Oscar Pulido

Aplicaciones del aprendizaje de máquinas en Actuaría y Finanzas

Primera versión Fase de Comprensión de los datos para el problema de estimación de provisiones

Control de versiones

Versión	Autor	Fecha	Aprobado por
0.1	Oscar Pulido	30/08/2023	Francisco Gómez

En este documento se presentará la primera versión de la fase de entendimiento de los datos. Primero, se presentará un breve apartado de recolección inicial de los datos, posteriormente se mostrarán los reportes de descripción de los datos, de exploración de los datos y de calidad de los datos.

Recolección inicial de los datos

La fuente de los datos para este proyecto es propia y proviene del histórico de siniestros de distintas compañías estadounidenses para el ramo de Auto Comercial, se puede encontrar en el siguiente enlace: https://www.casact.org/sites/default/files/2021-04/comauto_pos.csv. Esta fuente de información contiene información sobre la fecha ocurrencia de siniestros, fecha de desarrollo de los pagos, montos pagados, información de la prima ganada y monto de reserva contabilizada para un conjunto de compañías. Se ha determinado que con los datos existentes es suficiente para seguir con la parte restante del proyecto.

Reporte de descripción de los datos

Formato: Los datos están en formato CSV y son de acceso libre.

Descripción de variables: A continuación, se muestra la composición por variables del conjunto de datos.

GRCODE: Código NAIC de la empresa (incluidos grupos de aseguradoras y aseguradoras individuales)

GRNAME: Nombre de la empresa NAIC (incluidos grupos de aseguradoras y aseguradoras individuales)

AccidentYear: Año del accidente (1988-1997)

DevelopmentYear: Año de desarrollo (1988 a 1997)

DevelopmentLag: Año de desarrollo (AY-1987 + DY-1987 - 1)

IncurLoss_: Siniestros ocurridos y gastos imputados declarados al final del año

CumPaidLoss_: Pérdidas pagadas acumuladas y gastos asignados al final del año

BulkLoss_: Reservas a granel e IBNR sobre pérdidas netas y gastos de defensa y contención de costes comunicados al final del año

PostedReserve97_: Reservas contabilizadas en el año 1997 tomadas del Anexo de Suscripción e Inversión - Parte 2A, incluidas las pérdidas netas no pagadas y los gastos de ajuste de pérdidas no pagados

EarnedPremDIR_: Primas devengadas en el ejercicio - directas y asumidas

EarnedPremCeded_: Primas devengadas en el ejercicio - cedidas

EarnedPremNet_: Primas devengadas en el ejercicio - netas

Individual 1 indica una entidad individual, 0 indica un asegurador de grupo

Reporte de campos relevantes: Las columnas de la base de datos que parecen de mayor relevancia para el ejercicio son: Año del siniestro (AccidentYear), Año de desarrollo (DevelopmentYear, Pérdida incurrida (IncurLoss_C) y Perdida pagada acumulada (CumPaidLoss_C), ya que por lo general se suelen usar para la estimación de la reserva IBNR usando el método de Chain-Ladder, pues permiten generar el triángulo run off de pagos acumulados. Asimismo, hay un campo identificador de compañía, lo cual permite ver que se puede calcular la reserva para las distintas compañías (GRNAME).

Esta información parece la más relevante, sin embargo, es posible que las columnas restantes puedan complementar el análisis, de manera que, no se descartan variables por el momento. Por ejemplo, las columnas asociadas a las primas se usan por lo general para usar un método diferente a Chain-Ladder para la proyección de pagos, estos son los métodos Loss Ratio y Bornhuetter-Ferguson que no aplican por el momento. Asimismo, vale mencionar que este ejercicio no requiere la presencia de variables tipo “Primary Key

Análisis descriptivo de los datos: El análisis exploratorio de los datos se hizo por medio de RStudio y muestra que el conjunto de datos de siniestros del ramo de Auto Comercial para las compañías cuenta con 15800 observaciones y 13 variables. Las variables presentan los siguientes formatos:

Variable	Formato
GRCODE_format	numeric
GRNAME_format	character
AccidentYear_format	numeric
DevelopmentYear_format	numeric
DevelopmentLag_format	numeric
IncurLoss_C_format	numeric
CumPaidLoss_C_format	numeric
BulkLoss_C_format	numeric
EarnedPremDIR_C_format	numeric
EarnedPremCeded_C_format	numeric
EarnedPremNet_C_format	numeric
Single_format	numeric
PostedReserve97_C_format	numeric

Las compañías en cuestión son 157, acá se muestran 10:

British Amer Ins Co
Capacity Ins Co
First American Specialty Ins Co
Center Mut Ins Co
Sirius Amer Ins Co
Daily Underwriters Of Amer
Health Care Ind Inc
Service Ins Co
Nonprofits Ins Assn Interins
Governmental Interins Exch

El conjunto de datos no presenta codificaciones especiales que interfieran con el análisis, aunque se presenta una segmentación por grupos que puede permitir estimar la reserva IBNR y aplicar el método Chain-Ladder.

