Taller 2 - Análisis Estadístico de Datos (Estado civil y Salario)

Alex Javier Ramirez Pérez

Ulises Ortega Revelo

Corporación Universitaria Autónoma del Cauca

Ingeniería Electrónica

Electiva II

Cristhian Alejandro Cañar Muñoz

20 de febrero de 2025

1. INTRODUCCION

En el segundo taller de análisis de datos, se nos proporcionó un archivo CSV con información diversa, del cual seleccionamos dos columnas para su análisis: salario (una variable cuantitativa) y Estado civil (una variable cualitativa). Este trabajo tiene como objetivo explorar y comprender la distribución de estas variables, así como identificar posibles relaciones entre ellas, utilizando técnicas básicas de estadística descriptiva y visualización de datos.

Para llevar a cabo este análisis, se emplearon herramientas como Python y Jupyter, las cuales nos permitieron manipular, procesar y visualizar los datos de manera eficiente. A través de librerías como pandas, matplotlib y seaborn, realizamos un análisis detallado que incluyó el cálculo de medidas estadísticas (como medias, medianas y proporciones) y la creación de gráficos que facilitan la interpretación de los resultados.

2. CODIGO Y PROCEDIMIENTO

```
#Cargar las librerias
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#Cargar los datos del archivo CSV
data_importada = pd.read_csv(r'C:\Codigo_Python\Mi df minable\mi_df_minable.csv',sep=',',decimal='.')
df_data_csv = pd.DataFrame(data_importada)
```

- Se importa la librería pandas y se le asigna el alias pd
- Se iimporta la librería **matplotlib.pyplot** y se le asigna el alias **plt**
- Se importa la librería **seaborn** y se le asigna el alias **sns**
- **data_importada = pd.read_csv(r'C:\Codigo_Python\Mi df minable\mi_df_minable.csv', sep=',', decimal='.')** - Lee el archivo CSV ubicado en la ruta especificada y lo carga en la variable
- **df_data_csv = pd.DataFrame(data_importada)** - Convierte los datos cargados en un DataFrame y los asigna a la variable df_data_csv

```
# Contar la cantidad de cada categoría en Estado_civil
estado_civil_counts = df_data_csv["Estado_civil"].value_counts()

# Crear la gráfica de pastel
plt.figure(figsize=(8, 8))
plt.pie(estado_civil_counts, labels=estado_civil_counts.index, autopct='%1.1f%%', startangle=140,
        colors=["skyblue", "lightcoral", "lightgreen", "gold", "violet", "orange"])
plt.title("Distribución del Estado Civil")
plt.show()
```

CODIGO PARA GRAFICAR LA DISTRIBUCIÓN DEL ESTADO CIVIL DE ACUERDO AL DATASET

- **estado_civil_counts = df_data_csv["Estado_civil"].value_counts()** - Cuenta la cantidad de registros para cada categoría en la columna de Estado_civil y lo guarda en la variable.
- **plt.figure(figsize=(8, 8))** – Crea una figura con el tamaño 8x8
- **plt.pie(estado_civil_counts, labels=estado_civil_counts.index, autopct='%1.1f%%', startangle=140, colors=["skyblue", "lightcoral", "lightgreen", "gold", "violet", "orange"])** -Crea un grafico pastel donde se le pasa el conteo en porcentaje y las categorías de Estado_civil, especificando los colores y el angulo.
- **Plt.tittle(“Distribucion del Estado Civil”), plt.show()** – Añade un nombre al grafico y lo imprime

```python
# Verificamos que la columna "salario" sea numérica
df["salario"] = pd.to_numeric(df["salario"], errors="coerce")

# Configurar estilo de seaborn
sns.set_style("whitegrid")

# Crear figura con dos gráficos
fig, axes = plt.subplots(1, 2, figsize=(14, 5))

# Histograma para visualizar la distribución del salario
sns.histplot(df["salario"], bins=10, kde=True, ax=axes[0], color="skyblue")
axes[0].set_title("Distribución del Salario")
axes[0].set_xlabel("Salario")
axes[0].set_ylabel("Frecuencia")

# Boxplot para analizar la dispersión y valores atípicos
sns.boxplot(x=df["salario"], ax=axes[1], color="lightcoral")
axes[1].set_title("Caja de Bigotes del Salario")
axes[1].set_xlabel("Salario")

# Mostrar gráficos
plt.tight_layout()
plt.show()
```

