

Informe Análisis **JOVENES** **A LA E**

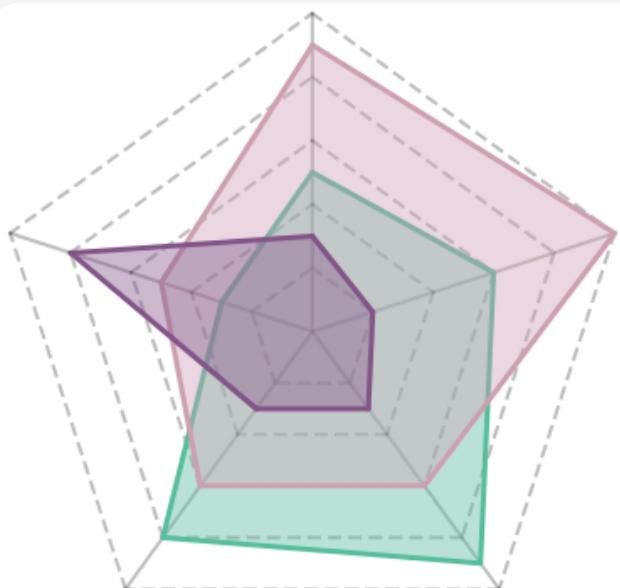
Presentado por VAY



ÍNDICE

<u>Nosotras</u>	01
<u>Equipo</u>	02
<u>Introducción</u>	03
<u>Justificación</u>	04
<u>Objetivo</u>	05
<u>Metodología</u>	06
<u>Modelo de Relación</u>	07
<u>Áreas de Análisis</u>	08
<u>Conclusiones</u>	09

NOSOTRAS



Somos una empresa especializada en las soluciones de analítica avanzada y visualización de datos. Nuestro logo representa el compromiso con el crecimiento, la precisión y la toma de decisiones basadas en datos

VAY nace de la visión conjunta de sus fundadoras: Vanesa, Valentina, Alejandra y Yuliana. Con una gráfica como logo y la analítica como motor, la empresa transforma datos en decisiones inteligentes.

VAY
ANALYTICS



EQUIPO



Alejandra Villa
[LinkedIn](#)
avillaposada9@gmail.com



Vanessa Paternina
[LinkedIn](#)
Vapaternina@outlook.com



Yuliana Gomez
[LinkedIn](#)
yulianago567@gmail.com



Valentina Suárez
[LinkedIn](#)
valentinasb99@gmail.com

INTRODUCCIÓN

Este documento acompaña la entrega de la base de datos estructurada y el dashboard de visualización desarrollado para el programa distrital Jóvenes a la E. Su propósito es describir las metodologías utilizadas en el proceso de limpieza y depuración de datos, así como justificar su importancia para garantizar la calidad, confiabilidad y utilidad de la información contenida en los sistemas desarrollados.

La limpieza de datos es una etapa crítica en cualquier proyecto de análisis o gestión de información, ya que asegura la integridad, coherencia y precisión de los registros, permitiendo una toma de decisiones más efectiva y basada en evidencia.

JUSTIFICACIÓN

La entrega de una base de datos estructurada y un dashboard de visualización para el programa distrital Jóvenes a la E responde a la necesidad de contar con herramientas que faciliten el análisis confiable y oportuno de la información. En este contexto, la limpieza y depuración de datos no solo constituyen un paso técnico, sino un proceso fundamental para asegurar la calidad de los productos entregables.

Por tanto, la justificación de este documento se basa en la necesidad de evidenciar y sustentar los procesos técnicos realizados, demostrando que la información presentada es el resultado de un tratamiento riguroso que prioriza la integridad, la coherencia y la precisión de los datos. Esta labor contribuye directamente a mejorar la efectividad del programa, fortaleciendo la gestión pública basada en evidencia.

ENUNCIADO

Diseño de Base de Datos para la Gestión del Programa “Jóvenes a la E” en Bogotá

Se busca diseñar una base de datos específica para el programa distrital “Jóvenes a la E”, una iniciativa de la Alcaldía de Bogotá orientada a facilitar el acceso de jóvenes a la educación técnica y tecnológica, a través de apoyos económicos, orientación vocacional y acompañamiento integral. El programa tiene como objetivo ampliar las oportunidades educativas de las juventudes bogotanas, especialmente aquellas que enfrentan barreras estructurales para continuar su formación académica.

La base de datos debe permitir el almacenamiento, organización y gestión eficiente de la información relacionada con los jóvenes beneficiarios, las instituciones de educación para el trabajo y el desarrollo humano (ETDH) y los programas técnicos o tecnológicos ofertados. Además, debe facilitar la generación de reportes y análisis que permitan hacer seguimiento a los procesos de inscripción, permanencia y egreso, y evaluar la equidad en el acceso y los resultados del programa.

ENUNCIADO

Cada beneficiario debe estar caracterizado por atributos como: ID y número de documento, edad, género, localidad de residencia, nivel socioeconómico, nivel educativo previo, programa al que accede, e institución educativa correspondiente. También debe registrarse si pertenece a poblaciones priorizadas, como jóvenes víctimas del conflicto armado, personas con discapacidad, comunidades étnicas, población LGBTIQ+, mujeres, entre otros.

Las instituciones educativas deben incluir atributos como ID, nombre, tipo de institución, y oferta académica. Los programas académicos deben estar descritos por su ID, nombre, modalidad.

OBJETIVO

Analizar cómo el proyecto "Jóvenes a la E" ha facilitado el acceso a la educación formal de los jóvenes beneficiados en Bogotá, especialmente de poblaciones vulnerables.

