FACULTAD DE INGENIERIA

ASIGNATURA: Procesamiento de datos a gran escala

Proyecto

PROFESOR: John Corredor Franco

AUTORES: Alejandro Salamanca, Andrés Salamanca, Alberto Vigna

Pontificia Universidad Javeriana
Bogotá
3 de abril, 2024

I. Entendimiento del negocio

El estado de Nueva York, reconocido como uno de los epicentros culturales, económicos y políticos de los Estados Unidos, se encuentra en un continuo estado de transformación y progreso. Con una población diversa y dinámica, así como una economía multifacética que abarca desde las finanzas hasta la tecnología, Nueva York enfrenta una serie de desafíos sociales y económicos que demandan un análisis meticuloso y una intervención estratégica.

Nueva York, con su densidad poblacional y diversidad étnica, presenta una serie de desafíos únicos. Desde las bulliciosas calles de la ciudad de Nueva York hasta las comunidades rurales en el norte del estado, la diversidad geográfica y demográfica es impresionante. Este estado alberga una rica mezcla de culturas, tradiciones y perspectivas, lo que lo convierte en un crisol de ideas e innovación. Además, Nueva York es un centro de comercio internacional y un importante destino turístico, lo que influye en su dinámica económica y social.

Nueva York, un estado emblemático de los Estados Unidos, enfrenta desafíos significativos en términos de pobreza y desigualdad económica. A pesar de ser un centro económico vibrante con oportunidades de empleo en constante evolución, la economía diversa de la ciudad de Nueva York se enfrenta a problemas persistentes de pobreza y desigualdad.

El proceso de revitalización urbana, un fenómeno económico y social en aumento, ha transformado muchos vecindarios de Nueva York, con un crecimiento económico y comercial significativo, pero con beneficios que no siempre se distribuyen equitativamente. Aunque la revitalización urbana ha impulsado el crecimiento del empleo en varios vecindarios, las oportunidades laborales rara vez llegan a los residentes locales, especialmente a las comunidades de color.

La pandemia de COVID-19 exacerbó aún más los desafíos económicos en Nueva York, con pérdidas masivas de empleo y un aumento en la inseguridad económica. Aunque la ciudad cuenta con una red de seguridad social robusta, las tasas de pobreza persisten, superando el promedio nacional y afectando desproporcionadamente a ciertos grupos, como niños, mujeres jefas de hogar y personas de color.

El estado ha reconocido la urgencia de abordar estos problemas y se ha comprometido a reducir la pobreza infantil a la mitad. Sin embargo, se necesitan enfoques más equitativos y basados en evidencia para lograr este objetivo. El análisis detallado de los perfiles económicos de los vecindarios, junto con un entendimiento de las dinámicas de la revitalización urbana y las implicaciones de la pandemia, proporciona una base para desarrollar estrategias efectivas que aborden la pobreza y promuevan un crecimiento económico más inclusivo en Nueva York.

El estado de Nueva York enfrenta desafíos significativos en términos de criminalidad y el sistema de justicia penal. Un estudio exhaustivo de las condenas penales y las disparidades raciales asociadas desde 1980 hasta 2021 revela una serie de tendencias preocupantes.

Desde 1980 hasta 2021, más de 6.6 millones de casos penales en Nueva York terminaron en condenas. Si bien hubo un aumento en las condenas durante la década de 1980, desde entonces ha habido fluctuaciones, con una notable disminución en 2019, seguida de un descenso continuo en 2020 y 2021, probablemente debido a la pandemia de COVID-19.

Más del 77% de las condenas fueron por delitos menores, con el restante 23% por delitos graves. La ciudad de Nueva York representó la mayoría de las condenas en el estado, aunque su participación ha disminuido con el tiempo, mientras que los condados suburbanos y del norte del estado han visto un aumento en su proporción de condenas.

Las disparidades raciales son evidentes en las condenas. A pesar de que las personas negras representan solo el 15% de la población del estado en 2019, representaron el 42% de las condenas desde 1980 hasta 2021. Las tasas de condena para personas negras son significativamente más altas que para personas blancas, con una disparidad que persiste a lo largo del tiempo.

Estos datos resaltan la necesidad de abordar las disparidades raciales y socioeconómicas en el sistema de justicia penal de Nueva York, así como la importancia de comprender las tendencias en la criminalidad para informar políticas efectivas de prevención del delito y justicia penal.

Indicadores macroeconómicos:

1. Población (2020):

- Nueva York (ciudad): 8,804,000
- Estado de Nueva York: 20,201,000
- Estados Unidos: 331,449,000

2. **Crecimiento de la Población (2020, cambio de 10 años):**

- Nueva York (ciudad): 7.7%
- Estado de Nueva York: 4.2%
- Estados Unidos: 7.4%

3. **Producto Interno Bruto (GDP) (2020, miles de millones de dólares):**

- Nueva York (ciudad): $1,022
- Estado de Nueva York: $1,725
- Estados Unidos: $20,894

4. **Crecimiento del Empleo (2021, cambio de 5 años):**

- Nueva York (ciudad): -3.0%
- Estado de Nueva York: -4.1%
- Estados Unidos: 1.2%

5. **Ingreso Medio por Hogar (2020):**

- Nueva York (ciudad): $70,000
- Estado de Nueva York: $72,600
- Estados Unidos: No proporcionado

6. **Precio Medio de Vivienda (2020):**

- Nueva York (ciudad): $650,000
- Estado de Nueva York: $350,000
- Estados Unidos: No proporcionado

7. **Porcentaje de la Población de 25 años o más con Licenciatura o Superior (2020):**

- Nueva York (ciudad): 40.3%
- Estado de Nueva York: 39.4%
- Estados Unidos: 35.1%

Estos indicadores proporcionan una visión general de la demografía, la economía y la educación en el estado de Nueva York y su ciudad más grande, Nueva York.

Objetivo:

El objetivo del negocio en este contexto es encontrar relaciones significativas entre la pobreza, la criminalidad y el nivel educativo en todo el estado de Nueva York. A través de un análisis exhaustivo de estos factores socioeconómicos, se busca comprender las complejas interacciones que influyen en los patrones de arresto y la actividad delictiva en diversas comunidades y grupos demográficos dentro del estado.

Este análisis abarca la recopilación y el análisis de datos sobre la distribución geográfica y demográfica de los arrestos en Nueva York, con el fin de identificar tendencias clave y factores de riesgo asociados con la criminalidad. Además, se investiga la relación entre la pobreza y la

delincuencia, examinando la desigualdad económica, el acceso a oportunidades laborales y la disponibilidad de recursos comunitarios en diferentes áreas urbanas y rurales del estado.

Asimismo, se lleva a cabo un estudio detallado sobre el nivel educativo de la población y su impacto en la criminalidad en todo el estado. Se busca identificar posibles vínculos entre el nivel de educación, las oportunidades laborales y el comportamiento delictivo, con el objetivo de determinar cómo mejorar el acceso a la educación puede contribuir a la reducción de la delincuencia en la región.

Los hallazgos y conclusiones de este análisis se utilizarán para informar el desarrollo de políticas y estrategias efectivas de prevención del delito, intervención comunitaria y aplicación de la ley en Nueva York. El objetivo final es mejorar la seguridad pública, reducir la delincuencia y promover el bienestar en todas las comunidades del estado.

II. Selección de los datos

Según los objetivos del proyecto principalmente se utilizarán los datasets de:

- Datos de los arrestos en Nueva York
- Datos de pobreza en Nueva York
- Datos de quejas de la policía de Nueva York actuales

III. Colección y descripción de los datos

Descripción set de datos arrestos en Nuevo York:

a. Tipo de datos:

- **arrest_key**: Un identificador único de tipo entero largo que representa la clave de arresto.

- **arrest_date**: Una cadena de texto que representa la fecha del arresto.

- **pd_cd**: Un número entero largo que indica el código del departamento de policía.

- **pd_desc**: Una cadena de texto que describe la descripción del departamento de policía.

- **ky_cd**: Un número entero largo que indica el código de clave.

- **ofns_desc**: Una cadena de texto que describe la descripción del delito.

- **law_code**: Una cadena de texto que representa el código de la ley.

- **law_cat_cd**: Una cadena de texto que indica la categoría de la ley.

- **arrest_boro**: Una cadena de texto que indica el distrito de arresto.

- **arrest_precinct**: Un número entero largo que indica el precinto de arresto.

- **jurisdiction_code**: Un número entero largo que indica el código de jurisdicción.

- **age_group**: Una cadena de texto que indica el grupo de edad del perpetrador.

- **perp_sex**: Una cadena de texto que indica el sexo del perpetrador.

- **perp_race**: Una cadena de texto que indica la raza del perpetrador.

- **x_coord_cd**: Un número entero largo que indica la coordenada X.

- **y_coord_cd**: Un número entero largo que indica la coordenada Y.

- **latitude**: Un número decimal de doble precisión que indica la latitud.

- **longitude**: Un número decimal de doble precisión que indica la longitud.

- **geocoded_column**: Una cadena de texto que representa la columna geocodificada.

b. Comprensión de cada atributo:

- ARREST_KEY: Es un identificador único y persistente generado aleatoriamente para cada arresto. Se utiliza para identificar de manera exclusiva cada registro de arresto en el conjunto de datos.

- ARREST_DATE: Representa la fecha y hora exactas en que ocurrió el arresto. Esta información es crucial para el análisis temporal de los datos y la identificación de patrones estacionales o tendencias a lo largo del tiempo.

- PD_CD: Es un código numérico que clasifica internamente el tipo específico de delito. Proporciona una categorización detallada del delito cometido en cada arresto, permitiendo un análisis más granular de los tipos de delitos involucrados.

- PD_DESC: Describe en detalle la naturaleza del delito correspondiente al código PD. Proporciona información sobre la descripción específica del delito cometido en cada arresto, lo que facilita la comprensión de los tipos de actividades delictivas que se están investigando.

- KY_CD: Similar al PD_CD, el KY_CD es un código numérico que clasifica internamente el delito, pero en una categoría más general. Proporciona una clasificación menos detallada pero más amplia de los tipos de delitos involucrados en cada arresto.

- OFNS_DESC: Describe en términos más generales la naturaleza del delito correspondiente al código KY. Proporciona una descripción más amplia de los tipos de delitos cometidos en cada arresto, lo que ayuda a contextualizar los datos en un nivel más alto.

- LAW_CODE: Representa el código de la ley bajo la cual se realizó el arresto, incluyendo las leyes penales del estado de Nueva York, VTL y otras leyes locales. Esta información es importante para comprender la base legal del arresto y las disposiciones legales específicas relacionadas con cada caso.

- LAW_CAT_CD: Indica el nivel de gravedad del delito, clasificándolo como delito grave, delito menor o violación. Esta clasificación proporciona información sobre la gravedad percibida de cada delito y puede ser útil para el análisis de tendencias en la prevalencia de diferentes tipos de delitos.

- ARREST_BORO: Identifica el distrito de la ciudad de Nueva York donde ocurrió el arresto, utilizando códigos de una sola letra para cada distrito. Esta información es importante para analizar la distribución geográfica de los arrestos dentro de la ciudad y puede revelar patrones de actividad delictiva en áreas específicas.

- ARREST_PRECINCT: Indica el precinto policial donde ocurrió el arresto. Esta información es crucial para el análisis de la distribución espacial de los arrestos y para entender cómo la actividad delictiva se distribuye en diferentes áreas de la ciudad.

- JURISDICTION_CODE: Identifica la jurisdicción responsable del arresto, con códigos numéricos que representan diferentes entidades policiales. Esta información es útil para distinguir entre arrestos realizados por el Departamento de Policía de Nueva York (NYPD) y otras jurisdicciones policiales en el estado.

- AGE_GROUP: Clasifica la edad del perpetrador en categorías predefinidas, lo que proporciona una visión general de la distribución de edades de los arrestados. Esta información es importante para comprender las características demográficas de los involucrados en la actividad delictiva.

- PERP_SEX: Indica el sexo del perpetrador, proporcionando información sobre la composición de género de los arrestados. Esta variable es importante para el análisis de género en relación con la actividad delictiva y puede revelar patrones de comportamiento delictivo diferencial entre hombres y mujeres.

- PERP_RACE: Describe la raza o etnia del perpetrador, permitiendo un análisis de las disparidades raciales en la actividad delictiva. Esta variable es crucial para comprender cómo las disparidades raciales pueden influir en la interacción con la aplicación de la ley y en la prevalencia de diferentes tipos de delitos entre diferentes grupos raciales o étnicos.

- X_COORD_CD y Y_COORD_CD: Representan las coordenadas X e Y del lugar donde ocurrió el arresto, en el sistema de coordenadas de Nueva York, Zona de Long Island. Estas coordenadas son importantes para el análisis espacial de los datos y la representación visual de la distribución geográfica de los arrestos en el estado.

- Latitude y Longitude: Representan la latitud y longitud del lugar donde ocurrió el arresto, en el sistema de coordenadas global WGS 1984. Estas coordenadas son cruciales para la georreferenciación de los datos y su visualización en mapas interactivos.

- New Georeferenced Column: Esta columna parece ser una nueva columna que contiene información georreferenciada adicional, pero no se proporciona una descripción detallada de su contenido en la documentación proporcionada. Sería necesario investigar más a fondo esta columna para comprender su significado y utilidad en el análisis de los datos de arrestos.

c. Descripción general:

El conjunto de datos proporciona una visión detallada de los arrestos en el estado de Nueva York, ofreciendo información sobre una amplia gama de incidentes. Cada entrada en el conjunto de datos representa un arresto específico y está identificado por un número único generado aleatoriamente. Además de la fecha y hora exactas en que ocurrió cada arresto, el conjunto de datos incluye detalles sobre la naturaleza del delito, proporcionando descripciones tanto detalladas como generales de los tipos de actividades delictivas. Estos datos son esenciales para comprender la distribución y gravedad de los delitos cometidos en la región, lo que puede ayudar a informar estrategias de aplicación de la ley y políticas de prevención del delito.

La información sobre la ubicación de cada arresto también está disponible en el conjunto de datos, incluyendo el distrito de la ciudad de Nueva York y el precinto policial donde ocurrió el incidente. Esta información geográfica permite un análisis detallado de la distribución espacial de la actividad delictiva en la región, lo que puede ayudar a identificar áreas de alto y bajo delito y orientar los recursos de aplicación de la ley de manera más efectiva.

Además de los detalles sobre el delito y la ubicación, el conjunto de datos también incluye información demográfica sobre los perpetradores de los arrestos, como su grupo de edad, sexo y raza o etnia. Estos datos demográficos son importantes para comprender las características de los individuos involucrados en la actividad delictiva y pueden ayudar a identificar posibles disparidades en el sistema de justicia penal. Al analizar estas variables demográficas en conjunto con los detalles del delito, los investigadores pueden obtener una comprensión más completa de los factores que contribuyen a la actividad delictiva y las interacciones entre diferentes grupos de personas y el sistema de justicia penal.

Descripción set de datos pobreza en Nuevo York:

a. Tipo de datos:

- serialno: Número de serie único asignado a cada individuo en la encuesta. (Long)

- sporder: Orden de aparición de la persona en la muestra de personas en la encuesta. (Long)

- pwgtp: Peso de la persona para generar estimaciones poblacionales. (Long)

- wgtp: Peso de la vivienda para generar estimaciones poblacionales. (Long)

- agep: Edad de la persona. (Long)

- cit: Ciudadanía de la persona. (Long)

- rel: Relación con el jefe de familia. (Long)

- sch: Asistencia escolar. (Long)

- schg: Nivel de grado o escolaridad. (Long)

- schl: Nivel de educación alcanzado. (Double)

- sex: Sexo de la persona. (Long)
- esr: Estado laboral. (Double)
- lanx: Idioma hablado en casa. (Double)
- eng: Dominio del idioma inglés. (Double)
- msp: Estado civil. (Double)
- mar: Estado civil. (Long)
- wkw: Semanas trabajadas en el último año. (Double)
- wkhp: Horas trabajadas por semana. (Long)
- dis: Limitación física o mental. (Long)
- jwtr: Medio de transporte al trabajo. (Double)
- np: Número de personas en la vivienda. (Long)
- ten: Tenencia de la vivienda. (Long)
- hht: Tipo de vivienda. (Long)
- agecateg: Categoría de edad. (Long)
- boro: Distrito municipal. (Long)
- citizenstatus: Estado de ciudadanía. (Long)
- educattain: Nivel de educación alcanzado. (Double)
- est_childcare: Estimación de los gastos de cuidado infantil. (Double)
- est_commuting: Estimación de los gastos de viaje al trabajo. (Double)
- est_eitc: Estimación de los créditos fiscales por ingresos del trabajo. (Double)
- est_ficatax: Estimación de los impuestos federales al ingreso. (Double)
- est_heap: Estimación de la asistencia energética para hogares de bajos ingresos. (Long)
- est_housing: Estimación del costo de vivienda. (Double)
- est_incometax: Estimación de los impuestos al ingreso. (Double)
- est_moop: Estimación del ingreso moderado para hogares de 1 a 4 personas. (Double)
- est_nutrition: Estimación del costo de una dieta nutritiva. (Double)
- est_povgap: Estimación del número de personas por debajo del umbral de pobreza. (Double)
- est_povgapindex: Índice de brecha de pobreza. (Double)

- ethnicity: Etnicidad. (Long)

- famtype_pu: Tipo de familia en unidades de pobreza. (Long)

- ftptwork: Estado laboral (trabajo a tiempo parcial o completo). (Long)

- intp_adj: Ingreso ajustado. (Double)

- mrgp_adj: Ingreso ajustado para casados. (Double)

- nycgov_income: Ingreso según el gobierno de la ciudad de Nueva York. (Double)

- nycgov_pov_stat: Estado de pobreza según el gobierno de la ciudad de Nueva York. (Long)

- nycgov_rel: Relación con el jefe de familia según el gobierno de la ciudad de Nueva York. (Long)

- nycgov_threshold: Umbral de pobreza según el gobierno de la ciudad de Nueva York. (Double)

- off_pov_stat: Estado de pobreza fuera de los programas de gobierno. (Long)

- off_threshold: Umbral de pobreza fuera de los programas de gobierno. (Long)

- oi_adj: Ingreso ajustado fuera de los programas de gobierno. (Double)

- pa_adj: Ingreso ajustado parcial. (Double)

- povunit_id: Identificación de la unidad de pobreza. (Long)

- povunit_rel: Relación con la unidad de pobreza. (Long)

- pretaxincome_pu: Ingreso antes de impuestos en unidades de pobreza. (Double)

- retp_adj: Ingreso ajustado para jubilación. (Double)

- rntp_adj: Ingreso ajustado para alquiler. (Double)

- semp_adj: Ingreso ajustado para empleo propio. (Double)

- ssip_adj: Ingreso ajustado para seguridad social suplementaria. (Double)

- ssp_adj: Ingreso ajustado para seguridad social. (Double)

- totalworkhrs_pu: Total de horas trabajadas en unidades de pobreza. (Long)

- wagp_adj: Ingreso ajustado por salario. (Double)

b. Comprensión de cada atributo:

- AgeCateg: Categoría de edad que indica si la persona tiene menos de 18 años, entre 18 y 64 años o 65 años o más.

- Boro: Distrito municipal de la ciudad de Nueva York, que puede ser Bronx, Brooklyn, Manhattan, Queens o Staten Island.

- CitizenStatus: Estado de ciudadanía que especifica si la persona es ciudadana por nacimiento, ciudadana naturalizada o no ciudadana.

- EducAttain: Nivel educativo alcanzado, que varía desde menos de secundaria hasta grado universitario o superior.

- EST_HousingStatus: Estimación del tipo de vivienda ocupada por el hogar, que incluye categorías como alquiler público, alquiler regulado y propiedad libre de hipoteca.

- EST_Childcare: Costos estimados de cuidado infantil proporcionados por el gobierno de la ciudad de Nueva York.

- EST_Commuting: Costos estimados de desplazamiento proporcionados por el gobierno de la ciudad de Nueva York.

- EST_FICAtax: Impuestos estimados de FICA (nómina) proporcionados por el gobierno de la ciudad de Nueva York.

- EST_HEAP: Asistencia estimada para calefacción proporcionada por el gobierno de la ciudad de Nueva York.

- EST_Housing: Ajuste estimado para subsidio de vivienda proporcionado por el gobierno de la ciudad de Nueva York.

- EST_IncomeTax: Impuestos estimados sobre el ingreso neto proporcionados por el gobierno de la ciudad de Nueva York.

- EST_MOOP: Gasto médico total estimado proporcionado por el gobierno de la ciudad de Nueva York.

- EST_PovGap: Diferencia estimada en dólares entre el ingreso familiar y el umbral de pobreza proporcionada por el gobierno de la ciudad de Nueva York.

- EST_PovGapIndex: La brecha de pobreza expresada como una proporción del umbral de pobreza proporcionada por el gobierno de la ciudad de Nueva York.

- Ethnicity: Raza/etnia de la persona.

- FamType_PU: Estructura familiar de la unidad de pobreza.

- FTPTWork: Experiencia laboral de la persona, ya sea a tiempo completo, parcial o sin trabajo.

- NYCgov_Income: Ingreso total estimado proporcionado por el gobierno de la ciudad de Nueva York.

- NYCgov_Pov_Stat: Estado de pobreza proporcionado por el gobierno de la ciudad de Nueva York.

- NYCgov_REL: Relaciones dentro del hogar según el gobierno de la ciudad de Nueva York.

- NYCgov_Threshold: Umbral de pobreza proporcionado por el gobierno de la ciudad de Nueva York.

- Off_Pov_Stat: Estado de pobreza oficial/federal.

- Off_Threshold: Umbral de pobreza oficial/federal.

- Povunit_ID: Identificador de unidades de pobreza dentro del hogar.

- Povunit_Rel: Relación dentro de la unidad de pobreza.

- PreTaxIncome_PU: Ingreso antes de impuestos en unidades de pobreza.

- TotalWorkHrs_PU: Horas trabajadas totales anuales por miembros de la unidad de pobreza.

- AGEP: Edad de la persona.

- CIT: Estado de ciudadanía.

- DIS: Recodificación de discapacidad.

- DS: Recodificación de discapacidad.

- ENG: Habilidad para hablar inglés.

- ESR: Recodificación del estado laboral.

- HHT: Tipo de hogar/familia.

- INTP_adj: Intereses, dividendos y alquiler neto en los últimos 12 meses, ajustado por factor de ajuste de ingresos.

- JWTR: Medio de transporte al trabajo.

- LANX: Idioma que se habla en casa además del inglés.

- MAR: Estado civil.

- MRGP_adj: Pago de hipoteca mensual, ajustado por factor de ajuste de ingresos.

- MSP: Estado civil con cónyuge presente/ausente.

- NP: Número de personas en la unidad de vivienda.

- OI_adj: Otros ingresos en los últimos 12 meses, ajustados por factor de ajuste de ingresos.

- PA_adj: Ingresos de asistencia pública en los últimos 12 meses, ajustados por factor de ajuste de ingresos.

- PWGTP: Peso de la persona.

