Instituto Tecnológico de Costa Rica

Escuela de Ingeniería en Computación

Visualización de Información

Casos de COVID-19 en Costa Rica

Realizado por:

Jennifer Alvarado Brenes c. 2020124171

Daniel Cornejo Granados c. 2019018538

Profesor:

Ing. Armando Arce Orozco

I Semestre 2024

Contenido

1. Introducción

En la era de la información, la capacidad de comprender y comunicar datos de manera efectiva se ha vuelto fundamental en diversos ámbitos, desde la ciencia hasta la toma de decisiones empresariales. En este contexto, la visualización de datos emerge como una poderosa herramienta que permite revelar patrones, identificar tendencias y comunicar perspectivas de manera clara y concisa. Es en este escenario que el lenguaje de programación R se posiciona como una de las herramientas más relevantes para el análisis y visualización de datos, gracias a su amplia gama de paquetes y su flexibilidad para manejar conjuntos de datos complejos.

En el contexto de la pandemia de COVID-19, la visualización de datos ha adquirido una importancia sin precedentes. En particular, en el caso de Costa Rica, la capacidad de analizar y comunicar eficazmente la evolución de la enfermedad ha sido crucial para informar a la población, orientar la toma de decisiones políticas y guiar las estrategias de salud pública. En este sentido, el uso de R para la visualización de datos ha jugado un papel fundamental, permitiendo a los analistas y científicos de datos generar gráficos informativos y dinámicos que ayudan a comprender la propagación del virus, identificar áreas de riesgo y evaluar el impacto de las medidas implementadas.

2. Descripción del problema

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 distribucción 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 bidimensionales (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 algún 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 seleccionados 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 compuestada: 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.)

3. Definición de fuentes de datos

Los datos utilizados para formas las respectivas gráficas en R fueron tomados de un archivo llamado “CasosCOVID.xlsx”.

Este archivo cuenta con varias hojas de datos, y estas hojas con varias columnas. En este proyecto se utilizó la hoja de datos generales.

Esta hoja específicamente cuenta con datos numéricos en muchas categorías, como casos positivos, por género, edad y nacionalidad. También se conocen los datos numéricos de los casos de covid hospitalizados, en UCI (Unidad de Cuidados Intensivos), los casos de fallecidos, recuperados, activos, entre otros.

Todos estos datos se proporcionan desde el 7 de marzo del 2020 hasta el 30 de mayo de 2022.

4. Descripción detallada

Se realizó un análisis exhaustivo con varios tipos de gráficas para muchos de los datos obtenidos. Esto se detalla a continuación:

4.1 Casos positivos

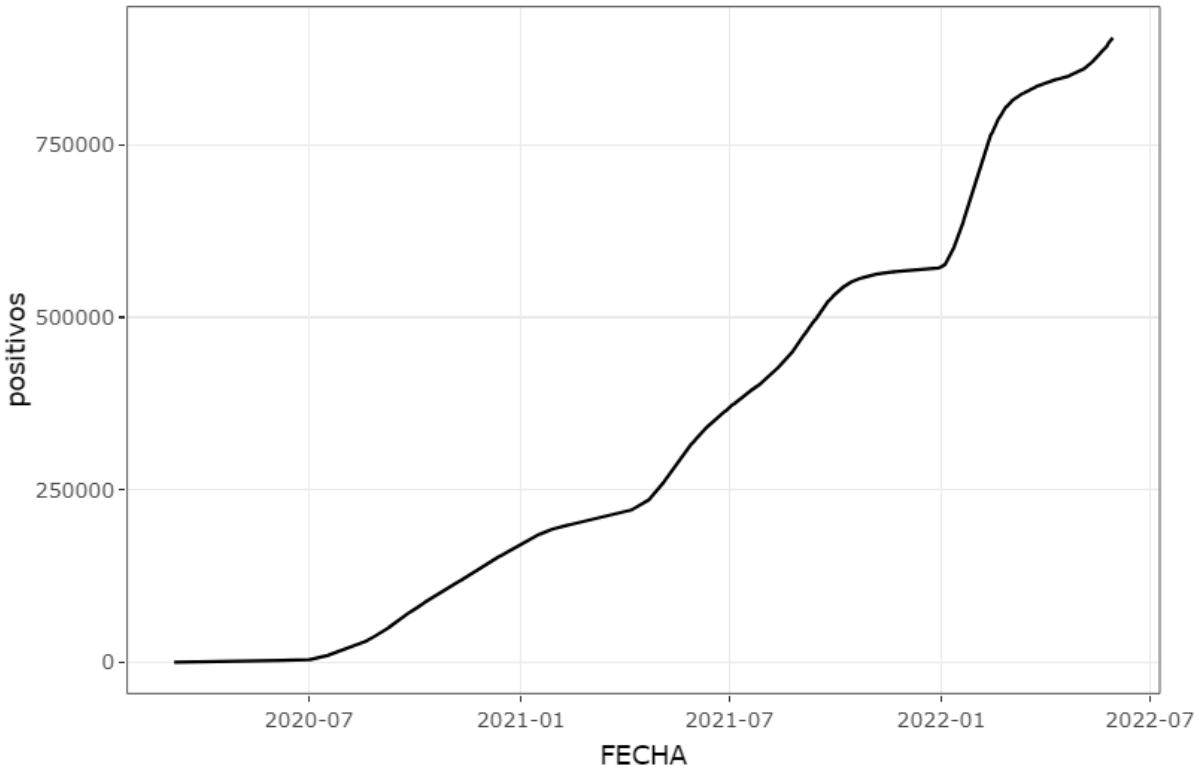
Se presenta la cantidad de casos positivos de COVID-19 en Costa Rica desde abril 2020 hasta mayo 2022.

