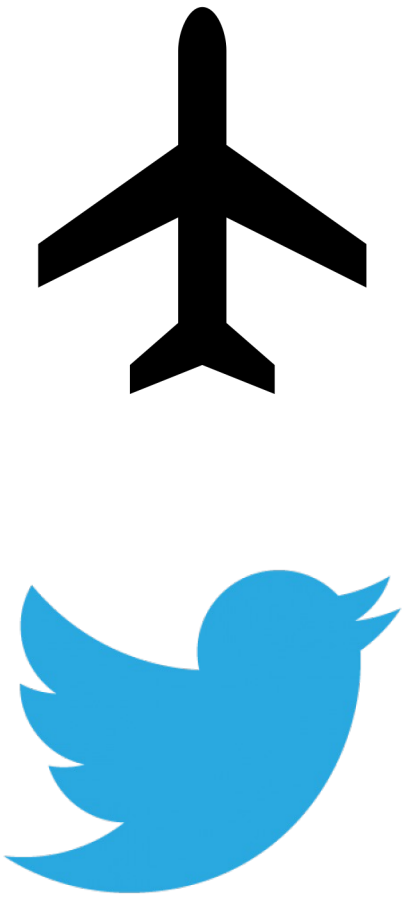
Twitter US Airlines Sentiment Analysis

Nate Rodriguez, **2023**



Introduction to Dataset and Cleaning

This is a Kaggle dataset that contains **tweets** collected from twitter regarding **US Airlines from February 16th to February 24th.** In the dataset you can find the following contents:

```
tweet_id
airline_sentiment
airline_sentiment_confidence
negativereason
negativereason_confidence
airline
airline_sentiment_gold
name
negativereason_gold
retweet_count
text
tweet_coord
tweet_created
tweet_location
user_timezone
```

As part of the cleaning process, and to better. understand the dataset, I transformed the data set to only contain these variables of interest:

tweet_id

airline_sentiment

negativereason

airline

text

tweet_coord

tweet_created

Introduction to Dataset and Cleaning (Cont.)

- Although selecting only the variables of interest helped clean the dataset, there was more to do:

Coordinates: In order to use the coordinates in a Power Bi dashboard, I split the coordinates by comma and hard bracket delimiters in order to have two columns, one with latitude and one with longitude. This mutation was done in Power Bi's Power Query.

Date: The tweet_created variable was given in date, time format but in order to plot tweets over time, I needed to split by a ” “ delimiter to get the YYYY/MM/DD format. Once I had this as its own variable, I appended it to the dataset. This mutation was done in Python using the Pandas library.

Airlines: Taking a look into the data, the airline ”Delta” was wrongly added as the airline since the corresponding tweets contained the “@JetBlue” tag. To fix this, I changed all airlines titled “Delta” to “JetBlue”. This was done in Power Bi and in Python.

Action Steps

1. Analyze dataset headers and values using filters and python tools.
2. Clean the dataset in order to reduce noise and remove unnecessary variables.
3. Using Python, conduct exploratory analysis to get familiar with dataset.
4. Create an interactive Power Bi dashboard that highlights high-level details and trends within the dataset.
5. Style the dashboard in a readable, intriguing manner.
6. Create visuals using Python to discover trends and develop insights.
7. Research any context that may be beneficial to the dataset time frame.
8. Combine all output and context to draw conclusions about the data set.

Methods of Analysis

Python

In order to better understand the dataset, I used the python libraries pandas, numpy, plotnine, and Word Cloud. These libraries are able to conduct numerical summaries and visualizations to help draw conclusions about the data.

<u>Python Deliverables:</u>

1. Positive and Negative Sentiment Word Cloud
2. Area Chart displaying tweets over time
3. Negative Reasons Table

Power Bi

Power Bi is a data visualization software that allows you to create dashboards to tell a story about a dataset. Using Power Bi, I created an analytical dashboard to summarize the dataset and identify trends. My user-friendly dashboard aims to be highly comprehensible but specific to airline needs.

Power Bi Deliverables:

1. Airline Sentiment Dashboard: Includes high level visualizations and KPI's that identify trends from the collection of tweets.

