Universidad de los Andes
Colombia

MIAD

ENTREGA DE LA GUÍA 4

Semana 6

Realizado por

Marlon Zambrano
Alberto Alejandro Molina
Ximena Gómez F.
Christian Beraun

Fecha

Noviembre 17 de 2024

Asignatura

Gerencia de Proyectos

Parte I: Característica de calidad de los datos

1.1. Contexto del Análisis y Evaluación de Calidad

En el proceso de análisis y perfilamiento de individuos, es fundamental garantizar que los datos sean accesibles, entendibles, relevantes y de calidad suficiente para generar valor en el problema de negocio. Esto implica trabajar con información proveniente de diversas fuentes, que pueden incluir bases de datos organizacionales, servidores virtuales o físicos, archivos distribuidos, y datos abiertos accesibles mediante API, entre otros.

Cuando los datos provienen de múltiples fuentes, es esencial llevarlos a un nivel de granularidad común, ajustándolos para asegurar su consistencia y coherencia con los objetivos del análisis. Este paso inicial facilita no solo la exploración de los datos, sino también su capacidad para responder preguntas de negocio.

En este contexto, se ha recibido información del CNC correspondiente a diversas encuestas realizadas entre 2012 y 2023 sobre apropiación digital, además de fuentes externas como resultados de Prueba Saber 11 (2010 a 2024) y la Encuesta Nacional de Hogares del DANE (2024). Sin embargo, considerando la granularidad y pertinencia de las fuentes así como datos que describan comportamientos y características propias de los individuos, se decidió trabajar únicamente con los resultados de la encuesta de campo de apropiación digital 2023 (Archivo XLSX) para esta primera etapa del análisis. Este enfoque permite focalizar el esfuerzo en segmentar individuos según sus características y niveles de apropiación digital, como el uso, frecuencia y accesibilidad.

Una vez seleccionada la fuente de datos principal, se procedió a evaluar su calidad en función de las dimensiones clave: totalidad, consistencia, claridad, formato y concordancia con el problema de negocio.

A continuación, se presentan los datos iniciales a ser procesados

Tabla 1.
Dataframe del dataset inicial

	REGISTRO	N_ENCUESTA	REGIONAL	PB1	SECTOR	REGION	MUNICIPIO_DEC	MUNICIPIO	PDET	PERSONAS	...	B26_2	B26_3	B26_6	B26_11	B26
0	1877	764678	5	1	526	5	8	8	2	2	...	3	NaN	NaN	2	
1	1885	900989	5	1	5002	5	13	13	2	2	...	2	2.0	4.0	5	
2	1889	968646	5	1	5002	5	13	13	2	2	...	2	1.0	4.0	3	
3	1898	180785	5	1	5006	5	13	13	2	6	...	3	NaN	NaN	8	
4	1901	734690	5	1	5006	5	13	13	2	3	...	1	2.0	6.0	2	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4174	8703	294431	1	1	1	4	32	32	2	5	...	3	NaN	NaN	3	
4175	8723	516103	1	1	1	4	32	32	2	4	...	2	2.0	20.0	2	
4176	8726	405383	1	1	1	4	32	32	2	5	...	1	1.0	7.0	4	
4177	8728	606911	1	1	2	4	32	32	2	3	...	1	2.0	17.0	4	
4178	8730	440739	1	1	2	4	32	32	2	8	...	1	2.0	2.0	1	

4179 rows × 296 columns

Elaborado por el equipo consultor, 2024.

➢ Nivel de granularidad

El *DataFrame* tiene una granularidad a nivel de individuo dentro del hogar, con 4,179 filas representando observaciones únicas que combinan datos individuales y del hogar. Las columnas incluyen identificadores (e.g., REGISTRO, N_ENCUESTA), variables geográficas (e.g., REGION, MUNICIPIO), y características demográficas y de comportamiento.

1.2. Características de Calidad de los Datos

Totalidad

En esta dimensión se procede a evaluar si la base de datos cuenta con toda la información necesaria y si los campos están completos. También se analiza la presencia de datos faltantes y su impacto en el análisis.

