WILDLIFE AIRSTRIKES IN THE UNITED STATES

REPORT SUMMARY

(Youtube Presentation Link)

DATASET ACQUISITION

The dataset was acquired from Kaggle (https://www.kaggle.com/datasets/dianaddx/aircraft-wildlife-strikes-1990-2023) and displays the data on airstrikes caused by wildlife in the USA.

ABOUT THE DATA

The original dataset contains multiple columns and records of airstrikes caused by wildlife from the year 1996 to 2023 and includes multiple damage sections for engines, propellers, etc. which were then further cleaned by using both Hive and Spark to get a final dataset that can be used for analysis purposes to derive meaningful insights.

DATA DICTIONARY

The dataset contains variables like Index ID, Airport, Airport Code, strike impact, etc. Here is the data dictionary:

- INDEX_NR: Individual record number
- INCIDENT_MONTH: Month strike occurred
- INCIDENT_YEAR: Year strike occurred (1996 to 2023)
- TIME_OF_DAY: Day or Night
- AIRPORT_ID: International Civil Aviation Organization airport code
- AIRPORT: Name of airport
- STATE: State where the airport is located
- FAAREGION: FAA Region where the airport is located
    - ASW – Southwest
    - ANM – Northwest Mountain
    - AEA – Eastern
    - ASO – Southern
    - AGL – Great Lakes
    - ACE – Central
    - ANE – New England
    - AWP – Western Pacific
    - AAL – Alaska
- OPERATOR: Airline Operator (includes private airlines, government, military and business aircraft)
- AMA: International Civil Aviation Organization code for Aircraft Make
- AMO: International Civil Aviation Organization code for Aircraft Model
- EMA: Engine Make Code
- EMO: Engine Model Code
- AC_CLASS: Type of aircraft
    - A: Airplane
    - B: Helicopter
    - C: Glider
    - D: Balloon
    - F: Dirigible
    - I: Gyroplane
    - J: Ultralight
    - Y: Other
- AC_MASS: Aircraft Mass
    - 1 = 2,250 kg or less
    - 2 = 2251-5700 kg
    - 3 = 5,701-27,000 kg
    - 4 = 27,001-272,000 kg
    - 5 = above 272,000 kg
- TYPE_ENG: Type of Engine Power
    - A = reciprocating engine (piston)

    - B = Turbojet
    - C = Turboprop
    - D = Turbofan
    - E = None (glider)
    - F = Turboshaft (helicopter)
    - Y = Other
- PHASE_OF_FLT: Phase of flight during which strike occurred
- HEIGHT: Feet of aircraft Above Ground Level during the strike
- SPEED: Knots (indicated airspeed)
- DISTANCE: Nautical miles from the airport
- SKY: Type of cloud cover, if any
- PRECIP: Precipitation
- DAMAGE: Level of damage
    - Blank: Unknown
    - N = None: No damage was reported.
    - M = Minor: When the aircraft can be rendered airworthy by simple repairs or replacements and an extensive inspection is not necessary.
    - M? = Undetermined level: The aircraft was damaged, but details as to the extent of the damage are lacking.
    - S = Substantial: When the aircraft incurs damage or structural failure which adversely affects the structure strength, performance, or flight characteristics of the aircraft
    - D = Destroyed: When the damage sustained makes it inadvisable to restore the aircraft to an airworthy condition.
- STR_RAD: Struck Radom? True or False
- DAM_RAD: Damaged Radom? True or False
- STR_WINDSHLD: Struck windshield? True or False
- DAM_WINDSHLD: Damaged windshield? True or False
- STR_NOSE: Struck nose? True or False
- DAM_NOSE: Damaged nose? True or False
- STR_ENG1: Struck Engine 1? True or False
- DAM_ENG1: Damaged Engine 1? True or False
- ING_ENG1: Ingested Engine 1? True or False
- STR_ENG2: Struck Engine 2? True or False
- DAM_ENG2: Damaged Engine 2? True or False
- ING_ENG2: Ingested Engine 2? True or False
- STR_ENG3: Struck Engine 3? True or False
- DAM_ENG3: Damaged Engine 3? True or False
- ING_ENG3: Ingested Engine 3? True or False
- STR_ENG4: Struck Engine 4? True or False
- DAM_ENG4: Damaged Engine 4? True or False
- ING_ENG4: Ingested Engine 4? True or False
- STR_PROP: Struck Propeller? True or False
- DAM_PROP: Damaged Propeller? True or False
- STR_WING_ROT: Struck Wing or Rotor? True or False
- DAM_WING_ROT: Damaged Wing or Rotor? True or False
- STR_FUSE: Struck Fuselage? True or False
- DAM_FUSE: Damaged Fuselage? True or False
- STR_LG: Struck Landing Gear? True or False
- DAM_LG: Damaged Landing Gear? True or False
- STR_TAIL: Struck Tail? True or False
- DAM_TAIL: Damaged Tail? True or False
- STR_LGHTS: Struck Lights? True or False
- DAM_LGHTS: Damaged Lights? True or False
- STR_OTHER: Struck other part? True or False
- DAM_OTHER: Damaged other part? True or False
- SPECIES_ID: International Civil Aviation Organization code for the type of bird or other wildlife
- SPECIES: Common name for bird or other wildlife
- REMARKS: Noticeable remarks made by cabin crew or officials
- REMAINS_COLLECTED: Indicates if bird or wildlife remains were found and collected
- REMAINS_SENT: Indicates if remains were sent to the Smithsonian Institution for identification
- SIZE: The size of the bird as reported by the pilot is a relative scale.
- WARNED: Pilot warned of birds/wildlife
- SOURCE: Type of report

