DESAUTELS | McGill

INSY 434-001

Final Project - Airline Industry
Professor Changseung Yoo
April 8th, 2019

Zachary Amar (260713997)
Alexandre Chahtahtinsky (260669581)
Aleem Damji (260713187)
Carly Matz (260652775)
Sean Mitro (260625474)
Nicholas Tariro Toronga (260715831)

Introduction

This project provides an analysis of the tweets and reviews from customers who write about their experiences with the top five airlines in the United States. These airlines, ranked by market share, include: Delta Airlines, American Airlines, United Airlines, Southwest and JetBlue. The main focus was placed on identifying the issues that customers complain about as well as the main issues associated with each of the five airlines. Leveraging text analytics approaches, including but not limited to Lift analysis, MultiDimensional Scaling, Sentiment analysis, Topic modelling and Network analysis, different graphs and diagrams were produced to guide the recommendations made and business insights that were uncovered.

Industry Overview

The domestic airline industry provides air transportation within the United States for passengers and cargo over regular routes and on regular schedules. This industry garners revenues of $142.3 billion per year and is expected to grow at an annualized rate of 2.5% for at least the next five years - with nearly 700 million domestic flights in 2015 alone. The industry is cyclical and very sensitive to key external drivers, predominantly factors that affect the number of domestic travellers (such as levels of per capita disposable income) and the world price of crude oil. As crude oil prices directly influence the price of jet fuel, airlines often implement fuel surcharges to increase revenue and shift some of the burden to consumers. As disposable income levels rise (as they are projected to in the United States), revenue in the industry will see growth, however, profit margins will be constrained due to rising fuel prices. This industry sees high competition with this trend expected to increase. Major players in the industry and their percent of market share are as follows: Delta Airlines (22.4%), American Airlines (21.3%), United Airlines (17.5%), Southwest (14.6%), and JetBlue (4.1%). With the high competition seen in this industry, smaller airlines are starting to capture market share through tough price competition. To fight back against these low-cost carriers (such as Spirit or Frontier), airlines with major market share are starting to focus more on their premium class segments to drive profit while offering some cost-cutting, no-frills tickets to consumers who prefer.

Network Analysis

We can visualize the flight network in the US using tools traditionally used for social network analysis. Using the online platform Polinode, two network graphs were drawn to compare the network architecture of two particular airlines with notably different network models: American Airlines with a traditional hub-and-spoke model, and Southwest who uses a unique point-to-point model.

