Resumen

Esta visualización a través de la información relativa a características propias de los asegurados, variables sociodemográficas e información referente a siniestros acaecidos en seguros de coches tiene como objetivo de identificación de patrones de comportamiento de distintos aspectos que conciernen al asegurado en su conducta frente al siniestro.

Datos básicos sobre la visualización

Título: Reclamación de seguros: Detención de fraudes

Descripción:

Respecto a los tipos de gráficas empleadas tenemos cuatro tipos. Para mostrar las dos cronologías se emplearán dos diagramas de líneas. Se emplearán dos mapas para localizar los estados donde se dieron de alta las pólizas y dónde se realizaron las reclamaciones. Por último, para mostrar los indicadores se utilizarán cinco diagramas de barras (divergentes, apiladas) y una vista de Gantt.

Respecto a la organización del contenido este se puede dividir principalmente en tres segmentos: el segmento superior contendrá toda la información relativa a las pólizas; el segundo segmento toda la información relativa a los asegurados; mientras el tercio restante contendrán la información de las reclamaciones.

Respecto a la interactividad se dispondrá de un filtro que nos permitirá seleccionar las reclamaciones que han sido fraudulentas o no, permitiéndonos ver los resultados reflejados en todas las gráficas de la visualización.

La visualización es accesible en la siguiente dirección:
https://public.tableau.com/app/profile/sonia.rodriguez.del.cerro.garcia/viz/srodriguezdelcerroPRACT2/Dashboard1?publish=yes.
Mientras que el proyecto del análisis de los datos se encuentra en el repositorio git:
https://github.com/soniauni/srodriguezdelcerro-PRACT2

Análisis del Trabajo Previo

Tras revisar los datos mencionados en la primera parte de la práctica, se ha decidido mantener la información que se había indicado

Utilizando un conjunto de datos proporcionado por Derrick Mwiti publicados en 31 Oct 2018 https://github.com/mwitiderrick.

La base de datos es un fichero de texto plano (.csv), fichero que se encuentra en el repositorio git con el nombre: insurance_claims.cvs, este formato permite trabajar con los datos en diferentes programas sin necesidad de realizar modificaciones. La base de datos contiene una muestra del total de siniestros ocurridos en un período concreto de tiempo de 1.000 expedientes. La base de datos recopilada 39 atributos con información relativa a características propias de los asegurados, variables sociodemográficas e información referente a siniestros acaecidos. Los atributos según su tipo de datos se dividen en dos tipos: 18 variables numéricas (dbl, int) y el resto cadenas.

