Trabajo Práctico Final

Ezequiel Victorel

Universidad Nacional de Luján
Departamento de Ciencias Básicas
Base de Datos Masivas (Gestión y Análisis) - 11088

2017

Resumen

El gran volumen de datos actual, lleva a considerar una muy confiable fuente de conocimiento. Pero el realizar un proceso de extracción manual por parte de distintos individuos ya no es eficiente ni eficaz, por esto, surge la necesidad de comenzar a utilizar un proceso automático que permita extraer conocimiento fidedigno de estos.

El presente trabajo de abarco como objetivo principal el analizar los crímenes reportados a lo largo de los años en la ciudad de Filadelfia, Pensilvania, Estados Unidos. De este objetivo principal, se determinaron otros que permitieron analizar las tasas de crímenes en los tipos de crímenes, las distintas subdivisiones geográficas de la ciudad, distintos lapsos de tiempo, obtención de zonas propensas a ciertos crímenes, etc.

Pero debido a la gran cantidad de crímenes reportados a lo largo de 10 años, para poder cumplir con estos objetivos, se debió utilizar el proceso *Knowledge Discovery in Databases* para así realizar: Un análisis exhaustivo de distintas fuentes de datos para satisfacer los objetivos, un proceso extracción, transformación y carga de tales fuentes que involucro uso de distintas técnicas (Por ejemplo geolocalización inversa), un proceso de minería de datos y, por último, una interpretación de resultados.

Índice general

Resumen II

1. Introducción 1

2. Objetivos 2
   2.1. Objetivo general 2
   2.2. Objetivos específicos 2

3. Materiales y métodos 3
   3.1. Herramientas 3
   3.2. Metodología 3

4. Desarrollo 5
   4.1. Selección 5
   4.2. Preprocesamiento 6
      4.2.1. Limitación de línea de tiempo 6
      4.2.2. Inconsistencias 6
         4.2.2.1. Distritos policiales 6
         4.2.2.2. Calles 7
      4.2.3. Ruido 9
         4.2.3.1. Temperaturas 9
         4.2.3.2. Archivos GeoJSON 10
      4.2.4. Datos faltantes 11
         4.2.4.1. Clasificación UCR 11
         4.2.4.2. Latitud y longitud 12
   4.3. Transformación 12
      4.3.1. Estructura lógica 12
      4.3.2. Transformación y carga 13
         4.3.2.1. Generación de Barrios 13
         4.3.2.2. Unión de crímenes con temperaturas 14
         4.3.2.3. Carga de calles 14
         4.3.2.4. Normalización de datos 15
         4.3.2.5. Generación de metadatos 15
         4.3.2.6. Unión y carga 17
         4.3.2.7. Generación y ejecución del trabajo 20
   4.4. Análisis de datos 21
      4.4.1. Ranking por cantidad de crímenes 21
      4.4.2. Ranking por tipo de crímenes 21
         4.4.2.1. Ciudad de Filadelfia 21

4.4.2.2. Barrios ... 21
4.4.2.3. Distritos policiales ... 22
4.4.3. Zonas de concentración ... 22
4.4.4. Divisiones geográficas con mayor y menor incremento ... 22
4.4.5. Análisis por intervalos de tiempos ... 23
4.4.5.1. Días y horas ... 23
4.4.6. Correlación Crimen / Temperaturas ... 23
4.4.6.1. Análisis de temperaturas ... 23
4.4.6.2. Correlación ... 24
4.4.7. Zonas propensas a la ocurrencia de crímenes ... 25

5. Resultados ... 28
5.1. Ranking por cantidad de crímenes ... 28
5.1.1. Barrios ... 28
5.1.2. Distritos policiales ... 29
5.1.3. Calles ... 30
5.2. Ranking por tipo de crímenes ... 30
5.2.1. Ciudad de Filadelfia ... 31
5.2.2. Barrios ... 32
5.2.3. Distritos policiales ... 34
5.3. Zonas de concentración ... 36
5.4. Divisiones geográficas con mayor y menor incremento ... 36
5.4.1. Barrios ... 37
5.4.2. Distritos policiales ... 38
5.5. Análisis por intervalos de tiempos ... 39
5.5.1. Años ... 39
5.5.2. Estaciones ... 40
5.5.3. Meses ... 40
5.5.4. Días ... 41
5.5.5. Horas ... 42
5.6. Correlación Crimen / Temperaturas ... 43
5.7. Zonas propensas a la ocurrencia de crímenes ... 45
5.7.1. Cluster 1 ... 45
5.7.1.1. Barrios ... 45
5.7.1.2. Tipos de crímenes ... 46
5.7.1.3. Tasas por lapsos de tiempo ... 47
5.7.1.4. Distribución de tasas por barrio ... 47
5.7.1.5. Tiempo ... 48
5.7.2. Cluster 2 ... 49
5.7.2.1. Barrios ... 49
5.7.2.2. Tipos de crímenes ... 50
5.7.2.3. Tasas por lapsos de tiempo ... 50
5.7.2.4. Distribución de tasas por barrio ... 51

5.7.2.5. Tiempo 52
5.7.3. Caracterización principal 52
5.7.3.1. Cluster 1 52
5.7.3.2. Cluster 2 52

6. Conclusiones 54

Bibliografía 55

Apéndice 56
Apéndice A: Estructura de “Philadelphia Crime Data” 56

Índice de ilustraciones

Ilustración 4.1: Diagrama de caja de las temperaturas históricas 9
Ilustración 4.2: Ejemplos de distribuciones de temperaturas diarias 10
Ilustración 4.3: Ejemplos de las diferencias entre las coordenadas y el archivo GeoJSON 10
Ilustración 4.4: Diagrama lógico de la base de datos principal 13
Ilustración 4.5: Barrios de la ciudad 14
Ilustración 4.6: Carga de calles 14
Ilustración 4.7: Normalización de crímenes 15
Ilustración 4.8: Eliminación temporal de coordenadas atípicas por distrito policial 16
Ilustración 4.9: Carga de metadatos 17
Ilustración 4.10: Unificación y carga de crímenes 18
Ilustración 4.11: Generación de mapping entre calle y calles fiables 19
Ilustración 4.12: Unión y carga - Crimen/Calles 20
Ilustración 4.13: Trabajo principal 20
Ilustración 4.14: Cantidad de crímenes en función de la temperatura 24
Ilustración 4.15: Cantidad óptima de clusters 26
Ilustración 4.16: Cantidad de elementos por clusters 26
Ilustración 5.1: Ranking 25 - Cantidad de crímenes por barrio 28
Ilustración 5.2: Ranking - Cantidad de crímenes por distrito policial 29
Ilustración 5.3: Ranking 25 - Cantidad de crímenes por calle 30
Ilustración 5.4: Cantidad de crímenes por tipo de crimen 31
Ilustración 5.5: Cantidad de crímenes por barrios más peligrosos (25) y tipo de crimen (3) 32
Ilustración 5.6: Mapa con el crimen más frecuente por barrio 33
Ilustración 5.7: Cantidad de crímenes por distrito policial y tipo de crimen (3) 34
Ilustración 5.8: Mapa con el crimen más frecuente por distrito policial - Solo reportes 35
Ilustración 5.9: Mapa con zonas de concentración 36
Ilustración 5.10: Barrio con mayor reducción en la tasa de crímenes 37
Ilustración 5.11: Barrio con mayor incremento en la tasa de crímenes 37
Ilustración 5.12: Distrito policial con mayor reducción en la tasa de crímenes 38
Ilustración 5.13: Distrito policial con mayor incremento en la tasa de crímenes 38
Ilustración 5.14: Cantidad de crímenes por año y tipo de crimen 39
Ilustración 5.15: Cantidad de crímenes por estación y tipo de crimen 40
Ilustración 5.16: Cantidad de crímenes por mes y tipo de crimen 40
Ilustración 5.17: Cantidad de crímenes por día 41
Ilustración 5.18: Cantidad de crímenes por hora 42
Ilustración 5.19: Cantidad de crímenes por hora y tipo de crimen 42
Ilustración 5.20: Cantidad de crímenes por temperatura y tipo de crimen - Parte 1 43
Ilustración 5.21: Cantidad de crímenes por temperatura y tipo de crimen - Parte 2 44
Ilustración 5.22: Mapa con cantidad de días observados por barrio - Cluster 1 45
Ilustración 5.23: Cantidad de crímenes por tipo de crimen - Cluster 1 46
Ilustración 5.24: Histogramas de tasas por lapsos de tiempo - Cluster 1 47

Ilustración 5.25: Mapa con distribución de tasas promedio por barrio - Cluster 1 48
Ilustración 5.26: Serie de tiempo de los promedios de tasas - Cluster 1..... 48
Ilustración 5.27: Mapa con cantidad de días observados por barrio - Cluster 2 49
Ilustración 5.28: Cantidad de crímenes por tipo de crimen- Cluster 2..... 50
Ilustración 5.29: Histogramas de tasas por lapsos de tiempo - Cluster 2..... 51
Ilustración 5.30: Mapa con distribución de tasas promedio por barrio - Cluster 2 51
Ilustración 5.31: Serie de tiempo de los promedios de tasas - Cluster 2..... 52

1. Introducción

A lo largo de la materia, se vio que en la actualidad existe un aumento masivo en la cantidad de datos disponibles sobre distintos contextos. La existencia de los mismos, llevo a considerar un nuevo paradigma de ciencia, donde a partir de los datos se puede generar conocimiento fidedigno.

Ahora, si bien antes era posible realizar un análisis manual por el propio usuario, tal aumento produjo la necesidad de crear un proceso que permita una automatización eficiente para poder extraer conocimiento fiable que permita en un futuro, una toma de dedición eficiente y efectiva.

Es por esto, que el trabajo consiste en la utilización de los distintos conocimientos adquiridos a lo largo de toda la materia sobre un caso real, con el objetivo de generar conocimiento que permita dar soporte en la toma de decisiones.

En este caso en particular, se toma como caso de estudio fuentes de datos que contengan información sobre crímenes reportados a lo largo de los años en la ciudad de Filadelfia, Pensilvania, Estados Unidos. A partir de estas fuentes, se procederá al análisis de las mismas con el objetivo de descubrir conocimiento como, por ejemplo, responder a tales como ¿Qué zonas de la ciudad son las más inseguras?, ¿En qué horarios es conveniente transitar por la ciudad?, etc.

2. Objetivos

2.1. Objetivo general

El principal objetivo del trabajo es encontrar posibles factores que influyen en la tasa de crímenes de la ciudad, junto con poder establecer zonas seguras e inseguras en la misma.

2.2. Objetivos específicos

Los objetivos específicos para llegar al objetivo principal son:

- Obtener y analizar un ranking según la cantidad de crímenes por:
  - Barrios
  - Distritos policiales
  - Calles
- Obtener y analizar un ranking de los tipos de crímenes por:
  - Ciudad de Filadelfia
  - Barrios
  - Distritos policiales
- Analizar las zonas con mayor concentración de crímenes, independientemente de las posibles subdivisiones geográficas del condado.
- Obtener según sea el caso, aquella subdivisión geográfica (Barrios y distritos policiales) con mayor reducción e incremento en la tasa de crímenes, y analizar su variación a lo largo de los años.
- Analizar la cantidad de crímenes por intervalos de tiempo, según sea:
  - Por año
  - Por estaciones
  - Por día del mes
  - Por hora del día
- En base a datos diarios de temperaturas de la ciudad, analizar en busca de una correlación entre la tasa de un crimen en particular y las temperaturas históricas.
- Por cada barrio de la ciudad, se analizarán los reportes para verificar si en los anteriores 5, 15, 30 días no ha habido ningún otro reporte del crimen en particular, para así luego, mediante una técnica de agrupamiento, analizar cada grupo en busca de zonas más y menos propensas a la ocurrencia de ciertos crímenes en particular.

3. Materiales y métodos

3.1. Herramientas

Las herramientas utilizadas fueron:

- **Data Integration - Kettle:** Es parte de una suite que posee herramientas para la realización de procesos de ETL. En este caso particular, fue utilizado como herramienta ETL en la etapa de transformación de datos junto con una parte de preprocesamiento.
- **Python**: Este lenguaje fue utilizado para la manipulación de elementos HTML y XML, donde los mismos fueron transformados al formato CSV. Además, se utilizó para la clusterización realizada en uno de los objetivos debido a su manejo de memoria.
- **Lenguaje R:** Este lenguaje fue aquel con mayor peso en todo el trabajo debido a que fue utilizado en la mayor parte del preprocesamiento, en las etapas iniciales de la transformación y, por último, en los análisis finales.
- **PostgreSQL:** Fue utilizado como motor de base de datos para poder realizar la carga de los resultados en la etapa de transformación, junto con su utilización en la etapa de minería de datos.

3.2. Metodología

La metodología utilizada estuvo basada en conceptos del proceso *Knowledge Discovery in Databases* (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) que permite establecer un conjunto de etapas que permitirán automatizar el proceso de minería de datos con el principal objetivo de extraer conocimiento de los datos de origen.

Para este trabajo se realizaron las siguientes etapas:

- **Identificación de objetivos:** Este paso ya fue descripto en la sección 2, donde se listan los distintos objetivos del proyecto.
- **Selección:** Esta etapa conllevo la selección y obtención de las distintas fuentes de datos que permitieron generar conocimiento. Se debió tener en cuenta los objetivos anteriores para así poder comprender los datos que son necesarios, y seleccionar las fuentes de los mismos. (Ver la sección 4.1)
- **Pre procesamiento:** En esta etapa se mejoró la fiabilidad de los datos removiendo ruidos, tratando los datos faltantes e inconsistencias. (Ver la sección 4.2)
- **Transformación:** En esta etapa, se realizaron múltiples eliminaciones de atributos, agregaciones, generalizaciones, normalización, etc. Luego, estos fueron

cargados en una estructura eficiente para así poder realizar la próxima etapa. (Ver la sección 4.3)

- **Minería de datos:** Mediante técnicas adecuadas según los objetivos, se generaron representaciones de alto nivel para poder ser analizadas con facilidad. (Ver la sección 4.4 y 5)
- **Evaluación e interpretación:** Los resultados de la etapa anterior fueron analizados e interpretados objetivamente para poder generar conocimiento. (Ver la sección 5)

4. Desarrollo

4.1. Selección

Antes que nada, debido a que en los objetivos propuestos existen análisis a nivel Tiempo/Año, las fuentes de datos estuvieron limitadas a las fechas entre 01/01/2006 y 31/12/2016.

Para poder cumplir con los objetivos propuestos de este trabajo fue necesario la selección de distintas fuentes de datos. Aquellos datos que permitieron tal fin son:

- Tipo de crimen
- Fecha
- Hora
- Calle
- Barrio
- Distrito policial
- Temperatura
- Coordenadas geográficas

En base a tales datos, y al dataset otorgado por los responsables de la materia, las fuentes seleccionadas fueron:

- **Philadelphia Crime Data**: Mediante el sitio web Kaggle[1] se obtuvo un dataset en formato CSV con la información necesaria para el establecimiento de los distintos tipos de crímenes registrados a lo largo del tiempo. En el Apéndice A: Estructura de "Philadelphia Crime Data" se puede observar la estructura del mismo. (Ver Crimenes Reportados.csv)
- **Datos históricos de temperatura:** Mediante un script personalizado en Python (Ver Descargar_Temperaturas_Historicas.py), se obtuvo del sitio web Weather Undeground[2] las temperaturas registradas en intervalos aproximados de horas a lo largo del tiempo en la ciudad. El archivo resultante se generó en formato CSV (Ver Temperaturas Historicas.csv) con la siguiente estructura:
    - Año
    - Mes
    - Día

[1] www.kaggle.com: En el sitio web Kaggle, existen un múltiples de datasets compartidos por los usuarios con el objetivo de poder realizar distintos tipos de análisis sobre los mismos.

[2] https://www.wunderground.com/history: Sitio que provee las temperaturas históricas de una ciudad por cada día y en intervalos aproximados de 1 hora, con la posibilidad de exportar los datos en formato CSV.

    - Hora (Formato: hh:mm a)
    - Temperatura
- **Barrios de la ciudad**: Mediante el repositorio publico GitHub, se obtuvo un archivo en formato GeoJSON[3] que contiene tanto nombres como información geográfica de los barrios de la ciudad de Filadelfia. (Ver Barrios de la ciudad de Filadelfia.geojson)

4.2. Preprocesamiento

Para poder realizar un preprocesamiento adecuado de todos los datos, mediante herramientas e investigación de distintas fuentes según sea el caso, se realizó un análisis exhaustivo para poder comprender la calidad de los datos. Por ejemplo: El dataset principal utilizado ("Philadelphia Crime Data") no contiene metadatos asociados a sus atributos, por lo que se realizó un análisis manual.

Como observación, se podrá ver que todos los posibles tratamientos de esta etapa están totalmente ligados ya que la decisión de cada uno estará sujeta de los otros y, además, aunque se agruparon los tratamientos según los posibles elementos de calidad de datos, será posible leer el desarrollo de manera adecuada. Todos los posteriores tratamientos de esta etapa fueron tratados en un mismo script junto con sus posteriores indicaciones a las secciones de este documento (Ver Preprocesamiento.R).

4.2.1. Limitación de línea de tiempo

En la sección 4.1, se estableció que los datos estarían limitados a un rango especifico de fechas. En cuanto a las fuentes de datos, el único que no cumple la condición es "Philadelphia Crime Data" debido a que en su obtención no permitía establecer los rangos de fechas, es por esto, que todas sus observaciones fueron analizadas y limitadas.

4.2.2. Inconsistencias

En este caso, se realizó un análisis exhaustivo observando los metadatos (Ver Apéndice A: Estructura de "Philadelphia Crime Data") de la fuente principal junto con los posibles valores que tomaba cada atributo.

4.2.2.1. Distritos policiales

En una primera observación, en los atributos Police_Districts y Dc_Dist, se observó que no había una correlación directa con la realidad, en cuanto a lo que representaban, dándose lo siguiente:

[3] GeoJSON: Estándar abierto diseñado para la representación de elementos geográficos mediante formato JSON.

- El atributo Police_Districts posee un total de 22 valores únicos, donde los mismos están enumerados secuencialmente, pero los distritos[4] históricos no siguen esta distribución.
- El atributo Dc_Dist se pensaba que no tenía correlación con una subdivisión directa de la policía ya que Police_Districts realmente lo indicaba según su nombre, pero existían valores tales como 77 que podrían indicar que este atributo realmente era el distrito policial. En una primera observación, y basándose en el punto anterior, se pensó en que existía una correlación entre un numero secuencial y el número de distrito de este atributo, pero, sin embargo, un numero secuencial podía pertenecer a múltiples valores de este atributo, por ende, el atributo Police_Districts no tenía significado alguno. Ahora partiendo de solo los valores de este atributo, se observó que había una relación entre los mismos y los distintos distritos históricos de la ciudad.

Por estas razones, tales atributos son considerados como inconsistentes. Como metodología para solucionar tal problema, simplemente se copiaron los datos del atributo Dc_Dist al atributo Police_Districts conservando los posibles datos históricos.

4.2.2.2. Calles

Como segunda observación, se advirtió que existían distintos formatos para la representación de las calles y su altura. Por ejemplo, para indicar el cruce entre dos calles se utilizaba tanto el carácter "/" como "&", y esto podía llevar a futuros errores de comprensión y alterar los objetivos claves.

Para poder solucionar estos problemas, se recurrió a la utilización de expresiones regulares y se trabajó en problemas como, por ejemplo:

- Distintos caracteres para las intersecciones de calles. Ejemplo: "/", "&", "and", "@ [Numero]".
- Distintas abreviaturas para la palabra avenida, como, por ejemplo: "Av.", "Ave", "Avenue", "Ave.".
- Distintas abreviaturas para la palabra Street, como, por ejemplo: "St", "Street".
- Existían caracteres basura como por ejemplo "´´", "@", "´´", etc.
- Múltiples espacios entre palabras.
- Entre otros

Debido a este procedimiento se redujo la cantidad de posibles de bloques de calles en un 28,17%. Si bien este resultado puede ser considerado como aceptable, se siguió

[4] A modo de ejemplo, en el enlace https://www.phillypolice.com/districts/ se obtienen los distritos policiales actuales, lo cual no implica que no haya habido otros en la historia de la ciudad.

observando que existían discrepancias entre palabras como, por ejemplo, diferir solo por un par de letras.

Para poder dar una solución final a este problema, se recurrió al concepto de Fuzzy Match (Unión borrosa), donde se trata de eliminar discrepancias mediante la maximización de un coeficiente de similitud obtenido de un algoritmo en específico, una cadena de caracteres con posibles discrepancias y un conjunto de cadenas de caracteres fiable. En este caso se utilizó:

- **Algoritmo:** De los 10 algoritmos de comparación que provee la herramienta Kettle, se seleccionó el algoritmo "Similitud de par de letras" ya que empíricamente se verifico que el mismo tenia resultados totalmente aceptables en estas circunstancias, donde puede haber ciertos errores debido a que la calle del crimen es inexistente.
- **Conjunto de cadenas fiables:** En este caso, y a consideración de los objetivos principales, solo es necesario un listado de calles de la ciudad de Filadelfia.

  Para este caso, se realizó un procedimiento distinto a los otros donde se obtuvieron datos externos, ya que se recurrió a la unión de tres fuentes distintas de datos debido a que las fuentes podrían estar incompletas[5] (Ver Inconsistencias - Calles\Pre-procesamiento y union.R). Para esto, las fuentes fueron:
  - **Open Street Map:** Mediante el sitio web www.overpass-api.de[6], se utilizó la interfaz de comandos del protocolo Overpass API para obtener una respuesta (Ver Inconsistencias - Calles\Respuesta - Calles Overpass API.xml) en formato XML que contiene las calles registradas en OpenStreetMap, utilizando la siguiente consulta:

    ```
    area[name="Philadelphia"];way(area)[highway][name]["ti-
    ger:name_type"];out;
    ```

    Luego, mediante un script personalizado en Python (Ver Inconsistencias - Calles\Parseador de respuesta Overpass API.py), se parseo la respuesta y se exporto a formato CSV.
  - **Geographic.org:** Mediante un script personalizado en Python (Ver Inconsistencias - Calles\Descargar calles de geographic.org.py), se descargó y exporto del sitio web www.geographic.org, un listado de calles en formato CSV.
  - **OpenDataPhily:** Mediante el repositorio abierto de la ciudad, se obtuvo un archivo en formato Shapefile con la mayoría de las calles existentes de la ciudad. (Ver Inconsistencias - Calles\CompleteStreetsTypesStndrds.shp).