OBJETIVOS ESPECÍFICOS

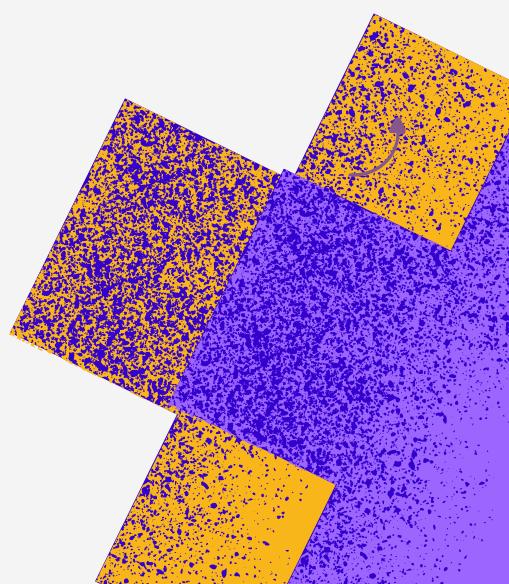
- Eliminar registros duplicados e inconsistentes.
- Completar, corregir campos faltantes o inválidos.
- Homogeneizar formatos de datos (fechas, texto, identificadores).
- Verificar la validez y coherencia interna de los registros (por ejemplo, edad vs. nivel educativo).
- Asegurar la correspondencia entre tablas relacionadas (integridad referencial).



METODOLOGÍA

Para llevar a cabo la limpieza de los datos se aplicaron diferentes metodologías, combinando procesos automáticos (algoritmos y scripts) con revisión manual en casos críticos.

FASES





1

ANÁLISIS INICIAL

Se identificaron y se accedieron a las fuentes de datos necesarias para el análisis. Una vez recopilada la información, se procedió con una exploración inicial orientada a comprender la estructura, volumen, calidad y naturaleza general del dataset disponible.

Durante esta etapa exploratoria se analizaron los siguientes aspectos clave:

- Número de registros y columnas presentes.
- Tipos de datos disponibles (numéricos, categóricos, fechas).
- Identificación de variables clave para el análisis.
- Presencia de datos faltantes, valores duplicados o inconsistencias.

El propósito de esta fase fue obtener una visión preliminar del estado actual y el potencial analítico de los datos, lo que permitió orientar adecuadamente las etapas siguientes de limpieza, transformación y análisis en profundidad.



GOV.CO

DATOS ABIERTOS BOGOTÁ

ALCALDÍA MAYOR DE BOGOTÁ D.C.

BOGOTÁ

Conjuntos de datos Entidades Temáticas Acerca de

Entramos al portal de datos abiertos de Bogotá

Seleccionamos la entidad

Descargamos el archivo

Entidad seleccionada

ATENEA
AGENCIA DISTRITAL PARA LA EDUCACIÓN SUPERIOR, LA CIENCIA Y LA TECNOLOGÍA

Agencia Distrital para la Educación Superior, la Ciencia y la Tecnología

Atenea es creada por medio del Decreto Distrital 73 de 2020 y concebida como una entidad pública de naturaleza a) social, descentralizada adscrita al sector de educación distrital y cuenta con autonomía financiera, administrativa y jurídica, y es el ente responsable de fomentar la articulación entre la educación media y la educación posmedia, para facilitar que los y las jóvenes de la capital puedan acceder a trayectorias de formación gratuita, pertinente y de calidad. Además, busca promover el ac

Buscar conjuntos de datos...

6 conjuntos de datos encontrados

Talento Capital

Agencia Distrital para la Educación Superior, la Ciencia y la Tecnología - Este conjunto de datos contiene la caracterización de los beneficiarios de las diferentes convocatorias del programa Talento Capital (antes Todos a la U). En el conjunto de...

Fecha de actualización: 2025-05-29

Beneficiarios Jóvenes a la E

Agencia Distrital para la Educación Superior, la Ciencia y la Tecnología - Este conjunto de datos contiene información acerca de la caracterización de los beneficiarios del programa Jóvenes a la E. Jóvenes a la E busca fomentar herramientas para el...

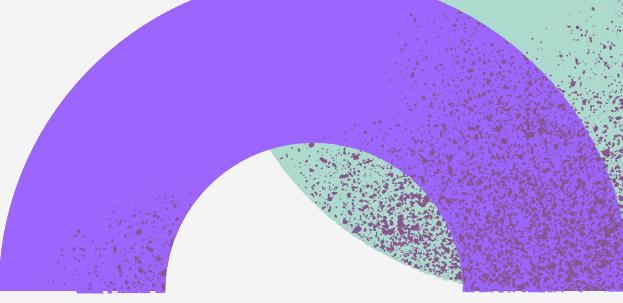
Fecha de actualización: 2025-05-29

Activar Windows

Ve a Configuración para activar Windows

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
CONVOC	DNI	CODLOC	LOCALIDA	CÓDIGO	NOMBRE	NUCLEO	MODALID	SECTOR	ZONA CO	SABER1	SEXO	EDAD AL F	GRUPO E	VICTIMA	DISCAPAC	GRUPOS	BENEFICI
1	JE1	681314054 01	USAQUÉN	1105	UNIVERSIDA EDUCACIÓN PRESENCIAL	NO APLICA	NO APLICA	NO APLICA	NINGUNO	NO	NO	18-21 AÑOS	NINGUNO	NO	NO	SIN SISBEN	2
2	JE1	980872585 01	USAQUÉN	1105	UNIVERSIDA EDUCACIÓN PRESENCIAL	NO APLICA	NO APLICA	NO APLICA	NINGUNO	NO	NO	22-25 AÑOS	NINGUNO	NO	NO	SIN SISBEN	1
4	JE1	88228517 01	USAQUÉN	1105	UNIVERSIDA EDUCACIÓN PRESENCIAL	NO APLICA	NO APLICA	NO APLICA	NINGUNO	NO	NO	26-28 AÑOS	NINGUNO	NO	NO	SIN SISBEN	1
5	JE1	1087703542 01	USAQUÉN	1105	UNIVERSIDA EDUCACIÓN PRESENCIAL	NO APLICA	NO APLICA	NO APLICA	NINGUNO	NO	NO	18-21 AÑOS	NINGUNO	NO	NO	SIN SISBEN	1
6	JE1	928378217 01	USAQUÉN	1701	PONTIFICIA EDUCACIÓN PRESENCIAL OFICIAL	URBANA	75-100	HOMBRE	22-25 AÑOS	NINGUNO	NO	22-25 AÑOS	NINGUNO	NO	NO	B - POBREZA	1
7	JE1	556549823 01	USAQUÉN	1704	UNIVERSIDA PSICOLOGÍA PRESENCIAL OFICIAL	URBANA	75-100	HOMBRE	MENOR DE 18 COMUNIDADES	NINGUNO	NO	NO	NO	NO	NO	B - POBREZA	1
8	JE1	62465 50 01	USAQUÉN	1711	UNIVERSIDA ADMINISTRA PRESENCIAL NO OFICIAL	URBANA	75-100	MUJER	MENOR DE 18 NINGUNO	NINGUNO	NO	NO	NO	NO	NO	C - VULNERA	1
9	JE1	292067672 01	USAQUÉN	1711	UNIVERSIDA ADMINISTRA PRESENCIAL OFICIAL	URBANA	75-100	MUJER	MENOR DE 18 NINGUNO	NINGUNO	NO	NO	NO	NO	NO	SIN SISBEN	1
10	JE1	721267212 01	USAQUÉN	1711	UNIVERSIDA INGENIERÍA I PRESENCIAL OFICIAL	URBANA	50-75	HOMBRE	22-25 AÑOS	NINGUNO	NO	NO	NO	NO	NO	B - POBREZA	1
11	JE1	45135348 01	USAQUÉN	1711	UNIVERSIDA INGENIERÍA I PRESENCIAL OFICIAL	URBANA	50-75	MUJER	18-21 AÑOS	NINGUNO	SI	NO	NO	NO	NO	B - POBREZA	1
12	JE1	5389239 01	USAQUÉN	1711	UNIVERSIDA SIN CLASIFIC PRESENCIAL NO OFICIAL	URBANA	75-100	MUJER	18-21 AÑOS	NINGUNO	SI	NO	NO	NO	NO	SIN SISBEN	1
13	IF1	93641466 01	IRAPUERO	1711	UNIVERSIDA TERAPIAS	PREFENCI	NO OFICIAL	URBANA	75-100	MUJER	18-21 AÑOS	NINGUNO	NO	NO	NO	B - POBREZA	1

