

Fabian Andres Diaz Martinez

Juan David Ramirez

David leon

Juan Diego Carreño Vásquez

Tomás Pinilla Florez

Analisis

Arrestos y accidentes en

NY

**01**

FILTROS Y TRANSFORMACIONES

**02**

RESPUESTA A PREGUNTAS

**03**

MODELO SELECCIONADO

**04**

APLICACION ML

**05**

RESULTADOS



# Filtros aplicados

## 01- Filtro de consistencia geográfica por borough

- Aseguramos datos ubicados correctamente para comparaciones territoriales confiables.
- Eliminamos registros sin borough para mantener análisis espaciales precisos.
- Garantizamos coherencia geográfica para interpretar patrones por zonas sin distorsiones.

## 02- Filtro de cobertura mensual válida

- Conservamos solo meses 1–12 para una línea temporal ordenada y comparable.
- Depuramos meses inválidos para asegurar tendencias mensuales consistentes.
- Estructuramos datos temporalmente correctos para análisis estables y sin ruido.

## 03- Filtro de eliminación de duplicados residuales

- Eliminamos duplicados mensuales para evitar inflar artificialmente los conteos.
- Aseguramos una sola observación por borough–año–mes para precisión analítica.
- Controlamos duplicaciones para mantener resultados temporales y territoriales fidedignos.

# Transformaciones

## 01- Índice de severidad ponderado

- Indicador que combina heridos y fallecidos para medir severidad real de colisiones.
- Métrica mensual que prioriza territorios según impacto y gravedad vial.
- Permite comparar incidentes no solo por cantidad, sino por severidad relativa.

## 02- Panel mensual unificado + razón operativo–arrestos

- Integramos arrestos, colisiones y contexto para un análisis territorial completo.
- Un panel único facilita comparar seguridad, movilidad y condiciones socioeconómicas.
- Razón arrestos/colisiones revela dónde la intervención policial es proporcionalmente mayor.



# Transformaciones

## 03- Suavizamiento temporal y rezagos

- Medias móviles reducen ruido mensual y resaltan tendencias reales.
- Rezagos permiten anticipar patrones y detectar cambios abruptos en el tiempo.
- Suavizar datos ayuda a entender comportamiento estable de arrestos y colisiones.

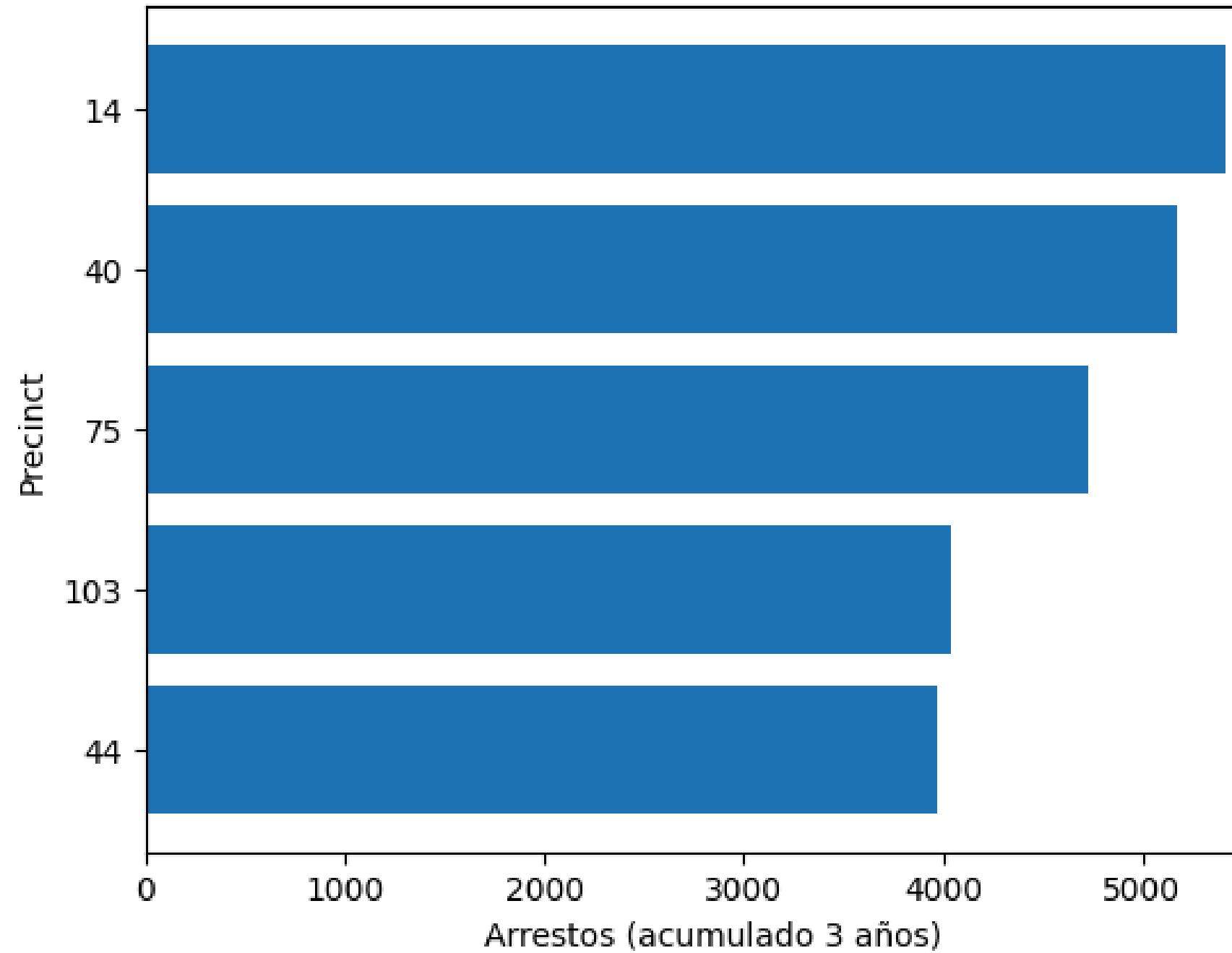
## 04 -Variación interanual (YoY)

- Comparamos cada mes con su año previo para detectar cambios relativos.
- YoY identifica mejoras, deterioros y efectos de políticas públicas.
- Resalta tendencias estructurales en seguridad y siniestralidad vial.



