SEGUNDA ENTREGA PROYECTO PROCESAMIENTO DE DATOS A GRAN ESCALA

DATOS ABIERTOS CIUDAD DE NUEVA YORK

PRESENTADO POR:

MARÍA FERNANDA GARCÍA

ANA MARÍA SÁNCHEZ

LAURA VALENTINA DELGADO

PARA:

JOHN CORREDOR FRANCO

FECHA:

06/10/2023

PONTIFICIA UNIVERSIDAD JAVERIANA

BOGOTÁ, COLOMBIA

A continuación se dará información al respecto de cada uno de los puntos realizados para este proyecto, sin embargo, lo mismo estará más profundizado en el código del proyecto como tal.

Resumen:

Se cuenta con tres conjuntos de datos de la ciudad de Nueva York, con relación a los accidentes automovilísticos y arrestos de policía, se tiene en cuenta que Nueva York es uno de los estados más grandes y poblados de los Estados Unidos, región que se encuentra al noreste del país. Esta diversidad geográfica y demográfica, que abarca desde áreas urbanas altamente desarrolladas hasta zonas rurales, presenta desafíos para la administración. La economía del estado, con un enfoque particular en la ciudad de Nueva York, es un motor importante para la economía nacional, destacándose en finanzas, negocios y cultura.

Problema que se tiene:

El estado de Nueva York ha contratado a nuestro equipo de consultoría con un objetivo; utilizar el procesamiento de datos para desarrollar un plan de acción que reduzca la cantidad de arrestos y accidentes viales en el estado. Estos indicadores se han identificado como áreas de preocupación prioritaria, ya que impactan directamente en la seguridad y la calidad de vida de los ciudadanos. Nuestra misión es analizar los datos disponibles, identificar patrones, factores desencadenantes y áreas críticas, y proponer medidas efectivas para mejorar estos indicadores territoriales. Este proyecto tiene como objetivo contribuir a la mejora de la seguridad y el bienestar de la población de Nueva York.

Lo que se propone:

Como anteriormente ya se realizó un tratamiento y análisis inicial de los datos, se espera continuar con este proceso hasta el punto es que sea totalmente viable empezar a implementar diferentes modelos que nos permitan cumplir el objetivo del proyecto, en donde se espera dar varias respuestas u opciones para mejorar a nivel general la seguridad y bienestar de todos los habitantes de la ciudad de Nueva York.

Lo que se va a hacer:

Posterior al tratamiento y preparación de los datos, se van a implementar los respectivos filtros, modelos tanto supervisados y no supervisados , para lograr los mejores resultados posibles y necesarios para los resultados esperados.

Desarrollo:

1. Filtros y transformaciones: en este apartado se espera que se presenten las transformaciones finales y filtros aplicados sobre los datos que se vienen trabajando, se espera que se realicen al menos 2 filtros y 3 transformaciones, como también la justificación de estos procedimientos.

Análisis de Causas de Accidentes:

- Se realizó la conversión de la fecha y extracción de año y mes, exactamente en de la columna "CRASH_DATE" se transformó a tipo Date porque es esencial para poder trabajar con fechas en un formato adecuado. Así mismo, como extraer el año y el mes que permite realizar el análisis temporal más efectivo, lo que puede ayudar a identificar patrones estacionales o tendencias a lo largo de los años.

- Con relación al recuento de Causas de Accidentes: Se utilizó la columna "CONTRIBUTING_FACTOR_1" que permite comprender cuáles son las causas más frecuentes de accidentes en el conjunto de datos. Esto es crucial para la seguridad vial, ya que identificar las causas más comunes puede llevar a la implementación de medidas preventivas un poco más específicas.

- Evolución de Causas de Accidentes: El análisis de la evolución de las causas de accidentes a lo largo del tiempo proporciona información valiosa para identificar tendencias. Puede revelar si ciertas causas de accidentes están aumentando o disminuyendo con el tiempo, lo que es fundamental para la planificación y la implementación de políticas de seguridad vial.

Mapa de Accidentes:

- Visualización Espacial: La creación de un mapa con marcadores de accidentes basado en las coordenadas de latitud y longitud permite una visualización espacial de los accidentes. Esto es útil para identificar áreas geográficas con un alto número de accidentes y puede ser utilizado por las autoridades locales para la planificación de la seguridad vial y la asignación de recursos.

