PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS SOCIALES

Título: Informe Lending - Entregable 1

Nombres y códigos:

Chipana Cangana, Leydi - 20202209

Chachayma Ynchicsana, Eliane - 20201113

Ttica Huanca, Maria Belen - 20206668

2026

ÍNDICE

INFORME

1. Análisis estadístico inicial (descriptivo)

El análisis estadístico inicial se realizó diferenciando entre variables numéricas y categóricas, con el objetivo de evaluar la estructura, distribución y calidad de la información disponible. En el caso de las variables numéricas vinculadas al perfil crediticio y a las características del préstamo, la proporción de valores faltantes es baja y no supera el 5% del total de observaciones. No obstante, se identifican algunas limitaciones relevantes, como la baja variabilidad en el número de moras recientes *(delinq_2yrs*) y en los registros de bancarrota *(pub_rec_bankruptcies)*, así como la limitada diversidad del plazo del préstamo *(term*), que solo adopta dos valores posibles (36 y 60 meses).

Desde el punto de vista distributivo, el ingreso anual *(annual_inc*) y la utilización del crédito revolvente (*revol_util*) presentan asimetría positiva y valores extremos elevados, mientras que el resto de las variables numéricas analizadas muestran rangos más acotados y consistentes con su naturaleza económica. El análisis de calidad confirma que las variables numéricas están correctamente definidas en términos de tipo de dato y que los valores faltantes en variables clave no superan el 2%, sin detectarse problemas de duplicidad.

En cuanto a las variables categóricas, el análisis evidencia aspectos más relevantes para el tratamiento posterior de los datos. Se identifican variables con variabilidad prácticamente nula, como plan de pago (*pymnt_plan*) y tipo de solicitud (*application_type*), cuyo aporte informativo es limitado. En contraste, el cargo laboral *(emp_title*) presenta una cardinalidad excesivamente alta, con cerca de 15,000 categorías distintas y una proporción considerable de valores faltantes, lo que dificulta su inclusión directa en modelos predictivos.

Adicionalmente, se observan categorías de baja frecuencia (menores al 1%) en variables como la ocupación (*emp_title*), la calificación crediticia (grade), el propósito del préstamo (*purpose*) y la ubicación geográfica (*addr_state)*. En particular, la ocupación concentra una elevada cantidad de categorías poco representadas, lo que sugiere la necesidad de aplicar técnicas de agrupación o recodificación en etapas posteriores del análisis para mejorar la estabilidad y capacidad explicativa de los resultados.

Tabla 1

Categorías	Variables
Variable Objetivo (Target)	‘loan_status’
Variables numéricas	'loan_amnt', 'term', 'int_rate', 'installment', 'annual_inc', 'dti', 'delinq_2yrs', 'revol_util', 'pub_rec_bankruptcies'
Variables categóricas	'grade', 'emp_title', 'emp_length', 'home_ownership', 'verification_status', 'pymnt_plan', 'purpose', 'addr_state', 'application_type'

2. Preprocesamiento

En cuanto al preprocesamiento de variables numéricas, el enfoque se centró en garantizar la estabilidad y normalidad de los datos. Primero, se aplicó una estrategia de imputación basada en la mediana para cubrir cualquier valor faltante, técnica que resulta más robusta frente a valores extremos que el promedio. Posteriormente, se utilizó una transformación logarítmica mediante ‘*FunctionTransformer’* en variables críticas como el monto del préstamo e ingreso anual. El resultado fue una mejora en el sesgo (skewness) del ingreso anual, pasando de 35.91 a 0.14, lo que acerca la distribución a una forma más acampanada y casi simétrica, facilitando la convergencia y precisión de los algoritmos de aprendizaje.

Respecto al tratamiento de variables categóricas, se realizó una limpieza mediante la eliminación de espacios en blanco accidentales y se transformaron caracteres como signos de interrogación ("?") en valores nulos reales. Finalmente, para evitar la pérdida de registros, se empleó un ‘*SimpleImputer’* con una estrategia constante, etiquetando los vacíos como "*Missing*".

