CS 458-558 Project 3

Most Popular Airports for Domestic Flights in the US from Chicago, Texas in 1997

Milan Thakkar, Conner Rhea, Grace Todd

For this project, we continued with our topic of US domestic flights from the year 1997. Due to the insufficient and incomplete data from our first project, we will be continuing to evaluate our dataset from project 2, which focuses on US domestic flight departures and arrivals. This topic is especially relevant in terms of location information because the data itself is reliant on the location of the airport, which airports a specific location services to, etc. In project 2, we speculated that the location of the airport is crucial to its number of serviced flights to other airports, specifically with regards to adjacent populous cities and relevant US monuments. We were able to come to these conclusions much easier with the visualization of airports on a map of the United States, where we could make connections between the most and least popular airports based on where they are located on the map.

For the purposes of this project, we have implemented several new methods of visualization in addition to visualizations from previous projects. We focus this report on how these visualizations were created, as well as a few potential directions for future visualizations of this data set.

Visualizing Geospatial Data

The following techniques were described in an article from [the humans of data](). These techniques provided a valuable insight into what kind of information our topic uses, and what meaning we can derive based on each visualization method. Some techniques were more applicable than others, for reasons explained below.

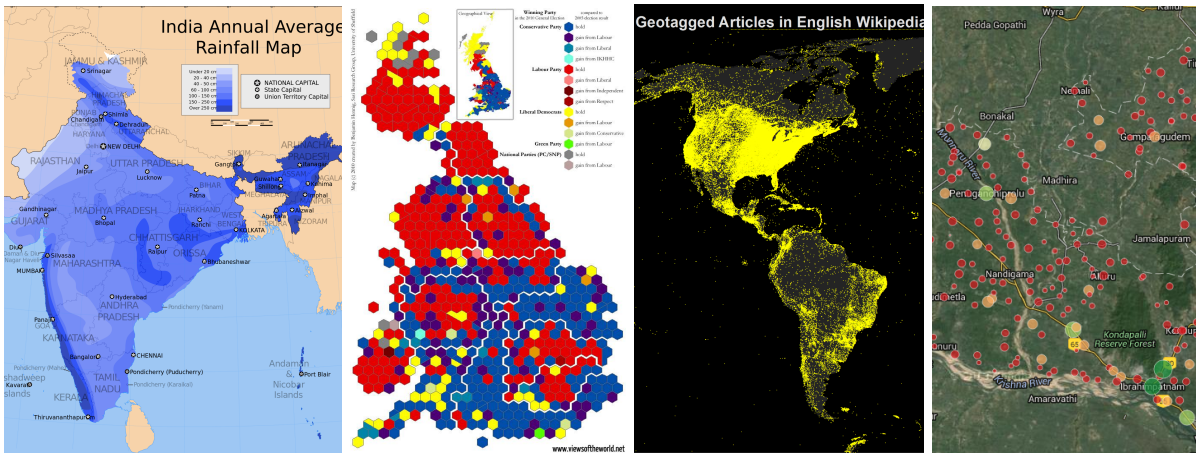


Figure 1 (from left to right): Examples of Heat map, Hexagonal binning, Dot map, and Cluster map from Humans of Data.

The first technique described in this article is the **Choropleth Map.** The Choropleth Map uses different shading patterns for different regions. If we were to do this with the different counties of the US for our data, it might be too specific to implement. Since counties are relatively small, it is not realistic to have data for every single county, since not every county in the US has its own airport. It might be interesting to see a Choropleth Map of each of the states, but we would have to average the popularity of each of the airports for a specific state, which might hide important information about the popularity of airports in each state.

The second technique described is the **Heat Map.** This technique is similar to Choropleth, but does not adhere to specific geospatial boundaries. This would probably be a step up for our dataset and could be interesting to see, especially considering the popular airports with regards to other nearby airports. It might be easier to visualize specific areas without needing to consider the boundaries of states or counties, since it is possible that airports in certain states also take in people from adjacent states.

The concept of **Hexagonal Binning** is particularly interesting, since it involves creating a grid of regular hexagons in the shape of the geographical region. It would be interesting to apply this to a map of the United States, but it might be tricky to apply this with such a consistent value as the number of flights in and out from each airport. This visualization method seems to work best for data which has a number of different data results (i.e. winning party when compared to results from a different year). We could apply this visualization using a range for each region (i.e. 0-500 service flights, 501-1000, 1001-2000, etc.), but this might not be a complex enough data set to do justice to this technique.

The **Dot map** visualization is familiar for us, since we essentially implemented this in project 2. We learned through visualizing this technique that depending on the size of the dots, the actual shape of the map might not even be visible. Because of the way that cities are located throughout the United States, there are large sections of the country that do not have any nearby airports. This could be a significant visualization of airport location in 1997, but the same data could be visualized while also deriving more relevant conclusions.

In a similar fashion, we also visualized **Cluster Maps** in project 2. We found this visualization to be especially useful, because it not only provided context to the location of each of the airports, but also the relevance of each airport in terms of number of serviceable flights. However, this visualization has a downside of neglecting less-popular airports and making them difficult to distinguish, which may make it easier to identify the most popular airports, but the less popular airports provide a valuable context to the importance of airport location. For instance, airports in the Alaskan wilderness are not distinguishable in this cluster map, where their location in context to the general US population has a credible impact on the number of their domestic serviceable flights.

Finally, a **Cartogram Map** visualization, where mapping variables are shown in a diagrammatic format, would likely be more relevant for a dataset with multiple variables, perhaps from a survey or something similar. With a dataset that purely focuses on a single variable– number of domestic serviceable flights in the US per airport– this visualization would likely not be as valuable.

Using Infogram

Out of the various programs listed on the Springfield article provided with the assignment, our team chose Infogram. We were drawn to the sleekness, the ease of use, and the drag-and-drop adaptability.

We used this program to create a multiview infographic about our dataset. While using Infogram, we noticed that even for the free version of this program, the features are expansive. There are a variety of graphics, shapes, icons, and much more that can be used for free, which are professional and sleek. The premium features for this program were intriguing as well; although the free features are high quality, the paid features of Infogram are even more so. One of the greatest disadvantages that we found to the free features is that the interactive interface of our infographic is not available with a free trial. For example, in the scatter plot provided in the infographic, a user should be able to hover over a point on the plot and see its airport name and the number of arriving and departing flights, but this is not an exportable feature for the free version.

Another disadvantage to Infogram is that working with data within the program is not ideal. While the program allows for easy import of Google spreadsheets and general CSV data, the data itself is difficult to apply to different visualizations within the program. It is easier to create graphics with this program by copying and pasting data into the provided spreadsheets for each chart instead of trying to import rows of data. This is a fairly trivial disadvantage, however, considering that we were able to work with hundreds of rows of data despite this minor setback.

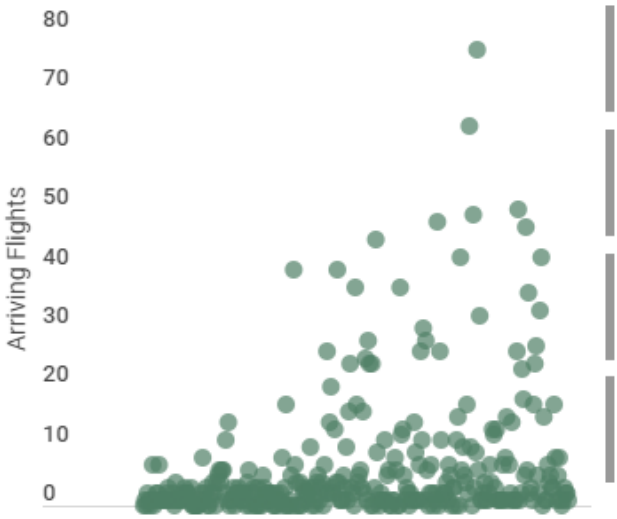
On top of providing a clean user interface for data visualization, Infogram provides a variety of templates and styles for displaying data. Although we built our infographic from scratch, there were a significant number of examples from which we drew inspiration.

