Assignment2

2025-02-07

```
flint <- read.csv(file = "flint.csv")

#1b
mean(flint$Pb >= 15)
```

```
## [1] 0.04436229
```

```
#1c
mean(flint$Cu[flint$Region == "North"])
```

```
## [1] 44.6424
```

```
#1d
mean(flint$Cu[flint$Pb >= 15])
```

```
## [1] 305.8333
```

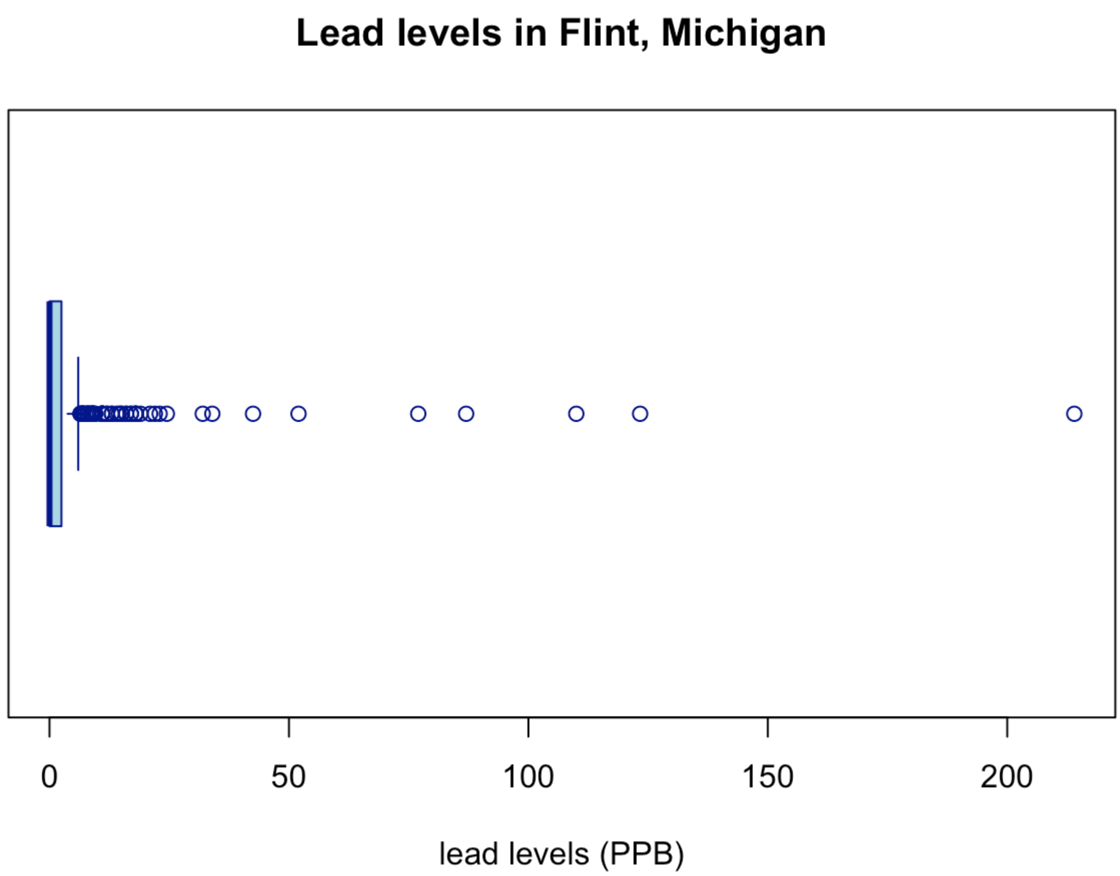
```
#1e
mean(flint$Pb)
```

```
## [1] 3.383272
```

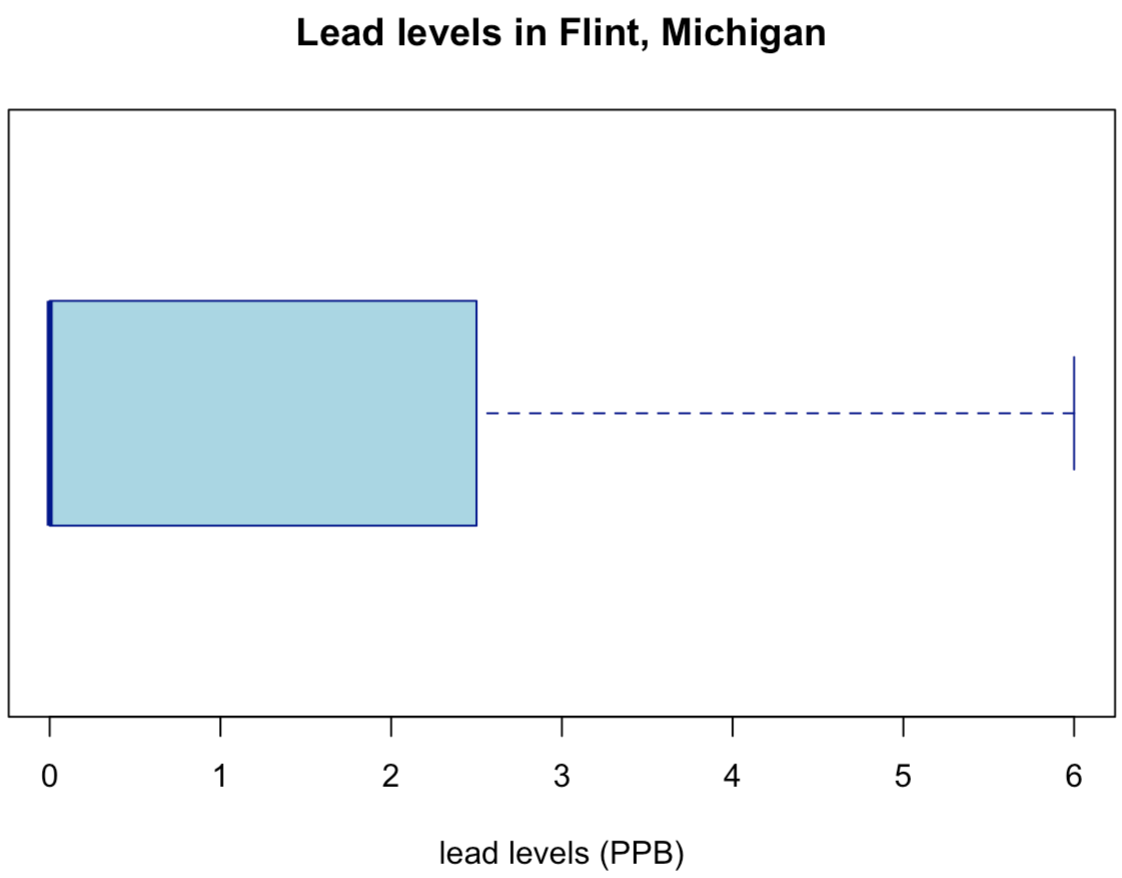
```
mean(flint$Cu)
```

```
## [1] 54.58102
```

