Lab 3, 732A98 Visualization

Jochen Schaefer, Jiawei Wu

2018/9/27

```
library(ggplot2)
library(readxl)
library(MASS)
library(plotly)
library(gridExtra)
library(maps)
library(akima)
library(sp)
library(sf)
library(stringr)
library(dplyr)
Sys.setenv('MAPBOX_TOKEN' = 'pk.eyJ1Ijoid3Vqd2dhcnkiLCJhIjoiY2ptOTlrY3cyMGFhaTN2cjFwNndkeHBqeiJ9._0kS0Ly5O5qJ-3Lj
YpnL6g')
```

Assignment 1

```
data1 <- read.csv("aegypti_albopictus.csv", header = TRUE)
str(data1)
```

```
## 'data.frame':    42066 obs. of  12 variables:
##  $ VECTOR       : Factor w/ 2 levels "Aedes aegypti",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ OCCURRENCE_ID: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ SOURCE_TYPE  : Factor w/ 2 levels "published","unpublished": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LOCATION_TYPE: Factor w/ 8 levels "Less than 10 km",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ POLYGON_ADMIN: Factor w/ 5 levels "-999","2","Less than 100km",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Y            : num  -3.22 -4.27 -4.27 -3.22 -3.04 ...
##  $ X            : num  40.1 15.3 15.3 40.1 40.1 ...
##  $ YEAR         : Factor w/ 57 levels "1958","1960",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ COUNTRY      : Factor w/ 151 levels "Afghanistan",..: 73 36 36 73 73 140 34 55 99 145 ...
##  $ COUNTRY_ID   : Factor w/ 150 levels "ABW","AFG","AGO",..: 70 32 32 70 70 136 34 50 96 140 ...
##  $ GAUL_AD0     : int  133 59 59 133 133 253 57 94 182 263 ...
##  $ STATUS       : Factor w/ 2 levels "E","T": NA NA NA NA NA NA NA NA NA NA ...
```

Task 1

```
dat1 <- data1[which(data1$YEAR==2004),]
p1 <-  plot_mapbox(dat1, x = ~X, y = ~Y) %>%
  add_trace(
    color = ~VECTOR,text = ~COUNTRY
  )
p1
```


Aedes aegypti mainly occurs in latin america, the south of the US and south-east asia. Mostly the coastal regions are affected. Aedes albopictus shows most instances in the middle-east of the US as well as some rare cases in Asia and Europe.

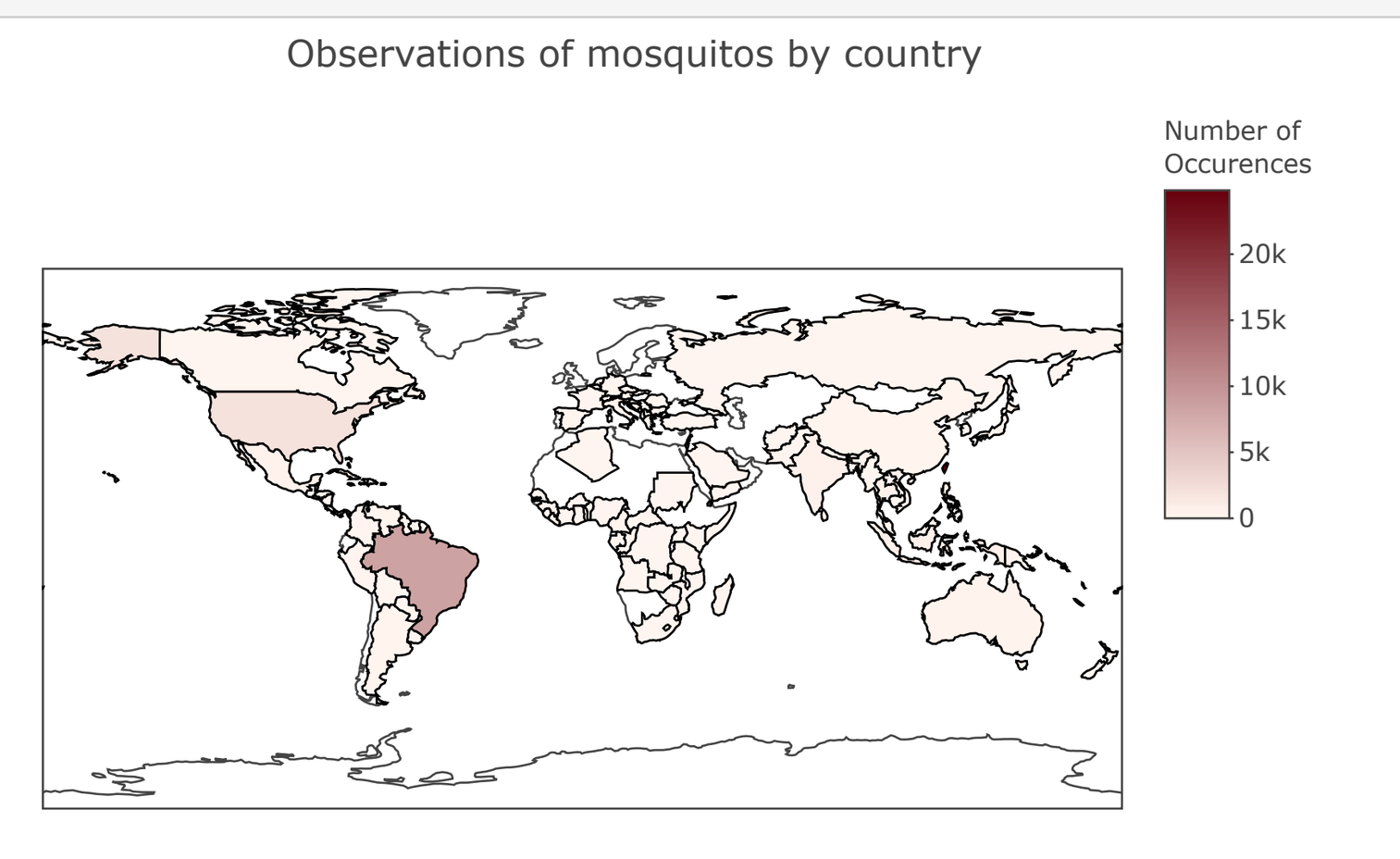
```
dat2 <- data1[which(data1$YEAR==2013),]
p2 <-  plot_mapbox(dat2, x = ~X, y = ~Y) %>%
  add_trace(
    color = ~VECTOR,text = ~COUNTRY
  )
p2
```