Overall, we agree that Infogram was invaluable in developing a more comprehensive view of our data. Displaying our data in a colorful, all-encompassing way is much more engaging than a simple scatter plot with a paragraph-long caption. Through this infographic we are able to portray our data in a more meaningful and interesting way. While we were not able to display all of our data through this technique, we agree that displaying relevant data can be more useful than displaying *all* of the data. The final infographic using Infogram can be seen below:

Most Popular Airports for Domestic Flights in the US

1997

Grace Todd, Milan Thakkar, Conner Rhea



Chicago O'Hare had the greatest number of serviceable flights (80 arriving).

Up to 78 airports across the US had reported only 1 serviceable domestic flight.

323 airports across the US were surveyed for this data set.

323 airports across the US were surveyed for this data set.

Figure 1: Number of arriving flights per airport in the US, 1997.

?

Uncertainty

Considering the information-sharing limitations of 1997 with respect to modern methods, there is a high chance that this dataset is incomplete.

Most Popular

Figure 2: While there were a number of airports in service during the year of 1997, only a select few serviced more than 45 domestic flights. Airports with the greatest number of serviceable flights were Chicago O'Hare and Dallas/Fort-Worth, with others residing in America's most populated cities.

1

Chicago O'Hare International Airport
80 arriving flights

2

Dallas/Forth Worth International
77 departing flights

3

William B. Hartsfield Atlanta International
64 arriving flights

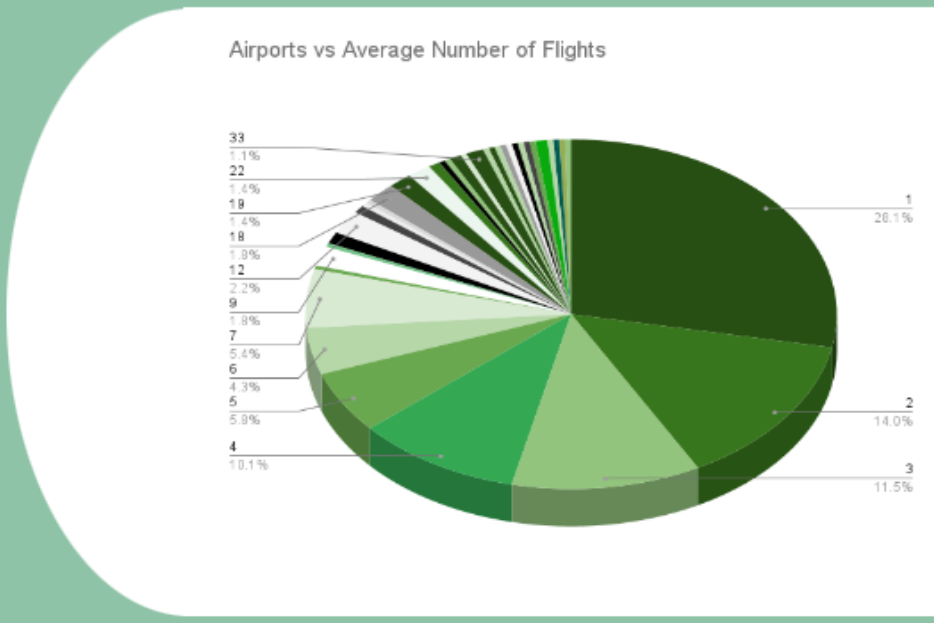
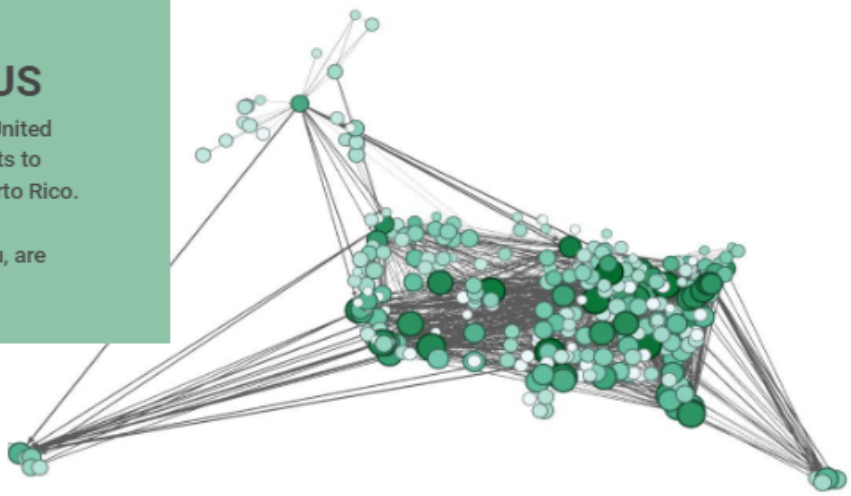


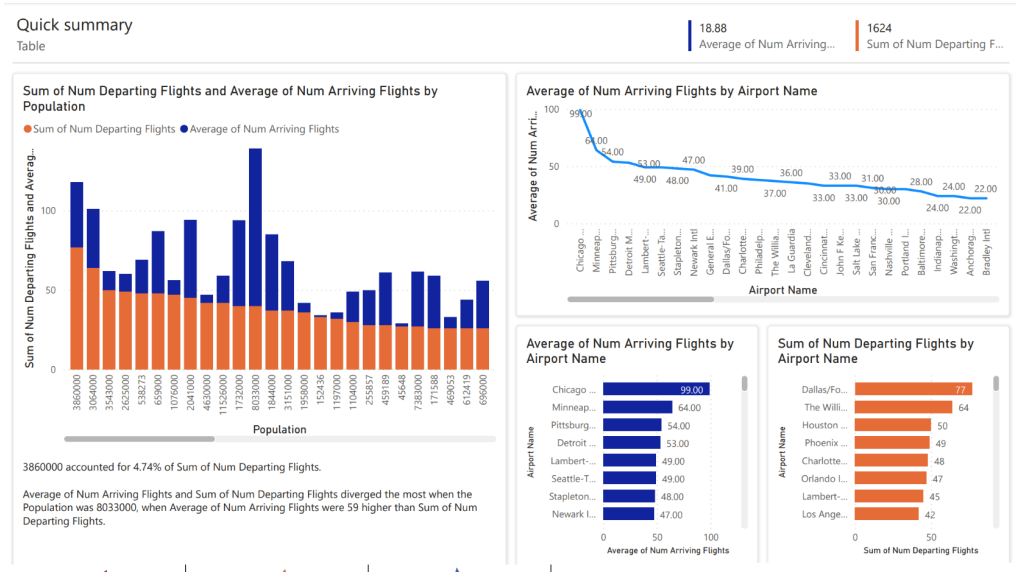
Figure 3: The majority of airports in the US averaged less than 4 serviceable domestic flights in the year 1997.

Beyond the US

Domestic flights in the United States also include flights to Hawaii, Alaska, and Puerto Rico. Some airports, such as Anchorage and Honolulu, are particularly prevalent.