CODIGO PARA GRAFICAR LA DISTRIBUCION DEL SALARIO DE ACUERDO AL DATASET

- **df["salario"] = pd.to_numeric(df["salario"], errors="coerce")** – Convierte la columna de Salario a tipo numerico, y si hay valores que no lo son, los convierte a NAN
- **sns.set_style("whitegrid")** – Configura el estilo de la cuadricula
- **fig, axes = plt.subplots(1, 2, figsize=(14, 5))** – Crea dos graficos al lado, con un tamaño de 14x5
- **sns.histplot(df["salario"], bins=10, kde=True, ax=axes[0], color="skyblue")** – Crea un histograma con 10 intervalos y una curva
- **axes[0].set_title("Distribución del Salario")** – Titulo de la primer grafica
- **axes[0].set_xlabel("Salario")** – Titulo eje X
- **axes[0].set_ylabel("Frecuencia")** – Titulo eje Y
- **sns.boxplot(x=df["salario"], ax=axes[1], color="lightcoral")** – Crea una caja de bigotes

- **axes[1].set_title("Caja de Bigotes del Salario")** – Titulo de la segunda grafica
- **axes[1].set_xlabel("Salario")** – Titulo eje X
- **plt.tight_layout()** – Ajustar espacio entre los gráficos
- **plt.show()** – Imprime

```
plt.figure(figsize=(10, 6))
sns.boxplot(x="Estado_civil", y="salario", data=df, palette="Set2")
plt.title("Distribución de Salarios por Estado Civil")
plt.xlabel("Estado Civil")
plt.ylabel("Salario")
plt.show()
```

CODIGO PARA GRAFICAR LA DISTRIBUCION DEL SALARIO EN RELACION AL ESTADO CIVIL

- **plt.figure(figsize=(10, 6))** – Crea una figura de 10x6 de tamaño
- **sns.boxplot(x="Estado_civil", y="salario", data=df, palette="Set2")** – Crea una caja de bigotes, y establecemos los títulos en el eje X y Y, especificamos que los datos vienen del dataframe y añadimos una paleta de colores.
- **plt.title("Distribución de Salarios por Estado Civil")** – Titulo del grafico
- **plt.show()** – Imprimimos el grafico

```python
# Convertir "salario" a numérico
df["salario"] = pd.to_numeric(df["salario"], errors="coerce")

# Calcular medidas estadísticas
media = df["salario"].mean()
mediana = df["salario"].median()
moda = df["salario"].mode()[0]  # Tomar la primera moda si hay más de una

# Configurar estilo de seaborn
sns.set_style("whitegrid")

# Crear el histograma con KDE
plt.figure(figsize=(10, 5))
sns.histplot(df["salario"], bins=30, kde=True, color="skyblue", alpha=0.7)

# Agregar líneas de media, mediana y moda
plt.axvline(media, color='red', linestyle='dashed', linewidth=2, label=f"Media: {media:.2f}")
plt.axvline(mediana, color='green', linestyle='dashed', linewidth=2, label=f"Mediana: {mediana:.2f}")
plt.axvline(moda, color='purple', linestyle='dashed', linewidth=2, label=f"Moda: {moda:.2f}")

# Etiquetas y título
plt.title("Distribución del Salario con Media, Mediana y Moda")
plt.xlabel("Salario")
plt.ylabel("Frecuencia")
plt.legend()

# Mostrar gráfico
plt.show()
```

CODIGO PARA GRAFICAR LA DISTRIBUCION DEL SALARIO, DE ACUERDO A LA MEDIA, MEDIANA Y MODA.

