

Introducción a la ciencia y análisis de datos con Python: una visión empresarial

German Andrés Holguín Londoño

(Pereira, Risaralda, Colombia, 1977)

M.Sc. en Ingeniería Eléctrica e Ingeniero Electricista de la Universidad Tecnológica de Pereira, es Profesor Titular de la Facultad de Ingenierías Eléctrica, Electrónica, Física y Ciencias de la Computación de la misma universidad. Es autor del libro *Principios y métodos combinatoriales en sistemas automáticos digitales* (2021) y ha publicado artículos en revistas especializadas nacionales e internacionales, entre los que destacan: *Enhanced heat tolerance of viral-infected aphids leads to niche expansion and reduced interspecific competition* en *Nature Communications* (2020); *Visual Servoing and Kalman Filter Applied to Parallel Manipulator 3-RRR* en *Electronics* (2024); *Using cameras for precise measurement of two-dimensional plant features: CASS*, en el libro *High-Throughput Plant Phenotyping: Methods and Protocols*, Springer US (2022); y *Pose Estimation of Robot End-Effecter using a CNN-Based Cascade Estimator* en la *IEEE 6th Colombian Conference on Automatic Control (CCAC)* (2023). Es integrante y cofundador del grupo de investigación en Gestión de Sistemas Eléctricos, Electrónicos y Automáticos.

gahol@utp.edu.co

Diego Alejandro Moreno Gallón

(Pereira, Risaralda, Colombia, 1997)

Ingeniero Electricista de la Universidad Tecnológica de Pereira, es profesor catedrático de la Facultad de Ingeniería de la misma universidad y actualmente cursa la Maestría en Ingeniería Eléctrica en la Universidad Tecnológica de Pereira. Ha publicado artículos en revistas especializadas nacionales e internacionales, como *An Industry 4.0 Based Data Analytics Framework for the Detection of Non-Technical Losses in a Smart Grid*, presentado en la *IEEE 6th Colombian Conference on Automatic Control (CCAC)* (2023), y *Analítica de datos para la detección de pérdidas no técnicas en los sistemas eléctricos de distribución inteligentes en el marco de la industria 4.0*, presentado en el V Congreso Internacional de Ingeniería Industrial (2022). Es integrante del grupo de investigación en Gestión de Sistemas Eléctricos, Electrónicos y Automáticos.

diego.moreno@utp.edu.co

Mauricio Holguín Londoño

(Pereira, Risaralda, Colombia, 1974)

Ph.D. en Ingeniería e Ingeniero Electricista de la Universidad Tecnológica de Pereira, es Profesor Titular de la Facultad de Ingenierías Eléctrica, Electrónica, Física y Ciencias de la Computación en la misma universidad. Es autor de los libros *Introducción al control de calidad y Seis Sigma* (2024), *Principios y métodos combinatoriales en sistemas automáticos digitales* (2021), *Pronóstico de vida útil remanente en rodamientos con base en la estimación de la probabilidad de la degradación* (2019), *Fundamentos Teóricos Para Los Autómatas Industriales* (2010) y *Automatismos Industriales* (2008). Ha publicado artículos en revistas especializadas nacionales e internacionales, y es Director del grupo de investigación en Gestión de Sistemas Eléctricos, Electrónicos y Automáticos.

mau.hol@utp.edu.co

Introducción a la ciencia y análisis de datos con Python: una visión empresarial

German Andrés Holguín Londoño

Diego Alejandro Moreno Gallón

Mauricio Holguín Londoño



Facultad de Ingenierías
Colección Textos Académicos
2024

Holguín Londoño, Germán Andrés
Introducción a la ciencia y análisis de datos con Python : una visión empresarial / Germán Andrés Holguín Londoño, Diego Alejandro Moreno Gallón y Mauricio Holguín Londoño. – Pereira : Editorial Universidad Tecnológica de Pereira, 2024.
147 páginas. – (Colección Textos académicos)
e-ISBN: 978-958-722-971-4
1. Ciencia de datos 2. Análisis de datos 3. Redes neuronales 4. Análisis de negocios 5. Fundamentos de Python 6. Estadística en la ciencia de datos.
CDD. 004.3

Introducción a la ciencia y análisis de datos con Python: una visión empresarial

© German Andrés Holguín Londoño

© Diego Alejandro Moreno Gallón

© Mauricio Holguín Londoño

eISBN: 978-958-722-971-4

Universidad Tecnológica de Pereira

Vicerrectoría de Investigaciones, Innovación y Extensión

Editorial Universidad Tecnológica de Pereira

Pereira, Colombia

Coordinador editorial:

Luis Miguel Vargas Valencia

luismvargas@utp.edu.co

Teléfono (606) 313 7381

Edificio 9, Biblioteca Central Jorge Roa Martínez

Cra. 27 No. 10-02 Los Álamos, Pereira, Colombia

www.utp.edu.co

Montaje y producción

Tomás Flórez Calle

Universidad Tecnológica de Pereira

Pereira, Risaralda, Colombia.

Reservados todos los derechos

Contenido

Prefacio	7
CAPÍTULO 1	
Introducción	
1.1. ¿Qué es el análisis de datos?	11
1.2. ¿Qué es la ciencia de datos?	13
CAPÍTULO 2	
Conceptos clave	15
2.1. Limpieza de datos (Data Cleaning)	17
2.2. Análisis estadístico	18
2.3. Parámetros estadísticos principales	19
2.4. Inteligencia artificial	19
2.5. Aprendizaje de máquina	20
2.6. <i>Aprendizaje profundo</i>	21
2.6.1. Data augmentation	22
2.6.2. Redes neuronales adversarias (Generative Adversarial Networks, GAN)	22
2.7. Interfaz de programación de aplicaciones (API)	19

CAPÍTULO 3

Recolección y preparación de los datos	25
3.1. La importancia de obtener datos y prepararlos	28
3.2. Python	30
3.3. Formatos de archivos en el manejo de datos	32
3.4. Ejemplos	33
3.4.1. <i>Formato CSV</i>	33
3.4.2. <i>Formato Excel</i>	36
3.4.3. <i>Formato XML</i>	37
3.4.4. <i>Formato JSON</i>	40
3.5. Pandas	43
3.5.1. <i>Estructuras de datos en Pandas</i>	44
3.5.2. <i>Manipulación de datos con pandas</i>	45
3.5.3. <i>Funciones estadísticas y de agregación</i>	47
3.5.4. <i>Agrupación y pivotaje</i>	48
3.5.5. <i>Combinar DataFrames</i>	48

CAPÍTULO 4

Recolección y preparación de los datos	51
4.1. Definición del flujo de análisis de datos	54
4.2. Importancia de un enfoque estructurado	54
4.3. Componentes de un flujo de análisis	56
4.4. Limpieza de datos	58
4.4.1. <i>Entendimiento general de los datos</i>	59
4.4.2. <i>Limpieza</i>	64
4.5. Análisis estadístico	67
4.5.1. <i>Tendencia central</i>	67
4.5.2. <i>Dispersión</i>	69
4.5.3. <i>Forma</i>	73

4.5.4. Visualización	75
4.5.4.1. Exploración inicial	76
4.5.4.2. Mapas de calor	76
4.5.4.3. Histogramas y diagramas de cajas	76
4.5.4.4. Gráficos de dispersión	76
4.5.4.5. Visualización geográfica	77
4.5.4.6. Visualizaciones interactivas	77
4.5.4.7. Visualizaciones uni-variadas	77
4.5.4.8. Visualizaciones multi-variadas	82

CAPÍTULO 5

Validación y análisis de integridad de bases de datos	93
5.1. Primer conjunto de datos	95
5.1.1. Carga y unión	95
5.1.2. Limpieza	96
5.1.3. Comprensión general de datos	97
5.1.4. Análisis geo-espacial	98
5.2. Segundo conjunto de datos	99
5.2.1. Proceso de carga de datos	99
5.2.2. Proceso de limpieza de datos	100
5.2.3. Análisis general de datos	101

CAPÍTULO 6

Validación y análisis de integridad de bases de datos	109
6.1. Proceso de limpieza	112
6.2. Proceso de comprensión general de datos	116
6.2.1. Cálculo de estadísticos de tendencia central	117
6.2.2. Cálculo de estadísticos de dispersión	118
6.2.3. Cálculo de estadísticos de forma	121

6.2.4. Gráficas uni-variadas	122
6.2.5. Análisis multi-variado	128
CAPÍTULO 7	
Consideraciones	137
7.1. Integración para analítica	140
7.1.1. Experimento	142
7.2. Temáticas relevantes	144
Referencias	149
Lista de figuras	151
Lista de tablas	153
Índice alfabético	155

Prefacio

La actualidad está siendo marcada por una avalancha de datos y una evolución constante en el mundo empresarial. Todos los sectores económicos avanzan a pasos agigantados y el análisis de datos se ha convertido en una herramienta indispensable. Este libro está diseñado para ser una guía introductoria para aquellos en el sector empresarial que buscan adentrarse en el mundo del análisis de datos, proporcionando una comprensión fundamental y ejemplos sencillos de cómo pueden ser utilizados para informar y mejorar las decisiones de negocios.

El primer propósito del libro, busca desmitificar el análisis de datos, presentándolo no como un campo exclusivo para expertos en TI o estadísticos, sino como una competencia accesible y valiosa para una amplia gama de profesionales. Segundo, se enfoca en la aplicación práctica, ofreciendo ejemplos que muestran cómo el análisis de datos puede resolver problemas.

A lo largo de esta obra, se explorará cómo recolectar, limpiar y manipular datos utilizando Python, un lenguaje de programación que se ha establecido firmemente como un estándar en el análisis de datos debido a su simplicidad y potencia. Se introducirá al lector en herramientas como Pandas y otras bibliotecas de Python, que simplifican enormemente este proceso. No obstante, más allá de

la mera técnica, este libro también se sumerge en la interpretación y el análisis crítico, habilidades clave para convertir los datos en perspectivas (*insights*) accionables.

El libro está pensado para ser una introducción práctica y accesible, apta para aquellos sin experiencia previa en programación. Ya sea que usted sea un gerente buscando mejorar la toma de decisiones en su empresa, un profesional de negocios interesado en aprovechar los datos, o simplemente alguien curioso sobre cómo el análisis de datos puede impulsar el éxito empresarial, encontrará en estas páginas un recurso valioso.

1

CAPÍTULO UNO

Introducción

1.1. ¿Qué es el análisis de datos?

Este libro está diseñado para brindar un apoyo esencial a cualquier empresa que busque realizar análisis de datos, y posteriormente, facilitar su camino hacia la ciencia de datos.

Primero, es imprescindible entender qué es el análisis de datos. Este es un proceso que implica examinar, limpiar y transformar datos con el objetivo de obtener información útil, descubrir patrones, identificar tendencias y tomar decisiones informadas. Aunque se puede aplicar en una amplia variedad de campos, en el ámbito empresarial es especialmente crucial. Por ejemplo, una empresa puede utilizar el análisis de datos para identificar patrones en el comportamiento de compra de sus clientes, optimizar sus procesos internos, reducir costos y aumentar la eficiencia operativa. En esencia, se puede utilizar en cualquier área donde se puedan obtener datos, pero su aplicación en los negocios puede ser determinante para lograr una ventaja competitiva en el mercado.

En términos generales, el análisis de datos implica la utilización de herramientas estadísticas, matemáticas y tecnológicas para obtener una plena comprensión del problema y el comportamiento

que está teniendo. Una vez que se lleva a cabo este proceso, se obtienen resultados que son utilizados para realizar *dashboards* (tablas interactivas) y gráficas que sean fáciles de entender, que pueden ser utilizados para mejorar la eficiencia, tomar decisiones empresariales más informadas, o para dar pie al desarrollo de modelos predictivos y analíticos avanzados en ciencia de datos.

El *pipeline* (cadena de procesos) de trabajo de un analista de datos generalmente se compone de los siguientes pasos:

1. **Definición del problema:** entender el problema que se quiere abordar, definir los objetivos de la tarea y determinar qué datos se necesitan para resolver el problema. De esta manera, se define el camino a seguir y se establece el punto de inicio y el punto final.
2. **Recopilación de datos:** recolectar los datos necesarios para abordar el problema. Estos pueden ser recopilados de diversas fuentes, como bases de datos públicas/privadas, archivos físicos, APIs, web scraping, entre otros.
3. **Limpieza y preparación de los datos:** los datos recolectados a menudo pueden contener errores, duplicados o valores faltantes. El analista de datos debe realizar una limpieza y preparación para asegurar que estén listos para el análisis. Esto puede incluir la eliminación de duplicados, la imputación de valores faltantes, el manejo de valores atípicos y la normalización de los datos.
4. **Análisis exploratorio de datos:** realizar un análisis exploratorio de los datos para identificar patrones, tendencias y relaciones interesantes. Esto puede incluir la realización de visualizaciones, estadísticas descriptivas y modelado básico.
5. **Modelado y análisis:** seleccionar y aplicar técnicas de modelado para analizar los datos y resolver el problema en cuestión. Esto puede incluir la aplicación de técnicas de aprendizaje automático, estadísticas inferenciales, análisis de redes, entre otros.
6. **Interpretación y comunicación de resultados:** interpretar los resultados del análisis y comunicarlos de manera efectiva al público objetivo. Esto puede incluir la creación de informes,

dashboards, presentaciones, visualizaciones y la generación de recomendaciones basadas en los resultados del análisis.

7. **Monitoreo y mantenimiento:** monitorear y mantener el análisis para asegurarse de que los resultados sigan siendo precisos y útiles. Esto puede incluir la actualización de los datos, la revisión de los modelos y el monitoreo de los resultados a lo largo del tiempo.

1.2. ¿Qué es la ciencia de datos?

La ciencia de datos es un campo interdisciplinario que combina métodos estadísticos y computacionales para extraer ideas y conocimientos de los datos. Implica la recopilación, el análisis, la interpretación de conjuntos de datos grandes y complejos mediante diversas técnicas, como la minería de datos, el aprendizaje automático y la visualización. Con los procesos descritos, se podrá realizar modelamientos que darán valor agregado dentro de la organización o a los interesados [1].

También, los científicos de datos utilizan una combinación de conocimientos matemáticos y de programación, así como conocimientos específicos del sector, para analizar los datos y extraer conclusiones que puedan apoyar la toma de decisiones y crear herramientas que puedan impulsar los resultados empresariales. Se trabaja con grandes cantidades de datos, tanto estructurados como no estructurados, para identificar patrones, tendencias y desarrollar modelos predictivos que puedan ayudar en la toma de decisiones informadas [1]. La ciencia de datos tiene aplicaciones en una amplia gama de campos, como la sanidad, las finanzas, el marketing, el análisis de redes sociales, entre otros. En los últimos años, se ha convertido en un campo cada vez más importante.

Los campos mencionados son reconocidos desde hace mucho tiempo y anteriormente se llamaba analista a la persona que trataba con los datos. Entonces, ¿por qué ha surgido la ciencia de datos y por qué hay una alta necesidad? Hay tres motivos que explican este fenómeno reciente: Primero, la emergencia de nuevas tecnologías que permiten recoger, etiquetar y guardar una cantidad ingente de datos procedentes de redes sociales, registros y sensores, lleva a cuestionarse qué apli-

caciones prácticas pueden tener esos datos acumulados. Segundo, los progresos en computación, que incluyen métodos innovadores de análisis de datos a escalas crecientes y la disponibilidad de poderosos recursos de computación en la nube, hacen que incluso las pequeñas empresas puedan acceder a estas herramientas de manera sencilla y a bajo costo. Esto, junto con desarrollos recientes en aprendizaje automático, ha impulsado soluciones impresionantes en áreas que llevaban mucho tiempo desafiando a los expertos, como la visión por computadora y el procesamiento de lenguaje natural. Tercero, el impacto real de la analítica de datos moderna ha sido demostrado por grandes compañías tecnológicas y fondos de inversión cuantitativos, y se han convertido en ejemplos inspiradores en campos tan variados como la gestión de deportes y las predicciones electorales, popularizando así la ciencia de datos entre el público general, sin mencionar los grandes modelos de lenguaje que han revolucionado la forma en que interactuamos con la máquina. Todo esto hace que las empresas cada vez necesiten más gente con una cantidad variada de conocimientos multidisciplinares para poder suplir todas las necesidades propias y del mercado.

2

CAPÍTULO DOS

Conceptos clave

Antes de avanzar en el desarrollo del análisis y ciencia de datos, es importante hablar sobre ciertos conceptos clave como *data cleaning* (Limpieza de datos), análisis estadístico, *data augmentation* (aumento de datos), entre otros, que son fundamentales para comprender los temas y argumentos que se presentan. Al definirlos de manera clara y concisa, se busca que el lector tenga una base sólida para la interpretación y comprensión amplia de lo que se discute. Así, se puede profundizar en los temas de manera más efectiva y se puede tener una visión más completa y profunda de las ideas presentadas en el texto.

2.1. Limpieza de datos (*Data Cleaning*)

La limpieza o depuración de datos, conocida ampliamente por su título en inglés (*data cleaning*), es el proceso de corrección de elementos incompletos, duplicados o erróneos en un conjunto de datos. Consiste en identificar los errores del conjunto de datos para luego modificarlos, actualizarlos o eliminarlos y, de esta forma, realizar una corrección. La limpieza de datos mejora la calidad de los

mismos y ayuda a, posteriormente, proporcionar información más precisa, coherente y fiable para la toma de decisiones [2]. Es importante recalcar que tener datos no significa tener información, para poder llegar a tenerla hay que realizarle preguntas adecuadas a los datos y tenerlos lo más limpios posibles (todo este tema se va ir abordando más adelante).

La limpieza de datos es una parte importante en el proceso general de gestión de datos y uno de los componentes principales del trabajo de preparación de los mismos; en ella, se disponen los conjuntos de datos para su uso en aplicaciones de inteligencia empresarial (*Business Intelligence* o *BI* por sus siglas en inglés,) y ciencia de datos. Suele ser realizada por analistas (*data analyst*) e ingenieros de datos (*data engineers*) u otros profesionales de la gestión de datos. Pero los científicos de datos, los analistas de BI y los usuarios empresariales también pueden limpiar datos o participar en el proceso de limpieza de datos para sus propias aplicaciones. En últimas, este proceso prepara el insumo que va a ser usado para generar herramientas que dan un mayor valor agregado a las empresas [2].

2.2. Análisis estadístico

Hay dos enfoques generales para abordar un análisis estadístico: el univariado y el multivariado. El análisis univariado es la forma más sencilla de analizar los datos, como otras formas de estadística, puede ser inferencial o descriptivo; donde el prefijo “uni” se refiere a uno y, por ende, los datos solo tienen una variable. No se ocupa de las causas ni de las relaciones y su principal objetivo es describir; toma los datos, los resume y encuentra patrones en ellos [3, 4, 5, 6].

El análisis multivariado o multivariante, por otro lado, se basa en los principios de la estadística multivariante. Normalmente, el análisis multivariante se utiliza para abordar situaciones en las que se realizan múltiples mediciones en diversas características de cada unidad experimental; las relaciones entre estas mediciones y sus estructuras son importantes. Una categorización moderna y superpuesta del análisis multivariado incluye modelos normales, generales multivariantes, teo-

ría de la distribución, el estudio y la medición de las relaciones, y cálculos de probabilidad, entre otros [3, 4, 5].

2.3. Parámetros estadísticos principales

Un parámetro estadístico, como la media o la desviación estándar, es una característica numérica que resume o describe un aspecto de una población entera. Estos parámetros son fundamentales en el análisis estadístico, pero suelen ser desconocidos ya que obtener datos de todos los miembros de una población es generalmente impracticable. Por lo tanto, recurrimos a estadísticas muestrales — estimaciones derivadas de una selección representativa de la población — para hacer inferencias sobre estos parámetros desconocidos [3, 4, 5].

2.4. Inteligencia artificial

La inteligencia artificial (IA) es un campo que busca emular funciones humanas a través de sistemas computacionales. Según [7, 8], hay diversas definiciones y enfoques sobre lo que es la IA. Estas definiciones se pueden categorizar en dos dimensiones:

1. Las que se preocupan por los procesos de pensamiento y razonamiento.
2. Las que se centran en el comportamiento.

Estas categorías pueden dividirse aún más, según si miden el éxito en función de la fidelidad al desempeño humano o en relación con una medida de rendimiento ideal llamada racionalidad. Un sistema se considera racional si hace lo “correcto” basándose en lo que sabe. Las aproximaciones históricas a la IA han variado, con algunos centrándose en emular el pensamiento humano y otros en lograr comportamientos racionales ideales. La IA ha sido influenciada por diversas disciplinas,

como la filosofía, que se ha preguntado sobre la naturaleza del pensamiento, el origen del conocimiento y cómo el conocimiento lleva a la acción. Personalidades históricas, desde Aristóteles hasta Descartes y Leibniz, han contribuido con ideas que han formado la evolución del concepto detrás de la IA, explorando temas como la lógica, la mecánica del pensamiento y la distinción entre mente y materia.

En resumen, la IA es un campo interdisciplinario que busca comprender, emular y mejorar las capacidades humanas mediante sistemas computacionales, tomando inspiración e ideas de diversas disciplinas históricas y filosóficas.

2.5. Aprendizaje de máquina

El aprendizaje de máquina es una subárea de la inteligencia artificial que se centra en el proceso por el cual los agentes mejoran su rendimiento en tareas futuras basándose en observaciones sobre el mundo. En esencia, se trata de sistemas que pueden aprender y adaptarse a nuevas circunstancias, detectando y extrapolando patrones de datos sin estar explícitamente programados para hacerlo [7, 8].

Existen diferentes formas y enfoques de aprendizaje en función de diversos factores, como el componente del agente que se desea mejorar, el conocimiento previo del agente, la representación utilizada para los datos y el tipo de retroalimentación disponible.

Las principales formas de aprendizaje del agente son:

1. **Aprendizaje no supervisado:** aprende patrones en la entrada sin retroalimentación explícita. Un ejemplo común es el agrupamiento o “clustering”.

2. **Aprendizaje supervisado:** observa ejemplos de entradas y salidas, aprende una función que mapea la entrada a la salida. Es como si tuviera un “profesor” que le proporciona la salida correcta para cada entrada.
3. **Aprendizaje por refuerzo:** aprende a partir de recompensas o castigos, adaptando su comportamiento en función de estas retroalimentaciones.

Además, existen formas intermedias como el aprendizaje semi-supervisado, donde el agente dispone de algunos ejemplos etiquetados y muchos sin etiquetar.

La necesidad del aprendizaje de máquina surge porque los diseñadores no pueden anticipar todas las situaciones posibles en las que un agente puede encontrarse, ni todos los cambios que pueden ocurrir con el tiempo. Además, hay tareas para las cuales los humanos simplemente no saben cómo programar una solución, y la única opción es permitir que la máquina aprenda por sí misma.

2.6. Aprendizaje profundo

El Aprendizaje profundo es una técnica avanzada dentro del aprendizaje automático que permite a las máquinas aprender y discernir patrones a partir de datos crudos. Mientras que los sistemas tradicionales de inteligencia artificial dependían del conocimiento codificado y de características específicamente diseñadas, el aprendizaje profundo desarrolla sus propias características a partir de los datos, un proceso conocido como aprendizaje de representación [8].

Las representaciones aprendidas a menudo ofrecen un rendimiento superior al que se puede obtener con características diseñadas manualmente y permiten a los sistemas de IA adaptarse rápidamente a nuevas tareas con mínima intervención humana. Las técnicas de aprendizaje profundo introducen representaciones que se expresan en términos de otras representaciones más simples, permitiendo a la computadora construir conceptos complejos a partir de conceptos más básicos.

Por ejemplo, un sistema de aprendizaje profundo puede representar la imagen de una persona combinando conceptos más simples, como esquinas y contornos, que a su vez están definidos en términos de bordes.

Un aspecto clave del aprendizaje profundo es su capacidad para desentrañar y discernir los diferentes factores de variación presentes en los datos, construyendo representaciones jerárquicas. Por ejemplo, un modelo profundo puede entender una imagen de un automóvil combinando características simples como bordes para formar contornos más complejos y, finalmente, una representación completa del automóvil.

El ejemplo principal de un modelo de aprendizaje profundo es la red profunda feedforward o perceptrón multicapa (MLP). Esta es una función matemática que mapea un conjunto de valores de entrada a valores de salida, compuesta por la combinación de muchas funciones más simples, cada una de las cuales proporciona una nueva representación de la entrada.

2.6.1. Data augmentation

El aumento de datos (*Data Augmentation*) es un conjunto de técnicas del análisis de datos y la estadística, que son usados para realizar un aumento en los datos obtenidos, al agregar datos que son copias ligeramente modificadas de los ya existentes; en otras palabras, se trata de técnicas que generan más datos a partir de datos existentes. En el área del aprendizaje de máquina (*Machine Learning*), esta técnica se utiliza para reducir el sobreajuste, que ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien a datos nuevos, al realizar entrenamientos. Esto está relacionado al **sobre-muestreo** [9].

2.6.2. Redes neuronales adversarias (*Generative Adversarial Networks, GAN*)

La idea central de una GAN se basa en el entrenamiento indirecto a través del discriminador, otra red neuronal que puede decir lo realista que parece la entrada, que a su vez se actualiza

dinámicamente. Esto significa que el generador no se entrena para minimizar la distancia a una imagen específica, sino para engañar al discriminador. Esto permite al modelo aprender de forma no supervisada [8].

La primera red es la generativa, y como su nombre lo indica, genera candidatos, mientras que la red discriminatoria evalúa. El concurso funciona en términos de distribuciones de datos; normalmente, la red generativa aprende a trazar un mapa desde un espacio a una distribución de datos de interés, mientras que la red discriminatoria distingue los candidatos producidos por el generador de la verdadera distribución de datos. El objetivo de entrenamiento de la red generativa es aumentar la tasa de error de la red discriminatoria [8].

Un conjunto de datos conocido sirve como datos de entrenamiento inicial para el discriminador y con esto mientras se entrena se le presentan muestras del conjunto de datos de entrenamiento hasta que consiga una precisión aceptable. El generador se entrena en función de si consigue engañar al discriminador. Por lo general, el generador se alimenta con datos aleatorios extraídos de un espacio predefinido (por ejemplo, una distribución normal multivariante). Luego, los candidatos sintetizados por el generador son evaluados por el discriminador. Se aplican procedimientos independientes de retro-propagación a ambas redes para que el generador produzca mejores muestras, mientras que el discriminador se vuelve más hábil a la hora de marcar las muestras sintéticas. Cuando se utiliza para la generación de imágenes, el generador suele ser una red neuronal deconvolucional y el discriminador una red neuronal convolucional [8].

2.7. Interfaz de programación de aplicaciones (API)

Las API son mecanismos que permiten que dos o más componentes de software se comuniquen entre ellos mediante un conjunto de definiciones y protocolos. Por ejemplo, el sistema de software de la oficina meteorológica contiene datos meteorológicos diarios. La aplicación meteorológica de un teléfono puede interactuar de manera rápida y sencilla con este sistema a través de las API y puede mostrar las actualizaciones meteorológicas diarias en dicho teléfono [10].

3

CAPÍTULO TRES

Recolección y preparación de los datos

En el corazón del análisis estadístico y la toma de decisiones basada en datos se encuentra un proceso crítico, pero a menudo subestimado: la recopilación y preparación de datos. Este capítulo se dedica a explorar en profundidad estas etapas fundamentales, reconociendo su papel vital en la determinación de la calidad y eficacia del análisis de datos subsiguiente.

La recopilación de datos es mucho más que simplemente reunir información; es una parte fundamental en sí misma que requiere un enfoque metódico y considerado. Los datos recogidos deben ser no solo relevantes y precisos, sino también representativos de la población o fenómeno en estudio. Este capítulo aborda las diversas metodologías y consideraciones implicadas en la recolección de datos, desde la selección de fuentes adecuadas hasta la implementación de técnicas que aseguren la integridad y la fiabilidad de los datos recopilados.

Una vez que los datos son recogidos, el siguiente desafío es prepararlos para el análisis. La preparación de datos es un proceso igualmente crucial, que incluye la limpieza, transformación y organización de los datos en un formato utilizable. A través de ejemplos prácticos y discusiones

detalladas, este capítulo guiará al lector en las mejores prácticas y técnicas para la preparación de datos, garantizando que estén listos para una exploración y análisis más profundos.

Al finalizar este capítulo, el lector tendrá un entendimiento sólido de cómo la recopilación y preparación de datos no son meras tareas preliminares, sino pasos esenciales que definen el camino hacia análisis de datos exitosos y decisiones informadas.

3.1. La importancia de obtener datos y prepararlos

El proceso de recopilación de datos, una piedra angular en el análisis estadístico, implica la meticulosa tarea de reunir y medir información para responder interrogantes de investigación y sustentar decisiones informadas. La calidad de los datos recabados es determinante, ya que influirá directamente en la precisión y la fiabilidad de las conclusiones extraídas. Aquí, la máxima *garbage in, garbage out* cobra especial relevancia, advirtiendo que la calidad del input determina la del output [11].

En el ámbito empresarial, las fuentes de datos son variadas e incluyen registros de ventas, información de clientes, datos financieros, inventarios y métricas de rendimiento, entre otros. La correcta recolección y preparación de estos datos es esencial para obtener *insights* valiosos y tomar decisiones informadas. Sin embargo, es común encontrar malas prácticas en las empresas, como la falta de estandarización en la captura de datos, almacenamiento en silos departamentales, y registros incompletos o inconsistentes. Por ejemplo, diferentes departamentos pueden utilizar formatos o sistemas distintos para registrar información similar, lo que dificulta la integración y el análisis global. Estas prácticas pueden conducir a errores en el análisis y a conclusiones erróneas. Por ello, es fundamental implementar procesos adecuados de recolección y preparación de datos, alineados con los objetivos específicos de la investigación y con la disponibilidad de recursos, para garantizar la calidad y fiabilidad de los datos utilizados en el análisis.

La recopilación de datos de alta calidad es imperativa, requiriendo la definición de criterios claros para la selección de la muestra, el diseño de instrumentos de medición no sesgados y la estan-

rización de procedimientos para la recolección de datos. Cuando se trabaja con fuentes existentes, es crucial validar la metodología de recolección y asegurar su fiabilidad.

El proceso de recopilación puede ser extenso y se debe invertir tiempo considerable en garantizar que los datos sean representativos y precisos. La tarea de recolección y posterior limpieza de datos es esencial, ya que la estructura del análisis y las decisiones que se derivan de este dependen intrínsecamente de la calidad de los datos. Además, es vital recoger los datos de manera ética, respetando la privacidad y autonomía de los participantes [11, 6].

Una vez obtenidos, los datos suelen requerir un procesamiento meticuloso para su preparación analítica. La limpieza de datos implica la corrección de errores y la eliminación de incoherencias, mientras que su transformación asegura la adecuación de los datos para el análisis o la integración de múltiples fuentes en un conjunto coherente de datos.

La eficacia en la recopilación de datos es crucial en el proceso de análisis, justificando la inversión de tiempo y recursos para asegurar la alta calidad y relevancia de los datos en relación con los objetivos investigativos o empresariales. Con datos confiables y verídicos en mano, el paso siguiente es su adecuación y preparación para el análisis.

La preparación de datos abarca su limpieza, tratamiento y transformación. La limpieza se centra en identificar y rectificar errores, valores omitidos y discrepancias. La eliminación de duplicados, la corrección de errores tipográficos o la imputación de valores faltantes son tareas comunes en este proceso [11].

El tratamiento y la transformación de datos adaptan la información a un formato idóneo para el análisis. Esto incluye la combinación de datos de distintas fuentes, la conversión de formatos y la generación de nuevas variables. Las operaciones matemáticas o estadísticas aplicadas durante la transformación de datos pueden revelar nuevas perspectivas o crear variables adicionales, como el cálculo de medias o la normalización de los datos.

Para garantizar una recolección y preparación de datos de alta calidad, se recomienda seguir las siguientes prácticas:

- Definir objetivos claros: Establecer con precisión las preguntas o metas investigativas orienta la recopilación de datos hacia resultados útiles y pertinentes.
- Seleccionar fuentes confiables: Emplear fuentes de datos oficiales o reconocidas contribuye a la precisión y actualización de la información recabada.
- Validar los datos: Una verificación previa al análisis asegura la exactitud y completitud de los datos, incluyendo la revisión de valores faltantes o inconsistencias.
- Cumplir con estándares de calidad de datos: Adoptar normativas como la ISO 8000-1:2022 garantiza la integridad, exactitud y coherencia de los datos.
- Documentar fuentes y métodos: Registrar las fuentes de datos y los procedimientos utilizados en su recolección, limpieza y transformación fomenta la transparencia y reproducibilidad del análisis.

3.2. Python

Python es un lenguaje de programación potente y cada vez más popular, muy adecuado para el análisis de datos. En primer lugar, Python tiene una sintaxis sencilla e intuitiva que es fácil de aprender, incluso para los principiantes. Esto lo convierte en un lenguaje atractivo para los analistas de datos que no tengan una sólida formación en programación. Con Python, los analistas de datos pueden escribir fácilmente scripts para realizar tareas de limpieza, manipulación y visualización de datos. Además, el paradigma de programación orientada a objetos de Python facilita la organización del código y la escritura de funciones reutilizables, lo que puede resultar especialmente útil para tareas complejas de análisis de datos [12].

Otra ventaja de Python es la amplia y activa comunidad de desarrolladores que han creado una gran variedad de bibliotecas y herramientas para el análisis de datos. Esto significa que hay una gran cantidad de recursos disponibles para ayudar a los analistas de datos a iniciarse en Python, así como para resolver problemas y encontrar soluciones. La comunidad Python también ofrece

apoyo a través de foros en línea, grupos de usuarios y conferencias, lo que facilita a los analistas de datos la creación de redes y la colaboración con otros profesionales del sector. Muy posiblemente las cosas que se quieran realizar ya estén publicadas en la comunidad y de esta manera se puede agilizar el trabajo [12].

