BSDS 100 Case Study

Jacques Sham & Charles Siu

April 28, 2018

Part 1: Executive Summary

!!DOTO: No more than 250 words explaining, with extreme concision, your data and findings!!

Part 2: Introduction to the Data

I. Overview

San Francisco is the 13th largest city in the United States and is the Pacific Gateway of the nation. The airport owned by this city is called San Francisco International Airport, SFO, located in the Northern part of San Mateo County, near Millbrae. SFO has operated since 1927 and the first international service started in 1946. The airport is the 2nd busiest airport in California, 7th busiest airport in the United States, and 23rd busiest airport in the world. At the same time, SFO received the a lot of awards from rating agencies, such as the 3rd Best Airport in North America in 2012 and 3rd Best Airport Worldwide in 2014 from SkyTrax. There are 3 domestic terminals and 1 international terminal to the massive flow of passengers travel to different part of the United States and the Globe.

San Francisco government offers open source data on the detail information on the passenger and air traffic in SFO between 2006 and 2017. The data we downloaded from the San Francisco government includes destination or origin, airlines, terminals, and passenger count between July, 2005 and December, 2017. The data set consists `17960` rows and `10` columns.

The report is going to investigate the following facts about SFO:

1) Overall annual and monthly passengers count between 2006 and 2017
2) Passengers Count by regions
3) Passengers Count by airlines
4) Passengers traffic travelled by Low Cost Carrier
5) Passengers traffic in the airport terminals
6) Passengers Count in 1 selected domestic carrier and 1 selected international carrier

Before analyze the dataset, we have to clean the data and double check the accuarcy of the data.

```
# First, let's clean the data.
# Rename the column names
names(data) <- c("date", "operAirline", "operCode", "airline", "code", "isDomestic",
                 "region", "type", "category", "terminal", "area", "pax")

# Drop operating airline and code columns since they are insignificant
data %<>%
  select(-c(operAirline, operCode))

# Convert isDomestic to boolean
```

```
data$isDomestic %<>% recode("Domestic" = T, "International" = F)

# Reformat the dates into Date objects
data$date %<>%
  as.character() %>%
  as.yearmon("%Y%m") %>%
  as.Date()

# Get month and year values
data %<>%
  mutate(month = date %>% format("%m") %>% factor(labels = month.name),
         year = date %>% format("%Y") %>% as.numeric())
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.2
```

```
# Remove data from 2005 for easy comparison
data %<>% filter(year != 2005)

data$region %<>% recode("Central America" = "Latin America",
                        "South America" = "Latin America")
data$category %<>% recode("Other" = "Full Service")

# Change terminal into factor
data %<>%
  mutate(terminal = factor(terminal, levels = c("Terminal 1", "Terminal 2", "Terminal 3", "International
```

We have rename the column names of the dataset and convert the data structure of some column for the convenient of the analysis. Here is the explaination of each column: airline [character]: Airline name code [character]: Airline Code, code for variable "airline" isDomestic [logical]: If the flight is `domestic`,T; if `international`: F, flights from/to Canada counts as International region [character]: Geom region including `US`, `Canada`, `Mexico`, `Latin America`, `Europe`, `Middle East`, `Asia`, and `Australia / Oceania` type [character]: activity type; `Deplaned` means arrival, `Enplaned` means departure, `Thru / Transit` means flight transit at SFO category [factor]: Airline price type; `Low fare` is Low cost carrier, else are others terminal [factor]: SFO terminal area [character]: area within SFO terminal pax [int]: passenger count of given row month [factor]: The `month` of the airline operates year [double]: The `year` of the airline operates

We also have to vertify the accuracy of the data, below is how we clean the data. Below is the steps we fix the dataset:

Step 1: Standardize `United Airlines` `United Airlines` merged with `Continent Airlines` in 2013 that some data from `United Airlines` are written in `United Airlines - Pre 07/01/2013` that we have to convert this to `United Airlines` as we can tell those flights were `United Airlines`' operation.

```
# Merge "United Airlines - Pre 07/01/2013" to "United Airlines"
data$airline %<>% recode("United Airlines - Pre 07/01/2013" = "United Airlines")
```

Step 2: Standardize `Emirates` Some of `Emirates` data were written followed with a space. It looks like a typing error but it creates error when analyzing the data, so that I have to standardize all `Emirates` related data.

```
# Remove the strip of "Emirates "
data$airline %<>% recode("Emirates " = "Emirates")
```

Step 3: Re-identify the category There are a lot of `full service airlines` idenify as `low cost carrier`, and vice versa. The international official aviation organization, International Civil Aviation Organization(ICAO), has an official definition on `low cost carrier` and provides a list of `low cost carrier`. We checked the airlines in the dataset and convert some of the wrongly identified airlines according to he their list.

This is the list of low cost carrier: https://www.icao.int/sustainability/Documents/LCC-List.pdf

(Explaination what we did)

```
# Some airlines were wrongly identified in category
# Convert the below airlines to full service/Low Fare
full_svc_airline <- c("Air China", "Air India Limited", "Air New Zealand", "Air Pacific Limited dba Fiji
                      "Emirates", "United Airlines", "Virgin America", "Volaris Airlines", "Delta Air Li
                      "US Airways")

lcc_svc_airline <- c("XL Airways France", "WOW Air", "WestJet Airlines")

data %<>%
  mutate(cat_temp = ifelse(airline %in% full_svc_airline, "Full Service",
                    ifelse(airline %in% lcc_svc_airline, "Low Fare", as.character(category)))) %>%
  mutate(category = as.factor(cat_temp))
```

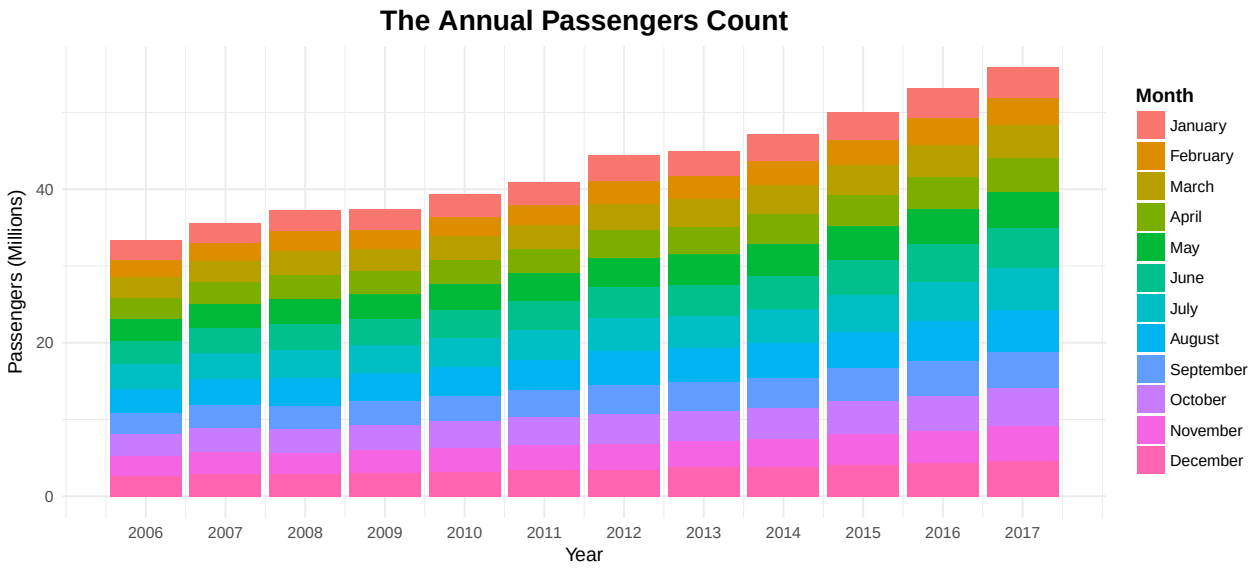
Part 3: Exploratory Analysis

I. Overview on Passenger traffic in SFO

Overall passenger traffic between 2006 and 2017

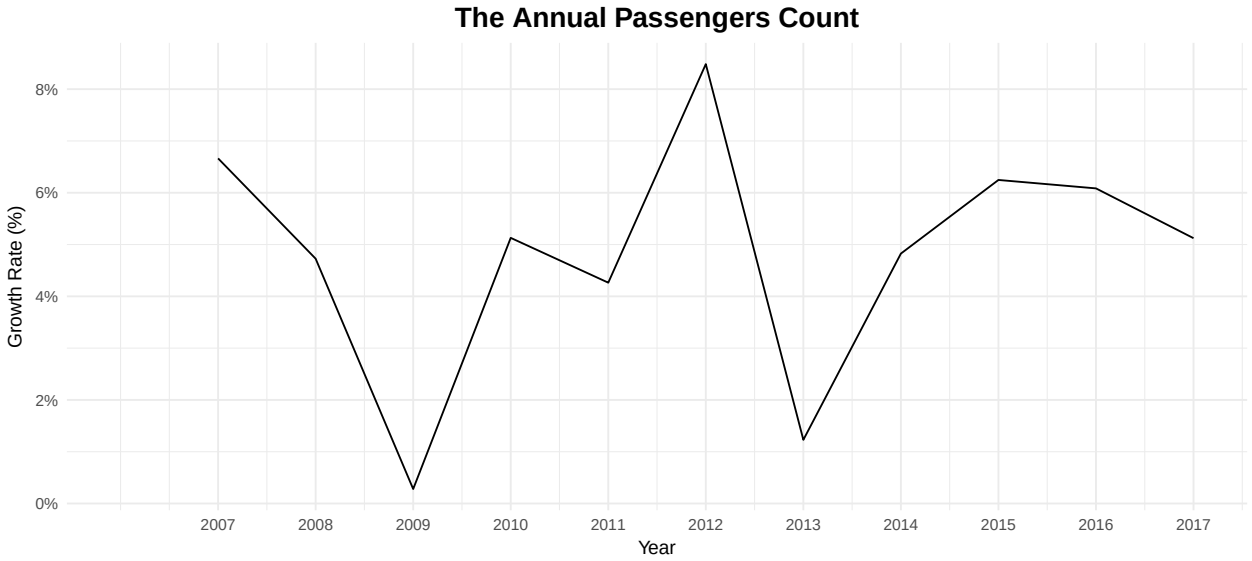
SFO is a busy airport and is a transpacific geteway in the West Coast of the United States, the passenger traffic is heavy. Below is the annual total passengers count in SFO.

```
data %>%
  ggplot() +
  geom_bar(aes(x = year, y = pax / million, fill = month), stat = "identity") +
  theme_minimal() +
  scale_x_continuous(name = "Year", breaks = seq(2006, 2017, by = 1)) +
  scale_y_continuous(name = "Passengers (Millions)") +
  scale_fill_discrete(name = "Month") +
  ggtitle("The Annual Passengers Count") +
  format_title +
  format_legend_title
```



