Problem Set 1

Imelda Finn, 22334657

Due: October 3, 2021

Question 1 (50 points): Education

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

```
iqData <- 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.*

```
# calculate sample statistics
    # capture the number of observations
    n <- length(iqData)

    # calculate mean
    iqSum <- sum(iqData)            # sum of IQ scores
    iqMean <-iqSum / n              # mean IQ score for sample

    # calculate variance and standard deviation
    iqVar <- sum((iqData - iqMean)^2)/(n-1)
    iqSD <-sqrt(iqVar)

    iqse <- iqSD / sqrt(n)          # standard error of sample
```

The code for the t-test at 90% is:[1]

```
t.val <- qt(alphaVal/2, df = n-1, lower.tail = FALSE)

CI_lower <- iqMean - t.val * iqse
CI_upper <- iqMean + t.val * iqse
```

The result was:

Our Confidence interval for the IQ of the students in the sample is:

93.96 < mean IQ < 102.92

with a confidence level of 90%. [2]

2. *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 $\alpha = 0.05$.

   (a) The number of observations is 25, which isn't ideal for t-test statistics (we would prefer at least 30 observations). The data isn't really random, as all are students from the same school.

   (b) H_0: the average IQ score in the sample is less than the population average ie $\mu_O \leq \mu$

   (c) $H\alpha$: the average IQ score in the sample is greater than the population average ie $\mu_O > \mu$

   (d) Calculate test statistic
   $$TS = \frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}}$$

   (e) Calculate p-value
   $$p = Pr(Z \leq -|\frac{\bar{Y} - \mu_O}{\sigma_{\bar{Y}}}|)$$

   (f) if p $\leq \alpha = 0.05$, we reject the null hypothesis

```
# Test our hypothesis
   alphaVal <- 0.05
```

[1] t.val = 1.71

[2] A Z-test gave a 90% CI of $94.13 < \mu < 102.75$.

```
popMean <- 100
#get test statistic
testStatistic <- (iqMean - popMean) / iqse

pValue <- pnorm(-abs(testStatistic))

# calculate t-test p-value
t_pValue <- pt(abs(testStatistic), df = n-1, lower.tail = FALSE)
```

Results

p-value for normal distribution is 0.276

t	df	p-value
-0.5957439	24	0.2784617

The p-value is greater than 5%, so we cannot reject the null hypothesis. The data does not support the suggestion that the school IQ scores are greater than the population average.

Question 2 (50 points): Political Economy

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*

