

# MEMORIA TÉCNICA

## Análisis Exploratorio de Datos (EDA)

### Proyecto:

Exploratory Data Analysis of [Adult Census Income(Kaggle)]

### Descripción:

Análisis de la estructura, calidad y patrones principales del conjunto de datos con fines exploratorios.

---

**Autor:** Nick Brown

**Curso / Programa:** Data Science Bootcamp

**Institución:** The Bridge Online School

**Fecha:** Diciembre 2025

---

### Herramientas utilizadas:

Python · pandas · numpy · matplotlib · seaborn

## **1. Introducción**

### ***Contexto del problema***

El presente documento recoge la memoria técnica del Análisis Exploratorio de Datos (EDA) realizado sobre un conjunto de datos socioeconómicos basado en el Adult Income Dataset. El objetivo principal del análisis es comprender cómo distintas características demográficas, educativas y laborales se relacionan con el nivel de ingresos de los individuos, diferenciando entre aquellos que ganan más o menos de 50.000 dólares anuales.

Este análisis permite identificar patrones estructurales y relaciones entre variables que influyen en los ingresos, sirviendo como base para análisis posteriores.

### ***Objetivo del análisis***

Explorar, describir e interpretar las relaciones entre las variables del dataset y el nivel de ingresos mediante un enfoque estrictamente exploratorio, sin construir modelos predictivos.

---

## **2. Hipótesis**

Saber cómo influyen los factores demográficos y laborales en el nivel de ingresos.

---

## **3. Preguntas de investigación**

- ¿Un mayor nivel educativo implica una mayor probabilidad de ganar más de 50K?
  - ¿Cómo influyen la edad y las horas trabajadas en los ingresos?
  - ¿Existen diferencias de ingresos según sexo, estado civil u ocupación?
  - Entre los que tienen capital gain, ¿qué factores se relacionan con mayores ganancias?
- 

## **4. Descripción del conjunto de datos**

El conjunto de datos utilizado corresponde al Adult Census Dataset, disponible en la plataforma Kaggle. Está basado en datos del censo de Estados Unidos

El dataset contiene más de 30.000 observaciones y 14 variables que describen características demográficas, educativas, laborales y económicas de los individuos. La

variable objetivo es income, categorizada en  $\leq 50K$  y  $> 50K$ , utilizada para analizar factores asociados a distintos niveles de ingresos.

---

## 5. Limpieza y preparación de los datos

Se trataron valores faltantes, se recodificar variables categóricas, se transformó capital gain mediante logaritmos y se discretizan variables como edad y horas trabajadas para facilitar la interpretación.

---

## 6. Análisis exploratorio

### 6.1 Análisis univariante

Se analizaron las distribuciones individuales de income, workclass, estado civil, ocupación, raza, sexo, edad, educación, horas trabajadas y capital gain, identificando desbalances y asimetrías relevantes.

