Case Study on US Flight Data

Chris Dong

```
library(tidyverse)
library(magrittr)
library(stringr)
library(lubridate)
library(maps)
library(ggmap)
```

The course website contains a sub-directory for `hw2` named `q3_data`. In that directory you will find US flight data for the years 1990 through 2017. A data dictionary is also provided. The `keys` subdirectory provides additional information you may choose to use, but are not required to. There are two objectives.

1. Write a block of code that, with a single execution, imports and stores all 28 `csv` in one data frame named `airlineData`.

```
year <- 1990:2017
csvAll <- function(year){
  paste0("q3_data/657240010_T_T100D_MARKET_US_CARRIER_ONLY_",year,"_All.csv")
}
airlineData <- data_frame()
for(i in 1:length(year)){
  airlineData <- rbind(airlineData,readr::read_csv(csvAll(year[i])))
}
```

I am making a vector for the years from 1990 to 2017.
Then, I make a function that has everything for the `.csv` file except the year. For example, `csvAll(2000)` will give the corresponding `.csv` file for the year 2000. Then, I create an empty dataframe and populate it one by one with a loop.

2. Using the data dictionary provided, explore the data and report on your findings. You are requied to employ as much `dplyr` and piping notation (from the `magrittr` package) as possible. Your use of these packages—or lack thereof—will impact your grade on this question. There is no right answer to this question. Your job is to explore and report with words, graphs and tables. You need not report on all variables. Use your judgement to distill the data and report back the most interesting information.Your work should not exceed 15 pages, but is not required to be that long. Length is not a criterion on which you will be graded. The best reports are often concise and direct.

First, I will convert all of the variable names to lowercase to make it easier.

```
names(airlineData) <- str_to_lower(names(airlineData))
```

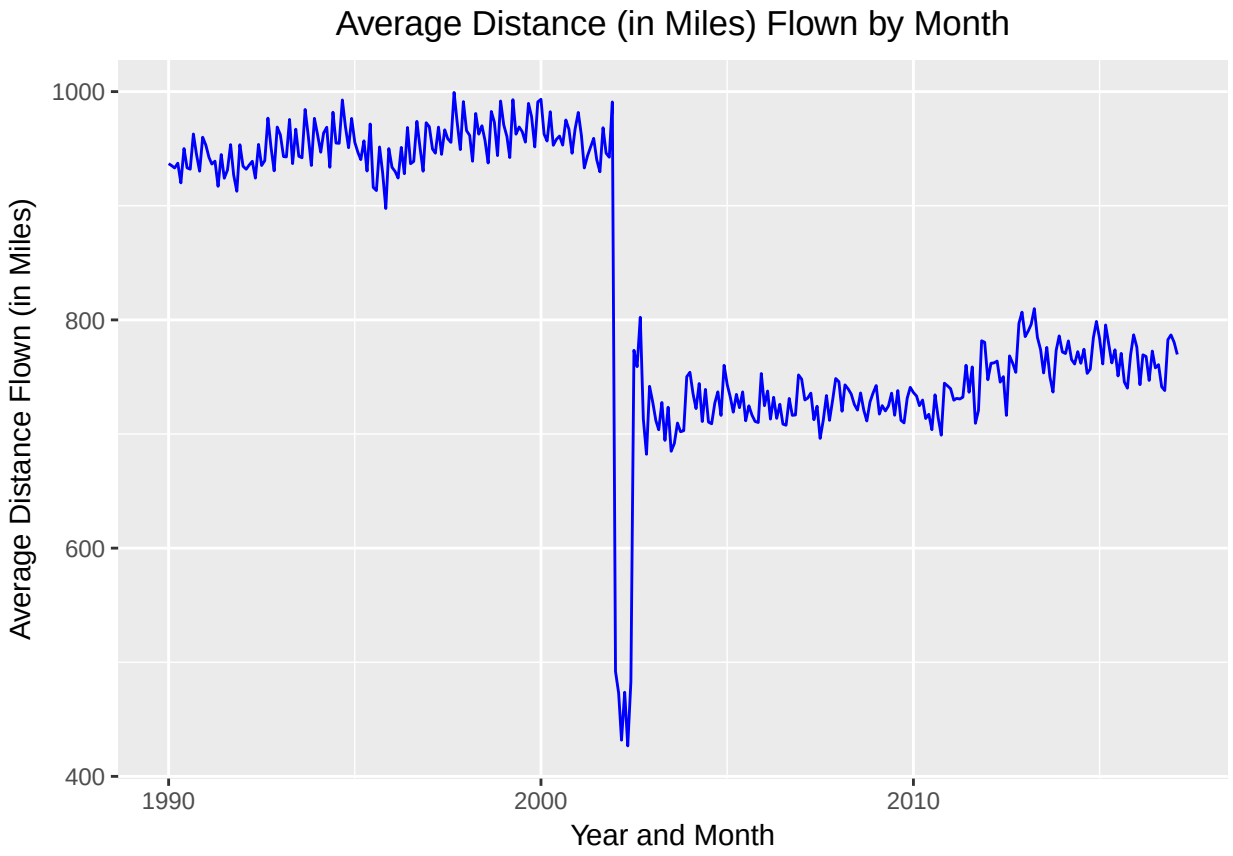
I am combining the `year` and `month` variable into the variable `year_month`.

```
airlineData %<>%
  mutate(year_month = make_date(year, month))
```

