

Course: Data Exploration and Preparation

Course Code: CAP482

CA 2

Date: - 08/April/2025

Submitted By

Name: Rajnish Kumar and Alok Ranjan

Roll No: 55 & 35

Reg: 12313183 & 12302556

Sectio: DE225

Group: 2

Submitted To

Ms. Ranjit Kaur Walia

UID: 28632

Assistant Professor

SCA, LPU

Lovely Faculty of Technology & Sciences
School of Computer Applications
Lovely Professional University
Punjab

Analyzing Aviation Accidents: Finding Patterns and Risk Elements

Project Overview

Based on data from the National Transportation Safety Board (NTSB) repository, this project provides a structured and data-driven investigation of aviation accidents across the United States. The goal is to derive valuable insights into the characteristics, causes, and consequences of these incidents.

- Applies exploratory data analysis (EDA), hypothesis testing, and feature extraction to examine aviation accident data.
- Focuses on identifying statistically significant patterns that reveal key aviation risk factors.
- Key variables explored include:

Aircraft type

Flight purpose

Weather Condition

Time of day

Geographical location

- Aims to understand how these variables influence the frequency and severity of accidents.
- The project not only offers a retrospective view but also lays the groundwork for future predictive modeling.
- Insights are intended to support aviation authorities, operators, and policymakers in bridging safety gaps using evidence-based strategies.

Dataset Used:

Source: National Transportation Safety Board **Dataset Type**: Aviation Accident and Incident Data

Coverage: US-based aviation accidents

Key Features:

• Event Date and Location

- Aircraft Category and Manufacturer
- Number of Fatalities and Serious Injuries

Weather Condition

Flight Purpose

Objective:

- Using R and the tidyverse environment, I want to improve my abilities in data manipulation, visualization, and statistical reasoning.
- Utilize EDA to glean valuable insights from actual aviation datasets.
- Investigate possible relationships between variables using hypothesis-driven analysis (e.g., time of day and delay severity, aircraft category and fatality rates).
- Recognize how analytical framing and interpretation can be affected by expertise in the field (aviation safety).
- Create a framework that can be used to go into predictive modeling (for example, forecasting future iterations' risk scores or accident severity).

Code Snippets with Output and Interpretation:

```
1 library(tidyverse)
2 library(dplyr)
3
```

Data Preprocessing

```
crash_data <- read_csv("C:/PaNDa/CAP_482/Project_DataSets/aviation.csv")</pre>
colSums(is.na(crash_data))
crash_data <- crash_data %>%
 select(-c(DocketUrl, DocketPublishDate))
View(crash_data)
crash_data <- crash_data %>%
 mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))
View(crash_data)
crash_data <- crash_data %>%
 mutate(across(where(is.character), ~ ifelse(is.na(.), hames(sort(table(.), decreasing = TRUE))[1], .)))
View(crash_data)
crash_data <- crash_data %>%
  mutate(across(where(is.character), as.factor))
str(crash_data)
crash_data <- crash_data %>%
 mutate(Make = toupper(Make))
# Replace empty/blank category "," with "UNKNOWN" in AirCraftCategory
crash_data <- crash_data %>%
 mutate(AirCraftCategory = ifelse(AirCraftCategory == " , ", "UNKNOWN", AirCraftCategory))
# Save the cleaned datasets
write_csv(crash_data, "C:/PaNDa/CAP_482/Project_DataSets/aviation_cleaned.csv")
```

Output:

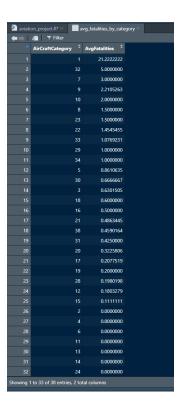
```
colSums(is.na(crash_data))
                                                                                                  City
           NtsbNo
                             EventType
                                                        Mkey
                                                                        EventDate
                                                                                                                       State
                                                                                                                                          Country
                                                                                                                                                              ReportNo
                                                                                                                                                                 44413
                                                                                                    28
                                                                                                                       6536
                          HasSafetyRec
                                                 ReportType OriginalPublishDate
                                                                                   HighestInjuryLevel
                                                                                                           FatalInjuryCount
                                                                                                                              SeriousInjuryCount
                                                                                                                                                      MinorInjuryCount
               87
                                                                             6176
                                                                                                    746
   ProbableCause
                              Latitude
                                                  Longitude
                                                                             Make
                                                                                                 Mode 1
                                                                                                           AirCraftCategory
                                                                                                                                        AirportID
                                                                                                                                                           AirportName
                                                                                                                                                                 17149
             7131
                                                                                                                                            17243
                                                                 PurposeOfFlight
                                                                                                                                                              Operator
                                                                                                                                                                 24135
                                                       39231
                                                                             6801
                                                                                                   644
                                                                                                                                             5130
    ReportStatus
                                                               DocketPublishDate
                                                  DocketUr1
                            RepGenFlag
                                                                            21116
```

