Advance Graphing in Airport data

Jiaqi Yao

Introduction

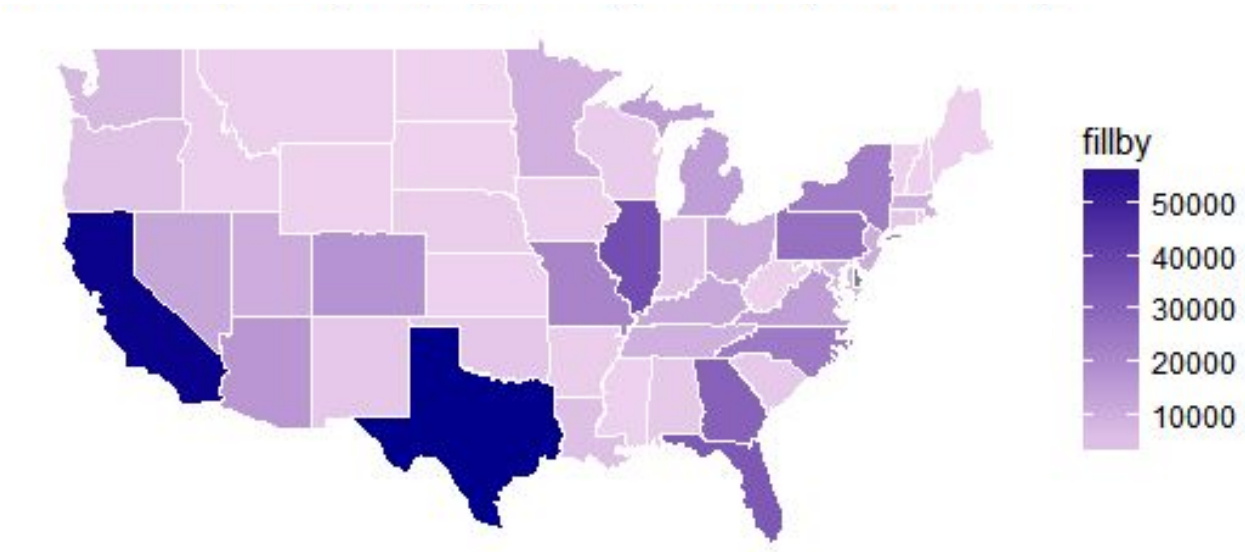
The main focus of this part is to have a combined understanding of association between delayed flights and its origin and destination airports. Departure delay, 'Depdelay', is one of the key variables in this section. We believe that departure delay would be more interesting to look at than arrival delay, based on the benefit of the early information gain. And we also pay more attention to origin airport than destination.

Data Preparation

Joining task is conducted in hive, where the 'orig' and 'dest' are used as id variable to match up with the 'idata' in airport.csv. Here we use the left outer join to ensure the completeness of the main data is reserved, even when the airport data is not inclusive. Since we are studying the delayed flight. Observation with positive 'depdelay' is discarded. During the joining process, we also keep track of the longitude and latitude information for map plotting.

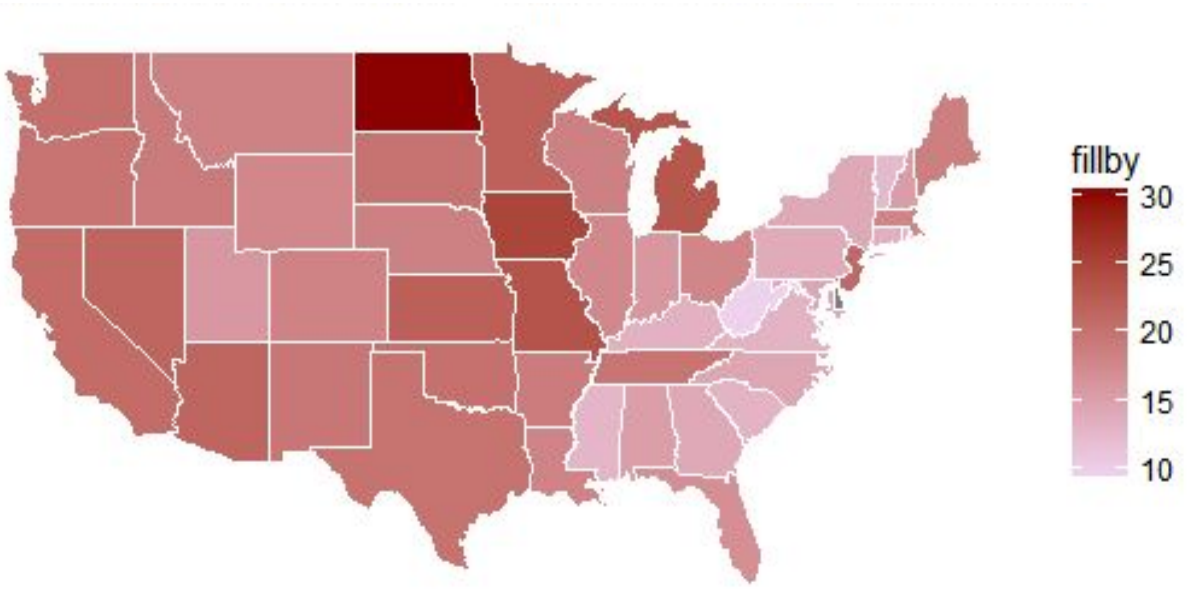
Results

a) Count of Delayed Flights per State during 1995 and 1996.