The variable `distance` measures the distance between airports. We see that, on average, people flew to further destinations prior to the year 2002. There is a drastic drop, followed by another increase. However, it never reaches as high as it did in the past.

```
airlineData %>%
  mutate(year_month = make_date(year, month)) %>%
    group_by(year_month) %>%
```

```
    summarise(avgDist = mean(distance,na.rm=T)) %>%
        ggplot(aes(x=year_month, y=avgDist)) + geom_line(color="blue") +
          xlab("Year and Month") + ylab("Average Distance Flown (in Miles)") +
          ggtitle("Average Distance (in Miles) Flown by Month") +
          theme(plot.title = element_text(hjust = 0.5))
```



Pinpointing the `year` and `month` of this low point, we see that starting the year `2002`, the average distance between airports for flights is less than 500. It begins to increase to the high 700's starting July 2002. I am displaying the `year` and `month` in a different format because year_month includes `01` for the day by default so it may be misleading and therefore I am not including the day when displaying it.

```
airlineData %>%
  mutate(year_month = make_date(year, month)) %>%
    group_by(year_month) %>%
      summarise(avgDist = mean(distance,na.rm=T)) %>%
        filter(year_month >= make_date(year = 2001L, month = 9L),
               year_month <= make_date(year = 2002L, month = 8L)) %>%
                mutate(year_month = format(year_month, "%b %Y")) %>%
                  knitr::kable(align=c('c','c'),
                               col.names = c("Time Period", "Average Distance"),
                               digits = 0)
```

Time Period	Average Distance
Sep 2001	968
Oct 2001	946
Nov 2001	943
Dec 2001	991

Time Period	Average Distance
Jan 2002	492
Feb 2002	473
Mar 2002	432
Apr 2002	474
May 2002	427
Jun 2002	483
Jul 2002	773
Aug 2002	759

In the table, we see a decrease in the number of flights starting the beginning of 2002 – it lasts for about half a year.

I suspect the cause of this is the infamous 9/11 event. Essentially, fears became instilled in people's minds, especially with regard to flying. Two of the airlines involved were American Airlines and United Airlines. Furthermore, Delta Air Lines was suspected of hijacking. We can do some analysis to see if there was any impact on their popularity.

Before 9/11 occurred, here are the top 6 most popular flight carriers

```
before <- airlineData %>% filter(year_month < make_date(year = 2001L, month = 9L)) %>%  group_by(unique
    mutate(percent = n * 100 / sum(n)) %>%
      arrange(desc(n)) %>% head(6)
before %>% knitr::kable(align=c('c','c'),
                    col.names = c("Carrier Name", "Count", "Percentage"),
                    digits = 2)
```

Carrier Name	Count	Percentage
Delta Air Lines Inc.	209382	12.75
US Airways Inc.	206633	12.59
United Air Lines Inc.	178682	10.88
American Airlines Inc.	158340	9.64
Southwest Airlines Co.	147667	8.99
Northwest Airlines Inc.	144875	8.82

After 9/11 occurred, here are the top 6 most popular flight carriers

```
after <- airlineData %>% filter(year_month >= make_date(year = 2001L, month = 9L)) %>%  group_by(unique
    mutate(percent = n * 100 / sum(n)) %>%
      arrange(desc(n)) %>% head(6)
after %>% knitr::kable(align=c('c','c'),
                    col.names = c("Carrier Name", "Count", "Percentage"),
                    digits = 2)
```

Carrier Name	Count	Percentage
Southwest Airlines Co.	436823	10.90
United Air Lines Inc.	182255	4.55
Hageland Aviation Service	181154	4.52
Delta Air Lines Inc.	173570	4.33
Federal Express Corporation	155514	3.88
American Airlines Inc.	145791	3.64

As we can see from the tables above, Delta Airlines used to be the most popular, consisting of 12.75% of all flights. After 9/11, they went from 1st to 4th, decreasing to 4.33% of all flights. Similarly, United Airlines had a decrease of 6.34% in their flights. Finally, American Airlines was also affected, going from 4th to 6th and having a 6.01% decrease in the percentage of flights.

Next, let's investigate the `carrier_group` variable. This variable indicates how much revenue these carriers are generating. It is labeled 1, 2, or 3. 1, or `Large Regional`, has an annual revenue from 20 to 100 million; 2, or `National` is from 100 million to 1 billion; 3, or `Major` has revenue over 1 billion. I will also remove `7` from our data because it is not so relevant and is a very tiny percentage of our data.

One question I am wondering is whether carriers with higher revenue will fly to farther airports, on average.

