**Airline and Accident Data Analysis**

**Group Name and Members:**

Group 2

Member Names

* Ningxin Kong
* Xulin Chen
* Huaiyuan Fan

**Executive summary (Business Overview)**

In the airline sector, large volumes of unprocessed data represent an untapped resource with the potential to enhance strategic decision-making. Our objective is to derive insights from airline data and accident data in terms of analysis on revenue, delay, and accident data. For revenue Analysis, we aim to investigate passenger booking patterns, and also to understand the revenue trends for different aircraft models, routes, and fare types. Insights derived can optimize decision-making processes for flight routes. For delay analysis, we focus on trends in flight performance, such as delays for different airports, operators, etc. Stakeholders can make more informed choices to enhance efficiency and reduce costs. For accident analysis, we focus on evaluating flight performance with respect to accident frequencies of different aircraft types, operators, location, routes, etc. The team aims to derive insights for safety enhancement and risk mitigation within the airline industry.

**Database & tools**

Python, MYSQL workbench, Tableau, Excel, Linux

**EDA**

The following 2 publicly available datasets will be used:

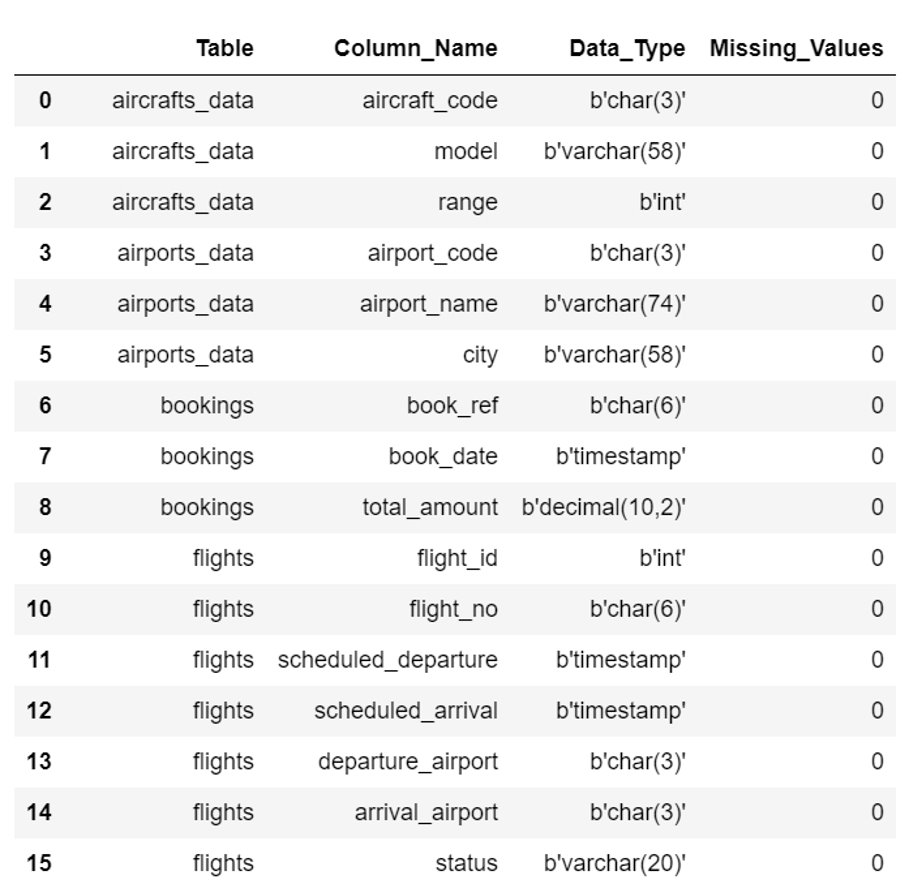
**1st Data Source: Airline Data**

* Data description: Airline operations and passenger data for analytics
* Data source: **(**<https://www.kaggle.com/datasets/mohammadkaiftahir/airline-dataset>**)**
* Dataset overview:

