2a. The following code is used to get summary statistics on the dataset. The output is also shown below:

a <- c(12,14,15)

summary(dataset[,a])

GDP medals athletes

Min. : 0.0110 Min. : 0.000 Min. : 0.00

1st Qu.: 0.1212 1st Qu.: 0.000 1st Qu.: 0.00

Median : 0.3849 Median : 0.000 Median : 2.00

Mean : 1.1691 Mean : 1.751 Mean : 18.17

3rd Qu.: 1.5127 3rd Qu.: 0.000 3rd Qu.: 13.00

Max. :14.5230 Max. :37.000 Max. :230.00

NA's :156

2b. The head function is used to print the first 5 observations for country, year, medals, athletes, GDP data:

b <- c(2,4,14,12,15)

head(dataset[,b],n = 5)

year country medals GDP athletes

1 1980 Albania 0 NA 0

2 1984 Albania 0 0.0641 0

3 1988 Albania 0 0.0637 0

4 1992 Albania 0 0.0206 0

5 1994 Albania 0 0.0587 0

2c. The table function is used to print the number of data points corresponding to each level:

table(dataset$year)

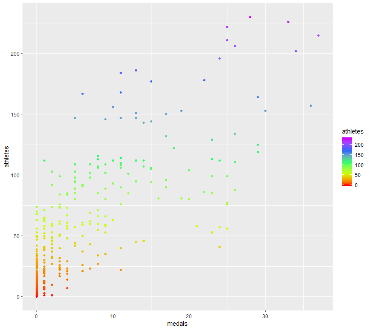
1980 1984 1988 1992 1994 1998 2002 2006 2010 2014

117 117 117 113 110 110 110 110 109 109

2d. The following code prints out a colorful scatterplot. There is a positive correlation between medals and athletes

c <- c(14,15)

dataset[,c] %>%

 ggplot(aes(x = medals, y = athletes, color = athletes)) +

scale\_color\_gradientn(colours = rainbow(5))+

geom\_point()

2e. It is possible that higher GDP means better trained athletes, which means more medals. Higher GDP means higher population. Therefore, more athletes

2f.

d <- c(12,14)

dataset[,d] %>%

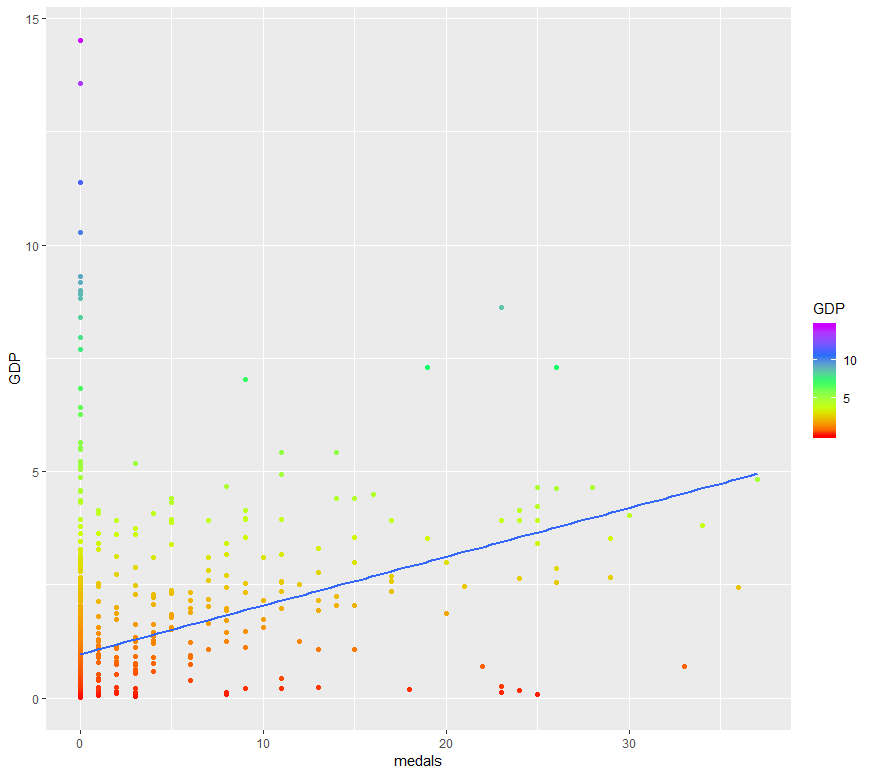
ggplot(aes(x = medals, y = GDP, color = GDP)) +