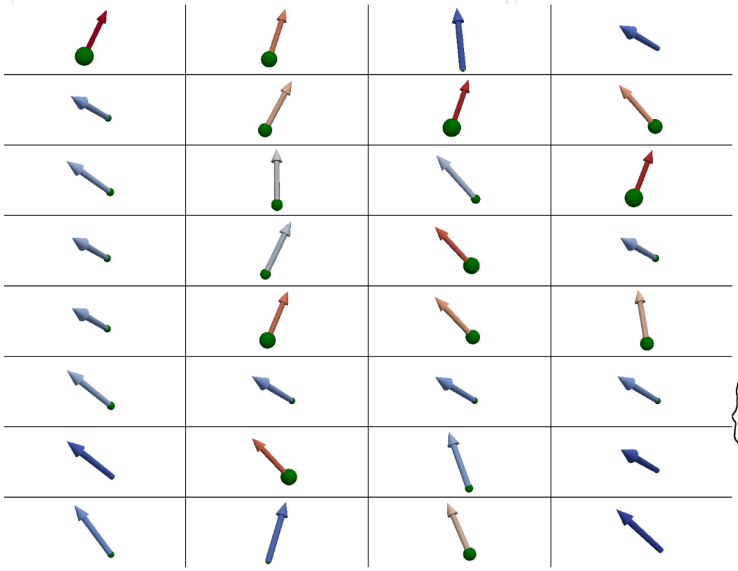


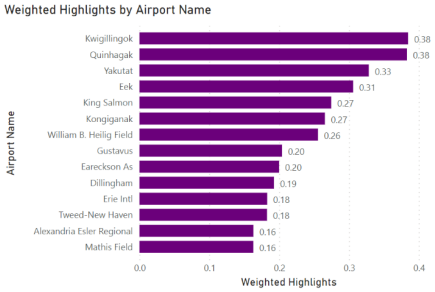
Multi-View Visualization System

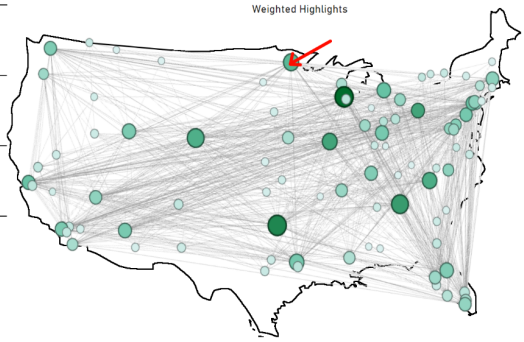


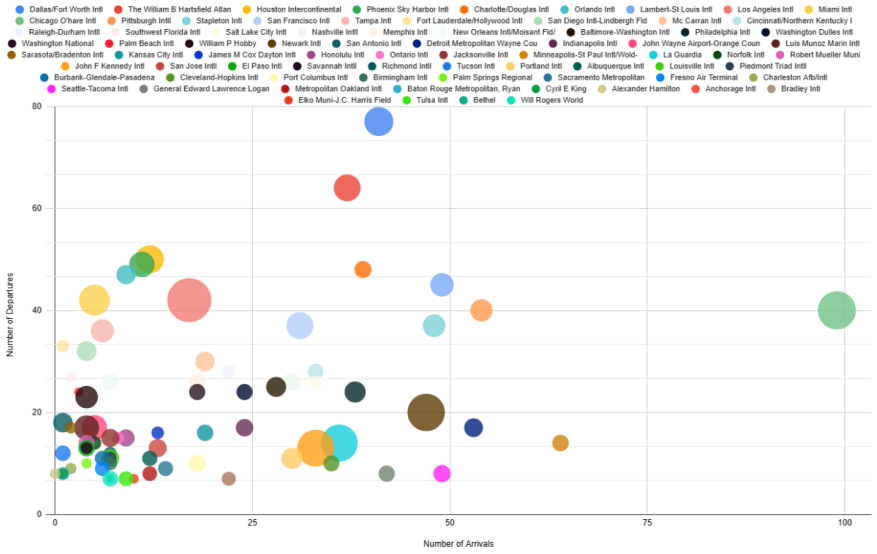
Summary Page: Select Map Location For Highlights

Next Page



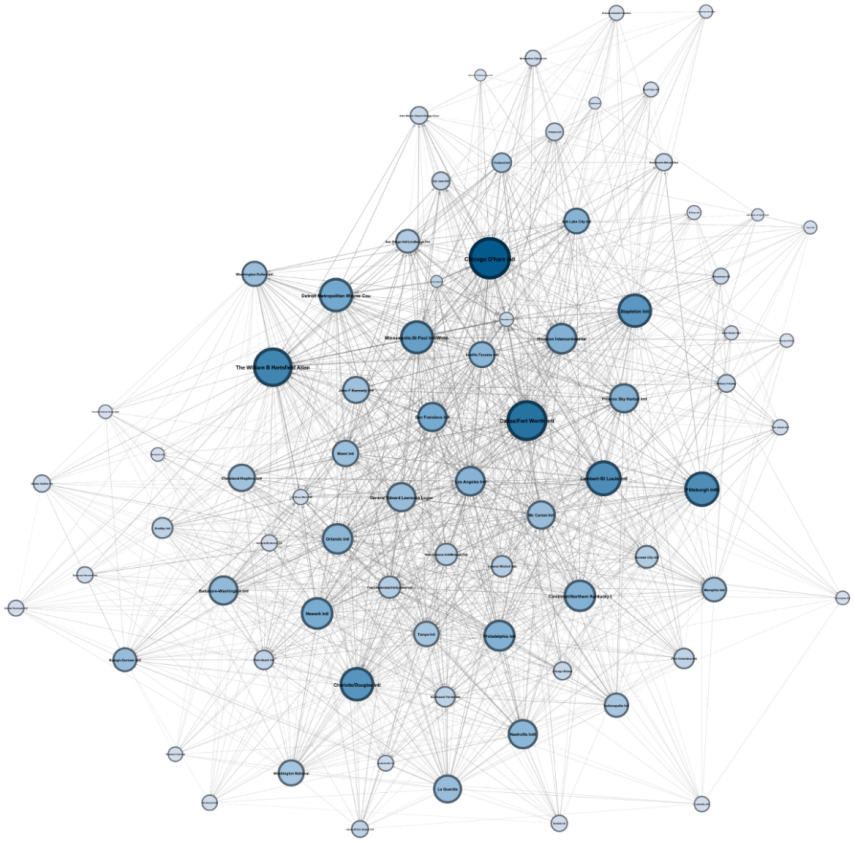






Overview Page: Data for all map locations

Prev Page



Descriptions of Motivations

This Data aims to examine the complex relationship between Airport Usage and Population within a subset of the United States reported data on flight within the country, excluding foreign flights. Through this data we can come to the conclusion that population size does have at least some effect on airport traffic. As population size increases so does the number of incoming and outgoing flights to regional airports. This makes logical sense as airports were likely built in major population hubs first and foremost so that correlation makes sense, however this is also an element that doesn't care about population size which adds uncertainty to the data, which is stopovers. Some smaller airports may end up being good layover points for planes to land over the course of a long journey, meaning that even a smaller airport may still register due to regional location. Take Raleigh-Durham International Airport for example in North Carolina (one of the medium nodes in the Gephi Graph), this node is fairly big despite North Carolina not having a particularly large population compared to most of the big states, however this destination serves as a great entry point before reaching into and out of the East coast, meaning that it still sees a larger number of flights compared to other smaller airports such as Tampa, Florida however it is close to a location such as Palm Beach, which is a know tourist destination despite low population in the surrounding area.

With this we can see that while population does account for the majority of airport hubs within the data set, external factors can still have an effect, meaning that there is still important information to be gleaned from this data set regarding the effect of population on air travel within the US, and what contributes to it.

With our Multi-View Visualization we choose a number of different data graph types to represent the effect of population on our data set, but also to highlight some of the uncertainty inherent to the problem. For example, the Quick Summary Bar chart does a good job showing the effects of population on air travel, making it easy to see at a glance that high population numbers correlate to higher flights, however this also shows that some airports for example have way more flights departing that arriving, which also shows the uncertainty of the system when an imbalance such as that is present.

The two graphs nearby also paint an interesting picture, with the Average Number of Arriving and Departing Flights having different Airports in the top 9 results only sharing 1 of them, being Lambert. This shows that airports can be big for different

reasons as well, as both Chicago and Dallas/Fort Worth are number one airports in their respective category, and both airports are large nodes within the network map graph generated from Gephi.

Looking at the Gephi adjacency chart we created you can also see how the large population airports dominate the network, acting as major hubs for the entire network overall. Through this we can also see many of the major connections between some of the minor airports and major airports, combined with the new population data puts into perspective the major effect that population density in the area does have for an airport.

Additionally we have a highlight for a subset of the data, and its weighted edges for a general area with the purple bar graph, in our case, Alaska. This provides an interesting at a glance view into the details of a selection of the data, allowing us to quickly draw conclusions about this chunk of our data and how it correlates to the bigger picture. Our Glyph provides a look into the flights out of the selected airport, with the degree corresponding to the location being traveled to, and the ball attached providing information on the number of flights between the two airports in question.

The bubble graph was an apt visualization to demonstrate the relationship between population and the amount of arrivals and departures from each airport, as we expected the larger population areas to service more flights in and out of the airport, it would be easy to see what the relationship was.