Revisando las variables numéricas, se ha detectado que el atributo policy_number tiene más de 900 valores distintos y que no es un dato relevante para el estudio.
Con respecto a las variables categóricas, se ha identificado que algunas columnas de cadenas tienen muchos valores distintos (900+): incident_location, atributos irrelevantes como policy_csl. Además, se existen variables que dependen unas de otras y que no aportan información relevante para este estudio, Auto_make y auto_model, no es necesario saber el modelo del coche. Incident_type y collision_type, el tipo de colisión solo está informado según el tipo de incidente.
Se eliminan las siguientes estas columnas del conjunto de datos para mejorar la precisión del modelo: policy_number, incident_location, policy_csl, auto_model, collision_type
Las variables que se introducen en el conjunto final son las siguientes:

```
## Observations: 1,000
## Variables: 34
## $ months_as_customer          <int> 328, 228, 134, 256, 228, 256, 137, 165,...
## $ age                         <int> 48, 42, 29, 41, 44, 39, 34, 37, 33, 42,...
## $ policy_bind_date            <fct> 2014-10-17, 2006-06-27, 2000-09-06, 199...
## $ policy_state                <fct> OH, IN, OH, IL, IL, OH, IN, IL, IL, IL,...
## $ policy_deductable           <int> 1000, 2000, 2000, 2000, 1000, 1000, 100...
## $ policy_annual_premium       <dbl> 1406.91, 1197.22, 1413.14, 1415.74, 158...
## $ umbrella_limit              <int> 0, 5000000, 5000000, 6000000, 6000000, ...
## $ insured_zip                 <int> 466132, 468176, 430632, 608117, 610706,...
## $ insured_sex                 <fct> MALE, MALE, FEMALE, FEMALE, MALE, FEMAL...
## $ insured_education_level     <fct> MD, MD, PhD, PhD, Associate, PhD, PhD, ...
## $ insured_occupation          <fct> craft-repair, machine-op-inspct, sales,...
## $ insured_hobbies             <fct> sleeping, reading, board-games, board-g...
## $ insured_relationship        <fct> husband, other-relative, own-child, unm...
## $ capital.gains               <int> 53300, 0, 35100, 48900, 66000, 0, 0, 0,...
## $ capital.loss                <int> 0, 0, 0, -62400, -46000, 0, -77000, 0, ...
## $ incident_date               <fct> 2015-01-25, 2015-01-21, 2015-02-22, 201...
## $ incident_type               <fct> Single Vehicle Collision, Vehicle Theft...
## $ incident_severity           <fct> Major Damage, Minor Damage, Minor Damag...
## $ authorities_contacted       <fct> Police, Police, Police, Police, None, F...
## $ incident_state              <fct> SC, VA, NY, OH, NY, SC, NY, VA, WV, NC,...
## $ incident_city               <fct> Columbus, Riverwood, Columbus, Arlingto...
## $ incident_hour_of_the_day    <int> 5, 8, 7, 5, 20, 19, 0, 23, 21, 14, 22, ...
## $ number_of_vehicles_involved <int> 1, 1, 3, 1, 1, 3, 3, 3, 1, 1, 1, 3, 1, ...
## $ property_damage             <fct> YES, ?, NO, ?, NO, NO, ?, ?, NO, NO, YE...
## $ bodily_injuries             <int> 1, 0, 2, 1, 0, 0, 0, 2, 1, 2, 2, 1, 1, ...
## $ witnesses                   <int> 2, 0, 3, 2, 1, 2, 0, 2, 1, 1, 2, 2, 0, ...
## $ police_report_available     <fct> YES, ?, NO, NO, NO, NO, ?, YES, YES, ?,...
## $ total_claim_amount          <int> 71610, 5070, 34650, 63400, 6500, 64100,...
## $ injury_claim                <int> 6510, 780, 7700, 6340, 1300, 6410, 2145...
## $ property_claim              <int> 13020, 780, 3850, 6340, 650, 6410, 7150...
## $ vehicle_claim               <int> 52080, 3510, 23100, 50720, 4550, 51280,...
## $ auto_make                   <fct> Saab, Mercedes, Dodge, Chevrolet, Accur...
## $ auto_year                   <int> 2004, 2007, 2007, 2014, 2009, 2003, 201...
## $ fraud_reported              <fct> Y, Y, N, Y, N, Y, N, N, N, N, N, N, N, ...
```

El tratamiento de los datos se puede ver en el documento pdf srodriguezdelcerro-PRA2 y el código R con el mismo nombre que se ha subido a repositorio git.
El fichero final utilizada para la visualización, es el fichero también subido al repositorio git con el nombre: insurance_claims (insurance_claims)_insurance_claims

Se ha descartado analizar otros indicadores como los datos de vehículo (Auto Make, Auto Year) ya que se considera que independientemente de cómo sea el vehículo no afecta a la identificación de fraudes.
Referente a variables de incidentes también se han descartado en la visualización las variables de número de vehículos implicados (Number Of Vehicles Involved), testigos (Witnesses), el tipo (Incident Type), la severidad (Incident Severity), fecha (Incident Date) y ciudad del incidente (Incident City), para localizar los incidentes hemos utilizado el estado donde se produjo y en lugar de analizar las fechas de incidente se ha visualizado la hora del día en el que ocurrió.

En cuanto a variables relacionadas con los asegurados tampoco se han utilizado el código postal (Insured Zip), relación con el asegurado (Insured Relationship), ocupación (Insured Occupation), nivel educativo (Insured Education Level) o hobbies (Insured Hobbies).

No obstante, durante el desarrollo de proyecto final se ha realizado algunos cambios en la base de datos creando nuevos campos a partir de otros para poder ser visualizados correctamente, a lo largo del documento se hace referencia a estos.

Herramientas utilizadas

Para desarrollar este proyecto se ha empleado dos herramientas: R y Tableau.

Tableau es una herramienta de visualización de datos potente utilizada en el área de la Inteligencia de negocios. Permite generar visualizaciones o dashboards (cuadros de mando) que permiten explorar y analizar conjunto de datos de forma interactiva.
El uso de la herramienta es fácil ya que se maneja a través de menús e iteraciones de drag & drop. Además, una vez indicados los tipos de datos que contine cada columna de nuestra tabla, Tableau te indica cuales son las mejores maneras para representarla, y ayudar al usuario en el proceso de creación de la visualización de datos.
Tiene una parte gratuita Tableau Public que guardará todos los trabajos de una manera pública en los servidores de Tableau y será compartido entre todos los usuarios.

