Metal Birds 🎸🐥 (Group E)

Abigail Newbury (amn2217@cumc.columbia.edu), Andrey Zaznaev (az2732@cumc.columbia.edu), Hsi-Yu Chen (hc3448@cumc.columbia.edu), Aparajita Kashyap (ak4885@cumc.columbia.edu)

Brief description:

In this project, we plan to examine the patterns of flights across various airlines in the United States. We are interested in common flight patterns that airlines may take (since some airlines "own" certain routes), and how efficiently those flights occur (in terms of delayed departure and delayed arrival). This data lends itself to geographic and network-based analyses especially, since we can think of airports (spread across the US) as nodes and flights between them as edges, and the frequency of flights as edge weights. We additionally plan to conduct sentiment analysis on reviews of various airlines to better understand the reasons people may like or dislike certain airlines. With this study, we will be able to answer questions such as:

- *What are the most common flight patterns across the US?*
- *Which parts of the country have the worst delays?*
- *Which airlines do people hate the most?*
- *Do certain airport features, like year of establishment or proximity to nearest city center, affect incoming and outgoing flight patterns?*

Data Source:

Flight Data:

Kaggle datasets (note: we may use one, several, or all of these datasets):

1. https://www.kaggle.com/datasets/tylerx/flights-and-airports-data
2. https://www.kaggle.com/datasets/mahoora00135/flights
3. https://www.kaggle.com/datasets/open-flights/flight-route-database/data
4. https://www.kaggle.com/datasets/chaudharyanshul/airline-reviews (airline review data)

Airports dataset from Dept of Transportation: https://catalog.data.gov/dataset/airports-5e97a

Visualizations:

We plan to include "traditional" static visualizations, geographic visualizations, network, and text-based visualizations to answer our questions. For the question of common flight patterns, we plan on visualizing flight-path data as a network to demonstrate the popularity of flights connecting different airports. We will also produce a heatmap with the largest US cities plotted on both axes and cell colors representing the frequency of flights between each pair of cities. For the question of the worst delays by region in the US, we plan on displaying airports on a map of the US and including information on the frequency of delay and average delay time. We also

plan to include bar graphs demonstrating the frequency and extent of delays by airline. To answer the question of which airlines are least liked by customers, we will create word clouds with popular airline complaints and perform sentiment analysis. Lastly, to answer the fourth question, we plan to create scatterplots with information from the department of transportation using each airport as an individual data point.

Are there any significant hurdles that you have doubts about? Would not solving them render the project incomplete?

The main limitation of this project is the fact that we are combining multiple disparate datasets – so it may be difficult to tell a cohesive story about the airlines we choose to focus on. For example, most of the airline reviews are not from common American airlines, so we may not have sufficient data to tie the results from the reviews to the results from the locations and delays datasets. However, we should still be able to communicate important (and universal) information about what aspects of flying tend to go wrong for people.