















 odhinto /  
dsc-phase-1-project-v3



 Code  Pull requests  Actions  Projects  Wiki  Security  Insights  Settings

[View license](#)

★ 0 stars    178 forks    0 watching    Branches    Activity  
 Tags

🌐 Public repository · Forked from [learn-co-curriculum/dsc-phase-1-project-v3](#)

  1 Branch  0 Tags  

 Go to file 







 Go to file 

 Add file ▾

Code 

This branch is **36 commits ahead of** learn-co-curriculum/dsc-phase-1-project-v3:master .

[Contribute](#) ▾[Sync fork](#) ▾**odhinto** notebook final editsabddbd0 · 50 minutes ago 

 data	final touches	1 hour ago
 .gitignore	add .gitignore and init repo	last year
 LICENSE.md	add contributing and license	last year
 README.md	final touches	1 hour ago
 encoded_aviation_data.csv	notebook final edits	50 minutes ago
 index.ipynb	notebook final edits	50 minutes ago

# Hi 🙌, I'm Anthony Odhiambo

Connect with me:



Github Repository: [Click to Open the Project Github Repository](#)

Tableau Dashboard: [Click to Open the Tableau Dashboard](#)

# Problem Definition

---

Our company's expansion and diversification plans include venturing into the aviation industry to own and operate airplanes for commercial and private enterprises. A key preliminary step for this consideration is risk assesment for different aircrafts to advise which aircrafts pose the lowest risk for the intended business endeavor. This projects seeks to assess risk potential from analysis of aviation accident data from 1962 to 2023.

## Business Understanding

---

The primary objective of this exercise is to identify the lowest-risk aircraft for our company to purchase and operate. The following are some of the key considerations we expect to make:

- Historical accident trends by aircraft type
- Severity and frequency of accidents
- Factors contributing to the accidents e.g. weather, pilot error or mechanical failures
- Any correlations between operational risk and aircraft characteristics

From the data, our company can also assess which types of aviation ventures pose the least risks e.g. personal flights, commercial flights, instructional flights etc.

## Data Preprocessing

---

This section prepare the provided aviation data for analysis. We intend to do the following:

- Dataset Overview - Load and understand the data
- Handling Missing Values using derived domain knowledge and imputation
- Data Cleaning e.g. standardizing categorical values, deriving useful date data, removing duplicates etc

## Dataset Overview

---

It is imperative for us to understand the aviation dataset first i.e.:

- The data structure e.g. available columns, data types and presence of missing values
- Establish the relevance of the data to our study
- Identify useful columns to focus on

Data Understanding will prescribe subsequent cleaning steps to be done in the **Data Cleaning** subsection

## Python Libraries Initialization

First, we initialize common libraries we project to utilize in this exercise:

- pandas to create and manipulate dataframes
- seaborn and matplotlib to facilitate any requisite visualizations within the notebook
- numpy for mathematical calculations

- etc

## Data Loading

We then load the dataset into python as a dataframe and embark on a data understanding exercise.

## Data Understanding

Sampling a few columns to assess the various categorical data present will help us define how to clean the data, e.g.:

- **Investigation.Type:** There are 71 different categories, but the only seemingly relevant categories here are 'Accident' and 'Incident'. The others seem like noise. Perusal of the 'noisy' data on excel showed that where the column "Event.Id" was blank, there was a noisy column in 'Investigation.Type'. Dropping rows with empty Event.ID may fix this
- **Injury.Severity:** Where there are fatalities, the number of fatalities is appended on the string. This is unnecessary, since there's another column that details the number of fatalities. There is a need to clean this column to only have **Fatal**, **Serious**, **Minor**, **Non-Fatal**, **Incident** and **Unavailable**. There is missing data.
- **Aircraft.damage:** The categorization of extent of damage seems okay as is i.e. 4 categories (**Destroyed**, **Substantial**, **Minor**, **Unknown**). There is no need for cleaning this column. There is missing data.
- **Aircraft.Category:** The categorization of Aircrafts seems okay as is. But there is a need to combine 'Unknown' and 'UNK' such that they are the same category. There is missing data.
- **Make:** Unstandardized capitalization is causing the same manufacturer to be split e.g. Cessna vs CESSNA. There is a need to regularize capitalization to prevent such an error. Also, there is noise introduced by having entries such as 'Cessna Aircraft' vs 'Cessna Aircraft Co' vs 'Cessna Aircraft Co.'. As they are too many possibilities, we may need to ignore this for now (and revisit once we can make use of fuzzy logic to normalize similar/equivalent data). There is missing data.

