Universidad Nacional de San Luis

Facultad de Ciencias Físico Matemáticas y Naturales

DEPARTAMENTO DE INFORMÁTICA

2025

LABORATORIO DE DATOS

GRUPO 36

TRABAJO PRÁCTICO N°1:

Introducción a los datos

ALUMNOS:

- PONCE VIRGINIA ALEJANDRA.
- WALTER AGUSTIN QUIROGA.

PROFESORES:

- BARRIONUEVO, MERCEDES.
- VILLEGAS, PAULA.
- VALLEJO, ENRIQUE.

INTRODUCCIÓN

Dataset de Recuperos de Autos:
https://docs.google.com/spreadsheets/d/1wiUSNb_XDS20cxHhc8dd-enwJc42l8MlBwWWER4Qii4/edit?usp=sharing

En este trabajo práctico, abordamos el análisis de datos utilizando el dataset proporcionado por la Dirección Nacional del Registro de la Propiedad Automotor, el cual contiene información sobre los vehículos recuperados tras haber sido denunciados como robados.

¿Categórico o Numérico? Por favor, especificar cada uno de los atributos.

Identificación de atributos y tipos de datos:

Atributo	Tipo de Dato	Descripción
tramite_tipo	Cadena de texto (String)	Tipo de trámite realizado.
tramite_fecha	Fecha	Fecha en la que se realizó el trámite.
fecha_inscripcion_inicial	Fecha	Fecha de inscripción inicial del automotor.
registro_seccional_codigo	Numérico (Entero)	Código del registro seccional.
registro_seccional_descripcion	Cadena de texto (String)	Descripción del registro seccional.
registro_seccional_provincia	Cadena de texto (String)	Provincia del registro seccional.
automotor_origen	Cadena de texto (String)	Origen del automotor (Ej: Nacional, Importado).
automotor_anio_modelo	Numérico (Entero)	Año del modelo del automotor.
automotor_tipo_codigo	Numérico (Entero)	Código del tipo de automotor.
automotor_tipo_descripcion	Cadena de texto (String)	Descripción del tipo de automotor.
automotor_marca_codigo	Numérico (Entero)	Código de la marca del automotor.
automotor_marca_descripcion	Cadena de texto (String)	Descripción de la marca del automotor.

automotor_modelo_codigo	Cadena de texto (String)	Código del modelo del automotor (puede contener letras y números).
automotor_modelo_descripcion	Cadena de texto (String)	Descripción del modelo del automotor.
automotor_uso_codigo	Numérico (Entero)	Código del uso del automotor.
automotor_uso_descripcion	Cadena de texto (String)	Descripción del uso del automotor.
titular_tipo_persona	Cadena de texto (String)	Tipo de persona titular (Física o Jurídica).
titular_domicilio_localidad	Cadena de texto (String)	Localidad del domicilio del titular.
titular_domicilio_provincia	Cadena de texto (String)	Provincia del domicilio del titular.
titular_genero	Cadena de texto (String)	Género del titular (Masculino, Femenino).
titular_anio_nacimiento	Numérico (Entero)	Año de nacimiento del titular.
titular_pais_nacimiento	Cadena de texto (String)	País de nacimiento del titular.
titular_porcentaje_titularidad	Numérico (Entero)	Porcentaje de titularidad del automotor.
titular_domicilio_provincia_id	Numérico (Entero)	Identificador numérico de la provincia del domicilio del titular.
titular_pais_nacimiento_id	Numérico (Entero)	Identificador numérico del país de nacimiento del titular.

Detección de anomalías y valores faltantes

Para hablar sobre las anomalías que identificamos en la base de datos, primero debemos tener en cuenta el “Proceso de recolección de datos” en el cual se obtienen los datos relativos a las denuncias de robo o hurto y el recupero de vehículos en Argentina, a través del sistema de la **Dirección Nacional de los Registros de la Propiedad del Automotor (DNRPA)**, correspondiente al año 2024.

Estos datos se obtienen por una “Denuncia policial”, la persona afectada por el robo o hurto de un vehículo realiza una denuncia en la policía, la cual se puede efectuar tanto en línea mediante el sistema de Denuncia Virtual, como de forma presencial en las comisarías. Esta denuncia incluye datos sobre el vehículo, como la marca, el modelo, el dominio (patente), y la fecha del robo,etc.

