

Trabajo Practico - Ejercicio 3 Seleccionen un resultado que les parezca relevante a partir de alguno de los ejercicios del entregable. Diseñe e implemente una comunicación en base a este mensaje, en un archivo PDF.

Elija las palabras y visualización más adecuada para que la comunicación sea entendible, efectiva para:

2. Una publicación científica o reporte técnico interno. No más de una página A4:

2. Ejemplo: La sección de resultados de [IZA DP No. 12914: The Impact of a Minimum Wage Change on the Distribution of Wages and Household Income](#).

3. Ejemplo: Alguna de las secciones de [Temporary reduction in daily global CO2 emissions during the COVID-19 forced confinement](#)

4. Puntos clave:

3. Nivel de detalle técnico requerido. Es necesario justificar la validez del análisis.

4. La idea presentada puede ser más compleja. Pueden asumir que la audiencia tiene conocimiento técnico y va a analizar las visualizaciones en detalle.

5. Pueden presentar más en detalle las limitaciones del análisis (significancia estadística, etc.)

6. No es necesario que mencionen objetivos ni descripciones del conjunto de datos, se supone que eso ya estaría explicado en otras secciones del informe.

```
1 import io
2 import matplotlib
3 import matplotlib.pyplot as plt
4 import numpy as np
5 import pandas as pd
6 import seaborn
7
8 pd.options.mode.chained_assignment = None
9
10 seaborn.set_context('talk')
```

▼ Descripción del DataSet

Para analizar la distribución de los salarios entre las personas que trabajan en el rubro de IT en Argentina con una perspectiva de género, partimos de la base de datos de la encuesta de salarios realizada por Open Qube y publicada en el blog de SysArmy¹ para el año 2020².

La descripción de la metodología de muestreo empleada para la construcción y la

✓ 0s completed at 12:14 AM



TABLA I

```

1 url = 'https://cs.famaf.unc.edu.ar/~mteruel/datasets/diplodatos/sysarmy_surve
2 df = pd.read_csv(url)
3 filas = df.shape[0]
4 columnas = df.shape[1]
5 print (f"La base de datos cuenta con {columnas} columnas y {filas} filas")
6 df[['profile_gender', 'salary_monthly_NETO']].groupby('profile_gender').descr
7

```

La base de datos cuenta con 48 columnas y 6095 filas

salary_monthly_NETO						
	count	mean	std	min	25%	50%
profile_gender						
Hombre	4944.0	98836.063558	102159.664237	1.0	53285.0	80000.0
Mujer	908.0	73501.414465	56256.880181	2.0	43818.0	65000.0
Otros	31.0	95441.354839	88635.674358	40.0	39500.0	74000.0

Exploración y Curación

No se cuenta con valores faltantes. Para el presente analisis se consideraron valores atipicos salarios netos declarados menores a ARS 1000. La descriptiva del nuevo dataset puede visualizarse en la Tabla II.

Los data sets curados a emplear se denominan Grupo A para los datos provenientes de individuos de género masculino y Grupo B para otros géneros, incluyendo el género femenino.

TABLA II

```

1 df[df.salary_monthly_NETO > 1000][['profile_gender', 'salary_monthly_NETO']].

```

salary_monthly_NETO						
	count	mean	std	min	25%	50%
profile_gender						

```
6
7 o_submil = sum((df.salary_monthly_NETO < 1000) & is_man)
8 h_submil = sum((df.salary_monthly_NETO < 1000) & ~is_man)
9
10 muestra_hombre = df[(df.salary_monthly_NETO > 1000) & is_man]
11 muestra_no_hombre = df[(df.salary_monthly_NETO > 1000) & ~is_man]
12
13 print (f"{o_submil} entradas de género masculino")
14 print (f"{h_submil} entradas de otros géneros")
```

```
Salarios Mensuales Netos < 1000AR$
389 Entradas Eliminadas (6%)
129 entradas de género masculino
47 entradas de otros géneros
```

```
1 muestra_hombre[['profile_gender', 'salary_monthly_NETO']].describe()
```

salary_monthly_NETO	
count	4.815000e+03
mean	1.014816e+05
std	1.022153e+05
min	1.100000e+03
25%	5.500000e+04
50%	8.000000e+04
75%	1.100000e+05
max	2.080000e+06