There are 55823712 passengers traveled in SFO 2017 compared to 33332970 in 2006. We can see a upward trend on passengers count in SFO. We can confirm the trend by plotting the annual passenger growth rate in the below line chart.

```
growth_rate_year %>%
 ggplot(aes(x = year, y = growth)) +
 geom_line() +
 theme_minimal() +
 scale_x_continuous(name = "Year", breaks = seq(2007, 2017, by = 1)) +
 scale_y_continuous(name = "Growth Rate (%)", labels = percent) +
 ggtitle("The Annual Passengers Count") +
 format_title +
 format_legend_title
```



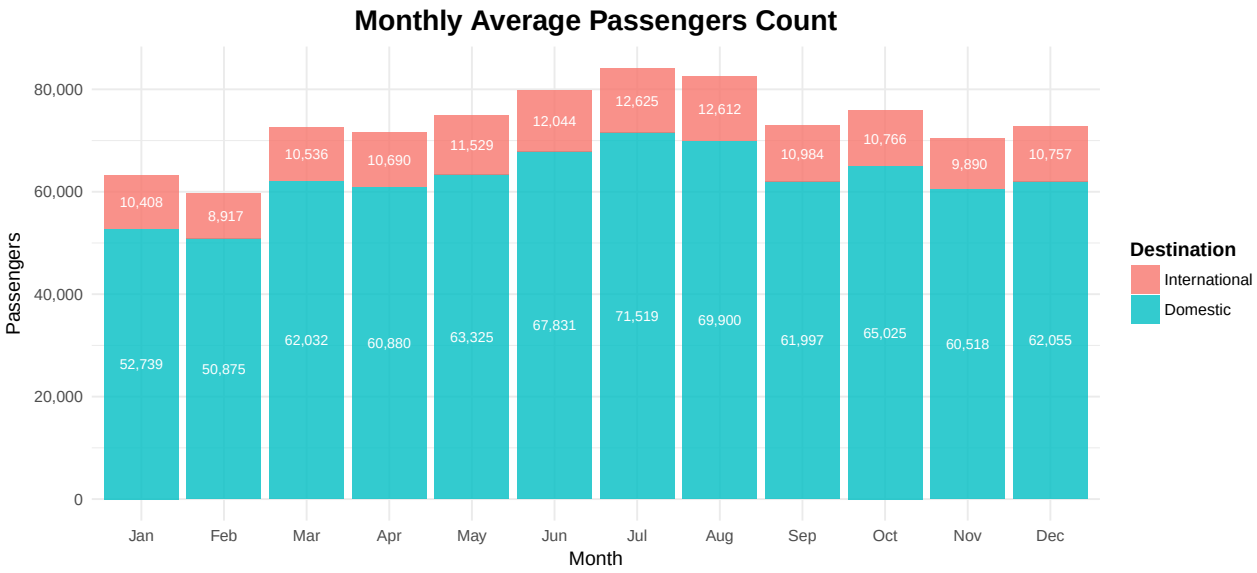
Beside in 2009 SFO experienced a slow year in passenger growth, we can confirm SFO has a upward trend on passenger growth. 2012 is the year SFO experienced the strongest growth between 2006 and 2017, with 0% growth. The average passenger growth in the period is 4.82%. It will take 15 years after 2006 for SFO to double the passenger traffic, in 2021.

Average passenger traffic between 2006 and 2017

The airline industry is season-sensitive industry. It means the season effects significant influence the passenger traffic. Generally, summer and Christmas holidays are two major peak season in the airline industry. Therefore, SFO experiences higher passenger traffic in those periods.

Below is the bar chart of monthly average total passenger count of SFO between 2006 and 2017.

```
data %>%
  group_by(isDomestic, month) %>%
  summarise(avg_pax = round(mean(pax), digit = 0)) %>%
  ggplot(aes(x = factor(month, labels = month.abb), y = avg_pax, fill = isDomestic)) +
  geom_bar(stat = "identity", alpha = 0.8) +
  theme_minimal() +
  scale_y_continuous(labels = comma) +
  scale_fill_discrete(name = "Destination", label = c("International","Domestic")) +
  labs(x = "Month", y = "Passengers") +
  ggtitle("Monthly Average Passengers Count") +
  geom_text(aes(label = format(avg_pax, big.mark = ",")), size = 2.75,
            position = position_stack(vjust = 0.5), colour = "white") +
  format_title +
  format_legend_title
```



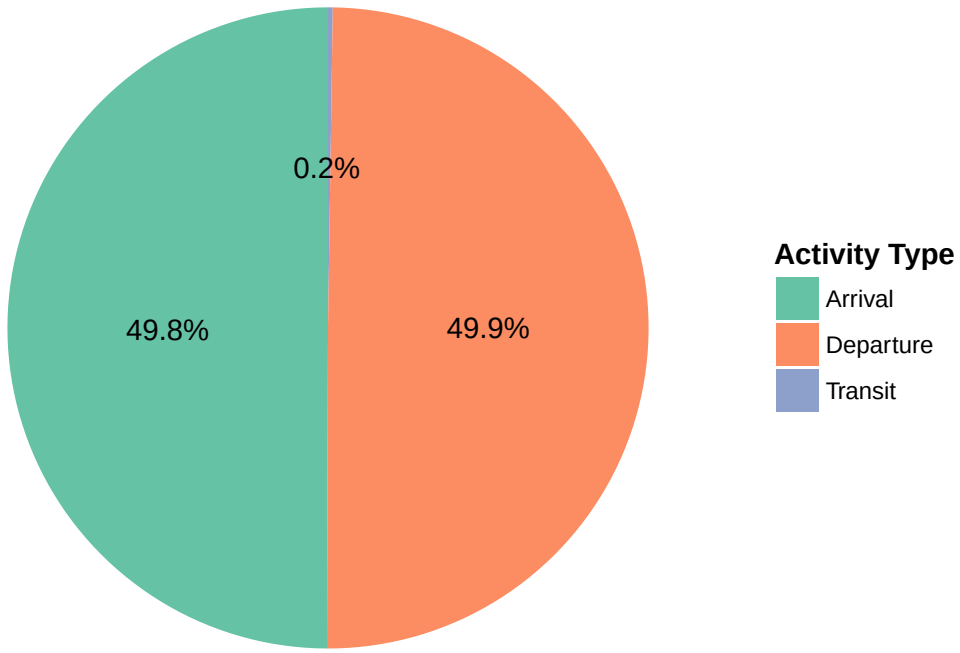
As seen on the bar chart, `June`, `July`, `August`, `October`, `December` are the months that SFO experience relatively higher passenger traffic. Beside `October`, the other high-traffic months are in the summer or Christmas holiday seasons. Interestingly, summer is the only busiest period of international traveling in or out from SFO as `June`, `July`, `August` are the only months that SFO handles more than `12000` international travelers.

Roughly 80% of passengers in SFO are domestic travelers.

Activity Type

The data set provided by the San Francisco government has the statistic on airplanes departure, arrival, and transit. Departure and arrival means airplanes depart and arrive in SFO, respectively. Transit means the passengers connect through SFO with the same airplane and the same flight number (If a passenger transfer from one plane to a different plane does not qualify "Transit" in the dataset). In the below pie chart shows the airplane activities between 2006 and 2017.

```r
total_traffic <- data %>%
  summarise(total_traffic = sum(pax)) %>%
  as.numeric()

data %>%
  group_by(type) %>%
  summarise(traffic = sum(pax)) %>%
  ggplot(aes(x = "", y = traffic, fill = type)) +
  geom_bar(width = 1 , stat = "identity") +
  geom_text(aes(x = 1, y = cumsum(traffic) - traffic / 2, label = percent(traffic / total_traffic))) +
  coord_polar(theta = "y") +
  theme_void() +
  scale_fill_brewer(palette = "Set2", name = "Activity Type", label = c("Arrival", "Departure", "Transi
  ggtitle("Percentage of Airplane Activities") +
  format_title +
  format_legend_title
```




```r
transit_list <- data %>% filter(type=="Thru / Transit") %>% count(airline) %>% arrange(desc(n))
```

There is no suprise on the close to 50%-50% split on departure and arrival passengers in SFO. Part of the reason is that the domestic carriers tend not to operate transiting flight via SFO, the portion of transit flights in SFO becomes significantly small. The top three airlines offers transit flights are United Airlines, Southwest Airlines and Alaska Airlines. At the same time, there are transit fligh operated by some foreign carriers such as Singapore Airlines' `Singapore - Hong Kong - San Francisco` route and China Southern Airlines' `Canton - Wuhan - San Francisco` route. Since SFO is the final destination of either route, therefore passengers traveled with those routes do not count toward `transit` flights.