Tras realizar la denuncia policial, el titular del vehículo se comunica con la DNRPA para registrar el robo o hurto en su base de datos. Este trámite también se puede realizar en línea a través de la página web de la DNRPA o de forma presencial en sus delegaciones. Los datos registrados incluyen los detalles del vehículo y la información sobre el denunciante. Cuando el vehículo robado es recuperado por las autoridades, el proceso de actualización de la base de datos se realiza en el sistema de la DNRPA, donde se reflejan los detalles del recupero, como la fecha y el lugar.

Al revisar la base de datos correspondiente a las denuncias y recuperos de vehículos, identificamos varias inconsistencias o anomalías en los registros.

1. **Fechas incorrectas:**

En la columna "automotor_anio_modelo", la mayoría de los registros contienen años de 4 dígitos, como 1965 o 2020, lo que es coherente con los datos de vehículos reales. Sin embargo, en algunos registros, se observan años representados solo por dos dígitos, como "20", lo que genera confusión, ya que no está claro si se refiere al año 1920 o al 2020.

Soluciones propuestas

Para corregir y mejorar la calidad de los datos, proponemos las siguientes soluciones, basadas en el supuesto de un sistema automatizado de registro virtual a través de una página web:

- Validación automática del año de modelo: Se implementará una validación automática en el formulario de la página web de la DNRPA para asegurar que el año del modelo ingresado esté en un rango lógico y realista. Por ejemplo que contenga 4 dígitos.

- Integración con el sistema de la Policía: Se propondrá una integración entre el sistema de la DNRPA y el sistema de denuncia policial para automatizar la entrada de datos en el sistema de la DNRPA. Esto reduciría los errores manuales y garantiza la consistencia en los registros, ya que los datos ingresados en la denuncia policial serían automáticamente transferidos a la base de datos de la DNRPA.
- Auditoría continua de los registros: Se implementará un sistema de auditoría automática que permita verificar que todos los datos ingresados cumplan con los criterios establecidos y alertar sobre posibles inconsistencias antes de que los registros sean finalizados.

Solución en la base de datos:
Bien, ahora suponiendo que un automóvil del año 1920 es altamente improbable en los registros actuales, un valor de "20" es lógico que sea interpretado como el año 2020. En este caso, nuestro grupo opta por modificar el valor 20 a 2020 en la base de datos, para luego poder mostrar de manera visual y prolija los datos y estadísticas.

Justificación del supuesto

La suposición de que el valor "20" en el campo "automotor_anio_modelo" se refiere al año 2020, en lugar de 1920, se basa en la observación de que los vehículos modernos más comunes tienen años de fabricación cercanos a la fecha actual (2024). Es ilógico que un vehículo registrado en el año 2024 tenga un modelo de hace más de 100 años. Por lo tanto, al encontrarnos con un valor de solo dos dígitos en este campo, asumimos que el valor hace referencia al siglo XXI, más específicamente al año 2020.

2. Valores faltantes o nulos:

En los registros de la base de datos, se identificaron campos vacios o valores nulos, en la columna "automotor_anio_modelo" y en “automotor_tipo_codigo”. Esta falta de información puede generar problemas en la integridad de los datos, ya que son campos claves para realizar análisis y generar estadísticas sobre los vehículos.

Soluciones propuestas

Para abordar los valores faltantes, proponemos las siguientes soluciones, basadas en el supuesto de que se trata de un sistema automatizado de registro virtual a través de una página web:

- Validación automática de los datos: En un futuro sistema de registro automatizado, se podría implementar una validación que obligue a los usuarios a completar el campo de "automotor_anio_modelo" y “automotor_tipo_codigo” antes de finalizar el trámite. Si el valor está vacío, el sistema podría alertar al usuario y solicitar que ingrese un dato válido. Además, en el caso de los años se validará que el valor ingresado esté dentro de un rango de años plausible, para evitar errores en el futuro.

Solución en la base de datos:

En el caso de los valores faltantes en el campo, podríamos llevar a cabo un proceso de imputación de valores. La imputación consiste en sustituir los valores faltantes por estimaciones basadas en los datos disponibles. Existen diferentes métodos de imputación:

- Imputación basada en la media o mediana: Si el campo tiene una distribución relativamente homogénea de años, podemos utilizar la mediana o la media de los valores existentes para imputar los registros faltantes. La mediana sería preferible si los datos contienen valores atípicos, ya que es menos sensible a valores extremos.

- Imputación por regresión: Si existe una relación significativa entre los campos que no tienen valor y otros campos del conjunto de datos (como la marca, el modelo o el uso del automotor), podemos utilizar un modelo de regresión para predecir el valor del año del modelo basándonos en esos otros atributos. Este enfoque permite hacer una imputación más informada, alineada con las características del vehículo.