Realizamos la inicial exploración de los datos



2

MYSQL

Se identificó y se accedió a la fuente de datos mediante MySQL como sistema de gestión de bases de datos relacional. A través de este entorno, se almacenó y organizó la información estructurada, lo que permitió realizar consultas SQL para acceder a los datos relevantes según los criterios del análisis.

Una vez conectados, se ejecutaron consultas para comprender la estructura general de las tablas y extraer subconjuntos de datos significativos. Esta exploración inicial incluyó:

- Número de registros y columnas.
- Tipos de datos (numéricos, categóricos, fechas).
- Identificación de claves primarias y relaciones entre tablas.
- Presencia de datos duplicados o inconsistentes a nivel de base.

El objetivo fue evaluar la calidad y organización de los datos disponibles para garantizar una extracción precisa y útil para el análisis posterior.





FORMULACIÓN DE PREGUNTAS Y CONSULTAS

1. ¿Cuál localidad tiene el mayor número de postulaciones?

```
SELECT l.nombre_localidad, COUNT(p.participacion_id) AS total_postulaciones
FROM participacion p
JOIN estudiante e ON p.estudiante_id = e.estudiante_id
JOIN localidad l ON e.localidad_id = l.localidad_id
GROUP BY l.nombre_localidad
ORDER BY total_postulaciones DESC
LIMIT 1;
```

nombre_localidad	total_postulaciones
USAQUEN	2255

Solo reporta USAQUEN

Pero al momento de contar cuantas localidades hay con registros arroja que hay 22

```
SELECT COUNT(*) AS total_localidades FROM localidad;
```

total_localidades
22

Indagamos a profundidad

codigo

2. ¿Qué instituciones reciben más beneficiarios?

```
SELECT i.nombre_inst_edu_superior, COUNT(p.participacion_id) AS total_beneficiarios
FROM participacion p
JOIN programa pr ON p.programa_id = pr.programa_id
JOIN inst_edu_superior i ON pr.inst_edu_superior_id = i.inst_edu_superior_id
GROUP BY i.nombre_inst_edu_superior
ORDER BY total_beneficiarios DESC;
```

nombre_inst_edu_superior	total_beneficiarios
UNIVERSIDAD NACIONAL DE COLOMBIA	138
UNIVERSIDAD DE LOS ANDES	111
UNIVERSIDAD EL BOSQUE	108
UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA	98
UNIVERSIDAD ANTONIO NARIÑO	91
FUNDACION UNIVERSITARIA LOS LIBERTADORES	87
UNIVERSIDAD DE LA SALLE	80
UNIVERSIDAD SANTO TOMAS	79
UNIVERSIDAD DISTRITAL-FRANCISCO JOSE DE CALDAS	73
UNIVERSIDAD SERGIO ARBOLEDA	72
UNIVERSIDAD MANUELA BELTRAN-UMB	68
POLITECNICO GRANCOLOMBIANO	66
UNIVERSIDAD EAN	66
COLEGIO MAYOR DE NUESTRA SEÑORA DEL RAYO	65

Resultado

FORMULACIÓN DE PREGUNTAS Y CONSULTAS

3. ¿Cuál es el porcentaje de beneficiarios según grupo étnico?

```
SELECT e.grupo_etnico,
       COUNT(p.participacion_id) AS total_beneficiarios,
       (COUNT(p.participacion_id) * 100.0 / (SELECT COUNT(*) FROM estudiante)) AS porcentaje
FROM estudiante e
JOIN participacion p ON e.estudiante_id = p.estudiante_id
GROUP BY e.grupo_etnico
ORDER BY total_beneficiarios DESC;
```

grupo_etnico	total_beneficiarios	porcentaje
NINGUNO	1163	51.57428
COMUNIDADES NEGRAS O APROCOLOMBIANAS	415	18.40355
SIN INFORMACION	345	15.29933
INDIGENA	330	14.63415
GITANO(A) O RROM	2	0.08869

Pregunta

Consulta en MySQL

5. ¿Qué porcentaje de estudiantes provienen de colegios públicos vs privados?

```
SELECT sector_colegio,
       COUNT(estudiante_id) AS total_estudiantes,
       (COUNT(estudiante_id) * 100.0 / (SELECT COUNT(*) FROM estudiante)) AS porcentaje
FROM estudiante
GROUP BY sector_colegio;
```

sector_colegio	total_estudiantes	porcentaje
NO APLICA	489	21.68514
OFICIAL	1240	54.98891
NO OFICIAL	526	23.32594