[5] También se observó que ciertas calles de los posibles valores del dataset principal no existían en las fuentes utilizadas.

[6] No se utilizó el mapa interactivo de OpenStreetMap directamente ya que posee un límite de respuesta.

Cabe aclarar que las fuentes sufrieron pre procesamientos que implicaban normalizar ciertas palabras y así obtener una unión correcta.

Si bien la primera parte del proceso (Uso de expresiones regulares) fue realizada sobre el lenguaje R, el proceso de Fuzzy Match fue incorporado dentro del proceso de transformación ya que era necesario realizar una generalización en las calles, y en la inversa de este proceso podrían perderse datos, además de repetir un proceso complejo innecesariamente para los objetivos del trabajo.

4.2.3. Ruido

De todos los atributos que componen a las fuentes de datos, aquellos que pueden resultar con ruido son: Temperatura (De las temperaturas históricas), Lat, Long. Pero estos dos últimos, no necesariamente indicaban ruido, sino que también se verá que los archivos GeoJSON también pueden tener ruido en sus límites geográficos.

4.2.3.1. Temperaturas

Como se puede observar en la Ilustración 4.1, en las temperaturas registradas existen valores atípicos, específicamente con valores -9999 (Tabla 4.1), que se asocian a temperaturas sin registro.

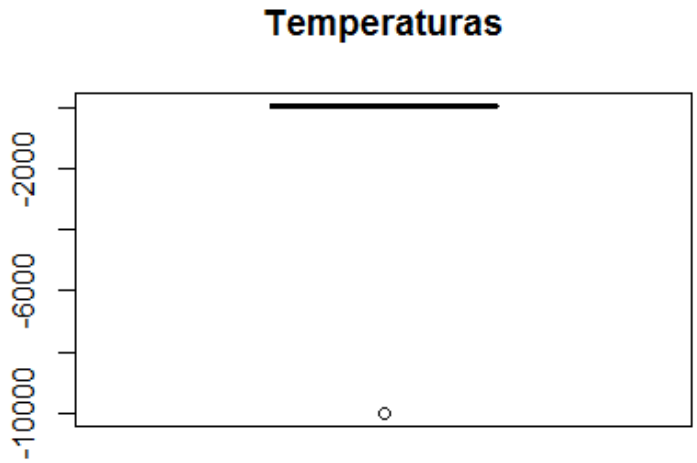


Ilustración 4.1: Diagrama de caja de las temperaturas históricas

Atributo	Mínimo	Promedio	Máximo
Temperatura	-9999	-5.726	38.9

Tabla 4.1: Medidas de tendencia de las temperaturas históricas

Como técnica para el tratamiento de datos faltantes (Ver Ruido - Temperaturas\Tratamiento de ruido.R) se utilizó la imputación por mediana según el día (Han, Kamber, & Pei, 2011), debido a que:

- Las temperaturas varían por muchos factores a lo largo de los años, y el día a día.
- Cada día puede tener distintas distribuciones (Ver Ilustración 4.2).

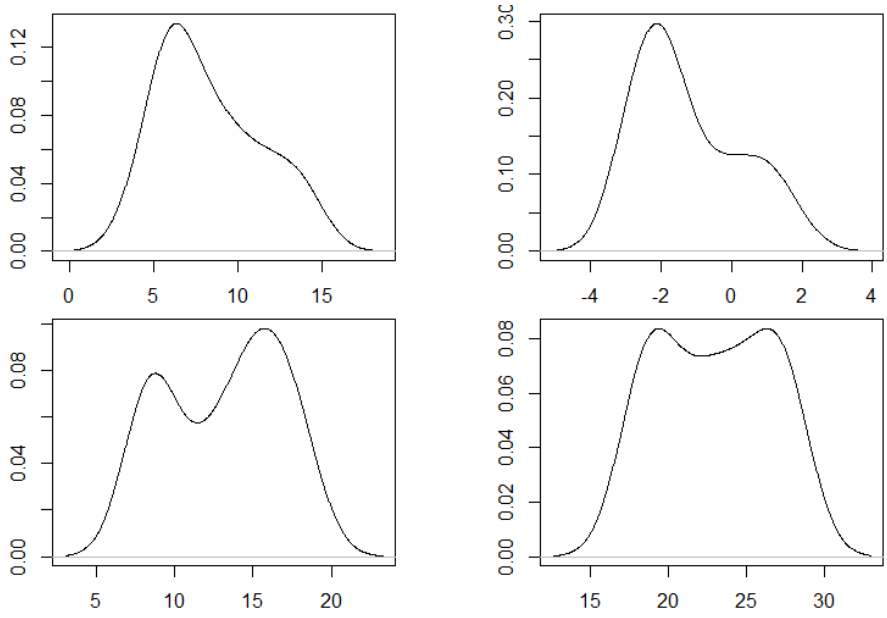


Ilustración 4.2: Ejemplos de distribuciones de temperaturas diarias

Ver código Ruido - Temperaturas\Ejemplos distribuciones Temperaturas.R

4.2.3.2. Archivos GeoJSON

Ya realizado el análisis de los atributos, se planteó una metodología necesaria para el ruido existente en los archivos GeoJSON utilizados en todo el trabajo. Estos archivos fueron utilizados en ciertos tratamientos para realizar la geolocalización inversa, pero no siempre se obtenían áreas asociadas a las coordenadas. Esto no ocurre debido a posibles ruidos en las coordenadas sino debido a que los archivos GeoJSON tienen pequeñas diferencias con la realidad propia, y entre otros archivos de la misma ciudad.

Por ejemplo, se observó en varios casos que los límites geográficos de los archivos se encontraban sobre las mitades de las calles limítrofes, pero existían coordenadas sobre la otra mitad de las mismas (Ver Ilustración 4.3 junto con su correspondiente código Ruido - GeoJSON\Ejemplo diferencias en coordenadas.html).

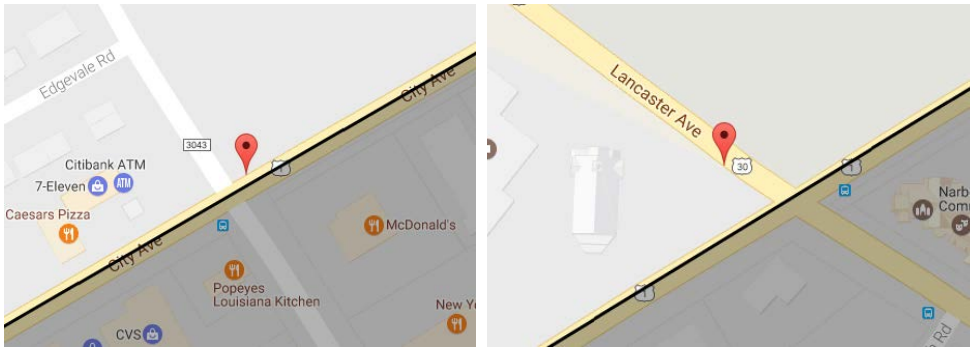


Ilustración 4.3: Ejemplos de las diferencias entre las coordenadas y el archivo GeoJSON

Entonces, para poder reducir las diferencias, se adiciono a la técnica de geolocalización el hecho de asignar aquellos polígonos más cercanos (Que por ejemplo pueden representar barrios, o distritos) a las coordenadas que no han tenido áreas asociadas, eliminándose el "ruido" de los atributos Lat y Lon.

4.2.4. Datos faltantes

Por un lado, en la Tabla 4.2 se muestra la cantidad de datos faltantes por cada atributo que compone al dataset principal, sobre los cuales se aplicaron los tratamientos adecuados. Cabe aclarar, que se consideró como dato faltante a aquellos campos con valores nulos, o simplemente espacios en blanco.

Atributo	Datos faltantes
Dc_Dist	0
Psa	0
Dispatch_Date_Time	0
Dispatch_Date	0
Dispatch_Time	0
Hour	0
Dc_Key	0
Location_Block	0
UCR_General	437
Text_General_Code	437
Police_Districts	0
Month	0
Lon	17092
Lat	17092

Tabla 4.2: Datos faltantes de "Philadelphia Crime Data"

Y por otro, en la Tabla 4.3 se muestran los datos faltantes para el único atributo no generado de los datos históricos de temperaturas.

Atributo	Datos faltantes
Temperatura	0

Tabla 4.3: Datos faltantes del histórico de temperaturas

Utilizando solamente las tuplas completas del dataset hasta el momento, solo se removería un 0,91% del total de tuplas.

4.2.4.1. Clasificación UCR

En los atributos UCR_General, Text_General_Code, donde ambos atributos representan lo mismo, sus datos faltantes coinciden en todos sus casos con el otro, donde el primero posee valores nulos y el segundo espacios en blanco.

En este caso, el primer atributo fue reemplazado por el valor 2600, y el segundo por "All Other Offenses", el cual es un tipo de crimen utilizado como una clasificación genérica.

4.2.4.2. Latitud y longitud

El tratamiento que se le dio al porcentaje de datos faltantes fue ignorar por completo a las tuplas incompletas debido a que representaban un porcentaje muy bajo para alterar todos los análisis posteriores (Solo existe un 0,77% de datos faltantes). Además, para poder realizar la mayoría de los análisis, eran totalmente necesarios ya que permiten generar en la etapa de transformación los barrios de la ciudad.

4.3. Transformación

Como ya se indicó, en esta etapa se realizaron varias alteraciones a la estructura de los datos y a sus valores con el objetivo de obtener datos y estructuras fiables para así explorarlos.

4.3.1. Estructura lógica

En la Ilustración 4.4, se observa un diagrama lógico de la base de datos que almacenara todos los datos necesarios para los posteriores análisis. Su utilización fue debido a que la cantidad de crímenes registrados es muy alta (Poco más de 2 millones) y su manipulación por parte de una simple estructura de datos seria ineficiente tanto para los análisis actuales como futuros.

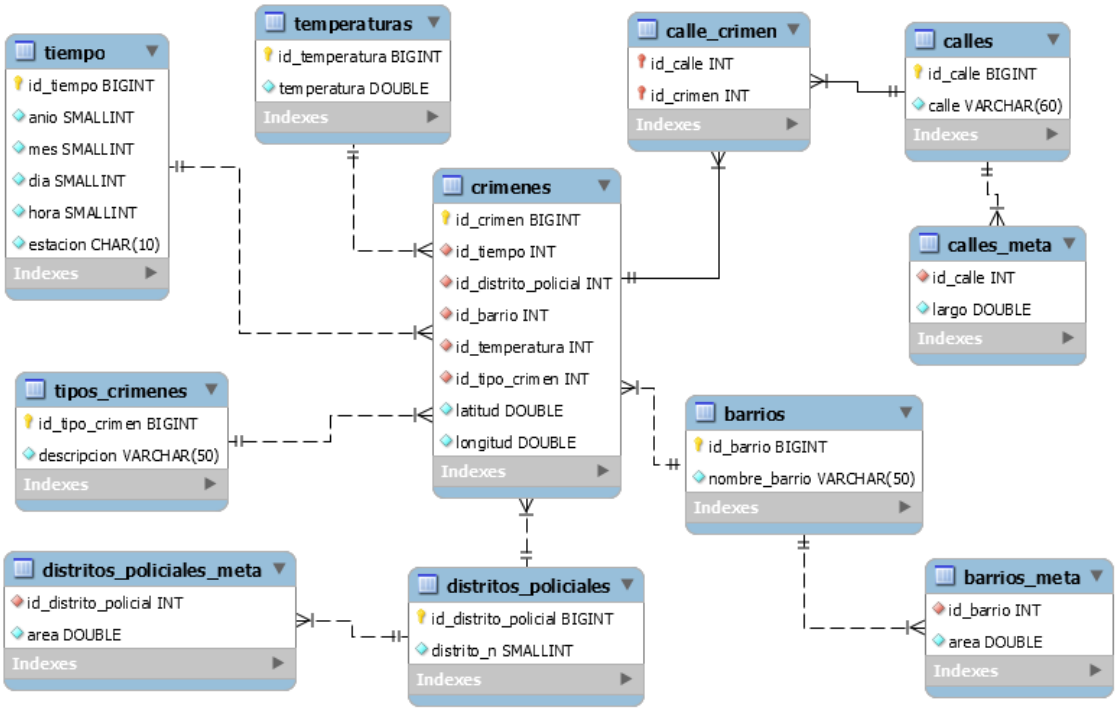


Ilustración 4.4: Diagrama lógico de la base de datos principal

Las tablas con el sufijo "_meta" hacen referencia a datos que fueron utilizados para poder satisfacer los objetivos con mayor precisión. En caso de más información de la Ilustración 4.4 ver el archivo tabla.sql.

Ya realizada la estructura de las tablas finales, se procedió a crear una base de datos denominada "TP-Final-BDM" en PostgreSQL, con la posterior ejecución del script SQL anteriormente nombrado.

4.3.2. Transformación y carga

4.3.2.1. Generación de Barrios

En base a las coordenadas y mediante la técnica de geolocalización inversa, donde la cual consiste que a partir de un par de coordenadas geográficas obtener un dirección o elemento representante del área asociada a las coordenadas, se adicionaron los barrios (Ver Ilustración 4.5) de la ciudad de Filadelfia a los crímenes reportados. La técnica consistió en base al archivo GeoJSON de los barrios, y se procedió a realizar lo siguiente:

- Con el sitio web www.mapshaper.org se cargó y exporto al mismo en formato ShapeFile para que sea válido en R.
- Con las coordenadas completas y el anterior archivo, se realizó una geolocalización inversa obteniéndose por cada par de coordenadas un barrio.
- Se adicionaron los nombres de los barrios en un nuevo atributo llamado Neighborhood.

- Se asignaron a los pares de coordenadas sin resultados en la geolocalización inversa, los barrios más cercanos a las mismas y se volcaron en el atributo Neighborhood (Ver la sección 4.2.3.2).

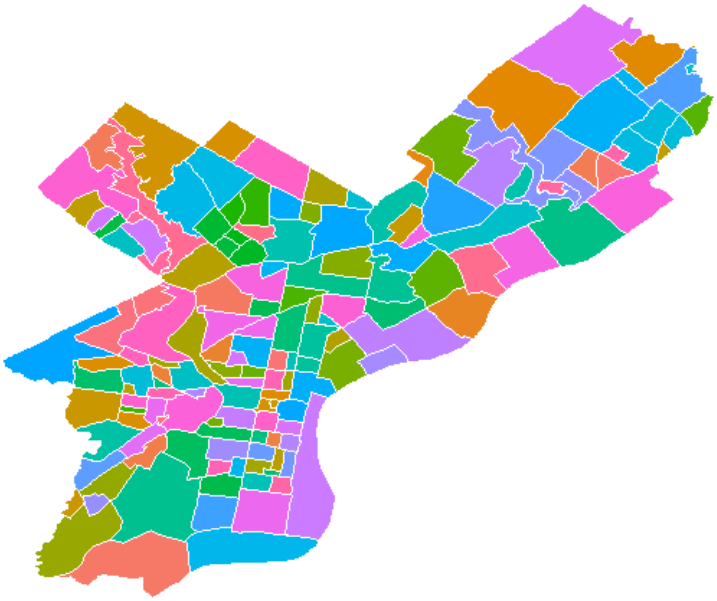


Ilustración 4.5: Barrios de la ciudad

Ver la sección 4.3.2.1 del archivo Script de transformaciones.r.

4.3.2.2. Unión de crímenes con temperaturas

En la sección 4.1 se vio que se necesitarían las temperaturas históricas de la ciudad, y fueron obtenidas en intervalos indefinidos de horas o minutos. Si bien en los minutos que no figuran, se podría considerar inconsistencias en los rangos o datos faltantes, se optó por utilizar una técnica particular para la unión de los datos (Temperatura / Reporte Crimen). La técnica utilizada consistió en asignar a cada reporte de crimen original, la lectura de temperatura más temprana al mismo según la fecha y hora de ambos, sin considerar las temperaturas con fechas y horas mayores al reporte.

Ver la sección 4.3.2.2 del archivo Script de transformaciones.r.

4.3.2.3. Carga de calles

Como se había dicho anteriormente, para resolver inconsistencias presentes en las calles es necesario utilizar un listado valido de las mismas. Es por esto que en la Ilustración 4.6 se muestra la carga de calles obtenidas en la sección 4.2.2.2.

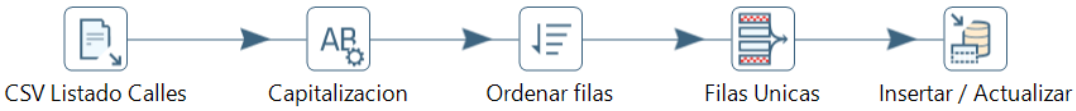


Ilustración 4.6: Carga de calles

Ver el archivo Carga - Calles.ktr.

4.3.2.4. Normalización de datos

En este paso, se realizó la normalización de los datos principales de las fuentes, para así luego poder unificarlos en la base de datos.

El proceso general planteado para cada atributo, básicamente consistió en:

- Obtener solo el/los atributo/s necesarios para la tabla
- Realizar las transformaciones necesarias
- Ordenar y obtener los valores únicos del/los atributo/s
- Insertar o actualizar los valores en la tabla correspondiente

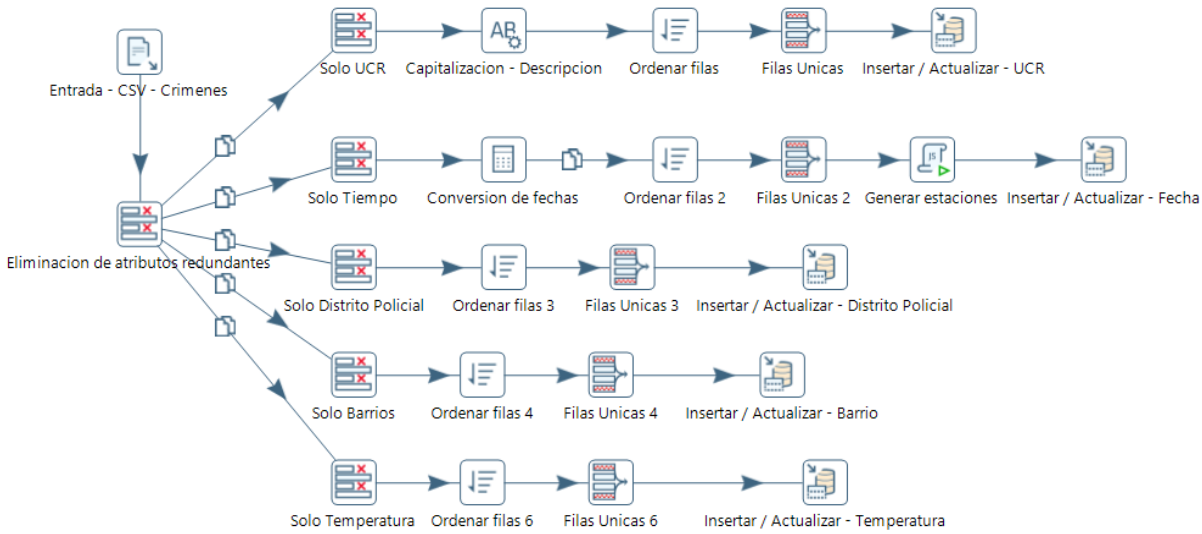


Ilustración 4.7: Normalización de crímenes

Ver el archivo [Normalizacion de atributos - Crimenes.ktr.]

4.3.2.5. Generación de metadatos

Como se vio anteriormente se adicionaron tres tablas con el sufijo "_meta" a la estructura principal, pero el sentido de las mismas es debido a una normalización de la tasa de crímenes necesaria en la etapa de análisis (Ver la sección 4.4).

Tales metadatos[7] consistieron en:

- **Área de los barrios:** En base al archivo obtenido en la sección 4.3.2.1, se calculó el área de cada barrio en kilómetros cuadrados.
- **Área de los distritos policiales:** Dado de que no se dispone de un archivo geográfico con los límites de los mismos, se calculó la envoltura convexa (*convex hull*) de las coordenadas asociadas a cada distrito y se le calculo el área en kilómetros cuadrados, la cual no coincide completamente con la real ya que es estimada. Previamente al cálculo de las áreas, además se aplicó el

[7] Solo se consideró el área de cada subdivisión geográfica debido a que no se disponía de datos fiables que indiquen la población en cada una de estas.

algoritmo de clustering DBSCAN[8] (Ester, Kriegel, Sander, & Xu, 1996) para eliminar coordenadas atípicas a cada distrito policial, ya que las mismas alteraban completamente el resultado. En la Ilustración 4.8 se muestra un caso particular del distrito 1, donde a la izquierda se muestran las coordenadas sin procesar, y a la derecha, aquellas que no fueron detectadas como ruido por el algoritmo.

- **Longitud de las calles:** En base al archivo que contiene la mayor parte de las calles de la ciudad (Utilizado previamente en la sección 4.2.2.2), se obtuvo la longitud de las calles en metros, y para aquellas que no disponía de tal medida, se imputo por la mediana de las demás.

El archivo Generacion de metadatos.R contiene el código para la generación de los mismos.