Make	Make
<input checked="" type="checkbox"/> Cesna	<input type="checkbox"/> Boeing
<input checked="" type="checkbox"/> Cessna	<input type="checkbox"/> Boeing 777-306Er
<input checked="" type="checkbox"/> Cessna Aircraft	<input type="checkbox"/> Boeing - Canada (De Havilland)
<input checked="" type="checkbox"/> Cessna Aircraft Co	<input type="checkbox"/> Boeing (Stearman)
<input checked="" type="checkbox"/> Cessna Aircraft Co.	<input type="checkbox"/> Boeing Commercial Airplane Gro
<input checked="" type="checkbox"/> Cessna Aircraft Company	<input type="checkbox"/> Boeing Company
<input checked="" type="checkbox"/> Cessna Ector	<input type="checkbox"/> Boeing Company, Long Beach Div
<input checked="" type="checkbox"/> Cessna Reems	<input type="checkbox"/> Boeing Helicopters Div.
<input checked="" type="checkbox"/> Cessna Reims	<input type="checkbox"/> Boeing Of Canada/Dehav Div
<input checked="" type="checkbox"/> Cessna Robertson	<input type="checkbox"/> Boeing Stearman
<input checked="" type="checkbox"/> Cessna Skyhawk li	<input type="checkbox"/> Boeing Vertol
<input checked="" type="checkbox"/> Cessna Soloy	<input type="checkbox"/> Boeing-Brown
<input checked="" type="checkbox"/> Cessna Wren	<input type="checkbox"/> Boeing-Stearman
<input checked="" type="checkbox"/> Cessna/Air Repair Inc	<input type="checkbox"/> Boeing-Vertol
<input checked="" type="checkbox"/> Cessna/Weaver	

*Dirty Makers - Needs Cleaning with Fuzzy Logic*

- **Model:** A bit of non-standardized capitalization introduces noise into the data. This will need to be corrected. There is missing data.
- **Amateur.Built:** There are 2 categories: **Yes** and **No** as well as missing data.
- **Engine.Type:** The column seems clean, with Reciprocating Engine Aircrafts accounting for majority of the accidents/incidents. Entries with "UNK" should be substituted with "Unknown" to clean the data. There are missing values.

- **Purpose.of.flight:** The column seems clean, with personal flights accounting for majority of the accidents/incidents.
- **Weather.Condition:** IMC, or **Instrument Meteorological Conditions**, are weather conditions that require a pilot to rely on flight instruments. On the other hand, VMC, or **Visual Meteorological Conditions**, are weather conditions that allow a pilot to navigate by visual reference to the ground and other landmarks. 91.6% of the accidents occurred during VMC i.e. weather is hardly a factor leading to accidents. We need to combine 'UNK' with 'Unk' i.e. standardize capitalization.
- **Broad.phase.of.flight:** Majority of the accidents occur during landing e.g. 24.995%.
- **Schedule:** There seems to be a rather even distribution between scheduled and non-scheduled flights where accidents/incidents occurred, although more accidents occur for non-scheduled flights (35% vs 31%). We will need to replace "UNK" with "Unknown" to maintain similar format.

Summarily, the dataset contains 31 columns and 90348 rows (including the header columns). There are several missing values in different columns. The only column without any missing data is the 'Investigation.Type' column, and this can form a good place to start with the data cleaning exercise.

Further perusal of the data in Microsoft Excel gave some preliminary insights that can advise the data cleaning exercise:

- Where 'Event.Id' is blank, all the other columns are also blank. These could be deleted from the onset
- 'Event.Date' uses a YYYY-MM-DD format whereas 'Publication.Date' uses a DD-MM-YYYY format. It would be better to standardize the date formats
- 'Investigation.Type' seems to have 2 relevant values, i.e. "Accident" and "Incident". The rest of the values seem to be dates, and where it is a date, the rest of the columns are empty. Such rows can also be deleted from the onset.
- In the column 'Injury.Severity', there are too many categories since the number of fatalities is appended beside the label 'Fatal'. This is repetitive since there is another independent column 'Total.Fatal.Injuries' that details the number of fatalities. It may be better to just define the category 'fatal' for this column.
- In the column 'Make', capitalization differences have been noted e.g. 'CESSNA' vs 'Cessna'. This could make python consider these as two different makers. This needs to be standardized/corrected.