6. ¿Cuál es la edad promedio de los beneficiarios? (Rango de edad)

```
SELECT e.rango_edad, COUNT(p.participacion_id) AS total_beneficiarios
FROM estudiante e
JOIN participacion p ON e.estudiante_id = p.estudiante_id
GROUP BY e.rango_edad
ORDER BY total_beneficiarios DESC;
```

rango_edad	total_beneficiarios
18-21 AÑOS	719
22-25 AÑOS	640
MENOR DE 18 AÑOS	515
26-28 AÑOS	335
SIN INFORMACION	46

FORMULACIÓN DE PREGUNTAS Y CONSULTAS

7. ¿Cuál es el porcentaje de estudiantes con discapacidad que participan y resultan beneficiarios?

```
SELECT e.discapacidad,
       COUNT(p.participacion_id) AS total_beneficiarios,
       (COUNT(p.participacion_id) * 100.0 / (SELECT COUNT(*) FROM estudiante WHERE discapacidad <> 'Ninguna')) AS porcentaje_beneficiarios
FROM estudiante e
JOIN participacion p ON e.estudiante_id = p.estudiante_id
WHERE e.discapacidad <> 'Ninguna'
GROUP BY e.discapacidad
ORDER BY porcentaje_beneficiarios DESC;
```

Result Grid		
discapacidad	total_beneficiarios	porcentaje_beneficiarios
NO	2126	94.27938
SI	129	5.72062

8. ¿Qué programas académicos tienen mayor cantidad de beneficiarios?

```
SELECT pr.nombre_programa, COUNT(p.participacion_id) AS total_beneficiarios
FROM participacion p
JOIN programa pr ON p.programa_id = pr.programa_id
GROUP BY pr.nombre_programa
ORDER BY total_beneficiarios DESC;
```

nombre_programa	total_beneficiarios
ADMINISTRACIÓN	196
INGENIERÍA DE SISTEMAS, TELEMÁTICA Y AFI...	146
SIN CLASIFICAR	142
INGENIERÍA INDUSTRIAL Y AFIRES	119
ECONOMÍA	106
CONTADURÍA PÚBLICA	101
INGENIERÍA AMBIENTAL, SANITARIA Y AFIRES	91
EDUCACIÓN	90
INGENIERÍA ELECTRÓNICA, TELECOMUNICACI...	88
DERECHO Y AFIRES	84
COMUNICACIÓN SOCIAL, PERIODISMO Y AFIRES	72
INGENIERÍA MECÁNICA Y AFIRES	71
PSICOLOGÍA	63
DISEÑO	60
ESTADÍSTICA ESTADÍSTICA Y AFIRES	57

9. ¿Hay diferencias en la tasa de beneficiarios entre víctimas y no víctimas del conflicto?

```
SELECT e.victima_conflicto_arm,
       COUNT(p.participacion_id) AS total_beneficiarios,
       (COUNT(p.participacion_id) * 100.0 / (SELECT COUNT(*) FROM estudiante)) AS porcentaje_beneficiarios
FROM estudiante e
JOIN participacion p ON e.estudiante_id = p.estudiante_id
GROUP BY e.victima_conflicto_arm
ORDER BY porcentaje_beneficiarios DESC;
```

victima_conflicto_arm	total_beneficiarios	porcentaje_beneficiarios
NO	1526	67.67184
SI	729	32.32816



FORMULACIÓN DE PREGUNTAS Y CONSULTAS

10. ¿Cómo afecta el grupo de Sisben al acceso a diferentes instituciones educativas?

```
SELECT i.nombre_inst_edu_superior,
       e.grupo_sisben,
       COUNT(p.participacion_id) AS total_beneficiarios,
       (COUNT(p.participacion_id) * 100.0 / SUM(COUNT(p.participacion_id)) OVER (PARTITION BY i.inst_edu_superior_id)) AS porcentaje_sisben
  FROM participacion p
 JOIN estudiante e ON p.estudiante_id = e.estudiante_id
 JOIN programa pr ON p.programa_id = pr.programa_id
 JOIN inst_edu Superior i ON pr.inst_edu_superior_id = i.inst_edu_superior_id
 GROUP BY i.nombre_inst_edu_superior, e.grupo_sisben, i.inst_edu_superior_id
 ORDER BY i.nombre_inst_edu_superior, total_beneficiarios DESC;
```

nombre_inst_edu_superior	grupo_sisben	total_beneficiarios	porcentaje_sisben
COLMUNA MAYOR DE RUSTICA FAVICIA DEL ROSARIO	E-POBREZA MODERADA	30	30.76473
COLMUNA MAYOR DE RUSTICA FAVICIA DEL ROSARIO	C-VULNERABLE	96	24.61538
COLMUNA MAYOR DE RUSTICA FAVICIA DEL ROSARIO	SIN SISBEN	12	18.46344
COLMUNA MAYOR DE RUSTICA FAVICIA DEL ROSARIO	A-POBREZA EXTREMA	90	15.38462
COLMUNA MAYOR DE RUSTICA FAVICIA DEL ROSARIO	D-NO POBRE	7	10.76523
CORPORACION INTERNACIONAL PARA EL DESARROLLO EDUCATIVO -CIDE-	B-POBREZA MODERADA	1	10.00000
CORPORACION TECNICO-INDUSTRIAL COLOMBIANA -TEICO	A-POBREZA EXTREMA	3	33.33333
CORPORACION TECNICO-INDUSTRIAL COLOMBIANA -TEICO	C-VULNERABLE	3	33.33333
CORPORACION TECNICO-INDUSTRIAL COLOMBIANA -TEICO	B-POBREZA MODERADA	3	33.33333
CORPORACION TECNICO-INDUSTRIAL COLOMBIANA -TEICO	D-NO POBRE	3	33.33333
CORPORACION TECNICO-INDUSTRIAL COLOMBIANA -TEICO	SIN SISBEN	3	33.33333
CORPORACION UNIFICADA NACIONAL DE EDUCACION SUPERIOR-CUN-	C-VULNERABLE	7	41.17647
CORPORACION UNIFICADA NACIONAL DE EDUCACION SUPERIOR-CUN-	B-POBREZA MODERADA	4	23.52341
CORPORACION UNIFICADA NACIONAL DE EDUCACION SUPERIOR-CUN-	SIN SISBEN	3	17.64706
CORPORACION UNIFICADA NACIONAL DE EDUCACION SUPERIOR-CUN-	A-POBREZA EXTREMA	2	11.76471
CORPORACION UNIFICADA NACIONAL DE EDUCACION SUPERIOR-CUN-	D-NO POBRE	1	5.88235

