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

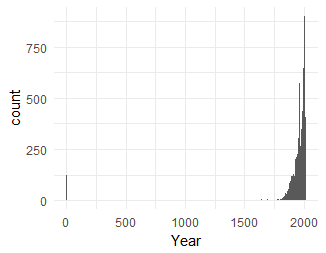
Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta es la introducción. Esta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

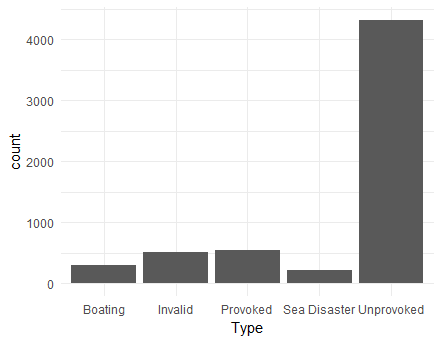
|  |  |  |
| --- | --- | --- |
| Case Number | Texto | Numero del caso, construido a partir de la fecha de ocurrencia del hecho |
| Date | Texto | Fecha de ocurrencia del hecho |
| Year | Numerico Entero | Año de ocurrencia del hecho |
| Type | Categorico (5) | Tipo de hecho del accidente |
| Country | 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:



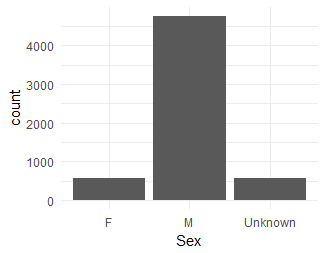
Para el preprocesamiento se trabajará sobre sus valores extremos y el valor cero como se ve en el gráfico

La variable categorica type esta distribuida de la siguiente manera:



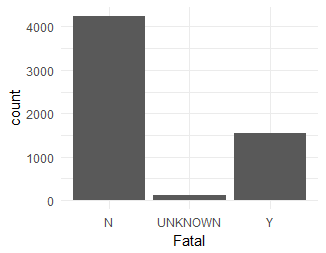
Claramente se observa que la amplia mayoría de los accidentes fueron no provocados.

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



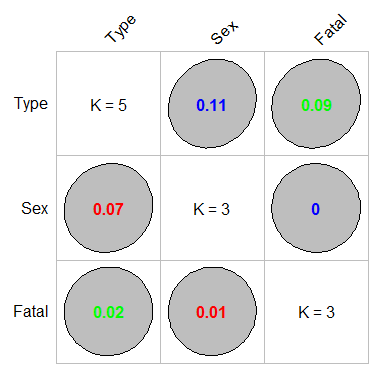
Nuevamente, se observa un amplio predominio, en este caso, el sexo masculino en los accidentes reportados.

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, 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 vale 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 tiene doble dirección, 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 alguna otra variable

Preprocesamiento:

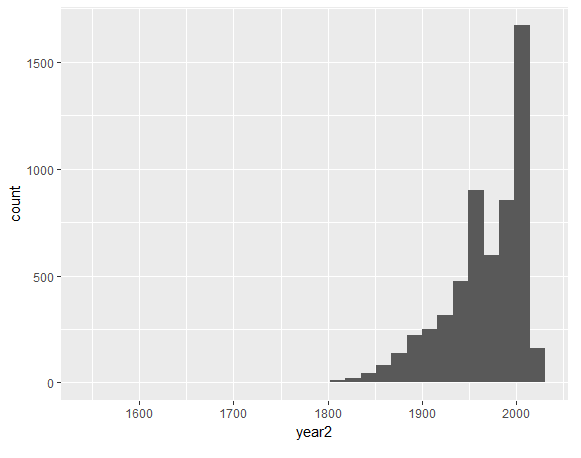
Las siguiente variables fueron eliminadas al final del proceso, ya que no aportaban al análisis, poseían un alto nivel de cardinalidad o eran texto puro:

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

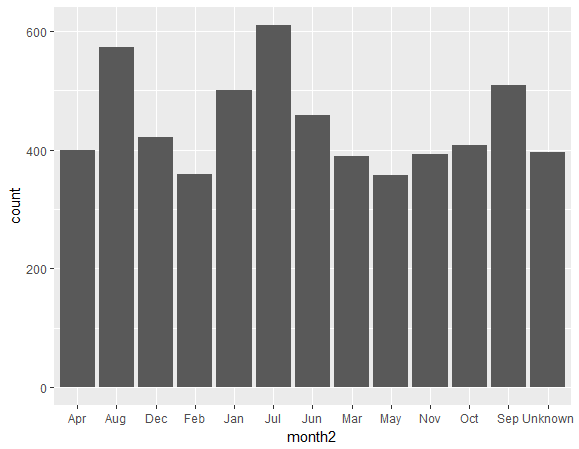
Sex: se reemplazo con el valor “unknown”, los valores que no eran ni M ni F.

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

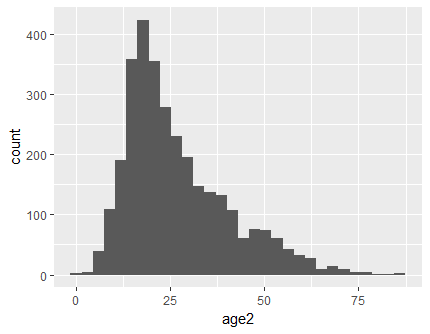
Year: solo nos quedamos con los registros donde Year era mayor a 1500, y no nos basamos en la variable Year original, sino que se creó una nueva Year donde se obtuvo el año a partir del parseo de case number, ya que se notaron algunas diferencias y el año dentro de case number era mas acertado. La distribución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



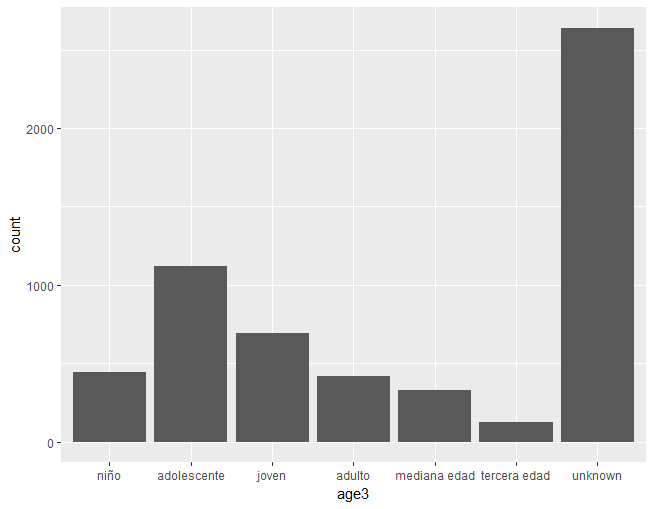
Month: se creó una variable month, a partir del parseo del mes en la variable case number, en los casos donde el mes no estaba disponible, se creó un valor “unknown”:



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:



La estructura del dataset no es optima para realizar un análisis de reglas de asociación

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 measure of dependence for fuzzy ordered categorical data. *Computational Statistics & Data Analysis*, *51*(1), 323-334.