## Data Cleaning

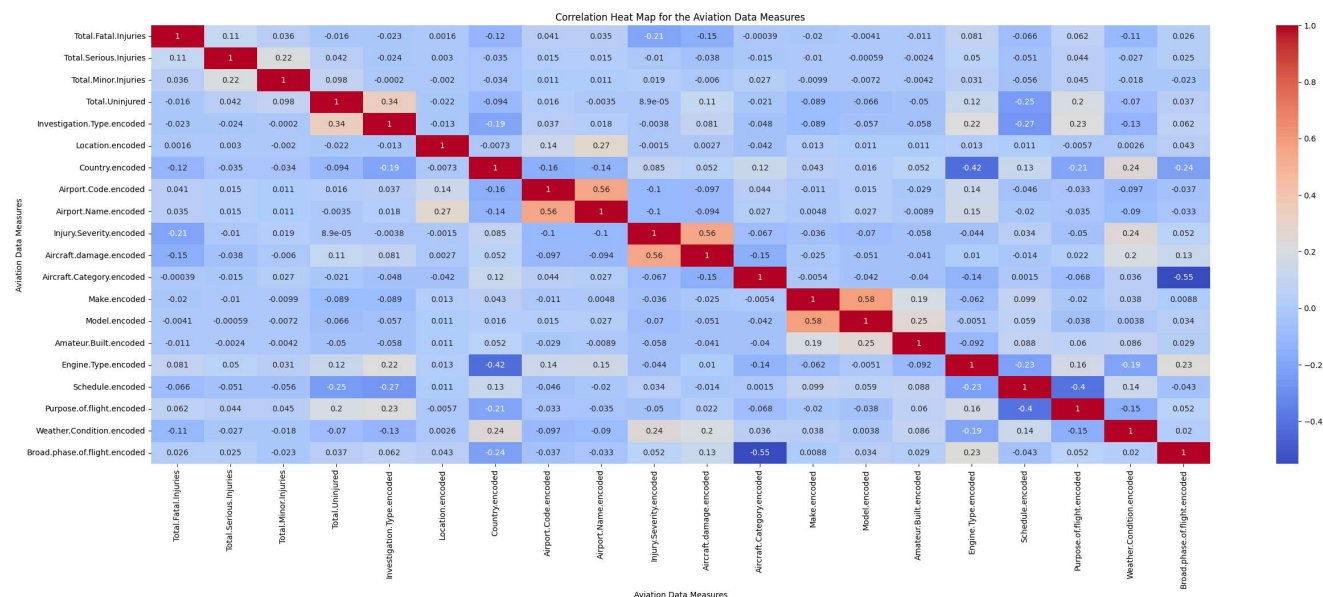
From the data cleaning requirements identified during the data understanding step, the following cleaning exercises were implemented:

- Removing rows with empty **Event.ID** values since for such cases, most of the other values were blank. This yielded clean **Investigation.Type** columns as well.
- Removing rows with empty **Make** values since the main objective of this study is to identify the lowest-risk air crafts to invest in.
- Removing the bracketed number of fatalities in the **Injury.Severity** column since the same can be evaluated from the **Total.Fatal.Injuries** column.
- Normalizing capitalization formatting in all the categorical fields to ensure clean categories.
- Replacing **UNK** with **Unknown** in all relevant fields to enhance understanding of the data.

- Converting Date Columns **Event.Date** and **Publication.Date** to Python Datetime format
- Handling missing values: We will use different strategies to handle different categories of missing data:
  - Most of the longitude and latitude data is missing, and it does not seem to be relevant to this study. Additionally, we also have location data which is more readily available. Thus, we can drop the longitude and latitude columns
  - FAR.Description data seems irrelevant to the study, and so it was dropped
  - For the few missing values of relevant categorical columns, we can **\*\*fill missing values with "Unknown"**. We may be able to extract insight even with "Unknown" parameters if others related parameters are known e.g. you may have an unknown model but know the manufacturer. This remains relevant to our study.
  - For the data on total number of injuries, it is best to assume that the data meant to be there is '0' e.g. if Total.Serious.Injuries is empty, it means there were no fatal injuries
- Checking and removing any duplicates
- Encoding categorical data into numbers to facilitate correlation calculations.

## Exploratory Data Analysis

We can check for correlation in the measures in our data:



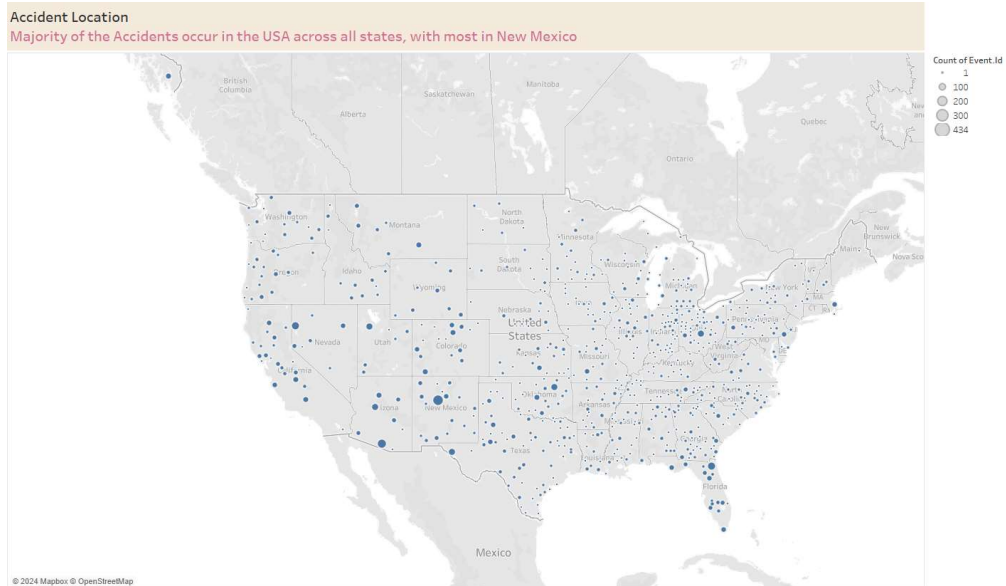
### Testing for Correlation in the Data Measures

There is no immediate correlational insight noted from correlation heat map. This could be due to the several cases of "Unknown" category. It is prudent to carry out further exploratory analysis using Tableau.

## Tableau EDA

Most of the accidents occur in the USA, distributed across almost all the states.

