Robert Baker
23371822
Postgraduate Diploma in Applied Social Data Science

Problem Set 1

29th September 2023

Question 1 - Education

A school counselor was curious about the average IQ of the students in her school and took a random sample of 25 students' IQ scores. The following is the data set:

```
y <− c(105,69,86,100,82,111,104,110,87,108,87,90,94,113,112,98,80,97,95,111,114,89,95,126,98)
```

1. Find a 90% confidence interval for the average student IQ in the school

(Please also see the code document for a full breakdown of the code and steps)

After loading the necessary packages I inspected and explored the data

To find the confidence interval:

Calculated the sample mean - 98.44

Calculated the sample standard dev - 13.09

Calculated the sample Square Root - 5 - sqrt(length(y))

Using the above (standard dev / square root) I calculated the standard error - 2.62. This helps estimate the range within which the average IQ of the students in the school population will probably fall

We then need to calculate the margin of error. To do so I used standard (z) normal distribution, noting that it was a two-tailed test (as the teacher was not explicitly looking at higher or lower than the average)

Here I used the qnorm function in R to find the upper and lower bounds e.g. lower bound -

```
lower_90   <- qnorm(0.05 / 2,
             mean = mean_y,
             sd = sd_y)
```

Using the above steps - and based on the sample data in the vector (y) - we can estimate with c. 90% confidence that the mean IQ of the students in the school population is likely to fall within the range of approximately 73 to 98.

Next, the school counselor was curious whether the average student IQ in her school is higher than the average IQ score (100) among all the schools in the country. Using the same sample, conduct the appropriate hypothesis test with α = 0.05.

First I set up the null hypothesis & alternative hypothesis:

Null hypothesis: average student IQ is not higher than nat. av. ($\mu <= 100$)

Alternative hypothesis: average student IQ is higher than nat. av. ($\mu > 100$)

I then calculated the Z-Statistic - of -0.60 - using the formula (mean_y - 100) / standard_error_y

This z-statistic of approx. -0.6 suggests the sample mean IQ is about 0.6 standard errors below the population mean of 100

I then worked out a probability value (of approx. 0.72) using the formula: 1 - pnorm(z_stat)

It is a right-tailed test as we are looking at 'higher than the average' i.e. so we are looking at the area to the right of the z-stat in a standard normal distribution

A p-value of 0.72 is relatively high. This means under the null hypothesis, there is a c. 72% chance of observing a sample mean IQ as low as this, or lower, from random sampling variability. Given our significance level was 0.05 we don't have enough evidence to support rejecting H0. This is backed up by our confidence interval of 90% for the estimated school population mean IQ which was 73 to 98 (i.e. below 100).

Question 2 - Political Science

Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

State - 50 states in US

Y - per capita expenditure on shelters/housing assistance in state

X1 - per capita personal income in state

X2 - Number of residents per 100,000 that are ”financially insecure” in state

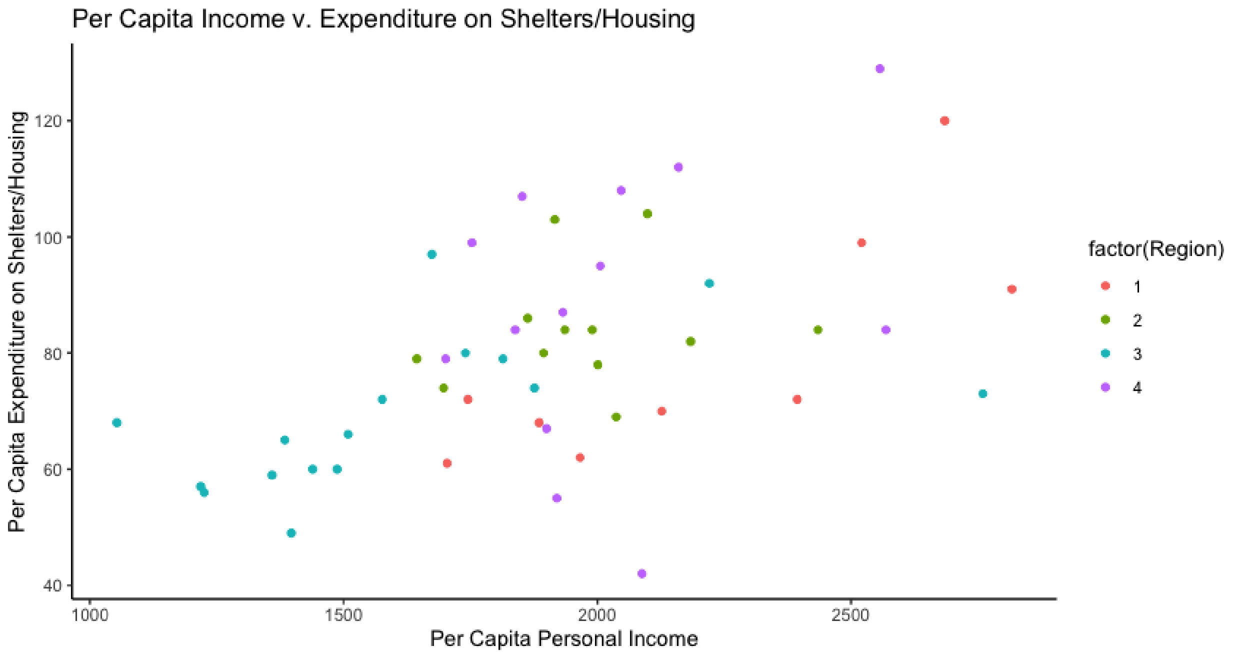
X3 - Number of people per thousand residing in urban areas in state