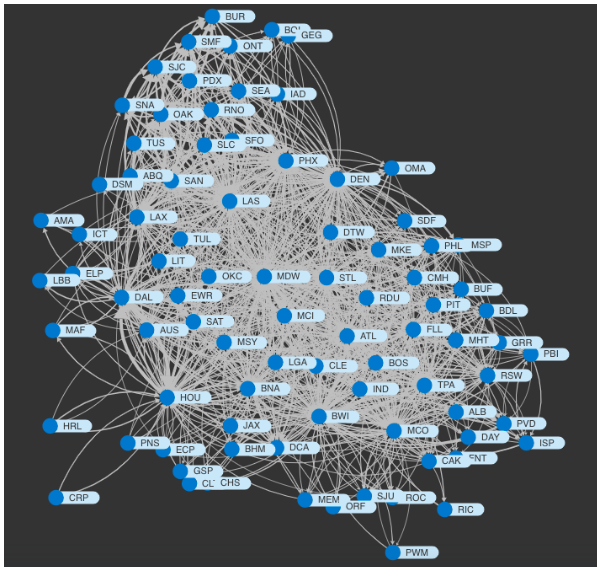


Figure 1: *Southwest Airlines Network*

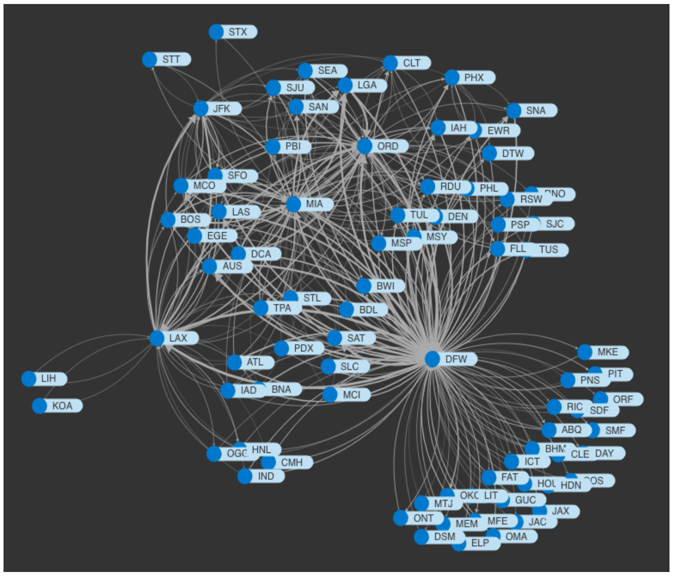


Figure 2: *American Airlines Network*

From observing the two network graphs, it is clear that Southwest's network has a higher density than American Airlines. While American has a major hub at Dallas Fort Worth (DFW), Southwest has no visible hub airports, with most airports being directly connected.

NodeXL was used to calculate individual and network level metrics for the entire US domestic air network. Overall, the airports of Atlanta, Dallas Fort Worth and Chicago O'Hare have the highest betweenness centralities of the network. Such a result is logical considering these three airports are the hubs for Delta, American Airlines and United Airlines, respectively and thus are responsible for connecting disparate parts of the States. Atlanta has the highest in-degree and out-degree centrality of the network at 164 airports, meaning it has the widest network of direct connections stemming from it.

Looking at the network edges, the most frequent flight routes can be identified. The three most common routes in the network are JFK-LAX, LAX-JFK and SFO-LAX. Despite ranking ninth in betweenness centrality, LAX airport is involved with all of the top six most flown routes in the US. In total, there are 4298 unique flight routes in the data which were flown at least once in 2015 connecting a total of 315 airports. The reciprocated edge ratio is 0.994, meaning that despite being a directed network, almost every node pair has flights operating both ways between the nodes. The overall network density is 0.04. This is low but natural as there is no possible way they every airport in the network could be directly connected with every other airport due to lack of demand. In fact, of the more than 400 US domestic airports with commercial service, less than 20 percent generate more than 50 directional passengers per day and only 5 percent of domestic city-pairs generate sufficient traffic to support non-stop service (Lott & Taylor, 2005).

Problem Statement

With the increase of user generated content, airlines are one of the industries that have been fundamentally affected as customers take to different forums their opinions and feelings. The industry is facing strong competition and understanding customers' feelings about an airline's service has become one of the key priorities for marketing managers. This provides the root of the problem under scrutiny. Centered on analyzing the tweets and reviews for different airlines, the main challenge revolves around discovering the service issues that customers or travelers are really concerned about. It is important to note that only the tweets from customers were considered and not those from the airlines themselves as the former accurately represent customers' opinions. In particular, most analysis aims to identify the main topics that customers complain about and the main problems that each airline face so that business insights can be drawn, and recommendations based on airlines' respective service delivery can be made.

Methodology

Determining the airlines

The IBIS World Industry Report 48111b on Domestic Airlines in the US guided the decision of which airlines to target for this project. Our project followed their report in choosing the top five airlines in the US (Delta, American Airlines, United, Southwest, and JetBlue) as defined in the report. In order to segment our chosen airlines appropriately, the dplyr package was used within R.

Web Scraping

To create a dataset, data was scraped from both Twitter as well as online forums regarding the airline industry (i.e. Tripadvisor). Using the "GetOldTweets" package in Python, tweets were collected from 2015 (the pre-specified date range) in English that mentioned each of the airline companies the project focuses on. Tweets that were made by the company were avoided as they were not the most representative ones. To collect data from online forums, the URLs of the five airlines chosen were identified to scrape. From there, the URLs were edited to scrape multiple pages using the Sitemap Metadata which allowed more reviews to be obtained than presented on a single web page as the code collected 2000 reviews, showing about 10 per page (Appendix 1). It was attempted to scrape the review directly from the main page, however, it was noticed that the webpage shows previews of the reviews and thus the scraping was only returning about half of the review (Appendix 2). To correct this, he scraper was set to click the link in the title, scrape the full review as well as the rating, and return to the main page, and repeat on the following reviews. Then identical sitemaps for each airline were created and let them run, simultaneously, for five hours. Each sitemap clicks the rating title link which opens the rating and review to be recorded (Appendix 3). It is important to note that pagination is done through the URL, not through a selector. The data was then able to be cleaned - removing unnecessary columns from the output csv files and concatenate them into one, cohesive file.

