Proyecto Procesamiento de Datos a Gran Escala

Alejandro Salamanca, Andrés Salamanca, Alberto Vigna

Índice

Contexto
Selección de los datos
Descripción de datos
Exploración de datos
Reporte de Calidad

Contexto

Nueva York: un estado de transformación y diversidad

- Epicentro cultural, económico y político de EE. UU.
- Diversidad geográfica y demográfica impresionante
- Centro de comercio internacional y turismo

Desafíos económicos y sociales

- Pobreza y desigualdad económica persistente
- Revitalización urbana con beneficios desiguales
- Impacto negativo de la pandemia de COVID-19

Contexto

Población (2020):

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

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

Precio Medio de Vivienda (2020):

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

Contexto

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

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

Ingreso Medio por Hogar (2020):

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

Crecimiento del Empleo

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

OBJETIVO

- Analizar la relación entre pobreza, criminalidad y nivel educativo en Nueva York.
- Recopilación y análisis de datos sobre la distribución geografica de arrestos para identificar tendencias y factores de riesgo.
- Investigación de la relación entre pobreza y delincuencia, examinando la desigualdad económica y el acceso a oportunidades laborales.

Datasets escogidos

Datos de los arrestos en Nueva York

Datos de pobreza en Nueva York

Datos de quejas de la policía de Nueva York actuales

Exploración de datos de los arrestos en Nueva York:

Mapa de los arrestos según longitud y latitud

Numero de arrestos por género

Exploración de datos de los arrestos en Nueva York:

Arrestos por BORO

Arrestos por edad

Exploración de datos de los arrestos en Nueva York:

Arrestos por delito

Arrestos por raza

Exploración de datos de los arrestos en Nueva York:

Arrestos en BORO según el sexo

Arrestos según sexo y raza

Análisis del comportamiento

Exploración de datos de los arrestos en Nueva York:

	count	missing	mean	std dev	zeros	min	median	max	custom
arrest_key	1,000	0%	263M	2.27M	0%	261M	262M	279M	data type: bigint
pd_cd	1,000	0%	423.99	283.91	0%	15	339	969	data type: bigint
ky_cd	1,000	0%	248.42	151.34	0%	101	235	995	data type: bigint
arrest_precinct	1,000	0%	66.68	32.92	0%	1	63	123	data type: bigint
jurisdiction_code	1,000	0%	0.82	7.1	91%	0	0	97	data type: bigint
jurisdiction_code	1,000	0%	0.82	7.1	91%	0	0	97	data type: bigint
x_coord_cd	1,000	0%	1.01M	22.1k	0%	915k	1.01M	1.06M	data type: bigint
y_coord_cd	1,000	0%	207k	31.4k	0%	126k	204k	272k	data type: bigint
latitude	1,000	0%	40.74	0.09	0%	40.51	40.73	40.91	data type: double
longitude	1,000	0%	-73.92	0.08	0%	-74.25	-73.92	-73.73	data type: double

log expand

Exploración de datos de pobreza en Nueva 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

1 Bronx 2 Brooklyn 3 Manhattan4 Queens 5 Staten Island

Mediana de nivel de educación en cada BORO:

1 menos que la escuela secundaria2 Título de escuela secundaria3 algo de universidad4 Licenciatura o superior

Exploración de datos de pobreza en Nueva York:

Declaración de estado de pobreza según raza:

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

Total de horas trabajadas por raza

Exploración de datos de pobreza en Nueva York:

Horas trabajadas según edad

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

Horas trabajadas por BORO

1 Bronx 2 Brooklyn 3 Manhattan4 Queens 5 Staten Island

Exploración de datos de pobreza en Nueva York:

Porcentaje de experiencia laboral del individuo

1 tiempo completo todo el año

2 Menos de tiempo completo durante todo el año

3 Sin trabajo

Exploración de datos de pobreza en Nueva York:

Porcentaje de niveles de educación

1 menos que la escuela secundaria

2 Título de escuela secundaria

3 algo de universidad

4 Licenciatura o superior

Análisis de comportamiento

Exploración de datos de pobreza en Nueva York:

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

Análisis de comportamiento

Exploración de datos de pobreza en Nueva York:

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

Exploración de Quejas en Nuevo York:

Total, de quejas por BORO

Raza del sospechoso agrupado por BORO

Brooklyn (azul), Manhattan (naranja), Queens (amarillo), Bronx (azul claro) y Staten Island (rojo).

Exploración de Quejas en Nuevo York:

Crimen fue completado o no

Lugares en los que ocurren los crímenes

Exploración de Quejas en Nuevo York:

Tipo de crimen:

Misdemeanor (delito menor): Crimen menos grave, penas incluyen multas o hasta un año de cárcel.

Felony (delito grave): Crimen grave, penas pueden incluir más de un año de cárcel.

