

IE 7374 - Cloud Project: Aviation Accident Analysis

Group 10 - Subhasree Vemprala Sathyanarayanan and Stuti Dhebar

Problem Definition

Over the years, the aviation industry has witnessed tragic accidents, claiming hundreds of lives due to various errors. These are primarily caused by factors such as pilot error, weather conditions, mechanical failure, fuel mismanagement, and more. As a result, analyzing air accident data becomes crucial in addressing important questions like how safety measures can be improved, what technological advancements should airline companies consider while designing aircrafts, and which areas and situations can be categorized as high-risk. Most importantly, this analysis plays a pivotal role in maintaining public trust in air travel.

Objective

The aim of this project is to reveal insights into the main causes of air crashes and other aviation accidents, with a focus on understanding how these can be prevented in future. Here, we plan to perform analysis on various attributes within the data files by implementing a data pipeline in AWS. This pipeline will merge data files and transform all necessary rows and columns for final analysis. Some of the columns we will focus on are accident location, number of injuries sustained, weather conditions, and aircraft design details like type, make, model, among other relevant factors.

Some key analytical points we will explore are:

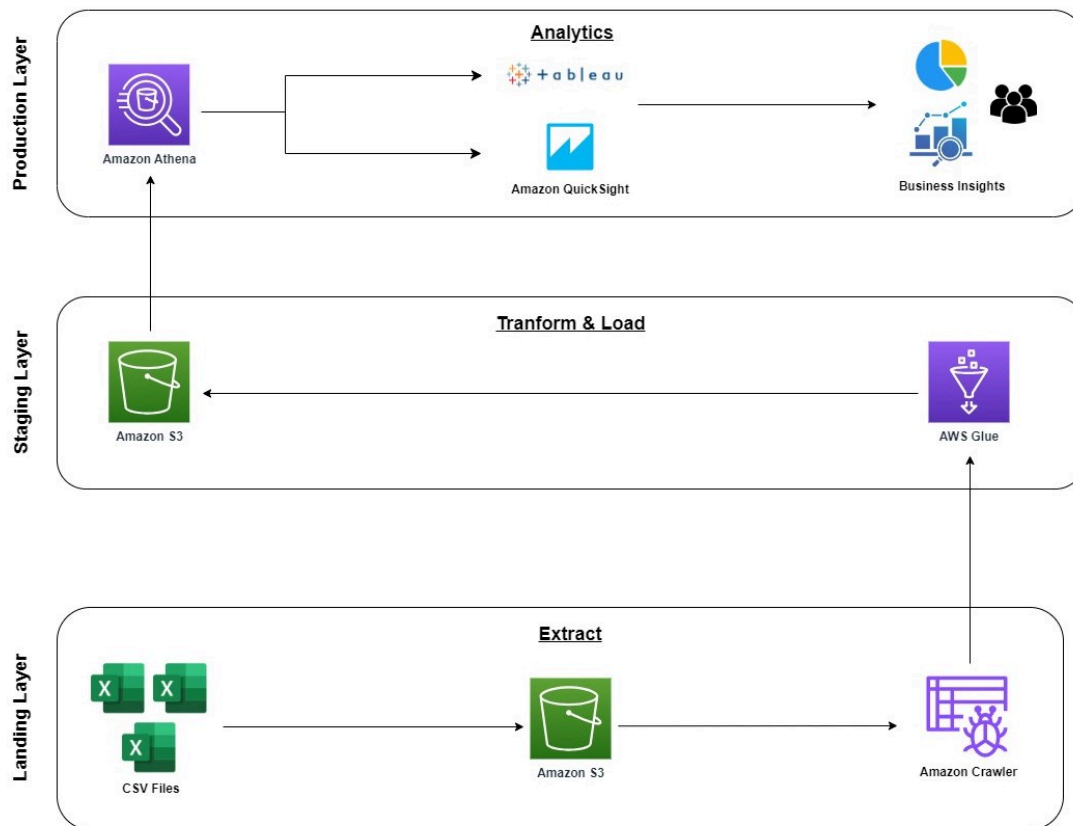
- What areas and weather conditions are more prone to airline accidents?
- Which type of accidents and incidents are common and why?
- What are common technical faults that lead to such tragedies?
- With advancements in aviation technology, have the number of accidents reduced over the years?