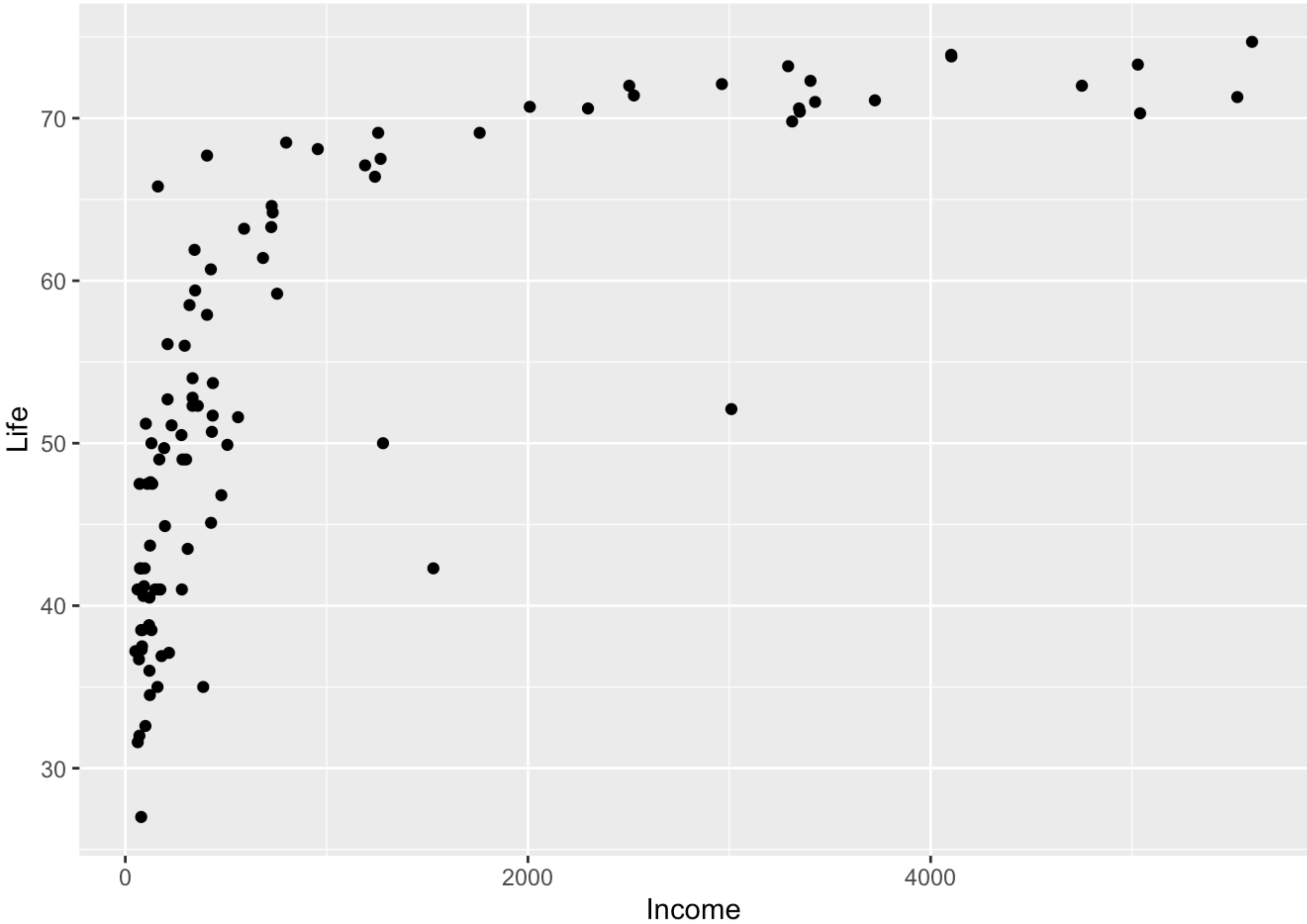
```
#1f
boxplot(flint$Pb,
        main="Lead levels in Flint, Michigan",
        xlab="lead levels (PPB)",
        col="lightblue",
        border="darkblue", outline=TRUE, horizontal=TRUE)
```



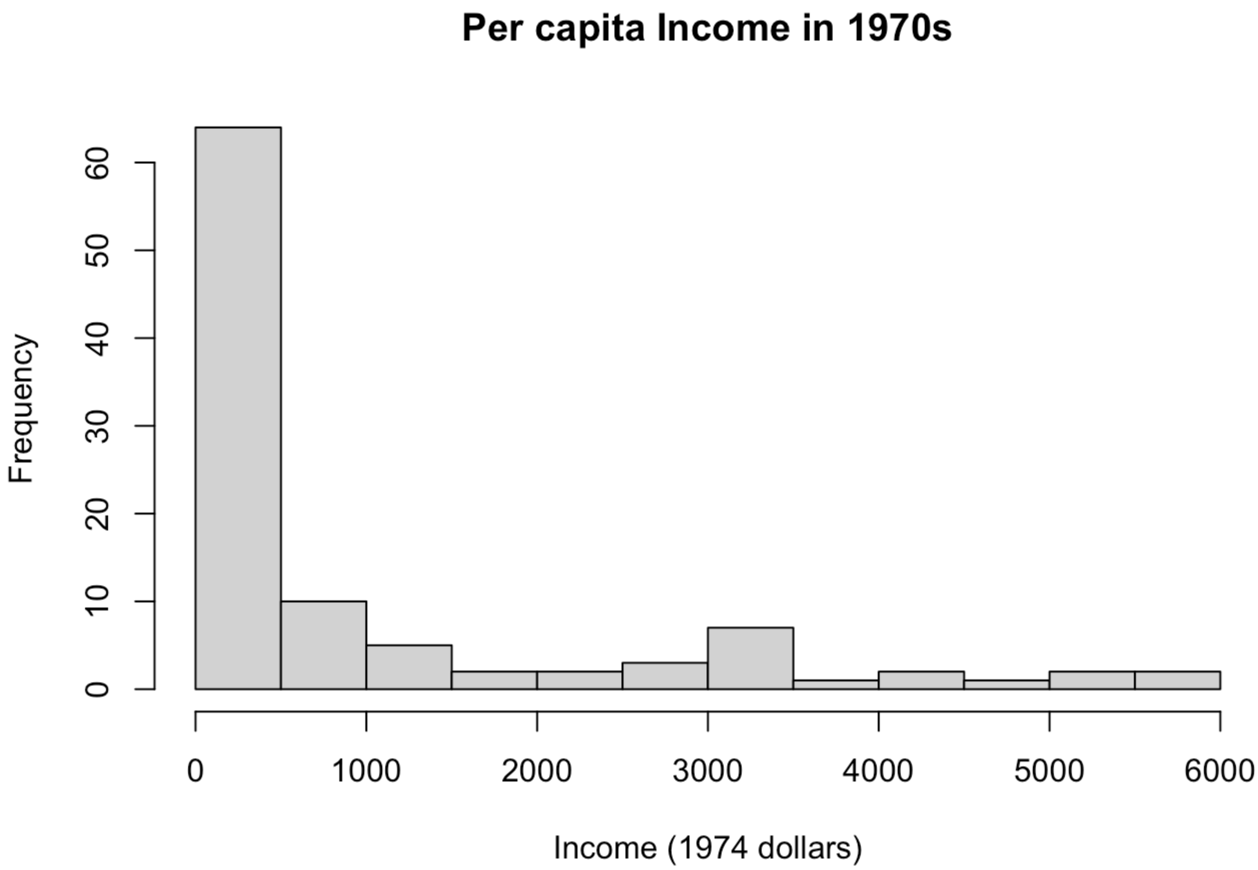

```
#checking the boxplot without the outliers to get a clearer picture
boxplot(flint$Pb,
        main="Lead levels in Flint, Michigan",
        xlab="lead levels (PPB)",
        col="lightblue",
        border="darkblue", outline=FALSE, horizontal=TRUE)
```



