Flight Data Visualization

Andrew Le
Ming Luo
Sajjad Kazemi
Venkata Sai Tarun Reddy Pongulaty

Table of Content

Introduction

Dataset Description

Method

Results

Shiny App

Conclusion

What is the problem?

Passengers want to get to their destinations on time but are often unable to, resulting in negative feedback

Airlines should be more proactive in their continual improvement processes

Why is it worth further research?

Brings awareness to delays that can be controlled by human processes

Saves passengers time and avoids unneeded stress

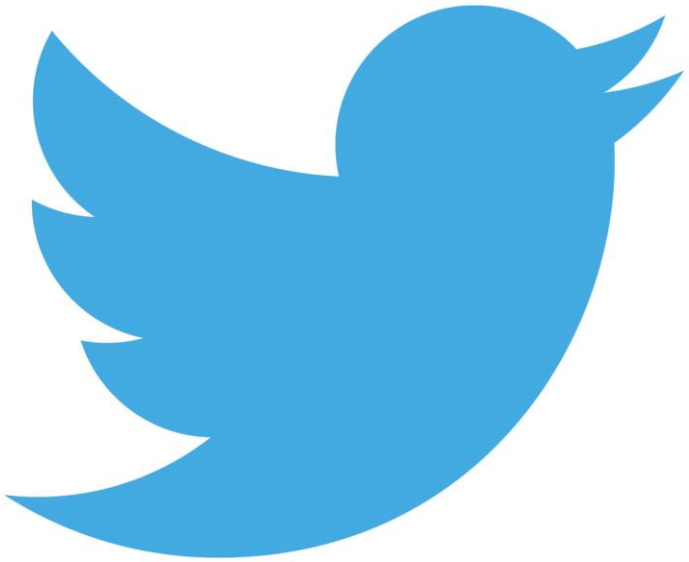
Provides feedback for airline companies so they can improve their practices

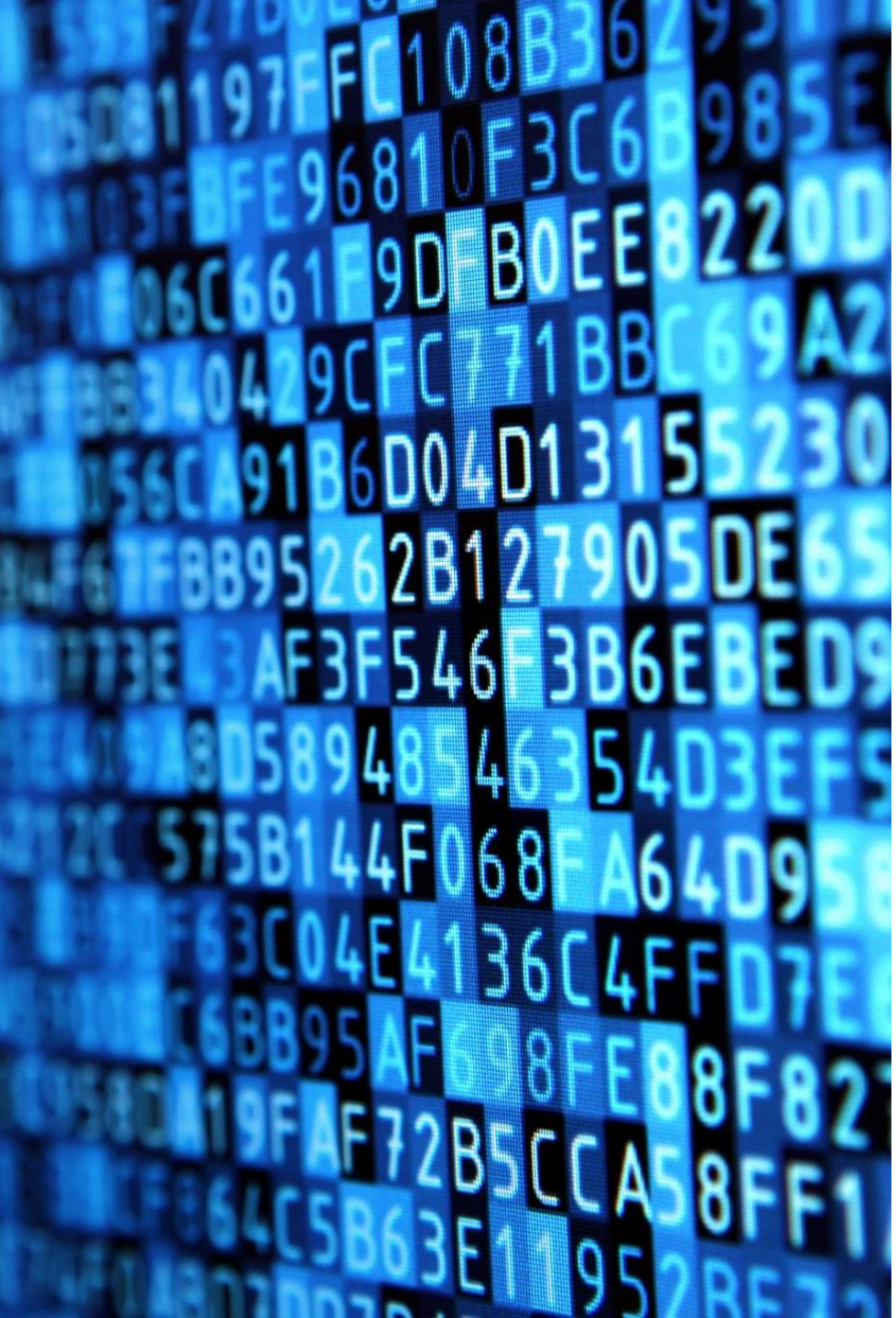
How could you solve this problem?