By diving into these attributes, we aim to discover patterns and identify correlations which have the potential to inform safety and mitigation strategies for similar incidents in the future.

Data Source

For this project, we will use Kaggle's [Aviation Accident Database](#). This dataset has more than 30 columns and ~80,000 records providing air accident information over the past 40 years. The attributes are both continuous and categorical in nature and since there are many missing values, we will be using different techniques to transform the data before performing analysis. Finally, the dataset will be divided into 2 or more files before loading in the pipeline.

Data Pipeline Architecture



For this project, we divided our data file of aviation accidents data into 3 files - accident investigation, accident location, and US states codes. A 3-layer pipeline was created to implement the ETL process using Amazon S3, AWS Crawler, AWS Glue, and Amazon Athena. Before implementing the pipeline, we performed data cleaning for a few columns in Python to impute null values. The most frequently occurring value (mode) was used for imputation. In some other cases implemented in the ETL Workflow, we

used 'Unknown' as enough details were not available to impute null values.

```
[ ] # Replacing NONE values with None
df.replace('NONE', 'None', inplace = True)

df['Weather_Condition'].replace('Unk', 'UNK')

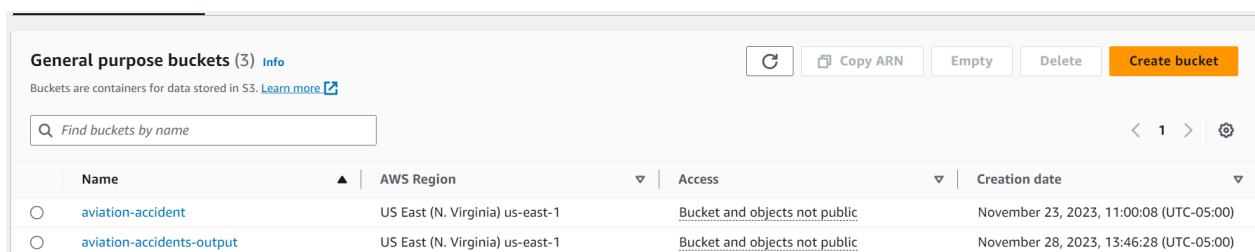
[ ] # Splitting Injury Severity to remove unnecessary details
df[['Injury_Severity', 'Num']] = df['Injury_Severity'].str.split(',', expand = True)
df.drop(['Num'], axis = 1, inplace = True)

# Imputing missing values in columns with the most frequently occurring values
df['Purpose_of_Flight'].value_counts()
df['Purpose_of_Flight'].fillna('Personal', inplace = True)
```

Fig 1. Sample of Python Data Cleaning Script

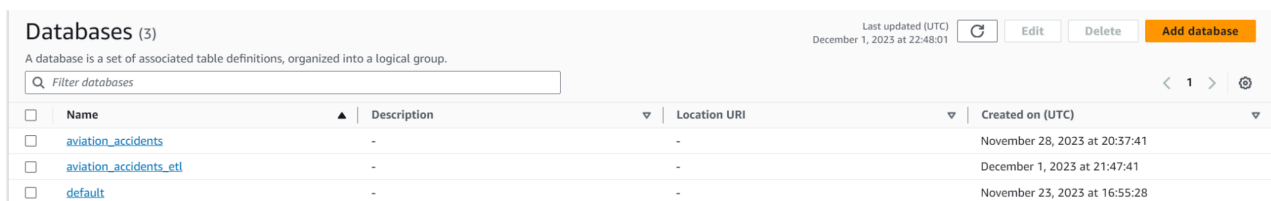
Landing Layer

In the first stage of the data pipeline, we created source and target S3 buckets to manage data prior to and post ETL, respectively. After creating S3 buckets, the CSV files were uploaded to specific folders within the source bucket. Next, to perform ETL and query the transformed data for analysis, we created source and target AWS Glue Databases using Amazon Athena. The data was loaded in the database with the help of AWS Crawlers which automatically detects the schema of the files and creates tables to store data in the database. The screenshots below show the steps described here.



