Tecnológico de Costa Rica

Visualización de información

Proyecto 1

Luis Felipe Calderon Perez

Yeri Estiven Porras Viquez

Introducción

El presente proyecto tiene como finalidad presentar diferentes análisis mediante gráficas realizadas en el lenguaje R, para ello se seleccionaron los accidentes de tránsito en Costa Rica ocurridos entre el año 2016 y el año 2022.

Se utilizaron diversas gráficas como las caras de chernoff, barras apiladas, barras, de pastel, facetas, matriz de dispersión, así como una gráfica compuesta utilizando barras y puntos, en estas se utilizaron diversas variables de diferentes dimensiones, univariable, bivariable y multivariable, principalmente categóricas.

Junto con estas gráficas se da un análisis exploratorio (EDA), en donde se pueden observar, como valores mínimos, valores máximos, densidad de los valores y proporciones dentro de los datos

Descripción del problema

Proyecto 1

El propósito de este proyecto es realizar un análisis exploratorio (EDA) de un conjunto de datos de interés. Para ello se deberá publicar una página Web en donde se describa el EDA realizado, incorporando una serie de visualizaciones que describen el comportamiento de los datos y los patrones identificados en los mismos.

Análisis exploratorio de datos

El análisis exploratorio de datos (EDA), que a menudo hace uso de técnicas de visualización de datos, es una herramienta utilizada por los científicos de datos para examinar, evaluar y resumir grandes conjuntos de datos. Facilita a los científicos de datos la búsqueda de patrones, la identificación de anomalías, la comprobación de hipótesis o la validación de suposiciones, ayudándoles a gestionar de forma óptima las fuentes de datos para obtener las respuestas necesarias.

EDA ofrece un mejor conocimiento de las variables del conjunto de datos y de las interacciones entre ellas. Se utiliza sobre todo para ver qué pueden revelar los datos más allá del trabajo formal de modelización o comprobación de hipótesis. También puede ayudar a determinar la idoneidad de los métodos estadísticos que se piensan utilizar para el análisis de datos. Los enfoques EDA fueron creados por primera vez en la década de 1970 por el matemático estadounidense John Tukey y siguen siendo un enfoque popular en el proceso de descubrimiento de datos.

Tipos de gráficas

- Gráfica unidimensionales (una sola variable): se deben crear tres o más gráficas en las que se muestre la distribución de los datos de una sola variable. Se debe describir en el documento dicha gráfica, así como cualquier tendencia detectada en la misma. Nótese que la idea no es mostrar la distribución de cualquier variable, sino escoger aquellas que presentan un patrón interesante.
- Gráfica bidimensional (dos variables): se deben crear dos o más gráficas en las que se muestre la combinación de dos variables. Se deben escoger aquellas combinaciones de variables en las que se detecte alguna interrelación entre las variables. Igualmente se debe describir en el documento dicha gráfica, así como cualquier tendencia detectada en la misma.
- Gráfica multidimensional: Se deberá elaborar al menos un tipo de gráfica multidimensional en donde se incorporen los valores de al menos 5 variables al mismo tiempo. Dichas variables deben ser seleccionadas cuidadosamente de forma que sean fáciles de observar las tendencias o casos particulares de los datos.
- Facetas: Se debe incorporar al menos una gráfica de facetas en el análisis. Dicha gráfica podrá involucrar tres o más variables que presenten algún comportamiento interesante.
- Imagen compuesta: Adicionalmente se debe incluir en el informe al menos una gráfica compuesta de otras gráficas.

- Interacción: Todas las gráficas que se presenten en la página Web deberán contar con capacidades de interacción. Para ello se pueden utilizar las librerías: Plotly, Giraffe, Bokeh, etc.

Generación de página Web

Para generar la página Web se utilizará un notebook de R (rnotebook). Note que se deben ocultar los segmentos de código en la página generada. La página generada debe ser subida al TecDigital y debe quedar publicada en algún sitio público (github, netlify, etc.)

Consideraciones generales

- Todo el desarrollo del proyecto debe realizarse en lenguaje R.
- Se deberá generar una documentación formal, en formato pdf, en donde se describan las diferentes etapas del desarrollo del proyecto, las decisiones de diseño que se tomaron, los mecanismos de programación (en R) utilizados, y los resultados de las diferentes pruebas al programa. Dicha documentación deberá incluir al menos las siguientes secciones:
    - Introducción
    - Descripción del problema (este enunciado)
    - Definición de fuentes de datos
    - Descripción detallada y explicación de las secciones principales del documento.
    - Conclusiones
- El proyecto puede realizarse en grupos de a lo más dos estudiantes. No se permite la copia entre grupos de estudiantes.

