ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA
DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA EN SISTEMAS
AUDIOVISUALES Y MULTIMEDIA

TRABAJO FIN DE GRADO

Análisis de la taquilla del cine español mediante la aplicación de técnicas de Data Science

Autor: Jorge Rodríguez Barrios

Tutor: Dr. Felipe Alonso Atienza

Cotutor: Dr. Rafael Linares Palomar

Curso académico 2014 / 2015

Resumen

El análisis de datos es un área en auge en el que a través de un proceso de inspección y transformación de los mismos, se permite dar valor añadido a productos y dar apoyo en la toma de decisiones.

Este proyecto trata sobre un estudio de la taquilla de los cines españoles mediante técnicas de *Data Science*. Para ello se han analizado los datos recopilados por la consultora Rentrak sobre los ingresos generados por cada película en proyección en cada determinada semana, así como otros datos de interés tales como número de cines que proyectan dicha película o el número de copias adquirido por todos los diferentes cines. Estos ficheros de datos han sido facilitados por la Biblioteca de la Universidad y posteriormente han sido analizados para extraer conclusiones, localizar patrones, realizar un análisis retrospectivo y tratar de llevar a cabo también uno predictivo.

El procesado y análisis de datos se ha realizado con R, un lenguaje y entorno de programación muy popular en análisis estadístico y gráfico. Este lenguaje proporciona una estructura de datos, los *DataFrames*, que permite manejar en forma de matriz listas de vectores de igual longitud.

De cara a disponer de una base de datos más completa, se ha ampliado la información semanal de cada película que ofrecen los ficheros de Rentrak con datos extraídos de IMDB (*Internet Movie DataBase*). Estos nuevos datos cubren aspectos como el género de una película o el país de procedencia de la misma y han sido incorporados a través de consultas a la API (*Application Program Interface*) de OMDB (*Open Movie Data Base*).

Después del procesado de datos, se han realizado análisis que abarcan desde los ingresos por determinados géneros o distribuidoras a la relación entre géneros y procedencia de las películas con su fecha de estreno. El atractivo de este análisis reside en la visión de mercado que pueda a ofrecer a una distribuidora que quiera estrenar una película con unas determinadas características y ayudar así a poder obtener los máximos beneficios con dicho estreno.

Acrónimos y siglas

API …….. Application Program Interface.

IMDB ……... Internet Movie DataBase.

JSON ……...….. JavaScript Object notation.

N/A ……. Not Available.

OMDB ……..……. Open Movie DataBase.

XML ……. Extensible Markup Language.

Índice

Capítulo 1 - Introducción y objetivos

1.1 Antecedentes y estado del arte

Este proyecto se enmarca dentro del contexto del *Big Data* y de *Data Science*. Estos términos emergen para designar la nueva profesión que se encarga de tomar conciencia de las cantidades ingentes de datos y crear significado y valor sobre ellos. Para realizar un buen trabajo en el área de *Data Science* [1] tal como muestra el diagrama de Venn de la figura 1.1.1, son necesarios simultáneamente conocimientos de matemáticas y estadística, experiencia en entornos y lenguajes de programación y un conocimiento práctico del área al que pertenecen los datos que se están estudiando.

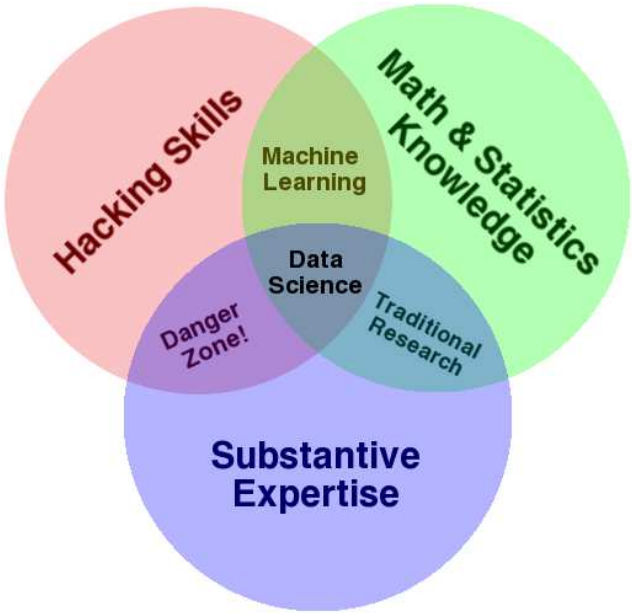


Figura 1.1.1 – Diagrama de Venn.

En concreto, empleando técnicas de *Data Science* se realizará un estudio sobre la industria cinematográfica española. A través de analizar patrones en los hábitos de consumo, se pretende constatar con datos cuál será la directriz que llevará a un proyecto a alcanzar el nivel más alto de éxito. Las industrias dedicadas al entretenimiento y cultura, como es la del cine [2], abarcan un rango muy amplio de consumidores con unas necesidades específicas variantes a lo largo del año. Es por este motivo, por el que un conocimiento de la evolución de dichas necesidades, se convierte en un elemento clave para definir el camino que deberá tomar un determinado proyecto.

Los motivos que impulsan este estudio son muy variados y las ideas a tratar se pueden resumir en las siguientes:

- Se analizará como afecta a una película la fecha de estreno en función de su país de procedencia y de su género.
- Estudio sobre las distribuidoras que más impacto tienen en las taquillas en lo que respecta a términos de recaudación.
- La relación entre los ingresos generados en taquilla la semana del estreno con la recaudación total del largometraje.
- Estimación de los posibles ingresos en base al género, distribuidora, país de procedencia y mes de estreno de una producción.

1.2 Objetivos del proyecto

El objetivo general de este proyecto consiste en ofrecer un análisis retrospectivo y mostrar un modelo predictivo que se conseguirá tras analizar los datos recopilados semanalmente por la consultora Rentrak sobre la repercusión que ha tenido cada película en proyección dicha semana en las taquillas de los cines en España.

La realización de este análisis tiene como fin la adquisición del conocimiento de cómo afectan las distintas variables que forman el contexto de una película, a saber: género, distribuidora, mes de estreno y país de procedencia como variables categóricas y en el caso de las numéricas, total de cines en los que se emite dicha producción, número de copias adquiridas por cada cine y datos sobre ingresos totales y acumulados y número de espectadores por semana.

Analizando retrospectivamente estos resultados, no sólo se podrá comprender como repercuten las variables mencionadas en aspectos tales como ingreso total generado por una película, sino que también se podrá obtener la información necesaria para evaluar qué directrices debe tomar un proyecto para que este alcance los mayores beneficios.

Para abordar el problema y realizar el análisis, se ha subdividido el objetivo principal en problemas más pequeños, cuya resolución por separado y de manera secuencial contribuye a la consecución del objetivo general:

1. Limpieza de datos:

- Los ficheros Excel de Rentrak no están listos para su procesado, por tanto en este apartado se trata la necesidad de un pre-procesado de los datos.

2. Procesado de datos:

- La entrada de este apartado es la salida del apartado anterior. Se cubren aspectos tales como la elección de la estructura de datos que contendrá toda la información de la taquilla por cada película.

3. Información adicional:

- Para un análisis más completo, en esta etapa se añade a la información disponible una serie de metadatos extraídos de la API de OMDB.

4. Simplificación de variables categóricas:

- Agrupamiento en variables de carácter más amplio de las categorías que por su especificación forman grupos reducidos que no son aptos para el estudio.

5. Representación de resultados:

- Mediante gráficas que faciliten la visualización, se generarán representaciones que ayuden a extraer los diversos resultados y dar respuesta a las preguntas que se planteen.

6. Modelo predictivo:

- A partir de la información recopilada durante los últimos años, se analizarán las variables que forman el contexto de una película con el fin de ver la relación que mantienen con los ingresos generados y crear así un modelo de predicción.

Para el desarrollo de los diversos objetivos propuestos, el presente proyecto se estructura en cuatro capítulos:

- El primero es una breve introducción que trata de situar en contexto y familiarizar al lector con los objetivos que se persiguen.

- El segundo capítulo contiene la parte técnica. Se explicarán los métodos empleados para que a través de unos datos de entrada, pueda crearse una base de datos que permita realizar un análisis con el fin de percibir hábitos o patrones y en base a estos tomar decisiones.

- En el tercer capítulo se emplearán diversas técnicas para poder visualizar los resultados que se han extraído en el capítulo dos y se analizarán las conclusiones a partir de las representaciones.

- Finalmente, en el capítulo cuatro se enunciarán las conclusiones generales del proyecto, proponiendo líneas futuras de investigación y señalando las competencias adquiridas durante el desarrollo del trabajo.

Capítulo 2 - Material y métodos

En este capítulo se realizará todo el proceso relacionado con el tratamiento de datos. Se partirá de la información original suministrada por la Biblioteca de la Universidad y se detallará cada procedimiento necesario hasta alcanzar una base de datos que permita un análisis de calidad. Se dividirá el capítulo en cuatro apartados.

