



Proyecto Final: Detección de Fraude en Tarjetas de Crédito

Materia: Estrategias empresariales basadas en Ciencia de Datos

Profesor: Dra. Dafne Angelica Rosso Pelayo

Nombres: Brian Antonio Aranda Mejía **ID:** 0209486

Carlos Cabrera Castrejón **ID:** 0217201

Esteban Mayen Soto **ID:** 0217201

Fecha de entrega: 21 de Octubre de 2023

Universidad Panamericana

Facultad de Ingeniería

Entendimiento del negocio

Determinar los Objetivos del negocio

Antecedentes

Este conjunto de datos de transacciones con tarjetas de crédito abarca un período de dos años y ha sido creado utilizando una herramienta desarrollada por Brandon Harris. El mismo comprende tanto transacciones legítimas como fraudulentas y se ha generado mediante la utilización de perfiles que definen las características de los clientes y las transacciones en cuestión.

Objetivos del negocio

1. **Detección y Prevención de Fraude:** Uno de los objetivos principales podría ser desarrollar sistemas y algoritmos efectivos para detectar y prevenir transacciones fraudulentas. Esto implica crear modelos de aprendizaje automático que puedan identificar patrones sospechosos de actividad y alertar a los equipos de seguridad para tomar medidas adecuadas.
2. **Optimización de la Experiencia del Cliente:** El negocio podría tener como objetivo garantizar que las transacciones legítimas se procesen sin problemas y que los clientes legítimos no se vean afectados por medidas excesivas de seguridad. Esto podría implicar el desarrollo de algoritmos que diferencien entre transacciones legítimas y fraudulentas, mejorando así la experiencia del cliente.
3. **Mejora de la Seguridad de las Transacciones:** Los objetivos podrían incluir la mejora constante de las medidas de seguridad de las transacciones. Esto podría incluir el monitoreo en tiempo real, la identificación temprana de posibles amenazas y la implementación de nuevas tecnologías para proteger las transacciones y los datos de los clientes.
4. **Desarrollo de Capacidades Técnicas:** El negocio podría aspirar a mejorar sus habilidades técnicas y conocimientos en el ámbito de la seguridad cibernética y la detección de fraude. Esto podría incluir la capacitación del personal en técnicas avanzadas de análisis de datos, aprendizaje automático y seguridad informática.

Criterios de Éxito

1. **Tasa de Detección de Fraude:** Un criterio fundamental sería la capacidad del sistema para detectar transacciones fraudulentas con alta precisión. La tasa de detección de fraude, es decir, la

proporción de transacciones fraudulentas identificadas correctamente en comparación con el total de transacciones fraudulentas, sería un indicador clave de éxito.

2. **Tasa de Falsos Positivos:** Además de la detección precisa de fraudes, es importante minimizar los falsos positivos, es decir, transacciones legítimas que son identificadas erróneamente como fraudulentas. Una baja tasa de falsos positivos asegura que los clientes legítimos no se vean afectados negativamente por medidas de seguridad excesivas.
3. **Experiencia del Cliente:** La satisfacción del cliente es un criterio esencial. La implementación de medidas de seguridad no debe afectar negativamente la experiencia del cliente. Un criterio de éxito podría ser mantener una experiencia de usuario fluida y sin problemas, incluso con las medidas de seguridad en su lugar.
4. **Rendimiento del Sistema:** El tiempo de respuesta del sistema y su capacidad para manejar la carga de transacciones son aspectos cruciales. Un sistema eficiente y rápido garantiza que las transacciones legítimas y sospechosas se procesen sin demoras innecesarias.
5. **Reducción de Pérdidas por Fraude:** Una métrica financiera importante sería la reducción efectiva de pérdidas debido a transacciones fraudulentas. Cuanto más exitoso sea el sistema en la detección temprana de fraudes, menor será el impacto financiero en el negocio.
6. **Cumplimiento Normativo:** Si el negocio logra cumplir con las regulaciones y estándares de seguridad pertinentes, esto se consideraría un criterio de éxito importante para garantizar la integridad y la confianza en el manejo de los datos de los clientes.
7. **Capacidad de Adaptación:** La capacidad del sistema y el equipo para adaptarse a nuevas amenazas y patrones de fraude emergentes también es un criterio de éxito. La detección temprana y la rápida respuesta a las nuevas tácticas de fraude son fundamentales.
8. **Reducción de Incidentes Críticos:** Si el negocio logra reducir o eliminar incidentes graves de fraude que podrían dañar la reputación y la confianza de los clientes, esto también sería un criterio de éxito significativo.

Situación Actual

En la detección de fraude se observa un aumento considerable en los fraudes relacionados a los pagos digitales, por lo cual se ha generado inquietud tanto para consumidores como para los vastos comercios del

mundo. A nivel mundial Brasil ocupa el primer lugar en detección de fraudes seguido de México, el impacto que esto representa para las empresas llega a ser 3.8 veces mayor que el valor de transacción original, con esto podemos ver el gran desafío al que nos enfrentados con el poder reducir los fraudes tanto para los pagos digitales que son un problema que sigue creciendo y creciendo.

Inventario de recursos

1. Dataset recopilando información de transacciones fraudulentas
2. Google Colaboratory para tratamiento y exploración de los datos que provengan del dataset, junto a esto utilizaremos ciertos modelos como lo son regresión logística, árboles de decisión o tal vez el uso de redes neuronales.

Requerimientos

1. Debemos de tener un dataset el cual proporcione información de movimientos fraudulentos y no fraudulentos para que de esta manera podamos entrenar modelos que funcionen en la vida real con situaciones reales
2. El tipo de tecnología que vamos a utilizar podemos sacarle provecho a Google Colab para hacer el tratamiento y exploración dentro de los datos.
3. Selección de modelos para la detección
4. Las evaluaciones y métricas
5. Implementación

Supuestos, restricciones, riesgos y contingencias

Supuestos:

1. **Calidad de los Datos:** Se asume que los datos utilizados en la simulación y en los sistemas de detección de fraudes son precisos y confiables para garantizar la efectividad de los algoritmos.
2. **Efectividad del Modelo:** Se supone que los algoritmos y modelos de detección de fraudes son efectivos para identificar patrones y anomalías en las transacciones, basados en los perfiles de simulación.
3. **Seguridad de los Sistemas:** Se asume que se han implementado medidas de seguridad sólidas para proteger los sistemas de detección de fraudes y los datos sensibles de los clientes.

Restricciones:

1. **Recursos Financieros:** Las restricciones presupuestarias pueden limitar la inversión en tecnologías avanzadas, personal especializado y capacitación.
2. **Capacidad de Procesamiento:** La capacidad de procesamiento de las infraestructuras tecnológicas podría limitar la rapidez y la cantidad de transacciones que se pueden analizar en tiempo real.
3. **Regulaciones y Cumplimiento:** Restricciones legales y normativas pueden afectar la forma en que se recopilan, almacenan y utilizan los datos de los clientes.

Riesgos:

1. **Falsos Positivos y Negativos:** Los sistemas de detección podrían generar falsos positivos que molesten a los clientes legítimos o falsos negativos que permitan actividades fraudulentas.
2. **Robo de Datos:** Existe el riesgo de que los datos de los clientes sean robados por ciberdelincuentes si no se implementan medidas de seguridad adecuadas.
3. **Tecnología Obsoleta:** La rápida evolución de la tecnología podría llevar a la obsolescencia de sistemas y algoritmos, lo que podría resultar en una detección menos efectiva.
4. **Reacción a Nuevas Tácticas de Fraude:** Los estafadores pueden desarrollar nuevas tácticas y patrones de fraude que los sistemas actuales no sean capaces de detectar.

Contingencias:

1. **Respuesta a Falsos Positivos:** Implementar un proceso de revisión humana para verificar transacciones identificadas como sospechosas antes de tomar medidas drásticas.
2. **Actualización de Tecnología:** Establecer un plan para la actualización y mejora continua de los sistemas de detección y la infraestructura tecnológica.
3. **Formación Continua:** Proporcionar capacitación continua al personal para asegurarse de que estén al tanto de las últimas tendencias en detección de fraudes.
4. **Monitoreo Constante:** Establecer un equipo de monitoreo en tiempo real para identificar anomalías y patrones de fraude emergentes.
5. **Comunicación con Clientes:** Preparar un plan de comunicación claro para informar a los clientes sobre medidas de seguridad y posibles interrupciones del servicio debido a actividades de detección de fraudes.

Terminología

1. **Transacciones Legítimas:** Transacciones financieras realizadas por clientes que siguen las normas y regulaciones, sin intenciones fraudulentas.
2. **Transacciones Fraudulentas:** Transacciones financieras engañosas o ilegales realizadas con la intención de obtener ganancias ilícitas o causar daño.
3. **Detección de Fraude:** Proceso de identificar y señalar transacciones que muestran patrones o características sospechosas indicativas de fraude.
4. **Falsos Positivos:** Transacciones legítimas que son incorrectamente identificadas como fraudulentas por el sistema de detección, lo que puede resultar en molestias para los clientes.
5. **Falsos Negativos:** Transacciones fraudulentas que no son detectadas por el sistema de detección, lo que puede resultar en pérdidas financieras para el negocio.
6. **Perfil de Simulación:** Características específicas utilizadas en la simulación para definir a un grupo de clientes, como edad, género, ubicación, etc.
7. **Aprendizaje Automático:** Uso de algoritmos y modelos para analizar datos y aprender patrones, permitiendo la detección automática de anomalías y comportamientos sospechosos.
8. **Tasa de Detección:** Porcentaje de transacciones fraudulentas que son identificadas correctamente por el sistema de detección.
9. **Tasa de Falsos Positivos:** Porcentaje de transacciones legítimas que son erróneamente identificadas como fraudulentas por el sistema de detección.
10. **Carga de Transacciones:** La cantidad total de transacciones que el sistema debe manejar en un período de tiempo específico.
11. **Algoritmos de Detección de Fraude:** Conjunto de reglas y patrones utilizados para identificar transacciones fraudulentas basadas en el análisis de datos.
12. **Regulaciones y Cumplimiento:** Normas y estándares legales que deben cumplirse para garantizar la protección de los datos de los clientes y la seguridad de las transacciones financieras.
13. **Modelos de Distribución:** Representaciones matemáticas de cómo se distribuyen ciertos parámetros, como la cantidad de transacciones por día o los montos de transacciones.
14. **Seguridad Cibernética:** Protección de los sistemas informáticos y la información de los ataques cibernéticos y otras amenazas de seguridad.

Costos y beneficios

Costos:

1. **Desarrollo de Algoritmos y Modelos:** Inversión en tiempo y recursos para desarrollar algoritmos y modelos de detección de fraudes, que incluye investigaciones, pruebas y ajustes continuos.
2. **Tecnología y Software:** Adquisición de tecnologías y software avanzados para el análisis de datos, aprendizaje automático y detección de patrones anómalos.
3. **Personal Especializado:** Contratación y capacitación de expertos en seguridad cibernética, análisis de datos y aprendizaje automático para gestionar y optimizar los sistemas de detección.
4. **Actualizaciones Continuas:** Costos recurrentes para mantener los sistemas actualizados con las últimas tendencias en técnicas de fraude y seguridad informática.
5. **Capacitación y Formación:** Costos asociados a la capacitación del personal en nuevas técnicas y tecnologías de detección de fraudes.
6. **Infraestructura Tecnológica:** Inversión en infraestructura de TI robusta y segura para manejar grandes volúmenes de transacciones y datos.
7. **Cumplimiento Normativo:** Costos relacionados con el cumplimiento de regulaciones y estándares de seguridad en el manejo de datos financieros.

Beneficios:

1. **Reducción de Pérdidas por Fraude:** La detección efectiva de transacciones fraudulentas conlleva a una disminución de las pérdidas financieras ocasionadas por actividades fraudulentas.
2. **Mejora de la Reputación:** La implementación exitosa de medidas de seguridad puede mejorar la percepción de los clientes sobre la confiabilidad del negocio.
3. **Protección del Cliente:** La detección temprana de actividades fraudulentas protege a los clientes de posibles robos de identidad y fraudes.
4. **Eficiencia Operativa:** La automatización de la detección de fraudes permite una mayor eficiencia en la identificación y respuesta a actividades sospechosas.
5. **Cumplimiento Legal:** El cumplimiento de las regulaciones y estándares de seguridad evita sanciones legales y protege la integridad del negocio.

Determinar las metas de la minería de datos

Metas de la minería de datos

Meta: Detección Precisa de Fraudes

Objetivo SMART

1. **Específico:** Desarrollar un algoritmo de detección de fraudes que logre identificar al menos el 95% de las transacciones fraudulentas entre un conjunto de datos de 1 millón de transacciones simuladas y mantener la tasa de falsos positivos por debajo del 1%.
2. **Medible:** Mediremos el éxito del objetivo mediante la tasa de detección de fraudes y la tasa de falsos positivos obtenidos después de implementar el algoritmo.
3. **Alcanzable:** Aunque somos un equipo de estudiantes de posgrado finalizando la última etapa del diplomado en ciencia de datos, contamos con la supervisión de nuestra profesora. Además, tenemos acceso a datos de transacciones simuladas y tecnologías de aprendizaje automático. A pesar de nuestra condición de estudiantes, tenemos los recursos necesarios para llevar a cabo este proyecto.
4. **Relevante:** La detección precisa de fraudes es esencial para proteger los activos financieros de la empresa y la confianza de los clientes. Además, contribuirá a la reducción de pérdidas financieras y a la mejora de la reputación del negocio.
5. **Temporal:** El objetivo debe lograrse en un plazo de tres meses a partir de la fecha de inicio del proyecto. Esto permitirá medir el rendimiento del algoritmo durante tres meses en un entorno de producción y realizar ajustes si es necesario antes de la fecha límite. El objetivo tiene como fecha límite el 21 de Octubre del 2023

Criterio de éxito de la minería de datos

1. **Tasa de Detección de Fraudes:** Un alto porcentaje de transacciones fraudulentas identificadas correctamente en comparación con el total de transacciones fraudulentas. Por ejemplo, una tasa de detección del 95% sería un indicador sólido de éxito.
2. **Tasa de Falsos Positivos:** Mantener una baja tasa de transacciones legítimas que son incorrectamente identificadas

como fraudulentas. Por ejemplo, una tasa de falsos positivos del 1% o menos sería deseable.

