Project 2 Writeup

Ankita Agrawal (aa653), Yezy Lim (yl647), Zachary Vinegar (zzv2)

Graph 1: US Domestic Flights by State

Data Sources

For my visualization, I downloaded data from three different sources: (1) Bureau of Transportation Statistics' Flight Origin and Destination Survey, (2) Bureau of Transportation Statistics' Airport Database, and (3) the United States Census Bureau.

(1) The Flight Origin and Destination Survey was used to get information on flight frequency between states. The dataset could only be downloaded in quarters, so I downloaded four datasets (1 for each quarter) so that I could have a dataset representing all of 2015. Also, the dataset was a random sample of 10% of total domestic flights in the U.S., so I multiplied all of the extracted values by 10. The variables I selected to analyze were 'OriginState', 'DestState', and 'Passengers'. 'OriginState' and 'DestState' were state abbreviations (e.g. 'NY') and 'Passengers' was the number of passengers for each flight booking, which usually ranged from 1 to 4. Each observation of this dataset represented 1 flight booking. The entire dataset for 2015 had more than 40 million observations, which is understandable since millions of flight bookings happen each year. Because my dataset had so many observations, I decided to filter it using R. I grouped the observations by the pair of 'OriginState' and 'DestState' to add the number of 'Passengers' to see the total number of people flying from and to different states in 2015. This condensed my dataset from +40 million rows to 2,667 rows. I then noticed that this dataset contained the Virgin Islands, U.S. military territories, and Puerto Rico in addition to the 50 states. I decided to keep Puerto Rico but omit the Virgin Islands and U.S. military territories from my data. However, representing 2,000+ observations when each row represented a unique flight route was still too hectic, so I calculated the percentage of people flying that specific route from the total number of people flying in 2015. Then I extracted observations with percentage greater than 0.001. My final dataset was 262 observations of unique 'OriginState' and 'DestState' pairs and number of 'Passengers' traveling that route.

Later in the project, I decided that I wanted to display the total number of people flying out of and into each state in 2015. I used the dataset of 2,667 rows from (1) and decided to create 2 queries and join them. For the first query, I grouped by 'OriginState' and summed up the number of 'Passengers' to calculate the number of people leaving each state ('outflow'). For the second query, I again grouped by 'DestState' and summed up the number of 'Passengers' to calculate the number of people entering each state ('inflow'). Then I joined those two queries by 'state' so I had a dataset of 'state', 'inflow', and 'outflow'.

(R script is attached for convenience.)

(2) The Airport Database was used to calculate how many airports were in each state. I decided to use the same source (Bureau of Transportation Statistics) for consistency. This dataset originally had a list of airport names for each state. I performed ROWS function on Excel, which calculated the number of rows (or airports) in a list. My final dataset had 'State' and 'AirportNum'.

(3) The United States Census Bureau was used to determine the number of people living in each state in 2015. This was an easy download that already had the dataset organized by 'State' and

'Population'. I combined this dataset with (2) to get a combined dataset of 'State', 'Population', and 'AirportNum'.

In total, I have 3 datasets: data_ext ('orig', 'dest', 'numPeople'), popData ('state', 'outflow', 'inflow'), and 2015data ('State', 'Population', 'AirportNum').

Mapping

For my chord diagram, the arcs represent the 50 states and Puerto Rico, and each region was mapped to a unique color using an ordinal scale which had a range of 51 colors. The chords represent the number of people flying between the two states. The width of the chord diagram is proportional to the number of people flying that route. Because chord diagrams are directional, I decided to use colors of the chord to represent direction. The color of the chord matches the color of the arc that has more passengers leaving than coming into that arc. For example, there are more people flying from Arizona to California than California to Arizona since the chord representing this route is purple, the same color as Arizona's arc (indeed, 539,887 fly from AZ to CA while 532,104 fly from CA to AZ). Finally, I have an axis on all of the arcs in units of millions. From the value range on the axis, you can see roughly how many people fly in and out of a state in total and how many people fly into or out of that state to another state specifically. For example, from looking at the axis on CA's arc, it's easy to see roughly 7 million people fly in and out of CA in total, and less than 1 million people fly between TX and CA. Finally, I display state names in the color corresponding to the state arc's color in the 'more information' section so that you can easily match the information to the arc of interest.

