PROYECTO

PROCESAMIENTO DE DATOS A GRAN ESCALA

INTEGRANTES

BRAYAN STEVEN CARRILLO MORA

SANTIAGO BOTERO PACHECO

SANTIAGO AVILÉS TIBOCHA

SANTIAGO RUEDA PINEDA

INDICE DE CONTENIDOS

PROYECTO

1/ MODELO DE NEGOCIO

1'732.561
MILLONES DE USD

NUEVA YORK PIB

$78.982 USD

PIB PER CAPITA EN NUEVA YORK

8'335.897 HABITANTES

POBLACIÓN EN NUEVA YORK

MODELO DE NEGOCIO

Nueva York, una de las ciudades más famosas y económicamente influyentes de los Estados Unidos, enfrenta una serie de desafíos socioeconómicos y territoriales. A pesar de ser una de las ciudades más prósperas y vibrantes del mundo, Nueva York también experimenta disparidades significativas en términos de ingresos, acceso a la educación, seguridad pública y calidad de vida.

OBJETIVO DEL NEGOCIO

El objetivo del equipo de consultoría contratado por el estado de Nueva York es desarrollar un plan de acción basado en el procesamiento de datos para mejorar los indicadores territoriales de interés para el gobierno. Esto incluye identificar áreas de oportunidad para abordar desigualdades socioeconómicas, mejorar la calidad de vida de los residentes y promover un desarrollo sostenible en todo el estado..

PROYECTO

2/ DATOS A ESCOGER

DATOS A ESCOGER

1. Datos de Pobreza en Nueva York: Estos datos ofrecen insights sobre la distribución de la pobreza en Nueva York, incluyendo métricas como ingresos, acceso a servicios sociales y necesidades básicas insatisfechas. Se utilizan para identificar áreas con altos niveles de pobreza y diseñar intervenciones específicas para mejorar las condiciones socioeconómicas en esas áreas.

*2.Datos de Arrestos en Nueva York: Estos datos contienen información detallada sobre los arrestos realizados por el Departamento de Policía de Nueva York (NYPD), incluyendo detalles sobre los delitos, las características demográficas de los arrestados y la ubicación de los arrestos. Se seleccionan para analizar la seguridad pública y la prevalencia de ciertos delitos en diferentes áreas de la ciudad.