scale\_color\_gradientn(colours = rainbow(5))+

geom\_point()+

geom\_smooth(method = 'lm', fill = NA)

##there is a positive correlation between medals and GDP



2g.

d <- c(11,14)

dataset[,d] %>%

ggplot(aes(x = medals, y = population, color = population)) +

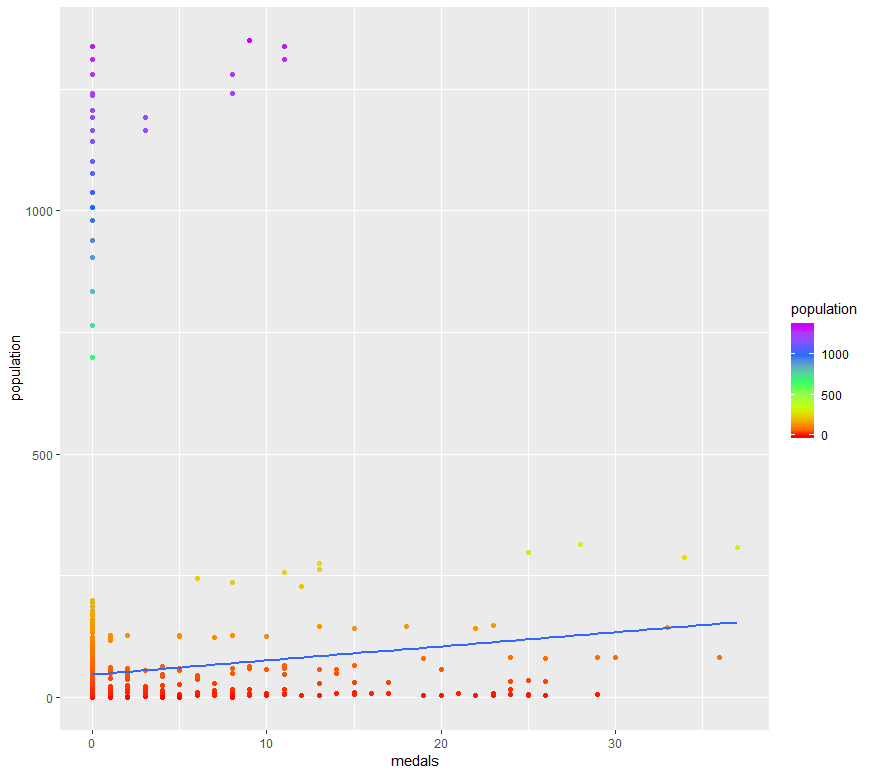
scale\_color\_gradientn(colours = rainbow(5))+

geom\_point() +

geom\_smooth(method = 'lm', fill = NA)

##there seems to be a positive correlation between medals and population

##There might be two different clusters of data points



2h.

