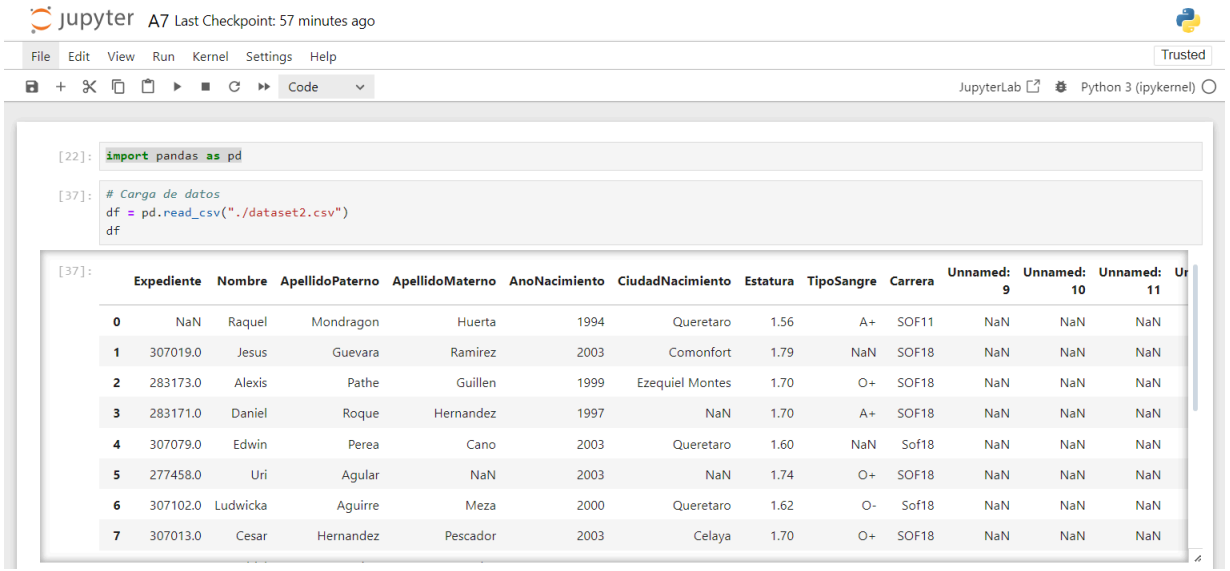
Actividad #7

Carga, limpieza y transformación

APLICACIONES DE BIG DATA

PATRYCK YAEL POUMIAN CAMACHO 307036

PATRYCK YAEL POUMIAN CAMACHO 307036



[37]:

	Expediente	Nombre	ApellidoPaterno	ApellidoMaterno	AnoNacimiento	CiudadNacimiento	Estatura	TipoSangre	Carrera	Unnamed: 9	Unnamed: 10	Unnamed: 11	Ur
0	NaN	Raquel	Mondragon	Huerta	1994	Queretaro	1.56	A+	SOF11	NaN	NaN	NaN	
1	307019.0	Jesus	Guevara	Ramirez	2003	Comonfort	1.79	NaN	SOF18	NaN	NaN	NaN	
2	283173.0	Alexis	Pathe	Guillen	1999	Ezequiel Montes	1.70	O+	SOF18	NaN	NaN	NaN	
3	283171.0	Daniel	Roque	Hernandez	1997	NaN	1.70	A+	SOF18	NaN	NaN	NaN	
4	307079.0	Edwin	Perea	Cano	2003	Queretaro	1.60	NaN	Sof18	NaN	NaN	NaN	
5	277458.0	Uri	Agular	NaN	2003	NaN	1.74	O+	SOF18	NaN	NaN	NaN	
6	307102.0	Ludwicka	Aguirre	Meza	2000	Queretaro	1.62	O-	Sof18	NaN	NaN	NaN	
7	307013.0	Cesar	Hernandez	Pescador	2003	Celaya	1.70	O+	SOF18	NaN	NaN	NaN	

```
# Limpieza de datos
# a
df = df.fillna({"ApellidoPaterno": "No especificado","ApellidoMaterno":"No especificado", "TipoSangre":"No especificado","Carrera":"No especificado"})
df[["Expediente","Nombre","ApellidoPaterno","ApellidoMaterno","AnoNacimiento","CiudadNacimiento","Estatura","TipoSangre","Carrera"]]

# b
df = df.fillna({"Edad": 0,"Estatura":0})

# c
df = df.loc[:, ~df.columns.str.startswith('Unnamed')]
df = df.dropna()
df
```