Temas a asignar

- Accidentes de tránsito en Costa Rica
- Casos de Covid en Costa Rica
- Temblores en Costa Rica
- UGGS Earthquake Catalog
- Incidencia tumores malignos en Costa Rica
- Estadísticas policiales en Costa Rica
- Uso de transporte - Asamblea Legislativa

Definición de fuentes de datos

La fuente de datos fue dada por el profesor, sin embargo escoger los accidentes de tránsito en Costa Rica fue una elección nuestra. Escogimos este tema, debido al aumento de accidentes de tránsito que son reportados diariamente en las noticias y a la necesidad de movilizarnos de un lugar a otro. Por lo que es importante saber y concientizar acerca de los problemas y accidentes que se pueden presentar en carretera.

En la fuente de datos proporcionada se nos presentan diversas variables como lo son:

- Clase de accidente: muestra los tipos y gravedad de los heridos
- Tipo de accidente: muestra que vehículos colisionaron o si involucran a terceros
- Año: muestra el año en el que ocurrió el accidente
- Hora: muestra la hora en la que se presentó el accidente
- Hora recodificada: muestra en qué intervalo de 6 horas se dio el accidente
- Provincia: muestra en qué provincia sucedió el accidente
- Cantón: muestra en qué cantón de la provincia sucedió el accidente
- Distrito: muestras en que distrito del cantón sucedió el accidente
- Ruta: muestra el número de ruta o si es cantonal
- Kilómetro: muestra el número del kilómetro de la ruta o si es cantonal
- Rural o urbana: muestra si es una ruta de una comunidad rural o urbana
- Calzada vertical: muestra si existe una pendiente
- Calzada horizontal: muestra si es un cruce, recta o curva
- Tipo de calzada: muestra si es en asfalto o concreto
- Tipo de circulación: muestra el sentido de circulación de los vehículos
- Estado del tiempo: muestra el clima del momento del incidente
- Estado de la calzada: muestra las condiciones en las que se encuentra la calzada
- Región Mideplan: muestra la región en la que sucedió según el MIDEPLAN
- Tipo de ruta: muestra si es una ruta nacional o cantonal
- Dia: indica el día del incidente
- Mes: indica el mes del incidente

Descripción

Dentro de los datos pudimos encontrar patrones y detalles interesantes entre las variables que se definirán con las siguientes gráficas.

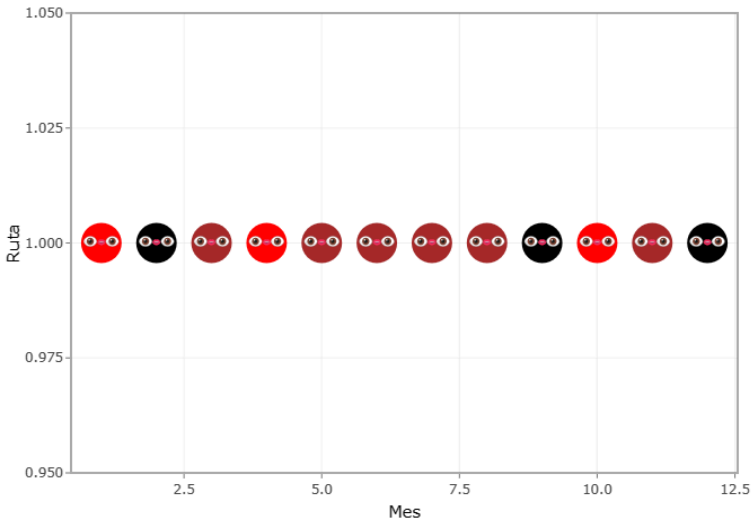

fig. 1

En esta gráfica de accidentes entre el 2016-2022 de caras de Chernoff podemos observar en un inicio que hay 12 caras, una por cada mes del año. Segundo hay variedad de colores, los cuales representan el día con más accidentes de ese mes. Y por último todas las caras tienen el mismo tamaño, ya que el tamaño determina la ruta con más accidentes. Teniendo esto en cuenta, ¿por qué tienen el mismo tamaño todas las caras? Pues los datos de análisis indican que a través del intervalo de años entre el 2016-2022 las rutas con más accidentes han sido las cantonales. Es muy interesante que en ese intervalo de años el día que suele tener más accidentes es el viernes, además el Lunes y Sábado suelen tener la misma cantidad de accidentes.

