KIBERNUM
IT Academy

Módulo 3 – Análisis Exploratorio y Programación Estadística

Análisis Univariado

Ciencia de Datos

Objetivos

- Explicar en qué consiste el análisis de datos univariado.
- Distinguir medidas de tendencia central, dispersión y posición.
- Reconocer diagramas para análisis univariado.
- Implementar análisis univariado en Python.

Análisis Univariado

Es la forma más simple de análisis, en donde se estudia cada variable de forma aislada. El propósito principal de un análisis univariado es describir los datos, valiéndose de la estadística descriptiva, para encontrar patrones y develar fenómenos subyacentes difíciles de encontrar solamente observando los datos de forma aislada.

Hagamos un análisis del dataset de Sueldos de San Francisco, en particular, de la columna sueldo base (**BasePay**) para ver cómo se comporta.

```
1 import pandas as pd
```

```
1 df = pd.read_csv('Salaries.csv')
```

```
1 df.head(2)
```

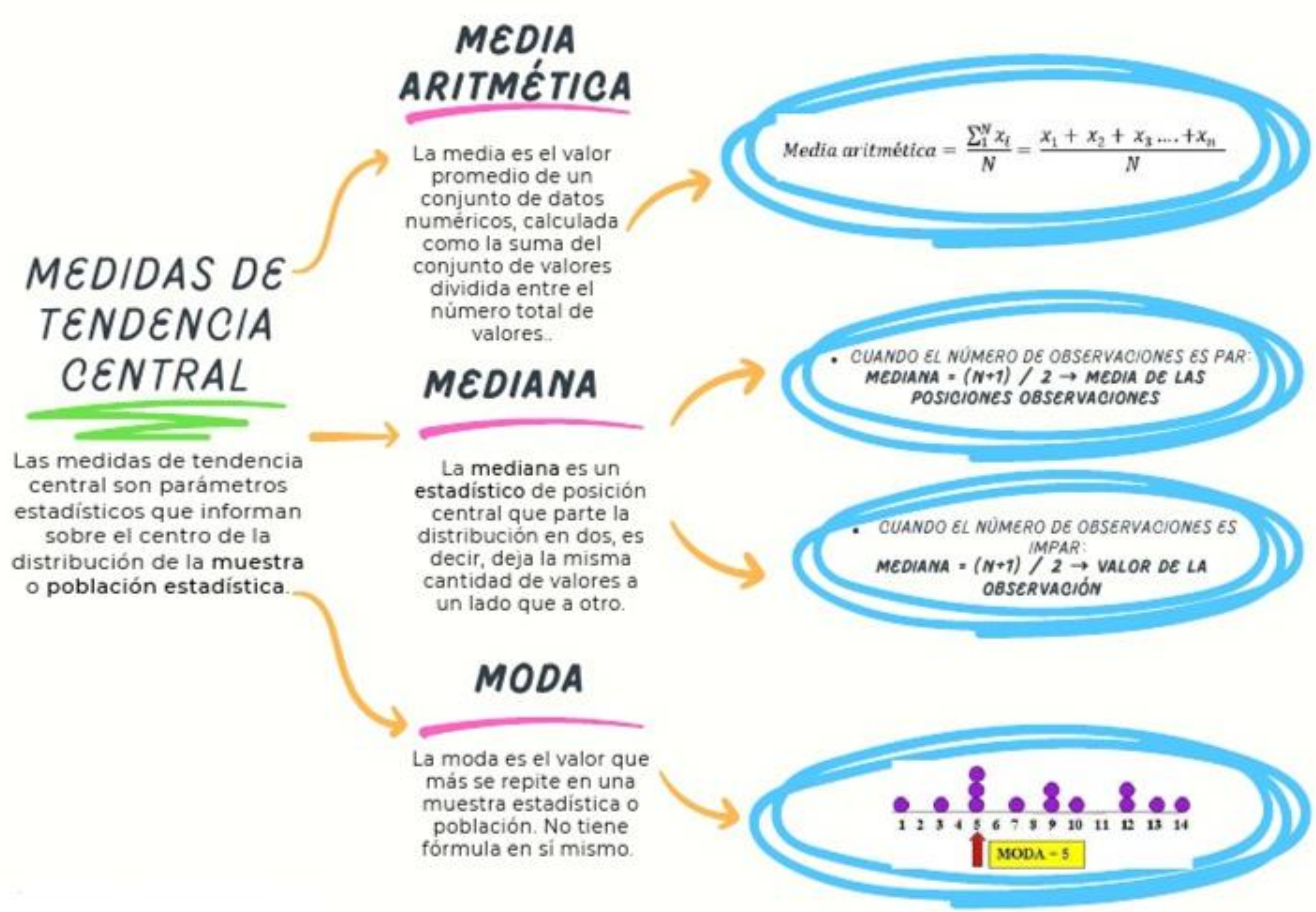
	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	567595.43	567595.43	2011	NaN	San Francisco	NaN
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	538909.28	2011	NaN	San Francisco	NaN

Medidas de Tendencia Central

Son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores, representando el centro en torno al cual se sitúan los datos.