Compared to 2004, Aedes aegypti spread all over Brazil and parts of Peru, while the occurence in other parts of the world decreased drastically. Aedes albopictus is now present in Italy and Taiwan.

If looking at the world map, there is a problem of overplotting, because there are several observations located very close to one another, at least in the 2013 map. This is problematic, because one can not estimate the actual number of individual occurences.

Task 2

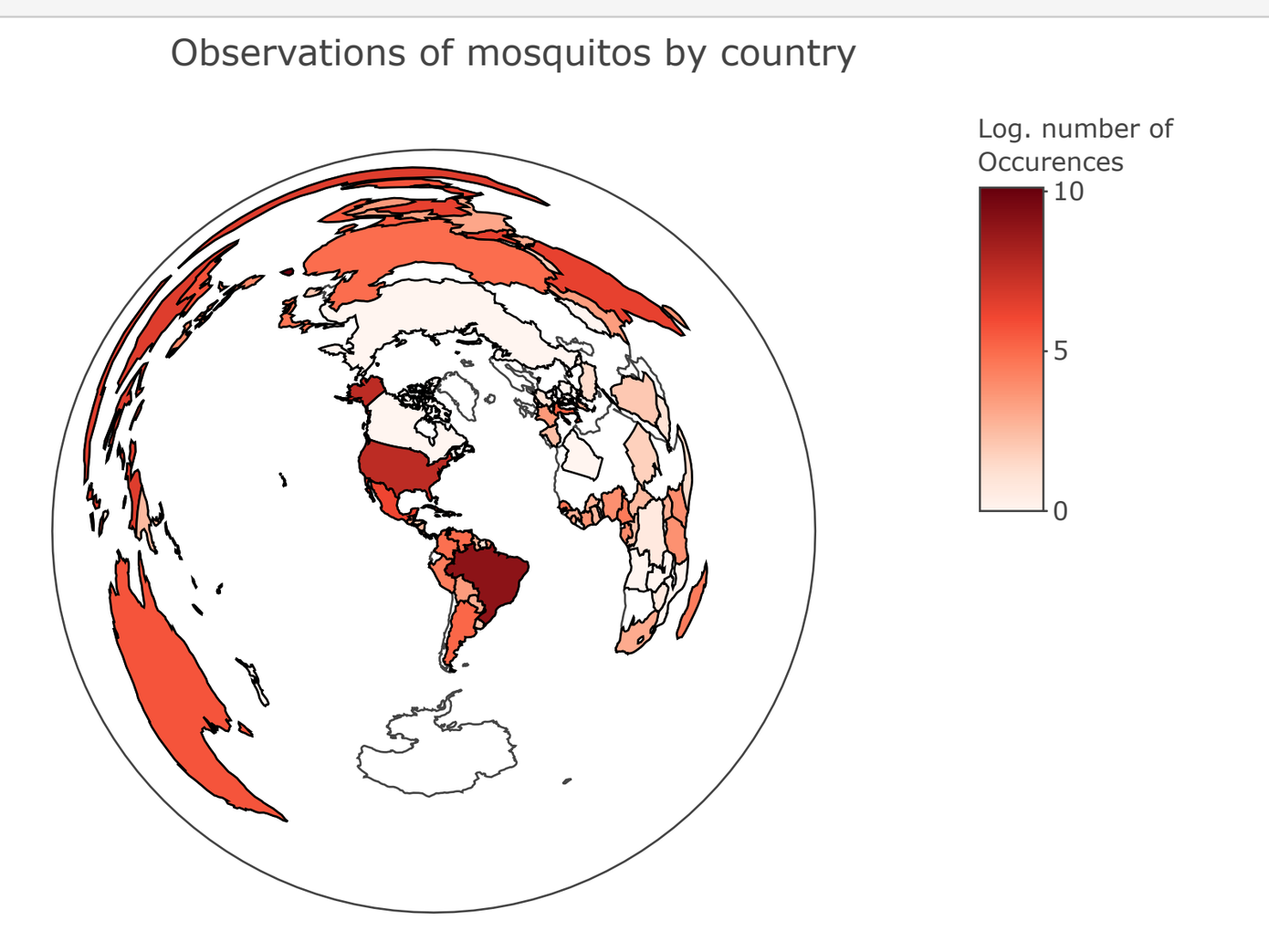
```
p3 <- data1 %>%dplyr::select(cou=COUNTRY_ID)%>% count(cou) %>%
  plot_geo() %>% add_trace(z = ~n, color = ~n, colors = 'Reds',text = ~cou, locations= ~cou) %>%colorbar(title="N
umber of <br>Occurences") %>% layout(title="Observations of mosquitos by country")
p3
```



