preprocesamiento

April 16, 2018

1 Pre-procesamiento del set de datos

Para cada set de datos analizamos los tipos de datos, la presencia de valores nulos o no validos, los identificadores usados y la posibilidad de relacionar los datos entre distintos sets con estos identificadores. Se encontraron casos de fechas nulas y con formato no valido (personas menores de edad y mayores a 100 años) por ser una cantidad despreciable se descartaron los casos irregulares. Para el caso de las columnas donde existian muy pocos datos validos para usar, se opto por descartar la columna.

```
In [3]: # importacion general de librerias y de visualizacion (matplotlib y seaborn)
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns

        %matplotlib inline

        plt.style.use('default') # haciendo los graficos un poco mas bonitos en matplotlib
        #plt.rcParams['figure.figsize'] = (20, 10)

        sns.set(style="whitegrid") # seteando tipo de grid en seaborn
```

1.1 1. Educacion de postulantes

1.1.1 Inspeccion rapida: Forma y calidad

```
In [41]: # %timeit sirve para evaluar el tiempo de ejecucion
         df_temp = pd.read_csv('../csv/datos_navent_fiuba/fiuba_1_postulantes_educacion.csv')
         df_temp.head()
```

```
Out[41]:   idpostulante         nombre    estado
         0         NdJl       Posgrado  En Curso
         1         8BkL  Universitario  En Curso
         2         1d2B  Universitario  En Curso
         3         NPBx  Universitario  En Curso
         4         NPBx         Master  En Curso
```

```
In [12]: df_temp['nombre'].value_counts()
```

```
Out[12]: Secundario           110256
         Universitario        104295
         Terciario/Técnico     47733
         Otro                  24748
         Posgrado               7387
         Master                 3598
         Doctorado               214
         Name: nombre, dtype: int64

In [11]: df_temp['estado'].value_counts()

Out[11]: Graduado      194474
         En Curso       78531
         Abandonado     25226
         Name: estado, dtype: int64

In [42]: df_temp.isnull().any()

Out[42]: idpostulante    False
         nombre          False
         estado          False
         dtype: bool

In [45]: (df_temp['idpostulante'].value_counts() > 1).any()

Out[45]: True
```

1.2 2. Genero y edad de postulantes

1.2.1 Inspeccion rapida: Forma y calidad

```
In [48]: df_temp = pd.read_csv('../csv/datos_navent_fiuba/fiuba_2_postulantes_genero_y_edad.csv
         df_temp.head()

Out[48]:   idpostulante fechanacimiento  sexo
         0         NM5M      1970-12-03   FEM
         1         5awk      1962-12-04   FEM
         2         ZaO5      1978-08-10   FEM
         3         NdJl      1969-05-09  MASC
         4         eo2p      1981-02-16  MASC

In [33]: df_temp['sexo'].value_counts()

Out[33]: FEM           101981
         MASC           94339
         NO_DECLARA      4568
         Name: sexo, dtype: int64

In [34]: df_temp.isnull().any()
```

```
Out[34]: idpostulante       False
         fechanacimiento     True
         sexo               False
         dtype: bool

In [165]: df_temp.isnull().sum()

Out[165]: idpostulante          0
          fechanacimiento    4750
          sexo                  0
          dtype: int64

In [39]: # ok, miro cuales son las fechas malas no nulas

In [35]: df_temp[
             df_temp['fechanacimiento'].notnull()][
             (pd.to_datetime(df_temp['fechanacimiento'].dropna(), errors='coerce').isnull())]

Out[35]:        idpostulante fechanacimiento        sexo
         56206       xkPwXwY      0031-12-11         FEM
         71458       LN85Y3b      0029-05-11        MASC
         130846      8M2R6pz      0024-02-09         FEM
         141832      A36Npjj      0033-09-14         FEM
         145683      dYjVOrb      0012-11-04  NO_DECLARA
         148638      GNZOvAv      0004-07-19        MASC
         149653      1QPQ8QL      0011-03-08        MASC

In [36]: # Las fechas malas pueden descartarse
         df_temp['fechanacimiento'] = pd.to_datetime(df_temp['fechanacimiento'], errors='coerc

In [37]: # Considero fechas anteriores al siglo XX como invalidas
         df_temp.loc[df_temp['fechanacimiento'] < '1900-01-01', 'fechanacimiento'] = pd.NaT

In [55]: # Considero fechas que implican edades menores a 15 anios como invalidas
         df_temp.loc[df_temp['fechanacimiento'] > '2003-01-01', 'fechanacimiento'] = pd.NaT

In [47]: (df_temp['idpostulante'].value_counts() > 1).any()

Out[47]: False

In [38]: df_temp.to_csv('../csv/datos_navent_fiuba/fiuba_2_postulantes_genero_y_edad_fix.csv')
```