En este caso es posible observar un comportamiento con tendencia lineal en la cantidad de casos positivos de COVID-19 en Costa Rica, aunque durante los años en los que se recopilaron los datos podemos notar algunas gradas es posible afirmar que el comportamiento fue mayormente lineal.

Código realizado para su producción:

```
casos_positivos <- ggplot(data, aes(x=FECHA, y=positivos)) +
geom_line() + theme_minimal() + theme_bw()
ggplotly(casos_positivos)
```

Gráfico:



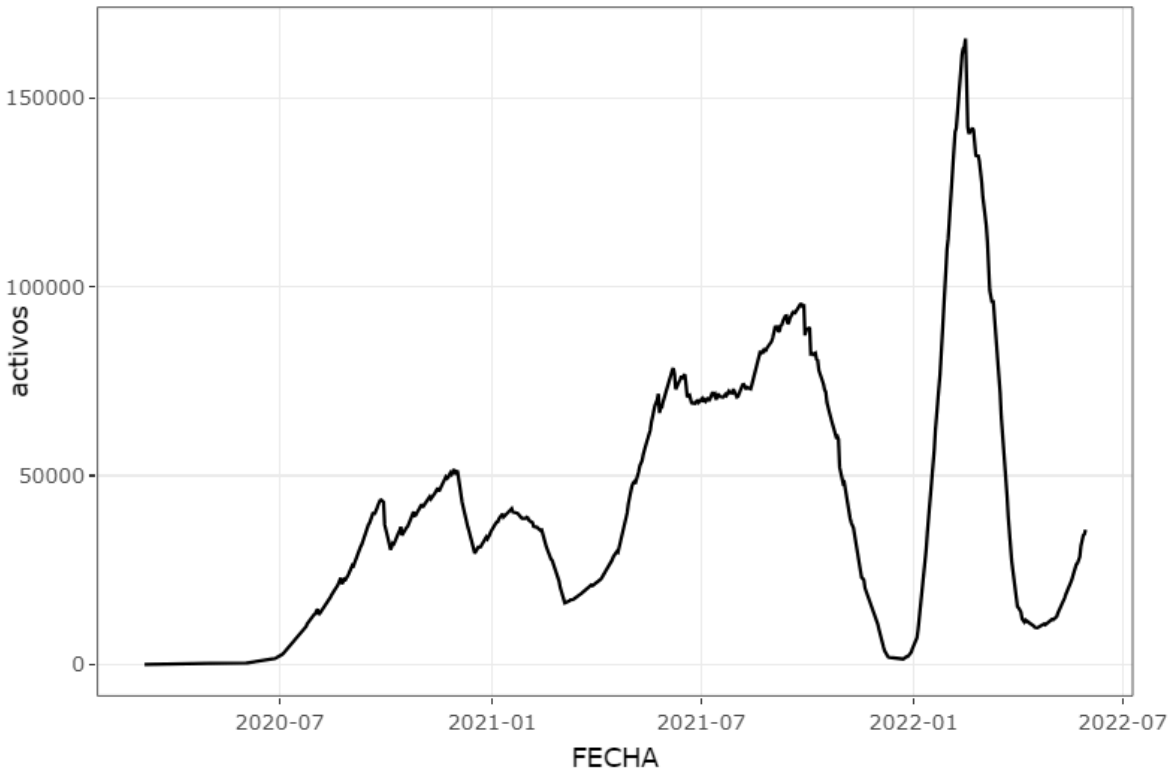
4.2 Casos activos

Se presenta la cantidad de casos activos de COVID-19 en Costa Rica desde abril 2020 hasta mayo 2022. Podemos notar un comportamiento común de jorobas en la cantidad de casos activos, este comportamiento se podría decir que fue por las medidas de restricción sanitaria que se tomaron en el país.