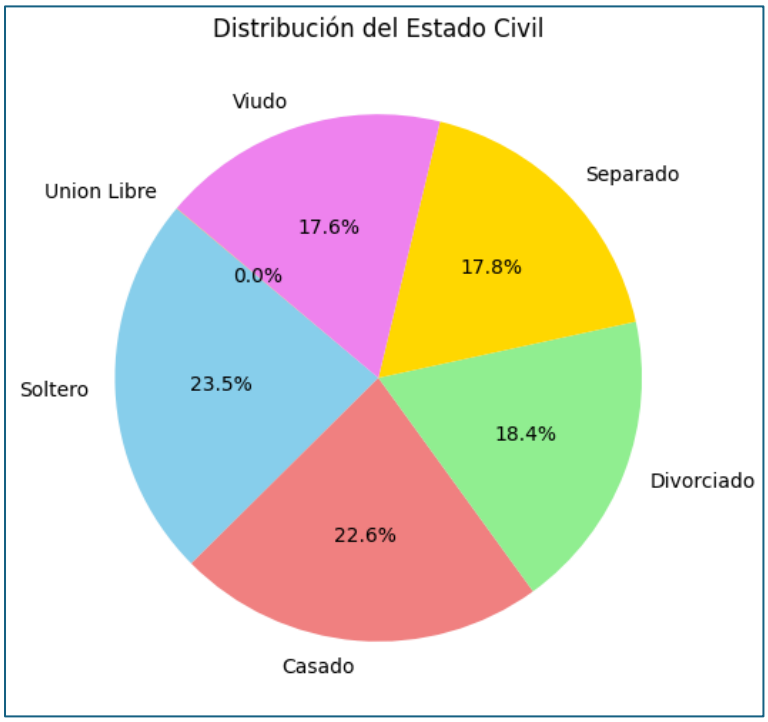
- **df["salario"] = pd.to_numeric(df["salario"], errors="coerce")** – Convierte la columna de Salario a tipo numerico, y si no valores de estos, los convierte en NAN
- **media = df["salario"].mean()** – Calcula la media de Salario
- **mediana = df["salario"].median()** – Calcula la mediana de Salario
- **moda = df["salario"].mode()[0]** – Calcula la moda de Salario, y si hay mas de una, solo se toma la primera
- **sns.set_style("whitegrid")** – Configura el grafico para que tenga una cuadricula
- **plt.figure(figsize=(10, 5))** – Crea una figura con un tamaño de 10x5
- **sns.histplot(df["salario"], bins=30, kde=True, color="skyblue", alpha=0.7)** – Crea un histograma con 30 intervalos y una curva

- **plt.axvline(media, color='red', linestyle='dashed', linewidth=2,**

label=f"Media: {media:.2f}") – Añade una línea vertical en la posición del valor de la media color rojo

- **plt.axvline(mediana, color='green', linestyle='dashed', linewidth=2, label=f"Mediana: {mediana:.2f}")** – Añade una linea vertical en la posición de la mediana color verde
- **plt.axvline(moda, color='purple', linestyle='dashed', linewidth=2, label=f"Moda: {moda:.2f}")** – Añade una línea vertical en la posición de la moda color morado
- **plt.title("Distribución del Salario con Media, Mediana y Moda")** – Titulo del grafico
- **plt.legend()** – Muestra la leyenda con las líneas de media, mediana y moda.

3. ANALISIS CUALITATIVO



El grupo con mayor representación es el de personas **solteras** con un **23.5%**, lo que sugiere que una proporción significativa de los solicitantes no tiene compromisos matrimoniales. Este factor podría estar relacionado con una mayor independencia financiera o, por el contrario, con una menor estabilidad económica en comparación con otros estados civiles. El grupo de **casados (22.6%)** también tiene una fuerte presencia, lo que podría indicar que

muchas personas en matrimonio recurren a préstamos para cubrir necesidades familiares o mejorar su calidad de vida.

Los estados de **divorciado (18.4%)** y **separado (17.8%)** representan un porcentaje considerable, lo que sugiere que un número importante de solicitantes ha pasado por una ruptura matrimonial, situación que podría afectar sus finanzas personales y su capacidad de pago.

