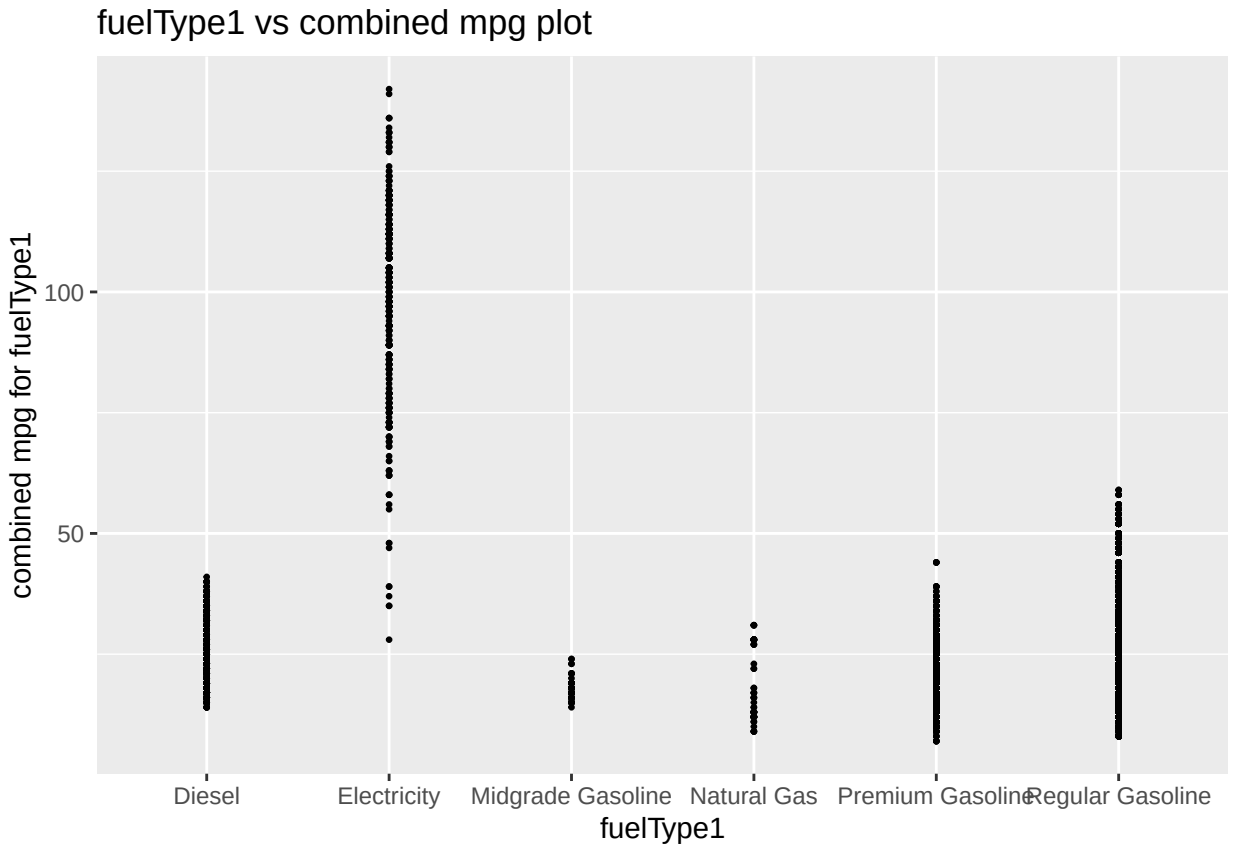
problemSet_2.R

2022-09-26

myName <- "JingJianGao"

#Problem 1a-Analysis on Fuel Economy

```
library(ggplot2)
vehicles <- read.csv("/Users/billg/Documents/vehicles.csv")
Regvehicle <- glm(comb08~fuelType1,data=vehicles)
ggplot <- ggplot(vehicles)
ggplot+aes(fuelType1,comb08)+
  geom_point(size=0.5)+xlab("fuelType1")+ylab("combined mpg for fuelType1")+
  ggtitle("fuelType1 vs combined mpg plot")
```

