Airplane Crashes over Time

Rachel Kudzin

Due: 12/07/2024

Introduction

In this report, we will investigate trends about plane crashes, going back as early as 1908. Our dataset includes over 5,000 aviation accidents, including the most recent plane crashes from 2024. All data is scraped from Plane Crash Info, a website maintained by Richard Kebabjian.

Data Overview

As the Plane Crash Info website states on its database overview page, this dataset includes all (or maybe most, according to the website's disclaimer) aviation accidents that meet the following criteria:[1]

- All civil and commercial aviation accidents of scheduled and non-scheduled passenger airliners worldwide, which resulted in a fatality (including all U.S. Part 121 and Part 135 fatal accidents)
- All cargo, positioning, ferry and test flight fatal accidents.
- All military transport accidents with 10 or more fatalities.
- All commercial and military helicopter accidents with greater than 10 fatalities.
- All civil and military airship accidents involving fatalities.
- Aviation accidents involving the death of famous people.
- Aviation accidents or incidents of noteworthy interest.

Plane Crash Info includes information, if known, for the date and time of the accident, the airline/operator and the flight number, the route, aircraft type and other specifics (such as serial number), the total number aboard and fatalities, broken up by passengers and crew, total killed on the ground, and a text description of the accident. Not all of these values are full for all rows, but the dataset is reasonably extensive.

Data Preparation

We scraped all data available from the database into a clean dataframe. Our scraping function first calculates all subfiles in the database, and then concatenates the results from each page together into one dataframe. We also cached each page to ensure we would not need to unnecessarily run the scraping calls. This data is now available in the project repository, under the `data` folder.

After we saved the data from Plane Crash Info, we cleaned the data to ensure we could accurately investigate the information. First, we ran a drop duplicate command to ensure each row was unique, although there were no duplicate rows.

We then completed a series of formatting steps. We ensured that both dates and times were standardized and recognizable as such by `tidyverse`. We made sure times were in 24 hour form.[2] We also split the aboard and fatality columns into total, passengers, and crew. Finally, we ran queries through the `ArcGIS API` for

[1] The criteria in the report are copied from the Plane Crash Info database overview page

[2] All times are local to the crash, as specified in the database overview page

each location description. Nearly all locations returned latitude and longitude points, but for those that did not, we cleaned the names of descriptors like "near" and re-ran the queries. When all queries finished, only twelve filled locations failed to include latitude and longitude marks. These twelve descriptions that did not find a latitude and longitude match are since-renamed parts of the former USSR. The cleaned dataset is available in the `cleaned_data` folder.

Findings

Crashes Over Time

To begin our investigation, it makes sense to look at the trend of plane crashes over time. It seems to be a reasonable assumption that crashes have decreased, but we will need to see the specific information to confirm. Since there are not that many crashes overall, we will aggregate by year.

When we aggregate by year, we can see that there was a steep increase in fatal airplane crashes up until the 1950s. Part of this increase is likely due to a few things, including the increasing prevalence of plane travel over the 1900s and the two world wars, which involved heavy air fights. Interestingly, we see that from about 1945- the late 1990s, there were consistently higher numbers of plane crashes per year, even though the number was decreasing slightly overall. Then, around the 2000s, the number of fatal plane crashes steeply decreased, to nearly the same levels as the early 1900s, when planes were incredibly rare. It certainly appears to confirm our intuition that plane travel has always been relatively safe, and is certainly even more safe now than it ever has been before.

However, it is possible that aggregating up by year could hide important trends. Thus, we can also look at a heatmap of the number of crashes by month over the entire time period in the dataset.

As we saw before, the bars at each edge, representing the beginning and most recent months, have the fewest fatal crashes, with many having zero or one. Beyond this overall pattern, there does not seem to be a strong trend in the figure, beyond the overall trend of the years in the middle having more crashes, which we saw before. Perhaps January and December have more crashes, but there is not a strong overall structure in the data when it comes to monthly patterns.

Fatality Rates

In addition to the overall crash numbers, the dataset contains information about the total number of people onboard, broken into crew and passenger when possible, as well as the number of fatalities, also broken into crew and passenger when available. We can look into the death rate overall, as well as for crew and passengers.