Muestra el %de beneficiarios por Sisben que tiene cada institución

12. ¿Cómo varía el porcentaje de beneficiarios mujeres según grupo étnico?

```
SELECT e.grupo_etnico,
       COUNT(p.participacion_id) AS total_beneficiarios,
       (COUNT(p.participacion_id) * 100.0 / (SELECT COUNT(*) FROM estudiante)) AS porcentaje
  FROM estudiante e
 JOIN participacion p ON e.estudiante_id = p.estudiante_id
 GROUP BY e.grupo_etnico
 ORDER BY total_beneficiarios DESC;
```

resultado	total_beneficiarios	porcentaje
NINGUNO	1163	51.57428
COMUNIDADES NEGRAS O AFROCOLOMBIANAS	415	18.40355
SIN INFORMACION	345	15.29933
INDIGENA	330	14.63415
GITANO(A) O RROM	2	0.08869

13. ¿Cuál es el programa más solicitado por convocatoria y localidad?

```
SELECT pr.nombre_programa, COUNT(p.participacion_id) AS total_beneficiarios
  FROM participacion p
 JOIN programa pr ON p.programa_id = pr.programa_id
 GROUP BY pr.nombre_programa
 ORDER BY total_beneficiarios DESC
 LIMIT 1;
```

nombre_programa	total_beneficiarios
ADMINISTRACIÓN	196

3

PYTHON

Se empleó como herramienta principal para el análisis, limpieza y procesamiento de los datos extraídos. Una vez importados desde MySQL, se realizó una exploración inicial utilizando bibliotecas como pandas, numpy y matplotlib, que permitieron una visualización y comprensión detallada del dataset.

Durante esta fase, se analizaron los siguientes aspectos:

- Número de registros y columnas.
- Tipos de datos presentes (numéricos, categóricos, fechas).
- Variables clave para el análisis.
- Presencia de datos faltantes, valores atípicos, duplicados o inconsistencias.

El propósito fue obtener una visión preliminar del estado de los datos, identificar posibles problemas de calidad y preparar el conjunto para los pasos posteriores del análisis exploratorio, estadístico o predictivo.

LIMPIEZA DE DATOS PREVIA

```
# Poblar la base de datos
# Esta sección contiene el código para insertar los datos limpios de los DataFrames en las tablas correspondientes de la base de datos.

# Crear un cursor para ejecutar consultas
cursor = database_connection.cursor()

# 1. Localidades
localidad_ids = {}
for i, row in df_localidad.iterrows():
    cursor.execute("INSERT IGNORE INTO localidad (nombre_localidad) VALUES (%s)", (row['nombre_localidad'],))
    database_connection.commit()
    localidad_ids[row['nombre_localidad']] = cursor.lastrowid # Guardar el ID de la localidad

# 2. Instituciones
for _, row in df_institucion.iterrows():
    cursor.execute("INSERT INTO institucion (inst_edu_superior_id, nombre_inst_edu_superior) VALUES (%s, %s)", (int(row['inst_edu_superior_id']), row['nombre_inst_edu_superior']))
    database_connection.commit()

# 3. Programas
for _, row in df_programa.iterrows():
    cursor.execute("INSERT INTO programa (modalidad, inst_edu_superior_id) VALUES (%s, %s)", (row['modalidad'], int(row['inst_edu_superior_id'])))
    database_connection.commit()

# 4. Convocatoria
convocatoria_info = {
    'J01': (2024, 1),
    'J06': (2023, 2),
    'J05': (2023, 1),
    'J04': (2022, 2),
    'J03': (2022, 1),
    'J02': (2021, 2),
    'J01': (2021, 1)
}

for _, row in df_convocatoria.iterrows():
    anio, semestre = convocatoria_info.get(row['convocatoria_id'], (2020, 1))
    cursor.execute("""
        INSERT INTO convocatoria (convocatoria_id, anio, semestre)
        VALUES (%s, %s, %s)
    """, (row['convocatoria_id'], anio, semestre))
```

```
# Limpieza de datos
# Esta sección se encarga de limpiar los DataFrames eliminando filas duplicadas y filas que contengan valores nulos (NaN).
def limpiar_y_reportar(df, nombre):
    """
    Elimina duplicados y nulos de un DataFrame, muestra info y retorna el DataFrame limpio.
    """
    df = df.drop_duplicates()
    df = df.dropna()
    logging.info(f"\nDataFrame {nombre} después de la limpieza:")
    df.info()
    return df

# Aplicar limpieza a cada DataFrame
df_localidad = limpiar_y_reportar(df_localidad, "localidad")
df_institucion = limpiar_y_reportar(df_institucion, "institucion")
df_programa = limpiar_y_reportar(df_programa, "programa")
df_convocatoria = limpiar_y_reportar(df_convocatoria, "convocatoria")

# 0:05
025-06-02 11:32:05,923 - INFO -
DataFrame localidad después de la limpieza:
class 'pandas.core.frame.DataFrame'
index: 22 entries, 0 to 6599
data columns (total 1 columns):
 # Column          Non-Null Count  Dtype  
--- 
 0) nombre_localidad  22 non-null   object 
types: object(1)
memory usage: 352.0+ bytes
025-06-02 11:32:05,939 - INFO -
DataFrame institucion después de la limpieza:
class 'pandas.core.frame.DataFrame'
index: 52 entries, 0 to 1613
data columns (total 2 columns):
 # Column          Non-Null Count  Dtype  
--- 
 0) inst_edu_superior_id  52 non-null   int64 
 1) nombre_inst_edu_superior  52 non-null   object 
types: int64(1), object(1)
memory usage: 1.2+ bytes
025-06-02 11:32:05,961 - INFO -
```