Refer to Data Dictionary Excel Workbook for detailed steps performed Data Dictionary

DATA CLEANING

The initial level of data cleaning was done using Hive using an Apache Hadoop cluster on the Google Cloud platform. After the initial level of cleaning that included dropping columns not fit for analysis or missing value records, the output file (combined_output.csv) was then exported to Google Cloud storage from where it was imported to Spark in Jupyter Notebook (renamed to strike_report_clean.csv).

The imported data in Jupyter Notebook Spark was then treated for missing and NULL values by using different techniques and then passed on to perform Analysis and to create visualizations.

- *Refer to Hive SSH Notebook for detailed steps performed under Hive*

DATA ANALYSIS AND VISUALISATION

Data Analysis is an important part of exploratory data analysis, performed under Jupyter Notebook using Spark with its libraries that included pandas, and matplot for the extensive approach to perform Exploratory Data Analysis and obtain the results used for extracting meaningful insights.

- *Refer to Aircraft Strikes Jupyter Notebook for detailed steps performed under data Analysis and visualizations*

FINER QUESTIONS

FINER stands for Focused, Interpretative, Narrow, Exploratory, and Reproducible, and it represents a set of criteria for formulating well-defined research questions.

We have developed 8 FINER Research questions that have been answered based on our analysis and visualizations performed in Jupyter Notebook.

Question: What is the damage distribution of aircraft based on the damage taken by different segments of the aircraft?

Interpretation: By visualizing damage distribution based on segments, we can interpret that Body Damage (42.5%) is the most common phenomenon during an airstrike caused by wildlife.

Question: Which years have seen the highest number of airstrikes?

Interpretation: By plotting a time series graph, we can analyze that the year 2019 and year 2022 has seen the highest airstrikes, also a downfall in the year 2020 was caused due to COVID lockdown protocols.

Question: Which aircraft class has incurred the most damage?

Interpretation: A graph plot between the number of incidents and Aircraft class shows that Class A aircraft i.e. Passenger Airplane has recorded the highest damage.

Question: Which species has caused the highest number of strikes and what size?

Interpretation: While the species of the birds remain unknown, the size of the bird that has caused the most airstrikes was a small-sized bird.

Question: Which FAA region has been the audience of most strikes?

Interpretation: Plotting a graph for FAA regions along with the number of incidents that occurred, ASO (Southern Airport Region) was the audience of the most airstrikes.

Further research and analysis on this can show what bird species are available in this region and upon comparing it with the months of highest airstrikes. We can find out more insights if high airstrikes were during the time of the bird migration period.

Question: Which month in the year 2022 caused the maximum strikes and in which FAA region?

Interpretation: A grouped stacked column graph legend interprets that August was the month with the maximum number of airstrikes in the Southern FAA region.

Based on research results, the Southerly Bird Migrations start in August and go on till October, hence the highest number of strikes were observed in August.

Question: Which airport observed the most number of airstrikes?

Interpretation: A graph between Airports (top 15) and the number of strikes DENVER International Airport observed the maximum number of airstrikes overall.

DENVER International Airport belongs to the Northwest Mountain Region (ANM)

Question: Damage analysis to the aircraft based on the aircraft mass.

Interpretation: The box plot helps to compare the distribution of damage levels across different aircraft mass categories. It provides insights into the central tendency, spread, and potential outliers in the damage level distribution for each mass category.

Damage level comparisons:

- The damage level D has the lowest median aircraft mass and also exhibits the least variability, with a very short box indicating that most of the data points are close to the median.
- The damage level S has a slightly higher median aircraft mass than D and a somewhat wider spread, indicating more variability in the aircraft mass for this damage category.
- The damage levels M? And M have similar median values of aircraft mass, which are higher than those for S and D. They also have wide boxes, suggesting a considerable spread in the aircraft mass data within these damage levels.
- Damage level N has the highest median aircraft mass and very little spread, which implies that aircraft within this damage category tend to have a consistently high mass.

The above-mentioned FINER questions can then be used for further research and be used to derive meaningful insights for the Airstrikes Data in the US.