**PROYECTO 03 – DATA ANALYST**

**Contexto**

La **Organización de Aviación Civil Internacional (OACI)**, organismo de la Organización de las Naciones Unidas, quiere investigar en profundidad los accidentes producidos desde inicios del siglo XX. Para ello, les solicita la elaboración de un informe y un dashboard interactivo que recopile tal información.

La OACI únicamente cuenta con un dataset sobre datos de accidentes de aviones, pero insta a la consultora de datos -de la que forman parte- que intente cruzar esta información con otras fuentes de su interés. Esto con el objetivo de obtener mayor claridad y consistencia en los fundamentos del estudio.

**Propuesta de trabajo**

A raíz de esta solicitud, nuestro Project Manager nos encarga una serie de tareas a cumplir:

* Realizar un EDA con el dataset provisto, junto con un reporte de calidad y diccionario de datos
* Buscar y relacionar información relevante con los eventos
* Crear una base de datos en un motor SQL e ingestar el csv procesado
* Elaborar un dashboard e idear un storytelling con el objetivo de presentarlo ante la OACI
* Adjuntar todo el trabajo en un repositorio de GitHub

**ARCHIVOS ADJUNTOS DEL TRABAJO:**

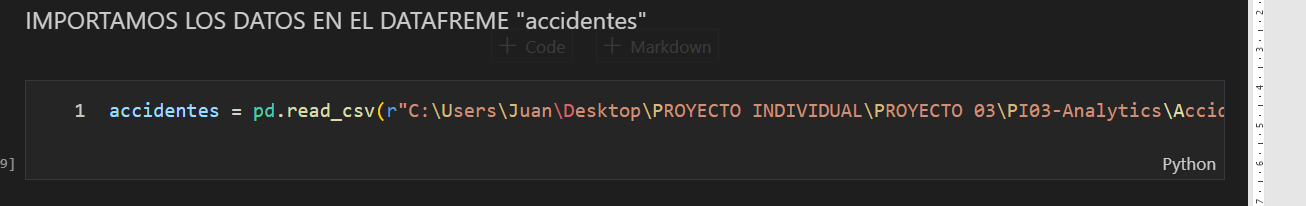
-PROYECTO03.ipynb

-Análisis de calidad, diccionario de datos y sustentación.pdf

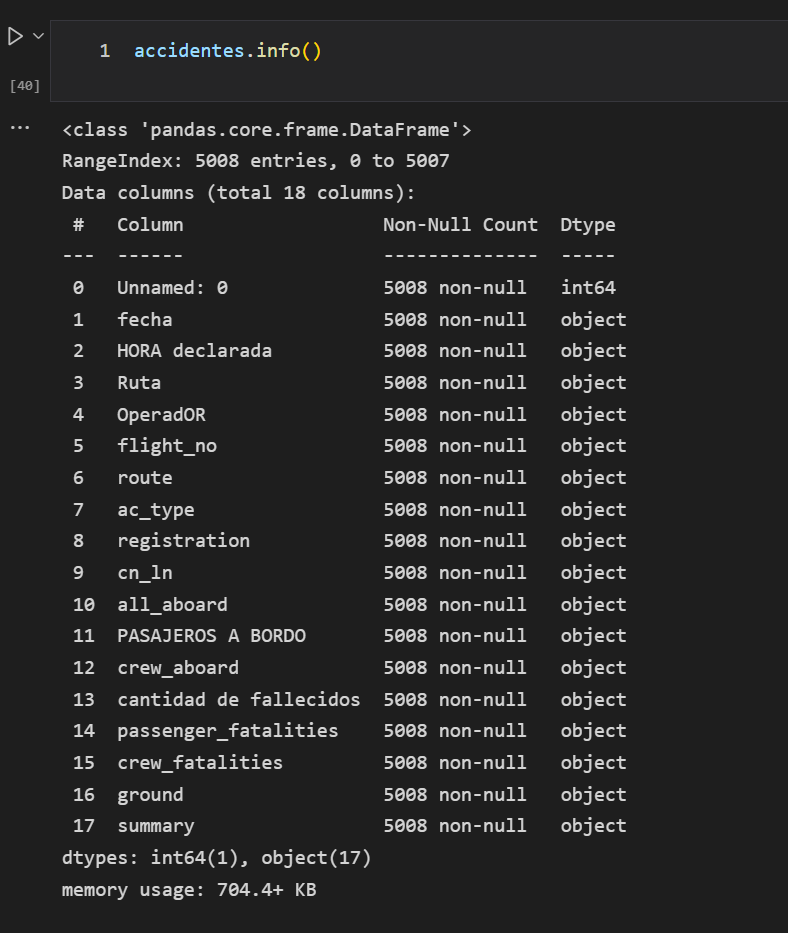
-ANALISIS.pbix

-AccidentesAviones.csv

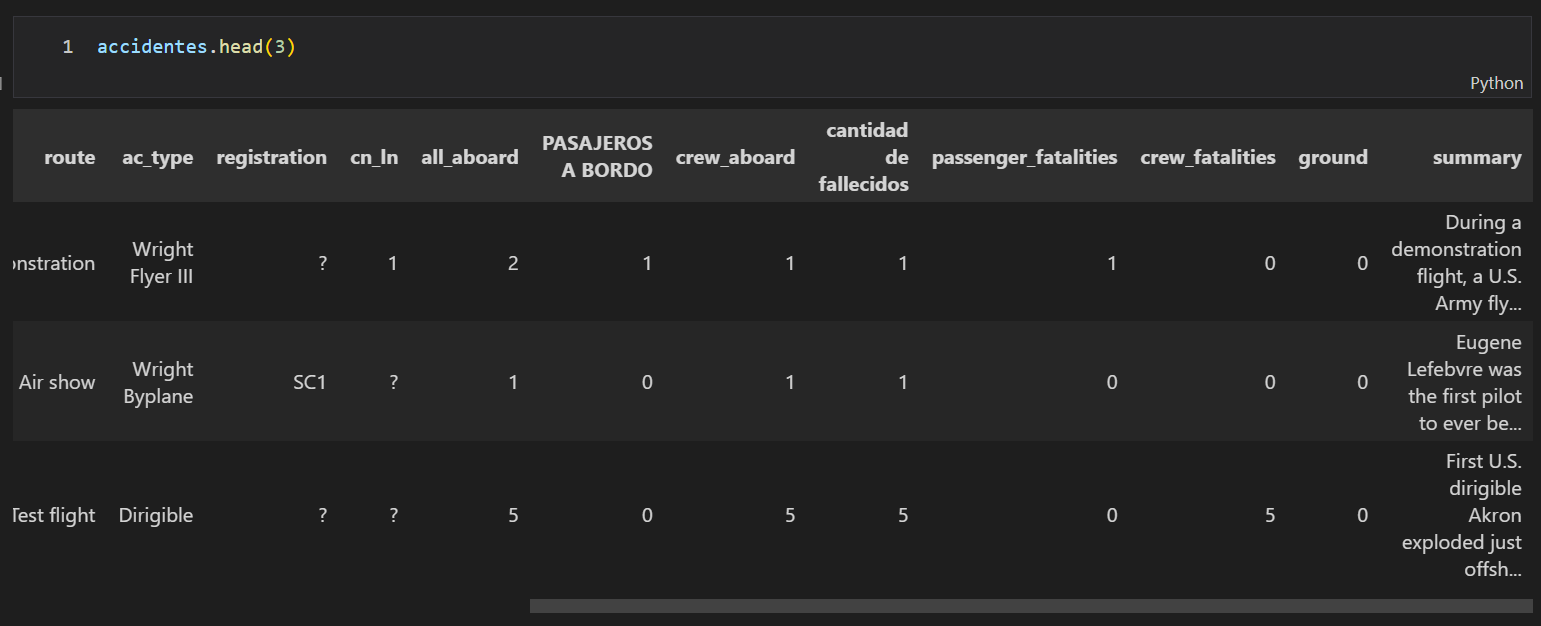
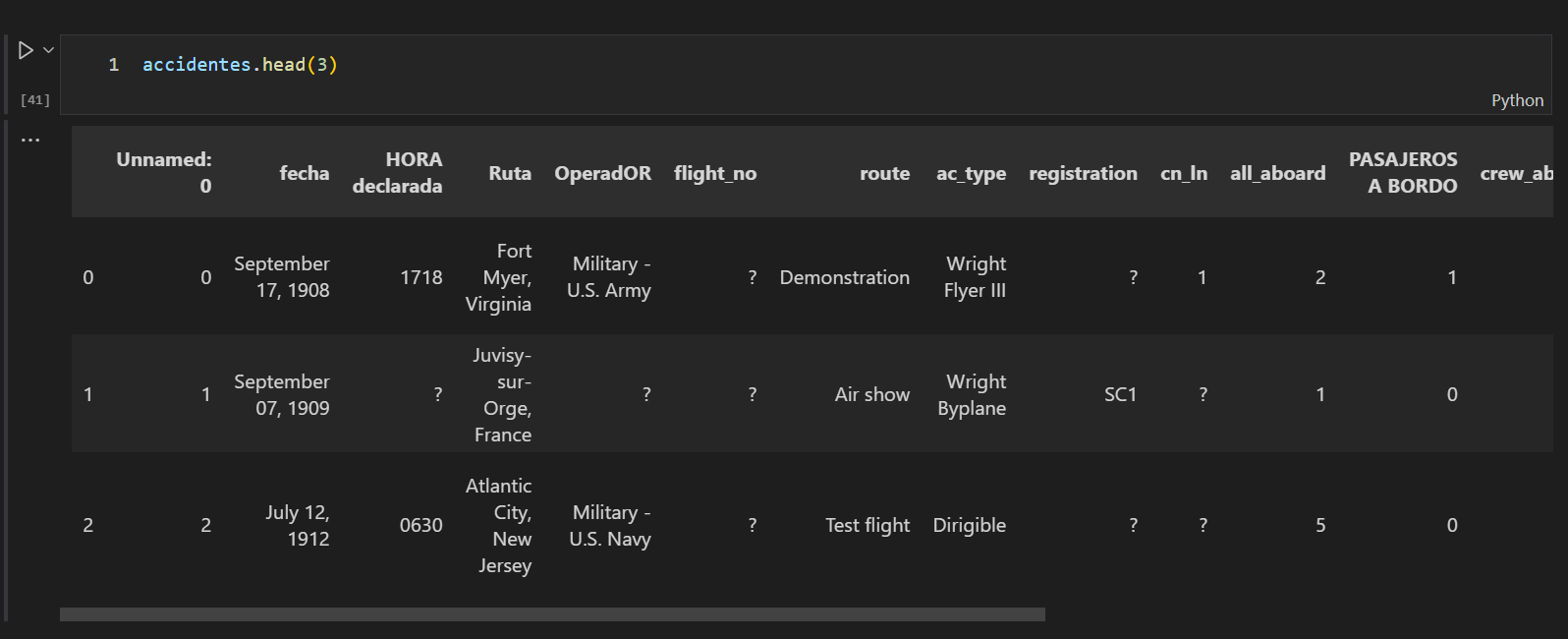
**EDA:**



