finalreport704193973

Vergil

November 22, 2015

```
library(knitr)
library(XML)
library(ggplot2)
```

1.

```
GroupEast <- read.csv("~/Desktop/GroupEast.csv", stringsAsFactors=FALSE)
dim(GroupEast)
```

```
## [1] 13548    41
```

```
irsCAzip2013 <- read.csv("~/Desktop/irsCAzip2013.csv", stringsAsFactors=FALSE)
dim(irsCAzip2013)
```

```
## [1] 1484  111
```

2.

A.

```
GroupEast1 <- GroupEast[,-which(names(GroupEast) %in% c("Suite.Number", "Neighborhood",
"Cross.Street", "Link.to.Map", "Link.to.Menu", "Hours", "Founded", "Private.Room", "Lin
k.to.Yelp", "Link.to.Yahoo..Local", "Link.to.Citysearch", "Link.to.Zagat"))]
dim(GroupEast1)
```

```
## [1] 13548    29
```

B.

```
GroupEast1 <- GroupEast1[-which(GroupEast1$Longitude %in% NA),]
GroupEast1$Address <- paste(GroupEast1$Street.Address,GroupEast1$City,GroupEast1$State,s
ep = ",")
zero <- which(GroupEast1$Longitude %in% 0)
locg <- read.csv("~/Documents/Academics/2015 Fall/Stats 20/Final Project/locg.csv", stri
ngsAsFactors=FALSE)
#locg <- geocode(GroupEast1[zero,]$Address, output="latlona", messaging=FALSE, source="g
oogle")
GroupEast1[zero,]$Longitude <- locg$lon
GroupEast1[zero,]$Latitude <- locg$lat
```

C.

```
a <- which(names(GroupEast1) %in% c("Phone.Number", "Email", "Alcohol", "Credit.Cards",
"Good.for.Kids", "Childrens.Menu", "Takeout", "Delivery", "Kosher", "Halal", "Vegan.Vege
tarian", "Gluten.Free.Options", "Organic.Options", "Wheelchair.Access", "Price", "Chef",
 "Reservations"))

for (i in a) {
  GroupEast1[,i][which(GroupEast1[,i]==" ")] <- NA
}
```

D.

```
GroupEast1$Ratings_new <- gsub("/ 5","",GroupEast1$Ratings)
GroupEast1$Ratings_new <- as.numeric(GroupEast1$Ratings_new)
```

E.

```
names(GroupEast1)[5] <- "Zip"
GroupEast1 <- GroupEast1[,31:1]
```

3

A.

```
kable(as.data.frame(table(GroupEast1$Ratings)))
```

Var1	Freq
	3976
1 / 5	432
1.5 / 5	106
2 / 5	424
2.5 / 5	408
3 / 5	1843
3.5 / 5	1296
4 / 5	1717
4.5 / 5	1476
5 / 5	1858

```
mean(GroupEast1$Ratings_new,na.rm = T)
```

```
## [1] 3.694979
```

B.

```
kable(as.data.frame(table(GroupEast1$Alcohol)))
```

Var1	Freq
beer & wine	1549
full bar	975
no	1926

```
kable(as.data.frame(table(GroupEast1$Credit.Cards)))
```

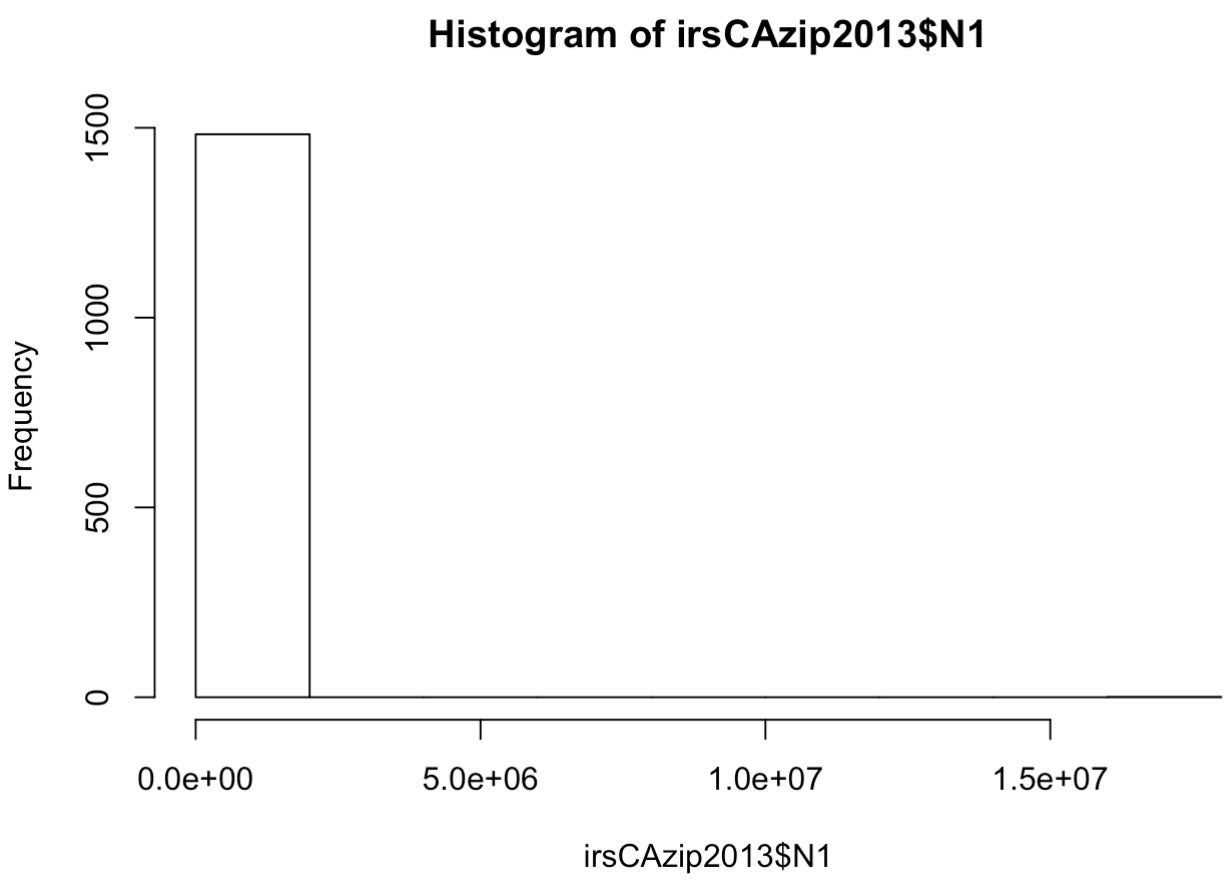
Var1	Freq
no	710
yes	5405

```
kable(as.data.frame(table(GroupEast1$Good.for.Kids)))
```