d <- c(5,14)

dataset[,d] %>%

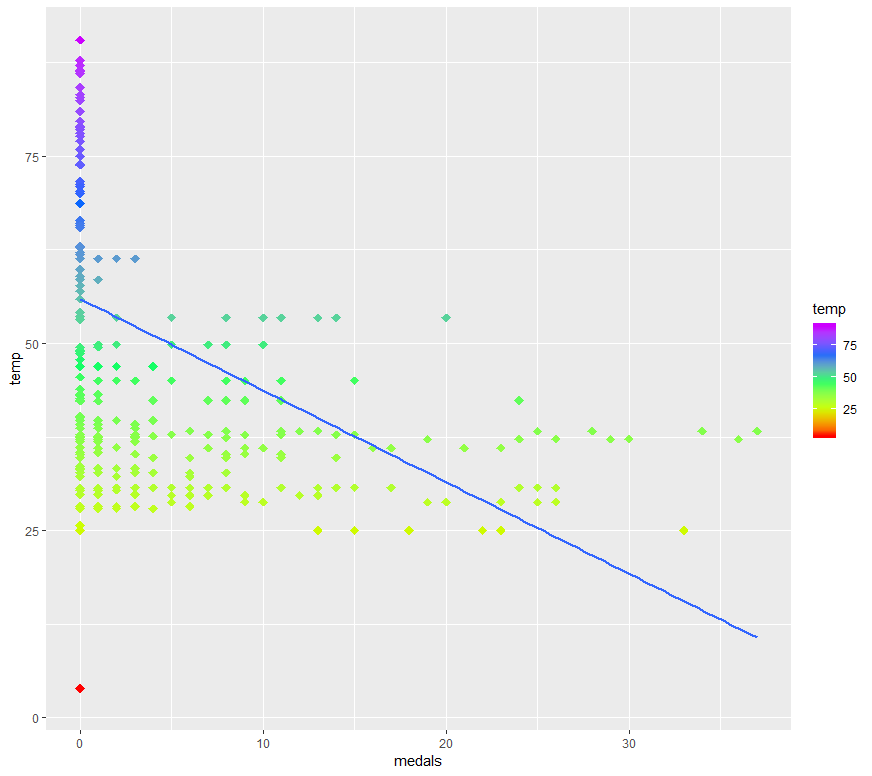
ggplot(aes(x = medals, y = temp, color = temp)) +

scale\_color\_gradientn(colours = rainbow(5))+

geom\_point(shape = 18, size = 3)+

geom\_smooth(method = 'lm', fill = NA)

##There is a negative correlation between medals and temperature



3a.

a <- c(7,12,13,16)

datasetSiblingsHeightWage <- dataset[,a]

summary(datasetSiblingsHeightWage)

##The mean height stayed relatively the same, which is odd because you would expect people to grow taller becaues of puberty

##Also, the max height dropped from 83 to 81 showing error in the data collection process

##The amount of NA's is very concerning because it will skew the results

##The mean siblings is 3, but the max siblings is 29 and highly unlikely. This eludes to the possibility that some data points are entered in wrong

##Also, the max wage is 1533 and highly unlikely. Possible the data was entered in wrong.

siblings height81 height85 wage96

Min. :-3.000 Min. :48.00 Min. :48.00 Min. : 0.000

1st Qu.: 2.000 1st Qu.:64.00 1st Qu.:64.00 1st Qu.: 6.743

Median : 3.000 Median :67.00 Median :67.00 Median : 10.783

Mean : 3.844 Mean :67.01 Mean :67.08 Mean : 14.177

3rd Qu.: 5.000 3rd Qu.:70.00 3rd Qu.:70.00 3rd Qu.: 16.213

Max. :29.000 Max. :83.00 Max. :81.00 Max. :1533.333

NA's :543 NA's :1823 NA's :5756

3b.

