# project_yutong

*Yutong Wang*

*4/23/2019*

# Any package that is required by the script below is given here:

```
inst_pkgs = load_pkgs =  c("ggplot2","ggplot2movies", "dplyr","babynames","data.table","Rcpp","devtools")
inst_pkgs = inst_pkgs[!(inst_pkgs %in% installed.packages()[,"Package"])]
if(length(inst_pkgs)) install.packages(inst_pkgs)

git_pkgs = git_pkgs_load = c("streamgraph","DT")

git_pkgs = git_pkgs[!(git_pkgs %in% installed.packages()[,"Package"])]

if(length(git_pkgs)){
  library(devtools)
  install_github('rstudio/DT')
  install_github('hrbrmstr/streamgraph')
}

load_pkgs = c(load_pkgs, git_pkgs_load)

# Dynamically load packages
pkgs_loaded = lapply(load_pkgs, require, character.only=T)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggplot2movies
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: babynames
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## Loading required package: Rcpp
## Loading required package: devtools
```

```
## Loading required package: streamgraph
```

```
## Loading required package: DT
```

# Downloading the necessary data (done in terminal):

```
wget https://raw.githubusercontent.com/coatless/stat490uiuc/master/airlines/airlines_data.sh
chmod u+x airlines_data.sh
./airlines_data.sh 2002 2002
mv airlines.csv groupproject.csv
./airlines_data.sh 1998 1998
tail -n+2 airlines.csv >> groupproject.csv
```

# Deleting columns with all data missing (done in terminal):

```
cut -d ',' -f 23,25,26,27,28,29 --complement groupproject.csv > groupprojectcl.csv
```

# Creating the necessary tables (done in hive):

Creating table with flight data:

```
CREATE TABLE flights (Year INT, Month INT, DayofMonth INT,  DayOfWeek INT, DepTime INT,
CRSDepTime INT, ArrTime INT, CRSArrTime INT, UniqueCarrier STRING, FlightNum INT, TailNum STRING,
ActualElapsedTime INT, CRSElapsedTime INT, AirTime INT, ArrDelay INT, DepDelay INT, Origin STRING,
Dest STRING, Distance INT, TaxiIn INT, TaxiOut INT,Cancelled INT, Diverted INT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
LOAD DATA LOCAL INPATH 'groupprojectcl.csv'
OVERWRITE INTO TABLE flights;
ALTER TABLE flights set tblproperties("skip.header.line.count"="1");
```

Creating table with the daily number of flights:

```
Hive -e "SELECT Year, Month, DayofMonth, count (DayofMonth) FROM flights GROUP BY Year, Month, DayofMonth" > coun
t_day.csv
```

Creating table with the daily number of flights for Carriers:

```
Hive -e "SELECT Year, Month, DayofMonth, UniqueCarrier,  count (UniqueCarrier) FROM flights GROUP BY Year, Month,
DayofMonth, UniqueCarrier" > count_carrier_day.csv
```

Creating table with the number of flights for destinations:

```
hive -e "SELECT Year, Dest, count (Dest) FROM flights GROUP BY Year, Dest" > dest.csv
```

# Data preparation

```
library(biganalytics)
```

```
## Loading required package: bigmemory
```

```
## Loading required package: foreach
```

```
## Loading required package: biglm
```

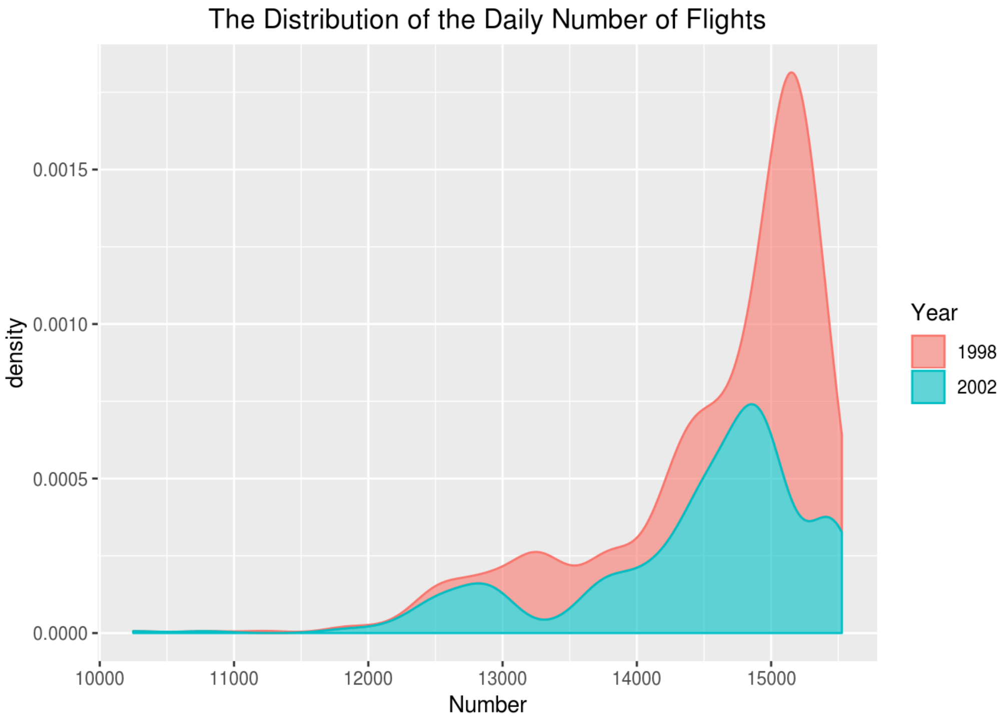
```
## Loading required package: DBI
```

```
x <-attach.big.matrix("gp.desc")
x<-x[-1,]
colnames(x)<- c("Year", "Month", "DayofMonth", "DayOfWeek",
                "DepTime", "CRSDepTime", "ArrTime", "CRSArrTime", "UniqueCarrier", "FlightNum",
                "TailNum", "ActualElapsedTime", "CRSElapsedTime", "AirTime", "ArrDelay",
                "DepDelay", "Origin", "Dest", "Distance", "TaxiIn", "TaxiOut", "Cancelled", "Diverted","Elapsed")
year98 <- x[which(x[,"Year"] == 1998),]
year02 <- x[which(x[,"Year"] == 2002),]

data <- read.csv("divert.csv", header=FALSE,sep = "\t")
colnames(data) <- c("Year","Origin", "Dest")
```

# Descriptive Analysis

## The distribution of the daily number of flights in these two years

```
day.n <- as.data.frame(read.csv("count_day.csv",sep = "\t", header = FALSE))
colnames(day.n) <- c("Year","Month","DayofMonth","Number")
day.n$Year <- factor(day.n$Year)
day.n$date <- as.Date(paste(day.n$Month, day.n$DayofMonth),format = "%m%d")
ggplot(day.n, aes(x = Number)) +
  geom_density(aes(group = Year, colour = Year, fill = Year), position="stack", alpha = 0.6) +
  labs(title = "The Distribution of the Daily Number of Flights") +
  theme(plot.title = element_text(hjust = 0.5))
```