Var1	Freq
no	757
yes	4811

C.

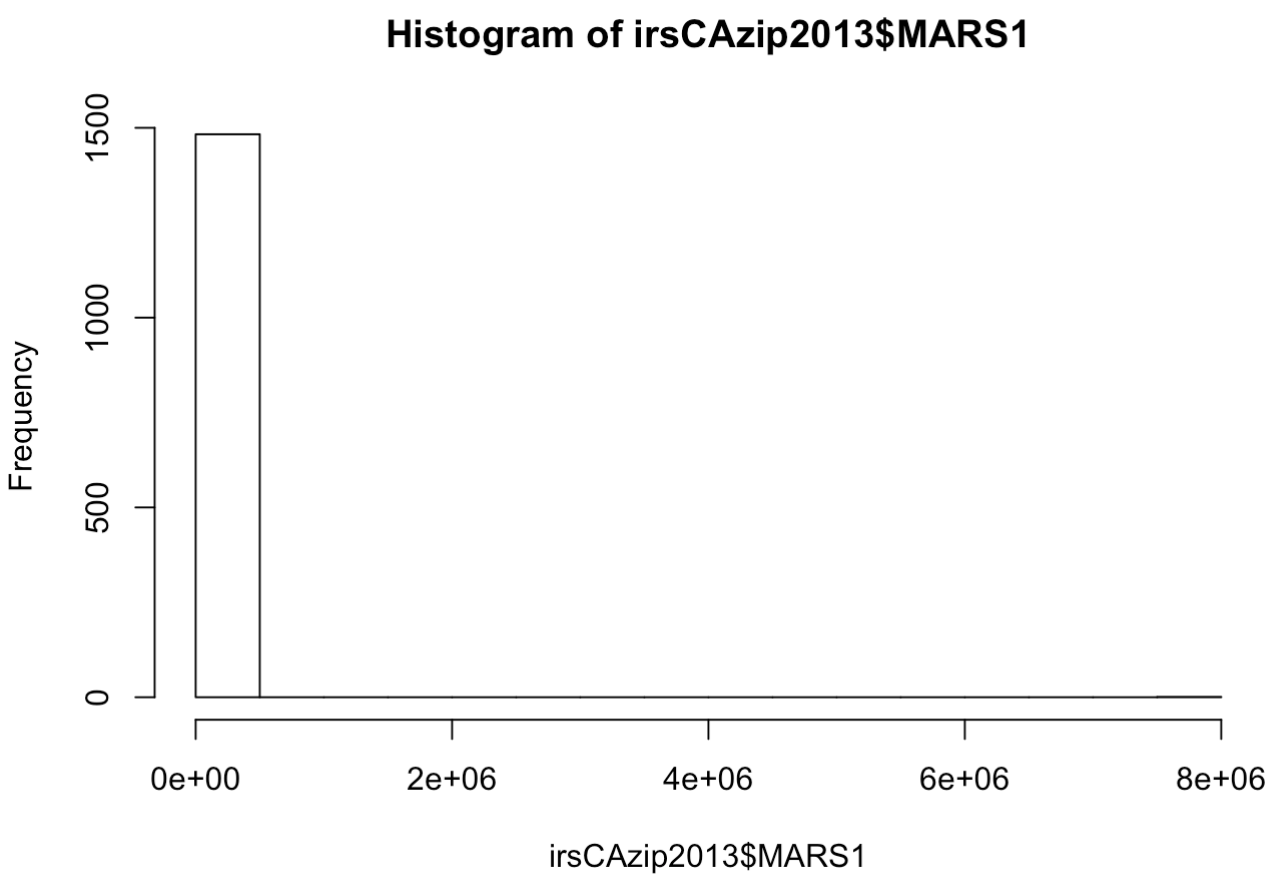
```
hist(irsCAzip2013$N1)
```




```
median(irsCAzip2013$N1,na.rm = T)
```

```
## [1] 10200
```

