Proyecto - Procesamiento de datos a gran escala.

Entrega 1 - Entendimiento del negocio y entendimiento de los datos

Procesamiento de Datos a Gran Escala

Santiago Botero Pacheco

Santiago Rueda Pineda

Santiago Avilés Tibocha

Brayan Steven Carrillo Mora

Tabla de contenido

1. Entendimiento del 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.

Cabe resaltar que la economía de Nueva York (estado) es la mayor economía regional de los Estados Unidos y la segunda economía urbana más grande del mundo después de Tokio. [1]

Con un PIB de 1,732,561 millones de dólares se avala dicha posición, aunque con un PIB per cápita de $78.982, se coloca en la posición número 12 globalmente por dicha estadística, y nuevamente en la segunda posición a nivel nacional por detrás de Washington D.C.

A pesar de estos indicadores, Nueva York, no se considera una megaciudad, debido a que su población se queda corta con un total estimado de 8,335,897 habitantes con corte de Julio del 2022.

Ya con un poco más de contexto sobre la ciudad de Nueva York, se obtiene acceso a conjuntos de datos que generan múltiples índices de interés donde podemos encontrar los siguientes campos a analizar:

- Datos de arrestos en Nueva York.
- Datos de pobreza en Nueva York.
- Datos de accidentes viales en Nueva York.
- Datos de niveles de educación en Nueva York.

Objetivo de negocio:

Teniendo en cuenta los desafíos en Nueva York, podemos generar nuestro objetivo, este consiste en que nuestro equipo de consultoría contratado por el estado de Nueva York desarrolle 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. El plan de acción debe aprovechar el análisis de datos para informar la toma de decisiones estratégicas y la asignación eficiente de recursos gubernamentales.

2. Selección de los datos a utilizar

En este apartado se listan los conjuntos de datos que se utilizarán durante el proyecto, explicando la razón de esta selección basada en el objetivo de negocio del equipo de consultoría.

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.

La selección de estos conjuntos de datos se basa en la necesidad de comprender y abordar una variedad de problemas socioeconómicos y territoriales en Nueva York, como la pobreza y la seguridad pública. Estos datos proporcionarán información valiosa para desarrollar un plan de acción efectivo que mejore los indicadores territoriales de interés para el gobierno.

3. Colección y descripción de datos

Debido a que los dos datasets principales que hemos elegido son los de pobreza y arrestos en la ciudad de NY, tuvimos que utilizar el API que brinda el 'NYC OpenData'. Una vez tenemos el API, utilizamos el siguiente código para extraer la información y transformarla en un dataframe en pandas y spark:

```
urlPobrezaNY= "https://data.cityofnewyork.us/resource/cts7-vksw.json"
pd_pobreza_df = pd.read_json(urlPobrezaNY)
spark_pobreza_df = spark.createDataFrame(pd_pobreza_df)
```

Pobreza en Nueva York

```
api_endpoint = "https://data.cityofnewyork.us/resource/uip8-fykc.csv"
df_arrestos_ny_pd = pd.read_csv(api_endpoint)
df_arrestos_ny = sql_c.createDataFrame(df_arrestos_ny_pd)
```

Arrestos en Nueva York

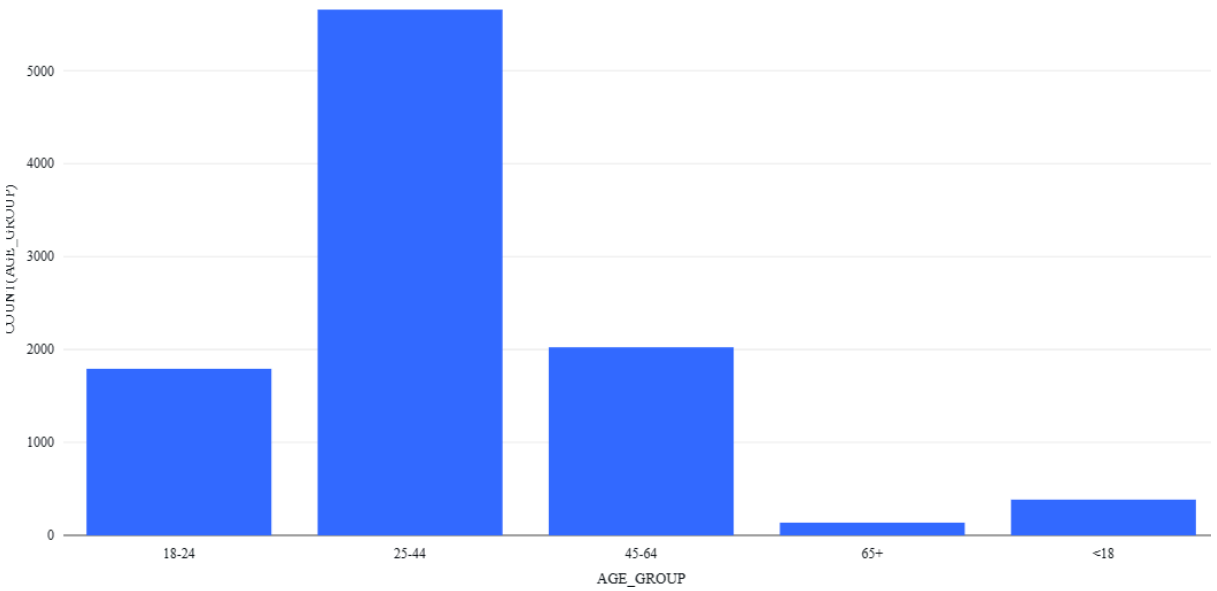
El dataframe de arrestos tenía 19 columnas y el de pobreza 61. Debido a que los nombres de las columnas estaban acortados y eran difíciles de leer, decidimos hacer un diccionario con una explicación de cada columna y su significado. El diccionario puede ser encontrado en el *Anexo 1* al final del documento

Acerca del primer dataset, este contiene información sobre arrestos realizados por la policía de Nueva York. Cada registro incluye detalles como la fecha y hora del arresto, la ubicación, el tipo de delito, la edad y el sexo del perpetrador, entre otros. Este dataset puede ser utilizado para analizar tendencias delictivas, patrones de comportamiento criminal y la eficacia de las estrategias de aplicación de la ley.