On the bubble scatter plot, we are able to see that, as a common trend, the airports that service smaller population areas were generally in the lower left corner of the graph, indicating that population shows a general correlation with the number of flights. On the other hand, there were airports that stood out as either having a low population relative to a higher than expected number of flights. Additionally, LAX stood as an outlier in having a very large population for a relatively low number of flights. While its clear population has a strong correlation with the number of flights, it's also very apparent that there are other variables at play here.

Components and Layout

With this data set we are mainly looking for information regarding airport flights and correlating that with the population of the surrounding area. To this end our data contains Airport Names and ID Labels, number of incoming and outgoing flights,

population data around those airports and hubs, and edge weights corresponding to the number of flights total the airport handles.

Through this we analyzed the effect of population density on airport usage within the United States and how it correlates with flights in and out of that specific airport. Higher number of flights both in or out indicates an airport that is well used, and though lining up those datasets together we got a closer look into how much population density affects our nation's airports.

The population data that we acquired for the areas that the airports service is very helpful in showing a correlation between population and frequency of flights to and from an airport. However, in this data lies a great amount of uncertainty. For example, it does not account for the fact that some of the airports, such as Seattle-Tacoma service areas much beyond Seattle and Tacoma, as many people from northwest Washington use that airport outside of those areas. Conversely, Newark, La Guardia, and JFK international airports all service the same area, and may report fewer flights per population than the correlation would indicate. Another factor that may contribute to uncertainty is that some of the airports service vacation hotspots such as Palm Beach. Palm Beach has a relatively low population compared to William P. Hobby airport that services Houston, with a much larger population than Palm Beach. This also once again, does not account for flights in and out of the United States, so it's possible there may be an airport under-represented due to it handling primarily out of country flights compared to in country ones.

The data is separated into two pages, one page providing a summary and map of the data in question, and allowing the user to select a location to see highlights of the selected location, and one page providing a general overview of the important data points and how they stack up on the scale overall. Using these two window panes, users can quickly search for the data they want to see about US Air Travel, and then compare it to the others in the data set.

Demonstration

Since we did not implement the same dataset in project 1, we have recreated a few figures that might have been used for the first project had our secondary data set been used:

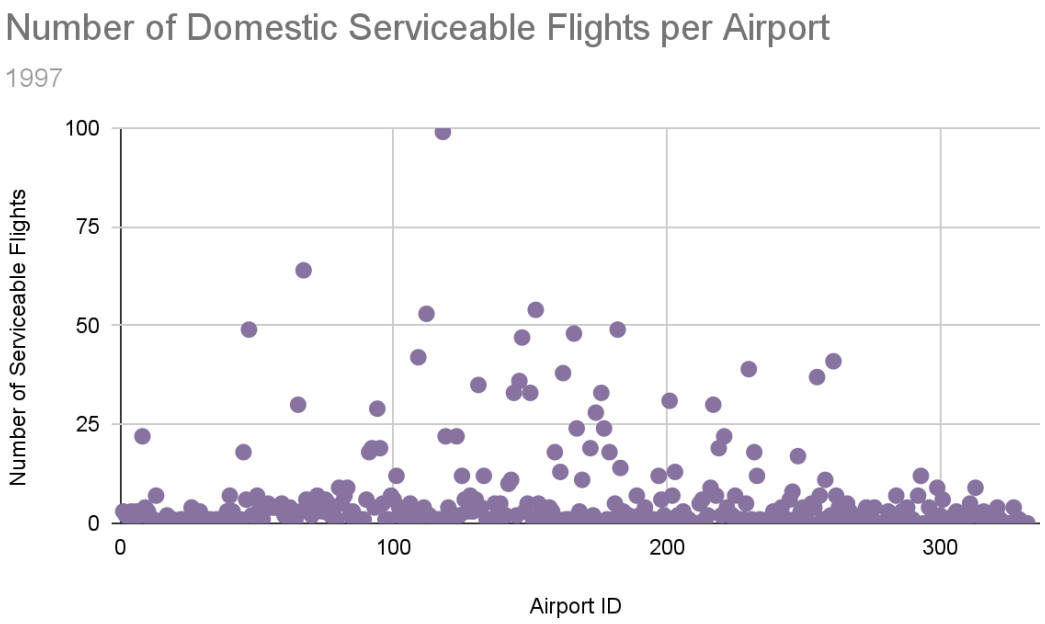


Figure 2: Number of domestic serviceable flights per airport. Number of flights calculated by adding up the number of edges sourced at each airport. The majority of airports report less than 25 domestic flights to other airports mentioned in the data set.

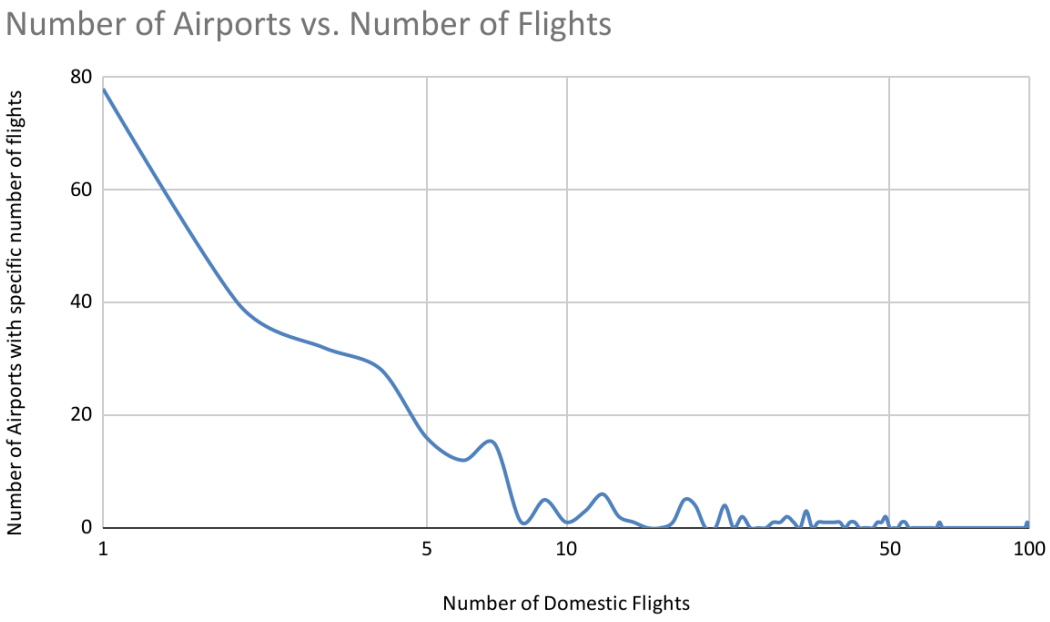


Figure 3: Average number of serviceable domestic flights per airport. Approximately 80 US airports service an average of one domestic flight, with approximately one airport servicing a maximum of 99 domestic flights (Chicago O'Hare Airport).

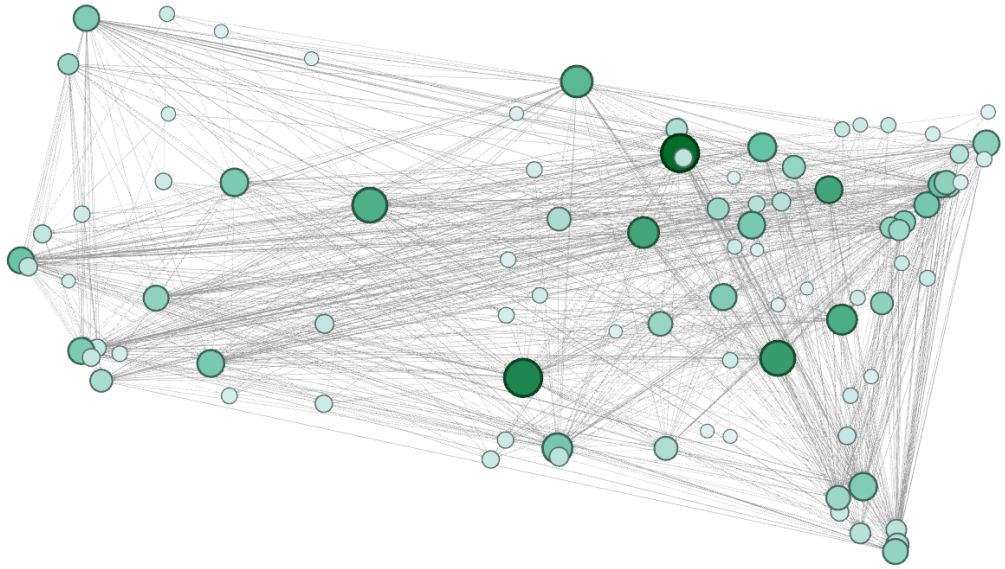


Figure 4: Map of airports with most domestic flights constrained to the continental United States. Airports with service to the most number of airports are larger and have a darker hue. Filters: Longitude, Edge weight.

