In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [33]:

```
# Read the CSV file
df_train = pd.read_csv("Train.csv")
```

In [35]:

```
# Display the first few rows of the dataframe
#print(df_train.info())
df_train['fecha_dato'] = pd.to_datetime(df_train['fecha_dato'])
df_train['fecha_alta'] = pd.to_datetime(df_train['fecha_alta'])

df_train['ind_empleado'] = df_train['ind_empleado'].astype('category')
df_train['sexo'] = df_train['sexo'].astype('category')

# Convert pais_residencia to category
df_train['pais_residencia'] = df_train['pais_residencia'].astype('categor

# Convert ind_nuevo to integer
df_train['ind_nuevo'] = df_train['ind_nuevo'].fillna(0).astype('int64')

# Convert ind_nuevo to category
df_train['ind_nuevo'] = df_train['ind_nuevo'].astype('category')

# Convert ult_fec_cli_1t to datetime
df_train['ult_fec_cli_1t'] = pd.to_datetime(df_train['ult_fec_cli_1t'], e

# Convert indrel to category
df_train['indrel'] = df_train['indrel'].astype('category')

# Convert canal_entrada to category
df_train['canal_entrada'] = df_train['canal_entrada'].astype('category')

# Convert tipodom to category
df_train['tipodom'] = df_train['tipodom'].astype('category')

# Step 1: Convert non-numeric values to NaN
df_train['antiguedad'] = pd.to_numeric(df_train['antiguedad'], errors='co

# Drop rows with NaN values in the antiguedad column and convert to int
df_train = df_train.dropna(subset=['antiguedad'])
df_train['antiguedad'] = df_train['antiguedad'].astype('int64')

# Verify the changes
print(df_train['antiguedad'].dtype)
print(df_train['antiguedad'].isnull().sum())
```

```
int64
0
```

In [36]:

```
# Convert to category
df_train['indrel_1mes'] = df_train['indrel_1mes'].astype('category')
df_train['tiprel_1mes'] = df_train['tiprel_1mes'].astype('category')
```

```
df_train['indext'] = df_train['indext'].astype('category')
df_train['conyuemp'] = df_train['conyuemp'].astype('category')
df_train['indfall'] = df_train['indfall'].astype('category')
df_train['cod_prov'] = df_train['cod_prov'].astype('category')
df_train['nomprov'] = df_train['nomprov'].astype('category')
df_train['ind_actividad_cliente'] = df_train['ind_actividad_cliente'].ast
df_train['segmento'] = df_train['segmento'].astype('category')

# Convert age to integer
df_train['age'] = df_train['age'].fillna(0).astype('int64')

# Summary Statistics to ensure data is merged correctly
print(df_train.info())

# Add 0 to the categories of 'conyuemp'
df_train['conyuemp'] = df_train['conyuemp'].astype('category')
df_train['conyuemp'] = df_train['conyuemp'].cat.add_categories([0])

# Substitute NaN values in 'conyuemp' column with 0
df_train['conyuemp'] = df_train['conyuemp'].fillna(0)

# Verify the changes
print(df_train['conyuemp'].isnull().sum())  # Should print 0 if all NaN v

# Check the number of rows and columns
num_rows = df_train.shape[0]
num_columns = df_train.shape[1]
print(f"Number of rows: {num_rows}")
print(f"Number of columns: {num_columns}")

# Remove rows where age is greater than 100
df_train = df_train[df_train['age'] <= 100]

# Add 'Missing' as a new category
df_train['canal_entrada'] = df_train['canal_entrada'].astype('category')
df_train['canal_entrada'] = df_train['canal_entrada'].cat.add_categories(
df_train['canal_entrada'] = df_train['canal_entrada'].fillna('No informat

# Remove the column 'ult_fec_cli_1t' in-place
df_train.drop(columns=['ult_fec_cli_1t'], inplace=True)

# Verify the changes
#print(df_train.columns)

# Drop rows with NaN values in the 'sexo' column
df_train = df_train.dropna(subset=['sexo'])

# Substitute NaN values in 'renta' column with 0 using .loc to avoid Sett
df_train.loc[:, 'renta'] = df_train['renta'].fillna(0)

#get the number of missing data points per column
missing_values_count = df_train.isnull().sum()
#look at the number of missing points in the 48 columns
missing_values_count[0:48]

# Drop rows with NaN values in the 'indrel_1mes' column
df_train = df_train.dropna(subset=['indrel_1mes'])
```

```python
df_train = df_train.dropna(subset=['tiprel_1mes'])

# Remove all rows with any null values
df_train = df_train.dropna()

# Verify the changes
print(df_train.isnull().sum())  # Should print 0 for all columns
# print(df_train.shape)  # Check the shape to see how many rows were drop
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 13619575 entries, 0 to 13647308
Data columns (total 48 columns):
 #   Column                 Dtype
---  ------                 -----
 0   fecha_dato             datetime64[ns]
 1   ncodpers               int64
 2   ind_empleado           category
 3   pais_residencia        category
 4   sexo                   category
 5   age                    int64
 6   fecha_alta             datetime64[ns]
 7   ind_nuevo              category
 8   antiguedad             int64
 9   indrel                 category
 10  ult_fec_cli_1t         datetime64[ns]
 11  indrel_1mes            category
 12  tiprel_1mes            category
 13  indresi                object
 14  indext                 category
 15  conyuemp               category
 16  canal_entrada          category
 17  indfall                category
 18  tipodom                category
 19  cod_prov               category
 20  nomprov                category
 21  ind_actividad_cliente  category
 22  renta                  float64
 23  segmento               category
 24  ind_ahor_fin_ult1      int64
 25  ind_aval_fin_ult1      int64
 26  ind_cco_fin_ult1       int64
 27  ind_cder_fin_ult1      int64
 28  ind_cno_fin_ult1       int64
 29  ind_ctju_fin_ult1      int64
 30  ind_ctma_fin_ult1      int64
 31  ind_ctop_fin_ult1      int64
 32  ind_ctpp_fin_ult1      int64
 33  ind_deco_fin_ult1      int64
 34  ind_deme_fin_ult1      int64
 35  ind_dela_fin_ult1      int64
 36  ind_ecue_fin_ult1      int64
 37  ind_fond_fin_ult1      int64
 38  ind_hip_fin_ult1       int64
 39  ind_plan_fin_ult1      int64
 40  ind_pres_fin_ult1      int64
 41  ind_reca_fin_ult1      int64
 42  ind_tjcr_fin_ult1      int64
 43  ind_valo_fin_ult1      int64
 44  ind_viv_fin_ult1       int64
 45  ind_nomina_ult1        float64
 46  ind_nom_pens_ult1      float64
 47  ind_recibo_ult1        int64
dtypes: category(16), datetime64[ns](3), float64(3), int64(25), object(1)
memory usage: 3.6+ GB
None
0
Number of rows: 13619575
Number of columns: 48
fecha_dato               0
ncodpers                 0
ind_empleado             0
pais_residencia          0
sexo                     0
age                      0
fecha_alta               0
ind_nuevo                0
antiguedad               0
indrel                   0
indrel_1mes              0
tiprel_1mes              0
indresi                  0
indext                   0
conyuemp                 0
canal_entrada            0
indfall                  0
tipodom                  0
cod_prov                 0
nomprov                  0
ind_actividad_cliente    0
renta                    0
segmento                 0
ind_ahor_fin_ult1        0
ind_aval_fin_ult1        0
ind_cco_fin_ult1         0
ind_cder_fin_ult1        0
ind_cno_fin_ult1         0
ind_ctju_fin_ult1        0
ind_ctma_fin_ult1        0
ind_ctop_fin_ult1        0
ind_ctpp_fin_ult1        0
ind_deco_fin_ult1        0
ind_deme_fin_ult1        0
ind_dela_fin_ult1        0
ind_ecue_fin_ult1        0
ind_fond_fin_ult1        0
ind_hip_fin_ult1         0
ind_plan_fin_ult1        0
ind_pres_fin_ult1        0
ind_reca_fin_ult1        0
ind_tjcr_fin_ult1        0
ind_valo_fin_ult1        0
ind_viv_fin_ult1         0
ind_nomina_ult1          0
ind_nom_pens_ult1        0
```

```
ind_recibo_ult1          0
dtype: int64
```

In [37]:

```
# Function to calculate the number of months between two dates
def calculate_month_diff(start_date, end_date):
    return (end_date.year - start_date.year) * 12 + end_date.month - star

# Apply the function to create the new column
df_train['months_between'] = df_train.apply(lambda row: calculate_month_d
```

In [38]:

```
# Handle negative values by setting them to zero
df_train['antiguedad'] = df_train['antiguedad'].apply(lambda x: 0 if x <

# Display the updated DataFrame
print(df_train.head())
```

```
  fecha_dato  ncodpers ind_empleado pais_residencia sexo  age fecha_alta
\
0 2015-01-28   1375586            N              ES    H   35 2015-01-12
1 2015-01-28   1050611            N              ES    V   23 2012-08-10
2 2015-01-28   1050612            N              ES    V   23 2012-08-10
3 2015-01-28   1050613            N              ES    H   22 2012-08-10
4 2015-01-28   1050614            N              ES    V   23 2012-08-10

  ind_nuevo  antiguedad indrel  ... ind_plan_fin_ult1 ind_pres_fin_ult1  \
0         0           6    1.0  ...                 0                 0
1         0          35    1.0  ...                 0                 0
2         0          35    1.0  ...                 0                 0
3         0          35    1.0  ...                 0                 0
4         0          35    1.0  ...                 0                 0

  ind_reca_fin_ult1 ind_tjcr_fin_ult1 ind_valo_fin_ult1 ind_viv_fin_ult1
\
0                 0                 0                 0                0
1                 0                 0                 0                0
2                 0                 0                 0                0
3                 0                 0                 0                0
4                 0                 0                 0                0

  ind_nomina_ult1 ind_nom_pens_ult1 ind_recibo_ult1 months_between
0             0.0               0.0               0              0
1             0.0               0.0               0             29
2             0.0               0.0               0             29
3             0.0               0.0               0             29
4             0.0               0.0               0             29

[5 rows x 48 columns]
```

In [40]:

```
# Map 'H' to 'Men' and 'V' to 'Women'
df_train['sexo'] = df_train['sexo'].map({'H': 'Men', 'V': 'Women'})
```

In [46]:

```
# Display the updated DataFrame
print(df_train)
```

```
         fecha_dato  ncodpers ind_empleado pais_residencia   sexo   age \
0        2015-01-28   1375586            N              ES    Men   35
1        2015-01-28   1050611            N              ES  Women   23
```

```
2        2015-01-28   1050612            N              ES  Women   23
3        2015-01-28   1050613            N              ES    Men   22
4        2015-01-28   1050614            N              ES  Women   23
...             ...       ...          ...             ...    ...  ...
13647303 2016-05-28   1166766            N              ES  Women   25
13647304 2016-05-28   1166765            N              ES  Women   22
13647305 2016-05-28   1166764            N              ES  Women   23
13647306 2016-05-28   1166763            N              ES    Men   47
13647307 2016-05-28   1166789            N              ES    Men   22

         fecha_alta ind_nuevo  antiguedad indrel  ... ind_plan_fin_ult1  \
0        2015-01-12         0           6    1.0  ...                 0   
1        2012-08-10         0          35    1.0  ...                 0   
2        2012-08-10         0          35    1.0  ...                 0   
3        2012-08-10         0          35    1.0  ...                 0   
4        2012-08-10         0          35    1.0  ...                 0   
...             ...       ...         ...    ...  ...               ...   
13647303 2013-08-14         0          33    1.0  ...                 0   
13647304 2013-08-14         0          33    1.0  ...                 0   
13647305 2013-08-14         0          33    1.0  ...                 0   
13647306 2013-08-14         0          33    1.0  ...                 0   
13647307 2013-08-14         0          33    1.0  ...                 0   

         ind_pres_fin_ult1 ind_reca_fin_ult1 ind_tjcr_fin_ult1  \
0                        0                 0                 0   
1                        0                 0                 0   
2                        0                 0                 0   
3                        0                 0                 0   
4                        0                 0                 0   
...                    ...               ...               ...   
13647303                 0                 0                 0   
13647304                 0                 0                 0   
13647305                 0                 0                 0   
13647306                 0                 0                 0   
13647307                 0                 0                 0   

         ind_valo_fin_ult1 ind_viv_fin_ult1 ind_nomina_ult1 ind_nom_pens_u
lt1  \
0                        0                0             0.0               
0.0   
1                        0                0             0.0               
0.0   
2                        0                0             0.0               
0.0   
3                        0                0             0.0               
0.0   
4                        0                0             0.0               
0.0   
...                    ...              ...             ...               
...   
13647303                 0                0             0.0               
0.0   
13647304                 0                0             0.0               
0.0   
13647305                 0                0             0.0               
0.0   
13647306                 0                0             0.0               
```

```
0.0
13647307                 0                0             0.0
0.0

         ind_recibo_ult1 months_between
0                      0              0
1                      0             29
2                      0             29
3                      0             29
4                      0             29
...                  ...            ...
13647303               0             33
13647304               0             33
13647305               0             33
13647306               0             33
13647307               0             33

[13379656 rows x 48 columns]
```