Y con respecto al segundo dataset, contiene información sobre variables relacionadas con la pobreza en los Estados Unidos, incluyendo datos sobre ingresos, trabajo, educación, vivienda, salud y nutrición. Esta información puede ser utilizada para analizar las características de las personas que viven en la pobreza, identificar grupos de población vulnerable, medir la profundidad de la pobreza y evaluar la eficacia de las políticas públicas y de seguridad.

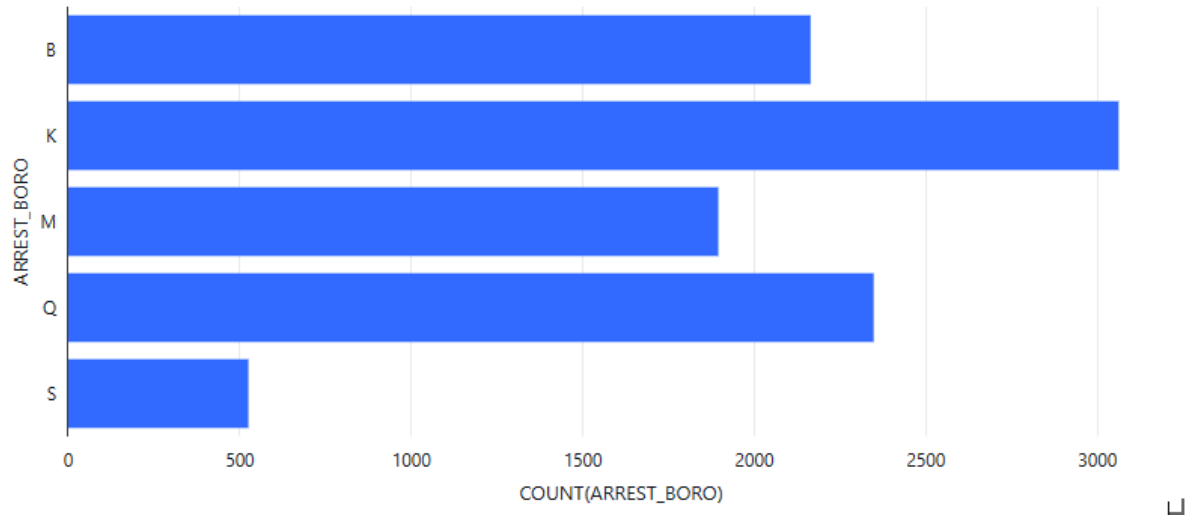
4. Exploración de los datos

Empezando por el primer dataset, mostraremos las gráficas que faciliten el entendimiento de la estructura y comportamiento de los datos de los arrestos. Para la mayoría de las gráficas, se utiliza el diagrama de barras, debido a la cantidad de registros disponibles en los datasets, igualmente, resulta en una gráfica de fácil comprensión para el público en general.



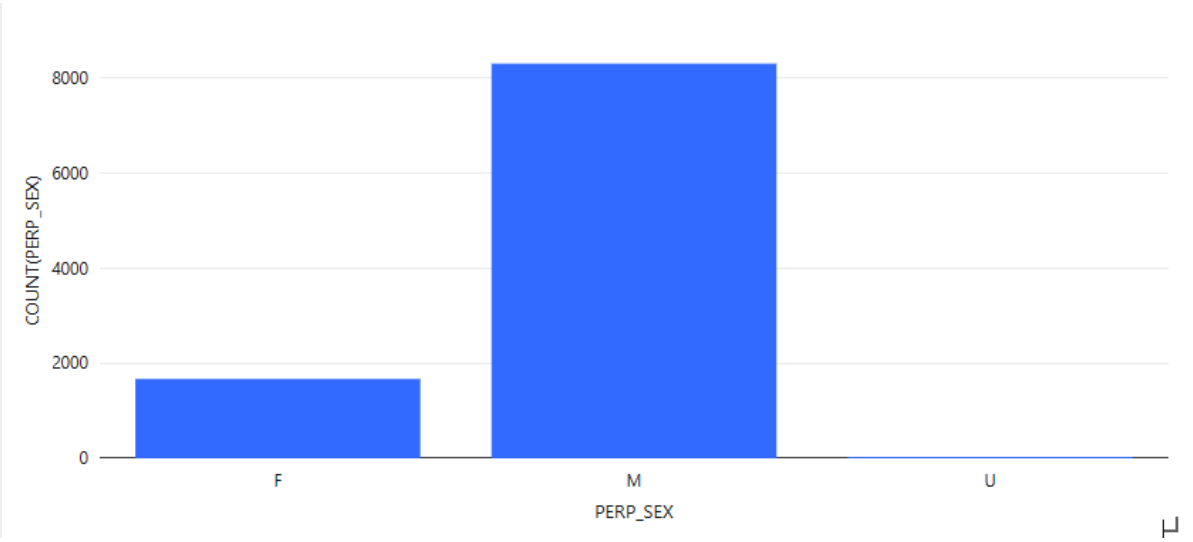
Gráfica 1 - Conteo de edades

La gráfica 1 muestra las edades más comunes asociadas con los arrestos en los registros policiales de Nueva York. Como se puede observar, hay una mayor tendencia de arrestos en personas de entre 25 y 44 años, y además, la mayoría de los arrestos ocurren en personas de entre 25 y 64 años. También se incluyen datos sobre los menores de edad arrestados, lo que abre posibilidades de análisis respecto a la población que aún no ha alcanzado la mayoría de edad.



