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:

```r
library(tidyverse)
library(dplyr)
```

Data Preprocessing

```r
# ---------------------------------- Data Preprocessing ----------------------------------

# Load the datasets
crash_data <- read_csv("C:/PaNDa/CAP_482/Project_DataSets/aviation.csv")

# Missing values
colSums(is.na(crash_data))
# Remove columns with more than 50% missing values
crash_data <- crash_data %>%
  select(-c(DocketUrl, DocketPublishDate))
View(crash_data)

# Fill Missing Numeric Values with Mean/Median
crash_data <- crash_data %>%
  mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))
View(crash_data)

# Fill Missing Categorical Values with Mode
crash_data <- crash_data %>%
  mutate(across(where(is.character), ~ ifelse(is.na(.), names(sort(table(.), decreasing = TRUE))[1], .)))
View(crash_data)

# Convert categorical variables to factors
crash_data <- crash_data %>%
  mutate(across(where(is.character), as.factor))
str(crash_data)

# Convert Manufacturer in uppercase
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:

```
> # Missing values
> colSums(is.na(crash_data))
         NtsbNo          EventType               Mkey          EventDate               City              State            Country           ReportNo
              0                  0                  0                  0                 28               6536                 43              44413
              N       HasSafetyRec         ReportType OriginalPublishDate HighestInjuryLevel   FatalInjuryCount SeriousInjuryCount  MinorInjuryCount
             87                  0                  0               6176                746                  0                  0                  0
  ProbableCause           Latitude          Longitude               Make              Model   AirCraftCategory          AirportID        AirportName
           7131                  0                  0                 52                 65                441              17243              17149
   AmateurBuilt    NumberOfEngines          Scheduled    PurposeOfFlight                FAR     AirCraftDamage   WeatherCondition           Operator
              0               5275              39231               6801                644                338               5130              24135
   ReportStatus         RepGenFlag          DocketUrl  DocketPublishDate
              0                  0              21116              21116
```