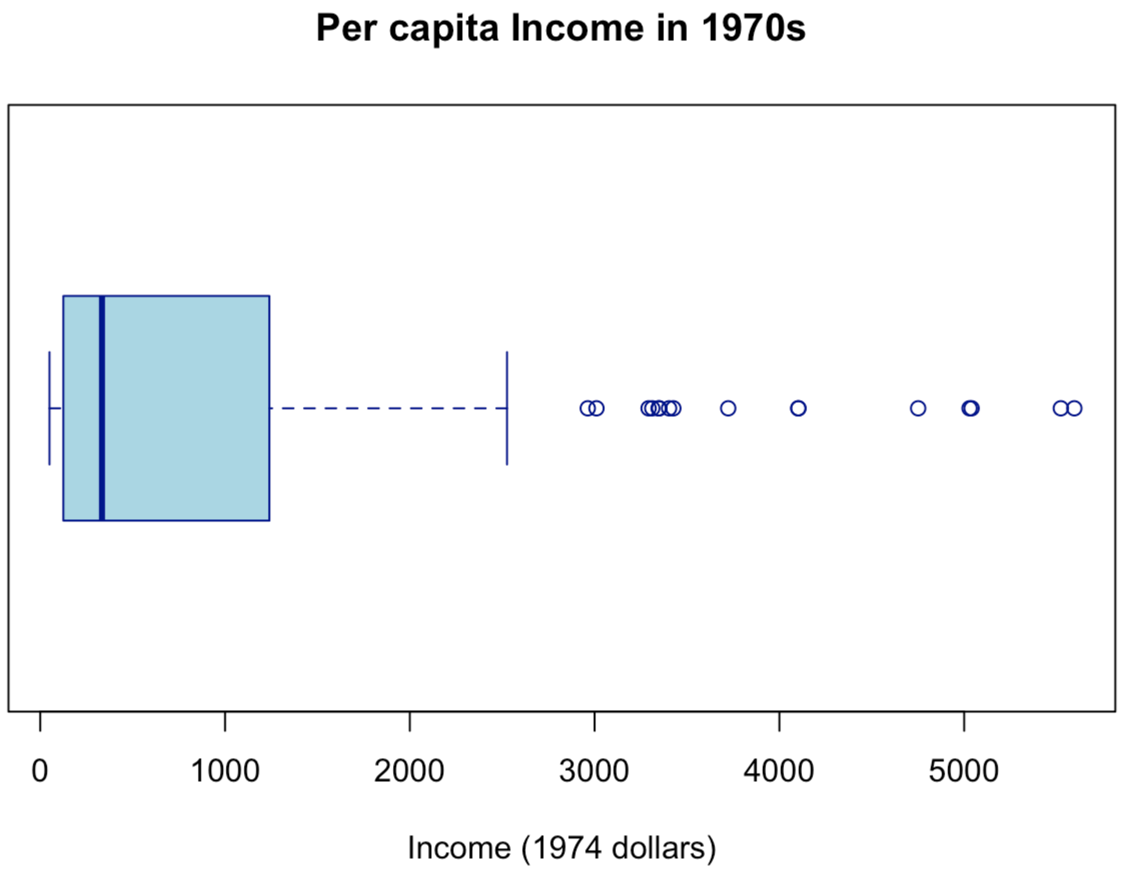

```
#1g
# As we can see in the boxplot the data contains a lot of outliers, i.e. areas
# with a much higher lead level so the mean might not be suitable way to
# measure the data. For skewed data distributions such as this the median tends
# to be a better measure to see the center of the data




#2
life <-read.table(
  "https://ucla.box.com/shared/static/rqk4lc030pabv30wknx2ft9jy848ub9n.txt",
  header = TRUE)
library(ggplot2)

#2a
ggplot(life, aes(x=Income, y=Life)) + geom_point()
```



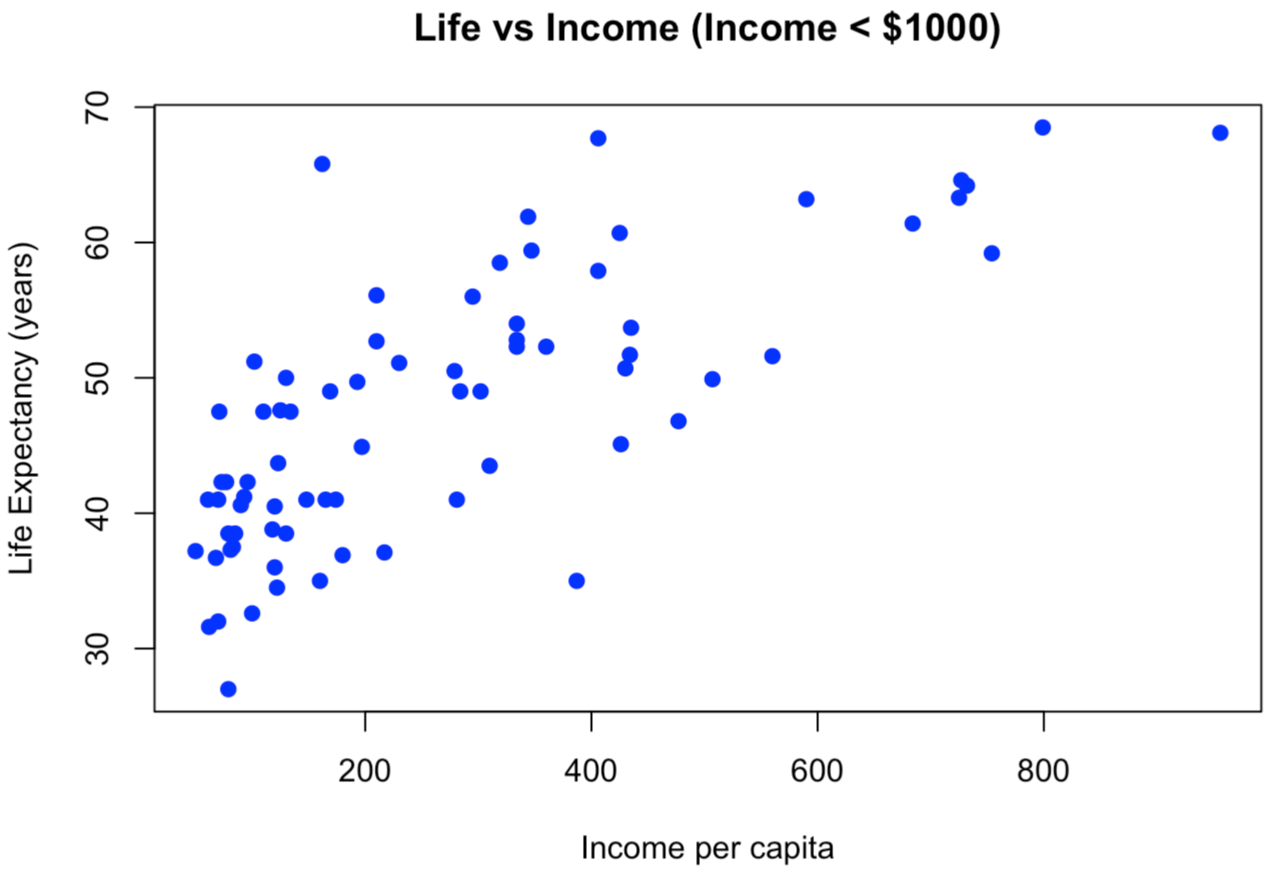

```
# as expected as income increases initially, life expectancy rises very rapidly
# but at a certain point the rate of increase plateaus and as income increases
# life expectancy increases at a much slower rate (this is almost an
# exponential growth)

#2b
hist(life$Income, xlab="Income (1974 dollars)",
     main="Per capita Income in 1970s")
```




```
boxplot(life$Income,
        main="Per capita Income in 1970s",
        xlab="Income (1974 dollars)",
        col="lightblue",
        border="darkblue", outline=TRUE, horizontal=TRUE)
```




```
# as we can see the income levels are highly right skewed. Most of the data
# lies between the 0-500 range but there are a lot of outliers with a much
# higher per capita income () (larger variability)

#2c
income_below_1000 <- life[life$Income < 1000, ]
income_atleast_1000 <- life[life$Income >= 1000, ]

#2d
plot(income_below_1000$Income, income_below_1000$Life,
     main="Life vs Income (Income < $1000)",
     xlab="Income per capita", ylab="Life Expectancy (years)",
     pch=19, col="blue")
```




```
correlation <- cor(income_below_1000$Income, income_below_1000$Life)
correlation
```

```
## [1] 0.752886
```

