Flights Departing From Boston Logan Airport

This project explores various aspects of domestic flights departing from Boston Logan Airport in 2015. It focuses on answering multi-faceted questions to better understand the reasons behind delayed flights.

Xinman Liu, Melissa Putur, Jiao Sun, Adil Wahab, Rui Xu

```
# loading libraries
library(dplyr)
library(ggplot2)
library(ggpubr)
library(tidyverse)
library(plotly)
library(httpuv)
library(lubridate)
library(RColorBrewer)
library(scales)
library(stringr)
library(readr)
library(gridExtra)
library(maps)

# modifying chart size
options(repr.plot.width=5, repr.plot.height=3)

#upload excel file
flights0 <- read_csv("flight_data/bos_flights_only - bq-results-20190809-141638-3zfmbbf1
em6h.csv")
airline <- read_csv("flight_data/airlines.csv")
airports <- read_csv("flight_data/airports.csv")

flights <- flights0 %>%
  left_join(airports, by = c("DESTINATION_AIRPORT" = "IATA_CODE")) %>%
  left_join(airline, by = c("AIRLINE" = "IATA_CODE"))
```

Project Overview: Boston Logan airport is the largest airport in New England with over 100,000 domestic flights annually. While the airport does offer exceptional ocean views and dining options, ie Dunkin Donuts, delayed and canceled flights can create financial and emotional burdens for passengers like increased trip costs and missed work or vacation days. To maximize our trips and avoid flight delays and/or cancellations out of Boston Logan we will explore which airlines, departure times, and destinations experience the fewest disruptions. We will also explore if a flight is delayed what factors contribute to whether or not the flight will ultimately arrive on time or late at the destination airport.

Data: To explore our question above we downloaded 3 datasets related to domestic flights in the United States in 2015. The full datasets can be found here: https://www.kaggle.com/usdot/flight-delays/version/1#flights.csv (https://www.kaggle.com/usdot/flight-delays/version/1#flights.csv)

Dataset 1 - Flights: The flights dataset contains a row for every domestic flight in the United States in 2015. The original file contained 5.8 million rows and was 1.1 GB. After uploading the file into Big Query, we ran a query to capture only flights that departed out of Boston to increase query processing time and reduce the amount of money we needed to pay per query. Our updated dataset contained 107847 rows and was 20 MB. We used the following columns for our analysis. Note: There was no data available for flights out of Boston in October in this dataset.

YEAR - year of the flight; dbl class; only includes data from 2015

MONTH - month of the flight; dbl class

DAY - day of the flight; a dbl class

AIRLINE - code of the airline; chr class

FLIGHT_NUMBER - flight number; dbl class

ORIGIN_AIRPORT - the departing airport; chr class; identical to the IATA_CODE column from the airport table

DISTANCE - the distance in miles between the origin airport and destination airport; dbl class

DESTINATION_AIRPORT - the destination airport; chr class; identical to the IATA_CODE column from the airport table

DEPARTURE_DELAY - the total departure delay time in minutes; dbl class

ARRIVAL_DELAY - the total arrival delay time in minutes; dbl class

AIR_SYSTEM_DELAY - the departure delay in minutes caused by an air system issue (non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc.); dbl class

SECURITY_DELAY - the delay in minutes caused by a security issue (evacuation of a terminal/concourse, re-boarding of aircraft because of a security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas); dbl class

AIRLINE_DELAY - the delay in minutes caused by an airline issue (technological, maintenance and fueling problems with the airline); dbl class

LATE_AIRCRAFT_DELAY - the delay in minutes caused by the late arrival of the aircraft to be used for the flight from a previous flight and congestion in air traffic; dbl class

WEATHER_DELAY - the delay in minutes caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival; dbl class

Dataset 2 - Airlines: The airline dataset contains two columns, the name of the 14 airlines with domestic flights within the U.S. in 2015 and their airline codes and 15 rows. Because the flights dataset did not contain full airline names, we had to join the airlines table to the flights table by airline code to reference airline names. The following columns were availble in the airlines dataset:

Airline_name - the full name of the airline; a nullable string

IATA_code - the associated code with the airline; a nullable string

Dataset 3 - Airports: The airports table contains seven columns and 322 rows. Because the flights dataset did not contain the full airport name or destination city, we used the IATA_CODE column to join the flights table and capture this information. The following columns were available in the flights dataset:

IATA_code - the associated code with the airport; a nullable string

Airport - full name of the airport; a nullable string

City - name of the city the airport is in; a nullable string

State - name of the state the airport is in; a nullable string

Country - name of the country the airport is in; a nullable string

Below is a sample of the raw data of the files:

```
head(flights)
```

Y…	M…	…	DAY_OF_W…	AIRLI…	FLIGHT_NUM…	TAIL_NUM…	ORIGIN_AIRPO…	DESTINATIO
<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<chr>	<chr>	<chr>
2015	8	31	1	B6	111	N629JB	BOS	ORD
2015	9	21	1	AA	147	N3MEAA	BOS	LAX
2015	2	9	1	DL	1901	N678DL	BOS	ATL
2015	2	9	1	WN	1061	N631SW	BOS	BWI
2015	2	16	1	DL	2549	N375DA	BOS	MSP
2015	7	13	1	AA	1488	N3MHAA	BOS	ORD

6 rows | 1-9 of 38 columns

```
head(airline)
```

IATA_CODE	AIRLINE
<chr>	<chr>
UA	United Air Lines Inc.
AA	American Airlines Inc.
US	US Airways Inc.
F9	Frontier Airlines Inc.
B6	JetBlue Airways
OO	Skywest Airlines Inc.

6 rows

```
head(airports)
```

IATA_C... <chr>	AIRPORT <chr>	CITY <chr>	ST... <chr>	COU... <chr>	LATIT... <dbl>	LO
ABE	Lehigh Valley International Airport	Allentown	PA	USA	40.65236	-
ABI	Abilene Regional Airport	Abilene	TX	USA	32.41132	-
ABQ	Albuquerque International Sunport	Albuquerque	NM	USA	35.04022	-1
ABR	Aberdeen Regional Airport	Aberdeen	SD	USA	45.44906	-
ABY	Southwest Georgia Regional Airport	Albany	GA	USA	31.53552	-
ACK	Nantucket Memorial Airport	Nantucket	MA	USA	41.25305	-

6 rows

Data Exploration

Question: How many airports can you fly to from Boston and what were the most popular airports to fly to?

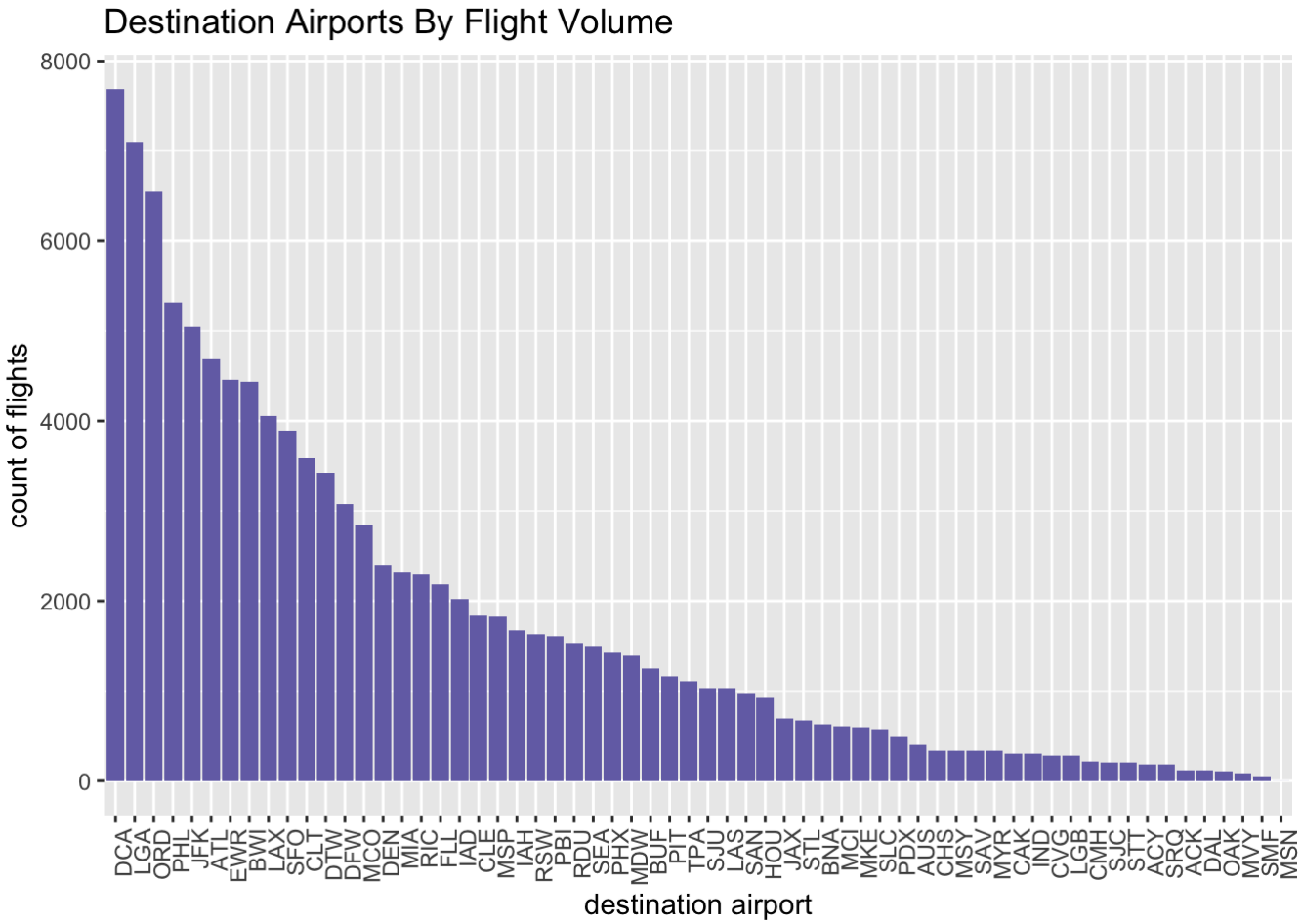
In 2015, Boston Logan had flights to 62 unique airports in the United States. Ronald Reagan Airport in Arlington, LaGuardia in New York, and Chicago O'Hare were the most popular destinations in 2015 with each airline receiving over 6,500 flights from Boston.

```
flights %>%
  count(AIRPORT, CITY, sort = TRUE)
```

AIRPORT <chr>	CITY <chr>
Ronald Reagan Washington National Airport	Arlington
LaGuardia Airport (Marine Air Terminal)	New York
Chicago O'Hare International Airport	Chicago
Philadelphia International Airport	Philadelphi
John F. Kennedy International Airport (New York International Airport)	New York
Hartsfield-Jackson Atlanta International Airport	Atlanta
Newark Liberty International Airport	Newark
Baltimore-Washington International Airport	Baltimore
Los Angeles International Airport	Los Angele
San Francisco International Airport	San Franci

1-10 of 62 rows Previous **1** 2 3 4 5 6 7 Next

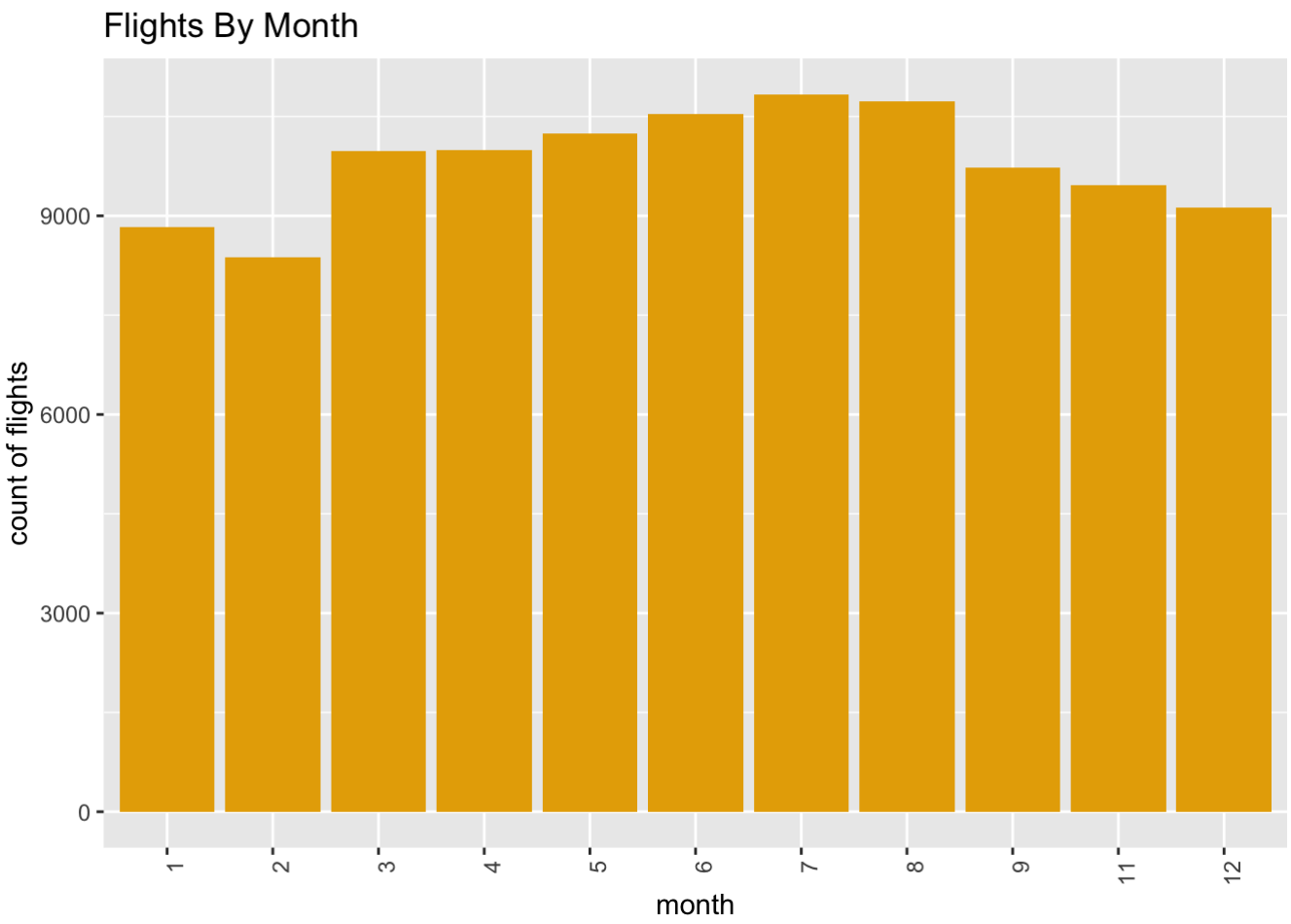
```
flights %>% count(DESTINATION_AIRPORT, sort = TRUE) %>%
  arrange(desc(n)) %>%
  ggplot() +
  geom_col(aes(x = reorder(DESTINATION_AIRPORT,-n), y = n),fill = "#7570B3") +
  theme(axis.text.x=element_text(angle=90, hjust=1)) +
  labs(title = "Destination Airports By Flight Volume", x = "destination airport", y =
"count of flights")
```



Question: How many flights are out of Boston each month?