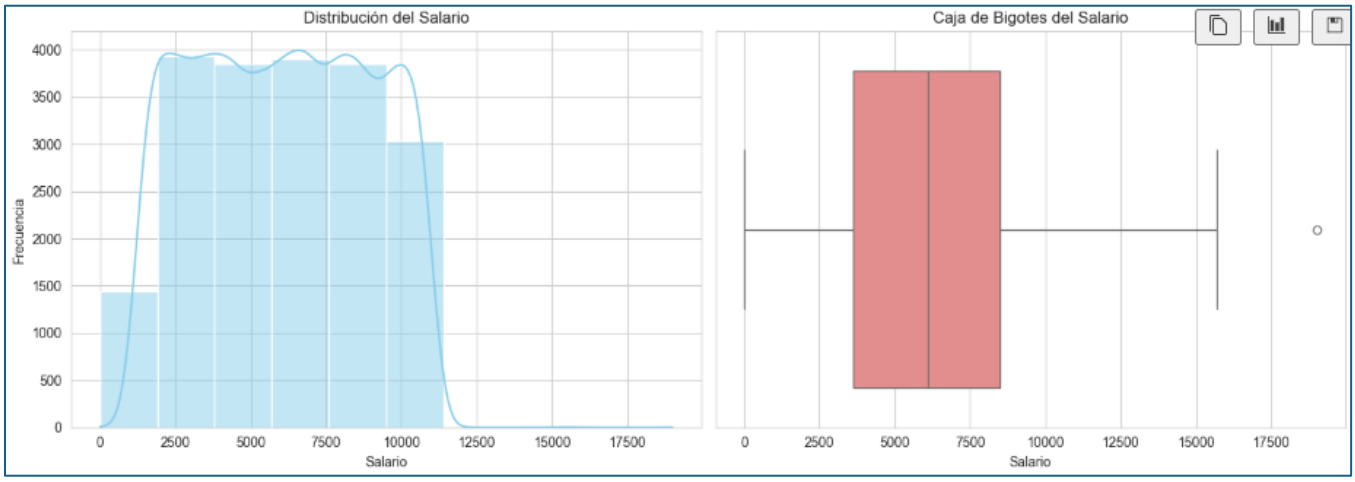
El **grupo de viudos (17.6%)** también tiene una presencia significativa, lo que podría reflejar que muchas de estas personas buscan crédito para afrontar gastos posteriores a la pérdida de su pareja o mantener su estabilidad financiera.

Finalmente, el estado de **unión libre (0.0%)** no tiene representación en los datos, lo que podría indicar una omisión en la recolección de información o que las personas en esta situación no suelen solicitar préstamos dentro de la muestra analizada.

Conclusiones

- La mayoría de los solicitantes de préstamos son solteros (23.5%), lo que podría estar asociado con una mayor necesidad de financiamiento individual.
- El matrimonio (22.6%) también es un estado civil relevante en la solicitud de créditos, posiblemente debido a las responsabilidades compartidas y necesidades familiares.
- La proporción de personas divorciadas y separadas (sumando un 36.2%) es alta, lo que podría indicar que la disolución del matrimonio impacta la estabilidad financiera y genera una mayor dependencia del crédito.
- La viudez (17.6%) refleja que un segmento de la población solicita préstamos tras la pérdida de su pareja, lo que podría relacionarse con la necesidad de ajustar su situación económica.
- La ausencia de personas en unión libre (0.0%) requiere una revisión de la base de datos o un análisis más profundo sobre la participación de este grupo en la solicitud de préstamos.

ANALISIS CUANTITATIVO



Histograma de la Distribución del Salario

- La mayoría de los salarios se encuentran concentrados entre **2,500 y 10,000**.
- Se observa una distribución relativamente uniforme en este rango, con ligeras variaciones en la frecuencia.
- Existe una caída abrupta en la frecuencia de los salarios que superan los **10,000**, lo que indica que son menos comunes.
- No se aprecian valores negativos ni una distribución sesgada hacia la derecha o izquierda de manera extrema.

