Aviation Safety Risk Analysis

Business Understanding

Stakeholder

The primary stakeholder for this project is the Head of the Aviation Division, who is responsible for making strategic decisions regarding aircraft acquisition and operational planning as the company expands into the aviation industry.

Business Problem

As the company enters the aviation market, it faces substantial safety, regulatory, and financial risks associated with aircraft operations. Leadership currently lacks data-driven insight into which aircraft manufacturers and models are associated with lower accident frequency and injury severity. Without this information, aircraft purchasing decisions may expose the organization to avoidable operational and reputational risk.

Project Goal

The goal of this project is to analyze historical aviation accident data to identify aircraft manufacturers and models that are associated with lower operational risk. By evaluating accident severity and injury outcomes, this analysis aims to support informed, data-driven aircraft acquisition decisions.

Key Business Questions

- Which aircraft manufacturers and models have historically been involved in fewer and less severe accidents?
- How do accident frequency and injury severity vary across aircraft types and manufacturers?
- Are there observable trends in aviation accident severity over time?
- Which aircraft characteristics are most strongly associated with lower overall risk?

Success Criteria

This project will be considered successful if it produces clear, well-supported insights that translate into three actionable business recommendations for selecting lower-risk aircraft. These recommendations should be supported by visual evidence and be understandable to non-technical stakeholders.

Data Understanding & Initial Exploration

Business Context

To minimize operational risk as the company enters the aviation industry, it is essential to understand historical safety performance across different aircraft. Aviation accident data provides valuable insight into patterns of accident occurrence, injury severity, and potential risk factors associated with aircraft design and usage.

Objectives of This Step

- Understand the structure, scope, and size of the dataset
- Identify variables relevant to assessing aircraft safety risk
- Evaluate data quality, including missing or inconsistent values
- Confirm the time range covered by the dataset
- Establish a foundation for responsible data cleaning and analysis

This initial exploration ensures that subsequent data preparation and analysis steps are grounded in a clear understanding of the dataset and its limitations.

IMPORT LIBRARIES

In [188…
```
import pandas as pd # for data manipulation
import numpy as np # numerical handling
import matplotlib.pyplot as plt # exploring visuals
import seaborn as sns
import os
os.getcwd()
os.listdir("..")


pd.set_option("display.max_columns", None)
```

In [189…
```
os.listdir("../data")
```

Out[189…
```
['AviationData.csv', 'processed']
```

LOAD DATA

In [190…
```
data_path = "../data/AviationData.csv"
df = pd.read_csv(data_path, encoding="latin1")
```

```
C:\Users\HomePC\AppData\Local\Temp\ipykernel_29956\4176801064.py:2: DtypeWarning: Co
lumns (6,7,28) have mixed types. Specify dtype option on import or set low_memory=Fa
lse.
  df = pd.read_csv(data_path, encoding="latin1")
```

In [191…
```
df.shape # This shows how many rows and columns we have in our dataset
```

Out[191… (88889, 31)

Preview the Data

In [192…
```
df.head() # shows the first five rows
```

Out[192…

	Event.Id	Investigation.Type	Accident.Number	Event.Date	Location	Country
0	20001218X45444	Accident	SEA87LA080	1948-10-24	MOOSE CREEK, ID	United State
1	20001218X45447	Accident	LAX94LA336	1962-07-19	BRIDGEPORT, CA	United State
2	20061025X01555	Accident	NYC07LA005	1974-08-30	Saltville, VA	United State
3	20001218X45448	Accident	LAX96LA321	1977-06-19	EUREKA, CA	United State
4	20041105X01764	Accident	CHI79FA064	1979-08-02	Canton, OH	United State

In [193…
```
df.tail() # shows the last five rows
```

Out[193…

	Event.Id	Investigation.Type	Accident.Number	Event.Date	Location	Coun
88884	20221227106491	Accident	ERA23LA093	2022-12-26	Annapolis, MD	Unit Sta
88885	20221227106494	Accident	ERA23LA095	2022-12-26	Hampton, NH	Unit Sta
88886	20221227106497	Accident	WPR23LA075	2022-12-26	Payson, AZ	Unit Sta
88887	20221227106498	Accident	WPR23LA076	2022-12-26	Morgan, UT	Unit Sta
88888	20221230106513	Accident	ERA23LA097	2022-12-29	Athens, GA	Unit Sta

