BIOS512_FinalProject_Williams (1) (4) (1) (1) (2)

November 29, 2025

BIOS512 FINAL PROJECT BELLA WILLIAMS

[1]:
```
library(tidyverse)
```

```
  Attaching core tidyverse packages                tidyverse
2.0.0
  dplyr     1.1.2       readr     2.1.4
  forcats   1.0.0       stringr   1.5.0
  ggplot2   3.4.2       tibble    3.2.1
  lubridate 1.9.2       tidyr     1.3.0
  purrr     1.0.1
  Conflicts
tidyverse_conflicts()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()
  Use the conflicted package
(<http://conflicted.r-lib.org/>) to force all conflicts to
become errors
```

[2]:
```
install.packages('fivethirtyeight')
install.packages('fivethirtyeightdata', repos = 'https://fivethirtyeightdata.
 ↪github.io/drat/', type = 'source')
```

```
Installing package into '/srv/rlibs'
(as 'lib' is unspecified)

Installing package into '/srv/rlibs'
(as 'lib' is unspecified)
```

[3]:
```
library(fivethirtyeight)
data("airline_safety")
```

Part I: Preparing the Data

[4]:
```
head(airline_safety)
```

A tibble: 6 × 9

airline <chr>	incl_reg_subsidiaries <lgl>	avail_seat_km_per_week <dbl>	incidents_85_99 <int>
Aer Lingus	FALSE	320906734	2
Aeroflot	TRUE	1197672318	76
Aerolineas Argentinas	FALSE	385803648	6
Aeromexico	TRUE	596871813	3
Air Canada	FALSE	1865253802	2
Air France	FALSE	3004002661	14

[5]:
```
str(airline_safety)

summary(airline_safety)

mean(airline_safety$fatalities_85_99)
mean(airline_safety$fatalities_00_14)
```

```
tibble [56 × 9] (S3: tbl_df/tbl/data.frame)
 $ airline               : chr [1:56] "Aer Lingus" "Aeroflot" "Aerolineas
Argentinas" "Aeromexico" …
 $ incl_reg_subsidiaries : logi [1:56] FALSE TRUE FALSE TRUE FALSE FALSE …
 $ avail_seat_km_per_week: num [1:56] 3.21e+08 1.20e+09 3.86e+08 5.97e+08
1.87e+09 …
 $ incidents_85_99       : int [1:56] 2 76 6 3 2 14 2 3 5 7 …
 $ fatal_accidents_85_99 : int [1:56] 0 14 0 1 0 4 1 0 0 2 …
 $ fatalities_85_99      : int [1:56] 0 128 0 64 0 79 329 0 0 50 …
 $ incidents_00_14       : int [1:56] 0 6 1 5 2 6 4 5 5 4 …
 $ fatal_accidents_00_14 : int [1:56] 0 1 0 0 0 2 1 1 1 0 …
 $ fatalities_00_14      : int [1:56] 0 88 0 0 0 337 158 7 88 0 …
```

```
  airline          incl_reg_subsidiaries avail_seat_km_per_week
 Length:56          Mode :logical         Min.   :2.594e+08
 Class :character   FALSE:40              1st Qu.:4.740e+08
 Mode  :character   TRUE :16              Median :8.029e+08
                                          Mean   :1.385e+09
                                          3rd Qu.:1.847e+09
                                          Max.   :7.139e+09
 incidents_85_99  fatal_accidents_85_99 fatalities_85_99 incidents_00_14
 Min.   : 0.000   Min.   : 0.000        Min.   :  0.0    Min.   : 0.000
 1st Qu.: 2.000   1st Qu.: 0.000        1st Qu.:  0.0    1st Qu.: 1.000
 Median : 4.000   Median : 1.000        Median : 48.5    Median : 3.000
 Mean   : 7.179   Mean   : 2.179        Mean   :112.4    Mean   : 4.125
 3rd Qu.: 8.000   3rd Qu.: 3.000        3rd Qu.:184.2    3rd Qu.: 5.250
 Max.   :76.000   Max.   :14.000        Max.   :535.0    Max.   :24.000
 fatal_accidents_00_14 fatalities_00_14
 Min.   :0.0000        Min.   :  0.00
 1st Qu.:0.0000        1st Qu.:  0.00
 Median :0.0000        Median :  0.00
 Mean   :0.6607        Mean   : 55.52
```

```
 3rd Qu.:1.0000         3rd Qu.: 83.25
 Max.   :3.0000         Max.   :537.00
```

112.410714285714

55.5178571428571

- Checking Missingness

[6]:
```
sum(is.na(airline_safety))
colSums(is.na(airline_safety))
```

0

airline 0 **incl_reg_subsidiaries** 0 **avail_seat_km_per_week** 0
incidents_85_99 0 **fatal_accidents_85_99** 0 **fatalities_85_99** 0
incidents_00_14 0 **fatal_accidents_00_14** 0 **fatalities_00_14** 0

- Checking Duplicates

[7]:
```
nrow(airline_safety)
nrow(distinct(airline_safety))

airline_safety %>%
  group_by(airline) %>%
  tally() %>%
  filter(n > 1)
```

56

56

A tibble: 0 × 2

airline	n
<chr>	<int>

[8]:
```
airline_safety %>%
  group_by(incl_reg_subsidiaries) %>%
  tally() %>%
  arrange(desc(n))
```

A tibble: 2 × 2

incl_reg_subsidiaries	n
<lgl>	<int>
FALSE	40
TRUE	16

Part II: Data Description (Codebook)

The airline_safety dataset contains safety records for 56 airlines from 1985–2014. The data were originally compiled by FiveThirtyEight and include operational exposure (airline size) and incident/fatality counts for two time periods. No missing values are present in this dataset.

Variable	Type	Description
airline	character	Airline name
incl_reg_subsidiaries	logical	TRUE if safety numbers include regional subsidiaries, otherwise FALSE
avail_seat_km_per_week	numeric	Airline size, measured as available seat–kilometers flown per week
incidents_85_99	integer	Total incidents from 1985–1999
fatal_accidents_85_99	integer	Fatal accidents from 1985–1999
fatalities_85_99	integer	Total fatalities from 1985–1999
incidents_00_14	integer	Total incidents from 2000–2014
fatal_accidents_00_14	integer	Fatal accidents from 2000–2014
fatalities_00_14	integer	Total fatalities from 2000–2014

Part III: Data Analysis/Visualization

Question 1. Which airlines have the highest total number of incidents between 1985 and 2014, and what is the range of total incidents across all airlines?

