Actividad 1 Proyecto de Python

El dataset seleccionado contiene información de 1000 usuarios, con información sobre su edad, estado civil y más. También incluye si compraron una bicicleta o no.

Con el resumen estadístico de las variables numéricas podemos observar que la edad promedio es de aproximadamente 44 años, el ingreso medio es de $60k y la mayoría de las personas tienen entre 0 y 3 hijos y autos. Hay una dispersión considerable en los ingresos y la edad.

```
df.describe()
```

	ID	Income	Children	Cars	Age
count	1000.000000	994.000000	992.000000	991.000000	992.000000
mean	19965.992000	56267.605634	1.910282	1.455096	44.181452
std	5347.333948	31067.817462	1.626910	1.121755	11.362007
min	11000.000000	10000.000000	0.000000	0.000000	25.000000
25%	15290.750000	30000.000000	0.000000	1.000000	35.000000
50%	19744.000000	60000.000000	2.000000	1.000000	43.000000
75%	24470.750000	70000.000000	3.000000	2.000000	52.000000
max	29447.000000	170000.000000	5.000000	4.000000	89.000000

Contamos los valores nulos:

```
df.isnull().sum()
```

```
ID                   0
Marital Status       7
Gender              11
Income               6
Children             8
Education            0
Occupation           0
Home Owner           4
Cars                 9
Commute Distance     0
Region               0
Age                  8
Purchased Bike       0
dtype: int64
```

Imputamos los valores nulos:

```
:
# Imputar valores nulos en variables numéricas con la mediana
df['Income'].fillna(df['Income'].median(), inplace=True)
df['Children'].fillna(df['Children'].median(), inplace=True)
df['Cars'].fillna(df['Cars'].median(), inplace=True)
df['Age'].fillna(df['Age'].median(), inplace=True)

# Imputar valores nulos en variables categóricas con la moda
for col in ['Marital Status', 'Gender', 'Home Owner']:
    df[col].fillna(df[col].mode()[0], inplace=True)

# Verificar que ya no haya valores nulos
print("Valores nulos después de la limpieza:")
print(df.isnull().sum())
```

```
Valores nulos después de la limpieza:
ID                  0
Marital Status      0
Gender              0
Income              0
Children            0
Education           0
Occupation          0
Home Owner          0
Cars                0
Commute Distance    0
Region              0
Age                 0
Purchased Bike      0
dtype: int64
```

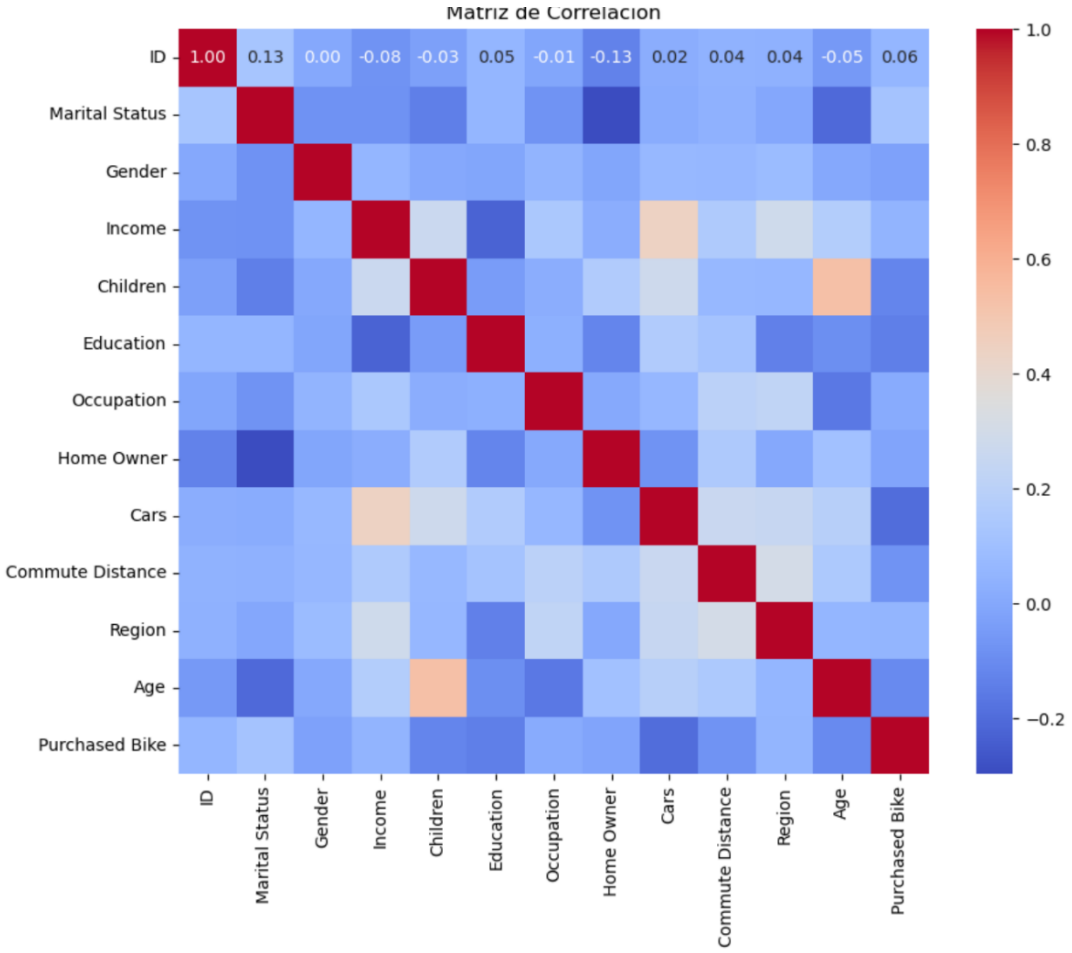
Transformamos las variables categóricas a numéricas:

```
5]:
#Convertir las variables categóricas a numericas
#Creamos una copia de nuestro dataframe
df_encoded = df.copy()

# Identificar variables categoricas
categorical_cols = df.select_dtypes(include=['object']).columns.tolist()

# Utilizamis LabelEncoder
le = LabelEncoder()
for col in categorical_cols:
    df[col] = df[col].astype(str)  # Asegurar que todos los valores sean string
    df[col] = le.fit_transform(df[col])
```

