problemSet_2

Danya Zhang

2022-09-25

#Fuel Economy 1.

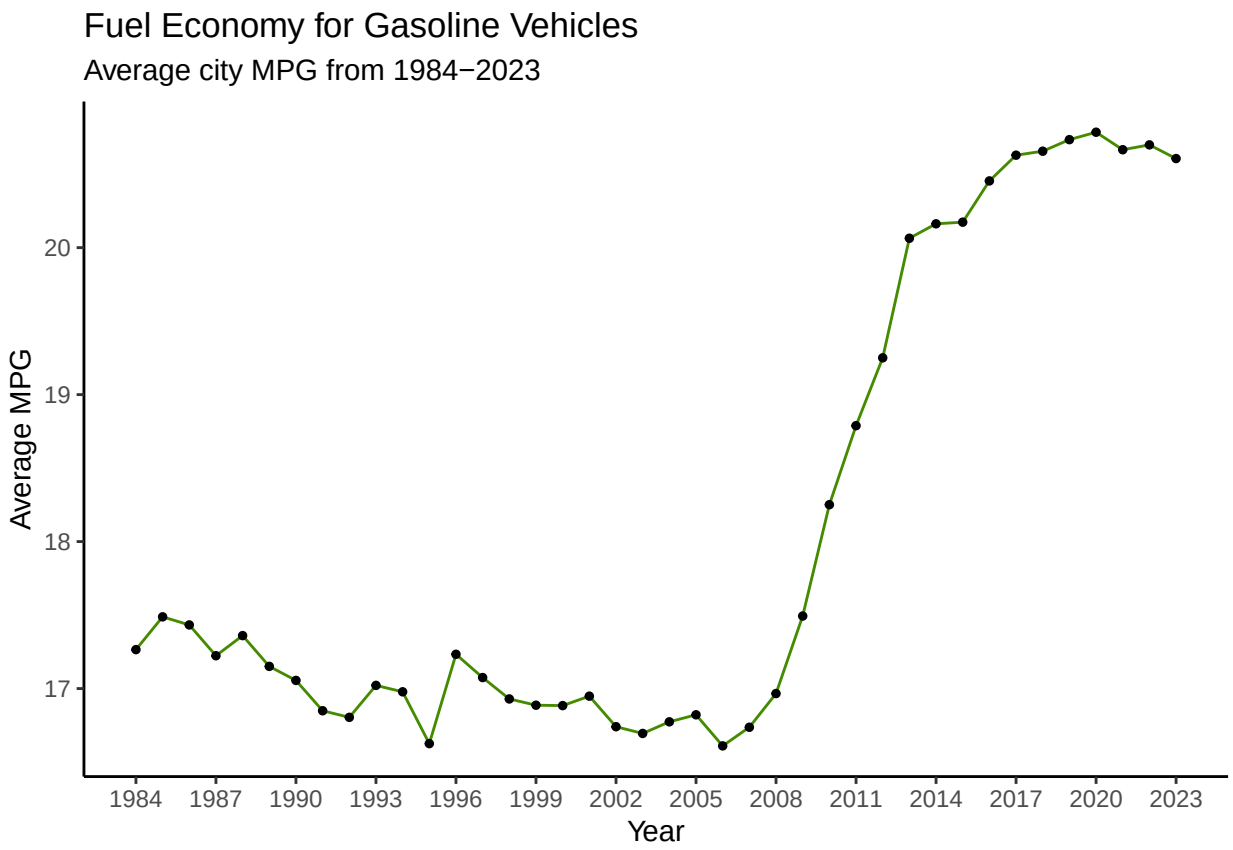
```
library(ggplot2)
vehicles <- read.csv("~/Documents/MSSP/MA615/MA615-R/vehicles.csv")

#only get columns related to fuel economy
vehsub <- vehicles[,c(5,7,16,18,32,35,37,47,49,59:62,64,71,81:83)]

#only get gas vehicles
hybrid_fuels <- c("Natural Gas", "E85","Propane","Electricity")
not_hybrid_ind <- grep(paste(hybrid_fuels,collapse="|"), vehsub$fuelType2, value=FALSE,invert=TRUE)
no_hybrid <- vehsub[not_hybrid_ind,]
gas_ind <- grep("Gasoline",no_hybrid$fuelType1)
gas <- no_hybrid[gas_ind,]

#calculate average city mpg of gasoline-only vehicles for each year
city08_mean <- data.frame(aggregate(gas$city08, list(gas$year), FUN=mean))
  colnames(city08_mean) <- c("year","mpg_mean")

ggplot(city08_mean, aes(year, mpg_mean)) +
  labs(title = "Fuel Economy for Gasoline Vehicles",
       subtitle = "Average city MPG from 1984-2023")+
  xlab("Year") + ylab("Average MPG")+
  geom_line(aes(x = year, y = mpg_mean), color = 'chartreuse4') +
  geom_point(size = 1) +
  scale_x_continuous(breaks=seq(1984, 2023, 3)) +
  theme_classic()
```