A continuación, se muestra un análisis de estadísticas descriptivas del conjunto de datos por variable:

```
     GRCODE         GRNAME          AccidentYear  DevelopmentYear DevelopmentLag  IncurLoss_C
Min.   :  266   Length:15800       Min.   :1988   Min.   :1988    Min.   : 1.0   Min.   :  -253
1st Qu.: 8281   Class :character   1st Qu.:1990   1st Qu.:1994    1st Qu.: 3.0   1st Qu.:     5
Median :15303   Mode  :character   Median :1992   Median :1997    Median : 5.5   Median :   315
Mean   :18655                      Mean   :1992   Mean   :1997    Mean   : 5.5   Mean   :  5156
3rd Qu.:28886                      3rd Qu.:1995   3rd Qu.:2000    3rd Qu.: 8.0   3rd Qu.:  1965
Max.   :44598                      Max.   :1997   Max.   :2006    Max.   :10.0   Max.   :256434
CumPaidLoss_C         BulkLoss_C       EarnedPremDIR_C    EarnedPremCeded_C EarnedPremNet_C
Min.   :  -253.00   Min.   : -1395.0   Min.   :  -445.0   Min.   :-3329.0   Min.   :  -128.0
1st Qu.:     2.75   1st Qu.:     0.0   1st Qu.:    42.8   1st Qu.:    0.0   1st Qu.:    19.8
Median :   229.00   Median :     0.0   Median :   732.0   Median :   73.0   Median :   549.5
Mean   :  4190.97   Mean   :   357.4   Mean   :  8985.4   Mean   : 1508.9   Mean   :  7476.6
3rd Qu.:  1535.25   3rd Qu.:    26.0   3rd Qu.:  3991.5   3rd Qu.:  656.2   3rd Qu.:  2810.5
Max.   :244280.00   Max.   :105923.0   Max.   :412331.0   Max.   :60233.0   Max.   :406609.0
    Single       PostedReserve97_C
Min.   :0.0000   Min.   :     0
1st Qu.:0.0000   1st Qu.:   121
Median :1.0000   Median :   938
Mean   :0.7215   Mean   : 10815
3rd Qu.:1.0000   3rd Qu.:  4711
Max.   :1.0000   Max.   :360657
```

Este análisis descriptivo permite entender que los siniestros más antiguos ocurren en 2018, y el desarrollo máximo va hasta 2006, para propósitos de este reporte, se asumirá que los montos de pagos conocidos van hasta 1997 por lo tanto, se buscará pronosticar desarrollos mayores a 1997 y hasta 2006.

Si se ven las variables de interés, es decir, IncurLoss_C y CumPaidLoss_C, se pueden evidenciar algunos problemas de calidad de la información iniciales, aunque este punto se abordará más en detalle en el informe de calidad, se puede evidenciar la presencia de valores atípicos al ver la diferencia entre los montos máximos de cada variable y la media. Asimismo, es posible ver valores negativos, los cuáles pueden tener un tratamiento especial, como se verá en el reporte de calidad más adelante.

En conclusión, el reporte permite ver que los datos desde un enfoque descriptivo son aptos para el cálculo de la reserva IBNR usando el método Chain-Ladder.

