UADY
UNIVERSIDAD AUTÓNOMA DE YUCATÁN

Minería de Datos

Proyecto

Parte II

Prof. Ernesto Guerrero

Equipo:

- Bañuelos Verónica
- Díaz Cristhian
- Colás Flor
- Mercado Pablo
- Carrillo Rafael
- Rubí Naila

Stop and Frisk

Stop and Frisk fue un programa que se implementó en Estados Unidos, particularmente en la ciudad de Nueva York, que llevó a cabo el departamento de policía. Éste consiste en detener, cuestionar y, si es necesario, registrar o cachear a los civiles que parezcan sospechosos de portar armas o sustancias relacionadas con el contrabando.

Dentro de la Constitución de los Estados Unidos se encuentra la Cuarta Enmienda, la cual protege a los civiles de arrestos arbitrarios. Ésta requiere que antes de detener al sospechoso, el policía debe tener una sospecha razonable de que el individuo ha cometido o está a punto de cometer un delito. Si el policía sospecha razonablemente que el sospechoso está armado y es peligroso, el policía podrá cachearlo, lo que significa que le dará una rápida palmada de la ropa exterior del sospechoso. El registro también se llama Terry Stop, derivado del caso de la Corte Suprema Terry vs. Ohio, 392 U.S. 1 (1968).

En 2017 se realizaron 10,861 paradas, sin embargo, el programa anteriormente tuvo lugar en una escala mayor. Entre 2003 y 2013, se realizaron más de 100,000 paradas por año, con 685,724 personas detenidas en el apogeo del programa en 2011.

Posteriormente, el programa fue objeto controversia racial. Según el reporte correspondiente al 2011 realizado por la New York Civil Liberties Union: las y los jóvenes afroamericanos y latinos fueron blanco de una cantidad enormemente desproporcionada de paradas.

En promedio, de 2002 a 2013, el número de personas detenidas sin ninguna condena fue del 87,6%. Esto quiere decir que un enorme porcentaje de los detenidos eran inocentes y no estaban realizando ninguna actividad delictiva.

Objetivo

Disminuir las tasas de crimen de cada localidad a través de acciones de prevención tales como la detención (stop) y cacheo (frisk) de personas cuyos actos y/o actitud levanten sospechas de que un crimen se haya cometido o esté por cometerse.

Recolección de Datos

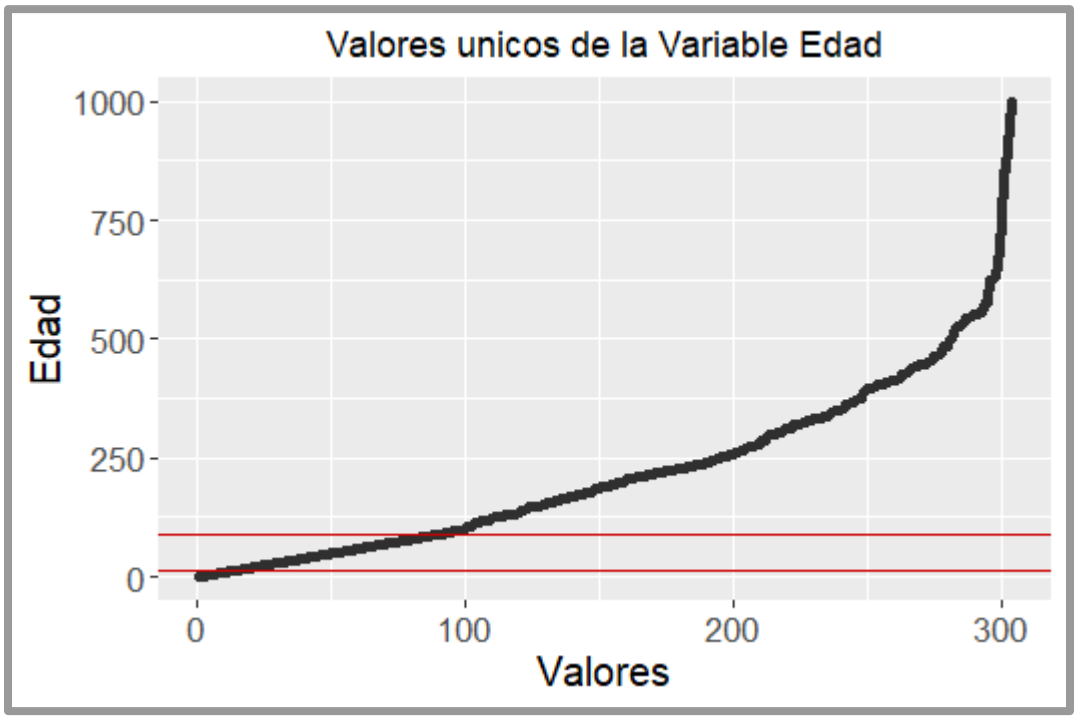
La metodología para la obtención de datos consistía en llenar un formato UF-250 posterior a la detención. Este incluía el llenado de determinadas variables, tales como características de la persona detenida y detalles de la detención. Posteriormente, estas variables se capturan en una base de datos. Gran parte de las variables que forman parte de dicho formato son registradas como de tipo dicotómico en la base, de lo que se puede intuir que el interés residía en observar si el detenido cumplía o no con la característica descrita. Un ejemplo de esto es si el detenido portaba o no un arma o si el detenido opuso resistencia o no. Además de esto, se registraban las características físicas de los detenidos, como son edad, sexo, color de ojos, entre otras.

Sección 2

Revisión de la Base de Datos

Se realiza una consulta de la estructura de la base de datos y se puede ver que todas las columnas son valores numéricos y enteros. La función str() también muestra en algunos casos los “levels” y en otros el rango.

Ciertas variables numéricas tienen comportamientos bastante irregulares. La variable edad, por ejemplo, tiene un comportamiento anormal en sus valores, ya que si se usa la función unique() y se grafican, se puede ver que hay algunos fuera de lugar, valores desde 0 hasta mayores a 100 años; se esperaría un comportamiento ordenado si se usa el comando sort() y, de esta forma, se apreciaría mejor este gráfico.



Se puede ver que más de la mitad de los datos únicos son datos inconsistentes, ya que no existen personas con más de 200 años. De todos estos datos únicos se quiere ver cuántos son los datos dañados del total, para ello se usa la función length() a un vector con los valores que son mayores a 89 y menores a 15.

```
> p=base$age
> length(p[(14>=p | p>=90)])
[1] 16848
>
```

Se puede ver que 16,848 datos son inconsistentes, es decir un 2.45% del total.

Para checar cuales de estos datos son los dañados se utiliza función which() y usando la columna **dob** (date of birth) se calcula la edad correcta de esos datos e incluso se puede volver a calcular la edad de todos los datos. Para hacer eso, se debe checar que la columna **dob** y la columna **datestop** sean consistentes.

Las columnas **dob** y **datestop** están dadas en formato numérico y se requiere convertir en formato-fecha para poder hacer un cálculo del número de días entre dos fechas y con ello lograr el cálculo de **age2** variable que contendrá una propuesta de edad. Previo a esto, se tiene que checar que estas columnas no muestran inconsistencias; para ello se checa que los valores de esta columna no superen por partes sus valores posibles, es decir, no exista un valor 13ddyyyy o mm32yyyy o mmdd2012 o de naturaleza parecida. Esto se puede checar calculando los valores únicos de las columnas con unique() y el mínimo y máximo de estos valores.

```
> a=unique(base$datestop)
> min(a)
[1] 1012011
> max(a)
[1] 12312011
> length(a)
[1] 365
```

Se puede ver que el valor mínimo y máximo coincide con el valor mínimo y máximo posibles para datestop. De igual forma se determina que hay 365 datos únicos por lo que se puede sospechar que no hay datos anormales entre ellos, mas sin embargo no es algo totalmente certero. Para estar seguros de la consistencia de todos los datos se tendrá que realizar un análisis elemento a elemento en un ciclo.

```
> p<-unique(base$dob)
> max(p)
[1] 12311998
> min(p)
[1] 1011900
```

Para la columna **dob** se tiene un máximo y mínimo consistente, pero esto solo es una sospecha, ya que se tendría que checar dato a dato si no se encuentran inconsistencias.

Usando el siguiente código se puede checar que cada elemento de **dob** y **datestop** sea consistente. El código también nos dice cuántos errores hay en los datos, valor diferente pero máximo del número de datos con errores. (Para usar ese código se buscó la existencia de NA’s, no hubo ninguno en ambas columnas).

```
a=base$dob
fecha=NULL
Errorres=0
for (i in seq_along(base$dob)) {
  tam=str_length(a[i])
  #realiza la convercion a fecha
  ifelse(as.numeric(str_sub(a[i],tam-5,tam-4))>31,(Errorres=Errorres+1),Errorres)
  ifelse(as.numeric(str_sub(a[i],1,tam-6))>12,(Errorres=Errorres+1),Errorres)
  ifelse(as.numeric(str_sub(a[i],tam-3,tam))>1996 | as.numeric(str_sub(a[i],tam-3,tam))<1920 ,
         (Errorres=Errorres+1),Errorres)

  fecha[i]=paste(str_sub(a[i],tam-5,tam-4),
                 str_sub(a[i],1,tam-6),
                 str_sub(a[i],tam-3,tam),
                 sep = '-')
}
```

Este código permite decir que hay 281,925 errores en los datos, es decir, máximo 281,925 datos de un total de 685,724 pueden presentar una inconsistencia en la columna **dob** (el 41.113%).

El mismo código cambiando **a** por la columna de **datestop** nos permite decir que no hay ninguna inconsistencia en la columna **datestop**, es decir se confirma la sospecha de que toda la columna **datestop** es consistente.