The above graph shows the average city mpg of gasoline-only vehicles for each year from 1984-2023.

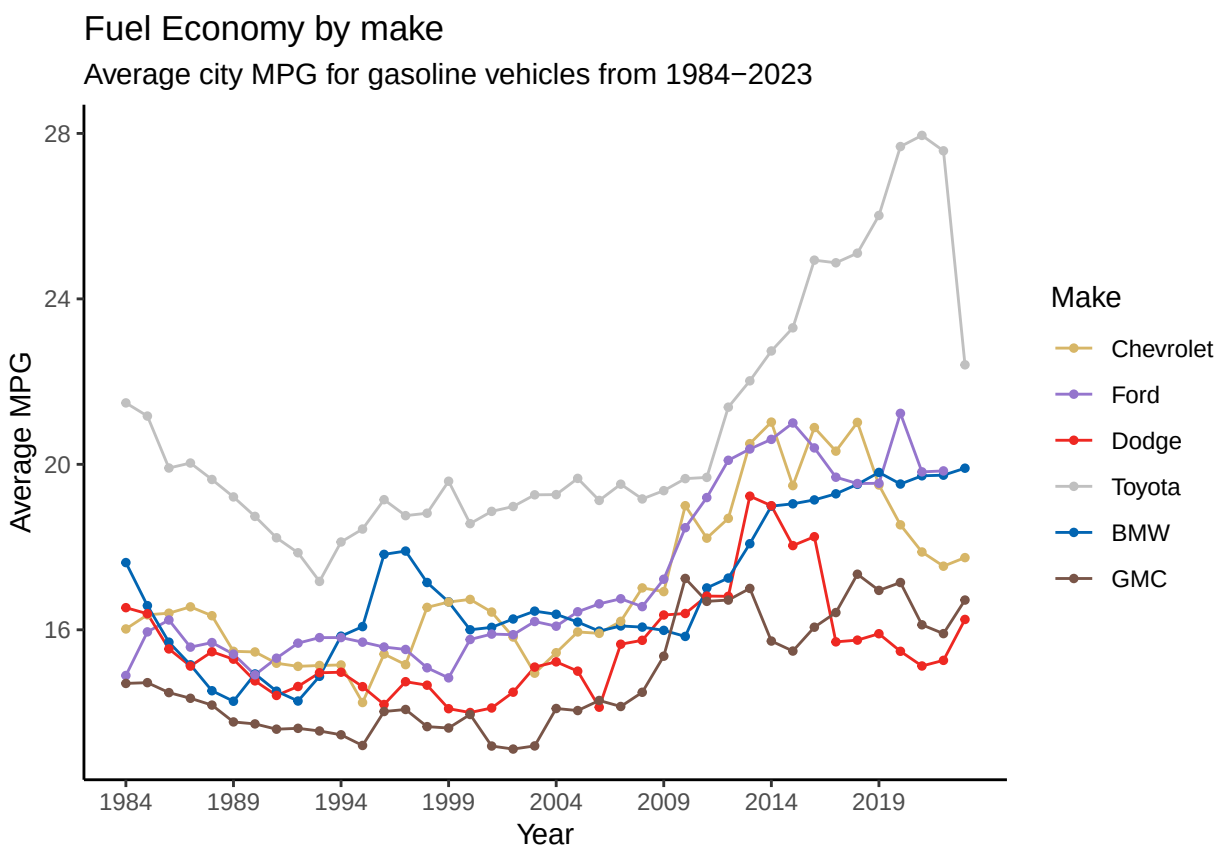
#Fuel Economy 2.

```
#get top 8 makes by frequency
data2 <- data.frame(gas$year,gas$make, gas$city08)
  colnames(data2) <- c("year","make","citympg")
brand_count <- as.data.frame(table(data2$make))
  colnames(brand_count) <- c("Make","Freq")
brand_count_sort <- brand_count[order(-brand_count$Freq),]
top10 <- brand_count_sort[1:6,]
cleaned <- data2[data2$make %in% top10$Make,]

#get mean city mpg for each year of most popular vehicles
mpg_mean <- data.frame(aggregate(cleaned$citympg, list(cleaned$year,cleaned$make), FUN=mean))
  colnames(mpg_mean) <- c("year","make","mean")
  mpg_mean$make <- factor(mpg_mean$make, levels = c("Chevrolet", "Ford",
        "Dodge","Toyota","BMW","GMC"))

ggplot(mpg_mean, aes(year, mean, color=factor(make))) +
  labs(title = "Fuel Economy by make",
       subtitle = "Average city MPG for gasoline vehicles from 1984-2023",
       x = "Year", y = "Average MPG") +
  labs(color = "Make") +
  geom_line() +
  geom_point(size = 1) +
  scale_x_continuous(breaks=seq(1984, 2023, 5)) +
  scale_color_manual(values =
```