Código realizado para su producción:

```
casos_activos <- ggplot(data, aes(x=FECHA, y=activos)) + geom_line() +
theme_minimal() + theme_bw()
ggplotly(casos_activos)
```

Gráfico:



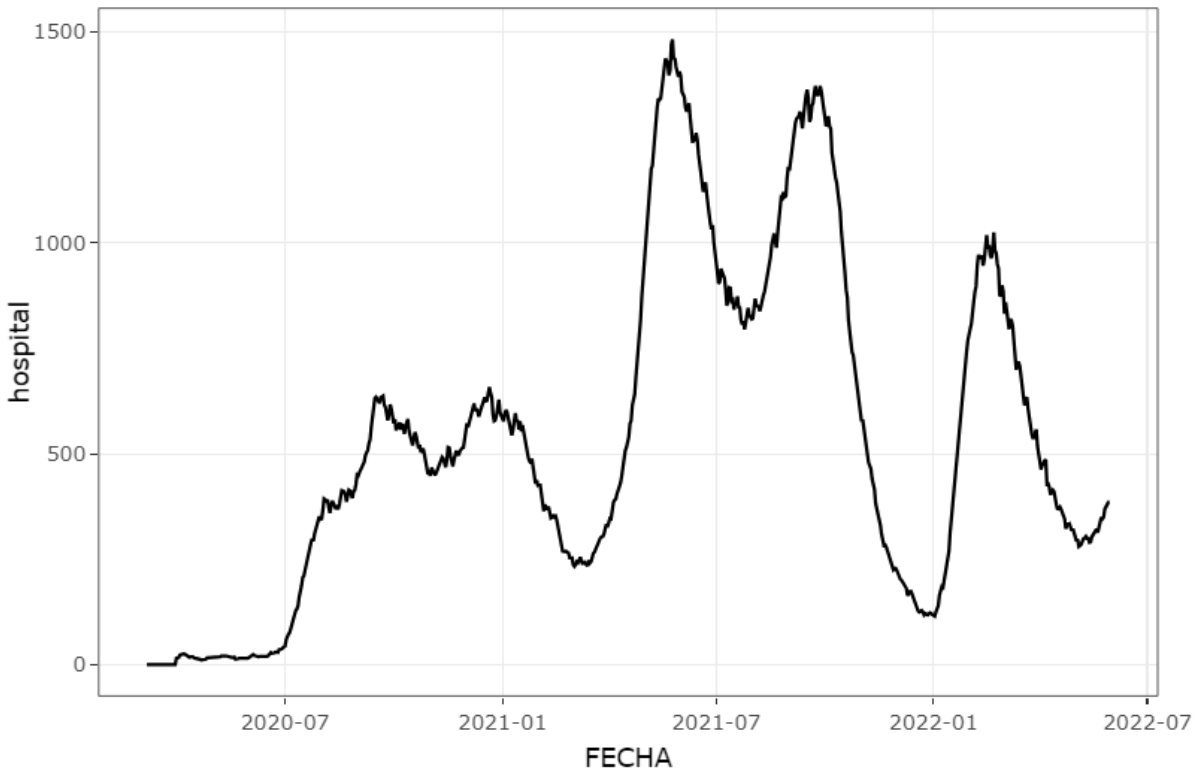
4.3 Casos hospitalizados

La cantidad de casos hospitalizados no parecen demasiados a simple vista de los ejes, sin embargo es necesario tener en cuenta la cantidad de espacios con los que cuenta cada hospital para todos sus pacientes, por lo que la cantidad de casos hospitalizados no es un dato menor. Durante la extensión de los datos podemos notar el parecido con los casos activos, en especial podemos notar la acentuación de las jorobas de julio 2021, donde se presentaron los picos más altos de casos hospitalizados.

Código realizado para su producción:

```
casos_hospitalizados <- ggplot(data, aes(x=FECHA, y=hospital)) +
geom_line() + theme_minimal() + theme_bw()
ggplotly(casos_hospitalizados)
```

Gráfico:



4.4 Casos de muertes por género

A continuación, podemos ver el comportamiento de las muertes por COVID-19 en Costa Rica. En este caso, se presentan los datos por género, donde podemos notar que la cantidad de muertes en hombres es mayor que en mujeres. Sin embargo, la diferencia no es tan grande como se podría esperar. En general, podemos observar un comportamiento lineal en la cantidad de muertes por COVID-19 en Costa Rica, y cómo a lo largo del tiempo la diferencia de muertes entre hombres y mujeres se incrementa parcialmente.