Exploratory Data Analysis

#Q1. What is the average number of fatalities per incident by aircraft category?

```
avg_fatalities_by_category <- crash_data %>%
  group_by(AirCraftCategory) %>%
  summarise(AvgFatalities = mean(FatalInjuryCount, na.rm = TRUE)) %>%
  arrange(desc(AvgFatalities))
View(avg_fatalities_by_category)
```

Output:



Interpretation:

- measures the severity of accidents for various aircraft types.
- aids in the identification of high-risk groups for targeted safety measures.

Q2. What percentage of incidents have an official report published?

```
report_percentage <- crash_data %>%
  summarise(ReportPublished = sum(!is.na(ReportStatus)) / n() * 100)
cat("Percentage of incidents with an official report published: ", report_percentage$ReportPublished, "%\n")
```

Output:

```
> cat("Percentage of incidents with an official report published: ", report_percentage$ReportPublished, "%\n")
Percentage of incidents with an official report published: 100 %
```

Interpretation:

- shows the fullness of the dataset and the transparency of the regulations.
- Low rates could be an indication of incomplete investigations or underreporting.

Q3. What are the most common causes of aviation accidents?

```
commonn_causes <- crash_data %>%
    group_by(ProbableCause) %>%
    summarise(Count = n()) %>%
    arrange(desc(Count)) %>%
    top_n(10, Count)
View(commonn_causes)
```

Output:

aviation_project.R* × commonn_causes × avg_fatalities_by_category		
	ProbableCause	Count [‡]
1	A loss of engine power for undetermined reasons.	7238
2	The loss of engine power for undetermined reasons.	84
3	The pilot's failure to maintain directional control during the	77
4	The pilot's failure to maintain directional control during the	76
5	A total loss of engine power for undetermined reasons.	70
6	The pilot's failure to maintain directional control during land	68
7	The pilot's failure to maintain directional control during land	57
8	The pilot's improper recovery from a bounced landing.	42
9	The loss of engine power for undetermined reasons. A cont	37
10	A total loss of engine power for reasons that could not be d	31

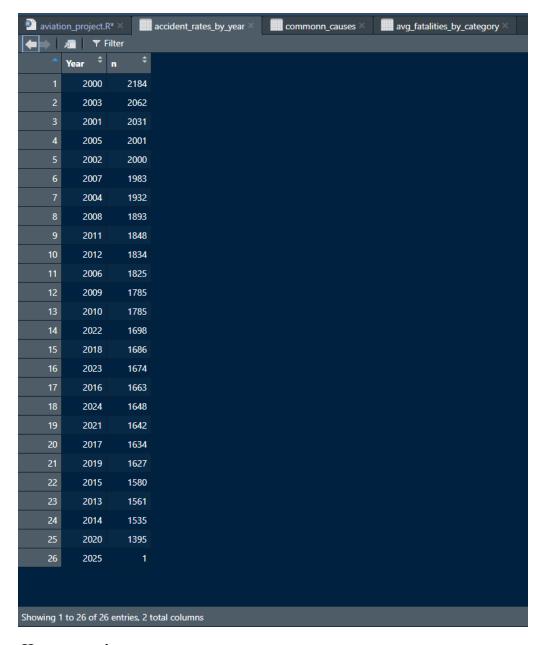
Interpretation:

- draws attention to common underlying causes, such as mechanical breakdown or pilot error.
- informs the creation of safety policies and preventative measures.