	Name	AWS Region	Access	Creation date
<input type="radio"/>	aviation-accident	US East (N. Virginia) us-east-1	Bucket and objects not public	November 23, 2023, 11:00:08 (UTC-05:00)
<input type="radio"/>	aviation-accidents-output	US East (N. Virginia) us-east-1	Bucket and objects not public	November 28, 2023, 13:46:28 (UTC-05:00)

Fig 2. Creation of Source and Target S3 Buckets



	Name	Description	Location URI	Created on (UTC)
<input type="checkbox"/>	aviation_accidents	-	-	November 28, 2023 at 20:37:41
<input type="checkbox"/>	aviation_accidents_etl	-	-	December 1, 2023 at 21:47:41
<input type="checkbox"/>	default	-	-	November 23, 2023 at 16:55:28

Fig 3. Creation of Source and Target AWS Glue Databases

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (4) Info								
View and manage all available crawlers.					Last updated (UTC) December 1, 2023 at 22:41:37	Refresh	Action	Run Create crawler
<input type="text" value="Filter crawlers"/>								
<input type="checkbox"/>	Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last...	
<input type="checkbox"/>	accident-crawler	Ready		Succeeded	December 1, 2023 at 05:4...	View log	1 created	
<input type="checkbox"/>	accident_etl_crawler	Ready		Succeeded	December 1, 2023 at 21:4...	View log	3 created	
<input type="checkbox"/>	location-data-crawler	Ready		Succeeded	December 1, 2023 at 04:1...	View log	1 created	
<input type="checkbox"/>	statecode-load-crawler	Ready		Succeeded	December 1, 2023 at 03:4...	View log	1 created	

Fig 4. Creation of Crawlers to Load Data into Source Database

Amazon Athena > Query editor												
Editor Recent queries Saved queries Settings												
Workgroup: primary												
Data												
Query 5: SELECT * FROM "aviation_accidents"."accident_details";												
Data source: AwsDataCatalog Database: aviation_accidents												
SQL Ln 1, Col 55												
Run again Explain Cancel Clear Create												
Query results Query stats												
Completed Time in queue: 105 ms Run time: 1.462 sec Data scanned: 14.87 MB												
Results (88,889) Copy Download results												
<input type="text" value="Search rows"/>												
#	acc_id	investigation_type	accident_number	event_date	location	country	injury_severity	aircraft_damage	aircraft_category	make	model	amateur_built
1	1	Accident	SEAB7LA080	24-10-1948	"MOOSE CREEK	ID"	United States	Fatal	Destroyed	Unknown	Stinson	10B-3
2	2	Accident	LAX94LA336	19-07-1962	"BRIDGEPORT	CA"	United States	Fatal	Destroyed	Unknown	Piper	PA24-180
3	3	Accident	NYC07LA005	30-08-1974	"Saltville	VA"	United States	Fatal	Destroyed	Unknown	Cessna	172M
4	4	Accident	LAX96LA321	19-06-1977	"EUREKA	CA"	United States	Fatal	Destroyed	Unknown	Rockwell	112

Fig 5. Preview of Accident Details Table

Amazon Athena > Query editor												
Editor Recent queries Saved queries Settings												
Workgroup: primary												
Data												
Query 4: SELECT * FROM "aviation_accidents"."location";												
Data source: AwsDataCatalog Database: aviation_accidents												
SQL Ln 1, Col 46												
Run again Explain Cancel Clear Create												
Query results Query stats												
Completed Time in queue: 68 ms Run time: 712 ms Data scanned: 476.31 KB												
Results (12,211) Copy Download results												
<input type="text" value="Search rows"/>												
#	location_id	stateid	location	country	latitude	longitude	airport_code	airport_name				
1		(N) SKWENTNA	United States				LAKE CREEK					
2		100MI S KING SLM	United States									
3		11NM EAST OF SI	United States									
4		18NM ESE KETCHI	United States				NOTCH LAKE					

Fig 6. Preview of Accident Location Table

Query 6

```
SELECT * FROM 'aviation_accidents'.usstate_codes' limit 10;
```

Run again Explain Cancel Clear Create

Query results Query stats

Completed Time in queue: 122 ms Run time: 458 ms Data scanned: 1.06 KB

Results (10)