### Variables categóricas

```
# Creamos dos variables para analizar las columnas categoricas y numericas

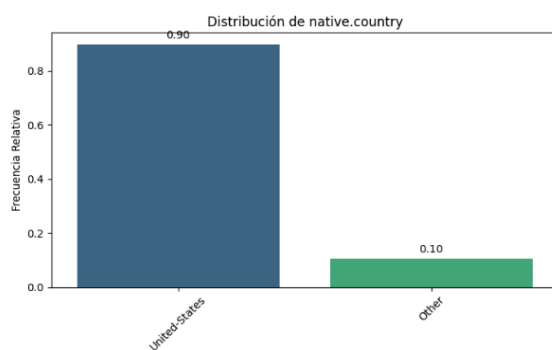
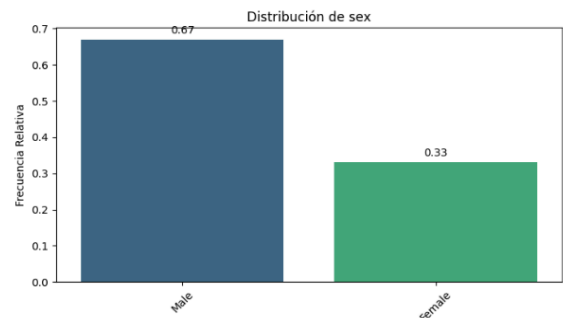
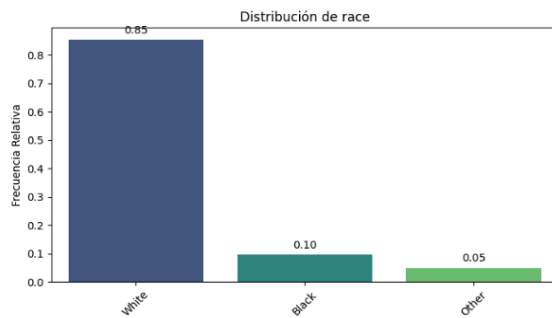
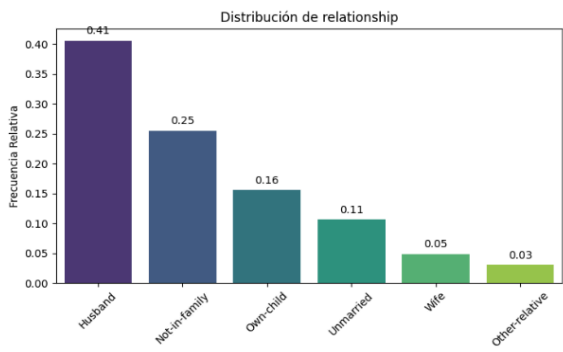
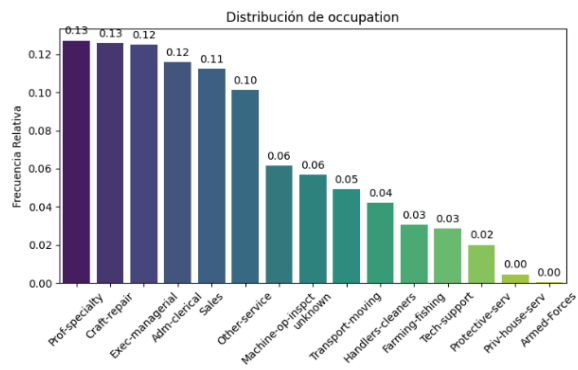
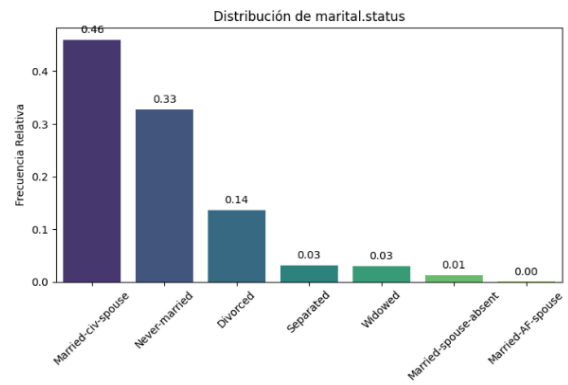
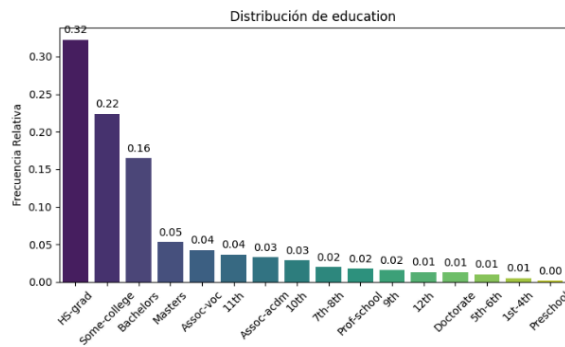
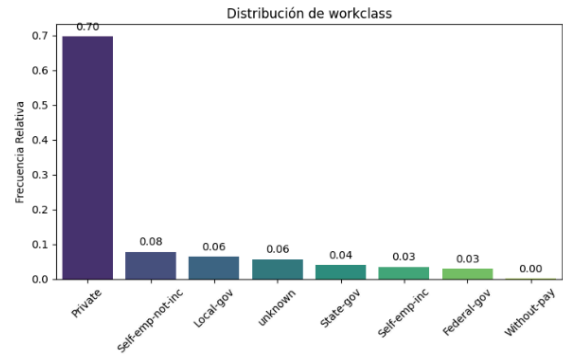
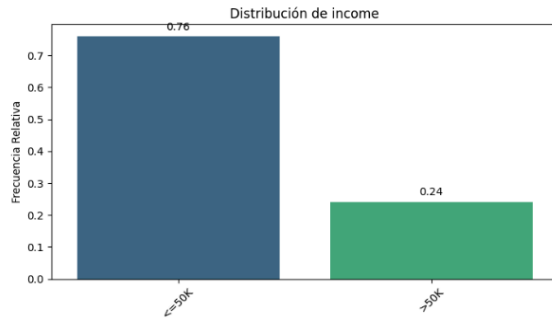
columnas_categoricas =
["income", "workclass", "education", "marital_status", "occupation", "relationships", "race", "sex", "native_country"]

columnas_numericas =
["age", "hours_per_week", "education_num", "capital_gain", "capital_loss"]
```

### *Representar las variables en gráficas*

```
# Utilizamos la función predefinida para representar gráficamente todas las variables categóricas

pinta_distribucion_categoricas(df, columnas_categoricas,
mostrar_valores=True)
```



## Notas variable categoricas

Observamos los gráficos y tomamos notas sobre lo que vemos en la visualización de cada variable. Anotamos cualquier observación interesante y posibles análisis adicionales que podrían resultar relevantes para profundizar más adelante.

Asimismo, decidimos si es necesario tratar o modificar los valores de alguna manera, por ejemplo, creando bins (agrupaciones), renombrando categorías, etc.