- La base de datos contiene información de 4,179 hogares, abarcando un total de 1,173 columnas que incluyen respuestas individuales y datos asociados.
- Se identificaron valores faltantes en aproximadamente el 80% de las columnas, en especial en preguntas relacionadas con hábitos digitales avanzados.

Consistencia

En esta dimensión analizamos si los datos son coherentes entre sí, identificando duplicados, contradicciones o errores lógicos en los valores registrados.

- Se detectaron inconsistencias en opciones de respuesta similares (columnas)que se encuentran duplicadas con ligeras variaciones de formato, como la representación de 'MUNICIPIO_DEC', 'MUNICIPIO', 'GENERO', 'GENERO_SEL' y 'B3_SEXO_1'.
- Será necesario realizar un proceso de estandarización y validación lógica para garantizar la coherencia de los datos.

Claridad

En esta dimensión se revisa si las etiquetas, nombres y categorías de los datos son comprensibles, y si la información es fácil de interpretar para generar resultados significativos.

- La documentación de la encuesta (PDF) proporciona descripciones detalladas de las variables, aunque algunas etiquetas en el dataset no coinciden completamente con los nombres utilizados en el informe.
- La gran cantidad de columnas puede dificultar la interpretación inicial, especialmente porque no todas las variables tienen una nomenclatura clara.
- Se precisa que no hubo un diccionario de datos que permita entender de forma más eficiente cada uno de los datos asociados a la encuesta.

Formato

En esta dimensión se examina si los datos están estructurados correctamente para ser utilizados en herramientas de análisis, incluyendo formatos uniformes y conversiones necesarias.

- Los datos están en formato Excel, lo que facilita su manipulación inicial. Sin embargo, los valores de respuesta contienen problemas de homogeneidad, como uso de separadores decimales inconsistentes y valores categóricos representados con diferentes códigos.
- Además, algunos datos numéricos aparecen como texto, requiriendo una conversión adecuada.
- Se realizará un ajuste del formato para asegurar que todas las variables sigan un estándar uniforme.

Todo el análisis realizado en las dimensiones planteadas, junto con el procesamiento del dataframe, se detalla en el Anexo 1: “Notebook”. Este documento contiene no solo el análisis efectuado, sino también las acciones implementadas para gestionar la gran cantidad de datos, abordando aspectos clave como la variabilidad, completitud y claridad tomando siempre en cuenta que las variables se asocien al problema de negocio.

1.3. Concordancia con el Problema de Negocio

En esta dimensión se valida si las variables y datos disponibles son relevantes y útiles para responder la pregunta de negocio planteada.

- Los datos relacionados con el nivel de apropiación digital (frecuencia de uso, accesibilidad y competencias) son relevantes para el objetivo de segmentar individuos según sus características.
- Sin embargo, algunas variables no están directamente relacionadas con el problema planteado (e.g., variables demográficas redundantes o variables no tan vinculadas con el problema de negocio, lo que subraya la necesidad de un filtrado cuidadoso.

Parte II: Pertinencia de aplicar procesos de limpieza y resultados

2.1. Relevancia del proceso de limpieza

La calidad de los datos es el cimiento sobre el cual se construyen análisis robustos y decisiones estratégicas. En este informe, detallamos las técnicas de limpieza de datos aplicadas, un proceso clave para garantizar que los insumos analíticos cumplan con los estándares necesarios para abordar el problema de negocio con precisión y claridad.

El proceso de limpieza, aunque a menudo considerado una tarea menos sofisticada en Analytics (Informs, 2014), es innegablemente crítico. En particular, cuando los datos provienen de fuentes preexistentes o fueron recolectados para propósitos distintos, como en este caso, es necesario ajustar su estructura y corregir posibles inconsistencias para alinearlos con los objetivos del análisis actual. Además, las particularidades del proyecto exigen un enfoque que considere valores inválidos, rangos atípicos, datos faltantes y correlaciones no deseadas entre campos.