- **Media:** corresponde al promedio aritmético, es decir, la suma de los valores dividido por la cantidad de valores.
- **Mediana:** corresponde al valor de la variable que ocupa la posición central en el conjunto de valores, es decir, el 50% de las observaciones tiene un valor igual o inferior a la mediana y el otro 50% un valor igual o superior a la mediana
- **Moda:** corresponde al valor que más se repite en un conjunto de datos.

Medidas de Tendencia Central



KIBERNUM

Medidas de Tendencia Central

Ahora veamos qué pasa con la variable sueldo base.

```
# media
df['BasePay'].mean()
```

```
66325.44884050643
```

El promedio de sueldo es de aproximadamente U$ 66.325

```
# mediana
df['BasePay'].median()
```

```
65007.45
```

La mitad de las personas tiene un sueldo igual o inferior a U$ 65.007

```
# moda
df['BasePay'].mode()
```

```
0    0.0
dtype: float64
```

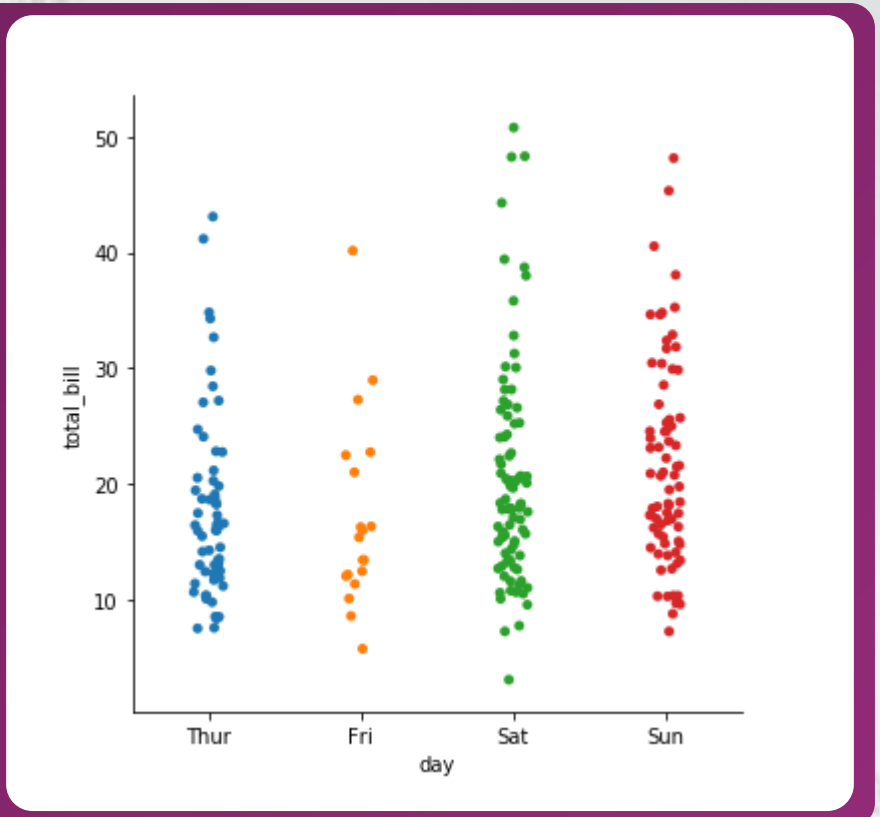
El valor que más se repite es U$0.0
Hemos descubierto un insight!!!

Medidas de Dispersión

Muchas veces vamos a preguntarnos si las mediciones tienes valores muy cercanos entre sí o muy dispersos. Para eso necesitamos **medidas de dispersión**.

Se calculan para describir la dispersión de los valores de una muestra en torno a un parámetro de ubicación. En pocas palabras, los parámetros de dispersión **son una medida de cuánto fluctúa una muestra en torno a un valor medio**.

