Air Traffic Data (bureau of Transportation Statistics)

Nicha Ruchirawat

July 24, 2017

```
# read 28 .csv files into one data frame
files <- dir("~/Desktop/q3_data", pattern = "\\.csv", full.names = TRUE)
airplane_data <- do.call(rbind, lapply(files, read.csv))
```

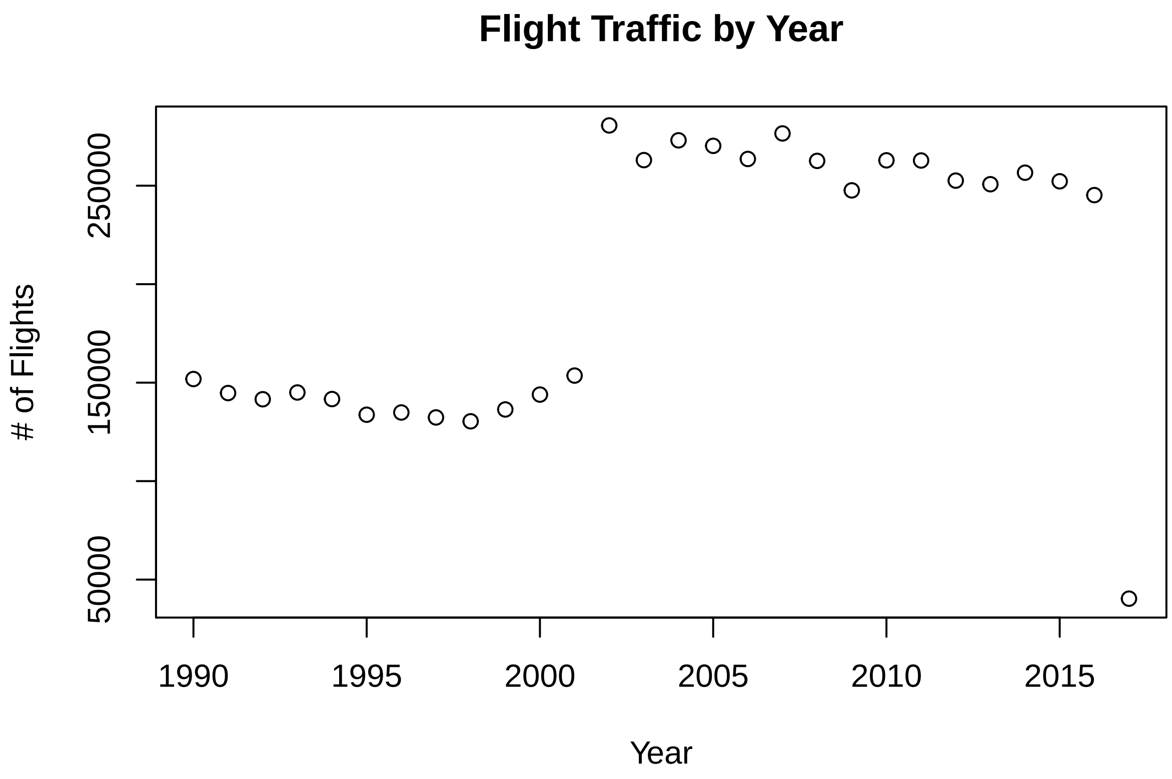
```
# load all necessary libraries
library(dplyr)
```

```
##
## 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
```

```
library(magrittr)
library(lattice)
```

We will first investigate into high-level trends over time over the years 1990-2017 for air traffic.