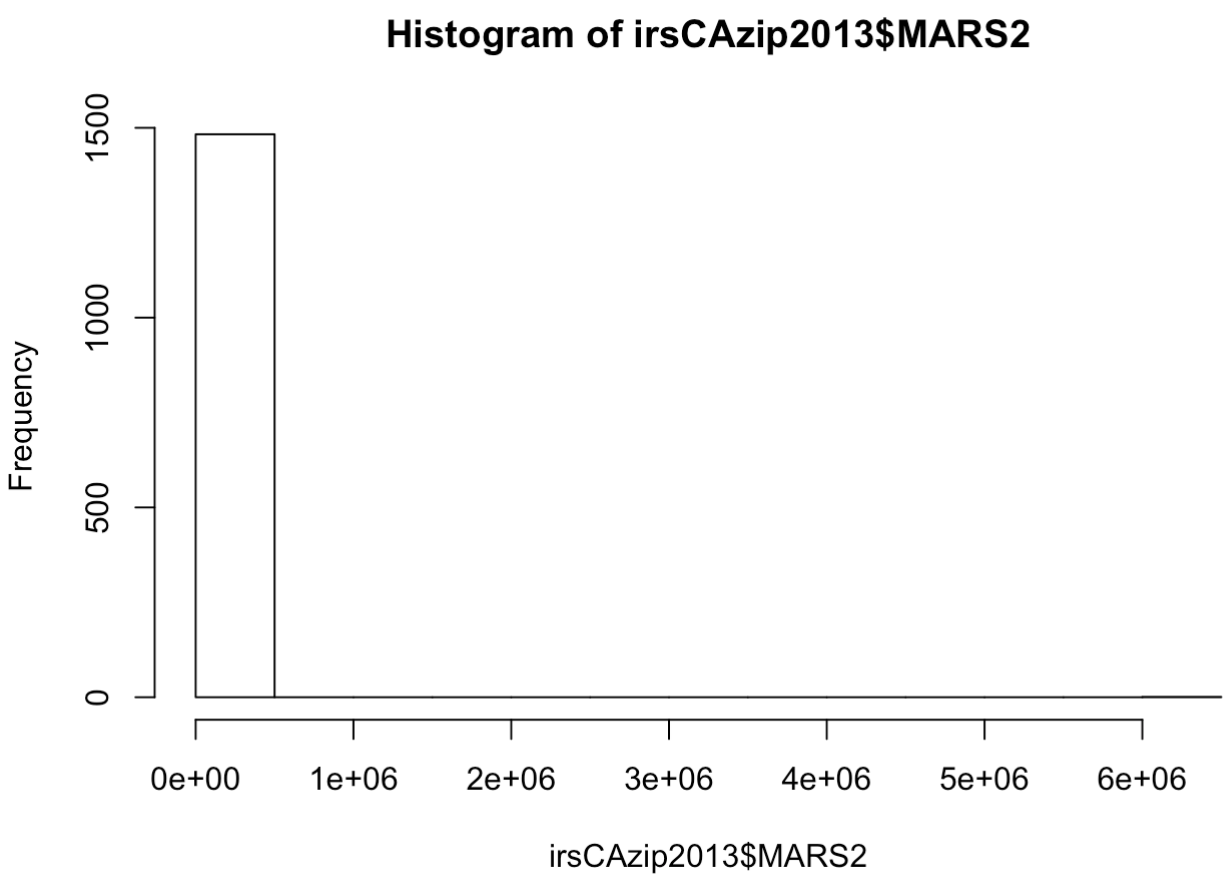
```
hist(irsCAzip2013$MARS1)
```




```
median(irsCAzip2013$MARS1,na.rm = T)
```

```
## [1] 4725
```

