RISK ASSESSMENT AND DATA-DRIVEN RECOMMENDATIONS FOR AIRCRAFT ACQUISITION

BY GODWIN MUTUMA MURUNGI

Overview

This research analyses data on aircraft damage to find patterns and trends in aviation-related accidents. The analysis's goals are to deliver important business answers and useful information to aviation sector stakeholders.

Business Problem

My company wants to diversify its portfolio by expanding into new industries, specifically aviation. They are considering purchasing and operating airplanes for commercial and private enterprises but lack knowledge about the risks. My task is to analyze aviation accident data and determine which aircraft present the lowest risk. These findings will help the head of the new aviation division make informed decisions on which aircraft to purchase.

Data Understanding

This data is from the NTSB aviation accident database and contains information from 1962 and later about civil aviation accidents and selected incidents within the United States, its territories and possessions, and in international waters.

Project Objectives

1. Data Management
Establish a framework to handle missing values and ensure data integrity, enabling accurate analysis of customer behavior and operational performance in aviation.

2. Data Aggregation and Visualization
Aggregate data from multiple sources and use simple visualizations (e.g., bar charts, line graphs) to present key trends, making insights accessible to stakeholders.

3. Give Recommendations
Provide three actionable recommendations to capitalize on the new aviation opportunity, supported by visual data representations to illustrate expected outcomes.

Handling Missing Values

In my dataset, I've calculated the percentage of missing values for each column. The following columns exhibit significant missing data:

Latitude: 61.32%
Longitude: 61.33%
Aircraft.Category: 63.68%
FAR.Description: 63.97%
Schedule: 85.85%
Air.carrier: 81.27%

Reason for Row Dropping

I aim to maintain data integrity and ensure the quality of my analysis. Columns with over 60% missing values present a challenge because:

1. Insufficient Data: High percentages of missing data can lead to unreliable conclusions, as the remaining data may not be representative of the whole dataset.
2. Data Quality: Keeping rows with excessive missing data can compromise the overall quality of my dataset. Removing these rows helps maintain a cleaner and more reliable dataset for further analysis.

Therefore, I will drop rows where any column has missing values exceeding 60% to enhance the quality of my dataset and the accuracy of my analysis.

Further cleaning

I decided to drop the following columns due to their significant percentages of missing data:
Airport.Code: 43.60% missing values
Airport.Name: 40.71% missing values

Reasons for Dropping These Columns

1. **Central Tendencies**: Measures of central tendency cannot appropriately replace the missing values in these columns. Airport codes and Airport names are categorical variables, and using central tendencies would not provide meaningful replacements. Therefore, adding these columns to the dataset would not be helpful.
2. **Data Quality**:Maintaining columns with a significant number of missing values may have a negative impact on my dataset's overall quality. To ensure proper analysis, it is necessary to work with data that is complete

Handling missing values for the Categorical Data

In this analysis, I opted to replace missing values in categorical data with the mode. This is because:

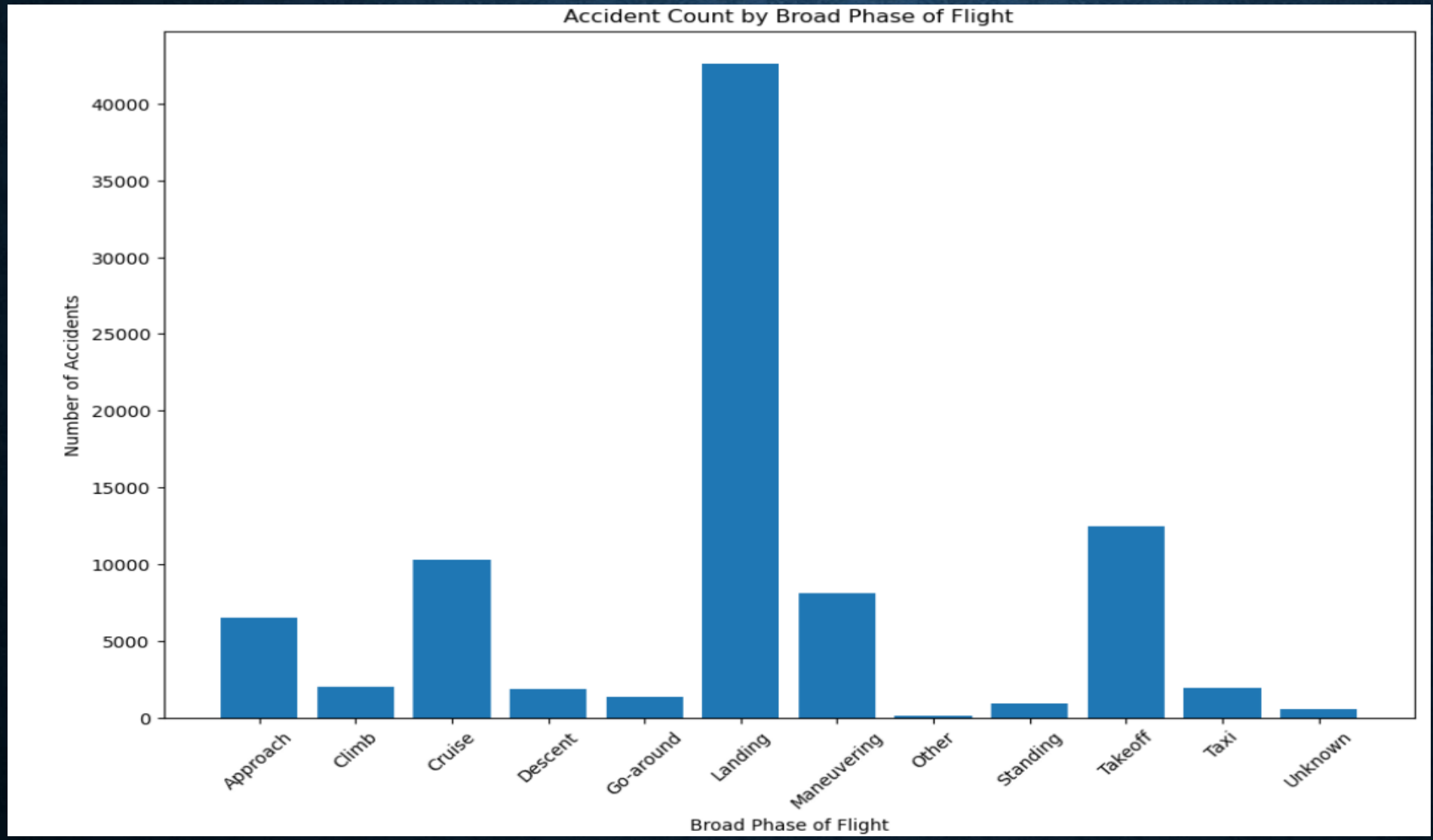
1. **Efficiency**: Mode imputation is a simple and effective computing technique, particularly when dealing with large datasets.
2. **Preserving Data Distribution**: Using the mode, the original distribution of the categorical variable is preserved, ensuring the accuracy of the data.
3. **Minimal Data Loss**: By filling in missing values with the most common category, we minimize the loss of information.
4. **Common Practice**: This method is a widely accepted and effective approach in data analysis and machine learning.