wageHeight <- c(3,4)

datasetSiblingsHeightWage[,wageHeight] %>%

ggplot(aes(x = wage96, y = height85, color = height85)) +

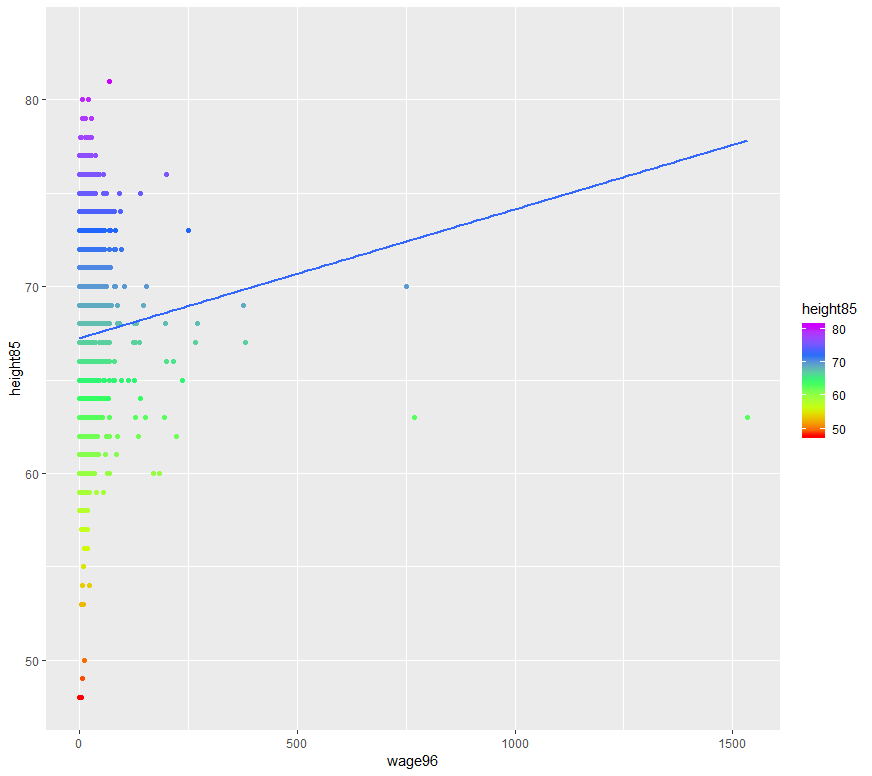
scale\_color\_gradientn(colours = rainbow(5))+

geom\_point()+

geom\_smooth(method = 'lm', fill = NA)

##the slope is positive which shows a positive correlation between wage and height

##there are outliers in the dataset and is shown in the graph



3d.

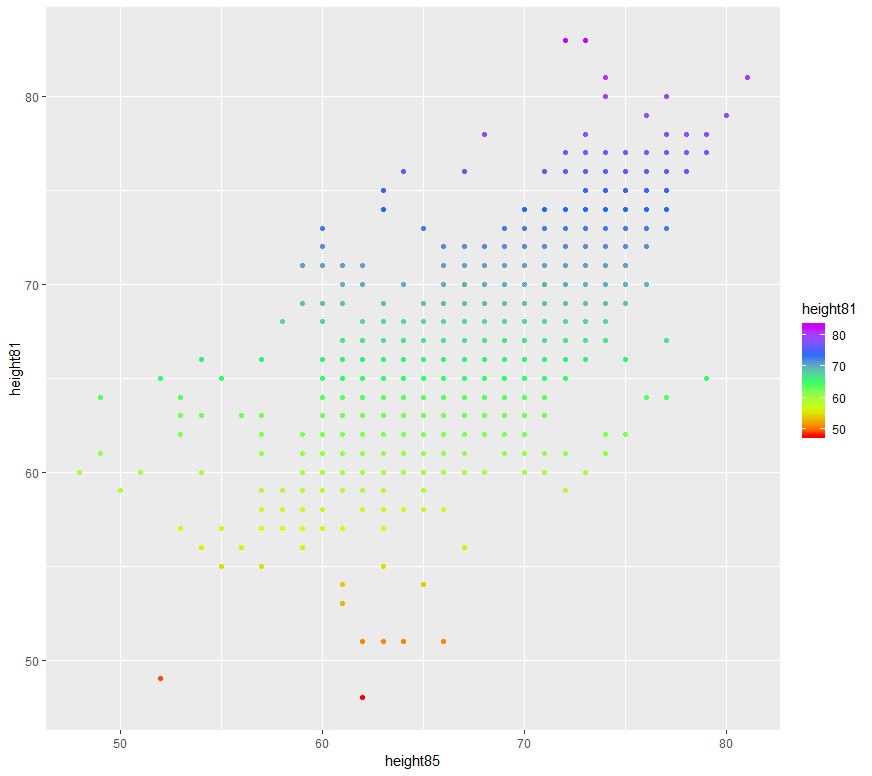
adolescentHeightVsAdultHeight <- c(2,3)