- Imputación por valores más frecuentes (moda): Si observamos que hay ciertos valores que se repiten frecuentemente, podríamos optar por imputar el

valor más frecuente o modal, lo que puede ser especialmente útil en campos categóricos con un número limitado de opciones.

- Mantenimiento de registros como nulos: Si no es posible imputar un valor adecuado de forma confiable, otra opción sería dejar los valores como "nulos" o "vacíos". De esta forma, podríamos aplicar técnicas de análisis que no se vean afectadas por la ausencia de estos datos.

Nosotros decidimos con respecto a estos valores faltantes, mantener los registros con los campos vacios o valores nulos (NULL) en lugar de asignarles un valor predeterminado. Esto se debe a que la asignación de un valor arbitrario, podría introducir sesgo en los análisis y no reflejar adecuadamente la realidad de los datos. De esta forma, los registros con valores nulos representan casos donde la información no está disponible o no se pudo determinar con precisión. Este enfoque permite mantener la integridad de la base de datos y facilita la toma de decisiones basada en datos completos y verificables.

3. Un campo con un valor no válido:

Durante el análisis de la base de datos, se identificó un valor no válido ("."), presente en el atributo "automotor_tipo_descripcion", el cual no corresponde a un tipo de vehículo reconocible ni aporta información útil al conjunto de datos. Para abordar este problema, decidimos aplicar la “imputación por regresión”. A partir de los valores presentes en otros campos completos (como marca, modelo y atributos relacionados), utilizamos un modelo de regresión para predecir y asignar el tipo de vehículo más apropiado. Este método nos permitió preservar la integridad y coherencia de la base de datos, completando correctamente los registros incompletos conforme a las tendencias observadas en el resto de los datos.

Como resultado de la investigación, determinamos que el tipo de vehículo correspondiente era "PICK-UP". Además, al realizar este cambio, observamos que existían dos variantes en los datos: "PICK UP" y "PICK-UP". Para garantizar la consistencia, tomamos la decisión de estandarizar todos los registros bajo el término "PICK-UP".

Otro campo similar a este que se identificó, es el atributo “automotor_marca_descripcion” con un valor no válido (*), el cual no corresponde a una marca reconocible ni aporta información útil al conjunto de datos. Para abordar este problema, también decidimos aplicar la “imputación por regresión”. A partir de los valores presentes en otros campos completos (como marca, modelo y atributos relacionados), utilizamos un modelo de regresión para predecir y asignar la marca del vehículo “FIAT”.

3. Corrección de errores en los campos:

En los atributos "automotor_tipo_descripcion", “titular_domicilio_localidad” y en “automotor_marca_descripcion”, identificamos inconsistencias en la forma en que se registraron los datos. Algunos valores presentaban variaciones debido a la presencia de signos de puntuación innecesarios (por ejemplo, palabras con un punto al final), mientras que otros registros contenían el mismo tipo de vehículo pero con diferencias en el orden de las palabras o en la forma de escritura. Para garantizar la uniformidad y mejorar la calidad de los datos, realizamos un proceso de normalización, en el cual:

- Eliminamos caracteres innecesarios, como puntos o espacios adicionales.
- Corregimos errores de ortografía, y signos de ortográficos, como los acentos.
- Unificamos la nomenclatura para mantener coherencia en la base de datos.

Este proceso asegura que los registros sean más homogéneos, reduciendo errores y mejorando la interpretación de los datos.

4. Manejo de valores inválidos:

Durante el proceso de limpieza y estandarización de datos, identificamos registros en los campos "automotor_marca_descripcion" y "automotor_modelo_descripcion" que contenían la etiqueta "Marva invalida" y “Modelo invalido”. Por lo cual, decidimos mantener estos valores sin alteración por las siguientes razones:

- Falta de información verificable: No contamos con información confiable para imputar de manera precisa la marca o el modelo del vehículo. Realizar una

asignación basada en suposiciones podría introducir errores en la base de datos y comprometer la calidad del análisis.

- Evitar sesgos en la interpretación de datos: Cualquier intento de rellenar estos valores con datos aproximados podría generar sesgos en el análisis, afectando estudios estadísticos o decisiones basadas en estos registros.
- Transparencia y trazabilidad de los datos: Mantener el valor como "Valor inválido" permite diferenciar claramente los datos registrados correctamente de aquellos que presentan inconsistencias en su origen.