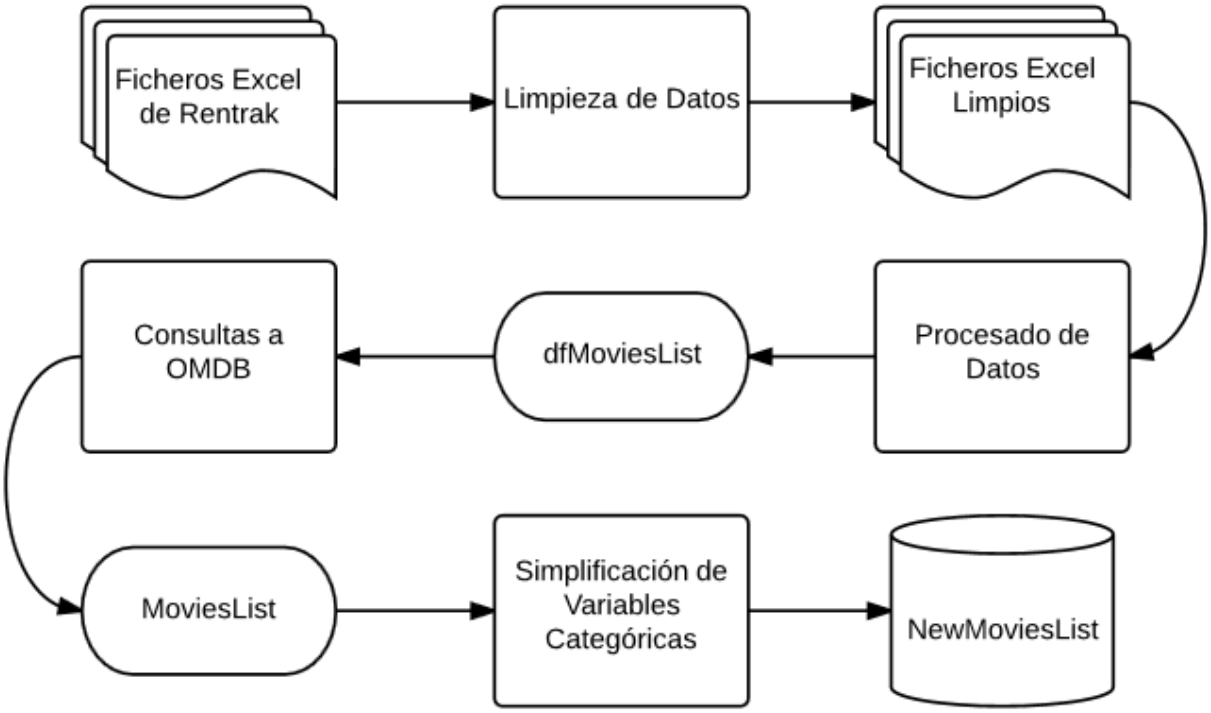


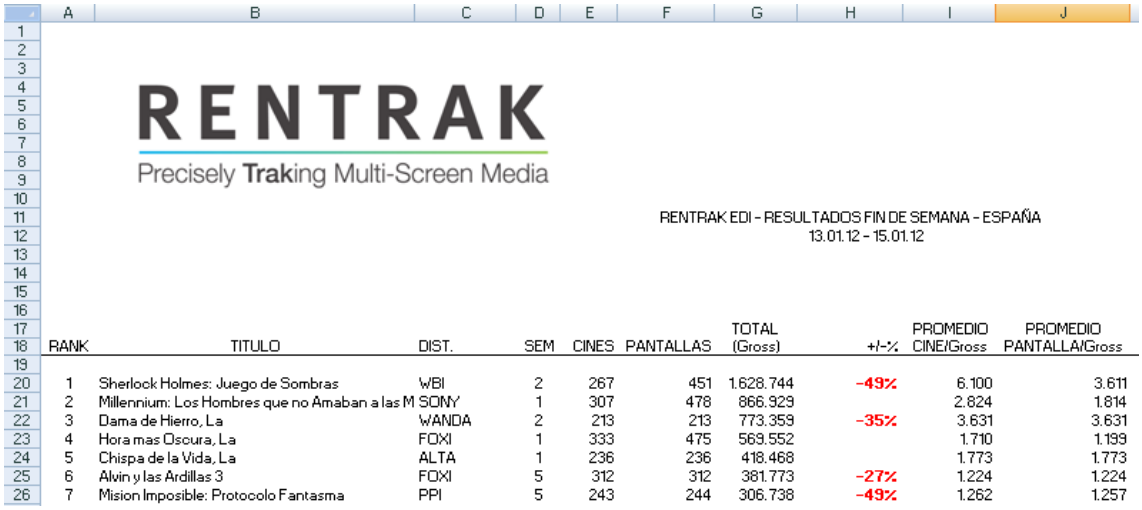
Figura 2.0.1 Diagrama de flujo.

Siguiendo el diagrama de flujo de la figura 2.0.1:

1. Inicialmente se eliminará la información no relevante de los ficheros Excel de Rentrak.
2. Los datos de interés que proporciona el proceso anterior se tratarán para obtener la primera aproximación a una base de datos de las películas disponibles.
3. La información disponible se verá ampliada con metadatos mediante peticiones a OMDB.
4. Finalmente se concluye la base de datos reorganizando la información capturada en el apartado previo.

2.1 Limpieza de datos

Los ficheros Excel que serán la entrada al sistema, inicialmente presentan unas características que no son óptimas para iniciar el análisis. Tal como se aprecia en la figura 2.1.1, el fichero muestra información sobre la consultora que recoge los datos, una imagen y numerosas celdas vacías de valor, que cuando sean leídas por la rutina de R se convertirán en valores N/A.



RENTRAK

Precisely Traking Multi-Screen Media

RENTRAK EDI - RESULTADOS FIN DE SEMANA - ESPAÑA
13.01.12 - 15.01.12

RANK	TITULO	DIST.	SEM	CINES	PANTALLAS	TOTAL (Gross)	+/-%	PROMEDIO CINE/Gross	PROMEDIO PANTALLA/Gross
1	Sherlock Holmes: Juego de Sombras	WBI	2	267	451	1.628.744	-49%	6.100	3.611
2	Millennium: Los Hombres que no Amaban a las M	SONY	1	307	478	866.929		2.824	1.814
3	Dama de Hierro, La	WANDA	2	213	213	773.359	-35%	3.631	3.631
4	Hora mas Oscura, La	FOXI	1	333	475	569.552		1.710	1.199
5	Chispa de la Vida, La	ALTA	1	236	236	418.468		1.773	1.773
6	Alvin y las Ardillas 3	FOXI	5	312	312	381.773	-27%	1.224	1.224
7	Mision Imposible: Protocolo Fantasma	PPI	5	243	244	306.738	-49%	1.262	1.257

Figura 2.1.1 Ejemplo de fichero Excel de Rentrak.

Por tanto el objetivo de esta primera etapa es lograr una transformación de estos ficheros con el fin de que únicamente contengan los valores correspondientes a la información semanal.

Como se había adelantado en la introducción, el lenguaje de programación con el que se desarrollará el proyecto será R [3]. Una de las ventajas que ofrece, son las numerosas y útiles librerías que hay disponibles. Para este paso se cargará la librería ‘xlsx’ que permite trabajar con ficheros Excel.

Para realizar el estudio se disponen de los ficheros Excel correspondientes a todas las semanas desde inicios de 2012 hasta mediados de 2015, por tanto es necesario emplear una rutina (de aquí en adelante script) que procese todos los ficheros automáticamente.

Después de cargar la librería oportuna y elegido el directorio que contenga los Excel de Rentrak, se añaden todos automáticamente a la lista ‘files’, para posteriormente y mediante la función ‘read.xlsx’, cargar todo su contenido en un *DataFrame* por cada fichero.

```
library(xlsx)

files = list.files(pattern='[.]xlsx')

dfList <- list()

for (i in 1:length(files)){
  dfList[[paste0("excel",i)]] <- read.xlsx(files[i],
  header=T,stringsAsFactors=FALSE,sheetIndex = 1)
}
```

El fragmento de código anterior genera una lista de *DataFrames*, dónde cada uno de estos últimos se corresponde con cada uno de los ficheros Rentrak. Un *DataFrame* [4] es una matriz, donde las columnas son vectores de igual longitud y pueden ser de diferentes tipos de datos. El atractivo de usar este tipo de estructura que proporciona R, es que la información semanal de una película presentará siempre los mismos campos, y se verá ampliada con nuevas filas, conforme transcurra su estancia en proyección y cabe mencionar también la facilidad para poder acceder o filtrar valores de los *DataFrames* aprovechando la potencia de R. Se procede a eliminar todo lo que no sea información esencial:

```
for (i in 1:length(files)){

  dfList[[i]] <- dfList[[i]][ -c(8,12) ]
  dfList[[i]] <- dfList[[i]][ -c(15:17) ]

  row.has.na <- apply(dfList[[i]], 1, function(x){any(is.na(x))})
  dfList[[i]] <- dfList[[i]][!row.has.na,]

  names(dfList[[i]]) <- dfList[[i]][1,]
  dfList[[i]] = dfList[[i]][-1,]

  names(dfList[[i]])[names(dfList[[i]]) ==
                       '(Gross)'] <- 'GrossTotal'
  names(dfList[[i]])[names(dfList[[i]]) ==
                       'PANTALLA/Adm'] <- 'AdmPromPerPANTALLA'
  names(dfList[[i]])[names(dfList[[i]]) ==
                       'ACUM'] <- 'EurosAcum'

  rownames(dfList[[i]]) <- NULL
  dfList[[i]] = dfList[[i]][,-1]

  dfList[[i]] <- data.frame("TituloOriginal" = c("-"), dfList[[i]])
  dfList[[i]] <- data.frame("Year" = c("-"), dfList[[i]])

  write.xlsx(dfList[[i]],paste0("excel",i,".xlsx"),
             sheetName = "sheet1", row.names = FALSE)

}
```

Ejecutando estas instrucciones:

- Se eliminan las columnas que no son de interés.
- Se eliminan todas las filas que contengan valores N/A. Esta son las que están por encima y por debajo de los datos de interés.
- Se modifican los nombres de las columnas. Necesario para poder acceder luego a cada valor específico, ya que sería imposible si el nombre contuviese un espacio o un '/' (como en el caso de 'PANTALLA/Adm').
- Se añaden las columnas 'Year' y 'TituloOriginal', que inicialmente no las incluye Rentrak, y serán de interés para análisis futuros.
- Se escriben los *DataFrames* procesados en nuevos ficheros .xlsx.

Como salida del script, se generan en el directorio de trabajo tantos ficheros Excel como se introdujeron a la entrada del sistema, con la gran diferencia de que estos nuevos ficheros están preparados para iniciar el procesado de datos. Por tanto, la salida de esta primera etapa, será la entrada de la siguiente. En la figura 2.1.2 se muestra el resultado de realizar el proceso de limpieza de datos sobre el fichero mostrado en la figura 2.1.1.

	Titulo	TituloOriginal	Year	DIST.	SEM	CINES	PANTALLAS	GrossTotal	GrossPromPerCINE
1	Sherlock Holmes: Juego de Sombras	-	2012	WBI	2	267	451	1628744	6100.16479400749
2	Millennium: Los Hombres que no Amaban a las Mujeres	-	2012	SONY	1	307	478	866929	2823.87296416938
3	Dama de Hierro, La	-	2012	WANDA	2	213	213	773359	3630.79342723005
4	Hora mas Oscura, La	-	2012	FOXI	1	333	475	569552	1710.36636636637
5	Chispa de la Vida, La	-	2012	ALTA	1	236	236	418468	1773.16949152542
6	Alvin y las Ardillas 3	-	2012	FOXI	5	312	312	381773	1223.63141025641
7	Mision Imposible: Protocolo Fantasma	-	2012	PPI	5	243	244	306738	1262.2962962963

Figura 2.1.2 Ejemplo de fichero procesado.

