HW2 R Notebook

2:

```
# Function to calculate Manhattan Distance
manhattan_distance <- function(vec1, vec2) { sum(abs(vec1 - vec2)) }

# Function to calculate Euclidean Distance
euclidean_distance <- function(vec1, vec2) { sqrt(sum((vec1 - vec2)^2)) }

# Example vectors
vec1 <- c(1, 2, 3, 4)
vec2 <- c(4, 3, 2, 1)

# Compute the distances
manhattan_dist <- manhattan_distance(vec1, vec2)
euclidean_dist <- euclidean_distance(vec1, vec2)

# Print the distances
print(paste("Manhattan distance:", manhattan_dist))
print(paste("Euclidean distance:", euclidean_dist))
```

4.

```
#Function to compute the correlation between miles per gallon, mpg, and weight, wt.
cor(mtcarsmpg, mtcarswt)

#Function to produce scatter plot
plot(mtcarswt, mtcarsmpg,
main = "Scatter Plot of MPG vs. Weight",
ylab = "mtcars$mpg",
xlab = "mtcars$wt", pch = 19)
```

5.

```
library(dplyr)

library(tidyr)


#Read the data from metabolite.csv into a dataframe
df <- read.csv("metabolite.csv")


#Remove columns with more than 75% missing values
threshold <- 0.75

df_clean <- df %>%

select_if(function(col) mean(is.na(col)) <= threshold)


#Replace missing values in the remaining columns with the median
df_final <- df_clean %>%

mutate(across(.cols = where(is.numeric),

.fns = ~ifelse(is.na(.), median(., na.rm = TRUE), .)))


#Print
print(df_final)
```

6.

```
library(dplyr)

library(tidyr)

library(ggplot2)

library(stats)


#Read the data from metabolite.csv into a dataframe

df <- read.csv("metabolite.csv")


#Remove columns with more than 75% missing values and replace missing values

threshold <- 0.75 df_clean <- df %>%

select_if(function(col) mean(is.na(col)) <= threshold) %>%

mutate(across(.cols = where(is.numeric),

.fns = ~ifelse(is.na(.), median(., na.rm = TRUE), .)))
```

df_pca_ready <- df_clean %>% select(-Label)

#Apply PCA

pca_result <- prcomp(df_pca_ready, center = TRUE, scale. = TRUE)

#Create a dataframe with PCA results and label for plotting

pca_data <- data.frame(PC1 = pca_result$x[, 1], $PC2 = pca_result$x[,2]) %>% bind_cols(df_clean %>% select(Label))

#Scatter plot using the first two principal components

ggplot(pca_data, aes(x = PC1, y = PC2, color = Label)) + geom_point() + theme_minimal() + labs(title = "PCA of Metabolites Data", x = "Principal Component 1", y = "Principal Component 2")