Visualization Lab3 Group2

Note: You can copy mapbox token into a file called “.mapbox_token” in the same directory of RMD file.

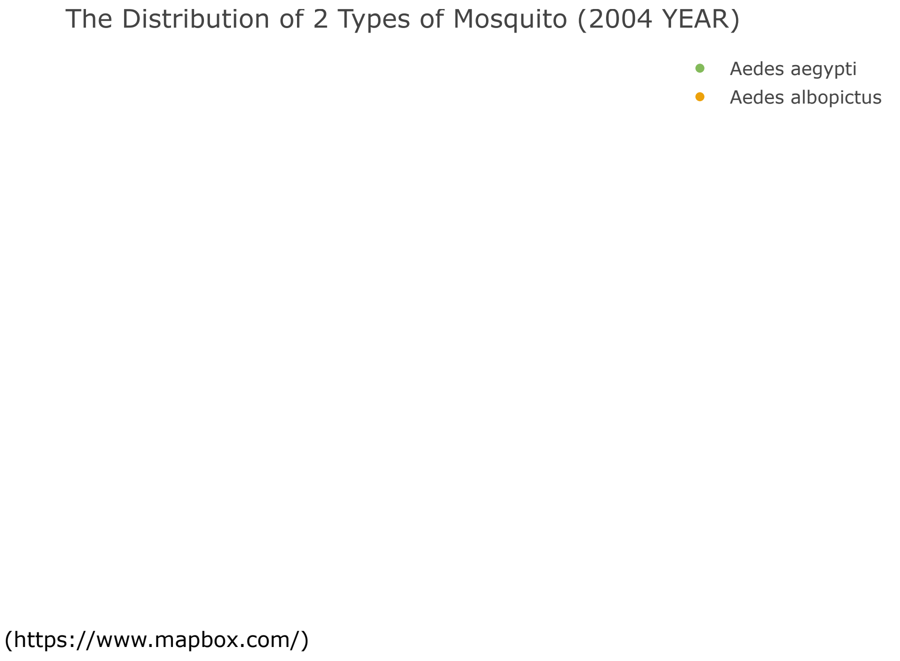
Assignment 1

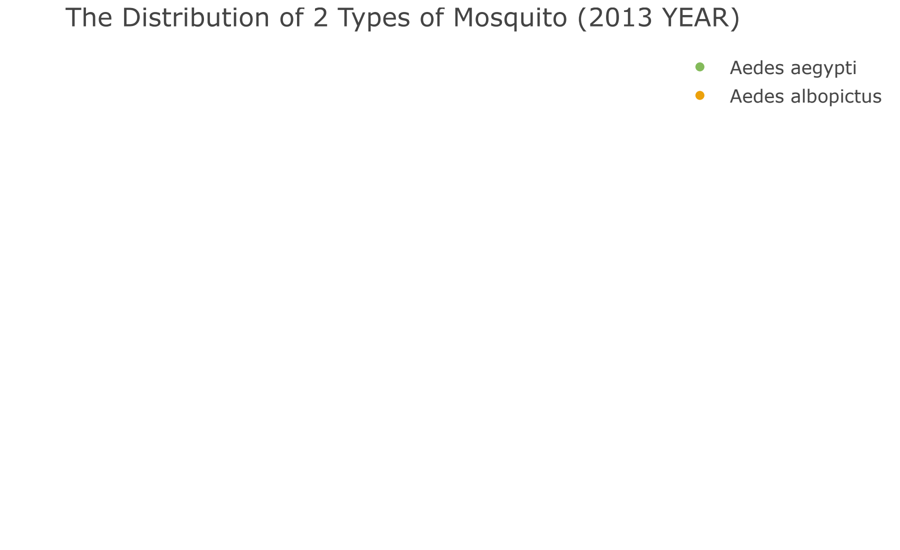
Test 1

```
aegypti_albopictus=read.csv("aegypti_albopictus.csv")

#filter() to filter out the specific row of data you want
A=filter(aegypti_albopictus,YEAR=="2004")
B=filter(aegypti_albopictus,YEAR=="2013")

fig1_2004<-plot_mapbox(A,x=~X,y=~Y) %>% add_markers(
  color=~VECTOR,colors = c("#83BA5A", "#ECA10B")
) %>% layout(
  title="The Distribution of 2 Types of Mosquito (2004 YEAR)"
)
fig1_2013<-plot_mapbox(B,x=~X,y=~Y) %>% add_markers(
  color=~VECTOR,colors = c("#83BA5A", "#ECA10B")
) %>% layout(
  title="The Distribution of 2 Types of Mosquito (2013 YEAR)"
)
```


