

Máster Universitario en Big Data y Ciencia de Datos

13MBID_10_A_2023-24

Metodologías de gestión y diseño de proyectos Big Data

Actividad 1

Alumnos:

- **Calampa Tantachuco, Colbert Moises Bryan**
- **Miranda Villalón, Elena**

Edición octubre 2023-24

Contenido

Actividades Prácticas - Aplicando técnicas ágiles para la gestión de proyectos de ciencia de datos 3

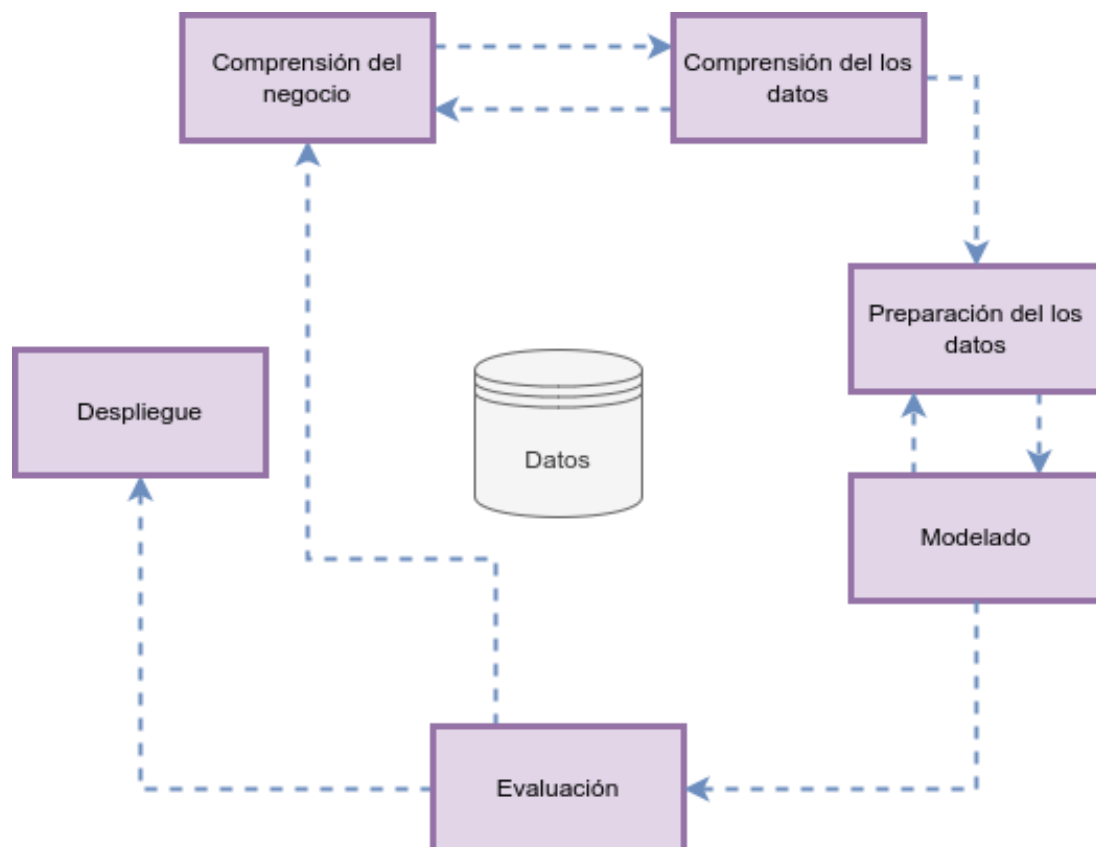
[A] Comprensión del negocio.....	4
Determinar los objetivos de la Organización	4
Evaluación de la situación	4
Determinación de los objetivos del proyecto.....	4
Definir plan del proyecto	5
[B] Comprensión de los datos	6
Recolección de datos iniciales	6
Descripción de los datos	6
Exploración de datos	6
Verificación de la calidad de los datos.....	16
Registro de metadatos de cada dataset	18
[C] Fase de preparación de los datos.....	22
Selección de datos.....	22
Limpieza de los datos.....	22
Integración de los datos.....	23
Construcción de datos.....	23
Formateo de los datos	26

Actividades Prácticas - Aplicando técnicas ágiles para la gestión de proyectos de ciencia de datos

El presente documento es una planilla que se utilizará para el desarrollo de la documentación correspondiente a las Actividades Prácticas I y II. El contenido será guiado según las fases y actividades de la metodología CRISP-DM.

Una vez completado con la información correspondiente al proyecto de ciencia de datos y complementado con los reportes de la ejecución de la libreta Jupyter desarrollada se podrán finalizar las tareas del proyecto.

La metodología CRISP-DM cuenta con 6 fases, ver figura 1, que forman un ciclo iterativo, con vistas a lo que se podrá considerar como un proceso iterativo-incremental de desarrollo de soluciones de ciencia de datos para un contexto en particular.



[A] Comprensión del negocio

Determinar los objetivos de la Organización

Las autoridades de una entidad financiera desean obtener conocimiento a partir de su base de datos histórica de créditos otorgados. Para esta tarea, los datos disponibles se agrupan en dos dimensiones:

- **Datos de créditos:** que contienen la información de los créditos solicitados por los clientes y si los mismos han sido considerados en mora en algún momento.
- **Datos de otros productos:** que contienen la información sobre otros productos (en particular tarjetas de crédito) que poseen los clientes con la entidad y un resumen de su actividad y características principales

Evaluación de la situación

Se cuenta con los siguientes recursos para la ejecución del proyecto:

- Los datos históricos de los créditos solicitados por los clientes y de otros productos obtenidos por ellos.
- Se cuenta con el personal adecuado para la ejecución de las tareas involucradas en el proyecto.
- Se cuenta con un experto en el dominio para abordar dudas o cuestiones de detalladas que pudieran surgir.
- Se cuenta con las herramientas tanto software como hardware para el desarrollo y despliegue de los productos que pudieran surgir del proyecto.

Determinación de los objetivos del proyecto

Considerando los datos disponibles, los objetivos planteados en esta etapa de trabajo son:

- Generar un producto de datos orientado a la **visualización** de aspectos de interés en los datos disponibles.
- Aplicar técnicas de **aprendizaje automático** para identificar grupos de interés entre los clientes de la entidad y poder comprender tal composición para generar campañas de fidelización personalizadas.

Como condición necesaria para el uso de los resultados obtenidos en una instancia de producción, se requiere que:

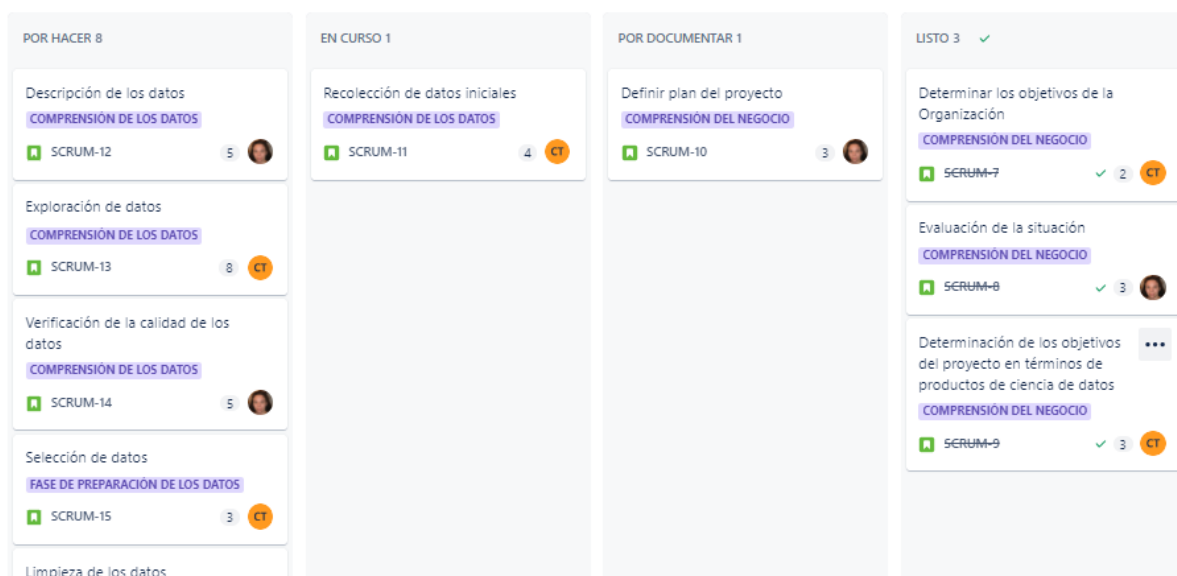
- Para la visualización: identificar indicadores clave que se desee visualizar mediante gráficos que permitan apreciar la distribución.
- Para el proceso de agrupamiento: identificar grupos de clientes que posean características similares y se pueda obtener un conjunto de reglas que definan su comportamiento.