Handling Missing Data for the Numerical Data

For all the numerical data with missing values, I opted to use the mean imputation method to handle the missing data.
The reason behind this is because this method provides an important principle that accurately summarizes the data without introducing bias.

Visualizing the Data

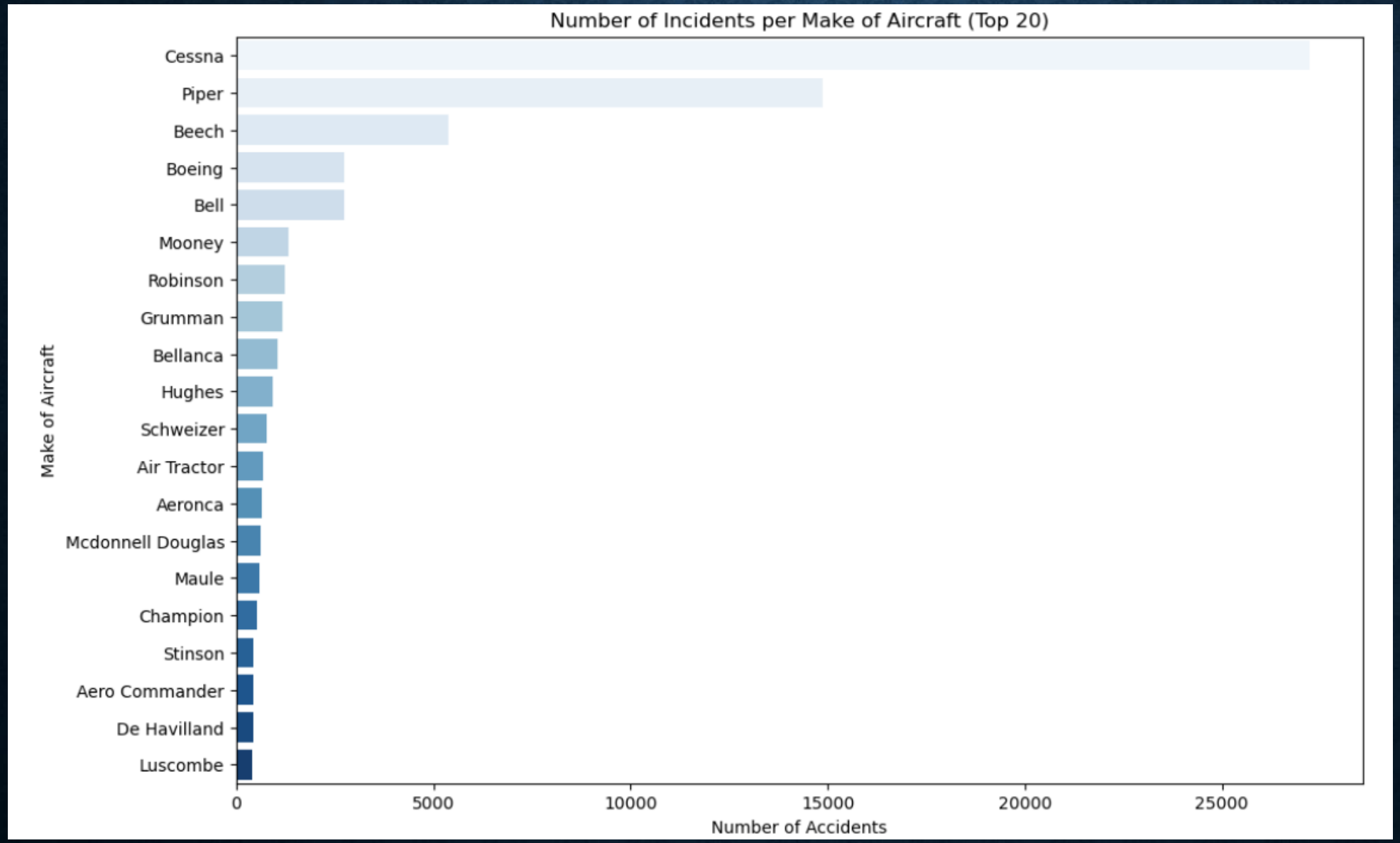
A Plot of Accident Count by Broad Phase of Flight



