Parcial Pandas 26.04.2021 102848

April 26, 2021

1 Parcial Pandas

(***) Una compañía de internet tiene una plataforma de sincronización de información que relaciona:

Producers: Identificados como event_producer_id, son sistemas que producen información a ser enviada a otras plataformas por intermedio de un consumer.

Consumers: Identificados como event_consumer_id son sistemas que consumen la información y la envían a las plataformas destino (identificada como event_consumer_target).

El archivo event_log.csv contiene todos los problemas que se detectaron, desde el inicio del funcionamiento del sistema: (event_id, event_type_id, event_status, country_id, event_producer_id, event_date)

Donde event_id es el identificador único del evento que sucedió, event_type_id indica el tipo de evento, event_status que indica el estado registrado de ese evento (puede ser ERROR, DELAYED, CAPTURED), country_id indicando el país de origen del evento y event_producer_id indicando cual ha sido el sistema "producer" que ha generado el evento.

Por otro lado contamos con event_types.csv con el siguiente formato: (event_type_id, event_type_name, event_consumer_id, event_consumer_target)

Donde event_type_id es el identificador único del tipo de evento, event_type_name el nombre del tipo de evento, event_consumer_id el identificador del sistema "consumer" que procesa la información y event_consumer_target indicando el nombre del la plataforma destino a la que envia la informacion el "consumer". Un tipo de evento solo tiene un target y consumer. Se desea:

a) Top 5 de Consumers que han tenido la mayor cantidad de eventos que resultaron en un event_status de ERROR.

b) De los eventos ocurridos para el country_id: BR indicar la cantidad de eventos totales por cada evento ocurridos por event_consumer_target.

Importo las librerias necesarias para resolver el parcial:

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Me genero un data frame para poder trabajar mas comodamente, en este caso ya que no tengo un archivo csv omito la linea read_csv(). En caso de tener un csv, para optimizar la carga del archivo utilizaria solo las columnas utiles e indicaria el tipo de dato de cada columna :

[48]:
```
event_log = pd.DataFrame({
    'event_id':[1,2,3,4,5,6,7],
    'event_type_id':[1, 2, 2, 1, 2, 3, 4],
    'event_status':['ERROR', 'DELAYED', 'ERROR', 'DELAYED', 'ERROR', 'CAPTURED',␣
 ↪'ERROR'],
    'country_id':['BR', 'BR', 'BR', 'BR', 'BR', 'BR', 'BR'],
    'event_producer_id':[1,6,3,4,5,12,8],
    'event_date':
 ↪['2021-03-16','2021-03-17','2021-03-18','2021-03-17','2021-03-17','2021-03-17','2021-03-16']

})
event_log
```

[48]:

	event_id	event_type_id	event_status	country_id	event_producer_id	event_date
0	1	1	ERROR	BR	1	2021-03-16
1	2	2	DELAYED	BR	6	2021-03-17
2	3	2	ERROR	BR	3	2021-03-18
3	4	1	DELAYED	BR	4	2021-03-17
4	5	2	ERROR	BR	5	2021-03-17
5	6	3	CAPTURED	BR	12	2021-03-17
6	7	4	ERROR	BR	8	2021-03-16

[49]:
```
event_types = pd.DataFrame({
    'event_type_id':[1, 2, 3, 4],
    'event_type_name':['Alta', 'Modificacion', 'Ingreso', 'Baja'],
    'event_consumer_id':['google', 'yt', 'google','yt'],
    'event_consumer_target':['salesforce', 'dynamodb', 'erp','salesforce']
})
event_types
```

[49]:

	event_type_id	event_type_name	event_consumer_id	event_consumer_target
0	1	Alta	google	salesforce
1	2	Modificacion	yt	dynamodb
2	3	Ingreso	google	erp

```
3              4            Baja                yt            salesforce
```

Primero filtro por los eventos que tuvieron error:

[50]:
```
errores = event_log.loc[event_log['event_status']== 'ERROR',['event_type_id']]
errores
```

[50]:
```
   event_type_id
0              1
2              2
4              2
6              4
```

Ahora unifico los dos df:

[51]:
```
consumidores_con_errores = errores.merge(event_types,how = 'inner')
consumidores_con_errores
```

[51]:
```
   event_type_id event_type_name event_consumer_id event_consumer_target
0              1            Alta            google            salesforce
1              2   Modificacion                yt              dynamodb
2              2   Modificacion                yt              dynamodb
3              4            Baja                yt            salesforce
```

Como mi dataframe es pequeño, en lugar de hacer el top5 de consumidores con mas errores hago el top 2:

[52]:
```
#Punto A
top_consumidores_con_errores=consumidores_con_errores.
 ↪groupby(['event_consumer_id']).count().reset_index()
top_consumidores_con_errores=top_consumidores_con_errores.
 ↪rename(columns={'event_type_id':'cantidad_errores'})
top_consumidores_con_errores.loc[:,['event_consumer_id','cantidad_errores']].
 ↪nlargest(2,'cantidad_errores')
```

[52]:
```
  event_consumer_id  cantidad_errores
1                yt                 3
0            google                 1
```

Para realizar el punto b, primero empiezo filtrando por los paises con codigo BR:

[53]:
```
eventos_brasil = event_log.loc[event_log['country_id']==␣
 ↪'BR',['event_type_id','event_status']]
eventos_brasil
```

[53]:
```
   event_type_id event_status
0              1        ERROR
1              2      DELAYED
2              2        ERROR
```

```
3               1      DELAYED
4               2        ERROR
5               3     CAPTURED
6               4        ERROR
```

[56]:
```
eventos = eventos_brasil.merge(event_types,how = 'inner')
eventos['cantidad_errores']=1
eventos
```

[56]:
```
   event_type_id event_status event_type_name event_consumer_id  \
0              1        ERROR            Alta            google
1              1      DELAYED            Alta            google
2              2      DELAYED   Modificacion                yt
3              2        ERROR   Modificacion                yt
4              2        ERROR   Modificacion                yt
5              3     CAPTURED         Ingreso            google
6              4        ERROR            Baja                yt

  event_consumer_target  cantidad_errores
0            salesforce                 1
1            salesforce                 1
2              dynamodb                 1
3              dynamodb                 1
4              dynamodb                 1
5                   erp                 1
6            salesforce                 1
```

No pude lograr eliminar los nombres event_consume_target y event_status, pero no deberian estar, intente con reset index pero no funciono :(

[80]:
```
#Punto B
pivot = eventos.pivot_table\
(index='event_status',\
 columns='event_consumer_target',\
 values='cantidad_errores',aggfunc='count')
pivot
```

[80]:
```
event_consumer_target  dynamodb  erp  salesforce
event_status
CAPTURED                    NaN  1.0         NaN
DELAYED                     1.0  NaN         1.0
ERROR                       2.0  NaN         2.0
```