# Documento del sistema

### Restricciones del sistema

En este apartado se van a explicar las restricciones que se han tenido que tener en cuenta a la hora de realizar el análisis de los datos y la generación de los modelos finales. Estas restricciones se pueden ver como limitaciones que marcan el desarrollo del proyecto así como los resultados finales obtenidos:

* + - * Este tipo de análisis y de sistemas se suelen desarrollar con unos lenguajes de programación específicos orientados al Big Data como pueden ser R, Python o Julia. Esto requiere de conocimientos muy específicos en este área que se han tenido que adquirir. A pesar de realizar un estudio sobre este lenguaje, al no tener ninguna experiencia en este campo, no se ha podido profundizar lo deseado en esta materia.
      * Se ha utilizado el entorno de desarrollo concreto llamado **Anaconda 3** para facilitar la ejecución del código que genera las gráficas utilizadas para realizar el análisis y la generación de los modelos. Este entorno de desarrollo resulta muy útil para todo lo concerniente al desarrollo de un sistema de este tipo.
      * Se ha utilizado un ordenador con procesador AMD R5 y 16Gb de RAM. Esto resulta una limitación a la hora de ejecutar análisis complejos con un conjunto masivo de datos por lo que obliga a utilizar subconjuntos relativamente pequeños del conjunto original de datos.
      * El conjunto de datos tiene una cantidad masiva de registros con datos como Datetime que se deben transformar para poder obtener un modelo válido.

### Entorno operacional

A continuación se describe lo necesario para realizar el análisis y desarrollo de los modelos de predicción:

* + - * Ordenador con un mínimo de 8Gb de RAM y un procesador como mínimo i5 para poder soportar la carga de procesamiento a realizar.
      * Entorno de desarrollo Anaconda versión 3 que contiene todo lo necesario para trabajar sobre este proyecto.

Gráfico

Descripción generada automáticamente

* + - * Paquetes con la funcionalidad relacionada con la generación de gráficos y con funciones matemáticas necesarias para evaluar los modelos. Estos paquetes se pueden descargar de internet con la ayuda de Anaconda.

## Algoritmo KNN vecinos más cercanos

Este algoritmo se basa en la comparación de los datos más parecidos al ejemplo que se trata para determinar el valor de la clase. Se puede determinar con cuantos “vecinos” se debe comparar el ejemplo actual y saca una media de los más “cercanos”.

Se ha podido comprobar que es un algoritmo que tarda mucho en clasificar cuando el conjunto de datos es muy grande por lo que se han tenido que seleccionar fracciones del

conjunto origen para reducir los tiempos y el coste de procesamiento. Este coste excesivo de recursos se basa en que guarda todas las instancias que utiliza para realizar la predicción cuando se utilice el conjunto de test.

La influencia excesiva de los atributos irrelevantes se a podido neutralizar en el análisis inicial de los datos eliminando aquellos que no resultan útiles para la generación del modelo final.

El ruido en los datos se ha intentado neutralizar realizando pruebas con distintos valores del parámetro K (vecinos cercanos a evaluar).

Para mejorar la clasificación de este algoritmo se pueden utilizar 2 métodos:

* Edición: elimina instancias o ejemplos engañosos. Estas instancias generan ruido que perturba la clasificación provocando una clasificación errónea.
* Condensación: elimina instancias superfluas. Estas instancias no suponen una gran ayuda a la hora de realizar una buena clasificación pero suponen un aumento en el tiempo empleado y en la carga de procesamiento.

A continuación se puede ver el código necesario para generar el modelo final con el algoritmo de knn vecinos más cercanos:

1. modelKnn <- function(datos, test){
2. modelo.knn <- train.kknn(Category~., data = datos, K=15)
3. pred <- predict(modelo.knn, test)
4. # Matriz de confusión
5. mc <- table(pred,test[,1], dnn = c("Asignado","Real"))
6. # Ordenar tabla alfabéticamente
7. mc <- mc[order(rownames(mc)),order(colnames(mc))]

8. cat("\*\* Knn: \n")

1. print(mc)
2. # Aciertos en %
3. aciertosKnn <- sum(diag(mc)) / sum(mc) \* 100
4. cat("\nCorrectamente clasificados:",round(aciertosKnn,2),"%\n\n\n")

13. }

**Figura 27. Generación del modelo final con un algoritmo Knn vecinos más cercanos**

Las líneas más importantes de la generación del modelo con este algoritmo son las siguientes:

* En la línea 2 se utiliza la función “**train.kknn**” para generar el modelo. El primer parámetro representa la variable por la que se va a clasificar, el segundo parámetro representa el conjunto de datos de entrenamiento y el tercer parámetro representa el número de vecinos más cercanos que se van a tener en cuenta.
* En la línea 3 se realiza la predicción a partir del modelo generado.

## Construcción del modelo final

Se quiere construir un sistema que pueda predecir la probabilidad de que un crimen sea de una categoría determinada a partir de los datos de entrada referentes a la localización y la fecha en la que se produjo el susodicho crimen.