Python cuenta con una amplia gama de librerías y herramientas muy usadas para el análisis de datos, como NumPy, Pandas, Matplotlib y SciPy. Estas librerías ofrecen una amplia gama de funcionalidades, desde la manipulación y limpieza de datos hasta la visualización y el análisis estadístico. Por ejemplo, Pandas es una popular librería de Python para la manipulación y el análisis de datos, y proporciona un potente objeto de marco de datos para trabajar con datos tabulares. Por su parte, Matplotlib es una librería de Python muy utilizada para crear visualizaciones de datos de alta calidad. Incluso, a partir de Matplotlib se han creado librerías como Seaborn que facilitan la integración con los DataFrames de pandas [6, 11, 12].

Además, Python se puede integrar fácilmente con otros lenguajes y herramientas de programación, lo que lo convierte en una opción ideal para los analistas de datos que trabajan en entornos multilenguaje. Python también puede utilizarse para proyectos de análisis de datos tanto a pequeña como a gran escala. Puede utilizarse para procesar y analizar pequeños conjuntos de datos en un único ordenador, así como para procesar y analizar grandes conjuntos de datos en sistemas distribuidos. Esto convierte a Python en una herramienta versátil y escalable para el análisis de datos.

En general, la facilidad de uso de Python, su comunidad amplia y activa, su amplia gama de librerías y herramientas, su interoperabilidad y su escalabilidad lo convierten en una opción ideal para el análisis de datos. Python es una potente herramienta que puede ayudar a los analistas de datos a procesar y analizar datos, visualizarlos y obtener información para la toma de decisiones en una amplia gama de campos e industrias.

Al buscar datos, se encuentran diferentes tipos de archivos que pueden ser tratados con Python. Algunos de los archivos que pueden aparecer incluyen CSV, Excel, JSON, XML y SQL [6, 11].

3.3. Formatos de archivos en el manejo de datos

En el análisis y la ciencia de datos, es habitual encontrar distintos tipos de archivos, cada uno con sus propias características y usos. Uno de ellos son los archivos CSV (Comma Separated Value o Valores Separados por Comas), que se utilizan para almacenar datos tabulares en formato de texto. Los archivos CSV pueden manejarse fácilmente con Python utilizando el módulo CSV incorporado o librerías de terceros como Pandas.

Otro tipo de archivo habitual en el análisis de datos son los archivos Excel, que son hojas de cálculo creadas con Microsoft Excel. Estos archivos pueden contener varias hojas de cálculo, fórmulas y formatos. Python tiene la librería openpyxl, que permite a los usuarios trabajar con archivos Excel. Los archivos JSON o JavaScript Object Notation son otro tipo de archivo de uso frecuente en el análisis de datos. Los archivos JSON son archivos de texto que utilizan un formato ligero de intercambio de datos, lo que los convierte en una opción ideal para intercambiar datos entre aplicaciones web y almacenar datos de configuración. Python incorpora un módulo json que proporciona métodos para codificar y decodificar datos JSON además se puede utilizar la librería Pandas.

Los archivos XML o Extensible Markup Language son archivos de texto que utilizan un lenguaje de marcado para definir elementos y atributos. Se utilizan habitualmente para almacenar e intercambiar datos entre aplicaciones. Python proporciona el módulo xml.etree.ElementTree para analizar y manipular datos XML o también la antes mencionada librería Pandas.

Los archivos SQL o Structured Query Language contienen consultas o sentencias SQL que se utilizan para manipular datos en una base de datos. Python proporciona varias bibliotecas para interactuar con bases de datos, incluyendo el módulo incorporado sqlite3, así como bibliotecas de terceros como SQLAlchemy, PyMySQL o Pandas.

Python dispone de varias bibliotecas incorporadas y paquetes de terceros para trabajar con estos tipos de archivos. Para leer y manipular archivos CSV, puede utilizar el módulo csv o la popular biblioteca Pandas. Para leer y manipular archivos Excel, puede utilizar la biblioteca openpyxl o Pandas. Para manejar archivos JSON y XML, puedes utilizar los módulos incorporados json,

xml.etree.ElementTree o Pandas. Y para trabajar con archivos DB y búsquedas SQL, se puede utilizar el módulo incorporado sqlite3 o bibliotecas de terceros como SQLAlchemy, PyMySQL y Pandas.

3.4. Ejemplos

Ahora que se ha discutido la importancia de la recopilación de datos y el papel de Python en el análisis de datos, se explorarán algunos tipos de archivos comunes en este contexto y cómo pueden ser manejados mediante el uso de Python. En esta sección, se presentarán ejemplos de cómo trabajar con archivos CSV, Excel, JSON, XML y bases de datos utilizando las librerías integradas de Python y paquetes de terceros.

Al concluir esta sección, se adquirirá una comprensión sólida de cómo leer, manipular y transformar datos en diversos formatos de archivo mediante el uso de Python. Ya sea que se esté abordando conjuntos de datos extensos o más pequeños, la capacidad para trabajar con una variedad de archivos resulta fundamental para cualquier analista de datos o científico. A continuación, se explorará cómo Python puede ser empleado efectivamente en el manejo de estos formatos comunes de archivos en el contexto del análisis de datos.

3.4.1. Formato CSV

Los archivos de valores separados por comas (CSV) son un formato de archivo popular para almacenar e intercambiar datos tabulares, como los datos de las hojas de cálculo. Python proporciona un módulo csv integrado que facilita la lectura, escritura y manipulación de archivos CSV.

Supóngase que se tiene un archivo CSV llamado “datos_ventas.csv” que contiene datos de ventas de una empresa. El archivo tiene cuatro columnas: “fecha”, “producto”, “region_ventas” e “cantidad_ventas”. He aquí cómo podemos utilizar Python para leer el archivo y extraer algunos datos:

Primero se tiene el archivo “datos_ventas.csv” que contiene la siguiente información:

```
fecha,producto,region_ventas,cantidad_ventas
01/04/2022,Ford,Medellín,24
02/04/2022,Pontiac,Medellín,34
03/04/2022,Mitsubishi,Pereira,22
04/04/2022,Mazda,Pereira,45
02/04/2022,Dodge,Bogotá,30
02/04/2022,Volkswagen,Bogotá,16
01/04/2022,Lexus,Bogotá,14
03/04/2022,Cadillac,Bogotá,1
03/04/2022,Lexus,Medellín,28
01/04/2022,Mitsubishi,Medellín,44
02/04/2022,MINI,Bogotá,7
```

Código Python:

```
1 import csv
2 # Abre el archivo CSV
3 with open('datos_ventas.csv', 'r') as file:
4     # Crea un objeto que lee CSV
5     lector = csv.reader(file)
6     # Omite el encabezado
7     next(lector)
8     # Recorre las filas y extrae algunas conclusiones
9     total_ventas = 0
10    for fila in lector:
11        cantidad_ventas = float(fila[3])
12        total_ventas += cantidad_ventas
13    # Imprime el total de ventas
14    print(f"Total de ventas: ${total_ventas}")
```

En este ejemplo, primero se importa el módulo CSV y abrimos el archivo CSV utilizando la función open(). A continuación, se crea un objeto lector CSV utilizando la función csv.reader() y luego se salta a la fila de cabecera utilizando la función next().

A continuación, se recorren las filas del archivo y se extrae algunos datos. En este caso, se calculan las ventas totales sumando la columna “cantidad_ventas” de cada fila. Se convierte el valor de “cantidad_ventas” en un valor flotante utilizando la función float(), ya que los valores se leen inicialmente como cadenas.

Por último, se imprime el valor total de las ventas.

Este es sólo un ejemplo sencillo de cómo trabajar con archivos CSV en Python. Hay muchas otras operaciones que se pueden realizar, como filtrar, ordenar y agregar datos. La librería Pandas también proporciona un potente conjunto de herramientas para trabajar con ficheros CSV y otros formatos de datos tabulares.

Otra forma de acercarse a este problema es con la librería Pandas que carga los datos de csv a un objeto DataFrame que es una tabla tabular. Para realizar el mismo trabajo antes mostrado pero ahora en pandas se haría:

Código Python:

```
1 import pandas as pd
2 # Lee el archivo CSV y lo vuelve un DataFrame
3 df = pd.read_csv('datos_ventas.csv')
4 # Suma todas las ventas
5 total_ventas = df['cantidad_ventas'].sum()
6 # Imprime el total de ventas
7 print(f"Total de ventas: ${total_ventas}")
```

Como se puede ver, Pandas proporciona una forma cómoda, potente de leer y manipular archivos CSV para el análisis de datos. Además logra hacer que el código sea diciente y a la hora de releer es fácil de interpretar qué es lo que está realizando.

3.4.2. Formato Excel

Supongamos que ahora tenemos un archivo de Excel llamado “edades.xlsx” que contiene una hoja llamada “Hoja1” con los siguientes datos:

TABLA 3.1
CONTENIDO DEL ARCHIVO “edades.xlsx”

Nombre	Edad	Género
Alice	24	Mujer
Bob	32	Hombre
Charlie	45	Hombre

Para leer estos datos en un script de Python, se puede utilizar el siguiente código:

```
1 import openpyxl
2 # Cargar el archivo Excel
3 workbook = openpyxl.load_workbook('example.xlsx')
4 # Seleccione la hoja de calculo
5 hoja_calculo = workbook['Hoja1']
6 # Recorrer las filas e imprimir los datos
7 edades = list()
8 for row in hoja_calculo.iter_rows(values_only=True):
9     nombre, edad, genero = row
10    edades.append(edad)
11    print(f"Nombre: {nombre}, Edad: {edad}" +
12          ", Genero: {genero}")
13 print("El promedio de edad es de" +
14 f"{sum(edades)/len(edades)} anos")
```

Aquí, primero se carga el archivo de Excel, la tabla 3.1, utilizando el método `load_workbook()` de `openpyxl`. A continuación, seleccionamos la hoja de cálculo por su nombre utilizando la sintaxis `workbook['Hoja1']` además de declarar la lista en la que guardaremos las edades para su posterior cálculo de media. Por último, recorremos las filas de la hoja de cálculo utilizando el método `worksheet.iter_rows()`, imprimimos los valores de cada fila y luego del ciclo se muestra el promedio. Observe que también pasamos el argumento `values_only=True` al método `iter_rows()`, que le indica que devuelva solo los valores de cada celda y al finalizar, esto optimiza el rendimiento y facilita el procesamiento de los datos.

Usando la librería `openpyxl` en Python, podemos leer fácilmente datos de archivos Excel y realizar varias operaciones sobre ellos según sea necesario. Pero si se usara la librería Pandas se haría de la siguiente forma:

```

1 import pandas as pd
2 # Cargar el fichero Excel en un DataFrame de Pandas
3 df = pd.read_excel('example.xlsx', sheet_name='Hoja1')
4 # Calcular la edad media
5 prom_edad = df ['Edad'].mean()
6 print("El promedio de edad es de:", prom_edad)

```

Para este caso ya no habría la necesidad de usar ciclos `for` para recorrer cada una de las filas y se puede hacer de manera rápida y eficiente el trabajo de análisis de datos.

3.4.3. Formato XML

XML (*Extensible Markup Language*) es un lenguaje de marcado ampliamente utilizado para almacenar e intercambiar datos. Python proporciona varias bibliotecas, como `xml.etree.ElementTree` y `lxml`, que facilitan el análisis sintáctico y la manipulación de datos XML.

Supongamos que usted está trabajando en un proyecto que involucra el procesamiento de datos de un archivo XML que contiene información sobre una lista de libros. Necesita leer los datos XML, extraer información relevante, realizar tareas de manipulación de datos y generar un informe resumido.

Archivo XML: "libros.xml"

```
<libros>
  <libro>
    <titulo>Python Programming</titulo>
    <autor>John Smith</autor>
    <genero>Programming</genero>
    <precio>29.99</precio>
  </libro>
  <libro>
    <titulo>Data Science for Beginners</titulo>
    <autor>Alice Johnson</autor>
    <genero>Data Science</genero>
    <precio>39.99</precio>
  </libro>
  <libro>
    <titulo>Introduction to Machine Learning</titulo>
    <autor>Bob Brown</autor>
    <genero>Machine Learning</genero>
    <precio>49.99</precio>
  </libro>
</libros>
```

El código que se procedería a hacer es el siguiente:

```
1 import xml.etree.ElementTree as ET  
2  
3 # Analisis de datos XML  
4 tree = ET.parse('libros.xml')  
5 root = tree.getroot()  
6  
7 # Extraccion de informacion pertinente  
8 total_libros = len(root.findall('libro'))  
9 total_precio = sum(float(libro.find('precio').text)  
10 for libro in root.findall('libro'))  
11 precio_medio = total_precio / total_libros  
12  
13 # Visualizacion de la informacion extraida  
14 print(f'El total de libros es: {total_libros}')  
15 print(f'El precio total es: ${total_precio:.2f}')  
16 print(f'El precio medio es: ${precio_medio:.2f}')
```

Utilizamos el módulo `xml.etree.ElementTree` para analizar los datos XML y obtener el elemento raíz del árbol XML. Extraemos la información relevante de los datos XML, como títulos de libros, autores, géneros y precios, y realizamos tareas de manipulación de datos, como calcular el número total de libros, el precio total y el precio medio.

Este ejemplo demuestra cómo analizar y manipular datos XML en Python utilizando el módulo `xml.etree.ElementTree`, y cómo realizar tareas de manipulación de datos con los datos utilizando las funcionalidades proporcionadas por el árbol XML. Muestra la versatilidad de Python para trabajar con datos XML en proyectos de análisis y ciencia de datos. Por otro lado, se podría manejar con la librería Pandas realizando el siguiente código:

```
1 import pandas as pd
2 # Lectura del archivo XML en un DataFrame
3 df = pd.read_xml('libros.xml', xpath='libro')
4 # Extraccion de informacion pertinente
5 total_libros = df.shape[0]
6 total_precio = df['precio'].sum()
7 precio_medio = df['precio'].mean()
8 # Visualizacion de la informacion extraida
9 print(f'El total de libros es: {total_libros}')
10 print(f'El precio total es: ${total_precio:.2f}')
11 print(f'El precio medio es: ${precio_medio:.2f}')
```

Como se puede observar en el código en donde se usa pandas ya es mucho más fácil de entender cada una de las líneas de código, además de ser sencilla la manera en la que se accede a la tabla generada por el tipo de formato usado.

3.4.4. Formato JSON

El formato JSON (JavaScript Object Notation) de intercambio de datos muy utilizado, ligero y legible. Python proporciona un módulo integrado llamado JSON que facilita el análisis sintáctico y la serialización de datos JSON.

Ahora supongamos que está trabajando en un proyecto de análisis de datos y ha recibido un archivo JSON que contiene información sobre pedidos de clientes. Necesita analizar los datos JSON para extraer información relevante con la que va tomar decisiones en la empresa.

Archivo JSON: “pedidos_clientes.json”

```
{"pedidos": [ {"pedido_id": 1001, "nombre_cliente": "John Smith", "nombre_producto": "iPhone 12", "cantidad": 2, "precio": 999.99}, {"pedido_id": 1002, "nombre_cliente": "Alice Johnson", "nombre_producto": "MacBook Pro", "cantidad": 1, "precio": 1999.99}, {"pedido_id": 1003, "nombre_cliente": "Bob Brown", "nombre_producto": "Apple Watch", "cantidad": 3, "precio": 399.99}]} }
```

Ahora para resolver el problema se crea el siguiente código en python:

```
1 import json  
2  
3 # Leer datos del archivo JSON  
4 with open('pedidos_clientes.json', 'r') as f:  
5     data = json.load(f)  
6
```

```
7 # Extraccion de informacion pertinente
8 pedidos = data['pedidos']
9 total_pedidos = len(pedidos)
10 total_cantidad = sum([pedido['cantidad']]
11 for pedido in pedidos])
12 total_ingresos = sum([pedido['cantidad'] * pedido['precio']
13 for pedido in pedidos])
14
15 # Visualizacion de la informacion extraida
16 print(f'Total de pedidos: {total_pedidos}')
17 print(f'Cantidad total de ventas: {total_cantidad}')
18 print(f'Total de ingresos: ${total_ingresos:.2f}')
```

Primero se utiliza la función json.load() para cargar los datos JSON del archivo y convertirlos en un objeto diccionario de Python. Se accede a la información relevante del objeto diccionario, como la lista de pedidos, y se realizan tareas de análisis de datos, como calcular el número total de pedidos, la cantidad total vendida y los ingresos totales. Finalmente, se muestra la información extraída utilizando el formato de cadena de Python.

Este ejemplo muestra cómo leer y analizar un archivo JSON en Python, extraer información relevante y realizar tareas de análisis de datos. Destaca la comodidad y flexibilidad de trabajar con datos JSON utilizando el módulo JSON integrado de Python en proyectos de análisis de datos y ciencia de datos. Pero como se ha venido trabajando en los ejemplos, también Pandas logra manejar este tipo de archivos como se puede observar a continuación:

```
1 import pandas as pd
2
3 # Leer datos JSON como un DataFrame
4 df = pd.read_json('pedidos_clientes.json',
5 orient='records')
6
7 # Extracting relevant information
8 total_pedidos = df.shape[0]
9 total_cantidad = df['cantidad'].sum()
10 total_ingresos = (df['cantidad'] * df['precio']).sum()
11
12 # Visualizacion de la informacion extraida
13 print(f'Total de pedidos: {total_pedidos}')
14 print(f'Cantidad total de ventas: {total_cantidad}')
15 print(f'Total de ingresos: ${total_ingresos:.2f}')
```

3.5. Pandas

Como se pudo observar, el módulo Pandas es una potente y versátil biblioteca de Python para el análisis y la manipulación de datos. Proporciona una amplia gama de funcionalidades para trabajar con diferentes tipos de datos, incluyendo archivos CSV, XML, Excel y JSON. Con Pandas, puede leer, escribir y manipular fácilmente datos de estos formatos de archivo populares, y realizar tareas de análisis de datos con facilidad.

Por ejemplo, la función `pandas.read_csv` le permite leer archivos CSV y cargarlos como `pandas DataFrames`, que luego puede utilizar para diversas tareas de análisis de datos. De forma similar, la función `pandas.read_excel` permite leer ficheros Excel y cargarlos como `DataFrames`, facilitando la extracción de información relevante y la realización de tareas de manipulación de datos.

Pandas también proporciona funcionalidades para analizar y convertir datos de un formato a otro y manipular datos XML utilizando la función `pandas.read_xml`, y de igual forma en datos JSON utilizando las funciones `pandas.read_json` y `pandas.to_json`. Estas funcionalidades facilitan el trabajo con estos populares formatos de intercambio de datos para proyectos de análisis y ciencia de datos.

3.5.1. Estructuras de datos en Pandas

Antes de sumergirnos en las funcionalidades específicas, es esencial comprender las dos principales estructuras de datos que Pandas ofrece:

- **Series:** Una estructura unidimensional similar a un array, pero con etiquetas que permiten almacenar cualquier tipo de datos.
- **DataFrame:** Una estructura bidimensional, similar a una tabla de base de datos, una hoja de cálculo de Excel o una tabla de datos en R.

Ejemplo:

```
1 import pandas as pd
2 # Creando una Serie
3 serie = pd.Series([1, 2, 3, 4])
4 print(serie)
5 # Creando un DataFrame
6 data = { 'Nombres': [ 'Ana', 'Juan', 'Luis' ] ,
7          'Edades': [ 25, 30, 22 ] }
8 df = pd.DataFrame(data)
9 print(df)
```

Obteniendo la tabla 3.2 en donde se puede ver la estructura del dataframe que se crea en el bloque de código anterior.

TABLA 3.2
DATAFRAME CREADO

	Edades	Doctor
0	Ana	25
1	Juan	30
2	Luis	22

Primero, se importa la biblioteca pandas con el alias **pd**. A continuación, se crea una Serie que es básicamente un arreglo unidimensional que puede contener cualquier tipo de dato. Esta serie tendrá índices asignados automáticamente que empiezan desde 0. Luego, se crea un **DataFrame**, que es una estructura bidimensional (similar a una tabla). Utilizamos un diccionario para definir las columnas y sus respectivos valores.

3.5.2. Manipulación de datos con pandas

La manipulación de datos es una etapa crucial en cualquier proceso de análisis de datos. Es aquí donde los datos crudos se transforman, limpian y estructuran para su posterior análisis. Pandas, con su versatilidad y funcionalidades, se ha establecido como una herramienta esencial para esta tarea. Proporciona un conjunto de herramientas que permiten la selección, filtrado, ordenación y modificación de conjuntos de datos de una manera intuitiva y eficiente.

En el contexto de pandas, un DataFrame es la estructura central que permite al usuario interactuar con los datos. Consta de filas y columnas, donde cada columna puede ser de un tipo de dato diferente. A través de simples comandos, es posible seleccionar columnas o filas específicas, filtrar datos según ciertos criterios, ordenar el conjunto según valores determinados, entre otras operaciones. Estas acciones son fundamentales para preparar los datos para análisis posteriores o para obtener *insights* (aspectos útiles) inmediatos de ellos. A continuación, se va a explorar algunas de estas operaciones clave en detalle.

Cuando se tiene un conjunto de datos dentro de un DataFrame, pandas ofrece múltiples formas de interactuar con él. Una tarea común es seleccionar columnas o filas específicas. Por ejemplo, para seleccionar solo la columna “Edades” o quizás un subconjunto de columnas como “Nombres” y “Edades”, se utiliza una notación simple y directa. Además, si se desea obtener una fila específica basada en su índice, se puede hacer uso del método `loc`. Por ejemplo, para obtener la primera fila que corresponde al índice 0, se utiliza “`df.loc[0]`”.

```
1 # Seleccionar una columna  
2 edades = df [ 'Edades' ]  
3 # Seleccionar multiples columnas  
4 subset = df [ [ 'Nombres' , 'Edades' ] ]  
5 # Seleccionar filas por indice  
6 primera_fila = df . loc [ 0 ]
```

El filtrado es otra operación esencial. Imaginemos que solo queremos observar registros de personas mayores de 25 años. En pandas, este filtrado se logra con una sintaxis que, aunque es concisa, es increíblemente poderosa. Además, si se desea reorganizar el conjunto de datos según ciertos criterios, como ordenar por edad, pandas proporciona funciones como `sort_values`.

```
1 # Filtrar por valores de una columna  
2 mayores_de_25 = df [ df [ 'Edades' ] > 25 ]
```

En el proceso de manipulación, a menudo se necesita agregar o modificar columnas. Por ejemplo, si se quisiera agregar una columna “Profesión” al DataFrame, se puede hacer de manera directa asignando una lista de valores a una nueva columna. Como se puede observar en la tabla 3.3 aparece una nueva columna con el nombre dado y cada valor entregado en la lista se ubica en su respectiva fila. Por último, en muchos conjuntos de datos, se pueden encontrar valores faltantes o nulos. Pandas ofrece métodos como `fillna` para tratar estos valores, permitiendo, por ejemplo, reemplazar cualquier valor nulo con un 0.

```

1 # Ordenar por una columna
2 df_ordenado = df.sort_values(by='Edades', ascending=False)
3 # Agregar columna
4 df['Profesion'] = ['Ingeniera', 'Doctor', 'Abogado']

```

TABLA 3.3
DATAFRAME CON COLUMNA EXTRA

	Nombres	Edades	Profesión
0	Ana	25	Ingeniera
1	Juan	30	Doctor
2	Luis	22	Abogado

3.5.3. Funciones estadísticas y de agregación

Una vez que se ha estructurado y limpiado el conjunto de datos, es posible que se desee obtener información estadística sobre él. Pandas proporciona el método **describe**, que da un resumen estadístico rápido de todas las columnas numéricas, mostrando medidas como el promedio, mediana, valores máximo y mínimo, entre otros. Además, si se desea realizar operaciones de agregación específicas, como calcular la suma total o el promedio de una columna particular, se pueden usar métodos como **sum** y **mean**.

```

1 # Proporciona un resumen estadistico de las columnas numericas
2 print(df.describe())
3 total_edades = df['Edades'].sum()
4 promedio_edades = df['Edades'].mean()

```

3.5.4. Agrupación y pivotaje

En ocasiones, se necesita agrupar datos según ciertos criterios para realizar análisis más detallados. En pandas, esto es similar a la operación GROUP BY en SQL. Por ejemplo, si se quisiera conocer el promedio de edades por profesión, se puede usar el método **groupby**. Además, para reorganizar los datos de una forma que facilite ciertos análisis o visualizaciones, pandas ofrece operaciones de pivotaje. Esto permite, por ejemplo, transformar los datos para que los nombres sean índices, las profesiones sean columnas y los valores sean las edades correspondientes, esto pudiéndose realizar a partir del método **pivot**.

```

1 # Agrupar por una columna y calcular la media de otra columna
2 promedio_por_profesion = df.groupby('Profesion')['Edades'].mean()
3 df_pivot = df.pivot(index='Nombres',
4                      columns='Profesion',
5                      values='Edades')

```

El resultado de este pivote de la tabla es la tabla 3.4 donde se puede observar que la estructura cambia y habla de la edad media que tiene cada nombre para una profesión específica, de esta forma, se cambia la estructura de la tabla original y encontrando nuevas relaciones.

TABLA 3.4
DATAFRAME LUEGO DE PIVOTEADO DE TABLA

Nombres	Abogado	Doctor	Ingeniera
Ana	NaN	NaN	25.0
Juan	NaN	30.0	NaN
Luis	22.0	NaN	NaN

3.5.5. Combinar DataFrames

Finalmente, en muchos proyectos de análisis de datos, se trabaja con múltiples fuentes de datos que necesitan ser combinadas. Imaginemos que se tiene un DataFrame con nombres y edades y

otro DataFrame con nombres y salarios. Con pandas, combinar estos conjuntos de datos es sencillo usando funciones como **merge**, que permite, por ejemplo, unir los dos DataFrames basados en la columna “Nombres” y se puede observar el resultado en la tabla 3.5.

Con estas herramientas y funcionalidades, pandas se convierte en una piedra angular para cualquier científico de datos que busque manipular, transformar y analizar datos de manera efectiva en Python.

```

1 df2 = pd.DataFrame({'Nombres': ['Ana', 'Juan', 'Pedro'],
2                     'Salarios': [50000, 60000, 55000]})  

3 df_combinado = pd.merge(df, df2, on='Nombres', how='outer')
```

TABLA 3.5
DATAFRAME LUEGO DEL MERGE

	Nombres	Edades	Profesión	Salarios
0	Ana	25.0	Ingeniera	50000.0
1	Juan	30.0	Doctor	60000.0
2	Luis	22.0	Abogado	NaN
3	Pedro	NaN	NaN	55000.0

En resumen, pandas es un módulo muy útil para el análisis de datos ya que proporciona funcionalidades para trabajar con cuatro tipos de ficheros muy utilizados como son CSV, XML, Excel y JSON. Simplifica el proceso de lectura, escritura y manipulación de datos a partir de estos formatos de archivo, y permite centrarse en las tareas reales de análisis de datos, en lugar de en las tareas de carga y análisis sintáctico de datos. Aunque dependiendo del caso, se podría llegar a necesitar herramientas para crear módulos propios para leer datos que fueron estructurados de maneras extrañas.

4

CAPÍTULO CUATRO

Flujo de análisis de datos

En la era de bases de datos masivas, las organizaciones encuentran la necesidad de aprovechar los datos para tomar decisiones informadas y obtener una ventaja competitiva. Sin embargo, extraer información valiosa de grandes cantidades de datos requiere un enfoque estructurado y sistemático del análisis de datos. El flujo de análisis de datos es un marco de referencia que describe los pasos y componentes clave que intervienen en la transformación de los datos en bruto en información procesable que pueda servir de base a las estrategias empresariales e impulsar el crecimiento.

En este capítulo, se profundizará en los detalles del flujo de análisis de datos, explorando las mejores prácticas, técnicas y herramientas para implementar un proceso de análisis de datos racionalizado en su organización. Desde la integración de datos hasta la selección de modelos, la validación y la interpretación de resultados, cubriremos cada etapa del proceso, proporcionando orientación y ejemplos prácticos para ayudarle a gestionar eficazmente sus esfuerzos de toma de decisiones basados en datos.

Además, examinaremos la importancia del análisis iterativo y la mejora continua para mantener la relevancia y precisión de sus conocimientos a medida que su organización evoluciona y se adapta

a nuevos retos. Al fomentar una cultura de aprendizaje continuo, puede asegurarse de que sus esfuerzos de análisis de datos sigan siendo ágiles y respondan a las necesidades cambiantes del negocio.

Al final de este capítulo, comprenderá en profundidad el flujo de análisis de datos y las habilidades necesarias para aplicarlo de forma eficaz en su organización. Armado con este conocimiento, estará bien equipado para navegar por las complejidades del análisis de datos, optimizar sus procesos de toma de decisiones basados en datos y, en última instancia, aprovechar el poder de los datos para impulsar el éxito y el crecimiento de su negocio.

4.1. Definición del flujo de análisis de datos

El flujo de análisis de datos es un enfoque estructurado que permite a las organizaciones navegar por el complejo proceso de transformación de datos brutos en información práctica. Siguiendo una secuencia bien definida de pasos, las empresas pueden garantizar que sus esfuerzos de análisis de datos sean eficientes, eficaces y alineados con sus objetivos estratégicos generales. Establecer un sólido flujo de análisis de datos no sólo simplifica el proceso de trabajar con datos, sino que también ayuda a maximizar el valor de los conocimientos generados, contribuyendo en última instancia a la toma de decisiones basada en datos y a la mejora de los resultados empresariales.

Al comprender los fundamentos del flujo de análisis de datos, estará mejor equipado para desarrollar e implementar una estrategia integral de análisis de datos para su organización, garantizando que sus esfuerzos de toma de decisiones basados en datos sean racionales, escalables y alineados con sus objetivos empresariales.

4.2. Importancia de un enfoque estructurado

Llevar un enfoque estructurado del análisis de datos ayuda a las organizaciones a gestionar y aprovechar eficazmente todo el potencial de sus datos. Mediante la implementación de un flujo de

análisis bien definido, las empresas pueden cosechar numerosos beneficios para tomar decisiones basadas en datos. A continuación, analizaremos las principales ventajas de adoptar un enfoque estructurado para el análisis de datos:

Eficacia y coherencia: Un enfoque estructurado garantiza que el proceso de análisis de datos sea coherente y repetible, lo que conduce a una mayor eficiencia y permite a los equipos trabajar más eficazmente en diferentes proyectos. Esta coherencia también garantiza que los datos generados sean fiables y puedan compararse entre distintos análisis.

Mejora de la calidad de los datos: Un flujo de análisis de datos bien definido incluye pasos para la limpieza y el preprocesamiento de datos, que son esenciales para garantizar la precisión y fiabilidad de los resultados del análisis. Al incorporar estos pasos al proceso, las organizaciones pueden minimizar el impacto de los problemas de calidad de los datos en su toma de decisiones.

Mayor colaboración y comunicación: Un enfoque estructurado del análisis de datos facilita la colaboración en los proyectos entre los miembros de equipos de diferentes departamentos y disciplinas. Al tener una comprensión clara de cada etapa del proceso, las partes interesadas pueden aportar más eficazmente su experiencia y compartir ideas, lo que conduce a decisiones mejor informadas.

Escalabilidad: La adopción de un enfoque estructurado permite a las organizaciones ampliar sus esfuerzos de análisis de datos de manera más eficaz. A medida que aumentan el volumen, la variedad y la complejidad de los datos, contar con un flujo de análisis bien definido garantiza que las empresas puedan realizar análisis cada vez más complejos y adaptarse a nuevas fuentes de datos y requisitos.

Trazabilidad y transparencia: Un enfoque estructurado del análisis de datos proporciona una hoja de ruta clara de los pasos que se dan desde los datos brutos hasta los conocimientos procesables. Esta trazabilidad permite a las organizaciones seguir el progreso y los resultados de sus análisis, garantizando la transparencia y permitiendo a las partes interesadas comprender mejor la lógica que subyace a las decisiones basadas en datos.

Al adoptar un enfoque estructurado para el análisis, las organizaciones pueden agilizar sus procesos de toma de decisiones basados en datos, mejorar la calidad de sus conocimientos y gestionar mejor los retos asociados con el manejo de grandes volúmenes de datos. Sabiendo la estructura y flujo de los datos logra hacer que las preguntas que se hacen diariamente para guiar la empresa puedan resolverse de manera efectiva. En última instancia, un flujo de análisis de datos bien definido es esencial para las organizaciones que desean maximizar la obtención de información relevante e impulsar un cambio significativo basado en datos.

4.3. Componentes de un flujo de análisis

Un flujo de análisis de datos sólido consta de varias secciones interconectadas que guían a las organizaciones a través del proceso de transformación de datos sin procesar en información procesable. Al comprender y gestionar eficazmente estos componentes, las empresas pueden garantizar que sus esfuerzos de análisis de datos sean exhaustivos, eficientes y estén alineados con sus objetivos estratégicos.

El primer componente, la recopilación e integración de datos, implica identificar las fuentes de datos pertinentes, adquirir los necesarios e integrarlos en un conjunto de datos cohesionado. Es esencial elegir fuentes que sean fiables, precisas y alineadas con los objetivos analíticos para garantizar la calidad y pertinencia de los conocimientos generados.

Al haber obtenido los datos se puede proceder con la limpieza y el preprocesamiento de estos. Ahora la idea se centra en resolver los problemas relacionados con la calidad de todo lo recolectado, como los valores que faltan, las incoherencias y las imprecisiones. Mediante la aplicación de técnicas de limpieza y preprocesamiento de datos, las organizaciones pueden garantizar que sus conjuntos de datos sean fiables, coherentes y adecuados para el análisis.

Durante la tercera fase, la exploración de datos y la ingeniería de características, los analistas exploran los datos para comprender mejor sus características, identificar patrones y relaciones, y

crear nuevas variables o características que puedan mejorar el análisis. Técnicas como la estadística descriptiva, la visualización y el análisis de correlaciones pueden ayudar a desvelar ideas y fundamentar el posterior proceso de modelización.