```
> str(crash_data)
tibble [44,507 × 34] (S3: tbl_df/tbl/data.frame)
 $ NtsbNo            : Factor w/ 44507 levels "ANC00FA024","ANC00FA052",..: 25234 2575 44506 11169 25169 15984 16017 16018 11168 11166 ...
 $ EventType         : Factor w/ 3 levels "ACC","INC","OCC": 1 1 1 1 1 1 1 1 1 1 ...
 $ Mkey              : num [1:44507] 199500 199498 199524 199496 199492 ...
 $ EventDate         : POSIXct[1:44507], format: "2025-01-01 02:20:00" "2024-12-31 14:30:00" "2024-12-31 14:16:00" "2024-12-31 13:20:00" ...
 $ City              : Factor w/ 13793 levels "40 nm vicinity south of Lake Jackson",..: 8366 307 7010 4333 9397 4842 8222 9088 8017 9317 ...
 $ State             : Factor w/ 57 levels "Alabama","Alaska",..: 12 2 35 50 39 6 6 6 19 20 ...
 $ Country           : Factor w/ 190 levels "Afghanistan",..: 179 179 179 179 179 179 141 128 179 179 ...
 $ ReportNo          : Factor w/ 94 levels "AAB0202","AAB0203",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ N                 : Factor w/ 42077 levels "(H-VISTA",",",..: 33019 14854 28743 18403 28076 26450 2592 39899 19803 29142 ...
 $ HasSafetyRec      : logi [1:44507] FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ ReportType        : Factor w/ 4 levels "BoardBrief","DirectorBrief",..: 2 2 2 2 2 2 3 2 2 2 ...
 $ OriginalPublishDate: POSIXct[1:44507], format: NA NA NA "2025-01-30 05:00:00" ...
 $ HighestInjuryLevel : Factor w/ 4 levels "Fatal","Minor",..: 3 3 3 3 1 4 1 3 3 4 ...
 $ FatalInjuryCount   : num [1:44507] 0 0 0 0 1 0 173 0 0 0 ...
 $ SeriousInjuryCount : num [1:44507] 0 0 0 0 0 1 2 0 0 1 ...
 $ MinorInjuryCount   : num [1:44507] 0 0 0 0 0 0 0 0 0 0 ...
 $ ProbableCause      : Factor w/ 34533 levels "'THIS CASE WAS MODIFIED MAY 30, 2006.'The airplane's inadvertent impact with one of several deer that had enter"| __truncated_
_,..: 1443 1443 1443 29635 1443 1443 1443 1443 1443 1443 ...
 $ Latitude          : num [1:44507] 26.2 61.2 35.9 29.3 39 ...
 $ Longitude         : num [1:44507] -81.8 -149.8 -106.3 -94.8 -83.4 ...
 $ Make              : Factor w/ 6463 levels ",","107.5 Flying Corporation",..: 652 1227 1227 4963 1227 812 812 812 1826 1553 ...
 $ Model             : Factor w/ 8860 levels "-","-269C","(EX) RV-6",..: 1739 260 365 6730 664 1354 1093 1093 3405 3340 ...
 $ AirCraftCategory  : Factor w/ 38 levels ",",",AIR,AIR",..: 3 3 3 21 3 3 3 3 3 3 ...
 $ AirportID         : Factor w/ 7691 levels "-","---","(AZ38)",..: 2154 2132 4953 5780 5781 5781 5781 5781 5622 5781 ...
 $ AirportName       : Factor w/ 14914 levels "---","-70.8301542",..: 9243 8574 10775 10815 10775 10775 10775 10775 10775 10775 ...
 $ AmateurBuilt      : Factor w/ 7 levels "FALSE","FALSE,FALSE",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ NumberOfEngines   : Factor w/ 34 levels ",",",,",",0",..: 11 11 11 11 18 11 11 11 11 11 ...
 $ Scheduled         : Factor w/ 3 levels "NSCH","SCHD",..: 1 1 1 1 1 2 2 1 1 1 ...
 $ PurposeOfFlight   : Factor w/ 99 levels ",",",,",",AOBV",..: 61 48 61 75 37 61 61 61 61 61 ...
 $ FAR               : Factor w/ 69 levels ",","091,ARMF",..: 39 39 39 39 39 26 48 39 39 39 ...
 $ AirCraftDamage    : Factor w/ 28 levels ",","Destroyed",..: 19 19 19 19 19 19 19 19 19 19 ...
 $ WeatherCondition  : Factor w/ 3 levels "IMC","Unknown",..: 1 3 3 3 3 3 3 3 3 3 ...
 $ Operator          : Factor w/ 16956 levels "--","- M/s Jindal Steel & Power Ltd.",..: 11965 11965 11965 6228 11965 15729 8175 9086 9736 11965 ...
 $ ReportStatus      : Factor w/ 3 levels "Completed","In work",..: 2 2 2 1 2 2 3 3 2 2 ...
 $ RepGenFlag        : logi [1:44507] FALSE FALSE FALSE FALSE FALSE FALSE ...
```

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:

aviation_project.R* | avg_fatalities_by_category

	AirCraftCategory	AvgFatalities
1	1	21.2222222
2	32	5.0000000
3	7	3.0000000
4	9	2.2105263
5	10	2.0000000
6	8	1.5000000
7	23	1.5000000
8	22	1.4545455
9	33	1.0769231
10	29	1.0000000
11	34	1.0000000
12	5	0.8610635
13	30	0.6666667
14	3	0.6301505
15	18	0.6000000
16	16	0.5000000
17	21	0.4863445
18	38	0.4590164
19	31	0.4250000
20	20	0.3225806
21	17	0.2077519
22	19	0.2000000
23	28	0.1980198
24	12	0.1803279
25	15	0.1111111
26	2	0.0000000
27	4	0.0000000
28	6	0.0000000
29	11	0.0000000
30	13	0.0000000
31	14	0.0000000
32	24	0.0000000

Showing 1 to 33 of 38 entries, 2 total columns

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

Filter

	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)
```

Output:

aviation_project.R* | accident_rates_by_year | commonn_causes | avg_fatalities_by_category

Filter

	Year	n
1	2000	2184
2	2003	2062
3	2001	2031
4	2005	2001
5	2002	2000
6	2007	1983
7	2004	1932
8	2008	1893
9	2011	1848
10	2012	1834
11	2006	1825
12	2009	1785
13	2010	1785
14	2022	1698
15	2018	1686
16	2023	1674
17	2016	1663
18	2024	1648
19	2021	1642
20	2017	1634
21	2019	1627
22	2015	1580
23	2013	1561
24	2014	1535
25	2020	1395
26	2025	1

Showing 1 to 26 of 26 entries, 2 total columns

IInterpretation:

- 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")
```

Output:

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

Interpretation:

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

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

```r
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?

```r
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?