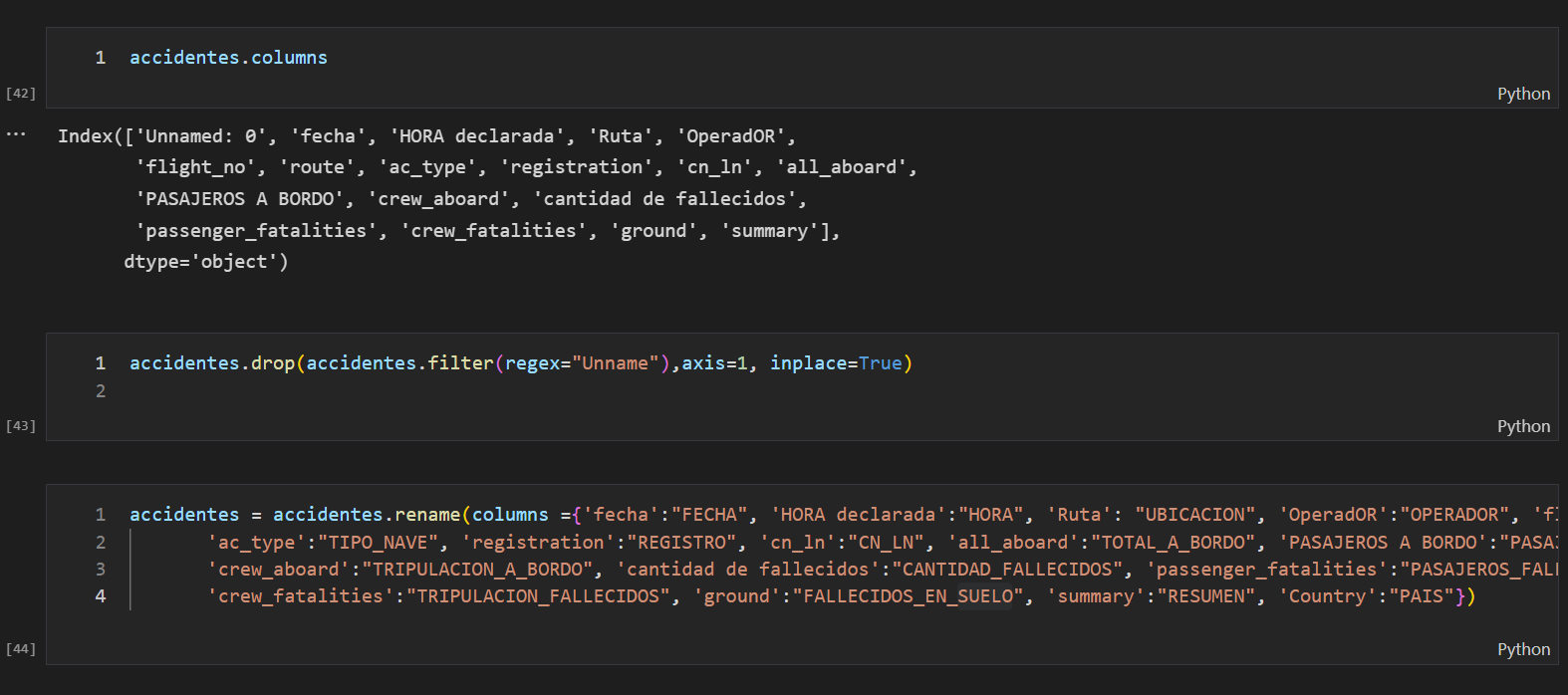
Primer importamos los datos en un data frame que llamaremos “accidentes”



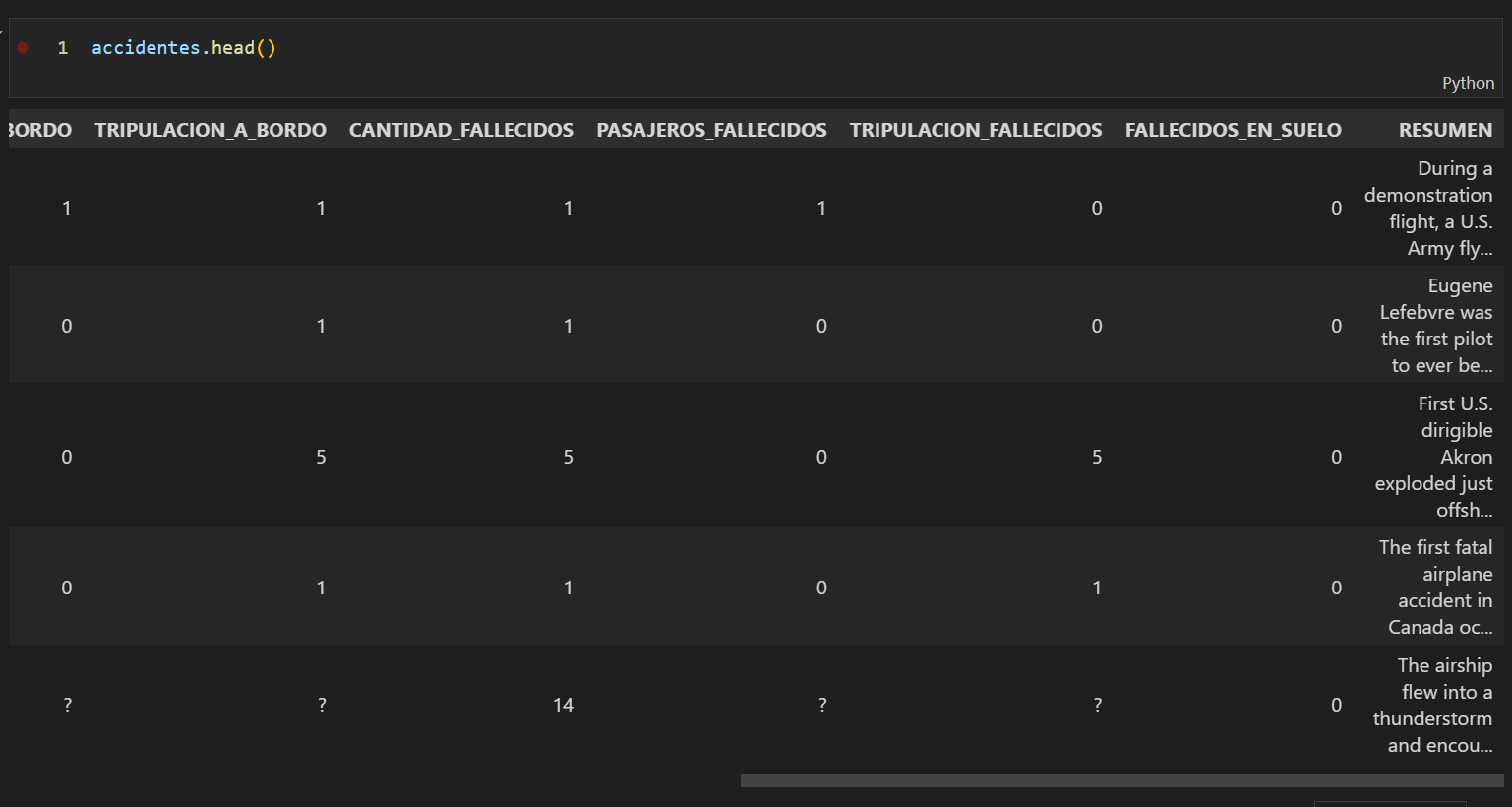
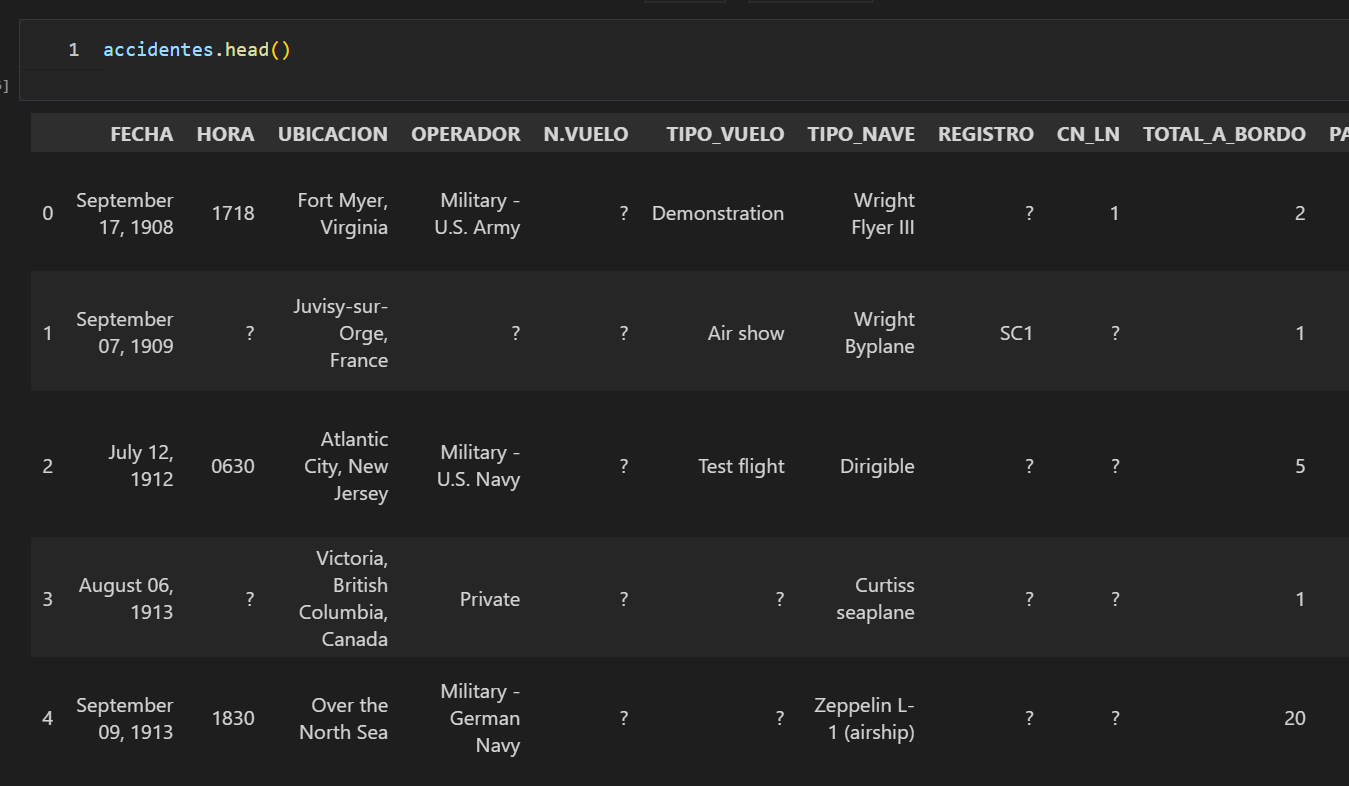
Mediante info() conocemos que tenemos 5,008 registros (filas) y 17 columnas. De las 17 columnas, solamente 1 es del tipo entero y las otras 16 son texto. Además, podemos observas que no hay datos nulos (eso es bueno), no obstante, vamos a visualizar el data frame