Throughout the year there were between 8,300 and 10,800 flights out of Boston each month. January had the least amount of flights, with 8,837 while July had the most flights with 10,837. In general, the winter months were the least busy while the summer months were the busiest.

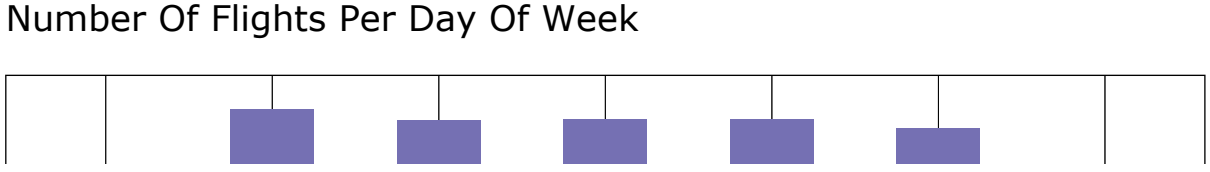
```
flights %>% count(MONTH, sort = TRUE) %>%
  arrange(desc(n)) %>%
  ggplot() +
  geom_col(aes(x = factor(MONTH), y = n),fill = "#E6AB02") +
  theme(axis.text.x=element_text(angle=90, hjust=1)) +
  labs(title = "Flights By Month", x = "month", y = "count of flights")
```

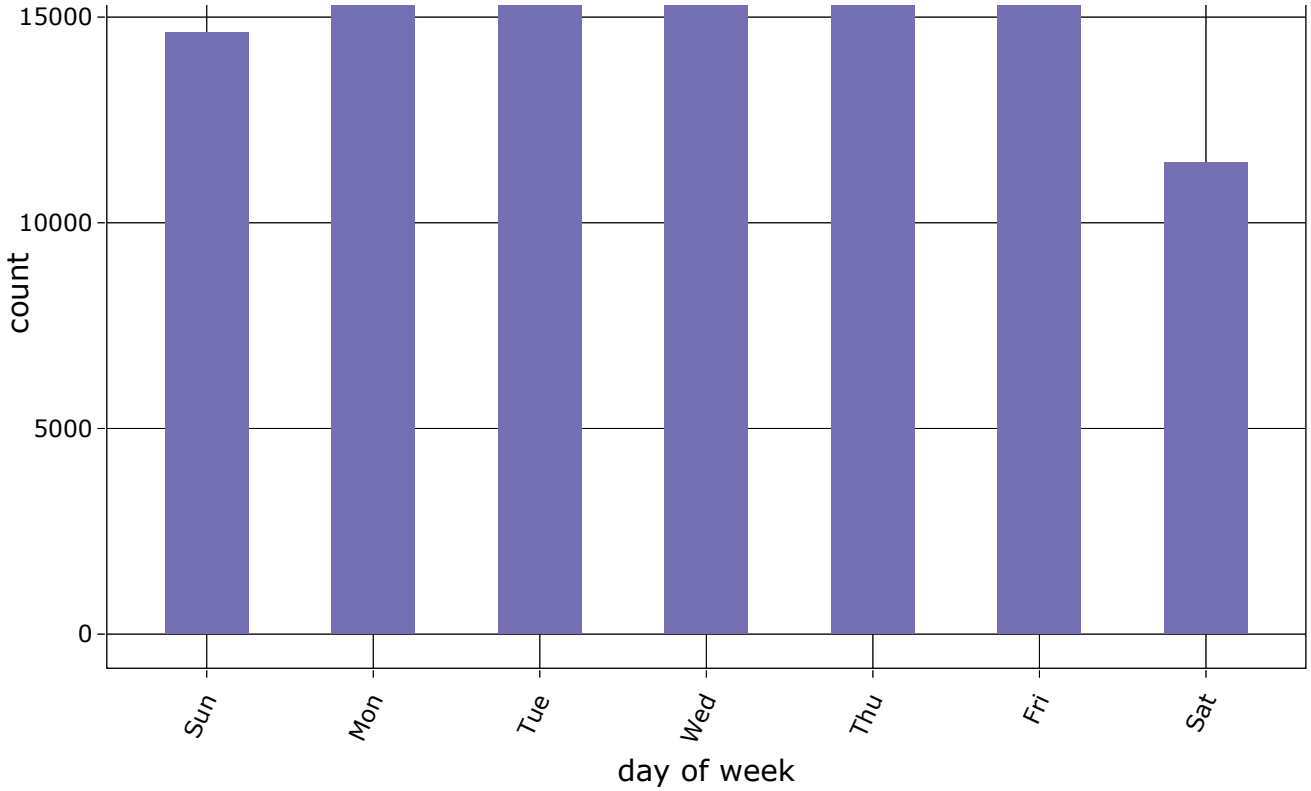


Question: How many flights departed from Boston on each day of the week?

Throughout the year, there were more flights out of Boston on weekdays compared to weekends, which is due to the notion that business trips are during the week and people who want to spend their weekend out of the city will usually fly Friday after work/school. Most flights out of Boston are on Mondays and the least are on Saturdays. Tuesdays, Wednesdays, and Thursdays have about the same number of flights departing Boston.

```
ggplotly(flights %>%
  mutate(date = make_datetime(YEAR, MONTH, DAY),
         weekday = wday(date, label = TRUE)) %>%
  group_by(weekday) %>%
  summarize(count = n(),  # counting the number of flights (per weekday)
            delay = mean(ARRIVAL_DELAY, na.rm = TRUE)  # average delay (per weekday)
           ) %>%
  ggplot(aes(weekday, count)) +
    geom_bar(stat="identity", width = 0.5, fill = "#7570B3") +
    labs(title="Number Of Flights Per Day Of Week",
         x = "day of week",
         y = "count") +
    theme_linedraw() +
    theme(axis.text.x = element_text(angle=65, vjust=0.6)))
```





Question: How many flights departed late out of Boston Logan in 2015?

In 2015 39,093, or 37% of flights departed late out of Boston.

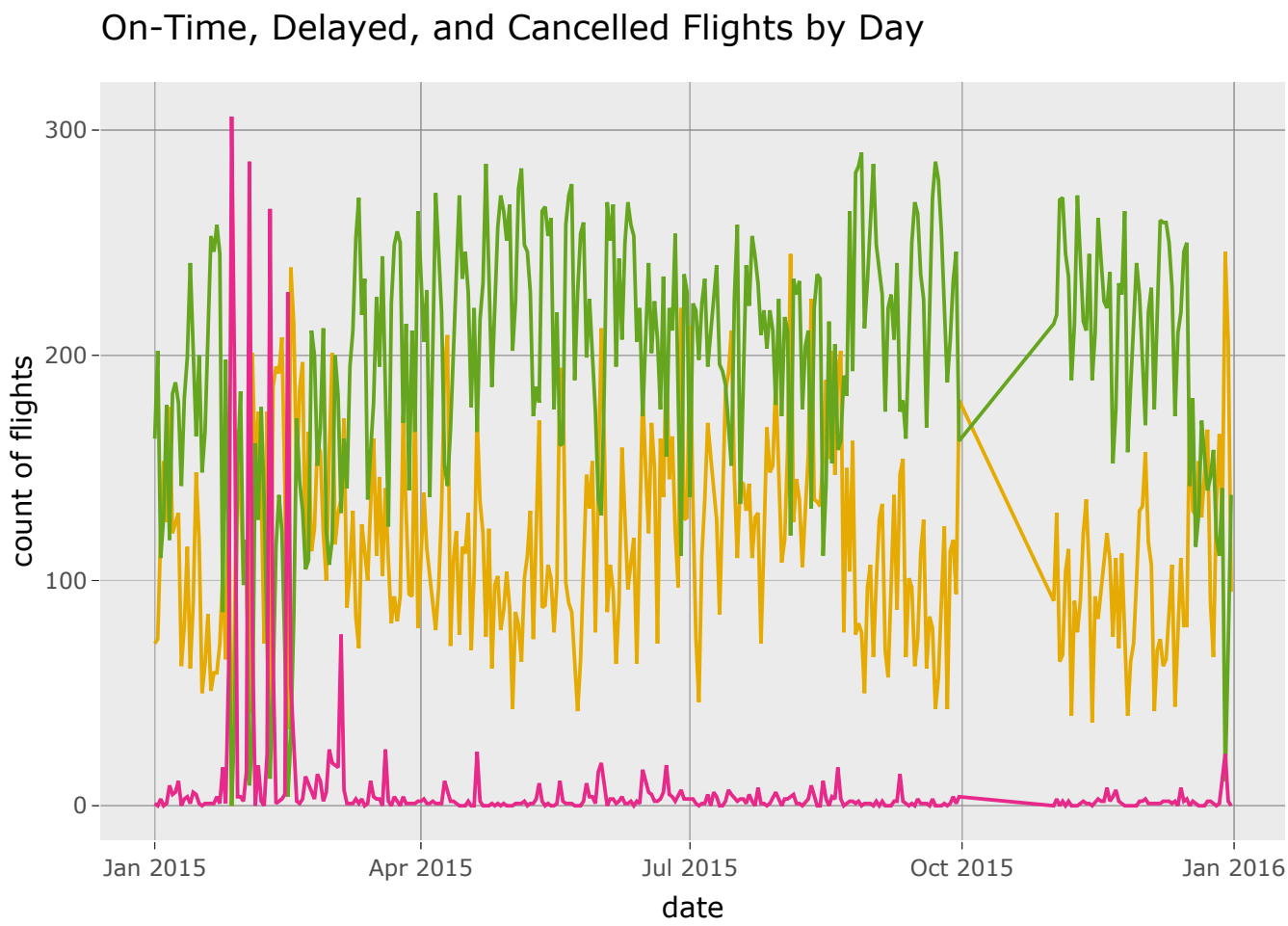
```
delay_count <- flights %>%
 transmute(DESTINATION_AIRPORT, DEPARTURE_DELAY, DELAYED = DEPARTURE_DELAY, NOT_DELAYED
= DEPARTURE_DELAY)
delayed_istrue <- which(delay_count$DEPARTURE_DELAY > 0)
delayed_isnottrue <- which(delay_count$DEPARTURE_DELAY <= 0)
delay_count$DELAYED[delayed_istrue] <- 1
delay_count$DELAYED[delayed_isnottrue] <- 0
delay_count$NOT_DELAYED[delayed_istrue] <- 0
delay_count$NOT_DELAYED[delayed_isnottrue] <- 1
delay_count %>%
 select(DESTINATION_AIRPORT, DELAYED, NOT_DELAYED) %>%
 filter(!is.na(DELAYED), !is.na(NOT_DELAYED)) %>%
 summarize(total_delays = sum(DELAYED), total_ontimes = sum(NOT_DELAYED)) %>%
 mutate(percent_delayed = total_delays/(total_delays + total_ontimes))
```

total_delays	total_ontimes	percent_delayed
<dbl>	<dbl>	<dbl>
39093	66183	0.3713382

1 row

In general, in 2015 there were more on-time flights than delayed flights on any given day out of Boston Logan. Typically canceled flights made up a very small portion of total flights, with the exception being 1/27, 2/2, 2/9, and 2/15 where we can see the majority of flights were canceled.

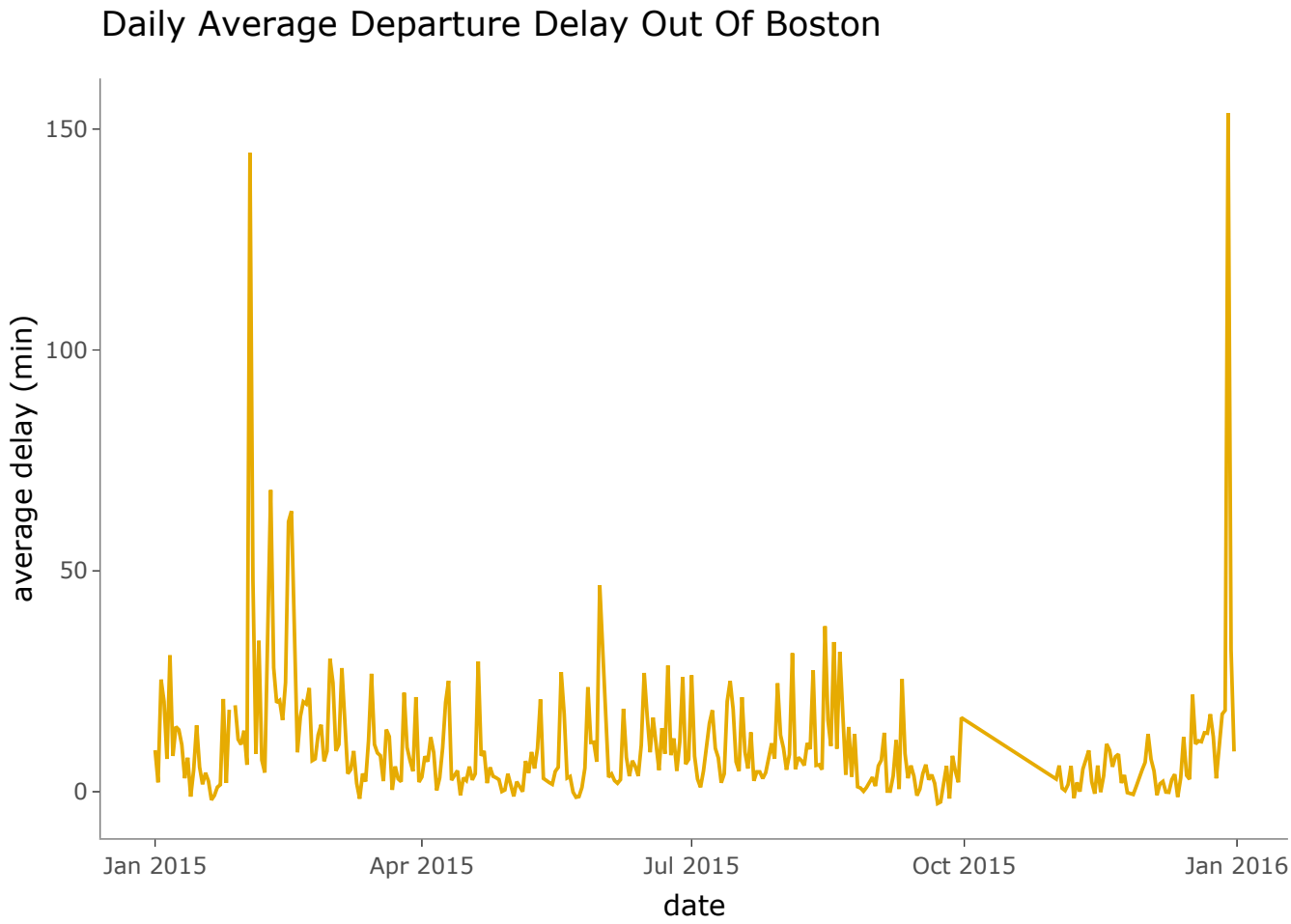
```
flights %>%
 select(YEAR,MONTH,DAY,DEPARTURE_DELAY,CANCELLED) %>%
 mutate(date = make_date(YEAR,MONTH,DAY), delay10 = ifelse(DEPARTURE_DELAY > 0, 1, 0), n
otdelay10 = ifelse(DEPARTURE_DELAY <= 0, 1, 0)) %>%
 group_by(date) %>%
 summarize(delayed = sum(delay10, na.rm = TRUE), notdelayed = sum(notdelay10, na.rm = TR
UE), cancelled = sum(CANCELLED, na.rm = TRUE)) -> flightsByDay
f <- ggplot(data = flightsByDay) +
 geom_line(aes(x = date, y = delayed), color = "#E6AB02") +
 geom_line(aes(x = date, y = notdelayed), color = "#66A61E") +
 geom_line(aes(x = date, y = cancelled), color = "#E7298A") +
 labs(title = "On-Time, Delayed, and Cancelled Flights by Day", x = "date", y = "count o
f flights")
ggplotly(f)
```