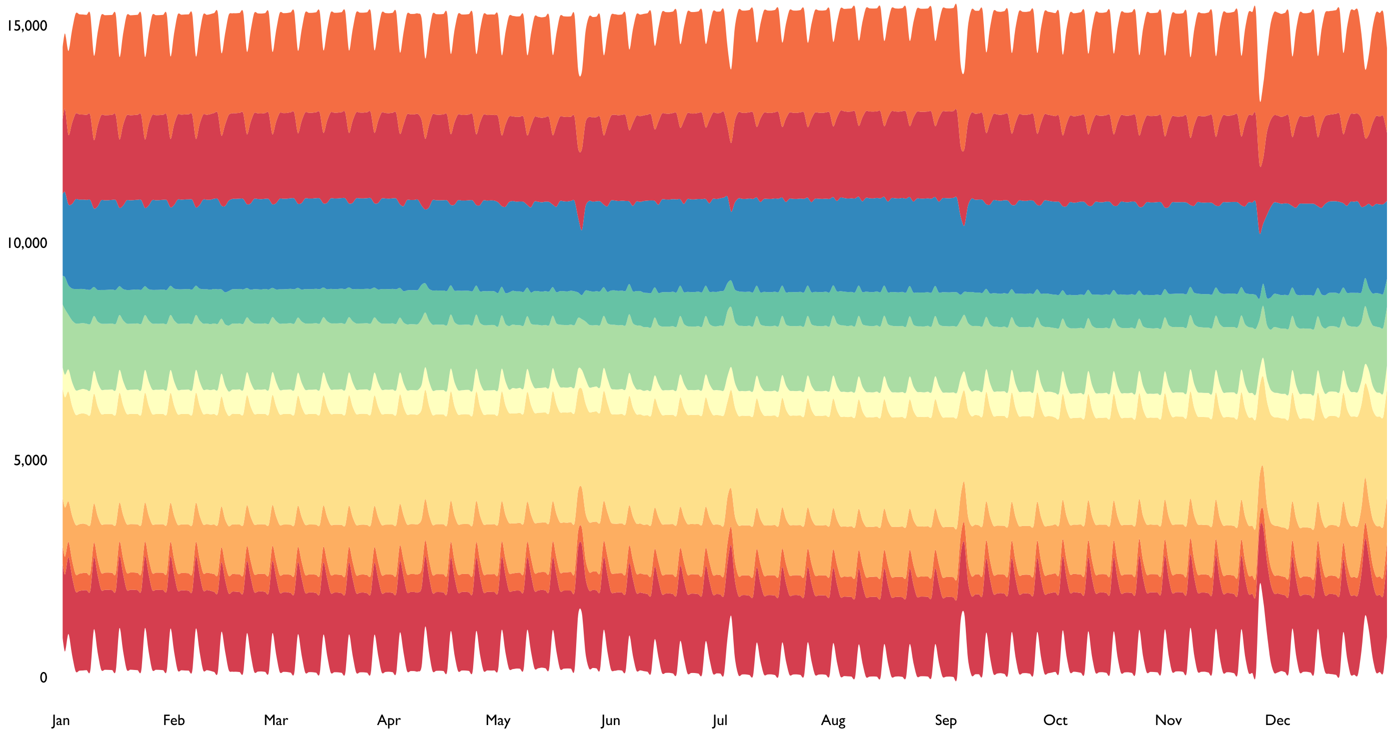


In the above plot, the red denotes the distribution of the daily number of flights in 1998, and the blue denotes that in 2002. We can see the distribution is left-skewed, which means the number of flights is larger than 14000 most of the time. Comparing the distributions for two years, the number of flights in 2002 is less than that in 1998.

# The daily number of flights for different carrier in two years.

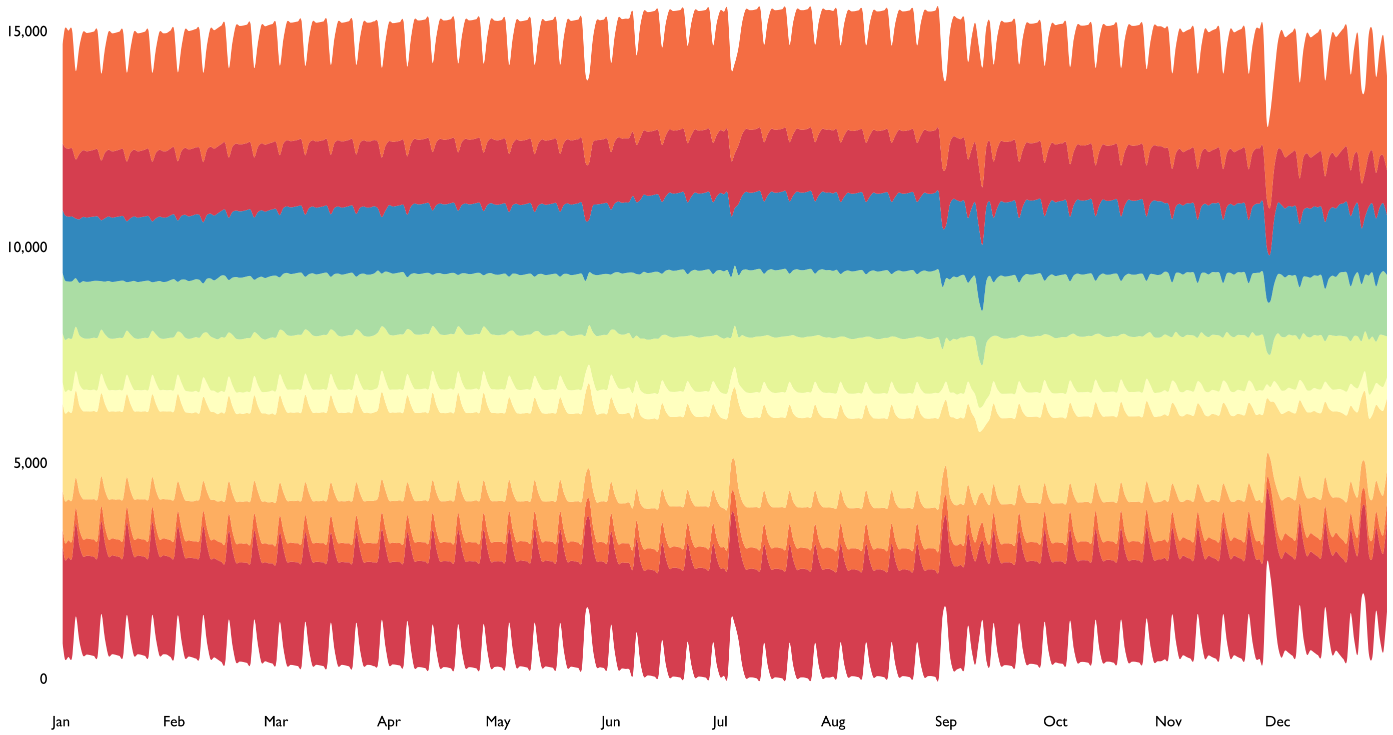
```
day.carrier.n <- as.data.frame(read.csv("count_carrier_day.csv",sep= "\t", header = FALSE))
colnames(day.carrier.n) <- c("Year","Month","DayofMonth","Carrier","Number")
day.carrier.n$date <- as.Date(paste(day.carrier.n $Year,day.carrier.n $Month, day.carrier.n $DayofMonth, sep = '-
'))
carrier98 <- day.carrier.n[which(day.carrier.n["Year"] == "1998"),]
carrier02 <- day.carrier.n[which(day.carrier.n["Year"] == "2002"),]

streamgraph(carrier98, "Carrier", "Number", "date") %>%
  sg_fill_brewer("Spectral") %>%
  sg_axis_x(tick_units = date) %>%
  sg_title(title ="The Number of Flights for Carriers Over time in 1998")
```

