "Black or African American"] <- "AfricanAmer"

Cancer$Race[Cancer$Race == "White"] <- "Caucasian"

Cancer$Race[Cancer$Race == "Other Races and Unknown combined"] <- "Other"

Cancer$Race[Cancer$Race == "Asian or Pacific Islander"] <- "AsianAmer"

glimpse(Cancer)

summary(Cancer)

CancerPovPerOver20 <- filter(Cancer, 'Poverty%' > 20)

CancerPovPerOver20

CancerPovPerUnder9 <- filter(Cancer, 'Poverty%' < 9)

CancerPovPerUnder9

# Box Plot

boxplot(Cancer$Race, main="Box plot", ylab="Poverty%")

# Have to make race numeric for boxplot to work

unique(Cancer$Race)

Cancer$RaceR <- NA

Cancer$RaceR[Cancer$Race=="NativeAmer"]<-0

Cancer$RaceR[Cancer$Race=="AfricanAmer"]<-1

Cancer$RaceR[Cancer$Race=="Caucasian"]<-2

Cancer$RaceR[Cancer$Race=="Other"]<-3

Cancer$RaceR[Cancer$Race=="AsianAmer"]<-4

# Basic Graphs

boxplot(Cancer$RaceR, main="Box plot", ylab="Poverty%")

hist(Cancer$RaceR)

hist(Cancer$`Poverty%`)

# Histogram with 12 Bins

hist(Cancer$`Poverty%`,

breaks=12,

col="red",

xlab="Poverty Population %",

main="Colored histogram with 12 bins")

# Histogram with Rug Plot and Density Curve

hist(Cancer$`Poverty%`,

freq = FALSE,

breaks = 12,

col = "red",

xlab = "Histogram, Rug Plot, Density Curve")

rug(jitter(Cancer$`Poverty%`))

lines(density(Cancer$`Poverty%`), col = "blue", lwd=2)

# Histogram with Normal Curve and Box

x <- Cancer$`Poverty%`

h <- hist(x,

breaks=12,

col="red",

xlab="Histogram with Normal Curve and Box")

xfit <- seq(min(x), max(x), length=40)

yfit<-dnorm(xfit, mean=mean(x), sd=sd(x))

yfit <- yfit\*diff(h$mids[1:2])\*length(x)

lines(xfit, yfit, col="blue", lwd=2)

box()

# Boxplot

head(Cancer)

d2 <- ggplot(Cancer, aes(x = "RaceR", y = "Poverty%"))

d2 + geom\_boxplot() + xlab("")

d2 <- ggplot(Cancer, aes(x = "", y = "RaceR"))

d2 + geom\_boxplot() + xlab("")

# Descriptive Statistics

myvars <- c("RaceR", "Poverty%")

head(Cancer[myvars])

summary(Cancer[myvars])

mystats <- function(x, na.omit=FALSE){

if(na.omit)

x <- x[!is.na(x)]

m <- mean(x)

n <- length(x)

s <- sd(x)

skew <- sum((x-m)^3/s^3)/n

kurt <- sum((x-m)^4/s^4)/n-3

return(c(n=n, mean=m, stdev=s, skew=skew, kurtosis=kurt))

}

myvars <- c("RaceR", "Poverty%")

sapply(Cancer[myvars], mystats, na.omit=TRUE)

summary(Cancer$State)

Cancer[c("State")]

glimpse(Cancer)

unique(Cancer$State)

StateR <- as.numeric(State)

StateR

myvars <- names(Cancer) %in% c("RaceR.f", "Poverty%.f")

Cancer2 <- Cancer[!myvars]

StateR <- c("Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colorado", "Connecticut",

"District of Columbia", "Georgia", "Idaho", "Indiana", "Kansas", "Louisiana", "Maryland",

"Michigan", "Mississippi", "Montana", "Nevada", "New Jersey", "New York", "North Dakota",

"Oklahoma", "Pennsylvania", "South Carolina", "Tennessee", "Utah", "Virginia", "West Virginia",

"Wyoming", "Delaware", "Florida", "Hawaii", "Illinois", "Iowa", "Kentucky", "Maine", "Massachusetts",

"Minnesota", "Missouri", "Nebraska", "New Hampshire", "New Mexico", "North Carolina", "Ohio",

"Oregon", "Rhode Island", "South Dakota", "Texas", "Vermont", "Washington", "Wisconsin"

)

unique(Cancer$RaceR)

RaceR2 <- c(0, 1, 2, 3, 4)

unique(Cancer$`Poverty%`)