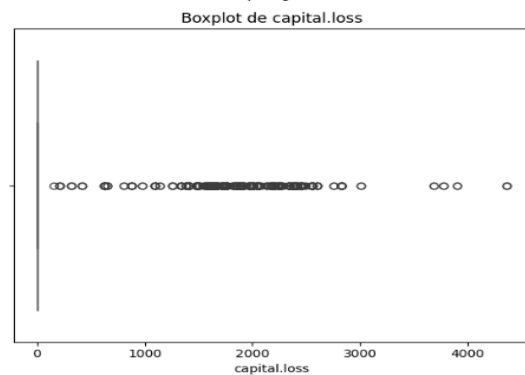
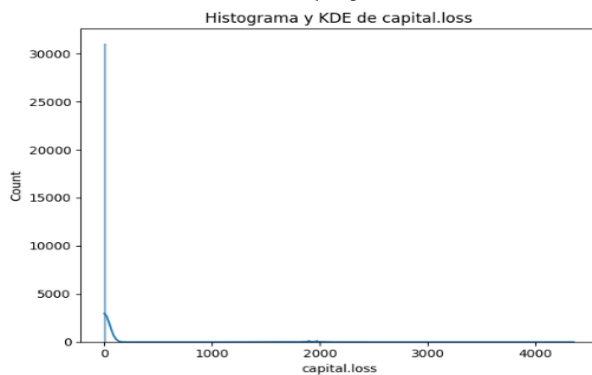
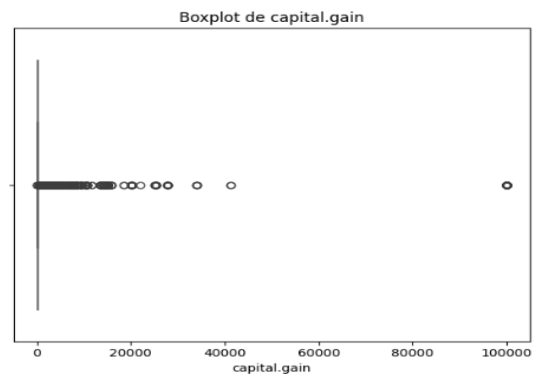
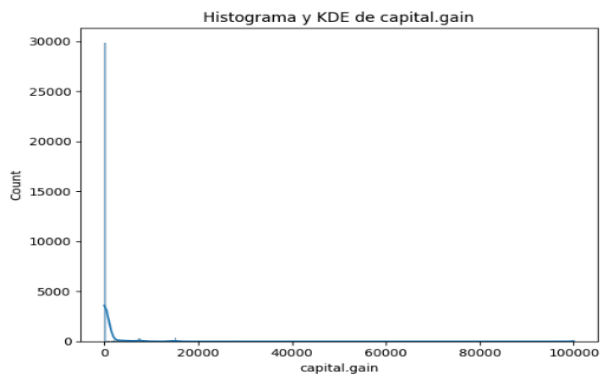
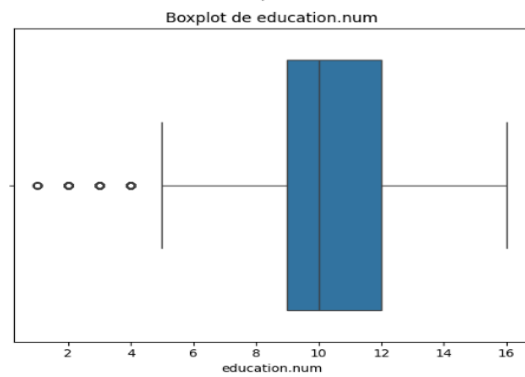
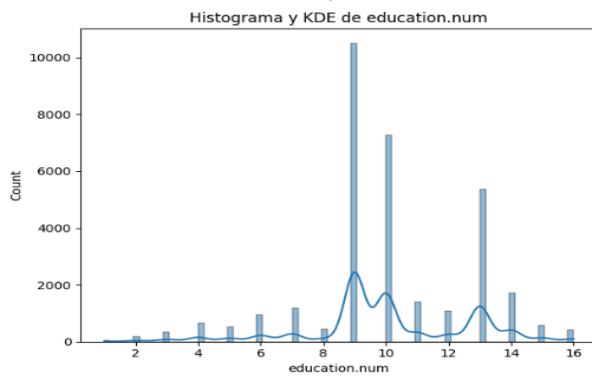
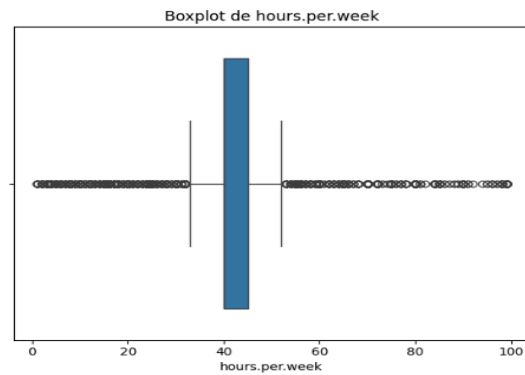
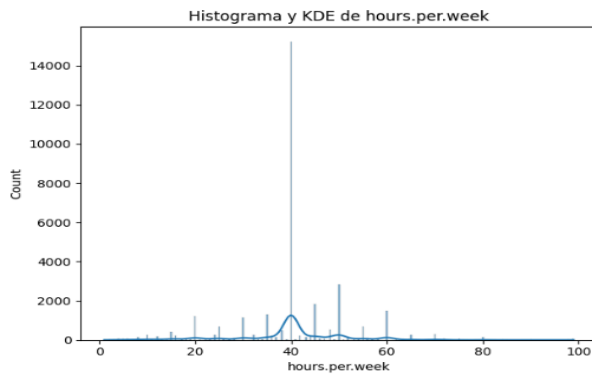
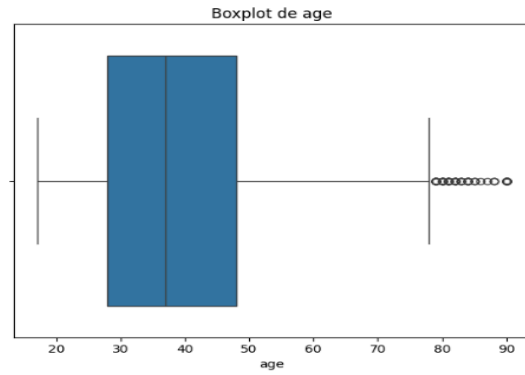
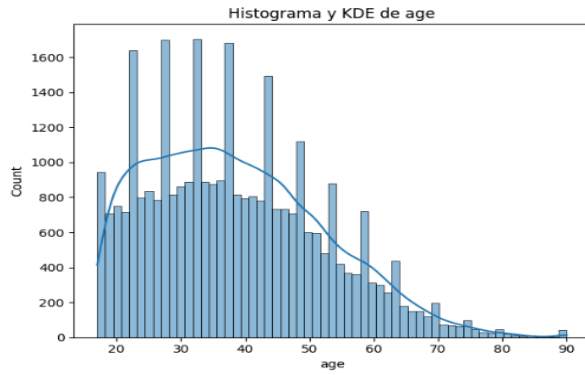
## Variables numericas

### *Mostrar estadísticas descriptivas*

```
# Mostrar estadísticas descriptivas de las variables numéricas  
  
df.describe()
```

### *Representar las variables en gráficos*

```
# Utilizamos la función predefinida para representar gráficamente todas  
las variables categóricas  
  
plot_combined_graphs(df, columnas_numericas)
```



## Notas sobre variables numéricas

Observamos los gráficos y tomamos notas sobre lo que vemos en la visualización de cada variable. Anotamos cualquier observación interesante y posibles análisis adicionales que podrían resultar relevantes para profundizar más adelante.

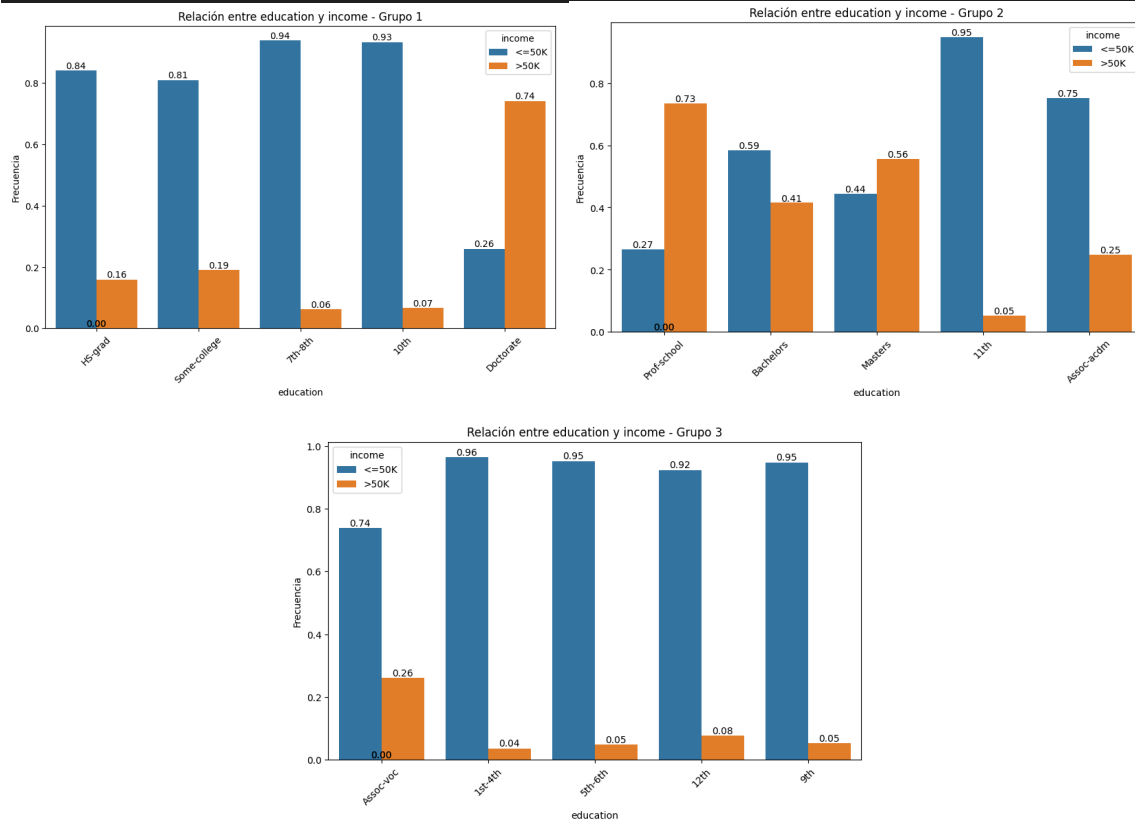
Asimismo, decidimos si es necesario tratar o modificar los valores de alguna manera, por ejemplo, creando bins (agrupaciones), renombrando categorías, etc.

## 6.2 Análisis bivariante

Se estudiaron relaciones entre educación, edad, horas trabajadas, sexo, estado civil, ocupación y capital gain frente al ingreso, observándose relaciones claras con educación, sexo y estado civil.

### Education VS income

```
plot_categorical_relationship_fin(df,"education","income", show_values=True,
relative_freq=True)
```



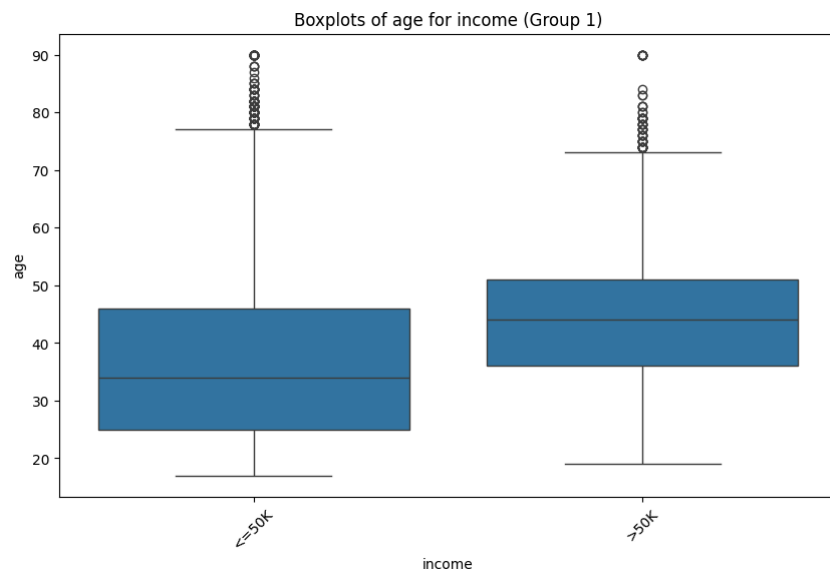
## Age VS income

```
# mostramos estadísticas descriptivas de las variables
```

```
df.groupby("income")["age"].describe()
```

```
# plot boxplots
```

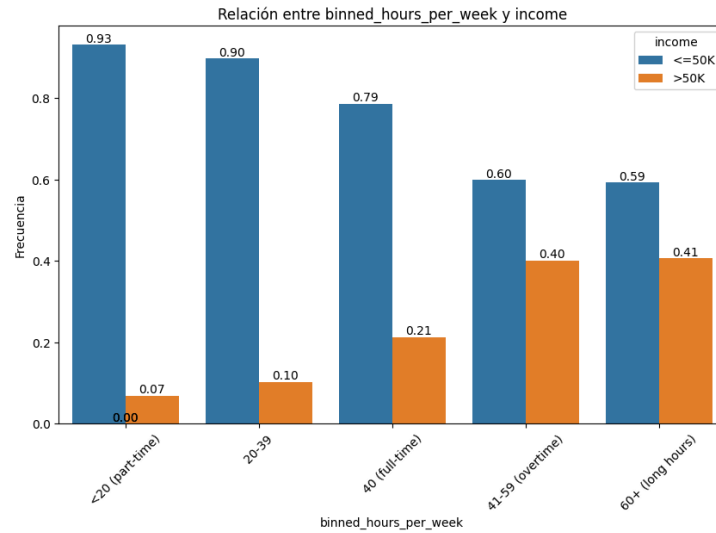
```
plot_grouped_boxplots(df,"income","age")
```



