Airline Data Analysis

Stephen H.

Part 1 Load the Packages

```
library(ggplot2)
library(tidyverse)
```

```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------
--------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
abluetheme <- theme(plot.background = element_rect(fill = "lightblue",
 colour = "black", size = 2, linetype = "solid"), legend.background=el
ement_rect(colour = "black", size = 1, linetype = "solid"), panel.back
ground=element_rect(colour = "black", size = 2, linetype = "solid"), p
lot.title=element_text(size=15))
```

Part 2 Read in the File

```
filenames <- list.files("/Users/shsu/Downloads/q3_data/", pattern="*.c
sv", full.names=TRUE)
airlineData <- do.call(rbind,lapply(filenames, read.csv))
```

Part 3 Top 10 For Entertainment

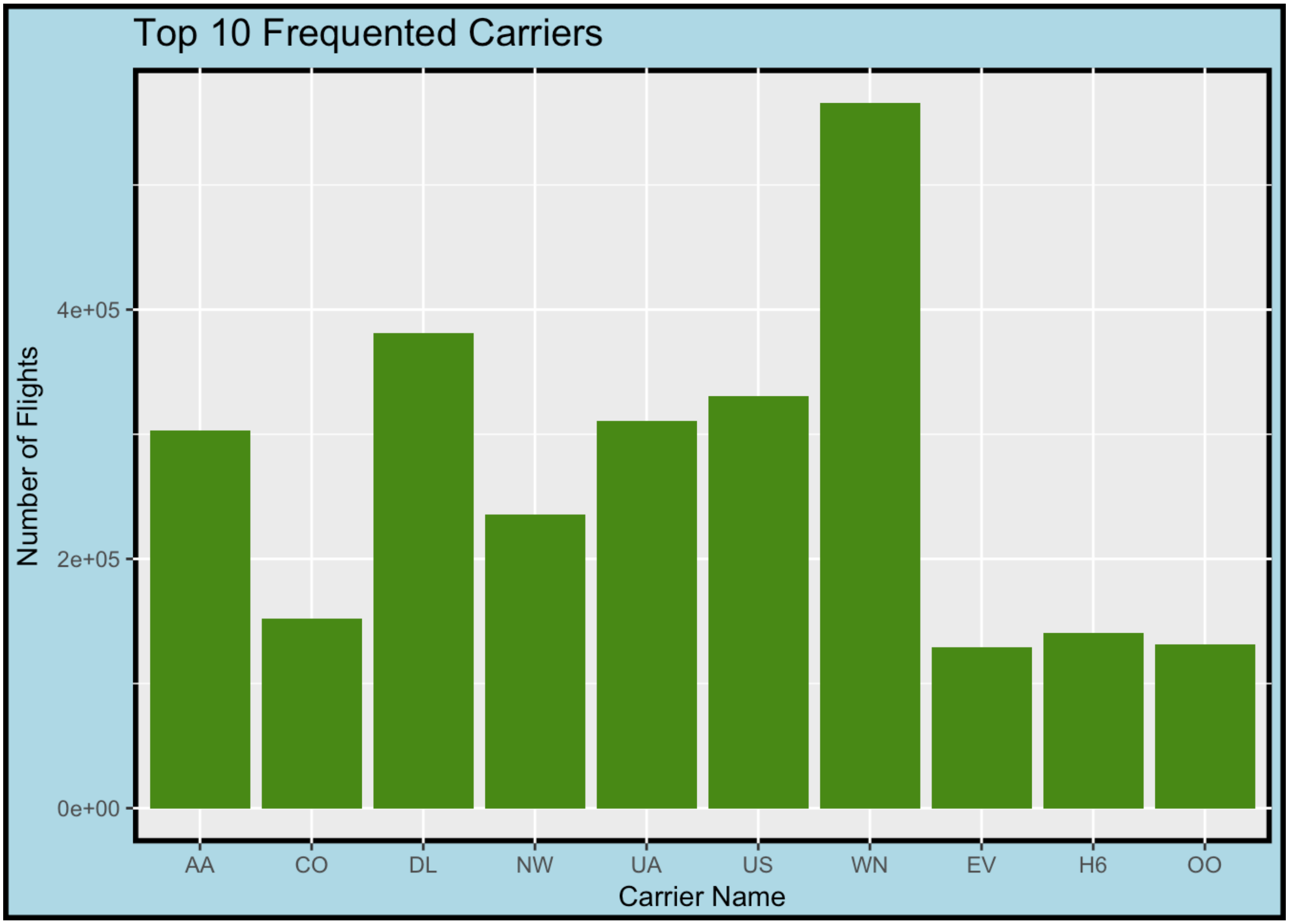
```
mainDF <- airlineData %>%
      select(PASSENGERS, FREIGHT, MAIL, DISTANCE, UNIQUE_CARRIER, CARR
IER_GROUP, CARRIER_GROUP_NEW, ORIGIN_CITY_NAME, ORIGIN_STATE_ABR, DEST
_CITY_NAME, DEST_STATE_ABR, YEAR, QUARTER, MONTH, DISTANCE_GROUP, CLAS
S)
```

```
totalFlights <- airlineData %>%
      filter(PASSENGERS != 0) %>%
      group_by(UNIQUE_CARRIER) %>%
      summarize(NumFlights = n()) %>%
      filter(NumFlights > mean(NumFlights)) %>%
      top_n(10)
```

```
## Warning: package 'bindrcpp' was built under R version 3.2.5
```

```
## Selecting by NumFlights
```

```
FlightGraph <- ggplot(totalFlights,
aes(x=UNIQUE_CARRIER,y=NumFlights)) + geom_bar(stat="identity", fill =
 "chartreuse4") + abluetheme + labs(title="Top 10 Frequented
Carriers") + xlab("Carrier Name") + ylab("Number of Flights")
FlightGraph
```

