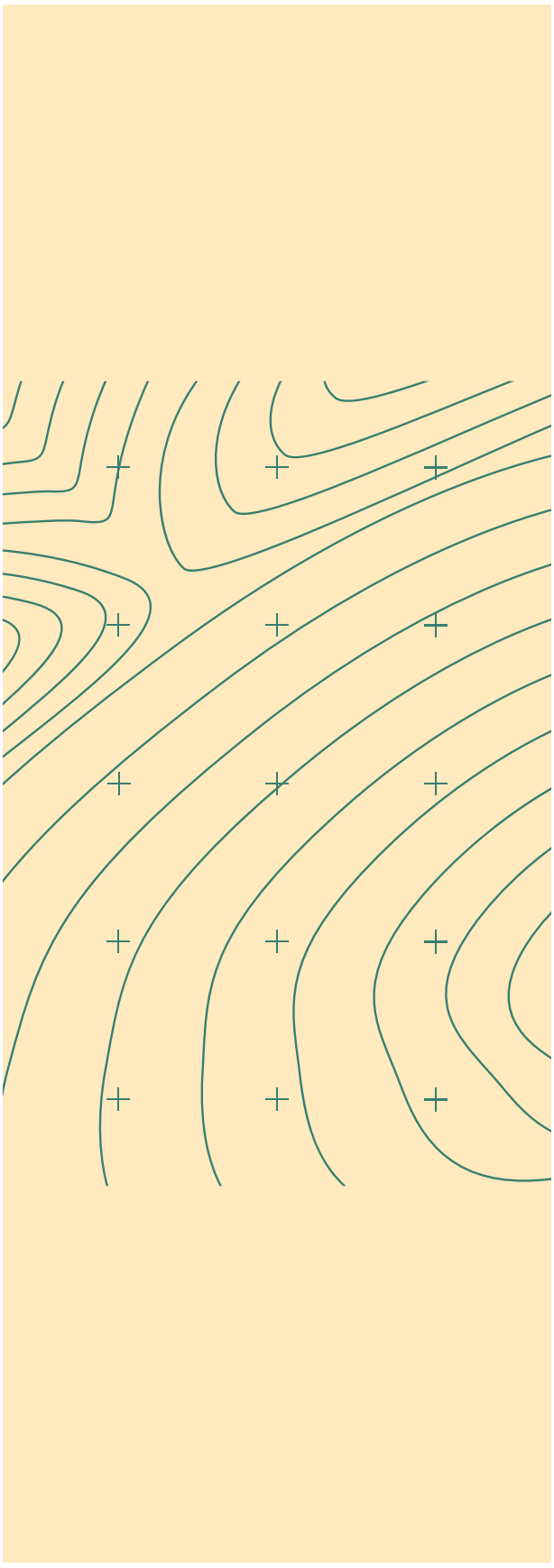
CAPÍTULO II

MINERÍA DE DATOS: UN EJERCICIO DE PRECISIÓN PARA EL PROYECTO CANALES

Por: Julia Castaño González y
Jaime Andrés Valencia



Introducción

El presente capítulo describe, específicamente, el proceso de minería de datos realizado para la primera parte metodológica del proyecto de investigación CANALES, que, al ser un proyecto adherido al Centro Analítico de Producciones Culturales CAPAZ, debía seguir su metodología.

Big data es una de las diversas técnicas desarrolladas para gestionar la avalancha de datos y las grandes cantidades de información que se generan a diario, propias de la era digital en la que vivimos. Sin embargo, es común encontrar que este concepto se trata como si fuera sinónimo de las técnicas análisis de datos y minería de datos, por lo que se hace necesario reconocer la diferencia, justamente, para dar cuenta del desarrollo metodológico del proyecto CANALES.

En síntesis, el *big data* se ocupa de una exploración general de todo tipo de datos que pueden ser procesados, a la vez y en grandes cantidades; mientras que la minería de datos y el análisis de datos implican una vista cercana y detallada sobre un tipo de dato. A partir de estas técnicas se descubren patrones y relaciones que no son tan evidentes y, tras su análisis, permiten sacar conclusiones y tomar decisiones (Data Visualization Project, 2023a; Data Visualization Project, 2023b; UNIR Revista, 2023).