The Distribution of 2 Types of Mosquito (2004 YEAR)
● Aedes aegypti
● Aedes albopictus
(https://www.mapbox.com/)



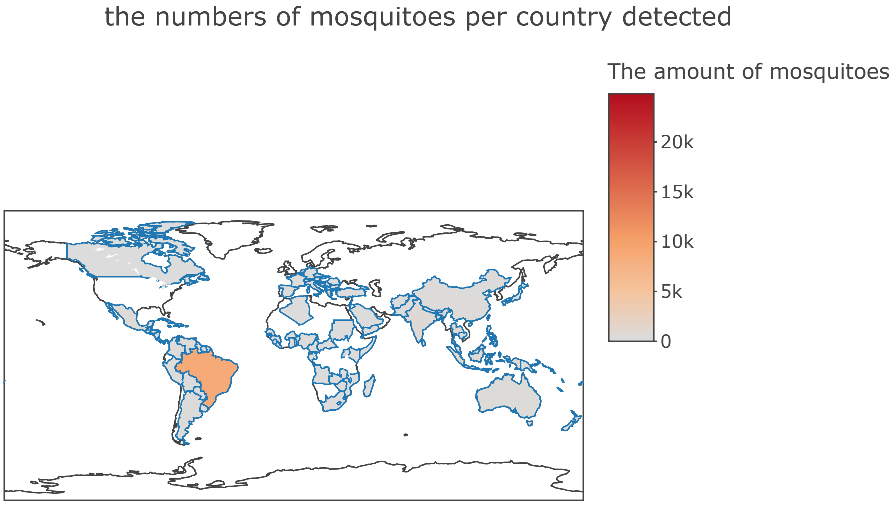
In 2004,the aedes albopitus distributed mainly near the tropic of cancer,especially in Taiwan,but some of them were spotted in Indonesia and the Southern Pacific Ocean, Meanwhile,Most of the aedes aegypti distributed in Mexico and South America,but some are found in Africa and southern Asia. In 2013,the amount of aedes albopictus had reduced rapidly,and mainly distributed in Taiwan and Northern Italy. Meanwhile,the amount of aedes aegypti had increased sharply,the distribution of aedes aegypti is concentrated in South America,especially in Brazil,where a great amount number of aedes aegypti had been spotted. In the 2nd figure,the amount of the points is too high,so that they would merge together,we could not see clearly,but we could still find the distribution of them roughly.

Test 2 (https://www.mapbox.com/)

```
#Import the id code of each nation
df <- read.csv("https://raw.githubusercontent.com/plotly/datasets/master/2014_world_gdp_with_codes.csv")
C=aegypti_albopictus %>% count(COUNTRY)
#The merge() function matches data or merges two data frames based on one or more columns
C=merge(df,C,by.x="COUNTRY",by.y="COUNTRY",all=TRUE)

g1 <- list(
  scope = 'world',
  projection = list(type = 'equirectangular'),
  showlakes = TRUE,
  lakecolor = toRGB('white')
)

fig2<-plot_geo(C) %>%
  add_trace(type="choropleth",z=~n,locations=~CODE,
            colorscale=c("#54BEF7","#183747"))  %>%
  colorbar(title="The amount of mosquitoes") %>%
  layout(
    title="the numbers of mosquitoes per country detected",
    geo=g1
  )
fig2
```