Definir plan del proyecto

El proyecto se gestionará para el seguimiento de sus tareas a través de la herramienta Jira. En su versión online están definidas todas las historias de usuario correspondientes a la generación del MVP del proyecto según los objetivos expresados anteriormente. El enlace a la herramienta se adjunta a continuación:

<https://colbert6.atlassian.net/jira/software/projects/SCRUM/boards/1/backlog>

El proyecto está planteado en dos iteraciones, donde se han repartido las fases de la metodología CRISP-DM, se adjuntan captura de pantalla de la planificación realizada para la primera iteración:



[B] Comprensión de los datos

Recolección de datos iniciales

Se cuenta con dos *datasets* exportados desde los sistemas transaccionales de la organización en formato *.csv*:

- Datos de créditos [*datos_creditos.csv*]: que contienen la información de los créditos solicitados por los clientes y si los mismos han sido considerados en mora en algún momento.
- Datos de otros productos [*datos_tarjetas.csv*]: que contienen la información sobre otros productos (en particular tarjetas de crédito) que poseen los clientes con la entidad y un resumen de su actividad y características principales.

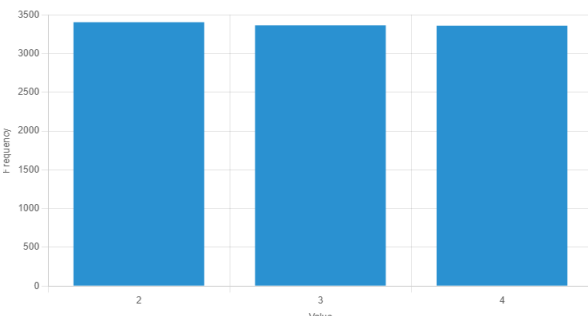

Descripción de los datos

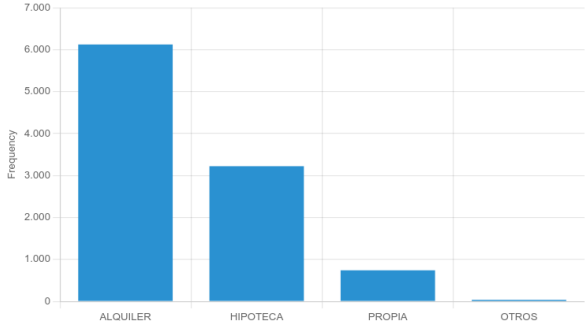
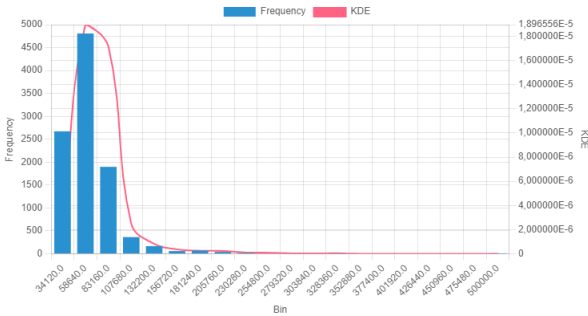
Se describen las características principales de cada dataset:

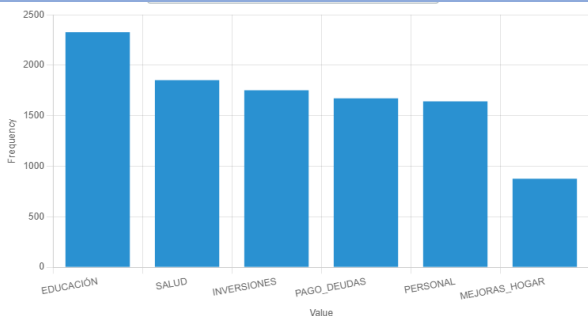
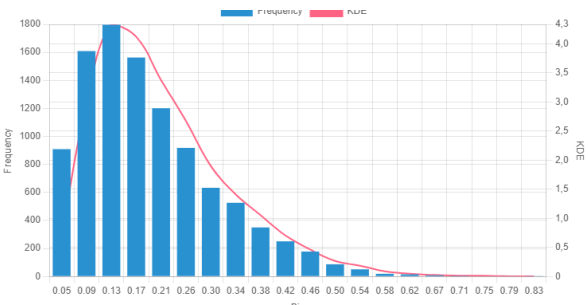
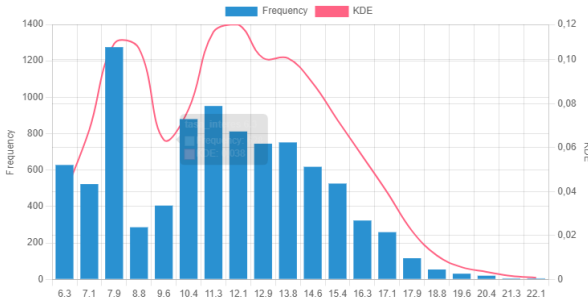
Dataset	Columnas / Atributos	Cantidad de filas
datos_creditos.csv	<ul style="list-style-type: none">• id_cliente• edad• importe_solicitado• duracion_credito• antiguedad_empleado• situacion_vivienda• ingresos• objetivo_credito• pct_ingreso• tasa_interes• estado_credito• falta_pago	Cantidad de filas: 10127
datos_tarjetas.csv	<ul style="list-style-type: none">• id_cliente• antiguedad_cliente• estado_civil• estado_cliente• gastos_ult_12m• genero• limite_credito_tc• nivel_educativo• nivel_tarjeta• operaciones_ult_12m• personas_a_cargo	Cantidad de filas: 10127

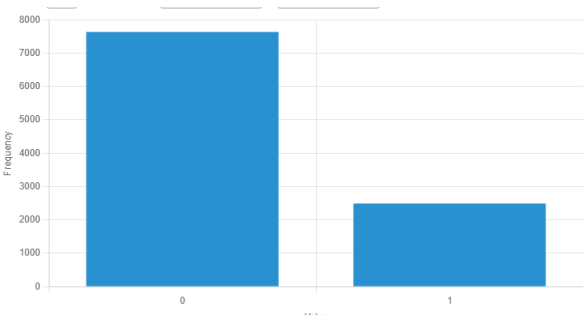
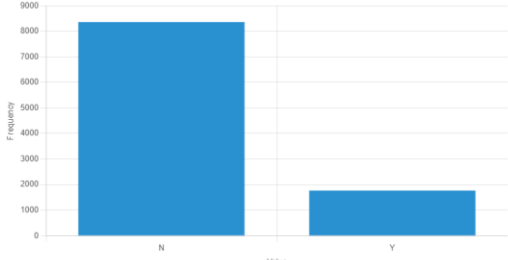
Exploración de datos