[11]:
```
airline_safety$total_incidents <- airline_safety$incidents_85_99 +
airline_safety$incidents_00_14

p3 <- ggplot(
airline_safety,
aes(x = reorder(airline, total_incidents), y = total_incidents)

) +
geom_col() +
coord_flip() +
labs(
title = "Total Incidents per Airline (1985-2014)",
x = "Airline",
y = "Total Incidents"
) +
theme_minimal()

ggsave("plot3_total_incidents_bar.png", p3, width = 6, height = 6)
```

```
print(p3)
```

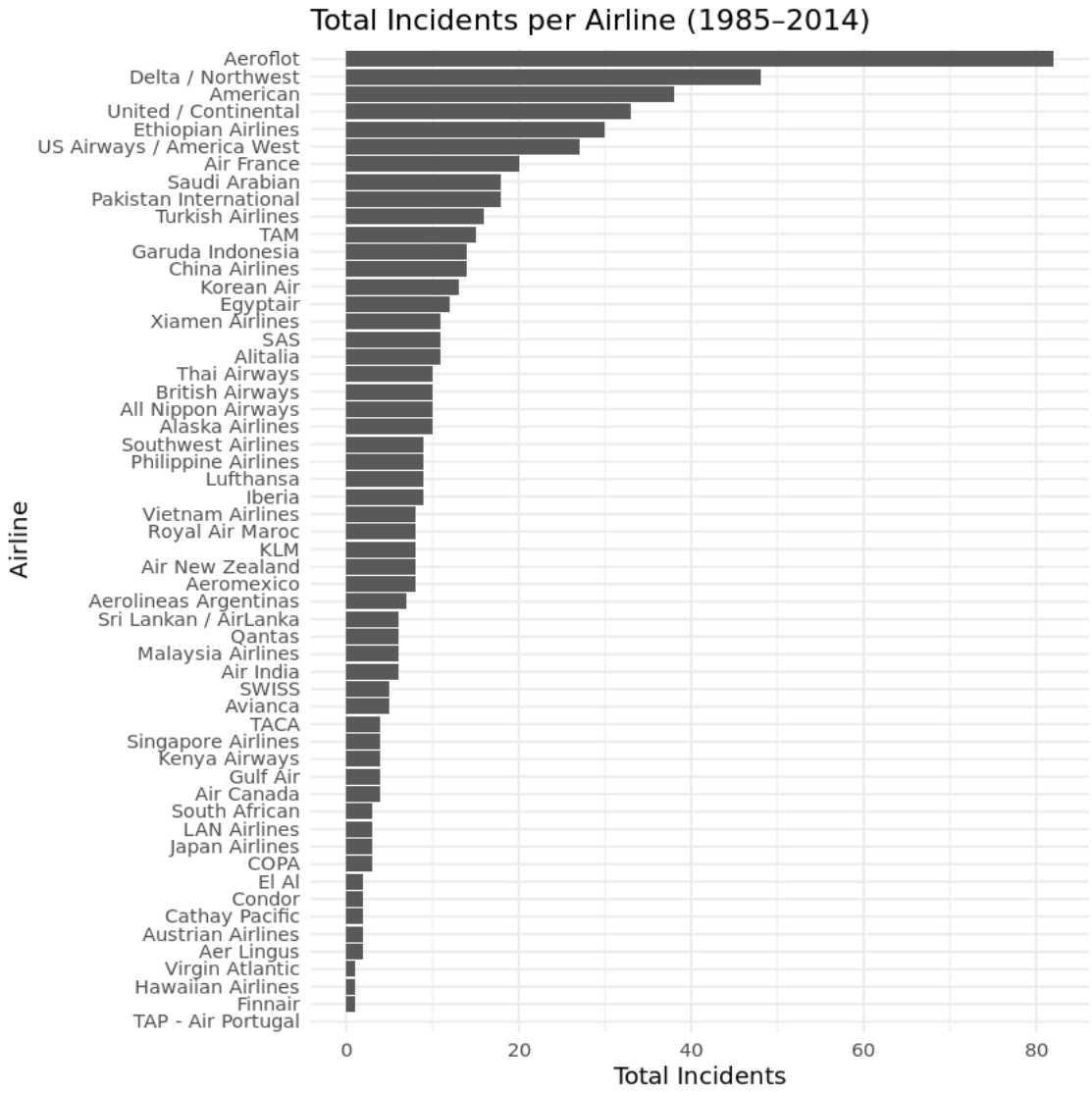


Figure 1. Horizontal bar chart displaying the Total Incidents recorded for each of the 56 airlines over the entire 30-year study period (1985–2014), sorted in descending order of incident count.

Conclusion: The bar chart of "Total Incidents per Airline (1985-2014)" shows that Aeroflot had the highest total number of incidents in the 30-year period. The next highest incidents belong to Delta/Northwest, American, and United/Continental. The total number of incidents across all 56 airlines ranges significantly, from a minimum of 0 for TAP Air Portugal to the maximum value of 76 for Aeroflot.

[12]:
```
library(dplyr)

airline_safety <- airline_safety %>%
```

```
    mutate(
        incidents_85_99_rate = incidents_85_99 / (avail_seat_km_per_week / 1e9),
        incidents_00_14_rate = incidents_00_14 / (avail_seat_km_per_week / 1e9)
    )
```

Question 2. Which airlines have the highest incident rates?

[13]:
```
install.packages('ggrepel')
```

```
Installing package into ‘/srv/rlibs’
(as ‘lib’ is unspecified)
```

[14]:
```
library(ggplot2)
library(ggrepel)

p1 <- ggplot(
    airline_safety,
    aes(
        x = avail_seat_km_per_week,
        y = incidents_85_99_rate,
        label = airline
    )
) +
    geom_point() +
    geom_text(aes(label = airline),
            vjust = -0.5,
            size = 3
    ) +
    labs(
        title = "Incident Rate (1985-1999) vs Operational Volume",
        x = "Available Seat KM per Week",
        y = "Incidents per Billion Seat-KM (1985-1999)"
    ) +
    theme_minimal()

print(p1)
```