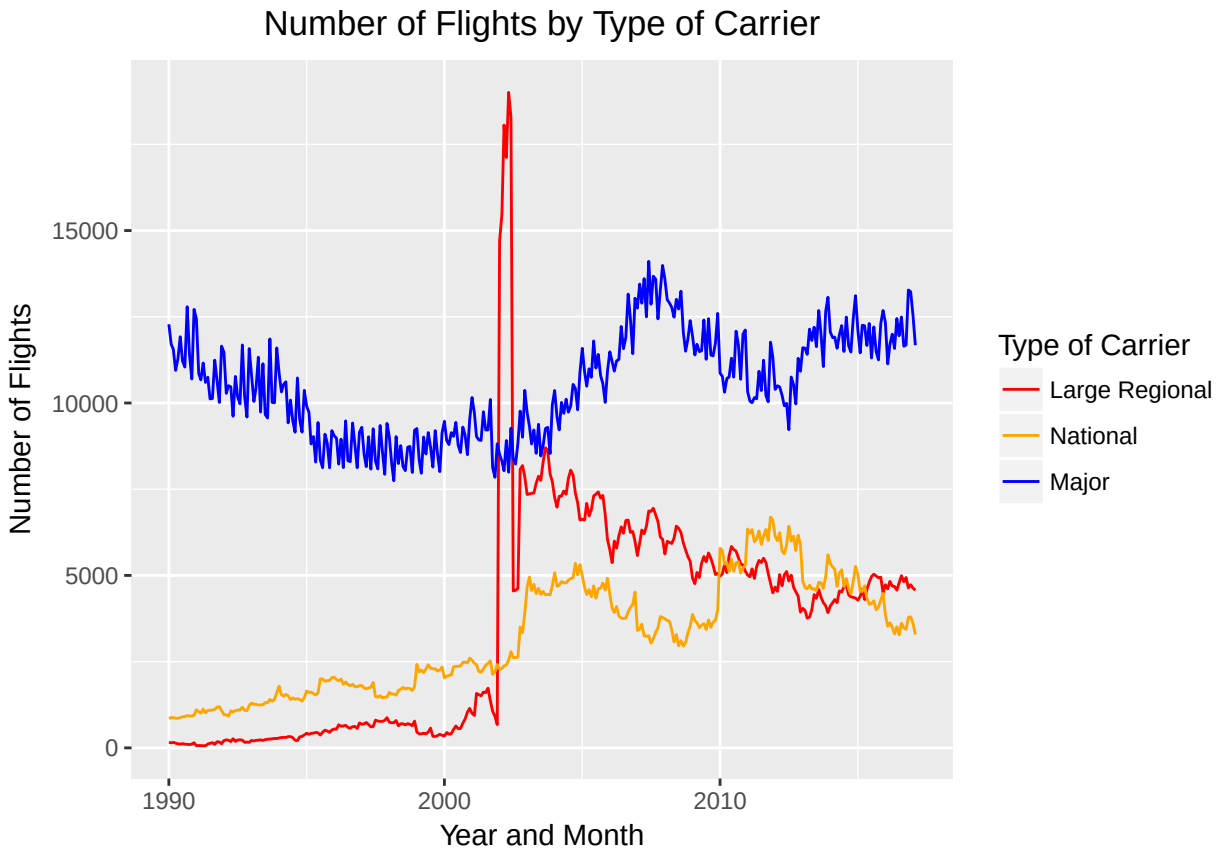
```
airlineData %>% filter(carrier_group!=7) %>% group_by(carrier_group) %>%
  summarise(avgDist = mean(distance, na.rm = T)) %>%
    knitr::kable(align=c('c','c'),
                 col.names = c("Carrier Group", "Average Distance"),
                 digits = 2)
```

Carrier Group	Average Distance
1	302.42
2	618.60
3	1019.66

The table agrees with my expectations – the bigger carrier companies seem to take care of the longer distance flights.

Next is a plot showing the number of flights over time for the three type of carriers.

```
airlineData %>% filter(carrier_group!=7) %>% group_by(carrier_group, year_month) %>%
  count() %>% ggplot(aes(x=year_month, y=n,
                         group=factor(carrier_group),
                         color=factor(carrier_group)))+
                         geom_line() +  xlab("Year and Month") +
                         ylab("Number of Flights") +
                         ggtitle("Number of Flights by Type of Carrier") +
                         theme(plot.title = element_text(hjust = 0.5)) +
                         labs(col = "Type of Carrier") +
                         scale_color_manual(labels = c("Large Regional", "National", "Major"),
                                            values = c("red","orange","blue"))
```



Interestingly, there is a large peak in large regional carriers. To see exactly when these peaks occured, we can look at which months had the most number of flights.

```
airlineData %>% filter(carrier_group==1) %>% group_by(carrier_group, year_month) %>%
  count() %>% arrange(desc(n)) %>% ungroup() %>% select(-carrier_group) %>%
  mutate(year_month = format(year_month, "%b %Y")) %>% head(6) %>%
      knitr::kable(align=c('c','c'), col.names = c("Time Period", "Number of Flights"))
```

Time Period	Number of Flights
May 2002	19005
Jun 2002	18285
Mar 2002	18056
Apr 2002	17121
Feb 2002	15469
Jan 2002	14699

From the table, we can see that the number of flights by `Large Regional` carriers increased significantly right around the time of 9/11. This makes sense logically because terrorists may be more likely to target `Major` airline carriers over smaller ones. People may feel that regional carriers would be safer to travel on and thus we see the peak in flights by large regional carriers.

Here, we can see where most of these flights are coming from and traveling to.

```
airlineData %>% group_by(origin_state_nm, dest_state_nm) %>% count() %>% arrange(desc(n))
```

```
## # A tibble: 2,686 x 3
## # Groups:   origin_state_nm, dest_state_nm [2,686]
```

```
##    origin_state_nm dest_state_nm      n
##              <chr>         <chr>  <int>
##  1          Alaska        Alaska 673396
##  2           Texas         Texas  85269
##  3      California    California  77592
##  4         Florida       Florida  44551
##  5           Texas    California  40887
##  6      California         Texas  40550
##  7          Hawaii        Hawaii  34758
##  8        New York       Florida  30110
##  9         Florida      New York  29803
## 10      Washington    Washington  25874
## # ... with 2,676 more rows
```

We can see that most of the data involves flights within the state of Alaska. We should be careful because our data might be skewed towards Alaska data; flights to, from, and within Alaska may not be representative of flights in general. From this table, we also see that many flights are in-state. The top 4, for example, are flights that go from one city to another within the same state.