Question: What is the trend line for the daily average departure delay throughout the year?

The month of February in 2015 had the highest daily average delay in minutes with February 2nd standing out for having average departure delays of 144 minutes. February of 2015 was recorded as the snowiest month on record in Boston since 1891 with almost 65 inches of snow which likely led to an increase in flight delays. Another day that stood out was December 29th, for having the highest average delay of 154 minutes, this was likely due to the mix of a high volume of flights returning from Christmas, departing to celebrate New Years, and flight delays due to snow.

```
flight_count <- flights %>%
  mutate(date = make_datetime(YEAR, MONTH, DAY)) %>%
  group_by(date) %>%
  summarize(count = n(),  # counting the number of flights/day
            delay = mean(DEPARTURE_DELAY, na.rm = TRUE)  # average delay/day
           )

p <- ggplot(flight_count, aes(date, delay)) +
  geom_line(color="#E6AB02") +
  labs(title = "Daily Average Departure Delay Out Of Boston", y = "average delay (min)")
+
  theme_classic()

ggplotly(p)
```



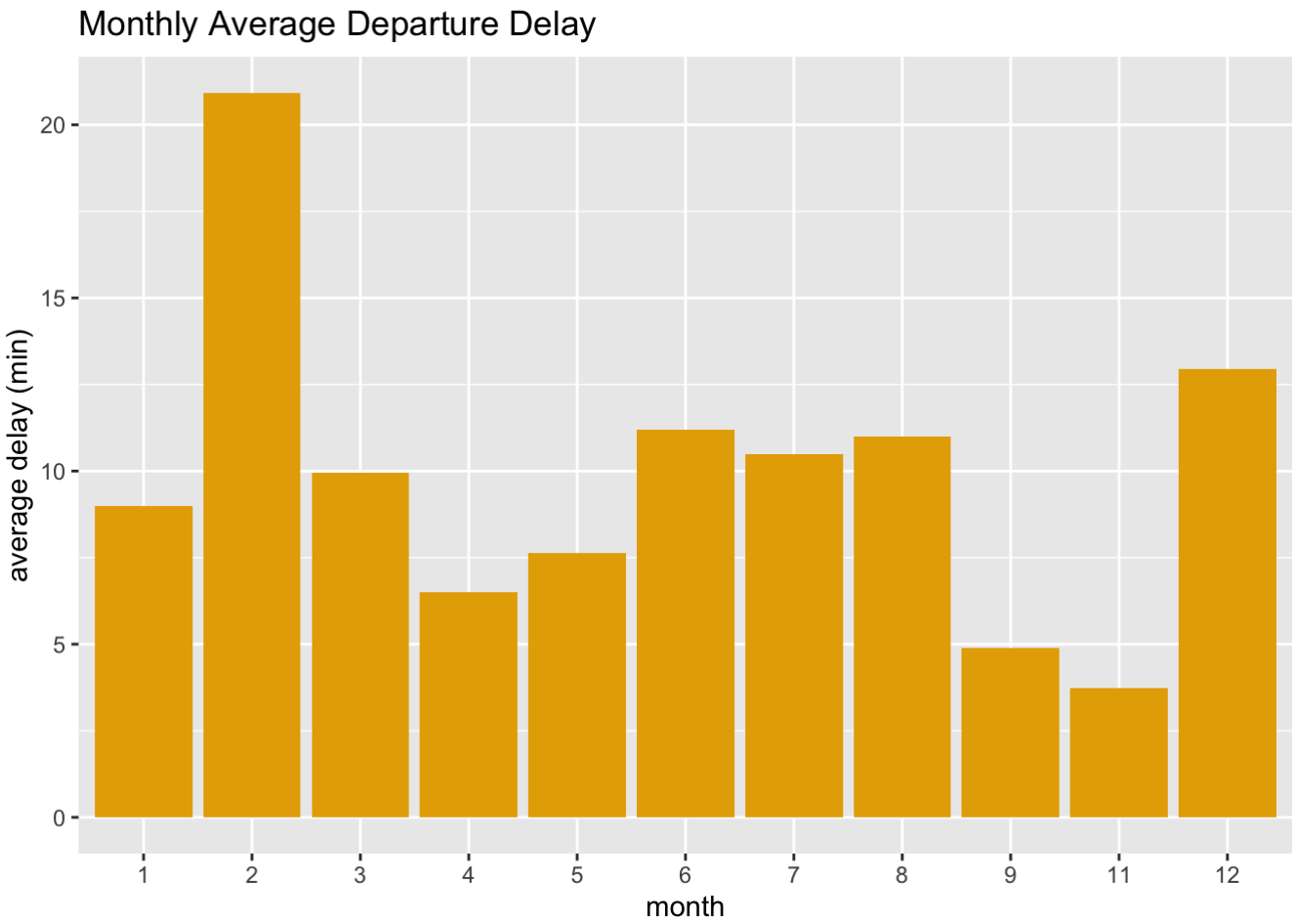
```
#by month
bosFlights <- read_csv("flight_data/bos_flights_only - bq-results-20190809-141638-3zfmbb
f1em6h.csv")
bosFlights %>%
  mutate(monthFac = as.factor(MONTH)) %>%
  group_by(monthFac) %>%
  summarize(count = n(),
            delay = mean(DEPARTURE_DELAY, na.rm = TRUE)) -> delayByMonth

delayByMonth
```

monthFac <fctr>	count <int>	delay <dbl>
1	8837	9.001463
2	8380	20.924010
3	9971	9.962028
4	10000	6.512877
5	10241	7.637633
6	10544	11.186064
7	10837	10.501626
8	10727	10.991732
9	9726	4.899845
11	9464	3.722776

1-10 of 11 rows

Previous **1** 2 Next

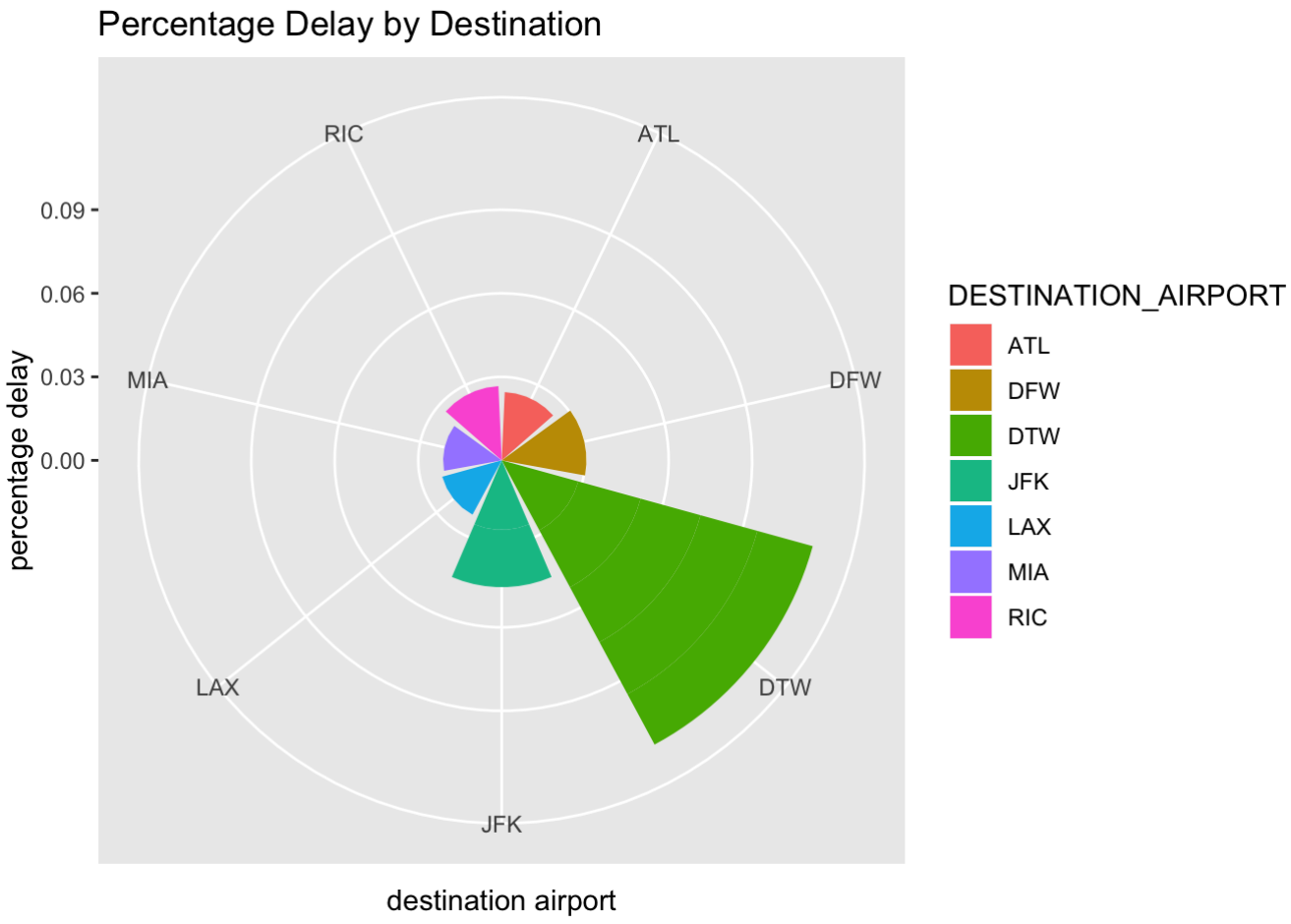
```
ggplot(delayByMonth, aes(x = monthFac, y = delay)) +
  geom_col(fill="#E6AB02") +
  labs(title = "Monthly Average Departure Delay", y = "average delay (min)", x = "month"
)
```



Question: Which destination has the highest proportion of departure delay out of all the destinations from Boston?

The chart below shows the top 7 destinations with the highest percentage of departure delays out of Boston Logan. Detroit Metro Airport appears to have a significantly higher percent of delay in comparison to all the other destinations out of Boston.

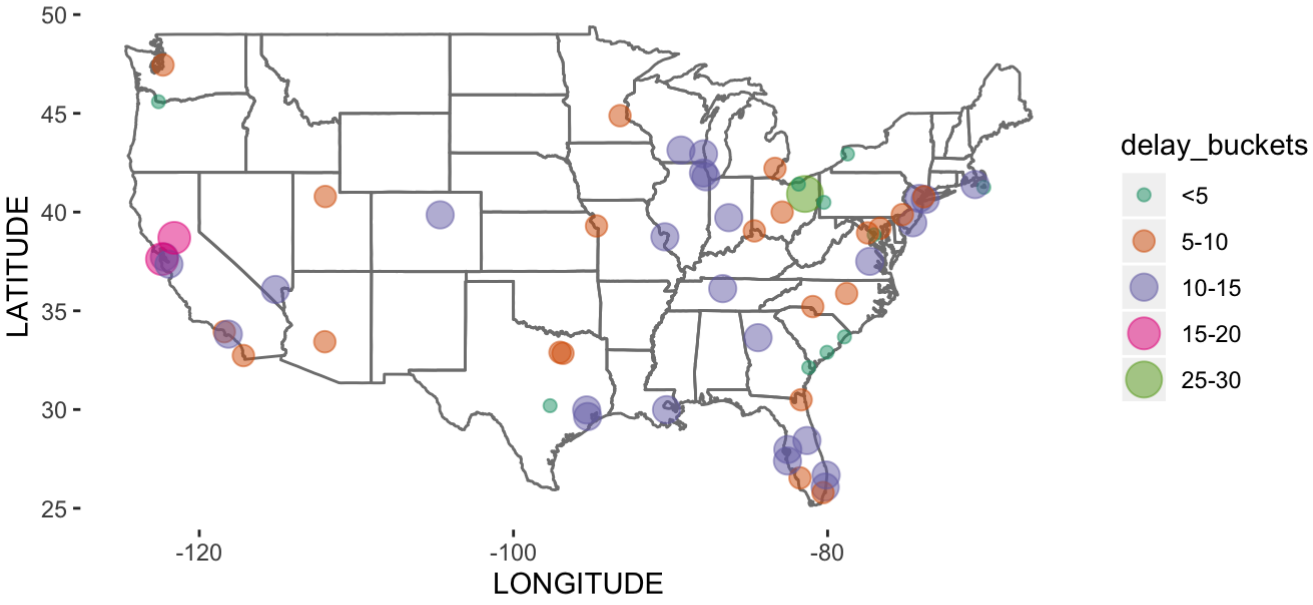
```
z <- flights %>%
 group_by(DESTINATION_AIRPORT) %>%
 summarise(avgDelay=mean(DEPARTURE_DELAY,na.rm=T)) %>%
 filter(avgDelay>10) %>%
 select(avgDelay,DESTINATION_AIRPORT) %>%
 arrange(desc(avgDelay)) %>%
 ggplot(aes(DESTINATION_AIRPORT,avgDelay))+
 geom_bar(stat="identity",fill="#E6AB02")+
 geom_hline(yintercept=20,color="#E7298A",size=1)+
 labs(x="destination airport",
        y="average delay",
        title="Average Delay by Destination") +
     theme(axis.text.x=element_text(angle=90, hjust=1))

flights %>%
 filter(!is.na(DEPARTURE_DELAY),DEPARTURE_DELAY>0) %>%
 mutate(percDealy=DEPARTURE_DELAY/n(),na.rm=T) %>%
 filter(percDealy>0.02) %>%
        select(percDealy,DESTINATION_AIRPORT) %>%
          arrange(desc(percDealy)) %>%
ggplot(aes(DESTINATION_AIRPORT,percDealy))+
geom_bar(stat = "identity",aes(fill=DESTINATION_AIRPORT))+
coord_polar()+
labs(x="destination airport",
      y="percentage delay",
      title="Percentage Delay by Destination")
```



Question: What cities experienced the longest average arrival delays?

