

General Electric Aerospace



Data Understanding

Flavio Ruvalcaba Leija - A01367631

Oscar Eduardo Nieto Espitia - A01705090

Eduardo Gonzalez Luna - A01658281

Cristian Rogelio Espinoza Diaz - A01702752

Anatanael Jesus Miranda Faustino - A01769232

04/10/2023

1. Overview

El siguiente documento es el análisis de los datos proporcionados por General Electric y el desarrollo de la extracción de los mismos.

El documento se divide en cuatro fases: colección, descripción, exploración y reporte de calidad de los datos

2. Carga de datos

Reto Datos

3. Colección de datos

Ubicación de los datos:

Los datos fueron otorgados por el socio formador y obtenidos de una simulación de FADEC y del avión completo. El conjunto de datos está sincronizado en una plataforma en la nube de Microsoft conocida como Box, donde se encuentran los 2718 archivos. Estos archivos contienen todas las variables a analizar.

Método para obtener los datos:

A través de una simulación aérea, se generaron datos que fueron almacenados en el dispositivo FADEC, posteriormente, se extrajeron a formato csv, fueron ajustados a un rango de tiempo normalizado entre los diferentes archivos, por cuestiones de seguridad y de privacidad se volvieron anónimos y finalmente, se subieron a box sync.

Para acceder a los datos utilizamos el siguiente proceso:

-  PRO02. USO DE LA COMPUTADORA

Problemas para la obtención de datos:

- La mayor limitante es que los datos son propiedad de General Electric, se nos ha concedido permiso para acceder a ellos y manipularlos a efectos de nuestro análisis. Sin embargo, debido a acuerdos de confidencialidad, no podemos extraerlos por medios convencionales de internet.
- Debido a los acuerdos de confidencialidad, los datos sólo pueden ser manipulados y observados de manera física, local en la computadora proporcionada por GE en horario de clase.
- El internet del tecnológico de Monterrey tiene un firewall que no nos permite acceder a los servidores de las aplicaciones de General Electric, por esta razón, es necesario conectarse a internet a través de un router no administrado.
- Ya que la computadora está completamente gestionada por General Electric, estamos limitados a las aplicaciones de la nube privada de General Electric.

Solución a los problemas:

- Para el primer problema, designamos un tiempo específico en clase para configurar la computadora, designamos tareas y lo agregamos al plan de iteración.

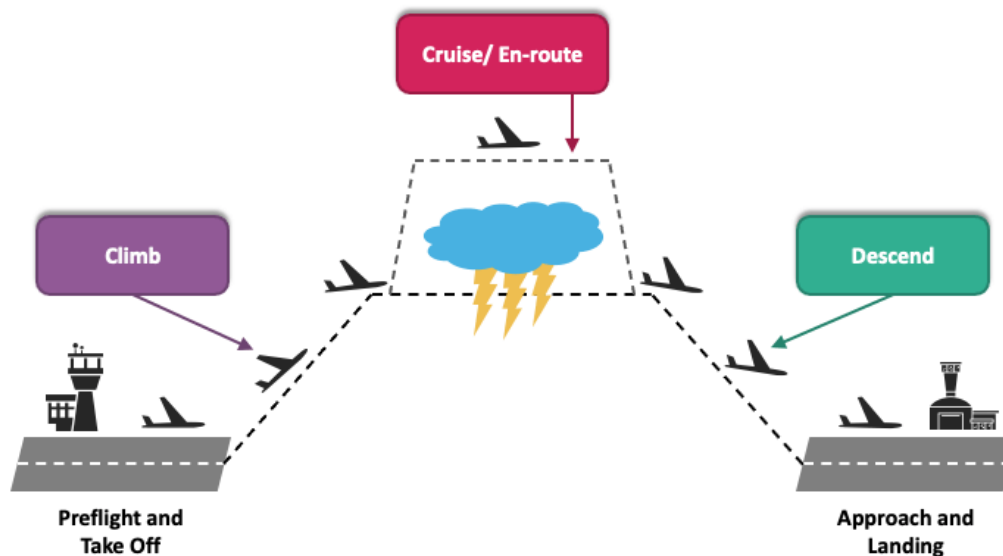
- Para el segundo problema, en las asignaciones del plan tenemos un horario específico para trabajar en la computadora, esto implica que hemos organizado nuestro tiempo y ajustado nuestra forma de trabajo.
- Para el tercer problema, hemos utilizado 2 métodos alternativos, el primero fue utilizar un hotspot de celular de algún miembro del equipo. El segundo método es emplear un router proporcionado por GE.
- Para el cuarto problema, hemos usado la página oficial con el catálogo de las aplicaciones para verificar que las aplicaciones que deseamos usar estén habilitadas. De lo contrario, se buscará algún reemplazo para dicha aplicación que se haya descubierto en el catálogo.