Se describen los metadatos de cada dataset:

		<ul style="list-style-type: none"> • min:2 • mode:2 
antiguedad_empleado	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):9,790 Count (missing):337 % Missing:3.33</p> <p><i>Estadísticas de los valores</i></p> <ul style="list-style-type: none"> • max:123 • mean:3.9385 • median:4 • min:0 
situacion_vivienda	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p><i>Distribución de valores (valor – cantidad - %):</i></p> <ul style="list-style-type: none"> • ALQUILER - 6,125 - 60.48% • HIPOTECA - 3,223 - 31.83% • PROPIA – 741 - 7.32% • OTROS – 38 - 0.38%

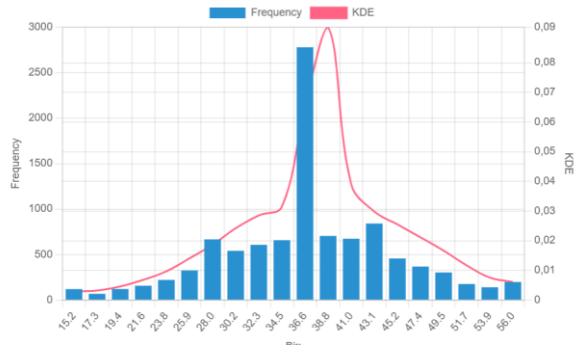
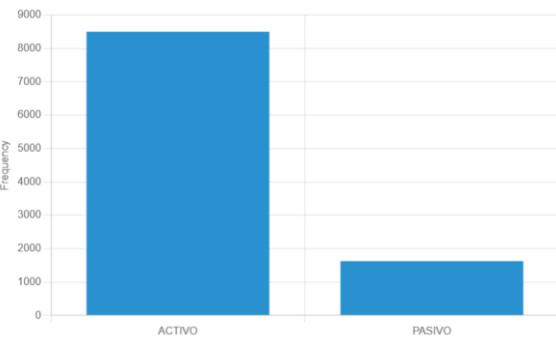
		
ingresos	Int64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores</i></p> <ul style="list-style-type: none"> • max:500,000 • mean:50,381.8976 • median:46,000 • min:9,600 • mode:60,000 
objetivo_credito	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p><i>Distribución de valores (valor – cantidad - %):</i></p> <ul style="list-style-type: none"> • EDUCACIÓN - 2,328 - 22.99% • SALUD - 1,853 - 18.30% • INVERSIONES - 1,753 - 17.31% • PAGO_DEUDAS - 1,673 - 16.52% • PERSONAL - 1,643 - 16.22% • MEJORAS_HOGAR – 877- 8.66%

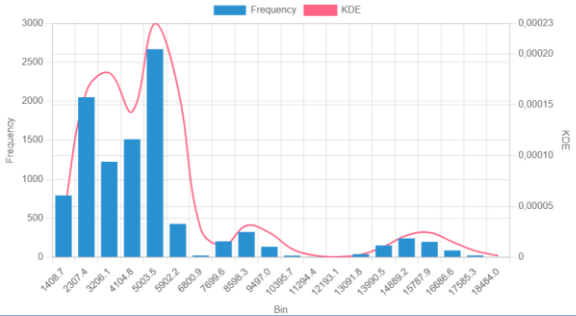
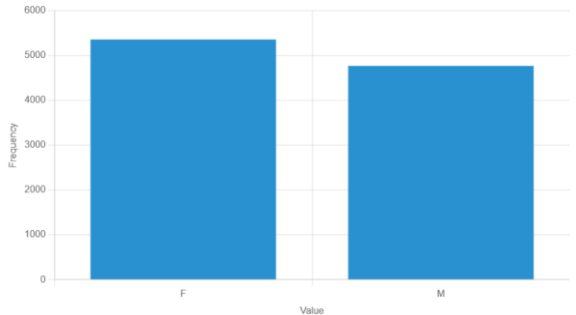
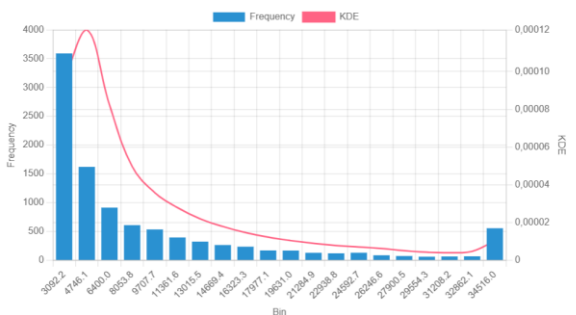
		
pct_ingreso	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores</i></p> <ul style="list-style-type: none"> • max:0.83 • mean:0.1772 • median:0.15 • min:0.01 
tasa_interes	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):9,215 Count (missing):912 % Missing:9.01</p> <p><i>Estadísticas de los valores</i></p> <ul style="list-style-type: none"> • max:22.11 • mean:10.9794 • median:10.99 • min:5.42 
estado_credito	Int64 (categorico)	<p>Total Rows:10,127</p>

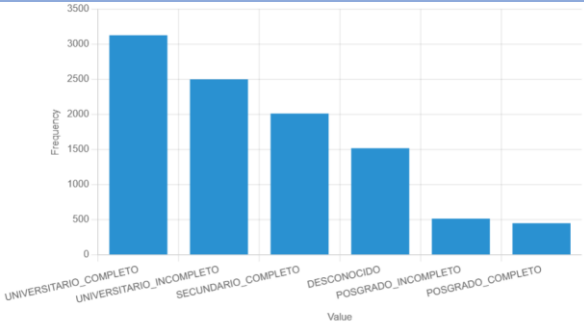
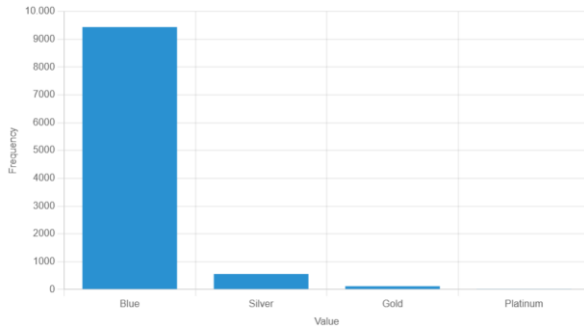
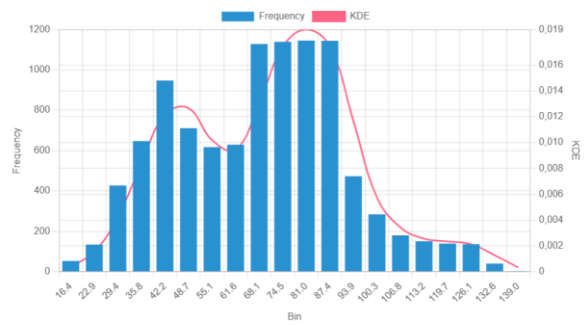
		<p>Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>):</p> <ul style="list-style-type: none"> 0 - 7,635 - 75.39% 1 - 2,492 - 24.61% 
falta_pago	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>):</p> <ul style="list-style-type: none"> N - 8,359 - 82.54% Y - 1,768 - 17.46% 