3. **Reducción de Pérdidas Financieras:** Una disminución significativa en las pérdidas financieras causadas por actividades fraudulentas en comparación con el período anterior a la implementación de la detección de fraudes.
4. **Mejora de la Experiencia del Cliente:** Medir la satisfacción de los clientes y la eficiencia de las transacciones legítimas, asegurando que la implementación de medidas de detección no cause molestias excesivas.
5. **Análisis de Patrones de Fraude:** Identificar nuevos patrones y tácticas de fraude que antes no se habían detectado, demostrando la capacidad de adaptación y prevención anticipada..
6. **Cumplimiento Legal:** Cumplir con las regulaciones y estándares de seguridad relacionados con el manejo de datos financieros y la prevención de fraudes.
7. **Eficiencia Operativa:** Evaluar la capacidad del sistema para manejar grandes volúmenes de transacciones y procesarlas en tiempo real sin demoras significativas.
8. **Reducción de Incidentes Críticos:** Una disminución en la cantidad de incidentes graves de fraude que podrían dañar la reputación y la confianza del cliente.

Generar el plan de trabajo

Plan del proyecto inicial

Plan de Proyecto Inicial: Detección de Fraudes en Transacciones de Tarjetas de Crédito

Objetivo del Proyecto: Desarrollar un algoritmo de detección de fraudes que logre identificar al menos el 95% de las transacciones fraudulentas y mantenga la tasa de falsos positivos por debajo del 1% en un conjunto de datos de 1 millón de transacciones simuladas.

Equipo del Proyecto:

- Estudiantes de posgrado finalizando la última etapa del diplomado en ciencia de datos.
- Supervisión de la profesora de ciencia de datos.

Duración del Proyecto: 7 semanas (desde el 29 de agosto de 2023 hasta el 21 de octubre de 2023)

Fases del Proyecto:

1. Definición y Planificación (1 semana):

- Establecer los objetivos
- Identificar y obtener acceso a los conjuntos de datos de transacciones simuladas y perfiles de clientes.
- Planificar la metodología de desarrollo, incluyendo las técnicas de detección y las herramientas a utilizar.
- Definir los indicadores clave de rendimiento (KPIs) para medir el éxito del proyecto.
- Elaborar el cronograma detallado y asignar responsabilidades.

2. Exploración y Preparación de Datos (3 semanas):

- Explorar y analizar los datos de transacciones y perfiles de clientes para identificar patrones y características.
- Realizar limpieza de datos, manejo de valores faltantes y transformaciones necesarias.
- Dividir los datos en conjuntos de entrenamiento y prueba.

3. Desarrollo y Entrenamiento del Algoritmo (2 semanas):

- Aplicar técnicas de aprendizaje automático supervisado y no supervisado para desarrollar el algoritmo de detección de fraudes.
- Entrenar el modelo utilizando el conjunto de entrenamiento y ajustar hiper parámetros.
- Evaluar el rendimiento del modelo utilizando el conjunto de prueba y ajustar el algoritmo según sea necesario.
- Realizar pruebas de rendimiento intermedias y ajustes iterativos.

4. Documentación y Presentación (1 semana):

- Documentar el proceso de desarrollo, metodología utilizada y resultados obtenidos.
- Preparar una presentación final que resuma el proyecto, el algoritmo desarrollado y los logros alcanzados.
- Presentar los resultados y conclusiones ante la profesora y la clase.

Entregables:

- Algoritmo de detección de fraudes implementado y evaluado.
- Documentación detallada del proceso, metodología y resultados.
- Presentación.

	Responsable	Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7
1. Establecer los objetivos	Brian							
2. Identificar y obtener acceso a los conjuntos de datos de transacciones simuladas y perfiles de clientes.	Esteban							
3. Planificar la metodología de desarrollo, incluyendo las técnicas de detección y las herramientas a utilizar	Carlos							
4. Definir los indicadores clave de rendimiento (KPIs) para medir el éxito del proyecto	Brian							
5. Elaborar el cronograma detallado y asignar responsabilidades	Brian							
6. Explorar y analizar los datos de transacciones y perfiles de clientes para identificar patrones y características	Esteban							
7. Realizar limpieza de datos, manejo de valores faltantes y transformaciones necesarias	Carlos							
8. Dividir los datos en conjuntos de entrenamiento y prueba	Carlos							
9. Aplicar técnicas de aprendizaje automático supervisado y no supervisado para desarrollar el algoritmo de detección de fraudes	Esteban							
10. Entrenar el modelo utilizando el conjunto de entrenamiento y ajustar hiper parámetros	Carlos							
11. Evaluar el rendimiento del modelo utilizando el conjunto de prueba y ajustar el algoritmo según sea necesario	Esteban							
12. Realizar pruebas de rendimiento intermedias y ajustes iterativos	Brian							
13. Documentar el proceso de desarrollo, metodología utilizada y resultados obtenidos	Todos							
14. Preparar una presentación final que resuma el proyecto, el algoritmo desarrollado y los logros alcanzados	Todos							
15. Presentar los resultados y conclusiones ante la profesora y la clase	Todos							

Inventario de técnicas y herramientas

Técnicas de Detección de Fraudes:

1. **Análisis de Reglas:** Utilizar reglas predefinidas para identificar patrones de transacciones sospechosas basadas en criterios específicos.
2. **Aprendizaje Automático Supervisado:** Entrenar modelos con datos etiquetados (transacciones legítimas y fraudulentas) para que el sistema aprenda a reconocer patrones y detectar fraudes.
3. **Aprendizaje Automático No Supervisado:** Identificar anomalías en los datos sin etiquetas, lo que puede ser útil para detectar nuevas tácticas de fraude.
4. **Modelos de Clasificación:** Utilizar algoritmos de clasificación (como Regresión Logística, Máquinas de Soporte Vectorial) para categorizar las transacciones como legítimas o fraudulentas.
5. **Análisis de Series Temporales:** Identificar patrones de transacciones en función del tiempo, lo que puede ayudar a detectar tendencias y anomalías.

Herramientas de Análisis de Datos:

1. **Python:** Un lenguaje de programación ampliamente utilizado para análisis de datos y desarrollo de algoritmos de detección de fraudes.
2. **Pandas:** Una biblioteca de Python para el manejo y análisis de datos en forma de estructuras de datos llamadas DataFrames.
3. **Scikit-learn:** Biblioteca de aprendizaje automático de código abierto para Python que ofrece herramientas para clasificación, regresión y más.

Herramientas de Visualización:

1. **Matplotlib:** Biblioteca de trazado de gráficos en 2D para Python, que permite visualizar datos y resultados.
2. **Seaborn:** Una biblioteca de visualización de datos basada en Matplotlib que proporciona gráficos estadísticos atractivos.

Recopilación de datos iniciales

Los datos iniciales utilizados en este proyecto fueron recopilados a partir de un conjunto de datos de transacciones con tarjetas de crédito simuladas. Estas transacciones abarcan el período desde el 1 de enero de 2019 hasta el 31 de diciembre de 2020 y comprenden tanto transacciones legítimas como fraudulentas. La fuente de estos datos proviene de un conjunto público disponible en Kaggle, al que se puede acceder a través del enlace proporcionado: <https://www.kaggle.com/datasets/kartik2112/fraud-detection>

Contiene información esencial sobre las características de las transacciones, los perfiles de los clientes y los patrones de actividad que pueden ser utilizados para diseñar y afinar estrategias de detección efectivas.

	trans_date_trans_time	cc_num	merchant	category	amt	first	last	gender	street	city	state	zip	lat
0	01/01/2019 00:00	2.70319E+15	fraud_Rippin, Kub and Mann	misc_net	4.97	Jennifer	Banks	F	561 Perry Cove	Moravian Falls	NC	28654	36.0788
1	01/01/2019 00:00	6.30423E+11	fraud_Heller, Gutmann and Zieme	grocery_pos	107.23	Stephanie	Gill	F	43039 Riley Greens Suite 393	Orient	WA	99160	48.8878
2	01/01/2019 00:00	3.88595E+13	fraud_Lind-Buckridge	entertainment	220.11	Edward	Sanchez	M	594 White Dale Suite 530	Malad City	ID	83252	42.1808
3	01/01/2019 00:01	3.53409E+13	fraud_Kutch, Hermiston and Farrell	gas_transport	45	Jeremy	White	M	9443 Cynthia Court Apt. 038	Boulder	MT	59632	46.2306
4	01/01/2019 00:03	3.75534E+14	fraud_Keeling-Crist	misc_pos	41.96	Tyler	Garcia	M	408 Bradley Rest	Doe Hill	VA	24433	38.4207
5	01/01/2019 00:04	4.76727E+13	fraud_Stroman, Hudson and Erdman	gas_transport	94.63	Jennifer	Conner	F	4655 David Island	Dublin	PA	18917	40.375
6	01/01/2019 00:04	3.00747E+13	fraud_Rowe-Vandervort	grocery_net	44.54	Kelsey	Richards	F	889 Sarah Station Suite 624	Holcomb	KS	67851	37.9931
7	01/01/2019 00:05	6.01136E+15	fraud_Corwin-Collins	gas_transport	71.65	Steven	Williams	M	231 Flores Pass Suite 720	Edinburg	VA	22824	38.8432
8	01/01/2019 00:05	4.92271E+15	fraud_Herzog Ltd	misc_pos	4.27	Heather	Chase	F	6888 Hicks Stream Suite 954	Manor	PA	15665	40.3359
9	01/01/2019 00:06	2.72083E+15	fraud_Schoen, Kuphal and Nitzsche	grocery_pos	198.39	Melissa	Aguilar	F	21326 Taylor Squares Suite 708	Clarksville	TN	37040	36.522
10	01/01/2019 00:06	4.64289E+12	fraud_Rutherford-Mertz	grocery_pos	24.74	Eddie	Mendez	M	1831 Faith View Suite 653	Clarinda	IA	51632	40.7491
11	01/01/2019 00:06	3.77234E+14	fraud_Kerluke-Abshire	shopping_net	7.77	Theresa	Blackwell	F	43576 Kristina Islands	Shenandoah Junction	WV	25442	39.3716
12	01/01/2019 00:06	1.80043E+14	fraud_Lockman Ltd	grocery_pos	71.22	Charles	Robles	M	3337 Lisa Divide	Saint Petersburg	FL	33710	27.7898
13	01/01/2019 00:07	5.55986E+15	fraud_Kiehn Inc	grocery_pos	96.29	Jack	Hill	M	5916 Susan Bridge Apt. 939	Grenada	CA	96038	41.6125
14	01/01/2019 00:09	3.51487E+15	fraud_Beier-Hyatt	shopping_pos	7.77	Christopher	Castaneda	M	1632 Cohen Drive Suite 639	High Rolls Mountain Park	NM	88325	32.9396
15	01/01/2019 00:09	6.012E+15	fraud_Schmidt and Sons	shopping_net	3.26	Ronald	Carson	M	870 Rocha Drive	Harrington Park	NJ	7640	40.9918
16	01/01/2019 00:10	6.01186E+15	fraud_Lebsack and Sons	misc_net	327	Lisa	Mendez	F	44259 Beth Station Suite 215	Lahoma	OK	73754	36.385
17	01/01/2019 00:10	3.56542E+15	fraud_Mayert Group	shopping_pos	341.67	Nathan	Thomas	M	4923 Campbell Pines Suite 717	Carlisle	IN	47838	38.9763
18	01/01/2019 00:11	2.34825E+15	fraud_Konopelski, Schneider and Hartmann	food_dining	63.07	Justin	Gay	M	268 Hayes Rue Suite 811	Harborcreek	PA	16421	42.1767
19	01/01/2019 00:12	4.95683E+18	fraud_Schultz, Simonis and Little	grocery_pos	44.71	Kenneth	Robinson	M	269 Sanchez Rapids	Elizabeth	NJ	7208	40.6747
20	01/01/2019 00:13	4.46978E+18	fraud_Bauch-Raynor	grocery_pos	57.34	Gregory	Graham	M	4005 Dana Glens	Methuen	MA	1844	42.728
21	01/01/2019 00:14	2.30534E+15	fraud_Harris Inc	gas_transport	50.79	Jeffrey	Rice	M	21447 Powell Circle	Moulton	IA	52572	40.6866
22	01/01/2019 00:17	1.80048E+14	fraud_Kling-Grant	grocery_net	46.28	Mary	Wall	F	2481 Mills Lock	Plainfield	NJ	7060	40.6152
23	01/01/2019 00:17	6.30442E+11	fraud_Pacocho-Bauch	shopping_pos	9.55	Susan	Washington	F	759 Erin Mount Suite 956	May	TX	76857	31.9571
24	01/01/2019 00:18	4.42878E+18	fraud_Lesch Ltd	shopping_pos	22.95	Richard	Waters	M	7683 Natasha Way Apt. 945	Waukesha	WI	53186	42.9993

Descripción de los datos

1. trans_date_trans_time

- Nombre de la variable: trans_date_trans_time
- Nombre de la variable acordado: Marca de tiempo de la transacción
- Unidades de medida: Fecha y hora
- Valores permitidos: Registros de fecha y hora válidos
- Definición de la variable: Este atributo representa la fecha y hora en que se realizó la transacción.

2. cc_num

- Nombre de la variable: cc_num
- Nombre de la variable acordado: Número de la tarjeta de crédito utilizada en la transacción
- Unidades de medida: Número de tarjeta de crédito
- Valores permitidos: Números de tarjeta de crédito válidos

- Definición de la variable: Este atributo almacena el número de tarjeta de crédito utilizada en la transacción.

3. **merchant**

- Nombre de la variable: merchant
- Nombre de la variable acordado: Nombre del comerciante involucrado en la transacción
- Unidades de medida: Texto (nombre del comerciante)
- Valores permitidos: Nombres de comerciantes válidos
- Definición de la variable: Este atributo contiene el nombre del comerciante que participó en la transacción.

4. **category**

- Nombre de la variable: category
- Nombre de la variable acordado: Categoría del producto o servicio relacionado con la transacción
- Unidades de medida: Texto (categoría)
- Valores permitidos: Categorías válidas de productos o servicios
- Definición de la variable: Este atributo describe la categoría del producto o servicio relacionado con la transacción.

5. **amt**

- Nombre de la variable: amt
- Nombre de la variable acordado: Monto de la transacción
- Unidades de medida: Moneda (por ejemplo, dólares)
- Valores permitidos: Montos válidos de transacción
- Definición de la variable: Este atributo representa el monto o valor de la transacción.

6. **first**

- Nombre de la variable: first
- Nombre de la variable acordado: Nombre de pila del titular de la tarjeta
- Unidades de medida: Texto (nombre)
- Valores permitidos: Nombres de personas válidos
- Definición de la variable: Este atributo contiene el nombre de pila del titular de la tarjeta de crédito.

7. **last**

- Nombre de la variable: last
- Nombre de la variable acordado: Apellido del titular de la tarjeta
- Unidades de medida: Texto (apellido)
- Valores permitidos: Apellidos de personas válidos
- Definición de la variable: Este atributo almacena el apellido del titular de la tarjeta de crédito.