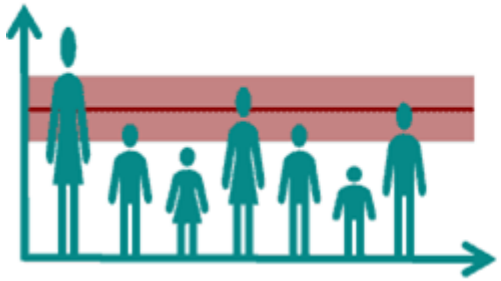
Nótese en este ejemplo que la variable cuantitativa Total de la Cuenta (total_bill) tiene una dispersión de valores levemente mayor el sábado.



Medidas de Dispersión

Existen varias formas de medir la dispersión de una variable. A continuación, algunos ejemplos:

Desviación estándar



Distancia media de todos los valores medios al valor medio.

Rango

Distancia entre el valor más bajo y el más alto de una distribución.

Rango intercuartílico

Espectro en el que se sitúa el 50% medio de los valores. Diferencia entre el primer y el tercer cuartil.

Medidas de Dispersión

A continuación, se describirán las siguientes medidas más utilizadas para caracterizar la dispersión de una variable.

- **Rango de variación**: es la diferencia entre el mayor valor de la variable y el menor valor.
- **Varianza**: es la suma de las diferencias entre el valor y el promedio, al cuadrado, dividido por la cantidad de valores.
- **Desviación Estándar**: corresponde a la raíz cuadrada de la varianza, para que la medida de dispersión quede en la misma unidad que los valores de la variable.

Rango de Variación

El **rango de variación**, también conocido simplemente como "rango", es una medida estadística que indica la diferencia entre el valor máximo y el valor mínimo en un conjunto de datos.

En otras palabras, es la amplitud o extensión total de los valores presentes en el conjunto de datos.

$$R = \text{Valor}_{\text{Max}} - \text{Valor}_{\text{Min}}$$

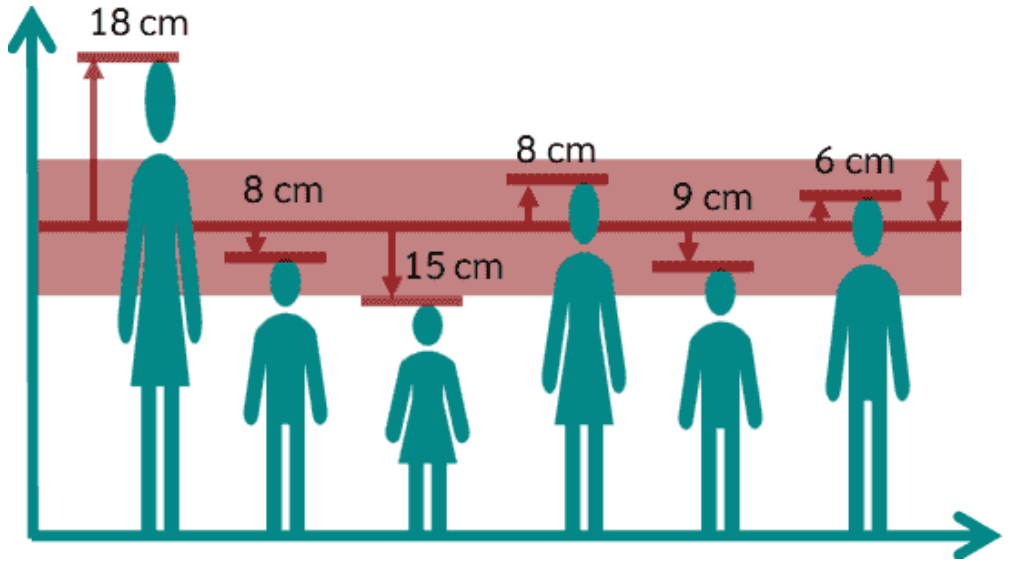
$$R = x_{max} - x_{min}$$

Varianza

La varianza es una medida importante en estadística porque proporciona información sobre la dispersión de los datos.