Este apartado abordamos el cómo se aplicaron técnicas específicas de limpieza de datos, justificadas por las características únicas del conjunto de datos y respaldadas por prácticas estándar en Analytics. En concreto, los datos correspondientes a los resultados de la encuesta de campo de apropiación digital del año 2023 cuentan con un total de 1,173 columnas, por lo que tenemos mucha información pero también, como se explicó anteriormente, de características que podrían no cumplir con los estándares de calidad para ser utilizados dentro de la problemática de *analytics* planteada. De manera específica, se proponen varios criterios para evaluar la usabilidad de los datos en base a las dimensiones de calidad evaluadas y la pertinencia de aplicar técnicas de limpieza que veremos en la siguiente sección.

2.2. Descripción de las técnicas empleadas y su pertinencia

En primer lugar, se revisaron todas las variables proporcionadas de forma individual, evaluando su contenido y determinando su nivel de pertinencia con la problemática que se busca resolver. Por dar un ejemplo, para la segmentación que se busca realizar no se necesita información del número del contacto del encuestado, ni se necesita su fecha de nacimiento si ya se tiene su edad. Estas variables se eliminarán del análisis y no serán candidatas a ningún proceso de limpieza. Siguiendo esta lógica, se identificaron 5 grupos de variables candidatas a eliminación.

a. Eliminación General:

Se descartaron observaciones con datos faltantes en variables clave donde la imputación no era razonable. Además, se eliminaron columnas completas que no aportaban información relevante al análisis.

- ✓ Información de contacto de los encuestados: 5 variables fueron eliminadas ya que correspondían a información de contacto de los encuestados, lo cual no resulta relevante para la segmentación a realizar.
- ✓ *Metadata* de encuesta: Información como hora de inicio de la encuesta, dispositivo desde el cuál se responde la encuesta, etc.; con un total de 39 variables eliminadas.

b. Eliminación por redundancia

- ✓ En este caso, se aplicó una técnica para identificar y eliminar columnas duplicadas en el conjunto de datos. El proceso comenzó comparando todas las columnas del DataFrame mediante la transposición de este (df.T), lo que permitió evaluar si alguna columna tenía valores idénticos a otra. Se detectaron aquellas columnas que compartían exactamente el mismo contenido en todas sus filas.
- ✓ Variables con información redundante: Se identificaron un total de 12 variables con información

redundante que ya se encontraba en otras variables.

c. Eliminación por su generación de valor con el problema de negocio

- ✓ Información personal irrelevante: Datos como nombres de los miembros del hogar, números de cédula, etc., que correspondían a información personal que no es de utilidad para determinar características de los grupos de individuos. Se eliminaron un total de 111 variables, ya que existían columnas para registrar datos de hasta 15 miembros del hogar por cada atributo, por ejemplo: nombre_1, nombre_2, …, nombre_15.

- ✓ Información muy desagregada: Esta categoría, la más extensa, agrupa variables con un nivel de detalle demasiado alto, lo que dificulta su uso en la segmentación. Por ejemplo, en lugar de mantener una variable para cada uno de los 25 servicios de *streaming* que indicara si el encuestado tenía acceso o no, se optó por una única variable que resumiera el número total de servicios a los que tenía acceso. También se descartaron variables que contenían la misma información en formatos no estructurados y difíciles de procesar. En total, se eliminaron 595 variables bajo este criterio.

- ✓ Variables sin una definición clara: Otras de las razones por las que las variables fueron eliminadas sin ser candidatas a ningún proceso de limpieza fueron las variables a las que no se les pudo asociar algún significado especifico, ya que tanto su nombre como su contenido resultaron confusos y no se contaba con un diccionario de datos adecuado para poder trabajarlas. En esta sección se identificaron 103 variables.

➢ Variables con un indicador de completitud inferior al 90%

Se consideraron las variables con una completitud total menor al 90% como atributos que no serán candidatos a limpieza por imputación de variables. De forma específica, de las 296 variables que quedaron del descarte del punto anterior, se descartaron 239 variables que no cumplían con este criterio, las otras 57 variables pueden ser sometidas a un proceso de imputación de datos mediante regresión.