Para realizar el desarrollo de la visualización del proyecto final se utilizó Tableau Public. Una vez finalizado el proyecto final se publicó para que este libre para cualquier persona.

Inicialmente a través de R se exploró los datos de dataset con el propósito de analizar los tipos datos, la dimensión de dataset (número de instancias y columnas).

Decisiones de diseño

A continuación, se realiza la descripción técnica del proyecto.

En primer lugar, se ha analizado los datos del dataset con el propósito de identificar los campos de interés. Una vez que tenemos una idea sobre la información a tratar se ha planteado una pregunta ¿Que indicadores me pueden permitir detectar fraudes?

En segundo lugar, se ha realizado un análisis inicial de los datos con R. En este punto se ha estudiado los tipos de campos y cuáles de ellos son campos nulos o vacíos. Además, se ha analizado como están distribuidas los valores de los campos categóricos.

En tercer lugar, se han identificado cuatro indicadores de interés: características de las pólizas contratadas, de los asegurados, de las reclamaciones y de los incidentes.

En cuarto lugar, se ha analizado que gráficas seleccionar para representar los datos. Para mostrar las cronologías se ha seleccionado los diagramas de líneas. Para mostrar los indicadores se ha seleccionado diagramas de barras. Además, se han creado mapas ya que este permite una localización rápida.

En quinto lugar, se empleó la herramienta Tableau Public para importar los datos ya tratados.

Por último, se creó un Dashboard que llamaremos “Dashboard1” en el que se añadieron las hojas creadas.

Finalmente, con Tableau Public publicamos Dashboard “Dashboard1”. Esta publicación permite el acceso libre al Dashboard.

Visualización

La idea sobre el tema para desarrollar este proyecto surgió del trabajo previo descrito en la práctica 1, la cual sirvió como base para desarrollar las ideas de la visualización. Este trabajo ha profundizado más en mostrar los indicadores de las reclamaciones fraudulentas, datos de los clientes y datos de las pólizas tratadas.

Para este trabajo emplearemos un fondo blanco ya que nos permitirá mostrar un contraste mayor al enseñar las distintas gráficas. Respecto a los colores del resto de elementos estos tendrán la siguiente configuración

- Se empleará el color amarillo fuerte #b6992d para todos los textos.
- Para relacionar los datos relativos a reclamaciones fraudulentas o no, en la parte inferior izquierda se muestra una caja con los colores relacionados a cada tipo de reclamaciones. Para las reclamaciones N fraudulentas se usarán el color gris oscuro #767676 y para las reclamaciones S fraudulentas se usará el color rosa #BB7693.
- En el mapa que visualiza los estados de las pólizas (Policy State), se ha utilizado el color amarillo fuerte #b6992d.
- En la visualización de Gantt, donde se representas las horas del día en las que se producen mayor número de siniestros (Incident Hour Of The Day), se ha utilizado una paleta personalizada divergente con dos rangos. Donde los que tiene menor valor se identifican con el color naranja #FBB04E y los que tienen mayor valor se identifican con el color café #B66353
- Las categorías utilizadas en el mapa que visualiza la suma total de los importes de las reclamaciones de siniestros (SumaTotalClaim) por estados (Incident State) se ha utilizado la paleta “piedra Miller”. La leyenda tiene tres rangos “Menos de 5000000”, donde se utiliza el color naranja #FBB04E, para el rango “Menos de 10000000” se utiliza el color café #B66353 y para el rango “Mas de 10000000” se utiliza el color verde amarillo #BFBB60.

Como se pueden observar se han utilizado siempre colores de la paleta “piedra Miller”. Usando 6 colores (amarillo fuerte, gris oscuro, rosa, naranja, café y verde-amarillo) para representar tanto los títulos, filtros y leyendas.