II. Destination

As the bar chart on the monthly average passengers count shows that roughly 80% of the SFO passengers are domestic passengers. One maybe interested the destination of the remaining 20% travelers. The map below shows the passengers count in North America between 2006 and 2017.

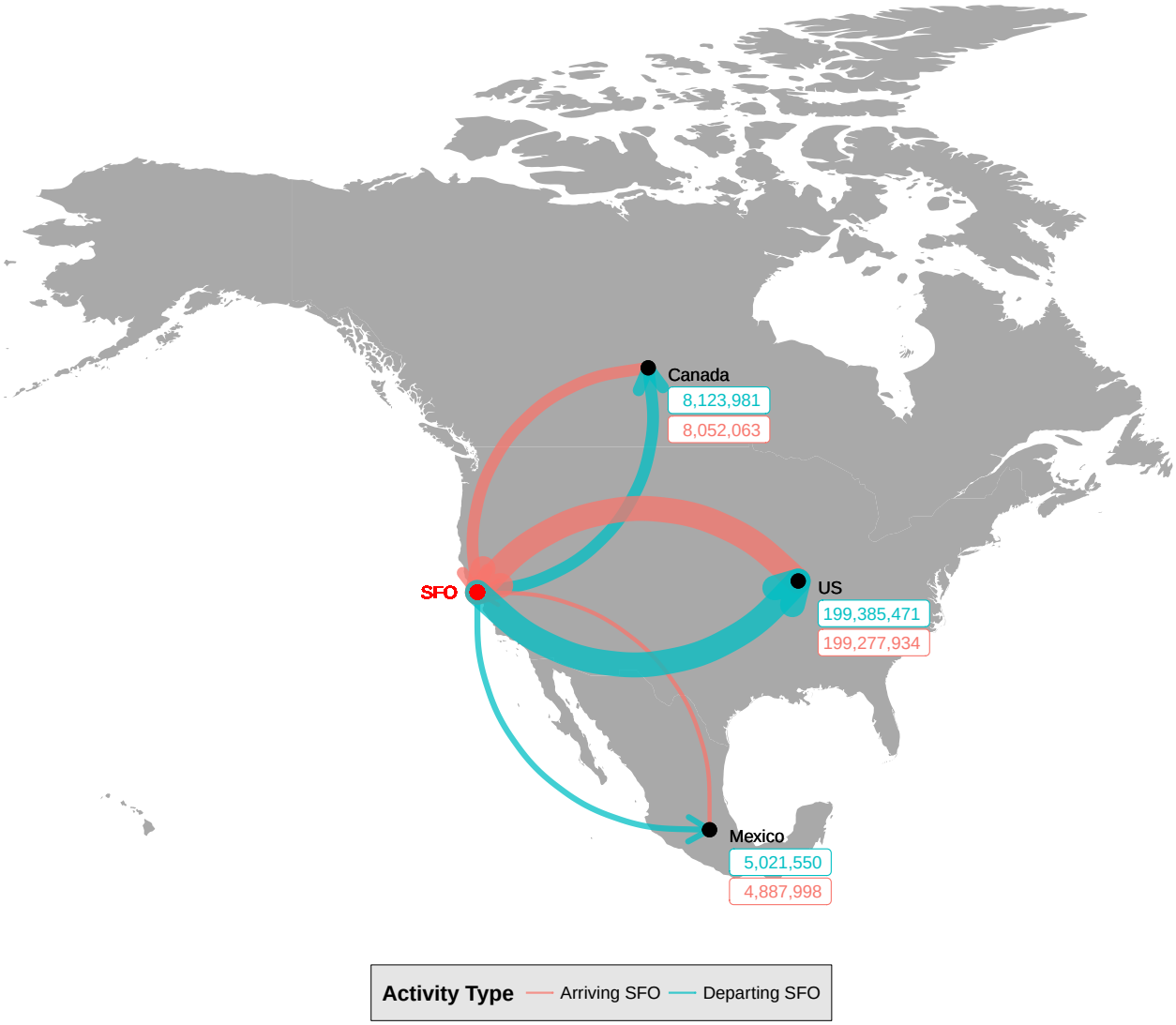
```
# Load the world map and cities
world <- map_data("world")
cities <- world.cities

# Retrieve the information of SF
sf <- world.cities %>%
  filter(name == "San Francisco" & country.etc == "USA")

# Derive the cities and corresponding regions
cities %<>%
  filter(
    (name == "Adelaide" & country.etc == "Australia") |
    (name == "La Paz" & country.etc == "Bolivia") |
    (name %in% c("Saint Louis", "La Ronge",
                 "Riyadh", "Mexico City", "Shenzhen", "Ostrava"))
    ) %>%
  mutate(region =
           ifelse(name == "Adelaide", "Australia / Oceania",
           ifelse(name == "Saint Louis", "US",
           ifelse(name == "La Ronge", "Canada",
           ifelse(name == "La Paz", "Latin America",
           ifelse(name == "Riyadh", "Middle East",
           ifelse(name == "Mexico City", "Mexico",
           ifelse(name == "Shenzhen", "Asia",
           ifelse(name == "Ostrava", "Europe", NA)))))))
  ) %>%
  full_join(y = data %>%
              mutate(region = as.character(region)) %>%
              group_by(region, type) %>%
              summarize(pax = sum(pax)), by = "region") %>%
  mutate(
    origin.lat = ifelse(type == "Enplaned", sf$lat, lat),
    origin.long = ifelse(type == "Enplaned", sf$long, long),
    dest.lat = ifelse(type == "Enplaned", lat, sf$lat),
    dest.long = ifelse(type == "Enplaned", long, sf$long)
  ) %>%
  filter(type != "Thru / Transit")

cities_na <- cities %>% filter(region %in% c("US", "Canada", "Mexico"))
cities_intl <- cities %>% filter(!(region %in% c("US", "Canada", "Mexico")))
```

```
# Geom objects for drawing static objects
draw_sf_point <- geom_point(x = sf$long, y = sf$lat, color = "red", size = 3)

draw_na_sf_label <- geom_text(aes(x = sf$long, y = sf$lat, label = "SFO"),
                              hjust = 1, nudge_x = -2, color = "red", size = 3)

draw_intl_sf_label <- geom_text(aes(x = sf$long, y = sf$lat, label = "SFO"),
```

```
                            hjust = 1, nudge_x = -5, color = "red", size = 3)

draw_color_legend <- scale_color_discrete(name = "Activity Type",
                                          labels = c("Arriving SFO", "Departing SFO"))

draw_size_legend <- scale_size_continuous(trans = "log10", guide = F)

# Functions for drawing curves, points and labels
draw_flight_curve <- function(d) {
  geom_curve(data = d %>% filter(type %in% c("Deplaned", "Enplaned")),
                            aes(x = origin.long, y = origin.lat,
                                xend = dest.long, yend = dest.lat,
                                color = type, size = pax),
                            curvature = 0.5, lineend = "round",
                            alpha = 0.75, arrow = arrow(length = unit(0.025, "npc")))
}

draw_city_points <- function(d) {
  return (geom_point(data = d, aes(x = long, y = lat), color = "black", size = 3))
}

draw_na_city_labels <- geom_text(data = cities_na, aes(x = long, y = lat, label = region),
                                 hjust = 0, nudge_x = 2, nudge_y = -0.5,
                                 color = "black", size = 3)

draw_intl_city_labels <- geom_text(data = cities_intl, aes(x = long, y = lat, label = region),
                                 hjust = 0, nudge_x = 3.5, nudge_y = -1,
                                 color = "black", size = 3)

draw_na_enplaned_labels <- geom_label(data = cities_na %>% filter(type == ("Enplaned")),
                                      aes(x = long, y = lat, label = format(pax, big.mark = ","), color = ty
                                      hjust = 0, nudge_x = 2, nudge_y = -2.5, size = 3, show.legend = F)

draw_na_deplaned_labels <- geom_label(data = cities_na %>% filter(type == ("Deplaned")),
                                    aes(x = long, y = lat - 4, label = format(pax, big.mark = ","), color =
                                    hjust = 0, nudge_x = 2, nudge_y = -.75, size = 3, show.legend = F)

draw_intl_enplaned_labels <- geom_label(data = cities_intl %>% filter(type == ("Enplaned")),
                                      aes(x = long, y = lat, label = format(pax, big.mark = ","), color = ty
                                      hjust = 0, nudge_x = 3.5, nudge_y = -7, size = 3, show.legend = F)

draw_intl_deplaned_labels <-geom_label(data = cities_intl %>% filter(type == ("Deplaned")),
                                    aes(x = long, y = lat - 4, label = format(pax, big.mark = ","), color =
                                    hjust = 0, nudge_x = 3.5, nudge_y = -9.5, size = 3, show.legend = F)

format_theme <- theme(
  axis.text = element_blank(),
  axis.line = element_blank(),
  axis.ticks = element_blank(),
  panel.border = element_blank(),
  panel.grid = element_blank(),
  axis.title = element_blank(),
  legend.position = "bottom",
```

```
  legend.background = element_rect(fill = "gray90", size = 0),
  legend.title = element_text(face = "bold")
  )
```

