

Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?

Group Project report using the Airlines Data

ROBERT COFFEY - 21399841

LOUIS THOMAS - 21315313

AARON CHURCH- 20331226

ERIN WEADICK



DS406 Topics in Data Science

Group Project

1. Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2. Method

We used the Airline Safety dataset provided for the project, which is also available on GitHub. The dataset contains safety records for a range of airlines across two time periods: 1985–1999 and 2000–2014. Each observation includes variables such as the number of incidents, fatal accidents, fatalities, and available seat kilometres per week (ASK for short), a measure of airline capacity.

We started with some Data Preparation. To support further comparisons, we mapped each airline to its corresponding country and continent. Inconsistent country names (e.g., “russian federation” or “Scandinavia”) were manually cleaned and standardized to lowercase for consistency.

To begin our analysis, we generated a pairwise correlation matrix using the `ggpairs()` function from the `GGally` package. Non-numeric columns (like: airline names, countries, and continents) were excluded from correlation analysis to focus on numeric relationships. To help visualize the relationships between key variables, names were manually renamed, and the plot was labelled. a background colour gradient was also added to highlight stronger correlations. As shown in Figure 5

We found that correlations between incidents or fatalities across time periods were generally weak. To explore this further, we examined the relationship between fatality counts in the two timeframes using a scatter plot, with log-transformed ASK used as a colour gradient to show airline size. This allowed us to investigate whether larger airlines showed different patterns in safety outcomes.

We continued looking at another Rate-Based Comparison calculating safety incident rates per billion ASK For both periods. We computed the Incident rate, Fatal accident rate, and Fatality rate. Each was calculated by dividing the relevant count by ASK and multiplying by 10^9 . These normalized rates allowed us to account for the different scale of each airline. We then focused on airlines that had more than 50 fatalities between 1985–1999 and computed the change in fatality rate from the earlier to the later period. These differences were visualized using a horizontal bar plot, with airlines ordered by the rate of change. This helped identify which higher risk airlines showed improvement or got worse in their safety performance over time.

Throughout the project, the `ggplot2` and `dplyr` packages were invaluable for data screening and visualization. Additionally, tools such as `GGally` (for pairwise exploration), `plotly` (for interactive graphics used in some behind the scenes analysis), and `tsibbledata` (for GDP and population augmentation) played key roles in our analysis.

3. Results

3.1 Correlation Matrix Analysis

Figure 5 shows a correlation matrix plot. This was made to check for any correlation between the airline data variables that were given between the periods 1985-1999 and 2000-2014. The variables included in the matrix were available seat kilometres flown every week, total number of incidents for 1985-1999 and 2000-2014, Total number of fatal accidents for 1985-1999 and 2000-2014, and total number of fatalities 1985-1999 and 2000-2014. From observing the graph there is a clear strong positive relationship ($\text{Corr} = 0.857$) between incidents and fatal accidents in the period 1985-1999. This shows that airlines within this period that had a higher number of incidents were also more likely to have fatal accidents. In the period 2000-2014 there was also a similar correlation for these variables also ($\text{Corr} = 0.598$) just slightly weaker. This plot confirmed that there was a correlation between incidents and fatal accidents. It can be seen also that there was a moderate correlation ($\text{Corr} = 0.726$) between the available seat kilometres flown every week and total number of incidents in the period 2000-2014 suggesting that airlines with greater flight volume are likely to have more incidents. This could be due to more use of the planes rather than poorer safety measures.