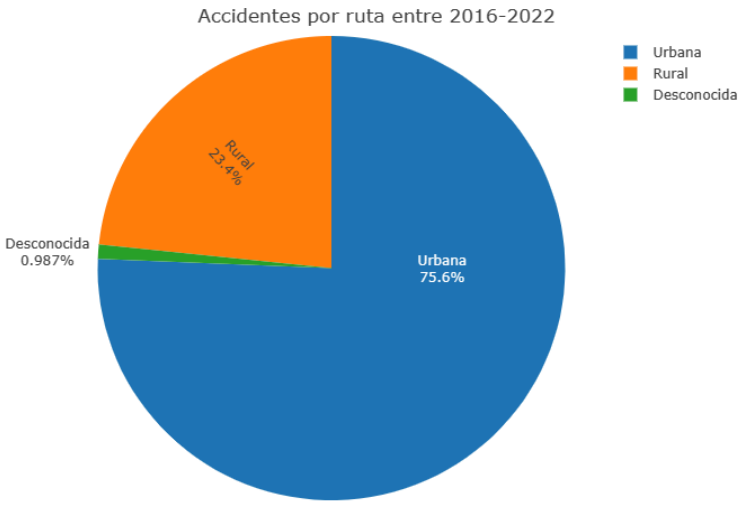

fig.2

En este gráfico de pastel se muestra el porcentaje de los accidentes entre el 2016-2022 por tipo de ruta. Se puede apreciar que el mayor porcentaje de accidentes en el intervalo de años se ha dado en rutas urbanas, eso quizá tenga que ver en parte porque hay menos tránsito en rutas rurales. Con lo anterior en cuenta, también puede ser que porque hay menos cantidad de tránsito haya menos factores psicoemocionales que afecten al momento de conducir.

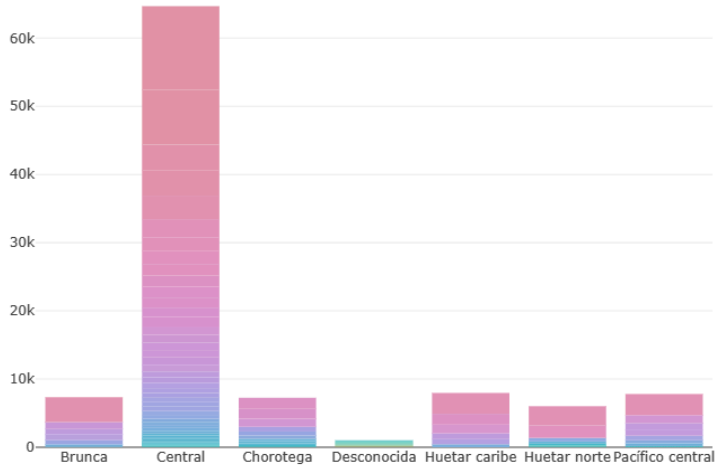

fig. 3

Entre 2016-2022, la gráfica apilada de barras muestra que las barras están agrupadas por región y apiladas por cantones con mayor número de accidentes. El color indica cuántos accidentes hubo en cada cantón. En la zona de Costa Rica, donde más accidentes ocurrieron, la región central y el cantón de San José tienen el mayor número de accidentes. Además, los cantones que comparten más accidentes son los más desarrollados urbanamente, mientras que los cantones con menos accidentes son los más pobres o menos desarrollados. Esto se podría deber a que hay menos vehículos en carretera, por la falta de poder adquisitivo.

