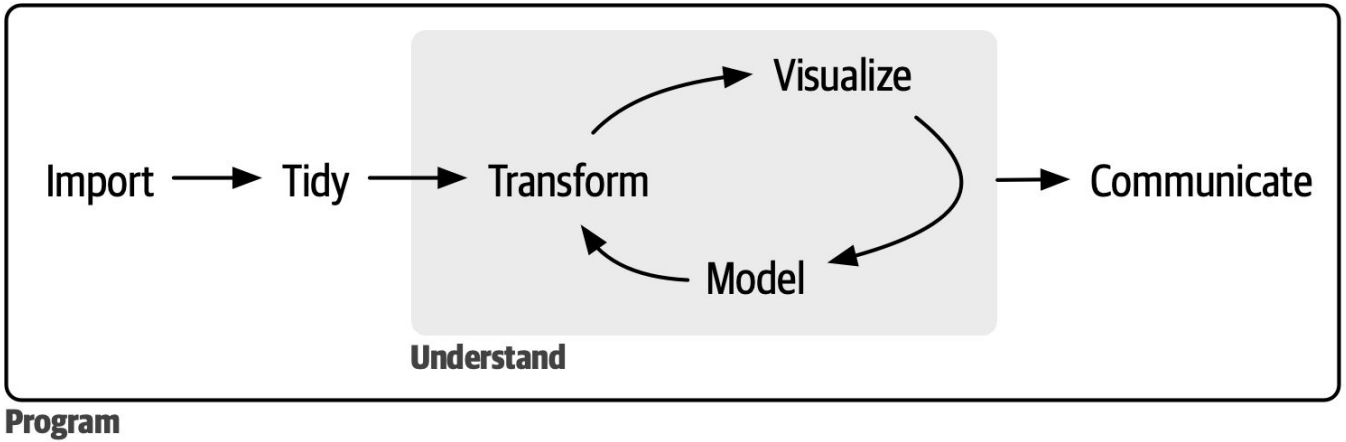
R y Big Data

Introducción a R y Big Data

Introducción a R y Big Data

R es un lenguaje de programación y un entorno de software ampliamente utilizado para el análisis de datos y la visualización. Su relevancia en el campo del Big Data se debe a su capacidad para manejar grandes conjuntos de datos y realizar análisis estadísticos avanzados.

Proceso de Data Science



Importar datos R

En primer lugar, hay que importar los datos a R. Esto suele significar que se toman los datos almacenados en un archivo, una base de datos o una interfaz de programación de aplicaciones web (API) y se cargan en un marco de datos en R. Si no se pueden introducir los datos en R, no se puede hacer ciencia de datos con ellos.

Limpiar datos R

Limpiar los datos significa almacenarlos de forma coherente, de modo que la semántica del conjunto de datos coincida con la forma en que se almacenan.

En resumen, cuando los datos están limpiados, cada columna es una variable y cada fila es una observación. Los datos limpios son importantes porque su estructura coherente permite centrar los esfuerzos en responder a preguntas sobre los datos, en lugar de luchar por darles la forma adecuada para las distintas funciones.

Transformar datos en R

La transformación incluye acotar las observaciones de interés (como todas las personas de una ciudad o todos los datos del último año), crear nuevas variables que sean funciones de variables existentes (como calcular la velocidad a partir de la distancia y el tiempo) y calcular un conjunto de estadísticas de resumen (como recuentos o medias).

Juntas, la limpieza y la transformación se denominan "wrangling".

Siguientes pasos

Una vez que se dispone de datos limpios con las variables necesarias, existen dos motores principales de generación de conocimiento: la visualización y la modelización. Ambos tienen puntos fuertes y débiles complementarios, por lo que cualquier análisis de datos real iterará entre ellos muchas veces.

Visualización de datos

Una buena visualización te mostrará cosas que no esperabas o te planteará nuevas preguntas sobre los datos. Una buena visualización también puede indicarte que estás haciendo la pregunta equivocada o que necesitas recopilar datos diferentes.

Las visualizaciones pueden sorprender, pero no se adaptan especialmente bien porque requieren que un ser humano las interprete.

Modelado de datos

Los modelos son herramientas complementarias de la visualización. Una vez que sus preguntas son lo suficientemente precisas, puede utilizar un modelo para responderlas.

Los modelos son fundamentalmente herramientas matemáticas o computacionales, por lo que suelen ser escalables. Incluso cuando no es así, suele ser más barato comprar más ordenadores que más cerebros.

Comunicación

El último paso de la ciencia de datos es la **comunicación**, una parte absolutamente crítica de cualquier proyecto de análisis de datos. No importa lo bien que tus modelos y tu visualización te hayan llevado a comprender los datos si no eres capaz también de comunicar tus resultados a los demás.

Programación

Alrededor de todas estas herramientas está la programación. La programación es una herramienta transversal que se utiliza en casi todas las partes de un proyecto de ciencia de datos.

No es necesario ser un programador experto para ser un científico de datos de éxito, pero aprender más sobre programación merece la pena porque convertirse en un mejor programador permite automatizar tareas comunes y resolver nuevos problemas con mayor facilidad.

FIN