Region - 1=Northeast, 2= North Central, 3= South, 4=West

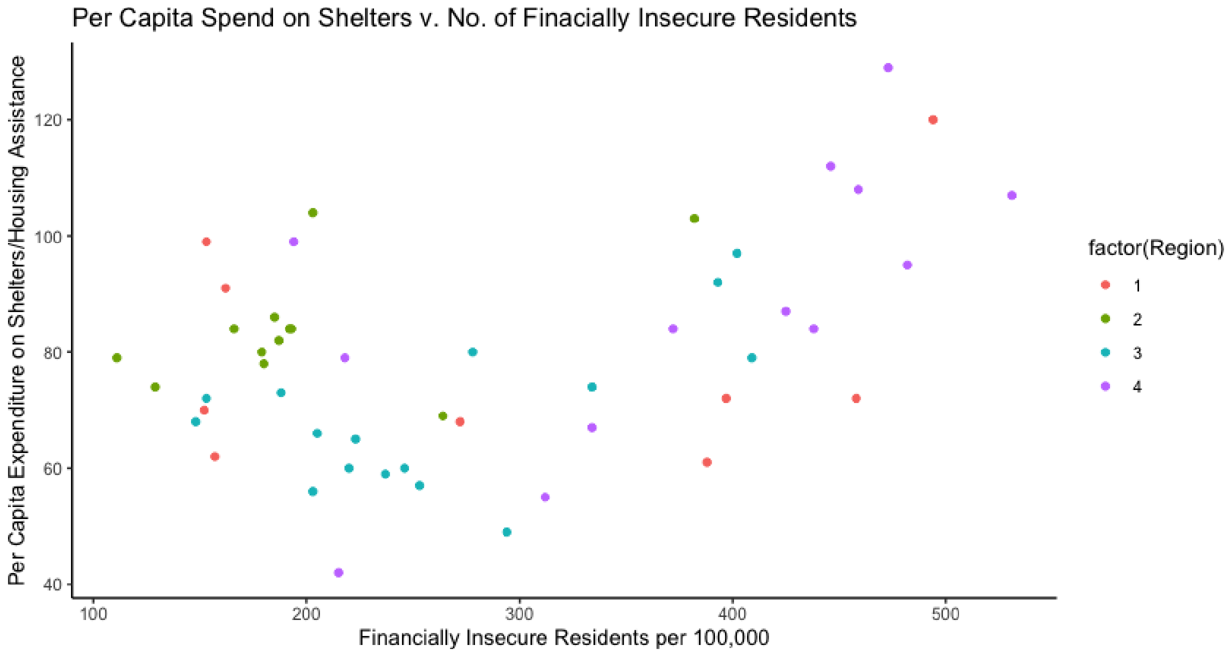
———————————

I first loaded the data - using read.table() - and the necessary packages. I then explored the data using head(), summary(), View(), str() & levels()