datasetSiblingsHeightWage[,adolescentHeightVsAdultHeight] %>%

ggplot(aes(x=height85, y = height81, color = height81)) +

scale\_color\_gradientn(colours = rainbow(5)) +

geom\_point()



##lets look at points where y = 70. If the point has x < 70, then those points signal a decrease in height as person ages

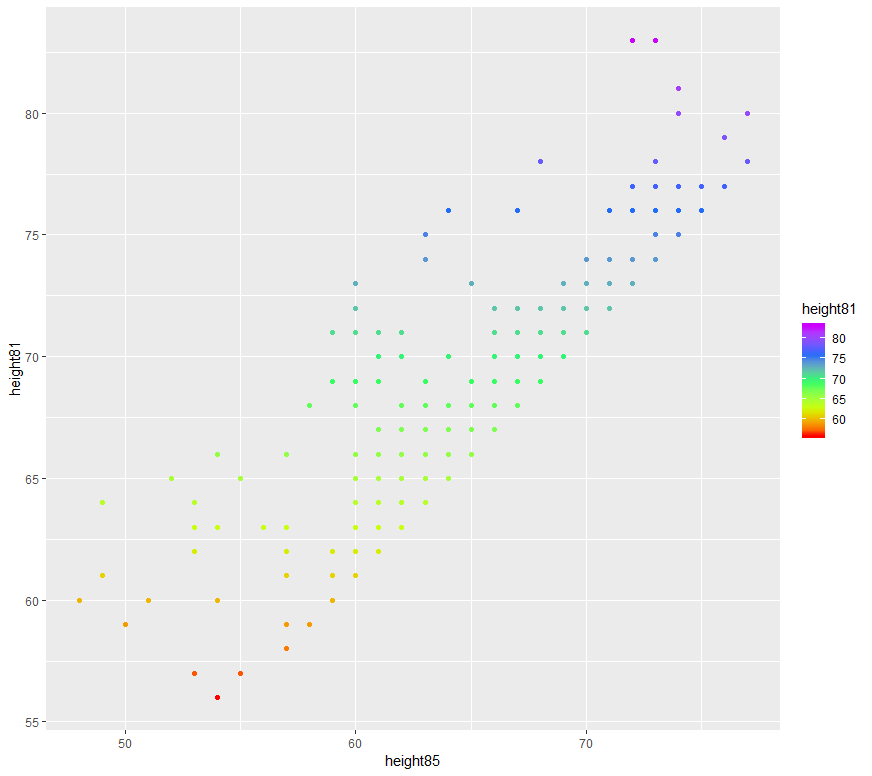
##therefore, we ignore those points where people's adolescent height is greater than their adult height

##Should not use these observations because it is unlikely that one's height will decrease as one ages.