A screenshot of a computer

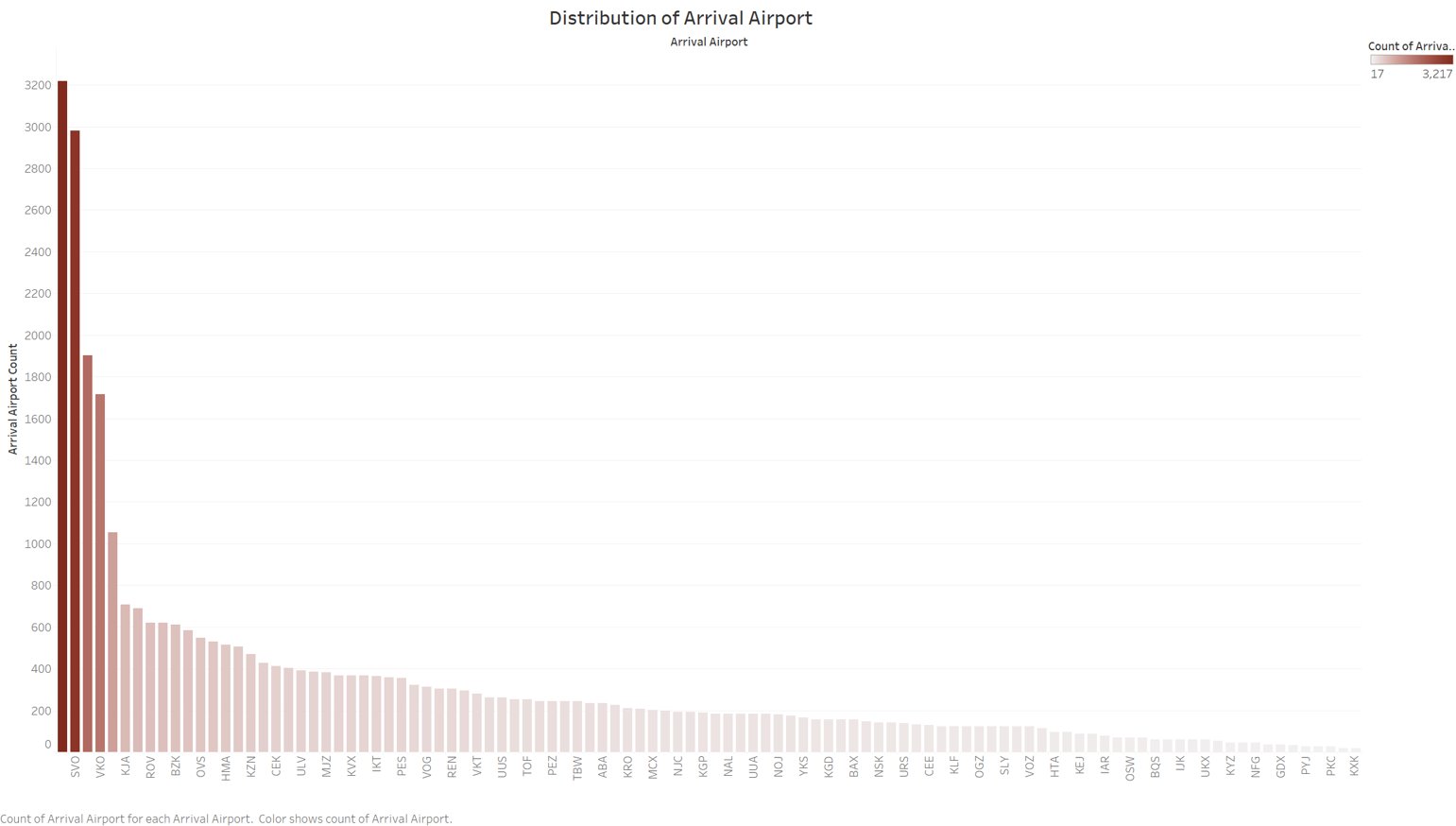
Description automatically generated

* Dataset key characteristics: data type & missing values

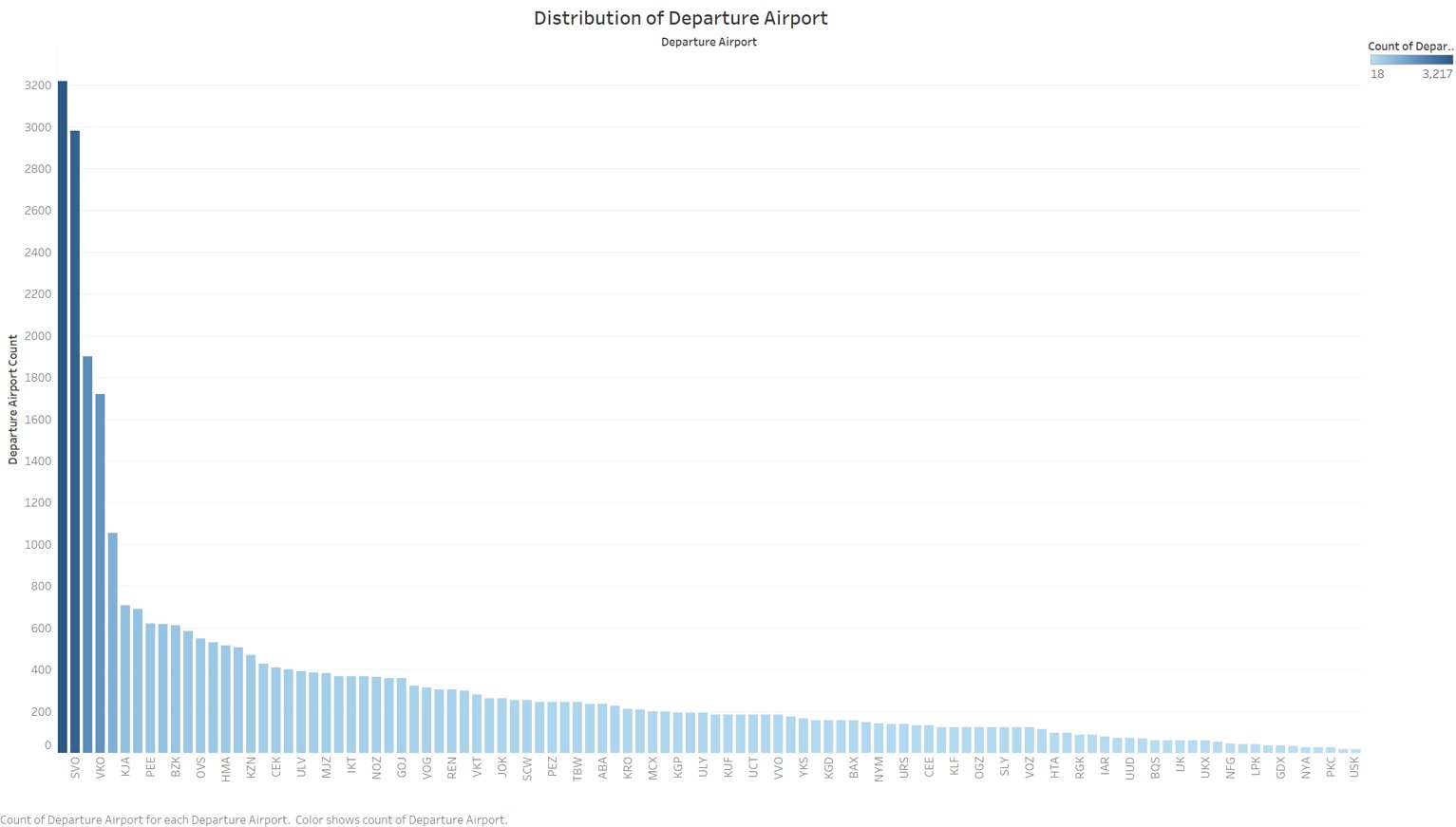




* Arrival airport distribution:
  + Extremely right skewed since there is a small number of very busy airports
  + Top 3 airports with the most arrival: DME (Moscow Domodedovo), SVO (Moscow Sheremetyevo Alexander S. Pushkin), LED (St. Petersburg Pulkovo Airport)



* Departure Airport Distribution
  + Extremely right-skewed distribution
  + DME and SVO (in Moscow) and LED (in St. Petersburg) have the most departures



2nd Data Source: Past Accidents Data

* Data description: Worldwide accidents and hijackings involving airliners (12+ passengers), corporate jets and military transport aircraft since 1919.
* Data source: (<https://www.kaggle.com/datasets/deepcontractor/aircraft-accidents-failures-hijacks-dataset>)
* Data Shape:
  + There are 23 Attributes and 23,519 Records in the Accident Dataset and Datatype for all columns besides Fatalities (Int) are strings

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

* Data Completeness: 6 attributes have substantial null values

A bar chart with text on it

Description automatically generated

* + Null values (%) of each attribute

A screenshot of a computer screen

Description automatically generated

* Data Cleaning & Key Fields
  + After removal of redundant attributes and attributes with substantial null values, Dataset contains 13 columns. 484 useful records remained to match with Airline Dataset records and analyze on commercial aviation records only.