MDS

The analysis began by first implementing a word frequency analysis to identify the major conversational topics each airline generates online as per our merged twitter and web-scraped dataset. Multidimensional scaling was next implemented from between airlines. To do so, all reviews were combined and tweets per airline and then used a find-and-replace word list to replace all the variation of airlines as well as combine similar topics. From here the lift and dissimilarity matrices were able to be calculated and generate the MDS plot. MDS was next used between airlines and major consumer issues. Running a word frequency per airline unveiled that most complaints were about the same topic, therefore it was able to be determined the following key words and treated them as main consumer discussion topics: “time” (time, delayed, hours), “service” (service, customer, attendants, crew, luggage, boarding), and “flight” (flight, seats, cancelled, food, snacks, pretzels) . Once again, the lift (Appendix 4), dissimilarity matrices as well as the MDS plot were generated as seen below.

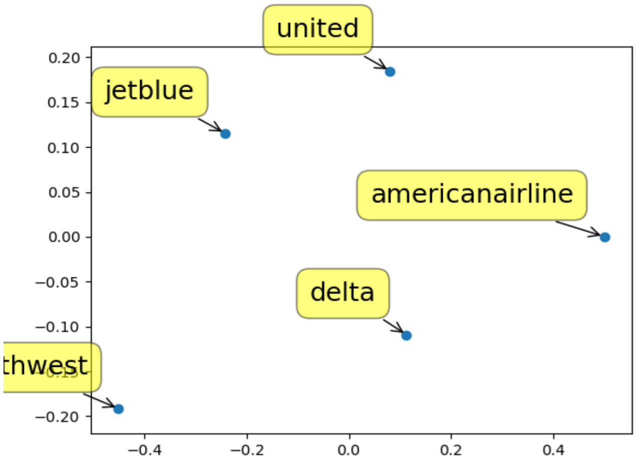


Figure 3: *MDS plot of the Airlines*

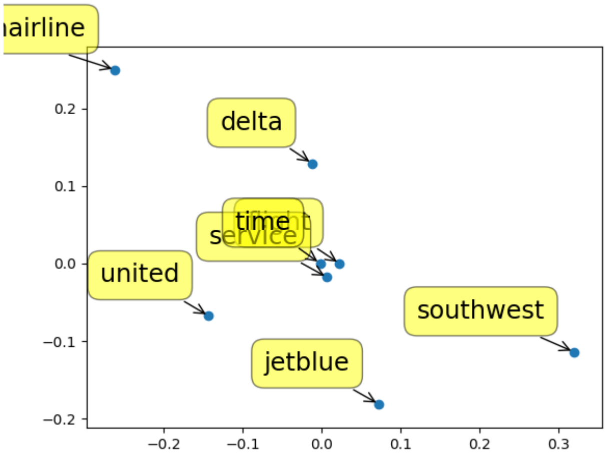


Figure 4: *MDR plot of the Airlines vs the Topics*

Sentiment Analysis

In order to determine how people, feel about each airline, sentiment analysis was performed on the data in order to calculate a sentiment score for each company. To do so, our dataset of reviews was analyzed using Sentistrength. After generating an output file, it was loaded into Excel, however, the output was not transferred cleanly and had to be manipulated - there were reviews that were not correctly separated by tab by using text-to-column that needed to be manually edited and then reintegrate thus reviews into the original output. While it was possible to generate the sentiment score for each airline, it is important to note the accuracy observed from Sentistrength. A problem encountered concerned the term “cheap” - Sentistrength would assign this word with a -1 despite the fact that when consumers were discussing how they got a cheap flight, this was overwhelmingly a positive attribute to their experience with this airline. Additionally, some reviews discussed the writers’ satisfaction with the airport, which cannot be attributed to the airline itself. Another clear error was seen in a review written “I have not had a bad experience on American Airlines” resulting in a sentiment score of -1. These errors cause the calculated sentiment scores to be interpreted with caution.

We also used the scripts parserforsentiment and sentiment in order to identify the sentiment for each of the main topics by airlines. This allowed us to determine what people thought about the main topics we had discovered above. Unfortunately, a lot of the lift between the topics and the airlines were less than 1 so while we did run the code for the sentiment, we did not include it in the report.