Organizing and analyzing information on airline flight delays, as well as passenger feedback

Create visual tools that will help people pick flights and for airlines to reduce issues causing flight delays

What data are we looking at?



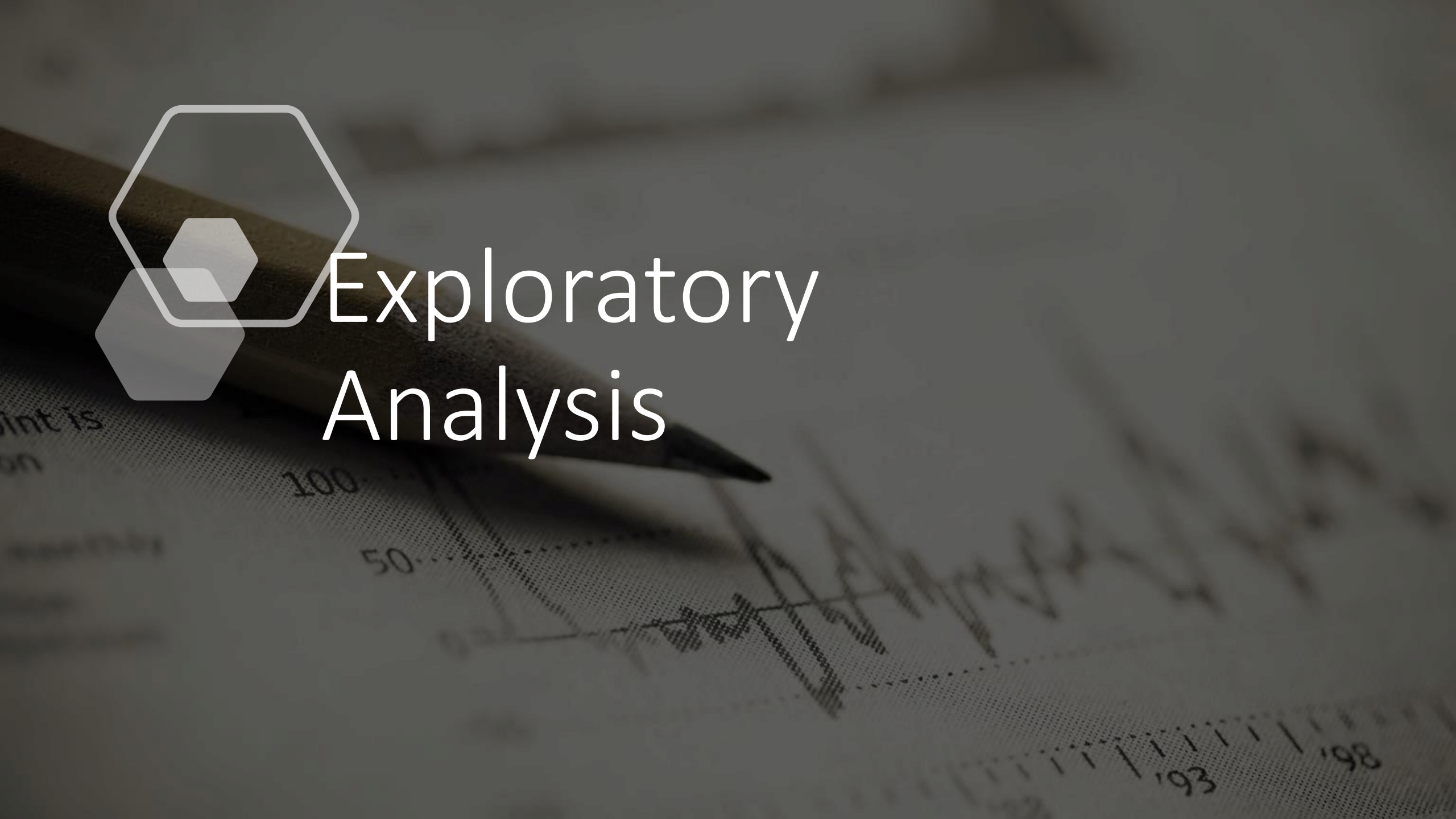


Flight Delays

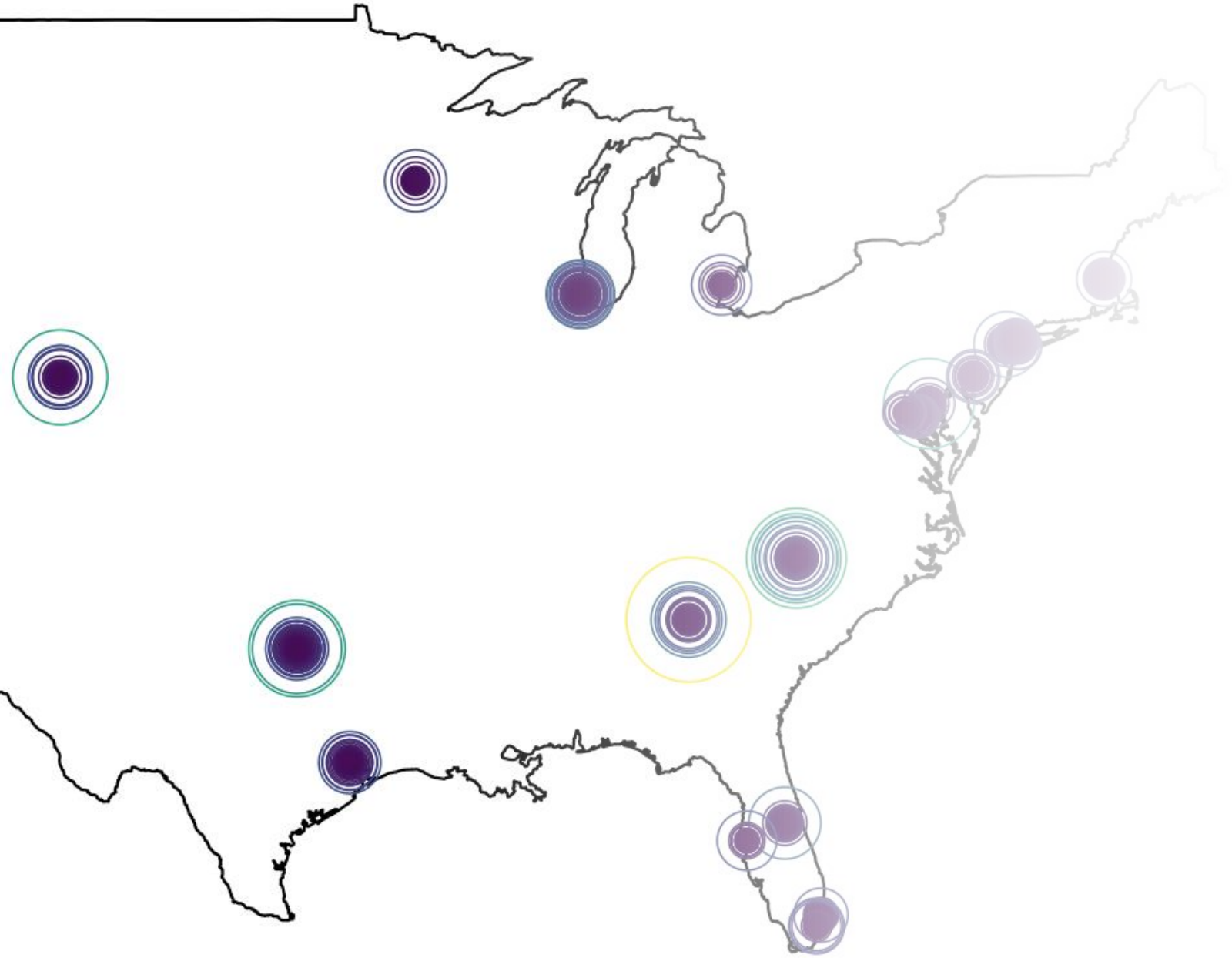
- **Size:** 27*21,804
- **Number of Numeric Variables:** 20
- **Missing Values:** 7%
- **Date:** Feb 2015

Twitter Sentiment

- **Size:** 15*14,600
- **Number of Numeric Variables:** 4
- **Missing Values:** 28%
- **Date:** Jan 2015 to Jan 2021



Exploratory Analysis



Airports Overview

- Locating airports based on different variables
- Finding whether an airport had an unusual cancelled/delayed flights