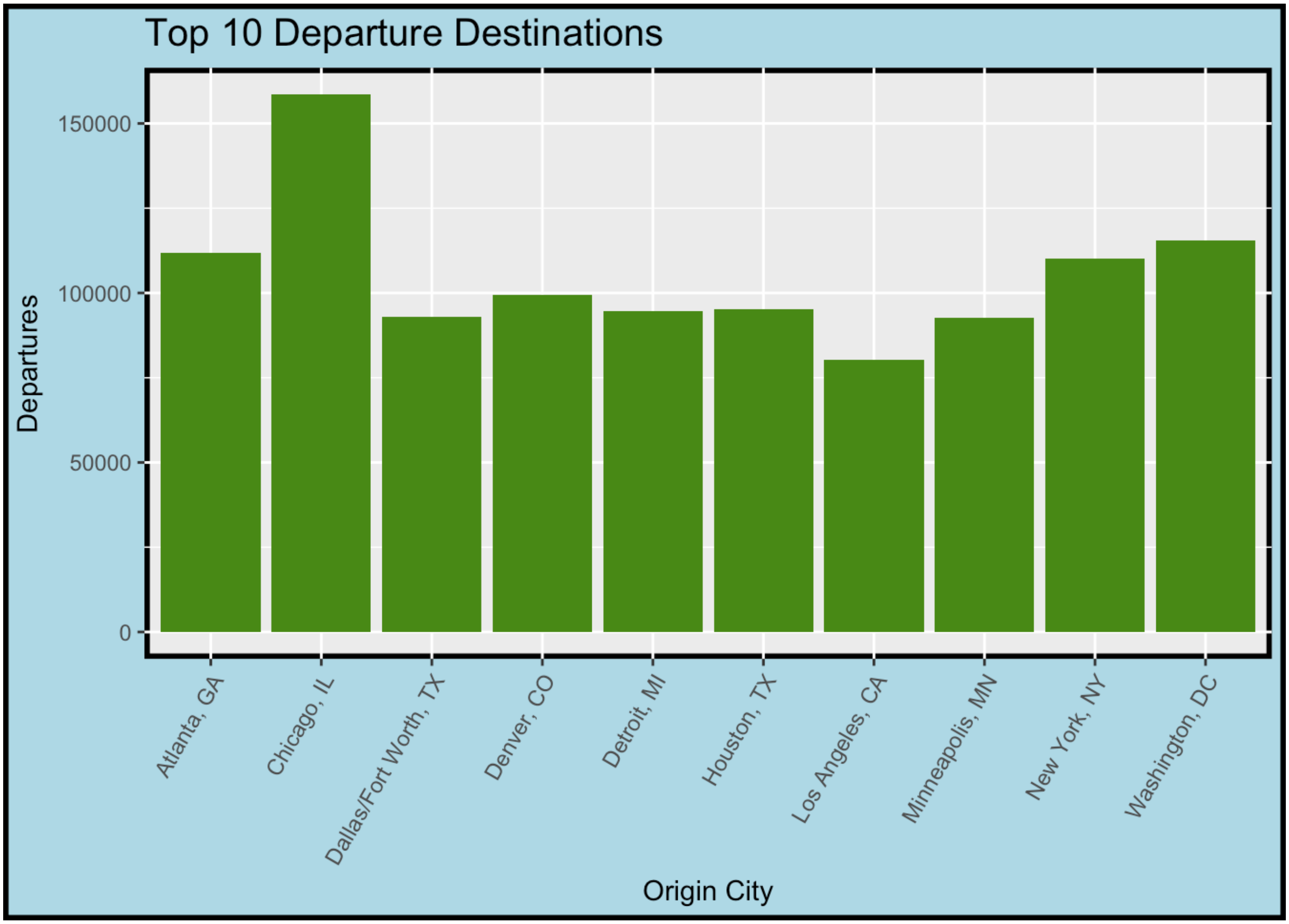


The above graph represents the carrier abbreviations with the most flights between the years of 1990 to 2017. We can see from the graph that `WN` otherwise known as Southwest is the carrier with the most flights.

```
popularAirport <- airlineData %>%
      filter(PASSENGERS != 0) %>%
      group_by(ORIGIN_CITY_NAME) %>%
      summarize(DestinCount = n()) %>%
      filter(DestinCount > mean(DestinCount)) %>%
      top_n(10)
```

```
## Selecting by DestinCount
```

```
OriginGraph <- ggplot(popularAirport, aes(x=ORIGIN_CITY_NAME,y=DestinCount)) + geom_bar(stat="identity", fill= "chartreuse4") + abluetheme + labs(title="Top 10 Departure Destinations") + theme(axis.text.x = element_text(angle=60, hjust=1)) + xlab("Origin City") + ylab("Departures")
OriginGraph
```