Ranking Airlines

As seen in Appendix 6, the average star rating of each airline based off of their web reviews as well as the average sentiment score for each airline were calculated, showing how people feel about each airline. In order to further analyze the problems faced by each airline, the following flight data was summarized (originally sourced from the United States Bureau of Transportation) regarding their average delay and arrival times (Appendix 7). Furthermore, in the data, each delay was segmented into one of three categories: late airplane, airline issue, and weather delay. The data also reports whether the flight was cancelled or diverted. For each airline, the percentage of their flights with each type of delay was calculated to present a relative delay rate. We were able to determined the most common and least common reasons for delay per airline and allocate a "biggest issue" to each company to focus their efforts on (Appendix 8).

Conclusions and Recommendation

The conclusions from the analysis demonstrate that each airline examined has a different set of issues and, therefore, has a different recommendation for their needs.

The recommendation for JetBlue would be to focus on delay reduction. JetBlue has the highest sentiment value and second highest rating of the five airlines, showing that JetBlue customers are generally happy with JetBlue's service. However, JetBlue has the second highest average departure delay and the highest average arrival delay, with airline-related issues being the largest cause of delayed flights. JetBlue also has the highest percentage of flights cancelled. To resolve these airline issues, JetBlue should reassess their flight schedules. Seventy percent of JetBlue flights are based out of John F. Kennedy Airport in New York or Boston's Logan Airport, both of which are located in the Northeastern United States ("JetBlue Heads for Worst Year of Flight Delays in a Decade", 2017). This region of the USA not only has the most congested airspace, but the two airports are close enough that one significant weather event would impact both airports and severely impact JetBlue's operations ("JetBlue Heads for Worst Year of Flight Delays in a Decade", 2017). This is in contrast to competitors, who have bases spread out throughout the United States, mitigating weather and flight congestion risk ("JetBlue Heads for Worst Year of Flight Delays in a Decade", 2017). Recognizing that JetBlue has found a successful niche in this region, the recommendation for JetBlue is to not undertake a costly and risky endeavor by expanding geographically. Instead, JetBlue should increase the scheduled time between flights for airplanes on the ground at these airports to create leeway that can absorb delays. JetBlue will then be more flexible in responding to both airspace congestion and weather events while continuing to provide excellent customer service.

Delta Airlines' recommendation is to shift their marketing to focus on their on-time performance. Currently, Delta has the lowest average delays for both arrival and departure with the largest cause of their delays being weather-related. At the same time, Delta has a 4-star rating on average alongside the second highest sentiment of 0.733 among customers. Delta is excelling in both customer service and delay mitigation, but they have an opportunity to combine the two in one marketing campaign. By advertising on-time performance, customers will associate Delta with reliability and the airline will be able to deliver on this metric. Additionally, this campaign will convince time-sensitive travelers to choose Delta for their flying needs rather than competitors.

Southwest Airlines should engage in the industry practice of "schedule padding". Southwest customers view the airline positively, with it receiving the highest rating of all airlines and a very high sentiment score of 0.732. However, Southwest is in the middle of the pack regarding flight delays with an average departure delay of 10.60 minutes. The majority of Southwest flights are delayed due to a late airplane. Late airplanes have a significant domino effect in Southwest's network as the airline is the only major American airline following a "point-to-point" strategy ("Factors That Have Strengthened Southwest's Domestic Presence", 2016). With this strategy, Southwest does not connect passengers through hub airports, but rather using direct flights between points ("Factors That Have Strengthened Southwest's Domestic Presence", 2016). However, unlike at a hub, this strategy means that delayed airplanes cannot be replaced with idle airplanes that are parked at the hub airport as point-to-point destinations do not necessarily have an idle Southwest plane waiting. However, the "point-to-point" strategy is highly successful for Southwest and is a key differentiating factor in their continued success ("Factors That Have Strengthened Southwest's Domestic Presence", 2016). As a result, Southwest should not rethink this strategy, but instead engage in "schedule padding" - the practice of adding extra time to a flight's departure and arrival time to add more latitude in the case of a delay (McCartney, 2017). Most major airlines follow this practice, but Southwest's schedule padding is the lowest of all major American airlines (McCartney, 2017). By padding schedules, Southwest can reduce airplane delays while maintaining their current operational structure.