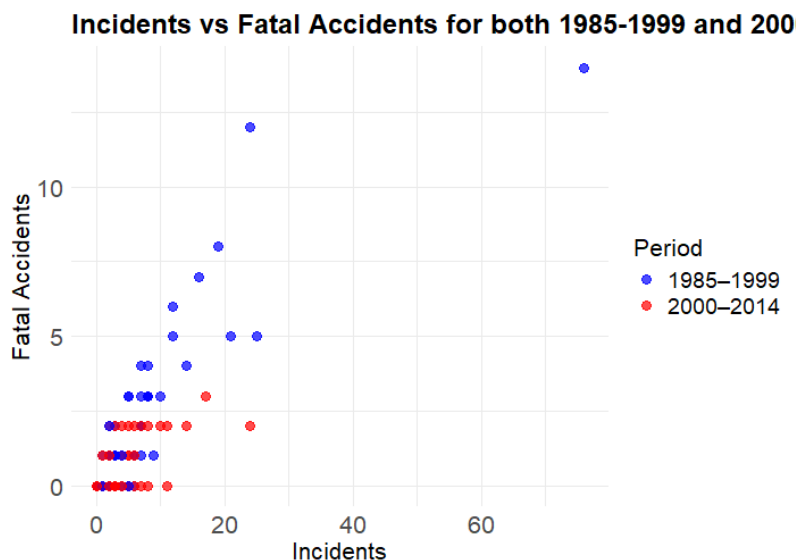


Figure 1: Plot of old fatalities versus new fatalities using GDP for

3.2 Comparing Incident Between Periods

In Figure 5 we found that there was a strong correlation between incidents and fatal accidents. Figure 1 shows that but for both periods 1985-1999 (blue) and 2000-2014 (red). Each data point represents a single airline or multiple as some have the same total number of incidents and fatal accidents. For 1985-1999 we can see that the data points are more widely distributed with multiple airlines showing high incident and fatal accident counts. An airline Aeroflot at (76 incidents and 14 fatal accidents) shown within this period that it was a significant safety risk.

For the 2000-2014 period the data points all cluster at the bottom left of the graph shown by the red data points. This indicates that the majority of airlines

experienced less incidents and fatal accidents showing an overall improvement in the safety of the airlines. These improvements may be due to the advancements in technology, and more safety measures. It was seen that Aeroflot that was the highest in the first period was no longer the highest in this period. It is also important to note that in both periods all american airlines (Delta/Northwest, United/Continental, US Airways/ American west, American) scored very high suggesting possibly american airlines are the least safest.

Overall there is evidence that there is a decrease in the total number of incidents and fatal accidents from the period 1985-1999 to 2000-2014 suggesting an improvement in airline safety.

This bar chart represents the fatalities per ASK (Available Seat Kilometer) change from 1985-1999 to 2000-2014 across the airlines. From looking at the graph it is evident that the majority of airlines showed a decrease in fatalities per ASK.

This indicates improved safety performance correlative to the volume of operations and suggests that even though fatalities still may occur for some airlines they are less frequent. Avianca and Pakistan international airlines both showed the largest decrease in fatalities per ASK. Airlines such as Air France and American did show an increase in fatalities per ASK which is alarming since the rest of the airlines showed a decreased. Overall this graph shows which airline safety outcomes have improved and which had not. It is also interesting to note that For all the American airlines as mentioned above in Figure 1 there has not been much improvement.

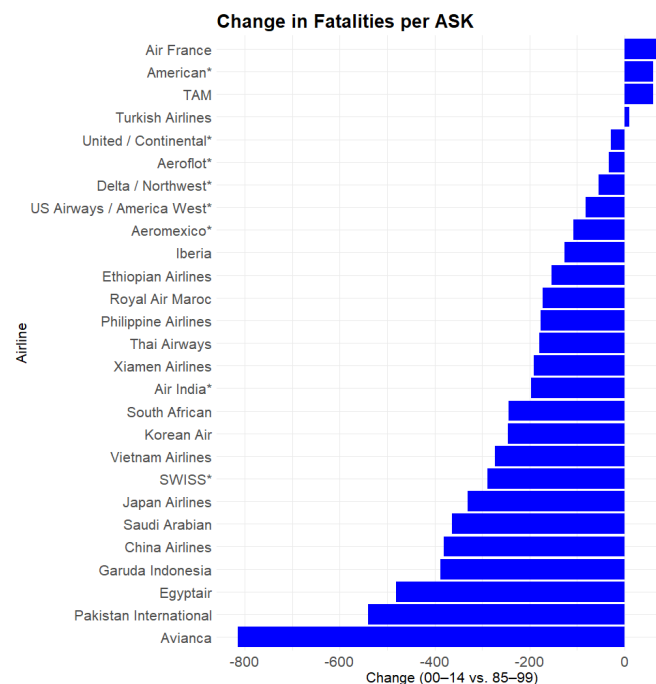


Figure 2: Plot of old fatalities versus new fatalities using GDP for

This bar chart, Figure 2, represents the fatalities per ASK (Available Seat Kilometer) change from 1985-1999 to 2000-2014 across the airlines. From looking at the graph it is evident that the majority of airlines showed a decrease in fatalities per ASK. This indicates improved safety performance correlative to the volume of operations and suggests that even though fatalities still may occur for some airlines they are less frequent. Avianca and Pakistan international airlines both showed the largest decrease in fatalities per ASK. Airlines such as Air France and American did show an increase in fatalities per ASK which is alarming since the rest of the airlines showed a decreased. Overall this graph shows which airline safety outcomes have improved and which hadn't. It is also interesting to note that For all the American airlines as mentioned above in figure 2 there hasn't been much improvement.

