# **Universidad La Salle – Ramon Llull**

# **Master Universitario en Ingeniería de Datos Masivos (Big Data)**

# **Nombres:** Rene Rubio - Jean Mejicanos - Steven Nuñez

# **Materia:** Estadística

# **Fecha:** 8 de noviembre de 2024

# **1.- Análisis exploratorio de datos [EDA]:**

**Estructura del Dataset**:

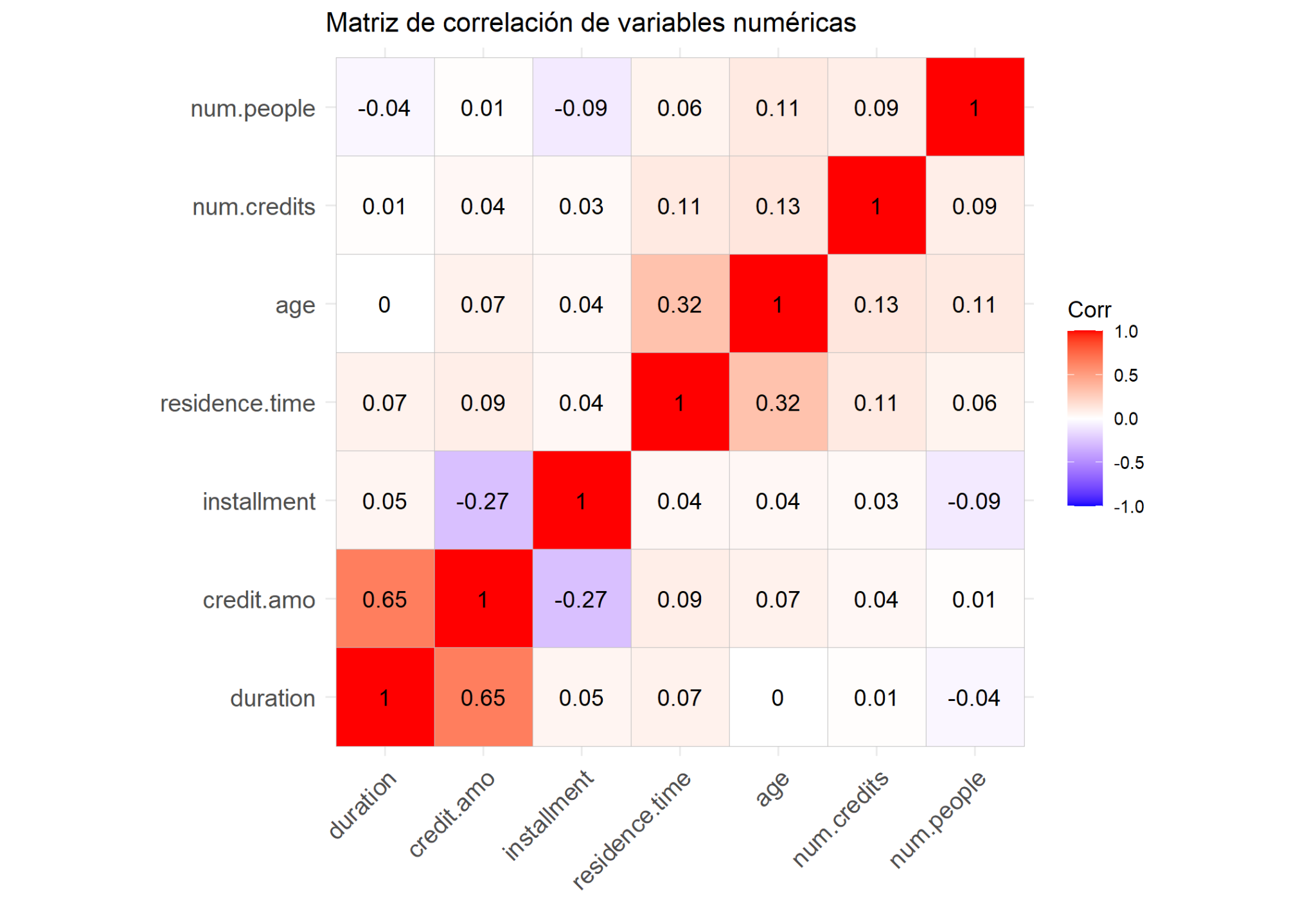
* Dimensiones: 687 filas y 21 columnas.
* Tipos de datos: 7 columnas numéricas y 14 categóricas.