2.2 Procesado de datos

En la introducción se adelantaba que la entrada a cada nuevo proceso de tratamiento de datos iba a ser la salida del proceso anterior. De esta forma, esta etapa se servirá de los ficheros generados en la limpieza de datos. Estos ficheros se encuentran almacenados en la lista nombrada 'dfList', cada uno como un *DataFrame* diferente.

Cada uno de estos *DataFrames* presenta la información de taquilla correspondiente a una semana concreta para cada una de las películas en proyección dicha semana. Con esta predisposición de los datos, no se podría llevar a cabo un estudio que fuese más profundo que

una comparativa entre las diferentes semanas. Va a ser en esta etapa donde se solucione este problema.

El propósito de este apartado se resume en transformar la lista de entrada, compuesta por *DataFrames* correspondientes a la información recopilada en taquilla para cada determinada semana, en una nueva lista que contenga para una determinada película, toda la información de su evolución temporal.

Para alcanzar la solución al problema planteado se declara una lista vacía que será la variable que almacene los nuevos *DataFrames* específicos por película y se propone el desarrollo de un script que sea capaz de leer los títulos uno a uno de cada *DataFrame* y comprobar si existe una entrada con ese mismo título en la lista que será la variable que retorne la ejecución de la rutina. De esta forma, se darán tres situaciones que se detallan a continuación y que están referenciadas en el código para una mejor comprensión del mismo:

- Situación 1:
  - La lista aún no está inicializada. Se crea el primer *DataFrame* y se añade la fila correspondiente a dicha película.

- Situación 2:
  - Encontramos una coincidencia en la lista destino. En este caso se actualiza el *DataFrame* añadiendo la información de la nueva semana.

- Situación 3:
  - Después de recorrer toda la lista no obtenemos ninguna coincidencia. Se crea un nuevo *DataFrame* y se añade a la lista.

Tras ejecutar el código que se muestra en la siguiente página y que contempla las situaciones detalladas anteriormente, la variable ‘dfMoviesList’ contendrá un *DataFrame* exclusivo por cada película con toda su información temporal y actuará como base de datos para futuros scripts.

```
dfMoviesList <- list()

for (i in 1:length(dfList)){
  for (j in 1:length(dfList[[i]]$Titulo)){

    Movie <- dfList[[i]]$Titulo[j]
    row_index <- which(dfList[[i]]$Titulo == Movie)
    Found = FALSE;

    #Sitaución 1

    if (length(dfMoviesList) == 0){
      DF1 <- data.frame()
      DF1 <- rbind(DF1, dfList[[i]][row_index,])
      dfMoviesList[["DF1"]] <- DF1
    }else{
      for (k in 1:length(dfMoviesList)){

        #Sitaución 2

        if (Movie == dfMoviesList[[k]]$Titulo[1]){
          Found = TRUE;
          DF <- dfMoviesList[[k]]
          DF <- rbind(DF, dfList[[i]][row_index,])
          dfMoviesList[[k]] <- DF
        }else{
          NULL
        }
      }

      #Sitaución 3

      if (Found == FALSE){
        DF <- data.frame()
        DF <- rbind(DF, dfList[[i]][row_index,])
        index <- length(dfMoviesList) + 1
        dfMoviesList[[paste0("DF",index)]] <- DF
      }
    }
  }
}
```

Para una mejor organización de los datos, se ordenan las filas de los diferentes *DataFrames* por el número de semana que presenten.

```
for (i in 1:length(dfMoviesList)){

  dfMoviesList[[i]] <- dfMoviesList[[i]][order(as.numeric
                       (dfMoviesList[[i]]$SEM)),]

}
```

En la figura 2.2.1 se muestra un ejemplo de cómo queda un *DataFrame* tras pasar por esta etapa de procesado. Comparando esta captura con la representación en la figura 2.1.2, se observa la evolución hacia unos datos listos para analizar.

	Titulo	TituloOriginal	Year	DIST.	SEM	CINES	PANTALLAS	GrossTotal	GrossPromPerCINE
1	Sherlock Holmes: Juego de Sombras	-	2012	WBI	2	267	451	1628744	6100.16479400749
2	Sherlock Holmes: Juego de Sombras	-	2012	WBI	3	363	430	939124	2587.11845730028
3	Sherlock Holmes: Juego de Sombras	-	2012	WBI	4	345	369	592462	1717.28115942029
4	Sherlock Holmes: Juego de Sombras	-	2012	WBI	5	306	306	368371	1203.82679738562
5	Sherlock Holmes: Juego de Sombras	-	2012	WBI	6	190	190	179145	942.868421052632
6	Sherlock Holmes: Juego de Sombras	-	2012	WBI	7	91	91	67224	738.725274725275
7	Sherlock Holmes: Juego de Sombras	-	2012	WBI	8	39	39	18331	470.025641025641

Figura 2.2.1 DataFrame específico por película.

En este punto cabe mencionar que los tiempos de ejecución de los scripts comienzan a ser cuestiones a tener en cuenta para un correcto desarrollo del proyecto. En la figura 2.2.2 se muestra la variable 'dfMoviesList' cargada en el entorno de R para poder trabajar con ella. Si fuera necesario generarla cada vez que se abra el entorno de programación para retomar el proyecto, no se estaría haciendo un buen uso de las posibilidades que ofrece este lenguaje con la consecuente pérdida de tiempo que conlleva.

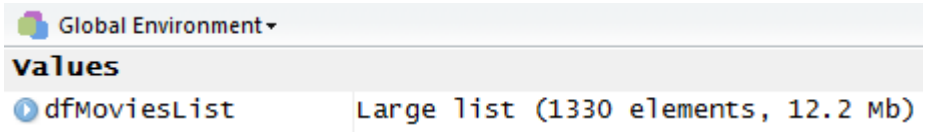


Figura 2.2.2 Variable cargada en el entorno R.

R contempla estos casos en los que las variables tienen un peso muy elevado y se requiere un alto tiempo de cómputo para generarlas y permite guardar dichas variables como objetos mediante la instrucción:

```
save(dfMoviesList, file = "dfMoviesList.Rdata")
```

Ya no será necesario generar más veces 'dfMoviesList', se tendrá siempre accesible en el directorio de trabajo para su uso instantáneo.

En la siguiente etapa de procesado, se resolverá el problema de la información repetida, como es el caso del título o del año, y se completará la información de cada película con metadatos que ofrezcan información sobre género y país de procedencia entre otros.

2.3 Información adicional

La lista generada en la etapa anterior está preparada para poder iniciar los procesos de análisis, sin embargo como esta lista actuará como base de datos y de ella dependen la variedad de enfoques que puedan llevarse a cabo, se añadirá información adicional que permita ampliar la forma de relacionar las características ya existentes.

También se cambiará la predisposición de mostrar los datos del apartado anterior, aunque por motivos puramente estéticos. En la definición de *DataFrame*, se mencionaba que podían almacenar vectores de igual longitud, cualidad que se aprovecha para organizar la estructura de la información semanal, pero que afecta negativamente a las variables que no hace falta ver repetidas, como son el título, año o distribuidora. Se propone entonces la estructura de almacenamiento de datos que se empleará para el resto del proyecto: una lista de listas de *DataFrames*. Entrando en detalles, se declarará una lista por cada película, que contendrá un *DataFrame* por cada variable categórica que sólo deba presentar un valor, caso del título o distribuidora, y un *DataFrame* con toda la información semanal, dado que aquí los vectores columna sí tendrán todos la misma longitud. Por último almacenaremos cada lista de *DataFrames* por película específica en una lista que contenga a todas las demás listas. La ventaja que ofrece este tipo de almacenamiento es que por un lado se evita tener datos repetidos y por otro que sigue garantizando un fácil acceso a cualquier valor que se requiera.

Para incorporar la nueva información y crear la nueva estructura, se propone un script que logra simultáneamente ambos objetivos. Previa a la ejecución del mismo, ha sido necesario incorporar a cada película su identificador de IMDB [5]. Con el identificador disponible, se realizan consultas a la API de OMDB [6] que retorna la ficha de la película solicitada en formato JSON, un formato ligero para el intercambio de datos que no requiere el uso de XML y la gran ventaja que ofrece es que para su manejo requiere *parsers* (analizadores sintácticos) muy sencillos. R ofrece una librería para trabajar con este tipo de datos, ‘rjson’, con funciones predefinidas que permiten la traducción instantánea de este formato, facilitando así la tarea de incorporar la nueva información.

El código necesario para hacer capturas de OMDB y generar la nueva estructura se detalla a continuación. Inicialmente, se carga la librería para tratar con JSON y se declara la variable que contendrá cada lista específica por película, que es también una lista.

```
library(rjson)

MoviesList <- list()
```

En esta primera parte del script, se declaran todos los *DataFrames* que van a componer cada película para acto seguido, rellenarlos con la información que requiera cada uno. Esta parte soluciona que la lista del apartado anterior tuviese variables repetidas tantas veces como longitud presentase el vector columna que formaba su matriz de datos.

```
for (i in 1:length(dfMoviesList)){

  Titulo <- data.frame()
  TituloOriginal <- data.frame()
  IMDB_ID <- data.frame()
  Mes <- data.frame()
  Year <- data.frame()
  Dist <- data.frame()
  Genero <- data.frame()
  Rating <- data.frame()
  Country <- data.frame()
  SEM <- data.frame()
  Others <- data.frame()

  List <- list()

  DF <- dfMoviesList[[i]]

  Titulo <- DF$Titulo[1]
  TituloOriginal <- DF$TituloOriginal[1]
  IMDB_ID <- DF$IMDB_ID[1]
  Mes <- DF$Mes[1]
  Year <- DF$Year[1]
  Dist <- DF$DIST.[1]
  SEM <- DF[-c(1:6)]

  if (is.null(IMDB_ID)){
    IMDB_ID <- "-"
  }
```

Toda la información disponible hasta ahora proviene de los ficheros de Rentrak. OMDB permite accesos a su API desde la ejecución de código a través de su URL seguida del valor del identificador, por tanto se realiza una búsqueda para todos los identificadores de IMDB distintos de un valor nulo asignado:

```
if (IMDB_ID != "-"){

  url <- paste0("http://www.omdbapi.com/?i=",IMDB_ID)

  json_info <-  fromJSON(readLines(url, warn=FALSE))
```