Python Visualizations: Word Clouds

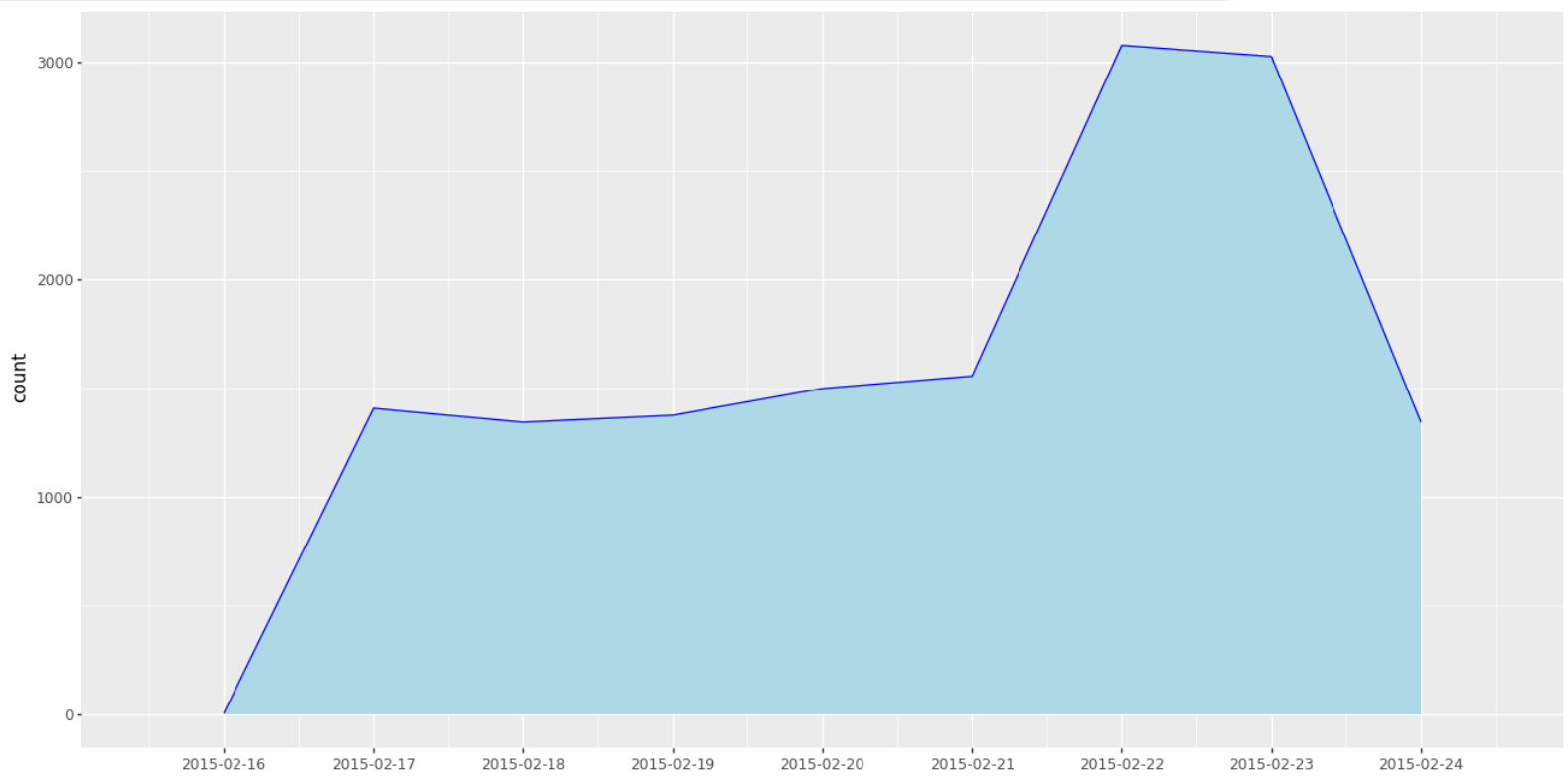
Positive Word Cloud

The most common words in positive tweets are: thank, flight, great, awesome, and love.