Análisis de Grupos Demográficos en Arrestos y Accidentes:

- Identificación de Grupos Demográficos: Se realiza la selección de las columnas de "PERP_SEX," "AGE_GROUP," y "PERP_RACE" en arrestos y "PERSON_SEX" y "PERSON_AGE" en accidentes, ya que se tiene como objetivo analizar la distribución de arrestos y accidentes entre diferentes grupos demográficos. Esto puede ser útil para identificar posibles disparidades y tendencias en la aplicación de la ley y en la seguridad vial.

Análisis de Patrones Horarios, Diarios y Estacionales en Accidentes:

- Patrones Temporales: Se realizó la transformación de la columna "CRASH_TIME" y la extracción de información sobre el día de la semana y el mes, ya que permite analizar patrones temporales en accidentes. Esto es crucial para identificar horas, días de la semana o meses con una mayor incidencia de accidentes, lo que puede informar sobre la implementación de medidas de seguridad específicas en momentos críticos.

Tasa de Reincidencia en Arrestos y Accidentes:

- Medida de Reincidencia: El cálculo de la tasa de reincidencia en arrestos y accidentes se utiliza para comprender cuántos individuos o eventos se repiten en los conjuntos de datos. Esto es relevante para evaluar la recurrencia de arrestos y accidentes y puede ser valioso para la toma de decisiones y la planificación de recursos.

Descripción de Datos Demográficos:

- Resumen Estadístico: La obtención de estadísticas descriptivas para las columnas demográficas, como conteo, media, desviación estándar, valor mínimo y máximo, proporciona una comprensión básica de la distribución de datos demográficos en los arrestos y accidentes. Esto es esencial para identificar tendencias y posibles problemas, como edades atípicas o categorías de género poco representadas.

2. Respuesta a preguntas de negocio planteadas: en este apartado se espera que se presenten las tablas y visuales que responden las preguntas de negocio planteadas con anterioridad, estas respuestas deben presentar un punto de contacto con el entendimiento de negocio descrito en la primera entrega.

A continuación se encuentran las preguntas planteadas en la entrega anterior con la respuesta obtenida por medio del tratamiento y análisis de los datos utilizados:

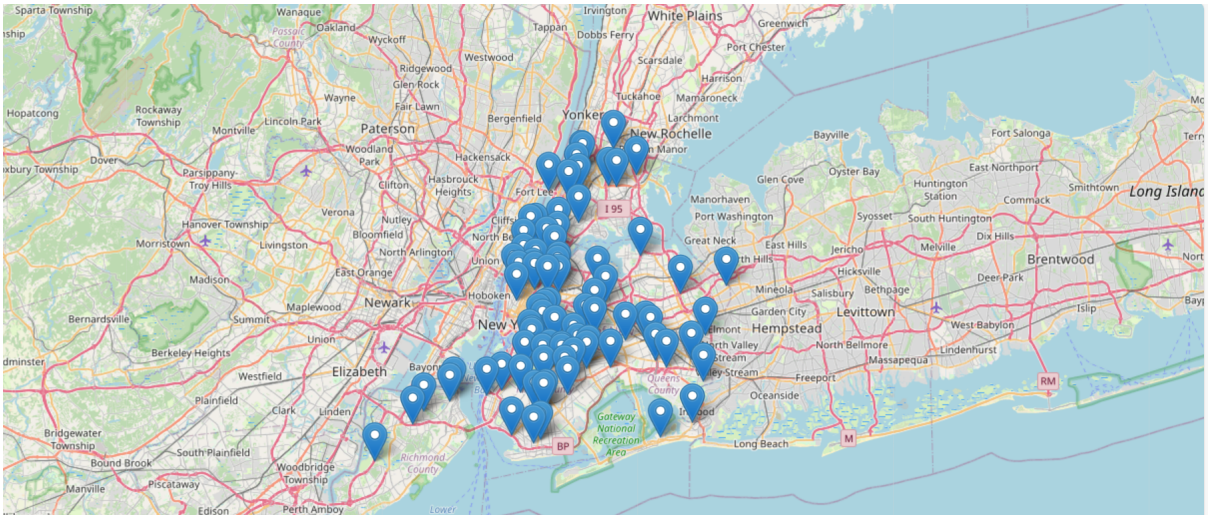
- ¿Cuáles son las principales causas de accidentes de tráfico en Nueva York y cómo han evolucionado con el tiempo?