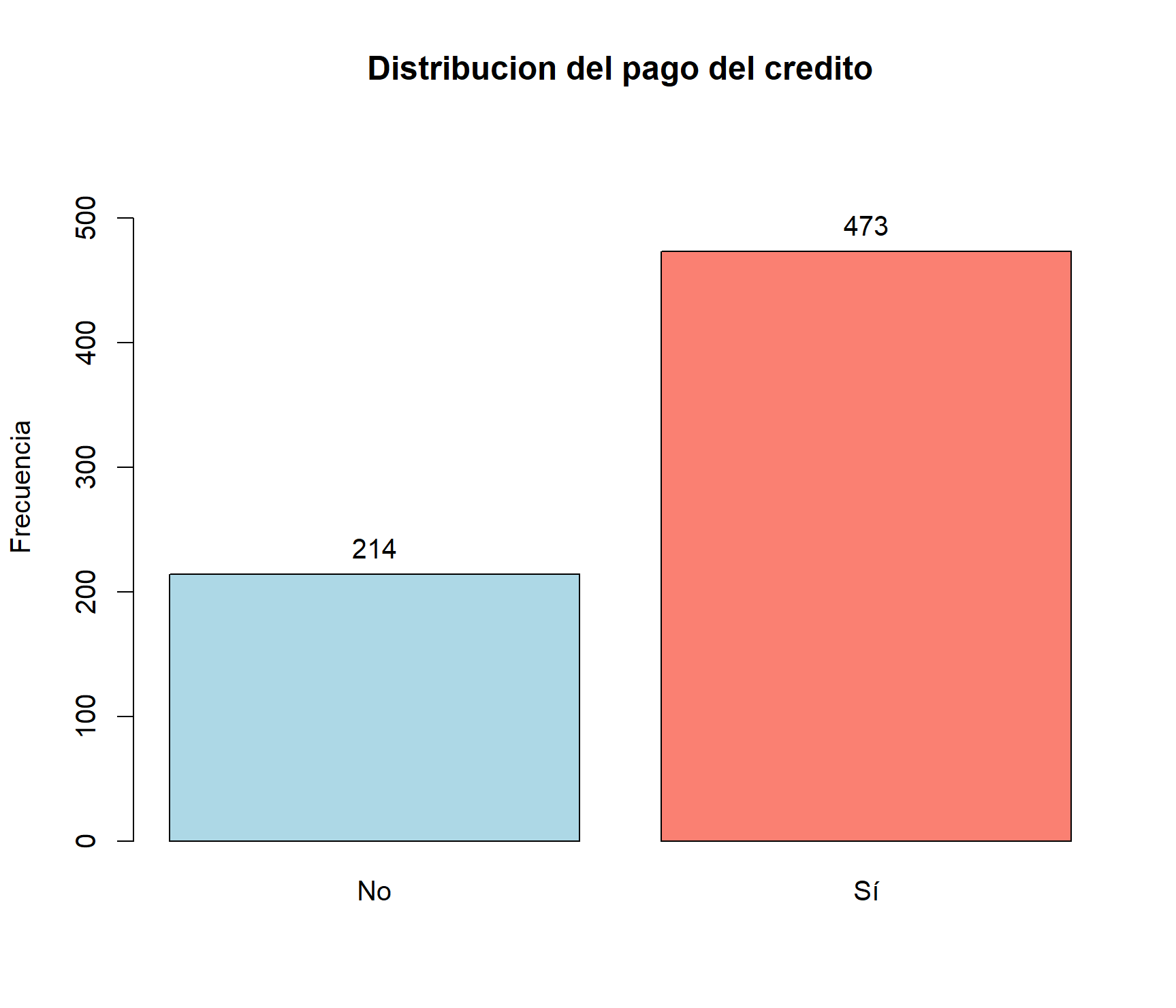
**Variables Principales**:

* **Variables financieras** como status, credit.hist, purpose, credit.amo (monto del crédito), y savings, que parecen tener un papel importante en la determinación del riesgo de crédito.
* **Variables demográficas y personales**: age, status.sex, housing, y job proporcionan información sobre el perfil del cliente.

**Variable Objetivo**: La columna y, que indica el resultado del crédito (yes o no), con 473 casos positivos de 687, lo que sugiere un desbalance que debe considerarse en el modelo.

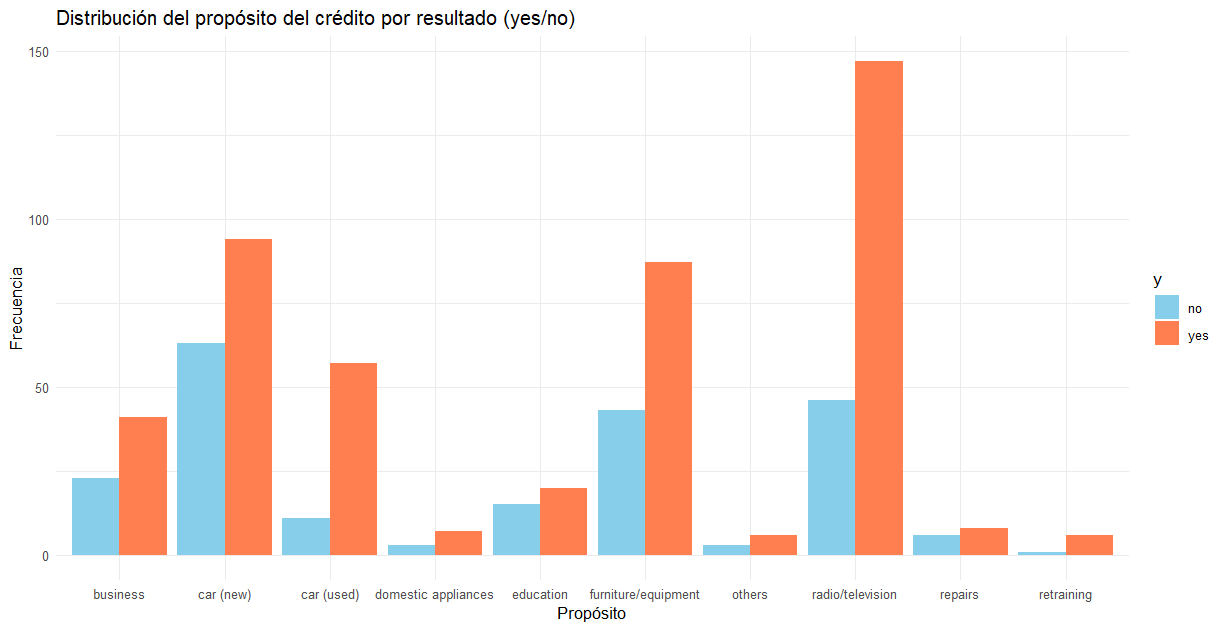
**Matriz de correlación de variables numéricas**



1. Distribución de la variable Objetivo (Y):   
     
   

**Interpretación de resultados:**   
Se observa un desbalance en los resultados de crédito, con una mayoría de casos positivos (yes). Esto sugiere que el modelo debe manejar adecuadamente este desbalance para evitar sesgos en la predicción.