8. **gender**

- Nombre de la variable: gender
- Nombre de la variable acordado: Género del titular de la tarjeta
- Unidades de medida: Texto (género)
- Valores permitidos: Valores que representan género (por ejemplo, "M" o "F")
- Definición de la variable: Este atributo indica el género del titular de la tarjeta de crédito.

9. **street**

- Nombre de la variable: street
- Nombre de la variable acordado: Dirección de la ubicación de la transacción
- Unidades de medida: Texto (dirección)
- Valores permitidos: Direcciones válidas
- Definición de la variable: Este atributo contiene la dirección de la ubicación donde se realizó la transacción.

10. **city**

- Nombre de la variable: city
- Nombre de la variable acordado: Ciudad de la ubicación de la transacción
- Unidades de medida: Texto (ciudad)
- Valores permitidos: Nombres de ciudades válidas
- Definición de la variable: Este atributo representa la ciudad donde tuvo lugar la transacción.

11. **state**

- Nombre de la variable: state
- Nombre de la variable acordado: Estado o provincia de la ubicación de la transacción
- Unidades de medida: Texto (estado o provincia)
- Valores permitidos: Nombres de estados o provincias válidos
- Definición de la variable: Este atributo indica el estado o provincia donde ocurrió la transacción.

12. **zip**

- Nombre de la variable: zip
- Nombre de la variable acordado: Código postal de la ubicación de la transacción
- Unidades de medida: Número (código postal)
- Valores permitidos: Códigos postales válidos
- Definición de la variable: Este atributo almacena el código postal de la ubicación de la transacción.

13. **lat**

- Nombre de la variable: lat

- Nombre de la variable acordado: Latitud de la ubicación de la transacción
- Unidades de medida: Grados decimales (por ejemplo, 40.7128)
- Valores permitidos: Valores válidos de latitud
- Definición de la variable: Este atributo contiene la latitud geográfica de la ubicación de la transacción.

14. long

- Nombre de la variable: long
- Nombre de la variable acordado: Longitud de la ubicación de la transacción
- Unidades de medida: Grados decimales (por ejemplo, -74.0060)
- Valores permitidos: Valores válidos de longitud
- Definición de la variable: Este atributo almacena la longitud geográfica de la ubicación de la transacción.

15. city_pop

- Nombre de la variable: city_pop
- Nombre de la variable acordado: Población de la ciudad donde se realizó la transacción
- Unidades de medida: Número (población)
- Valores permitidos: Números válidos de población
- Definición de la variable: Este atributo representa la población de la ciudad donde se efectuó la transacción.

16. job

- Nombre de la variable: job
- Nombre de la variable acordado: Ocupación del titular de la tarjeta
- Unidades de medida: Texto (ocupación)
- Valores permitidos: Ocupaciones válidas
- Definición de la variable: Este atributo contiene la ocupación o profesión del titular de la tarjeta de crédito.

17. dob

- Nombre de la variable: dob
- Nombre de la variable acordado: Fecha de nacimiento del titular de la tarjeta
- Unidades de medida: Fecha
- Valores permitidos: Fechas de nacimiento válidas
- Definición de la variable: Este atributo indica la fecha de nacimiento del titular de la tarjeta de crédito.

18. trans_num

- Nombre de la variable: trans_num

- Nombre de la variable acordado: Número único de la transacción
- Unidades de medida: Texto (número único)
- Valores permitidos: Números únicos válidos
- Definición de la variable: Este atributo contiene un número único que identifica cada transacción de manera exclusiva.

19. **unix_time**

- Nombre de la variable: unix_time
- Nombre de la variable acordado: Marca de tiempo en formato Unix de la transacción
- Unidades de medida: Segundos desde la época Unix
- Valores permitidos: Marcas de tiempo Unix válidas
- Definición de la variable: Este atributo representa la marca de tiempo de la transacción en formato Unix, que cuenta los segundos desde la época Unix (1 de enero de 1970).

20. **merch_lat**

- Nombre de la variable: merch_lat
- Nombre de la variable acordado: Latitud de la ubicación del comerciante
- Unidades de medida: Grados decimales (por ejemplo, 40.7128)
- Valores permitidos: Valores válidos de latitud
- Definición de la variable: Este atributo almacena la latitud geográfica de la ubicación del comerciante involucrado en la transacción.

21. **merch_long**

- Nombre de la variable: merch_long
- Nombre de la variable acordado: Longitud de la ubicación del comerciante
- Unidades de medida: Grados decimales (por ejemplo, -74.0060)
- Valores permitidos: Valores válidos de longitud
- Definición de la variable: Este atributo contiene la longitud geográfica de la ubicación del comerciante involucrado en la transacción.

22. **is_fraud**

- Nombre de la variable: is_fraud
- Nombre de la variable acordado: Indicador binario de fraude
- Unidades de medida: 0 (legítima) o 1 (fraudulenta)
- Valores permitidos: 0 (legítima) o 1 (fraudulenta)
- Definición de la variable: Este atributo es un indicador binario que señala si la transacción es fraudulenta (1) o legítima (0).

Verificación de la calidad de los datos

El proceso de verificación de calidad de datos revela que el conjunto de datos está bien estructurado y no presenta valores nulos en ninguna de sus 23 columnas. Cada uno de los 1,296,675 registros contiene información válida y completa en todas las categorías.

Las columnas contienen una variedad de tipos de datos que reflejan adecuadamente la naturaleza de la información que representan, incluyendo números enteros, valores de punto flotante y cadenas de caracteres. La uniformidad y consistencia en la presentación de los datos en todas las filas y columnas sugieren que el conjunto de datos ha sido cuidadosamente preparado y está listo para ser utilizado en análisis posteriores, como la implementación de técnicas de detección de fraudes y modelado predictivo.

```
[ ] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1296675 entries, 0 to 1296674
Data columns (total 23 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0             1296675 non-null   int64
1   trans_date_trans_time  1296675 non-null   object
2   cc_num                 1296675 non-null   int64
3   merchant               1296675 non-null   object
4   category               1296675 non-null   object
5   amt                    1296675 non-null   float64
6   first                  1296675 non-null   object
7   last                   1296675 non-null   object
8   gender                 1296675 non-null   object
9   street                 1296675 non-null   object
10  city                   1296675 non-null   object
11  state                  1296675 non-null   object
12  zip                    1296675 non-null   int64
13  lat                    1296675 non-null   float64
14  long                   1296675 non-null   float64
15  city_pop               1296675 non-null   int64
16  job                    1296675 non-null   object
17  dob                    1296675 non-null   object
18  trans_num              1296675 non-null   object
19  unix_time              1296675 non-null   int64
20  merch_lat              1296675 non-null   float64
21  merch_long             1296675 non-null   float64
22  is_fraud               1296675 non-null   int64
dtypes: float64(5), int64(6), object(12)
memory usage: 227.5+ MB
```

Haciendo un análisis más detallado en la calidad de los datos, en este conjunto de datos se revela información valiosa sobre diversas características de las columnas.

La columna trans_date_trans_time presenta un total de 1,296,675 entradas, de las cuales existen 1,274,791 únicas.

La fecha y hora más común de transacciones es el 22 de abril de 2019 a las 16:02:01, que ocurre en 4 ocasiones.

El atributo amt muestra que la media del monto de transacción es de aproximadamente 70.35, con una desviación estándar de 160.32, lo que indica una variabilidad considerable en los montos.

En la columna gender, el valor más frecuente es "F" (femenino), con una frecuencia de 709,863, sugiriendo que la mayoría de las transacciones involucran a mujeres.

El atributo city_pop tiene un valor medio de alrededor de 88,824, lo que indica la población promedio de las ciudades involucradas en las transacciones.

En relación con la detección de fraudes, la columna is_fraud muestra que el 0.578% de las transacciones son fraudulentas, ya que tienen un valor de 1.

Esto nos ofrece una visión más profunda de las características de los datos y su distribución, lo que es esencial para el posterior procesamiento y análisis.

	trans_date_trans_time	cc_num	merchant	category	\
count	1296675	1.296675e+06	1296675	1296675	
unique	1274791	NaN	693	14	
top	2019-04-22 16:02:01	NaN	fraud_Kilback LLC	gas_transport	
freq	4	NaN	4403	131659	
mean	NaN	4.171920e+17	NaN	NaN	
std	NaN	1.308806e+18	NaN	NaN	
min	NaN	6.041621e+10	NaN	NaN	
25%	NaN	1.800429e+14	NaN	NaN	
50%	NaN	3.521417e+15	NaN	NaN	
75%	NaN	4.642255e+15	NaN	NaN	
max	NaN	4.992346e+18	NaN	NaN	

	amt	first	last	gender	\
count	1.296675e+06	1296675	1296675	1296675	
unique	NaN	352	481	2	
top	NaN	Christopher	Smith	F	
freq	NaN	26669	28794	709863	
mean	7.035104e+01	NaN	NaN	NaN	
std	1.603160e+02	NaN	NaN	NaN	
min	1.000000e+00	NaN	NaN	NaN	
25%	9.650000e+00	NaN	NaN	NaN	
50%	4.752000e+01	NaN	NaN	NaN	
75%	8.314000e+01	NaN	NaN	NaN	
max	2.894890e+04	NaN	NaN	NaN	

	street	city	...	lat	\
count	1296675	1296675	...	1.296675e+06	
unique	983	894	...	NaN	
top	0069 Robin Brooks Apt.	695 Birmingham	...	NaN	
freq	3123	5617	...	NaN	
mean	NaN	NaN	...	3.853762e+01	
std	NaN	NaN	...	5.075808e+00	
min	NaN	NaN	...	2.002710e+01	
25%	NaN	NaN	...	3.462050e+01	
50%	NaN	NaN	...	3.935430e+01	
75%	NaN	NaN	...	4.194040e+01	
max	NaN	NaN	...	6.669330e+01	

	long	city_pop	job	dob	\
count	1.296675e+06	1.296675e+06	1296675	1296675	
unique	NaN	NaN	494	968	
top	NaN	NaN	Film/video editor	1977-03-23	
freq	NaN	NaN	9779	5636	
mean	-9.022634e+01	8.882444e+04	NaN	NaN	
std	1.375908e+01	3.019564e+05	NaN	NaN	
min	-1.656723e+02	2.300000e+01	NaN	NaN	
25%	-9.679800e+01	7.430000e+02	NaN	NaN	
50%	-8.747690e+01	2.456000e+03	NaN	NaN	
75%	-8.015800e+01	2.032800e+04	NaN	NaN	
max	-6.795030e+01	2.906700e+06	NaN	NaN	

	trans_num	unix_time	merch_lat	\
count	1296675	1.296675e+06	1.296675e+06	
unique	1296675	NaN	NaN	
top	0b242abb623afc578575680df30655b9	NaN	NaN	
freq	1	NaN	NaN	
mean	NaN	1.349244e+09	3.853734e+01	
std	NaN	1.284128e+07	5.109788e+00	
min	NaN	1.325376e+09	1.902779e+01	
25%	NaN	1.338751e+09	3.473357e+01	
50%	NaN	1.349250e+09	3.936568e+01	
75%	NaN	1.359385e+09	4.195716e+01	
max	NaN	1.371817e+09	6.751027e+01	

	merch_long	is_fraud
count	1.296675e+06	1.296675e+06
unique	NaN	NaN
top	NaN	NaN
freq	NaN	NaN
mean	-9.022646e+01	5.788652e-03
std	1.377109e+01	7.586269e-02
min	-1.666712e+02	0.000000e+00
25%	-9.689728e+01	0.000000e+00
50%	-8.743839e+01	0.000000e+00
75%	-8.023680e+01	0.000000e+00
max	-6.695090e+01	1.000000e+00

Exploración de los datos

Durante la etapa de exploración de los datos, se observó que la columna "Unnamed: 0" contiene los mismos valores que el número de índice asignado por Python a cada fila. Como resultado, se tomó la decisión de eliminar esta columna del DataFrame.

Unnamed: 0	trans_date_trans_time	cc_num	merchant	category	amt	first	last	gender	street	...	lat	long	city_pop	job	dob
0	2019-01-01 00:00:18	2703186189652095	fraud_Ripkin, Kub and Mann	misc_net	4.97	Jennifer	Banks	F	561 Perry Cove	...	36.0788	-81.1781	3495	Psychologist, counselling	1988-03-09
1	2019-01-01 00:00:44	630423337322	fraud_Heller, Gutmann and Zieme	grocery_pos	107.23	Stephanie	Gill	F	43039 Riley Greens Suite 393	...	48.8878	-118.2105	149	Special educational needs teacher	1978-06-21
2	2019-01-01 00:00:51	38859492057661	fraud_Lind-Buckridge	entertainment	220.11	Edward	Sanchez	M	594 White Dale Suite 530	...	42.1808	-112.2620	4154	Nature conservation officer	1962-01-19
3	2019-01-01 00:01:16	3534093764340240	fraud_Kutch, Hermiston and Farrell	gas_transport	45.00	Jeremy	White	M	9443 Cynthia Court Apt. 038	...	46.2306	-112.1138	1939	Patent attorney	1967-01-12
4	2019-01-01 00:03:06	375534208663984	fraud_Keeling-Crist	misc_pos	41.96	Tyler	Garcia	M	408 Bradley Rest	...	38.4207	-79.4629	99	Dance movement psychotherapist	1986-03-28

```
df = df.drop(['Unnamed: 0'],axis=1)
df.head()
```

trans_date_trans_time	cc_num	merchant	category	amt	first	last	gender	street	city	...	lat	long	city_pop	job	dob
2019-01-01 00:00:18	2703186189652095	fraud_Ripkin, Kub and Mann	misc_net	4.97	Jennifer	Banks	F	561 Perry Cove	Moravian Falls	...	36.0788	-81.1781	3495	Psychologist, counselling	1988-03-09
2019-01-01 00:00:44	630423337322	fraud_Heller, Gutmann and Zieme	grocery_pos	107.23	Stephanie	Gill	F	43039 Riley Greens Suite 393	Orient	...	48.8878	-118.2105	149	Special educational needs teacher	1978-06-21
2019-01-01 00:00:51	38859492057661	fraud_Lind-Buckridge	entertainment	220.11	Edward	Sanchez	M	594 White Dale Suite 530	Malad City	...	42.1808	-112.2620	4154	Nature conservation officer	1962-01-19
2019-01-01 00:01:16	3534093764340240	fraud_Kutch, Hermiston and Farrell	gas_transport	45.00	Jeremy	White	M	9443 Cynthia Court Apt. 038	Boulder	...	46.2306	-112.1138	1939	Patent attorney	1967-01-12
2019-01-01 00:03:06	375534208663984	fraud_Keeling-Crist	misc_pos	41.96	Tyler	Garcia	M	408 Bradley Rest	Doe Hill	...	38.4207	-79.4629	99	Dance movement psychotherapist	1986-03-28

Posteriormente, utilizamos la librería `pandas_profiling` para crear un informe exhaustivo denominado "Detección de Fraude en Tarjetas de Crédito", que a lo largo de este documento, exploraremos las estadísticas y gráficas generadas por `pandas_profiling` para las variables que contiene nuestro set de datos, brindando una visión en profundidad de su distribución, singularidad, correlación y otras características clave.