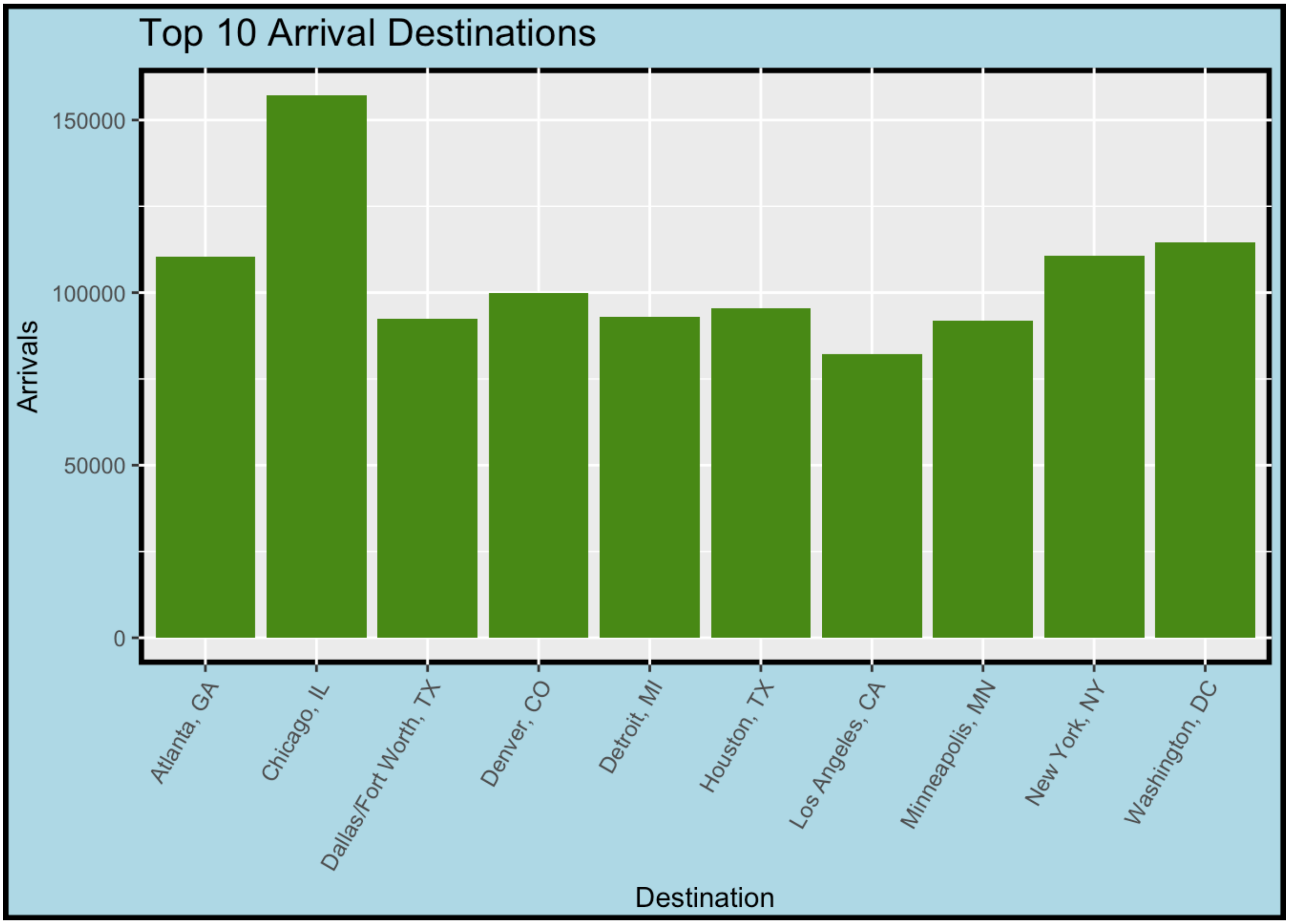


The above graph represents the origin city with the most amount of departures.

```
popularDest <- airlineData %>%
      filter(PASSENGERS != 0) %>%
      group_by(DEST_CITY_NAME) %>%
      summarize(DestinCount = n()) %>%
      filter(DestinCount > mean(DestinCount)) %>%
      top_n(10)
```