3/ COLECCIÓN Y DESCRIPCIÓN DE DATOS
&
4/ EXPLORACIÓN DE LOS DATOS

NYC OpenData

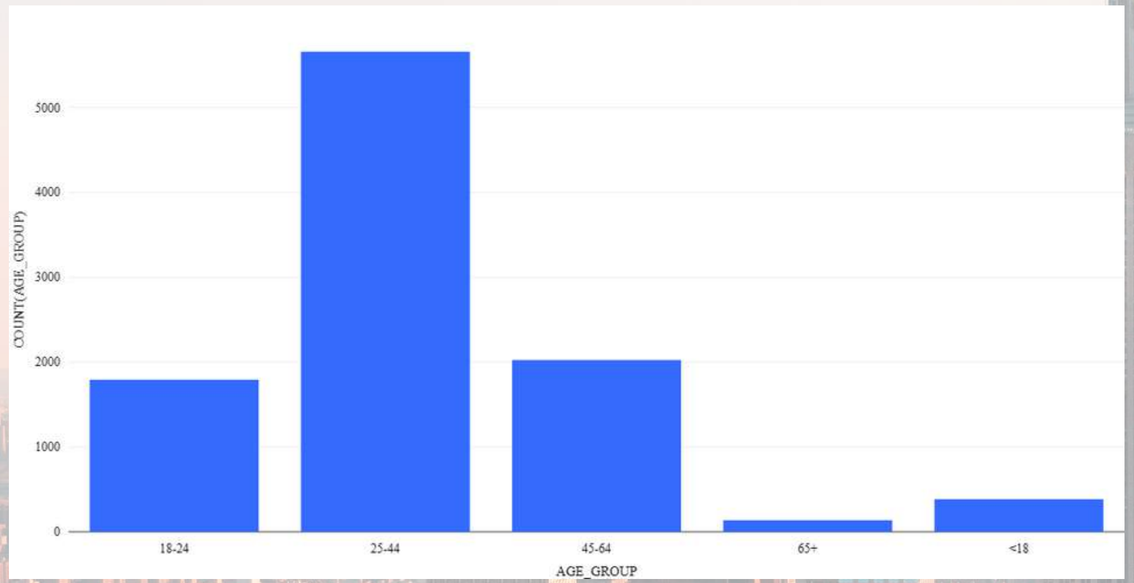
ARRESTOS EN NUEVA YORK

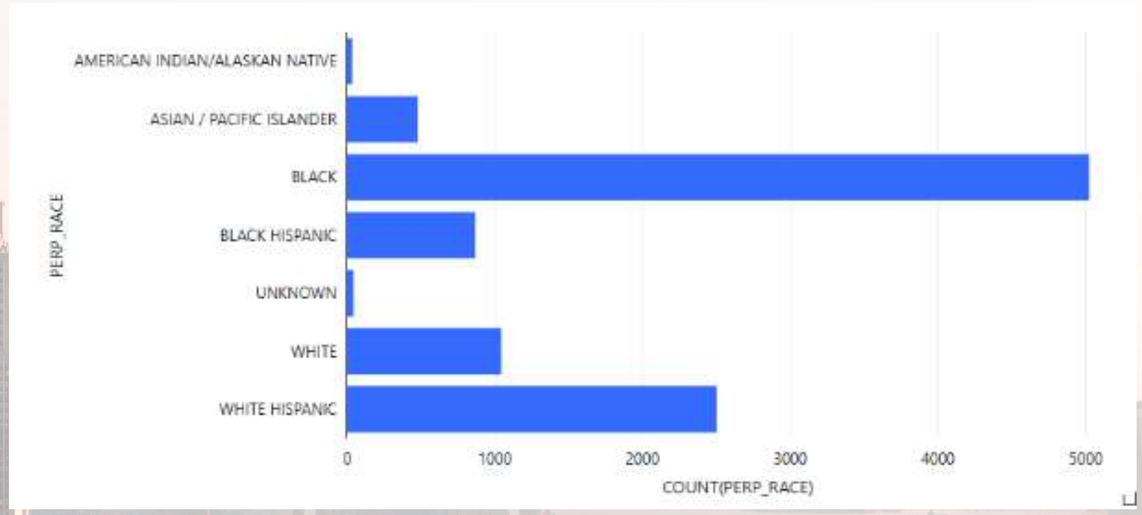
Rows	Columns	Each row is a
227K	19	Arrest in NYC by NYPD