Este análisis es esencial para tomar decisiones informadas y desarrollar estrategias efectivas en la lucha contra los fraudes en transacciones con tarjetas de crédito.

En la variable "trans_date_trans_time" se observa que existen 1,274,791 valores distintos en esta columna, lo que representa aproximadamente el 98.3% de la totalidad de los datos. Esto indica una alta variabilidad en las fechas y tiempos de transacción registrados. Además, no se encuentran

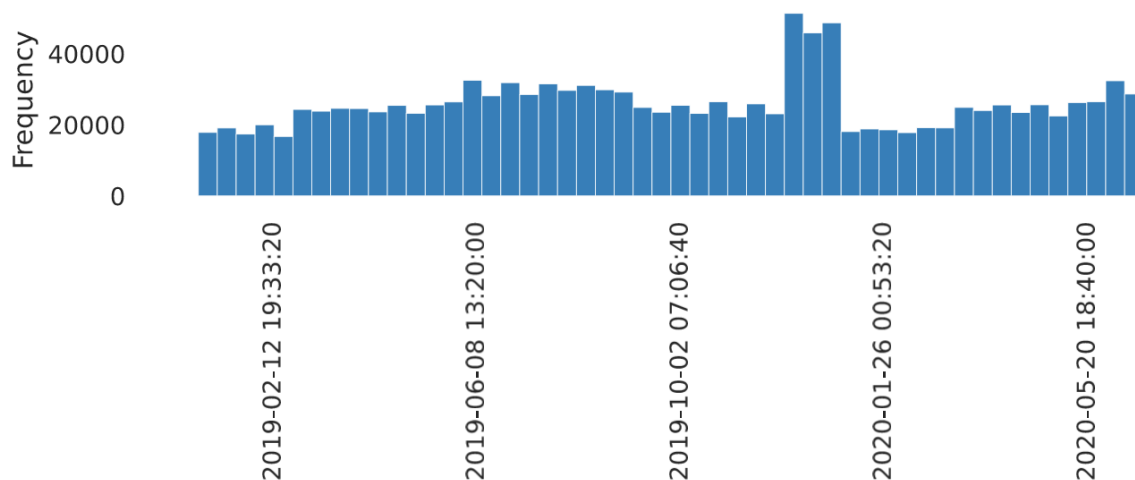
valores faltantes en esta columna, lo que sugiere que todos los registros tienen una fecha y hora asociada. El tamaño de memoria utilizado por esta columna es de 9.9 MiB, lo que es relevante para consideraciones de eficiencia en el almacenamiento de datos. Por último, se identifica que la fecha y hora mínima en esta columna es el 1 de enero de 2019 a las 00:00:18, mientras que la fecha y hora máxima es el 21 de junio de 2020 a las 12:13:37, lo que proporciona información sobre el período de tiempo abarcado por los datos.

trans_date_trans_time

Date

Distinct	1274791	Minimum	2019-01-01 00:00:18
Distinct (%)	98.3%	Maximum	2020-06-21 12:13:37
Missing	0		
Missing (%)	0.0%		
Memory size	9.9 MiB		

Histogram



Histogram with fixed size bins (bins=50)


En la variable "merchant" se identifican 693 valores distintos en esta columna, lo que representa un escaso 0.1% de variabilidad, lo que indica que la mayoría de los registros comparten los mismos valores de "merchant". No se encontraron valores faltantes en esta columna, lo que sugiere que todos los registros tienen información para esta variable. Además, el tamaño de memoria utilizado para almacenar esta columna es de 9.9 MiB, lo que es importante considerar para la gestión de recursos computacionales.

merchant
Text

Distinct	693
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

En la sección relacionada con las palabras más comunes en esta variable, se observa que las tres palabras que se repiten con mayor frecuencia son "and" con una cuenta de 474,111 veces, lo que representa un notable 15.7% del contenido total de la columna. Le sigue "lic" con una cuenta de 97,780 veces, lo que equivale al 3.2% de los datos, y "inc" con una cuenta de 91,939 veces, representando un 3.0%

Overview	Words	Characters
Value	Count	Frequency (%)
and	474111	15.7%
lic	97780	3.2%
inc	91939	3.0%
sons	73145	2.4%
ltd	70853	2.3%
plc	66475	2.2%
group	50447	1.7%
fraud_kutch	10560	0.3%
fraud_schaefer	9394	0.3%
fraud_streich	9250	0.3%
Other values (804)	2069403	68.4%



La variable "category" se identifican 14 categorías distintas en esta columna, lo que representa un modesto 0.1% de variabilidad en las categorías presentes, indicando que esta variable tiende a agrupar datos en un número limitado de categorías.

No se encontraron valores faltantes en esta columna, lo que demuestra que todos los registros tienen una categoría asignada. Además, el tamaño de memoria utilizado para almacenar esta columna es de 9.9 MiB, lo que puede ser relevante para la gestión de recursos en el análisis de datos.

category

Categorical

Distinct	14
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

En la sección relacionada con las categorías más comunes, se observa que las tres categorías con mayor frecuencia son "gas_transport" con una cuenta de 131,659 veces, lo que representa un sólido 10.2% del total de categorías. A continuación, "grocery_pos" cuenta con 123,638 registros, equivalente al 9.5%, y "home" con 123,115 registros, también un 9.5%. Estos resultados sugieren que la variable "category" se distribuye de manera desigual, con algunas categorías siendo significativamente más frecuentes que otras, lo que podría ser indicativo de patrones de gasto o comportamiento de compra en el conjunto de datos

Overview	Categories	Words	Characters
Common Values			
Value	Count	Frequency (%)	
gas_transport	131659	10.2%	
grocery_pos	123638	9.5%	
home	123115	9.5%	
shopping_pos	116672	9.0%	
kids_pets	113035	8.7%	
shopping_net	97543	7.5%	
entertainment	94014	7.3%	
food_dining	91461	7.1%	
personal_care	90758	7.0%	
health_fitness	85879	6.6%	
Other values (4)	228901	17.7%	

En la variable "amt" se identifican 52,928 valores distintos en esta columna, lo que representa un 4.1% de variabilidad en los montos registrados, lo que sugiere que existen diversas cantidades de transacciones en el conjunto de datos. No se encontraron valores faltantes ni infinitos, lo que garantiza la integridad de los datos.

amt

Real number (R)

SKEWED

Distinct	52928	Minimum	1
Distinct (%)	4.1%	Maximum	28948.9
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	70.351035	Memory size	9.9 MiB

En cuanto a las estadísticas descriptivas, se observa que el monto mínimo es de 1, mientras que el máximo alcanza los 28,948.9, lo que indica una amplia gama de valores en esta variable. No se encontraron valores iguales a cero ni valores negativos, lo que podría ser relevante para el análisis. El valor promedio (mean) de la variable es de aproximadamente 70.35, y la mediana (median) es de 47.52, lo que sugiere una posible asimetría en la distribución de los montos.

El rango (range) entre el valor mínimo y máximo es de 28,947.9, y el rango intercuartil (IQR) es de 73.49, lo que indica una dispersión sustancial en los datos. La desviación estándar (standard deviation) es de alrededor de 160.32, lo que sugiere una variabilidad significativa en los montos. El coeficiente de variación (CV) es de 2.28, lo que sugiere una moderada variabilidad relativa en comparación con la media.

Además, el valor de la curtosis (kurtosis) es de 4545.645, indicando una distribución leptocúrtica con colas pesadas, lo que puede sugerir una concentración de montos alrededor de la media con valores extremos infrecuentes. La suma (sum) de todos los montos en esta variable es de 91,222,429, y la varianza (variance) es de 25,701.232, lo que proporciona información sobre la dispersión y la agrupación de los datos.

Finalmente, se destaca que la variable "amt" no sigue una tendencia monótona. Estos resultados son esenciales para comprender la distribución, la dispersión y la variabilidad de los montos de transacción en el conjunto de datos, lo que es fundamental para análisis posteriores y la toma de decisiones basadas en datos.

Statistics	Histogram	Common values	Extreme values
Quantile statistics		Descriptive statistics	
Minimum	1	Standard deviation	160.31604
5-th percentile	2.44	Coefficient of variation (CV)	2.2788014
Q1	9.65	Kurtosis	4545.645
median	47.52	Mean	70.351035
Q3	83.14	Median Absolute Deviation (MAD)	37.5
95-th percentile	196.31	Skewness	42.277874
Maximum	28948.9	Sum	91222429
Range	28947.9	Variance	25701.232
Interquartile range (IQR)	73.49	Monotonicity	Not monotonic

En la variable "first" se identifican 351 valores distintos, lo que representa un porcentaje menor al 0.1% de variabilidad, indicando que esta variable tiende a contener una variedad limitada de nombres en los datos. Es importante destacar que no se encontraron valores faltantes en esta columna, lo que garantiza que todos los registros tienen un nombre en la variable "first". Además, el tamaño de memoria utilizado para almacenar esta columna es de 9.9 MiB, lo que puede ser relevante para la gestión de recursos en el análisis de datos.

first

Text

Distinct	352
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

En la sección relacionada con las palabras más comunes en esta variable, se observa que los tres nombres que se repiten con mayor frecuencia son "christopher" con una cuenta de 26,669 veces, lo que representa un 2.1% del total de nombres. Le sigue "robert" con una cuenta de 21,667 veces, equivalente al 1.7%, y "jessica" con 20,581 veces, también un 1.6%. Estos resultados indican que los nombres en la columna "first" están dominados por unos pocos nombres propios, lo que podría ser relevante para identificar tendencias o patrones en los datos.

Overview			Words	Characters
Value	Count	Frequency (%)		
christopher	26669	2.1%		
robert	21667	1.7%		
jessica	20581	1.6%		
james	20039	1.5%		
michael	20009	1.5%		
david	19965	1.5%		
jennifer	16940	1.3%		
william	16371	1.3%		
mary	16346	1.3%		
john	16325	1.3%		
Other values (342)	1101763	85.0%		



La variable "last" se identifican 481 valores distintos en esta columna, lo que representa un porcentaje menor al 0.1% de variabilidad en los apellidos registrados, indicando que esta variable tiende a contener una variedad limitada de apellidos en los datos. Es importante destacar que no se encontraron valores faltantes en esta columna, lo que garantiza que todos los registros tienen un apellido en la variable "last". Además, el tamaño de memoria utilizado para almacenar esta columna es de 9.9 MiB, lo que puede ser relevante para la gestión de recursos en el análisis de datos.

last


Text

Distinct	481
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

En la sección relacionada con las palabras más comunes en esta variable, se observa que los tres apellidos que se repiten con mayor frecuencia son "smith" con una cuenta de 28,794 veces, lo que representa un 2.2% del total de apellidos. Le sigue "williams" con una cuenta de 23,605 veces, equivalente al 1.8%, y "davis" con 21,910 veces, también un 1.7%. Estos

[Overview](#)
[Words](#)
[Characters](#)

Value	Count	Frequency (%)
smith	28794	2.2%
williams	23605	1.8%
davis	21910	1.7%
johnson	20034	1.5%
rodriguez	17394	1.3%
martinez	14805	1.1%
jones	13976	1.1%
lewis	12753	1.0%
gonzalez	11799	0.9%
millar	11698	0.9%
Other values (471)	1119907	86.4%

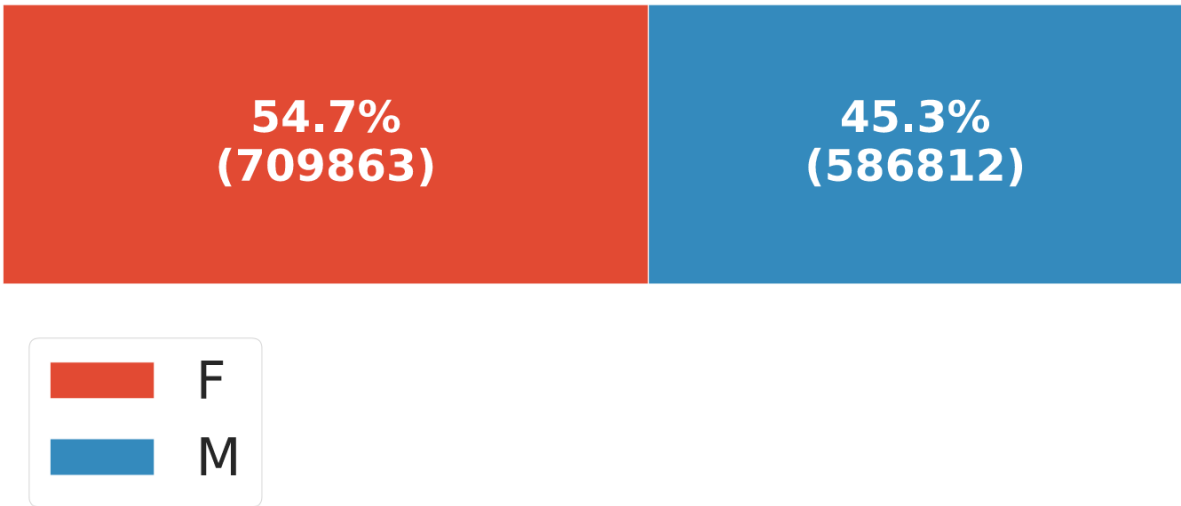


A word cloud visualization of the names from the dataset. The most prominent names are 'williams', 'smith', 'davis', 'johnson', 'rodriguez', 'martinez', 'jones', 'lewis', 'gonzalez', and 'millar', which correspond to the top entries in the table. The words are arranged in a circular pattern, with larger words being more central and smaller words towards the periphery. The colors of the words are varied, including shades of blue, green, yellow, and red.

No se encontraron valores faltantes en esta columna, lo que indica que todos los registros tienen información de género. Además, el tamaño de memoria utilizado para almacenar esta columna es de 9.9 MiB, lo que es relevante para consideraciones de eficiencia en el almacenamiento de datos.