Akron experienced the longest arrival delays of any US city with an average arrival delay of 26 minutes followed by Sacramento with a delay of 19 minutes and San Francisco at 15 minutes.

```
flights %>%
  group_by(DESTINATION_AIRPORT) %>%
  summarise(avg_dest_delay = mean(DEPARTURE_DELAY, na.rm = TRUE)) %>%
  inner_join(airports, by = c("DESTINATION_AIRPORT" = "IATA_CODE")) %>%
  filter(LONGITUDE < 0 & between(LATITUDE, 25, 50))  -> delayByState # limiting the map
 to US mainland boarders    for clarity  # only destinations with delays more than one h
our

delay_buckets <- cut(x = delayByState$avg_dest_delay, breaks = c(0,5,10,15,20,25,30))
levels(delay_buckets) <- c("<5", "5-10", "10-15", "15-20", "20-25", "25-30")

delayByState$delay_buckets <- delay_buckets

  ggplot(data = delayByState, aes(y = LATITUDE, x = LONGITUDE, color = delay_buckets, si
ze = delay_buckets)) +
  borders("state") +
  geom_point(alpha = .5) +
  coord_quickmap() +
  scale_color_brewer(palette = "Dark2") +
  theme(panel.background = element_blank())
```



```
delayByState %>%
  select(CITY, avg_dest_delay) %>%
  arrange(desc(avg_dest_delay)) %>%
  head(5)
```

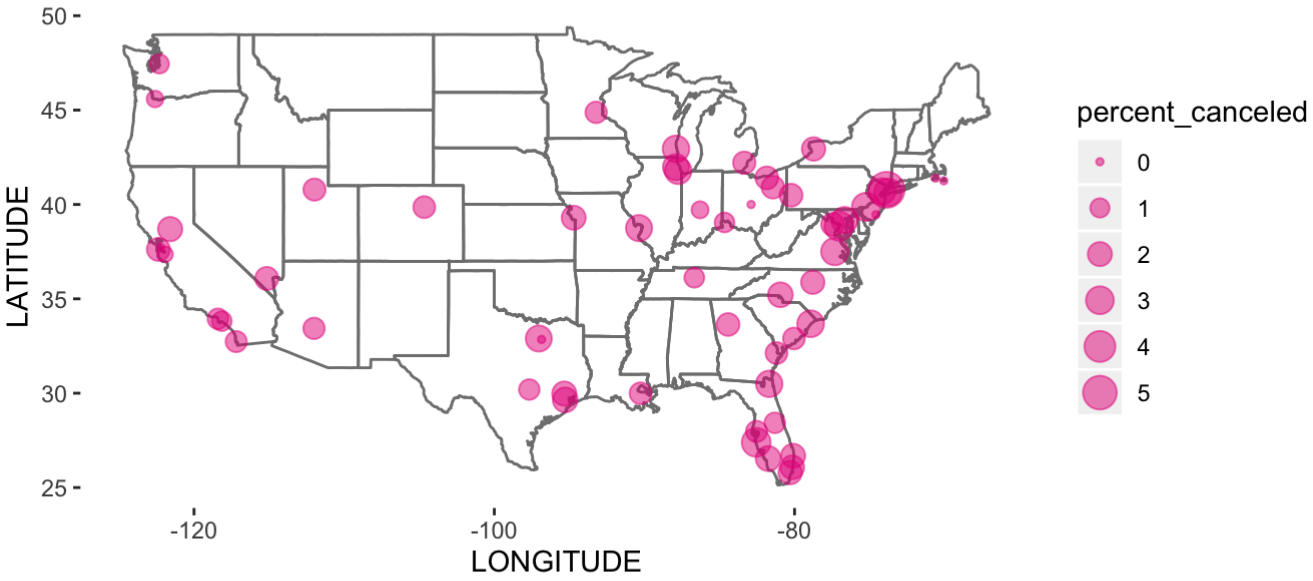
CITY <chr>	**avg_dest_delay** <dbl>
Akron	26.83333
Sacramento	19.31250
San Francisco	15.33316
Marthas Vineyard	14.78571
New Orleans	14.74164

5 rows

Question Which airport had the most flights canceled?

LGA, LaGuardia, had the most flights canceled to it with almost 6% of flights departing for LGA ultimately being canceled.

```
flights %>%
  group_by(DESTINATION_AIRPORT) %>%
  summarise(percent_canceled = sum(is.na(DEPARTURE_DELAY))/n() * 100, n = n()) %>%
  filter(n >= 5) %>%
  inner_join(airports, by = c("DESTINATION_AIRPORT" = "IATA_CODE")) %>%
  filter(LONGITUDE < 0 & between(LATITUDE, 25, 50)) %>%
  ggplot(aes(y = LATITUDE, x = LONGITUDE, size = percent_canceled)) +
  borders("state") +
  geom_point(alpha = .5, color = "#E7298A") +
  coord_quickmap() +
  theme(panel.background = element_blank())
```




```
flights %>%
  group_by(DESTINATION_AIRPORT) %>%
  summarise(percent_canceled = sum(is.na(DEPARTURE_DELAY))/n() * 100, n = n()) %>%
  filter(n >= 5) %>%
  select(DESTINATION_AIRPORT,percent_canceled) %>%
  arrange(desc(percent_canceled))
```

DESTINATION_AIRPORT <chr>	**percent_canceled** <dbl>
LGA	5.8906426
DCA	3.3823338
SRQ	3.3149171
PHL	3.1055901
JFK	2.9341792
RIC	2.8733130
MYR	2.7027027
MKE	2.6622296
MDW	2.6599569

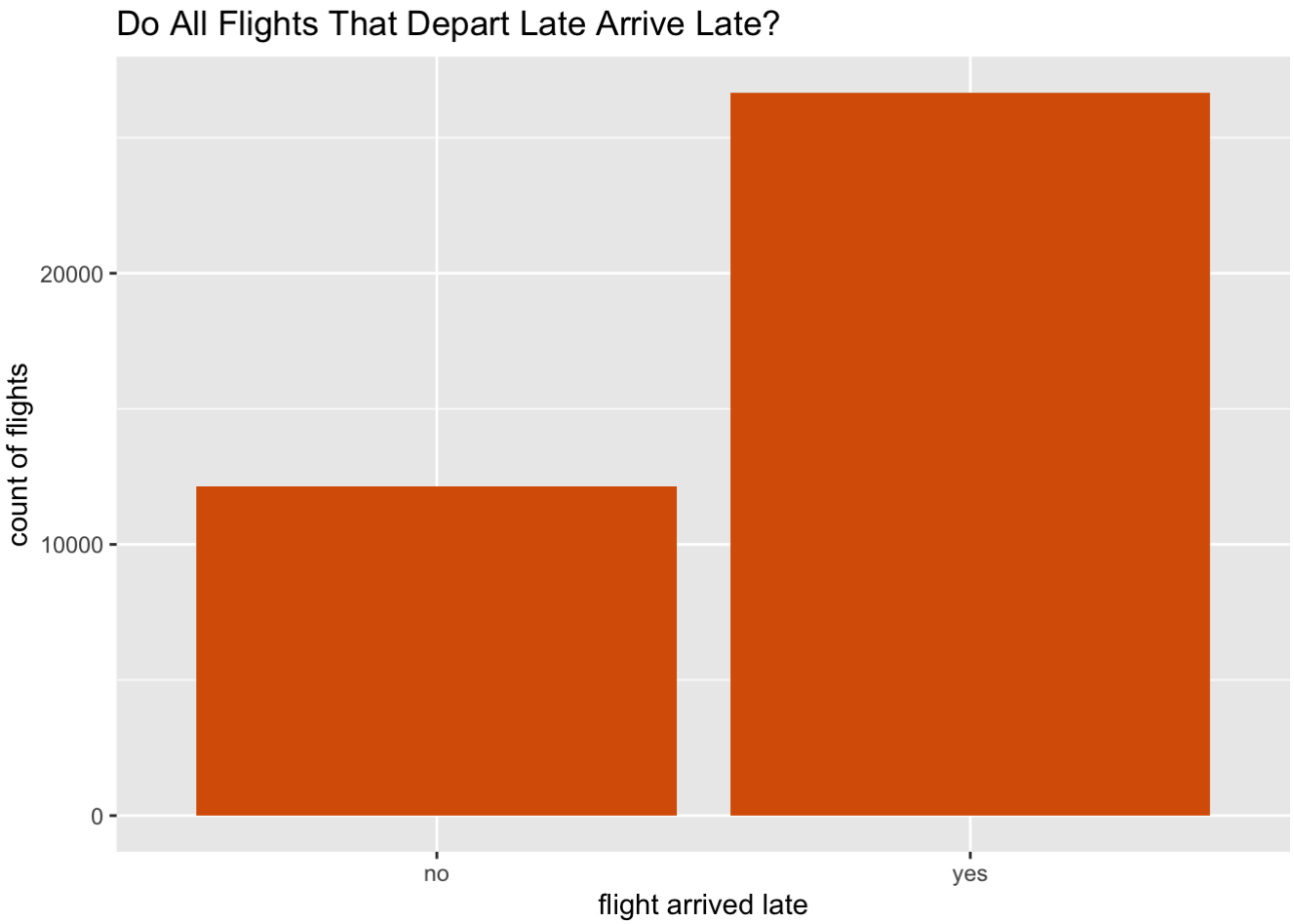
DESTINATION_AIRPORT <chr>	percent_canceled <dbl>
JAX	2.6086957

1-10 of 61 rows Previous **1** 2 3 4 5 6 7 Next

Question: Does departing late guarantee that your flight will land late?

Even if your flight departs late it is possible that your flight will still land on time at the destination airport. About 12,000 flights that departed late out of Boston in 2015 landed on time.

```
bos_flights_delay_expand <- flights
yes <- which(bos_flights_delay_expand$DEPARTURE_DELAY > 0)
bos_flights_delay_expand$DEPARTURE_DELAY[yes] <- "yes"
no <- which(bos_flights_delay_expand$DEPARTURE_DELAY <= 0)
bos_flights_delay_expand$DEPARTURE_DELAY[no] <- "no"

bos_flights_delay_expand <- select(bos_flights_delay_expand,DEPARTURE_DELAY, ARRIVAL_DEL
AY, DISTANCE, AIR_TIME, DESTINATION_AIRPORT)

yes <- which(bos_flights_delay_expand$ARRIVAL_DELAY > 0)
bos_flights_delay_expand$ARRIVAL_DELAY[yes] <- "yes"
no <- which(bos_flights_delay_expand$ARRIVAL_DELAY <= 0)
bos_flights_delay_expand$ARRIVAL_DELAY[no] <- "no"

delayed_flights_only <- filter(bos_flights_delay_expand, DEPARTURE_DELAY == "yes")

#chart one - how many flights that depart late also arrive late
ggplot(data = na.omit(delayed_flights_only), aes(x = ARRIVAL_DELAY)) +
  geom_bar(fill = "#D95F02") +
  labs(title = "Do All Flights That Depart Late Arrive Late?", x = "flight arrived late"
, y = "count of flights")
```



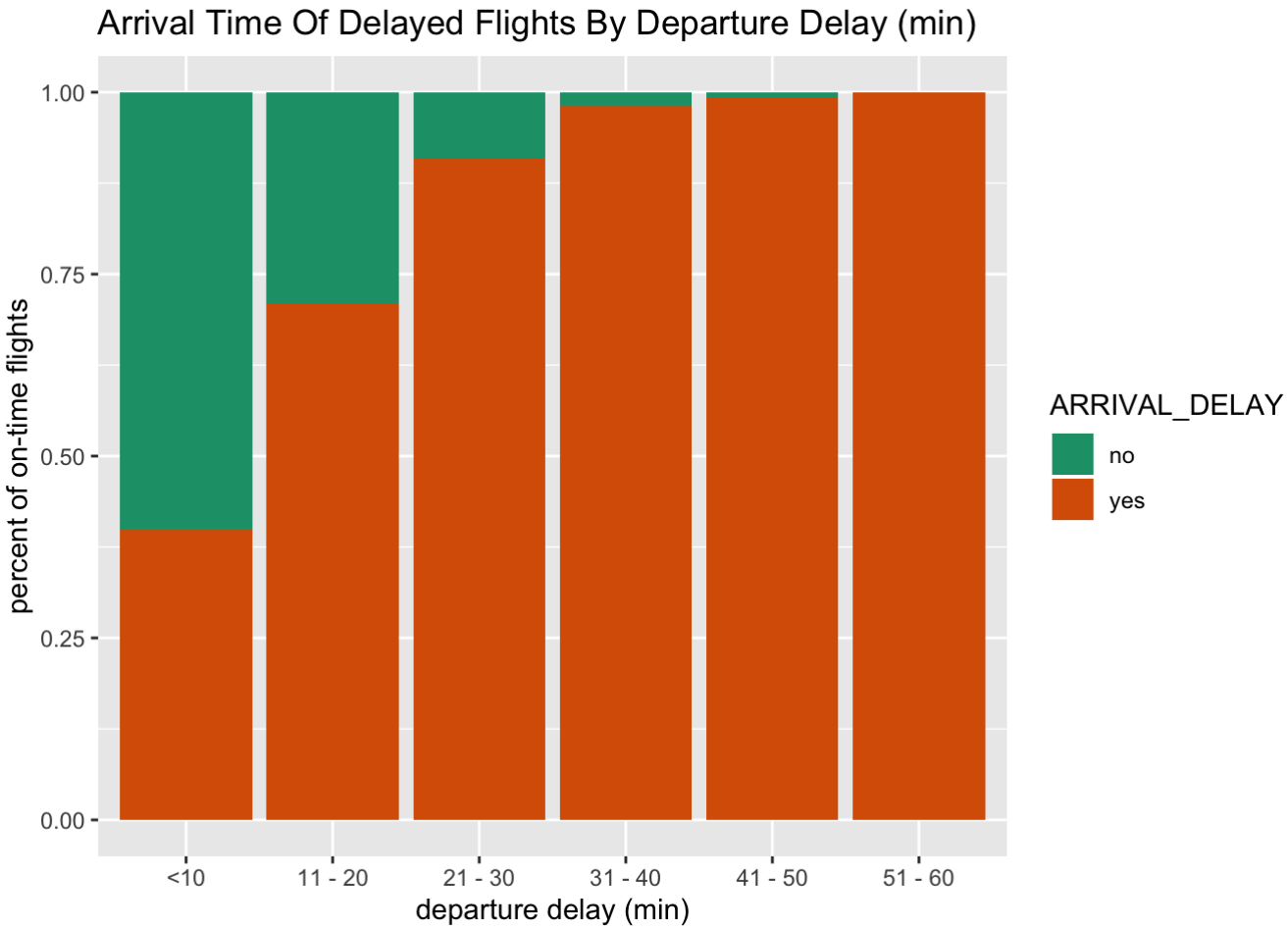
Question What impacts on whether or not a delayed flight will land on time?