Next, we can see the most popular filghts by origin and destination city.

```
airlineData %>% group_by(origin_city_name, dest_city_name) %>% count() %>% arrange(desc(n))
```

```
## # A tibble: 74,756 x 3
## # Groups:   origin_city_name, dest_city_name [74,756]
##    origin_city_name  dest_city_name     n
##               <chr>           <chr> <int>
##  1     New York, NY Los Angeles, CA  3795
##  2   Washington, DC    New York, NY  3728
##  3     New York, NY  Washington, DC  3681
##  4      Chicago, IL    New York, NY  3672
##  5  Los Angeles, CA    New York, NY  3566
##  6     New York, NY     Chicago, IL  3308
##  7  Minneapolis, MN     Chicago, IL  3305
##  8      Chicago, IL Minneapolis, MN  3267
##  9      Detroit, MI     Chicago, IL  3267
## 10      Chicago, IL     Detroit, MI  3163
## # ... with 74,746 more rows
```

This coincides with common sense. New York, Los Angeles, Washington DC, and Chicago are all very popular places to travel to, for both work and tourism.

I will focus on California for now – what are the most popular travel destinations in California?

```
airlineData %>% filter(dest_state_nm=="California") %>%
  group_by(dest_city_name) %>% count() %>% arrange(desc(n))
```

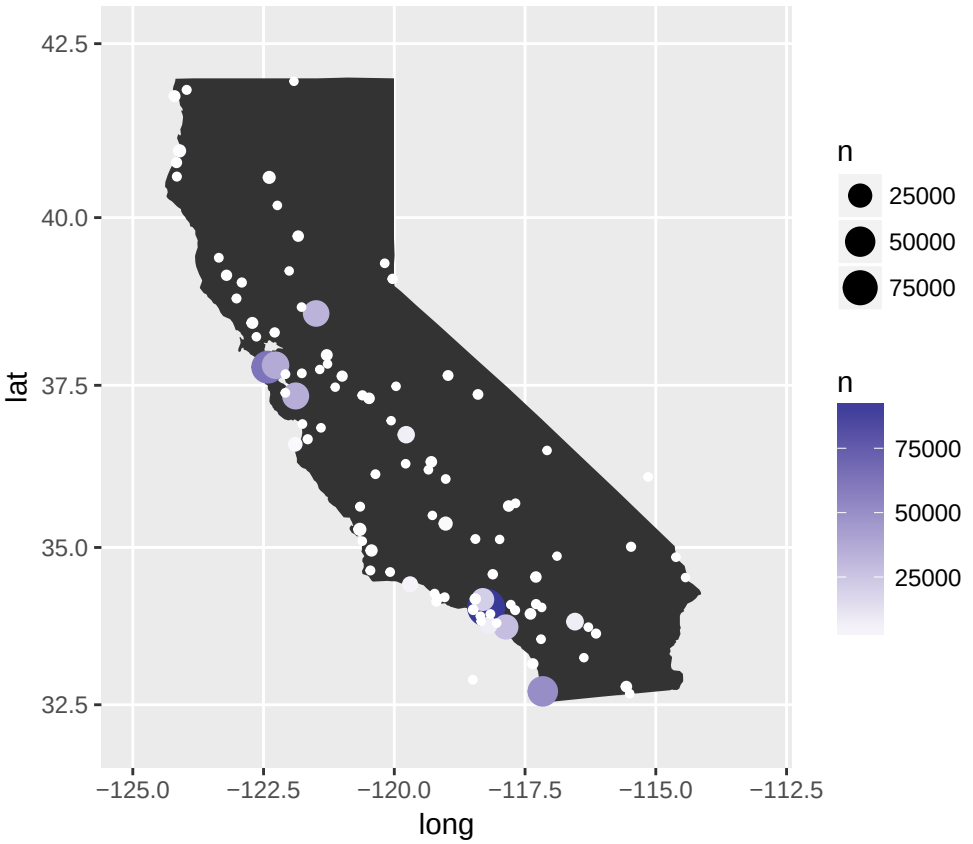
```
## # A tibble: 109 x 2
## # Groups:   dest_city_name [109]
##       dest_city_name     n
##                <chr> <int>
##  1   Los Angeles, CA 92549
##  2 San Francisco, CA 62893
##  3     San Diego, CA 50851
##  4       Oakland, CA 37138
##  5      San Jose, CA 35386
##  6       Ontario, CA 35021
```

```
##  7    Sacramento, CA 33506
##  8     Santa Ana, CA 28410
##  9       Burbank, CA 20932
## 10  Palm Springs, CA  7845
## # ... with 99 more rows
```

I can use `library(map)` to make a plot that displays the most popular travel destinations for cities in California. First, I use `map_date` to get the latitude and longitude in order to draw California on `ggplot`. Then, I am extracting only the cities name in my dataset. Next, I want to remove the `, CA` so I use the regular expression function `gsub()`. Then, I will create a new data frame that contains all of this information. Finally, I will use `geocode()` to extract the latitude and longitude of the cities in California to plot their popularity by size and color.