Limpieza de
Dataframes

Verificación de datos duplicados

```
# Validación de datos
# Verificar si hay datos duplicados y nulos en cada DataFrame

def validar_y_reportar(df, nombre):
    """
    Reporta cantidad de nulos, duplicados y muestra info del DataFrame.
    """
    logging.info(f"\nDatos nan estudiantes: {df.isna().sum()}")
    logging.info(f"\nDatos null estudiantes: {df.isnull().sum()}")
    logging.info(f"\nDuplicados estudiantes: {df.duplicated().sum()}")
    df.info()
```

2025-06-02 11:32:05,792 - INFO - Datos nan estudiantes:

```
rango_edad          0
genero             0
grupo_etnico       0
victima_conflicto_arm 0
discapacidad       0
grupo_sischen      0
percentil_saber1   0
zona_colegio        0
sector_colegio     0
dtype: int64
```

2025-06-02 11:32:05,807 - INFO - Datos null estudiantes:

```
rango_edad          0
genero             0
grupo_etnico       0
victima_conflicto_arm 0
discapacidad       0
grupo_sischen      0
percentil_saber1   0
zona_colegio        0
sector_colegio     0
dtype: int64
```

2025-06-02 11:32:05,834 - INFO - Duplicados estudiantes:

```
# Definir el cursor
cursor = database_connection.cursor()

# Crear base de datos y tablas
cursor.execute("CREATE DATABASE IF NOT EXISTS jovenes_a_la_e")
cursor.execute("USE jovenes_a_la_e")

cursor.execute("""
CREATE TABLE localidad (
    localidad_id INT PRIMARY KEY AUTO_INCREMENT,
    nombre_localidad VARCHAR(100) NOT NULL
)
""")

cursor.execute("""
CREATE TABLE institucion (
    inst_edu_superior_id INT PRIMARY KEY AUTO_INCREMENT,
    nombre_inst_edu_superior VARCHAR(100) NOT NULL
)
""")

cursor.execute("""
CREATE TABLE programa (
    programa_id INT PRIMARY KEY AUTO_INCREMENT,
    nombre_programa VARCHAR(100) NOT NULL,
    modalidad VARCHAR(50) NOT NULL,
    inst_edu_superior_id INT NOT NULL,
    FOREIGN KEY (inst_edu_superior_id) REFERENCES institucion(inst_edu_superior_id) ON DELETE CASCADE
)
""")

cursor.execute("""
CREATE TABLE estudiantes (
    estudiante_id INT PRIMARY KEY AUTO_INCREMENT,
    documento_identidad VARCHAR(20) UNIQUE NOT NULL,
    rango_edad VARCHAR(20) NOT NULL,
    genero VARCHAR(20) NOT NULL,
    grupo_etnico VARCHAR(200) NOT NULL,
    victima_conflicto_arm VARCHAR(10) NOT NULL,
    discapacidad VARCHAR(50) NOT NULL,
    grupo_sischen VARCHAR(100) NOT NULL,
    percentil_saber1 VARCHAR(20) NOT NULL,
    ...
)
""")
```

4

POWER BI

se utilizó como herramienta de visualización y análisis interactivo de los datos procesados. Una vez preparados y depurados los datos en Python, fueron importados a Power BI para construir dashboards y reportes dinámicos que facilitaran la interpretación y comunicación de los resultados.

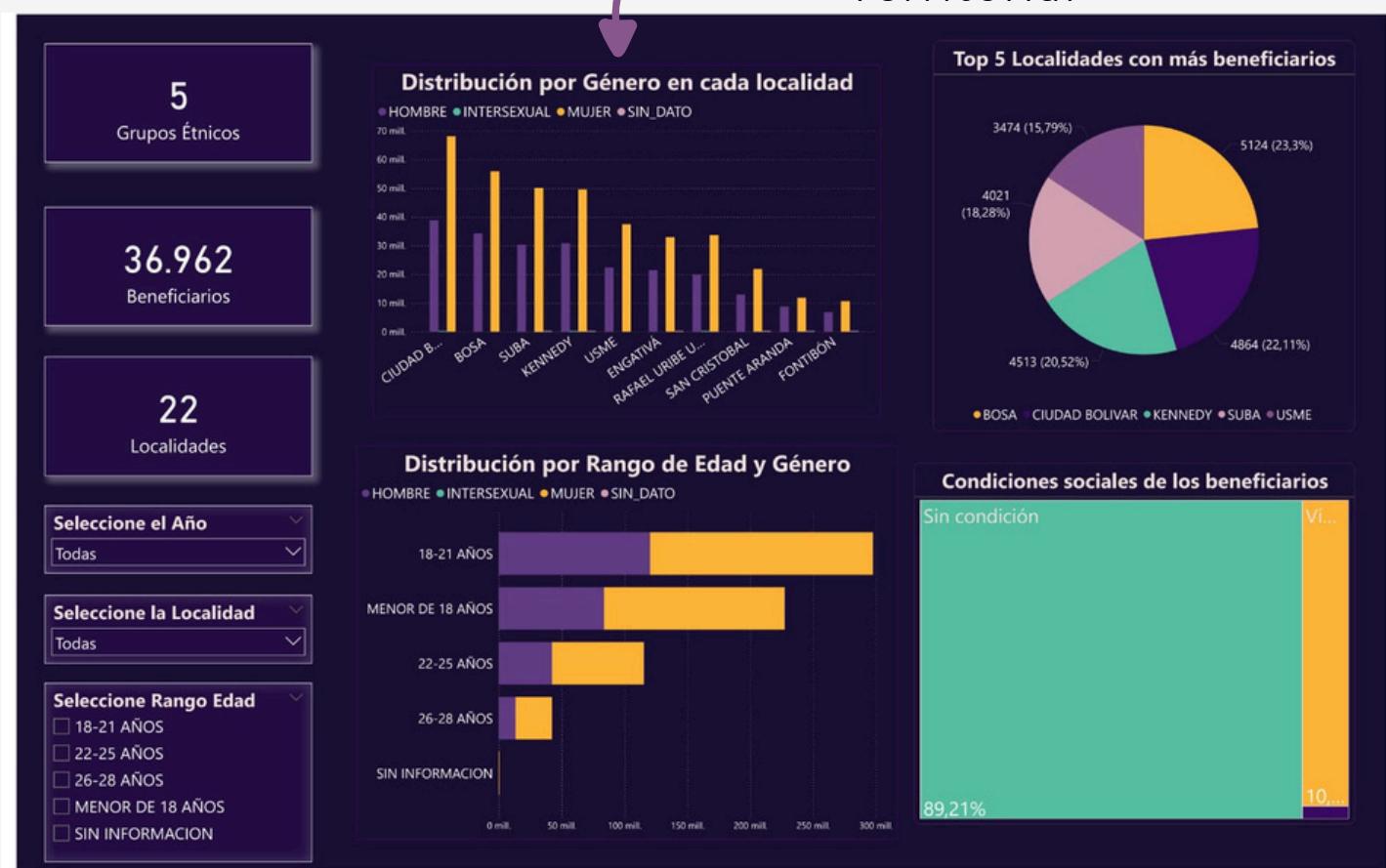
Durante la fase inicial de trabajo con Power BI, se llevó a cabo una exploración general del conjunto de datos dentro del entorno visual, con el fin de entender:

- Número de registros y columnas disponibles.
- Tipos de datos y su correcta detección (numéricos, texto, fechas).
- Variables clave y su relación para la creación de gráficos.
- Detección de datos atípicos o inconsistentes mediante visualizaciones preliminares.

El objetivo fue obtener una visión clara y estructurada del estado de los datos desde una perspectiva visual, permitiendo identificar patrones, tendencias y relaciones significativas que orientaran la toma de decisiones de manera efectiva y accesible para usuarios no técnicos.

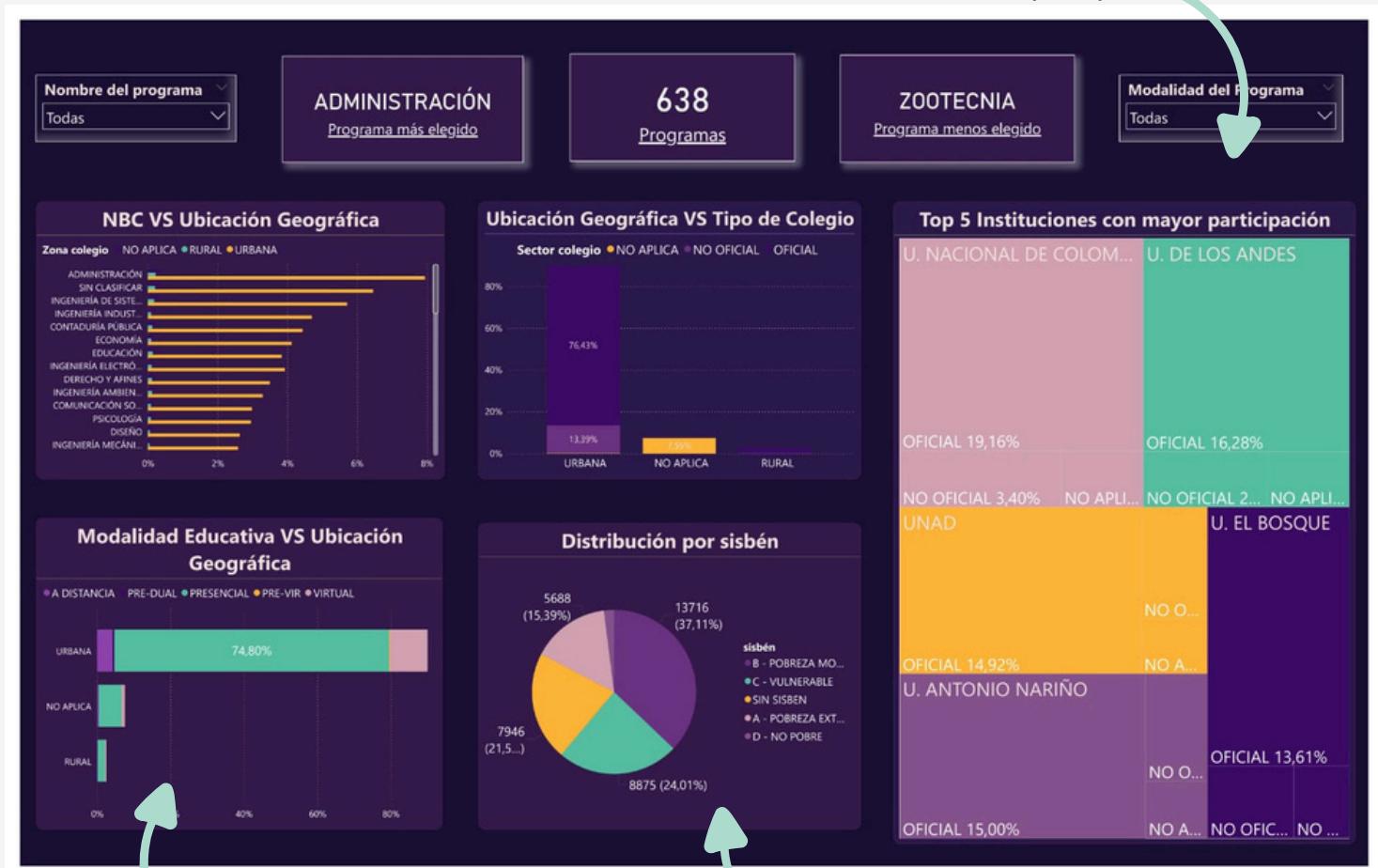
TABLERO 1

Análisis de Cobertura y Representatividad Territorial



TABLERO 2

Distribución por
Institución de
Educación Superior
(IES)