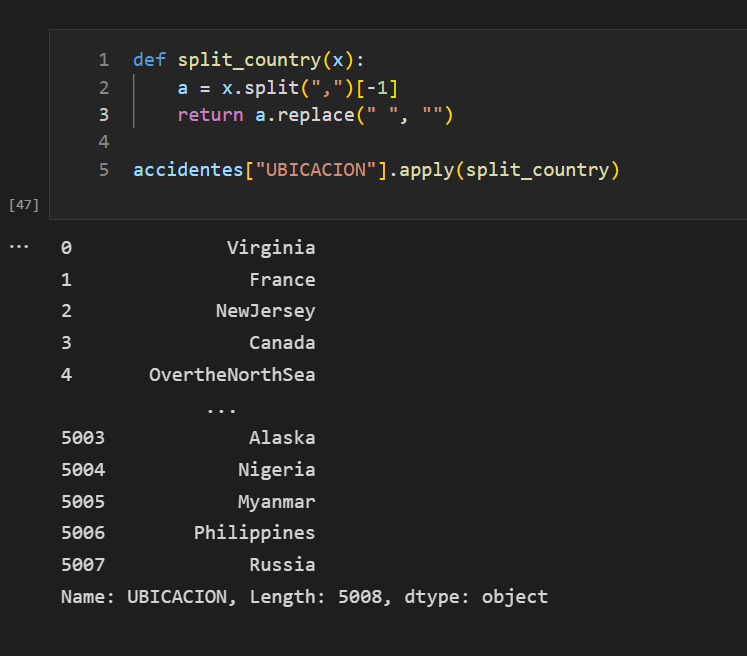
Imprimimos las 3 primeras filas de accidentes, vemos que hay columnas que eran texto pero que contienen números. Además, si bien es cierto que supuestamente no hay datos nulos, estos aparentemente fueron sustituidos por “?”. Por último la columna “Unmamed: 0” está como un índice.



En este paso, dropearemos la columna “Unnamed: 0”, ya que no aporta en el análisis. Luego pasamos a estandarizar los nombres de las columnas y las vamos a renombrar de acuerdo al dato que tiene.