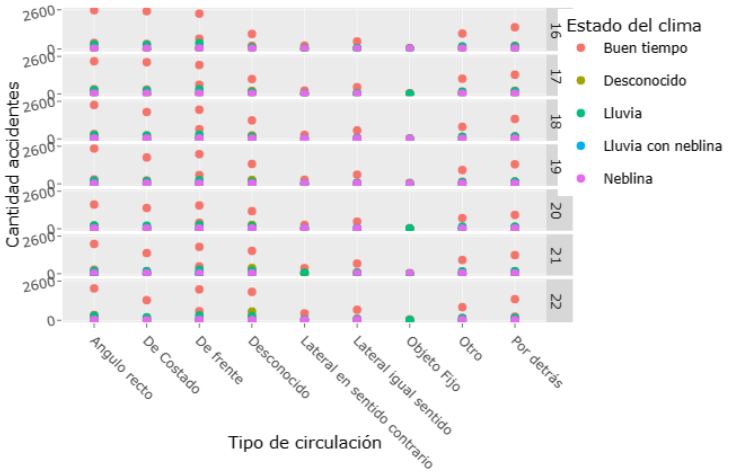

fig. 4

Esta gráfica muestra el mínimo de accidentes por año, tipo de circulación, estado del clima y son agrupados por la gravedad del accidente. La mayoría de los accidentes curiosamente se dan con buen tiempo, contrariamente a lo que se esperaría que sea más bien en un mal clima como la lluvia. Esto quiere decir que si hay un buen clima los conductores se relajan y no andan lo suficientemente atentos al ambiente. Esto tiene un efecto inmediato con el tipo de circulación, ya que al no tomar la suficiente precaución se terminan accidentado, y estos en su mayoría son de extrema gravedad.

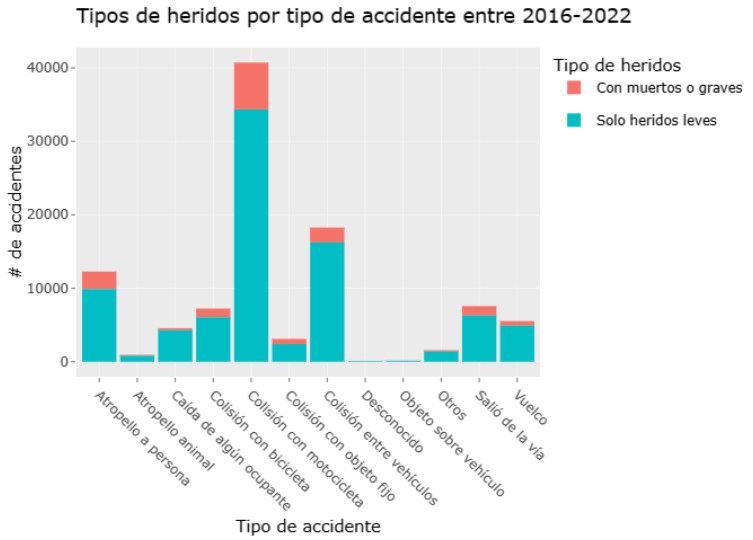

fig. 5

En esta gráfica de barras apiladas podemos ver los heridos en accidentes de tránsito que se presentaron entre 2016-2022. gracias a esta gráfica podemos conceptualizar de una mejor manera la razón por la cual las autoridades son insistentes con respecto al cuidado que se tiene que tener al conducir una motocicleta ya que como se puede observar son los accidentes con mayor cantidad de heridos leves y con la mayor cantidad de heridos graves o muertos, que si bien es cierto no todos los heridos son motociclistas pueden ser una gran mayoría de los heridos graves, además, de que hay incluso más muertes que los atropellos directos hacia personas. Asimismo, podemos ver que son pocos los accidentes de los cuales no se sabe que los causó.