Se expone aquí, en primera medida, cómo se resuelve el primer nivel metodológico

para CANALES, que indica la utilización de *big data*: "analizar datos y generar enfoques cuantitativos que permitan la toma de decisiones y la inteligencia colectiva de la investigación" (UCP, 2021, p. 21).

En este proceso, posterior a la juiciosa revisión de la información existente sobre conectividad a partir de los *datasets* encontrados, se realizan análisis preliminares de tipo exploratorio de datos EDA *(Exploratory Data Analysis)*, se hacen las correspondientes visualizaciones, se grafican y, finalmente, se crea un *dashboard* o tablero de datos para ver el comportamiento y las características de cada municipio. Esto con cada uno de los diecisiete municipios involucrados en el proyecto CANALES.

Big data, minería de datos o análisis de datos: decisión para CANALES

En esta ocasión, por tratarse de información limitada, para el proyecto CANALES no se logró realizar *big data*, solo se llegó a la minería de datos. Y es que para que un procesamiento de datos sea considerado *big data*, el conjunto de datos tiene que surtir unos condicionantes conocidos como las 3V: volumen, variedad y velocidad (UNIR Revista, 2020).

Se puede decir que la definición del término *big data* varía dependiendo de cada autor. Para Dan Kusnetzky (2010), "aplica a la información que no puede ser procesada o analizada mediante procesos tradicionales". Por su parte, para UNIR Revista (2020), "se refiere a volúmenes tan grandes de datos que no pueden ser accedidos, almacenarse, y procesarse en una única máquina, por lo que requieren de sistemas específicos para ello".

Entonces, por no tratarse de una cantidad masiva o incluso desproporcionada de datos, recolectados a través del tiempo, con cierta complejidad de análisis, la información encontrada para saldar la primera parte de la metodología de CANALES no equivale a *big data*. Antes bien, correspondería a los conceptos de análisis de datos o, más exactamente, al de minería de datos, teniendo en cuenta que, según Amazon Web Services (2023a) el análisis de datos:

> Convierte datos sin procesar en información práctica. Incluye una serie de herramientas, tecnologías y procesos para encontrar tendencias y resolver problemas mediante datos. Los análisis de datos pueden dar forma a procesos empresariales, mejorar la toma de decisiones e impulsar el crecimiento empresarial.

Por otra parte, la minería de datos:

> Es una técnica asistida por computadora que se utiliza en los análisis para procesar y explorar grandes conjuntos de datos. Gracias a las herramientas y métodos de minería de datos, las organizaciones pueden descubrir patrones y relaciones ocultas en

> sus datos. La minería de datos transforma datos en bruto en conocimiento práctico. Las compañías utilizan dicho conocimiento para resolver problemas, analizar las consecuencias en el futuro de decisiones empresariales y aumentar sus márgenes de beneficio. (Amazon Web Services, 2023b)

Así pues, el análisis de datos y la minería de datos comparten un objetivo común: transformar datos sin procesar en información valiosa y práctica. Ambos utilizan procesos, herramientas y tecnologías para explorar grandes conjuntos de datos, en busca de patrones, tendencias y relaciones ocultas. De igual manera, ambos enfoques tienen el potencial de mejorar la toma de decisiones en diferentes contextos y permiten resolver problemas de manera más eficiente.

Dentro de las diferencias clave entre estos dos términos, está que, mientras el análisis de datos se centra en encontrar tendencias y resolver problemas mediante datos, y puede implicar una gama más amplia de enfoques y técnicas, la minería de datos se enfoca específicamente en el procesamiento y la exploración de grandes conjuntos de datos para descubrir conocimiento práctico. La minería de datos es una técnica asistida por computadora, lo que implica el uso de algoritmos y métodos específicos para extraer información valiosa de los datos.

Por todo lo anterior, y a pesar de que la instrucción del nivel 1 de la metodología de CANALES, correspondiente a la "obtención de los datos", debía recolectar, a partir de *big data*: "a. Obtención de la información existente sobre conectividad (…) b. Identificación de datos sobre el consumo cultural de contenidos audiovisuales y digitales en los territorios (…) c. Recopilación de relatos de los prosumidores (…)" (UCP, 2021, p. 22), la información arrojada para este proyecto obedece más a la minería de datos.