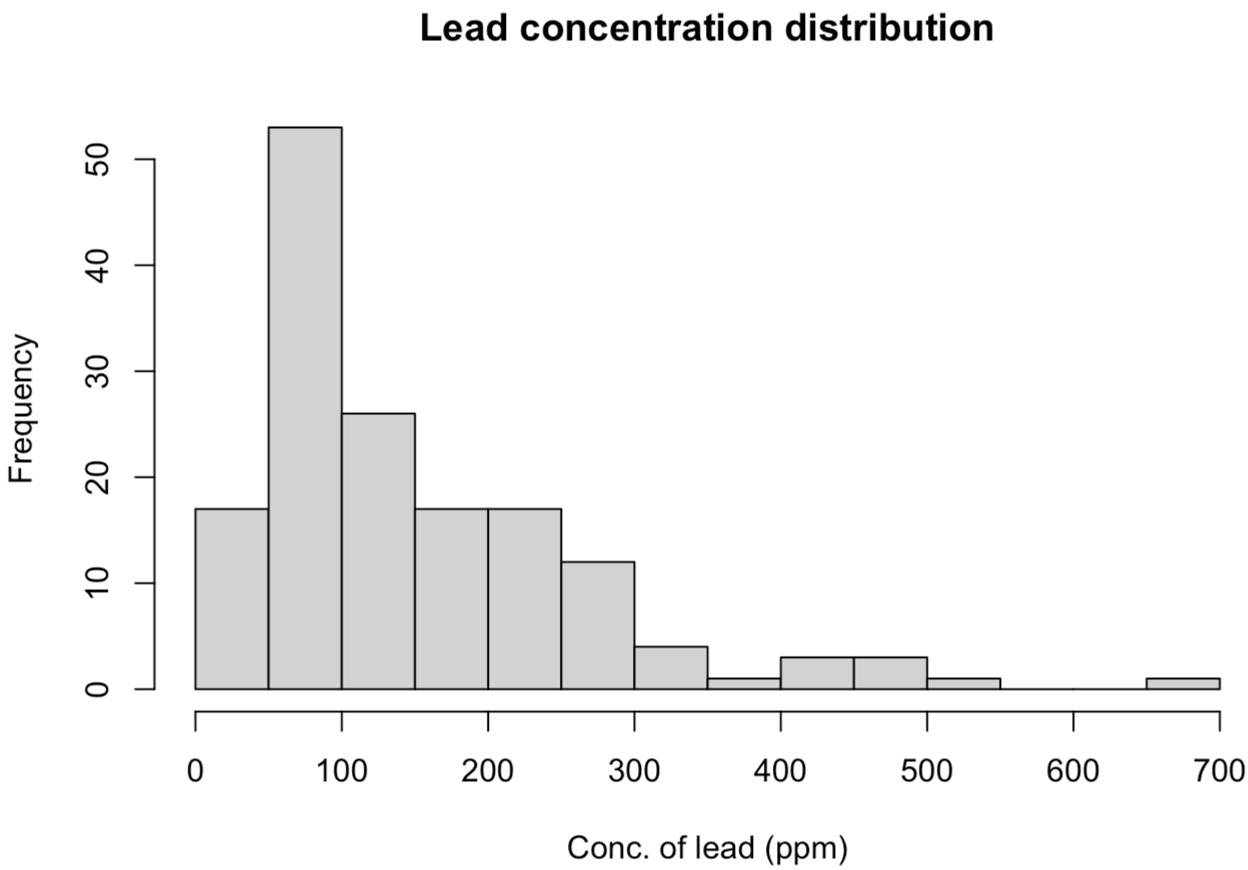
```
#3
maas <- read.table(
  "https://ucla.box.com/shared/static/tv3cxooyp6y8fh6gb0qj2cxihj8klg1h.txt",
  header = TRUE)

#3a
summary(maas$lead)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    37.0    72.5   123.0   153.4   207.0   654.0
```

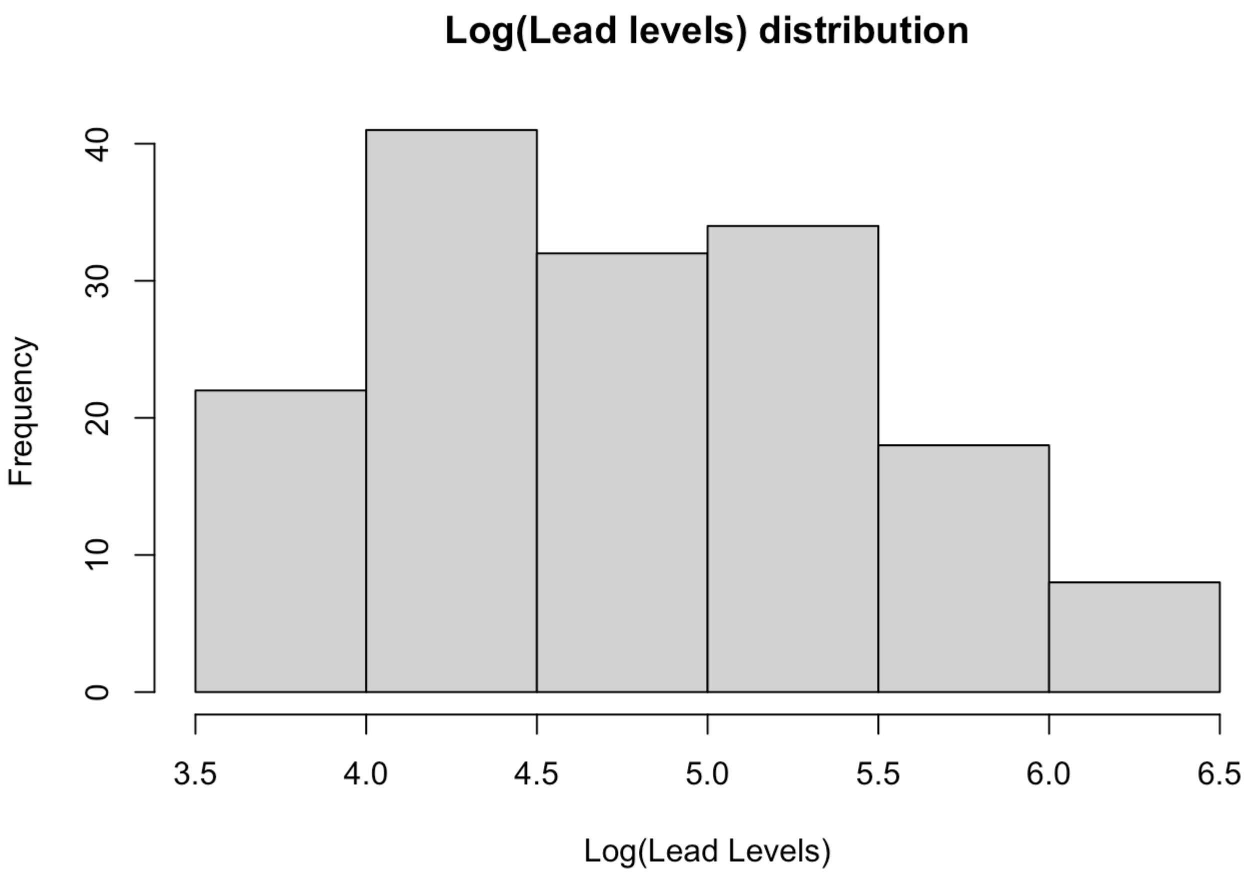
```
summary(maas$zinc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   113.0   198.0   326.0   469.7   674.5  1839.0
```

