KIBERNUM
IT Academy

Módulo 3
Clase 3

Análisis Univariado

Análisis Exploratorio de Datos

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.

```
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

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

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

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

Entregan información sobre la variación de la variable, en donde se busca resumir en un sólo valor la dispersión que tiene un conjunto de datos.

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 (*)

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

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.

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

Medidas de Dispersión

Cuando se quiere estimar la desviación estándar como indicador estadístico de una población a partir de una muestra, es importante distinguir los conceptos de población y muestra.

En general, las muestras subestiman la variabilidad de la población. 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

Medidas de Dispersión

A continuación, algunos ejemplos:

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

```
319441.02
```

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

```
1828814049.0424156
```

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

```
1828801695.9467416
```

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

```
42764.63549525958
```

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

```
42764.49106381066
```

Medidas de Posición

Las medidas de posición son indicadores estadísticos que permiten resumir los datos en uno sólo, o dividir su distribución en intervalos del mismo tamaño. Las medidas más habituales son las siguientes:

- **El cuartil:** divide la distribución en 4 partes iguales. De esta forma, existen tres cuartiles, Q1, Q2 y Q3.
- **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.

Medidas de Posición

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

```
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!!!!

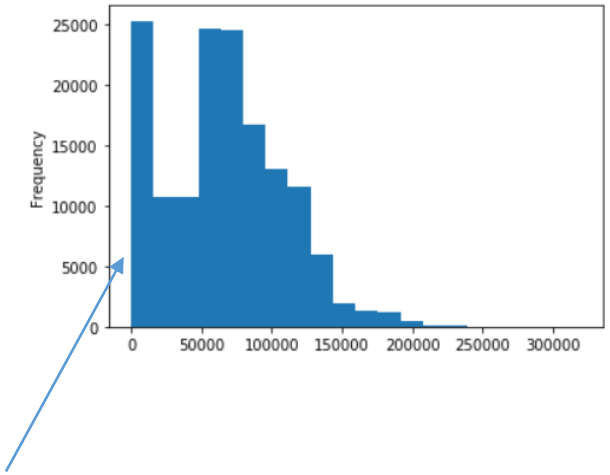
Histograma

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)
```

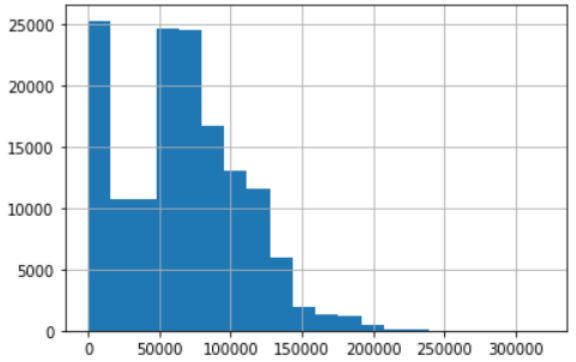
```
<matplotlib.axes._subplots.AxesSubplot at 0x19d2579e148>
```



Otra forma más resumida de generar un histograma

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

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



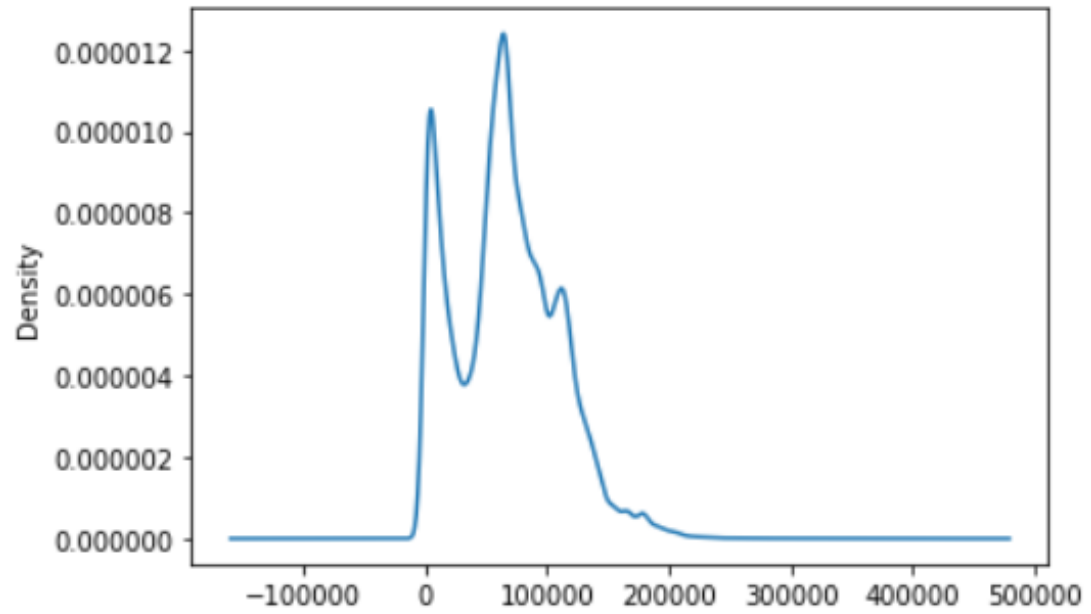
Nótese que los intervalos de mayor frecuencia se encuentran en torno a los 60 mi dólares, y que también hay un intervalo de valores próximos a cero que tiene una frecuencia alta. **Este puede ser otro insigth!!!**

Diagrama de Distribución

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

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

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

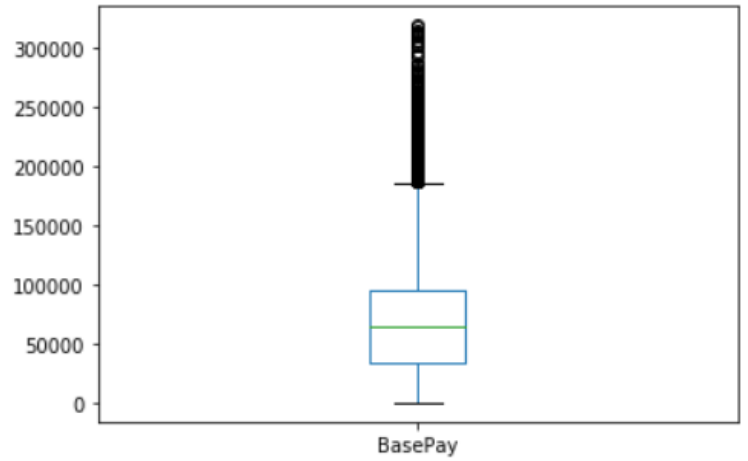


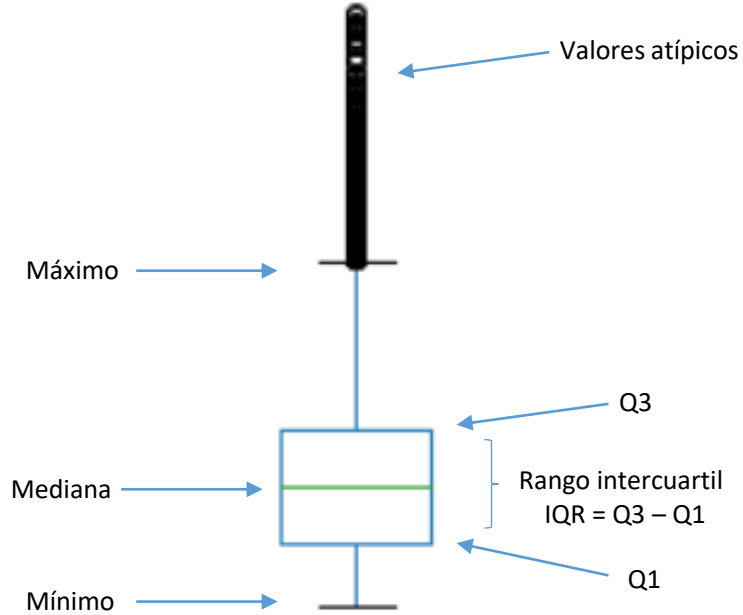
Medidas de Posición

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

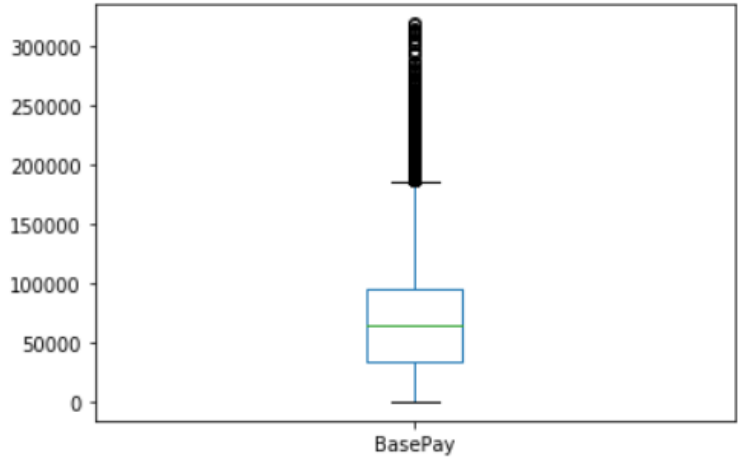
Un valor atípico en un set de datos corresponde a una observación que es numéricamente distantes del resto de los datos, extremadamente grande o extremadamente pequeña. Un valor atípico puede generar un efecto desproporcionado en los resultados estadísticos, como la media, lo cual puede conducir a interpretaciones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



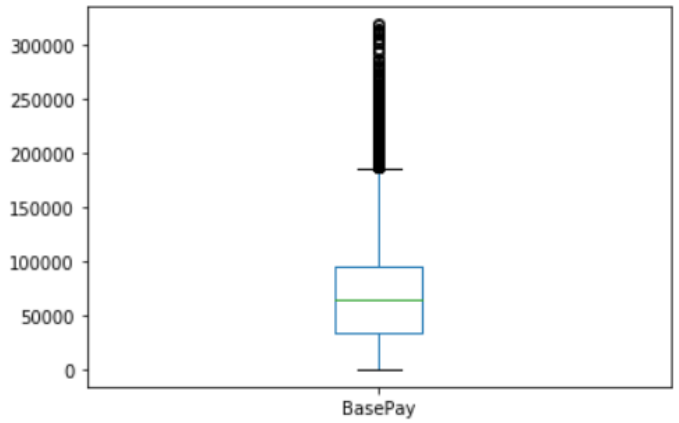
KIBERNUM

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

Gracias

KIBERNUM