Bootcamp Data Science - The Bridge

PROYECTO EDA

Antonio Pulido Jerez, **abril 2021**

MEMORIA

ÍNDICE

0 – Introducción

1 – Fuentes de Información

2 – Desarrollo del proyecto

3 – Hipótesis y Curiosidades

4 – Librerías utilizadas

5 – Documentación del proyecto

0. INTRODUCCIÓN

Dado que ha sido el primer proyecto EDA al que nos hemos enfrentado, el inicio ha sido duro, debido sobre todo a que, por desconocimiento, no es fácil medir la dificultad o el reto que puede suponer la elección de unos datos u otros. Dicho esto, la experiencia, aunque dura, merece la pena porque enseguida te das cuenta de que la única forma de aprender es 'peleando' con los datos.

Así pues, en el momento de elegir la temática no conocíamos todavía los criterios de evaluación, donde se menciona que es conveniente que este mismo trabajo pueda tener continuidad en los siguientes módulos. Debido a la naturaleza de los datos de los datasets que elegí en su día, es posible que para mí no sea posible, pero aun así he decididio seguir adelante con ello ya que, en fin y al cabo, **es un EDA y la única forma de limpiar datos es que estén sucios,** aunque luego no servirán para ML.

El reto me lo he planteado de la siguiente forma:

- Recibo un dataset sucio y limpio los campos que voy a usar.
- En este nuevo dataset, además, incorporo información de otras fuentes de datos
- Analizo los datos y formulo hipótesis
- Creo dataframes y visualizaciones para la validación de las hipótesis
- Creo un "Story Telling" y lo plasmo en una presentación de negocio
- Defiendo mi presentación en público.

Como he comentado anteriormente, es la primera vez, también para generar una cantidad significativa de código. Por cómo ha ido el proyecto, finalmente mi estructura de carpetas no se ajusta a la perfección con lo mencionado en los criterios de corrección, eso, por supuesto, no significa que no haya un orden y un buen criterio en mi código, que queda de la siguiente forma:

- **src/código**: En esta carpeta están los dos notebooks que he trabajado:
    - Aircrash_Investigation: Análisis de datos y limpieza de los mismos.
    - Aircrash_Visualization: Creación de df's para visualizaciones y visualizaciones
- **src/data**: csv's iniciales y csv's procesados.

  No aporto /utils ni /main, ya que en los propios notebooks está todo el código explicado y ordenado, con índices claros, al principio de los mismos, que permiten el acceso a cada parte del proceso.

Muchas gracias,
Antonio Pulido

1. FUENTES DE INFORMACIÓN

1.1 Kaggle (Principal): https://www.kaggle.com/mihirsethi007

Método de obtención: Descarga de fichero ".csv" directa desde la web

Archivo utilizado: aircrash1908-2008.csv

Descripción: Registro de accidentes aéreos entre 1908 y 2009

Estructura:

Variable	Description	Type	Nulls
Date	Date of accident, in the format - January 01, 2001	Object	0
Time	Local time, 24h format unless otherwise specified	Object	2219
Location	Region and/or Country where the accident took place	Object	20
Operator	Airline or operator of the aircraft	Object	18
nFlight	Flight number assigned by the aircraft operator	Object	4199
Route	Complete/partial route flown prior to the accident	Object	1706
Type	Aircraft type	Object	27
Regist	ICAO registration of the aircraft	Object	335
cn/In	Construction-serial number / Line-fuselage number	Object	1228
Aboard	Total aboard (passengers / crew)	Float	22
Fatal	Total fatalities aboard (passengers / crew)	Float	12
Ground	Total killed on the ground	Float	22
Summary	Brief description of the accident and cause if known	Object	390

1.2. Plane Crash Info (Complementaria): http://www.planecrashinfo.com

Método de obtención: web scraping

Descripción: Sitio web con base de datos de accidentes aéreos desde 1908

1.3. Airfleets (Complementaria): https://www.airfleets.net/home/

Método de obtención: web scraping

Descripción: Información sobre aeronaves civiles. Cubre la mayoría de fabricantes del mundo (Airbus, Boeing, Embraer, Bombardier, Sukhoi, Fokker, ...). Airfleets informa sobre movimientos de aeronaves entre diferentes operadores y estado de flotas de la mayoría de aerolíneas del mundo.