TOP 5 PRECINCTS CON MÁS ARRESTOS EN LOS  
ÚLTIMOS TRES AÑOS

Top-5 precincts por arrestos (2023-2025)

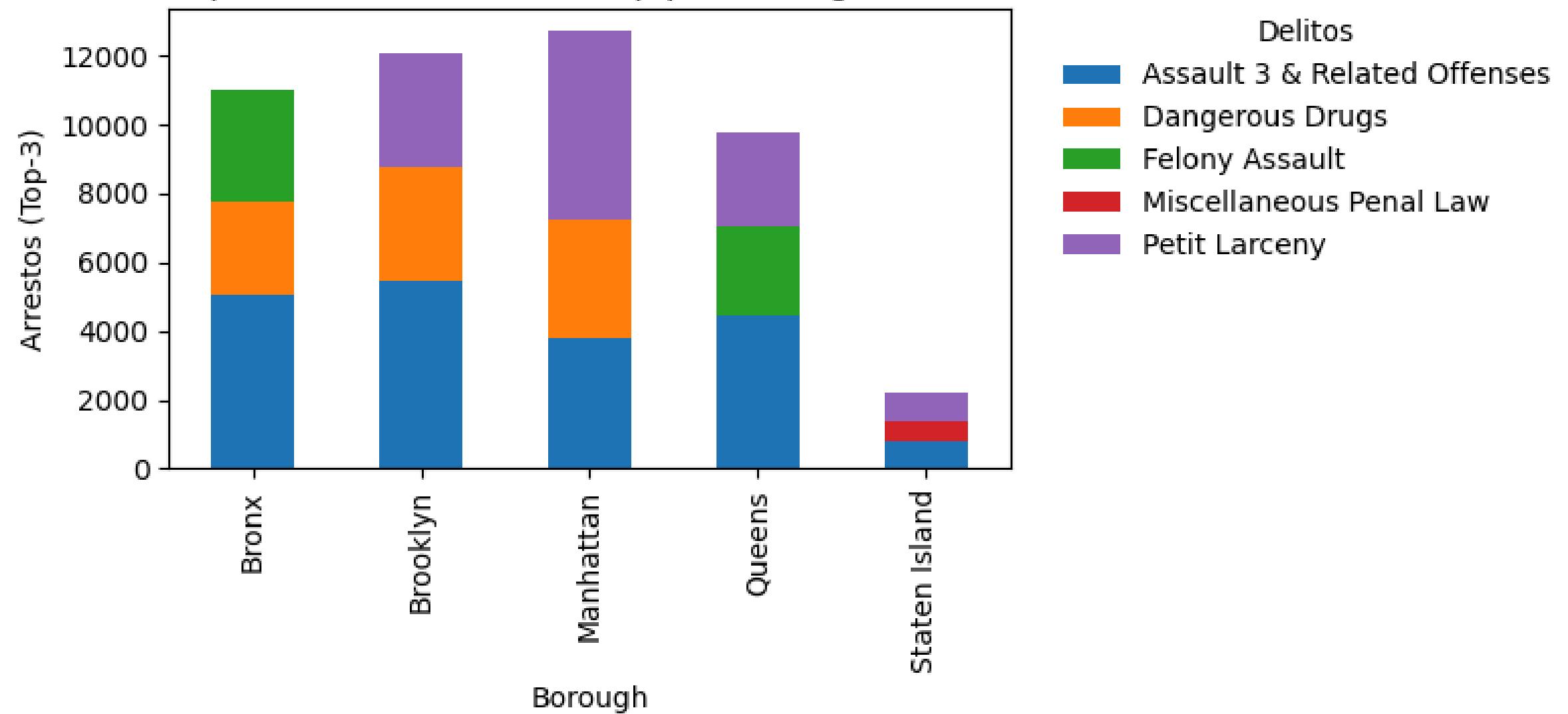


# Respuesta 1

1. Los precincts 14, 40, 75, 103 y 44 registran más arrestos recientes.
2. El precinct 14 es el distrito con el volumen de arrestos más alto.
3. Estos cinco precincts concentran la mayor actividad policial de toda la ciudad.