Negative Word Cloud

The most common words in negative tweets are: flight, United, bag, delayed, and plane.

Python Visualizations (Cont.): Area Chart

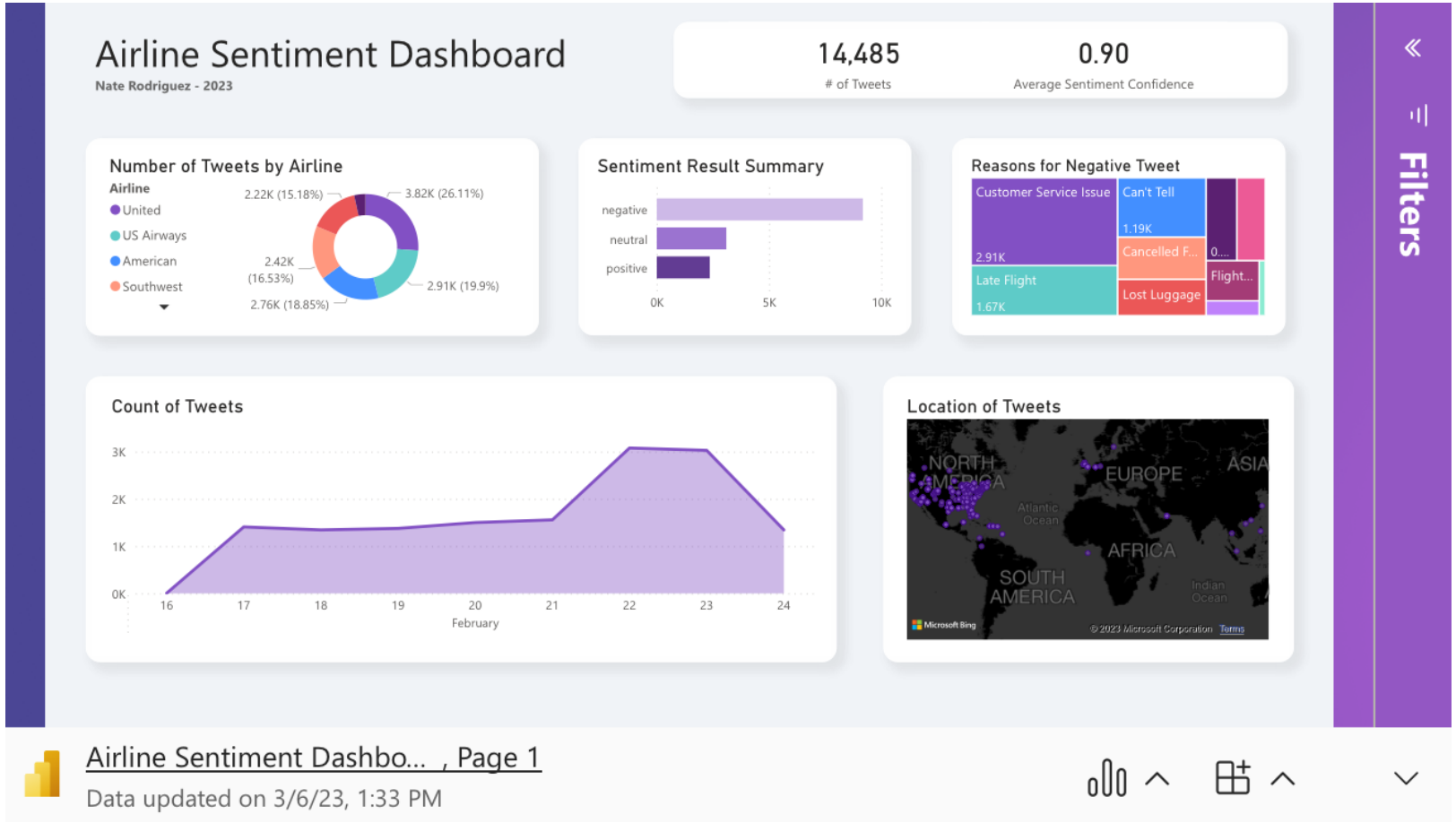


Area Chart

This Area Chart shows the number of tweets per day ranging from February 16th to February 24th.