Para conseguir un modelo suficientemente válido se han realizado los siguientes pasos:

* Se ha dividido el conjunto de datos entre el 50% para entrenamiento y 50% para validación.
* Se han tratado los datos del conjunto de ejemplos y se han extraído nuevas variables de entrada que se han unido a las variables originales.
* Se han realizado diversos análisis de las variables de entrada frente a la variable de salida para determinar que variables son más útiles y formarán parte del modelo.
* Se han ajustado varios modelos con hora del día, día de la semana, mes, año, distrito policial y combinaciones de variables independientes.
* Se insertaron los datos de cada modelo en el conjunto de datos de entrenamiento y se comprobó su precisión en un conjunto de test.
* Se usó el algoritmo de agregación bootstrap para generar modelos de clasificación.
* Se usó un algoritmo de clasificación KNN por los vecinos más cercanos y se generaron varios modelos comparando sus resultados.
* Se usó una máquina de vectores de soporte para generar modelos de clasificación.
* Se usó el algoritmo de árboles de decisión C5.0 para generar modelos de clasificación.
* Se ha utilizado el método de Naive Bayes de clasificación para generar modelos finales.
* Se han utilizado Redes de Neuronas para generar un modelo de clasificación.
* Se comparó los resultados de cada uno de los mejores modelos generados por cada uno de los algoritmos de clasificación para obtener el mejor.

El modelo final contendrá las siguientes variables de entrada:

* Hora del día
* Día de la semana
* Mes
* Distrito policial

## Análisis de los datos

En este apartado se pretende hacer un análisis de los datos obtenidos de la página del INEI con la inclusión de gráficas que demuestren la distribución de los datos y la relación entre ellos. Se usarán muestras aleatorias debido a la cantidad de los registros y a la cantidad de atributos y de valores que estos tienen. Esto agilizará el análisis y la ejecución de las pruebas.

### Distribución de los crímenes en relación con las variables únicas

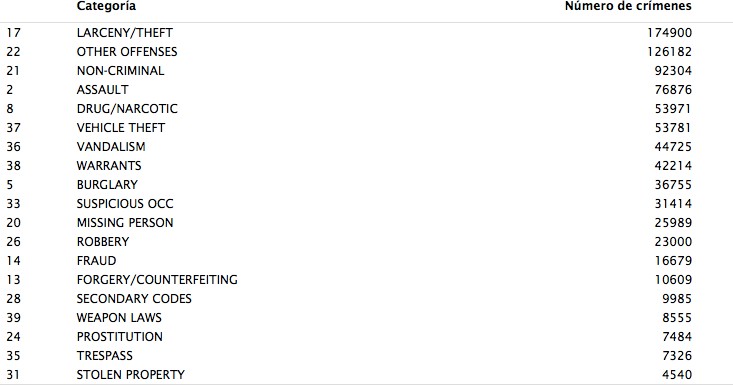
#### Variable Categoría del Crimen

En la siguiente gráfica se muestra el número de crímenes para cada categoría. En este caso se pueden observar dos detalles principales:

* + - * + Aparece una categoría, “Larceny/Theft”, en el top en cuento a número de crímenes. Se da alrededor de un 20% de los ejemplos.
        + Las 19 categorías que se dan en más crímenes aglutinan cerca del 97% de todos los ejemplos del conjunto de datos. Por lo que se ha decido eliminar los ejemplos que devuelven como variable de salida uno de los otros 20 valores.

"Top 19 crímenes"

Se puede observar, como se había expuesto, que las 19 categorías con más incidencia se dan en el 96.50% de los ejemplos y la categoría más importante se da casi el 20% de los ejemplos.



Como se ha indicado, se va a proceder a eliminar los ejemplos que contengan alguna de las 20 categorías con menos incidencia. Esto contribuirá a aligerar el análisis ya que se eliminarán la mitad de los posibles valores de salida y no tendrá un impacto demasiado grande dada la poca incidencia que tienen estas categorías.

También se ha realizado una transformación de los datos para extraer ciertos elementos de la fecha como pueden ser el año, el mes, el día del mes y la hora. El conjunto de datos ha quedado distribuido en 16 columnas:

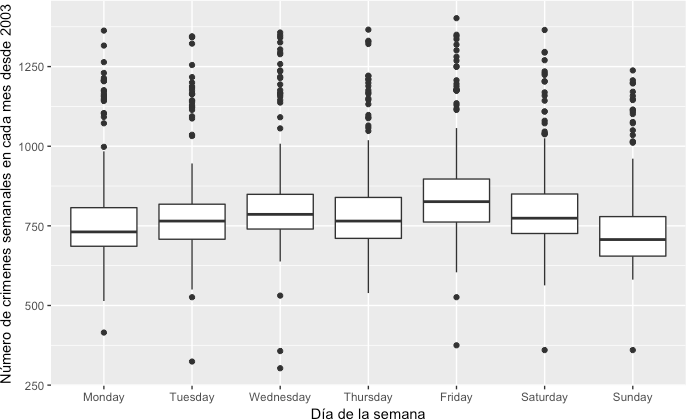
* **Dates**: Fecha en la que se produjo el crimen.
* **Category**: Categoría en la que se produjo el crimen. Se podrán observar 19 valores diferentes teniendo en cuenta que se han eliminado las 20 categorías que menos incidencia tenían.
* **Descript**: Descripción del crimen.
* **DayOfWeek**: Día de la semana en el que se produjo el crimen.
* **PdDistrict**: Distrito policial en el que se produjo el crimen.
* **Resolution**: Resultado del crimen. Arrestado o no arrestado.
* **Address**: Dirección donde se desarrollo el crimen.
* **X**: Longitud de la localización del crimen.
* **Y**: Latitud de la localización del crimen.
* **Years**: Año en el que ocurrió el crimen.
* **Month**: Mes en el que ocurrió el crimen expresado en números.
* **DayOfMonth**: Día del mes en el que ocurrió el crimen.
* **Hour**: Hora del día del incidente.
* **YearsMo**: Combinación del año y el mes.
* **Weekday**: Variable que indica si el día de la semana en el que ocurrió el crimen fue fin de semana o no.
* **AddressType**: Variable que indica si el crimen ocurrió en una intersección o en otra localización.

