STAGE 2.1 task

# Load necessary libraries

library(ggplot2)

library(dplyr)

library(tidyr)

# Import the data

data\_url <- "https://raw.githubusercontent.com/HackBio-Internship/2025\_project\_collection/main/Python/Dataset/mcgc\_METADATA.txt"

df <- read.table(data\_url, header = TRUE, sep = "\t", stringsAsFactors = FALSE)

# Preview the dataset

print(head(df))

# Convert to long format

df\_long <- df %>%

pivot\_longer(cols = -Strain, names\_to = "Condition", values\_to = "OD600") %>%

mutate(

Mutant = ifelse(grepl("MUT", Condition), "+", "-"), # Identify knock-in (+) vs. knock-out (-)

Time = rep(seq(0, 10, length.out = nrow(df)), each = ncol(df) - 1) # Assuming time series data

)

# Convert OD600 to numeric (if needed)

df\_long$OD600 <- as.numeric(gsub("[A-Za-z]", "", df\_long$OD600)) # Extract numeric values

df\_long <- na.omit(df\_long) # Remove any missing values

# Function to determine the time to reach carrying capacity

find\_carrying\_capacity\_time <- function(strain\_data) {

if (nrow(strain\_data) == 0) return(NA)

max\_od600 <- max(strain\_data$OD600, na.rm = TRUE)

time\_to\_capacity <- strain\_data$Time[which.max(strain\_data$OD600)]

return(time\_to\_capacity)

}

# Compute carrying capacity time for each strain

capacity\_times <- df\_long %>%

group\_by(Strain, Mutant) %>%

summarise(Carrying\_Capacity\_Time = find\_carrying\_capacity\_time(cur\_data()), .groups = "drop")

# Ensure carrying capacity times are correctly merged

df\_long <- left\_join(df\_long, capacity\_times, by = c("Strain", "Mutant"))

# 1️⃣ \*\*Plot Growth Curves\*\*

ggplot(df\_long, aes(x = Time, y = OD600, color = Mutant)) +

geom\_line(aes(group = interaction(Strain, Condition)), alpha = 0.8) +

facet\_wrap(~ Strain) +

labs(title = "Growth Curves of OD600 vs Time",

x = "Time (hours)",

y = "OD600",

color = "Mutation Type") +

theme\_minimal()

# 2️⃣ \*\*Scatter Plot: Time to Reach Carrying Capacity\*\*

ggplot(capacity\_times, aes(x = factor(Mutant), y = Carrying\_Capacity\_Time, color = factor(Mutant))) +

geom\_jitter(width = 0.2, size = 3, alpha = 0.7) +

labs(title = "Scatter Plot: Time to Reach Carrying Capacity",

x = "Mutation Type (- = Knock Out, + = Knock In)",

y = "Time to Carrying Capacity (hours)") +

theme\_minimal()

# 3️⃣ \*\*Box Plot: Carrying Capacity Comparison\*\*

ggplot(capacity\_times, aes(x = factor(Mutant), y = Carrying\_Capacity\_Time, fill = factor(Mutant))) +

geom\_boxplot(alpha = 0.5) +

labs(title = "Comparison of Carrying Capacity Time",

x = "Mutation Type",

y = "Time to Carrying Capacity (hours)") +

theme\_minimal()

# 4️⃣ \*\*Statistical Test (t-test)\*\*

knock\_out\_times <- capacity\_times$Carrying\_Capacity\_Time[capacity\_times$Mutant == "-"]

knock\_in\_times <- capacity\_times$Carrying\_Capacity\_Time[capacity\_times$Mutant == "+"]

if (length(knock\_out\_times) > 1 & length(knock\_in\_times) > 1) {

t\_test\_result <- tryCatch({

t.test(knock\_out\_times, knock\_in\_times)

}, error = function(e) {

print("Error in t-test: Check data consistency.")

})

print(t\_test\_result)

} else {

print("Not enough data for t-test")

}

# \*\*Observations (Included in Code)\*\*

# - The \*\*growth curve plot\*\* shows OD600 over time for different strains.

# - The \*\*scatter plot\*\* visualizes how long each strain took to reach carrying capacity.

# - The \*\*box plot\*\* compares the distributions of knock-out (-) and knock-in (+) strains.

# - The \*\*t-test\*\* determines if there is a statistically significant difference.