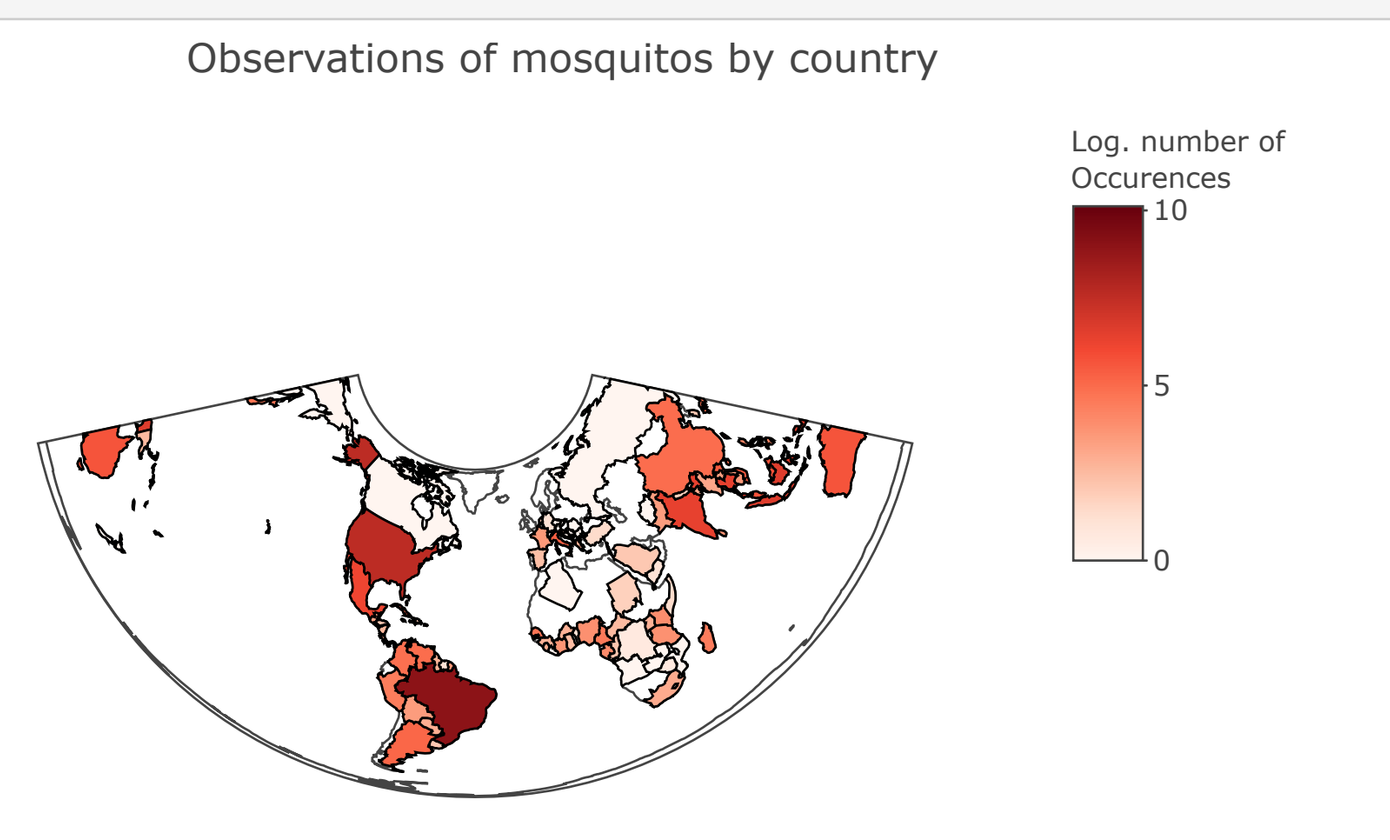
The color scale covers a very wide scale, ranging from 1 in Angola to 8,500 in Brazil. Because of that, it is hard to distinguish between the lower values, because they are similar on the scale, even though they are quite different in reality. For example, Afghanistan (1 case) and Malaysia (303 cases) have the same color on the map.