- REL: Relación con la persona de referencia.

- RETP_adj: Ingresos por jubilación en los últimos 12 meses, ajustados por factor de ajuste de ingresos.

- RNTP_adj: Alquiler mensual, ajustado por factor de ajuste de ingresos.

- SCH: Inscripción escolar.

- SCHG: Nivel de grado que está cursando.

- SCHL: Logro educativo.

- SEMP_adj: Ingresos por trabajo por cuenta propia en los últimos 12 meses, ajustados por factor de ajuste de ingresos.

- SERIALNO: Número de serie del hogar del censo.

- SEX: Sexo.

- SPORDER: Número de cada persona en el hogar del censo.

- SSIP_adj: Ingresos suplementarios de seguridad en los últimos 12 meses, ajustados por factor de ajuste de ingresos.

- SSP_adj: Ingresos de seguridad social en los últimos 12 meses, ajustados por factor de ajuste de ingresos.

- TEN: Tenencia de la vivienda.

- WAGP_adj: Ingresos por salario en los últimos 12 meses, ajustados por factor de ajuste de ingresos.

- WGTP: Peso de la vivienda.

- WKHP: Horas trabajadas habituales por semana en los últimos 12 meses.

- WKW: Semanas trabajadas en los últimos 12 meses.

- WKWN: Semanas en las que trabajó en los últimos 12 meses.

c. Descripción general:

El conjunto de datos proporciona una amplia gama de información socioeconómica y demográfica sobre los residentes del estado de Nueva York. Incluye atributos como edad, género, estado de ciudadanía, nivel educativo, ingresos estimados, estado de pobreza, estructura familiar, estado laboral, discapacidad, idiomas hablados en casa, entre otros. Estos datos están organizados en una estructura tabular con cada fila representando a un individuo dentro de un hogar censal, identificado por un número de serie único. Los atributos cubren aspectos como la composición demográfica de los hogares, los niveles de ingresos, el estatus de ciudadanía, la situación laboral y educativa, así como los costos estimados de vivienda, transporte y otros gastos. Este conjunto de datos proporciona una visión detallada de la población del estado de

Nueva York y puede ser utilizado para realizar análisis socioeconómicos, estudios demográficos y para informar políticas públicas orientadas a la equidad y el bienestar social.

Descripción Datos de quejas de la policía de Nueva York actuales:

a. Tipo de datos:

Contiene información sobre todos los delitos reportados a la Policía de la Ciudad de Nueva York (NYPD). Aquí está una explicación de las columnas presentes en el conjunto de datos:

- **CMPLNT_NUM**: Es un ID generado aleatoriamente para cada queja.
- **ADDR_PCT_CD**: Es el código del precinto en el que ocurrió el incidente.
- **BORO_NM**: Es el nombre del distrito en el que ocurrió el incidente.
- **CMPLNT_FR_DT**: Es la fecha exacta del evento reportado.
- **CMPLNT_FR_TM**: Es la hora exacta del evento reportado.
- **CMPLNT_TO_DT**: Es la fecha de finalización del evento reportado, si se conoce.
- **CMPLNT_TO_TM**: Es la hora de finalización del evento reportado, si se conoce.
- **CRM_ATPT_CPTD_CD**: Indica si el crimen se completó con éxito, o si se intentó pero falló o fue interrumpido prematuramente.
- **HADEVELOPT**: Es el nombre del desarrollo de viviendas de NYCHA donde ocurrió el incidente, si corresponde.
- **HOUSING_PSA**: Código de nivel de desarrollo.
- **JURISDICTION_CODE**: Jurisdicción responsable del incidente.
- **JURIS_DESC**: Descripción de la jurisdicción.
- **KY_CD**: Código de clasificación del delito de tres dígitos.
- **LAW_CAT_CD**: Nivel del delito: delito grave, delito menor, violación.
- **LOC_OF_OCCUR_DESC**: Ubicación específica del incidente.
- **OFNS_DESC**: Descripción del delito correspondiente con el código clave.
- **PARKS_NM**: Nombre del parque, área de recreo o espacio verde de la ciudad de Nueva York donde ocurrió el incidente, si corresponde.
- **PATROL_BORO**: Nombre del distrito de patrulla en el que ocurrió el incidente.
- **PD_CD**: Código de clasificación interna de tres dígitos.
- **PD_DESC**: Descripción de la clasificación interna correspondiente con el código PD.
- **PREM_TYP_DESC**: Descripción específica de las instalaciones.
- **RPT_DT**: Fecha en que se informó el evento a la policía.
- **STATION_NAME**: Nombre de la estación de tránsito.
- **SUSP_AGE_GROUP**: Grupo de edad del sospechoso.
- **SUSP_RACE**: Descripción de la raza del sospechoso.
- **SUSP_SEX**: Descripción del sexo del sospechoso.
- **TRANSIT_DISTRICT**: Distrito de tránsito en el que ocurrió el delito.
- **VIC_AGE_GROUP**: Grupo de edad de la víctima.
- **VIC_RACE**: Descripción de la raza de la víctima.

- **VIC_SEX**: Descripción del sexo de la víctima.
- **X_COORD_CD**: Coordenada X para el Sistema de Coordenadas del Plano Estatal de Nueva York, Zona de Long Island, NAD 83, unidades de pies.
- **Y_COORD_CD**: Coordenada Y para el Sistema de Coordenadas del Plano Estatal de Nueva York, Zona de Long Island, NAD 83, unidades de pies.
- **Latitude**: Coordenada de latitud para el Sistema de Coordenadas Globales, WGS 1984, grados decimales.
- **Longitude**: Coordenada de longitud para el Sistema de Coordenadas Globales, WGS 1984, grados decimales.
- **Lat_Lon**: Ubicación.

b. Comprensión de cada atributo:

- **CMPLNT_NUM (Número de queja)**: Es un ID único generado aleatoriamente para cada queja reportada a la Policía de la Ciudad de Nueva York (NYPD). Proporciona una identificación persistente para cada incidente registrado.
- **ADDR_PCT_CD (Código de Precinto)**: Es el código numérico del precinto policial en el que ocurrió el incidente. Los precintos son las divisiones territoriales de la policía que cubren áreas específicas de la ciudad.
- **BORO_NM (Nombre de Distrito)**: Es el nombre del distrito o borough en el que ocurrió el incidente. Nueva York tiene cinco distritos: Manhattan, Brooklyn, Queens, Bronx y Staten Island.
- **CMPLNT_FR_DT (Fecha de Inicio del Evento)**: Es la fecha exacta en la que ocurrió el evento reportado.
- **CMPLNT_FR_TM (Hora de Inicio del Evento)**: Es la hora exacta en la que ocurrió el evento reportado.
- **CMPLNT_TO_DT (Fecha de Finalización del Evento)**: Es la fecha de finalización del evento reportado, si se conoce. Esto se registra si el evento se extendió por un período de tiempo.
- **CMPLNT_TO_TM (Hora de Finalización del Evento)**: Es la hora de finalización del evento reportado, si se conoce.
- **CRM_ATPT_CPTD_CD (Estado del Crimen)**: Indica si el crimen fue completado con éxito ("COMPLETED") o si fue intentado pero falló o fue interrumpido prematuramente ("ATTEMPTED").
- **HADEVELOPT (Desarrollo de Viviendas NYCHA)**: Es el nombre del desarrollo de viviendas del New York City Housing Authority (NYCHA) donde ocurrió el incidente, si es aplicable.
- **HOUSING_PSA (Código de Nivel de Desarrollo)**: Es un código numérico que representa el nivel de desarrollo de viviendas.

- **JURISDICTION_CODE (Código de Jurisdicción)**: Indica la jurisdicción responsable del incidente, ya sea interna (como Policía, Tránsito y Vivienda) o externa (como Correccionales, Autoridad Portuaria, etc.).
- **JURIS_DESC (Descripción de Jurisdicción)**: Proporciona una descripción de la jurisdicción indicada por el código de jurisdicción.
- **KY_CD (Código de Clasificación de Delito)**: Es un código numérico de tres dígitos que clasifica el tipo de delito reportado.
- **LAW_CAT_CD (Nivel de Delito)**: Indica el nivel de gravedad del delito, categorizándolo como "felony" (delito grave), "misdemeanor" (delito menor) o "violation" (infracción).
- **LOC_OF_OCCUR_DESC (Descripción de Ubicación del Incidente)**: Proporciona una descripción específica de la ubicación del incidente, como "dentro", "frente a", "detrás de", etc.
- **OFNS_DESC (Descripción del Delito)**: Proporciona una descripción detallada del delito correspondiente al código de clasificación de delito.
- **PARKS_NM (Nombre del Parque)**: Indica el nombre del parque, área de recreo o espacio verde de la Ciudad de Nueva York donde ocurrió el incidente, si es aplicable.
- **PATROL_BORO (Distrito de Patrulla)**: Es el nombre del distrito de patrulla en el que ocurrió el incidente.
- **PD_CD (Código de Clasificación Interna)**: Es un código numérico de tres dígitos que proporciona una clasificación interna más detallada que el código de clasificación de delito.
- **PD_DESC (Descripción de Clasificación Interna)**: Proporciona una descripción detallada de la clasificación interna correspondiente al código de clasificación interna.
- **PREM_TYP_DESC (Descripción de Tipo de Instalación)**: Ofrece una descripción específica de las instalaciones donde ocurrió el incidente, como "tienda de comestibles", "residencia", "calle", etc.
- **RPT_DT (Fecha de Reporte)**: Es la fecha en que el incidente fue reportado a la policía.
- **STATION_NAME (Nombre de la Estación de Tránsito)**: Indica el nombre de la estación de tránsito más cercana al lugar donde ocurrió el incidente.
- **SUSP_AGE_GROUP (Grupo de Edad del Sospechoso)**: Describe el grupo de edad del sospechoso.
- **SUSP_RACE (Raza del Sospechoso)**: Proporciona una descripción de la raza del sospechoso.
- **SUSP_SEX (Sexo del Sospechoso)**: Indica el sexo del sospechoso.
- **TRANSIT_DISTRICT (Distrito de Tránsito)**: Es el distrito de tránsito en el que ocurrió el delito.
- **VIC_AGE_GROUP (Grupo de Edad de la Víctima)**: Describe el grupo de edad de la víctima.
- **VIC_RACE (Raza de la Víctima)**: Proporciona una descripción de la raza de la víctima.

- **VIC_SEX (Sexo de la Víctima)**: Indica el sexo de la víctima.
- **X_COORD_CD (Coordenada X)**: Es la coordenada X para el Sistema de Coordenadas del Plano Estatal de Nueva York, Zona de Long Island.
- **Y_COORD_CD (Coordenada Y)**: Es la coordenada Y para el Sistema de Coordenadas del Plano Estatal de Nueva York, Zona de Long Island.
- **Latitude (Latitud)**: Es la coordenada de latitud para el Sistema de Coordenadas Globales, WGS 1984.
- **Longitude (Longitud)**: Es la coordenada de longitud para el Sistema de Coordenadas Globales, WGS 1984.
- **Lat_Lon (Ubicación)**: Proporciona la ubicación precisa del incidente en formato de coordenadas latitud-longitud.

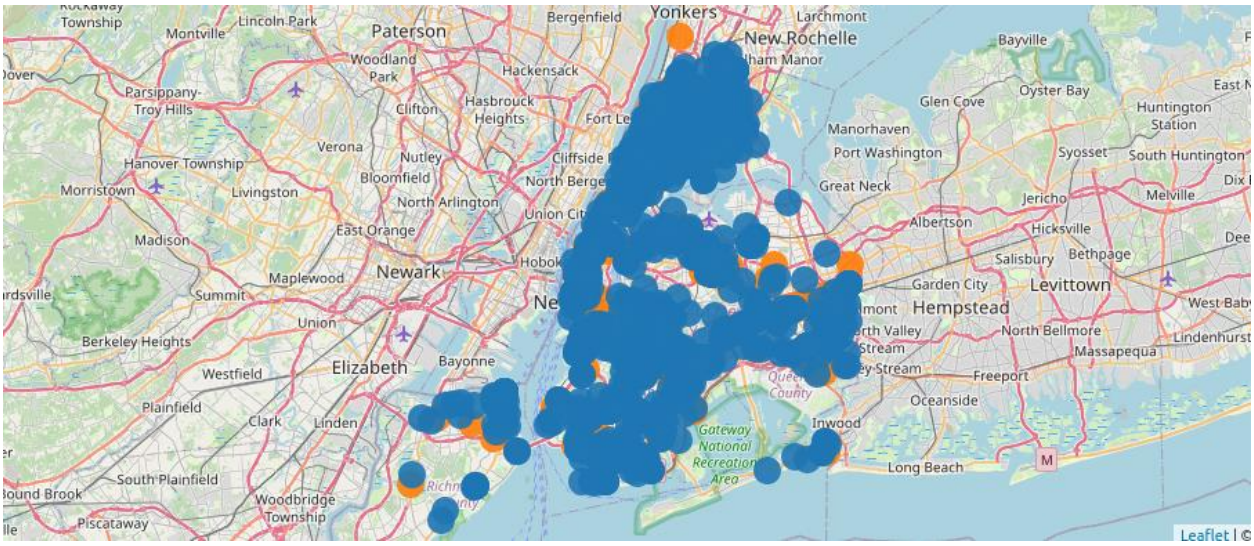
c. Descripción general:

El conjunto de datos "NYPD Complaint Data Current (Year To Date)" contiene información detallada sobre los delitos reportados a la Policía de la Ciudad de Nueva York (NYPD) durante el año en curso (2019). Incluye una amplia gama de variables que describen cada incidente, como la fecha y hora de ocurrencia, la ubicación específica, la naturaleza del delito, el nivel de gravedad, detalles sobre los sospechosos y víctimas, así como información geoespacial. Con más de 30 columnas, este conjunto de datos proporciona una visión completa de la actividad delictiva en la ciudad, permitiendo análisis detallados sobre patrones de criminalidad, distribución geográfica de los delitos y características demográficas de los involucrados. Esta información puede ser utilizada por analistas de seguridad pública, investigadores criminales y formuladores de políticas para comprender mejor la dinámica del crimen y desarrollar estrategias efectivas de prevención y aplicación de la ley.

IV. Exploración de los datos

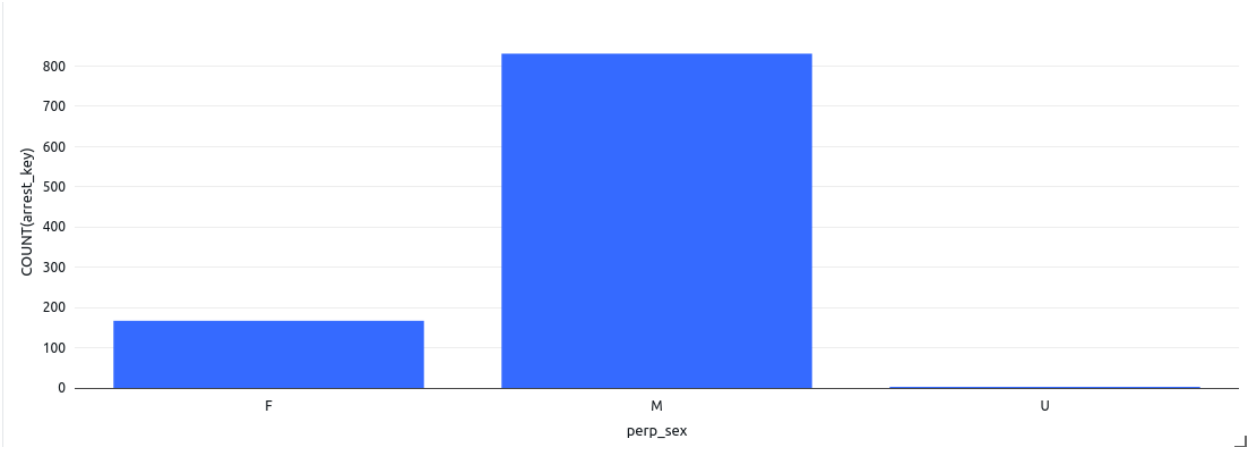
Exploración de arrestos en Nuevo York:

Mapa de los arrestos según longitud y latitud:



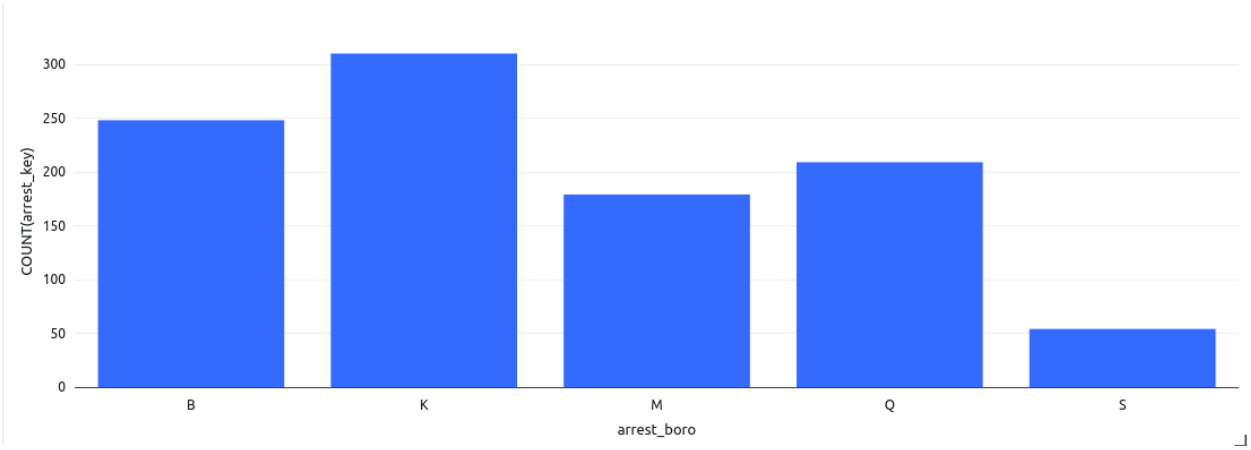
Esta visualización muestra la distribución geográfica de los arrestos en el área de interés. Es útil para identificar patrones espaciales y áreas de alta actividad delictiva. Un mapa de calor o un mapa de puntos pueden revelar áreas con mayor concentración de arrestos, lo que puede ser útil para la toma de decisiones en políticas de seguridad pública y asignación de recursos policiales.

Numero de arrestos por género:



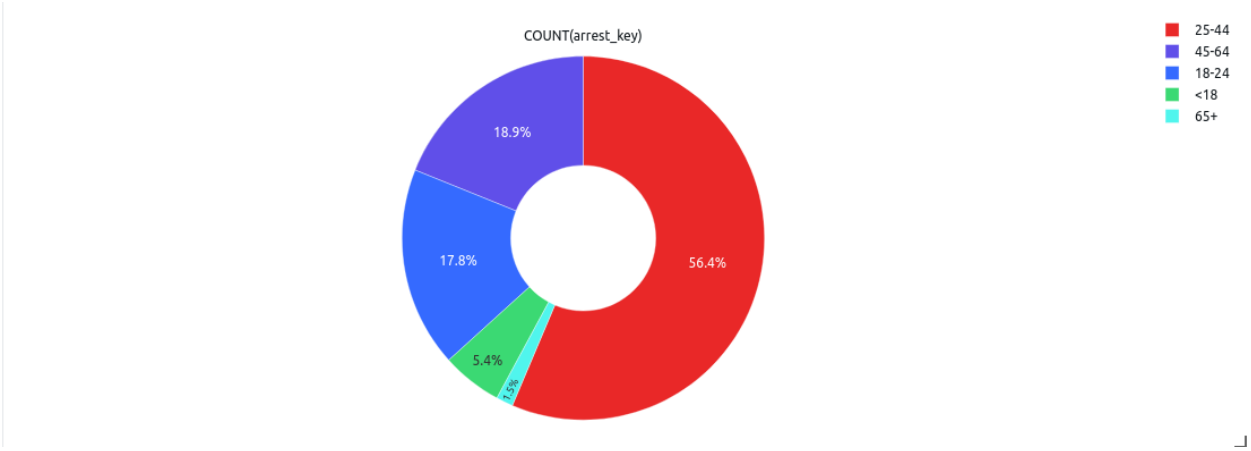
Este gráfico de barras muestra la cantidad total de arrestos desglosados por género. Es útil para visualizar la proporción de arrestos entre hombres y mujeres, lo que puede ser importante para comprender las disparidades de género en la delincuencia y para orientar políticas específicas de intervención.

Arrestos por BORO:



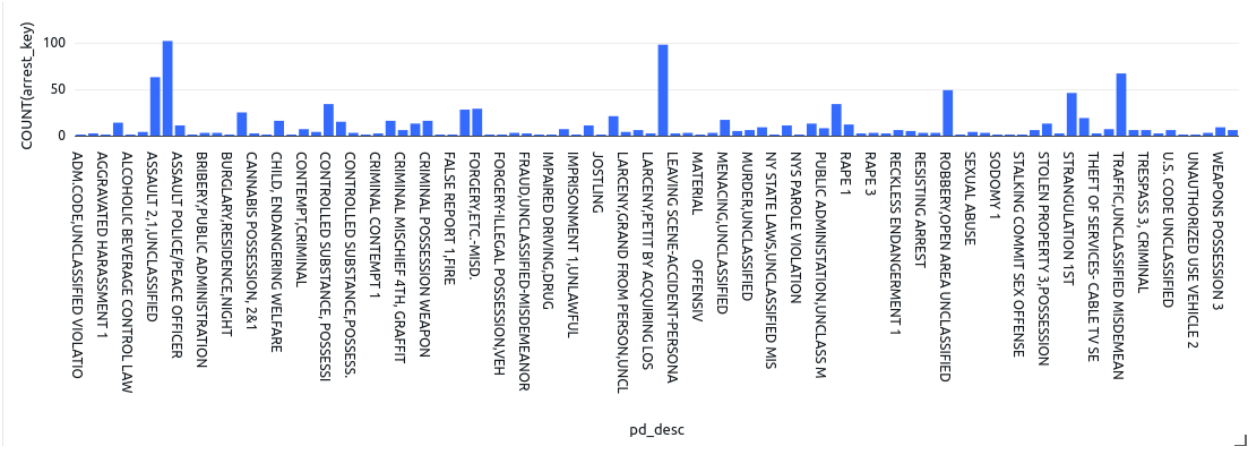
Esta visualización muestra la cantidad de arrestos en cada borough o distrito de la ciudad. Proporciona una idea de la distribución geográfica de la actividad delictiva en diferentes áreas de la ciudad, lo que puede ser útil para la planificación y asignación de recursos policiales.