Posibles soluciones:
El sistema debería incorporar reglas de validación que impidan el ingreso de valores inválidos. Por ejemplo: Solo permitir marcas y modelos de vehículos que existan en un catálogo predefinido y evitar el uso de valores genéricos como "Valor inválido", "-", ".", o espacios en blanco.

5. Automotores sin especificación:

Algunos registros en los campos "automotor_marca_descripcion" y "automotor_modelo_descripcion" aparecen con el valor "sin especificación", lo que indica una falta de información concreta sobre el vehículo. Para mantener la coherencia en la base de datos, decidimos estandarizar estos valores y categorizarlos como "Marca inválida" y "Modelo inválido".

Esta decisión se fundamenta en la necesidad de evitar ambigüedades y facilitar el análisis de datos. Al utilizar una nomenclatura uniforme, podemos identificar con mayor claridad los registros incompletos y establecer estrategias para su posterior corrección o exclusión en determinados análisis.

6. Valores cero/nulos:

En el atributo "automotor_modelo_codigo", identificamos registros con valores iguales a 0 y otros con campos vacíos. Para garantizar la coherencia en la base de datos, decidimos estandarizar estos casos convirtiendo los valores 0 en valores nulos (vacíos).

Justificación:

- Evitar interpretaciones incorrectas: Un valor 0 en este contexto no representa un modelo válido, sino una falta de información.

- Consistencia en la limpieza de datos: Al tratar los valores 0 y vacíos de la misma manera, aseguramos uniformidad en el tratamiento de datos faltantes.

- Mejora en el análisis: Los valores nulos pueden ser identificados fácilmente para su imputación o exclusión en procesos de análisis, evitando sesgos en los resultados.

Esta estandarización permite un manejo más preciso de los datos, facilitando futuras correcciones y análisis.

7. No declarado/Vacío:

En el atributo "automotor_uso_descripcion", identificamos registros con el valor "NO DECLARADO" y otros con campos vacíos o nulos. Para garantizar la uniformidad en la base de datos, decidimos reemplazar los valores vacíos con "NO DECLARADO".

Justificación:

- Estandarización de datos: Unificar la manera en que se representan los datos faltantes facilita el análisis y evita interpretaciones ambiguas.

- Mejor legibilidad: Un campo vacío puede generar confusión, mientras que "No declarado" indica explícitamente que la información no fue proporcionada.

- Consistencia en la limpieza: Mantener un criterio homogéneo ayuda a futuras imputaciones y análisis sin afectar la integridad de los datos.

En el atributo "titular_domicilio_localidad", encontramos registros con valores que contienen guiones, otros vacíos o nulos. Para mantener la coherencia en la base de datos, decidimos reemplazar estos valores con "NO DECLARADO".

Justificación:
Tenemos la misma justificacion que el punto anterior. Estandarización de datos para unificar los valores en un único término, como "NO DECLARADO", facilitando el análisis ya que los valores vacíos o con guiones pueden ser interpretados de manera errónea, mientras que "NO DECLARADO" aclara que la información no fue proporcionada, mejorando la comprensión de los datos.

8. No identificado:

En los atributos "titular_genero" y en “titular_pais_nacimiento”, encontramos los valores "no aplica" y "no identificado". Para estandarizar los datos y mejorar la consistencia, decidimos unir los valores "no aplica" y "no identificado" bajo un único término: "no identificado".

Justificación:

- Estandarización de datos.
- Mejor consistencia.
- Mejor legibilidad.

9. País y Provincia:

En el atributo "titular_pais_nacimiento", identificamos un valor que indica "Córdoba, Argentina". Dado que Córdoba es una provincia dentro de Argentina, decidimos unificar este valor como "Argentina" para mantener la consistencia en la base de datos.

Justificación:

1. Uniformidad de datos: La provincia Córdoba ya está representada en el atributo de provincia. Al asignar "Argentina" al atributo de país, evitamos la redundancia y aseguramos que el país se registre correctamente.

2. Mejor claridad: Al estandarizar el campo, evitamos posibles confusiones que podrían surgir al tener información repetida o inconsistente.

3. Facilidad en el análisis: La consolidación de los valores del país en una única categoría facilita el análisis posterior, sin distorsionar los registros con información innecesaria.

10. Porcentaje de Titularidad:

En el atributo "titular_porcentaje_titularidad", se observó que los valores varían desde 8 hasta 6977. Sin embargo, al analizar estos valores, encontramos que algunos superan ampliamente el 100%, lo que es inconsistente con la definición estándar de un porcentaje. Un porcentaje representa una fracción de un total, y por convención, no puede exceder el 100%.