#	state_id	us_state	abbreviation
1	1	Alabama	AL
2	2	Alaska	AK
3	3	Arizona	AZ
4	4	Arkansas	AR

Fig 7. Preview of US States Code Table

Staging Layer

In the second stage of the pipeline, ETL transformations were performed on the data. Three jobs were created to transform the three tables. For the first table i.e accident investigation details, the data was loaded from AWS Glue Data Catalog and the following transformations were performed:

- Split Transform for splitting date column into day, month, and year arrays to perform analysis at different levels of time granularity. These arrays were converted using Array to Columns Transform.
- Split Transform for location column to store the exact location (eg. city) of accident and the state of accident in different arrays. Again, the Array to Columns Transform was used.
- Derived Column Transform was used to create a new column that stores a bucket value for injuries in each accident based on conditions in existing column. For ex, '0-100 injuries'.
- Finally, schema for some columns was changed and final data was stored in target S3 bucket.

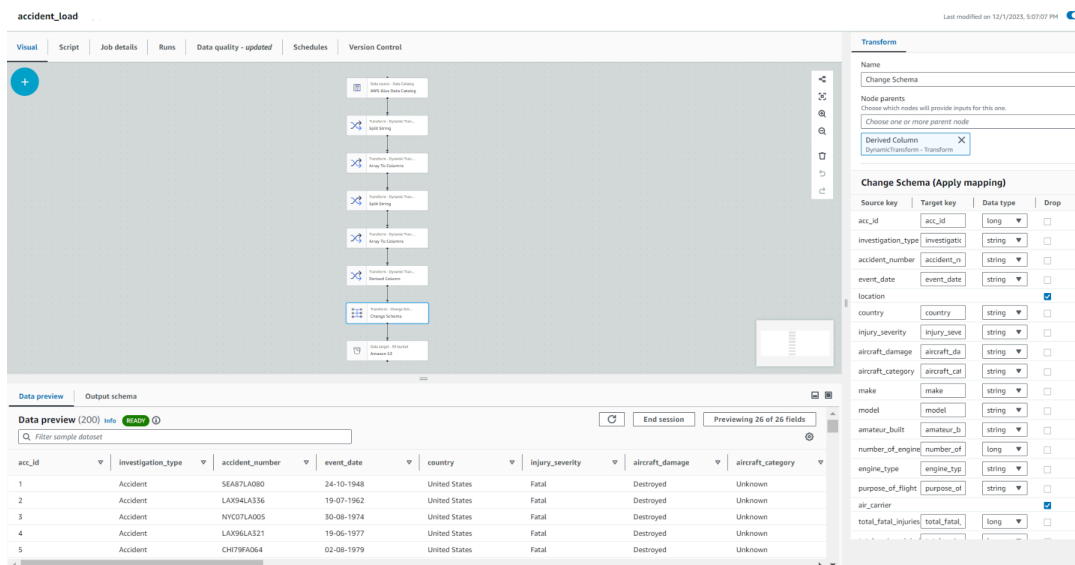


Fig 8. Accident Details - ETL Workflow

Next, for the accident location table, the following transformations were performed before loading data into target S3 bucket:

- Change of Schema to drop latitude and longitude columns
- Derived Column Transform to impute missing values with 'Unknown' for airport name and code columns.