Como fuente de letra se selecciona es Tableau Book y color amarillo fuerte #b6992d. Respecto a los textos del proyecto estos tendrán la siguiente configuración:

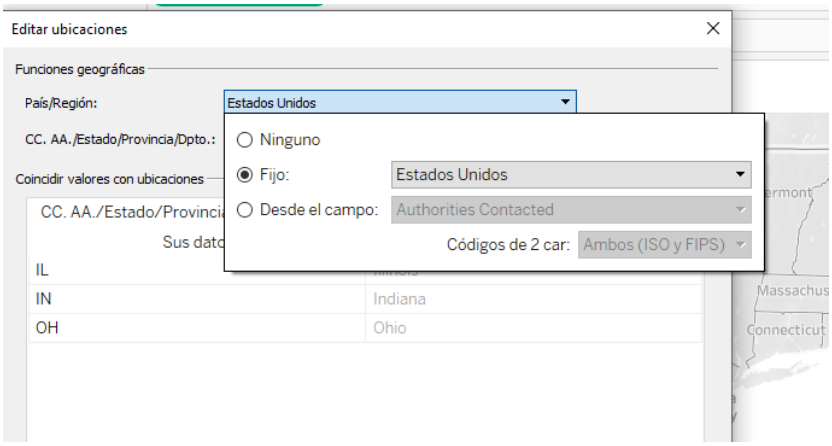
- El título del proyecto se mostrará centrado, en negrita (bold), tendrá un tamaño de 16.
- Los subtítulos se mostrarán centrados, en negrita (bold), tendrá un tamaño de 10.
- Los títulos específicos asociados a cada gráfica se mostrarán centrados, en negrita (bold), tendrá un tamaño de 10.
- Los textos de cada una de las gráficas se mostrarán en negrita (bold) y con un tamaño de 8.
- En las gráficas de barras el texto de la coordenada x se mostrará tendrá un tamaño de 8.
- Las descripciones emergentes de cada gráfica se mostrarán en negrita (bold) y con un tamaño de 8.

El diseño de la visualización de datos divide el contenido de forma horizontal en cuatro segmentos:

- En la parte superior se mostrará el título del proyecto de visualización y en la parte superior izquierda se muestra el filtro para realizar operaciones de selección de reclamaciones fraudulentas o no.

- En el siguiente segmento se tratarán los indicadores relacionados con las pólizas.

- En el primer gráfico podremos visualizar el número de pólizas dadas de alta por año (Policy Bind Date) se mostrarán cronológicamente desde 1990 hasta 2014, mediante diagrama de líneas. El indicador tenía el formato de fecha y hemos creado un nuevo campo, usando la opción crear “fecha personalizada” solo con el año de contratación, Policy Bind Date (Años), para la visualización. En la descripción emergente se podrá ver el año de contratación y el número de pólizas.
- A continuación, se informará del número total de pólizas tratadas.
- En el siguiente gráfico, se mostrarán el indicador del importe de primas anuales (Policy Annual Premium), se mostrarán cronológicamente por año (Policy Bind Date) desde 1990 hasta 2014 mediante un diagrama de líneas. En el título del gráfico se muestra SUMA(Policy Annual Premium) de los siniestros fraudulentos y de los que no. En la descripción emergente se podrá ver el año de contratación y el importe anual de las primas.
- Por último, se muestra un mapa con la localización (estado) de las pólizas (Policy State). Para que reconociera los estados que estamos tratando: IL (Illinois), IN (Indiana) y OH (Ohio), se ha tenido que editar las ubicaciones, para que considerara las latitudes y longitudes generadas a través del campo geográfico (Policy State) en el país de EEUU. En la descripción emergente se podrá ver el estado de contratación y el número de pólizas. En el gráfico del mapa el usuario podrá realizar acciones típicas como: zoom (aumento y disminución), mover el mapa, selección de un país, selección de áreas y búsquedas por nombre de país.



Para la representación de datos que muestran tendencias a lo largo de un periodo de tiempo hemos utilizado gráficos de líneas, que muestra como un dato cuantitativo en este caso REC(insurance_claim) y Policy Annual Premium en un intervalo de tiempo Policy Bind Date (Años) y que nos permite tener una visión global del intervalo de tiempo. Los gráficos de líneas se representan mediante dos ejes, el X y el Y. En el X colocamos los intervalos temporales, y en el eje Y los datos cuantitativos que tenemos que mostrar.

Para la representación del dato geográfico Policy State hemos utilizado un mapa político que muestra las fronteras entre los estados, para posicionarnos más fácilmente donde se dieron de alta las pólizas que estamos tratando.

- A continuación, en el siguiente segmento se tratarán los indicadores relacionados con los asegurados.

  Utilizamos tres gráficos de barras para visualizar los indicadores de:
  - Insured Sex. En la descripción emergente se podrá ver el porcentaje del número de pólizas.
  - Age. En la descripción emergente se podrá ver el número de pólizas.
  - Months As Customer. En descripción la emergente se podrá ver el número de pólizas. Este último indicador se ha utilizado para crear un campo calculado “Years as customer” que es el que se utiliza en la gráfica, para que se representen los años que han sido clientes.