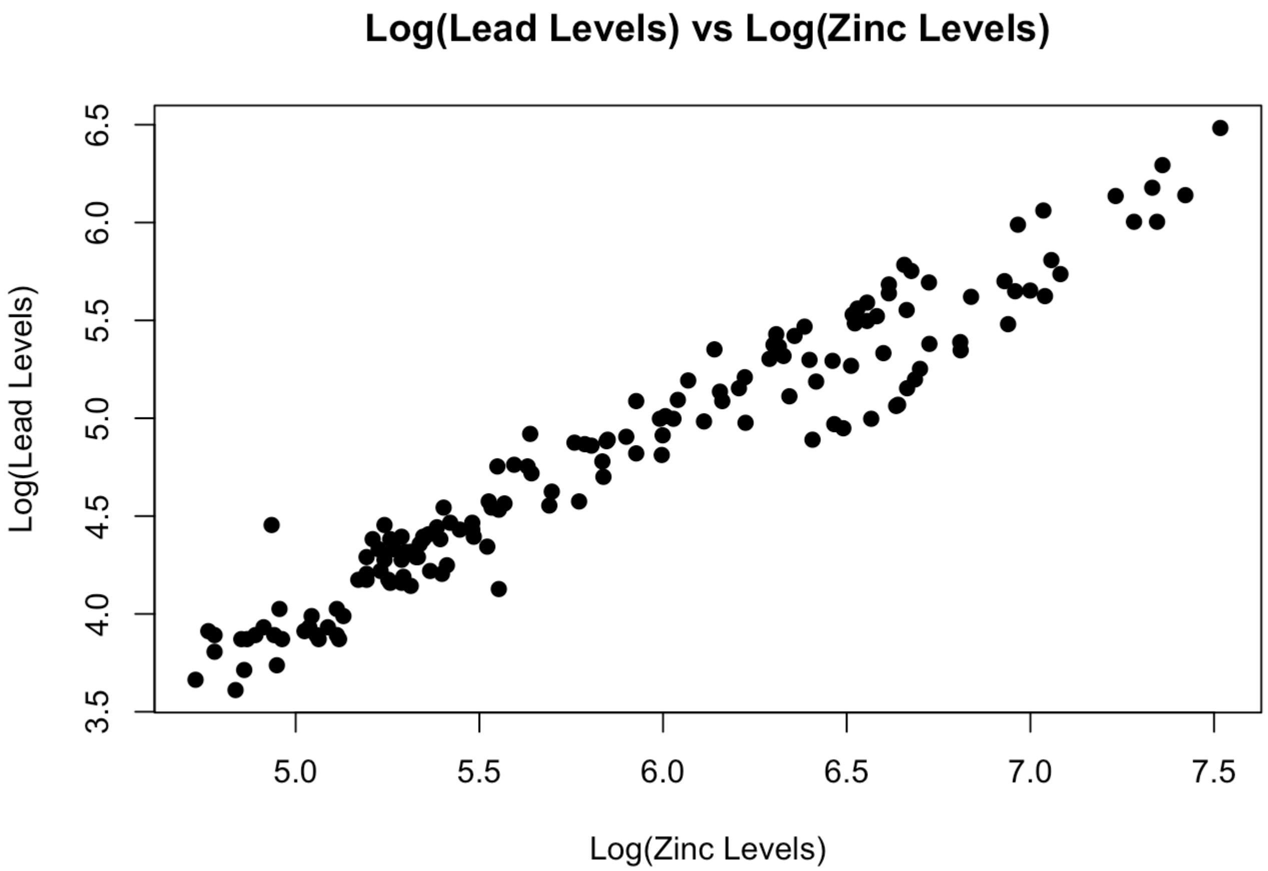
```
#3b
hist(maas$lead, xlab="Conc. of lead (ppm)",
     main="Lead concentration distribution")
```



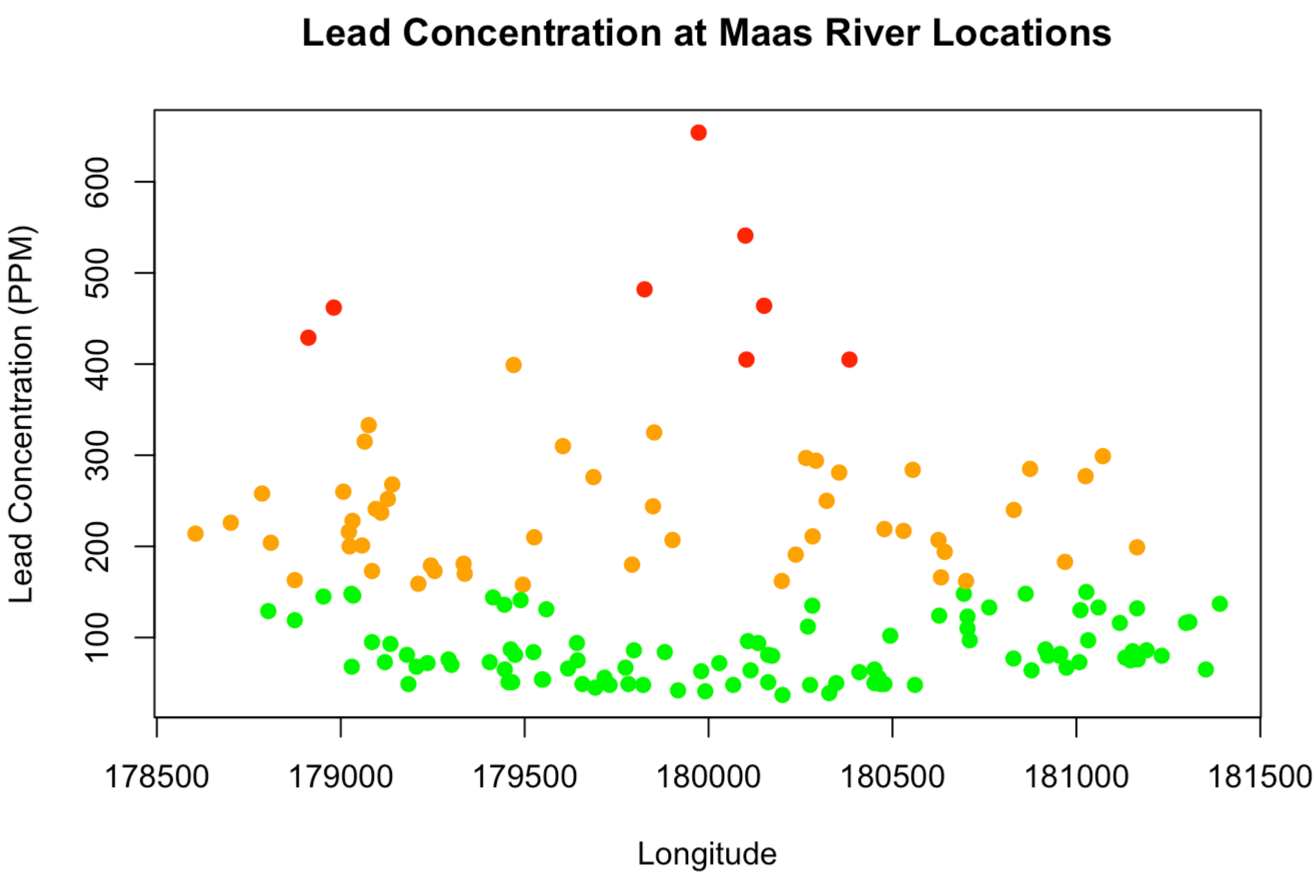

```
hist(log(maas$lead), xlab="Log(Lead Levels)",
     main="Log(Lead levels) distribution")
```



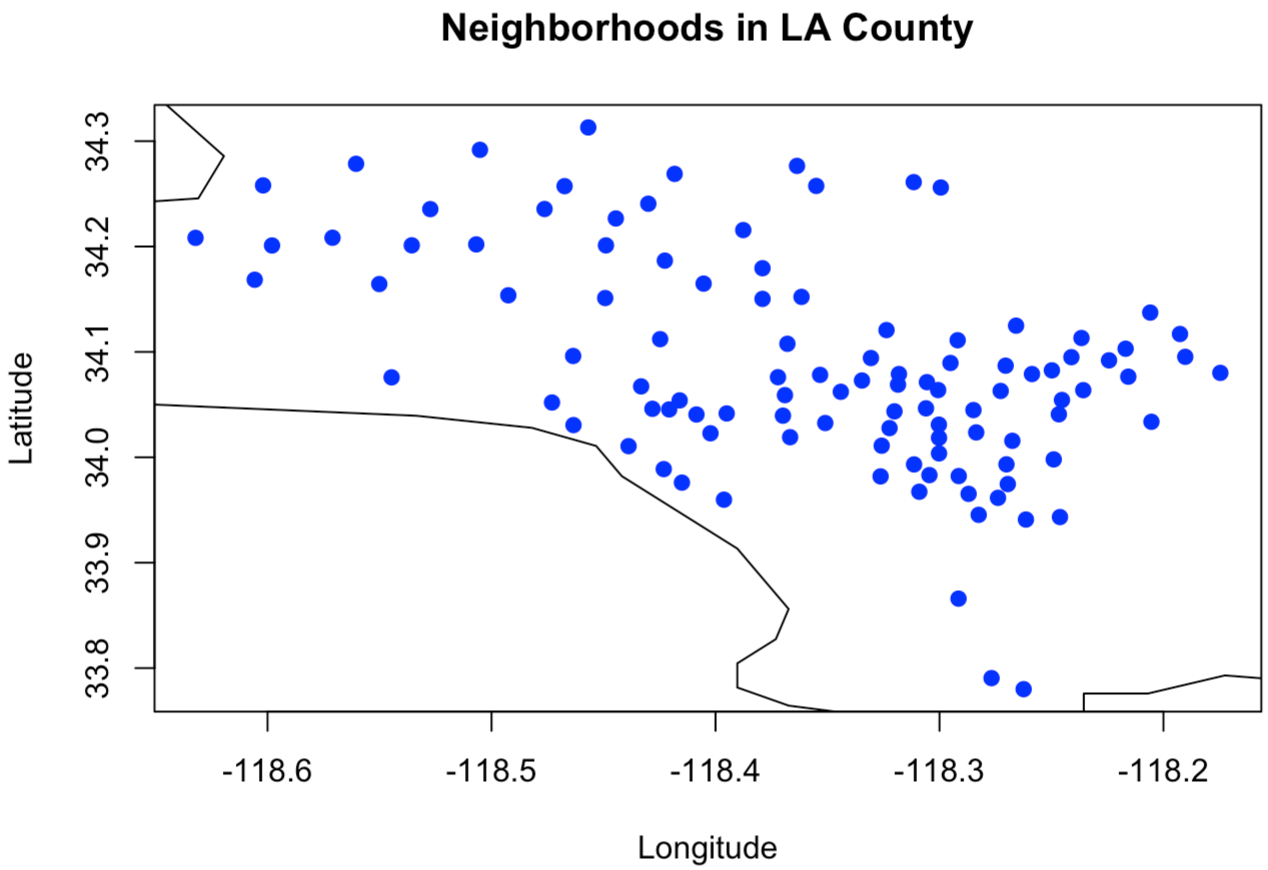
```
#3c
plot(log(maas$zinc), log(maas$lead), pch=19,
     xlab="Log(Zinc Levels)", ylab="Log(Lead Levels)",
     main="Log(Lead Levels) vs Log(Zinc Levels)")
```



