Visualization Activity

sunil salunke

6/19/2022

Section 1: Description of the data

This is the data set which information about airlines and their safety incidents. The data is classified in safety incidents, fatal accidents, and fatalities. There are 2 duration of 1985 to 1999 and 2000 to 2014 for range of 56 airlines.

Section 2: Reading the data into R

```
#using read.csv to read data from csv file from a URL,
url <-
↪  "https://raw.githubusercontent.com/fivethirtyeight/data/master/airline-safety/airline-safety.csv"
airline_safety_df <- read.csv(url)
```

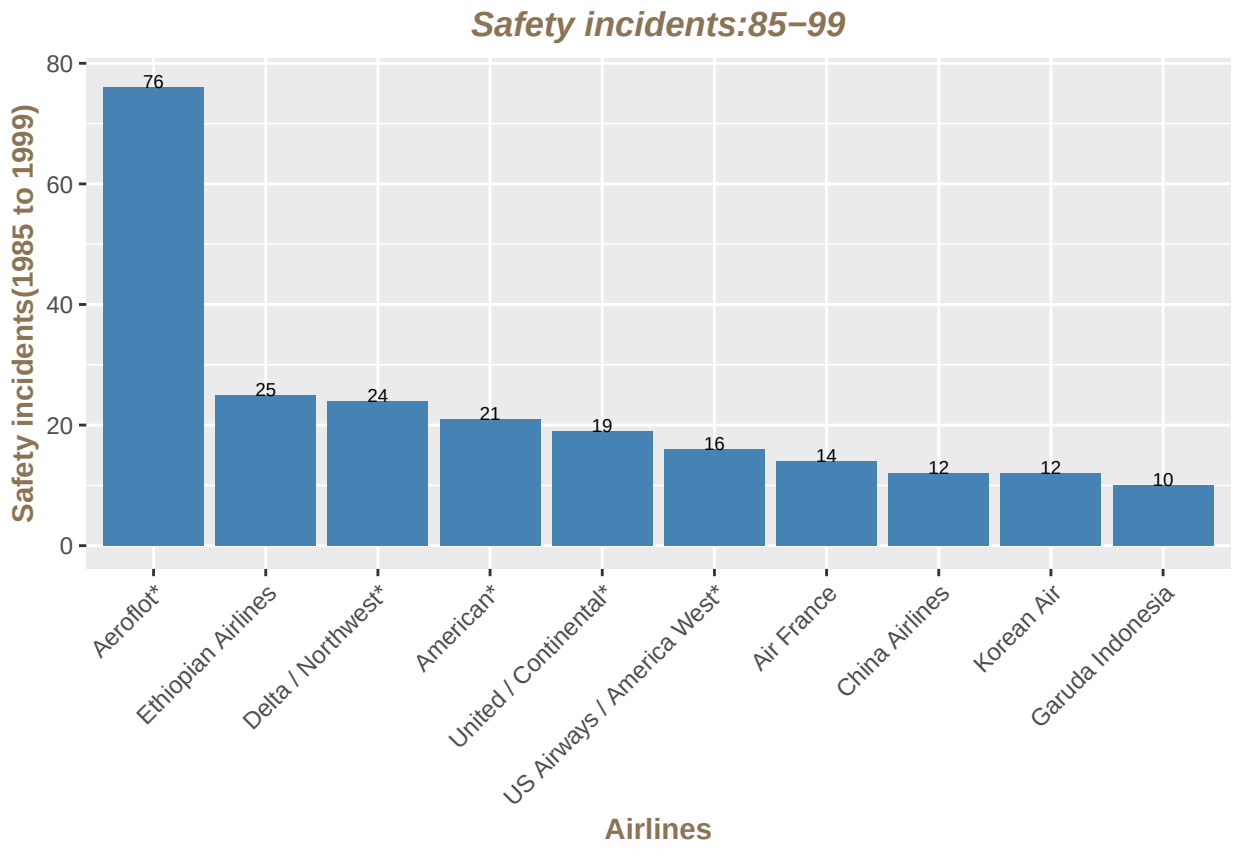
```
#Libraries needed for analysis
library(ggplot2)
library(tidyverse)
library(patchwork)
library(treemapify)
```

