Exploracion 4 DF ID\_VISITANTE - Ana

Tratamiento de datos al hacer un merge de las 4 tablas de t\_id y id\_visitante

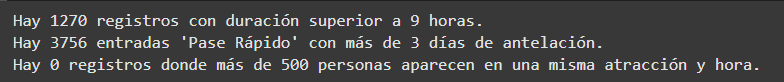
Al hacer un analisis exploratorio vemos que tenemos varios problemas

#### **Problemas:**

* duracion: mínimo de **-39**, lo cual no tiene sentido.
* coste: mínimo de **-2.25**, no debería haber precios negativos.
* tiempo\_de\_espera: mínimo de **-3**, también inválido.
* comienzo\_atraccion: tiene valores negativos

Filtraremos los valores inválidos negativos.

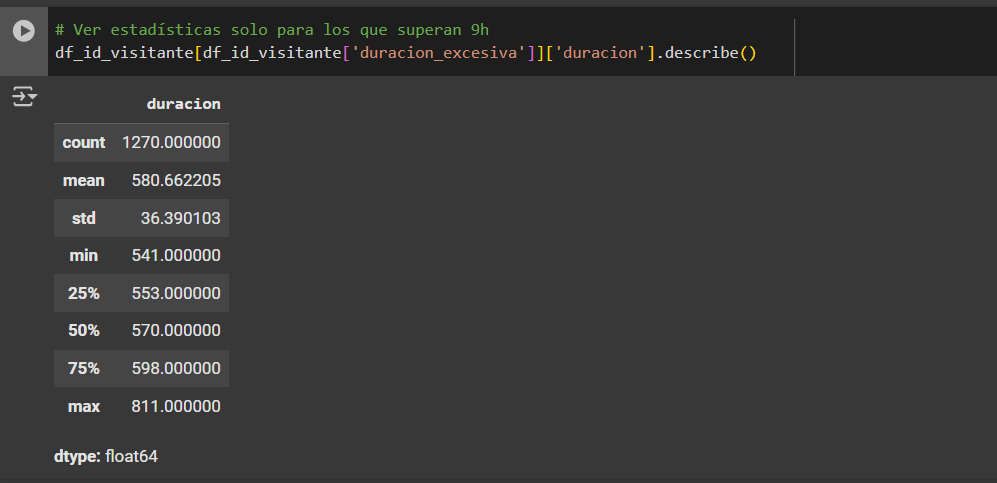
Aplicamos las reglas de negocio con el siguiente resultado



**Miraremos los datos con más detalles:**

**Registros de duración superior a 9 hr:**

Haciendo un analisis estadistico podemos ver que:



La **media es 580 min (~9h 40min)** → está muy cerca del límite

La mitad no se pasa más de 30 min

|  |
| --- |

| El 75% está por debajo de 10 horas |
| --- |

El **máximo es 811 min (~13h 30min)** → eso sí es un outlier claro

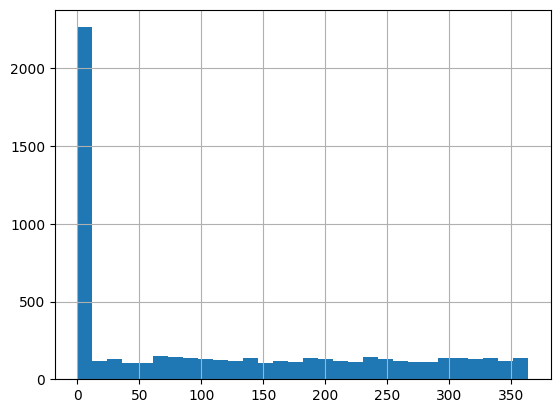
El **mínimo es 541 min** → solo 1 minuto por encima

Decision:

Parece más una desviación natural, error menor de sistema o redondeo, salvo algunos casos extremos

Mantendremos parte de los registros : hasta 1hr por encima de las 9 hr, considerándolos como errores leves

Valores de FAST PASS



Confirmamos con este hist visualmente que es un error. Se ve que la distribucion de los dias es sistematica e igual, y no tiene sentido. La mayoría están entre 4 y 365 con distribución artificial

Son muchos registros para eliminarlos (más de 3.700 registros). El error es **claramente sistemático** y se repite asi que vamos a corregir estos dias a 3 para no perder los registros.

El tema de los duplicados

t\_id representa una foto, y cada foto solo debería estar asociada a un visitante, entonces hay un problema de datos. No puede haber dos visitantes con la misma imagen. En ese caso:

* Uno de los registros puede estar mal asignado
* O hay un error en la generación del t\_id en el sistema del negocio



**Tenemos dos pares de registros** donde el campo t\_id (nombre de la foto) se repite, pero el resto de los datos **NO son iguales**.

Esto puede significar:

1. La **misma foto** fue tomada a **más de una persona o grupo** (ej. dos personas en la misma atracción)
2. Un **error de etiquetado o asignación** del t\_id - mas probable
3. Una **duplicación técnica** en la carga de datos - mas probable

Conclusion: los registros son diferentes pero válidos. Así que **no vamos a eliminar por ahora**.

**Valores unicos**

**Atracciones**: los nombres parecen correctos y no hay tildes o repeticiones: hay 35

No hay errores ortográficos evidentes ni duplicados

**Procedencia:** Los nombres son correctos y no hay repeticiones. Hay 34 paises

Tipo de entrada: Los nombres correctos y 6 tipos de pases

Hemos observado que:

Total de visitantes únicos: 1785

Visitantes con múltiples procedencias: 1742

* Tienes 1.785 visitantes únicos (id\_visitante).
* 1.742 de ellos (¡el 98%!) tienen múltiples procedencias asignadas.