```
#expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2022/main/datasets/expenditure.txt", header=T)
```

- In Figure 1, there appears to be some positive correlation between the per capita expenditure on Shelters/Housing Assistance (SHA) and each of the three numerical variables. In each case, an increase in the response variable accompanies an increase in the candidate explanatory variable.

  The weakest relationship is between SHA expenditure and the number of residents who are financially insecure. When looked at in more detail, Figure 2 shows that there is a negative correlation between the two variables (ie SHA expenditure per capita decreases as the number of financially insecure residents increases) until $X2$ reaches approximately 300 per 100,000; after that point, the values increases are positively correlated. (line fitted using: `‘geom_smooth()‘ using method = 'loess' and formula 'y ~ x'`)

  The income per capita shows some correlation with expenditure and with the number of urban residents, but not with financially insecure residents. Similarly, the plot of urban residents against financially insecure residents appears to be uniformly distributed, (i.e. uncorrelated), while expenditure and income both increase with the increase in urban population. Figure 3 shows that the slope of the line relating expenditure on SHA to the urban population turns negative after $X3$ reaches approximately 750 per 1,000, however there are very few data points in this region of the plot.

-

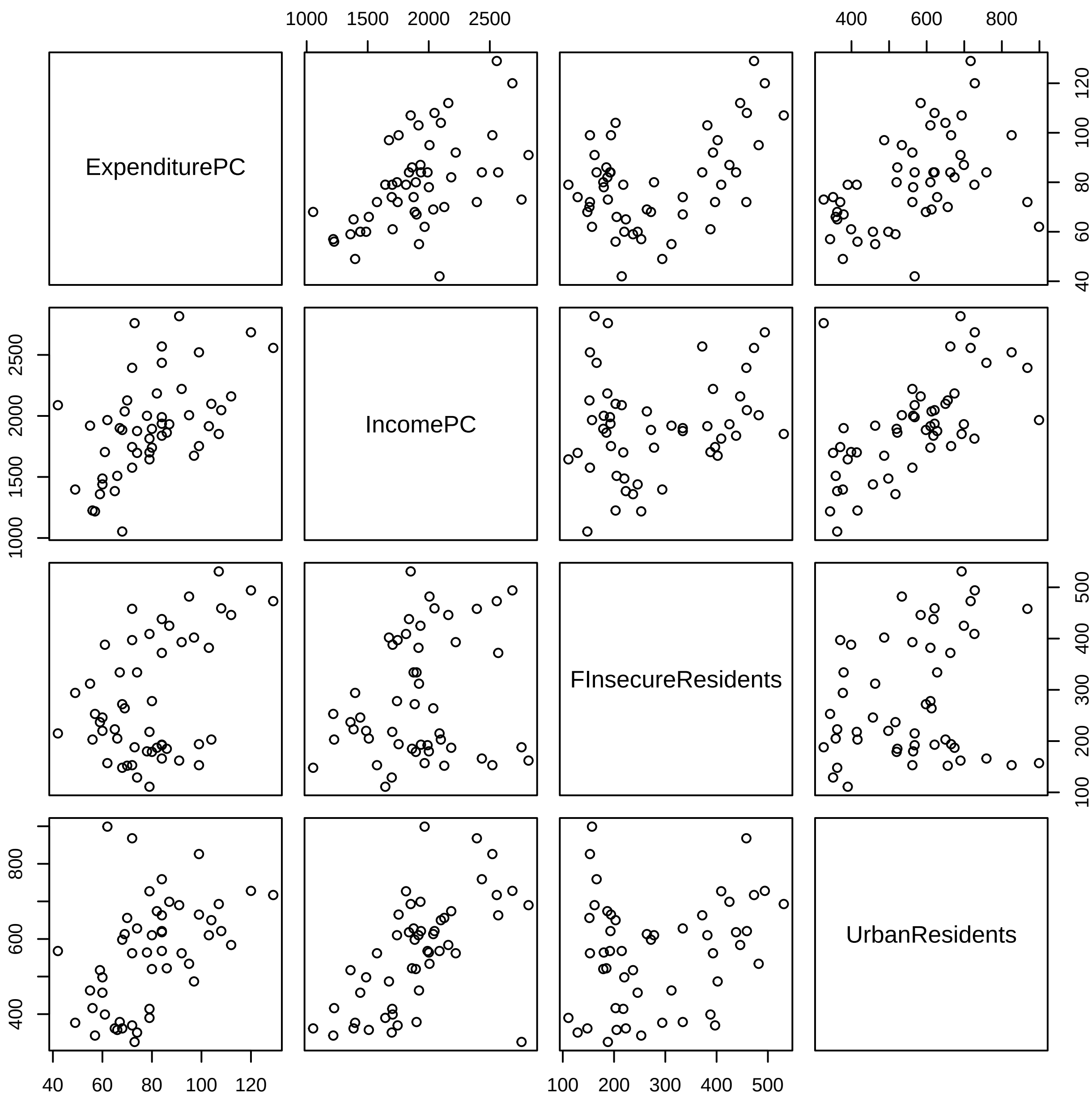

Figure 1: Pair-wise comparison of expenditure variables

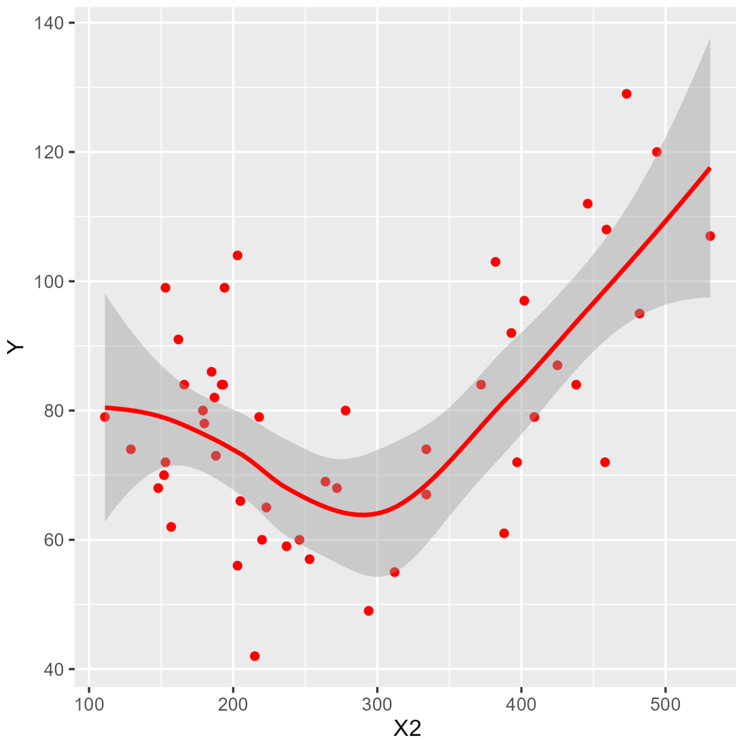

Figure 2: Relationship between expenditure on SHA and financially insecure population

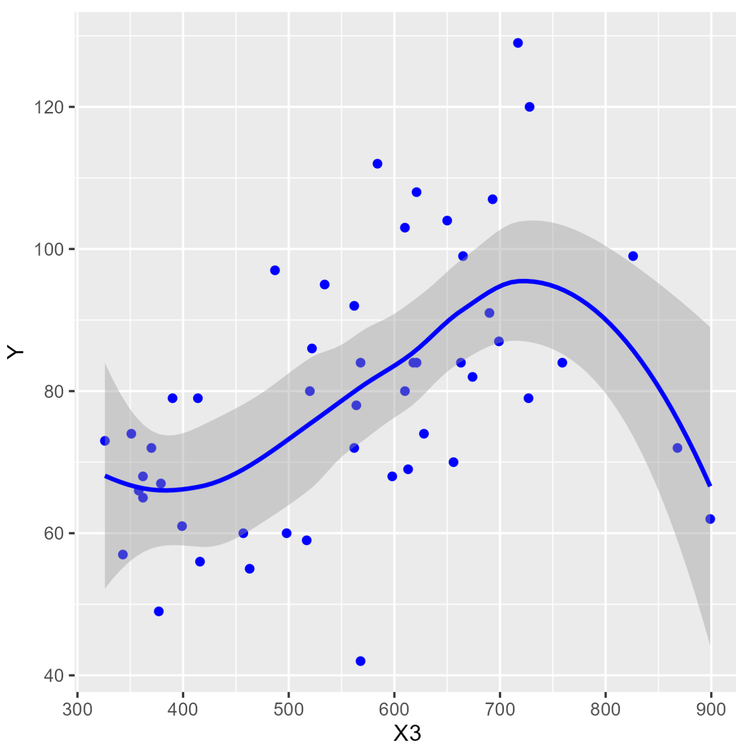

Figure 3: Relationship between expenditure on SHA and urban population

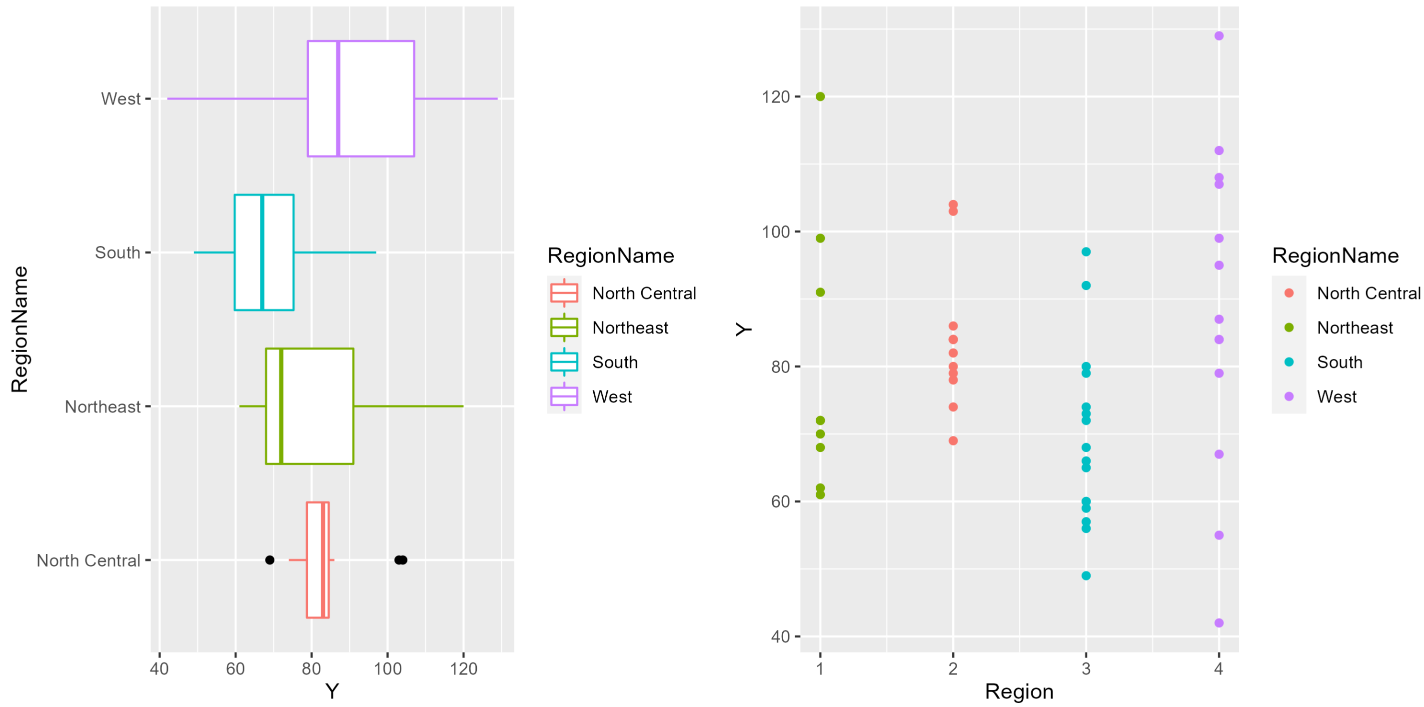

Figure 4: SHA expenditure by geographical region

The relationship between expenditure on SHA and region is shown in Figure 4