In the chart below we can see that the longer the departure delay is in minutes, the less likely a flight is to land on time at the destination airport. For flights that departed less than 10 minutes late from Boston in 2015, almost 60% still landed on time at their destination airport. For flights that departed between 10 and 20 minutes late, over a quarter landed on time. However, 100% of flights that departed an hour or later also landed late. As we saw earlier, more than 1/3 of flights departed late out of Logan in 2015. Because of the likeliness of departing late airlines likely pad their flight times so that they are able to appear on schedule.

```
distance_buckets <- cut(x = delayed_flights_only$DISTANCE, breaks = c(0,500,1000,1500,20
00,2500,3000))
levels(distance_buckets) <- c("<500","501 - 1000", "1001 - 1500", "1501 - 2000", "2001 -
2500","2501 - 3000",">3000")
delayed_flights_only$distance_buckets <- distance_buckets

bos_flights_delay_expand$dep_delay_minutes <- flights$DEPARTURE_DELAY
bos_flights_delay_expand$arr_delay_minutes <- flights$ARRIVAL_DELAY

summary(na.omit(flights$DEPARTURE_DELAY))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  -35.000   -5.000   -2.000    9.606    7.000 1190.000
```

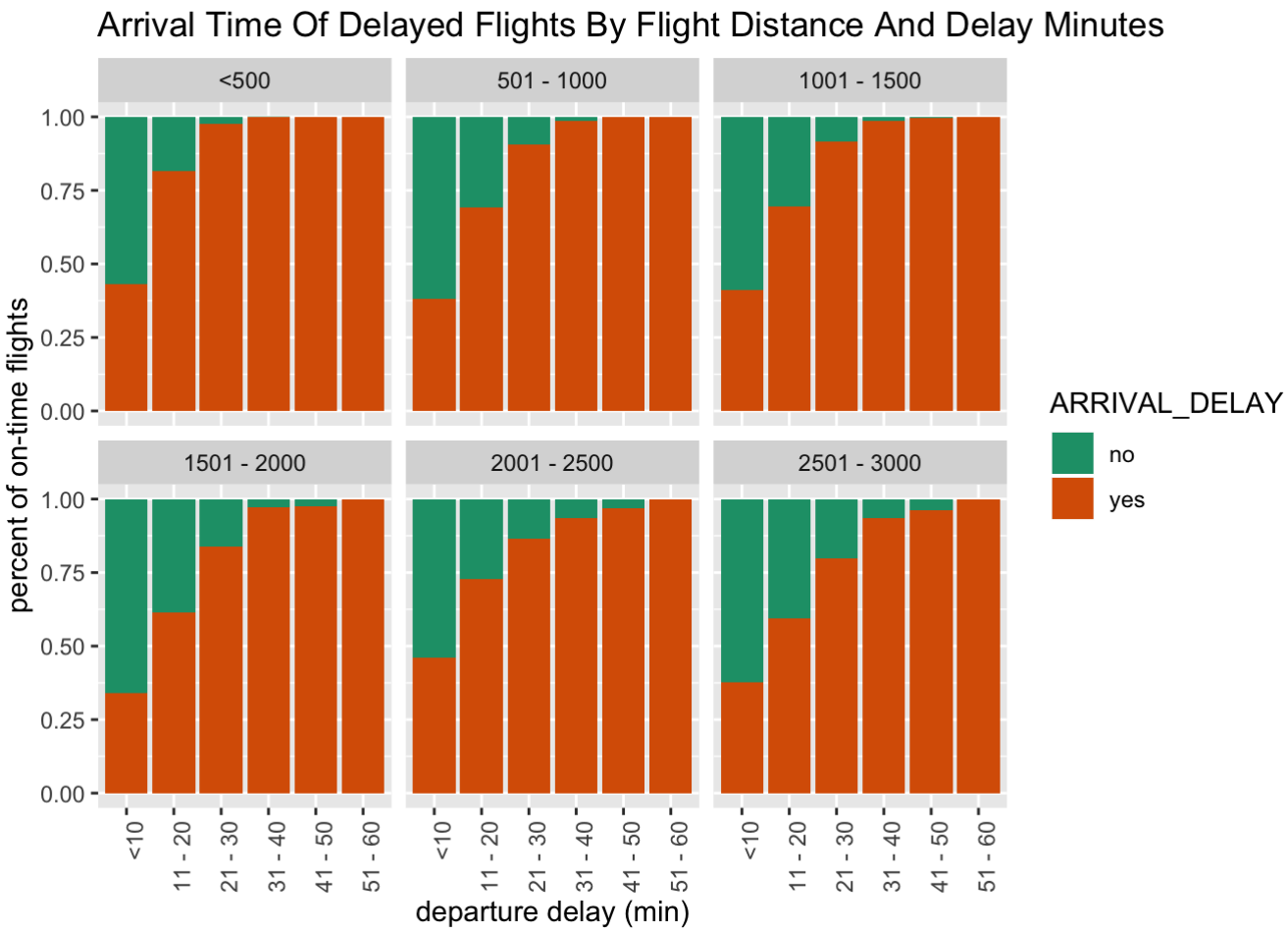
```
bos_flights_delay_lt60 <- filter(bos_flights_delay_expand, dep_delay_minutes > 0 & dep_d
elay_minutes <= 60)

dep_delay_buckets <- cut(x = bos_flights_delay_lt60$dep_delay_minutes, breaks = c(0,10,2
0,30,40,50,60))
levels(dep_delay_buckets) <- c("<10", "11 - 20", "21 - 30", "31 - 40","41 - 50","51 - 6
0")
bos_flights_delay_lt60$dep_delay_buckets <- dep_delay_buckets

ggplot(data = na.omit(bos_flights_delay_lt60)) +
  geom_bar(mapping = aes(x = dep_delay_buckets, fill = ARRIVAL_DELAY), position = "fill"
) +
  labs(title = "Arrival Time Of Delayed Flights By Departure Delay (min)", x = "departur
e delay (min)", y = "percent of on-time flights") +
  scale_fill_brewer(palette = "Dark2")
```



Question: Are flight times for longer distance flights padded more than for shorter flights?

As the distance of flights increases, there is a slightly higher chance that the flight will land on-time despite a departure delay. However, no flights that had a departure delay of over 50 minutes were able to overcome landing late no matter the flight distance.

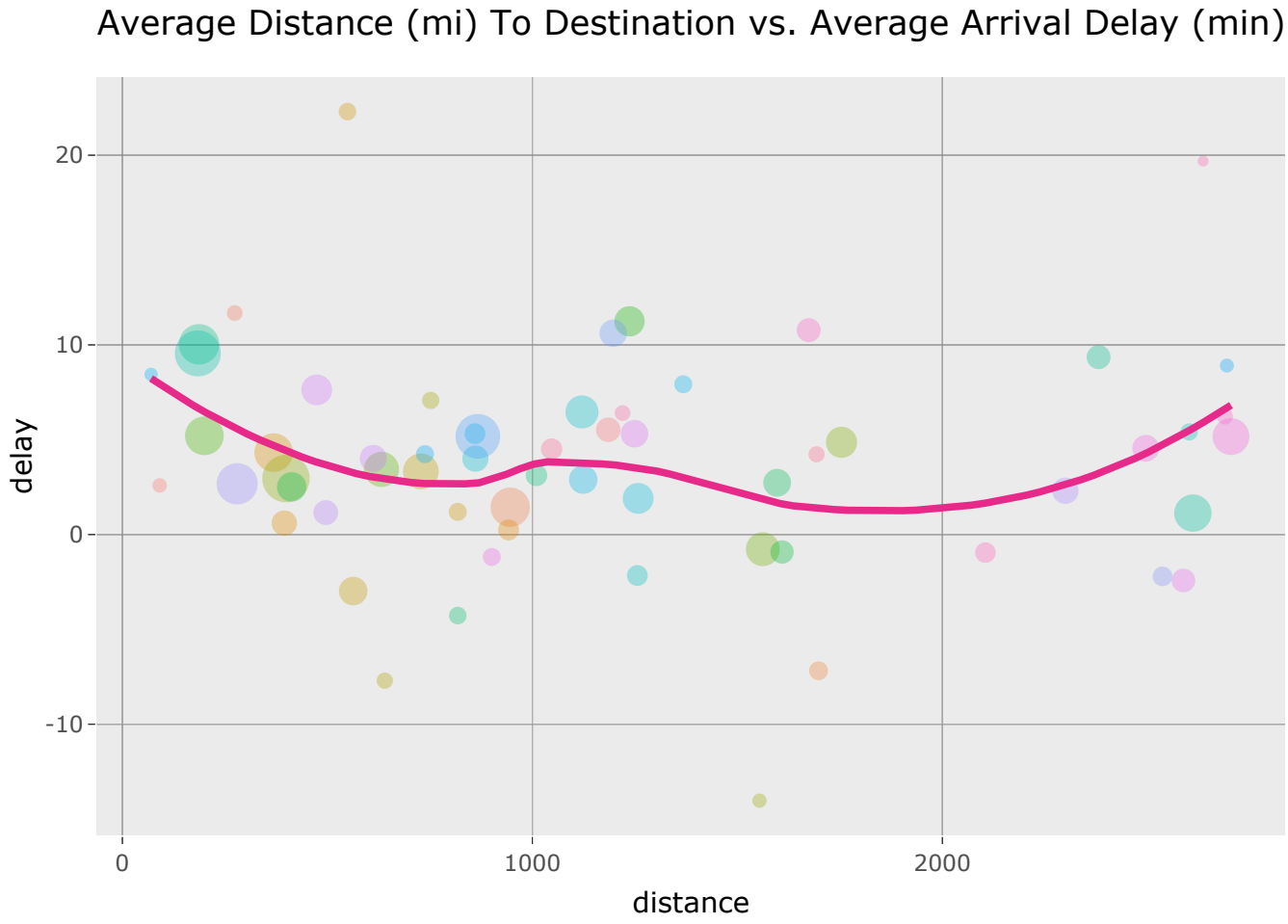
```
distance_buckets_2 <- cut(x = bos_flights_delay_lt60$DISTANCE, breaks = c(0,500,1000,150
0,2000,2500,3000))
levels(distance_buckets_2) <- c("<500","501 - 1000", "1001 - 1500", "1501 - 2000", "2001
- 2500","2501 - 3000",">3000")
bos_flights_delay_lt60$distance_buckets <- distance_buckets_2

ggplot(data = na.omit(bos_flights_delay_lt60)) +
  geom_bar(mapping = aes(x = dep_delay_buckets, fill = ARRIVAL_DELAY), position = "fill"
)+
  facet_wrap(~ distance_buckets) +
  scale_fill_brewer(palette = "Dark2") +
  labs(title = "Arrival Time Of Delayed Flights By Flight Distance And Delay Minutes", x
= "departure delay (min)", y = "percent of on-time flights") +
  theme(axis.text.x=element_text(angle=90, hjust=1))
```



Question: What is the relationship between average distance and the average arrival delay of flights from Boston to another domestic airport?

The relationship between average distance and arrival delay based on all delayed flights appears to be split in different directions based on a cut-off distance with flights up to 1000 miles having a steeper negative relationship with arrival delay compared to the distance between 1000 and 2000 miles. However, flights that fly more than 2000 miles have a positive relationship with arrival delay that explains arrival delay increases for flights that have destinations over 200 miles from Boston.