¿QUÉ DELITOS SE CONCENTRAN EN MAYOR MEDIDA  
EN CADA BOROUGH

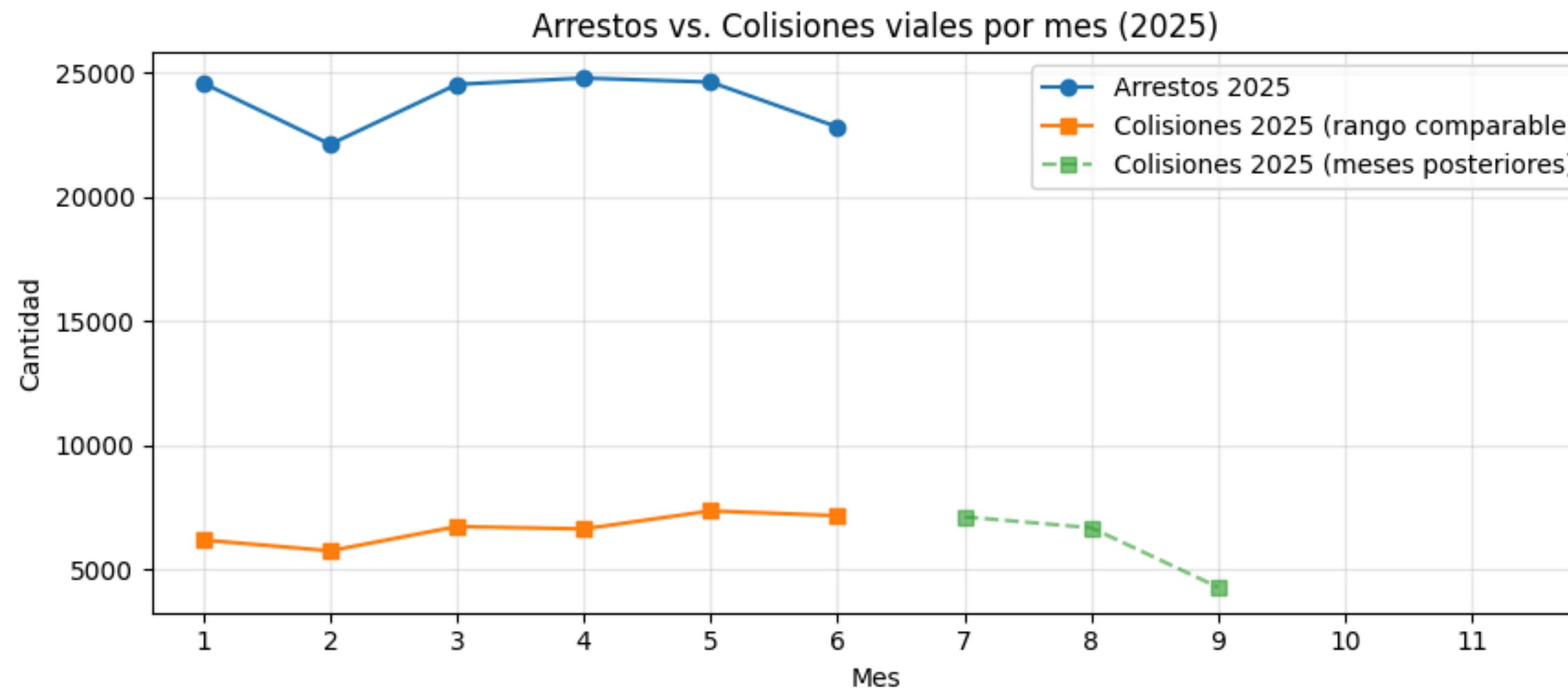
Participación de los 3 delitos Top por borough (2025 YTD)



# Respuesta 2

1. Bronx y Queens concentran principalmente agresiones y violencia interpersonal.
2. Brooklyn combina agresiones con hurtos menores; Manhattan domina en Petit Larceny.
3. Staten Island presenta niveles bajos y delitos dispersos sin un patrón dominante.

¿EN QUÉ MESES DEL AÑO SE INTENSIFICAN LOS ARRESTOS Y CÓMO SE RELACIONA ESA ESTACIONALIDAD CON LA DE LOS ACCIDENTES VIALES?



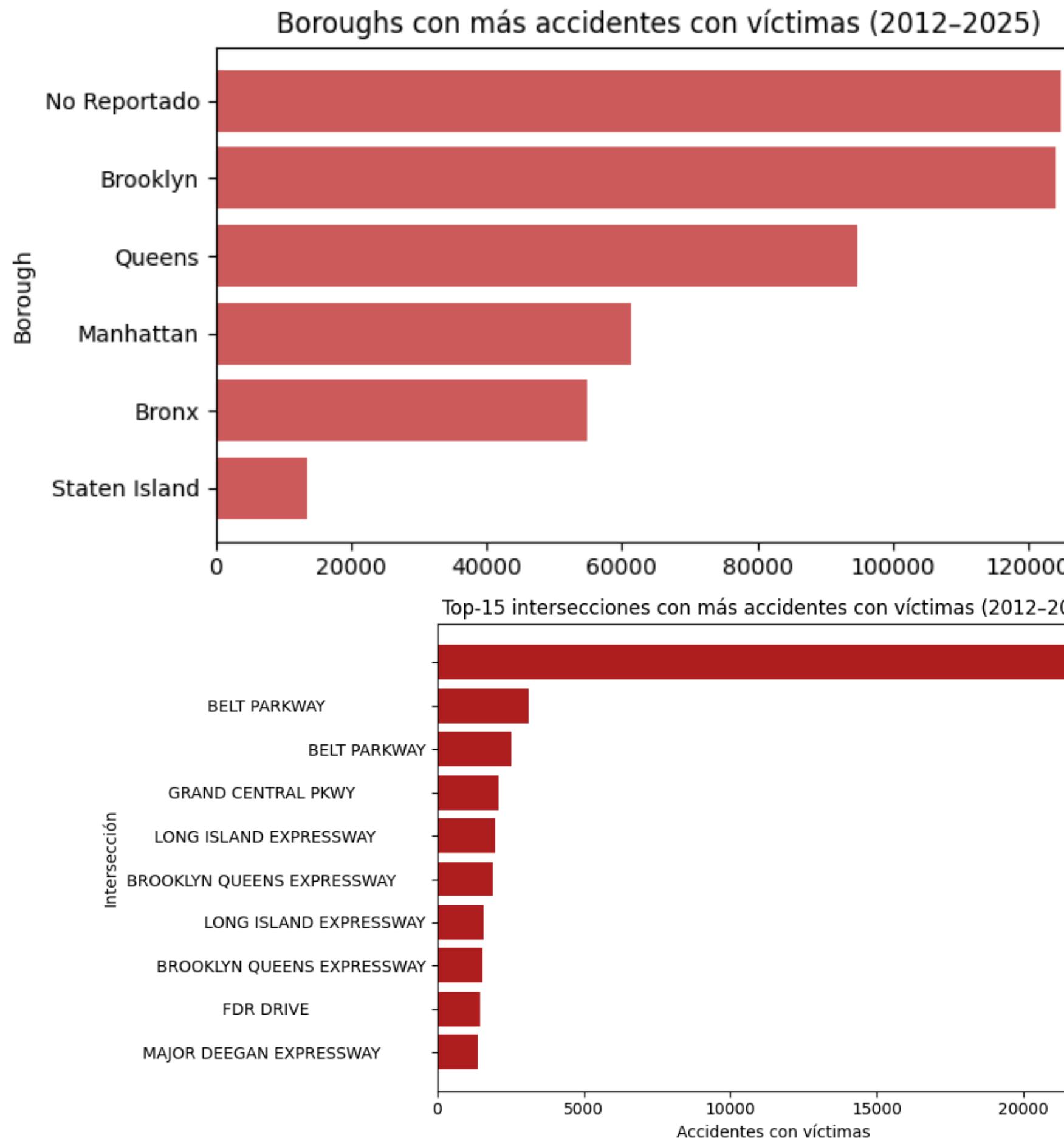
Intensidad de correlación Arrestos ↔ Colisiones (2025)

