# INFORME ANALISIS DE CORRESPONDENCIA ANACOR

**Máster en Data Science**

**Juan Manuel López Zafra**

**Marta Divassón Carribero**

**Fecha: 1 noviembre 2018**

El estudio realizado tiene como objetivo llevar a cabo un análisis exploratorio de datos para determinar la posible relación existente entre un rango de edad de individuos determinado y distintos géneros cinemáticos.

El análisis de correspondencia o ANACOR, al igual que el análisis de componentes principales o análisis factorial, es una técnica de reducción de dimensión a través de una realidad multidimensional, en el que los factores vienen dados por correspondencias entre categorías de filas o columnas. Este persigue una pregunta: ¿a que se debe la ausencia de independencia?

Para ello, la prueba de independencia (*Chi cuadrado*) mide la posibilidad de rechazar la hipótesis nula en la que se plantea que las categorías sean independientes y afirmar una relación dada por la tabla de contingencia. Ahora bien, al realizarlo, se refleja la probabilidad de independencia de Pearson con 8 grados de libertad y un *pvalor* de 0. Por tanto, se rechaza la hipótesis nula y se observa una posible relación. No basta con esto ya que hay que fundamentarlo con un análisis más exhaustivo y así clarificar lo fundamentado.

En el siguiente grafico construido se representa mediante rectángulos el tamaño relativo de cada categoría y lo que aporta al modelo, mientras que el tamaño del circulo o "globo" indica la relación existente entre la variable de la fila y la columna a la que está asociada. En primer lugar, hay una gran asociación entre individuos de 25 y 50 años que ven películas de acción. También existe una relación notoria entre jóvenes menores de 25 años y películas de terror y una relación menos significativa en la que la comedia ocupa las categorías de menos de 25 años y 25 y 50 años de forma homogénea.

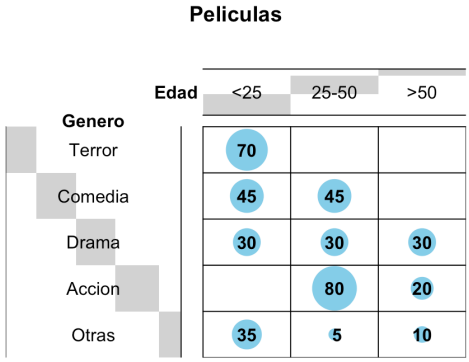
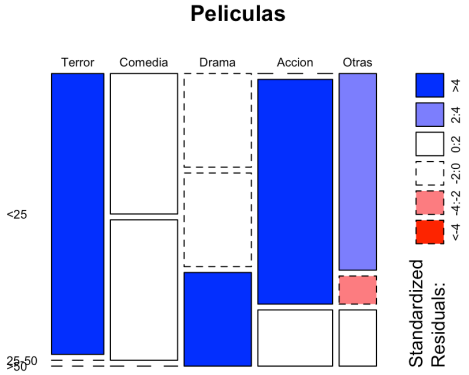



*Gráfico: Test de globos*

Se podría hacer una visualización más estricta, pero con el gráfico anterior hay datos detallados y claros. El segundo gráfico mide la importancia de cada categoría y su aportación a cada fila. Es decir, el color azul indica una mayor relación, esto es en las películas de terror y acción para los jóvenes menores de 25 y drama entre los jóvenes mencionados y los que están entre 25 y 50 años y una menor relación de otras categorías para aquellos jóvenes.

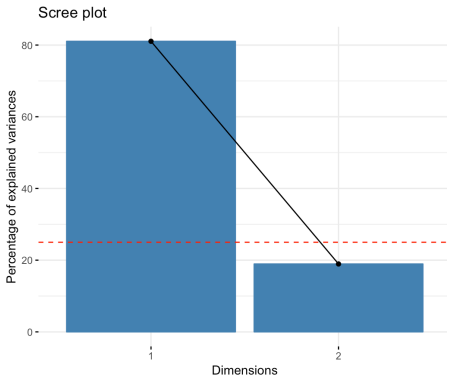



Este análisis de correspondencias de variables cualitativas muestra la clara relación de variables. Por tanto, se puede formalizar una reducción de la dimensión, siempre de forma en la que se pierda la

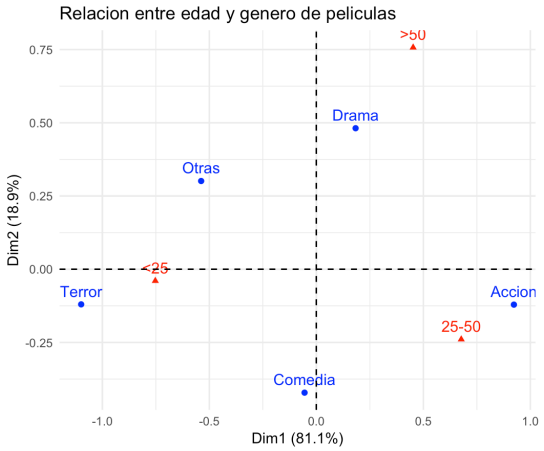
menor información posible, ya que hay datos que apenas aportan al modelo. Para ello, se utiliza "*FactoMineR*" ya que permite emplear técnicas de reducción bastante aplicadas.