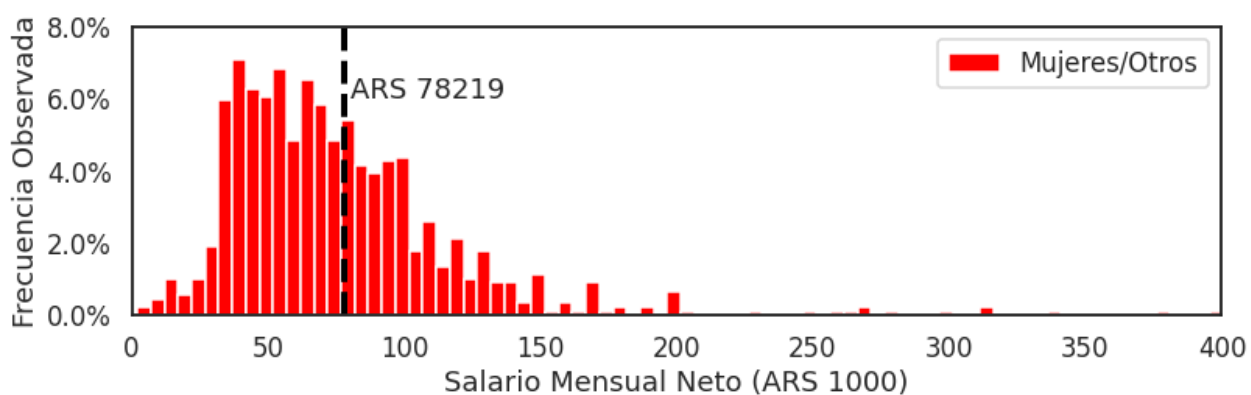
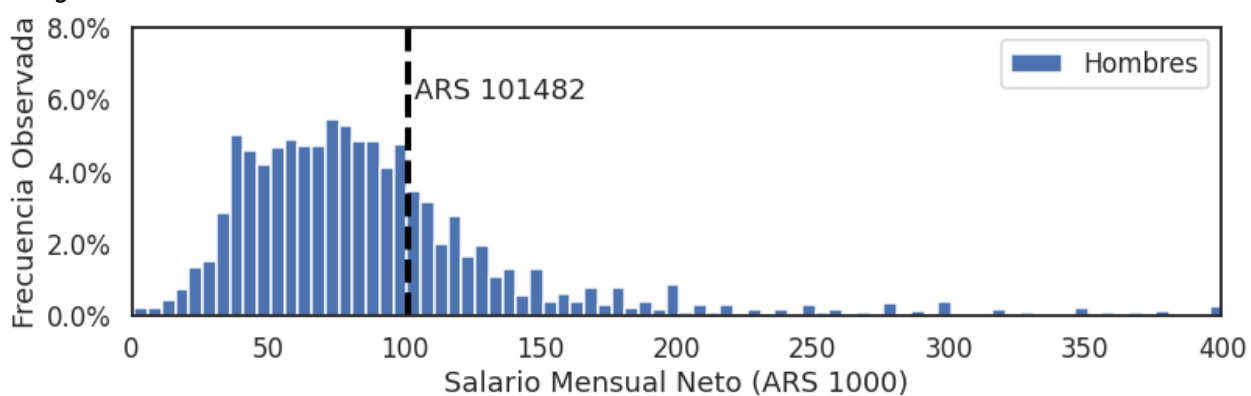
```
1 muestra_no_hombre[['profile_gender', 'salary_monthly_NETO']].describe()
```

salary_monthly_NETO	
count	891.000000
mean	78219.250943

Figura 1 - Histograma de Frecuencias

```
Text(32.0, 0.5, 'Frecuencia Observada')
```

```
<Figure size 432x288 with 0 Axes>
```

