**Análisis de Datos - ETL**

Tenemos 8 tablas

bookings

flights

tickets

tickets\_flights

boarding\_passes

seats

aircrafts\_data

airports\_data

***Conclusión***:

Las tablas contiene informacíón de un solo mes completo: Julio 2017, el final de junio 2017 y principios de agosto 2017.

En la tabla de flights , la columnas actual\_departure y actual\_arrival tienen un 50% de valores ‘/N’. Se asume que son vuelos que están en horario. Del resto de las tablas, ninguna tabla presenta nulos que haya que rellenar, ni filas duplicadas. En la de bookings, que tiene campo numérico, encontramos un 2% de outliers, pero consideramos que son eventos puntuales de la operatoria normal de la empresa. De ser necesario, estos outliers se identificarán al momento de hacer el ETL del Dashboard en PBI.

Las tablas de boarding\_passes, ticket\_flights y seats no tienen identificador único. Hay que crearles un id.

* **Bookings**
* Shape

Filas, columnas

(262788, 3)

* columnas

['book\_ref', 'book\_date', 'total\_amount']

* Describe de columna total\_amount

Sum 20,766,980,900

|  |  |
| --- | --- |
| **total\_amount** |  |
| count | 262,788.00 |
| mean | 79,025.61 |
| std | 77,621.92 |
| min | 3,400.00 |
| 25% | 29,000.00 |
| 50% | 55,900.00 |
| 75% | 99,200.00 |
| max | 1,204,500.0 |

Llama la atención su rango y su desviación std respecto a la media.

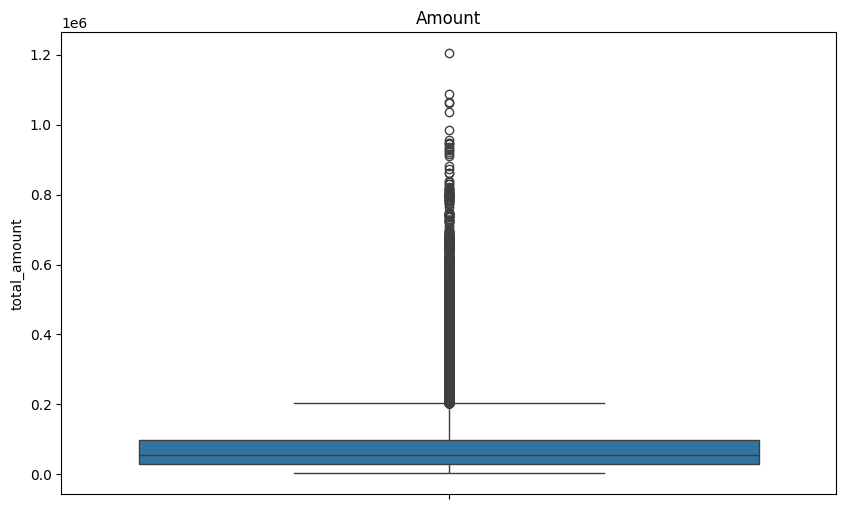
* Nulos

La tabla no presenta nulos.

Sin embargo llama la atención su rango y su desviación std.

* Outliers

De 262.788 registros que tiene la base , 5.141 se consideran outliers porque están a 3 desviaciones standard de la media, y es casi un 2% de la base.

****

Describe de total\_amount outliers (5.141 registros)

count 5,141.00

**mean 428,091.34**

std 96,236.15

min 311,900.00

25% 369,000.00

50% 398,600.00

75% 476,000.00

max 1,204,500.00

Describe de total\_amount not\_outliers (257.647 registros)

count 257,647.00

**mean 72,060.47**

std 58,998.39

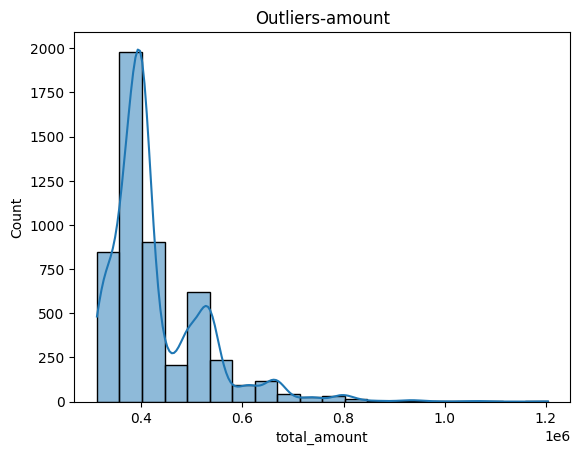
min 3,400.00

25% 28,800.00

50% 55,800.00

75% 94,200.00

max 311,700.00



No se corregirán los outliers por ser eventos particulares de la operatoria normal de la empresa. De ser necesario, se agregará una columna al momento de realizar el dashboard para ´poder identificarlos y filtrarlos.

