HW1 park 1

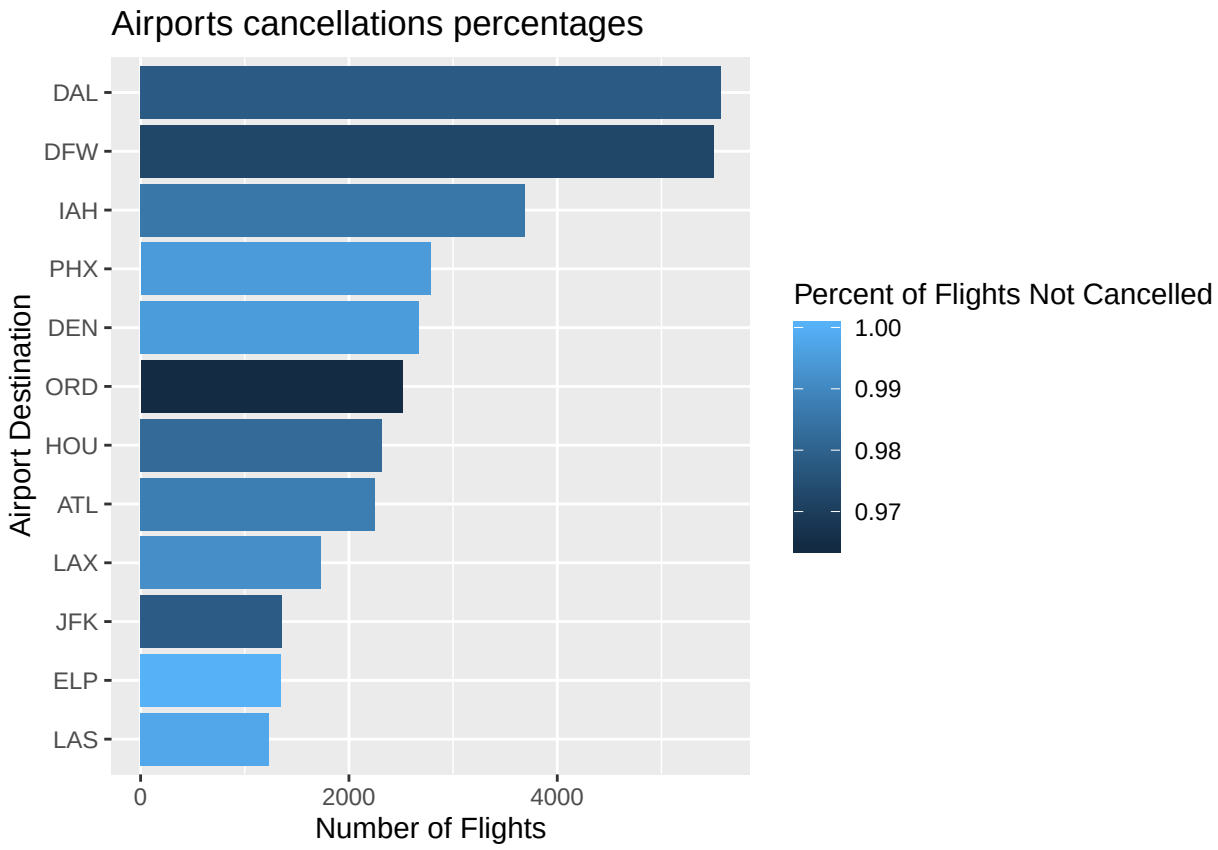
Alexander Rados, Sri Jonnalagadda, Kenny Kato

2/6/2020

The first plot we have is chart which compares airports by number of flights. The shading of the respective bars shows the percentage of cancelled flights. Dallas Love airport and Dallas Fort-Worth airport have the most flights and some of the higher percentages of cancellations. This reasons that the more flights you have the more cancellations you may occur hence the higher percentage of cancellations. The highest percentage of cancellations come out of Orlando which has the average amount of flights. On the lower tail of quaintly of flights, JFK airport has a high percentage of cancellations per number of flights.

What we can deduce from our plot is that one maybe cautious about cancellations when going through Orlando. Similar logic follows for the Dallas airports but we should keep in mind this maybe a repercussion from the higher quantity of flights. This theory would also lead us to be cautious with JFK and see with more data(flights) if the percentage of cancellations holds or decreases since the lack of comparable data may provide harsher results compared to other airports.

```
r  print(Cancels + ggtitle("Airports cancellations percentages"))
```



This plot is a break down of delays in minutes by each month. The coloration shows the individual airports. The goal of this is to see which month and time have the highest delays and which airports.