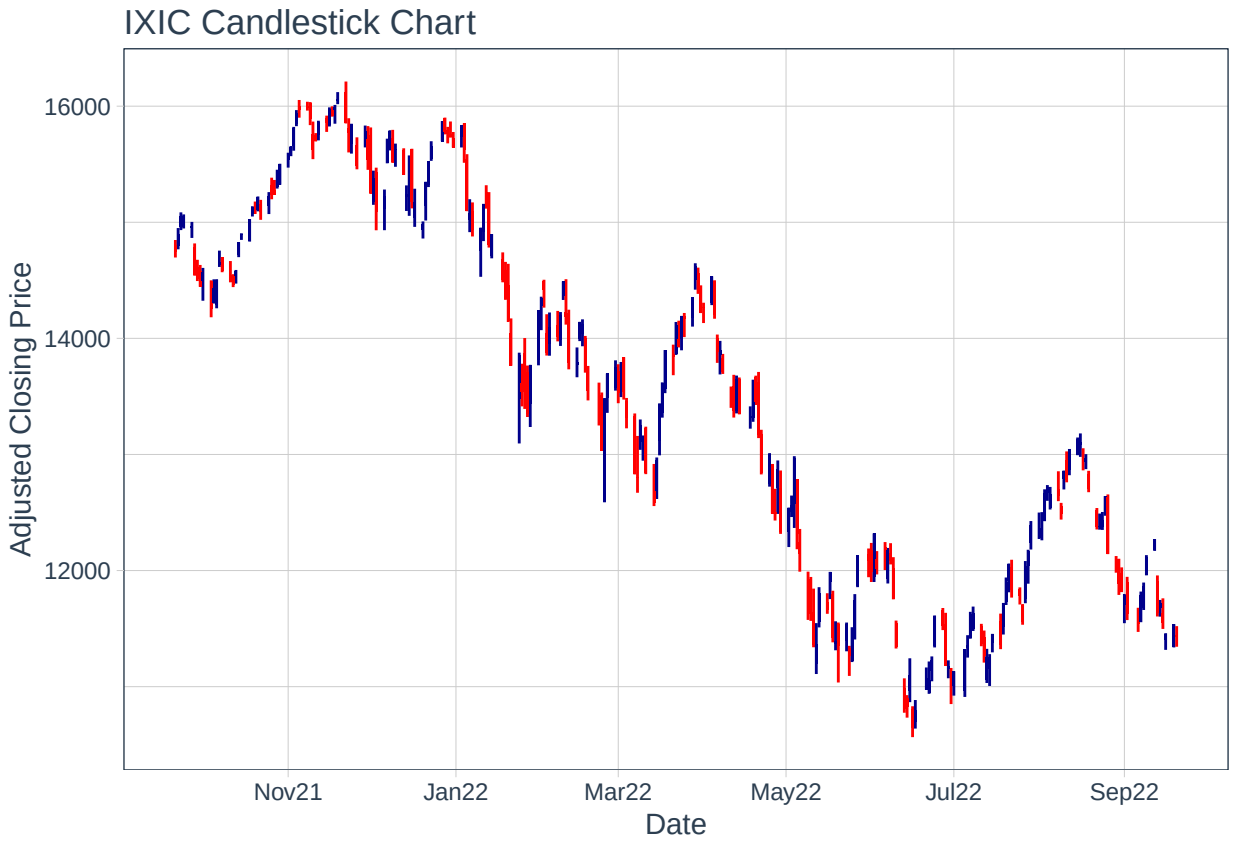
```
    c("#D7B667","#9575CD","#EE2822","#C0C0C0","#0265B2","#795548")) +
  theme_classic()
```