Arrestos por edad:



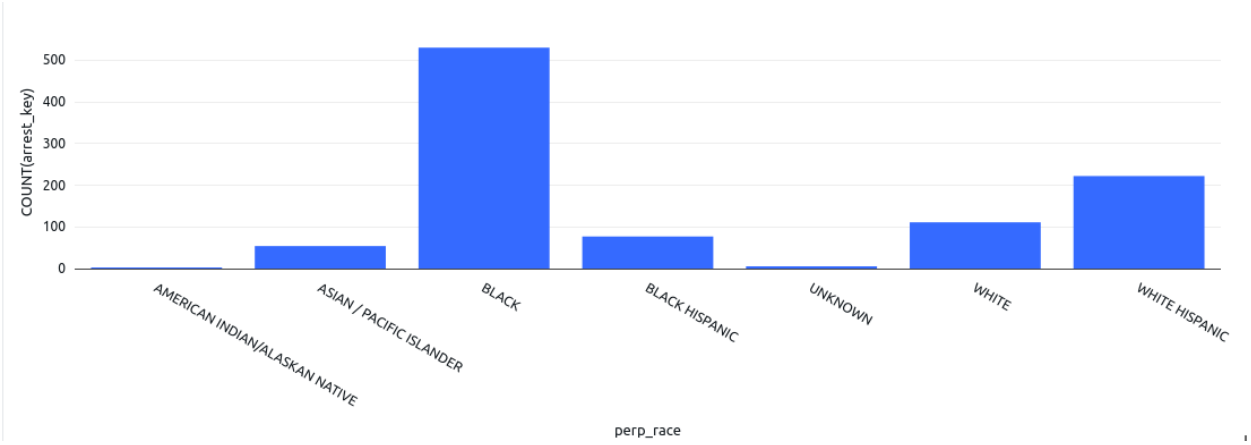
Este gráfico de barras muestra la cantidad de arrestos agrupados por rango de edad. Permite identificar las edades más comunes entre los arrestados y puede ser útil para comprender las tendencias de la delincuencia en diferentes grupos de edad y para informar sobre programas de prevención dirigidos a grupos específicos.

Arrestos por delito:



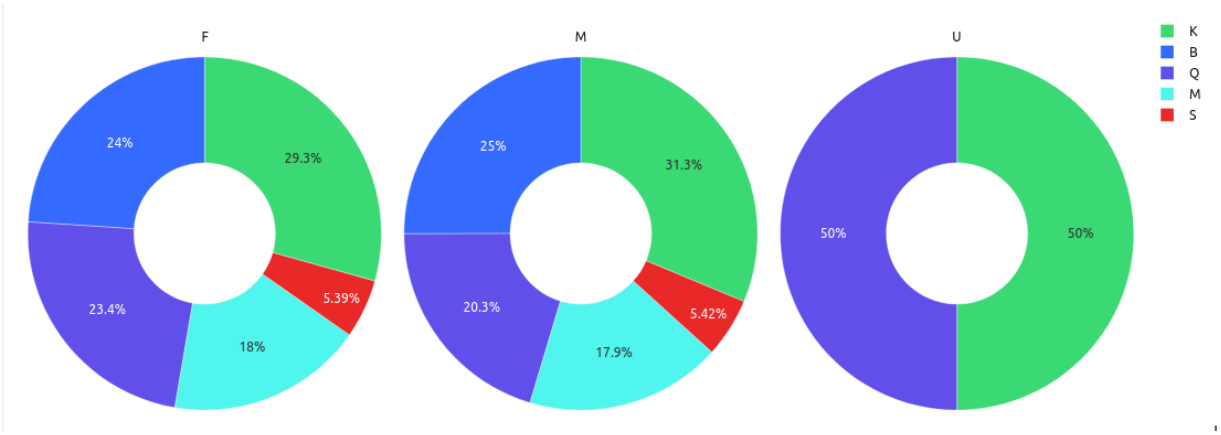
Esta visualización muestra la cantidad de arrestos para diferentes tipos de delitos. Es útil para identificar los delitos más comunes y su incidencia en el conjunto de datos, lo que puede proporcionar información valiosa para la formulación de políticas de seguridad y prevención del delito.

Arrestos por raza:



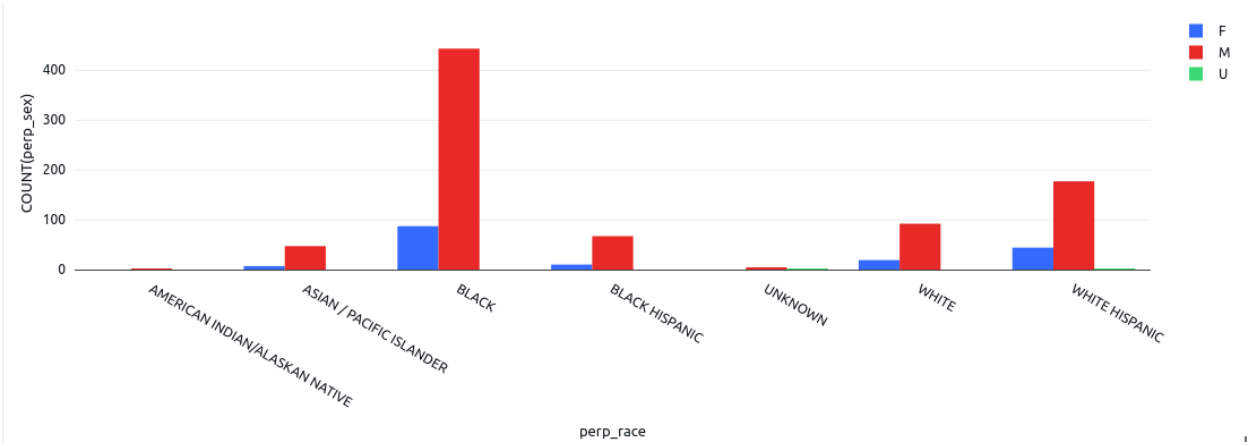
Este gráfico de barras muestra la cantidad de arrestos agrupados por raza o etnia. Es útil para comprender las disparidades raciales en la delincuencia y puede ayudar a identificar posibles sesgos en la aplicación de la ley.

Arrestos en BORO según el sexo:



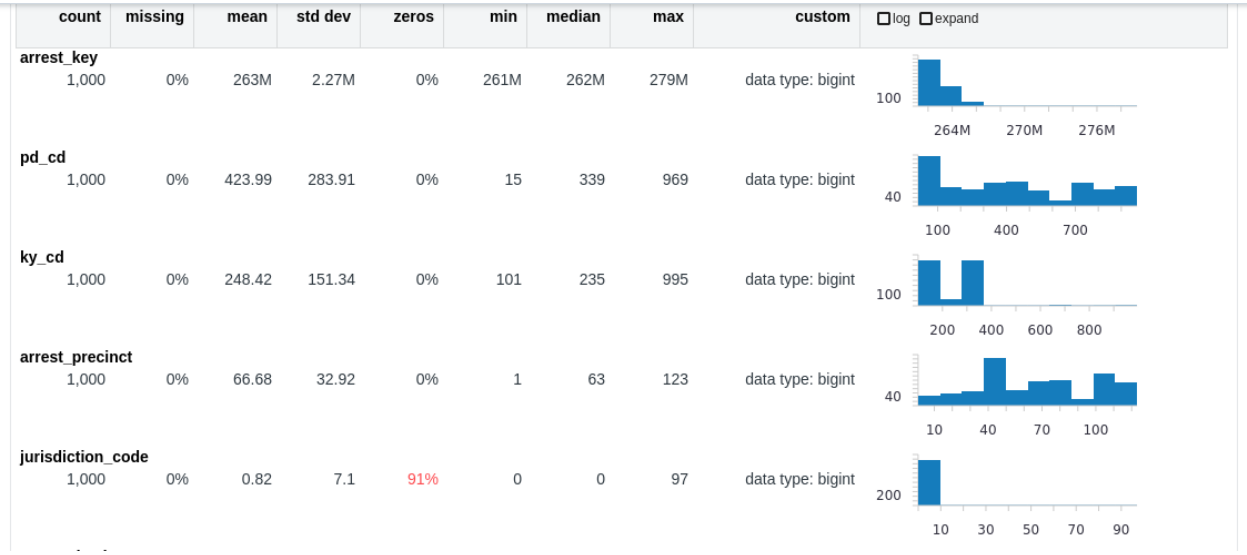
Esta visualización muestra la cantidad de arrestos desglosados por BORO y género. Permite comparar la distribución de arrestos entre hombres y mujeres en cada área geográfica, lo que puede ser útil para comprender las diferencias de género en la delincuencia a nivel local.

Arrestos según sexo y raza:



Este gráfico de barras agrupado muestra la cantidad de arrestos desglosados por género y raza. Permite comparar la distribución de arrestos entre diferentes grupos raciales y étnicos, así como entre hombres y mujeres, lo que puede proporcionar información valiosa sobre las disparidades en la aplicación de la ley.

Análisis del comportamiento:



	count	missing	mean	std dev	zeros	min	median	max	custom	log expand
arrest_key	1,000	0%	263M	2.27M	0%	261M	262M	279M	data type: bigint	100 / 264M 270M 276M
pd_cd	1,000	0%	423.99	283.91	0%	15	339	969	data type: bigint	40 / 100 400 700
ky_cd	1,000	0%	248.42	151.34	0%	101	235	995	data type: bigint	100 / 200 400 600 800
arrest_precinct	1,000	0%	66.68	32.92	0%	1	63	123	data type: bigint	40 / 10 40 70 100
jurisdiction_code	1,000	0%	0.82	7.1	91%	0	0	97	data type: bigint	200 / 10 30 50 70 90
jurisdiction_code	1,000	0%	0.82	7.1	91%	0	0	97	data type: bigint	200 / 10 30 50 70 90
x_coord_cd	1,000	0%	1.01M	22.1k	0%	915k	1.01M	1.06M	data type: bigint	50 / 940K 980K 1M
y_coord_cd	1,000	0%	207k	31.4k	0%	126k	204k	272k	data type: bigint	40 / 140K 180K 220K 260K
latitude	1,000	0%	40.74	0.09	0%	40.51	40.73	40.91	data type: double	40 / 40.55 40.7 40.85
longitude	1,000	0%	-73.92	0.08	0%	-74.25	-73.92	-73.73	data type: double	50 / -74.2 -74.05 -73.9

El análisis del comportamiento de los datos, que incluye estadísticas descriptivas como datos faltantes, media, desviación estándar, mínimo y máximo de cada atributo del conjunto de datos, sirve para proporcionar una visión detallada y cuantitativa de la estructura y las características de los datos. Estas métricas son fundamentales para comprender la distribución, la variabilidad y las tendencias dentro de los datos, lo que a su vez ayuda a identificar patrones, anomalías y posibles sesgos. Además, este análisis proporciona información útil para la toma de decisiones en el procesamiento de datos, la selección de técnicas de modelado adecuadas y la formulación de hipótesis para investigaciones posteriores

Descripción set de datos pobreza en Nuevo York:

Promedio de estatus según el BORO:

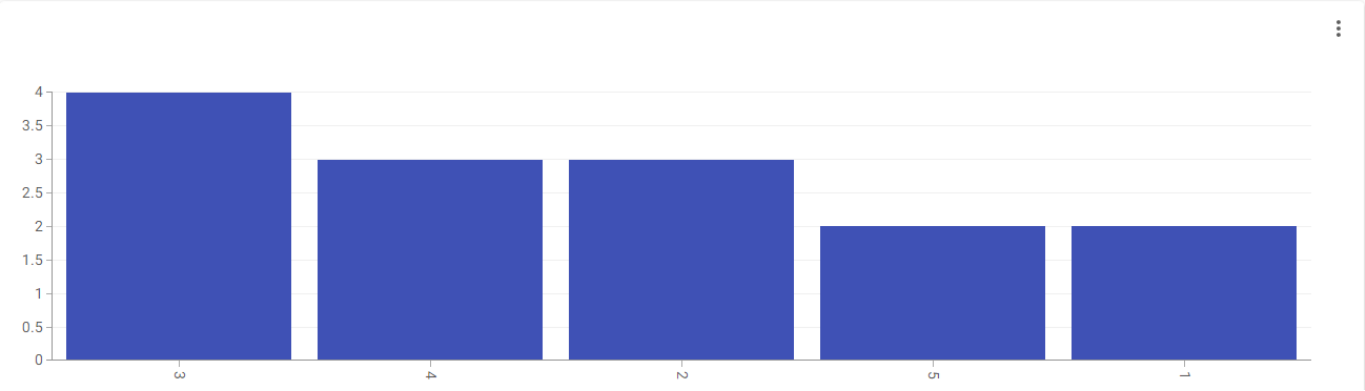
Boro	CitizenStatus (Average)
4	1.643692801207661
1	1.522231286967163
2	1.4852795236520013
3	1.4303364589078875
5	1.3186567164179104

BORO:

1 Bronx
2 Brooklyn
3 Manhattan
4 Queens
5 Staten Island

Las gráficas representan el promedio del estatus ciudadano según el BORO (distrito) en Nueva York. Al dividir el análisis por distritos y estatus ciudadano, proporciona una visión detallada de cómo varía la distribución del estatus ciudadano en cada área de la ciudad. Esta información es crucial para comprender las dinámicas socioeconómicas y demográficas dentro de Nueva York, ya que puede revelar disparidades en la ciudadanía y ofrecer perspectivas sobre la diversidad y la inmigración en diferentes distritos. Además, al comparar los promedios de estatus ciudadano entre los distritos, se pueden identificar posibles áreas de interés para políticas públicas y programas de inclusión social.

Mediana de nivel de educación en cada BORO:

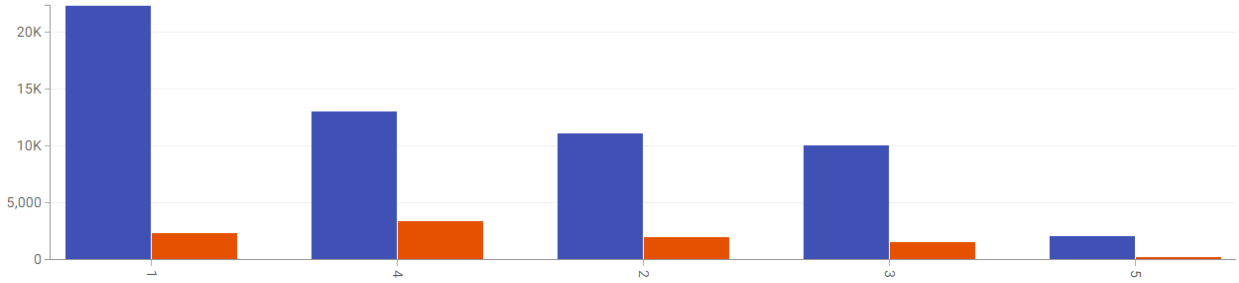


EducAttain:

1 menos que la escuela secundaria
2 Título de escuela secundaria
3 algo de universidad
4 Licenciatura o superior

Las gráficas muestran la mediana del nivel de educación en cada BORO (distrito) de Nueva York. Este análisis proporciona una perspectiva sobre el nivel educativo promedio en diferentes áreas de la ciudad. Al comparar las medianas de educación entre los distritos, se pueden identificar disparidades en el acceso a la educación y las oportunidades educativas. Esta información es fundamental para comprender las necesidades educativas de cada comunidad y puede orientar la asignación de recursos y políticas destinadas a mejorar la equidad educativa y promover el acceso a la educación en toda la ciudad.

Declaración de estado de pobreza según raza:

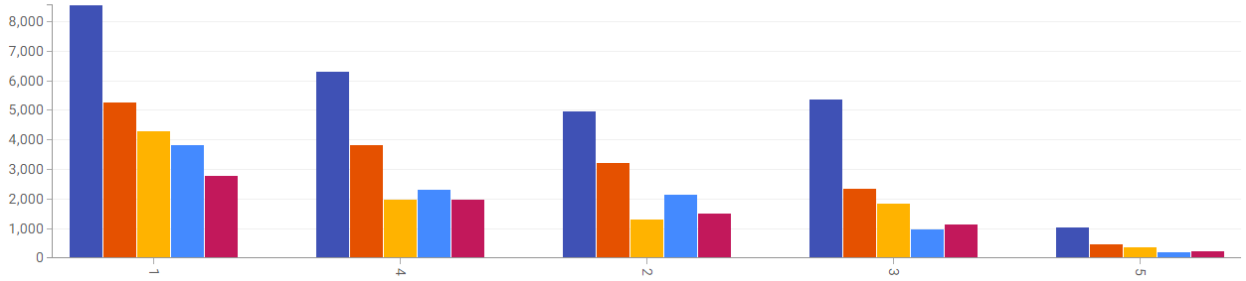


Parámetros:

1 blanco no hispano
2 negros no hispanos
3 asiáticos no hispanos
4 hispano, cualquier raza
5 Otra raza/grupo étnico

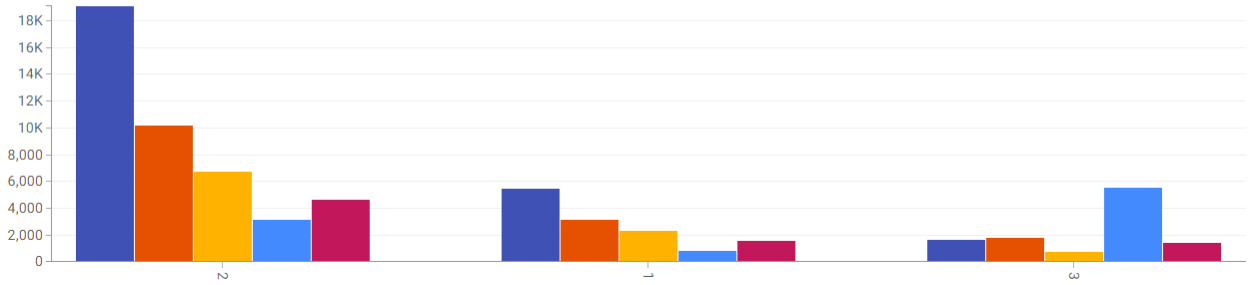
El análisis de la declaración del estado de pobreza según la raza proporciona una visión de las disparidades económicas entre diferentes grupos étnicos en Nueva York. Al examinar la distribución de la declaración de pobreza en función de la raza, podemos identificar inequidades en el acceso a recursos económicos y oportunidades para distintas comunidades. Esto es crucial para comprender y abordar las disparidades económicas y sociales basadas en la raza, lo que puede informar políticas y programas dirigidos a reducir la pobreza y promover la equidad racial en la ciudad.

Total de horas trabajadas por raza:



Analizar el total de horas trabajadas por raza ofrece una perspectiva importante sobre las disparidades laborales entre diferentes grupos étnicos en Nueva York. Al examinar esta métrica, podemos entender mejor cómo se distribuye el trabajo y la contribución económica en función de la raza. Esto puede ayudar a identificar posibles desigualdades en el acceso al empleo, las oportunidades laborales y los ingresos entre diferentes grupos raciales. Este análisis es fundamental para comprender y abordar las disparidades laborales y económicas basadas en la raza, lo que puede informar políticas y programas destinados a promover la equidad laboral y reducir las brechas de ingresos en la ciudad.

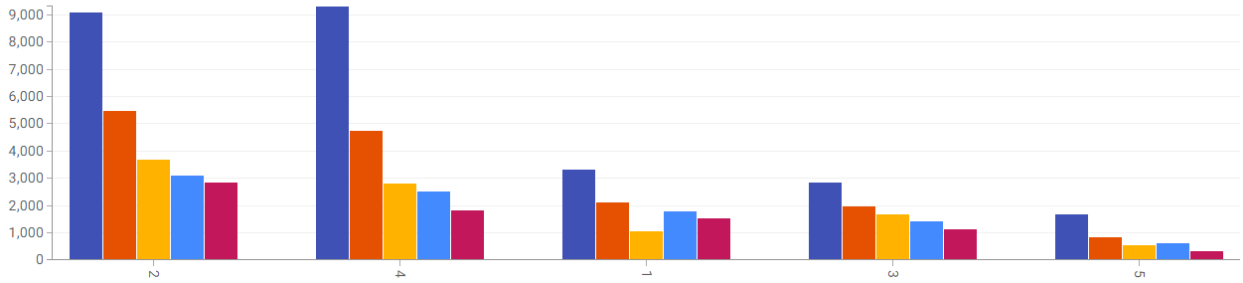
Horas trabajadas según edad:



Edad: 1. Under 18 years; 2.18 to 64 years;3. 65+ years

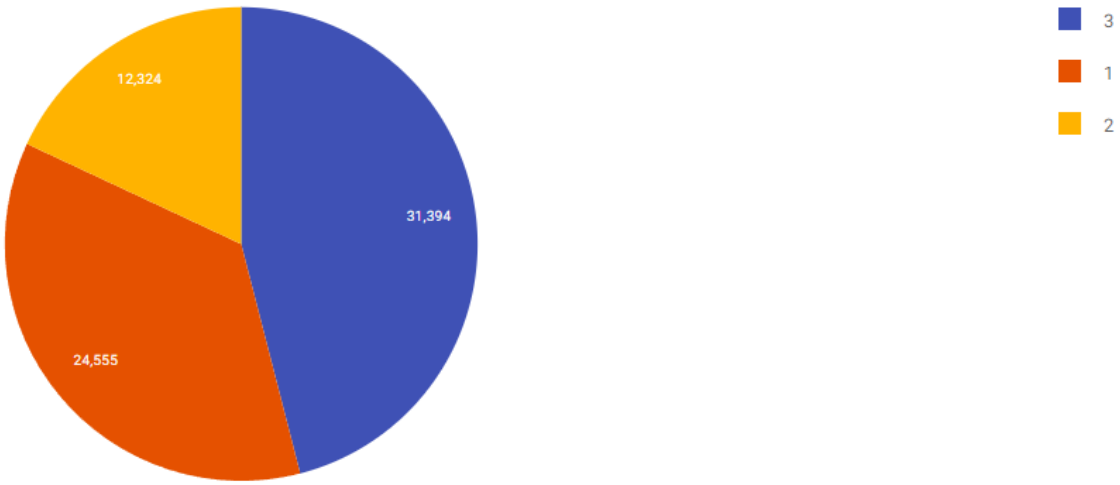
Analizar las horas trabajadas según la edad proporciona información crucial sobre la participación laboral en diferentes etapas de la vida. Este análisis puede revelar patrones de trabajo, como la distribución de horas laborales entre grupos de edad, las tasas de empleo y desempleo en diferentes etapas de la vida, y cómo cambia la participación laboral a lo largo del tiempo. Además, puede ayudar a identificar posibles desafíos o necesidades específicas de empleo para grupos de edad particulares, como los jóvenes que ingresan al mercado laboral, los trabajadores de mediana edad o los adultos mayores que buscan empleo o transiciones laborales.

Horas trabajadas por BORO:



El análisis de las horas trabajadas por BORO proporciona una visión detallada de la actividad laboral en diferentes áreas de la ciudad de Nueva York. Al observar cómo se distribuyen las horas trabajadas en cada BORO, se pueden identificar disparidades en la participación laboral entre las diferentes regiones de la ciudad. Esto puede ayudar a comprender mejor la dinámica laboral local, incluida la disponibilidad de empleo, los tipos de trabajo disponibles y las características del mercado laboral en cada área.