When we look at the violin plots, it is clear that most crashes in our dataset have a complete fatality rate, with the majority of instances at above a 50% fatality rate. The most interesting trend in the graph appears to be that crew have a higher fatality rate than passengers, even if this rate is only slightly more likely.

We can, therefore, look at the difference in fatality rates between crews and passengers for the same crashes. There are some planes with no passengers, but there are never planes without a crew, so we should expect some rates to be negative solely due to this reality.

Perhaps unsurprisingly, most crashes have the same death rate between crew and passengers, which makes sense when there are no survivors. It is interesting to see that most of the rest of the data is relatively evenly spread around the 0 axis, suggesting that passengers and crew survive generally at similar rates, although crew do seem to die slightly more when there are survivors, even though this trend is still slight.

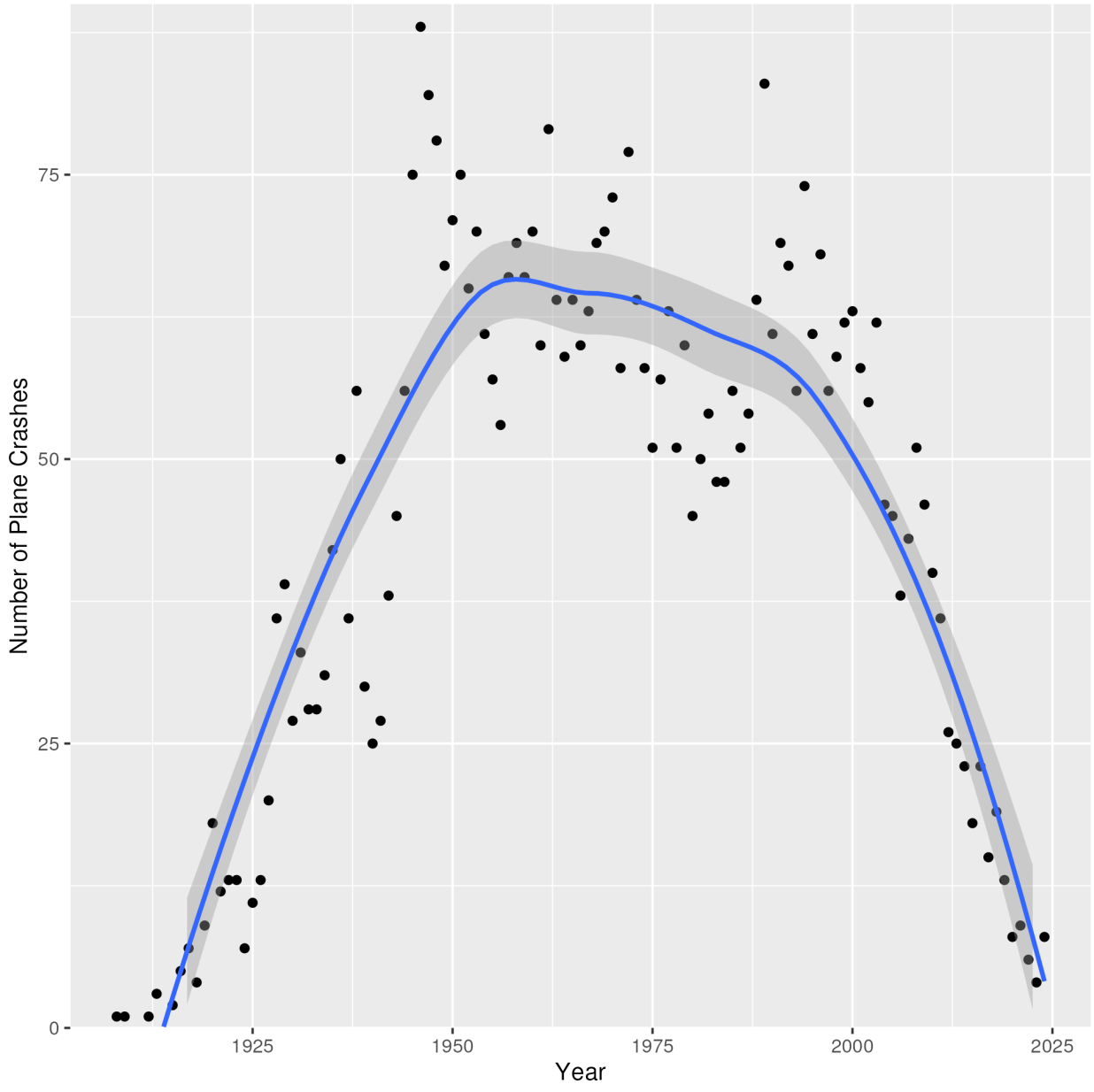

Figure 1: Number of Fatal Plane Crashes by Year

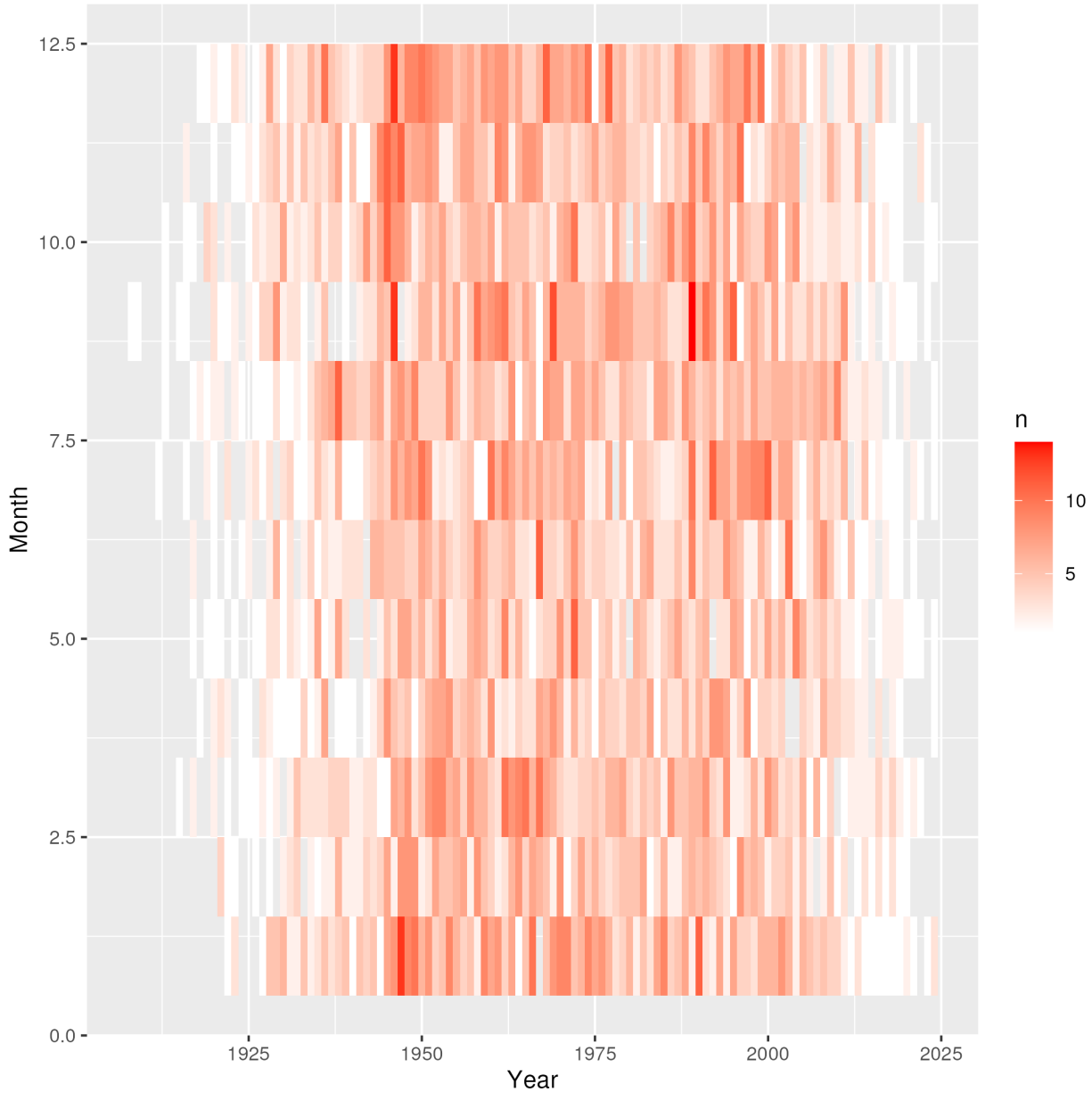

Figure 2: Fatal Plane Crashes By Month, Over Time

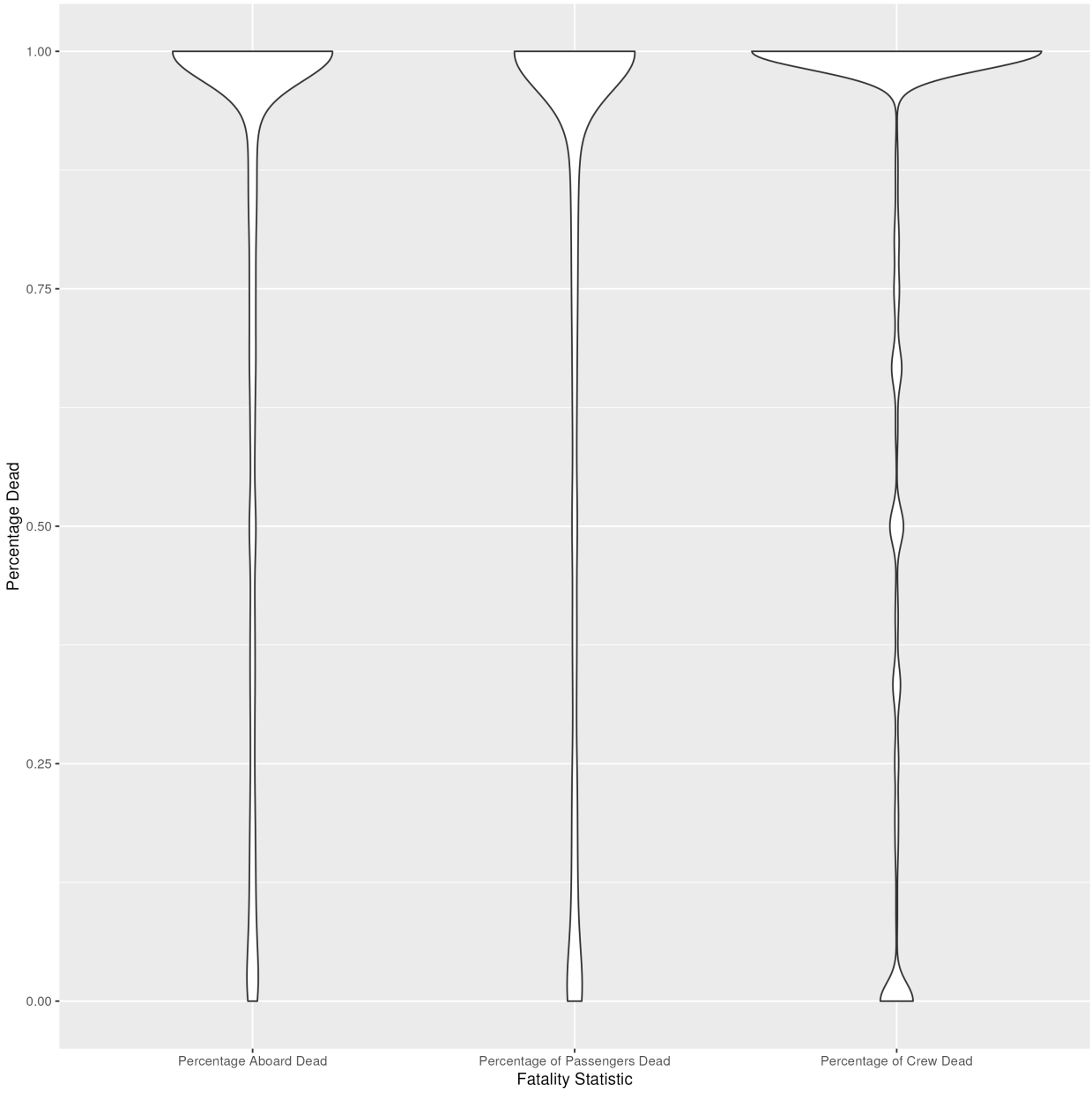