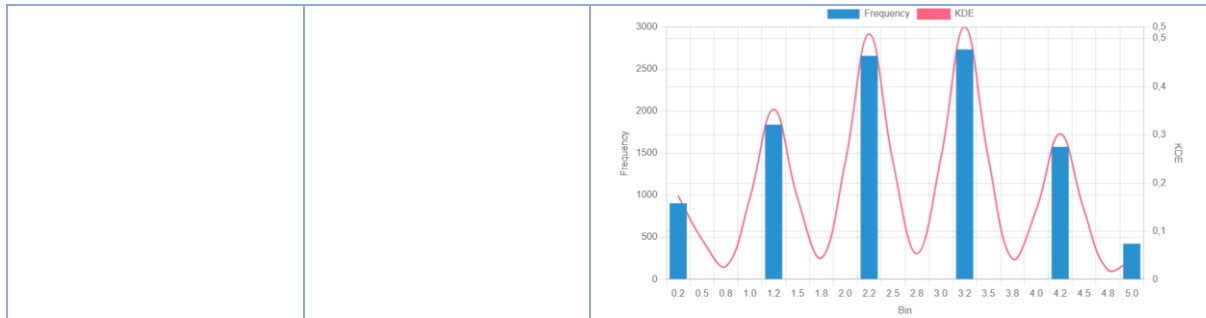
- Dataset: datos_tarjetas.csv

Columnas / Atributos	Tipo de datos	Observaciones
id_cliente	Float64 (numérico)	Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0
antigüedad_cliente	Float64 (numérico)	Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0

																	
estado_civil	String (nominal)	<div>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</div> <table><tr><td>CASADO</td><td>4,687</td><td>46.28%</td></tr><tr><td>SOLTERO</td><td>3,943</td><td>38.94%</td></tr><tr><td>DESCONOCIDO</td><td>749</td><td>7.40%</td></tr><tr><td>DIVORCIADO</td><td>748</td><td>7.39%</td></tr><tr><td>TOTAL</td><td>10,127</td><td>100.00%</td></tr></table>	CASADO	4,687	46.28%	SOLTERO	3,943	38.94%	DESCONOCIDO	749	7.40%	DIVORCIADO	748	7.39%	TOTAL	10,127	100.00%
CASADO	4,687	46.28%															
SOLTERO	3,943	38.94%															
DESCONOCIDO	749	7.40%															
DIVORCIADO	748	7.39%															
TOTAL	10,127	100.00%															
estado_cliente	String (nominal)	<div>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</div> 															
gastos_ult_12m	Float64 (numérico)	<div>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</div>															

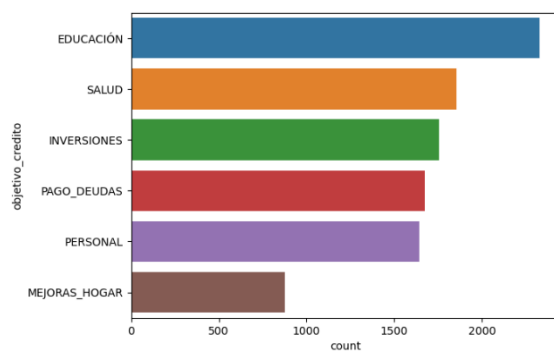
		
genero	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> 
limite_credito_tc	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Estadística de valores:</p> <ul style="list-style-type: none"> • max:34,516 • mean:8,631.9537 • median:4,549 • min:1,438.3 
nivel_educativo	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p>

		
nivel_tarjeta	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> 
operaciones_ult_12m	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> 
personas_a_cargo	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p>

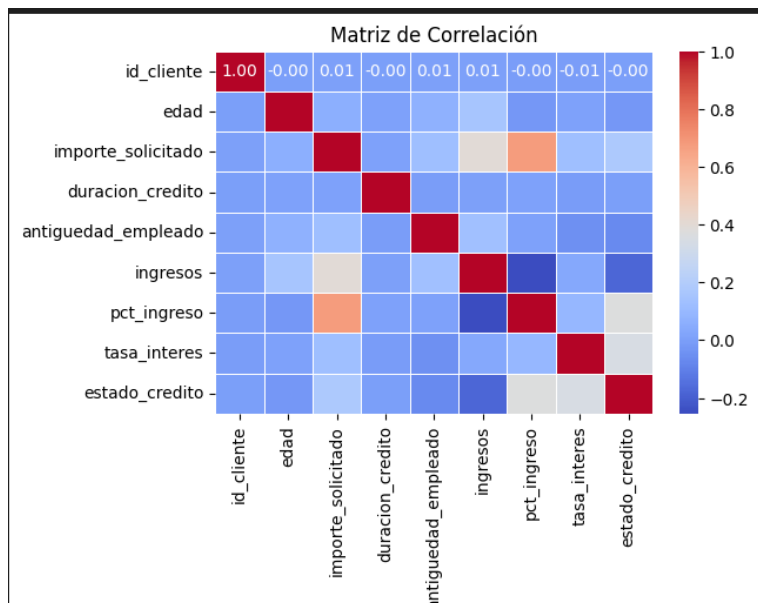


Se adjuntan a continuación algunos gráficos de interés:

- Distribución del tipo de crédito solicitado por los clientes

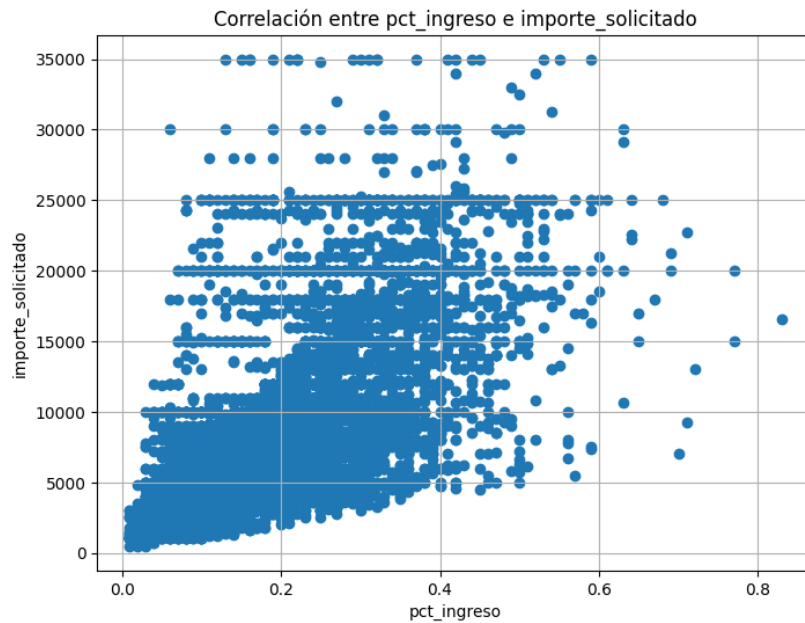


- Matriz de correlación de las variables aplicada al Dataset: datos_creditos.csv



Mediante la matriz de correlación aplicada podemos identificar una relación medianamente fuerte entre las variables “pct_ingreso” y “importe_solicitado”. Se procederá a realizar un análisis de estas dos variables en conjunto.

- Gráfico de dispersión entre las variables pct_ingreso vs importe_solicitado



El gráfico de dispersión nos muestra cierta correlación positiva entre las variables analizadas, también se identifica un amplio grupo entre pct_ingreso menores a 0.4 y importe_solicitado menores a 15000.

Verificación de la calidad de los datos

Definición de objetivos y características de la evaluación inicial

Descripción del uso propuesto

Las autoridades de una entidad financiera desean obtener conocimiento a partir de su base de datos histórica de créditos otorgados.

Según este objetivo, se desarrolla un proyecto de ciencia de datos para desarrollar un producto de datos que sea una propuesta de solución para este escenario.

En este contexto, se requiere realizar un análisis de calidad de los datos disponibles para dar cumplimiento a lo establecido en la fase Comprensión de los Datos de la metodología CRISP-DM con la que se está gestionando el mencionado proyecto.