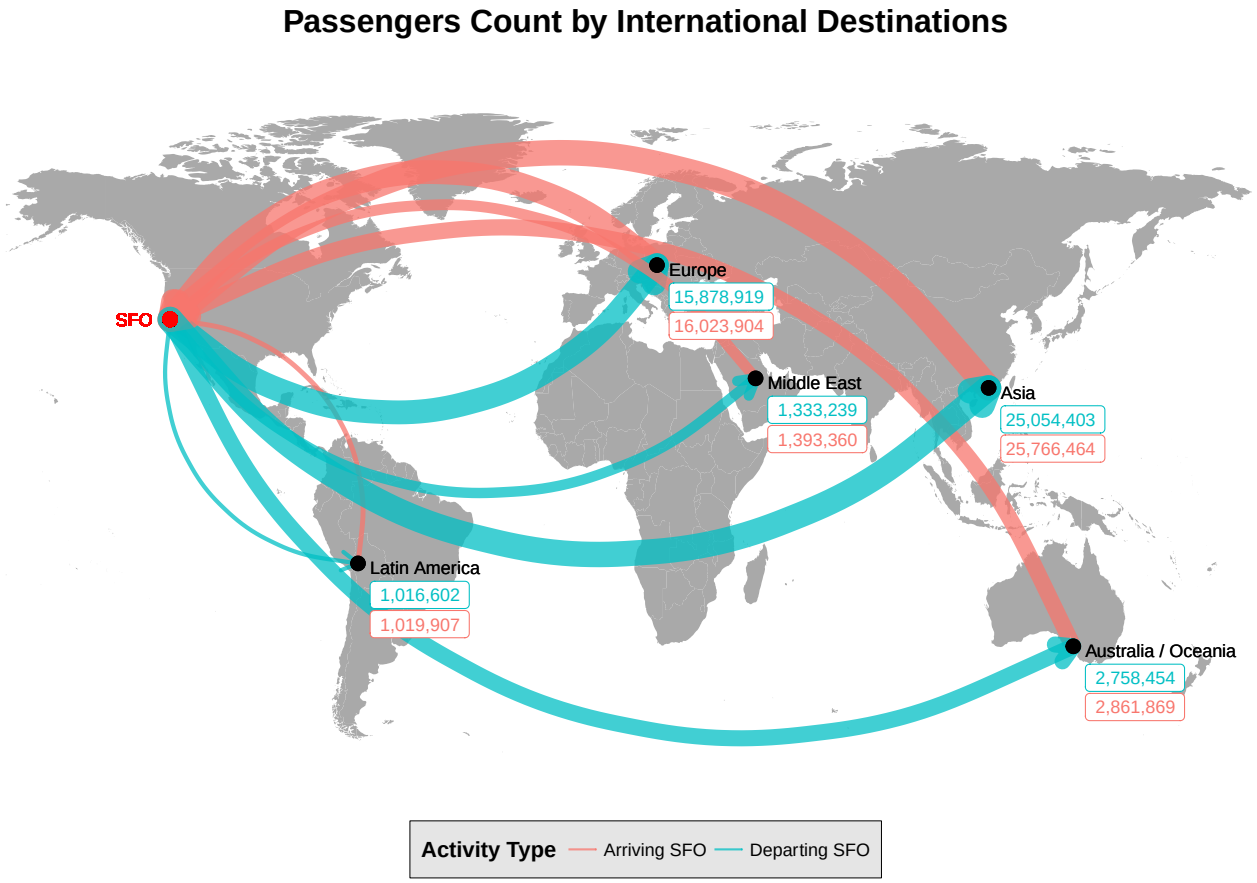
```
world %>%
  filter(region %in% c("USA", "Canada", "Mexico")) %>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, group = group), fill = "darkgray") +
  draw_flight_curve(cities_na) +
  draw_city_points(cities_na) +
  draw_na_city_labels +
  draw_sf_point +
  draw_na_sf_label +
  draw_na_enplaned_labels +
  draw_na_deplaned_labels +
  coord_fixed(1.3) +
  theme_minimal() +
  draw_color_legend +
  draw_size_legend +
  scale_fill_brewer(palette = "Set2") +
  scale_x_continuous(limits = c(-170, -50)) +
  ggtitle("Passengers Count by North American Destinations") +
  format_theme +
  format_title
```



And the map below shows the passengers count on other regions outside North America between 2006 and 2017.

```
world %>%
  filter(region != "Antarctica") %>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, group = group), fill = "darkgray") +
  draw_flight_curve(cities_intl) +
  draw_city_points(cities_intl) +
  draw_intl_city_labels +
  draw_sf_point +
  draw_intl_sf_label +
  draw_intl_enplaned_labels +
  draw_intl_deplaned_labels +
  coord_fixed(1.3) +
  theme_minimal() +
```

```
  draw_color_legend +
  draw_size_legend +
  scale_x_continuous(limits = c(-170, 200)) +
  scale_y_continuous(limits = c(-60, 90)) +
  ggtitle("Passengers Count by International Destinations") +
  format_theme +
  format_title
```



The map shows that `Asia` is the continent the most passengers flying from or to after the United States, followed by `Europe`. Surprisingly there are less passengers coming from or going to `Canada` and `Mexico` than they coming from or going to `Asia` or `Europe`, and the passenger traffic from or to `Latin America` is very little.

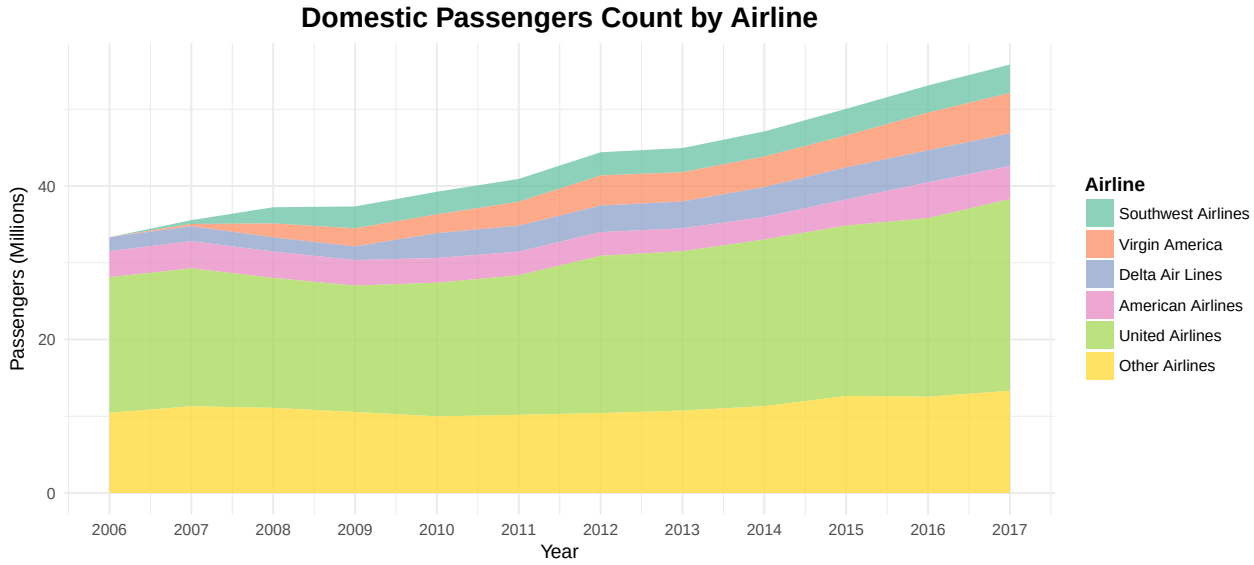
III. Airline Overview

There are a lot of airlines serving travelers in SFO. Below is the stacked line chart of the domestic passengers count by airlines.

```
# Filter the top 5 airlines by domestic passenger count
top5_dom_list <- data %>%
  filter(isDomestic) %>%
  group_by(airline) %>%
  summarise(total_pax = sum(pax)) %>%
  top_n(5, total_pax) %>%
  arrange(total_pax) %>%
```

```r
  select(-total_pax)

# Combine and compute the other airlines
other_dom_airline <- data %>%
  filter(!(airline %in% top5_dom_list$airline)) %>%
  group_by(year) %>%
  summarise(sum = sum(pax)) %>%
  mutate(airline = "Other Airlines") %>%
  select(airline, year, sum)
```

```r
data %>%
  group_by(airline, year) %>%
  summarize(sum = sum(pax)) %>%
  right_join(top5_dom_list, by = "airline") %>%
  ungroup() %>%
  rbind(other_dom_airline) %>%
  rbind(data %>%
          group_by(airline) %>%
          right_join(top5_dom_list, by = "airline") %>%
          summarize(year = min(year) - 1) %>%
          filter(year == min(data$year)) %>%
          mutate(sum = 0)) %>%
  mutate(airline = factor(airline, levels = rbind(top5_dom_list, "Other Airlines")$airline)) %>%
  ggplot() +
  geom_area(aes(x = year, y = sum / million, fill = airline), alpha = 0.75) +
  scale_x_continuous(name = "Year", breaks = seq(min(data$year), max(data$year), by = 1)) +
  scale_y_continuous(name = "Passengers (Millions)") +
  scale_fill_brewer(name = "Airline", palette = "Set2") +
  theme_minimal() +
  ggtitle("Domestic Passengers Count by Airline") +
  format_title +
  format_legend_title
```



```
airline_dom_pax <- data %>%
                      filter(isDomestic) %>%
                      group_by(airline) %>%
                      summarise(total_pax = sum(pax))


top5_dom_pax <- airline_dom_pax %>%
                   top_n(5, total_pax) %>%
                   summarise(pax = sum(total_pax))

dom_pax <- airline_dom_pax %>% summarise(pax = sum(total_pax))

top5_dom_pax_num <- top5_dom_pax[1,1]
all_dom_pax <- dom_pax[1,1]
non_top5_dom_pax <- dom_pax[1,1] - top5_dom_pax[1,1]
top5_dom_portion<- paste(round(top5_dom_pax_num/all_dom_pax,4)*100,"%",sep="")
non_top5_dom_portion <- paste(round(non_top5_dom_pax/all_dom_pax,4)*100,"%",sep="")
```

