# 

# Aircraft Collisions in the USA

Heer Bhavesh Parekh

Department of Applied Sciences, San Jose State University

DATA 230: Data Visualization

Prof. Andrew Bond

December 12, 2021

# Abstract

Since the late 1950s, the number of air crashes has remained consistent. In the United States of America, detailed study of the causal elements is not done regularly. Nevertheless, the examination of flying catastrophes is a critical foundation for flight safety. Except for operations involving commercial transportation services, general aviation encompasses all civilian flying. Unfortunately, this aviation branch has a poor safety record, accounting for 94% of all civil aircraft fatalities. Plane crashes are among the foremost damaging mischances causing 400-600 calamities. Major accident risk factors include adverse weather, geographical region, post-impact fires, etc.

The Convention on International Civil Aviation defines an air transport disaster as an event occurring related to the operation of an aircraft that occurs between the moment any person boards the airplane to fly and the period all such individuals depart, in which:

a) an individual is fatally or deeply wounded,

b) the plane perpetuates significant damage or malfunctions.

An aviation incident is defined as an occurrence related to an airplane's functioning that impacts or might influence the integrity of operating the aircraft, besides an accident. I want to visualize and evaluate the data in this project by discussing the different causes of aviation incidents and accidents that have been caused within the United States and how we could increase the quality and safety of air travel. Visualizations would make finding the large dataset's patterns, trends, and outliers easier.

***Github: github.com/Heer-14/Data\_Visualization***

**Table of Contents**

# Objectives

The goal of the dashboard is to attempt to answer the following questions about the aircraft collisions in the USA. Below are interactive visualizations on the tableau are created:

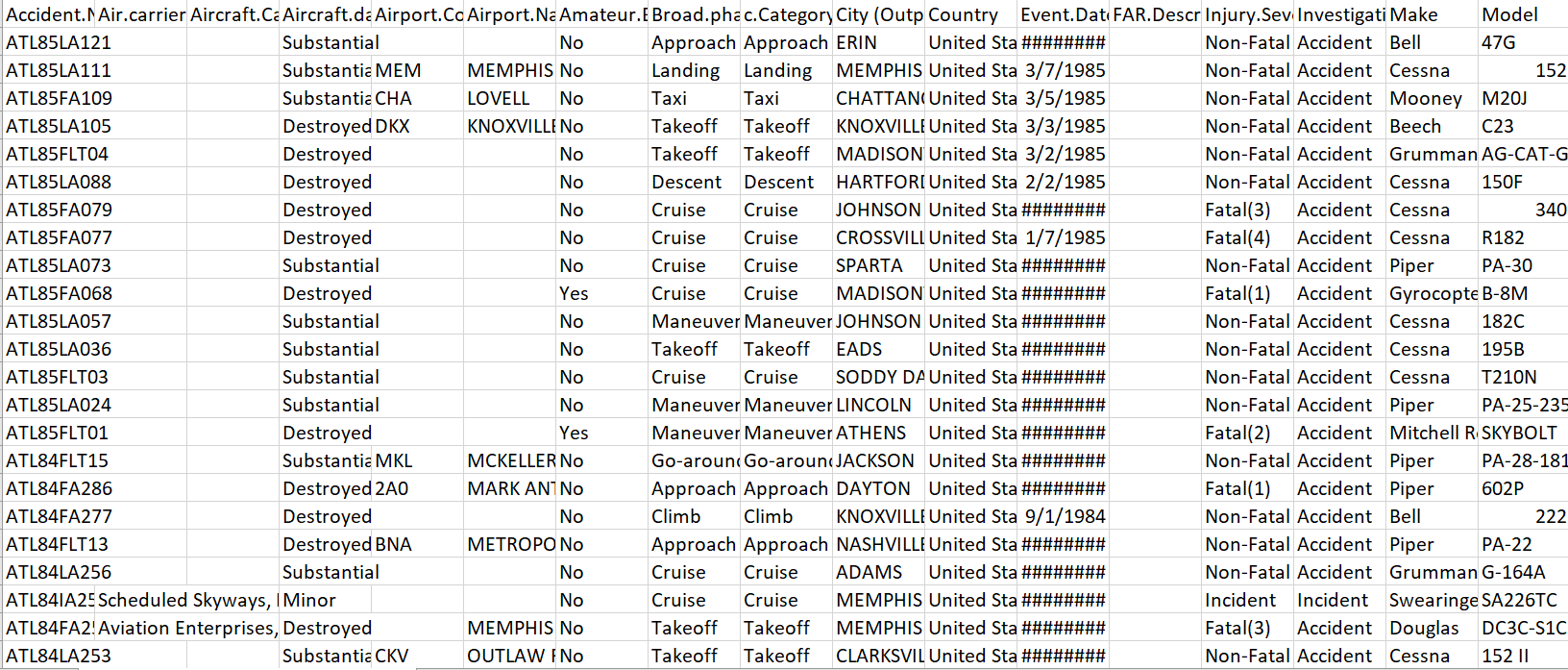
1. Total statistics of aircraft accidents from 1982-2021
2. The most common phase of accidents
3. Amount of damage caused to an aircraft during a crash
4. State in the USA with the most accidents
5. The exact locations where the collision was detected
6. The fatality and injury rate during the hits per year

# Dataset Description

The chosen dataset provides data on civil aviation accidents and incidents in the United States, its regions and possessions, and global waterways from 1982 to the present. For this project, I have taken two datasets, the primary and the other one is the secondary. Both the datasets are taken from Kaggle.com and are in .csv format. The first dataset is 'output.csv .' The other is 'Hexagon\_map.csv.'

Output.csv dataset consists of 30 attributes and has approximately eighty thousand records. It consists of records from the year 1982 to 2020. Some of the attributes of the datasets are provided below in figure 1:

**Figure 1**



**Event.Id:** ID of the event

**Investigation.Type:** Type of investigation, whether accident or incident.

**Accident.Number:** Accident number

**Event.Date:** Date of Event

**Location:** The location of the event where aircraft collided.

**Country:** Name of Country

**Latitude:** Latitudinal points of the location

**Longitude:** Longitudinal points of the location

**Airport. Code:** The codes associated with the Airports

**Airport. Name:** Name of the Airport

**Injury.Severity:** The severity of harm caused to the passengers after the collision. The severity of injuries is divided into:

* Fatal
* Non-Fatal
* Minor
* Serious

**Aircraft. Damage:** Aircraft level of damage. It is divided into three:

* Substantial
* Destroyed
* Minor

**Aircraft. Category:** The category of air carriers. The types are:

* Airplane
* Balloon
* Gliders
* Gyrocrafts
* Gyroplane
* Helicopter
* Powered parachute
* Rocket
* Ultralights
* Weight-shifts

**Registration. Number:** Registration Number of ID

**Make:** Name of the manufacturer

**Model:** Name of the model of the air carrier

**Amateur. Built:** The air carrier was amateur-built or created by a pro. This field has boolean values.

**Number of Engines:** The number of engines that the air carrier was built with

**Engine\_Type:** Type of Engine.

**FAR. Description:** Far Description

**Schedule:** Schedule the aircraft whether it was on time or not.

**Purpose. Of. Flight:** The purpose of the passengers to board the air carrier

**Air.Carrier:** Name of the air carrier

**Total. Fatal. Injuries:** Number of passengers that died in the collision.

**Total. Serious Injuries:** Number of passengers that were seriously injured in the collision.

**Total. Minor.Injuries:** Number of passengers that were minorly injured in the collision.

**Total. Uninjured:** Number of passengers that were uninjured in the collision.

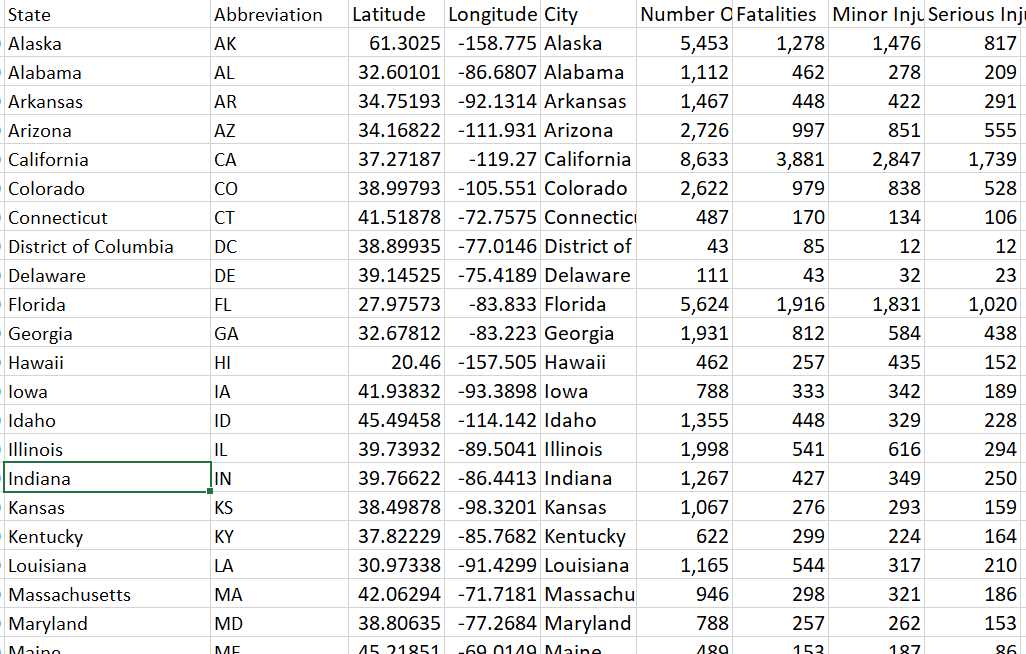
**Weather. Condition:** The condition of the weather at the time of the collision.