4. Discussion

The best message that we get from the results is the fact that, fatalities are decreasing. This is a really good trend to see, as it shows us that airlines are getting safer, made clear from Figure 1* and Figure 2 plots. The variability of fatalities between airlines have gotten better. These results are similar to the results the article came to, where there isn't much of a sign of past fatalities implying future fatalities, and how global fatal accident rate has decreased significantly over the few decades.

4.1 Fatalities and ASK

Looking at our correlation plot, Figure 5. We realised there isn't much of a correlation between old and new incidents. Some variables would have been impossible to find a relationship between, for example - new fatal incidents. So we created plots of old Fatal Accidents and new Fatal Accidents, with Fatalities, in relation to the ASK variable (available seat km per week). As shown in Figure 3.

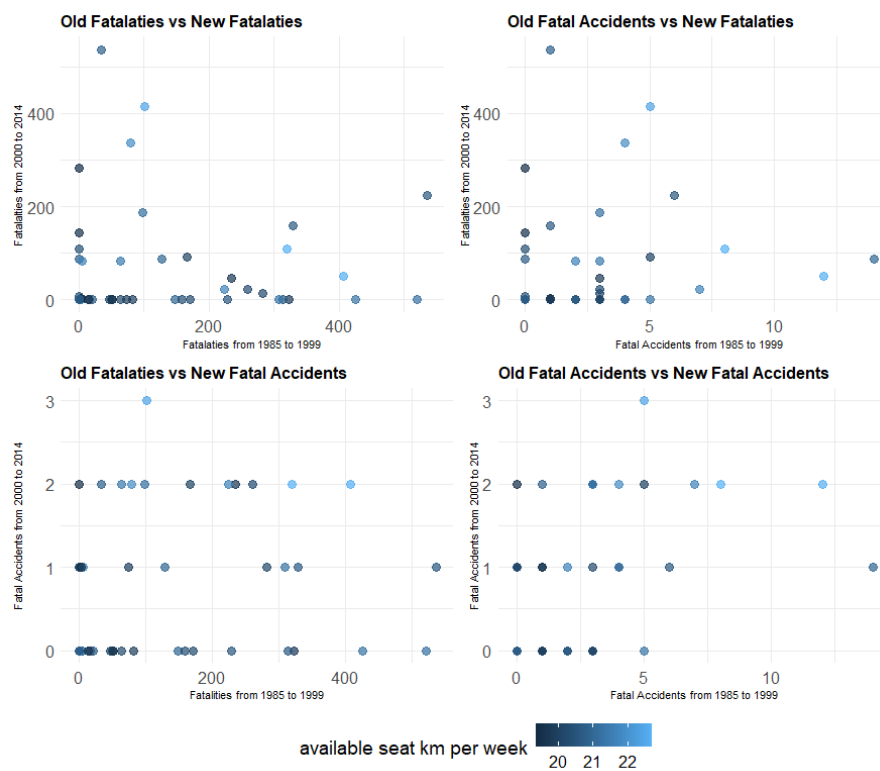


Figure 3: Plot of old fatalities versus new fatalities using logged available seat km per week

The main trends is that not much airlines are in the upper right quadrant. Where you would high old and new fatalities or high new and old fatal accidents. These can be seen in the first to plots in the **fig 1**. What we are seeing in the top-left and bottom-right quadrants are more volatile airlines. Where they would have high and low fatalities from one period to the other.

Where the the lighter blue would be higher ASK variables. Where there seems to be some interesting trends saying that bigger the seat per distance the larger the fatal-accidents. We found that there might be issues with the ASK variable, mainly being the fact that it would not be constant over the 30 year period. Which is further supported with data points on top-left and bottom-right quadrants. Furthermore, saying that more seats on means more fatalities is a little redundant.