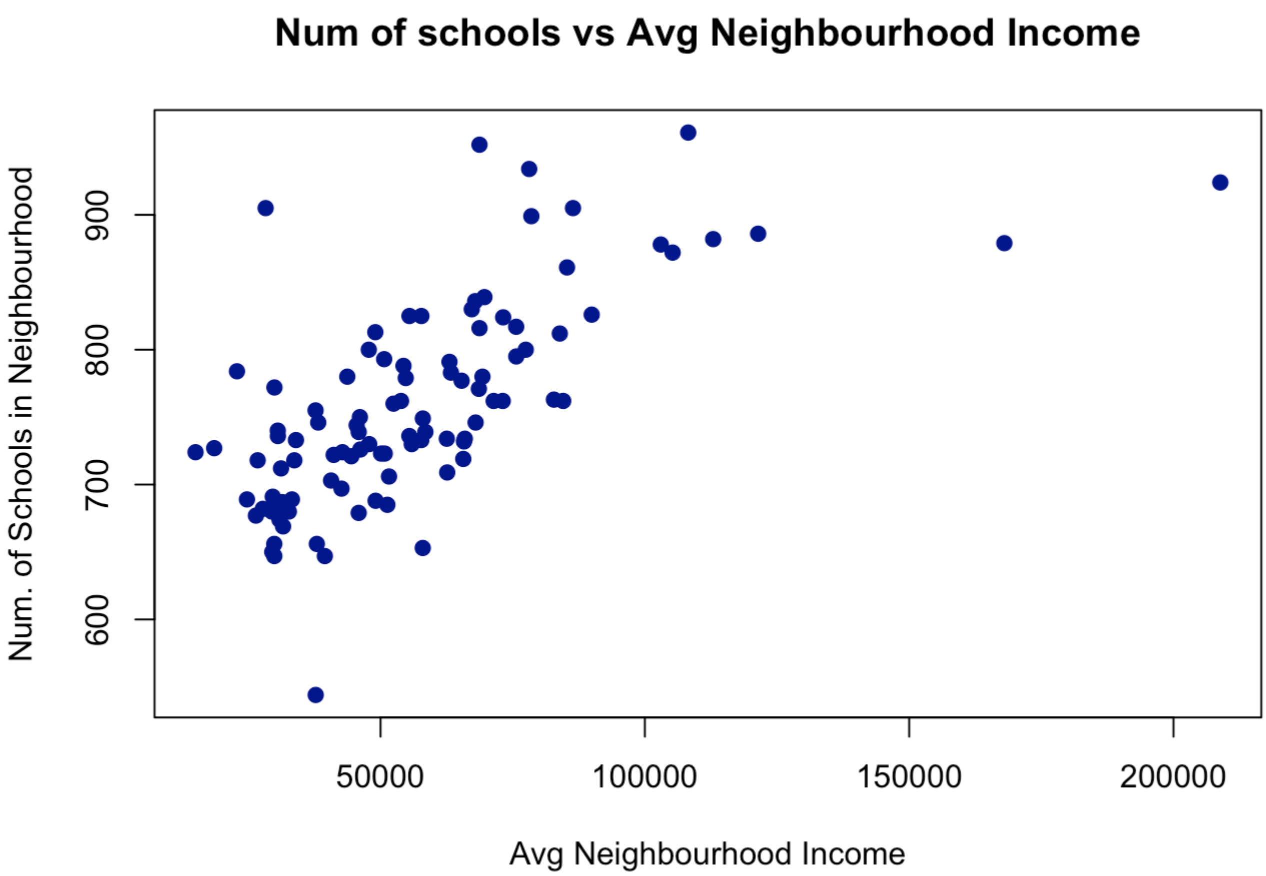

```
# the graph shows us a strong positive linear relationship between
# log(lead) and log(zinc)

#3d
x <- maas$x
y <- maas$lead

mycolors <- c("green", "orange", "red") #can be changed to other colors
mylevels <- cut(y, c(0, 150, 400, Inf))

plot(x, y,
     col=mycolors[as.numeric(mylevels)],
     pch= 19,
     main="Lead Concentration at Maas River Locations",
     xlab="Longitude",
     ylab="Lead Concentration (PPM)")
```




```
#4
LA <- read.table(
  "https://ucla.box.com/shared/static/d189x2gn5xfmcic0dmnhj2cw94jwvqpa.txt",
  header=TRUE)
library(maps)
library(mapdata)

#4a
#install.packages("maps")
#library(maps)

plot(x = LA$Longitude, y = LA$Latitude, pch=19,
     xlab="Longitude", ylab="Latitude",
     main="Neighborhoods in LA County",
     col="blue")
map("county", "california", add = TRUE)
```




```
#4b
LA_neighbourhoods_with_schools <- LA[LA$Schools > 0, ]
plot(LA_neighbourhoods_with_schools$Income,
     LA_neighbourhoods_with_schools$Schools,
     xlab="Avg Neighbourhood Income", ylab="Num. of Schools in Neighbourhood",
     main="Num of schools vs Avg Neighbourhood Income",
     pch=19, col="darkblue")
```




```
# There's a moderate positive linear relationship between income and LA school
# performance. That means that neighbourhoods with higher incomes tend to have
# better performing schools.

#5
customer_data <- read.csv(
  "https://ucla.box.com/shared/static/y2y8rcie7mjw2h5t92x9dfcp133tc90h.csv")

#5a
colSums(is.na(customer_data))
```

```
##        cust_id            age         gender         income      education
##              0             10              0              5              0
## marital_status   purchase_amt
##              0              7
```

```
# there are 22 missing values
# age, income, and purchase_amt have missing values 10, 5, and 7 respectively

#5b
class(customer_data$cust_id) #character
```

```
## [1] "character"
```

```
class(customer_data$age) #integer
```

```
## [1] "integer"
```

```
class(customer_data$gender) #character
```

```
## [1] "character"
```

```
class(customer_data$income) #integer
```

```
## [1] "integer"
```

```
class(customer_data$education) #character
```

```
## [1] "character"
```

```
class(customer_data$marital_status) #character
```

```
## [1] "character"
```