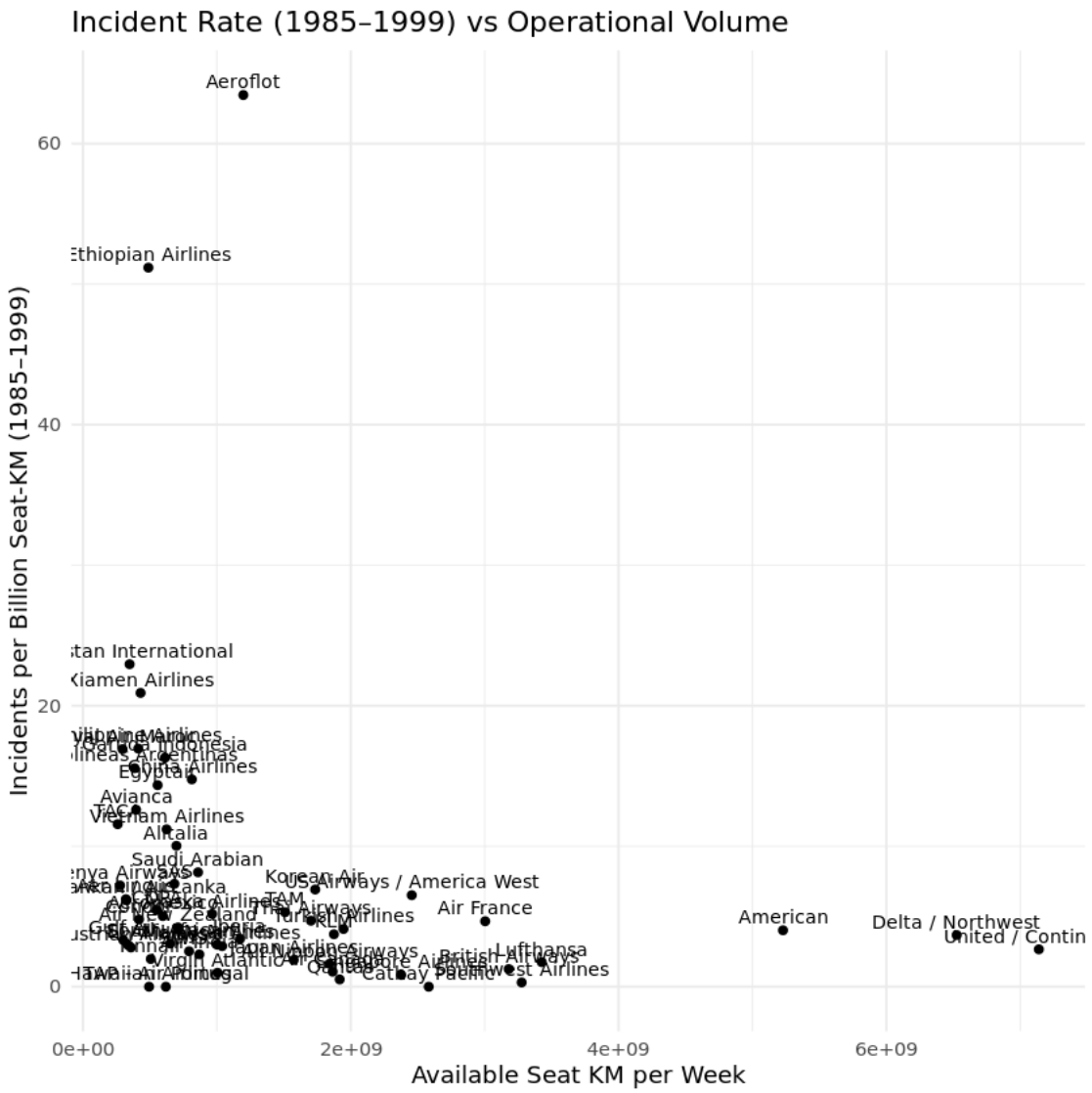


Figure 2. This scatterplot displays the relationship between each airline's operational volume, measured as available seat-kilometers per week, and its incident rate during 1985–1999, expressed per billion seat-kilometers.

[15]:
```
p2 <- ggplot(
    airline_safety,
    aes(
        x = avail_seat_km_per_week,
        y = incidents_00_14_rate,
        label = airline
    )
) +
```

```
    geom_point() +
    geom_text(aes(label = airline),
              vjust = -0.5,
              size = 3)  +
    labs(
        title = "Incident Rate (2000-2014) vs Operational Volume",
        x = "Available Seat KM per Week",
        y = "Incidents per Billion Seat-KM (2000-2014)"
    ) +
    theme_minimal()

print(p2)
```

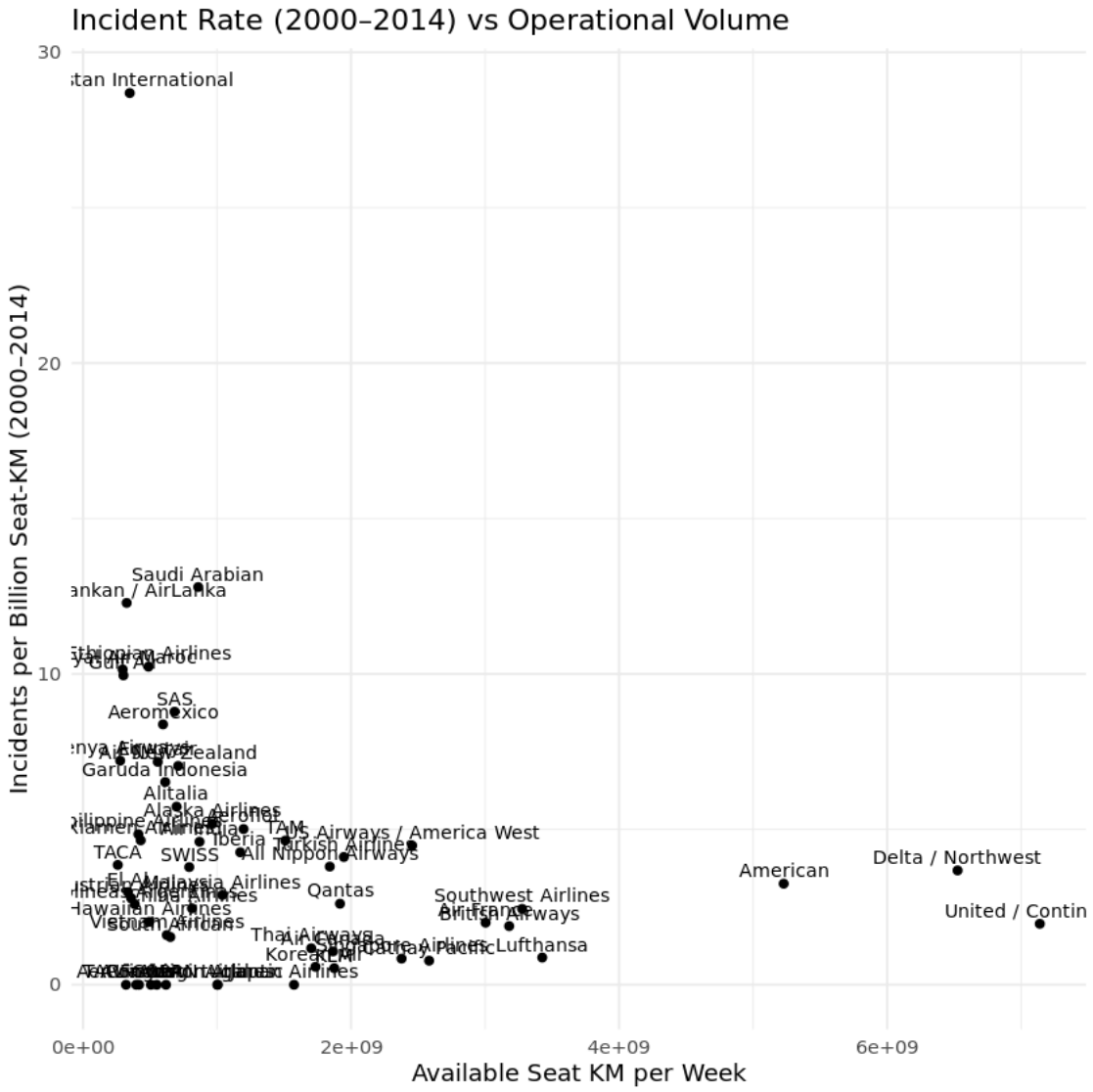


Figure 3. This scatterplot displays the relationship between each airline's operational volume, measured as available seat-kilometers per week, and its incident rate during 2000-2014, expressed per billion seat-kilometers.

Conclusion: The incident rate scatterplots reveal that the highest incident rates are generally associated with lower operational volume (Available Seat KM per Week). For the period 1985–1999, Aeroflot had the highest incident rate, followed by Ethiopian Airlines and Pakistan International. For the period 2000–2014, Pakistan International had the highest rate, followed by Saudi Arabian and Sri Lankan/AirLanka. Notably, Aeroflot's incident rate appears to have dropped significantly in the second period, moving from the highest rate in 1985–1999 to a much lower, clustered position in 2000–2014. Conversely, American, Delta/Northwest, and United/Continental maintain low incident rates relative to their high operational volume in both periods.

Question 3. How consistently did safety improvements occur across airlines from 1985–1999 to 2000–2014 after adjusting for operational volume?