Task 3

```
g1 <- list(
  projection = list(type = "azimuthal equidistant")
)

p4 <- data1 %>%dplyr::select(cou=COUNTRY_ID)%>% count(cou) %>% plot_geo() %>%
  add_trace(z = ~log(n), color = ~log(n), colors = 'Reds',text = ~cou, locations= ~cou) %>%
  layout(geo=g1, title="Observations of mosquitos by country") %>% colorbar(title="Log. number of <br>Occurences"
)
p4
```



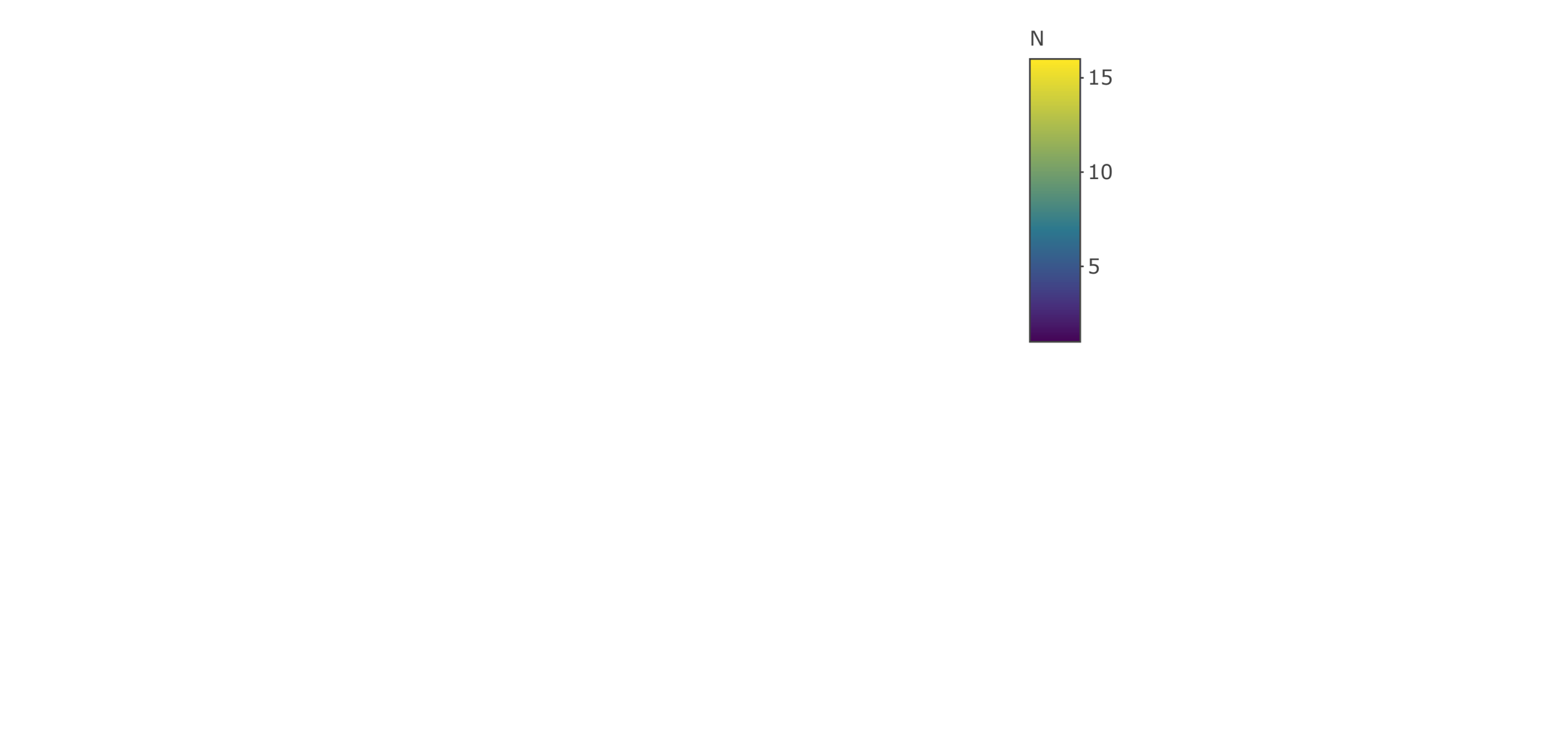

```
g2 <- list(
  projection = list(type = "conic equal area")
)

p5 <- data1 %>%dplyr::select(cou=COUNTRY_ID)%>% count(cou) %>% plot_geo() %>%
  add_trace(z = ~log(n), color = ~log(n), colors = 'Reds',text = ~cou, locations= ~cou) %>%
  layout(geo=g2 ,title="Observations of mosquitos by country") %>% colorbar(title="Log. number of <br>Occurences"
)

p5
```