```
regional_mean_table <-expenditure %>% # Tidyverse method for grouping
      group_by(Region) %>%
      summarise(mean = round(mean(Y), 2))
```

Table 1: Regional spending on SHA

	Region	mean	regions
1	1	79.44	Northeast
2	2	83.92	North Central
3	3	69.19	South
4	4	88.31	West

The West region has the highest average per capita spending ($88.31) on Shelters/Housing Assistance (Table 1).

- The graph of income per capita (X1, $) against expenditure on SHA per capita (Y, $) appears to show a positive correlation between the two variables, i.e. it suggests that higher pc income could be a predictor for higher spending on housing assistance (Figure 5).

  When the data is subdivided by region the relationship between the variables is much less convincing. The individual regions show clear differences between the interaction of the two values - North Central and South show no positive correlation; in the Northeast and West it is possible that some kind of relationship exists, but it's not consistent (Figure 6).

  The apparent relationship when the data are combined clearly masks underlying differences between the regions.

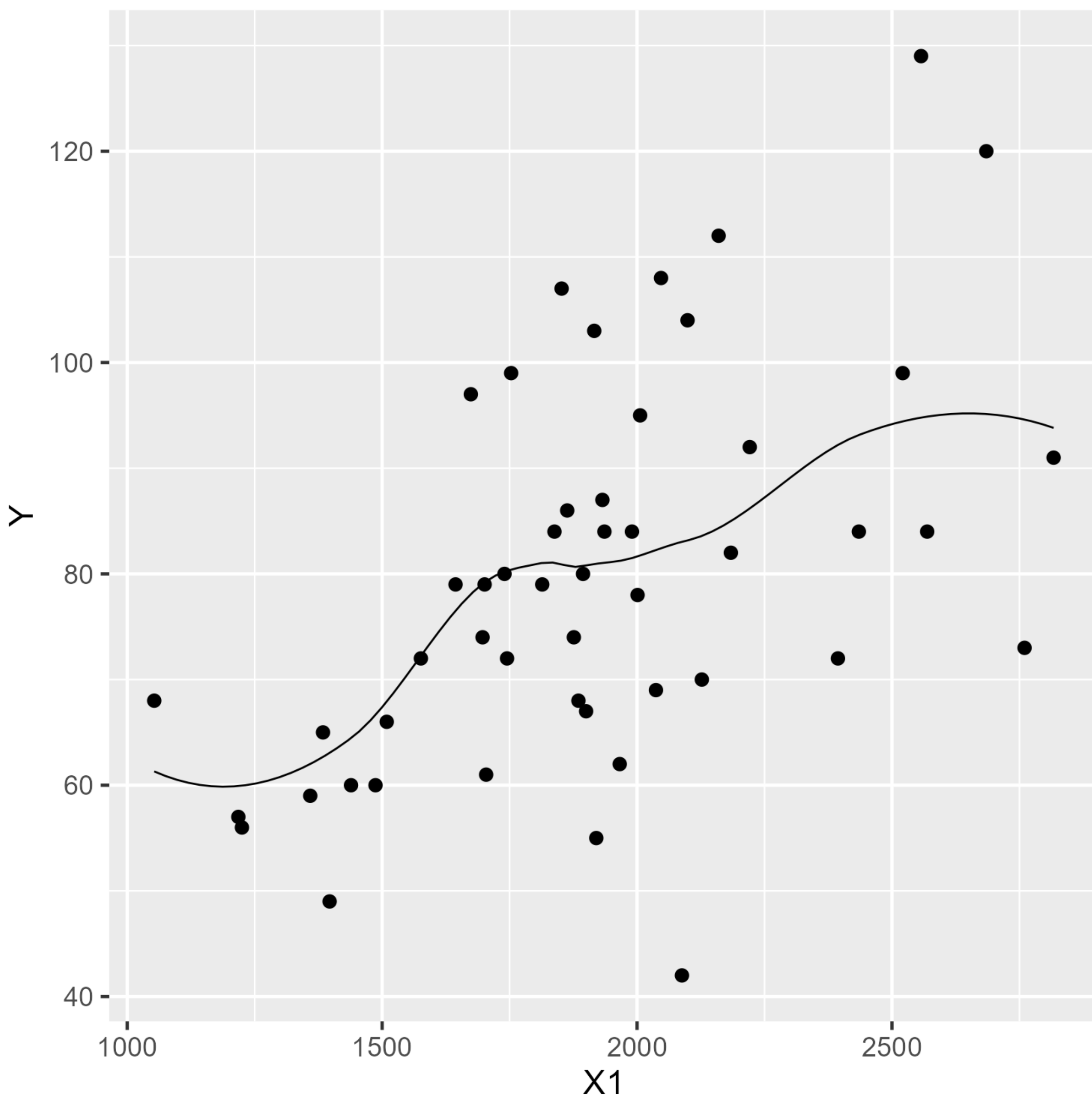

Figure 5: Income vs SHA expenditure, per capita

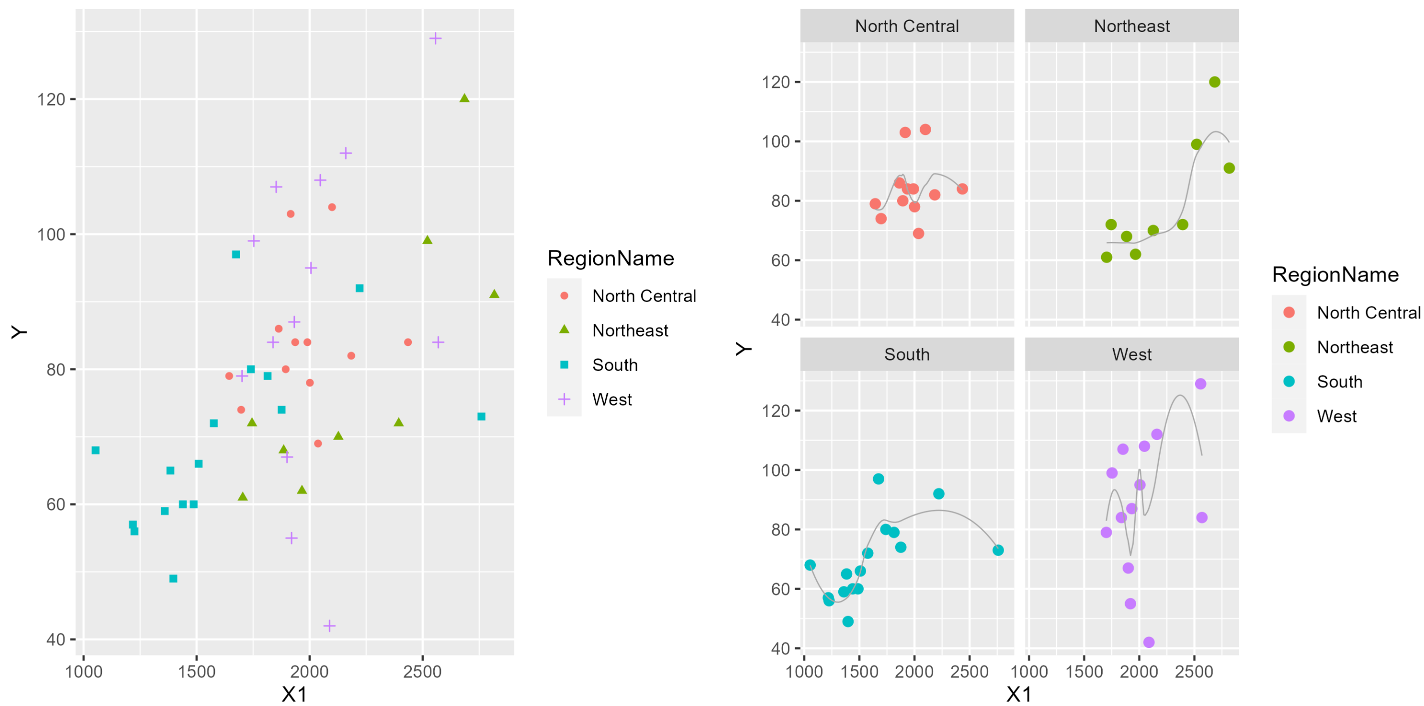

Figure 6: Income per capita vs SHA expenditure, by Region

Appendix - R code