Desarrollo de la metodología

Independientemente de no haber sido clasificado el proceso para CANALES como *big data*, se siguieron las instrucciones de revisión de "información existente sobre conectividad a través de técnicas de análisis de la información multivariado de tipo sociodemográfico, técnico y tecnológico" (UCP, 2021, p. 22). Esto se hizo a partir de *datasets* o recursos estadísticos extraídos de entidades nacionales, regionales o locales.

Por ello, se auscultaron múltiples datos históricos relacionados con las variables del estudio de CANALES: acceso, uso y consumo. Como punto de partida, dicha revisión se inició en *datasets*, entendidos como los conjuntos de datos a analizar, alojados en la Plataforma Nacional de Datos Abiertos de Colombia (Gov.co. Datos Abiertos, 2023). Entre ellos se encuentran Internet Fijo Accesos por tecnología y segmento, Internet Fijo Penetración por Municipio e Internet Fijo Penetración Departamentos (Gov.co. Datos Abiertos, 2023). Este último fue descartado después de los

análisis preliminares de tipo exploratorio de datos, en adelante EDA por su denominación en inglés *(Exploratory Data Analysis)*, por considerar que su aporte no era de "utilidad", como los otros dos *datasets*.

Una vez identificados los *datasets*, se realizó un breve proceso EDA para caracterizar los datos y hallar "anómalos" o información interesante para el estudio, y así poder reconocer qué variables se encontraban presentes en estos *datasets* y cómo estaban categorizadas. Esto dio como resultado una tabla guía para tener una referencia de cómo representar los datos.

En este orden de ideas, la tabla 1 muestra qué variables van a ordenar la información, cuáles van a permitir la agrupación de los datos y cuáles son los datos por representar; en este caso, el indicador correspondiente. Con esta información definida, se pasó a rastrear los datos anómalos (Data Viz Catalogue, 2023), ya que estos hablan de eventos específicos dentro de los municipios.

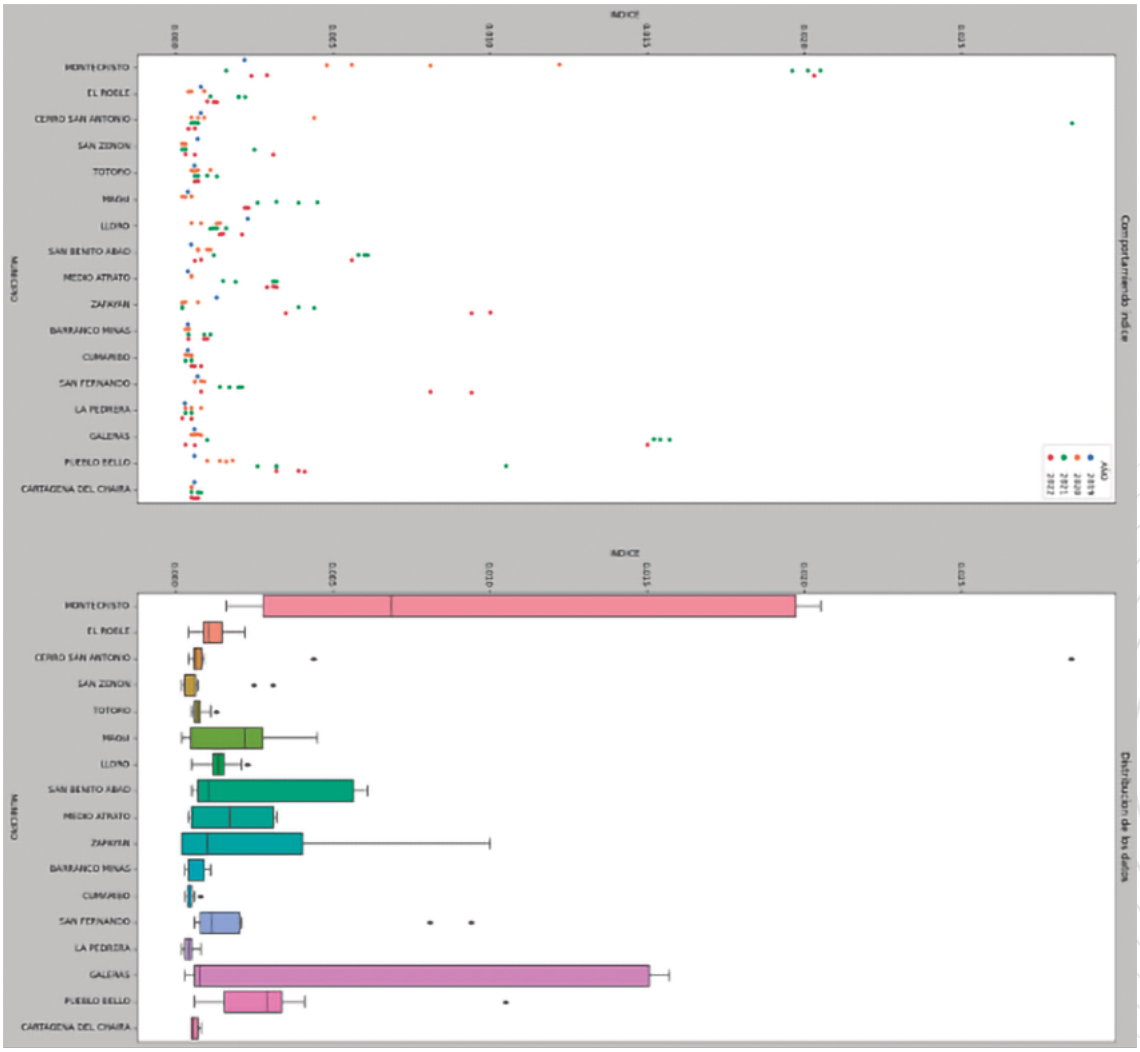
Los datos anómalos pueden ser ruido en otros contextos. Sin embargo, en este proyecto son importantes, ya que señalan cambios por fuera de lo usual en los indicadores. Por ejemplo, en la figura 1 se muestra como dato anómalo que en Cerro San Antonio, en el año 2021, se presentó un índice de penetración de internet en comparación con los otros años revisados.

