Project 2 - Airline Safety

Zachary Palmore

10/2/2020

Objective

In this part of project 2, we will find out which airline is the most lethal, of selected airlines. This file contains five airlines with the number of fatalities from accidents in the intervals 1985 to 1999 and from 2000 to 2014. Our objective here is to;

- Quantify the total fatalities from accidents over 1985 - 2014 for each airline
- Find the highest and lowest number of fatalities from accidents over the entire duration
- Check for any trends in the fatalities from accidents in the intervals 1985 - 1999 and 2000 - 2014

To get those answers, the data needs to be cleaned and tidied before we can make use of it. We will start by importing the data.

```
require(tidyverse)
require(magick)
require(tesseract)
require(reshape2)
```

Importing

The data we are using is a small sample of the real data. The original can be used for comparison once imported. It is imported using a link to the GitHub repository maintained by FiveThirtyEight.

```
raw <- read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/airline-safety/airline-sa
View(raw)

original <- raw %>%
  filter(airline == "Ethiopian Airlines" |
           airline == "Garuda Indonesia" |
           airline == "Pakistan International" |
           airline == "TAM" |
           airline == "Turkish Airlines")
```

The sample was given in a table but the image was difficult to save. Instead of taking a screenshot of the image and trying to read it in as it was, the information was copied into a spreadsheet first. Then, a snip was taken of the spreadsheet which we will use to pull the data in for analysis.

As a joint photographics group (.JPG) it is of little functional use to us. All the data is stored in pixels and cannot be accessed until the characters we can see are extracted. But before we try that we need to import the information into some functional format. We should also clean the image and prepare it for tidying. There is a package called *magick* that can help us do that.

```
# Turning the jpg into a 'magick' format
airlines <- image_read("airlines.jpg") %>%
  # Creating a transparent background -
  # Because of the white in the image
  # this lightly darkens lines and characters
  image_negate() %>%
  image_transparent("transparent", fuzz=-10) %>%
  image_background("transparent") %>%
  # Convert the block magick format into readable letters
  # Using contrast and the negative of the image
  # Encircle the character edges and thin where possible
  image_morphology(method = "Thinning",
                   kernel = "Rectangle") %>%
  # Enhance contrast again
  image_negate() %>%
  # Cut off what is not needed in the image
  image_crop(geometry_area(0, 0, 2, 2))
# View the image stats
airlines
```

airline	fatal_accidents_85_99	fatal_accidents_00_14
Ethiopian-Airlines	5	2
Garuda-Indonesia	3	2
Pakistan-International	3	2
TAM	3	2
Turkish-Airlines	3	2

This image has now been imported from its jpg format and is ready for tidying; as if by, shall we say, magic? Thankfully, there were no changes to the data within the image. From here, we can use another function to read and extract the information into a data frame.

Tidying

Ultimately, we need to be able to share this data as a spreadsheet (specifically a .csv). To do so, we can use another package called *tesseract* which uses optical character recognition to identify and extract individual characters of an image. With our magick alterations, the image should be clear enough to run it through the tidying machine.

```
## Warning: 5 parsing failures.
## row col  expected    actual         file
##   1  -- 4 columns 3 columns literal data
##   2  -- 4 columns 3 columns literal data
##   3  -- 4 columns 3 columns literal data
##   4  -- 4 columns 3 columns literal data
##   5  -- 4 columns 3 columns literal data
```

```
## [1] 5 3 3 3 3
```

```
## [1] 2 2 2 2 2
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## Using Airline as id variables
```

At this point, we could use the function write.csv() to create a csv from the data frame. For purposes of review, this is left to the discretion of the user. For comparison, we can see the original raw data and visually see if there are any differences.

```
# The original data with selected airlines
original[,c(1,4,7)]
```

```
##                  airline fatal_accidents_85_99 fatal_accidents_00_14
## 1     Ethiopian Airlines                     5                     2
## 2       Garuda Indonesia                     3                     2
## 3 Pakistan International                     3                     2
## 4                    TAM                     3                     2
## 5       Turkish Airlines                     3                     2
```

```
# The imported and tidyed data
head(airsafety, 4)
```

```
##                  Airline      Time Fatalities
## 1     Ethiopian-Airlines 1985-1999          5
## 2       Garuda-Indonesia 1985-1999          3
## 3 Pakistan-International 1985-1999          3
## 4                    TAM 1985-1999          3
```

Thanks to the ORC, that was far fewer steps than trying to enter the data by hand. Now that the data is parsed through with all information extracted identically to how it was in the image, we can begin the analysis.