Valores de varianza más altos indican una mayor dispersión de los datos alrededor de la media, mientras que valores de varianza más bajos indican una menor dispersión. Nótese que la varianza se incrementa de forma cuadrática en la medida que los puntos más se alejan de su media.

$$\mathrm{Var}(x) = \frac{\sum_1^n (x_i - \bar{X})^2}{n}$$

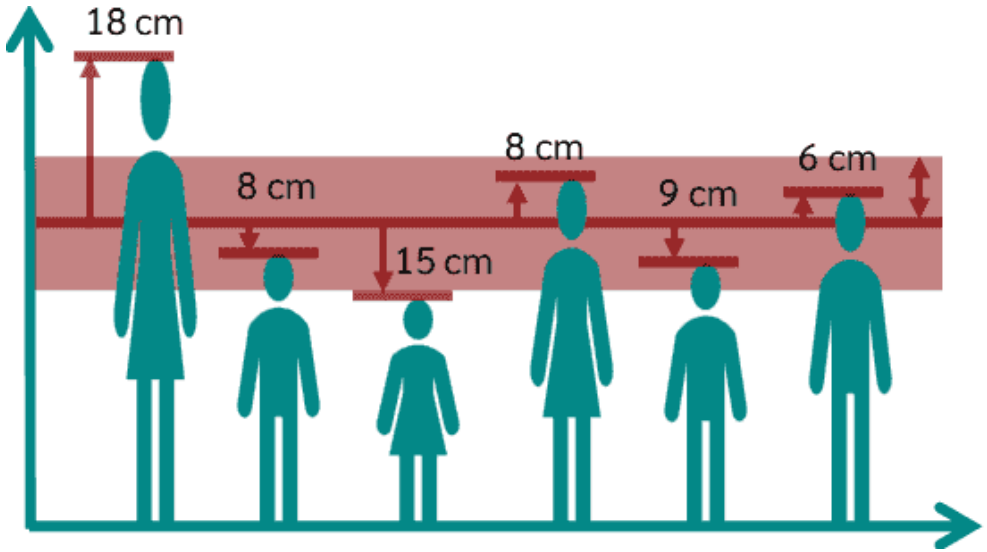


Desviación Estándar

La desviación estándar se calcula como la raíz cuadrada de la varianza. La varianza, como mencioné anteriormente, es la media de las diferencias al cuadrado entre cada valor individual y la media del conjunto de datos.

Al tomar la raíz cuadrada de la varianza, obtenemos una medida de dispersión en las mismas unidades que los datos originales.

$$\sigma^2 = Var(x)$$



Corrección de Bessel

Cuando se quiere estimar la desviación estándar como indicador estadístico de una población a partir de una muestra, es importante considerar que, en general, las muestras tienden a subestimar la variabilidad de la población. Sin embargo, esto se puede mejorar con la corrección de Bessel:

- Si estamos estimando la desviación estándar (o la varianza) de la población a partir de una muestra, debemos dividir por n -1
- Si estamos midiendo la desviación estándar (o la varianza) de la población, debemos dividir por n.

*Corrección de Bessel
https://es.wikipedia.org/wiki/Correcci%C3%B3n_de_Bessel

Corrección de Bessel

Desviación estándar de la población

$$\sigma = \sqrt{\frac{\sum_1^n (x_i - \bar{X})^2}{n}}$$

Desviación estándar de una muestra

$$s = \sqrt{\frac{\sum_1^n (x_i - \bar{X})^2}{n - 1}}$$

Medidas de Dispersión

A continuación, algunos ejemplos de cómo realizar cálculos de **medidas de dispersión** en Python con la librería Pandas.

```python
# rango de variación
rv = df['BasePay'].max() - df['BasePay'].min()
print(rv)
```

```
319441.02
```

```python
# varianza, por defecto utiliza n - 1 (estimacion de la población a partir de una muestra)
df['BasePay'].var()
```

```
1828814049.0424156
```

```python
# varianza utilizando n (medición de la desviación de la población)
df['BasePay'].var(ddof=0)
```

```
1828801695.9467416
```

```python
# desviación estándar, por defecto utiliza n - 1
df['BasePay'].std()
```

```
42764.63549525958
```

```python
# desviación estándar, utilizando n
df['BasePay'].std(ddof=0)
```

```
42764.49106381066
```