Tabla 1
Caracterización de datos

Variable	Tipo
DATE	Cualitativa-Ordinal
AÑO	Cuantitativa-Discreta
TRIMESTRE	Cuantitativa-Discreta
DEPARTAMENTO	Cualitativa-Nominal
PROVEEDOR	Cualitativa-Nominal
MUNICIPIO	Cualitativa-Nominal
SEGMENTO	Cualitativa-Nominal
TECNOLOGIA	Cualitativa-Nominal
No accesos fijos (tecnologia, proveedor, municipio,departamento,segmento)	Cuantitativa-Continua
Poblacion Dane (Municipio, Departamento)	Cuantitativa-Continua
VELOCIDAD DE SUBIDA	Cuantitativa-Continua
VELOCIDAD DE BAJADA	Cuantitativa-Continua
INDICE DE PENETRACION (Municipio Departamento)	Cuantitativa-Continua

Figura 1

Distribución de índice de penetración a internet por municipio



En la figura 1 se pueden ver dos tipos de visualización de datos (en adelante VIZ). A la izquierda se denota un *jitter plot* o diagrama de fluctuación, también denominado gráfica de tira fluctuante o gráfica de valor individual fluctuante, que se utiliza para comparar o ver distribuciones de datos (Data Visualization Project, 2023a).

En concreto, la figura 1 muestra, de manera general, cómo se comportó el índice en cada municipio. Por su parte, a la derecha se encuentra un *box plot* o gráfico de cajas y bigotes. Se trata de:

> Una forma conveniente de representar gráficamente grupos de datos numéricos a través de sus cuartiles. Un diagrama de caja muestra la mediana, los cuartiles superior/inferior y el máximo/mínimo. Los valores atípicos se pueden trazar como puntos individuales. Los