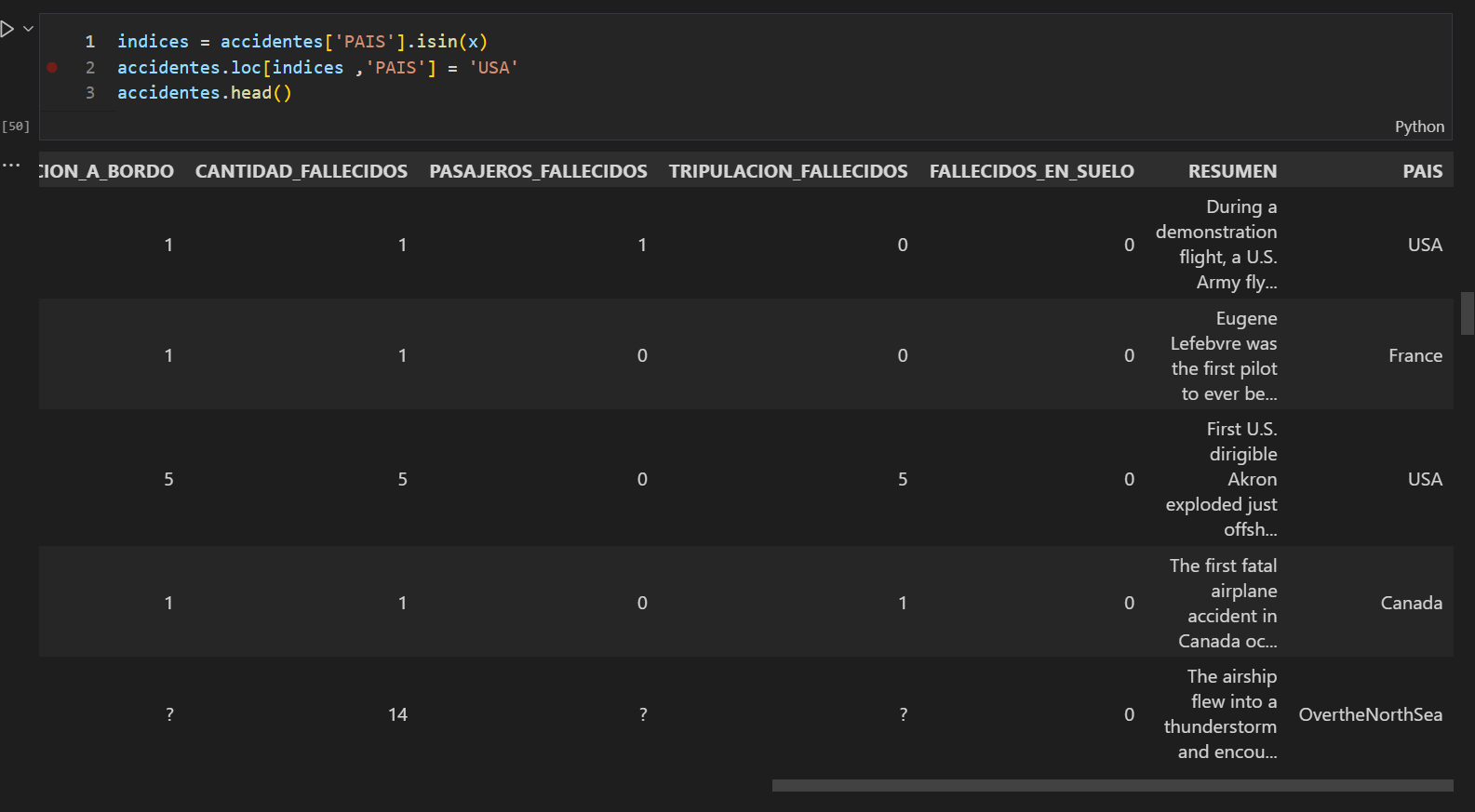
Ahora como vimos las columnas ya fueron renombradas de acuerdo al dato que contienen. El siguiente paso es ir estandarizando los datos, empezamos por la columna “UBICACIÓN”, nos damos cuenta que existen ciudades con sus respectos países, pero por ejemplo para cuando está una ubicación de Estados Unidos, sólo están como ciudades y estado. Para el análisis vamos a requerir las ubicaciones por países.



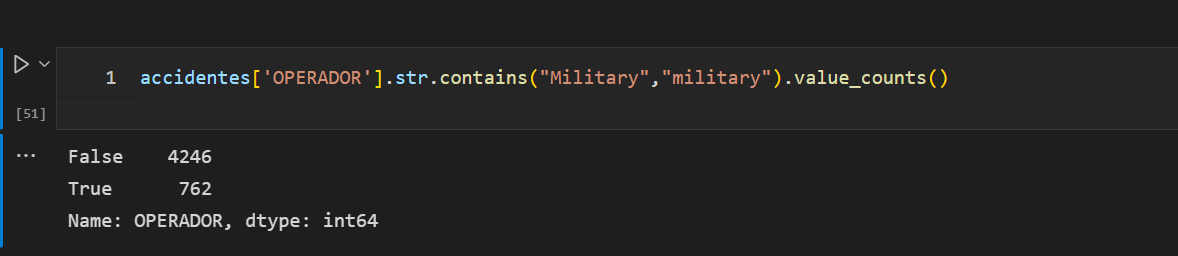
Creamos una función que nos permita obtener la última palabra de la cela, observamos que se guardaron Virginia, France, NewJersey, Canada, entre otros.



Ahora vemos a cambiar los datos que están como nombres de estado, por el de “USA”, luego imprimos