**Broad.phase.of. Flight:** The stage of flight when the crash took place.

**Report.Status:** The status of the report generated after the crash

The second dataset is the Hexagon\_map.csv. This dataset has nine attributes with sixty records. It has been collected from Kaggle.com. It has the below attributes as shown in figure 2:

Figure 2



**State**: Name of the states in the USA

**Abbreviation:** State abbreviations

**Latitude:** Latitudinal points of the states

**Longitude:** Longitudinal points of the states

**City:** Name of the city in the USA

**Number. Of.Accidents:** Accident number for the crashes

**Fatalities:** Passengers died in the crash

**Minor\_Injuries:** Passengers injured during the crash

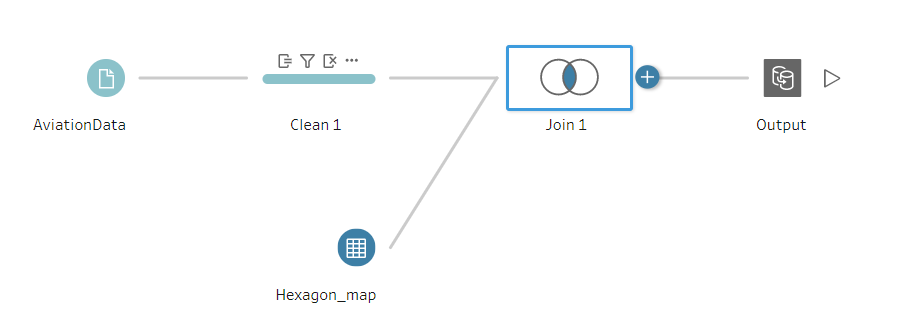
**Serious Injuries:** Passengers seriously injured during the crash

**Country:** Name of Country

# Data Processing

Kaggle.com provided the data for this research. For feature extraction and data quality checks, datasets are rigorously inspected. However, the raw datasets required a lot of cleaning and preparation because they were skewed and unprepared. Data preparation techniques include categorizing linear data into discrete categories and keeping categorical variables. As a result, there are many missing values, null values, and redundant fields in the output data. In addition, because the datasets comprise textual, numerical, invalid, and zero-valued information, they will need to be modified and cleaned. So we utilized Tableau Prep Builder for data cleaning and preparation. They filter, convert, separate, and merge datasets with Tableau Prep Builder. Tableau is a software application that makes data preparation easy and painless. Tableau Prep Builder is a tool for preparing data for analysis by collecting, structuring, and categorizing it. At this stage, I have removed all the unwanted columns that would not be required or are not available. These attributes are Events. Id, Registration.Number, Location, etc. I have split some of the columns into two to get better insights; the dataset was subsequently cleansed, the datatypes of the fields were changed, and the city and state variables were given geographical roles. Figure 2 depicts the data preparation and merging sequence for the two datasets by country. It was created using Tableau Prep Builder.

**Figure 2**



# Challenges Faced

The significant challenges that were faced during this project are:

* Tableau Dashboards' Screen Resolution:

The tableau display resolution differs from the end user's screen resolution; the dashboard layout is disrupted. In addition, the arrangement of the dashboards will be slightly distorted.

* Data collection was a complicated process since the latest airlines dataset is difficult to find.
* Data cleaning and processing have taken a lot of time.

# Scope of work

All cities in the US don't have Airports, so that it could show only a few countries in the geodata. Creating multiple dashboards and making them available to the public and Airport authorities will be the most extensive scope of this project.

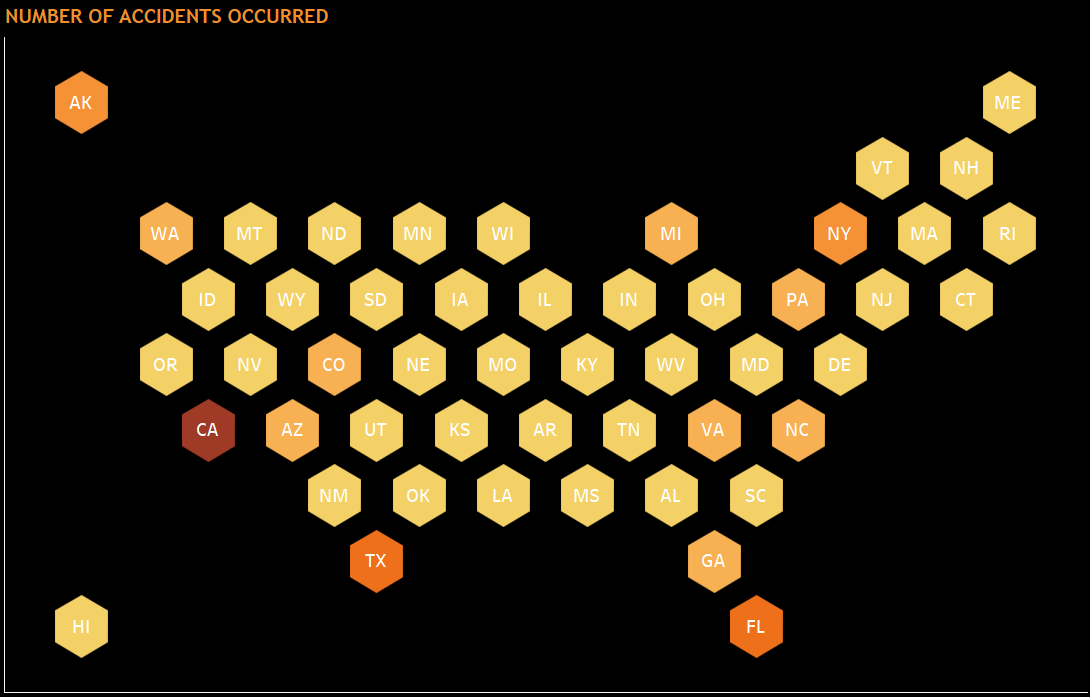
# Visualization Patterns and diagrams

**Choropleth Map:**

Choropleth Maps shows colored, shaded, or patterned geographical areas or regions connected to a data variable. This allows you to visualize data over a large geographic area, allowing you to see variations or patterns across the exhibited area.

The map below shows the total number of accidents in each state. This technique is used to visualize the geodata about the total number of accidents submitted across the United States.

**Figure**

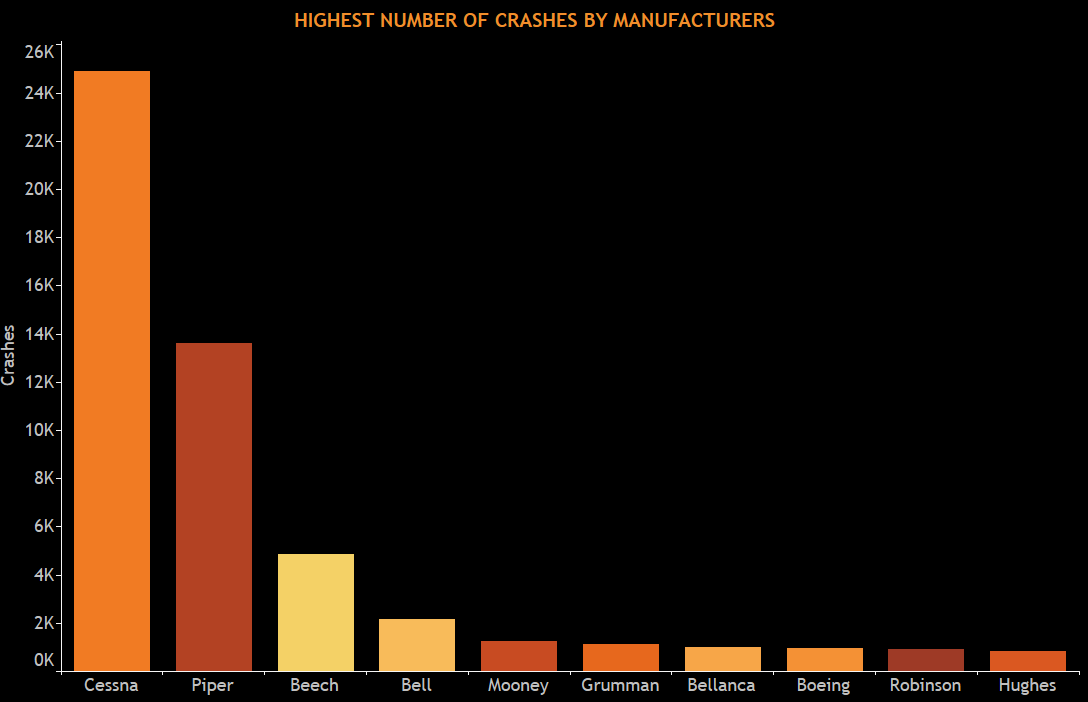


**Bar Charts:**

The traditional Bar Chart shows numerical comparisons across categories using horizontal or vertical bars (column chart). The map's one axis depicts the exact variables being contrasted, while the other is a discontinuous value range.