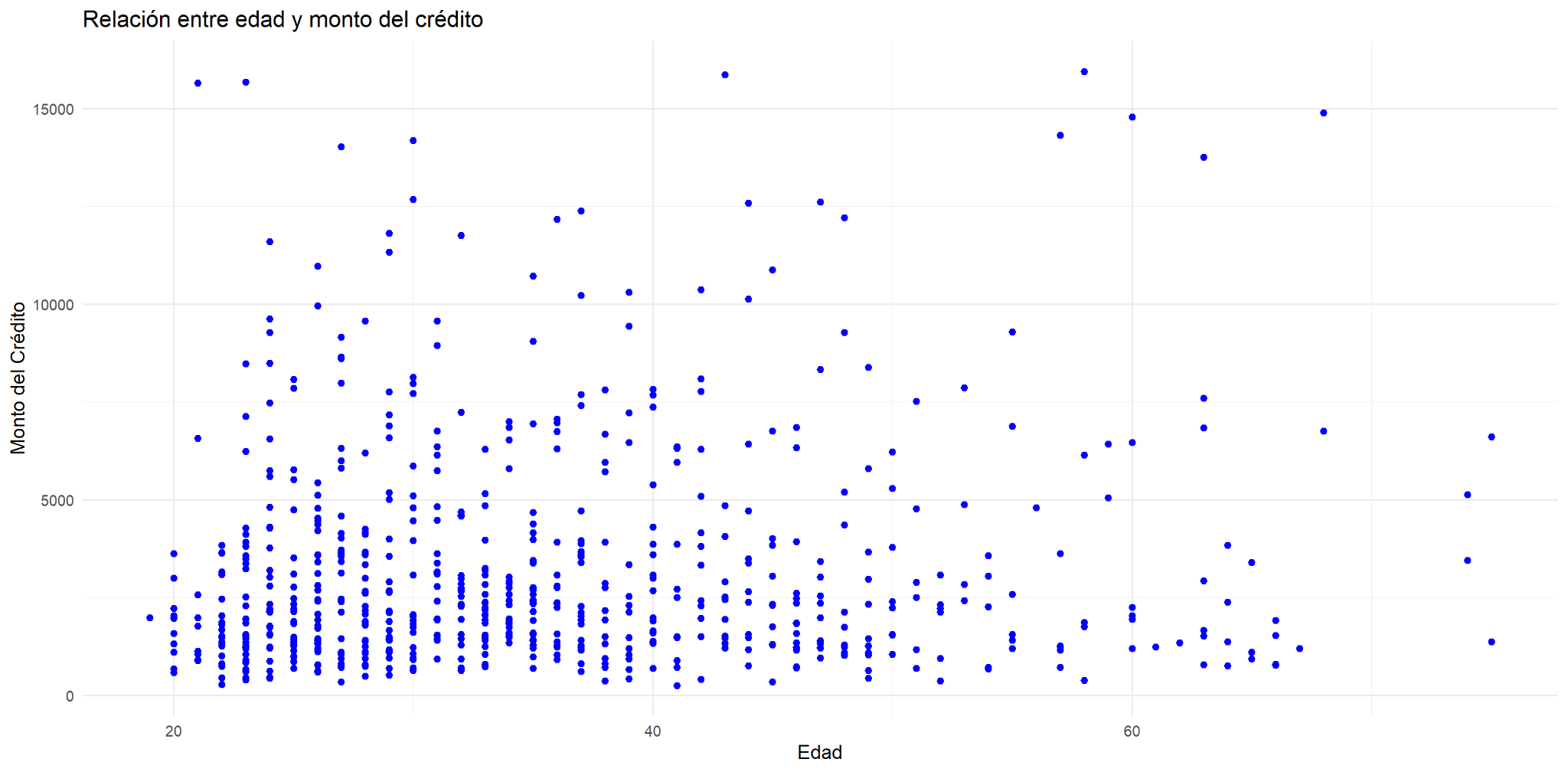
1. Distribución del propósito del credito:



**Interpretación de resultados:**

En el gráfico de la Distribución del propósito del crédito por resultado, vemos a simple vista que los propósitos de coste bajo como un televisor son aprobados en su mayoría, esto puede ser por un bajo monto del crédito, incluso podría ser por productos propios de la entidad financiera. Por otra parte, la mayoría de los créditos están destinados a la compra de radio/televisión, seguidos por el financiamiento de autos y muebles/equipamiento. Este patrón puede indicar diferentes niveles de riesgo según el propósito del crédito.