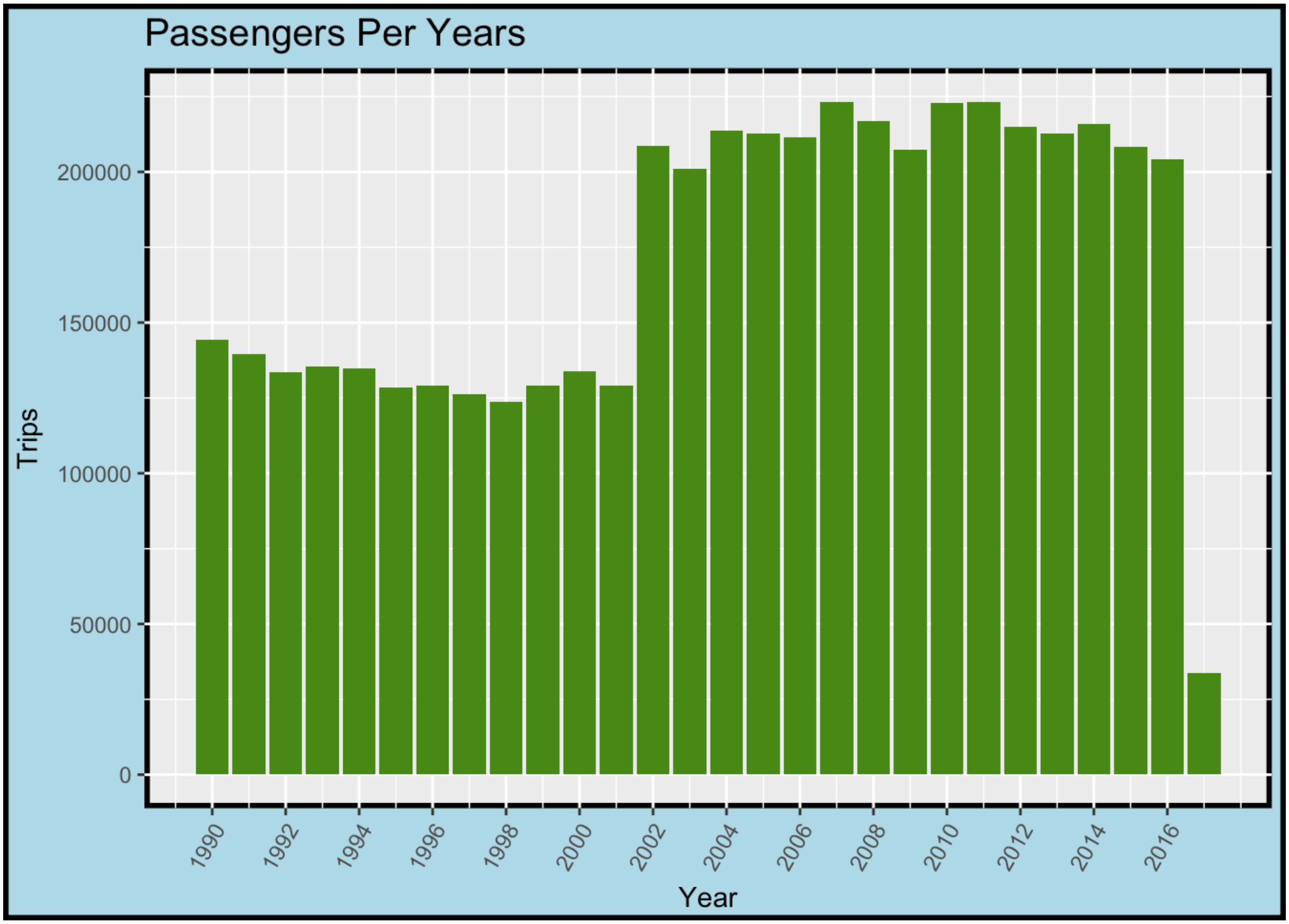
```
## Selecting by DestinCount
```

```
DestGraph <- ggplot(popularDest, aes(x=DEST_CITY_NAME,y=DestinCount)) + geom_bar(stat="identity", fill= "chartreuse4") + abluetheme + labs(title="Top 10 Arrival Destinations") + theme(axis.text.x = element_text(angle=60, hjust=1)) + xlab("Destination") + ylab("Arrivals")
DestGraph
```



The above graph represents the origin city with the most amount of arrivals. Note that the top ten of the departures is the same as top ten of the arrivals.