En el resumen del análisis, al presentar 2 dimensiones, se acumula el 100% de la varianza explicada con la información que aportan las 2 dimensiones, ya que la varianza es el momento de inercia o fuerza centrifuga. Con esto, se puede representar la categoría de películas en un espacio de 2 dimensiones, es decir, la combinación de coseno cuadrado para las dos dimensiones es significativa ya que cada genero aporta información a la dimensión 1 o a la otra. Otro valor que observar es la contribución de cada género, siendo *Terror* y *Acción* los más contribuyentes a la Dimensión 1. Si estuviésemos en presencia de independencia, no se podría dar un porcentaje de explicación y este estudio no tendría sentido.

En la interpretación de los resultados, al medir el coeficiente de correlación (aquel que mide la asociación entre variables métricas y no métricas) obtenemos un valor de 0.7605, lo cual implica una correlación elevada y significativa (todo aquello por encima de 0.2 lo es). En el siguiente gráfico se muestra como el 100% de la explicación se da en las dos primeras dimensiones, lo cual la reducción es muy evidente, pero lo representamos además añadiendo la línea que indica el porcentaje explicativo si las variables fuesen independientes.

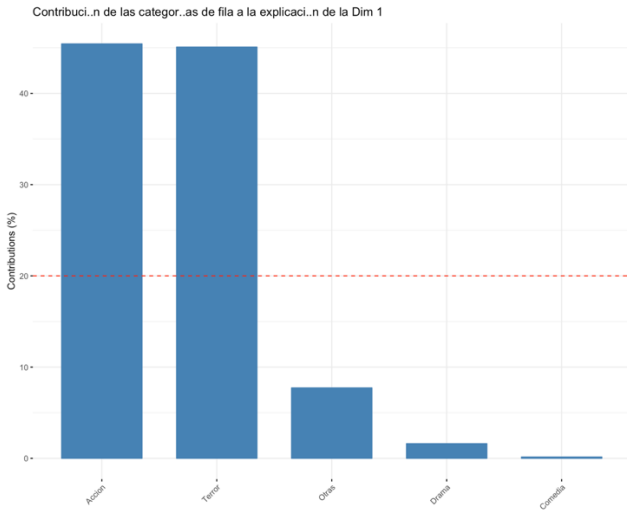



En el siguiente gráfico se muestra una relación entre variables y porcentaje explicativo a cada dimensión. Pero la representación análoga con un "biplot" es la más significativa ya que las flechas azules y rojas indica que cuanto menor es la distancia angular, mayor es la asociación entre colores(último gráfico).

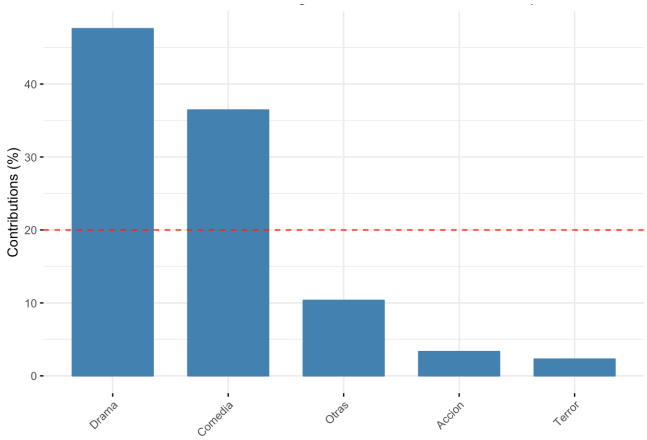



***Conclusión***

Tras un análisis mediante gráficos explicativos de la contribución de cada variable a las dimensiones, se obtiene con un gráfico en el que se representan las categorías mas importantes, se refleja la importancia de la variable "Terror" y "Acción", con respecto a la primera dimensión.




Mientras que en el caso de la segunda dimensión, las contribuciones son mayores por parte de las variables Drama y Comedia, como se muestra en el siguiente gráfico.




Finalmente, en el caso de un “biplot” se observa una serie de ángulos entre categorías mediante la representación de flechas. En el caso de las variables “Acción” y “Terror” distan por un ángulo de 180 grados, lo cual indica una asociación inversa lo cual implica relación, igualmente ocurre entre “Drama” y “Comedia”. Las variables con flechas horizontales(Acción y Terror) son las que más aportan a la primera dimensión y las verticales(Drama y Comedia) a la segunda dimensión. Otros géneros de peliculas aportan a ambas dimensiones.