La columna **city** no se encuentra en la base de datos con ese nombre, por lo que se propone usar la columna de nombre **borough** la cual parece ser city ya que tiene los niveles que city debería tener.

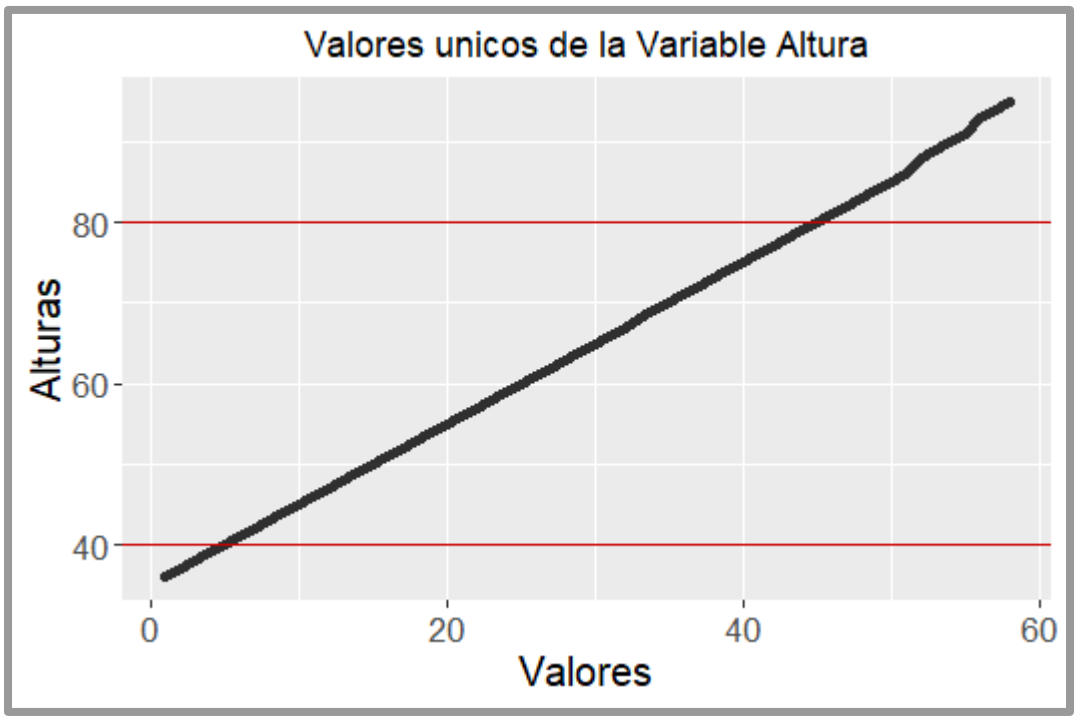
La columna **timestop** tiene valores como 0, lo cual podría ser una inconsistencia ya que significa que la parada le tomo 0 minutos. Se propone verificar estos datos y todos aquellos que tengan inconsistencia de la forma 25mm o hh65.

En la columna **sex** se tienen 12,110 datos faltantes por lo que se tendría que revalidar ya que es un dato que es difícil de no ser anotado en el rellenado de la form UF- 250.

Como ya se había insinuado, **age** es una columna con datos inconsistentes y hasta se puede pensar que es una columna redundante si ya se cuenta con el día de nacimiento y la fecha de la detención, aunque estas columnas tienen más datos inconsistentes que el age.

La columna **height** tiene valores enteros y como máximo un valor de 95, lo cual es algo medio común en ciertas personas, por otro lado, **height** tiene un mínimo de 36 pulgadas, que es una estatura muy pequeña pero no evidentemente inconsistente.

Usando sort() se puede ver que:



Si se consideran los datos sobre el tamaño máximo y mínimo promedio [1] varios datos pueden ser atípicos, pero no inconsistentes.

En el caso de **weight** se tiene que el valor máximo es de 999 libras, lo cual es el peso de una de las personas más gordas del mundo[2], esto da mucha sospecha de unos datos inconsistentes ya que una persona en esa condición no se movería de casa. El valor mínimo de **weight** es de 0, lo cual también es inconsistente. Se propone ir a checar con el creador de la base de datos para ver esas inconsistencias; también se puede hacer una búsqueda de regresión lineal que relacione peso con altura ya que esa columna no parece ser tan inconsistente y se puede proponer valores de peso para cada "raza" según cada regresión lineal que se haga.

De hecho 1,122 datos de la columna peso tiene valores debajo de los 27 kg que es el récord de peso mínimo y 99,340 tienen peso arriba del máximo promedio. Se debe considerar que Estados Unidos es el país con mayor porcentaje de obesidad. Considerando lo anterior, el número de datos en la columna **weight** mayores a 300 libras son 1,506.

La columna **othfeatr** ofrece nula información, ya que debería ser un string y usando unique() se puede ver que no tiene ningún carácter y los valores que tiene o son NA o valores

[1] Se puede checar el globo interactivo en la siguiente pagina http://www.ncdrisc.org/height-mean-map.html

[2] De hecho, la persona más gorda en la historia pesa poco más, pero ese peso se puede ubicar entre los 10 más grandes de la historia. https://www.guinnessworldrecords.com/world-records/heaviest-man

numéricos inconsistentes, esta columna deberá volverse a solicitar o proponer que se descarte ya que tal vez no hay otras características anotadas en los forms.

La columna **arstoffn** al realizar una consulta muestra tener solo 4476 levels, siendo que se esperaban más, esta columna tiene 644,844 datos vacíos, por lo que podría decirse que no proporciona mucha información al programa, solo el 5.96% de los datos son no vacíos en esta columna.

La columna **perstop** muestra tener datos mínimos de 0 y máximo de 999, lo cual no es muy consistente, se propone comparar esta columna con la columna **timestop** y ver coincidencias y nuevas inconsistencias.

La columna **crimecode** parece estar en reemplazo de la columna **detailcm** por lo que se propone cambiar el nombre de esa columna, posible error al nombrar la variable.. El máximo y mínimo elemento de esta columna es consistente.

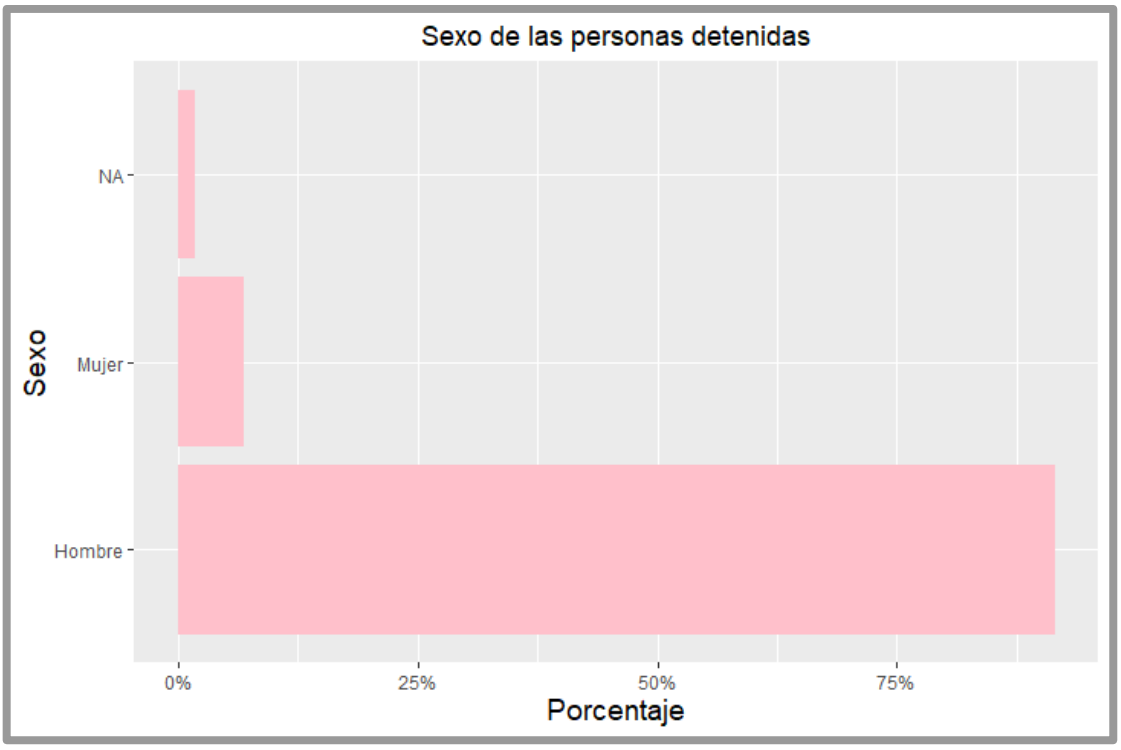
La columna **forceuse** no se encontró, pero se puede verificar que existe una columna igual dicotómica llamada **weapon** que podría ser **forceuse** y que por un error se nombró erróneamente.

Las columnas **borough, race, sex, haircolor, eyecolor, build y premtype** (aunque solo uno) tiene NA’s

En general se espera volver a validar la base de datos porque no se puede tener confianza de los datos en ella ya que muestra tener muchas inconsistencias. Todas las demás variables que no se mencionaron previamente no presentan errores o inconsistencia aparentes.