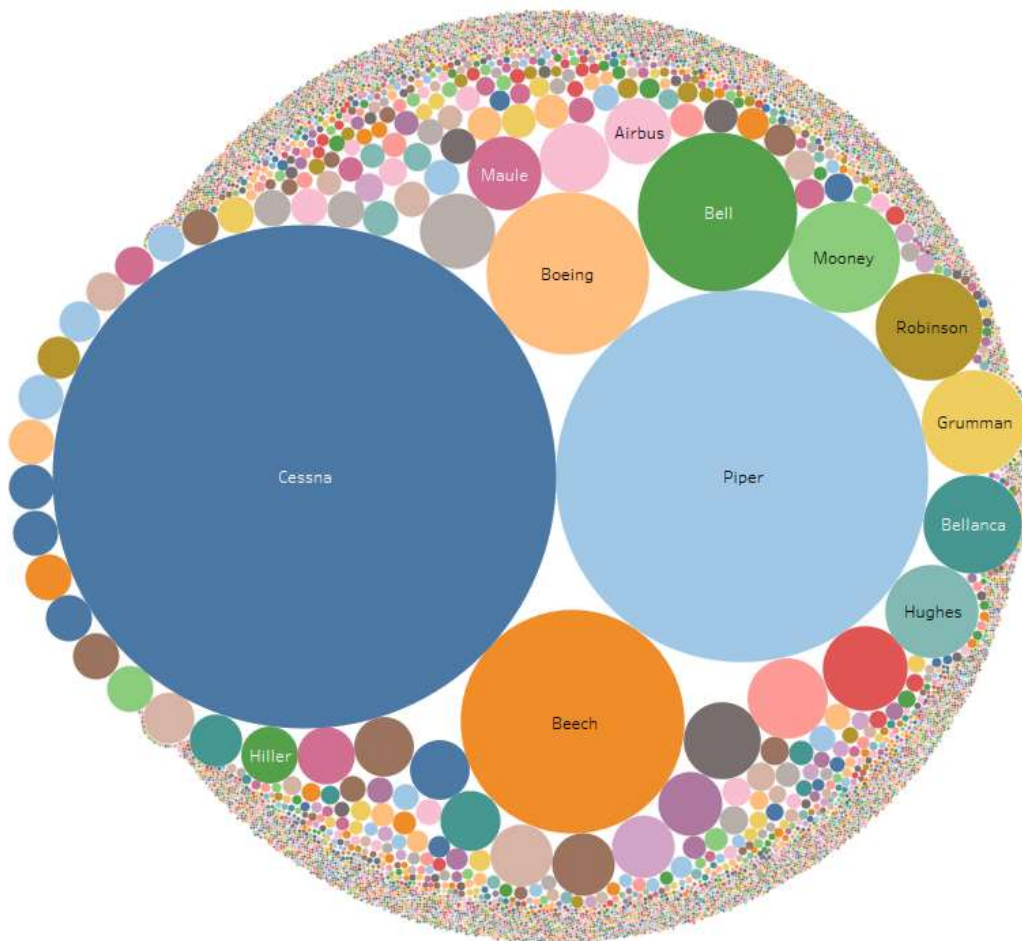




*Recorded Accident Locations*

Cessna, Piper, Beech, Boeing and Bell registered the most accidents.

