Introduccion al Business Analytics

Clase 2: Tareas y Roles de BA

Eduard F. Martinez Gonzalez, Ph.D.

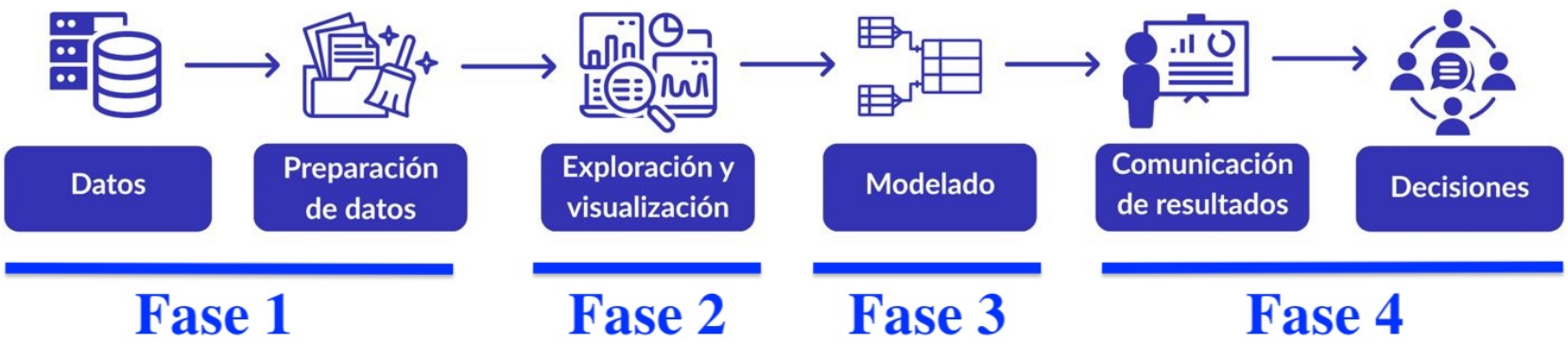
Departamento de Economía, Universidad Icesi

August 8, 2025

¿Qué vimos en la Clase 01?

- Discutimos cómo el crecimiento exponencial de los datos ha transformado la toma de decisiones en las organizaciones.
- Introdujimos el concepto de **Business Analytics** como puente entre los datos y la acción estratégica.
- Analizamos casos reales como el **Netflix Prize**, donde los datos fueron clave para personalizar recomendaciones.
- Presentamos la estructura del curso, incluyendo unidades temáticas, metodología y evaluación.
- Conversamos sobre los distintos tipos de tareas analíticas: clasificación, predicción, segmentación, detección de anomalías y optimización.
- Finalizamos con una visión general del **flujo analítico**, desde la pregunta de negocio hasta la decisión informada.

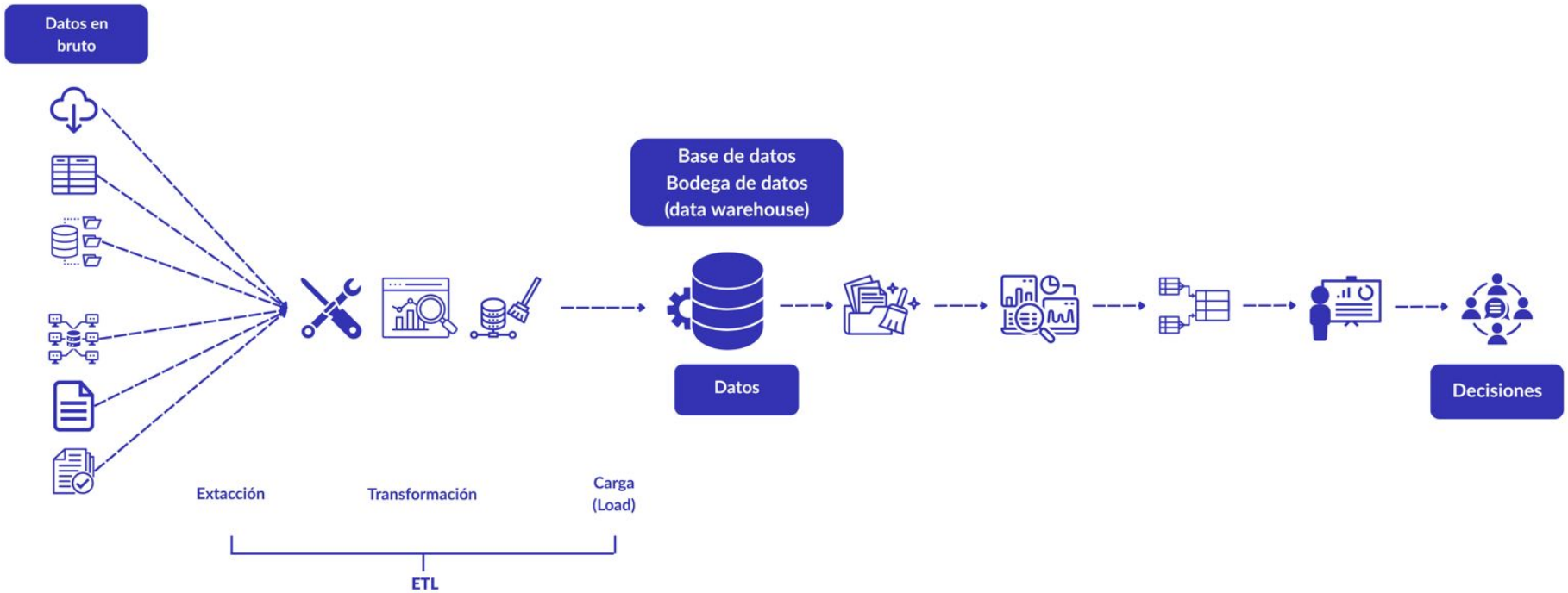
Fases del Proceso Analítico de Datos



Fase 1

Objetivo: Obtener un conjunto de datos limpio, homogéneo y bien documentado, listo para el análisis exploratorio y la construcción de modelos.