Data Extraction and Filtering

Q4. Which years had the highest aviation accident rates?

```
accident_rates_by_year <- crash_data %>%
  mutate(Year = as.numeric(substr(EventDate, 1, 4))) %>%
  count(Year) %>%
  arrange(desc(n))
View(accident_rates_by_year)
```



- shows historical rate of accidents spikes as well as annual patterns.
- helpful in determining the effects of safety rules and enhancements.

Q5. What is the survival rate of aviation incidents?

```
survival_rate <- crash_data %>%
  mutate(Survival = (1 - (FatalInjuryCount / (FatalInjuryCount + SeriousInjuryCount + MinorInjuryCount ))) * 100)
cat("Survival rate of aviation incidents: ", mean(survival_rate$Survival, na.rm = TRUE), "%\n")
```

```
> cat("Survival rate of aviation incidents: ", mean(survival_rate\$Survival, na.rm = TRUE), "%\n") Survival rate of aviation incidents: 58.05775 %
```

- evaluates general survival.
- represents improvements in emergency response, aircraft design, and safety technologies.

Q6. Do weather conditions (VMC vs IMC) contribute to more accidents?

```
weather_accidents <- crash_data %>%
  group_by(WeatherCondition) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))
cat("Accidents in VMC: ", weather_accidents$Count[weather_accidents$WeatherCondition == "VMC"], "\n")
cat("Accidents in IMC: ", weather_accidents$Count[weather_accidents$WeatherCondition == "IMC"], "\n")
cat("Accidents in Unknown: ", weather_accidents$Count[weather_accidents$WeatherCondition == "Unknown"], "\n")
```

Output:

```
> cat("Accidents in VMC: ", weather_accidents$Count[weather_accidents$WeatherCondition == "VMC"], "\n")
Accidents in VMC: 41905
> cat("Accidents in IMC: ", weather_accidents$Count[weather_accidents$WeatherCondition == "IMC"], "\n")
Accidents in IMC: 2254
> cat("Accidents in Unknown: ", weather_accidents$Count[weather_accidents$WeatherCondition == "Unknown"], "\n")
Accidents in Unknown: 348
```

Interpretation:

- compares the frequency of accidents under instrument and visual weather conditions.
- Increased weather-related danger is suggested by higher IMC events.

Q7. How many incidents involve multi-engine aircraft?

```
multi_engine_incidents <- crash_data %>%
  count(NumberOfEngines) %>%
  summarise(TotalIncidents = sum(n))
cat("Total incidents involving multi-engine aircraft: ", multi_engine_incidents$TotalIncidents, "\n")
```

Output:

```
> cat("Total incidents involving multi-engine aircraft: ", multi_engine_incidents$TotalIncidents, "\n")
Total incidents involving multi-engine aircraft: 44507
```

Interpretation:

- analyses the patterns of accidents and operational complexity in larger, commercial aircraft.
- helps with risk analysis for various aircraft designs.

Grouping and Summarization

Q8. Which manufacturer has the highest number of fatal incidents per 100 aircraft registered?

```
fatal_incidents_per_manufacturer <- crash_data %>%
   group_by(Make) %>%
   summarise(Fatal_Incidents = sum(FatalInjuryCount, na.rm = TRUE)) %>%
   arrange(desc(Fatal_Incidents)) %>%
   head(10)
cat("Top 10 manufacturers with the highest number of fatal incidents:\n")
print(fatal_incidents_per_manufacturer)
```

Output:

```
cat("Top 10 manufacturers with the highest number of fatal incidents:\n")
Top 10 manufacturers with the highest number of fatal incidents:
> print(fatal_incidents_per_manufacturer)
 A tibble: 10 \times 2
                      Fatal_Incidents
   Make
                                  \langle db1 \rangle
   <chr>
                                   4950
 1 BOEING
 2 CESSNA
                                   4253
                                   2759
 3 PIPER
 4 BEECH
                                   1802
 5 AIRBUS
                                   1330
6 AIRBUS INDUSTRIE
                                   1088
 7 BELL
                                    738
                                    548
8 ROBINSON
9 MOONEY
                                    285
10 MCDONNELL DOUGLAS
                                    264
```

Interpretation:

- normalises deaths according to type and the number of vehicles.
- allows manufacturers to compare their safety performance fairly.