PovPerR <- c(18.4, 10.1, 17.7, 18.8, 15.8, 12.2, 10.4, 12.0, 17.9, 16.1, 17.8, 10.8, 15.2, 14.0, 15.0, 12.3, 13.3, 19.7, 13.5, 9.9, 11.4, 16.3, 22.3, 15.3, 14.9, 12.4, 8.5, 10.9, 20.9, 15.5, 16.8, 11.2, 15.4, 16.5, 15.7, 13.8, 17.2, 16.7, 11.7, 11.6, 12.7)

Cancer3 <- data.frame(RaceR2, StateR, PovPerR)

glimpse(Cancer2)

Pov\_num <- as.numeric(PovPerR)

Pov\_num

print(State)

Cancer$StateR <- NA

Cancer$StateR[Cancer$State=='Alabama'] <- 0

Cancer$StateR[Cancer$State=='Alaska'] <- 1

Cancer$StateR[Cancer$State=='Arizona'] <- 2

Cancer$StateR[Cancer$State=='Arkansas'] <- 3

Cancer$StateR[Cancer$State=='California'] <- 4

Cancer$StateR[Cancer$State=='Colorado'] <- 5

Cancer$StateR[Cancer$State=='Connecticut'] <- 6

Cancer$StateR[Cancer$State=='District of Columbia'] <- 7

Cancer$StateR[Cancer$State=='Georgia'] <- 8

Cancer$StateR[Cancer$State=='Idaho'] <- 9

Cancer$StateR[Cancer$State=='Indiana'] <- 10

Cancer$StateR[Cancer$State=='Kansas'] <- 11

Cancer$StateR[Cancer$State=='Louisiana'] <- 12

Cancer$StateR[Cancer$State=='Maryland'] <- 13

Cancer$StateR[Cancer$State=='Michigan'] <- 14

Cancer$StateR[Cancer$State=='Mississippi'] <- 15

Cancer$StateR[Cancer$State=='Montana'] <- 16

Cancer$StateR[Cancer$State=='Nevada'] <- 17

Cancer$StateR[Cancer$State=='New Jersey'] <- 18

Cancer$StateR[Cancer$State=='New York'] <- 19

Cancer$StateR[Cancer$State=='North Dakota'] <- 20

Cancer$StateR[Cancer$State=='Oklahoma'] <- 21

Cancer$StateR[Cancer$State=='Pennsylvania'] <- 22

Cancer$StateR[Cancer$State=='South Carolina'] <- 23

Cancer$StateR[Cancer$State=='Tennessee'] <- 24

Cancer$StateR[Cancer$State=='Utah'] <- 25

Cancer$StateR[Cancer$State=='Virginia'] <- 26

Cancer$StateR[Cancer$State=='West Virginia'] <- 27

Cancer$StateR[Cancer$State=='Wyoming'] <- 28

Cancer$StateR[Cancer$State=='Delaware'] <- 29

Cancer$StateR[Cancer$State=='Florida'] <- 30

Cancer$StateR[Cancer$State=='Hawaii'] <- 31

Cancer$StateR[Cancer$State=='Illinois'] <- 32

Cancer$StateR[Cancer$State=='Iowa'] <- 33

Cancer$StateR[Cancer$State=='Kentucky'] <- 34

Cancer$StateR[Cancer$State=='Maine'] <- 35

Cancer$StateR[Cancer$State=='Massachusetts'] <- 36

Cancer$StateR[Cancer$State=='Minnesota'] <- 37

Cancer$StateR[Cancer$State=='Missouri'] <- 38

Cancer$StateR[Cancer$State=='Nebraska'] <- 39

Cancer$StateR[Cancer$State=='New Hampshire'] <- 40

Cancer$StateR[Cancer$State=='New Mexico'] <- 41

Cancer$StateR[Cancer$State=='North Carolina'] <- 42

Cancer$StateR[Cancer$State=='Ohio'] <- 43

Cancer$StateR[Cancer$State=='Oregon'] <- 44

Cancer$StateR[Cancer$State=='Rhode Island'] <- 45

Cancer$StateR[Cancer$State=='South Dakota'] <- 46

Cancer$StateR[Cancer$State=='Texas'] <- 47

Cancer$StateR[Cancer$State=='Vermont'] <- 48

Cancer$StateR[Cancer$State=='Washington'] <- 49

Cancer$StateR[Cancer$State=='Wisconsin'] <- 50

Cancer2 <- Cancer[, 5:11]

head(Cancer2, n=10)