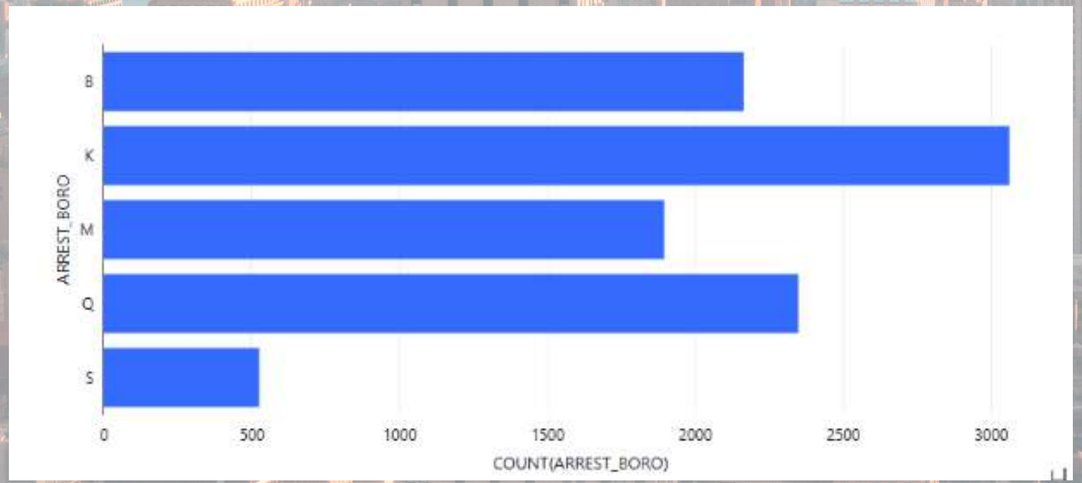
POBREZA EN NUEVA YORK

Rows	Columns
68.3K	61

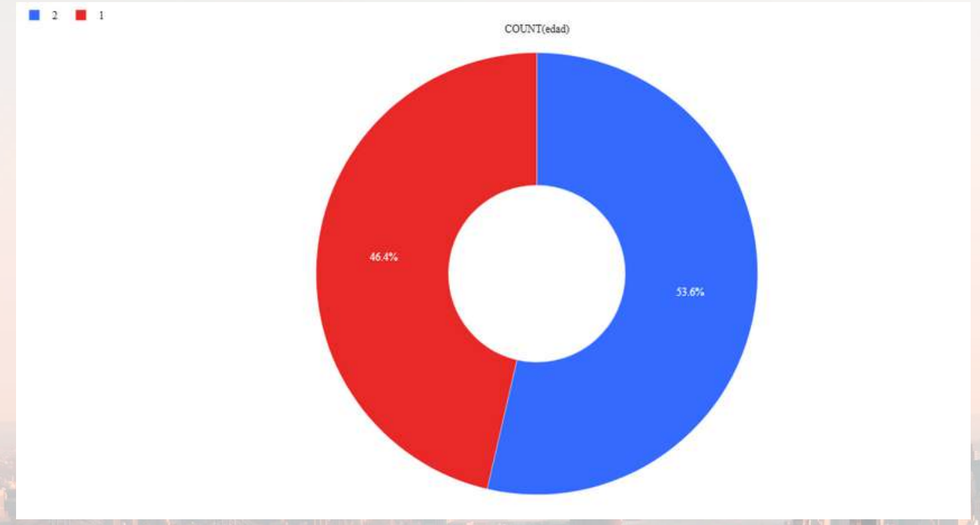
ARRESTOS EN NUEVA YORK

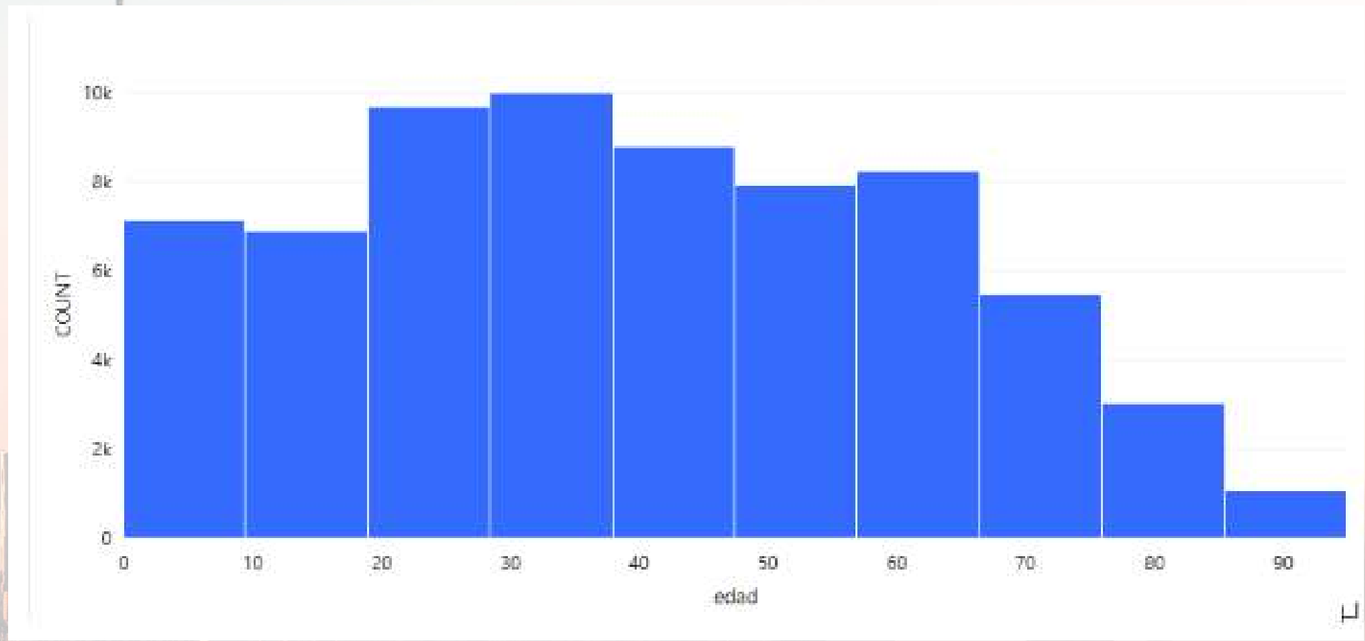


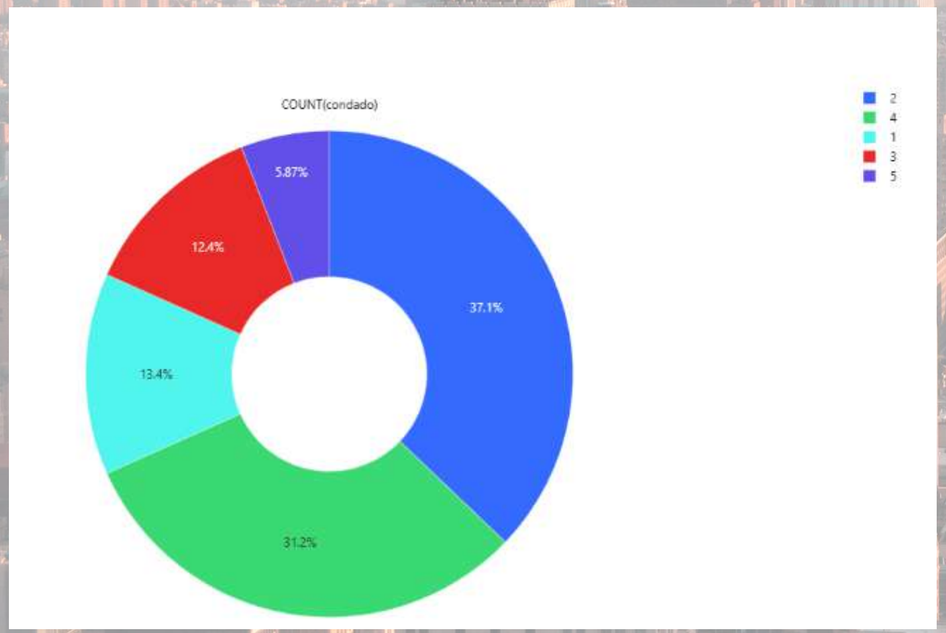