➢ Valores extremos

Al ser una encuesta donde principalmente se miden preferencias, no existen variables donde se presenten valores extremos, ya que las respuestas vienen dadas de un listado de opciones.

A continuación se muestra el dataframe resultando luego de la aplicación del proceso de limpieza:

Tabla 2.
Dataframe final

	REGISTRO	N_ENCUESTA	REGIONAL	PB1	SECTOR	REGION	MUNICIPIO	PDET	PERSONAS	GENERO	...	B13_2	B14_1	B14_2	ESTRATO_B26_1	B
0	1877	764678	5	1	526	5	8	2	2	2	...	1	8	2.0	3	
1	1885	900989	5	1	5002	5	13	2	2	2	...	3	1	2.0	2	
2	1889	968646	5	1	5002	5	13	2	2	2	...	1	8	2.0	2	
3	1898	180785	5	1	5006	5	13	2	6	2	...	1	6	1.0	1	
4	1901	734690	5	1	5006	5	13	2	3	1	...	3	8	1.0	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4174	8703	294431	1	1	1	4	32	2	5	1	...	2	3	1.0	3	
4175	8723	516103	1	1	1	4	32	2	4	1	...	3	48	1.0	2	
4176	8726	405383	1	1	1	4	32	2	5	1	...	3	8	2.0	2	
4177	8728	606911	1	1	2	4	32	2	3	2	...	1	3	2.0	2	
4178	8730	440739	1	1	2	4	32	2	8	2	...	1	0	NaN	2	

4179 rows × 58 columns

Elaborado por el equipo consultor, 2024.

Parte III: Resultados del proceso de entendimiento de los datos

El análisis se centra en 58 variables clave, seleccionadas por su relevancia para la investigación. Estas variables, descritas en la Tabla 3, abarcan información demográfica, socioeconómica, y sobre el acceso, uso y percepción del internet.

Tabla 3.

Variables del Dataframe resultante

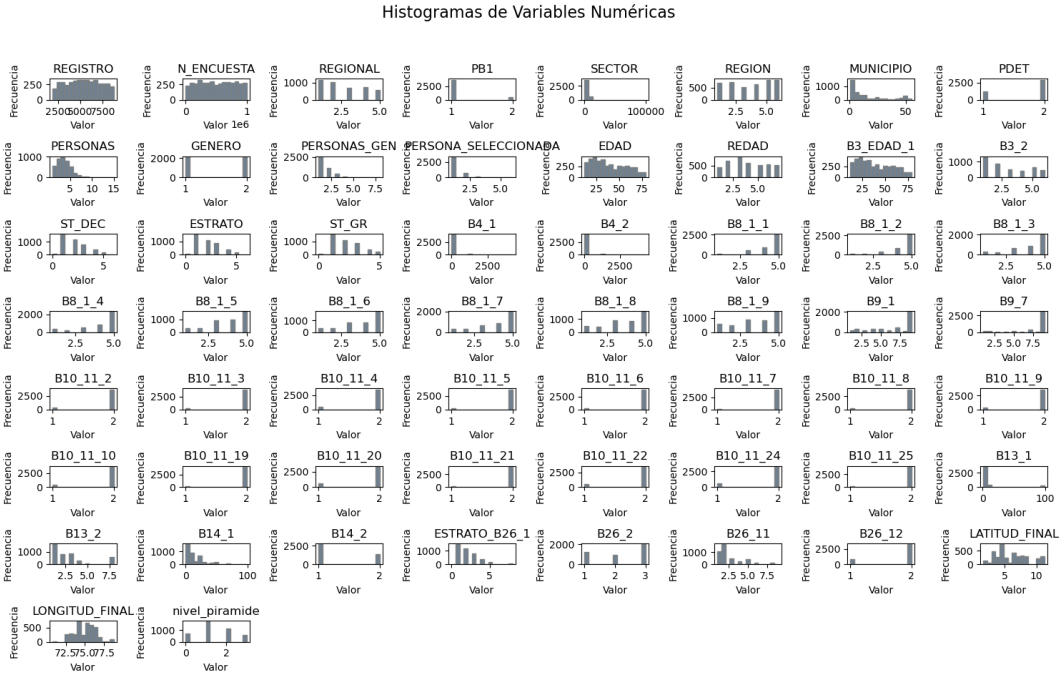
Nombre Variable	Descripción
REGISTRO	Numero de registro
N_ENCUESTA	Número de encuesta
REGIONAL	Regional del CNC
PB1	área geográfica
SECTOR	Sector geográfico
REGION	Región de Colombia
MUNICIPIO	Municipio
PDET	-
PERSONAS	Cantidad de Personas
GENERO	Genero
PERSONAS_GEN	Cantidad de Personas del hogar
PERSONA_SELECCIONADA	Persona que responde la encuesta
EDAD	Edad Persona que responde la encuesta
REDAD	Rango Edad
B3_EDAD_1	Edad 1ra persona
B3_2	Categoría de edad
ST_DEC	-
ESTRATO	Estrato socioeconómico
ST_GR	-
B4_1	Servicios telecomunicaciones