Luego se tiene la sección para la selección y validación de modelos, implica elegir las técnicas analíticas más apropiadas para el problema específico y validar su rendimiento. Es posible que los analistas tengan que entrenar y ajustar varios modelos, comparar su rendimiento utilizando diversas métricas y seleccionar el mejor modelo en función de su capacidad para realizar la generalización de la problemática asociada.

Una vez que se ha seleccionado y validado un modelo adecuado, el componente final consiste en interpretar los resultados, obtener información práctica y comunicarla eficazmente a las partes interesadas. Este paso requiere técnicas claras de narración y visualización de datos para garantizar que los conocimientos generados sean fácilmente comprensibles y puedan servir de base para la toma de decisiones basada en datos. Dentro de este apartado entran conceptos de *storytelling* (contar historias), buenas prácticas en la visualización, teoría del color, entre otras.

Un flujo de análisis de datos sólido no es un proceso único, sino un marco iterativo y en continua evolución. Las organizaciones deben revisar y perfeccionar periódicamente sus esfuerzos de análisis de datos, incorporando comentarios, actualizando modelos y fomentando una cultura de aprendizaje continuo para garantizar que sus procesos de toma de decisiones basados en datos sigan siendo ágiles y respondan a las cambiantes necesidades empresariales.

Al incorporar estos componentes a su flujo de análisis de datos, las organizaciones pueden garantizar que sus esfuerzos de análisis de datos sean exhaustivos, eficientes y eficaces. Además, un flujo de análisis de datos sólido permite a las empresas obtener información significativa a partir de sus datos, impulsando la toma de decisiones informadas, de esta manera se promueve el crecimiento y el éxito a corto, mediano y largo plazo.

4.4. Limpieza de datos

La limpieza de datos es un paso fundamental en el proceso de análisis, ya que garantiza la calidad y fiabilidad de los datos utilizados para generar ideas y fundamentar la toma de decisiones. Con el crecimiento exponencial de los datos en el entorno empresarial actual, las organizaciones se encuentran a menudo con una amplia gama de problemas relacionados con la calidad de los datos, como valores que faltan, incoherencias, duplicados e imprecisiones. Descuidar estos problemas puede conducir a resultados sesgados o engañosos, lo que en última instancia socava el valor y el impacto de las iniciativas basadas en datos.

En este capítulo, profundizaremos en los fundamentos de la limpieza de datos, explorando diversas técnicas, mejores prácticas y herramientas de Python para identificar y abordar eficazmente los problemas de calidad. Discutiremos la importancia de la limpieza de datos para garantizar la precisión y fiabilidad de sus análisis, así como los retos y dificultades que las organizaciones pueden encontrar al tratar con conjuntos de datos grandes y complejos. Además, ofreceremos ejemplos prácticos y orientación práctica para ayudar a desarrollar e implantar procesos sólidos de limpieza de datos en su organización.

Al comprender los principios y técnicas esenciales de la limpieza de datos, estará mejor preparado para garantizar la calidad y fiabilidad de sus datos y, en última instancia, aumentar la eficacia de sus esfuerzos de toma de decisiones basadas en datos. Con los componentes limpios y precisos a su disposición, puede aprovechar con confianza el poder del análisis de datos para tomar decisiones informadas, optimizar las estrategias empresariales e impulsar el crecimiento y el éxito de su organización.

De aquí en adelante se utilizará la base de datos libre de airbnb que se encuentra en la plataforma kaggle en el siguiente link <https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata> a partir de dicha base se realizarán los análisis y desarrollos que darán un ejemplo base para sus futuras implementaciones.

4.4.1. Entendimiento general de los datos

Una de las primeras acciones que hay que hacer antes de empezar a limpiar los datos es comprender su estructura. Incluso si ya se tiene una idea de la recopilación y la fusión, debe explorar el conjunto de datos más a fondo para identificar cualquier matiz o problema que pueda haberse pasado por alto durante las fases iniciales. Conocer en profundidad la estructura de los datos, es vital para diseñar una estrategia eficaz de limpieza, que aborde los retos y requisitos específicos de su conjunto de información.

Primero se realiza un Análisis Exploratorio de Datos o también conocido por sus siglas en inglés como **EDA**. Es un paso crucial en el proceso de análisis de datos que implica el uso de una combinación de métodos estadísticos, técnicas de visualización y resumen de datos, para explorar y comprender las principales características, patrones y relaciones dentro de un conjunto de datos. El **EDA** se centra principalmente en obtener información de los datos, identificar anomalías o valores atípicos y detectar posibles problemas de calidad de la información, antes de pasar a un análisis más complejo, como el aprendizaje automático o el modelado estadístico.

Para empezar, se van a usar jupyter notebooks (cuadernos de jupyter) en los que se van a ir realizando paso por paso los desarrollos. Imagine que es un trabajador o dueño de una empresa de bienes raíces o un inversor inmobiliario, y a partir de datos de Airbnb piensa mejorar la estrategia de compra para la empresa. Como ya se realizó la obtención y unificación de la información, ahora el paso a realizar es la importación de los módulos de python y cargar la base de datos.

```

1 import pandas as pd
2 df_raw = pd.read_csv('data_raw/Airbnb_Open_Data.csv')
3 print(df_raw.head())

```

Al inicio se importa el módulo de Pandas para manejar los datos, que en este caso es de valores separados por comas (CSV) y está ubicado en una carpeta llamada “raw_data”. Se utiliza el méto-

do “read_csv” para leer el archivo CSV y guardarla en una variable tipo “DataFrame” (Se puede ver como una especie de tabla de excel de python) y posteriormente se muestran las primeras diez lineas de la tabla. A primera vista se puede observar que hay 26 columnas y se podría ver que contiene cada una. Para ver qué tipo de variable tiene cada columna se usa un método de los dataframes llamado “dtypes” el cual arroja el tipo por cada columna y de dicha forma se obtiene la tabla 4.1. En la tabla mencionada se puede percibir mayormente variables numéricas (del tipo entero y flotante) y el resto parecen ser cadenas de caracteres.

Como se tiene conocimiento del sitio de donde provienen los datos es decir Estados Unidos y específicamente en Nueva York se puede eliminar la columna de “country” y “country_code” ya que no aportarían información de relevancia ya que es algo obvio para el análisis y no hay más comparaciones. De igual forma hay que verificar que no contenga datos extraños como que aparezcan nombres de otras regiones o ciudades que no sean parte del análisis. Dado el caso en el que se tuvieran más países o ciudades y el análisis abarcara todos esos elementos se podría mantener las columnas.

```
1 # Eliminar columnas no necesarias  
2 df_raw = df.drop(["country", "country_code"], axis='columns')
```

Algo extra que se puede apreciar de la tabla 4.1 es que se necesita acondicionar las columnas “price” y “service fee” ya que aparecen como cadenas de caracteres, y específicamente aparecen de la forma \$1,200 por lo que es necesario tratar las columnas para cambiar a un tipo numérico. Para lograr el objetivo se utiliza el atributo “str” que tiene un método de remplazo y se le pasan los caracteres que se quieren reemplazar en particular se está eliminando los caracteres (“\$”, “,”) y los caracteres de reemplazo que para el ejemplo es de eliminación. Luego se guardan los cambios y al final se pasa la columna modificada a numérico.

```

1 import pandas as pd
2
3 cols = ['price', 'service fee']
4
5 for col in cols:
6     df_raw[col] = df_raw[col].str.replace('$', '')
7     df_raw[col] = df_raw[col].str.replace(',', '')
8
9 df_raw[col] = pd.to_numeric(df_raw[col])

```

Al tener una comprensión general de qué columnas se dispone, se puede ahora sacar información general de las variables (columnas) que son del tipo numéricas. Para sacar datos descriptivos rápidos se utilizaría el método “describe” y se obtendrían las tablas 4.2. Las tablas son un resumen estadístico del conjunto de datos de Airbnb, que ofrece una visión general de alto nivel de los atributos clave relacionados con los anuncios en la ciudad de Nueva York. Incluye el recuento, la media, la desviación estándar, el mínimo, el primer cuartil (25 %), la mediana (50 %), el tercer cuartil (75 %) y los valores máximos para cada atributo. Estos estadísticos pueden ayudarnos a comprender la distribución, las tendencias centrales y la dispersión de los datos, que son cruciales para los pasos posteriores del análisis de datos.

```

1 # Tipos de columnas
2 print(df_raw.dtypes)
3
4 # Descriptores generales para columnas
5 # de tipo numerico
6
7 print(df_raw.describe())

```

Los atributos de la tabla son los siguientes

TABLA 4.1
TIPOS POR COLUMNA

Nombre de la Columna	Tipo
id	int64
NAME	object
host id	int64
host_identity_verified	object
host name	object
neighbourhood group	object
neighbourhood	object
lat	float64
long	float64
country	object
country code	object
instant_bookable	object
cancellation_policy	object
room type	object
Construction year	float64
price	object
service fee	object
minimum nights	float64
number of reviews	float64
last review	object
reviews per month	float64
review rate number	float64
calculated host listings count	float64
availability 365	float64
house_rules	object

- id: Identificador único para cada listado.
- host id: Identificador único para cada host.
- lat: Latitud de la ubicación del listado.
- long: Longitud de la ubicación del listado.
- Año de construcción: Año de construcción del inmueble.
- Noches mínimas: El número mínimo de noches requerido para una reserva.

- Número de opiniones: El número total de reseñas de un anuncio.
- Opiniones por mes: El número medio de opiniones que recibe un anuncio al mes.
- Número de valoración: Un número de valoración basado en las reseñas.

TABLA 4.2
PARÁMETROS GENERALES DE LAS COLUMNAS NUMÉRICAS

	id	host id	lat	long	Construction year	minimum nights	number of reviews
count	1.025990e+05	1.025990e+05	102591	102591	102385	102190.	102416
mean	2.914623e+07	4.925411e+10	40.73	-73.95	2012.49	8.14	27.48
std	1.625751e+07	2.853900e+10	0.06	0.05	5.77	30.55	49.513
min	1.001254e+06	1.236005e+08	40.50	-74.25	2003	-1223	0
25 %	1.508581e+07	2.458333e+10	40.69	-73.98	2007	2	1
50 %	2.913660e+07	4.911774e+10	40.72	-73.95	2012	3	7
75 %	4.320120e+07	7.399650e+10	40.76	-73.93	2017	5	30
max	5.736742e+07	9.876313e+10	40.92	-73.70	2022	5645	1024

	reviews per month	review rate number	calculated host listings count	availability 365
count	86720	102273	102280	102151
mean	1.37	3.28	7.94	141.13
std	1.75	1.29	32.22	135.43
min	0.01	1	1	-10
25 %	0.22	2	1	3
50 %	0.74	3	1	96
75 %	2	4	2	269
max	90	5	332	3677

- Número calculado de anuncios de anfitrión: El número de anuncios que un anfitrión tiene en Airbnb.
- Disponibilidad 365: El número de días que un anuncio está disponible para reservar en un año.

La tabla 4.2 ofrece un resumen de las principales características del conjunto de datos, que puede ser útil para identificar problemas de calidad de los datos o tendencias que justifiquen una investigación más profunda. Por ejemplo, el valor mínimo de -1223 para “noches mínimas” y de -10 para “disponibilidad 365” podría indicar errores de introducción de datos o incoherencias que deben abordarse durante la limpieza de datos. Del mismo modo, el alto valor máximo de “número de opiniones” o “recuento calculado de listados de anfitriones” podría sugerir la presencia de valores atípicos o anfitriones populares con muchas propiedades.

Comprender la distribución y las características del conjunto de datos, como se muestra en la tabla 4.2 de estadísticas resumidas, es un primer paso crucial en el proceso de análisis de datos. Nos ayuda a identificar posibles problemas de calidad, como valores atípicos o incoherencias, que pueden afectar a los resultados de nuestro análisis. Un problema común de calidad de los datos es la presencia de valores perdidos o nulos en el conjunto de datos.

Identificar y tratar los valores nulos es vital por varias razones. Los datos incompletos o ausentes pueden dar lugar a resultados sesgados o inexactos al realizar análisis o crear modelos predictivos. Por ejemplo, si falta el valor “Año de construcción” en un número significativo de listados, nuestro análisis de la relación entre la antigüedad del inmueble y el precio puede estar sesgado. Por otra parte, algunos métodos estadísticos y algoritmos de aprendizaje automático no pueden manejar los datos que faltan, e intentar utilizarlos con conjuntos de datos que contienen valores nulos puede dar lugar a errores o comportamientos inesperados. De otro lado, la comprensión de los patrones de omisión en el conjunto de datos puede revelar posibles problemas con el proceso de recopilación de datos o relaciones entre variables que pueden no ser evidentes a primera vista.

4.4.2. Limpieza

Para garantizar la fiabilidad y validez de nuestros análisis, es esencial identificar y tratar los valores nulos en el conjunto de datos. Las estrategias habituales para tratar los datos que faltan incluyen la eliminación, la imputación y el empleo de técnicas de análisis que tengan en cuenta los datos que faltan. La supresión consiste en eliminar del conjunto de datos los registros con valores que faltan y es adecuada cuando la cantidad de datos que faltan es pequeña y no conlleva una pérdida significativa de información. Por otro lado, la imputación rellena los valores que faltan con estimaciones o aproximaciones, como la media, la mediana o la moda del atributo respectivo, o también al utilizar métodos más avanzados como las técnicas basadas en la regresión, k-nearest neighbors o modelos de aprendizaje automático entrenados con los datos disponibles. Por último, algunos métodos estadísticos y algoritmos de aprendizaje automático, como los árboles de decisión o los bosques aleatorios, pueden tratar los datos que faltan sin imputación o eliminación explícitas.

Como parte de la etapa de pre-procesamiento de datos, es crucial evaluar el alcance y la naturaleza de los valores nulos en el conjunto de datos y aplicar las técnicas adecuadas para abordarlos. De este modo, nos aseguramos de que nuestros análisis y modelos posteriores se basan en información completa y precisa, lo que nos permite obtener ideas y recomendaciones más fiables.

En la figura 4.1 se muestra el porcentaje de valores nulos o ausentes (Ratio de nulos %) para cada atributo del conjunto de datos. Esta información es crucial para comprender la exhaustividad de

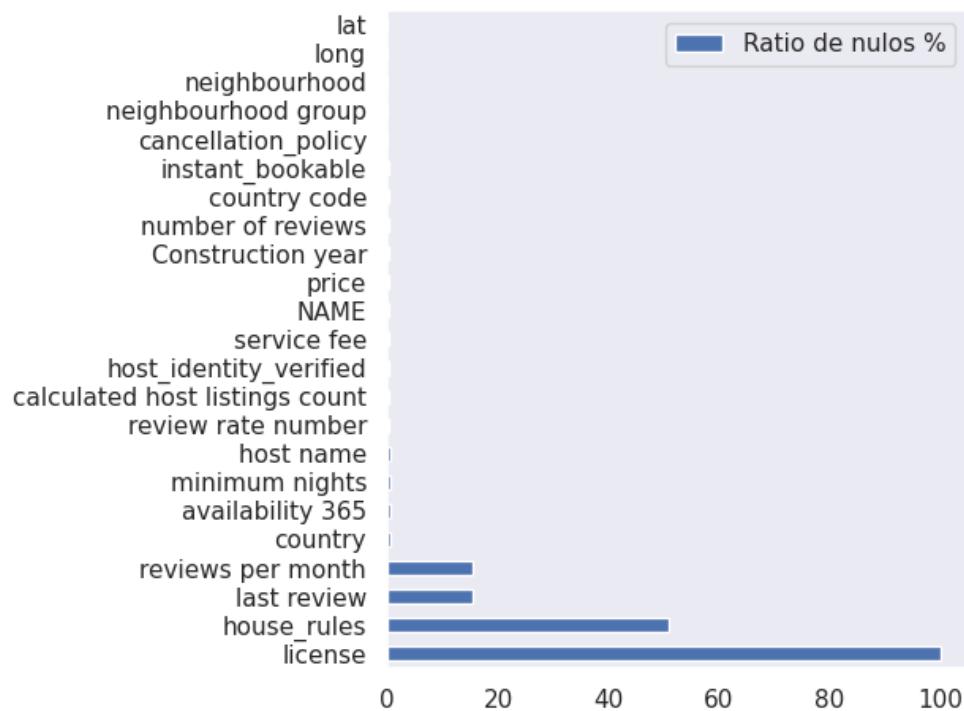


Fig. 4.1: Ratio de nulos por variable

los datos e identificar posibles problemas de calidad de los mismos. El hallazgo más notable es que el atributo “licencia” tiene un porcentaje sorprendentemente alto de valores ausentes, casi el 100 %. Esto indica que casi todos los registros del conjunto de datos carecen de información sobre la licencia, y podría no ser útil para el análisis o la modelización. Por tal motivo lo más efectivo sería eliminar dicha variable de nuestro análisis. Otro atributo con un número significativo de valores nulos es “house_rules”, con aproximadamente un 50.8 % de los registros con valores nulos. Esto podría afectar a los análisis relacionados con las normas internas y podría ser necesario abordarlo

con una eliminación de variable. La eliminación va depender si se va hacer procesamiento de lenguaje natural sobre esta variable y tratar de extraer algún tipo de característica relevante, para el caso de análisis no se va realizar y por ende se va eliminar dicha columna.

Además, “última revisión” y “revisiones por mes” tienen valores perdidos en torno al 15.4 %, lo que puede afectar a los análisis relacionados con las revisiones y los comentarios de los clientes. Los demás atributos tienen un porcentaje relativamente menor de valores perdidos, la mayoría por debajo del 0.5 %. Sin embargo, es esencial tratar adecuadamente estos valores perdidos para garantizar la precisión y fiabilidad de los análisis. Teniendo esto en cuenta se pueden eliminar las muestras que contienen los valores nulos. Para dicho proceso se realiza lo siguiente:

```
1 df = df_raw.copy()
2 # Columnas que mantendremos
3 keep_cols = ['NAME', 'host id', 'host_identity_verified',
4 'host name', 'neighbourhood group', 'neighbourhood',
5 'lat', 'long', 'instant_bookable', 'cancellation_policy',
6 'room type', 'Construction year', 'price', 'service fee',
7 'minimum nights', 'number of reviews', 'last review',
8 'reviews per month', 'review rate number',
9 'calculated host listings count', 'availability 365']
10 df = df[keep_cols]
11 # Eliminacion de muestras con nulos
12 df = df.dropna()
13 print(df.shape[0]/df_raw.shape[0],
14       df.shape[1]/df_raw.shape[1])
15 # Guardar base limpia
16 df.to_csv('data/clean/Airbnb_Open_Data_clean.csv')
17 df.head()
```

De esta manera se estima que se eliminaron 5 columnas y alrededor del 18 % de datos que contenían algún tipo de valor nulo. Además se procede a realizar un guardado de la base de datos al pasar por la limpieza. En este punto se podría ir pensando en que información extra se podría buscar y que esté relacionada con lo que se quiere realizar. Por ejemplo encontrar en un área específica el número de restaurantes, centros comerciales, hospitales, teatros, actividades, etc. Estos datos podrían dar información extra y se tendría que buscar cómo levantar esa información.

4.5. Análisis estadístico

En esta sección nos adentraremos en el proceso de realizar un análisis estadístico exhaustivo del conjunto de datos depurados de Airbnb. El poder del análisis estadístico reside en su capacidad para desvelar patrones ocultos, tendencias y relaciones dentro de los datos, proporcionando en última instancia información valiosa para la toma de decisiones. Mediante el empleo de diversas técnicas y medidas estadísticas, examinaremos la estructura subyacente de los datos, identificaremos características clave y descubriremos relaciones entre variables. Los conocimientos adquiridos a partir de este análisis servirán de base sólida para posteriores exploraciones, incluido el desarrollo de modelos predictivos o la aplicación de estrategias basadas en datos. Esta sección le guiará a través de los pasos esenciales de un análisis estadístico exhaustivo, destacando la importancia de comprender los datos y sus complejidades antes de pasar a técnicas más avanzadas. Antes de empezar a realizar los diferentes análisis estadísticos entraremos a desglosar los diferentes conceptos.

4.5.1. Tendencia central

Media: Dentro de los estadísticos de tendencia central, el más común es la media aritmética ordinaria. Si se hace la consideración de que a la mayoría de conjuntos de datos se les toma como muestras, entonces se habla de media muestral [3, 4, 5]. Si la muestra es de tamaño n , es decir que sus elementos son x_1, x_2, \dots, x_n , la media está definida por la ecuación (4.1).

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (4.1)$$

Mediana: Esta medida de tendencia central indica el punto en el que la muestra es dividida en dos mitades de igual tamaño [3, 4, 5]. Para encontrar ese valor de separación, primero se debe realizar el ordenamiento del conjunto de datos de manera creciente; de este modo, el valor puede ser hallado al realizar las operaciones mostradas en (4.2).

$$x_{mediana} = \begin{cases} X_{[n+1]/2} & \text{si } n \text{ impar} \\ \frac{x_{n/2} + x_{n/2+1}}{2} & \text{si } n \text{ par} \end{cases} \quad (4.2)$$

Moda: la moda es un estadístico que indica la observación que ocurre con mayor frecuencia en la muestra [3, 4, 5].

Aplicación de conceptos: Luego de haber realizado la limpieza de la base de datos se vuelve a realizar el cálculo de los estadísticos principales para ya tener idea de cómo es el comportamiento general de los datos. Sabiendo que en este punto ya se podrían obtener conclusiones que sean más significativas y libres del sesgo.

TABLA 4.3
ESTADÍSTICOS DE TENDENCIA CENTRAL

Estadísticos	Construction year	price	service fee	minimum nights	#reviews
Media	2012.51	626.05	125.21	7.43	32.21
Mediana	2012.00	625.00	125.00	3.00	11.00
Moda	2006.00	206.00	216.00	2.00	1.00

Estadísticos	reviews month	review rate	calculated host listings count	availability 365
Media	1.37	3.28	7.03	141.87
Mediana	0.74	3.00	1.00	101.00
Moda	0.03	5.00	1.00	0.00

Tras el proceso de limpieza de datos, ahora podemos analizar la información estadística actualizada de las propiedades de Airbnb en Nueva York se obtuvo la tabla 4.3. El año medio de construcción de estas propiedades es 2012.51, con un valor mediano de 2012, lo que sugiere que la mayoría de las propiedades se han construido en la última década. El precio medio por noche de estos alojamientos es de 626.05\$, y el precio medio es de 625.00\$. La tarifa de servicio media que se cobra es de 125.21\$, con una tarifa mediana de 125.00\$.

Los huéspedes suelen reservar un alojamiento por una media de 7.43 noches, mientras que el valor mediano indica una estancia más típica de 3 noches. El número medio de opiniones sobre un alojamiento es de 32.21, y el valor mediano es de 11, lo que indica que algunos alojamientos tienen un número significativamente mayor de opiniones, lo que influye en la media. Por término medio, los alojamientos reciben 1.37 opiniones al mes, con un valor medio de 0.74, lo que muestra una tendencia similar. El índice medio de opiniones (una valoración de 1 a 5) es de 3.28, y la mediana es de 3, lo que nos da a interpretar que la mayoría de los alojamientos reciben opiniones moderadamente positivas.

El recuento medio calculado de anuncios de alojamientos es de 7.03, con una mediana de 1, lo que sugiere que algunos alojamientos tienen un número significativo de anuncios, lo que influye en el valor medio. Por último, la disponibilidad media de los anfitriones a lo largo del año es de 141,87 días, y la mediana es de 101 días, lo que indica que algunos anfitriones tienen una mayor disponibilidad, lo que afecta a la media.

Esta información estadística actualizada proporciona una comprensión más clara del mercado de Airbnb en Nueva York y ayuda a tomar decisiones informadas relacionadas con los precios, la duración de las reservas y la disponibilidad de las propiedades.

4.5.2. Dispersión:

Los estadísticos de tendencia central por sí solos no entregan suficiente información a la hora de describir los datos de manera adecuada; por tal motivo, también se debe observar cómo es el comportamiento de la dispersión del conjunto de información [3, 4, 5].

Varianza: La varianza está dada por s^2 de un conjunto de datos y se define como la suma de los cuadrados de las desviaciones respecto a su media y , como es una muestra, se divide por $n - 1$; lo anterior, se puede observar en la ecuación (4.3). Esta medida nos indica qué tan separados están los puntos respecto a la media aritmética, el problema es que para comparar los datos de la varianza tienen unidades elevadas al cuadrado, por tal motivo está la desviación estándar [3, 4, 5].

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \quad (4.3)$$

Desviación estándar: La desviación estándar está dada por la raíz cuadrada de la varianza, como se observa en (4.4). Esta medida se encuentra en las mismas unidades que los datos y , por ende, se pueden comparar directamente [3, 4, 5].

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} \quad (4.4)$$

Cuartiles: En esta medida de dispersión, se tienen en cuenta las posiciones del arreglo de datos ordenado, separando al arreglo en cuatro secciones del mismo tamaño. El primer cuartil separa el primer 25 % de las muestras, y el tercer cuartil el primer 75 %. El segundo cuartil separa el primer 50 %, pero esto vendría siendo lo mismo que obtener la mediana [3, 4, 5].

Rango intercuartílico: Es el valor entre el primer y tercer cuartil, indicando la dispersión entre el primer 25 % y el 75 %. El rango intercuartílico se realiza con la ecuación (4.5) [3, 4, 5].

$$IQR = Q_3 - Q_1 \quad (4.5)$$

Coeficiente de variación: El coeficiente de variación (CV) es una medida estadística de la dispersión de los puntos de una serie de datos en torno a la media, se puede calcular con la ecuación (4.6). Este coeficiente representa la relación entre la desviación estándar y la media; es una estadística

útil para comparar el grado de variación de una serie de datos con otra, incluso si las medias son drásticamente diferentes entre sí; además, permite identificar la representatividad de la media en el conjunto de datos, esto se hace al comparar el valor CV con un límite calculado de comparación. Si el CV es inferior al 20 % entonces la media es representativa, dado el caso contrario, significa que la media no es representativa para el conjunto de muestras [3, 4, 5].

$$CV = \frac{s * 100}{\bar{X}} \quad (4.6)$$

Aplicación de conceptos: Continuando con el análisis ahora se va a calcular los estadísticos relacionados con la dispersión, siendo crucial para una comprensión global de la distribución de los datos y del nivel de variabilidad dentro del conjunto de datos. Las medidas de dispersión, como el rango, la varianza y la desviación estándar, proporcionan información sobre la dispersión y la coherencia de los puntos de datos en torno a los valores de tendencia central (media, mediana o moda). El análisis de las estadísticas de dispersión permite identificar posibles valores atípicos, evaluar la fiabilidad de los datos, comparar diferentes conjuntos de datos.

TABLA 4.4
ESTADÍSTICOS DE DISPERSIÓN

Estadísticos	Construction year	price	service fee	minimum nights	#reviews
Min	2003.00	50.00	10.00	-365.00	1.00
Q2	2007.00	340.00	68.00	2.00	3.00
Q3	2017.00	914.00	183.00	5.00	38.00
Max	2022.00	1200.00	240.00	5645.00	1024.00
std	5.76	331.69	66.34	27.99	51.84
IQR	10.00	574.00	115.00	3.00	35.00
var	33.19	110017.43	4401.16	783.27	2686.94

Estadísticos	reviews month	review rate	calculated host listings count	availability 365
Min	0.01	1.00	1.00	-10.00
Q2	0.22	2.00	1.00	6.00
Q3	2.01	4.00	2.00	266.00
Max	90.00	5.00	332.00	3677.00
std	1.75	1.28	29.47	133.92
IQR	1.79	2.00	1.00	260.00
var	3.05	1.65	868.32	17935.29

La tabla 4.4 proporcionada ofrece estadísticas de dispersión para varias características del conjunto de datos, como el año de construcción, el precio, la tarifa de servicio, las noches mínimas, el número de opiniones, las opiniones por mes, el índice de opiniones, el recuento calculado de listados de alojamientos y la disponibilidad 365. Estas estadísticas nos permiten comprender la dispersión y la coherencia de los datos.

Por ejemplo, el año de construcción oscila entre 2003 y 2022, con un rango intercuartílico (IQR) de 10 años, lo que indica que la mayoría de las propiedades se construyeron en un periodo relativamente reciente. La característica precio tiene un amplio rango de 50 a 1200, con una desviación estándar 331,69 y un IQR mayor 574, lo que sugiere una variabilidad significativa en los precios. Esta variabilidad podría deberse a factores como la ubicación, el tipo de propiedad o los servicios ofrecidos.

La tarifa de servicio también varía, con valores que oscilan entre 10 y 240 y un IQR de 115. La característica de noches mínimas muestra cierta incoherencia, con un valor mínimo negativo -365, lo que podría indicar posibles errores en la introducción de datos que podrían requerir una mayor investigación.

Las características número de revisiones y revisiones por mes muestran amplios rangos, con valores máximos elevados 1024 y 90, respectivamente, y grandes desviaciones estándar 51.84 y 1.75, respectivamente. Esto podría indicar la presencia de algunas propiedades con muchas reseñas, que podrían ser populares o haber estado listadas durante más tiempo.

El índice de valoración tiene un rango relativamente estrecho de 1 a 5, con una desviación estándar de 1.28 y un IQR de 2, lo que sugiere un patrón de valoración más consistente entre las propiedades. El recuento calculado de propiedades gestionadas por anfitriones tiene un rango amplio de 1 a 332 y una desviación estándar alta 29.47, lo que indica variabilidad en el número de propiedades gestionadas por anfitriones.

Por último, la característica disponibilidad 365 tiene un amplio rango de -10 a 3677, con una desviación estándar 133.92 y un IQR 260. El valor mínimo negativo puede apuntar de nuevo a posibles errores en la introducción de datos.

En resumen, las estadísticas de dispersión que figuran en la tabla nos ayudan a comprender la variabilidad y la coherencia de las distintas características del conjunto de datos. Algunas características presentan una variabilidad significativa, mientras que otras muestran una mayor coherencia. Puede que sea necesario investigar más a fondo posibles errores de introducción de datos o valores atípicos para garantizar un análisis y una interpretación precisos de los datos.

4.5.3. Forma

Coeficiente de simetría o *skewness*: El coeficiente de simetría, o tercer momento estadístico, es una medida de la distorsión de la distribución, indicando simetría o asimetría en un conjunto de datos. La asimetría se demuestra en una curva de campana cuando los puntos de datos no se distribuyen simétricamente a la izquierda y a la derecha de la mediana en dicha curva. Si la curva de campana se desplaza hacia la izquierda o la derecha, se dice que está sesgada [3, 4, 5].

La asimetría puede cuantificarse como una representación de la medida en que una determinada distribución varía de una distribución normal. Una distribución normal tiene sesgo cero, mientras que una distribución \log_{normal} , por ejemplo, mostraría cierto sesgo a la derecha. Para realizar el cálculo del tercer momento estadístico se utiliza la ecuación (4.7).

$$CA = \frac{1}{N-1} \frac{\sum_{i=1}^N (x_i - \bar{x})^3}{S^3} \quad (4.7)$$

Curtosis: La curtosis es una medida del tipo de colas de una distribución. La cola es la frecuencia con la que se producen los valores atípicos. El exceso de curtosis es la comparación de la cola en la distribución analizada en relación con una distribución normal. Las distribuciones con curtosis media (colas medianas) son llamadas mesocúrticas, las distribuciones con curtosis baja (colas finas) son platicúrticas, y las distribuciones con curtosis alta (colas gruesas) son leptocúrticas [3, 4, 5].

Las colas son los extremos que se estrechan a ambos lados de una distribución. Representan la probabilidad o la frecuencia de los valores que son extremadamente altos o bajos en comparación

con la media. En otras palabras, las colas representan la frecuencia de los valores atípicos. Se debe tener en cuenta que esta medida representa ambas colas al mismo tiempo y se conoce como el cuarto momento estadístico, que está dado por la ecuación (4.8)

$$K = \frac{1}{N-1} \frac{\sum_{i=1}^N (x_i - \bar{x})^4}{S^4} \quad (4.8)$$

Aplicación de conceptos: Los estadísticos de forma, como la asimetría y la curtosis, proporcionan más información sobre la distribución de los datos. El sesgo mide la asimetría de la distribución, mientras que la curtosis indica las colas de la distribución, lo que nos ayuda a identificar la presencia de valores atípicos o extremos. El análisis de estos estadísticos puede revelar patrones y características subyacentes en los datos, lo que resulta esencial para realizar predicciones precisas y extraer conclusiones significativas.

TABLA 4.5
ESTADÍSTICOS DE FORMA

Estadísticos	Construction year	price	service fee	minimum nights	#reviews
Asimetría	0.004	0.002	0.002	112.29	3.60
Kurtosis	-1.214	-1.19	-1.19	20612.78	22.56

Estadísticos	reviews month	review rate	calculated host listings count	availability 365
Asimetría	7.11	-0.13	7.984	0.68
Kurtosis	221.62	-1.13	72.36	4.47

En la tabla 4.5, los valores de asimetría y curtosis para el año de construcción, el precio y la tasa de servicio son cercanos a 0, lo que sugiere que estas características tienen datos relativamente simétricos, distribuidos normalmente o uniformemente. Sin embargo, la característica Noches mínimas presenta una asimetría positiva muy elevada 112.3 y una curtosis extremadamente alta 20612.78, lo que indica la presencia de valores atípicos y una distribución muy sesgada. Esta observación corrobora el hallazgo anterior de posibles errores de introducción de datos en la característica de noches mínimas.

Las características número de opiniones y opiniones por mes muestran una asimetría positiva 3.6 y 7.11, respectivamente y valores de curtosis de 22.56 y 221.62, lo que indica la presencia de valores atípicos y una distribución de cola larga. Esto sugiere que podría haber algunas propiedades con un número inusualmente alto de revisiones o revisiones por mes.

La tasa de reseñas tiene una asimetría ligeramente negativa -0,13 y una curtosis negativa -1.13, lo que indica una distribución relativamente simétrica con colas ligeras. El recuento calculado de listados de hosts presenta una asimetría positiva elevada 7.98 y una curtosis 72.36, lo que sugiere una distribución de colas largas con posibles valores atípicos. Por último, la característica disponibilidad 365 tiene una asimetría positiva 0.68 y una curtosis moderadamente alta 4.47, lo que indica una distribución algo sesgada con colas más pesadas.