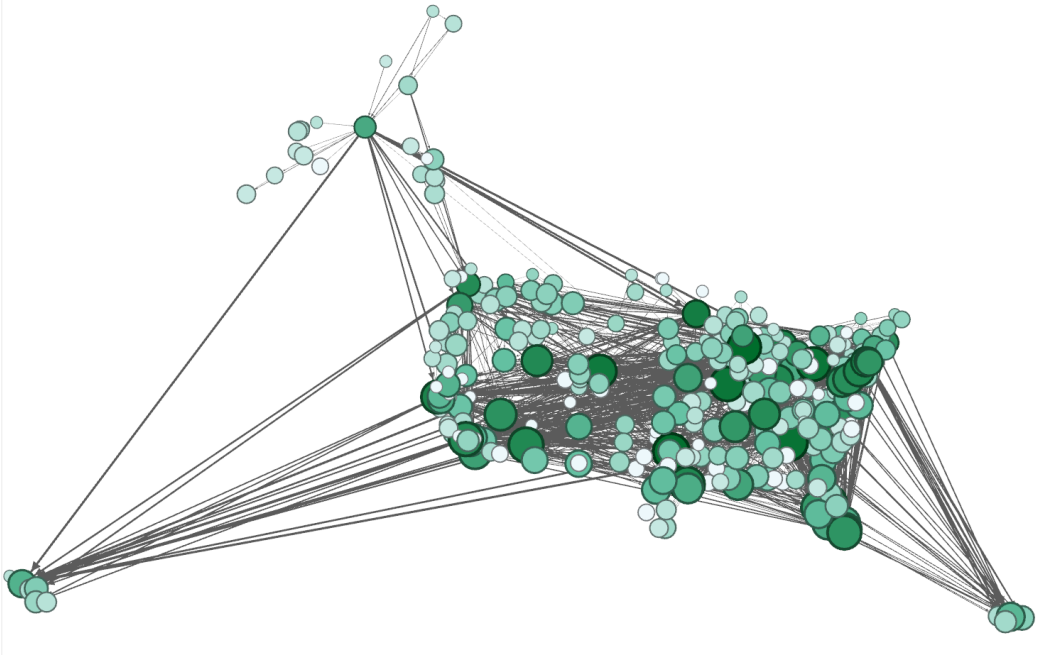


Figure 5: Map of airports with domestic flights to or from the United States with a constrained longitude, from Hawaii to Puerto Rico. Airports with higher frequency of service to other airports have a darker hue and a larger node size, with a heavier edge

thickness. Airport nodes have not been labeled to increase readability. Filters: Edge weight, range.

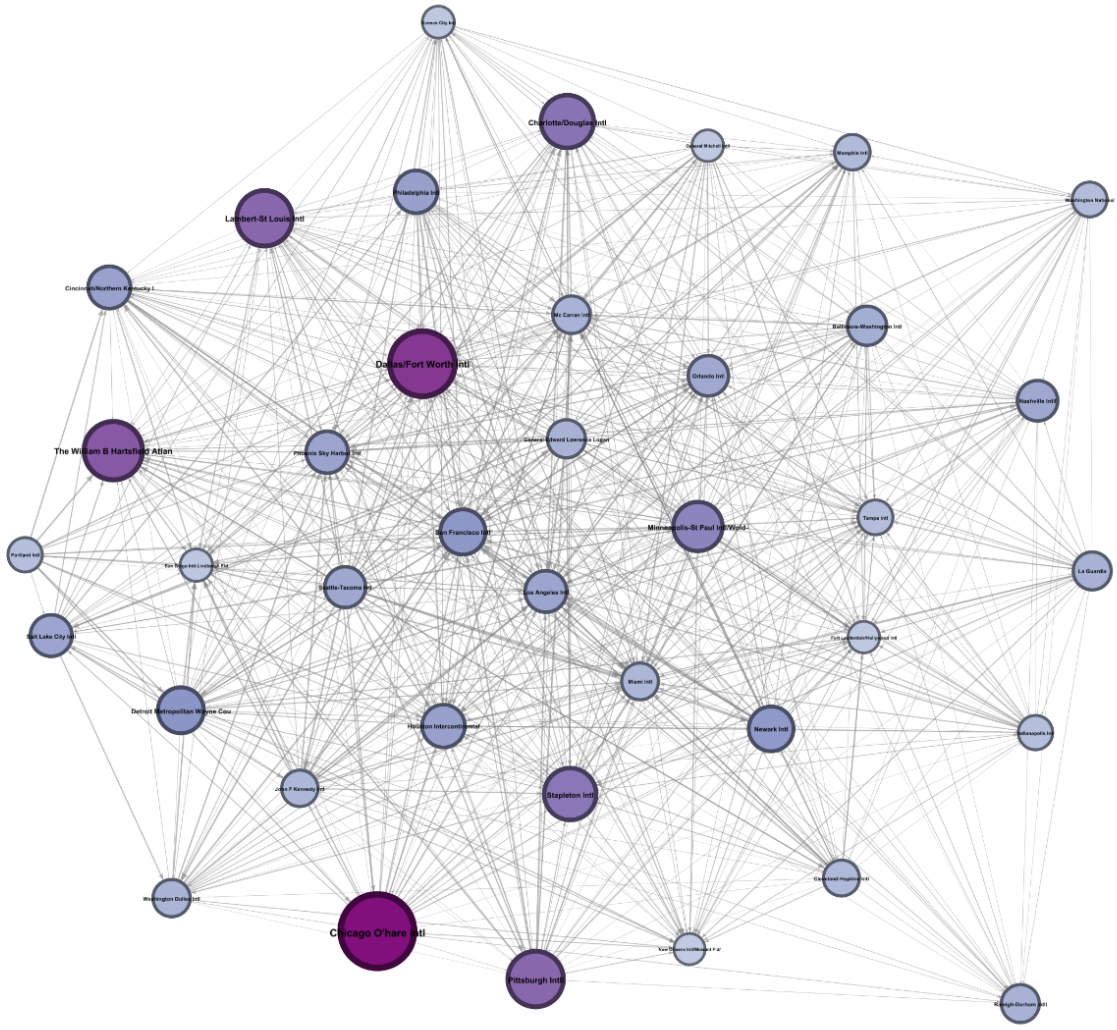


Figure 6: All airports in the United States with service to more than 30 other airports. Airports with service to the highest number of airports in bright purple. Edges connect each airport to its serviceable airports. Filter: Edge weight, degree range.