Nombre Variable	Descripción
B4_2	Cual servicio tiene en el hogar
B8_1_1	Percepción del internet
B8_1_2	Percepción del internet
B8_1_3	Percepción del internet
B8_1_4	Percepción del internet
B8_1_5	Percepción del internet
B8_1_6	Percepción del internet
B8_1_7	Percepción del internet
B8_1_8	Percepción del internet
B8_1_9	Percepción del internet
B9_1	Compra por internet
B9_7	Hace cuento se venció el internet
B10_11_2	Dispuesto a suscribirse?
B10_11_3	Dispuesto a suscribirse?
B10_11_4	Dispuesto a suscribirse?
B10_11_5	Dispuesto a suscribirse?
B10_11_6	Dispuesto a suscribirse?
B10_11_7	Dispuesto a suscribirse?
B10_11_8	Dispuesto a suscribirse?
B10_11_9	Dispuesto a suscribirse?

Nombre Variable	Descripción
B10_11_10	Dispuesto a suscribirse?
B10_11_19	Dispuesto a suscribirse?
B10_11_20	Dispuesto a suscribirse?
B10_11_21	Dispuesto a suscribirse?
B10_11_22	Dispuesto a suscribirse?
B10_11_24	Dispuesto a suscribirse?
B10_11_25	Dispuesto a suscribirse?
B13_1	Nivel educativo del padre o madre
B13_2	Nivel de ingles
B14_1	Horas de descanso
B14_2	Descanso suficiente?
ESTRATO_B26_1	Estrato del servicio publico
B26_2	¿Trabaja?
B26_11	Nivel educativo del padre o madre
B26_12	¿Estudia actualmente?
LATITUD_FINAL	Latitud del hogar
LONGITUD_FINAL	Longitud del hogar
nivel_piramide	Pirámide de apropiación digital

Elaborado por el equipo consultor, 2024. (FORMULARIO 2023 - CC_9139-01Apropiacion digital)

- ### Histograma de las variables



Variables Demográficas:

- REGISTRO, N_ENCUESTA: Identificadores únicos.
- REGION, MUNICIPIO: Ubicación geográfica de los encuestados. Permite analizar diferencias en el acceso a internet y su uso entre regiones.
- GENERO: Variable categórica importante para identificar posibles brechas de género en el acceso a la tecnología.
- PERSONAS, PERSONAS_GEN: Tamaño del hogar y número de personas por género.
- EDAD, REDAD, B3_EDAD_1, B3_2: Edad de los encuestados y posiblemente de otros miembros del hogar. Permite analizar cómo varía el uso de internet según la edad.

Variables Socioeconómicas:

- PDET: Indica si la persona reside en un área de Programas de Desarrollo con Enfoque Territorial. Útil para analizar la efectividad de políticas públicas en zonas vulnerables.
- ST_DEC, ESTRATO: Estrato socioeconómico. Permite evaluar la relación entre el nivel socioeconómico y el acceso/uso de internet.
- B13_1, B26_11: Nivel educativo del padre o madre. Puede ser un indicador del capital cultural y su influencia en el uso de la tecnología.
- B26_2, B26_12: Situación laboral y educativa del encuestado. Información relevante para comprender el contexto socioeconómico de los encuestados.