4. Descripción de datos

Los 2718 archivos csv de los datos pueden representar una o más fases de vuelo. Dichas fases de vuelo son:

FLIGHT PHASES

Aircraft Journey or Phases of Flight



En el documento [DiccionarioDatos.xlsx](#), se analizaron las 37 columnas que posee todo csv de General Electric. Toda columna está en formato numérico, por lo tanto, no hubo motivo para el uso de técnicas como One Hot Encoding para convertir variables categóricas a numéricas.

Dentro de ese archivo se encontrarán diversas estadísticas descriptivas como el promedio, su porcentaje de datos faltantes, su correlación con la variable dependiente conocida como “stableCruise_boolean”, etc.

El total de filas que se analizaron para esta etapa fue de 116,129,496 filas, al multiplicarlas por el número de columnas (37), nos da que el total de datos numéricos que se analizaron en esta fase fue de 4,292,097,352.

Evaluación de que los datos adquiridos satisfacen los requerimientos:

Por cuestiones de seguridad por parte del socio formador, General Electric, los datos fueron anonimizados. Por lo tanto, desconocemos qué representan los datos en el tema aeroespacial. No obstante, en una de las diversas pláticas que obtuvimos con el representante Ignacio Mendieta, surgió el comentario que con este set de datos el equipo de desarrollo del proyecto interno llegó a la posible solución de este problema.

En consecuencia, consideramos que los datos proporcionados por GE son suficientes para satisfacer los requerimientos.

5. Exploración de datos

1. Partimos los datos y solo usamos los de stable cruise con un valor 1
2. Sacamos estadísticas descriptivas de las columnas y las registramos en el diccionario de datos.
3. Hicimos un plot de correlación con stable cruise.
4. Análisis de multicolinealidad (Descartamos las columnas pares)
5. Hicimos gráficas de distribuciones.
6. Responder preguntas de minería e interpretar los resultados.
 - a. Evalúa los resultados
 - b. Determinar si existe una relación significativa entre las columnas
 - c. numéricas y StableCruise
 - d. Considerar si la relación es positiva o negativa y la magnitud de las diferencias
 - e. observadas