En general, los valores de asimetría y curtosis nos ayudan a comprender mejor la distribución subyacente de cada característica. Mientras que algunas características parecen ser relativamente simétricas y tener una distribución normal o uniforme, otras presentan valores elevados de asimetría y curtosis, lo que indica la presencia de valores atípicos, distribuciones sesgadas o colas pesadas. Esta información puede servir de guía para posteriores procesos de análisis y una segunda limpieza de datos, garantizando una interpretación más precisa y significativa de los mismos.

4.5.4. Visualización

La visualización de datos es un componente esencial de la analítica de datos y una poderosa herramienta para entender, comunicar patrones y relaciones dentro de un conjunto de datos. En el contexto de Airbnb, una empresa global de hospedaje en línea, la visualización de datos puede proporcionar una visión valiosa de una variedad de factores clave, como los precios, la ubicación de los listados, las valoraciones de los clientes y mucho más. En esta sección, exploraremos diferentes técnicas y herramientas de visualización de datos, desde tablas y gráficos hasta mapas de calor, histogramas, diagramas de cajas, gráficos de dispersión y visualizaciones geográficas e

interactivas. Cada una de estas técnicas tiene sus propias fortalezas y puede revelar diferentes aspectos de los datos de Airbnb [13, 14].

4.5.4.1. Exploración inicial

La exploración inicial debe consistir en una descripción general de los datos. Para nuestro conjunto de datos de Airbnb, esto podría implicar mostrar un resumen de las variables, como el número de listados, la ubicación geográfica de los listados, los precios medios, etc. Se pueden utilizar tablas y gráficos de barra para visualizar estos datos [13, 14].

4.5.4.2. Mapas de calor

Los mapas de calor pueden ser muy útiles para identificar correlaciones entre diferentes variables. Por ejemplo, podemos crear un mapa de calor para explorar la relación entre el precio de un listado de Airbnb y otras variables como la ubicación, el número de habitaciones, las instalaciones disponibles, las valoraciones de los clientes, etc.

4.5.4.3. Histogramas y diagramas de cajas

Los histogramas y los diagramas de cajas son herramientas eficaces para visualizar la distribución de los datos. Un histograma de precios de Airbnb, por ejemplo, nos ayudaría a entender la gama de precios en la plataforma, mientras que un diagrama de cajas nos permitiría identificar valores atípicos [13, 14].

4.5.4.4. Gráficos de dispersión

Los gráficos de dispersión pueden ayudar a identificar las relaciones entre dos variables. Por ejemplo, podríamos usar un gráfico de dispersión para explorar la relación entre el precio de un listado y el número de valoraciones que ha recibido [13, 14].

4.5.4.5. Visualización geográfica

Dado que Airbnb opera en todo el mundo, las visualizaciones geográficas pueden proporcionar una visión valiosa. Podríamos usar un mapa para mostrar la densidad de listados en diferentes áreas o para mostrar el precio medio de los listados en cada área [13, 14].

4.5.4.6. Visualizaciones interactivas

Las visualizaciones interactivas pueden ser particularmente útiles para manejar grandes conjuntos de datos. Por ejemplo, podríamos usar una tabla interactiva que permita a los usuarios ordenar y filtrar los listados de Airbnb según diferentes criterios [13, 14].

4.5.4.7. Visualizaciones uni-variadas

Al final, la visualización de los datos es un trabajo complejo, y la elección de las visualizaciones apropiadas dependerá en gran medida de los datos específicos y las preguntas que estemos tratando de responder. Para este caso particular se va a empezar el análisis de las características con gráficas uni-variadas en donde se tiene la frecuencia, el acumulado y un diagrama de caja [13, 14].

En relación a las figuras 4.7.4.2 y 4.3, se observa un comportamiento que se asemeja a una distribución exponencial, con una mayor frecuencia de valores cercanos a cero. En cuanto a los diagramas de caja, es necesario analizarlos individualmente dadas las diferencias notables que se presentan en cada caso. En primer lugar, en el caso de las noches mínimas requeridas para una reserva, el 50 % de los datos se sitúa en valores inferiores a 4. Seguidamente, en lo que respecta al número de reseñas, más de la mitad de los datos se ubican por debajo de 30 reseñas. Finalmente, en el caso de las reseñas por mes, la mitad de los datos se halla por debajo de 1.5. Estas observaciones nos permiten entender mejor el comportamiento de los usuarios y las tendencias de uso del servicio de Airbnb.

En las figuras 4.4, 4.5 y 4.6, observamos una distribución uniforme a través de múltiples variables esenciales. Esto implica que los precios totales y los precios de servicio se distribuyen de manera equilibrada en todo el espectro considerado, lo que refleja una consistencia en los costos

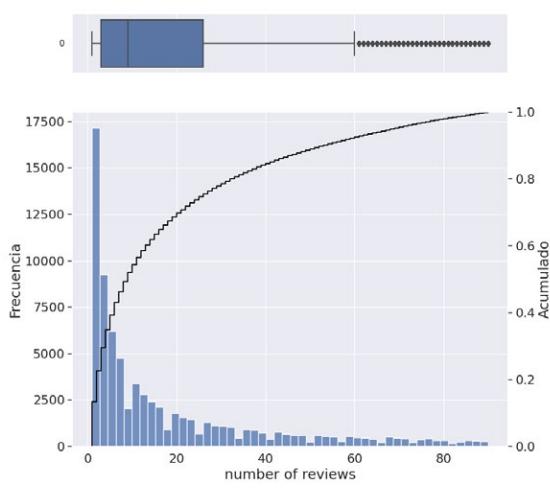


Fig. 4.2: Gráfica uni-variada número de reseñas

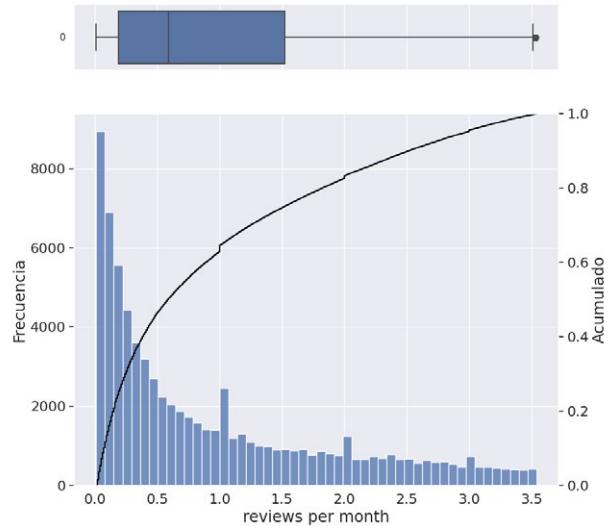


Fig. 4.3: Gráfica uni-variada número de reseñas por mes

a los que se enfrentan los usuarios. Simultáneamente, se aprecia una distribución relativamente uniforme en la construcción de edificios a lo largo del periodo analizado, desde 2003 hasta 2022, lo que sugiere un ritmo constante de desarrollo de infraestructuras durante estos años. Tal como se discutió en la sección donde se calculan las medidas de forma, consulte la tabla 4.5, se había planteado que la distribución podría ser uniforme o normal. A través de estas visualizaciones, logramos identificar que la distribución efectivamente sigue un patrón uniforme, corroborando así nuestras suposiciones iniciales [13, 14].

En lo que respecta al ratio de reseñas y al recuento calculado de listados por anfitrión, ambos toman valores enteros entre 1 y 5. El ratio de reseñas no parece seguir un patrón específico, con la mitad de los datos fluctuando entre 2 y 4. Por otro lado, el recuento calculado de listados por anfitrión exhibe una distribución que se asemeja a la exponencial, con la mayoría de los datos concentrados entre uno y tres. Esta observación sugiere que, en general, los anfitriones tienden a

tener un número reducido de listados, lo que puede ser indicativo de la prevalencia de anfitriones individuales frente a empresas de alojamiento en la plataforma de Airbnb.

```

1 import seaborn as sns
2 import matplotlib.pyplot as plt
3 def univariado_hist(datos, num_bar, x_label=',',

```

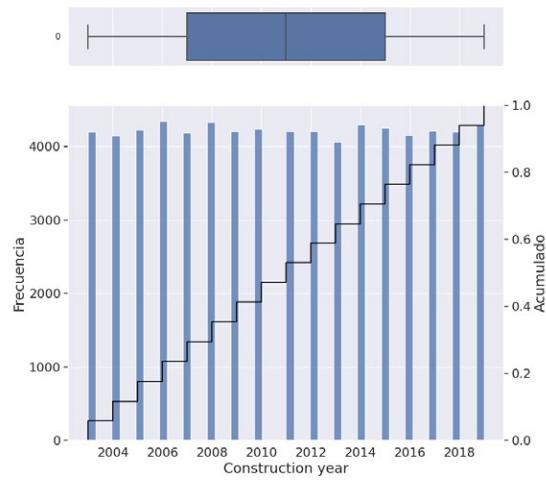


Fig. 4.4: Gráfica uni-variada años de construcción

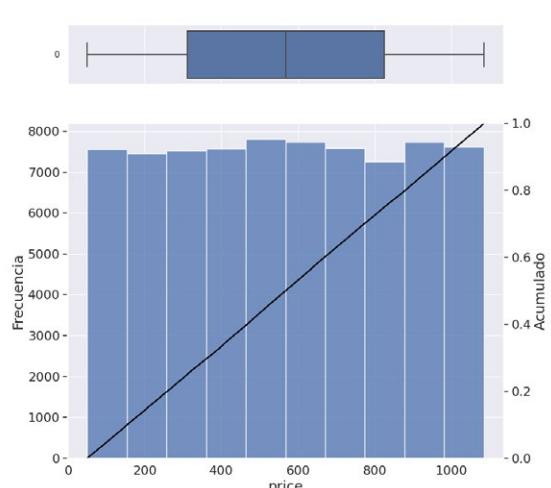


Fig. 4.5: Gráfica uni-variada precio

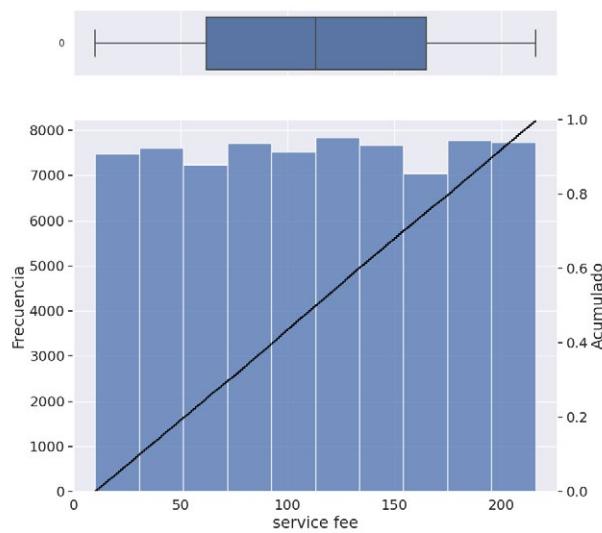


Fig. 4.6: Gráfica uni-variada cuota de servicio

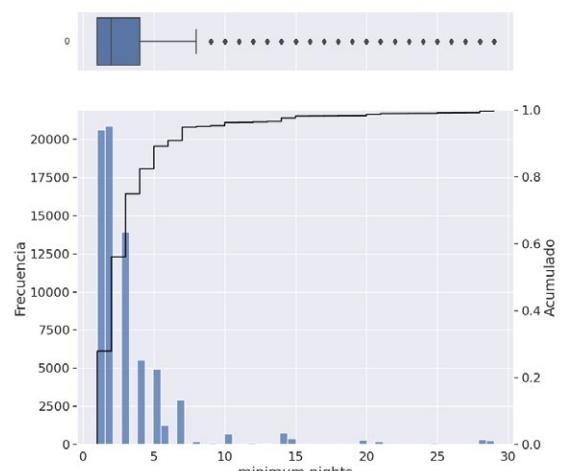


Fig. 4.7: Gráfica uni-variada mínimo número de noches

```
4             y_label='',
5             save_name='default.png'):
6     f, (ax_box, ax_hist) = plt.subplots(2,
7             figsize=(10,10),
8             sharex=True,
9             gridspec_kw={"height_ratios": (.15, .85)})
10    ax_cum = ax_hist.twinx()
11
12    # Asignando la grafica a cada ax
13    sns.boxplot(data=datos, orient='h', ax=ax_box)
14    sns.histplot(data=datos,
15                  bins=int(abs(num_bar)),
16                  ax=ax_hist)
17
18    sns.ecdfplot(data=datos, color='black', ax=ax_cum)
19
20    # Se elimina la etiqueda del eje x para el
21    # grafica de caja
22
23    ax_box.set(xlabel='')
24    ax_box.set(ylabel='')
25
26    ax_hist.tick_params(axis='x', labelsize=16)
27    ax_hist.tick_params(axis='y', labelsize=16)
28    ax_hist.set_ylabel(y_label, fontsize=18)
29    ax_hist.set_xlabel(x_label, fontsize=18)
30
31    ax_cum.tick_params(axis='y', labelsize=16)
32    ax_cum.set_ylabel('Acumulado', fontsize=18)
33
34    plt.grid()
35
36    plt.savefig('images/' + save_name)
37    plt.tight_layout()
38
39    plt.show()
```

```

32 for col in filter_cols:
33     if col != 'availability_365':
34         threshold = df[col].quantile(0.9)
35         filter_one = (df[col] < threshold)
36         filter_two = (df[col] > 0)
37         data = df[filter_one & filter_two].reset_index()
38     else:
39         data = df[df[col] < df[col].max()].reset_index()
40     if df.shape[0] != 0:
41         save_name = col.replace(' ', '_').lower()
42         univariado_hist(data[col], 50,
43                           x_label=col,
44                           y_label='Frecuencia',
45                           save_name=save_name+'.png')

```

El código proporcionado es un ejemplo de cómo se generan las visualizaciones discutidas anteriormente utilizando Python, y más específicamente, los módulos seaborn y matplotlib. Estos módulos son ampliamente utilizados en la ciencia de datos para la visualización gráfica y la exploración de datos. El script comienza importando los módulos seaborn y matplotlib. Luego, define una función llamada “univariado_hist” que crea un histograma univariado con una gráfica de caja (boxplot) y una gráfica de función de distribución acumulada empírica (ecdf) en el mismo marco. Esta función toma varios argumentos, incluyendo los datos a visualizar, el número de barras en el histograma, las etiquetas de los ejes (x, y), y el nombre del archivo donde se guardará la figura.

Dentro de esta función, se inicializan dos subgráficas en el mismo marco: “ax_box” para la gráfica de caja y “ax_hist” para el histograma. También se establece un tercer eje, “ax_cum”, para la gráfica ecdf, que comparte el eje x con el histograma. Se utilizan las funciones “boxplot”, “histplot” y

“ecdfplot” de seaborn para dibujar las gráficas correspondientes en sus respectivos ejes. Luego, se personalizan las etiquetas de los ejes, las fuentes, se guarda y se muestra la figura. Al observar el código anterior se puede observar la facilidad de creación y diseño de las visualizaciones dentro de Python.

4.5.4.8. Visualizaciones multi-variadas

Matriz de varianzas y covarianzas: La covarianza es una medida de la variabilidad conjunta de dos variables aleatorias. Si los valores más grandes de una variable corresponden principalmente con los valores grandes de la variable con la que se compara, y los valores pequeños corresponden entre ambas variables, la covarianza es positiva. En el caso contrario, cuando los valores mayores de una variable se relacionan principalmente con los valores menores de la otra, la covarianza es negativa. El signo de la covarianza muestra la tendencia de la relación lineal entre las variables. La magnitud de la covarianza no es fácil de interpretar porque no está normalizada y, por tanto, depende de las magnitudes de las variables. La covarianza esta dada por la ecuación (4.9) y si se desea la varianza entonces $j = k$ [3, 4, 5].

$$s_{jk} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \text{ donde } j \neq k \quad (4.9)$$

La matriz de covarianza es cuadrada, simétrica y en su diagonal principal se encuentran las varianzas, mientras que fuera de la diagonal están las covarianzas. Para el cálculo de la matriz de covarianzas se utiliza la matriz de datos centrados \tilde{X} que es la matriz de datos en donde cada característica se le resta la media. Para calcular la matriz de covarianza \tilde{S} se aplica la ecuación (4.10).

$$\tilde{S} = \frac{1}{n-1} \tilde{X}^\top \tilde{X} \quad (4.10)$$

Coeficiente de correlación: En estadística, la correlación es toda relación estadística, que sea causal o no, entre dos variables aleatorias o datos multivariados. Aunque de manera general, la correlación puede indicar cualquier tipo de asociación, en estadística comúnmente hace referencia al grado de relación lineal entre variables.

Para determinar la dependencia de las variables aleatorias se debe comprobar si no satisfacen una propiedad matemática de independencia probabilística. Comúnmente, se asocia la correlación como sinónimo de dependencia, sin embargo, cuando se utiliza de manera técnica, hace referencia a que hay una combinación lineal, o no lineal (dependiendo del coeficiente), entre las variables probadas y sus respectivos valores esperados. Puntualmente, la correlación es la medida de qué tanta relación hay entre dos o más variables. Existen varios coeficientes de correlación, que miden el grado de esta. El coeficiente de correlación más común es el de Pearson, que solo es sensible a la relación lineal entre dos variables; aunque, también hay otros tipos de coeficientes que tratan de ser más sensibles ante relaciones no lineales de las variables aleatorias.

Visualizaciones de correlación, varianza y covarianza: Finalmente, el script recorre todas las columnas especificadas en “filter_cols”. Si la columna no es “availability 365”, se filtran los datos para excluir los valores por encima del percentil 90 y los valores igual a 0. Si la columna es “availability 365”, se filtran los datos para excluir el valor máximo. Luego, si el dataframe filtrado no está vacío, se llama a la función “univariado_hist” para crear y guardar una figura para la columna. Este código, por tanto, proporciona una forma eficiente de explorar y visualizar rápidamente las distribuciones univariadas de múltiples variables en un conjunto de datos. Pero todo no se queda en el mundo univariado, también hay que ver las relaciones existentes en diferentes dimensiones para ver el comportamiento entre las diferentes variables un caso particular y muy fácil para observar el comportamiento de los datos en dos dimensiones. Para esto se utiliza la librería seaborn de python y específicamente el método “pairplot” como se podrá ilustrar en la figura 4.8 en donde se pasa el dataframe al método. Para el caso de nuestra base de datos de ejemplo no se logra

observar ninguna correlación directa y es por eso que tocaría afectar los ejes con logaritmo natural para ver si hay relaciones diferentes a la lineal.

```
1 plt.figure()  
2 sns.pairplot(df, diag_kind="hist")  
3 plt.show()
```

El diagrama de dispersión por pares (pairplot) es una herramienta valiosa para visualizar las relaciones entre múltiples variables. Sin embargo, puede no siempre evidenciar de manera clara si existe una relación lineal directa entre las variables. Por lo tanto, es recomendable complementar esta visualización con una matriz de correlación, la cual cuantifica la relación lineal entre pares de variables. Además, un mapa de calor de esta matriz puede ayudar a destacar visualmente las correlaciones más fuertes como se puede percibir en la figura 4.9b, a comparación de la gráfica de la figura 4.9a que puede dar una intuición pero, no indica completamente la relación como se apreciar al comparar las gráficas en la figura 4.9.

Es importante recordar que, aunque este libro se enfoca en la correlación lineal, existen otras medidas de asociación que pueden ser más adecuadas en distintos contextos. Por ejemplo, la correlación de Spearman, la correlación de Kendall, entre otras, pueden captar relaciones no lineales entre las variables. Aunque estas técnicas más avanzadas no se abordarán en detalle en este libro, siempre es útil expandir los horizontes y profundizar en el estudio de estos métodos a partir de las técnicas más sencillas aquí presentadas.

A medida que avanzamos en el análisis de datos, es fundamental recordar que la información contenida en nuestro conjunto de datos trasciende la simple relación bidimensional entre las variables.

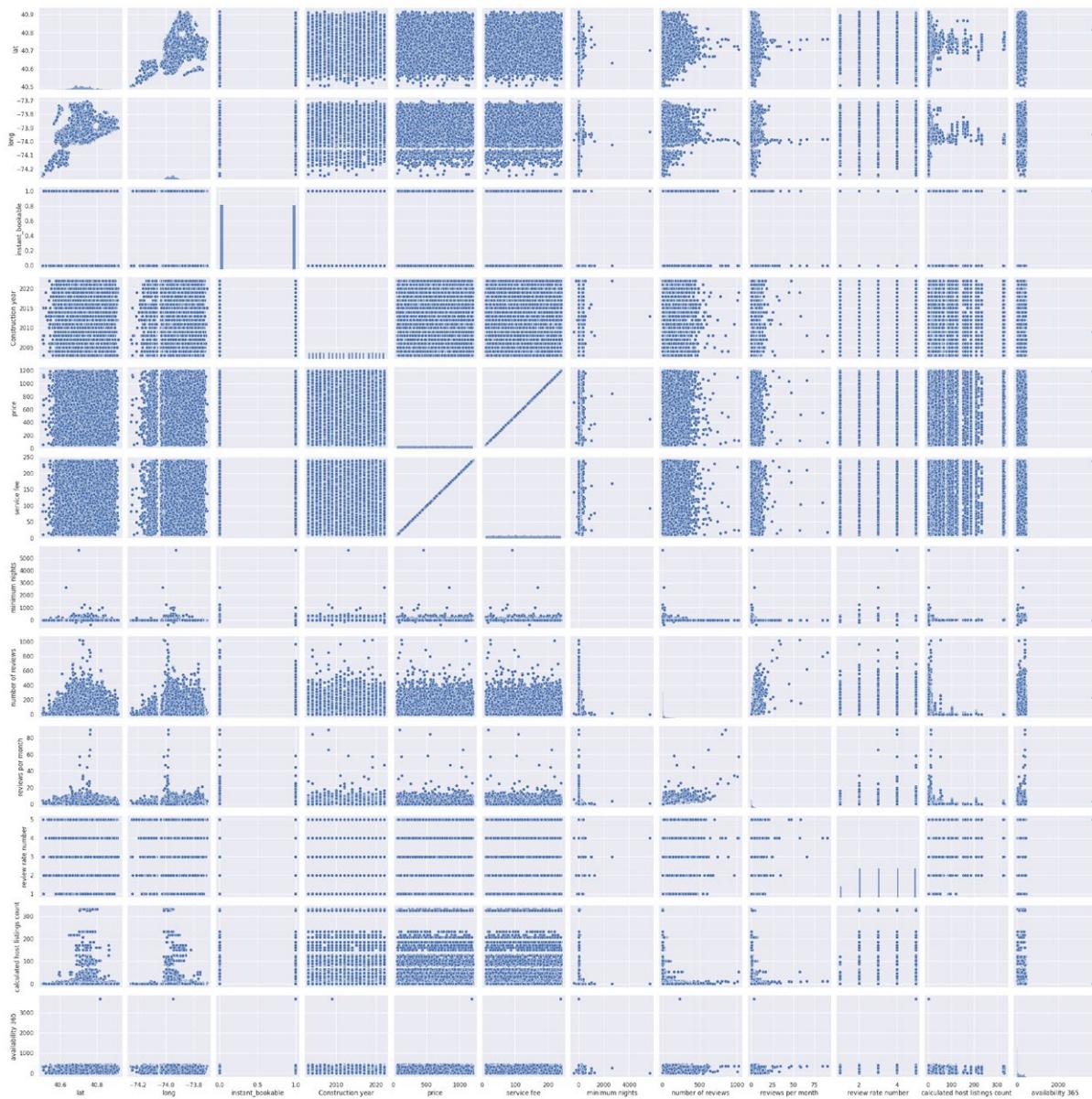


Fig. 4.8: Gráfica multivariada

De hecho, existen interacciones más complejas en un espacio multidimensional que pueden revelar patrones o características interesantes. Una herramienta visual útil para explorar estas interacciones más allá de las dos dimensiones son las Caras de Chernoff. Este método permite visualizar múltiples dimensiones de datos mediante el uso de características faciales humanas modificadas, lo que puede ayudar a identificar similitudes y diferencias entre las variables en estudio.

Análisis de componentes principales: El análisis de componentes principales (PCA) es una técnica muy popular para analizar grandes conjuntos de datos que contienen un elevado número de características (o también denominado dimensiones) por observación, aumentando la interpretabilidad de los datos y conservando la máxima cantidad de información, permitiendo la visualización

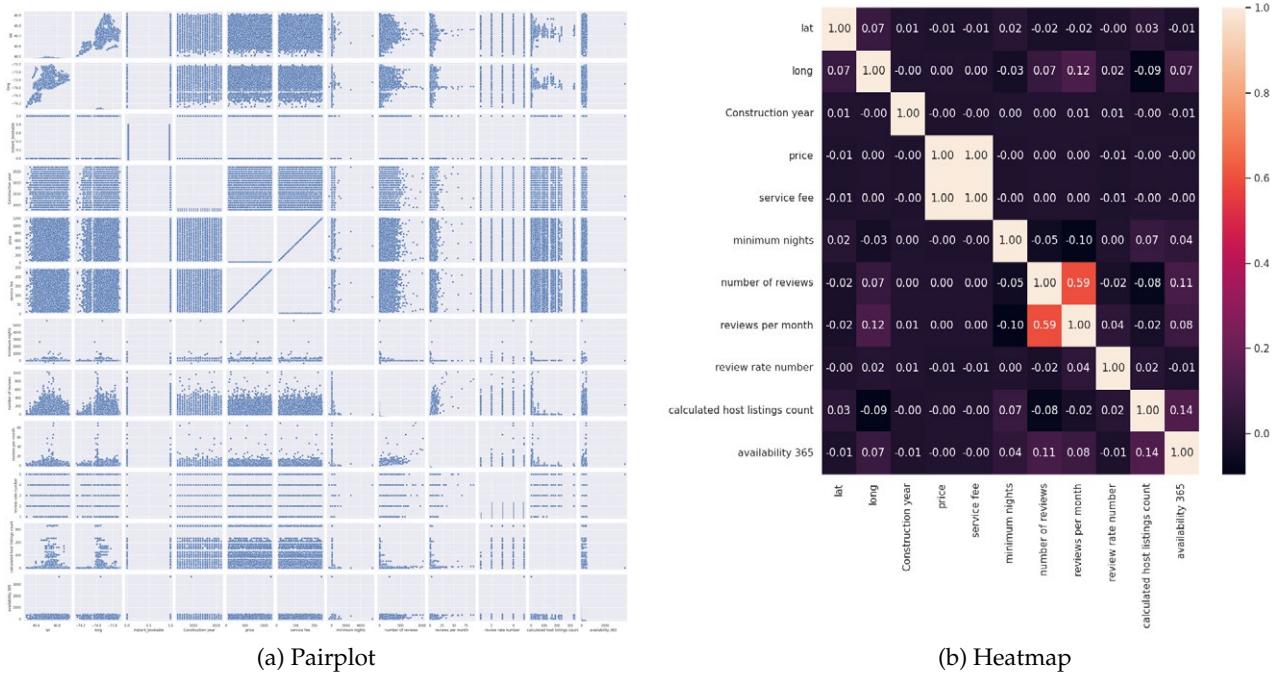


Fig. 4.9: Histograma Latitud

ción de datos multidimensionales. Formalmente, el PCA es una técnica estadística para reducir la dimensionalidad de un conjunto de datos; esto se consigue transformando linealmente los datos en un nuevo sistema de coordenadas en el que la mayor parte de la variación de los datos puede describirse con menos dimensiones que los datos iniciales. Muchos estudios utilizan los dos primeros componentes principales para representar los datos en dos dimensiones e identificar visualmente grupos de puntos de datos estrechamente relacionados.

Los componentes principales de una colección de puntos en un espacio de coordenadas reales son una secuencia de p vectores unitarios, donde el i -ésimo vector es la dirección de una línea

que mejor se ajusta a los datos mientras es ortogonal a los primeros $i - 1$ vectores. Aquí, la línea que mejor se ajusta se define como aquella que minimiza la distancia perpendicular media al cuadrado desde los puntos a la línea. Estas direcciones constituyen una base orto-normal en la que las diferentes dimensiones individuales de los datos están linealmente no correlacionadas. El análisis de componentes principales (PCA) es el proceso de calcular los componentes principales y utilizarlos para realizar un cambio de base en los datos, a veces utilizando solo los primeros componentes principales e ignorando el resto.

Visualización con PCA: Además de las Caras de Chernoff, existen diversas técnicas que pueden ser útiles para manejar la alta dimensionalidad de los datos. Una de ellas es la reducción de dimensionalidad, como el Análisis de Componentes Principales (PCA) o el Análisis Discriminante Lineal (LDA). Estas técnicas transforman los datos originales en un nuevo conjunto de variables (componentes) que son combinaciones lineales de las variables originales, y que mantienen la mayor cantidad de variabilidad presente en los datos. Esta transformación puede facilitar la identificación de las variables más importantes y ayudar a eliminar el ruido presente en los datos.

Otra técnica que puede utilizarse para explorar las relaciones entre las muestras es el clustering. Los métodos de clustering, como el K-means, la propagación de afinidad o la agrupación jerárquica, agrupan los datos en función de sus similitudes, lo que puede revelar patrones y relaciones ocultas entre los datos. Es importante señalar que la elección del método de clustering a utilizar dependerá en gran medida de la naturaleza de los datos y del problema en estudio.

En resumen, para obtener una comprensión más completa y detallada de los datos, es necesario explorar y considerar su naturaleza multidimensional. Las técnicas mencionadas anteriormente, junto con un enfoque de análisis visual como las Caras de Chernoff, pueden ser muy útiles en este sentido. Para facilitar la comprensión de múltiples variables, se implementa el siguiente código, que utiliza el Análisis de Componentes Principales (PCA, por sus siglas en inglés) de la biblioteca de aprendizaje automático de Python, scikit-learn. PCA es una técnica de reducción de dimensionalidad frecuentemente empleada en análisis de datos para visualizar la estructura de los mismos en un espacio de menor dimensión.

```
1 # Importando las bibliotecas necesarias
2
3 from sklearn.decomposition import PCA
4 from sklearn.preprocessing import StandardScaler
5 import pandas as pd
6 import matplotlib.pyplot as plt
7 df_encoded = pd.get_dummies(df,
8                             columns=[ 'neighbourhood group'])
9
10 elements = list(df_encoded.columns[20:])
11
12 # Separando las características
13 x = df_encoded.loc[:, cols + elements].values
14
15
16 # Definiendo PCA con 2 componentes
17 pca2 = PCA(n_components=2)
18 principalComponents2 = pca2.fit_transform(x)
19 df_principal2 = pd.DataFrame(data = principalComponents2,
20                               columns = [ 'PC1', 'PC2'])
21
22 # Definiendo PCA con 3 componentes
23 pca3 = PCA(n_components=3)
24 principalComponents3 = pca3.fit_transform(x)
25 df_principal3 = pd.DataFrame(data = principalComponents3,
26                               columns = [ 'PC1', 'PC2', 'PC3'])
27
```

```
28 # Creando la grafica para PCA de 2 componentes
29 plt.figure(figsize=(8,6))
30 plt.scatter(df_principal2['PC1'], df_principal2['PC2'])
31 plt.title('PCA con 2 Componentes')
32 plt.xlabel('Primer componente principal')
33 plt.ylabel('Segundo componente principal')
34 plt.show()
35
36 # Creando la grafica para PCA de 3 componentes
37 fig = plt.figure(figsize=(8,6))
38 ax = fig.add_subplot(111, projection='3d')
39 ax.scatter(df_principal3['PC1'],
40             df_principal3['PC2'],
41             df_principal3['PC3'])
42 ax.set_title('PCA con 3 Componentes')
43 ax.set_xlabel('Primer componente principal')
44 ax.set_ylabel('Segundo componente principal')
45 ax.set_zlabel('Tercer componente principal')
46 ax.view_init(35, 0)
47 plt.show()
```

El procedimiento inicia con la importación de las bibliotecas requeridas: scikit-learn, pandas y matplotlib. A continuación, se preprocesa el conjunto de datos original “df”, convirtiéndolo en una representación de variables dummy para las categorías presentes en la columna “neighbourhood group” mediante el uso de la función “get_dummies” de pandas. El resultado se almacena en “df_encoded”.

Posteriormente, se seleccionan las columnas de interés, las cuales se almacenan en las listas “elements” y “cols”, y se separan como características, ‘x’, para la implementación de PCA. Estas características son estandarizadas utilizando “StandardScaler” de scikit-learn, ya que PCA es sensible a la escala de las variables. “StandardScaler” estandariza las características para que tengan una media de 0 y una desviación estándar de 1.

Después, se definen dos instancias de PCA: una con dos componentes y otra con tres, utilizando ‘PCA(n_components=2)’ y ‘PCA(n_components=3)’. Estas instancias se ajustan y transforman a las características estandarizadas ‘x’ mediante la función ‘fit_transform()’. Los resultados se almacenan en ‘df_principal2’ y ‘df_principal3’, respectivamente.

En última instancia, el código proyecta las características en dos y tres componentes principales, empleando para ello la biblioteca matplotlib. Se utiliza la función “scatter()” para crear los diagramas de dispersión y las funciones “set_xlabel()”, “set_ylabel()”, y “set_zlabel()” para asignar las etiquetas a los ejes respectivos. En el caso del diagrama tridimensional, se ajusta un ángulo de visión específico utilizando la función “view_init()”. Ambos diagramas se visualizan con la función “plt.show()”. Este procedimiento facilita la representación visual de la estructura de los datos en espacios bidimensional y tridimensional, lo que puede ser especialmente útil para identificar patrones o estructuras subyacentes en los datos, como se puede apreciar en las Figuras 4.10 y 4.11.