Observation

From the above we can conclude that most of the accidents occur when landing

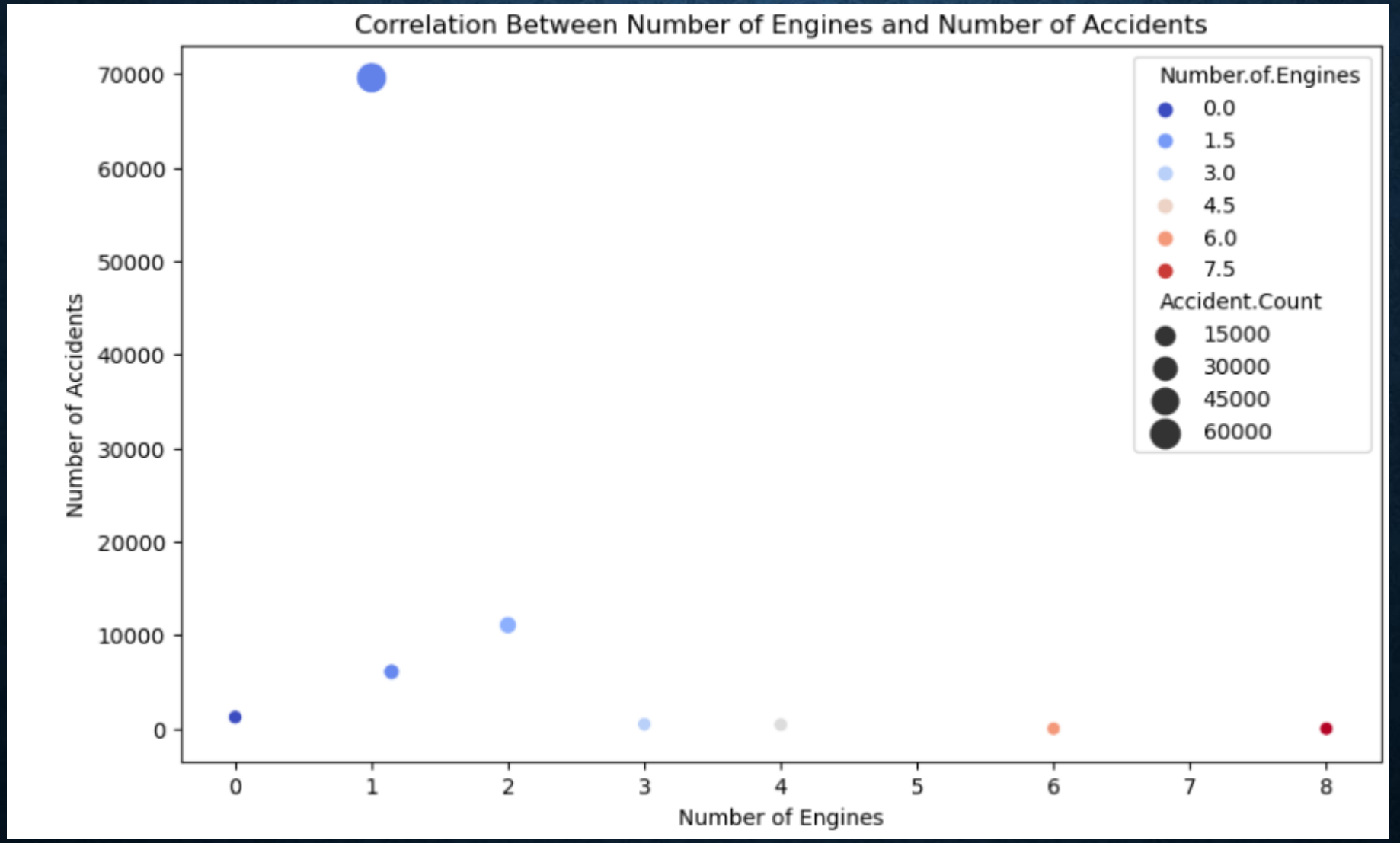
Number of Incidents per Make of Aircraft for the Top 20 aircraft Makes



Observation

- **Cessna** leads with **27,212** incidents, reflecting its popularity in general aviation.
- **Piper** follows with **14,870** incidents, also common among pilots.
- **Beech** has **5,372** incidents, while **Boeing** reports **2,745**, indicating fewer incidents for larger commercial aircraft.