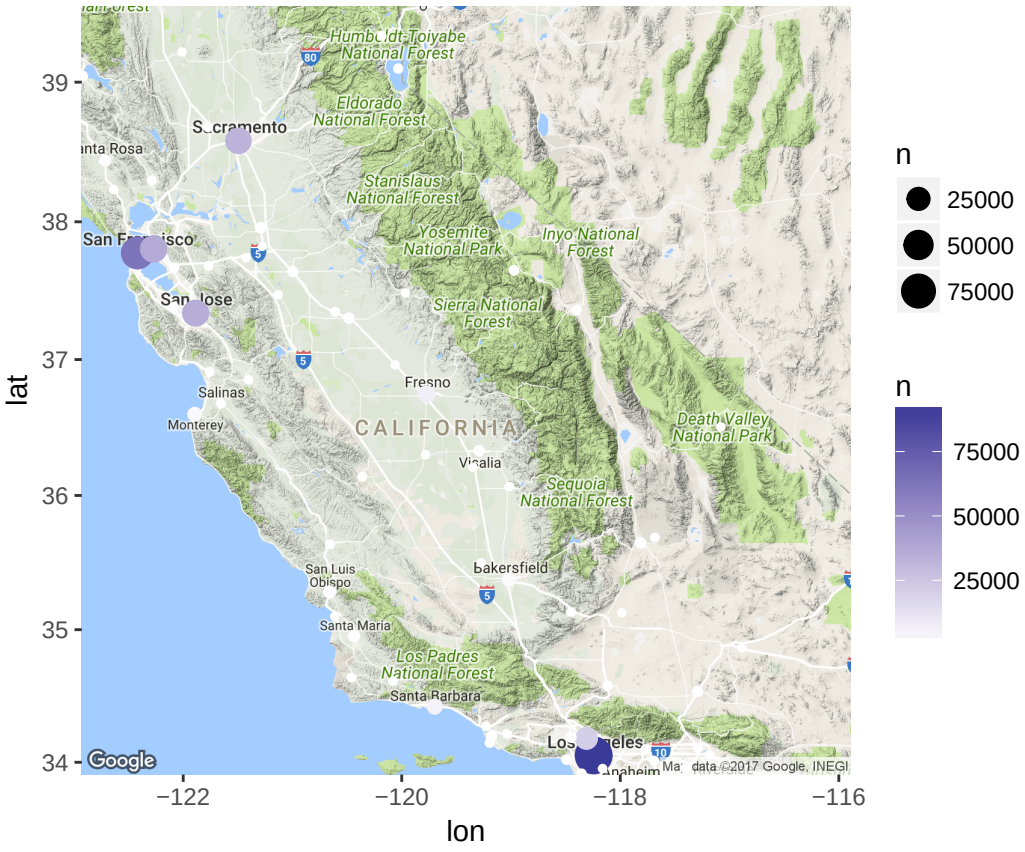
Source: I used Stack Overflow (https://stackoverflow.com/questions/29037851/how-do-i-plot-us-cities-using-ggplot) to assist in making the plot, though I learned this in undergraduate before. I also had to do some data cleaning before it would work properly.

```
CA <- map_data("state", region = "california")

cities <- airlineData$dest_city_name

cityCA <- gsub('.{4}$', '', cities)

dataframeCA <- airlineData %>% mutate(City = cityCA) %>%
  filter(dest_state_nm=="California") %>%
  group_by(City) %>% count() %>% arrange(desc(n))

geographyCA <- geocode(dataframeCA$City)

dataframeCA$lon <- geographyCA$lon
dataframeCA$lat <- geographyCA$lat

ggplot(CA, aes(x=long, y=lat))+
  geom_polygon()+
  coord_map() +
  geom_point(data = dataframeCA, aes(x=lon, y=lat, size=n, color = n))+
  xlim(c(-125,-113)) +ylim(c(32, 42.5))+scale_color_gradient2()
```



We can plot similar information on Google Maps. From this plot, we can easily get a visualization and see how San Francisco, Sacramento, San Jose, and Los Angeles have high counts for the number of flights that heads towards there.

```
ggmap(get_map(location='California', zoom = 7))+
  geom_point(data = dataframeCA, aes(x=lon, y=lat, size=n, color = n))+
  scale_color_gradient2()
```

