FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1

[75.06 / 95.58] Organización de Datos
Primer Cuatrimestre de 2018

Alumno	Padrón	E-mail
Soro, Lucas Gustavo	95665	lugusor@gmail.com
Hazan, Pablo Nehuén	96522	phazan.fiuba@gmail.com
Núñez Leyes, Javier Damián	94455	javier.nunezleyes@gmail.com
Montes, Marcelo	81397	mdmontes@gmail.com

Índice

1_ Introducción

El informe expone los resultados del análisis exploratorio de los datos del registro histórico de avisos de busquedas laborales en https://www.zonajobs.com.ar

Link al set de datos:
https://drive.google.com/file/d/1K4uRag5nmGtfuvzyJV9RL_73lzsh_iTO/view?usp=sharing

Link al repositorio con los notebook en GitHub:
https://github.com/javierleyes/OrganizacionDeDatos

2_ 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.

1. Educacion de postulantes

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
```

Bitacora: Todos los datos sanos y bien categorizados. Algunos postulantes tienen varios niveles de educacion.
