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

ERIN

DS406 Topics in Data Science

Group Project

1. Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer 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 kilometers 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 GGallypackage. 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 1

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. Report

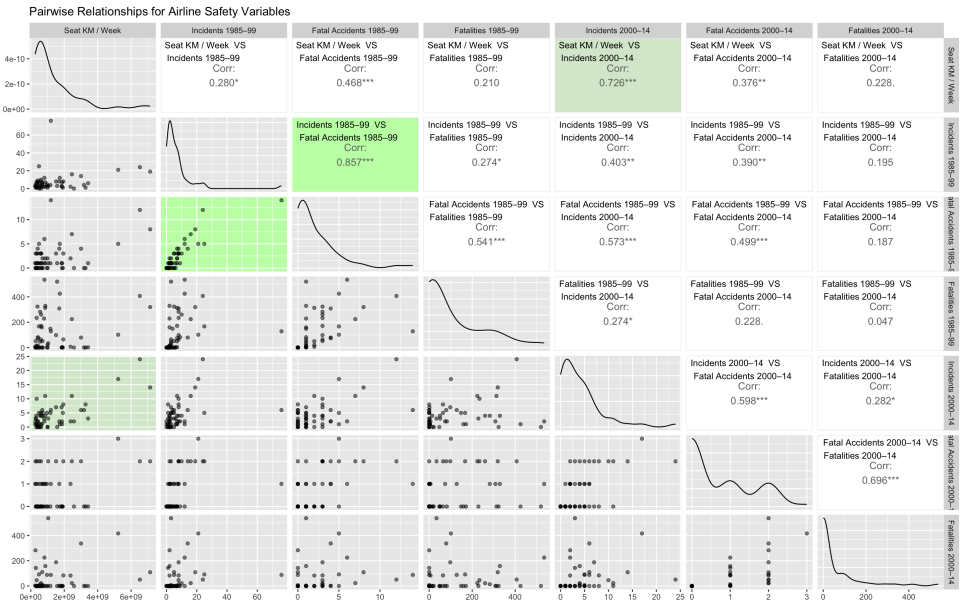

Figure 1: *Plot of old fatalities versus new fatalities using GDP for*

Figure 1 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 (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 (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 (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.

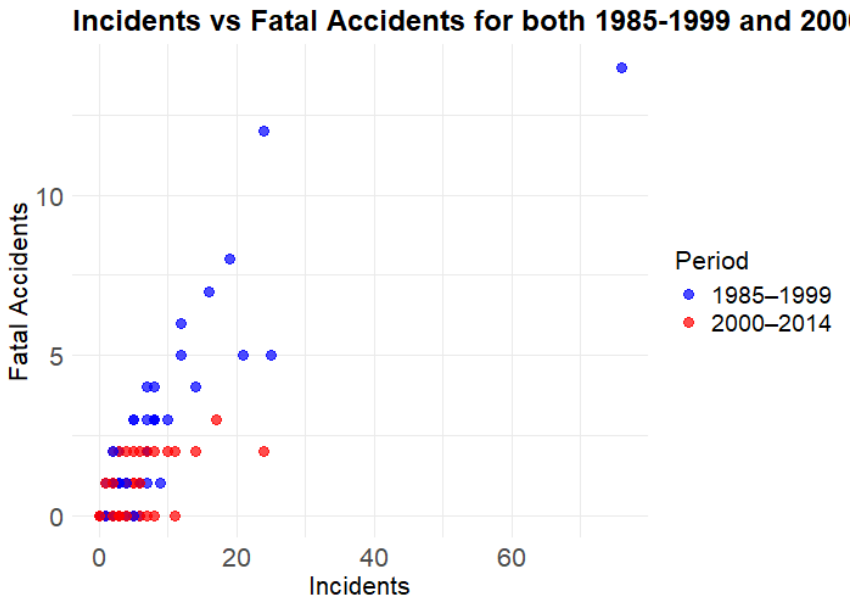

Figure 2: *Plot of old fatalities versus new fatalities using GDP for*

4. Discussion

4.1 Initial

Looking at our correlation plot, Figure 1. We realised there much of a correlation between old and new incidents. And some data would have been impossible to find a meaningful relation between variables, for example - new fatal incidents. So we created plots of old Fatal Accidents and new Fatal Accidents, with Fatalities. As shown in Figure 3.