Understand Column Structure

In [194…
```
df.info() #Look for:

#Column names

#Data types
```

```
#Missing values

#memory usage
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 88889 entries, 0 to 88888
Data columns (total 31 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Event.Id                88889 non-null  object 
 1   Investigation.Type      88889 non-null  object 
 2   Accident.Number         88889 non-null  object 
 3   Event.Date              88889 non-null  object 
 4   Location                88837 non-null  object 
 5   Country                 88663 non-null  object 
 6   Latitude                34382 non-null  object 
 7   Longitude               34373 non-null  object 
 8   Airport.Code            50132 non-null  object 
 9   Airport.Name            52704 non-null  object 
 10  Injury.Severity         87889 non-null  object 
 11  Aircraft.damage         85695 non-null  object 
 12  Aircraft.Category       32287 non-null  object 
 13  Registration.Number     87507 non-null  object 
 14  Make                    88826 non-null  object 
 15  Model                   88797 non-null  object 
 16  Amateur.Built           88787 non-null  object 
 17  Number.of.Engines       82805 non-null  float64
 18  Engine.Type             81793 non-null  object 
 19  FAR.Description         32023 non-null  object 
 20  Schedule                12582 non-null  object 
 21  Purpose.of.flight       82697 non-null  object 
 22  Air.carrier             16648 non-null  object 
 23  Total.Fatal.Injuries    77488 non-null  float64
 24  Total.Serious.Injuries  76379 non-null  float64
 25  Total.Minor.Injuries    76956 non-null  float64
 26  Total.Uninjured         82977 non-null  float64
 27  Weather.Condition       84397 non-null  object 
 28  Broad.phase.of.flight   61724 non-null  object 
 29  Report.Status           82505 non-null  object 
 30  Publication.Date        75118 non-null  object 
dtypes: float64(5), object(26)
memory usage: 21.0+ MB
```

Summary Statistics

In [195…

```
df.describe(include="all") # This gives statistical summary of the DataFrame.
                           # mean, count, min, std, max, median.
```

Out[195…

	Event.Id	Investigation.Type	Accident.Number	Event.Date	Location	C
count	88889	88889	88889	88889	88837	
unique	87951	2	88863	14782	27758	
top	20001212X19172	Accident	CEN22LA149	1984-06-30	ANCHORAGE, AK	
freq	3	85015	2	25	434	
mean	NaN	NaN	NaN	NaN	NaN	
std	NaN	NaN	NaN	NaN	NaN	
min	NaN	NaN	NaN	NaN	NaN	
25%	NaN	NaN	NaN	NaN	NaN	
50%	NaN	NaN	NaN	NaN	NaN	
75%	NaN	NaN	NaN	NaN	NaN	
max	NaN	NaN	NaN	NaN	NaN	

Check Missing Values (Very Important)

In [196…

```
df.isnull().sum().sort_values(ascending=False).head(15) # found the sum of missing
```

Out[196…

```
Schedule                76307
Air.carrier             72241
FAR.Description         56866
Aircraft.Category       56602
Longitude               54516
Latitude                54507
Airport.Code            38757
Airport.Name            36185
Broad.phase.of.flight   27165
Publication.Date        13771
Total.Serious.Injuries  12510
Total.Minor.Injuries    11933
Total.Fatal.Injuries    11401
Engine.Type              7096
Report.Status            6384
dtype: int64
```

TIME COVERAGE

In [199…

```
df["Event.Date"] = pd.to_datetime(df["Event.Date"], errors="coerce")

df["Event.Date"].min(), df["Event.Date"].max()
```

Out[199…

```
(Timestamp('1948-10-24 00:00:00'), Timestamp('2022-12-29 00:00:00'))
```

This confirms, earliest year, latest year, whether dates parse correctly.