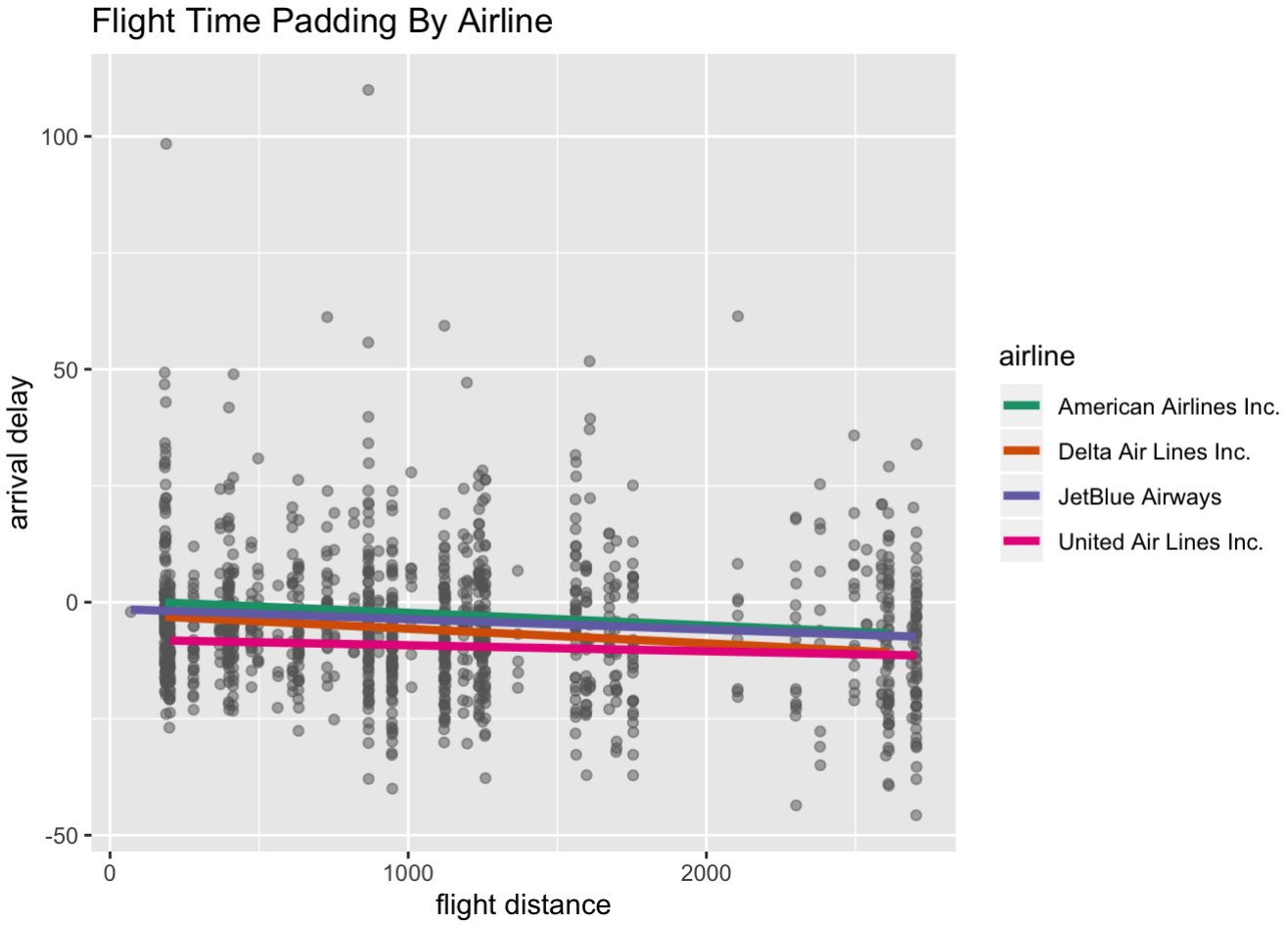
```
p <- ggplot(data = delay, mapping = aes(x = distance, y = delay)) +
  geom_point(aes(size = count, color = DESTINATION_AIRPORT), alpha = 1/3) +
  geom_smooth(se = FALSE, color = "#E7298A") +
  labs(title = "Average Distance (mi) To Destination vs. Average Arrival Delay (min)",
      caption = "The circle size shows the number of flights to that destination.")

ggplotly(p + theme(legend.position = "none") + scale_fill_brewer(palette = "Dark2"))
```



Question Which airline pads their flight times the most?

Of the top four most popular airlines, it appears that all four airlines pad their flight times. If a flight departs on time we can see in general that it will land early, also the longer a flight is the more likely it is to land early if it departs on time. The pad time amongst airlines seems to be fairly similar, but we can see in the chart below that the estimated arrival line for United flights that depart on time is completely in the negative, meaning in general, if a United flight departs on time it will most likely land early. In contrast, American Airlines seem to pad their flights the least.

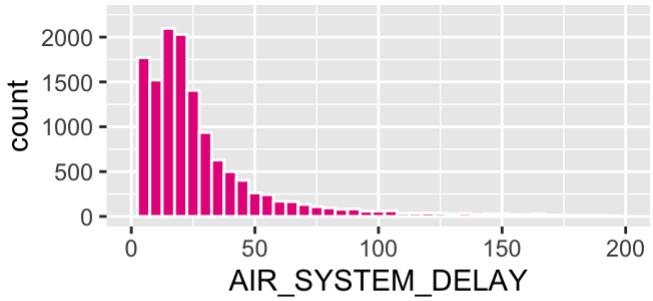
```
on_time_flights <- filter(flights, DEPARTURE_DELAY == 0)
on_time_flights_top_airlines <- filter(on_time_flights, AIRLINE == c("AA","B6","DL","UA"
))

#ontime flights by airline
ggplot(data = on_time_flights_top_airlines, aes(x = DISTANCE, y = ARRIVAL_DELAY, color =
AIRLINE.y)) +
  geom_jitter(alpha = .5, color = "#666666") +
  geom_smooth(method = "glm", se = FALSE, size = 1.5 ) +
  scale_colour_brewer(palette = "Dark2") +
  labs(title = "Flight Time Padding By Airline", x = "flight distance", y = "arrival del
ay", color = "airline")
```

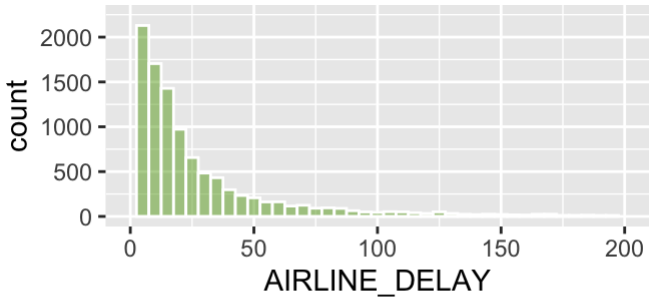


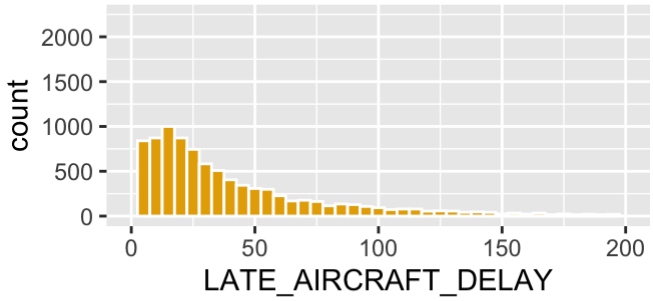
Question Which type of delay is the worst?

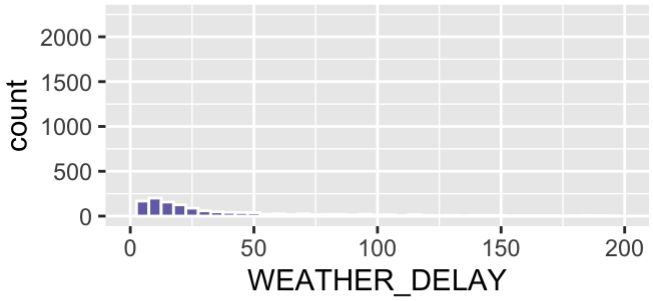
The dataset includes 5 major types of weather delays. The most common delay type is an air system delay with 14,179 flights impacted by this delay type at Logan in 2015. The average air system delay was 28 minutes long. The second most common delay type was airline delay which impacted 10,934 flights with an average delay time of 29 minutes. The third and fourth most common delay types, late inbound aircraft and weather delay, while being less common, had a longer average delay time of 43 and 44 minutes respectively.

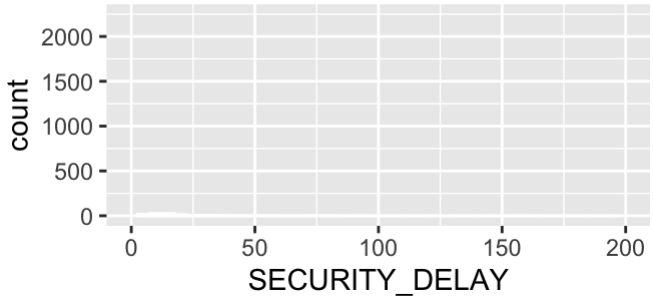
```
air_system_delay <- filter(flights, AIR_SYSTEM_DELAY > 0, AIR_SYSTEM_DELAY < 300)
weather_delay <- filter(flights, WEATHER_DELAY > 0, WEATHER_DELAY < 300)
late_aircarft_delay <- filter(flights, LATE_AIRCRAFT_DELAY > 0, LATE_AIRCRAFT_DELAY < 30
0)
security_delay <- filter(flights, SECURITY_DELAY > 0, SECURITY_DELAY < 300)
airline_delay <- filter(flights, AIRLINE_DELAY > 0, AIRLINE_DELAY < 300)

air_system_delay <- filter(flights, AIR_SYSTEM_DELAY > 0, AIR_SYSTEM_DELAY < 300)
weather_delay <- filter(flights, WEATHER_DELAY > 0, WEATHER_DELAY < 300)
late_aircarft_delay <- filter(flights, LATE_AIRCRAFT_DELAY > 0, LATE_AIRCRAFT_DELAY < 30
0)
security_delay <- filter(flights, SECURITY_DELAY > 0, SECURITY_DELAY < 300)
airline_delay <- filter(flights, AIRLINE_DELAY > 0, AIRLINE_DELAY < 300)

airsystemPlot <- ggplot() +
  geom_histogram(data = air_system_delay, aes(x = AIR_SYSTEM_DELAY),binwidth = 5, fill =
"#E7298A", color = "white") +
  ylim(0, 2250) + xlim(0,200)
airlinePlot <- ggplot() +
  geom_histogram(data = airline_delay, aes(x = AIRLINE_DELAY), binwidth = 5, fill = "#66
A61E", alpha = .5, color = "white") +
  ylim(0, 2250) + xlim(0,200)
lateAircraftPlot <- ggplot() +
  geom_histogram(data = late_aircarft_delay, aes(x = LATE_AIRCRAFT_DELAY),binwidth = 5,
 fill = "#E6AB02", color = "white") +
  ylim(0, 2250) + xlim(0,200)
weatherDelayPlot <- ggplot() +
  geom_histogram(data = weather_delay, aes(x = WEATHER_DELAY), binwidth = 5, fill = "#75
70B3", color = "white") +
  ylim(0, 2250) + xlim(0,200)
securityDelayPlot <- ggplot() +
  geom_histogram(data = security_delay, aes(x = SECURITY_DELAY),binwidth = 5, fill = "#6
66666", color = "white") +
  ylim(0, 2250) + xlim(0,200)

grid.arrange(airsystemPlot, airlinePlot, lateAircraftPlot, weatherDelayPlot, securityDel
ayPlot, nrow = 3)
```












```
summary(air_system_delay$AIR_SYSTEM_DELAY)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.00   19.00   28.77   33.00  299.00
```

```
nrow(air_system_delay)
```

```
## [1] 14179
```

```
summary(airline_delay$AIRLINE_DELAY)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   15.00   29.01   33.00  298.00
```

```
nrow(airline_delay)
```

```
## [1] 10934
```

```
summary(late_aircarft_delay$LATE_AIRCRAFT_DELAY)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   14.00   28.00   43.34   56.00  298.00
```

```
nrow(late_aircarft_delay)
```

```
## [1] 9317
```

```
summary(weather_delay$WEATHER_DELAY)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   11.00   22.00   43.64   54.00  297.00
```

```
nrow(weather_delay)
```

```
## [1] 1371
```

```
summary(security_delay$SECURITY_DELAY)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    8.00   14.00   17.84   21.00   81.00
```

```
nrow(security_delay)
```

```
## [1] 117
```

Analysis based on popular airlines

Question: Which airline has a better solution for departure delay?

```
flights %>%
 count(AIRLINE, sort = TRUE) %>%
 head(10)
```

AIRLINE <chr>	**n** <int>
B6	37962
AA	17629
DL	14256
UA	11282
WN	9956

AIRLINE <chr>	n <int>
US	8984
EV	2300
NK	2226
VX	1593
AS	1401

1-10 of 10 rows

We chose the 5 airlines based on a combination of the highest number of flights out of Boston in 2015 and general popularity in the last 2 years. The 5 airlines include JetBlue Airways (B6), American Airlines (AA), Delta Air Lines (DL), United Airlines (UA), and Alaska Airlines (AS).

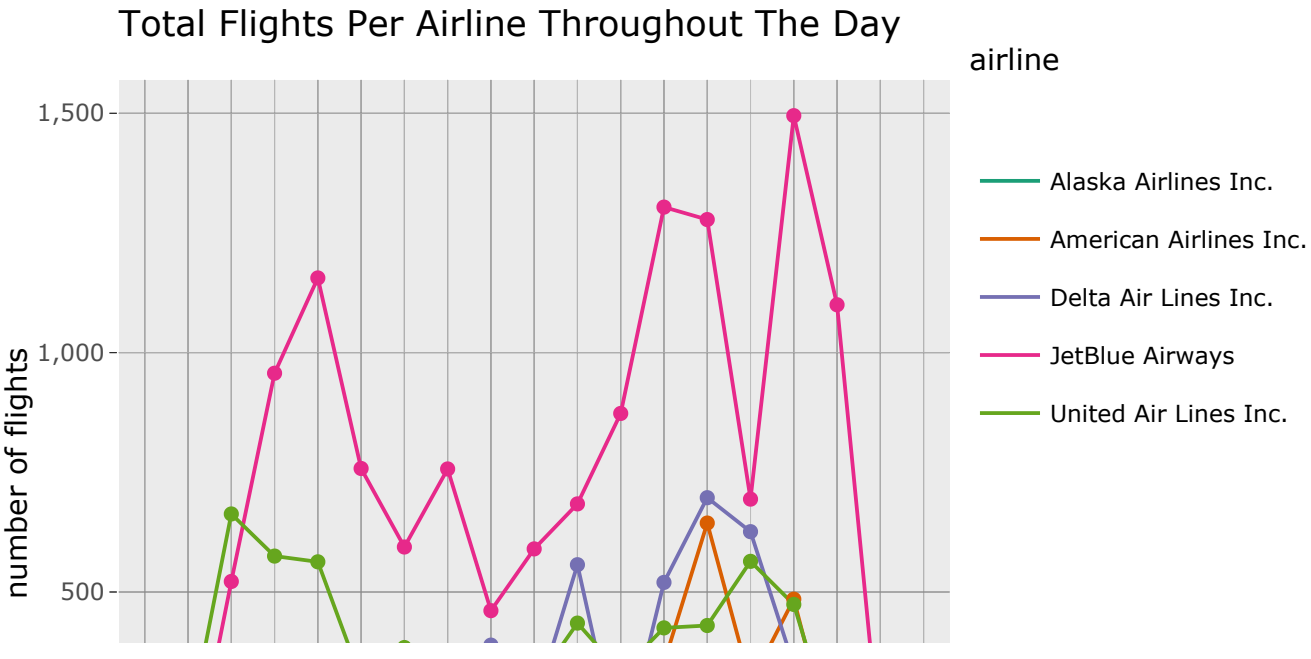
Plot 1 shows the total number of flights by the 5 airlines throughout the day. We found Alaska Airlines with both the least total number of flights and the least time scale for departure. Delta Airlines, JetBlue Airways, and American Airlines have a large number of flights departing in the afternoon and evening. United Airlines, on the other hand, has a relatively flat distribution of flight departure time.