#### Variables Relacionadas con el Tiempo

Entre la lista de variables de entrada que contienen los ejemplos se encuentran algunas variables relacionadas con el momento en el que se produjo el crimen. Se presenta la fecha completa del crimen y el día de la semana. Pero se considera que la fecha completa es un concepto demasiado amplio y se han extraído variables que aclaran un poco más el análisis, como se ha explicado en el punto anterior.

* + - * 1. Variable Día de la Semana

Se va a mostrar la relación entre el número de crímenes y el día de la semana en el que fueron cometidos.

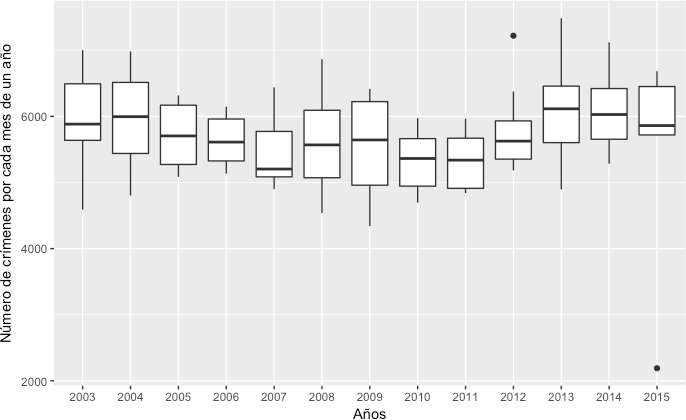


Se puede observar como los Miércoles y los Viernes se produce un repunte de los crímenes y los Domingos se produce un descenso del número de crímenes.

* + - * 1. Variable Año del Crimen

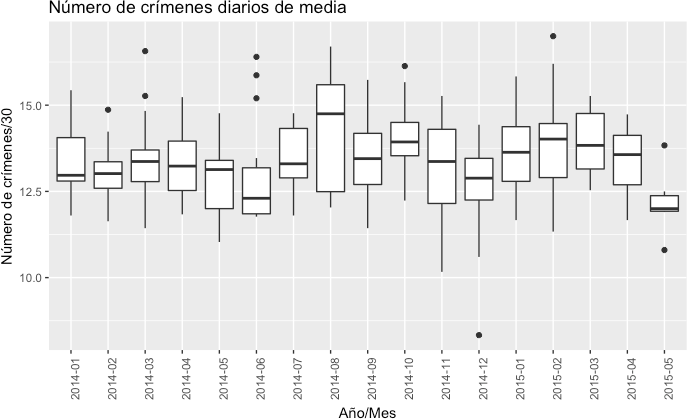
Se ha analizado la distribución de los crímenes a través de los años de los que se ha sacado la muestra utilizada. Se puede observar una caída en 2017, lo que es debido a que no se tienen datos de todo el año si no solamente hasta Mayo. Por otro lado se puede ver un descenso en el

número de incidentes desde 2003 hasta 2010 y un posterior incremento a partir de este año. Se ha realizado este análisis para comprobar la progresión de los crímenes a lo largo de los años y lo que se puede esperar pero la variable “**Years**” no se incluirá en el modelo final.



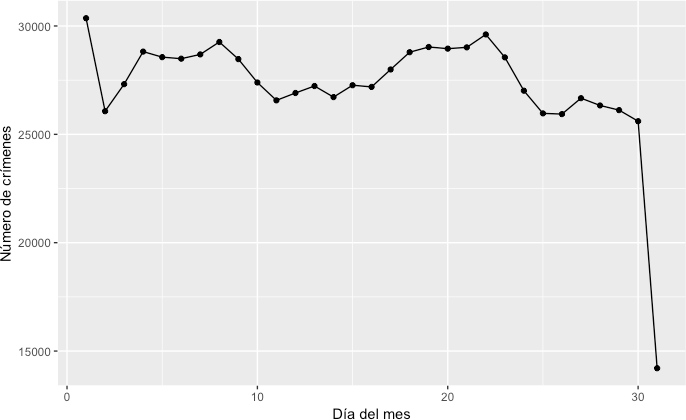
* + - * 1. Variable combinación Año-Mes

Se ha realizado un análisis para comprobar si se tienen datos de todo el año 2015. Para ello se ha creado un gráfico que muestra la relación entre el número de crímenes por mes del año desde 2014. Se puede ver que en Mayo de 2015 la muestra se reduce de forma drástica y no aparecen más datos a partir de este momento, lo que quiere decir que sólo se tienen datos hasta ese mes de ese año. Aunque se comprueba que se tienen datos hasta Mayo de 2015, este mes no se incluirá en el modelo final por el escaso conjunto de datos que contiene.