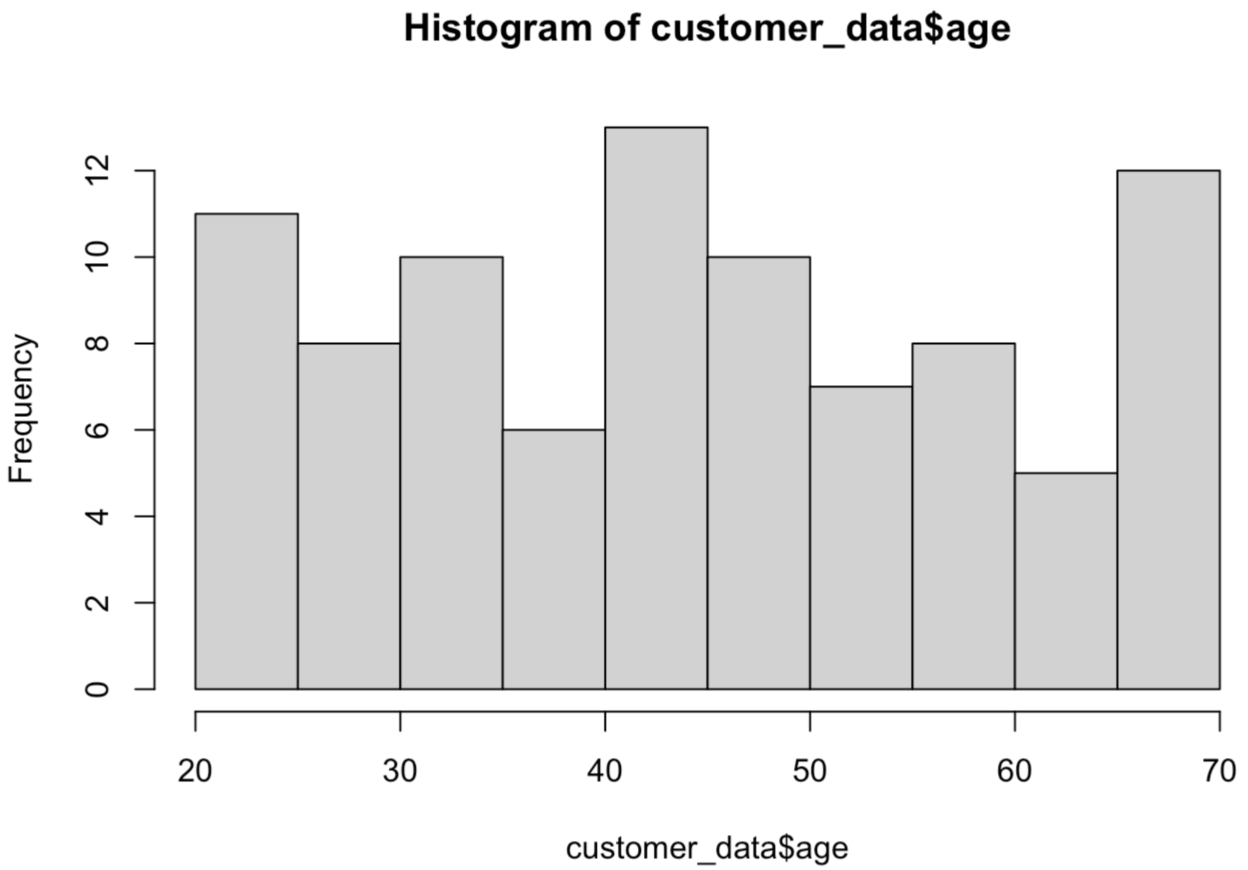
```
class(customer_data$purchase_amt) #integer
```

```
## [1] "integer"
```

```
# as gender, education, and marital_status have limited options as to what
# they could be it might be better to convert them to factor

customer_data$gender <- as.factor(customer_data$gender)
customer_data$education <- as.factor(customer_data$education)
customer_data$marital_status <- as.factor(customer_data$marital_status)

#5c
summary(customer_data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   20.00   32.00   44.00   44.99   56.75   70.00      10
```