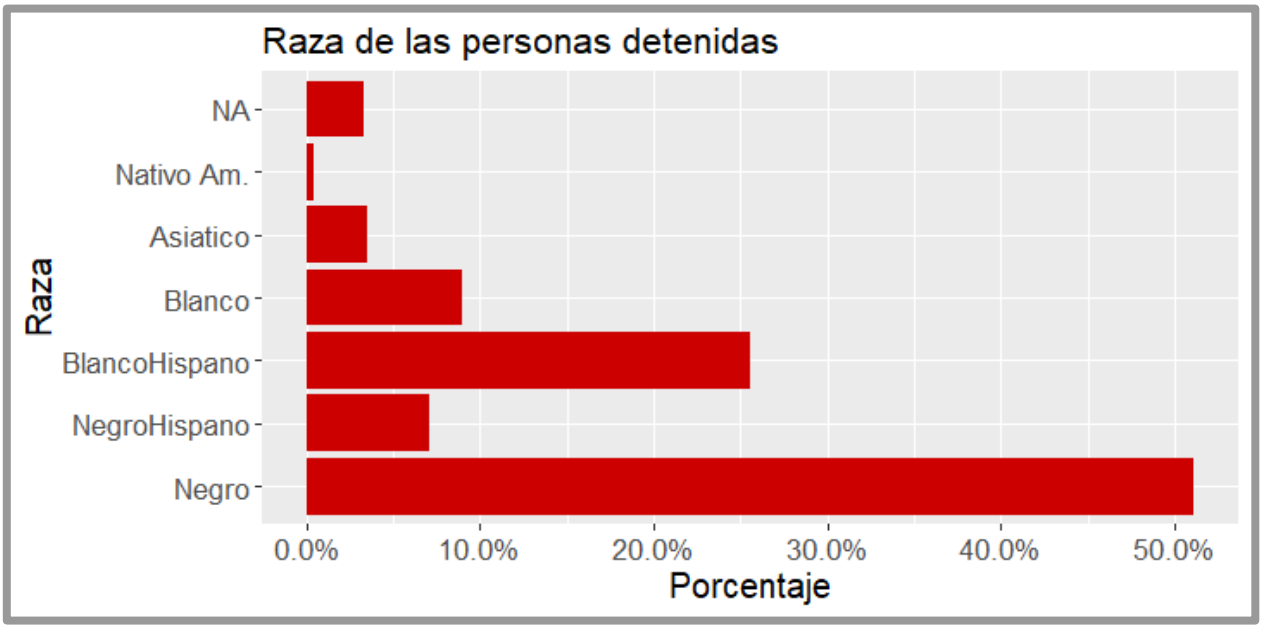
Análisis Descriptivo

Posterior a la revisión de la base, se procede a analizar algunas variables de interés. Inicialmente se analiza la variable sexo. En el siguiente gráfico de barras puede ver el porcentaje de detenidos y detenidas del programa durante el 2011:



La columna NA representa las personas detenidas a las que no se les capturó la variable sexo o bien no se especificó si era mujer u hombre. Se puede observar que la mayoría de las personas detenidas son hombres, sumando un total de 626 830 hombres detenidos y 46 784 mujeres detenidas, correspondientes al 93.05% y el 6.95% del total (sin contar a las personas cuyo registro de sexo este vacío) respectivamente.

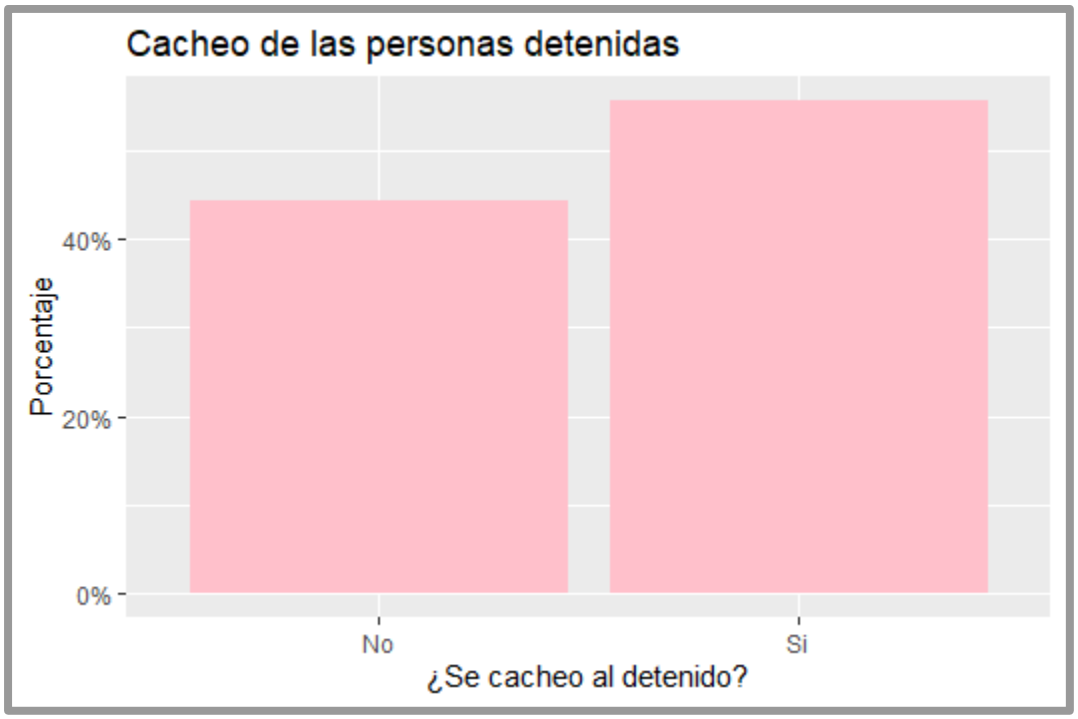
Por otro lado, se tiene el siguiente gráfico con base en la variable “race”:



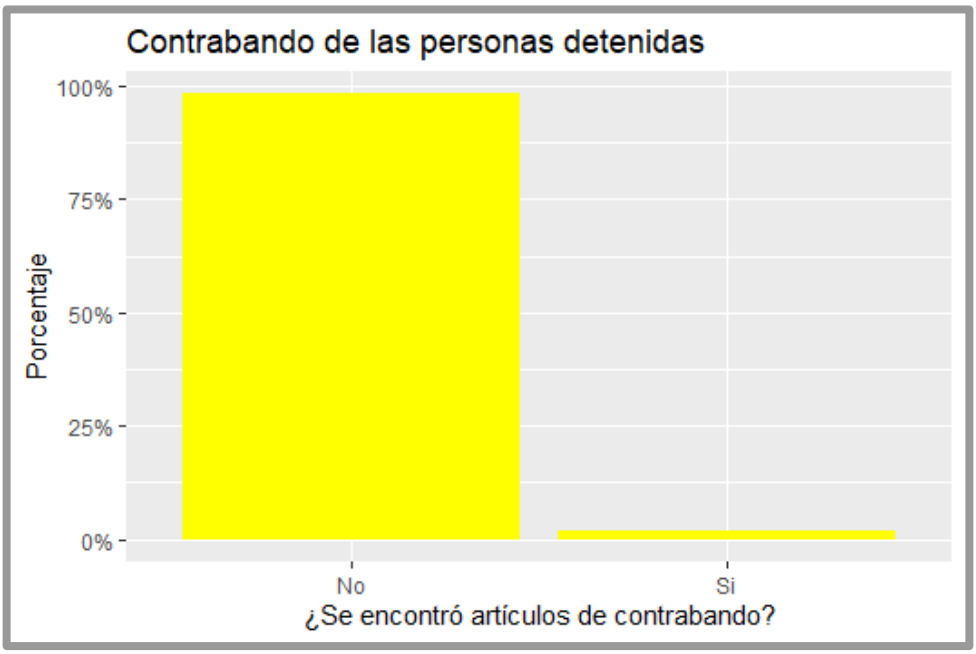
De acuerdo con el gráfico se observa que la mayoría de las personas detenidas son gente negra (52.89%) seguidas de la gente blanca hispana (26.43%), mientras que la menor cantidad de detenidos es la india americana/alasqueños nativos (0.43%), sin embargo, esto puede deberse a la reducida proporción de personas de esa etnia que habitan en la ciudad de Nueva York. De acuerdo con el U.S Census Bureau solo el 0.4 por ciento de la población neoyorquina se considera gente india americana/alasqueños nativos.

Aunque representan solo el 4.7 por ciento de la población de la ciudad, las personas negras e hispanas entre las edades de 14 y 24 representaron casi el 80% de las paradas en 2011. El número de personas negras paradas excedió a toda la población de personas no negras.

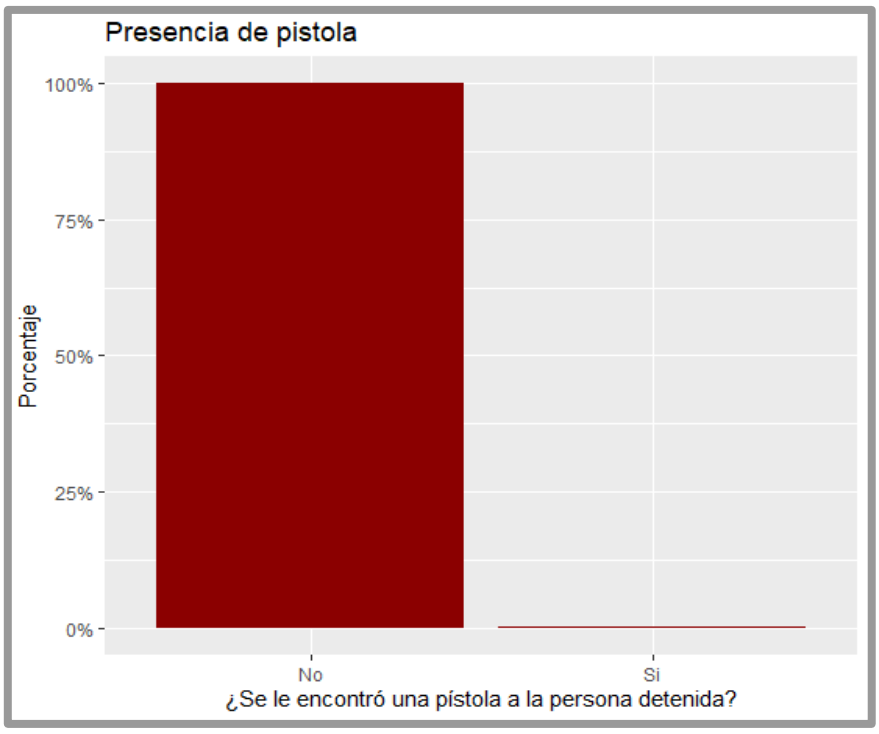
Se procede a realizar un gráfico para determinar la proporción de personas que son cacheadas/registradas:



Alrededor de un 55.66% de las personas detenidas son cacheadas por los oficiales, esto equivale a 381 704 personas de las 685 724.