[18]:
```
library(dplyr)
library(tibble)

airline_safety <- airline_safety %>%
  mutate(
    fatal_accidents_85_99_rate = fatal_accidents_85_99 /␣
 ↪(avail_seat_km_per_week / 1e9),
    fatalities_85_99_rate      = fatalities_85_99      /␣
 ↪(avail_seat_km_per_week / 1e9),

    fatal_accidents_00_14_rate = fatal_accidents_00_14 /␣
 ↪(avail_seat_km_per_week / 1e9),
    fatalities_00_14_rate      = fatalities_00_14      /␣
 ↪(avail_seat_km_per_week / 1e9)
  )
```

[19]:
```
p2 <- ggplot(
    airline_safety,
    aes(x = fatalities_85_99_rate,
        y = fatalities_00_14_rate,
        label = airline)
) +
    geom_point() +
    geom_text(
        vjust = -0.5,
        size = 3
    ) +
    labs(
        title = "Fatality Rates: 1985-1999 vs 2000-2014 (Standardized by SKM)",
        x = "Fatality Rate 1985-1999",
        y = "Fatality Rate 2000-2014"
    ) +
```

```
    theme_minimal()

ggsave("plot2_fatality_rates_comparison.png", p2, width = 6, height = 4)

print(p2)
```

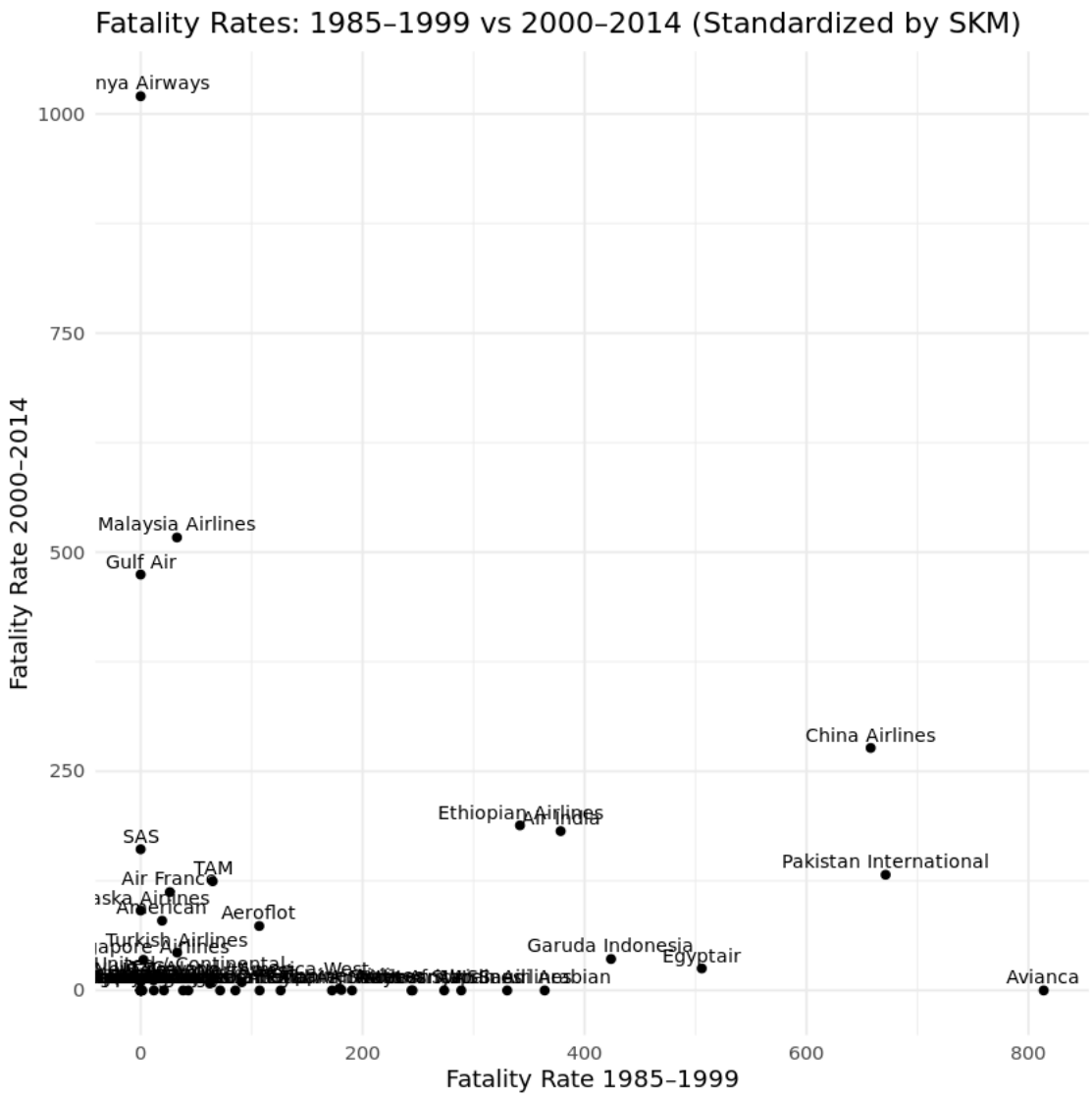


Figure 4. Scatterplot comparing the raw number of fatalities between the period 1985–1999 (x-axis) and 2000–2014 (y-axis). This visualization helps identify airlines that experienced a change in the total number of fatalities between the two time periods.

Analysis Method 1: Dimensionality Reduction

[20]:
```
airline_safety$class <- cut(
  airline_safety$fatalities_00_14_rate,
  breaks = c(-Inf, 0, 2, Inf),
  labels = c("none", "low", "high")
)

airline_safety$class <- as.factor(airline_safety$class)

sapply(airline_safety, class)
```

airline 'character' **incl_reg_subsidiaries** 'logical' **avail_seat_km_per_week** 'numeric' **incidents_85_99** 'integer' **fatal_accidents_85_99** 'integer' **fatalities_85_99** 'integer' **incidents_00_14** 'integer' **fatal_accidents_00_14** 'integer' **fatalities_00_14** 'integer' **total_incidents** 'integer' **incidents_85_99_rate** 'numeric' **incidents_00_14_rate** 'numeric' **fatal_accidents_85_99_rate** 'numeric' **fatalities_85_99_rate** 'numeric' **fatal_accidents_00_14_rate** 'numeric' **fatalities_00_14_rate** 'numeric' **class** 'factor'

[21]:
```
sapply(
  airline_safety %>% select(where(is.numeric)),
  function(x) {
    c(
      Mean = mean(x),
      SD   = sd(x),
      Var  = var(x),
      Min  = min(x),
      Max  = max(x)
    )
  }
)
```

A matrix: 5 × 14 of type dbl

	avail_seat_km_per_week	incidents_85_99	fatal_accidents_85_99	fa
Mean	1.384621e+09	7.178571	2.178571	1
SD	1.465317e+09	11.035656	2.861069	1
Var	2.147154e+18	121.785714	8.185714	2
Min	2.593733e+08	0.000000	0.000000	0
Max	7.139291e+09	76.000000	14.000000	5

[22]:
```
airline_scaled <- airline_safety %>%
  mutate(across(
    .cols = where(is.numeric),
    .fns  = ~ as.numeric(scale(.))
  ))
```