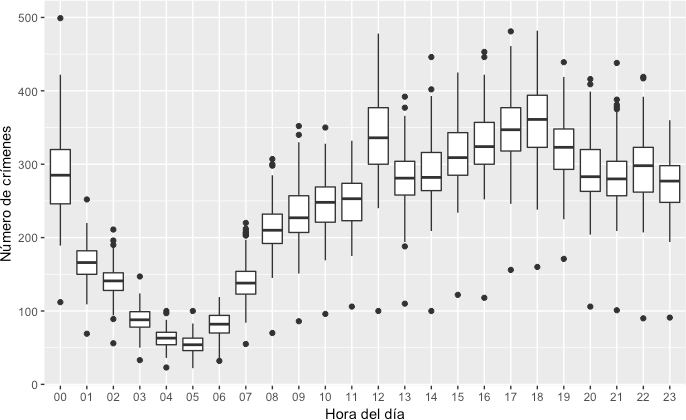
* + - * 1. Variable Día del Mes

Lo primero que salta a la vista es que el día 31 del mes se produce un descenso muy acusado del número de crímenes. Esto se debe a que durante el año solo existen 7 meses con 31 días por lo que la incidencia ese día es mucho menor. También se puede ver que el día 1 del mes se producen una mayor cantidad de crímenes que en el resto de los días del mes. Esto se puede deber a que cuando no se conoce el día exacto del crimen, por defecto, se establece el día 1 como la fecha del crimen. Si se excluyen los días 1 y 31 del mes se puede ver una mayor incidencia entre los días 4 y 9, y entre los días 18 y 23.



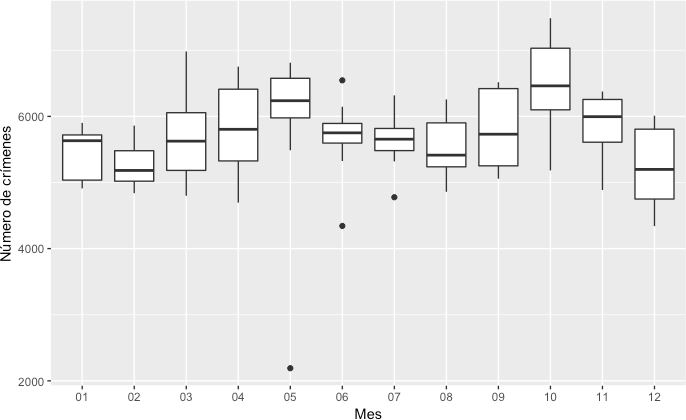
* + - * 1. Variable Hora del Día

Se puede observar que el punto más bajo a lo largo de todo el día son las 5 de la mañana. En general las horas de la madrugada, de 1 de la mañana a 7 de la mañana, son las horas donde se producen menos incidencias. Por otro lado, por la tarde es cuando aparece la mayor cantidad de crímenes, entre las 12-13 y las 22.



* + - * 1. Variable Mes del Año

Se pueden observar crecimientos de las incidencias en los meses de Mayo y Octubre y un descenso importante en Diciembre.



#### Resultados iniciales de análisis por variable única

Se ha realizado este análisis del número de crímenes por variable única o variable independiente para tratar de predecir la categoría de un crimen a partir de la fecha y la localización en el que se produjo tratando cada variable de entrada de forma independiente y acumulando los resultados de cada una, por lo tanto, el elemento más importante en este sistema será la categoría del crimen y los dos elementos secundarios más importantes serán la fecha y la localización del crimen, y el modelo a desarrollar deberá predecir la categoría del crimen a partir de su fecha y su localización.

Cabe destacar la importancia que pueden tener las variables que se han extraído de la variable “**Dates**” de las que se han podido obtener las siguientes conclusiones:

* + - * + El número de crímenes descienden de forma acusada entre la 1 y las 7 de la mañana.
        + La criminalidad es menor en Diciembre y aumenta en Octubre y Mayo.
        + La mayoría de los crímenes pertenecen a la categoría “**Larceny/Theft**”.
        + Los distritos de “**Southern**” y “**Mission**” tienen los índices de criminalidad más altos.

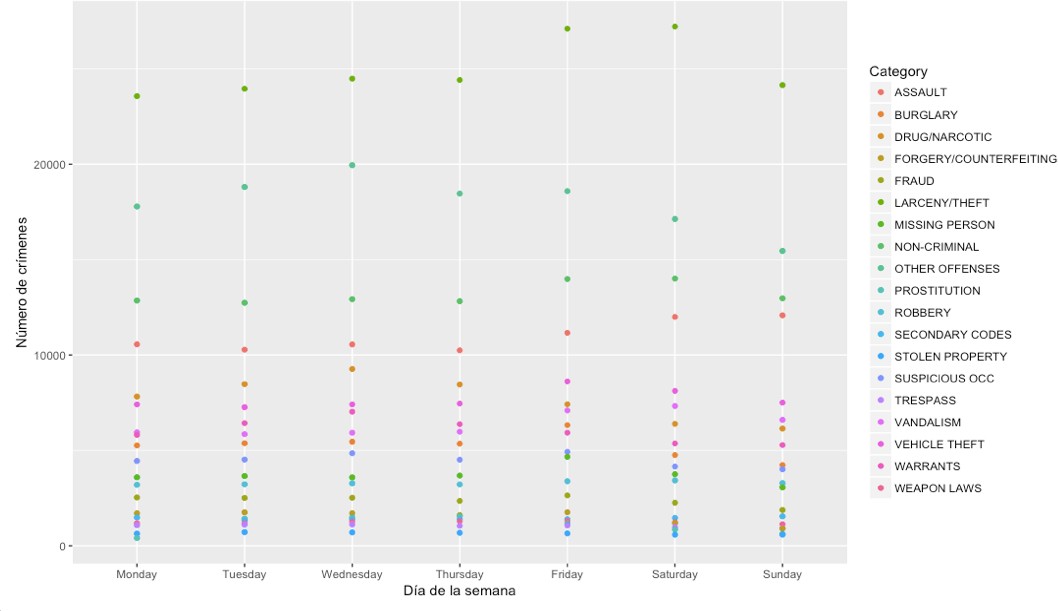
Como se ha explicado antes se han extraído nuevas variables de las ya existentes que han podido tener efectos importantes en la tendencia de los datos dados. Estas nuevas variables serán importantes en el análisis de las muestras y simplificarán el análisis.