Código realizado para su producción:

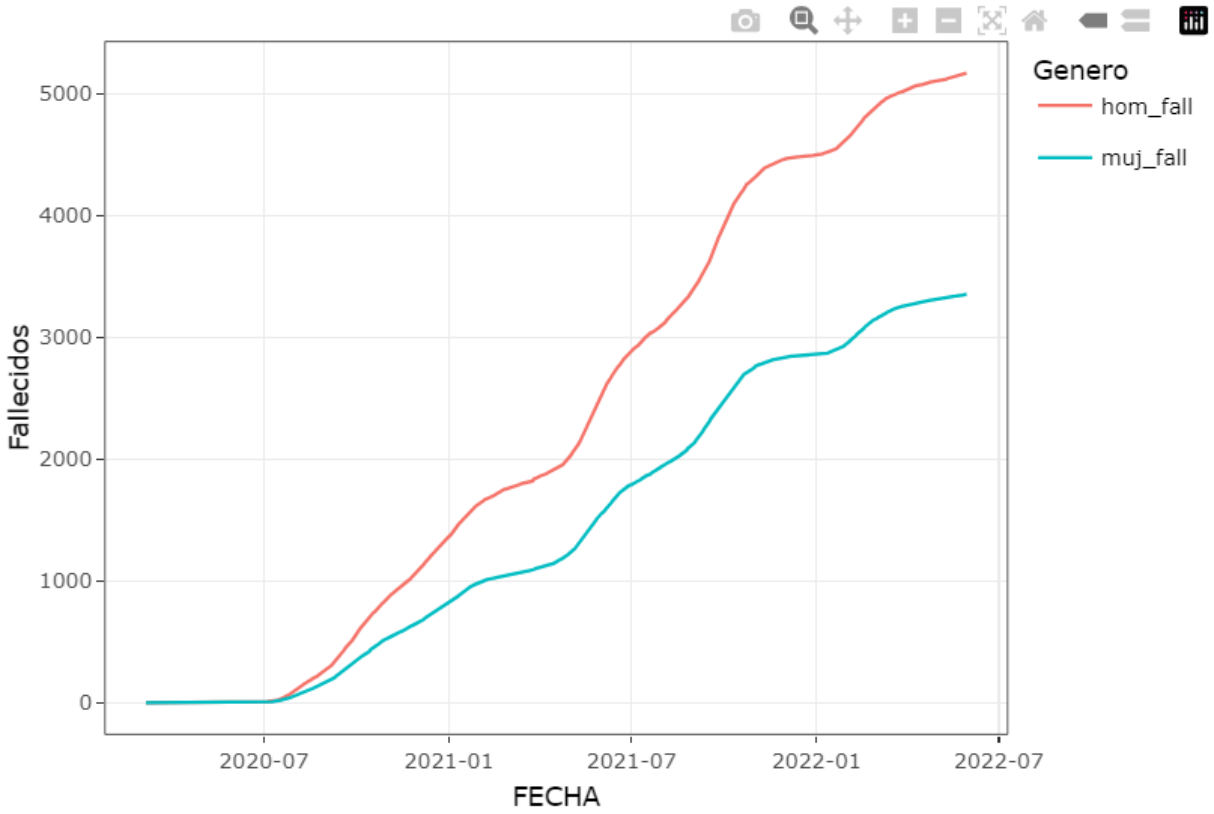
```
names(data)[names(data) == "muj_fall"] <- "Mujeres_fallecidas"
names(data)[names(data) == "hom_fall"] <- "Hombres_fallecidos"

casos_fallecidos <- pivot_longer(data, cols = c(Mujeres_fallecidas,
Hombres_fallecidos), names_to = "Genero", values_to="Fallecidos")

fallecidos_genero <- ggplot(casos_fallecidos, aes(x=FECHA,
y=Fallecidos)) + geom_line(aes(color=Genero))+theme_minimal() +
theme_bw() + labs(title="Fallecidos por COVID-19 en Costa Rica",
x="Fecha", y="Fallecidos") + scale_color_manual(values=c("blue",
"red"))
```

```
ggplotly(fallecidos_genero)
```

Gráfico:



4.5 Casos de recuperaciones por género

En este caso, se presentan los datos de recuperaciones por COVID-19 en Costa Rica. Se muestran los datos por género, donde se puede observar que la cantidad de recuperaciones en hombres es prácticamente la misma que en mujeres, excepto por una anomalía en los datos para la fecha del 9 de junio de 2021. En general, se puede observar un comportamiento lineal en la cantidad de recuperaciones por COVID-19 en Costa Rica, y cómo a lo largo del tiempo la diferencia de recuperaciones entre hombres y mujeres se mantiene similar, aunque al final del diagrama se nota un aumento en la recuperación de las mujeres en comparación con los hombres.