Porcentaje de experiencia laboral del individuo:

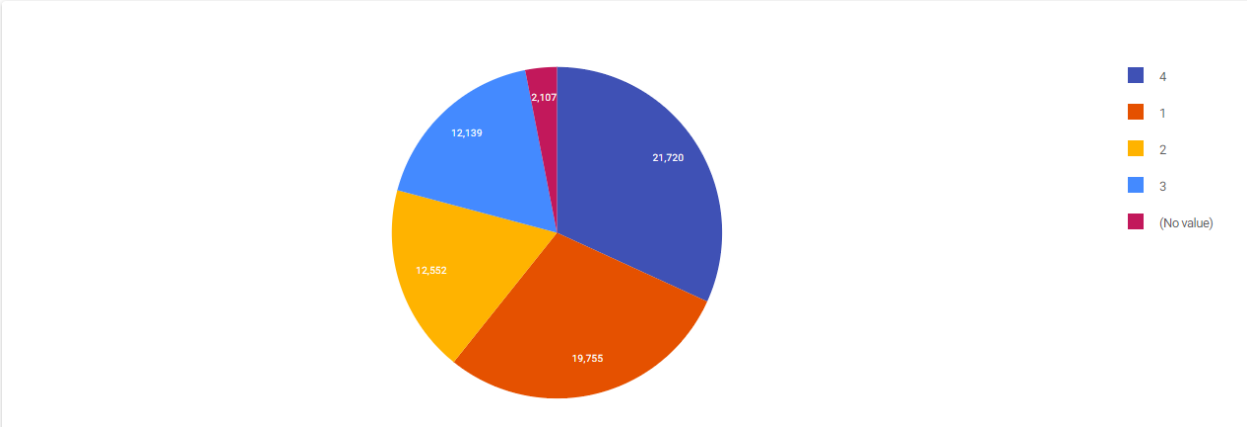


Parámetros:
1 tiempo completo todo el año
2 Menos de tiempo completo durante todo el año
3 Sin trabajo

El porcentaje de experiencia laboral del individuo muestra la distribución de la población según su experiencia laboral en términos de tiempo completo o parcial durante todo el

año, así como aquellos que están sin trabajo. Este análisis proporciona una visión general de la participación laboral de la población en la muestra de datos y ayuda a comprender las tendencias de empleo y desempleo en la población estudiada

Porcentaje de niveles de educación:



EducAttain:

1 menos que la escuela secundaria
2 Título de escuela secundaria
3 algo de universidad
4 Licenciatura o superior

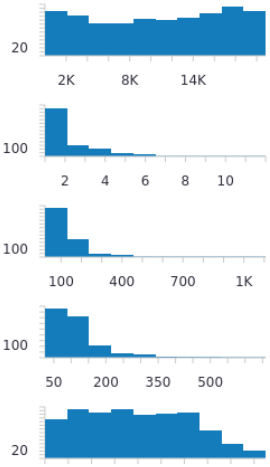
Este análisis permite entender la composición educativa de la población en términos generales y identificar las proporciones de personas con diferentes niveles de educación en toda la muestra de datos. Al examinar estos porcentajes, podemos obtener una visión general de la educación de la población en el conjunto de datos y entender mejor la estructura educativa de la muestra.

Análisis de comportamiento:

Numeric Features (61)

Chart to show: Standard ☐log ☐expand

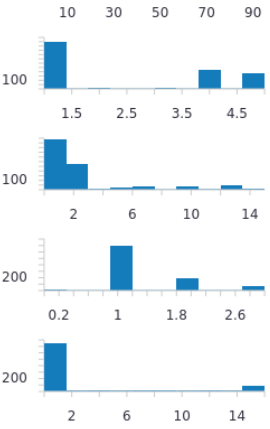
	count	missing	mean	std dev	zeros	min	median	max	custom
serialno	1,000	0%	10.9k	6,251.53	0%	1	11.0k	20.8k	data type: bigint
sporder	1,000	0%	2.34	1.57	0%	1	2	12	data type: bigint
pwgtp	1,000	0%	124.69	98.9	0%	21	99	1,107	data type: bigint
wgtp	1,000	0%	121.4	94.43	0%	25	97	659	data type: bigint
agep	1,000	0%	39.35	23.09	0.5%	0	38	95	data type: bigint



Numeric Features (61)

Chart to show: Standard ☐log ☐expand

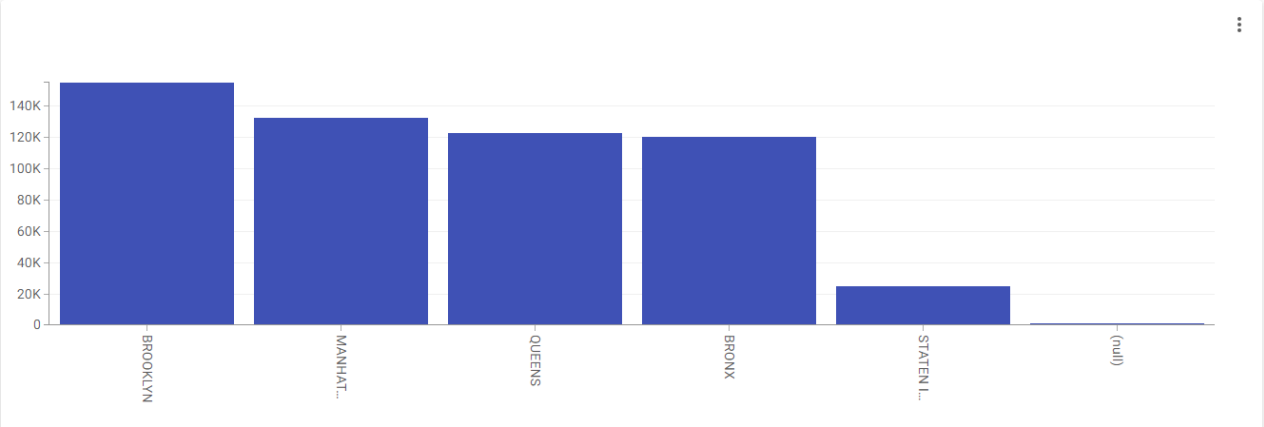
	count	missing	mean	std dev	zeros	min	median	max	custom
cit	1,000	0%	2.49	1.71	0%	1	1	5	data type: bigint
rel	1,000	0%	2.43	3.68	38.3%	0	1	15	data type: bigint
sch	1,000	0%	1.32	0.65	2.4%	0	1	3	data type: bigint
schg	1,000	0%	2.66	5.07	73.4%	0	0	16	data type: bigint



El análisis del comportamiento de los datos para el conjunto de datos de pobreza en Nueva York se realizó enfocándose en un subconjunto seleccionado de atributos en lugar de examinar todos los atributos disponibles. Esta selección puede deberse a la relevancia de los atributos para los objetivos específicos del estudio, las limitaciones de tiempo y recursos, así como la complejidad del conjunto de datos. Al centrarse en un grupo representativo de atributos, el análisis pudo proporcionar una comprensión adecuada del conjunto de datos en su conjunto, permitiendo identificar patrones, tendencias y relaciones importantes para el estudio de la pobreza en Nueva York.

Exploración de Quejas en Nuevo York:

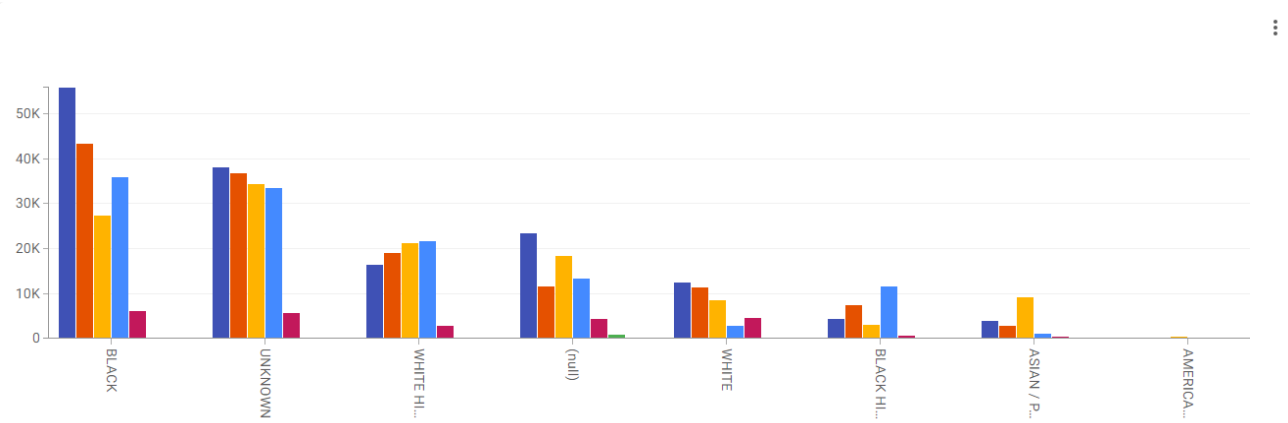
Total, de quejas por BORO



Esta distribución sugiere que Brooklyn tiene el mayor número de incidentes de robo, seguido de cerca por Manhattan, mientras que Queens tiene la menor cantidad de robos en comparación con los otros dos distritos.

Esta información proporciona una visión general de la distribución de robos en la ciudad y puede ser útil para comprender las tendencias delictivas y asignar recursos de seguridad de manera más efectiva en cada distrito

Raza del sospechoso agrupado por BORO



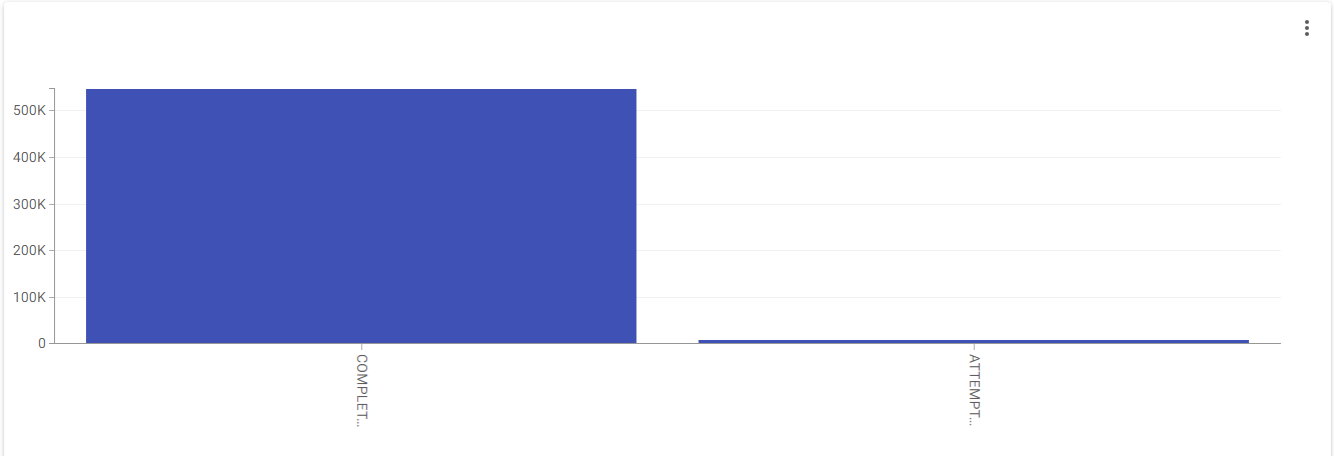
Los datos se presentan en una tabla donde las columnas están coloreadas para representar los diferentes distritos de la ciudad: Brooklyn (azul), Manhattan (naranja), Queens (amarillo), Bronx (azul claro) y Staten Island (rojo).

Se observa que la mayoría de los sospechosos son de raza negra en el distrito de Brooklyn.

También se destaca que la raza hispana/blanca es más prominente como sospechosos en áreas como el Bronx y menos prominentes en Brooklyn.

A partir de estos hallazgos, se plantea la hipótesis de que la distribución de sospechosos por raza podría estar influenciada por la segregación residencial o la territorialidad de los grupos étnicos en la ciudad.

Crimen fue completado o no

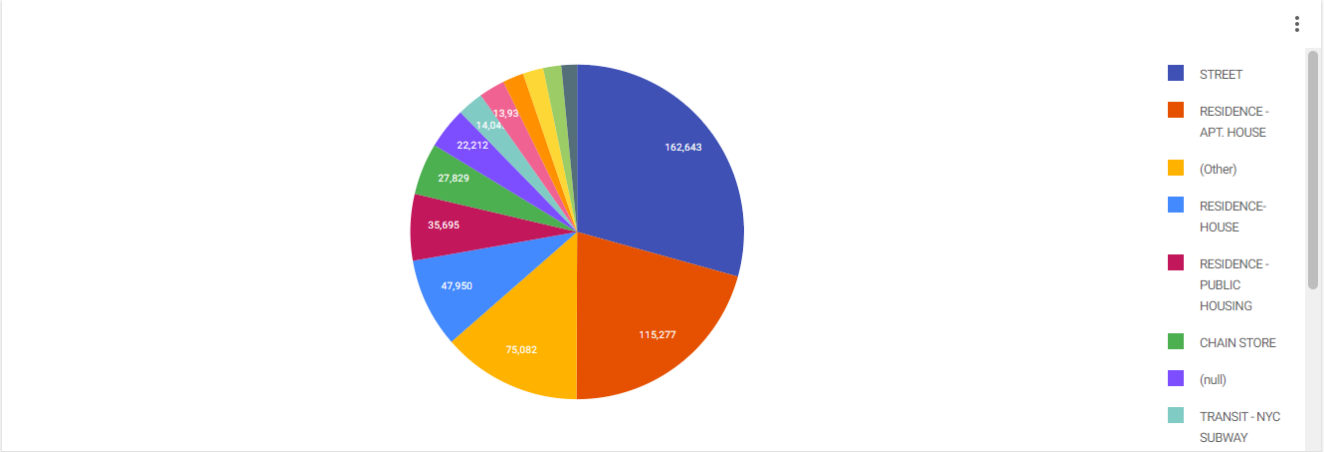


La gráfica muestra que la mayoría de los robos son completados en comparación con los robos no completados.

Esta diferencia podría sugerir que los robos completados son más propensos a ser reportados a las autoridades en comparación con los robos no completados.

Es posible que los robos no completados no se reporten con la misma frecuencia debido a una variedad de razones, como la falta de pruebas suficientes, la percepción de que no es necesario o la falta de conocimiento sobre cómo reportarlos.

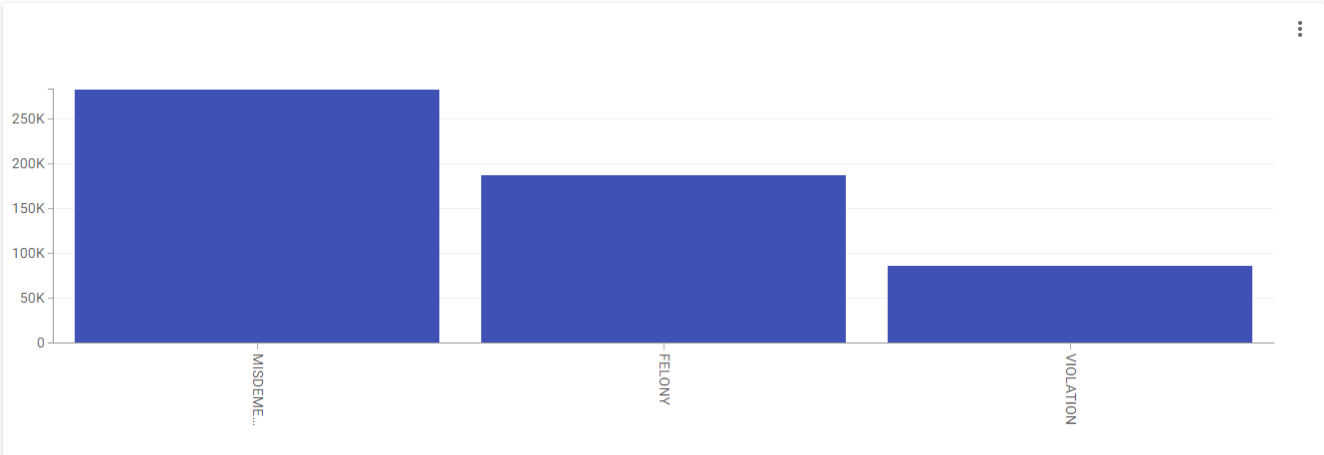
Lugares en los que ocurren los crímenes



La tabla indica que los lugares donde ocurren más crímenes son en la calle, seguido por residencias en apartamentos y residencias en casas.

Esta información sugiere que las ubicaciones públicas, como las calles, y los entornos residenciales son más propensos a ser escenarios de actividad criminal en comparación con otros tipos de instalaciones, como tiendas de abarrotes, parques o en el mismo transporte público.

Tipo de crimen:

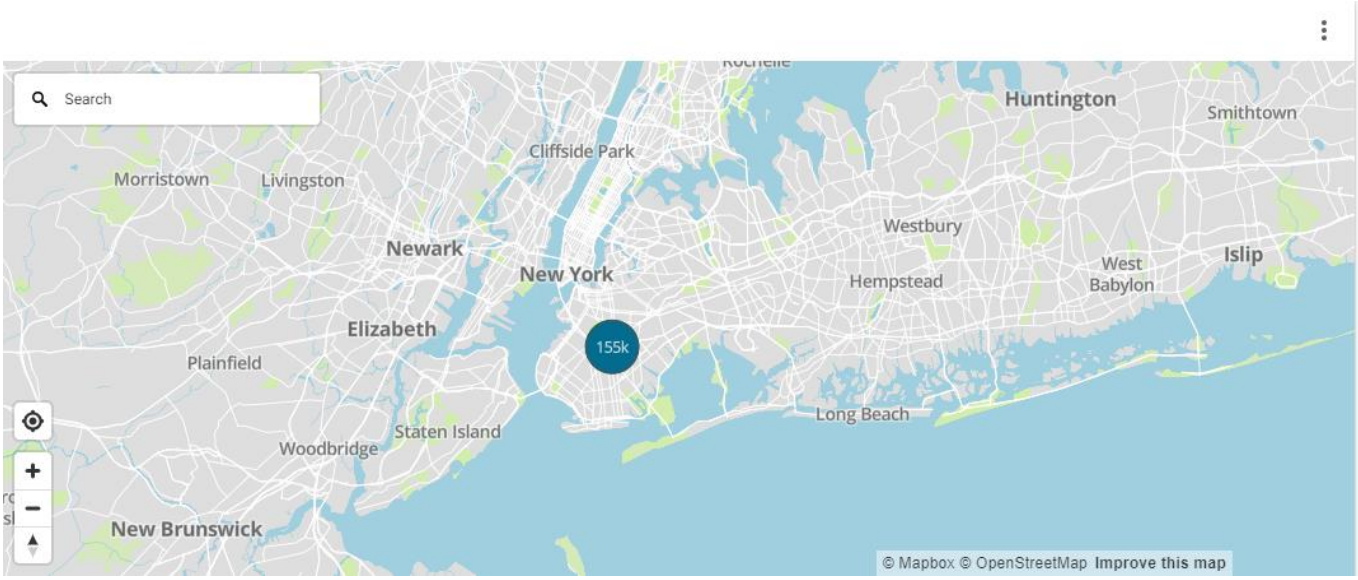


La gráfica indica que la mayoría de las quejas son por delitos menores (misdemeanor), seguidas por delitos graves (felony) y, finalmente, violaciones.

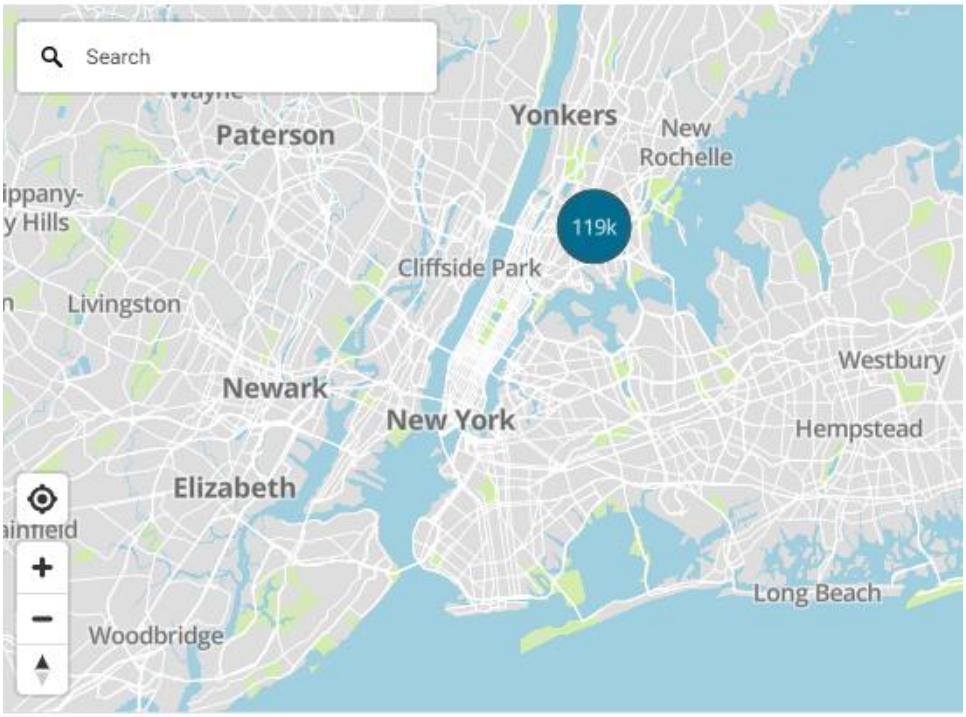
Esta distribución es coherente con las expectativas comunes, ya que los delitos menores tienden a ser más frecuentes en comparación con los delitos graves y las violaciones.

Los delitos menores suelen incluir una variedad de infracciones menos graves, como robos menores o vandalismo, que pueden ocurrir con más frecuencia en comparación con delitos más serios como el asalto con arma de fuego o el asesinato, que caen en la categoría de delitos graves.