Tabla Resumen

Análisis	**Resultados**
Imputación de valores con la mediana	Porcentaje de valores imputados: ● annual_inc: 0.02% ● dti: 0.06% ● delinq_2yrs: 0.05% ● revol_util: 0.14% ● pub_rec_bankruptcies: 1.61%
Transformación logarítmica log(1+x)	Reducción de sesgo en porcentaje: ● loan_amnt: 53.97% ● annual_inc: 99.61%
Unificación de categorías	Diferencia de categorías antes y después ● emp_title: -113 (Reducción) ● emp_length: +1 ● home_ownership: +1

Análisis	Resultados
	● verification_status: +1 ● purpose: +1 ● addr_state: +1
Limpieza de missings codificados	No se encontraron missings codificados “?” en todas las variables
Imputación de missings	No se imputaron missings

3. Análisis Exploratorio (EDA univariado) y Visualización

Basado en el análisis univariado de variables numéricas, se identificaron algunos puntos críticos. En primer lugar, las variables *annual_inc* y *revol_util* continúan con valores extremos con una disparidad masiva mostrada en el boxplot e histograma (véase Anexo). Las variables *pymnt_pla*n y *application_type* no varían. Por otro lado, observando las distribuciones de *delinq_2yrs* y *pub_rec_bankruptcies* , la gran mayoría de los clientes (más del 75%) tiene un valor de 0.0. Es decir, estás dos variables ocurren, pero en muy pocos casos.

En cuánto a las variables categóricas, *emp_title* presentan 15,047 valores únicos, lo que podría general problemas de cardinalidad. En el caso de *emp_length*, presenta 502 valores nulos. Al ser una variable ordinal (de "*< 1 year*" a "*10+ years*"), estos nulos podrían representar a personas desempleadas o que omitieron el dato.

En cuanto a la variable target, el gráfico de barras demuestra que la clase "*Fully Paid*" supera por mucho a "*Charged Off*". En otras palabras, se presenta el problema de clases desbalanceadas, lo cual debe ser tomado en cuenta más adelante.

De tal manera, se realizó un segundo procesamiento, en el que el objetivo principal de estas acciones fue reducir el "ruido" del dataset y preparar las estructuras del dataset.

Tabla resumen de preprocesamiento post análisis univariado

Acción	Variables	Resultados
Feature Selection	pymnt_plan, application_type, emp_title, revol_util, delinq_2yrs, pub_rec_bankruptcies	Se eliminaron seis variables que presentaban problemas críticos detectados en el análisis exploratorio:

Acción	Variables	Resultados
Categorización	term	se convirtió a tipo categoría binaria, ya que solo existen dos plazos posibles: 36 y 60 meses.
Limpieza de Target	loan_status	Reducción de valores nulos de 3 a 0.

4. EDA bivariado: variables explicativas vs target

El análisis exploratorio bivariado entre las variables numéricas y la variable objetivo (*loan_status*) permite evaluar la relación entre las características financieras de los solicitantes y la probabilidad de que un préstamo sea completamente pagado. A partir de los boxplot por clase, se identificó los principales determinantes del cumplimiento (*Fully Paid*) frente al incumplimiento *(Charged Off*). La tasa de pago global se sitúa en 0.852.

Al analizar los gráficos de segmentación por quintiles, existe una relación lineal negativa entre la tasa de interés (*int_rate*) y la probabilidad de pago, reflejando un mayor riesgo crediticio en préstamos con condiciones más onerosas. De manera contraria, la variable *annual_inc* presenta una tendencia ascendente (a medida que subimos de quintil de ingresos, la tasa de éxito de pago aumenta). Por último, las variables *dti*, *loan_amnt* e *installment* muestran una fluctuación mínima.