Decisión tomada:

Para mantener la coherencia de los datos y garantizar la integridad de la base de datos, podemos aplicar una regla de estandarización en la que: Los valores mayores a 100 sean ajustados automáticamente a 100 para que representen un porcentaje completo de titularidad.

Justificación de la decisión:

- Lógica de porcentaje: Un porcentaje representa una parte de un total, y cualquier valor superior a 100 es inconsistente con esta definición. El supuesto que tomamos es que cualquier valor superior al 100% fue un error de entrada de datos o una interpretación incorrecta de los mismos.

- Estandarización y consistencia: Al limitar los valores a un rango de 0 a 100, garantizamos que todos los registros sigan la misma lógica y sean fácilmente interpretables sin la necesidad de realizar correcciones adicionales en el análisis.

- Facilidad para análisis posteriores: Al unificar estos valores en un rango de 0 a 100, se facilita la comparación y análisis de los datos.

Sin embargo, es importante aclarar que, dado que los datos provienen de la Dirección Nacional del Registro de la Propiedad Automotor (DNRPA), y los valores observados no siguen una convención de porcentaje común, debemos investigar más a fondo el significado de estos valores antes de realizar ajustes definitivos. Es posible que estos números no representan porcentajes de titularidad tradicionales, sino que corresponden a un sistema de codificación o formato específico utilizado por la DNRPA.

Así que antes de llevar a cabo nuestra lógica, vamos a consultar con la DNRPA o revisar su documentación para obtener una explicación clara sobre cómo se estructuran estos valores y si representan alguna medida distinta a un porcentaje.Si se confirma que estos valores representan realmente porcentajes, podremos justificar los cambios. En caso contrario, ajustaremos nuestro enfoque en base a la verdadera naturaleza de los datos.

Transformación de datos:

Se agregaron tres columnas adicionales para descomponer la variable de fecha en: **día, mes y año**, tal cual lo pide el punto N°6 del práctico, en la cual se utiliza la fórmula TEXTO, para extraer la información del día, mes y año de la fecha indicada.

Evaluación del proceso de recolección de datos:

- En primer lugar, al evaluar nuestro dataset, pudimos darnos cuenta que es un formulario en el cual diferentes funcionarios lo completaron, tal vez en un principio en físico, para luego ser pasado a un sistema en línea en el cual se lo pueda consultar, lo cual se puede evidenciar, por las inconsistencias antes mencionadas, como por ejemplo, el ingresar el año con 4 dígitos y en algunos casos con solo 2 dígitos, o también espacios vacíos, guiones, etc.

- Del análisis de datos podemos sugerir las siguientes mejoras, a la hora de la recolección de los datos:

1. Validación y Control en el Ingreso de Datos

- Implementar validaciones en formularios para asegurar la integridad de los datos.
- Usar listas desplegables y opciones predefinidas para evitar errores de escritura.
- Incluir reglas de formato (ejemplo: el año debe ser de cuatro dígitos).

2. Automatización del Proceso de Recolección

- Integración de sistemas (Ej: DNRPA y Policía) para minimizar la entrada manual.
- Uso de APIs para sincronizar datos y reducir duplicaciones o inconsistencias.
- Registro automático de eventos con fecha y hora para mejorar la trazabilidad.

3. Mecanismos de Auditoría y Monitoreo

- Implementar auditorías automáticas para detectar anomalías antes de registrar los datos.
- Utilizar herramientas de monitoreo para identificar patrones inusuales o errores recurrentes.
- Generar reportes de inconsistencias y anomalías para su revisión manual.

4. Manejo de Datos Faltantes

- Aplicar estrategias de imputación adecuadas (media, mediana, regresión, moda).
- Marcar datos faltantes como "No declarado" o "No identificado" para diferenciarlos de errores.
- Evitar asignaciones arbitrarias que puedan sesgar los análisis.

5. Estandarización y Normalización de Datos

- Unificar formatos en descripciones (Ej: "PICK UP" y "PICK-UP" → "PICK-UP").
- Eliminar caracteres innecesarios como espacios en blanco, guiones o signos de puntuación.
- Aplicar reglas de escritura uniforme (Ej: "Córdoba, Argentina" → "Argentina").

6. Consistencia entre Atributos Relacionados

- Asegurar que el país, provincia y localidad sean congruentes en la base de datos.
- Aplicar restricciones para que los valores de ciertas categorías no sean contradictorios.