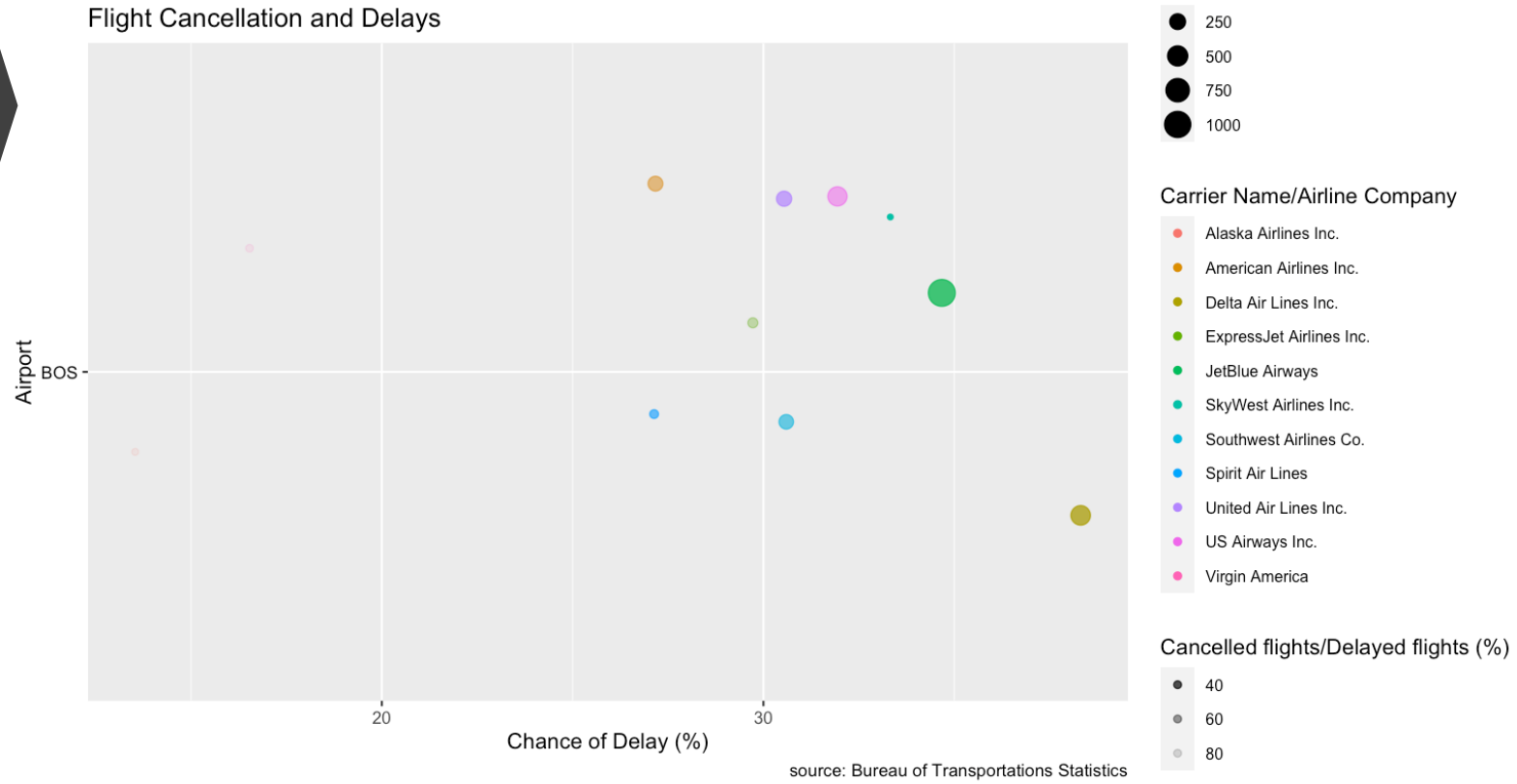
Cancelled Flights

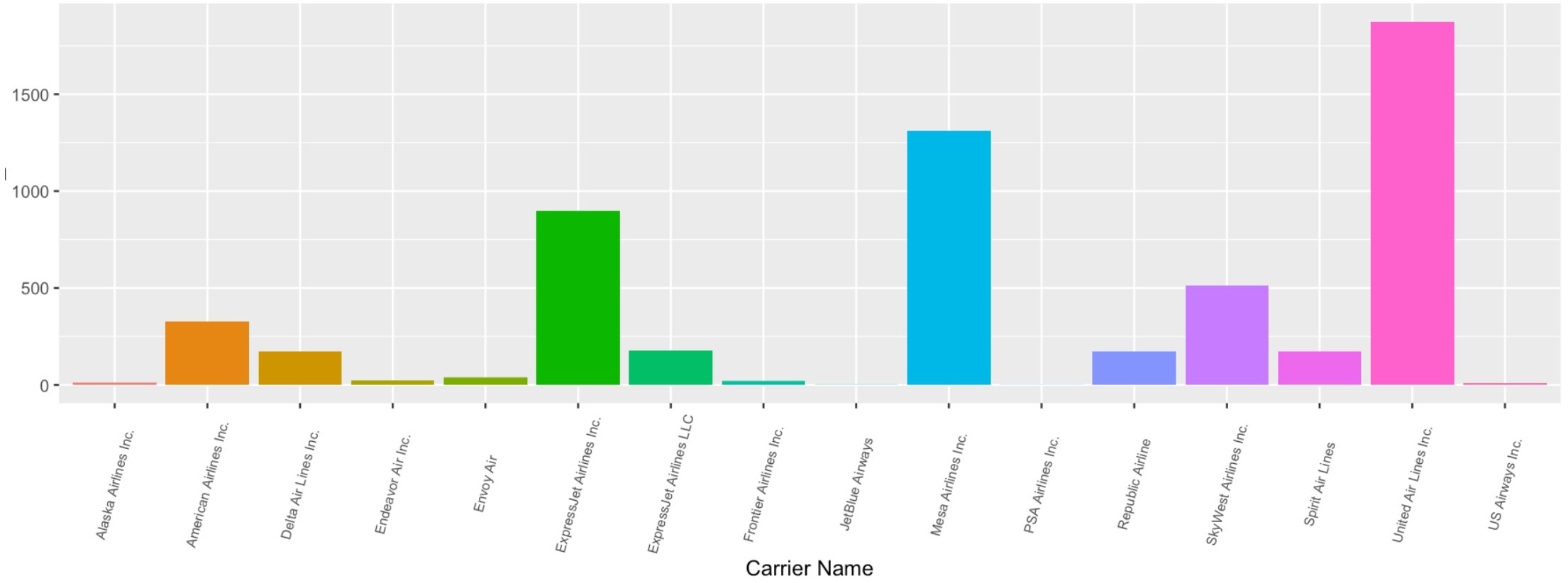
Delayed Flights

Boston Logan Airport