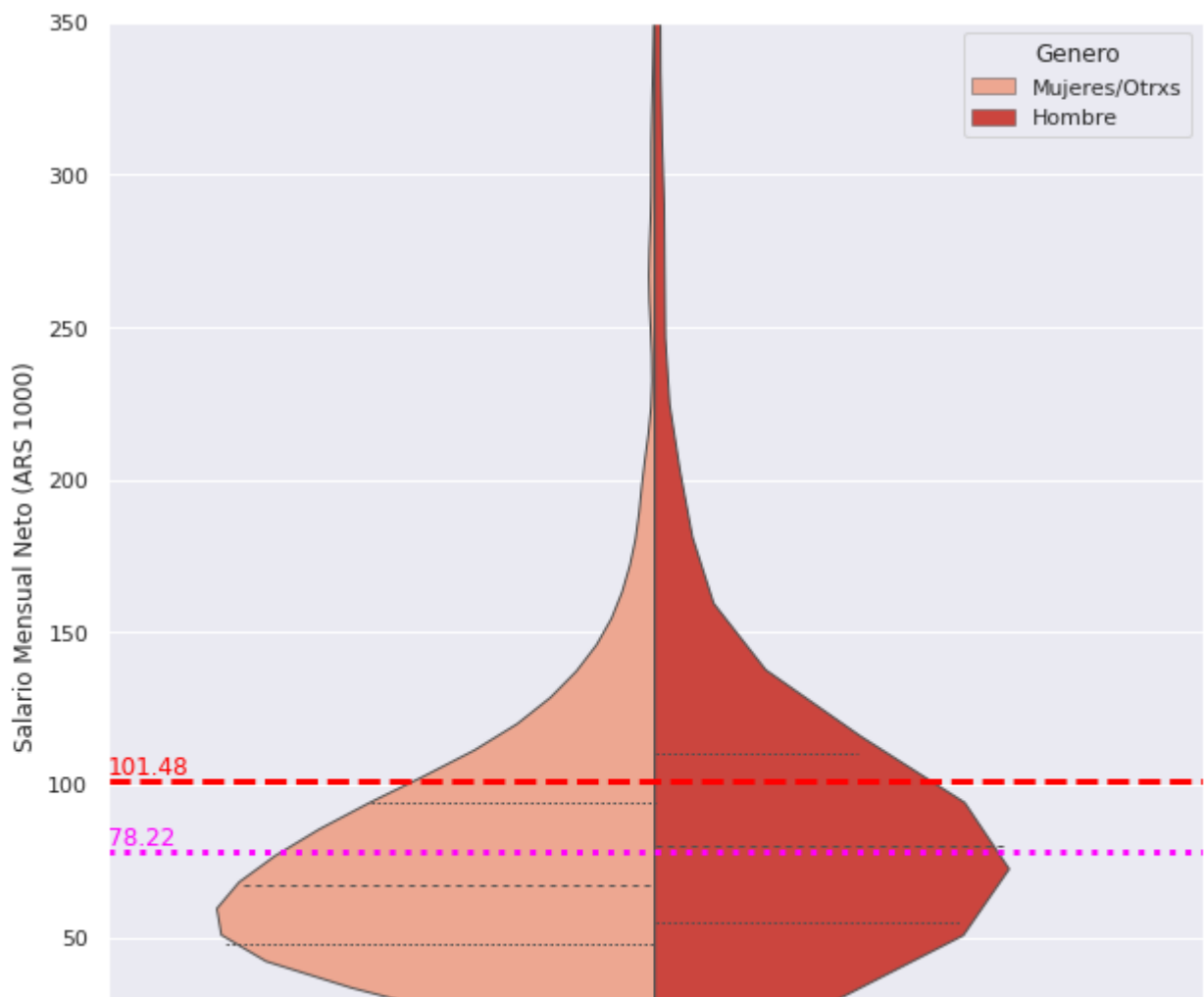


Análisis de datos de la encuesta de @sysarmy a programadxres de @Argentina revela que el salario neto para mujeres y disidencias es un 23% menor respecto al de los hombres. Cuando manipular la información es un negocio al servicio del poder, recurrir a los datos originales es una necesidad.

