INFM600

Viral Shah, Phaneendra Kumar

March 21, 2016

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

##setting directory  
setwd("C:/Viral/Courses/INFM 600/Information Discovery/2016\_presidential\_election")  
getwd()

## [1] "C:/Viral/Courses/INFM 600/Information Discovery/2016\_presidential\_election"

##reading datasets  
countyfacts = read.csv("county\_facts.csv",na.strings = c("NA","","NULL"))  
results = read.csv("primary\_results.csv")  
countyfacts$state\_abbreviation = factor(countyfacts$state\_abbreviation)  
results$county=paste(results$county,"County")  
  
##calculating the population of educated people of each county  
countyfacts$educatedpeople = countyfacts$PST045214\*countyfacts$EDU635213  
  
##subsetting the states to only include Iowa, New Hampshire, Nevada, South Carolina  
fourstates = subset(countyfacts, state\_abbreviation %in% c("IA","NH","NV","SC"))  
  
##calculating the total votes per county  
votecount=aggregate(votes ~ county, FUN=sum,data=results)  
colnames(votecount) <- c("area\_name","votes")  
  
##merging fourstates and votecount dataframe  
combineresult=merge(x=fourstates,y=votecount)  
  
## descriptive statistics of both variables  
summary(combineresult$votes)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 95 1248 3088 12480 11930 148700

summary(combineresult$educatedpeople)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 68470 1010000 1874000 6165000 4529000 173600000

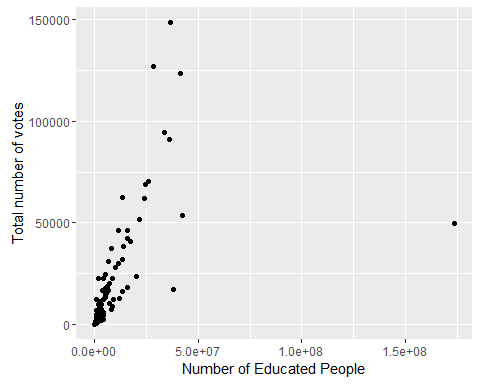
##Correlation test and writing the file to the system  
cor.test(combineresult$educatedpeople,combineresult$votes)

##   
## Pearson's product-moment correlation  
##   
## data: combineresult$educatedpeople and combineresult$votes  
## t = 8.9888, df = 168, p-value = 4.441e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4587033 0.6634779  
## sample estimates:  
## cor   
## 0.5698728

write.csv(combineresult,"combineresult.csv")  
  
##plotting the correlation  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.2.4

ggplot(combineresult,aes(educatedpeople,votes))+geom\_point()+labs(x="Number of Educated People",y="Total number of votes")



##plotting the correlation statewise  
p <- ggplot(combineresult,aes(educatedpeople,votes))+geom\_point()+labs(x="Number of Educated People",y="Total number of votes")  
p+facet\_grid(state\_abbreviation ~ .)