Story

The visualization confirms a lot of our intuition. First and foremost, there are A LOT of traveling domestically! The most popular states for traveling were Texas, Virginia, Washington, Arizona, California, Colorado, Florida, Georgia, Hawaii, Illinois, Nevada, New York, North Carolina, and Pennsylvania. This was expected because most of these states are major hubs for traveling, important cities, or vacation spots. California was the most popular state shortly followed by Texas, which can be explained by their large sizes and major cities. This means that not only do state residents fly within their state (LA to SF or Austin to Dallas) but also many people come for tourism and work. It was surprising to see Hawaii beat many other states in terms of popularity, but this is probably because Hawaii is a major vacation destination. Finally, it was interesting to see that the number of airports did not have an obvious correlation with the number of people traveling to and from that state. For example, Georgia has 8 airports and more than 4 million flyers, but New York has 21 airports and only 237,000 flyers. This can probably be explained by the state's geographic remoteness and size. Because NY is surrounded by other populated states and is relatively small, there might be a smaller demand to fly than in Georgia, which is more remote and larger. Another interesting discovery was that for each state, the numbers of people flying into and out of that state were very similar. For example, the number of people flying into NY was 237,943 and out of NY was 237,695. Likewise, the number of people flying into CA was 7,784,978 and out of CA was 7,755,611.

Graph 2: Flight Prices over Time

Data Sources

1) The average flight prices from 1993-2015 of all airports in the US:
   - Collected from: http://www.transtats.bts.gov/AverageFare/.
   - Details: To be more specific, the prices represented the Average Domestic Airline Itinerary Fares By Origin City for Year (Inflation Adjusted to 2009 dollars)
   - Filtering**:** We choose the 8 airports with highest passenger traffic, and didn't choose more since we didn't want to clutter the line graph

The next data sets were used in the pop-up box.

2) State GDP growth:
   - Collected from:
   - http://www.bea.gov/iTable/iTable.cfm?reqid=70&step=1&isuri=1&acrdn=2#reqid=70&step=5&isuri=1&7003=200&7004=naics&7035=-1&7006=00000&7001=1200&7036=-1&7002=1&7090=70&7007=-
   - Details: Since the top 8 airports depicted in the graph came from different states, I wanted to represent how the health of the state could affect the prices of flights. I thought that maybe more growth represented a richer state, therefore more demand for flying, and therefore more expensive flight prices. This was the case for some years, but not all of them.
   - Filtering: to the 8 states from which the 8 airports resided in

3) The number of flights departing the airport:
   - Collected from: http://www.transtats.bts.gov/Data_Elements.aspx?Data=1
   - Details: Similar to GDP growth, I wanted to see if the # of flights leaving that state in that year affected the price. This also helps related it to the graph above which depicts frequency of flights
   - Filtering: Once again, I filtered to only the 8 airports I was using

Mapping

I used a line plot that appended a path element with given x and y coordinates. The scale of the x axis was from the years 1993-2015 and the y axis was an appropriate range that captured the price ranges. Each colored line represents a separate airport. I used a tooltip box to create a popup and html code to format the information. I also appended circles at each data point so that I could use the functions on-mouseover and on-mouseout to append the min and max lines, tooltip div box, and make the line bolder.

The Story

What was really interesting about this graph was that flight prices actually went down over time! Please note that prices are inflation adjusted. People are always complaining about expensive flight prices, but they are cheaper than they have ever been. It is probably due to technology improvements. Also most airports have similar trends in flight prices, especially the giant dip in flight prices around the Great Recession (2009). Unfortunately, the correlation between the number of flights departing the airport and

also state GDP growth doesn't have an apparent correlation with flight price by scrolling over various data points.