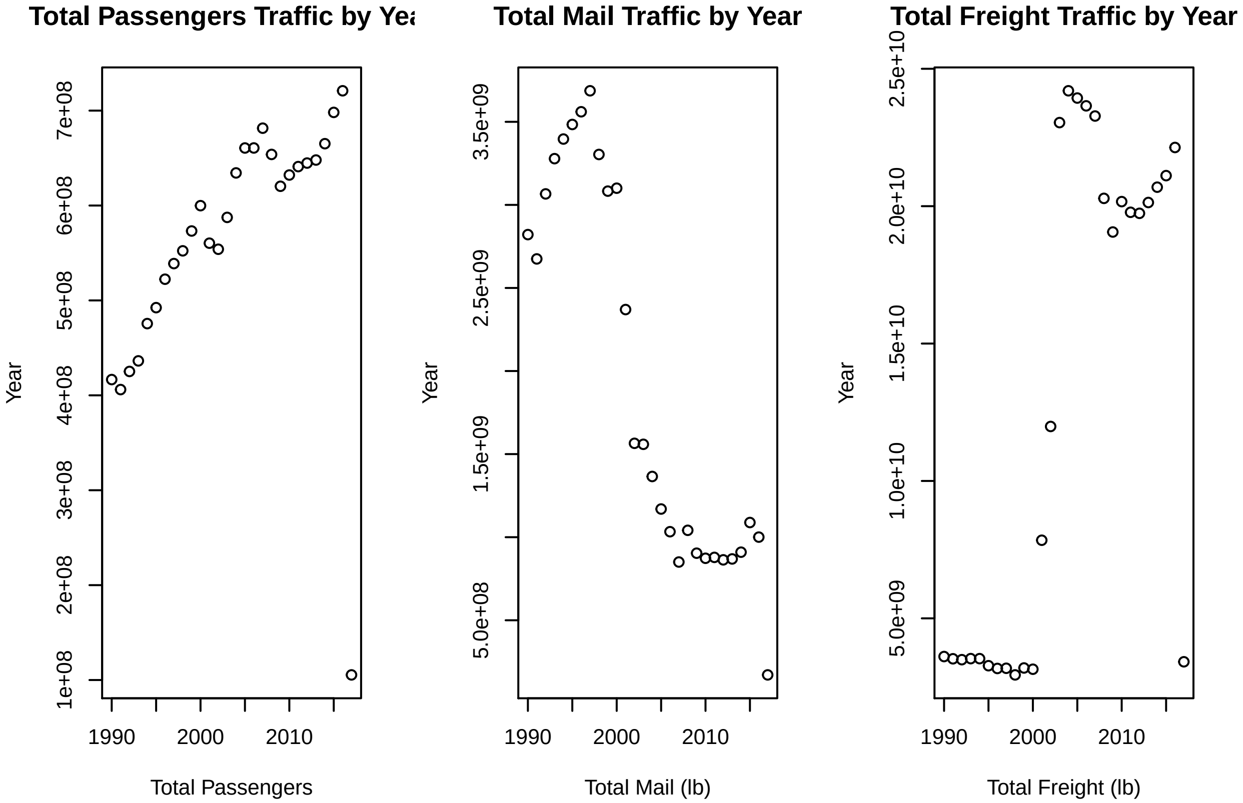
```
# create data frame that accumulate # of flights by year
flightbyyear <- airplane_data %>% group_by(YEAR) %>% count(.)

# plot number of flights by year
plot(flightbyyear$YEAR, flightbyyear$n, main = "Flight Traffic by Year",
    xlab = "Year", ylab = "# of Flights")
```



There is a big jump in air traffic between year 2001 and 2002, with a stable but slightly decreasing trend in the following years.

We will then explore how this trend in air traffic translates into total passengers, mail, and freight, traffic by year:

```r
# sum number of passengers by year
passengerByYear <- airplane_data %>% group_by(YEAR) %>% summarise(totalPassengers = sum(PASSENGERS))
# sum lbs of mail by year
mailByYear <- airplane_data %>% group_by(YEAR) %>% summarise(totalMail = sum(MAIL))
# sum lbs of freight by year
freightByYear <- airplane_data %>% group_by(YEAR) %>% summarise(totalFreight = sum(FREIGHT))
# create 3 side-by-side plot space
par(mfrow = c(1, 3))
# plot passengers by year
plot(passengerByYear$YEAR, passengerByYear$totalPassengers, xlab = "Total Passengers",
    ylab = "Year", main = "Total Passengers Traffic by Year")
# plot mail by year
plot(mailByYear$YEAR, mailByYear$totalMail, xlab = "Total Mail (lb)",
    ylab = "Year", main = "Total Mail Traffic by Year")
# plot freight by year
plot(freightByYear$YEAR, freightByYear$totalFreight, xlab = "Total Freight (lb)",
    ylab = "Year", main = "Total Freight Traffic by Year")
```



The Passengers Traffic plot shows that passengers traffic have been steadily increasing over the years, with downward trends only during 2000-2002 and 2007-2009.