The following are glyph visualizations as depicted in project 2:

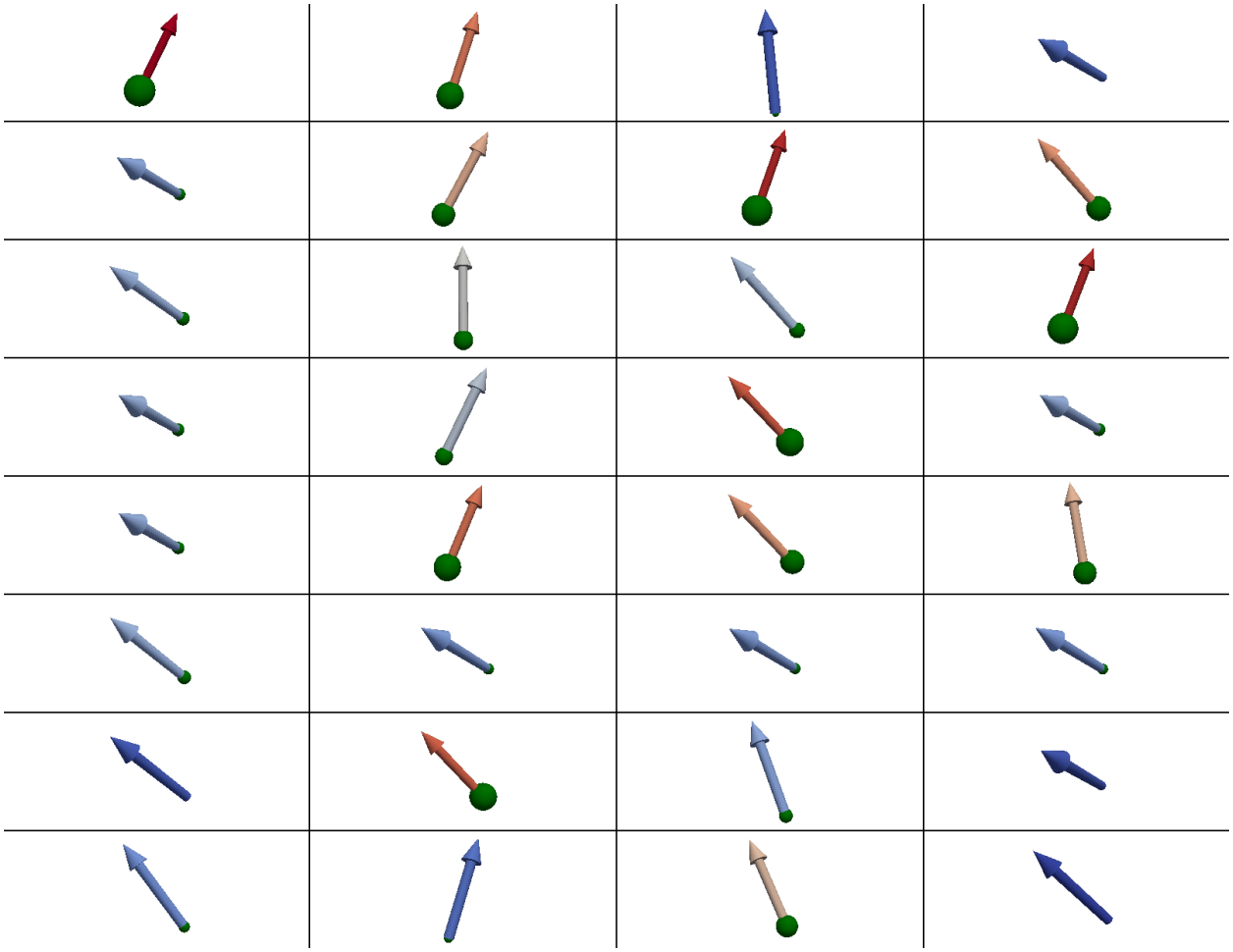


Figure 7: A subset of the Airport Dataset described through the use of a glyph. This Data Set specifically looks at the Chicago O'Hare International Airport, and all the significant edges (degree > 30) that connect to it. The Direction is determined through the vector of the longitude and latitude of the two airports, and the color and size of the green ball corresponds to the number of flights between the two airports, the ball being bigger for the larger number of flights between them.

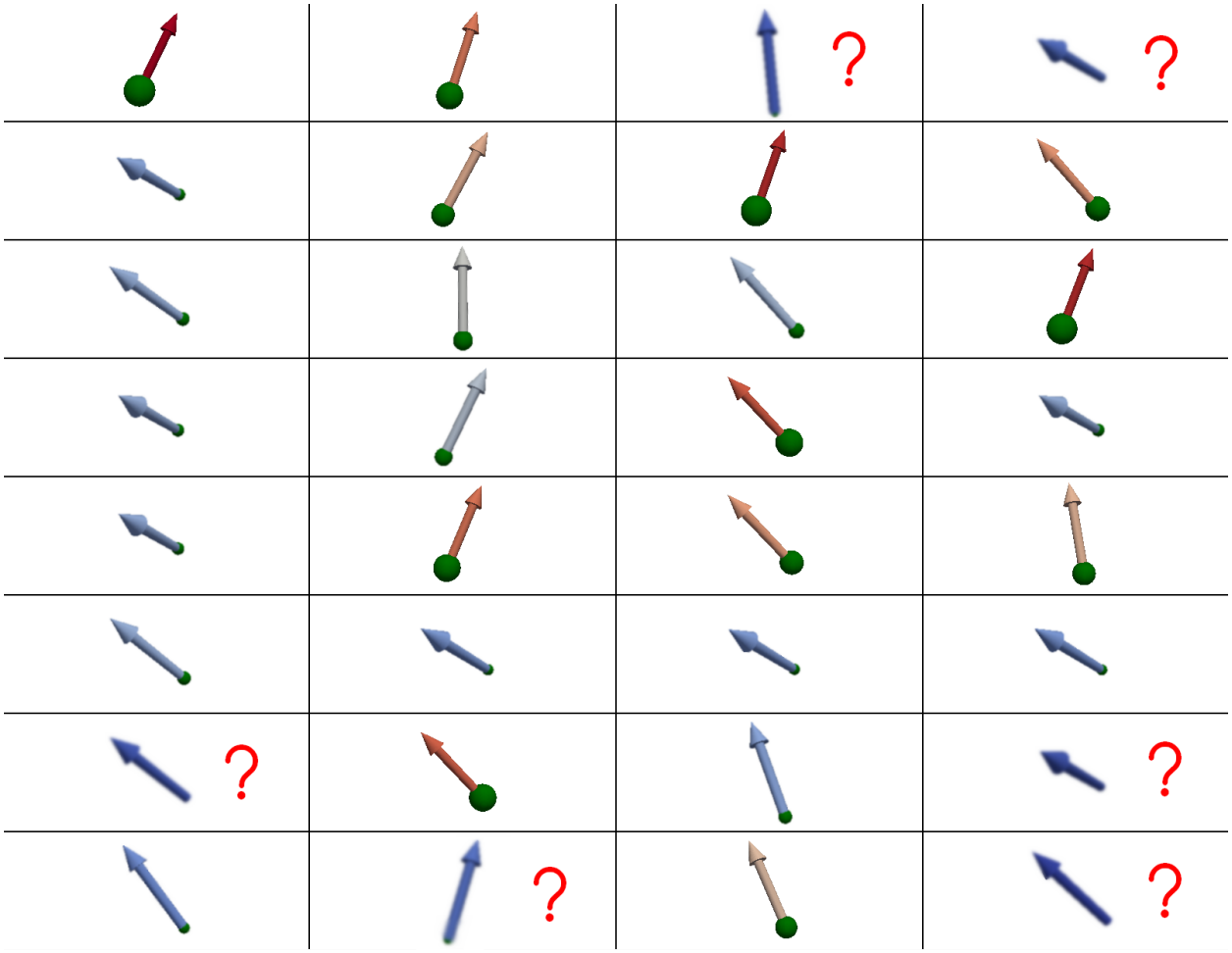


Figure 8: Glyphs with uncertainty added. Airports with smaller numbers of weights have increased blurriness to indicate the uncertainty of accurate representation of incoming and outgoing flights. Question marks denote ultimate blurriness and uncertainty in specific vectors.

The visualizations below are magnified versions of the graphics used in the multi-view visualization, for viewing convenience.