```r
##################################################################
# Imelda Finn, 22334657
# POP77003 - Stats I
# clear global .envir, load libraries, set wd
##################################################################

# 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)
}

# load necessary packages
lapply(c("ggplot2", "stargazer", "tidyverse", "stringr", "quantreg"),  pkgTest
  )

# set working directory to current parent folder
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

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

# load data as vector  - in .tex file - update if move from 38
iqData <- 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)

## Save our data to a .csv file in the data directory
write.csv(iqData, file = "Data/iq.csv", row.names = FALSE)

# Explore data
summary(iqData)
str(iqData)
```

```r
head(iqData)

# look at sampling from sample
meanIQ <- vector("double", length = 1000)
for (i in 1:1000) {
  meanIQ[i] <- mean(sample(iqData, 25, replace=TRUE))
}
summary(meanIQ)
boxplot(meanIQ, iqData, xlab=c("averaged vs original sample"))


# Visually inspect the data
hist(iqData, breaks = 10, main = "Histogram of IQ", xlab = "IQ")

plot(density(iqData), main = "PDF of IQ", xlab = "IQ")

# Use a QQ plot to determine if our IQ variable is normally distributed
qqnorm(iqData)
qqline(iqData, distribution = qnorm)
# Sample values fall away from normal line at upper end

##----------------------------------------------------------------------------
# calculate sample statistics
# capture the number of observations
n <- length(iqData)

# calculate mean
iqSum <- sum(iqData)            # sum of IQ scores
iqMean <-iqSum / n              # mean IQ score for sample

# calculate variance and standard deviation
iqVar <- sum((iqData - iqMean)^2)/(n-1)
iqSD <-sqrt(iqVar)

iqse <- iqSD / sqrt(n)          # standard error of sample

##----------------------------------------------------------------------------
## Confidence Intervals
# Calculate 90 percent confidence intervals using normal distribution
# assuming  iqMean ~ N(mu, iqse)
alphaVal = 0.1
CI_lower <- qnorm(alphaVal/2, mean = iqMean, sd =  iqse)

CI_upper <- qnorm(1-alphaVal/2, mean = iqMean, sd = iqse)

# output
cat(str_glue("{(1-alphaVal)*100}% Confidence Intervals, two-sided z-test"))
matrix(c(CI_lower, CI_upper), ncol = 2,
       dimnames = list("",c("Lower", "Upper")))

# Calculate 90 percent confidence intervals using t-distribution
```

```r
# degrees of freedom = n-1 = 24 - should be >30
t.val <- qt(alphaVal/2, df = n-1, lower.tail = FALSE)

CI_lower <- iqMean - (t.val * iqse)
CI_upper <- iqMean + (t.val * iqse)

# calculate using t-test
cat(str_glue("{(1-alphaVal)*100}% Confidence Intervals, two-sided t-test"))
matrix(c(CI_lower, CI_upper), ncol = 2,
       dimnames = list("",c("Lower", "Upper")))

# t-test results in (slightly) wider confidence interval

# Check our working
#t.test(iqData, conf.level = 1-alphaVal, alternative = "two.sided")

cat("Our Confidence interval for the IQ of the students in the sample is: ")
cat(str_glue("  {round(CI_lower,2)} < mean IQ < {round(CI_upper,2)} "))
cat(str_glue("with a confidence level of {(1-alphaVal)*100}%"))


##-----------------------------------------------------------------------------
## Hypothesis Testing
# Wrangling our data
class(iqData) # What class of vector is our IQ variable? - numeric

# Hypothesis test:
# H0 : average IQ of students in school is less than or equal to  national average
# Ha : average IQ of students in school is greater than national average

# alpha = 0.05, 1-tail test, single population

# don't have variance of population, only have mean to compare against

# Test our hypothesis
alphaVal <- 0.05
popMean <- 100
#get test statistic
testStatistic <- (iqMean - popMean) / iqse

pValue <- pnorm(-abs(testStatistic))

# calculate t-test p-value
t_pValue <- pt(abs(testStatistic), df = n-1, lower.tail = FALSE)

matrix(c(testStatistic, n-1, t_pValue  ), ncol = 3,
       dimnames = list("",c("t", "df", "p-value")))
cat(str_glue("p-value for normal distribution is {round(pValue,3)}"))

# check result
```

```
t.test( iqData  ,
       mu = 100, # population mean
       var.equal = TRUE, # The default is FALSE - don't have var for popn
       alternative = "less", # H0: sample mean > population mean
       conf.level = .95) #


# How do we interpret the output?
# for confidence level of 95%, we cannot reject the hypothesis (p-value >
   alpha)


#####################
# Problem 2
#####################
rm(list=ls())
# set working directory to current parent folder
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

# function to save output to a file that you can read in later to your docs
output_stargazer <- function(outputFile, appendVal=TRUE, ...) {
  output <- capture.output(stargazer(...))
  cat(paste(output, collapse = "\n"), "\n", file=outputFile, append=appendVal)
}

# read in expenditure data
#expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_
   Fall2022/main/datasets/expenditure.txt", header=T)
#write.table(expenditure, "Data/expenditure.txt")
expenditure <- read.table("Data/expenditure.txt", header=T)