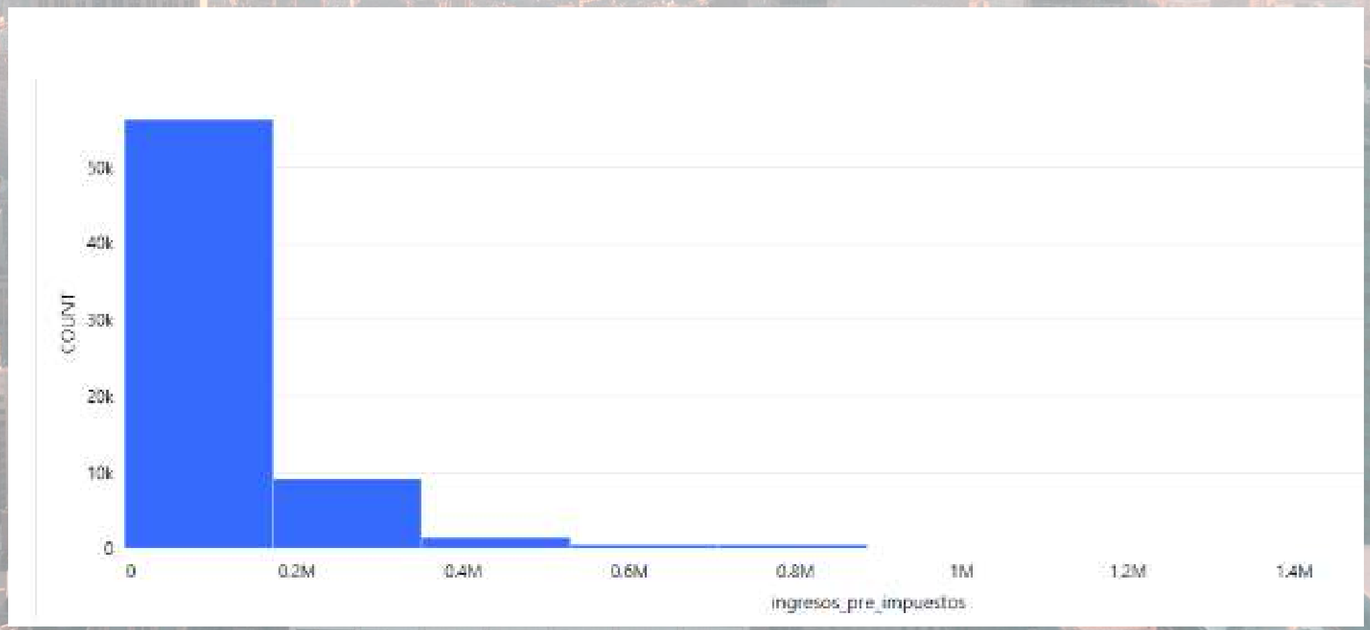


POBREZA EN NUEVA YORK









5/ REPORTE DE CALIDAD DE DATOS

REPORTE DE CALIDAD DE DATOS

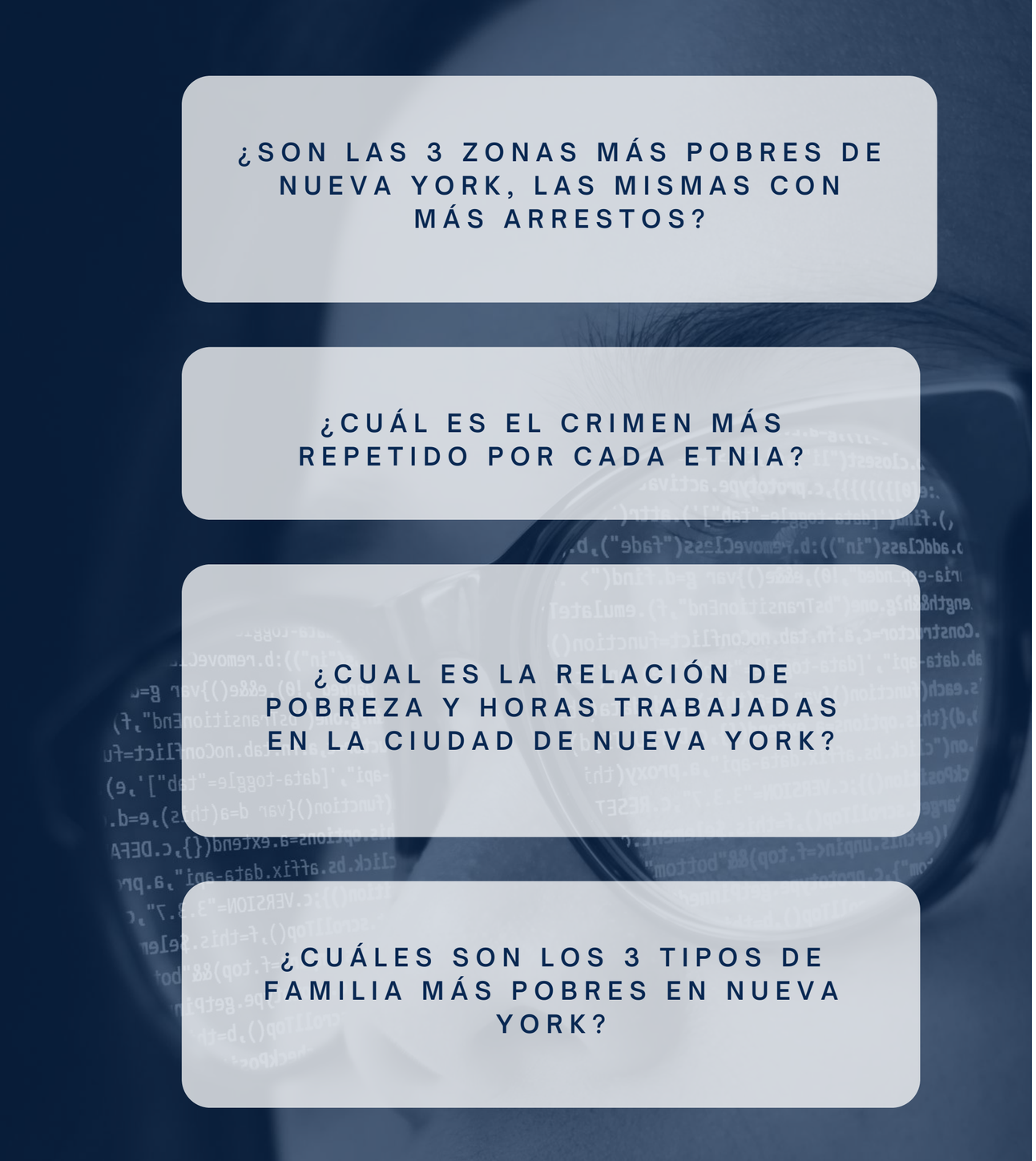
NULOS **DENTRO DE LA POBREZA EN NUEVA YORK**