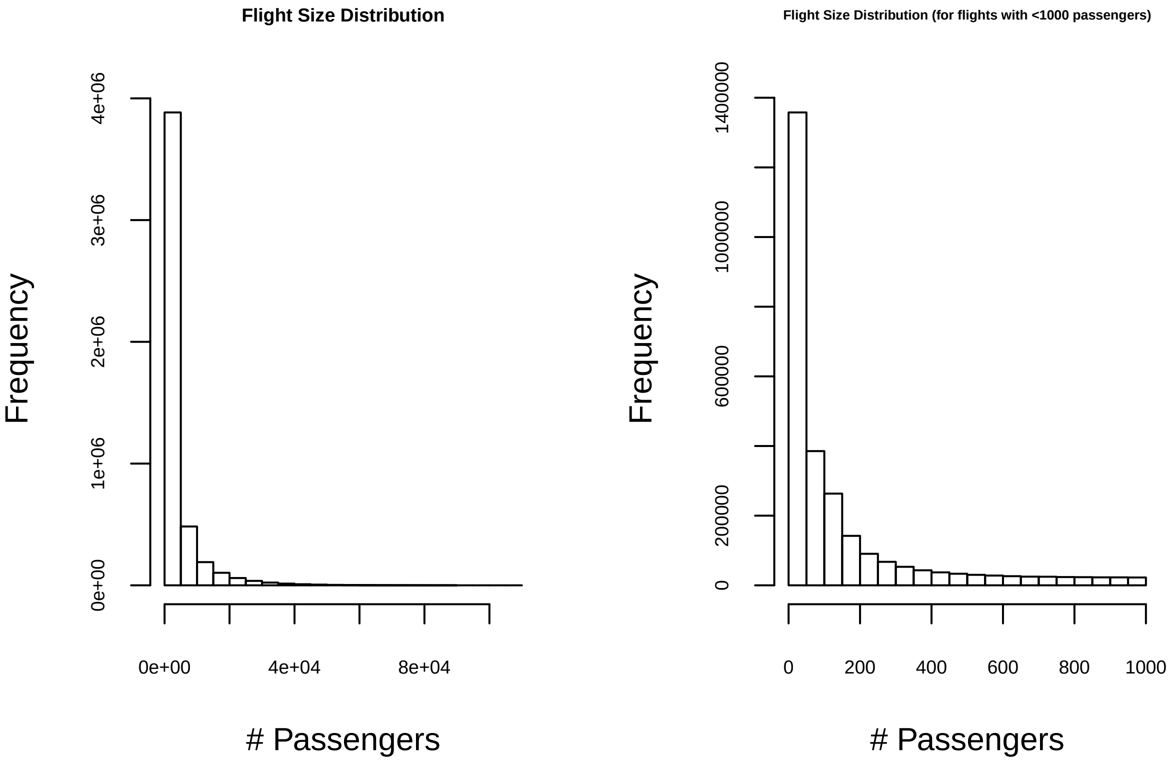
The Mail Traffic plot shows that mail traffic significantly dropped from 1997 onwards. This could be explained by the boom of the internet and rise of electronic mail.

The Freight Traiffic plot shows that there is an increase in freight volume from 2000, which stabilized from around 2004, with a slight drop during 2007-2009.

However, when compared with the total flight traffic plot, it seems that the big rise in flight traffic between 2001-2002 did not align with decrease in passenger traffic nor mail traffic in the same period. This could mean that there is a rise in smaller flight size with more planes carrying less people, or this rise in flight traffic could be attributed to rise in planes carrying freight.