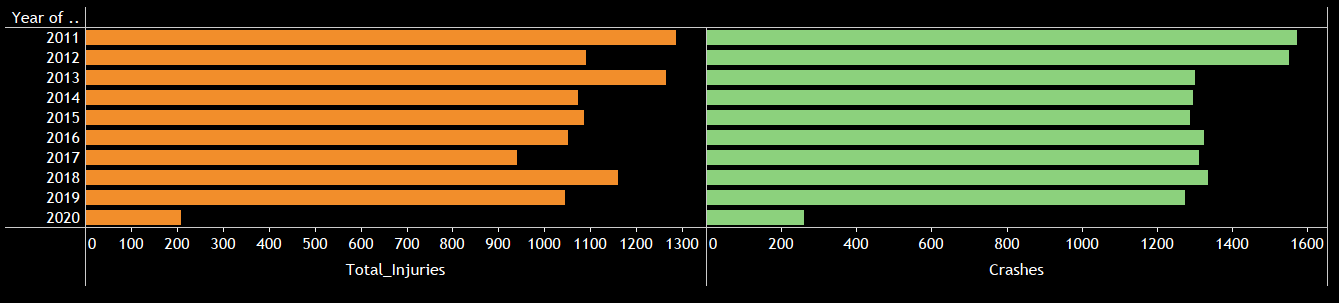
Vertical bars illustrate the top 10 air crashes by the manufacturing companies.

**Figure 3**



Horizontal bars are used to illustrate the count of total\_injuries and crashes per year.

**Figure 4**



**Line Graph:**

Quantitative values are displayed using line graphs across a continuous interval or period. The most common application of a line graph is to demonstrate trends and analyze how data has changed over time.

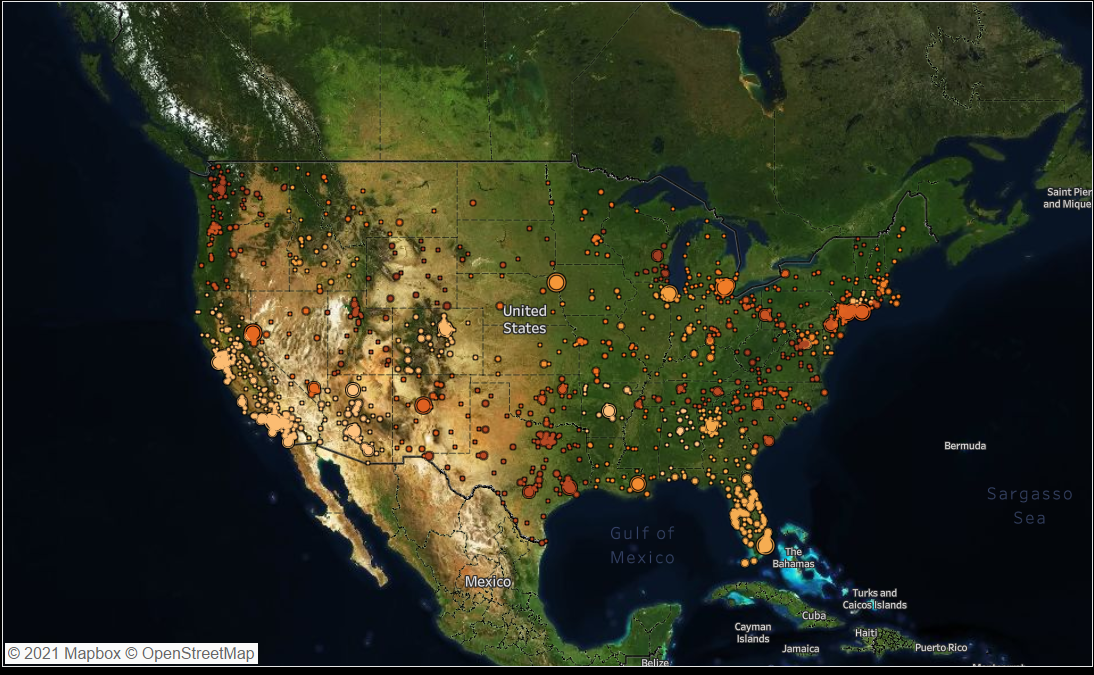
Charts display the number of crashes each year from 1982 to 2020.



**Dot Map**

Dot Maps are a method of discovering spatial patterns or data distribution over a geographic region by scattering evenly sized points across it.

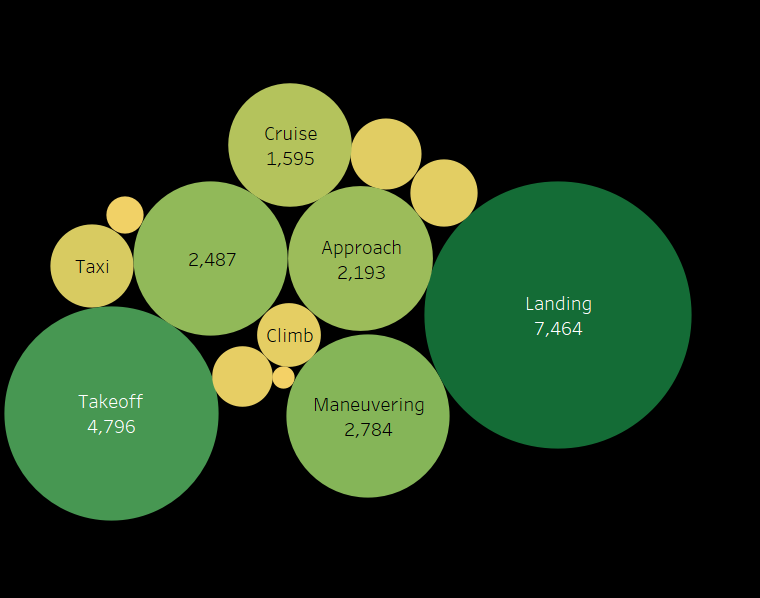
The Dot map illustrates the exact location of air crashes detected in different cities in the United States.



## **Packed Bubble chart:**

The packed bubbles view is a way of displaying relational value without having to worry about axes. The bubbles are packed in as firmly as possible to make the most of the available area.

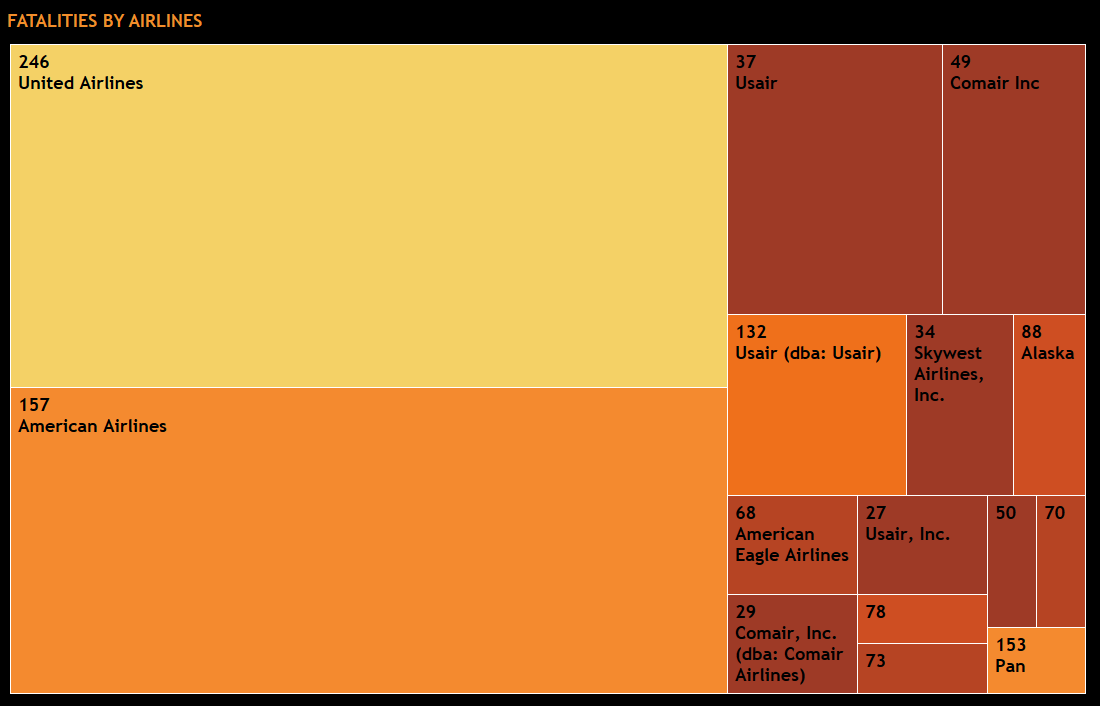
The bubble chart shows the distribution of total crashes as a function of the phase of the plane crash. The total number of crashes determines the size of the packed bubble; the higher the number, the larger the bubble and its color intensity, and vice versa.