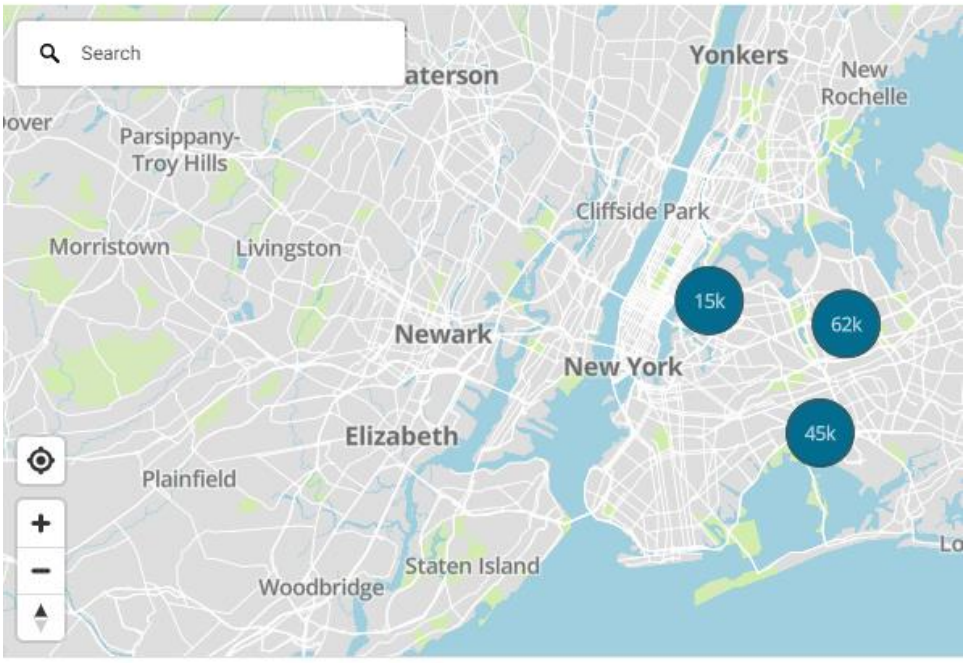
Mapa cantidad de crímenes Brooklyn:



Mapa cantidad de crímenes Bronx:



Mapa cantidad de crímenes Queens:



en la primera gráfica, se muestra la distribución de crímenes en Brooklyn, con un total de 155 mil quejas.

En la segunda gráfica, se muestra la distribución de crímenes en el Bronx, con un total de 119 mil quejas.

En la tercera gráfica, se muestra la distribución de crímenes en Queens, con un total de 122 mil quejas.

Estas gráficas proporcionan una representación visual de cómo se dividen los crímenes en cada distrito de la Ciudad de Nueva York. Los números totales de quejas indican la cantidad total de incidentes reportados en cada distrito, lo que puede ser útil para comprender la carga de trabajo de las autoridades locales y las necesidades de seguridad en cada área.

V. Reporte de calidad

Reporte de calidad para data set arrestos en NY:

Se utilizó la función **`select`** de Spark con una lista de expresiones para contar los valores nulos o inválidos en cada columna del conjunto de datos de arrestos. La expresión **`when(isnan(c) | col(c).isNull(), c)`** se encarga de verificar si un valor es NaN (no es un número) o nulo en

cada columna **c**, y devuelve **c** si es así, lo que permite contar los valores nulos o inválidos en cada columna.

Este reporte de calidad de datos sirve para evaluar la integridad y completitud del conjunto de datos de arrestos. Al calcular el número de valores nulos o inválidos en cada columna, podemos identificar posibles problemas o deficiencias en los datos que podrían afectar la calidad del análisis posterior o los resultados del modelo.

ARREST_KEY	ARREST_DATE	PD_CD	PD_DESC	KY_CD	OFNS_DESC	LAW_CODE	LAW_CAT_CD	ARREST_BORO	ARREST_PRECINCT
0	0	2	17	17	17	2	1601	0	0

JURISDICTION_CODE	AGE_GROUP	PERP_SEX	PERP_RACE	X_COORD_CD	Y_COORD_CD	Latitude	Longitude	New Georeferenced Column
0	0	0	2711	0	0	0	0	0

Los resultados del análisis de calidad de datos indican que la mayoría de las columnas en el conjunto de datos de arrestos no contienen valores nulos o inválidos. Sin embargo, se observa la presencia de valores faltantes en varias columnas específicas. Específicamente, se han identificado 2 registros con valores faltantes en la columna "PD_CD" (código de delito policial), 17 registros con valores faltantes en las columnas "PD_DESC" (descripción del delito policial), "KY_CD" (código de delito principal) y "OFNS_DESC" (descripción del delito principal), y 2 registros con valores faltantes en la columna "LAW_CODE" (código de ley). Además, se encontraron 1601valores faltantes en la columna " LAW_CAT_CD " (Categoría de Ley) y 2711 en la columna “PERP_RACE” (Raza del delincuente)

Este análisis destaca la importancia de garantizar la integridad de los datos antes de proceder con cualquier análisis adicional. Aunque la mayoría de las columnas están completas, la presencia de valores faltantes en estas columnas específicas sugiere la necesidad de una investigación más profunda para comprender la causa de estos faltantes. Esto resalta la importancia de llevar a cabo una revisión exhaustiva de la calidad de los datos como parte fundamental del proceso analítico, lo que garantiza la confiabilidad de los resultados obtenidos.

Imputación por moda para PD_CD y PERP_RACE: Dado que se identificaron valores faltantes en las columnas "PD_CD" (código de delito policial) y "PERP_RACE" (raza del delincuente), se puede utilizar la imputación por moda para completar estos valores faltantes. Esto implica reemplazar los valores faltantes con el valor más comúnmente observado en cada una de estas columnas.

Eliminación de registros con valores faltantes en PD_DESC, KY_CD, OFNS_DESC y LAW_CODE: Como se encontraron valores faltantes en estas columnas específicas, otra estrategia sería eliminar los registros correspondientes a estos valores faltantes. Esto garantiza que los datos restantes sean completos y no estén sesgados por la presencia de valores faltantes.

Imputación por moda para LAW_CAT_CD: Para la columna "LAW_CAT_CD" (categoría de ley), donde se encontraron 1601 valores faltantes, se puede aplicar la imputación por moda para completar estos valores faltantes. Esto implica reemplazar los valores faltantes con la categoría de ley más comúnmente observada en el conjunto de datos.

Reporte de calidad para data set pobreza en NY:

SCHL	ESR	LANX	ENG	MSP	WKW	JWTR	EducAttain
2107	11798	3533	37557	11021	31394	34956	2107

Únicos columnas con datos faltantes

El análisis de calidad de datos en el conjunto de datos de pobreza revela la presencia de valores nulos en varias columnas, lo que requiere atención para garantizar la integridad de los datos. En particular, se identificaron 2107 valores faltantes en la columna que indica el nivel educativo alcanzado (SCHL). Además, se observaron 11798 valores faltantes en la columna que representa el estado de empleo de la persona (ESR), mientras que la columna que indica el idioma principal hablado en el hogar (LANX) contiene 3533 valores faltantes. La columna que refleja el dominio del idioma inglés (ENG) muestra la mayor cantidad de valores faltantes, con un total de 37557. En cuanto al estado civil de las personas, se encontraron 11021 valores faltantes en la columna correspondiente (MSP). Por otro lado, la columna que describe la semana laboral habitual (WKW) presenta 31394 valores faltantes, y la columna que representa el modo de transporte principal utilizado para viajar al trabajo (JWTR) contiene 34956 valores faltantes. Finalmente, se identificaron 2107 valores faltantes en la columna que indica el nivel educativo alcanzado (EducAttain).

Estos hallazgos son fundamentales para comprender la calidad de los datos y deben abordarse adecuadamente antes de realizar análisis o modelado posteriores. La imputación de datos u otras técnicas de limpieza pueden ser necesarias para completar los valores faltantes y garantizar la fiabilidad de los resultados obtenidos a partir de este conjunto de datos.

Técnicas de limpieza:

- schl (Asistencia escolar): Dado que este atributo representa el nivel de grado escolar asistido, una posible técnica para tratar los valores faltantes podría ser la imputación por la moda, es decir, completar los valores faltantes con el grado escolar más comúnmente reportado en el conjunto de datos.

- esr (Estado de empleo de la persona): Para esta columna, podría ser apropiado imputar los valores faltantes utilizando la moda o la categoría de empleo más común entre las personas en el conjunto de datos. Alternativamente, si los datos no son demasiado escasos, se podría utilizar la media o la mediana de la columna para imputar los valores faltantes.
- lanx (Idioma distinto al inglés hablado en casa): Una estrategia para manejar los valores faltantes en esta columna podría ser la imputación por la moda, donde los valores faltantes se completan con el idioma más comúnmente hablado en el conjunto de datos.
- eng (Capacidad de hablar inglés): Para esta columna, la imputación por la moda también podría ser una opción adecuada, donde los valores faltantes se completan con el nivel de habilidad en inglés más comúnmente reportado en el conjunto de datos.
- msp (Estado civil): Dado que esta columna indica el estado civil de la persona, una técnica de imputación viable podría ser completar los valores faltantes con el estado civil más frecuente observado en el conjunto de datos.
- wkw (Semanas trabajadas el año anterior): Aquí, la imputación por la media o la mediana de la columna podría ser una opción adecuada para completar los valores faltantes, proporcionando una estimación razonable de las semanas trabajadas basada en el comportamiento general de la población en el conjunto de datos.
- JWTR (Modo de transporte principal utilizado para viajar al trabajo): Una posible técnica sería la imputación por la moda, donde los valores faltantes se completan con el modo de transporte más comúnmente utilizado para viajar al trabajo en el conjunto de datos.
- EducAttain (Nivel educativo alcanzado): Dado que esta columna representa el nivel educativo alcanzado, podemos aplicar la imputación por la moda. Completar los valores faltantes con el nivel educativo más comúnmente reportado en el conjunto de datos.

Reporte de calidad para data set quejas de la policía de Nueva York actuales:

cmplnt_num	addr_pct_cd	boro_nm	cmplnt_fr_dt	cmplnt_fr_tm	cmplnt_to_dt	cmplnt_to_tm	crm_atpt_cptd_cd	hadevelopt	housing_psa	jurisdiction_code	juris_desc	ky_cd	law_cat_cd	loc_of_occur_desc	ofns_desc	parks_nm	patrol_boro
0	70	1019	0	0	34738	34325	0	553360	519092	0	0	0	0	111769	17	551791	0

pd_cd	pd_desc	prem_typ_desc	rpt_dt	station_name	susp_age_group	susp_race	susp_sex	transit_district	vic_age_group	vic_race	vic_sex	x_coord_cd	y_coord_cd	latitude	longitude	lat_lon	geocoded_column
386	386	22212	0	540384	286537	220259	195712	540384	160142	173461	0	12	12	12	12	12	12

Como se puede observar, varias columnas tienen valores faltantes, algunos en gran cantidad por lo que remplazar estos valores faltantes podría ser difícil.

Imputación de Moda: Para columnas categóricas como **LOC_OF_OCCUR_DESC, OFNS_DESC, PREM_TYP_DESC, SUSP_AGE_GROUP, SUSP_RACE, SUSP_SEX, TRANSIT_DISTRICT, VIC_AGE_GROUP, VIC_RACE, y VIC_SEX,** se puede considerar la imputación de moda para rellenar los valores faltantes. Esto implica reemplazar los valores faltantes con el valor más frecuente en esa columna.

Eliminación de Registros: Para columnas con un alto número de valores faltantes, como **PARKS_NM**, se podría considerar la eliminación de esta columna del análisis si no es esencial para los objetivos del estudio.

VI. Planteamiento de preguntas

¿Existe alguna relación entre la ubicación específica del incidente y el tipo de delito?

¿Cuál es la tendencia temporal de los delitos en Nueva York a lo largo de los años, basada en la fecha de reporte? ¿Hay algunos meses en los cuales los delitos aumentan y cuál sería la razón?

¿Existe una correlación entre la cantidad de delitos reportados en un año específico y los indicadores económicos clave de ese mismo año, como el PIB per cápita, la tasa de desempleo o el índice de pobreza, en el área metropolitana de Nueva York?

¿Cuáles son las áreas de mayor concentración de delitos en Nueva York, basadas en las coordenadas geográficas y dividido por (borough), y cómo se comparan estas áreas con los datos de demografía y nivel socioeconómico?

¿Cómo varía la situación de pobreza en Nueva York en función de la raza y etnicidad de los residentes?

¿Existe alguna correlación entre la tasa de delitos reportados en un área específica de Nueva York y el nivel de pobreza en ese mismo lugar?

¿Cómo varía la situación de pobreza entre los diferentes grupos étnicos y raciales en áreas con altos y bajos índices de criminalidad en Nueva York?

¿Hay relación entre la raza de quienes cometen delitos y la proporción de personas de esa misma raza que están en situación de pobreza en Nueva York?

VII. Filtros, limpieza y transformación

Limpieza del set de datos de arrestos:

Para la limpieza del set de datos se implementó un código en Pyspark que hace lo siguiente:

Este código implementa un proceso de limpieza de datos utilizando Apache Spark para el conjunto de datos de arrestos. La primera etapa consiste en calcular la moda de dos columnas específicas: "LAW_CAT_CD" y "PERP_RACE". Para esto, se agrupan los datos por cada categoría única en estas columnas y se cuenta la frecuencia de cada una. Luego, se selecciona la categoría con la mayor frecuencia como la moda. Una vez calculadas las modas, se procede a reemplazar los valores nulos en estas columnas con las modas correspondientes utilizando la función **withColumn**. Si una celda está vacía en "LAW_CAT_CD", se sustituye por la moda previamente calculada de esa columna; de manera similar, para los valores nulos en "PERP_RACE", se utiliza la moda correspondiente. Además, se eliminan las filas que contienen valores faltantes en columnas críticas como "PD_DESC", "KY_CD", "OFNS_DESC" y "LAW_CODE", garantizando así que el conjunto de datos esté completo y listo para análisis posteriores.

Limpieza de set de datos pobreza:

Este código realiza varias operaciones para limpiar el conjunto de datos de pobreza utilizando Apache Spark. En primer lugar, se calcula el total de filas en el DataFrame y se cuenta el número de valores nulos en cada columna, incluyendo valores como "(No value)", "UNKNOWN" y "(null)". Esto se hace utilizando una lista de comprensión en PySpark para iterar sobre cada columna y aplicar la función **count**, **when** y **isnan** para identificar y contar los valores nulos. Luego, se convierte el DataFrame de PySpark resultante en un DataFrame de pandas para facilitar su manipulación y visualización. Se calcula el porcentaje de valores nulos en cada columna dividiendo el número de valores nulos por el total de filas y multiplicándolo por 100. Se configura pandas para mostrar todas las filas y columnas, y se imprime el DataFrame resultante que muestra los porcentajes de valores nulos en cada columna.

Después de analizar los porcentajes de valores nulos, se decide eliminar algunas columnas específicas que tienen una cantidad significativa de valores nulos, como "JWTR", "ENG" y "WKW". Esto se logra utilizando el método **drop** en el DataFrame de pobreza.

Finalmente, se calcula la moda de la columna "SCHL" (nivel de educación) agrupando los datos por cada valor único en esta columna y seleccionando el valor con la mayor frecuencia. Luego, se reemplazan los valores nulos en la columna "SCHL" con la moda correspondiente utilizando la función **withColumn** en Spark, asegurando así que no haya valores faltantes en esta columna crítica para el análisis de datos de pobreza.

Limpieza del set de datos de quejas:

Eliminación de columnas con valores nulos predominantes: Primero, se identificaron las columnas que tenían una cantidad significativa de valores nulos. Estas columnas fueron: **TRANSIT_DISTRICT, STATION_NAME, PARKS_NM, HOUSING_PSA y HADEVELOPT**. La decisión de eliminar estas columnas se tomó porque una gran proporción de sus valores eran nulos.

Imputación de valores faltantes utilizando la moda: Luego, se abordó la imputación de valores faltantes en la columna **ADDR_PCT_CD, BORO_NM**. En lugar de eliminar toda la fila, se optó por imputar los valores faltantes en esta columna utilizando la moda de esta.

Eliminación de filas con ubicación desconocida (NULL):

Primero, se identificaron las filas en las cuales tanto la longitud (LONGITUDE) como la latitud (LATITUDE) eran desconocidas (NULL). Estas filas representaban instancias donde la ubicación no estaba registrada.

Llenado de valores en la columna LOC_OF_OCCUR_DESC basado en PREM_TYP_DESC:

Se decidió llenar los valores faltantes en la columna LOC_OF_OCCUR_DESC utilizando información de la columna PREM_TYP_DESC. La idea era inferir la ubicación del incidente a partir del tipo de establecimiento reportado.

Si el tipo de establecimiento reportado era "STREET", se asumió que el incidente ocurrió "FRONT OF" ese establecimiento.

Si el tipo de establecimiento reportado era uno de los valores específicos como "RESIDENCE - APT. HOUSE", "RESIDENCE-HOUSE", "RESIDENCE - PUBLIC HOUSING", "CHAIN STORE", o "DEPARTMENT STORE", se asumió que el incidente ocurrió "INSIDE" del establecimiento.

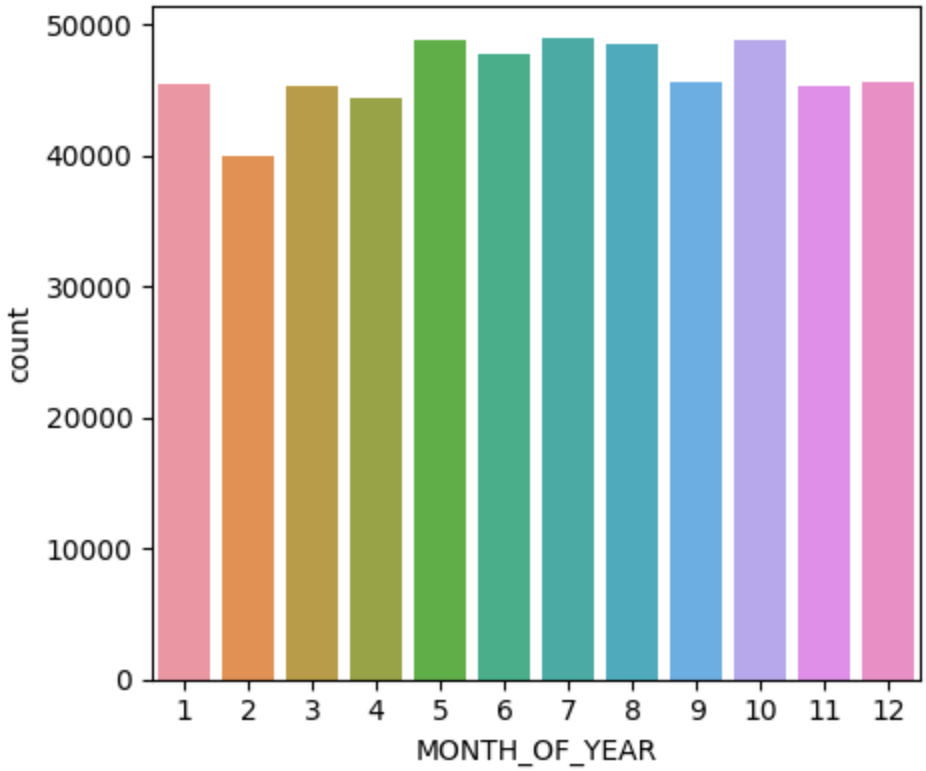
Para cualquier otro tipo de establecimiento, no se realizaron cambios en la columna LOC_OF_OCCUR_DESC.

Creación de la nueva característica MONTH_OF_YEAR

La nueva característica MONTH_OF_YEAR permite analizar los datos según la estacionalidad de los incidentes.

Al agregar esta característica, ahora podemos examinar si ciertos tipos de incidentes tienden a ocurrir más en ciertos meses del año.

Esto podría ser útil para identificar patrones estacionales en la incidencia de ciertos tipos de crímenes o quejas, así como para planificar medidas de seguridad o recursos de manera más efectiva en diferentes momentos del año.



Como se puede observar de la gráfica los meses que hay más robos es en diciembre y enero.

Preparación de datos y modelado:

Set de datos de pobreza:

Filtro 1: Variable "AGEP"

La variable "AGEP" representa la edad de las personas encuestadas. Se decidió filtrar el conjunto de datos para incluir únicamente a personas mayores de edad. Esta decisión se basa en dos consideraciones principales:

- Representatividad del Dataset: La mayoría de los datos recopilados pertenecen a personas mayores de edad, lo que garantiza una representación significativa de la población.
- Relevancia para el Análisis de Pobreza: Las personas mayores de edad son más propensas a enfrentar situaciones de pobreza, por lo que centrar el análisis en este grupo demográfico es relevante para los objetivos del estudio.

Filtro 2: Estado del Ciudadano

La variable relacionada con el estado del ciudadano incluía tres categorías: ciudadano actual, ciudadano por nacimiento y no ciudadano. Para enfocar el análisis en ciudadanos legalmente

reconocidos, se optó por excluir del conjunto de datos a aquellas personas que no son ciudadanos. Esta decisión se fundamenta en la necesidad de mantener la coherencia y relevancia del análisis dentro del contexto de la población objetivo: los ciudadanos legalmente reconocidos en Nueva York.

Al aplicar estos filtros, se garantiza que el conjunto de datos resultante esté optimizado para el análisis de la situación de pobreza en Nueva York, centrándose en la población adulta y ciudadana. Este proceso sienta las bases para la etapa de transformación y modelado de inteligencia artificial, que se abordará en secciones posteriores de este documento.

Transformación 1: Indexación de Variable "MAR"

La variable "MAR" representa el estado matrimonial de las personas encuestadas. Dado su carácter categórico y su potencial importancia para el análisis, se decidió indexar esta variable. Esta indexación permite codificar las diferentes categorías de estado matrimonial para su uso futuro en modelos de inteligencia artificial, facilitando así la inclusión de esta información en el análisis predictivo.

Transformación 2: Normalización de Variable "PreTaxIncome_PU"

La variable "PreTaxIncome_PU" representa el ingreso previo a impuestos de las personas encuestadas, medida en dólares. Dado que esta variable abarca una amplia gama de valores y puede afectar significativamente el rendimiento de los modelos de inteligencia artificial, se aplicó la normalización. Este proceso ajusta los valores de la variable para que estén en una escala común, lo que ayuda a mejorar la estabilidad y la eficacia de los modelos de análisis posteriores.