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

El análisis gráfico de comparación proporciona una vista clara de la distribución de género en la base de datos. Se observa que hay más mujeres que hombres, con un total de 709,863 mujeres, lo que representa el 54.7% del total, en comparación con 586,812 hombres, que constituyen el 45.3%. Estos hallazgos sugieren que las mujeres están sobrerrepresentadas en la base de datos de detección de fraudes en comparación con los hombres, lo cual es información importante a tener en cuenta en el análisis y la toma de decisiones relacionadas con el fraude.



En la variable "street" se identifican 983 valores distintos en esta columna, lo que representa un modesto 0.1% de variabilidad en las direcciones registradas, indicando que esta variable tiende a contener una variedad limitada de calles en los datos. Es importante destacar que no se encontraron valores faltantes en esta columna, lo que garantiza que todas las observaciones tienen una dirección en la variable "street". Además, el tamaño de memoria utilizado para almacenar esta columna es de 9.9 MiB, lo que puede ser relevante para la gestión de recursos en el análisis de datos.

street

Text

Distinct	983
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

En la sección relacionada con las palabras más comunes en esta variable, se observa que las tres palabras que se repiten con mayor frecuencia son "apt" con una cuenta de 327,791 veces, lo que representa un 6.4% del total de palabras en las direcciones. A continuación, "suite" cuenta con 305,467 repeticiones, equivalente al 5.9%, y "island" con 22,954 repeticiones, es decir, el 0.4%. Estos resultados sugieren que ciertas palabras clave relacionadas con la descripción de unidades o ubicaciones específicas, como "apt" y "suite", son comunes en las direcciones registradas. La presencia de "island" en un pequeño porcentaje de direcciones podría indicar ubicaciones particulares o características geográficas relevantes.

[illegible]

En la variable "city" se identifican 894 valores distintos en esta columna, lo que representa un modesto 0.1% de variabilidad en las ciudades registradas, indicando que esta variable tiende a contener una variedad limitada de nombres de ciudades en los datos. Es importante destacar que no se encontraron valores faltantes en esta columna, lo que garantiza que todas las observaciones tienen una ciudad en la variable "city". Además, el tamaño de memoria utilizado para almacenar esta columna es de 9.9 MiB, lo que puede ser relevante para la gestión de recursos en el análisis de datos.

city
Text

Distinct	894
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

En la sección relacionada con las palabras más comunes en esta variable, se observa que las tres palabras que se repiten con mayor frecuencia en los nombres de las ciudades son "city" con una cuenta de 21,314 veces, lo que representa un 1.3% del total de palabras en las direcciones. Luego, "west" cuenta con 19,473 repeticiones, equivalente al 1.2%, y "north" con 14,425 repeticiones, es decir, el 0.9%. Estos resultados sugieren que ciertas palabras clave relacionadas con la ubicación geográfica, como "city," "west," y "north," son comunes en los nombres de las ciudades registradas. Estas palabras pueden proporcionar pistas sobre la orientación geográfica o la estructura de nomenclatura utilizada para describir las ciudades en el conjunto de datos.

Overview

Words

Characters

Value	Count	Frequency (%)
city	21314	1.3%
west	19473	1.2%
north	14425	0.9%
saint	14363	0.9%
falls	12794	0.8%
new	11842	0.7%
mount	11375	0.7%
lake	11249	0.7%
san	10260	0.6%
springs	8727	0.5%
Other values (918)	1482445	91.6%

A word cloud visualization of city names. The words are arranged in a circular pattern, with the most frequent words being the largest. The words include: saint, north, city, west, falls, new, mount, lake, san, springs, and many others. The colors are primarily blue, green, and yellow.



En la variable "state" se identifican 51 valores distintos en esta columna, lo que representa un porcentaje menor al 0.1% de variabilidad en los estados registrados, indicando que esta variable tiende a contener una cantidad limitada de estados en los datos. Es relevante destacar que no se encontraron valores faltantes en esta columna, lo que asegura que todas las observaciones tienen un estado en la variable "state". Además, el tamaño de memoria utilizado para almacenar esta columna es de 9.9 MiB, lo que puede ser relevante para la gestión de recursos en el análisis de datos.

state

Text

Distinct	51
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

En la sección relacionada con las palabras más comunes en esta variable, se observa que los tres estados que se repiten con mayor frecuencia son "tx" (Texas) con una cuenta de 94,876 veces, lo que representa un 7.3% del total de estados. Luego, "ny" (Nueva York) cuenta con 83,501 repeticiones, equivalente al 6.4%, y "pa" (Pensilvania) con 79,847 repeticiones, es decir, el 6.2%. Estos resultados indican que ciertos estados, identificados por sus códigos de dos letras, son mucho más frecuentes que otros en el conjunto de datos.

En la variable "zip" se identifican 970 valores distintos en esta columna, lo que representa un 0.1% de variabilidad en los códigos postales registrados, indicando que esta variable tiende a contener una diversidad limitada de códigos postales en los datos. El hecho de que se destaque en rojo con la etiqueta "High_correlation" sugiere la existencia de una correlación significativa entre los valores de esta variable y posiblemente otras en el conjunto de datos.

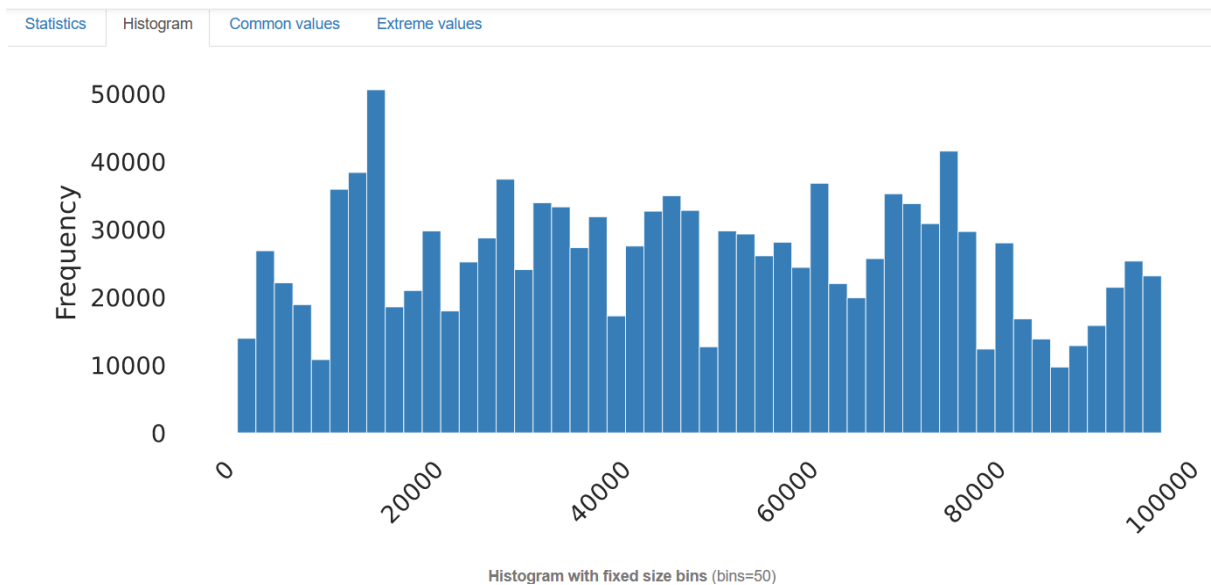
Real number (\mathbb{R})

Distinct	970
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	48800.671

No se encontraron valores faltantes ni infinitos en esta columna, lo que garantiza la integridad de los datos. El análisis estadístico revela que el valor mínimo es 1,257, mientras que el máximo es 99,783, con una amplia gama de valores entre ellos, lo que indica una variabilidad sustancial en los códigos postales. La mediana se sitúa en 48,174, lo que sugiere una distribución no sesgada hacia ningún extremo.

Statistics	Histogram	Common values	Extreme values
Quantile statistics		Descriptive statistics	
Minimum	1257	Standard deviation	26893.222
5-th percentile	7208	Coefficient of variation (CV)	0.55108305
Q1	26237	Kurtosis	-1.0964493
median	48174	Mean	48800.671
Q3	72042	Median Absolute Deviation (MAD)	23068
95-th percentile	94569	Skewness	0.079680758
Maximum	99783	Sum	6.327861×10^{10}
Range	98526	Variance	7.2324542×10^9
Interquartile range (IQR)	45805	Monotonicity	Not monotonic

La gráfica de barras muestra los tres códigos postales más comunes: 82514 con una cuenta de 1,589, que representa el 0.3% del total; 48088 con una cuenta de 1,518, equivalente al 0.3%; y 34112 con una cuenta de 1,495, también el 0.3%. Estos hallazgos son fundamentales para comprender la distribución de los códigos postales en el conjunto de datos y pueden ser útiles en análisis geográficos o para segmentar los datos según ubicaciones específicas. La alta correlación resaltada indica que esta variable podría estar relacionada con otras en el conjunto de datos, lo que puede tener implicaciones importantes en el análisis.



En la variable "lat" es importante destacar que esta variable se resalta en rojo con la etiqueta "High_correlation", lo que sugiere una correlación significativa con otras variables en el conjunto de datos.

Se identifican 968 valores distintos en esta columna, lo que representa un 0.1% de variabilidad en las latitudes registradas, indicando que esta variable tiende a contener una diversidad limitada de latitudes en los datos. No se encontraron valores faltantes ni infinitos en esta columna, lo que asegura la integridad de los datos.

lat

Real number (R)

HIGH CORRELATION

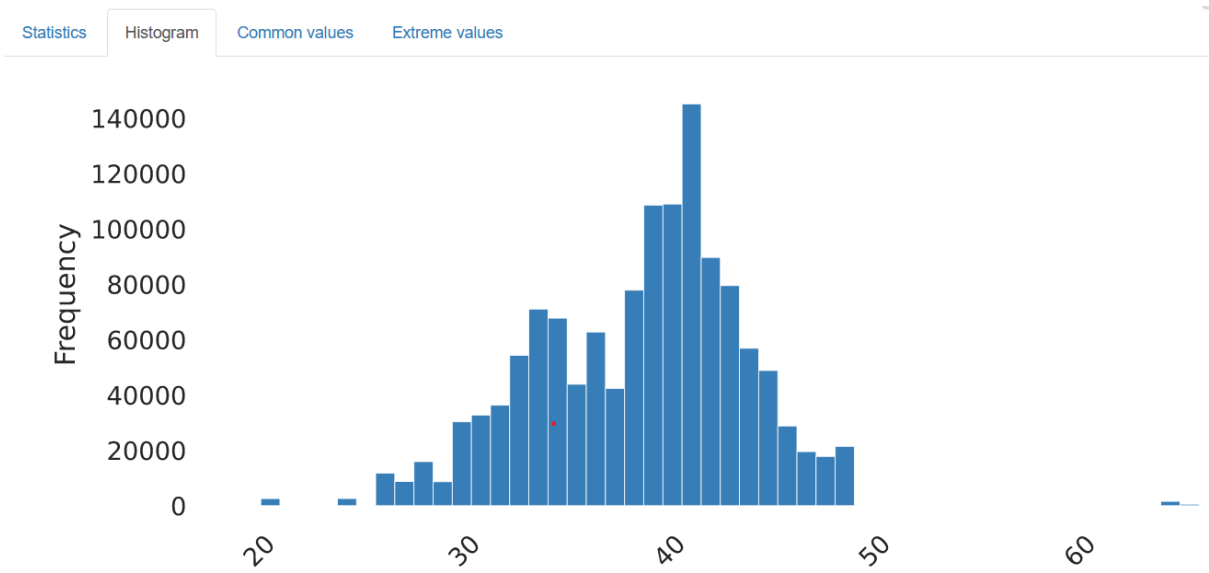
Distinct	968	Minimum	20.0271
Distinct (%)	0.1%	Maximum	66.6933
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	38.537622	Memory size	9.9 MiB

El análisis estadístico revela que el valor mínimo es 20.0271, mientras que el máximo es 66.6933, con una amplia gama de valores entre ellos, lo que indica una variabilidad sustancial en las latitudes. La mediana se sitúa en 39.3543, y el rango intercuartil (IQR) es de 7.3199, lo que sugiere una dispersión moderada en los datos.

El coeficiente de variación (CV) es de 0.13171047, lo que indica una variabilidad relativa baja en comparación con la media. Además, la kurtosis es de 0.81296795, lo que sugiere una distribución moderadamente puntiaguda alrededor de la media.

Statistics	Histogram	Common values	Extreme values
Quantile statistics		Descriptive statistics	
Minimum	20.0271	Standard deviation	5.0758084
5-th percentile	29.8826	Coefficient of variation (CV)	0.13171047
Q1	34.6205	Kurtosis	0.81296795
median	39.3543	Mean	38.537622
Q3	41.9404	Median Absolute Deviation (MAD)	3.3597
95-th percentile	45.8433	Skewness	-0.18602768
Maximum	66.6933	Sum	49970771
Range	46.6662	Variance	25.763831
Interquartile range (IQR)	7.3199	Monotonicity	Not monotonic

La gráfica de barras muestra las tres latitudes más comunes: 36.385 con una cuenta de 3,646, que representa el 0.3% del total; 26.1184 con una cuenta de 3,613, equivalente al 0.3%; y 42.5164 con una cuenta de 3,597, también el 0.3%.



En la variable "long" se identifican 969 valores distintos en esta columna, lo que representa un 0.1% de variabilidad en las longitudes registradas, indicando que esta variable tiende a contener una diversidad limitada de longitudes en los datos. No se encontraron valores faltantes ni infinitos en esta columna, lo que garantiza la integridad de los datos.