  Para representar estos indicadores se han utilizado gráficos de columnas que permiten mostrar comparaciones numéricas entre categorías. Un eje del gráfico muestra las categorías que se comparan, en este caso el eje X (Insured Sex, Age y Years as customer), y el otro eje representa la escala de valores en este caso el eje Y (REC(insurance_claim)).

- Finalmente, en el último segmento se tratarán los indicadores relacionados con las reclamaciones registradas.
  - En el primer gráfico se representa el capital ganado (Capital.Gains) y perdido (Capital.Loss). En la descripción emergente se podrá ver el importe en millones.
  - El segundo gráfico se representa las horas en las que ocurrieron los siniestros (Incident Hour Of The Day). En la descripción emergente se podrá ver el número de pólizas en cada hora. Tal y como se ha comentado anteriormente se ha utilizado una paleta personalizada divergente con dos rangos. Donde los que tiene menor valor se identifican con el color naranja y los que tienen mayor valor se identifican con el color café.
  - El tercer gráfico se muestra las ocurrencias de los siniestros según los indicadores de las autoridades contactadas (Authorities Contacted) y si existe reporte policial (Police Report Available). En la descripción emergente se podrá ver el número de pólizas para cada caso.
  - El cuarto gráfico se muestra la localización de los siniestros. En la descripción emergente se puede ver el estado (Incident State) y el importe de la suma de importes total de los siniestros (SumaTotalClaim). Tal y como se ha comentado anteriormente se utilizado una paleta de colores para identificar los estados con tiene tres rangos “Menos de 5000000”, donde se utiliza el color naranja, para el rango “Menos de 10000000” se utiliza el color café y para el rango “Mas de 10000000” se utiliza el color verde amarillo. En el gráfico del mapa el usuario podrá realizar acciones típicas como: zoom (aumento y disminución), mover el mapa, selección de un país, selección de áreas y búsquedas por nombre de país.

Para la representación de los capitales se ha utilizado un gráfico divergente o también llamado gráfico espejo. Colocando en una barra izquierda los valores negativos y en una derecha los positivos y marcando la divergencia mediante una línea vertical en el medio.

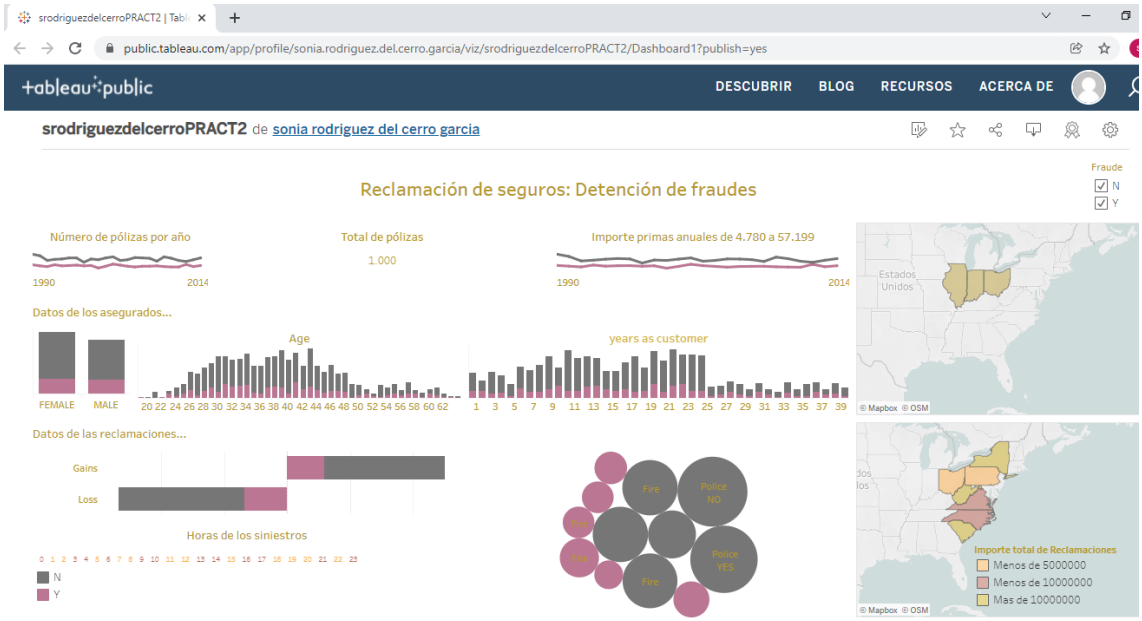
Para la representación de las horas de los siniestros se ha utilizado una visualización de Gantt. Este tipo de visualización plasma en una línea horizontal, los diferentes elementos a analizar, en nuestro caso se muestra cada hora del día y según el color de cada día se muestran los dos rangos de valores, de 29 a 41 el color naranja y de 41 a 54 el color café.