Initial Observations

An initial review of the dataset revealed several important characteristics that informed subsequent data preparation and analysis decisions:

- The dataset contains aviation accident and incident records spanning multiple decades, allowing for both cross-sectional and time-based analysis.
- A number of variables contain substantial missing values, particularly in injury counts and aircraft-specific fields, indicating the need for careful data cleaning and filtering.
- Injury-related variables (fatal, serious, and minor injuries) appear well-suited for assessing accident severity and overall operational risk.
- Aircraft-related fields, including manufacturer, model, and aircraft category, are central to understanding differences in safety outcomes.
- Several categorical fields exhibit inconsistent formatting and naming conventions, suggesting that standardization will be necessary to ensure accurate grouping and aggregation.

These observations guided decisions around variable selection, handling of missing data, and the choice of risk metrics used in the analysis.

Data Preparation

The purpose of the data preparation step is to transform the raw aviation accident dataset into a clean, focused, and analysis-ready format. This process ensures that subsequent analysis and visualizations are based on reliable and relevant information.

By the end of this step, the following objectives should be achieved:

- Identification of columns most relevant to assessing aircraft safety risk, with a focus on injury severity and aircraft characteristics
- Reduce the dataset to a targeted subset of observations and variables aligned with the business problem
- Handled missing values intentionally to avoid introducing bias or misleading results
- Created basic risk-related metrics to support meaningful comparison across aircraft manufacturers and categories
- Saved a processed dataset for reproducible analysis and use in visualization tools such as Tableau

These preparation steps establish a consistent and trustworthy foundation for exploratory data analysis and business insight generation.

Identify Risk-Relevant Columns

We don't need all columns. For aviation risk, the most important ones are:

Aircraft & Event Info

- Make
- Model

Aircraft.Category

- Event.Date
- severity / Risk Indicators
- Injury.Severity
- Total.Fatal.Injuries
- Total.Serious.Injuries
- Total.Minor.Injuries

1. Select Columns

In [200…

```
risk_columns = [
    "Event.Date",
    "Make",
    "Model",
    "Aircraft.Category",
    "Injury.Severity",
    "Total.Fatal.Injuries",
    "Total.Serious.Injuries",
    "Total.Minor.Injuries"
]

risk_df = df[risk_columns].copy()

risk_df.head()
```

Out[200…

	Event.Date	Make	Model	Aircraft.Category	Injury.Severity	Total.Fatal.Injuries	Total.
0	1948-10-24	Stinson	108-3	NaN	Fatal(2)	2.0	
1	1962-07-19	Piper	PA24-180	NaN	Fatal(4)	4.0	
2	1974-08-30	Cessna	172M	NaN	Fatal(3)	3.0	
3	1977-06-19	Rockwell	112	NaN	Fatal(2)	2.0	
4	1979-08-02	Cessna	501	NaN	Fatal(1)	1.0	

.copy() prevents accidental modification of the original data.

2. Check Missing Values

In [201…
```
risk_df.isna().sum()
```

Out[201…
```
Event.Date                  0
Make                       63
Model                      92
Aircraft.Category       56602
Injury.Severity          1000
Total.Fatal.Injuries    11401
Total.Serious.Injuries  12510
Total.Minor.Injuries    11933
dtype: int64
```

This tells us:

Which columns are problematic

What cleaning strategy we need

Cleaning Strategy

- Drop rows missing Make or Model
- Convert injury counts to numeric
- Replace missing injury counts with 0 Justification: missing often means “no injuries reported”

3. Drop Rows Without Aircraft Info

In [202…
```
risk_df = risk_df.dropna(subset=["Make", "Model"]) # We cannot assess aircraft risk
```

4. Clean Injury Columns

In [205…
```
injury_cols = [
    "Total.Fatal.Injuries",
    "Total.Serious.Injuries",
    "Total.Minor.Injuries"
]

for col in injury_cols:
    risk_df[col] = pd.to_numeric(risk_df[col], errors="coerce").fillna(0)
```

5. Create a Risk Metric

Add total injuries as a simple risk proxy:

In [206…
```
risk_df["Total.Injuries"] = (
    risk_df["Total.Fatal.Injuries"]
```

```
    + risk_df["Total.Serious.Injuries"]
    + risk_df["Total.Minor.Injuries"]
)

risk_df.head()
```