Section 3:

Visualization 1: Top 10 airlines by number of safety incidents for 1985-1999

```
pct_format = scales::percent_format(accuracy = .1) #label format

top_safety_incidents_85_99 <- airline_safety_df %>%
  select(airline,incidents_85_99) %>% #selecting necessary columns
  top_n(10, incidents_85_99) %>%  #selecting top 10 records
  arrange(desc(incidents_85_99)) %>% #arranging records in decending order
  ggplot(aes(x=reorder(airline,-incidents_85_99), y = incidents_85_99)) + #x and y axis
  ↪  data
  geom_bar(stat = "identity", fill = "steelblue") + #define chart type
  geom_text(aes(label = sprintf('%d', incidents_85_99)),
            nudge_y =  1, size = 2.5) + #define bar labels
  xlab("Airlines") + #define x axis label
  ylab("Safety incidents(1985 to 1999)") + #define y axis label
```

```
  ggtitle("Safety incidents:85-99") + #define plot title
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title.x = element_text(color = "burlywood4", face = "bold"),
        axis.title.y = element_text(color = "burlywood4", face = "bold"),
        plot.title = element_text(color = "burlywood4", face = "bold.italic", hjust =
        ↪  0.5))

top_safety_incidents_85_99
```



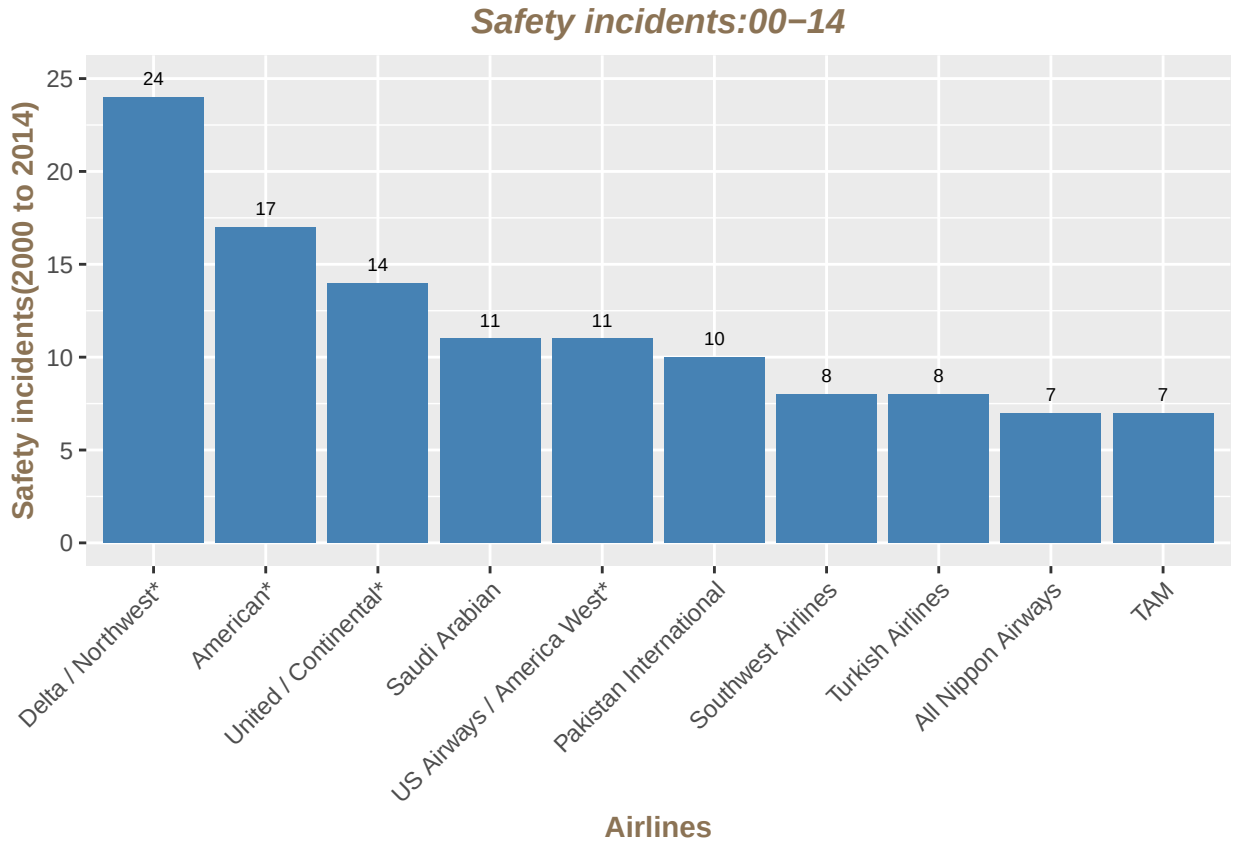
Comment: This visualization gives information about the top 10 Airlines by the safety incidents across the time from 1985 to 1999.