Definición de calidad

Se van a analizar los siguientes atributos de calidad:

Atributo	Observaciones
----------	---------------

Exactitud	Grado en el que los datos de un atributo representan un valor verdadero.
Compleitud	Grado en el que los datos de un registro tienen valores asociados a cada una de sus columnas y el dataset en general aplica el mismo criterio para todas sus filas.
Consistencia	Grado en el cual los datos son coherentes con otros datos del contexto y con los conjuntos de datos disponibles para este proyecto.

Cada una de las dimensiones definidas por los atributos antes listados, será relacionada con una o más características a analizar a fin de establecer la calidad de los datos disponibles:

Atributo	Características que analizar
Exactitud	Cumplimiento de reglas de formateo. Cumplimiento de reglas del negocio.
Compleitud	Compleitud de registros y del dataset.
Consistencia	Unicidad en atributo clave. Cumplimiento de integridad referencial.

Características que deben cumplir los datos

Dimensión	Característica	Granularidad	Umbral de aceptación
Compleitud	Compleitud a nivel de filas	Filas	20%
	Compleitud a nivel del dataset	Dataset	10%
Exactitud	Cumplimiento de reglas de formateo	Dataset	10%
	Cumplimiento de reglas de valores	Filas	0%
	Cumplimiento de reglas de negocio	Dataset	10%
Consistencia	Unicidad en atributos clave	Dataset	0%
	Integridad referencial	Dataset	10%

Registro de metadatos de cada dataset

Tarea resuelta en las actividades anteriores.

Evaluación inicial de los datos disponibles

Se inicia expresando la definición de las métricas aplicables para la medición de las características mencionadas en la sección anterior.

Identificador	Descripción	Forma de realizar la medición	Umbral de aceptación
completitud_f	Compleitud a nivel de filas	atributos_vacios / total_atributos	20%
completitud_d	Compleitud a nivel del dataset	filas_con_vacios / total_filas	10%
formato_valido	Cumplimiento de reglas de formateo	filas_no_cumplen_formato / total_filas	10%
valores_ajustados	Cumplimiento de reglas de valores	filas_fuera_rango / total_filas	0%
valores_errores	Cumplimiento de reglas de negocio	filas_con_errores / total_filas	10%
claves_unicas	Unicidad en atributos clave	filas_claves_duplicadas / total_filas	0%
integridad_referencial	Integridad referencial	filas_con_problemas_relacion / total_filas	10%

Una vez aplicados los cálculos descritos en la tabla anterior se obtendrán los valores necesarios para realizar la evaluación de calidad de los datos en sí, los resultados se registran en las siguientes tablas.

Dimensión: Completitud

datos_créditos:

Indicador	Umbral de aceptación	Resultados obtenidos	Evaluación
completitud_f	20%	Filas que incumplen el umbral de nulos en columnas [completitud_f] - datos_créditos: 0 (0.0)%	OK
completitud_d	10%	Filas que presentan nulos en el dataset [completitud_d] - datos_creditos: 1225 (12.1)%	No cumplimiento

datos_tarjetas:

Indicador	Umbral de aceptación	Resultados obtenidos	Evaluación
completitud_f	20%	Filas que incumplen el umbral de nulos en columnas [completitud_f] - datos_tarjetas: 0 (0.0)%	OK
completitud_d	10%	Filas que presentan nulos en el dataset [completitud_d] - datos_tarjetas: 0 (0.0)%	OK

Dimensión: Exactitud

Indicador	Umbral de aceptación	Resultados obtenidos	Evaluación
formato_valido	10%	No se encuentran atributos que requieran análisis para este indicador.	OK
valores_ajustados	0%		No cumplimiento (3/11)
Atributo "edad"		Cantidad de filas con valores fuera de rango en atributo edad (%): 4 (0.04%)	No cumplimiento
Atributo "antigüedad_empleado"		Cantidad de filas con valores fuera de rango en atributo edad: 339 (3.35%)	No cumplimiento
Atributo "importe_solicitado"		Cantidad de filas con valores fuera de rango en atributo importe_solicitado (%): 0 (0.00%)	OK

Atributo "duracion_credito"		Cantidad de filas con valores fuera de rango en atributo duracion_credito (%): 0 (0.00%)	OK
Atributo "situacion_vivienda"		Cantidad de filas con valores fuera de rango en atributo situacion_vivienda (%): 0 (0.00%)	OK
Atributo "objetivo_credito"		Cantidad de filas con valores fuera de rango en atributo objetivo_credito (%): 0 (0.00%)	OK
Atributo "ingresos"		Cantidad de filas con valores fuera de rango en atributo ingresos (%): 0 (0.00%)	OK
Atributo "pct_ingreso"		Cantidad de filas con valores fuera de rango en atributo pct_ingreso (%): 0 (0.00%)	OK
Atributo "tasa_interes"		Cantidad de filas con valores fuera de rango en atributo tasa_interes: 930(9.18%)	No cumplimiento
Atributo "estado_credito"		Cantidad de filas con valores fuera de rango en atributo estado_credito (%): 0 (0.00%)	OK
Atributo "falta_pago"		Cantidad de filas con valores fuera de rango en atributo falta_pago (%): 0 (0.00%)	OK
valores_errores	10%		OK (2/2)
Regla 1: <i>"Para aquellos casos en que los créditos constituyan un porcentaje de los ingresos del cliente mayor al 50% sus ingresos deberán ser mayores a 20.000."</i>		Cantidad de filas que no cumplen la regla: 15 (~0.15%)	OK
Regla 2: <i>"Para aquellos créditos cuya duración sea la mínima permitida el porcentaje de los ingresos del cliente (con respecto al importe solicitado) no podrá exceder el 60% salvo en los casos en los que sea propietario de su vivienda.."</i>		Cantidad de filas que no cumplen la regla: 7 (~0.07%)	OK

datos_tarjetas:

Indicador	Umbral de aceptación	Resultados obtenidos	Evaluación
formato_valido	10%	No se encuentran atributos que requieran análisis para este indicador.	OK
valores_ajustados	0%		No cumplimiento (1/10)
Atributo "antigüedad_cliente"		Cantidad de filas con valores fuera de rango en atributo antigüedad_cliente (%): 103 (1.02%)	No cumplimiento
Atributo "personas_a_cargo"		Cantidad de filas con valores fuera de rango en atributo personas_a_cargo: 0(0%)	OK
Atributo "gastos_ult_12m"		Cantidad de filas con valores fuera de rango en atributo gastos_ult_12m: 0(0%)	OK
Atributo "limite_credito_tc"		Cantidad de filas con valores fuera de rango en atributo limite_credito_tc: 0(0%)	OK
Atributo "operaciones_ult_12m"		Cantidad de filas con valores fuera de rango en atributo operaciones_ult_12m: 0(0%)	OK
Atributo "estado_civil"		Cantidad de filas con valores fuera de rango en atributo estado_civil: 0(0%)	OK
Atributo "estado_cliente"		Cantidad de filas con valores fuera de rango en atributo estado_cliente: 0(0%)	OK
Atributo "genero"		Cantidad de filas con valores fuera de rango en atributo genero: 0(0%)	OK
Atributo "nivel_educativo"		Cantidad de filas con valores fuera de rango en atributo nivel_educativo: 0(0%)	OK
Atributo "nivel_tarjeta"		Cantidad de filas con valores fuera de rango en atributo nivel_tarjeta: 0(0%)	OK

Dimensión: Consistencia

Indicador	Umbral de aceptación	Resultados obtenidos	Evaluación
claves_unicas	0%	Dataset: datos_creditos - No se detectaron claves duplicadas	OK
		Dataset: datos_tarjetas - No se detectaron claves duplicadas	OK
integridad_referencial	10%	Reporte general: - Filas del dataset credits (inicial): 10127 - Filas del dataset tarjetas (inicial): 10127 - Errores detectados en la operación de unión: 0 - Filas del dataset unificado: 10127	OK