Código realizado para su producción:

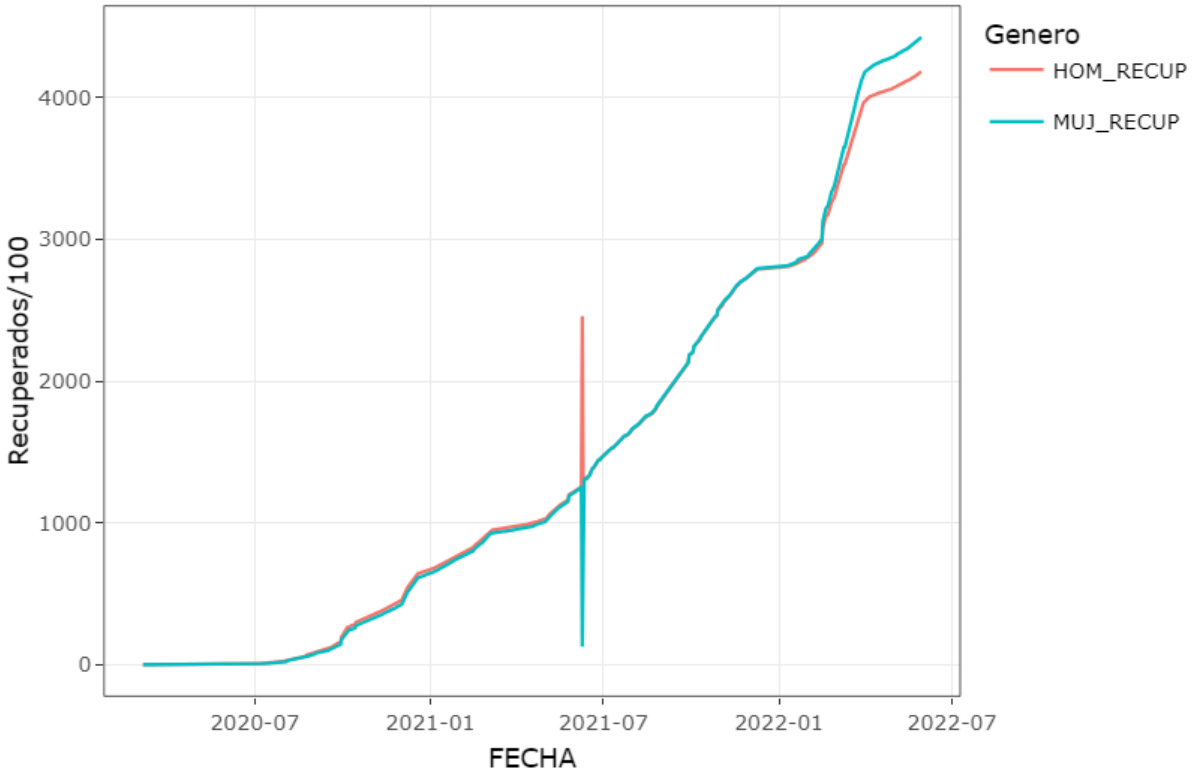
```
names(data)[names(data) == "MUJ_RECUP"] <- "Mujeres_recuperadas"
names(data)[names(data) == "HOM_RECUP"] <- "Hombres_recuperados"

casos_recuperados <- pivot_longer(data,
cols=c(Mujeres_recuperadas,Hombres_recuperados), names_to="Genero",
values_to="Recuperados")
```

```
recuperados_genero <- ggplot(casos_recuperados, aes(x=FECHA,
y=Recuperados/100)) +
geom_line(aes(color=Genero)) +
theme_minimal() + theme_bw() +
scale_color_manual(values = c("blue", "red")) +
labs(title="Recuperados por COVID-19 en Costa Rica", x="Fecha",
y="Recuperados")

ggplotly(recuperados_genero)
```

Gráfico:



4.6 Evolución de casos

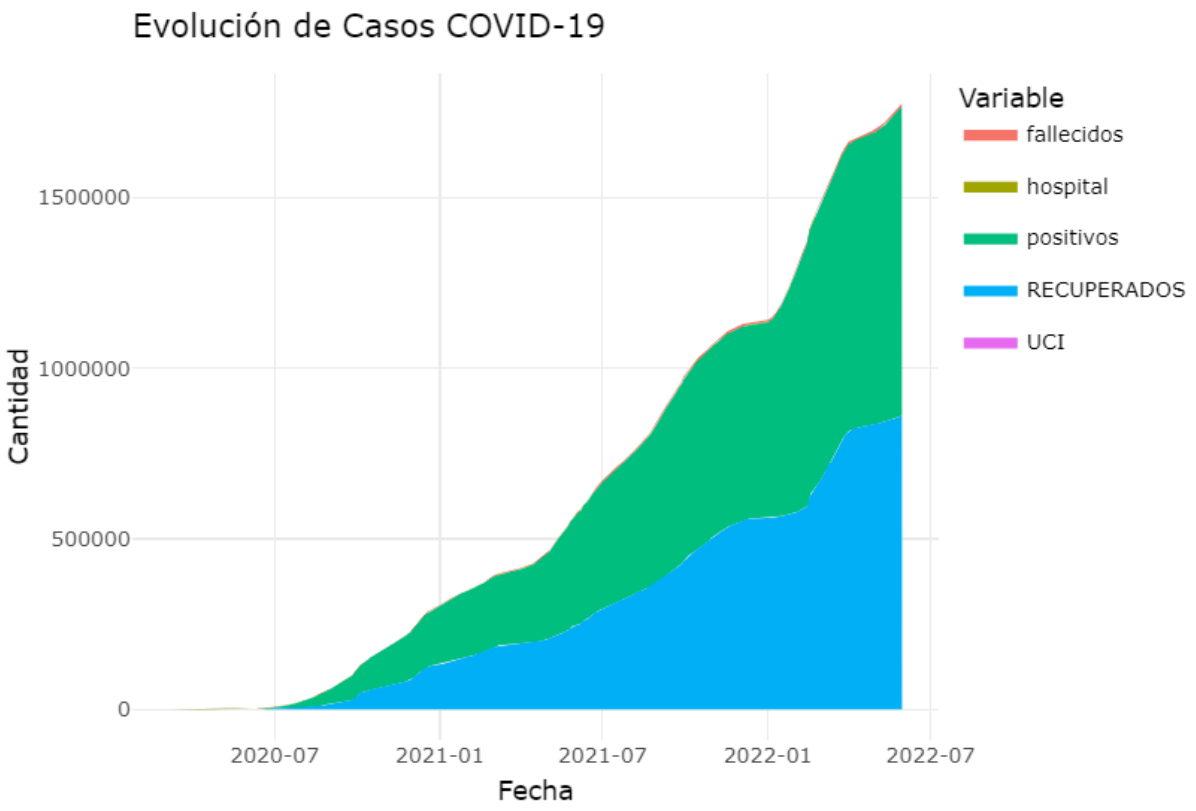
Se visualiza la evolución de los casos de COVID-19 en Costa Rica referentes a las categorías de casos positivos, fallecidos, hospitalizados, hospitalizados en el área de UCI (Unidad de Cuidados Intensivos) y casos recuperados, entre abril 2020 y mayo 2022.

Se puede observar que en mucha mayor medida, hay casos positivos y recuperados, ya que no todas las personas que fueron contagiadas en algún momento murieron, fueron hospitalizadas o llegaron a la UCI. Estas últimas tres categorías representan una minoría en comparación a la gran cantidad de casos positivos y posteriormente, recuperados. Es fácil de visualizar en esta gráfica de áreas apiladas, que entra en la categoría de gráfica multidimensional.

Código realizado para su producción:

```
data_long <- pivot_longer(data, cols = c("positivos", "fallecidos",
"hospital", "UCI", "RECUPERADOS"),names_to = "Variable", values_to =
"Valor")
evolucion_casos <- ggplot(data_long, aes(x = FECHA, y = Valor, fill =
Variable)) + geom_area() + labs(title = "Evolución de Casos COVID-19",
x = "Fecha", y = "Cantidad", fill = "Variable") + theme_minimal()
ggplotly(evolucion_casos)
```

Gráfico:



4.7 Casos de covid en adultos vs menores de edad

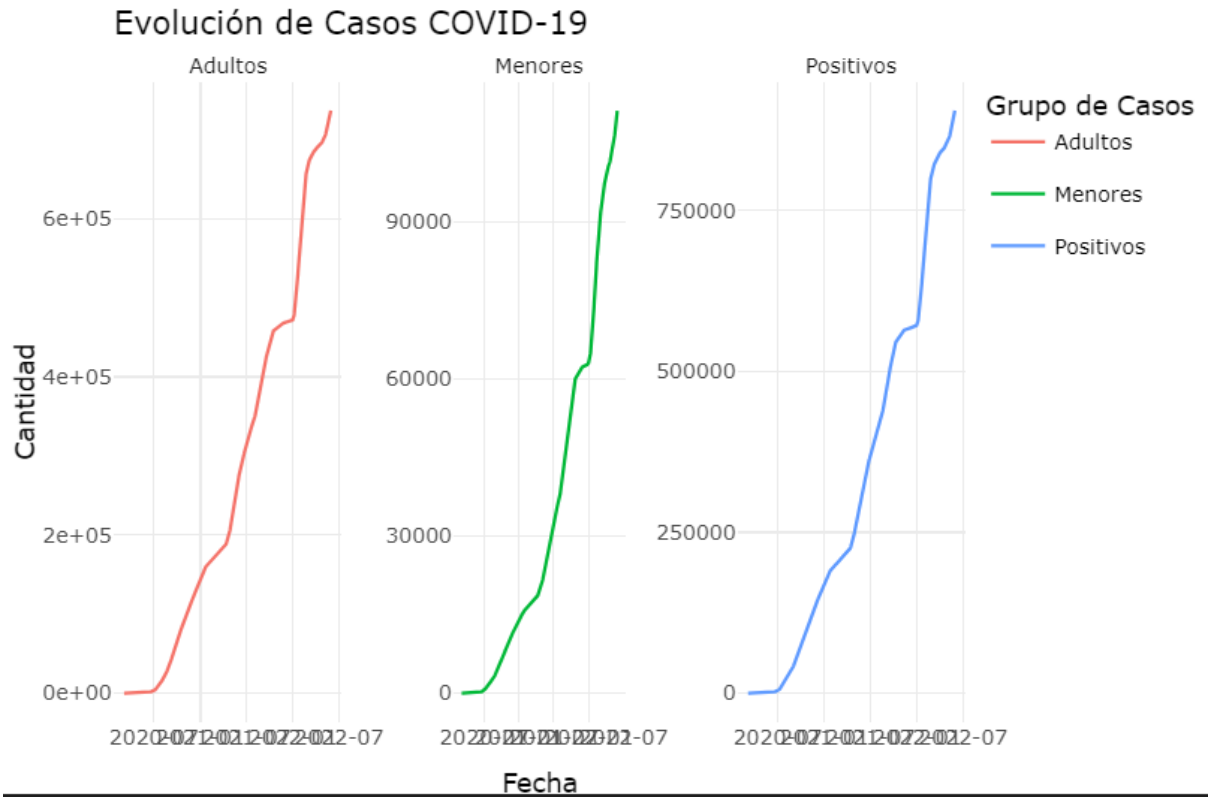
La población de personas menores de edad se vio afectada en mucha menor medida que la población de adultos, dato que es lógico ya que los adultos deben salir más de casa y los menores son una población altamente resguardada y cuidada.

Se puede observar un menor volumen de los datos en los menores de edad en la siguiente gráfica de facetas.

Código realizado para su producción:

```
names(data)[names(data) == "positivos"] <- "Positivos"
names(data)[names(data) == "adul_posi"] <- "Adultos"
names(data)[names(data) == "menor_posi"] <- "Menores"
data_long <- pivot_longer(data, cols = c("Positivos", "Adultos",
"Menores"), names_to = "Grupo de Casos", values_to = "Cantidad")
facetas_casos <- ggplot(data_long, aes(x = FECHA, y = Cantidad, color =
`Grupo de Casos`)) + geom_line() + labs(title = "Evolución de Casos
COVID-19", x = "Fecha", y = "Cantidad", color = "Grupo de Casos") +
facet_wrap(~ `Grupo de Casos`, scales = "free_y") + theme_minimal()
ggplotly(facetas_casos)
```

Gráfico:



4.8 Resumen de casos

Para el análisis de los datos es necesario hacer comparaciones de todas sus categorías. Aquí se muestran algunas, como los datos para los casos positivos, casos activos y casos hospitalizados entre abril de 2020 y mayo de 2022.