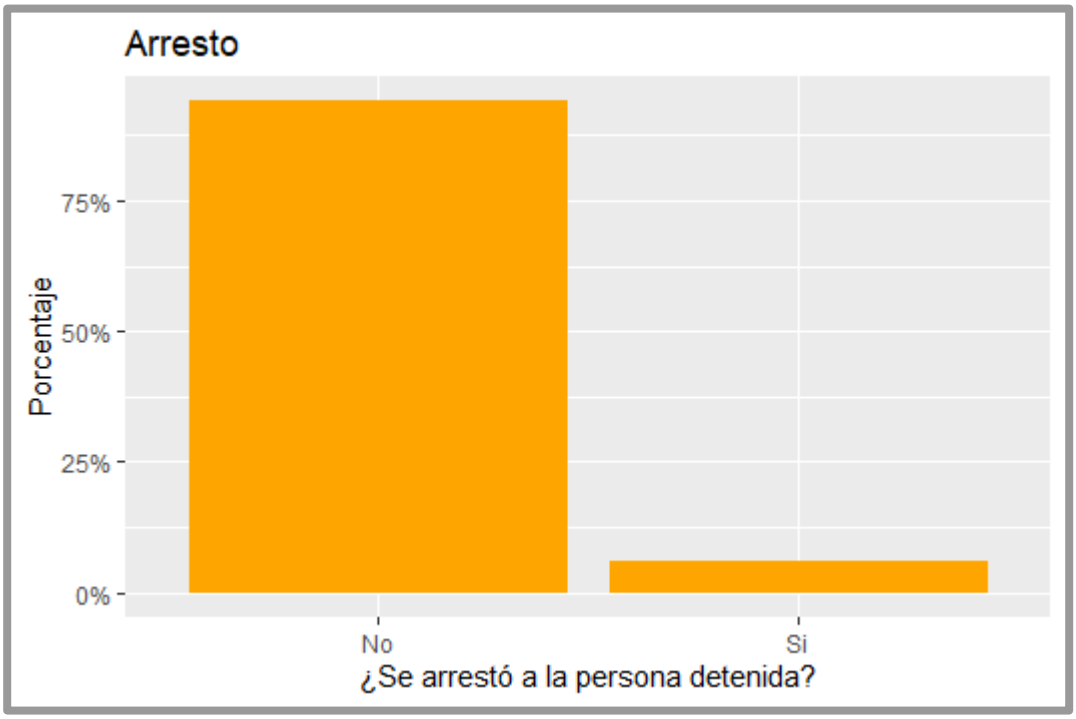


Del total de personas detenidas se observa que aproximadamente a más del 95% no se les encontró artículos de contrabando, solo a 11 803 se les encontró este tipo de artículos.



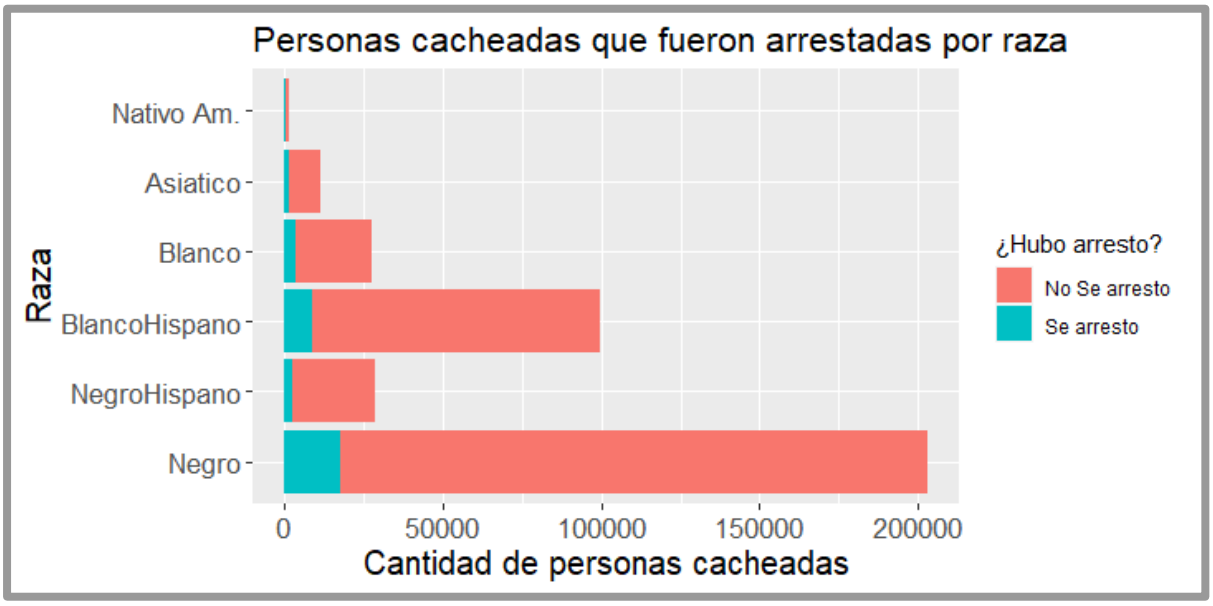
Al igual que en contrabando no se encontraron armas en casi ninguna de las personas detenidas, esto se puede deber a que, en su mayoría, las detenciones fueron arbitrarias y por esto la probabilidad de encontrar algún artículo de contrabando o armas es muy baja, en el gráfico no se aprecia tanto debido a que el porcentaje es muy pequeño, pero solo a 745 personas (0.1%) se les encontró una pistola.

También nos interesaría saber el total de arrestos que se dieron en el 2011.



De las 685 724 detenidas solamente 40 883 fueron arrestadas, esto es equivalente al 5.96% del total. Como ya se había mencionado, esto se puede deber a que las detenciones procedieron sin realmente tener suficiente evidencia o razón para que se llevarán a cabo.

A continuación se muestra un gráfico de 3 variables: “raza”, cantidad de personas cacheadas y si fueron arrestadas (1) o no (0), donde el eje x representa la “raza”:

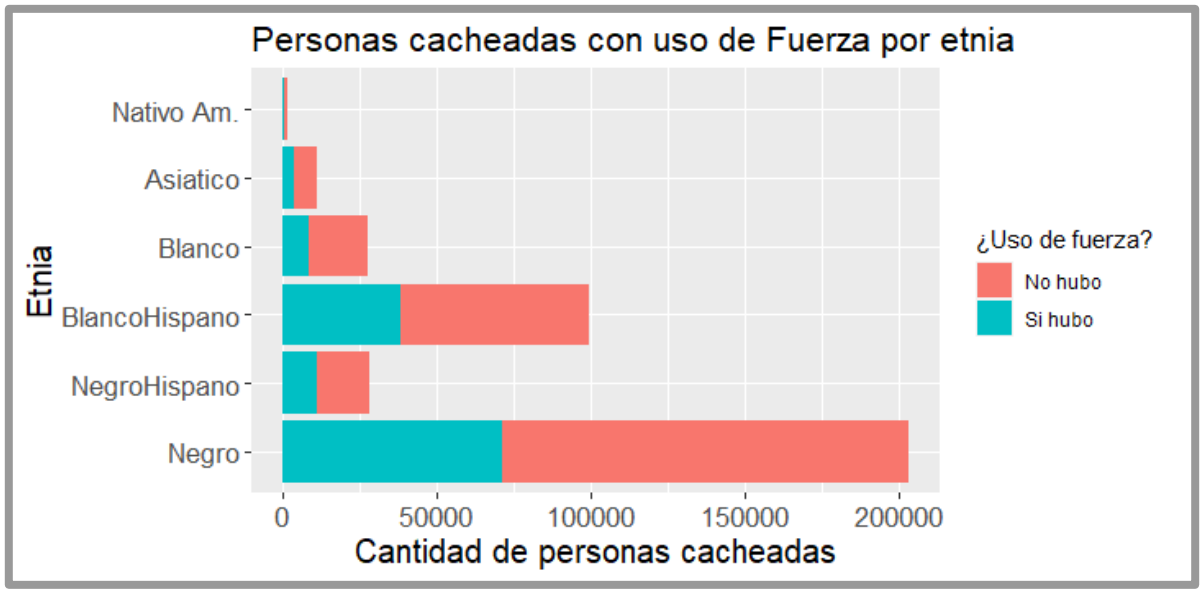


A primera instancia es notorio que la cantidad de personas arrestadas es muy pequeña comparada con el total de personas detenidas. Asimismo se nota la concentración de las personas cacheadas en la cuarta y ùltima barra del gráfico, correspondiente a las personas blanca hispanas y personas negras respectivamente . Del gráfico se puede extraer la siguiente información:

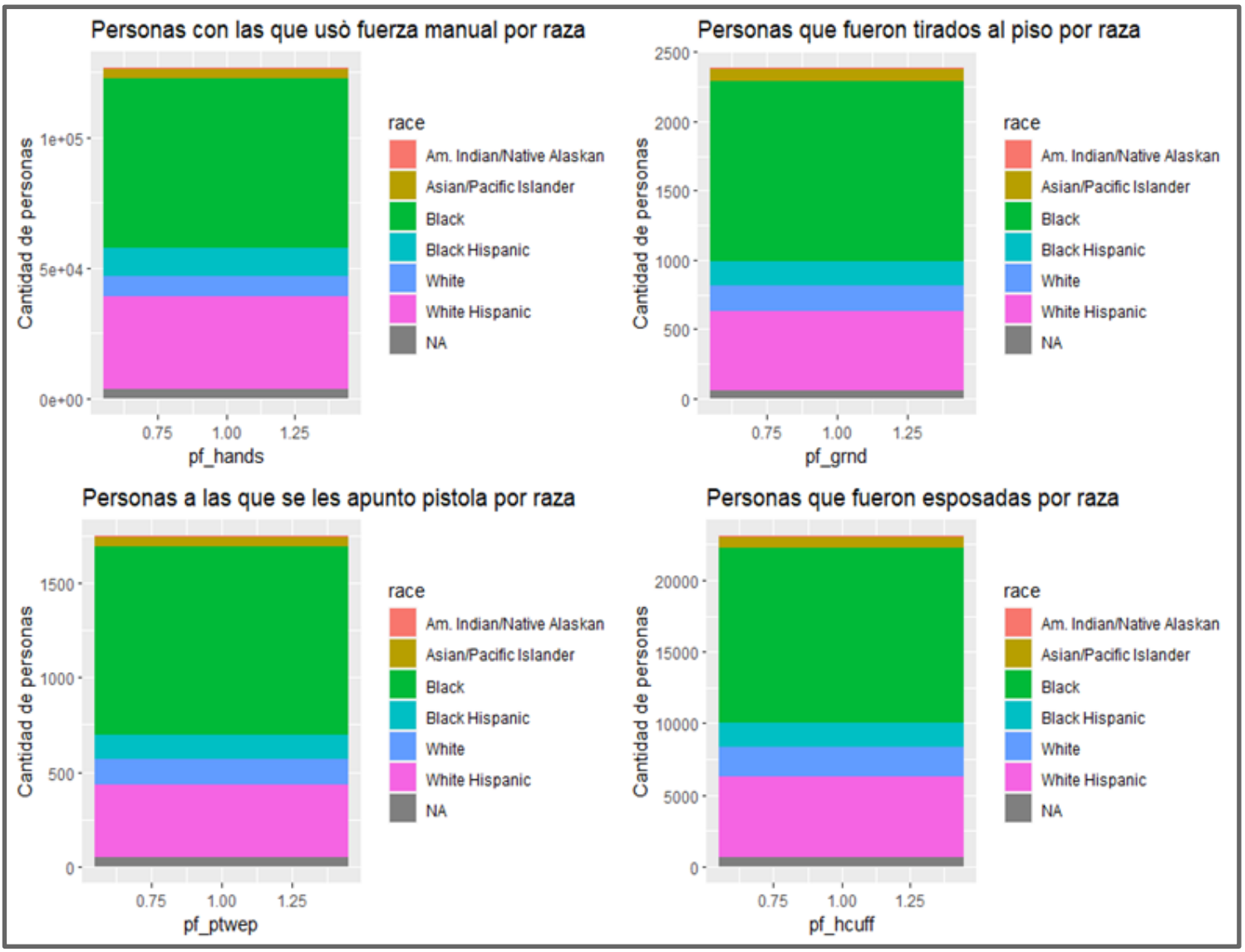
- De las 202 925 personas negras cacheadas solo el 8.65% fueron arrestadas
- De las 28 320 personas hispanas negras cacheada el 8.88% fueron arrestadas
- De las 99 393 personas hispanas blancas cacheadas el 8.79% fueron arrestadas
- De las 27 341 personas blancas cacheadas el 12.88% fueron arrestadas
- De las 11 262 personas asiáticas cacheadas el 10.73% fueron arrestadas
- De las 1 430 personas nativo americanas/alaskeñas cacheadas el 8.67% fueron arrestadas.

La lista anterior nos muestra como parece haber una tasa de crímenes cometidos por personas de diferentes etnias similar, es decir, a pesar de que se detienen más personas de un etnia que de otra las tasas de arresto son similares. Pudiera deberse a que sin importar la etnia la proporción de personas con razones para ser arrestadas es la misma. Por lo que no debería tenerse una idea de que cierta etnia tenga propensión al crimen frente de otra etnia.

A continuación se mostrará la cantidad de personas en las que hubo **uso de fuerza** por etnia, esta gráfica la obtenemos usando la familia de variables **Physical Force used by Officer** en las que se detallan los tipos de **Forceuse** que se aplicaron al detenido, creando una indicadora basada en dichas variables se obtiene una propuesta a la variable perdida llamada **Forceuse.**



Como podemos ver, en general el uso de fuerza en las etnias es proporcional al número de detenidos, es decir, parece no haber diferencia entre las proporciones de personas a las que se les aplicó fuerza por cada etnia. Sin embargo como ya vimos si hay una distinción grande entre las personas detenidas por etnia, siendo tal vez de interés el ver cómo se ejerce fuerza por el total de detenidos.



En general, se observa que casi el 80% de las personas a las que se les aplicó fuerza o alguna táctica de retención son personas negras o hispanas blancas. A pesar de que la tasa de arrestos de la gente negra es menor que la de la gente blanca, es evidente la desproporcionalidad del uso de fuerza hacia la gente negra.

Además del racismo institucionalizado, un motivo por el que se reflejan estas proporciones tan sesgadas se debe a que el número de gente negra detenida abona a más de la mitad de la gente detenida.

Sección 3

Estrategias para evaluar los objetivos del programa

Gracias a la sección 2 del presente documento, se pueden identificar varios problemas con el programa de Stop and Frisk. Siendo el objetivo del programa disminuir las tasas de crimen a través de acciones preventivas hacia personas que levanten sospechas, un primer indicador para evaluar qué tan bien se está cumpliendo este objetivo, sería evaluar la proporción de arrestos realizados al total de las personas detenidas. Esto, para probar la efectividad del programa, ya que si esta proporción fuera elevada indicaría que las detenciones son preventivas a raíz de sospechas objetivas y no arbitrarias.

Del mismo modo se identifica un sesgo muy pronunciado en las detenciones de acuerdo a la variable raza. Un primer indicador para controlar este sesgo es vigilar la proporción de personas detenidas por etnia (race); en grupos demográficos grandes, esta proporción no debería diferir mucho. Por ejemplo, en la ciudad de Nueva York, de acuerdo al U.S Census Bureau, el 42.7% de la población es gente blanca, mientras que el 24.3% es gente negra. Sin embargo, durante el 2011, de las 685 724 personas detenidas, 350 743 fueron gente negra y sólo 61 805 personas blancas fueron detenidas. Tomando en cuenta que Nueva York tiene una población aproximada de 8.3 millones de personas, casi un 18% de la población negra en Nueva York fue arrestada. Mantener estos porcentajes proporcionales al tamaño del grupo demográfico es un primer paso para evitar el sesgo.

Afirmación clasificación

Para esta sección tomaremos la siguiente regla y buscaremos verificar con un modelo de clasificación si existe dicha regla. Nos basamos en el posible hecho de que existe una actitud racista en el programa *"Stop and Frisk"* para hacer la elección de parámetros clasificadores para la regla, al igual que cierto estereotipo físico que ciertos medios de comunicación han propagado en la sociedad de E.U.A.

> Regla: Si se conoce la etnia, el sexo, la complexión y si la persona fue registrada entonces podremos saber si será arrestada.

La regla anterior se buscará confirmar usando las variables **Race**, **Sex**, **Build y Frisked.** La variable que se busca predecir es **arstmade.** Por eventos actuales resulta factible pensar que exista el uso de fuerza (**forceuse**) en los sujetos con las características anteriormente mencionadas, por lo que hacer una verificación para dicha regla también sería de interés, la dejaremos como un resultado extra.

La matriz de predicción para este modelo fue la siguiente:

Predicción\Referente	**No**	**Sí**
No	184,113	11,877
Sí	0	0

La matriz de predicción podría parecer un poco extraña, pero la razón de que en las predicciones ninguna persona será arrestada se debe a que, para el modelo que se entrenó, nadie debería ser arrestado únicamente por su etnia, sexo, complexión física y por el hecho de haber sido registrado por un policía.

El modelo califica a los usuarios en su totalidad como si estos no debieran ser arrestados siendo que, si hay casos en los que son arrestados, este tipo de detalles en el modelo se pueden ver en los índices de sensibilidad y especificidad, con valores de 1 y 0 respectivamente, se habla de ellos más adelante.

De igual forma, de la matriz de predicción notamos que de las 195,990 personas que se predice que no serán arrestadas, en realidad 11,877 sí lo serán.

Ahora, a partir de la matriz de predicción, se pueden obtener los indicadores de exactitud del modelo:

Exactitud	0.9394
Sensibilidad	1
Especificidad	0
Precisión	0.9394

Observamos que el 93.94% de las tuplas fueron correctamente clasificadas, por lo que la tasa de verdaderos positivos es 1 (sensibilidad) y se tiene una precisión del 93.94%. Lo que quiere decir que nuestro modelo es confiable en el sentido de que no cometerá errores de injusticia, como dijimos parece que no predice arresto a ningún miembro de la base con la que se testeó.

Con lo anterior se puede confirmar que tristemente vivimos en una sociedad en el que los prejuicios aún existen y, más triste aún, se encuentran presenten en autoridades como la policía. Pues aunque nadie debería ser arrestado por las razones previamente mencionadas, un 6.06% de las personas sí lo fueron; pareciera un porcentaje bajo pero representa a 11,877 personas en total, las suficientes para llenar un estadio de fútbol de mediana capacidad.