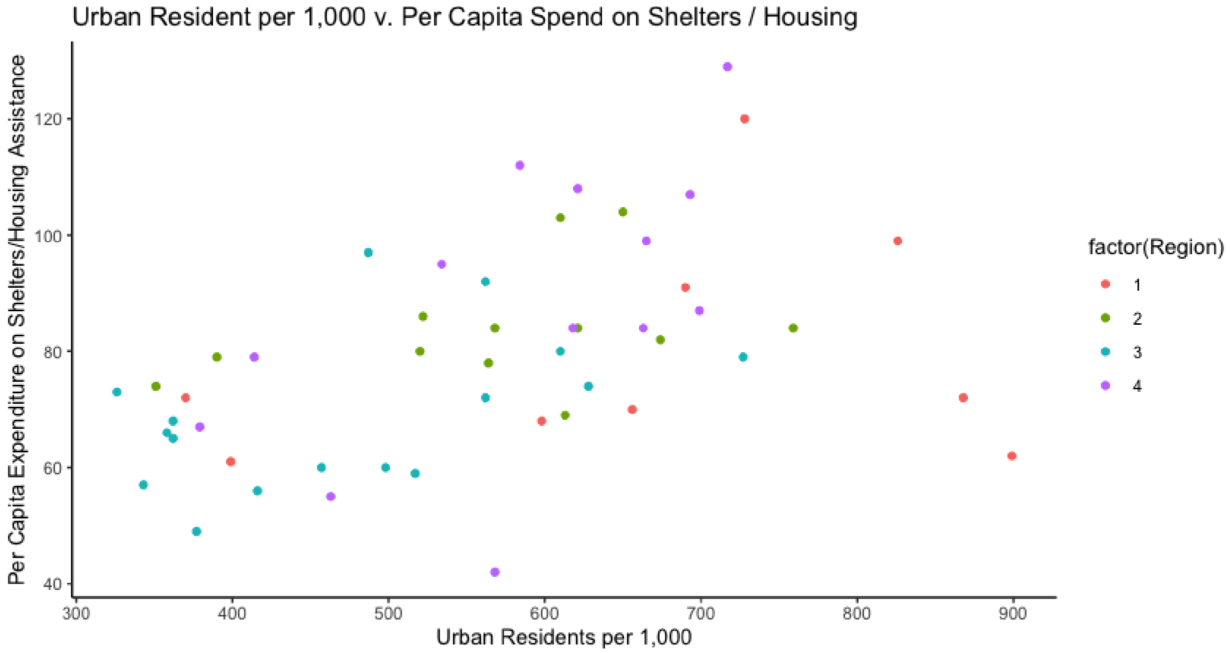
Please plot the relationships among Y, X1, X2, and X3 ? What are the correlations among them (you just need to describe the graph and the relationships among them)?



The above scatterplot shows the relationship between per capita personal income and per capita expenditure on shelters / housing (in each US state). We can, broadly-speaking, see that the higher the per capita income the higher the per capita spend on shelters / housing - there is a positive correlation. The colour is indicating which region the state is - and each region also more-or-less appears to be following this trend



The above scatterplot shows the relationship between Financially Insecure Residents (x-axis) and per capita Spend on Shelters / Housing (y-axis). There looks to be a positive correlation, but it is less clear-cut - and it would be difficult to draw conclusions with more information. We don’t know what’s cause and what’s effect. The West (purple) appears to have the most Financially Insecure Residents and North Central the least.



The above scatterplot plots Urban Residents per 1000 v Per Capita Expenditure on Shelters / Housing. Overall there looks to be a positive correlation across all regions. There appears to be a lot of variance in the urban residents per 1,000 figures - and also quite a few outliers - which would make any exact interpretation of the chart challenging.