**The Number of Flights for Carriers Over time in 1998**



```
streamgraph(carrier02, "Carrier", "Number", "date") %>%
  sg_fill_brewer("Spectral") %>%
  sg_axis_x(tick_units = date) %>%
    sg_title(title ="The Number of Flights for Carriers Over time in 2002")
```

**The Number of Flights for Carriers Over time in 2002**



These plots show the daily number of different carriers. Both of the plots contain 10 Carriers, but we can see the carriers are different in these two years.

The MQ joined in 2002. MQ denotes the Carrier of *American Eagle Airlines Inc*. By checking the information of this carrier, Flagship Airlines and Wings West were merged in 1998, with the new carrier named American Eagle Airlines. This brand name was used until being discontinued in 2014. The TW disappeared in 2002, which denote the *Trans World Airways LLC*. This company suffered financial problems in 2001, whose assets were acquired by the parent company of American Airlines. They quickly changed their name and formed a new company. So, this carrier only appears in 1998.

From the plots, we can see the number of flights is periodic. There is not much difference in the daily number of flights.

Comparing the plots for two years, the shape of the plots are similar, but the daily number is more stable in 1998. The number of flights increases a little in the summer of 2002. In 1998, DL occupied most of the flights, and the flights' number of TW is the least. Then, UA and WN occupied most of the flights.

# The Number of Flights at Different Times

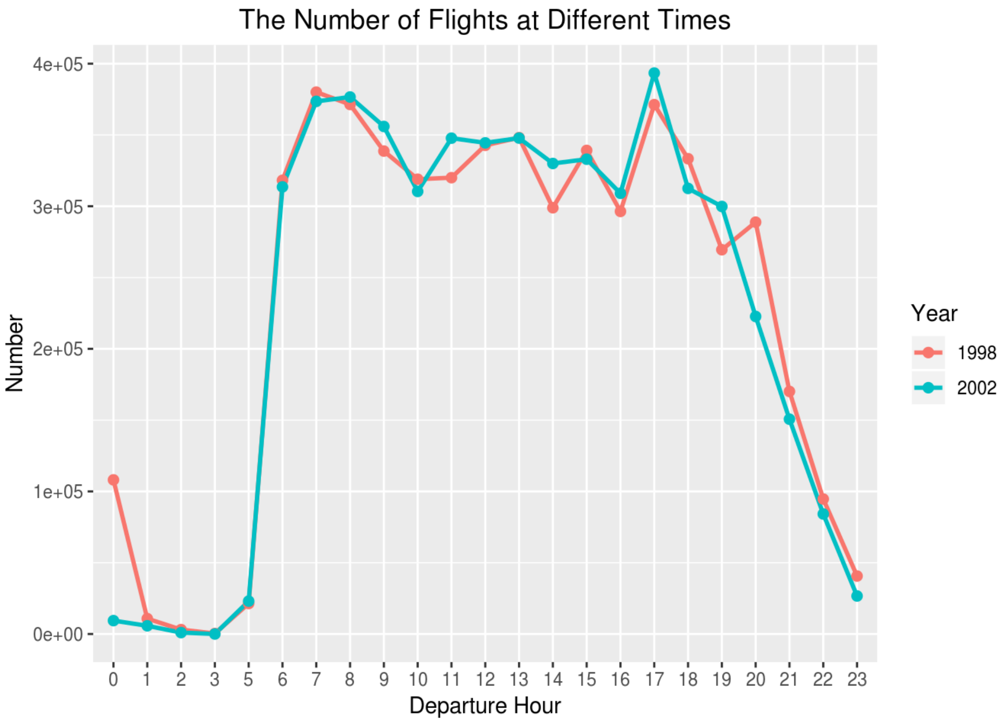
```
hour <- floor(x[,"CRSDepTime"]/100)
hour[which(hour == 24)] <- 0
library(reshape)
```

```
##
## Attaching package: 'reshape'
```

```
## The following object is masked from 'package:data.table':
##
##     melt
```

```
## The following object is masked from 'package:dplyr':
##
##     rename
```

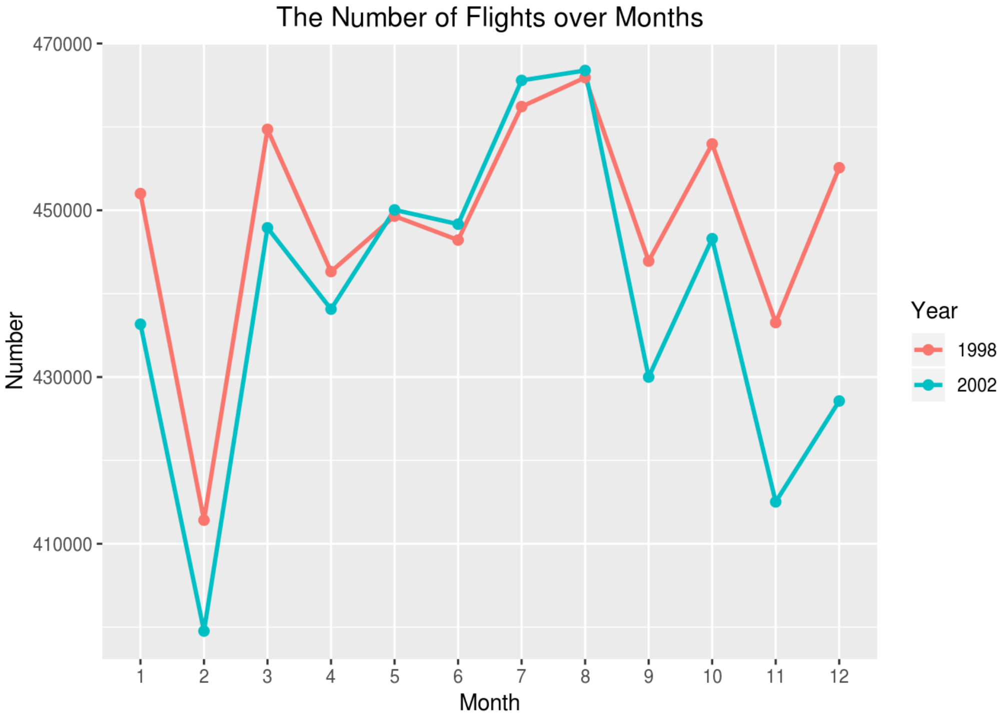
```
hour <- melt(table(hour,x[,"Year"]))
colnames(hour) <- c("Hour","Year","Number")
hour$Year <- as.factor(hour$Year)
ggplot(data = hour,aes(x = as.factor(Hour),y = Number,group = Year, col = Year))+
  geom_line(size = 1)+
  geom_point(size = 2)+
  labs(x="Departure Hour",title ="The Number of Flights at Different Times")+
  theme(plot.title = element_text(hjust = 0.5))
```