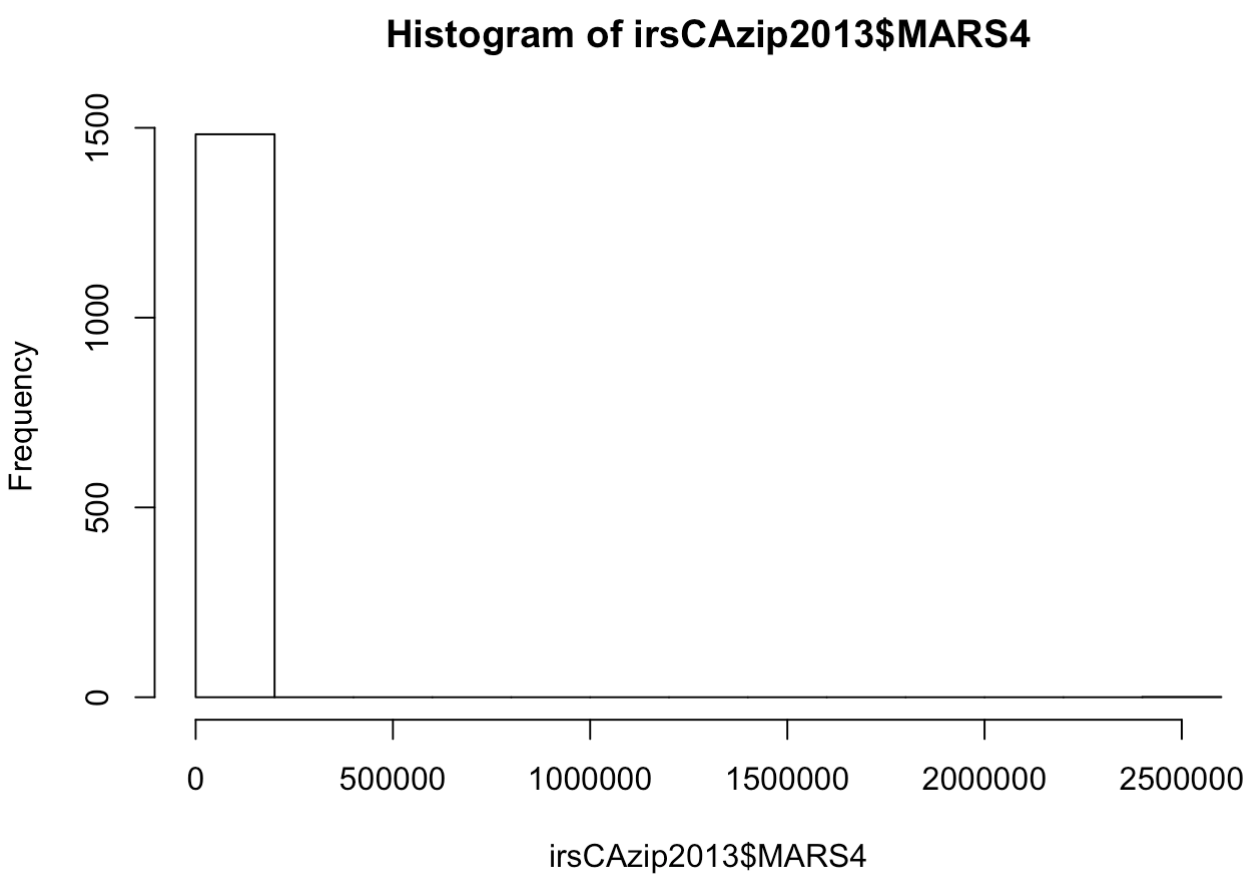
```
hist(irsCAzip2013$MARS2)
```




```
median(irsCAzip2013$MARS2,na.rm = T)
```

```
## [1] 3555
```

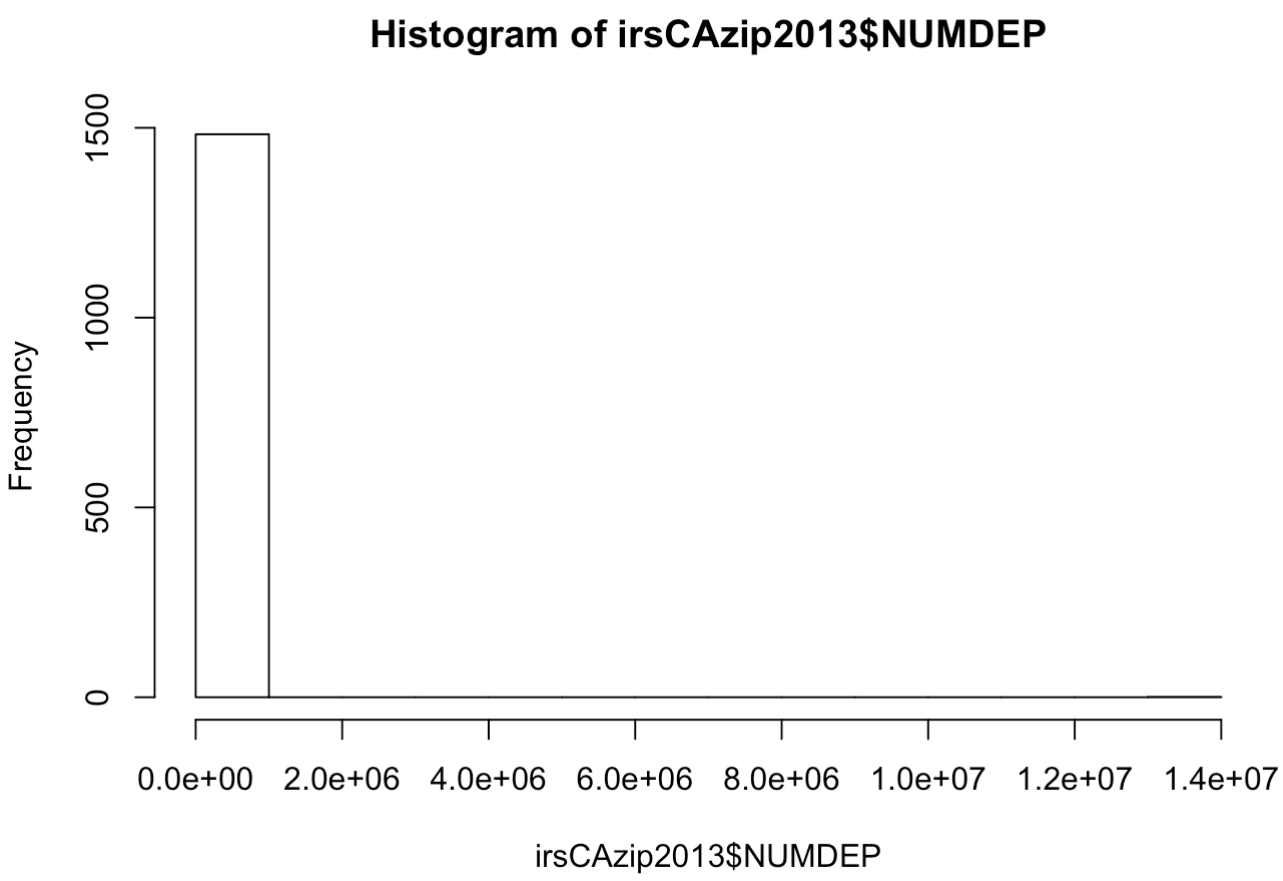
```
hist(irsCAzip2013$MARS4)
```




```
median(irsCAzip2013$MARS4,na.rm = T)
```

```
## [1] 1080
```