Para la representación de los indicadores de las autoridades contactadas durante el siniestro y si existe reporte policial, se ha utilizado un gráfico de barras apiladas, representadas por círculos, donde cada círculo tiene un diámetro mayor según el número de pólizas.

Para la representación de la localización de los siniestros se ha utilizado un mapa coropléticos que son un tipo de mapa temático de carácter cuantitativo que tiene las áreas coloreadas donde cada área tiene diferentes colores para representar diferentes escalas de valores de la variable suma de importes total de los siniestros.

Respecto a la interactividad el mapa juega un papel importante en la visualización. En los mapas podremos realizar operaciones como: zoom (aumento y disminución), mover el mapa, selección de un país, selección de áreas y búsquedas por nombre de país. Además, como hemos ido indicando, en cada gráfico cuando el usuario para el ratón por encima se mostrará la descripción emergente de cada uno de los gráficos. Si selecciona alguno de los datos el resto de los datos de la gráfica se quedará opaco. Por último, la visualización tiene un filtro que permite seleccionar las reclamaciones fraudulentas o no en todos los gráficos.

La composición final de este proyecto se puede ver en la figura.



Conclusiones

A lo largo de desarrollo del proyecto he aprendido que la visualización de los datos es una herramienta de comunican de información a las personas. Dado que las visualizaciones pueden llevar a las tomas de decisiones importantes estas deben entre otras cosas responder a las preguntas para las que fueron desarrolladas, ser auto explicativos, ser funcionales, en muchos casos debe permitir interactividad con los usuarios.

El fraude en los seguros es un gran problema en la industria. En este ejemplo, trabajaremos con algunos datos de seguros de automóviles para demostrar cómo podemos saber si una reclamación de seguro es fraudulenta o no.

Una de las opciones es cruzar constantemente bases de datos de reclamaciones actuales con bases de datos de reclamaciones históricas fraudulentas, datos de los clientes, y estudiar los casos en los que existan coincidencias. Cuantas más variables tengamos y podamos cruzar de manera eficiente, más opciones de éxito tiene este sistema. Las variables en estas bases de datos no provienen únicamente de las pólizas y las reclamaciones, sino que también de los datos de los clientes, su comportamiento, su entorno… Este proceso es tremendamente complicado para ser realizado por una persona y cada caso consumiría muchos recursos, en cambio utilizando análisis de Big Data es algo mucho más factible.

En el desarrollo del trabajo encontramos evidencias que, en un total de 1000 pólizas tratadas, a lo largo de los años entre 1990 y 2014 siempre el número de pólizas no fraudulentas ha sido superior al de fraudulentas. Aunque como podemos ver en la gráfica hay dos años donde el número se ha acercado son los años 1998 y 2012. Igual ocurre con el importe anual de las primas siempre ha sido mayor el importe de las pólizas no fraudulentas.

Si analizamos las características de los asegurados podemos concluir, que, aunque hay más pólizas contratadas por mujeres, el porcentaje de fraudes es el mismo tanto en mujeres como en hombres. Si podemos identificar que hay mas fraudes en edades de 31 y 41 años y en asegurados con una antigüedad menor de 25 años.

En cuanto a las reclamaciones de los siniestros, el capital de perdido en los siniestros (6,8M) es superior al ganado (5,98M), las horas donde más se producen siniestros fraudulentos son de madrugada (a las 21, 23, 1, 3 y 6 horas), durante la mañana (a las 10 y 11 horas), al medio día (a las 14, 15 y 16 horas) y por la tarde (a las 18 horas).
Adicionalmente se encontró evidencias de que los fraudes se han producido cuando en los incidentes se han llamado a los bomberos y ambulancias sin reporte. Finalmente, la localización de cada siniestro y el importe de la suma de importes total de los siniestros, nos hacen ver que en todos los estados el importe ha sido menor de 5M, aunque el total del importe de siniestros en estados como new york, virginia occidental o carolina del sur superen los 10M.