**TreeMap:**

To present the data in layered rectangles, use treemaps. Dimensions are used to establish the treemap's structure, whereas measures define the size and color of the individual rectangles. Treemaps are a simple data visualization that can provide information in a visually appealing manner.

The figure shows the treemap depicting the count of fatalities per airline.



**Combination Chart:**

Views that combine numerous mark kinds in the exact visualization are known as combination charts. Combination charts can also display many levels of detail in a single view.

The below chart shows the number of passengers injured with respect to the years.



# Use Cases

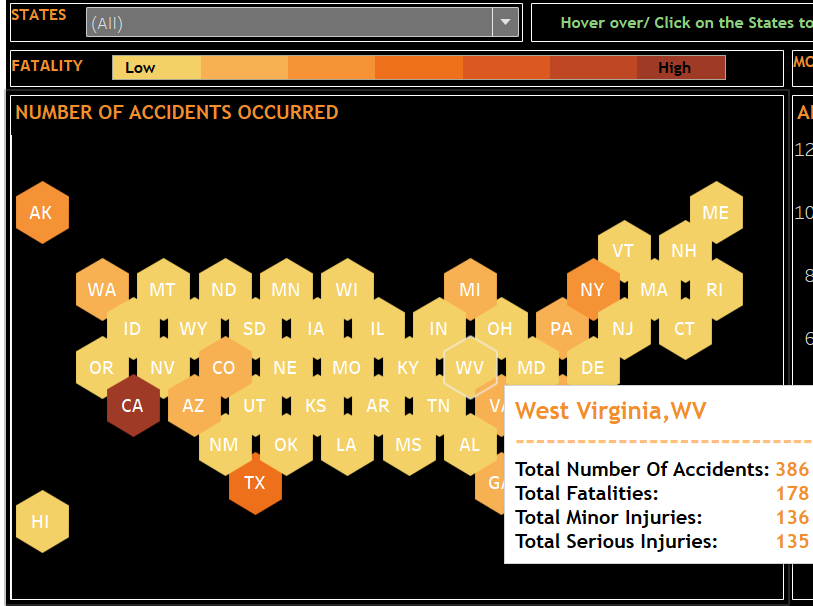
In this project, I have created three dashboards compiled into a story. The dashboards consist of many different kinds of graphs addressing the below cases:

1. Number of Accidents occurred in the USA
2. Injury count based on the aircraft models
3. Total number of injuries from 1982-2020
4. The exact location of air crashes detected in the cities.
5. Which phase was the aircraft in when the crash took place?
6. Top n aircraft carriers
7. Crashes detected with the aircraft damage
8. Number of passengers injured and crashes taken place over the years
9. The highest number of crashes by manufacturers
10. Number of crashes over the years
11. Statistics of aircraft Accidents
12. Fatality by aircraft category

**Use Case 1: Number of Accidents occurred in the USA**

The graph shows the total number of accidents that have taken place over the years in each state. The states have been displayed in the form of hexagons. The user can select the conditions from the state filter to indicate the injuries for that particular state. It also shows the count of fatalities, severe and minor injuries. The more intense the color, the more are the number of accidents occurred.

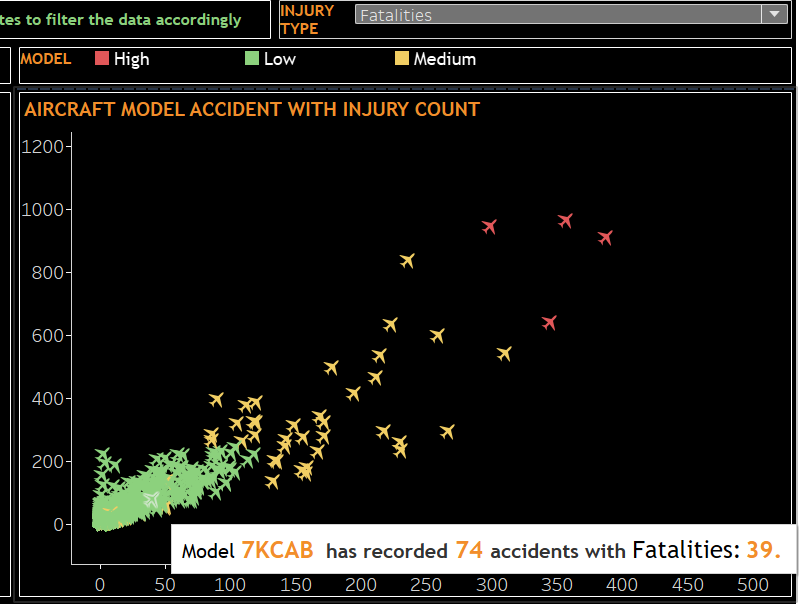
For example, this map shows that California has the highest number of accidents with more than 8k accidents and Vermont with minor accidents. So, also, In West Virginia (WV). Total accidents are 386, with 178 passengers dead, 136 passengers with minor injuries, and 135 passengers with serious injuries.



**Use Case 2: Injury count based on the aircraft models**

In the graph, the total injury count has been shown according to the aircraft model. Here, I have divided the charts into three segments, green as low accident count, high accident count as red, and medium accident count as yellow. Depending on the injury type, we can see the accident number and its corresponding aircraft model that had crashed. Also, depending on the states too, we can find the count of accidents with its injury type and model name.

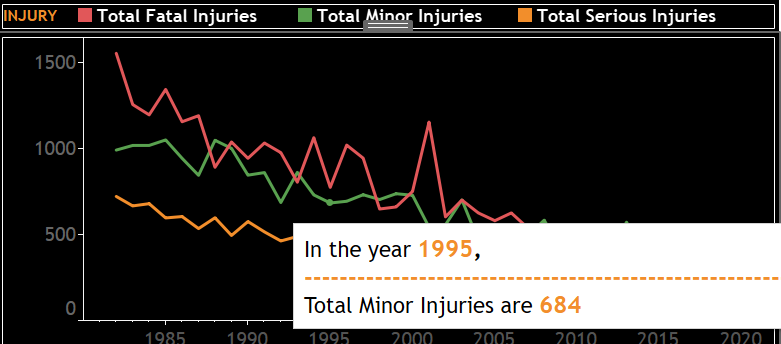
For example: With filter as fatality and for all states, we can see that aircraft Model 7KCAB has recorded a total of 74 accidents with 39 passenger fatalities.

****

**Use Case 3: Total number of injuries from 1982-2020**

The combination line graph below shows the total number of injuries over the years, Depending on the injury type. The red line shows the total fatal injuries, green shows the minor injuries, and orange is the serious injuries.

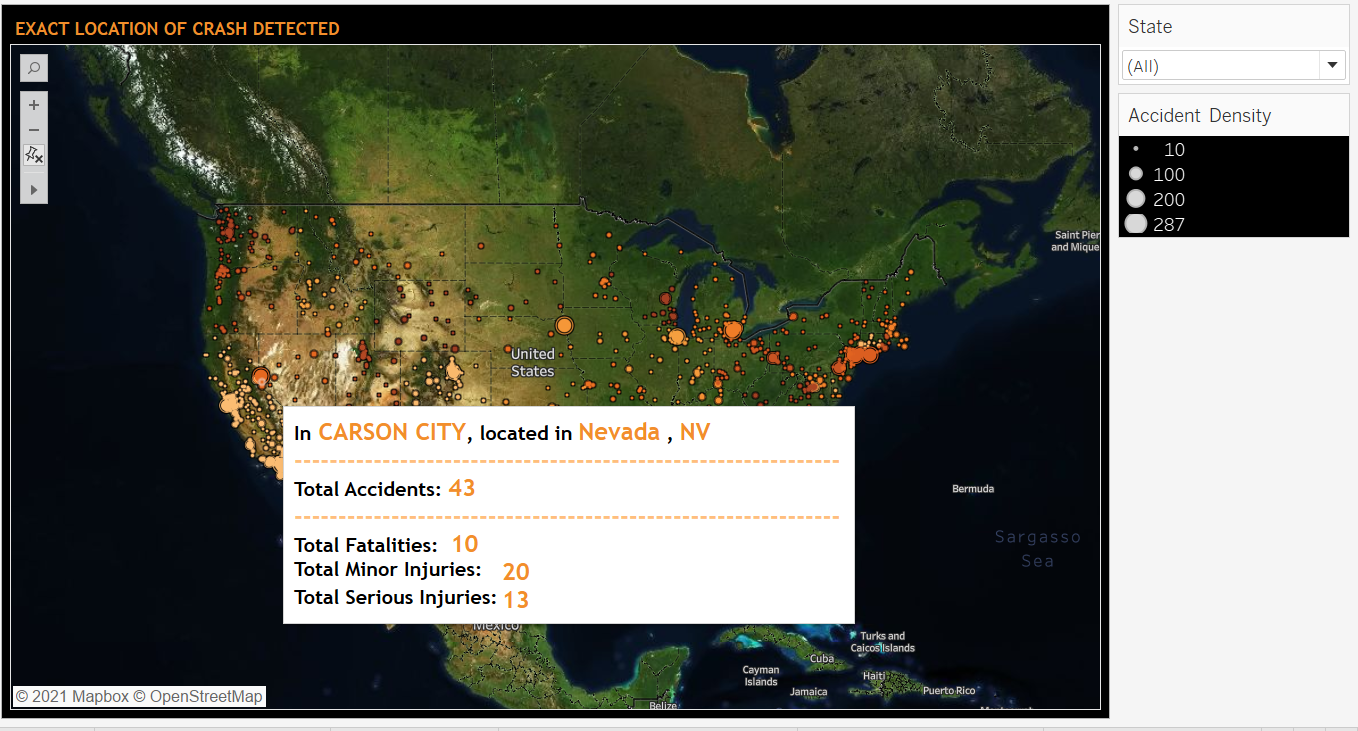
For example, in 1995, Total minor injuries were 684, fatal injuries were 775, and serious injuries were 449. We can also depict that the injury count has decreased over the years.