Transformación 3: Creación de Columna de Ingreso por Nivel de Educación

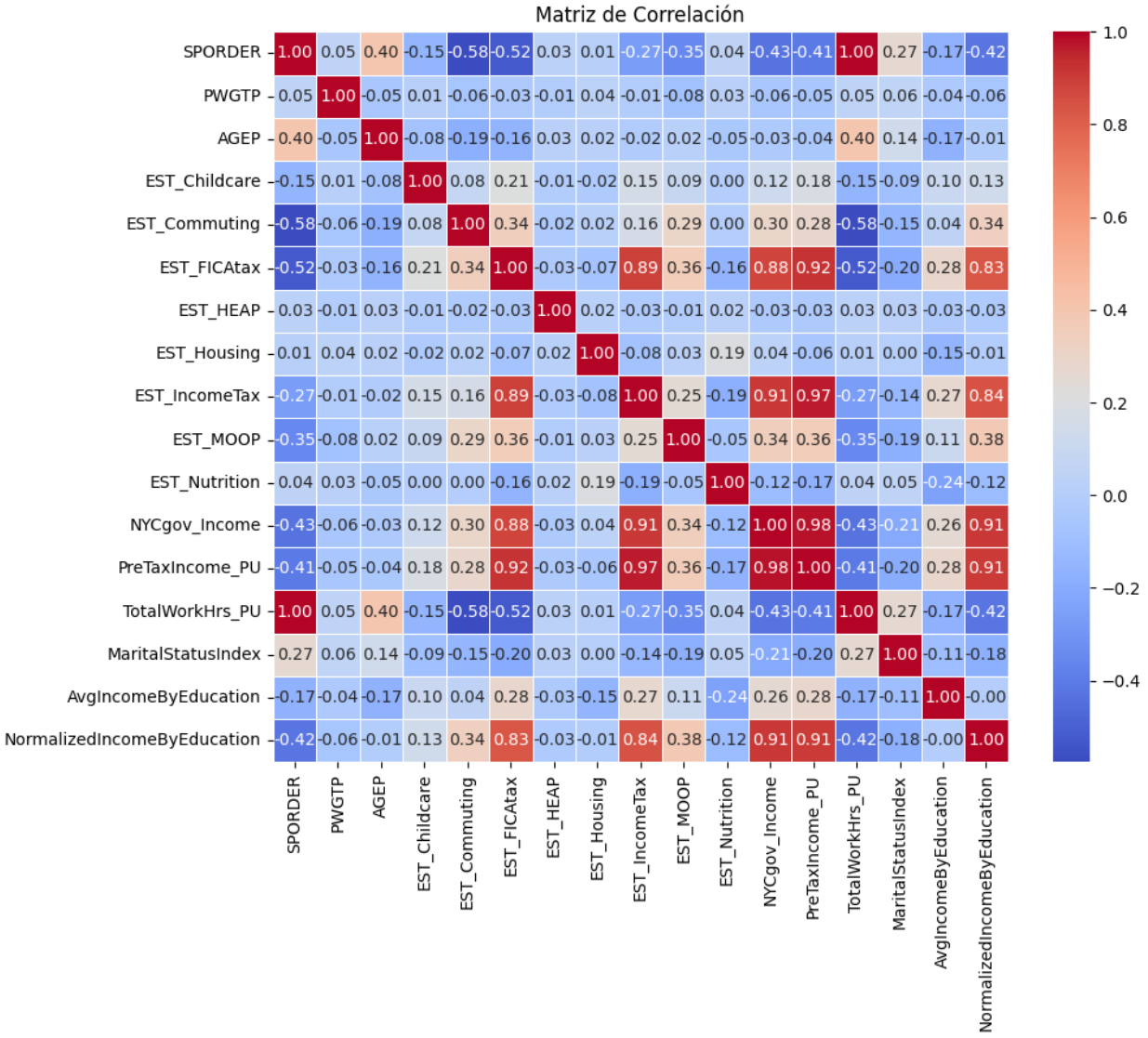
Se creó una nueva columna en el conjunto de datos que representa el ingreso de las personas según su nivel de educación. Esta transformación implica agrupar los datos según el nivel educativo de las personas y calcular el ingreso promedio para cada grupo. La inclusión de esta columna proporciona una perspectiva adicional sobre la distribución del ingreso en relación con el nivel educativo de la población, lo que puede ser útil para el análisis de la pobreza y la toma de decisiones.

Al aplicar estas transformaciones, se mejora la calidad y la utilidad del conjunto de datos para su análisis posterior. Estas modificaciones sientan las bases para la fase de modelado de inteligencia artificial, que se abordará en secciones posteriores de este documento.

Matriz de correlación:

En esta sección se presenta el análisis de la matriz de correlación realizada sobre todas las variables numéricas del conjunto de datos sobre la pobreza en Nueva York. El propósito de este análisis fue identificar las relaciones más significativas entre las variables y determinar aquellas

que estaban altamente correlacionadas, lo que permitiría descartar algunas de ellas para mejorar la calidad del modelo.



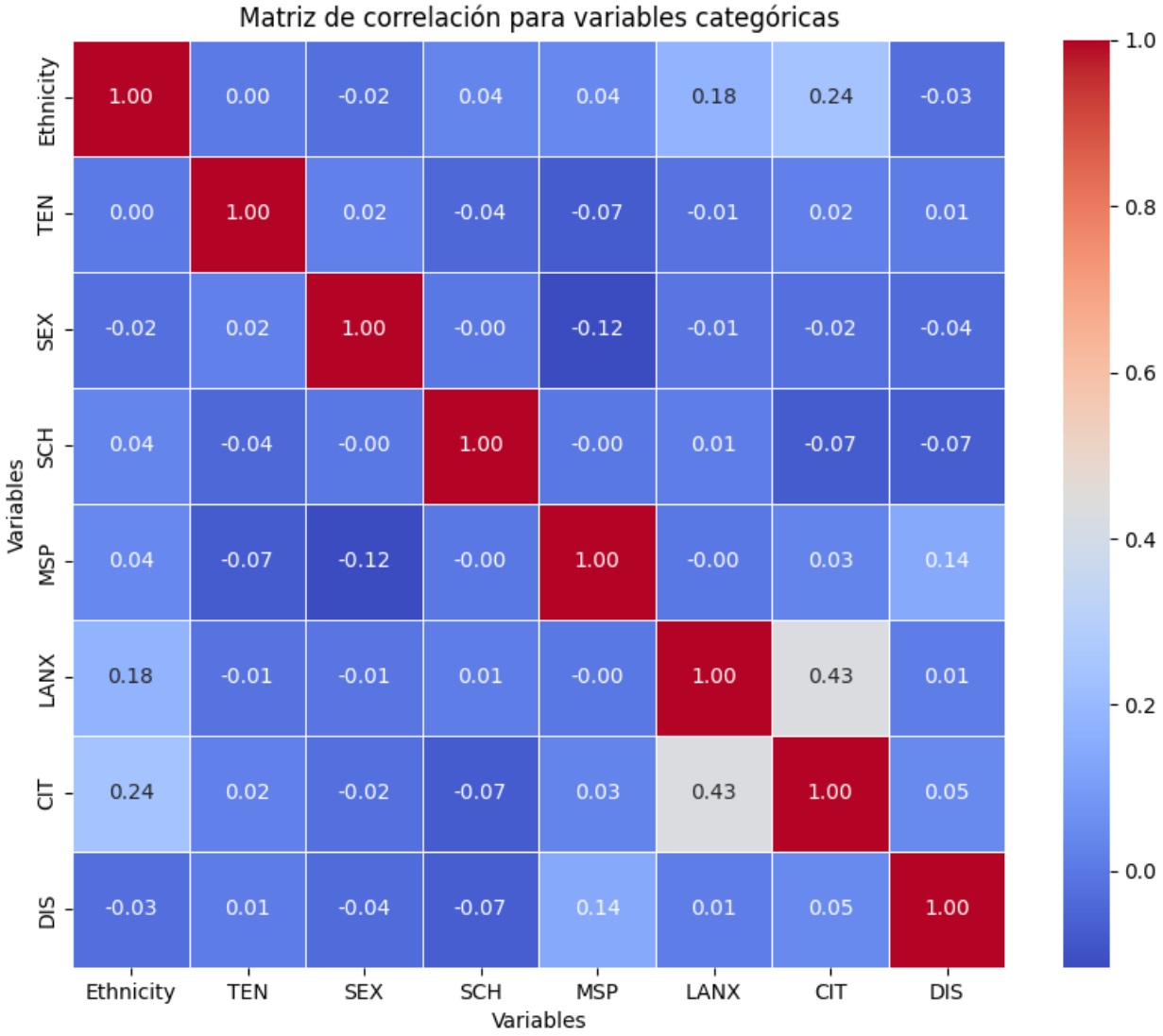
Correlación entre Variables de Ganancias:

Se encontró una correlación significativa entre las variables relacionadas con las ganancias, como "EST_Ficatex" y "PreTaxIncome_PU". Esto sugiere una relación directa entre estas variables, indicando que las variables tipo "EST" están asociadas con los ingresos y gastos de dinero por diversas causas, como finca raíz, cuidado de los niños, salud, entre otras.

Implicaciones de la Correlación:

La alta correlación entre estas variables sugiere que podrían estar proporcionando información redundante al modelo. Dado que las variables relacionadas con las ganancias son fundamentales

para comprender la situación económica de los encuestados, se debe tener cuidado al seleccionar qué variables eliminar para evitar pérdida de información relevante.



Tras realizar la matriz de correlación de las variables categóricas, no se encontraron correlaciones significativas entre estas variables. Esto sugiere que, en el conjunto de datos analizado, no hay relaciones lineales fuertes entre las variables categóricas.

Durante el proceso de análisis y selección de variables para el estudio sobre la pobreza en Nueva York, se identificaron varias variables que fueron descartadas por dos razones principales: alta correlación con otras variables y falta de relevancia para el estudio. Las variables descartadas son:

Variables Descartadas por Alta Correlación:

- EST_FICATEX

- EST_IncomeTax
- NYCgov_Income
- PreTaxIncome_PU
- EST_FICAtax

Estas variables demostraron una correlación significativa con otras variables incluidas en el estudio, lo que sugiere una redundancia en la información que proporcionan.

Variables Descartadas por Falta de Relevancia:

- EST_PovGap
- EST_PovGapIndex
- NYCgov_REL
- NYCgov_Threshold
- Off_Threshold
- Povunit_ID
- Povunit_Rel
- AgeCateg
- DS
- ESR
- HHT
- INTP_adj
- JWTR
- MRGP_adj
- NP
- REL
- RELP
- RELSHIPP
- RETP_adj
- RNTP_adj
- SERIALNO
- WGTP
- WKHP
- WKW
- JWTRNS
- WKWN

Estas variables fueron consideradas no relevantes para el estudio sobre la pobreza en Nueva York y, por lo tanto, fueron eliminadas del conjunto de datos para simplificar el análisis y el modelado.

Set de datos de arrestos:

Filtro por Edad (AGE_GROUP):

El primer filtro se aplicó a la variable "AGE_GROUP", que indica el grupo de edad al que pertenece la persona arrestada. Se decidió aplicar un filtro para incluir únicamente a personas mayores de edad, lo que implica eliminar del conjunto de datos el grupo de edad menor a 18 años.

Este filtro se fundamenta en la necesidad de centrar el análisis en el grupo demográfico de adultos, ya que las leyes y procesos judiciales para menores de edad pueden diferir significativamente de los de los adultos. Además, al centrarse en adultos, se puede obtener una perspectiva más precisa sobre la situación de los arrestos en Nueva York en el contexto de la población adulta.

Al aplicar este filtro, se garantiza que el conjunto de datos resultante esté optimizado para el análisis de los arrestos en Nueva York entre adultos, lo que facilita la identificación de tendencias y patrones relevantes en el análisis posterior.

Transformación 1: Creación de Columna de Meses de Arresto

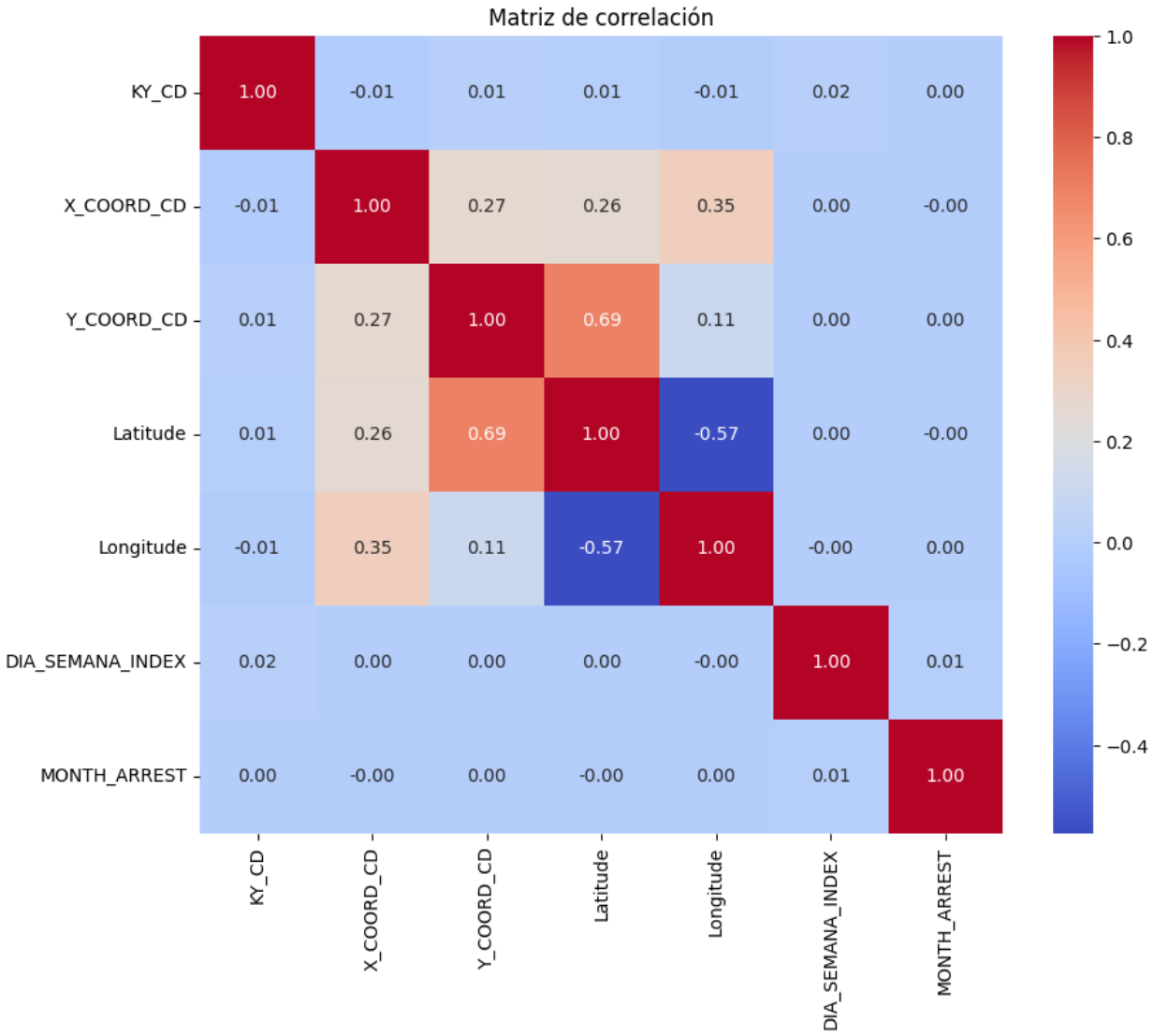
Se creó una nueva columna en el conjunto de datos para separar y registrar los meses en los que se realizaron los arrestos. Esta transformación permite una mejor comprensión y análisis de los patrones temporales de los arrestos en Nueva York, facilitando la identificación de posibles estacionalidades o tendencias a lo largo del año.

Transformación 2: Creación de Columna de Día de la Semana de Arresto

Se creó una nueva columna para diferenciar y registrar el día de la semana en el que se realizaron los arrestos. Esta transformación proporciona información adicional sobre la distribución temporal de los arrestos, lo que puede ser útil para identificar patrones específicos de comportamiento delictivo según el día de la semana.

Al aplicar estas transformaciones, se enriquece el conjunto de datos de arrestos en Nueva York con información temporal adicional, lo que facilita un análisis más profundo y una mejor comprensión de los factores que influyen en los arrestos en la ciudad.

Matriz de correlación:



Tras realizar la matriz de correlación, se encontró que no hay ninguna correlación significativa entre las variables numéricas del conjunto de datos de arrestos en Nueva York. Esto indica que no existe una relación lineal fuerte entre estas variables y que no hay dependencias lineales directas entre ellas.

Set de datos de quejas:

Imputación de datos faltantes:

1. **Reemplazo de valores nulos por la moda**:
   Se identificó la moda de las columnas *ADDR_PCT_CD* (precinto de la dirección) y *BORO_NM*. Los valores nulos en estas columnas se reemplazaron con la moda correspondiente.

2. **Eliminación de filas con valores nulos en columnas críticas**:

Las filas que tenían valores nulos en las columnas *Latitude* y *PD_DESC* fueron eliminadas del conjunto de datos se eliminaron ya que tenían un gran porcentaje de valores nulos.

3. **Imputación de LOC_OF_OCCUR_DESC basada en PREM_TYP_DESC**:
   Para la columna **LOC_OF_OCCUR_DESC** (ubicación del incidente), se asignaron valores basados en los tipos de locales en **PREM_TYP_DESC**. Por ejemplo, si el tipo de local era "STREET", se imputó "FRONT OF". Para locales como "RESIDENCE - APT". HOUSE" o "CHAIN STORE", se imputó "INSIDE", esto se hizo para poder reducir el número de valores en la columna además de que también es una buena forma para reducir el número de valores nulos.

4. **Conversión de formatos de fecha y creación de nueva columna**:
   La columna *CMPLNT_FR_DT* (fecha de la queja) se convirtió a un formato de fecha adecuado y se creó una nueva columna **MONTH_OF_YEAR** que extrae el mes de la fecha de la queja.

5. **Reemplazo de valores nulos en LOC_OF_OCCUR_DESC y PREM_TYP_DESC con la moda**:
   Nuevamente, se calcularon las modas de las columnas **LOC_OF_OCCUR_DESC** y **PREM_TYP_DESC** y se imputaron los valores nulos con estas modas.

6. **Eliminación de columnas con alto porcentaje de nulos**:
   Se eliminaron las columnas **SUSP_AGE_GROUP** y **SUSP_RACE** del conjunto de datos, ya que tenían un valor mayor a 50% de datos faltantes, por lo que era difícil usar un método adecuado para imputar los valores.

7. **Eliminación de valores erróneos en VIC_AGE_GROUP**:
   Se identificaron y eliminaron filas con valores erróneos en la columna **VIC_AGE_GROUP,** ya que había valores negativos, Hasta valores mayores a 1000.

8. **Imputación de VIC_AGE_GROUP con la moda**:
   La moda de la columna **VIC_AGE_GROUP** se calculó y se usó para reemplazar los valores nulos en esa columna.

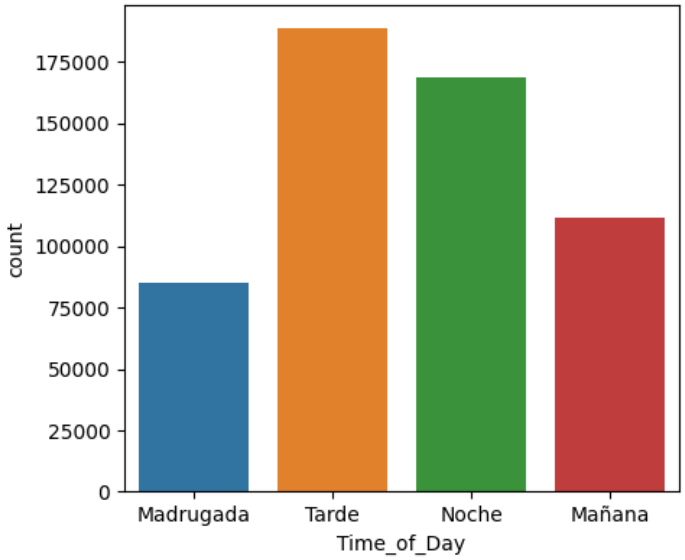
9. **Imputación de VIC_RACE basada en BORO_NM**:
   Se definió una lógica para imputar la raza de la víctima (**VIC_RACE**) basada en el **BORO_NM**. Por ejemplo, si el condado era "STATEN ISLAND", la raza se imputó como "WHITE"; si era "BROOKLYN", se imputó como "BLACK", Esto se realizó investigando cual es la raza más común en ese boro.

10. **Imputación de SUSP_SEX con la moda**:

La moda de la columna **SUSP_SEX** se calculó y se usó para reemplazar los valores nulos en esa columna.

Creación de Variables Derivadas

11. **Categorías de tiempo del día**:
    Se creó una nueva columna **Time_of_Day** para categorizar las horas del día en segmentos como "Madrugada", "Mañana", "Tarde" y "Noche" basándose en la columna **CMPLNT_FR_TM**.



12. **Día de la semana y fin de semana**:
    Se derivaron nuevas columnas para el día de la semana (**Day_of_Week**) y si el día es fin de semana (**Is_Weekend**).

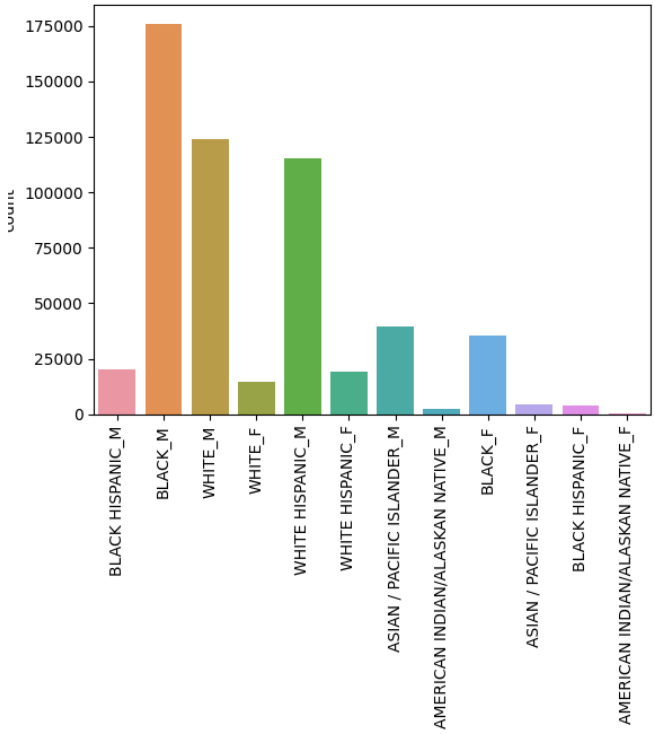
13. **Trimestres del año**:
    Se creó una columna **Quarter** que indica el trimestre del año en el que ocurrió la queja.

14. **Combinación de categorías legales y de ubicación**:
    Se crearon columnas combinadas como **Detailed_Law_Cat** (combinación de la categoría legal y código) y **Location_Desc** (combinación de la descripción del lugar del incidente y el tipo de local). Esto se realizó para poder tener un variable de crímenes más específicos

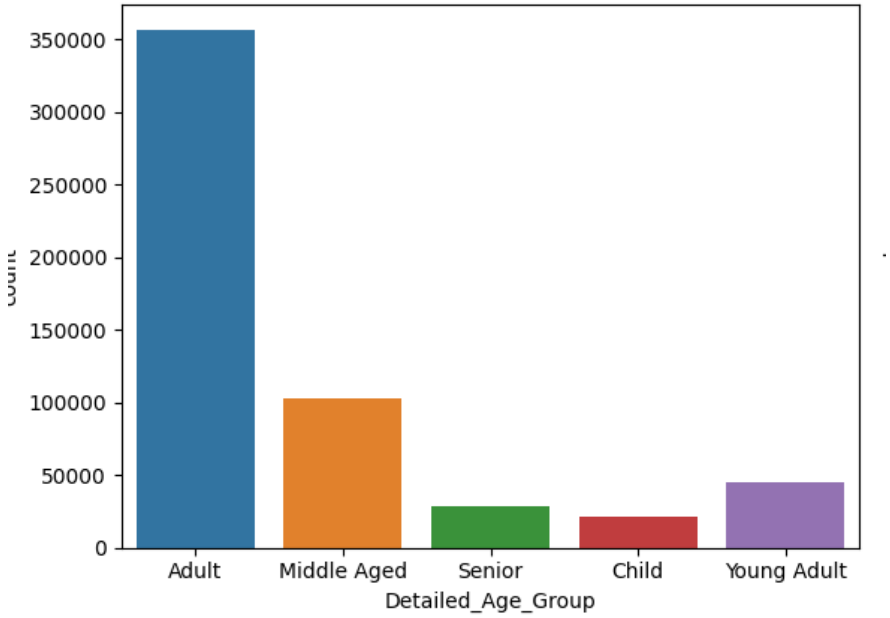
15. **Estado del crimen (completado o intentado)**:
    Se derivó una columna **Is_Completed** para indicar si el crimen fue completado (**COMPLETED**).