Es importante destacar que esta variable se resalta en rojo con la etiqueta "High_correlation", lo que indica una correlación significativa con otras variables en el conjunto de datos

long

Real number (R)

HIGH CORRELATION

Distinct	969	Minimum	-165.6723
Distinct (%)	0.1%	Maximum	-67.9503
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	1296675
Infinite (%)	0.0%	Negative (%)	100.0%
Mean	-90.226335	Memory size	9.9 MiB

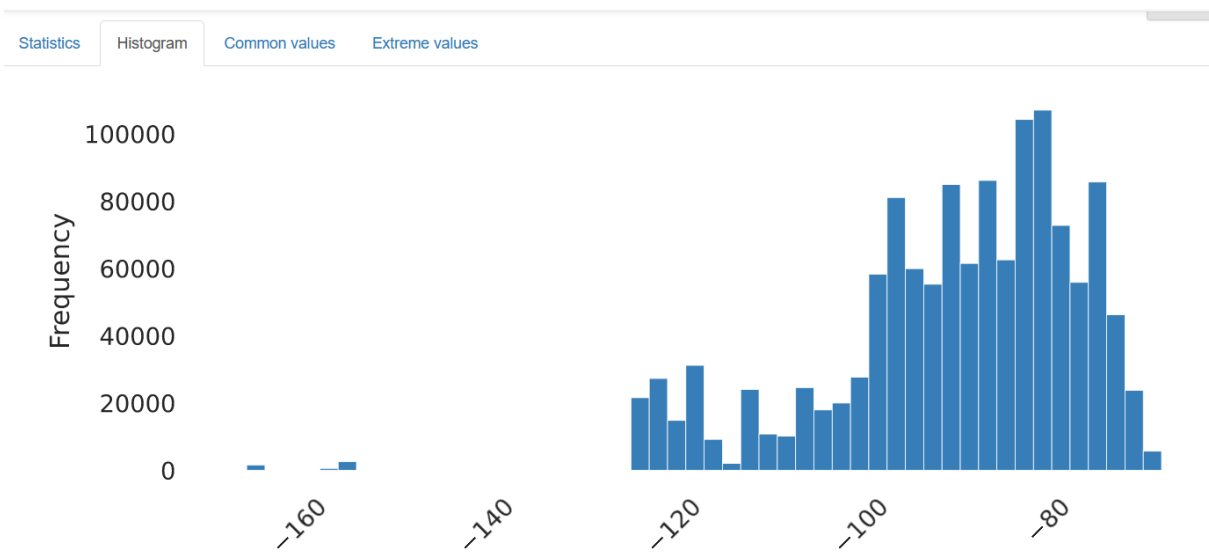
El análisis estadístico revela que el valor mínimo es -165.6723, mientras que el máximo es -67.9503, con una amplia gama de valores entre ellos, lo que

indica una variabilidad sustancial en las longitudes. La mediana se sitúa en -87.4769, y el rango intercuartil (IQR) es de 16.64, lo que sugiere una dispersión moderada en los datos.

El coeficiente de variación (CV) es -0.15249513, lo que indica una variabilidad relativa baja en comparación con la media. Además, la kurtosis es 1.8558923, lo que sugiere una distribución ligeramente puntiaguda alrededor de la media.

Statistics	Histogram	Common values	Extreme values
Quantile statistics		Descriptive statistics	
Minimum	-165.6723	Standard deviation	13.759077
5-th percentile	-119.0825	Coefficient of variation (CV)	-0.15249513
Q1	-96.798	Kurtosis	1.8558923
median	-87.4769	Mean	-90.226335
Q3	-80.158	Median Absolute Deviation (MAD)	8.1527
95-th percentile	-73.5112	Skewness	-1.1501077
Maximum	-67.9503	Sum	-1.1699423 × 10 ⁸
Range	97.722	Variance	189.3122
Interquartile range (IQR)	16.64	Monotonicity	Not monotonic

La gráfica de barras muestra las tres longitudes más comunes: -98.0727 con una cuenta de 3,646, que representa el 0.3% del total; -81.7361 con una cuenta de 3,613, equivalente al 0.3%; y -82.9832 con una cuenta de 3,597, también el 0.3%.



En la variable "job" se identifican 494 valores distintos en esta columna, lo que representa un porcentaje menor al 0.1% de variabilidad en los trabajos registrados, indicando que esta variable tiende a contener una variedad limitada de ocupaciones en los datos. Es relevante destacar que no se encontraron valores faltantes en esta columna, lo que garantiza que todas las observaciones tienen información sobre la ocupación en la variable "job". Además, el tamaño de memoria utilizado para almacenar esta columna es

de 9.9 MiB, lo que puede ser relevante para la gestión de recursos en el análisis de datos.

job
Text

Distinct	494
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

En la sección relacionada con las palabras más comunes en esta variable, se observa que las tres palabras que se repiten con mayor frecuencia en los trabajos son "engineer" con una cuenta de 131,756 veces, lo que representa un 4.6% del total de palabras en las ocupaciones. A continuación, "officer" cuenta con 110,915 repeticiones, equivalente al 3.9%, y "manager" con 61,124 repeticiones, es decir, el 2.1%. Estos resultados sugieren que ciertas palabras clave relacionadas con ocupaciones específicas, como "engineer", "officer" y "manager", son comunes en las descripciones de trabajo registradas.

Overview

Words

Characters

Value	Count	Frequency (%)
engineer	131756	4.6%
officer	110915	3.9%
manager	61124	2.1%
scientist	55878	1.9%
designer	52218	1.8%
surveyor	49062	1.7%
teacher	38126	1.3%
psychologist	32600	1.1%
research	29754	1.0%
editor	28725	1.0%
Other values (456)	2289024	79.5%

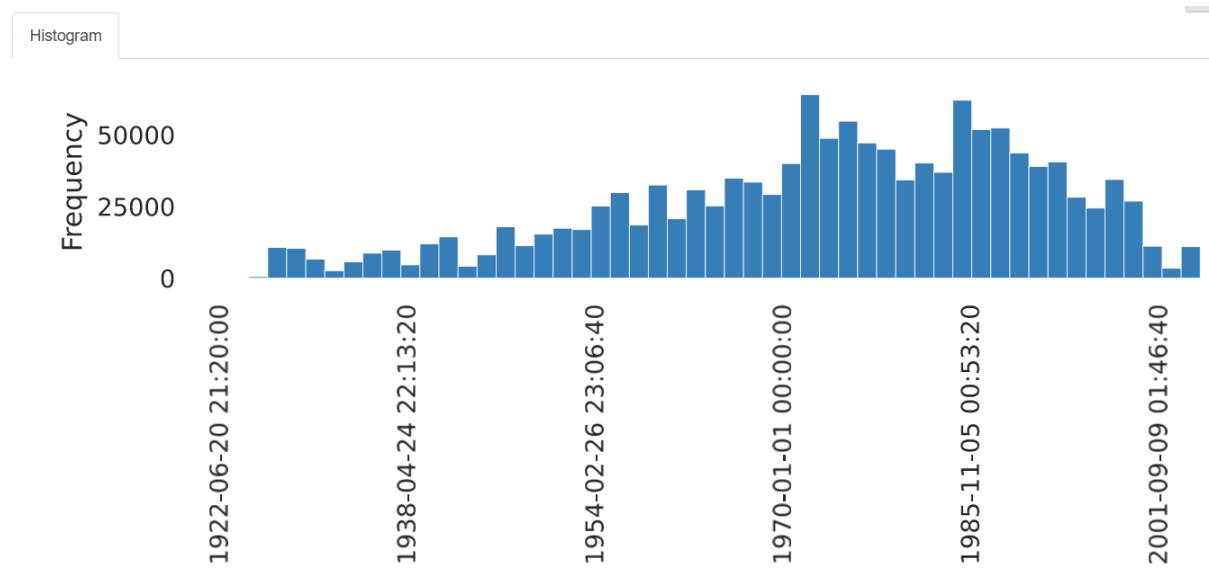
En la variable "dob", que representa la fecha de nacimiento en el conjunto de datos, existen 968 fechas de nacimiento distintas en esta columna, lo que equivale a un 0.1% de variabilidad en las fechas de nacimiento registradas.

No se encontraron valores faltantes en esta variable, lo que asegura que todas las observaciones tienen información sobre la fecha de nacimiento.

El análisis de las fechas de nacimiento también incluye estadísticas sobre los valores mínimo y máximo. La fecha de nacimiento más temprana registrada es el 30 de octubre de 1924, mientras que la más reciente es el 29 de enero de 2005. Esto muestra un rango amplio de fechas de nacimiento en los datos.

dob		Date	
Distinct	968	Minimum	1924-10-30 00:00:00
Distinct (%)	0.1%	Maximum	2005-01-29 00:00:00
Missing	0		
Missing (%)	0.0%		
Memory size	9.9 MiB		

En cuanto al histograma, se observa que la mayoría de las personas registradas en el conjunto de datos tienen fechas de nacimiento que caen en el período comprendido entre los años 1970 y 1985. Este patrón sugiere una concentración significativa de personas nacidas en esa década específica. El histograma es una representación visual útil para comprender la distribución de las fechas de nacimiento y puede ser valioso para análisis demográficos o segmentación por edad en futuros estudios



En la variable "trans_num" se identifican 1,296,675 valores distintos en esta columna, lo que significa que cada número de transacción es único en el conjunto de datos. Esto se refleja en un porcentaje del 100.0% de

singularidad, lo que indica que no se repiten números de transacción en los datos analizados.

No se encontraron valores faltantes en esta variable, lo que garantiza que todas las transacciones tienen un número de transacción registrado. Además, el tamaño de memoria utilizado para almacenar esta columna es de 9.9 MiB, lo que puede ser útil para la gestión de recursos en el análisis de datos.

trans_num

Text

UNIQUE

Distinct	1296675
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

La singularidad de los números de transacción es un hallazgo importante, ya que cada transacción se puede identificar de manera única en función de este número. Esto es esencial para tareas de seguimiento y verificación de transacciones en un conjunto de datos, ya que cada transacción puede ser rastreada de manera individual sin ambigüedad.

6d294ed2cc447d2c71c7171a3d54967c
83ec1cc84142af6e2acf10c44949e720
189a841a0a8ba03058526bcfe566aab5
fc28024ce480f8ef21a32d64c93a29f5
0b242abb623afc578575680df30655b9
c1d9a7ddb1e34639fe82758de97f4abf
3c74776e558f1499a7824b556e474b1d
8a6293af5ed278dea14448ded2685fea
6b849c168bdad6f867558c3793159a81
3b9014ea8fb80bd65de0b1463b00b00e
413636e759663f264aae1819a4d4f231
7bb25a43205191eb7344282b88fc54d3

En la variable "merch_lat" se identifican 1,247,805 valores distintos, lo que representa un 96.2% de variabilidad en las latitudes de los comercios registrados en los datos. No se encontraron valores faltantes ni infinitos en esta columna, lo que garantiza la integridad de los datos.

El análisis de estadísticas revela que el valor mínimo es 19.027785, mientras que el máximo es 67.5102267, lo que indica una amplia gama de valores en las latitudes de los comercios.

merch_lat

Real number (R)

HIGH CORRELATION

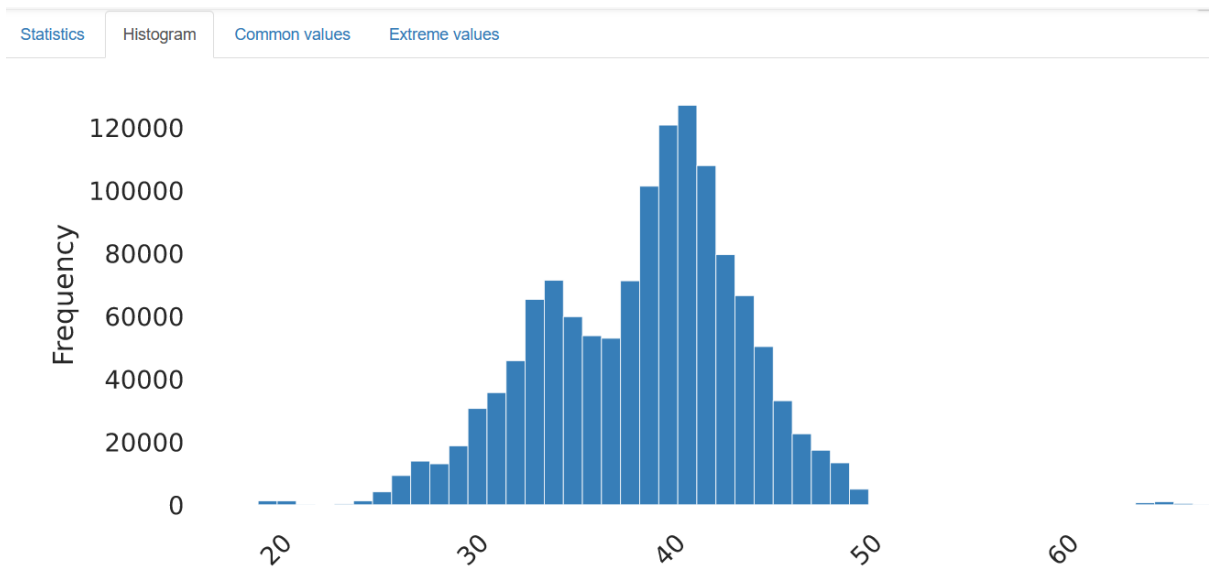
Distinct	1247805	Minimum	19.027785
Distinct (%)	96.2%	Maximum	67.510267
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	38.537338	Memory size	9.9 MiB

La mediana se encuentra en 39.36568, y el rango intercuartil (IQR) es de 7.223592, lo que sugiere una dispersión moderada en los datos. El coeficiente de variación (CV) es de 0.13259318, lo que indica una variabilidad relativa baja en comparación con la media.

Además, la kurtosis es 0.79599391, lo que sugiere una distribución ligeramente puntiaguda alrededor de la media.

Statistics	Histogram	Common values	Extreme values	
Quantile statistics		Descriptive statistics		
Minimum	19.027785		Standard deviation	5.1097884
5-th percentile	29.751653		Coefficient of variation (CV)	0.13259318
Q1	34.733572		Kurtosis	0.79599391
median	39.36568		Mean	38.537338
Q3	41.957164		Median Absolute Deviation (MAD)	3.397536
95-th percentile	46.00353		Skewness	-0.18191543
Maximum	67.510267		Sum	49970403
Range	48.482482		Variance	26.109937
Interquartile range (IQR)	7.223592		Monotonicity	Not monotonic

La gráfica de barras muestra las tres latitudes más comunes en los datos: 41.937796 con una cuenta de 3,613, lo que representa menos del 0.1% del total; 42.265012 con una cuenta de 3,597, también menos del 0.1%; y 41.305966 con una cuenta de 4, lo que también representa menos del 0.1%. Estos resultados son fundamentales para comprender la distribución de las latitudes de los comercios y pueden ser relevantes en análisis geográficos o en la segmentación de datos según ubicaciones específicas. La alta correlación resaltada sugiere que esta variable podría estar relacionada con otras en el conjunto de datos, lo que puede tener implicaciones importantes en el análisis.



En la variable "merch_long" se cuenta con 1,275,745 valores distintos, lo que representa un 98.4% de variabilidad en las longitudes de los comercios registrados en los datos. No se encontraron valores faltantes ni infinitos en esta columna, lo que garantiza la integridad de los datos.