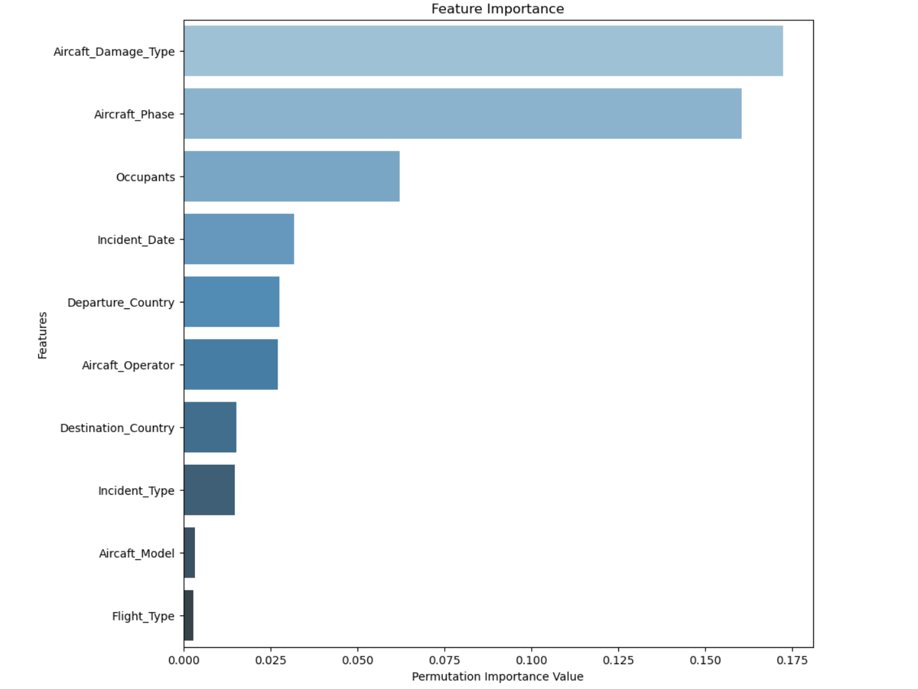
A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

* Attribute importance level
  + Feature Test indicates attributes that have high importance relation with accident fatalities: Aircraft Damage Type, Aircraft Phase, Occupants, Incident Date, Departure Country, Aircraft Operator



**Extraction, Transformation and Loading (ETL)**

**Airlines Dataset:**

* Extraction: The dataset is originally in SQLite database (109 MB). We converted it to an SQL dump (166 MB) using Linux commands. However, the file size was too large for MySQL Workbench, leading us to split it into 32 smaller SQL files.
* Transformation:
  + We noticed the SQL dump was compatible with PostgreSQL than MySQL. To resolve this, we manually edited the SQL dump in a text editor, adjusting data types, syntax, and constraints for MySQL compatibility.
  + As the dataset is about Russian airlines, it contains some Russian expression along with English expression. We removed Russian terms from the ‘city’ and ‘airport name’ columns in the ‘airports dataset’ and ‘model’ in the aircraft dataset, converting data types from JSON to VARCHAR.
  + We drop the table `seats` as it contains no primary key, and the content is nothing related to our analysis. We also omitted the coordinates and time zone columns in the ‘airports dataset’ to reduce redundancy.
* Load: The final step involved loading the 32 SQL files into MySQL Workbench.

Note: Please refer to ` Airline SQL Files to be loaded into MySQL ` folder in `Airline Dataset ETL process` folder for more information.

**Accident Dataset:**

* Extraction: The accident dataset was downloaded as a CSV and loaded into Python for transformation.
* Transformation:
  + Remove irrelevant columns from the aircraft accident datasets: Date (duplicate of incident date), Arit (duplicate of incident date), Aircraft Registration, Ground casualties (99% null), Collision casualties (99% null), Time (61% null), Aircraft Engine (55% null),"Incident\_Cause(es)", "Onboard\_Crew", "Onboard\_Passengers", "Aircaft\_First\_Flight"
  + Filter out only passenger flights from the incident dataset (since the airline dataset only contains passenger flights)
  + Mutate two new columns ‘Departure\_Country’ and ‘Destination\_ Country’ based on information from departure airport and arrival airport
  + Mutate a column to indicate whether the flight is domestic or international
  + The ‘Aircraft\_Model’ was reformatted to match the ‘model’ in the aircraft dataset (In airline dataset), filtering the data accordingly. This is the attribute that we aim to connect with the ‘aircaft\_dataset’ from the other data source
  + Extract occupant number from ‘Onboard\_Total’ and mutate a new column ‘Occupants’
  + Split the 'Incident\_Category' column into two new columns ‘Incident\_Type’ and 'Incident\_Damage'.
  + Drop 'Aircaft\_Nature', ‘Departure\_Airport’, ‘Destination\_Airport’, ‘Onboard\_Total’ and 'Incident\_Category' columns.
  + Removing all rows where any column contains 'Unknown' or NaN values
  + A total of 399 records with 13 columns.
  + Normalize the clean dataset to 11 separate datasets
* Loading: Finally, these datasets were loaded into MySQL Workbench.

Note: Please refer to `Incident Dataset Cleaning & Normalizing Process.ipynb` file in `Accident Dataset ETL process` folder for specific cleaning and normalization steps.