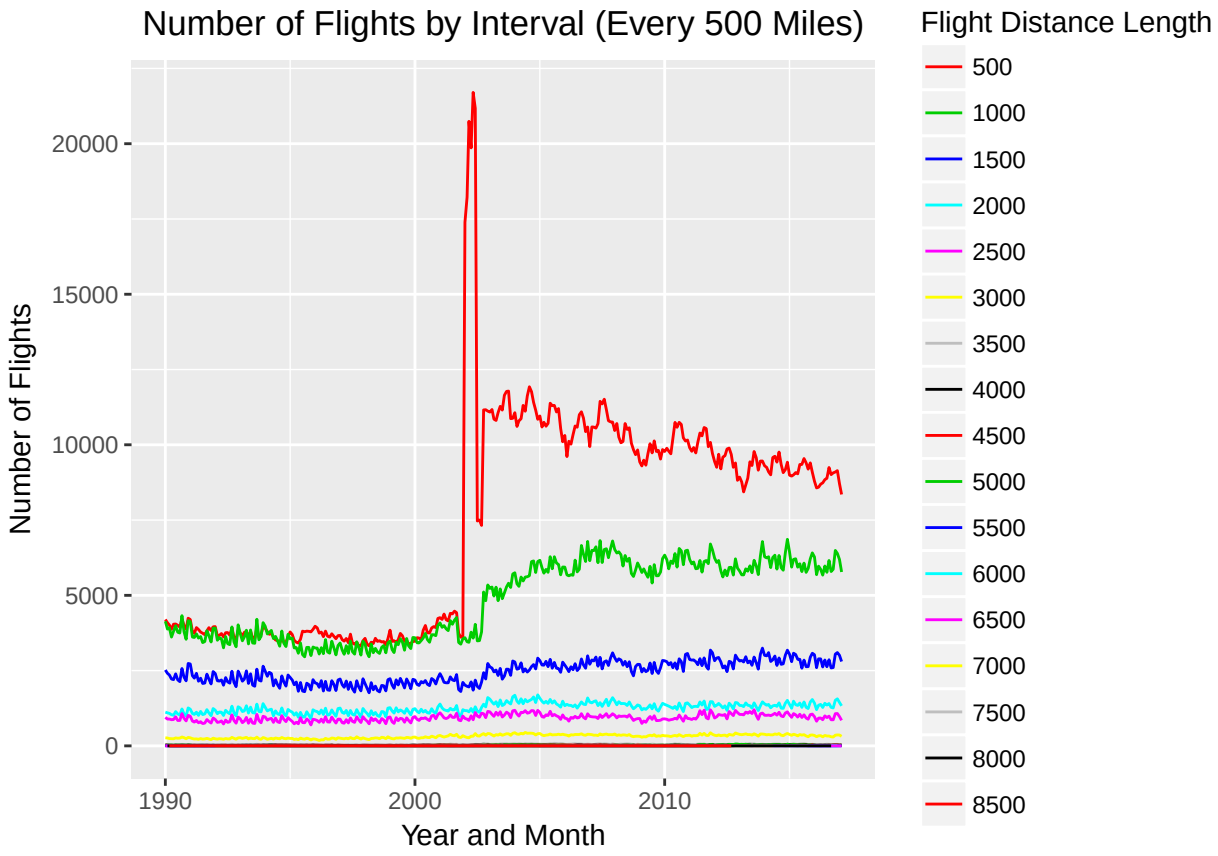


I find that many of the variables are repetitive, so I will remove some of them for easier analysis.

```
condenseData <- airlineData %>%
  select(passengers:unique_carrier, unique_carrier_name, region, carrier_group_new, origin,
         origin_state_nm, dest, dest_state_nm, quarter, distance_group,
         class, year, month, year_month)
```

I noticed a variable called `distance_group`, which measures the distance of the flights and turns it into a categorical variable that counts every 500 miles. I can plot the number of flights within the categorical variable distance_group over time.

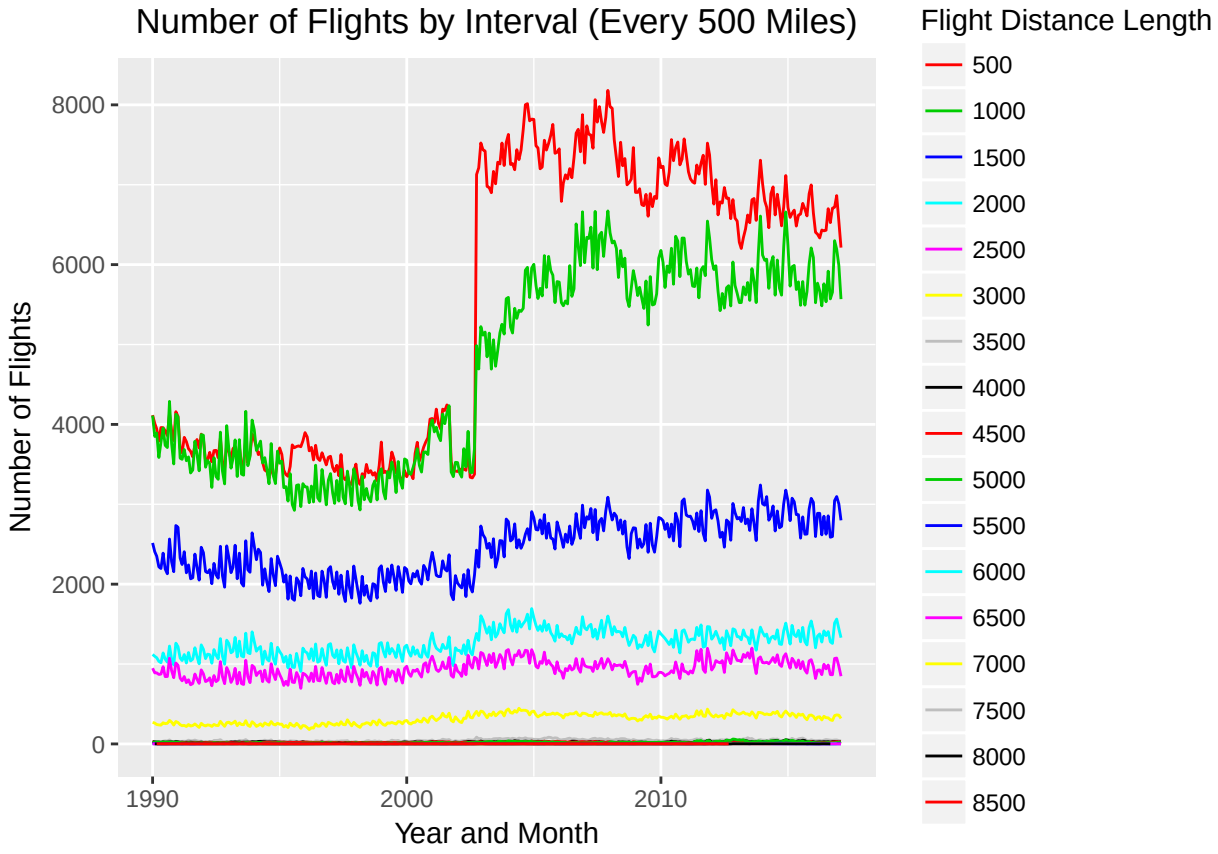
```
condenseData %>% group_by(year_month) %>% count(distance_group) %>% arrange(desc(n)) %>%
  ggplot(aes(x=year_month, y=n, color = factor(distance_group))) + geom_line()+
                        xlab("Year and Month") +
                        ylab("Number of Flights") +
                        ggtitle("Number of Flights by Interval (Every 500 Miles)") +
                        theme(plot.title = element_text(hjust = 0.5)) +
                        labs(col = "Flight Distance Length") +
                        scale_color_manual(labels = c(seq(500,8500,by=500)),
                                           values = 2:18)
```



We get similar results as previously, with a peak in shorter flights during the 9/11 period. Something new that I discovered is the increase in flights from 500 miles to 1000 miles over time. According to the plot, the distance of flights seem to stay more or less the same over time.