February 2015

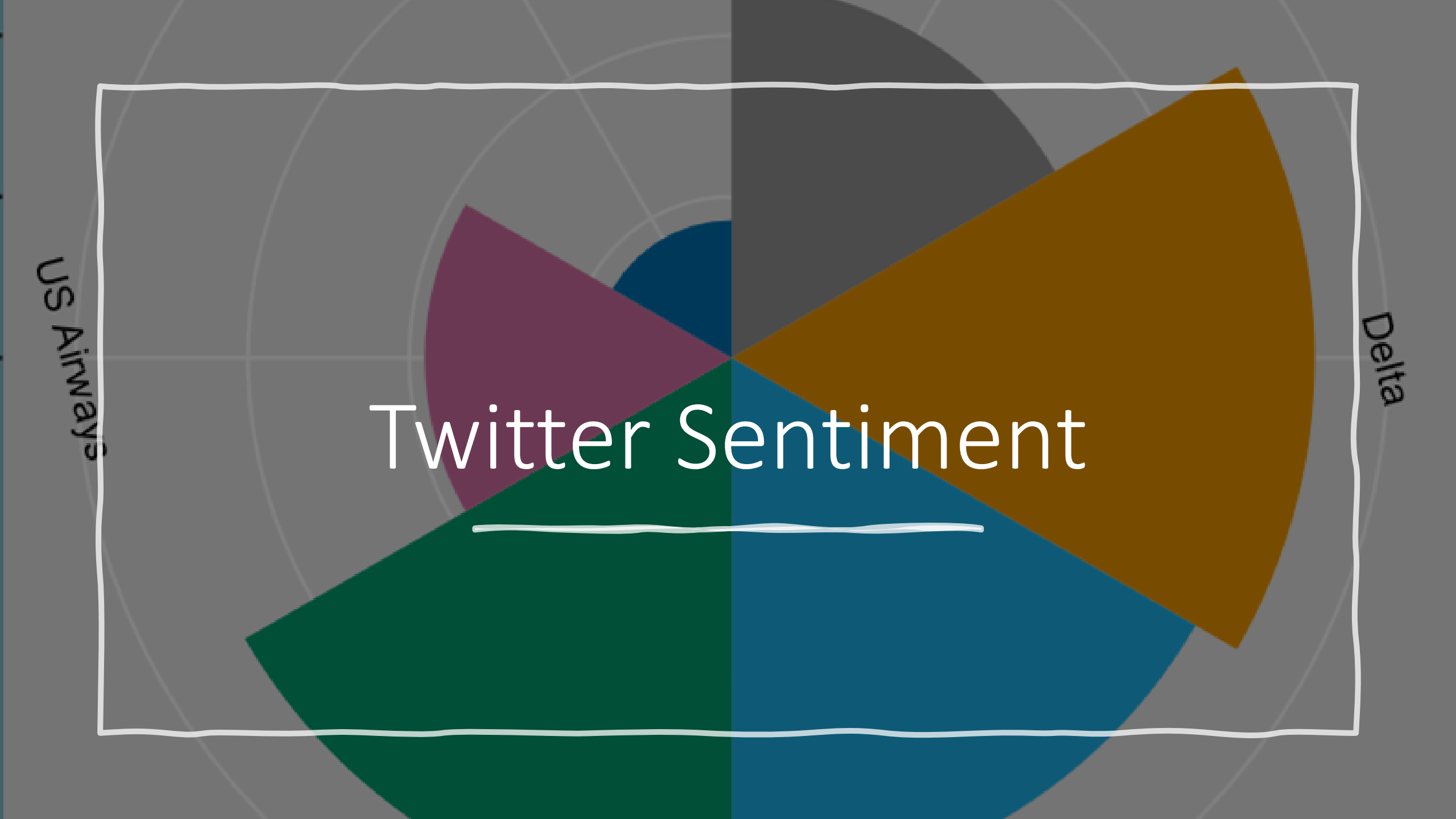
Virgin America and Alaska Airline near 80% for cancelled flights if a delay occurred.