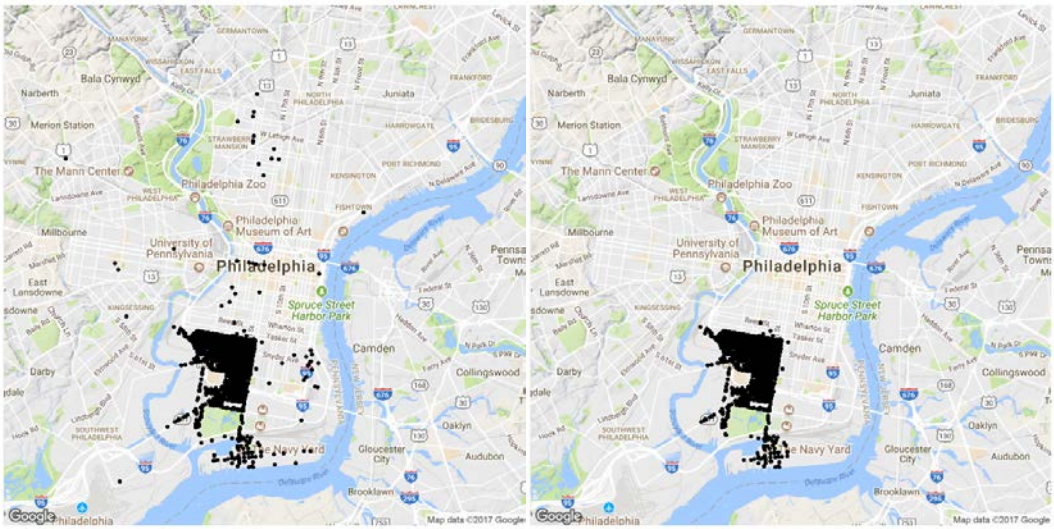


Ilustración 4.8: Eliminación temporal de coordenadas atípicas por distrito policial

Ya generado los metadatos, se procedió a la carga de los mismos en la base de datos principal (Ver la siguiente ilustración).

[8] Como máximo se aplicó el algoritmo a 65000 coordenadas seleccionadas aleatoriamente del distrito por cuestiones de memoria RAM.

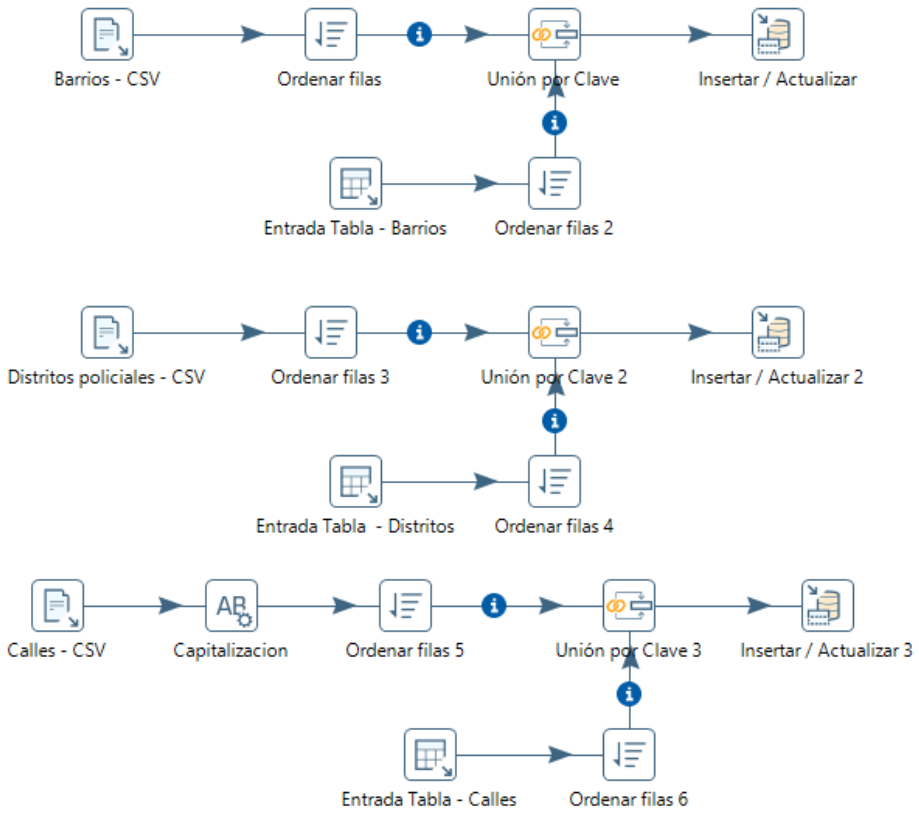


Ilustración 4.9: Carga de metadatos

Ver el archivo Carga - Metadatos.ktr.

4.3.2.6. Unión y carga

Ya realizado las normalizaciones, fue necesario realizar la última etapa de la transformación. La misma consistió en volver unir los datos ya precargados de la etapa anterior debiendo utilizar los mismos pasos de transformación de la misma para obtener una unión precisa.

A continuación, se procedió a la normalización de los atributos para así poder generar relaciones 1 a muchos. El proceso general planteado en la Ilustración 4.10 para cada unión con una tabla es:

- Realizar las mismas transformaciones sobre el/los atributo/s de la etapa de normalización de datos.
- Ordenar los valores por la clave principal del atributo (Nombre de barrio, numero de distrito, etc.).
- Obtener los datos de la tabla de manera ordenada por la misma clave que el paso anterior.
- Realizar una unión por la clave entre los dos pasos anteriores.
- Eliminar las claves utilizadas para la unión, quedándose solo con próximos atributos a utilizar y la clave primaria de la tabla.

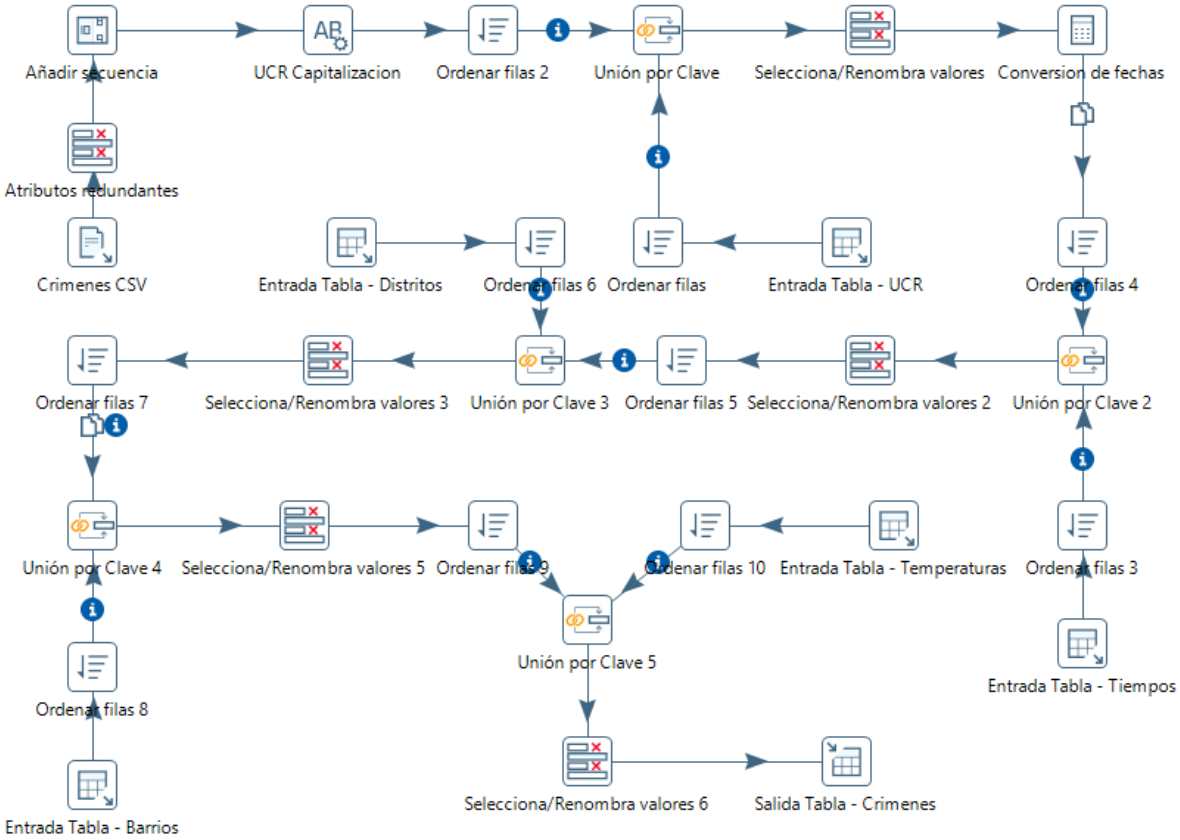


Ilustración 4.10: Unificación y carga de crímenes

El archivo mostrado en la Ilustración 4.10 se corresponde a Union y carga - Crimenes.ktr.

Si bien la tabla principal "crimenes" posee una clave primaria, y es de tipo autoincremental, en la Ilustración 4.10 se observa que el tercer paso se añade una secuencia numérica. La misma, es debido a que para poder realizar las uniones de la relación muchos a muchos con la tabla "Calles" es necesario establecer un proceso inverso que permita obtener un identificador según un crimen particular para así relacionarlo con sus posibles calles.

En una segunda parte, se procedió a realizar el proceso de unión entre crimen/calles mediante una previa generalización de las calles ya que contienen alturas e intersecciones entre calles. El proceso se dividió en dos partes debido a que realizar un "Fuzzy match" entre dos millones de calles conllevaba aproximadamente entre 6-8 horas (Estimado a partir de una previa ejecución parcial).

A continuación se describe el proceso de generalización de calles utilizando tanto en la Ilustración 4.11 e Ilustración 4.12. Para esto se procedió a realizar lo siguiente:

- Añadir una secuencia numérica para poder establecer la relación con la tabla de crímenes anteriormente cargada.
- Si la calle contiene una barra indicadora de intersección:
    - Dividir la fila en tantas veces aparezca una barra de intersección.
    - Eliminar posibles espacios al principio o al final.
    - Quitar alturas innecesarias.

    - Filtrar las filas para no valores nulos o solo espacios.
    - Renombrar los atributos para una posterior unión entre flujos.
- Si la calle no contiene una barra indicadora de intersección:
    - Quitar alturas y bloques.
    - Renombrar los atributos para una posterior unión entre flujos.
- Unir los flujos de los dos pasos anteriores.
- Agregar la palabra St a las letras únicas para un mejor macheo.

La primera transformación para la unión entre calles y crímenes, consto de los siguientes pasos:

- Obtención de calles con alturas únicas para eliminar redundancias en el proceso.
- Realizar el proceso de generalización anteriormente dicho.
- Obtener solo las calles únicas.
- Realizar un "Fuzzy match" entre las calles únicas, y las calles cargadas en la base de datos (Las cuales son fiables).
- Ignorar aquellas calles sin resultados, las cuales son 15 de 16244.
- Guardar en un archivo temporal la relación entre Calle (Posiblemente con errores) y la clave primaria de la calle asignada en el "Fuzzy match".

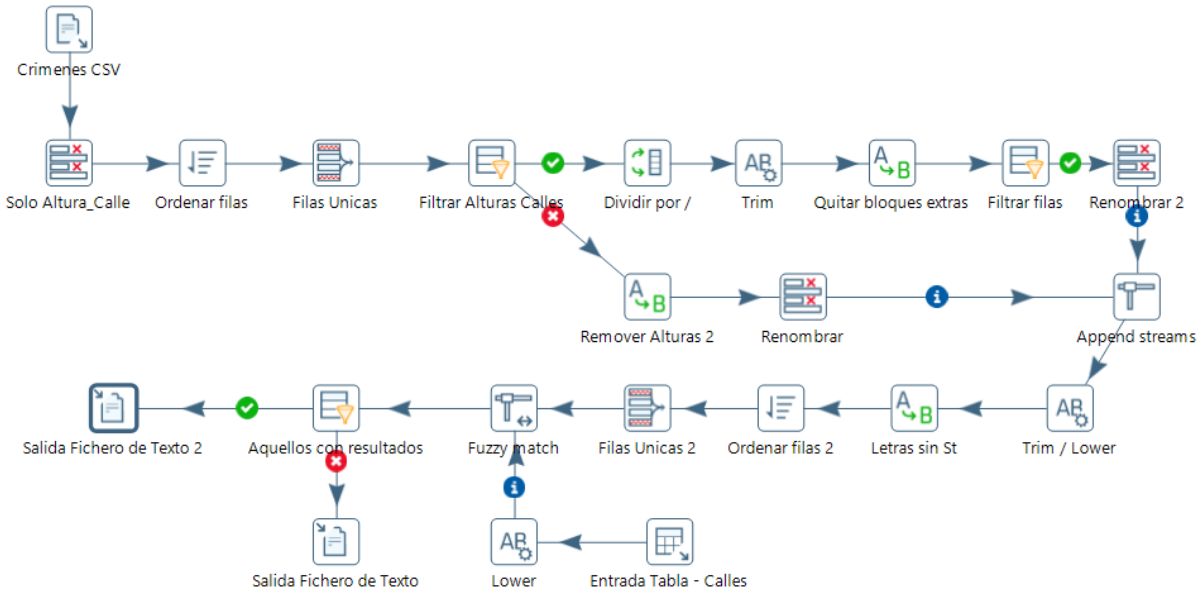


Ilustración 4.11: Generación de mapping entre calle y calles fiables

Ver el archivo Generacion de mapping - Calles.ktr.

Partiendo del archivo temporal generado de la transformación anterior, se realizó los siguientes pasos:

- Añadir a cada crimen reportado un identificador secuencial numérico para luego relacionar una calle con sus posibles crímenes.
- Realizar el proceso de generalización anteriormente dicho.
- Ordenar las filas para poder realizar una unión por calle.

- Realizar una unión a partir de la calle entre los flujos para obtener el id_calle correspondiente.
- Obtener los pares únicos de id_crimen, id_calle, ya que se podría dar el caso de que el "Fuzzy match" hizo una misma asignación a las calles para un crimen en particular.
- Cargar los pares en la base de datos.

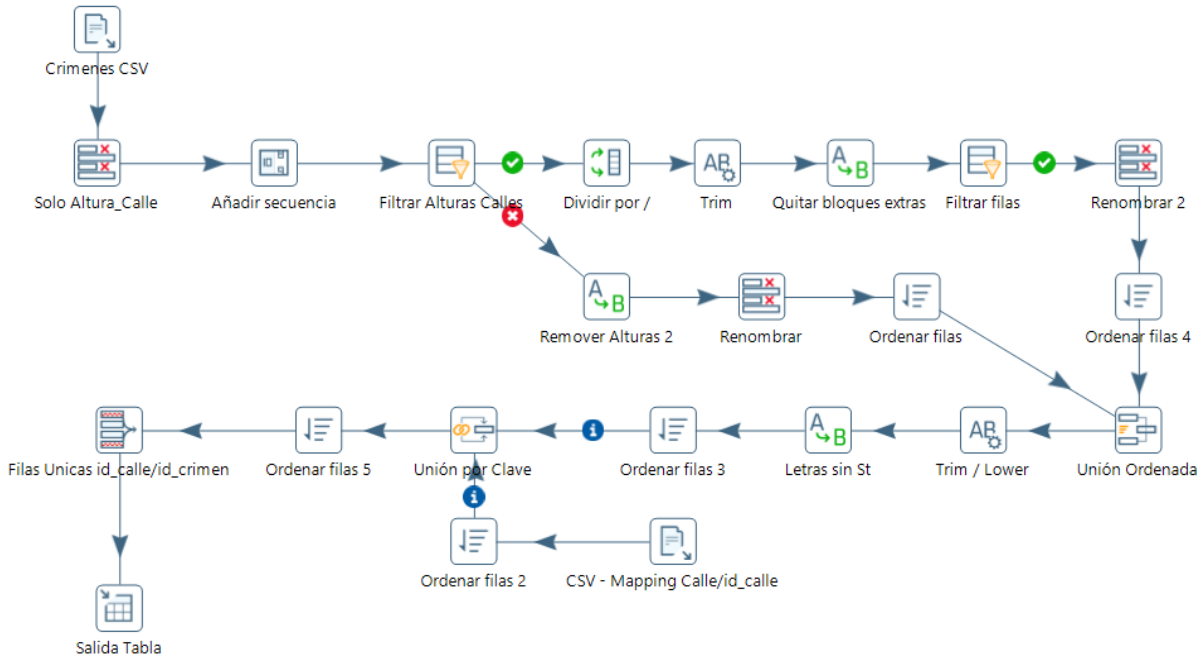


Ilustración 4.12: Unión y carga - Crimen/Calles

Ver el archivo Union y carga - Crimenes-Calles.ktr.

4.3.2.7. Generación y ejecución del trabajo

Ya definidas todas las transformaciones, se procedió a realizar un único trabajo para poder ejecutar todas las transformaciones en el orden correcto, siendo el resultado la siguiente ilustración:

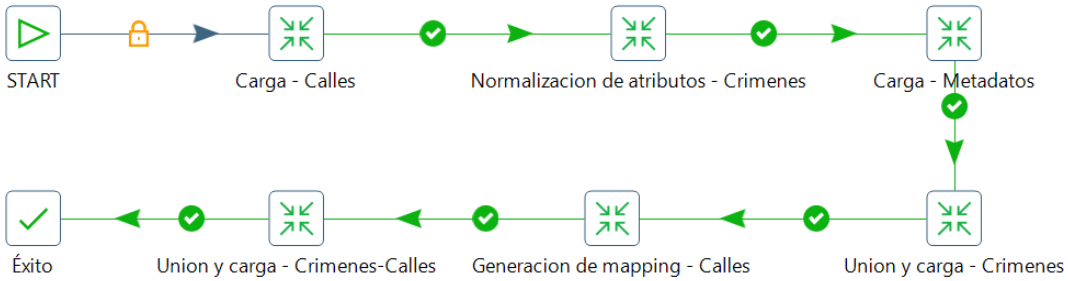


Ilustración 4.13: Trabajo principal

Ver el archivo Trabajo principal.kjb.

4.4. Análisis de datos

4.4.1. Ranking por cantidad de crímenes

La metodología a utilizar en este objetivo consistió en:

- Agrupar por el atributo correspondiente, y ordenar los valores decrecientemente según la tasa de crímenes normalizada por:
  - Barrios y distritos policiales: El área en kilómetros cuadrados.
  - Calles: La longitud en metros.

  Limitar el resultado solo a los 25 valores con mayor tasa de crímenes.
- Realizar un gráfico de barras horizontal, donde por cada posible valor del atributo, mostrar la tasa normalizada. Además, agregar el promedio y la mediana de las tasas normalizadas para el total de grupos que existan.

Ver el archivo Objetivo 1 - Ranking por cantidad de crimenes

4.4.2. Ranking por tipo de crímenes

Ver el archivo Objetivo 2 - Ranking por tipo de crimen

4.4.2.1. Ciudad de Filadelfia

Para este sub-objetivo se debió realizar lo siguiente:

- Agrupar el total de crímenes reportados según el tipo de crimen, y ordenar los grupos según la cantidad de crímenes de manera decreciente.
- Realizar un gráfico de barras horizontal, mostrando la cantidad de crímenes por tipo mismo junto con el porcentaje representativo en el total de la muestra. Además, agregar el promedio y la mediana de las cantidades por tipo de crimen.

4.4.2.2. Barrios

Para este sub-objetivo se debió realizar lo siguiente:

- Según los barrios obtenidos en el ranking de la sección 4.4.1, realizar una especialización por tipo de crimen para así obtener otro subranking, que indique los 3 tipos de crímenes más frecuentes por barrio, normalizando la cantidad por el área de los mismos.
- Realizar un gráfico de barras, donde para cada posible barrio, mostrar un ranking de los tipos de crímenes.

Adicionalmente, se realizó un mapa de los barrios de la ciudad, donde se muestra mediante un color especifico el tipo de crimen más frecuente en los mismos.

4.4.2.3. Distritos policiales

Para este sub-objetivo se debió realizar lo siguiente:

- Por cada distrito policial, realizar una especialización por tipo de crimen para así obtener otro subranking, que indicara los 3 tipos de crímenes más frecuentes por distrito, normalizando la cantidad por el área de los mismos.
- Realizar un gráfico de barras, donde para cada posible distrito, mostrar un ranking de máximo 3 tipos de crímenes.

Adicionalmente, se realizó un mapa de los crímenes reportados según sus coordenadas específicas, donde cada color representa el tipo de crimen más frecuente por distrito policial. Cabe aclarar que, al ser datos históricos, pueden superponerse los colores.

4.4.3. Zonas de concentración

Se debió realizar un mapa de calor que represente las zonas de concentración en base a las coordenadas geográficas de cada crimen reportado, donde a mayor concentración, más cercano al color rojo se representara la zona.[9]

Ver el archivo Objetivo 3 - Zonas de concentración

4.4.4. Divisiones geográficas con mayor y menor incremento

Tanto para los barrios como distritos policiales se realizaron los siguientes pasos:

- Obtener una agrupación por división geográfica, con la cantidad de crímenes por año en tal división.
- Calcular los porcentajes de incremento año a año según la división, siendo el mismo calculado a partir de la división de la cantidad de crímenes por la cantidad inmediata anterior.
- Calcular los promedios[10] de los porcentajes de incrementos por cada división, y obtener aquella con el mínimo promedio, la cual representa la división con mayor reducción en la tasa de crímenes a lo largo de los años, y aquella con máximo promedio es la que tuvo mayor incremento.
- Realizar para ambas divisiones un gráfico de líneas, con las cantidades de crímenes por año junto con una línea de tendencia[11] y los porcentajes de incrementos correspondientes a cada año.

Ver el archivo Objetivo 4 - Divisiones geográficas con mayor y menor incremento

---

[9] Por cuestiones de limitación de hardware solo se procesó un total de 1.500.000 de reportes al azar ya que el proceso de generación de un mapa de calor con el total de reportes requería mayor capacidad computacional.

[10] En el caso de los barrios, se vio que la mediana daba mejores resultados por las distribuciones por año.

[11] Se agregó tanto la correlación como el coeficiente de determinación para comprender estadísticamente el comportamiento lineal.

4.4.5. Análisis por intervalos de tiempos

Ver el archivo Objetivo 5 - Analisis por tiempo

Para los rankings por año, estación y mes, se realizaron los siguientes pasos:

- Agrupar los crímenes por intervalo de tiempo, y a su vez, por tipo de crimen.
- Realizar un gráfico de barras vertical, donde por cada posible intervalo agrupado, se mostrará una barra apilada por tipo de crimen y la cantidad de repeticiones del mismo en el intervalo de tiempo. Además, agregar el promedio y la mediana de los totales por intervalos de tiempo.

4.4.5.1. Días y horas

El grafico a realizar consistió en:

- Obtener la cantidad de crímenes por cada valor posible del intervalo de tiempo.
- Realizar un gráfico de líneas, mostrando la cantidad de crímenes en función del intervalo de tiempo. Además, agregar la media y la mediana de los totales por grupo.

