Assignment 3

Hunter GIles

3/9/2022

### Loading Data

# install.packages("readxl")  
library(readxl)  
data <- read\_excel("./data/Engineers.xlsx")  
# View(data)  
  
sum(is.na(data))

*readxl* is need to load the necessary excel file for this assignment.

### Question 1

1. Create a new variable “Age” that contains the engineers’ ages as of March 1, 2022. What is the average age of the engineers?

Birth year vector must be in the same format as end\_date to apply mathematical transformation

data$BirthDate <- as.Date(data$BirthDate, format="%m/%d/%y")  
end\_date <- as.Date("03/01/2022", format="%m/%d/%y")

Create *Age* column by taking the difference between BirthDate and end\_date. Divide by 365.25 to account for leap years. Then remove decimals.

data$Age <- difftime(end\_date, data$BirthDate)/365.25  
data$Age <- as.numeric(floor(data$Age))  
head(data$Age, 10)

## [1] 46 52 50 59 62 46 40 47 46 63

Output the average Age.

mean(data$Age)

## [1] 50.93

### Question 2

1. Bin the age values into three equal-size groups. Label the groups using numbers 1 (lowest age values) to 3 (highest age values). How many observations are in group 3?

data$AgeEqualSizeBin <- cut(data$Age, breaks=3, labels=c('1','2','3'), include.lowest=T, right=FALSE)  
table(data$AgeEqualSizeBin)

##   
## 1 2 3   
## 33 34 33

Group 3 frequency is 3.

### Question 3

1. Bin the annual salary values into four equal interval groups. Label the groups using numbers 1 (lowest salary values) to 4 (highest salary values). How many engineers are assigned to group 4?

bins <- quantile(data$Age, probs=seq(0,1,by=.25))  
data$AgeEqualIntervalBin <- cut(data$Age, breaks=bins, labels=c('1','2','3','4'), include.lowest=T, right=F)  
table(data$AgeEqualIntervalBin)

##   
## 1 2 3 4   
## 25 24 26 25

Group 4 frequency is 25

### Question 4

1. Bin the number of professional certificates achieved into the following three groups: < 2, between 2 and 4, and over 4. Label the groups “Low,” “Medium,” and “High.” How many engineers are in the “High” group?

data$CertBin <- cut(data$Certficates, breaks=c(0,2,4,Inf))  
table(data$CertBin)[3]

## (4,Inf]   
## 20

There are 20 engineers that have over 4 certificates (>4).

### Question 5

Transform the Personality variable into dummy variables. Use the most frequent category as the reference category.

Get the unique values in the personality vector to be transformed to dummies.

unique(data$Personality)

## [1] "Explorer" "Diplomat" "Analyst" "Sentinel"

for (i in unique(data$Personality)) {  
 col = paste("Personality", i, sep="\_")  
   
}