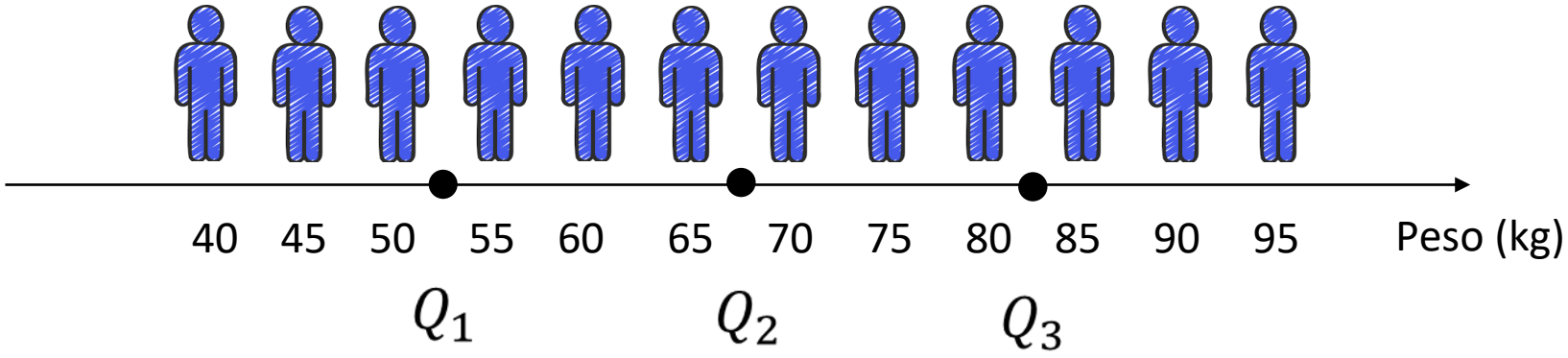
Medidas de Posición

Las medidas de posición, también conocidas como **cuantiles**, son estadísticas que dividen un conjunto de datos ordenados en partes iguales o proporcionales. Estas medidas son útiles para comprender la distribución de los datos y para identificar valores atípicos. Las medidas más habituales son las siguientes:

- **El cuartil:** divide la distribución en 4 partes iguales, por lo tanto hay 3 cuartiles.
- **El quintil:** divide la distribución en 5 partes, por lo tanto hay 4 quintiles.
- **El decil:** divide la distribución en 10 partes iguales.
- **El percentil:** divide la distribución en 100 partes iguales.

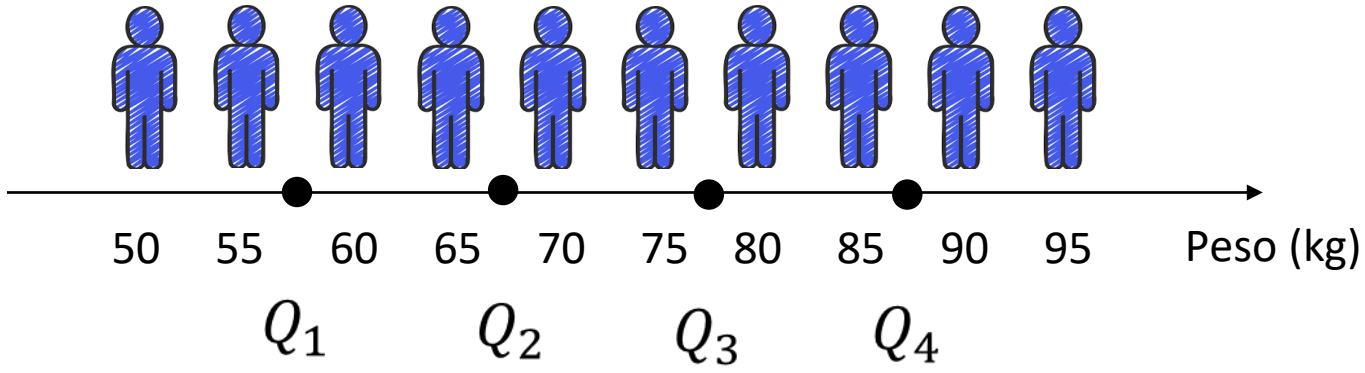
Cuartiles

Los Cuartiles dividen el conjunto de datos en cuatro partes iguales. El primer cuartil (Q1) representa el 25% de los datos más bajos, el segundo cuartil (Q2) es la mediana y representa el 50% de los datos, y el tercer cuartil (Q3) representa el 75% de los datos más bajos.



