Tarea1

August 27, 2025

[35]:
```
import pandas as pd
import numpy as np
```

[36]:
```
salarios=pd.read_csv("/Users/marcobarragan/Documents/
↪Seminario-de-Ciencia-de-datos/salary.csv")
```

[37]:
```
salarios.head()
```

[37]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

[38]:
```
salarios.tail()
```

[38]:

	age	workclass	fnlwgt	education	education-num
32556	27	Private	257302	Assoc-acdm	12
32557	40	Private	154374	HS-grad	9
32558	58	Private	151910	HS-grad	9
32559	22	Private	201490	HS-grad	9
32560	52	Self-emp-inc	287927	HS-grad	9

```
           marital-status          occupation relationship    race      sex  \
32556   Married-civ-spouse        Tech-support         Wife   White   Female
32557   Married-civ-spouse   Machine-op-inspct      Husband   White     Male
32558              Widowed        Adm-clerical    Unmarried   White   Female
32559        Never-married        Adm-clerical    Own-child   White     Male
32560   Married-civ-spouse     Exec-managerial         Wife   White   Female

       capital-gain  capital-loss  hours-per-week  native-country  salary
32556             0             0              38   United-States   <=50K
32557             0             0              40   United-States    >50K
32558             0             0              40   United-States   <=50K
32559             0             0              20   United-States   <=50K
32560         15024             0              40   United-States    >50K
```

[39]:
```
salarios.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64
 5   marital-status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital-gain    32561 non-null  int64
 11  capital-loss    32561 non-null  int64
 12  hours-per-week  32561 non-null  int64
 13  native-country  32561 non-null  object
 14  salary          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

[47]:
```
print("Situación inicial ")
print("Shape:", salarios.shape)

# Conteo de los "?" , estoo porque en el csv, pareciera que los nulos tienen ?
qmarks = (salarios == " ?").sum()

# Nulos reales
nulls_inicio = salarios.isnull().sum()
```

```
print("\nValores '?' por columna:")
print(qmarks[qmarks > 0])

print("\nNulos reales al inicio:")
print(nulls_inicio[nulls_inicio > 0] if (nulls_inicio > 0).any() else "No hay␣
 ↪nulos directos en CSV")

# Duplicados
dup_inicio = salarios.duplicated().sum()
print(f"\nDuplicados al inicio: {dup_inicio}")
```

```
Situación inicial
Shape: (32561, 15)

Valores '?' por columna:
Series([], dtype: int64)

Nulos reales al inicio:
workclass         1836
occupation        1843
native_country     583
dtype: int64

Duplicados al inicio: 24
```

[50]:
```
# Limpiar espacios
obj_cols = salarios.select_dtypes(include="object").columns
for c in obj_cols:
    salarios[c] = salarios[c].str.strip()

# Reemplazamos lso  "?" por NaN para tratarlos como nulos
salarios.replace("?", np.nan, inplace=True)

# Renombrar columnas
salarios.rename(columns={
    "fnlwgt":"final_weight",
    "education-num":"education_num",
    "marital-status":"marital_status",
    "capital-gain":"capital_gain",
    "capital-loss":"capital_loss",
    "hours-per-week":"hours_per_week",
    "native-country":"native_country"
}, inplace=True)
```

[52]:
```
print("\nDespués de limpieza")

nulls_despues = salarios.isnull().sum()
```

```
print("\nNulos después de limpieza:")
print(nulls_despues[nulls_despues > 0] if (nulls_despues > 0).any() else "No␣
 ↪hay nulos después de limpieza")

dup_despues = salarios.duplicated().sum()
print(f"\nDuplicados después de limpieza: {dup_despues}")
```

```
Después de limpieza

Nulos después de limpieza:
workclass         1836
occupation        1843
native_country     583
dtype: int64

Duplicados después de limpieza: 24
```

[53]:
```
# Eliminar duplicados
salarios_clean = salarios.drop_duplicates().copy()

print("\nShape final sin duplicados:", salarios_clean.shape)

# Exportar versión limpia
salarios_clean.to_csv("salary_clean.csv", index=False)
print("Archivo limpio exportado: salary_clean.csv")
```

```
Shape final sin duplicados: (32537, 15)
Archivo limpio exportado: salary_clean.csv
```

0.0.1 Descripción de los datos

El conjunto de datos contiene 32,561 registros y 15 variables con información demográfica y laboral: edad, educación, estado civil, ocupación, país de origen, horas trabajadas por semana y nivel salarial (<=50K o >50K). La edad promedio es de 39 años y la mayoría trabaja alrededor de 40 horas semanales. Predomina la categoría de ingresos <=50K. Se identificaron valores inconsistentes en `workclass`, `occupation` y `native_country`, representados con "?", además de 24 registros duplicados.

0.0.2 Preguntas de análisis propuestas

1. ¿Cómo se relaciona el nivel educativo con la probabilidad de tener un salario >50K?

2. ¿Qué diferencias existen en la distribución de salarios entre hombres y mujeres, y según el estado civil?

3. ¿Qué ocupaciones concentran más ingresos >50K y cómo influyen las horas trabajadas por semana?

4. ¿Existen diferencias salariales según el país de origen, considerando educación y ocupación?

5. ¿Cómo se comportan las variables `capital_gain` y `capital_loss` entre los grupos de salario <=50K y >50K?