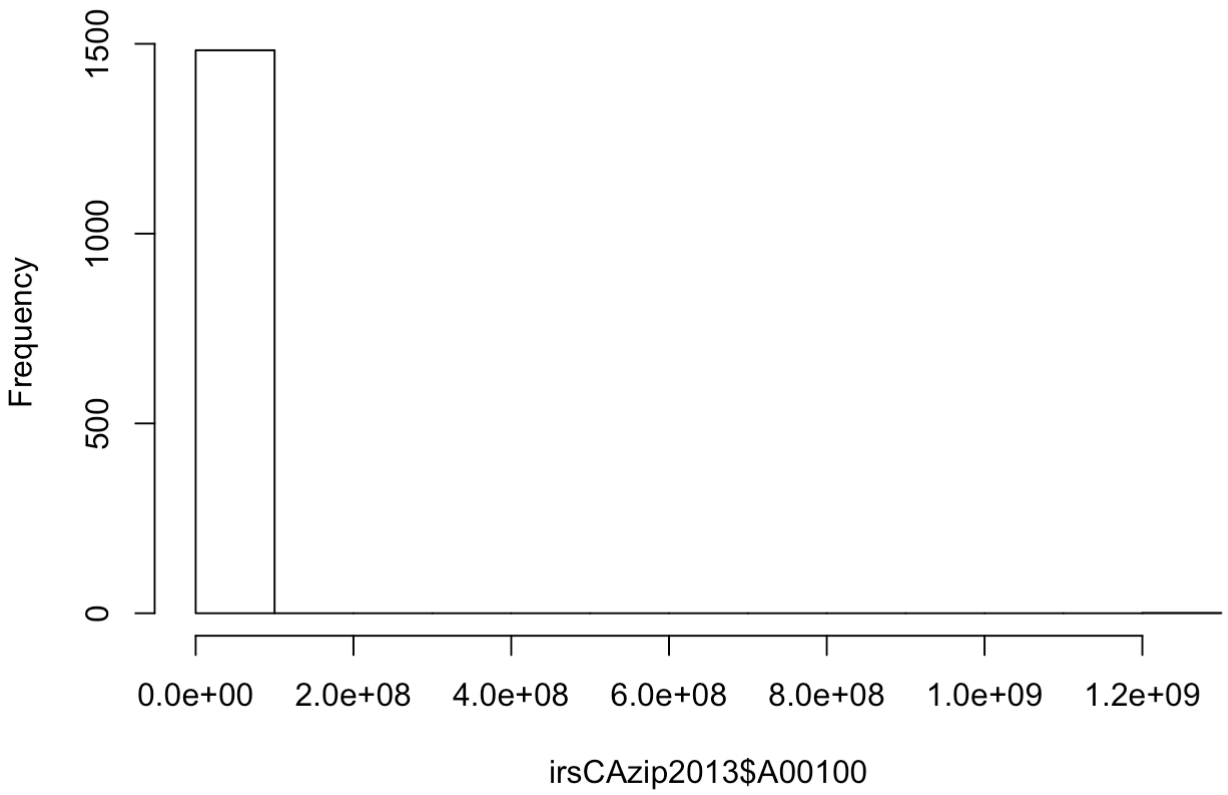
```
hist(irsCAzip2013$NUMDEP)
```




```
median(irsCAzip2013$NUMDEP,na.rm = T)
```

```
## [1] 6400
```

```
hist(irsCAzip2013$A00100)
```




```
median(irsCAzip2013$A00100,na.rm = T)
```

```
## [1] 603562
```

4.

```
names(irsCAzip2013)[1] <- "Zip"
GroupEast1_irs <- merge(GroupEast1,irsCAzip2013[,which(names(irsCAzip2013) %in%
c("N1","MARS1","MARS2", "MARS4", "NUMDEP", "A00100","Zip"))],by="Zip")
dim(GroupEast1_irs)
```

```
## [1] 13439    37
```

5.