Gráfica 2 - Arrestos por zonas en NY

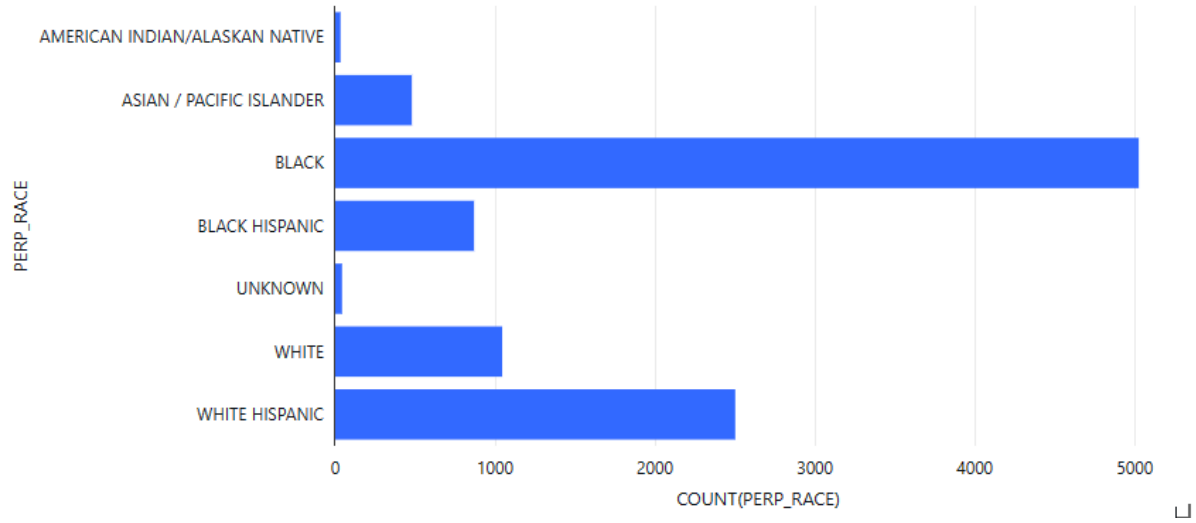
Una particularidad de la ciudad de Nueva York, es su muy marcada división por distritos o localidades. Acá encontramos a Brooklyn, Staten Island, Manhattan, El Bronx y Queens; respectivamente se encuentran en la Gráfica 2 como (K), (S), (M), (B) y (Q). Esto puede ayudar a identificar cómo las distintas poblaciones y políticas dentro de cada uno afectan en temas de seguridad.



Gráfica 3 - Arrestos por sexo

La estadística del sexo de las personas arrestadas en la ciudad es de gran importancia, ya que esta información nos permite dividir todo tipo de análisis en dos poblaciones distintas y examinar cómo afecta a cada una por separado. Claramente, la gráfica muestra un mayor número de hombres siendo arrestados en la ciudad de Nueva York. Esta distinción por sexo

nos proporciona una comprensión más profunda de las dinámicas y los factores subyacentes detrás de los arrestos, lo que puede ser crucial para diseñar políticas y estrategias de prevención del delito efectivas y equitativas.



Gráfica 4 - División por Etnia

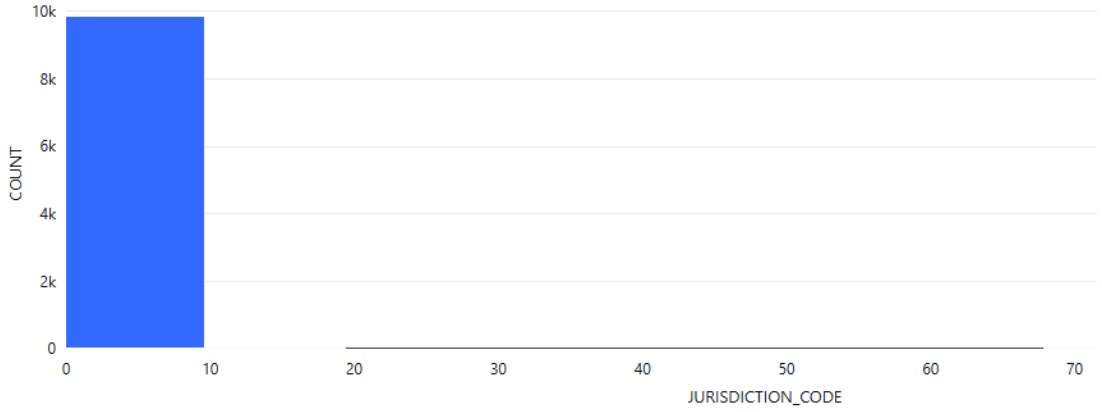
La consideración de las divisiones étnicas en casos de seguridad y arrestos es de suma importancia en el contexto actual, especialmente en un mundo cada vez más conectado a través de las redes sociales. Las suposiciones y percepciones de la gente sobre diferentes grupos étnicos pueden influir significativamente en la forma en que se aplican las políticas de seguridad y cómo se llevan a cabo los arrestos.

La falta de atención a estas divisiones puede llevar a la discriminación y al tratamiento desigual de ciertas comunidades. Por otro lado, un análisis cuidadoso de las estadísticas étnicas de arrestos puede ayudar a identificar patrones de comportamiento delictivo, así como posibles sesgos en la aplicación de la ley. Además, puede servir como una herramienta para abordar las causas subyacentes de la delincuencia en comunidades específicas, como la falta de oportunidades económicas o el acceso limitado a servicios sociales. En resumen, la consideración de las divisiones étnicas en casos de seguridad y arrestos es esencial para promover la equidad, la justicia y la efectividad en las políticas y prácticas de aplicación de la ley.

Gráfica 5 - Crímenes más comunes de los detenidos

Contar con información sobre los delitos más comunes dentro de los arrestos es crucial para comprender los patrones de delincuencia y priorizar recursos en materia de seguridad pública. Identificar los delitos más frecuentes proporciona una visión clara de los problemas de seguridad específicos que enfrenta una comunidad, lo que permite a las autoridades enfocar sus esfuerzos en áreas críticas y diseñar estrategias de prevención más efectivas. Además, este conocimiento puede ayudar a mejorar la asignación de recursos policiales y a desarrollar programas de intervención dirigidos a reducir la incidencia de los delitos más comunes.

Debido a la naturaleza de la variable, se optó por representar la información en una nube de palabras, forma la cual resulta más interactiva para este tipo de información, además que refleja fortaleza en las áreas a tratar.

