In [ ]:
```
import pandas as pd
import numpy as np
```

In [ ]:
```
# Leer el archivo nombres.csv y convertirlo en un df
df_base = pd.read_csv('nombres.csv')
# Ver los primeros registros del df
df_base.head()
```

Out[ ]:

	state	sex	year	name	quantity
0	MO	F	1910	Mary	611
1	MO	F	1910	Helen	313
2	MO	F	1910	Dorothy	270
3	MO	F	1910	Mildred	267
4	MO	F	1910	Ruth	237

In [ ]:
```
# Mostrar información sobre el df creado
df_base.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5933561 entries, 0 to 5933560
Data columns (total 5 columns):
 #   Column    Dtype 
---  ------    ----- 
 0   state     object
 1   sex       object
 2   year      int64 
 3   name      object
 4   quantity  int64 
dtypes: int64(2), object(3)
memory usage: 226.3+ MB
```

In [ ]:
```
#Filtrar todas los hombres con nombre Mary
#Crear un nuevo dataframe
df_mary = df_base[(df_base["sex"] == "M") & (df_base["name"] == "Mary")]
df_mary
```

Out[ ]:

	state	sex	year	name	quantity
76848	MO	M	1913	Mary	6
78179	MO	M	1916	Mary	5
79914	MO	M	1920	Mary	9
80430	MO	M	1921	Mary	7
81976	MO	M	1924	Mary	5
...	...	...	...	...	...
5815488	CT	M	1955	Mary	5
5816465	CT	M	1958	Mary	5
5816806	CT	M	1959	Mary	5
5817816	CT	M	1962	Mary	6
5899574	KS	M	1936	Mary	5

1054 rows × 5 columns

In [ ]:
```
#Ya no necesitamos el sexo y el nombre, hacemos una copia
nuevo_df = df_mary[["state", "quantity"]]
nuevo_df.head()
```

Out[ ]:

	state	quantity
76848	MO	6
78179	MO	5
79914	MO	9
80430	MO	7
81976	MO	5

In [ ]:
```
#Obtener la suma de la columna quantity
nuevo_df[["quantity"]].sum()
```

Out[ ]:
```
quantity    9638
dtype: int64
```

In [ ]:
```
estado_df = nuevo_df.groupby("state").agg("sum")
estado_df
```

Out[ ]:

state	quantity
AL	519
AR	159
AZ	6
CA	540
CO	18
CT	26
DC	51
FL	127
GA	482
IA	35
IL	570
IN	213
KS	5
KY	343
LA	237
MA	220
MD	92
MI	293
MN	59
MO	236
MS	392
NC	576
NE	10
NJ	137
NM	69
NY	785
OH	543
OK	92
PA	668
SC	361
TN	316
TX	969
UT	5
VA	263
WA	10

state	quantity
WI	92
WV	119

In [ ]:
```
# Reto:
# Obtener la cantidad promedio de Mary por estado usando estados_df
promedio_df = nuevo_df.groupby("state").agg("mean")
promedio_df
```

Out[ ]:

state	quantity
AL	10.380000
AR	6.625000
AZ	6.000000
CA	8.852459
CO	6.000000
CT	5.200000
DC	12.750000
FL	6.350000
GA	10.041667
IA	5.833333
IL	10.363636
IN	6.454545
KS	5.000000
KY	9.026316
LA	6.771429
MA	6.875000
MD	6.133333
MI	7.918919
MN	5.900000
MO	7.375000
MS	9.333333
NC	12.521739
NE	5.000000
NJ	6.227273
NM	6.272727
NY	11.544118
OH	8.901639
OK	6.133333
PA	11.133333
SC	9.025000
TN	9.028571
TX	13.842857
UT	5.000000
VA	7.514286
WA	5.000000

state	quantity
WI	6.133333
WV	6.263158

In [ ]:
```
# Ordenar el set de datos por la mayor cantidad
estado_df.sort_values('quantity', ascending=False)
```

Out[ ]:

state	quantity
TX	969
NY	785
PA	668
NC	576
IL	570
OH	543
CA	540
AL	519
GA	482
MS	392
SC	361
KY	343
TN	316
MI	293
VA	263
LA	237
MO	236
MA	220
IN	213
AR	159
NJ	137
FL	127
WV	119
MD	92
OK	92
WI	92
NM	69
MN	59
DC	51
IA	35
CT	26
CO	18
NE	10
WA	10
AZ	6

state	quantity
KS	5
UT	5

In [ ]:
```
# Reto
# Obtener los 10 registros con la mayor cantidad de Mary
estado_df.sort_values('quantity', ascending=False).head(10)
```

Out[ ]:

state	quantity
TX	969
NY	785
PA	668
NC	576
IL	570
OH	543
CA	540
AL	519
GA	482
MS	392

In [ ]:
```
# Para convertir el índice en una columna del df, se puede generar uno nuevo aleato
# Nota al uso de inplace=True para modificar el df
estado_df.reset_index(inplace=True)
estado_df
```

Out[ ]:

	state	quantity
0	AL	519
1	AR	159
2	AZ	6
3	CA	540
4	CO	18
5	CT	26
6	DC	51
7	FL	127
8	GA	482
9	IA	35
10	IL	570
11	IN	213
12	KS	5
13	KY	343
14	LA	237
15	MA	220
16	MD	92
17	MI	293
18	MN	59
19	MO	236
20	MS	392
21	NC	576
22	NE	10
23	NJ	137
24	NM	69
25	NY	785
26	OH	543
27	OK	92
28	PA	668
29	SC	361
30	TN	316
31	TX	969
32	UT	5
33	VA	263
34	WA	10
35	WI	92

	state	quantity
36	WV	119

In [ ]:
```
estado_df.to_excel('estados_mary.xlsx', index=None)
```