5.1 Descripción general de los datos

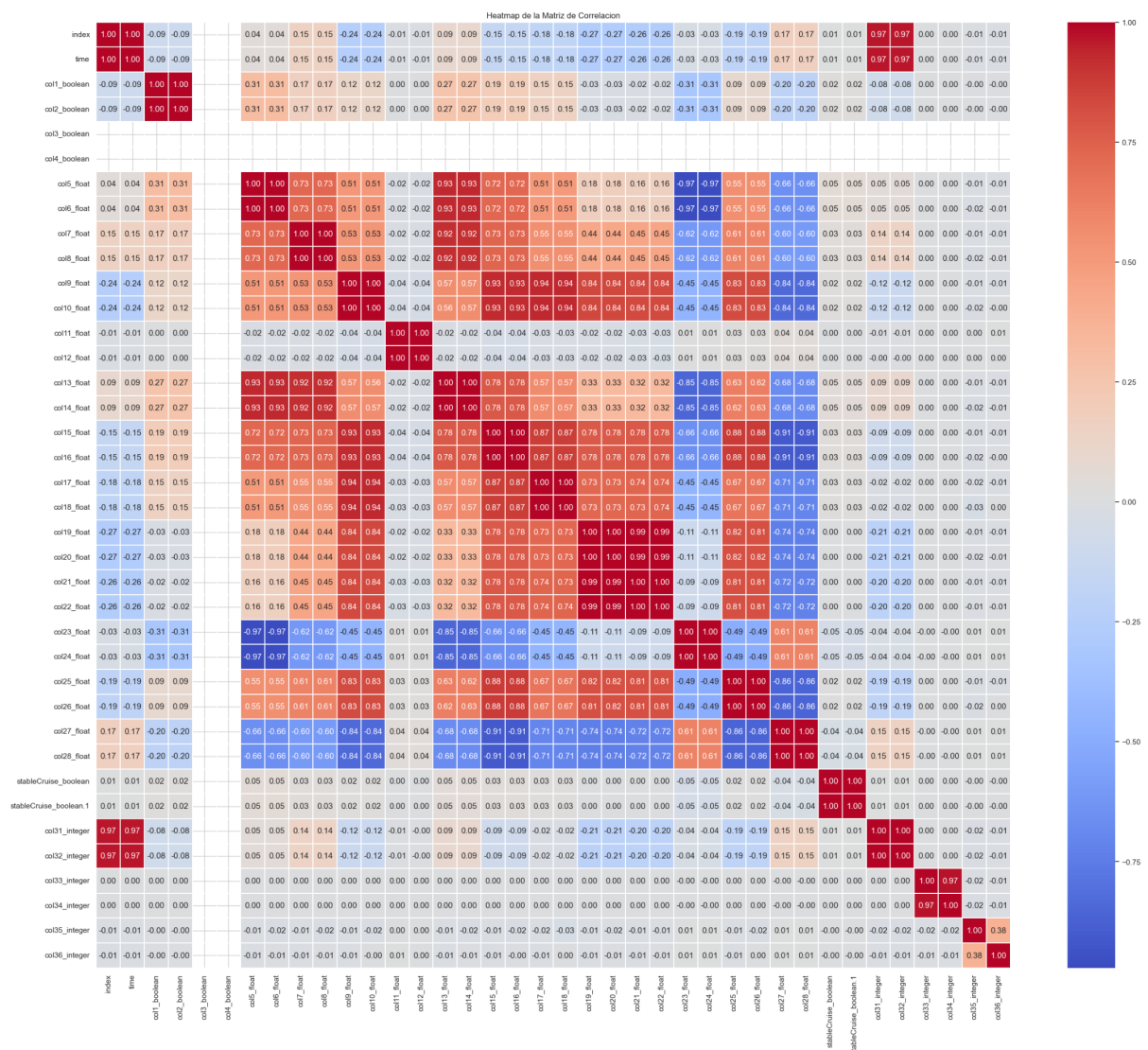
En primeros pasos, con la cantidad de 2,718 archivos CSV; cada uno de ellos tiene una cantidad de 37 columnas y un aproximado de 200,000 filas. Con estos datos, sin ninguna modificación por parte del equipo, realizamos el experimento de ejecutar una gráfica de correlación. El experimento terminó de ejecutarse en una hora y cuarenta minutos. Con la restricción de tiempo del equipo sobre la computadora proporcionada por el socio formador y el tiempo que un simple script tardaba en accionar, decidimos proceder con un análisis que implicaba restar el tiempo de la ejecución. Con nuestras opciones, decidimos recortar las columnas duplicadas (aquellas con el número par), por otro lado, inutilizar las simulaciones de vuelos que no tengan una estabilidad de crucero (archivos CSV que no tengan 1 en la columna stableCruise_boolean). Al proceder con estos cambios, el tiempo de ejecución del script fue acortado de una hora cuarenta a tan solamente veinte minutos. Al enfocarnos en los archivos que tienen estabilidad en el crucero podemos obtener de primera mano los cambios y las razones por la que un crucero es estable.

5.2 Estadísticas descriptivas

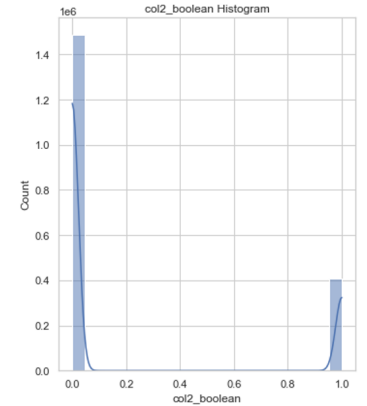
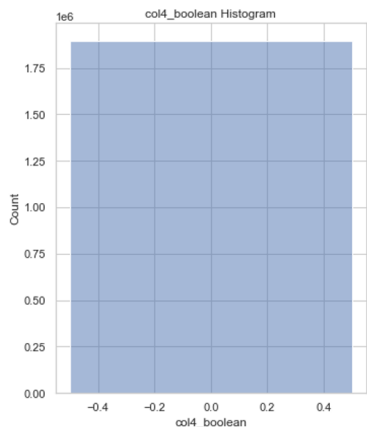
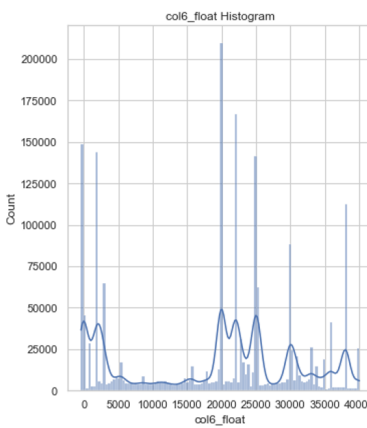
Para un mejor entendimiento de las variables, se decidió obtener estadísticas descriptivas básicas para tener un mejor entendimiento de estas variables. Se decidió registrar la siguiente información en el [X DiccionarioDatos.xlsx](#) : Como NaN, Porcentaje de NaN, Valor mínimo, Valor máximo, Promedio, Mediana, Varianza y la desviación estándar. Cabe mencionar que estas estadísticas descriptivas se hicieron sobre los datos filtrados que contenían Estabilidad de Crucero, no se hizo sobre todos los datos proporcionados. Gracias a este análisis comenzamos a sospechar que no era necesario incluir las columnas pares y que probablemente nos estábamos enfrentando a un problema de multicolinealidad en el caso de que las incluyamos. Otro descubrimiento relevante fue que las variables col3_boolean y col4_boolean no cambian a lo largo de los 800 archivos analizados.