### Distribución de los crímenes en una relación bivariante

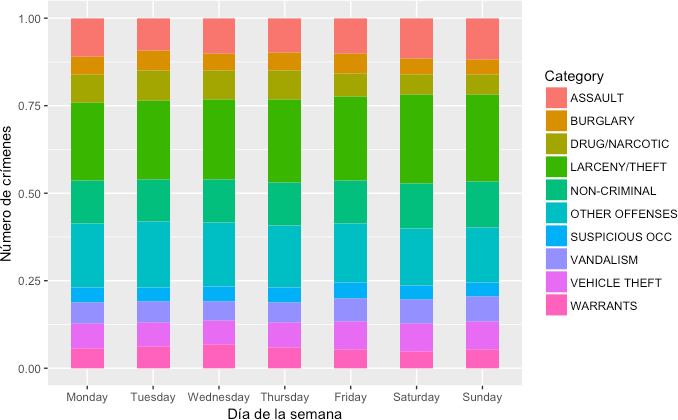
En esta parte se realizará una comparación de varias variables, 2 o más variables frente a 1. Esto servirá de ayuda para determinar la relación entre variables. En concreto se comparará la categoría del crimen, que es la variable cuyo valor se pretende predecir, frente a otras variables para determinar cuales de ellas tienen mayor impacto en su resultado.

#### Categoría del crimen por día de la semana

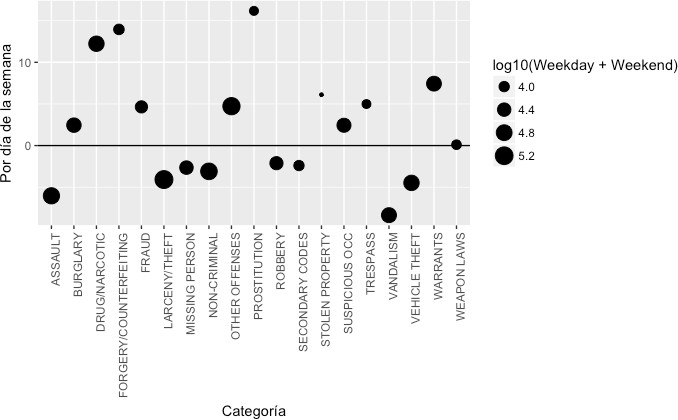
Se va a establecer la relación entre la categoría del crimen y el día de la semana de forma gráfica. Inicialmente se mostrará el total de crímenes por día de la semana dependiendo de la categoría.



Para comprobar la relación entre el día de la semana y la categoría del crimen, se van a fraccionar los valores de la variable categoría utilizando solo los 10 más importantes. Se puede ver una tendencia dependiendo de los días de la semana. En concreto se puede observar que para el valor “**Larceny/Theft**”, en los fines de semana, el número de incidencias aumenta mucho.



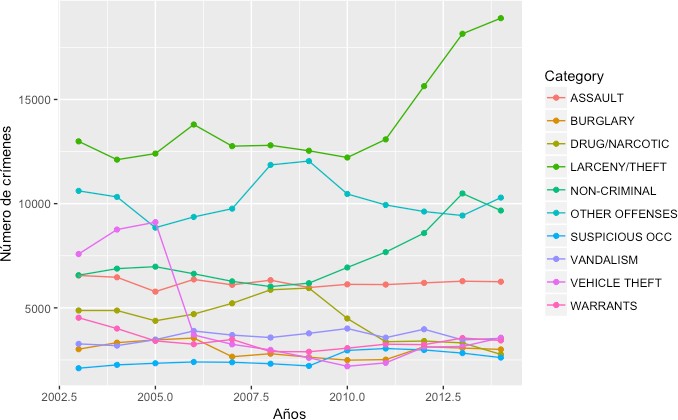
Las gráficas anteriores indican que se produce un incremento en el ratio de crímenes en los fines de semana. Por este motivo se ha considerado dividir los días en dos conjuntos, los días de fin de semana (de Viernes a Sábado) y los días de diario (de Lunes a Jueves). Se ha obtenido la media de los crímenes dividiendo el total de crímenes por el número de días de la variable “**weekday**” (4 para los días de diario y 3 para los fines de semana). Entonces, se calcula el porcentaje de la diferencia de la media del número de crímenes. Un valor positivo indica una mayor incidencia de los crímenes en los días de diario y un valor negativo indica una incidencia menor en los días de diario. Se puede observar una gran influencia de la variable “**weekday**” en el número de incidencias.



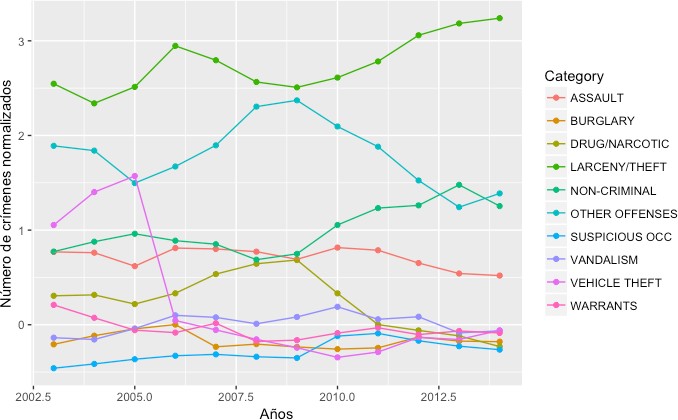
#### Categoría del crimen por año

Se va a realizar un análisis de las incidencias a través de los Años y por Categoría. Se utilizarán los datos desde 2003 y hasta 2014 ya que no se tiene un conjunto completo de datos de 2015.