$r = 0.33$

# Respuesta 3

1. Los arrestos aumentan entre enero y junio, con picos en marzo y mayo.
2. Las colisiones muestran repuntes similares, reflejando una estacionalidad parcialmente coincidente.
3. La relación entre ambos es positiva pero débil, sin dependencia lineal fuerte.

¿QUÉ INTERSECCIONES Y BOROUGHS PRESENTAN  
MAYOR NÚMERO DE ACCIDENTES CON VÍCTIMAS?

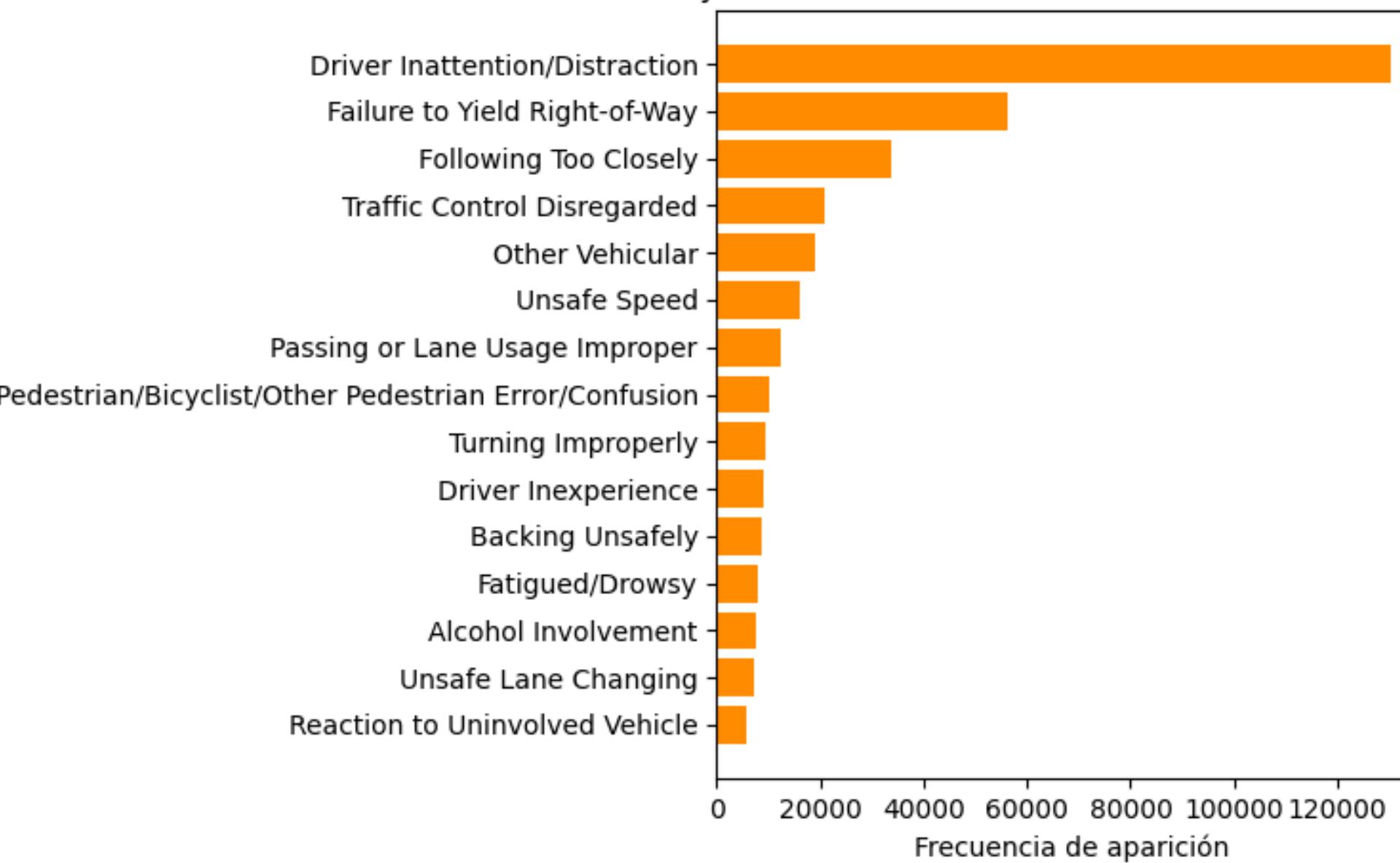


## Respuesta 4

1. Brooklyn y Queens registran la mayor cantidad de accidentes con víctimas en la ciudad.
2. Corredores como Belt Parkway y Long Island Expressway concentran los incidentes más graves.
3. Manhattan y Bronx muestran niveles intermedios, mientras Staten Island presenta la menor incidencia.