La función 'fromJSON' almacena en la variable 'json_info' la ficha de la película solicitada, con un campo que indica si la petición ha tenido éxito. Para estos casos positivos, se añaden a los datos disponibles información sobre género, calificación IMDB y país de procedencia del largometraje entre otros.

```
  if (json_info$Response == "True"){

    List[["Titulo"]] <- Titulo
    List[["TituloOriginal"]] <- TituloOriginal
    List[["IMDB_ID"]] <- IMDB_ID
    List[["Mes"]] <- Mes
    List[["Year"]] <- Year
    List[["Dist"]] <- Dist
    List[["Genero"]] <- json_info$Genre
    List[["Rating"]] <- json_info$imdbRating
    List[["country"]] <- json_info$Country
    List[["SEM"]] <- SEM


    Others <- as.data.frame(json_info)
    List[["Others"]] <- Others[-c(1,2,4,6:12,14:20)]
```

En el caso de que no se disponga del identificador de una producción o de que la respuesta tras la petición sea negativa, se mantiene la información previa sin ampliar:

```
}else{
  List[["Titulo"]] <- Titulo
  List[["TituloOriginal"]] <- TituloOriginal
  List[["IMDB_ID"]] <- IMDB_ID
  List[["Mes"]] <- Mes
  List[["Year"]] <- Year
  List[["Dist"]] <- Dist
  List[["SEM"]] <- SEM
}
```

Finalmente, la lista que contiene todas las características sobre cada determinada película se inserta en la lista 'MoviesList', y se guarda esta estructura para su posterior uso en la etapa de análisis.

```
  MoviesList[[paste0("DFL",i)]] <- List
}

save(MoviesList, file = "MoviesList.Rdata")
```

2.4 Simplificación de variables categóricas

En los problemas con gran volumen de datos es frecuente retocar la disposición de los mismos después de realizar un análisis de resultados. Está secuencia de reorganización-análisis se aplica con el fin de solventar posibles anomalías que se presenten a la hora visualizar representaciones o para corregir defectos en el modelo planteado.

En este punto del procesamiento de datos se procede a reestructurar la distribución de los géneros y de los países de procedencia de las películas capturados de OMDB. Es necesario someter a los datos a este cambio porque para el caso concreto de los géneros, la información capturada se corresponde con subgéneros con la consecuente dispersión de las películas en diferentes grupos con menor frecuencia absoluta de aparición. La solución propuesta se basa en unificar los subgéneros en géneros más amplios que engloben más volumen de producciones.

Para no perder los géneros originales, la lista obtenida en el apartado anterior se mantendrá intacta, pero se duplicará para realizar los cambios sobre la nueva del modo que sigue:

```
NewMoviesList <- MoviesList

for (i in 1:length(NewMoviesList)){

  Movie <- NewMoviesList[[i]]

  if (!is.null(Movie$Genero) && Movie$Genero[1] != "N/A"){

    Genero <- strsplit(Movie$Genero[1],', ')[[1]]
    l = length(Genero)

    for (j in 1:l){

      if (Genero[j] == "Crime"   || Genero[j] == "Horror" ||
          Genero[j] == "Mystery" || Genero[j] == "Thriller"){

        NewMoviesList[[i]]$Genero[1] <- "Thriller"

        break
```

En el ejemplo de código se comprueba si alguno de los subgéneros *Crime*, *Horror*, *Mystery* o *Thriller* está entre los datos proporcionados por OMBD, en caso afirmativo el género se

establecerá como Thriller, considerando este último como el que engloba a todos los demás. Siguiendo este procedimiento, se contemplarán los siguientes géneros:

- Aventuras
- Drama
- Comedia
- Animación
- Thriller
- Documental

Se realizará también una nueva clasificación para los países de procedencia. Teniendo en cuenta que los dos países que proporcionan mayor material cinematográfico a las taquillas españolas son Estados Unidos y España, se crearán tres grupos para organizar los países, los dos primeros se corresponden con los ya nombrados y un tercer grupo que englobe a todos los demás. El código es similar al desarrollado para el agrupamiento de géneros. En este caso vamos a filtrar por la variable 'Country':

```
if (!is.null(Movie$country) && (Movie$country != "")){

  Country_List <- strsplit(Movie$country[1],', ')[[1]]
  l = length(Country_List)

  Found = FALSE

  for (j in 1:l){

    if (Country_List[j] == "Spain"){

      NewMoviesList[[i]]$country[1] <- "Spain"
      Found = TRUE
      break

    }else if(Country_List[j] == "USA"){

      NewMoviesList[[i]]$country[1] <- "USA"
      Found = TRUE
      break

    }else{
      NULL
    }
  }

  if (Found == FALSE){

    NewMoviesList[[i]]$country[1] <- "Other"
  }
}
```

Con este retoque a la clasificación de los metadatos concluye la etapa de procesado.

Mediante los scripts propuestos a lo largo de este capítulo se consigue alcanzar el objetivo de disponer de unos datos listos para el análisis y numerosas posibilidades de representaciones con diferentes enfoques que se tratarán en el siguiente capítulo.

Capítulo 3 – Resultados

3.1 Herramientas de representación

Con los datos extraídos se realizarán diversos estudios que se dividirán en secciones a lo largo de este capítulo. A modo de introducción se detallarán los recursos y modo de empleo de los mismos para lograr las representaciones en las que se apoyará el análisis. Contrariamente a la estructuración y desarrollo del capítulo previo, que se comentaba el código necesario según fuese siendo utilizado, para no entorpecer la explicación y visualización de los resultados en este capítulo se comenzará explicando el material y métodos empleados, para poder analizar los datos en los apartados posteriores sin interrupción.

Todas las gráficas han sido generadas a través de R y se diferenciará dos tipos de visualizaciones en base a la complejidad que requiera cada implementación. Para representar un único valor de eje Y por valor de eje X, será necesaria la librería 'ggplot2' [7]. El siguiente fragmento de código carga dicha librería y mediante una de las funciones que nos ofrece, recibiendo como parámetro el *DataFrame* en cuestión (DF) y las variables que corresponden con los valores a representar en cada eje (número de semana y *gross total*), se genera en la variable 'g' la gráfica deseada. Con la multitud de parámetros que acepta este tipo de representación, es posible personalizar cada aspecto de la gráfica. En este caso se especifican los valores del eje Y para que se muestren en millones en lugar de en unidades, se añade un título y se establece un tipo de fuente más atractiva que la que viene por defecto. Estas instrucciones generan la figura 3.2.1:

```
library(ggplot2)

g <- ggplot(DF, aes(SEM, GrossTotal)) + geom_bar(stat="identity")
g <- g + scale_y_continuous(label=function(x){return(paste( x/10^6,"M"))})
g <- g + ggtitle('The Dark Knight Rises')
g <- g + theme(plot.title = element_text(size=20, face="bold", vjust=1))
```

En el caso de que se requieran varios valores del eje Y para un único valor del eje X, muy útil para comparar diversos valores para una misma variable, la implementación aumenta en complejidad. Para la representación de la gráfica se mantendrá el uso de la librería 'ggplot2' y para poder organizar las distintas variables e indicar a R el conjunto de valores de eje Y que

tendrá asociado cada valor de eje X, se empleará la librería 'reshape2'. Para ver el código, se muestra el ejemplo de la figura 3.3.2.

Primero se crea un *DataFrame* con la información de estrenos de los tres géneros que más títulos abarcan, con una cuarta columna que contenga los meses del año y se carga la librería necesaria:

```
df <- data.frame(Drama,Comedy,Adventure,Mes)
library(reshape2)
```

El siguiente paso es hacer una fusión del *DataFrame* con la función 'melt' que proporciona la librería cargada, convertir los valores que devuelve a numéricos y asignar nombres a las diferentes columnas resultado de la fusión, donde *Fi* indica frecuencia absoluta, para poder hacer referencia a ellas posteriormente:

```
df <- melt(df, id.vars='Mes')
df$value <- as.numeric(df$value)
names(df) <- c("Mes","Genre","Fi")
```

Cuando se hace una fusión, básicamente se crea una columna con el valor y otra columna con las variables, para el ejemplo, con *Drama*, *Comedy* y *Adventure*. Finalmente se representa la gráfica como en el apartado anterior, con la novedad de que ahora además de pasar los correspondientes parámetros, con la opción 'fill' se indica que para cada mes concreto, se representará el valor correspondiente del eje Y para las variables contenidas en 'Genre':

```
g <- ggplot(df, aes(x=Mes, y=Fi, fill=Genre)) +
     geom_bar(stat='identity', position='dodge')
g <- g + ggtitle('Genres from Spain')
g <- g + theme(plot.title = element_text(size=20, face="bold", vjust=1))
```

Se procede a analizar a lo largo del capítulo las gráficas que se han considerado de interés para realizar el estudio retrospectivo y que se han desarrollado a través de las instrucciones propuestas.

3.2 Representaciones básicas

Una primera toma de contacto con la representación de datos situará en contexto para poder profundizar más adelante en el análisis.

3.2.1 Evolución temporal de los ingresos de una película

El punto de partida se sitúa en visualizar la evolución temporal de los ingresos (de ahora en adelante *gross*) de una película. En la figura 3.2.1.1 se aprecia como la semana del estreno es la que más impacto tiene en aspectos referentes a recaudación, siendo en este caso concreto, más del doble del *gross* de la semana 2 y más del triple de correspondiente a la semana 3. Este modelo se repite en la representación del *gross* total frente a la semana para todas las películas en condiciones normales.

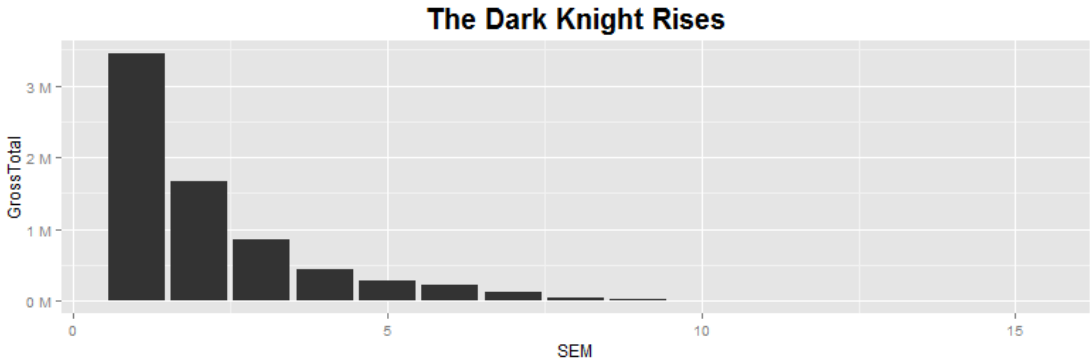


Figura 3.2.1.1 – Gross total respecto de la semana – The Dark Knight Rises.