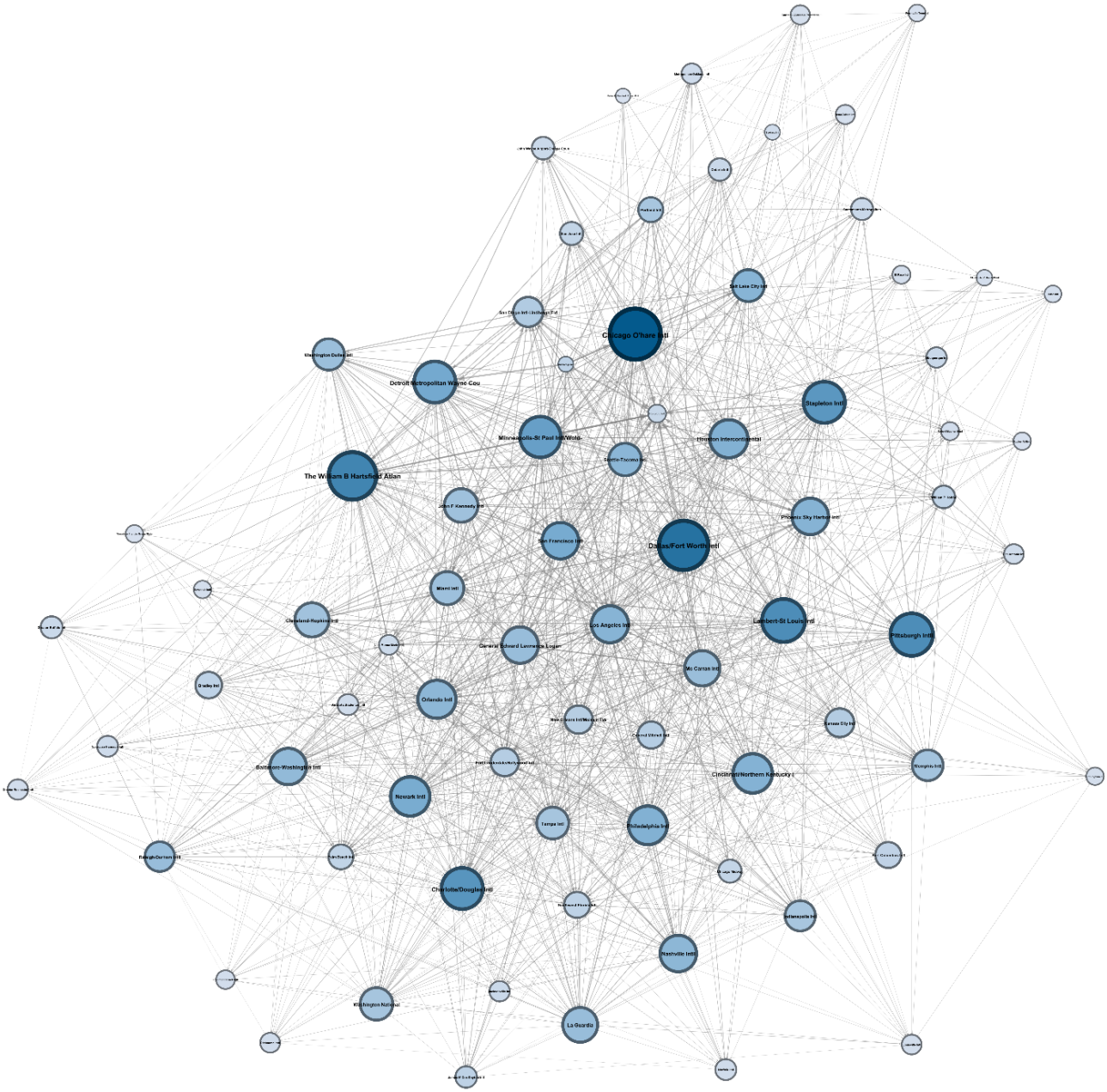


Figure 8: Updated network graph of airports as seen in the Multi-View visualization. The larger and darker the color of the circle, the more domestic serviceable flights that airport has.

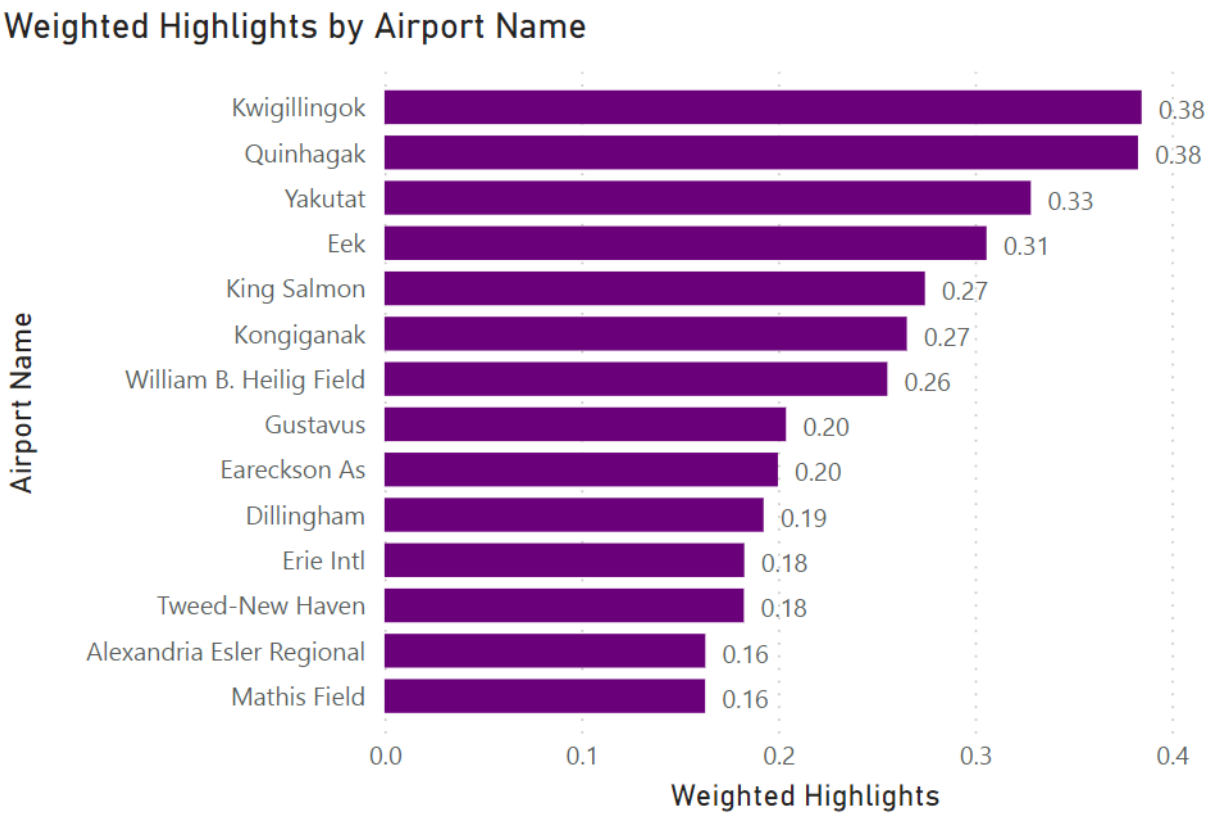


Figure 9: Weighted highlights for top 14 airports in data set, as seen in Multi-View visualization.