A.

```
mcdonaldsG <- readHTMLTable("http://www.stat.ucla.edu/~vlew/datasets/mcdonaldsG.html")
mcdonaldsG <- as.data.frame(mcdonaldsG)
names(mcdonaldsG) <-
c("store","Address","City","State","zipcode","Phone","longtitude","latitude","INWALMART","
LAYPLACE")
dim(mcdonaldsG)
```

```
## [1] 14044    10
```

B.

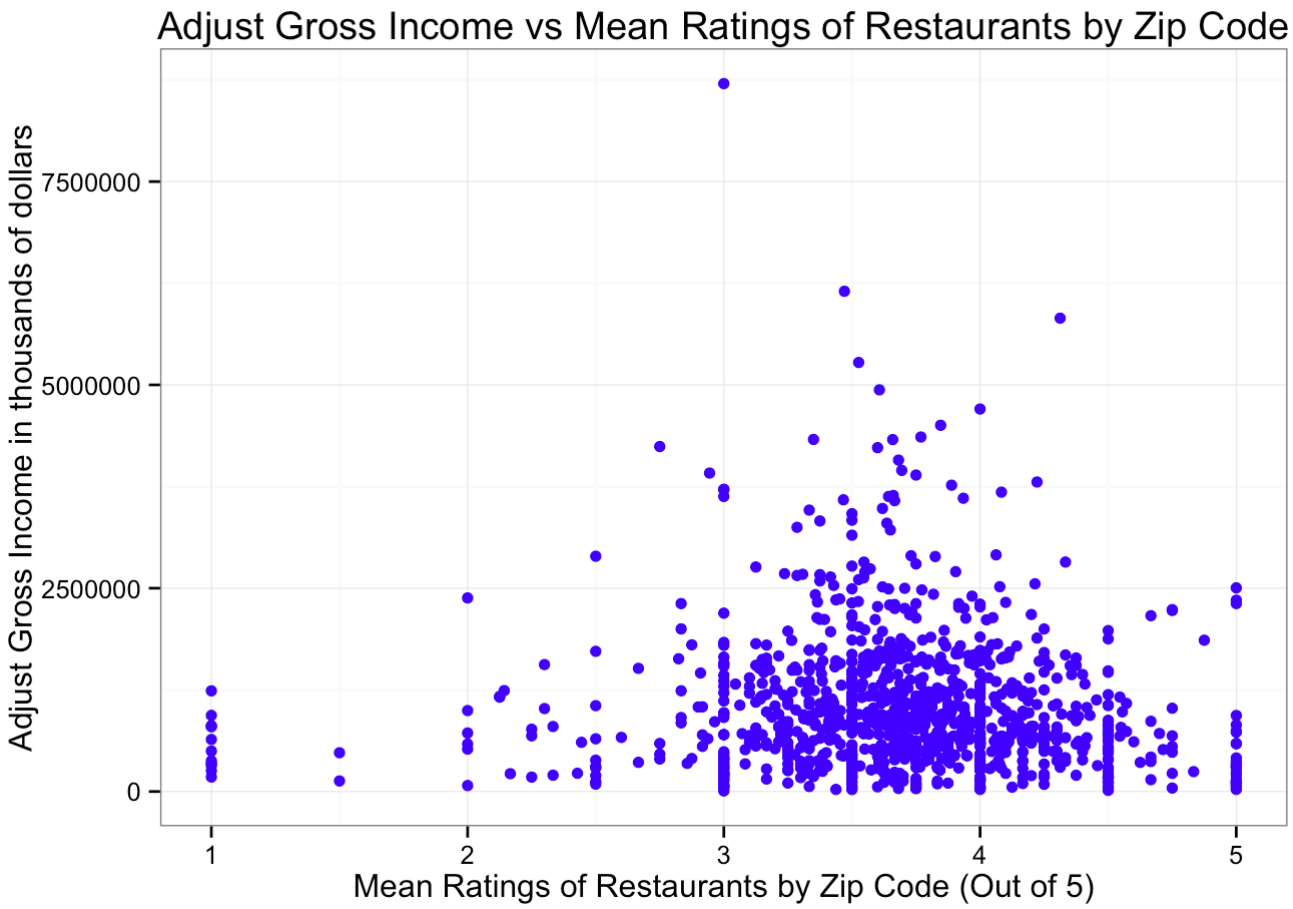
```
names(mcdonaldsG)[5] <- "Zip"
irs_mcdonaldsG <- merge(irsCAzip2013[,which(names(irsCAzip2013) %in% c("N1","MARS1","MAR
S2", "MARS4", "NUMDEP", "A00100","Zip"))],mcdonaldsG,by="Zip")
d <- merge(mcdonaldsG,GroupEast1,by="Zip")
```

A way to fix d?

