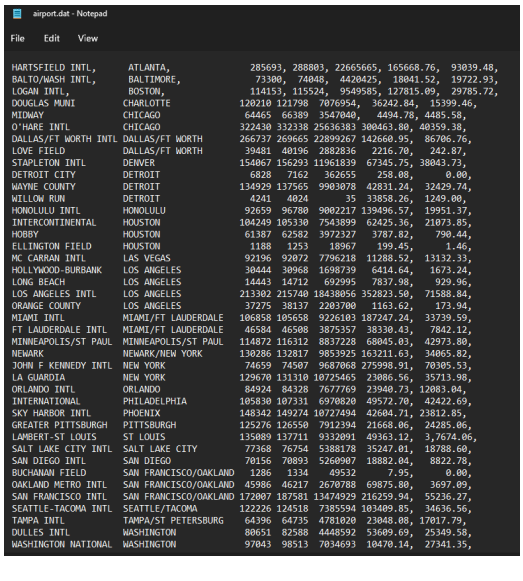
Quentin Hardison

My Information Visualization Report

Information visualization is defined as “the practice of representing data in a meaningful, visual way that users can interpret and easily comprehend” (*What is information visualization?*). The human body is able to detect patterns and trends in the data of a visualization and can be used to make decisions after interpretation of the data. These decisions are usually solutions to problems or areas that needed improvement identified in the visualization. A visualization should allow the viewers to gain insight and find the most important information from the data without the viewer using many resources. This project presentation was about presenting a data set and creating my own visualization.

To start with my information visualization, I had to first find an interesting dataset that I wished to create a visualization for. The dataset was required to have at least 30 to 1,000 points with at least 3 to 6 variables. For my dataset, I chose a dataset that contains the various United States of America’s airport statistics. In this dataset, it is measuring the data through 4 different variables: the *Total Amount of Scheduled Departures*, the *Total Amount of Departures Performed*, the *Total Amount of Passengers*, and the *Total Amount of Cargo Moved* in tons. This dataset includes data from each airport of each major city, however, in order to condense the information for the visualization, I grouped together all of the data for the same city. For example, Houston has 3 different airports: George Bush Intercontinental Airport (IAH), William P. Hobby Airport (HOU), and Ellington Airport (EFD). With this information, I grouped the data for Houston together for each variable, such as adding the *Total Amount of Passengers* for each airport and that became the full Total Amount of Passengers for the city of Houston.



After I went through each variable for each city, I had created my own sub-dataset using the full data set. In my dataset, each city had one total for each variable and generated a list of this information. This information was inserted in Excel, and I was presented with a number of ways of displaying the information through visualization. Creating a visualization is like the

method you would use to solve a problem. When approaching a new problem, you must first figure out what the problem is, which in this report is how to transfer a dataset into a my own visualization.