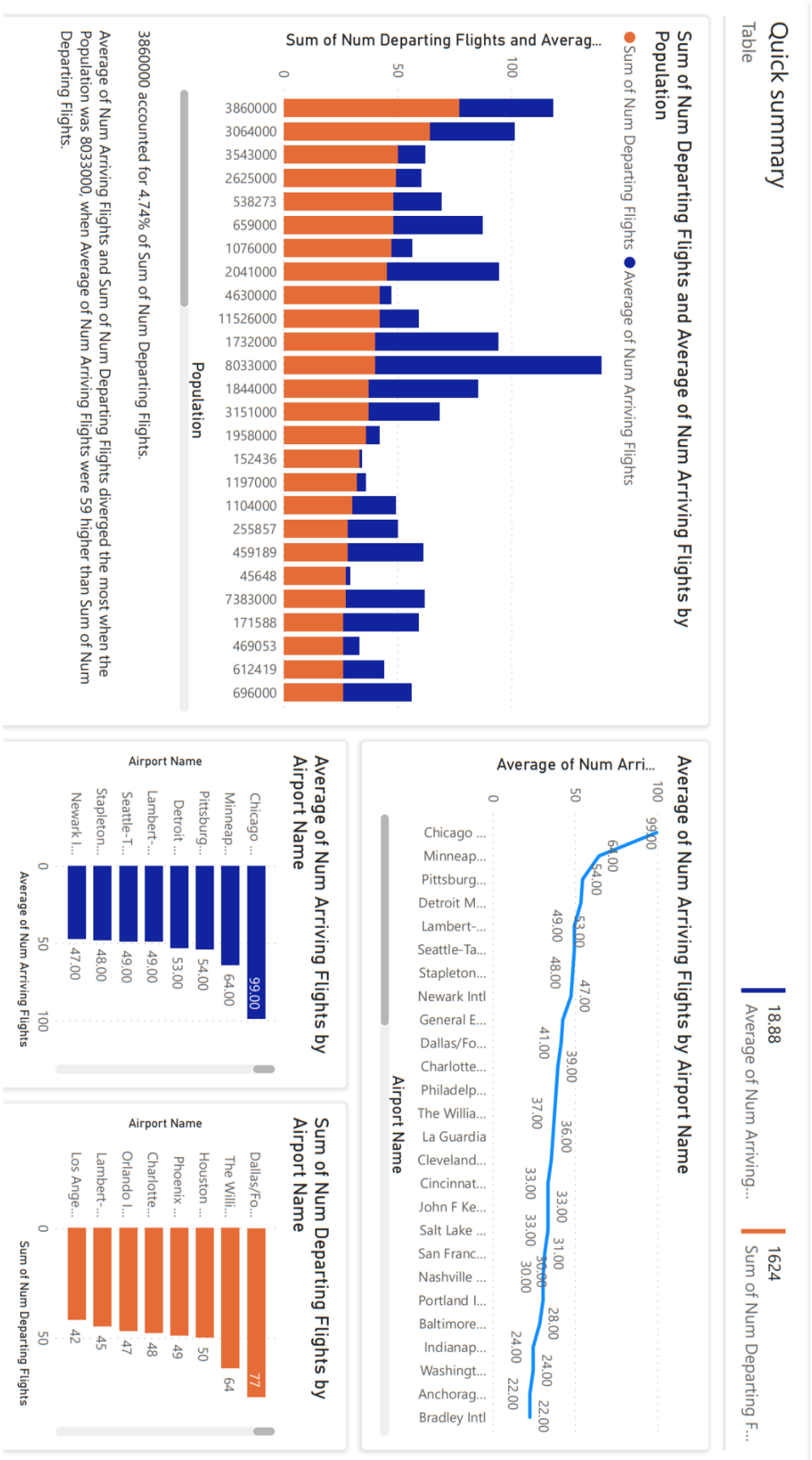


Figure 10: Screenshot of interactive visualization interface, as seen in multi-view visualization.

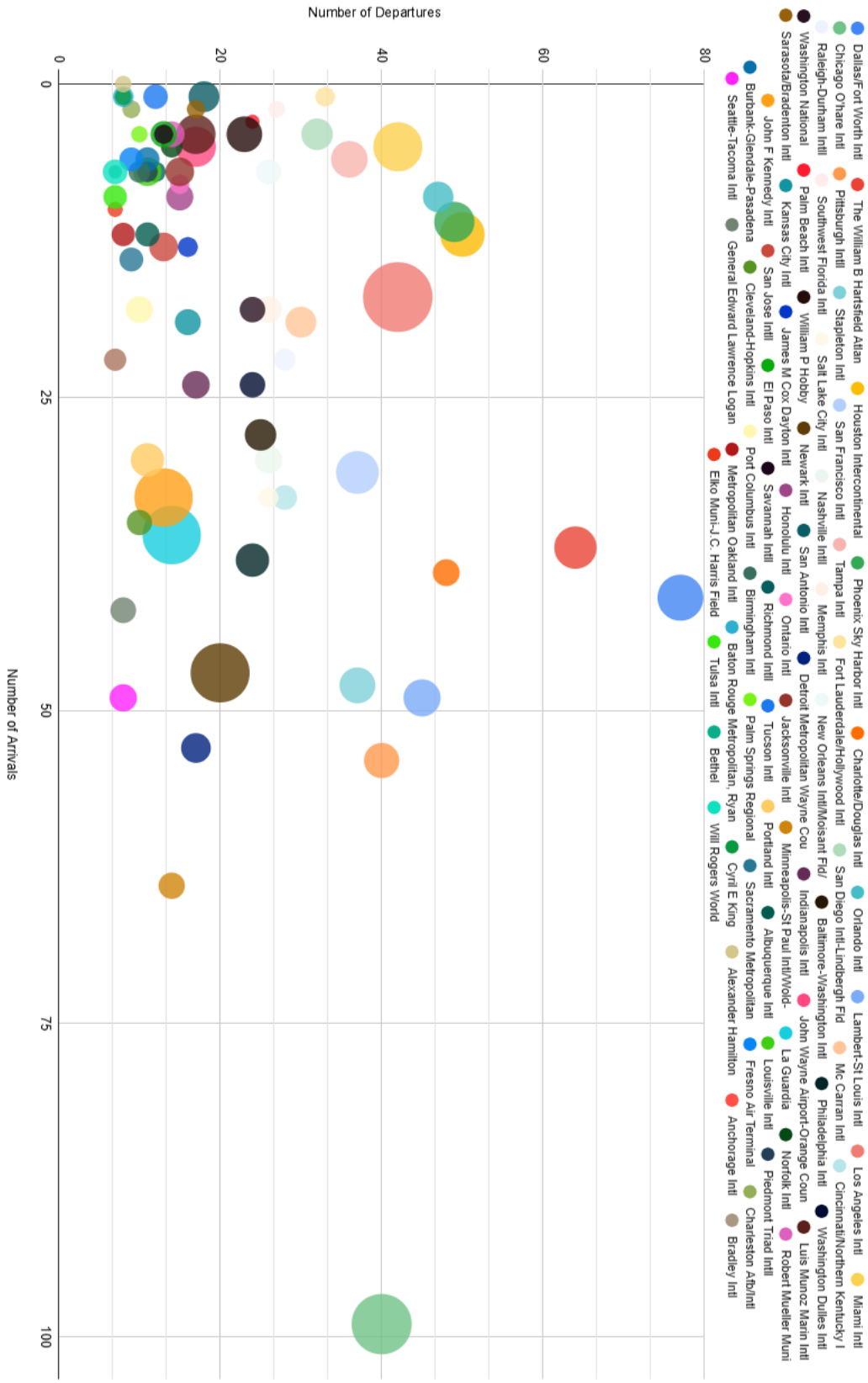


Figure 11: Bubble chart of serviceable flights per airport versus population surrounding the airport, as seen in the multi-view visualization

Design Principles

The following design principles were implemented to create the most effective visuals:

1. **Clarity:** To properly clarify our data visualizations, we have included captions for each figure describing what the visualization is depicting. In cases where the visualization is a graph, we provide axis labels, graph titles, and– where appropriate– legends to identify each airport.
2. **Simplicity:** Specifically with the Infogram infographic and our bubble plots, we aimed to keep our visualizations as simple as we could without obscuring the context of the data. To do this, we reduced the level of information displayed, such as the names of the airports in the bubble plots and reducing the overall frequency of text. Since our dataset includes more than 300 airports, it would be difficult for our visualizations to depict the values of every single airport without the visualization becoming crowded and unreadable. Therefore, we have kept most of our visualizations with a minimal amount of text and let the visualization speak for itself.
3. **Effective use of color:** Also with the Infogram infographic in mind, we used color to effectively and consistently communicate the information to the viewer. The design of this infographic aimed to direct the viewer's eye from one visualization to the next, all the way to the bottom of the page. The contrast between the green and the whitespace separates the large block into smaller sections, which are more digestible.
4. **Appropriateness:** To ensure that our visualizations appropriately communicate the context of our data, we relied mostly on geolocation-based visualizations. The subject matter of our data relies heavily on the interconnectedness of airports across the country, and so we were more easily able to visualize these relationships on a map. Where our visualizations relied more on quantities (i.e. comparing the number of arriving versus departing flights for each of the airports), we implemented bar charts and scatter plots.

With these design principles in mind, we were able to create more effective visualizations of domestic flights in the United States in 1997.