En general, en la representación bidimensional de los componentes principales, parecen discernirse tres agrupaciones o “nubes” de datos. En la representación tridimensional, por su parte, se vuelve más desafiante distinguir cada conjunto individualmente, como se puede observar en las Figuras 4.10 y 4.11. Al haberse identificado algunas posibles agrupaciones de datos, se vuelve factible el uso de técnicas adicionales, como el clustering, para explorar estos posibles grupos de datos y determinar qué características comparten. Esta estrategia permite una mayor profundidad en la interpretación y comprensión de la estructura de los datos.

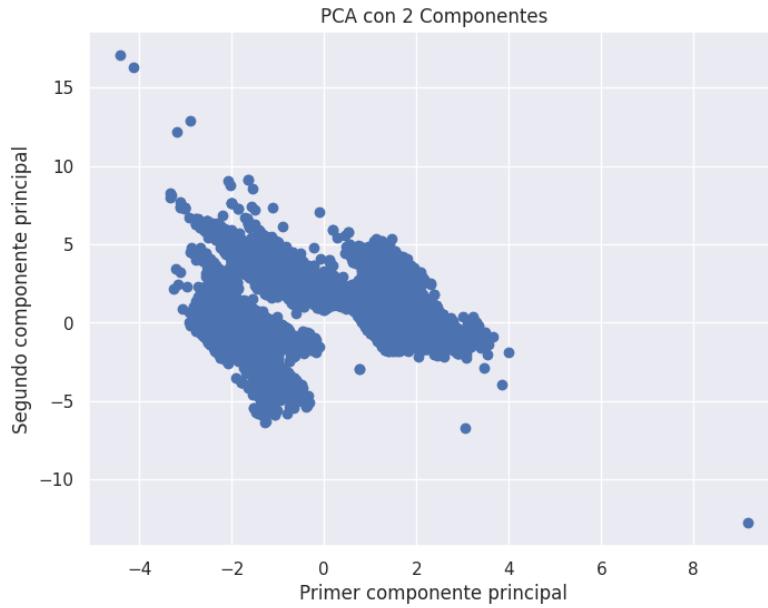


Fig. 4.10: Gráfica PCA 2d

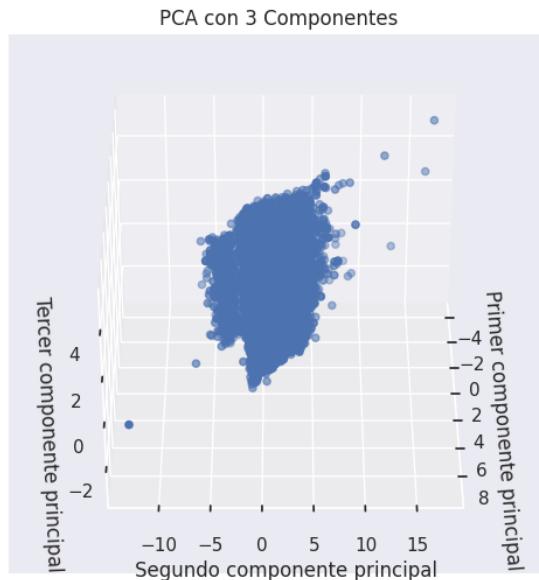


Fig. 4.11: Gráfica PCA 3d

Cabe mencionar que, aunque PCA es una herramienta poderosa y ampliamente utilizada para la reducción de la dimensionalidad, también tiene sus limitaciones. La interpretación de los componentes principales puede ser compleja y no siempre es fácil de entender en términos de las

variables originales. Además, PCA asume una relación lineal entre las variables, lo cual puede no ser el caso en todos los conjuntos de datos. Por lo tanto, es crucial entender que la visualización obtenida mediante PCA es un primer paso en el análisis de datos, no el último. Tras la identificación de posibles agrupaciones a través de PCA, es recomendable realizar un análisis más detallado mediante otras técnicas de aprendizaje automático o estadístico, como clustering o clasificación supervisada, para profundizar en el entendimiento de los patrones identificados en los datos. Adicionalmente, siempre es relevante considerar el conocimiento del dominio al interpretar los resultados, para asegurar que estos tengan sentido en el contexto de estudio.

5

CAPÍTULO CINCO

Validación y análisis de integridad de bases de datos

En esta sección se va a desarrollar un ejemplo que continúe con todo lo que se ha tratado hasta el momento. Para el desarrollo del ejemplo se usó información de vehículos de transporte eléctrico que transmitían datos registrados en el vehículo durante diferentes recorridos. Para este apartado se añade un link de acceso a los datos que se van a utilizar, es importante tener en cuenta que se van a entregar datos sintéticos ya que la base de datos original no se puede publicar, [link](#).

5.1. Primer conjunto de datos

El primer conjunto de datos está compuesto por cuatro archivos de los cuales dos son tipo Excel y el resto son de valores separados por comas (csv), los cuales contienen datos de las variables obtenidas a través de sensores tales como: memoria, tiempos de comunicación, localización GPS (latitud y longitud), estado de las baterías (Temperatura, Nivel de Carga), kilómetros del odómetro, revoluciones del motor, entre otros.

5.1.1. Carga y unión

Haciendo uso de un *jupyter notebook* se puede desarrollar el proceso de limpieza y unión de las tablas que se proporcionaron. A grandes rasgos, el procedimiento inicia cargando y decodificando las tablas, para identificar el número de muestras por tabla y posteriormente hacer un *full outer join*; de la tabla resultante, se obtiene el total de características y el número total de muestras.

5.1.2. Limpieza

Posteriormente es necesario realizar la limpieza de los datos, en donde se eliminan las características que se encuentran sin información, además de eliminar características que no son relevantes o no aportan para el análisis. Las características que se proceden a eliminar por no ser relevantes para el análisis son las siguientes:

- “Memeoria Disco Sts”
- “Uso Cpu Sts”
- “Memoria Ram Sts”
- “Versión Trama”
- “Tipo Trama”
- “VersiÃ³n Trama”

Dentro de las variables se identifica una completa ausencia de datos o un solo valor repetido a través de todas las muestras. Es importante observar que los nombres tienen errores obvios de ortografía o ya de por sí están mal escritos, esto se puede dar a múltiples razones. La primera razón es que hubo un error al cargar los datos como por ejemplo una mala selección de codificación. Por otro lado está en el diseño de la captación de datos y en la comunicación entre piezas de software en donde se cometieron errores a la hora de configurar y codificar de los datos. Como los datos provienen de una red de sensores en vehículos de transporte masivo pueden existir problemas en la codificación.

5.1.3. Comprensión general de datos

En el estudio de la analítica de datos, es esencial llevar a cabo un análisis detallado de las variables involucradas en el conjunto de datos. A continuación, se presenta un enfoque metodológico para abordar este proceso en el contexto de un caso particular.

Inicialmente, se realiza una transformación de la variable en donde se indica la fecha de lectura a una nueva característica en la que se estandariza el nombre, esto con el motivo de facilitar el acceso. La transformación consiste en representar la fecha en el formato estandarizado "YYYY/MM/DD". A continuación, se identifican los valores únicos de esta nueva variable para determinar los días en los que se llevaron a cabo registros. En este caso específico, se encontraron registros correspondientes a tres días del mes de junio.

No obstante, el texto no proporciona información detallada sobre a qué día corresponden exactamente estos registros. Por lo tanto, será necesario realizar un análisis adicional para establecer una correlación más precisa entre los registros y las fechas.

Como siguiente paso, se lleva a cabo un agrupamiento de los datos en función de los días de operación. Este procedimiento permite obtener la cantidad de muestras recopiladas en cada uno de los días involucrados. A través de este análisis, se determinó que se contabilizaron 53, 328 y 767 muestras para los tres días de operación respectivamente.

El proceso de analítica de datos realizado en este ejemplo incluye la transformación y agrupación de variables temporales, así como la identificación de los registros únicos. Aunque se han obtenido resultados preliminares, es necesario realizar investigaciones adicionales para obtener una comprensión más profunda de la relación entre las fechas y los registros. Esta metodología se puede aplicar a otras situaciones similares en el ámbito de la analítica de datos.

5.1.4. Análisis geo-espacial

En este apartado, se aborda el análisis geo-espacial de los datos para obtener una comprensión más profunda de las áreas geográficas en las que se recopilaron las muestras. Asumiendo que el sistema de referencia utilizado por los sensores GPS es WGS84, dado que es el estándar en la mayoría de los dispositivos, se procede a cargar la información de localización en un marco de datos geopandas y generar la visualización.

Al examinar los datos, se identifican dos trayectos principales, uno en Ciudad 1 y otro en Ciudad 2. Además, se detectan registros que corresponden a períodos de estacionamiento del vehículo en cuestión, por lo que se procede a descartar estos puntos. También se encuentran registros correspondientes a una tercera ciudad, Ciudad 3, dentro de los datos del trayecto de Ciudad 1, por lo que se eliminan estos puntos del análisis.

Realizando una inspección detallada de los datos, se detectan vacíos en la información del trayecto de Ciudad 1. Ante esta situación, surgen preguntas acerca de las posibles causas de estos vacíos, como la posibilidad de que el dispositivo de cómputo se haya apagado durante el recorrido, haya ocurrido un reinicio, entre otros factores. También se observa la presencia de aproximadamente cuatro grupos discontinuos en el trayecto de Ciudad 1, lo cual sugiere la necesidad de un análisis adicional.

Tras eliminar los datos atípicos y repetidos, se almacenan las tablas resultantes de los trayectos en archivos separados. Posteriormente, se realizan limpiezas adicionales de los registros únicos que no aportan información relevante y de las columnas sin información. A partir de las tablas resultantes, se elaboran gráficos que visualizan los puntos del recorrido de Ciudad 2 con información relacionada con la aceleración, el odómetro y la velocidad. Sin embargo, no se realiza un gráfico similar para Ciudad 1 debido a la falta de información en algunas áreas.

Finalmente, se obtienen algunas estadísticas descriptivas sobre la ruta de Ciudad 2. En particular, se analiza el consumo de energía en función de los datos disponibles, y se calcula el consumo por

kilómetro del recorrido. Esta información puede ser útil para comprender mejor el rendimiento del vehículo en términos de eficiencia energética y otros aspectos relevantes.

5.2. Segundo conjunto de datos

El segundo conjunto de datos consta de dos archivos CSV (valores separados por comas). Estos archivos incluyen información sobre diversas variables recopiladas mediante sensores, como: memoria, tiempos de comunicación, localización GPS (latitud y longitud), estado de las baterías (temperatura y nivel de carga), distancia recorrida según el odómetro, revoluciones del motor y otros parámetros similares.

5.2.1. Proceso de carga de datos

Para el análisis y procesamiento de los datos obtenidos de los autobuses en Noruega, se utiliza un cuaderno Jupyter Notebook. Este cuaderno se centra en el estudio y la limpieza de la información recolectada durante dos días específicos: el 8 y el 9 de septiembre.

El conjunto de datos corresponde a un viaje de ida y vuelta entre dos ubicaciones en Noruega. Para el análisis, se consideran variables como las posiciones geográficas del vehículo y los desplazamientos, los cuales se obtienen a partir de diversas fuentes de información. Adicionalmente, se emplea una técnica de geocodificación inversa para determinar la altitud asociada a cada par de coordenadas de latitud y longitud registradas.

Cabe mencionar que los datos fueron proporcionados originalmente en formato Excel, lo que generó ciertas dificultades en el proceso de carga y conversión. Algunos registros se perdieron y ciertos valores numéricos fueron malinterpretados por el software, lo que resultó en la asignación de formato de texto y la adición de puntos innecesarios en los datos originales. Por lo tanto, fue necesario realizar una limpieza previa de los datos para garantizar su correcta interpretación y manipulación. Luego de realizar la limpieza se guarda todo en archivos csv para su posterior procesamiento.

Una vez cargado el conjunto de datos, se observa que cuenta con veinticinco características y trescientos sesenta y siete registros. Este conjunto de datos se utilizará como base para el análisis y la extracción de información relevante en el contexto del transporte público en Noruega.

5.2.2. Proceso de limpieza de datos

El proceso de limpieza de datos implica evaluar cada columna y determinar su relevancia para el análisis. Inicialmente, se revisan las columnas para identificar aquellas que contienen un único dato o están vacíos, y se eliminan debido a su falta de contribución al análisis. Como resultado, se reduce el número de columnas de 25 a 10, las cuales incluyen características como ubicación geográfica, distancia recorrida, niveles de energía y temperatura de las baterías, entre otras.

No obstante, no todas las columnas restantes son relevantes para el estudio. Algunas contienen información de identificación de registros o marcas temporales asociadas al momento de envío de los datos a la base de datos. Estas columnas no aportan información valiosa para el análisis y, por lo tanto, se eliminan.

Posteriormente, se convierte la columna que contiene la información de tiempo de lectura de datos en un objeto de tipo Datetime, siguiendo una estructura de tiempo específica. Aunque la norma ISO 8601 establece un formato preferido (AAAA-MM-DD), se utiliza un formato alternativo en este caso.

En cuanto a las columnas de latitud y longitud, se identifican problemas derivados de la exportación de los datos a archivos de Excel. Las conversiones entre formatos pueden generar pérdidas de información, especialmente cuando se trata de archivos binarios dependientes de variables de configuración específicas. Por lo tanto, es necesario ajustar la latitud y la longitud, desplazando el separador decimal seis posiciones hacia la izquierda. De esta manera, se obtienen valores que se encuentran dentro de los límites establecidos para la latitud (0 a 90 grados) y la longitud (de -180 a 180 grados). Las ecuaciones (5.1) y (5.2) muestran el ajuste necesario para las coordenadas:

$$Latitud_{ajustada} = \frac{Latitud_{Sinajustar}}{1 \times 10^6} \quad (5.1)$$

$$Longitud_{ajustada} = \frac{Longitud_{Sinajustar}}{1 \times 10^6} \quad (5.2)$$

Con estos ajustes, se finaliza el proceso de limpieza de datos, lo que permite avanzar en el análisis con un conjunto de información más adecuado y coherente.

5.2.3. Análisis general de datos

Los ajustes realizados previamente se almacenan en sus respectivas columnas para que se reflejen en la tabla; de esta forma, se obtiene un conjunto de datos limpio, con un total de siete características, para el día 8 un total de 275, y para el día 9 unos 367 registros.

En el análisis, el experto sugiere realizar un subconjunto de tamaño 3 de la tabla resultante, con las características que representan la fecha y hora de lectura de datos, kilómetros recorridos y nivel de energía restante. Dentro de este subconjunto, se encuentran datos atípicos que son detectados rápidamente, ya que el odómetro comienza alrededor de los 9500 para la prueba del día 8; por lo tanto, los valores que estén en cero no son interesantes ya que pueden ser posibles errores de registro.

Al nuevo subconjunto se le puede extraer información sobre la distancia recorrida por el autobús en la prueba, su velocidad promedio y el comportamiento de la carga durante todo el recorrido de prueba. De tal manera, se obtiene el valor máximo y el mínimo de la columna que representa los kilómetros recorridos para calcular la diferencia entre estos valores. Al final, se obtiene que el total de kilómetros recorridos por el autobús en Noruega fue de 113 kilómetros para el día 8 y 141 kilómetros para el día 9; donde el odómetro marca como mínimos para los dos recorridos un valor de 9501 y 9615 kilómetros acumulados con un máximo de unos 9615 y 9755 kilómetros acumulados para los días 8 y 9, respectivamente.

Con los datos del odómetro, se puede realizar una propagación del acumulado de las diferencias en el tiempo, siempre que se tenga la tabla de manera ordenada en la columna que representa la fecha y hora de lectura de datos, como se puede observar en la ecuación (5.3) en donde se logra obtener el total de kilómetros acumulados por recorrido.

$$\begin{aligned}
 & \text{tabla[Posicion''][Indice]} = \text{tabla[Posicion''][Indice - 1]} \\
 & + (\text{tabla[Odometro''][Indice]} - \text{tabla[Odometro''][Indice - 1]}) \tag{5.3}
 \end{aligned}$$

En la figura 5.1, se realiza una gráfica de la posición versus el tiempo de duración de la prueba; a partir de esta gráfica, se puede observar que el autobús estuvo detenido durante casi dos horas al inicio de la prueba y una hora al final de esta. Mientras que se estuvo moviendo alrededor de tres horas de manera quasi constante y con una velocidad media de 14 kilómetros por hora para un recorrido de 114 kilómetros.

En la figura 5.2, se realiza una gráfica de la posición versus el tiempo de duración de la prueba en el día nueve. En esta gráfica se puede observar que en la prueba estuvo detenido durante casi dos horas desde las 14:20 de la tarde hasta las 16:20, para luego realizar un recorrido de alrededor de 23 kilómetros. De las 18:15 a las 19:00 ocurre otro movimiento del vehículo, con un desplazamiento de alrededor de 60 kilómetros, otro descanso para finalizar con el último desplazamiento

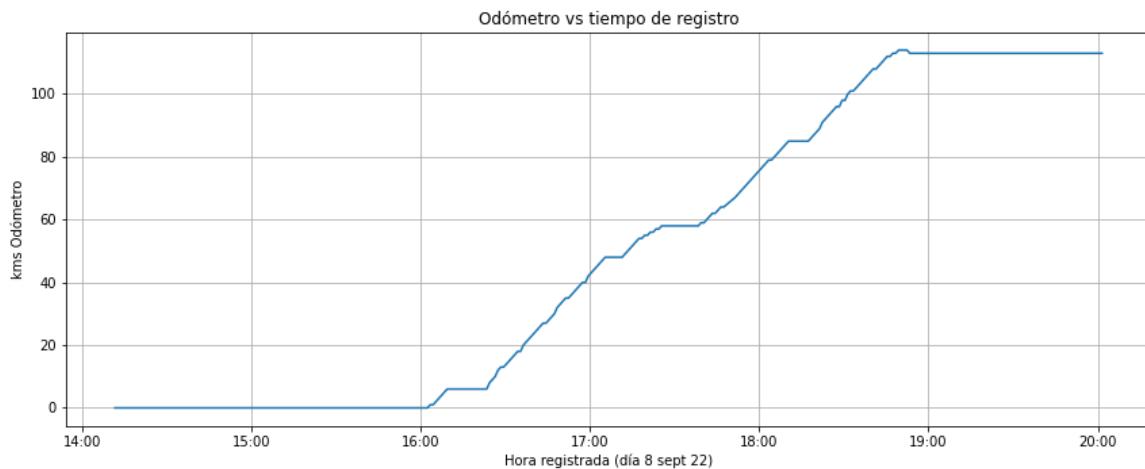


Fig. 5.1: Posición vs Tiempo día ocho

alrededor de 60 kilómetros. Teniendo los desplazamientos y la duración de tiempo del desplazamiento, se puede realizar el cálculo de la velocidad promedio de todo el trayecto, siendo esta de 14,55 kilómetros por hora.



Fig. 5.2: Posición vs Tiempo día nueve

Al comparar la gráfica del nivel de carga del vehículo en Noruega en cada instante de tiempo, se espera que siga la misma tendencia que la figura 5.2, pero de manera descendente. Por lo tanto, se procede a graficar la carga del vehículo en función del tiempo de duración del trayecto, lo que permite obtener las figuras 5.3 y 5.4, correspondientes a los días 8 y 9, respectivamente. En la figura 5.3, correspondiente al día 8, se observa que el nivel de las baterías sigue una tendencia inversa a la del movimiento del vehículo; al invertir la gráfica, ambas tendrían la misma tendencia.

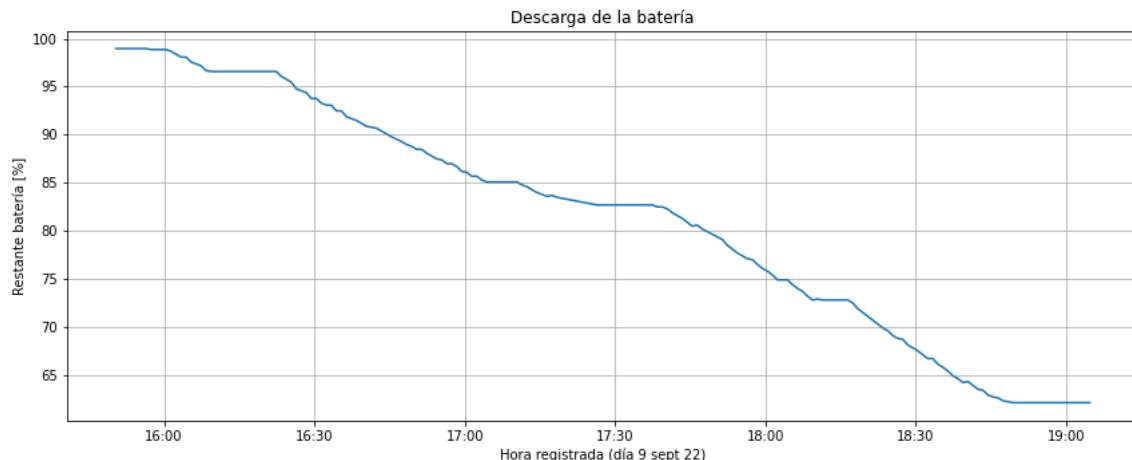


Fig. 5.3: Nivel de Carga vs Tiempo Día Ocho

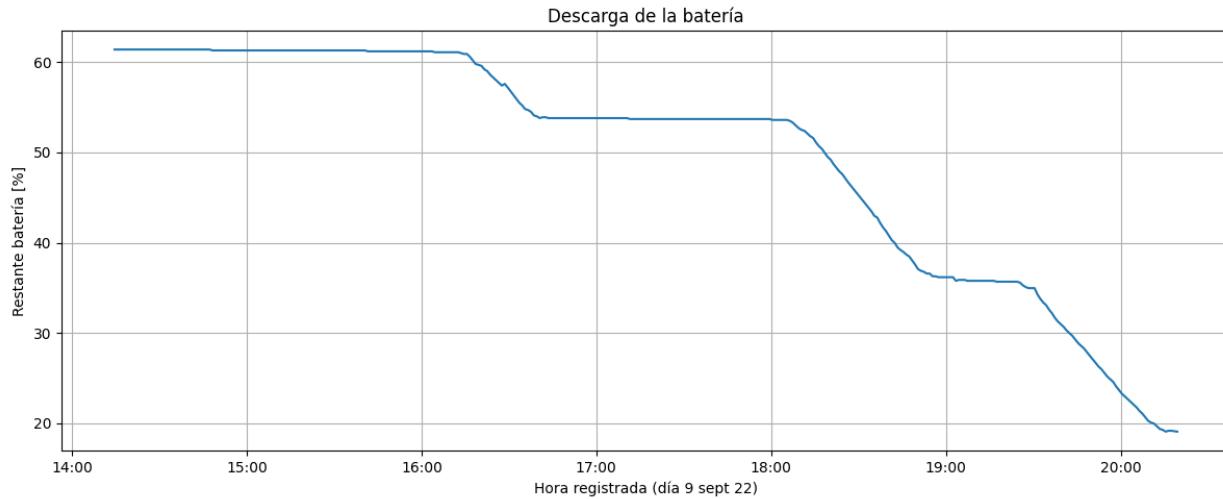


Fig. 5.4: Nivel de Carga vs Tiempo Día Nueve

Cabe mencionar que las figuras 5.3 y 5.4 presentan valores en términos de porcentaje. Para realizar un análisis posterior, es necesario transformar estos valores en el eje correspondiente. Para ello, se requiere conocer el valor nominal de potencia de las baterías y aplicar dicha información en la propagación por todo el conjunto de datos. Tras realizar la transformación, se pueden graficar la potencia frente al tiempo del recorrido, obteniendo las figuras 5.5 y 5.6. Mediante una inspección visual, se puede confirmar que la transformación no afectó la forma de las gráficas al comparar la figura 5.3 con la figura 5.5, y la figura 5.4 con la figura 5.6.

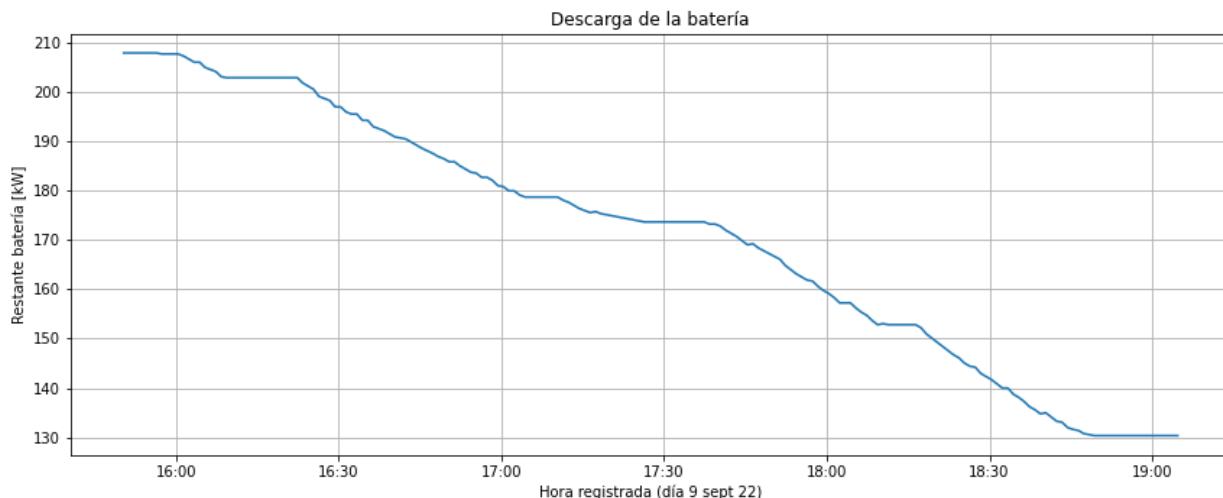


Fig. 5.5: Potencia vs Tiempo Día Ocho

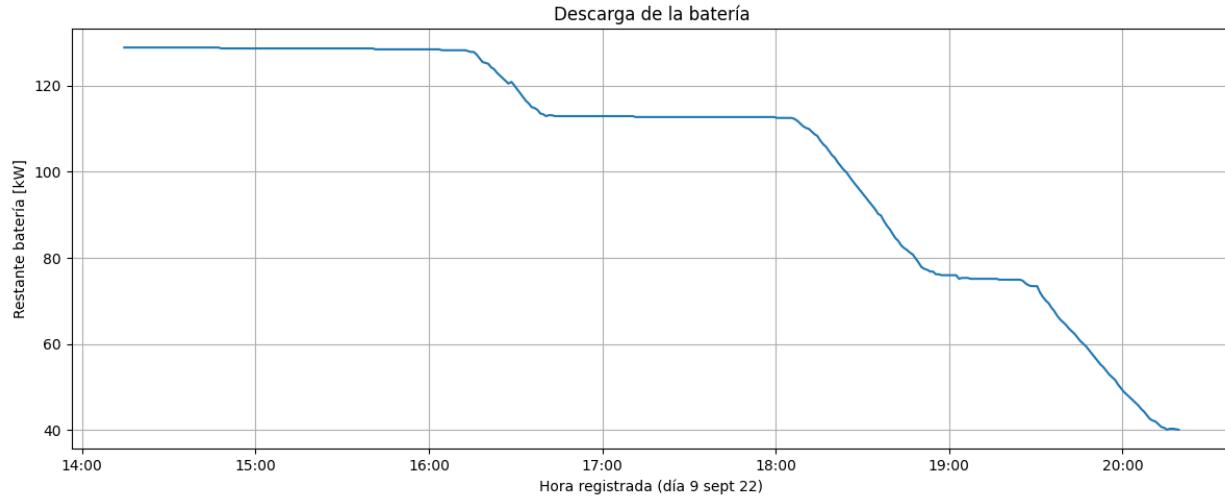


Fig. 5.6: Potencia vs Tiempo Día Nueve

Examinando las figuras 5.3 y 5.5, relacionadas al día 8, es posible extraer información relevante sobre ciertas características, como el rango de variación del nivel de carga a lo largo del recorrido, el tiempo total del trayecto, la potencia y la energía involucradas. En este contexto, se observa un rango de [99,36 % 61,87 %] para la batería y una potencia de 77,448 kilovatios en un recorrido de tres horas, lo que se traduce en una energía aproximada de 250,81 kilovatios hora. Por otro lado, para el día 9, las figuras 5.4 y 5.6 indican un rango de [61,87 19,09], una potencia de 88,79 kilovatios en un recorrido de seis horas y una energía cercana a 539,87 kilovatios hora.

Utilizando las coordenadas de cada punto del recorrido y la herramienta **open-elevation API** (una Interfaz de Programación de Aplicaciones de código abierto alternativa a las versiones de pago), se pueden obtener los perfiles de altitud a partir de las coordenadas de longitud y latitud. Los perfiles de altitud se presentan en las figuras 5.7 y 5.8. Para el día 8, el rango de altura está entre 60 y 185 m.s.n.m, mostrando un solo valle que indica un descenso y un ascenso durante el trayecto. En el caso del día 9, el rango de altura también se encuentra entre 60 y 185 m.s.n.m, pero con dos valles que indican descensos y ascensos en el recorrido. En el primer valle, se desciende aproximadamente 40 metros, mientras que en el segundo valle se desciende 120 metros.

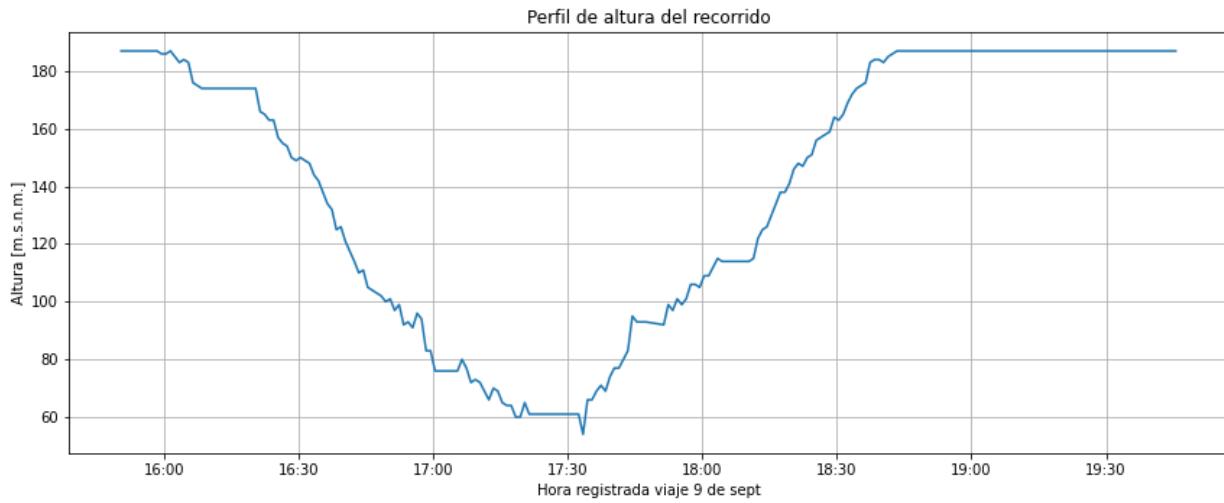


Fig. 5.7: Altura vs Tiempo Día Ocho

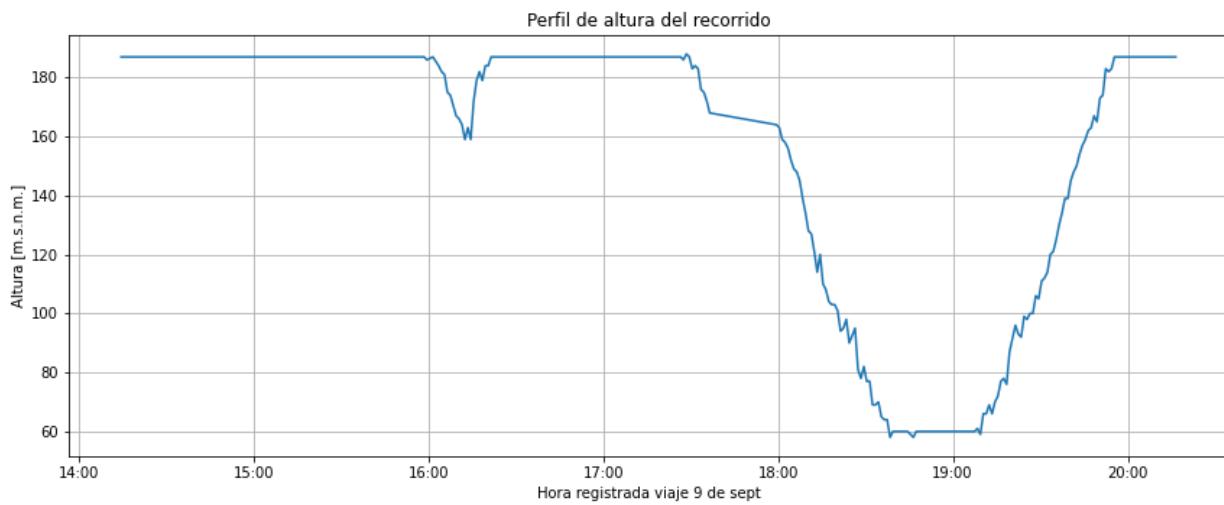


Fig. 5.8: Altura vs Tiempo Día Nueve

Con el perfil de altura obtenido, es posible llevar a cabo una comparación entre este y la distancia acumulada, además de añadir dicha distancia al análisis. Aún falta visualizar el nivel de energía en función de la distancia acumulada, lo cual puede observarse en la figura 5.9. En esta figura, se aprecia cómo la disminución del nivel de energía en relación a la distancia acumulada sigue una línea recta descendente, donde la pendiente determina el rendimiento de la batería por cada kilómetro recorrido. A partir de estos resultados, es posible realizar una estimación del máximo recorrido que un autobús de estas características puede alcanzar bajo dichas condiciones.

En un futuro, cuando se disponga de diferentes modelos de autobuses en diversas ubicaciones, se podrá establecer de manera más precisa el rendimiento de un autobús, considerando ciertas condiciones específicas. Además, se podría contemplar la inclusión de datos ambientales o buscar información relacionada con las áreas de despliegue para tener en cuenta estos factores en los análisis realizados.