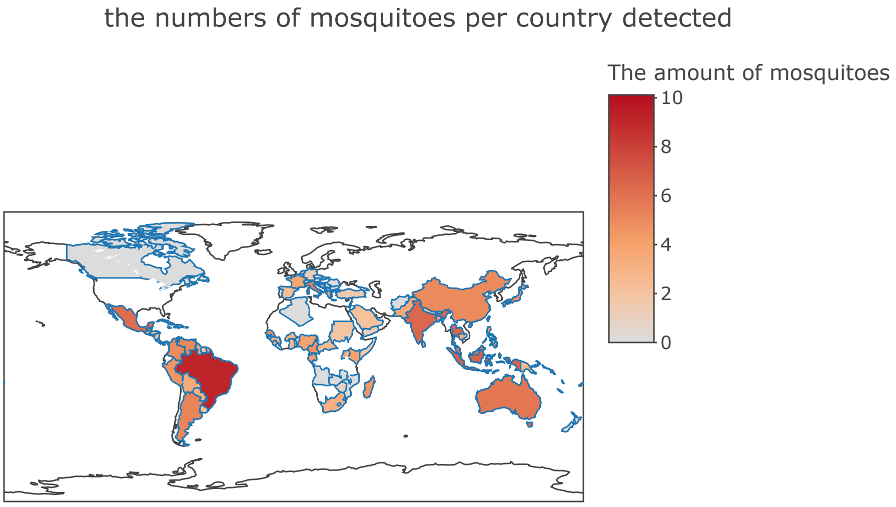


There is certain information in the map,because Brazil have had 8501 samples and Taiwan have had nearly 25k,compare to other countries,the numbers of mosquitos was much higher,so other countries finally turn out grey in the map compare to Brazil and Taiwan. We could barely see the difference between the numbers of mosquitoes of those countries.

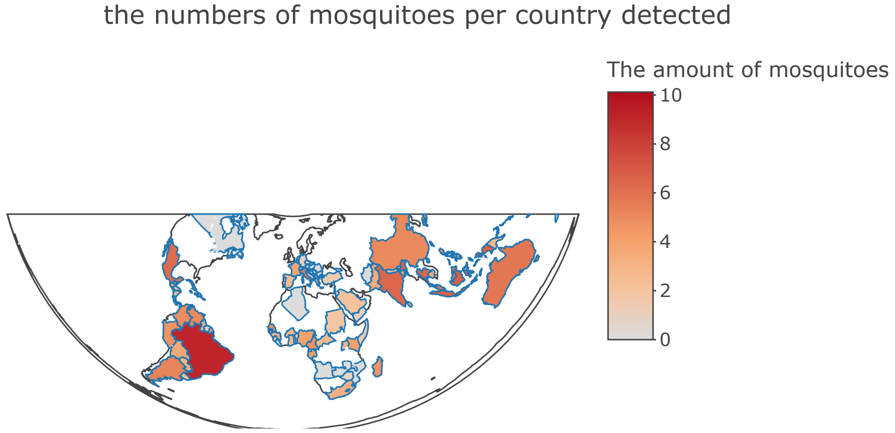
Task 3

```
#feature scaling
D<-C
D$n<-log(D$n)
#a
fig3<-plot_geo(D) %>%
  add_trace(type="choropleth",z=~n,locations=~CODE,
            colorscale=c("#54BEF7","#183747"))  %>%
  colorbar(title="The amount of mosquitoes") %>%
  layout(
    title="the numbers of mosquitoes per country detected",
    geo=g1
  )
#b
g2 <- list(
  scope = 'world',
  projection = list(type = 'conic equal area'),
  showlakes = TRUE,
  lakecolor = toRGB('white')
)

fig4<-plot_geo(D) %>%
  add_trace(type="choropleth",z=~n,locations=~CODE,
            colorscale=c("#54BEF7","#183747"))  %>%
  colorbar(title="The amount of mosquitoes") %>%
  layout(
    title="the numbers of mosquitoes per country detected",
    geo=g2
  )
```



the numbers of mosquitoes per country detected



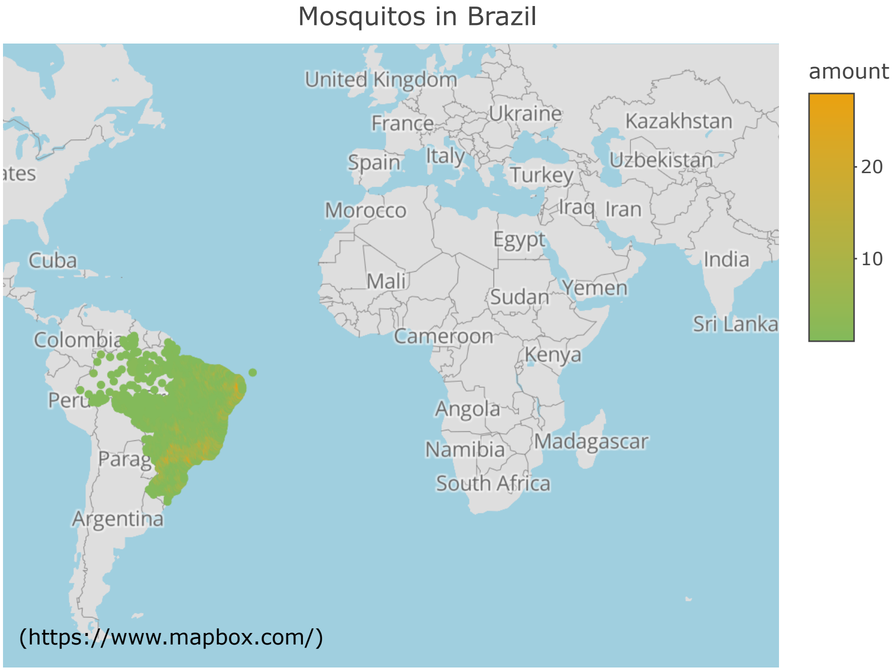
the numbers of mosquitoes per country detected