Vemos que ahora ya cambiaron los nombres de los estados por USA



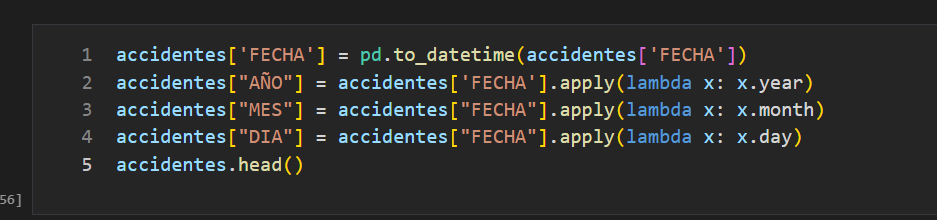
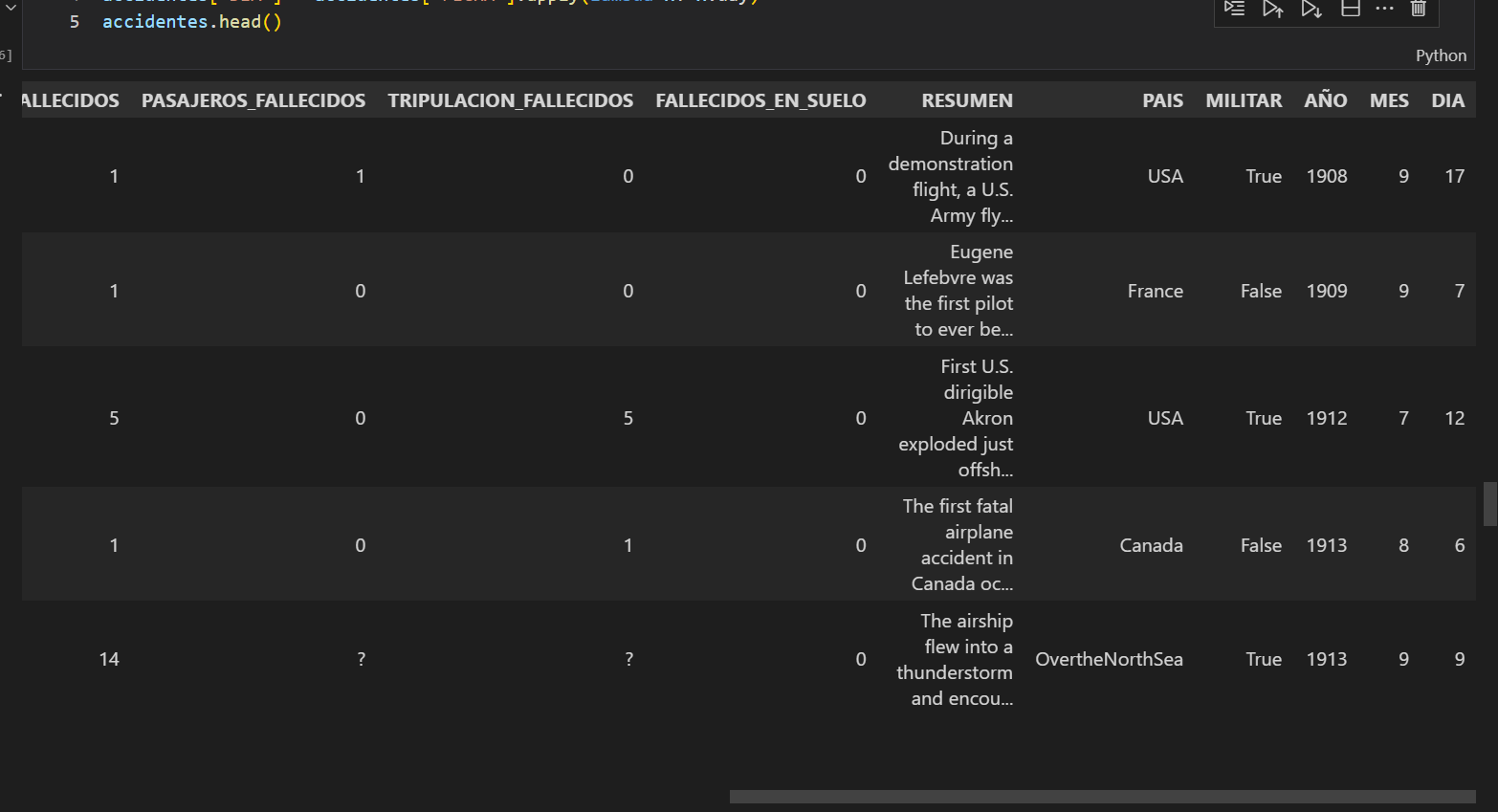
Ahora queremos extraer y conocer cuántos vuelos fueron de clase militar o no. Acá vamos a contar las filas que contengan “Military” y “military”. Nos brinda un resultado de 4,246 como falsos y 762 como verdaderos. Nos damos cuenta que efectivamente podemos usar esta información puede aportar en el análisis, para determinar si fueron vuelos militares o no militares.



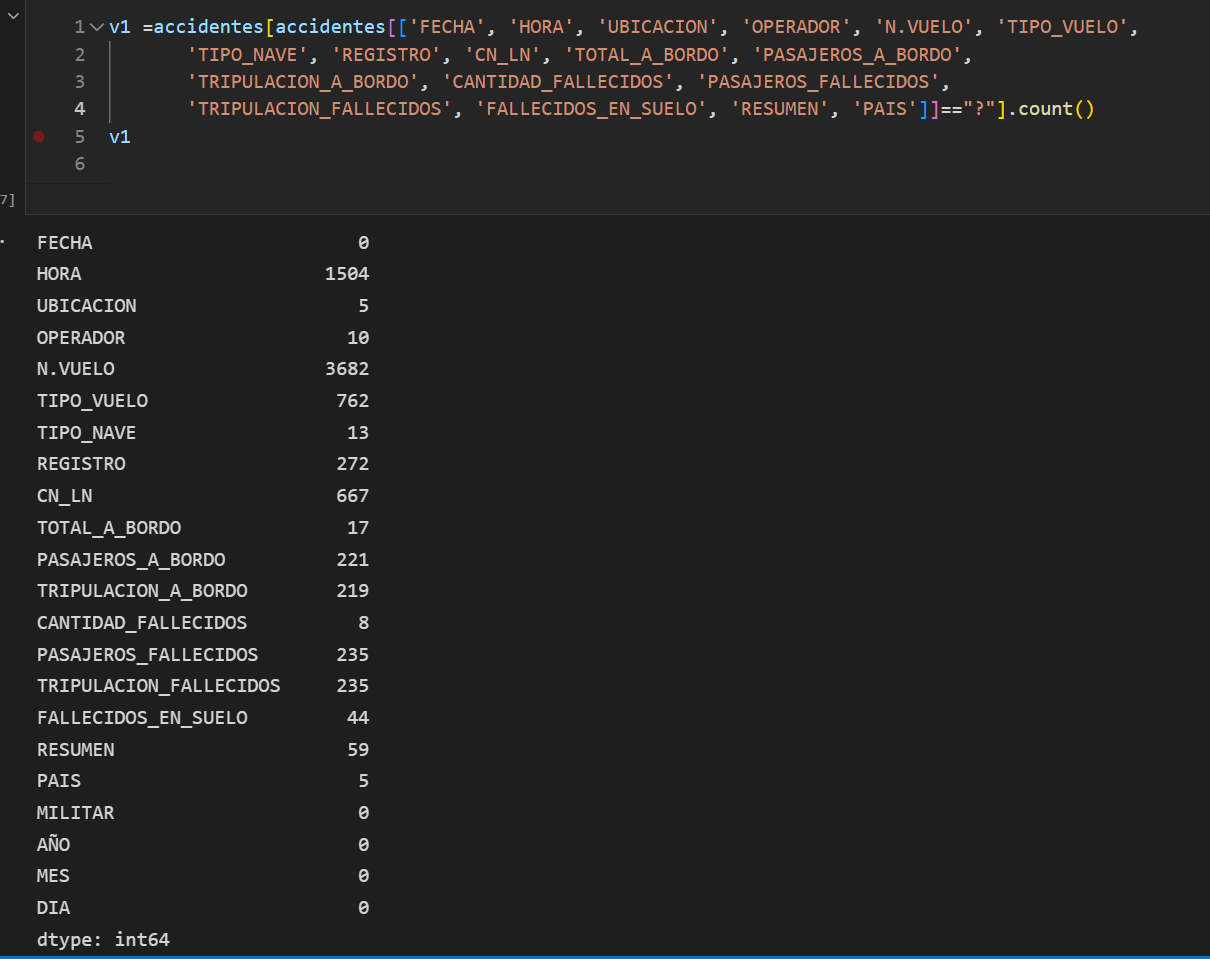
Entonces creamos una variable que contenga los resultados llamada “a”, luego pasamos a guardar los resultados en una nueva columna que llamaremos “MILITAR”, finalmente para imprimir usamos gráficos de barras para apreciar mejor las cantidades y la diferencia.