Adicionalmente, para el análisis por hora, se realizó un segundo grafico que consiste en:

- Agrupar los crímenes por hora, y luego por tipo de crimen obteniendo la cantidad de crímenes por grupo.
- Realizar un gráfico de líneas, donde se muestre en el eje de las abscisas las 24 horas diarias, y con múltiples líneas que representen las cantidades en los distintos tipos de crímenes.

4.4.6. Correlación Crimen / Temperaturas

4.4.6.1. Análisis de temperaturas

Como método de comparación inicial se calculó la cantidad de crímenes en función de las temperaturas (Ver Ilustración 4.14) según los reportes. A partir de esto, se observó primero que existen ciertas temperaturas con cantidad de crímenes atípicas, y segundo, que se sigue una distribución normal.

Debido a estas dos razones, se determinó dos vías para poder calcular una correlación lineal por cada tipo de crimen:

- Para solucionar las cantidades atípicas se crearon 50[12] rangos discretos de temperaturas según un binning por igual anchura, y a su vez, para calcular la correlación, se calculó la temperatura media de este.
- Dado que existe una distribución normal, existe un comportamiento creciente hasta una cierta temperatura, y luego de la misma, un comportamiento decreciente. Por esto, mediante múltiples pruebas, y según la Ilustración 4.14, se determinó que una temperatura critica para poder realizar una regresión lineal es 26°C.

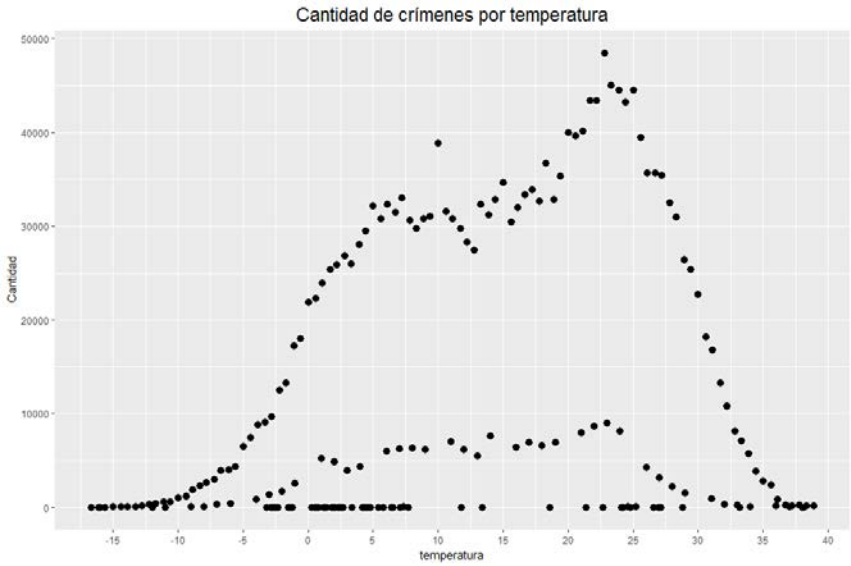


Ilustración 4.14: Cantidad de crímenes en función de la temperatura

4.4.6.2. Correlación

Ya definida la temperatura crítica y la cantidad de rangos, se procedió por cada tipo de crimen a:

- Dividir todas las muestras en 50 rangos según la temperatura media de cada uno de estos y contar la cantidad de crímenes.
- Graficar mediante puntos y líneas la cantidad de crímenes en función de la temperatura media del rango.
- Según las observaciones que no superan a la temperatura critica:
    - Obtener una recta mediante una regresión lineal.
    - Calcular y mostrar en el grafico la correlación y el coeficiente de determinación.

Adicionalmente, se ordenaron los tipos de crímenes de mayor a menor correlación. Ver el archivo Objetivo 6 - Correlacion entre tasas de crimenes y temperaturas.R.

[12] Un compromiso entre perdida de precisión y reducción del ruido.

4.4.7. Zonas propensas a la ocurrencia de crímenes

El objetivo pretende analizar las tasas de crímenes por cada tipo de crimen, y en cada barrio de la ciudad según los días anteriores a un reporte en distintos lapsos de tiempo. Por esto, en una primera parte se realizaron los siguientes pasos:

- Realizar un producto cartesiano entre los posibles barrios, tipos de crímenes y el rango de días desde 01/01/2006 hasta 31/12/2016.
- Obtener la cantidad de crímenes por día de cada tipo, en cada barrio de la ciudad.
- Realizar una unión del tipo *LEFT* entre los dos resultados anteriores, indicando por cada fila si se trata de un crimen real, ya que se obtienen días que no han ocurrido crímenes.
- Calcular la cantidad de crímenes que han ocurrido en los lapsos de tiempo de 5, 15 y 30 días por cada tipo de crimen y en cada barrio, mediante el uso de ventanas deslizantes. A su vez, normalizar las cantidades por el área de cada barrio, para obtener la cantidad de crímenes por kilómetro cuadrado.
- Eliminar aquellas filas que no pertenezcan a un día valido según los reportes de crímenes de la ciudad.
- Exportar el resultado a un archivo en formato CSV.

El código que realiza los pasos anteriores es Queries para generar datos.sql.

Luego, la segunda parte es agrupar los reportes según las tasas en los lapsos de tiempo. Entonces, mediante k-Means[13] se procedió a:

- Cargar el resultado en formato CSV.
- Transformar las variables categóricas "id_barrio" e "id_tipo_crimen" en variables de tipo "dummy" para una mejor agrupación.
- Normalizar las variables numéricas (Fechas y ventanas) mediante la técnica z-score.
- Ejecutar k-Means con k=[2; 7] y calcular el coeficiente de silueta[14] teniendo en cuenta los dos conjunto de datos siguientes:
    - Considerar las fechas (Año, mes y día)
    - No considerar las fechas[15]

[13] Por cuestiones de memoria RAM se utilizó una versión optimizada de k-Means denominada Mini-Batch k-Means (Sculley, 2010) que según Béjar (2013) las diferencias entre ambos algoritmos para k menor a 10 es del 2%.

[14] Solo se consideró 12000 observaciones al azar para calcular tal coeficiente por cuestiones de memoria RAM.

[15] Se determinó que no tenerlas en cuenta, no tenía una influencia lo suficientemente considerable en la agrupación final, sino que las diferencias en su mayoría estaban dadas por las inicializaciones y cálculos aleatorios en las ejecuciones del algoritmo.

- Tomar el agrupamiento final como aquel con mayor coeficiente. En este caso, el mejor resultado estuvo dado por un k igual a 2 sin considerar las fechas (Ver la Ilustración 4.15).

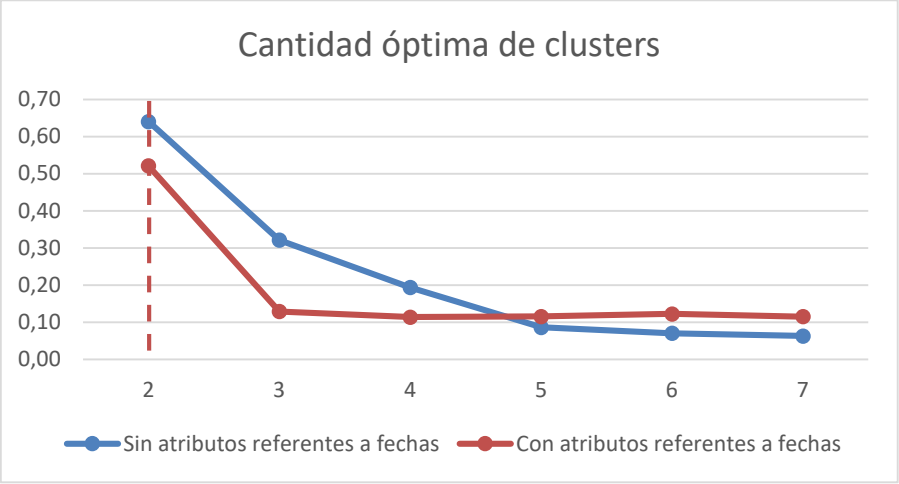


Ilustración 4.15: Cantidad óptima de clusters

El código de la clusterización se encuentra en el archivo Mini-Batch k-Means.py.

Adicionalmente, en la Ilustración 4.16 se muestra la cantidad de elementos por cluster.

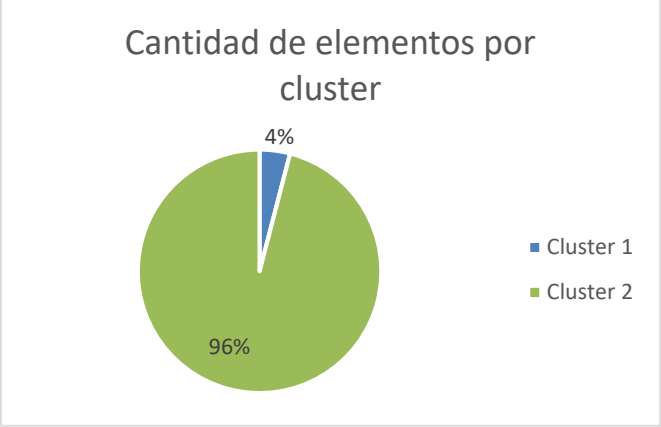


Ilustración 4.16: Cantidad de elementos por clusters

En cuanto a las características de los clusters, los análisis planteados fueron:

- Realizar y analizar la distribución de días en 5 intervalos discretos en un mapa por cada barrio de la ciudad.
- Analizar los tipos de crímenes mediante los pasos utilizados en la sección 4.4.2.1.
- Realizar y analizar un histograma por cada lapso de tiempo para determinar la distribución de las tasas de crímenes en cada uno de estos.
- Realizar y analizar los promedios de las tasas mediante un mapa por barrio para el lapso de 15 días. De igual manera, se determinó que solo era necesario

analizar para un único lapso de tiempo ya que, a mayor cantidad de días por lapso de tiempo, mayor será el promedio en cada barrio.

- Graficar una serie de tiempo a partir de los promedios de tasas en un lapso de 15 días por año y mes, junto con la línea de tendencia de la misma.

5. Resultados

En esta sección, se verá los resultados de la ejecución de los algoritmos vistos la sección 4.4, para satisfacer cada objetivo junto con la evaluación e interpretación de los mismos.

5.1. Ranking por cantidad de crímenes

5.1.1. Barrios

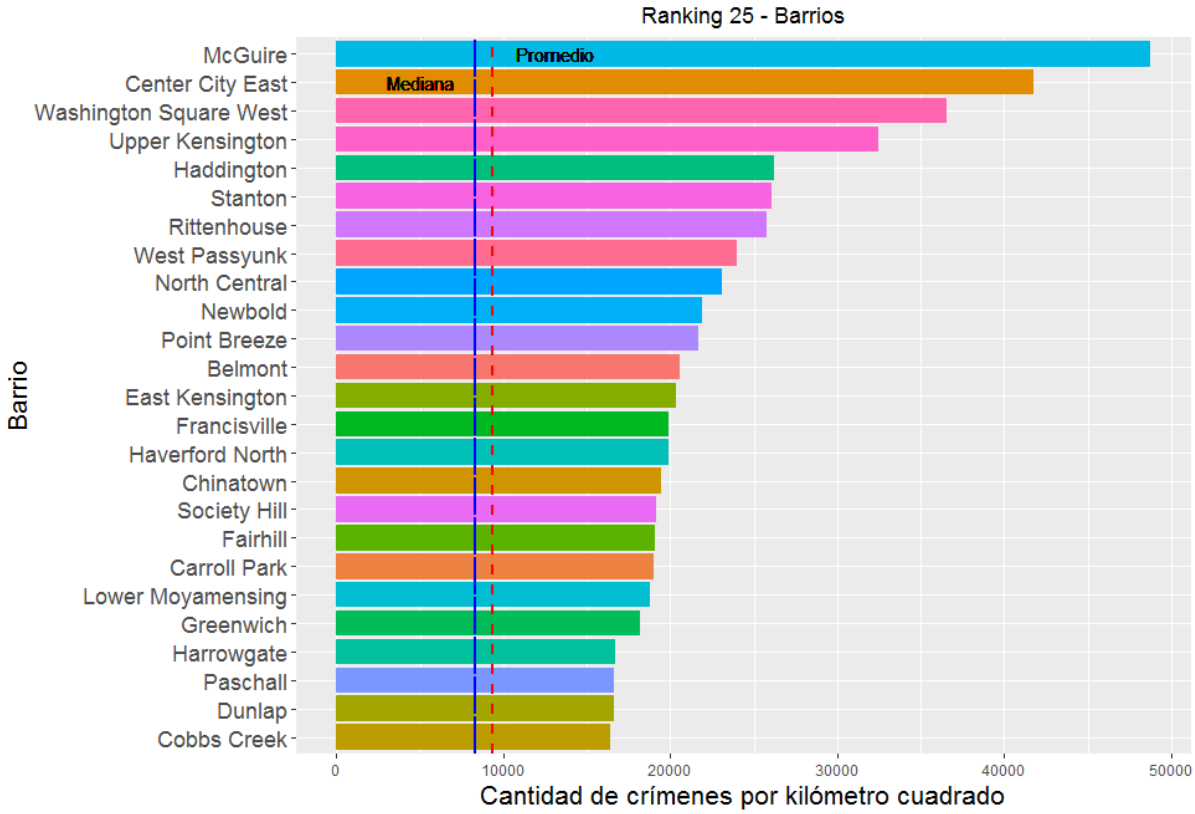


Ilustración 5.1: Ranking 25 - Cantidad de crímenes por barrio

Evaluación e interpretación: Los 25 barrios más peligrosos demuestran una tasa de crímenes por kilómetro cuadrado superior a la media y mediana en todos sus casos. Además, a partir del barrio en la posición 20 (Lower Moyamensing), las tasas según el promedio general se duplican hasta llegar a ser cuatro veces mayor al mismo (McGuire).

5.1.2. Distritos policiales

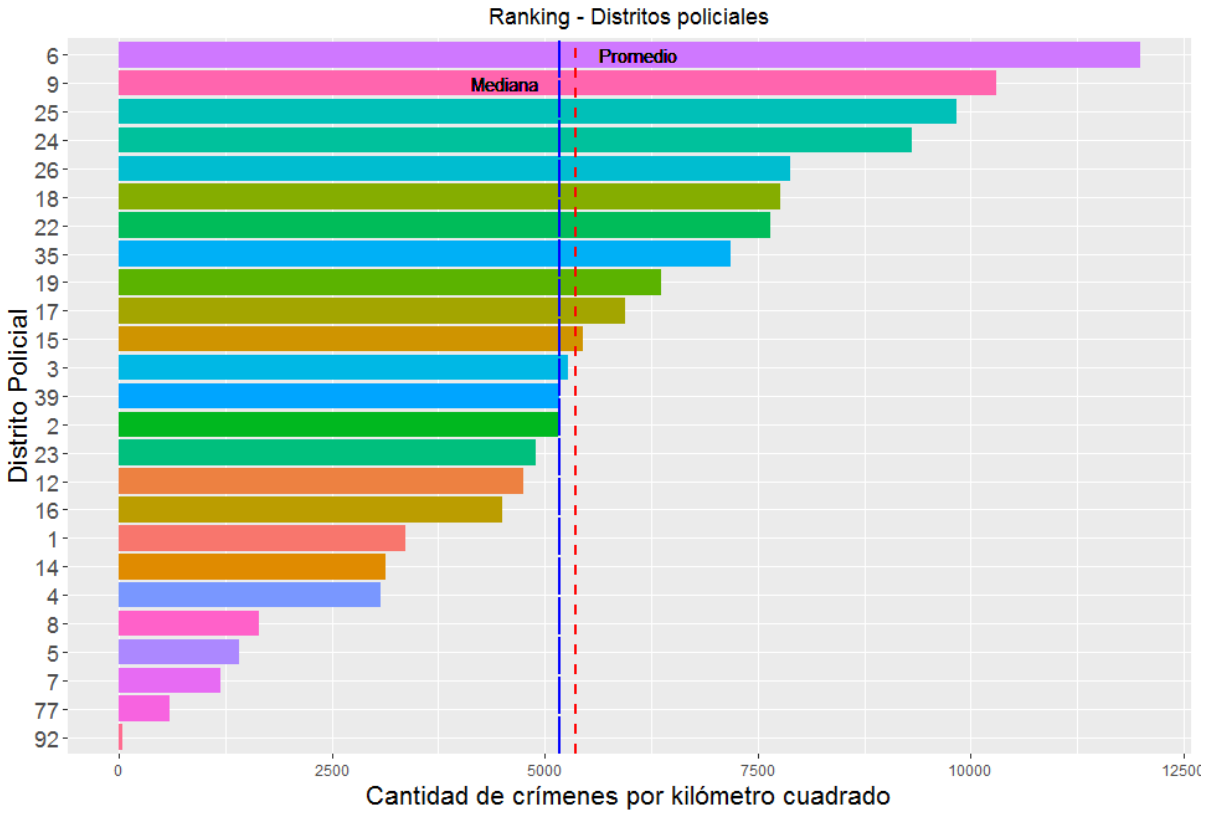


Ilustración 5.2: Ranking - Cantidad de crímenes por distrito policial

Evaluación e interpretación: En la totalidad de distritos policiales, al menos el 56% de estos se encuentran por debajo de la media, indicando que existe un crecimiento positivo lineal. Además, según la media, un solo distrito la duplica por completo una única vez (Distrito 6), y también, los distritos 77 y 92 la reducen en casi su totalidad.

5.1.3. Calles

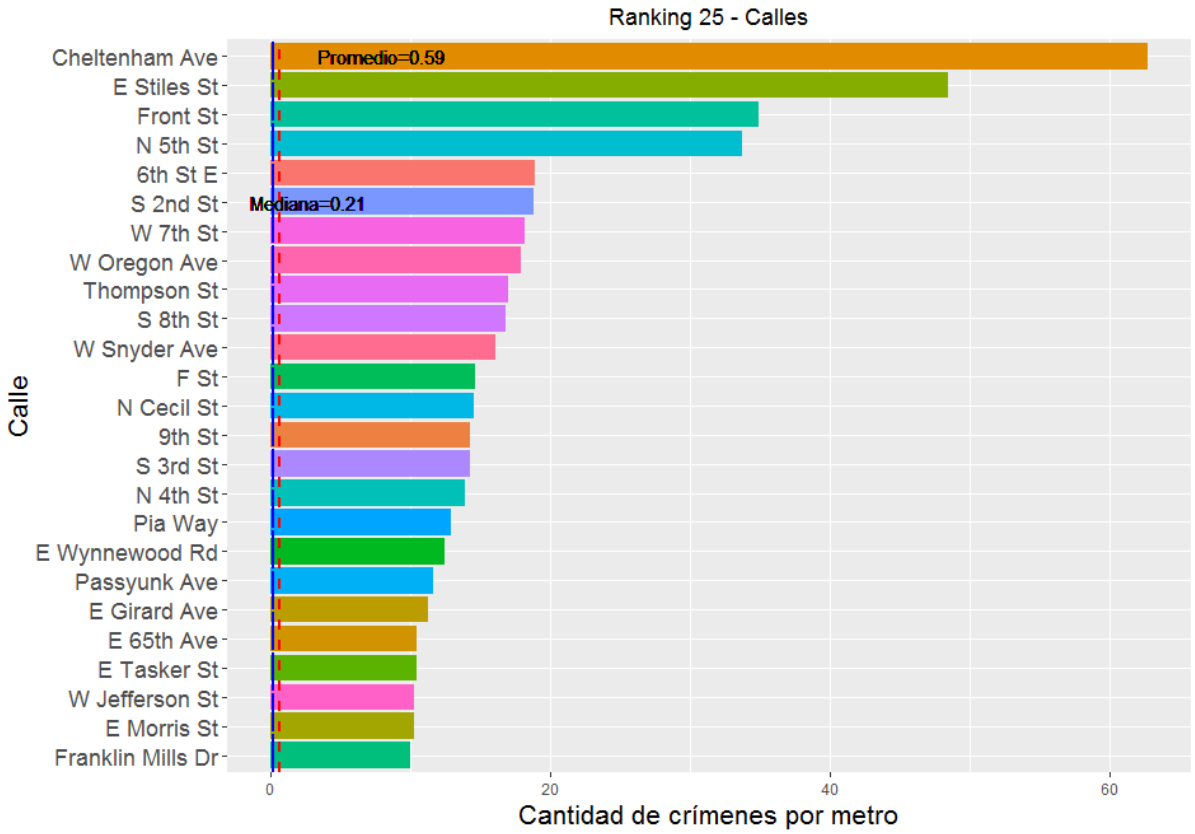


Ilustración 5.3: Ranking 25 - Cantidad de crímenes por calle

Evaluación e interpretación: Las cuatro primeras calles presentan una diferencia notable con respecto a las demás dentro de este ranking, ya que las demás tienen un incremento pequeño entre calles sucesivas. En cuanto a las medidas de tendencia, existe un gran incremento en cuanto las mismas y las calles de este ranking ya que, por ejemplo, la primera calle (Cheltenham Ave) presenta un incremento de 106 veces la media, y la última de 17 veces (Franklin Mills Dr). Junto con que solo el 20% de las calles de la ciudad supera a la media, las calles de este ranking se pueden considerar con tasas atípicas.