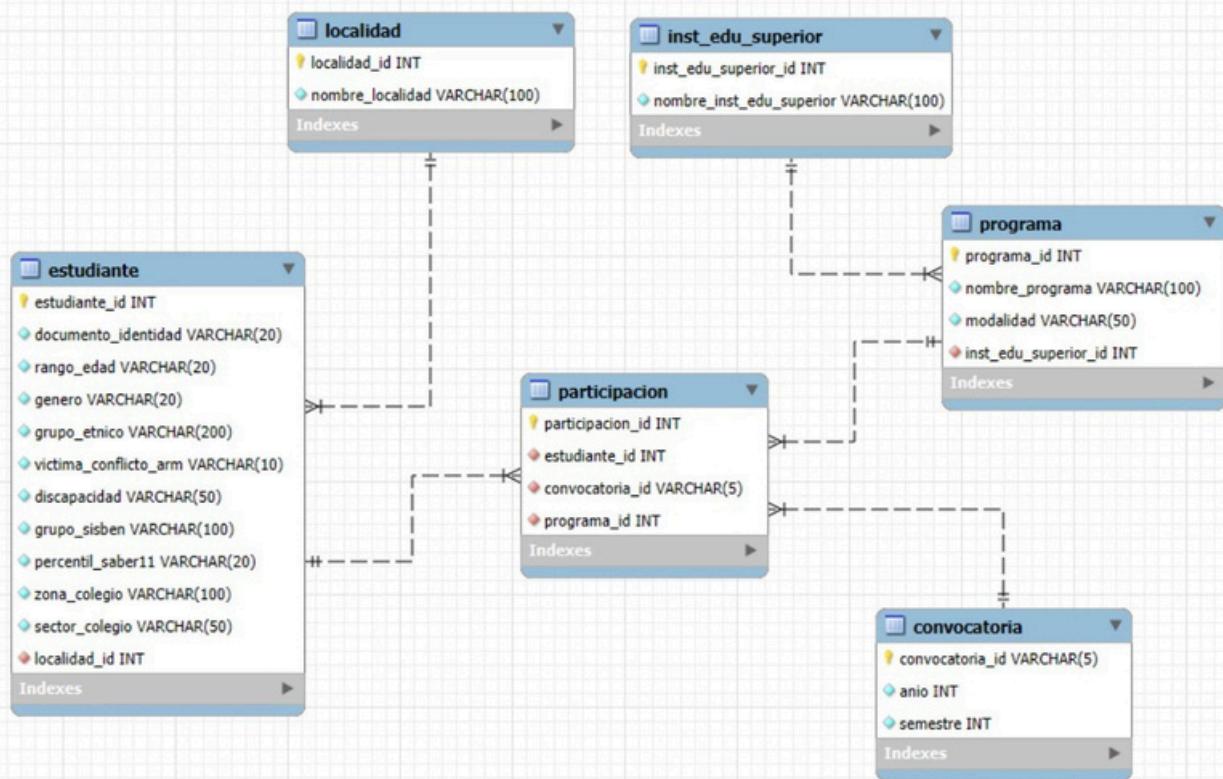


Elección de
Programa
Académico y
Modalidad

Acceso a la Educación
Superior según
condiciones sociales

MODELO DE RELACIÓN

Con el fin de estructurar adecuadamente la información y facilitar su análisis, se definió y estableció un modelo de relación entre las distintas tablas del conjunto de datos. Esta etapa fue fundamental para asegurar la integridad de los datos y permitir un análisis eficiente y coherente en herramientas como Power BI o mediante consultas SQL.



ARÉAS DE ANÁLISIS

Una vez preparada y comprendida la estructura de los datos, se definieron las áreas clave de análisis en función de los objetivos del proyecto. Estas áreas permitieron enfocar los esfuerzos analíticos en dimensiones relevantes para la toma de decisiones estratégicas.

Análisis de Cobertura y Representatividad Territorial:

Verificar equidad territorial y focalización geográfica efectiva.

Acceso a la Educación Superior según condiciones sociales:

Evaluar la inclusión y enfoque diferencial del programa

Análisis de Calidad Académica de los Beneficiarios:

Medir si el programa llega a estudiantes con mayores barreras educativas.

Elección de Programa Académico y Modalidad:

Identificar preferencias formativas y posibles limitantes de acceso.

Distribución por Institución de Educación Superior (IES):

Analizar el papel de las IES en el alcance del programa

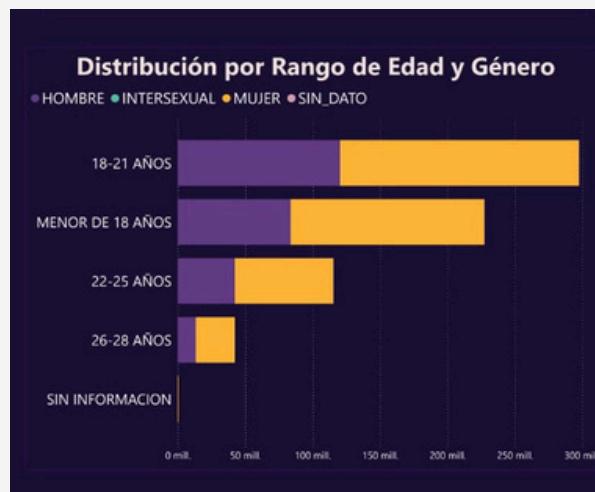


CONCLUSIONES

La base de datos y el dashboard desarrollados para “Jóvenes a la E” mejoran significativamente la gestión y el análisis del programa. La limpieza de datos garantiza información confiable, lo que facilita la toma de decisiones basadas en evidencia y contribuye a una gestión más eficiente y equitativa de las oportunidades educativas para los jóvenes de Bogotá.

PERFIL DEL BENEFICIARIO

La mayoría son **mujeres entre 18 y 21 años**, de colegios oficiales en zonas urbanas y estratos bajos, lo que confirma que el programa llega a población vulnerable.



CONCLUSIONES

ELECCIÓN DE PROGRAMAS E INSTITUCIONES

Los **programas más elegidos** son Administración, Contaduría e Ingeniería, y las universidades con más beneficiarios son la U. Nacional, UNAD y U. de los Andes.



COBERTURA TERRITORIAL

Se impactaron las 20 localidades de Bogotá, destacando **Bosa, Suba, Kennedy y Ciudad Bolívar** como las de mayor participación.



CONCLUSIONES

OPORTUNIDADES DE MEJORA

Falta mayor inclusión de **jóvenes en zonas rurales**, mejorar el registro de condiciones sociales especiales y diversificar la oferta académica promovida.

En general, los datos indican que el programa está logrando un impacto territorial y social relevante



**CREEMOS EN LA DIVULGACIÓN ABIERTA DE LOS DATOS,
POR ESO TE DEJAMOS EL [LINK](#) PARA QUE CONOZCAS
MÁS SOBRE NUESTRO PROYECTO.**