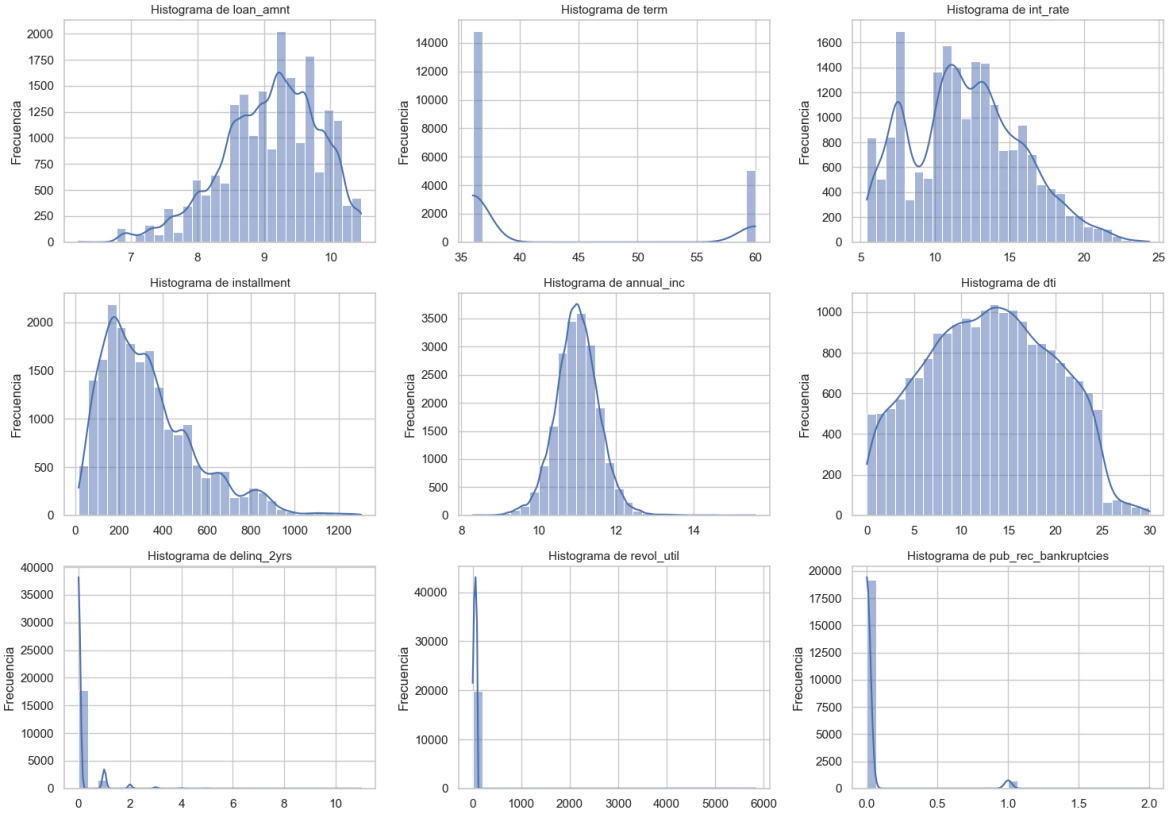
En cuanto a las variables categóricas, se puede resaltar la señal predictiva por grupo de *term*. Los préstamos a 36 meses tienen una tasa de pago muy superior al promedio global, mientras que los de 60 meses caen drásticamente por debajo de la línea de referencia. Otros indicadores son grade (la tasa de pago desciende escalonadamente desde las categorías A hasta G) y purpose (curiosamente aunque la mayoría pide para *debt_consolidation*, su tasa de pago se mantiene cercana al promedio)

Asimismo, observamos que algunas variables contienen pocos registros en sus categorías. Por ejemplo, la variable *grade* contiene 159 y 540 registros en las categoría “G2 y “F” , respectivamente. También, *purpose* o propósito de préstamo contiene categorías casi invisibles como "*renewable_energy*", "*educational*", "*house*" y "*vacation*". En este caso, se podría agrupar en una categoría como “other”, con el fin de reducir la dispersión, evitar sesgos y mejorar la estabilidad del modelo predictivo.