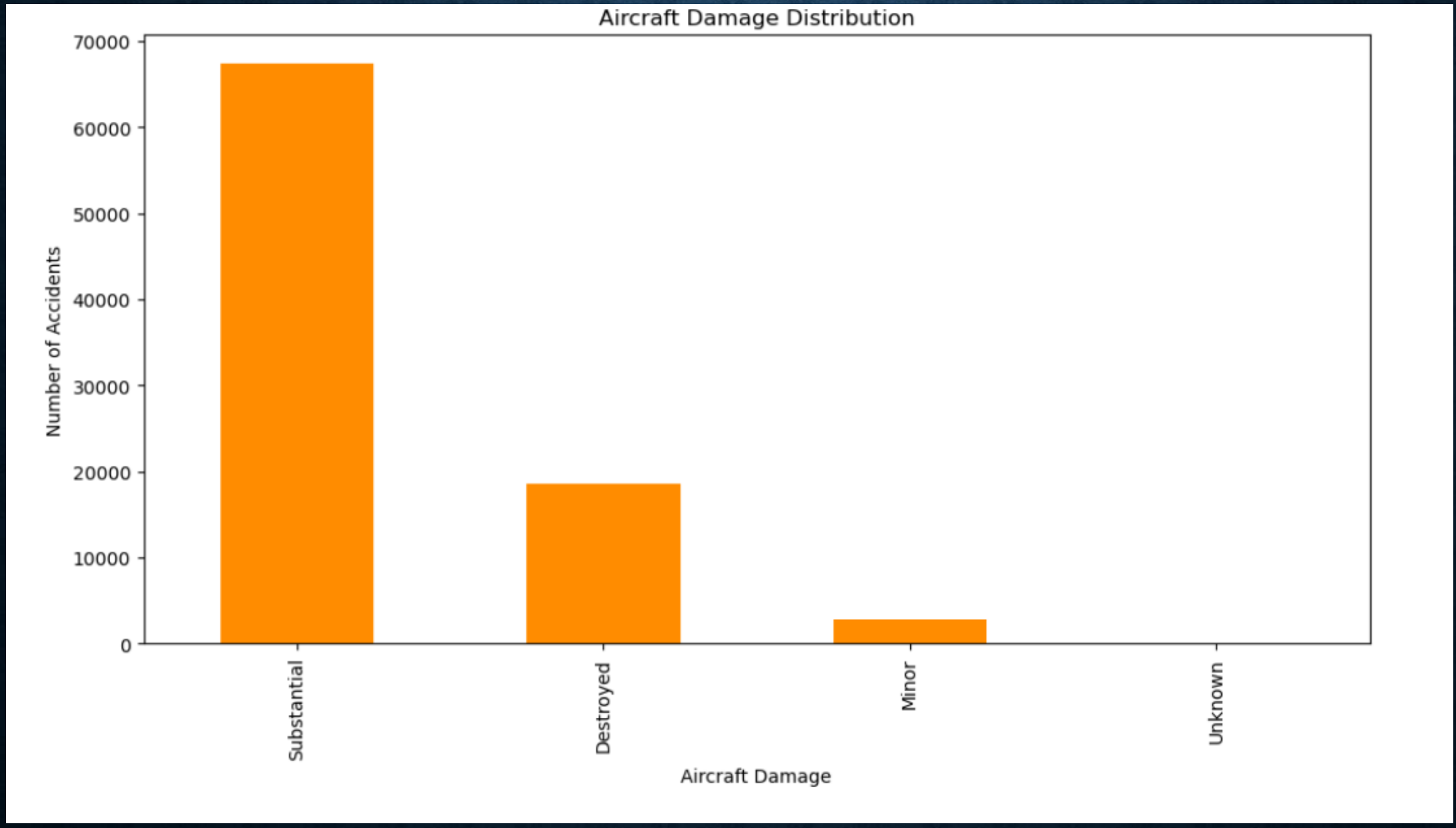
Scatter Plot to view the Correlation Between Number of Engines and Number of Accidents



Observation

Number of accidents decrease as the number of engines increases, which could be due to the fact that larger aircraft with more engines have more safety measures and are used in more controlled conditions.

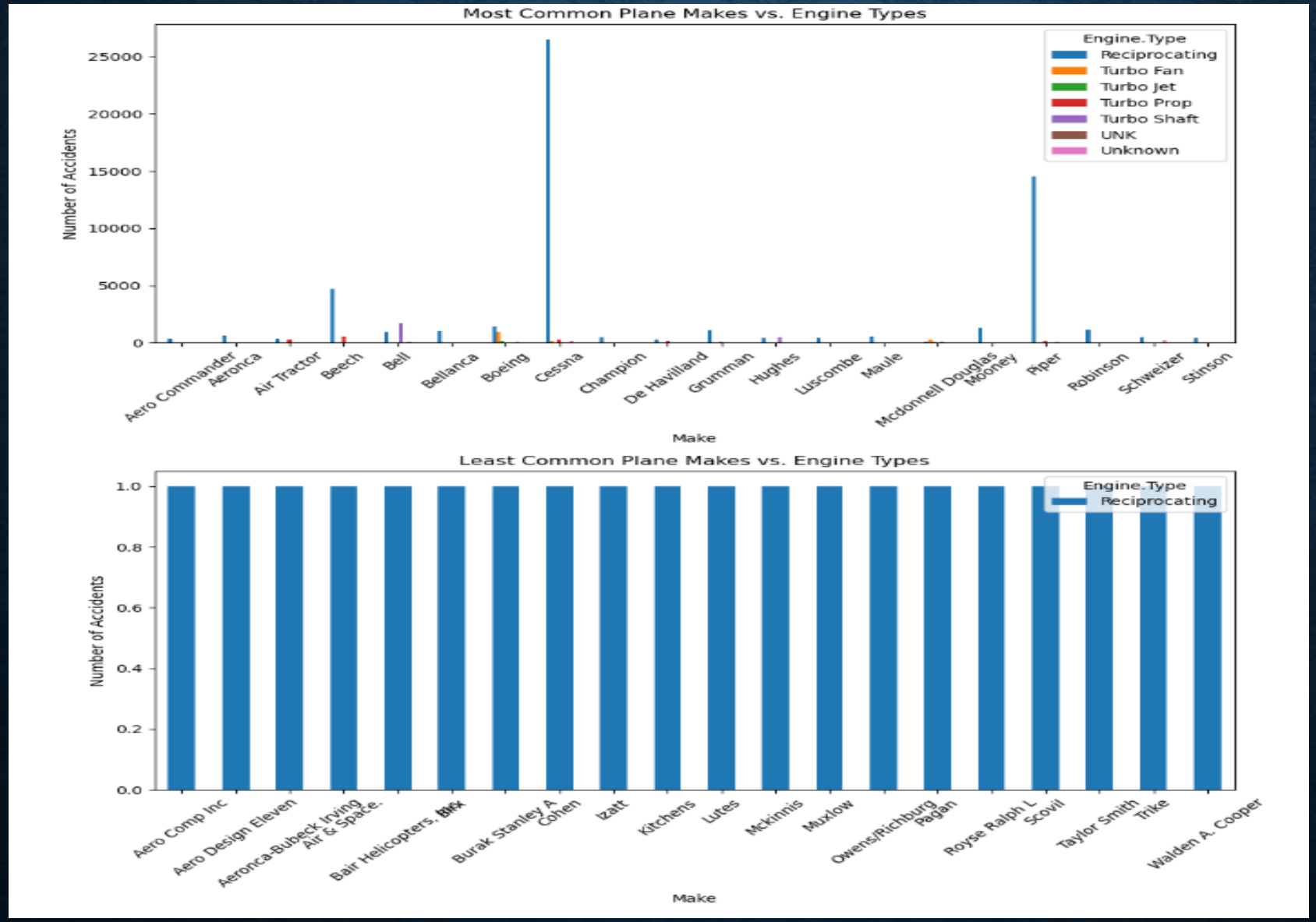
Aircraft Damage Distribution



Observation

From the chart, it can be observed that the majority of accidents result in substantial damage to the aircraft. A smaller number of accidents lead to destroyed or minor damage. The category "Unknown" has a relatively low frequency, indicating that the extent of damage is often determined and reported.