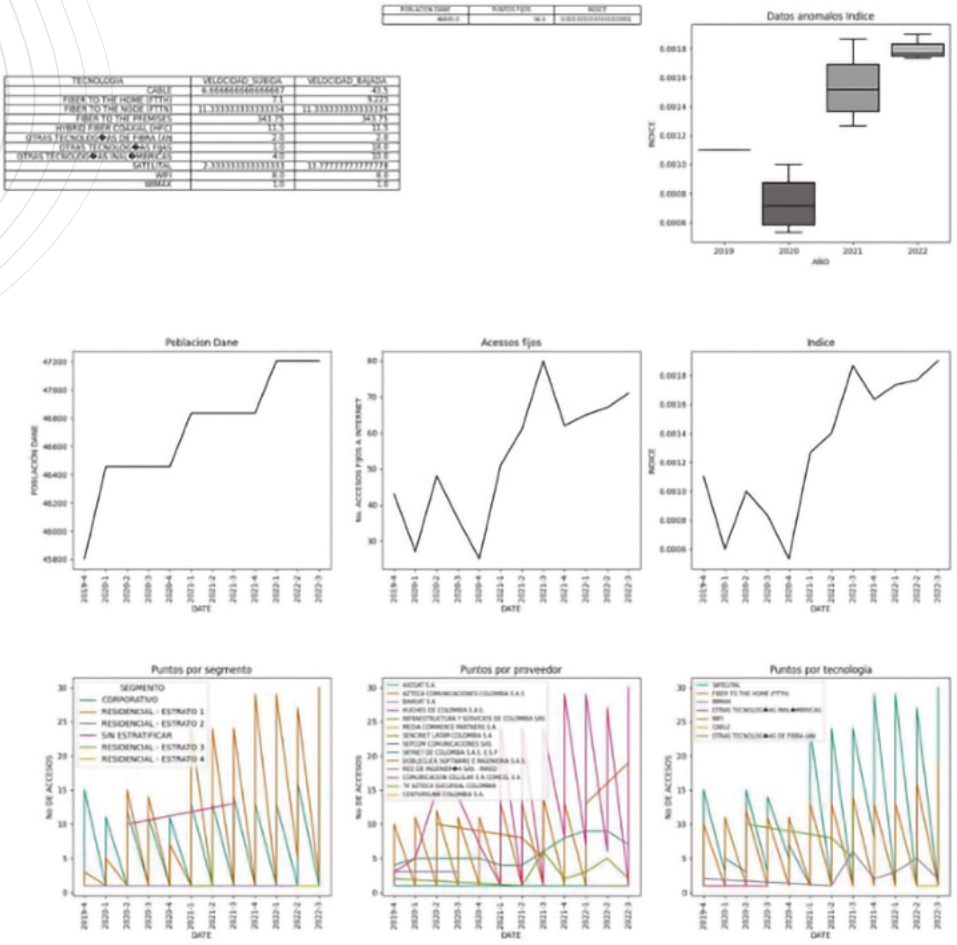
> espacios entre las diferentes partes del cuadro indican el grado de dispersión (propagación) y sesgo en los datos, y muestran valores atípicos. Los diagramas de caja se pueden dibujar horizontal o verticalmente. (Data Visualization Project, 2023b)

En este caso, se puede ver cómo están distribuidos los datos en cada municipio y cuáles son los datos anómalos. Por tanto, se interpreta de la siguiente manera: cuanto más grande la caja, más dispersos y variados los datos; cuanto más pequeña, más juntos, y se considera un dato anómalo si está por fuera de la caja.

Una vez identificados los datos se construyó un *dashboard* o tablero de datos (véase figura 2), para ver, en general, el comportamiento y las características de cada municipio.

Figura 2

Ejemplo de *dashboard* región Pacífica



De la misma forma se diseñaron *dashboards* para el Caribe y la Amazonía, además de uno por municipio. Es decir, diecisiete *dashboards* que posteriormente fueron interpretados, para lo que, además, fue necesaria la revisión de los siguientes documentos: Recomendaciones sobre la velocidad de conexión a internet (Netflix, 2023); Requisitos del sistema y dispositivos compatibles con YouTube (Google, 2023a); Problemas con reproducciones en *streaming* en directo en Prime Video (Prime Video, 2023) y Requisitos del *hardware* de Meet (Google, 2023b).

Dichos documentos fueron importantes, ya que especifican las características técnicas —dadas en velocidad de subida y de bajada de datos de internet— requeridas para usar aplicaciones como Netflix, YouTube y Prime Video, consideradas en los instrumentos del proyecto CANALES. Al momento de comparar los resultados de la encuesta de CANALES y la información presente en el *dashboard* contra estos documentos, se llega a preguntas como las siguientes: ¿qué características tienen los municipios?, ¿dichas características les permiten a los pobladores acceder a las aplicaciones mencionadas?