Esto no tiene sentido lógico si la procedencia representa el país de origen del visitante. En un dataset coherente, cada visitante debería tener una única procedencia estable.

Pare eso vamos a aplicar la MODA , extrayendo la procedencia mas frecuente asociada a cada ID\_Visitante, de esa maner obtenermos un valor asignado unico.

Procedencia es el que tiene mas sentido aqui que tenga un valor asignado a cada ID\_visitante. A diferencia de la procedencia, donde esperas un solo valor por visitante, en duración, atraccion, ticket es perfectamente normal que varíe.

Vamos a ver visitantes unicos/dia y sus estadisticas



Es totalmente coherente con:

* Visitantes que **vuelven varios días** (con pase anual, por ejemplo).
* Visitantes que **entran y salen varias veces en el mismo día** (si se cuenta por ticket).
* **Varias fotos**  por visitante.(35K aprox registros)

**Vamos a analizar el COSTE**

**Utilizaremos el IQR (rango intercuartílico) para ver entre que valores deberia estar nuestro coste:**

**Al ejecutar, tenemos**

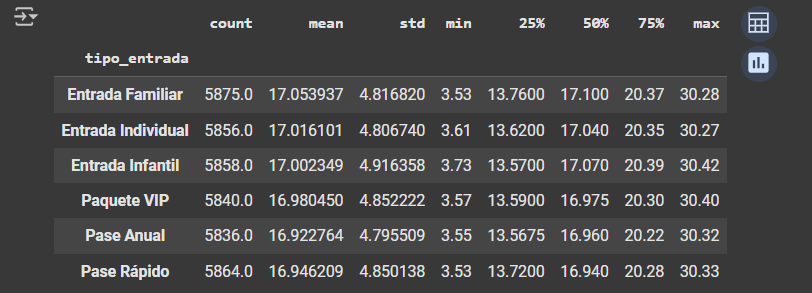
3.5217389060147433

30.434612667293717

0.61% son outliers de coste

Tenemos 217 outliers. Los vamos a eliminar ya que representan muy poco de nuestros datos





Aunque ciertos valores del campo coste o antelacion de compra parecen poco realistas (como entradas infantiles más caras que un pase anual), he decidido mantener la variabilidad del dataset sin imponer una lógica externa