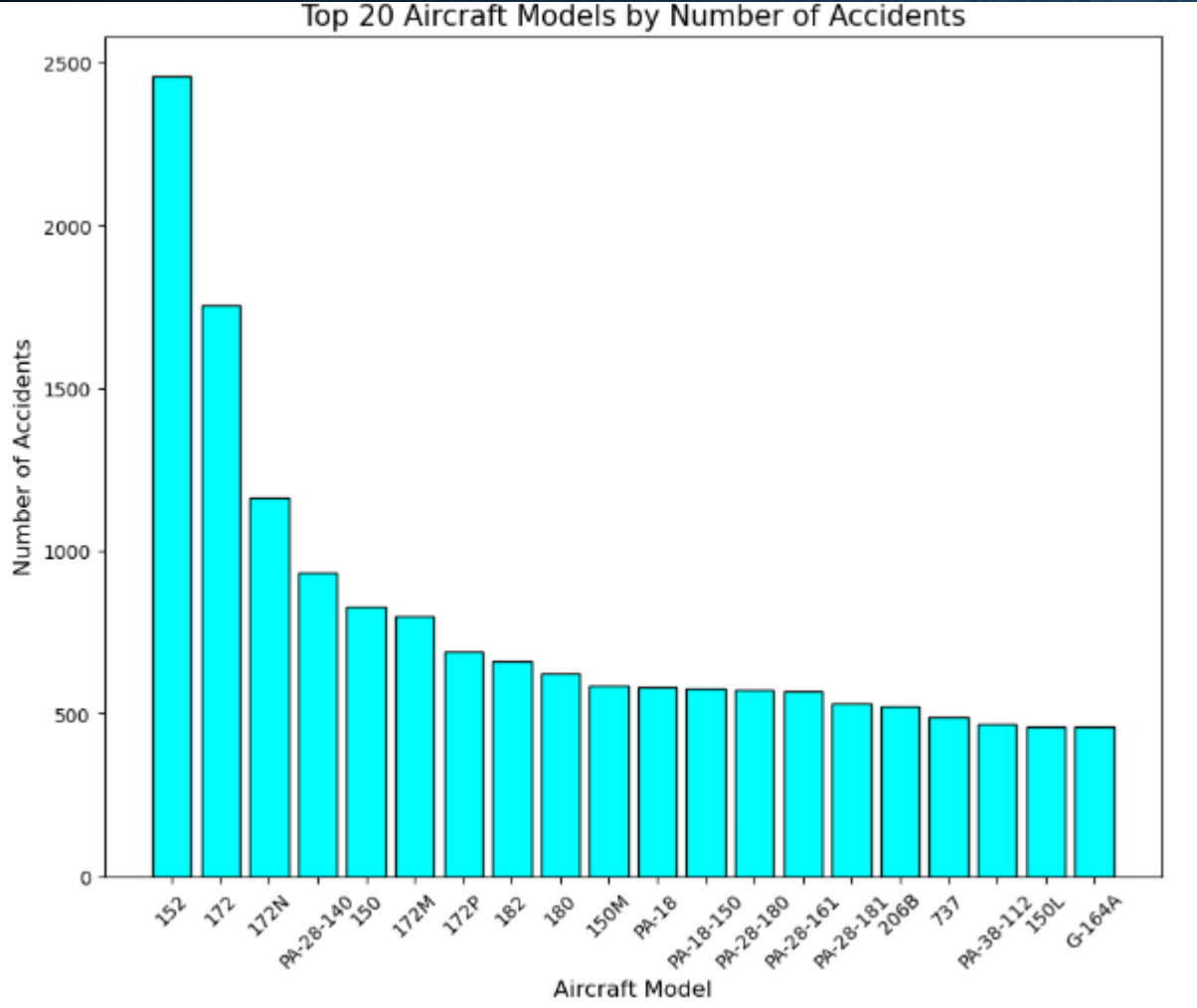
Visualizing the Top 20 and Bottom 20 Planes Based on Their Make Frequency and Their Injury Severity