1.3 3. Vista de avisos online y offline

1.3.1 Inspeccion rapida: Forma y calidad

```
In [201]: df_temp = pd.read_csv('../csv/datos_navent_fiuba/fiuba_3_vistas.csv')
          df_temp.head()
```

```
Out[201]:       idAviso                    timestamp idpostulante
          0  1111780242  2018-02-23T13:38:13.187-0500      YjVJQ6Z
          1  1112263876  2018-02-23T13:38:14.296-0500      BmVpYoR
          2  1112327963  2018-02-23T13:38:14.329-0500      wVkBzZd
          3  1112318643  2018-02-23T13:38:17.921-0500      OqmP9pv
          4  1111903673  2018-02-23T13:38:18.973-0500      DrpbXDP
```

```
In [172]: df_temp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 961897 entries, 0 to 961896
Data columns (total 3 columns):
idAviso         961897 non-null int64
timestamp       961897 non-null object
idpostulante    961897 non-null object
dtypes: int64(1), object(2)
memory usage: 22.0+ MB
```

```
In [206]: pd.to_datetime(df_temp['timestamp']).sort_values().head(10)
```

```
Out[206]: 2373   2018-02-23 18:38:10.808
          1041   2018-02-23 18:38:12.173
          1352   2018-02-23 18:38:12.581
          1691   2018-02-23 18:38:12.790
          1692   2018-02-23 18:38:12.945
          0      2018-02-23 18:38:13.187
          2029   2018-02-23 18:38:13.269
          2030   2018-02-23 18:38:13.343
          351    2018-02-23 18:38:13.849
          1      2018-02-23 18:38:14.296
          Name: timestamp, dtype: datetime64[ns]
```

1.4 4. Postulaciones hasta 1 de marzo

1.4.1 Inspeccion rapida: Forma y calidad

```
In [208]: df_temp = pd.read_csv('../csv/datos_navent_fiuba/fiuba_4_postulaciones.csv')
          df_temp.head()
```

```
Out[208]:       idaviso idpostulante     fechapostulacion
          0  1112257047         NM5M  2018-01-15 16:22:34
          1  1111920714         NM5M  2018-02-06 09:04:50
          2  1112346945         NM5M  2018-02-22 09:04:47
          3  1112345547         NM5M  2018-02-22 09:04:59
          4  1112237522         5awk  2018-01-25 18:55:03
```

```
In [209]: df_temp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3401623 entries, 0 to 3401622
Data columns (total 3 columns):
idaviso             int64
idpostulante        object
fechapostulacion    object
dtypes: int64(1), object(2)
memory usage: 77.9+ MB
```

```
In [212]: pd.to_datetime(df_temp['fechapostulacion']).sort_values().head(10)
```

```
Out[212]: 1525012   2018-01-15 00:00:01
          1269880   2018-01-15 00:00:02
          1842775   2018-01-15 00:00:09
          1525013   2018-01-15 00:00:10
          3348905   2018-01-15 00:00:11
          222799    2018-01-15 00:00:16
          1812230   2018-01-15 00:00:16
          1558135   2018-01-15 00:00:16
          2435961   2018-01-15 00:00:16
          3159078   2018-01-15 00:00:18
          Name: fechapostulacion, dtype: datetime64[ns]
```

```
In [199]: df_temp['idpostulante'].apply(len).value_counts()
```

```
Out[199]: 7    2763243
          6     632698
          5       5278
          4        404
          Name: idpostulante, dtype: int64
```