El análisis de estadísticas revela que el valor mínimo es -166.67124, mientras que el máximo es -66.950902, lo que indica una amplia gama de valores en las longitudes de los comercios.

merch_long

Real number (R)

HIGH CORRELATION

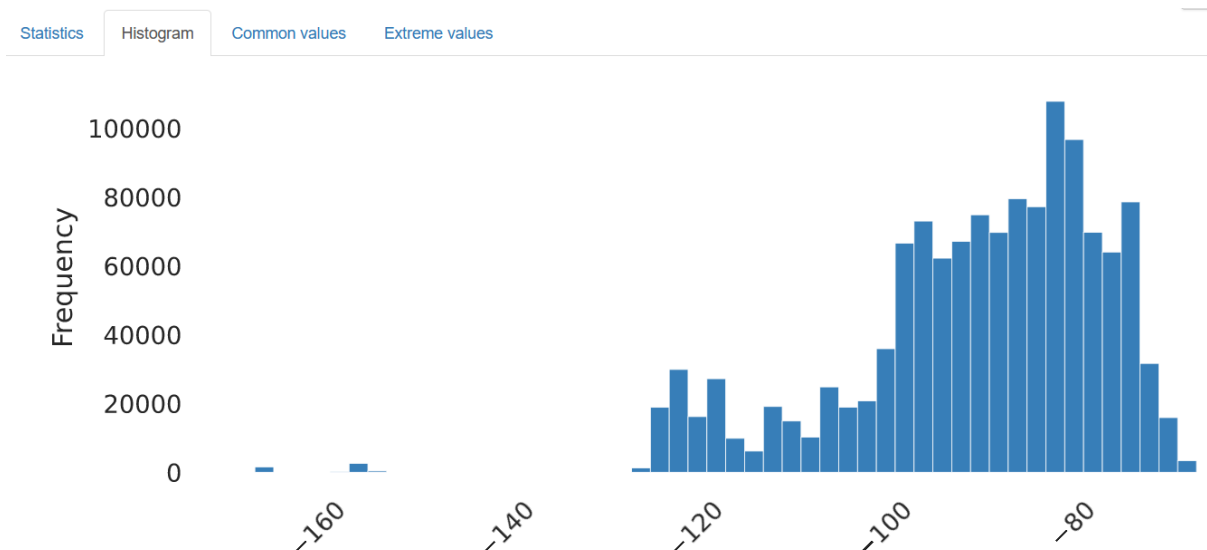
Distinct	1275745	Minimum	-166.67124
Distinct (%)	98.4%	Maximum	-66.950902
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	1296675
Infinite (%)	0.0%	Negative (%)	100.0%
Mean	-90.226465	Memory size	9.9 MiB

La mediana se encuentra en -87.438392, y el rango intercuartil (IQR) es de 16.660479, lo que sugiere una dispersión moderada en los datos. El coeficiente de variación (CV) es de -0.15262806, lo que indica una variabilidad relativa baja en comparación con la media.

La kurtosis es 1.8484792, lo que sugiere una distribución ligeramente puntiaguda alrededor de la media.

Statistics	Histogram	Common values	Extreme values
Quantile statistics		Descriptive statistics	
Minimum	-166.67124	Standard deviation	13.771091
5-th percentile	-119.33009	Coefficient of variation (CV)	-0.15262806
Q1	-96.897276	Kurtosis	1.8484792
median	-87.438392	Mean	-90.226465
Q3	-80.236796	Median Absolute Deviation (MAD)	8.227889
95-th percentile	-73.354218	Skewness	-1.1469599
Maximum	-66.950902	Sum	-1.169944 × 10 ⁸
Range	99.72034	Variance	189.64294
Interquartile range (IQR)	16.660479	Monotonicity	Not monotonic

La gráfica de barras muestra las tres longitudes más comunes en los datos: -81.219189 con una cuenta de 3,613, lo que representa menos del 0.1% del total; -74.618269 con una cuenta de 3,597, también menos del 0.1%; y -87.116414 con una cuenta de 4, lo que también representa menos del 0.1%. Estos resultados son fundamentales para comprender la distribución de las longitudes de los comercios y pueden ser relevantes en análisis geográficos o en la segmentación de datos según ubicaciones específicas. La alta correlación resaltada sugiere que esta variable podría estar relacionada con otras en el conjunto de datos, lo que puede tener implicaciones importantes en el análisis.



En la variable "is_fraud" se tiene 2 categorías distintas, lo que equivale a menos del 0.1% de variabilidad en las categorías. No se encontraron valores faltantes en esta variable, lo que garantiza que todos los casos tienen una etiqueta de fraude o no fraude.

is_fraud

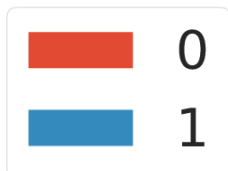
Categorical

IMBALANCE

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	9.9 MiB

El análisis de comparación revela un desequilibrio significativo en las clases. Hay 1,289,186 casos etiquetados como "reales" o no fraudulentos, lo que representa el 99.4% del total de casos. Por otro lado, hay solo 7,506 casos etiquetados como "fraudulentos", que ocupan solo el 0.6% del total. Este desequilibrio en la distribución de clases es un hallazgo importante y puede tener implicaciones significativas en la construcción de modelos de detección de fraudes. Es común que los modelos se sesguen hacia la clase mayoritaria (en este caso, "reales"), lo que podría resultar en una menor capacidad para identificar casos de fraude.

99.4%
(1289169)



En conclusión, el análisis detallado de las variables clave en el "Reporte: Detección de Fraude en Tarjetas de Crédito" generado a través de la librería `pandas_profiling` nos proporciona una visión profunda y valiosa de los datos relacionados con transacciones con tarjetas de crédito.

Hemos examinado minuciosamente las estadísticas y gráficas que revelan información esencial sobre las características de estas variables, su distribución y comportamiento.

En el contexto de la detección de fraudes en tarjetas de crédito, este informe se convierte en una herramienta fundamental para comprender y abordar los patrones fraudulentos. Además, hemos identificado desequilibrios en la distribución de clases y alta correlación en algunas variables, aspectos críticos a considerar en la construcción de modelos de detección de fraudes.

Estos hallazgos respaldan la toma de decisiones informadas y estrategias efectivas en la lucha contra los fraudes en las transacciones con tarjetas de crédito, ayudando así a proteger a los usuarios y las instituciones financieras.

Limpieza de los datos

Afortunadamente, nos percatamos de que nuestro conjunto de datos no presenta ningún valor faltante. Por lo tanto, lo único que consideramos viable para su mejora fue la variable `"trans_date_trans_time"`. Decidimos primero dividirla en dos atributos distintos, `"trans_date"` y `"trans_time"`, con el

propósito de separar la fecha y la hora en que se llevó a cabo la transacción en dos categorías independientes.

```
df['trans_date_trans_time'] = df['trans_date_trans_time'].astype('datetime64[ns]')
df = df.assign(trans_date = df['trans_date_trans_time'].dt.strftime("%d/%m/%Y"))
df = df.assign(trans_time = df['trans_date_trans_time'].dt.strftime("%H:%M:%S"))
```

Después de dividir la variable "trans_date_trans_time" en dos atributos separados, se convirtió la variable "trans_time" primero en un tipo de dato datetime y luego en una cadena de caracteres (string) para facilitar su manipulación. Finalmente, se redondearon los minutos a la hora más cercana con el propósito de eliminar los minutos y los segundos.

```
# Convertir la columna 'trans_time' a tipo datetime.time y luego a un string para manipularla más fácilmente
df['trans_time'] = pd.to_datetime(df['trans_time'], format='%H:%M:%S').dt.time.astype(str)

# Redondear los minutos a la hora más cercana (eliminando los minutos y segundos)
df['trans_time_rounded'] = df['trans_time'].str[:2] + ':00'
```

Para la variable "trans_date", se procedió a convertirla en un tipo de dato datetime y luego se descompuso en cuatro variables distintas: "year," "month," "day," y "weekday."

```
# Convertir la columna 'trans_date' a tipo datetime
df['trans_date'] = pd.to_datetime(df['trans_date'], format='%d/%m/%Y')

# Descomponer la columna 'trans_date' en componentes temporales
df['year'] = df['trans_date'].dt.year
df['month'] = df['trans_date'].dt.month
df['day'] = df['trans_date'].dt.day
df['weekday'] = df['trans_date'].dt.weekday
```

Finalmente, se llevó a cabo un proceso de depuración en el conjunto de datos. Esto implicó la eliminación de varias variables que no eran necesarias para nuestro análisis o que podrían introducir ruido en los modelos de detección de fraudes. Las variables eliminadas incluyen información como números de tarjeta de crédito, nombres de usuarios, género, ubicación y otros detalles personales que no estaban relacionados directamente con la detección de fraudes.

```
df = df.drop(['cc_num'],axis=1)
df = df.drop(['first'],axis=1)
df = df.drop(['last'],axis=1)
df = df.drop(['gender'],axis=1)
df = df.drop(['city'],axis=1)
df = df.drop(['zip'],axis=1)
df = df.drop(['street'],axis=1)
df = df.drop(['job'],axis=1)
df = df.drop(['dob'],axis=1)
df = df.drop(['trans_num'],axis=1)
df = df.drop(['unix_time'],axis=1)
df = df.drop(['merchant'],axis=1)
df = df.drop(['merch_lat'],axis=1)
df = df.drop(['merch_long'],axis=1)
df = df.drop(['trans_date_trans_time'],axis=1)
```

```
df = df.drop(['year'],axis=1)
```

Como resultado final del proceso de limpieza y selección de variables, hemos depurado el conjunto de datos para retener únicamente las características esenciales para nuestro análisis de detección de fraudes en transacciones con tarjetas de crédito. Las variables seleccionadas incluyen: "category" (categoría), "amt" (monto de la transacción), "state" (estado), "lat" (latitud), "long" (longitud), "city_pop" (población de la ciudad), "is_fraud" (indicador de fraude), "trans_time_rounded" (hora de la transacción redondeada), "month" (mes de la transacción), "day" (día de la transacción) y "weekday" (día de la semana de la transacción).

Estas variables representan la información crítica necesaria para identificar y analizar patrones de fraude de manera efectiva, al tiempo que reducen el riesgo de divulgar información sensible o redundante en el proceso

	category	amt	state	lat	long	city_pop	is_fraud	trans_time_rounded	month	day	weekday
0	entertainment	170.06	FL	28.5697	-80.8191	54767	0	16:00	12	5	3
1	gas_transport	76.92	MI	44.2529	-85.0170	1126	0	09:00	8	19	0
2	grocery_pos	268.73	WI	42.8035	-88.4092	9679	1	01:00	7	3	2
3	grocery_pos	345.91	AL	32.2844	-86.9920	800	1	03:00	4	12	4
4	grocery_pos	290.95	KY	37.1046	-83.5706	467	1	00:00	1	13	0

Modelado

Regresión Logística

La detección de fraude puede beneficiarse enormemente del empleo de un modelo de regresión logística debido a múltiples factores clave. Específicamente, este método se alinea idealmente con la esencia binaria del tema, ya que nuestra preocupación es si las transacciones se clasifican como legítimas o fraudulentas. En consecuencia, la regresión logística está diseñada específicamente para este tipo de discriminación, prediciendo la probabilidad de que una transacción sea fraudulenta o no. La detección de fraude puede beneficiarse enormemente de la regresión logística debido a sus coeficientes fácilmente interpretables. Estos coeficientes están vinculados a características específicas, por ejemplo, el historial del cliente o el monto de la transacción, y demuestran cómo cada característica influye en la probabilidad de que una transacción sea fraudulenta. Como resultado, esta metodología resalta los factores más importantes para la detección de fraude.

Incluir términos de interacción o características polinomiales permite modelar relaciones no lineales, lo cual es un gran beneficio. La regresión logística también puede manejar eficazmente la multicolinealidad, que es una circunstancia en la que ciertas características están significativamente correlacionadas.

El atributo de la regresión logística de ser altamente interpretable se destaca como un rasgo valioso, que permite explicar claramente las decisiones del modelo a los reguladores y otras partes interesadas. Esta cualidad es particularmente importante en las aplicaciones financieras, donde la necesidad de transparencia en las decisiones de detección de fraude es de suma importancia.

Incluso en grandes conjuntos de datos, la regresión logística puede entrenar rápidamente el modelo, haciéndolo computacionalmente eficiente para entornos en tiempo real.

Por lo cual, emplear la regresión logística como táctica para combatir el fraude es un enfoque eficaz en este proyecto gracias a su ajuste de cuestiones binarias, su comprensibilidad, su capacidad para representar conexiones no lineales y su hazaña computacional. Sin embargo, debemos tener en cuenta que ningún modelo es impecable y es mejor impulsar este método junto con otros enfoques y prácticas, como seleccionar datos con habilidad y perfeccionar atributos, para aumentar la precisión en circunstancias reales.

Grid

Los resultados obtenidos en la detección de fraudes a través de un modelo de regresión logística son altamente prometedores. El elevado valor de Accuracy del 90.25% indica que el modelo clasifica correctamente la gran mayoría de las transacciones, lo cual es esencial para minimizar los errores de clasificación. La precisión, con un valor del 90.48%, señala que la gran mayoría de las transacciones clasificadas como fraudulentas son genuinamente fraudulentas, lo que refuerza la confiabilidad del modelo. Además, el Recall del 73.51% indica que el modelo captura apropiadamente más del 70% de las transacciones fraudulentas, un aspecto crucial para detectar la mayoría de los fraudes. Por último, el puntaje de F1, que es de 81.12%, refleja un equilibrio efectivo entre precision y recall.