[C] Fase de preparación de los datos*Selección de datos*

En función de los resultados del análisis de calidad de datos, se ha determinado remover las siguientes columnas de los datasets disponibles:

- **Datos_creditos:** no se eliminaron columnas
- **Datos_tarjetas:** se eliminó la columna “nivel_tarjeta”

Limpieza de los datos

En esta actividad se aplican filtros a nivel de las filas de cada dataset según las recomendaciones de los expertos en el dominio sobre los resultados de la evaluación de calidad realizada previamente:

Dataset	Atributo	Filtro aplicado	Observaciones
datos_creditos	“edad”	Se filtran aquellas filas donde la edad registrada supera los 90 años	Cantidad de filas filtradas: 4
	“antigüedad_employado”	Se filtran aquellas filas donde la antigüedad_employado registrada supera los 50 años o tenga valores nulos	Cantidad de filas filtradas: 339
	“tasa_interes”	Se filtran aquellas filas donde la tasa_interes supera 20 o tenga valores nulos	Cantidad de filas filtradas: 904

	"regla_pct_ingresos"	Se filtran aquellas filas que no cumplen con la especificación de la regla	Cantidad de filas filtradas: 14
	"regla_pct_ingresos_credito_rapido_no_vienda"	Se filtran aquellas filas cuya duración sea la mínima permitida (2) el porcentaje de los ingresos del cliente (con respecto al importe solicitado) no podrá exceder el 60% salvo en los casos en los que sea propietario de su vivienda.	Cantidad de filas filtradas: 5
datos_tarjetas	"antiguedad_clientes"	Se filtran aquellas filas donde la antiguedad_clientes supera 50.	Cantidad de filas filtradas: 103

Integración de los datos

Se ha realizado la operación de integración (mezcla) de los datasets del escenario en uno nuevo denominado "datos_integrados" tomando como referencia al atributo "id_cliente".

El resultado de dicha operación presenta las siguientes dimensiones:

- Cantidad de columnas: **23**
- Cantidad de filas: **8861**

Construcción de datos

Se documentan a continuación los cambios realizados y la correspondiente generación de nuevos atributos sobre el dataset integrado:

Atributo	Transformación aplicada
"estado_civil"	Cambios realizados para una mejor lectura de los datos: <ul style="list-style-type: none"> • 'CASADO': 'C', • 'SOLTERO': 'S', • 'DESCONOCIDO': 'N', • 'DIVORCIADO': 'D'
"estado_credito"	Cambios realizados para una mejor lectura de los datos: <ul style="list-style-type: none"> • Valor 0 (crédito pendiente de cancelación): 'P', • Valor 1 (crédito cancelado): 'C',
"edad"	Valores numéricos convertidos a ordinales aplicando rangos (<i>etiqueta – valores del rango asociado</i>): <ul style="list-style-type: none"> • 'menor_25': [0, 24], • '25_a_30': [24, 50] (*) <p>(*) Se toma como valor superior del rango a uno que incluye todo el conjunto de datos disponible según el análisis de metadatos.</p>

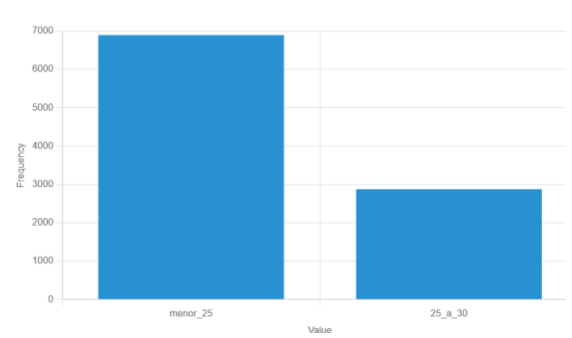
"antigüedad_empleado"	<p>Valores numéricos convertidos a ordinales aplicando rangos (<i>etiqueta – valores del rango asociado</i>):</p> <ul style="list-style-type: none"> • 'menor_5': [0, 4], • '5_a_10': [5, 9], • 'mayor_10': [10, 50] (*) <p>(*) Se toma como valor superior del rango a uno que incluye todo el conjunto de datos disponible según el análisis de metadatos.</p>
"pct_ingreso"	<p>Valores numéricos convertidos a ordinales aplicando rangos (<i>etiqueta – valores del rango asociado</i>):</p> <ul style="list-style-type: none"> • 'hasta_20': [0, 0.19], • '20_a_40': [0.20, 0.39], • '40_a_60': [0.40, 0.59], • 'mayor_60': [0.60, 0.99] (*) <p>(*) Se toma como valor superior del rango a uno que incluye todo el conjunto de datos disponible según el análisis de metadatos</p>
"ingresos"	<p>Valores numéricos convertidos a ordinales aplicando rangos (<i>etiqueta – valores del rango asociado</i>):</p> <ul style="list-style-type: none"> • 'hasta_20k': [0, 19999] • '20k_a_50k': [20000, 49999], • '50k_a_100k': [50000, 99999], • 'mayor_100k': [100000, 999999] (*) <p>(*) Se toma como valor superior del rango a uno que incluye todo el conjunto de datos disponible según el análisis de metadatos</p>
"tasa_interes"	<p>Valores numéricos convertidos a ordinales aplicando rangos (<i>etiqueta – valores del rango asociado</i>):</p> <ul style="list-style-type: none"> • 'hasta_7p': [0, 6.99], • '7p_a_15p': [7, 14.99], • '15p_a_20p': [15, 19.99], • 'mayor_20p': [20, 100] (*) <p>(*) Se toma como valor superior del rango a uno que incluye todo el conjunto de datos disponible según el análisis de metadatos.</p>
"antigüedad_cliente"	<p>Valores numéricos convertidos a ordinales aplicando rangos (<i>etiqueta – valores del rango asociado</i>):</p> <ul style="list-style-type: none"> • 'menor_2y': [0,23], • '2y_a_4y': [24, 47], • 'mayor_4y': [48, 100] (*) <p>(*) Se toma como valor superior del rango a uno que incluye todo el conjunto de datos disponible según el análisis de metadatos.</p>