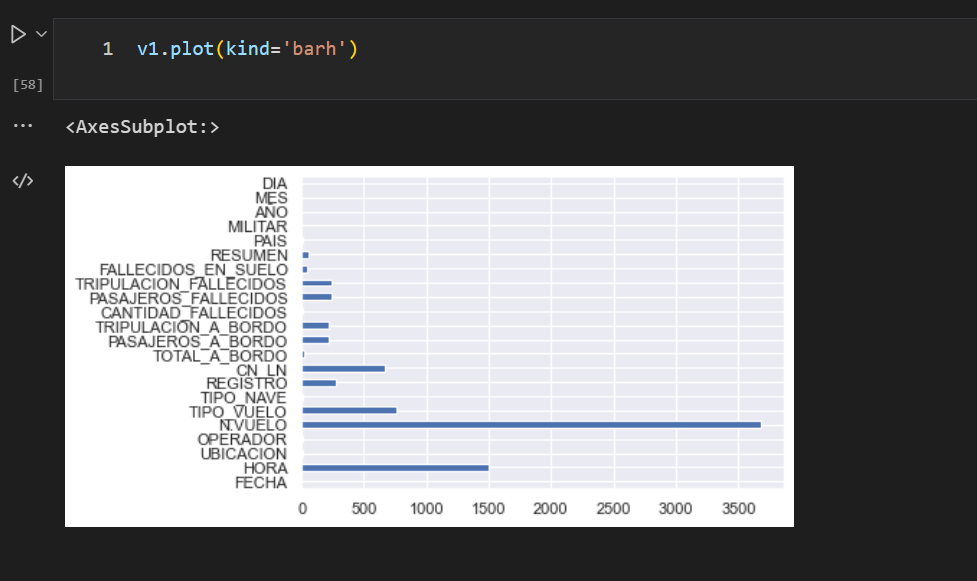
Vemos que se creó la nueva columna y se guardaron los resultados correctamente

Ahora queremos extraer el año, mes y día de la columna “FECHA” y luego imprimimos 

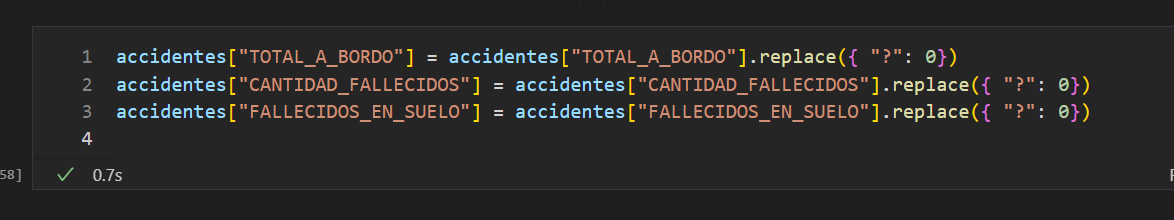
Vemos que se crearon nuevas columnas con los datos respectivos para cada una.



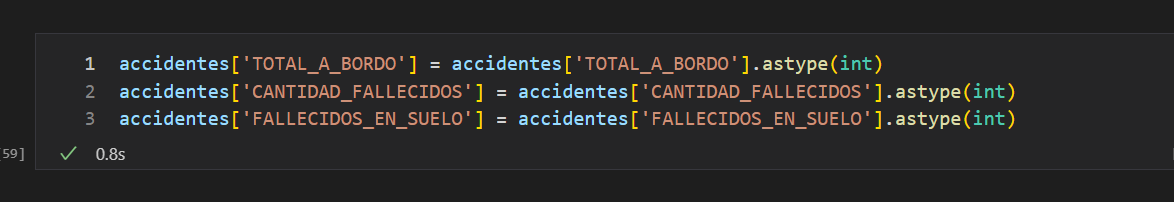
Ahora para conocer con qué columnas vamos a trabajar queremos conocer cuántos “?” hay por cada una. Podemos ver que las columnas que no tienen menos de 50 “?” son FECHA, PAIS, OPERADOR, TIPO\_NAVE, TOTAL\_A\_BORDO, CANTIDAD\_FALLECIDOS, FALLECIDOS\_EN\_SUELO, MILITAR, AÑO, MES, DIA.