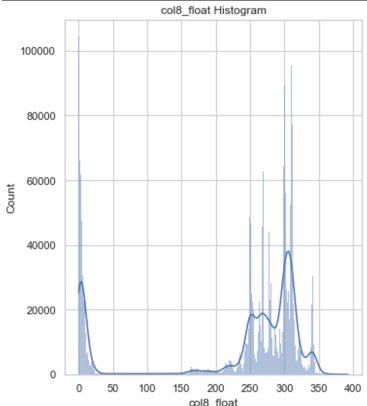
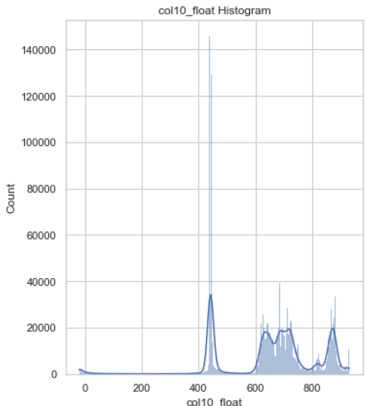
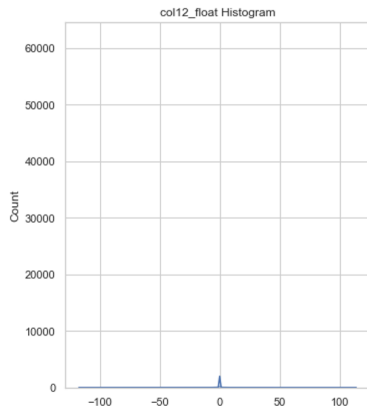
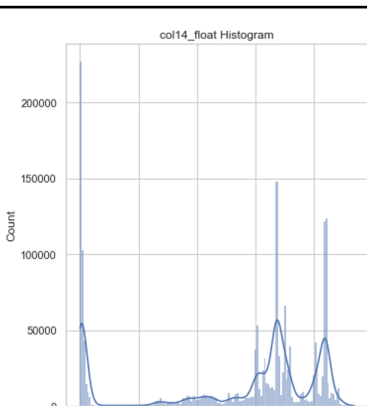
5.3 Correlación con “Stable Cruise”

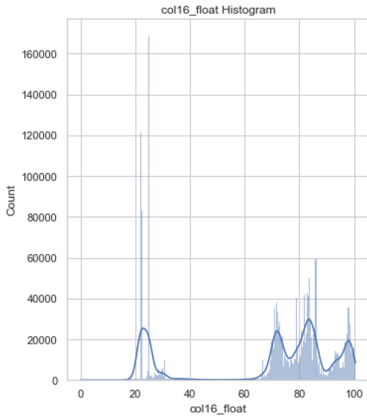
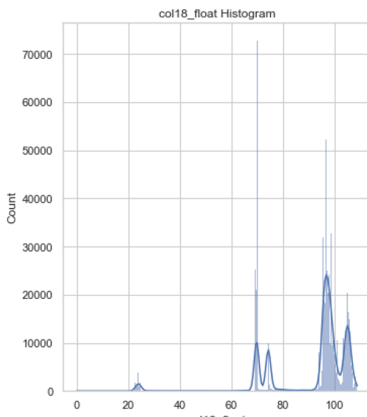
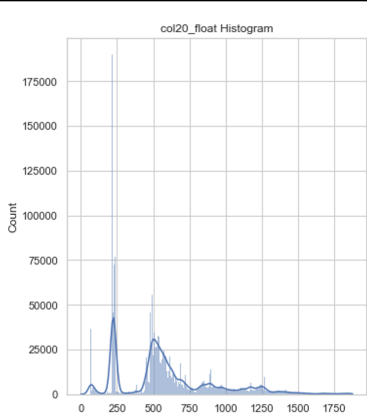
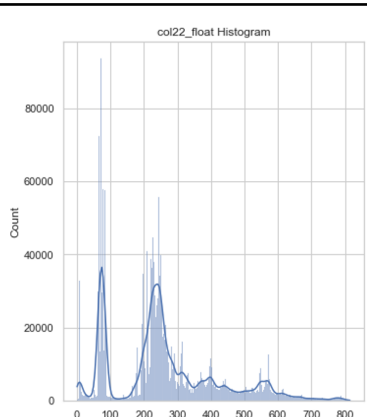
Para examinar la relación o asociación entre las variables realizamos una tabla de correlación. Su propósito principal es resumir la relación entre estas variables de una manera que sea fácil de entender y que ayude a identificar patrones o tendencias en los datos.