# 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

data_headers <- c("State", "ExpenditurePC", "IncomePC", "FInsecureResidents",
              "UrbanResidents", "Region")
regions <- c("Northeast","North Central", "South", "West")
names(expenditure)
expenditure$RegionName<-regions[expenditure$Region]

# Inspect the data
head(expenditure)
str(expenditure)
summary(expenditure)

#investigate spending on Housing assistance
# Visualise
```

```r

onefile <- FALSE
#pdf( file = if(onefile) "expenditure_plots.pdf" else "expenditure_plots%03d.pdf")

hist(expenditure$Y, main = "Histogram of spending on HA ", xlab = "$, per capita")

plot(density(expenditure$Y),
     main = "PDF of spending on HA ", xlab = "$, per capita")

qqnorm(expenditure$Y)
qqline(expenditure$Y, distribution = qnorm)

#dev.off() # close pdf file
#---------------------------------------------------------------------------
# plot the numerical variables against each other
pdf("expenditure_pairs.pdf" )
pairs(~Y + X1 + X2 + X3, expenditure, labels = data_headers[2:5])
dev.off() # close output


# look at detail of some relationships
ggplot(expenditure) +
  geom_point(aes(X2, Y), colour = "red" ) +
  geom_smooth(aes(X2, Y), colour = "red")
ggsave("y_x2.png", width = 5, height = 5)


ggplot(expenditure) +
  geom_point(aes( X3, Y, ), colour = "blue" ) +
  geom_smooth(aes( X3, Y), colour = "blue")
ggsave("y_x3.png", width = 5, height = 5)

#-------------------------------------------------------------
# regional expenditure on housing assistance
# look at plots
# by state
x <- seq(1, length(expenditure$Y))
ggplot(expenditure) +
  geom_point(aes( x, Y , colour = RegionName))

# grouped by region
ggplot(expenditure) +
  geom_point(aes(Region, Y,colour = RegionName))
ggsave("region_y.png", width = 5, height = 5)
# more spread in r4, least in r2

ggplot(expenditure) +
  geom_boxplot(aes(Y, RegionName, colour=RegionName), outlier.colour = "black")
```

```r
ggsave("region_boxplot.png", width = 5, height = 5)


# calculate regional means
regional_mean_table <-expenditure %>% # Tidyverse method for grouping
  group_by(Region) %>%
  summarise(mean = round(mean(Y), 2))

regional_mean_table <- cbind(regional_mean_table, regions)

ggplot(regional_mean_table) +
  geom_point(aes(regions, mean, colour = regions, shape = regions), size=3)
ggsave("region_means.png", width = 5, height = 5)

# West region has highest per capita mean expenditure on housing assistance

matrix(regional_mean_table$mean, ncol = 4,
       dimnames = list("", c(regional_mean_table$regions)))
cat(str_glue("Highest average pc spending on SHA is ${regional_mean_table[4, 2]} in the West region"))


output_stargazer("regional_means.tex", appendVal = FALSE, regional_mean_table,
                 title="Regional spending on SHA", #column.labels=regional_mean_table$regions,
                 label="tab:region_mean", summary=FALSE, digits = 2
                 )   # file fragment


#-----------------------------------------------------------------------
# look at income vs expenditure on HA, by region
#factor(expenditure$Region) <- regions
ggplot(expenditure) +
  geom_point(aes( X1, Y)) +
  geom_smooth(aes(X1, Y), colour = "black", se=FALSE, size=0.3)
ggsave("y_x1.png", width = 5, height = 5)

ggplot(expenditure) +
  geom_point(aes( X1, Y, colour= RegionName, shape= RegionName))
ggsave("y_x1_region.png", width = 5, height = 5)

ggplot(expenditure, ) +
  geom_point(aes(X1, Y, colour=RegionName), size = 2) +
  facet_wrap(~ RegionName, nrow = 2) +
  geom_smooth(aes(X1, Y) , colour = "darkgrey", size = 0.3, se=FALSE)
ggsave("y_x1_region_facet.png", width = 5, height = 5)
```