1.4. Banco Mundial (Complementaria): https://www.bancomundial.org/es/home

Descripción: Asociación internacional que trabaja proyectos para el desarrollo y facilita, de forma gratuita, un repositorio abierto de conocimiento sobre múltiples temáticas, como: social, urbanístico, economía y finanzas, sector privado y público.
Método de obtención: Descarga de ficheros ".csv" directa desde la web.
Archivos utilizados:

- API_IS.AIR.DPRT_DS2_en_csv_v2_2169474.csv:
  Transporte aéreo, **salidas de vuelos**/país/año a nivel mundial

- API_IS.AIR.PSGR_DS2_en_csv_v2_2252261.csv:
  Transporte aéreo, **pasajeros transportados**/país/año a nivel mundial

- API_SP.POP.TOTL_DS2_en_csv_v2_2252106.csv:
  Población global, **evolución**/país/año de población mundial

1.5. Estadística para todos (Complementaria):
http://www.estadisticaparatodos.es/taller/loterias/loterias.html

Descripción: Web especializada en estadística
Archivo utilizado: Página con probabilidad de diferentes loterías
Método de obtención: Directa desde la web

2. DESARROLLO DEL PROYECTO – CRONOLOGÍA

2.1. Temática elegida: Accidentes Aéreos

2.2. Análisis inicial:
El primer reto fue encontrar fuentes de datos completas y fiables. En dicha búsqueda encontré 3 fuentes: **planecrashinfo.com, kaggle y airfleets.net**

Planecrashinfo.com: Fue la primera que evalué, tiene datos de accidentes aéreos desde 1908 a 2009. Dispone de una base de datos principal, basada en "url's", una por cada año del periodo mencionado.
Además, dispone de diferentes secciones con información diversa sobre los accidentes, como fotos, accidentes de famosos, historia, estadísticas, etc.

- Kaggle: Haciendo una búsqueda en el propio kaggle, encontré un dataset de accidentes aéreos de 1908 a 2009, me llamó la atención que el periodo coincidía con el de planecrashinfo.com. Al comparar ambas fuentes, enseguida me di cuenta de que es la misma base de datos. Sin embargo, la base de datos de planecrashinfo.com disponía de menos campos que la de kaggle, lo cuál es extraño, ya que lo normal es que sea el usuario de Kaggle el que ha obtenido le info de planecrashinfo.com, y no al revés. En cualquier caso me di cuenta de que, obviamente, el dataset de kaggle sería la fuente principal.

- Airfleets.net: Esta web dispone de todo tipo de información sobre aviación, entre la que está una base de datos de accidentes aéreos, de la que me interesaba la información de fabricante y tipo de avión, la cuál, en principio, también estaba en kaggle, pero en airfleets estaba mucho más limpia.

- Bancomundial.org: Al evaluar las fuentes anteriores, me percaté de que, para algunas de las hipótesis que me planteaba, necesitaría información sobre cantidad total de vuelos, de forma que pudiera calcular proporciones. En esta página hay información muy limpia y completa de cantidad de vuelos por país y año, cantidad de pasajeros transportados por país y año y evolución de la población mundial por país y año. Esto me abría un abanico de posibilidades muy grande a la hora de hacer análisis, dado que había dos variables coincidentes con el dataset principal; el país y el año.

Una vez decididas las fuentes de datos, planteé unas hipótesis iniciales. Obviamente, tras hacer un análisis más profundo, tuve que iterar y cambiar algunas de ellas, como la principal, ya que con la información disponible no se podía validar.

2.3. Preparación de fuentes de datos, retos y dificultades:

– **Planecrashinfo.com**: Dado que la base de datos era incompleta, decidí utilizar la URL de 'Accidents by Category', en la que, a priori, estaban los accidentes por tipo de causa, una url por causa. Una vez obtenida la información, observé que algunos valores estaban en el campo equivocado, lo cuál me obligó a corregirlo, la mayoría con código y los últimos flecos manualmente.
Web Scraping, librería utilizada: ***pandas***

– **Airfleets.net**: Tal como he mencionado anteriormente, de aquí obtuve las tablas de accidente por año con detalle de fabricante y tipo de avión. Este fue el primer reto relevante, ya que el web scraping era más complejo y, de hecho, fui baneado por la página y tuve que hacer uso humano, durante un periodo, para poder volver a intentarlo. Aún así, en la segunda tentativa, habiendo añadido 'time lapses', para simular comportamiento humano, el scraping se cortó cuando llevaba un 70% aproximadamente, finalmente me apoyé en un compañero para que ejecutar el código desde su ordenador y así completé el fichero.
Web scraping, librerías utilizadas ***urllib.request*** y ***bs4.BeautifulSoup***

Al hacer el análisis, ya en Jupyter Notebook, observé que la cantidad de accidentes por año era muy inferior al del dataset de Kaggle. Observando los datos se intuye que es porque solo aparecen los de grandes compañías aéreas comerciales. Dado que la muestra era tan inferior, finalmente decidí **no utilizar la información de airfleets en la presentación del proyecto**.

– **Kaggle:** Enseguida me di cuenta de que los datos eran muy sucios y que, por tanto, la limpieza de los datos iba a suponer un reto importante, el cual decidí acometer ya que, en fin y al cabo, es el primer proyecto EDA y era importante este aprendizaje. Así pues, decido limpiar el 'dataset' y crear uno más limpio y completo.

