Proyecto Individual III

Data Analytics

Data Sciences DTS04

Henry

Oscar Mario Mariño Arias

November 11, 2022

1. **Introduction**:

The project consists of understanding the key factors that produced flight crashes from September 1909 to July 2021 and explaining it in a concise and cohesive presentation. This project corresponds to the third individual project of the Data Science career of Henry cohort 04 (DS04)

1. Data dictionary:

This section is focused primarily in explaining the structure and content of data and provides meaningful descriptions for each of the features that characterize the dataset used.

* 1. Crashes:

This dataset is made of 5008 columns and 17 rows. The features of this table are the final characteristics of the dataframe.

Table . Description of features for dataset Crashes

|  |  |  |
| --- | --- | --- |
| **Column** | **Description** | **Origin** |
| **date** | Date of the accident. | Came with the initial dataset |
| **time** | Time of the accident. | Came with the initial dataset |
| **time\_of\_day** | Time of the day in which the accident happened. | Feature engineering. |
| **crash\_site** | Location of the accident. | Came with the initial dataset |
| **country** | Country of the accident | Feature engineering. |
| **continent** | Continent of the accident | Feature engineering. |
| **latitude** | Latitude of the accident. | Feature engineering. |
| **longitude** | Longitude of the accident. | Feature engineering. |
| **operator** | Airline or operator of the aircraft. | Came with the initial dataset |
| **Flight\_type** | Type of flight | Feature engineering. |
| **Flight\_no** | Flight number assigned by the aircraft operator. | Came with the initial dataset |
| **route\_flight\_type** | Complete or partial route flown prior to the accident. | Came with the initial dataset |
| **ac\_type** | Aircraft type. | Came with the initial dataset |
| **registration** | ICAO registration of the aircraft. | Came with the initial dataset |
| **cn\_ln** | Construction or serial number / Line or fuselage number. | Came with the initial dataset |
| **people\_on\_board** | Total people aboard. | Came with the initial dataset |
| **crew\_aboard** | Crew aboard. | Came with the initial dataset |
| **passengers\_on\_board** | Passengers aboard. | Came with the initial dataset |
| **fatalities** | Total fatalities. | Came with the initial dataset |
| **crew\_fatalities** | Crew fatalities. | Came with the initial dataset |
| **passenger\_fatalities** | Passengers’ fatalities. | Came with the initial dataset |
| **ground** | Total killed on the ground. | Came with the initial dataset |
| **summary** | Description of the accident and cause if known. | Came with the initial dataset |

1. EDA:

The Exploratory Data Analysis (EDA) is the process by which the data analyst becomes acquainted, usually using statistics and visualizations, to start formulating testable hypothesis.

The following are results from applying an EDA:

* Maximize insight into a data set.
* Uncover underlying structure.
* Extract and transform important variables.
* Detect outliers and anomalies.
* Test underlying assumptions.
  1. Null values:

In the Table 2 you will find the number of null records for each of the features. It is important to note that in the initial file, the missing values come as the character ?.

Table . Null values for the features of the dataset Crashes

| **Column** | **Null values (quantity)** | **Column** | **Null values (quantity)** |
| --- | --- | --- | --- |
| **date** | 0 | **people\_on\_board** | 17 |
| **time** | 1054 | **crew\_aboard** | 219 |
| **crash\_site** | 5 | **passengers\_on\_board** | 221 |
| **Operator** | 10 | **fatalities** | 8 |
| **Flight\_no** | 3682 | **crew\_fatalities** | 235 |
| **route\_flight\_type** | 762 | **passenger\_fatalities** | 235 |
| **ac\_type** | 13 | **ground** | 44 |
| **registration** | 272 | **summary** | 59 |
| **cn\_ln** | 667 | - | - |

The null values were handled differently according to each column. For the columns date and time, the module datetime was used after extracting some alphanumeric characters.

For the number of people in air crashes and its fatalities, the sum of the passengers and crew onboard and deaths were used. Buckles were implemented to accomplish this.

* 1. Duplicates:

There are no duplicate records in the dataset Crashes (Figure 2).

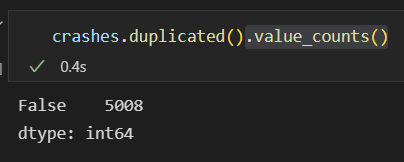


Figure . Duplicate records of the data Crashes

* 1. Feature engineering:

For this section six additional features were added. They mainly consist of location features in order to filter the dataframe according to different characteristics.

* Time\_of\_day: feature obtained using the time column.
* country, continent, latitude, longitude: Characteristics acquired with the tools geopandas and geopy from the column crash\_site.
* Flight\_type: Classification done using key words in the column.