Figure 3: Violin Plots for Percentage of Fatalities by Crew and Passenger

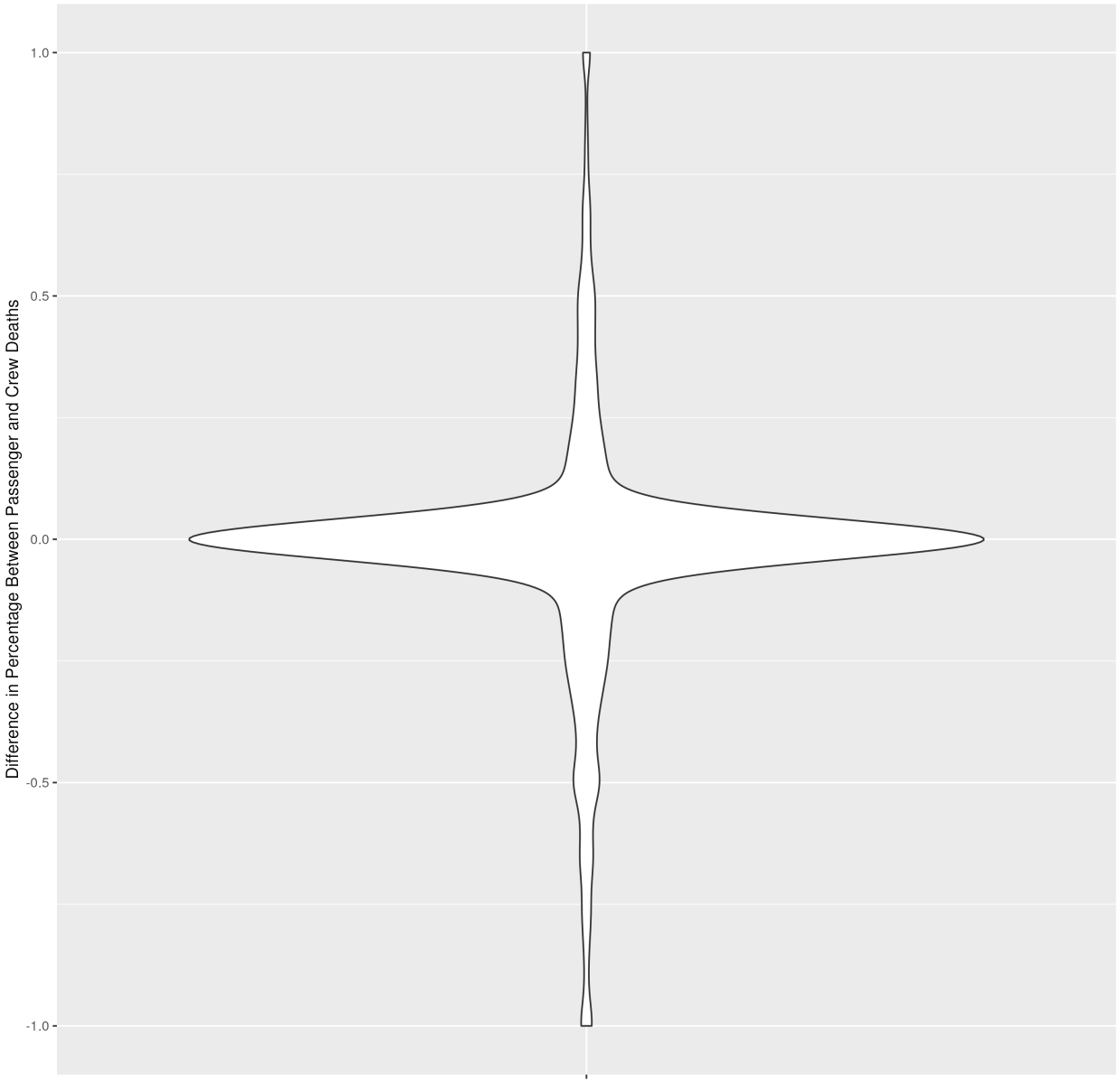

Figure 4: Difference in the Percentage of Passenger and Crew Deaths

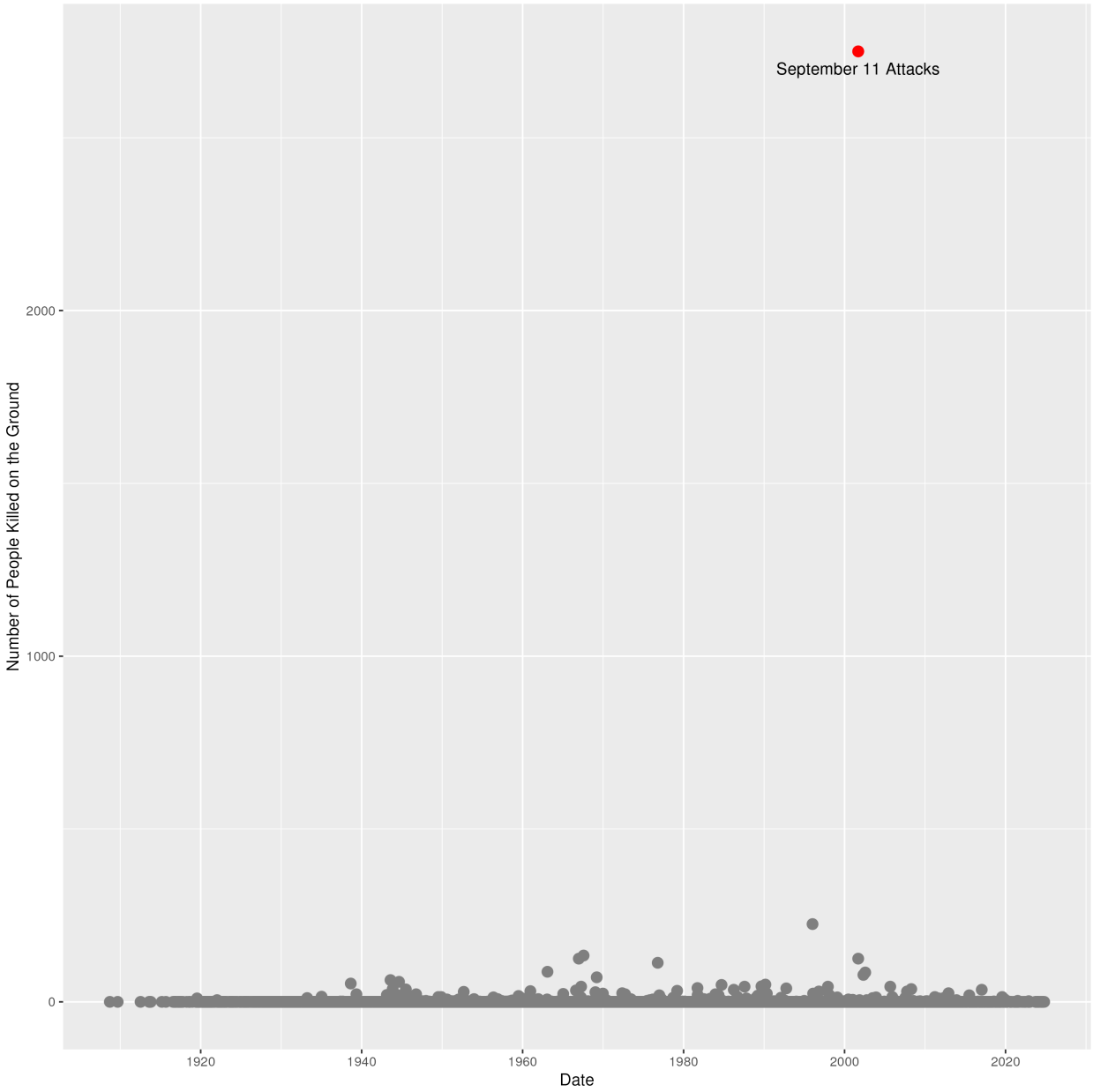

Figure 5: Number Killed on the Ground over Time