	Expediente	Nombre	ApellidoPaterno	ApellidoMaterno	AnoNacimiento	CiudadNacimiento	Estatura	TipoSangre	Carrera
1	307019.0	Jesus	Guevara	Ramirez	2003	Comonfort	1.79	No especificado	SOF18
2	283173.0	Alexis	Pathe	Guillen	1999	Ezequiel Montes	1.70	O+	SOF18
4	307079.0	Edwin	Perea	Cano	2003	Queretaro	1.60	No especificado	Sof18
6	307102.0	Ludwicka	Aguirre	Meza	2000	Queretaro	1.62	O-	Sof18
7	307013.0	Cesar	Hernandez	Pescador	2003	Celaya	1.70	O+	SOF18
8	275890.0	Gabiel	Feregrino	Hernandez	2003	Queretaro	2.10	O-	SOF18
9	276777.0	Andrei	Garcia	Bautista	2003	DF	1.72	O+	SOF18
10	307092.0	Pablo	Camorlinga	Vazquez	2003	San jose Iturbide	1.70	O-	SOF18
11	307110.0	Einar	Rodriguez	Valle	2001	DF	0.00	O-	SOF18
13	307051.0	Diego	Pescador	No especificado	2003	Salamanca	1.78	O+	SOF18
15	307042.0	Samuel	Serrato	Loyola	2003	Queretaro	1.70	B+	SOF18

```python
# 3
def calcularEdad(AnoNacimiento):
    edad = 2024 - AnoNacimiento
    return edad

df["Edad"] = df["AnoNacimiento"].apply(calcularEdad)
df[["Expediente","Nombre","ApellidoPaterno","ApellidoMaterno","AnoNacimiento","CiudadNacimiento","Estatura","TipoSangre","Carrera","Edad"]]

# 4
df[["Expediente","Nombre","ApellidoPaterno","CiudadNacimiento","Edad"]]
```

	Expediente	Nombre	ApellidoPaterno	CiudadNacimiento	Edad
1	307019.0	Jesus	Guevara	Comonfort	21
2	283173.0	Alexis	Pathe	Ezequiel Montes	25
4	307079.0	Edwin	Perea	Queretaro	21
6	307102.0	Ludwicka	Aguirre	Queretaro	24
7	307013.0	Cesar	Hernandez	Celaya	21
8	275890.0	Gabiel	Feregrino	Queretaro	21
9	276777.0	Andrei	Garcia	DF	21
10	307092.0	Pablo	Camorlinga	San jose Iturbide	21
11	307110.0	Einar	Rodriguez	DF	23
13	307051.0	Diego	Pescador	Salamanca	21
15	307042.0	Samuel	Serrato	Queretaro	21

```python
# 5
df_filtrado = df[df['CiudadNacimiento'] != 'Queretaro']

df_filtrado.to_csv('ResultadoNoQueretano.csv')
df_filtrado
```

	Expediente	Nombre	ApellidoPaterno	ApellidoMaterno	AnoNacimiento	CiudadNacimiento	Estatura	TipoSangre	Carrera	Edad
1	307019.0	Jesus	Guevara	Ramirez	2003	Comonfort	1.79	No especificado	SOF18	21
2	283173.0	Alexis	Pathe	Guillen	1999	Ezequiel Montes	1.70	O+	SOF18	25
7	307013.0	Cesar	Hernandez	Pescador	2003	Celaya	1.70	O+	SOF18	21
9	276777.0	Andrei	Garcia	Bautista	2003	DF	1.72	O+	SOF18	21
10	307092.0	Pablo	Camorlinga	Vazquez	2003	San jose Iturbide	1.70	O-	SOF18	21
11	307110.0	Einar	Rodriguez	Valle	2001	DF	0.00	O-	SOF18	23
13	307051.0	Diego	Pescador	No especificado	2003	Salamanca	1.78	O+	SOF18	21