¿QUÉ FACTORES CONTRIBUYENTES APARECEN CON MAYOR FRECUENCIA EN ACCIDENTES GRAVES?

Factores contribuyentes más frecuentes en accidentes con víctimas (2012-2025)

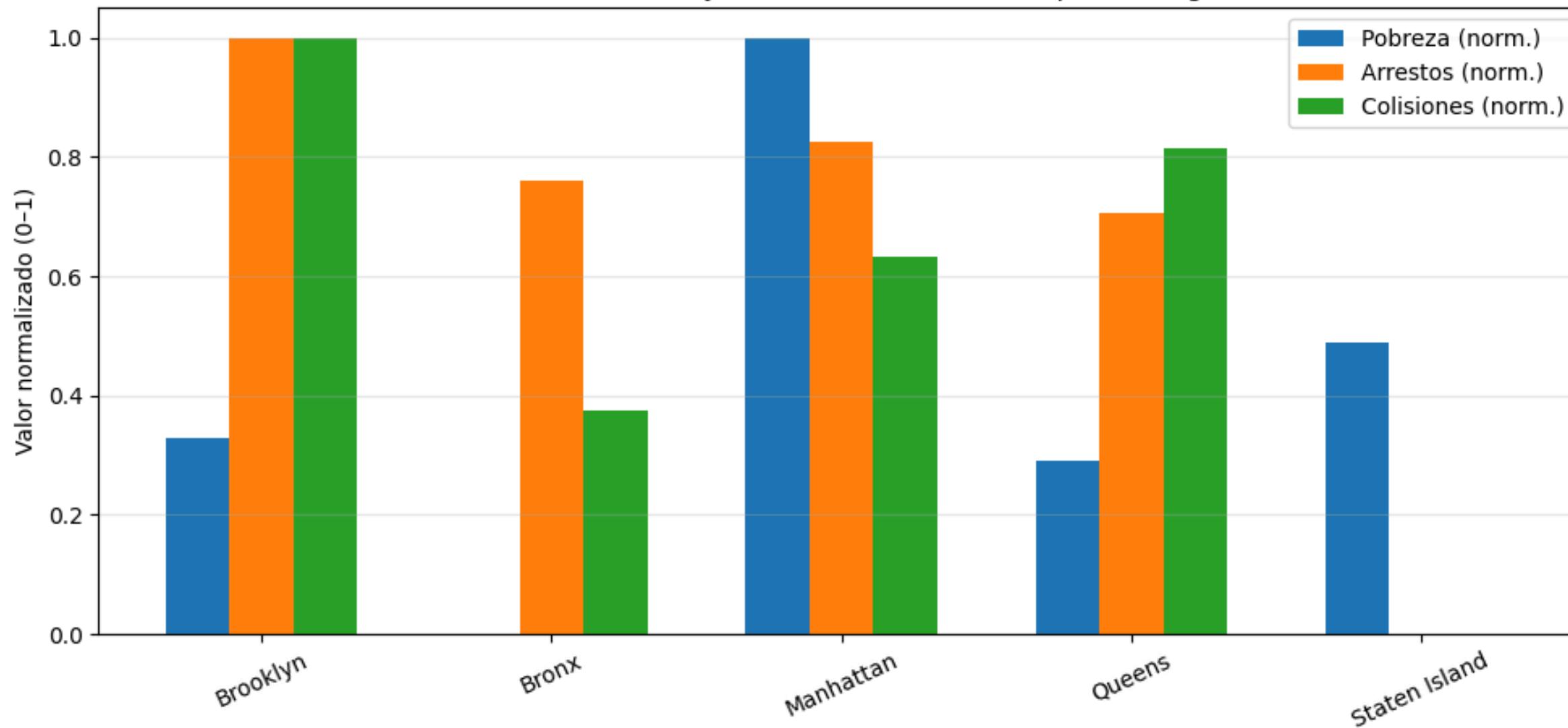


# Respuesta 5

1. La distracción del conductor es el factor más frecuente en accidentes graves.
2. No ceder el paso y seguir muy cerca destacan como causas recurrentes.
3. Los accidentes graves se explican principalmente por fallas humanas, no mecánicas o ambientales.

¿QUÉ BOROUGHS PRESENTAN SIMULTÁNEAMENTE  
MAYOR POBREZA Y MAYORES NIVELES DE ARRESTOS  
O COLISIONES?