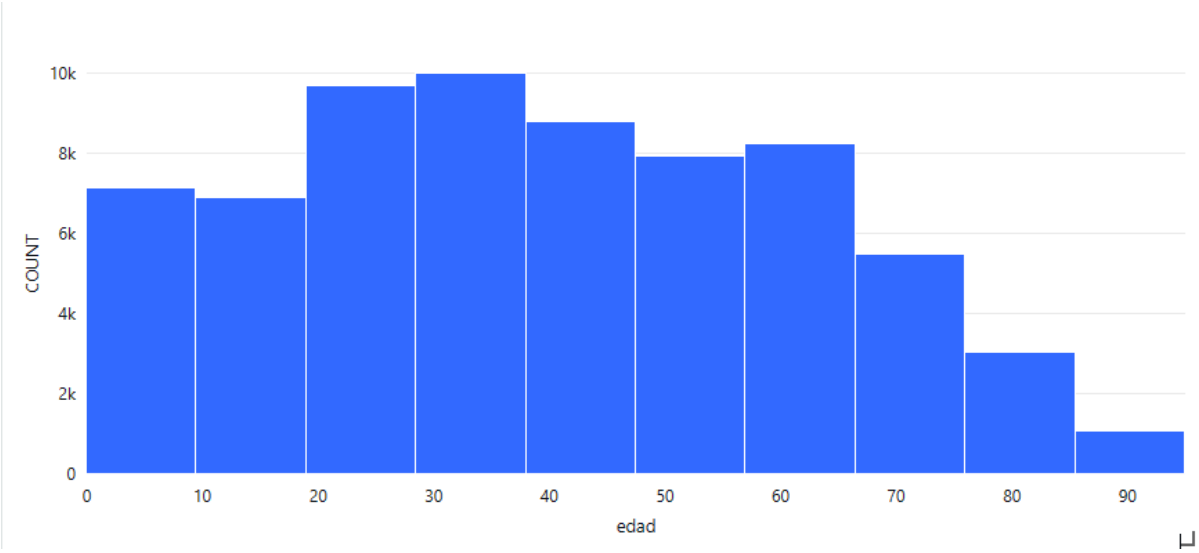


Gráfica 6 - División de código de jurisdicción

La consideración del código de jurisdicción en los datos de arrestos es de suma importancia, especialmente cuando la enorme mayoría de los datos tiene el código 0, que representa la jurisdicción de la patrulla de la Policía de Nueva York (NYPD). Esto puede indicar la predominancia de arrestos llevados a cabo por la NYPD en comparación con otras jurisdicciones. Esta información es crucial para comprender cómo se distribuyen los recursos policiales y cómo se asignan las responsabilidades de aplicación de la ley en diferentes áreas de la ciudad. Además, puede arrojar luz sobre posibles disparidades en el tratamiento de casos dentro de la misma ciudad, dependiendo de la jurisdicción responsable.

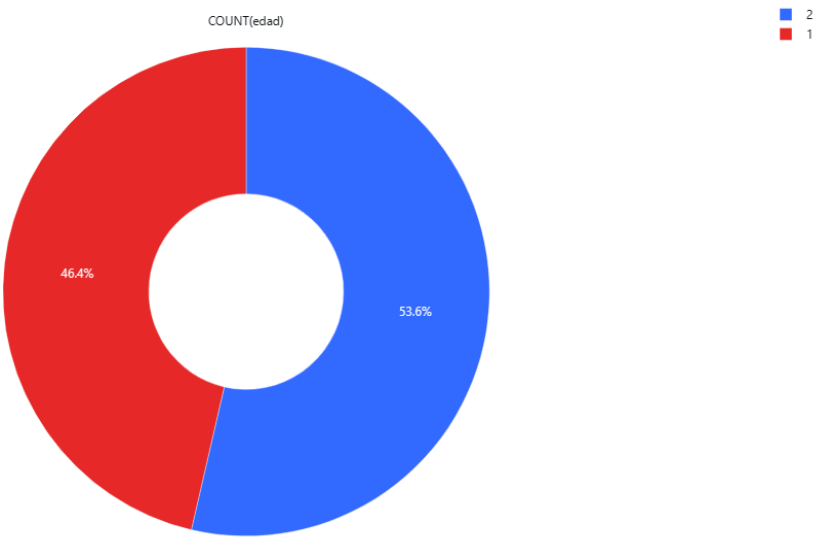
Ahora a continuación mostraremos los datos que obedecen al interés del proyecto pertenecientes al dataset de pobreza en NY.

Debido a la naturaleza del dataset, en este caso todos los datos son numéricos y son mostrados en diagramas de barras.