Graph 3 – Flight Delays

Data

I used data from 2 different sources:

1. Bureau of Transportation Statistics' Flight On-Time Performance
2. aggdata.com Complete List of US Airport Locations

1. Flight On-Time Performance was used to get information on flight delays between airports. The dataset could only be downloaded by month, so we downloaded twelve datasets (1 for each month) so that we could have a dataset representing all of 2015. The variables we selected to analyze were OriginAirportID, DestAirportId, and ArrDelayMinutes. OriginAirportID and DestAirportId were airport codes (e.g. 'LAX') and ArrDelayMinutes was the difference in minutes between scheduled and actual arrival time with early arrivals set to 0. We decided to filter it using R. We grouped the observations by the pair of OriginAirportID and DestAirportId to average the amount of delay. This condensed our dataset from +6 million rows to 48,747 rows. We then noticed that this dataset contained too many airports to show on a map without clutter. So we decided to only show the top 8 airports and their routes between each other.
   (R script is attached for convenience.)

2. The Complete List of US Airport Locations was used to calculate the geographic coordinates for each airport code. We then used this data to plot the airport on a map of the United States.

Mapping

For my map, the nodes represent the 8 top airports plotted at the correct longitude and latitude. The edges represent the flight routes between airports and the color and weight of the route indicates the average delay over the selected time period in 2015. The color gradient goes from green to yellow to red where shorter delays are green (good) and longer delays are red (bad). This 3 color gradient really helps distinguish the different delay values by providing a good amount of contrast. Because these flight routes are directional, I decided to use the curvature of the route to identify the forward and backward direction. Routes going East to West are curved downward, while routes going West to East are curved upward. Similarly, routes going North to South are curved to the right, while routes going South to North are curved to the left.

Story

The visualization confirms our most intuitive guesses. The most delay happens in the coldest months of the winter – December, January, and February. This is most likely due to the snowy weather and bad runway conditions that normally cause delays. In, addition, June seems to have a good amount of delay for airports in the Midwest. We hypothesized that this could be due to the increase in travel over the summer. As for days of the week, Monday seems to be the day with the most delay. This is because Mondays are the most popular day for business travel.