2.3. Proceso de limpieza de la fuente de datos principal:

2.3.1. Dataset inicial (input):

```
RangeIndex: 5268 entries
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype  Nulls
---  ------        --------------  -----  -----
 0   Date          5268 non-null   object    0
 1   Time          5268 non-null   object 2219
 2   Location      5248 non-null   object   20
 3   Operator      5250 non-null   object   18
 4   Flight #      1069 non-null   object 4199
 5   Route         3562 non-null   object 1706
 6   Type          5241 non-null   object   27
 7   Registration  4933 non-null   object  335
 8   cn/In         4040 non-null   object 1228
 9   Aboard        5246 non-null   float64  22
 10  Fatalities    5256 non-null   float64  12
 11  Ground        5246 non-null   float64  22
 12  Summary       4878 non-null   object  390
dtypes: float64(3), object(10)
```

En rojo los campos descartados, en verde los utilizados:
Registration, cn/In y Flight #: Campos no relevantes para el análisis, como el número de vuelo.

Summary: Este campo contiene, a priori, una descripción del accidente y/o la causa del mismo. Usando las causas de accidente de planecrashinfo, busco apariciones de las mismas y solo obtengo **322** registros, un **6%** del total, por lo que descarto el análisis por causa de accidente en este dataset.

Time: Misma info que campo Date, pero con información de hora, con estructura MM/DD/AAA HH:mm.
Lo descarto porque el **42%** de los registros tienen hora 00:00, lo cuál lo hace poco fiable.

Route: En principio es la ruta que cubría el avión, lo cuál lo hacía interesante a priori. Sin embargo, decido descartarlo, ya que **1/3** de los datos son desconocidos y, además, no viene el país de origen y destino, por lo que no hubiera podido hacer el análisis.

Date: Fecha del accidente, en formato string, lo utilizo para análisis por año y por meses.

Location: Lugar del accidente, lo uso para obtener el país, pero supone un trabajo engorroso, ya que aparecen distintos niveles, población, ciudad, país, etc. Separados por comas, el problema es que no tiene orden y, además, no aparecen todos los niveles en todas las celdas.

Operator: Teóricamente, compañía aérea que operaba el avión, sin embargo hay 2476 valores únicos, dado que el mismo operador está escrito de diferente forma y, a veces, con las iniciales, lo que hace imposible su limpieza. Sin embargo, cuando es un vuelo militar lo dice explícitamente y cuando es un vuelo de servicio de trasporte de mercancías, también. Lo utilizo para crear un campo de tipo de vuelo.

Type: Fabricante y tipo de avión. Campo vital para mis hipótesis, sin embargo pasa igual que con el campo operator, tiene 2445 valores únicos. Decido exportar los valores únicos a csv y, utilizando Excel, hago un campo con el fabricante y con el país del fabricante, los cuáles añado al dataset.

Numéricos:
Aboard: Número de personas que iban a bordo del avión.
Fatalities: Número de fallecidos en el accidente, que iban a bordo del avión.
Ground: Número de fallecidos en el accidente que NO iban a bordo del avión.

2.3.2. Dataset resultante de la limpieza (output):

En verde los campos nuevos.

```
Int64Index: 5256 entries, 0 to 5255
Data columns (total 26 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Date                      5256 non-null   object
 1   Time                      5256 non-null   object
 2   Location                  5256 non-null   object
 3   Operator                  5256 non-null   object
 4   Route                     5256 non-null   object
 5   Type                      5256 non-null   object
 6   Aboard                    5256 non-null   float64
 7   Fatalities                5256 non-null   float64
 8   Ground                    5256 non-null   float64
 9   Summary                   5256 non-null   object
 10  unos                      5256 non-null   float64
 11  year                      5256 non-null   float64
 12  type_flight               5256 non-null   object
 13  country                   5256 non-null   object
 14  airfleets_operator        5256 non-null   object
 15  pci_accident_cause        5256 non-null   object
 16  date2                     5256 non-null   datetime64[ns]
 17  fatal_ratio               5256 non-null   float64
 18  aircraft_manufacturer     5256 non-null   object
 19  aircraft_type             5256 non-null   object
 20  manufacturer_country      5256 non-null   object
 21  country_code              5256 non-null   object
 22  region                    5256 non-null   object
 23  income_group              5256 non-null   object
 24  manufacturer_region       5256 non-null   object
 25  manufacturer_income_group 5256 non-null   object
dtypes: datetime64[ns](1), float64(6), object(19)
```

Descripción campos nuevos:
unos: columna de 1's, la creo para hacer 'groupings' y agregaciones

year: Año del accidente, formato numérico para usar en series temporales, extraído del campo Date.