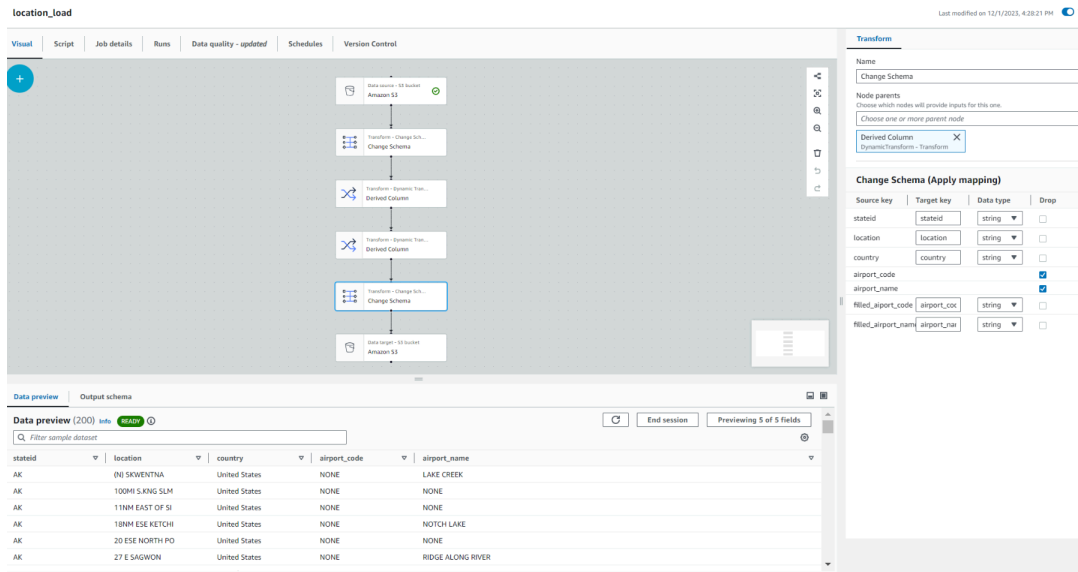


Fig 9. Accident Location - ETL Workflow

Finally, for the US state codes table, as it has very few columns and no transformations were needed, the data was directly loaded into target S3 bucket.

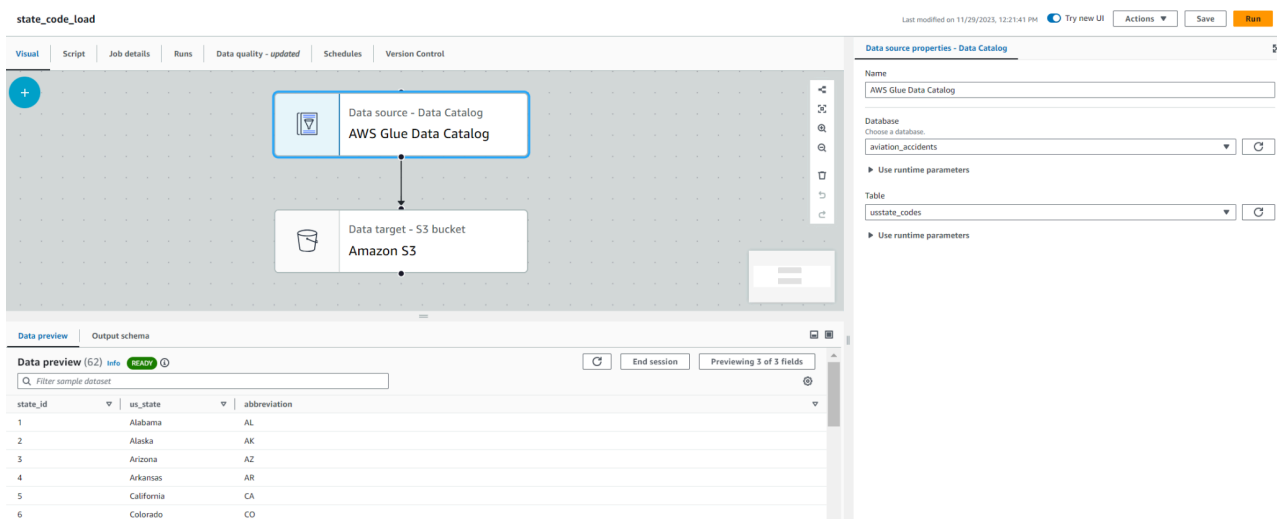


Fig 10. US States Code - ETL Workflow

Amazon Athena > Query editor

Editor Recent queries Saved queries Settings Workgroup: primary

Data

Data source: AwsDataCatalog Database: aviation_accidents_etl

Tables and views: Filter tables and views

Tables (3): accident_details, accident_location, state_codes

Views (0)

SQL: `SELECT * FROM "aviation_accidents_etl"."accident_details"`

Run again Explain Cancel Clear Create

Query results Query stats

Completed Time in queue: 66 ms Run time: 1.801 sec Data scanned: 15.20 MB

Results (88,889)