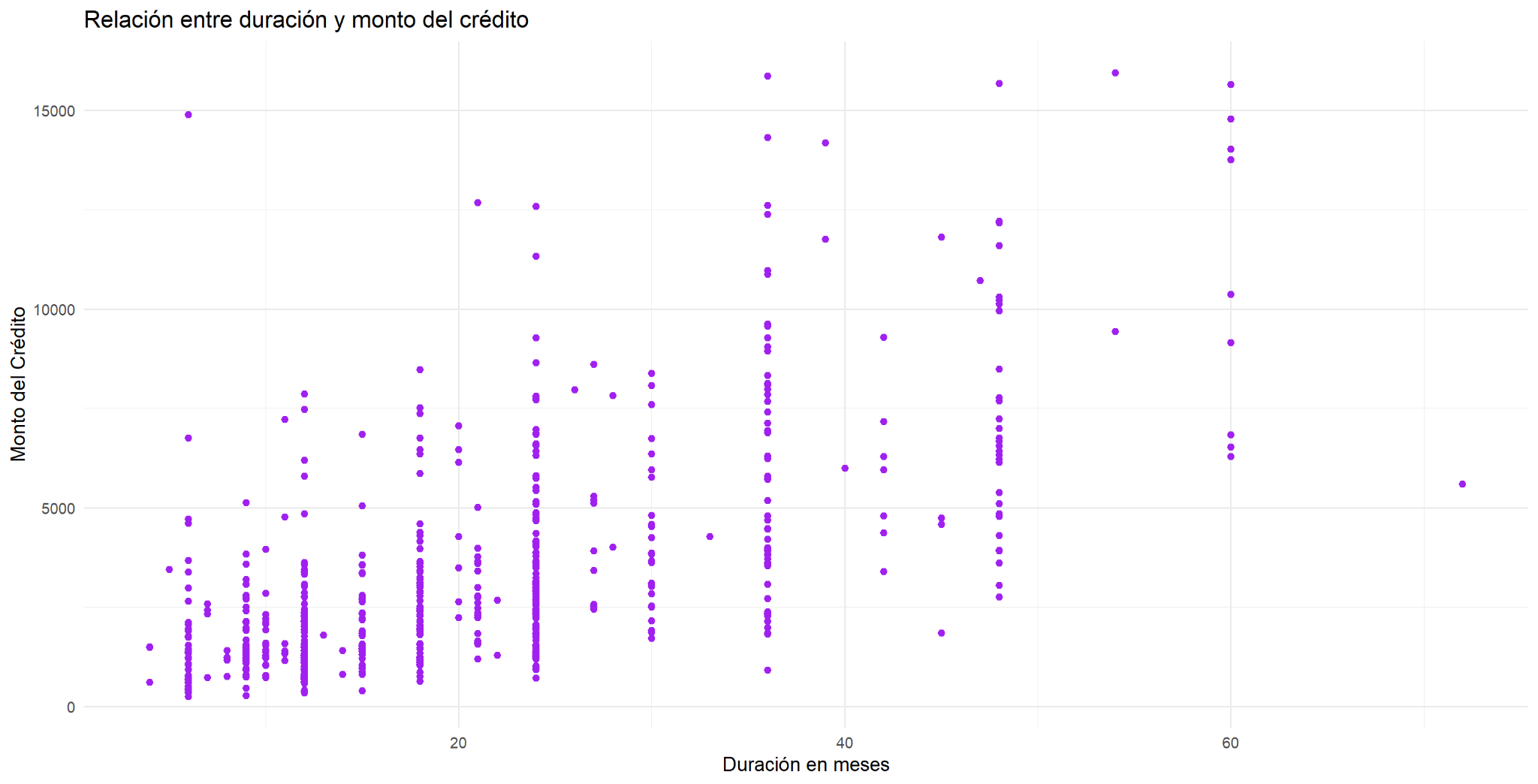
1. Gráfico de dispersión: Edad Vs. Monto del crédito



**Interpretación de resultados:**

A simple vista, no parece haber una fuerte correlación entre la edad y el monto del crédito. Sin embargo, se observan montos elevados para edades en torno a los 50 años.