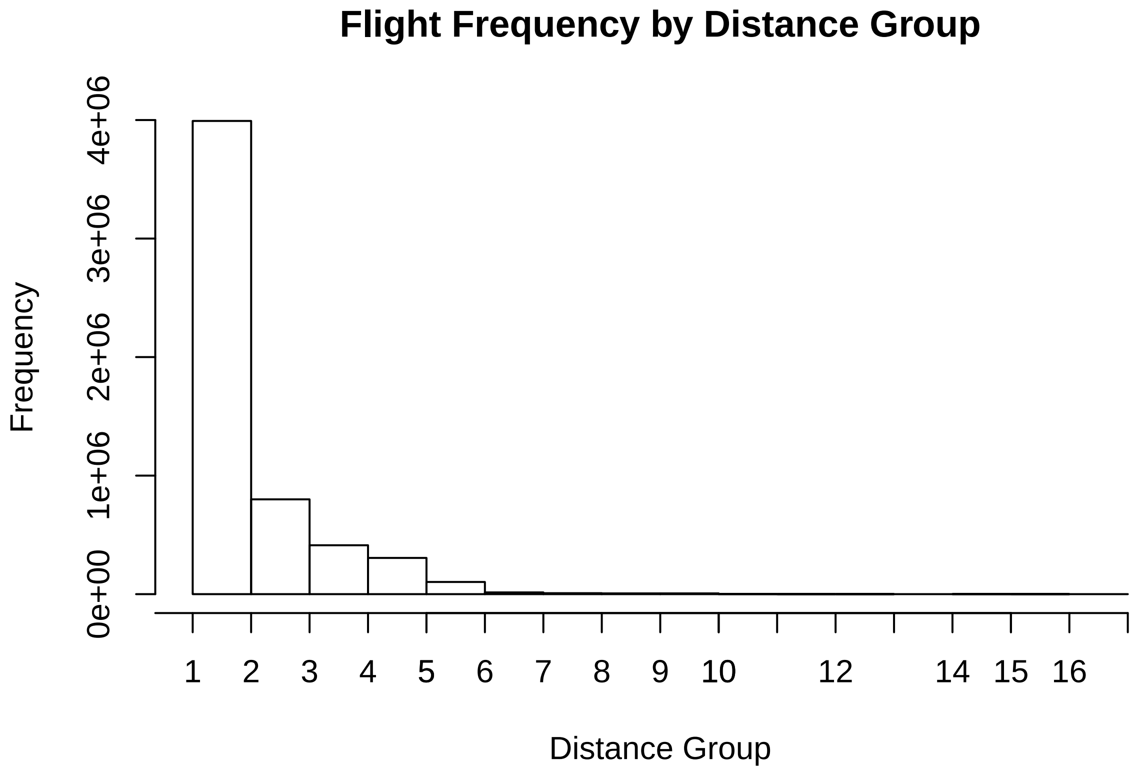
```
# create data frame of passenger flights by filtering out those
# with no passengers
passengerflight <- airplane_data[airplane_data$PASSENGERS > 0, ]
# create data frame of passenger flights by filtering out those
# with no passengers and subsetting only those with less than 1000
# passengers
passengerflight2 <- airplane_data[airplane_data$PASSENGERS > 0 & airplane_data$PASSENGERS <
    1000, ]

# create 2 by 2 space for histogram
par(mfrow = c(1, 2))
# create histogram for flight size distribution
hist(passengerflight$PASSENGERS, main = "Flight Size Distribution",
    xlab = "# Passengers", ylab = "Frequency", cex.main = 0.6, cex.axis = 0.6)
# create histogram for flight size distribution for flight size
# with less than 1000 passengers
```

```r
hist(passengerflight2$PASSENGERS, main = "Flight Size Distribution (for flights with <1000 passengers)"
    xlab = "# Passengers", ylab = "Frequency", cex.main = 0.4, cex.axis = 0.6)
```