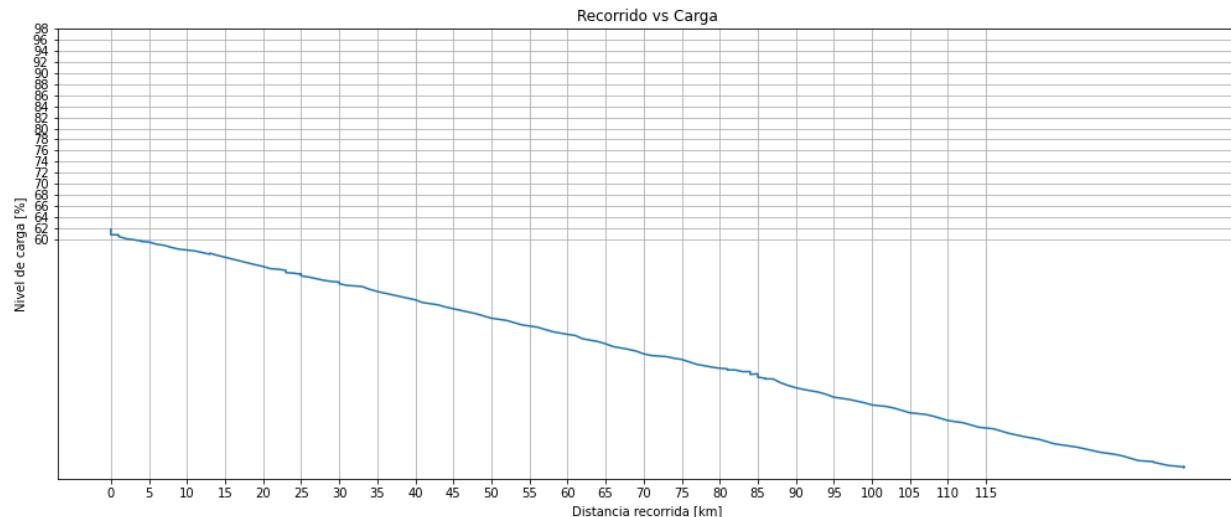


Fig. 5.9: Nivel de Energía vs Distancia Acumulada

6

CAPÍTULO SEIS

Visualizaciones de datos y análisis de relaciones en bases de datos

La información, catalizador del entorno empresarial moderno, adquiere valor sin precedentes con la evolución tecnológica. Este progreso facilita no solo la acumulación y el procesamiento de datos sino también su conversión en recursos monetizables. Al utilizar los datos estratégicamente, las empresas pueden innovar, ajustar precios y asignar recursos eficientemente, fomentando una toma de decisiones informada y estratégica [15]. Sin embargo, la dependencia de datos en estado puro puede inducir a conclusiones erróneas y decisiones desafortunadas. Este capítulo se enfoca en métodos estadísticos y técnicas de visualización avanzados para refinar y comprender dichos datos, maximizando su potencial.

La visualización de datos surge como un puente entre el análisis numérico y la comprensión intuitiva, permitiendo no solo discernir patrones y anomalías sino también facilitar la comunicación efectiva de resultados. Profundizaremos en cómo las técnicas de visualización pueden interpretar y narrar las historias que yacen en las complejidades de las bases de datos. Con este conocimiento, el lector podrá incorporar visualizaciones en la toma de decisiones, asegurando resultados validados y comprensibles para un espectro amplio de audiencias.

Centrándonos en un conjunto de datos reciente, analizaremos las métricas de las pruebas de autobuses noruegos del 8 y 9 de septiembre. Posteriormente, simularemos un entorno de big data, con sus inherentes desafíos de volumen, velocidad y variedad, para demostrar la aplicabilidad de estas técnicas en contextos de mayor escala.

6.1. Proceso de limpieza

Como se describe en la sub-sección 2.1, data cleaning, la importancia de esta etapa es significativa para la obtención de conclusiones de mayor profundidad y relevancia. Inicialmente, para limpiar los datos, es necesario realizar y responder preguntas como: ¿De qué tipo son los datos a tratar? ¿cuántos datos se tienen?, ¿cuántas características hay en el conjunto de datos?, ¿cuántos valores únicos hay por característica? y ¿cuántos baches de información hay por característica? Estas preguntas marcan el punto de inicio y sirven para comparar al momento de finalizar el proceso. Para el caso de los conjuntos del día 8 y 9 de septiembre de la prueba con el conjunto de datos 2 se obtiene la tabla 6.1.

Los acrónimos utilizados en la tabla para las columnas son:

- **NTC: Nivel Tanque Combustible**
- **PAM: Presión Aceite Motor**
- **TM: Temperatura Motor**
- **CC: Consumo Combustible**

TABLA 6.1
TABLA DATOS INICIALES

	Registros	Características	NTC	PAM	TM	CC
Día 8	275	26	275	275	275	275
Día 9	367	26	367	367	367	367

En la tabla 6.1 se encuentra que ambas pruebas tienen 26 características con un total de 275 muestras para el día 8 y 367 para el día 9. De estas características, se tienen cuatro columnas en las que no hay ningún dato. Estas columnas son: "nivelTanqueCombustible", "presionAceiteMotor", "temperaturaMotor", y "consumoCombustible". La ausencia de datos en las cuatro columnas anteriores indica que no son relevantes para el análisis y, por ende, pueden ser removidas. Al remover las columnas vacías se queda con dos tablas de 22 características con información; ahora, el objetivo es identificar cuántos valores únicos hay por variable, esto resulta en la tabla 6.2.

TABLA 6.2
TABLA DATOS ÚNICOS

	Día 8	Día 9
id	275	367
fechaHoraEnvioDato	275	367
fechaHoraLecturaDato	275	367
idConductor	1	1
idOperador	1	1
idRegistro	275	367
idRuta	1	1
idVehiculo	1	1
latitud	139	142
longitud	140	143
kilometrosOdometro	105	112
nivelRestanteEnergia	126	140
revolucionesMotor	118	121
tecnologiaMotor	1	1
temperaturaBaterias	6	10
tipoBus	1	1
tipoTrama	1	1
versionTrama	1	1
sentidoMarcha	1	1
tipoFreno	1	1
tramaRetransmitida	1	1
estadoDesgasteFrenos	1	1

Tras realizar el análisis se obtienen los resultados de la tabla 6.2, de la cual se sabe que hay doce columnas en las que solamente hay un valor único; es decir, la información que aporta estas variables al análisis es prácticamente nula, por tal motivo, se puede proceder a eliminar las columnas: "idConductor", "idOperador", "idRuta", "idVehiculo", "tecnologiaMotor", "tipoBus", "tipoTrama", "versionTrama", "sentidoMarcha", "tipoFreno", "tramaRetransmitida", y "estadoDesgaste-

Frenos". De esta forma se queda con dos conjuntos de datos de 10 variables con 275 muestras para el día 8, y 367 muestras para el día 9. Al final del proceso de eliminación se queda con la tabla 6.3

A medida que se va haciendo limpieza de los conjuntos de datos estos van disminuyendo de tamaño; ahora, lo que falta del proceso general es revisar el problema de negocio, los intereses

TABLA 6.3
TABLA DATOS ÚNICOS LUEGO DE LA ELIMINACIÓN

	Día 8	Día 9
id	275	367
fechaHoraEnvioDato	275	367
fechaHoraLecturaDato	275	367
idRegistro	275	367
latitud	139	142
longitud	140	143
kilometrosOdometro	105	112
nivelRestanteEnergia	126	140
revolucionesMotor	118	121
temperaturaBaterias	6	10

del stakeholder y el conocimiento del negocio, para identificar qué variables de las restantes son importantes para el análisis. De esta forma, se va por cada columna que se observa en la tabla 6.3, revisando su importancia.

Se identificaron las columnas: "id", "fechaHoraEnvioDato", y "idRegistro". En el primer caso, el del id, los valores de la columna no aportan nada al análisis ya que son identificadores únicos de la base de datos; además, no hay otras tablas para realizar el cruce de la información. La fecha de envío es el dato del momento en el que se envió el registro a la base de datos remota, y la diferencia con la fecha de lectura es muy poca, por ende, se puede descartar. Por último, el id registro también es un identificador único el cual no sirve ya que no hay otra tabla para realizar el cruce de información. Pero estos identificadores pueden ser útiles en un momento más avanzado del proceso, ya que estos podrían identificar los datos de un bus particular dentro de una flota de buses. Por los motivos antes dichos, entonces se procede a realizar la eliminación de las columnas seleccionadas, obteniendo la tabla 6.4.

TABLA 6.4
TABLA DATOS ÚNICOS LUEGO DE LA ELIMINACIÓN Y DEPURACIÓN

	Día 8	Día 9
fechaHoraLecturaDatos	275	367
latitud	139	142
longitud	140	143
kilometrosOdómetro	105	112
nivelRestanteEnergía	126	140
revolucionesMotor	118	121
temperaturaBaterías	6	10

Tras una inspección general de las características, se observa que los valores de “latitud” y “longitud” están en el orden de los millones. El rango de los valores de “latitud”, y “longitud” deberían estar en valores de $[-90, 90]$ y $[0, 360]$, respectivamente. Por este motivo, hay que realizar un tratamiento a los datos, en este caso se perdió el punto decimal y hay que volver a llevarlo al valor real. También, se debe tener en cuenta en qué rangos deberían estar las otras variables para evitar muestras que sean erróneas; por ejemplo, el odómetro marca un acumulado de kilómetros del vehículo y, por tal motivo, no es posible que tenga datos de 0 al igual que el nivel de la batería, ya que durante el trayecto si se queda en cero no podría moverse más. Al realizar esta depuración por cómo se comportan los datos, se obtiene que para el día 8 hay un total de 260 muestras y para el día 9 hay 344.

También se puede observar que hay datos que se pueden usar para crear nuevas columnas. Por ejemplo, obtener la altura en una ubicación a partir de la longitud y latitud. Otra posible columna serían los kilómetros acumulados en el trayecto realizado, o también separar de la fecha una columna con la hora de la toma de los datos; esto en particular ayudaría al momento de realizar gráficas y comparar los dos trayectos. Por último, se puede tener los metros recorridos entre muestra y muestra. Es bueno desde este punto pensar en qué variables se pueden calcular para analizar el problema.

Lo primero que se hace para añadir una nueva columna es nombrarla como “kilometrosAcum-Trayecto”, en esta se va hacer la diferencia del mínimo valor del acumulado de kilómetros del

bus contra todos los elementos de la columna del odómetro, como se puede observar en la ecuación (6.1).

$$kilometrosAcumTrayecto = kilometrosOdometro - \min(kilometrosOdometro) \quad (6.1)$$

Ya para la nueva columna, denominada “altitud”, hay que utilizar una API gratuita en la que se ingresan los valores de latitud y de longitud para recibir el valor de la altitud. Con la información de cada muestra se le va agregando a la columna altitud su debido valor en [m.s.n.m.]. Por último, se va a crear la columna “kilometrosEntreMuestra” para hacer la diferencia entre el valor actual de la muestra del odómetro, contra la muestra anterior; y de esta forma, saber cuánto recorrió el vehículo entre muestras. De esta manera se obtiene la tabla 6.5 con valores únicos de los conjuntos de información.

TABLA 6.5
TABLA DATOS ÚNICOS LUEGO DE AGREGAR NUEVAS COLUMNAS

	Día 8	Día 9
fechaHoraLecturaDatos	275	367
latitud	139	142
longitud	140	143
kilometrosOdometro	105	112
nivelRestanteEnergia	126	140
revolucionesMotor	118	121
temperaturaBaterias	6	10
kilometrosAcumTrayecto	101	110
kilometrosEntreMuestra	4	5
altitud	71	72

6.2. Proceso de comprensión general de datos

Al haber realizado el proceso de limpieza y tener limpio el conjunto de datos, se sigue con el proceso de compresión de la información. Este proceso trata de entender el comportamiento univariado y bivariado de las características que componen el conjunto de datos. La primera parte del

análisis es obtener los componentes estadísticos que describen a cada una de las características, es decir, los estadísticos principales. Estos estadísticos principales se dividen en tres grupos: los estadísticos de tendencia central, los estadísticos de dispersión, y los estadísticos de forma.

6.2.1. Cálculo de estadísticos de tendencia central

Antes de entrar a realizar cálculos y análisis de los estadísticos de tendencia central o cualquier otro, es necesario establecer qué variables son cualitativas, cuantitativas y cuáles definitivamente no se les aplicará estos cálculos. Al observar las tablas resultantes del día 8 y 9 de las pruebas del bus desarrollado en terreno, los datos que son cuantitativos continuos son “latitud”, “longitud”, “nivelRestanteEnergia”, “kilometrosOdómetro”, “temperaturaBaterias”, “kilometrosAcumTrayecto”, “kilometrosEntreMuestra”, “altitud”. Por otro lado, a la fecha no se les va hacer ningún cálculo y no hay variables cualitativas.

Luego del ajuste de los datos que tenían un comportamiento extraño, se puede proceder hacer el cálculo de la media, mediana y la moda de los conjuntos de datos. Al realizar los cálculos se obtienen las tablas [6.6](#) y [6.7](#).

TABLA 6.6
ESTADÍSTICOS DE TENDENCIA CENTRAL DÍA 8

Variable	Media	Mediana	Moda
latitud	29.00	29.06	29.09
longitud	-111.18	-111.09	-111.04
nivelRestanteEnergia	81.12	85.07	62.08
kilometrosOdómetro	9199.76	9540.0	9501
temperaturaBaterias	30.85	32.0	32.0
revolucionesMotor	737.92	0.00	0.0
kilometrosAcumTrayecto	46.34	45.50	0.0
kilometrosEntreMuestra	0.44	0.00	0.0
altitud	148.11	174.0	187

TABLA 6.7
ESTADÍSTICOS DE TENDENCIA CENTRAL DÍA 9

Variable	Media	Mediana	Moda
latitud	29.03	29.09	29.09
longitud	-111.14	-111.04	-111.04
nivelRestanteEnergia	45.92	53.68	-
kilometrosOdometro	9235.88	9639.0	9614
temperaturaBaterias	31.84	33.0	32.0
revolucionesMotor	686.48	0.00	0.0
kilometrosAcumTrayecto	43.00	25.0	0.0
kilometrosEntreMuestra	0.40	0.00	0.0
altitud	158.41	187.0	187

De las tablas 6.6 y 6.7, se puede observar cómo es el comportamiento de sus tres variables de tendencia central, y dar una primera impresión de cómo se distribuyen los datos. En el caso de la “latitud”, “longitud”, “temperaturaBaterias”, y “nivelRestanteEnergía”, indica a primera vista que su comportamiento es normal, ya que los tres datos de tendencia central están muy cerca entre sí, y como la única distribución en donde las tres medidas de tendencia central son iguales es la normal por tal motivo se podría ver una similitud. Por otro lado, se tienen las variables “kilometrosOdometro” y “revolucionesMotor”, en donde la media es inferior a la mediana para la primera variable y en la segunda, la media es mayor que la mediana. Esto nos quiere indicar que las distribuciones están sesgadas a la izquierda y a la derecha respectivamente. Para finalizar se tiene el caso de “kilometrosAcumTrayecto”, “kilometrosEntreMuestra”, y “altitud”.

6.2.2. Cálculo de estadísticos de dispersión

Ahora que se tienen los datos de los estadísticos de tendencia central, estos no indican completamente cómo es el comportamiento de los datos. Solo se saben algunos indicios de sus tendencias pero hace falta más información; es por eso que ahora calcularemos los estadísticos de dispersión para entender cómo están espaciadas las diferentes variables. Se pueden ver los resultados de los cálculos en las tablas 6.8 y 6.9.

**TABLA 6.8
ESTADÍSTICOS DE DISPERSIÓN DÍA 8**

Variable	Varianza	DesvEstandar	Cuartil 1	Cuartil 2	IQR	Mínimo	Máximo
latitud	0.01	0.10	28.9	29.09	0.18	28.83	29.09
longitud	0.025	0.158	-111.31	-111.04	0.27	-111.49	-111.04
nivelRestanteEnergia	186.7	13.66	72.77	98.96	26.19	61.87	99.36
kilometrosOdometro	1820.41	42.67	9501	9586	85.0	9501	9615
temperaturaBaterias	1.30	1.14	30.0	32.0	2.0	29.0	33.0
revolucionesMotor	835848.86	914.24	0.0	1901.5	1901.5	0.0	2472.0
kilometrosAcumTrayecto	1820.41	42.66	0.0	85.0	85.0	0.0	114.0
kilometrosEntreMuestra	0.36	0.60	0.0	1.0	1.0	0.0	3.0
altitud	2099.42	45.81	106.0	187.0	81.0	54.0	187.0

**TABLA 6.9
ESTADÍSTICOS DE DISPERSIÓN DÍA 9**

Variable	Varianza	DesvEstandar	Cuartil 1	Cuartil 2	IQR	Mínimo	Máximo
latitud	0.01	0.097	28.97	29.09	0.118	28.83	29.092
longitud	0.026	0.16	-111.22	-111.04	0.182	-111.50	-111.038
nivelRestanteEnergia	203.11	14.25	35.855	61.27	25.415	19.09	61.87
kilometrosOdometro	2209.15	47.00	9614.0	9699.0	85.0	9614.0	9755.0
temperaturaBaterias	7.36	2.71	30.0	35.0	5.0	28.0	36.0
revolucionesMotor	1048644.86	1024.03	0.0	1920.5	1920.5	0.0	2480.0
kilometrosAcumTrayecto	2209.15	47.00	0.0	85.0	85.0	0.0	141.0
kilometrosEntreMuestra	0.48	0.69	0.0	1.0	1.0	0.0	5.0
altitud	2100.73	45.83	139.0	187.0	48.0	58.0	188.0

Como se puede observar en las tablas 6.8 y 6.9, ya se tienen los datos de dispersión y ahora es interpretar qué dicen en conjunto con los centrales variable por variable. Los datos de “latitud” y “longitud”, nos indican que el recorrido ocurre relativamente cerca al punto en donde se guarda el carro y que su distribución es, al parecer, normal o uniforme. Estas distribuciones tienen lógica ya que el trayecto es ida y vuelta por el mismo camino. Si no hubiese sido por el mismo camino la distribución sería diferente. Los valores de los cuartiles, mediana, mínimo y máximo, indican que los datos están muy concentrados; dando un rango total de los datos de 0,26 con una desviación estándar de 0,1 para la latitud, y 0,45 con una desviación estándar de 0,158 para la longitud en el día ocho de septiembre. Para el día nueve de septiembre dan unos valores similares.

Para ambos días en el “nivelRestanteEnergia” la media es inferior a la mediana, pero no están alejados de manera significativa por lo que todavía podría seguir una distribución normal. El rango total de los datos está en 37,49 en el día ocho y 42,78 en el día nueve, dando que la diferencia

entre las dos pruebas es muy poca. Además, es de anotar que el mínimo del trayecto para el día ocho es el máximo para el día nueve, indicando que no se volvió a realizar la carga de las baterías del vehículo en la prueba. El rango inter-cuartílico es de 26,19 y 25,415, con los datos del **IQR** se puede sacar cuánto representa el rango del 50 % de los datos en el rango completo y obtener que para el día ocho representa un 0,7, y para el día nueve un 0,6 del rango total. Los valores de cuánto ocupa el cincuenta porciento de los datos en el rango total indican que la mayoría del rango es ocupado por la mitad de los datos y con el valor de la desviación estándar dando indicios de estar no tan dispersos.

En la variable de “kilometrosOdometro” se observa que el mínimo y el cuartil 1 tienen igual valor para las dos pruebas, indicando que efectivamente hay una distribución sesgada a la derecha. Ahora, para establecer qué tan dispersos están los datos, se obtiene el valor de representación del **IQR** en el rango total en cada trayecto, obteniendo una representación del 0,75 para el día ocho y del 0,6 para el día nueve. El análisis es el mismo para “kilometrosAcumTrayecto” ya que ambas columnas representan lo mismo solo que se cambia el punto de referencia y esta variable va a ser importante al momento de graficar. Por otro lado, en el caso de las “revolucionesMotor”, se tiene algo similar; el cuartil 1 y el valor mínimo son iguales en ambos casos, además los valores del cuartil 2 y el máximo para ambas pruebas son similares por ende la representación del **IQR** en el rango total en ambos días es similar y da alrededor de 0,77 del rango total.

Con la “temperaturaBaterias” parece tener una tendencia normal, ya que los datos de tendencia central son muy similares. La dispersión de los datos no es alta ya que los rangos totales e intercuartílicos no son muy amplios y están cerca del valor de la desviación estándar. El mismo proceso pasa con “kilometrosEntreMuestra” solo que este está sesgado a la derecha por lo que el mínimo y el cuartil 1 son iguales para los dos trayectos pero indicando una representación del **IQR** sobre el rango total de un 0,33 y de un 0,2 dando indicios de una mayor dispersión para el resto de los datos.

Por último se tienen los datos de “altitud”, en el que se halla que las medias para ambos trayectos son inferiores a la mediana, indicando un sesgo a la izquierda. Además, la representación del **IQR**

en el rango total es del 0,6 y del 0,37 para cada día respectivamente, indicando que los datos están más repartidos en el día ocho, y para el día nueve el cincuenta porciento están muy condensados.

La representatividad de la media para cada variable se evalúa utilizando el coeficiente de variación (*CV*), previamente descrito en el marco teórico. Los resultados, presentados en la Tabla 6.10, indican que las variables “revolucionesMotor”, “kilometrosAcumTrayecto”, “kilometrosEntreMuestra” y “altitud” presentan un *CV* superior al 20 %, lo cual sugiere que no son representativas para los días analizados. Esta conclusión es coherente con la presencia de distribuciones sesgadas observadas en los datos, proporcionando un fundamento adicional para esta determinación.

TABLA 6.10
COEFICIENTE DE VARIACIÓN PARA LAS VARIABLES ANALIZADAS EL 8 Y 9 DE SEPTIEMBRE

Variable	Día 8	Día 9
latitud	0.35 %	0.34 %
longitud	-0.14 %	-0.15 %
nivelRestanteEnergia	16.18 %	29.51 %
kilometrosOdometro	0.45 %	0.49 %
temperaturaBaterias	3.63 %	8.3 %
revolucionesMotor	121.98 %	145.24 %
kilometrosAcumTrayecto	92.07 %	109.28 %
kilometrosEntreMuestra	136.3 %	169.32 %
altitud	30.94 %	28.93 %

6.2.3. Cálculo de estadísticos de forma

Con el cálculo de los dos estadísticos de forma más importantes, que son el tercer y cuarto momento estadístico, podremos comprobar hasta cierto punto la forma de las características. De este modo, se aplican las ecuaciones (4.7) y (4.8) a las diferentes columnas de cada trayecto, obteniendo la tabla 6.11.

De la tabla 6.11 se puede observar que, a excepción de los dos días de la características “kilometrosEntreMuestra”, el resto de las características tienen una curtosis que indica que las colas son platycúrticas, en cambio para “kilometrosEntreMuestra” indicaría que es leptocúrtica.

Por otro lado, tenemos los valores de asimetría: Primero, se tiene que “latitud”, “longitud”, “nivelRestanteEnergia”, “temperaturaBaterias” y “altitud”, tienen valores negativos que indican un

TABLA 6.11
TERCER Y CUARTO MOMENTO ESTADÍSTICO

Variable	Skewness		Curtosis	
	Día 8	Día 9	Día 8	Día 9
latitud	-0.608	-1.171	1.690	2.60
longitud	-0.708	-1.277	1.988	2.994
nivelRestanteEnergia	-0.414	-0.78	1.717	2.225
kilometrosOdometro	0.304	0.804	1.581	2.276
temperaturaBaterias	-0.291	-0.371	1.921	1.696
revolucionesMotor	0.552	0.862	1.458	1.870
kilometrosAcumTrayecto	0.304	0.803	1.581	2.276
kilometrosEntreMuestra	1.120	2.077	3.742	9.252
altitud	-0.722	-1.282	1.932	2.962

sesgo hacia la izquierda. En el caso de “kilometrosOdometro”, “revolucionesMotor”, “kilometrosAcumTrayecto” y “kilometrosEntreMuestra”, tienen un resultado de asimetría positiva que indica una asimetría a la derecha, de esta forma se confirma lo que se venía estableciendo a partir de los otros estadísticos. Luego de verificar los supuestos, es importante obtener visualizaciones que muestren las distribuciones de las diferentes características.

6.2.4. Gráficas uni-variadas

La obtención de los principales estadísticos de las diferentes características permite la visualización de las distribuciones de todas las variables en consideración. El trazado de estas distribuciones facilita la comprensión de la estructura y el comportamiento de las variables, revelando patrones y tendencias que no son fácilmente perceptibles a través del análisis estadístico básico.

La representación gráfica de las distribuciones comienza con la segmentación de las variables en función de los resultados de la asimetría. Por consiguiente, las figuras 6.1, 6.2, 6.3, 6.4, y 6.5 corresponden a valores de asimetría negativos, mientras que las figuras 6.6, 6.7, 6.8, y 6.9 representan

valores de asimetría positivos. Cada conjunto de gráficos proporciona una visión integral de la variable en estudio, incluyendo el histograma de frecuencias, el porcentaje acumulativo de datos y un diagrama de caja construido a partir de los cuartiles previamente calculados. Este enfoque visual enriquece nuestra comprensión del comportamiento de las variables y sirve de base para la propuesta de posibles modelos que puedan optimizar el modelo de negocio existente o explorar nuevas oportunidades comerciales.

La comparación de las trayectorias representadas en las figuras 6.1, 6.2, y 6.5 sugiere una similitud notable en el desplazamiento del autobús desarrollado en el terreno. Los histogramas muestran dos picos de frecuencia en los extremos de la distribución, probablemente reflejando las paradas prolongadas del autobús. Esta suposición se refuerza al observar la distribución de los niveles de energía en la figura 6.3, donde se observa una concentración de datos en los extremos de energía. En la figura 6.3b, en particular, se aprecian picos de frecuencia entre el 30 %-40 % y 50 %-60 %, lo que indica un mayor número de paradas en comparación con la figura 6.3a.

A la luz de los análisis gráficos, es importante destacar la relevancia de comprender y utilizar apropiadamente las técnicas de visualización de datos. La habilidad para interpretar correctamente estas representaciones visuales puede proporcionar una mayor comprensión de los datos, destacando patrones ocultos, tendencias y relaciones que no se pueden descubrir mediante métodos puramente estadísticos. Además, los gráficos permiten una comunicación más efectiva y accesible de los resultados, facilitando la discusión entre distintos miembros del equipo o partes interesadas, incluyendo a aquellos sin una formación técnica o estadística extensa.

En el contexto de análisis de *Big Data*, la visualización de datos también desempeña un papel crucial en la exploración y análisis preliminar de grandes volúmenes de datos. Puede ayudar a los analistas a entender la estructura subyacente de los datos, detectar anomalías y errores, y tomar decisiones informadas sobre las técnicas de análisis y modelado más adecuadas. Por lo tanto, la capacidad de crear y utilizar eficazmente las representaciones visuales de datos se ha convertido en una habilidad esencial en el mundo moderno de la analítica de datos.

El análisis de la figura 6.4 revela un comportamiento intrigante, pues se espera una distribución normal para la temperatura de la batería del autobús. En la figura 6.4a, donde el autobús hace pocas paradas, emergen dos estados distintos. Podemos denominar un estado como "caliente" el otro como "moderadamente caliente", posiblemente correlacionándose con los momentos de actividad y reposo del vehículo, respectivamente. En contraste, la figura 6.4b muestra una mayor presencia de valores intermedios, lo que sugiere una transición más gradual entre los estados de reposo y movimiento del autobús.

En relación a los kilómetros marcados por el odómetro, estos se pueden comparar con el acumulado de kilómetros, ambos exhibiendo una forma de distribución similar. Como se mencionó

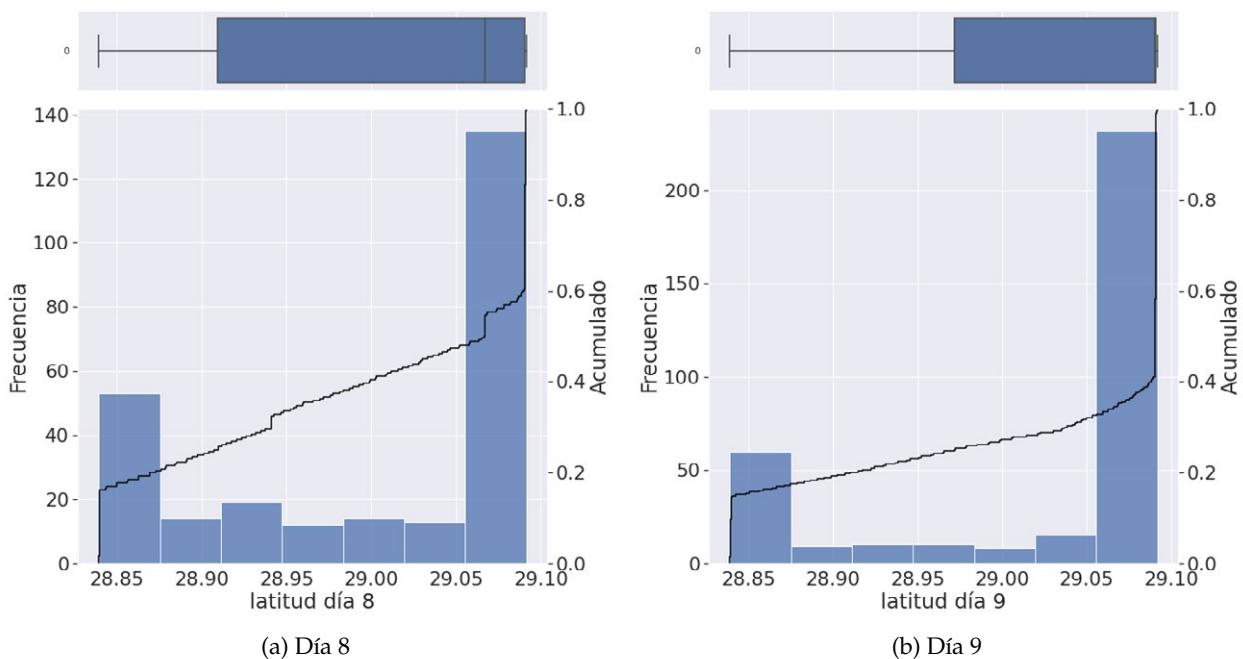


Fig. 6.1: Histograma Latitud

previamente, los kilómetros acumulados por trayecto se introdujeron para facilitar el análisis de los recorridos; en realidad, solo alteran el punto de referencia de los kilómetros del odómetro, como se puede apreciar en las figuras 6.6 y 6.8. Una inspección detallada de las gráficas revela

que los acumulados siguen las mismas tendencias observadas en las gráficas de las baterías. En efecto, durante el día ocho, se observan dos largas paradas del vehículo al inicio y al final, mientras que en el día nueve, además de las paradas al principio y al final, se observan dos paradas significativas a lo largo del recorrido.

Finalmente, las variables “revolucionesMotor” y “kilometrosEntreMuestra” muestran un comportamiento comparable para ambos días. En ambos casos, se observan dos rangos con alta frecuencia en las gráficas: uno entre 0 y 250 y otro entre 1750 y 2500. Esto nos proporciona indicios sobre el estilo de conducción. Otro indicador es la distancia recorrida entre cada muestra, que podría ser una huella característica de cada conductor.