This plot shows the average city mpg vs. year for gasoline vehicles from the 6 makers whose frequency is the highest in the vehicles dataset. From the plot, it seems that Toyota, Ford, and BMW have made the most progress. BMW initial and final average mpg hasn't changed much, but their upward trend can be seen. Toyata's and Ford's recent mpg have fluctuated and decreased quite a bit.

#NASDAQ Composite

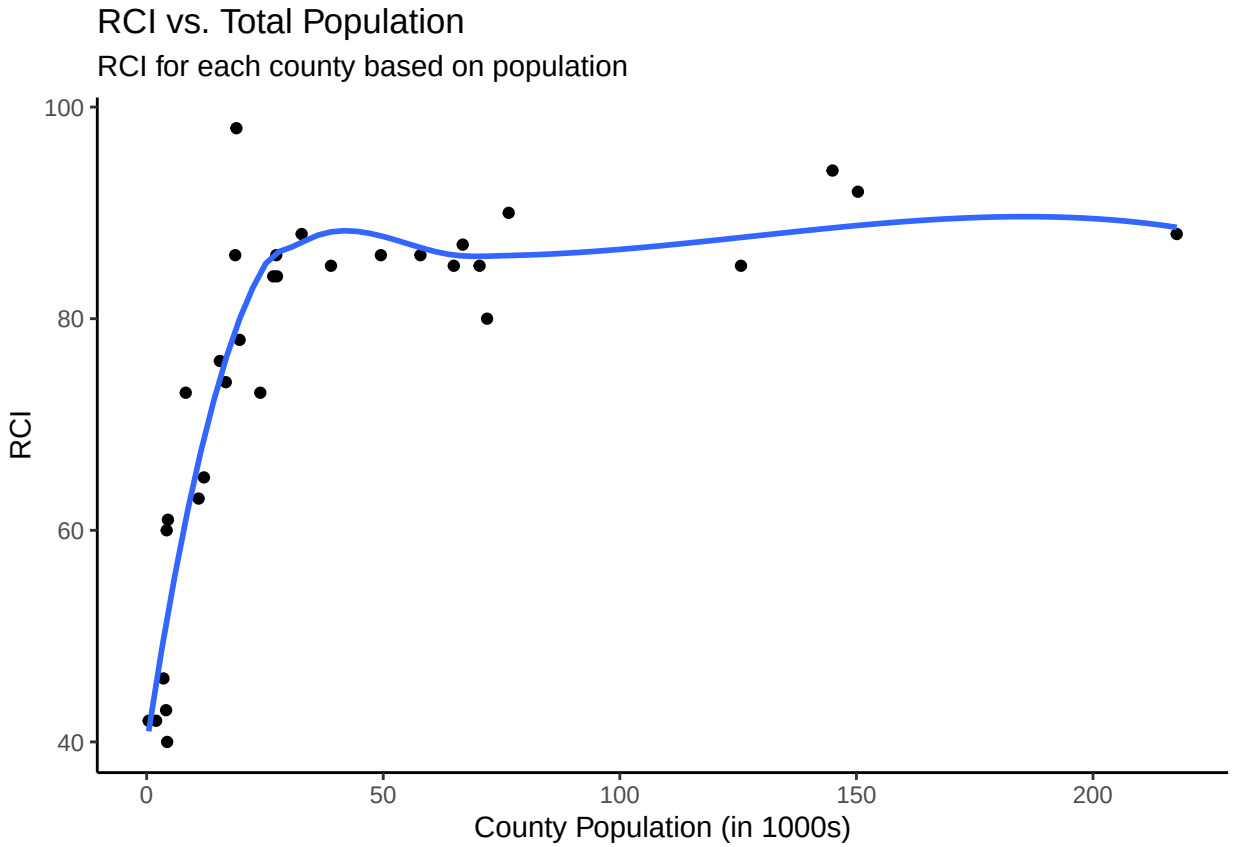
```
library(tidyverse)
library(tidyquant)
ixic <- read.csv("/Users/dz/Documents/MSSP/MA615/MA615-R/IXIC21-22.csv",header=TRUE)
ixic %>%
    ggplot(aes(x = as.Date(Date), y = Adj.Close)) +
    labs(y= "Adjusted Closing Price", x = "Date") +
    geom_candlestick(aes(open = Open, high = High,low= Low,close=Close)) +
    scale_x_date(date_breaks = "2 months", date_labels = "%b%y") +
    labs(title = "IXIC Candlestick Chart", y = "Adjusted Closing Price") +
    theme_tq()
```



