Análisis exploratorio de datos con `tidyverse`

Parte 1: Fundamentos de análisis exploratorio en `R`

Rodrigo Zepeda-Tello

2022-10-07

Mostramos cómo funciona `dplyr` para filtrar (`filter`), seleccionar (`select`), mutar (`mutate`), agrupar (`group_by`), y resumir (`summarise`) bases de datos en `R`

> **Note**
>
> Los datos están disponibles en el Github y en Dropbox

> **Warning**
>
> Si aún no cuentas con una instalación de `tidyverse` dentro de `R` corre la siguiente instrucción:
>
> ```
> install.packages("tidyverse")
> ```

El flujo de trabajo de trabajo con datos

En general el flujo de trabajo de un análisis de datos se divide en tres componentes principales:

1. **Preparación de los datos** lo que incluye la recolección (no discutida aquí), la **importación** de los datos y la **limpieza** inicial (por ejemplo el poner nombres a

las columnas u homologar mayúsculas y minúsculas) para generar una base de trabajo.

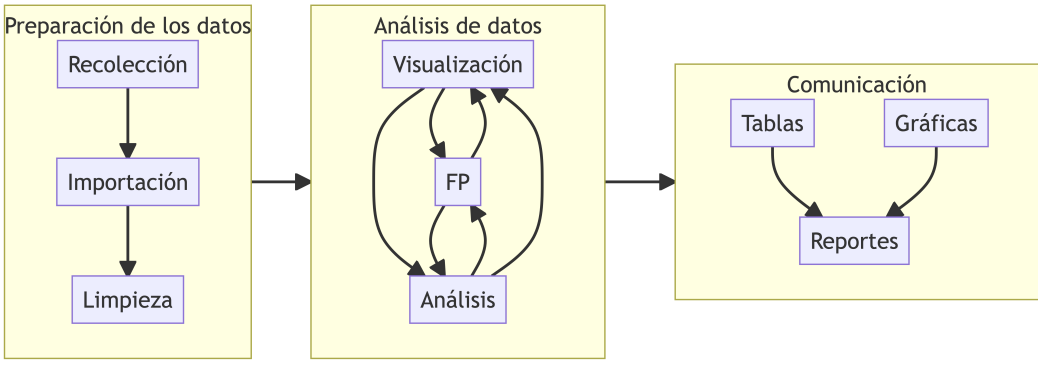
> ⚠ Warning
>
> Una vez recolectados la recomendación es que las bases de datos **no se toquen** una vez se tiene la información. Entre menos modifiquemos la base de datos hay menor probabilidad de cometer errores y accidentes que resulten en **pérdida de información**.

2. **Análisis de datos** incluye tres pasos que fluyen en cualquier orden:

a. Formulación de preguntas *¿qué me gustaría saber de mis datos?*
b. Visualización de los datos *¿puede una gráfica ayudarme a responder mi pregunta u orientarme hacia qué analizar?*
c. Análisis de los datos: *¿qué resumen de la información (por ejemplo una tabla o un promedio) resulta eficaz para presentar lo que me interesa?*

3. Una vez se tienen los datos analizados continuamos a la parte de **comunicación** donde buscamos **generar tablas, gráficas y reportes** (entre otros) que comuniquen nuestros datos al público.

El siguiente diagrama intenta resumir el flujo de información



Armado de un proyecto

Lectura de bases de datos

`R` sirve para abrir cualquier tipo de dato. En particular es posible leer bases de datos desde `Excel`, `csv`, `txt`, `Stata`. También (aunque no lo veremos) puede leer imágenes, videos, datos provenientes de documentos `pdf`, páginas web, mapas, etc. Al día de hoy no he encontrado un sólo tipo de dato que no pueda leer `R`.

Las bases de datos