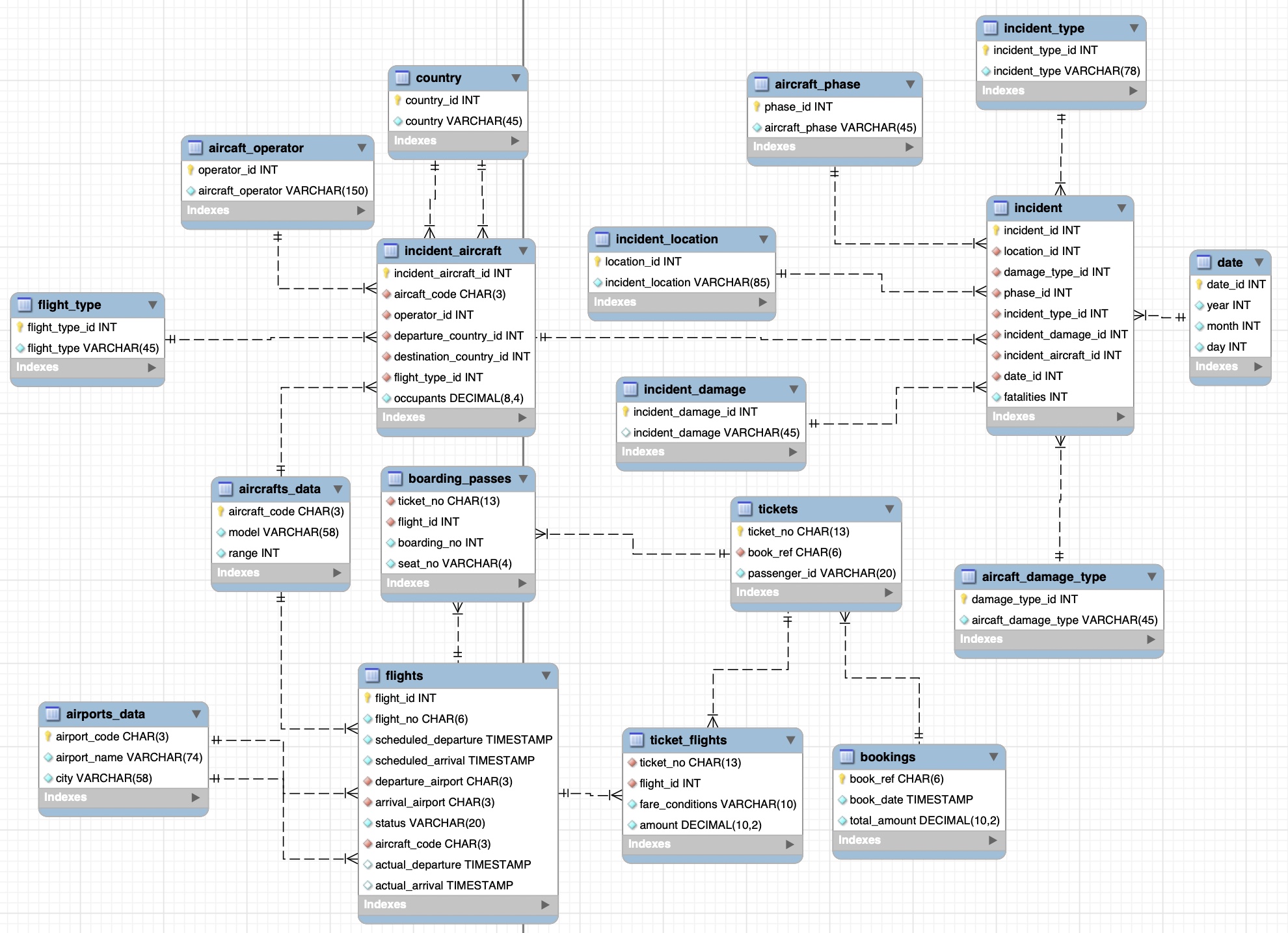
**Database Model**

**Design considerations**

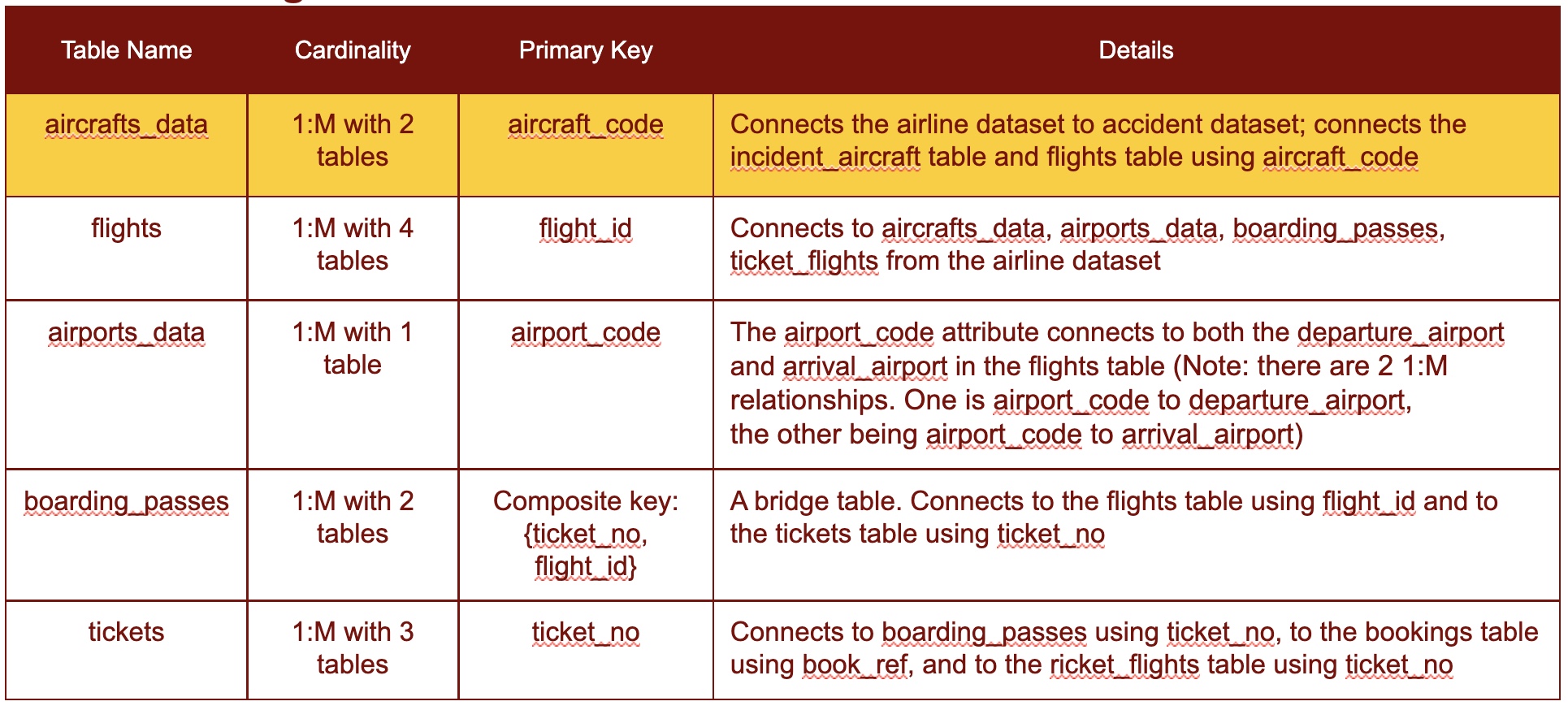
The team has decided to develop a relational database for our analysis, considering several key factors to inform this decision.