From this graph, we can find out the darkest region indicating a most frequent occurrence of flight delay is Texas, California and Illinois. The coastal area in U.S seem to have on average more delayed flights than the middle part. Colder northern part seems to have comparable more delays than the warmer southern region.

b) Average of Delayed time per Flight per Origin State



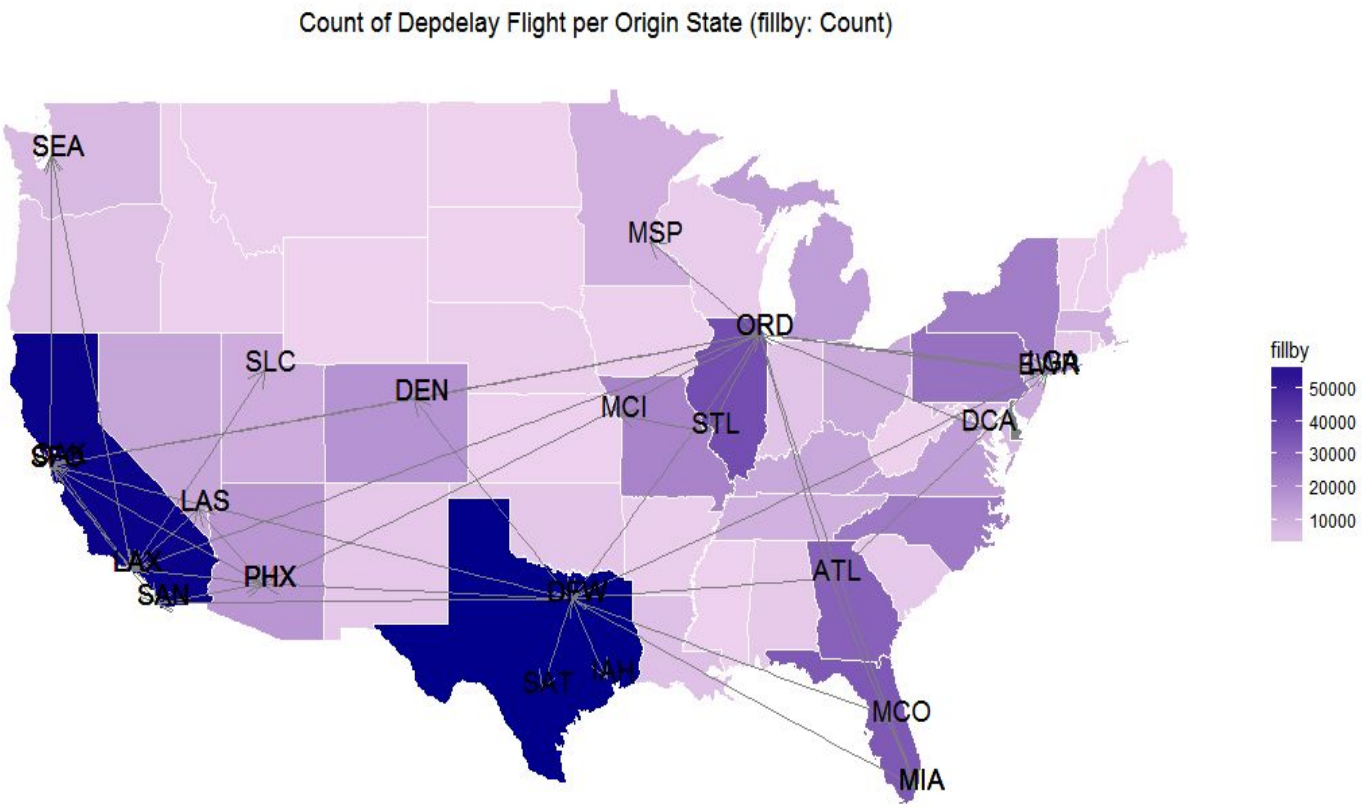
Besides the interest to know how frequent a delayed flight might happen, we probably also want to learn about how bad the delay would be. The result shown on this red map is telling a somewhat different story than the previous one.

As we can tell, the east coastal regions are having an average delayed time way better than the other parts. West coast states like Texas and California do not stand out as before. The middle part and the colder northern part however, does show a sign of a worse delay. Especially North Darkola, an average of over 10 minute longer wait is likely to be significant, given our large dataset.