Normalized makes it easy to see the difference of the number of mosquitoes in each countries by color regardless how big the difference is between one country and the others. In fig3,The Equirectangular projection (i.e. the Mercator projection) zooms in too much at high latitudes and zooms out too much at low latitudes, so we would feel that Greenland is about the size of Australia, but it fits countries around the Tropic of Cancer, In contrast, in fig4,the Conic equal area projection (i.e. Albers projection) has the same problem, so it is only suitable for mid-latitude countries with large east-west spacing

Task 4

```
E=filter(aegypti_albopictus,YEAR %in% c("2013","2014") & COUNTRY=="Brazil")
E$X1=cut_interval(E$X,100)
E$Y1=cut_interval(E$Y,100)
#group_by() takes an existing tbl and converts it into a grouped tbl where operations are performed "by group".
F=E %>% select(X1,Y1,X,Y) %>%
  group_by(X1,Y1) %>%
  summarize(mean_X=mean(X),mean_Y=mean(Y),amount=n())

plot_mapbox(F,x=~mean_X,y=~mean_Y)  %>%
  add_markers(color=~amount,colors = c("#83BA5A", "#ECA10B"))  %>%
  layout(title = "Mosquitos in Brazil")
```



The most severe area infested by mosquitoes are Natal,Rio de Janeiro,Sao Paulo and Maringa. Such discretization will undoubtly help in analyzing the distribution of mosquitoes,cause the sample points won't merge together and we could see the amount of mosquitoes in the areas by the color of the scatter points.

Assignment 2

1. Load data

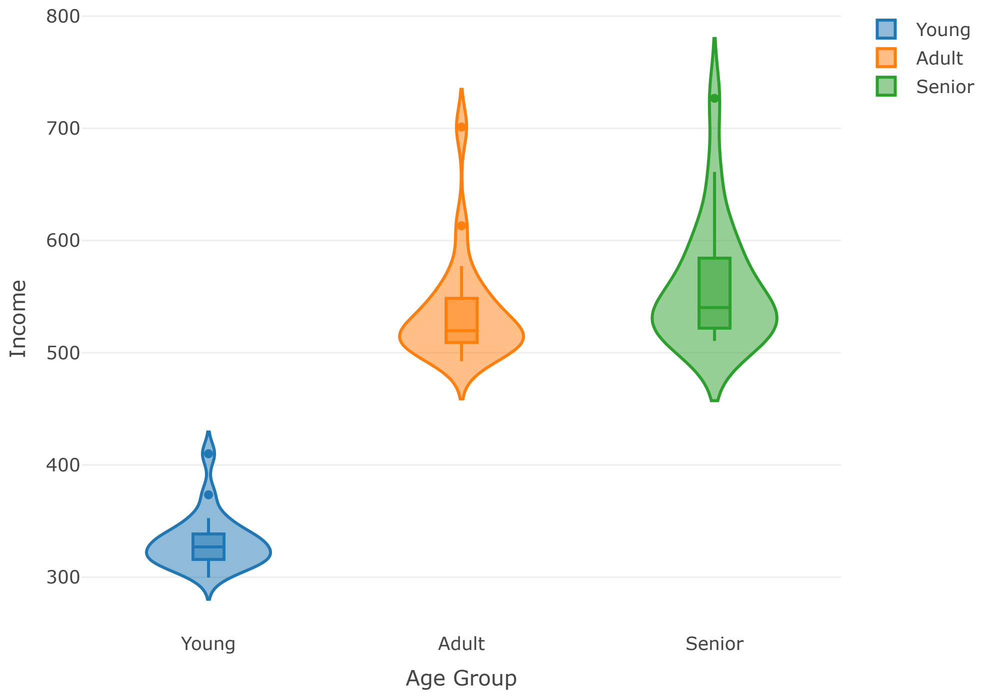
```
rds2 <- readRDS("gadm36_SWE_1_sf.rds")

df2 <- read.csv("000000KD_20220914-153718-UTF8.csv")
df2$type.of.household <- NULL
colnames(df2) <- c("region", "age", "income")

# group
df2 <- df2 %>% mutate("group" = case_when(
    age == "18-29 years" ~ "Young",
    age == "30-49 years" ~ "Adult",
    age == "50-64 years" ~ "Senior"))
```

2. Violin plot

```
# ensure plot order
ordered <- factor(df2$group, levels = c("Young", "Adult", "Senior"))
plot_ly(df2, x = ordered, y = ~income, split = ordered, type = "violin", box = list(visible = TRUE)) %>%
  layout(xaxis = list(title = "Age Group"), yaxis = list(title = "Income"))
```



Q1: Analyze this plot and interpret your analysis in terms of income.