Es importante destacar que la identificación de patrones en las distribuciones de los datos, como los observados en las figuras mencionadas anteriormente, no solo proporciona una visión detallada del comportamiento del sistema en estudio, sino que también informa la construcción y la afinación de modelos predictivos. Estos patrones, que reflejan el funcionamiento subyacente del

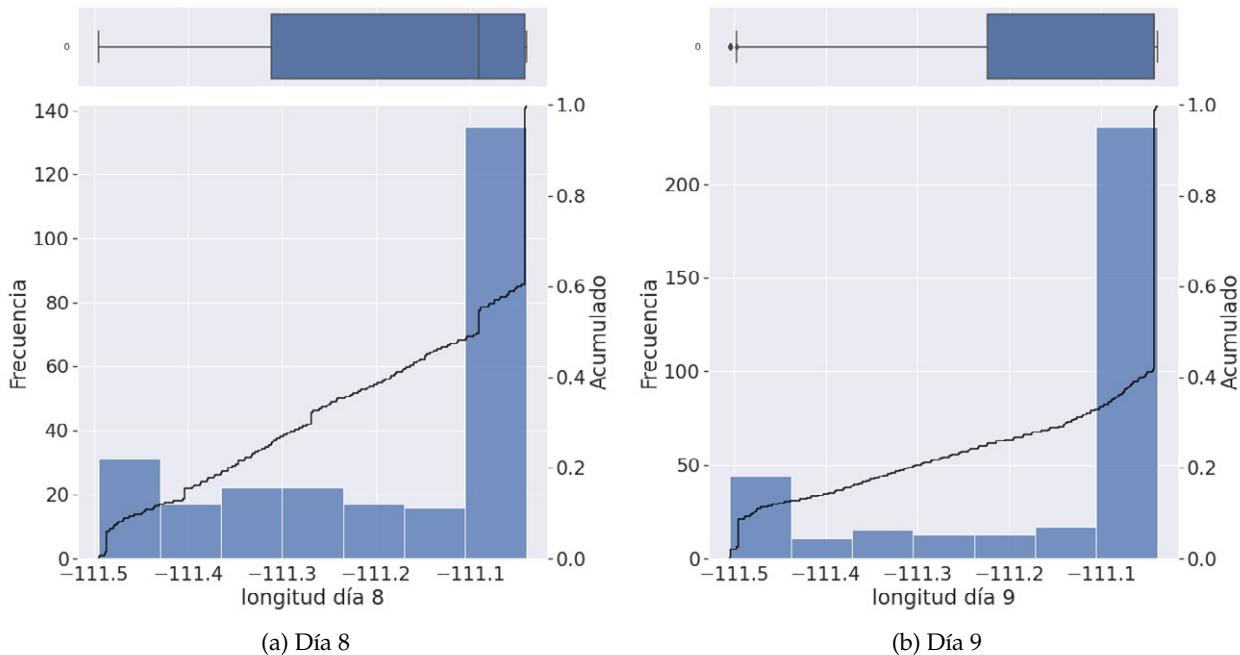
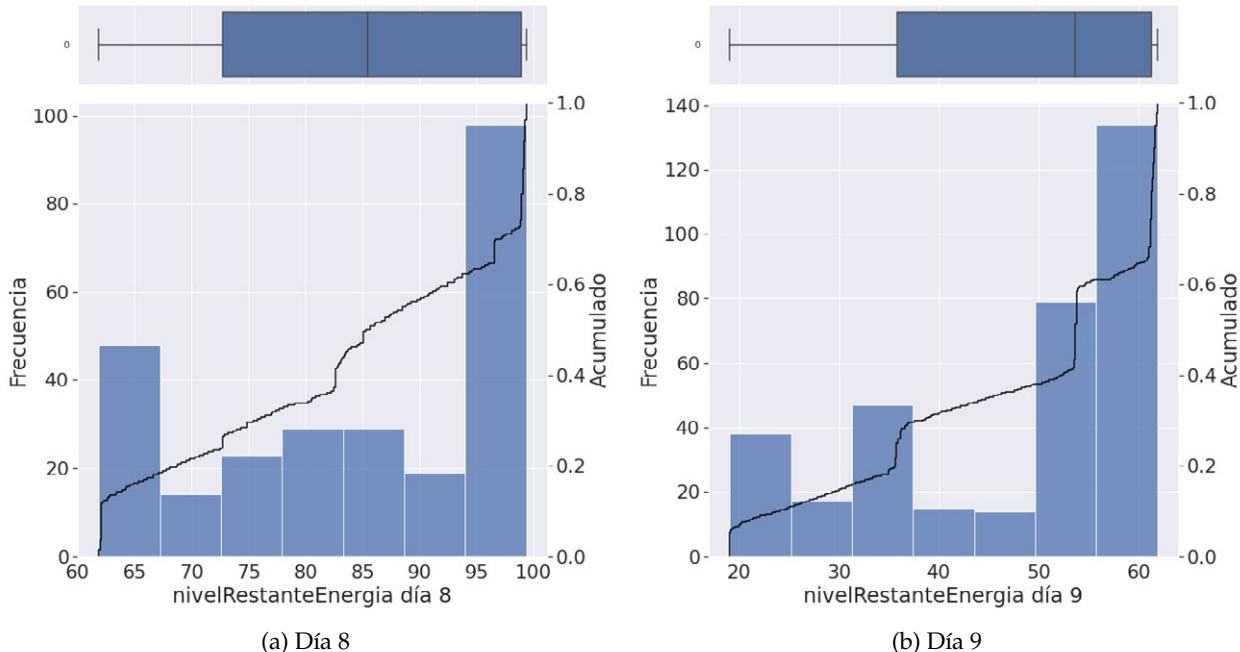


Fig. 6.2: Histograma Longitud

sistema, guían la selección de las características apropiadas y la estructura del modelo para capturar eficazmente las dinámicas del sistema.

Además, el análisis de los datos en la escala temporal proporciona información crítica sobre la evolución del sistema. Por ejemplo, la comparación de los datos entre los días revela las fluctuaciones diarias y podría indicar la presencia de ciclos diarios en el sistema. A su vez, esto podría utilizarse para mejorar el pronóstico y la planificación, permitiendo prever cuándo es probable que ocurran ciertas condiciones y programar las operaciones de acuerdo a ello.

Finalmente, es valioso señalar que el análisis detallado realizado aquí sienta las bases para investigaciones más profundas. Por ejemplo, podríamos investigar más a fondo las causas de los dos estados de temperatura de la batería y las paradas prolongadas del autobús, y estudiar su impacto en el rendimiento general del sistema. Asimismo, los patrones observados en las revoluciones del motor y los kilómetros recorridos entre cada muestra podrían analizarse más a fondo para desarrollar perfiles de conducción más detallados y entender cómo los diferentes estilos de conducción afectan al sistema.



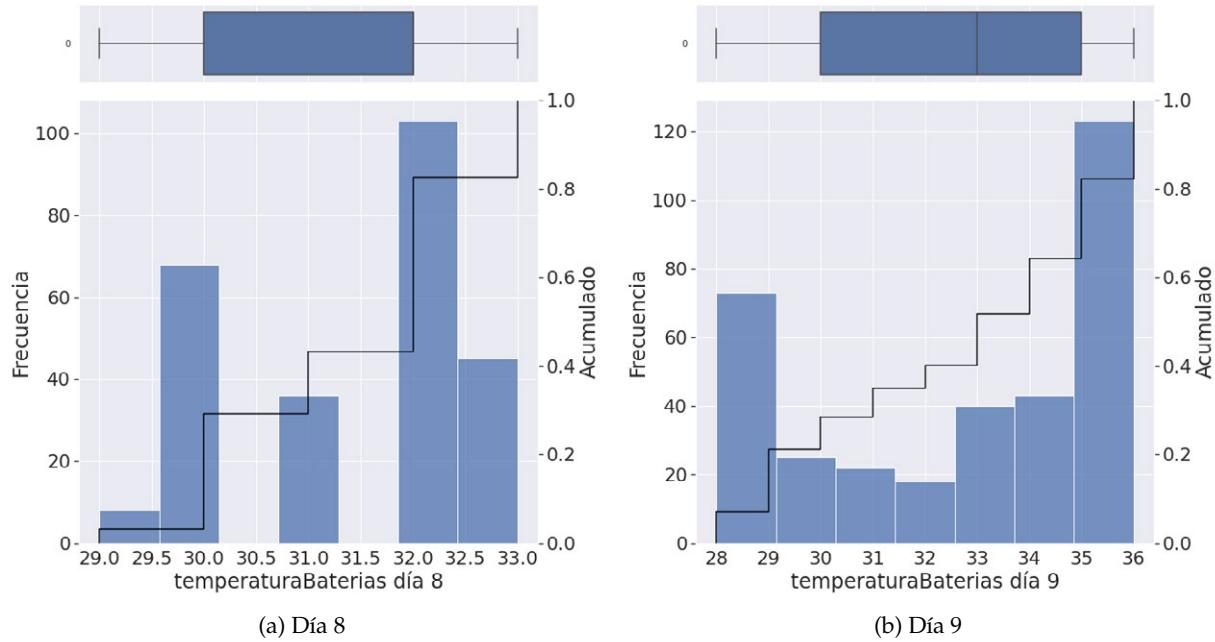


Fig. 6.4: Histograma Temperatura Baterías

A partir del entendimiento uni-variado se pueden hacer relaciones entre variables y esto nos da paso a entender cómo se relacionan las variables de manera multi-variada, es decir, cómo se comportan unas contra otras.

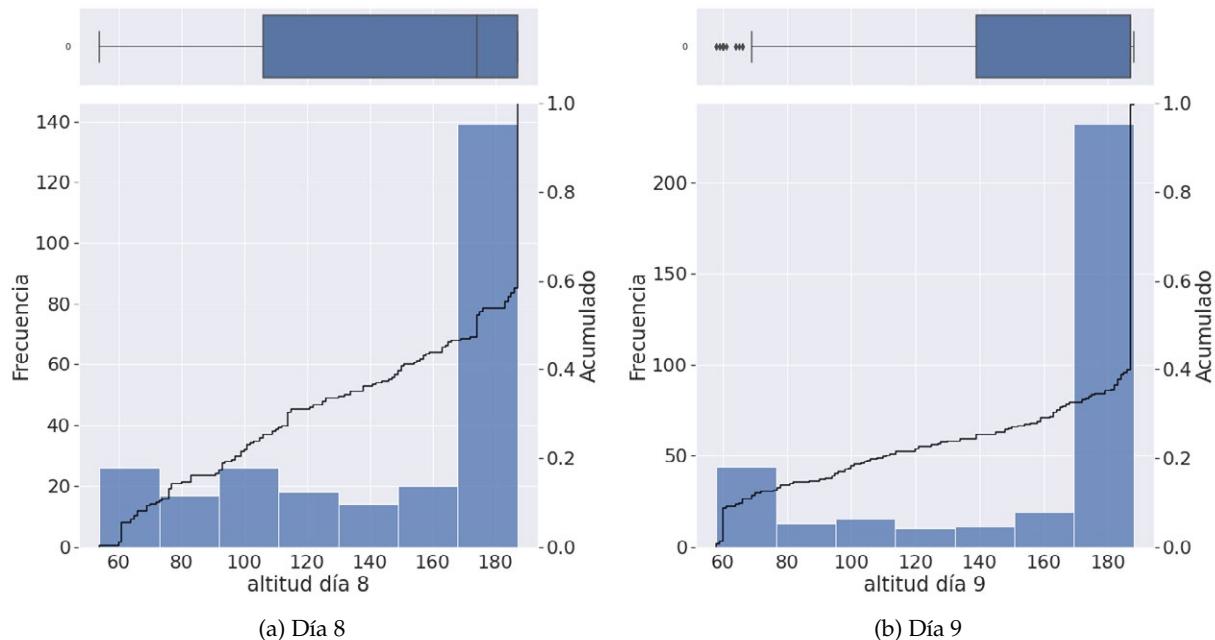


Fig. 6.5: Histograma Altitud

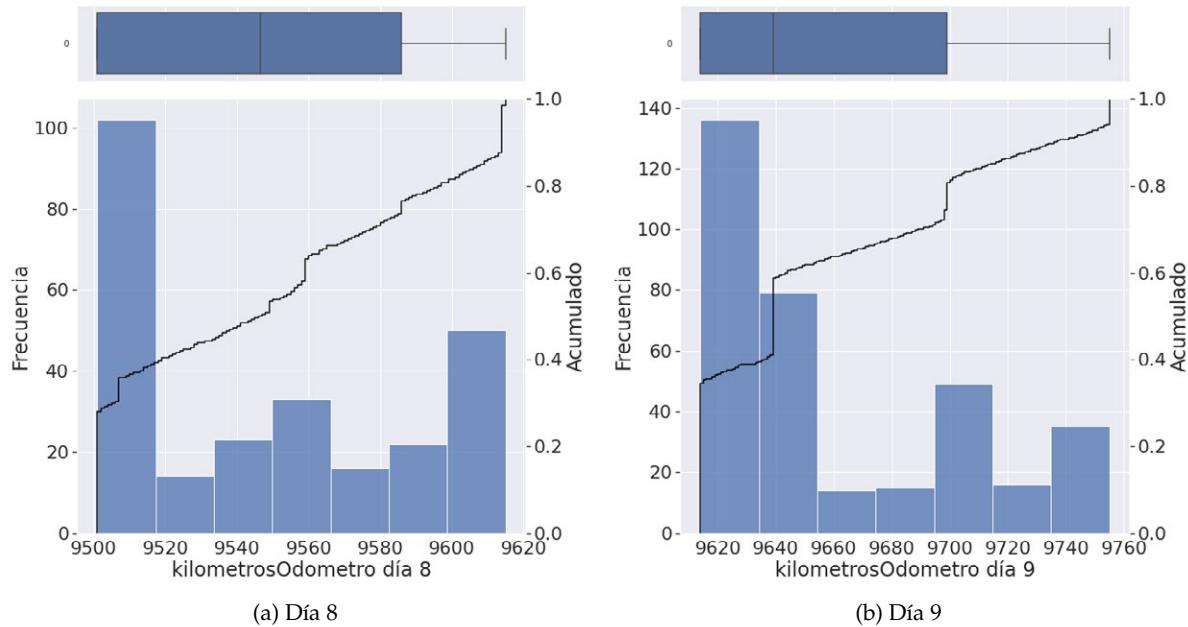


Fig. 6.6: Histograma Kilómetros Odómetro

6.2.5. Análisis multi-variado

Como en un experimento se puede medir, calcular, y tomar diferentes variables se empieza entendiendo una por una su comportamiento individual; luego de saber ese comportamiento, se

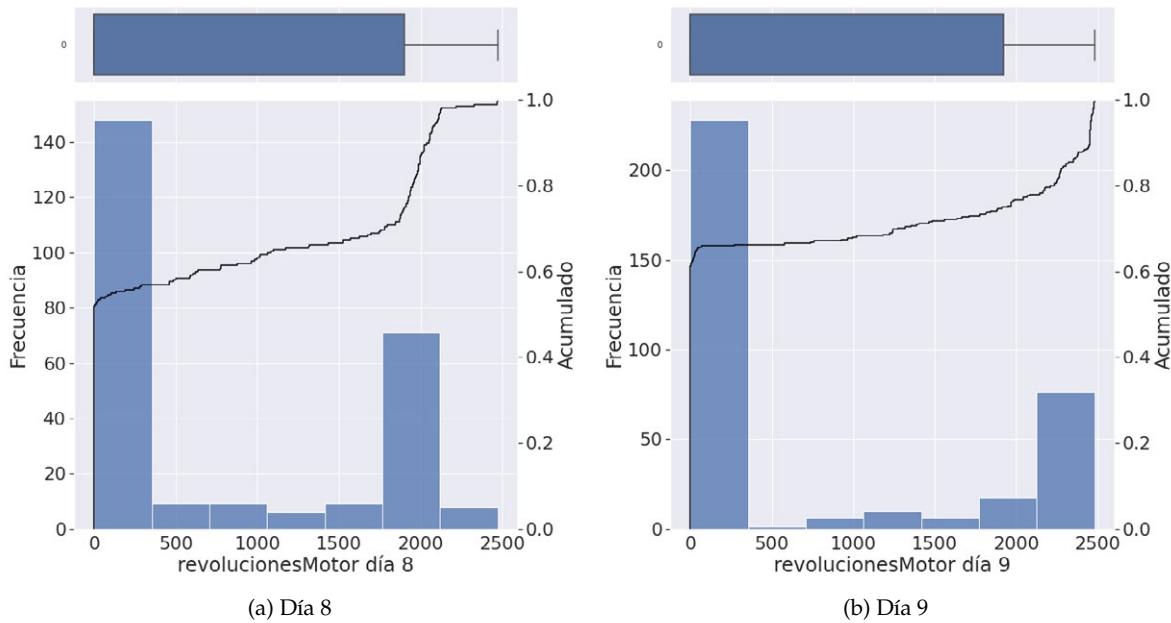


Fig. 6.7: Histograma Revoluciones del Motor

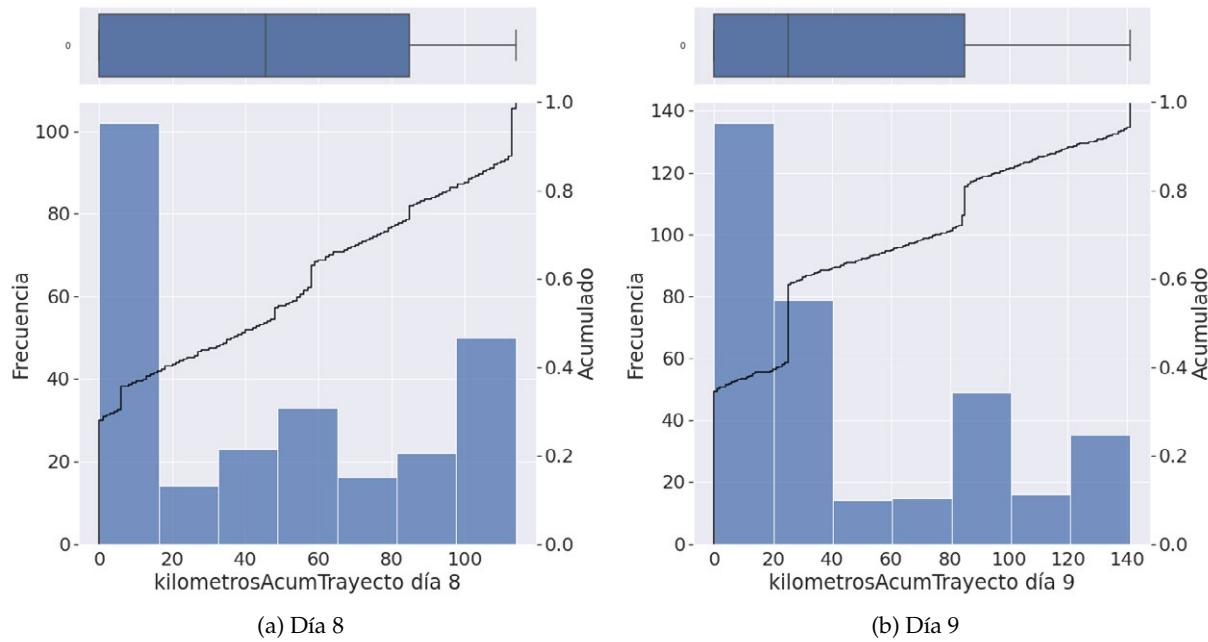


Fig. 6.8: Histograma Kilómetros Acumulados por Trayecto

prosigue con la interacción que hay entre las variables, para esto se visualiza y simplifica el conjunto, se realiza búsqueda de relaciones, y semejanzas.

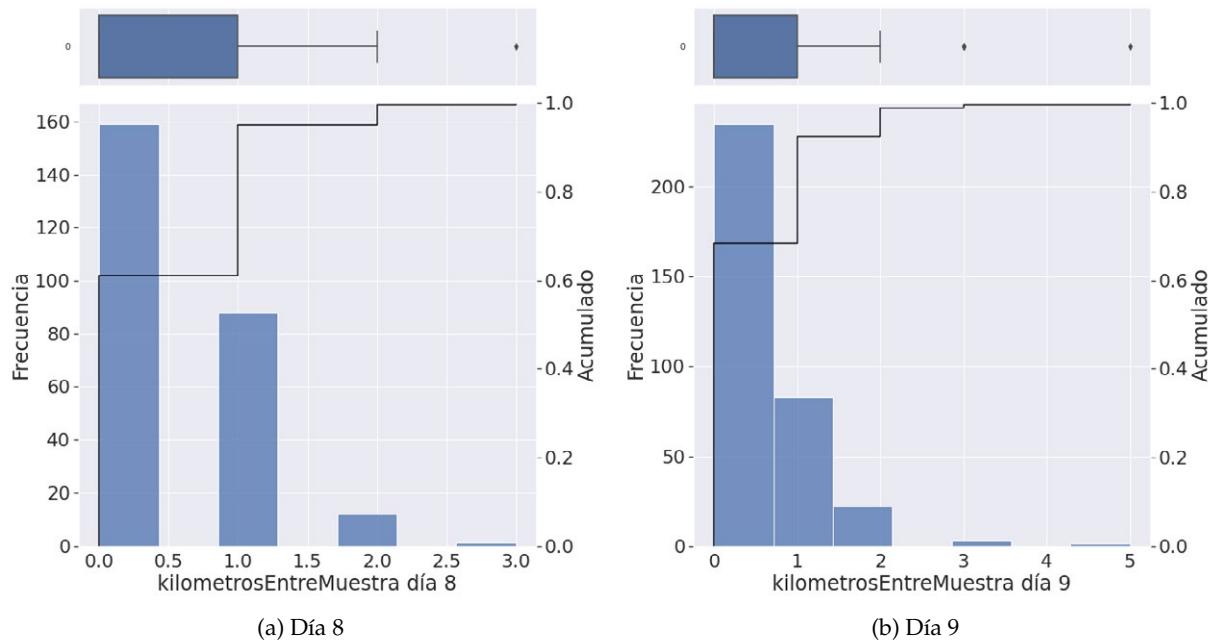


Fig. 6.9: Histograma Kilómetros entre Muestras

Como ya se tiene de la sección de estadísticos de tendencia central las medias para cada característica, se puede ahora construir el vector de medias. Se debe tener cuidado al construir el vector de medias, ya que cada valor debe ir en la posición de su respectiva característica; si la matriz de datos es $n \times m$ entonces el vector de medias debe ser $m \times 1$ y se expresa de la forma mostrada en (6.2).

$$\bar{x} = \frac{1}{n} \sum_i^n x_{ij} = \begin{bmatrix} x_1 \\ \dots \\ x_m \end{bmatrix} \quad (6.2)$$

Con el vector de medias se calcula la matriz corregida y se realiza el cálculo de la matriz de covarianzas.

Como se puede observar en las tablas 6.12, 6.13, 6.14, 6.15, el signo nos da indicios de qué tipo de relación hay, pero la magnitud se vuelve difícil de interpretar por las magnitudes de las características. Por ejemplo en “revolucionesMotor”, “kilometrosAcumTrayecto”, y “kilometrosOdometro” varios de los valores que los tienen en cuenta son bastante grandes, y difíciles de interpretar. También se puede observar que los valores para “kilometrosAcumTrayecto”, y “kilometrosOdometro” son exactamente iguales ya que simplemente se hizo un cambio de referencia y ambas características nos hablan de lo mismo; esto se debe tener en cuenta, dado el caso que se desee desarrollar modelos para predecir o interpretar conjuntos de muestras.

TABLA 6.12
MATRIZ DE COVARIANZA DÍA 8 PARTE I

	latitud	longitud	kilometrosOdometro	nivelRestanteEnergia
latitud	0.010079	0.015808	-0.987312	0.241424
longitud	0.015808	0.025162	-1.545663	0.374196
kilometrosOdometro	-0.987312	-1.545663	1820.411331	-581.726156
nivelRestanteEnergia	0.241424	0.374196	-581.726156	186.699406
revolucionesMotor	-35.595142	-52.135424	7993.125379	-2557.412325
temperaturaBaterias	-0.050719	-0.079886	43.662890	-13.848398
kilometrosAcumTrayecto	-0.987312	-1.545663	1820.411331	-581.726156
kilometrosEntreMuestra	-0.023724	-0.034871	5.226388	-1.618762
altitud	4.506879	7.133644	-287.069216	58.889142

**TABLA 6.13
MATRIZ DE COVARIANZA DÍA 8 PARTE II**

	revolucionesMotor	temperaturaBaterias	kilometrosAcumTrayecto	kilometrosEntreMuestra	altitud
latitud	-35.595142	-0.050719	-0.987312	-0.023724	4.506879
longitud	-52.135424	-0.079886	-1.545663	-0.034871	7.133644
kilometrosOdometro	7993.125379	43.662890	1820.411331	5.226388	-287.069216
nivelRestanteEnergia	-2557.412325	-13.848398	-581.726156	-1.618762	58.889142
revolucionesMotor	835848.860870	421.224874	7993.125379	369.066305	-14641.671176
temperaturaBaterias	421.224874	1.302331	43.662890	0.257871	-19.803698
kilometrosAcumTrayecto	7993.125379	43.662890	1820.411331	5.226388	-287.069216
kilometrosEntreMuestra	369.066305	0.257871	5.226388	0.363454	-9.887363
altitud	-14641.671176	-19.803698	-287.069216	-9.887363	2099.427666

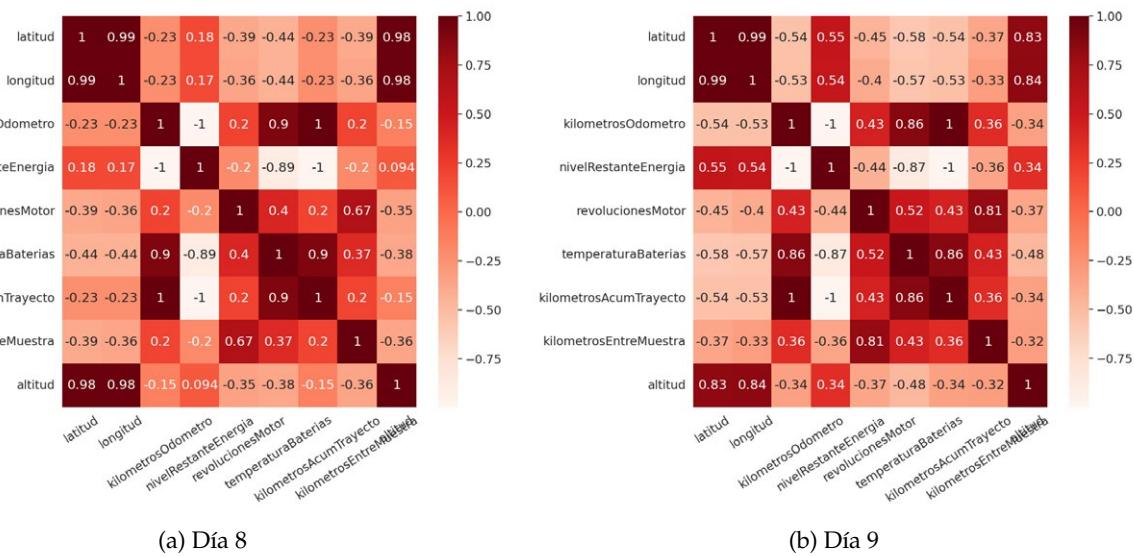
**TABLA 6.14
MATRIZ DE COVARIANZA DÍA 9 PARTE I**

	latitud	longitud	kilometrosOdometro	nivelRestanteEnergia
latitud	0.009465	0.015692	-2.477161	0.759537
longitud	0.015692	0.026378	-4.074255	1.246603
kilometrosOdometro	-2.477161	-4.074255	2209.151527	-669.712607
nivelRestanteEnergia	0.759537	1.246603	-669.712607	203.108813
revolucionesMotor	-45.189633	-66.290172	20758.465837	-6466.544679
temperaturaBaterias	-0.152906	-0.250028	109.877254	-33.541039
kilometrosAcumTrayecto	-2.477161	-4.074255	2209.151527	-669.712607
kilometrosEntreMuestra	-0.024905	-0.036706	11.626153	-3.606463
altitud	3.710876	6.244401	-732.207717	224.284404

**TABLA 6.15
MATRIZ DE COVARIANZA DÍA 9 PARTE II**

	revolucionesMotor	temperaturaBaterias	kilometrosAcumTrayecto	kilometrosEntreMuestra	altitud
latitud	-4.518963e+01	-0.152906	-2.477161	-0.024905	3.710876
longitud	-6.629017e+01	-0.250028	-4.074255	-0.036706	6.244401
kilometrosOdometro	2.075847e+04	109.877254	2209.151527	11.626153	-732.207717
nivelRestanteEnergia	-6.466545e+03	-33.541039	-669.712607	-3.606463	224.284404
revolucionesMotor	1.048645e+06	1446.225905	20758.465837	576.317064	-17313.307267
temperaturaBaterias	1.446226e+03	7.359067	109.877254	0.816962	-59.866423
kilometrosAcumTrayecto	2.075847e+04	109.877254	2209.151527	11.626153	-732.207717
kilometrosEntreMuestra	5.763171e+02	0.816962	11.626153	0.481651	-10.252517
altitud	-1.731331e+04	-59.866423	-732.207717	-10.252517	2100.733397

Ahora se puede realizar el cálculo del coeficiente de correlación; para este caso particular se va a utilizar el coeficiente de correlación de Pearson, que es sensible solo ante las relaciones lineales, y se puede observar en la figura 6.10. Como se puede percibir en las matrices con los coeficientes de correlación, los valores que se acercan más a rojo representan los valores que tienen una correlación lineal positiva, y los que se aproximan a menos uno, una correlación lineal negativa. Entre “revolucionesMotoR”, “temperaturaBaterias”, “KilometrosAcumTrayecto”, y “kilometrosEntreMuestra” hay una relación positiva moderada y fuerte. Por otro lado, está la relación entre las variables antes mencionadas con la longitud y latitud, entregando una relación moderadamente negativa.



(a) Día 8

(b) Día 9

Fig. 6.10: Mapa de Calor para el Coeficiente de Correlación

Pero es bueno poder ver cómo es la relación entre variables, por eso se hace una gráfica en la que se muestran los puntos al comparar dos variables y en las distribuciones cuando se compara consigo misma, ver figura 6.11. Como se puede observar entre “kilometrosAcumTrayecto” y “kilometrosOdometro” junto con “nivelRestanteEnergia”, hay relaciones lineales obvias igual que en el caso de la longitud con la latitud, y altitud.

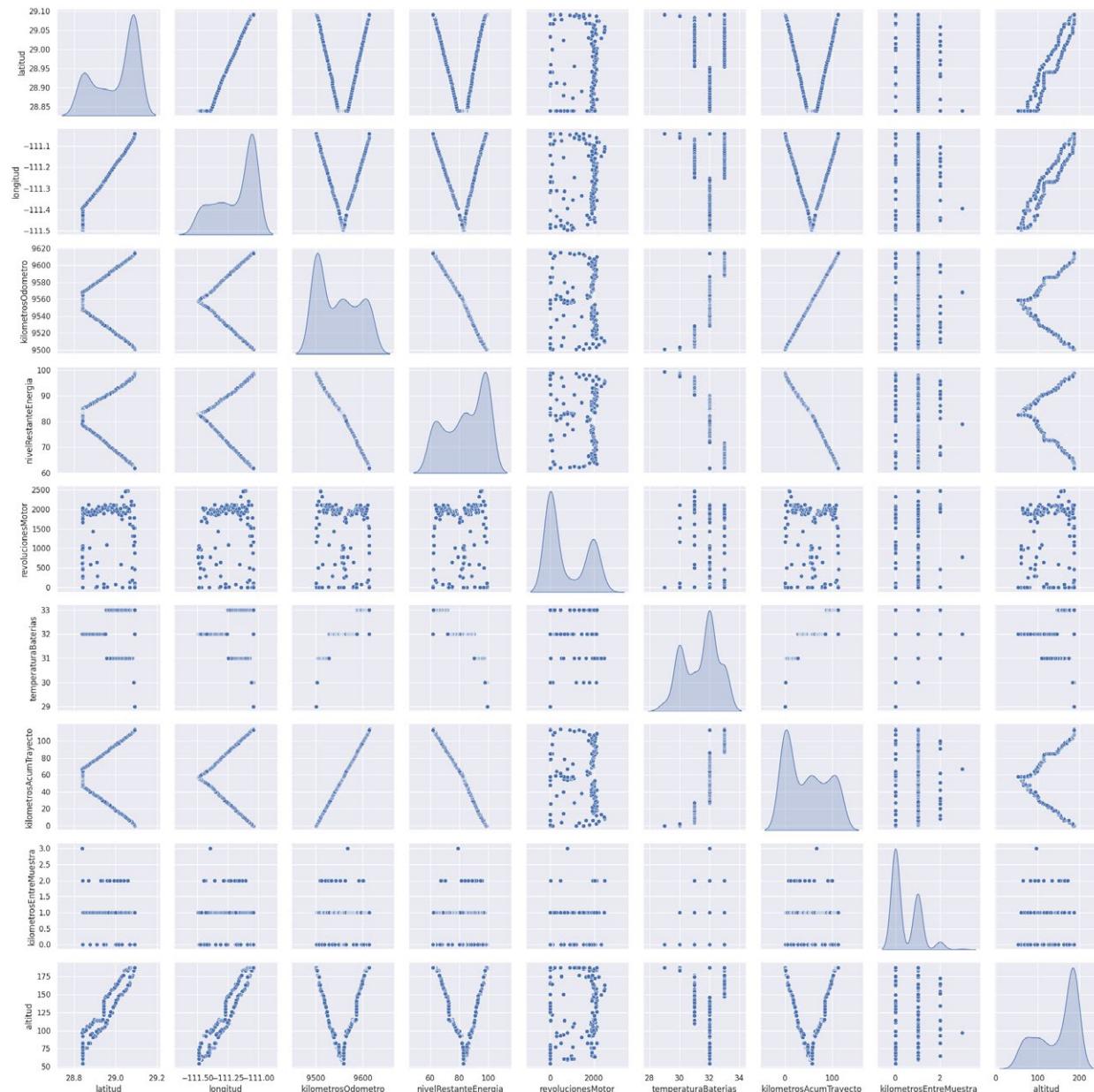


Fig. 6.11: Gráfica Correlación

Ahora que se tiene una idea de cómo es el comportamiento de las variables, se puede tratar de buscar reducirlas manteniendo el total de la información. Una técnica es a través de los autovalores y los vectores propios, estableciendo si una de estas características es una combinación lineal de las otras. Lo que se busca con la reducción de dimensionalidad es tratar de obtener grupos separables de manera visual, una de estas técnicas es PCA. Como se puede observar, en las figuras 6.12, 6.13 se presentan al parecer tres grupos de datos de las gráficas de dos componentes, en el caso de las gráficas de tres componentes no logra encontrarse una forma de agrupar visualmente los conjuntos de datos. Luego de establecer los grupos, se hace la respectiva marcación y se analizan los grupos de datos por separado para poder interpretar a qué hacen referencia estos conjuntos. Cuando se tienen conjuntos de datos pequeños como los que se utilizan en el presente documento, se puede apoyar en este tipo de metodologías para obtener análisis, resultados, y además poder establecer si se puede aplicar otro tipo de modelos que ayuden a la compañía, como metodologías de detección de atípicos, agrupamientos más sofisticados o predicciones de comportamiento de fallos.