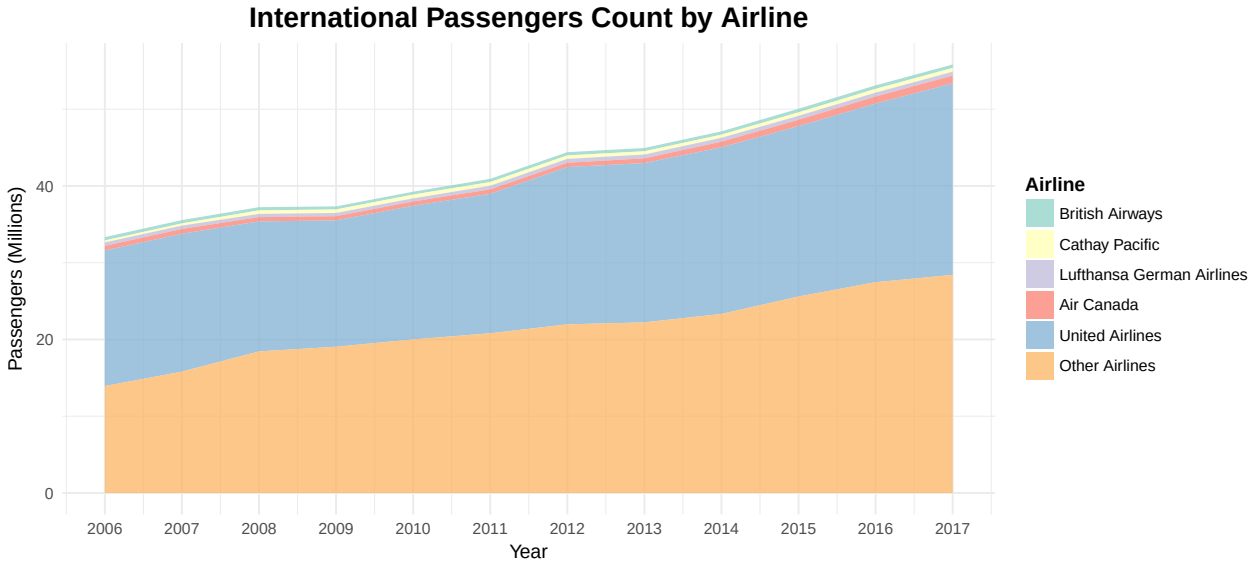
The chart shows that a large portion of SFO travelers traveling by United Airlines as SFO is one of the hub of United Airlines in the West Coast. Southwest Airlines, Virgin America(Recently Merged with `Alaska Airlines`), Delta Air Lines, American Airlines are the other top 5 major carrier in SFO in passenger count. A large portion of travelers in SFO travel with those airlines, 85.43% of all travelers traveled with those airlines between 2006 and 2017.

Below is the stacked line chart of the international passengers count by airlines.

```
# Filter the top 5 airlines by international passenger count
top5_intl_list <- data %>%
  filter(!isDomestic) %>%
  group_by(airline) %>%
  summarise(total_pax = sum(pax)) %>%
  top_n(5, total_pax) %>%
  arrange(total_pax) %>%
  select(-total_pax)

# Combine and compute the other airlines
other_intl_airline <- data %>%
  filter(!(airline %in% top5_intl_list$airline)) %>%
  group_by(year) %>%
  summarise(sum = sum(pax)) %>%
  mutate(airline = "Other Airlines") %>%
  select(airline, year, sum)
```

```
data %>%
  group_by(airline, year) %>%
  summarize(sum = sum(pax)) %>%
  right_join(top5_intl_list, by = "airline") %>%
  ungroup() %>%
  rbind(other_intl_airline) %>%
  rbind(data %>%
          group_by(airline) %>%
          right_join(top5_intl_list, by = "airline") %>%
          summarize(year = min(year) - 1) %>%
```

```
          filter(year == min(data$year)) %>%
          mutate(sum = 0)) %>%
  mutate(airline = factor(airline, levels = rbind(top5_intl_list, "Other Airlines")$airline)) %>%
  ggplot() +
  geom_area(aes(x = year, y = sum / million, fill = airline), alpha = 0.75) +
  scale_x_continuous(name = "Year", breaks = seq(min(data$year), max(data$year), by = 1)) +
  scale_y_continuous(name = "Passengers (Millions)") +
  scale_fill_brewer(name = "Airline", palette = "Set3") +
  theme_minimal() +
  ggtitle("International Passengers Count by Airline") +
  format_title +
  format_legend_title
```




```
airline_intl_pax <- data %>%
                      filter(!isDomestic) %>%
                      group_by(airline) %>%
                      summarise(total_pax = sum(pax))


top5_intl_pax <- airline_intl_pax %>%
                  top_n(5, total_pax) %>%
                  summarise(pax = sum(total_pax))

intl_pax <- airline_intl_pax %>% summarise(pax = sum(total_pax))

all_intl_pax <- intl_pax[1,1]
non_top5_intl_pax <- intl_pax[1,1] - top5_intl_pax[1,1]
non_top5_intl_portion <- paste(round(non_top5_intl_pax/all_intl_pax,4)*100,"%",sep="")
```

Air Canada is largest foreign carrier operating international flights in SFO, serving flights between San Francisco and Vancouver, Victoria, Edmonton, Calgary, Toronto, and Montréal. The other major foreign carrier including Lufthansa German Airlines, British Airways, and Cathay Pacific from Frankfurt and Munich, Germany, London Heathrow, Britain, and Hong Kong, respectively.

A very large portion of international travelers not traveling with the top 5 airlines: 44.85%. The portion of international travelers not traveling with the top 5 airlines is significantly higher than the portion of domestic

travelers not traveling with the top 5 airlines, which is 14.57%. The airline industry monopoly effect is more dominant in domestic flights than in the international flights.

```
top_area <- data %>%
  group_by(region, airline) %>%
  summarise(pax_flow = sum(pax)) %>%
  top_n(1, pax_flow)
```

III. Low Cost Carrier vs Full Service Carrier

Due to the advance of avaition technology, the cost of air travel significantly decrease in the last 100 years. Low cost carriers provide low cost traveling by significant ticket price to attract passengers; to make up the loss of revenue, `low cost carriers` cut cost on meals, seating comfort, onboard entertainment, and seating priority. Alternatively, airlines that provide decent meals, seating priority, and concern on seating comfort and onboard entertainment are called `full Service Carrier` For travelers travel without too much conceern on those feature, traveling with low cost carrier is a good choice for those travelers. Southwest Airlines is one of the sucessful story: Since the deregulation on domestic airline industry in United States, Southwest Airlines provides frequent short-haul domestic flights and became one of the largest airline in the United States. After the success of the Southwest Airlines story, there are more low cost carriers flying into SFO for the last 40 years. In 2016, Iceland-based low cost carrier, Wow Air, began scheduled service from Reykjavik, Iceland to SFO that makes Wow Air the first low cost carrier operate regular scheduled service to SFO outside of the North America.

```
lcc_pax <- data %>% filter(year==2017) %>% group_by(category) %>% summarise(sumpax = sum(pax)) %>% muta

lcc_pax_prct <- paste(round(lcc_pax[2,"prct_pax"],4)*100,"%",sep="")
lcc_pax_count <- lcc_pax[2,"sumpax"]

southwest_count <- data %>% filter(airline=="Southwest Airlines") %>% group_by(year) %>% summarise(year

southwest_pax_growth <- southwest_count %>% filter(year >= 2010) %>% mutate(growth = growth_rate(yearpa
southwest_pax_y_growth <- southwest_pax_growth  %>% summarise(mean_growth = mean(growth, na.rm=T))

wowair_pax_growth <- data %>% filter(airline=="WOW Air") %>%
  group_by(year) %>% summarise(yearpax = sum(pax)) %>%
  mutate(growth = growth_rate(yearpax)) %>% summarise(mean_growth = mean(growth, na.rm=T))
```

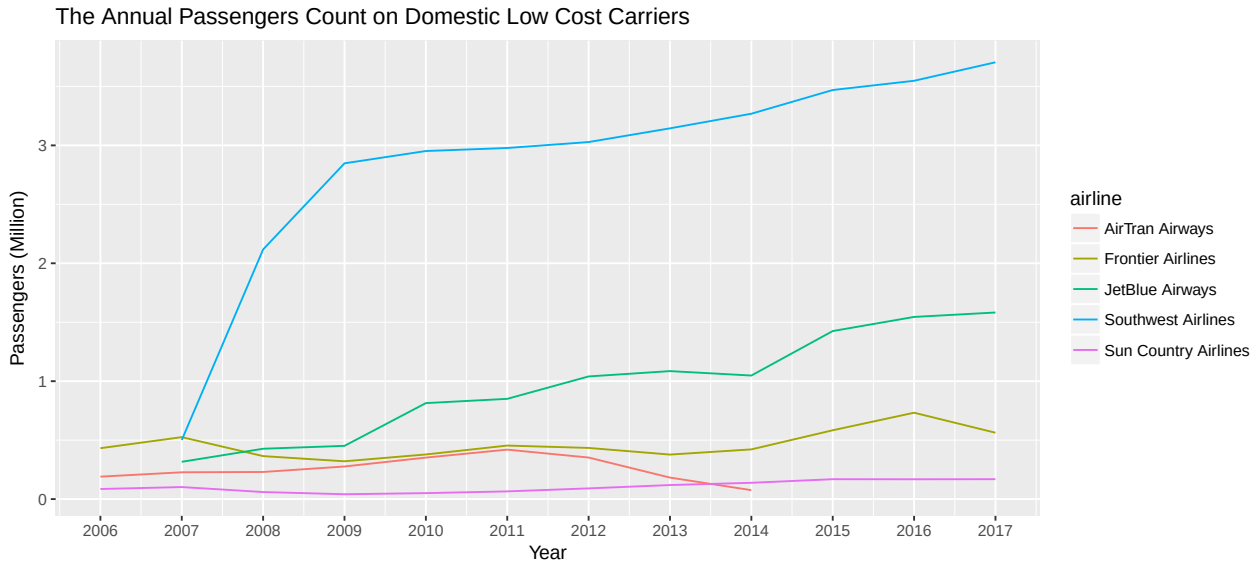
In 2017, about 6317794 passengers traveled via SFO by low cost carriers, which makes up about 11.32% of the total passengers count.