Search rows

#	acc_id	investigation_type	accident_number	event_date	country	injury_severity	aircraft_damage	aircraft_category	make	model	amateur_built	number_of_engines	engine_type
101	101	Accident	LAX82FUJ24	17-01-1982	"United States"	Non-Fatal	Destroyed	Airplane	Cessna	182P	No	1	Reciprocating
102	102	Accident	LAX82DA042	17-01-1982	"United States"	Non-Fatal	Destroyed	Airplane	Convair	440	No	2	Reciprocating
103	103	Accident	FTW82FRG21	17-01-1982	"United States"	Fatal	Destroyed	Airplane	Piper	PA-31	No	2	Reciprocating
104	104	Accident	DEN82FA020	17-01-1982	"United States"	Non-Fatal	Destroyed	Airplane	Piper	PA-31T	No	2	"Turboprop"

Fig 11. Preview of Accidents Table in Target S3

Production Layer

In this layer, the transformed data was used to perform analysis in Amazon Athena.

1) Countries with the highest number of accidents or incidents by year - This query gives an idea about the countries with the most number of aviation accidents. Also, note that our data is limited to a time frame and does not represent all time aviation accidents.

Editor Recent queries Saved queries Settings Workgroup: primary

Data

Data source: AwsDataCatalog Database: aviation_accidents_etl

Tables and views: Filter tables and views

Tables (3): accident_details, accident_location, state_codes

Views (0)

SQL: `SELECT country, e_year, COUNT(*) AS occurrences
FROM accident_details
GROUP BY country, e_year
order by e_year desc, occurrences desc;`

Run again Explain Cancel Clear Create

Query results Query stats

Completed Time in queue: 112 ms Run time: 1.074 sec Data scanned: 15.20 MB

Results (2,073)

Search rows

#	country	e_year	occurrences
1	"United States"	2022	1250
2	"American Samoa"	2022	35
3	Brazil	2022	31

2) Number of non-fatal and fatal injuries across different accident types - This query gives the total number of injuries that could possibly lead to death indicating the severity of accidents.

Amazon Athena > Query editor

Editor Recent queries Saved queries Settings Workgroup: primary

Data

Data source: AwsDataCatalog Database: aviation_accidents_etl

Tables and views

Filter tables and views

Tables (3): accident_details, accident_location, state_codes

Views (0)

SQL

```
1 SELECT investigation_type, injury_severity, COUNT(*) AS inj_sev
2 FROM accident_details
3 WHERE injury_severity IN ('Fatal', 'Non-Fatal')
4 GROUP BY investigation_type, injury_severity
5 ORDER BY investigation_type;
```

Run again Explain Cancel Clear Create

Query results Query stats

Completed Time in queue: 116 ms Run time: 839 ms Data scanned: 15.20 MB

Results (4)

Search rows

#	investigation_type	injury_severity	inj_sev
1	Accident	Non-Fatal	67212
2	Accident	Fatal	17790
3	Incident	Fatal	19
4	Incident	Non-Fatal	3852

3) Top 10 makes of aircrafts that were destroyed during aviation accidents - Analyzing the makes of aircrafts can be useful in identifying if a certain type of make/model is leading to such accidents over time.

Amazon Athena > Query editor

Editor Recent queries Saved queries Settings Workgroup: primary

Data

Data source: AwsDataCatalog Database: aviation_accidents_etl

Tables and views

Filter tables and views

Tables (3): accident_details, accident_location, state_codes

Views (0)

SQL

```
1 SELECT LOWER(make) AS make, COUNT(*) AS num_destroyed
2 FROM accident_details
3 WHERE aircraft_damage IN ('Destroyed')
4 GROUP BY LOWER(make)
5 ORDER BY num_destroyed DESC
6 LIMIT 10;
```

Run again Explain Cancel Clear Create

Query results Query stats

Completed Time in queue: 104 ms Run time: 701 ms Data scanned: 15.20 MB

Results (10)