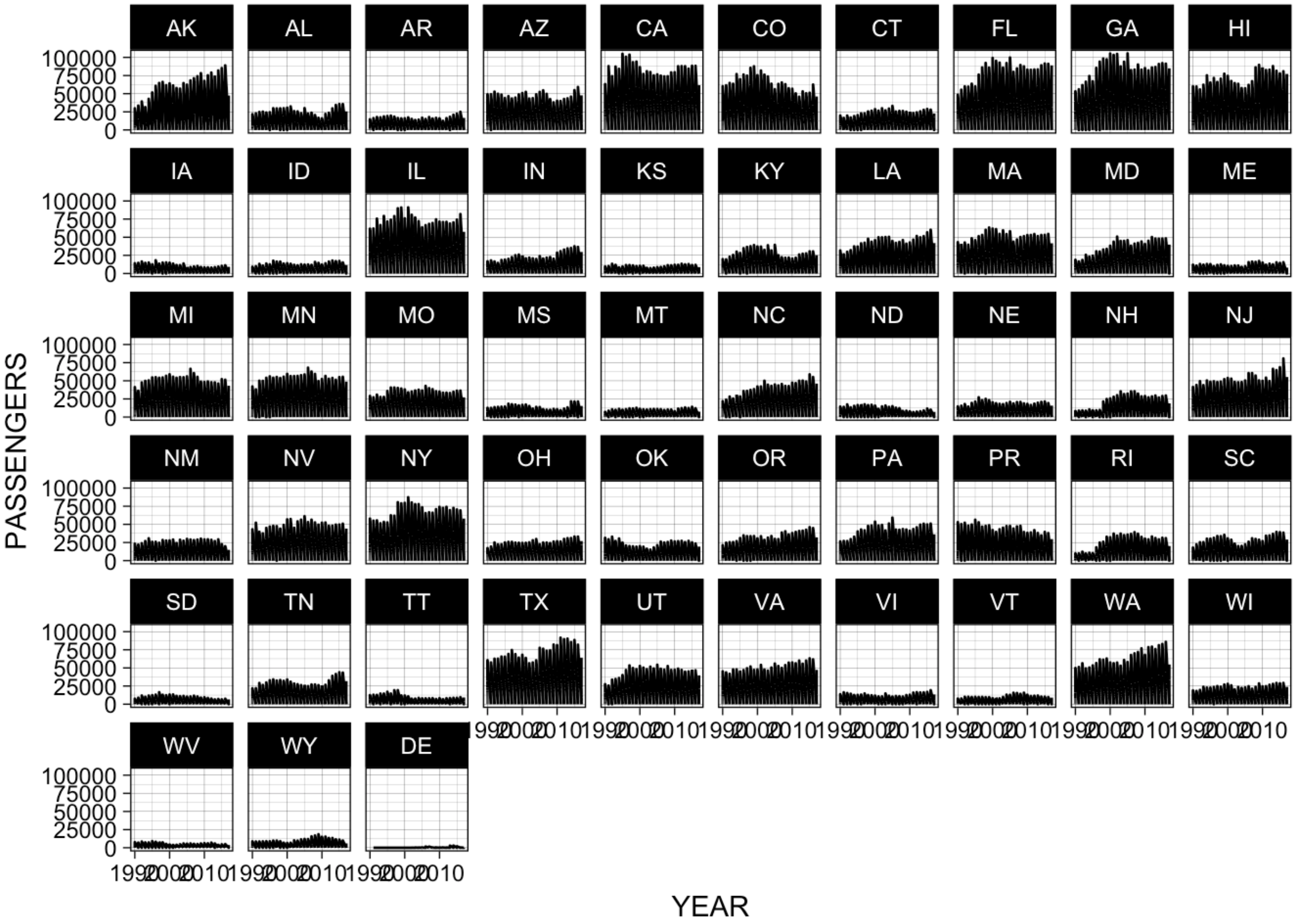
```
YearTrend <- airlineData %>%
      filter(PASSENGERS != 0) %>%
      group_by(YEAR) %>%
      summarize(DestinCount = n())
YearPlot <- ggplot(YearTrend, aes(x=YEAR,y=DestinCount)) + geom_bar(st
at="identity", fill= "chartreuse4") + abluetheme + labs(title="Passeng
ers Per Years") + theme(axis.text.x = element_text(angle=60, hjust=1))
 + xlab("Year") + ylab("Trips") +
scale_x_continuous(breaks=c(seq(1990,2016,2)))
YearPlot
```



The above plot represents the passengers per year from 1990 to 2017.

Using airline data, I began my analysis with a catchy overview analysis like “TOP 10 MOST POPULAR CARRIERS”, “TOP 10 DESTINATION AIRPORTS”, “TOP 10 DEPARTURE AIRPORTS”, and “PASSENGERS OVER THE YEARS”. Through this beginning analysis, we can see several insights: the top departure airports are also the top arrival airports, Southwest Airlines is a clear leader in flights, and also our yearly data shows a clear jump in passengers from 2001 to 2002.

```
StateFlightGraph <- mainDF %>%
      ggplot(aes(YEAR, PASSENGERS)) +
      geom_line() +
      theme_linedraw() +
      facet_wrap(~ORIGIN_STATE_ABR, ncol=10)
StateFlightGraph
```