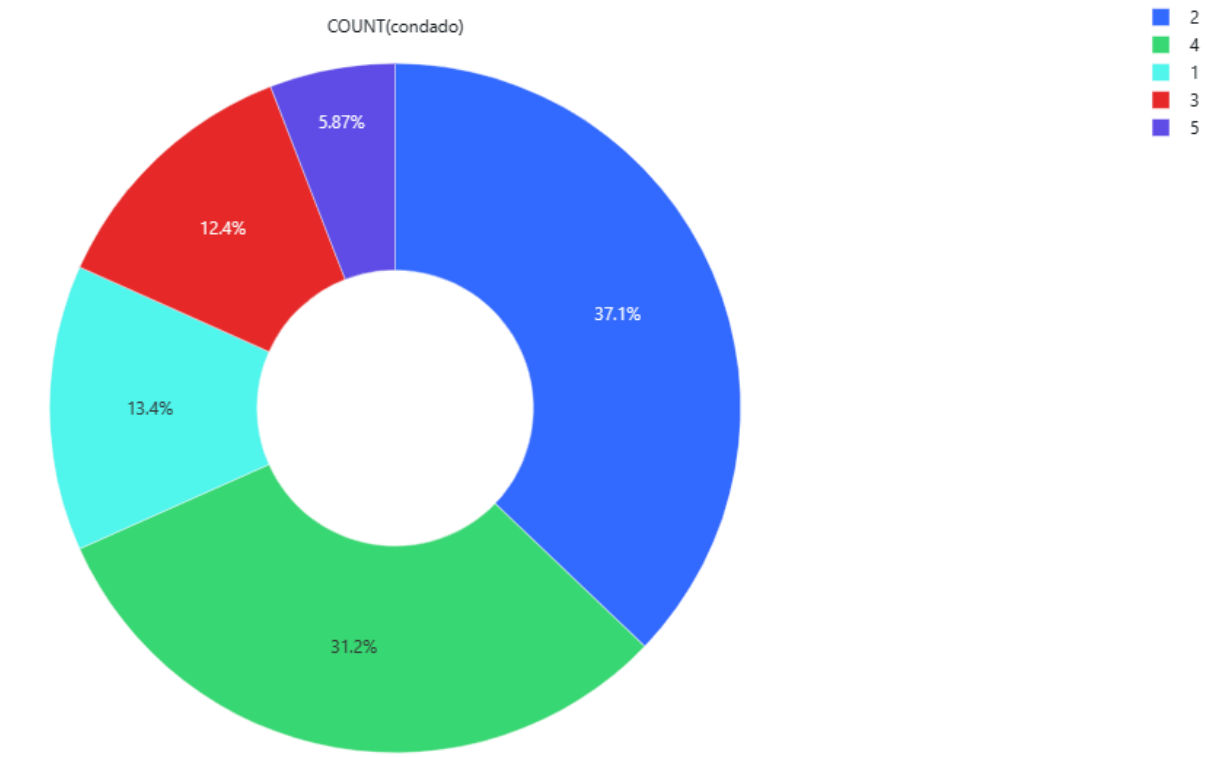
Gráfica 7 - Edades de pobreza

El análisis de las edades en los datos de pobreza en Nueva York es crucial para identificar grupos demográficos vulnerables y dirigir recursos y programas de ayuda de manera efectiva. Revela quiénes están más afectados por la pobreza y dónde se necesitan intervenciones específicas para abordar sus necesidades económicas y sociales. Por ejemplo, el análisis de las edades puede revelar la prevalencia de la pobreza entre los jóvenes en edad de trabajar, lo que podría indicar dificultades para acceder a oportunidades laborales o educativas.



Gráfica 8 - División por sexo

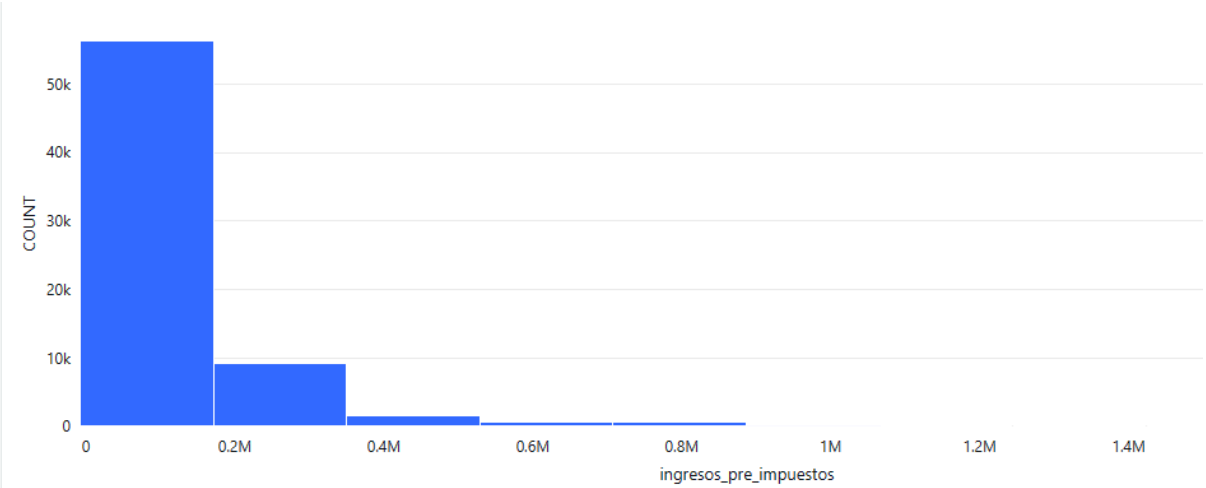
El análisis del sexo en los datos de pobreza en Nueva York es esencial para comprender las disparidades de género en el acceso a recursos y oportunidades económicas. Esto puede incluir iniciativas para cerrar la brecha salarial de género, proporcionar acceso equitativo a servicios de cuidado infantil y promover oportunidades laborales inclusivas. Es fundamental para el entendimiento de la gráfica 8, comprender que en el dataset, se utiliza el código 1, para el género femenino, y el código 2, para el género masculino.



Gráfica 9 - División por distritos

El análisis de los distritos en los datos de pobreza en Nueva York es fundamental para entender las disparidades geográficas en la distribución de la pobreza y dirigir recursos hacia las áreas que más lo necesitan. La asignación de códigos específicos a cada distrito, como Manhattan (1), Brooklyn (2), Queens (3), Bronx (4) y Staten Island (5), en la gráfica proporciona una clara visualización de la distribución de la pobreza en la ciudad. Permite identificar qué distritos enfrentan los mayores desafíos económicos y sociales y desarrollar estrategias personalizadas para abordar sus necesidades únicas.

Por ejemplo, puede revelar si hay áreas específicas con tasas de pobreza más altas y qué servicios o programas son necesarios para mejorar la calidad de vida de los residentes en esos distritos. En resumen, comprender los datos de pobreza desglosados por distritos permite una planificación más efectiva y una respuesta más precisa a los desafíos económicos dentro de la diversa metrópolis de Nueva York.



Gráfica 10 - Ingresos pre-impuestos por persona

El análisis de los ingresos pre impuestos en la gráfica de pobreza en Nueva York es crucial para comprender la magnitud de la desigualdad económica dentro de la ciudad. Aunque pueda parecer obvio que los datos se agrupen predominantemente en la parte baja de la escala de ingresos, esta visualización ofrece una representación clara y cuantitativa de la brecha económica entre los diferentes estratos de la población. Identifica las personas y hogares que están en la parte inferior de la distribución de ingresos y resalta las disparidades en el acceso a recursos económicos y oportunidades. Además, este análisis puede ayudar a informar políticas y programas destinados a reducir la desigualdad de ingresos, como aumentar el salario mínimo, mejorar la accesibilidad a la educación y capacitación laboral, y proporcionar un mayor apoyo a los trabajadores de bajos ingresos. En la gráfica 10, se muestran los ingresos en términos de miles de dólares.

5. Reporte de calidad de datos

Se revisarán los diferentes campos de ambos conjuntos de datos para determinar su integridad y adecuación para su uso en futuros análisis.

Sección de Arrestos

Los datos de arrestos contienen la siguiente información:

Column Name	Value
arrest_key	0
arrest_date	0
pd_cd	0
pd_desc	0
ky_cd	0
ofns_desc	0
law_code	0
law_cat_cd	**8**
arrest_precinct	0
jurisdiction_code	0
age_group	0
perp_sex	0
perp_race	0
x_coord_cd	0
y_coord_cd	0
latitude	0
longitude	0
geocoded_column	0

Gráfica 11 - Tabla datos nulos en el dataset de arrestos en la ciudad de Nueva York.