```r
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 × 2
   Make               Fatal_Incidents
   <chr>                        <dbl>
 1 BOEING                        4950
 2 CESSNA                        4253
 3 PIPER                         2759
 4 BEECH                         1802
 5 AIRBUS                        1330
 6 AIRBUS INDUSTRIE              1088
 7 BELL                           738
 8 ROBINSON                       548
 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?

```r
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)
```

Output:

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

Interpretation:

- 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)
```

Output:

```
> cat("Top aircraft categories involved in fatal incidents:\n")
Top aircraft categories involved in fatal incidents:
> print(fatal_incidents_by_aircraft)
# A tibble: 38 × 2
   AirCraftCategory Fatal_Incidents
              <int>           <dbl>
 1                3           23613
 2               21            2315
 3                5             502
 4                1             191
 5               17             134
 6               38              84
 7               20              80
 8               12              55
 9                9              42
10               23              24
# ℹ 28 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Interpretation:

- 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)
```

Output:

```
> 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 × 2
   State          Total_Incidents
   <fct>                    <dbl>
 1 California               16400
 2 Florida                    860
 3 Texas                      836
 4 New York                   713
 5 Alaska                     507
 6 Arizona                    443
 7 Colorado                   418
 8 Georgia                    388
 9 North Carolina             296
10 Utah                       272
```

Interpretation:

- 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 × 2
  Make                       Total_Incidents
  <chr>                                <int>
1 ,                                        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)
```

Output:

```
> cat("Top 5 deadliest years for aviation accidents:\n")
Top 5 deadliest years for aviation accidents:
> print(deadliest_years)
# A tibble: 5 × 2
   Year Total_Fatalities
  <dbl>            <dbl>
1  2000             1716
2  2005             1674
3  2001             1564
4  2010             1374
5  2003             1347
```

Interpretation:

- 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)
```

Output:

aviation_project.R* | crash_data | accident_rates_by_year | commonn_causes | avg_fatalities_by_category

Filter

	AmateurBuilt	NumberOfEngines	Scheduled	PurposeOfFlight	FAR	AirCraftDamage	WeatherCondition	Operator	ReportStatus	RepGenFlag	IncidentSeverity
	FALSE	1	NSCH	PERS	91	Substantial	IMC	Pilot	In work	FALSE	No Injury
	FALSE	1	NSCH	INST	91	Substantial	VMC	Pilot	In work	FALSE	No Injury
	FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	No Injury
	FALSE	1	NSCH	POSI	91	Substantial	VMC	Galveston Helicopter Adventures, LLC	Completed	FALSE	No Injury
	FALSE	2	NSCH	FERY	91	Substantial	VMC	Pilot	In work	FALSE	Catastrophic
	FALSE	1	SCHD	PERS	121	Substantial	VMC	UNITED AIRLINES INC	In work	FALSE	Serious
	FALSE	1	SCHD	PERS	NUSC	Substantial	VMC	Jeju Air	N/A	FALSE	Catastrophic
	FALSE	1	NSCH	PERS	91	Substantial	VMC	KLM - Royal Dutch Airline	N/A	FALSE	No Injury
	FALSE	1	NSCH	PERS	91	Substantial	VMC	M U D Y PROPERTIES LLC	In work	FALSE	No Injury
	FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	Serious
	FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	Completed	FALSE	Serious
	FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	No Injury
	FALSE	2	NSCH	PERS	91	Unknown	VMC	Pilot	In work	FALSE	No Injury
	FALSE	2	NSCH	PERS	91	Substantial	IMC	Metroplex Flight Services	In work	FALSE	No Injury
	FALSE	1	NSCH	PERS	NUSN	Substantial	VMC	Pilot	N/A	FALSE	No Injury
	FALSE	2	NSCH	INST	91	Substantial	VMC	MELBOURNE FLIGHT TRAINING LLC	In work	FALSE	Serious
	FALSE	1	SCHD	PERS	121	Substantial	VMC	ALASKA AIRLINES INC	In work	FALSE	Serious
	FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	Serious
	FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	No Injury
	FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	No Injury
	TRUE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	Minor

Showing 1 to 21 of 44,507 entries, 35 total columns

Interpretation:

- 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

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

Output:

aviation_project.R* | crash_data

Filter

AmateurBuilt	NumberOfEngines	Scheduled	PurposeOfFlight	FAR	AirCraftDamage	WeatherCondition	Operator	ReportStatus	RepGenFlag	IncidentSeverity	FatalityRate
FALSE	1	NSCH	PERS	91	Substantial	IMC	Pilot	In work	FALSE	No Injury	NaN
FALSE	1	NSCH	INST	91	Substantial	VMC	Pilot	In work	FALSE	No Injury	NaN
FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	No Injury	NaN
FALSE	1	NSCH	POSI	91	Substantial	VMC	Galveston Helicopter Adventures, LLC	Completed	FALSE	No Injury	NaN
FALSE	2	NSCH	FERY	91	Substantial	VMC	Pilot	In work	FALSE	Catastrophic	100.00000
FALSE	1	SCHD	PERS	121	Substantial	VMC	UNITED AIRLINES INC	In work	FALSE	Serious	0.00000
FALSE	1	SCHD	PERS	NUSC	Substantial	VMC	Jeju Air	N/A	FALSE	Catastrophic	98.85714
FALSE	1	NSCH	PERS	91	Substantial	VMC	KLM - Royal Dutch Airline	N/A	FALSE	No Injury	NaN
FALSE	1	NSCH	PERS	91	Substantial	VMC	M U D Y PROPERTIES LLC	In work	FALSE	No Injury	NaN
FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	Serious	0.00000
FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	Completed	FALSE	Serious	0.00000
FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	No Injury	NaN
FALSE	2	NSCH	PERS	91	Unknown	VMC	Pilot	In work	FALSE	No Injury	NaN
FALSE	2	NSCH	PERS	91	Substantial	IMC	Metroplex Flight Services	In work	FALSE	No Injury	NaN
FALSE	1	NSCH	PERS	NUSN	Substantial	VMC	Pilot	N/A	FALSE	No Injury	NaN
FALSE	2	NSCH	INST	91	Substantial	VMC	MELBOURNE FLIGHT TRAINING LLC	In work	FALSE	Serious	0.00000
FALSE	1	SCHD	PERS	121	Substantial	VMC	ALASKA AIRLINES INC	In work	FALSE	Serious	0.00000
FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	Serious	0.00000
FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	No Injury	NaN
FALSE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	No Injury	NaN
TRUE	1	NSCH	PERS	91	Substantial	VMC	Pilot	In work	FALSE	Minor	0.00000

Showing 1 to 21 of 44,507 entries, 36 total columns

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?

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

Output:

```
> cat("Accidents by month:\n")
Accidents by month:
> print(accidents_by_month)
# A tibble: 43,657 × 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
 7 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
```

Interpretation:

- 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?

```
  # Step 1: Keep only rows with valid TRUE/FALSE values
crash_data$AmateurBuilt <- ifelse(crash_data$AmateurBuilt == "TRUE", TRUE,
                                  ifelse(crash_data$AmateurBuilt == "FALSE", FALSE, NA))

aviation_clean <- crash_data %>%
  filter(!is.na(AmateurBuilt))
t.test(FatalInjuryCount ~ AmateurBuilt, data = crash_data)
```

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?

```r
# 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?

```r
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 × 2
   AirportID Total_Fatalities
   <fct>                <dbl>
 1 OPRN                   157
 2 FMCH                   152
 3 URSS                   113
 4 MUHA                   112
 5 OLBA                    90
 6 XUBS                    89
 7 CGK                     62
 8 RCQC                    58
 9 WIHH                    44
10 ZYLD                    42
# 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.