In general, adults and seniors are at the same level of income, while young people have the lowest income, which is much less than the other two groups. Young people are just starting their careers, the less income is reasonable. As getting older and more experienced, salary may go up and there will be more ways to manage their money, so the disposable income for households would be more.

From the median value, we can see the median of the young is almost in the middle of the box, and with the age growing, the median is closer to q1. It may because the income disparities become larger after working for years.

The peaks of the density curves in all three groups are below the median value. It shows that most people’s income is lower than the median. It may be an evidence of that the few hold the majority of the wealth.

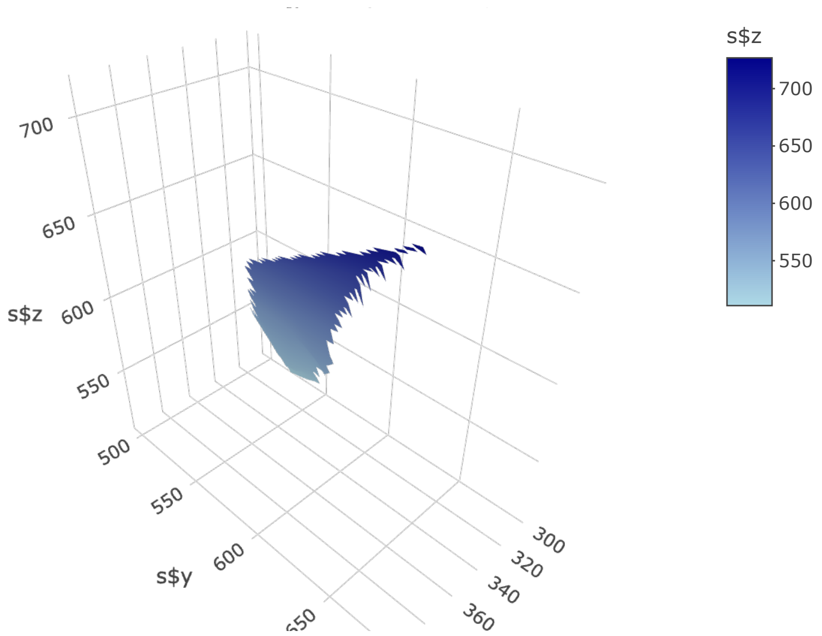
3. Surface plot

```
grouped <- split(df2, df2$group)

df_income <- data.frame(
  region = grouped$Young$region,
  young  = grouped$Young$income,
  adult  = grouped$Adult$income,
  senior = grouped$Senior$income)

attach(df_income)
s <- interp(young, adult, senior, duplicate = "mean")
detach(df_income)

plot_ly(x = ~s$x, y = ~s$y, z = ~s$z, type = "surface", colors = colorRamp(c("lightblue","darkblue")))
```



Q1: What kind of trend can you see and how can this be interpreted?

If the income of young people is high and the income of adults is high, the income of the senior will be high. The disposable income of senior is positively correlated with two other groups. If a person has sufficient disposable income in their youth and adulthood, it is evidence that they may have a substantial family or be a very capable person, either working or investing, which will provide a good basis for income in old age.

Q2: Do you think that linear regression would be suitable to model this dependence?

