Título del trabajo

Nombre y Apellido Alumno1, Nombre y Apellido Alumno 2

**Resumen.** Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen. Este es el resumen.

*Síntesis del trabajo en no más de 200 palabras*

1 Introducción

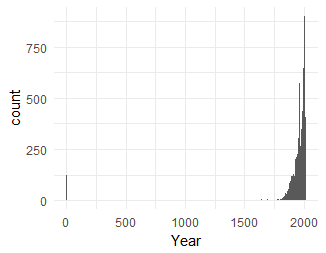
Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción.

*Toda la información de contexto que permita entender el trabajo que se realizó. No más de una carilla.*

2 Materiales y Métodos

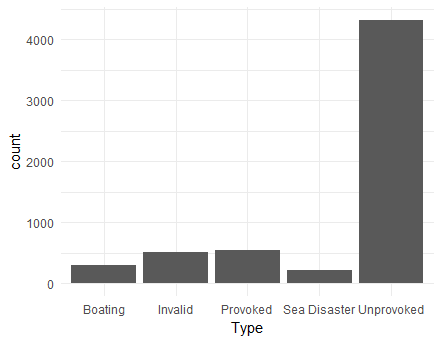
|  |  |  |
| --- | --- | --- |
| Case Number | Texto | Numero del caso, construido a partir de la fecha de ocurrencia del hecho |
| Date | Texto | Fecha de ocurrencia del hecho |
| Year | Numerico Entero | Año de ocurrencia del hecho |
| Type | Categorico (5) | Tipo de hecho del accidente |
| Country | Texto | Pais de ocurrencia del hecho |
| Area | Texto | Area del país de ocurrencia del hecho |
| Location | Texto | Lugar dentro del área de ocurrencia del hecho |
| Activity | Texto | Actividad realizada en el momento del hecho |
| Name | Texto | Nombre de la persona afectada |
| Sex | Categorica (5) | Sexo de la persona afectada |
| Age | Texto | Edad discretizada de la persona afectada |
| Injury | Texto | Lesión resultante del ataque |
| Fatal(Y/N) | Categorica (4) | Si el ataque resultó en muerte |
| Time | Texto | Horario del ataque |
| Species | Texto | Especie de tiburón involucrada |
| Investigator or Source | Texto | Fuente que registró el ataque |
| pdf | Texto | Nombre del archivo para descargar información del ataque |
| href | Texto | URL del archivo de ataque para descargar |
| original order | Numerico entero | Id del ataque |

La única variable numérica es la variable Year, que sin realizar ningún tipo de transformación tiene la siguiente distribución:



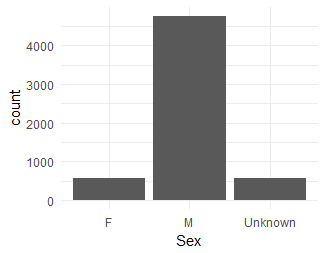
Para el preprocesamiento se trabajará sobre sus valores extremos y el valor cero como se ve en el gráfico

La variable categorica Type esta distribuida de la siguiente manera:



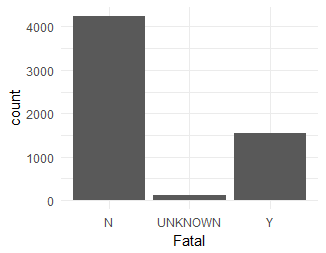
Claramente se observa que la amplia mayoría de los accidentes fueron no provocados.

La siguiente variable categórica Sex, poseía valores que no eran ni M ni F, esos valores fueron sustituidos por “unknown”:



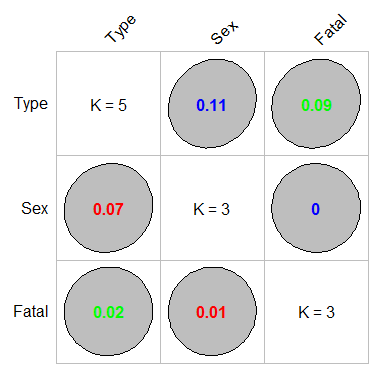
Nuevamente, se observa un amplio predominio, en este caso, el sexo masculino en la mayoría de los accidentes reportados.