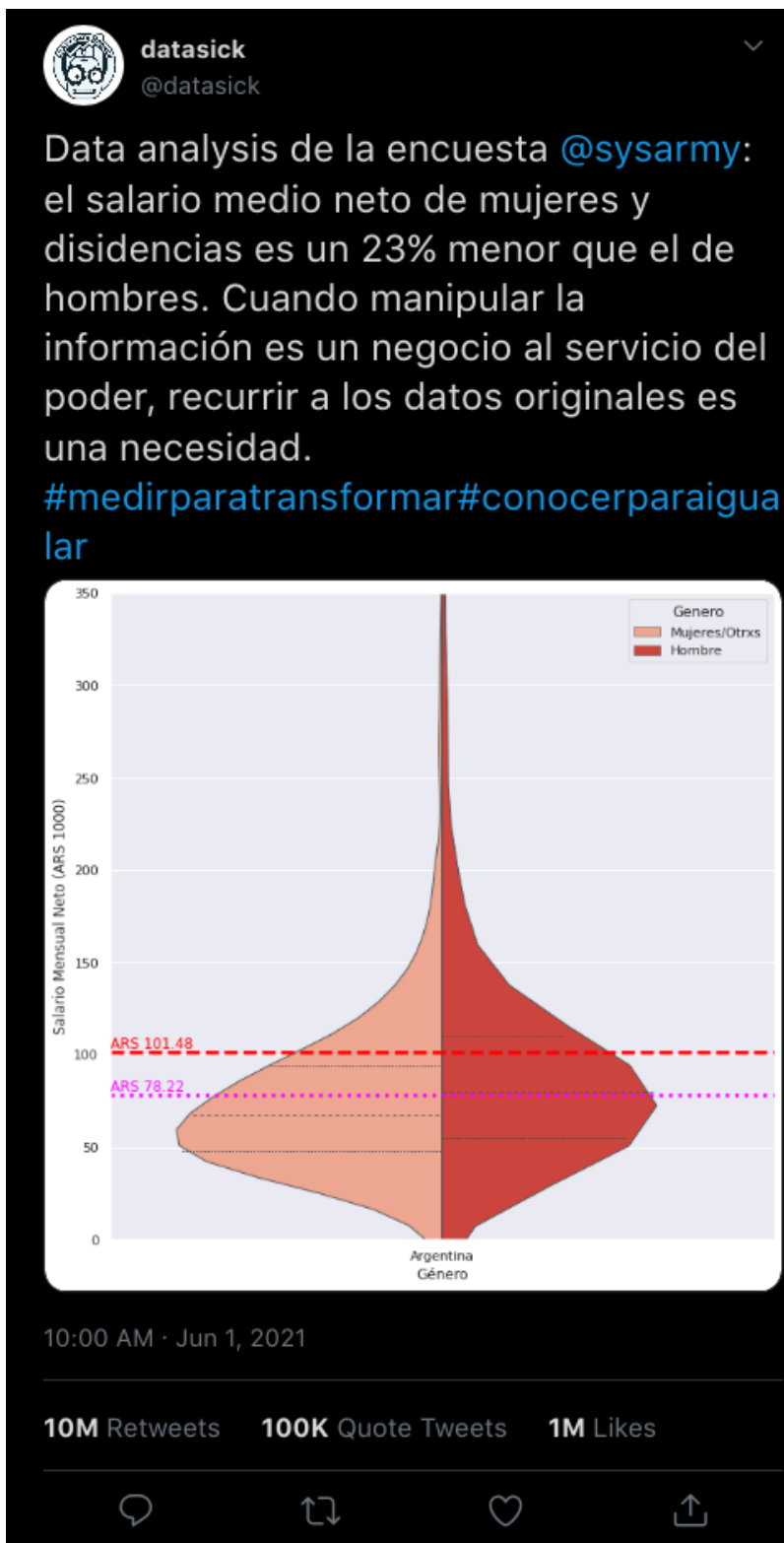
#datos como una heramienta para la igualdad #medirparatransformar #conocerparaigualar #analizarparaopinar #paracambiarlarealidadhayqueconocerla

Default title text

```
Text(0, 0.5, 'Salario Mensual Neto (ARS 1000)')
```



SHOW CODE



Para el análisis supondremos la independencia de ambas muestras de acuerdo a la metodología empleada por Open Qube.

Diferencia Estadísticamente Significativa

Para descartar efectos aleatorios en la estimación de la diferencia de las medias muestrales se emplea una Prueba de Hipótesis para Diferencia de Medias Muestrales, para dos muestras independientes.

Hipótesis Nula

$$H_0 \rightarrow \mu_a - \mu_b = 0$$

Estadístico

$$(t - 1) = (\bar{X}_a - \bar{X}_b) - (\mu_a - \mu_b)$$

Estimación

En caso de existir evidencias de una diferencia estadísticamente significativa para las dos subpoblaciones se procede a realizar una estimación.

Por las propiedades de la función Esperanza $E(X)$ la esperanza de la diferencia entre dos variables aleatorias es la diferencia de sus esperanzas marginales i.e.

$$E(\overline{X}_a - \overline{X}_b) = E(\overline{X}_a) - E(\overline{X}_b) = \mu_a - \mu_b.$$

()


```
3 (_, pvalue) = scipy.stats.levene(groupA, groupB)
4
5 print (f"Se {'rechaza' if pvalue < alpha else 'Acepta'} H0 (Varianzas Iguales
    Se rechaza H0 (Varianzas Iguales) para una significancia del 5% con pvalue
```