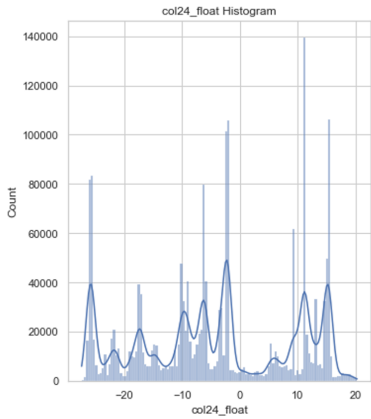
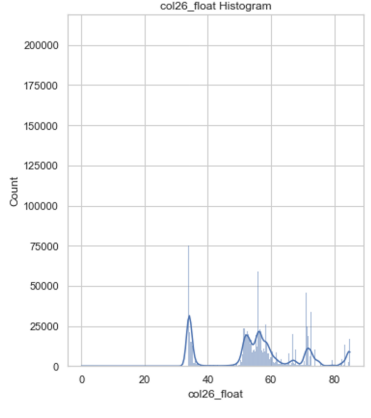
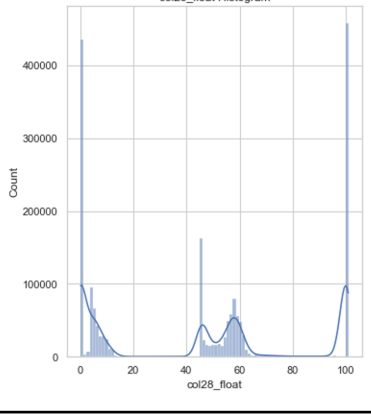
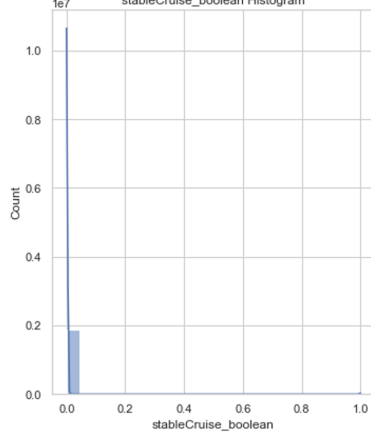


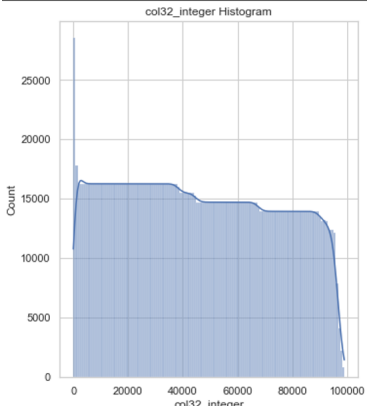
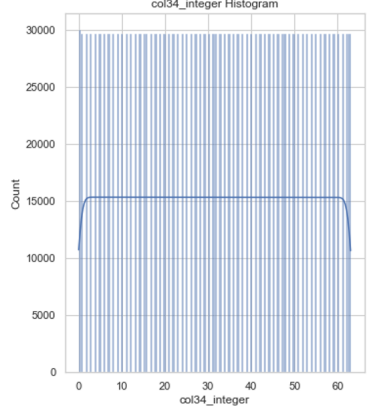
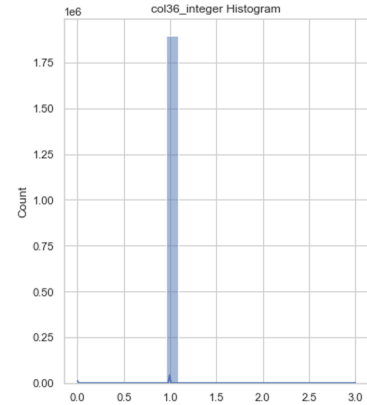
5.4 Gráficas de distribución

