Problem Set 2

```
################################################################################
# Problem Set 2
################################################################################

# For this problem set, you will analyze and combine the following two
# state-level datasets:
# - econ_data.xlsx (Economic data)
# - nbd_data.csv (Neighborhood data)
# Variables are described in the problem set description on Canvas.

################################################################################

# Uncomment the commands below to load needed packages
#library(tidyverse)
#library(readxl) # Install if not already installed

################################################################################
# 1. Import both Economic and Neighborhood data into R
################################################################################

# To import .xlsx files:  mydata <- read_excel("filename.xlsx")
# To import .csv files:   mydata <- read.csv("filename.csv")

# You need to install and load readxl to use read_excel()
# Look at the names of the files in the folder and replace filename with those names
# Replace mydata with more intutive names (eg. econ_data, nbd_data)
# If you are successful, both datasets will appear under data on the top-right

################################################################################
# 2. Use either the mean() or summarize() function to find the average median
#    household income (medHHinc) and the average percentage of the population
#    living within half a mile of a park (parks).
################################################################################



################################################################################
# 3. Merge both the datasets based on state names.
################################################################################

# To merge data: merged_data <- merge(data_name1, data_name2, by="unit")

# Replace unit with the name of the variable used for merging.
# Replace data_name1 and data_name2 with names you assigned to the datasets
# while importing
```

```
################################################################################
# 4. Find and remove observations with missing values on the variable poverty
#    from the merged data
################################################################################

# If a variable var has missing values, the argument is.na(var) returns TRUE.

# Which states had missing values on poverty? Use filter() and is.na() as follows:
# merged_data %>% filter(is.na(var)==TRUE)

# To remove missing observations, once again use filter() and is.na()

################################################################################
# 5. Find the state with the highest poverty rate using the functions filter()
#    and max().
################################################################################

# Which state has highest poverty rate?

################################################################################
# 6. Save your resulting dataset in R format.
#    (R's native format has extension .rda or .Rdata)
################################################################################

# save(merged_data, file="merged_data.rda")
# Check if you did indeed save this data in your folder.

################################################################################
# 7. Use the mutate() and log() functions to create a new variable called
#    log_hhinc that is equal to log of medHHinc. Save this new variable to your
#    existing merged data.
################################################################################


################################################################################
# 8. Create a scatter plot with log_hhinc on the x-axis and fine_partc_mttr on
#    the y-axis. Use your favorite color for the points on the plot.
################################################################################


################################################################################
# 9. Use the mutate() function to create a new variable called high_income
#    that takes value 1 whenever medHHinc is above 45,000 and 0 otherwise.
#    Save this new variable to your existing merged data.
# Note: Income is reported in 1000s of dollars.
################################################################################

# How many states are categorized as high income? You can use table() or count().

################################################################################
# 10. Use the functions group_by() and summarize() to find the average fine
#     particulate matter for states with high_income=1 and high_income=0.
################################################################################
```

```
################################################################################

# Make sure to remove all unnecessary comments, such as hints or notes I've provided,
# so that your final script for submission only contains the essential commands
# required for executing tasks and any comments you've added.
```