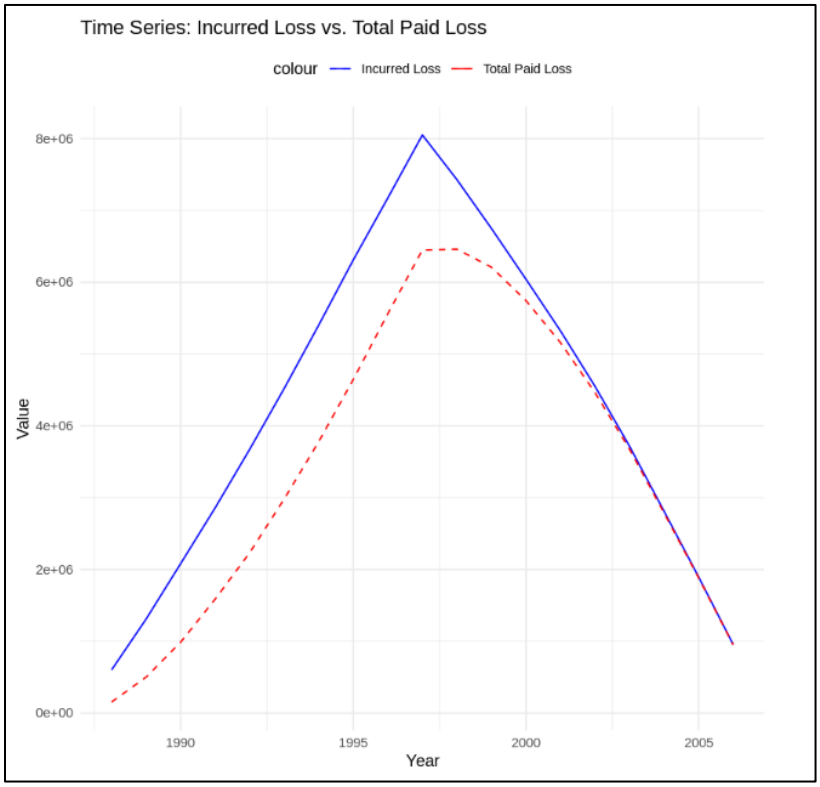
Reporte de exploración de los datos

Dada la naturaleza de los datos, y del ejercicio de este proyecto, no se requiere un ejercicio con gran robustez, ya que básicamente las variables de mayor interés se comportan como series de tiempo con temporalidad para la ocurrencia de los siniestros y el desarrollo de los pagos.

Análisis exploratorio: Las variables de interés permiten ver la magnitud que tendrá la reserva a estimar. A continuación, se muestra un resumen de los montos incurridos y pagados por año de desarrollo, lo cual permite tener un punto de partida para la construcción de los triángulos run-off

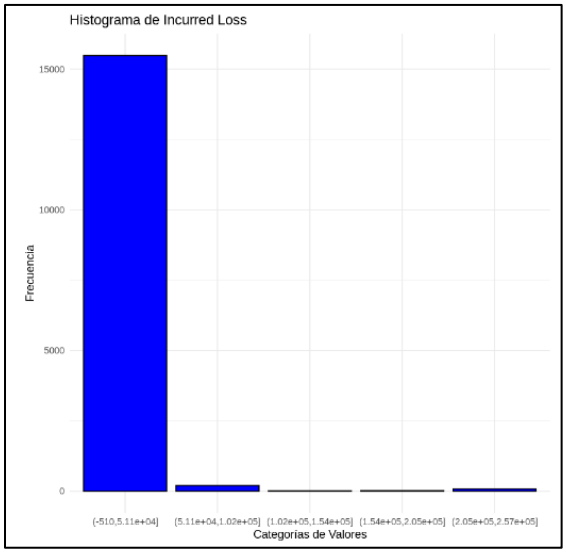
DevelopmentYear	Incurred_Loss	Cummulate_Loss
<dbl>	<dbl>	<dbl>
1988	600839	154058
1989	1309342	500772
1990	2081606	987503
1991	2859155	1587324
1992	3674297	2234030
1993	4522315	2982593
1994	5404961	3781285
1995	6311670	4646588
1996	7174345	5562037
1997	8051238	6449562
1998	7429344	6460829
1999	6749233	6209861
2000	6042077	5744583
2001	5322303	5158822
2002	4550222	4462499
2003	3715769	3674079
2004	2810161	2786136
2005	1895680	1883709
2006	956165	951111

