Group Members: Azka Javaid and Caleb Ki (Group3)
Title: Aviation Incident Analysis

Motivation: 2 months after 9/11, American Airlines flight 587 crashed. Flight 587 was the second most deadliest aviation incident in American history. Yet, American Airlines flight 587 is generally forgotten today. Dozens of plane crashes happen every year, yet we never hear about them. Contrast that with the incidents concerning Malaysian Airlines flights in the past couple of year. The flight that went missing and the flight that was shot down received mass media coverage. What makes these flights different from the ones that go unreported?

Purpose: For our project, we would like to explore airplane crashes. In particular, we want to understand why certain airplane crashes get more news coverage than others. We plan to combine the aviation incident dataset obtained from OpenData titled, “Airplane Crashes and Fatalities Since 1908.” The dataset contains variables indicating Date of an airplane crash along with the Operator (Military/Air/Private), Number of people Aboard, Number of Fatalities, Location and Summary of the incident (whether the plane crashed/exploded/shot down/disappeared/pilot error along with other details like whether the plane was cargo). We plan to explore disparities between actual crash events (as measured by the OpenData) and the number of aviation incidents reported on news media (as gauged from the GDELT dataset). The main purpose of this analysis is to explore why certain aircraft crashes are reported more publicly than others.

Data: The data will come from two different sources. The first source of data will come from the Gdelt dataset, and the other is a public dataset “Airplane Crashes and Fatalities Since 1908” hosted by Open Data by Socrata.

Analytic Ideas/End Products: We would like to build a model to predict the rate at which an aviation incident is reported by news media average based on aircraft features like the Operator, Number Aboard, Number of Fatalities, Crash Location and Summary of the incident. The response variable, NumMentions/NumSources/NumArticles, will be provided by the GDELT dataset while the mentioned predictor variables will be supplied by the Aircraft OpenData. We will merge the two data sources in a way such that each row is a plane crash incident with associated features like Operator, Number Aboard, Number of Fatalities, Crash Location, Summary and NumMentions/NumSources/NumArticles. This model should also incorporate current news media coverage at the time of an aviation incident that can possibly overshadow certain aviation crashes (i.e. 9/11 incident that overshadowed the American Airlines flight crash described in the motivation above). BigQuery will be used to query the GDELT dataset to find relevant NumMentions regarding the aviation incidents.

Besides building a model to predict the rate at which a future aviation accident is reported by news media, we will also use Alchemy to perform text analytics to better understand news media coverage surrounding aviation incident by scraping the urls returned from the GDELT data. We would then perform clustering to better understand keywords regarding aviation

incident reporting. This can be translated in a ShinyApp where we would show top keywords/sentiment surrounding aviation coverage.