As we can see in this visualization Aeroflot, Ethiopian Airlines, and Delta/ Northwest Airlines are the airlines with the highest number of safety incidents between 1985 and 1999.

Visualization 2: Top 10 airlines by number of safety incidents for 2000-2014

```
pct_format = scales::percent_format(accuracy = .1) #label format

top_safety_incidents_00_14 <- airline_safety_df %>%
  select(airline,incidents_00_14) %>% #selecting necessary columns
  top_n(10, incidents_00_14) %>%  #selecting top 10 records
  arrange(desc(incidents_00_14)) %>% #arranging records in decending order
```

```
  ggplot(aes(x=reorder(airline,-incidents_00_14), y = incidents_00_14)) + #x and y axis
   ↪  data
  geom_bar(stat = "identity", fill = "steelblue") + #define chart type
  geom_text(aes(label = sprintf('%d', incidents_00_14)),
            nudge_y =  1, size = 2.5) + #define bar labels
  xlab("Airlines") + #define x axis label
  ylab("Safety incidents(2000 to 2014)") + #define y axis label
  ggtitle("Safety incidents:00-14") + #define plot title
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title.x = element_text(color = "burlywood4", face = "bold"),
        axis.title.y = element_text(color = "burlywood4", face = "bold"),
        plot.title = element_text(color = "burlywood4", face = "bold.italic", hjust =
         ↪  0.5))

top_safety_incidents_00_14
```