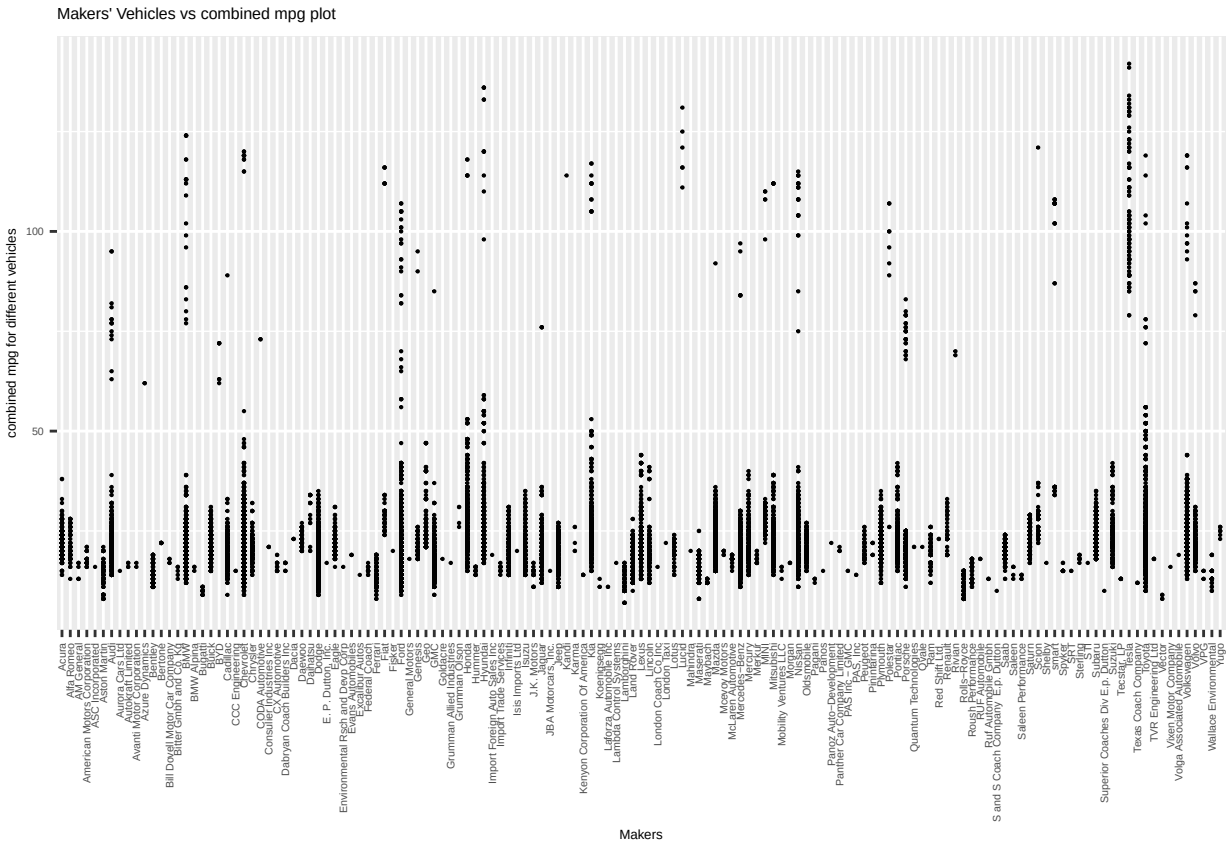



```
#This plot is showing the variety of fuels in fuelType1 and their relationship with combined mpg.
#We can see the difference between different kinds of fuel and their MPG performance.
```