5. ANEXO

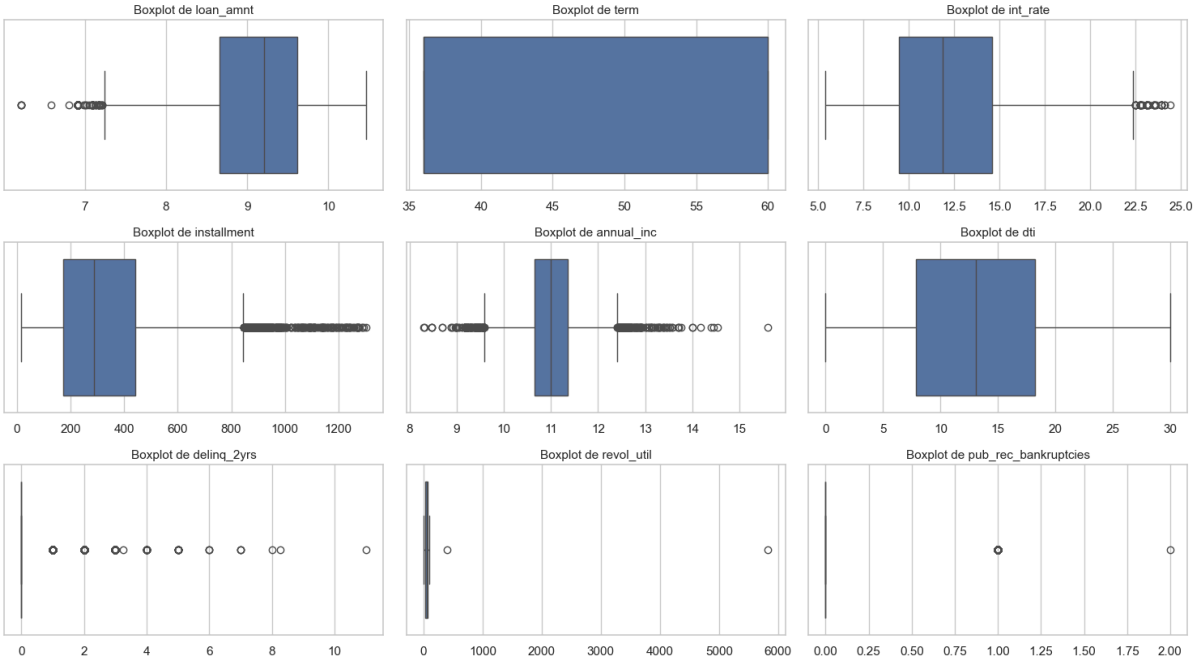
Visualizaciones de EDA univariado

1.1 Histogramas



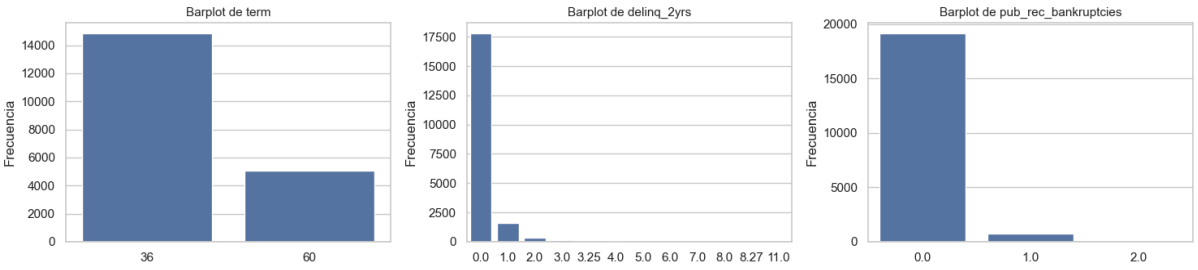
Fuente: Elaboración propia

1.2 Boxplots