The x-axis is the hour of the departure time, the y-axis is the number of flights departed at this time. From the above plot, we conclude that most of the flights departed between 6 am and 7 pm. In 2002, the Flights departed in the daytime is more than that of 1998. Besides, there are fewer flights departed during the evening in 2002.

# The monthly number of flights in two years

```
f1998 <- day.n[which(day.n["Year"] == "1998"),]
f2002 <- day.n[which(day.n["Year"] == "2002"),]
month <- matrix(0,12,2)
colnames(month)<- c("1998","2002")
month[,"1998"] <- tapply(f1998$Number,f1998$Month, sum)
month[,"2002"] <- tapply(f2002$Number,f2002$Month, sum)
month <- melt(month)
colnames(month) <- c("Month","Year","Number")
month$Year <- factor(month$Year)
ggplot(data=month,aes(x = as.factor(Month), y = Number,group = Year, col = Year))+
    geom_line(size = 1)+
    geom_point(size = 2)+
    labs(x ="Month",title ="The Number of Flights over Months")+
    theme(plot.title = element_text(hjust = 0.5))
```

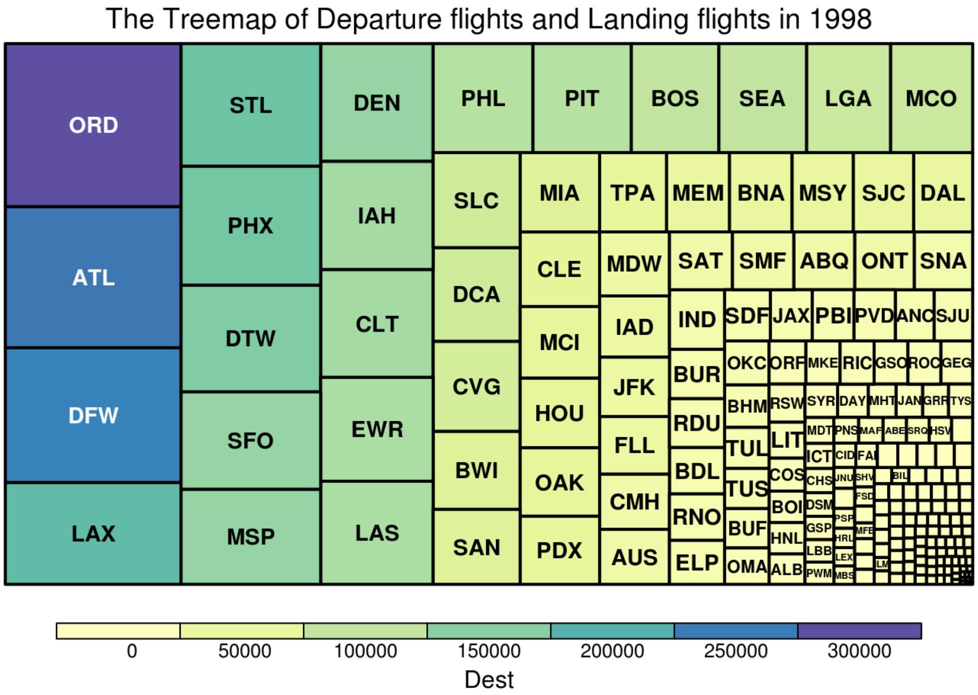


The x-axis of the plot is the month of the flights, and the y-axis denotes the number of flights each month. The trend is similar in these two years. From May and August, the number of flights in 2002 is larger than that of 1998. In other time, We can see the number of flights in 2002 is less than that in 1998.

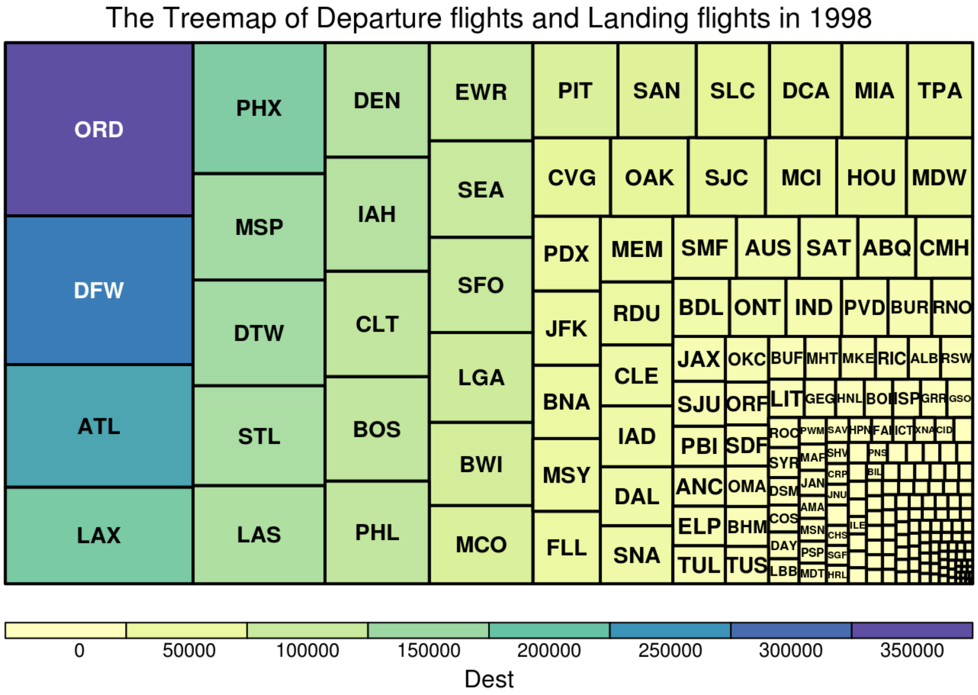
# Advanced Analysis

## The number of departure flights and landing flights for airports in the two years

```
library(treemap)
org98 <- as.data.frame(table(as.character(data[which(data["Year"] == "1998"),"Origin"])))
des98 <- as.data.frame(table(as.character(data[which(data["Year"] == "1998"),"Dest"])))
d98 <- merge(org98,des98, by.x="Var1",by.y="Var1")
colnames(d98) <- c("Name","Origin","Dest")
org02 <- table(as.character(data[which(data["Year"] == "2002"),"Origin"]))
des02 <- table(as.character(data[which(data["Year"] == "2002"),"Dest"]))
d02 <- merge(org02,des02, by.x="Var1",by.y="Var1")
colnames(d02) <- c("Name","Origin","Dest")
treemap(d98,
        index = c("Name"),
        vSize = "Origin",
        vColor = "Dest",
        type = "value",
        palette = "Spectral",
        title = "The Treemap of Departure flights and Landing flights in 1998")
```



```
treemap(d02,
        index = c("Name"),
        vSize = "Origin",
        vColor = "Dest",
        type = "value",
        palette = "Spectral",
        title = "The Treemap of Departure flights and Landing flights in 1998")
```