Because an overwhelming majority of the flights involve the state of Alaska, I will create another plot by first filtering out Alaska to see if there is any changes.

```
condenseData %>%
  filter(origin_state_nm!="Alaska" | dest_state_nm!= "Alaska") %>%
  group_by(year_month) %>% count(distance_group) %>% arrange(desc(n)) %>%
  ggplot(aes(x=year_month, y=n, color = factor(distance_group))) + geom_line()+
                          xlab("Year and Month") +
                          ylab("Number of Flights") +
                          ggtitle("Number of Flights by Interval (Every 500 Miles)") +
                          theme(plot.title = element_text(hjust = 0.5)) +
                          labs(col = "Flight Distance Length") +
                          scale_color_manual(labels = c(seq(500,8500,by=500)),
                                             values = 2:18)
```



This plot seems to be more accurate and representative after removing Alaska from it. In particular, we see that flights under 500 or 1000 miles became more common over time. There is less of a sudden drop as we see in the previous plot. This shows that when analyzing data, we need to make sure that our results are not biased. Without filtering out Alaska, we may get misleading results as such.

Although fairly obvious because the origin and destination variable indicate that our data is mostly domestic flights, we can confirm by looking at the `region` variable.

```
condenseData %>% group_by(region) %>% summarise(n=n()) %>%
  mutate(percent = n*100/sum(n)) %>% arrange(desc(n))
```

```
## # A tibble: 7 x 3
##   region       n      percent
##    <chr>   <int>        <dbl>
## 1      D 5611582 99.312178311
## 2      I   19254  0.340751802
## 3      L    9392  0.166216938
## 4      P    6011  0.106380964
## 5      A    2430  0.043005447
## 6      S    1569  0.027767715
## 7   <NA>     209  0.003698822
```

Indeed, our data is nearly 100% domestic filghts.

Next, I will explore Carrier History from the keys data.

```
carrierHistory <- readr::read_csv("q3_data/keys/L_CARRIER_HISTORY.csv")
```

```
## Parsed with column specification:
## cols(
```

```
##   Code = col_character(),
##   Description = col_character()
## )
```

```
carrierHistory %>% glimpse()
```

```
## Observations: 1,876
## Variables: 2
## $ Code        <chr> "02Q", "04Q", "05Q", "06Q", "07Q", "09Q", "0BQ", "...
## $ Description <chr> "Titan Airways (2006 - )", "Tradewind Aviation (20...
```

I am using `glimpse()` to see the format of the data, particularly, the `Description` variable.

I want to extract the data from `Description`, in particular the starting year. I also want the ending year, if applicable. I can also create a logical variable indicating whether the carrier is still in business.

```
carrierStartDate <- carrierHistory %>%
  select(Description) %>% str_extract_all("[[:digit:]]{4} -") %>% unlist()
carrierStartDate <- carrierStartDate %>%
  str_extract_all("[[:digit:]]{4}") %>% unlist() %>% as.numeric()

outOfBusiness <- str_detect(carrierHistory$Description, "- [[:digit:]]{4}")
indexOutofBusiness <- which(str_detect(carrierHistory$Description, "- [[:digit:]]{4}"))

lastYear <- str_extract_all(carrierHistory$Description, "- [[:digit:]]{4}") %>% unlist()
lastYear <- str_extract_all(lastYear, "[[:digit:]]{4}") %>%
            unlist() %>% as.numeric()
carrierEndDate <- rep(2017, length(carrierStartDate))
carrierEndDate[indexOutofBusiness] <- lastYear

yearsInService <- carrierEndDate - carrierStartDate

airlineName <- str_extract(carrierHistory$Description, "[[:alpha:]].* \\(")
airlineName <- gsub(" \\(","",airlineName)

carrierInfo <- data_frame(Code = carrierHistory$Code, airlineName, outOfBusiness,
                          carrierStartDate, carrierEndDate, yearsInService)

combineData <- left_join(condenseData, carrierInfo,
                         c("unique_carrier"="Code",
                           "unique_carrier_name"="airlineName"))
```

First, in `carrierStartDate`, I extracted the first part of the year (2006 - 2012) to get `"2006 -"` in a vector that is the same length as `carrierHistory`, or 1876. Then, I performed it again to remove the `-` and then turned it into a numeric vector.

The next part is a little complicated. Carriers that are out of business will have an end date. It would be the `2012` in the example above. However, Carriers that are still in business will simply have `(2006 - )`. So, first I will create a logical vector in `outOfBusiness` to indicate whether or not the business still exists. Then, I will use `which()` to get the index number of it. Next, I will extract the `- 2012` part and do it again to get `2012` as numeric. I will utilize the index numbers I found earlier to fill a `2017` vector with the respective years. Note, that those carriers that are still in business will have `2017` as an end date simply so I can calculate how long they have been in service. I will then extract the name of the airline using regular expressions by extracting up to `(` and then `gsub()` to replace the `(` with an empty string.