Afirmación Predicción

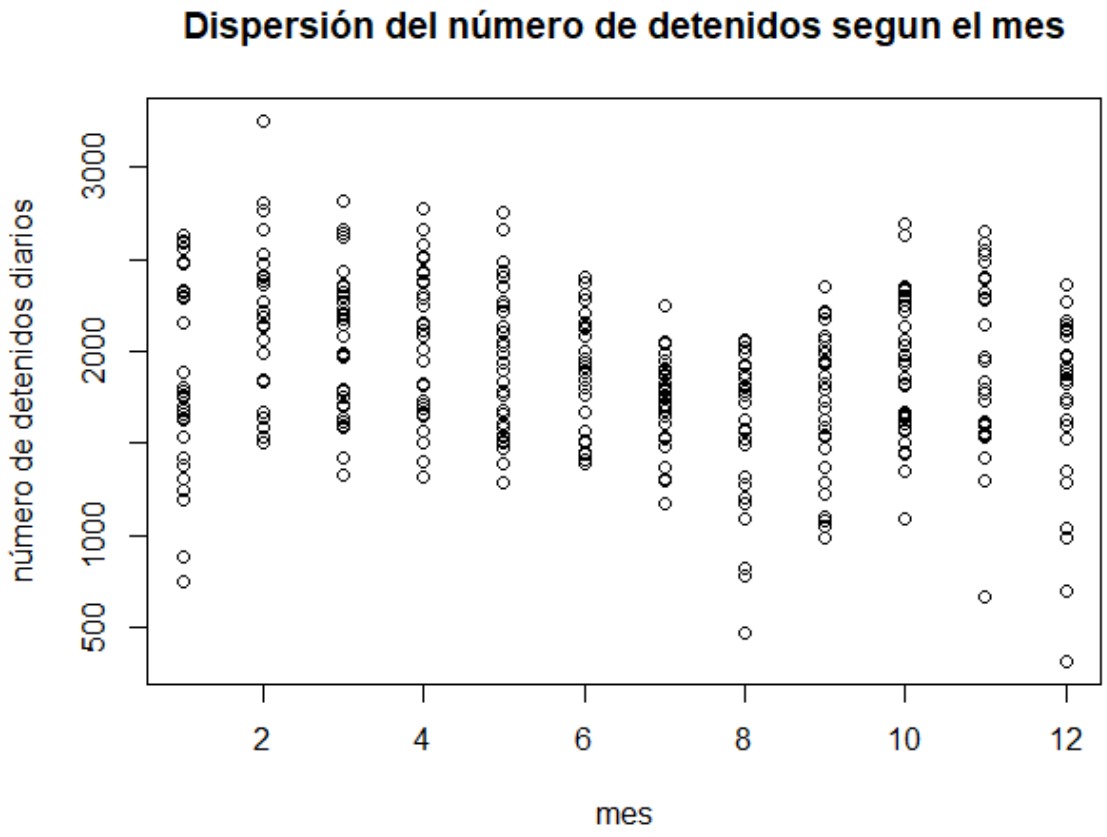
Se propone como afirmación que *el número de personas arrestadas por día en función del número de detenidos de ese mismo día y del mes correspondiente a dicho día se predice a través del modelo*.

La afirmación anterior daría pauta a reflexionar sobre si existe algún efecto por temporadas en los arrestos, como por ejemplo, si suele haber mayor número de arrestos en el mes de abril y por ende ese mes las estaciones policiales deberían prepararse para tener atender a una mayor demanda de recursos.

Se realizan los cálculos pertinentes para conocer el número de detenidos por día, además del mes al que corresponde dicho día.

Análisis de la relación entre variables

En la siguiente gráfica podemos observar la dispersión entre la variable mes y detenidos diarios.



Podemos notar que hay meses en los que en general hay menos detenidos que otros meses, como por ejemplo agosto. En cambio en febrero el número de detenidos parece ser de los más altos del año. Sería interesante verificar si estas "altas" y bajas" coinciden con las del número de arrestos, en cuyo caso podría haber una correlación entre la variable número de detenidos diarios y mes correspondiente.

Generación del Modelo

Se tiene el siguiente modelo de regresión lineal múltiple:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \epsilon$$

De R obtenemos las estimaciones de los parámetros para formar el primer modelo (todas las variables), por lo que la ecuación del primer modelo ajustado es:

$$y = -13.15151 + 0.066079x_1 + 14.13979x_2 + 9.524790x_3 + 8.133871x_4 + 0.385993x_5 + 0.233243x_6$$

$$-3.090633x_7 + 1.379571x_8 + 7.048617x_9 - 6.823077x_{10} - 19.812731x_{11} - 9.629802x_{12} + \epsilon$$

Donde:

x_1: numstop

x_2: enero

x_3: febrero

x_4: marzo

x_5: abril

x_6: mayo

x_7: junio

x_8: julio

x_9: agosto

x_{10}: septiembre

x_{11}: octubre

x_{12}: noviembre

Se observa que el p-valor del estadístico F (< 2.2e-16)fue muy pequeño , por lo que el modelo fue significativo. Es decir, el número de arrestos diarios se relaciona con el número de detenidos diarios y el mes correspondiente.

Por otra parte, se obtuvo un valor de la R^2 ajustada de 0.6997, lo que quiere decir que al incluir estas trece variables independientes en el análisis se explica un 69.97% de la varianza del número de detenidos por día.

Selección de los mejores predictores

Se utilizó el método de eliminación hacia atrás en combinación con selección hacia adelante con el comando *step(regresion, direction = "backward", trace = 1)* para seleccionar el

mejor modelo. Este método consiste en empezar con un modelo que contenga todas las variables regresoras y en cada paso se reduce el modelo hasta que llegue a un nivel de ajuste óptimo, con las menor cantidad de regresoras y el mejor grado de ajuste. De las salidas de R se obtuvo el modelo y el AIC correspondiente a cada modelo:

- Start: AIC= 1757.2

arstmade = numstop + ene + feb + mar + abr + may + jun + jul + ago + sep + oct + nov

- Step 1: AIC= 1755.2

arstmade = numstop + ene + feb + mar + abr + jun + jul + ago + sep + oct + nov

- Step 2: AIC= 1753.2

arstmade = numstop + ene + feb + mar + jun + jul + ago + sep + oct + nov

- Step 3: AIC= 1751.27

arstmade = numstop + ene + feb + mar + jun + ago + sep + oct + nov

- Step 4: AIC= 1749.94

arstmade = numstop + ene + feb + mar + ago + sep + oct + nov

Parámetros del modelo final:

Variable independiente	**Coeficiente**
intercepto	-13.29489
numstop	0.06602
ene	14.39071
feb	9.79896
mar	8.39855
ago	7.28613
sep	-6.58008
oct	-19.55469
nov	-9.37396

El modelo resultante conservó las variables numstop, ene, feb, mar, ago, sep, oct y nov; el resto de las variables no fueron significativas para el modelo. Se selecciona este modelo ya que es el que tiene el AIC más bajo. El criterio de información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico.

Adecuación del modelo

Posteriormente se obtiene el p-valor del modelo, así como los valores de R cuadrada y R ajustada:

R^2	R^2 ajustada	p-valor
0.7072	0.6989	< 2.2e-16

Se observa que el valor de la R^2 ajustada es mayor que la que se obtuvo en el primer modelo (0.6997) a pesar de que se eliminaron 4 variables del modelo. Lo que quiere decir que al quedarnos con dichas ocho variables independientes en el análisis se explica un 69.89% de la varianza de la variable dependiente. Esto indica que hubo una correcta selección de variables que se mantienen en el modelo ya que la proporción de la variación explicada por los regresores no solo no disminuyó sino aumento.

La ecuación del modelo final es:

$$\widehat{arstmade} = -13.29489 + 0.06602x_1 + 14.39071x_2 + 9.79896x_3 + 8.39855x_4 + 7.28613x_5 - 6.58008x_6 - 19.55469x_7 - 9.37396x_8$$

El cual refleja que por una persona más detenida se incrementa la cantidad de arrestos en un 0.06602 si las otras variables permanecen constantes; para la variable enero, tenemos que hay 14.39071 arrestos más por día en el mes de enero siempre y cuando las demás variables permanezcan constantes; para la variable febrero, tenemos que hay 9.79896 arrestos más por día en el mes de febrero siempre y cuando las demás variables permanecen constantes; para la variable marzo, tenemos que hay 8.39855 arrestos más por día en el mes de marzo siempre y cuando las demás variables permanezcan constantes; para la variable agosto, tenemos que hay 7.28613 arrestos más por día en el mes de agosto siempre y cuando las demás variables permanezcan constantes; para la variable septiembre, tenemos que hay 6.58 arrestos más por día en el mes de septiembre siempre y cuando las demás variables permanezcan constantes; para la variable octubre, tenemos que hay 19.5546 arrestos más por día en el mes de octubre siempre y cuando las demás variables permanezcan constantes; para la variable noviembre, tenemos que hay 9.37396 arrestos más por día en el mes de enero siempre y cuando las demás variables permanezcan constantes.

De igual manera el p-valor del estadístico F es muy pequeño (< 2.2e-16) para un nivel de significancia de 0.05, por lo que el modelo es significativo.