In the tree plot, the color denotes the number of landing flights, and the size denotes the number of the departure flights in airports. The size and color are similar in these two years. We can see the ORD is the busiest airports in these two years. Compare to the plot of 1998, we can see the size of the blocks in the middle is smaller than the above plot. Although the total flights in 2002 is less than that in 1998, the flights in ORD and DFW is more. So, the flights might cluster in the main airports. We will check that in the following analysis in map.

We can see the number of the departure flights in proportion to that of the landing flights. So, in the following anlysis, we only focus on the destination of the flights.

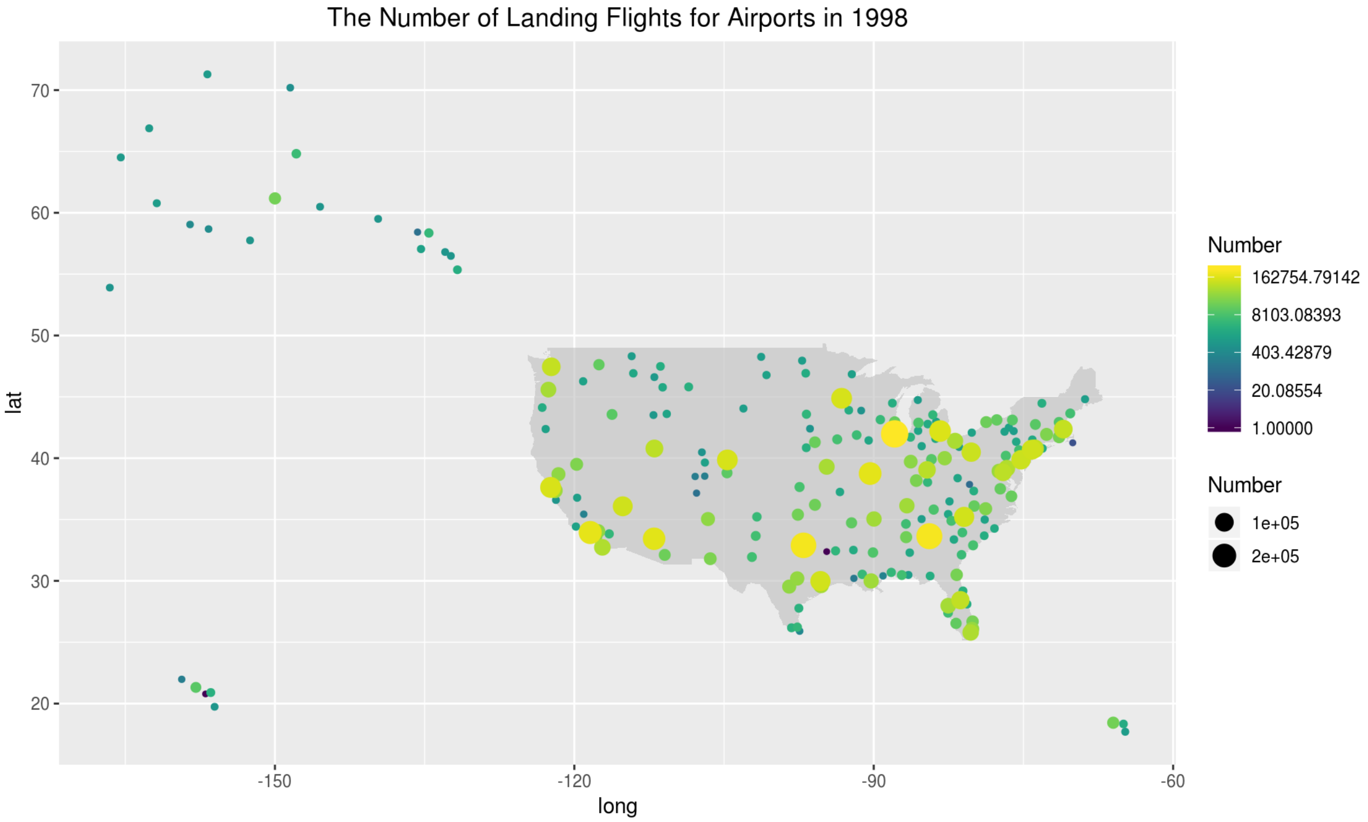
# The number of Des in the two years

> reference https://www.r-graph-gallery.com/330-bubble-map-with-ggplot2/ (https://www.r-graph-gallery.com/330-bubble-map-with-ggplot2/)