## Education VS income

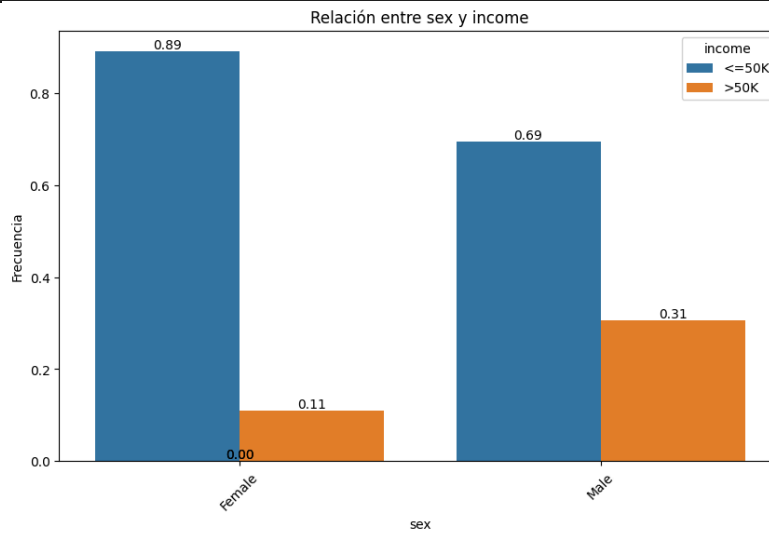
```
# representamos gráficamente la relación entre las variables
```

```
plot_categorical_relationship_fin(df,"binned_hours_per_week","income",relative_freq=True,  
show_values=True)
```



## Sex VS income

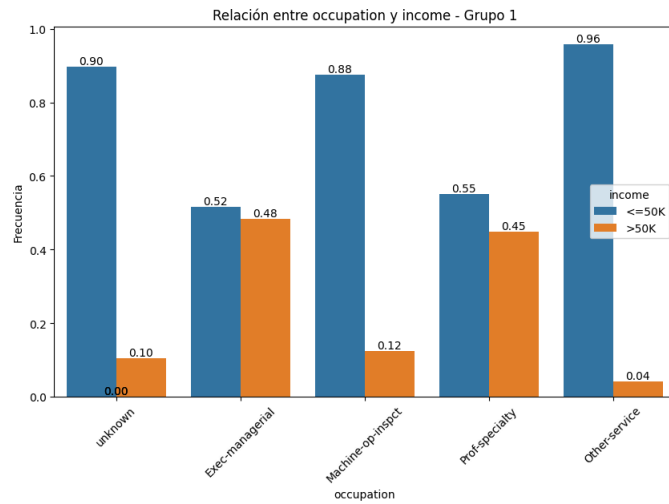
```
# representamos gráficamente la relación entre las variables
plot_categorical_relationship_fin(df,"sex","income",relative_freq=True,show_values=True)
)
```



## Occupation VS income

```
# representamos gráficamente la relación entre las variables

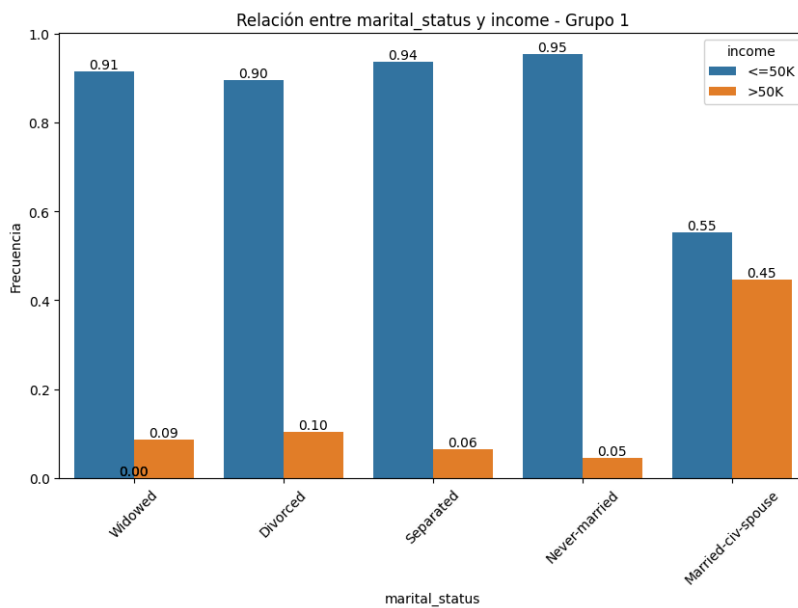
plot_categorical_relationship_fin(df,"occupation","income",relative_freq=True,show_values=True)
```



## Marital status VS income

```
# representamos la relación entre las variables

plot_categorical_relationship_fin(df,"marital_status","income",relative_freq=True,show_values=True)
```

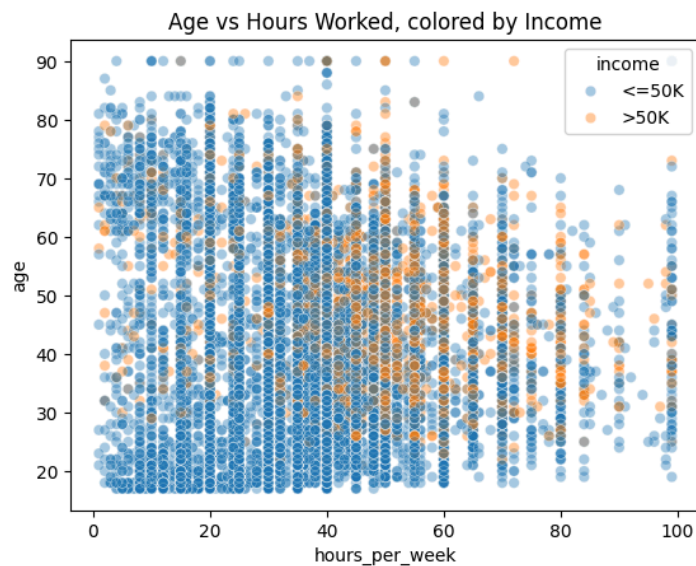


### 6.3 Análisis multivariante

Se analizaron interacciones entre educación, sexo, horas trabajadas e ingresos, así como un análisis específico de capital gain condicionado por nivel de ingresos.