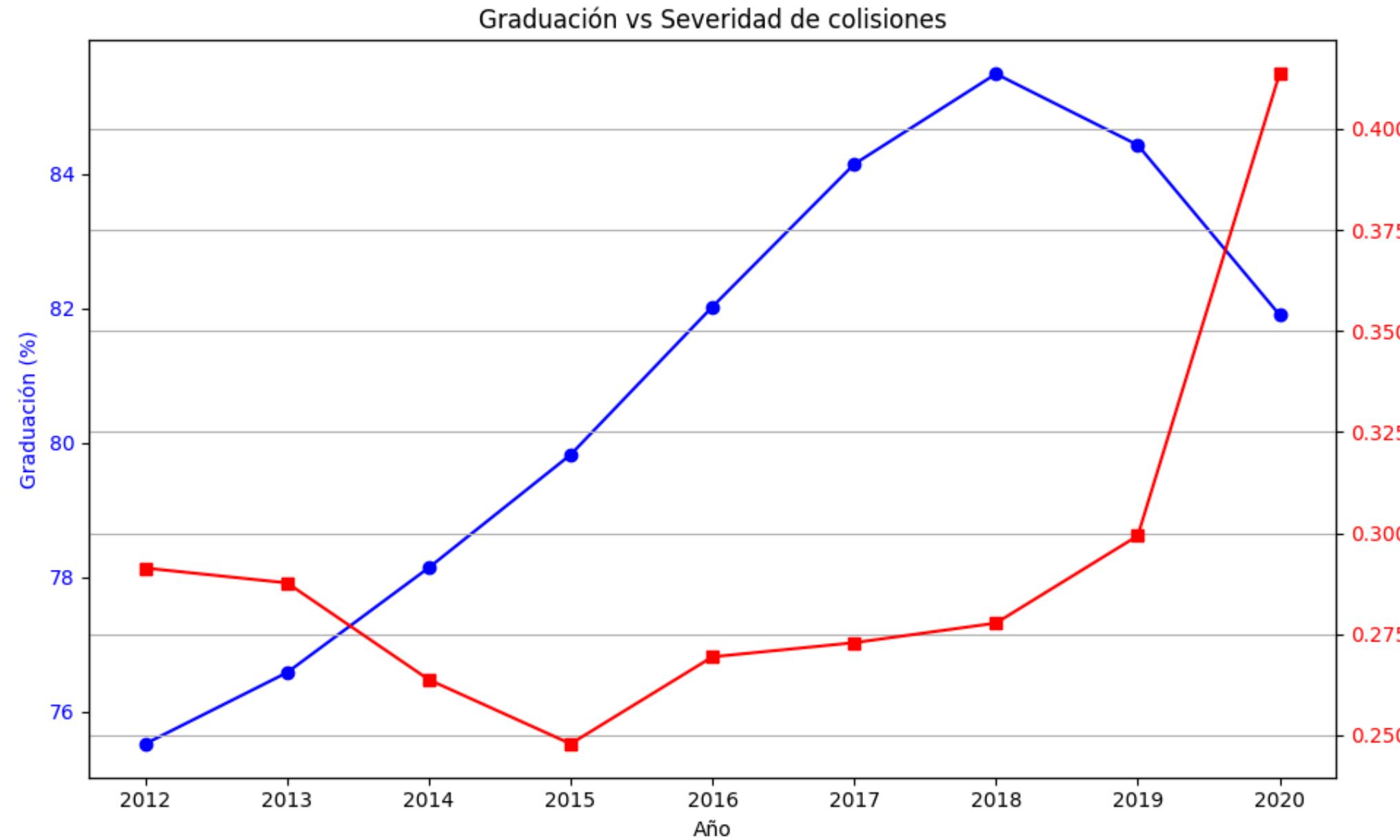
Pobreza, arrestos y colisiones normalizados por borough



# Respuesta 6

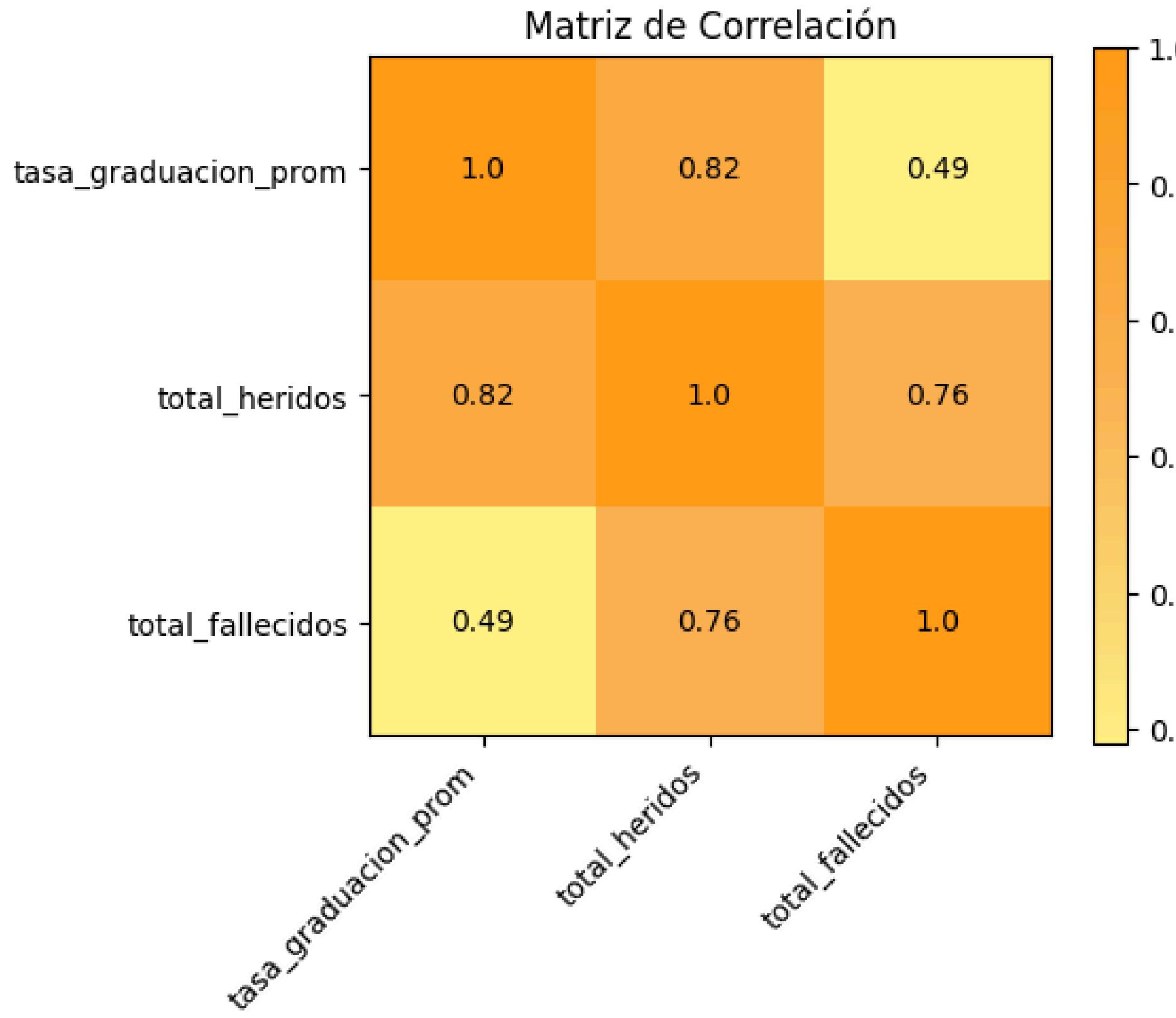
1. Brooklyn y Manhattan lideran arrestos y colisiones, pero no la pobreza.
2. El Bronx es el más pobre, pero muestra niveles intermedios de incidentes.
3. No hay relación directa: movilidad y actividad urbana explican más los eventos.