**Use Case 4: The exact location of air crashes detected in the cities.**

The satellite view of the US map shows the exact location of air crashes detected per city and state. It also shows the number of accidents and injury count for the selected states. The size of the circles depends on the total number of passengers injured.

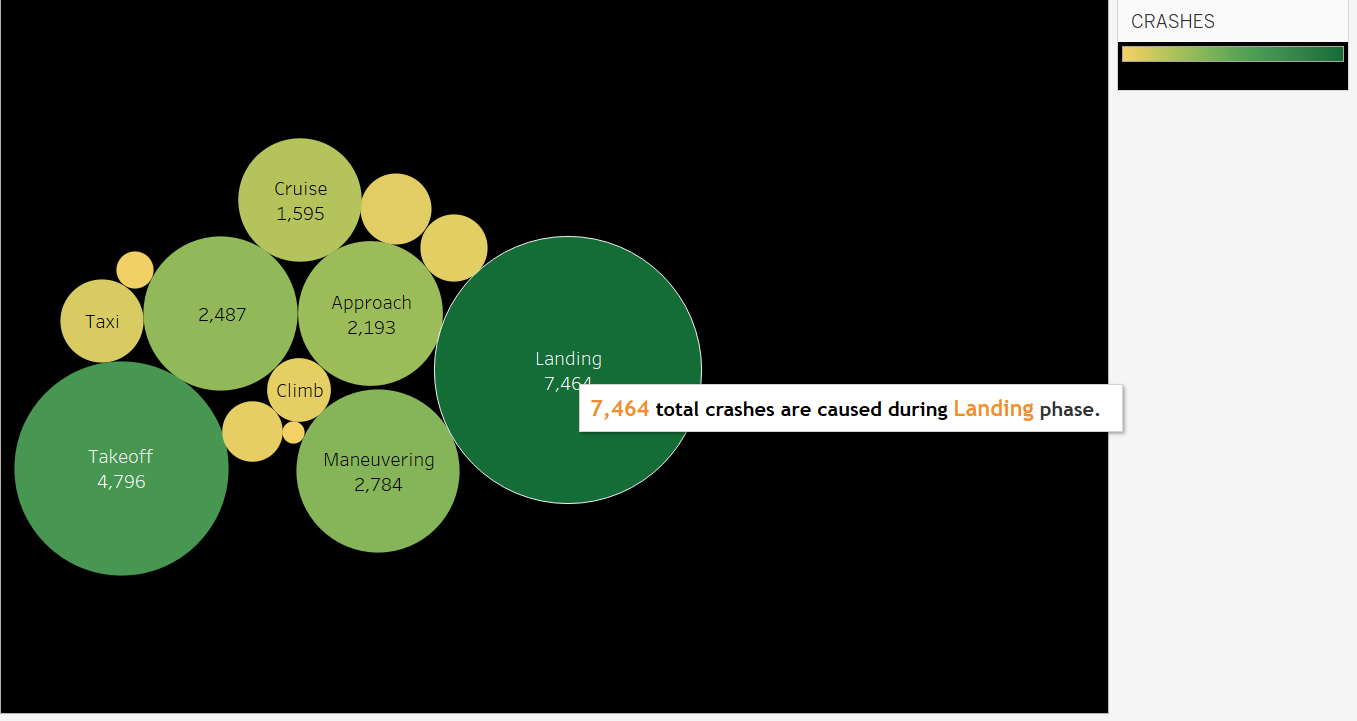
Example: In Carson City, located in Nevada, NV, the total number of accidents recorded is 43. Out of which ten passengers have died, 20 have experienced minor injuries, and 13 have serious injuries.



**Use Case 5: Which phase was the aircraft in when the crash took place?**

The bubble chart shows the distribution of total crashes as a function of the phase of the plane crash. The total number of crashes determines the size of the packed bubble; the higher the number, the larger the bubble and its color intensity, and vice versa.

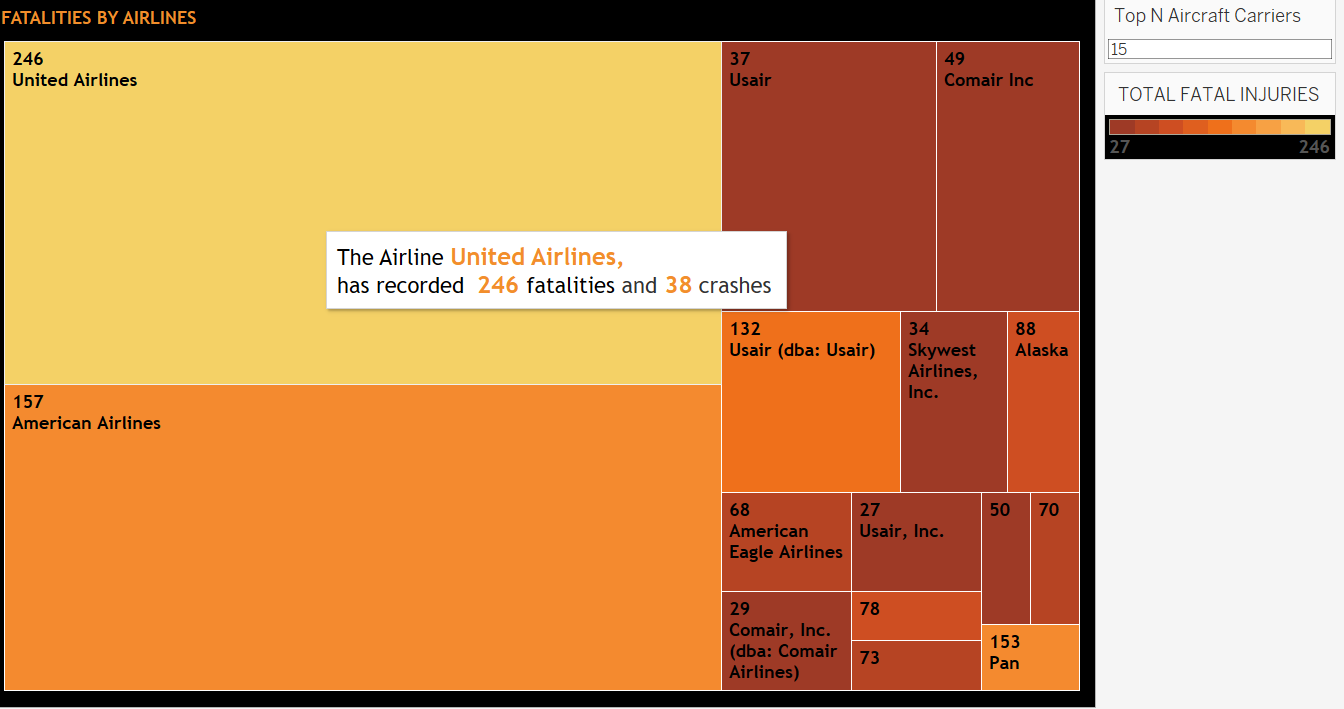
Example: Highest crashes have taken place in the landing phase.



**Use Case 6: Fatalities by Airlines**

The treemap depicts the top N aircraft carriers that have experienced fatalities and crashes. The intensity of the color depends on the fatality count. We can select the top N aircraft from the filters mentioned.