Estimación de los intervalos de confianza para los coeficientes del modelo

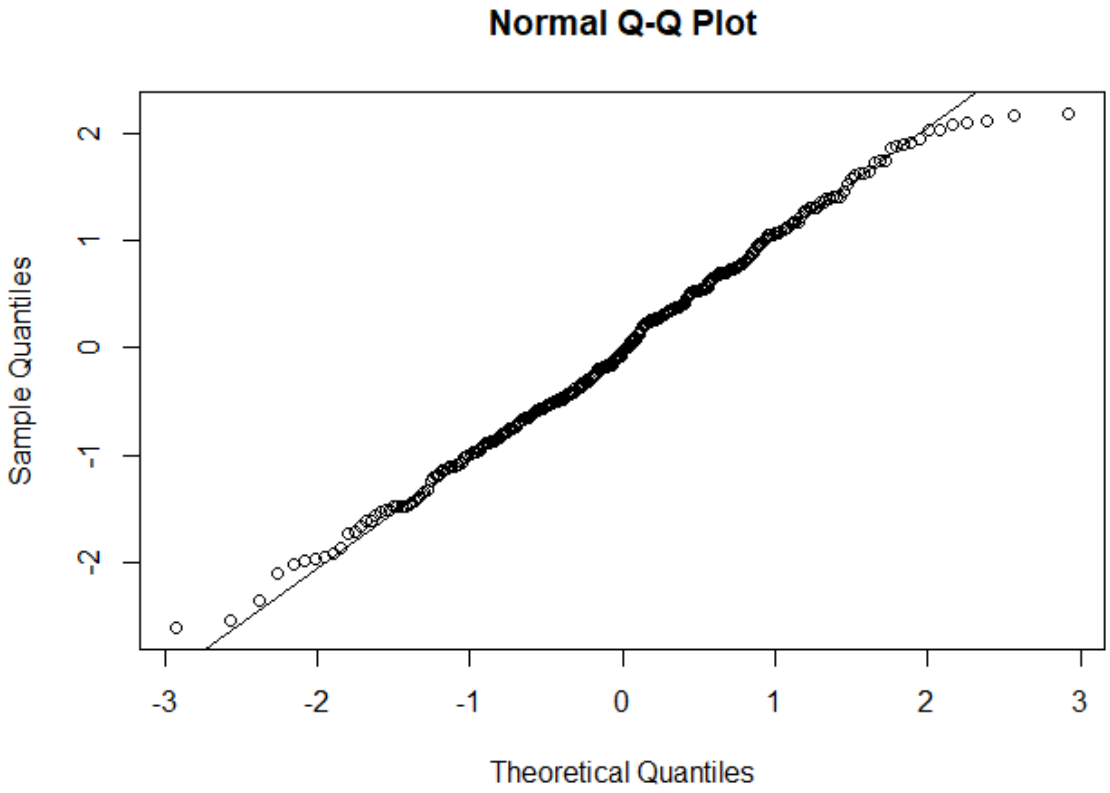
Los intervalos de confianza con un alpha de 2.5% para los parámetros del modelo resultan de la siguiente manera:

Parámetro	Límite inferior	Límite superior
β_0	-25.534	-3.7086
β_1	0.0601	0.0713
β_2	7.7979	24.6625
β_3	2.6556	20.8431
β_4	1.6654	18.9420
β_5	-6.76	11.9171
β_6	0.5058	17.6175
β_7	-13.7014	4.1447
β_8	-26.4095	-8.9529

Comprobación de supuestos

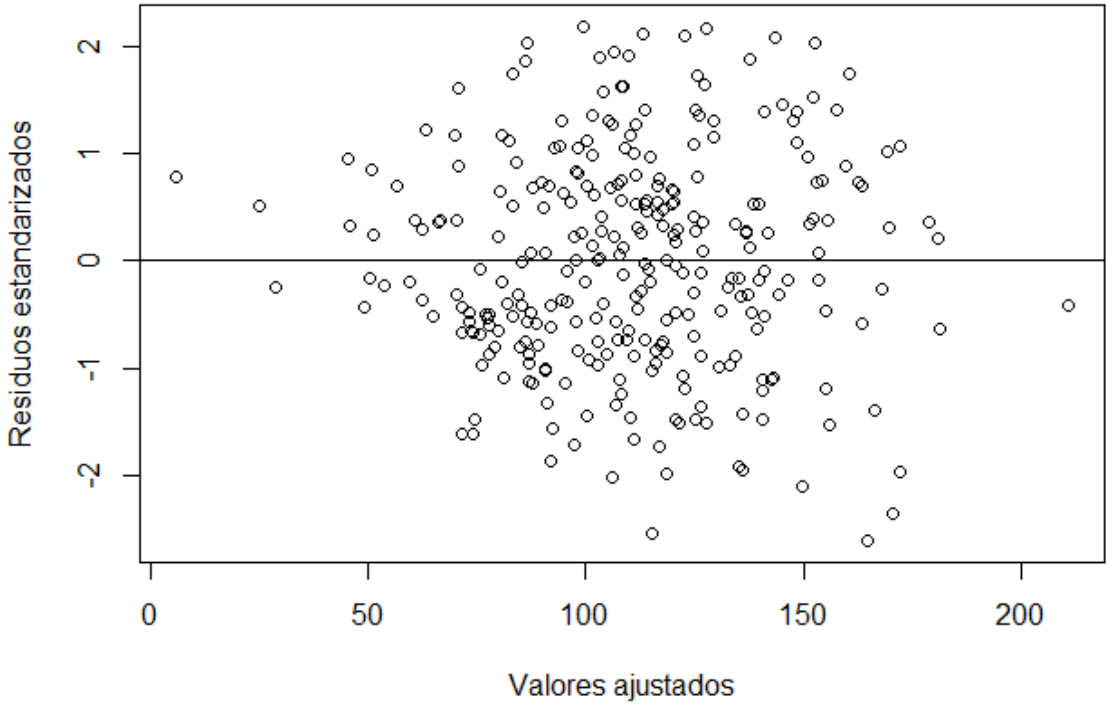
Es importante verificar los supuestos que requiere una regresión lineal para conocer si el modelo propuesto tiene validez. Dichos supuestos a verificar son:

- Normalidad



En el gráfico de probabilidad normal puede verse que la mayoría de puntos se encuentran sobre la línea de referencia, por lo que parece suponer que se cumple el supuesto de normalidad. Para verificar de manera más certera, se aplicó una prueba de Shapiro Wilks, donde se obtuvo un p-valor de 0.1984, por lo que al ser mayor que Alpha = 0.05 no se rechaza la hipótesis nula y se concluye que los residuos cumplen con el supuesto de normalidad.

- Homocedasticidad, Linealidad e Independencia de los residuos



En el gráfico anterior, correspondiente a *valores ajustados vs residuos*, se puede observar que la varianza de los errores es constante debido a que en su mayoría los puntos se encuentra entre la franja del 2 y el -2, por lo que el gráfico parece sugerir que se cumple el supuesto de homoestacidad. Asimismo no se observa ningún patrón en los residuales, indicando que se cumple el supuesto de independencia. Finalmente se tiene una cantidad similar de puntos arriba y debajo de la línea, comprobando así la linealidad en los errores. Por lo que se concluye que se cumplen los supuestos necesarios.

Predicción

_____El comando utilizado en R para obtener la predicción fue:

prediccion_IC = predict(modelo, test, interval = "prediction", level = 0.95)

Mes correspondiente	**Número de detenidos**	**Predicción**	**Límite Abril,inferior**	**Límite superior**	**Real**
oct	1566	70.63	30.70	110.57	76

oct	1653	76.35	36.44	116.27	92
ene	2483	164.82	124.82	204.83	178
sep	1938	107.99	68.04	147.94	117
ago	818	48.21	8.15	88.26	53
may	1391	76.81	37.49	116.13	67
dic	696	31.13	-8.63	70.89	36

Dicho comando para obtener la predicción para la el número de arrestados diarios y sus correspondientes intervalos de confianza para la respuesta media y predicción, generado para un número de detenidos entre 696 a 2483 y entre distintos meses a lo largo del año. Cabe mencionar que estos valores corresponden a las primeras ocho tuplas que forman parte de nuestra base de datos de prueba o test set. Se seleccionaron ocho de manera ilustrativa.

En la tabla, las dos primeras columnas corresponden a los valores que toman las variables independientes, número de detenidos diario y el mes correspondiente a ese día. Seguidas de estas, están las tres columnas correspondientes a el valor que se predice para la variable independiente (número diario de arrestos) seguido de los extremos del intervalo. Por último, se encuentra la columna correspondiente al valor real para la variable dependiente. Con dicha columna podemos darnos una idea de la confiabilidad de nuestras predicción. Podemos notar que hay valores que no son enteros e incluso negativos, lo cual no tiene sentido al ser nuestra unidad de medida número de personas. Sin embargo esto se arregla fácilmente tomando el entero más cercano o parte entera (dependiendo de los objetivos del modelo) y tomando como límite inferior mínimo o al número cero.

La tabla nos brinda, por ejemplo, la siguiente información: el número de arrestados diarios en un día correspondiente al mes de octubre y al haber sido detenidas 1566 personas es de 70-71 personas con un intervalo para la respuesta media de [30.71, 110.57]. Además, podemos notar que el error en dicha predicción fue de 5-6 personas, dado que el valor real de arrestados en dicho día fue de 76 personas.

Conclusión

A modo de conclusión regresamos a reflexionar sobre los efecto por temporadas en los arrestos. El modelo parece indicar que efectivamente hay una estacionalidad en el número de arrestos en conjunto con el número de detenidos. Por lo que esto se podría tomar en cuenta para que las estaciones policiales se preparen para atender a una mayor demanda en los meses de enero, febrero y marzo. Lo cual se podría lograr al hacer una mejor distribución de los recursos anuales que reciben en función de la demanda que sugiere este modelo.