The below line chart illurates the passenger counts among domestic low cost carriers, includes `AirTran Airways`, `Frontier Airlines`, `JetBlue Airways`, `Southwest Airlines`, `Sun Country Airlines`.

```
dom_lcc_plot <- data %>%
              filter(isDomestic &
                       category=="Low Fare" &
                       airline != "ATA Airlines" &
                       airline != "Allegiant Air" &
                       airline != "Servisair" &
                       airline != "Spirit Airlines" &
                       airline != "Trego Dugan Aviation") %>%
              group_by(year,airline) %>%
              summarise(sumpax = sum(pax)/1000000) %>%
```

```
              arrange(desc(sumpax))

## dataframes for plotting
intl_lcc_plot <- data %>%
                  filter(!isDomestic &
                           category=="Low Fare" &
                           airline != "ATA Airlines" &
                           airline != "Servisair") %>%
                  group_by(year, airline) %>%
                  summarise(sumpax = sum(pax)) %>%
                  arrange(desc(sumpax))

intl_lcc_plot_2017 <- data %>%
                  filter(!isDomestic &
                           category=="Low Fare" &
                           airline != "ATA Airlines" &
                           airline != "Servisair" &
                           airline != "Sun Country Airlines" &
                           year == 2017) %>%
                  group_by(airline) %>%
                  summarise(sumpax = sum(pax)) %>%
                  arrange(desc(sumpax))
```

```
dom_lcc_plot %>% ggplot(aes(x = year, y = sumpax, group = airline, color = airline)) +
        geom_line() +
        labs(x = "Year", y = "Total Passenger Count")+
        ggtitle("The Annual Passengers Count on Domestic Low Cost Carriers") +
        scale_x_continuous(name = "Year", breaks = seq(min(data$year), max(data$year), by = 1)) +
        scale_y_continuous(name = "Passengers (Million)", breaks = seq(0,max(dom_lcc_plot$sumpax), by =
```

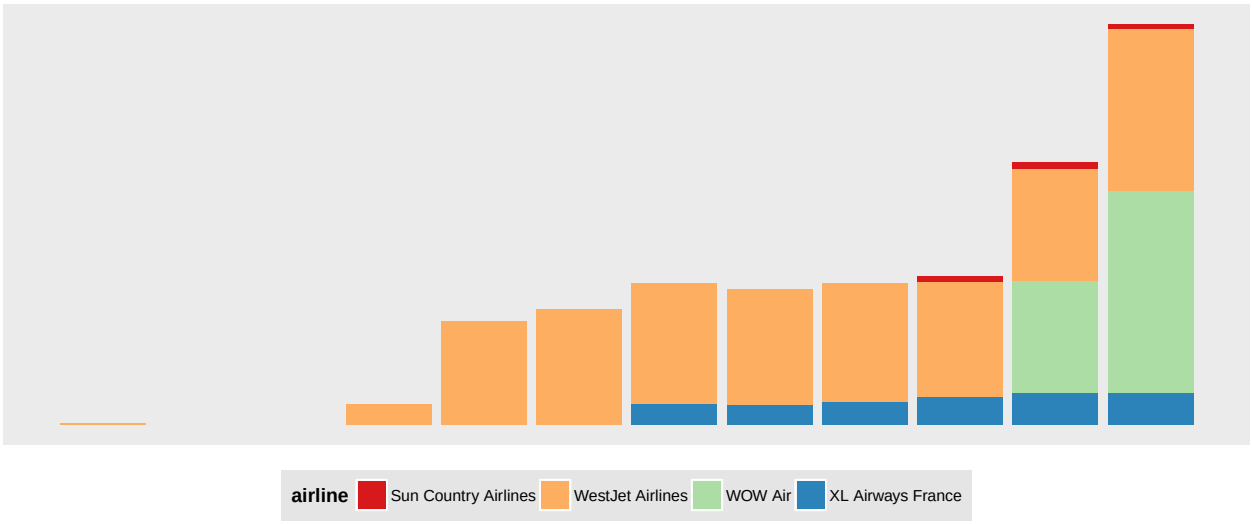


Prior to 2007, not a lot of passengers traveled by low cost carriers via SFO. The passengers count skyrocketed between 2007 and 2009 due to low cost carrier giant `Southwest Airlines` began service in SFO. `Southwest Airlines` served 500926,2117364,2847732 passengers in `r 2007, 2008, and 2009, respectively. After 2009, the annual passenger growth 3.31%. In 2017, 3704789 passengers traveled with`Southwest Airlines‘.

JetBlues is the second largest low cost carrier and provides regular service to Long Beach, CA, Fort Lauderdale, FL, Boston, and New York JFK. However, JetBlues only provides about 10 daily flights between SFO and above destinations. However, the JetBlues'flight frequency in SFO is significantly less than Southwest Airlines' flight frequency in SFO; for example, Southwest Airlines provides 58 daily flights between SFO and all Los Angeles area airports include Los Angeles, Burbank, Santa Ana, and Ontario, CA.

There were not a lot of international low cost carrier service to SFO. Before 2016, WestJet Airlines and XL Airways France the only international low cost carrier operate seasonal route from Calgary and Vancouver to SFO, and Paris de Gauelle to SFO. In 2016, Wow Air annouced began scheduled service from Reykjavik, Iceland to SFO, about 5 times per week. Below is the bar chart on the passengers count on international low cost carriers.

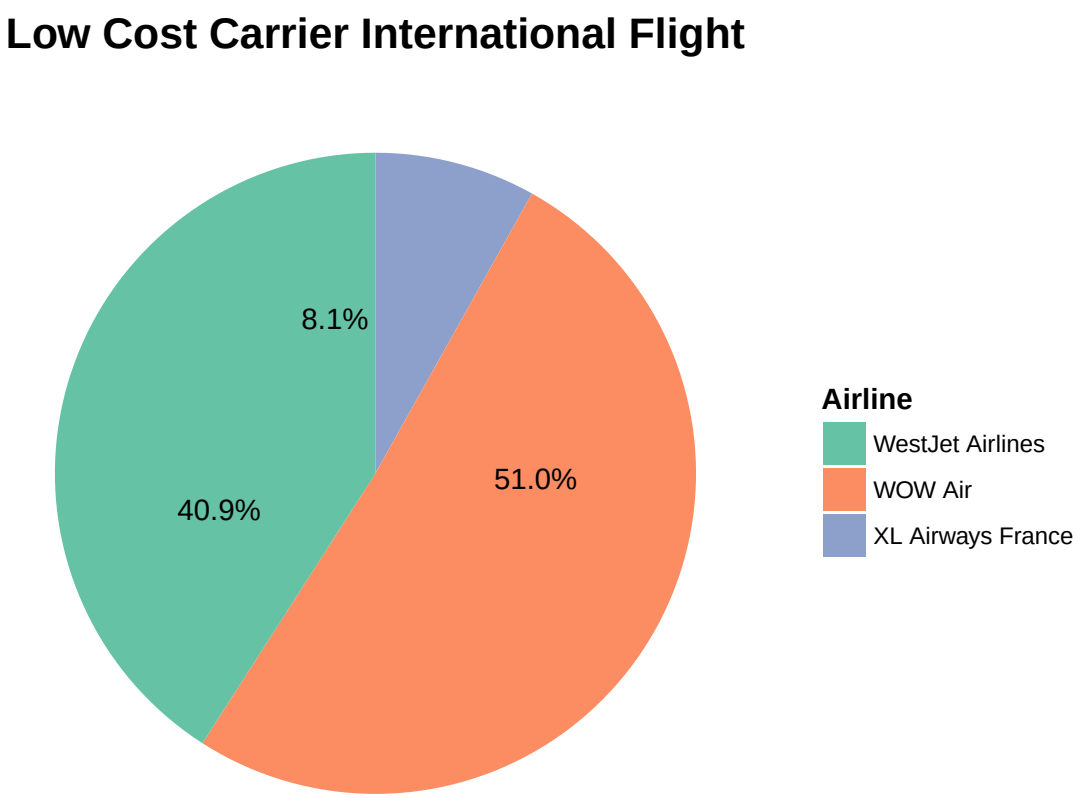
```
intl_lcc_plot %>% ggplot(aes(x = year, y = sumpax, fill = airline)) +
                    scale_fill_brewer(palette = "Spectral") +
                    geom_bar(stat = "identity") +
                    scale_x_continuous(name = "Year", breaks = seq(min(data$year), max(data$year), by =
                    scale_y_continuous(name = "Passengers", breaks = seq(0,500000,by=50000))+
                    ggtitle("The Annual Passengers Count on International Low Cost Carrier")+
                    format_theme +
                    format_title
```



There were not a lot of passengers traveled with international low cost carrier prior to 2016. The passengers count almost double when Wow Air began service to SFO. The first year passenger growth of Wow Air was 81.2%. However, Wow Air's service to SFO is new, we expect the growth rate would smooth out to single digit by 2019 if we observe the pattern from Southwest Airlines' first 3 years of service in SFO. Below is the market share on international low-cost flight service.

```
total_intl_lcc_pax_2017 <- intl_lcc_plot_2017 %>%  summarise(total_pax = sum(sumpax))
total_intl_lcc_pax_2017 <- as.numeric(total_intl_lcc_pax_2017)

intl_lcc_plot_2017 %>%
  ggplot(aes(x = "", y = sumpax, fill = airline)) +
  geom_bar(width = 1 , stat = "identity") +
  geom_text(aes(x = 1, y = cumsum(sumpax) - sumpax / 2, label = percent(sumpax / total_intl_lcc_pax_201
  coord_polar(theta = "y") +
  theme_void() +
```

```
  scale_fill_brewer(palette = "Set2", name = "Airline") +
  ggtitle("Low Cost Carrier International Flight") +
  format_title +
  format_legend_title
```



The above pie chart confirmed `Wow Air` has the largest market share on interntional low cost flights, follow by `WestJet Airlines` and `XL Airways France`.