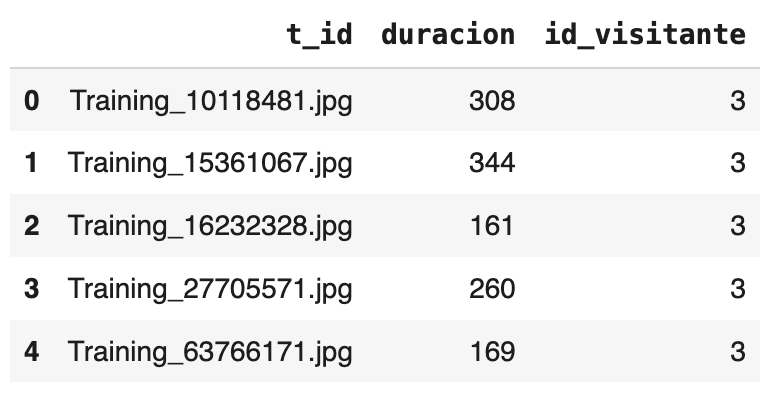
Exploracion individual Ines

🧭 Exploración de id\_visitante\_duracion

# 🧭 Exploración de id\_visitante\_duracion.json

### ⌗ Sobre los datos

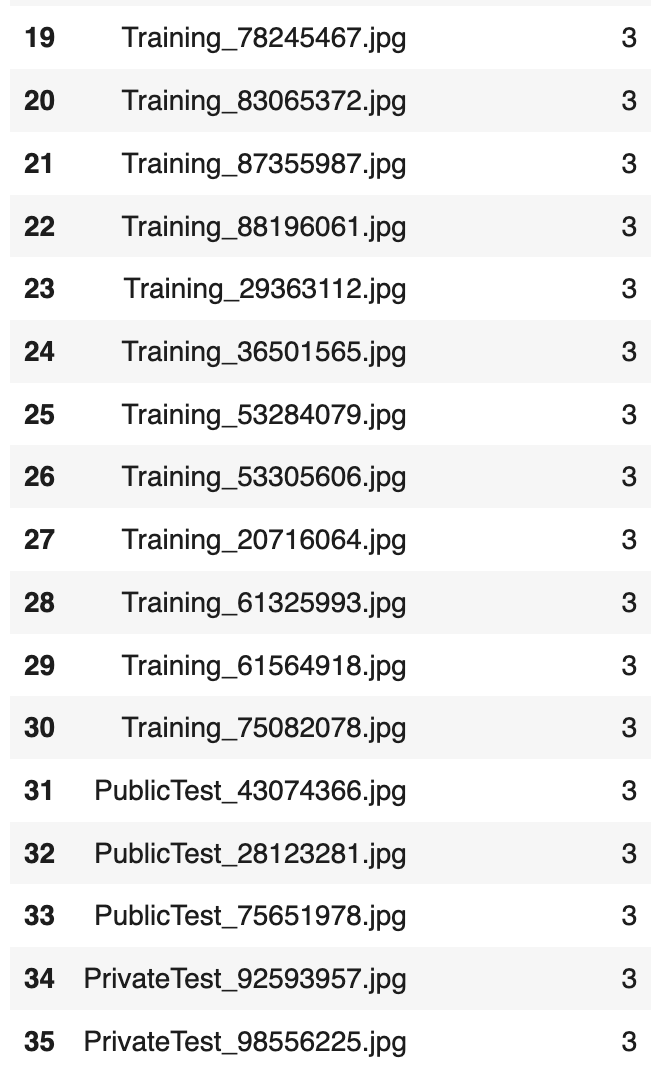
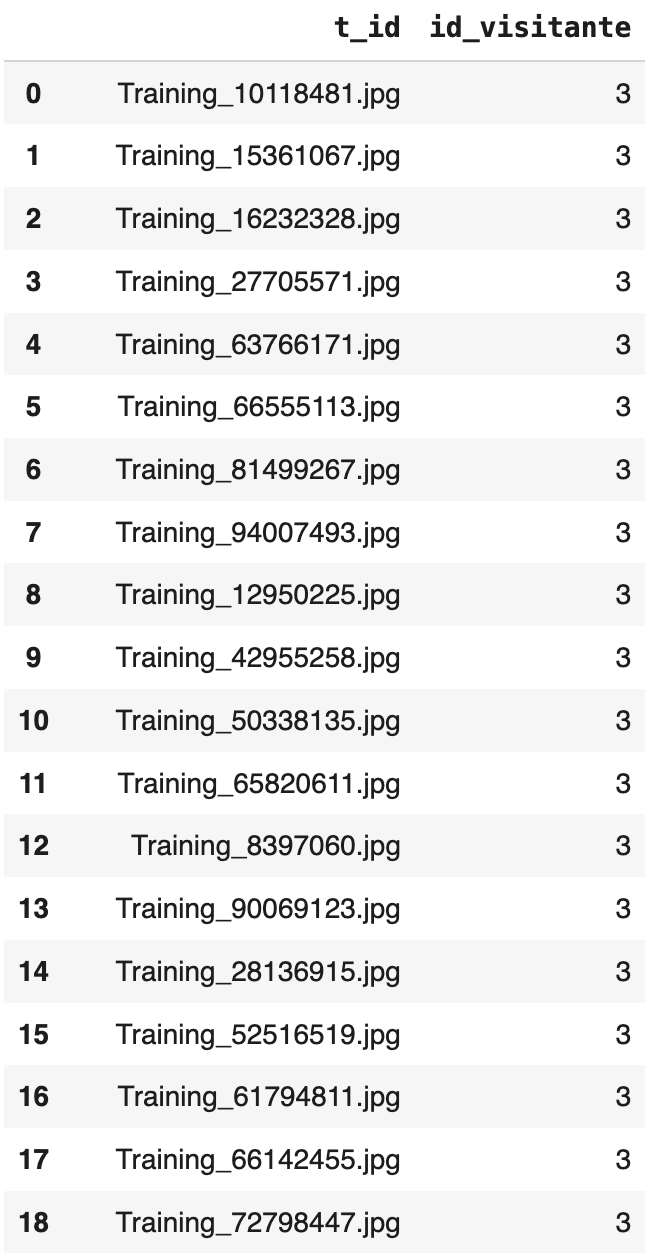
Haciendo un head vemos que la duración está asociada a MULTIPLES fotografias y SOLO A UN id\_visitante.



### ❌ Problema

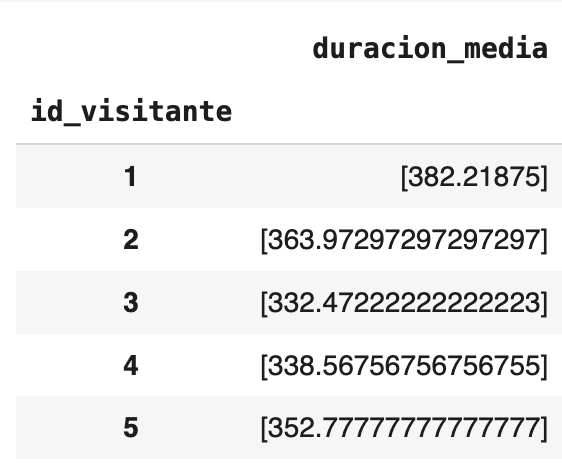
* ¿Por qué hay una duración asociada a cada imagen?
* ¿Por qué los id\_visitante repetidos no tienen la misma duración siempre?

Podemos entender que cada visitante tenga varias fotos, pero no varias duraciones, ya que estas no coinciden. Y no se pueden asociar a un día concreto porque hay visitantes que tendrían más de 30 días. Vamos a ver el ejemplo del visitante 3:



### ✅Posibles soluciones

* Hacer una media de duración por visitante:



* ¿Se puede tener en cuenta que ‘duración’ sea otra cosa y esté mal categorizado?
* ¿Puede ser el id\_visitante erróneo?

### ❌ Problema

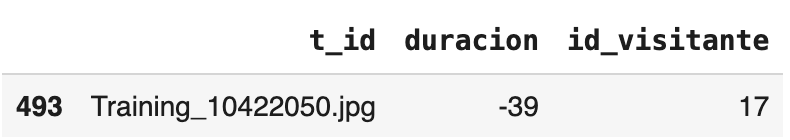
Hay 1279 datos que superan las 9 horas, que supone el 3.56% del df\_duracion

### ✅Posibles soluciones

* Borrar esas filas.
* Dejar aquellos valores que duran menos de 10h y considerarlos como errores del sistema.