Como se puede ver en la gráfica siguiente se produce un ascenso de las categorías “**Larceny/Theft**” y “**Non-Criminal**”. También se puede observar una clara caída en la categoría “**Vehicle Theft**” en 2006.

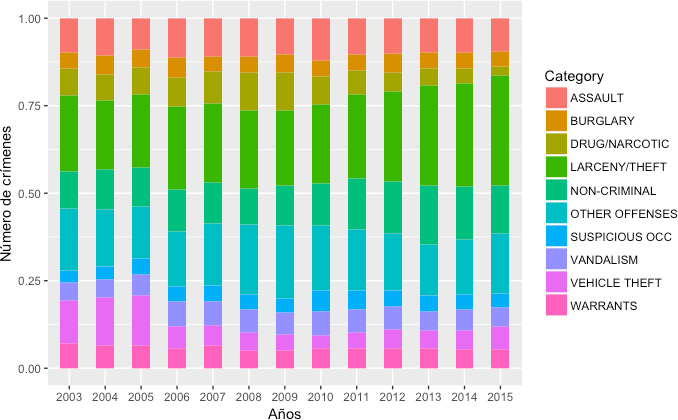


Para intentar extraer información más útil se ha normalizado el número de crímenes, lo que se puede ver en la siguiente gráfica, restando al número de crímenes la media de cada año y dividiéndolo por la desviación estándar (raíz cuadrada de la varianza (la media aritmética del cuadrado de las desviaciones de una variable con respecto a su media)).



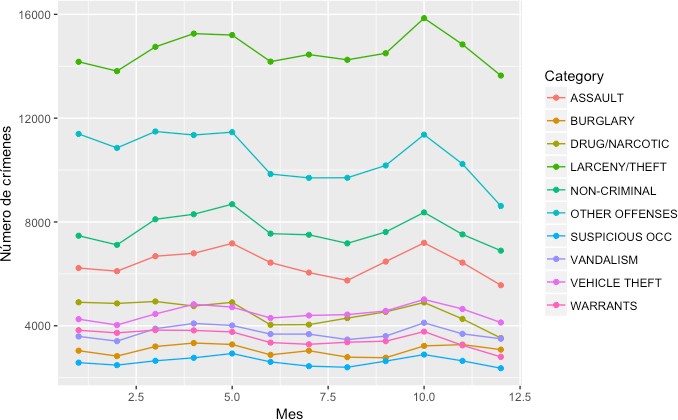
La gráfica que sigue a este texto representa la proporción de crímenes de cada categoría por el total de los crímenes del año. Esto nos puede dar una idea de la dispersión de cada número de crímenes a lo largo de los años y su tendencia en la muestra. En este caso se puede ver un ascenso de las categorías “**Larceny/Theft**” y “**Non-Criminal**” desde 2010. Se puede observar un descenso en las categorías “**Vehicle Thefts**” y “**Drug/Narcotics**”. Como se está explicando y se muestra en estos gráficos, la distribución de los crímenes cambia dependiendo de los años.

Se puede determinar que no serían demasiado útiles datos sobre crímenes de años muy alejados de la fecha actual por lo tanto se usarán datos de crímenes de la muestra a partir de 2012. Este análisis no ha conseguido encontrar una tendencia real y consistente de las distintas categorías a través de los años.

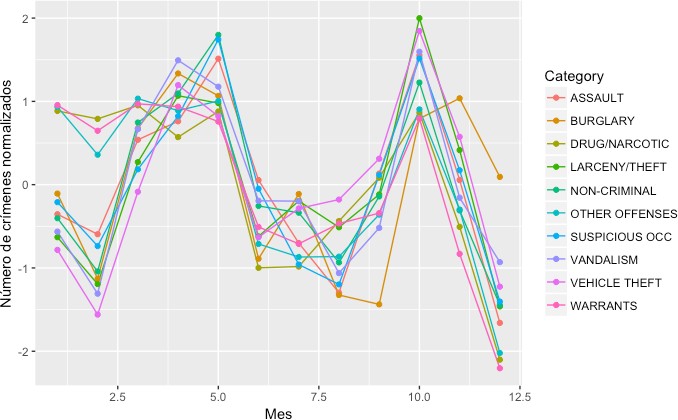


#### Categoría del crimen por mes

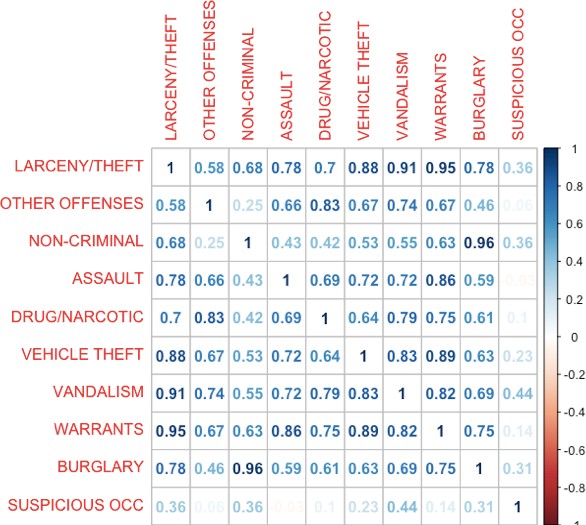
En la siguiente gráfica se puede ver la relación entre la categoría de los crímenes y el mes en el que se desarrollaron junto a su tendencia en cuanto a número de crímenes. Se puede observar una tendencia en el número de crímenes a lo largo de los meses. Esta tendencia se puede ver de forma más clara en las categorías “**Assault**”, “**Non-Criminal**”, “**Other-Offenses**” y “**Larceny/Theft**”.