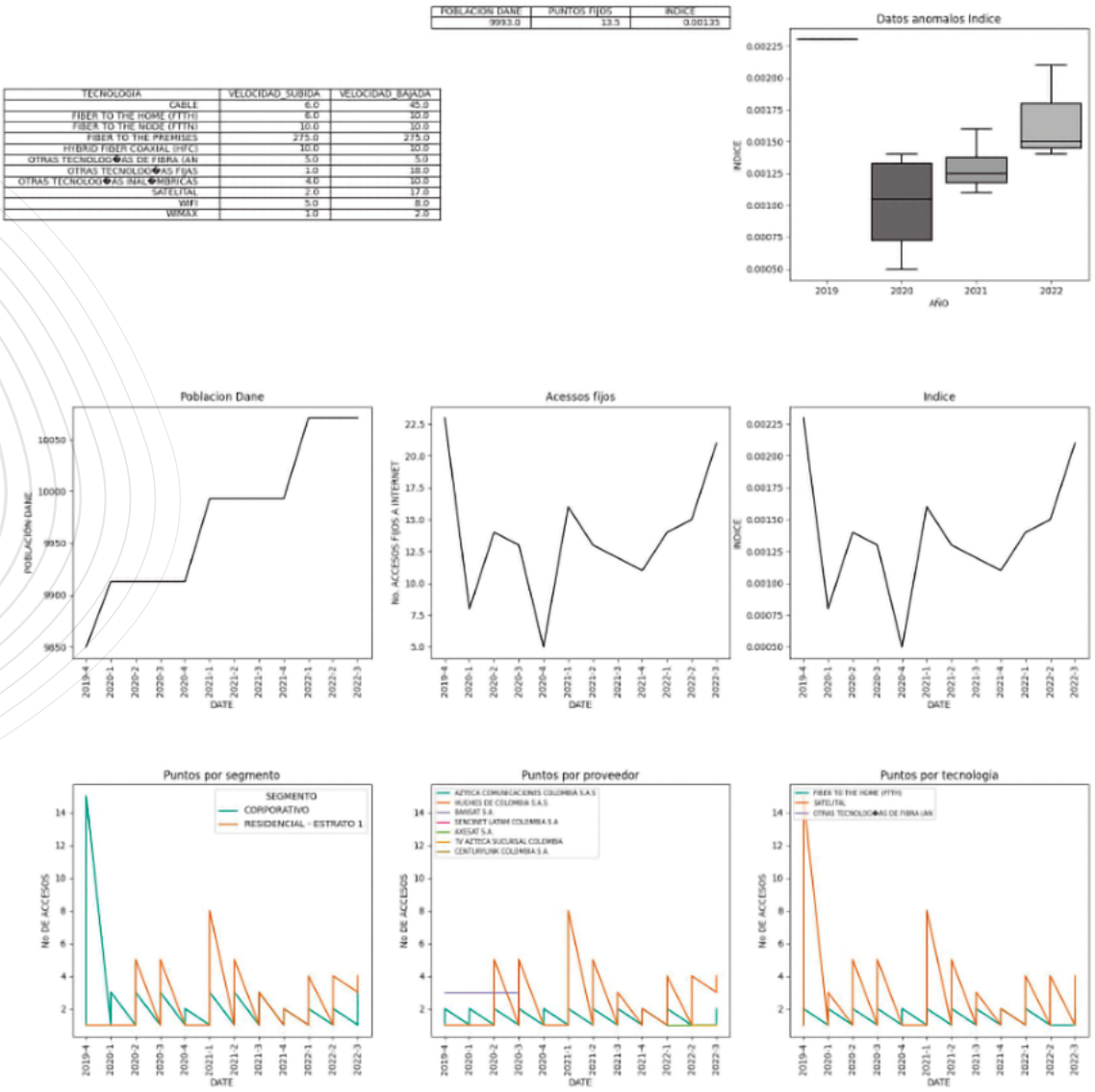
Por ejemplo, veamos el caso de la pregunta n.º 29 en los resultados de la región Pacífica: Si tuviera que calificar la importancia del internet a la hora de consumir los siguientes contenidos audiovisuales, ¿qué valoración le daría? En la opción: Producciones de plataformas como Netflix, Prime u otros, los porcentajes de respuesta fueron: Alta: 34 %, Media: 18 %, Baja: 47 %, NR: 2 %. Por su parte, en la pregunta n.º 37. Califique el uso que usted le da a las siguientes plataformas de televisión por *streaming* de 0 a 5. En la opción Netflix se obtuvieron los siguientes resultados: 31 % (5), 11 % (4), 6 % (3), 3 % (2), 4 % (1), 45 % (0), 1 % NR.