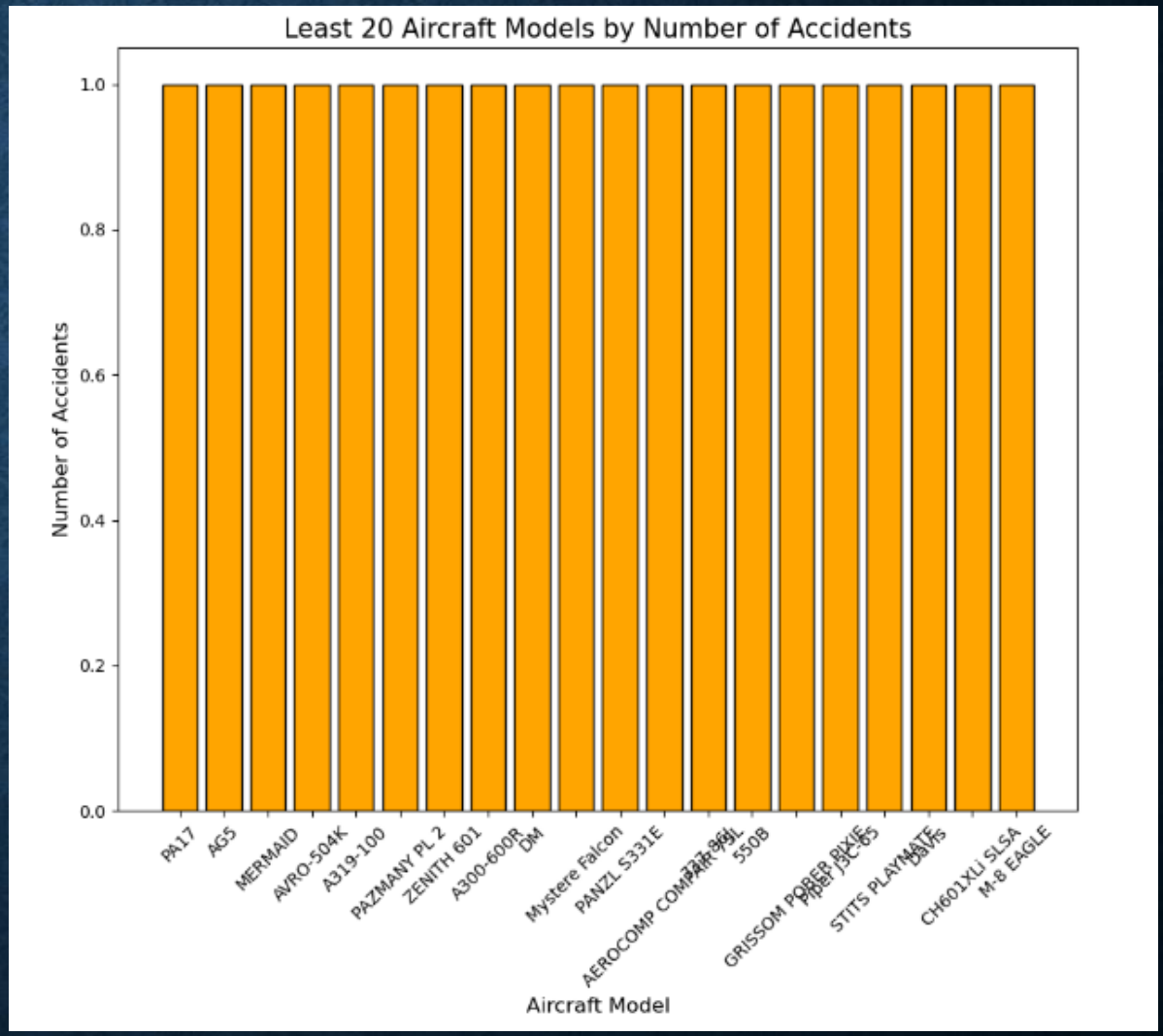


Observation

1. **Cessna** and **Piper** are highly represented in the data set with most non-fatal incidents.
2. For the least frequent planes, most incidents are reported as **minor injuries**.
3. **Fatalities** are more prominent in frequently used aircraft, but **minor injuries** dominate in less common aircraft.