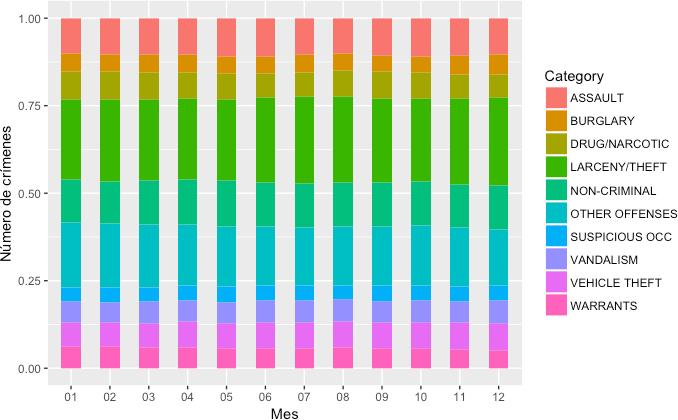
A continuación, se normaliza el número de crímenes en cada mes. Después de esta operación se puede observar una clara dependencia con el mes.



La matriz a continuación muestra una fuerte correlación entre las categorías de los crímenes a través de los meses.

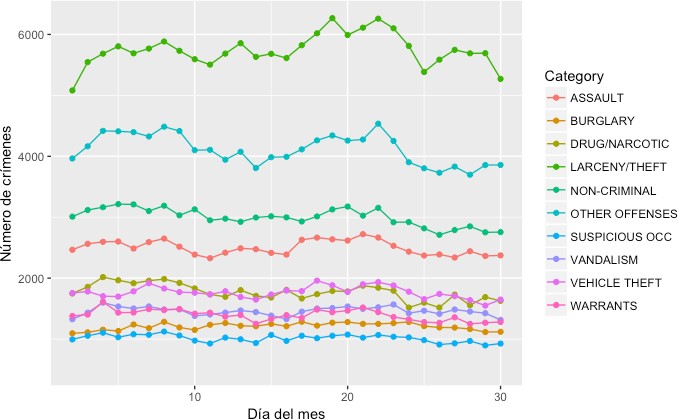


Por lo tanto, podría ser posible eliminar el mes de la predicción final. A continuación se muestra una gráfica que pone de manifiesto que las proporciones relativas por meses no sufren a penas cambios y el efecto del mes en la categoría del crimen es demasiado débil para utilizarlo en la predicción, por lo que se eliminará del modelo de predicción.

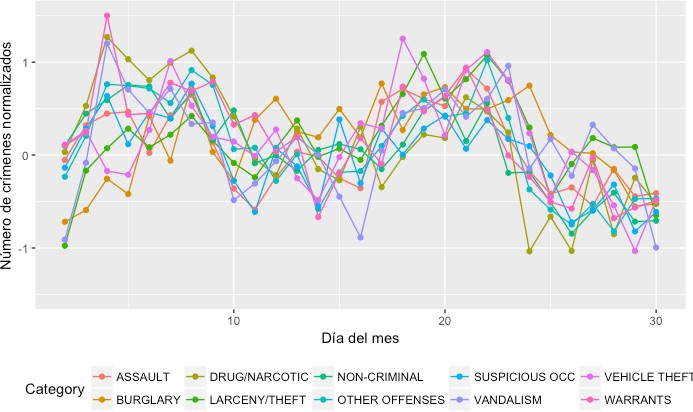


#### Categoría del crimen por día del mes

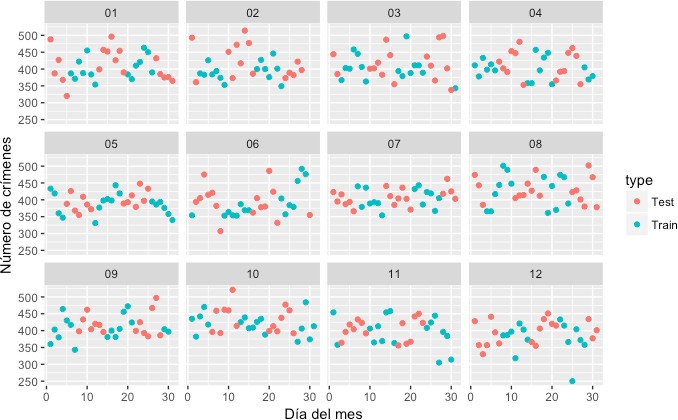
En la siguiente gráfica se puede ver la variación de los crímenes por día de la semana y por categoría del crimen. En este caso no se ha incluido el día 31 porque sólo hay 7 meses con 31 días y tampoco se ha incluido el día 1 porque este día ha sido usado como día por defecto cuando no se conoce la fecha exacta del crimen. La gráfica muestra una tendencia en el número de crímenes a lo largo de los días del mes.



Para poder visualizar de forma más clara esta tendencia, se han normalizado los valores. Se puede ver que de los días 10 al 16 y de los días 25 al 30 tienen una menor incidencia que de los días 5 al 10 y de los días 15 al 20.