Variable	Valor
Edad	686
Educación	2101
Sexo	0
Semanas trabajadas	31384
Horas trabajadas	0
Estado civil	11021
Tipo de hogar	0

NULOS **DENTRO DE LOS ARRESTOS EN NUEVA YORK**

Column Name	Value
arrest_key	0
arrest_date	0
pd_cd	2
pd_desc	0
ky_cd	17
ofns_desc	0
law_code	0
law_cat_cd	2210

6/ PLANTEAMINETO SOBRE PREGUNTAS DE LOS DATOS



¿SON LAS 3 ZONAS MÁS POBRES DE NUEVA YORK, LAS MISMAS CON MÁS ARRESTOS?

¿CUÁL ES EL CRIMEN MÁS REPETIDO POR CADA ETNIA?

¿CUAL ES LA RELACIÓN DE POBREZA Y HORAS TRABAJADAS EN LA CIUDAD DE NUEVA YORK?

¿CUÁLES SON LOS 3 TIPOS DE FAMILIA MÁS POBRES EN NUEVA YORK?

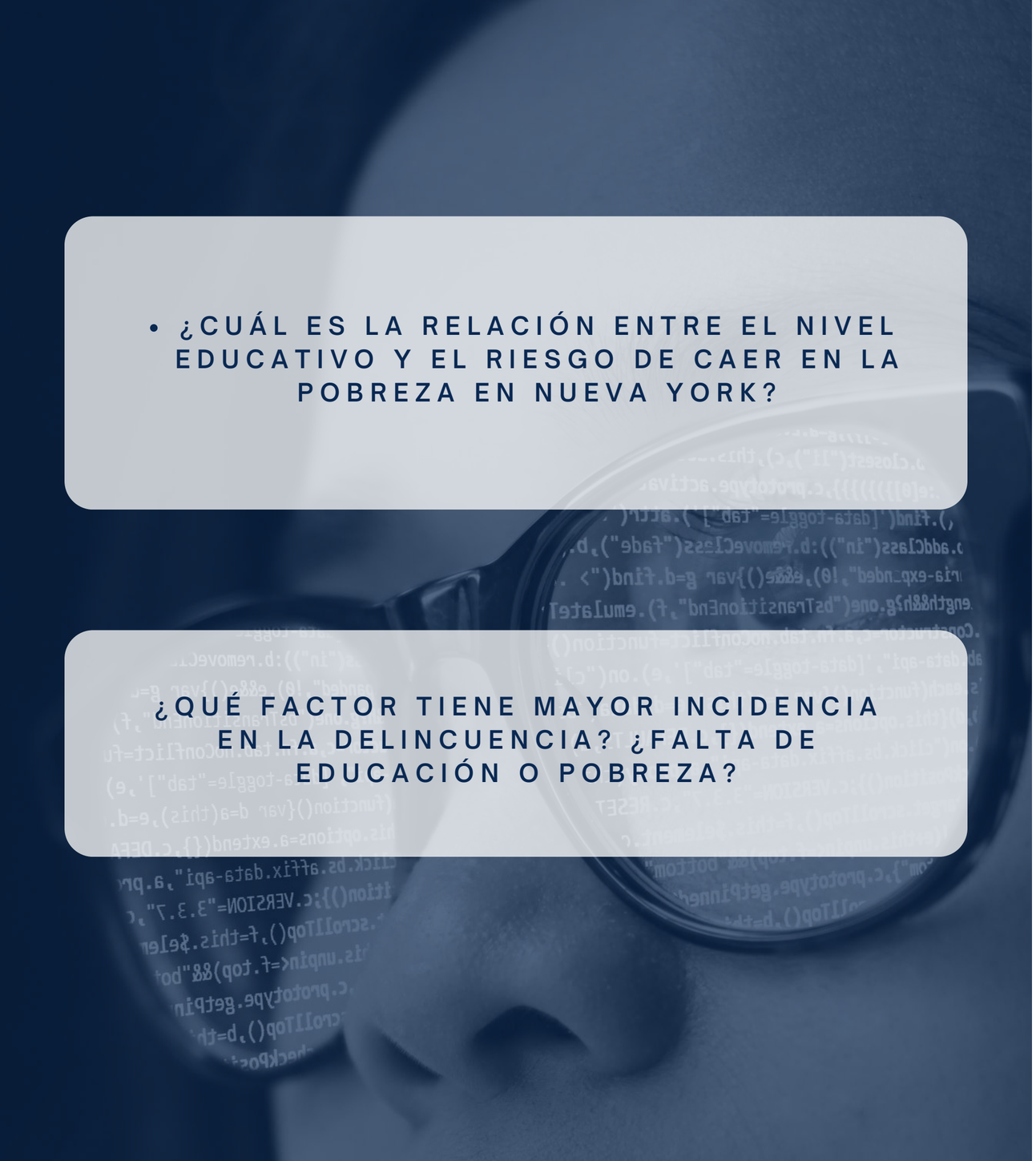
PREGUNTAS

¿CUAL ES LA MODA PARA EL RANGO DE EDAD DEL DELINCUENTE EN NUEVA YORK?