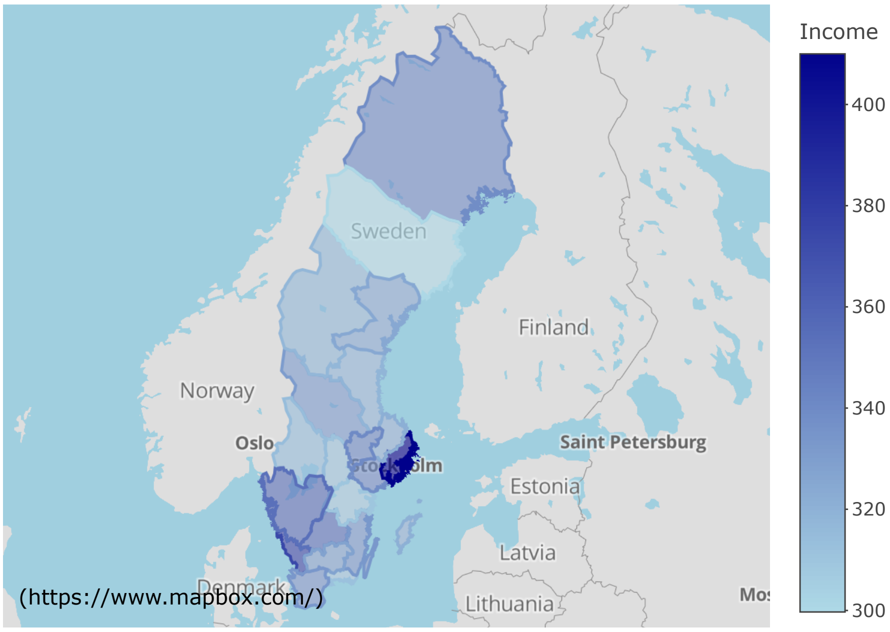
Linear regression would be suitable because we can see from the plot that there is a linear relationship between x, y and z, and x and y are independent of each other.

4. Choropleth maps

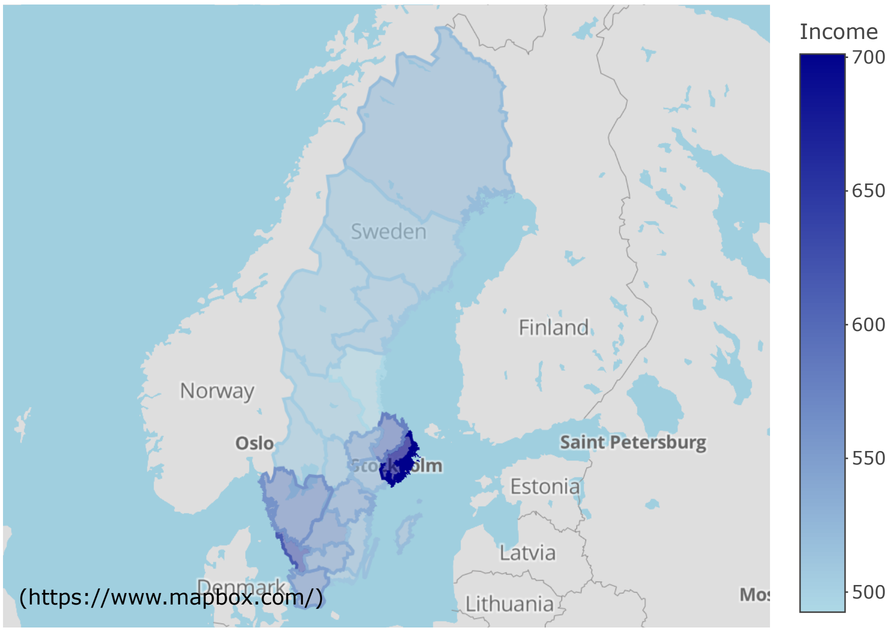
```
rds2_4_young <- rds2
rds2_4_young$Income <- df_income[rds2_4_young$NAME_1, "young"]
rds2_4_young$Income[is.na(rds2_4_young$Income)] <- 0
p_young <- plot_mapbox() %>%
  add_sf(data  = rds2_4_young,
         split = ~NAME_1,
         color = ~Income,
         showlegend = FALSE,
         alpha  = 0.6,
         colors = colorRamp(c("lightblue","darkblue")))

rds2_4_adult <- rds2
rds2_4_adult$Income <- df_income[rds2_4_adult$NAME_1, "adult"]
rds2_4_adult$Income[is.na(rds2_4_adult$Income)] <- 0
p_audlt <- plot_mapbox() %>%
  add_sf(data  = rds2_4_adult,
         split = ~NAME_1,
         color = ~Income,
         showlegend = FALSE,
         alpha  = 0.6,
         colors = colorRamp(c("lightblue","darkblue")))
```

Income of Young people



Income of Audlts



Q1: Analyze these maps and make conclusions.

For young people, there is no particularly large north-south difference. But Västerbotten and Blekinge has significantly less income than their neighbor regions. Norrbotten is prominent in the northern provinces. Östergötland, Värmland, Örebro form a relatively low-income zone in the south-central.

For adults, most southern provinces have more disposable income than northern provinces.