Se especificaba previamente que la representación corresponde con la trayectoria en condiciones normales por la posibilidad de que el largometraje reciba un premio mientras aún está en proyección. Eligiendo como ejemplo Birdman (figura 3.2.1.2), que recibió cuatro premios Oscar mientras aún estaba en taquilla, se observa en la gráfica la repercusión que tuvo sobre el público dicho éxito.

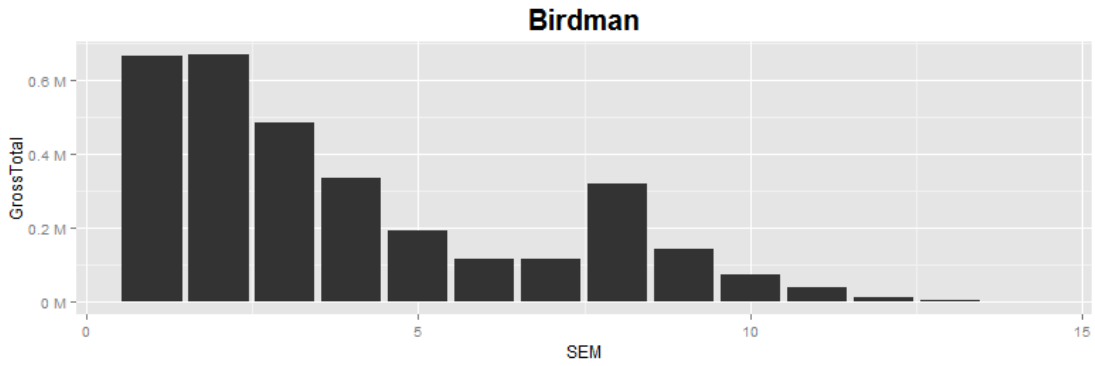


Figura 3.2.1.2 – Gross total respecto de la semana – Birdman.

3.2.2 Ingresos por distribuidora

Cambiando el enfoque, la siguiente variable a tener en cuenta es la distribuidora. La figura 3.2.2.1 muestra los ingresos acumulados por cada distribuidora desde 2012 hasta mediados de 2015. Con una cantidad superior a 350 millones de euros, WBI encabeza la lista, pero se deben tener en cuenta las producciones que llevan a taquilla cada distribuidora para poder elegir la que más rentabiliza los proyectos.

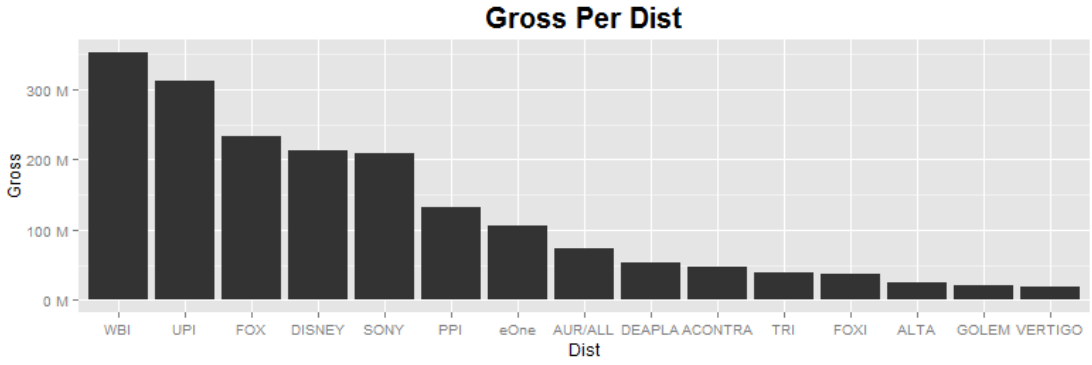


Figura 3.2.2.1 – Gross generado por distribuidora.

En la figura 3.2.2.2, se puede comprobar que WBI pese a ser la que más recauda, no es la que más películas distribuye, como se podría haber pensado inicialmente.

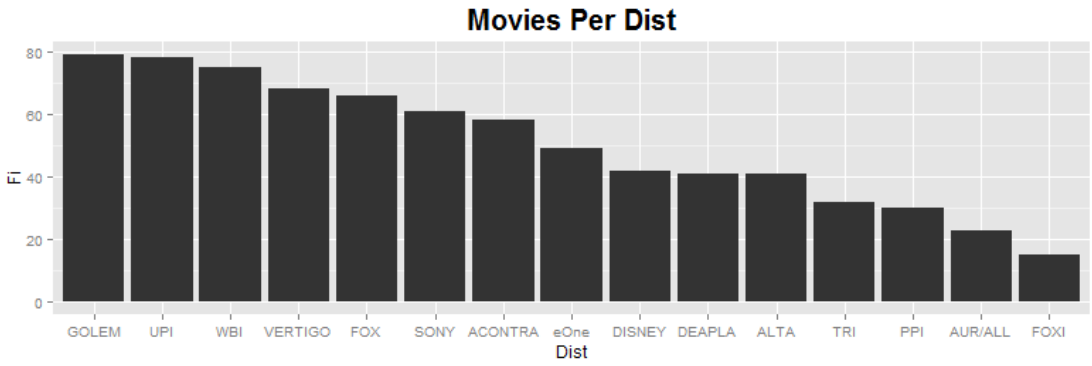


Figura 3.2.2.2 – Número de estrenos por distribuidora.

Se determinará qué distribuidora genera más ingresos por producción realizando un promedio entre ingresos totales y número de estrenos. La figura 3.2.2.3 muestra el resultado de la operación anterior, obteniendo como resultado que las producciones de Disney recaudan una media de 5 millones de euros convirtiéndola en la distribuidora más atractiva en este aspecto.

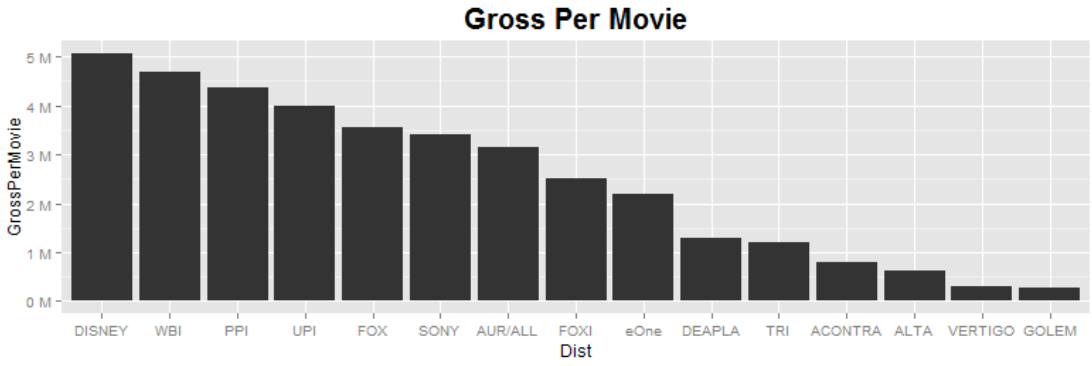


Figura 3.2.2.3 – Gross promedio por película de cada distribuidora.

3.2.3 Ingresos por género

Replicando el análisis previo para el estudio del género, la figura 3.2.3.1 muestra como las películas de aventuras obtienen unos beneficios considerablemente superiores sobre los demás géneros. Como se demostró en el caso de las distribuidoras, este hecho no significa que sea el género más rentable, sólo indica que es el que más recauda.

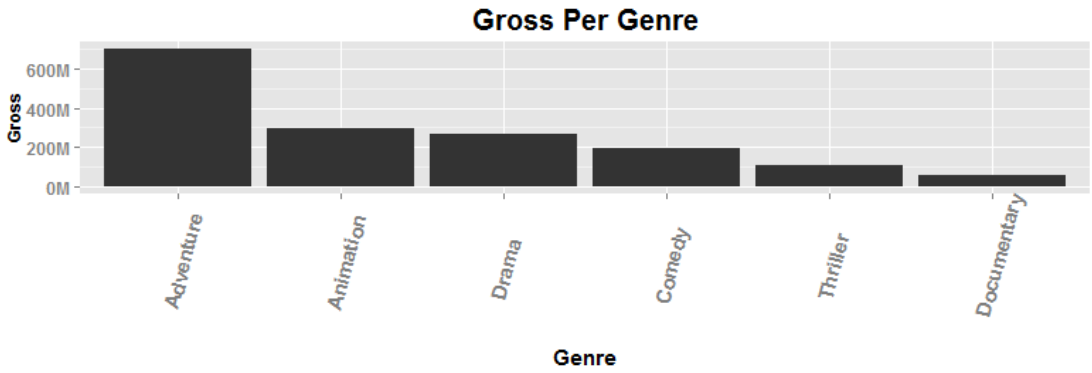


Figura 3.2.3.1 – Gross generado por género.

Para ver qué género ofrece mejor promedio de recaudación, se analizan cuántos estrenos corresponden a cada uno. La figura 3.2.3.2 muestra el consecuente resultado, declarando las películas de drama como las más abundantes en las taquillas, seguidas de las de aventuras. Por último en la figura 3.2.3.3, se realizan las operaciones necesarias para visualizar qué género ofrece mayor media de ingresos por cada título que se estrena.

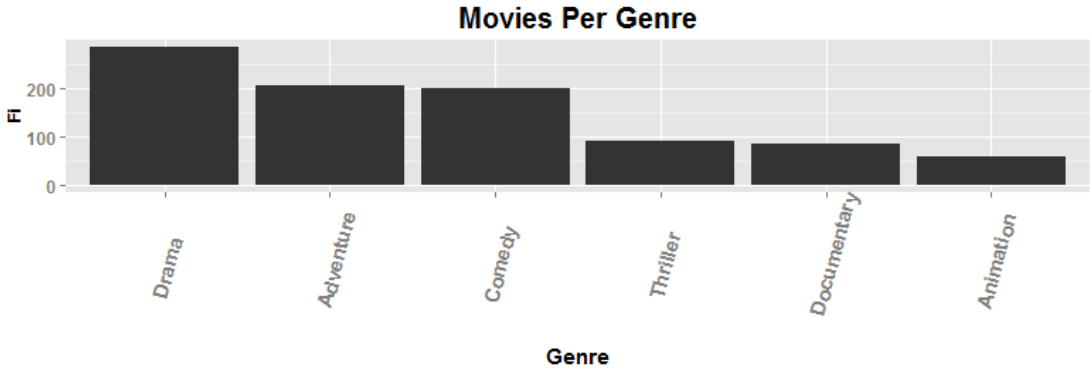


Figura 3.2.3.2 – Número de estrenos por género.