[23]:
```
airline_numeric <- airline_scaled %>%
  select(
    incidents_85_99_rate,
```

```
    fatal_accidents_85_99_rate,
    fatalities_85_99_rate,
    incidents_00_14_rate,
    fatal_accidents_00_14_rate,
    fatalities_00_14_rate
  )
```

[24]:
```
airline_pca <- prcomp(
  airline_numeric,
  center = FALSE,
  scale. = FALSE
)

summary(airline_pca)
airline_pca$rotation
```

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.6948 1.2464 0.8236 0.76597 0.41874 0.3658
Proportion of Variance 0.4787 0.2589 0.1130 0.09779 0.02922 0.0223
Cumulative Proportion  0.4787 0.7377 0.8507 0.94848 0.97770 1.0000
```

A matrix: 6 × 6 of type dbl

	PC1	PC2	PC3	PC4
incidents_85_99_rate	0.4534195	0.2620169	0.60339132	0.10198439
fatal_accidents_85_99_rate	0.4789720	0.3838092	0.21237815	-0.08083012
fatalities_85_99_rate	0.3779975	0.3060301	-0.60239564	-0.56383010
incidents_00_14_rate	0.4268943	-0.1328973	-0.44871505	0.71068383
fatal_accidents_00_14_rate	0.4322747	-0.4836178	0.01518499	-0.01321358
fatalities_00_14_rate	0.2302484	-0.6624465	0.16235194	-0.39989397

[25]:
```
pc_scores <- as_tibble(airline_pca$x) %>%
  mutate(class = airline_safety$class)
```

[26]:
```
ggplot(pc_scores, aes(x = PC1, y = PC2, color = class, label =␣
↪airline_safety$airline)) +
    geom_point() +
  geom_text(aes(label = airline_safety$airline),
            vjust = -0.5,
            size = 3
  ) +
  labs(
    title = "PCA of Airline Safety Data",
    x = "Principal Component 1",
    y = "Principal Component 2",
    color = "Risk Class"
  ) +
  theme_minimal() +
```

```
coord_fixed()
```

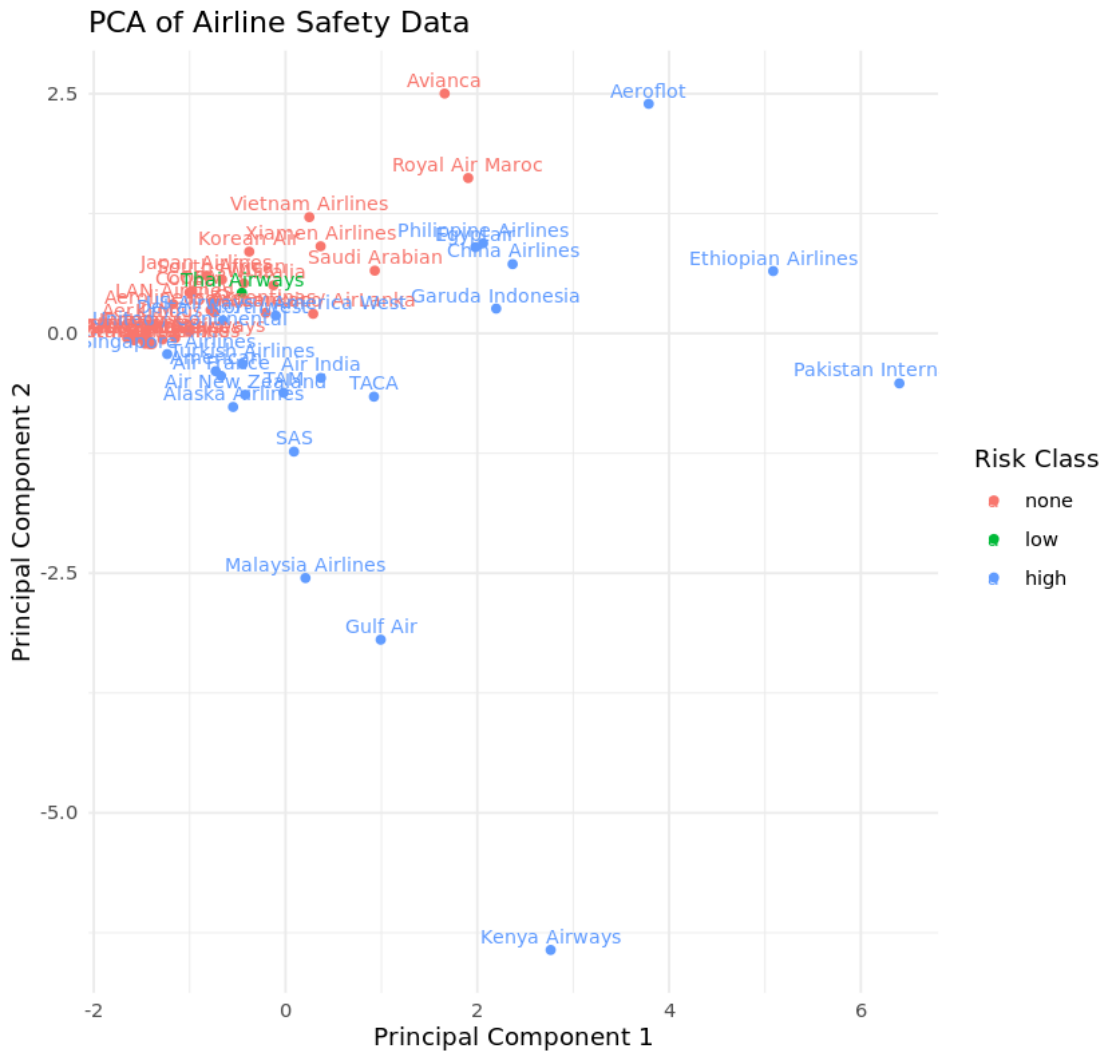


Figure 5. Scatterplot of the airline safety data reduced to the first two principal components (PC1 and PC2). The points are colored based on the assigned Risk Class (none, low, or high), which was derived by binning the fatality rate for the 2000–2014 period.

Conclusion: The principal component analysis (PCA) successfully captured a large portion of the variance in the first two components, with 73.77% of the total variance explained by PC1 and PC2 combined. The "high" risk class (airlines with high recent fatality rates) is visually separated in the lower-right quadrant of the scatterplot. This indicates that high overall historical risk (high PC1) combined with high recent fatality risk (low PC2) effectively distinguishes the most unsafe airlines.

Analysis Method 2: Clustering

[27]:
```r
label_randomly <- function(n_points, n_clusters){
  sample(1:n_clusters, size = n_points, replace = TRUE)
}

get_cluster_means <- function(data, labels){
  data %>%
    mutate(label__ = labels) %>%
    group_by(label__) %>%
    summarize(across(where(is.numeric), mean), .groups = "drop") %>%
    rename(label = label__) %>%
    arrange(label)
}

assign_cluster <- function(data, means){
  X <- as.matrix(data)
  C <- as.matrix(means %>% select(-label))
  X_sq <- rowSums(X*X)
  C_sq <- rowSums(C*C)
  dist_sq_matrix <- outer(X_sq, C_sq, "+") - 2*(X %*% t(C))
  max.col(-dist_sq_matrix)
}

kmeans_done <- function(old_means, new_means, eps = 1e-6){
  om <- as.matrix(old_means %>% select(-label))
  nm <- as.matrix(new_means %>% select(-label))
  mean(sqrt(rowSums((om - nm)^2))) < eps
}

mykmeans <- function(data, n_clusters, eps = 1e-6, max_iter = 100){
  labels <- label_randomly(nrow(data), n_clusters)
  old_means <- get_cluster_means(data, labels)
  iter <- 0
  repeat {
    labels <- assign_cluster(data, old_means)
    new_means <- get_cluster_means(data, labels)
    if (kmeans_done(old_means, new_means, eps) || iter >= max_iter) break
    old_means <- new_means
    iter <- iter + 1
  }
  if (iter >= max_iter) warning("K-means did not converge in max_iter.")
  cat("K-means converged after", iter, "iterations.\n")
  list(labels = labels, means = new_means)
}
```

[28]:
```r
n_clusters <- 3
my_results <- mykmeans(airline_numeric, n_clusters)
```