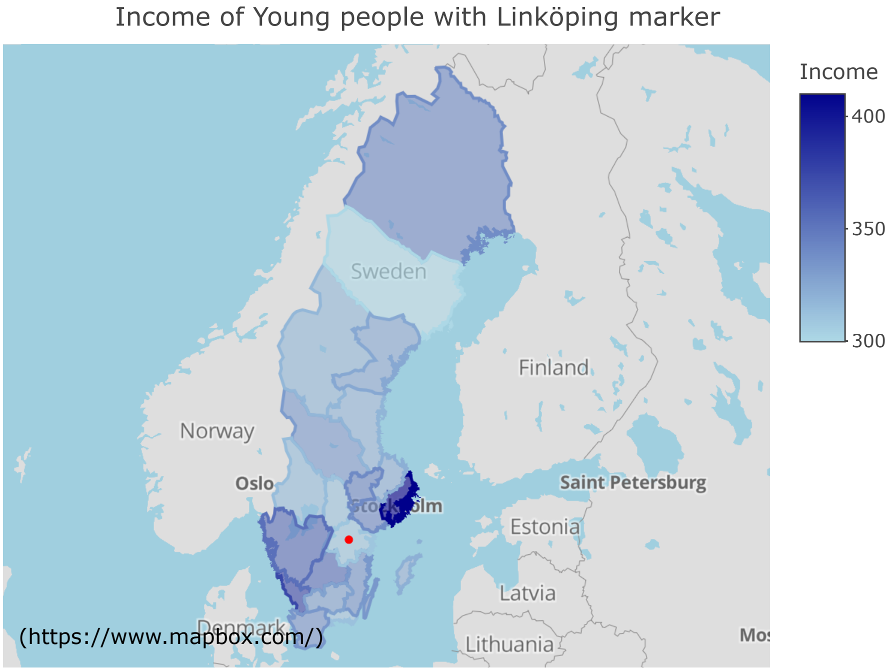
Stockholm has the highest disposable income, both among young people and adults.

Q2: Is there any new information that you could not discover in previous statistical plots?

Among some northern provinces, like Norrbotten, Västernorrland, Västerbotten and Jämtland, the income disparities for adults has narrowed, unlike the income disparities for young people. In other provinces, the income of adults has lost the advantage it showed in the data of young people, like Gävleborg.

5. Where is Linköping

```
p_young %>%
  add_trace(lat = 58.409814,
            lon = 15.624522,
            mode  = "markers",
            color = I("red"),
            text  = "Linköping",
            hoverinfo = 'text') %>%
  layout(title = 'Income of Young people with Linköping marker')
```



Appendix

Codes For Assignment 1

```r
library("ggplot2")
library("plotly")
library("dplyr")
library("rjson")

#1
read.csv("~/aegypti_albopictus.csv")
Sys.setenv("MAPBOX_TOKEN"="{mapbox_toekn}")
A=filter(aegypti_albopictus,YEAR=="2004")
B=filter(aegypti_albopictus,YEAR=="2013")

fig1_2004<-plot_mapbox(A,x=~X,y=~Y) %>% add_markers(
  color=~VECTOR,colors = c("#83BA5A", "#ECA10B")
) %>% layout(
  title="The Distribution of 2 Types of Mosquito (2004 YEAR)"
)
fig1_2013<-plot_mapbox(B,x=~X,y=~Y) %>% add_markers(
  color=~VECTOR,colors = c("#83BA5A", "#ECA10B")
) %>% layout(
  title="The Distribution of 2 Types of Mosquito (2013 YEAR)"
)

#2
df <- read.csv("https://raw.githubusercontent.com/plotly/datasets/master/2014_world_gdp_with_codes.csv")
C=aegypti_albopictus %>% count(COUNTRY)
C=merge(df,C,by.x="COUNTRY",by.y="COUNTRY",all=TRUE)

g1 <- list(
  scope = 'world',
  projection = list(type = 'equirectangular'),
  showlakes = TRUE,
  lakecolor = toRGB('white')
)

fig2<-plot_geo(C) %>%
  add_trace(type="choropleth",z=~n,locations=~CODE.x,
            colorscale=c("#54BEF7","#183747"))  %>%
  colorbar(title="The amount of mosquitos") %>%
  layout(
    title="the numbers of mosquitos per country detected",
    geo=g1
  )

#3
#feature scaling
D<-C
D$n<-log(D$n)
#a
fig3<-plot_geo(D) %>%
  add_trace(type="choropleth",z=~n,locations=~CODE.x,
            colorscale=c("#54BEF7","#183747"))  %>%
  colorbar(title="The amount of mosquitos") %>%
  layout(
    title="the numbers of mosquitos per country detected",
    geo=g1
  )
#b
g2 <- list(
  scope = 'world',
  projection = list(type = 'conic equal area'),
  showlakes = TRUE,
  lakecolor = toRGB('white')
)

fig4<-plot_geo(D) %>%
  add_trace(type="choropleth",z=~n,locations=~CODE.x,
            colorscale=c("#54BEF7","#183747"))  %>%
  colorbar(title="The amount of mosquitos") %>%
  layout(
    title="the numbers of mosquitos per country detected",
    geo=g2
  )
```

```
#4
E=filter(aegypti_albopictus,YEAR %in% c("2013","2014") & COUNTRY=="Brazil")
E$X1=cut_interval(E$X,100)
E$Y1=cut_interval(E$Y,100)
#group_by() takes an existing tbl and converts it into a grouped tbl where operations are performed "by group".
F=E %>% select(X1,Y1,X,Y) %>%
  group_by(X1,Y1) %>%
  summarize(mean_X=mean(X),mean_Y=mean(Y),amount=n())

plot_mapbox(F,x=~mean_X,y=~mean_Y)  %>%
  add_markers(color=~amount,colors = c("#83BA5A", "#ECA10B"))  %>%
  layout(title = "Mosquitos in Brazil")
```