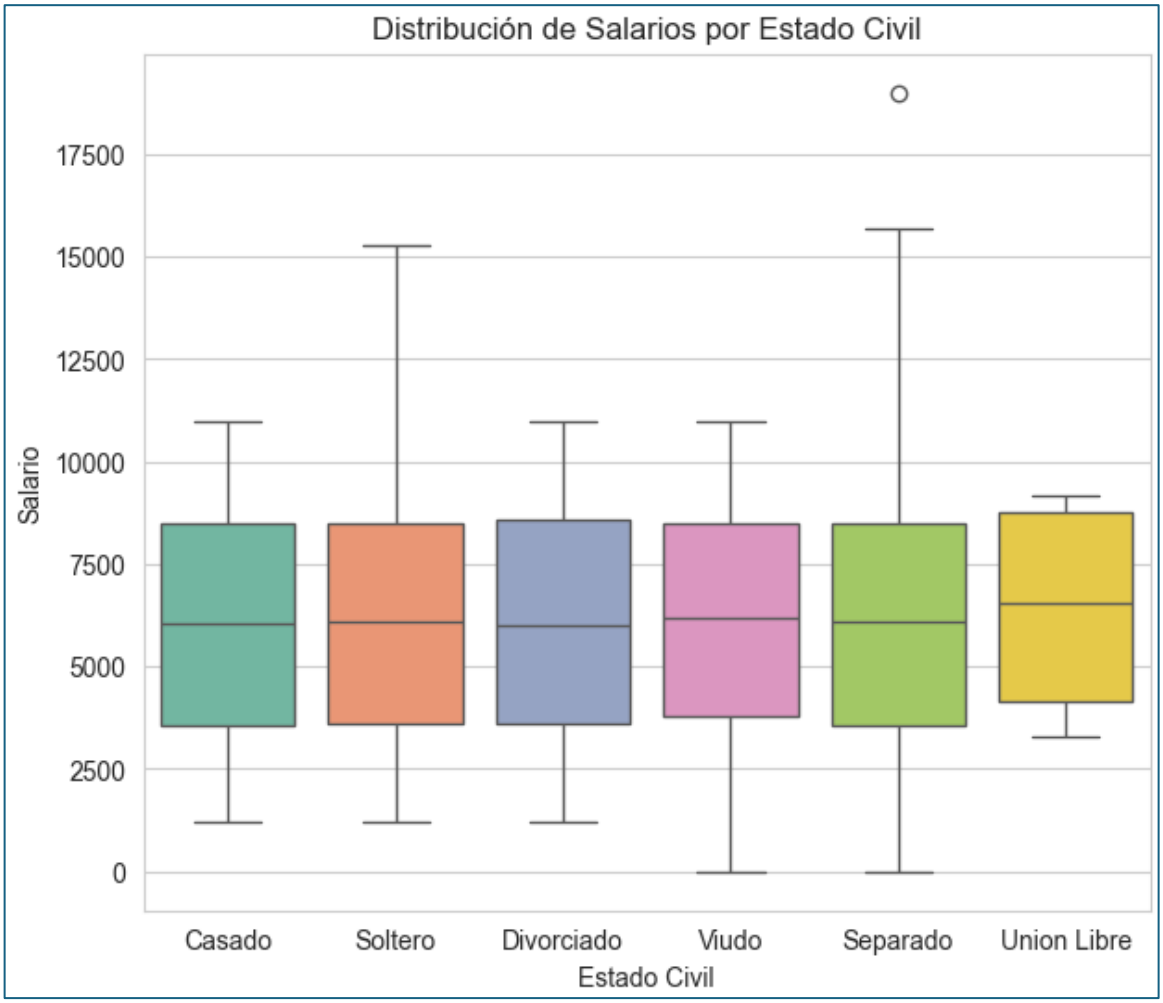
Diagrama de Caja y Bigotes

- La mediana del salario se encuentra aproximadamente en **5,000 a 7,500**.
- Se observa una distribución relativamente simétrica con bigotes que se extienden hasta valores cercanos a **15,000**.
- Se identifica la presencia de un valor atípico superior a **17,500**, lo que sugiere la existencia de un salario significativamente mayor al resto.
- El rango intercuartílico (IQR) está comprendido entre **alrededor de 3,000 y 10,000**, indicando que la mayoría de los salarios se encuentran en este intervalo.