¿CÓMO HA EVOLUCIONADO LA TASA DE  
GRADUACIÓN ESCOLAR Y CÓMO SE COMPARA CON  
LA SEVERIDAD DE ACCIDENTES?



# Respuesta 7

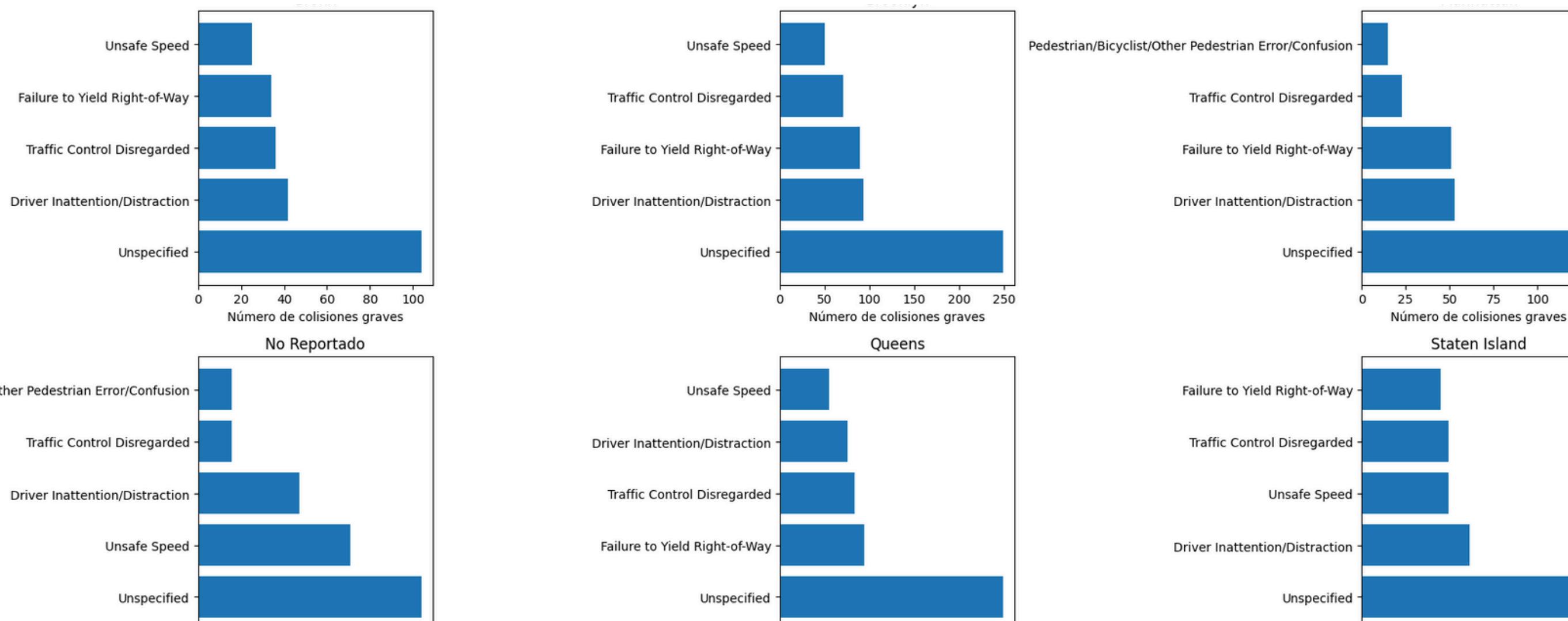
1. La graduación aumentó hasta 2018, mientras la severidad volvió a crecer desde 2016.
2. Ambas series muestran correlaciones positivas, aumentando en años similares.
3. No hay relación causal: comparten factores externos que afectan ambas tendencias.



# Respuesta 7

1. La graduación aumentó hasta 2018, mientras la severidad volvió a crecer desde 2016.
2. Ambas series muestran correlaciones positivas, aumentando en años similares.
3. No hay evidencia suficiente para argumentar relación causal: comparten factores externos que afectan ambas tendencias.