Variable	Descripción	Distribución
col2_boolean	<p>Frecuencia: Un millón cuatrocientos mil valores son negativos y cuatrocientos mil son positivos.</p> <p>Distribución: Predominan por mucho los valores negativos.</p> <p>Patrones: No hay ningún patrón aparente por el tipo de dato.</p> <p>Conclusión: Si se llega a utilizar esta variable deberíamos tener una distribución más equilibrada.</p>	 <p>The histogram for col2_boolean shows a distribution that is heavily skewed towards 0.0. The y-axis represents the count, scaled by 1e6, ranging from 0.0 to 1.4. The x-axis represents the value of col2_boolean, ranging from 0.0 to 1.0. There is a very high frequency of values near 0.0, with a count exceeding 1.4 million. A much smaller frequency is observed near 1.0, with a count around 0.4 million.</p>
col4_boolean	<p>Frecuencia: Un millón ochocientos mil valores son negativos.</p> <p>Distribución: Se conforma completamente por valores negativos</p> <p>Patrones: Sólo tiene valores negativos</p> <p>Conclusión: Esta variable no aporta información para nuestro modelo debido a la falta de una distribución normal.</p>	 <p>The histogram for col4_boolean shows a distribution entirely concentrated on negative values. The y-axis represents the count, scaled by 1e6, ranging from 0.00 to 1.75. The x-axis represents the value of col4_boolean, ranging from -0.4 to 0.4. The distribution is very narrow and centered around -0.4, with a count reaching approximately 1.75 million.</p>
col6_float	<p>Distribución: No tiene una distribución clara.</p> <p>Patrones: No posee un patrón claro</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>The histogram for col6_float shows a complex, multi-modal distribution. The y-axis represents the count, ranging from 0 to 200,000. The x-axis represents the value of col6_float, ranging from 0 to 40,000. The distribution is highly irregular, with many sharp peaks and valleys. The highest peak is around 20,000, with a count exceeding 200,000. Other significant peaks are visible at approximately 5,000, 15,000, 25,000, and 35,000.</p>

col8_float	<p>Frecuencia: Los valores van de 0 a 415, de los cuales la mayoría están en 300.</p> <p>Distribución: Tiene una distribución cargada a la derecha.</p> <p>Patrones: La mayoría de los valores se encuentran entre los 300</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>col8_float Histogram</p>
col10_float	<p>Frecuencia: Los valores van de 0 a 800, de los cuales la mayoría están en 600.</p> <p>Distribución: Tiene una distribución cargada a la derecha.</p> <p>Patrones: Outliers del 0 a 350. Tiene un gran pico de valores a los 430</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>col10_float Histogram</p>
col12_float	<p>Frecuencia: Los valores van de -118 a 113, de los cuales la mayoría están en 0.</p> <p>Distribución: A pesar de los valores extremos, podría tener una distribución normal.</p> <p>Patrones: Se compone mayormente por valores extremos</p> <p>Conclusión: Esta variable se debe analizar con más detalle. Eliminando los outliers y volviendo a graficar para encontrar más patrones en esta variable</p>	 <p>col12_float Histogram</p>
col14_float	<p>Frecuencia: Los valores van de 0 a 1.18.</p> <p>Distribución: Tiene una distribución cargada hacia los extremos.</p> <p>Patrones: La mayoría de los valores se encuentran entre el 0. Los valores suelen ser cargados hacia los extremos</p> <p>Conclusión: Esta variable se debe analizar con más detalle para identificar si el comportamiento de la distribución de valores es necesario modificarla o no.</p>	 <p>col14_float Histogram</p>

col16_float	<p>Frecuencia: Los valores van de 0 a 102.</p> <p>Distribución: Tiene una distribución cargada a la derecha.</p> <p>Patrones: La mayoría de los valores se encuentran entre el 76</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>col16_float Histogram</p> <p>The histogram shows the frequency distribution of col16_float. The x-axis ranges from 0 to 100, and the y-axis (Count) ranges from 0 to 160,000. There is a prominent peak around 25 and a long tail extending towards 100.</p>
col18_float	<p>Frecuencia: Los valores van de 0 a 140.</p> <p>Distribución: Tiene una distribución cargada a la derecha.</p> <p>Patrones: La mayoría de los valores se encuentran entre el 97</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>col18_float Histogram</p> <p>The histogram shows the frequency distribution of col18_float. The x-axis ranges from 0 to 140, and the y-axis (Count) ranges from 0 to 70,000. There is a prominent peak around 70 and a long tail extending towards 140.</p>
col20_float	<p>Frecuencia: Los valores van de 0 a 1877.</p> <p>Distribución: Tiene una distribución cargada a la izquierda.</p> <p>Patrones: La mayoría de los valores se encuentran entre el 510</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>col20_float Histogram</p> <p>The histogram shows the frequency distribution of col20_float. The x-axis ranges from 0 to 1877, and the y-axis (Count) ranges from 0 to 175,000. There is a prominent peak around 250 and a long tail extending towards 1877.</p>
col22_float	<p>Frecuencia: Los valores van de 0 a 819.</p> <p>Distribución: Tiene una distribución cargada a la izquierda.</p> <p>Patrones: La mayoría de los valores se encuentran entre el 216</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>col22_float Histogram</p> <p>The histogram shows the frequency distribution of col22_float. The x-axis ranges from 0 to 819, and the y-axis (Count) ranges from 0 to 80,000. There is a prominent peak around 100 and a long tail extending towards 819.</p>