##Problem #1b-Analysis on Vehicle makers

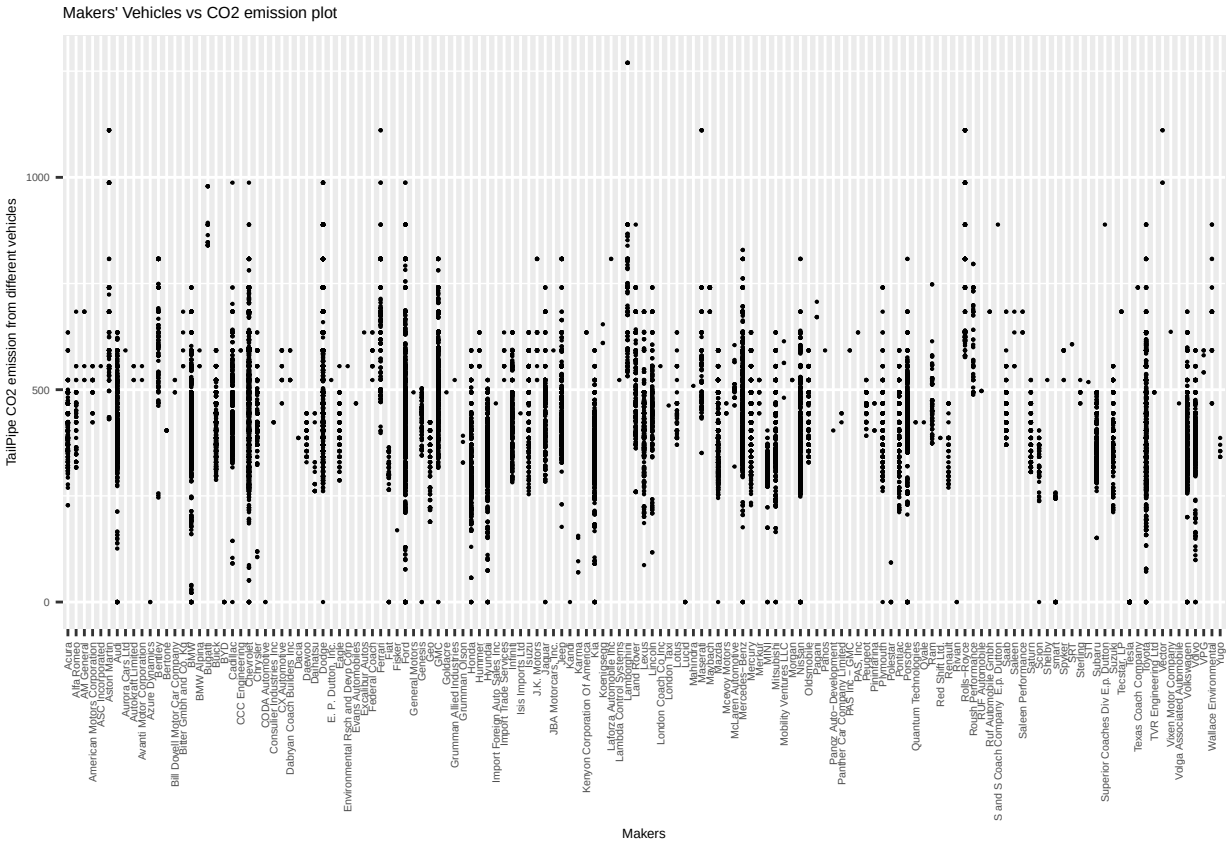
```
Regmaker <- glm(comb08~make,data=vehicles)
ggplot+aes(make,comb08)+geom_point(size=0.1)+xlab("Makers")+
  ylab("combined mpg for different vehicles")+
```

```
  ggtitle("Makers' Vehicles vs combined mpg plot")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  theme(text = element_text(size = 5))
```




```
#This plot is about the relationship between combined miles per gallon and the
#variety of vehicles made from different makers.
#We can see the difference in performance of mpg among makers to know who made the most progress.


Regmaker2 <- glm(co2TailpipeGpm~make,data=vehicles)
ggplot+aes(make,co2TailpipeGpm)+geom_point(size=0.1)+xlab("Makers")+
  ylab("TailPipe CO2 emission from different vehicles")+
  ggtitle("Makers' Vehicles vs CO2 emission plot")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  theme(text = element_text(size = 5))
```




```
#This plot is about the relationship between CO2 emission and
#the variety of vehicles made from different makers.
#We can see how much CO2 is produced by cars from certain makers to know how much progress
#they made on reducing CO2 emission.
```