En general, los datos de arrestos parecen tener una buena calidad, con la mayoría de los campos completos. El único campo con un porcentaje de valores nulos más alto es "*law_cat_cd*" con 8 valores.

Sección de Pobreza

Los datos de pobreza contienen la siguiente información:

Variable	Valor
Edad	0
Educación	**24**
Sexo	0
Semanas trabajadas	**511**
Horas trabajadas	0
Estado civil	**181**
Tipo de hogar	0
Condado	0
Estado de cuidado infantil	0
Estado de transporte	0
Estado de vivienda	0
Brecha de pobreza	0
Índice de brecha de pobreza	0
Estado de pobreza NYC	0
Estado de pobreza oficial	0
Ingresos pre-impuestos	0
Total de horas trabajadas por persona en la unidad familiar	0

Gráfica 12 - Datos nulos presentes en el dataset de pobreza en Nueva York.

En los datos de pobreza, se observa un mayor porcentaje de valores nulos en los campos "educación" (24), "semanas_trabajadas" (511) y "estado_civil" (181). Estos campos deberán ser revisados con atención antes de realizar cualquier análisis pero en general se hallan en muy buen estado de integridad

Una vez revisados y tratados, incluyendo aquellos que sean 0 en contextos incorrectos, los datos podrán ser utilizados de manera efectiva para realizar análisis sobre los temas de arrestos y pobreza en la ciudad de Nueva York.

6. **Planteamiento de preguntas sobre los datos**
    - ¿Son las 3 zonas más pobres de Nueva York, las mismas con más arrestos?
    - ¿Cual es la moda para el rango de edad del delincuente en Nueva York?
    - ¿Cuál es el crimen más repetido por cada etnia?
    - ¿Cuál es la proporción de género en los crímenes en Nueva York?
    - ¿Cuáles son los 3 tipos de familia más pobres en 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?
    - ¿Cómo varía la tasa de pobreza en función del estado civil de los residentes de Nueva York?
    - ¿Cual es la relación de pobreza y horas trabajadas en la ciudad 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?
7. **Filtros, limpieza y transformación inicial**

    Una vez planteadas las preguntas que se tienen en relación a los datos podemos empezar a hacer el proceso de filtrado y limpieza de datos, ya que estas sirven de enfoque y guía para determinar el cómo modificar o limpiar los datos. En este caso, tenemos dos datasets conformados por un gran número de datos, que según lo expresado en el reporte de calidad de datos, tienen muy pocas incosnsitencias.

    Sin embargo, contamos con varios valores nulos que pueden dificultar el procesamiento futuro de los datos, por esto es que las principales transformaciones y filtros realizados fueron implementados en pro de limpiar los datos.

    Sección de Pobreza

Para el dataset de Pobreza en Nueva York nos encontramos con varias columnas que contienen valores nulos como lo son las columnas de *schl* , *msp* y *wkw*. Además, podemos encontrar una inconsistencia en la columna nombrada agep, esto debido a que según el diccionario de datos esta columna hace referencia a la edad, y el valor mínimo de la columna es 0, lo cual a simple vista no tiene sentido.

Partiendo de estos fenómenos evidenciados en el dataset, empezamos por cambiar el nombre de la columnas a unos los cuales fueran las interpretables. Los cambios se expresan en la siguiente tabla:

Nombre Anterior	Nombre Nuevo
agep	edad
schl	educacion
sex	sexo
wkw	semanas_trabajadas
wkhp	horas_trabajadas
msp	estado_civil
hht	tipo_hogar
boro	condado
est_childcare	est_cuidado_infantil
est_commuting	est_transporte
est_housing	est_vivienda
est_povgap	brecha_pobreza
est_povgapindex	indice_brecha_pobreza
nycgov_pov_stat	estado_pobreza_nyc
off_pov_stat	estado_pobreza_oficial
pretaxincome_pu	ingresos_pre_impuestos

totalworkhrs_pu	tot_hrs_trabajo_pu

Gráfica 13 - Nuevas denominaciones de los datos en el dataset de pobreza

Una vez teniendo estos nuevos nombres se procedió a averiguar si los valores para la columna *edad* que son 0 eran debido al contexto o eran valores nulos. Para determinar esto se usó un resumen de datos agrupado, el cual arrojó como resultado que el dato en sí era una inconsistencia, ya que no existen personas de 0 años que tengan ingresos previos a impuestos por encima de 0 o un nivel de educación. Por ende se procedió a convertir todos los datos de la columna que eran 0 a nulos con Pyspark:

```
clean_spark_poverty=clean_spark_poverty.withColumn("edad",when(clean_spark_poverty["edad"]==0,None).otherwise(clean_spark_poverty["edad"]))
```

Posterior a esto resultamos con un nuevo reporte de nulos, el cual arroja las cantidades precisas de los datos que no tenemos a nuestra disposición dentro del dataset:

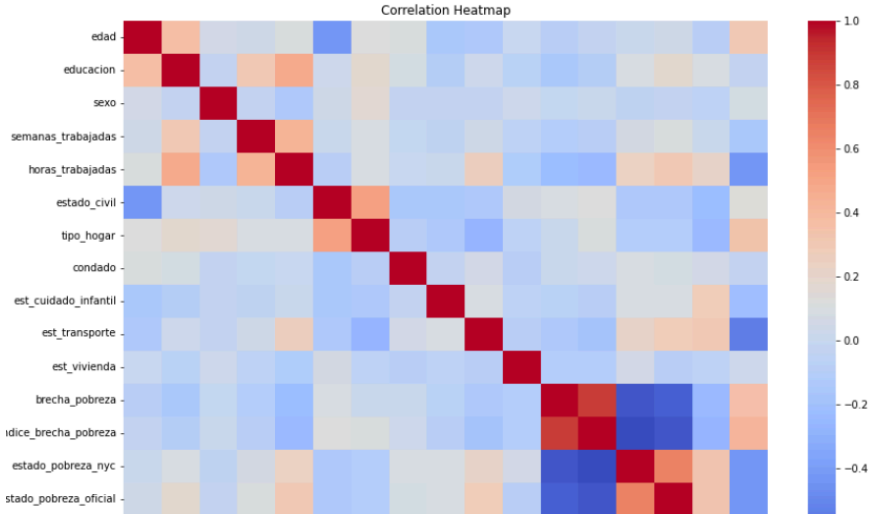
Variable	Valor
Edad	**5**
Educación	**24**
Sexo	0
Semanas trabajadas	**511**
Horas trabajadas	0
Estado civil	**181**
Tipo de hogar	0
Condado	0
Estado de cuidado infantil	0
Estado de transporte	0
Estado de vivienda	0

Brecha de pobreza	0
Índice de brecha de pobreza	0
Estado de pobreza NYC	0
Estado de pobreza oficial	0
Ingresos pre-impuestos	0
Total de horas trabajadas por persona en la unidad familiar	0

Gráfica 14 - Tabla datos nulos del dataset de pobreza en la ciudad de Nueva York.