### ❌ Problema

Hay un valor negativo:

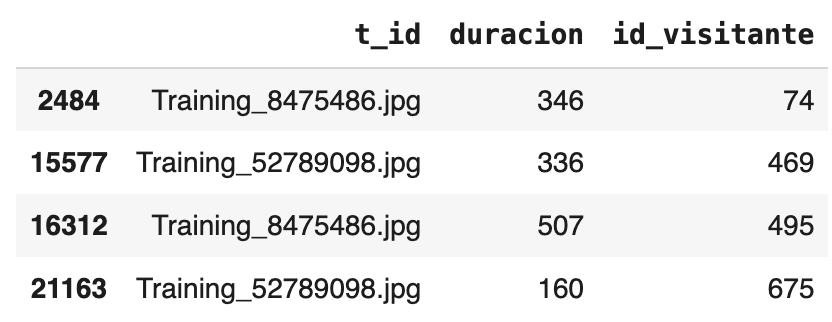


### ✅Posibles soluciones

* Borrar esa fila.
* Sustituir ese valor con la media de duración de ese id\_visitante - de todas maneras ese negativo no harbia que considerarlo en la media ya que es un claro outliner

### ❌ Problema

Hay dos duplicados, que son los mismos que en el resto de json:



### ✅Posibles soluciones

* Borrar esas filas.
* Como son dos identificadores diferentes, podemos añadirles el id\_visitante detrás del nombre para que sean únicos.

🎫 Exploración de id\_visitante\_ticket

# 🎟️ Exploración de id\_visitante\_ticket.json

### ⌗ Sobre los datos

Haciendo un head vemos que tenemos estos datos.



### ❌ Problema

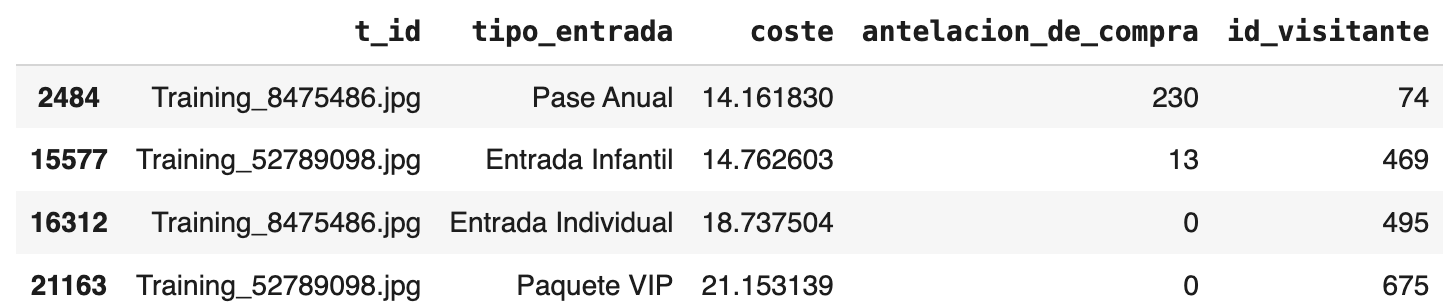
* Hay 3780 entradas del TOTAL vendidas con más de 3 días de antelación, que supone el 10.53% del df\_ticket
* Esto supone que un 63.15% de las entradas de Pase Rápido se hayan vendido con más de 3 días de antelación

### ✅Posibles soluciones

* Forzar esos datos y poner 3 días de antelación.

### ❌ Problema

Hay dos duplicados, que son los mismos que en el resto de json:



### ✅Posibles soluciones

* Borrar esas filas.
* Como son dos identificadores diferentes, podemos añadirles el **id\_visitante** detrás del nombre para que sean únicos - crearemos una columna aparte que concatena **t\_id + id\_visitante** como indicador unico de cada registro para preservar los duplicados

### ❌ Problema

* Hay 13 datos negativos en 'coste', que supone el 0.04% del df.



### ✅Posibles soluciones

* Borrar esas filas, ya que supone muy poco solo un 0,04% del df.

Exploracion individual William Ganem

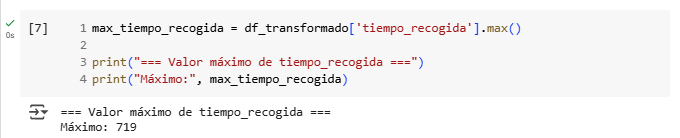
🕵🏻 Exploración de emocion.json

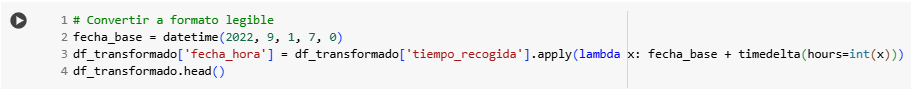
🕵🏻 Exploración de emocion.json

**tiempo\_recogida**: Representa el momento en que se toma cada fotografía, medido en horas desde las 07:00 del 1 de septiembre de 2022.

**Rango posible:**

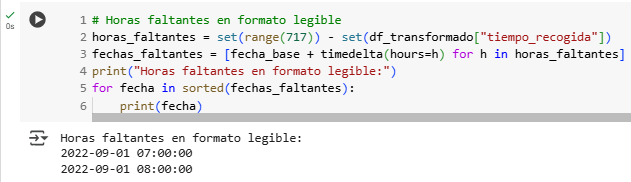
* El rango posible de horas que manejamos es de 0 a 719, ya que 719 es el valor máximo que encontramos después de hacer el conteo



* El valor 0 representa en forma de entero la fecha 01-09-2022 a las 7AM y el 719 la fecha 01-10-2022 a las 6AM

**Problema:**

* Dentro de la columna nos hacen falta los primeros 2 valores (0-1), por ende dentro de esa primera franja horaria no existen registros de fotos