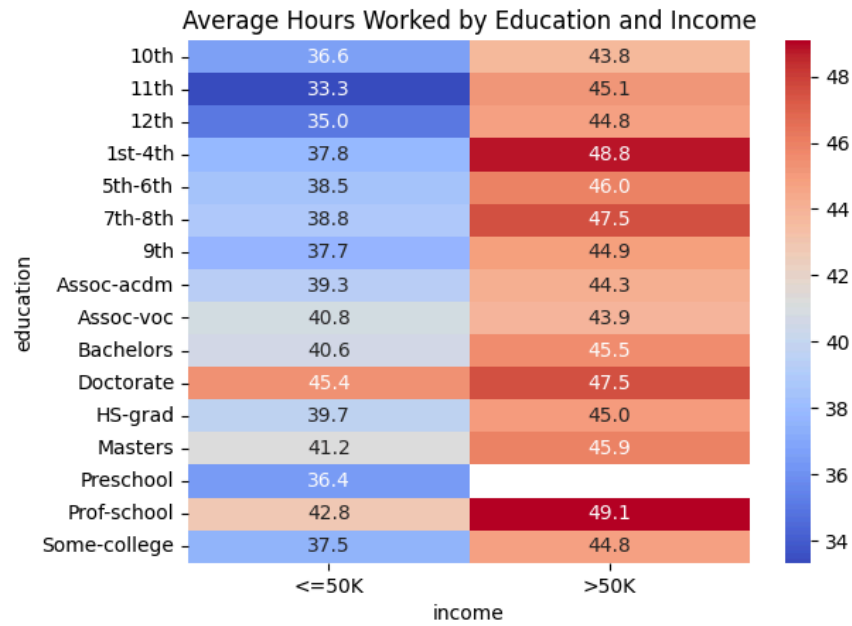
#### age, hours per week, income

```
# representamos la relación entre la variable en un scatterplot usando seaborn
sns.scatterplot(
    data=df,
    y='age',
    x='hours_per_week',
    hue='income',
    alpha=0.4)
plt.title('Age vs Hours Worked, colored by Income')
plt.show()
```



#### education, hours per week, income

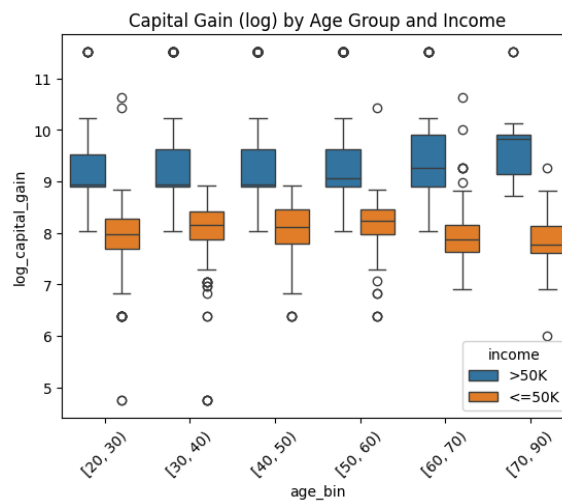
```
# representamos la relación entre las variables con un heatmap
heat = pd.pivot_table(
    df,
    values="hours_per_week",
    index="education",
    columns="income",
    aggfunc="mean")
sns.heatmap(heat, annot=True, fmt=".1f", cmap="coolwarm")
plt.title("Average Hours Worked by Education and Income")
plt.show()
```