	Y	**X1**	**X2**	**X3**
Y	1.00	0.53	0.45	0.46
X1	0.53	1.00	0.21	0.60
X2	0.45	0.21	1.00	0.22
X3	0.46	0.60	0.22	1.00

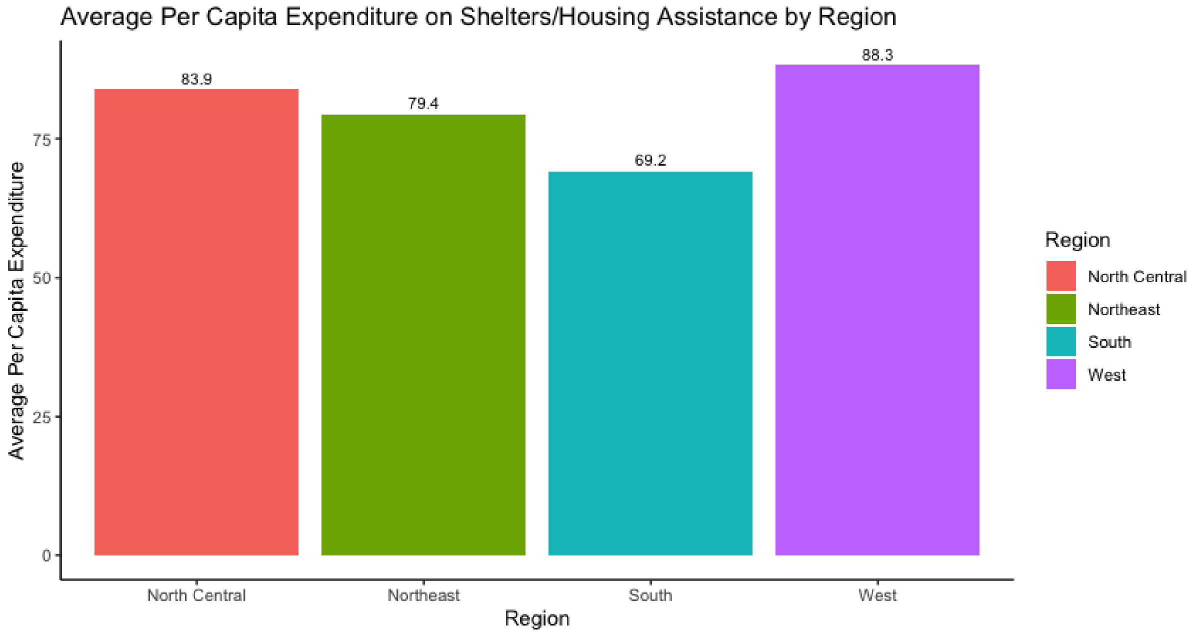
Above is a correlation matrix on our data. This was called using the cor() function in R.

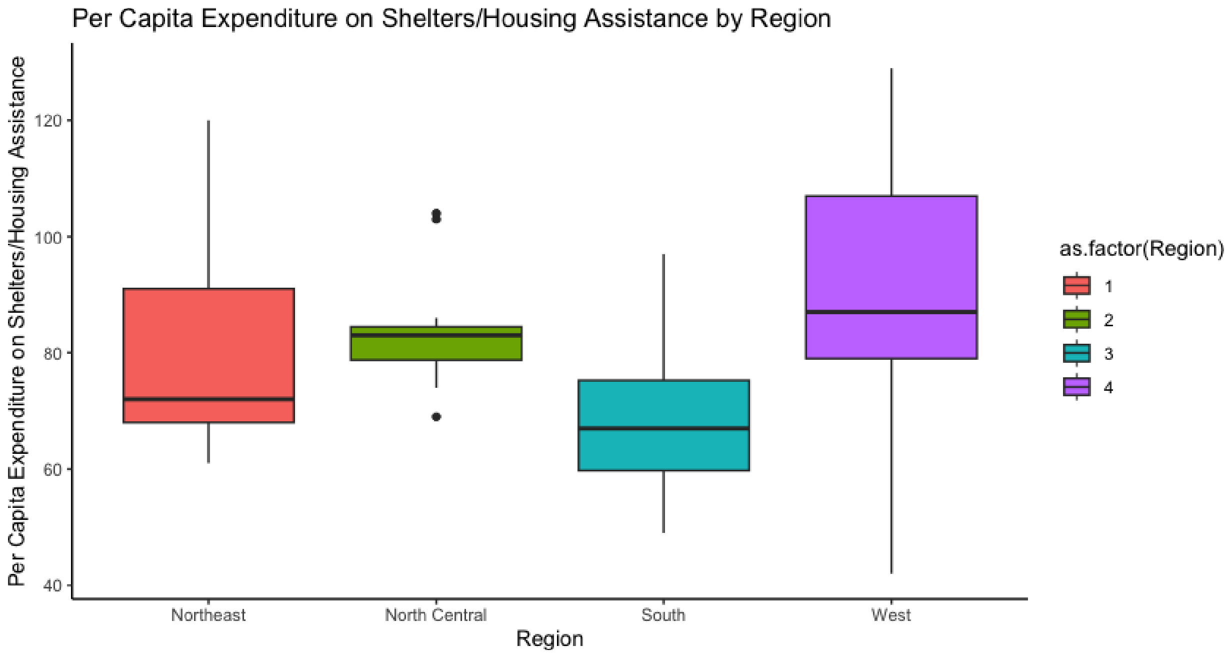
We can see that X1 (per capita personal income) and X3 (no. of people living in urban areas) is the most positively correlated relationship. It appears states with more people living in towns and cities per capita earn more on average

X1 (per capita personal income) and X2 (% of financially insecure) have a correlation of 0.21, which was lower than I expected.