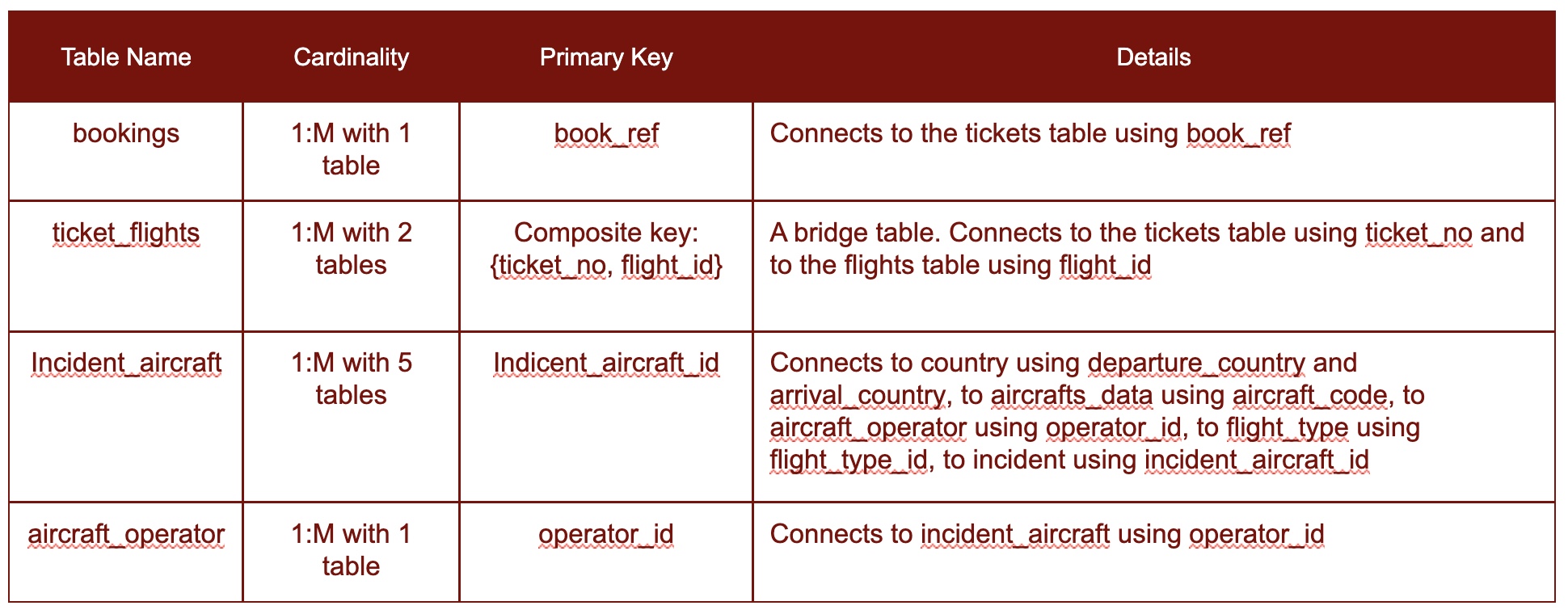
* Handling Complex Relationships: Our two data sources were originally designed for distinct purposes, resulting in complex relationships. We chose a relational database for its capability to manage these complexities effectively.
* Versatility for Broad Use: A relational database isn't limited to data analysis; it has a variety of applications. Its ability to efficiently process diverse data operations makes it an adaptable, comprehensive solution.
* Clarity and Ease of Interpretation: The structured layout of relational databases, with data neatly organized in tables, rows, and columns, enhances interpretability.
* Efficiency through Data Normalization: By obeying data normalization principles, our relational database minimizes redundancy, which further promotes both efficiency and integrity.