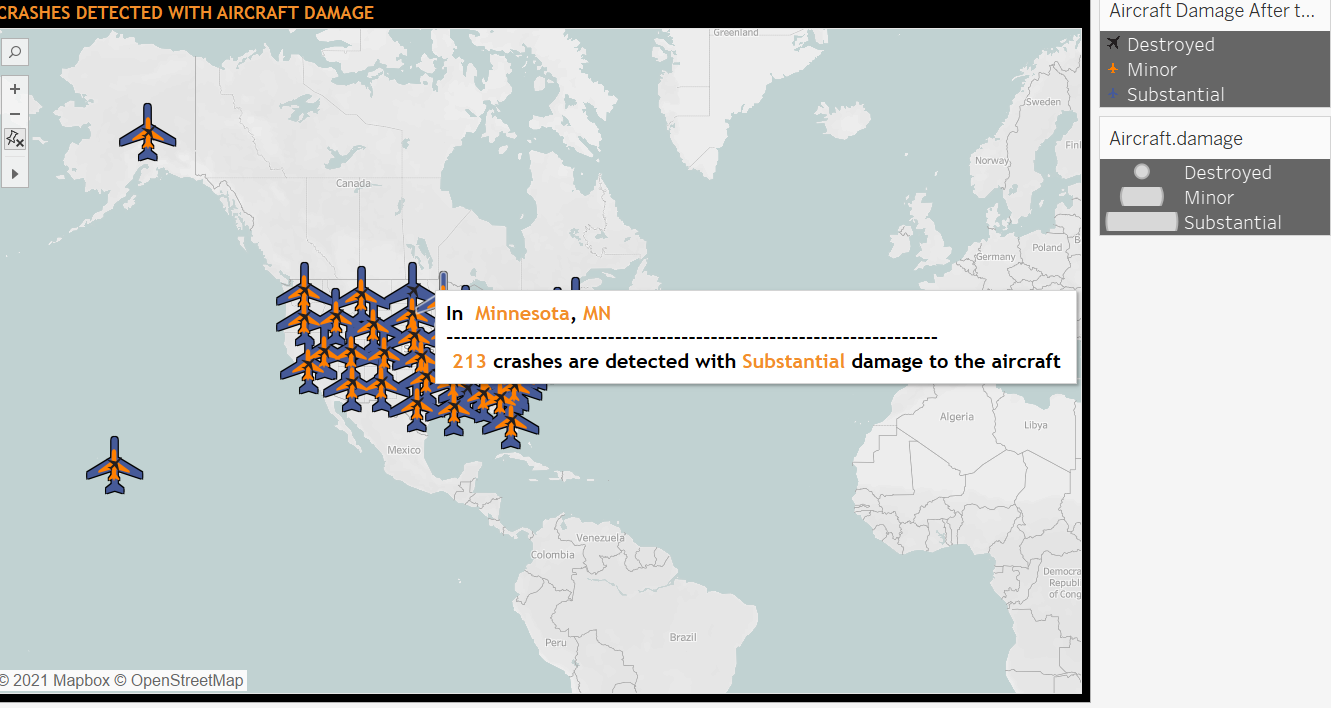
Example: The below graph shows the top 15 aircraft carriers with the highest fatalities. United Airlines has the highest fatality rate compared to others. It is almost 1.5 times the amount the American airlines.



**Use Case 7: Aircraft Damage after the crash**

The following map shows the intensity of damage the aircraft experienced in a particular state after the crash. The black airplane depicts the destroyed aircraft, orange as minor damage caused and blue with substantial damage.

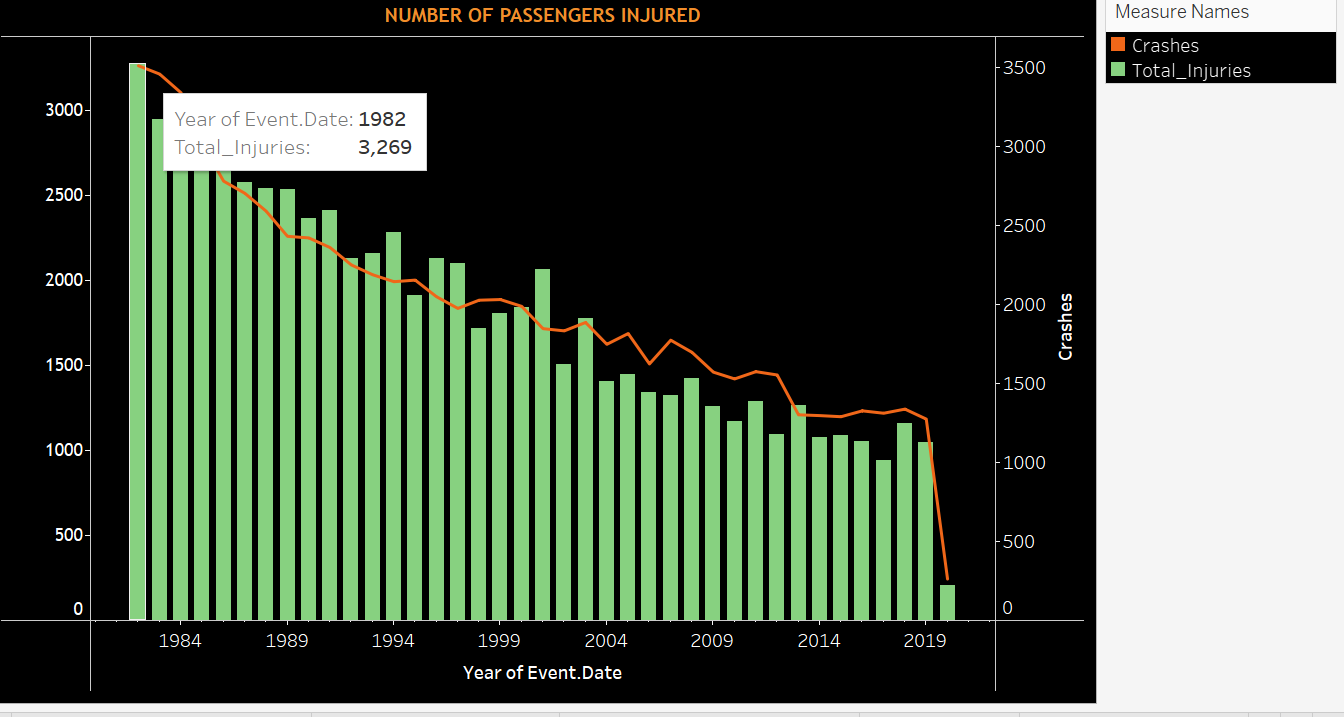
Example: Minnesota, MN has 213 crashes detected with substantial damage done to the plane after the crash and 06 collisions with minor damage, and 19 crashes where the aircraft is destroyed.



**Use Case 8:** Number of passengers injured and crashes taken place over the years

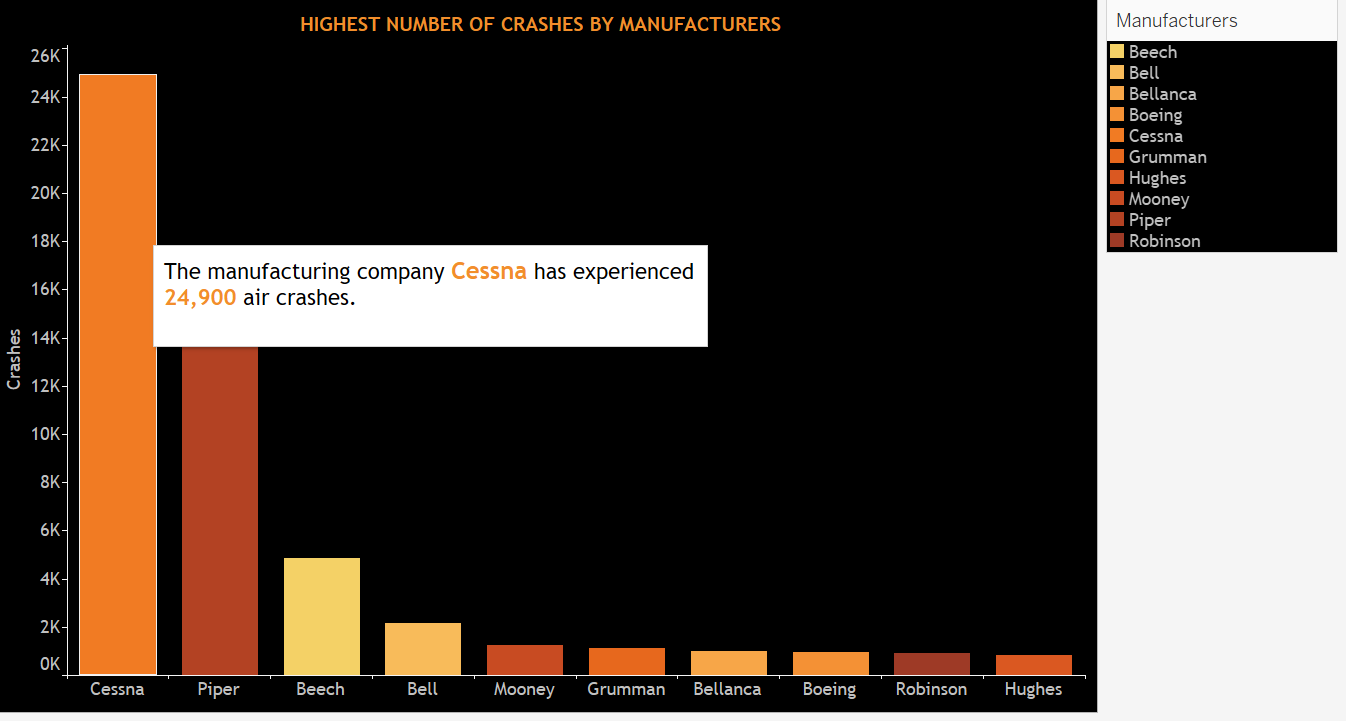
The time dual combination graph shows the total accidents and the crashes for the years. The orange line shows the number of crashes, and the green histograms show the total number of injuries.