5.2. Ranking por tipo de crímenes

5.2.1. Ciudad de Filadelfia

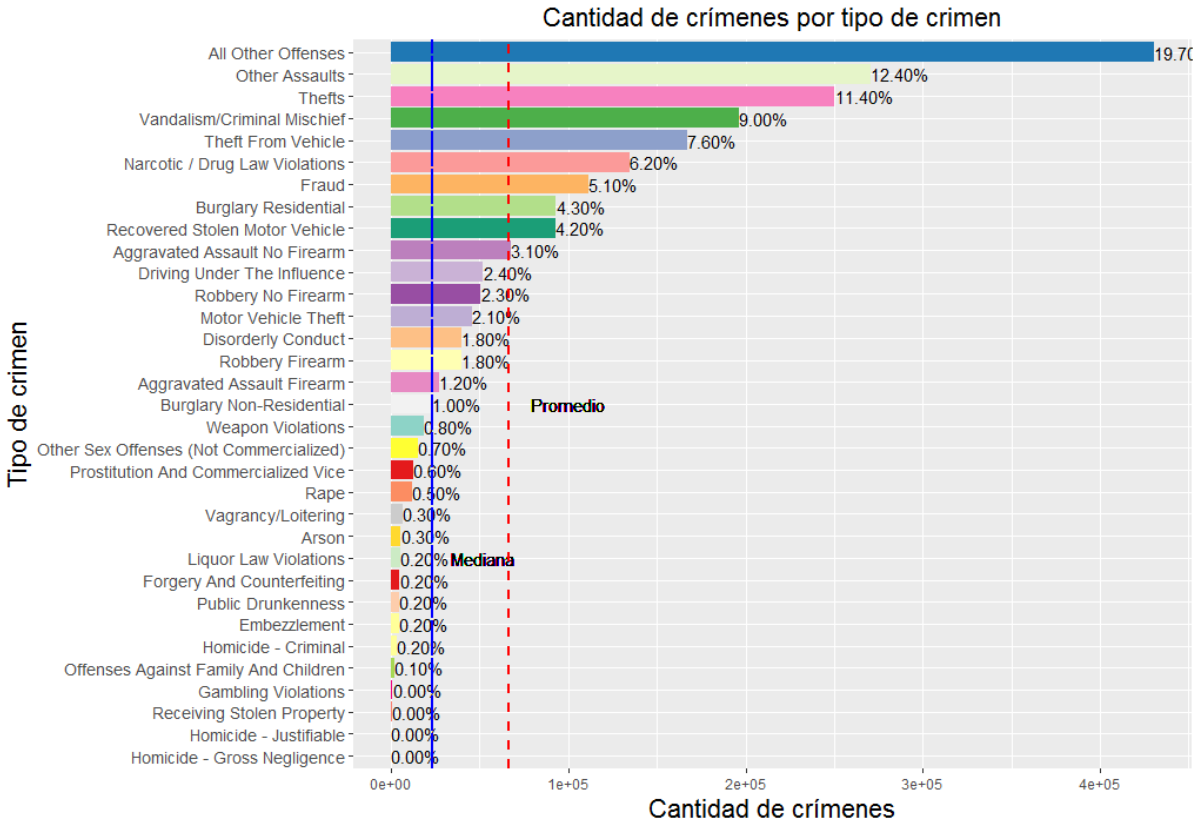


Ilustración 5.4: Cantidad de crímenes por tipo de crimen

Evaluación e interpretación: El 24% de los distintos tipos de crímenes supera a la media. En cuanto a los tipos más frecuentes, los primeros cuatro abarcan el 52.5% de la totalidad de reportes en la ciudad, donde un 19.7% se corresponde únicamente a "All Other Offenses", el tipo de crimen más frecuente. Junto con que al menos los primeros 13 crímenes representan el 80% de todos los reportes.

En los casos de los últimos cuatro tipos, representan un muy bajo porcentaje de la totalidad de crímenes.

5.2.2. Barrios

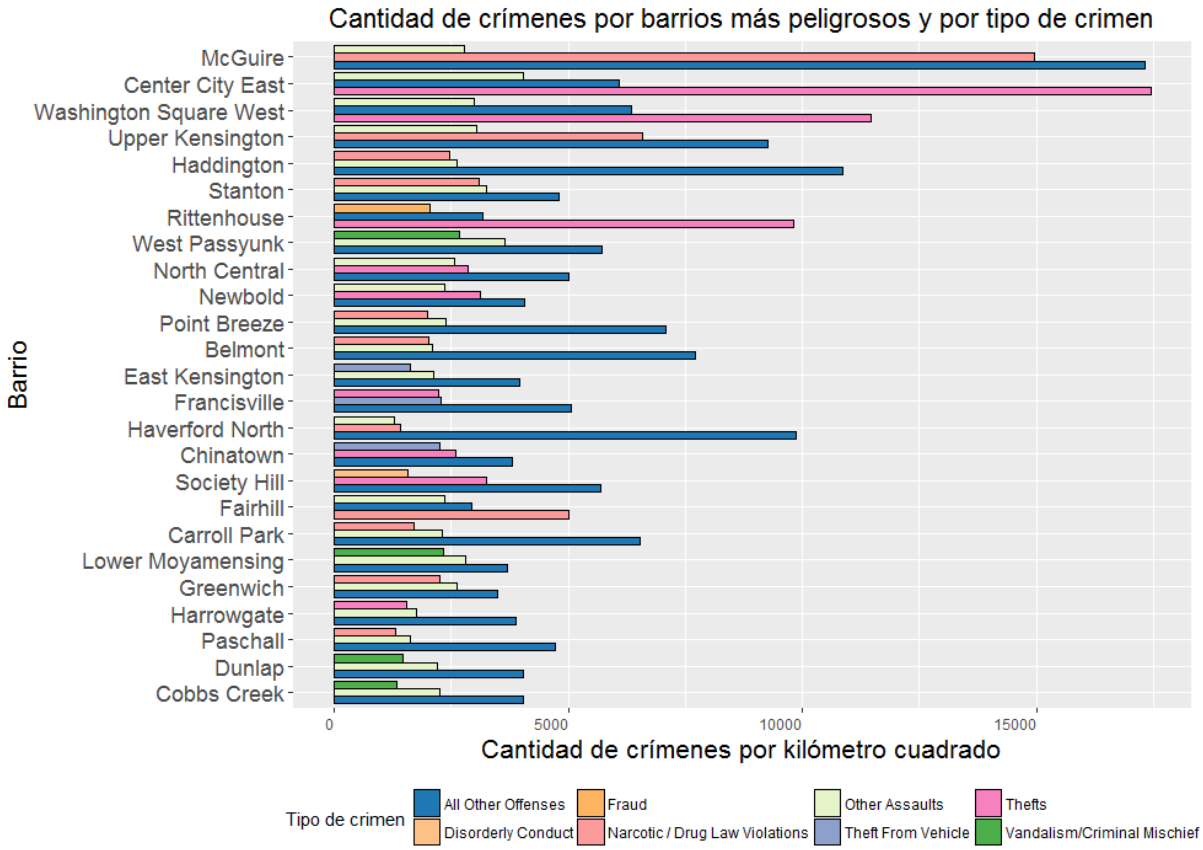


Ilustración 5.5: Cantidad de crímenes por barrios más peligrosos (25) y tipo de crimen (3)

Evaluación e interpretación: El barrio más peligroso, se caracteriza por tener una tasa alta de "All Other Offenses" y "Narcotic / Drug Law Violations", donde este último no es tan frecuente en los demás barrios.

En la mayor parte de los barrios, el tipo de crimen más frecuente por kilómetro cuadrado es "All Other Offenses", con la salvedad de que al menos tres barrios (Center City East, Washington Square West, Rittenhouse) se caracterizan por tener una gran tasa de "Thefts", y el barrio Fairhill se caracteriza por "Narcotic / Drug Law Violations". De igual manera, "All Other Offenses" siempre se encuentra presente en los demás barrios, y siendo otro tipo de crimen muy frecuente, "Other Assaults", ya que se encuentra en la mayoría de los subrankings de los barrios más peligrosos.

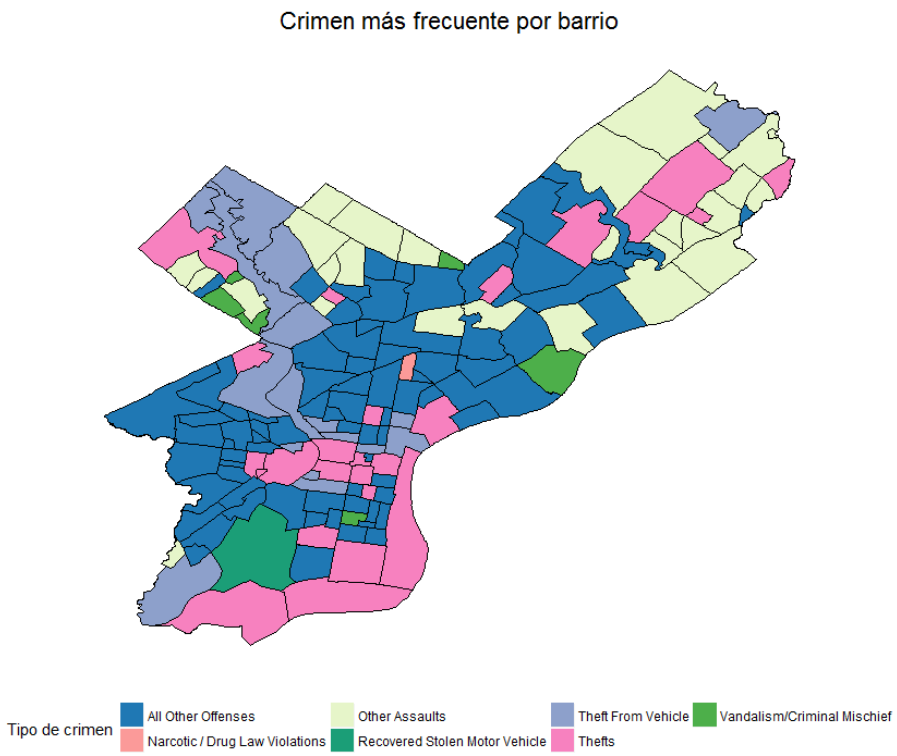


Ilustración 5.6: Mapa con el crimen más frecuente por barrio

Evaluación e interpretación: Si bien no existe un patrón claro en la distribución de tipos de crímenes en la ciudad, se puede ver algunos subpatrones como: En el norte y noroeste hay una secuencia de barrios adyacentes que se caracterizan por "Theft From Vehicle", a lo largo del rio por la parte sur se caracteriza por "Thefts", el oeste y la mayor parte del norte por "All Other Offenses" y el noreste por "Other Assaults".

5.2.3. Distritos policiales

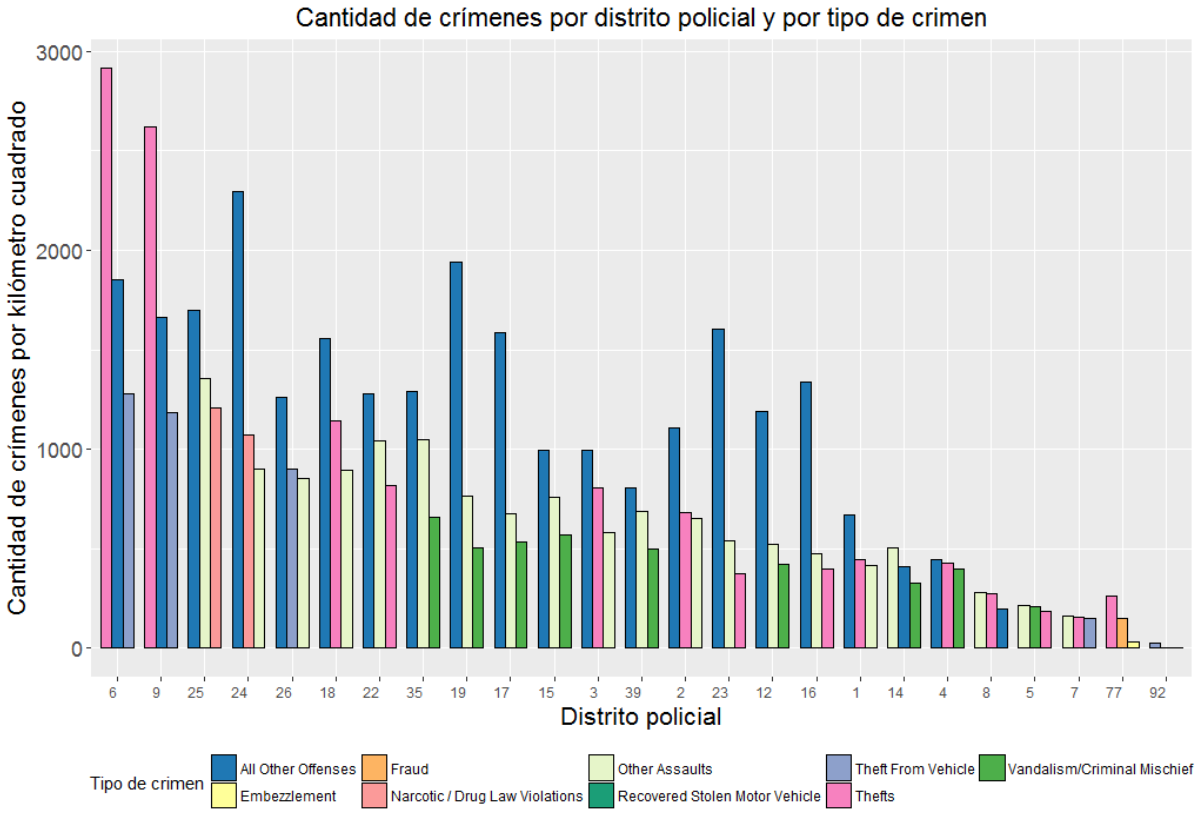


Ilustración 5.7: Cantidad de crímenes por distrito policial y tipo de crimen (3)

Evaluación e interpretación: Los dos distritos (6 y 9) con mayor tasa criminal, se caracterizan por tener una tasa alta de "Thefts" en comparación de los demás tipos, seguidos de "All Other Offenses" y "Theft From Vehicle".

Los demás distritos intermedios (25 a 1) están caracterizados por tener en primer lugar "All Other Offenses", y en las demás posiciones, "Other Assaults". De los cuales, los dos primeros, y los únicos distritos, poseen también "Narcotic / Drug Law Violations". Y, por último, otro tipo de crimen frecuente en estos distritos es "Vandalism/Criminal Mischief", el cual no era tan frecuente en los análisis por barrio.

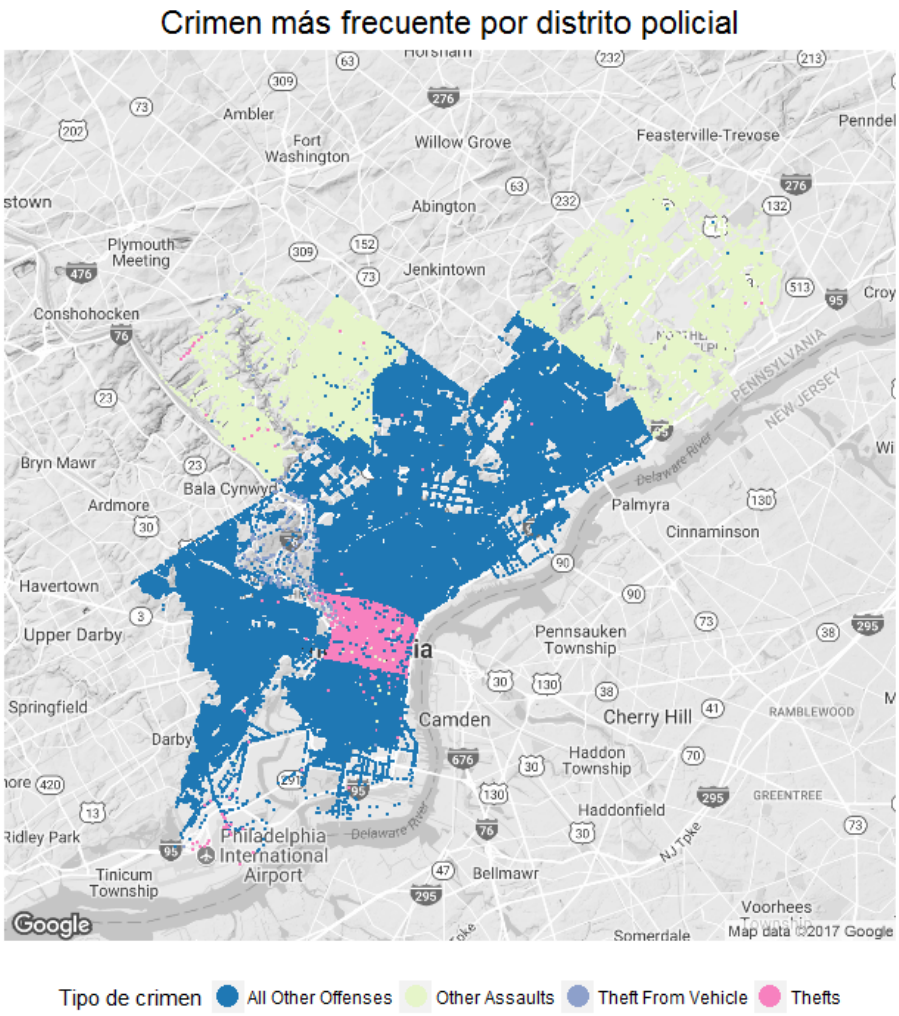


Ilustración 5.8: Mapa con el crimen más frecuente por distrito policial - Solo reportes

Evaluación e interpretación: En cuanto a la ubicación de los distritos en el mapa, la distribución es bastante clara, donde claramente resaltan 4 tipos de crímenes:

- "Thefts", se encuentran en los distritos pertenecientes a la región céntrica de la ciudad.
- "Theft From Vehicle" a lo largo del centro del noroeste.
- "Other Assaults" en los distritos del noroeste y noreste.
- "All Other offenses", en las zonas restantes.

5.3. Zonas de concentración

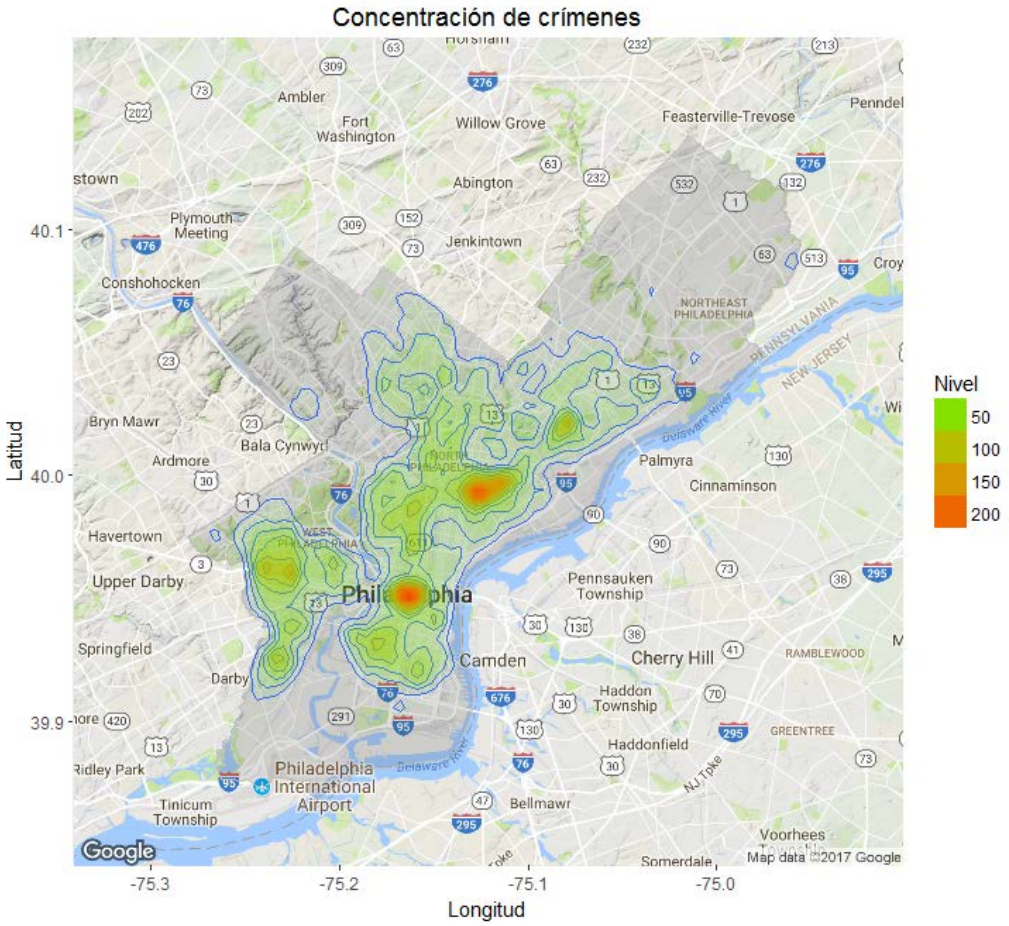


Ilustración 5.9: Mapa con zonas de concentración

Evaluación e interpretación: Las zonas de mayor concentración se encuentran claramente en dos puntos principales: En la región céntrica y en los barrios Kensington, Bridesburg y Richmond. Luego, existen ciertas zonas parciales y menores como en oeste (Con una leve zona en suroeste y noroeste), en el norte, una sección del sur y noreste de la ciudad. Pero la concentración general de crímenes, está dada en el centro de los límites de la ciudad.

5.4. Divisiones geográficas con mayor y menor incremento

5.4.1. Barrios

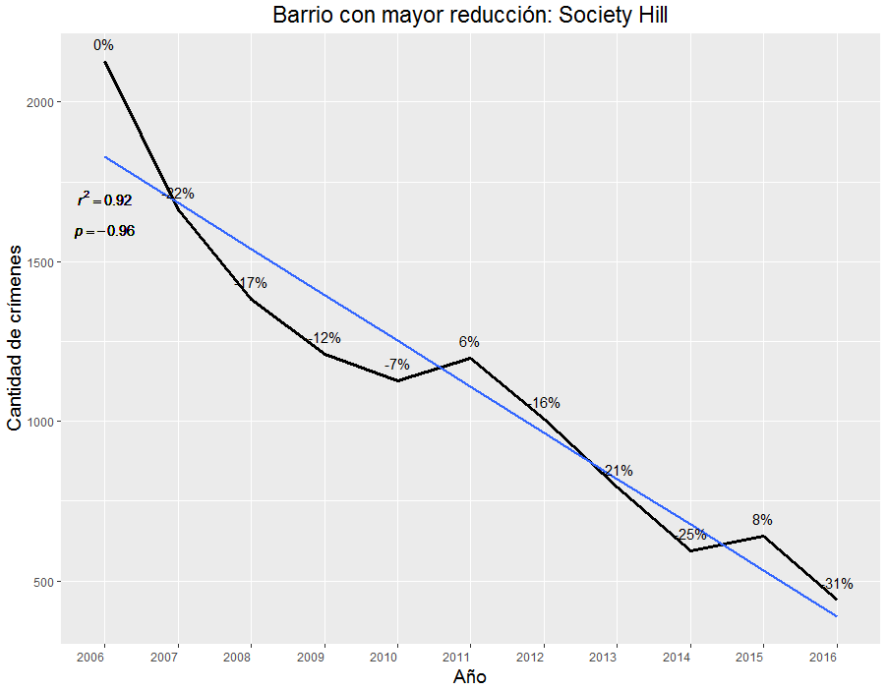


Ilustración 5.10: Barrio con mayor reducción en la tasa de crímenes

Evaluación e interpretación: El barrio obtuvo una reducción mayormente continua a lo largo de los años, con la excepción de los años 2011 y 2015, donde hubo un incremento del 7% en promedio. Además, tiene un comportamiento lineal ya que se obtuvo correlación lineal negativa fuerte.