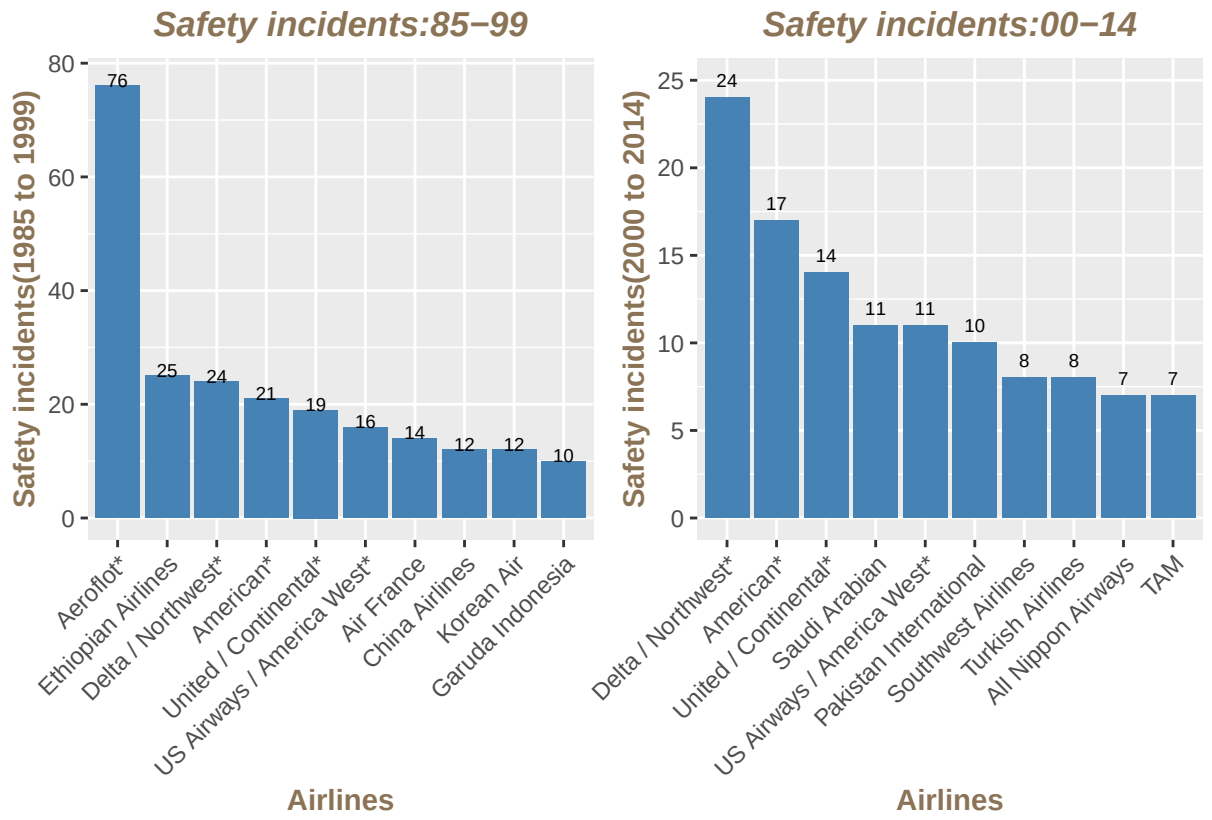


Comment: This visualization gives information about the top 10 Airlines by the safety incidents across the time from 2000 to 2014.

As we can see in this visualization Delta/ Northwest Airlines, American, and United/Continental are the airlines with the highest number of safety incidents between 2000 and 2014.

Visualization 3: Comaprison of top 10 airlines by number of safety incidents across 1985-1999 and 2000-2014

```
#airlines with top safety incidents across time periods
#grid.arrange(top_safety_incidents_85_99, top_safety_incidents_00_14)
#dev.off()
top_safety_incidents_85_99 + top_safety_incidents_00_14
```



Comment: This visualization compares the top 10 Airlines by the highest number of safety incidents between 1985 to 1999 and 2000 to 2014. This visualization helps to understand whether the same Airlines have high safety incidents across both periods.

As we can see Delta/ Northwest, American, and United/Continental airlines appear in the top 10 airlines with safety incidents across both periods.

This observation gives us insight that these three Airlines are less safer to travel at they have a relatively higher number of safety incidents across both time periods between 1985 to 1999 and 2000 to 2014.

Safety Score calculation for 2000 - 2014 period:

Safety Score: Safety score is a measure calculated by considering safety incidents, fatal accidents, and fatalities

Safety score is calculated by cosidering below steps.

1. Step 1: In order to calculate safety score for incidents_00_14, mean of the column is calcualted and each record is substracted from the mean.

2. If the number of incidents are higher than mean then it gives negative score indicating less safe airline. If the number of incidents are lower than mean then it gives positive score indicating safer airline. Higher the positive value, safer the airline. Higher the negative value, less safer the airline.

3. Step 2: Multiply the result obtained from step 1 by the square root of the number of seat kilometers flown. This will be helpful to give higher credit to airlines that have achieved a good safety record over the larger kilometers flown.

4. Step 3: Standardize the score for the category i.e. incidents_00_14 in order to understand how many standard deviation airline is above or below mean.