Conclusiones

- La mayor parte de la distribución de los salarios se encuentra dentro del rango de **2,500 a 10,000**, lo que sugiere una concentración de ingresos en este intervalo.
- La distribución presenta una forma relativamente homogénea dentro del rango principal, sin sesgos extremos.

- La presencia de un valor atípico indica que algunos individuos reciben salarios significativamente superiores al resto de la población.
- El análisis de la mediana y el rango intercuartílico sugiere que el 50% de los salarios se encuentran en un rango de **3,000 a 10,000**, proporcionando una mejor comprensión de la tendencia central de los ingresos.



El grupo con mayor representación es el de personas **solteras** con un **23.5%**, seguido por los **casados (22.6%)**. La suma de los divorciados y separados representa un **36.2%**, lo que sugiere una alta incidencia de disoluciones matrimoniales. El grupo de **viudos (17.6%)** también es significativo, mientras que la ausencia de personas en **unión libre (0.0%)** podría indicar una inconsistencia en la recolección de datos.

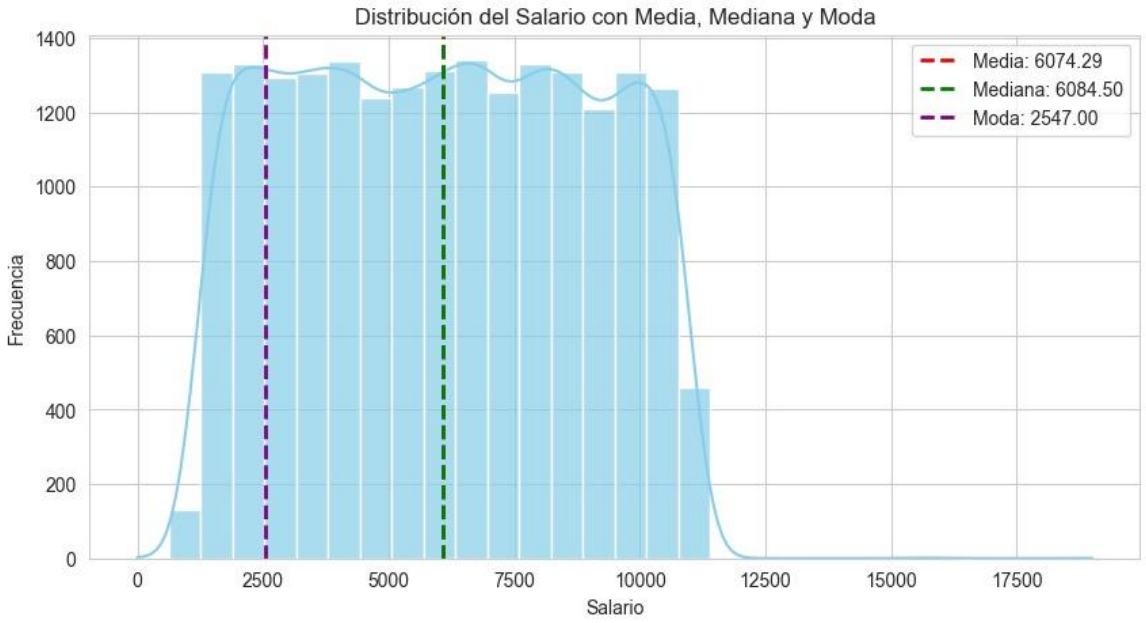
Análisis de la Distribución Salarial

Se han utilizado gráficos estadísticos para representar la distribución de los salarios en la población estudiada. A continuación, se analizan los resultados obtenidos:

1. **Distribución General del Salario**
    - El histograma muestra que los salarios se concentran principalmente en un rango entre **2,500 y 10,000**, con una frecuencia alta en este intervalo.
    - El diagrama de caja revela que la mediana del salario se encuentra aproximadamente en **6,000 - 7,000**, con valores atípicos que superan los **15,000**.
2. **Distribución de Salarios por Estado Civil**
    - Se observa una tendencia similar en la distribución de salarios para los diferentes estados civiles.
    - La mediana del salario en cada categoría es cercana a **6,000 - 7,500**, sin diferencias significativas entre los grupos.
    - Se identifican valores atípicos en la categoría de **casados y separados**, con salarios que superan los **15,000**.
    - Los rangos intercuartílicos (50% de los datos) son similares para todas las categorías, lo que indica que la variabilidad salarial es homogénea en la población analizada.

Conclusiones

- La mayoría de la población es **soltera (23.5%)** y **casada (22.6%)**, lo que sugiere que estas categorías predominan en la sociedad analizada.
- La distribución salarial es relativamente homogénea entre los diferentes estados civiles, con una mediana cercana a **6,000 - 7,500** en todos los casos.
- Se presentan valores atípicos en la distribución de salarios, especialmente en la categoría de **casados y separados**, lo que indica la presencia de personas con ingresos significativamente más altos.
- La falta de representación en la categoría de **unión libre (0.0%)** sugiere una posible inconsistencia en los datos recopilados.



Análisis de la Distribución del Salario con Media, Mediana y Moda

Este informe presenta un análisis detallado de la distribución de los salarios en función de tres medidas estadísticas fundamentales: media, mediana y moda. Los datos utilizados provienen del archivo de Excel proporcionado, lo que permite una evaluación cuantitativa de la estructura salarial.

1. Media (Promedio)

- **Valor:** $6,074.29
- Representada por la línea **roja punteada** en la gráfica.
- Refleja el salario promedio de la muestra analizada.
- Puede estar influenciada por valores atípicos, lo que puede generar una percepción distorsionada de la realidad salarial.

2. Mediana

- **Valor:** $6,084.50
- Representada por la línea **verde punteada**.
- Es el valor central de la distribución; el 50% de los salarios son menores y el otro 50% son mayores.
- Su cercanía a la media sugiere una distribución relativamente equilibrada, aunque con una ligera dispersión.

3. Moda

- **Valor:** $2,547
- Representada por la línea **morada punteada**.
- Es el salario que se repite con mayor frecuencia en la muestra.
- Su valor considerablemente menor que la media y la mediana indica una mayor concentración de empleados con salarios más bajos.

4. Forma de la Distribución

- La distribución presenta características **bimodales**, lo que sugiere la existencia de **dos grupos salariales predominantes**.
- La mayoría de los salarios se encuentran en el rango de **$2,500 a $10,000**, con una disminución en los valores extremos.
- La diferencia entre la moda y la media/mediana sugiere que un grupo de empleados tiene salarios considerablemente más bajos que el promedio.
- La existencia de valores atípicos elevados influye en la media, elevándola artificialmente.

5. Comparación por Estado Civil

- El análisis de la distribución de salarios según el estado civil muestra que no hay diferencias significativas en la mediana salarial entre los grupos.
- Sin embargo, algunos estados civiles presentan una mayor variabilidad salarial, con valores atípicos que superan los $15,000.
- Se observa una ligera tendencia a que los empleados en **unión libre y separados** tengan una mayor dispersión salarial en comparación con otros grupos.

Conclusión

La distribución de salarios en esta muestra refleja una disparidad significativa. Aunque la media y la mediana se mantienen en valores similares, la moda indica que una gran cantidad de empleados perciben salarios bajos. Esto sugiere que, aunque algunos empleados tienen ingresos elevados, una parte considerable de la población tiene remuneraciones reducidas, lo que podría implicar desigualdad en la estructura salarial. Además, la distribución de salarios por estado civil muestra patrones similares entre los grupos, aunque con diferencias en la dispersión de los datos.