```
print("--- Custom mykmeans Results (Cluster Labels) ---")
print(my_results$labels)

print("--- Custom mykmeans Results (Cluster Means) ---")
print(my_results$means)
```

```
K-means converged after 6 iterations.
[1] "--- Custom mykmeans Results (Cluster Labels) ---"
 [1] 2 3 2 2 2 2 2 2 2 2 2 2 2 3 2 2 3 2 2 2 3 2 3 2 3 1 2 2 2 1 2 2 2 2 1 3 3 2
[39] 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[1] "--- Custom mykmeans Results (Cluster Means) ---"
# A tibble: 3 × 7
  label incidents_85_99_rate fatal_accidents_85_99_rate fatalities_85_99_rate
  <int>                <dbl>
<dbl>                <dbl>
1     1               -0.309
-0.667
-0.619
2     2               -0.302
-0.363
-0.278
3     3                1.58                        2.00
1.57
#   3 more variables: incidents_00_14_rate <dbl>,
#   fatal_accidents_00_14_rate <dbl>, fatalities_00_14_rate <dbl>
```

[29]:
```
airline_matrix <- as.matrix(airline_numeric)

r_results <- kmeans(airline_matrix, centers = n_clusters, nstart = 25)

print("--- R's K-means Cluster Labels ---")
print(r_results$cluster)

print("--- R's K-means Means ---")
print(r_results$centers)
```

```
[1] "--- R's K-means Cluster Labels ---"
 [1] 2 1 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 1 2 1 2 1 3 2 2 2 3 2 2 2 2 3 1 1 2
[39] 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[1] "--- R's K-means Means ---"
  incidents_85_99_rate fatal_accidents_85_99_rate fatalities_85_99_rate
1            1.5785394                  1.9983701             1.5661191
2           -0.3018007                 -0.3632788            -0.2781350
3           -0.3092078                 -0.6670213            -0.6190439
  incidents_00_14_rate fatal_accidents_00_14_rate fatalities_00_14_rate
1            0.9006838                  0.8719603             0.1000995
2           -0.2215244                 -0.3279659            -0.2672675
3            0.5469733                  2.1942863             3.6196243
```

[30]:
```
library(ggplot2)

pc_cluster_plot <- pc_scores %>%
  mutate(cluster = as.factor(my_results$labels))

ggplot(pc_cluster_plot, aes(PC1, PC2, color = cluster, label =␣
 ↪airline_safety$airline)) +
      geom_point() +
      geom_text(aes(label = airline_safety$airline),
            vjust = -0.5,
            size = 3
    ) +
  labs(
    title = "PCA of Airline Safety Data Colored by K-means Cluster",
    color = "Cluster"
  ) +
  theme_minimal()
```

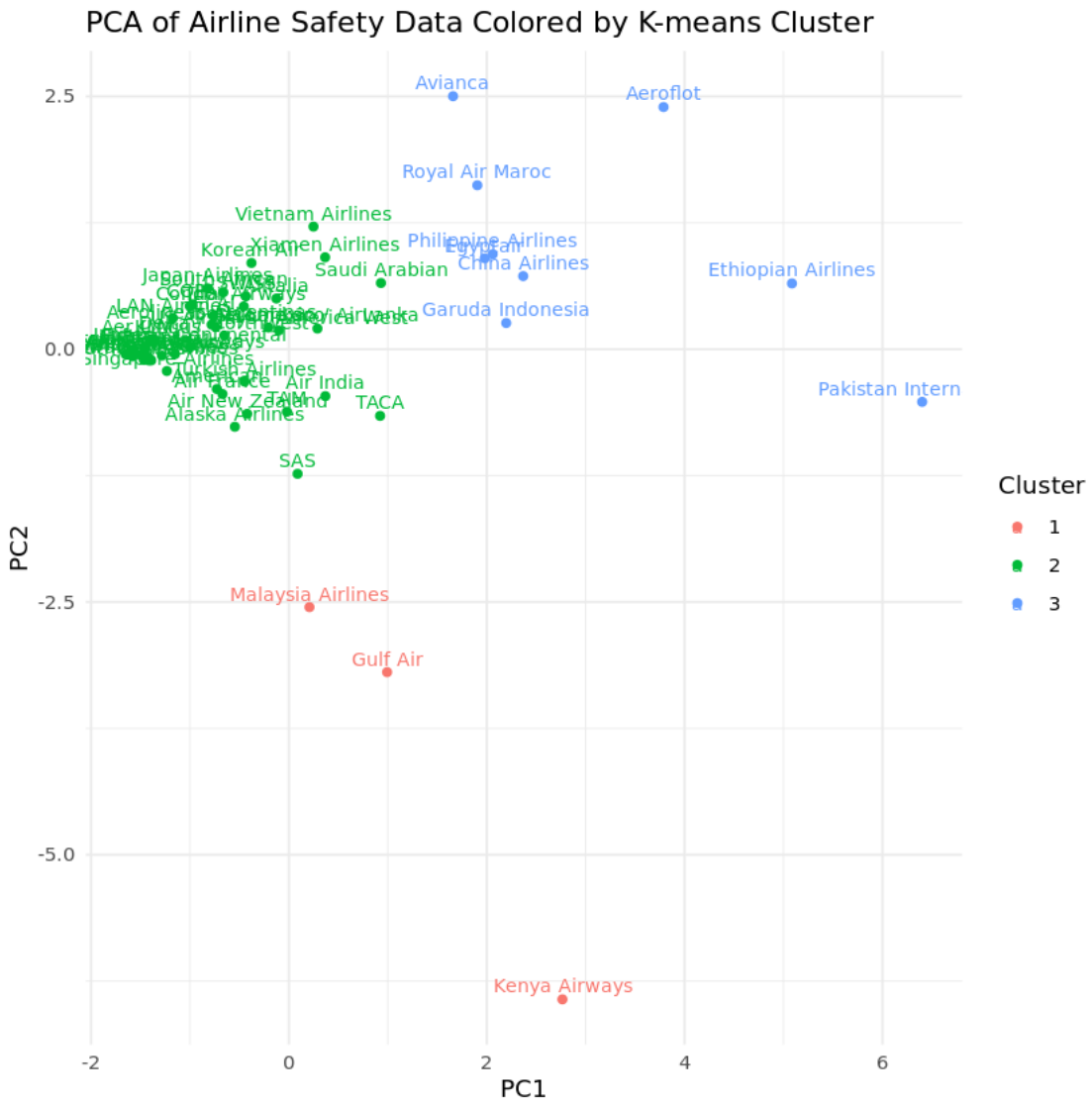


Figure 6. Scatterplot of the PCA results (PC1 vs. PC2) with the airlines colored according to the three clusters generated by the K-means algorithm.

Conclusion: K-means clustering (with k=3) separated the airlines into three meaningful safety groups based on their scaled incident and fatality rates. The groups are defined by the timing and severity of their safety issues. Cluster 2 group is characterized by high historical risk during the 1985–1999 period. Cluster 1 group is defined by relatively low historical risk scores, but a higher recent fatality risk during the 2000-2014 period. Cluster 3, the largest cluster, represents the low risk airlines that remained consistently low risk across all monitoring periods.

Analysis Method 1: Regression

[31]:
```
d <- airline_numeric
d$airline <- airline_safety$airline
```