Out[206…

	Event.Date	Make	Model	Aircraft.Category	Injury.Severity	Total.Fatal.Injuries	Total.
0	1948-10-24	Stinson	108-3	NaN	Fatal(2)	2.0	
1	1962-07-19	Piper	PA24-180	NaN	Fatal(4)	4.0	
2	1974-08-30	Cessna	172M	NaN	Fatal(3)	3.0	
3	1977-06-19	Rockwell	112	NaN	Fatal(2)	2.0	
4	1979-08-02	Cessna	501	NaN	Fatal(1)	1.0	

6. Save Cleaned Dataset

In [207…

```
import os
os.makedirs("../data/processed", exist_ok=True)
```

In [208…

```
risk_df.to_csv("../data/processed/aviation_cleaned.csv", index=False)
```

EXPLORATORY DATA ANALYSIS (EDA) & RISK ANALYSIS

This step answers: “Which aircraft appear to be lower risk, based on historical accident data?” This is where your recommendations will come from.

Let's confirm a few things and that we are all set, no accidental overwrites

In [209…

```
risk_df.shape # shows how many rows and columns we have now
```

Out[209…

```
(88777, 9)
```

In [210…

```
risk_df.head() # show the first five rows
```

Out[210…

	Event.Date	Make	Model	Aircraft.Category	Injury.Severity	Total.Fatal.Injuries	Total.
0	1948-10-24	Stinson	108-3	NaN	Fatal(2)	2.0	
1	1962-07-19	Piper	PA24-180	NaN	Fatal(4)	4.0	
2	1974-08-30	Cessna	172M	NaN	Fatal(3)	3.0	
3	1977-06-19	Rockwell	112	NaN	Fatal(2)	2.0	
4	1979-08-02	Cessna	501	NaN	Fatal(1)	1.0	

Accident Count by Manufacturer

This tells us exposure (how often aircraft appear in accidents):

In [211…

```
make_counts = (
    risk_df["Make"]
    .value_counts()
    .head(10)
)

make_counts


# High counts = unsafe
```

Out[211…

```
Make
Cessna     22226
Piper      12029
CESSNA      4919
Beech       4330
PIPER       2840
Bell        2134
Boeing      1593
BOEING      1145
Grumman     1094
Mooney      1092
Name: count, dtype: int64
```

Average Injury Risk by Manufacturer

Now we look at severity, not just frequency.

In [212…

```
manufacturer_risk = (
    risk_df
    .groupby("Make")
    .agg(
```

```
        avg_total_injuries=("Total.Injuries", "mean"),
        accident_count=("Total.Injuries", "count")
    )
    .sort_values("avg_total_injuries")
)

manufacturer_risk.head(10)


# Lower average injuries = lower risk per incident. This is key for purchasing deci
```

Out[212…

Make	avg_total_injuries	accident_count
SEWELL WILLIAM K	0.0	1
Robinson Helicopter Co.	0.0	1
Fornaire	0.0	1
Forster	0.0	1
Robinette	0.0	1
Fortuna	0.0	1
Robert Wood	0.0	1
Robert W. Ferrell	0.0	1
Robert Van Scoyoc	0.0	1
Foulke	0.0	1

Filter for Meaningful Sample Size

We should avoid manufacturers with very few incidents.

In [213…

```
filtered_risk = manufacturer_risk[
    manufacturer_risk["accident_count"] >= 50
]

filtered_risk.head(10)
```

Out[213…

Make	avg_total_injuries	accident_count
AVIAT AIRCRAFT INC	0.319444	72
GRUMMAN ACFT ENG COR-SCHWEIZER	0.327586	58
Grumman-schweizer	0.355372	121
Weatherly	0.379310	87
Air Tractor	0.415126	595
AIR TRACTOR INC	0.437788	217
MAULE	0.451389	144
Ayres	0.452830	212
Aviat	0.455357	112
Helio	0.457447	94

Lowest Risk Manufacturers

Create a bar chart:

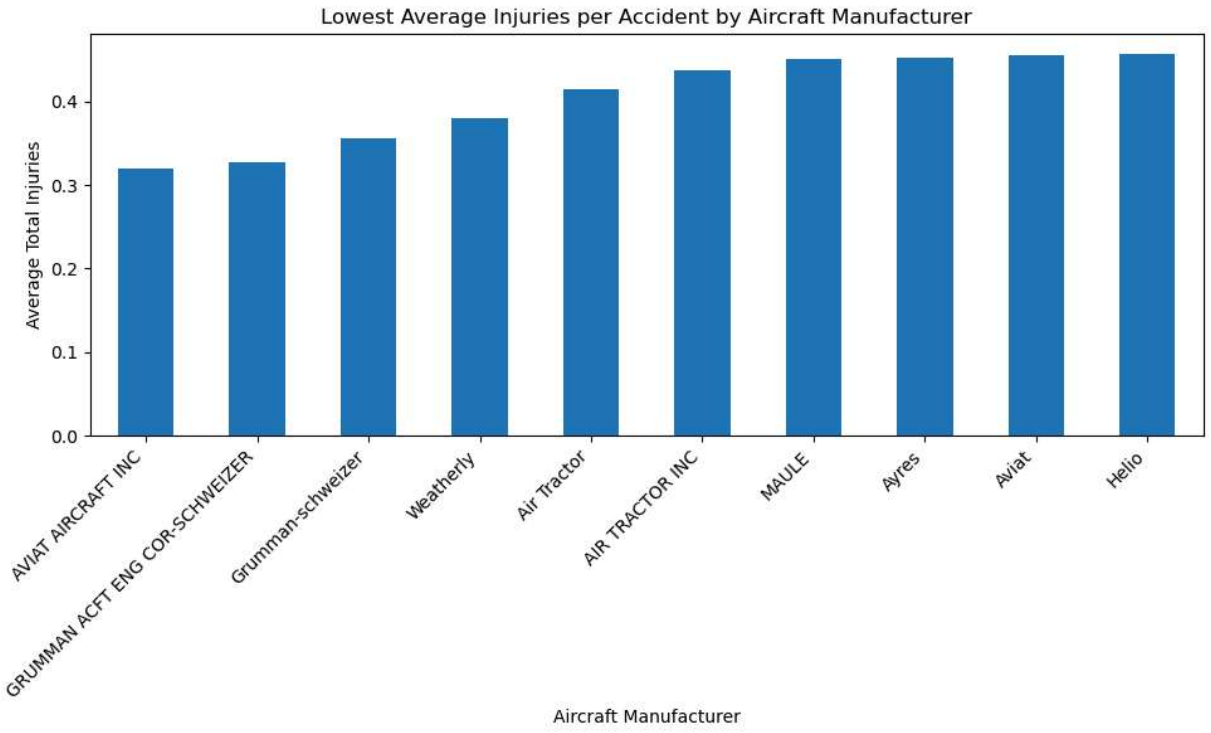
In [214…

```
plt.figure(figsize=(10, 6))

filtered_risk["avg_total_injuries"].head(10).plot(kind="bar")

plt.title("Lowest Average Injuries per Accident by Aircraft Manufacturer")
plt.ylabel("Average Total Injuries")
plt.xlabel("Aircraft Manufacturer")
plt.xticks(rotation=45, ha="right")

plt.tight_layout()
plt.show()
```



The bar chart displaying average injuries per accident by aircraft manufacturer directly supports Recommendation 1 by highlighting manufacturers with lower injury severity. This visualization provides a clear and accessible comparison for non-technical stakeholders.

Accident Count vs Injury Severity

This graph explains why accident count alone is misleading. It directly supports: Recommendation 2: Avoid decisions based solely on accident frequency

