Problem Set 2 Solution

```r
# Housekeeping
library(tidyverse)
library(readxl)

# 1. Import data
econ_data <- read_excel("econData.xlsx")
nbd_data <- read.csv("nbd_data.csv")

# 2. Average medHHinc and parks
econ_data %>% summarise(mean = mean(medHHinc))
mean(nbd_data$parks)

# 3. Merge data
merged_data <- merge(econ_data, nbd_data, by="state")

# 4. Remove missing values
merged_data %>% filter(is.na(poverty)==TRUE) # Alaska and Hawaii have missing values
merged_data <- merged_data %>% filter(is.na(poverty)==FALSE)

# 5. State with the highest poverty rate is Tennessee
merged_data %>% filter(poverty==max(poverty))

# 6. Saving data in R format
save(merged_data, file="merged_data.rda")

# 7. Create log_hhinc
merged_data <- merged_data %>%
  mutate(log_hhinc = log(medHHinc))

# 8. Scatter plot
ggplot(merged_data, aes(x = log_hhinc, y=fine_partc_mttr)) +
  geom_point(color="orange") +
  theme_classic()

# 9. Create a new variable called high_income
merged_data <- merged_data %>%
  mutate(high_income = ifelse(medHHinc>45, 1, 0))
merged_data %>% count(high_income) # 22 high income states

# 10. Average particulate matter by high_income
merged_data %>%
  group_by(high_income) %>%
  summarise(mean = mean(fine_partc_mttr))
```