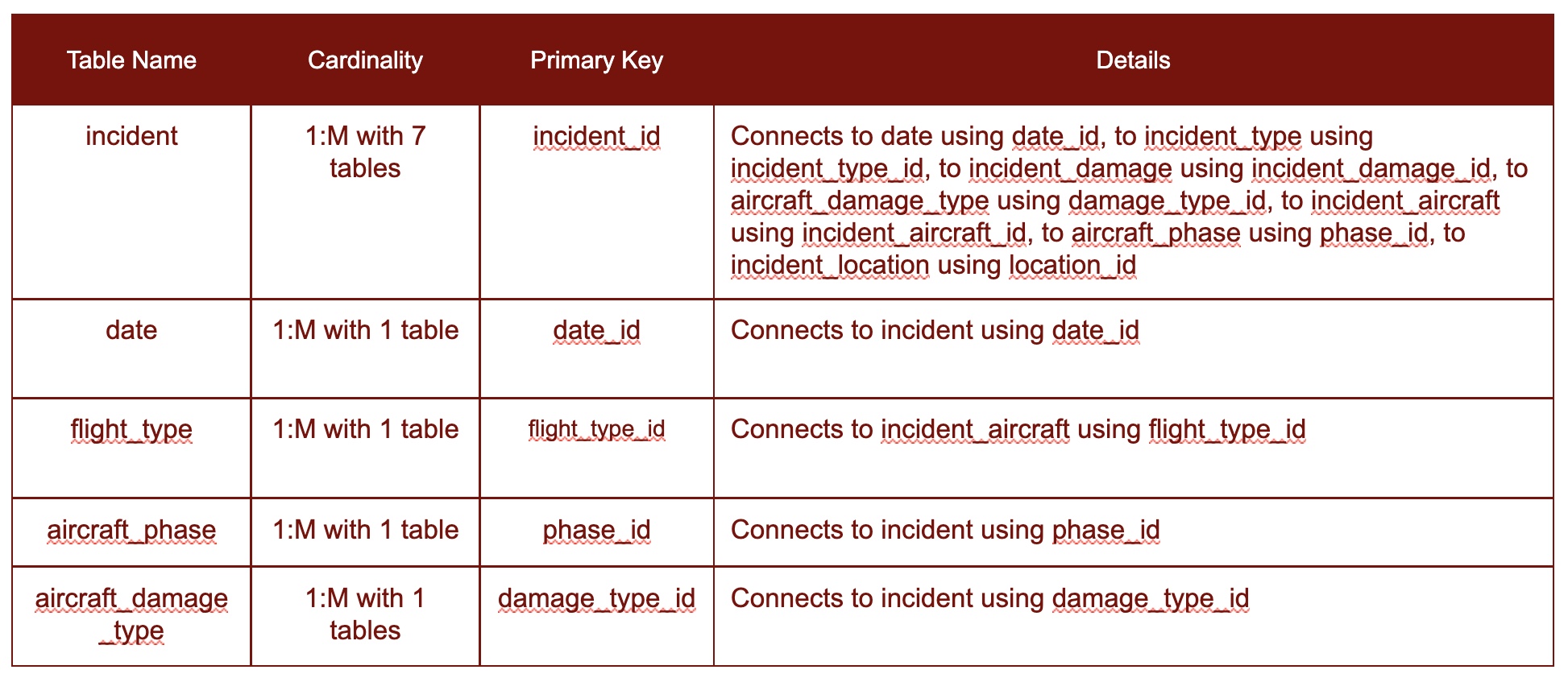
**Design Model**

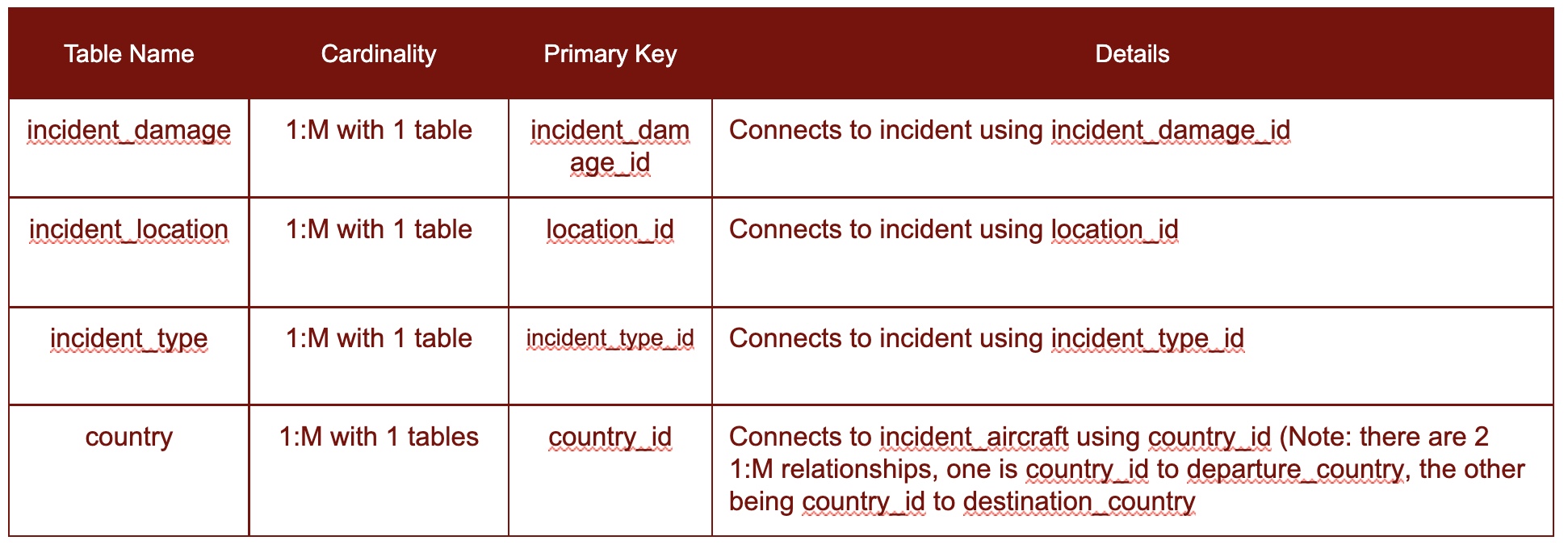


**Database Schema**



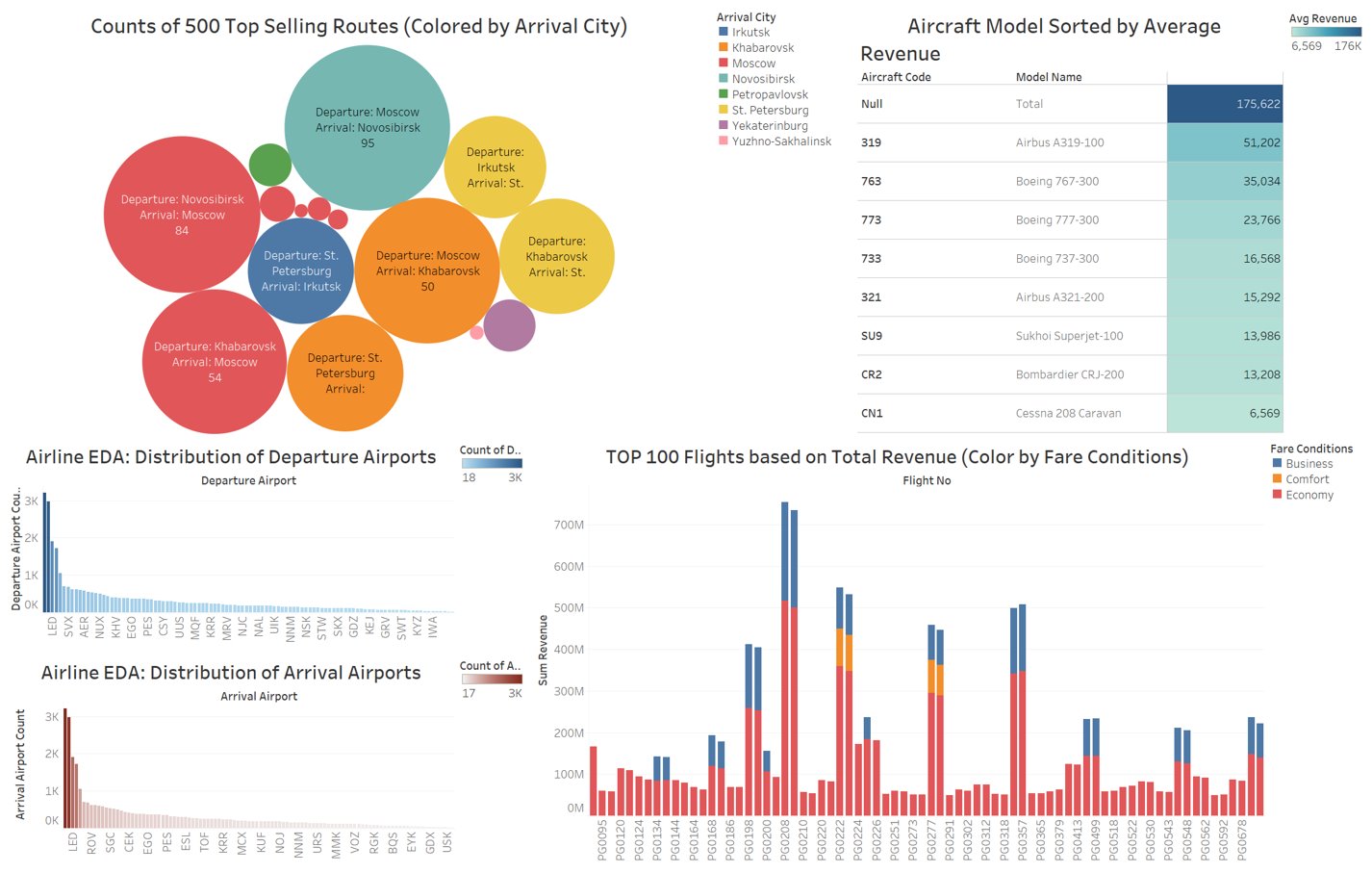






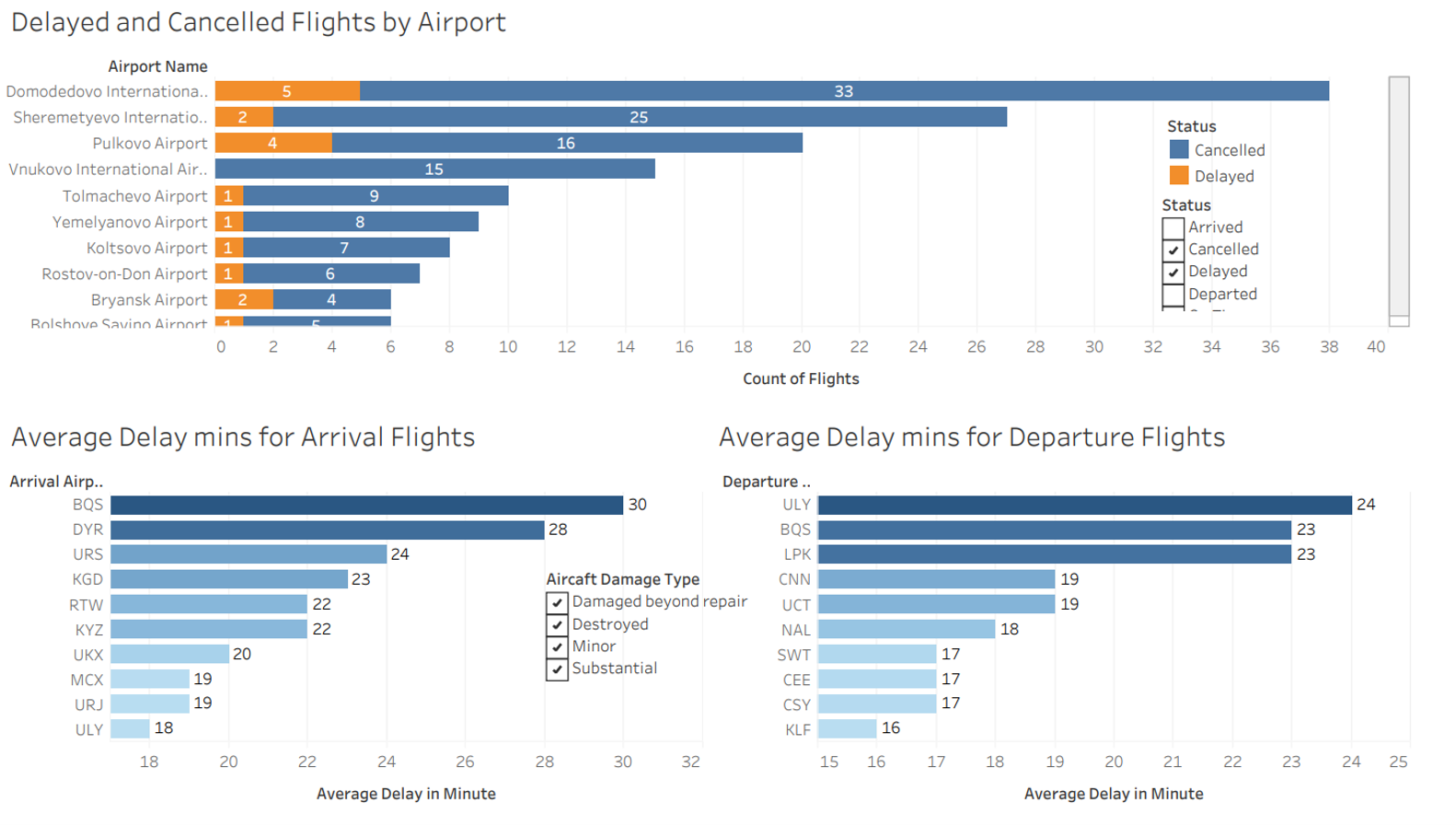
**Business use case**

**Business Use Case 1: Revenue Analysis**



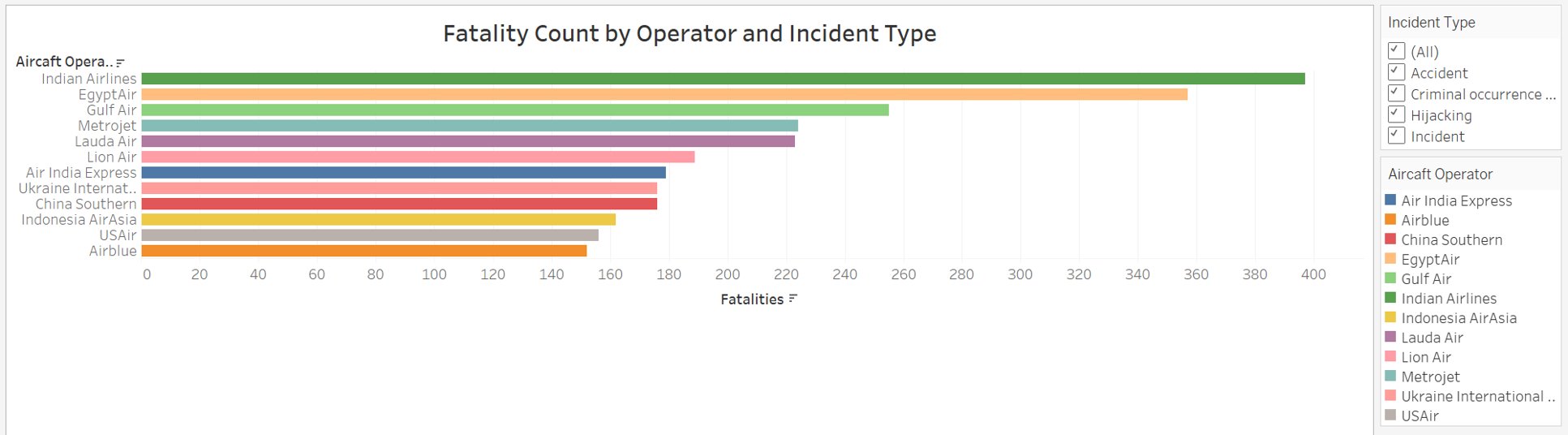
* The Moscow to Novosibirsk route stands out as our top-selling route, with a remarkable total of 95 bookings.
* Airbus A319-100, Boeing 767-300, and Boeing 777-300 are the top 3 models in terms of average revenue.
* Economy tickets make up the majority of total revenue, with flight PG0208.
* Flight PG0208 has the highest total revenue.