Northern Europe and Northern Africa seem to have few occurences of mosquitos, whereas the Americas and southern Asia have more. When zoomed out, the conic equal area projection seems to ignore some places, e.g. the eastern part of Russia and also cuts countries in half. Still, it has the advantage of giving correct representations of the area. The azimuthal equidistant projection shows all countries at once, but is heavily distorted depending on the viewing angle.

Task 4

```
data1 %>% filter(COUNTRY_ID=='BRA',YEAR == "2013") %>%
  mutate(X1 = cut_interval(X,n=100), Y1 = cut_interval(Y,n=100))%>%
  group_by(X1,Y1) %>%
  summarise(mX=mean(X),mY=mean(Y),N=n()) %>%
  plot_mapbox(x=~mX,y=~mY,color=~N)
```



Yes, it helps since there is now less overplotting and one can get a more accurate idea of the distribution of mosquitos.

Assignment 2

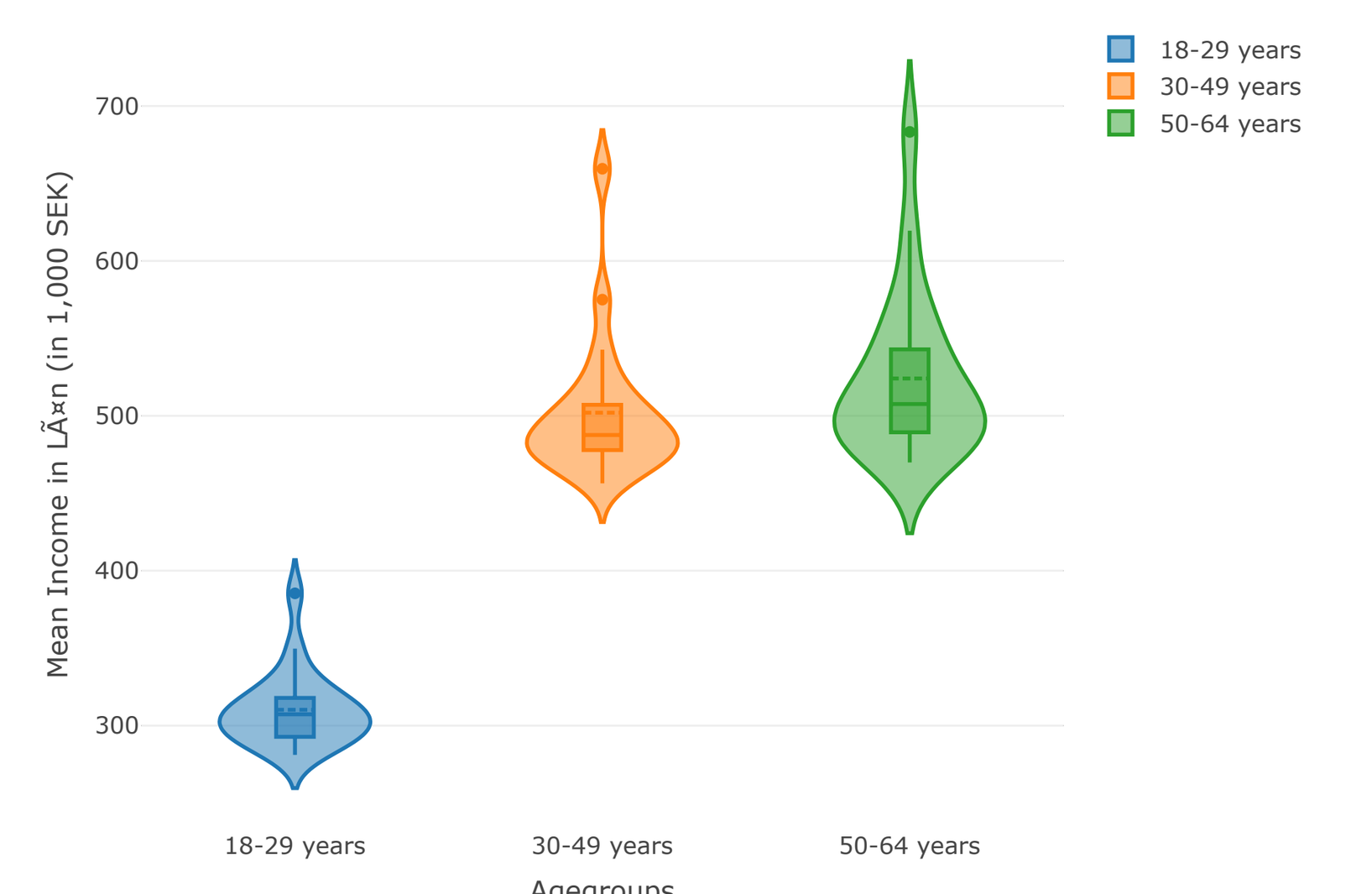
Task 1

```
data2 <- read.csv('KD.csv',header = TRUE, encoding = "latin1")
rds<-readRDS("gadm36_SWE_1_sf.rds")

datan<-data2[data2$age=='18-29 years',1:2]
datan$young<-data2$X2016[data2$age=='18-29 years']
datan$adult<-data2$X2016[data2$age=='30-49 years']
datan$senior<-data2$X2016[data2$age=='50-64 years']
head(datan)
```

```
##                    region type.of.household young adult senior
## 1      01 Stockholm county    all households 385.4 659.5  683.3
## 4        03 Uppsala county    all households 300.9 542.7  580.4
## 7   04 Södermanland county    all households 317.5 489.4  507.6
## 10 05 Östergötland county    all households 290.3 502.7  532.8
## 13     06 Jönköping county    all households 330.8 518.8  556.7
## 16     07 Kronoberg county    all households 307.3 503.2  530.3
```

Task 2

```
p6<-plot_ly(data2, x=~age, y=~X2016, split=~factor(age),
            type="violin", box=list(visible=T), meanline = list(visible = T)) %>% layout(yaxis=list(title="Mean I
ncome in LÃ¤n (in 1,000 SEK)"), xaxis=list(title="Agegroups"))
p6
```



The older you get the more you earn, where the difference between young and adult is greatest. Also variance seems to increase with age.