Although many travelers do not travel with low cost carriers, there are increasingly more passengers travel with low cost carriers. Therefore, we expect the portion of passengers traveled with low cost carrier increase.

IV. Terminal Traffic

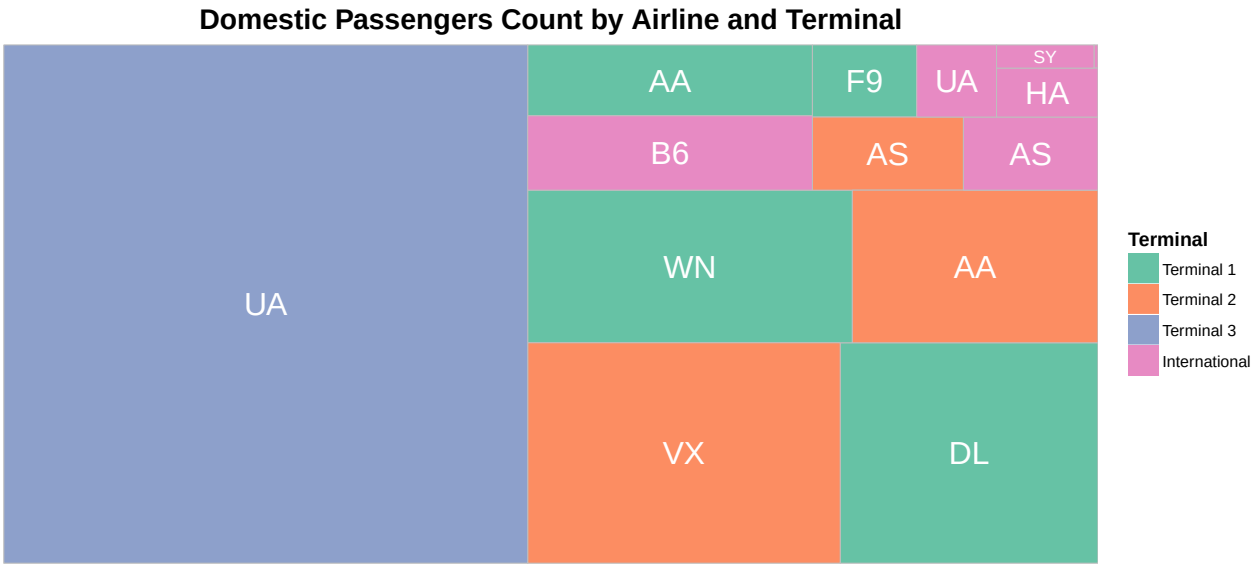
There are 4 terminals in SFO along with 115 gates: Terminal 1 (T1), Terminal 2 (T2), Terminal 3(T3), and International Terminal(IT). T1, T2, and T3 are designed to handle domestic and precleared flights from Canada, and IT are to handle international flgiths.

The oldest terminal is T2, formerly Center Terminal, which completed in 1954. Followed by is T1, formerly South Terminal, which was built in 1963. SFO expanded for the third time in 1979 by adding T3, formerly North Terminal. IT is the newest terminal in SFO, that is built in 2000. Recently, SFO is rebuilding T1, and scheduled to be completed by 2024.

Below is the tree map on domestic passenger count by terminal in 2017.

```
data %>%
  filter(isDomestic, !is.na(code) & year == 2017) %>%
  group_by(terminal, airline, code) %>%
  summarise(all_pax = sum(pax)) %>%
  ggplot(aes(area = all_pax, fill = terminal, label = code, group = airline)) +
```

```
  geom_treemap() +
  geom_treemap_text(colour = "white", place = "centre") +
  scale_fill_brewer(name = "Terminal", palette = "Set2") +
  ggtitle("Domestic Passengers Count by Airline and Terminal") +
  format_title +
  format_legend_title
```




```
pax_terminal <- data %>% filter(isDomestic & year == 2017 & terminal != "other") %>%
                  group_by(terminal) %>%
                  summarise(sumpax = sum(pax)) %>%
                  mutate(prct_pax = sumpax/sum(sumpax)) %>%
                  arrange(desc(prct_pax))
```

As we can see from the tree map: `United Airlines` is assigned to dock at T3, and some flights dock at IT. `Delta Airlines`, `Southwest Airlines`, `Frontier Airlines` and partial `American Airlines` flights were docking at T1, and T2 serves `Alaska Airlines`, `Virgin Airlines`, and the remaining `American Airlines` flights. Due to the constraint capcity of T1, T2, T3, the remaining domestic carriers with less flight frequency to SFO, including `Hawaiian Airlines`, `Jetblue Airways`, `Sun Country Airlines`, and some `Alaska Airlines` flights, are docking in IT even those are domestic flights.

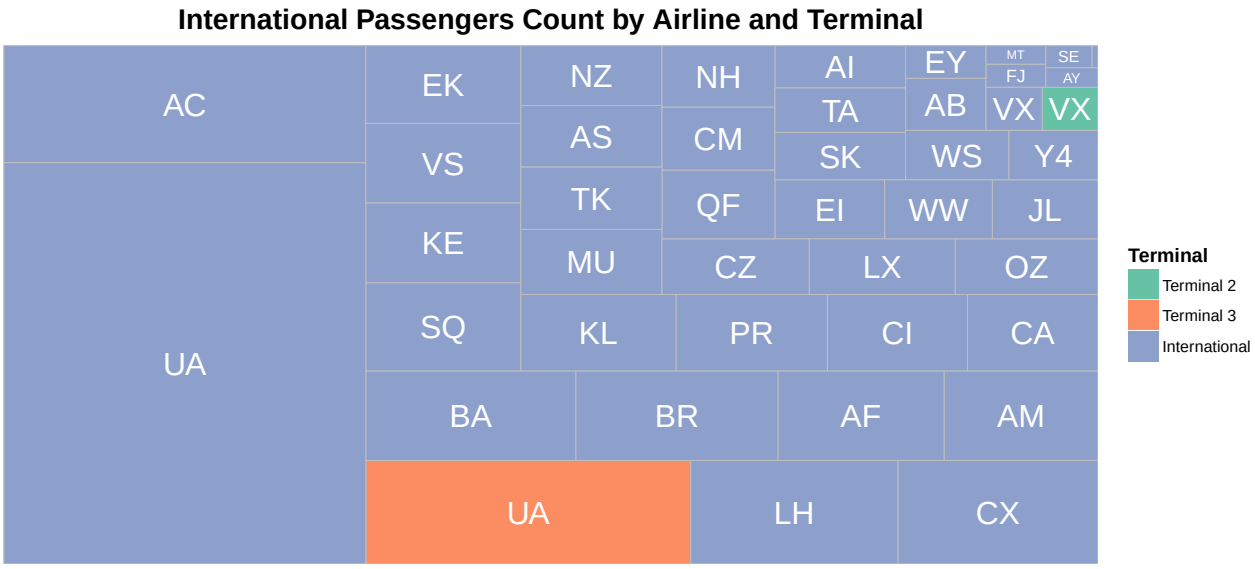
About 47.88% passengers traveled via Terminal 3 in 2017 that makes Terminal 3 the busiest terminal in SFO, followed by Terminal 1 with 23.64% of the domestic passengers. The remaining domestic passenger traveled in Terminal 2 and International, those make up of 20.7% and 7.77% of the domestic passenger traffic, respectively.

Almost half of the domestic passengers in 2017 traveled with `United Airlines` which makes up the largest share of the domestic flight market, followed by the major domestic airlines includes `American Airlines`, `Alaska Airlines`, `Delta Airlines`, `Virgin America`, and the low cost carrier giant `Southwest Airlines`

The below tree map illruates international passenger count by terminal in 2017.

```
data %>%
  filter(!isDomestic, !is.na(code) & year == 2017) %>%
  group_by(terminal, airline, code) %>%
  summarise(all_pax = sum(pax)) %>%
  ggplot(aes(area = all_pax, fill = terminal, label = code, group = airline)) +
```

```
  geom_treemap() +
  geom_treemap_text(colour = "white", place = "centre") +
  scale_fill_brewer(name = "Terminal", palette = "Set2") +
  ggtitle("International Passengers Count by Airline and Terminal") +
  format_title +
  format_legend_title
```




```
intl_table <- data %>% filter(!isDomestic & terminal != "International" & year ==2017) %>%
                       group_by(airline,region) %>%
                       summarise(sumpax =sum(pax)) %>%
                       arrange(desc(sumpax))
```

There is no surpise on the majority of the international passengers travel via IT. However, we see a small portion of `United Airlines` and `Virgin America` passengers travel via in T3 and T2. Those international flights dock at T2 and T3 are precleared flights between Canada and Mexico. That means passengers' passport and custom are cleared in Canada or Mexico by both the US Federal and either Canadian or Mexican government, so that those precleared flights are allowed to dock at domestic terminals in SFO.

```
# data %>%
#   filter(!is.na(code)) %>%
#   group_by(terminal, airline, code) %>%
#   summarise(all_pax = sum(pax)) %>%
#   ggplot(aes(area = all_pax, fill = terminal, label = code, group = airline)) +
#   geom_treemap() +
#   geom_treemap_text(colour = "white", place = "centre") +
#   scale_fill_brewer(name = "Terminal", palette = "Set2") +
#   ggtitle("Passengers Count by Airline and Terminal") +
#   format_title +
#   format_legend_title
```

We can see most of the domestic passengers traveled with `United Airlines` at T3. The remaining domestic passengers travel via T1, T2, and IT. Although we expect all international passengers travel via IT, we found the interesting fact that there are partial portion of 2017 international passengers travel via T2 and T3 because precleared flights to or from Canada or Mexico are allowed to dock in domestic terminals.

