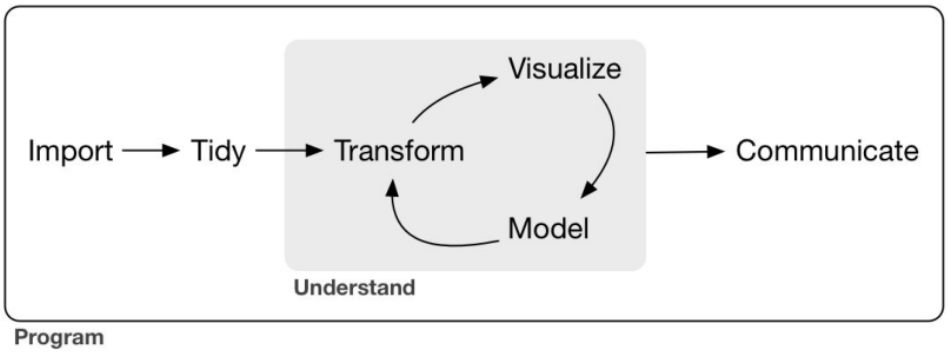
ICO 187 ANÁLISIS DE DATOS

Clase 06: ¿Qué es análisis de datos?

Año: 2021
Profesor: Sebastián Egaña

1. ¿Qué es análisis de datos?

Se puede definir análisis de datos como la búsqueda y generación de datos significativos que puedan ayudar a la toma de decisiones. Dicha definición sería principalmente orientada a los negocios. Por otra parte, en términos procedimentales, análisis de datos se define como una serie de pasos para la generación de información. El siguiente cuadro, responde a dicho a procedimiento en base a Wickham & Grolemound (2017):



Fuente: Wickham & Grolemound (2017)

Estos pasos, a pesar de estar pensados en torno a Rstudio aplica para cualquier herramienta orientada al análisis de datos. Hablemos un poco de cada uno de estos pasos:

- Import: Importar refiere a la toma de la información ya sea por parte de alguna base de datos, conexión SQL, webscrapping o web API para ser cargados dentro del programa en donde trabajarémos.
- Tidy: Ordenar, refiere al desarrollo de procedimientos para que la data que se tiene sea consistente con la semántica de los datos que se obtuvieron. Recordar que cada columna es una variable y cada fila una observación. En dicho punto ya hemos avanzado en el curso.
- Transform: Transformar, refiere a la generación de nuevas varriables y calcular estádisticas descriptivas. Ordenar y transformar, a veces se denomina wrangling o disputar o reñir, porque a veces dichos procedimientos para genarar la data se trasnforman en una pelea.

Después de esto, existen dos formas de poder obtener información de la base de datos: visualizar o modelar.

- Visualization: Visualizar, refiere a la generación de una descripción visual de los datos obtenidos. Una buena visualización puede resumir de buena manera un set de datos más complejo, pero también puede esconder las preguntas relevantes sobre los datos. Su escalabilidad no es alta, debido a que necesitan de alguien para interpretar.
- Models: Modelar, es un paso complementario a la visualización. Una vez se tiene una pregunta específica para el set de datos, se desarrolla la modelación para contestarla. Por lo general se necesita de modelos matemáticos o

computacionales para esto.

- Communication: Comunicar, es la parte critica de cada proyecto, debido a que dan lo mismo las visualizaciones o los modelamientos si no se pueden comunicar los resultados.

Todos estos pasos, pueden ser desarrollados en programas como Excel y Rstudio, con mayor o menor complejidad. Sed ebe tener en claro, que por lo general una sola herramienta no nos dejará desarrollar el 100 % de nuestro trabajo como análista; siempre deberemos estar recurriendo a otras herramientas como para desarrollar nuestros análisis.

2. Sobre el Big data

Por lo general, el término de big data es super utilizado por la supuesta importancia de la gran cantidad de datos que se generan. Desde nuestro punto de vista, o sea el mío, lo que importa es quién tiene la capacidad o no de poder procesar una gran cantidad de datos. Estamos hablando de set de datos mayores a 10Gb.

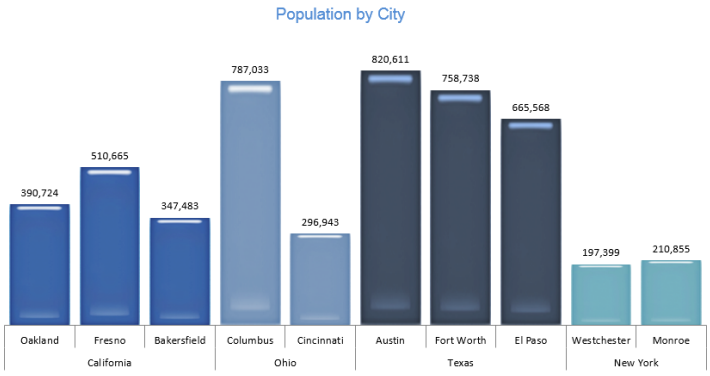
Para el caso de Excel, no existe un limitante en términos de capacidad, pero si en términos de filas y columnas (1.048.576 filas y 16.384 columnas). Por lo tanto, Excel no es una herramienta para el análisis de Big Data.