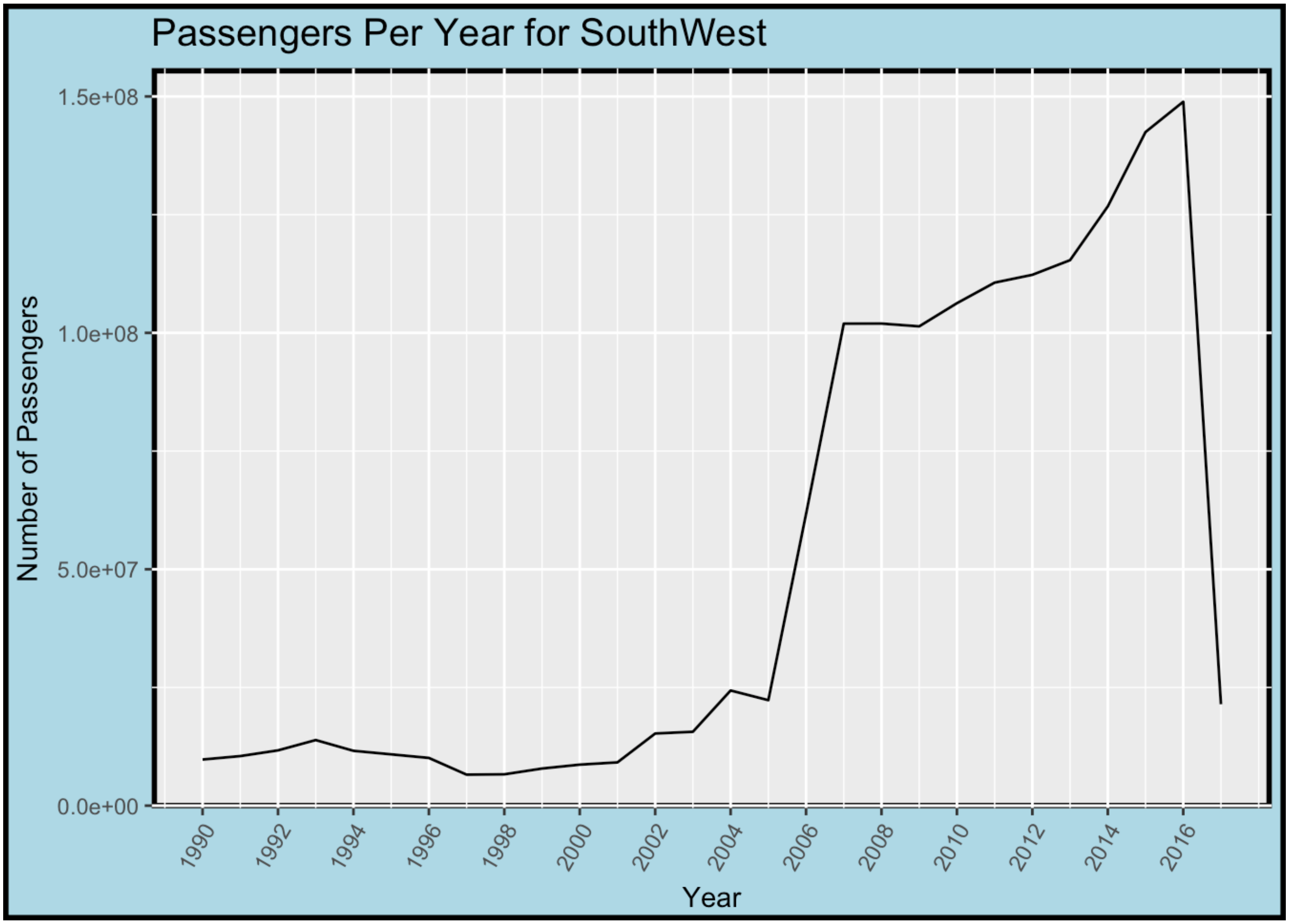
With the state overview above, we can see that over the years, states like Alaska, Washington, and New Jersey have been showing an increase in travelers. Meanwhile, other states such as Colorado have gone down. Finally, some states have reamined popular in visitors like California, Florida, Georgia, and Hawaii.

While these analyses provide a general feel for the data, I felt the best way to report was to focus on certain aspects of the data; as such, the remaining analyses are case studies specific to (1) an airplane company like Southwest and (2) an everyday passenger from California. I chose these two values appeared to stand out in terms of passengers and data amongst the other airlines and states.

Part 4 Airline Businesses

```
SouthWestDF <- mainDF %>%
      filter(UNIQUE_CARRIER=='WN' & MAIL == 0)

SouthWestPlot <- SouthWestDF %>%
      filter(PASSENGERS != 0) %>%
      group_by(YEAR) %>%
      summarise(n=sum(PASSENGERS)) %>%
      ggplot(aes(x=YEAR, y=n)) +
      geom_line() + xlab("Year") + ylab("Number of Passengers") +
      labs(title="Passengers Per Year for SouthWest") +
      scale_x_continuous(breaks=c(seq(1990,2016,2))) + theme(axis.tex
t.x = element_text(angle=60, hjust=1)) + abluetheme
SouthWestPlot
```