Example: The number of airplane crashes has decreased as the year goes by. However, the max air crashes and passenger injuries were experienced in 1982.



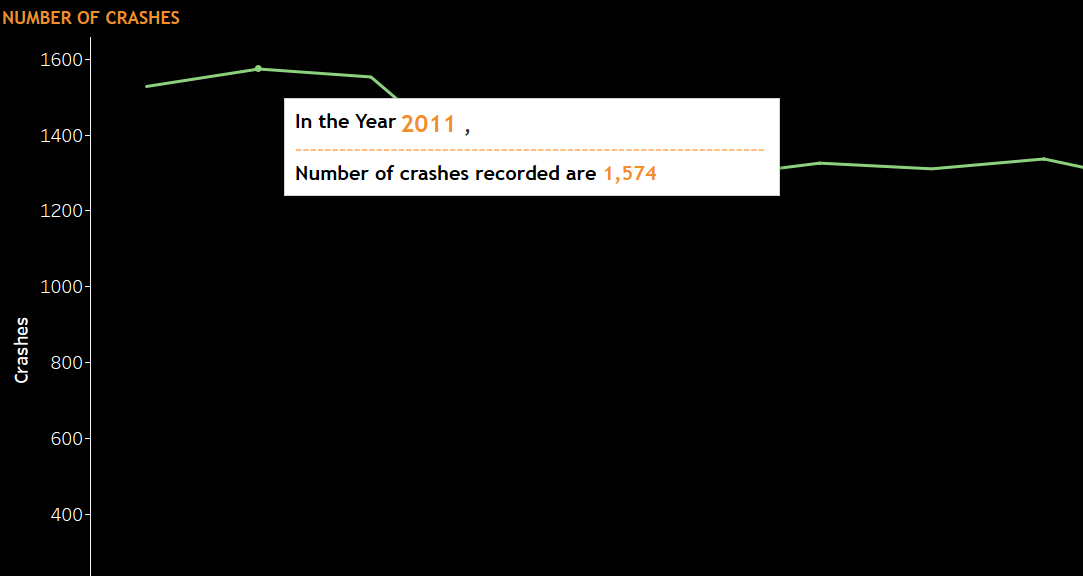
**Use Case 9:** The highest number of crashes by manufacturers

The Bar chart shows the top 10 manufacturing companies that have experienced maximum air crashes, with Cessna topping the list. The color intensity ranges from maximum crashes to minimum crashes



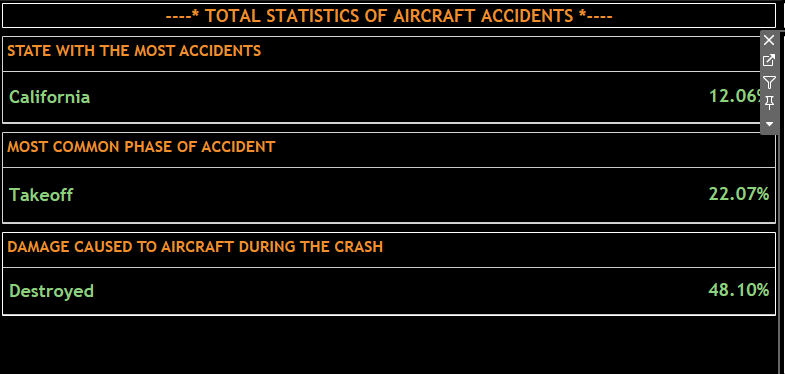
**Use Case 11: Number of crashes over the years**

The below line graph shows the number of crashes over **the years.** Maximum crashes were seen during 2011 and decreased over the years.



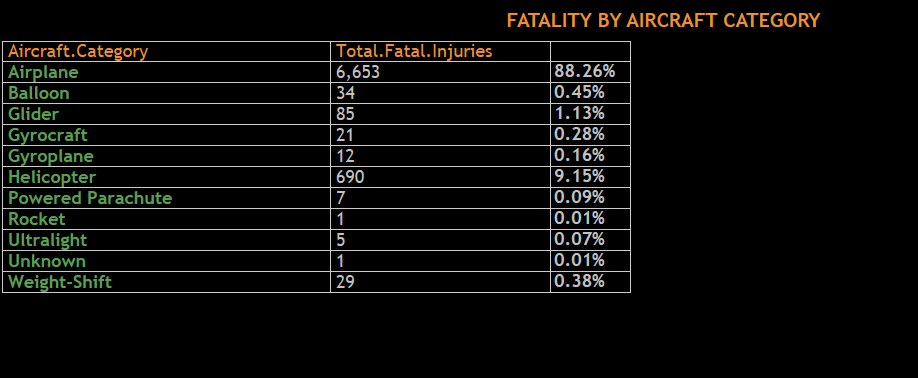
**Use case 11: Statistics of aircraft Accidents**

From the total statistics, CA has experienced the most accidents with 12.06%. In addition, 22.07% of accidents occurred during the takeoff phase of the aircraft, and 8.10% of the aircraft were destroyed after the crash.



**Use Case 12: Fatality by aircraft category**

The table shows the fatality rate as per the category with airplanes taking the lead with more than six thousand six hundred passengers dying, comprising 88.26%.

