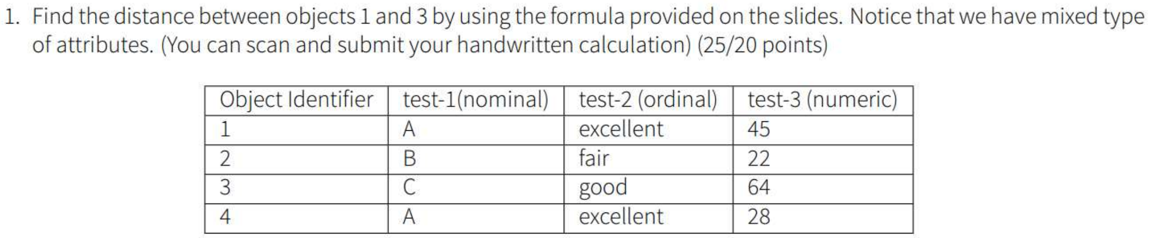
HW2

Saturday, February 15, 2025 2:03 PM

Q1



1. Find the distance between objects 1 and 3 by using the formula provided on the slides. Notice that we have mixed type of attributes. (You can scan and submit your handwritten calculation) (25/20 points)

Object Identifier	test-1(nominal)	test-2 (ordinal)	test-3 (numeric)
1	A	excellent	45
2	B	fair	22
3	C	good	64
4	A	excellent	28

Test-1

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ \boxed{1} & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Test-2

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ \boxed{.5} & .5 & 0 & \\ 0 & 1.0 & .5 & 0 \end{bmatrix}$$

Test-3

$$\begin{bmatrix} 0 & & & \\ .55 & 0 & & \\ \boxed{.45} & 1.0 & 0 & \\ .40 & .14 & .86 & 0 \end{bmatrix}$$

$$\frac{1\cdot(1) + 1(.5) + 1(.45)}{3} = \boxed{.65}$$

Q2

#R program to calculate manhattan Distance, a and b are both vectors of the same length

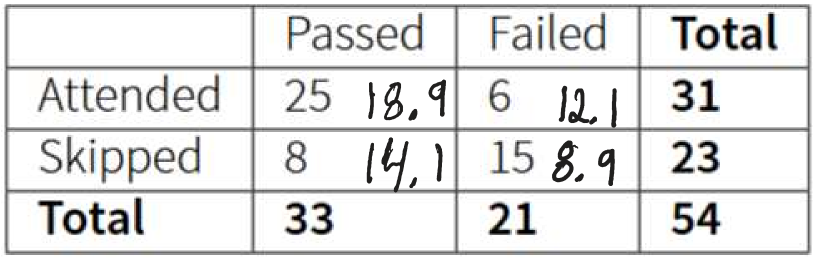
```
len <- 2
a <- sample(1:100, size=len)
b <- sample(1:100, size=len)

sum(abs(a-b))
```

#R program to calculate euclidian distance

```
sqrt(sum(((a-b)^2)))
```

Q3

	Passed	Failed	Total



	Passed	Failed	Total
Attended	25 18.9	6 12.1	**31**
Skipped	8 14.1	15 8.9	**23**
Total	**33**	**21**	**54**

$$\frac{(25-18.9)^2}{18.9}+\frac{(6-12.1)^2}{12.1}+\frac{(8-14.1)^2}{14.1}+\frac{(15-8.9)^2}{8.9}=11.864$$

Degrees of Freedom = 1

11.864 > 10.83 so *p* <.001. Reject Null hypothesis. *p* shows that there is a correlation between attending class and passing.

Q4

```
cor(mtcars$mpg, y = mtcars$wt)
```

Q5

```
#Read file
dat = read.csv("metabolite.csv")

#remove columns missing >75% of row data
dat_wout_missing = dat[, colMeans(is.na(dat)) < .75]

#replace NA values with column median
dat_cleaned <- lapply(dat_wout_missing, function(x) {
  if (is.numeric(x) | is.logical(x)) {
    x[is.na(x)] <- median(x, na.rm = TRUE)
  }
  return(x)
})

#converts the list back to a dataframe
dat_cleaned <- as.data.frame(dat_cleaned)


#Check to ensure no values remain as NA
sum(is.na(dat_cleaned))
```

Q6

```
#PCA

pca_results <- prcomp(dat_cleaned[2:188], retx = TRUE, center = TRUE, scale = TRUE)


# Create a new dataframe for plotting
pca_df <- data.frame(
  Class = dat_cleaned[, 1],    # Class labels
  PC1 = pca_results$x[, 1],    # First Principal Component
  PC2 = pca_results$x[, 2]     # Second Principal Component
)

# Scree plot

pca_var <- pca_results$sdev^2
pca_var_per <- round(pca_var/sum(pca_var)*100, 1)
barplot(pca_var_per, main = "Scree plot", xlab = "Principle Component", ylab = "Percent Variation")

# Create the PCA scatter plot
ggplot(pca_df, aes(x = PC1, y = PC2, color = Class)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(title = "PCA Plot",
       x = paste("PC1 ", pca_var_per[1], "%"),
       y = paste("PC2 ", pca_var_per[2], "%"))
```