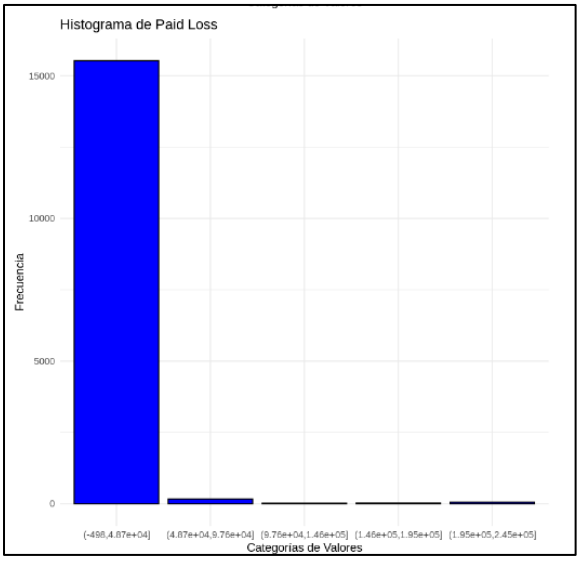
El cuadro permite ver que los años de desarrollo se concentran hacia los años de 1994 hasta 1999, lo cual indica que los pagos por lo general no son inmediatos. Esto se ve mejor con una gráfica de la perdida incurrida y la perdida pagada con respecto al año de desarrollo, la cual permite ver que los mayores montos incurridos se tienen en 1996, sin embargo, no es un monto mucho mayor en comparación con los otros años, a excepción de los años de 1988 a 1992, donde los montos incurridos son menores, esto se puede explicar porque en general los pagos no son inmediatos:



Una conclusión que se puede obtener de estos gráficos es que en general, las proyecciones se basarán en la información disponible, por lo tanto, aquellos años de desarrollo con mayores montos concentrarán la mayor cantidad de proyección, es decir, para estos segmentos en particular se espera aprovisionar mayor cantidad de reserva.

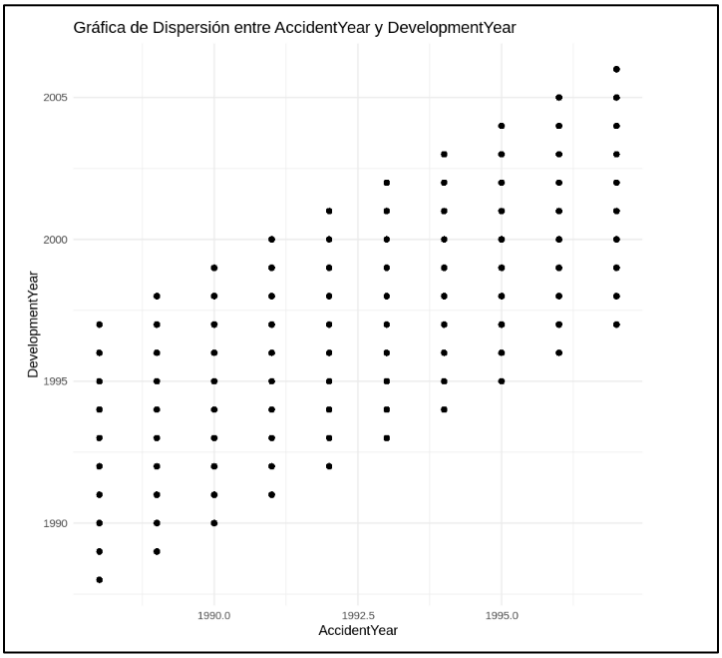
Ahora bien, un histograma permite entender la composición de la información:





Las gráficas anteriores permiten concluir que la distribución de las observaciones se concentra en aquellos siniestros de monto bajo, por lo tanto, los otros se consideran atípicos.

Otra gráfica de interés muestra la relación entre los históricos de ocurrencia y desarrollo para este las compañías, esto se puede ver de la siguiente manera:



Esto permite ver que el desarrollo de los siniestros tiene una duración similar independiente del año de ocurrencia. Esto parece una característica deseable, ya que se puede pensar en una mejor proyección basada en el método de Chain-Ladder cuando el desarrollo de los siniestros es similar. Finalmente, un análisis exploratorio de la estructura de triángulo de los datos para la variable pagos acumulados de la compañía Public Underwritters Grp luce de la siguiente manera:

	DevelopmentYear									
AccidentYear	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
1988	0	0	0	0	0	0	0	0	0	0
1989	NA	6	20	21	23	24	24	24	24	24
1990	NA	NA	52	95	105	117	128	128	128	128
1991	NA	NA	NA	127	258	301	332	372	376	377
1992	NA	NA	NA	NA	120	225	277	339	361	370
1993	NA	NA	NA	NA	NA	427	1005	1111	1263	1306
1994	NA	NA	NA	NA	NA	NA	584	1511	1716	1817
1995	NA	NA	NA	NA	NA	NA	NA	369	741	931
1996	NA	NA	NA	NA	NA	NA	NA	NA	294	595
1997	NA	NA	NA	NA	NA	NA	NA	NA	NA	312

Reporte de calidad de los datos

Dada la fuente de origen de los datos, se sabe que la base de datos contó con un proceso de limpieza y verificación de la calidad previo.

