San Francisco International Airport Passenger Traffic

Jacques Sham & Charles Siu

May 01, 2018

Part 1: Executive Summary

San Francisco is the Pacific gateway of the United States. The San Francisco International Airport is one of the greatest expense of the city and county government. We would like to know passenger traffic in the airport to check whether the tax revenue is spent wisely. San Francisco City and County Government provides open source dataset on passenger counts on airlines and regions in the airport. The dataset consists of more than 17000 entries between 2005 and 2017 on passenger count. Each entry consists the name of the airline and destination or origin region. We summarised the dataset to find out the passenger traffic related with airlines, origin or destination, and time. This analysis allows San Francisco taxpayers on understanding the usage on air travel infrastructure and airport authoritative to propose an investment on airport expansion.

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. SFO is the IATA airport code of the San Francisco International Airport defined by the International Air Transport Association (IATA), and is also the abbreviation of the airport. 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 the air travel 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.

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 of `17,959` rows and `12` columns. Here is the link of the dataset: https://data.sfgov.org/Transportation/Air-Traffic-Passenger-Statistics/rkru-6vcg

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

1) Average monthly passengers traffic between 2006 and 2017
2) Passengers traffic by destination/origin regions
3) Overview on passengers traffic by domestic airlines
4) Passengers traffic travelled by Low Cost Carrier
5) Passengers traffic in airport terminals
6) Passengers traffic on 1 selected domestic carrier

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

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.

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.

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. For example, the dataset was wrongly identify `Wow Air` as `full service carrier` but the airline is identify as `low cost carrier` by ICAO.

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

Part 3: Exploratory Analysis

I. Overview on Passenger traffic in SFO

Overall passenger traffic between 2006 and 2017

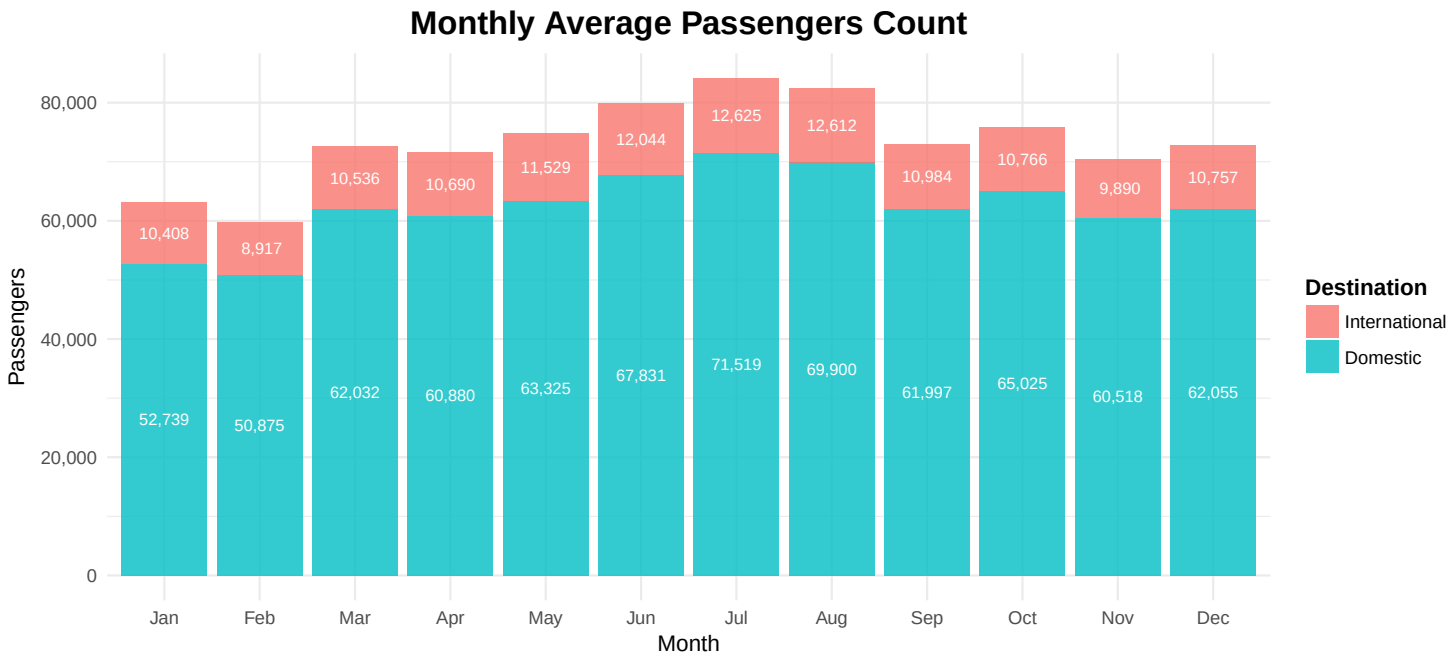
SFO is a busy airport and is a transpacific geteway in the West Coast of the United States, the passenger traffic is heavy. There are 55,823,712 passengers traveled in SFO 2017, compared to 33,332,970 in 2006. The average annual passenger growth in the period is 4.82%. It will take 15 years after 2006 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 passenger traffic varies by seasons. 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. About 76.96% of passengers via SFO were domestic travlers.