Top 20 Aircraft Models by Number of Accidents

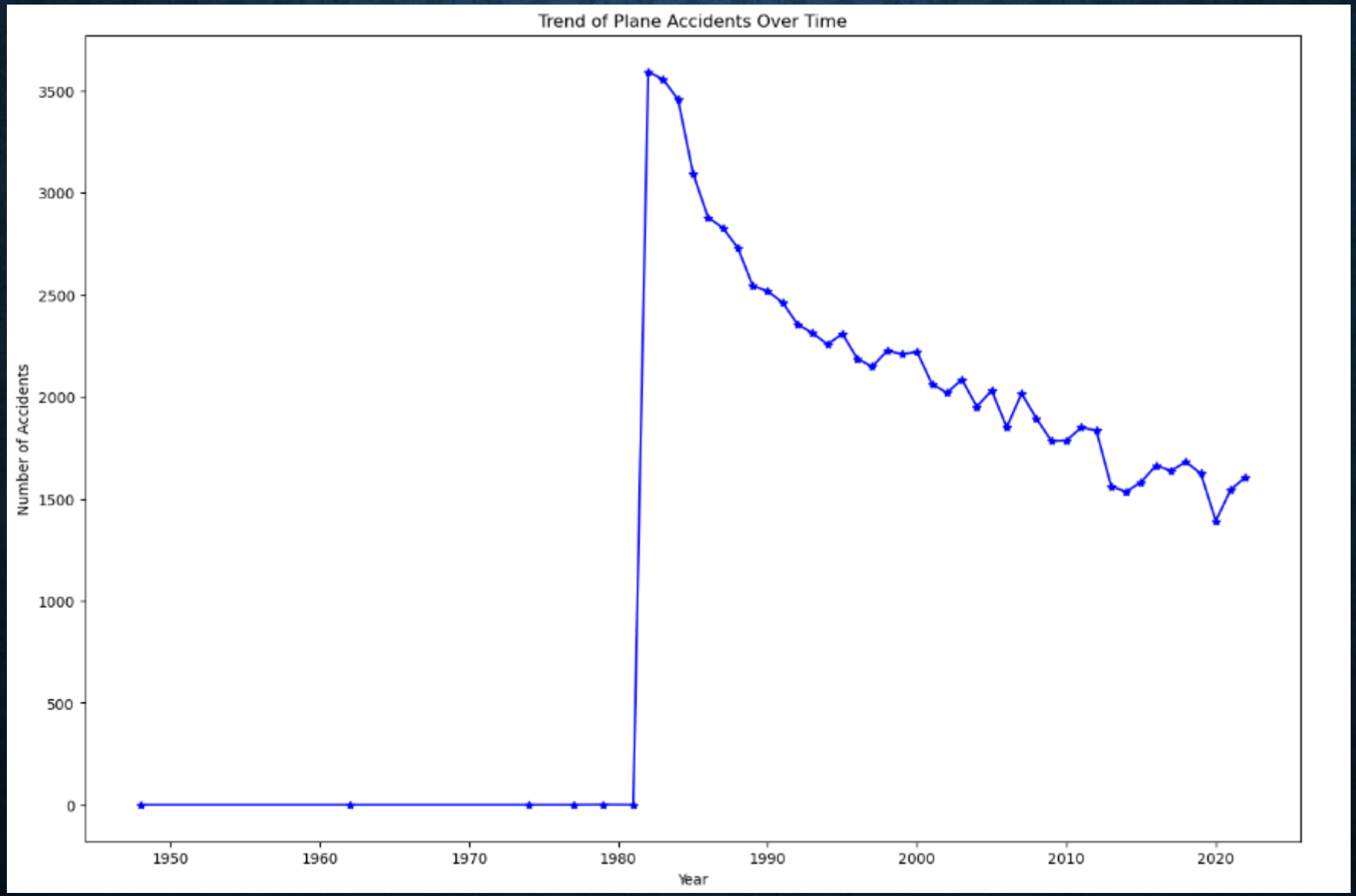




Observation

Model 152, 172 and 172N have relatively very high incidents hence I can term them as riskier.

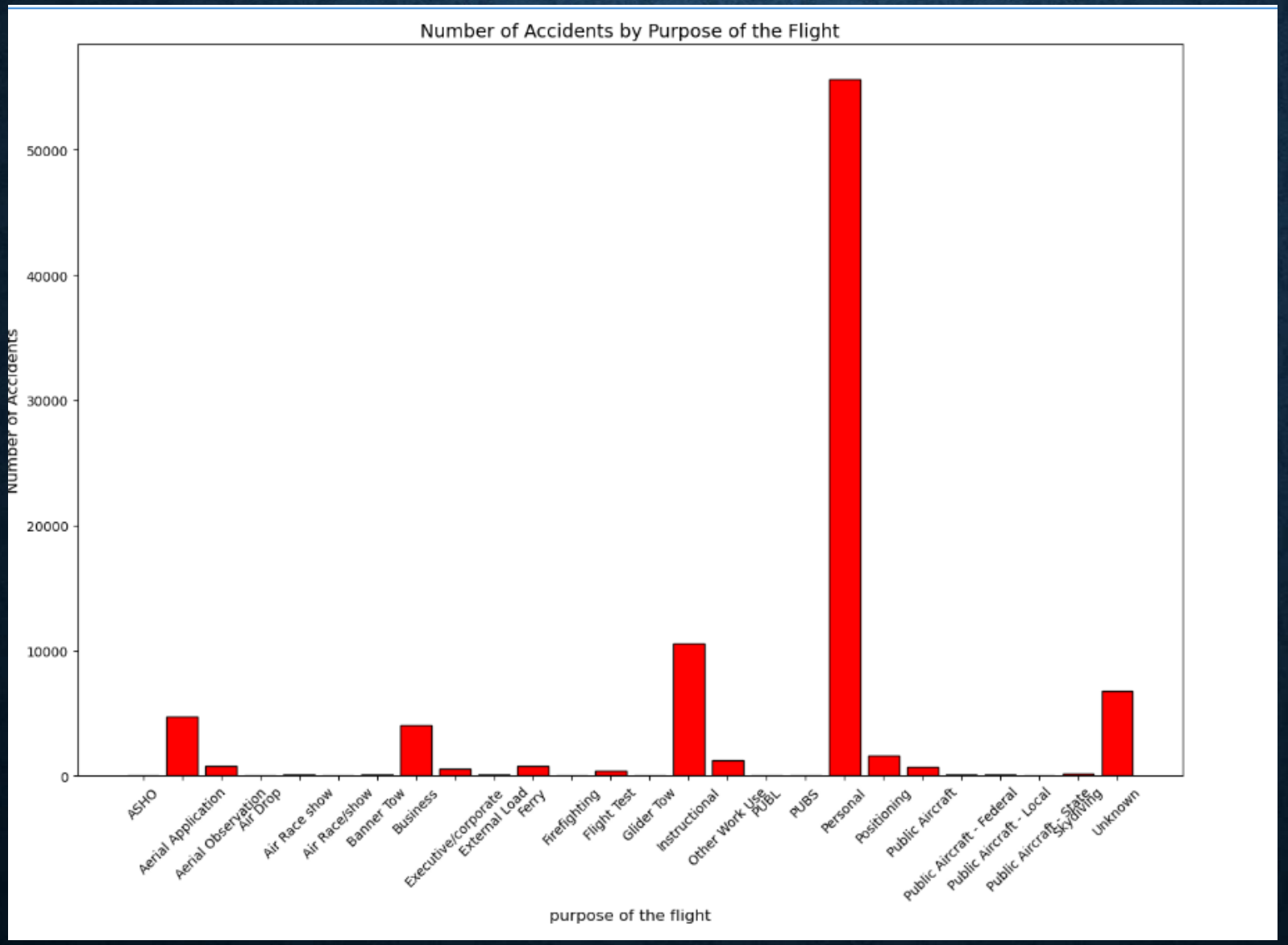
Trend of Plane Accidents Over Time



Observation

From the graph above, I can conclude that plane accidents have been reducing over the last Forty years despite us knowing that there has been a significant increase in the number of flights