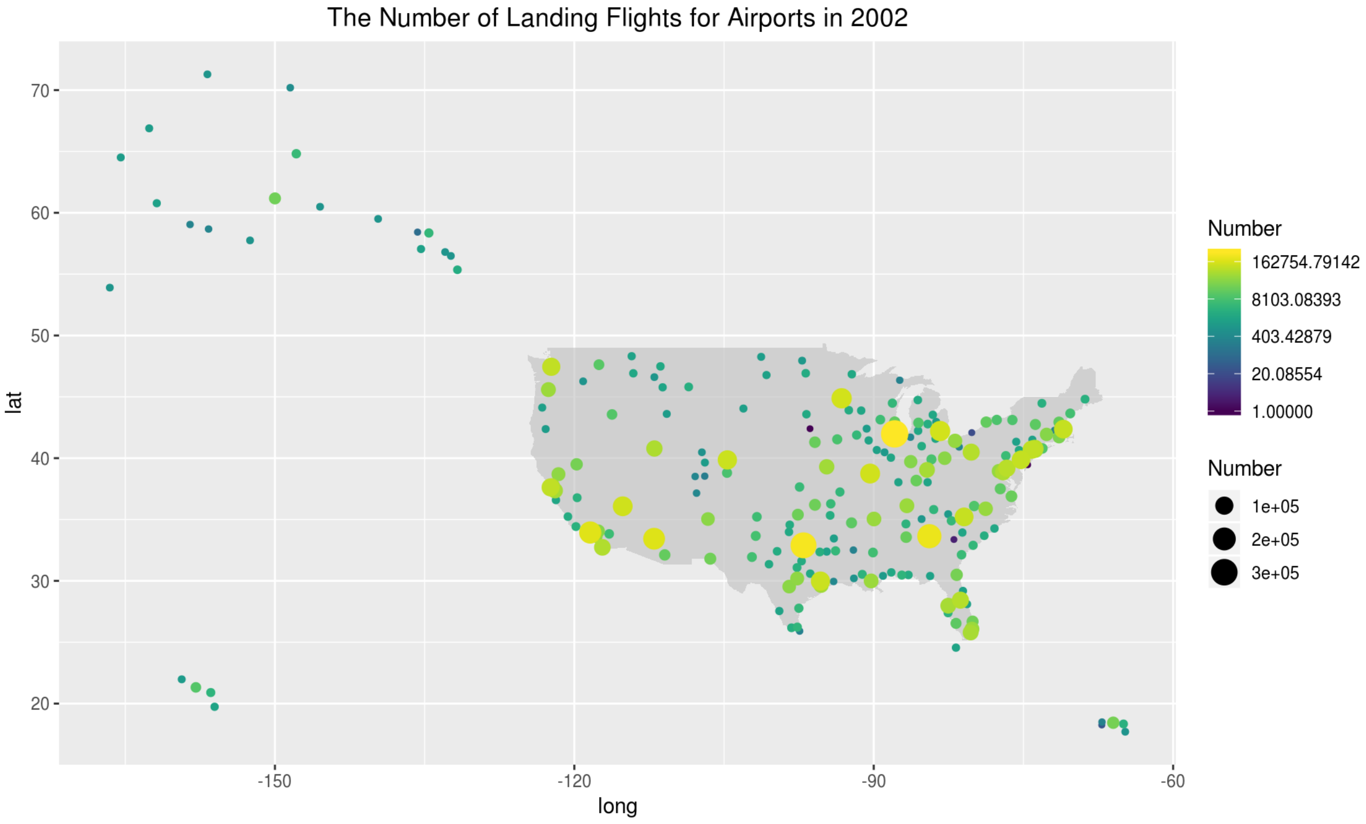
```
dest <- read.csv("dest.csv", header=FALSE, sep="\t" )
colnames(dest) <- c("Year","Dest","Number")
airport <- read.table("airports.csv",header=T, sep=',')
destl <- merge(dest,airport, by.x="Dest", by.y="iata")
destl98 <- destl[which(destl[,"Year"] == 1998),]
destl98<- destl98[order(destl98[,"Number"]),]
destl02 <- destl[which(destl[,"Year"] == 2002),]
destl02 <- destl02[order(destl02[,"Number"]),]

usam <- map_data("usa")
library(maps)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
par(mfrow = c(2,1))
ggplot() +
  geom_polygon(data = usam, aes(x = long, y = lat, group = group), fill = "grey", alpha = 0.6) +
  geom_point(data = destl98, aes(x = long, y = lat, size = Number, color = Number)) +
  scale_size_continuous(range = c(1,6))+
  scale_color_viridis(trans = "log")+
  labs(title = "The Number of Landing Flights for Airports in 1998")+
  theme(plot.title = element_text(hjust = 0.5))
```



```
ggplot() +
  geom_polygon(data = usam, aes(x = long, y = lat, group = group), fill = "grey", alpha = 0.6) +
  geom_point(data = destl02, aes(x = long, y = lat, size = Number, color = Number)) +
  scale_size_continuous(range = c(1,6))+
  scale_color_viridis(trans = "log")+
  labs(title = "The Number of Landing Flights for Airports in 2002")+
  theme(plot.title = element_text(hjust = 0.5))
```



# Netplot

## Data Preparation

> refrence https://kateto.net/wp-content/uploads/2018/06/Polnet%202018%20R%20Network%20Visualization%20Workshop.pdf (https://kateto.net/wp-content/uploads/2018/06/Polnet%202018%20R%20Network%20Visualization%20Workshop.pdf)

```
airport <- airport[which(airport$iata %in% union(data$Origin,data$Dest)),]
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

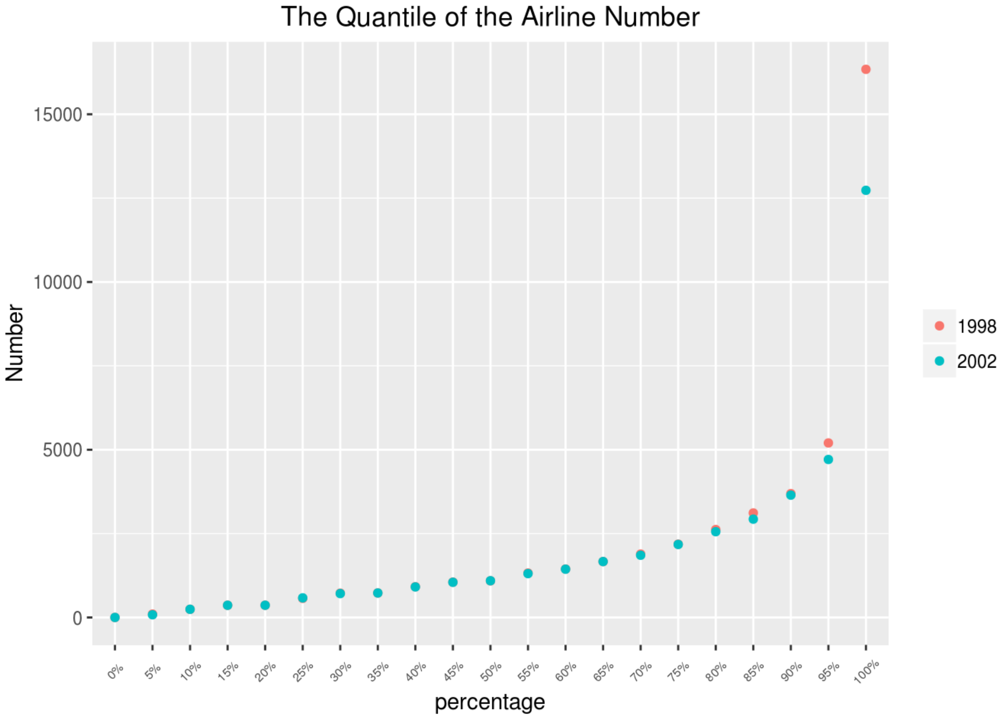
```
## The following objects are masked from 'package:reshape':
##
##     colsplit, melt, recast
```

```
## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
divert <- melt(table(data$Year,data$Origin,data$Dest))
divert <- divert[which(divert$value!=0),]
colnames(divert) <- c("Year","From","To","Number")

divert1998 <- divert[which(divert$Year== "1998"),]
divert2002 <- divert[which(divert$Year== "2002"),]

##generate a color gradient
library(grDevices)
col.1 <- adjustcolor("turquoise2", alpha=0.4)
col.2 <- adjustcolor("orange", alpha=0.4)
edge.pal <- colorRampPalette(c(col.1, col.2), alpha = TRUE)
edge.col <- edge.pal(100)
```

## Select the principal airports