Dadas estas cantidades se optó por los siguientes métodos para poder imputar o deshacerse de los valores faltantes de forma efectiva:

- Edad: Debido a su poca cantidad de registros que son nulos se optó por deshacerse de los registros, ya que no van a impactar de manera significativa
- Semanas trabajadas: Para calcular las semanas trabajadas nos basaremos en información que ya existe en el dataset y con contexto. Ya que usando la columna de *horas_trabajadas* y la reglamentación de la ciudad de Nueva York con relación al trabajo, podemos llegar a un dato cercano al real.
- Estado Civil: Debido a que este atributo es bastante variable y no existen patrones dentro de la población, se optó por usar el modelo de inteligencia artificial Kmean's con el fin de clusterizar o agrupar los datos según parámetros que sí existen y así poder predecir el valor faltante. Para poder definir qué columnas se usan para predecir se realizó el siguiente mapa de calor, el cual muestra que tan relacionadas están las variables del dataset, las más relacionadas son las usadas como parámetros.



Gráfica 15 - Mapa de calor - Correlación de Variables - Dataset pobreza en la ciudad de Nueva York.

- Educación: Para esta columna se usó primero el mismo procedimiento que para la columna de Estado Civil, sin embargo por bajas métricas en el desempeño del modelo se optó por usar la moda del grupo de personas con la misma edad que la del registro nulo que se está buscando.

Ya con estos procedimiento hechos, incluido evaluar el modelo de clustering con tres métricas principales (Silhouette Score, Calinski-Harabasz Index y Davies-Bouldin Index), logramos que el dataset quedará completamente limpio y estandarizado en el caso de la columna de semanas trabajadas.

Sección de Accidentes

Dado que este dataset según lo observado en el reporte de calidad de datos está bastante completo, como tal ninguna transformación o filtro muy grande fue requerido. Sin embargo, hay algunos registros que sí deben ser estandarizados, sobre todo teniendo en cuenta que estos se deben relacionar al otro dataset en alguna columna. Es por eso que los procesos que se aplicaron al dataset fueron los siguientes:

- One-hot-encoding: Para las variables categóricas que responden a las preguntas planteadas en el anterior numeral, se les aplicó una transformación de *one hot encoding* , donde cada una de las variables se vuelve una columna, para así de cierta forma discretizar la variable.
- Eliminación de la única columna que posee nulos, esto debido a que no es de interés para el análisis futuro y además es redundante

8. Bonos

Como objetivos extra a la realización de la primera entrega del proyecto, se muestra a continuación dos procedimientos para la extracción de datos:

- Realizar una ejecución de proceso web scraping para extraer la información de la población contenida en la tabla del siguiente vínculo y generar al menos un gráfico o tabla con dichos datos.
- Extraer información climática contenida en la siguiente API y generar al menos un gráfico o tabla con dichos datos.

Iniciando con el procedimiento de Web Scrapping, se utilizó para su cumplimiento la librería BeautifulSoup, la cual es especializada en este tipo de metodología para la extracción de información.

A través de requests, se realiza una petición GET a la página web, y dicha petición se decodifica a través de BeautifulSoup, para obtener los elementos del DOM que conforman la página, y en estos elementos, buscamos el elemento table, para obtener específicamente la información registrada en la tabla, presente en la página web.

Una vez obtenida la tabla, se deben buscar las filas, presentes en la tabla, debido a que no todos los elementos de la tabla son de interés, se identificó de forma manual cuales son los elementos a ignorar, en este caso, los títulos de cada columna en general, adicionalmente, se identificó que las localidades en varias ocasiones retornan cadenas vacías, por lo cual, se debe ajustar el codigo, para atender esta situación.

Finalmente, se identificó, que los elementos a tener en cuenta al interior del elemento tr, es decir, las respectivas filas de la tabla, son td y th, este último con un atributo en específico, el cual es el header, ya que dependiendo de su valor ‘b’ o ‘c’, puede representar una localidad o un barrio respectivamente.

Una vez obtenidos todos los elementos y extraída la información en listas locales, se procede a construir un dataframe de tipo pandas, a partir de dichas listas, para posteriormente transformarlo en un dataframe de tipo PySpark con la finalidad de mostrar la tabla con la información extraída como se muestra a continuación:

	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

Gráfica 16 - Tabla población de Nueva York - WebScrapping

Para el segundo bono, se utilizó la API proporcionada por la plataforma OpenWeather, realizando una consulta a través de requests, con la finalidad de obtener los datos, para este caso, se decidió realizar tres consultas a la API, y armar una tabla con base en la información obtenida.

En relación a lo anteriormente mencionado, se decidió utilizar los datos correspondientes a las ciudades de Nueva York, Bogotá y Madrid, utilizando sus respectivas coordenadas geográficas en formato decimal.

Debido a que se utiliza la versión gratuita de la API, las opciones están limitadas, y no permite ignorar algunos campos, por lo cual el JSON obtenido a través de la respuesta, contiene varios campos los cuales no son del interés del proyecto, para este caso, se extrajo, al interior de listas, la información de interés para la conformación de una tabla como, la temperatura, temperatura máxima, temperatura mínima, sensación térmica, presión y humedad.