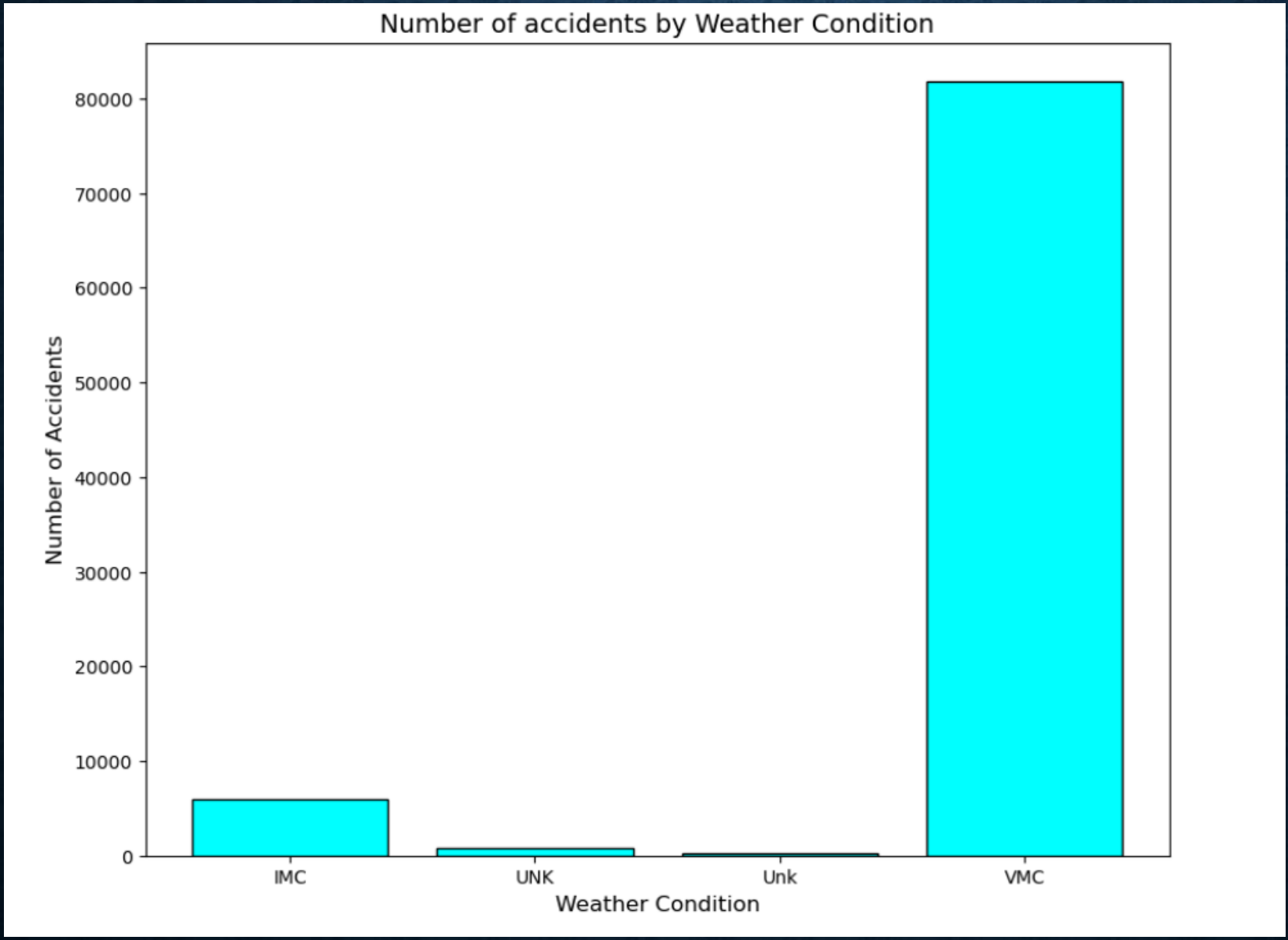
Number of Accidents by Purpose of the Flight



Observation

The bar graph shows the number of accidents based on the purpose of the flight. "Personal" flights have by far the highest number of incidents, with a significantly taller bar compared to all other categories. Other flight purposes, such as instructional, business, and aerial application, also have some accidents but at much lower rates. A few categories, like "Unknown" and "Public Aircraft," have fewer incidents. This suggests that personal flights are involved in the most accidents compared to other types of flights.

Number of accidents by Weather Condition



Observation

Many accidents tend to happen in Visual Metrological Conditions (*VMC*) where the pilots are advised to fly by sight compared to the Instrumental Metrological Conditions (*IMS*) where the pilots are required to use the instruments due to limited visibility. Both (*UNK*) and (*unk*) represents **unknown data**.

CONCLUSION

From my analysis, I have come up with the following observations:

1. **Aircraft Make**: Some aircraft models have higher accident rates with Cessna, Piper, and Beech being on top each with over 5000 cases.
2. **Number of Engines**: Single-engine aircraft are riskier than those with multiple engines. This might be caused by a lack of a backup in case of an engine failure which is something common with aircraft.
3. **Engine Type**: Aircraft with reciprocating engines have recorded more accidents and can be considered riskier.
4. **Purpose of Flight**: Personal and business flights often have higher accident rates. This might be caused by the fact that most of them have single engines.
5. **Weather Conditions**: Flying in clear weather *(Visual Metrological Conditions(VMC))* can still have risks.

Proposed Guides for Buying an Aircraft

1. **Choose Multi-Engine Aircraft**: Buy planes with two or more engines to lower the risk of engine failure and in case of an engine failure there can be a backup.
2. **Consider Turbo Jet Engines**: Jet engines may cost more but usually have fewer accidents than reciprocating engines.
3. **Avoid High-Risk Models**: From the analysis, some aircraft models tend to pose more risk with Model 152, 172 and 172N having relatively very high incidents hence I can term them as riskier.
4. **Focusing on Commercial and Cargo Flights**: These types of operations generally have fewer accidents compared to personal or business flights.
5. **Monitor Weather**: The company can invest more in weather monitoring tools as well as improving the pilots' skills since from the analysis where I was comparing the weather conditions and their relative number of accidents, I noted that many accidents occurred in the *(Visual Metrological Conditions(VMC))* where the pilots are advised to fly by sight compared to *(Instrumental Metrological Conditions (IMS))* where the pilots use weather monitoring tools.
6. **Maintain Aircraft Carefully**: I propose this because, having some aircraft with only a single engine, hence no backups, requires a fully functional engine and this can be achieved by regular inspections to make sure everything is well.
7. **Acquire Insurance**: From the analysis, some aircraft are damaged, some have substantial damage and others have minor damage. These damages can be costly to repair and in case of any serious injuries from those on board, the treatment too can be costly hence getting insurance is advised.

THANK YOU.