```
## Plot the quantile of the two years
myProbs <- seq(0, 1, by=0.05)
q1998 <- quantile(divert1998[,"Number"],myProbs,na.rm=TRUE)
q2002 <- quantile(divert2002[,"Number"],myProbs,na.rm=TRUE)
divertq <- as.matrix(cbind(q1998,q2002), rownames=myProbs)
colnames(divertq) <- c("1998","2002")
mdivert <- melt(divertq)
library(ggplot2)
ggplot(mdivert, aes(x=Var1, y=value, color=as.factor(mdivert$Var2), group=as.factor(mdivert$Var2))) +
  geom_point() +
  theme(plot.title = element_text(hjust = 0.5), legend.title = element_blank(), axis.text.x = element_text(size =
6, angle = 45, hjust = 0.5, vjust = 0.5))+
  xlab("percentage") + ylab("Number") + labs(title = "The Quantile of the Airline Number")
```

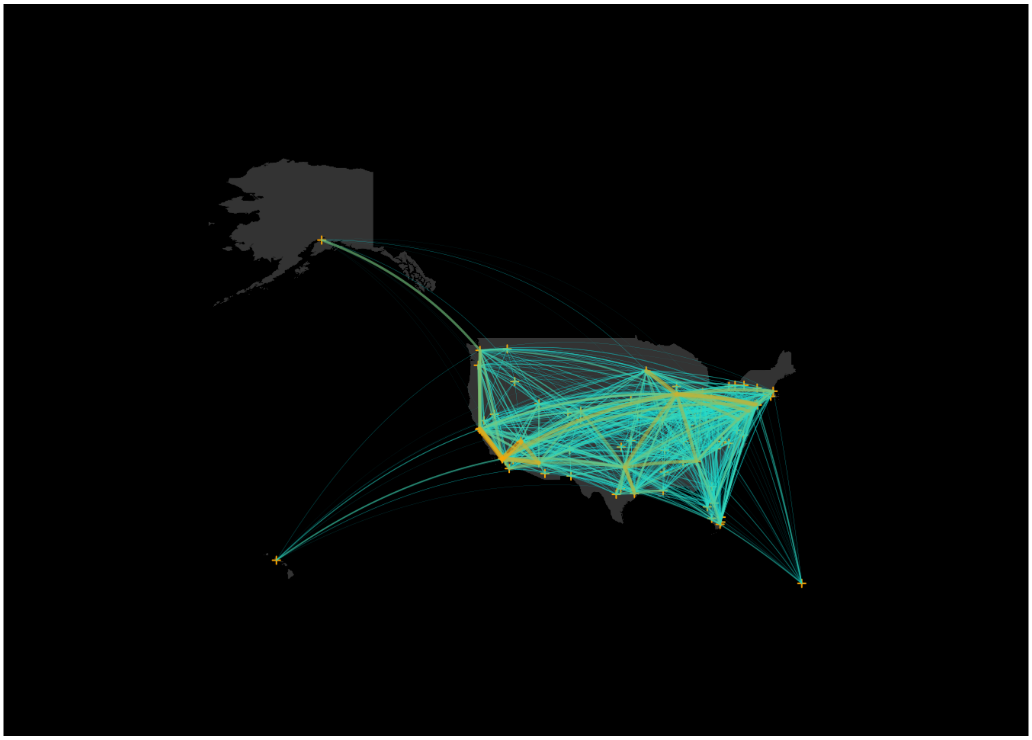


```
##By the quntile, we can find the busier and less busy airlines
divert1998 <- divert1998[order(divert1998[,"Number"]),]
divert2002 <- divert2002[order(divert2002[,"Number"]),]
dvt98lb <- divert1998[c(1:max(which(divert1998[,"Number"]<3601))),]
dvt02lb <- divert2002[c(1:max(which(divert1998[,"Number"]<3601))),]
dvt98mb <- divert1998[-c(1:max(which(divert1998[,"Number"]<5203))),]
dvt02mb <- divert2002[-c(1:max(which(divert1998[,"Number"]<4710))),]
```

## Plot the Net on map

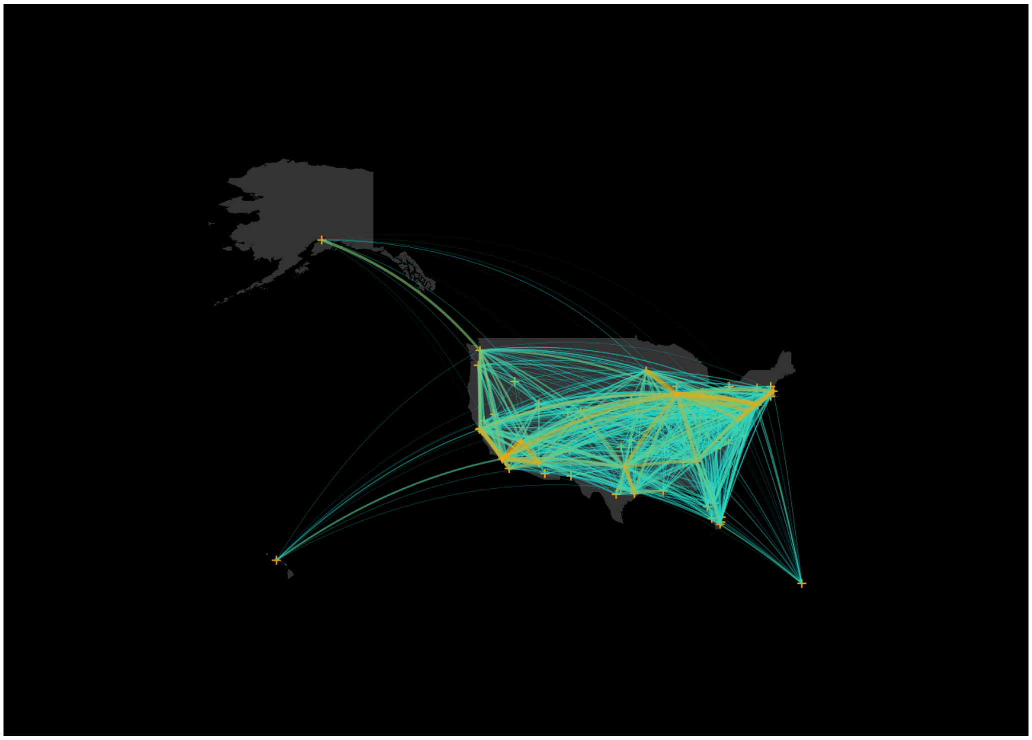
The greneral situation in 1998

```
id <- table(divert1998$From)
bid <- names(id)[id >10]
airports1998 <- airport[airport$iata %in% bid,]
flight1998 <- divert1998[divert1998$From %in% bid &
                           divert1998$To %in% bid, ]
airports1998[,"iata"] <- factor(airports1998[,"iata"])