##the following plot shows the points where people’s adolescent height is greater than adult height

datasetSiblingsHeightWage[which(datasetSiblingsHeightWage$height81 > datasetSiblingsHeightWage$height85),] %>%

ggplot(aes(x=height85, y = height81, color = height81)) +

 scale\_color\_gradientn(colours = rainbow(5)) +

geom\_point()

library(tidyr)

library(ggplot2)

library(dplyr)

##make sure the data file is saved in the working directory

dataset <- read.csv("olympics.txt", header = TRUE, sep = " ")

##A

a <- c(12,14,15)

summary(dataset[,a])

##B

b <- c(2,4,14,12,15)

head(dataset[,b],n = 5)

##C

table(dataset$year)

##D

c <- c(14,15)

dataset[,c] %>%

ggplot(aes(x = medals, y = athletes, color = athletes)) +

scale\_color\_gradientn(colours = rainbow(5))+

geom\_point()

##looks like a positive correlation between medals and athletes

##E

##it is possible that higher gdp means better trained athletes, which means more medals

##higher gdp means higher population. therefore, more atheletes

##F

d <- c(12,14)

dataset[,d] %>%

ggplot(aes(x = medals, y = GDP, color = GDP)) +

scale\_color\_gradientn(colours = rainbow(5))+

geom\_point()+

geom\_smooth(method = 'lm', fill = NA)

##there is a positive correlation between medals and GDP

##G

d <- c(11,14)

dataset[,d] %>%

ggplot(aes(x = medals, y = population, color = population)) +

scale\_color\_gradientn(colours = rainbow(5))+

geom\_point() +

geom\_smooth(method = 'lm', fill = NA)

##there seems to be a positive correlation between medals and population

##h

d <- c(5,14)

dataset[,d] %>%

ggplot(aes(x = medals, y = temp, color = temp)) +

scale\_color\_gradientn(colours = rainbow(5))+

geom\_point(shape = 18, size = 3)+

geom\_smooth(method = 'lm', fill = NA)

##There is a negative correlation between medals and temperature

library(tidyr)

library(ggplot2)

library(dplyr)

dataset <- read.csv("HeightWage\_MenWomenUS\_HW.csv", header = TRUE, sep = ",")

##A

a <- c(7,12,13,16)

datasetSiblingsHeightWage <- dataset[,a]

summary(datasetSiblingsHeightWage)

##The mean height stayed relatively the same, which is odd because you would expect people to grow taller becaues of puberty

##Also, the max height dropped from 83 to 81 showing error in the data collection process

##The amount of NA's is very concerning becaues it will skew the results

##The mean siblings is 3, but the max siblings is 29 and highly unlikely. This eludes to the possibility that some data points are entered in wrong

##Also, the max wage is 1533 and highly unlikely. Possible the data was entered in wrong.

##B

wageHeight <- c(3,4)

datasetSiblingsHeightWage[,wageHeight] %>%

ggplot(aes(x = wage96, y = height85, color = height85)) +

scale\_color\_gradientn(colours = rainbow(5))+

geom\_point()+

geom\_smooth(method = 'lm', fill = NA)

##the slope is positive which shows a positive correlation between wage and height

##there are outliers in the dataset and is shown in the graph

##C

datasetSiblingsHeightWage[which(datasetSiblingsHeightWage$wage96 < 500),wageHeight] %>%

ggplot(aes(x= wage96, y = height85, color = height85)) +

scale\_color\_gradientn(colours = rainbow(5))

geom\_point()+

geom\_smooth(method = 'lm', fill = NA)

##there is a positive correlation between wage and height

##D

adolescentHeightVsAdultHeight <- c(2,3)

datasetSiblingsHeightWage[,adolescentHeightVsAdultHeight] %>%

ggplot(aes(x=height85, y = height81, color = height81)) +

scale\_color\_gradientn(colours = rainbow(5)) +

geom\_point()

##lets look at points where y = 70. If the point has x < 70, then those points signal a decrease in height as person ages

##therefore, we ignore those points where people's adolescent height is greater than their adult height

##Should not use these observations because it is unlikely that one's height will decrease as one ages.

datasetSiblingsHeightWage[which(datasetSiblingsHeightWage$height81 > datasetSiblingsHeightWage$height85),] %>%

ggplot(aes(x=height85, y = height81, color = height81)) +

scale\_color\_gradientn(colours = rainbow(5)) +

geom\_point()