##Problem #2-Nasdaq

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidyquant)
```

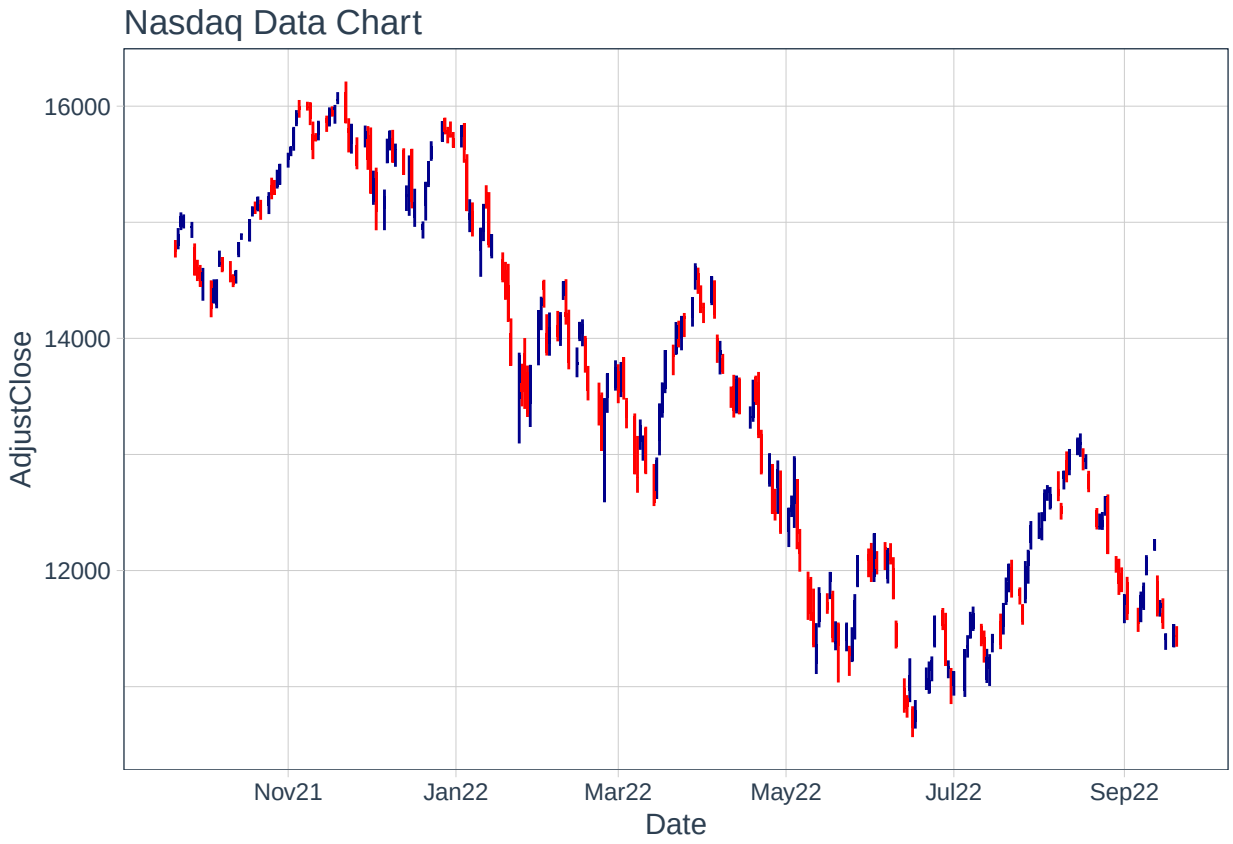
```
## Loading required package: lubridate
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
##
## Loading required package: PerformanceAnalytics
```

```
## Loading required package: xts
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
##
## Attaching package: 'xts'
##
## The following objects are masked from 'package:dplyr':
##
##     first, last
##
##
## Attaching package: 'PerformanceAnalytics'
##
## The following object is masked from 'package:graphics':
##
##     legend
##
## Loading required package: quantmod
## Loading required package: TTR
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

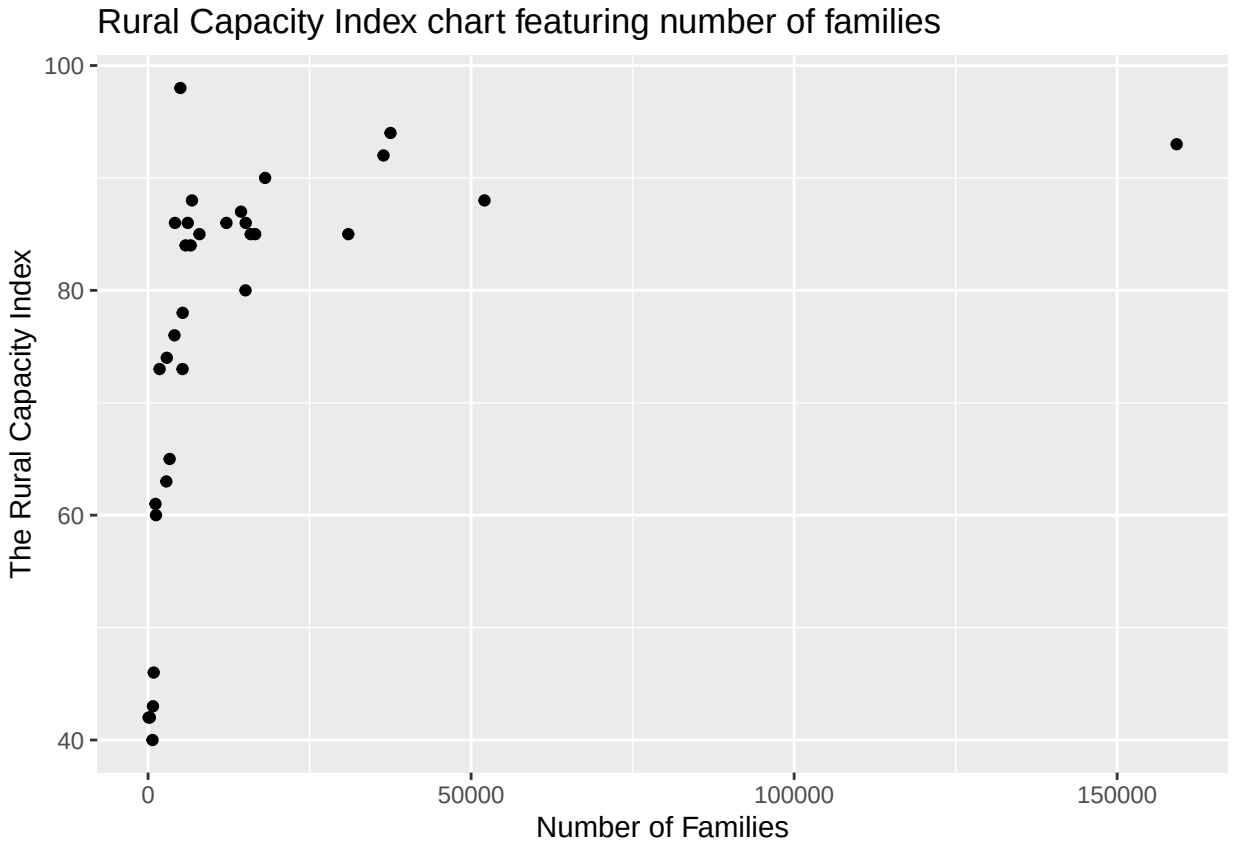
```
nasdaq <- read.csv("/Users/billg/Downloads/IXIC21-22.csv")
ggplot <- ggplot(nasdaq)
ggplot+aes(x=as.Date(Date),y= Adj.Close)+geom_candlestick(aes(open=Open,high=High,low=Low,close=Close))+
  labs(title="Nasdaq Data Chart",y="AdjustClose",x="Date")+
  theme_tq()+scale_x_date(date_breaks = "2 months", date_labels = "%b%y")
```




```
#This candlestick plot shows the Nasdaq Stock Market's open, high, low, and close price
#from 2021-9-21 to 2022-9-20
```