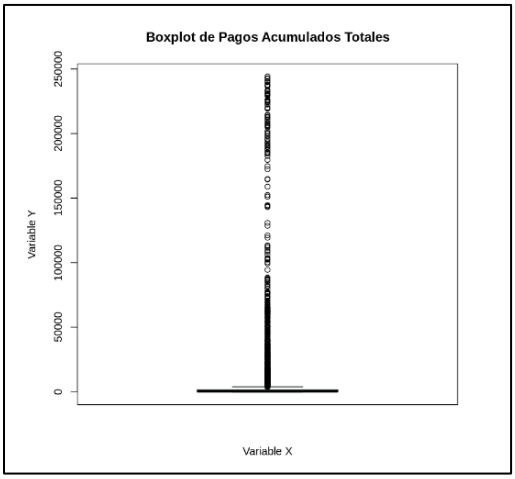
Análisis de valores nulos: El análisis de valores nulos de los datos permite ver que la calidad de los datos parece estar garantizada en ese sentido, dado que no se evidencian valores nulos:

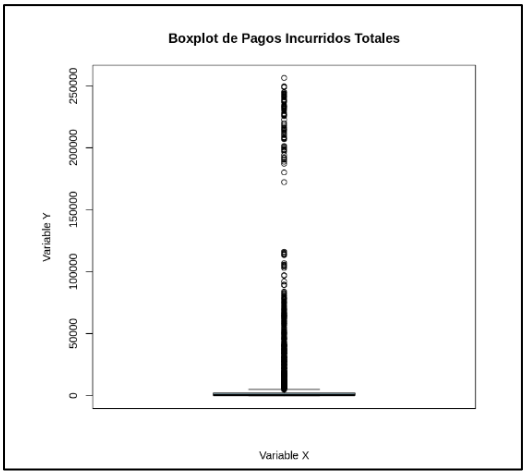
	Variable	ValoresNulos
GRCODE	GRCODE	0
GRNAME	GRNAME	0
AccidentYear	AccidentYear	0
DevelopmentYear	DevelopmentYear	0
DevelopmentLag	DevelopmentLag	0
IncurLoss_C	IncurLoss_C	0
CumPaidLoss_C	CumPaidLoss_C	0
BulkLoss_C	BulkLoss_C	0
EarnedPremDIR_C	EarnedPremDIR_C	0
EarnedPremCeded_C	EarnedPremCeded_C	0
EarnedPremNet_C	EarnedPremNet_C	0
Single	Single	0
PostedReserve97_C	PostedReserve97_C	0

Análisis de errores en los datos o atípicos: Los errores tipográficos aplicarían a las variables numéricas, ya que las categóricas no son relevantes dentro del análisis, pues indican una segmentación por grupos. En ese sentido, interesa buscar aquellos valores que sean negativos o que sean muy altos, dado que indicarían siniestros atípicos, los cuales podrían ser excluidos al momento de estimar la reserva IBNR. Para ello vale la pena revisar nuevamente el cuadro de estadísticas descriptivas:

```
     GRCODE         GRNAME          AccidentYear  DevelopmentYear DevelopmentLag  IncurLoss_C
 Min.   :  266   Length:15800       Min.   :1988   Min.   :1988    Min.   : 1.0   Min.   :  -253
 1st Qu.: 8281   Class :character   1st Qu.:1990   1st Qu.:1994    1st Qu.: 3.0   1st Qu.:     5
 Median :15303   Mode  :character   Median :1992   Median :1997    Median : 5.5   Median :   315
 Mean   :18655                      Mean   :1992   Mean   :1997    Mean   : 5.5   Mean   :  5156
 3rd Qu.:28886                      3rd Qu.:1995   3rd Qu.:2000    3rd Qu.: 8.0   3rd Qu.:  1965
 Max.   :44598                      Max.   :1997   Max.   :2006    Max.   :10.0   Max.   :256434
 CumPaidLoss_C        BulkLoss_C       EarnedPremDIR_C    EarnedPremCeded_C EarnedPremNet_C
 Min.   :  -253.00   Min.   : -1395.0   Min.   :  -445.0   Min.   :-3329.0   Min.   :  -128.0
 1st Qu.:     2.75   1st Qu.:     0.0   1st Qu.:    42.8   1st Qu.:    0.0   1st Qu.:    19.8
 Median :   229.00   Median :     0.0   Median :   732.0   Median :   73.0   Median :   549.5
 Mean   :  4190.97   Mean   :   357.4   Mean   :  8985.4   Mean   : 1508.9   Mean   :  7476.6
 3rd Qu.:  1535.25   3rd Qu.:    26.0   3rd Qu.:  3991.5   3rd Qu.:  656.2   3rd Qu.:  2810.5
 Max.   :244280.00   Max.   :105923.0   Max.   :412331.0   Max.   :60233.0   Max.   :406609.0
     Single       PostedReserve97_C
 Min.   :0.0000   Min.   :     0
 1st Qu.:0.0000   1st Qu.:   121
 Median :1.0000   Median :   938
 Mean   :0.7215   Mean   : 10815
 3rd Qu.:1.0000   3rd Qu.:  4711
 Max.   :1.0000   Max.   :360657
```