Attached R code

```
library(sqldf)

setwd("/Users/yezylim/Documents/CS3300/project2")


data <- read.csv("flight_frequencies.csv")

data <- data[, !colnames(data) %in% c('ORIGIN_COUNTRY','DEST_COUNTRY',
'X','ORIGIN_STATE_NM','DEST_STATE_NM', 'YEAR')]

data <- data[complete.cases(data), ]


#data1 <- data[, !colnames(data) %in% c('ORIGIN_STATE_NM','DEST_STATE_NM', 'YEAR')]


data1 <- sqldf("SELECT data.ORIGIN_STATE_ABR, data.DEST_STATE_ABR, SUM(data.PASSENGERS) AS
numPeople FROM data GROUP BY data.ORIGIN_STATE_ABR, data.DEST_STATE_ABR")

write.csv(data1, file = "final_data.csv")


#q2 data

q2 <- read.csv("q2.csv")

q2 <- q2[, colnames(q2) %in% c('ORIGIN_STATE_ABR','DEST_STATE_ABR', 'PASSENGERS')]

q2 <- q2[complete.cases(q2), ]

q2 <- sqldf("SELECT q2.ORIGIN_STATE_ABR, q2.DEST_STATE_ABR, SUM(q2.PASSENGERS) AS
numPeople FROM q2 GROUP BY q2.ORIGIN_STATE_ABR, q2.DEST_STATE_ABR")


#q3 data

q3 <- read.csv("q3.csv")

q3 <- q3[, colnames(q3) %in% c('ORIGIN_STATE_ABR','DEST_STATE_ABR', 'PASSENGERS')]

q3 <- q3[complete.cases(q3), ]

q3 <- sqldf("SELECT q3.ORIGIN_STATE_ABR, q3.DEST_STATE_ABR, SUM(q3.PASSENGERS) AS
numPeople FROM q3 GROUP BY q3.ORIGIN_STATE_ABR, q3.DEST_STATE_ABR")


#q4 data
```

```
q4 <- read.csv("q4.csv")

q4 <- q4[, colnames(q4) %in% c('ORIGIN_STATE_ABR','DEST_STATE_ABR', 'PASSENGERS')]

q4 <- q4[complete.cases(q4), ]

q4 <- sqldf("SELECT q4.ORIGIN_STATE_ABR, q4.DEST_STATE_ABR, SUM(q4.PASSENGERS) AS
numPeople FROM q4 GROUP BY q4.ORIGIN_STATE_ABR, q4.DEST_STATE_ABR")



#combine data

totalFlightData <- rbind(data1, q2, q3, q4)



#get rid of dups

totalFlightData1 <- sqldf("SELECT totalFlightData.ORIGIN_STATE_ABR, totalFlightData.DEST_STATE_ABR,
sum(totalFlightData.numPeople) AS numPeople FROM totalFlightData GROUP BY
totalFlightData.ORIGIN_STATE_ABR, totalFlightData.DEST_STATE_ABR")



#get rid of same orig/dest

totalFlightData2 <- sqldf("SELECT totalFlightData1.ORIGIN_STATE_ABR as orig,
totalFlightData1.DEST_STATE_ABR as dest, totalFlightData1.numPeople AS numPeople FROM totalFlightData1
WHERE totalFlightData1.ORIGIN_STATE_ABR <> totalFlightData1.DEST_STATE_ABR")



#checked to make sure it got rid of the right rows

#totalFlightData2 <- sqldf("SELECT totalFlightData1.ORIGIN_STATE_ABR as orig,
totalFlightData1.DEST_STATE_ABR as dest, totalFlightData1.numPeople AS numPeople FROM totalFlightData1
WHERE totalFlightData1.ORIGIN_STATE_ABR = totalFlightData1.DEST_STATE_ABR")



#get rid of TT territories

totalFlightData3 <- sqldf("SELECT totalFlightData2.orig as orig, totalFlightData2.dest as dest,
totalFlightData2.numPeople AS numPeople FROM totalFlightData2 WHERE totalFlightData2.orig <> 'TT' AND
totalFlightData2.dest <> 'TT' ")

totalFlightData3_ext <- sqldf("SELECT totalFlightData1.ORIGIN_STATE_ABR as orig,
totalFlightData1.DEST_STATE_ABR as dest, totalFlightData1.numPeople AS numPeople FROM totalFlightData1
WHERE totalFlightData1.ORIGIN_STATE_ABR <> 'TT' AND totalFlightData1.DEST_STATE_ABR <> 'TT' ")



#write without same orig/dest
```

```
write.csv(totalFlightData3, file = "FINALDATA.csv")


#write with same orig/dest

write.csv(totalFlightData3_ext, file = "FINALDATA_ext.csv")


#too many rows with few flights. condense to flights with popularity > 0.001

data4 <- read.csv("FINALDATA_ext.csv")

data4 <- sqldf("SELECT data4.orig, data4.dest, data4.numPeople FROM data4 WHERE data4.percentage > .001")

write.csv(data4, file = "short_final_data.csv")


#find incoming and outgoing population

data <- read.csv("FINALDATA_ext.csv")

data <- data[, !colnames(data) %in% c('percentage')]

outpopData <- sqldf("SELECT data.orig AS state, SUM(numPeople) as outflow FROM data GROUP BY data.orig
ORDER BY data.orig DESC")

inpopData <- sqldf("SELECT data.dest AS state, SUM(numPeople) as inflow FROM data GROUP BY data.dest
ORDER BY data.orig DESC")

write.csv(outpopData, file = "outpopData.csv")

write.csv(inpopData, file = "inpopData.csv")
```