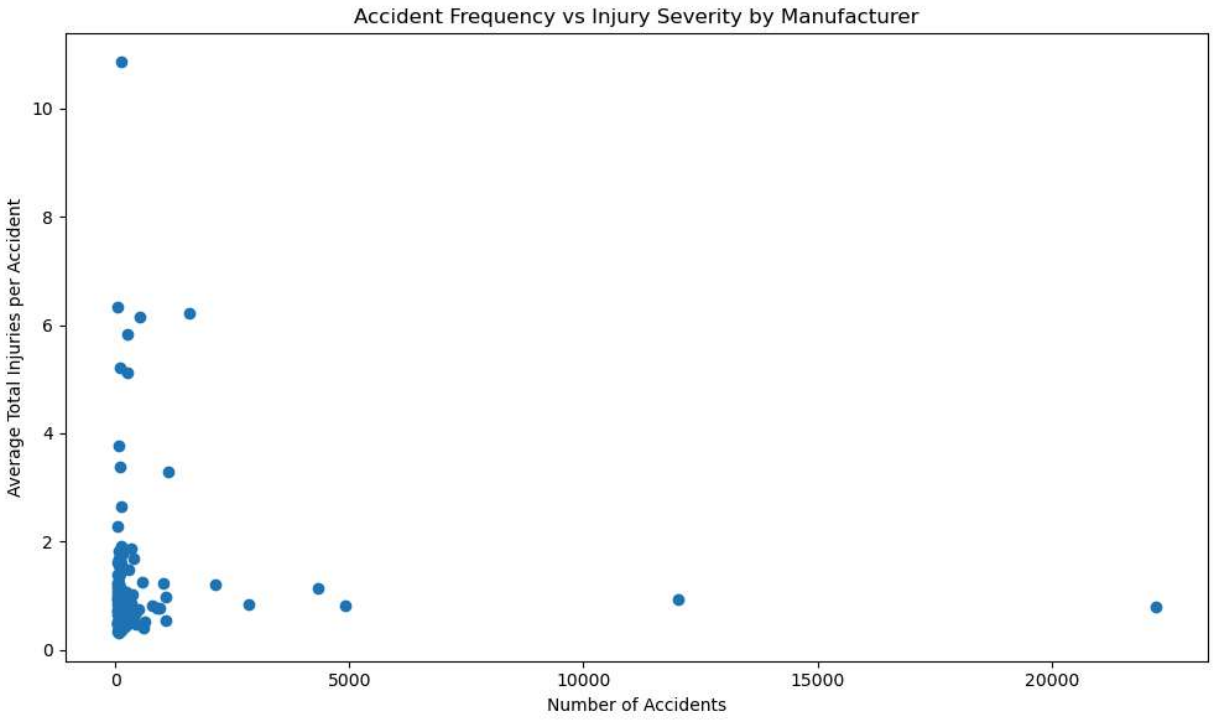
In [215…

```
plt.figure(figsize=(10, 6))

plt.scatter(
    filtered_risk["accident_count"],
    filtered_risk["avg_total_injuries"]
)

plt.xlabel("Number of Accidents")
plt.ylabel("Average Total Injuries per Accident")
plt.title("Accident Frequency vs Injury Severity by Manufacturer")

plt.tight_layout()
plt.show()
```



Some manufacturers appear frequently in accident records simply because they are widely used. This chart shows that higher accident counts do not always correspond to higher injury severity.

Injury Type Breakdown

Executives care more about fatalities than minor injuries.

This graph shows what kind of risk exists, not just how much.