Visualizamos nuestra matriz de correlación, podemos observar que los factores más asociados con la compra de bicicletas son “Edad”, “Numero de autos”, “Distancia”.



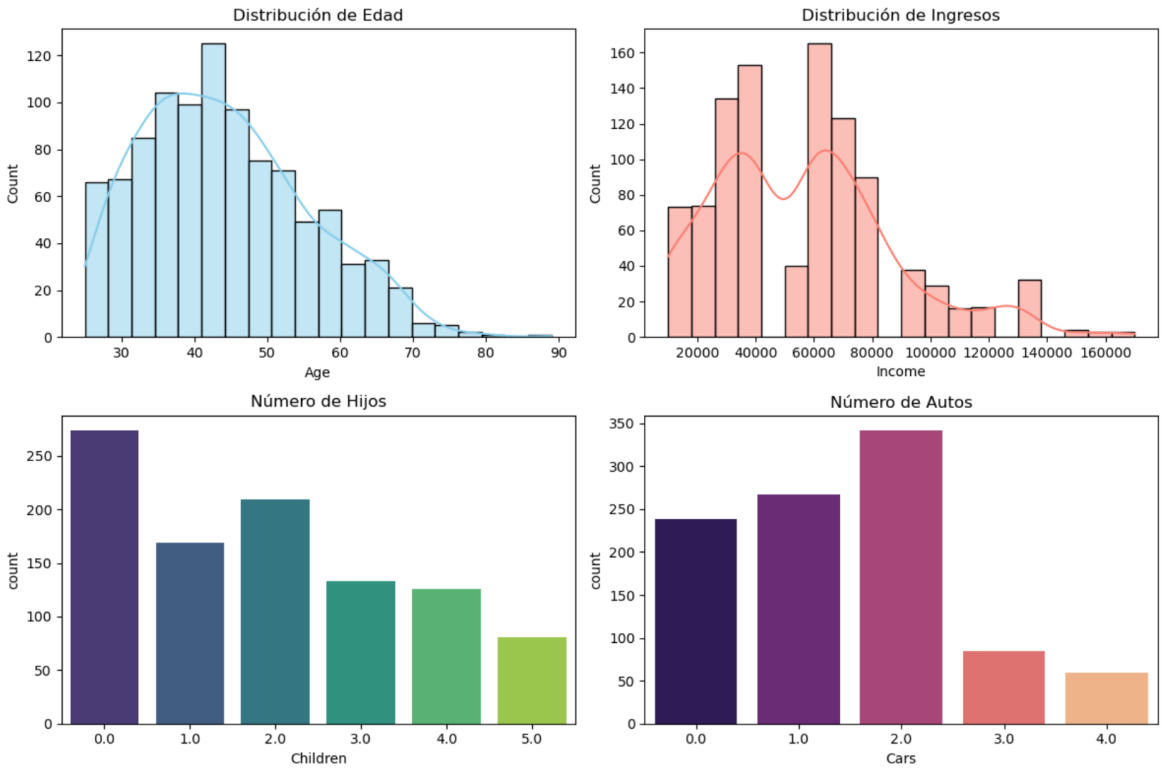
Con las siguientes graficas podemos observar lo siguiente:

1. El grupo demográfico más fuerte para ventas está entre 35 y 50 años.

2. Las bicicletas son más compradas por personas de ingresos medios.

3. Las personas con menos hijos tienden a comprar más.

4. La mayoría de los compradores tienen 2 autos.



Conclusión

Las bicicletas parecen ser más populares entre adultos de mediana edad, con ingresos medios, pocas responsabilidades familiares (0-2 hijos) y que poseen 1 o 2 autos. por lo cual la compra de bicicleta es para ejercicio, recreación y no para sustituir su transporte principal.

Para el siguiente paso se puede tomar en cuenta las variables más relevantes, realizar un mejor tratamiento con los datos nulos y crear un modelo predictivo para saber si un cliente comprará o no.