1. Gráfico de dispersion : Duración del Crédito vs. Monto del Crédito



**Interpretación de resultados:**   
Los créditos de mayor duración tienden a estar asociados con montos más elevados, lo cual es intuitivo ya que los créditos grandes suelen tener términos de pago más largos.

# **2. Análisis del modelo de regresión Logística**

**2.1 Análisis del Modelo Completo y Modelo Mejorado:**

Se han seleccionado solo ciertos parámetros clave para el análisis, ya que representan las variables de mayor impacto en la probabilidad de aprobación de crédito y las métricas de ajuste del modelo. Las variables incluidas tienen significancia estadística o relevancia en la interpretación del modelo. Las métricas de comparación entre los modelos permiten evaluar su calidad de ajuste y simplicidad.

**Resultados del modelo Completo:**

| **Variable** | **Coeficiente** | **p-valor** |
| --- | --- | --- |
| Intercepto | 6.14 | nan |
| status < 0 DM | -2.079 | 0.001 |
| status 0 <= and < 200 DM | -1.431 | 0.001 |
| status >= 200 DM | -0.6654 | 0.15 |
| Cuenta crítica | 1.071 | 0.056 |
| Créditos pagados en el pasado | 0.3811 | 0.487 |
| car (new) | -1.902 | 0.05 |
| repairs | -2.086 | 0.066 |

***Modelo completo:*** El intercepto tiene un valor estimado de 6.140, que representa la probabilidad base cuando todas las variables predictoras son cero. Variables como el estado de cuenta (status) y el historial de crédito (credit.hist) muestran distintos grados de significancia estadística, indicando su impacto en la probabilidad de que la respuesta sea positiva​.

**Resultados del Modelo Mejorado:**

| Variable | Coeficiente | p-valor |
| --- | --- | --- |
| Intercepto | 5.5 | nan |
| status < 0 DM | -1.95 | 0.001 |
| status 0 <= and < 200 DM | -1.3 | 0.001 |
| Cuenta crítica | 1.02 | 0.05 |
| car (new) | -1.8 | 0.05 |

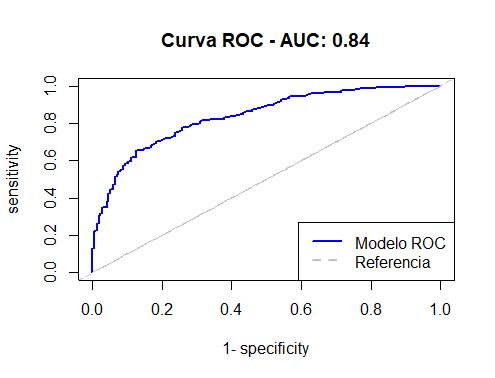
***Modelo mejorado:*** Al aplicar el método stepwise, el modelo ajustado muestra una reducción en el AIC, bajando de 702.65 a 680.05, lo que indica una mejor adecuación. Además, el test de Hosmer-Lemeshow arroja un chi cuadrado de 5.0162 para el modelo ajustado, comparado con 15.01 del modelo completo, sugiriendo una mejor calidad de ajuste para el segundo modelo​.