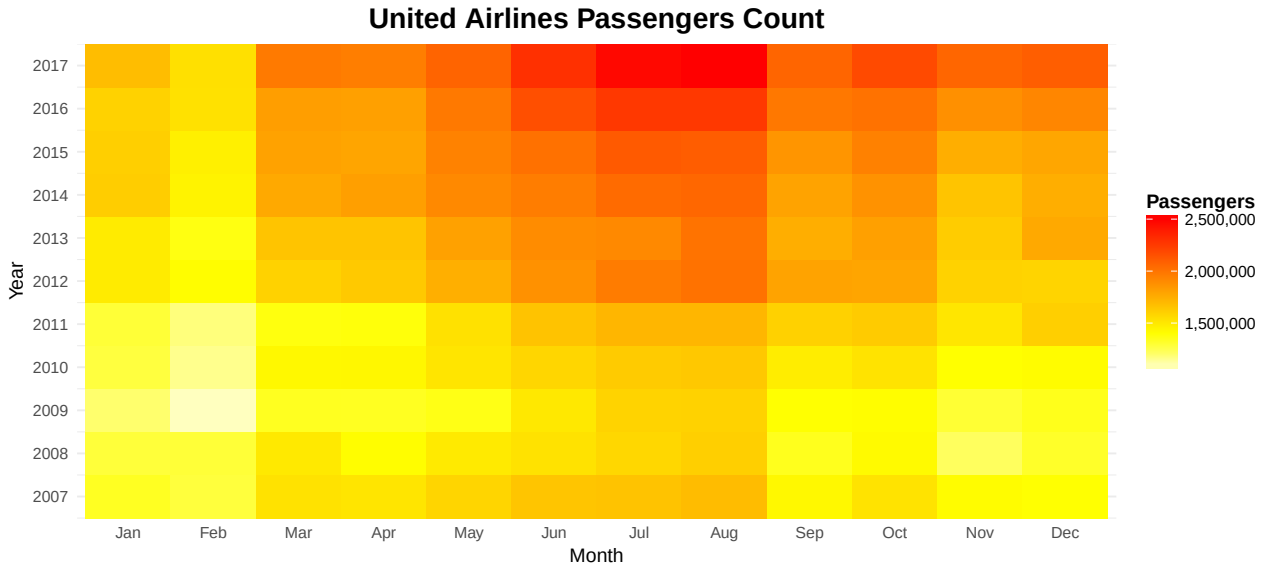
VI. Airlines

`United Airlines` is largest carrier in SFO in terms of passengers count in 2017, no matter in domestic or international flights since `United Airlines` set SFO as one of its hub that `United Airlines` has assigned a lot of flights fly in and out from SFO. In `United Airlines`' prespective, SFO is the 5th largest hub in terms of number of flights, and the primary hub for West Coast.

`United Airlines`' footstep in SFO can be traced back in 1937, `United Airlines` operated scheduled service to Los Angeles and New York in January 1937 after it was formed in 1934. And `United Airlines` has one of the largest single aircraft maintence bases in SFO.

The below heated map show the passenger traffic between 2007 and 2017 for `United Airlines`:

```
data %>%
  filter(code == "UA" & year >= 2007) %>%
  group_by(month, year) %>%
  summarise(Passengers = sum(pax)) %>%
  ggplot(aes(x = factor(month, labels = month.abb), y = year)) +
  geom_tile(aes(fill = Passengers)) +
  scale_x_discrete(name = "Month") +
  scale_y_continuous(expand = c(0, 0), name = "Year", breaks = seq(min(data$year), max(data$year), by =
  scale_fill_gradientn(colours = rev(heat.colors(10)), labels = comma) +
  theme_minimal() +
  ggtitle("United Airlines Passengers Count") +
  format_title +
  format_legend_title
```



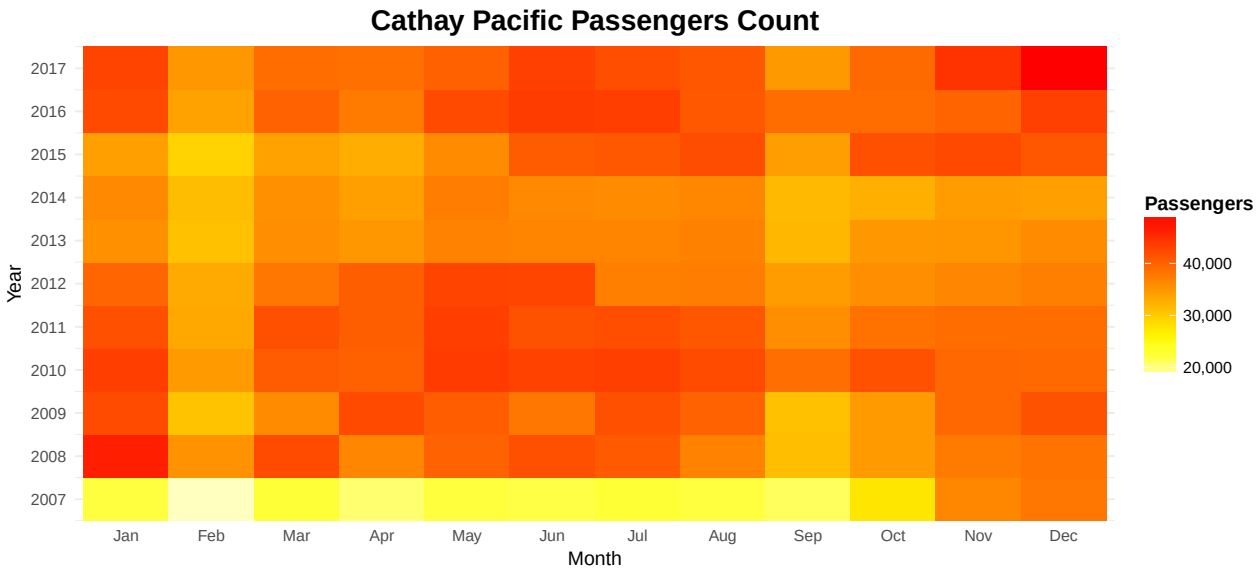
The above heated map shows the passenger traffic is relatively low by 2012. Also, we can also confirm that there are more passengers travel by `United Airlines` in summer period as `June`, `July`, and `August` of every year tend to have a darker colored pattern.

`Cathay Pacific` is a Scotish, Hong Kong-based airline. It was founded in 1946 in Hong Kong after the founders fled from Shanghai with their DC-3s, `Besty` and `Nikki`. Rather than calling "Hong Kong Airlines", the founders choose "Cathay" which is a medival word to represent "China"; Pacific is added in the name as the founders help one day `Cathay Pacifc` operates route across the Pacific Ocean from Hong Kong. Their dream came true in 1983, `Cathay Pacific` operated its first trans-Pacific flight from Hong Kong to Vancouver; in 1986, `Cathay Pacific` added SFO as the final destination in 1986 making SFO the first US destination of

`Cathay Pacific`. In the 90's, `Cathay Pacific` discontinue the flight to SFO; `Cathay Pacific` started to operate daily direct flight from Hong Kong to SFO in 1999.

The below heated map show the passenger traffic between 2007 and 2017 for `Cathay Pacific`:

```
data %>%
  filter(code == "CX" & year >=2007) %>%
  group_by(month, year) %>%
  summarise(Passengers = sum(pax)) %>%
  ggplot(aes(x = factor(month, labels = month.abb), y = year)) +
  geom_tile(aes(fill = Passengers)) +
  scale_x_discrete(name = "Month") +
  scale_y_continuous(expand = c(0, 0), name = "Year", breaks = seq(min(data$year), max(data$year), by =
  scale_fill_gradientn(colours = rev(heat.colors(10)), labels = comma) +
  theme_minimal() +
  ggtitle("Cathay Pacific Passengers Count") +
  format_title +
  format_legend_title
```



The passenger traffic of `Cathay Pacific` has a big jump between 2007 and 2008 due to `Cathay Pacific` increase their flight frequency from 1 daily flight to 2 daily flights. While `Cathay Pacific` announced expansion in 2017, `Cathay Pacific` will operate 3 daily flights between Hong Kong and SFO and we expect more passengers traveling `Cathay Pacific` in the future.

The seasonal effect on `Cathay Pacific` is little. The colored pattern across each year are very similiar that suggests the passengers count with `Cathay Pacific` is relatively more consistent across a year.

Part 4: Conclusions and Future Analysis

(Some conclusion on our analysis)

However, the dataset we have for this report focus mainly on region the flights fly to or from, there was not a lot of information on destination/origin cities and country, aircraft models. In this report, we are able to trace the destination/origin cities and country of foreign carriers from their website; we are not able to trace destination/origin cities and country for domestic carriers. In the future, we would like to explore more on destination/origin cities and country and aircraft models each flight.