Tras el análisis de estos datos, se puede concluir que en la región Pacífica las tecnologías más predominantes (figura 2) son conexión satelital y conexión por fibra. Al momento de comparar sus características técnicas con lo recomendado en el documento de Netflix (2023), estas conexiones sí permiten acceder a dicha plataforma en una resolución de 720p, lo que significa que las personas suelen acceder de manera adecuada a dicha plataforma. Esto coincide con los resultados de la encuesta, que señala que las personas tienden a darle un uso alto a este servicio, como se puede observar en la pregunta n.º 37, en donde el 31 % de los encuestados da una calificación de 5 al uso de Netflix, en una escala de 0 a 5.

Adicional a esto, en los *dashboards* individuales por municipio, se logró un análisis como el expuesto en la figura 3.

Figura 3

***Dashboard* Lloró, Chocó - región Pacífica**



POBLACION_DANE	PUNTOS_FIJOS	INDICE
9993.0	13.5	0.00135

TECNOLOGIA	VELOCIDAD_SUBIDA	VELOCIDAD_BAJADA
CABLE	6.0	45.0
FIBER TO THE HOME (FTTH)	6.0	10.0
FIBER TO THE NODE (FTTN)	10.0	10.0
FIBER TO THE PREMISES	275.0	275.0
HYBRID FIBER COAXIAL (HFC)	10.0	10.0
OTRAS TECNOLOG◆AS DE FIBRA (AN	5.0	5.0
OTRAS TECNOLOG◆AS FIJAS	1.0	18.0
OTRAS TECNOLOG◆AS INAL◆MBRICAS	4.0	10.0
SATELITAL	2.0	17.0
WIFI	5.0	8.0
WIMAX	1.0	2.0

En general, la figura 3 muestra una caída muy marcada en el índice y el número de accesos fijos entre 2019-4 y 2020-1 (véase VIZ número de accesos fijos y índice), lo cual se debe a una caída en el segmento poblacional corporativo con conexión satelital (véase VIZ: Puntos por segmento / Puntos por tecnología [Gov.co. Datos Abiertos, 2023b]). Se presenta también caída entre el 2020-2 y el 2020-4 en el índice y el número de accesos fijos (véase VIZ número de accesos fijos y índice), debido a la desaparición del proveedor Bansat S.A. (véase VIZ: Puntos por proveedor [Gov.co. Datos Abiertos, 2023b]).

Conclusiones

Big data se refiere al proceso de recolectar, almacenar y analizar grandes volúmenes de datos que superan la capacidad de las herramientas tradicionales de gestión de datos. Los desafíos principales en *big data* son la variedad, velocidad y volumen de la información que se maneja.

El análisis de datos es el proceso de examinar, limpiar, transformar e interpretar datos con el objetivo de descubrir patrones, tendencias, correlaciones, y así obtener conclusiones útiles para la toma de decisiones. Por su parte, la minería de datos se trata de una rama específica del análisis de datos que se centra en descubrir patrones y relaciones no evidentes en grandes conjuntos de datos. Esta utiliza técnicas estadísticas y de aprendizaje automático para extraer conocimiento.

La primera revisión de datos requerida para el proyecto CANALES no se ajusta a la técnica de *big data*, por lo que el procesamiento de datos obedeció más al concepto de minería de datos. Sin embargo, el procesamiento de la encuesta, en sus tres categorías (acceso, uso y consumo) sí se adapta a las condiciones para que sea *big data*, por la gran cantidad de cruces posibles, pero ese es un tema de otro capítulo.