II. Destination

As the bar chart on the monthly average passengers count shows that 76.96% of the SFO travelers are domestic travelers. One maybe interested the destination of the remaining 23.04% travelers. The map below shows the total passengers count in North America between `2006` and `2017`.

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

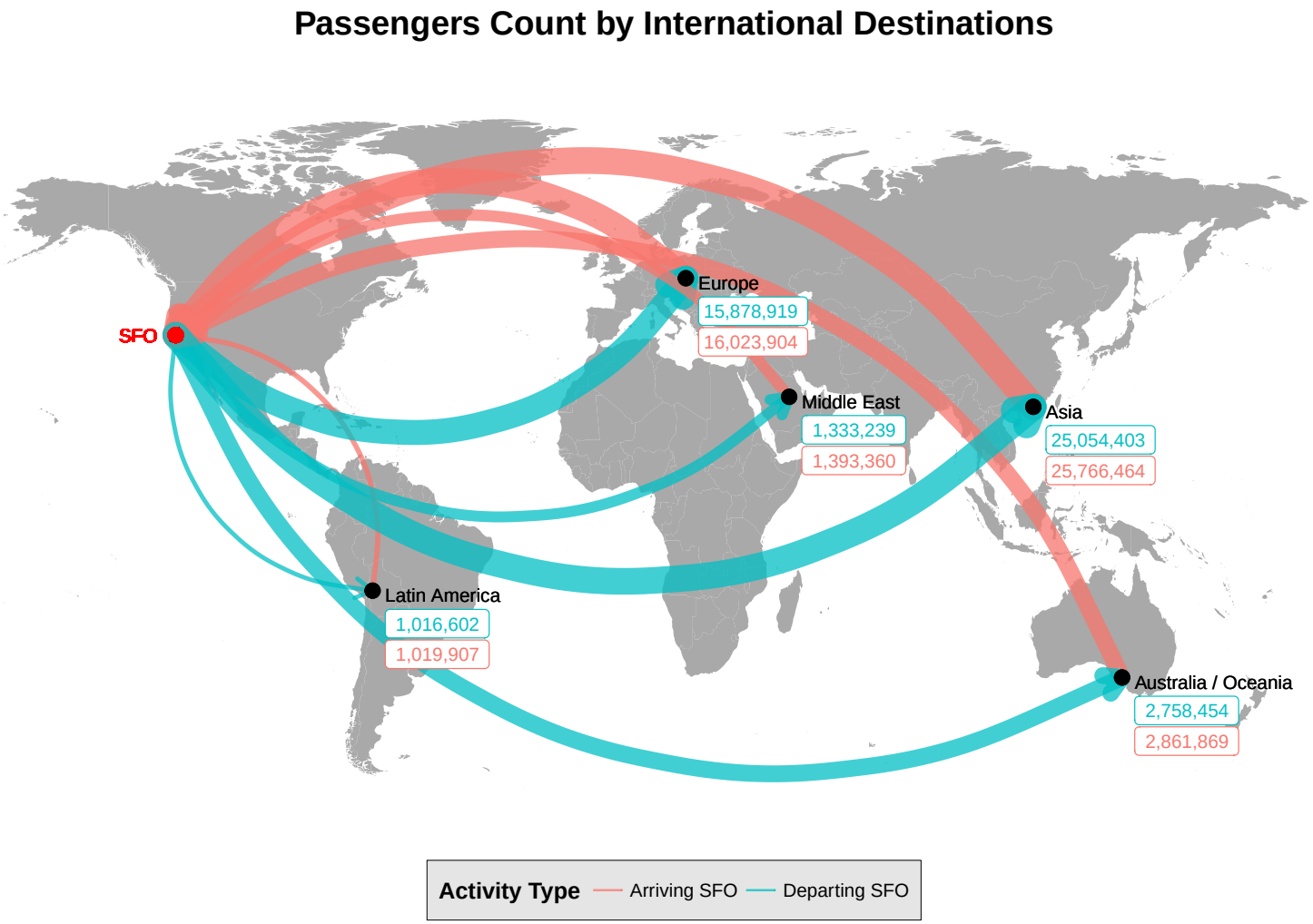
```
# Load the world map and cities
world <- map_data("world")
cities <- world.cities
# Retrieve the information of SF
sf <- world.cities %>%
  filter(name == "San Francisco" & country.etc == "USA")
# Derive the cities and corresponding regions
cities %<>%
  filter(
    (name == "Adelaide" & country.etc == "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_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_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_intl_enplaned_labels <- geom_label(data = cities_intl %>% filter(type == ("Enplaned")),
      aes(x = long, y = lat, label = format(pax, big.mark = ","), color = type),
      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 = type),
      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 != "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
```




```
region_pax <- data %>% group_by(region) %>% summarise(sumpax = sum(pax)) %>% arrange(desc(sumpax))
```