Analysis

Recall that we wanted to know how many fatalities occurred with each selected airline and which one had the worst record. It might also be interesting to sort them by most to least dangerous, and see if there are any trends in the fatalities from the intervals 1985-1999 and 2000-2014. It might be helpful to start with some basic statistics of the fatalities overall.

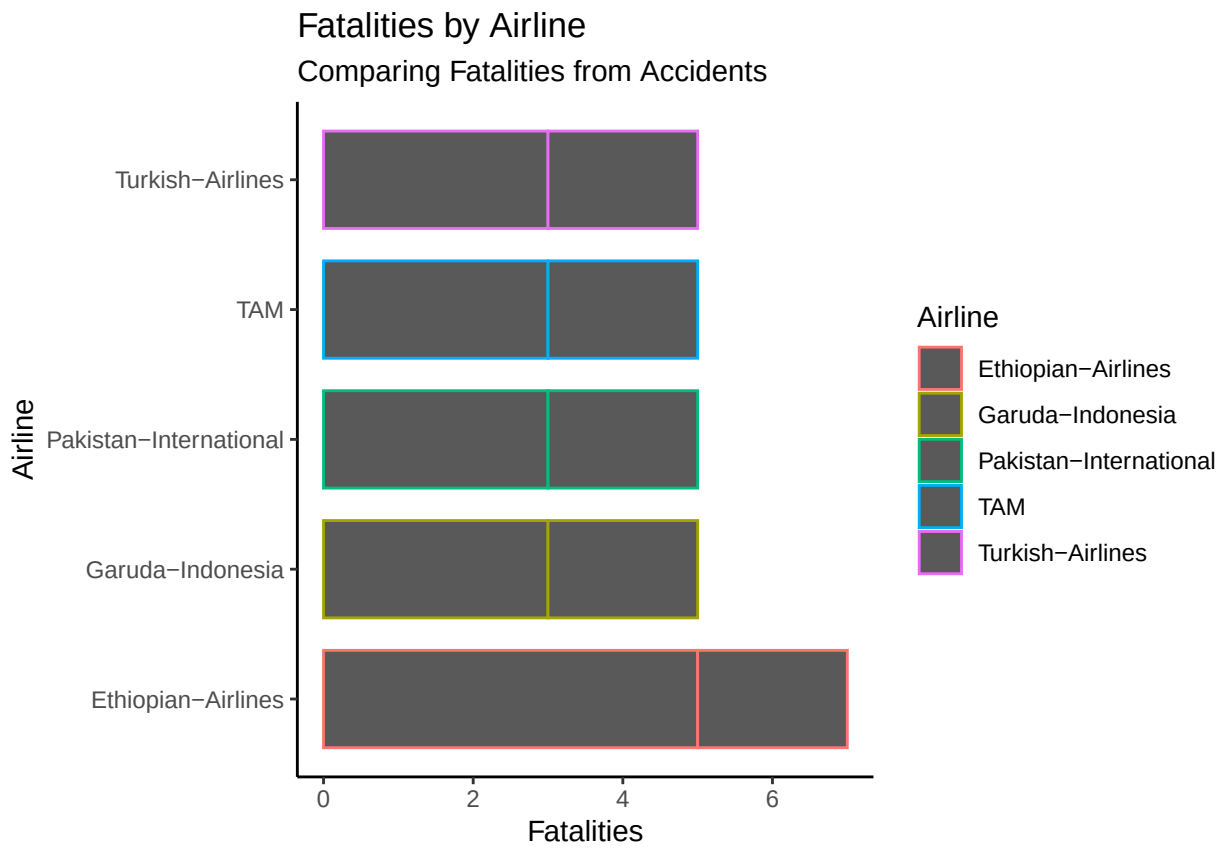
```
summary(airsafety$Fatalities)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0     2.0     2.5     2.7     3.0     5.0
```

That gives us a maximum number of fatalities over the 28 year span of 5.0 and a minimum of 2.0. It is surprising that none of these airlines had zero fatalities given that airlines are one of the safest ways to travel (statistically speaking). We can also see the average (mean) number of fatalities from these airlines is 2.7, while the median is 2.5. There must be an airline that is pulling the average fatalities slightly upwards.