#Rural Capacity Index 1.

```
rci1 <- read.csv("/Users/dz/Documents/MSSP/MA615/MA615-R/ruralCapacityData.csv",
    header=TRUE)
rci2 <- rci1[-1,]
rci2$pop1000 <- rci2$pop_total/1000

rci2 %>%
ggplot(aes(pop1000, cap_index)) +
  labs(title = "RCI vs. Total Population",
       subtitle = "RCI for each county based on population",
       x = "County Population (in 1000s)", y = "RCI") +
  geom_point() +
  geom_smooth(se=FALSE) +
  theme_classic()
```

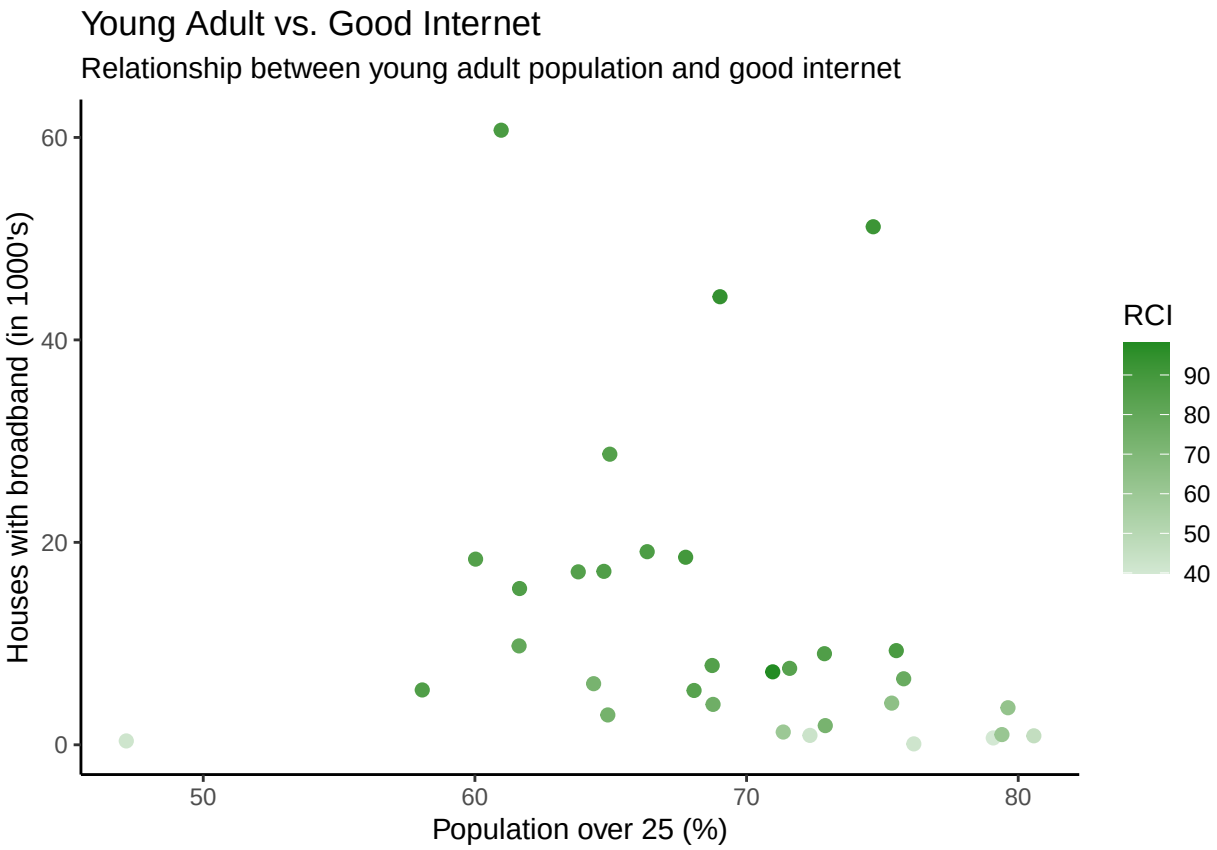


The plot above shows the RCI for every observation based on population.

#Rural Capacity Index 2.

```
rci2$per_over_25 <- rci2$pop_over_25/rci2$pop_total*100

rci2 %>%
ggplot(aes(per_over_25, house_broadband/1000)) +
  labs(title = "Young Adult vs. Good Internet",
       subtitle = "Relationship between young adult population and good internet",
       x = "Population over 25 (%)", y = "Houses with broadband (in 1000's)") +
  geom_point(aes(per_over_25,house_broadband/1000, color = cap_index),size=2) +
    scale_color_gradient(name = "RCI",low="#d3e8d3", high="#228B22") +
  theme_classic()
```



The above plot shows the relationship between houses with good internet and population in percentage of young adults over 25. RCI is shown as a color gradient; dark green represents high RCI and light green represents low RCI.

#Rural Capacity Index 3.