Please plot the relationship between Y and Region? On average, which region has the highest per capita expenditure on housing assistance?

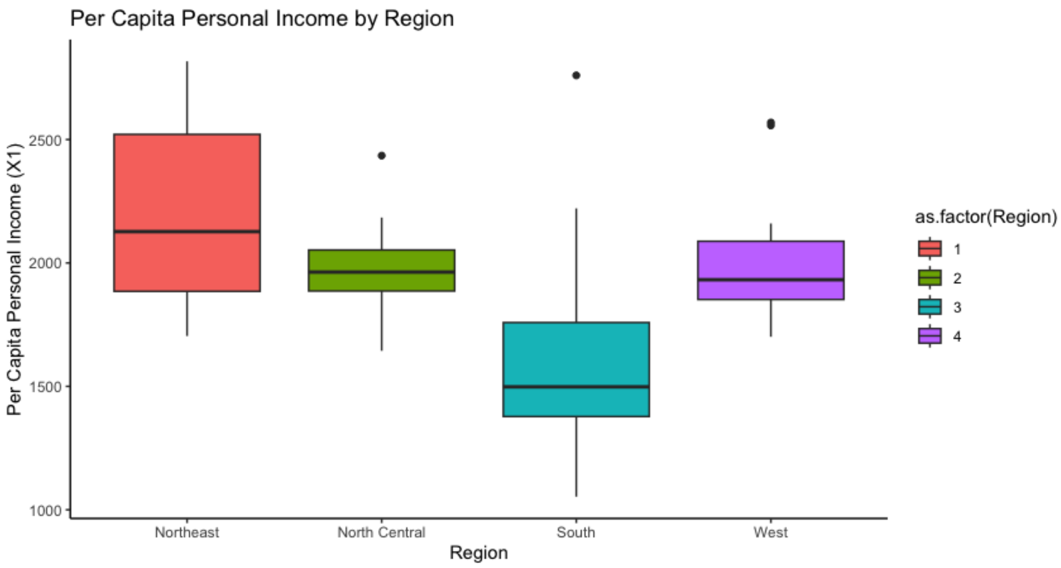
We can see in the below bar graph that overall the West spends most on shelters and housing assistance.