Violation (infracción): Infracción menor, penas suelen ser multas sin cárcel.

Exploración de Quejas en Nuevo York:

Mapa cantidad de crímenes Brooklyn:

Mapa cantidad de crímenes Bronx:

Mapa cantidad de crímenes Queens:

REPORTES DE CALIDAD

Los reportes de calidad son cruciales para garantizar la fiabilidad y la precisión de los análisis y decisiones basadas en datos.

Dataset de Arrestos en NY

- Evaluación de la calidad de los datos usando “select” de Spark y aplicando expresiones para identificar valores nulos o invalidos por columna, usando “when(isnan(c) | col(c).isNull(), c)”
- La mayoría de las columnas presentan una baja incidencia de valores nulos o inválidos, indicando una alta integridad en gran parte del dataset.
- Se identificaron valores faltantes específicos en columnas críticas como "PD_CD", "KY_CD", "OFNS_DESC", y "LAW_CAT_CD"

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

Dataset de Pobreza en NY

- Evaluación de la calidad de los datos usando “select” de Spark y aplicando expresiones para identificar valores nulos o invalidos por columna, usando “when(isnan(c) | col(c).isNull(), c)”
- La mayoría de las columnas presentan una alta incidencia de valores nulos o inválidos, indicando una baja integridad en gran parte del dataset.
- En todas las columnas hay valores faltantes

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

Únicos columnas con datos faltantes

Estrategias de limpieza de valores nulos o inválidos:

Objetivo: Mejorar la integridad y utilidad de los datos mediante la aplicación de técnicas para tratar valores faltantes

- **Técnicas de Limpieza Aplicadas:**
  - **SCHL & EducAttain (Nivel Educativo):** Imputación por la moda, usando el grado escolar más reportado.
  - **ESR (Estado de Empleo):** Imputación por la moda o la categoría de empleo más común. Alternativamente, uso de la media o mediana para datos no escasos.
  - **LANX (Idioma en Casa):** Imputación por la moda, seleccionando el idioma más comúnmente hablado.
  - **ENG (Dominio del Inglés):** Imputación por la moda con el nivel de habilidad en inglés más frecuentemente reportado.
  - **MSP (Estado Civil):** Imputación por la moda, utilizando el estado civil más frecuente.
  - **WKW (Semanas Trabajadas):** Imputación por la media o mediana, estimando las semanas trabajadas basándose en el comportamiento general de la población.
  - **JWTR (Modo de Transporte al Trabajo):** Imputación por la moda, eligiendo el medio de transporte más usado.

Planteamiento de preguntas

1

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

2

¿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?

3

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

Planteamiento de preguntas

4

¿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?

5

¿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?

Planteamiento de preguntas

6

¿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?

7

¿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?

8

¿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?

FILTROS, LIMPIEZA Y TRANSFORMACIÓN

DATASET DE ARRESTOS

- Cálculo de la moda de las columnas "LAW_CAT_CD" y "PERP_RACE".
- Agrupación de datos por cada categoría única en estas columnas y conteo de la frecuencia de cada una.
- Selección de la categoría con la mayor frecuencia como la moda.
- Reemplazo de valores nulos en estas columnas con las modas correspondientes utilizando la función withColumn.
- Sustitución de celdas vacías en "LAW_CAT_CD" por la moda previamente calculada de esa columna.
- Sustitución de valores nulos en "PERP_RACE" por la moda correspondiente.
- Eliminación de filas que contienen valores faltantes en columnas críticas como "PD_DESC", "KY_CD", "OFNS_DESC" y "LAW_CODE".

DATASET DE POBREZA

- Identificación de valores nulos
- Análisis de porcentajes de valores nulos
- Eliminación de columnas con valores nulos significativos ("JWTR", "ENG" y "WKW")
- Tratamiento de valores nulos en la columna "SCHL" (Nivel de Educación)

LIMPIEZA DEL SET DE DATOS DE QUEJAS

- Eliminación de columnas con valores nulos predominantes: TRANSIT_DISTRICT, STATION_NAME, PARKS_NM, HOUSING_PSA y HADEVELOPT gran proporción de valores nulos.

- Imputación de valores faltantes utilizando la moda: ADDR_PCT_CD y BORO_NM utilizando la moda.

- Eliminación de filas con ubicación desconocida (NULL): Filas donde la longitud y latitud eran desconocidas.

- Llenado de valores en la columna LOC_OF_OCCUR_DESC basado en PREM_TYP_DESC: Se llenaron los valores faltantes en LOC_OF_OCCUR_DESC utilizando información de PREM_TYP_DESC.

- Creación de la nueva característica MONTH_OF_YEAR: Permite analizar la estacionalidad de los incidentes para identificar patrones y planificar medidas de seguridad o recursos de manera más efectiva.

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