The map shows that Asia is the continent the most passengers flying from or to after the United States, with 51,040,819 passengers, 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`, 16,211,371 and 9,983,551 passengers, respectively, and the passenger traffic from or to `Latin America` is very little, with 2,036,509.

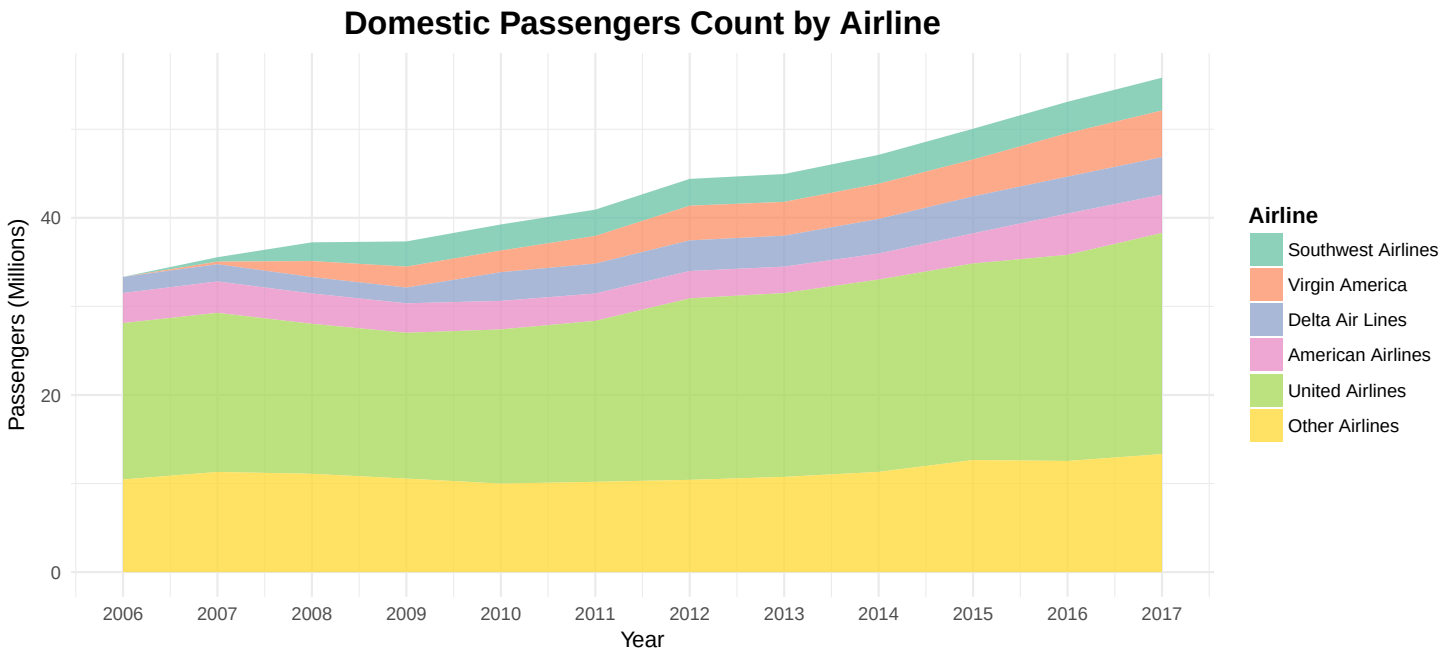
III. Domestic Carriers Overview

There are a lot of airlines serving travelers in SFO. Below is the stacked line chart of the domestic passengers traffic 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) %>%
  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)
```