La variable Fatal poseía valores que no eran ni Y ni N, se creó un nivel mas para esos valores, el valor “unknown”, pero en este caso, esos valores son infimos:



El análisis de otras variables que podrían ser categóricas, pero son texto, no tiene sentido alguno, ya que por ejemplo, la variable Country posee 174 valores, muy alta cardinalidad para analizarla y graficarla.

Al disponer solamente de una variable numérica, no podemos analizar su correlacion con otra variable numérica, pero si, lo que se realizó, fue analizar la asociacion de las variables categóricas mediante los estadísticos de Goodman-Kruskal, obteniendo la métrica tau (τ) (también, se podría haber utilizado el test de chi cuadrado):



Cuando el valor tau entre 2 variables es 1, significa que hay una asociación alta entre esas variables, caso contrario, el valor 0 indica la nula asociación entre esas variables, el valor K de la diagonal, denota la cantidad de valores de esa variable. Cabe mencionar que el valor tau es bidireccional, por ejemplo, Sex explicada por Type tiene un valor de 0.11, pero Type explicada por Sex tiene un valor de 0.07. Claramente se puede observar que no hay asociación entre las variables categóricas, ninguna de las 3 variables es explicada por ninguna otra variable.

Preprocesamiento:

Las siguientes variables fueron eliminadas al final del proceso, ya que no aportaban al análisis, poseían un alto nivel de cardinalidad o eran texto puro:

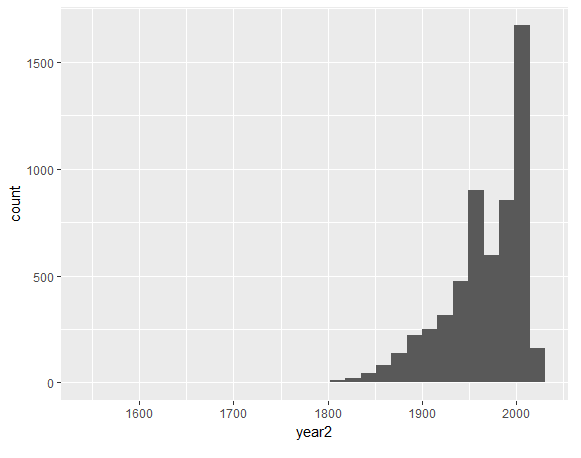
**Href, Href\_formula, pdf, investigator or source, species, name, injury, location, area, activity, case number, date, original order**

La variable Specie es candidata a realizarle text minning y transformarla en categórica, pero por una cuestión de tiempos no fue posible.

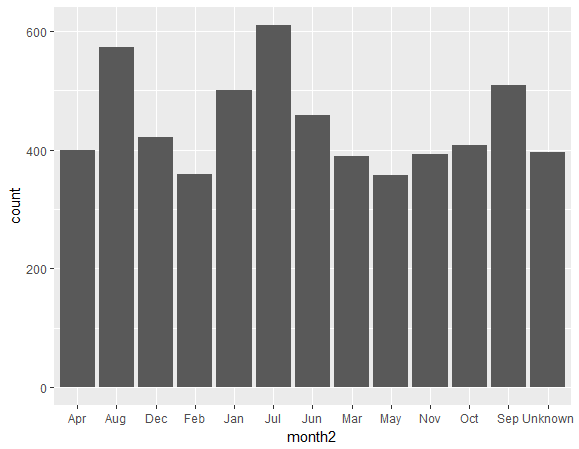
Sex: se reemplazo con el valor “unknown”, los valores que no eran ni M ni F.