type_flight: Vuelo comercial, militar o de servicio (cargo), extraido del campo 'Operator'

country: País donde se produjo el accidente, extraído de 'Location' y estandarizado con los países de bancomundial.org

airfleets_operator: Compañías aéreas extraídas de Airfleets.net, sólo identifica el **20%** de los registros, lo mantengo pero no lo uso.

pci_accident_cause: Causas de accidente de planecrashinfo.com, sólo identifica el **6%**, lo mantengo pero no lo uso.

date2: Misma información que el campo fecha pero en formato ***datetime.*** Lo utilicé cuando llegamos a esa materia en el bootcamp.

fatal_ratio: porcentaje de víctimas abordo respecto a personas abordo.

aircraft_manufacturer: Para poder hacer análisis por el fabricante del avión, exporto los 2445 valores únicos de la columna original 'Type' para, utilizando Excel, obtener el fabricante, país del fabricante y tipo de aeronave.

country_code: Una vez extraído el país del campo original 'Location' y haberlo modificado para que fueran los mismos valores que aparecen en bancomuncial.org, hago un merge para incorporar, mediante el campo país, la región y el nivel de ingresos de os países en los que se produjeron los accidentes.

region Región a la que pertenece el país en el que se produjo el accidente, incorporado en el merge descrito con bancomundial.org

income_group Nivel de ingresos del país donde se produjo el accidente, incorporado en el merge con bancomundial.org

manufacturer_region Región a la que pertenece el país del fabricante del avión, proviene de merge con bancomundial.org, pero esta vez usando el campo del país del fabricante.

manufacturer_income_group Nivel de ingresos del país del fabricante del avión, proviene de merge con bancomundial.org, pero esta vez usando el campo del país del fabricante.

3. HIPÓTESIS PLANTEADAS

3.1. Hipótesis principal:

Proporcionalmente, se producen más accidentes en países pobres que en países ricos.

3.2 Otras hipótesis

- Volar cada vez es más seguro.
- El porcentaje de víctimas, respecto a pasajeros, disminuye con el tiempo
- Es más probable que te toque la lotería a tener un accidente aéreo provocado por un rayo
- El mes con más accidentes es agosto

3.3 Curiosidades

- El año de más accidentes de vuelos militares coincide con el de más accidentes de vuelos comerciales
- Es más probable tener un accidente provocado por un rayo que por un fallo de diseño del avión.
- El accidente con más víctimas de la historia fue en Kinshasa, en la actual República Democrática del Congo.
- Martes, ni te cases ni te embarques.

4. INSTALACIONES Y LIBRERÍAS UTILIZADAS

4.1 Instalaciones:

```
!pip install squarify
```

4.2 Librerías:

```
import pandas as pd
import numpy as np
import matplotlib as mlb
import matplotlib.pyplot as plt
import seaborn as sns
import squarify
from sklearn import preprocessing
import urllib.request
from bs4 import BeautifulSoup
from datetime import timedelta,date,datetime
import time
import random
import warnings
warnings.filterwarnings("ignore")
plt.style.use('seaborn-bright')
```

5. DOCUMENTACIÓN DEL PROYECTO

5.1 Repositorio GitHub:

repo.txt
https://github.com/Ant14DS/Bootcamp_DATA_SCIENCE-2021/tree/main/Proyecto_EDA

5.2 Documentos:

-Presentación: Presentación_EDA_Aricrash_Antonio-Pulido_Abr2021.pptx
-Memoria: Memoria_Proyecto_EDA_Antonio-Pulido-Abr2021.docx

5.3 Código:

Los notebooks ya contienen las pruebas, módulos y explicación de la analítica
-Aircrash_Investigation.ipynb: Limpieza de datos
-Aircrash_Visualization.ipynb: Dataframes para visualizaciones y visualizaciones

5.4 Archivos csv:

Input	Fuente	Output
aircrash1908-2009.csv	Kaggle	aircrash1908-2009_procesado.csv
airfleets_accidents_1967-1997	airfleets.net	airfleets_accidents_1967-2019.csv
airfleets_accidents_1998-2019	airfleets.net	airfleets_accidents_1967-2019.csd
bancomundial_org_flights-country-year.csv	bancomundial.org	bancomundial_org_flights-country-year_procesado.csv
bancomundial_org_country-passengers-year.csv	bancomundial.org	bancomundial_org_country-passengers-year_procesado.csv
bancomundial_org_global-population.csv	bancomundial.org	bancomundial_org_global-population_procesado.csv
bancomundial_org_country-code_income-level.csv	bancomundial.org	-
pci_causes.csv	planecrashinfo.com	-
Probabilidad_Lotería.csv	estatisticaparatodos.es	-
manufacturers_before.csv	airfleets_accidents_1967-2019.csv	manufacturers_after.csv
united_states.csv	propio	-