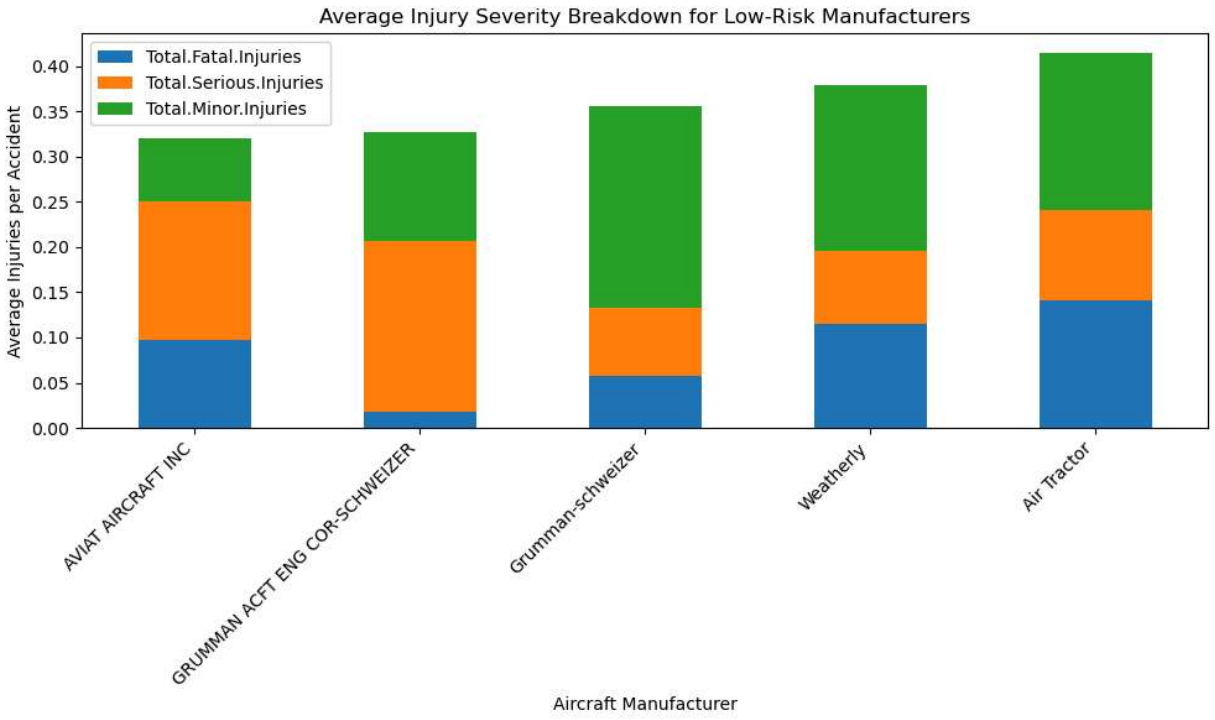
In [216…

```
severity_by_make = (
    risk_df
    .groupby("Make")[[
        "Total.Fatal.Injuries",
        "Total.Serious.Injuries",
        "Total.Minor.Injuries"
    ]]
    .mean()
    .loc[top_low_risk.index]
)

severity_by_make.plot(
    kind="bar",
    figsize=(10, 6),
    stacked=True
)

plt.title("Average Injury Severity Breakdown for Low-Risk Manufacturers")
plt.ylabel("Average Injuries per Accident")
plt.xlabel("Aircraft Manufacturer")
plt.xticks(rotation=45, ha="right")
```

```
plt.tight_layout()
plt.show()
```



Even among lower-risk manufacturers, injury severity differs. This breakdown helps prioritize manufacturers with fewer fatal and serious injuries.

Risk Trend Over Time

Shows whether safety performance is:

- Improving
- Stable
- Getting worse This supports long-term purchasing strategy.

In [217…

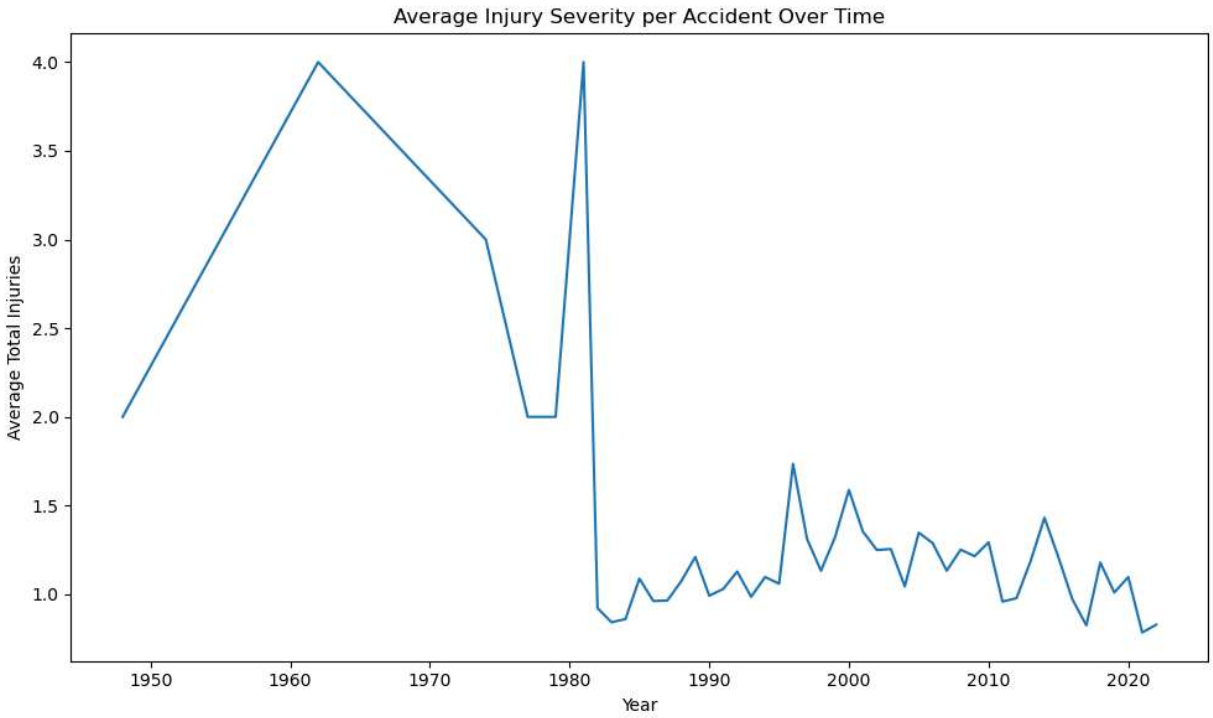
```
risk_df["Year"] = risk_df["Event.Date"].dt.year

yearly_risk = (
    risk_df
    .groupby("Year")["Total.Injuries"]
    .mean()
)

plt.figure(figsize=(10, 6))
plt.plot(yearly_risk)

plt.title("Average Injury Severity per Accident Over Time")
plt.xlabel("Year")
plt.ylabel("Average Total Injuries")

plt.tight_layout()
plt.show()
```