1.5 5. Avisos online al 8 de marzo

1.5.1 Inspeccion rapida: Forma y calidad

```
In [4]: df_temp = pd.read_csv('../csv/datos_navent_fiuba/fiuba_5_avisos_online.csv')
        df_temp.head()
```

```
Out[4]:       idaviso
        0  1112355872
        1  1112335374
        2  1112374842
        3  1111984070
        4  1111822480
```

```
In [5]: df_temp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5028 entries, 0 to 5027
```

```
Data columns (total 1 columns):
idaviso    5028 non-null int64
dtypes: int64(1)
memory usage: 39.4 KB
```

```
In [6]: df_temp['idaviso'].isnull().any()
```

```
Out[6]: False
```

1.6 6. Detalle de avisos online y offline

1.6.1 Inspeccion rapida: Forma y calidad

```
In [8]: df_temp = pd.read_csv('../csv/datos_navent_fiuba/fiuba_6_avisos_detalle.csv')
        df_temp.head()
```

```
Out[8]:       idaviso  idpais                                          titulo  \
        0     8725750       1               VENDEDOR/A PROVINCIA DE SANTA FE
        1    17903700       1                                     Enfermeras
        2  1000150677       1                                  Chofer de taxi
        3  1000610287       1  CHOFER DE CAMIONETA BAHIA BLANCA - PUNTA ALTA
        4  1000872556       1  Operarios de Planta - Rubro Electrodomésticos

                                              descripcion        nombre_zona  \
        0  <p><strong><strong>Empresa:</strong></strong> ...  Gran Buenos Aires
        1  <p>Solicitamos para importante cadena de farma...  Gran Buenos Aires
        2  <p>TE GUSTA MANEJAR? QUERES GANAR PLATA HACIEN...    Capital Federal
        3  <p><strong>Somos una empresa multinacional que...  Gran Buenos Aires
        4  <p><strong>OPERARIOS DE PLANTA</strong></p><p>...  Gran Buenos Aires

          ciudad       mapacalle tipo_de_trabajo         nivel_laboral nombre_area  \
        0    NaN             NaN       Full-time  Senior / Semi-Senior   Comercial
        1    NaN             NaN       Full-time  Senior / Semi-Senior       Salud
        2    NaN  Empedrado 2336       Full-time  Senior / Semi-Senior  Transporte
        3    NaN             NaN       Full-time  Senior / Semi-Senior  Transporte
        4    NaN             NaN       Full-time  Senior / Semi-Senior  Producción

             denominacion_empresa
        0                  VENTOR
        1  Farmacias Central Oeste
        2             FAMITAX SRL
        3     Wurth Argentina S.A
        4      ELECTRO OUTLET SRL
```

```
In [9]: df_temp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13534 entries, 0 to 13533
```

```
Data columns (total 11 columns):
idaviso                 13534 non-null int64
idpais                  13534 non-null int64
titulo                  13534 non-null object
descripcion             13534 non-null object
nombre_zona             13534 non-null object
ciudad                  47 non-null object
mapacalle               872 non-null object
tipo_de_trabajo         13534 non-null object
nivel_laboral           13534 non-null object
nombre_area             13534 non-null object
denominacion_empresa    13529 non-null object
dtypes: int64(2), object(9)
memory usage: 1.1+ MB
```

```
In [10]: df_temp.isnull().sum()

Out[10]: idaviso                    0
         idpais                     0
         titulo                     0
         descripcion                0
         nombre_zona                0
         ciudad                 13487
         mapacalle              12662
         tipo_de_trabajo            0
         nivel_laboral              0
         nombre_area                0
         denominacion_empresa       5
         dtype: int64

In [13]: df_temp['idpais'].value_counts()

Out[13]: 1    13534
         Name: idpais, dtype: int64

In [15]: df_temp['nombre_zona'].value_counts()

Out[15]: Gran Buenos Aires              12654
         Capital Federal                  876
         Buenos Aires (fuera de GBA)        2
         GBA Oeste                          2
         Name: nombre_zona, dtype: int64

In [16]: df_temp['ciudad'].value_counts()

Out[16]: Buenos Aires               14
         Argentina                  13
         CABA                        3
```

```
San Isidro                  2
Capital Federal             2
paternal                    1
Santa Rosa                  1
Microcentro                 1
República Argentina         1
Tortuguitas                 1
Buenos Aires Province       1
Parque Patricios            1
La Plata                    1
Barracas                    1
Mendoza                     1
caba                        1
Vicente Lopez               1
Zárate, Campana, Escobar    1
Name: ciudad, dtype: int64
```

```
In [18]: df_temp['tipo_de_trabajo'].value_counts()
```

```
Out[18]: Full-time          12339
         Part-time            863
         Teletrabajo          110
         Pasantia              63
         Por Horas             63
         Temporario            42
         Por Contrato          37
         Fines de Semana       14
         Primer empleo          3
         Name: tipo_de_trabajo, dtype: int64
```

```
In [23]: df_temp['nivel_laboral'].value_counts()
```

```
Out[23]: Senior / Semi-Senior                   9407
         Junior                                 2216
         Otro                                    921
         Jefe / Supervisor / Responsable         809
         Gerencia / Alta Gerencia / Dirección    181
         Name: nivel_laboral, dtype: int64
```

```
In [24]: df_temp['nombre_area'].value_counts()
```

```
Out[24]: Ventas                                  1659
         Comercial                                983
         Administración                           901
         Producción                               821
         Programación                             576
         Contabilidad                             416
         Tecnologia / Sistemas                    388
         Atención al Cliente                      347
```