Similar a lo obtenido en lo que respecta a las distribuidoras, el género que mayores beneficios genera por película tampoco encabezaba ninguna de las representaciones previas.

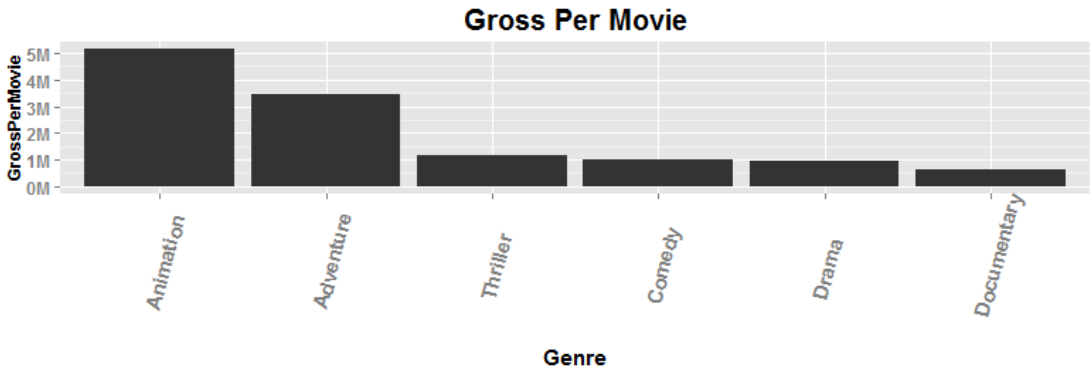


Figura 3.2.3.3 – Gross promedio por película de cada género.

Estas representaciones básicas sirven de introducción para tener una pequeña visión de cómo afectan determinadas variables a los ingresos totales. En el siguiente apartado se profundiza el análisis teniendo en cuenta varias variables simultáneamente, ofreciendo así un mayor nivel de detalle que conlleve a resultados más precisos.

3.3 Representaciones de varias variables

Para las gráficas de este capítulo se han seleccionado los géneros que más películas engloban, a saber, drama, aventura y comedia, tal como muestra la figura 3.2.3.2 y en cuanto a la procedencia de las producciones, para que el análisis sea más preciso se tendrán en cuenta como países productores España y Estados Unidos.

3.3.1 Estrenos por mes según el género y el país

Con las consideraciones anteriores, se procede a visualizar el volumen de estrenos que hay a lo largo del año para cada uno de los distintos géneros. En la figura 3.3.1.1 se muestra dicha evolución para las películas de origen estadounidense y en la figura 3.3.1.2 para las españolas.

Es notable la predilección de Estados Unidos por inundar las taquillas de forma constante con películas de aventuras/acción, siendo prácticamente el género dominante en lo que respecta a número de estrenos durante todo el año. Del mismo modo, se puede apreciar cierta ventaja de las películas de drama frente a las de comedia durante los primeros cuatro meses, situación que se da la vuelta en el segundo cuatrimestre.

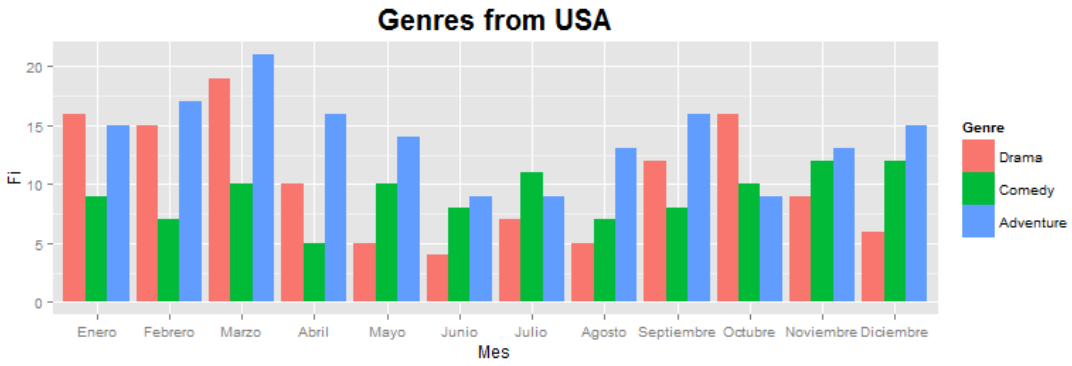


Figura 3.3.1.1 – Estrenos por mes de películas de Estados Unidos.

En España, sin embargo, es el género de drama el que tiene más impacto en taquilla. Se observa cierta tendencia a sacar a la luz películas con estas características en junio, septiembre y noviembre.

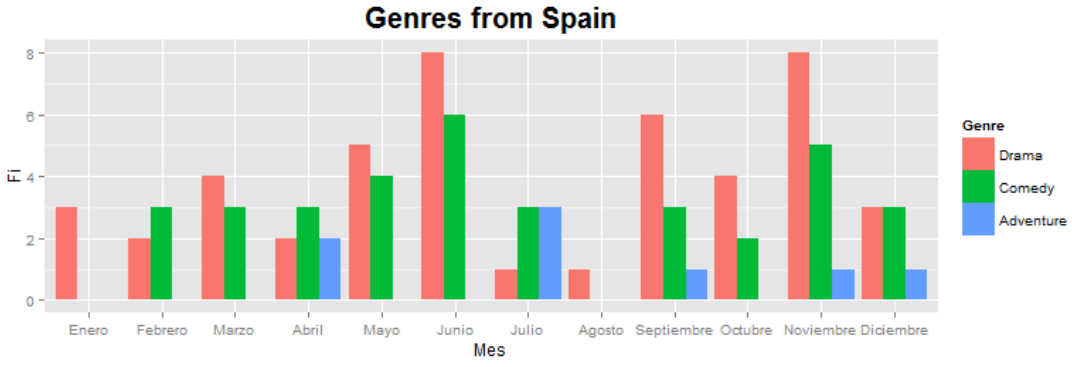


Figura 3.3.1.2 – Estrenos por mes de películas de España.

El motivo de las representaciones anteriores resulta muy atractivo de cara a la elección del mes para estrenar una película. Asumiendo constante el volumen de estrenos para el transcurso de los años y con acceso a esta información, si se desease llevar a taquilla una producción de origen español y de género drama, cabría considerar que los meses de junio, septiembre y noviembre no serían los más atractivos. El público no aumenta porque haya más oferta, sino que se reparte entre dicha oferta, por tanto convendría fijarse en los meses que dejan este sector libre, que serían julio y agosto seguidos de febrero y abril. Con este simple estudio de mercado se puede tratar de satisfacer las necesidades de los consumidores y alcanzar el máximo éxito para un proyecto dado.

3.3.2 Promedio de copias por cine según el género y país

Profundizando, se puede mostrar la relevancia de los estrenos visualizando el número promedio de copias que adquiere cada cine para su proyección simultánea. En la figura 3.3.2.1 se puede comprobar cómo las películas de aventuras procedentes de los Estados Unidos alcanzan casi una media de 2 proyecciones por cada cine, indicativo de que el largometraje está tendiendo cierta repercusión en taquilla.

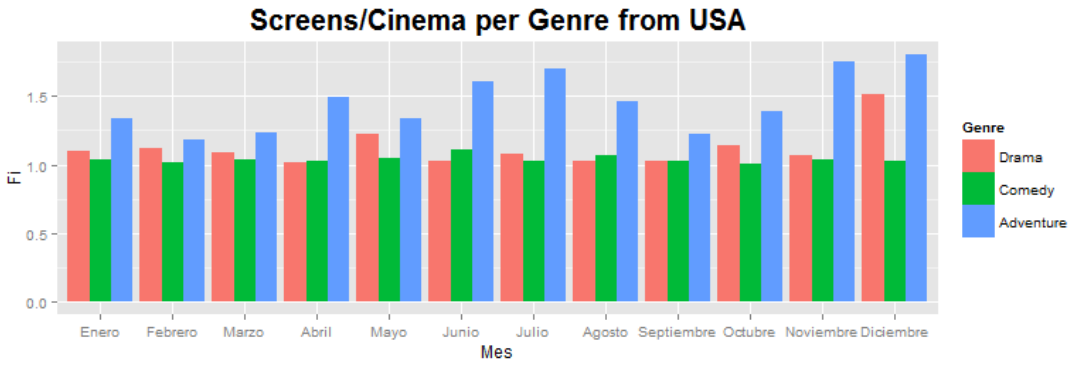


Figura 3.3.2.1 – Promedio de copias por cine y por mes de películas de Estados Unidos.

Un suceso interesante que permiten realizar los análisis retrospectivos, es la explicación de los valores atípicos en las representaciones. Es curioso, por ejemplo, como para la gráfica que muestra la media de copias por cine para las películas españolas (figura 3.3.2.2) se observa en el mes de septiembre un pico para el impacto de las películas de drama. Después de investigar el por qué de dicho suceso, se descubrió que a finales de ese mes de septiembre, es costumbre la celebración de un festival de cine en San Sebastián [8], donde las películas que quieren competir por los diferentes reconocimientos, salen a taquilla ese mismo mes, y de ahí ese pico de la respectiva gráfica.

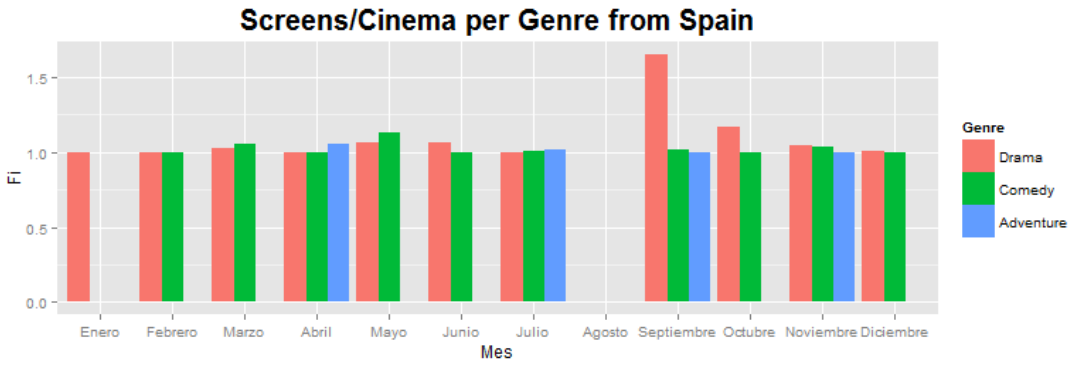


Figura 3.3.2.2 – Promedio de copias por cine y por mes de películas de España.