Fatal: se reemplazaron los valores faltante con el valor “unknown”

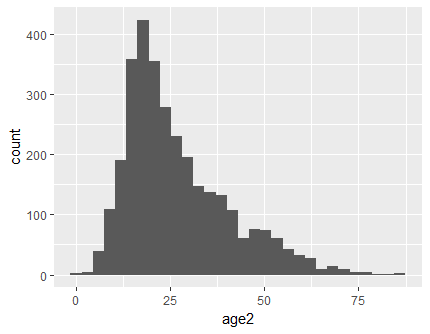
Year: solo nos quedamos con los registros donde Year era mayor a 1500, y no nos basamos en la variable Year original, sino que se creó una nueva Year donde se obtuvo el año a partir del parseo de case number, ya que se notaron algunas diferencias y el año dentro de case number era mas acertado. La distribución de year, luego del proceso, fue la siguiente:



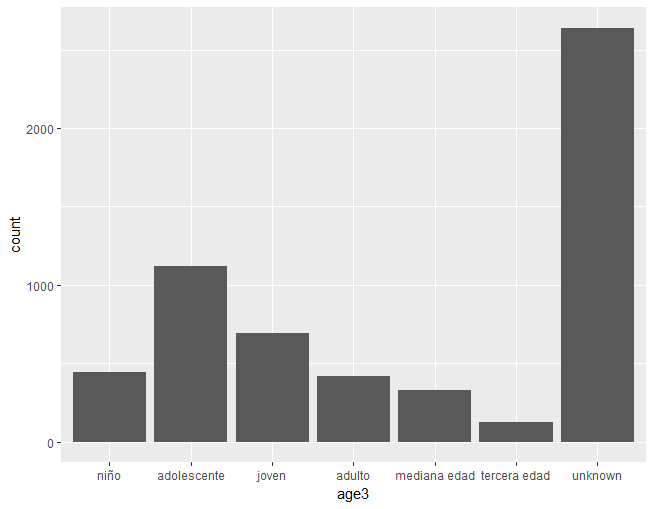
Month: se creó una variable month, a partir del parseo del mes en la variable case number, en los casos donde el mes no estaba disponible, se creó un valor “unknown”:



Age: la variable age original disponía no solo de valores numéricos, sino de otras pseudo discretizaciones, años de nacimiento, etc. Solo el 55% de los valores eran numéricos. Primero solo se extrajeron los valores numéricos, quedándonos un 45% de valores NA. Se intentó reemplazar los missing values por la media o mediana, pero dado el alto porcentaje de missing values, la distribución era altamente distorsionada. La distribución sin reemplazo era la siguiente:

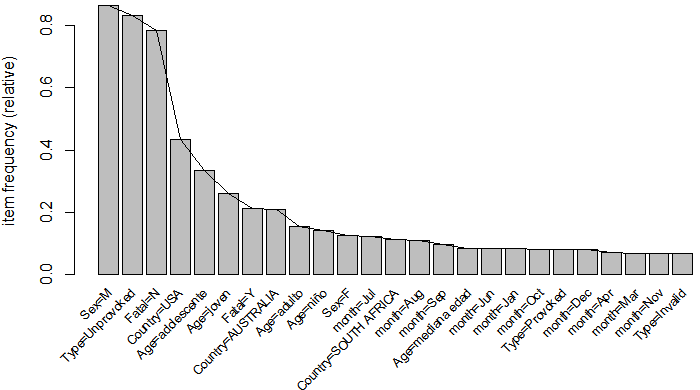


Distribución “tipica” de edad. Pero necesitábamos discretizar esta variable para poder ser utilizada como variable para el tipo de clase transactions para la minería de reglas de asociación, por ende, se realizó una discretizacion basada en clustering, específicamente kmeans, donde la discretizacion respetó lo mas posible a la distribución original. Los bins o segmentos fueron renombrados según un criterio para marcar las etapas de la edad y se agregó un valor (“unknown”) para los valores faltantes:



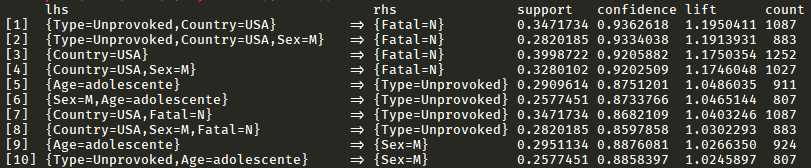
Una vez tenida la estructura preparada se procedió a crear las transacciones a partir del dataset. Todas las variables eran del tipo categóricas, por lo que la creación de las transacciones finalizó con 5772 transacciones y 447 columnas, ya que se crea una columna binaria por cada valor de la variable categórica, y en este caso, los países eran demasiados. Al evaluar los distintos ítemsets frecuentes dado un soporte de 8%, ya que buscábamos un soporte interesante de por lo menos 500 transacciones / 5772 = 0.08, notábamos la preponderancia de la edad con valor “unknown”, todas las reglas ordernadas por confidence y lift, resultaban como consecuenta la edad “unknown”. Por lo tanto se decidió procesar nuevamente sin tener en cuenta la edad “unknown”, solo con la discretizacion efectuada y evitando este valor.