¿CUÁL ES LA PROPORCIÓN DE GÉNERO EN LOS CRÍMENES EN NUEVA YORK?

¿CÓMO VARÍA LA TASA DE POBREZA EN FUNCIÓN DEL ESTADO CIVIL DE LOS RESIDENTES DE NUEVA YORK?

¿CUÁL ES LA RELACIÓN DE GASTOS EN IMPUESTOS RESPECTO A LOS INGRESOS EN LOS HOGARES DE LA ZONA MÁS POBRE DE NUEVA YORK?



- ¿CUÁL ES LA RELACIÓN ENTRE EL NIVEL EDUCATIVO Y EL RIESGO DE CAER EN LA POBREZA EN NUEVA YORK?

¿QUÉ FACTOR TIENE MAYOR INCIDENCIA EN LA DELINCUENCIA? ¿FALTA DE EDUCACIÓN O POBREZA?

PREGUNTAS OPCIONALES

ESTAS PREGUNTAS CLASIFICAN COMO OPCIONALES, DEBIDO A SU RELACIÓN EL CONJUNTO DE DATOS PROPORCIONADO ACERCA DEL NIVEL DE EDUCACIÓN PRESENTE EN LA CIUDAD DE NUEVA YORK.

POR DISTINTOS FACTORES, NO SE HA ESTABLECIDO ORIGINALMENTE DENTRO DEL ALCANCE DEL PROYECTO, SIN EMBARGO, EXISTE UNA POSIBILIDAD PARA LLEGAR A RESPONDER ESTAS INQUIETUDES.

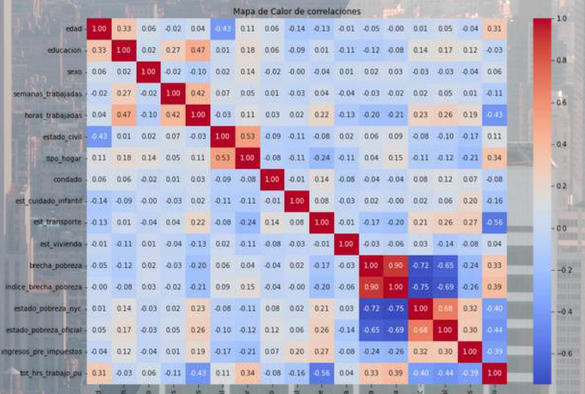
7/ FILTROS, LIMPIEZA Y TRANSFORMACIÓN INICIAL

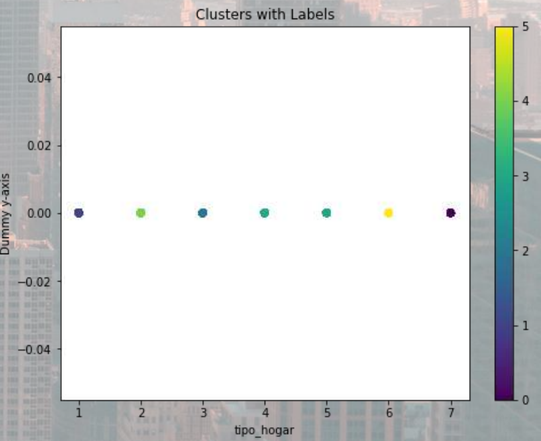
POBREZA EN NUEVA YORK

- TRATAMIENTO DE INCONSISTENCIA. **EDAD 0**
- **ELIMINACION DE DATOS NULOS**
  - EDAD (1%)
  - EDUCACIÓN (3%)
- **INTERPOLACION SEMANAS TRABAJADAS**
  - HORAS TRABAJADAS NO TIENE NULOS
  - EN NUEVA YORK LA NORMA ES TRABAJAR 40 HORAS A LA SEMANA