Search rows

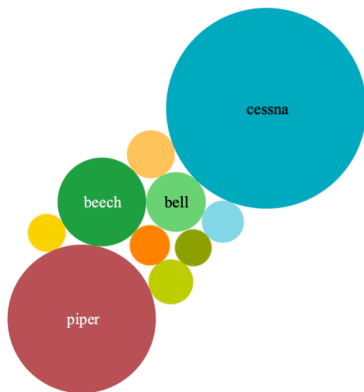
#	make	num_destroyed
1	cessna	5212
2	pipper	3428
3	beech	1585
4	bell	707
5	mooney	373
6	grumman	299
7	robinson	282
8	bellanca	233
9	hughes	189
10	boeing	169

Tableau Dashboard:

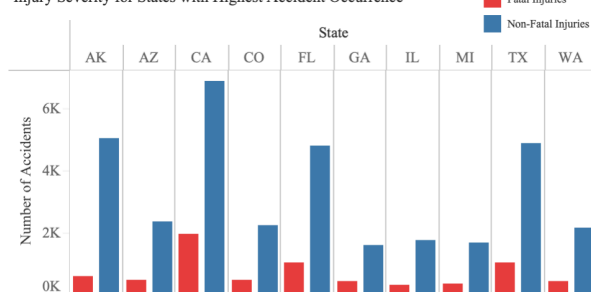
Aviation Accidents in the United States

Summary Statistics
Accidents : 82,248
Fatal Accidents : 15,024
Fatalities : 30,190

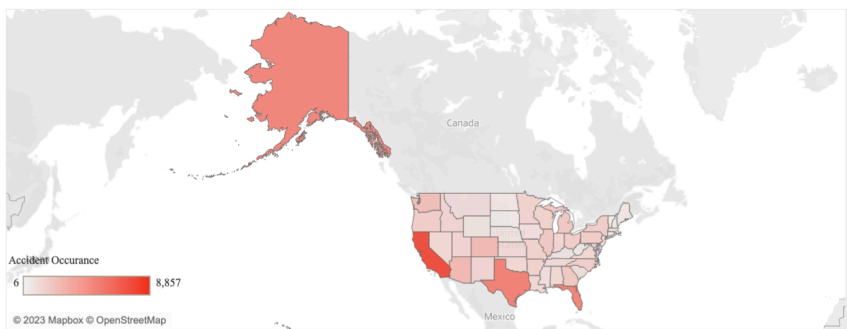
Aircraft Makes Most Frequently Involved in Accidents



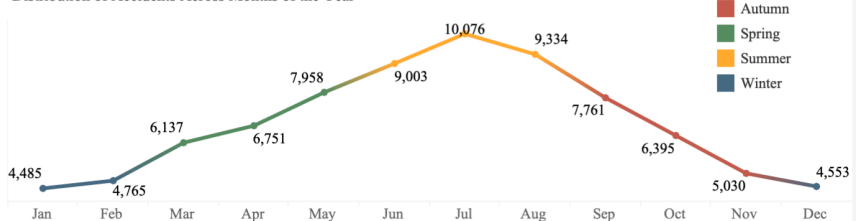
Injury Severity for States with Highest Accident Occurrence



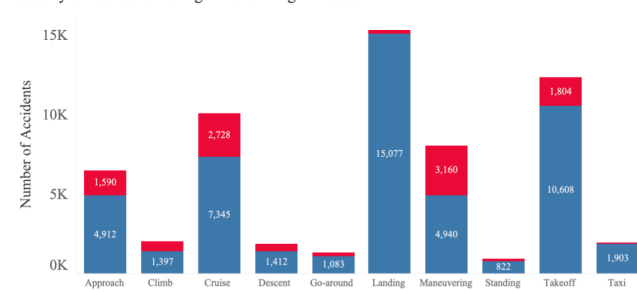
Accident Frequency by States in the United States



Distribution of Accidents Across Months of the Year



Severity of Accidents During Different Flight Phases



The Tableau dashboard on U.S. aviation accidents provides key insights into prevalent aircraft makes, accident frequencies by states, seasonal patterns, injury severity in high-incident states, and severity across flight phases. These insights could aid stakeholders in identifying potential safety improvements, directing attention to specific aircraft models, geographical areas, and times of the year prone to higher accident rates. Understanding injury severity by state informs emergency response planning, while insights into accidents during different flight phases contribute to targeted safety measures during critical stages of air travel. Overall, the dashboard helps equip decision-makers with actionable information to enhance aviation safety protocols and mitigate risks effectively.