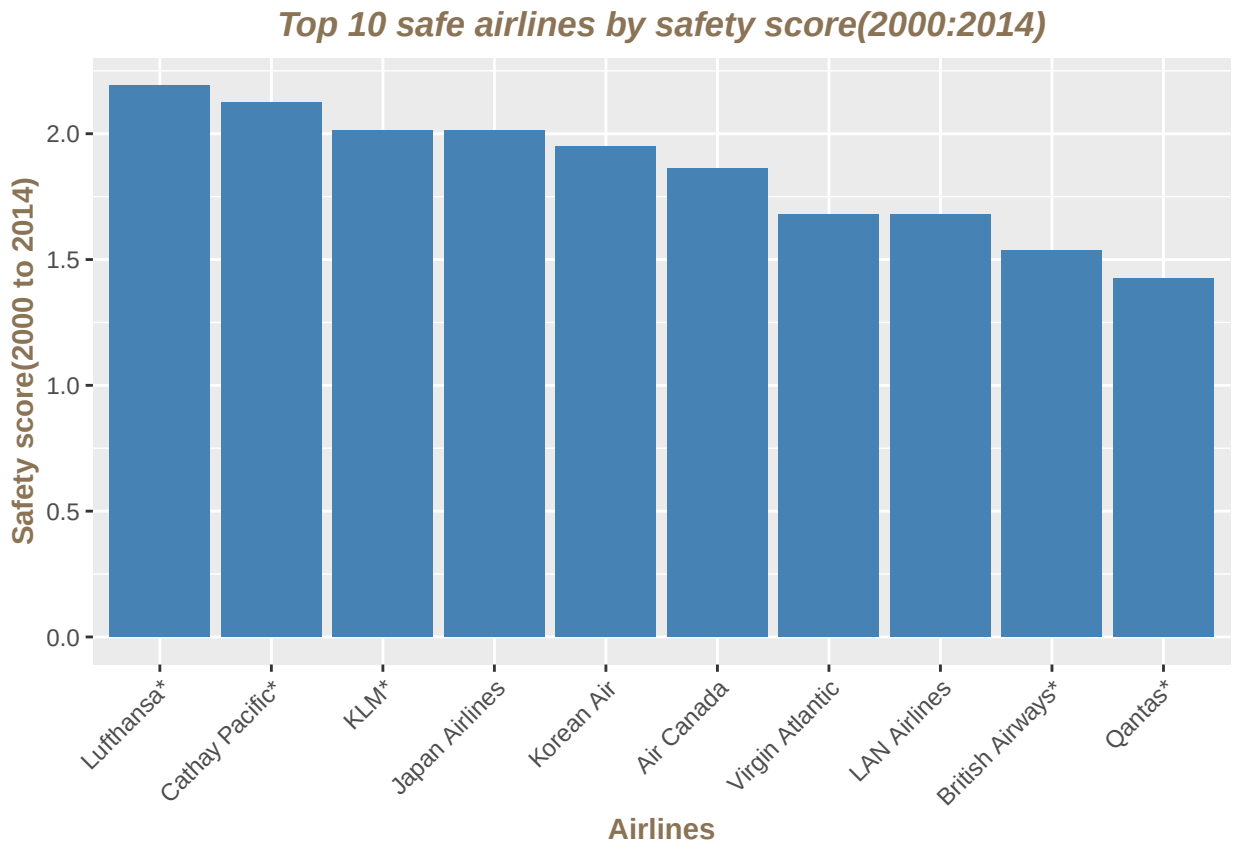
5. Step 4: Sum the score from 3 categories i.e. incidents, fatal_accidents, and fatalities

```
#let's develop a safety score by considering safety incidents, fatal accidents, and
↪  fatalities

#safety score for incidents
incidents_00_14_safety_score <- airline_safety_df %>%
  select(airline,incidents_00_14,avail_seat_km_per_week) %>%
  summarise(incidents_00_14_safety =
              (mean(incidents_00_14) - incidents_00_14 )* sqrt(avail_seat_km_per_week))
↪  %>%
  summarise(incidents_00_14_safety_score = (incidents_00_14_safety -
   ↪  mean(incidents_00_14_safety))/sd(incidents_00_14_safety))

#safety score for fatal accidents
fatal_accidents_00_14_safety_score <- airline_safety_df %>%
  select(airline,fatal_accidents_00_14,avail_seat_km_per_week) %>%
  summarise(fatal_accidents_00_14_safety =
              (mean(fatal_accidents_00_14) - fatal_accidents_00_14 )*
↪  sqrt(avail_seat_km_per_week)) %>%
  summarise(fatal_accidents_00_14_safety_score =
    (fatal_accidents_00_14_safety -
↪  mean(fatal_accidents_00_14_safety))/sd(fatal_accidents_00_14_safety))

#safety score for fatalities
fatalities_00_14_safety_score <- airline_safety_df %>%
  select(airline,fatalities_00_14,avail_seat_km_per_week) %>%
  summarise(fatalities_00_14_safety =
              (mean(fatalities_00_14) - fatalities_00_14 )* sqrt(avail_seat_km_per_week))
↪  %>%
  summarise(fatalities_00_14_safety_score =
    (fatalities_00_14_safety -
↪  mean(fatalities_00_14_safety))/sd(fatalities_00_14_safety))


airline_safety_df =  as.data.frame(cbind(airline_safety_df,
↪  incidents_00_14_safety_score))
airline_safety_df =  as.data.frame(cbind(airline_safety_df,
↪  fatal_accidents_00_14_safety_score))
airline_safety_df =  as.data.frame(cbind(airline_safety_df,
↪  fatalities_00_14_safety_score))
```

```
#overall safety score for airlines for 2000-2014 period
airline_safety_df$safety_score_00_14 = airline_safety_df$incidents_00_14_safety_score +
↪  airline_safety_df$fatal_accidents_00_14_safety_score +
↪  airline_safety_df$fatalities_00_14_safety_score
```