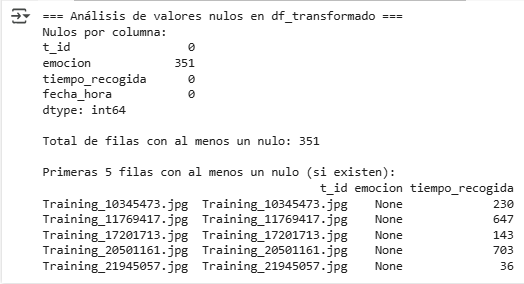
****

**Posibles causas:**

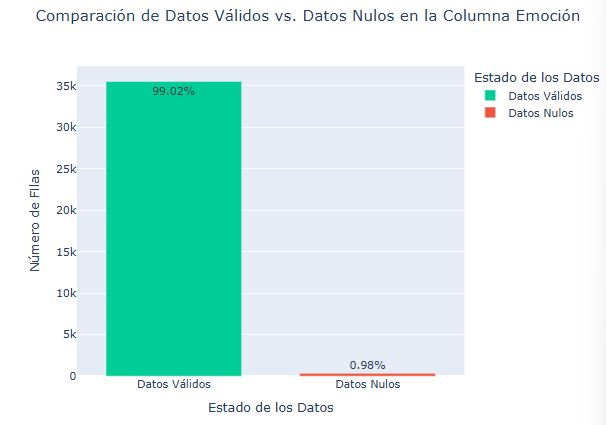
* Que la compañía no ha querido proporcionar esos datos, considerando que en otras fechas a la misma hora ya hay un flujo constante de visitas.
* Que durante esas 2 horas haya ocurrido un corte de electricidad
* Etc…

Consideramos que estas horas faltantes son irrelevantes al momento de realizar un análisis y en consecuencia para la toma de futuras decisiones ya que no aportarían ningún valor por lo corto que es el intervalo de tiempo.

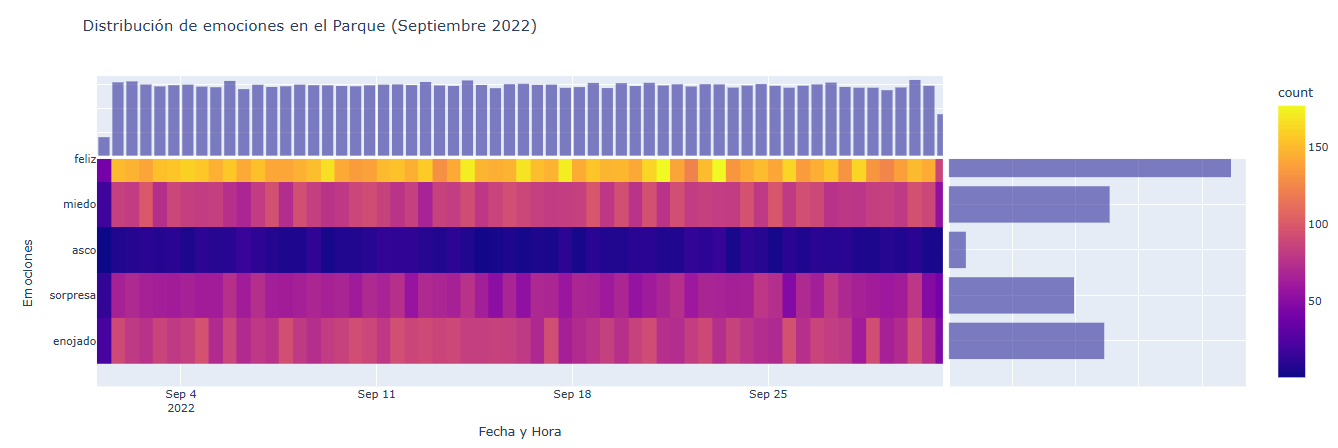
**Comparación de datos nulos y no nulos en C.Emoción:**



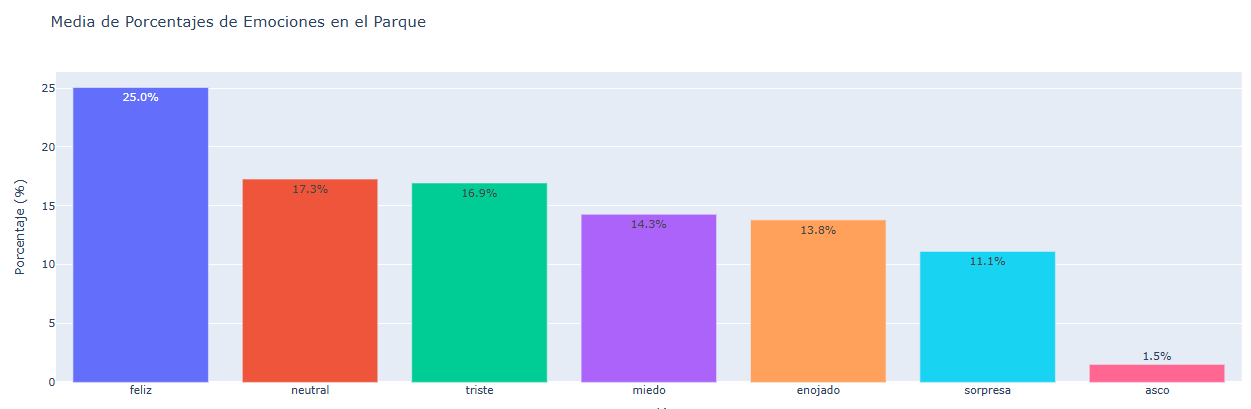
Encontramos 351 valores nulos dentro de un registro de más de 35mil filas por lo tanto no vemos necesario conservar o imputar esos datos nulos ya que representan menos del 1% de los datos.