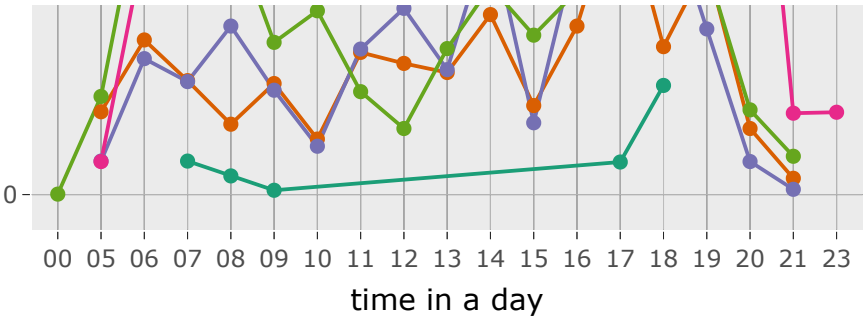
Plot 2 shows the average delay time throughout the day for departure and arrival, categorized by 5 popular airlines. The plot shows an increasing trend of delay time from morning to evening, which is the same as the data we got in the last project. An airline with a good solution for flight delay means the passenger will acquire a relatively low time cost when their flight is delayed. A low time cost if the arrival delay is less than the departure delay. In this plot, United Airlines and Alaska Airlines both have their arrival delay time less than departure delay time. The low time cost for American Airlines only affects their morning flights. And the other two airlines do not show an obvious difference in the arrival delay time and departure delay time.

Plot 1: The total number of flights in 2015, sorted by the time throughout the day and airline.

```
flights %>%
 filter( (AIRLINE == "UA"|
           AIRLINE == "AA"|
           AIRLINE == "AS"|
           AIRLINE == "DL"|
           AIRLINE == "B6"), DEPARTURE_DELAY > 0) %>%
 select(airline = AIRLINE, flight_num = FLIGHT_NUMBER,
        depart_time = SCHEDULED_DEPARTURE,
        depart_delay = DEPARTURE_DELAY,
        arrive_delay = ARRIVAL_DELAY) -> a
# time bucket
as.character(a$depart_time)
a$depart_time <- str_pad(a$depart_time, width = 4, side = "left", pad = 0)
a$depart_time <- str_sub(a$depart_time, start = 1, end = 2)
a %>%
 group_by(depart_time, airline) %>%
 summarise(count = n()) -> e
sum(e$count)
names(e) <- c("day_time", "iata_code", "num_of_flights")
names(airline) <- tolower(names(airline))
fl_num <- merge(e, airline, by = "iata_code")
```

```
p1 <- ggplot(fl_num, aes(x = day_time, y = num_of_flights,
                         color = airline,
                         group = airline)) +
  geom_line(lineend = "butt") +
  geom_point() +
  scale_y_continuous(labels = comma) +
  xlab("time in a day") +
  ylab("number of flights") +
  labs(title = "Total Flights Per Airline Throughout The Day") +
  scale_color_brewer(palette = "Dark2")

ggplotly(p1)
```

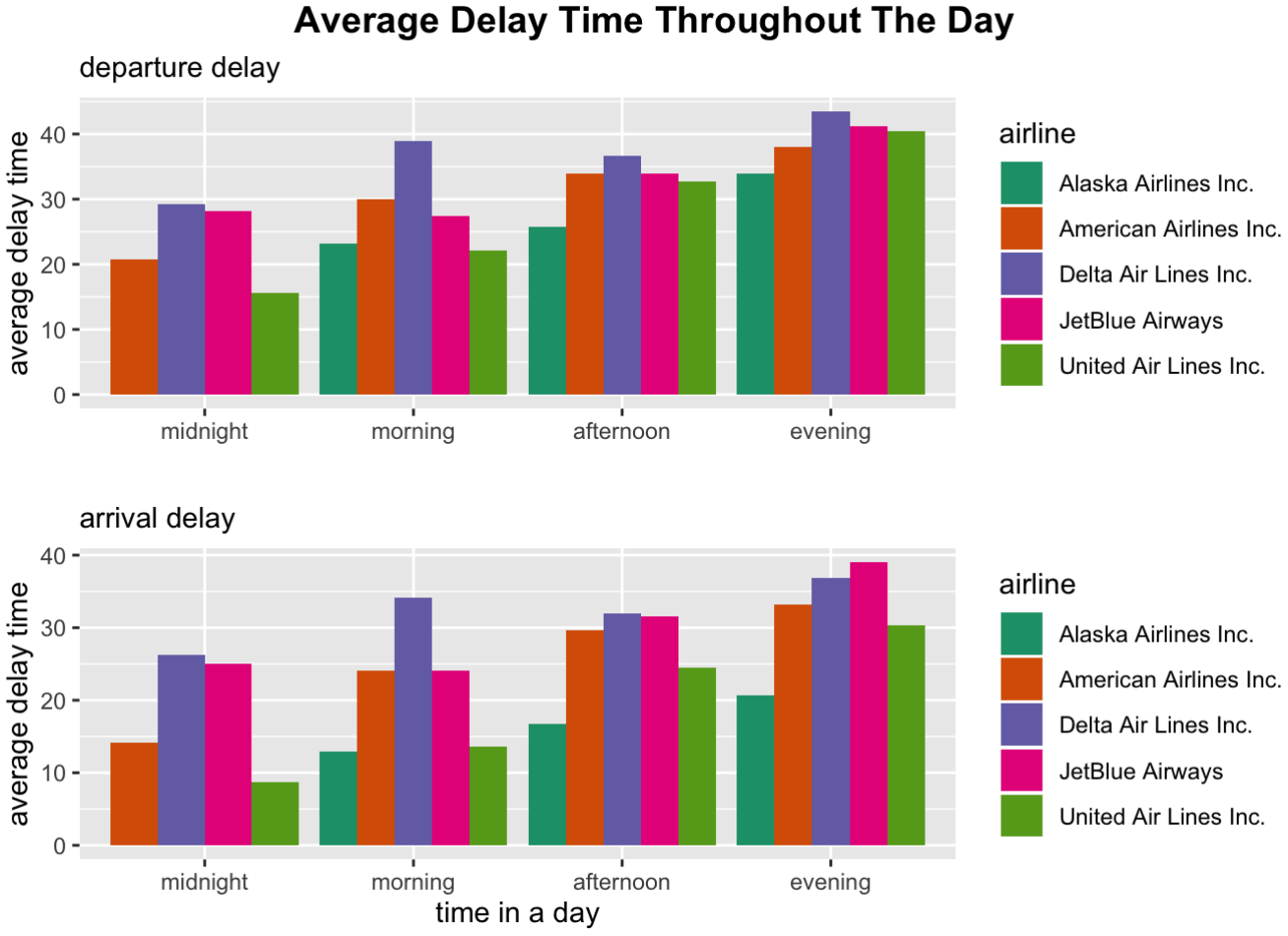




Plot 2: Compare the departure delay time and the arrival delay time of 5 popular airlines throughout the day.

```
a <- flights %>%
  filter( AIRLINE == "UA"|
            AIRLINE == "AA"|
            AIRLINE == "AS"|
            AIRLINE == "DL"|
            AIRLINE == "B6") %>%
  filter(DEPARTURE_DELAY > 0)

b2 <- data.frame(a$AIRLINE, a$FLIGHT_NUMBER, a$SCHEDULED_DEPARTURE, a$DEPARTURE_DELAY,
                 a$ARRIVAL_DELAY)
names(b2) <- c("airline","flight_num","depart_time","depart_delay","arrive_delay")

# time divided to four buckets
b2$depart_time <- as.numeric(b2$depart_time)
time_buckets <- cut(x=b2$depart_time, breaks = c(0, 600, 1200, 1800, 2400))
levels(time_buckets) <- c("midnight", "morning", "afternoon", "evening")
b2$depart_time <- time_buckets

# caculate average for delay time
c <- b2 %>%
  group_by(depart_time,airline) %>%
  summarise(avg_depart = mean(depart_delay, na.rm = TRUE))

d <- b2 %>%
  group_by(depart_time, airline) %>%
  summarise(avg_arrive = mean(arrive_delay, na.rm = TRUE))

fl <- merge(c,d, by = c("depart_time","airline"))
names(fl) <- c("day_time", "iata_code", "avg_ddt", "avg_art")
fl_air <- merge(fl, airline, by = "iata_code")
names(fl_air) <- tolower(names(fl_air))
```

```
# p2
departure_delay <- ggplot(fl_air) +
  geom_bar(aes(x = day_time, y=avg_ddt, fill = airline),
           position = "dodge",
           stat = "identity") +
  labs(x = "", y = "average delay time", subtitle = "departure delay") +
  scale_fill_brewer(palette = "Dark2")

arrive_delay <- ggplot(fl_air) +
  geom_bar(aes(x = day_time, y=avg_art, fill = airline),
           position = "dodge",
           stat = "identity") +
  xlab("time in a day") +
  ylab("average delay time") +
  labs(x = "time in a day", y = "average delay time", subtitle = "arrival delay") +
  scale_fill_brewer(palette = "Dark2")

dba <- ggarrange(departure_delay, arrive_delay,
          ncol = 1, nrow = 2) +
       scale_fill_brewer(palette = "Dark2")

annotate_figure(dba,
                top = text_grob("Average Delay Time Throughout The Day", face = "bold",
 size = 14))
```

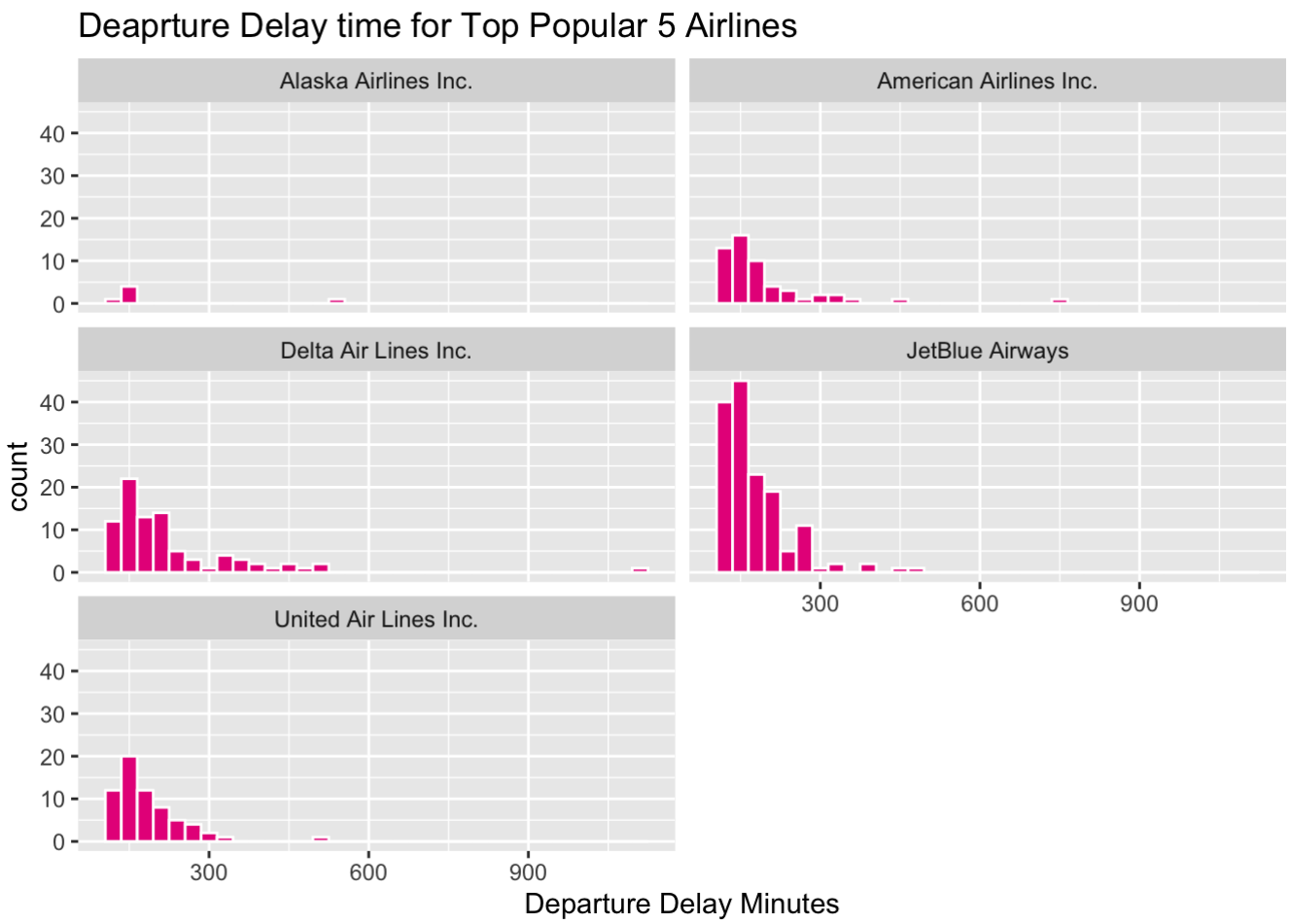


Question: Which airlines had the most delayed flights in the whole year?

We have filtered five most popular airlines, which is American Airlines (AA), Alaska Airlines (AS), JetBlue Airways (B6), Delta Air Lines (DL), and United Airlines (UA). Firstly, let us focus on when the departure delay time is less than 300 minutes. From the above charts, we could see that JetBlue has the most number of delayed flights. The main reason is that we have more observations of JetBlue, which is almost twice that of American Airlines and four times that of Delta. Delta, United, and American have similar observations.