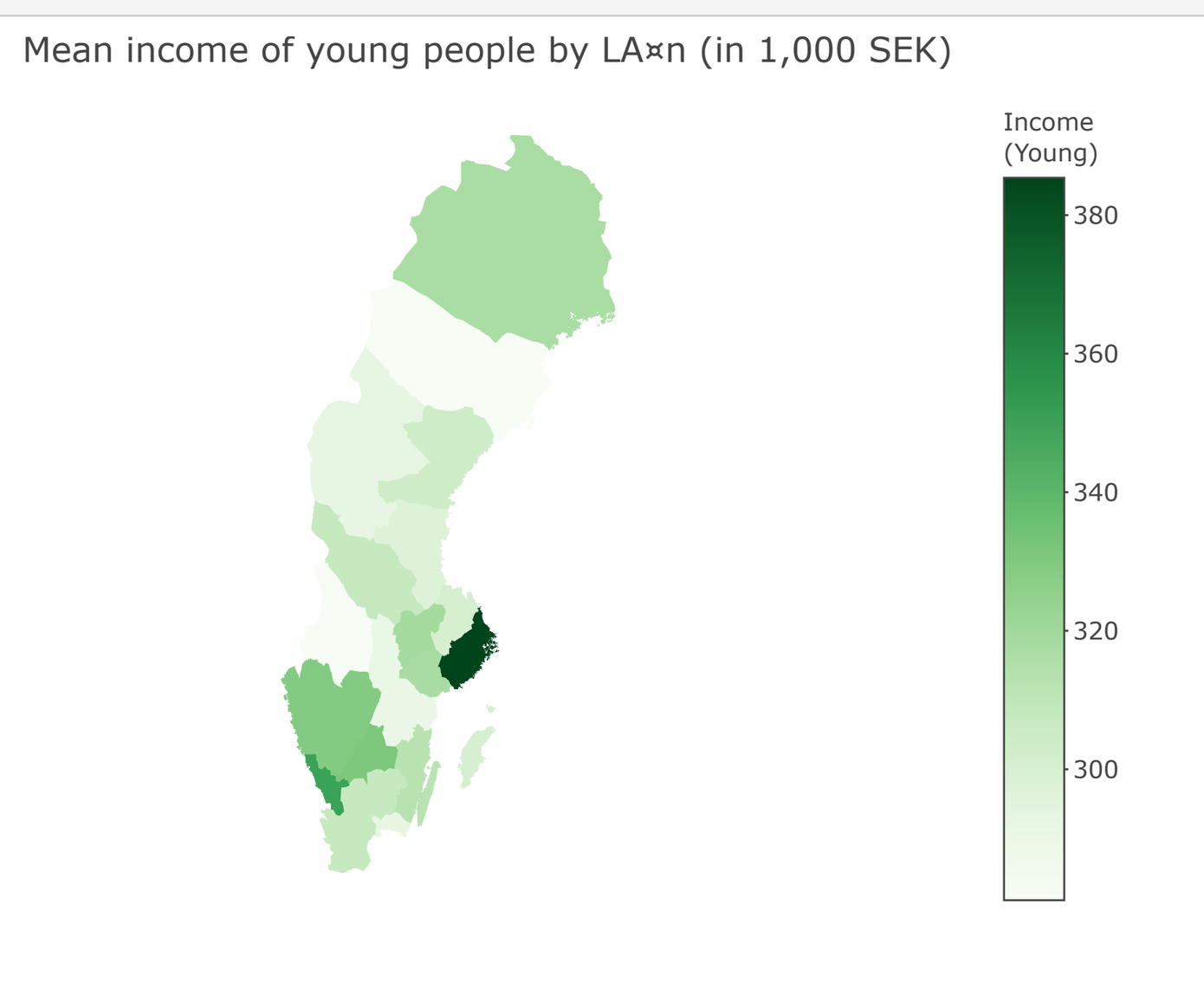
Task 3

```
attach(datan)
s=interp(senior,adult,young, duplicate = "mean")
detach(datan)
plot_ly(x=~s$x, y=~s$y, z=~s$z, type="surface", colors = 'Reds') %>%  layout(scene=list(xaxis=list(title="senior"
),yaxis=list(title="adult"),zaxis=list(title="young"))) %>% colorbar(title="young") %>% layout(title="Mean income
of young, adult and senior people in Swedish LÃ¤ns. <br> (in 1,000 SEK)")
```