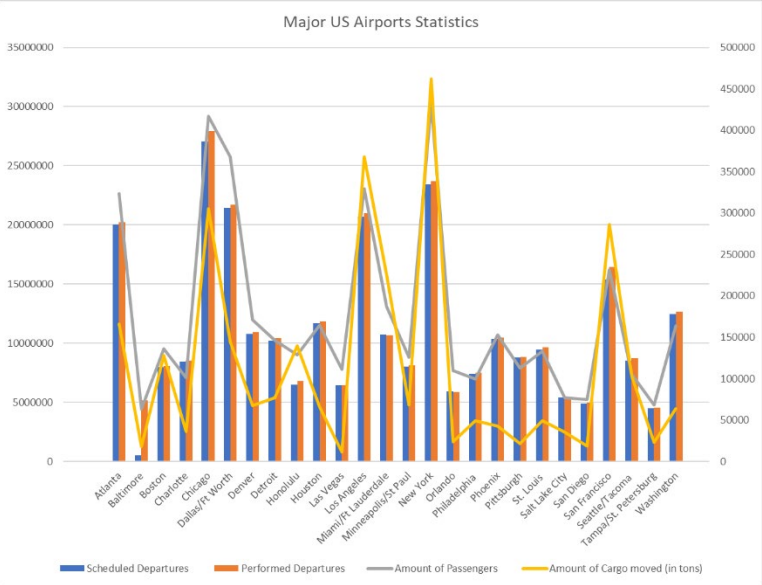
City	Scheduled Departures	Performed Departures	Amount of Passengers	Amount of Cargo moved (in tons)
Atlanta	285693	288803	22665665	165668.76
Baltimore	7330	74048	4420425	18041.52
Boston	114153	115524	9549585	127815.09
Charlotte	120210	121798	7076954	36242.84
Chicago	386895	398727	29183423	304958.58
Dallas/Ft Worth	306218	309861	25782103	144877.65
Denver	154067	156293	11961839	67345.75
Detroit	145998	148751	10265768	76947.58
Honolulu	92659	96780	9002217	139496.57
Houston	166824	169165	11535193	66412.63
Las Vegas	92196	92072	7796218	11288.52
Los Angeles	295464	299557	23033490	368239.74
Miami/Ft Lauderdale	153442	152166	13101460	225577.67
Minneapolis/St Paul	114872	116312	8837228	68045.03
New York	334615	338634	30266458	462297.1
Orlando	84924	84328	7677769	23940.73
Philadelphia	105830	107331	6970820	49572.7
Phoenix	148342	149274	10727494	42604.71
Pittsburgh	125276	126550	7912394	21668.06
St. Louis	135089	137711	9332091	49363.12
Salt Lake City	77368	76754	5388178	35247.01
San Diego	70156	70893	5260907	18882.04
San Francisco	219279	235132	16195249	286143.69
Seattle/Tacoma	122226	124518	7385594	103409.85
Tampa/St. Petersburg	64396	64795	4781020	23048.08
Washington	177694	181101	11483285	64079.83

After we know what the problem is, it is simpler to find information to find a solution. For our problem, we would need to find information regarding a dataset, which is the sub-dataset I created from the United States of America's airport statistics full dataset. After discovering the problem and finding information pertaining to it, the next step is to implement a solution to the problem. For our problem, the solution is to create a good visualization that will give a viewer insight to the data included.

Edward R. Tufte was an expert in the field of information visualization and presented his theories and practices in designing graphics for data. From his ideas, he introduced the term of data-ink ratio. "Tufte defines the data-ink ratio as the portion of ink (i.e., pixels) that make up data-information on the view" (*Deep Little-known ways to make your data visualization awesome*). Along with this definition, he provided an equation for the data ink ratio. The data-ink ratio can also be written as "the amount of data-ink divided by the total ink required to print the graphic" (*Deep Little-known ways to make your data visualization awesome*). In other words, from the aspect of data-ink, data visualization should have all the necessary information, however, elements that does not provide information should be limited or removed. Using less

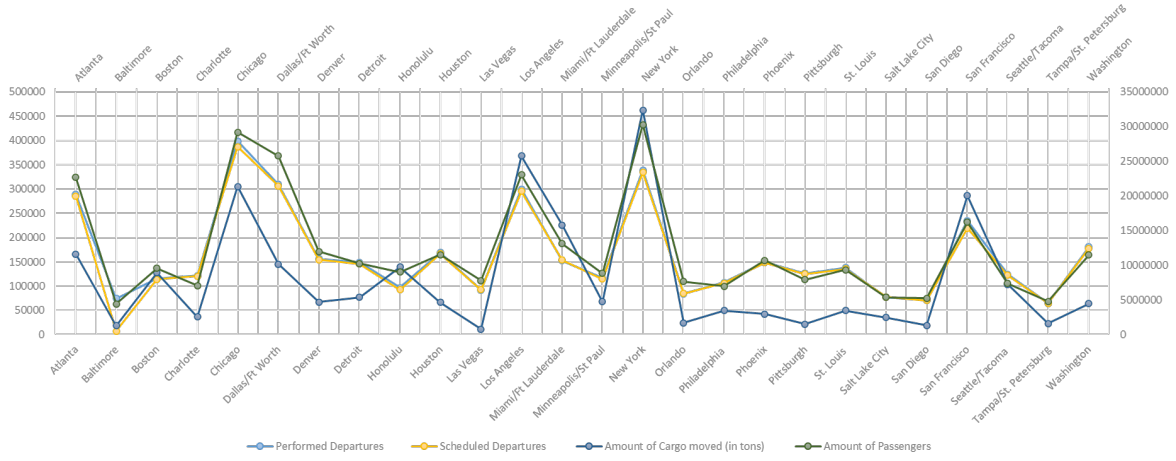
ink, such as various chart elements, could be more effective for the viewers to find the relevant information first, without reading elements that provide no information.