##Problem #3a-Rural capacity index

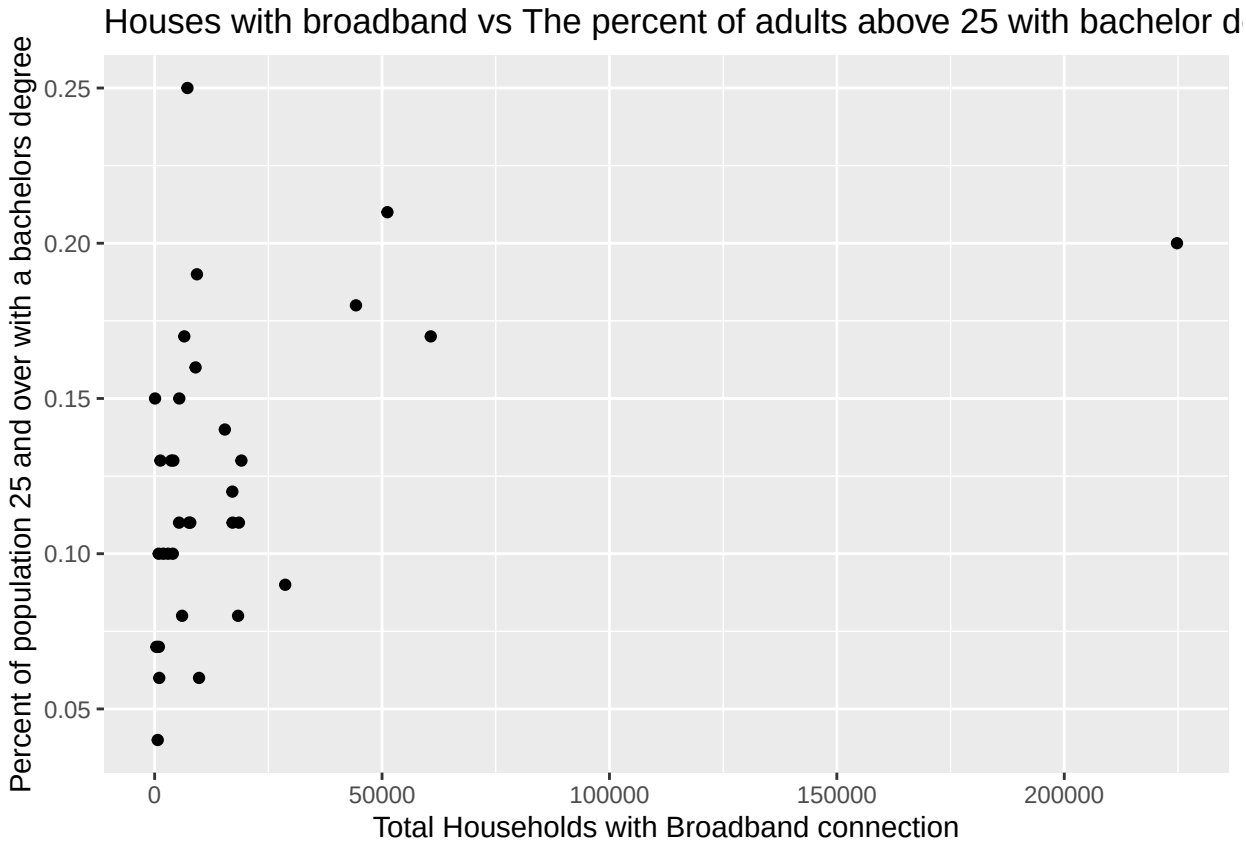
```
ruralcapacity <- read.csv("/Users/billg/Downloads/ruralCapacityData.csv")
ggplot <- ggplot(ruralcapacity)
ggplot+(aes(num_fam,cap_index))+geom_point()+xlab("Number of Families")+
  ylab("The Rural Capacity Index")+
  ggtitle("Rural Capacity Index chart featuring number of families")
```




```
#The number of families doesn't affect the Rural Capacity Index a lot.
#The number of families has few contributions to the Index
```