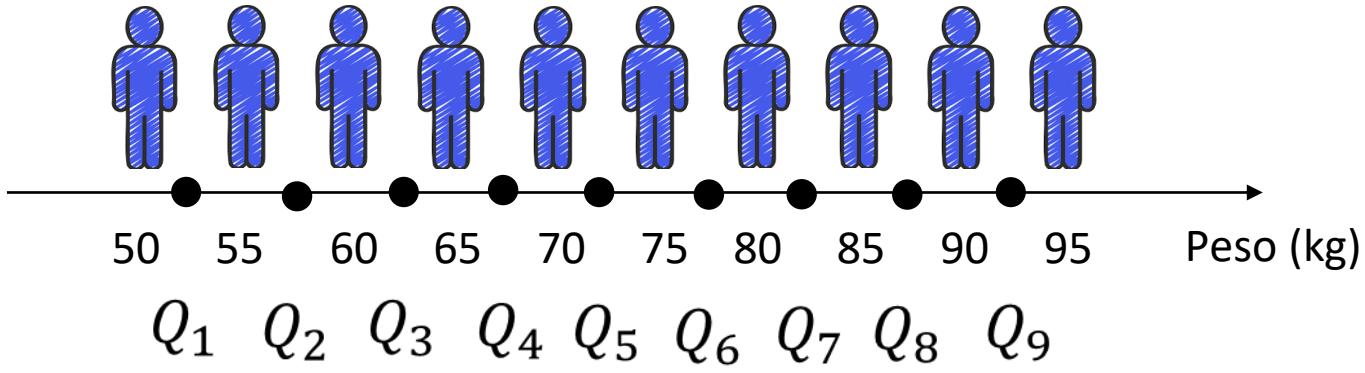
Quintiles

Los Quintiles dividen el conjunto de datos en cinco partes iguales, lo que significa que hay cuatro puntos de división. Los quintiles se usan menos comúnmente que los cuartiles, pero son útiles cuando se necesita una mayor granularidad en la división de los datos.



Deciles

Los Deciles dividen un conjunto de datos ordenados en diez partes iguales, de manera que cada parte representa el 10% del total de los datos. Esto significa que hay nueve puntos de corte que dividen los datos en diez grupos, cada uno con aproximadamente el 10% de los datos.



Medidas de Posición en Python

En este ejemplo utilizamos el método quantile para calcular los cuartiles de un conjunto de datos con la librería Pandas.

```
1  # Cálculo de los cuartiles
2  q1 = df['BasePay'].quantile(q=.25)
3  q2 = df['BasePay'].quantile(q=.5)
4  q3 = df['BasePay'].quantile(q=.75)
5
6  print('Q1:',q1)
7  print('Q2:',q2)
8  print('Q3:',q3)
```

```
Q1: 33588.2
Q2: 65007.45
Q3: 94691.05
```

Sumario de estadísticas

Con el método describe() obtenemos un sumario de estadísticas de la variable, lo cual es un excelente punto de partida para obtener insights.

```
1  # sumario de estadísticas
2  df['BasePay'].describe()
```

```
count    148045.000000
mean      66325.448841
std       42764.635495
min        -166.010000
25%       33588.200000
50%       65007.450000
75%       94691.050000
max      319275.010000
Name: BasePay, dtype: float64
```

¿¿Sueldo base negativo??
Insight!!!!

KIBERNUM

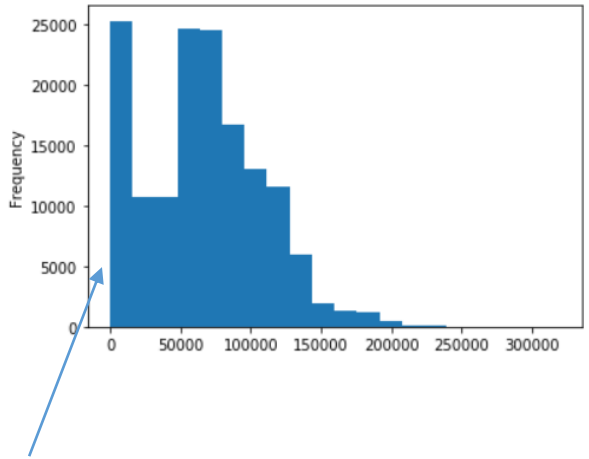
Histograma

Un histograma permite visualizar la frecuencia de ocurrencia de los distintos valores del conjunto de datos. Esta división se hacer en intervalos regulares entre el valor mínimo y máximo del conjunto.

Con este parámetro indicamos la cantidad de intervalos de acumulación

```
1 df['BasePay'].plot(kind='hist', bins=20)
```

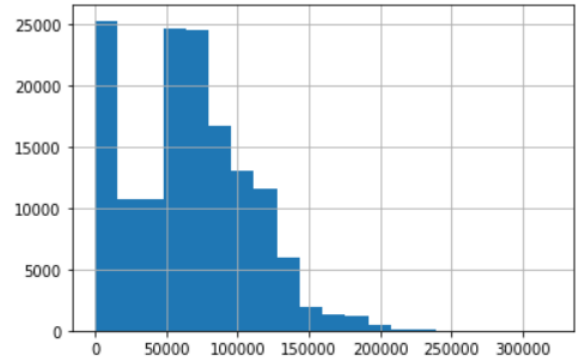
```
<matplotlib.axes._subplots.AxesSubplot at 0x19d2579e148>
```



Otra forma más resumida de generar un histograma

```
1 df['BasePay'].hist(bins=20)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x19d2542f908>
```