- Técnicas para el análisis de Big data: solo como datos, una posibilidad corresponde a la extracción de una parte del gran se de datos, y así poder realizar los análisis para replicarlos en la data completa.

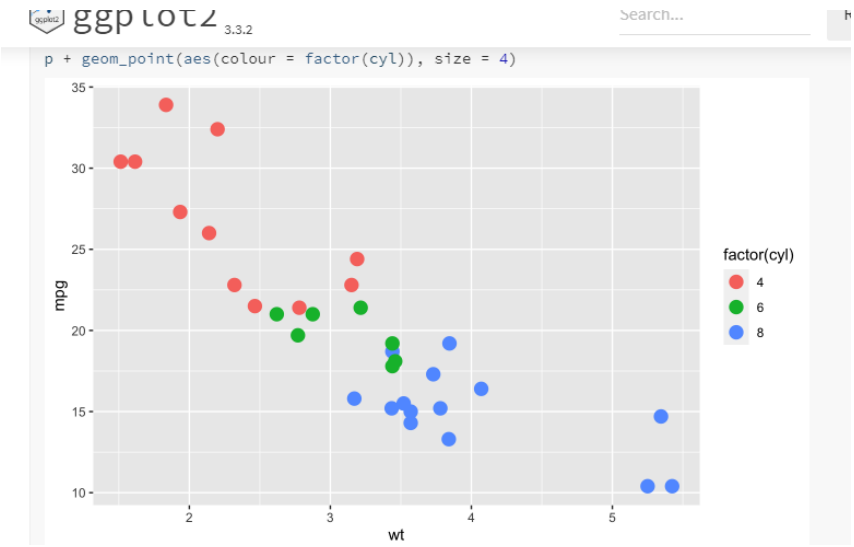
3. Sobre Excel

Ya hemos visto algunas de las limitantes de Excel para el análisis de datos (filas y columnas), pero hay otros puntos que deben ser abordados de manera particular:

- Visualización: A pesar de tener una gran capacidad de este campo, Excel comienza a tener problemas cuando existen dos o más variables de agrupamiento. Veamos la siguiente imagen:



Este problema, R lo soluciona a través de la aplicación de distintos colores a cada grupo:



Otro tema puede estar relacionado con la inserción de datos, anotaciones dentro de los gráficos.

Por último, Excel no es propiamente una herramienta desarrollada para generar reportes. Se puede intentar forzarla a esto, o desarrollar en conjunto con Word o Power Point.

4. Importar

En el caso particular de Excel, la importación de datos no es generalmente un problema. Se debe tener en cuenta los tipos de datos que puede leer: txt, csv, xls, xlsx, etc.

Otro caso corresponde al querer importar datos de bases de datos o desde la web.

Veamos un pequeño ejemplo, considerando el siguiente link:

https://www.previred.com/web/previred/indicadores-previsionales

Veamos otro ejemplo desde la página del Banco Central:

https://si3.bcentral.cl/Siete/ES/Siete/Cuadro/CAP_ESTADIST_MACRO/MN_EST_MACRO_IV/PEM_TC

La simplicidad de estos ejemplos, no debe confundirlos; por lo general el Webscrapping no es tan fácil; menos en Excel.

5. Ordenar

Se debe intentar realizar comprobaciones de los datos de ser necesario.

- Comprobar algún calculo realizado, generandolo de manera distinta.
- Generar una celda que confirme el calculo realizado.
- Funciones que nos permitan validar un dato, o varios datos.

En la mayoría de los casos, es necesario realizar un reemplazo de múltiples valores a la vez, ¿sabe usted como hacer esto en Excel?

Si no lo sabe, find it!

Por otra parte, dependiendo el tipo de data con la que trabajamos pueden existir valores que se repiten como también valores que no deberían repetirse nunca (dependiendo de la naturaleza de la base de datos).

Por ejemplo, dentro del Censo se maneja el concepto de Hogar e Individuo, ¿cuál de los dos debería tener un identificador único? ¿nuestro análisis será en base al hogar o al individuo?

6. Fechas Relevantes

Unidad	Evaluación	Ponderación	Fecha
Unidad I	Evaluación diagnóstica		25/03/2021
	Evaluación Individual Participación	(5 %)	05/04/2021
	Evaluación Grupal	(15 %)	27/04/2021 - 04/05/2021
	Evaluación Individual - Sumativa I	(30 %)	11/05/2021
Unidad II	Evaluación Formativa		13/05/2021
	Evaluación Individual Participación	(5 %)	27/05/2021
	Evaluación Grupal	(10 %)	08/06/2021 - 15/06/2021
	Evaluación Individual - Sumativa II	(15 %)	17/06/2021
Unidad III	Evaluación Formativa		22/06/2021
	Evaluación Individual Participación	(5 %)	24/06/2021
	Evaluación Individual Sesión I- Sumativa III	(15 %)	08/07/2021
	Evaluación Individual Sesión II- Sumativa III	(15 %)	13/07/2021