Con el estudio previo, se pueden sacar varias conclusiones de cara a elegir el mejor mes para llevar un largometraje a proyección. En el siguiente apartado del capítulo, se tratará de encontrar la relación entre los ingresos totales de una película con el ingreso generado en la primera semana en taquilla.

3.4 Recta de regresión lineal

En la industria del cine es común estimar el ingreso total de una película a partir de la recaudación de la primera semana en proyección y para ello se recurrirá a las rectas de regresión lineal. En estadística, la regresión lineal o ajuste lineal es un método matemático que modela la relación entre variables dependientes y variables independientes. Para este caso, la variable independiente será el ingreso de la semana 1 y la variable dependiente será la recaudación total (porque dependerá del ingreso de la primera semana). Con el fin de una mejor visualización de las gráficas, se han aplicado logaritmos tanto para los valores del eje X como para los del eje Y.

En la figura 3.4.1 se muestra la nube de puntos y la respectiva recta de regresión que relaciona el *gross* de la primera semana con el *gross total* de todas las películas que hay disponibles en la base de datos.

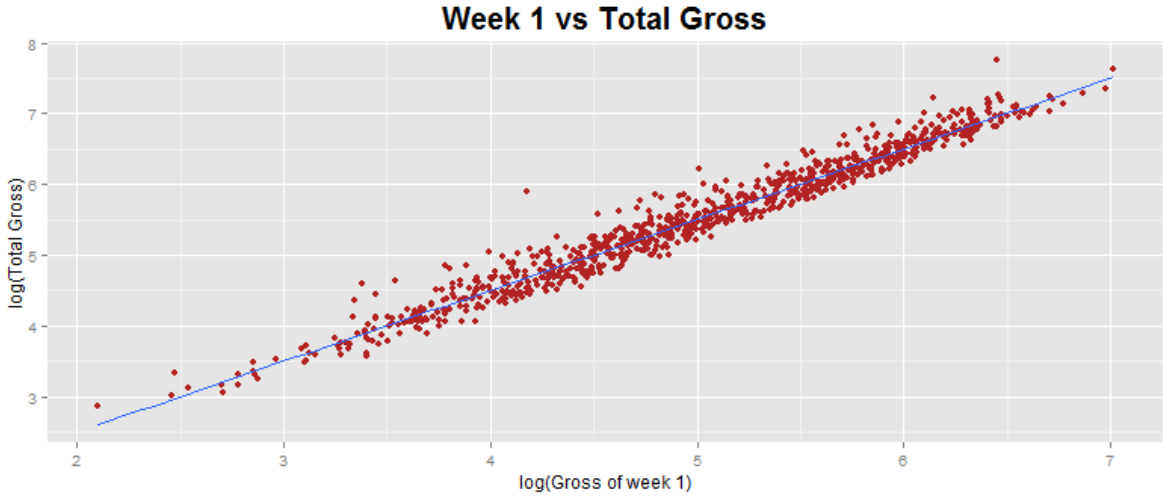


Figura 3.4.1 – Ingresos totales frente a ingresos en la semana 1.

Para hallar los parámetros de la recta se hace uso de la función 'lm' (*linear models)* [9] y se le pasan los parámetros de interés:

```
m <- lm(TotalGross ~ Sem1)
a <- signif(coef(m)[1], digits = 2)
b <- signif(coef(m)[2], digits = 2)
equation <- paste("y = ",b,"x + ",a, sep="")
```

Se obtiene como ecuación de la recta:

$$y = 1x + 0.5$$

En las figuras 3.4.2 y 3.4.3 se filtra la base de datos para únicamente visualizar los tres géneros y los países que más películas engloban, respectivamente, con el fin de demostrar si varía la pendiente de la recta de regresión.

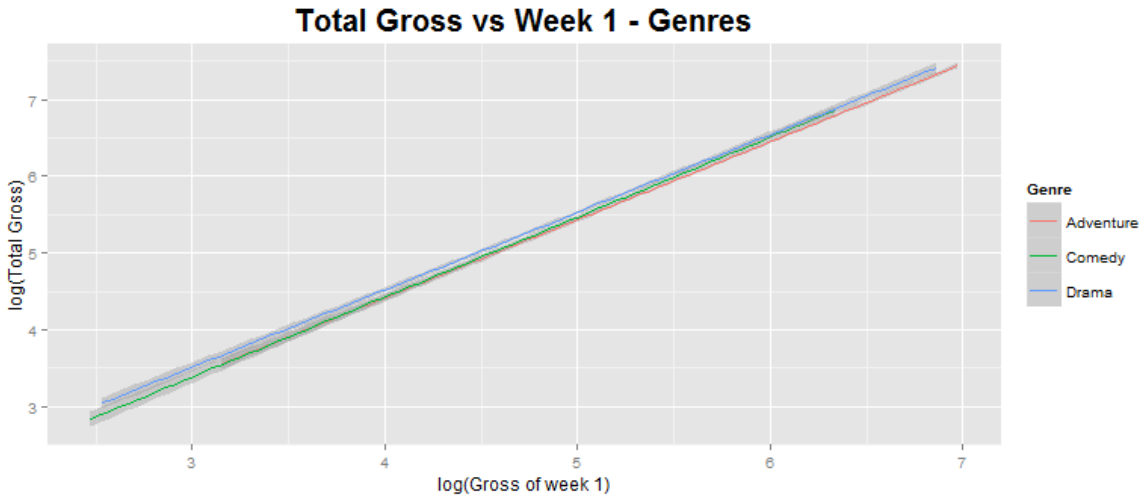


Figura 3.4.2 – Ingresos totales frente a ingresos en la semana 1 por géneros.

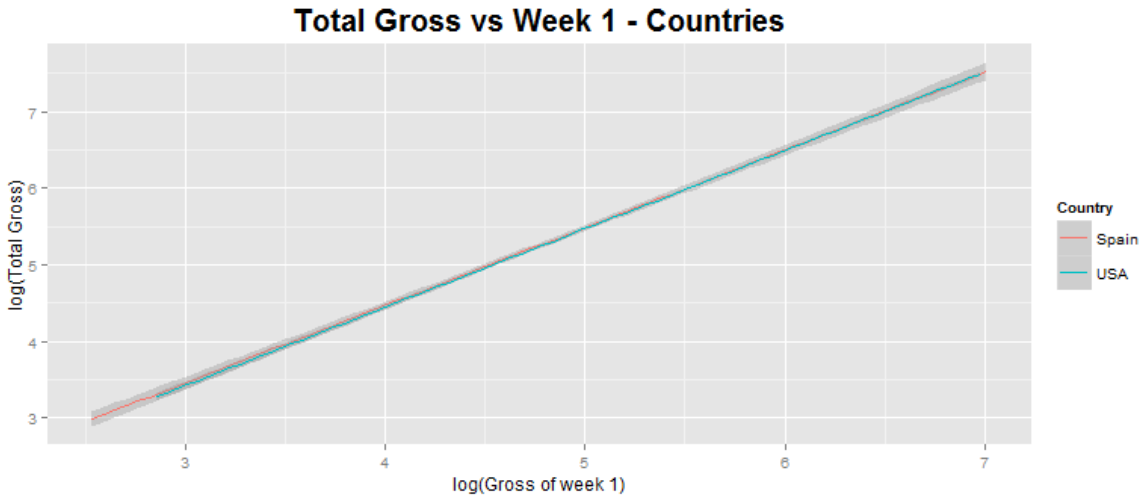


Figura 3.4.3 – Ingresos totales frente a ingresos en la semana 1 por países.

Se demuestra así que la pendiente no varía en función del país o del género para los datos analizados y que la ecuación anterior puede emplearse para realizar una estimación del *gross total* a partir del *gross* de la primera semana. En el siguiente apartado se tratará de realizar una predicción de los ingresos de la primera semana a partir de las características de un estreno potencial.

3.5 Árbol de regresión lineal

El objetivo de este apartado es predecir el *gross* de la primera semana a partir de las variables que definen una película, en concreto el género, el país de procedencia, el mes de estreno y la distribuidora. Para ello se hará uso de la librería 'rpart' [9] de R que proporciona los recursos necesarios para generar árboles de clasificación y de regresión lineal [10]. En el caso de la predicción de variables categóricas se emplearía un árbol de clasificación pero el ingreso de una película es una variable continua, por tanto este apartado se servirá de un árbol de regresión lineal.

Se ha decidido limitar el número de distribuidoras a aquellas que presenten un volumen superior a 20 títulos con el fin de evitar que el árbol tuviese en consideración variables tan específicas. El código necesario para cargar la librería y generar el árbol se detalla a continuación:

```
library(rpart)

df <- table(DataFrame$Dist)>20
NewDataFrame <- subset(DataFrame, Dist %in% names(df)[df])
rownames(NewDataFrame) <- NULL

tree <- rpart(Week1 ~ Genre + Country + Month + Dist,
              method="anova", data=NewDataFrame)

plot(tree, uniform=TRUE, main="Regression Tree for Movies")
text(tree, use.n=TRUE, all=TRUE, cex=.8)
```

La función 'rpart' se encarga de generar en la variable 'tree' el árbol de regresión lineal (el método 'anova' es para regresión lineal y el método 'class' para clasificación) que predecirá el *gross* de la primera semana a través de las variables que contienen el género, el país de procedencia, el mes de estreno y la distribuidora. Los datos empleados para el desarrollo del árbol son los de la base de datos creada en el capítulo 2 y filtrados para que presentasen los datos de interés para el árbol. Las estructuras de datos 'DataFrame' y 'NewDataFrame' contienen dichas variables de interés. En la figura 3.5.1 se muestra el árbol resultante y en la figura 3.5.2 la explicación complementaria que resulta necesaria para comprender las ramificaciones del árbol. Cabe destacar que los criterios de elección de una determinada rama dependen de si se cumple la condición que caracteriza a dicha rama, se toma el camino de la izquierda y si no se cumple, se toma el de la derecha. En la figura 3.5.2 los nodos terminales se identifican con un '*' al final de la línea.