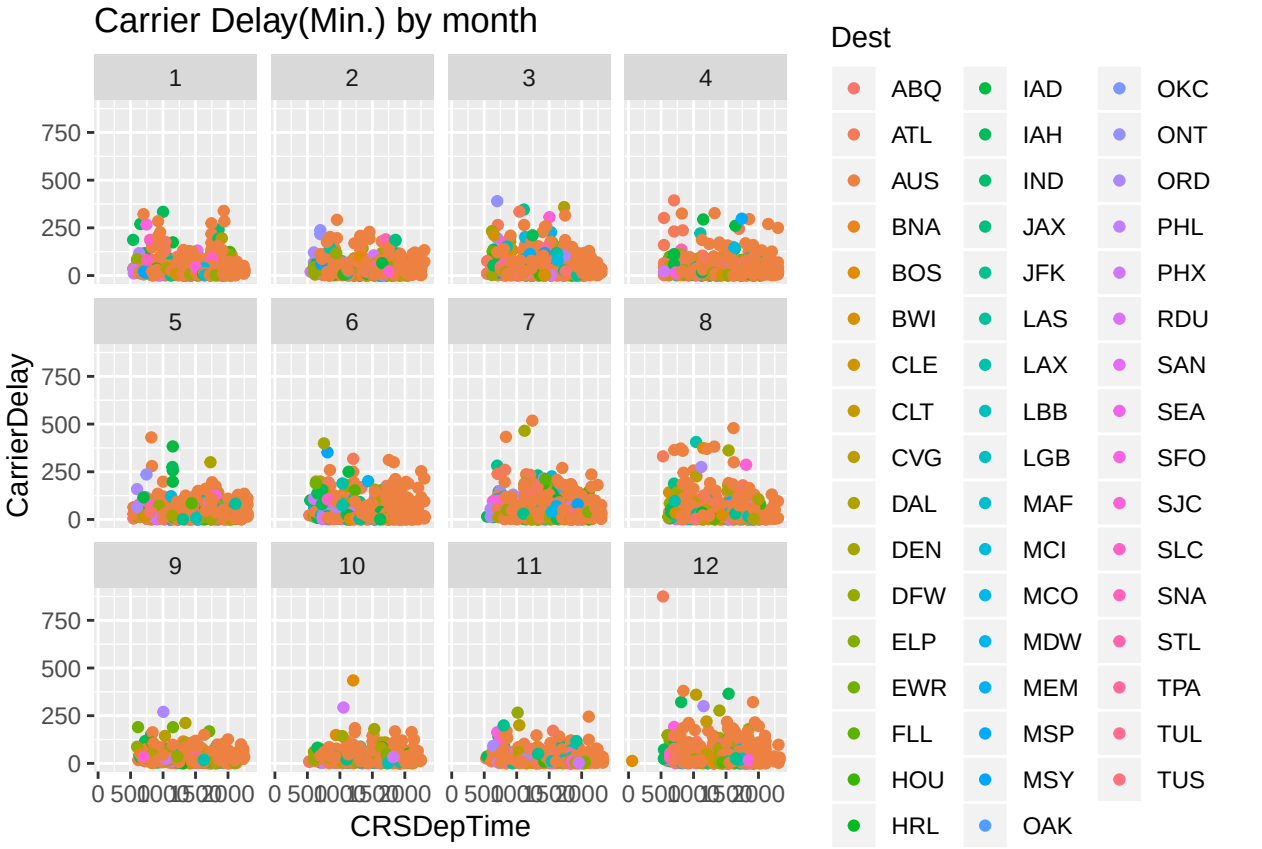
The general trend is that between midnight and 5 am there are little to no delays in any month by any airport. This reasons as fewer flights and few operations are occurring so we would expect a greater precision in on time flights.

The majority of our delayed flights happen during the daytime, 5 am to 11 PM. This logic also checks out

since more flights are occurring during the day. The longest delays are between 5am and 7pm so we could rationalize that this is the highest traffic period of the day.

As far as individual airports we cannot say one has a higher delay average than another. Nor do we see a month truly exceed another in delay times.

```
print(Carrier1 + ggtitle("Carrier Delay(Min.) by month"))
```



Here is our code for part 1.

library(tidyverse) library(tidyr) library(maps) # creating a dataset specifically for Austin departures ABIA <- read.csv('/Users/akhiljonnalagadda/Desktop/R/ABIA.csv') airport.codes_csv = read.csv('/Users/akhiljonnalagadda/Desktop/R/airport-codes.csv') aus_depart = filter(ABIA, Origin == "AUS") # grouping by destination and getting number of cancelled flights # figuring out portion of flights cancelled by destination by_dest = aus_depart %>% group_by(Dest) %>% summarise(count = n(), cancel = sum(Cancelled)) %>% filter(count>1000)

creating a new column based on proportion of flights that weren't cancelled

by destination

by_dest = mutate(by_dest, percent_flown = (count)/(count + cancel))

plotting flights and successful flights by destination

Cancels= ggplot(data = by_dest) + geom_bar(mapping = aes(x = reorder(Dest, count), y = count, fill = percent_flown), stat = "identity") + xlab("Airport Destination") + ylab("Number of Flights") + labs(fill = "Percent of Flights Not Cancelled") + coord_flip() # creating a map of the US usa = map_data(('state'), region = '.') # deleting unnecessary columns from airport codes keeps = c("iata_code", "coordinates") air_codes = airport.codes_csv[keeps] air_codes

dropping all the flights that don't have any delays and putting it into a new

data frame titled ABIA_no_na

ABIA_no_na = drop_na(ABIA)

filtering out all data that doesn't have a carrier delay or weather delay

car_del = filter(ABIA_no_na, CarrierDelay > 0) weather_del = filter(ABIA_no_na, WeatherDelay > 0) # plotting weather delayed flights on departure time while taking into account month ### edit

Carrier1= ggplot(data = car_del) + geom_point(mapping = aes(x = CRSDepTime, y = CarrierDelay, color = Dest)) + facet_wrap(~ Month)