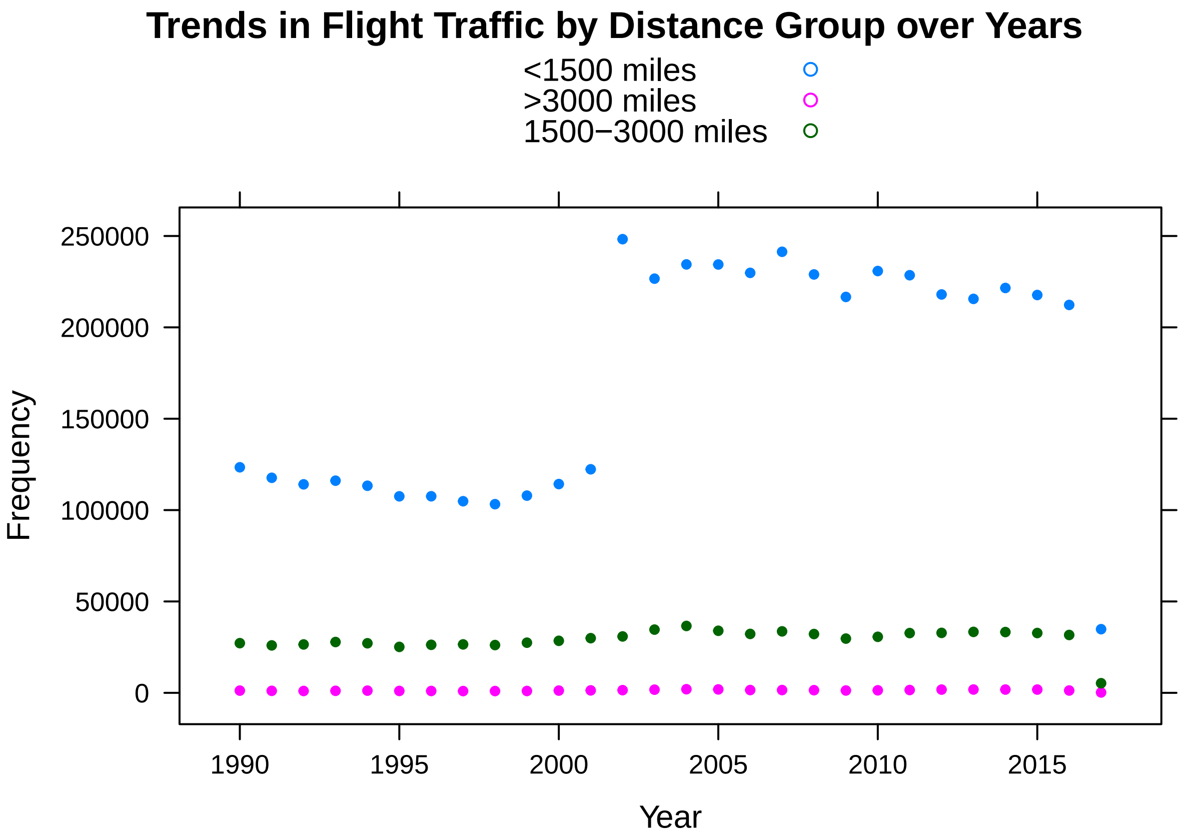


Histogram shows that flights carrying passengers mostly carries <50 passengers, followed by flights carrying passengers with 50-400 passengers.

```r
# get frequency data for flight traffic by distance group
dist_group <- airplane_data %>% group_by(DISTANCE_GROUP) %>% count(.)

# histogram for flight frequency by distance group
hist(airplane_data$DISTANCE_GROUP, breaks = 16, xlim = c(1, 17), xlab = "Distance Group",
    ylab = "Frequency", main = "Flight Frequency by Distance Group")
axis(side = 1, at = seq(0, 17, 1), labels = seq(0, 17, 1))
```



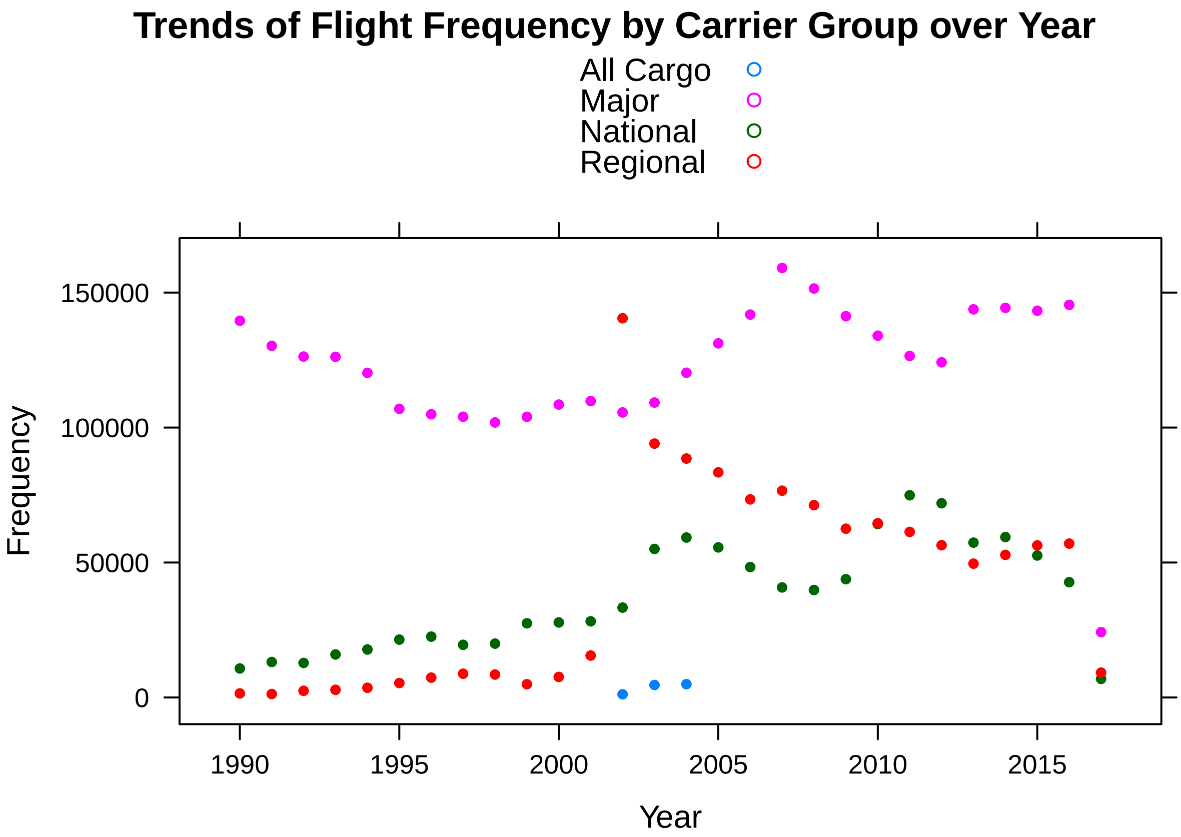
```
# group flight traffic data by year and distance group and its
# relative frequency
dist_group_byYear <- airplane_data %>% group_by(YEAR, DISTANCE_GROUP) %>%
    count(.)

# create smaller number of bins for distance group into 3 main
# groups
dist_new <- dist_group_byYear %>% mutate(dist = replace(DISTANCE_GROUP,
    DISTANCE_GROUP < 4, "<1500 miles")) %>% mutate(dist = replace(dist,
    DISTANCE_GROUP > 3 && DISTANCE_GROUP <= 6, "1500-3000 miles")) %>%
    mutate(dist = replace(dist, DISTANCE_GROUP > 6, ">3000 miles"))

# summarize total flight traffic by year and newly create distance
# group bins
dist_newgroup_byYear <- dist_new %>% group_by(YEAR, dist) %>% summarise(total = sum(n))

# plot trends of the 3 newly created distance group bins
xyplot(total ~ YEAR, dist_newgroup_byYear, groups = dist, pch = 20,
    auto.key = T, main = "Trends in Flight Traffic by Distance Group over Years",
    xlab = "Year", ylab = "Frequency")
```