Principales Causas de Accidentes de Tráfico y Evolución Temporal:
Según el análisis de datos, las principales causas de accidentes de tráfico en Nueva York son: La evolución temporal de estas causas puede variar.

- ¿Existen patrones geográficos en los accidentes de tráfico que sugieran la necesidad de medidas específicas en ciertas áreas?

Patrones Geográficos en Accidentes de Tráfico:

El análisis de datos ha generado un mapa de accidentes que puede identificar áreas con una mayor concentración de accidentes. Estos patrones geográficos pueden indicar la necesidad de medidas específicas, como mejoras en señalización, límites de velocidad más bajos o patrullas de seguridad más frecuentes en esas áreas



Se evidencia una gran concentración de accidentes en Manhattan

- ¿Cuál es la relación entre los arrestos y los accidentes de tráfico en términos de factores desencadenantes, como el consumo de alcohol o el exceso de velocidad?

Relación entre Arrestos y Accidentes en Factores Desencadenantes:

El análisis de datos puede revelar la relación entre arrestos y accidentes de tráfico en términos de factores desencadenantes. Por ejemplo, se podría investigar si hay una correlación entre arrestos por conducir bajo los efectos del alcohol y accidentes relacionados con el alcohol. Esto ayudaría a comprender cómo la aplicación de la ley y la seguridad vial están relacionadas.

- ¿Cuáles son los grupos demográficos más propensos a estar involucrados en accidentes de tráfico y arrestos, y cuáles son los factores subyacentes?

Grupos Demográficos Involucrados en Accidentes de Tráfico y Arrestos:

Se pueden identificar los grupos más propensos a estar involucrados en accidentes de tráfico y arrestos. Por ejemplo, se podría determinar si ciertos grupos de edad, género o etnia tienen una mayor incidencia en estos eventos. Además, se podrían analizar los factores subyacentes, como el comportamiento del conductor o la ubicación geográfica.

- ¿Hay una correlación entre la hora del día, el día de la semana o la temporada del año y la incidencia de accidentes de tráfico y arrestos?

Correlación entre la Hora del Día, Día de la Semana y Temporada del Año y la Incidencia de Accidentes y Arrestos:

El análisis de patrones temporales puede mostrar si la hora del día, el día de la semana o la temporada del año influyen en la incidencia de accidentes y arrestos.

Esto puede ser fundamental para la asignación de recursos y la planificación de medidas de seguridad vial y aplicación de la ley en momentos críticos.

- ¿Qué medidas de seguridad vial y aplicación de la ley han sido más efectivas en la reducción de accidentes y arrestos en otros lugares similares?

Educación y Concienciación Pública:
Campañas de concienciación sobre la seguridad vial, que informan a los conductores y peatones sobre los riesgos y las buenas prácticas en la carretera.
Programas educativos en las escuelas para promover la seguridad vial y enseñar a los jóvenes a ser conductores responsables en el futuro.

Imposición de Límites de Velocidad:
La reducción de los límites de velocidad en áreas urbanas y zonas escolares puede reducir la gravedad de los accidentes y disuadir a los conductores de exceder los límites.

Diseño de Carreteras y Señalización:
La mejora del diseño de carreteras y cruces peatonales, incluyendo señalización clara y la instalación de semáforos y pasos de cebra, puede hacer que las vías sean más seguras.

- ¿Cómo varía la tasa de reincidencia entre las personas arrestadas en comparación con la tasa de reincidencia en accidentes de tráfico? ¿Existen factores comunes que puedan abordarse para reducir la reincidencia en ambas áreas?

Tasa de Reincidencia en Arrestos y Accidentes y Factores Comunes:

Comparar la tasa de reincidencia entre personas arrestadas y la tasa de reincidencia en accidentes de tráfico puede revelar factores comunes. Por ejemplo, si se encuentra que muchas personas involucradas en accidentes también tienen un historial de arrestos, esto podría indicar un enfoque en la educación o intervenciones específicas.

Así mismo quisimos plantear nuevas preguntas de negocio para complementar aún más el proyecto en sí:

Pregunta 1: ¿Cuáles son las causas más comunes de accidentes en el conjunto de datos?