##create basemap
library(maps)
map("world", regions=c("usa"), col="grey20", fill=TRUE, bg="black", lwd=0.1, ylim=c(15.0,75.0), xlim=c(-169.0,-6
3.0))
points(airports1998$long,airports1998$lat, pch=3, cex=0.5, col="orange")
## install.packages("geosphere")
library(geosphere)
## add lines on the map
for(i in 1:nrow(flight1998)) {
  node1 <- airports1998[as.character(airports1998$iata) == as.character(flight1998[i,]$From),]
  node2 <- airports1998[as.character(airports1998$iata) == as.character(flight1998[i,]$To),]
  arc <- gcIntermediate( c(node1[1,]$long, node1[1,]$lat),
                         c(node2[1,]$long, node2[1,]$lat),
                         n=1000, addStartEnd=TRUE )
  edge.ind <- round(100*flight1998[i,]$Number / max(flight1998$Number))
  lines(arc, col=edge.col[edge.ind], lwd=edge.ind/30)
}
```



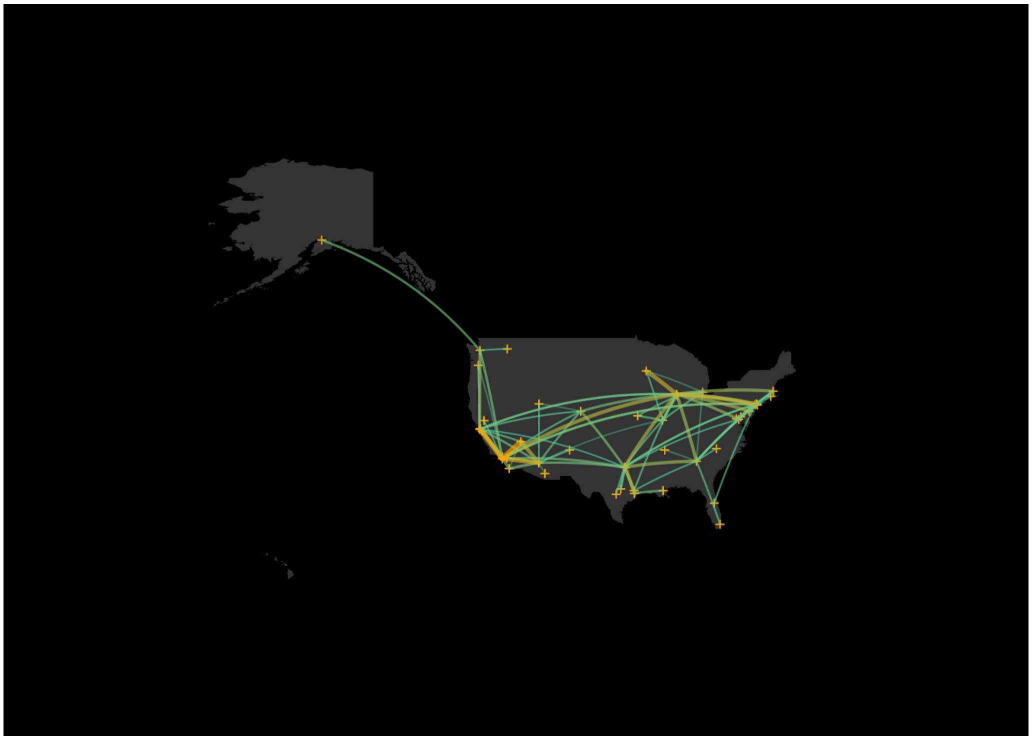
The general situation in 2002

```
id <- table(divert2002$From)
bid <- names(id)[id >10]
airports2002 <- airport[airport$iata %in% bid,]
flight2002 <- divert2002[divert2002$From %in% bid &
                           divert2002$To %in% bid, ]
airports2002[,"iata"] <- factor(airports2002[,"iata"])

##create basemap
library(maps)
map("world", regions=c("usa"), col="grey20", fill=TRUE, bg="black", lwd=0.1, ylim=c(15.0,75.0), xlim=c(-169.0,-6
3.0))
points(airports2002$long,airports2002$lat, pch=3, cex=0.5, col="orange")

library(geosphere)
## add lines on the map
for(i in 1:nrow(flight2002)) {
  node1 <- airports2002[as.character(airports2002$iata) == as.character(flight2002[i,]$From),]
  node2 <- airports2002[as.character(airports2002$iata) == as.character(flight2002[i,]$To),]
  arc <- gcIntermediate( c(node1[1,]$long, node1[1,]$lat),
                         c(node2[1,]$long, node2[1,]$lat),
                         n=1000, addStartEnd=TRUE )
  edge.ind <- round(100*flight2002[i,]$Number / max(flight2002$Number))
  lines(arc, col=edge.col[edge.ind], lwd=edge.ind/30)
}
```



The more busy flights in 1998

```
airports98mb <- airport[airport$iata %in% union(dvt98mb[,"From"], dvt98mb[,"To"]),]
airports98mb[,"iata"] <- factor(airports98mb[,"iata"])

library(maps)
map("world", regions=c("usa"), col="grey20", fill=TRUE, bg="black", lwd=0.1, ylim=c(15.0,75.0), xlim=c(-169.0,-6
3.0))
points(airports98mb $long,airports98mb $lat, pch=3, cex=0.5, col="orange")
library(geosphere)
## add lines on the map
for(i in 1:nrow(dvt98mb)) {
  node1 <- airports98mb[as.character(airports98mb$iata) == as.character(dvt98mb[i,]$From),]
  node2 <- airports98mb[as.character(airports98mb$iata) == as.character(dvt98mb[i,]$To),]
  arc <- gcIntermediate( c(node1[1,]$long, node1[1,]$lat),
                         c(node2[1,]$long, node2[1,]$lat),
                         n=1000, addStartEnd=TRUE )
  edge.ind <- round(100*dvt98mb[i,]$Number / max(dvt98mb$Number))
  lines(arc, col=edge.col[edge.ind], lwd=edge.ind/30)
}
```



The more busy case in 2002

```
airports02mb <- airport[airport$iata %in% union(dvt02mb[,"From"], dvt02mb[,"To"]),]
airports02mb[,"iata"] <- factor(airports02mb[,"iata"])

library(maps)
map("world", regions=c("usa"), col="grey20", fill=TRUE, bg="black", lwd=0.1, ylim=c(15.0,75.0), xlim=c(-169.0,-6
3.0))
points(airports02mb$long,airports02mb$lat, pch=3, cex=0.5, col="orange")
library(geosphere)

## add lines on the map
for(i in 1:nrow(dvt02mb)) {
  node1 <- airports02mb[as.character(airports02mb$iata) == as.character(dvt02mb[i,]$From),]
  node2 <- airports02mb[as.character(airports02mb$iata) == as.character(dvt02mb[i,]$To),]
  arc <- gcIntermediate( c(node1[1,]$long, node1[1,]$lat),
                         c(node2[1,]$long, node2[1,]$lat),
                         n=1000, addStartEnd=TRUE )
  edge.ind <- round(100*dvt02mb[i,]$Number / max(dvt02mb$Number))
  lines(arc, col=edge.col[edge.ind], lwd=edge.ind/30)
}
```