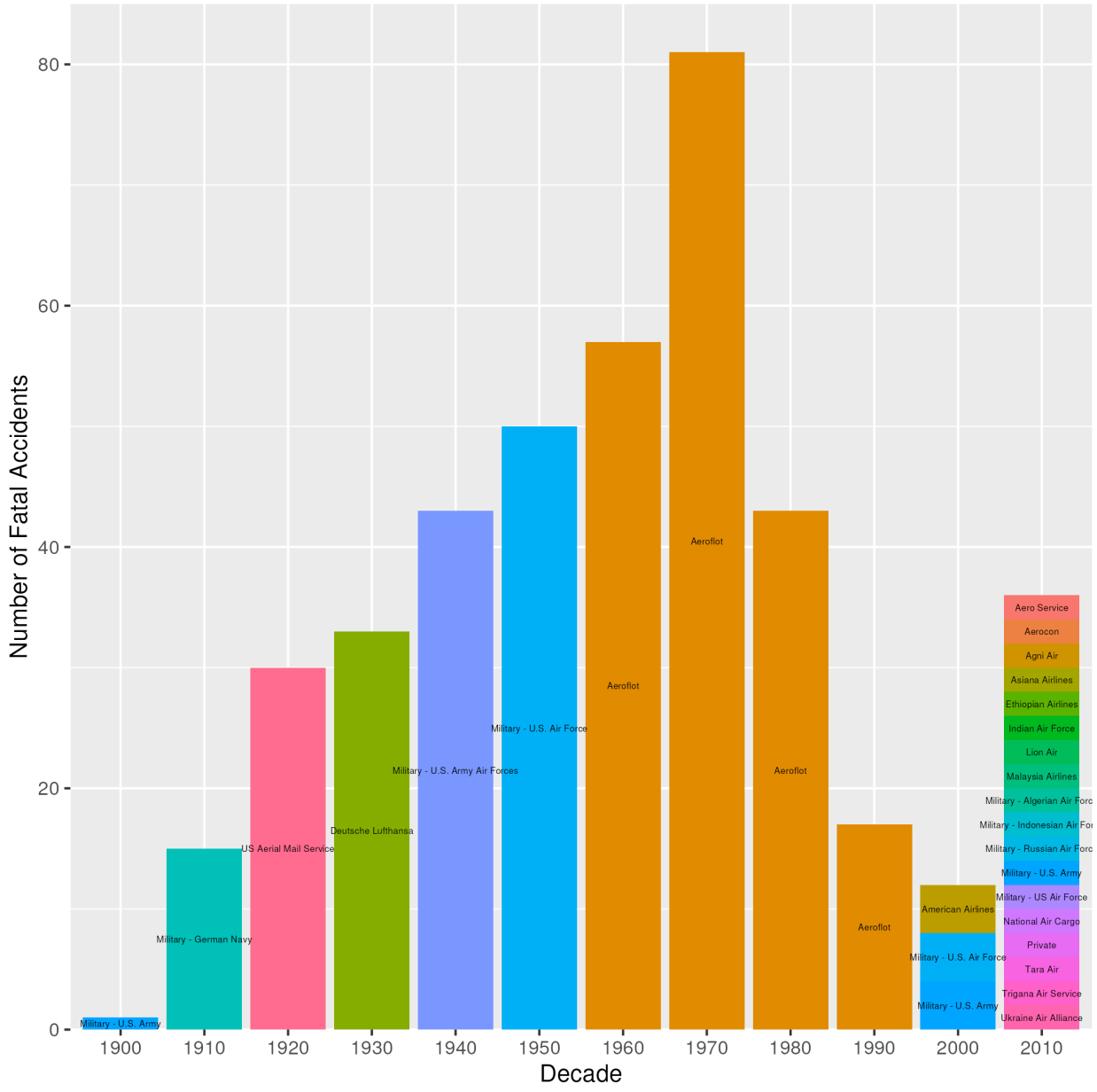

Figure 6: Airplane Operator with Most Accidents by Decade

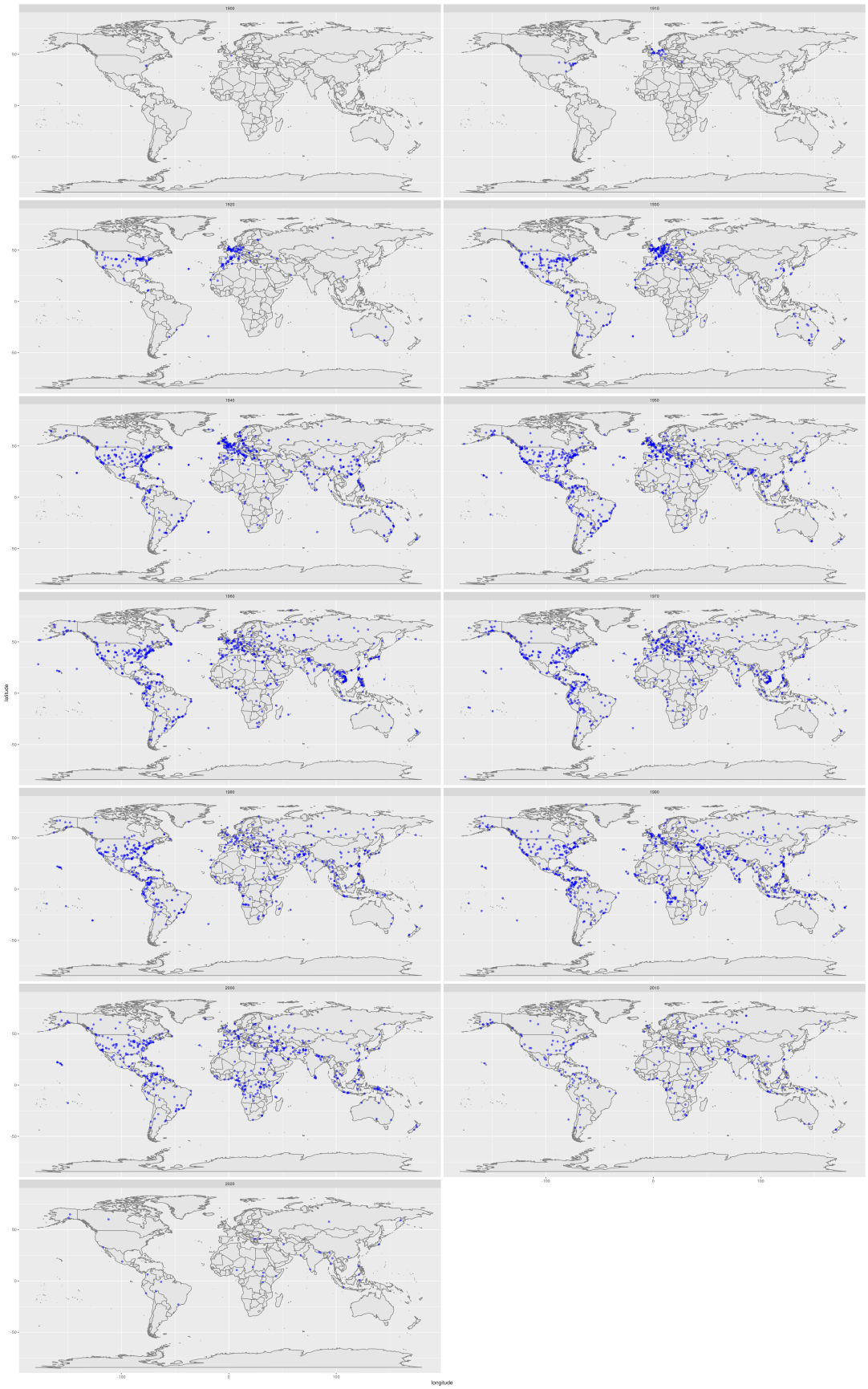

Figure 7: Locations of Airline Crashes

Figure 8: First Two Principle Components for Plane Crash Descriptions