Como primera aproximación a los ítems (top 25) de las transacciones, se observa:



Que el mayor soporte está dado por masculinos, este dato ya fue visto por la distribución de la variable Sex. Luego, el tipo de accidente Unprovoked, predomina también en los datos. Y por último que no haya sido fatal, hay un gran predominio de estos valores en las transacciones, que seguramente se notaran en las reglas mineadas. También puede observarse a USA y Australia como los 2 países con mas accidentes reportados.

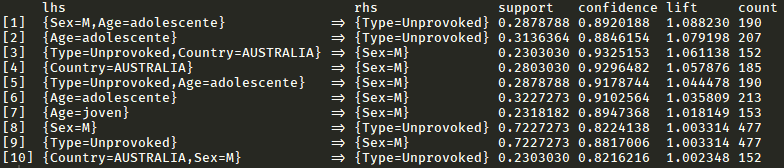
Al minar las primeras reglas con un soporte de 25%, ordenadas por lift, se obtuvieron reglas que daban como consecuente al accidente como no fatal, de sexo masculino y del tipo no provocado:

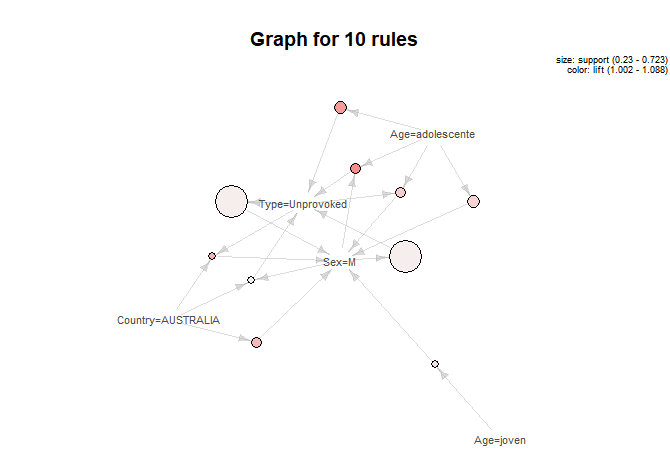


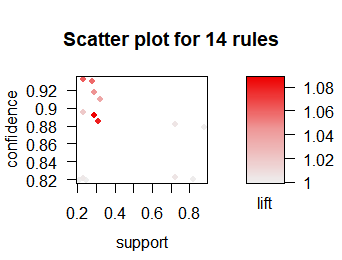
Se puede observar que los accidentes no fatales, no son provocados y son originarios de USA. Los adolescentes masculinos están involucrados en gran parte de los accidentes.

Dada la mayor cantidad de transacciones con accidentes no fatales, masculinos y no provocados era un comportamiento esperable el resultado de estas reglas, por lo que se decidió analizar las subpoblaciones del dataset, para obtener reglas con mejor conocimiento.