The histogram shows that most flights are in the shorter distance, so the distance groups are simplified to 3 main groups: <1500 miles, 1500-3000 miles, and >3000 miles.

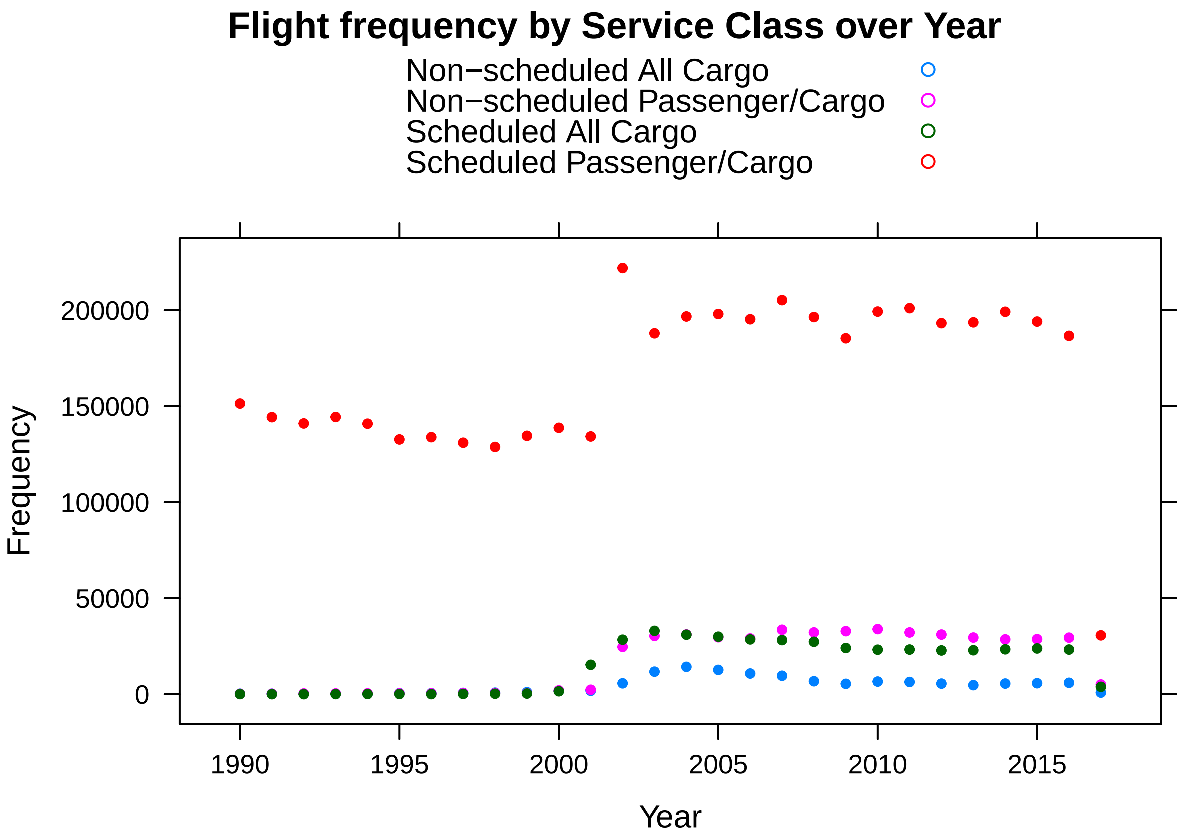
The scatter plot shows that there is a jump in short distance flights (<1500 miles) from year 2001-2002, which stayed high onwards. Mid-range and long distance flights frequency remained stable. Overall, short flights <1500 miles are most frequent, followed by flights between 1500-3000 miles and flights >3000 miles.

```
# isolate data for year and carrier group
carrier_group_frame <- data.frame(airplane_data$YEAR, airplane_data$CARRIER_GROUP)

# count frequency of flights by carrier group
carrierByYear <- carrier_group_frame %>% group_by(airplane_data.YEAR,
    airplane_data.CARRIER_GROUP) %>% count(.)

# label carrier group
carrier_new <- carrierByYear %>% mutate(carrier_group = replace(airplane_data.CARRIER_GROUP,
    airplane_data.CARRIER_GROUP == 1, "Regional")) %>% mutate(carrier_group = replace(carrier_group,
    airplane_data.CARRIER_GROUP == 2, "National")) %>% mutate(carrier_group = replace(carrier_group,
    airplane_data.CARRIER_GROUP == 3, "Major")) %>% mutate(carrier_group = replace(carrier_group,
    airplane_data.CARRIER_GROUP == 7, "All Cargo"))

# plot carrier group flight frequency over the years
xyplot(n ~ airplane_data.YEAR, carrier_new, groups = carrier_group,
    pch = 20, auto.key = T, main = "Trends of Flight Frequency by Carrier Group over Year",
    xlab = "Year", ylab = "Frequency")
```