```
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
```



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 domestic travelers traveled with those airlines between `2006` and `2017`.

III. Low Cost Carrier vs Full Service Carrier

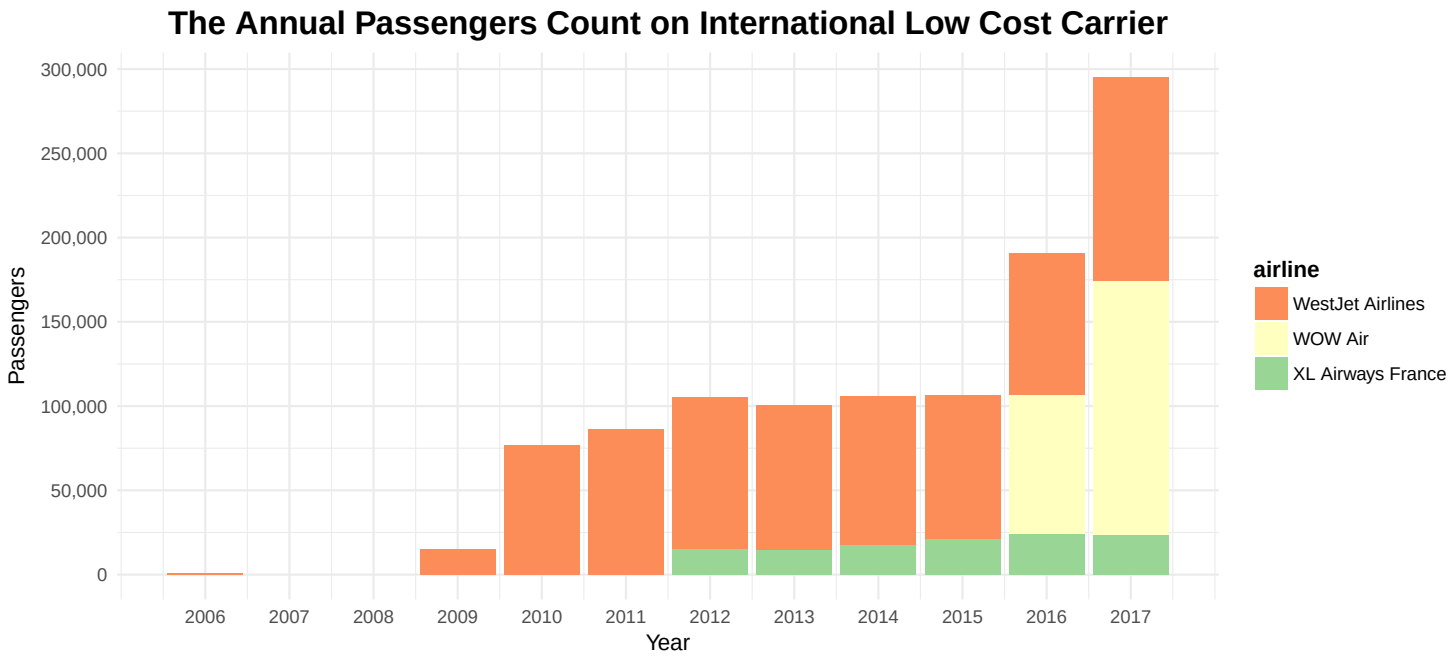
Due to the advance of avaition technology, the cost of air travel significantly decrease in the last 100 years. Low cost carriers provide low cost traveling by significant ticket price to attract passengers; to make up the loss of revenue, `low cost carriers` cut cost on meals, seating comfort, onboard entertainment, and seating priority. Alternatively, airlines that provide decent meals, onboard entertainment, and offer comfortable seats and seating priority are called `full Service Carrier`. For travelers travel without too much concern on those feature, traveling with low cost carrier is a good choice for those travelers. Southwest Airlines is one of the sucessful story and is a low cost carrier giant in the United Staets. 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.

In 2017, about 6,317,794 passengers traveled via SFO by low cost carriers, which makes up about 11.32% of the total passengers count.

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 500,926, 2,117,364, 2,847,732 passengers in 2007, 2008, and 2009, respectively. After 2009, the annual passenger growth 3.31%. In 2017, 3,704,789 passengers traveled with `Southwest Airlines`.

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 routes in SFO. In 2016, `Wow Air` annouced began scheduled service from Reykjavik, Iceland. 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 = 1)) +
  scale_y_continuous(name = "Passengers", breaks = seq(0, 500000, by = 50000), labels = comma)+
  ggtitle("The Annual Passengers Count on International Low Cost Carrier")+
  theme_minimal() +
  format_title +
  format_legend_title
```