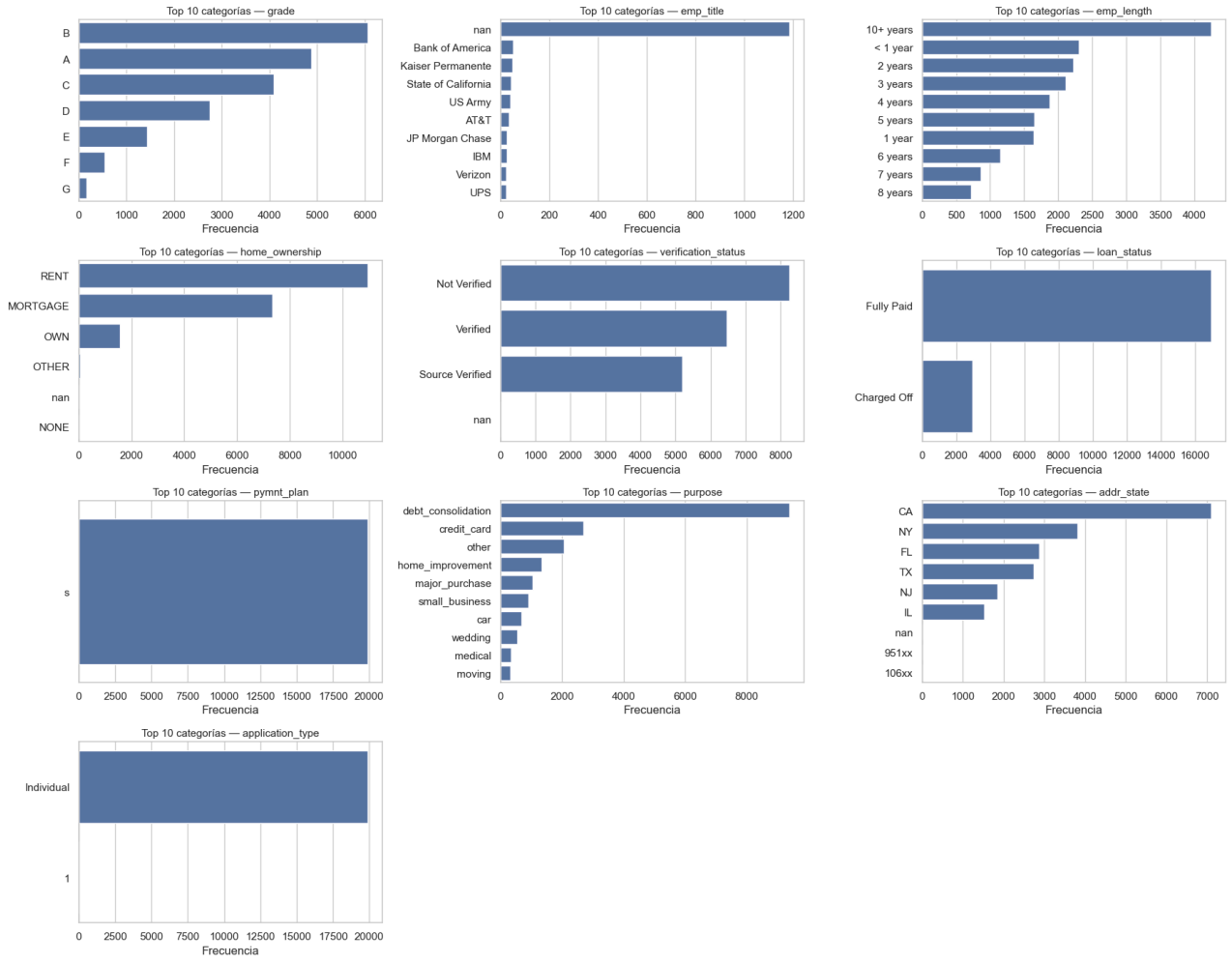
Fuente: Elaboración propia

1.3 Barplots



Fuente: Elaboración propia

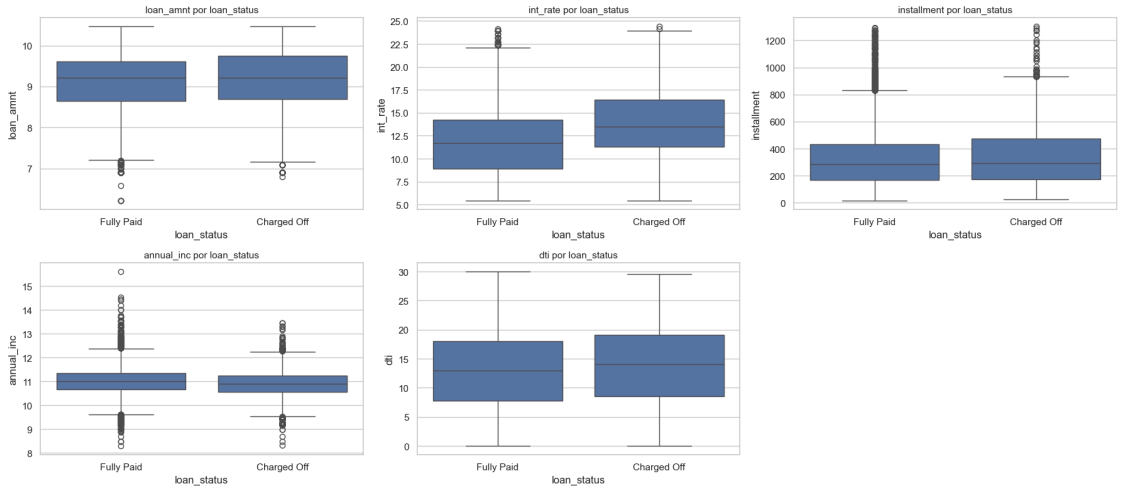
1.4 Countplots



Fuente: Elaboración propia