**Accident Frequency Per Maker**  
Cessna, Piper, Beech, Boeing and Bell Register Top 5 Accidents

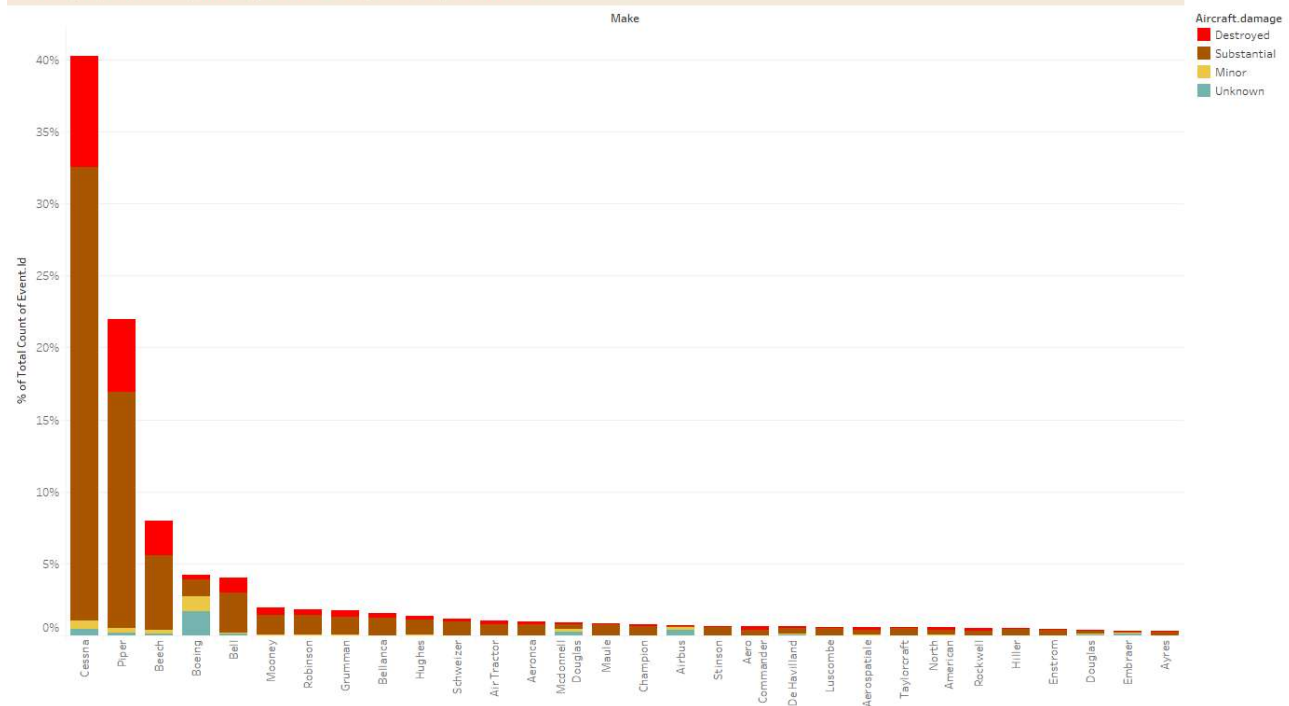


*Aviation Accidents Per Maker*

In the event of an accident, it is almost guaranteed that the aircraft damage will be substantial to totally damaged for most of the makers.

#### Aircraft Damage Per Maker

Cessna, Piper, Beech, Boeing and Bell Register Top 5 Accidents

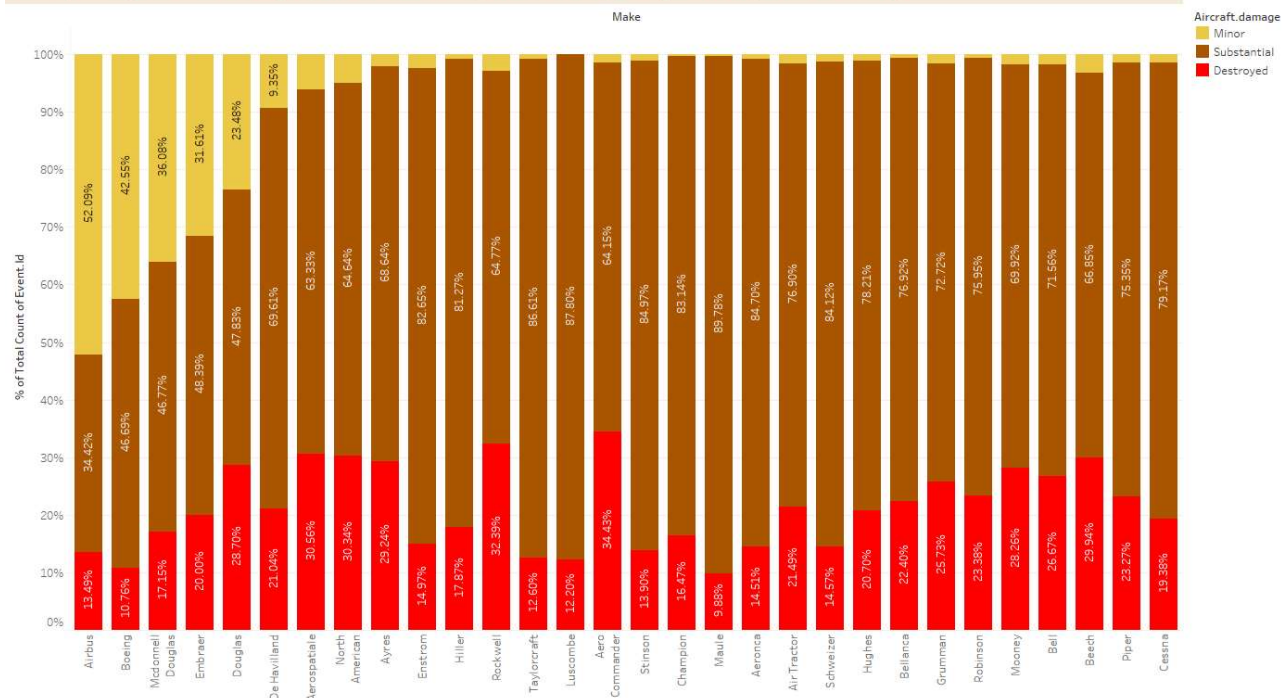


If you remove events where the aircraft damage is "Unknown", the following makers emerge as better candidates where there is a high likelihood of minor the accident only resulting in minor damage:

- Airbus
- Boeing
- McDonnell Douglas
- Embraer
- Douglas

## Aircraft Damage Per Maker %

Airbus, Boeing, McDonnell Douglas, Embraer and Douglas offer the highest probability of minor damage