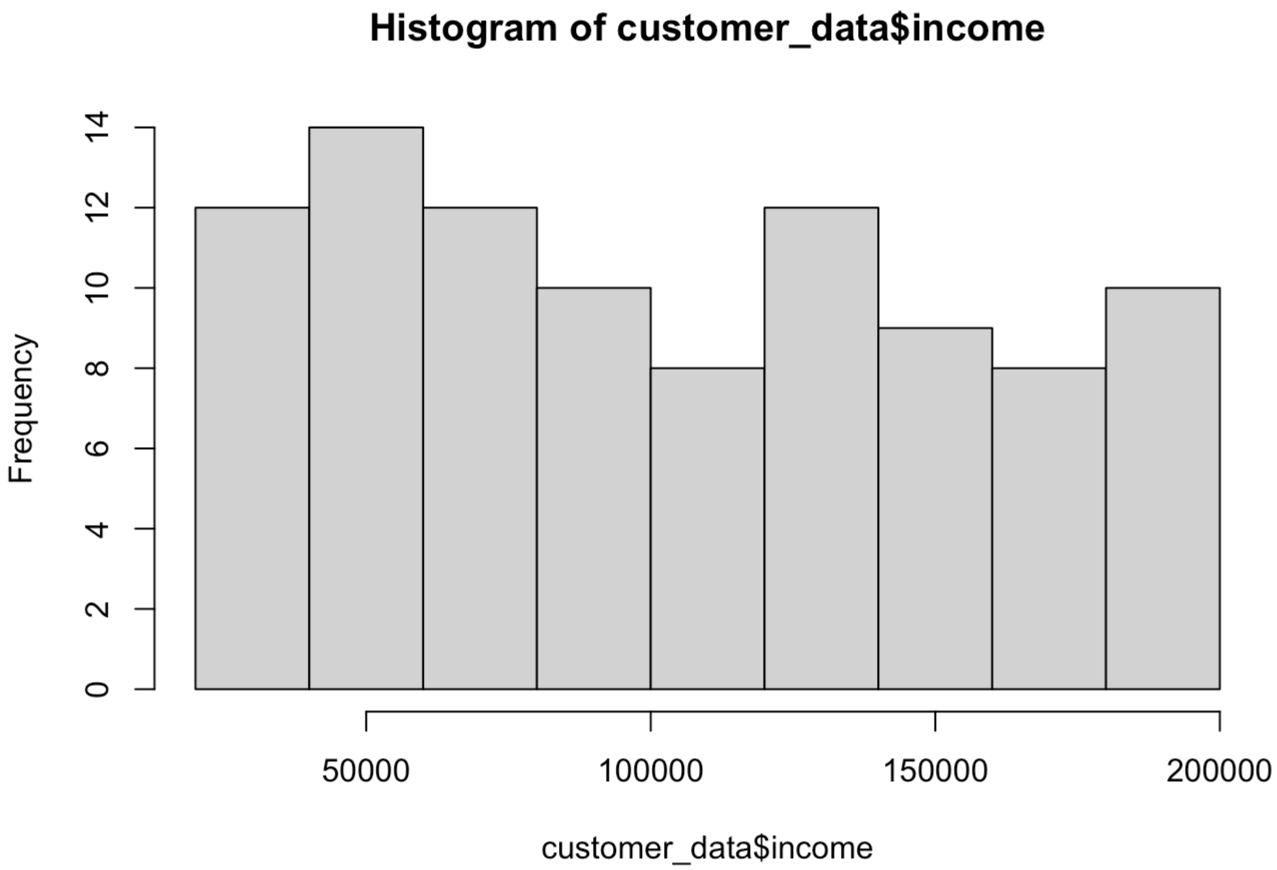
```
hist(customer_data$age)
```



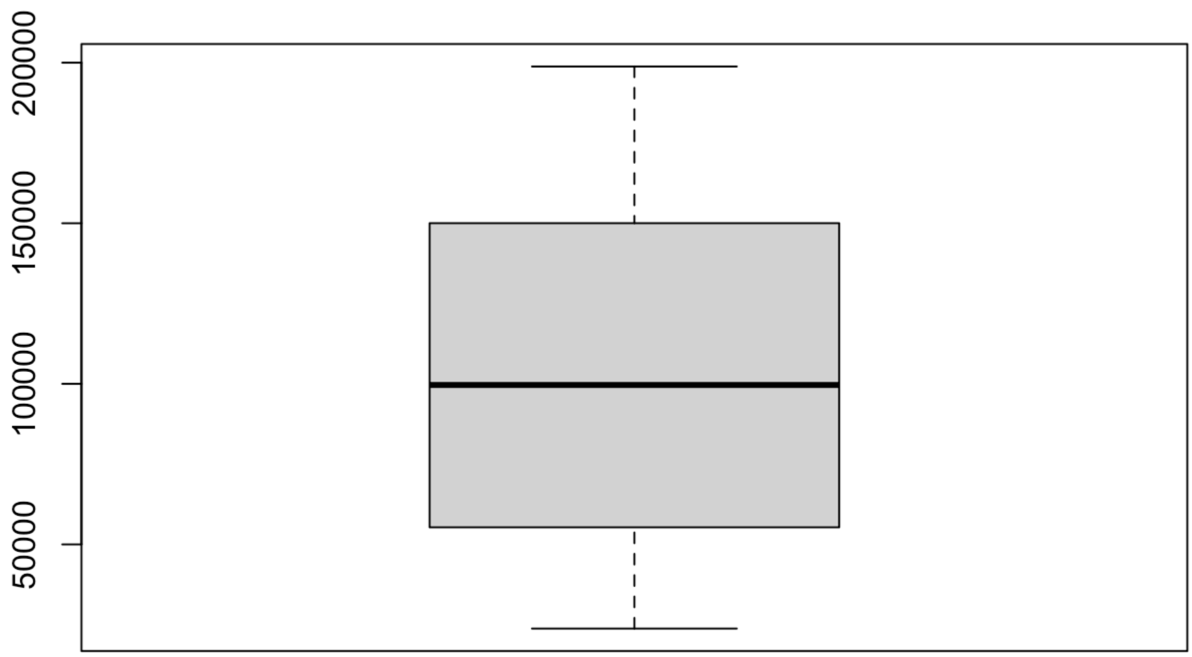

```
# a pretty even spread with seemingly no outliers

summary(customer_data$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   23798   55320   99637  103425  150030  198808       5
```

```
hist(customer_data$income)
```



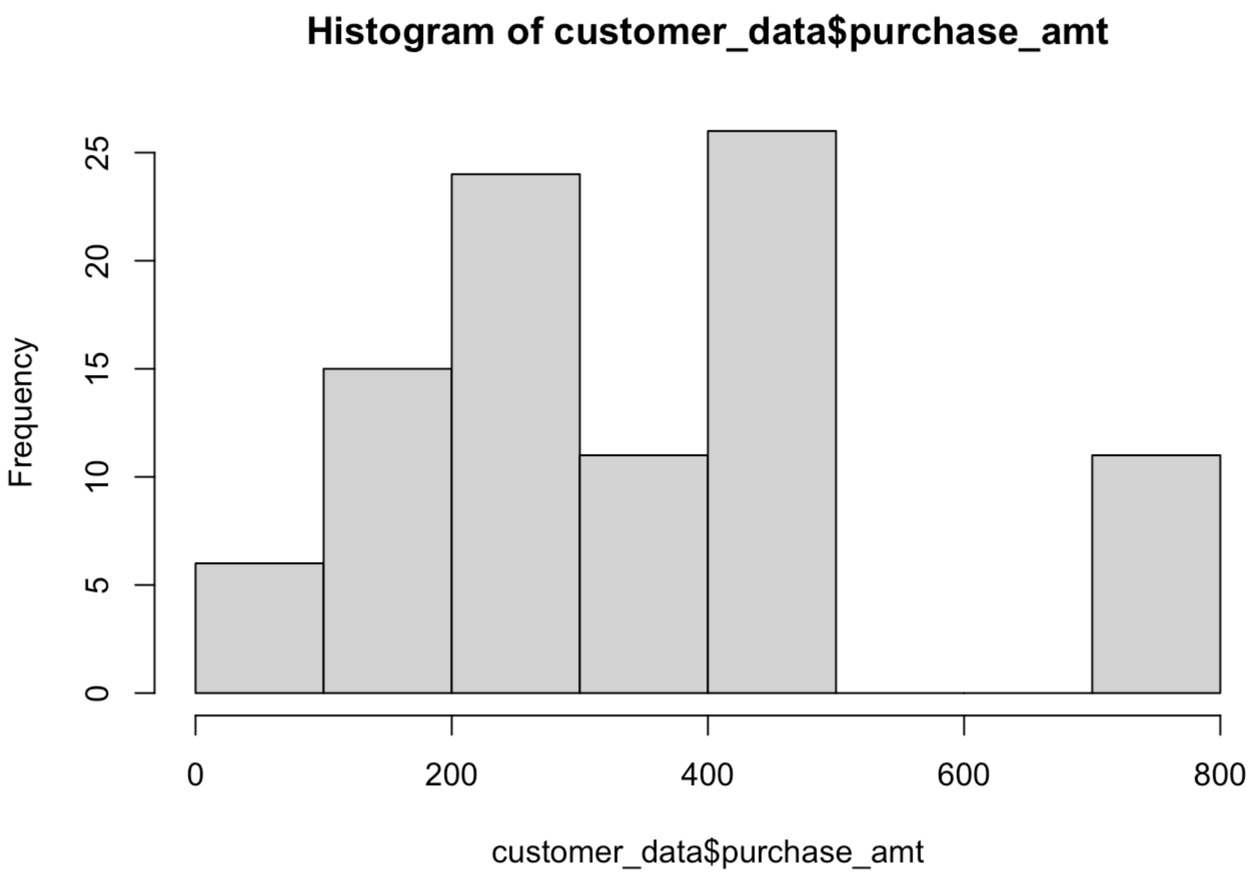

```
boxplot(customer_data$income)
```



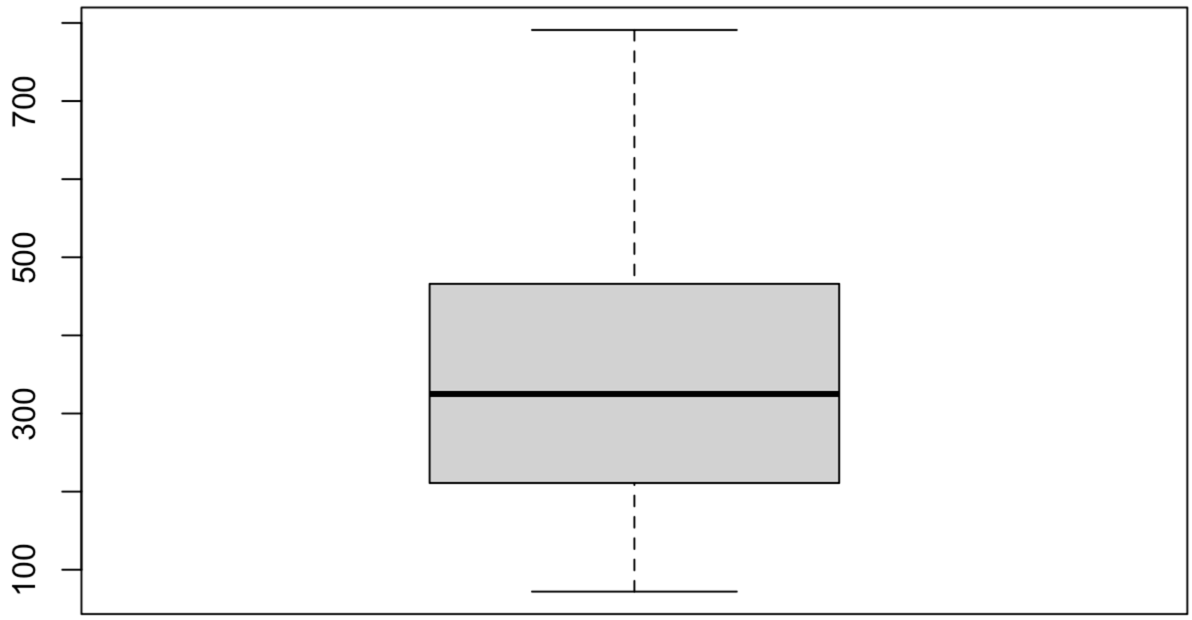
```
# a pretty even spread with seemingly no outliers

summary(customer_data$purchase_amt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    72.0   211.0   325.0   356.2   466.0   791.0       7
```

```
hist(customer_data$purchase_amt)
```




```
boxplot(customer_data$purchase_amt)
```




```
# the histogram shows us there are some gaps in the distribution but
# the boxplot shows us that there are no outliers and all the values lie within
# 1.5IQR above and below the 3rd and 1st upper quartile
```