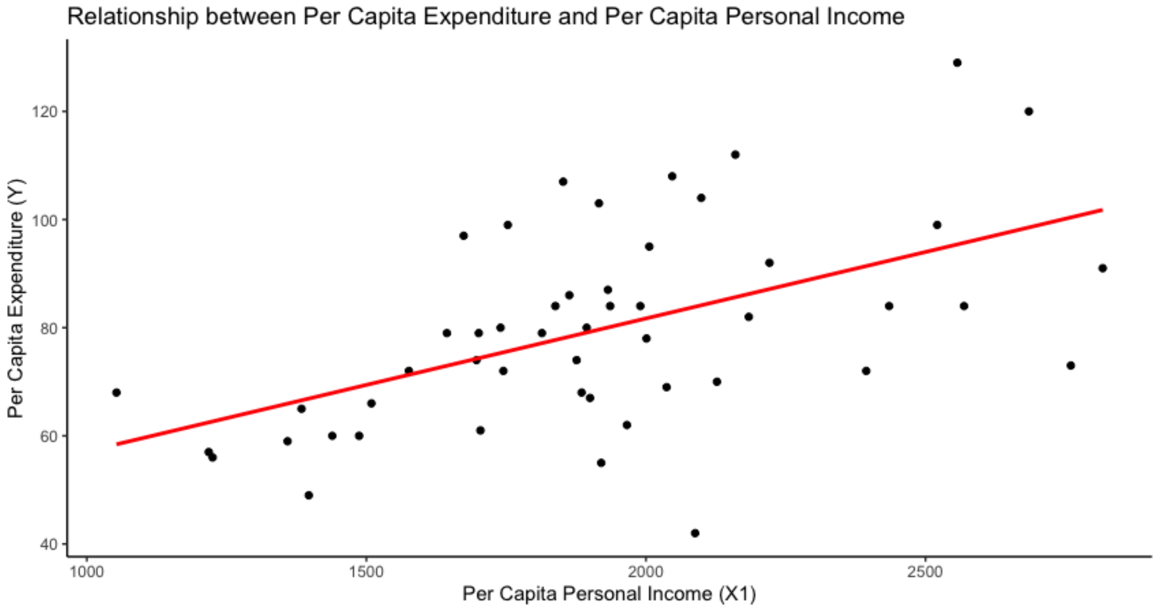


Looking at this in more detail, above is a box plot illustrating per capita spend on shelters and housing assistance by region. It was called using the ggplot() function in R. We can see that overall the West spends most on shelters and housing assistance - although there is a clearly a wide spread at state level. North Central spends the next highest per capita - and has less spread than West

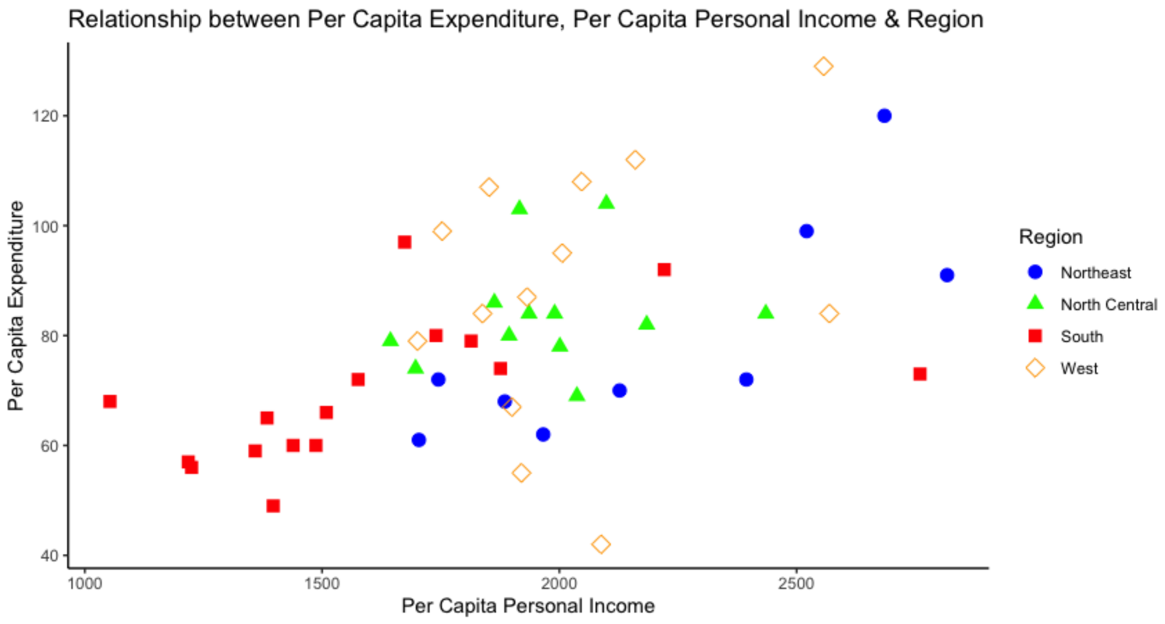
Please plot the relationship between Y and X1 ? Describe this graph and the relationship. Reproduce the above graph including one more variable Region and display different regions with different types of symbols and colors.



According to the data, the Northeast has the highest per capita - c. 2,150. It also has the largest interquartile range. The South looks to have the lowest (c. 1,450), although interestingly it also appears to have the state with both the highest per capita income and the lowest. Spend per capita in the West and North Central are broadly similar at just under 2,000. North Central looks to be the most symmetric of the four regions, with fewer outliers and less variance.



The above scatterplot explores the relationship between per capita personal income and per capita spend on shelters / housing assistance. We can see from the plots and the regression line that there is a positive correlation between them.



The above scatterplot uses symbols and colours to add regionality to the same data. We can see how states in the southern region (red) in general are quite clustered together vs. say the West (blue).