16. **Demografía de víctima y sospechoso**:

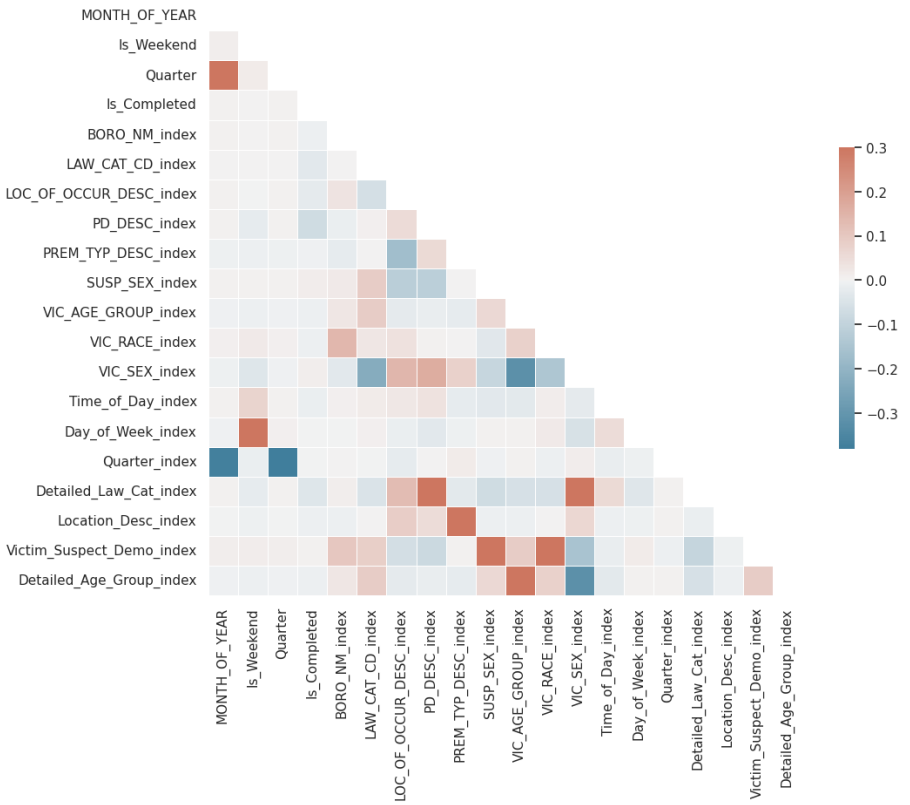
Se creó una columna **Victim_Suspect_Demo** que combina la raza de la víctima y el sexo del sospechoso.



17. **Grupo de edad detallado**:
    Se creó una columna **Detailed_Age_Group** que categoriza los grupos de edad de la víctima en categorías como "Child", "Young Adult", "Adult", "Middle Aged" y "Senior".



Matriz de correlación:



También se realizó la matriz de correlación del conjunto de datos. En este caso las variables que se revisaron fueron las variables luego de haber. Como se puede observar no hay gran correlación entre los datos esto es de esperar ya que al indexar los valores mucha información se pierde.

Técnicas de Inteligencia artificial:

Set de datos de pobreza:

En el análisis de pobreza en Nueva York, se empleó el modelo de RandomForestClassifier como técnica de aprendizaje automático para identificar las características más influyentes en el estado de pobreza de los individuos en la ciudad. Este modelo fue elegido por su capacidad para determinar las características más importantes en un conjunto de datos, particularmente en conjuntos de datos con múltiples variables y muestras.

Se seleccionó un conjunto de variables relevantes, entre ellas "Ethnicity", "AGEP" (edad), "SEX" (género), "NYCgov_Pov_Stat" (estado de pobreza según el gobierno de la ciudad de Nueva York), "Boro" (distrito), "CitizenStatus" (estado de ciudadanía), "EducAttain" (nivel de educación), "FamType_PU" (tipo de familia), "FTPTWork" (tipo de empleo: tiempo completo o tiempo parcial), "TotalWorkHrs_PU" (total de horas de trabajo), "CIT" (condición de ciudadanía), "DIS" (discapacidad), "SCH" (asistencia escolar) y "TEN" (situación de vivienda).

Tras entrenar el modelo, se analizó la importancia de cada característica en la predicción del estado de pobreza. Esto proporcionó información valiosa sobre qué características son más relevantes para identificar a las personas en situación de pobreza en Nueva York. A través de este análisis, se obtuvo una comprensión más profunda de los factores que contribuyen a la pobreza en la ciudad.

Métrica utilizada:

En cuanto a la métrica utilizada para evaluar el modelo de inteligencia artificial, se optó por el MulticlassClassificationEvaluator. Esta métrica es particularmente adecuada para evaluar modelos de clasificación en los que hay más de dos clases o categorías. En el contexto del análisis de pobreza en Nueva York, donde se busca predecir el estado de pobreza de los individuos, esta métrica proporciona una medida comprehensiva del rendimiento del modelo.

El MulticlassClassificationEvaluator evalúa la precisión de la clasificación para cada clase individual, así como la precisión general del modelo. Esto significa que no solo se tiene en cuenta la precisión general del modelo en predecir las clases correctas, sino también la precisión para cada clase específica, lo que permite identificar posibles desequilibrios en la predicción entre las diferentes categorías de pobreza.

Al utilizar el MulticlassClassificationEvaluator, se puede obtener una comprensión más completa del rendimiento del modelo de inteligencia artificial en la tarea de clasificación de la pobreza en Nueva York. Esto permite una evaluación más precisa de la capacidad del modelo para predecir con precisión el estado de pobreza de los individuos y proporciona información valiosa para ajustar y mejorar el modelo en futuras iteraciones.

Resultados y análisis:

En la evaluación del modelo de inteligencia artificial para el análisis de la pobreza en Nueva York, se llevaron a cabo varios experimentos utilizando diferentes hiperparámetros en el RandomForestClassifier. Los resultados fueron evaluados utilizando la métrica MulticlassClassificationEvaluator, específicamente la precisión (accuracy) del modelo.

Primer Experimento

En el primer experimento, se configuró el RandomForestClassifier con los hiperparámetros numTrees=100 y maxDepth=10. Este experimento resultó en una precisión (accuracy) de 0.8681. Esta configuración mostró un buen rendimiento, indicando que el modelo pudo clasificar correctamente el estado de pobreza de los individuos en un 86.81% de los casos.

Segundo Experimento

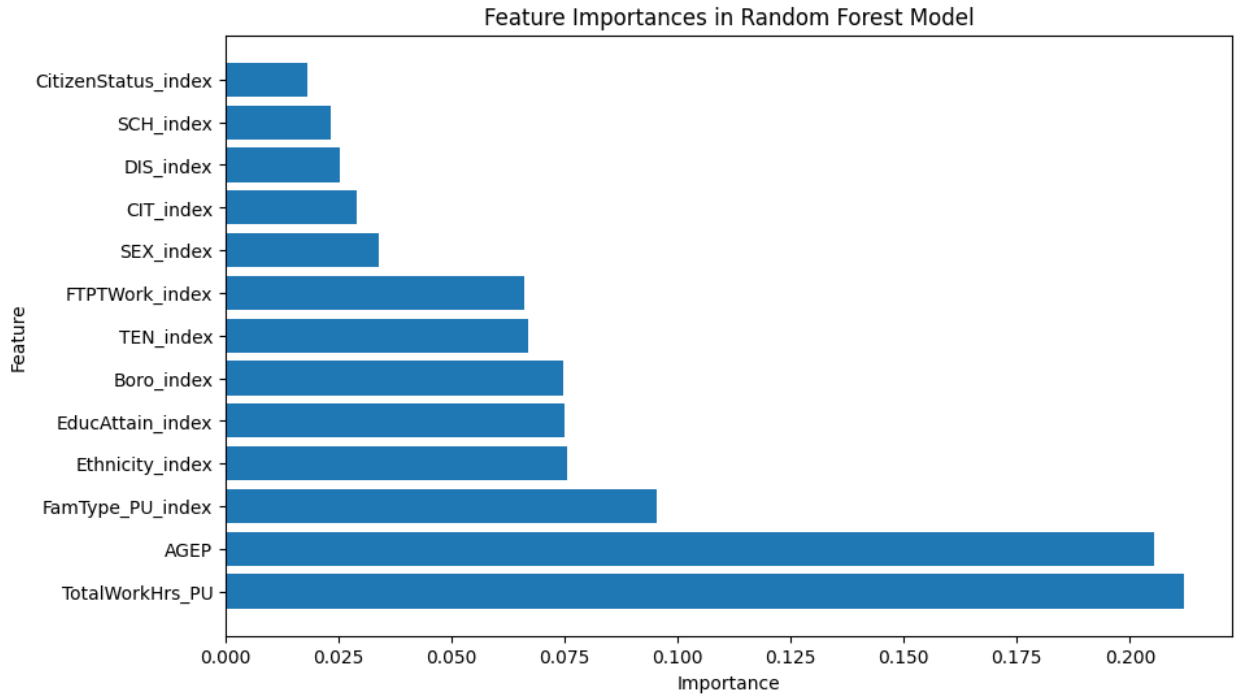
El segundo experimento utilizó una configuración de numTrees=100 y maxDepth=30. El resultado fue una precisión de 0.8661. A pesar de aumentar la profundidad máxima del árbol, la precisión del modelo no mejoró significativamente y de hecho mostró una ligera disminución.

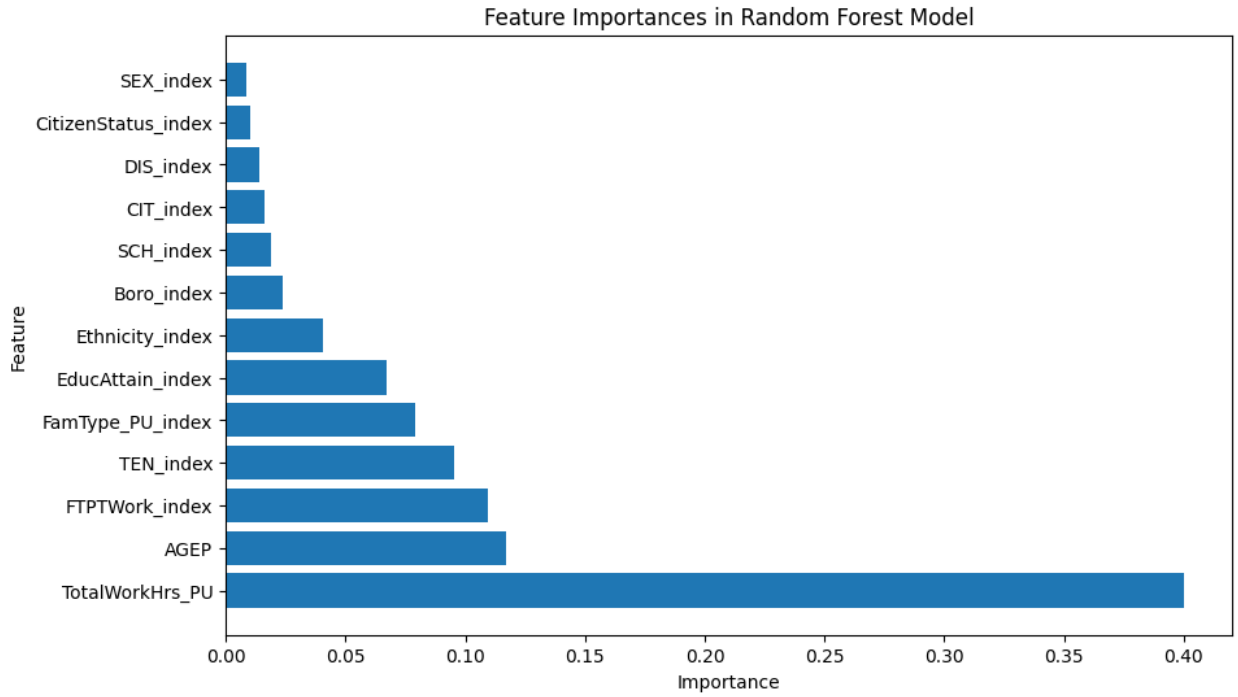
Esto sugiere que incrementar la profundidad más allá de un cierto punto no necesariamente mejora la capacidad predictiva del modelo y puede llevar a un sobreajuste (overfitting).

Tercer Experimento

En el tercer experimento, se establecieron los hiperparámetros en numTrees=50 y maxDepth=30. La precisión obtenida fue nuevamente de 0.8661. Reducir el número de árboles a 50 no afectó la precisión en comparación con el segundo experimento, lo que indica que en este caso, el número de árboles tuvo un impacto menor en la precisión del modelo en comparación con la profundidad del árbol.

Análisis de la problemática:





Los resultados obtenidos mediante el algoritmo de RandomForestClassifier proporcionan una visión detallada de la importancia de las diferentes características en la predicción del estado de pobreza de las personas en Nueva York. Este análisis permite entender cómo varían las condiciones económicas en función de varias características clave.

En el análisis de la importancia de las características, se observó que la característica más influyente es TotalWorkHrs_PU, que representa el total de horas trabajadas. Este resultado sugiere que una mayor cantidad de horas trabajadas está fuertemente asociada con una mejor situación económica. Esto puede ser interpretado como un indicio de que las personas que trabajan más horas tienen mayores ingresos, lo que reduce su probabilidad de encontrarse en situación de pobreza.

Además de las horas trabajadas, otras características también mostraron una influencia significativa en la predicción del estado de pobreza. AGEP (edad) y FamType_PU (tipo de familia) fueron identificadas como variables importantes. La edad puede estar relacionada con la experiencia laboral y, por ende, con el nivel de ingresos. Por otro lado, el tipo de familia puede influir en los recursos disponibles y en la estructura de apoyo económico, afectando así el índice de pobreza.

La identificación de TotalWorkHrs_PU como la característica más importante resalta la relación directa entre el empleo y la pobreza. Aquellos con empleos que permiten más horas de trabajo tienden a tener mejores condiciones económicas, lo que sugiere que políticas que promuevan el empleo y la creación de trabajos con suficientes horas podrían ser efectivas en reducir la pobreza.

La edad y el tipo de familia también aportan información valiosa. Las políticas dirigidas a mejorar las oportunidades laborales para personas en diferentes etapas de su vida y a apoyar diversas estructuras familiares pueden ayudar a mitigar los riesgos de pobreza. Por ejemplo, programas de capacitación y educación continua pueden ser particularmente beneficiosos para aumentar la empleabilidad y los ingresos de individuos de diversas edades.

Set de datos de arrestos:

Para el análisis del conjunto de datos de arrestos en Nueva York, se empleó el algoritmo RandomForestClassifier. Este método se utilizó para identificar las características más importantes que pueden predecir la gravedad del delito cometido por una persona. El análisis se centró en determinar cómo diferentes características influyen en la gravedad del delito, categorizada según el código de la ley (LAW_CAT_CD).

Se estudiaron varias características relevantes, incluyendo PERP_RACE (raza del delincuente), MONTH_ARREST (mes del arresto), DIA_SEMANA_INDEX (día de la semana del arresto), AGE_GROUP (grupo de edad), ARREST_BORO (distrito del arresto) y PERP_SEX (sexo del delincuente). Estas características fueron seleccionadas por su importancia teórica y disponibilidad en el conjunto de datos.

El RandomForestClassifier se utilizó para evaluar la importancia de cada una de estas características. Este algoritmo es robusto y eficaz para identificar relaciones complejas entre variables, y es especialmente útil en conjuntos de datos con múltiples características. Su capacidad para manejar grandes cantidades de datos y proporcionar una medida de la importancia de las características lo hace ideal para este tipo de análisis.

Métricas utilizadas:

Para evaluar el modelo de RandomForestClassifier aplicado al conjunto de datos de arrestos en Nueva York, se utilizaron varias métricas de evaluación proporcionadas por la biblioteca PySpark. Estas métricas permiten una comprensión integral del rendimiento del modelo en términos de precisión, precisión ponderada, recall ponderado y la puntuación F1.

La primera métrica utilizada fue la precisión (accuracy). La precisión mide la proporción de predicciones correctas hechas por el modelo respecto al total de predicciones realizadas. Esta métrica se calculó utilizando el evaluador de clasificación multiclase (MulticlassClassificationEvaluator). La precisión proporciona una visión general del rendimiento del modelo, indicando qué tan bien está prediciendo en general.

Además de la precisión, se evaluó la precisión ponderada (weighted precision). Esta métrica considera la precisión de cada clase y la pondera por el número de instancias en cada clase. Es particularmente útil en conjuntos de datos desbalanceados, donde algunas clases pueden tener muchas más instancias que otras. La precisión ponderada proporciona una medida más

equilibrada del rendimiento del modelo, asegurando que no se favorezca indebidamente a las clases más grandes.

El recall ponderado (weighted recall) también fue una métrica clave en la evaluación. El recall mide la capacidad del modelo para identificar correctamente todas las instancias de una clase en particular, ponderado por el tamaño de cada clase. Esta métrica es crucial para evaluar la sensibilidad del modelo, especialmente en aplicaciones donde es importante capturar todos los casos relevantes de una clase específica.

Por último, se utilizó la puntuación F1 (F1 score), que es la media armónica entre la precisión y el recall. La puntuación F1 proporciona un equilibrio entre precisión y recall, siendo especialmente útil cuando es necesario balancear entre falsos positivos y falsos negativos. Esta métrica es vital para tener una medida única que refleje tanto la precisión como la capacidad del modelo para recuperar las instancias correctas.

Análisis de resultados:

Para evaluar el rendimiento del modelo de RandomForestClassifier en el conjunto de datos de arrestos en Nueva York, se realizaron tres experimentos utilizando diferentes hiperparámetros. A continuación se presenta el análisis de los resultados obtenidos en cada experimento.

Primer Experimento

En el primer experimento, se configuró el modelo con los hiperparámetros numTrees=10. Los resultados obtenidos fueron los siguientes:

- Precisión (Accuracy): 0.5813
- Precisión Ponderada (Weighted Precision): 0.5632
- Recall Ponderado (Weighted Recall): 0.5813
- Puntuación F1 (F1 Score): 0.4781

Estos resultados indican un rendimiento insatisfactorio del modelo. La precisión y las métricas ponderadas no alcanzaron valores elevados, y la puntuación F1, que equilibra precisión y recall, también fue baja, reflejando la incapacidad del modelo para manejar adecuadamente las diferentes clases del conjunto de datos.

Segundo Experimento

En el segundo experimento, se ajustaron los hiperparámetros a numTrees=50. Los resultados fueron:

- Precisión (Accuracy): 0.5810
- Precisión Ponderada (Weighted Precision): 0.5639

- Recall Ponderado (Weighted Recall): 0.5810
- Puntuación F1 (F1 Score): 0.4734

A pesar del aumento en el número de árboles, los resultados no mejoraron significativamente. La precisión y las métricas ponderadas mostraron ligeras variaciones, pero la puntuación F1 continuó siendo baja, indicando que el ajuste no mejoró el equilibrio entre precisión y recall.

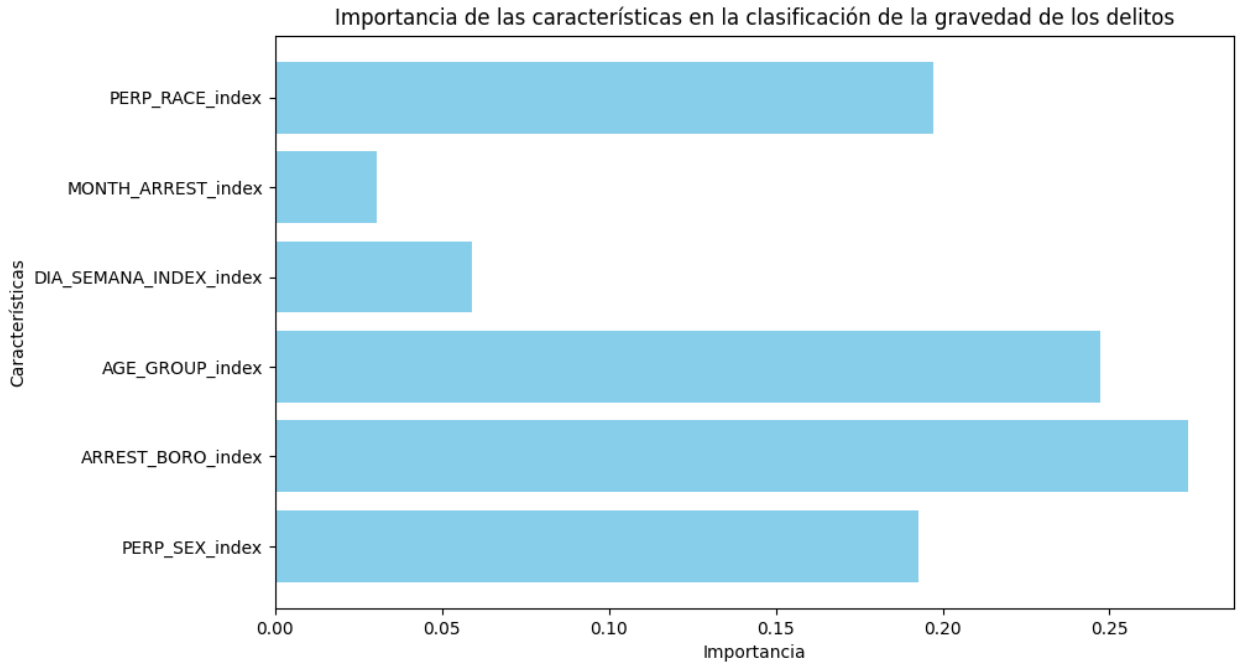
Tercer Experimento

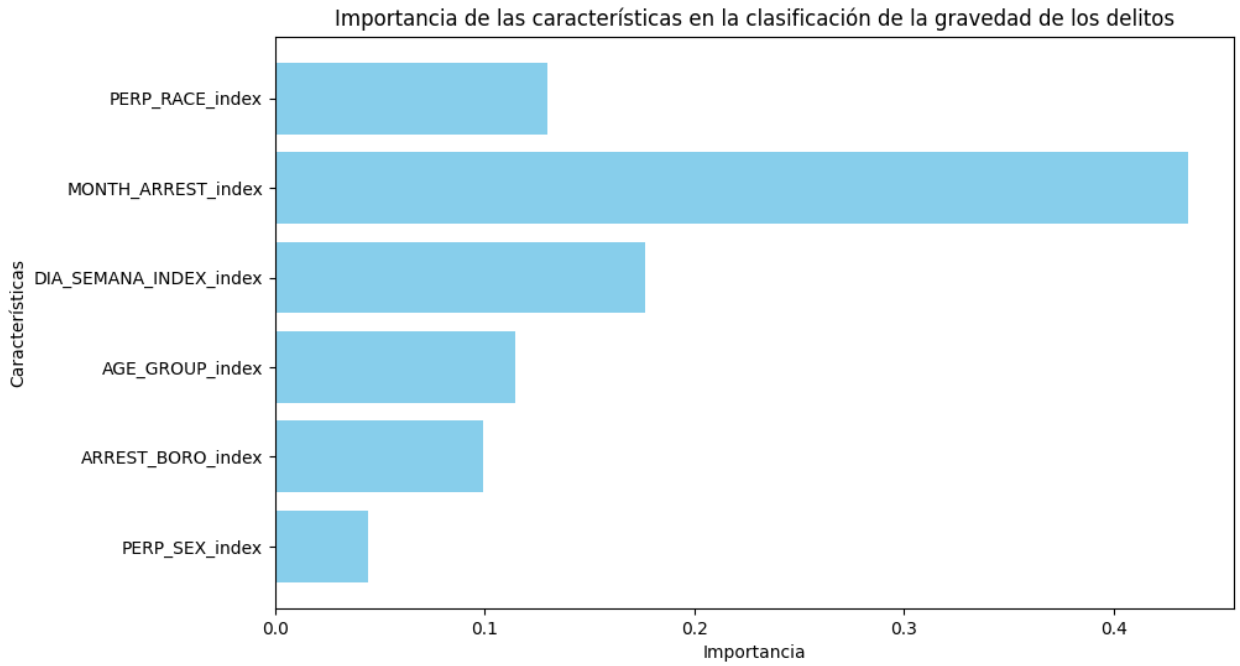
El tercer experimento se llevó a cabo con los hiperparámetros numTrees=20 y maxDepth=20. Los resultados obtenidos fueron:

- Precisión (Accuracy): 0.5705
- Precisión Ponderada (Weighted Precision): 0.5471
- Recall Ponderado (Weighted Recall): 0.5705
- Puntuación F1 (F1 Score): 0.5433

Este experimento también resultó en un rendimiento insatisfactorio. La precisión y las métricas ponderadas fueron incluso más bajas que en los experimentos anteriores. La puntuación F1 mostró una ligera mejora en comparación con los experimentos anteriores, pero todavía no alcanzó un nivel aceptable.