Flight traffic is most dominated by Major airlines, with the exception of year 2002, which faced a big jump in Regional airlines flights. Regional airlines became more frequent than National airlines after year 2002 but steadily decreased to be at similar levels after year 2010.

```
# group flight traffic by year and service class
service_class_byYear <- airplane_data %>% group_by(YEAR, CLASS) %>%
    count(.)
# convert service class data from factor to character
service_class_byYear$CLASS <- as.character(service_class_byYear$CLASS)
# label service classes
serviceclass_new <- service_class_byYear %>% mutate(service_group = replace(CLASS,
    CLASS == "F", "Scheduled Passenger/Cargo")) %>% mutate(service_group = replace(service_group,
    CLASS == "G", "Scheduled All Cargo")) %>% mutate(service_group = replace(service_group,
    CLASS == "L", "Non-scheduled Passenger/Cargo")) %>% mutate(service_group = replace(service_group,
    CLASS == "P", "Non-scheduled All Cargo"))

# plot service class frequency over the years
xyplot(n ~ YEAR, serviceclass_new, groups = service_group, pch = 20,
    auto.key = T, main = "Flight frequency by Service Class over Year",
    xlab = "Year", ylab = "Frequency")
```



There was a big jump in frequency of scheduled passenger/cargo flight between the years 2001-2002, which remained high onwards. There is also a rise in scheduled all cargo flights and non-scheduled passenger/cargo flights from year 2000, which stabilized after year 2003.

Key Findings in Flight Traffic over Years 1990-2017

1. Flight traffic increased significantly from year 2001-2002 and stabilized at the higher level ever since in the following years after 2002.
2. The big jump in flight traffic between years 2001-2002 follows the same trend as and may be driven by:
   - Sharp increase in scheduled passenger/cargo flights in that period
   - Sharp increase in short distance flights <1500 miles in that period
   - Sharp increase in freight traffic
   - Sharp increase in Regional flights
3. The generally higher flight frequency following 2002 onwards follows the same trend as and may be driven by:
   - Increase in <1500 mile flights
   - Increase in all carrier group flights, mostly in Major airline flights
   - Increase in all service class flights, mostly in scheduled passenger/cargo flights, with some increase in non-scheduled passenger/cargo and scheduled all cargo flights

Top 5 carriers in passenger and passenger growth in recent years (2010-2016):

```r
# filter data after year 2010, sum passengers by carrier and year
carrier_data <- airplane_data %>% filter(YEAR >= 2010) %>% group_by(UNIQUE_CARRIER_NAME) %>%
    summarise(totalPassenger = sum(PASSENGERS)) %>% arrange(desc(totalPassenger))
# output table of top 5 carriers
knitr::kable(carrier_data[1:5, ])
```

UNIQUE_CARRIER_NAME	totalPassenger
Southwest Airlines Co.	884093143
Delta Air Lines Inc.	734707919
American Airlines Inc.	554695427
United Air Lines Inc.	434583097
US Airways Inc.	264382952

```r
# filter data for year 2010 and 2016, sum passengers by carrier
# and year
carrier_byPassenger <- airplane_data %>% filter(YEAR == 2010 | YEAR ==
    2016) %>% group_by(YEAR, UNIQUE_CARRIER_NAME) %>% summarise(totalPassenger = sum(PASSENGERS))

# reshape to turn 2010 and 2016 rows into columns
casted_carrier <- reshape::cast(carrier_byPassenger, UNIQUE_CARRIER_NAME ~
    YEAR, mean)
```

```
## Using totalPassenger as value column.  Use the value argument to cast to override this choice
```

```r
# remove carriers that did not exist in 2010 to not divide by 0
casted_carrier <- casted_carrier[casted_carrier$"2010" != 0, ]
# calculate growth in column 'growth'
casted_carrier$growth <- (casted_carrier$"2016"/casted_carrier$"2010") *
    100
# sort by growth rate
sorted_carrier <- casted_carrier[order(-casted_carrier$growth), ]
colnames(sorted_carrier) <- c("UNIQUE_CARRIER_NAME", "2010", "2016",
    "Growth %")
# output table of top 5
knitr::kable(sorted_carrier[1:5, c(1, 4)])
```