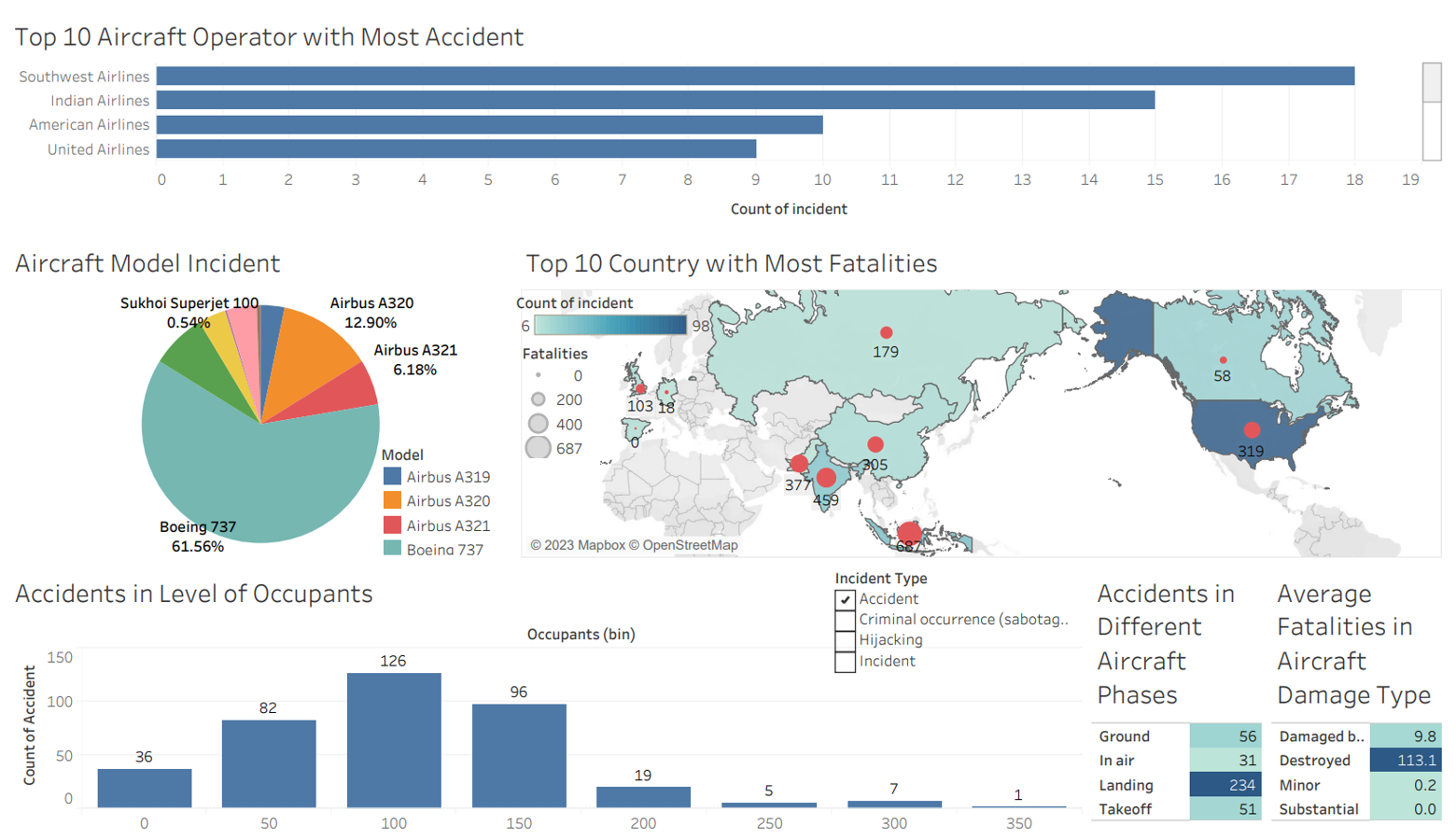
**Business Use Case 2: Delay Analysis**



* DME and SVO have a higher chance of delay or cancellation.

**Business Use Case 3: Accident Analysis**





* Insights: Most Aviation incident happened during Landing phase and Plane with smaller capacity tend to have higher incident rate
* Recommendation For Consumer: Consider that smaller planes have a higher incident rate especially during the landing phase. Check aviation safety record and try to choose larger planes if possible.
* Recommendation For Aviation Corporations: Focus on training programs for pilots, especially emphasizing the landing phase, continue monitor the accident record and implement thorough maintenance schedule.

**Conclusion**

Leveraging vast amounts of unprocessed airline data, our analysis aimed to enhance strategic decision-making across the industry. We delved into passenger booking patterns, revealing that the Moscow to Novosibirsk route is exceptionally popular, with 95 bookings. Our revenue analysis highlighted the Airbus A319, Boeing 767, and Boeing 777 as top revenue generators, primarily through economy class tickets, notably flight PG0208.

In examining flight delays, we identified higher incidences at DME and SVO airports. This information is crucial for improving operational efficiency and reducing costs. The safety analysis focused on the landing phase, where most aviation incidents occur, especially in smaller aircraft.

These insights are instrumental for stakeholders. For consumers, we recommend choosing larger aircraft for safer travel, considering their lower incident rates. Aviation corporations should prioritize pilot training for landings and maintain rigorous safety monitoring and maintenance schedules.

Overall, our approach to harnessing and analyzing airline data has yielded valuable insights for optimizing flight routes, enhancing operational efficiency, and improving safety measures within the airline industry.