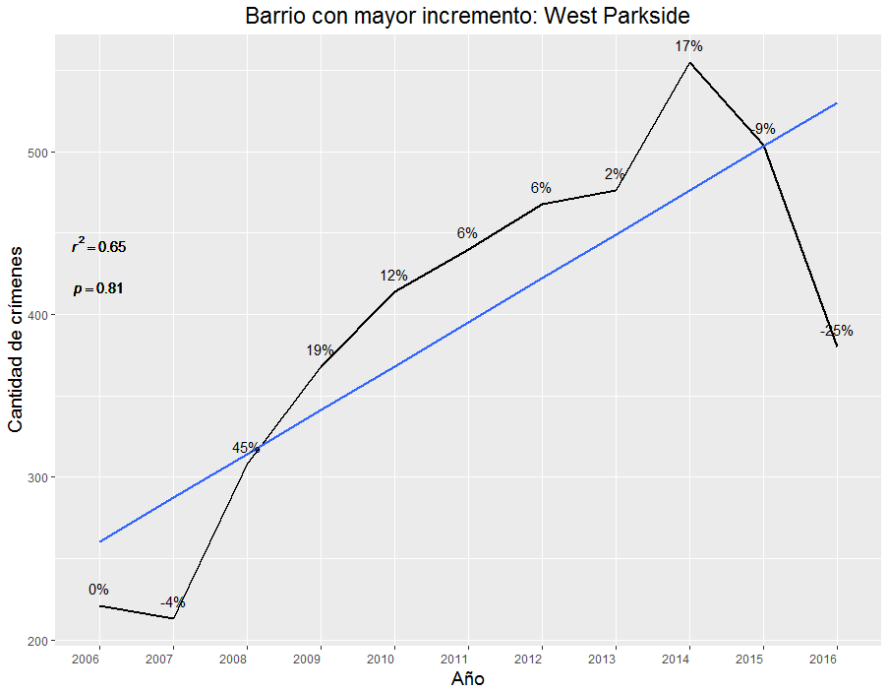


Ilustración 5.11: Barrio con mayor incremento en la tasa de crímenes

Evaluación e interpretación: Si bien obtuvo una reducción del 4% en el año 2007, la tasa comenzó con un incremento del 45% (El mayor en la historia de este barrio) y

se mantuvo el mismo durante 7 años. Pero luego del año 2015, comenzó una reducción considerable de la tasa, hasta ser similar al año 2009.

5.4.2. Distritos policiales

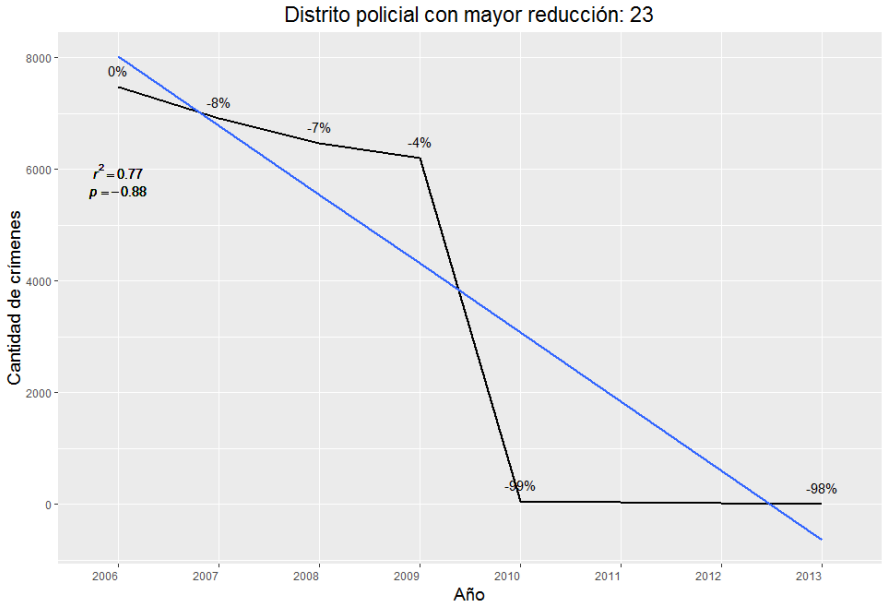


Ilustración 5.12: Distrito policial con mayor reducción en la tasa de crímenes

Evaluación e interpretación: A diferencia de los barrios, la línea de tiempo se desarrolló durante un lapso de 8 años, ya que a partir del año 2013 no se registraron más crímenes. Dado esto, la reducción fue continua, con una diferencia notable en los años 2010 y 2013, que indican que se redujo las tasas de crímenes en un 99% y 98%, casi en su totalidad.

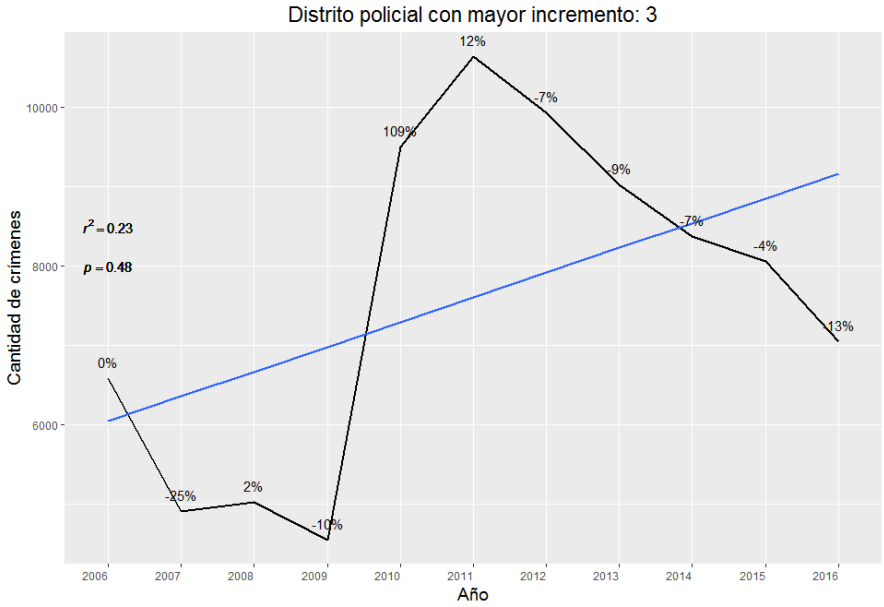


Ilustración 5.13: Distrito policial con mayor incremento en la tasa de crímenes

Evaluación e interpretación: No se observa un comportamiento lineal, tanto gráficamente como estadísticamente, porque hubo alternaciones entre incrementos y reducciones con diferencias notables en los distintos años.

La mayor reducción se dio en el año 2007 con un 25% con respecto al año anterior, y el mayor incremento fue del 109%, algo sumamente considerable ya que se duplico la tasa. En el año 2011, comenzó una reducción continua, pero, sin embargo, las tasas de crímenes no se redujeron a menos que el año 2006.

5.5. Análisis por intervalos de tiempos

5.5.1. Años

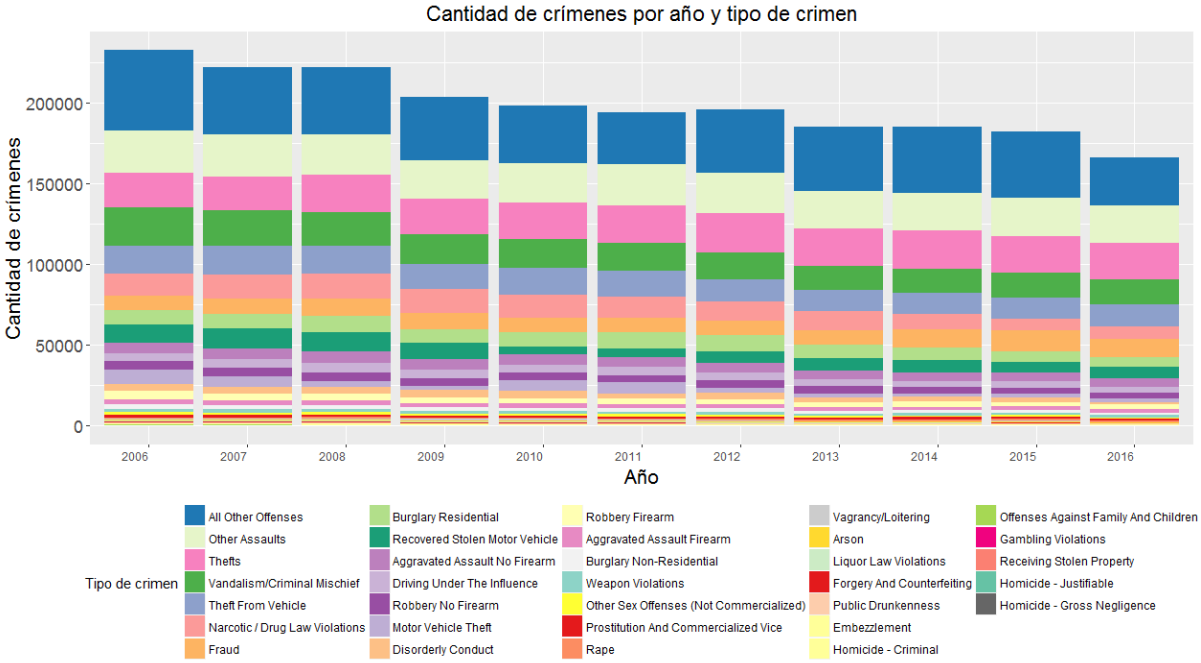


Ilustración 5.14: Cantidad de crímenes por año y tipo de crimen

Evaluación e interpretación: La primera observación es la decreciente tasa de crímenes a lo largo de los años. La segunda, es que los tipos de crímenes se mantienen en las mismas posiciones del subranking por año. Y, por último, existen ciertos cambios a lo largo de los años, como en el rango de 2006 a 2008, 2009 a 2012, 2013 a 2016 y 2016.

5.5.2. Estaciones

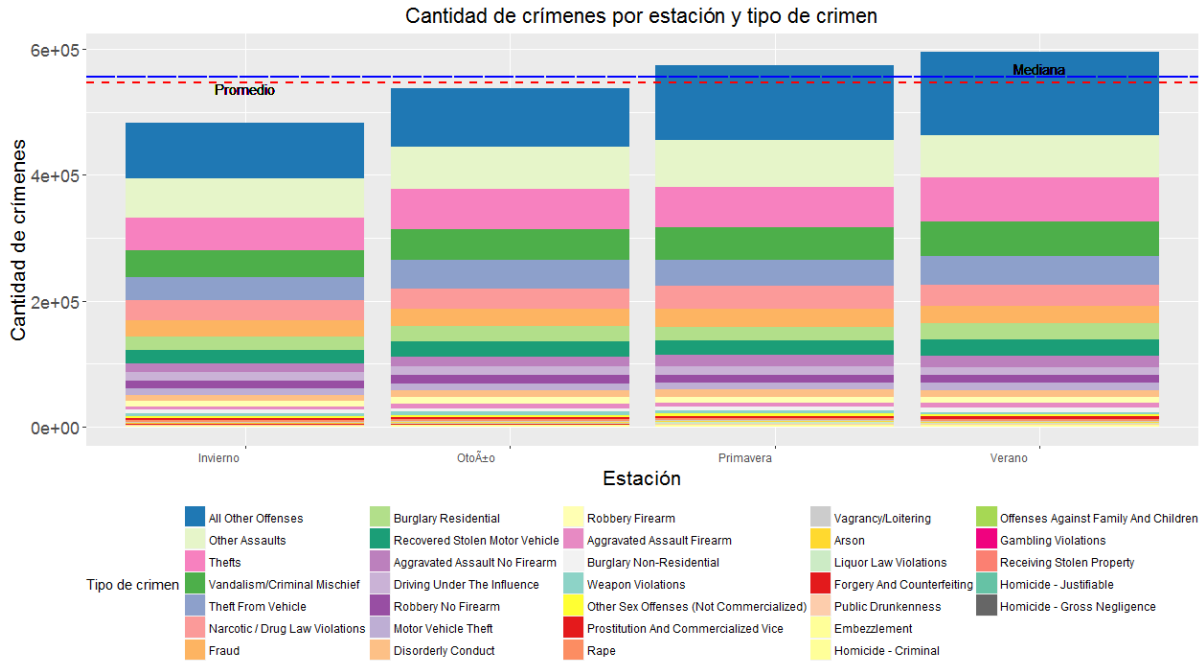


Ilustración 5.15: Cantidad de crímenes por estación y tipo de crimen

Evaluación e interpretación: De igual manera que la distribución por año, los tipos de crímenes se mantienen sin modificaciones a lo largo de las estaciones. En cuanto a la media y mediana, las estaciones con menor temperatura (invierno y otoño) se encuentran por debajo de las mismas, y aquellas estaciones con mayor temperatura, por arriba de estas. Esto puede llevar a considerar una asociación entre la tasa y la temperatura, donde a mayor temperatura, mayor la tasa de crímenes.

5.5.3. Meses

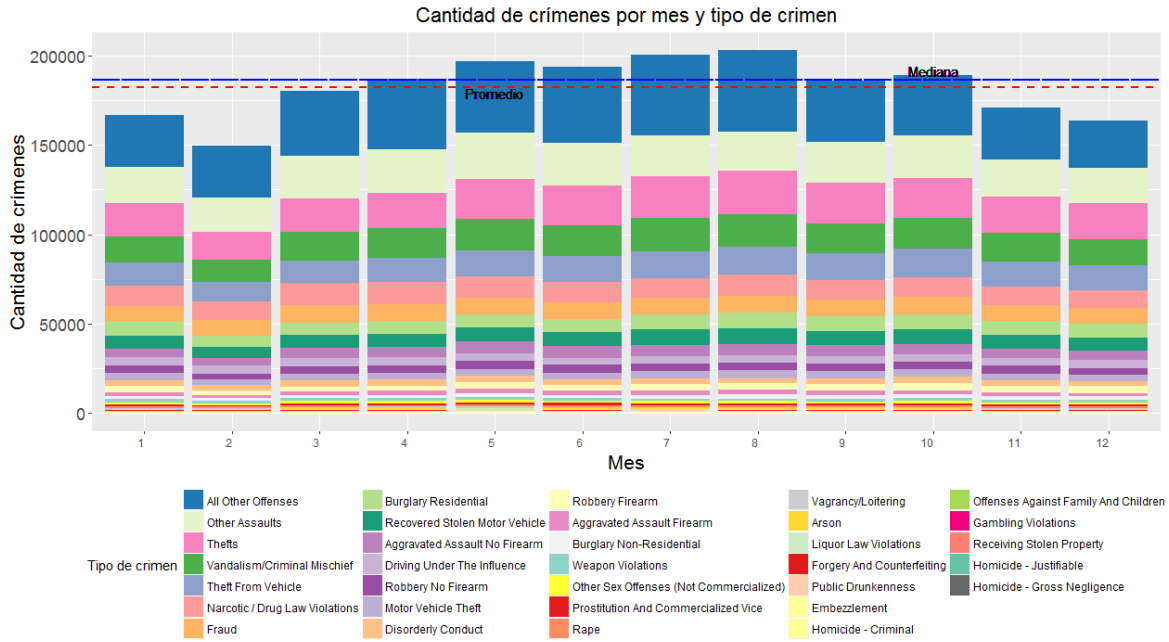


Ilustración 5.16: Cantidad de crímenes por mes y tipo de crimen

Evaluación e interpretación: Como en las estaciones, existe una diferencia entre los meses con menor y mayor temperatura. En este caso, tomando como medida de referencia a la media, los meses que la superan son desde abril a octubre, de los cuales, agosto es el mes con mayor cantidad de crímenes en todo el año. Con la salvedad de marzo, los demás meses que no superan a la media poseen una diferencia notable con respecto a los demás, junto con que el mes con menor cantidad de crímenes es febrero.

5.5.4. Días

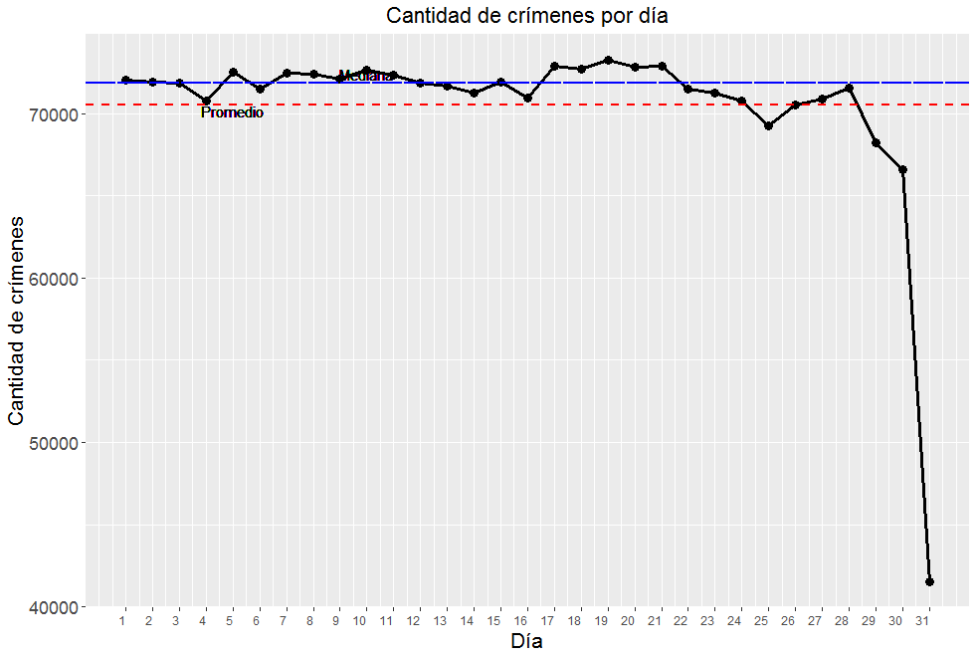


Ilustración 5.17: Cantidad de crímenes por día

Evaluación e interpretación: Si bien no todos los meses del año poseen 31 días (El 42%), se observa que el día 31 es aquel con menor tasa criminal, y el día 19, es aquel con la mayor tasa.

Dado que el día 31 posee una gran diferencia con los demás, la media se ve sesgada, por lo que la comparación es debidamente correcta con la mediana. Ya esto, se observan tres rangos de días que la superan: 1 a 3, 7 a 11, 17 a 21. También, se observa un decaimiento de la tasa a partir del día 22, ya que hay varios días que tienen tasas bajas en comparación de las demás.

5.5.5. Horas

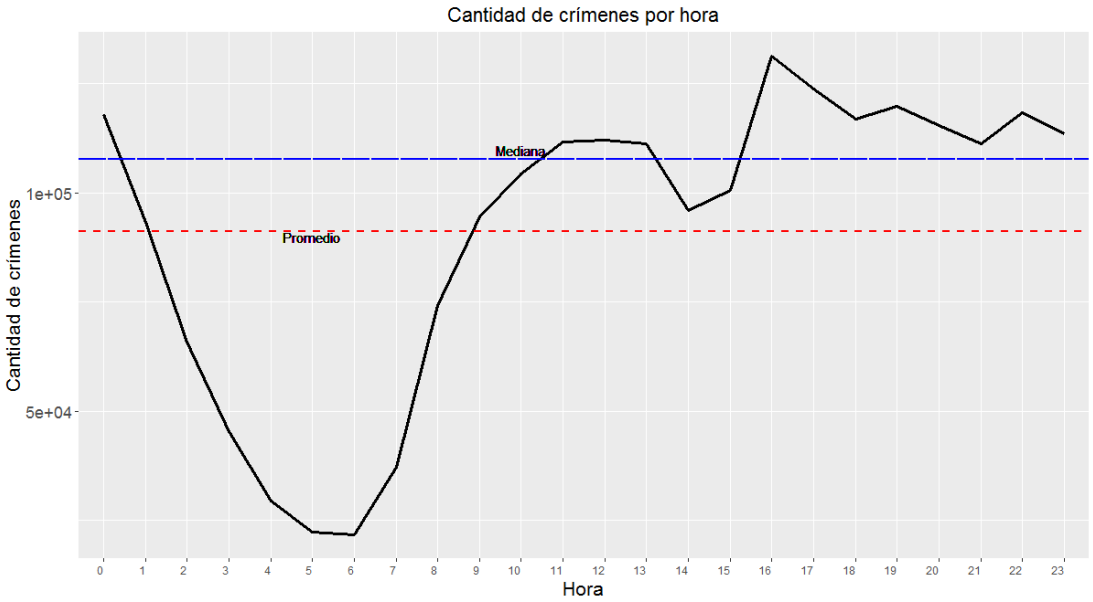


Ilustración 5.18: Cantidad de crímenes por hora

Evaluación e interpretación: Dado la poca cantidad de crímenes en los horarios de madrugada, se utiliza la mediana como medida de comparación. Entonces, los horarios que la superan se encuentran en los rangos de 11 a 13 horas y 16 a 0 horas.

La hora con mayor cantidad de crímenes es a las 16, y la menor cantidad, es a las 6.