Causing us to explore alternatives to ASK. Variables, that could reflect the safety of an airline, would be the amount of money spent on safety. Here due to constraints we couldn't get this value, instead we opted for GDP and Population. Where GDP would capture disposable income of a country, as well as the limit of how much an airline can spend. As shown in Figure 4

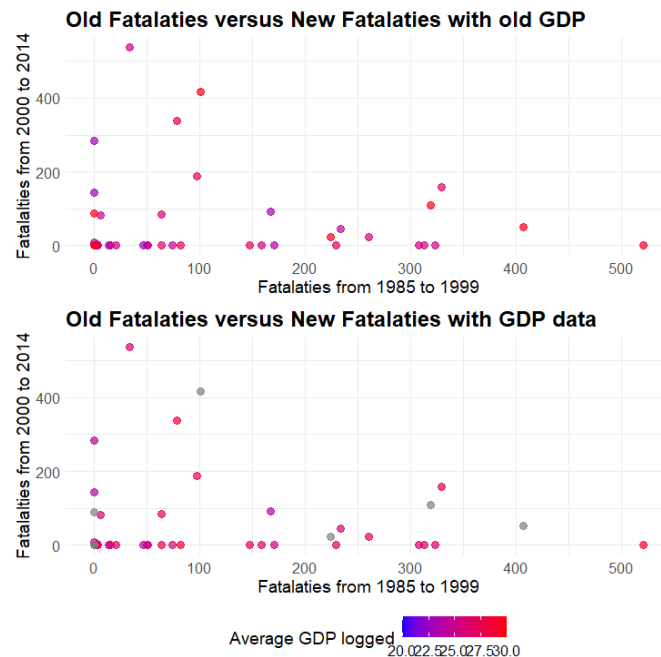


Figure 4: Plot of old fatalities versus new fatalities using logged available seat km per week

There are some interesting trends. Where higher old fatalities and low new fatalities would have higher old GDP means. There seemingly is less of a trend for the new GDP plot.

But what seems to be clear is that we have a similar trend to the ASK plots. Where lower GDP would be around the low-left quadrant, i.e. low new and old fatalities.

Probably showing us the same information as ASK, where less possible flying resulted in fewer deaths. As discussed before previous fatalities and metrics will not accurately predict future fatalities. More data driven research is needed by using audit data, airline ratings data and whistle-blower information. **Finally**, looking at Figure 3, we can see that modern fatal-accidents have significantly reduced. Making flying with these airlines a very safe experience.

5. Appendix

For all the images generated you can find the code in the reports section of the [github](#). There are also more plots we did; but due to constraints not talked about further.

5.1 Correlation Matrix Plot

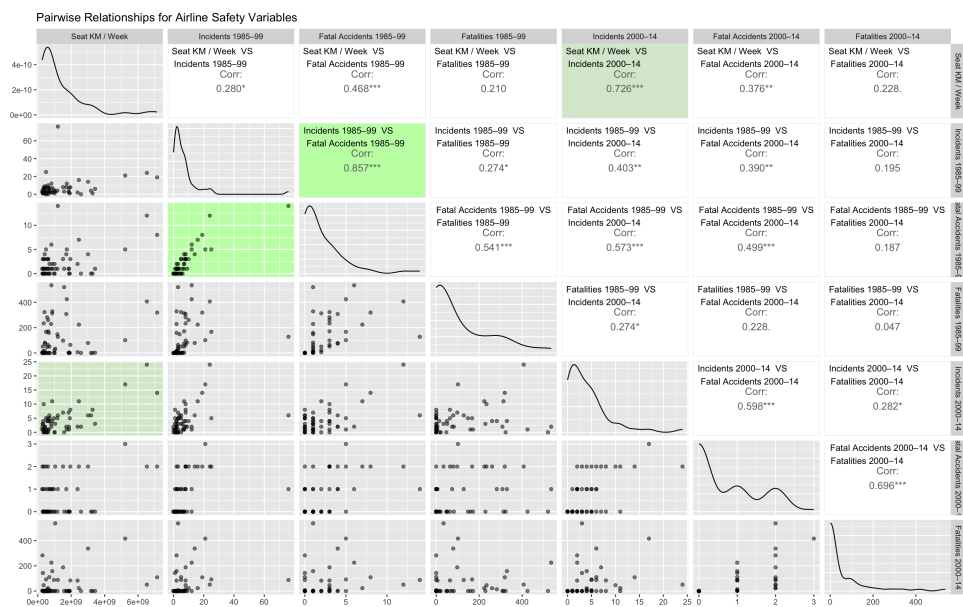


Figure 5: Correlation Matrix Plot