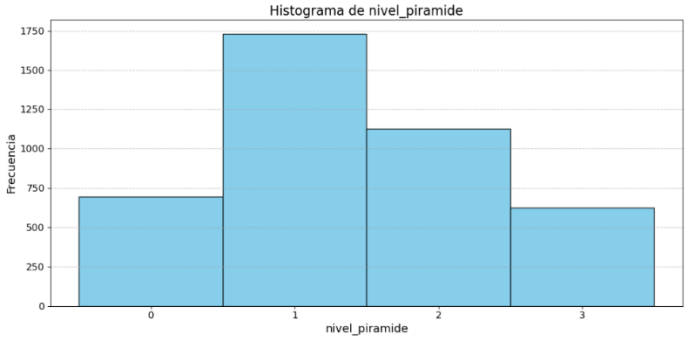
Variables relacionadas con el acceso y uso de Internet:

- PB1, SECTOR: Área y sector geográfico. Similar a "REGION" y "MUNICIPIO", puede ser útil para identificar zonas con mayor o menor acceso a internet.
- ST_GR, B4_1, B4_2: Disponibilidad de servicios de telecomunicaciones en el hogar. Indica el tipo de acceso a internet y otros servicios.
- B8_1_1 a B8_1_9: Percepción del internet. Recoge la opinión de los encuestados sobre la calidad, utilidad e importancia del internet.
- B9_1, B9_7: Uso de internet para compras y frecuencia de uso. Permite analizar hábitos de consumo online.
- B10_11_2 a B10_11_26: Disposición a suscribirse a un servicio. Muestra el interés en adquirir o mejorar el acceso a internet.
- nivel_piramide: Nivel en la pirámide de apropiación digital. Indica el grado de competencia digital de los encuestados.

Otras Variables:

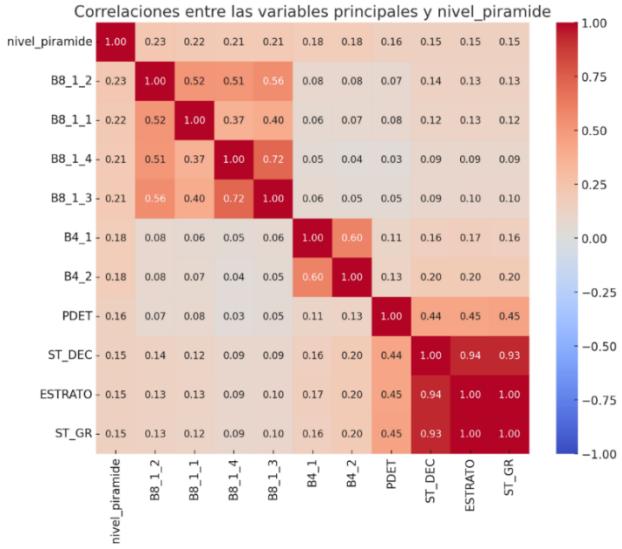
- B13_2: Nivel de inglés. Podría estar relacionado con el acceso a información en línea y la participación en comunidades digitales globales.
- B14_1, B14_2: Horas de descanso. No parece estar directamente relacionado con el tema principal de la encuesta.
- ESTRATO B26_1: Estrato del servicio público. No tengo claro a qué se refiere esta variable.
- LATITUD_FINAL, LONGITUD_FINAL: Coordenadas geográficas del hogar. Información útil para la visualización espacial de los datos.
- NIVEL_PIRAMIDE: Esta variable hace referencia al de uso que hacen las personas al conectarse a internet. Entre mayor sea el número de este índice, mayor es el nivel de apropiación.