##Problem #3b

```
ggplot+(aes(house_broadband,per_over_25_with_bach))+geom_point()+
  xlab("Total Households with Broadband connection")+
  ylab("Percent of population 25 and over with a bachelors degree")+
  ggtitle("Houses with broadband vs The percent of adults above 25 with bachelor degrees")
```

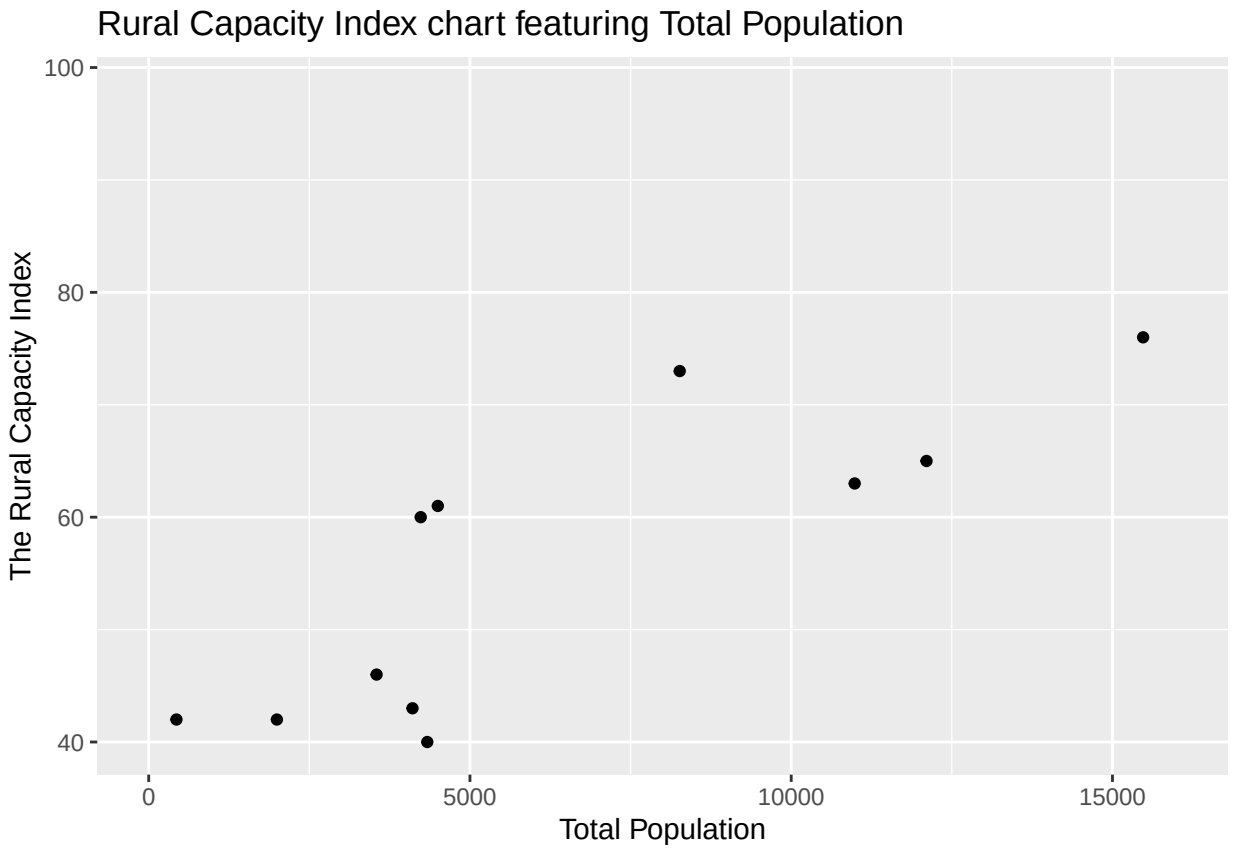



```
#This plot shows that there is not a strong correlation between
#the total households with Broadband connection and the percentage of population.
#But together, they have good impacts on the Capacity Index.
```