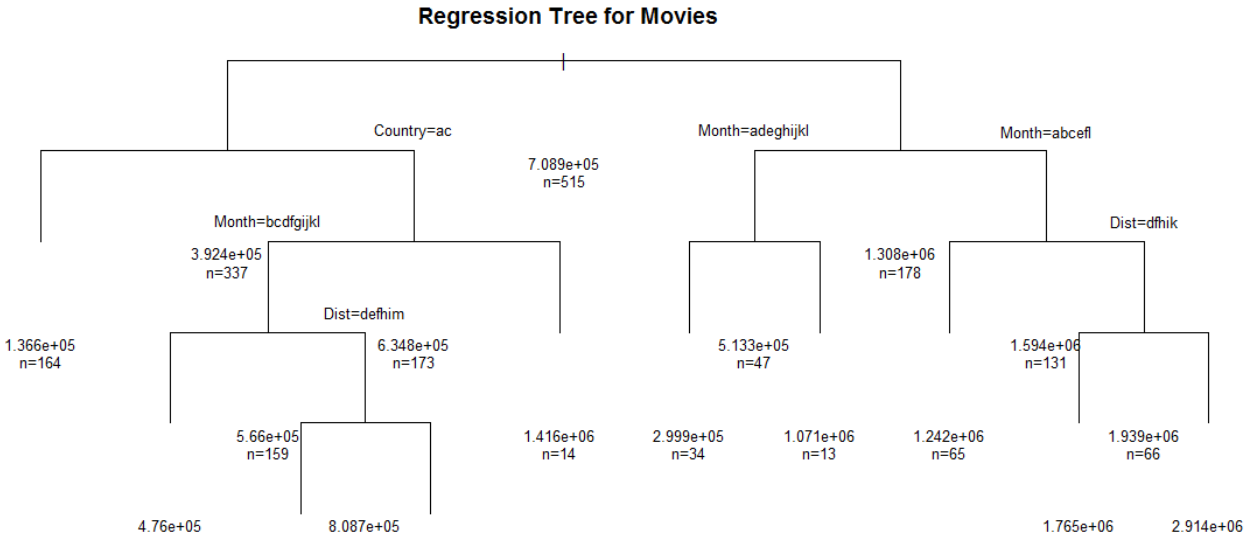


Figura 3.5.1 – Árbol de regresión lineal.

```
1) root 515 5.460429e+14  708944.7
  2) Genre=Comedy,Documentary,Drama,Thriller 337 1.983852e+14  392365.0
    4) Dist=ACONTRA,AVALON,DEAPLA,GOLEM,TRI,VERTIGO 164 6.604671e+12  136645.3 *
    5) Dist=DISNEY,eOne,FOX,PPI,SONY,UPI,WBI 173 1.708897e+14  634781.3
     10) Country=Other,USA 159 8.966456e+13  565970.2
       20) Month=10,11,12,3,4,6,7,8,9 116 2.527597e+13  475978.3 *
       21) Month=1,2,5 43 6.091488e+13  808739.1
         42) Dist=DISNEY,eOne,FOX,PPI,SONY,WBI 34 1.018161e+13  575118.4 *
         43) Dist=UPI 9 4.186727e+13 1691306.0 *
     11) Country=Spain 14 7.192194e+13 1416279.0 *
  3) Genre=Adventure,Animation 178 2.499377e+14 1308312.0
    6) Dist=ACONTRA,AVALON,DEAPLA,eOne,GOLEM,TRI,VERTIGO 47 3.375781e+13  513296.2
     12) Month=1,12,2,4,5,6,7,8,9 34 2.204451e+12  299933.5 *
     13) Month=11,3 13 2.595745e+13 1071322.0 *
    7) Dist=DISNEY,FOX,PPI,SONY,UPI,WBI 131 1.758156e+14 1593547.0
     14) Month=1,10,11,2,3,9 65 4.785196e+13 1242322.0 *
     15) Month=12,4,5,6,7,8 66 1.120485e+14 1939450.0
       30) Dist=DISNEY,FOX,PPI,SONY,UPI 56 7.113274e+13 1765390.0 *
       31) Dist=WBI 10 2.971810e+13 2914182.0 *
```

Figura 3.5.2 – Ramificaciones del árbol.

Otro de los atractivos de este árbol, es que retorna la importancia de las variables en el conjunto de datos estudiado para el objetivo solicitado. Para este conjunto de datos, el resultado es que el género de la película tiene un peso del 42%, la distribuidora un 37%, el mes de estreno un 12% y finalmente el país de procedencia un 9% en lo que respecta a la decisión del ingreso estimado. Siguiendo las ramas de interés y hasta llegar a un nodo terminal, se obtiene la predicción del *gross*, designado por el valor previo al símbolo '*'.

Para realizar una prueba del árbol de regresión, se tomará una película de test que no haya formado parte del conjunto que se ha utilizado para generarlo. Se va a predecir el *gross* estimado para la primera semana de la película 'Ahora o nunca'. Se trata de una película de comedia, de procedencia de España, que se ha estrenado en junio y es de la distribuidora Sony. El código necesario para crear esta prueba es:

```
test <- data.frame()
test <- rbind(test, c('Comedy','Spain','6','SONY'))
names(test) <- c("Genre","Country","Month","Dist")
```

Con la película de test disponible, se procede a predecir el posible ingreso cotejando con el árbol generado previamente:

```
> predict(tree, test)
      1
1416279
```

Como resultado se obtiene que una película de tales características podría obtener un beneficio la primera semana de 1.416.279 €. Comprobando con la información de Rentrak, la película 'Ahora o nunca' ha generado una cantidad exacta de 1.549.433 € en la semana de su estreno, por tanto se puede considerar una buena aproximación la estimación realizada. Para predecir el *gross total* a través de la predicción se emplea la recta del apartado 3.4:

$$y = 1 \cdot (\log 1416279) + 0.5 = 6.65$$

$$Total\ Gross = 10^{6.65} = 4478667\ €$$

La estimación del *total gross* para dicha película es de 4.478.667 €.

Con este último análisis concluye el capítulo de resultados. En el siguiente capítulo se extraerán las conclusiones del proyecto y se propondrán líneas futuras de trabajo.

Capítulo 4 – Conclusiones

4.1 Conclusiones generales del proyecto

Tras los análisis propuestos y resultados obtenidos en el capítulo 3, se pretende ofrecer un estudio de mercado del entorno del cine, de las variables que lo caracterizan y de su evolución a lo largo del año. Se ha tratado de dar respuesta a cómo es la evolución temporal de los ingresos de una película en proyección. También se aporta la información que permite comprender la distribución de los estrenos en función de su género y país de procedencia para cada mes, otorgando así la posibilidad de saber qué mes resulta más atractivo estrenar una película, analizando la oferta de la que dispone el público al que va dirigida dicha producción.

Mediante el análisis predictivo realizado en el punto 3.5 se puede estimar el ingreso que va a generar la película su primera semana en taquilla partiendo de unos parámetros dados. Utilizando esta predicción del *gross* de la primera semana y recurriendo al punto 3.4, se puede comprobar la estimación del *gross total* evaluando el valor deseado en la recta de regresión. Finalmente después de hacer la predicción del ingreso total a través de los apartados mencionados, se puede estudiar la viabilidad del proyecto analizando el apartado 3.3 y comprobar así si el mes elegido para el estreno es el óptimo o por el contrario, en dicho mes habrá más películas con características similares entre las que el público se repartirá.

De cara a la mejora de los resultados de los análisis, se proponen como líneas futuras de investigación el seguir recopilando ficheros con la información semanal, con objeto de seguir ampliando la base de datos y que lo resultados partan de más cantidad de datos procesados y analizados. También sería interesante disponer de las inversiones que se realizan sobre cada película y considerar esta nueva variable para realizar los análisis predictivos. Contando con las inversiones para cada producción el árbol de regresión lineal mejoraría considerablemente. De la misma manera, también se podría tener en cuenta información geográfica, pudiendo relacionarse las variables analizadas respecto a una determinada zona geográfica. Otro aspecto interesante a valorar sería la repercusión en redes sociales. Se podría definir una nueva variable que diese una indicación de cuánto éxito está teniendo una película en redes sociales como puedan ser Twitter o Facebook. Por último, cabría considerar la opción de acudir a las distribuidoras para recoger información sobre qué preguntas sería interesante responder mediante el análisis de datos y poder así realizar estudios más objetivos.

4.2 Competencias

A lo largo de este trabajo de fin de grado, he mencionado los términos *Big data* y *Data Science* para hacer referencia a la nueva rama que surge para analizar volúmenes masivos de datos y crear valor y significado sobre ellos. Después de ver la evolución y relevancia que esta nueva ciencia está tomando y puesto que durante mi grado de ingeniería no hay materias que incluyan esta temática, decidí que me iba a adentrar en el área del *Data Science* para desarrollar este proyecto.

Elementos clave en el transcurso del proyecto han sido los conocimientos adquiridos durante la carrera en el contexto de la estadística. Con la formación que recibimos a través de varias asignaturas, en concreto en Tratamiento Digital del Sonido, me ha sido posible entender los nuevos conceptos que me han sido necesarios para lograr los objetivos del proyecto y me han servido de base para ampliar mi conocimiento en este campo. De forma paralela, ha sido igual de importante estar familiarizado con entornos y lenguajes de programación. El hecho de haber cursado varias asignaturas de programación y entender diferentes lenguajes, ha tenido su repercusión en la fluidez con la que he aprendido R, el lenguaje en el que he implementado la totalidad de rutinas de este proyecto.

Entre las competencias adquiridas a destacar en el desarrollo del trabajo, puedo hacer mención a los conocimientos ampliados en lo que respecta a la estadística, destacando el trabajo realizado con árboles de regresión lineal, y a una buena introducción al análisis y procesado de datos. Como hecho más importante, he aprendido R, un lenguaje de programación que aún no se da como asignatura en la Universidad y que es el entorno de referencia para trabajar en el ámbito del Big Data.

Bibliografía

[1] http://drewconway.com

[2] El uso del marketing cinematográfico en la industria del cine español (2008). Rafael Linares Palomar.

[3] http://r-project.org y http://rstudio.com

[4] R Cookbook (2011). O'Reilly. Paul Teetor.

[5] http://imdb.com

[6] http://omdbapi.com

[7] R in a nutshell (2012). O'Reilly. Joseph Adler.

[8] http://sansebastianfestival.com

[9] http://statmethods.net

[10] An Introduction to Statistical Learning: with Applications in R (2014). Springer.