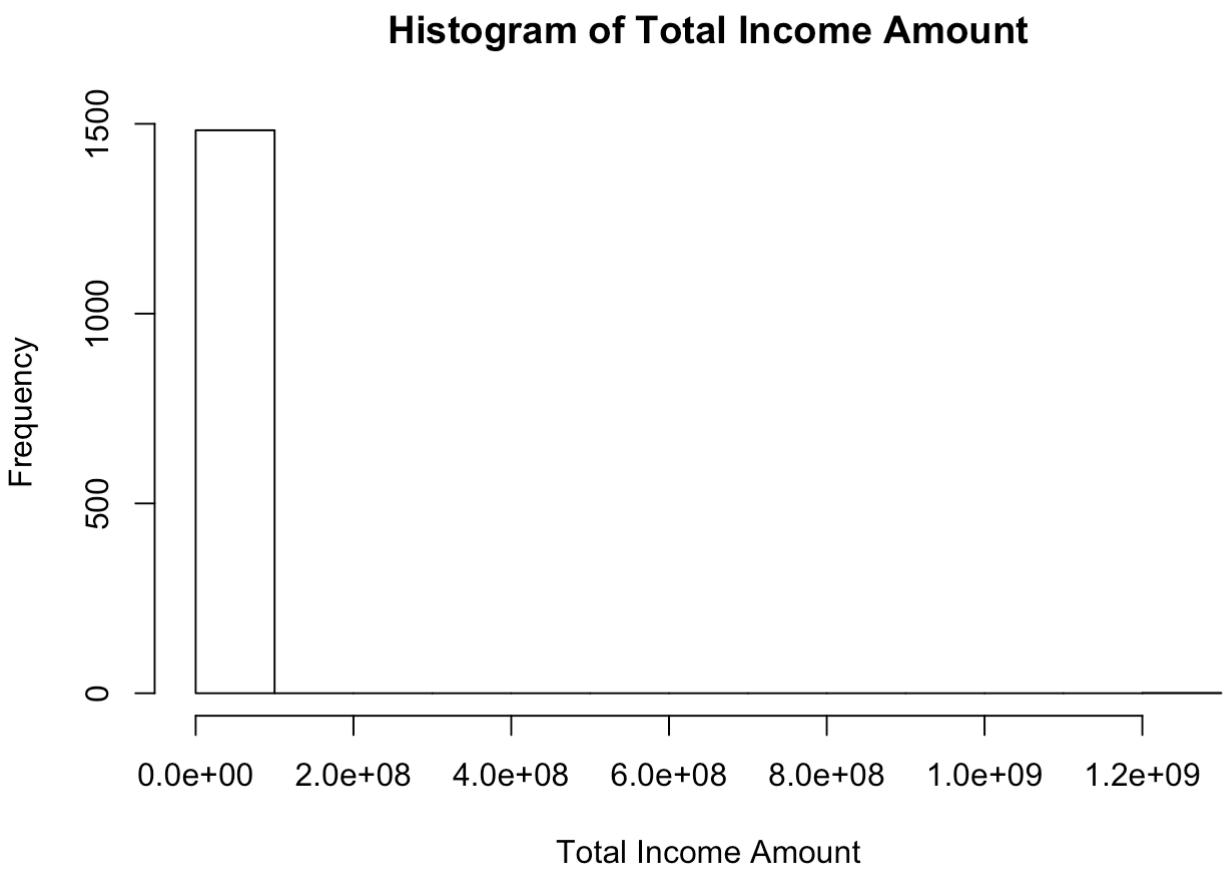
6.

```
# ggplot(irsCAzip2013_new,aes(x=A00100,y=A10300)) + geom_point()
# ggplot(GroupEast1_irs,aes(x=factor(Ratings_new),y=A00100))+geom_boxplot()
e <- as.data.frame(tapply(GroupEast1_irs$Ratings_new,GroupEast1_irs$Zip,mean,na.rm=T))
e$Zip <- row.names(e)
A00100_Ratings <- merge(irsCAzip2013[,c("A00100","Zip")],e,by="Zip")
names(A00100_Ratings)[3] <- "Ratings_mean"
ggplot(A00100_Ratings,aes(x=Ratings_mean,y=A00100))+geom_point(color="blue")+labs(title="A
just Gross Income vs Mean Ratings of Restaurants by Zip Code", x="Mean Ratings of Restau
rants by Zip Code (Out of 5)",y="Adjust Gross Income in thousands of
dollars")+theme_bw()
```

```
## Warning: Removed 39 rows containing missing values (geom_point).
```




```
hist(irsCAzip2013$A02650,main = "Histogram of Total Income Amount",xlab="Total Income Am
ount")
```



make variables continuous, then plot or linear regression mean of ratings by zipcode etc.

7.