The first visualization I created with my sub-dataset consisted of bar charts and line graphs combined into one. While the visualization contained all the necessary information, it also had a very low data-ink ratio, meaning it contained a high number of unnecessary elements of information. The placement of the measurements on both sides of the visualization gives a confusing and inaccurate measurement of the data being measured. While I used different colors to separate the different variables the data was being measured against, the placement of the overall graphs gives a messy visualization. For example, the line graph for the *Amount of Passengers* is set to gray, unique from the other variables. However, when looking at the visualization, it is sometimes difficult to follow where exactly the line is set to for each city when it is mixed in with other colors. Also, because of the measurement style placement and increment style, the viewer does not a have of knowing whether the variable is being measured by the right incremental side or the left incremental side.



A good visualization is able to show a viewer the most important information and support the insight that the designer intended. My first version of my visualization did not accomplish my goal of support the insight intended as it left the viewer confused and with questions. I created another visualization for my project, this time using the same type of line graph for all the variables. Each line graph also has dots included throughout the line, with each dot signifying

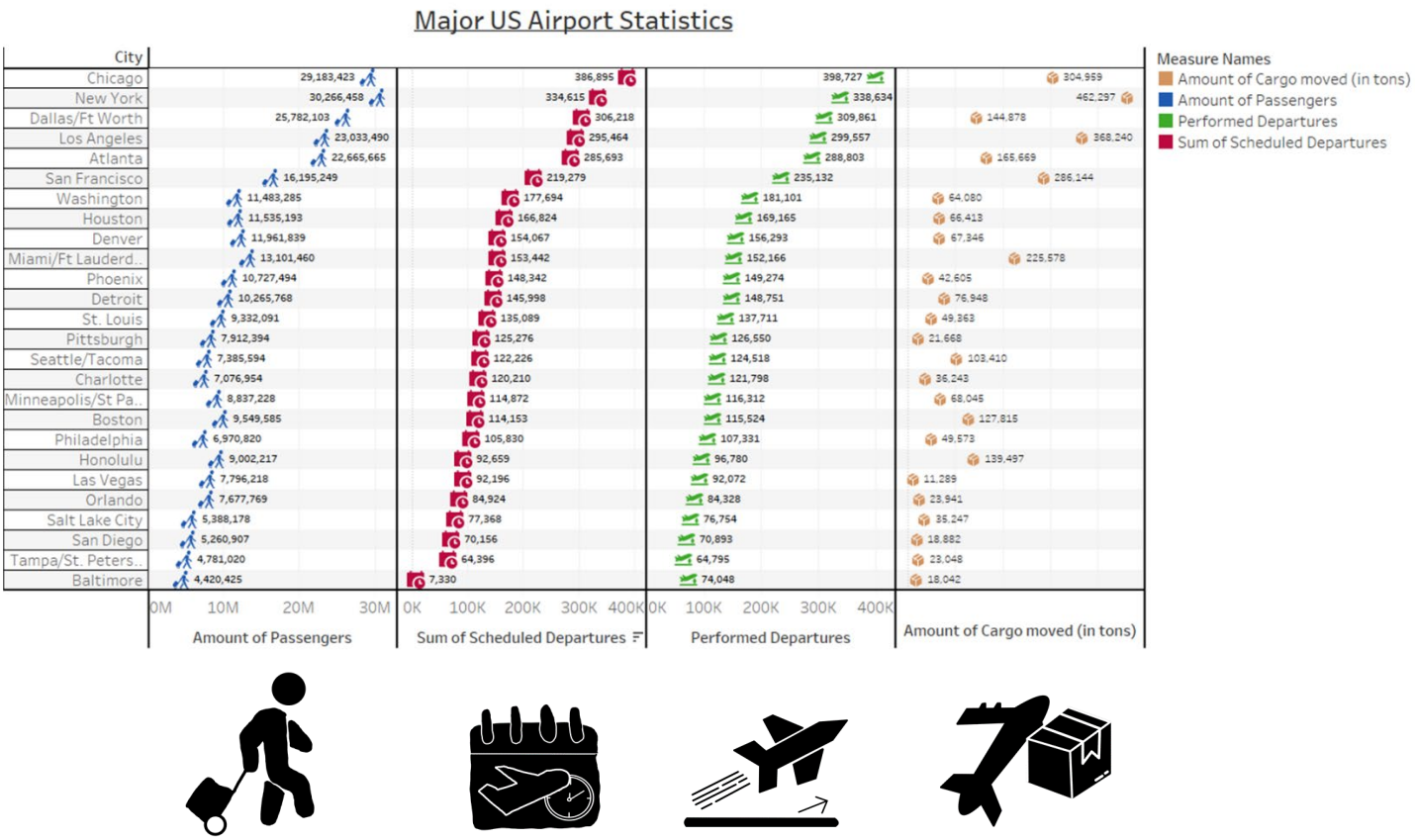
what city that specify part of the line graph is referring to. The cities are listed below at an angle, to make it them easier to read and easier for a viewer to separate. As with the first visualization, I color-coded the different variables as a way for a viewer to see the distinction in the visualization. While this visualization is cleaner and easier to read than the first visualization, it can still look messy with the lines being close to one another. For example, looking at the variables for the cities of Houston, Las Vegas, and Los Angeles, the lines begin to intertwine and creates a confusing, unclear image. It would be time-consuming as well for a viewer to follow any trends or patterns in the variables. Redundant data also contributes to the time-consuming.