Visualization 4: Top 10 airlines by their safety scores

```
airline_safety_df %>%
  select(airline,safety_score_00_14) %>% #selecting necessary columns
  top_n(10, safety_score_00_14) %>%  #selecting top 10 records
  arrange(desc(safety_score_00_14)) %>% #arranging records in descending order
  ggplot(aes(x=reorder(airline,-safety_score_00_14), y = safety_score_00_14)) + #x and y
  ↪  axis data
  geom_bar(stat = "identity", fill = "steelblue") + #define chart type
  xlab("Airlines") + #define x axis label
  ylab("Safety score(2000 to 2014)") + #define y axis label
  ggtitle("Top 10 safe airlines by safety score(2000:2014)") + #define plot title
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title.x = element_text(color = "burlywood4", face = "bold"),
        axis.title.y = element_text(color = "burlywood4", face = "bold"),
        plot.title = element_text(color = "burlywood4", face = "bold.italic", hjust =
        ↪  0.5))
```



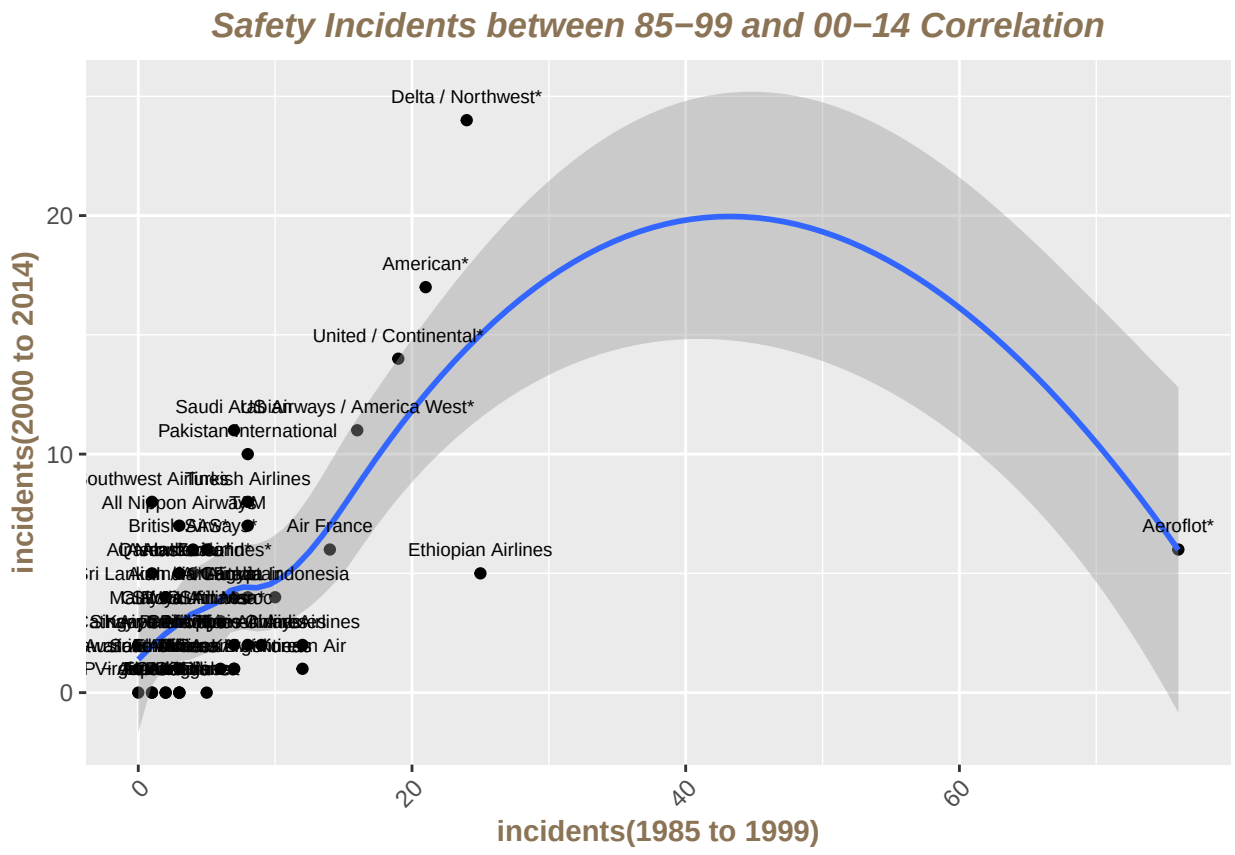
Comment: This visualization gives information about the top 10 Airlines which are safe to travel to by considering their safety scores across the period of 2000 to 2014.

As per the visualization, Lufthansa, Cathay Pacific, KLM, Japan Airlines, Korea Air, and Air Canada are some of the highly safe Airlines to travel via.

Visualization 5: Scatter plot of incidents by airline across 1985-99 and 2000-14

```
airline_safety_df %>%