##Problem #3c

```
ggplot+(aes(pop_total,cap_index))+geom_point()+xlab("Total Population")+
  ylab("The Rural Capacity Index")+
  ggtitle("Rural Capacity Index chart featuring Total Population")+xlim(0,16000)
```

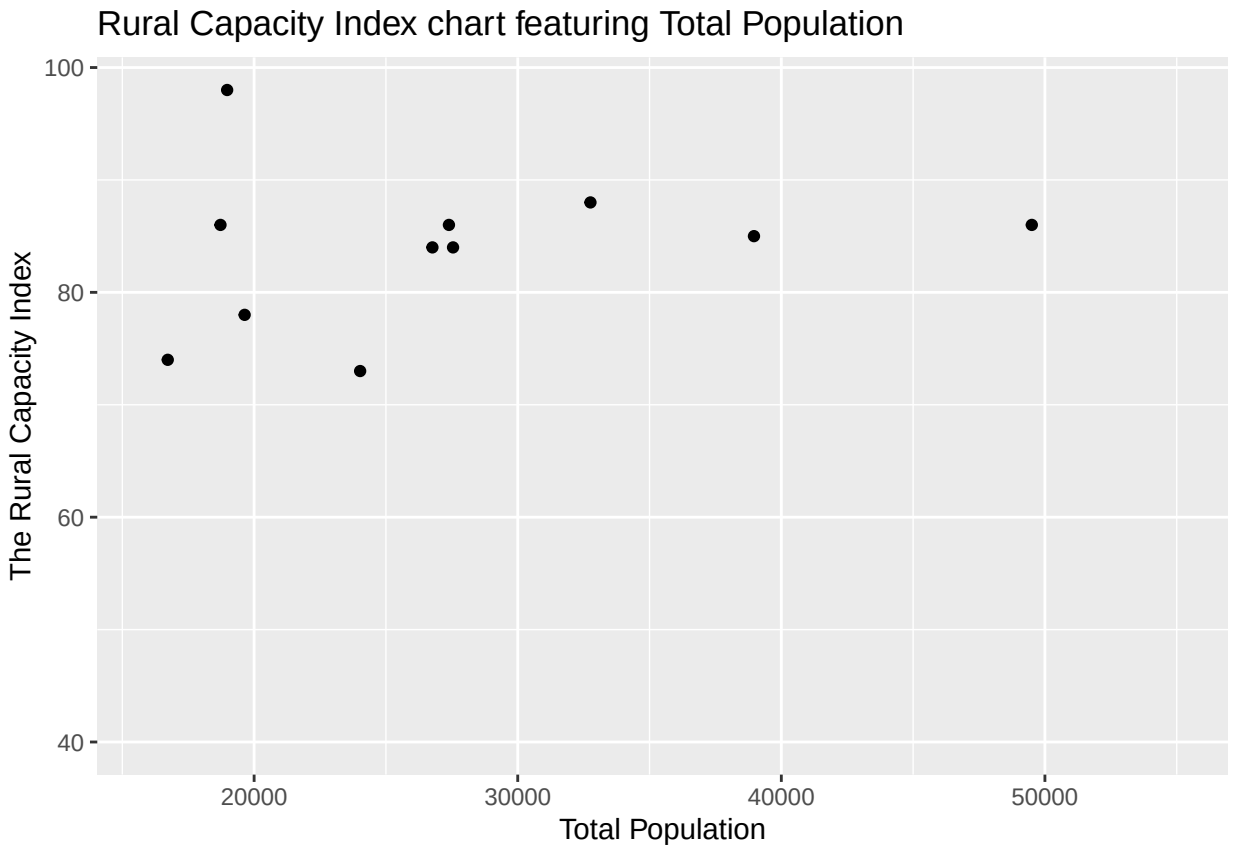
```
## Warning: Removed 22 rows containing missing values (geom_point).
```




```
#This plot shows the relationship between total population and Capacity index for
#total population <16000

ggplot+(aes(pop_total,cap_index))+geom_point()+xlab("Total Population")+
  ylab("The Rural Capacity Index")+
  ggtitle("Rural Capacity Index chart featuring Total Population")+xlim(16000,55000)
```