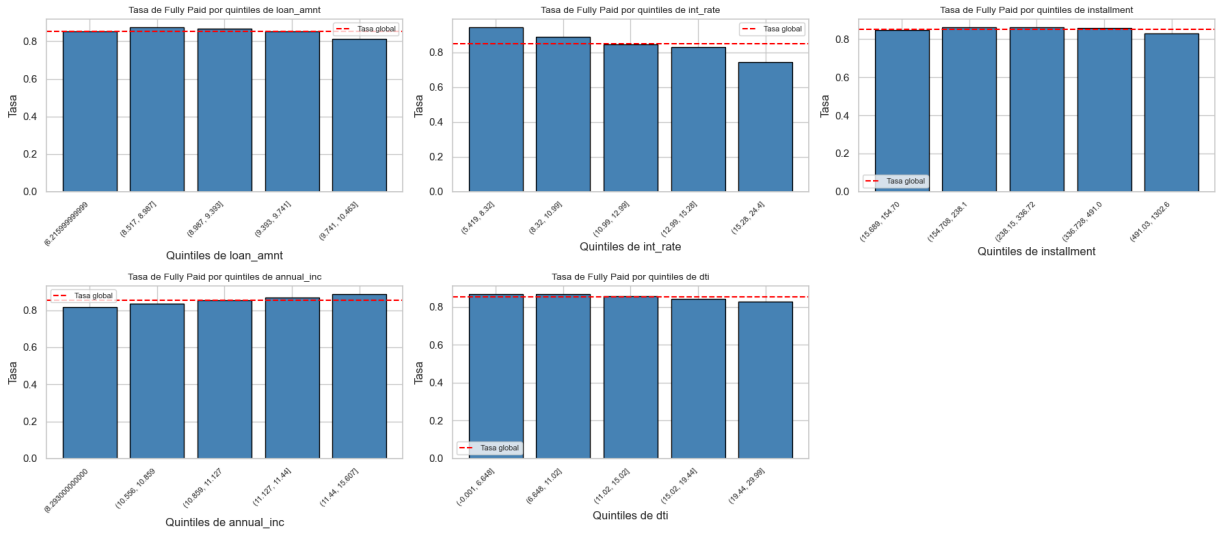
Visualizaciones de EDA bivariado

1.5 Boxplots



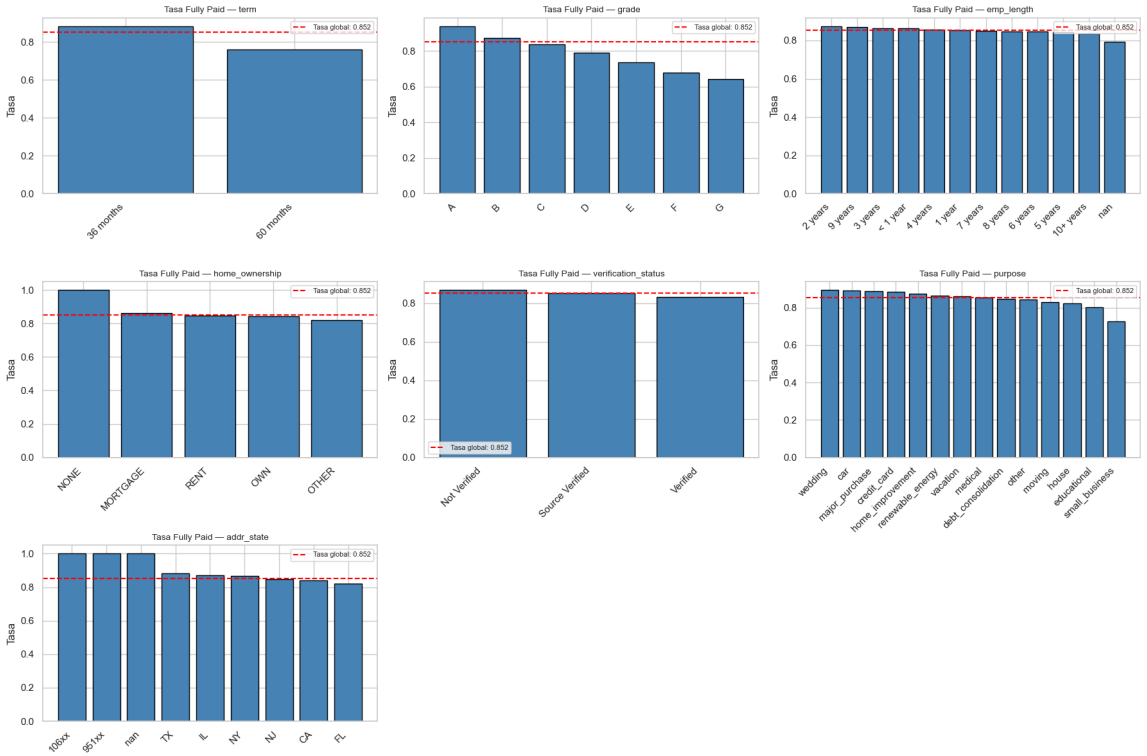
Fuente: Elaboración propia

1.6. Tasa del target por buckets