When trying to merge `carrierHistory` with `condenseData` via `unique_carrier` and `Code`, there does not seem to be a one-to-one matching. Therefore, I created two set of keys by including `unique_carrier_name`

and `airlineName`. I will use `left_join()` to keep only data from the main dataset.

Now, we can do some analysis with our new variables that contain information about the carrier's start and end dates(if applicable), their number of years in service, and whether or not they are still in business.

```
inBusiness <- combineData %>% group_by(outOfBusiness) %>% summarise(n=n()) %>%
  mutate(prop = n/sum(n))
inBusiness %>% knitr::kable(align=c('c','c','c'),
                            col.names = c("Still Operating?", "Number","Proportion"))
```

Still Operating?	Number	Proportion
FALSE	4007905	0.7093076
TRUE	1402008	0.2481234
NA	240534	0.0425690

About 24.81% went of business.

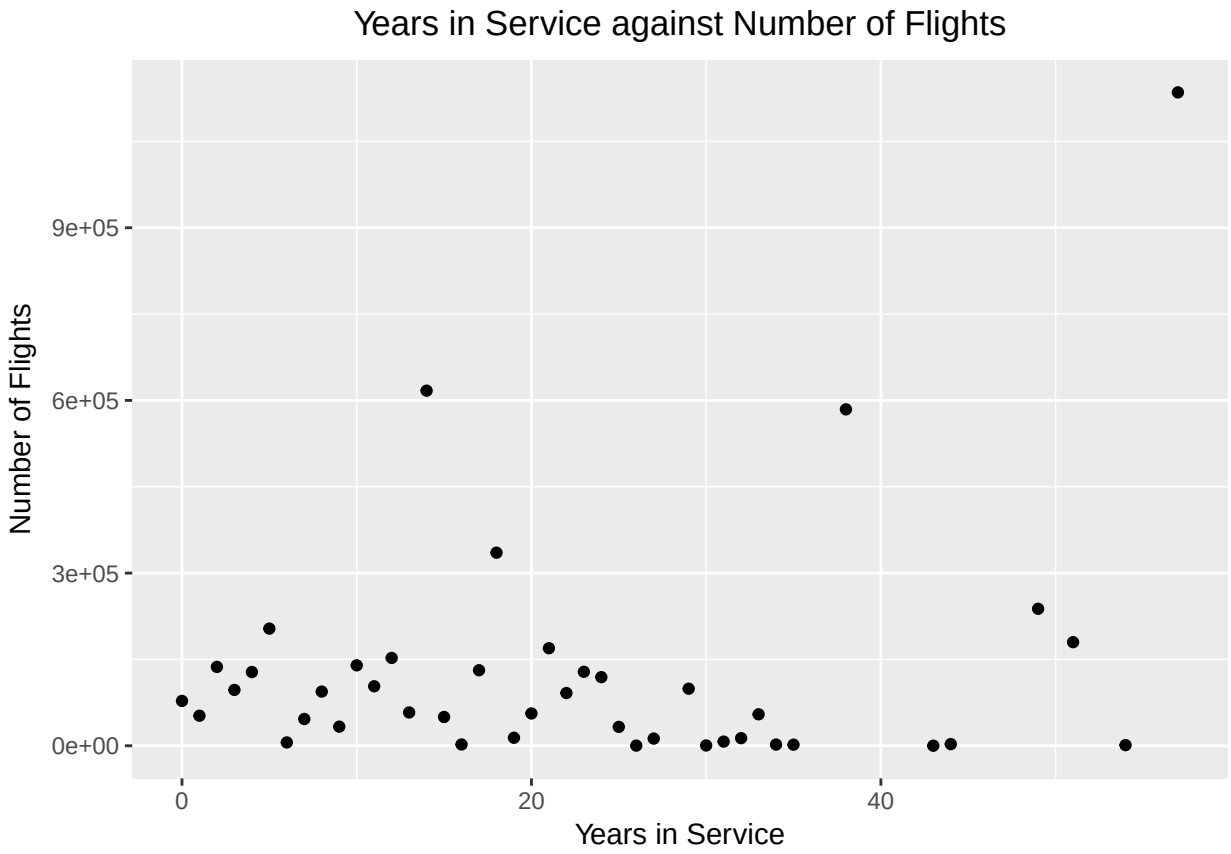
Is there any relationship between a carrier's number of years in service and its popularity?

```
combineData %>% count(yearsInService) %>% arrange(desc(n)) %>% head(3) %>%
  knitr::kable(align=c('c','c'),
               col.names = c("Years in Service", "Number"))
```

Years in Service	Number
57	1135187
14	616957
38	584490

Carriers that have been operating for many years are more popular, on average.

```
combineData %>% count(yearsInService) %>%
  ggplot(aes(x=yearsInService,y=n)) + geom_point() +
                          xlab("Years in Service") +
                          ylab("Number of Flights") +
                          ggtitle("Years in Service against Number of Flights") +
                          theme(plot.title = element_text(hjust = 0.5))
```



Yes, we do see a relationship. Most of the flights are from carriers who have been operating for a long time. However, the relationship is not that strong overall.

The top right point represents 57 years in service and are the following carriers:

```
combineData %>% filter(yearsInService==57) %>%
  group_by(unique_carrier_name) %>% count() %>% arrange(desc(n)) %>% head(3)
```

```
## # A tibble: 3 x 2
## # Groups:   unique_carrier_name [3]
##      unique_carrier_name      n
##                    <chr>  <int>
## 1   Delta Air Lines Inc. 382952
## 2  United Air Lines Inc. 360937
## 3 American Airlines Inc. 304131
```

Again, we see the same major carriers in our table of the top 6 most popular flight carriers.