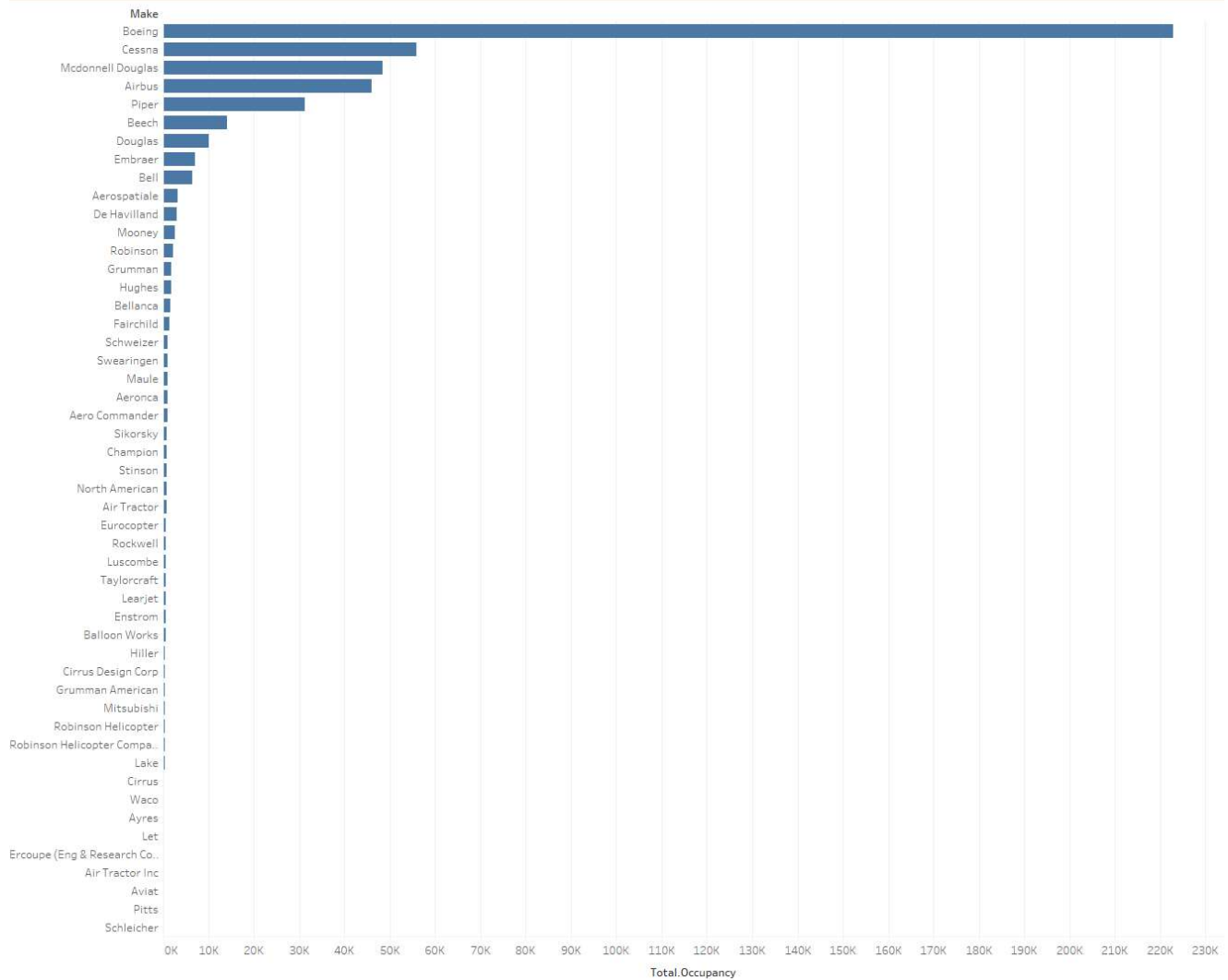
Aircraft Damage Per Maker

To analyze fatalities in the event of an accident, it is important to understand the total number of passengers involved in this data. This can be done by summing up all the fatalities, serious injuries, minor injuries and uninjured passengers.



### Total Occupancy

Boeing, Cessna, McDonnell Douglas, Airbus and Piper have carried the most passengers in total