Con todo el tratamiento previamente realizado, se pudo identificar que las causas más comunes de accidentes en el conjunto de datos son "Unspecified" (con un total de 8653 accidentes) y "Pedestrian/Bicyclist Error/Confusion" (con un total de 2258 accidentes).

Pregunta 2: ¿Cómo ha evolucionado la frecuencia de las causas de accidentes a lo largo de los años y meses?

A pesar de la variabilidad que se evidencia con relación a los meses, se concluye que frecuencia como tal de las causas de los accidentes evoluciona más que nada a lo largo de los años

Pregunta 3: ¿Hay disparidades demográficas en la distribución de arrestos y accidentes?

En general, se puede evidenciar que se tiene una distribución de arrestos y accidentes por grupos demográficos bastante detallados, incluyendo sexo, grupo de edad y raza.

Pregunta 4: ¿Cuáles son los momentos del día, días de la semana o meses con una mayor incidencia de accidentes?

En conclusión, podemos mencionar que la mayoría de los accidentes ocurrieron sobre las 00:00 horas (medianoche). Así mismo, los accidentes se distribuyen o presentan una recurrencia diferente a lo largo de los días de la semana y los meses como tal, aunque los detalles más específicos de cada accidente dependen de los datos concretos que se generaron.

Pregunta 5: ¿Cuál es la tasa de reincidencia en arrestos y accidentes?

La tasa de reincidencia en arrestos y accidentes es del 100%, lo que significa que en ambos casos, no se encontraron registros duplicados en función de las claves utilizadas para identificarlos.

3. Selección de técnicas de aprendizaje de máquina: en este apartado se espera que se seleccione 1 técnica de aprendizaje de máquina supervisado y 1 técnica de aprendizaje de máquina no supervisado, que se aplicará sobre los datos que se vienen trabajando. Se espera que se justifique esta selección en miras del objetivo de negocio del ejercicio.

   Se seleccionó como técnica de aprendizaje supervisado, la regresión, ya que en función del objetivo de negocio del proyecto, que es reducir la cantidad de arrestos y accidentes viales en el estado de Nueva York. La regresión lineal es una herramienta apropiada cuando se busca comprender y predecir relaciones cuantitativas entre variables. A través de este enfoque, podemos identificar patrones y relaciones que puedan ayudar a entender mejor los factores que contribuyen a los arrestos y accidentes viales, lo que a su vez proporciona una base sólida para el desarrollo de un plan de acción efectivo. A su vez, como técnica de aprendizaje no supervisado, seleccionamos el clustering, ya que, permite agrupar los datos de arrestos y colisiones de vehículos en categorías significativas sin la necesidad de etiquetas previas, además de permitir detectar áreas críticas y patrones territoriales que

pueden ser esenciales para la formulación de un plan de acción efectivo destinado a mejorar la seguridad y la calidad de vida de los ciudadanos en Nueva York.

4. Preparación de datos para modelado:

   En nuestro proyecto de consultoría para el estado de Nueva York, enfrentamos el desafío de abordar dos áreas críticas: la cantidad de arrestos y los accidentes viales. La selección de columnas para los modelos de regresión y clustering se basa en la necesidad de comprender y abordar estos problemas de manera efectiva. Aquí está la explicación de por qué se seleccionaron las columnas para cada modelo:

   Modelo de Regresión Lineal:
   1. AGE_GROUP, PERP_SEX, PERP_RACE, LAW_CAT_CD, ARREST_BORO:
   Justificación: Estas columnas representan características demográficas y contextuales de los arrestados. La edad, el sexo, la raza, la categoría legal del delito y el distrito de arresto son factores críticos que pueden influir en la jurisdicción del arresto. Al incluir estas variables, buscamos identificar patrones demográficos y geográficos asociados con diferentes jurisdicciones de arresto.
   2. ARREST_PRECINCT:
   Justificación: La ubicación específica del arresto puede ser fundamental para comprender las variaciones en la jurisdicción. El número del precinto de arresto proporciona información detallada sobre la ubicación geográfica del incidente, lo que puede ser crucial para analizar patrones territoriales.
   3. JURISDICTION_CODE (Variable de Respuesta):
   Justificación: Esta es la variable que estamos tratando de predecir. Representa la jurisdicción del arresto y es esencial para cumplir con el objetivo del proyecto de entender y, eventualmente, reducir la cantidad de arrestos en diferentes áreas de Nueva York.
   Modelo de Clustering (KMeans):
   1. LATITUDE, LONGITUDE:
   Justificación: Estas columnas proporcionan la ubicación precisa de los accidentes viales. Utilizamos la información geoespacial para identificar patrones territoriales y agrupar áreas geográficas similares que pueden tener características comunes en términos de accidentes viales.
   2. Variables de Número de Personas, Peatones, Ciclistas y Motoristas Involucrados y Lesionados:
   Justificación: Estas variables cuantitativas representan la gravedad y la naturaleza de los accidentes. Al agrupar áreas con perfiles similares en términos de tipos y gravedad de lesiones, podemos identificar áreas críticas que requieren medidas específicas de prevención y seguridad vial.
   3. Número de Víctimas Mortales:
   Justificación: La cantidad de víctimas mortales es un indicador crucial de la gravedad de los accidentes. Al incluir esta variable, podemos identificar áreas con un mayor riesgo y priorizar intervenciones para reducir la mortalidad en accidentes viales.