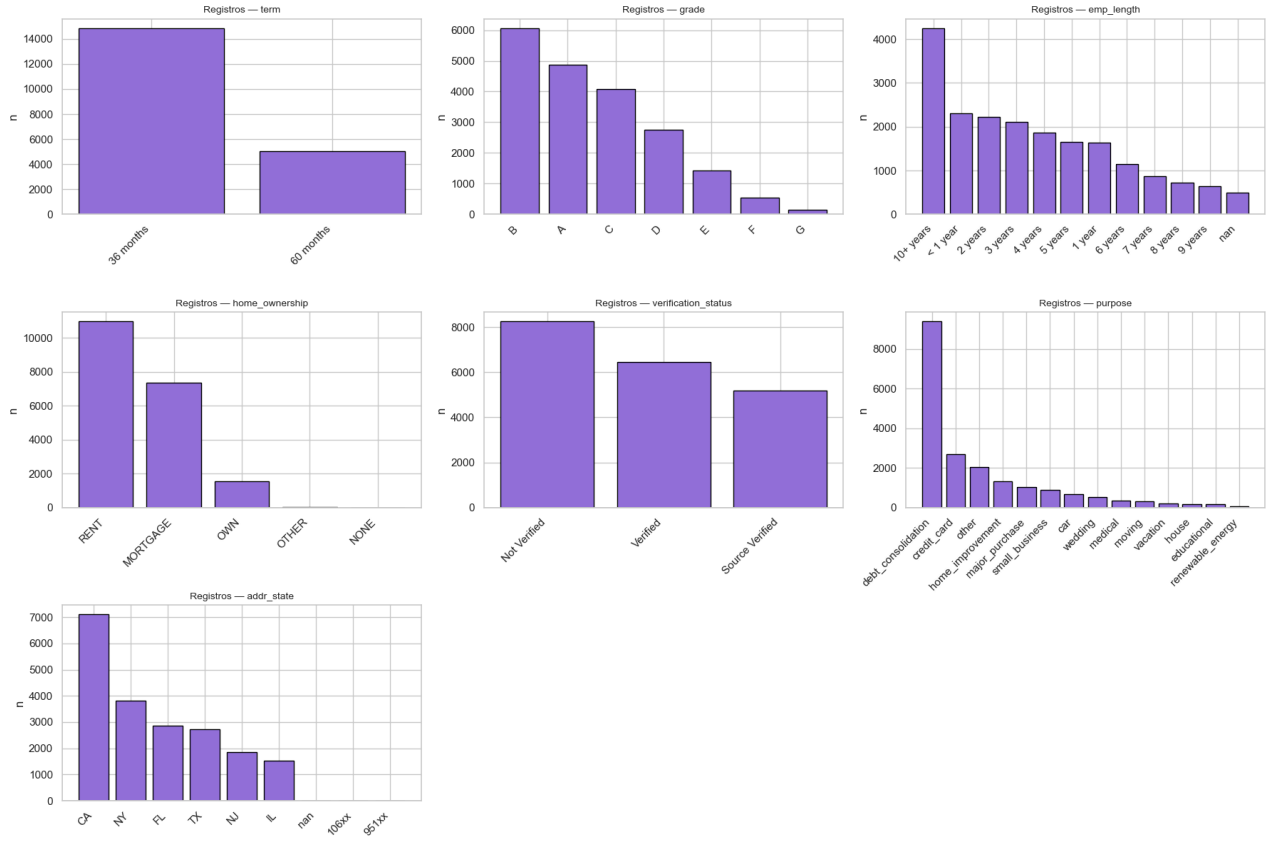
Fuente: Elaboración propia

1.7 Barplots: Tasa por categoría



Fuente: Elaboración propia

1.8 Barplots: Registros por categoría



Fuente: Elaboración propia