When we compare them, we will find that Delta has the most delayed flights, followed by United, then American. Alaska only has 1400 observations; thus, Alaska has the fewest number of delayed flights. Let us move to the departure delay time greater than 300 minutes, Delta has the most delayed flights. One thing that caught our attention was that Delta had the longest delay, over 1,000 minutes in February. We dug up the data and found that the longest delay was caused by poor weather. This is because the snowiest month on record in Boston was in February 2015, and snowfall reached 65 inches.

```
flights %>%
 filter(DEPARTURE_DELAY > 120,AIRLINE.y==c("American Airlines Inc.",
                                           "JetBlue Airways",
                                           "Delta Air Lines Inc.",
                                           "United Air Lines Inc.",
                                           "Alaska Airlines Inc.")) %>%
 ggplot()+
 geom_histogram(aes(x=DEPARTURE_DELAY),fill = "#E7298A",binwidth=30, color = "white")+
 facet_wrap(~AIRLINE.y,ncol = 2)+
 labs(x="Departure Delay Minutes",
      title="Deaprture Delay time for Top Popular 5 Airlines",hjust=0.5)
```



```
flights %>%
  count(AIRLINE.y) %>%
  arrange(desc(n))
```

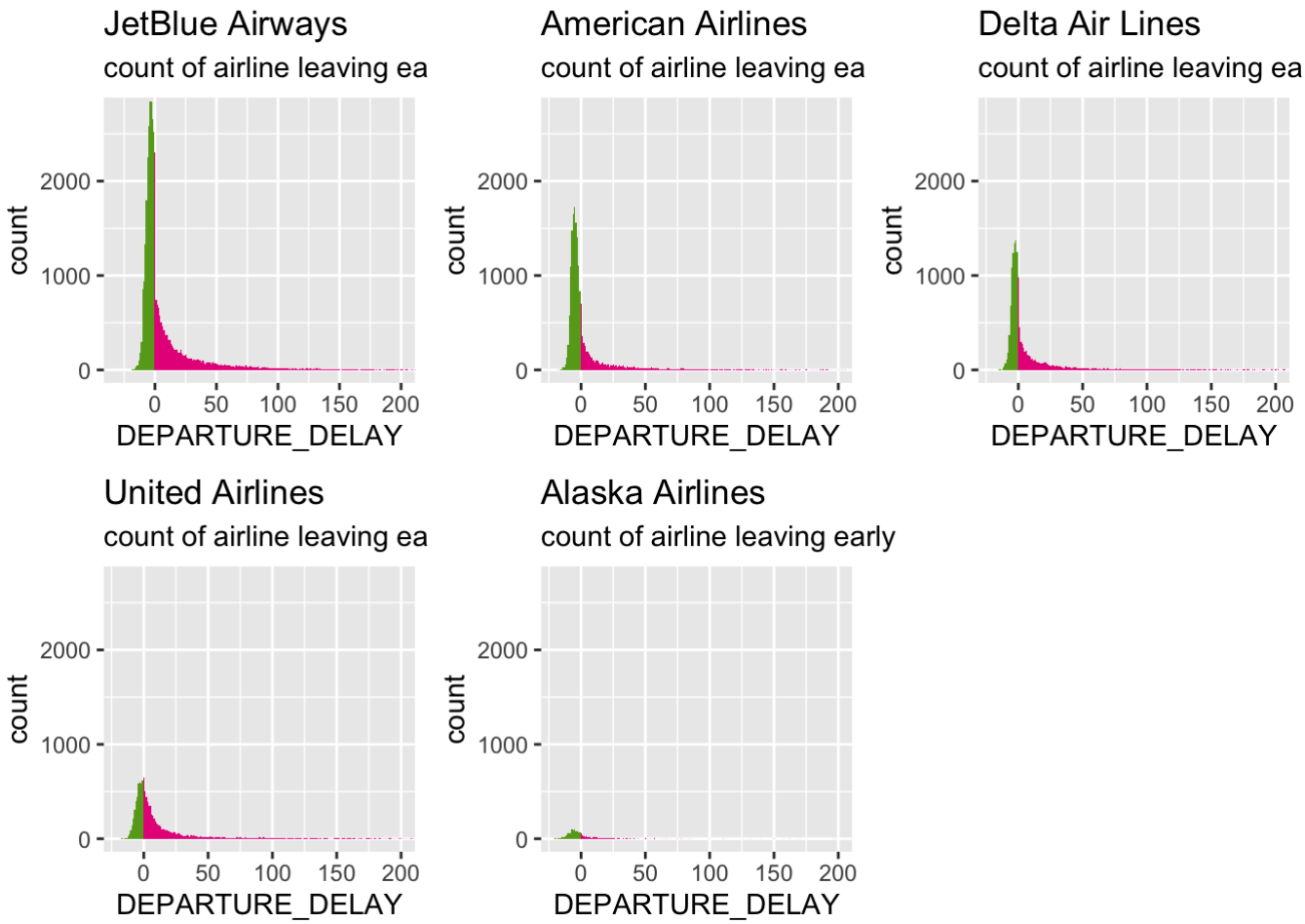
AIRLINE.y <chr>	n <int>
JetBlue Airways	37962
American Airlines Inc.	17629
Delta Air Lines Inc.	14256
United Air Lines Inc.	11282
Southwest Airlines Co.	9956
US Airways Inc.	8984
Atlantic Southeast Airlines	2300
Spirit Air Lines	2226
Virgin America	1593
Alaska Airlines Inc.	1401

1-10 of 11 rows Previous **1** 2 Next

Question: How many flights depart early and late for each airline?

In the charts below, we can see that for Jetblue, American, Delta, and Alaska a large number of their flights depart either on time or early. United has a fewer percentage of flights departing on-time or early as compared to the other airlines. Each histogram has a long right tail which tells us that if a flight departs late, instead of on-time or early, there is a large variation in how late the flight will depart.

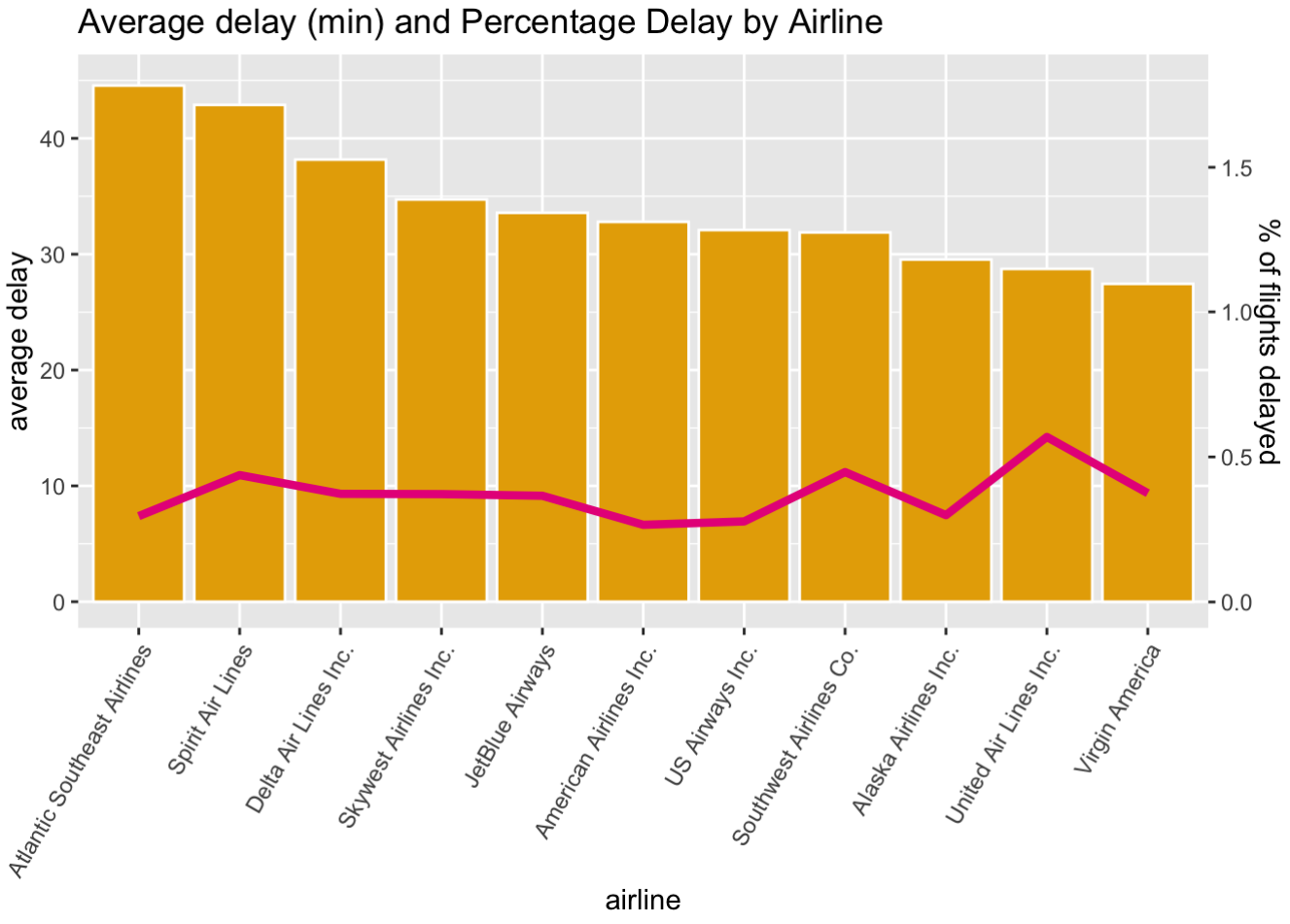
```
flights_B6 <- filter(flights,AIRLINE=="B6")
d1 <- ggplot(flights_B6,aes(x=DEPARTURE_DELAY))+
  geom_histogram(aes(fill = DEPARTURE_DELAY < 0), binwidth = 1, show.legend = F)+
  labs(title="JetBlue Airways",
       subtitle = "count of airline leaving early")+
  coord_cartesian(ylim = c(0,2750), xlim = c(-30,200))+
  scale_fill_manual(values = c("#E7298A","#66A61E"))

flights_AA <- filter(flights,AIRLINE=="AA")
d2 <- ggplot(flights_AA,aes(x=DEPARTURE_DELAY))+
  geom_histogram(aes(fill= DEPARTURE_DELAY < 0), binwidth = 1, show.legend = F)+
  labs(title="American Airlines",
       subtitle = "count of airline leaving early")+
  coord_cartesian(ylim = c(0,2750), c(-20,200))+
  scale_fill_manual(values = c("#E7298A","#66A61E"))

flights_DL <- filter(flights,AIRLINE=="DL")
d3 <- ggplot(flights_DL,aes(x=DEPARTURE_DELAY))+
  geom_histogram(aes(fill= DEPARTURE_DELAY < 0), binwidth = 1, show.legend = F)+
  labs(title="Delta Air Lines",
       subtitle = "count of airline leaving early")+
  coord_cartesian(ylim = c(0,2750), c(-20,200))+
  scale_fill_manual(values = c("#E7298A","#66A61E"))

flights_UA <- filter(flights,AIRLINE=="UA")
d4 <- ggplot(flights_UA,aes(x=DEPARTURE_DELAY))+
  geom_histogram(aes(fill= DEPARTURE_DELAY < 0 ), binwidth = 1, show.legend = F)+
  labs(title="United Airlines",
       subtitle = "count of airline leaving early")+
  coord_cartesian(ylim = c(0,2750), c(-20,200))+
  scale_fill_manual(values = c("#E7298A","#66A61E"))

flights_AS <- filter(flights,AIRLINE=="AS")
d5 <- ggplot(flights_AS,aes(x=DEPARTURE_DELAY))+
  geom_histogram(aes(fill= DEPARTURE_DELAY < 0 ), binwidth = 1, show.legend = F)+
  labs(title="Alaska Airlines",
       subtitle = "count of airline leaving early")+
  coord_cartesian(ylim = c(0,2750), c(-20,200))+
  scale_fill_manual(values = c("#E7298A","#66A61E"))

grid.arrange(d1,d2,d3,d4,d5, nrow = 2)
```



Question: Which airlines have the shortest average departure delays and least amount of flights delayed?

Atlantic Southeast Airlines had the longest delays on average with the average length of a delay being around 45 minutes. However, United had the most delayed flights with over 50% of their flights experiencing a delay.

```
flights %>%
    filter(DEPARTURE_DELAY > 0 ) %>%
    group_by(AIRLINE.y) %>%
    summarize(avg_delay = mean(DEPARTURE_DELAY, na.rm = TRUE)) -> avgDelayAirline

  flights %>%
    group_by(AIRLINE.y) %>%
    mutate(delay10 = ifelse(DEPARTURE_DELAY > 0, 1, 0), notdelay10 = ifelse(DEPARTURE_DE
LAY <= 0, 1, 0)) %>%
    summarize(percentDelay = (sum(delay10, na.rm = TRUE)/(sum(delay10, na.rm = TRUE)+sum
(notdelay10, na.rm = TRUE)))) %>%
    left_join(avgDelayAirline, by = "AIRLINE.y") %>%

  ggplot()+
    geom_bar(aes(reorder(AIRLINE.y, -avg_delay),avg_delay),stat = "identity",fill="#E6AB
02",color="white")+
    geom_line(aes(AIRLINE.y,percentDelay*25,group=1), col = "#E7298A", size = 1.5)+
    labs(title = "Average delay (min) and Percentage Delay by Airline",
         x="airline",
         y="average delay")+
    theme(axis.text.x=element_text(angle=60, hjust=1)) +
    scale_y_continuous(limits = c(0, 45),
                       sec.axis = sec_axis(~./25, name = "% of flights delayed"))
```



Summary

Our project investigated flight delays out of Boston Logan International Airport for domestic flights in 2015 to identify the most time-efficient trips for travelers from Boston. By summarizing and organizing the original dataset, we examined the relationships between flight delays out of Boston with the destination, airline, time of day, and also examined whether or not departing late guarantees a flight will also arrive late.

In 2015 flights from Boston departed to 62 airports within the United States. The most popular airports by flight volume were DCA, LGA, and ORD with each airport receiving over 6,500 flights from Boston Logan. The summer months had the most flights, and Saturdays and Sundays were the least popular days to fly.

We then explored what months, time of day, and airlines experienced the most delays. February had the highest delay time on average in minutes for the year of 2015, this was likely caused by the fact that February was the snowiest month on record in Boston since 1891. Also, flights typically experienced longer departure and arrival delays in the morning as compared to the evening. Finally Jet Blue had the most number of flights departing early.

We dug in several different ways to explore the relationship between departing late and arriving late. Of the 39,000 flights that departed late, 12,000 still managed to arrive at their destination airport on time. We found that in general, the longer a flight is the more likely it would be to arrive on time despite experiencing a flight delay.