Después de eliminar los nulos decidimos hacer un mapa de calor para apreciar la distribución de emociones por día



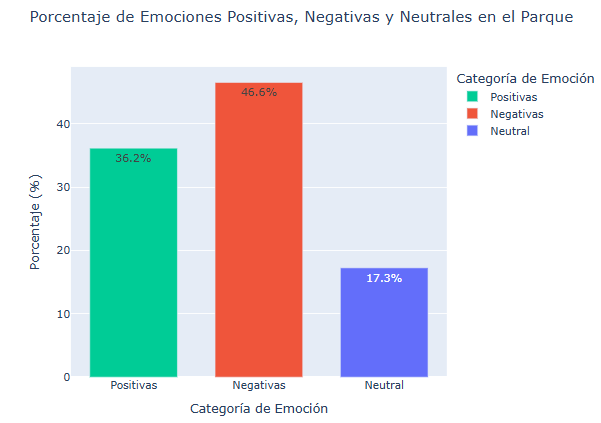
Básicamente a primeras la lectura nos indica que la emoción felicidad (“feliz”) es la que predomina y lo podemos apreciar de mejor manera con esta gráfica en donde distribuimos las emociones por porcentajes



Esto reitera que la emoción felicidad predomina sobre las otras.

Sin embargo al ver esta distribución decidimos realizar una agrupación de emociones positivas y emociones negativas, separando felicidad y sorpresa de tristeza, asco, miedo y enojo, dejando aparte los neutros y nos dimos cuenta que las emociones que predominan en el parque son las negativas.

Entendemos que la administración de Pontia World aceptará y valorará positivamente el planteamiento de métricas, por eso a pesar de no estar dentro de las preguntas de negocio, consideramos prudente mostrar este dato.

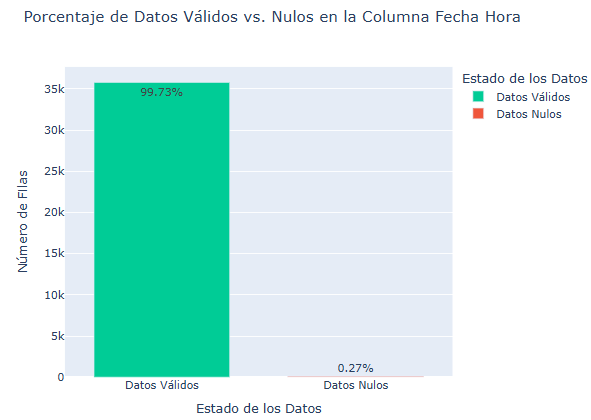
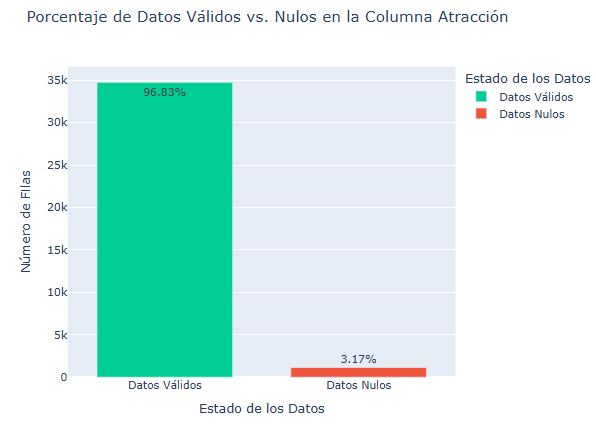
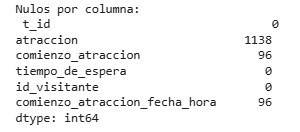


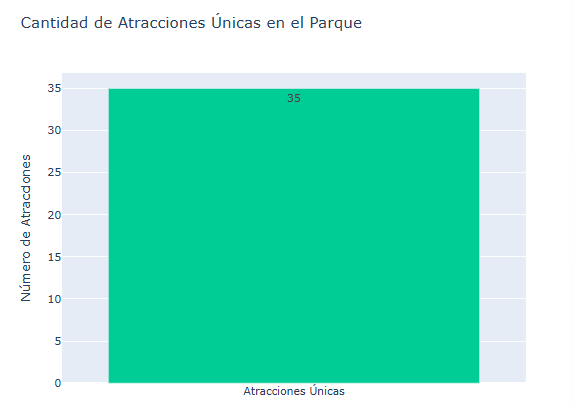
🎢 Exploración de id\_visitante\_atracciones.json