TRATAMIENTO DE NULOS COLUMNA ESTADO CIVIL

- EL 16% DE LOS DATOS SON NULOS
- USO DE MODELO KMEANS PARA PREDECIR LA COLUMNA
  - COLUMNA TIPO DE HOGAR



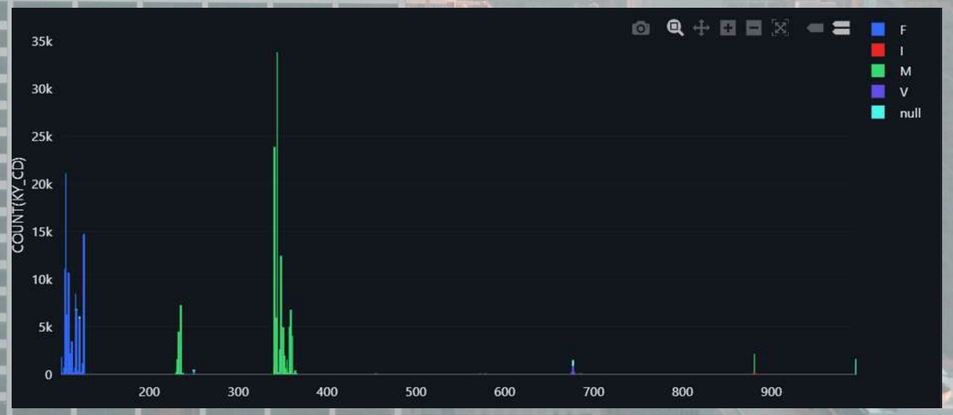


ARRESTOS EN NUEVA YORK

- ELIMINACIÓN DE LAS COLUMNAS QUE POSEAN NULOS

- LA MAYORIA DE NULOS NO ALCANZAN EL 1%
- LA COLUMNA LAW _CAT_CD TIENE UN 9% DE NULOS

- INTERPOLACION CON MODA



8/ BONOS

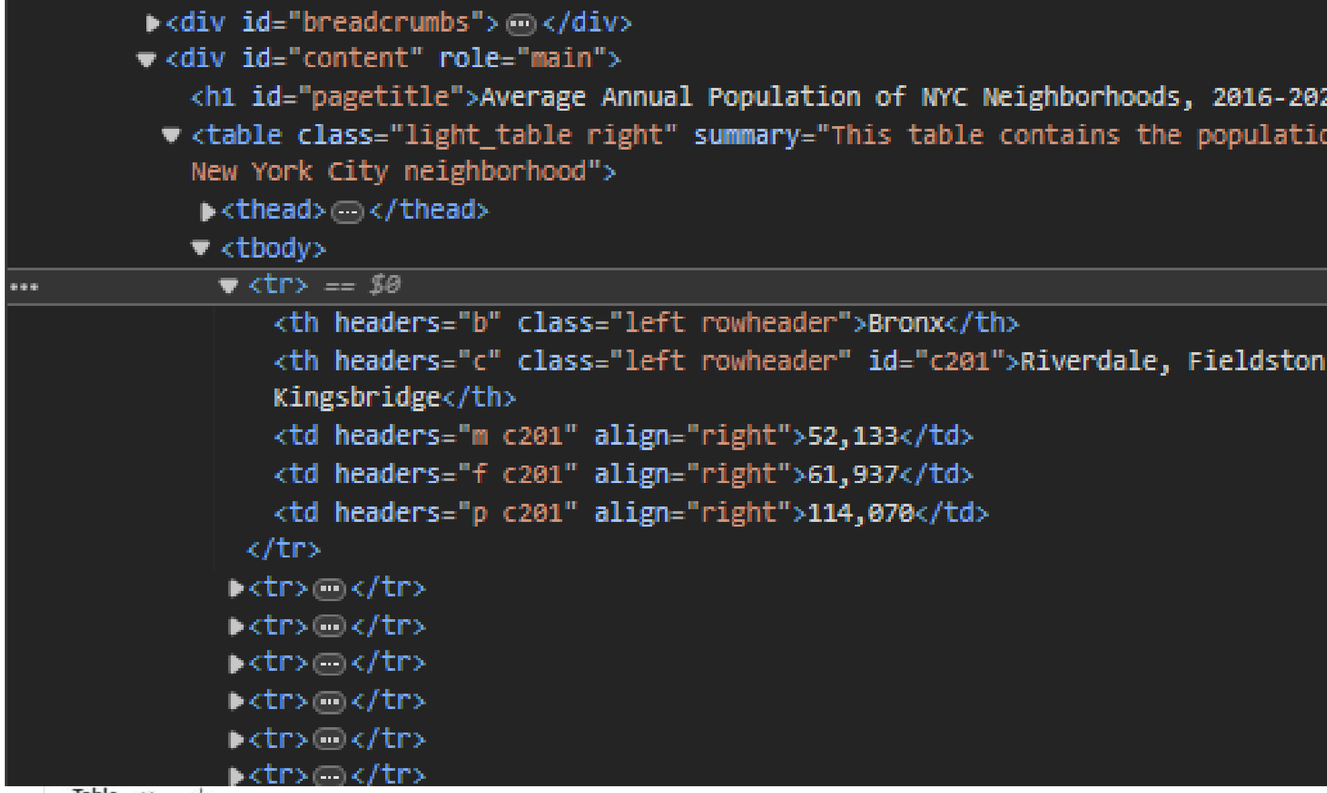
BONOS

Se utiliza requests y la libreria especializada para Web Scrapping BeautifulSoup, identificando si existe un elemento table en la web indicada, en este caso:

https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoodpop.htm

Una vez obtenida la tabla con ayuda de inspeccionar elemento, obtenemos el tipo de elemento en el cual se encuentra la información ya sea <tr>,<th>,<td>, o en casos especiales como las localidades y barrios con los headers ‘b’ y ‘c’.

Ya obtenidos los datos, se almacenan en listas, para construir un dataframe de pandas y finalmente transformarlo en una tabla de PySpark.