There were not a lot of option on international traveling with 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.

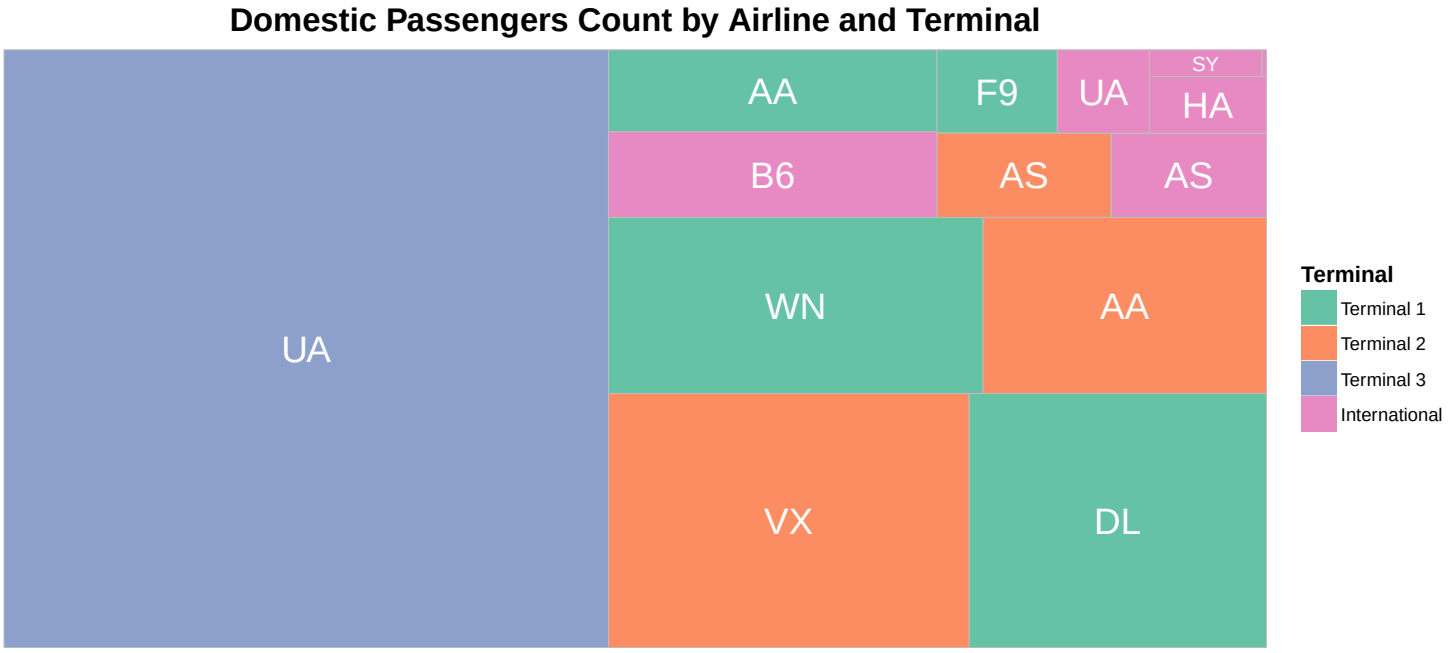
Although many travelers do not travel with low cost carriers, there are increasingly more passengers travel with low cost carriers. Therefore, we expect the portion of passengers traveled with low cost carrier increase.

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.