Now let's see how each airline fared against one another. We asked the question, who had the most fatalities over the 28 year duration? To find out, we can use a horizontal bar graph.

```
ggplot(data = airsafety, aes(x= Airline, y = Fatalities, col = Airline)) + geom_bar(stat = "identity", w
  theme_classic() +
  labs(title = "Fatalities by Airline",
       subtitle = "Comparing Fatalities from Accidents",
       x = "Airline",
       y = "Fatalities") + coord_flip()
```



This shows a clear difference. The airline with the highest number of fatalities from accidents is Ethiopian Airlines. They had 7 fatalities while the rest of the airlines had 5.

- Quantify the total fatalities from accidents over 1985 - 2014 for each airline
- Find the highest and lowest number of fatalities from accidents over the entire duration
- Check for any trends in the fatalities from accidents in the intervals 1985 - 1999 and 2000 - 2014

We also made it an objective to find any changes in the fatalities from accidents

```
# Filter data by time intervals
Interval1 <- airsafety %>%
  filter(Time == "1985-1999")
# Selecting where time is 2000-2014
Interval2 <- airsafety %>%
  filter(Time == "2000-2014")
summary(Interval1)
```

```
##    Airline                Time     Fatalities
##  Length:5           1985-1999:5   Min.   :3.0
##  Class :character   2000-2014:0   1st Qu.:3.0
##  Mode  :character                 Median :3.0
##                                   Mean   :3.4
##                                   3rd Qu.:3.0
##                                   Max.   :5.0
```

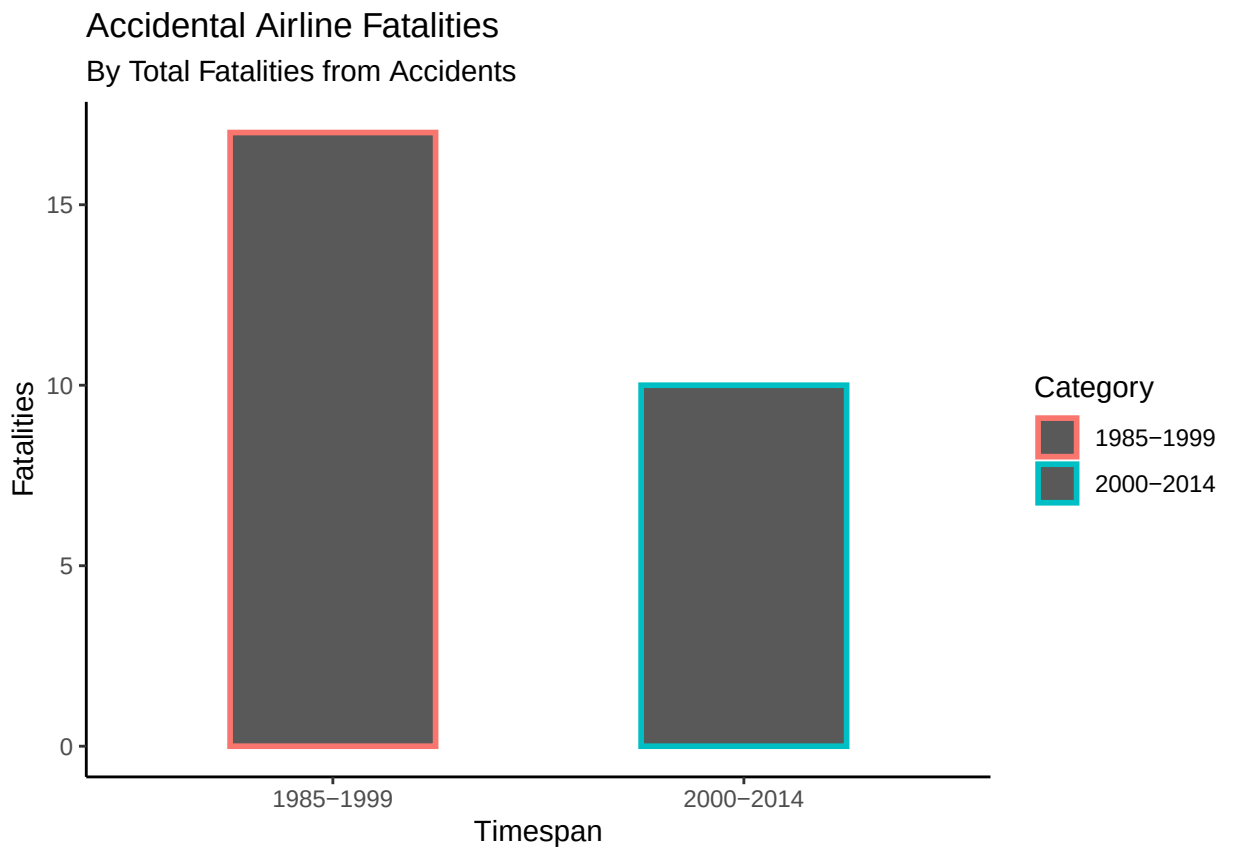
Of the 5 airlines in the first interval from 1985 - 1999 there were a maximum of 5 fatalities from accidents with a minimum of 3 and a median of 3. Given that our total fatalities from accidents had a maximum of 7, this interval makes up the majority of fatalities from accidents.

```
summary(Interval2)
```

```
##    Airline                Time     Fatalities
##  Length:5           1985-1999:0   Min.   :2
##  Class :character   2000-2014:5   1st Qu.:2
##  Mode  :character                 Median :2
##                                   Mean   :2
##                                   3rd Qu.:2
##                                   Max.   :2
```