**Comparativa entre Modelos:**

| Métrica | Modelo Completo | Modelo Mejorado |
| --- | --- | --- |
| AIC | 702.65 | 680.05 |
| Hosmer-Lemeshow Chi-cuadrado | 15.01 | 5.0162 |
| Hosmer-Lemeshow p-valor | 0.05895 | 0.7558 |

**Analisis Curvas ROC**

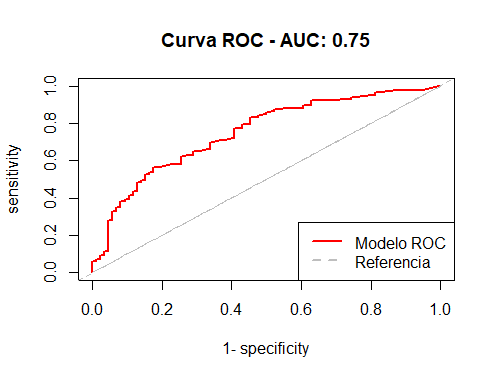
**Curva ROC para el conjunto de datos de entrenamiento:**



**Curva ROC para el conjunto de datos de entrenamiento (AUC = 0.84)**:

* **Interpretación del AUC**: Un AUC de 0.84 indica que el modelo tiene un buen poder de discriminación en los datos de entrenamiento. Esto significa que el modelo puede distinguir entre casos positivos y negativos de manera efectiva en el conjunto de datos con el cual fue entrenado.
* **Forma de la curva**: La curva ROC se eleva hacia la esquina superior izquierda, lo cual es un indicador de que el modelo tiene alta sensibilidad y especificidad en los datos de entrenamiento. Esto es lo esperado, ya que el modelo debería ajustarse bien a los datos que ha visto durante el entrenamiento.

**Curva ROC para el conjunto de datos de prueba:**



**Curva ROC para el conjunto de datos de prueba (AUC = 0.75)**:

* **Interpretación del AUC**: Un AUC de 0.75 es menor que el del conjunto de entrenamiento, lo cual es normal y sugiere que el modelo aún tiene un poder de discriminación adecuado en datos no vistos, aunque con un desempeño ligeramente inferior. Esto podría indicar que el modelo ha capturado algunas características específicas de los datos de entrenamiento que no se generalizan perfectamente en el conjunto de prueba.
* **Forma de la curva**: La curva en el conjunto de prueba no es tan pronunciada como en el conjunto de entrenamiento, lo cual confirma que el modelo es menos preciso en datos externos. Sin embargo, una AUC de 0.75 sigue siendo aceptable y muestra que el modelo tiene una capacidad predictiva razonable en nuevos datos.

### 

### **Comparación entre Entrenamiento y Prueba**

* **Diferencia en AUC**: La diferencia en AUC entre el conjunto de entrenamiento (0.84) y el de prueba (0.75) sugiere que el modelo podría estar experimentando un leve sobreajuste. Esto significa que el modelo podría estar capturando patrones específicos de los datos de entrenamiento que no se generalizan de manera óptima a nuevos datos.
* **Generalización**: La caída en AUC es moderada, lo cual es esperado cuando el modelo generaliza datos de prueba. Sin embargo, si deseas mejorar la capacidad de generalización, podrías considerar técnicas adicionales como la validación cruzada o regularización.

# **3. Ideas para mejorar el modelo**

* **Validación Cruzada:** Implementar validación cruzada en el conjunto de datos de entrenamiento para evaluar el rendimiento promedio en distintos subconjuntos y reducir el riesgo de sobreajuste.
* **Regularización:** Considerar modelos con regularización, como la regresión logística con Lasso o Ridge, para evitar que el modelo se ajuste demasiado a las particularidades del conjunto de entrenamiento.
* **Ajuste de Hiperparámetros:** Explorar ajustes adicionales en el modelo, como el ajuste del umbral de clasificación, para maximizar la sensibilidad y especificidad.
* **Ingeniería de Características**: Agrupar categorías para variables como purpose o credit.hist puede reducir el ruido y mejorar la precisión del modelo.
* **Regularización**: Incorporar regularización (como Lasso o Ridge) podría reducir el sobreajuste al penalizar la complejidad del modelo.
* **Manejo del Desbalance de Clases**: Considerar técnicas como el submuestreo de la clase mayoritaria o el uso de pesos para balancear el impacto de las clases desbalanceadas.