	UNIQUE_CARRIER_NAME	Growth %
114	National Air Cargo Group Inc d/ba National Airlines	864850.0000
107	Caribbean Sun Airlines, Inc. d/b/a World Atlantic Airlines	1309.3653
80	Kalinin Aviation LLC d/b/a Alaska Seaplanes	900.0763
133	Via Airlines d/b/a Charter Air Transport	865.6848
126	Multi-Aero, Inc. d/b/a Air Choice One	698.1595
#Top 5	carriers by passengers carried between 2010-2016:#	
1. Sou	thwest Airlines	
2. Del	ta Airlines	
3. Ame	rican Airlines	
4. Uni	ted Airlines	
5. US	Airways	

Top 5 carriers growth by passengers carried between 2010-2016:

1. National Air Cargo Group
2. Caribbean Sun Airlines
3. Kalinin Aviation LLC
4. Via Airlines Charter Air Transport
5. Multi-Aero, Inc.

Top 5 routes in passenger in recent years (2010-2016):

```r
# isolate airport, year, and passenger data
origin_dest <- data.frame(airplane_data$ORIGIN_AIRPORT_ID, airplane_data$DEST_AIRPORT_ID,
    airplane_data$YEAR, airplane_data$PASSENGERS)
# merge origin and destination ID to form column identifying route
origin_dest$origindest <- paste(origin_dest$airplane_data.ORIGIN_AIRPORT_ID,
    origin_dest$airplane_data.DEST_AIRPORT_ID, sep = "-")
# sum passengers for each route
origin_dest_freq <- origin_dest %>% group_by(origindest) %>% summarise(totalPassenger = sum(airplane_da
    arrange(desc(totalPassenger))
colnames(origin_dest_freq) <- c("Route", "# flights")
# output frequency table
knitr::kable(origin_dest_freq[1:5, ])
```

Route	# flights
12892-14771	36374459
14771-12892	36285667
13830-12173	35991658
12173-13830	34780078
12892-12478	32900107

```r
# filter passenger sum for each route for year 2010 and 2016
route_byPassenger <- origin_dest %>% filter(airplane_data.YEAR ==
    2010 | airplane_data.YEAR == 2016) %>% group_by(airplane_data.YEAR,
    origindest) %>% summarise(totalPassenger = sum(airplane_data.PASSENGERS))
# reshape data to turn 2010 and 2016 rows into columns
casted_route <- reshape::cast(route_byPassenger, origindest ~ airplane_data.YEAR,
    mean)
```

```
## Using totalPassenger as value column.  Use the value argument to cast to override this choice
```

```r
# remove flights that did not exist in 2010 to avoid dividing by 0
casted_route <- casted_route[casted_route$"2010" != 0, ]
# calculate growth in 'growth' column
casted_route$growth <- (casted_route$"2016"/casted_route$"2010") *
    100
# sort by growth
sorted_route <- casted_route[order(-casted_route$growth), ]
colnames(sorted_route) <- c("Route", "2010", "2016", "Growth %")
# output table of top 5
knitr::kable(sorted_route[1:5, c(1, 4)])
```

	Route	Growth %
21056	13232-11986	8241500
9289	11298-11413	4913000
8680	11267-13232	3948500
6992	11042-13577	2536400
14531	12191-10994	1684850

Top 5 routes in passenger between 2010-2016 (according to data key):

1. LAX - SFO
2. SFO - LAX
3. Kahului Airport - Honolulu Intl Airport
4. Honolulu Intl Airport - Kahului Airport
5. LAX - JFK

Top 5 routes growth in passenger between 2010-2016 (according to data key):

1. Chicago O'Hare - Grand Rapids, MI: Gerald R. Ford International
2. Dallas/Fort Worth International - Durango, CO: Durango La Plata County
3. Dayton, OH: James M Cox/Dayton International - Chicago O'Hare
4. Carlsbad, CA: McClellan-Palomar - Myrtle Beach, SC: Myrtle Beach International
5. Houston, TX: William P Hobby - Charleston, SC: Charleston AFB/International