Below is the tree map on domestic passenger traffic 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 from/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`

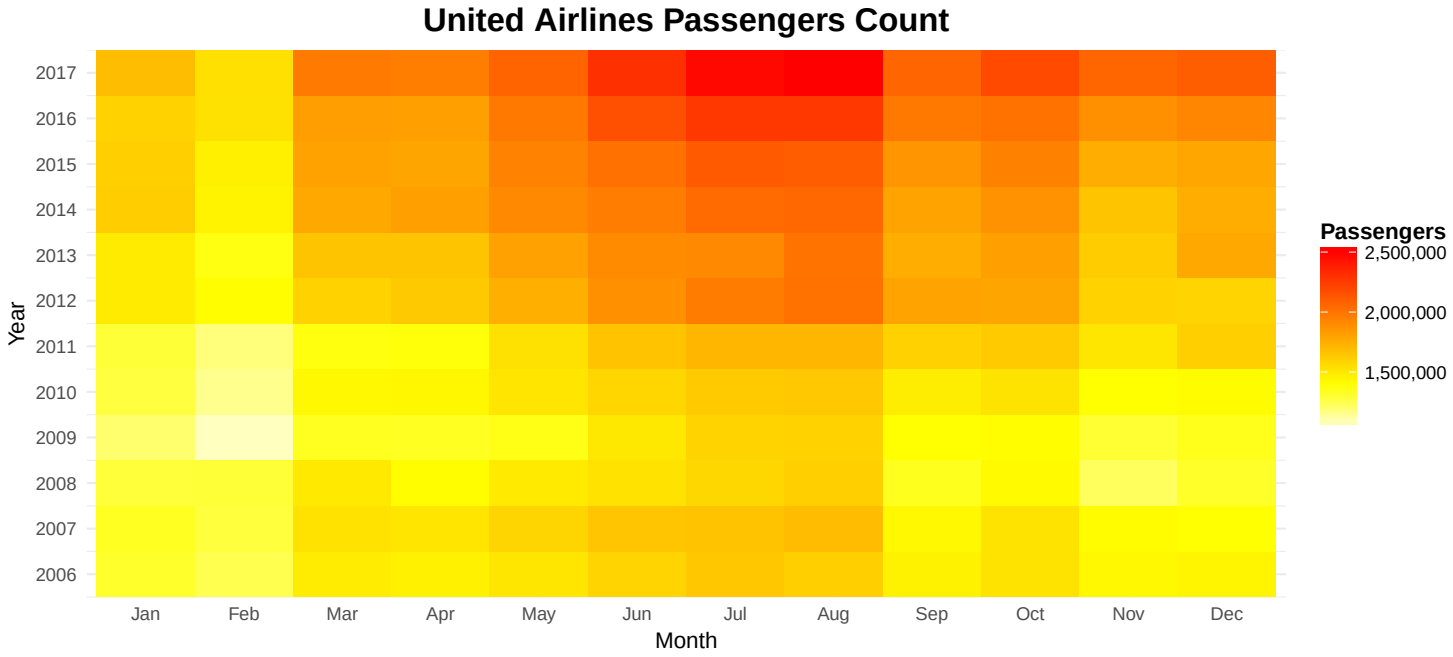
VI. Airlines

`United Airlines` is largest carrier in SFO in terms of passengers count in 2017, both 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") %>%
  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 = 1)) +
  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; the passenger traffic jumped after 2012 and continue to grow. 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 compare to the months of the same years, that means `United Airlines`' operation is very sensitive to seasonal effect.

Part 4: Conclusions and Future Analysis

In the report, we found out that the passenger traffic growth was continuous between 2006 and 2017 and we expect the growth is to be continue. In this report, we found out 76.96% of the passengers are domestic travelers and most of the domestic travelers travel with United Airlines. There are increasingly more travelers travel out of the United States by low cost carriers. And the airline industry seasonal effect is very sentative, especially the largest operator in SFO, `United Airlines`. After received the result from the analysis, we can see that the tax money was spent reasonable on air travel infrastructure in San Francisco as the passenger traffic was growing between the time period, realize the need for long-term investment in SFO to

increase the capcity for passenger travel to serve more travelers. At the same time, SFO adminstrative staffs can accommodate the short-term high demand on the airport service in the summer period, `June`, `July`, `August` after learning from the analysis.

However, the dataset we have for this report focus mainly on the passenger counts of region the flights fly to or from, there was no information on destination/origin cities and country, aircraft models. In this report; we are not able to trace any of destination/origin cities and country of the airlines. In the future, we would like to explore more on destination/origin cities and country and aircraft models each flight. Additionally, we are surprised on the passenger traffic of `United Airlines` continue to grow although the `United Airlines`' reputation worsened in recent years. We would like to observe the passenger traffic change on `United Airlines` in the next few years if the reputation effect occurs.