```r
glimpse(d)
summary(d)

set.seed(123)
train <- runif(nrow(d)) < 0.75
test  <- !train

f_model <- fatalities_00_14_rate ~
    incidents_85_99_rate +
    fatal_accidents_85_99_rate +
    fatalities_85_99_rate +
    incidents_00_14_rate +
    fatal_accidents_00_14_rate

m_improve <- lm(f_model, data = d %>% filter(train))
summary(m_improve)

dx <- d %>% filter(test)
dx <- dx %>% mutate(pred_fatal_rate = predict(m_improve, dx))

ggplot(dx, aes(x = fatalities_00_14_rate,
               y = pred_fatal_rate,
               label = airline)) +

  geom_point() +
  geom_text(vjust = -0.5, size = 3) +

  labs(
    x = "Actual Fatality Rate (00-14)",
    y = "Predicted Fatality Rate (00-14)",
    title = "Predicted vs Actual Fatality Rates (00-14)"
  ) +
  theme_minimal()

ggplot(dx, aes(x = fatalities_00_14_rate - pred_fatal_rate)) +
  geom_density(fill = "coral", alpha = 0.5) +
  labs(
    x = "Residuals (Actual - Predicted Fatality Rate)",
    y = "Density",
    title = "Residual Density Plot (Fatality Rate Model)"
  ) +
  theme_minimal()
```

```
Rows: 56
Columns: 7
$ incidents_85_99_rate       <dbl> -0.1510810, 4.9876633,
```

```
0.6858237, -0.259392…
$ fatal_accidents_85_99_rate <dbl> -0.77289269, 3.08524002,
-0.77289269, -0.21…
$ fatalities_85_99_rate      <dbl> -0.6758657, -0.1190439,
-0.6758657, -0.1172…
$ incidents_00_14_rate       <dbl> -0.87882643, 0.18990017,
-0.32587492, 0.908…
$ fatal_accidents_00_14_rate <dbl> -0.5522797003,
0.0002180767, -0.5522797003,…
$ fatalities_00_14_rate      <dbl> -0.38717841, 0.05192897,
-0.38717841, -0.38…
$ airline                    <chr> "Aer Lingus", "Aeroflot",
"Aerolineas Argen…
```

```
 incidents_85_99_rate fatal_accidents_85_99_rate fatalities_85_99_rate
 Min.   :-0.7107      Min.   :-0.7729            Min.   :-0.6759
 1st Qu.:-0.4891      1st Qu.:-0.7729            1st Qu.:-0.6759
 Median :-0.3118      Median :-0.3449            Median :-0.4905
 Mean   : 0.0000      Mean   : 0.0000            Mean   : 0.0000
 3rd Qu.: 0.0628      3rd Qu.: 0.1742            3rd Qu.: 0.2790
 Max.   : 4.9877      Max.   : 3.0852            Max.   : 3.5639
 incidents_00_14_rate fatal_accidents_00_14_rate fatalities_00_14_rate
 Min.   :-0.8788      Min.   :-0.5523            Min.   :-0.3872
 1st Qu.:-0.6606      1st Qu.:-0.5523            1st Qu.:-0.3872
 Median :-0.2529      Median :-0.5523            Median :-0.3872
 Mean   : 0.0000      Mean   : 0.0000            Mean   : 0.0000
 3rd Qu.: 0.1990      3rd Qu.: 0.1521            3rd Qu.:-0.1619
 Max.   : 5.2415      Max.   : 4.2183            Max.   : 5.7094
   airline
 Length:56
 Class :character
 Mode  :character
```

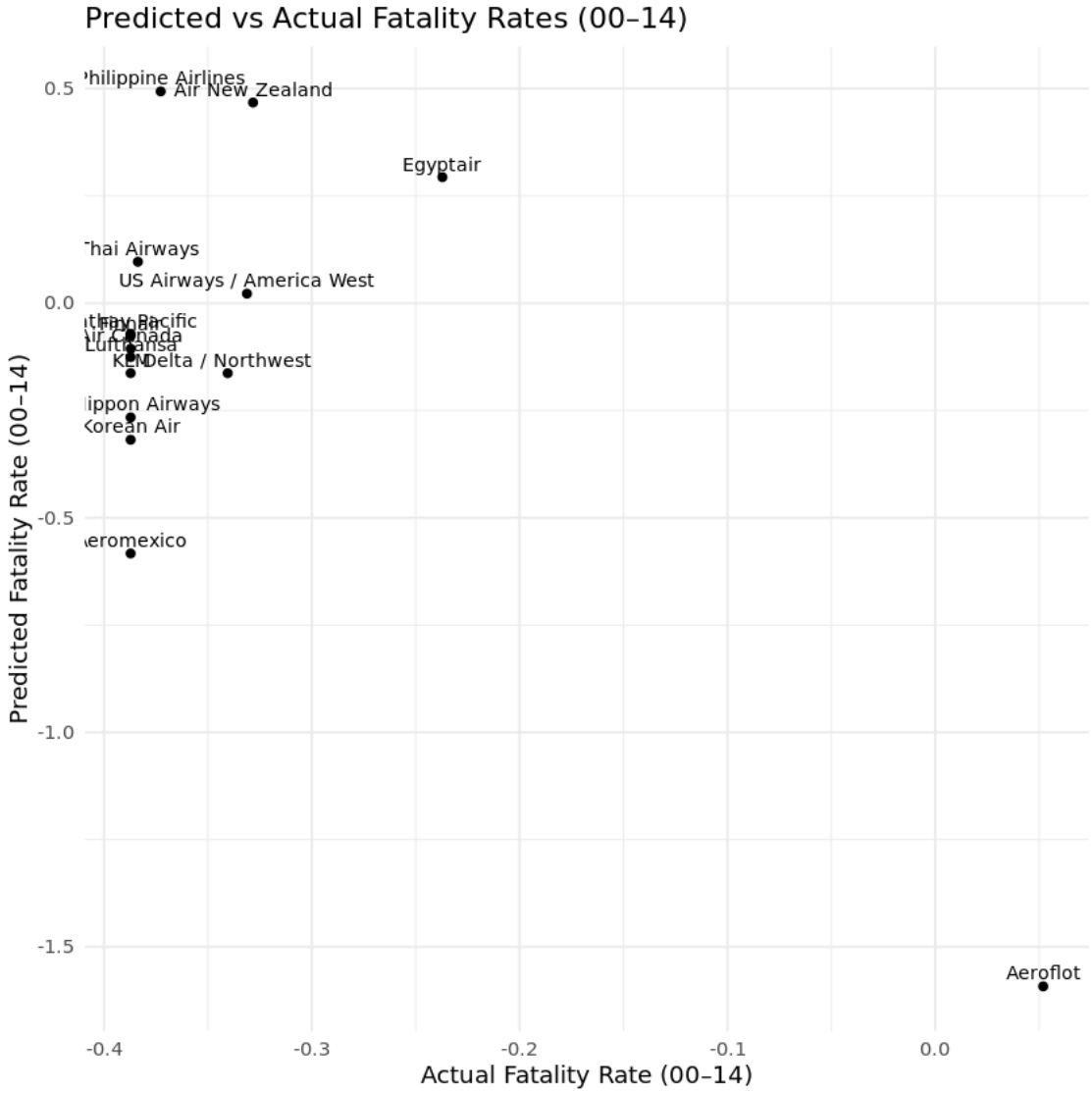
```
Call:
lm(formula = f_model, data = d %>% filter(train))