🎢Exploración de id\_visitante\_atracciones.json

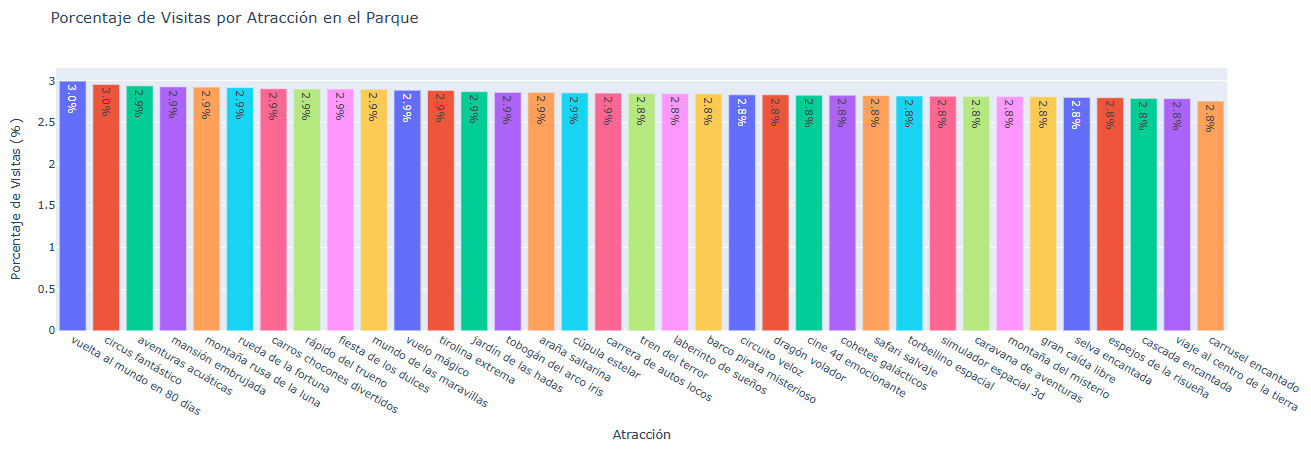
* **Datos Nulos:**

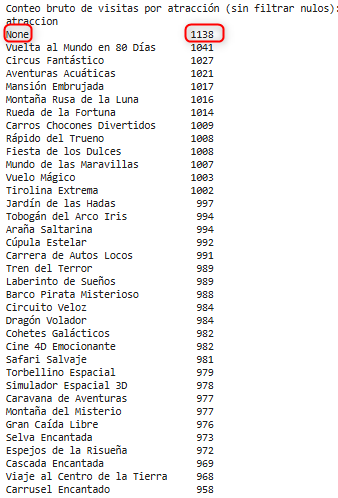
Al hacer un conteo de nulos dentro de este json a simple vista se podrían considerar irrelevantes, sin embargo



La distribución de ese 3% de atracciones en relación a la cantidad única de atracciones que existen el en parque que es de 35 atracciones si que tienen bastante peso 

Ya que las distribuciones de visitas de atracciones son sumamente parejas y ese porcentaje de nulos podría cambiar la percepción real de los datos que manejamos





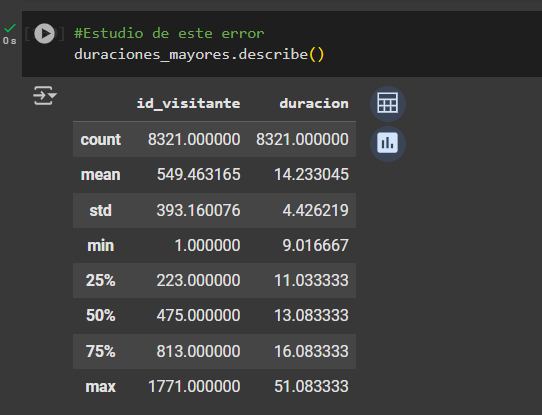
¿Qué podrían significar esos nulos?

1. Falta de registros de visita distribuidos en las distintas atracciones existentes
2. Una atracción que no está reflejada en los datos
3. Etc.

Nueva exploracion

Comprobando las reglas de negocio

#Podemos ver que 8321 instancias donde se supera esa duracion de 9 horas



Análisis del describe():

Duración

* **Media**: 14,23 horas (muy por encima del límite de 9 hr)
* **Mínima violación**: 9,02 h (muy justo)
* **25% de los casos** superan 11 horas
* **Máxima**: 51 horas en un día. Muy probablemente un error de datos

La media de duración es 14,23 horas, y la desviación estándar es 4,43 horas, lo cual es bastante alta

Un total de 1.527 visitantes únicos violaron la regla de las 9 horas por día al menos una vez. Lo que quiere decir que la regla de negocio NO SE CUMPLE la mayoria de las veces.

La regla de fast pass < 3 dias

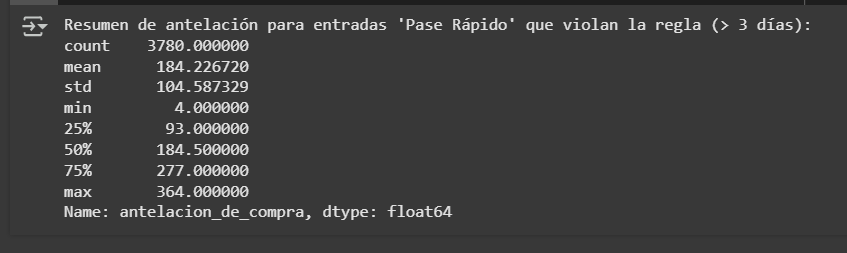
Resultado:

3.780 entradas violan la regla

Representan 10,53% del total

Y lo más preocupante: 63,15% de las "Pase Rápido" están mal vendidas

Haciendo un describe de estas violaciones de la regla se confirma una violación masiva y sistemática de la regla de negocio, que no sabemos porque pudo pasar



3.780 entradas violan la regla (más de 3 días de antelación).

Media ≈ 184 días: se están comprando con 6 meses de antelación en promedio.

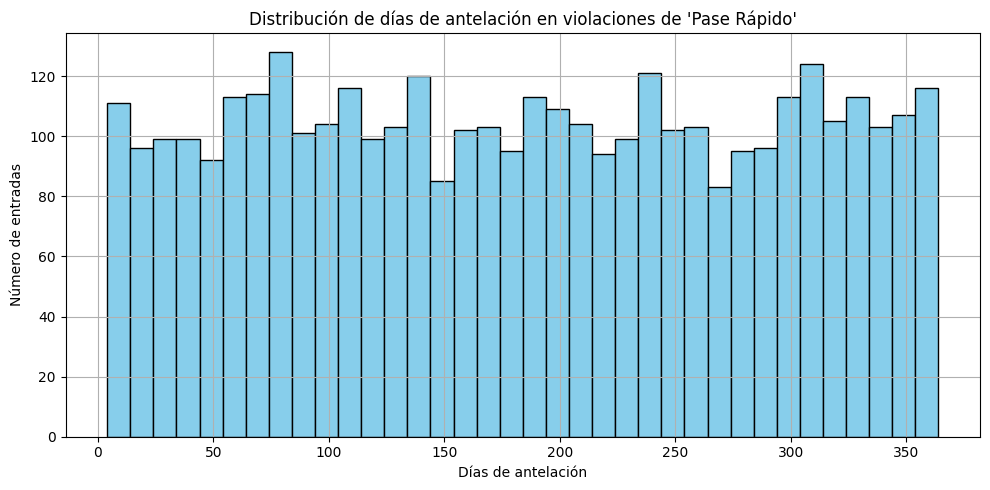
Mínimo = 4 días: justo por encima del umbral.