Mantenimiento	324
Recursos Humanos	235
Gastronomia	234
Oficios y Profesiones	209
Soporte Técnico	203
Logística	200
Call Center	191
Almacén / Depósito / Expedición	184
Compras	170
Marketing	153
Otros	153
Administración de Personal	152
Recepcionista	151
Transporte	148
Mantenimiento y Limpieza	141
Telemarketing	138
Finanzas	138
Tesorería	137
Créditos y Cobranzas	132
Salud	127
Desarrollo de Negocios	126
Medicina	119
	...
Auditoría Médica	3
Instrumentación	2
Topografía	2
Data Warehousing	2
Educación especial	2
Trabajo Social	2
Trabajo social	2
Diseño Multimedia	2
Mercadotecnia Internacional	2
Otras áreas técnicas en salud	2
Ingeniería Geológica	2
Diseño 3D	2
Medicina Laboral	2
Dirección	2
Responsabilidad Social	2
Farmacia comercial	2
Bienestar Estudiantil	1
Urbanismo	1
Comunicaciones Externas	1
Farmacia hospitalaria	1
Traduccion	1
Idiomas	1
Exploración Minera y Petroquimica	1
Otras Especialidades médicas	1
Emergentología	1

```
Arte y Cultura                              1
Telefonista                                 1
Instrumentación quirúrgica                  1
Quimica                                     1
Ingeniería en Petróleo y Petroquímica       1
Name: nombre_area, Length: 173, dtype: int64
```

```
In [26]: df_temp['nombre_area'].str.upper().value_counts().count()
```

```
Out[26]: 172
```

```
In [29]: df_temp[df_temp['denominacion_empresa'].isnull()]
```

```
Out[29]:          idaviso  idpais                                titulo  \
267    1111960305       1        VENDEDORA- PORTSAID- ZONA NORTE
268    1111960330       1   VENDEDORA- PORTSAID- CAPITAL FEDERAL
3850   1112289439       1                 Muestrista  Desiderata
6262   1112243714       1  Analista Comercial Senior - PORTSAID-
11096  1111946024       1          VENDEDOR DE SALON - PORTSAID-

                                                descripcion       nombre_zona  \
267    <p align="center"><strong>EXPERIENCIA PORTSAID...  Gran Buenos Aires
268    <p align="center"><strong>EXPERIENCIA PORTSAID...  Gran Buenos Aires
3850   <p>En Mazalosa s.a. nos encontramos en la búsq...  Gran Buenos Aires
6262   <p>En Mazalosa nos encontramos en la búsqueda ...  Gran Buenos Aires
11096  <p>En Portsaid estamos buscando al mejor vende...  Gran Buenos Aires

      ciudad mapacalle tipo_de_trabajo         nivel_laboral nombre_area  \
267      NaN       NaN       Full-time                 Otro      Ventas
268      NaN       NaN       Full-time                 Otro      Ventas
3850     NaN       NaN       Part-time  Senior / Semi-Senior  Producción
6262     NaN       NaN       Full-time  Senior / Semi-Senior   Comercial
11096    NaN       NaN       Full-time  Senior / Semi-Senior      Ventas

      denominacion_empresa
267                    NaN
268                    NaN
3850                   NaN
6262                   NaN
11096                  NaN
```