In this summary we can see that of the 5 airlines in the second interval from 2000-2014, there were a maximum of 2 fatalities from accidents with a minimum of 2 and a median of 2. All airlines in this section had a total of 2 fatalities from accidents. Since this is fewer than the first interval, this second interval is the least dangerous time of this study to fly. We can see this in another bar chart.

```
# Creating a table of stats with totals by interval
airsafety_stats <- aggregate(airsafety$Fatalities, by = list(Category = airsafety$Time), FUN = sum)
# Plotting the difference in fatalities by interval
ggplot(data = airsafety_stats, aes(x= Category, y = x, col = Category)) + geom_bar(stat = "identity", wi
  theme_classic() +
  labs(title = "Accidental Airline Fatalities",
       subtitle = "By Total Fatalities from Accidents",
       x = "Timespan",
       y = "Fatalities")
```



We can also review a table to compare exact numbers over each time interval. To do so, we should rename the new columns to match the rest of air safety statistics. Then we can either pass the data frame back through or, in this case, demonstrate how to make a table from the data frame.

```
# Generate table from airsafety stats
airsafety_stats <- airsafety_stats %>%
  dplyr::rename(Time = Category,
                Fatalities = x)
data.table::as.data.table(airsafety_stats)
```

```
##         Time Fatalities
## 1: 1985-1999         17
## 2: 2000-2014         10
```

Although this table only has two rows, if more data were added it would be better represented using this method. We can see in this table that there is a difference of 7 fatalities from accidents between the two with the timespan of 1985 - 1999 having the most.

Importantly, the comparisons from the intervals 1985 - 1999 and 2000 - 2014 used the same duration between them. Each has 14 years in its interval. This makes 2000 - 2014 the clear safest time to have flown.

Conclusion

As we are aware, safety is something we should all consider. From this brief analysis we can conclude that Ethiopian Airlines is the most dangerous airline in terms of the number of fatalities from accidents. They had 7 total deaths in 28 years from 1985 - 2014. That is the highest among any of the airlines.

We also noted a trend that the number of fatalities from accidents decreased overall from 1985 - 2014. The first 14 year interval had a total of 17 fatalities while second from 2000 - 2014 only had 10.

All airlines in 2000 - 2014 had 2 fatalities from accidents. Over the 28 year duration of the five flights included in this study we saw a maximum of 7 fatalities from accidents and minimum of 2. If you were going to fly during this time, it would have been best to do so within the timespan of 2000 - 2014 and to avoid flying with Ethiopian Airlines.