SkyWest Airlines Co. had the smallest cancellation rate per delayed flights



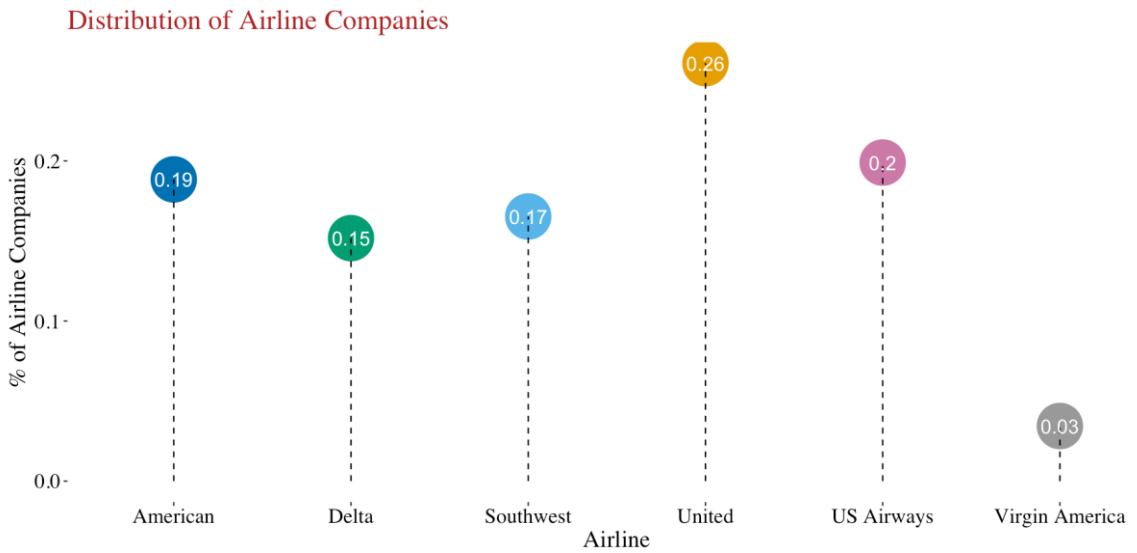


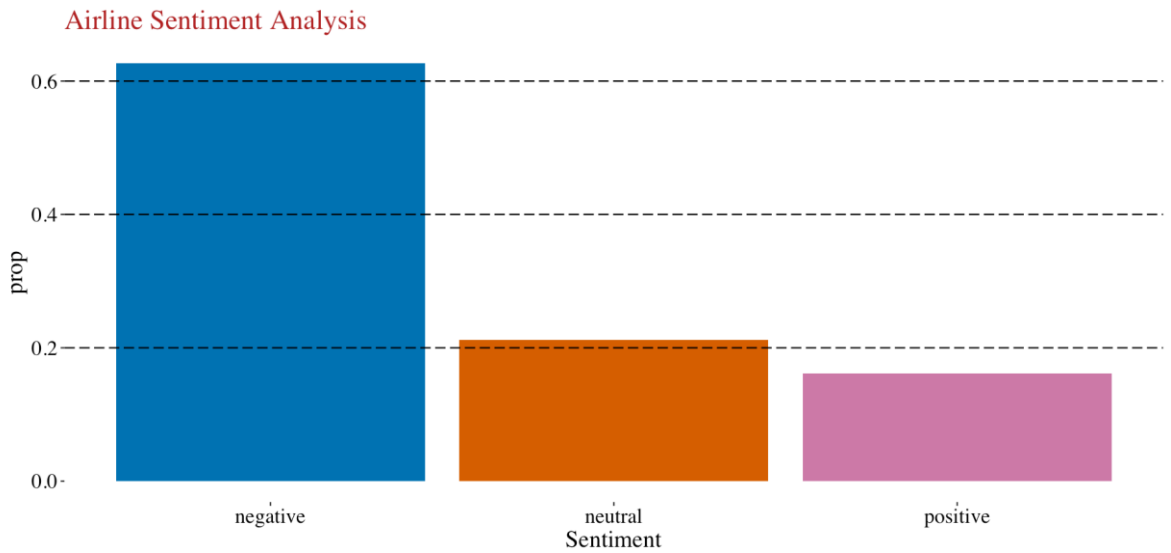
Airport Specific Analysis



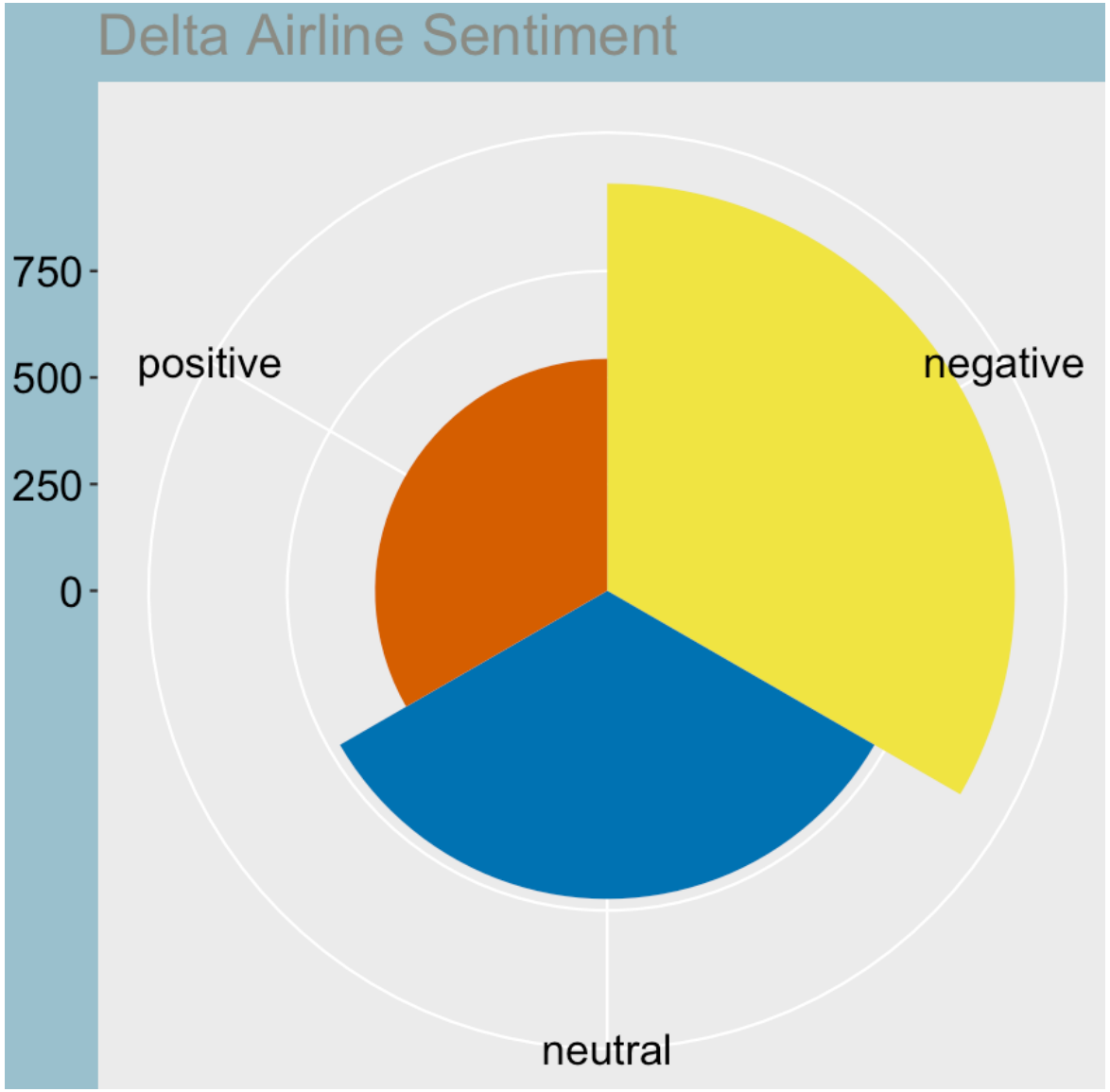
Twitter Sentiment

General Tweet Analysis





Airline Specific Tweet Analysis

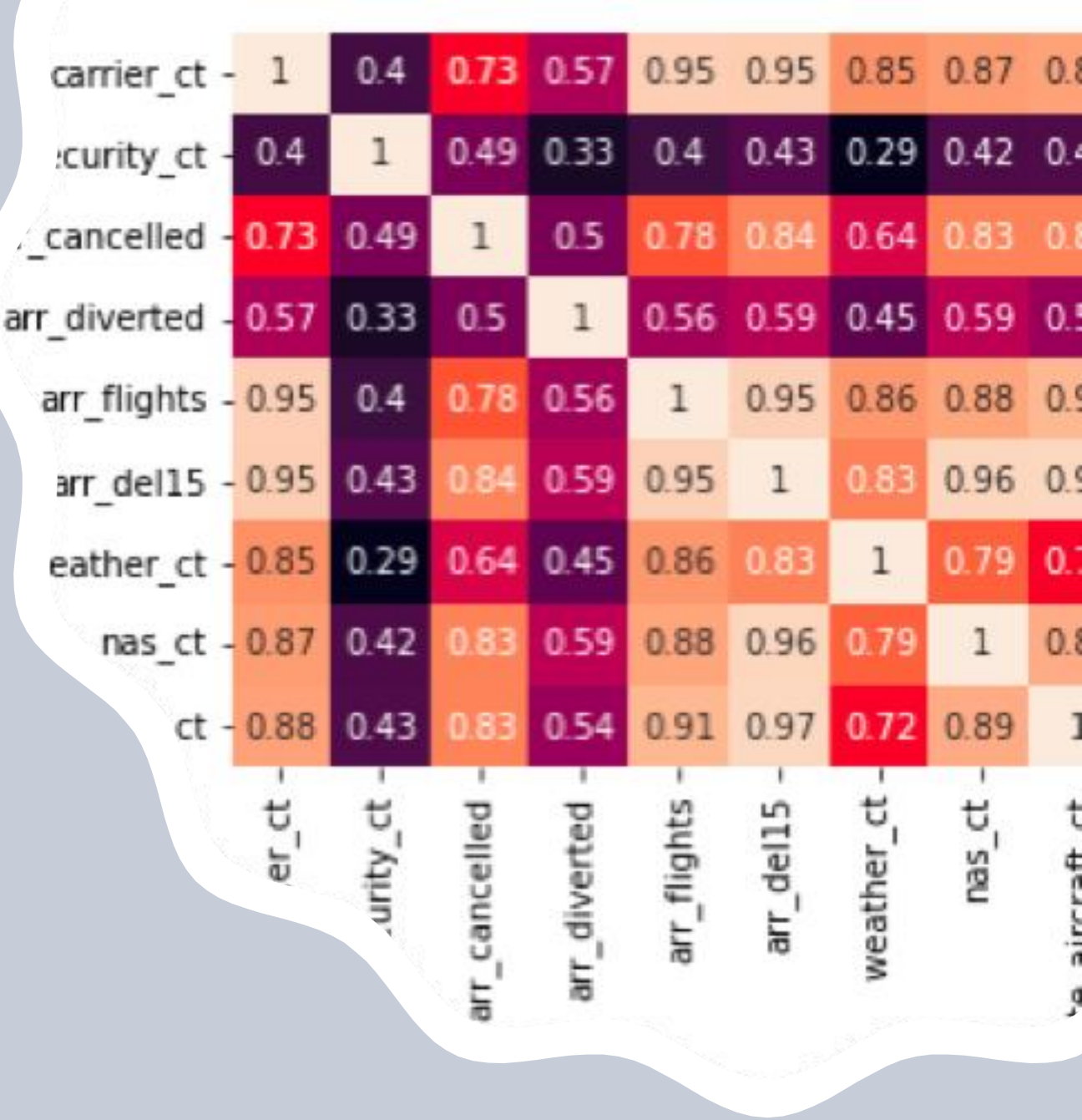


Data Mining Models

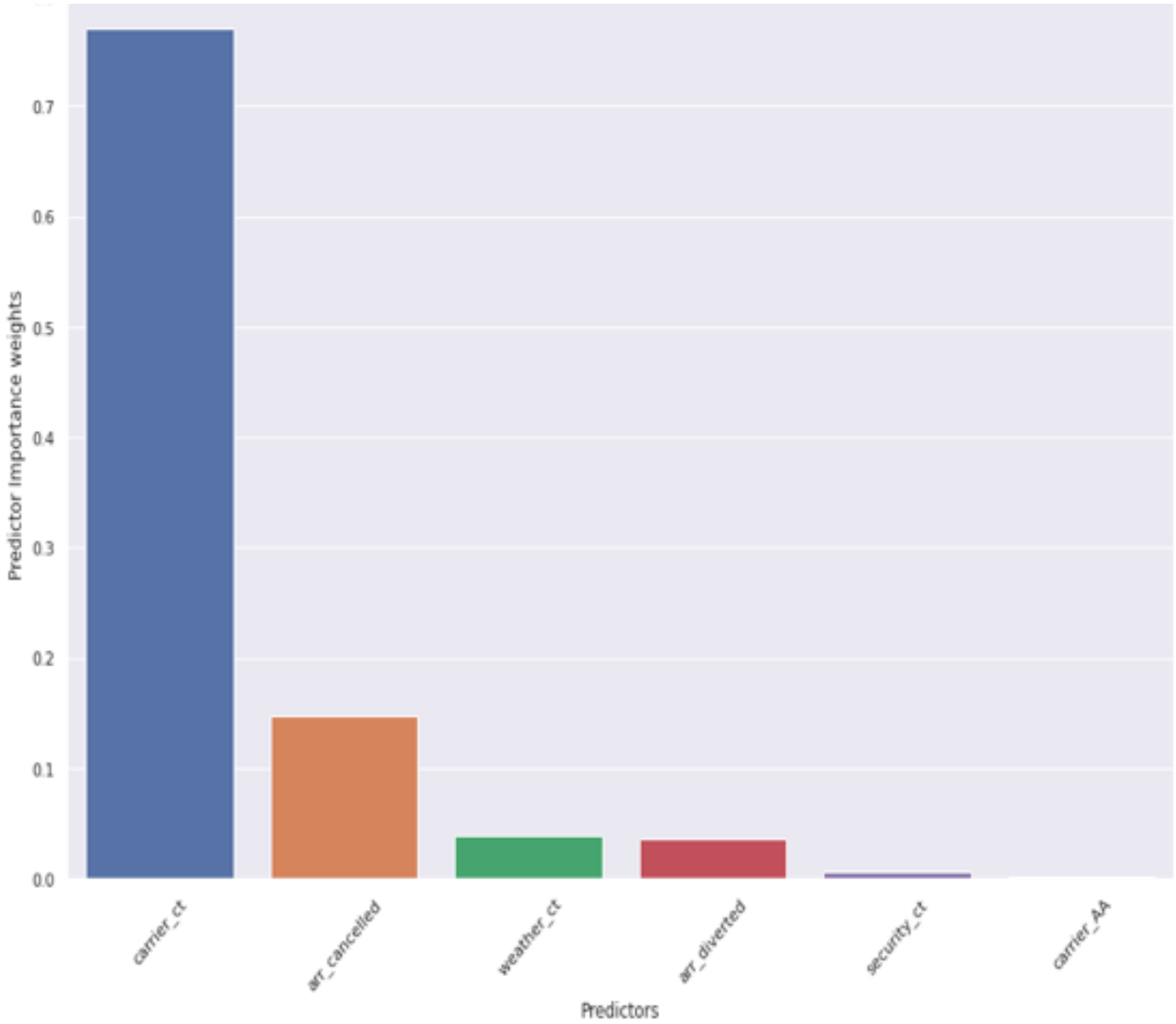
- Target Variable – arr_delay – The total delay of arrival
- Models used
  - Random Forest Classifier

Data processing and dimension reduction

- Correlation heat map was plotted across all numeric predictors
- Highly correlated pairs were removed to avoid overfitting
- 75% training and 25% testing combination for model building



Random Forest Classifier



Shiny App

Conclusion

We saw that twitter sentiments had some correlations to delayed/canceled flights; cannot prove any causation.

Using airline data, we can saw various trends based off different months and general patterns of different airlines and airports. This should help future passengers in planning their trips in order to minimize the chance of encountering delays or cancelled flights.

Team Contribution

All team members have equally contributed to this project in all the tasks from R-programming, shiny app implementation to project presentation.

Thank you!