  select(airline, incidents_85_99, incidents_00_14) %>% #selecting necessary columns
  ggplot(aes(x=incidents_85_99, y=incidents_00_14)) + #x and y axis data
  geom_point(fill = "steelblue") +  #define chart type
  geom_smooth() +
  geom_text(label = airline_safety_df$airline,
            nudge_y =  1, size = 2.5) + #define bar labels
  xlab("incidents(1985 to 1999)") + #define x axis label
  ylab("incidents(2000 to 2014)") + #define y axis label
  ggtitle("Safety Incidents between 85-99 and 00-14 Correlation") + #define plot title
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title.x = element_text(color = "burlywood4", face = "bold"),
        axis.title.y = element_text(color = "burlywood4", face = "bold"),
        plot.title = element_text(color = "burlywood4", face = "bold.italic", hjust =
        ↪  0.5))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Comment: This visualization gives information about whether there is a correlation between safety incidents from 1985 to 1999 time period versus 2000 to 2014 time period across different airlines.

This correlation visualization looks into all available data points to get a better picture of whether the incidents by airlines are predictable across both periods.

Based on the visualization, it can be concluded that incidents by airlines are somewhat predictable. The line shows the relationship between safety incidents from 1985 to 1999 time period and 2000 to 2014 time period across different airlines.