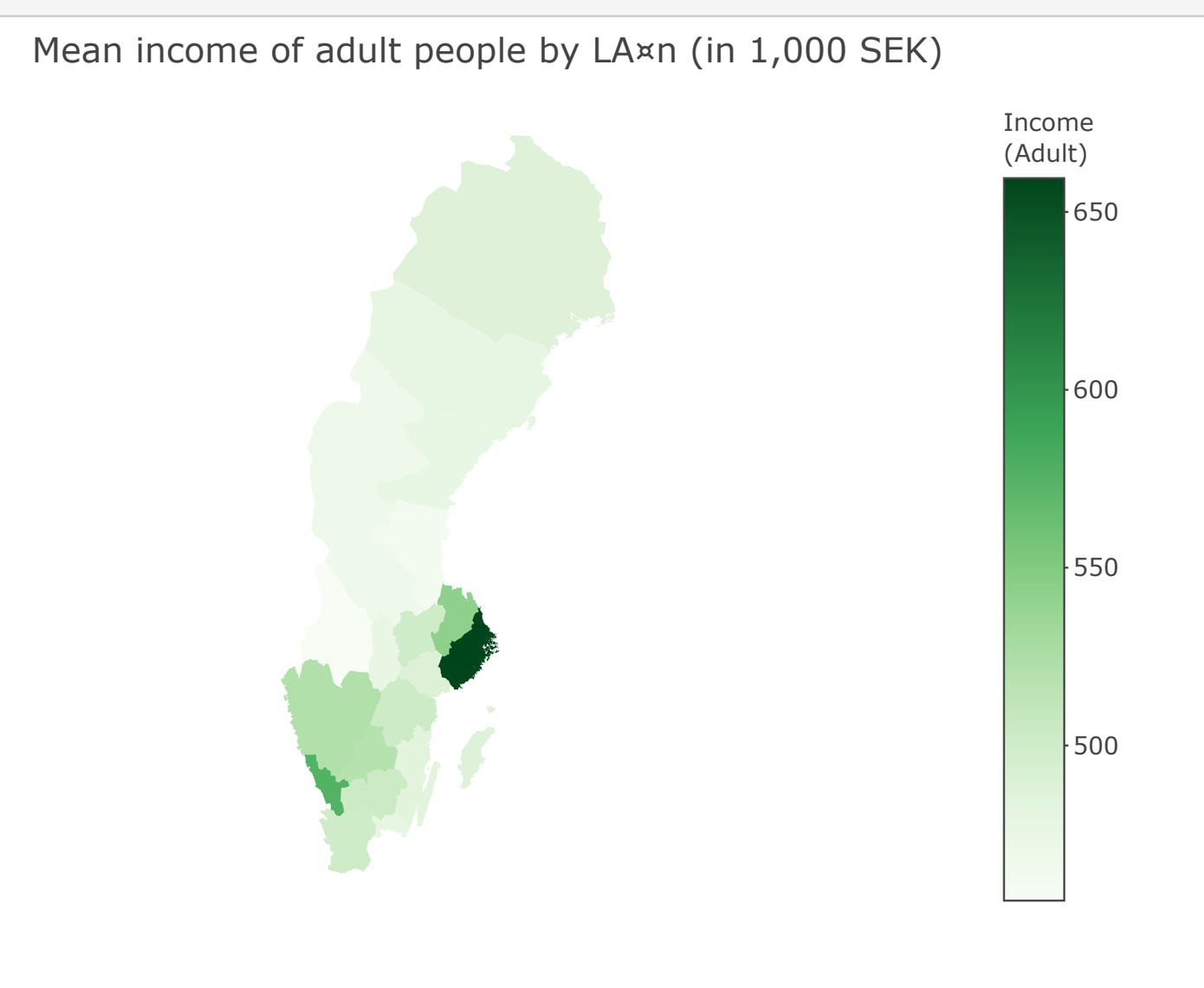
There seems to be a dependency between all three variables. Regions with high adult- and senior-salaries also show high salaries for the young. Since the graph appears to have a more or less monotonically incrasing shape, linear regression could be used.

Task 4

```
rownames(datan)=str_conv(datan$region,'latin1')
rds$young=as.numeric(sapply(rds$NAME_1, function(name) datan[grep(name,x=rownames(datan)),'young']))
rds$adult=as.numeric(sapply(rds$NAME_1, function(name) datan[grep(name,x=rownames(datan)),'adult']))
rds[11,'young']=290.9
rds[11,'adult']=478.1
p7<-plot_ly() %>% add_sf(data=rds, split=~NAME_1, color=~young,showlegend=F, alpha=1,colors = 'Greens') %>%
  layout(title="Mean income of young people by LÃ¤n (in 1,000 SEK)") %>%  colorbar(title="Income <br> (Young)")
p7
```




```
p8<-plot_ly() %>% add_sf(data=rds, split=~NAME_1, color=~adult,showlegend=F, alpha=1,colors = 'Greens') %>%
  layout(title="Mean income of adult people by LÃ¤n (in 1,000 SEK)") %>%  colorbar(title="Income <br> (Adult)")
p8
```



You can now see the geographic distribution of income. Regions close to larger cities like Stockholm and Uppsala as well as Vastra Gotaland (Gothenburg) and Halland show higher incomes. This difference is more significant for young people than for adults.

Task 5

```
ryd<-c(58.414871,15.56744)
p9<-plot_ly() %>% add_sf(data=rds, split=~NAME_1, color=~young,showlegend=F, alpha=1,colors = 'Greens')%>% add_ma
rkers(y=ryd[1],x=ryd[2],size=5, colors='red')%>%
  layout(title="Mean income of young people by LÃ¤n (in 1,000 SEK)") %>%  colorbar(title="Income <br> (Young)")
p9
```