Si bien el proyecto CANALES no se ajustó al uso de *big data,* el uso de técnicas y herramientas cercanas a esta área permitió procesar los datos existentes y generar conclusiones e información en torno a dichos datos.

Referencias

Amazon Web Services. (2023a). *¿Qué es la minería de datos?* Amazon Web Services. https://aws.amazon.com/es/what-is/data-mining/

Amazon Web Services. (2023b). *¿Qué es la analítica de datos?* Amazon Web Services. https://aws.amazon.com/es/what-is/data-analytics/

Data Viz Catalogue. (2023). *Box and Whisker Plots - Learn About this Chart and Its Tools*. Data Viz Catalogue. https://datavizcatalogue.com/methods/box_plot.html

Data Visualization Project. (2023a). *Jitter Plot.* Data Visualization Project. https://datavizproject.com/data-type/jitter-plot/

Data Visualization Project. (2023b). *Box plot*. Data Visualization Project. https://datavizproject.com/data-type/box-plot/

Gov.co. Datos Abiertos. (2023). Datos abiertos. MinTIC. https://www.datos.gov.co/

Gov.co. Datos Abiertos. (2023). *Internet Fijo Accesos por tecnología y segmento*. MinTIC. https://www.datos.gov.co/Ciencia-Tecnolog-a-e-Innovaci-n/Internet-Fijo-Accesos-por-tecnolog-a-y-segmento/n48w-gutb

Gov.co. Datos Abiertos. (2023). *Internet Fijo Penetración por Municipio*. MinTIC. https://www.datos.gov.co/Ciencia-Tecnolog-a-e-Innovaci-n/Internet-Fijo-Penetraci-n-Municipio/fut2-keu8

Gov.co. Datos Abiertos. (2023). *Internet Fijo Penetración Departamentos*. MinTIC. https://www.datos.gov.co/Ciencia-Tecnolog-a-e-Innovaci-n/Internet-Fijo-Penetraci-n-Departamentos/4py7-br84

Google. (2023a). *Requisitos del sistema y dispositivos compatibles con YouTube*. Ayuda de YouTube. https://support.google.com/youtube/answer/78358?hl=es-419

Google. (2023b). *Requisitos del hardware de Meet*. Ayuda de Administrador de Google Workspace. https://support.google.com/a/answer/4541234?hl=es#zippy=%2Crequisitos-generales-de-la-red%2Crequisitos-de-ancho-de-banda

Kusnetzky, D. (2010, 15 de febrero). *What is "Big Data"?* ZDNet. https://www.zdnet.com/article/what-is-big-data/

Netflix. (2023). *Recomendaciones sobre la velocidad de conexión a internet*. Netflix. Centro de ayuda. https://help.netflix.com/es/node/306#:~:text=Para%20ver%20

Prime Video. (2023). Problemas con reproducciones en streaming en directo en Prime Video. Prime Video Ayuda. https://www.primevideo.com/help/ref=atv_hp_nd_cnt?language=es_ES&nodeId=GP57SKQ7CB5DRS6F#:~:text=Prime%20Video%20recomienda%20una%20velocidad,de%20ancho%20de%20banda%20disponible.

UCP - Universidad Católica de Pereira. (2021). Formulario proyecto: Análisis del consumo audiovisual, acceso reciente a internet e interpretación pedagógica-narrativa de la población colombiana a través del Centro Analítico de Producciones Culturales - CAPAZ (fase II). Convocatoria n.o 908 de 2021: Ministerio de Ciencia, Tecnología e Innovación y Comisión de Regulación de Comunicaciones (CRC).

UNIR Revista. (2020). Las tres V del big data: Todo un reto por su volumen, variedad y velocidad. UNIR - Universidad Internacional de La Rioja. https://www.unir.net/ingenieria/revista/3-v-big-data/#:~:text=Las%20tres%20V%20del%20Big%20Data%20se%20refiere%20a%20los,adem%C3%A1s%20de%20sus%20principales%20retos

UNIR Revista. (2023). Diferencias entre minería de datos y big data. UNIR - Universidad Internacional de La Rioja. https://www.unir.net/ingenieria/revista/diferencias-mineria-datos-big-data/