col24_float	<p>Frecuencia: Los valores van de -41 a 24.</p> <p>Distribución: No tiene una distribución clara.</p> <p>Patrones: La mayoría de los valores se encuentran entre el -7</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>col24_float Histogram</p>
col26_float	<p>Frecuencia: Los valores van de 0 a 85.</p> <p>Distribución: Tiene una distribución cargada a la derecha.</p> <p>Patrones: La mayoría de los valores se encuentran entre el 55. Tiene valores extremos a la izquierda de la gráfica</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>col26_float Histogram</p>
col28_float	<p>Frecuencia: Los valores van de 0 a 101.</p> <p>Distribución: No tiene una distribución clara.</p> <p>Patrones: La mayoría de los valores se encuentran entre el 47</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>col28_float Histogram</p>
stableCruise_boolean	<p>Distribución: La distribución se concentra mayormente en los falsos.</p> <p>Patrones: La proporción de positivos es uno de cada 300.</p> <p>Conclusión: Probablemente tenemos que recortar el dataset para tener unos datos más equilibrados.</p>	 <p>stableCruise_boolean Histogram</p>

col32_integer	<p>Frecuencia: Los valores van de 0 a 200015.</p> <p>Distribución: No tiene una distribución clara.</p> <p>Patrones: La mayoría de los valores se encuentran entre el 75881</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>The histogram for col32_integer shows a distribution starting at 0 and ending at 100,000. The y-axis represents the count, ranging from 0 to 25,000. The distribution is relatively flat between 20,000 and 80,000, with a slight peak around 10,000 and a sharp drop-off after 80,000.</p>
col34_integer	<p>Frecuencia: Los valores van de 0 a 63.</p> <p>Distribución: No tiene una distribución clara.</p> <p>Patrones: La mayoría de los valores se encuentran entre el 31. La distribución de datos parece ser similar, oscilando en los mismos valores.</p> <p>Conclusión: Esta variable se debe analizar con más detalle.</p>	 <p>The histogram for col34_integer shows a distribution from 0 to 60. The y-axis represents the count, ranging from 0 to 30,000. The distribution is very narrow, with most values concentrated between 0 and 10, and a sharp peak around 31.</p>
col36_integer	<p>Frecuencia: Los valores van de 0 a 3.</p> <p>Distribución: No tiene una distribución clara.</p> <p>Patrones: La mayoría de los valores se encuentran entre el 1. La mayoría de los valores son 1.</p> <p>Conclusión: Esta variable se debe analizar con más detalle. Se debe analizar los casos en los que es diferente a 1.</p>	 <p>The histogram for col36_integer shows a distribution from 0.0 to 3.0. The y-axis represents the count, ranging from 0.00 to 1.75e6. The distribution is extremely narrow, with a single sharp peak at 1.0.</p>

5.6 Necesidad de datos adicionales

Con la información que hemos recolectado anteriormente podemos identificar que las columnas que tenemos son suficientes para cumplir con el objetivo de minería de datos.

Al conocer que los datos proporcionados por GE son específicos de una simulación de vuelo con datos almacenados por FADEC, sabemos que no se pueden agregar más datos que puedan apoyar en encontrar la relación de las variables con el stableCruise_boolean.

Las columnas donde hemos encontrado una relación directa con stable cruise son:

col23_float	Esta columna float tiene un valor de -0.05 en la tabla de correlación.
col13_float	Esta columna float tiene un valor de 0.05 en la tabla de correlación.
col5_float	Esta columna float tiene un valor de .05 en la tabla de correlación.
col7_float	Esta columna float tiene un valor de 0.03 en la tabla de correlación.
col27_float	Esta columna float tiene un valor de -0.04 en la tabla de correlación.
col15_float	Esta columna float tiene un valor de 0.03 en la tabla de correlación.
col17_float	Esta columna float tiene un valor de 0.03 en la tabla de correlación.

5.6 Interpretación de resultados

- ¿Qué características son más influyentes para la estabilidad de crucero?

Con la intención de contestar esta pregunta hicimos una investigación de herramientas para relacionar las columnas con estabilidad de crucero (stableCruise_boolean). Lo que encontramos posterior a la investigación fue la herramienta chi-cuadrado (χ^2). Con esta herramienta podemos evaluar si existe una relación significativa entre dos variables.