¿CUÁLES SON LOS FACTORES CONTRIBUYENTES MÁS FRECUENTES EN ACCIDENTES DE SEVERIDAD ALTA POR BOROUGH?



# Respuesta 8

1. Distracción del conductor y no ceder el paso son las causas más frecuentes.
2. Brooklyn y Queens concentran los mayores volúmenes de factores humanos críticos.
3. Manhattan destaca además por errores de peatones y ciclistas en colisiones graves.
4. Listo para continuar.

# Selección de técnicas de aprendizaje de máquina

## Técnica Supervisada: Random Forest Regressor

El objetivo es predecir el volumen de arrestos y el índice de severidad vial para el mes siguiente.

Justificación:

- Manejo robusto de variables mixtas (categóricas + continuas).
- Reduce el overfitting frente a árboles simples.
- Permite extraer la importancia de variables para explicar causas al gobierno.

## Técnica No Supervisada: K-Means Clustering

El objetivo es la segmentación territorial de riesgos.

Justificación:

- Agrupa boroughs o zonas con comportamientos similares (ej. "Alta Pobreza + Baja Siniestralidad").
- Permite diseñar políticas diferenciadas por clúster en lugar de una estrategia genérica.

# Preparación de datos para modelado

## Desafío de Integración (Data Quality)

- Problema: Las fuentes de Pobreza y Colisiones no tenían registros actualizados para 2025 (a diferencia de Arrestos), generando nulos.
- Solución: Implementación de Estrategia de Imputación (Medias históricas y constantes) en lugar de eliminación, salvando el 100% de la data reciente.

## Transformaciones con PySpark MLLib

- Limpieza: Eliminación de variables redundantes (multicolinealidad  $r > 0.90$ ).
- Vectorización: Indexación de Borough (StringIndexer) y ensamblaje de vectores (VectorAssembler).
- Normalización (Crítico): Aplicación de StandardScaler para estandarizar magnitudes (Ingresos vs. Conteos) y asegurar la convergencia correcta de K-Means.
- Resultado: Feature Vectors listos para entrenamiento.

# Evaluación de Modelos Predictivos

La evaluación de modelos es crítica para determinar la capacidad predictiva y utilidad práctica de las técnicas implementadas. Se presentan métricas, experimentos con diferentes configuraciones y la comparación que sustenta la selección del modelo óptimo.

## Métricas para Árbol de Decisión

### RMSE

Raíz cuadrada del promedio de errores al cuadrado. Penaliza fuertemente desviaciones grandes.

### MAE

Promedio de diferencias absolutas entre valores reales y predichos. Robusto ante valores atípicos.

### $R^2$ (Coeficiente)

Proporción de varianza explicada por el modelo. Valores cercanos a 1 indican mejor ajuste.

# Experimentos con Árbol de Decisión

Exp.	maxDepth	maxBins	RMSE	R <sup>2</sup>
1	3	16	245.82	0.412
2	5	16	198.34	0.584
3	7	32	176.91	0.651
4	10	32	181.23	0.638
5	15	64	189.47	0.619

Hallazgos Clave

- Profundidad 3: subajuste evidente
- Óptimo: maxDepth=7, maxBins=32
- Profundidades >10: sobreajuste
- maxBins >32: sin mejoras significativas



# Selección del Modelo Óptimo

## Experimentos con K-means

k	WSSSE	Silhouette	Iteraciones
2	8,947.23	0.523	8
3	6,142.87	0.587	12
4	4,821.56	0.548	15
5	3,956.18	0.512	18
6	3,287.45	0.478	21

### Observaciones

- k=3: punto de inflexión óptimo
- Máximo coeficiente de silueta (0.587)
- k>4: fragmentación excesiva
- 3 perfiles territoriales diferenciados

# Modelos Seleccionados

1

## Árbol de Decisión

**Configuración:** maxDepth=7, maxBins=32,  
minInstancesPerNode=10

- Menor RMSE (176.91) y mayor R<sup>2</sup> (0.651)
- 43 nodos terminales interpretables
- Variables clave: total\_crashes, month, borough\_index, lag\_arrests

2

## K-means

**Configuración:** k=3 clusters, maxIter=50, seed=42

- Máximo coeficiente de silueta (0.587)
- Cluster 1: Alta intensidad (Brooklyn/Manhattan)
- Cluster 2: Intensidad media (Queens/Bronx)
- Cluster 3: Baja intensidad (Staten Island)

# Integración Complementaria



## Capacidad Predictiva

Anticipar niveles de severidad bajo diferentes escenarios para asignación proactiva de recursos.



## Segmentación Estratégica

Identificar territorios con dinámicas similares para diseñar intervenciones diferenciadas.



## Evidencia para Política

Herramientas de priorización y marcos de implementación para decisiones basadas en datos.



# Conclusiones

## 01- No hay una relación simple entre pobreza, arrestos y accidentes

Brooklyn y Manhattan concentran muchos arrestos y colisiones sin ser los más pobres, mientras el Bronx es el más pobre pero con niveles intermedios de incidentes, lo que muestra que influyen también movilidad, actividad urbana y otros factores

## 02- Los modelos de ML aportan lectura, más que “predicciones perfectas”

el Árbol de Decisión ayuda a entender qué variables (borough, tiempo, volumen de choques, contexto) explican mejor los niveles de arrestos, y K-means permite crear perfiles territoriales (alta, media y baja intensidad) útiles para segmentar la ciudad

## 03- Las políticas deben ser focalizadas y basadas en datos

los resultados muestran que la mayoría de accidentes graves se explican por fallas humanas (distracción, no ceder el paso, seguir muy cerca), por lo que las intervenciones más efectivas combinan control policial, diseño vial y campañas de comportamiento en los boroughs