Appendix - Code

```
####################
# Robert Baker - 23371822 - Problem Set 1
# 29 / 09 / 23

####################
# load libraries
# set wd
# clear global .envir
####################

# remove objects
rm(list=ls())
# detach all libraries
detachAllPackages <- function() {
  basic.packages <- c("package:stats", "package:graphics", "package:grDevices", "package:utils", "package:datasets", "package:methods", "package:base")
  package.list <- search()[ifelse(unlist(gregexpr("package:", search()))==1, TRUE, FALSE)]
  package.list <- setdiff(package.list, basic.packages)
  if (length(package.list)>0)  for (package in package.list) detach(package,  character.only=TRUE)
}
detachAllPackages()

# load libraries
```

```
pkgTest <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[,  "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg,  dependencies = TRUE)
  sapply(pkg,  require,  character.only = TRUE)
}

# here is where you load any necessary packages
# ex: stringr
# lapply(c("stringr"),  pkgTest)

library(ggplot2)
lapply(c(),  pkgTest)

library(dplyr)


#####################
# Problem 1
#####################

y <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113, 112, 98, 80, 97, 95, 111, 114, 89, 95, 126, 98)

# checking data
```

```
summary(y)


class(y)



# Finding a 90% confidence interval for the average student IQ


# Getting the mean


mean_y <- mean(y)

mean_y


# Getting the standard dev.

sd_y <- sd(y)

sd_y


# Getting the Sqr. Root

sqrt_y <- sqrt(length(y))

sqrt_y


# Now the standard error (standard dev / square root)

standard_error_y <- sd(y)/sqrt_y

standard_error_y


# Calculating the margin of error for a 90% confidence interval using the
# standard (z) normal distribution (two-tails)

lower_90   <- qnorm(0.05 / 2,
                    mean = mean_y,
                    sd = sd_y)
upper_90  <- qnorm(0.95 / 2,
                   mean = mean_y,
                   sd = sd_y)

# print
message <- paste("Based on the sample data provided in the vector (y),",
                 "we can estimate with c. 90% confidence",
                 "that the mean is likely to fall within the range of approximately",
                 round(lower_90), "to", round(upper_90))
cat(message)


# Null hypothesis: average student IQ is not higher than nat. av. (μ <= 100)
# Alternative hypothesis: average student IQ is higher than nat av. (μ > 100)

# Calculating the z-statistic for the hypothesis test
z_stat <- (mean_y - 100) / standard_error_y

# Calculating the p-value for the hypothesis test (it's a right-tailed test)
p_value <- 1 - pnorm(z_stat)
```

```
# Displaying
z_stat
p_value

message2 <- paste("The z-statistic of approx.", round(z_stat, 2),
              "suggests that the sample mean IQ is about",
              round(abs(z_stat), 2), "standard errors below",
              "the hypothesized population mean of 100.\n",
              "The p-value of approx.", round(p_value, 2),
              "is relatively high.",
              "This means under the null hypothesis, there is a",
              round(p_value * 100, 2), "% chance",
              "of observing a sample mean IQ as low as this, or lower,",
              "just from random sampling variability.")

cat(message2)
```

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

# Problem 2

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


expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/expenditure.txt", header=T)



# Loading packages

# Adopted from: https://stackoverflow.com/questions/4090169/elegant-way-to-check-for-missing-packages-and-install-them

# with thanks to Hannah Frank for sharing this!


if(!require(tidyverse)){

  install.packages("tidyverse")

  library(tidyverse)

}



# Exploring the data


head(expenditure)
```

```
summary(expenditure)

View(expenditure)

str(expenditure)

levels(expenditure)

# Plotting the relationships among Y, X1, X2, and X3
# What are the correlations among them?

# Scatterplot 1 - Y v. X1
plot1 <- ggplot(expenditure, aes(x = X1, y = Y, color = factor(Region))) +
  geom_point() +
  labs(title = "Per Capita Income v. Expenditure on Shelters/Housing", x = "Per Capita Personal Income", y = "Per Capita Expenditure on Shelters/Housing") +
  theme_classic()
plot1

# Scatterplot 2 - Y v. X2
plot2 <- ggplot(expenditure, aes(x = X2, y = Y, color = factor(Region))) +
  geom_point() +
  labs(title = "Per Capita Spend on Shelters v. No. of Finacially Insecure Residents", x = "Financially Insecure Residents per 100,000", y = "Per Capita Expenditure on Shelters/Housing Assistance") +
  theme_classic()
plot2
```