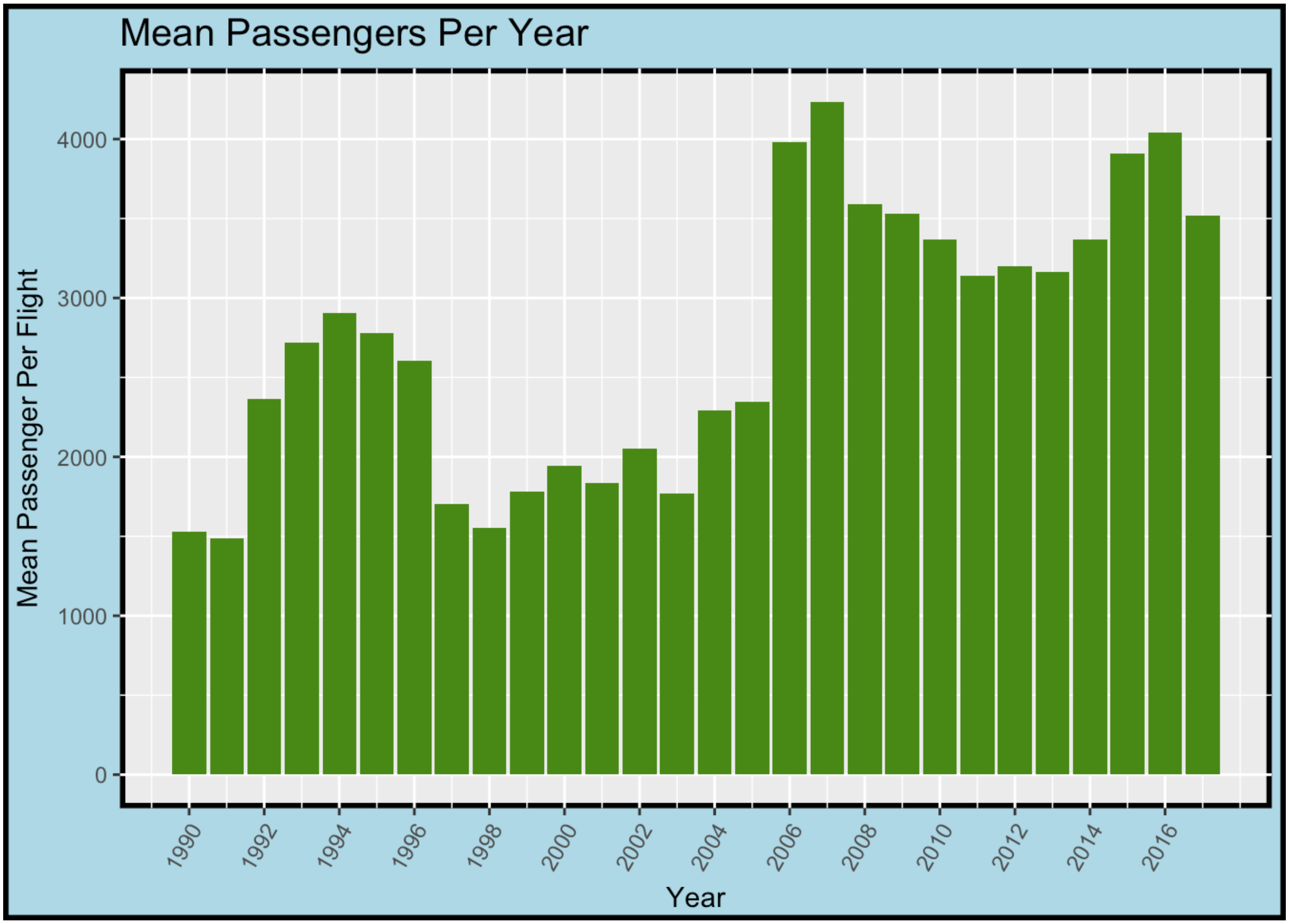
The above plot about passengers per year shows that Southwest had a dramatic increase passengers starting around 2006 and has kept that uphill trend. 2017 is most likely an incomplete set as the current date is still in the month of July 2017.

```
HowManyPeopleDF <- SouthWestDF %>%
      group_by(YEAR) %>%
      summarise(n=sum(PASSENGERS))

HowManyFlightsDF <- SouthWestDF %>%
      group_by(YEAR) %>%
      summarise(Count=n())

MeanPassenger <- data.frame(Year = c(1990:2017),Mean = as.integer(HowManyPeopleDF$n/HowManyFlightsDF$Count))

MeanPlot <- MeanPassenger %>%
      ggplot(aes(x=Year,y=Mean)) +
      geom_bar(stat="identity", fill= "chartreuse4")+
      abluetheme+
      scale_x_continuous(breaks=c(seq(1990,2016,2))) +
      theme(axis.text.x = element_text(angle=60, hjust=1)) +
      labs(title="Mean Passengers Per Year") + xlab("Year") + ylab("Mean Passenger Per Flight")
MeanPlot
```



The above plot shows the mean `PASSENGERS` from Southwest per `Month` per `Year`. Interestingly, the data shows that flights averaged over 4000 passengers in 2007. Upon analysis, it appears as though the data is summing up the number of passengers to one destination per month. For example, `mainDF[5650438,1]` returns the value 6.39410^{4} , which cannot be the passengers for one flight. Through this analysis, we can see that the mean passengers per `Month` is steadily growing from 1990.