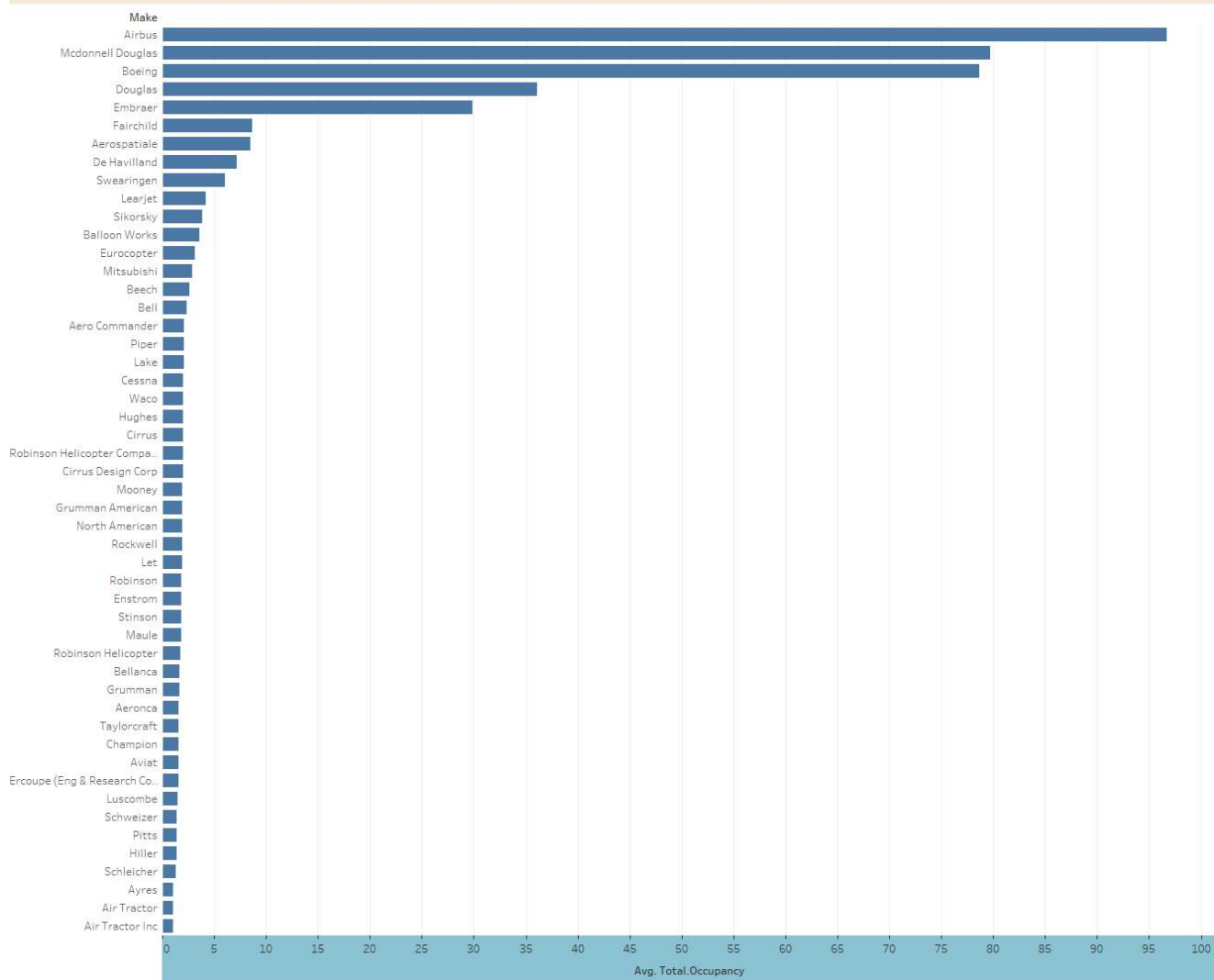


*Total Occupancy Per Maker*

We go a step further to assess the average occupancy per flight from the data. This will tell us the relative sizes of the aircrafts.

### Average Occupancy Per Maker

Airbus, McDonnell Douglas, Boeing, Douglas and Embraer have the highest average occupancy