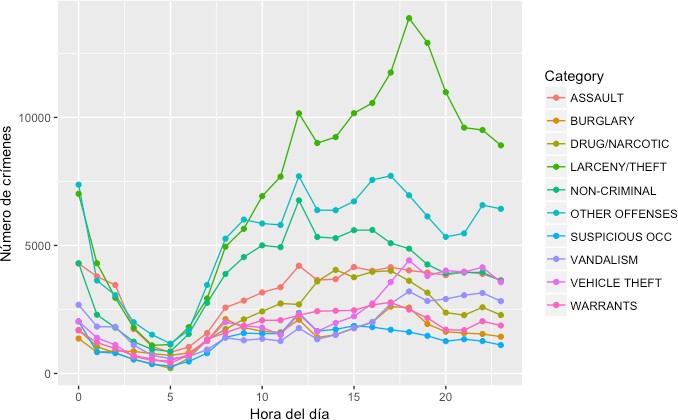


Para comprobar esta tendencia y poder sacar una conclusión, se han comparado los datos del conjunto de test con los datos del conjunto de entrenamiento desde 2017 y a través de los 12 meses. En esta gráfica se puede comprobar que se han usado datos de diferentes semanas para formar los conjuntos de test y de entrenamiento por lo que no tendría forma de validar los resultados obtenidos en el entrenamiento. Por este motivo no se usará el día del mes en el modelo a desarrollar.

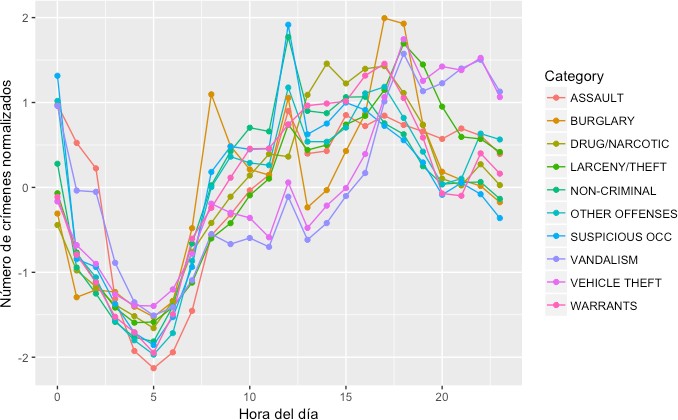


#### Categoría del crimen por hora

En la siguiente gráfica se puede observar la variación del número de crímenes a través de las diferentes horas del día y por categoría. Se observa la mayoría de la actividad desde las 10 de la mañana hasta la media noche y el punto más bajo alrededor de las 5 de la mañana.



Si se normaliza la muestra se observa la misma tendencia que en el gráfico anterior.



Queda de manifiesto que la hora del día afecta de manera determinante a los crímenes por lo que esta variable se incluirá en el modelo de predicción.

### Pruebas con el algoritmo Knn de vecinos más cercanos

Se han realizado varias pruebas modificando el parámetro K (número de vecinos a comparar) y el porcentaje de la muestra completa que se ha utilizado. Los resultados han sido los siguientes:

* + - * Como primer paso se realizaron pruebas con todos los elementos del conjunto de datos y con un K por defecto de 1. La prueba no se pudo terminar con éxito ya que el número de elementos era demasiado elevado y el tiempo de procesamiento necesario era impracticable.
      * Para la primera prueba se configuraron los parámetros de la siguiente forma:
        + Se obtuvo un subconjunto de datos del 20% de la muestra inicial.
        + Se estableció el valor de K en 20.

El porcentaje de acierto fue demasiado bajo para lo esperado. Posiblemente la muestra de los datos seleccionada no tenía suficiente información para generar un modelo válido.

* + - * Para la segunda prueba se configuraron los parámetros de la siguiente forma:
        + Se obtuvo un subconjunto de datos del 50% de la muestra completa.
        + Se estableció el valor de K en 5.

El porcentaje de éxito aumentó de forma visible con un mínimo aumento en el consumo de recursos.

* + - * Para la tercera prueba se configuraron los parámetros de la siguiente forma:
        + Se obtuvo un subconjunto de datos del 25% de la muestra original.
        + Se estableció el valor de K en 15.

El porcentaje de aciertos disminuyó con respecto a la segunda prueba pero se mantuvo por encima de la primera prueba. Parece ser que el porcentaje de la muestra utilizado tiene mucho que ver con los resultados de las predicciones.

* + - * Para la cuarta prueba se configuraron los parámetros de la siguiente forma:
        + Se obtuvo un subconjunto de datos del 20% de la muestra original.
        + Se estableció el valor de K en 10.

El porcentaje de aciertos es similar al modelo de la tercera prueba. Es posible que el valor de K sea demasiado elevado y sea mejor un valor más bajo.

* + - * Para la quinta prueba se configuraron los parámetros de la siguiente forma:
        + Se obtuvo un subconjunto de datos del 25% de la muestra original.
        + Se estableció el valor de K en 7.

El porcentaje de aciertos mejora muy ligeramente con respecto a la cuarta prueba pero se queda por debajo de la segunda prueba.

* + - * Para la sexta prueba se configuraron los parámetros de la siguiente forma:
        + Se obtuvo un subconjunto de datos del 30% de la muestra original.
        + Se estableció el valor de K en 20.

El porcentaje de aciertos es prácticamente igual que el de la tercera prueba. Esto indica que es importante aumentar el valor de K pero es aun más importante, en este caso, utilizar una muestra amplía de datos.

Se ha llegado a la conclusión de que el algoritmo de los K vecinos más cercanos no es el mejor para el problema que nos ocupa. No se ha conseguido superar los resultados de otros algoritmos como el de SVM incluyendo una muestra mayor de los datos si bien el tiempo de procesamiento ha sido menor.