

Tarea1

December 10, 2021

1 Tarea 1

Autor: Harvey Rodriguez Gil

```
[ ]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

1.0.1 Descripción:

El archivo “Baseball Salaries 2011.xlsx” contiene datos sobre los salarios de los beisbolistas de las ligas mayores en USA desde 2002 a 2011.

```
[ ]: df = pd.read_excel("Baseball Salaries 2011.xlsx")
df.head(10)
```

```
[ ]:
```

	Player	Team	Position	Experience	Salary
0	A.J. Burnett	New York Yankees	Pitcher	2	16500000
1	A.J. Ellis	Los Angeles Dodgers	Catcher	1	421000
2	A.J. Pierzynski	Chicago White Sox	Catcher	6	2000000
3	Aaron Cook	Colorado Rockies	Pitcher	4	9875000
4	Aaron Crow	Kansas City Royals	Pitcher	14	1400000
5	Aaron Harang	San Diego Padres	Pitcher	4	3500000
6	Aaron Heilman	Arizona Diamondbacks	Pitcher	14	2000000
7	Aaron Hill	Toronto Blue Jays	Second Baseman	7	5000000
8	Aaron Laffey	Seattle Mariners	Pitcher	2	431600
9	Aaron Miles	Los Angeles Dodgers	Second Baseman	9	500000

Q: ¿Cuáles variables de las presentadas para cada jugador son cuantitativas y cuáles categóricas?

A: Al revisar los datos encontramos que tenemos 5 columnas: **Player**, **Team**, **Position**, **Experience** y **Salary**. Por lo tanto podemos indicar que tenemos dos variables categóricas (**Team** y **Position**) y dos variables cuantitativas (**Experience** y **Salary**). El detalle en la siguiente celda.

```
[ ]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 843 entries, 0 to 842
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype
#   ...          ...
```

```

---  -----  -----  ---
0   Player      843 non-null  object
1   Team        843 non-null  object
2   Position    843 non-null  object
3   Experience   843 non-null  int64
4   Salary      843 non-null  int64
dtypes: int64(2), object(3)
memory usage: 33.1+ KB

```

Q: Haga un histograma del salario de todos los jugadores, y caracterice visualmente su forma (¿sesgada a la derecha o a la izquierda?, ¿achatada?, ¿Grandes colas? Si es sesgada, ¿qué relación puede tener esto respecto a la mediana y la media de la serie de datos?).

A: Siguiendo celda

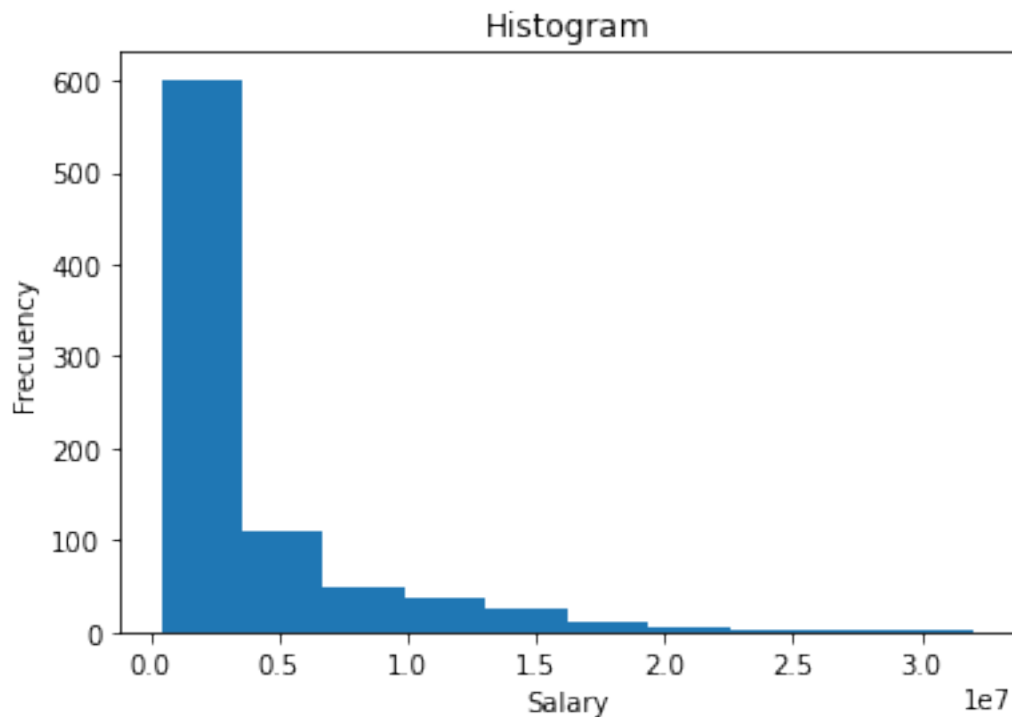
```

[ ]: n, bins, patches = plt.hist(df['Salary'])

plt.xlabel('Salary')
plt.ylabel('Frequency')
plt.title('Histogram')

plt.show();

```



[]: