Project Management Credit Customers

En este proyecto analizamos las características de los clientes y su relación con la aprobación de créditos, con la posibilidad de construir un modelo predictivo para predecir si un cliente será o no aprobado para un crédito.

Objetivo

Introducción y antecedentes del conjunto de datos: El conjunto de datos es compuesto de 1000 filas y 21 columnas. Son 13 columnas categóricas, 7 columnas numéricas y una columna objetivo 'Class'." el objetivo es ayudar al banco a comprender las características de los clientes que influyen si un cliente será o no aprobado para un crédito.

Limpieza y Exploración de datos

Se realizó un proceso de limpieza de datos para facilitar el análisis.

El objetivo está desbalanceado y se convirtió a numérico en el preprocesamiento.

Algunas columnas categóricas tuvieron valores inconsistentes, mientras que otras tuvieron demasiadas categorías y se agruparon en menos categorías.

Limpieza y Exploración de datos

Las columnas 'Checking_status', 'employment', 'other_parties', 'other_payment_plans', 'own_telephone' y 'foreign_worker' se eliminaron porque no eran relevantes para el modelo. Los valores de las columnas que se eliminaron eran inconsistentes y en algunos casos con casi todos los valores iguales.

```
# eliminar la columna 'checking_status'
df_1.drop('checking_status', axis=1, inplace=True)
# eliminar la columna 'employment'
df_1.drop('employment', axis=1, inplace=True)
# eliminar la columna 'other_parties'
df_1.drop('other_parties', axis=1, inplace=True)
# eliminar la columna 'other_payment_plans'
df_1.drop('other_payment_plans', axis=1, inplace=True)
# eliminar la columna 'own_telephone'
df_1.drop('own_telephone', axis=1, inplace=True)
# eliminar la columna 'foreign_worker'
df_1.drop('foreign_worker', axis=1, inplace=True)
```

Limpieza y Exploración de datos

Las columnas:

'credit_history': Fueron reemplazados valores inconsistentes: 'no credits/all paid' por 'no credits'

'purpose': Fueron agrupadas categorías con menos de 90 entradas en la categoría 'other' para disminuir el número de categorías

'savings_status': Fueron agrupadas valores poco relevantes en '> 100' y se mantuvieron: '< 100' y 'no known savings' para disminuir el número de categorías.

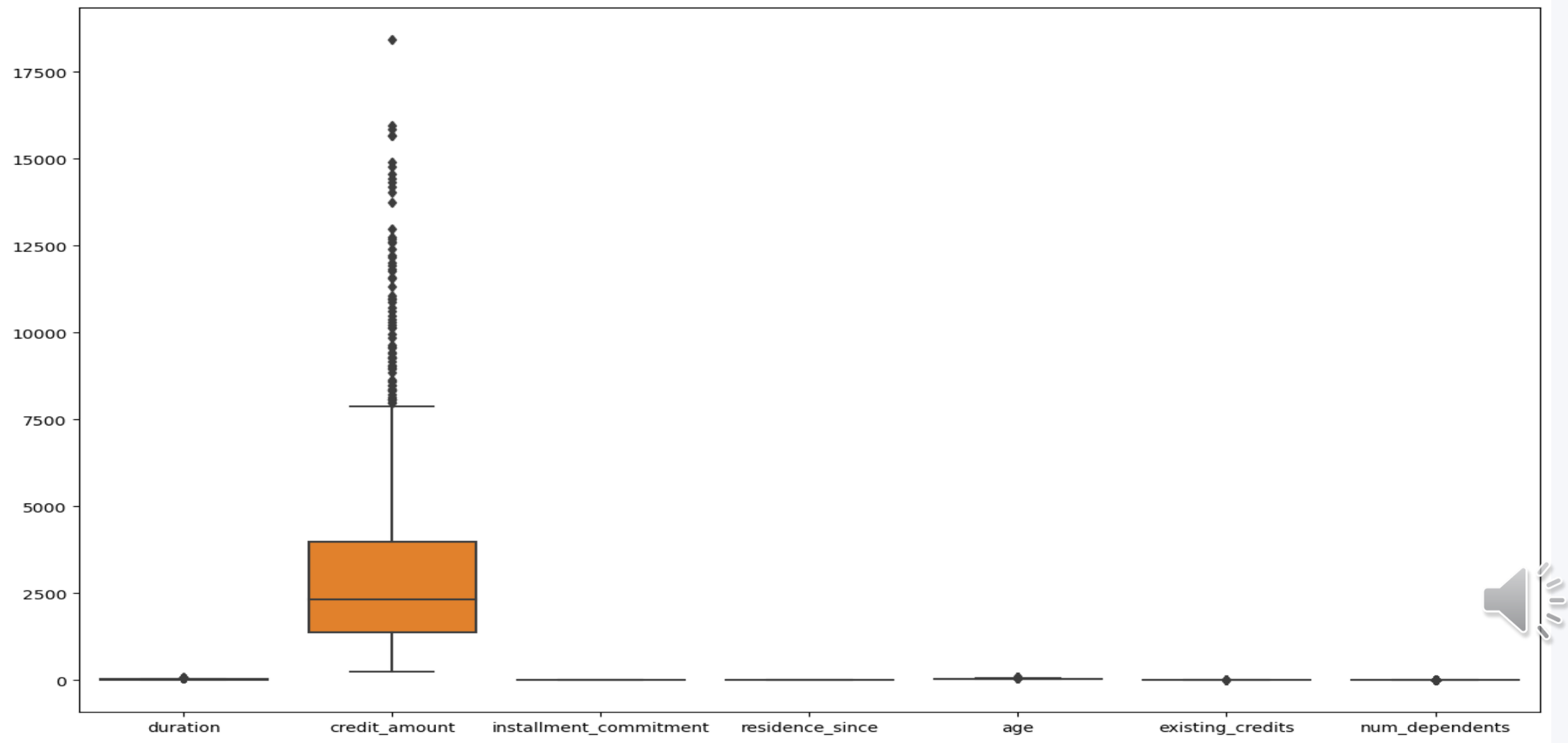
'personal_status' se renombró a 'gender' y y se convertió en binaria.

'job' se convirtió en binaria: skilled y unskilled.

```
# Reemplazar valores inconsistentes en la columna 'credit_history': 'no credits/all paid' por 'no credits'
df_1['credit_history'] = df_1['credit_history'].replace('no credits/all paid', 'no credits')
df_1['credit_history'] = df_1['credit_history'].replace('all paid', 'no credits')
# Reemplazar valores con mwnos de 90 en la columna 'purpose' por 'other'
df_1['purpose'] = df_1['purpose'].replace('education', 'other')
df_1['purpose'] = df_1['purpose'].replace('repairs', 'other')
df_1['purpose'] = df_1['purpose'].replace('domestic appliance', 'other')
df_1['purpose'] = df_1['purpose'].replace('retraining', 'other')
# Reemplazar valores por > 100 y mantener < 100 y 'no known savings'
df_1['savings_status'] = df_1['savings_status'].replace('100<=X<500', '>100')
df_1['savings_status'] = df_1['savings_status'].replace('500<=X<1000', '>100')
df_1['savings_status'] = df_1['savings_status'].replace('>=1000', '>100')
# Convertir la columna personal_status en una columna binaria
df_1['personal_status']=df_1['personal_status'].replace('male mar/wid', 'male')
df_1['personal_status']=df_1['personal_status'].replace('male div/sep', 'male')
df_1['personal_status']=df_1['personal_status'].replace('male single', 'male')
df_1['personal_status']=df_1['personal_status'].replace('female div/dep/mar', 'female')
# Cambiar el nombre de la columna 'personal_status' a gender
df_1 = df_1.rename(columns={'personal_status': 'gender'})
# convertir la columna 'job' en una columna binaria skilled y unskilled
df_1['job']=df_1['job'].replace('unemp/unskilled non res', 'unskilled')
df_1['job']=df_1['job'].replace('unskilled resident', 'unskilled')
df_1['job']=df_1['job'].replace('high qualif/self emp/mgmt', 'skilled')
```

Limpieza y Exploración de datos

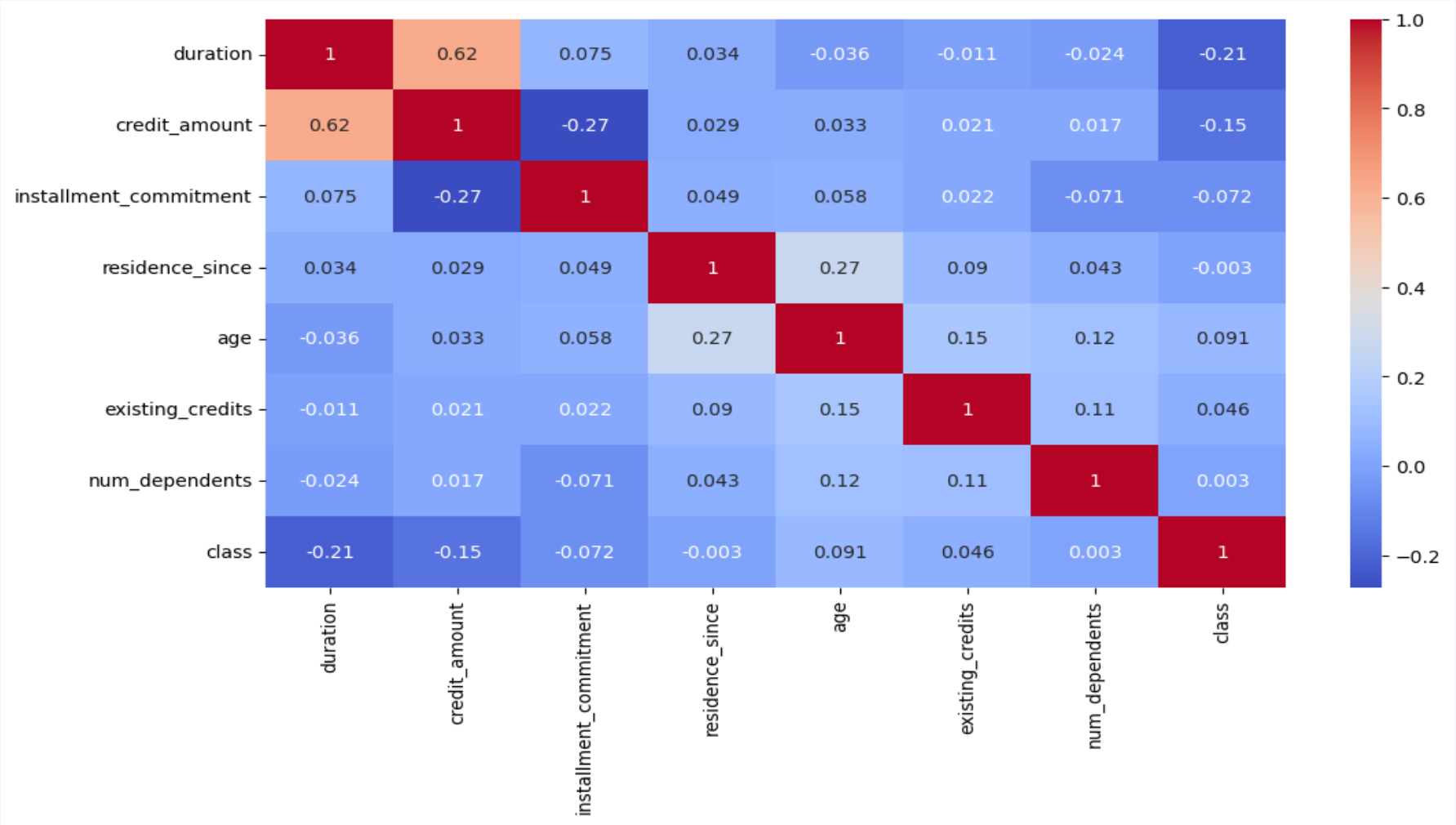
Outliers: La única columna numérica con valores atípicos es credit_amount. Decidí mantenerlos ya que pueden ser valores reales, no encontré razón para reemplazarlos.



Visualización de los datos

Matriz de Correlación

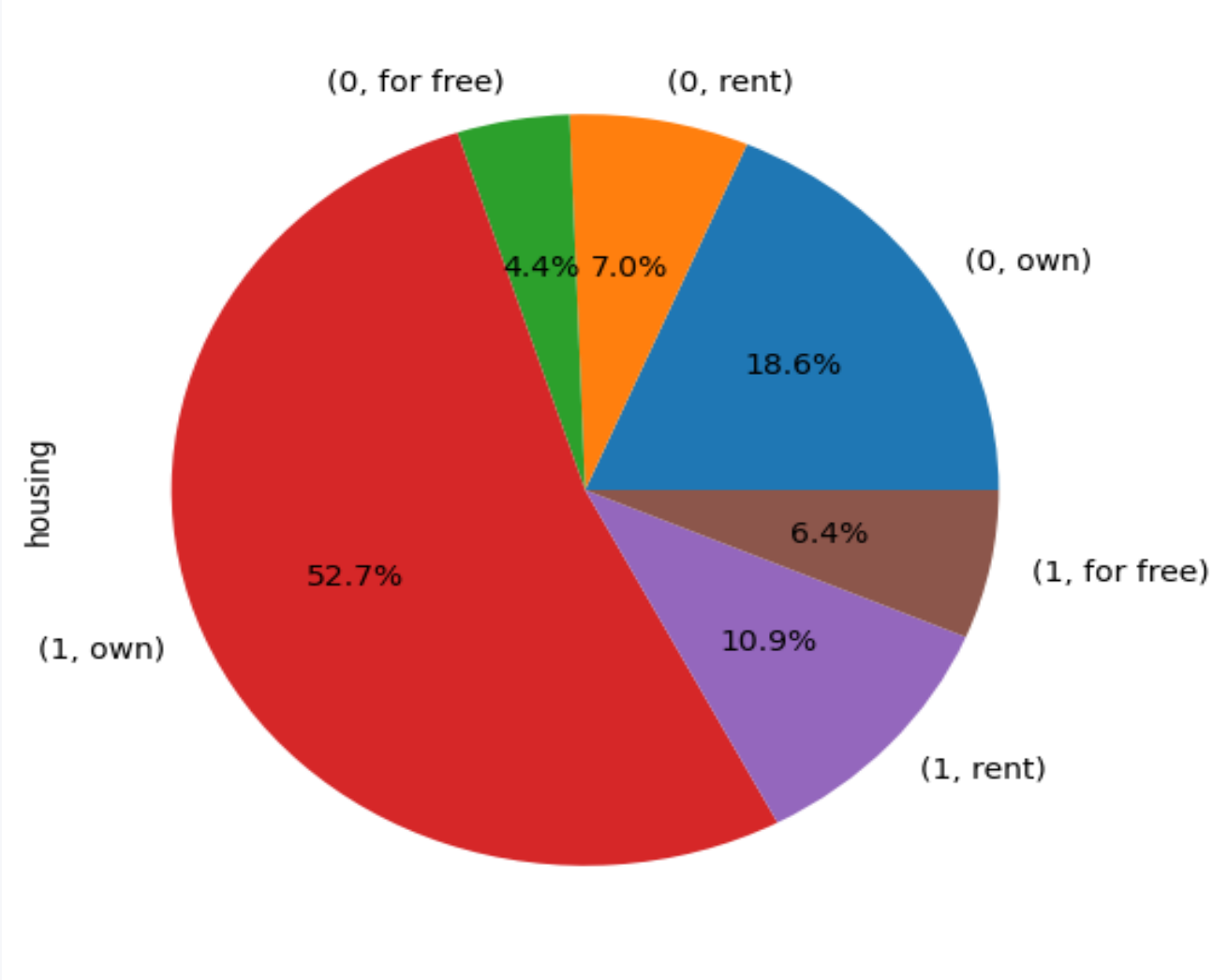
La matriz de correlación muestra que la variable que tiene la mayor correlación con la variable objetivo es 'credit_amount'. El resto de las variables numéricas no tienen una correlación significativa con el objetivo.



Visualización de los datos

Grafico de pie de la variable Casa Propia por clase

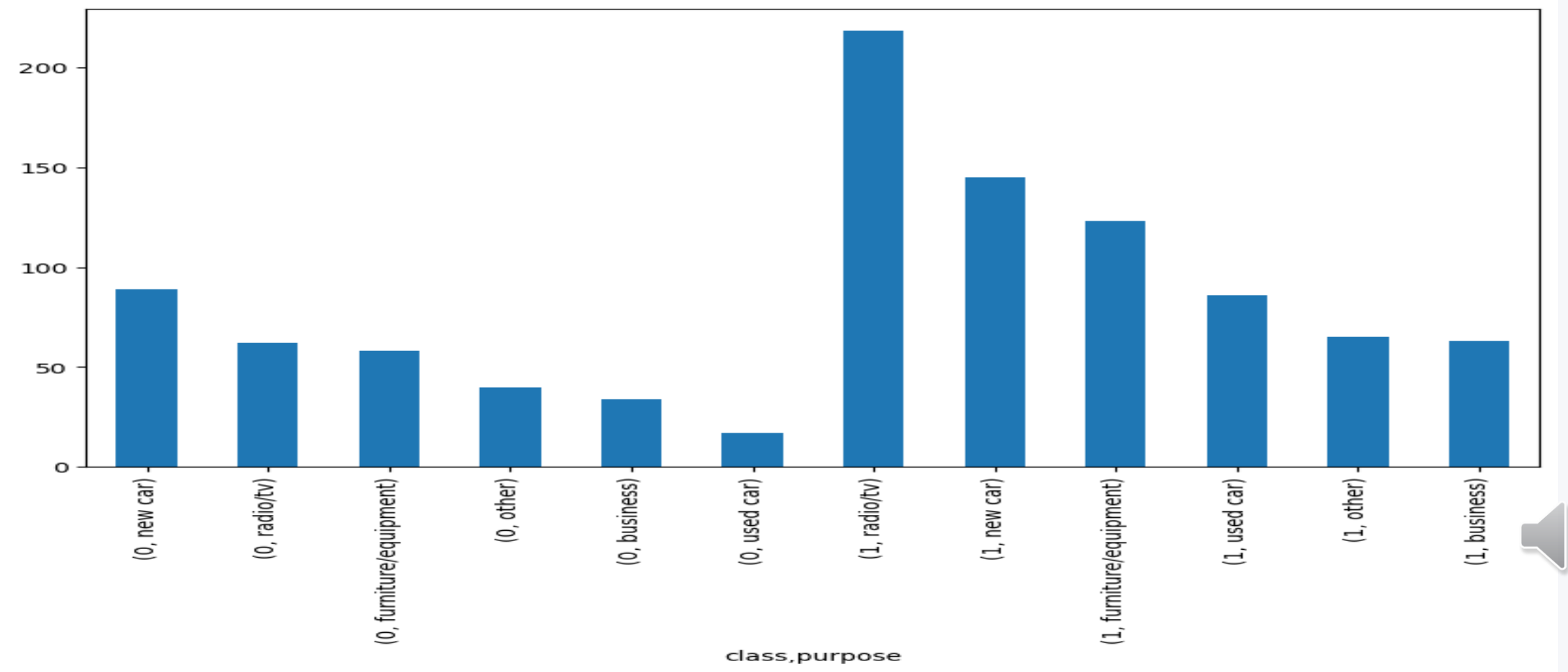
En el gráfico de pastel podemos ver que la mayoría de los clientes buenos o malos tienen casa propia. Esto indica que aquellos que tienen casa propia son más propensos a tomar créditos.



Visualización de los datos

Grafico de barras de la variable propósito por clase

En el gráfico de barras podemos ver que la mayoría de los créditos aprobados son para comprar televisores, seguidos por créditos para comprar autos nuevos. Los créditos más rechazados por el banco son para comprar autos nuevos.



Visualización de los datos

Grafico pairplot, relación entre todas las variables

En el gráfico pairplot podemos ver que la variable 'credit_amount' tiene una distribución asimétrica positiva, lo que indica que la mayoría de los clientes toman créditos pequeños. También hay más clientes buenos que malos en el data set.

Conclusion

En conclusión, este conjunto de datos ofrece una buena oportunidad para analizar las características de los clientes y su relación con la aprobación de créditos, con la posibilidad de construir un modelo predictivo para predecir si un cliente será o no aprobado para un crédito.

Además, es importante destacar que la limpieza de datos y la conversión de variables categóricas en numéricas o binarias permitió un análisis más preciso y una mejor interpretación de los resultados. También es importante mencionar que el objetivo desbalanceado puede requerir técnicas de muestreo estratificado o de ponderación para obtener resultados más precisos en el modelo.