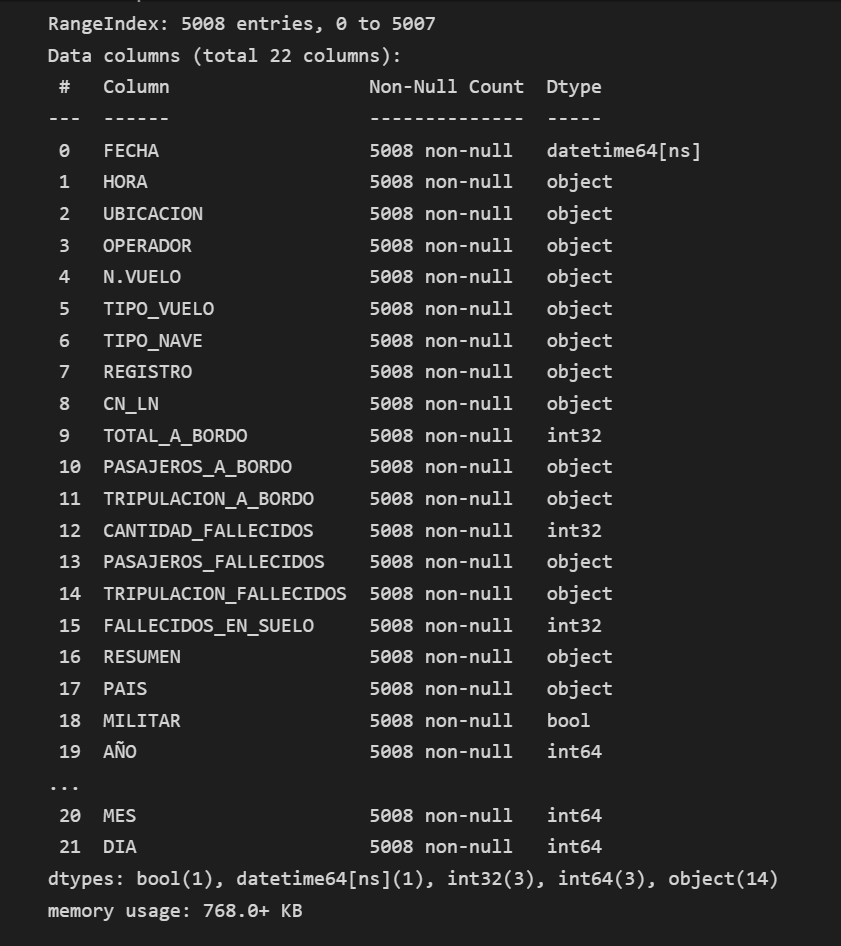
Imprimimos para apreciar mejor las cantidades, por lo antes descrito usaremos esas columnas para el análisis, ya que contienen la menor cantidad de datos faltantes o “?”.



Ahora vamos a pasar estas columnas que están como strings a int, no obstante para cambiar el tipo de datos, debemos cambiar los “?” a 0, esto no afectará en el análisis ya que la columna que tiene la mayor cantidad de ese valor es FALLECIDOS\_EN\_SUELO con 44, no representa ni el 0.88% del total.



Ahora los pasamos a enteros para que en power bi podamos analizar correctamente.



Por último verificamos que se cambiaron los tipos de datos, llegamos al final de esta parte que es la preparación para un correcto análisis.

**DICCIONARIO DE DATOS:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **NOMBRE COLUMNA** | **CANTIDAD DE DATOS “?”** | **TIPO** | **DESCRIPCIÓN** | **NOMBRE ANTERIOR** | **¿SE USARÁ PARA ANÁLISIS?** |
| COLUMNA ELIMINADA | 0 | - | Se eliminó, no aporta | Unnamed: 0 | No |
| FECHA | 0 | Fecha | Fecha del vuelo | fecha | Sí |
| HORA | 1504 | Texto | Hora del vuelo | HORA declarada | No |
| UBICACIÓN | 5 | Texto | Ubicación de vuelo | Ruta | Sí |
| OPERADOR | 10 | Texto | Nombre del operador de vuelo | OperadOR | Sí |
| N. VUELO | 3682 | Texto | N. de vuelo | flight\_no | No |
| TIPO\_VUELO | 762 | Texto | Tipo, clase de vuelo | route | No |
| TIPO\_NAVE | 13 | Texto | Tipo, modelo de nave | ac\_type | Sí |
| REGISTRO | 272 | Texto | Código registro | registration | No |
| CN\_LN | 667 | Texto | - | cn\_ln | No |
| TOTAL\_A\_BORDO | 17 | Int | Total personas a bordo de nave | all\_aboard | Sí |
| PASAJEROS\_A\_BORDO | 221 | Int | Total pasajeros a bordo en nave | PASAJEROS A BORDO | No |
| TRIPULACION\_A\_BORDO | 219 | Int | Total tripulación a bordo en nave | crew\_aboard | No |
| CANTIDAD\_FALLECIDOS | 8 | Int | Total de fallecidos en nave | cantidad de fallecidos | Sí |
| PASAJEROS\_FALLECIDOS | 235 | Int | Total de pasajeros fallecidos en nave | Passenger\_fatalities | No |
| TRIPULACION\_FALLECIDOS | 235 | Int | Total de tripulación fallecida en nave | crew\_fatalities | No |
| FALLECIDOS\_EN\_SUELO | 44 | Int | Total fallecidos en suelo | ground | Sí |
| RESUMEN | 59 | Texto | Descripción del accidente en inglés | Summary | No |
| PAIS | 5 | Texto | País del vuelo | - | Sí |
| MILITAR | 0 | Bool | Si vuelo fue militar o no | - | Sí |
| AÑO | 0 | Int | Año de vuelo | - | Sí |
| MES | 0 | Int | Mes de vuelo | - | Sí |
| DIA | 0 | Int | Dia de vuelo | - | Sí |

**CONCLUSIÓN:**

Habiendo conocidos los tipos de datos, cantidad y calidad de datos por cada columna, creado columnas para analizar. Se determinó que se usarán 12 del total de 22 columnas, ya que cumplen con los requisitos de tener pocos registros nulos (representados con “?”) la mayoría de ellas tiene 0 datos nulos , sin embargo algunas si tienen, aunque no son más de 44 datos de los 5008 que hay por cada columna, esto quiere decir que estos dato nulos representan un 0.8786%, esto quiere decir que no representan ni siquiera el 1%, por lo tanto el análisis que se hará tendrá poco ruido y será bastante preciso. Analizaremos las cantidades de fallecidos según tipo de vuelo (militar o no), por países, por operador, por fecha de vuelo, tipo de nave, total de personas a bordo y si fallecieron en suelo u otra zona.