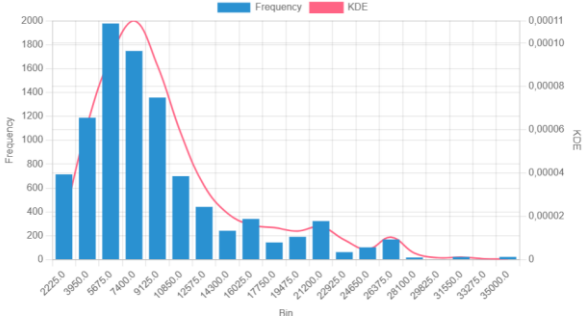
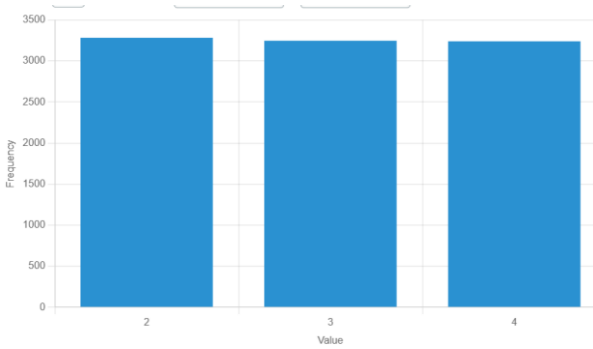
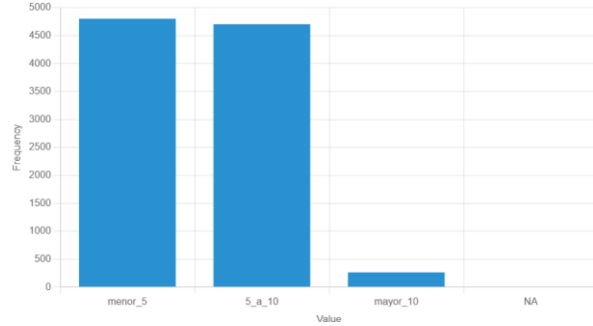
"limite_credito_tc"	<p>Valores numéricos convertidos a ordinales aplicando rangos (<i>etiqueta – valores del rango asociado</i>):</p> <ul style="list-style-type: none"> • 'menor_3k': [0, 2999], • '3k_a_5k': [3000, 4999], • '5k_a_10k': [5000, 9999], • 'mayor_10k': [10000, 100000](*) <p>(*) Se toma como valor superior del rango a uno que incluye todo el conjunto de datos disponible según el análisis de metadatos.</p>
"gastos_ult_12m"	<p>Valores numéricos convertidos a ordinales aplicando rangos (<i>etiqueta – valores del rango asociado</i>):</p> <ul style="list-style-type: none"> • 'menor_1k': [0, 999], • '2k_a_4k': [1000, 3999], • '4k_a_6k': [4000, 5999], • '6k_a_8k': [6000, 7999], • '8k_a_10k': [8000, 9999], • 'mayor_10k': [10000, 100000](*) <p>(*) Se toma como valor superior del rango a uno que incluye todo el conjunto de datos disponible según el análisis de metadatos.</p>
"operaciones_ult_12m"	<p>Valores numéricos convertidos a ordinales aplicando rangos (<i>etiqueta – valores del rango asociado</i>):</p> <ul style="list-style-type: none"> • 'menor_15': [0, 14], • '15_a_30': [15, 29], • '30_a_50': [30, 49], • '50_a_75': [50, 74], • '75_a_100': [75, 99], • 'mayor_100': [100, 1000] (*) <p>(*) Se toma como valor superior del rango a uno que incluye todo el conjunto de datos disponible según el análisis de metadatos.</p>

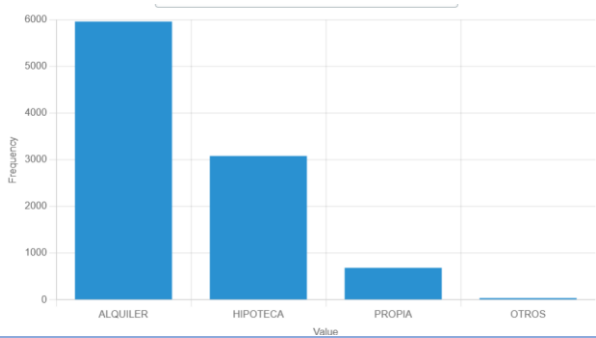
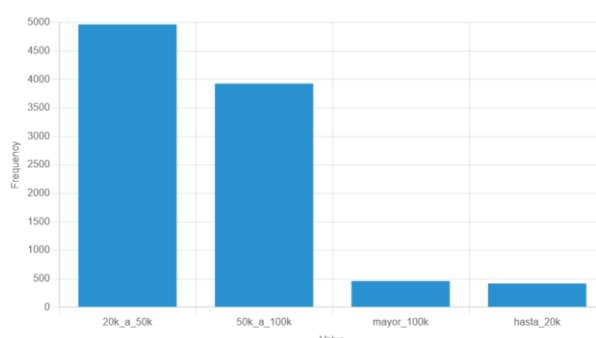
Formateo de los datos

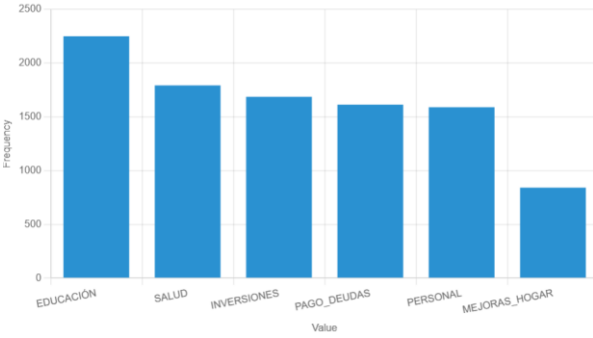
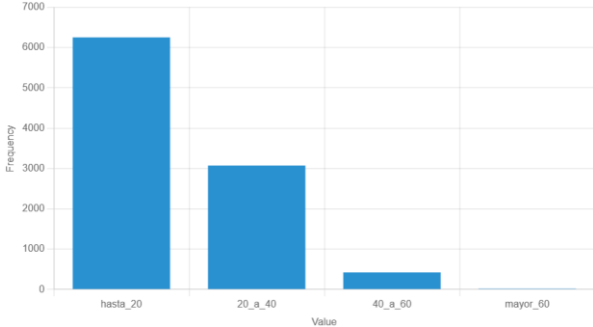
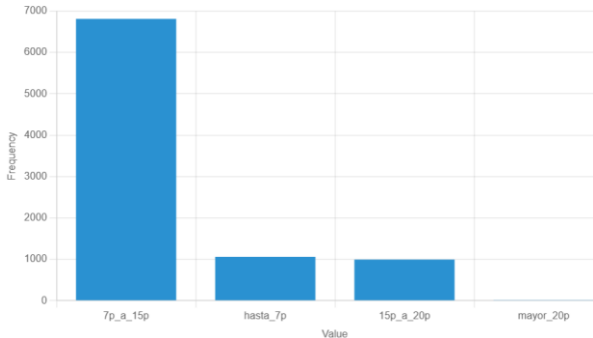
Se han especificado las operaciones del apartado anterior donde se han modificado cuestiones de formato sobre algunos atributos, pasando de un valor numérico a uno nominal basado en rangos. Estas modificaciones han cambiado los metadatos del dataset final a utilizar, por lo que se procede a documentarlos a continuación:

- Cantidad de columnas: 20
 - Se aclara que se ha eliminado el atributo “**id_cliente**” al haberse utilizado para realizar la integración de los datasets originales y no ser requerido en los pasos posteriores del proyecto.
- Cantidad de filas: (*se mantiene sin cambios sobre lo reportado al momento de la integración*)
- Dataset: datos_creditos.csv

Columnas / Atributos	Tipo de datos	Observaciones
edad	Category	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</p> 
importe_solicitado	Int64 (numérico)	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores</i></p> <ul style="list-style-type: none"> • max:35,000 • mean: 8,180.3199 • median:6,500 • min:500 • mode:5,000

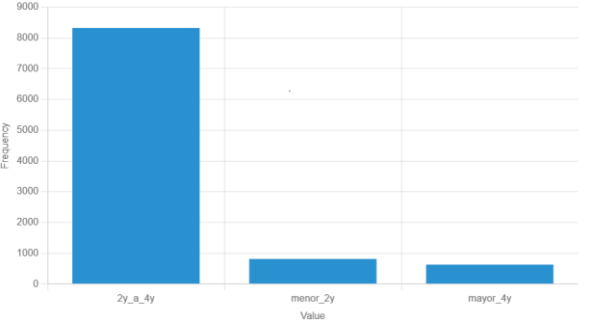
		
duracion_credito	Int64 (numérico)	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores</i></p> <ul style="list-style-type: none"> • max:4 • mean:2.9956 • median:3 • min:2 • mode:2 
antigüedad_emplead o	Category	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</p> 
situacion_vivienda	String (nominal)	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0</p>