Base on the previous two maps, we can roughly infer that if we travel from middle part of U.S, it's not likely to have a delayed flight, but once we have it, treat yourself to coffee maybe. If we are traveling from coastal area, we probably do not need to bother too much with the delays.

However, you might still want to avoid the following routes.

c) The most frequent delayed Flight Route



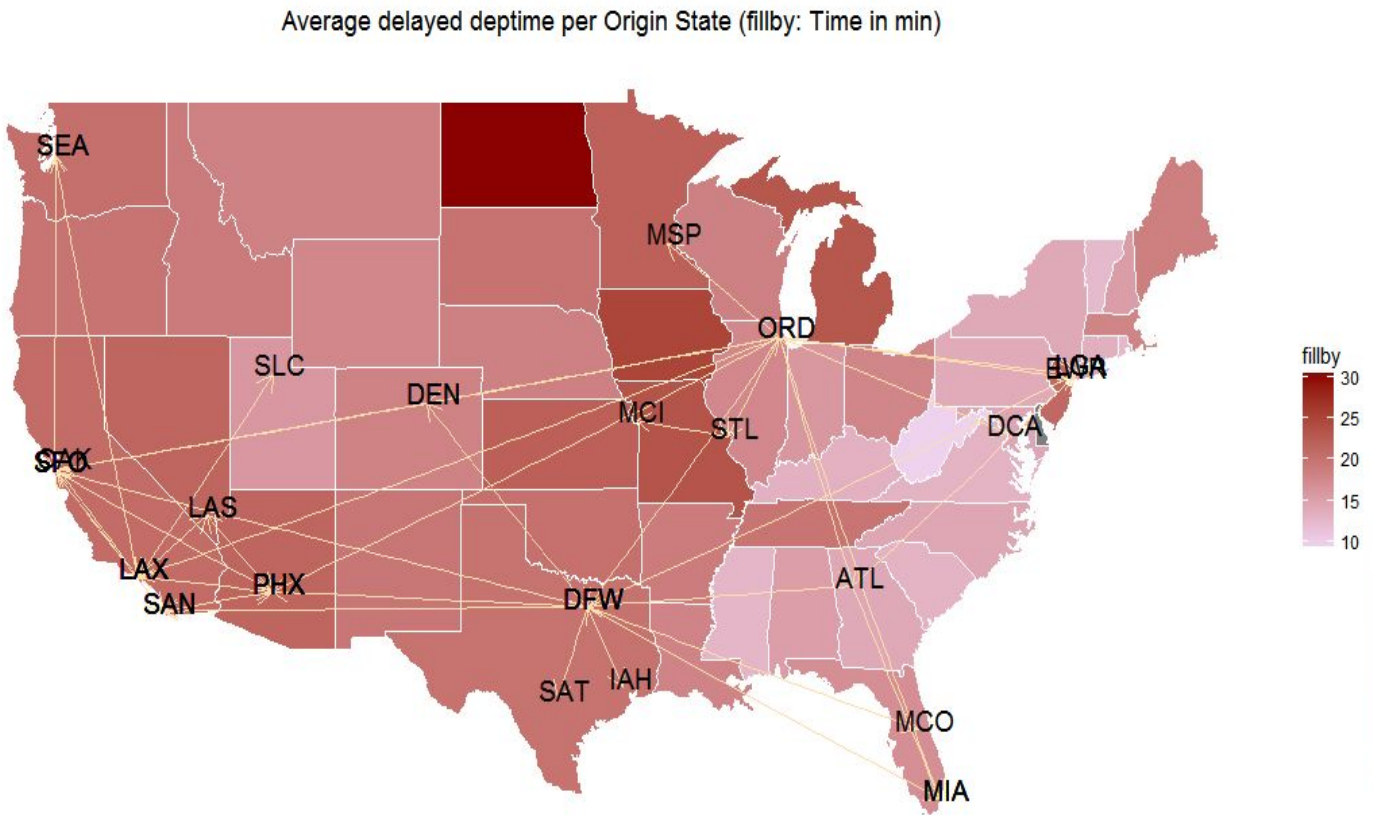
Utilizing the 'groupby' in hive, we obtained the delayed flight counts and average delayed time per route (direction does matter and is represented by arrow in map). Most of the routes are fairly long, thus those from ORD to LAX, DFW to MIA. A few short line route to take notice of is DFW to SAT/IAH. Thus we conclude that longer flight might seems to delayed more easily.

d) The worst delayed time Flight Route

Generally speaking, the line shape in the red map is identical to the lineshape in the blue map. All of these interesting discover in previous plot are still there. But they are different if we look at the table below.

Most Frequent Delayed Flight Route (Origin -- Destination)	Worse Delayed Time Flight Route (Origin -- Destination)
LAX SFO LAX PHX	PHX SFO SFO SEA

LAS LAX SFO LAX ORD LGA	LAX LAS LAX PHX DFW SFO



We might be able to conclude that the most frequent flight route is usually consistent with the worst delayed route. This is not true when we are flying one state to another when the origin and destination airport is not specified.