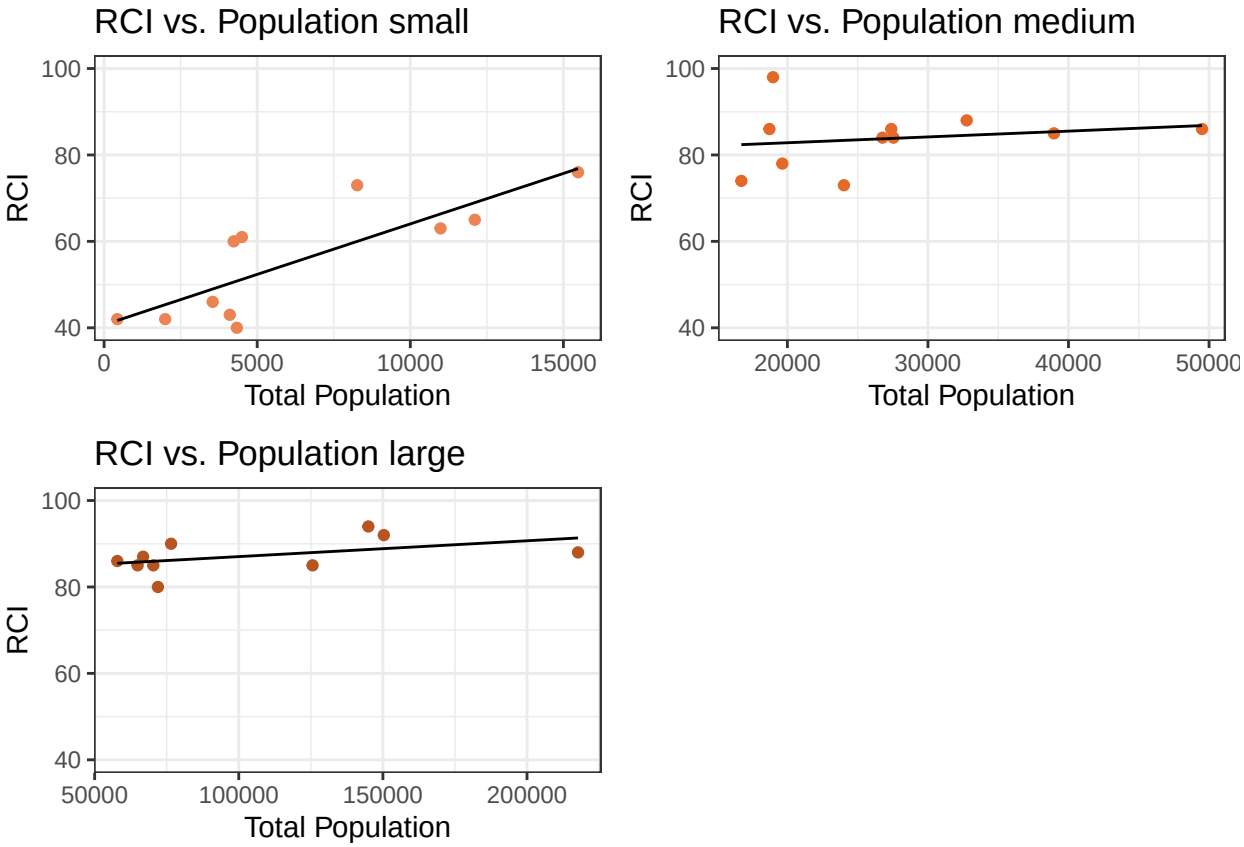
```
library(gridExtra)
par(mfrow=c(1,3))
plot1 <- ggplot(data=rci2[rci2$pop_total<16000,], aes(pop_total, cap_index)) +
  labs(title = "RCI vs. Population small",
       x = "Total Population", y = "RCI") +
  geom_point(color="#ea8553") +
  geom_smooth(method='lm', formula= y~x,color="black",se=FALSE,size=0.5)+
  ylim(40,100)+
  theme_bw()

plot2 <- ggplot(data=rci2[rci2$pop_total>16000 & rci2$pop_total<55000, ], aes(pop_total,
      cap_index)) +
  labs(title = "RCI vs. Population medium",
       x = "Total Population", y = "RCI") +
  geom_point(color="#e46828") +
  geom_smooth(method='lm', formula= y~x,color="black",se=FALSE,size=0.5)+
  ylim(40,100)+
  theme_bw()

plot3 <- ggplot(data=rci2[rci2$pop_total>55000, ], aes(pop_total, cap_index)) +
  labs(title = "RCI vs. Population large",
       x = "Total Population", y = "RCI") +
```

```
  geom_point(color="#b75420") +
  geom_smooth(method='lm', formula= y~x,color="black",se=FALSE,size=0.5)+
  ylim(40,100)+
  theme_bw()

grid.arrange(plot1,plot2,plot3,nrow=2, ncol=2)
```



The three graphs above show the relationship between RCI and population. The first graph is subsetted to a population sizes of less than 16000, the second to between 16000 and 55000, and the third to greater than 55000. It seems that as population increases, the correlation between population size and RCI become less strong. This could perhaps be a demonstration of something similar to the law of diminishing returns in economics.