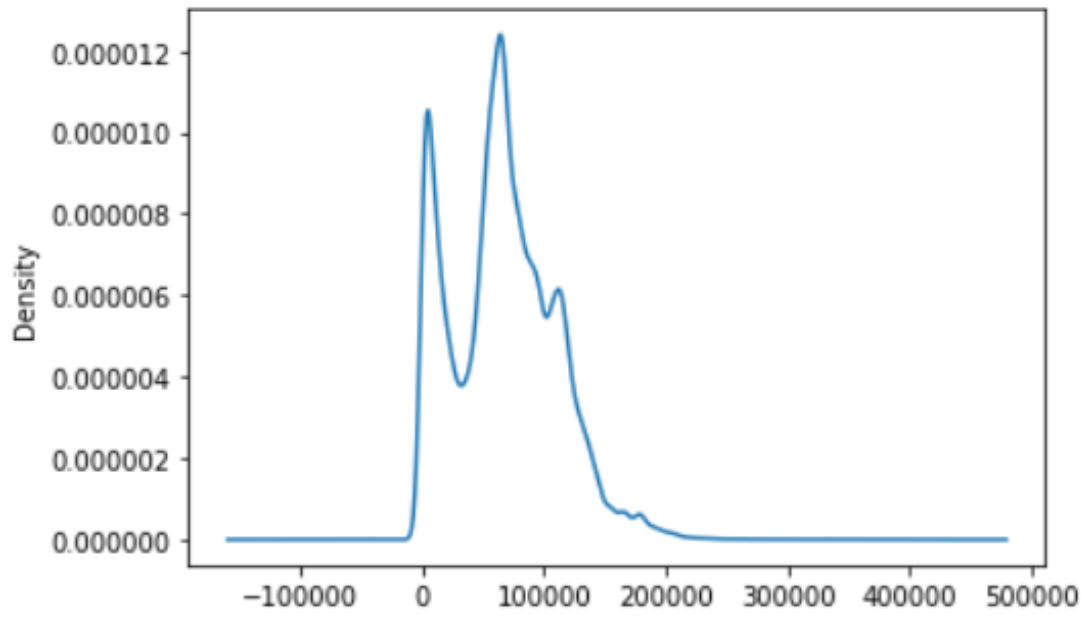


Nótese que los intervalos de mayor frecuencia se encuentran en torno a los 60 mil dólares, y que también hay un intervalo de valores próximos a cero que tiene una frecuencia alta. **Este puede ser otro insigth!!!**

Diagrama de Distribución

```
1 df['BasePay'].plot(kind='kde')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x19d25a18248>
```



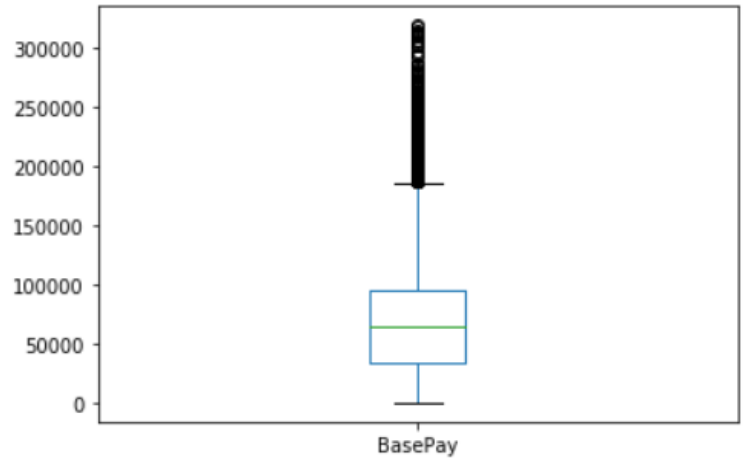
Un diagrama de distribución estima la función de densidad de probabilidad a partir de un conjunto de datos. (KDE: kernel density estimation).