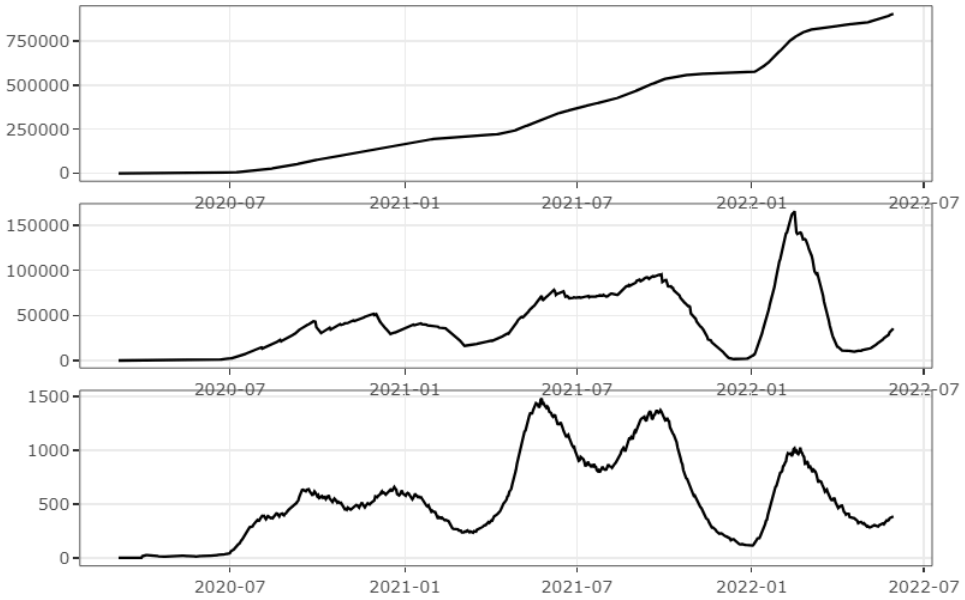
Se puede visualizar una correlación entre los casos activos y los casos hospitalizados, ya que estos calzan con los picos de personas contagiadas durante la pandemia. Por otro lado, los casos positivos siempre son crecientes.

Puede visualizarse de forma completa en la siguiente imagen compuesta, en donde el primero corresponde a casos positivos, el segundo a casos activos y el tercero a casos hospitalizados.

Código realizado para su producción:

```
casos_positivos <- ggplot(data, aes(x=FECHA, y=positivos)) +
geom_line() + theme_minimal() + theme_bw()
casos_activos <- ggplot(data, aes(x=FECHA, y=activos)) + geom_line() +
theme_minimal() + theme_bw()
casos_hospitalizados <- ggplot(data, aes(x=FECHA, y=hospital)) +
geom_line() + theme_minimal() + theme_bw()
grafico_compuesto <- subplot(casos_positivos, casos_activos,
casos_hospitalizados, nrows = 3)
ggplotly(grafico_compuesto)
```

Gráfico:



4.7 Librerías utilizadas

- **readxl**: Esta librería se utiliza para leer datos de archivos Excel. En este caso, se usa para leer el archivo "CasosCOVID.xlsx" y cargar los datos en el entorno de R.
- **ggplot2**: Es una librería muy popular en R para la creación de gráficos estadísticos. Se utiliza en varias partes del script para crear gráficos de líneas que muestran la evolución de diferentes variables relacionadas con los casos de COVID-19 en Costa Rica.
- **plotly**: Esta librería se utiliza para hacer que los gráficos de ggplot2 sean interactivos. Con la función ggplotly() se convierten los gráficos estáticos de ggplot2 en gráficos interactivos, lo que permite explorar los datos con más detalle.
- **tidyr**: Esta librería se utiliza para realizar operaciones de limpieza y transformación de datos. En el script, se utiliza principalmente para transformar el formato de los datos de ancho a largo, lo que facilita su manipulación y visualización.
- **patchwork**: Esta librería se utiliza para crear gráficos compuestos, es decir, combinar varios gráficos en una sola imagen. En el script, se utiliza para crear un gráfico compuesto que muestra la evolución de casos positivos, activos y hospitalizados en Costa Rica.

5. Conclusiones

Se exploró la importancia de la visualización de datos como una herramienta fundamental para comprender y comunicar la evolución de la pandemia de COVID-19 en Costa Rica. A través del análisis exploratorio de datos (EDA) y la creación de diversas visualizaciones utilizando el lenguaje de programación R, se obtuvieron perspectivas valiosas sobre la propagación del virus, la respuesta de las autoridades y el impacto en la población.

La utilización de librerías como ggplot2, plotly y tidyr permitieron generar gráficos informativos y dinámicos que ayudan a identificar tendencias, patrones y relaciones entre variables clave, como el número de casos positivos, activos, hospitalizados y fallecidos.

A través de gráficos unidimensionales, bidimensionales, multidimensionales, de facetas y compuestos, se exploraron diferentes aspectos de la pandemia, desde la evolución temporal de los casos hasta la distribución por género y edad. Estas visualizaciones han proporcionado una visión completa y detallada de la situación epidemiológica en Costa Rica, permitiendo una mejor comprensión de la magnitud y el impacto de la crisis sanitaria.

En resumen, este proyecto destaca el papel crucial de la visualización de datos y el análisis exploratorio en la comprensión y respuesta a crisis sanitarias como la pandemia de COVID-19.

En el siguiente link, se encuentran publicados todas las gráficas generadas: https://casoscovid-proyecto1.netlify.app