```
# Scatterplot 3 - Y v. X3
plot3 <- ggplot(expenditure, aes(x = X3, y = Y, color = factor(Region))) +
  geom_point() +
  labs(title = "Urban Resident per 1,000 v. Per Capita Spend on Shelters / Housing", x = "Urban Residents per 1,000", y = "Per Capita Expenditure on Shelters/Housing Assistance") +
  theme_classic()
plot3


# Correlation matrix
cor_matrix <- round(cor(expenditure[, c("Y", "X1", "X2", "X3")]), 2)
cor_matrix


# Creating box plot for Per Capita Expenditure on Shelters/Housing Assistance by Region

# Defining region labels
region_labels <- c("Northeast", "North Central", "South", "West")

ggplot(expenditure, aes(x = as.factor(Region), y = Y, fill = as.factor(Region))) +
  geom_boxplot() +
  labs(title = "Per Capita Expenditure on Shelters/Housing Assistance by Region",
       x = "Region",
       y = "Per Capita Expenditure on Shelters/Housing Assistance") +
  scale_x_discrete(labels = region_labels) +
  theme_classic()




# Creating Box plot for Financially Insecure Residents per 100,000 by Region
ggplot(expenditure, aes(x = as.factor(Region), y = X2, fill = as.factor(Region))) +
  geom_boxplot() +
  labs(title = "Financially Insecure Residents per 100,000 by Region",
       x = "Region",
       y = "Financially Insecure Residents per 100,000 (X2)") +
  scale_x_discrete(labels = region_labels) +
  theme_classic()

# Box plot for No. of People per Thousand Residing in Urban Areas by Region
ggplot(expenditure, aes(x = as.factor(Region), y = X3, fill = as.factor(Region))) +
  geom_boxplot() +
  labs(title = "No. of People per Thousand Residing in Urban Areas by Region",
       x = "Region",
       y = "Number of People per Thousand Residing in Urban Areas (X3)") +
  scale_x_discrete(labels = region_labels) +
  theme_classic()

# Explopring relationship between Y (per capita expenditure on shelters/housing assistance per state) & Region.
# On average, which region has the highest per capita expenditure on housing assistance?


# Calculating the mean expenditure by region
mean_expenditure_by_region <- expenditure %>%
  mutate(Region_Name = case_when(
    Region == 1 ~ "Northeast",
    Region == 2 ~ "North Central",
    Region == 3 ~ "South",
    Region == 4 ~ "West"
  )) %>%
  group_by(Region_Name) %>%
  summarize(Mean_Expenditure = mean(Y))

# Renaming columns
colnames(mean_expenditure_by_region) <- c("Region", "Per Capita Mean Expenditure")

# Printing
print(mean_expenditure_by_region)


# Creating a bar plot for mean expenditure by region
ggplot(mean_expenditure_by_region, aes(x = Region, y = `Per Capita Mean Expenditure`, fill = Region)) +
  geom_bar(stat = "identity") +

  geom_text(aes(label = round(`Per Capita Mean Expenditure`, 1)), vjust = -0.5, size = 3) +

  labs(title = "Average Per Capita Expenditure on Shelters/Housing Assistance by Region",

       x = "Region",

       y = "Average Per Capita Expenditure") +

  scale_x_discrete(labels = mean_expenditure_by_region$Region) +

  theme_classic()
```

```
# Creating a scatterplot of Y v. X1

ggplot(expenditure, aes(x = X1, y = Y)) +

  geom_point() +

  geom_smooth(method = "lm", se = FALSE, color = "red") +

  labs(title = "Relationship between Per Capita Expenditure and Per Capita Personal Income",

       x = "Per Capita Personal Income (X1)",

       y = "Per Capita Expenditure (Y)") +

  theme_classic()
```

```
# Creating a scatterplot of Y vs. X1 - with different symbols and colors for regions

ggplot(expenditure, aes(x = X1, y = Y, color = factor(Region), shape = factor(Region))) +

  geom_point(size = 3) +
  labs(title = "Relationship between Per Capita Expenditure, Per Capita Personal Income & Region",

       x = "Per Capita Personal Income",

       y = "Per Capita Expenditure",

       color = "Region",

       shape = "Region") +

  scale_color_manual(values = c("blue", "green", "red", "orange"),

                     labels = c("Northeast", "North Central", "South", "West")) +

  scale_shape_manual(values = c(19, 17, 15, 5),

                     labels = c("Northeast", "North Central", "South", "West")) +

  theme_classic()
```