Codes For Assignment 2

```r
library(dplyr)
library(ggplot2)
library(plotly)
library(akima)

Sys.setenv("MAPBOX_TOKEN" = "{mapbox_token}")

# Task 1

rds2 <- readRDS("gadm36_SWE_1_sf.rds")

df2 <- read.csv("000000KD_20220914-153718-UTF8.csv")
df2$type.of.household <- NULL
colnames(df2) <- c("region", "age", "income")

# group
df2 <- df2 %>% mutate("group" = case_when(
    age == "18-29 years" ~ "Young",
    age == "30-49 years" ~ "Adult",
    age == "50-64 years" ~ "Senior"))

# Task 2

ordered <- factor(df2$group, levels = c("Young", "Adult", "Senior"))
plot_ly(df2, x = ordered, y = ~income, split = ordered, type = "violin", box = list(visible = TRUE)) %>%
  layout(xaxis = list(title = "Age Group"), yaxis = list(title = "Income"))

# Task 3

grouped <- split(df2, df2$group)

df_income <- data.frame(
  region = grouped$Young$region,
  young  = grouped$Young$income,
  adult  = grouped$Adult$income,
  senior = grouped$Senior$income)

attach(df_income)
s <- interp(young, adult, senior, duplicate = "mean")
detach(df_income)

plot_ly(x = ~s$x, y = ~s$y, z = ~s$z, type = "surface", colors = colorRamp(c("lightblue","darkblue")))

# Task 4

mapped_names <- lapply(df2$region, function(region) {
  parts <- strsplit(region, " ")[[1]]
  name <- paste(parts[-c(1, length(parts))], collapse = ' ')
  if (name == "Örebro") { name <- "Orebro" }
  return(name)
})
rownames(df_income) <- mapped_names[!duplicated(mapped_names)]

rds2_4_young <- rds2
rds2_4_young$Income <- df_income[rds2_4_young$NAME_1, "young"]
rds2_4_young$Income[is.na(rds2_4_young$Income)] <- 0
p_young <- plot_mapbox() %>%
  add_sf(data  = rds2_4_young,
         split = ~NAME_1,
         color = ~Income,
         showlegend = F,
         alpha  = 0.6,
         colors = colorRamp(c("lightblue","darkblue")))

rds2_4_adult <- rds2
rds2_4_adult$Income <- df_income[rds2_4_adult$NAME_1, "adult"]
rds2_4_adult$Income[is.na(rds2_4_adult$Income)] <- 0
p_audlt <- plot_mapbox() %>%
  add_sf(data  = rds2_4_adult,
         split = ~NAME_1,
         color = ~Income,
         showlegend = F,
         alpha  = 0.6,
         colors = colorRamp(c("lightblue","darkblue")))

p_young %>%
  layout(title = 'Income of Young people')
p_audlt %>%
  layout(title = 'Income of Audlts')

# Task 5

p_young %>%
  add_trace(lat = 58.409814,
            lon = 15.624522,
            mode  = "markers",
            color = I("red"),
            text  = "Linköping",
            hoverinfo = 'text') %>%
  layout(title = 'Income of Young people with Linköping marker')
```