Esto permite ver la presencia de valores negativos y de valores lejanos a la media de las variables, por lo tanto, valdría la pena considerar excluir estos siniestros en la fase de preparación de los datos. El punto anterior se verifica con los siguientes boxplots:





En el caso de siniestros por montos negativos se puede pensar en casos específicos como cuando tras un siniestro, el beneficiario paga un deducible o copago mayor al monto del sinestro. Por otro lado, un siniestro atípico puede presentar un periodo de desarrollo muy largo que influya negativamente sobre el patrón de pagos obtenido con el triángulo run-off. Es por eso que se considera pertinente tratar de excluir los siniestros atípicos en las partes siguientes del proyecto.

Ahora bien, si se pasa analizar la calidad de los datos desde la perspectiva de los triángulos obtenidos.

	DevelopmentYear									
AccidentYear	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
1988	0	0	0	0	0	0	0	0	0	0
1989	NA	6	20	21	23	24	24	24	24	24
1990	NA	NA	52	95	105	117	128	128	128	128
1991	NA	NA	NA	127	258	301	332	372	376	377
1992	NA	NA	NA	NA	120	225	277	339	361	370
1993	NA	NA	NA	NA	NA	427	1005	1111	1263	1306
1994	NA	NA	NA	NA	NA	NA	584	1511	1716	1817
1995	NA	NA	NA	NA	NA	NA	NA	369	741	931
1996	NA	NA	NA	NA	NA	NA	NA	NA	294	595
1997	NA	NA	NA	NA	NA	NA	NA	NA	NA	312

La imagen anterior permite entender que en 1988 no se hicieron pagos por siniestros ocurridos en ese año. Por lo tanto, es válido afirmar que no hubo siniestros registrados en 1988 por Public Underwritters Grp. Como se puede evidenciar en uno de los gráficos anteriores, hubo compañías donde si hubo siniestros ocurridos en 1988 y pagados en los años posteriores. De cara a la siguiente fase de este proyecto, esto podría indicar que en el marco de la revisión de la calidad de datos es necesario excluir este año de accidentalidad, o aquellos años donde no hubo ocurrencia de siniestros, por lo tanto, no hubo pagos.

El análisis anterior permite plantear una pregunta sobre la existencia de pagos en cero para ciertas aseguradoras. Dentro de esta fase de calidad convendría ver si existen reportes en cero, que vienen siendo el equivalente a valores nulos, en este caso serían valores nulos que pueden ser eliminados en la fase siguiente. Un extracto de una tabla de valores en 0 muestra lo siguiente:

GRNAME	AccidentYear	DevelopmentYear	CumPaidLoss_C
<chr>	<dbl>	<dbl>	<dbl>
Public Underwriters Grp	1988	1988	0
Public Underwriters Grp	1988	1989	0
Public Underwriters Grp	1988	1990	0
Public Underwriters Grp	1988	1991	0
Public Underwriters Grp	1988	1992	0
Public Underwriters Grp	1988	1993	0
Public Underwriters Grp	1988	1994	0
Public Underwriters Grp	1988	1995	0
Public Underwriters Grp	1988	1996	0

Esto permite rastrear los datos con valor cero que se muestran el triángulo de pagos. El tratamiento de estos se mostrará en la fase siguiente del proyecto.