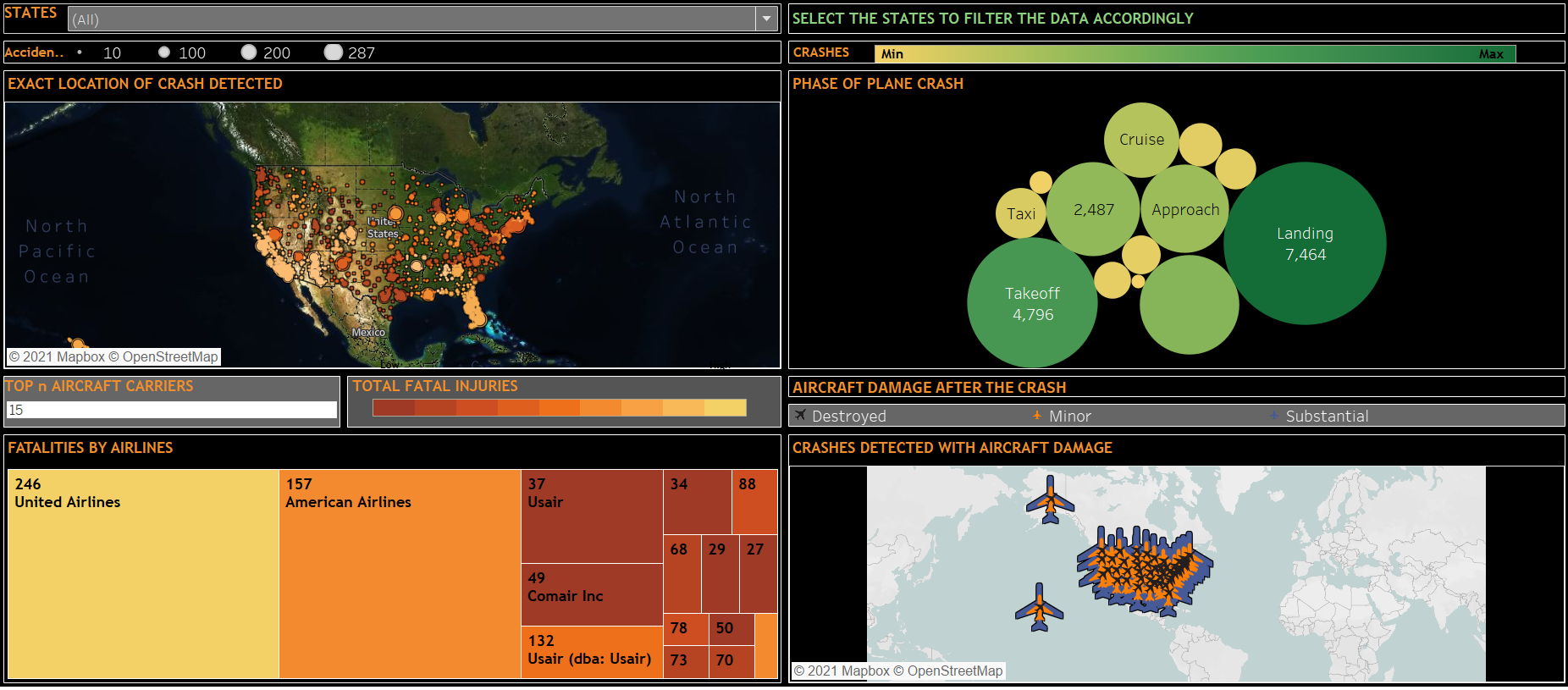


# Dashboards

1. Dashboard 1



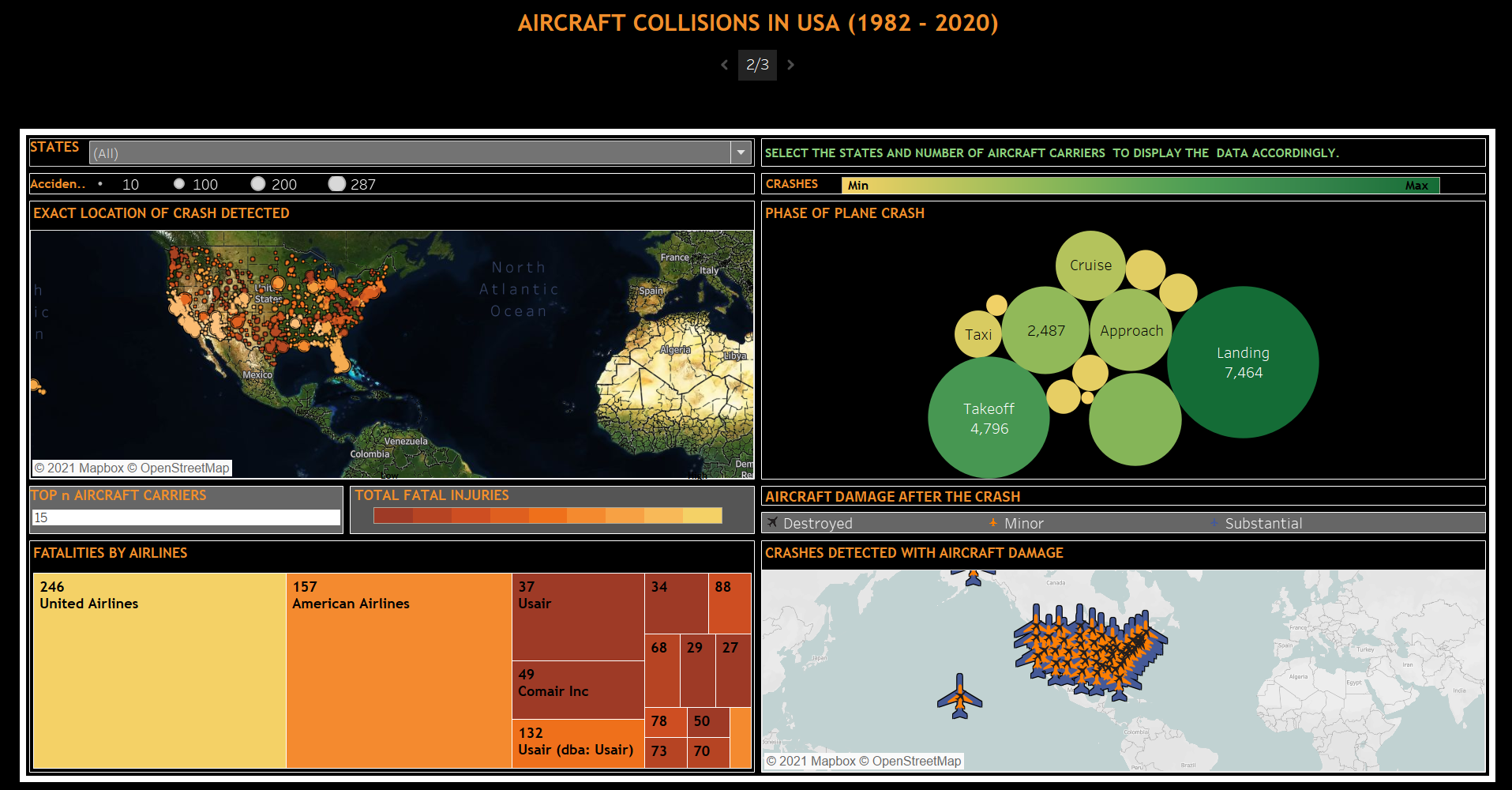
1. Dashboard 2

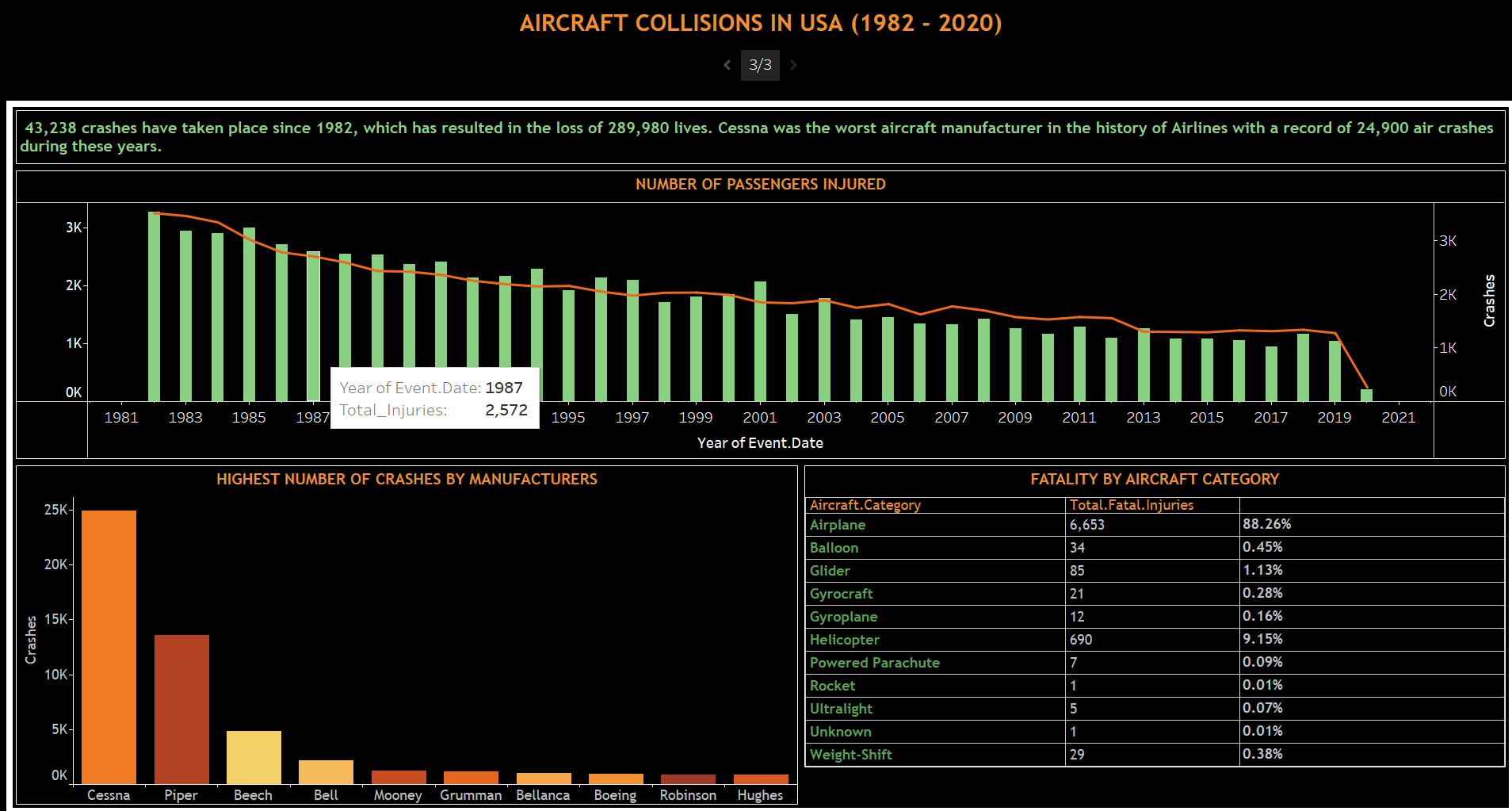


1. Dashboard 3



# Stories





# Conclusion

We can conclude from these stories and dashboards that 43,238 crashes have taken place since 1982, which has resulted in the loss of 289,980 lives. In addition, Cessna was the worst aircraft manufacturer in the history of Airlines, with a record of 24,900 air crashes during these years.

# Future Scope

In the future, I would like to build a machine learning model to detect and predict any anomalies in the flight journey by analyzing the flight condition, flight hours, the experience of the aircraft pilot, and weather conditions.