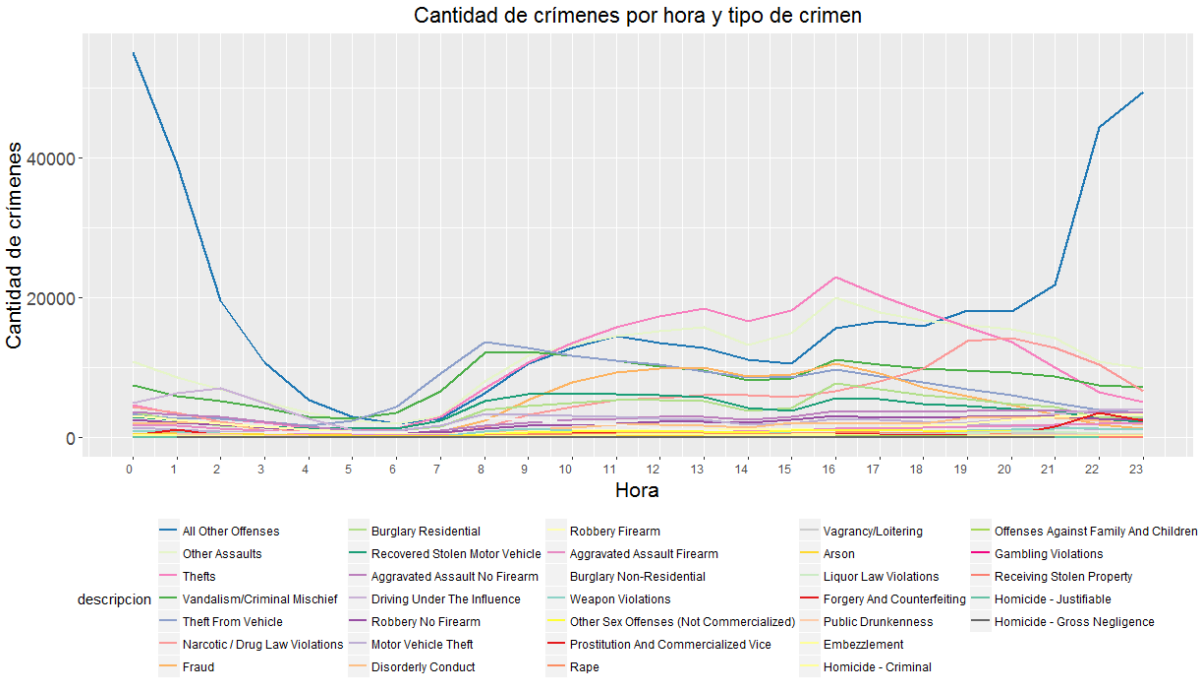


Ilustración 5.19: Cantidad de crímenes por hora y tipo de crimen

Evaluación e interpretación: A diferencia de la distribución anterior, no se sigue la misma por cada tipo de crimen.

Las distinciones más importantes son:

- “All Other Offenses” tiene una mayor tasa de crímenes en la franja horaria de 19 a 5 horas, con una gran distinción de los demás tipos durante la misma.
- “Thefts From Vehicle” tiene una mayor tasa de crímenes en la franja horaria de 6 a 9 horas.
- “Thefts” tiene una mayor tasa de crímenes en la franja horaria de 10 a 18 horas.

5.6. Correlación Crimen / Temperaturas

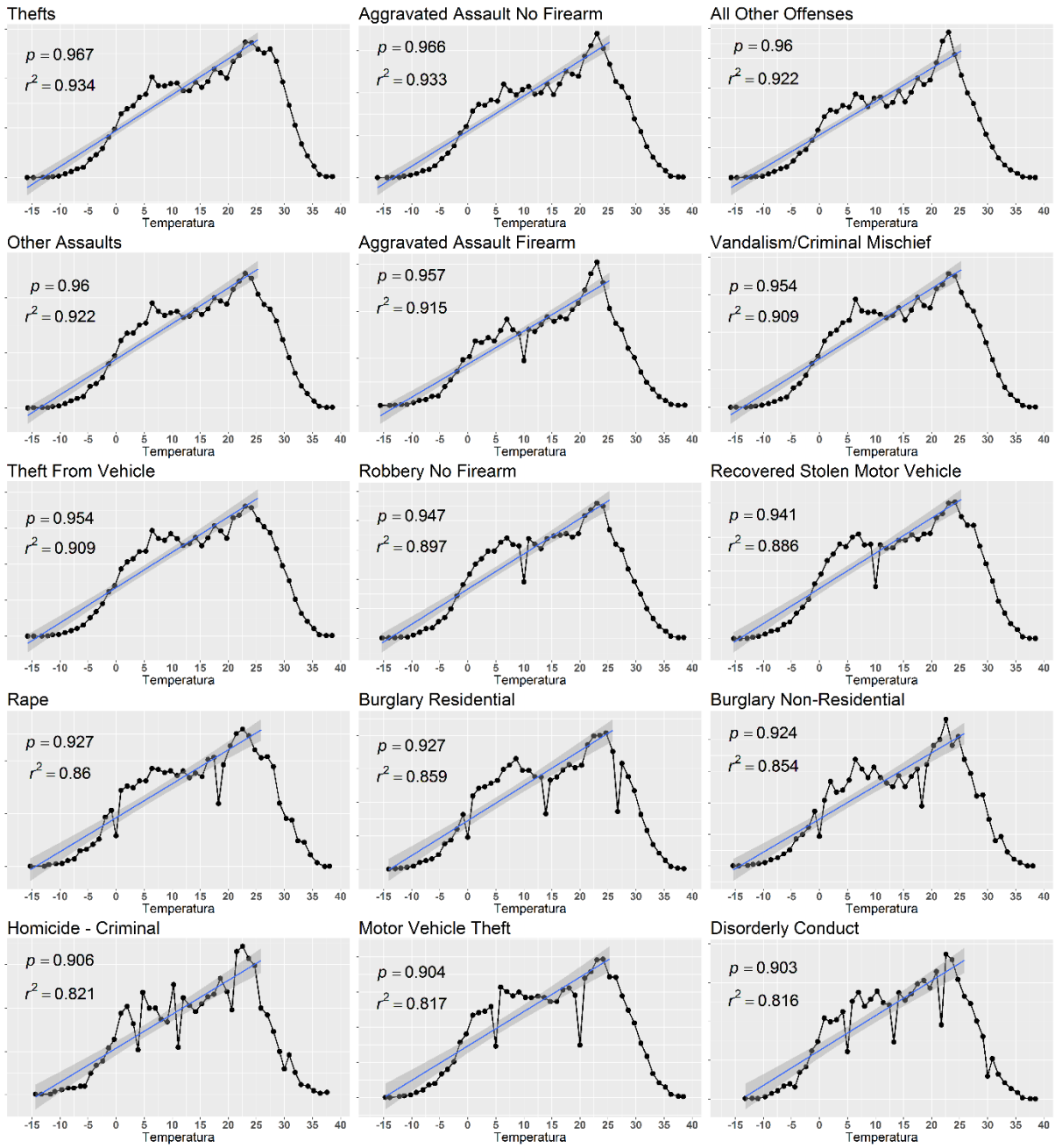


Ilustración 5.20: Cantidad de crímenes por temperatura y tipo de crimen - Parte 1

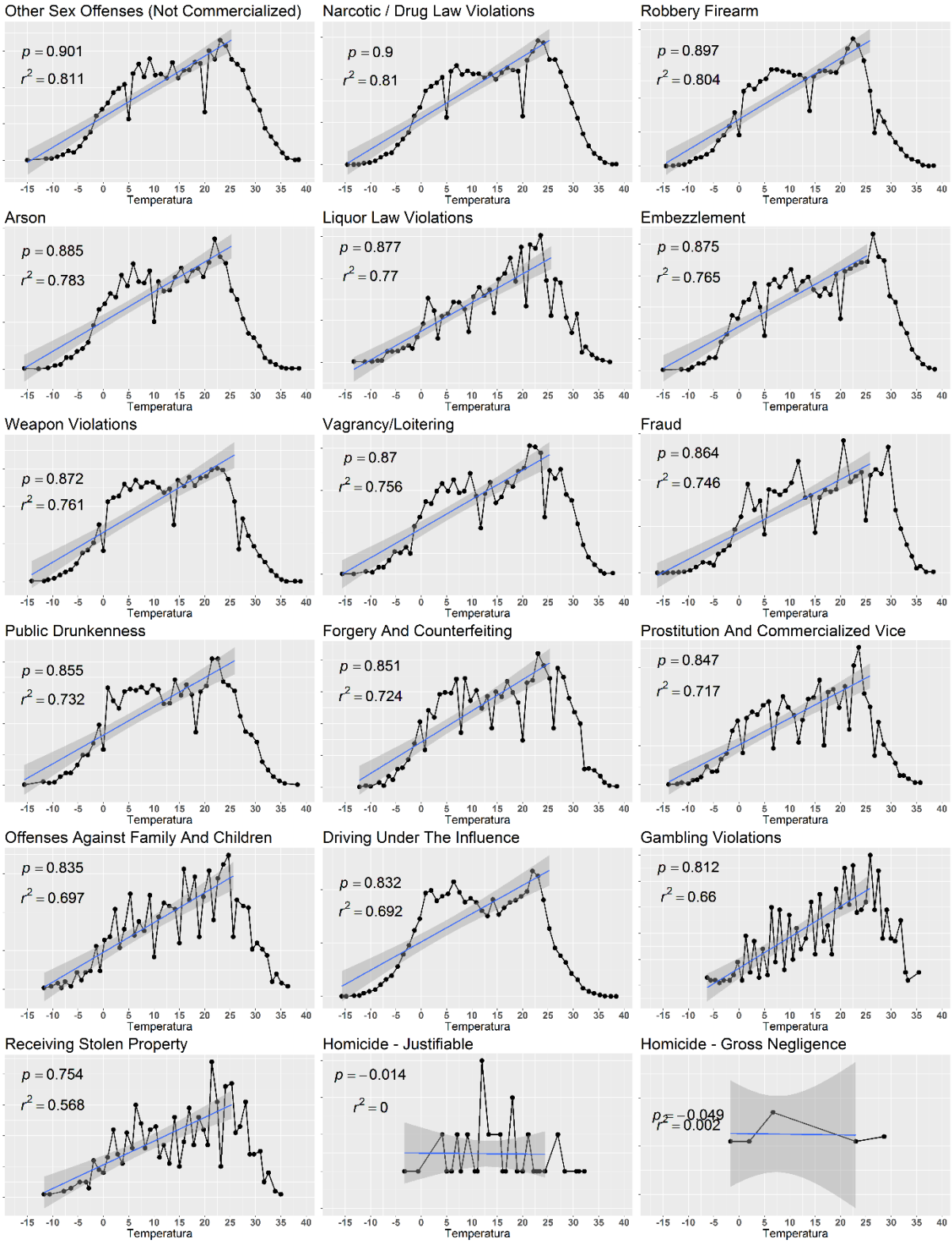


Ilustración 5.21: Cantidad de crímenes por temperatura y tipo de crimen - Parte 2

Evaluación e interpretación: La primera observación, y en correspondencia a lo que se dijo en la sección 5.5.2, se observa una correlación positiva entre la cantidad de crímenes y la temperatura. Además, como se vio en la sección 4.4.6, a partir de los 26°C en promedio y dependiendo del tipo de crimen, hay un decaimiento en las tasas.

La segunda observación, es que los primeros cuatro tipos junto con “Vandalism/Criminal Mischief” y “Theft From Vehicle” tienen tasas de crímenes con baja variación en temperaturas consecutivas, por eso las correlaciones más fuertes en todo el conjunto. Además de que por un lado, tres de estos son los primeros en el ranking general (Ver Ilustración 5.4), y por otro, las mayores tasas de crímenes se encuentran en el rango 20°C a 25°C.

Otra observación, se corresponde a que los crímenes “Weapon Violations” y “Driving Under The Influence”, que poseen un comportamiento más estable en la parte central de la distribución (En el rango promedio de 0°C a 22.5°C).

También se observa que crímenes tales como “Receiving Stolen Property”, “Gambling Violations”, “Offenses Against Family And Children”, “Forgery And Counterfeiting” y “Prostitution And Commercialized Vice” tienen una baja correlación en relación con los primeros, debido a que las tasas de crímenes poseen un comportamiento mucho más disperso entre temperaturas consecutivas.

Por último, los crímenes “Homicide - Justifiable”, “Homicide - Gross Negligence” no tienen un comportamiento lineal, debido a la poca cantidad de reportes de los mismos. Pero en relación con la temperatura, se puede establecer que el primero tiene una mayor tasa en el rango de 12°C a 18°C y el segundo, en 6.7°C.

5.7. Zonas propensas a la ocurrencia de crímenes

5.7.1. Cluster 1

Este cluster pertenece al 4% de la totalidad reportes diarios según la Ilustración 4.16.

5.7.1.1. Barrios

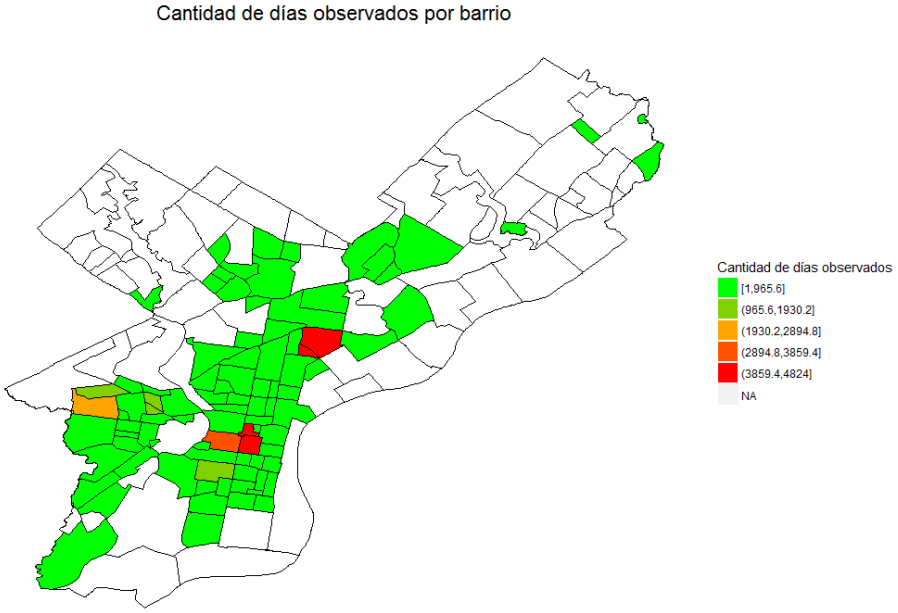


Ilustración 5.22: Mapa con cantidad de días observados por barrio - Cluster 1

Evaluación e interpretación: Este cluster solo contiene 92 barrios de la ciudad, donde la ubicación de los mismos es mayormente en los límites centrales de la misma, con ciertos barrios en el oeste. Existe ciertos barrios aislados, en el noreste, pero poseen poca cantidad de días observados.

De todos los barrios, solo 4 se caracterizan por tener el rango con mayor cantidad de días reportados.

5.7.1.2. Tipos de crímenes

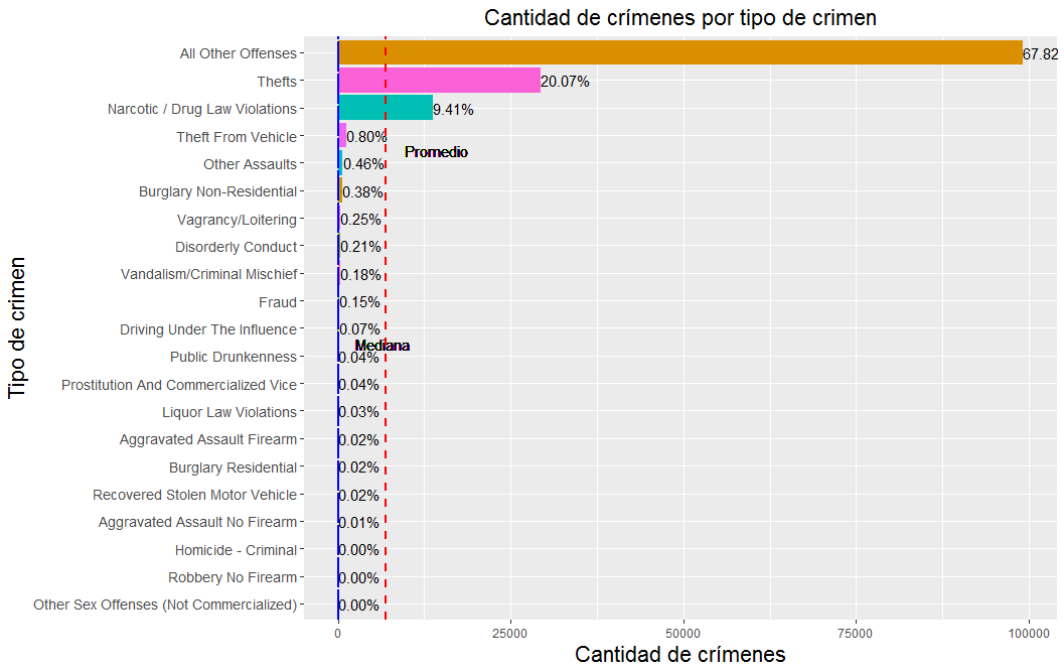


Ilustración 5.23: Cantidad de crímenes por tipo de crimen - Cluster 1

Evaluación e interpretación: De los 33 tipos de crímenes, solo 21 estuvieron presentes en este cluster, y de estos, solo 3 ("All Other Offenses", "Thefts" y "Narcotic / Drug Law Violations") hacen la diferencia ya que representan el 97.3% del cluster, donde el primer tipo ("All Other Offenses") abarca el 67.82% del mismo.

5.7.1.3. Tasas por lapsos de tiempo

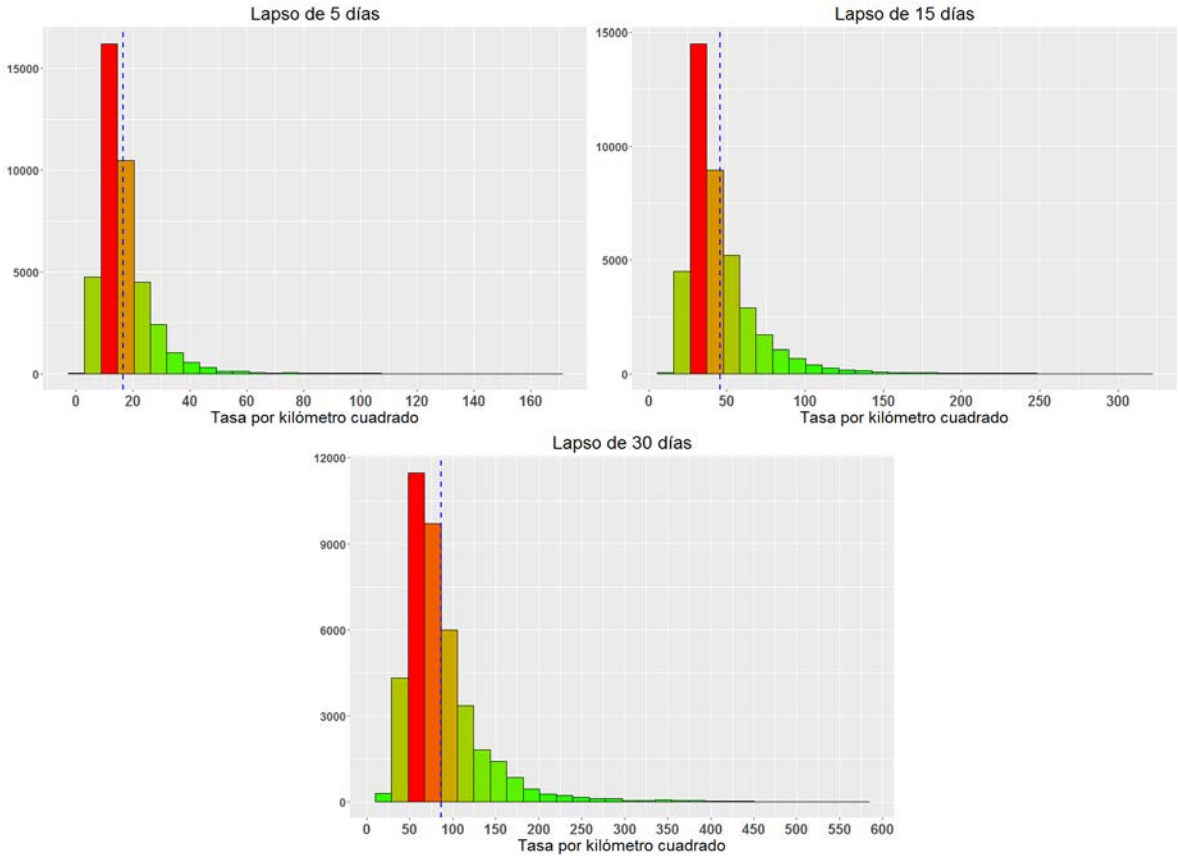


Ilustración 5.24: Histogramas de tasas por lapsos de tiempo - Cluster 1

Evaluación e interpretación: Las tasas en todos los lapsos siguen una distribución sesgada a la izquierda, con valores promedios aproximados en 16, 45 y 86 crímenes por kilómetro cuadrado en los anteriores 5, 15 y 30 días, donde las mayores concentraciones se encuentran cercanas a las medias.

5.7.1.4. Distribución de tasas por barrio

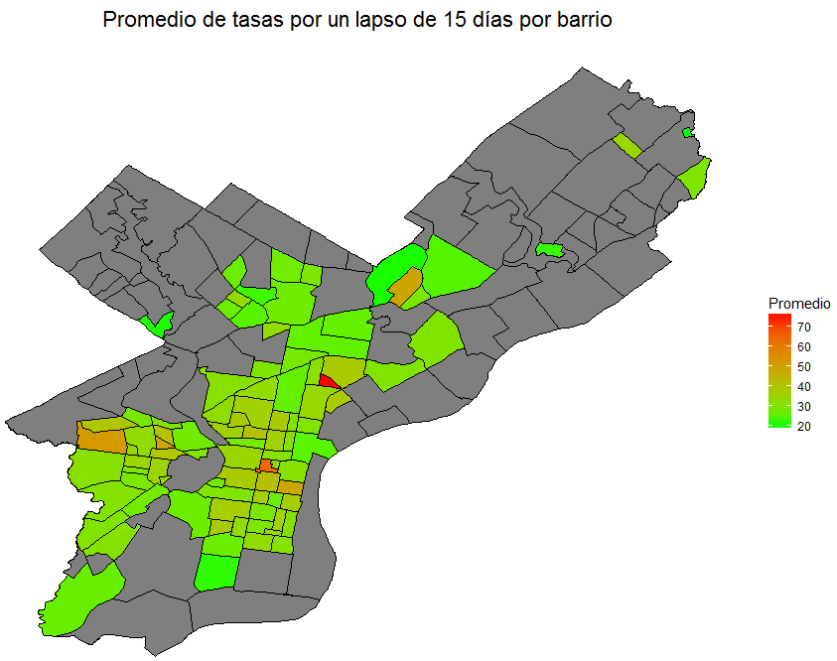


Ilustración 5.25: Mapa con distribución de tasas promedio por barrio - Cluster 1

Evaluación e interpretación: La distribución es bastante uniforme a lo largo de la ciudad, ya que no se observa una única región con concentración de tasas promedios altas. Solo se observan dos barrios (McGuire y Center City East) con una mayor tasa promedio en los 15 días anteriores del cluster, los cuales son aquellos con las mayores tasas en toda la ciudad (Ver la Ilustración 5.1).