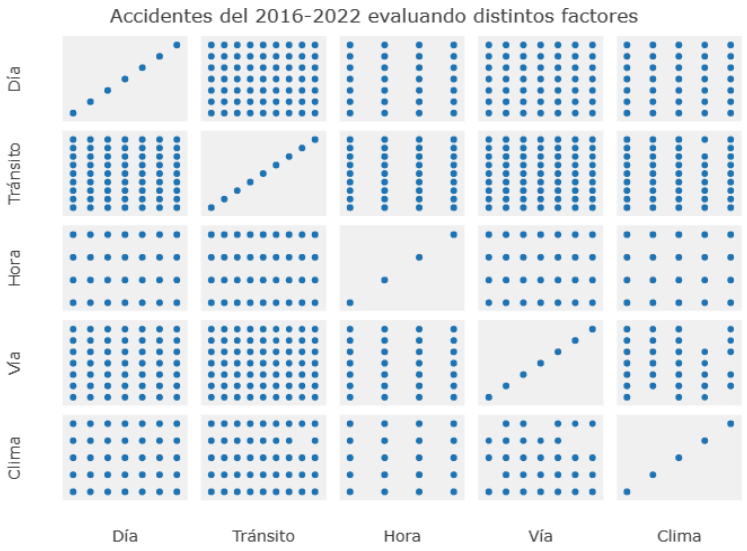

fig. 6

Esta matriz de dispersión multivariada evalúa las condiciones climáticas, el tipo de vía, el tiempo registrado, el tipo de tránsito y el día con más accidentes en Costa Rica del 2016 al 2022. Como muestra la matriz, el número de accidentes no solo en un año, sino en varios años. Depende del lugar por donde conduzcas, si hay prohibiciones de circulación, del tipo de tráfico o de la naturaleza de la carretera. Esto significa que estos tres factores, que varían de un año a otro, no cambian. Además, el día, las condiciones de la carretera y el clima muestran un patrón de accidentes debido a los días ocupados, lo que puede hacer que la conducción en ciertos tipos de vías sea más difícil y peor debido a las condiciones climáticas. Por último, el patrón entre día, carretera y hora puede ser muy similar al anterior, pero puede variar en función de si es hora punta o un indicio de que el conductor no se encuentra en óptimas condiciones de conducción. Curiosamente, el tráfico, las condiciones meteorológicas y las carreteras son claramente inestables, como se esperaba.

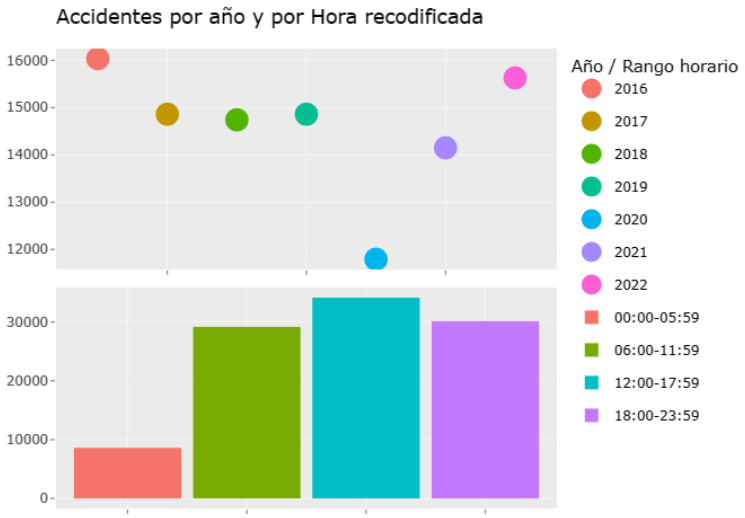

fig. 7

Esta gráfica compuesta nos muestra, en los puntos, los datos segmentados por años y, en la de barras, podemos ver los rangos horarios entre 2016-2022 que presentaron más accidentes. En la de puntos podemos ver como el año 2016 fue el año que tuvo mayor incidencia, además, un detalle que se puede rescatar es que en el año 2020 los accidentes tuvieron una menor presencia en nuestro país, que pudo haber sido debido a las restricciones vehiculares que las autoridades pusieron durante el periodo de pandemia y al haber menos vehículos ocurrieron menos accidentes. En la de barras se puede apreciar como los horarios entre las 12:00 y las 17:59 horas son los que mayor incidencia presentan posiblemente debido a que es el horario en el que la mayoría de las personas salen de trabajar y esto incrementa el tráfico. Asimismo, podemos ver como la madrugada presenta una cantidad mucho menor de accidentes posiblemente debido a que existe un tráfico menor durante esas horas.

Conclusión

Con el Análisis Exploratorio de los Datos se evidencian diversos factores que influyen a través de los años, en patrones de accidentes que podrían suponerse como influencia en el aumento de accidentes de tránsito, pero en realidad no son así. Un ejemplo claro es el análisis de la gráfica de facetas, en donde se observa que el buen estado del clima se relaciona con la mayor cantidad de accidentes.

Por otro lado, se notaron hallazgos que curiosamente desafían la intuición. Por ejemplo, se descubrió que el tipo de ruta más peligroso en Costa Rica es la cantonal, y que los cantones con mayor desarrollo presentan la incidencia más alta de accidentes. Además, se identificó un patrón inesperado en cuanto al horario recodificado de accidentes, en donde el rango de horas con menor cantidad de incidentes es entre la medianoche y las 6 de la mañana, posiblemente debido al menor tráfico en las carreteras durante ese periodo. Este fenómeno también se refleja en la disminución significativa de accidentes en el año 2020, atribuible al confinamiento durante la pandemia, lo cual también se muestra con la reducida cantidad de accidentes en zonas rurales.

Página web

En el siguiente link podrá observar el sitio web generado

Accidentes de transito en CR entre 2016-2022 (niceri88.github.io)