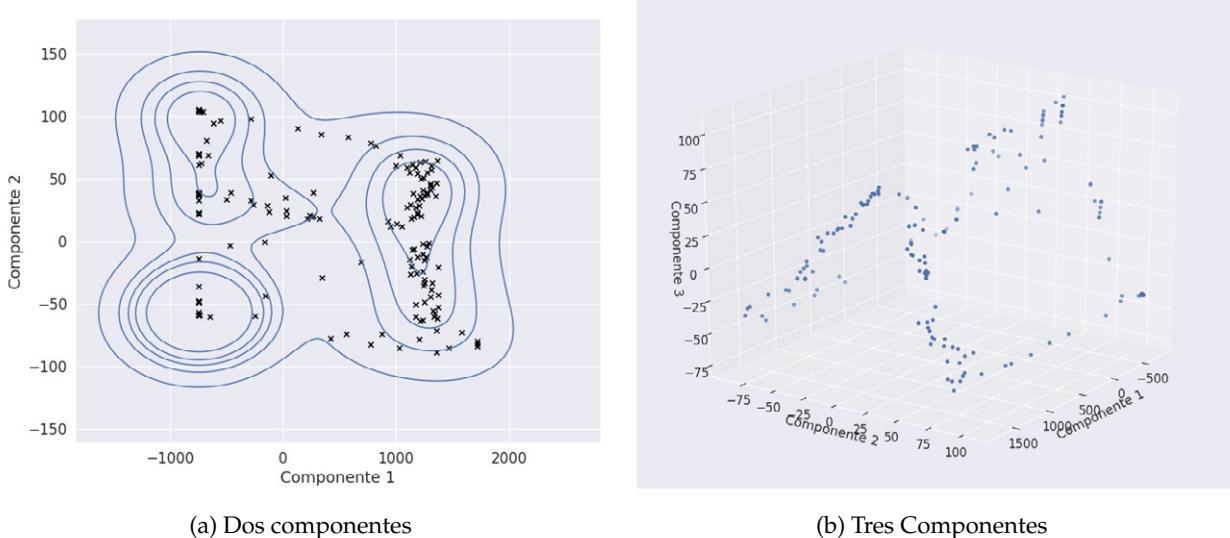


Fig. 6.12: Datos Obtenidos de PCA Día 8

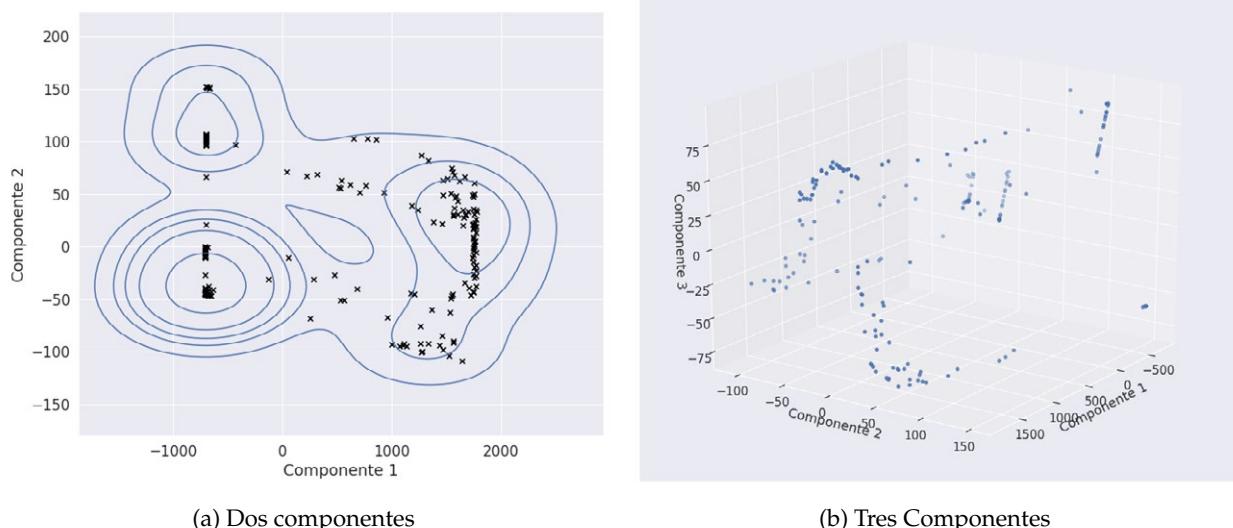


Fig. 6.13: Datos Obtenidos de PCA Día 9

7

CAPÍTULO SIETE

Consideraciones

A medida que se atraviesa la era de la información, nos enfrentamos a grandes desafíos y oportunidades en el campo del análisis de datos. Este capítulo aborda consideraciones críticas y temas relevantes que emergen en la intersección de la teoría estadística, la computación de alto rendimiento y la toma de decisiones basada en datos. Profundizaremos en la importancia de la integración de datos y analítica avanzada, la necesidad de estrategias efectivas para manejar el creciente flujo de información y el uso de herramientas para procesar grandes volúmenes de datos sin comprometer la memoria volátil.

Para enriquecer nuestro análisis y superar las limitaciones de un conjunto de datos reducido, hemos implementado técnicas de inteligencia artificial para generar datos sintéticos a través de diferentes enfoques. Utilizando métodos de Data Augmentation, como redes generativas adversarias tabulares (*Tabular GANs*), logramos expandir significativamente nuestro conjunto de datos original.

La expansión de nuestro conjunto de datos, desde un modesto inicio hasta una colección de 118,079 muestras, ilustra el poder de la Data Augmentation y establece el escenario para discu-

siones avanzadas sobre su correcta aplicación y validación. En este contexto, exploraremos cómo la ley de los grandes números se convierte en una aliada indispensable, y cómo metodologías como el *bootstrapping*, una técnica estadística que permite estimar la precisión de los estimadores al re-muestrear con reemplazo las muestras existentes, pueden adaptarse para el análisis en entornos de *Big Data*. El bootstrapping es especialmente útil para validar modelos y obtener intervalos de confianza cuando se trabaja con datos aumentados, complementando así las técnicas de Data Augmentation.

A su vez, consideraremos la utilidad práctica de tecnologías como Apache Spark, un motor de análisis unificado diseñado para el procesamiento de datos a gran escala. Apache Spark permite manejar y procesar eficientemente grandes volúmenes de información mediante su capacidad de computación en memoria y distribución de tareas. Su inclusión en este contexto es fundamental para abordar los desafíos de almacenamiento y procesamiento que surgen al trabajar con conjuntos de datos expansivos, como los generados a través de Data Augmentation. Además, discutiremos la implementación de *Dashboards* interactivos para el seguimiento y visualización en tiempo real de métricas clave. Estas herramientas no solo facilitan el análisis, sino que también democratizan el acceso a la información, permitiendo a una audiencia más amplia participar en el proceso de descubrimiento de conocimiento.

7.1. Integración para analítica

La fase inicial de nuestro análisis se basó en una muestra de aproximadamente seiscientos conjuntos de datos. Para fortalecer la robustez estadística y enriquecer el análisis, recurrimos a la herramienta *tapgan* [16], la cual nos permitió incrementar el volumen de datos hasta alcanzar un total de 118,079 muestras, cada una con nueve características distintas. La efectividad de la técnica de *Data Augmentation* se evidencia en la Tabla 7.1, donde se muestra que los datos aumentados mantienen una similitud estadística considerable con el conjunto original, preservando así la integridad de la distribución de los datos.

Mirando hacia el futuro, se espera un aumento en el volumen de datos a medida que se incrementa el número de autobuses y la duración de su operatividad. Para analizar eficazmente este crecimiento de datos, es imperativo aplicar metodologías y teorías escalables, como las técnicas de *resampling*, y en particular, el método de *bootstrapping*, que se fundamenta en la ley de los grandes números.

Además, la generación de nuevos datos a través de la augmentación no es suficiente por sí sola; es esencial validar la pertinencia y utilidad de estos datos. Los conjuntos aumentados deben ser un reflejo fiel de las propiedades y tendencias de los datos originales y, al mismo tiempo, proporcionar la diversidad necesaria para mejorar la precisión y robustez de los modelos predictivos. Esto requiere un balance entre la autenticidad y la representatividad, para prevenir el sobreajuste. Por lo tanto, se necesita un meticuloso proceso de validación para asegurar que los datos aumentados cumplan con estas condiciones. A medida que el Big Data continúa su trayectoria exponencial y las técnicas de augmentación de datos avanzan, se abren nuevas oportunidades para el desarrollo de métodos de análisis de datos más precisos y sofisticados, expandiendo las fronteras de nuestra capacidad para extraer conocimiento significativo y accionable de vastos conjuntos de datos.

Ley de los grandes números:

Si la probabilidad de que ocurra un hecho en una prueba única es p , y se hacen varias pruebas, independientemente y en las mismas condiciones, la proporción más probable de que ocurran los hechos en el número total de pruebas es también p ; aún más, la probabilidad que la proporción en cuestión difiera de p en menos de una cantidad dada, por pequeña que sea, aumenta al mismo tiempo que aumenta el número de pruebas [17].

La Ley de los Grandes Números incorpora varios teoremas que describen cómo un estadístico cambia a medida que se incrementa el número de muestras. Se lleva a cabo un promedio del estadístico de interés hasta que la muestra $k + 1$ no provoque un cambio significativo en la estimación previa.

Como se ilustra en la figura 7.1, al seleccionar muestras aleatorias del conjunto de datos, nos acercamos al valor esperado. Si extrapolamos este escenario a uno donde estamos enfrentando un problema de *Big Data* (es decir, cuando los datos son de gran volumen, rapidez y variedad, o en términos generales, cuando los datos no pueden almacenarse en la memoria flash), estas metodologías simplifican la realización de procesos en dichas bases de datos, proporcionando resultados con significancia estadística. Para la implementación de la estimación de parámetros en *Big Data*, se puede seguir el algoritmo presentado en la figura 7.2, que describe de forma sencilla cómo establecer la configuración para tal estimación.

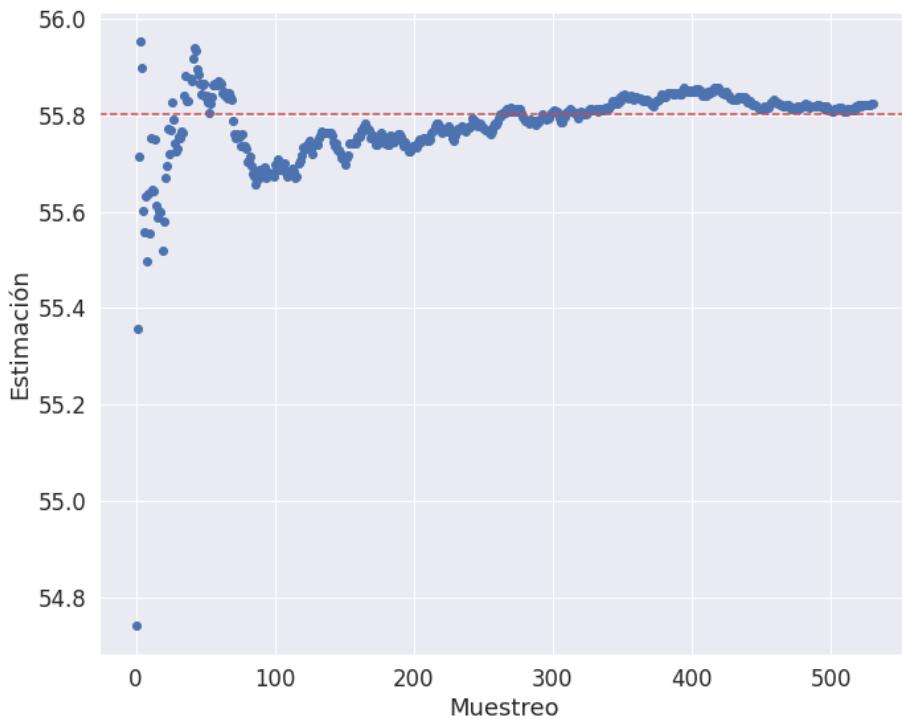


Fig. 7.1: Gráfico para la Ley de los Grandes Números

7.1.1. Experimento

El análisis exhaustivo de los trayectos realizados por cada autobús nos permite recopilar muestras detalladas que contienen información relevante de un autobús en un recorrido específico,

incluyendo diversas variables de interés. Estas variables pueden incluir, por ejemplo, la velocidad promedio, el consumo de combustible, la distancia recorrida y los tiempos de parada. A partir de esta información condensada, es posible crear visualizaciones que ilustren el comportamiento del autobús y su rendimiento a lo largo del trayecto.

Un ejemplo de este tipo de base de datos se puede observar en la tabla 7.2. Estas visualizaciones y tablas nos permiten entender mejor el desempeño de cada autobús y nos ayudan a identificar posibles áreas de mejora en términos de eficiencia y sostenibilidad.

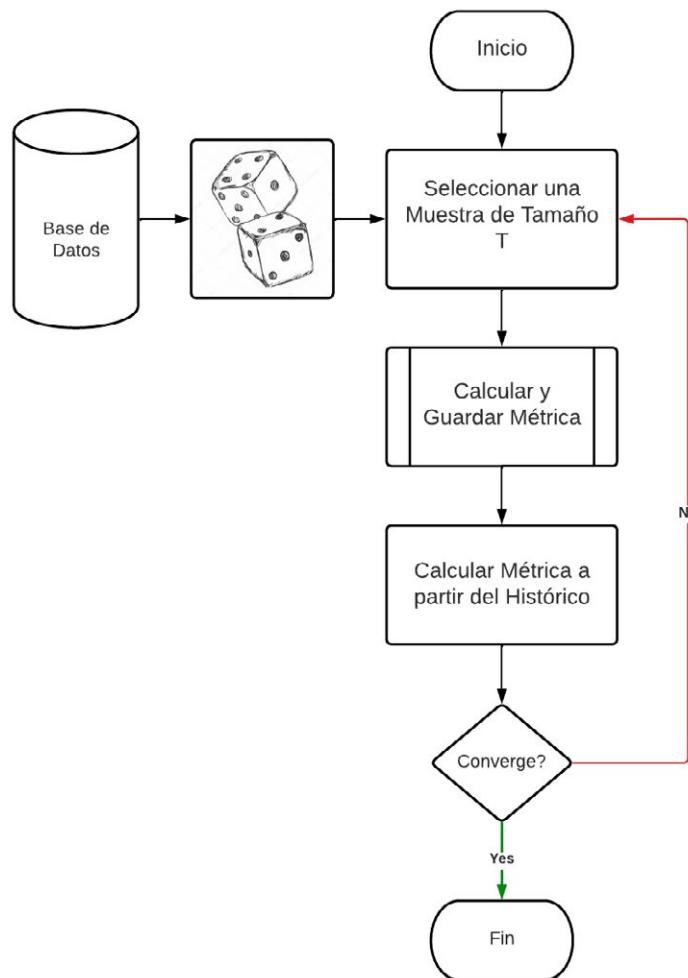


Fig. 7.2: Algoritmo Para Determinar Métricas en Bases que no Entran en Memoria Volátil

Este mismo principio se puede aplicar independientemente de la cantidad de pruebas o de trayectos analizados. En otras palabras, a medida que se recojan más datos de más recorridos, podremos seguir utilizando este enfoque para condensar y visualizar la información, lo que nos permitirá mantener un entendimiento claro y actualizado del rendimiento de los autobuses. Esto nos brinda una metodología robusta y escalable para el análisis de rendimiento en el transporte público.

7.2. Temáticas relevantes

- En el momento en el que se comience a hablar de mucho flujo de información, y para procesar información en grandes volúmenes existe también herramientas para no colapsar la memoria volátil. Una herramienta que a día de hoy es indispensable es Apache Spark; que es un motor de análisis unificado para el procesamiento de datos a gran escala, proporciona APIs en Java, Scala, Python y R, junto a un motor optimizado que soporta gráficos de ejecución general. También es compatible con un amplio conjunto de herramientas de alto nivel, como Spark SQL, para el procesamiento de datos estructurados y SQL, la API de pandas en Spark para las cargas de trabajo de pandas, MLlib para el aprendizaje automático, GraphX para el procesamiento de gráficos, Structured Streaming para el cálculo incremental y el procesamiento de flujos. Mezclando con bases de datos no estructuradas se logra sacar provecho a grandes volúmenes de información [18].
- También se pueden realizar *DashBoards* que estén conectados a la base de datos en donde esté contenida la información de los buses. El *DashBoard*, al alimentarse directamente de la base, puede obtener información visual de los indicadores que se identifiquen como necesarios. De esta manera se puede llevar un seguimiento adecuado del comportamiento de los dispositivos, algunas herramientas para este propósito se pueden desarrollar en “python”, también hay herramientas abiertas como “PowerBI”, y herramientas de pago como “Tableau” [13].

En conclusión, el análisis de datos se ha convertido en un pilar fundamental para las empresas que buscan mantenerse competitivas en un mercado en constante evolución. La capacidad de transformar datos en *insights* accionables permite a las organizaciones tomar decisiones informadas, optimizar procesos y anticiparse a las tendencias del mercado. Sin embargo, para que el análisis de datos sea realmente efectivo, es esencial que las empresas desarrollen políticas internas que promuevan la calidad y la integridad de los datos, fomenten una cultura de toma de decisiones basada en evidencia y aseguren el cumplimiento de regulaciones relacionadas con la privacidad y la seguridad de la información. Además, invertir en la formación continua de los empleados en habilidades analíticas y en la implementación de infraestructuras tecnológicas adecuadas es crucial. Estas consideraciones prácticas no solo mejoran el rendimiento operativo, sino que también pueden impulsar la innovación y el crecimiento sostenible a largo plazo.

TABLA 7.1
COMPARACIÓN CON DATA AUGMENTATION

Datos	media	min	Cuartil1	mediana	Cuartil3	max	IQR	STD	Varianza
dia8 latitud	29.002681	28.839405	28.909201	29.066874	29.090761	29.091413	0.181560	0.100394	1.00788e-02
dia9 latitud	29.028413	28.838931	28.972033	29.090410	29.090823	29.091820	0.118790	0.097290	9.465306e-03
ficti latitud	28.969615	28.838993	28.893465	28.944571	29.064358	29.091656	0.170893	0.085097	7.241445e-03
dia8 longitud	-111.181643	-111.494813	-111.311866	-111.090998	-111.041951	-111.040508	0.269915	0.158626	2.516235e-02
dia9 longitud	-111.145347	-111.504281	-111.224059	-111.042040	-111.041915	-111.038758	0.182144	0.162415	2.637848e-02
ficti longitud	-111.178431	-111.503011	-111.279473	-111.116803	-111.062732	-111.040306	0.216741	0.136744	1.869903e-02
dia8 kilometrosOdometro	9547.342308	9501.000000	9501.000000	9546.500000	9586.000000	9615.000000	85.000000	42.666279	1.820411e+03
dia9 kilometrosOdometro	9657.008721	9614.000000	9614.000000	9639.000000	9699.000000	9755.000000	85.000000	47.001612	2.209152e+03
ficti kilometrosOdometro	9616.471853	9501.000000	9578.673707	9616.430665	9649.417630	9755.000000	70.743924	59.031149	3.484676e+03
dia8 revolucionesMotor	749.511538	0.000000	0.000000	0.000000	1901.500000	2472.000000	1901.500000	914.247702	8.358489e+05
dia9 revolucionesMotor	705.072674	0.000000	0.000000	0.000000	1920.500000	2480.000000	1920.500000	1024.033623	1.048645e+06
ficti revolucionesMotor	985.986231	0.000000	115.413077	906.399231	1772.861807	2474.662951	1657.448730	834.389048	6.962051e+05
dia8 temperaturaBaterias	31.419231	29.000000	30.000000	32.000000	33.000000	2.000000	1.141197	1.302331e+00	
dia9 temperaturaBaterias	32.700581	28.000000	30.000000	33.000000	36.000000	5.000000	2.712760	7.359067e+00	
ficti temperaturaBaterias	32.310230	28.000000	31.547732	32.270911	33.483696	36.000000	1.935964	1.842672	3.395438e+00
dia8 kilometrosAcumTrayecto	46.342308	0.000000	0.000000	45.500000	85.000000	114.000000	85.000000	42.666279	1.820411e+03
dia9 kilometrosAcumTrayecto	43.008721	0.000000	0.000000	25.000000	85.000000	141.000000	85.000000	47.001612	2.209152e+03
ficti kilometrosAcumTrayecto	55.802796	0.000000	16.565479	45.919081	92.457438	141.000000	75.891959	42.037458	1.767148e+03
dia8 kilometrosEntreMuestra	0.442308	0.000000	0.000000	0.000000	1.000000	3.000000	1.000000	0.602872	3.634541e-01
dia9 kilometrosEntreMuestra	0.409884	0.000000	0.000000	0.000000	1.000000	5.000000	1.000000	0.694011	4.816513e-01
ficti kilometrosEntreMuestra	0.905875	0.000000	0.056421	0.957365	1.352813	2.879651	1.296391	0.860047	7.396810e-01
dia8 altitud	148.111538	54.000000	106.000000	174.000000	187.000000	187.000000	81.000000	45.819512	2.099428e+03
dia9 altitud	158.415698	58.000000	139.000000	187.000000	188.000000	48.000000	45.833758	2.100733e+03	
ficti altitud	148.122862	60.000000	120.903473	160.033170	182.287137	187.879877	61.383665	38.463058	1.479407e+03
dia8 nivelRestanteEnergia	84.439038	61.870000	72.770000	85.470000	98.960000	99.360000	26.190000	13.663799	1.866994e+02
dia9 nivelRestanteEnergia	48.291395	19.090000	35.855000	53.680000	61.270000	61.870000	25.415000	14.251625	2.031088e+02
ficti nivelRestanteEnergia	63.179699	43.436390	57.810142	63.964306	68.751110	77.355515	10.940968	6.884821	4.740076e+01

TABLA 7.2
INFORMACIÓN POR BUS Y POR TRAYECTO

Bus	Trayecto	Velocidad_media	Velocidad_std	Temperatura_media	Temperatura_std	Rendimiento_BK	Kilo_Totales	Tiempo
2	3	65.6	22.52	30.14	3.87	2.57	125	7
2	2	66.17	37.12	30.49	4.47	2.58	130	6
3	3	63.15	21.5	30.87	1.71	3.3	118	5
2	2	61.89	22.99	30.71	4.3	2.74	114	5
1	3	78.6	35.47	31.0	4.01	3.0	137	7
2	2	67.19	36.91	33.59	2.99	2.6	148	7
3	2	79.18	24.16	28.82	3.18	3.16	133	7
1	2	67.27	31.91	33.67	3.03	3.19	132	5
1	2	71.43	39.32	28.28	2.29	2.73	131	5
1	2	72.18	30.41	28.72	4.92	2.52	142	5

Referencias

- [1] J. VanderPlas, *Python data science handbook: Essential tools for working with data.* O'Reilly Media, Inc., 2016. [13](#)
- [2] J. D. Miller, *Hands-On Machine Learning with IBM Watson: Leverage IBM Watson to implement machine learning techniques and algorithms using Python.* Packt Publishing Ltd, 2019. [18](#)
- [3] D. S. Moore, *Estadística aplicada básica.* Antoni Bosch editor, 2005. [18, 19, 67, 68, 68, 70, 71, 73, 82](#)
- [4] W. W. Hines, D. C. Montgomery, and G. tr Nagore, *Probabilidad y Estadística para ingeniería y administración.* Compañía Editorial Continental, S.A, 1994. [18, 19, 67, 68, 69, 70, 71, 73, 82](#)
- [5] B. S. Everitt and A. Skrondal, *The Cambridge dictionary of statistics.* Cambridge University Press, New York, 2010. [18, 19, 67, 68, 69, 70, 71, 73, 82](#)
- [6] W. McKinney, *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython.* O'Reilly Media, Inc.", 2012. [18, 29, 31](#)
- [7] S. J. Russell, *Artificial intelligence a modern approach.* Pearson Education, Inc., 2010. [19, 20](#)
- [8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning.* MIT press, 2016. [19, 20, 21, 23](#)
- [9] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019. [22](#)
- [10] A. Inc, "What is an API?" https://aws.amazon.com/what-is/api/?nc1=h_ls, (accessed: 30.08.2022). [23](#)

- [11] J. Kazil and K. Jarmul, *Data wrangling with python: tips and tools to make your life easier*. O'Reilly Media, Inc.", 2016. [28](#), [29](#), [31](#)
- [12] C. Severance, *Python for Everybody Exploring Data Using Python 3*. Open Textbook Library, 2016. [30](#), [31](#)
- [13] S. Wexler, J. Shaffer, and A. Cotgreave, *The big book of dashboards: visualizing your data using real-world business scenarios*. John Wiley & Sons, 2017. [76](#), [77](#), [78](#), [144](#)
- [14] I. Milovanovic, D. Foures, and G. Vettigli, *Python Data Visualization Cookbook*. Packt Publishing Ltd, 2015. [76](#), [77](#), [78](#)
- [15] P. Simon, *Too big to ignore: the business case for big data*. John Wiley & Sons, 2013, vol. 72. [111](#)
- [16] I. Ashrapov, "Tabular gans for uneven distribution," *arXiv preprint arXiv:2010.00638*, 2020. [140](#)
- [17] B. Gnedenko, "Course in the theory of probability: Tutorial," 1998. [141](#)
- [18] B. Chambers and M. Zaharia, *Spark: The definitive guide: Big data processing made simple*. O'Reilly Media, Inc., 2018. [118](#)
- [19] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?" in *2016 international conference on digital image computing: techniques and applications (DICTA)*. IEEE, 2016, pp. 1–6.

Lista de figuras

4.1. Ratio de nulos por variable	65
4.2. Gráfica uni-variada número de reseñas	78
4.3. Gráfica uni-variada número de reseñas por mes	78
4.4. Gráfica uni-variada años de construcción	79
4.5. Gráfica uni-variada precio	79
4.6. Gráfica uni-variada cuota de servicio	79
4.7. Gráfica uni-variada mínimo número de noches	79
4.8. Gráfica multivariada	85
4.9. Histograma Latitud	86
4.10. Gráfica PCA 2d	91
4.11. Gráfica PCA 3d	91
5.1. Posición vs Tiempo día ocho	102
5.2. Posición vs Tiempo día nueve	103
5.3. Nivel de Carga vs Tiempo Día Ocho	103
5.4. Nivel de Carga vs Tiempo Día Nueve	104
5.5. Potencia vs Tiempo Día Ocho	104
5.6. Potencia vs Tiempo Día Nueve	105
5.7. Altura vs Tiempo Día Ocho	106
5.8. Altura vs Tiempo Día Nueve	106
5.9. Nivel de Energía vs Distancia Acumulada	107

6.1. Histograma Latitud	124
6.2. Histograma Longitud	125
6.3. Histograma Nivel Restante de Energía	126
6.4. Histograma Temperatura Baterías	127
6.5. Histograma Altitud	127
6.6. Histograma Kilómetros Odómetro	128
6.7. Histograma Revoluciones del Motor	128
6.8. Histograma Kilómetros Acumulados por Trayecto	129
6.9. Histograma Kilómetros entre Muestras	129
6.10. Mapa de Calor para el Coeficiente de Correlación	132
6.11. Gráfica Correlación.....	133
6.12. Datos Obtenidos de PCA Día 8	134
6.13. Datos Obtenidos de PCA Día 9	135
7.1. Gráfico para la Ley de los Grandes Números	142
7.2. Algoritmo Para Determinar Métricas en Bases que no Entran en Memoria Volátil	142

Lista de tablas

3.1. CONTENIDO DEL ARCHIVO “edades.xlsx”	36
3.2. DATAFRAME CREADO	45
3.3. DATAFRAME CON COLUMNA EXTRA	47
3.4. DATAFRAME LUEGO DE PIVOTEADO DE TABLA	48
3.5. DATAFRAME LUEGO DEL MERGE	49
4.1. TIPOS POR COLUMNA	62
4.2. PARÁMETROS GENERALES DE LAS COLUMNAS NUMÉRICAS	63
4.3. ESTADÍSTICOS DE TENDENCIA CENTRAL	68
4.4. ESTADÍSTICOS DE DISPERSIÓN	71
4.5. ESTADÍSTICOS DE FORMA	74
6.1. TABLA DATOS INICIALES	112
6.2. TABLA DATOS ÚNICOS	113
6.3. TABLA DATOS ÚNICOS LUEGO DE LA ELIMINACIÓN	114
6.4. TABLA DATOS ÚNICOS LUEGO DE LA ELIMINACIÓN Y DEPURACIÓN	115
6.5. TABLA DATOS ÚNICOS LUEGO DE AGREGAR NUEVAS COLUMNAS	116
6.6. ESTADÍSTICOS DE TENDENCIA CENTRAL DÍA 8	117
6.7. ESTADÍSTICOS DE TENDENCIA CENTRAL DÍA 9	118
6.8. ESTADÍSTICOS DE DISPERSIÓN DÍA 8	119
6.9. ESTADÍSTICOS DE DISPERSIÓN DÍA 9	119

6.10. COEFICIENTE DE VARIACIÓN PARA LAS VARIABLES ANALIZADAS EL 8 Y 9 DE SEPTIEMBRE	121
6.11. TERCER Y CUARTO MOMENTO ESTADÍSTICO	122
6.12. MATRIZ DE COVARIANZA DÍA 8 PARTE I	130
6.13. MATRIZ DE COVARIANZA DÍA 8 PARTE II	131
6.14. MATRIZ DE COVARIANZA DÍA 9 PARTE I	131
6.15. MATRIZ DE COVARIANZA DÍA 9 PARTE II	131
7.1. COMPARACIÓN CON DATA AUGMENTATION	146
7.2. INFORMACIÓN POR BUS Y POR TRAYECTO	147

Índice alfabético

- Business intelligence*, 14
- Acondicionar, 49
- Análisis de datos, 9
- Análisis exploratorio de datos, 48
- Análisis iterativo, 43
- Archivos csv, 25
- Archivos json, 26
- Archivos sql, 26
- Archivos xml, 26
- Bases de datos masivas, 43
- Cadena de proceso, 10
- Calidad de los datos, 45
- Características claves, 56
- Cargar la base de datos, 49
- Ciencia de datos, 11
- Coherencia, 44
- Colaboración, 45
- Comma separated value, 25
- Comprensión general, 50
- Comprensión global, 59
- Comunicación, 45
- Comunidad de Python, 24
- Conjunto de datos, 52
- Dashboards, 9
- Data analyst, 14
- Data cleaning, 13
- Dataframe, 36, 37
- Datos depurados, 56
- Decisiones basadas en datos, 45
- Definición de criterios, 22
- Definir objetivos, 23
- Depuración de datos, 13
- Distribución, 50
- Documentar fuentes, 24
- Documentar métodos, 24
- EDA, 48
- Eficacia, 44
- Eliminación, 53
- Enfoque estructurado, 44
- Errores, 52, 53, 60, 61
- Escalabilidad, 45
- Excel, 25
- Exploración de datos, 46
- Extensible markup language, 26

Fiabilidad, 45, 53
Flujo de análisis, 46
Flujo de análisis de datos, 43
Formato JSON, 33
Fuentes de datos, 22
Garbage in, garbage out, 22
Gran escala, 25
Imputación, 53
Ingeniería de características, 46
Inteligencia empresarial, 14
JavaScript Object Notation, 33
JavaScript object notation, 26
jupyter notebooks, 48
Librerías de Python, 24
Limpieza, 55
Limpieza de datos, 13, 46, 47
Manipulación de archivos CSV, 27
Manipulación de datos XML, 31
Matplotlib, 69
Mejora continua, 43
Multilenguaje, 25
Módulo csv, 27
Módulo Pandas, 35
Normas de calidad, 24
Pandas, 29, 30, 32, 35
Pequeña escala, 25
Pipeline, 10
Pre-procesamiento, 53
Preprocesamiento, 46
Python, 24
Recopilacion de datos, 22
Recopilación e integración de datos, 46
Recopilación eficaz, 23
Resumen estadístico, 50
Scripts, 24
Seaborn, 69
Selección y validación de modelos, 46
Structured query language, 26
Subgráficas, 69
Tablas interactivas, 9
Tendencias centrales, 50
Tipos de archivos, 25
Toma de Decisiones, 45
Transformación de datos, 46
Transformación de los datos, 43
Transparencia, 45
Tratamiento de datos, 23
Trazabilidad, 45
Utilizar fuentes fiables, 23
Validar los datos, 23
Validez, 53
Valores separados por comas, 25

La Editorial de la Universidad Tecnológica de Pereira tiene como política la divulgación del saber científico, técnico y humanístico para fomentar la cultura escrita a través de libros y revistas científicas especializadas.

Las colecciones de este proyecto son:
Trabajos de Investigación, Ensayos,
Textos Académicos y Tesis Laureadas.

Este libro pertenece a la Colección
de Textos Académicos.

La actualidad está siendo marcada por una avalancha de datos y una evolución constante en el mundo empresarial. Todos los sectores económicos avanzan a pasos agigantados y el análisis de datos se ha convertido en una herramienta indispensable. Este libro está diseñado para ser una guía introductoria, práctica y accesible para aquellos en el sector empresarial, o curiosos del tema, que buscan adentrarse en el mundo del análisis de datos, proporcionando una comprensión fundamental y ejemplos sencillos de cómo los datos pueden ser utilizados para informar y mejorar las decisiones de negocios.

Como primer propósito del libro, se busca desmitificar el análisis de datos, presentándolo no como un campo exclusivo para expertos en TI o estadísticos, sino como una competencia accesible y valiosa para una amplia gama de profesionales. Como segundo propósito, el libro se enfoca en la aplicación práctica, ofreciendo ejemplos que muestran cómo el análisis de datos puede resolver problemas; se explora cómo recolectar, limpiar y manipular datos utilizando Python, un lenguaje de programación que se ha establecido firmemente como un estándar, debido a su simplicidad y potencia. Se introduce al lector en herramientas como Pandas y otras bibliotecas de Python, que simplifican enormemente el proceso de análisis de datos. No obstante, más allá de la mera técnica, este libro también se sumerge en la interpretación y el análisis crítico, habilidades clave para convertir los datos en perspectivas (insights) accionables.

Today's world is marked by a data deluge and a constant evolution in the business landscape. All economic sectors are advancing at a rapid pace, and data analysis has become an indispensable tool. This book is designed to be an introductory, practical, and accessible guide for those in the business sector, or simply curious about the topic, who seek to delve into the world of data analysis. It provides a fundamental understanding and simple examples of how data can be used to inform and improve business decisions.

The book's primary purpose is to demystify data analysis, presenting it not as a field exclusive to IT experts or statisticians, but as an accessible and valuable skill for a wide range of professionals. Secondly, the book focuses on practical application, offering examples that demonstrate how data analysis can solve problems. It explores how to collect, clean, and manipulate data using Python, a programming language that has firmly established itself as a standard due to its simplicity and power. The reader is introduced to tools like Pandas and other Python libraries that greatly simplify the data analysis process. However, beyond mere technique, this book also delves into interpretation and critical analysis, key skills for turning data into actionable insights.

Facultad de Ingenierías

Colección de Textos Académicos

eISBN :