Tufte included five principles that are important in regard to data-ink. First, show the data above all else. Second, you should always try to maximize the data-ink ratio. Third, your visualization should not have any non-data ink, elements that provides no new information. Fourth, your visualization should not contain redundant data-ink, elements that repeat the same information over again. Lastly, you should also go back to revise and edit your visualization to make it better.

With these principles in mind and an idea of the visualization I wanted to create, I went back to revise my second version visualization. With this version, I used the application Tableau to help create the visualization. This application came equipped with many different features that aided in the making the visualization display the insight I intended. One of the features included a measurement system unique to each variable, according to my data. For example, the *Total Amount of Passengers* are going by tens in the millions, whereas the *Total Sum of Scheduled* and *Performed Departures* are measured by hundreds in the thousands. In the previous

visualizations, I only used color to separate the variables, however, within Tableau, I used colors and glyphs to separate the variables. I drew the glyphs and used Tableau to insert the icons and color-coded them to give them meaning. The *Total Amount of Passengers* is represented by a blue person with a suitcase. The *Total Amount of Scheduled Departures* is represented by a red calendar with a time clock on it, as passengers generally set flight dates and times on calendars. The *Total of Amount of Performed Departures* is represented by a plane taking off, with an arrow pointing up. It is colored green to represent profits as the flights were successfully flew. Lastly, the *Total Amount of Cargo* moved by each city is represented by a brown package box.



References

TIBCO. "What Is Information Visualization?" *TIBCO Software*, https://www.tibco.com/reference-center/what-is-information-visualization.

Deep, Mala. "Little Known Ways to Make Your Data Visualization Awesome." *Medium*, Towards Data Science, 14 Oct. 2020, https://towardsdatascience.com/little-known-ways-to-make-your-data-visualization-awesome-890d71b6e365.

Dataset:

http://jse.amstat.org/datasets/airport.dat.txt

http://jse.amstat.org/jse_data_archive.htm