* Duplicados - No hay
* **Tickets**
* Shape

Filas, columnas

(366733, 3)

* Columnas

'ticket\_no', 'book\_ref', 'passenger\_id'

* No hay columnas numéricas para analizar sus datos estadísticos
* Nulos

La tabla no presenta nulos.

* Outliers. No hay columnas numéricas.
* Duplicados - No hay
* Boarding\_passes
* Shape

Filas, columnas

(579686, 4)

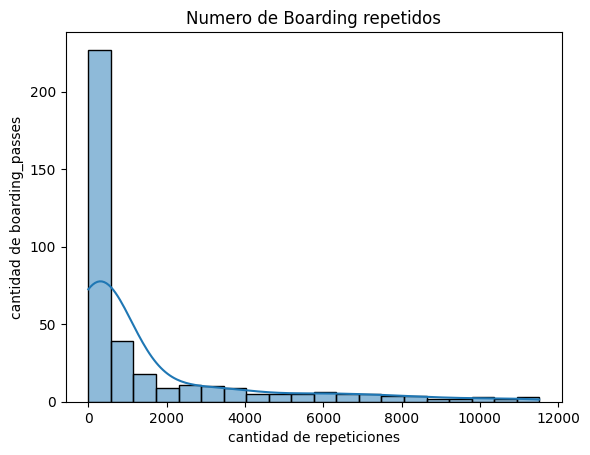
* Columnas

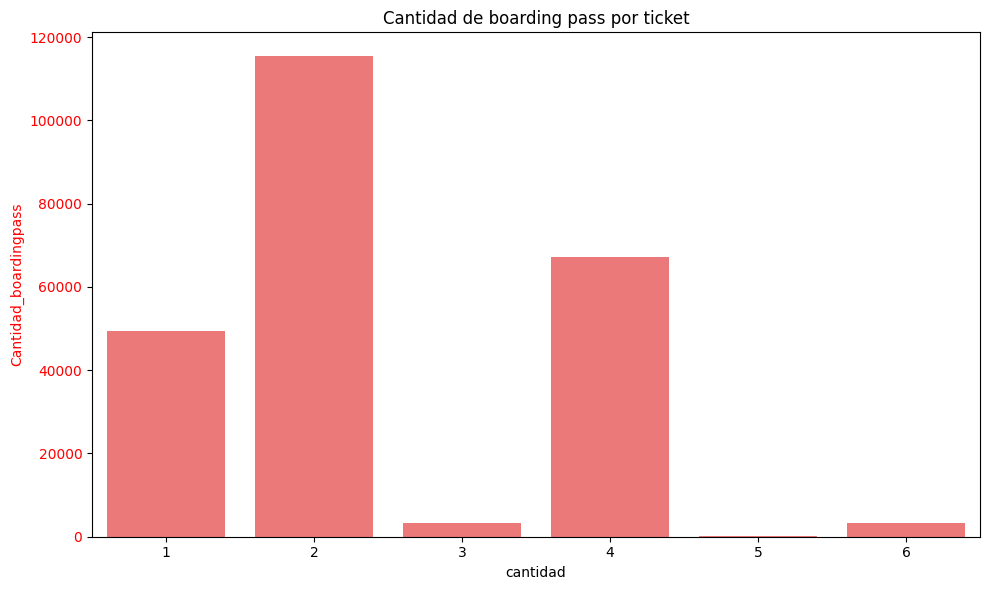
'ticket\_no', 'flight\_id', 'boarding\_no', 'seat\_no'

* No hay columnas numéricas para analizar sus datos estadísticos
* Nulos

La tabla no presenta nulos.

* Outliers. No hay columnas numéricas.
* Duplicados - No hay
* La tabla presenta mismo 'boarding\_no' para diferentes tickets. Hay más de un 'boarding\_no' por ticket, coincidente con la cantidad de vuelos asociados a ese ticket.





* **Ticktes\_flights**
* Shape

Filas, columnas

(1045726, 4)

* Columnas

'ticket\_no', 'flight\_id', 'fare\_conditions', 'amount'

* Describe de campo ‘amount’

Sum 20,766,980,900

count 1,045,726.00

mean 19,858.91

std 22,612.39

min 3,000.00

25% 7,200.00

50% 13,400.00

75% 23,100.00

max 203,300.00

* Duplicados - No hay
* **Fights**
* Shape

Filas, columnas

(33121, 10)

* Columnas

‘Flight\_id’, ‘flight\_no’, ‘scheduled\_departure’, ‘scheduled\_arrival’, ‘departure\_airport’, ‘arrival\_airport’, ‘status’, ‘aircraft\_code’, ‘actual\_departure’, ‘actual\_arrival’.

* Duplicados - No hay
* Nulos . Hay un 50% de los campos ‘actual\_departure’, ‘actual\_arrival’ en ‘/N’