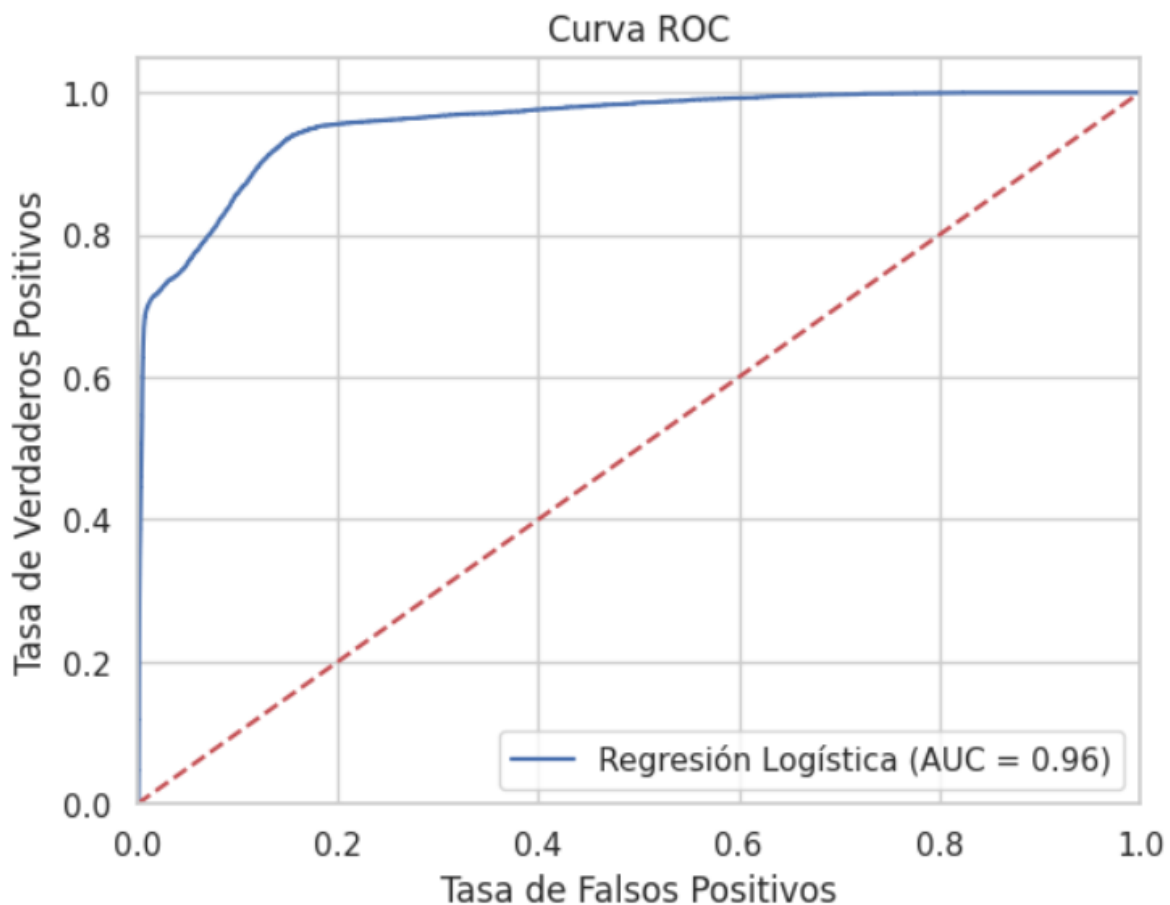
Estos logros se lograron al optimizar el modelo mediante la técnica de Grid Search. El Grid Search es un proceso que permite buscar la combinación más adecuada de hiper parámetros para el modelo, en este caso, la regresión logística, maximizando su rendimiento.

```
Accuracy: 0.9025924735710644  
Precision: 0.9048263473053892  
Recall: 0.7351873655940137  
F1 Score: 0.8112333370913796
```

Es importante destacar que, para obtener los parámetros previamente mencionados (Accuracy, Precision, Recall, F1 Score), se llevó a cabo un proceso de balanceo de datos. Esto implicó generar un 40% de datos adicionales, conservando el 60% restante de datos originales. Este enfoque se implementó debido a que la mayoría de los datos originales tenían una probabilidad de ser fraudulentos igual o mayor al 98%. Esto propiciaba un sobreajuste en el modelo de regresión logística, lo que limitaba su capacidad para realizar predicciones precisas de nuevas transacciones como fraudulentas o legítimas. Por lo tanto, el balanceo de datos se convirtió en una estrategia efectiva para mejorar el rendimiento del modelo y garantizar una detección confiable de transacciones fraudulentas.

Resultados

La evaluación de la eficacia de nuestro modelo de Regresión Logística se llevó a cabo mediante la Curva ROC. Esta herramienta gráfica representa la Tasa de Verdaderos Positivos frente a la Tasa de Falsos Positivos en diferentes umbrales de clasificación. Según los resultados obtenidos, el Área Bajo la Curva (AUC) alcanzó un valor de 0.96, lo que indica un alto grado de precisión y eficacia del modelo. La notable separación de la curva respecto a la línea de no-discriminación subraya la capacidad del modelo para distinguir adecuadamente entre las clases.



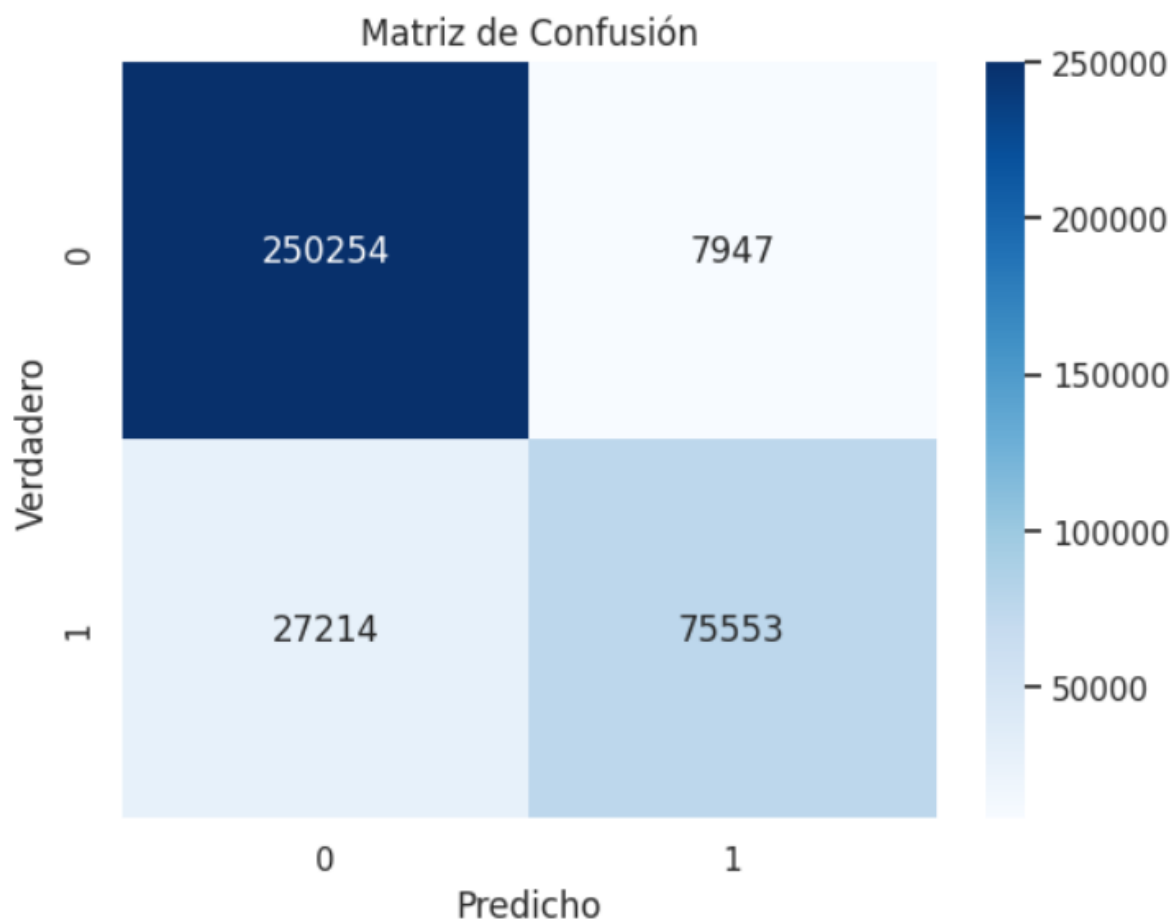
La matriz de confusión es utilizada para evaluar el rendimiento de nuestro modelo de clasificación. De un total de 250,254 instancias reales de la clase 0, nuestro modelo predijo correctamente 250,254 de ellas, mientras que erróneamente clasificó 7,947 como clase 1. Por otro lado, de las instancias reales de la clase 1, el modelo predijo correctamente 75,553 y erróneamente clasificó 27,214 como clase 0. Estos resultados nos proporcionan una perspectiva detallada de dónde el modelo está acertando y dónde se pueden hacer mejoras.

Verdaderos Negativos (VN): 250,254 - Estas son las instancias que fueron correctamente clasificadas como la clase 0.

Falsos Positivos (FP): 7,947 - Estas instancias fueron incorrectamente clasificadas como clase 1 cuando en realidad son de la clase 0.

Falsos Negativos (FN): 27,214 - Estas instancias fueron incorrectamente clasificadas como clase 0 cuando en realidad son de la clase 1.

Verdaderos Positivos (VP): 75,553 - Estas son las instancias que fueron correctamente clasificadas como la clase 1.



El código presentado es una interfaz gráfica construida usando la biblioteca `tkinter` en Python. Su objetivo principal es realizar predicciones de fraude utilizando un modelo previamente entrenado.

Funciones Principales:

load_csv(): Esta función permite al usuario cargar un archivo CSV. Una vez cargado, los datos se envían a la función `batch_predict` para realizar predicciones en lote.

batch_predict(data): Toma el conjunto de datos cargado y realiza predicciones utilizando el modelo preentrenado. Las probabilidades y predicciones se almacenan y se actualizan los gráficos correspondientes.

update_prediction_distribution_plot(): Muestra un histograma que representa la distribución de las predicciones realizadas, diferenciando entre transacciones fraudulentas y no fraudulentas.

predict_and_metrics(input_df): Realiza una predicción basada en datos proporcionados y devuelve la predicción junto con la probabilidad asociada.

predict(amt, city_pop, lat, long, category, state, day, time, month, weekday): Simula una función de predicción utilizando las entradas proporcionadas por el usuario y devuelve la predicción junto con la probabilidad de fraude.

button_predict(): Esta función se activa al presionar el botón "Predecir" en la interfaz. Recopila las entradas del usuario, realiza la predicción y actualiza los gráficos. Finalmente, muestra el resultado en una ventana separada.

update_probability_distribution_plot(): Representa un histograma que muestra la distribución de las probabilidades de fraude de las transacciones analizadas.

clear_entries(): Limpia todas las entradas en la interfaz, permitiendo al usuario ingresar nuevos datos.

Diseño de la Interfaz:

La interfaz gráfica es intuitiva, proporcionando campos para introducir detalles como cantidad, población de la ciudad, latitud, longitud, hora, mes, día y día de la semana. Estos detalles son fundamentales para que el modelo genere sus predicciones. Tras introducir los datos, el usuario puede presionar "Predecir" para determinar la naturaleza de la transacción.

Las gráficas desempeñan un papel crucial en esta herramienta. No solo visualizan las predicciones y probabilidades en tiempo real, sino que también ofrecen retroalimentación sobre el comportamiento del modelo. Al observar las gráficas, el usuario puede tener una idea clara de cómo el modelo está interpretando los datos ingresados, permitiendo una comprensión más profunda y una calibración potencial del modelo en el futuro.

Detección de Fraude

Cantidad:

Población de la Ciudad:

Latitud:

Longitud:

Hora:

Mes:

Día:

Día de la semana:

Estado:

Categoría:

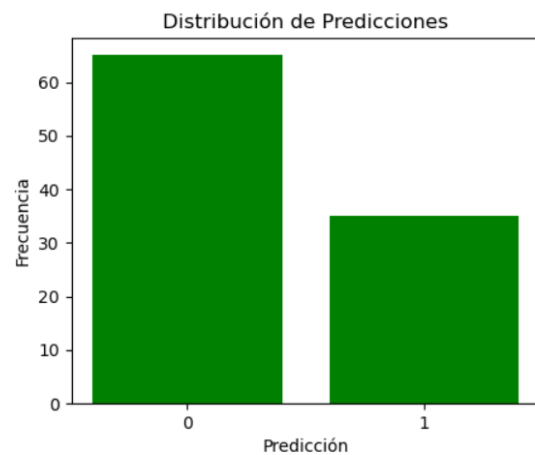
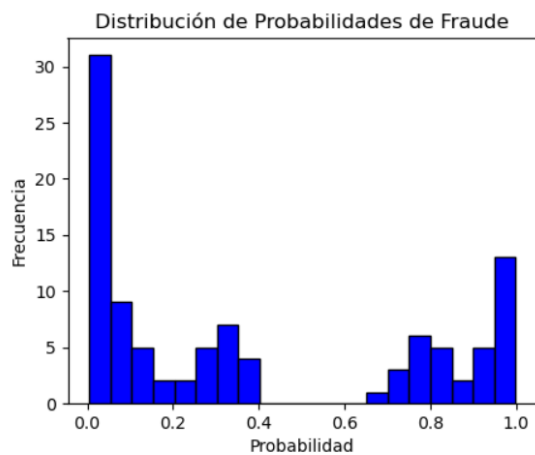
Predecir

Limpiar

Cargar CSV

Detección de Fraude

Cantidad:	<input type="text"/>
Población de la Ciudad:	<input type="text"/>
Latitud:	<input type="text"/>
Longitud:	<input type="text"/>
Hora:	<input type="text"/>
Mes:	<input type="text"/>
Día:	<input type="text"/>
Día de la semana:	<input type="text"/>
Estado:	<input type="text"/>
Categoría:	<input type="text"/>
<input type="button" value="Predecir"/>	<input type="button" value="Limpiar"/>



Predicción de Fraude

Resultado de la Predicción...

Predicción: No Fraudulento

Probabilidad de fraude: 0.04

Cerrar

Detección de Fraude

Cantidad:170.06

Población de la Ciudad:54767

Latitud:28.5697

Longitud:-80.8191

Hora:16:00

Mes:12

Día:5

Día de la semana:3

Estado:FL

Categoría:entertainment

Predicir

Limpiar

Predicción de Fraude

Resultado de la Predicción...

Predicción: Fraudulento

Probabilidad de fraude: 0.71

Cerrar

Detección de Fraude

Cantidad:268.73

Población de la Ciudad:9679

Latitud:42.8035

Longitud:-80.8191

Hora:01:00

Mes:7

Día:3

Día de la semana:2

Estado:WI

Categoría:grocery_pos

Predicir

Limpiar

Conclusiones

En conclusión, este proyecto resalta la valiosa contribución de la ciencia de datos en la toma de decisiones críticas, como la detección de transacciones fraudulentas, a pesar de que algunos datos fueran simulados.

Las metas establecidas se lograron exitosamente, incluso superando el desafío del sobreajuste inicial. Hemos alcanzado un impresionante nivel de precisión, con un accuracy del 90%, y la interfaz gráfica ha demostrado su capacidad para generar distribuciones de probabilidad precisas.

En el proyecto se ha logrado implementar un modelo de detección de fraudes eficiente y altamente interpretable mediante regresión logística. Aunque el modelo ha mostrado un rendimiento notable, es esencial considerar su integración con otros métodos y técnicas para aumentar aún más la precisión y robustez en situaciones reales. La herramienta diseñada no solo es eficiente en la detección, sino que también es amigable y educativa para el usuario. Es una combinación efectiva de ciencia de datos y diseño de interfaz para abordar un problema financiero crítico.

El siguiente paso implica implementar este modelo en un entorno en vivo y continuar afinándolo con actualizaciones periódicas. Esto permitirá mejorar aún más su precisión y eficacia en la detección de fraudes, destacando así el valor de la ciencia de datos en la industria financiera.