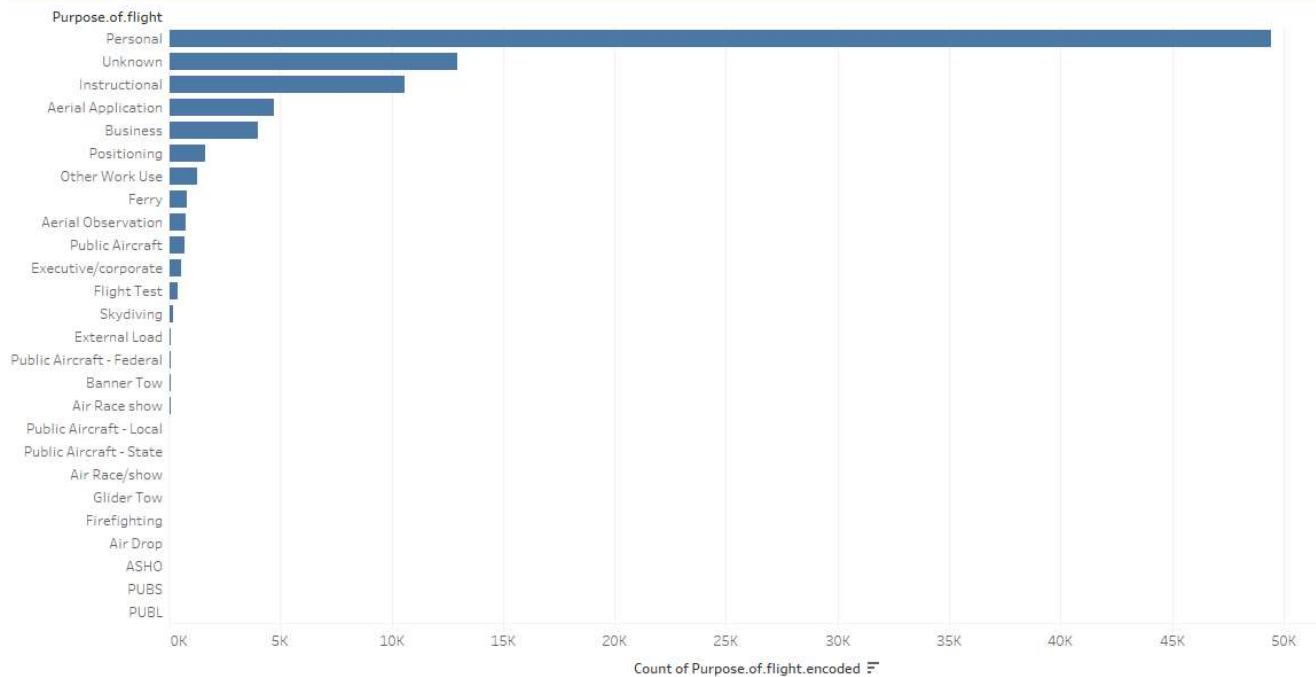
Average Occupancy Per Maker

Airbus, McDonnell Douglas, Boeing, Douglas and Embraer are generally bigger planes carrying more people, hence their high number of average occupancy. Cessna has a very low average occupancy, i.e. they make small aircrafts carrying very few people (around 2).

It is important to consider the purpose of the flight from the accident data.

### Accident Frequency Based on Purpose of Flight

56% of accidents occur on personal flights



### Accident Distribution Based on Purpose of Flight

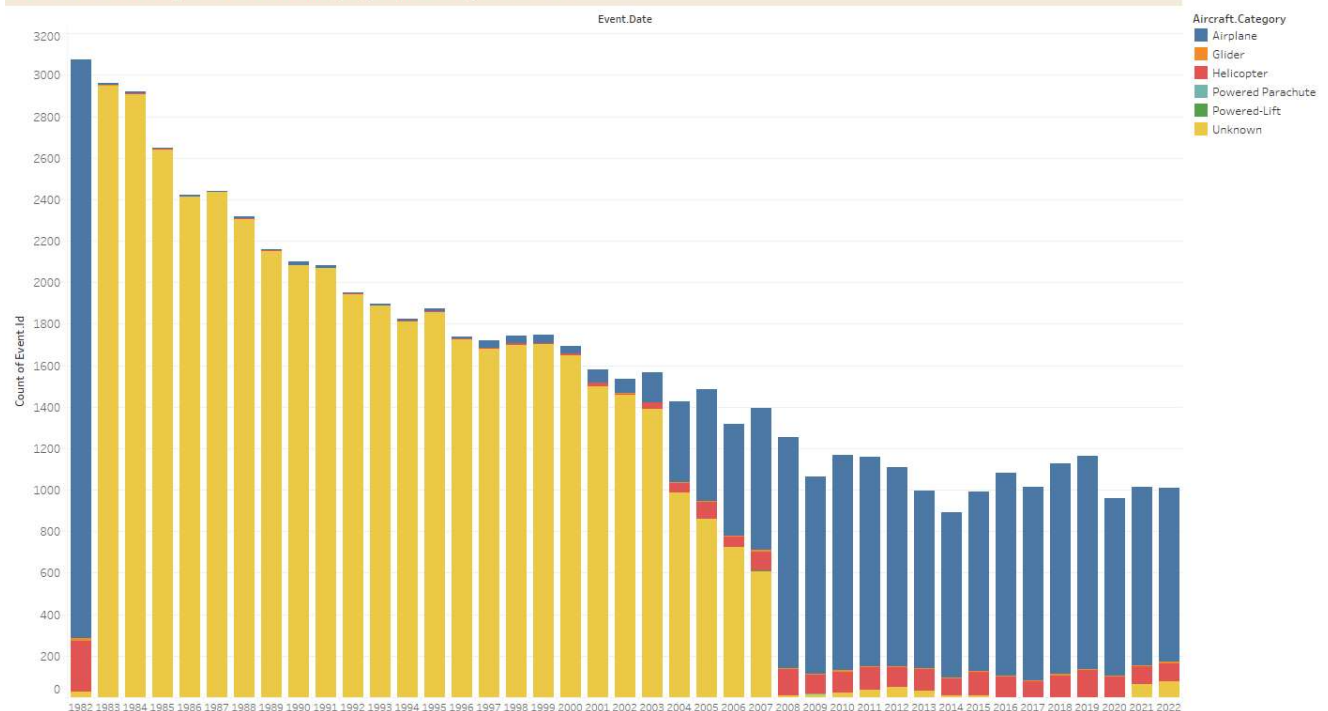
Most of the accidents occurred during personal flights.

**NB:-** Thus, this branch of aviation seems the most risky that our company should steer clear from, or only engage in with extreme caution.

Next, it is important to understand distribution of accidents across different aircraft categories over the years.

### Accident Trends Per Aircraft Category

From 1983 to 2007, the Aircraft Category is primarily unknown



### Accidents Trends per Aircraft Category

Majority of the aircraft category data is "unknown". Filtering out unknown data will give a better indication.

Accident Trends Per Aircraft Category (less Unknown Categories)

Removing data where Aircraft Category is Unknown, Airplanes register the most accidents



Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Languages

- Jupyter Notebook 100.0%