American Airlines should focus on rebranding themselves as a fully low-cost carrier. American has the second lowest customer rating and sentiment scores of the airlines. At the same time, American's average departure and arrival delays are also the second lowest of the five airlines. American customers do not view the airline favorably, and the airline has responded by reducing seat sizes and removing entertainment options in all classes of service (Matyszczyk, 2018). This is in contrast with other airlines, such as Delta, choosing to do the opposite (Matyszczyk, 2018). American's actions have been undertaken with the goal of reducing costs. American should pass on some of the savings this cost reduction to customers and rebrand themselves as a low-fare airline. This will allow the airline to continue with their current course of action and reduce their wasted operational costs. As a result, American customers will be willing to tolerate poor conditions and service as their airfares will be relatively cheap. This strategy allows American to continue with their present course of action while ensuring that customer expectations meet reality.

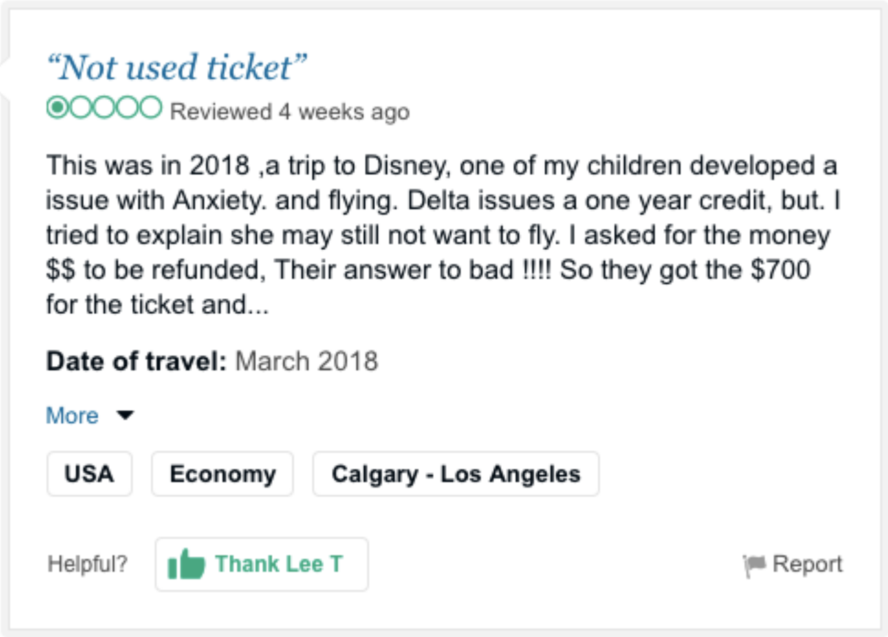
United Airlines' focus should be on repairing relationships with their consumers. United has the lowest rating and the lowest sentiment score of all five airlines. On top of this, United has the highest departure delay and the second highest arrival delay. The combination of these significant issues is a dilemma that United must address. The best option for United Airlines would be to focus on one issue at a time and attempt to solve it before tackling the second issue. United should focus on improving customer sentiment because it can be solved without changing the airline's operational structure, unlike delay reduction. United can improve customer relations by focusing on a customer-first atmosphere, from the top down. Currently, United executives publicly blame customers' focus on cost as being the reason for a reduction in passenger service (Matyszczyk, 2019). A rebranding of executive statements focusing on customers as partners, not opponents, would improve customer relations and help create a customer-first company culture. This culture can be propagated by front-line employees by rewarding them based on customer service and empowering them to help customers by giving employees the discretion to provide benefits such as flight changes or credits in case of a delay.

All in all, each of the five airlines face different issues as evidenced by the analysis performed. As such, the airlines have different recommendations targeted at both their marketing campaigns and their operational performance.

Appendix

Appendix 1: Links to TripAdvisor Reviews
https://www.tripadvisor.ca/Airline_Review-d8729020-Reviews-or10-American-Airlines#REVIEWS shows the first 10 reviews.
Change to: https://www.tripadvisor.ca/Airline_Review-d8729020-Reviews-or[10-2000]-American-Airlines#REVIEWS will go to review 2000

Appendix 2: Example TripAdvisor Review



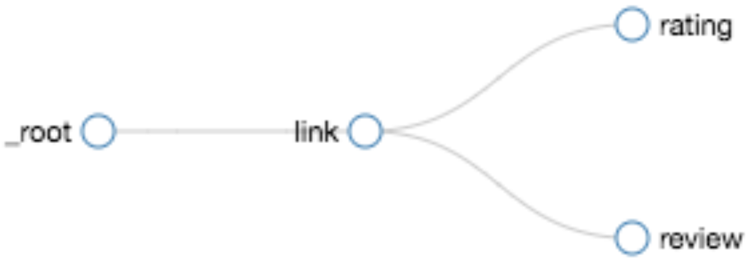
“Not used ticket”

●○○○○ Reviewed 4 weeks ago

This was in 2018 ,a trip to Disney, one of my children developed a issue with Anxiety. and flying. Delta issues a one year credit, but. I tried to explain she may still not want to fly. I asked for the money $$ to be refunded, Their answer to bad !!!! So they got the $700 for the ticket and...

Date of travel: March 2018

Appendix 3: Website Map for Web Scraper



Appendix 4: Lift Matrix by Airlines

	American Airline	Delta	Jetblue	Southwest	United
American Airline		0.21	0.08	0.07	0.20
Delta			0.20	0.13	0.21
Jetblue				0.20	0.27
Southwest					0.11
United					

Appendix 5: Lift Matrix and Sentiment for Airlines & Topics

	Time		**Service**		**Flight**	
	Lift	*Sentiment*	*Lift*	*Sentiment*	*Lift*	*Sentiment*
American Airline	0.61	NA	0.54	NA	0.48	NA
Delta	0.89	NA	0.82	NA	1.07	0,13
Jetblue	1.09	0.12	1.08	0,23	0.20	NA
Southwest	0.72	NA	0.80	NA	0.84	NA
United	0.11	NA	0.95	NA	0.94	NA

Appendix 6: Average Review Rating versus Sentiment for Airlines

Airline	**Avg. Sentiment**	**Avg. Rating**
JetBlue	0.810	4.06
Delta	0.733	4.00
Southwest	0.732	4.27
American Airlines	0.257	3.40
United Airlines	0.116	3.24

Appendix 7: Average Delays per Airline

Airline	**Average Departure Delay (mins)**	**Averaged Arrival Delay (mins)**
Southwest	10.60	4.37
JetBlue	11.50	6.68
Delta	7.37	0.187
American Airlines	8.90	3.45
United Airlines	14.40	5.43

Appendix 8: Delay Analysis per Airline

Airline	**Ranking By Delay Type (1 = most delays of this type, 5 = least)**			**Percentage of all flights with Issue**		**Biggest Issue**
	Late Airplane	**Airline Issue**	**Weather Delay**	**Flight Cancelled**	**Flight Diverted**	
Southwest	1 (12.9%)	3 (11.5%)	5 (0.79%)	3 (1.27%)	3 (0.270%)	Late Airplane
JetBlue	2 (11.9%)	1 (14.5%)	4 (0.81%)	1 (1.6%)	2 (0.273%)	Airline Issue / Cancelled Flight
Delta	5 (5.7%)	5 (7.2%)	3 (1.35%)	5 (0.44%)	5 (0.203%)	Weather Delay
American Airlines	4 (7.9%)	4 (9.5%)	2 (1.37%)	2 (1.5%)	1 (0.293%)	Diverted Flight
United Airlines	3 (9.6%)	2 (12.8%)	1 (1.46%)	3 (1.27%)	4 (0.269%)	Weather Delay

References

Factors That Have Strengthened Southwest's Domestic Presence. (2016). Retrieved from https://www.forbes.com/sites/greatspeculations/2016/09/14/factors-that-have-strengthened-southwests-domestic-presence/#14d83ece2cbe

JetBlue Heads for Worst Year of Flight Delays in a Decade. (2017). Retrieved from http://fortune.com/2017/12/30/jetblue-flight-delays/

Matyszczyk, C. (2018). American Airlines Just Quietly Admitted That Economy Class Passengers Will Be Squeezed Further (First Class, Too). Retrieved from https://www.inc.com/chris-matyszczyk/american-airlines-quietly-admits-its-shoving-more-seats-into-even-more-planes.html

Matyszczyk, C. (2019). United Airlines President Says It's Customers' Own Fault the Airline Is Shoving More Seats Into Planes. Retrieved from https://www.inc.com/chris-matyszczyk/united-airlines-president-says-its-customers-own-fault-that-airline-is-shoving-more-seats-onto-planes.html

McCartney, S. (2017). Which Airlines Pad Their Schedules the Most?. Retrieved from https://www.wsj.com/articles/which-airlines-pad-their-schedules-the-most-1498662950

Lott, S., & Taylor A. (2005). Hub networks hurt by handling capacity. Aviation Week & Space Technology, 163(18), 56.