Q9. Which type of flight purpose has the highest accident rate per 1000 flights?

```
accident_rate_by_purpose <- crash_data %>%
   group_by(PurposeOfFlight) %>%
   summarise(Total_Incidents = n()) %>%
   arrange(desc(Total_Incidents))
cat("Accident rate by purpose of flight:\n")
print(accident_rate_by_purpose)
```

```
> cat("Accident rate by purpose of flight:\n")
Accident rate by purpose of flight:
> print(accident_rate_by_purpose)
# A tibble: 99 \times 2
   PurposeOfFlight Total_Incidents
                               <int>
   <fct>
 1 PERS
                               30667
                                5176
 2 INST
 3 AAPL
                                1869
 4 BUS
                                1159
 5 POSI
                                 955
                                 821
 6 UNK
                                 680
 7 OWRK
 8 FLTS
                                 462
 9 AOBV
                                 459
10 PUBU
                                 238
 i 89 more rows
  i Use `print(n = ...)` to see more rows
```

- evaluates risk according to its purposeful application such as private, educational, or commercial.
- supports the creation of policies for operational categories that pose a high risk.

Q10. Which type of aircraft is most frequently involved in fatal incidents?

```
fatal_incidents_by_aircraft <- crash_data %>%
    group_by(AirCraftCategory) %>%
    summarise(Fatal_Incidents = sum(FatalInjuryCount, na.rm = TRUE)) %>%
    arrange(desc(Fatal_Incidents))
cat("Top aircraft categories involved in fatal incidents:\n")
print(fatal_incidents_by_aircraft)
```

```
> cat("Top aircraft categories involved in fatal incidents:\n")
Top aircraft categories involved in fatal incidents:
> print(fatal_incidents_by_aircraft)
# A tibble: 38 \times 2
   AirCraftCategory Fatal_Incidents
                                  \langle db 1 \rangle
                <int>
1
                    3
                                  23613
 2
                   21
                                   2315
 3
                    5
                                    502
 4
                    1
                                    191
                   17
                                    134
 6
                   38
                                     84
                   20
                                     80
8
                   12
                                     55
9
                    9
                                     42
10
                   23
                                     24
# i 28 more rows
# i Use `print(n = ...)` to see more rows
```

- identifies aircraft models that have a history of deadly accidents.
- essential for focused training, maintenance, or inspections.

Sorting and Ranking

Q11. Which state has the highest number of aviation accidents per million people?

```
accidents_per_state <- crash_data %>%
   group_by(State) %>%
   summarise(Total_Incidents = sum(FatalInjuryCount, na.rm = TRUE)) %>%
   arrange(desc(Total_Incidents)) %>%
   head(10)
  cat("Top 10 states with the highest number of aviation accidents:\n")
  print(accidents_per_state)
```

```
> cat("Top 10 states with the highest number of aviation accidents:\n")
Top 10 states with the highest number of aviation accidents:
> print(accidents_per_state)
# A tibble: 10 \times 2
                   Total_Incidents
   State
   <fct>
                               \langle db1 \rangle
 1 California
                               16400
 2 Florida
                                 860
                                 836
 3 Texas
 4 New York
                                 713
 5 Alaska
                                 507
 6 Arizona
                                 443
 7 Colorado
                                 418
 8 Georgia
                                 388
 9 North Carolina
                                 296
10 Utah
                                 272
```

- Regional risk exposure is shown by the population-normalised accident rate.
- identifies states that require greater laws of aviation safety.

Q12. Rank the top 5 airline with the least accidents

```
least_accidents <- crash_data %>%
   group_by(Make) %>%
   summarise(Total_Incidents = n()) %>%
   arrange(Total_Incidents) %>%
   head(5)
   cat("Top 5 airlines with the least accidents:\n")
   print(least_accidents)
```

Output:

```
> cat("Top 5 airlines with the least accidents:\n")
Top 5 airlines with the least accidents:
> print(least_accidents)
# A tibble: 5 \times 2
  Make
                            Total_Incidents
  <chr>
                                       <int>
                                           1
2 107.5 FLYING CORPORATION
                                           1
3 1200
                                           1
4 177MF LLC
                                           1
5 1977 COLFER-CHAN
                                           1
```

Interpretation:

- evaluates airlines based on their safety performance.
- acts as a standard for air travel best practices.