As you can see from the visualizations, there was an increase in tweets beginning February 22nd lasting until February 24th. I will get more into the interpretation of this increase in Slide 9.

Microsoft Power BI Dashboard



Power Bi Dashboard

This interactive dashboard shows a variety of insights such as:

1. Customer Service Issues were the most common reason for negative tweet.
2. Tweets were located mostly in Eastern United States.
3. Majority of tweets were negative.
4. United Airlines received highest number of mentions.

Also take note of KPI metrics such as **# of tweets** and **sentiment confidence** and how they change as you select only particular airlines.

Although a simple dashboard, this is a great way of communicating the data to any manager.

Drawing Conclusions from Context

This twitter data was collected from February 2015, ranging from the 16th to the 24th. In Data Science, is vital to exploit outside context in order to better understand the dataset. From a simple google search of key events in February 2015, I discovered that there were significant environmental and societal events that occurred.

February 2015 Cold Wave:

Record Setting cold weather and storms hit the eastern United States causing several flights to be grounded and delayed [1]. This would explain the **influx of complaints** as customer service issues and delays were two of the more common phrases found in the World Cloud. More specifically, a severe storm hit on Feb. 22nd leading to spike in complaints from airline customers [2].

United Airlines Mistaken Low Fares:

In February 2015, United Airlines mistakenly posted low-fare first class flights onto their website. Those who exploited their mistake were expecting to be given their low-fare but instead, the deals were not honored. This is supported by the data since United Airlines had the greatest number of tweets directed at them [3]. United not only had to deal with weather influenced delays, but also angry customers whose expectations were not met.

Drawing Conclusions (Continued)

- During the time frame of this data collection, **significant weather** and **outside events** caused massive delays across the airline industry, seeming to affect **United Airlines** most severely. This caused an influx of negative tweets mostly targeted towards **customer service**, **late flights**, and **cancelled flights**. This insight is beneficial because it allows airlines to target the struggling services and implement change.
- If I were a professional in the airline industry, I would utilize my **Power Bi dashboard** to track how the public is reacting to ongoing changes. Twitter can be a complex environment that disperses a large variety of opinions, but it can also serve as a generally accurate reflection of what the public thinks of your organization. By using **sentiment analysis** to track what the public is saying about you, organizations can stay on top of problems and implement change on a large scale. Any of the airlines that were included in the data would greatly benefit from this sort of analysis.

References

1. National Weather Service. (2015, February 14). Blizzard of 2015. Retrieved from https://www.weather.gov/iwx/20150214_blizzard

2. National Weather Service. (2015). Winter Storm Summary: February 20-21, 2015. Retrieved from https://www.weather.gov/media/phi/reports/WSS21Feb2015.pdf

3. USA Today. (2015, April 13). Long tarmac delays at U.S. airports spiked in February. Retrieved from https://www.usatoday.com/story/todayinthesky/2015/04/13/long-tarmac-delays-at-us-airports-spiked-in-february/25702235/

Appendix

Link to GitHub Repository that includes Google Collab Notebook, Dashboard, and this presentation.

Other Graphics

Figure 1. Tabulation showing most common reasons for negative tweets.

```
Customer Service Issue          2910
Late Flight                     1665
Can't Tell                      1190
Cancelled Flight                 847
Lost Luggage                     724
Bad Flight                       580
Flight Booking Problems          529
Flight Attendant Complaints      481
longlines                        178
Damaged Luggage                   74
Name: negativereason, dtype: int64
```

Figure 2. Tabulation showing variables and there corresponding number of null values.

```
tweet_id                        0
airline_sentiment               0
airline_sentiment_confidence    0
negativereason               5462
negativereason_confidence    4118
airline                         0
airline_sentiment_gold      14600
name                            0
negativereason_gold         14608
retweet_count                   0
text                            0
tweet_coord                 13621
tweet_created                   0
tweet_location               4733
user_timezone                4820
dtype: int64
```