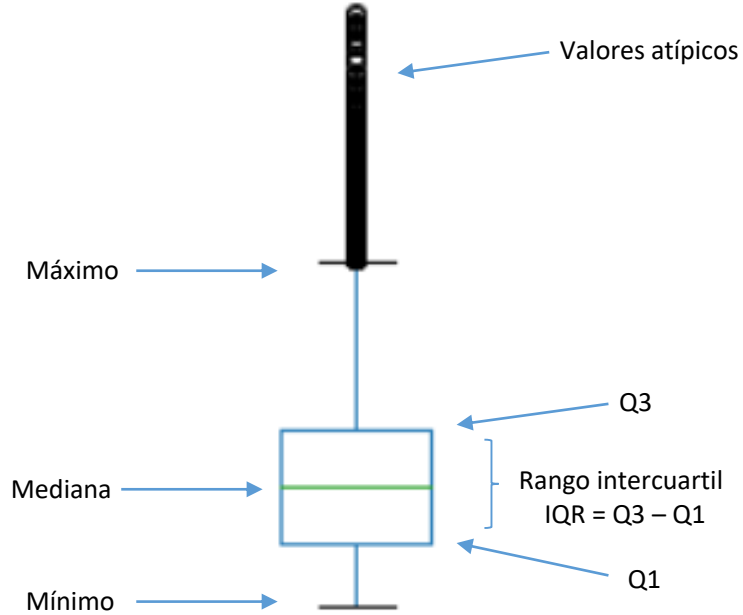
Medidas de Posición

El boxplot, o diagrama de caja y bigote, entrega información de cómo se distribuyen los valores en un conjunto de datos, identificando también los cuartiles y puntos atípicos.

```
1 df['BasePay'].plot(kind='box')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x19d25963c88>
```





Valores Atípicos

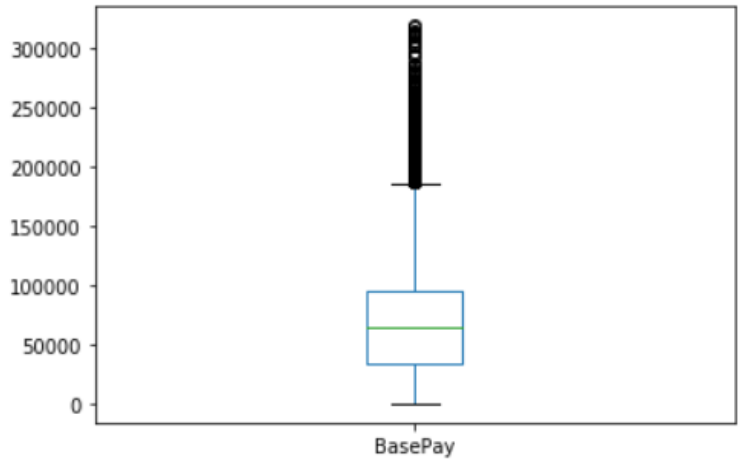
Un valor atípico en un set de datos corresponde a una observación que es numéricamente distante del resto de los datos, extremadamente grande o extremadamente pequeña. Un valor atípico puede generar un efecto desproporcionado en los resultados estadísticos, como la media, lo cual puede conducir a interpretaciones engañosas.

Un valor atípico, llamado a veces anomalía, podría deberse a un fenómeno único (por ejemplo, una lista de clientes de un banco con una persona de 124 años) o bien podría ser producido por un error (por ejemplo, tipearon 124 en vez de 24 años al momento de crear al cliente).

Sea cual sea el dato, es conveniente identificar la presencia de valores atípicos o anomalías en los datos.

```
1 df['BasePay'].plot(kind='box')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x19d25963c88>
```



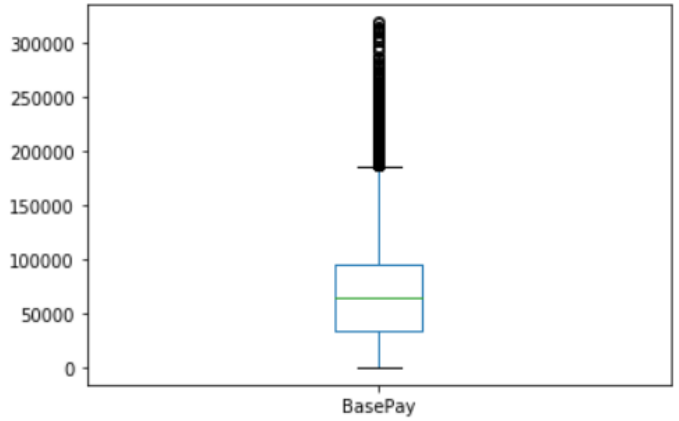
KIBERNUM

Sumario de estadísticas

En este caso, los sueldos más altos son distinguidos como valores atípicos, lo cual tiene sentido puesto que hay pocos sueldos muy altos.

```
1 df['BasePay'].plot(kind='box')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x19d25963c88>
```



Cálculo de los límites

LSUP = Q3 + 1.5*IQR

LINF = Q1 - 1.5*IQR

Dudas y consultas

KIBERNUM

Fin de la Presentación

KIBERNUM