		<p>% Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>):</p> <ul style="list-style-type: none"> ● ALQUILER - 5,963 - 61.03% ● HIPOTECA - 3,083 - 31.56% ● PROPIA – 686 - 7.02% ● OTROS – 38 - 0.39% 
ingresos	Category	<p>Total Rows:9,770</p> <p>Count (non-nan):9,770</p> <p>Count (missing):0</p> <p>% Missing:0</p> 
objetivo_credito	String (nominal)	<p>Total Rows:9,770</p> <p>Count (non-nan):9,770</p> <p>Count (missing):0</p> <p>% Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>):</p> <ul style="list-style-type: none"> ● EDUCACIÓN - 2,249 - 23.02% ● SALUD - 1,792 - 18.34% ● INVERSIONES - 1,686 - 17.26% ● PAGO_DEUDAS - 1,613 - 16.51% ● PERSONAL - 1,589 - 16.26% ● MEJORAS_HOGAR – 841- 8.61%

		
pct_ingreso	Category	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</p> 
tasa_interes	Category	<p>Total Rows:9,770 Count (non-nan):8,883 Count (missing):887 % Missing:9.08</p> 
estado_credito_N	String (nominal)	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>):</p> <ul style="list-style-type: none"> ● C - 7,396 - 75.70% ● P - 2,374 - 24.30%

		
falta_pago	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>):</p> <ul style="list-style-type: none"> ● N - 8,043 - 82.32% ● Y - 1,727 - 17.68% 

- Dataset: datos_tarjetas.csv

Columnas / Atributos	Tipo de datos	Observaciones
antigüedad_cliente	Category	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</p> 
estado_civil_N	String (nominal)	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0</p>

		<div>% Missing:0</div> <table><tr><td>CASADO</td><td>4,525</td><td>46.32%</td></tr><tr><td>SOLTERO</td><td>3,809</td><td>38.99%</td></tr><tr><td>DESCONOCIDO</td><td>718</td><td>7.35%</td></tr><tr><td>DIVORCIADO</td><td>718</td><td>7.35%</td></tr><tr><td>TOTAL</td><td>10,127</td><td>100.00%</td></tr></table>	CASADO	4,525	46.32%	SOLTERO	3,809	38.99%	DESCONOCIDO	718	7.35%	DIVORCIADO	718	7.35%	TOTAL	10,127	100.00%
CASADO	4,525	46.32%															
SOLTERO	3,809	38.99%															
DESCONOCIDO	718	7.35%															
DIVORCIADO	718	7.35%															
TOTAL	10,127	100.00%															
estado_cliente	String (nominal)	<div>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</div>  <table><tr><th>Value</th><th>Frequency</th></tr><tr><td>ACTIVO</td><td>8200</td></tr><tr><td>PASIVO</td><td>1570</td></tr></table>	Value	Frequency	ACTIVO	8200	PASIVO	1570									
Value	Frequency																
ACTIVO	8200																
PASIVO	1570																
gastos_ult_12m	Category	<div>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</div>  <table><tr><th>Value</th><th>Frequency</th></tr><tr><td>2k_a_4k</td><td>4800</td></tr><tr><td>4k_a_6k</td><td>3300</td></tr><tr><td>mayor_10k</td><td>700</td></tr><tr><td>6k_a_8k</td><td>300</td></tr><tr><td>8k_a_10k</td><td>250</td></tr><tr><td>menor_1k</td><td>100</td></tr></table>	Value	Frequency	2k_a_4k	4800	4k_a_6k	3300	mayor_10k	700	6k_a_8k	300	8k_a_10k	250	menor_1k	100	
Value	Frequency																
2k_a_4k	4800																
4k_a_6k	3300																
mayor_10k	700																
6k_a_8k	300																
8k_a_10k	250																
menor_1k	100																
genero	String (nominal)	<div>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</div>															

		 <p>Frequency</p> <p>Value</p> <p>F M</p>
limite_credito_tc	Category	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</p>  <p>Frequency</p> <p>Value</p> <p>menor_3k mayor_10k 5k_a_10k 3k_a_5k</p>
nivel_educativo	String (nominal)	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</p>  <p>Frequency</p> <p>Value</p> <p>UNIVERSITARIO_COMPLETO UNIVERSITARIO_INCOMPLETO SECUNDARIO_COMPLETO DESCONOCIDO POSGRADO_INCOMPLETO POSGRADO_COMPLETO</p>
operaciones_ult_12m	Category	<p>Total Rows:9,770 Count (non-nan):9,770 Count (missing):0 % Missing:0</p>



A partir de la ejecución de estas actividades, se da por **finalizado el primer sprint del proyecto**. Se adjuntan a continuación capturas de pantalla obtenidas desde la herramienta de seguimiento de incidencias utilizada (Jira) que registran lo mencionado:

Proyectos / Metodologías de gestión y diseño de proyectos Big Data

Backlog

CT [Avatar] [Avatar] [Avatar] Epic ▾












0 49 Completar sprint ...

Realizar las acciones correspondiente a las fases de comprensión de los datos y preparacion de los datos.





ID	Título	COMPRESIÓN...	FINALIZADA ▾	Estimación	Asignado a
SCRUM-7	Determinar los objetivos de la Organización	COMPRESIÓN...	FINALIZADA ▾	2	CT
SCRUM-8	Evaluación de la situación	COMPRESIÓN...	FINALIZADA ▾	3	[Avatar]
SCRUM-9	Determinación de los objetivos del proyecto en términos de productos de ciencia de d...	COMPRESIÓN...	FINALIZADA ▾	3	CT
SCRUM-10	Definir plan del proyecto	COMPRESIÓN...	FINALIZADA ▾	3	[Avatar]
SCRUM-11	Recolección de datos iniciales	COMPRESIÓN...	FINALIZADA ▾	4	CT
SCRUM-12	Descripción de los datos	COMPRESIÓN...	FINALIZADA ▾	5	[Avatar]
SCRUM-13	Exploración de datos	COMPRESIÓN...	FINALIZADA ▾	8	CT
SCRUM-14	Verificación de la calidad de los datos	COMPRESIÓN...	FINALIZADA ▾	5	[Avatar]
SCRUM-15	Selección de datos	FASE DE PREP...	FINALIZADA ▾	3	CT

Proyectos / Metodologías de gestión y diseño de proyectos Big Data

Backlog

<input type="text"/>	   	Epic ▾	  
SCRUM-16 Limpieza de los datos	FASE DE PREP...	FINALIZADA ▾	3 
SCRUM-17 Construcción de datos	FASE DE PREP...	FINALIZADA ▾	5 
SCRUM-18 Integración de los datos	FASE DE PREP...	FINALIZADA ▾	2 
SCRUM-19 Formateo de los datos	FASE DE PREP...	FINALIZADA ▾	3 
+ Crear incidencia			






colbert6.atlassian.net/jira/software/projects/SCRUM/boards/1/backlog

Tu trabajo ▾ Proyectos ▾ Filtros ▾ Paneles ▾ Equipos ▾ Más ▾ **Crear**  Mejorar   

Backlog

Completar Sprint 1

Este sprint contiene 13 incidencias en estado completado.
Ya está todo listo. ¡Bien hecho!

SCRUM-10 Definir plan del proyecto	COMPRESIÓN...	FINALIZADA ▾	3 
SCRUM-11 Recolección de datos iniciales	COMPRESIÓN...	FINALIZADA ▾	4 
SCRUM-12 Descripción de los datos	COMPRESIÓN...	FINALIZADA ▾	5 
SCRUM-13 Exploración de datos	COMPRESIÓN...	FINALIZADA ▾	8 
SCRUM-14 Verificación de la calidad de los datos	COMPRESIÓN...	FINALIZADA ▾	5 
SCRUM-15 Selección de datos	FASE DE PREP...	FINALIZADA ▾	3 