Máximo = 364 días: ¡un año entero!

Mediana = 184,5 días: **muy cercana a la media, lo que indica que la mayoría están en ese rango.**

Distribución muy consistente con ventas programadas a medio-largo plazo, probablemente automatizadas o permitidas erróneamente.

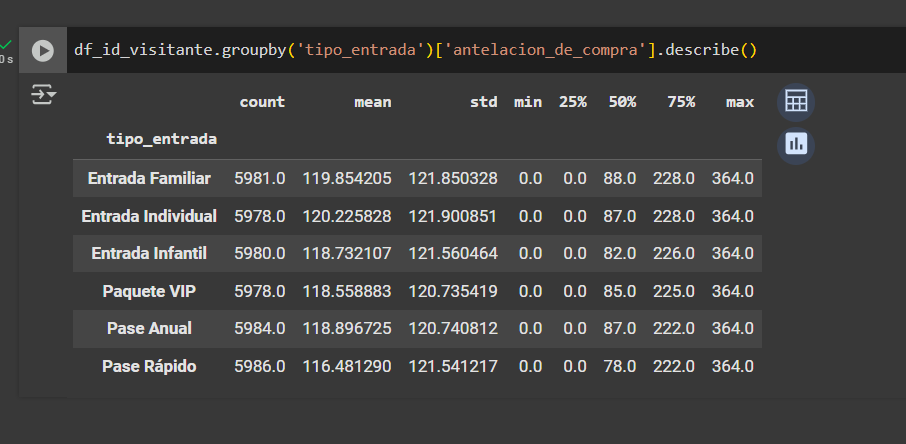
Viendo la distribucion de dias de estas entradas



El histograma muestra concentraciones muy marcadas cada cierto número de días (ej. en torno a 90, 180, 270 y 360 días). La regularidad sugiere que el sistema permitió comprar ciertas entradas "Pase Rápido" para fechas futuras muy lejanas en bloques fijos, como si se hubiera habilitado la venta para 3, 6, 9 o 12 meses adelante de golpe.Observamos que no hay nada cerca del umbral de 3 días A diferencia de lo esperado (una cola que comienza justo en 4, 5, 6 días...), no hay casi violaciones leves. La mayoría son compras mucho más anticipadas, lo cual apunta a un problema de lógica de negocio, no de comportamiento del usuario.

Conclusion: Podriamos forzar a 3 dias, pero la mayoria parecen estar muy por encima de ese rango, asi que no tiene sentido. Lo mas probable es que se haya asignado mal, asi que lo ideal seria renombrarlas con F.P. MAL VENDIDO, indicando a la empresa que algo ha fallado en el sistema para estas entradas, ha sido un error de carga o datos erroneous

Si analizamos que ocurre para los distintos tipos de entradas:



**Conclusiones importantes:**

Todas las entradas tienen antelación exagerada, no solo 'Pase Rápido'.La media está alrededor de 4 meses (120 días) para todas y no parecen seguir ninguna logica

La mediana es muy inferior a la media, lo que indica una distribución sesgada hacia valores muy altos. Muchas compras anticipadas a largo plazo. El valor máximo (364 días) está presente en todas. Esto indica que probablemente se habilitó un calendario completo (1 año) para todos los tipos de entrada, y no fue una excepción para 'Pase Rápido'.

El problema no es exclusivo de 'Pase Rápido'. Toda la lógica de antelación de compra parece estar mal configurada o demasiado laxa o la empresa no tiene realmente unas glas definidas para los distintos tipos de entradas. Si la política del parque define restricciones (como el límite de 3 días en 'Pase Rápido'), no se están aplicando en los datos correctamente ya que todas las entradas parece que se puedan comprar en taquilla (min=0) o de manera anticipada

Regla de no puede haber mas de 500 visitantes/ hora/atraccion

A partir del análisis que hemos realizado sobre la regla de aforo máximo de 500 visitantes por atracción y hora, estas son las principales conclusiones:

* No hay ninguna violación de la regla
* El análisis de df\_id\_visitante muestra que ninguna atracción supera los 500 visitantes en una hora, lo que indica que el sistema respeta adecuadamente esta limitación de aforo.
* La ocupación real es extremadamente baja. El promedio de visitantes por atracción y hora está entre 1,85 y 1,89 visitantes/hora, lo cual está muy por debajo del umbral crítico.

Por lo tnato

La distribución de visitantes en el tiempo está muy dispersa.

La demanda por atracción/hora es baja.

---

⚠️ 3. El umbral de 500 parece innecesariamente alto

Con menos de 2 visitantes por hora en promedio, el umbral de 500 no se acerca ni remotamente a ser un riesgo real.

Podrías replantear la regla para un control más útil, por ejemplo:

Activar alertas si una atracción supera los 100 visitantes/hora

O usar percentiles reales (por ejemplo, el 95%) para definir umbrales dinámicos

---

✅ 4. El sistema parece bien diseñado o los datos están filtrados

El hecho de que no haya ni un caso cercano al límite indica que:

O bien el sistema de reservas/datos está bien regulado,

O bien el conjunto de datos representa una muestra limpia o controlada.

---

¿Siguientes pasos recomendables?

Analizar la ocupación por día completo para validar si hay saturación diaria.

Ver si ciertas atracciones concentran más público o tienen más demanda.

Evaluar si el bajo volumen por hora es real o artefacto de cómo se genera el dato.

¿Quieres que haga el análisis por día para detectar saturaciones acumuladas?