Table

	Localidad	Barrio	Hombres	Mujeres	Total
1	Bronx	Riverdale, Fieldston & Kingsbridge	52,133	61,937	114,070
2	Bronx	Wakefield, Williamsbridge & Woodlawn	65,087	77,848	142,935
3	Bronx	Co-op City, Pelham Bay & Schuylerville	55,615	65,929	121,544
4	Bronx	Pelham Parkway, Morris Park & Laconia	61,233	67,896	129,130
5	Bronx	Belmont, Crotona Park East & East Tremont	75,963	87,740	163,704
6	Bronx	Bedford Park, Fordham North & Norwood	62,664	68,016	130,681
7	Bronx	Morris Heights, Fordham South & Mount Hope	64,748	71,644	136,391
8	Bronx	Concourse, Highbridge & Mount Eden	67,535	74,968	142,504
9	Bronx	Castle Hill, Clason Point & Parkchester	87,605	99,401	187,006
10	Bronx	Hunts Point, Longwood & Melrose	80,447	78,645	159,091
11	Kings (Brooklyn)	Greenpoint & Williamsburg	75,780	76,636	152,416
12	Kings (Brooklyn)	Bushwick	65,255	65,960	131,215
13	Kings (Brooklyn)	Bedford-Stuyvesant	63,462	73,197	136,658
14	Kings (Brooklyn)	Brooklyn Heights & Fort Greene	57,864	65,514	123,378
15	Kings (Brooklyn)	Park Slope, Carroll Gardens & Red Hook	52,939	58,276	111,216
16	Kings (Brooklyn)	Crown Heights North & Prospect Heights	57,113	69,223	126,336

BONOS

Se utiliza requests para realizar el llamado al End Point.

https://api.openweathermap.org/data/2.5/weather

Adicionalmente, debemos especificar como parámetros, la latitud y longitud de la ubicación a la cual deseamos consultar el clima, en este caso, realizamos 3 peticiones con las coordenadas correspondientes a Nueva York, Bogotá y Madrid.

También debe especificarse, como parámetros, la API Key proporcionada por la plataforma OpenWeatherMap, así como las unidades, en este caso métricas, ya que por defecto el End Point devuelve unidades imperiales.

Finalmente, reunimos todas las respuestas de las peticiones, en una lista, de la cual extraeremos a listas auxiliares los datos correspondientes a la temperatura, sensación térmica, mínima, máxima, presión y humedad.

Una vez conformadas las listas, son utilizadas para construir un dataframe de pandas y finalmente transformarlo en una tabla de PySpark.

```
temperatura = []
termica = []
minima = []
maxima = []
presion = []
humedad = []
for resultado in resultados:
    temperatura.append(resultado['main']['temp'])
    termica.append(resultado['main']['feels_like'])
    minima.append(resultado['main']['temp_min'])
    maxima.append(resultado['main']['temp_max'])
    presion.append(resultado['main']['pressure'])
    humedad.append(resultado['main']['humidity'])
```

ble

	Ciudad	Temperatura (C°)	Sensación termica (C°
	Nueva York	12.97	11.99
	Bogotá	7.68	6.96
	Madrid	6.25	3.95

3 rows | 0.41 seconds runtime

REFERENCIAS

- J. HAWKSWORTH, T. HOEHN Y A. TIWARI. “WHICH ARE THE LARGEST CITY ECONOMIES IN THE WORLD AND HOW MIGHT THIS CHANGE BY 2025?” ECONOMIC OUTLOOK NOVEMBER 2009. ACCEDIDO EL 9 DE ABRIL DE 2024. [EN LÍNEA]. DISPONIBLE: HTTPS://PWC.BLOGS.COM/FILES/GLOBAL-CITY-GDP-RANKINGS-2008-2025.PDF

- “WEATHER API - OPENWEATHERMAP”. WEATHER API - OPENWEATHERMAP. ACCEDIDO EL 9 DE ABRIL DE 2024. [EN LÍNEA]. DISPONIBLE: HTTPS://API.OPENWEATHERMAP.ORG

- “AVERAGE ANNUAL POPULATION OF NYC NEIGHBORHOODS, 2016-2020”. NEW YORK STATE DEPARTMENT OF HEALTH. ACCEDIDO EL 9 DE ABRIL DE 2024. [EN LÍNEA]. DISPONIBLE: HTTPS://WWW.HEALTH.NY.GOV/STATISTICS/CANCER/REGISTRY/APPENDIX/NEIGHBORHOODPOP.HTM

- “NYPD ARREST DATA (YEAR TO DATE)”. NYC OPEN DATA -. ACCEDIDO EL 9 DE ABRIL DE 2024. [EN LÍNEA]. DISPONIBLE: HTTPS://DATA.CITYOFNEWYORK.US/PUBLIC-SAFETY/NYPD-ARREST-DATA-YEAR-TO-DATE-/UIP8-FYKC/ABOUT_DATA

- “NYCGOV POVERTY MEASURE DATA (2018)”. NYC OPEN DATA -. ACCEDIDO EL 9 DE ABRIL DE 2024. [EN LÍNEA]. DISPONIBLE: HTTPS://DATA.CITYOFNEWYORK.US/CITY-GOVERNMENT/NYCGOV-POVERTY-MEASURE-DATA-2018-/CTS7-VKSW/ABOUT_DATA


MUCHAS GRACIAS