Una vez extraída la información y almacenada en listas, se procede a crear un dataframe de pandas, con el propósito de agrupar la información creando registros veraces, para posteriormente transformar este dataframe, en un dataframe de tipo PySpark, el cual finalmente se muestra en forma de tabla como se muestra a continuación:

	Ciudad	Temperatura (C°)	Sensación termica (C°)	Temperatura minima (C°)
1	Nueva York	12.56	11.86	10.53
2	Bogotá	11.68	10.76	11.68
3	Madrid	17.89	16.73	15.68

Gráfica 17 - Tabla - Clima 10/Abril - 3 AM - API OpenWeather Forecast

9. Referencias

[1]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

10. Anexos

TABLA 1:

Arrestos en Nueva York		
Nombre	**Tipo de Datos**	**Descripción**
ARREST_KEY	Texto Plano	ID persistente generado aleatoriamente para cada arresto.
ARREST_DATE	Fecha y Hora	Fecha exacta del arresto para el evento

		reportado.
PD_CD	Número	Código de clasificación interna de tres dígitos (más granular que el código clave).
PD_DESC	Texto Plano	Descripción de la clasificación interna correspondiente al código PD (más granular que la descripción del delito).
KY_CD	Número	Código de clasificación interna de tres dígitos (categoría más general que el código PD).
OFNS_DESC	Texto Plano	Descripción de la clasificación interna correspondiente al código KY (categoría más general que la descripción PD).
LAW_CODE	Texto Plano	Códigos legales de cargos correspondientes a la Ley Penal de NYS, VTL y otras leyes locales diversas.
LAW_CAT_CD	Texto Plano	Nivel de la ofensa: delito grave, delito menor, violación.
ARREST_BORO	Texto Plano	Distrito de arresto. B (Bronx), S (Staten Island), K (Brooklyn), M (Manhattan), Q (Queens).
ARREST_PRECINCT	Número	Precinto donde ocurrió el arresto.
JURISDICTION_CODE	Número	Jurisdicción responsable del arresto. Los códigos de jurisdicción 0 (Patrulla), 1 (Tránsito) y 2 (Vivienda) representan a la policía de Nueva York (NYPD), mientras que los códigos 3 y superiores representan jurisdicciones que no pertenecen a la policía de Nueva York.
AGE_GROUP	Texto Plano	Edad del perpetrador dentro de una categoría.
PERP_SEX	Texto Plano	Descripción del sexo del perpetrador.

PERP_RACE	Texto Plano	Descripción de la raza del perpetrador.
X_COORD_CD	Número	Coordenada X de punto medio para el Sistema de Coordenadas del Plano del Estado de Nueva York, Zona de Long Island, NAD 83, unidades pies (FIPS 3104).
Y_COORD_CD	Número	Coordenada Y de punto medio para el Sistema de Coordenadas del Plano del Estado de Nueva York, Zona de Long Island, NAD 83, unidades pies (FIPS 3104).
Latitude	Número	Coordenada de latitud para el Sistema de Coordenadas Globales, WGS 1984, grados decimales (EPSG 4326).
Longitude	Número	Coordenada de longitud para el Sistema de Coordenadas Globales, WGS 1984, grados decimales (EPSG 4326).
New Georeferenced Column	Objeto Geocodificación (Punto)	Sin descripción.

TABLA 2:

Pobreza en Nueva York		
Nombre	**Tipo de Datos**	**Descripción**
serialno	Número	Número de serie o identificador del registro individual.
sporder	Número	Orden secuencial de la persona en el hogar.
pwgtp	Número	Peso de la persona.
wgtp	Número	Peso del hogar.
agep	Número	Edad de la persona.
cit	Texto	Estado de ciudadanía.
rel	Texto	Relación con el cabeza de familia.
sch	Texto	Asistencia a la escuela.
schg	Texto	Grado al que asiste.
Schl	Texto	Nivel de estudios.
sex	Texto	Sexo.
esr	Texto	Situación laboral.
lanx	Texto	Idioma hablado en casa.

msp	Texto	Estado civil.
mar	Texto	Historial matrimonial.
wkw	Número	Semanas trabajadas.
wkhp	Número	Horas trabajadas por semana.
dis	Texto	Situación de discapacidad.
jwtr	Texto	Medio de transporte al trabajo.
np	Número	Número de personas en el hogar.
ten	Texto	Tenencia (propietario/alquilador).
hht	Texto	Tipo de hogar.
agecateg	Texto	Categoría de edad.
boro	Texto	Municipio.
citizenstatus	Texto	Estado de nacionalidad.
educattain	Texto	Nivel de estudios.
est_childcare	Número	Gastos estimados de guardería.
est_commuting	Número	Gastos estimados de desplazamiento al trabajo.
est_eitc	Número	Estimación del EITC (crédito fiscal por ingresos del trabajo).
est_ficatax	Número	Estimación del impuesto FICA.
est_heap	Número	Estimación HEAP (Home Energy Assistance Program).
est_housing	Número	Gastos estimados de vivienda.

est_incometax	Número	Estimación del impuesto sobre la renta.
est_moop	Número	Estimación del MOOP (gasto máximo de bolsillo).
est_nutrition	Número	Gastos estimados de nutrición.
est_povgap	Número	Brecha de pobreza estimada.
est_povgapindex	Número	Índice estimado de la brecha de pobreza.
ethnicity	Texto	Etnia.
famtype_pu	Texto	Tipo de familia.
ftptwork	Texto	Situación laboral a tiempo completo o parcial.
intp_adj	Número	Ajuste de los ingresos por intereses.
mrgp_adj	Número	Ajuste del pago de la hipoteca.
nycgov_income	Número	Ingresos del gobierno de NYC.
nycgov_pov_stat	Texto	Estado de pobreza del gobierno de NYC.
nycgov_rel	Texto	Relación con el gobierno de NYC.
nycgov_threshold	Número	Umbral del gobierno de NYC.
Off_pov_stat	Texto	Estado oficial de pobreza.
off_threshold	Número	Umbral oficial de pobreza.
oi_adj	Número	Ajuste por otros ingresos.

pa_adj	Número	Ajuste de ingresos por asistencia pública.
Povunit_id	Número	ID de la unidad de pobreza.
povunit_rel	Texto	Relación con la unidad de pobreza.
pretaxincome_pu	Número	Renta antes de impuestos.
retp_adj	Número	Ajuste de ingresos por jubilación.
rntp_adj	Número	Ajuste de la renta de alquiler.
semp_adj	Número	Ajuste de los ingresos por trabajo autónomo.
ssip_adj	Número	Ajuste del Supplemental Security Income (SSI).
ssp_adj	Número	Ajuste de los ingresos de la Seguridad Social.
totalworkhrs_pu	Número	Total de horas de trabajo por unidad de pobreza.
wagp_adj	Número	Ajuste de los ingresos salariales.
eng	Texto	Dominio del inglés.