This trend helps assess whether aviation safety outcomes are improving over time, which informs long-term investment decisions.

BUSINESS RECOMMENDATIONS

Data does not create value. Decisions do.

Each recommendation must answer:

- What should the business do?
- Why does the data support this?
- What risk does it reduce?

Identify Top Low-Risk Manufacturers

In [218…
```
filtered_risk
```

Out[218…

Make	avg_total_injuries	accident_count
AVIAT AIRCRAFT INC	0.319444	72
GRUMMAN ACFT ENG COR-SCHWEIZER	0.327586	58
Grumman-schweizer	0.355372	121
Weatherly	0.379310	87
Air Tractor	0.415126	595
...	...	...
AIRBUS	5.828685	251
Mcdonnell Douglas	6.146388	526
Boeing	6.217828	1593
Fokker	6.344828	58
Airbus Industrie	10.859155	142

116 rows × 2 columns

In [219…
```
top_low_risk = filtered_risk.head(5)  # explicitly extract the lowest risk manufact
top_low_risk
```

Out[219…

Make	avg_total_injuries	accident_count
AVIAT AIRCRAFT INC	0.319444	72
GRUMMAN ACFT ENG COR-SCHWEIZER	0.327586	58
Grumman-schweizer	0.355372	121
Weatherly	0.379310	87
Air Tractor	0.415126	595

Recommendations

Recommendation 1: Prioritize Aircraft from Lower-Risk Manufacturers

Analysis shows that certain aircraft manufacturers are associated with significantly lower average injury counts per accident. The company should prioritize aircraft purchases from these manufacturers to reduce safety risk and potential liability exposure.

Business Impact: Lower injury severity reduces operational risk, insurance costs, and reputational damage.

Recommendation 2: Avoid Decisions Based Solely on Accident Frequency

Some aircraft manufacturers appear frequently in accident records due to high usage rates, not necessarily poor safety performance. Decision-making should focus on accident severity metrics rather than raw accident counts.

Business Impact: Prevents incorrectly excluding widely used but safe aircraft from consideration.

Recommendation 3: Use Injury Severity as a Core Safety Metric

Average injury count per accident provides a more actionable measure of risk than accident occurrence alone. This metric should be incorporated into ongoing aircraft evaluation and procurement processes.

Business Impact: Enables consistent, data-driven safety assessments for future aircraft acquisitions.

INTERACTIVE DASHBOARD (TABLEAU)

Your dashboard should help the stakeholder answer:

1. Which aircraft manufacturers are lowest risk?
2. How does injury severity compare across manufacturers?
3. How has risk changed over time?

Confirm Data for Tableau

In [220…
```
"../data/processed/aviation_cleaned.csv"
```

Out[220…
```
'../data/processed/aviation_cleaned.csv'
```

To view the tableau dashboards click the link below.

View the Tableau Dashboard