Fig. 1. Histograma de las variables

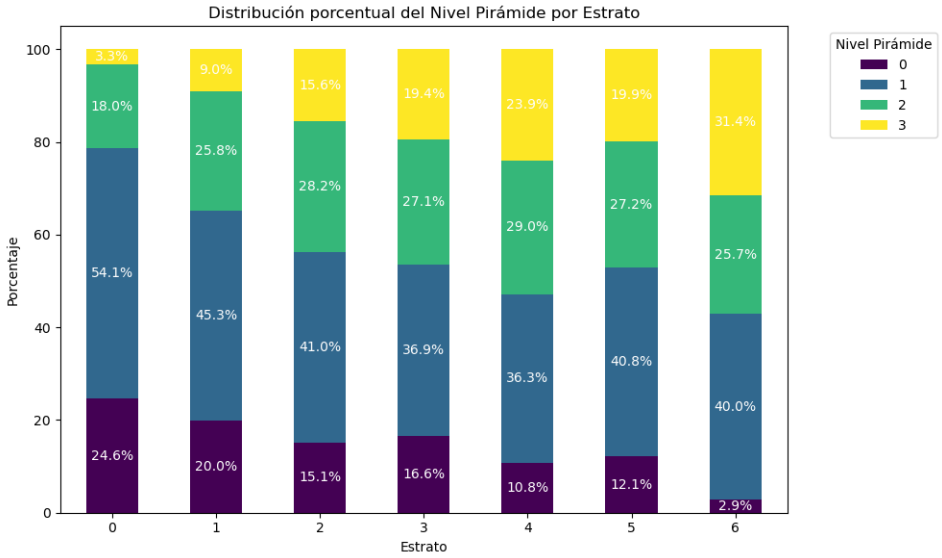


Los datos proporcionados muestran cómo se distribuyen los encuestados en los diferentes niveles: el nivel 1 es el más frecuente, con 1,731 personas, seguido del nivel 2, con 1,126 encuestados. El nivel 0, que podría reflejar una adopción digital muy baja o nula, cuenta con 697 personas, mientras que el nivel 3, el más alto y avanzado, incluye únicamente 625 individuos. La distribución de estas categorías tiene implicaciones importantes para el diseño de políticas públicas y estrategias de intervención. Los niveles bajos (0 y 1) pueden considerarse prioritarios para la implementación de programas enfocados en cerrar la brecha digital.

Encontrar las correlaciones más altas con la variable `nivel_piramide` es crucial porque estas relaciones ayudan a identificar los factores que influyen más significativamente en el nivel de adopción digital. Las variables con alta correlación ofrecen pistas sobre los aspectos que tienen mayor impacto en el nivel de adopción digital. Si encontramos que variables como el estrato socioeconómico o el acceso a internet tienen una fuerte correlación, las políticas pueden enfocarse en cerrar esas brechas.



Descubrimos que variables como `B8_1_1` (Percepción sobre el internet) y `ESTRATO` (nivel socioeconómico) están altamente correlacionadas con `nivel_piramide`, esto sugeriría la necesidad de aumentar la disponibilidad de dispositivos o servicios tecnológicos e implementar subsidios o programas de acceso a internet para personas en estratos bajos.



El análisis de la encuesta sobre estas variables principales permitirá obtener información relevante sobre la apropiación digital en la población estudiada, fundamental para:

- **Diseñar políticas públicas que promuevan la inclusión digital:** Se podrán identificar las zonas y los grupos poblacionales con mayor necesidad de intervención.
- **Desarrollar programas de capacitación en TIC:** Se podrán diseñar programas que se ajusten a las necesidades y características de la población objetivo.
- **Orientar la inversión en infraestructura tecnológica:** Se podrán identificar las zonas donde se requiere una mayor inversión en infraestructura para mejorar el acceso a internet.
- **Promover el desarrollo de contenidos digitales relevantes:** Se podrán identificar las necesidades de la población en términos de contenidos digitales.

Este análisis proporciona información valiosa para comprender la apropiación digital en Colombia y sus determinantes. Los hallazgos destacan la importancia de abordar las brechas digitales existentes a través de políticas públicas integrales que promuevan el acceso equitativo a la tecnología y el desarrollo de habilidades digitales en toda la población.

Anexos: ANEXOS TÉCNICOS