## Age, income, capital gain

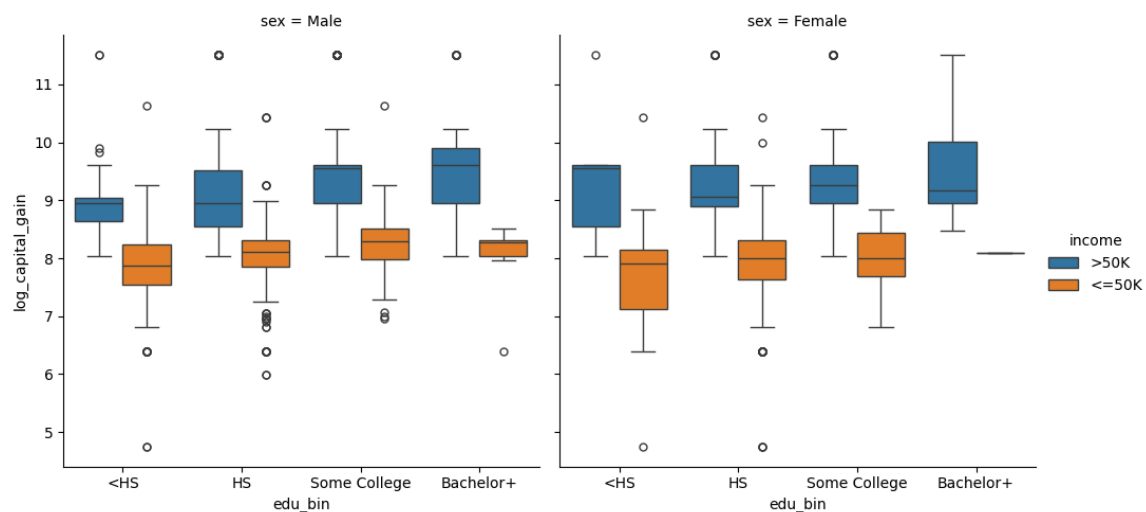
```
# creamos un boxplot por el análisis multivariante entre age, capital gain and income
# Create boxplot for multivariate analysis between age, capital gain and income
sns.boxplot(
    data=df_gain,
    x='age_bin',
    y='log_capital_gain',
    hue='income')

plt.xticks(rotation=45)
plt.title('Capital Gain (log) by Age Group and Income')
plt.show()
```



## Education, income, sex, capital gain

```
# representamos la relación entre las variables edu_bin, log capital gain, income and sex
sns.catplot(
    data=df_gain,
    x='edu_bin',
    y='log_capital_gain',
    hue='income',
    col='sex',
    kind='box')
plt.show()
```



## 7. Visualizaciones

Para el análisis exploratorio se utilizaron **histogramas**, **box plots**, **gráficos de barras** y **heatmaps**, seleccionados en función del tipo de variable y del objetivo analítico. Todas las visualizaciones fueron **correctamente etiquetadas**, con títulos, ejes y leyendas claras, y se acompañaron de una **interpretación explícita**, facilitando la identificación de patrones, comparaciones entre grupos y relaciones entre variables.

## 8. Conclusiones

Los resultados del análisis exploratorio confirman que el **nivel educativo** es el factor con mayor asociación al nivel de ingresos. Asimismo, entre los individuos que presentan **capital gain**, el **nivel de ingresos** se muestra como el principal determinante de mayores

ganancias. La **edad** y las **horas trabajadas** presentan una relación positiva, aunque de menor magnitud, actuando como **factores secundarios** en la explicación de los ingresos.

---

## 9. Recomendaciones

- Priorizar el **nivel educativo** como variable clave en análisis posteriores.
  - Considerar de forma explícita las **interacciones entre variables** demográficas, educativas y laborales.
  - Profundizar en el análisis del **capital gain**, especialmente en su relación con el nivel de ingresos y la educación.
- 

## 10. Cierre

Este documento consolida de forma estructurada el trabajo realizado durante el proceso de **análisis exploratorio de datos**, proporcionando una visión clara y fundamentada de los principales **factores asociados al nivel de ingresos**. Los resultados obtenidos sientan una base sólida para análisis posteriores y para la toma de decisiones informadas basadas en los datos.

---

## ANEXOS

### Anexo A – Diccionario de variables

El diccionario completo de variables del conjunto de datos, incluyendo su descripción, tipo y posibles categorías, se encuentra documentado en el notebook del proyecto. Dicho diccionario recoge variables de tipo demográfico, educativo, laboral y económico utilizadas a lo largo del análisis exploratorio.

---

### Anexo B – Estadísticos descriptivos

Los estadísticos descriptivos completos de las principales variables numéricas del dataset (media, mediana, desviación estándar, cuartiles y valores extremos) se calcularon durante el análisis exploratorio y se encuentran disponibles en el notebook del proyecto.

---

## **Anexo C – Visualizaciones adicionales**

Durante el proceso de análisis se generaron visualizaciones adicionales como apoyo al estudio exploratorio, incluyendo boxplots por subgrupos, distribuciones por ocupación y heatmaps de correlación. Estas visualizaciones no se incluyen en el cuerpo principal del documento para mantener la claridad expositiva y se encuentran implementadas en el notebook del proyecto.

---

## **Anexo D – Transformaciones de datos**

Las transformaciones aplicadas con fines exploratorios, tales como la transformación logarítmica de *capital gain*, la discretización de variables continuas y la recodificación de variables categóricas, se documentan detalladamente en el notebook del proyecto. Dichas transformaciones se utilizaron exclusivamente para facilitar la interpretación de los datos.

---

## **Anexo E – Código del análisis**

El análisis exploratorio se desarrolló en **Python**, utilizando las librerías *pandas*, *numpy*, *matplotlib* y *seaborn*. El código completo del análisis, junto con todas las visualizaciones y resultados intermedios, se encuentra disponible en el notebook del proyecto, garantizando la trazabilidad y reproducibilidad del trabajo realizado.