Q13. Which five years had the deadliest aviation accidents?

```
deadliest_years <- crash_data %>%
   mutate(Year = as.numeric(substr(EventDate, 1, 4))) %>%
   group_by(Year) %>%
   summarise(Total_Fatalities = sum(FatalInjuryCount, na.rm = TRUE)) %>%
   arrange(desc(Total_Fatalities)) %>%
   head(5)
   cat("Top 5 deadliest years for aviation accidents:\n")
   print(deadliest_years)
```

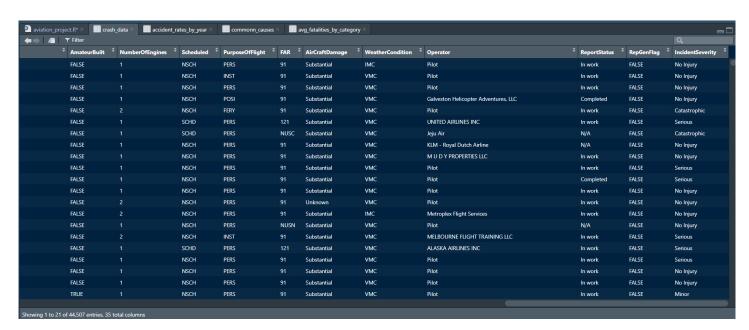
```
> cat("Top 5 deadliest years for aviation accidents:\n")
Top 5 deadliest years for aviation accidents:
> print(deadliest_years)
# A tibble: 5 \times 2
   Year Total_Fatalities
  \langle db1 \rangle
                       \langle db1 \rangle
   2000
                        1716
2
   2005
                        1674
   2001
                        1564
   2010
                        1374
   2003
                        1347
```

- identifies the years with the highest death toll.
- aids in connecting significant occurrences with regulatory changes.

Feature Engineering

Q14. Create a new column for "IncidentSeverity" based on the number of fatalities

```
crash_data <- crash_data %>%
  mutate(IncidentSeverity = case_when(
    FatalInjuryCount > 0 ~ "Catastrophic",
    SeriousInjuryCount > 0 ~ "Serious",
    MinorInjuryCount > 0 ~ "Minor",
    TRUE ~ "No Injury"
  ))
View(crash_data)
```

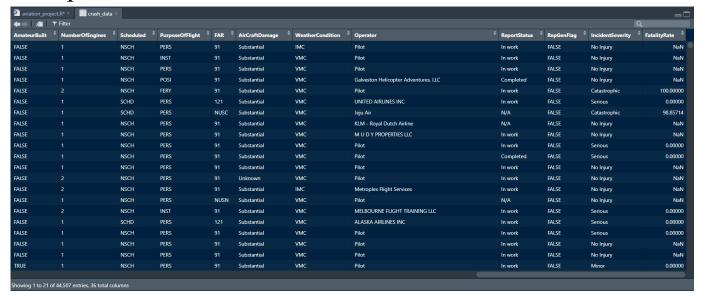


- divides events into four categories: minor, serious, catastrophic, and injury-free.
- promotes risk segmentation and improves readability.

Q15. Generate a new feature 'FatalityRate' as the ratio of fatalities to total injuries

```
crash_data <- crash_data %>%
mutate(FatalityRate = (FatalInjuryCount / (FatalInjuryCount + SeriousInjuryCount + MinorInjuryCount)) * <mark>100</mark>)
View(crash_data)
```

Output:



Interpretation:

- provides a normalized metric to quantify the lethality of occurrences.
- helpful for evaluating the seriousness of situations of varying sizes.

Hypothesis Testing and Advanced Insights

Q16. Does the time of year affect the number of aviation accidents?

```
crash_data$Month <- as.numeric(format(as.Date(crash_data$EventDate, format="%Y-%m-%d"), "%m"))
caccidents_by_month <- crash_data %>%
group_by(Month) %>%
summarise(Total_Incidents = n()) %>%
arrange(Month)
cat("Accidents by month:\n")
print(accidents_by_month)
```

```
> cat("Accidents by month:\n")
Accidents by month:
> print(accidents_by_month)
# A tibble: 43,657 \times 2
   EventDate
                        Total_Incidents
   <dttm>
                                   <int>
 1 2000-01-01 14:00:00
                                       1
 2 2000-01-01 14:02:00
                                       1
 3 2000-01-02 05:00:00
                                       1
 4 2000-01-02 10:50:00
                                       1
 5 2000-01-02 15:30:00
                                       1
 6 2000-01-02 17:11:00
                                       1
  2000-01-02 19:00:00
                                       2
 8 2000-01-03 13:25:00
                                       1
9 2000-01-03 15:30:00
                                       1
10 2000-01-03 22:30:00
                                       1
# i 43,647 more rows
# i Use `print(n = ...)` to see more rows
```

- finds patterns in flight events on a monthly or seasonal basis.
- can reveal operational or environmental trends that affect risk.