6.
   Regresión:

RMSE (Root Mean Squared Error):

Todos los modelos tienen valores muy cercanos, lo que sugiere un rendimiento similar en términos de predicciones.
MAE (Mean Absolute Error):

Nuevamente, los modelos tienen valores cercanos, indicando consistencia en la precisión de las predicciones, pero el Modelo 2 tiene el MAE más alto.

R-squared: Los valores son muy bajos para todos los modelos, sugiriendo que la variabilidad en la variable dependiente no está bien explicada por los modelos.
Explained Variance: El Modelo 3 tiene el valor más alto, lo que indica que explica una mayor proporción de la varianza en los datos.

Aunque los modelos tienen un rendimiento similar en RMSE y MAE, el Modelo 3 parece ser el mejor en términos de explicar la varianza en los datos.

Clustering:
El Modelo 3 tiene la mayor puntuación de silueta, lo que sugiere que los clusters son más claramente definidos en este modelo.
Inercia:

La inercia más baja se encuentra en el Modelo 2, indicando que los centroides de los clusters están más cerca entre sí en comparación con los otros modelos.
El Modelo 3 parece tener un rendimiento superior en términos de la métrica de silueta, lo que sugiere clusters más cohesivos. Sin embargo, el Modelo 2 tiene una inercia más baja, lo que indica una compactación más fuerte de los clusters.

Conclusiones:

El proyecto de análisis de datos para la ciudad de Nueva York sobre accidentes automovilísticos y arrestos policiales ha proporcionado una visión completa y detallada de la compleja interacción entre la seguridad vial y la aplicación de la ley en este estado. La convergencia de análisis exploratorio, visualizaciones y técnicas de aprendizaje de máquina ha permitido identificar áreas críticas, grupos demográficos específicos y patrones temporales, proporcionando una comprensión sólida de los desafíos.

Los hallazgos respaldan la formulación de estrategias específicas, tanto en términos de políticas públicas como de medidas prácticas. La importancia de la prevención, la educación y la concienciación destaca en el análisis temporal y demográfico, subrayando la necesidad de medidas específicas en momentos y lugares clave. Además, la aplicación de técnicas de clustering ha proporcionado una base para el desarrollo de estrategias territoriales, identificando áreas con perfiles similares de accidentes y orientando la asignación eficiente de recursos.

La evaluación de modelos ha revelado la necesidad de una validación continua y ajuste de enfoques, reconociendo la complejidad y dinámica de los datos de accidentes y arrestos. La intersección de datos demográficos, temporales y geográficos destaca la necesidad de

estrategias holísticas que aborden la diversidad de factores contribuyentes. Además, se destaca la importancia de consideraciones éticas en la implementación de estrategias basadas en datos, asegurando la equidad y evitando sesgos en la aplicación de políticas y acciones.

En resumen, el proyecto no solo ofrece respuestas a preguntas específicas, sino que también proporciona una plataforma sólida para la toma de decisiones informada y la acción proactiva en la mejora de la seguridad vial y la eficacia de la aplicación de la ley en Nueva York.

Referencias:

- Datasets utilizados:

1. NYPD Arrests Data (Historic) | NYC Open Data. (2023, 27 abril). https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u
2. Motor vehicle collisions - crashes | NYC Open Data. (2023, 10 noviembre). https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95
3. Motor vehicle collisions - person | NYC Open Data. (2023, 11 noviembre). https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6u