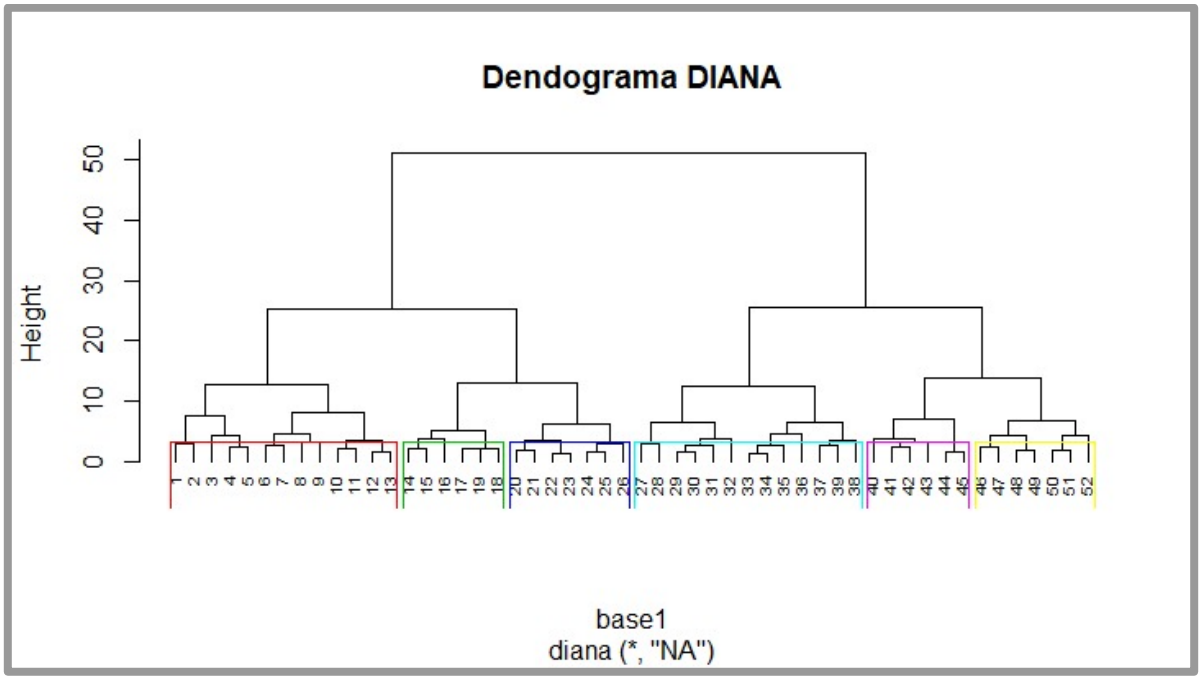
Afirmación clustering

Por la naturaleza del análisis de cluster se pueden proponer afirmaciones que busquen encontrar grupos que talvez no habíamos considerado. Tomando en cuenta que nuestro número de variables es grande, de aproximadamente 100 en la base de datos, de los cuales ya sabemos que no todos son consistentes, entonces procederemos a buscar grupos filtrando las variables a considerar, es decir, no usaremos todas las variables de la base y mucho menos aquellas que consideramos tienen problemas.

Se realizó una modificación a la base de datos, ya que el número de tuplas de la base de datos complicaba el análisis de clusters. Utilizando el atributo **Datestop** se clasificaron los datos en las 52 semanas del año, de acuerdo con la fecha en la que se realizó el registro. De esta manera, se redujo a una base de datos de 52 tuplas. De igual forma se contabilizaron los casos por cada semana del año agrupándolos por la familia de variables **razón del registro**, teniendo al final una base con el número de detenidos agrupados por todos los motivos posibles en cada una de las semanas.

Una vez reducida la base de datos, estandarizamos los valores, ya que en algunos casos, se registraron una gran cantidad de detenidos por algún motivos, mientras que en otros era menor la cantidad de detenidos. De modo que estandarizar los datos nos ayuda a hacer un mejor análisis.

El dendograma resultante fue el siguiente:

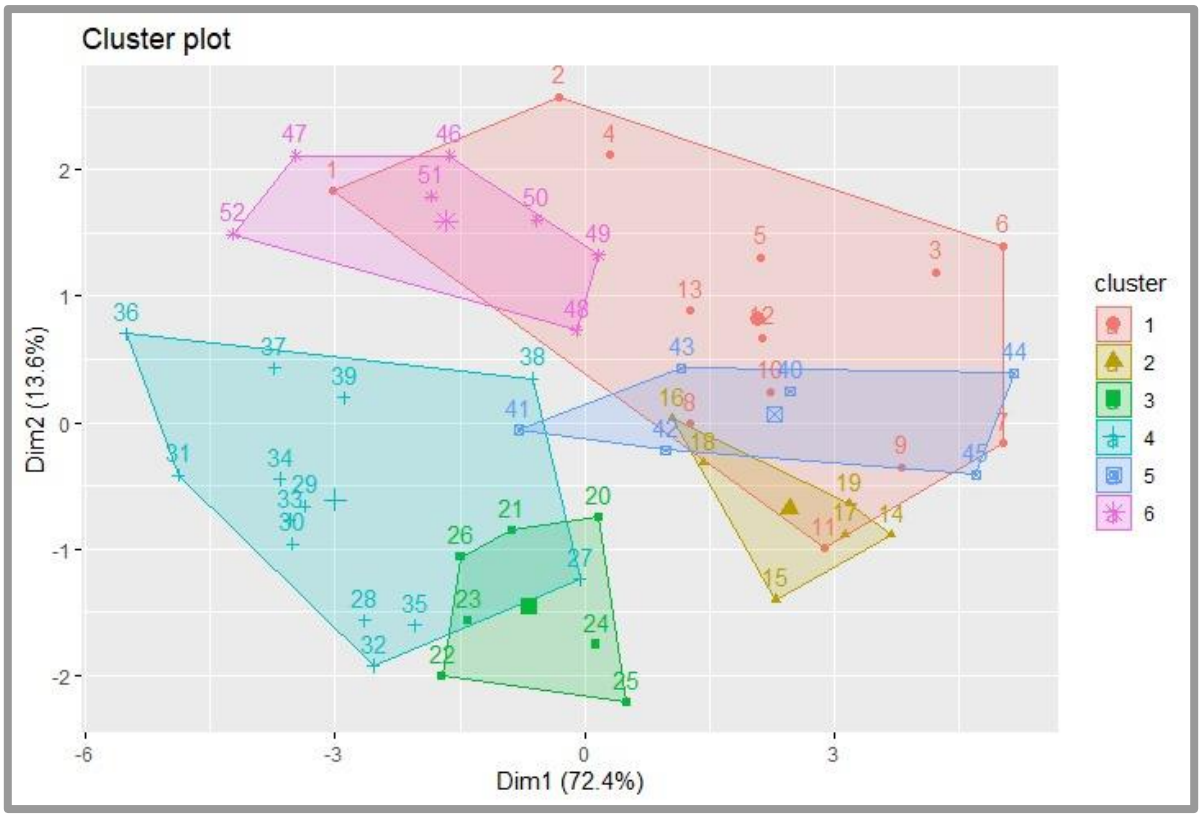


Con base en el dendograma, se puede ver que los clusters están integrados de la siguiente forma:

Cluster	Semanas	Tamaño	Meses
1	1 a 13	13	Enero, Febrero, Marzo
2	14 a 19	6	Abril, primera quincena de Mayo
3	20 a 26	7	Segunda quincena de Mayo, Junio
4	27 a 39	13	Julio, Agosto, Septiembre
5	40 a 45	6	Octubre, primera quincena de Noviembre
6	46 a 52	7	Segunda quincena de Noviembre, Diciembre

Algo que llama la atención es que los clusters se dividieron de tal forma que en cada semestre del año (Enero-Junio y Julio-Diciembre) se formaron 3 clusters de tamaños 13, 6 y 7, en ese orden. Mostrando tal vez la presencia de un patrón semestral.

También puede analizarse el gráfico de los clusters:



Los clusters 1 y 6, que corresponden al inicio y fin del año son aquellos que tienen más celebraciones en la ciudad de Nueva York:

- Cluster 1: Año Nuevo, Día de Martin Luther King, Nacimiento de Lincoln, Día del Presidente, Día de la Marmota, Año Nuevo Lunar (o Año Nuevo chino), San Valentín y San Patricio, para un total de 8 celebraciones
- Cluster 6: Día de Acción de Gracias, Encendido del Árbol de Navidad en el Rockefeller Center, Día de Navidad, Carrera de Medianoche y Año Nuevo en Times Square.

Además, en el gráfico de los clusters, se puede ver que la semana uno también podría quedar dentro del cluster 6 en el plano de componentes principales. Esto sugiere que puede existir una relación entre la criminalidad entre semanas y el número de celebraciones que se llevan a cabo en dicho periodo de tiempo.

Conclusión Final

A lo largo del reporte se tuvo como interés encontrar patrones y tendencias en los datos del programa, en esta última sección de afirmaciones hemos encontrado algunas que interesantes y reafirmado otras que nos parecían ciertas. Desde el comienzo de las afirmaciones habíamos visto la discriminación que se tenía en el programa a las personas Negras y como esto según los datos no mostraban tener un sustento real.

Sin embargo alejándonos algo de esa primera parte intentamos buscar patrones con temática algo diversificada, desde la sección de clasificación donde quisimos probar si había una regla que me pudiera justificar si una persona por características físicas y culturales sería arrestada hasta la última sección de clustering donde buscamos patrones entre las semanas del año con base a las razones de arresto.

Podemos decir que la limitante computacional al inicio fue un problema para trabajar con estos datos pero al final con el uso de mejor código y preguntas más concisas pudimos ver que el programa tiene un problema Stop and Frisk tiene mucho que mejorar y cosas interesantes que enseñarnos sobre lo que ser neoyorkino se refiere.