Análisis de la problemática:





El análisis de las gráficas de importancia de características obtenidas a partir de los diferentes experimentos con el modelo de RandomForestClassifier revela inconsistencias en los resultados. En los primeros dos experimentos, se sugiere que ARREST_BORO (el distrito del arresto) es la característica más importante para determinar la gravedad del delito. Sin embargo, en el tercer experimento, se concluye que el MONTH_ARREST (mes en el que se cometió el arresto) es la característica predominante.

Estas diferencias en las características más importantes pueden deberse a la variabilidad en los hiperparámetros utilizados en cada experimento. No obstante, dado que el rendimiento del modelo en todos los experimentos fue insatisfactorio, con métricas como la precisión, la precisión ponderada, el recall ponderado y la puntuación F1 todas en niveles bajos, es razonable cuestionar la fiabilidad de estas conclusiones.

El bajo rendimiento del modelo implica que no tiene una capacidad predictiva adecuada, lo que a su vez sugiere que las características identificadas como importantes no son consistentemente confiables. Por lo tanto, las diferencias observadas en las gráficas de importancia de características no proporcionan una base sólida para concluir que ARREST_BORO o MONTH_ARREST sean efectivamente los factores determinantes de la gravedad del delito.

Dado que las métricas indican que el modelo no está realizando bien su tarea de clasificación, es prudente descartar las conclusiones sobre la importancia de estas características basadas en este

análisis. Para abordar adecuadamente esta problemática, sería necesario mejorar el modelo mediante una optimización más rigurosa de los hiperparámetros, considerar el uso de técnicas de preprocesamiento de datos más avanzadas o explorar otros algoritmos de clasificación que puedan ofrecer un mejor rendimiento en este contexto específico.

Set de datos de quejas:

Primero que todo se pasar los valores categóricos a valores indexados para que puedan ser usados por el modelo de inteligencia artificial

Se usó **OneHotEncoder** para transformar las características categóricas indexadas en vectores binarios, lo que permitió que el modelo de aprendizaje automático pudiera procesarlas.

Para entrenar el modelo se entrenó un modelo de **RandomForestClassifier** utilizando el conjunto de datos de entrenamiento. El modelo se ajustó para predecir **LAW_CAT_CD_index** basado en las columnas de características ensambladas.

Métricas utilizadas Primer experimento:

Hiper parámetros:

numTrees=20, maxDepth=5

Precisión:

La precisión es una métrica que indica la proporción de predicciones correctas realizadas por el modelo en relación con el total de predicciones. Se calcula como:

Para este modelo, la precisión es:

Precisión=0.845

Esto significa que el modelo clasifica correctamente el 84.5% de las instancias en el conjunto de prueba.

Acurracy:

La precisión ponderada tiene en cuenta la precisión de cada clase y la pondera por el número de instancias de cada clase en el conjunto de datos. Es útil en conjuntos de datos desequilibrados donde algunas clases pueden ser más frecuentes que otras. La precisión se calcula para cada clase y luego se hace una media ponderada de estas precisiones.

Para este modelo, la precisión ponderada es:

Precisión Ponderada=0.726

Recall:

El recuerdo es la capacidad del modelo para encontrar todos los casos positivos dentro de una clase específica. Se calcula como:

Para este modelo, el recuerdo es:

Recuerdo=0.845

Esto indica que el modelo es capaz de identificar el 84.5% de las instancias positivas.

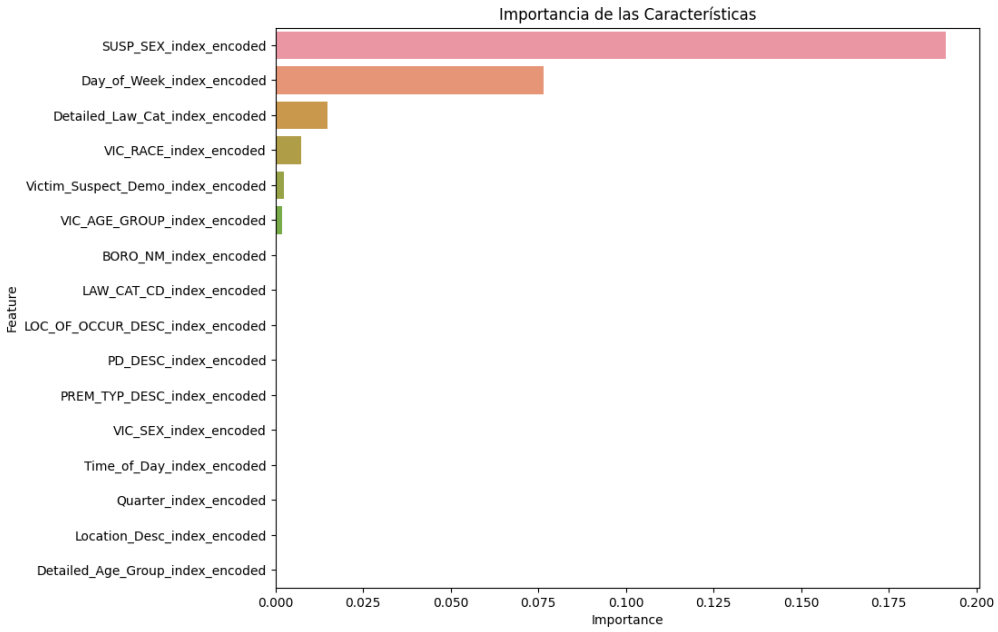
F1-score:

La puntuación F1 es la media armónica de la precisión y el recuerdo. Es una métrica útil cuando se necesita un equilibrio entre precisión y recuerdo, especialmente en conjuntos de datos desequilibrados.

Para este modelo, la puntuación F1 es:

F1 Score=0.778

La combinación de estas métricas y la matriz de confusión sugiere que, aunque el modelo funciona bien en general, hay espacio para mejorar, especialmente en la clasificación de la clase 2



Métricas utilizadas segundo experimento:

Hiper parámetros:

Al ajustar el modelo de Bosque Aleatorio con parámetros diferentes (número de árboles numTrees=50 y profundidad máxima maxDepth=20), los resultados de las métricas mejoraron significativamente

Precisión

Precision= 0.9998

Esto significa que el modelo clasifica correctamente el 99.98% de las instancias en el conjunto de prueba, lo que representa una mejora notable en comparación con el modelo anterior.

Acurracy

Precisión Ponderada=0.9998

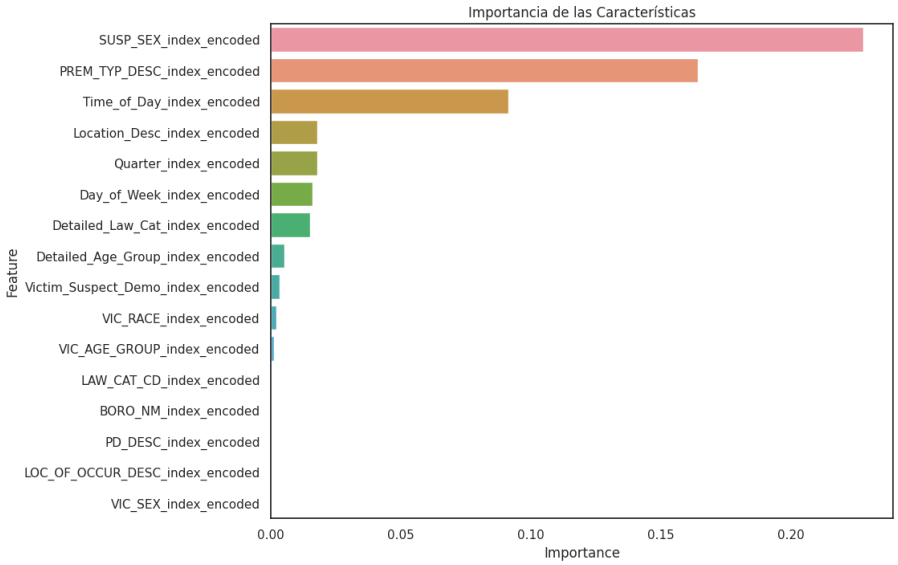
Recall

Recuerdo=0.9998

F1 Score

La puntuación F1 es la media armónica de la precisión y el recuerdo, proporcionando un equilibrio entre ambas métricas.*F1 Score=0.9998*

La mejora de los parámetros del modelo (incrementar el número de árboles y la profundidad máxima) ha resultado en un rendimiento significativamente mejor, con todas las métricas principales (precisión, precisión ponderada, recuerdo y F1 Score) alcanzando valores cercanos a 1.0



Métricas utilizadas Tercer experimento:

En el tercer experimento, se ajustaron los parámetros del modelo de Bosque Aleatorio con un número de árboles (numTrees=20) y una profundidad máxima (maxDepth=10).

Accuracy

Precision=0.9942

Esto significa que el modelo clasifica correctamente el 99.42% de las instancias en el conjunto de prueba.

Precisión:

Precision = 0.9942

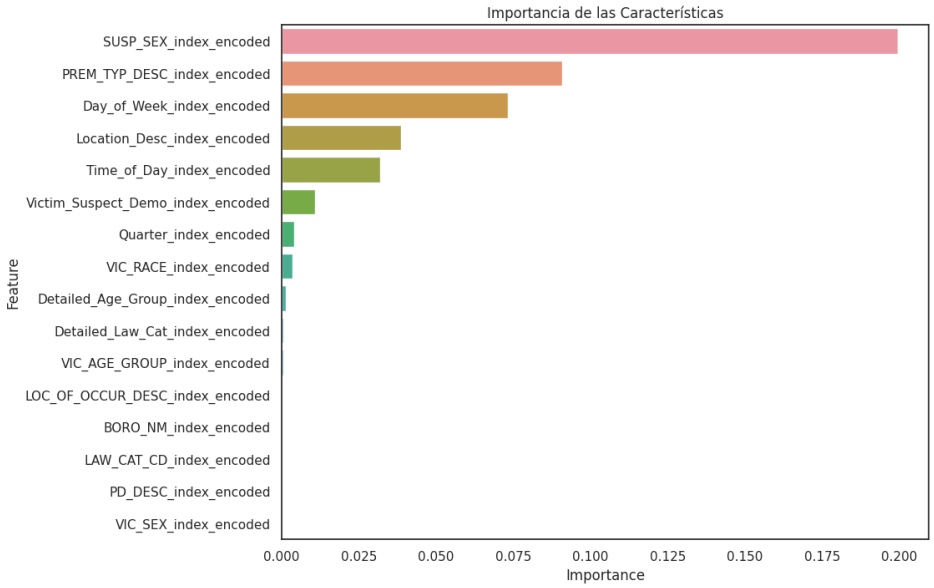
Recall:

Recuerdo=0.9942

F1 Score

F1 Score=0.9941

El tercer modelo muestra un excelente rendimiento general, aunque no tan impresionante como el segundo modelo que casi alcanzó la perfección.



Modelo de Aprendizaje No Supervisado: K-means Clustering

En este último experimento, se entrenó un modelo de aprendizaje no supervisado utilizando el algoritmo de clustering K-means

Hiper parámetros:

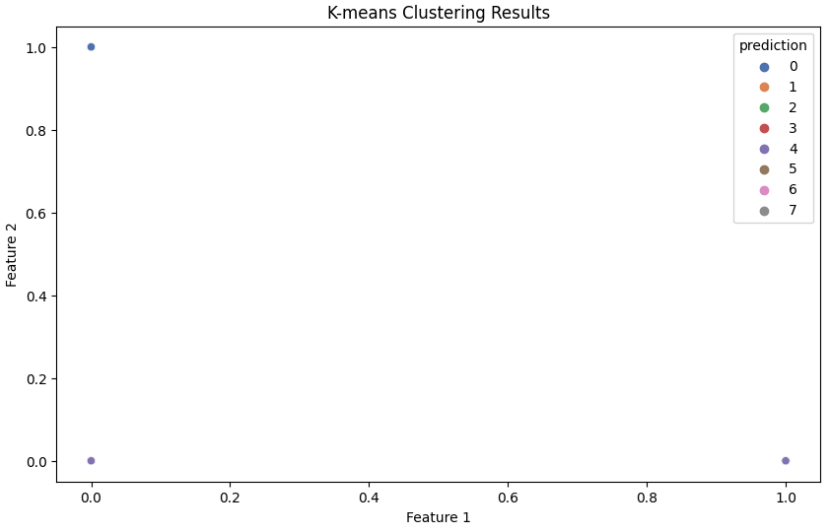
Los parámetros utilizados para entrenar al modelo fueron de k = 2, k=8 y k = 20:

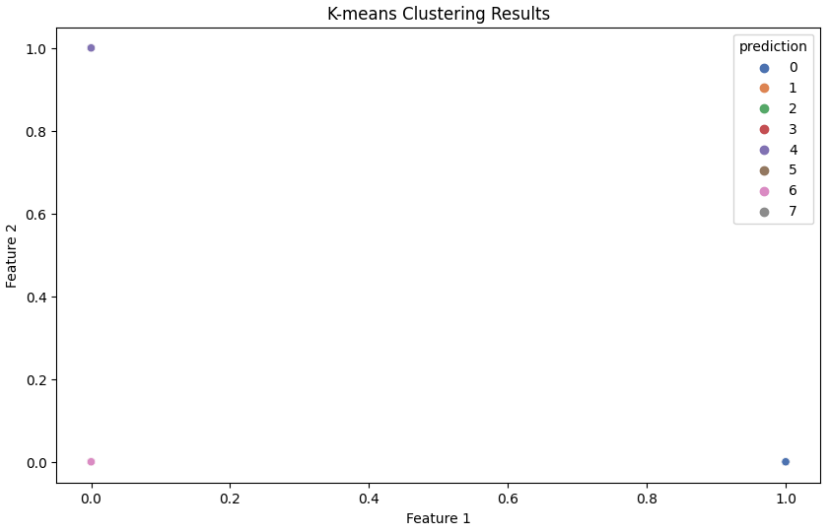
Al entrenar el modelo se usó la métrica de Silhouette:

Valor de 0.1387: Un valor de Silhouette de 0.1387 indica que los clusters formados por el modelo K-means no están claramente separados y solapados.

El valor de la métrica de Silhouette obtenido, 0.1387, indica que los clusters formados por el modelo K-means no están claramente separados y presentan una considerable superposición. Un valor de Silhouette cercano a 1 indica clusters bien definidos y distintos, mientras que valores cercanos a 0 sugieren que los clusters están mezclados y no se distinguen claramente entre sí. En este caso, el valor de 0.1387 revela que las instancias dentro de cada cluster no son significativamente más cercanas a los puntos de su propio cluster en comparación con los puntos de otros clusters, lo que sugiere que el modelo no ha logrado identificar patrones de agrupamiento claros en los datos.

Al grficar los resultados se puede observar que no son muy prometedores:





Resultados finales y conclusiones:

Conclusiones set de datos pobresa:

El análisis del conjunto de datos de pobreza en Nueva York mediante el modelo de RandomForestClassifier ha proporcionado varias conclusiones clave. En primer lugar, la evaluación de importancia de características reveló que TotalWorkHrs_PU (total de horas trabajadas) es la característica más significativa para predecir el estado de pobreza de un individuo. Esto sugiere que una mayor cantidad de horas trabajadas está asociada con una mejor situación económica y una menor probabilidad de estar en pobreza.

Además de las horas trabajadas, otras características como la edad (AGEP) y el tipo de familia (FamType_PU) también demostraron ser importantes en la determinación del estado de pobreza. Estos factores contribuyen significativamente a la variabilidad en la pobreza, indicando que tanto la estructura familiar como la etapa de vida de una persona influyen en su situación económica.

El modelo de RandomForestClassifier utilizado demostró ser efectivo para identificar las características importantes. Sin embargo, se observó que la precisión del modelo varía según los hiperparámetros utilizados. Es crucial optimizar estos parámetros para mejorar la capacidad predictiva del modelo y obtener resultados más robustos.

Las transformaciones aplicadas, como la normalización de la variable PreTaxIncome_PU y la creación de nuevas columnas basadas en el nivel de educación, ayudaron a mejorar la calidad de los datos y la efectividad del modelo. Estas transformaciones permitieron un mejor manejo de las variables y facilitaron la interpretación de los resultados.

A pesar de los hallazgos significativos, el análisis tiene limitaciones. Es posible que otros factores no incluidos en el conjunto de datos también influyan en la pobreza. Se recomienda realizar un análisis más profundo utilizando técnicas adicionales y considerar la inclusión de más variables contextuales que puedan proporcionar una visión más completa del problema.

Conclusiones set de datos arrestos:

El análisis del conjunto de datos de arrestos en Nueva York mediante el modelo de RandomForestClassifier ha revelado varias conclusiones importantes. Inicialmente, se observó una inconsistencia en los resultados obtenidos a partir de los diferentes experimentos, lo que

plantea preguntas sobre la fiabilidad del modelo y la importancia de las características identificadas.

En los primeros dos experimentos, la característica ARREST_BORO (el distrito del arresto) se destacó como la más importante para determinar la gravedad del delito. Sin embargo, en el tercer experimento, MONTH_ARREST (mes en el que se cometió el arresto) fue identificado como el factor predominante. Esta variabilidad en los resultados puede ser atribuida a los cambios en los hiperparámetros del modelo, lo que sugiere una falta de estabilidad y consistencia en la identificación de las características más relevantes.

Las métricas de evaluación del modelo, incluyendo precisión (accuracy), precisión ponderada (weighted precision), recall ponderado (weighted recall) y la puntuación F1 (F1 score), todas reflejaron un rendimiento insatisfactorio. Ninguno de los experimentos logró métricas elevadas, indicando que el modelo no es adecuado para predecir de manera efectiva la gravedad del delito basándose en las características proporcionadas.

Estas limitaciones del modelo sugieren que las conclusiones sobre la importancia de ARREST_BORO o MONTH_ARREST deben ser tomadas con cautela. Dado el bajo rendimiento del modelo, es probable que otros factores no considerados o una mayor optimización del modelo podrían mejorar la capacidad predictiva y proporcionar resultados más fiables.

Para abordar adecuadamente esta problemática, es recomendable realizar una optimización más rigurosa de los hiperparámetros, considerar el uso de técnicas de preprocesamiento de datos más avanzadas y explorar otros algoritmos de clasificación que puedan ofrecer un mejor rendimiento. Además, la inclusión de variables adicionales que puedan influir en la gravedad del delito podría proporcionar una visión más completa y precisa.

Conclusiones set de datos arrestos:

La variable más importante para predecir LAW_CAT_CD resultó ser SUSP_SEX_index, que indica el sexo del sospechoso. Este hallazgo sugiere que el sexo del sospechoso tiene una fuerte relación con el tipo de delito reportado. Además, la segunda variable más importante es PREM_TYP_DESC_index, que describe el tipo de localización del incidente (por ejemplo, residencia, tienda, calle). Esto indica que el entorno en el que ocurre el delito también es un factor crítico para determinar la categoría de ley del incidente. Otras variables como BORO_NM_index (distrito de ocurrencia), VIC_AGE_GROUP_index (grupo de edad de la víctima) y LOC_OF_OCCUR_DESC_index (descripción del lugar de ocurrencia) también mostraron una relevancia considerable en la predicción del tipo de delito.

Rendimiento del Modelo

El modelo de Bosque Aleatorio con parámetros numTrees=50 y maxDepth=20 mostró un rendimiento casi perfecto con una precisión del 99.98% y una puntuación F1 del 99.98%, lo que

indica una capacidad excepcional del modelo para clasificar correctamente las categorías de ley. Otros modelos, aunque menos precisos que el óptimo, también mostraron un rendimiento muy alto. Por ejemplo, con numTrees=20 y maxDepth=10, el modelo logró una precisión del 99.42% y una puntuación F1 del 99.41%. Estos resultados sugieren que el modelo es muy efectivo para la clasificación de los datos de quejas.

Referencias:

FX Empire, «Estados Unidos Nueva-York-Imperio-Estado-Fabricación-Índice 2001-2024 | FX Empire», *FX Empire*, 3 de septiembre de 2021. https://www.fxempire.es/macro/united-states/ny-empire-state-manufacturing-index

«Poverty Simulation | NYCOURTS.GOV». https://ww2.nycourts.gov/ip/nya2j/povertysimulation.shtml#:~:text=In%20New%20York%20State%2C%20under,poverty%20in%202021%20was%2013.9%25.

«Criminal justice Reports & Statistics», *NYS Division Of Criminal Justice Services*. https://www.criminaljustice.ny.gov/crimnet/ojsa/stats.htm

A. Schiller, «New York, NY Crime rates», *NeighborhoodScout*, 19 de marzo de 2024. https://www.neighborhoodscout.com/ny/new-york/crime