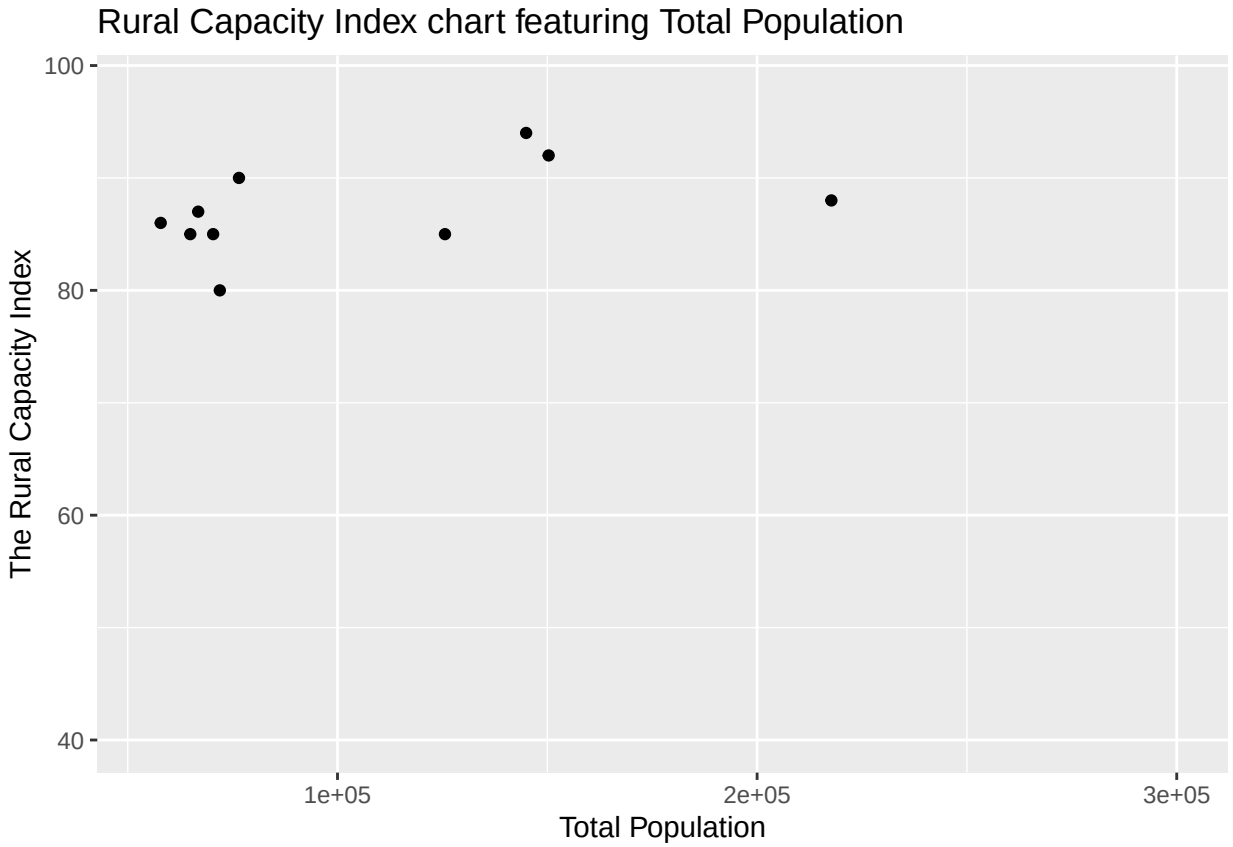
```
## Warning: Removed 22 rows containing missing values (geom_point).
```




```
#This plot shows the relationship between total population and Capacity index for
#16000< total population <55000


ggplot+(aes(pop_total,cap_index))+geom_point()+xlab("Total Population")+
  ylab("The Rural Capacity Index")+
  ggtitle("Rural Capacity Index chart featuring Total Population")+xlim(55000,300000)
```

```
## Warning: Removed 23 rows containing missing values (geom_point).
```




```
#This plot shows the relationship between total population and Capacity index for
#total population >55000
```

#The Rural capacity Index seems to have a linear relationship with total population in graph 1

#The Rural capacity Index seems to maintain stable once the total population is #greater than 16000 as shown in graph 2

#The Rural capacity Index is greater than 80 for all total populations greater than 55000