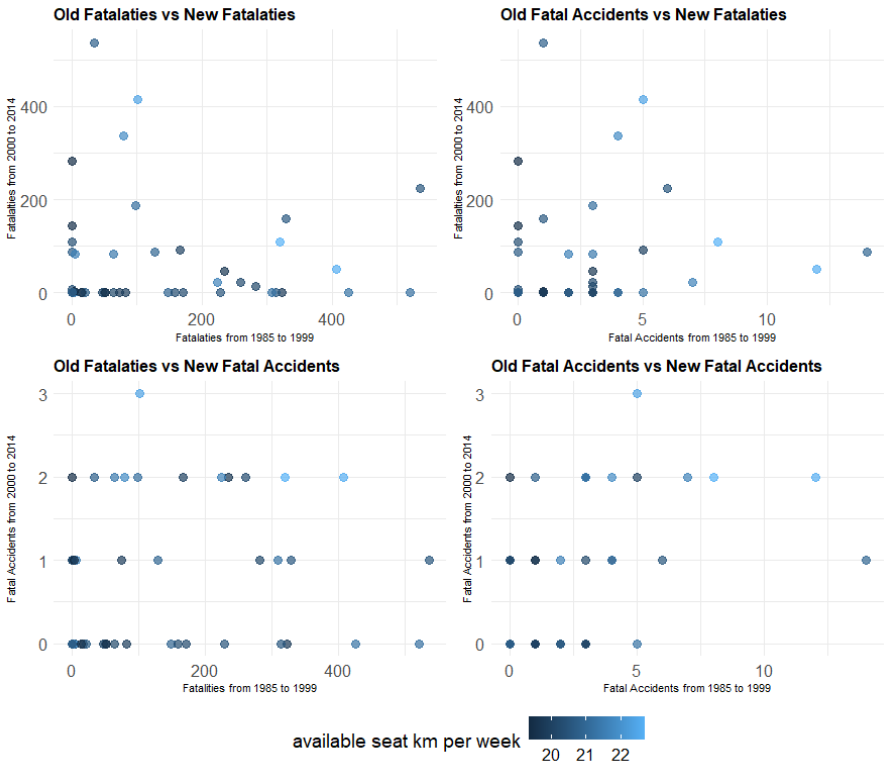

Figure 3: *Plot of old fatalities versus new fatalities using logged available seat km per week*

Where the the lighter blue would be higher available seat km per week. Where there seems to be some interesting trends saying that bigger the seat per distance the larger the fatal-accidents. Two issues with this is that we don't know much available seat per km (ASK) this would definitely not be constant over the 30 years. Which can be seen in the plots, where near 0 old fatalities have high new fatalities and visa versa.

Results: Not much of a correlation to work off of. Resulting in the interest in GDP data. We can extract this from `tsibbledata`. And merge it with relevant countries. We aren't saying much by saying more seats means that there were more fatalities.

Now that we have this data let's plot to see if there is any interesting trends from GDP and fatalities or incident data: For reasons presented above we will be dropping plots with new and old fatal accidents data. As shown in Figure 4

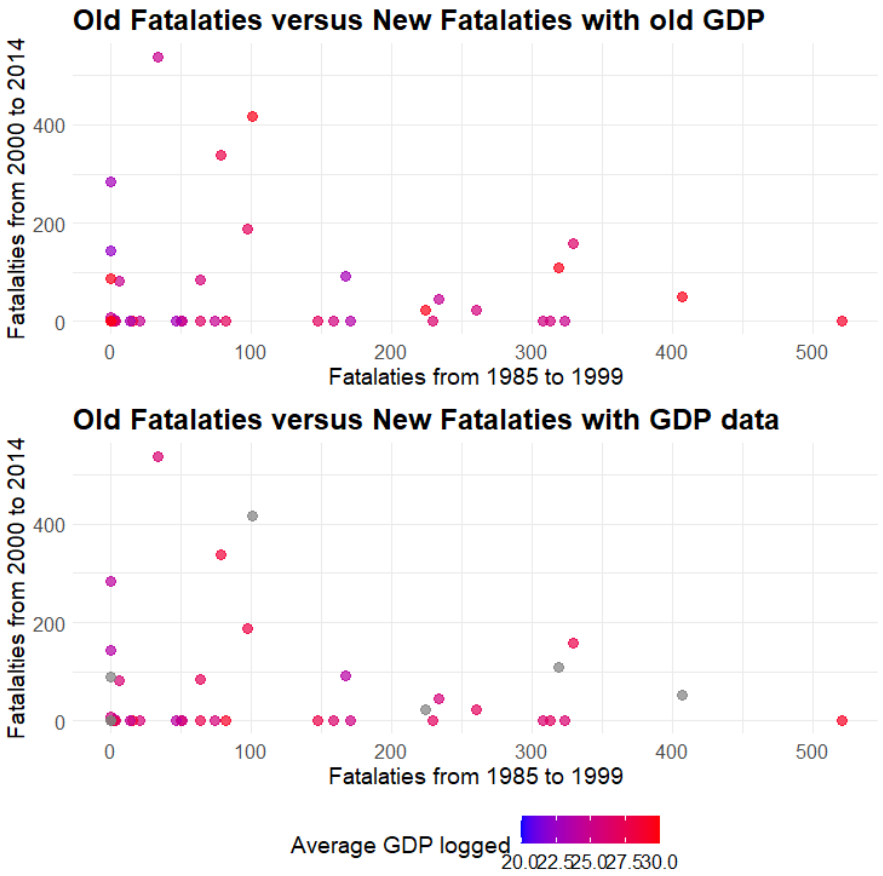

Figure 4: *Plot of old fatalities versus new fatalities using logged available seat km per week*

There is no clear pattern involved with the countries GDP. Airlines with high old fatalities and high new fatalities aren't correlation with any gdp.