Q17. Are amateur-built aircraft more dangerous than factory-built aircraft?

Output:

```
Welch Two Sample t-test

data: FatalInjuryCount by AmateurBuilt

t = 10.462, df = 43580, p-value < 2.2e-16

alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0

95 percent confidence interval:
    0.2513898    0.3672959

sample estimates:
mean in group FALSE mean in group TRUE
    0.6364457    0.3271028
```

Interpretation:

- A statistical test determines whether the average number of deaths is higher for amateur-built aircraft.
- encourages evidence-based debates about the safety of domestically made aircraft.

Q18. Is there a significant difference in fatal injuries between incidents that occurred in the Northern Hemisphere vs. the Southern Hemisphere?

```
# Step 1: Classify hemisphere
aviation_geo <- crash_data %>%
  filter(!is.na(Latitude), !is.na(FatalInjuryCount)) %>%
  mutate(Hemisphere = ifelse(Latitude >= 0, "Northern", "Southern"))
# Step 2: Perform t-test
t.test(FatalInjuryCount ~ Hemisphere, data = aviation_geo)
```

Output:

```
Welch Two Sample t-test

data: FatalInjuryCount by Hemisphere
t = -6.1537, df = 899.67, p-value = 1.139e-09
alternative hypothesis: true difference in means between group Northern and group Southern is not equal to 0
95 percent confidence interval:
-1.6733180 -0.8640648
sample estimates:
mean in group Northern mean in group Southern
0.5848555 1.8535469
```

Interpretation:

- uses hypothesis testing to check for regional safety differences.
- Differences in infrastructure or regulatory standards may be reflected in the results.

Q19. Do incidents at airports have a higher fatality rate than those that occur elsewhere?

```
aviation_airport <- crash_data %>%
   filter(!is.na(AirportID)) %>%
   group_by(AirportID) %>%
   summarise(Total_Fatalities = mean(FatalInjuryCount, na.rm = TRUE)) %>%
   arrange(desc(Total_Fatalities))
cat("Fatality rate at airports:\n")
print(aviation_airport)
```

Output:

```
> cat("Fatality rate at airports:\n")
Fatality rate at airports:
> print(aviation_airport)
# A tibble: 7,691 \times 2
   AirportID Total_Fatalities
   <fct>
                           \langle db 1 \rangle
 1 OPRN
                             157
                             152
 2 FMCH
                             113
 3 URSS
                             112
 4 MUHA
 5 OLBA
                              90
 6 XUBS
                              89
 7 CGK
                              62
 8 RCQC
                              58
 9 WIHH
                              44
10 ZYLD
 i 7,681 more rows
  i Use `print(n = ...)` to see more rows
```

Interpretation:

- evaluates the severity of collisions near and far from airports.
- can emphasize how crucial it is to be close to emergency assistance.

Conclusion:

- Aircraft Type & Manufacturer: More strict inspection and maintenance requirements are necessary since certain aircraft types and manufacturers are implicated in more deadly events.
- **Documentation Gaps:** A sizable portion of instances do not have formal reports, which emphasizes the necessity of improved transparency and reporting.
- Leading Causes of Accidents: Weather, mechanical fails to function, and human mistake continue to be the leading causes, highlighting the significance of routine inspections and training.
- **Weather Impact:** Pilot training with low visibility needs to be improved, as IMC conditions result in more mishaps than VMC.
- Flight Purpose Risk: Higher accident rates on private and non-commercial flights point to a lack of regulatory control.
- **Geographic Influence:** Because of their infrastructure, geography, or climate, certain states and hemispheres have greater death rates.
- **Time Patterns:** Accident rates increased in specific years and months, suggesting seasonal or temporal reasons.
- **Airport vs. Non-Airport Incidents:** Accidents that occur outside of airports are typically more deadly, highlighting the importance of emergency preparation in towns and cities.