5.7.1.5. Tiempo

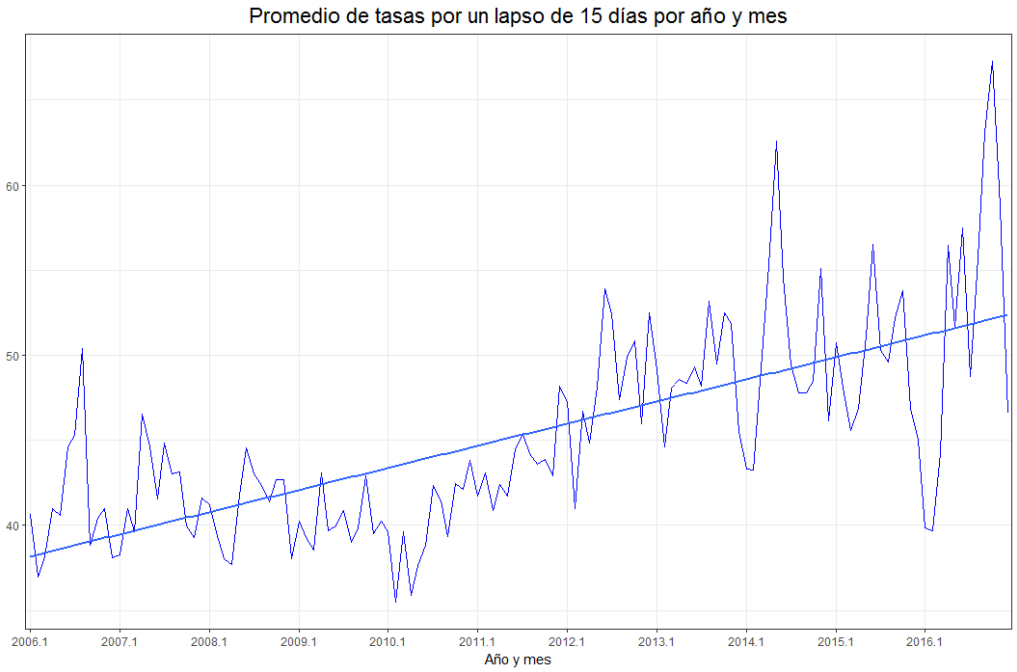


Ilustración 5.26: Serie de tiempo de los promedios de tasas - Cluster 1

Evaluación e interpretación: Se observa un decaimiento de las tasas promedio desde el año 2006 hasta el 2010, y luego comienza un incremento hasta el año 2016,

donde en a mediados de los años 2014 y 2016, se observan dos picos pronunciados con respecto a los demás años (mayo y octubre). Además, en promedio, desde el inicio de la serie se tenía una tasa de 40 crímenes en un lapso de 15 días, y a fines de la misma, se observa una tasa promedio de 55 crímenes, una mayor propensión en los crímenes de este cluster.

5.7.2. Cluster 2

Este cluster pertenece al 96% de la totalidad reportes diarios según la Ilustración 4.16.

5.7.2.1. Barrios

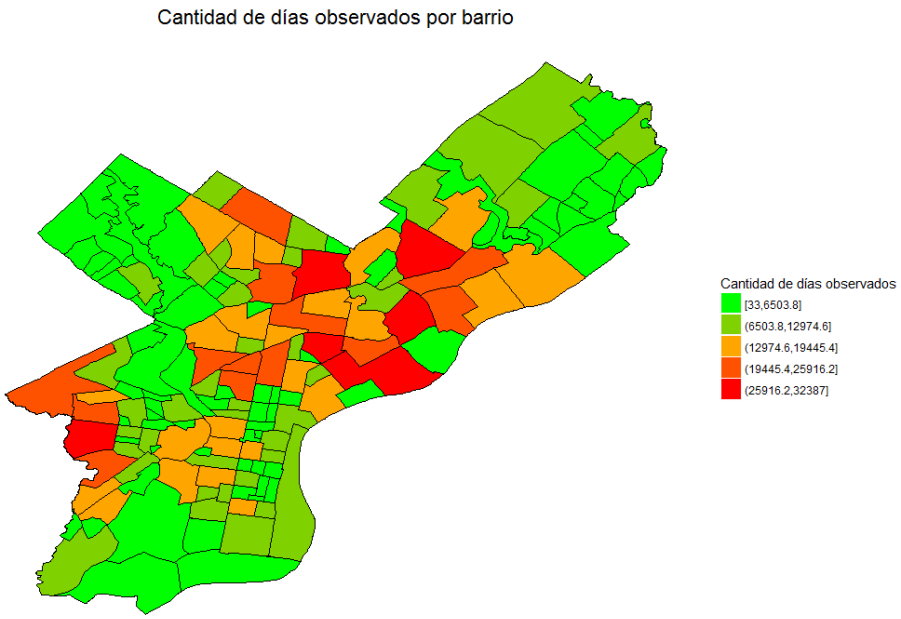


Ilustración 5.27: Mapa con cantidad de días observados por barrio - Cluster 2

Evaluación e interpretación: A diferencia del cluster anterior, las distribuciones de reportes de crímenes se encuentran en todos los barrios de la ciudad. De los cuales, aquellos con mayor presencia se pueden subdividir en dos zonas claras: Una en el oeste de la ciudad, y otra con varias zonas parciales en las fronteras de las regiones del norte, noreste, noroeste, y los barrios Kensington, Bridesburg y Richmond.

5.7.2.2. Tipos de crímenes

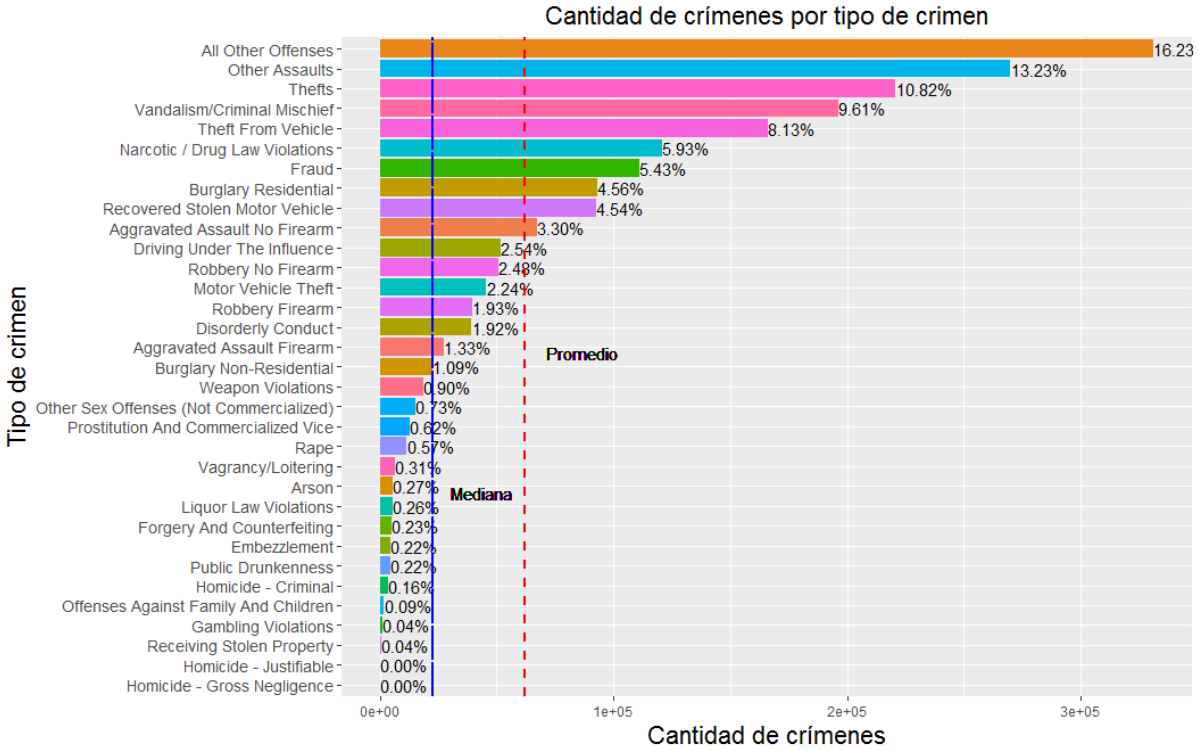
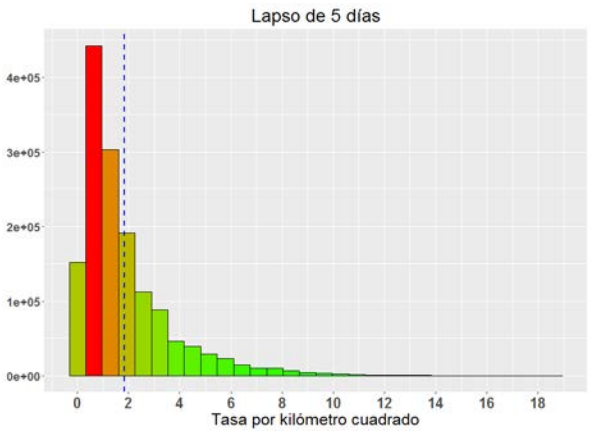
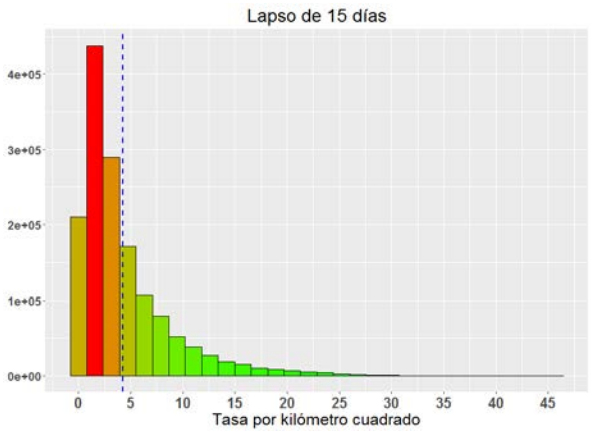


Ilustración 5.28: Cantidad de crímenes por tipo de crimen- Cluster 2

Evaluación e interpretación: A diferencia del cluster anterior, se sigue una distribución similar al ranking visto en la sección 5.2.1 (Tanto por la distribución dentro del ranking como la misma cantidad de tipos existentes).

5.7.2.3. Tasas por lapsos de tiempo





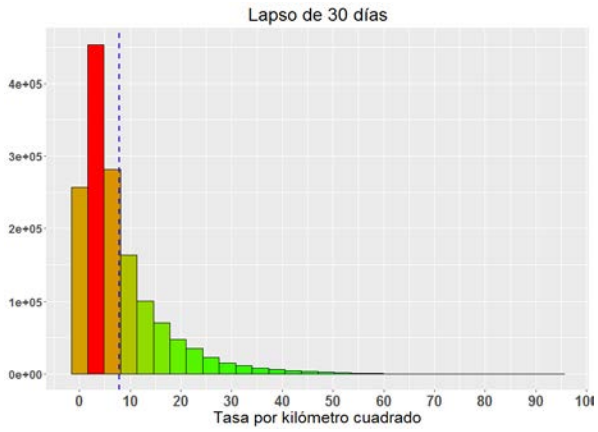


Ilustración 5.29: Histogramas de tasas por lapsos de tiempo - Cluster 2

Evaluación e interpretación: De igual manera que el cluster 1, se sigue una distribución sesgada a la izquierda, pero las tasas promedios en los distintos lapsos siempre son menores con correspondencia a tal cluster. Los promedios aproximados se corresponden a tasas de 1.8, 4.2 y 7.8 en los lapsos de 5, 15 y 30 días.

5.7.2.4. Distribución de tasas por barrio

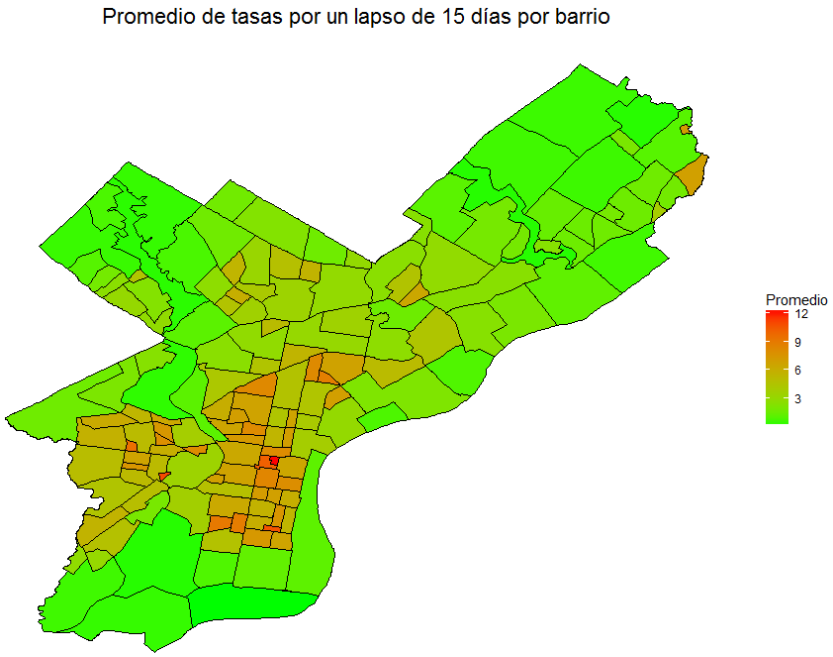


Ilustración 5.30: Mapa con distribución de tasas promedio por barrio - Cluster 2

Evaluación e interpretación: En este cluster, las mayores tasas de crímenes promedio se encuentran distribuidas dos posibles zonas: el oeste y la región céntrica, con limites en el sur y norte.

5.7.2.5. Tiempo

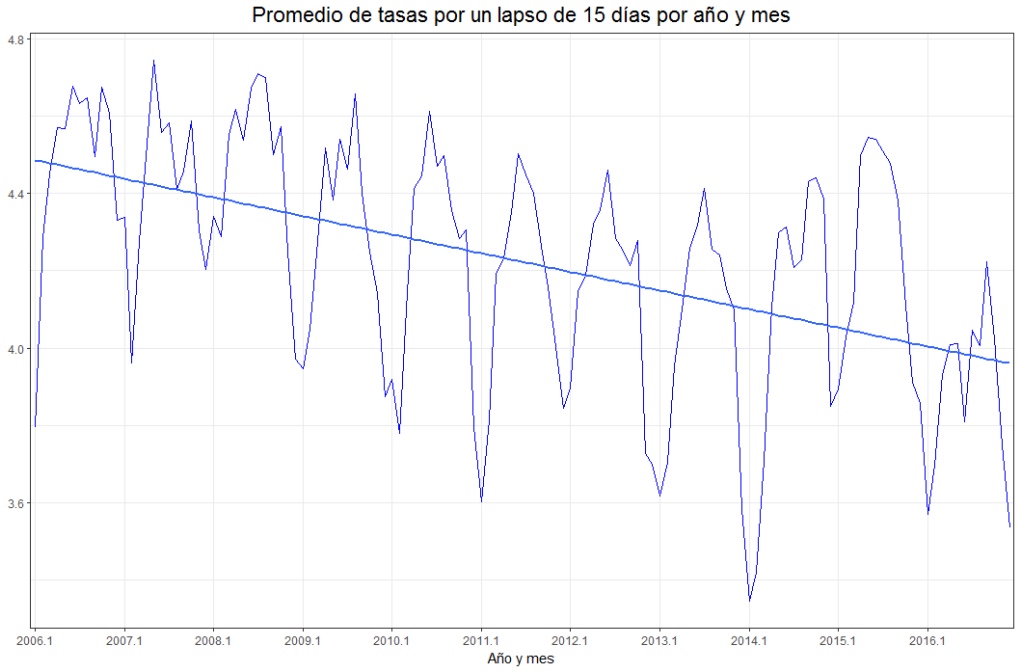


Ilustración 5.31: Serie de tiempo de los promedios de tasas - Cluster 2

Evaluación e interpretación: A diferencia con la serie de tiempo del cluster anterior, y en correspondencia a la Ilustración 5.14, se observa un decaimiento en las tasas promedios, con la salvedad de los años 2014 y 2015 que poseen un incremento en comparación de su año anterior. Por ende, se observa que los crímenes tienden a ser menos propensos a ocurrir a medida que pasan los años.

5.7.3. Caracterización principal

5.7.3.1. Cluster 1

Dado los análisis anteriores, este cluster se caracteriza básicamente por:

- Una concentración de crímenes en el centro de los límites de la ciudad, con zonas mayormente propensas en los barrios: McGuire y Center City East.
- Tres tipos de crímenes que superan a la media: “All Other Offenses”, “Thefts”, “Narcotic / Drug Law Violations”.
- Una tasa de crímenes alta en los distintos lapsos de tiempo, lo cual hace que los crímenes sean mucho más propensos a ocurrir.
- Una tasa de crímenes promedio por mes en crecimiento.

5.7.3.2. Cluster 2

Dado los análisis anteriores, este cluster se caracteriza básicamente por:

- Una distribución de crímenes en todos los barrios de la ciudad, con zonas mayormente propensas en el oeste y la región céntrica con limites en el sur y norte.
- Una tasa baja de crímenes en los distintos lapsos de tiempo.
- Al menos 10 tipos de crímenes que superan a la media.
- Una tasa de crímenes promedio por mes en decrecimiento.

6. Conclusiones

En el presente trabajo se demostró la suma importancia de un proceso *KDD*, dado que la confiabilidad de los resultados emerge de la capacidad de crear un proceso automático y con etapas bien definidas, el cual a su vez permite solventar la mayor cantidad de problemas que surjan de las fuentes de datos de origen de una manera eficaz.

En cuanto a los objetivos principales, se demostró que uno de los factores más importantes que influye en la tasa de crímenes es la temperatura en el momento de los reportes, ya que, entre ambos existe una correlación positiva. Esto se observó tanto para las temperaturas de por si, como las estaciones y meses con mayores temperaturas promedio.

Y, por otro lado, se pudo observar que las zonas más inseguras de la ciudad son en los alrededores de los barrios McGuire y Center City East. A su vez se observó que, si se agrupan todos los reportes en dos categorías según la tasa de crímenes en sus días anteriores, para las tasas de ocurrencias bajas existe un decrecimiento anual y para las tasas altas, un crecimiento.

Sin embargo, este trabajo estuvo sujeto a ciertas limitaciones a como la falta de la población para las distintas subdivisiones geográficas de la ciudad con el objetivo de calcular la densidad poblacional (Un factor importante que permitiría comparar los resultados con otras ciudades), el desconocimiento de otros métodos para el análisis de la correlación entre la tasa criminal y la temperatura, técnicas que permitan calcular las áreas por distrito policial y, por último, una mayor capacidad de cómputo para obtener con mayor confiabilidad las aproximaciones de las áreas de los distritos policiales, las zonas de concentración de crímenes y la ejecución total de los algoritmos de clustering.

Bibliografía

Béjar, J. (2013). *K-means vs Mini Batch K-means: A comparison.*

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. En *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (págs. 226-231). Portland, Oregon: AAAI Press.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). Advances in Knowledge Discovery and Data Mining. En U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining* (págs. 1-34). Menlo Park, CA, USA: American Association for Artificial Intelligence.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques.* Amsterdam: Morgan Kaupmann.

Sculley, D. (2010). Web-scale K-means Clustering. En *Proceedings of the 19th International Conference on World Wide Web* (págs. 1177-1178). New York, NY, USA: ACM.

Apéndice

Apéndice A: Estructura de "Philadelphia Crime Data"

El dataset consta de los siguientes atributos:

Atributo	Categoría	Tipo	Faltantes
Dc_Dist	División policial	Entero	No
Psa	División policial	String	No
Dispatch_Date_Time	Temporal	String (Formato: YYYY-MM-DD hh-mm-ss)	No
Dispatch_Date	Temporal	String (Formato: YYYY-MM-DD)	No
Dispatch_Time	Temporal	String (Formato: hh-mm-ss)	No
Hour	Temporal	String (Formato: hh)	No
Dc_Key	Identificación	Entero	No
Location_Block	Localización	String	No
UCR_General	Crimen	Entero	Si
Text_General_Code	Crimen	String	Si
Police_Districts	División policial	Entero	Si
Month	Temporal	String (Formato: YYYY-MM)	No
Lon	Localización	Real	Si
Lat	Localización	Real	Si

Donde cada uno de estos representa lo siguiente:

Atributo	Descripción
Dc_Dist	Aunque se indique como un distrito, se trata de las identificaciones de los distritos policiales reales.
Psa	Contiene el código del área de servicio policial correspondiente al distrito indicado en el atributo Police_Districts.
Dispatch_Date_Time	Contiene la fecha completa del crimen reportado.
Dispatch_Date	Contiene la fecha del crimen reportado.
Dispatch_Time	Contiene la hora del día del crimen reportado.
Hour	Contiene la hora del crimen reportado.

Dc_Key	Contiene un identificador único compuesto por el año, distrito y un identificador único.
Location_Block	Contiene el nombre de la calle del crimen reportado.
UCR_General	Corresponde a una categorización de crímenes establecida por el FBI.
Text_General_Code	Corresponde a una descripción del código contenido en el atributo UCR_General.
Police_Districts	Contiene el número secuencial, pero el mismo no indica realmente el distrito policial.
Month	Contiene el mes del crimen reportado.
Lon	Contiene la longitud del crimen reportado.
Lat	Contiene la latitud del crimen reportado.