With the total and mean, we can see that Southwest has shown a positive trend in all aspects, from the total number of flyers to the average they see per year. Interestingly, we can compare the total and the mean and notice that the mean plot shows a bimodal distribution but the total shows an almost linear graph. This might mean that the from 1992 to 1996, there were not that many passengers flying, but also equally less planes flying. Other insights are that South West boomed in 2006 whereas the `Passengers Per Years` graph shows the overall airline industry boom occurring in 2002. Southwest did not follow the industry, but instead somehow did something in 2006 that distinguished it from the rest of its competitors. A quick search on Google shows that prior to 2006, Southwest was only allowed to fly its 737’s to Little Rock and Tulsa. Bigger destinations such as New York and Los Angeles were off limits for Southwest. However, “in 2006, Congress voted to repeal the Wright Amendment—effective October 2014. So as of late last year, Southwest has finally been free “to move about the country” without restrictions.”(http://fortune.com/2015/09/23/southwest-airlines-business-travel/ (http://fortune.com/2015/09/23/southwest-airlines-business-travel/)). This then explains why Southwest boomed in 2006 while the airline industry did not have any significant changes. From graphs, we can learn more about history and figure out trends we did not know before.

Part 5 CA Resident

```
StateFlightDF <- airlineData %>%
      select(PASSENGERS, FREIGHT, MAIL, DISTANCE, UNIQUE_CARRIER, CARR
IER_GROUP, CARRIER_GROUP_NEW, ORIGIN_CITY_NAME, ORIGIN_STATE_ABR, DEST
_CITY_NAME, DEST_STATE_ABR, YEAR, QUARTER, MONTH, DISTANCE_GROUP, CLAS
S) %>%
      filter(ORIGIN_STATE_ABR=="CA" & MAIL == 0)

UniqueDestinations <- StateFlightDF %>%
      group_by(ORIGIN_CITY_NAME) %>%
      distinct(DEST_CITY_NAME) %>%
      summarise(Destinations = n()) %>%
      filter(Destinations > mean(Destinations)) %>%
      arrange(desc(Destinations)) %>%
      top_n(20)
```

```
## Selecting by Destinations
```

```
Mostairlines <- StateFlightDF %>%
      group_by(ORIGIN_CITY_NAME) %>%
      distinct(UNIQUE_CARRIER) %>%
      summarise(Carriers = n()) %>%
      top_n(20)
```

```
## Selecting by Carriers
```

```
CAcustomer <- inner_join(UniqueDestinations,Mostairlines, by="ORIGIN_C
ITY_NAME")
```

ORIGIN_CITY_NAME	Destinations	Carriers

Burbank, CA	354	55
Los Angeles, CA	330	128
San Diego, CA	303	98
Oakland, CA	292	102
San Francisco, CA	284	100
Ontario, CA	263	94
Sacramento, CA	262	77
San Jose, CA	261	89
Santa Ana, CA	225	41
Palm Springs, CA	193	47
Long Beach, CA	186	62
Fresno, CA	182	55
Santa Barbara, CA	144	39
Bakersfield, CA	122	41
Victorville, CA	120	36
Monterey, CA	108	39
Riverside, CA	81	23

The above `CAcustomer` table represents the (1) amount of unique destinations an origin city has (2) and the number of unique carriers at that airport. For example, San Francisco has 284 unique destinations and 100 unique carriers whereas another city like San Diego has 303 unique destinations and 98 unique carriers. Using this graph, we can see which airport cities are the most accessible to new locations with the most diversity of carriers.

```
MostExperienced <- StateFlightDF %>%
      group_by(UNIQUE_CARRIER) %>%
      summarise(Flight = sum(DISTANCE)) %>%
      arrange(desc(Flight))%>%
      top_n(20)
```

```
## Selecting by Flight
```

UNIQUE_CARRIER	Flight
WN	97172054
UA	50930819
AA	33410111
DL	21457582
HP	14369889
US	14323088
NW	14203569
CO	13222894
OO	13095129
AS	11300610
FX	10881913
5X	7562776
B6	5951336
YV	5586366
F9	5354005
TZ	5095130
VX	4494491
0WQ	4334972
TW	4060346
QX	3636290

Knowing the cities with the most destinations and carriers is good, but as a cautious California individual, I am skeptical of airplanes and their safety. Therefore, using the above table, I can gain information about which airline carriers have the most experience flying.

Airlines with more experience are most likely in control of the air flight process; thus a safer choice.

Combined, the `CAcustomer` and `MostExperienced` tables can be useful for deciding which nearby city in California has access to the most new locations and which carrier I should take to feel most at ease. For future projects, this could be expanded by creating a joined table of the places with the most carriers and a list of those carriers which are the most experienced for a more concise table. Additionally, the joining would differentiate between specific airlines, for example if Hawaiian Airlines was a main destination with a lot of experience, it needs to be handled from showing up at continental origin states. Given more knowledge on joins and time, a more insightful graph could be made to serve the typical Californian (and more) traveler.

From the results of the three reports (overall trends, case study of a single carrier, and case study of a California resident), we can see that having general graphs give a direction of where to analyze further, the further graphs reveal history otherwise unknown, and summarized tables offer a way to help ease customer's decisions about which airports and airlines to take.