Para utilizar esta herramienta, creamos un script que, con stable cruise como la principal variable, identificara las relaciones de las demás columnas con respecto a ella. El siguiente es un ejemplo de la ejecución de código:

```

1 Variable booleana: col1_boolean
2 Estadístico de chi-cuadrado: 59995.98241015045
3 Valor p: 0.0
4 Grados de libertad: 1
5 Frecuencia esperada:
6 [[1.03567710e+08 3.32913686e+05]
7  [1.19903127e+07 3.85423138e+04]]
8 -----
9 La variable col1_boolean está relacionada de manera significativa con stableCruise_boolean.
10 -----
11 Variable booleana: col2_boolean
12 Estadístico de chi-cuadrado: 59996.188318692715
13 Valor p: 0.0
14 Grados de libertad: 1
15 Frecuencia esperada:
16 [[1.03567726e+08 3.32913737e+05]
17  [1.19902967e+07 3.85422626e+04]]
18 -----
19 La variable col2_boolean está relacionada de manera significativa con stableCruise_boolean.
20 -----
21 Variable booleana: col3_boolean
22 Estadístico de chi-cuadrado: 0.0
23 Valor p: 1.0
24 Grados de libertad: 0
25 Frecuencia esperada:
26 [[1.15558023e+08 3.71456000e+05]]
27 -----
28 No hay evidencia significativa de relación entre col3_boolean y stableCruise_boolean.
29 -----
30 Variable booleana: col4_boolean
31 Estadístico de chi-cuadrado: 0.0
32 Valor p: 1.0
33 Grados de libertad: 0
34 Frecuencia esperada:
35 [[1.15558023e+08 3.71456000e+05]]
36 -----
37 No hay evidencia significativa de relación entre col4_boolean y stableCruise_boolean.
38 -----

```

- ¿Se observan diferencias estadísticamente significativas entre los datos registrados con redundancia?

Con el diccionario de datos que hemos creado, identificamos si las columnas pares eran efectivamente columnas duplicadas o tenían datos diferentes.

 DiccionarioDatos.xlsx

Como podemos observar en el diccionario, efectivamente, no existen datos distintos en las filas de las columnas duplicadas.

6. Reporte de calidad de los datos

Se realizaron análisis de las 37 columnas de información presentes en cada archivo CSV. Es importante destacar que, en la fase de exploración de datos, se optó por reducir el número de archivos a 737.

Esta decisión se basó en la presencia de un valor igual a 1 en al menos una fila de las columnas relacionadas con “Stable Cruise.” En otras palabras, se aplicó la técnica de submuestreo por mayoría para obtener una representación más concisa pero aún significativa de los datos. Además, se busca tener un número equivalente de datos positivos y negativos de stable cruise, sin embargo, cómo se observó en el diccionario de datos 1 de cada 300 datos contienen positivo el stable cruise. Por estas razones, no se descarta que se reduzca aún más el dataset.

El submuestreo por mayoría es una técnica empleada en la gestión de conjuntos de datos desequilibrados, en los cuales una de las clases (la clase mayoritaria) cuenta con un número considerablemente mayor de ejemplos en comparación con la otra clase (la clase minoritaria).

El objetivo principal de esta técnica es equilibrar las proporciones de clases en el conjunto de datos al eliminar estratégicamente algunos ejemplos de la clase mayoritaria.

Las ventajas de aplicar el submuestreo por mayoría son diversas:

- Contribuye a abordar el desafío del desequilibrio de clases en conjuntos de datos, lo que puede llevar a mejoras en el rendimiento de los modelos de aprendizaje automático.
- A diferencia del submuestreo aleatorio, el submuestreo por mayoría se caracteriza por su selectividad en la elección de ejemplos a eliminar, minimizando así la pérdida de información valiosa.
- Esta técnica puede potenciar la capacidad del modelo para aprender patrones en la clase minoritaria al reducir la influencia abrumadora de la clase mayoritaria.

Otro motivo fue el tiempo, ya que usar todos los archivos para esta fase de “Data Understanding” era muy costoso.

No obstante, esto no quiere decir que hayamos descartado los 1981 CSV para las siguientes fases de la creación y evaluación del modelo. Al ejecutar este código obtuvimos diversos hallazgos significativos que se registraron en `DiccionarioDatos.xlsx` como:

- Ninguna columna posee datos faltantes conocidos como NaN 's.
- Se confirma que los datos hay duplicados a excepción de la columna “time”
- Se confirma que los archivos van de 0 a 3000 (medida de tiempo a confirmar) en la columna “time”
- Las dos columnas con etiqueta “StableCruise_boolean”

Teniendo en cuenta todo lo anterior, podemos concluir que hasta el momento es posible lograr los objetivos de minería de datos, por lo tanto, podemos proseguir a la siguiente fase.

7. Forma de trabajo en Github

EST01. ESTÁNDAR DE BRANCHES

EST02. ESTANDAR DE COMMIT