Residuals:
    Min      1Q  Median      3Q     Max
-2.1574 -0.2410 -0.1021  0.2128  1.7337

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.05264    0.11513   0.457    0.650
incidents_85_99_rate       -0.22765    0.23799  -0.957    0.345
fatal_accidents_85_99_rate -0.15132    0.24225  -0.625    0.536
```

```
fatalities_85_99_rate       0.02293    0.14869   0.154    0.878
incidents_00_14_rate       -0.20880    0.13688  -1.525    0.136
fatal_accidents_00_14_rate  0.97016    0.13154   7.376 1.26e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7281 on 35 degrees of freedom
Multiple R-squared:  0.6466,        Adjusted R-squared:  0.5961
F-statistic: 12.81 on 5 and 35 DF,  p-value: 4.122e-07
```



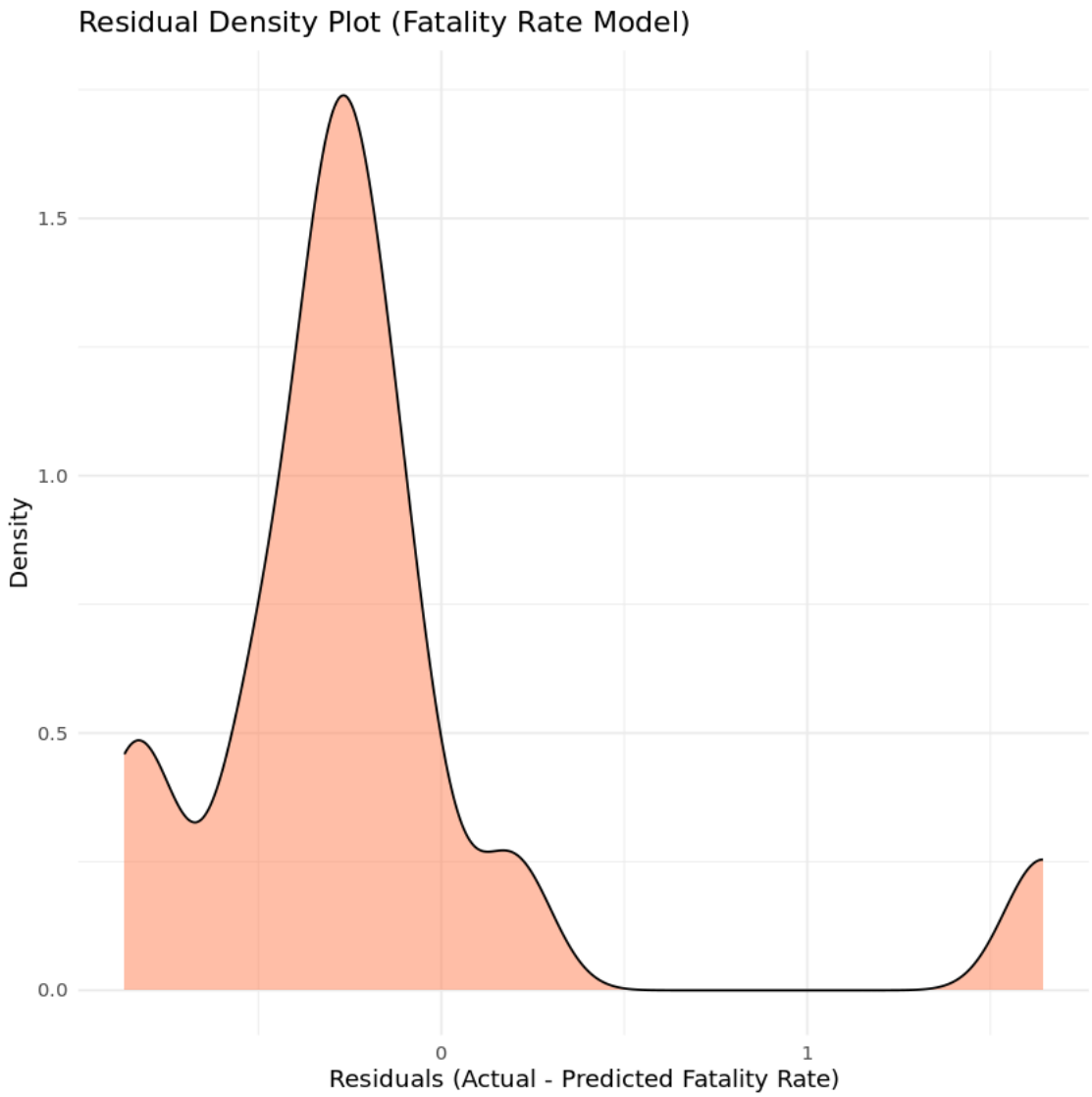


Figure 7. This figure presents the results of the linear regression model. The top panel shows a scatterplot of the Actual Fatality Rate (2000–2014) versus the Predicted Fatality Rate, illustrating the model's performance and highlighting outliers like Aeroflot. The bottom panel displays the density plot of the residuals (the difference between the actual and predicted rates), which assesses the model's assumptions and shows a non-normal, multi-modal distribution, confirming the presence of significant prediction errors for high-risk airlines.

Conclusion: The regression study, while confirming the strong and expected relationship that a higher 2000–2014 Fatal Accident Rate is the main driver of the Fatality Rate (coefficient 0.97016, $p < 0.001$), proved worthwhile by highlighting two crucial findings. First, it quantitatively disconfirmed the predictive power of long-term history, showing that all 1985–1999 safety rates were not statistically significant predictors of the recent fatality rate ($p > 0.8$ for fatalities_85_99_rate), suggesting an airline's safety profile is dynamic and can change drastically over 15 years. Second, the residual analysis was critical, as it identified specific airlines (e.g., Aeroflot) whose actual

fatality rates were extreme outliers compared to the model's prediction, thereby directing future investigative efforts towards the unique, non-rate-based factors that drive the highest-risk events.

Next Steps: To extend the analysis, the next step is to perform a separate cluster analysis using the raw counts of incidents and fatalities rather than the standardized rate variables. While the earlier PCA and K-means clustering focused on comparing airlines based on safety profiles normalized for operational volume, examining clusters formed from unscaled counts will help identify whether airlines with exceptionally high incident numbers—such as Aeroflot during 1985–1999—exhibit fundamentally different characteristics from those with high incident rates. By comparing the raw-count clusters with the previously generated rate-based clusters, we can assess whether high-volume carriers experience a distinct pattern of safety risk compared with smaller airlines, thereby providing deeper insight into how operational scale influences observed safety performance.