In [48]:

```
# Handle negative values in 'antiguedad' by setting them to zero
df_train['antiguedad'] = df_train['antiguedad'].apply(lambda x: 0 if x <

# Ensure 'ind_nuevo' contains only 1 or 0
df_train['ind_nuevo'] = pd.to_numeric(df_train['ind_nuevo'], errors='coer
df_train['ind_nuevo'] = df_train['ind_nuevo'].fillna(0)  # Replace NaN wi
df_train['ind_nuevo'] = df_train['ind_nuevo'].apply(lambda x: 1 if x == 1
```

In [61]:

```
# # Display the updated DataFrame
# print(df_train)
```

In [52]:

```
# Ensure 'indrel_1mes' contains only 1, 2, 3, 4, or 'P'
valid_values = {'1', '2', '3', '4', 'P'}
df_train['indrel_1mes'] = df_train['indrel_1mes'].astype(str)  # Convert
df_train['indrel_1mes'] = df_train['indrel_1mes'].apply(lambda x: x if x
```

In [55]:

```
# Display the updated DataFrame
print(df_train)
```

```
         fecha_dato  ncodpers ind_empleado pais_residencia   sexo  age  \
0        2015-01-28   1375586            N              ES    Men   35
1        2015-01-28   1050611            N              ES  Women   23
2        2015-01-28   1050612            N              ES  Women   23
3        2015-01-28   1050613            N              ES    Men   22
4        2015-01-28   1050614            N              ES  Women   23
...             ...       ...          ...             ...    ...  ...
13647303 2016-05-28   1166766            N              ES  Women   25
13647304 2016-05-28   1166765            N              ES  Women   22
13647305 2016-05-28   1166764            N              ES  Women   23
13647306 2016-05-28   1166763            N              ES    Men   47
13647307 2016-05-28   1166789            N              ES    Men   22

         fecha_alta  ind_nuevo  antiguedad indrel  ... ind_plan_fin_ult1
\
0        2015-01-12          0           6    1.0  ...                 0
1        2012-08-10          0          35    1.0  ...                 0
2        2012-08-10          0          35    1.0  ...                 0
3        2012-08-10          0          35    1.0  ...                 0
```

```
4        2012-08-10          0          35    1.0  ...                 0
...             ...        ...         ...    ...  ...               ...
13647303 2013-08-14          0          33    1.0  ...                 0
13647304 2013-08-14          0          33    1.0  ...                 0
13647305 2013-08-14          0          33    1.0  ...                 0
13647306 2013-08-14          0          33    1.0  ...                 0
13647307 2013-08-14          0          33    1.0  ...                 0

         ind_pres_fin_ult1 ind_reca_fin_ult1 ind_tjcr_fin_ult1  \
0                        0                 0                 0
1                        0                 0                 0
2                        0                 0                 0
3                        0                 0                 0
4                        0                 0                 0
...                    ...               ...               ...
13647303                 0                 0                 0
13647304                 0                 0                 0
13647305                 0                 0                 0
13647306                 0                 0                 0
13647307                 0                 0                 0

         ind_valo_fin_ult1 ind_viv_fin_ult1 ind_nomina_ult1 ind_nom_pens_u
lt1  \
0                        0                0             0.0
0.0
1                        0                0             0.0
0.0
2                        0                0             0.0
0.0
3                        0                0             0.0
0.0
4                        0                0             0.0
0.0
...                    ...              ...             ...
...
13647303                 0                0             0.0
0.0
13647304                 0                0             0.0
0.0
13647305                 0                0             0.0
0.0
13647306                 0                0             0.0
0.0
13647307                 0                0             0.0
0.0

         ind_recibo_ult1 months_between
0                      0              0
1                      0             29
2                      0             29
3                      0             29
4                      0             29
...                  ...            ...
13647303               0             33
13647304               0             33
13647305               0             33
13647306               0             33
```

```
13647307                0             33

[13379656 rows x 48 columns]
```

In [59]:

```
# # Print the first 5 and last 5 values of 'indrel_1mes' column
# first_5_values = df_train['indrel_1mes'].head(5)
# last_5_values = df_train['indrel_1mes'].tail(5)

# print("First 5 values of 'indrel_1mes':")
# print(first_5_values)

# print("\nLast 5 values of 'indrel_1mes':")
# print(last_5_values)
```

In [70]:

```
# Specify the filename for the new CSV file
filename = 'df_train_new.csv'

# Save the DataFrame to a new CSV file
df_train.to_csv(filename, index=False)

print(f'DataFrame saved to {filename}')
```

```
DataFrame saved to df_train_new.csv
```

In [ ]:

```
```