**Accidentes fatales**: soporte de 20%, se encontraron 14 reglas, analizamos las mejores 10 según lift

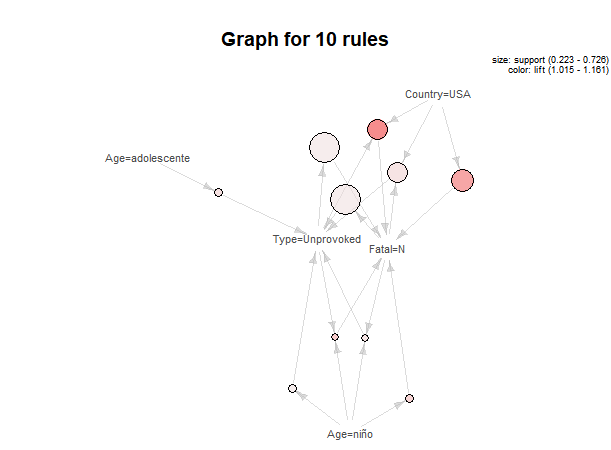


El siguiente grafo de las reglas nos permite observar las reglas encontradas, pero de una manera mucho mas intuititiva. El tamaño de los circulos indica el soporte y la intensidad de color, indica mayor lift. Todas las reglas tienen como consecuente a un hombre o un accidente no provocado. Podemos identificar a Australia como el país donde hay mas accidentes fatales, y vemos que esta involucrado en 3 reglas (3 flechas salen de su nodo). Otra observación que podemos hacer es que los adolescentes masculinos estan involucrados en los accidentes fatales, como asi tambien los jovenes, pero en menor medida. Podria ser que este relacionado al surf.

De las 14 reglas encontradas para accidentes fatales que se observan en el gráfico (las mas coloradas, son las de mayor lift). Vemos que, como métrica, el soporte no es bueno, ya que, aun teniendo alto soporte, esa regla puede tener un confidence muy bajo y por ende, lift también bajo.

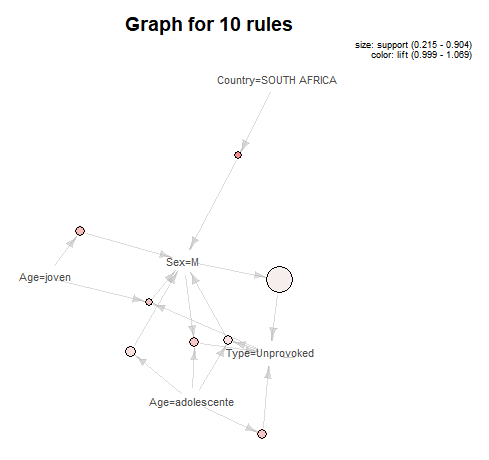
**Mujeres:** soporte del 20%

Respecto a las mujeres como dato interesante podemos observar mediante el siguiente gráfico que los accidentes de las mujeres ocurren a adolescentes y niñas, teniendo preponderancia el origen de estos accidentes en USA.



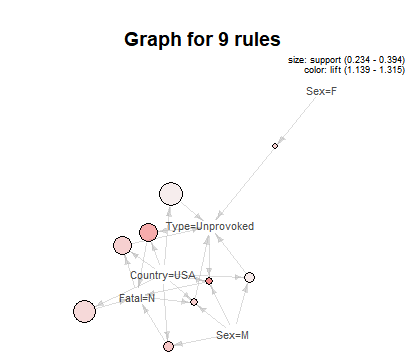
**No fatales fuera de USA**: soporte 20%

Al analizar los accidentes no fatales fuera de los Estados Unidos, detectamos a Sudafrica como el país donde ocurren los accidentes no fatales en hombres. Se puede ver en el grafico, un soporte bajo de la regla, pero con alto lift. Otro dato no llamativo es que las victimas son adolescentes y jóvenes.



**Tercera edad**: soporte 20%

En esta subpoblación puede observarse como antecedentes de las reglas a USA pero también aparece el sexo femenino, podría deducirse que en el rango etareo de tercera edad, las mujeres concurren mas al mar en proporción a los hombres que en otros rangos etareos.



Toda la descripción del dataset y las técnicas que utilizaron, acá pueden utilizar tablas y gráficos que ayuden a explicar los datos pero no se incluye nada de resultados. Es solo el análisis exploratorio.

Acá también se describe todo el workflow de preprocesamientos y todas las transformaciones.

Además, en está sección también se explican los experimentos que se realizaron. Ejemplo:

Se realizó una reducción de dimensionalidad con el método X y quedaron estas N variables, la variable objetivo utilizada fue la generada con el criterio de referencia dada en el enunciado del TP. Se corrió con esos datos un ajuste con J48 y se evaluó con la métrica Y.

3 Resultados obtenidos

Los resultados de la configuración de preprocesamiento X con las N variables se muestran en la tabla XX.....

Referencias

Hryniewicz, O. (2006). Goodman–Kruskal γ (tau) measure of dependence for fuzzy ordered categorical data. *Computational Statistics & Data Analysis*, *51*(1), 323-334.