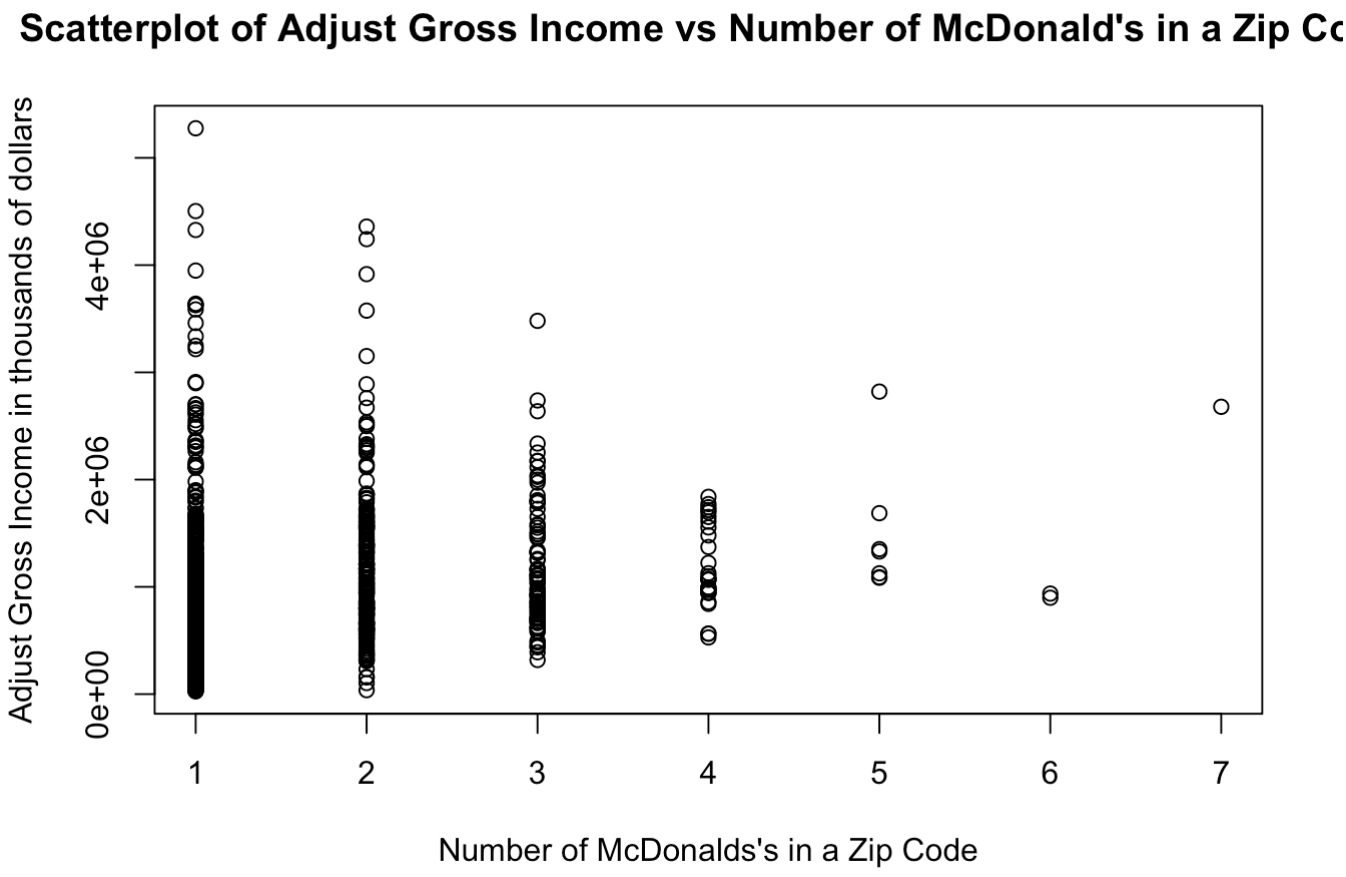
```
irsCAzip2013$A00100_sqrt <- sqrt(irsCAzip2013$A00100)
```

8.

```
g <- as.data.frame(table(mcdonaldsG$Zip))
names(g) <- c("Zip","Number_of_stores")
A00100_stores <- merge(g,irsCAzip2013[,c("A00100","Zip")],by="Zip")
lm1 <- lm(A00100~Number_of_stores,data=A00100_stores)
summary(lm1)
```

```
##
## Call:
## lm(formula = A00100 ~ Number_of_stores, data = A00100_stores)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1051324  -491752  -156127   319875  4321349
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        820490      52219  15.712  < 2e-16 ***
## Number_of_stores   133872      27573   4.855 1.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 720400 on 794 degrees of freedom
## Multiple R-squared:  0.02883,    Adjusted R-squared:  0.02761
## F-statistic: 23.57 on 1 and 794 DF,  p-value: 1.449e-06
```

```
plot(A00100~Number_of_stores,data = A00100_stores,main="Scatterplot of Adjust Gross Inco
me vs Number of McDonald's in a Zip Code",ylab = "Adjust Gross Income in thousands of do
llars",xlab="Number of McDonalds's in a Zip Code")
```



9.

```
# sort(tapply(GroupEast1$Ratings_new,GroupEast1$Zip,median))
```

Format:

1. Abstract: tell data set names+size. tell what you scrape and resulting size
2. Introduction: “graph this”,“regression on this”,“table this” (why did you choose these or what did choose to do?)

income level~number of mcdonalds in a zipcode

3. data + method: what did you do with the data? eg. did you drop a lot of observations

talk about geocode. and how many addresses were geocoded

4. result: table, graphs, regression,

log(AGI) for boxplot

airband around scatterplot

hide the R code in the result sections. Rmd file will have R code twice, html should have R code only in the last section

geocode can be written to a file that then be read