Fase 1: Extract, Transform, and Load (ETL)



Fase 1: Extract, Transform, and Load (ETL)

- **Extract (Extracción):**
  - Obtención de datos de múltiples fuentes: bases de datos relacionales, archivos CSV, APIs, sistemas legados.
  - Importancia: asegurar amplitud y relevancia de la información disponible.
- **Transform (Transformación):**
  - Limpieza: manejo de valores nulos, eliminación de duplicados, corrección de errores.
  - Enriquecimiento: agregaciones, normalizaciones, codificaciones categóricas.
  - Integración: unificación de formatos y esquemas entre diferentes fuentes.
- **Load (Carga):**
  - Almacenamiento en data warehouses o data lakes optimizados para análisis.
  - Garantía de disponibilidad y rendimiento para usuarios analíticos.

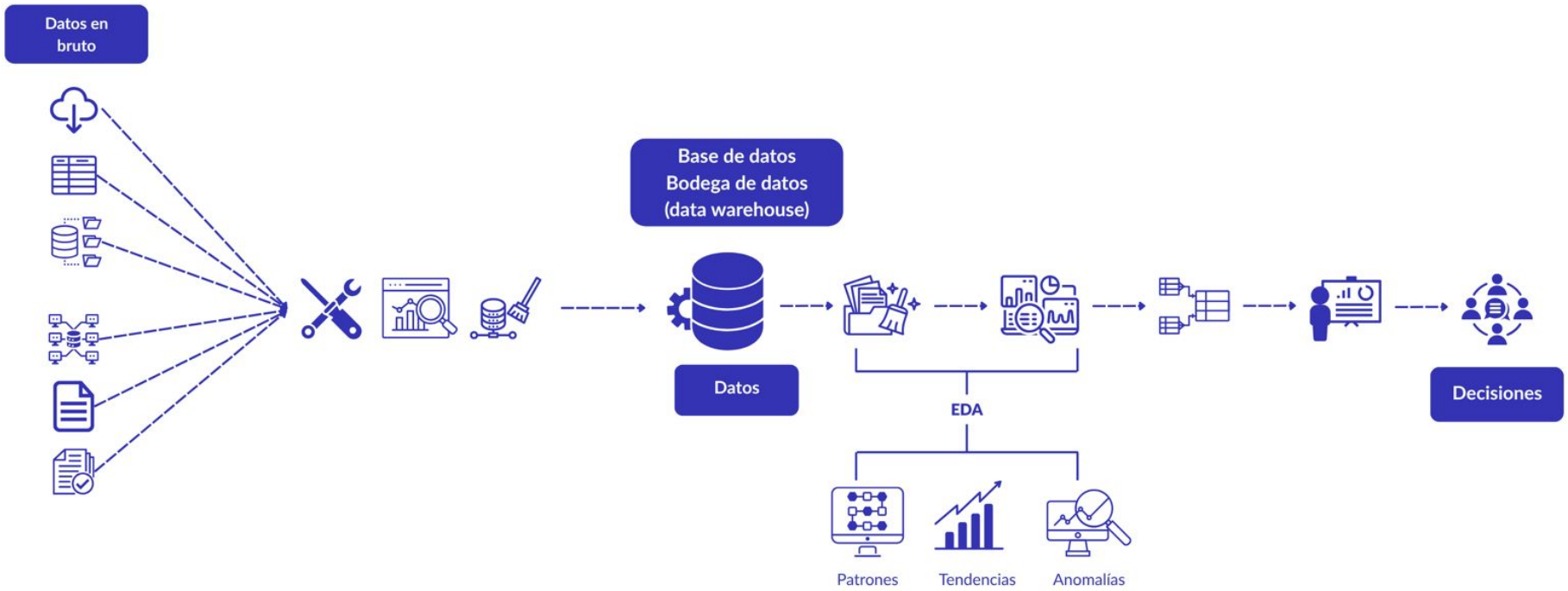
Fase 1: Limpieza y Preparación de Datos

- **Corrección de errores:** Identificar valores fuera de rango o inconsistentes (fechas futuras en registros históricos, edades negativas, montos excesivamente altos).
- **Manejo de valores faltantes:**
  - Cuantificar porcentaje de datos faltantes por variable.
  - Eliminar observaciones completas, completar con promedios o tendencias, o mantener el dato faltante como categoría especial.
- **Eliminación de duplicados:** Detectar casos repetidos por identificadores únicos (identificadores de cliente, número de transacción).
- **Integración de múltiples fuentes:** Consolidar datos internos con encuestas u otros conjuntos de datos externos (proveedores, datos públicos).
- **Transformaciones de variables:**
  - Escalar o normalizar valores si las magnitudes difieren ampliamente.
  - Convertir variables numéricas en rangos o categorías.
- **Validación de calidad de datos:** Analizar estadísticas básicas para cada variable (media, mediana, cuartiles, proporción de categorías).

Fase 2: Exploratory Data Analysis (EDA)

Objetivo: Descubrir patrones, tendencias y relaciones clave que guíen la estrategia de modelado.

Fase 2: Exploratory Data Analysis (EDA)



Fase 2: Exploratory Data Analysis (EDA) / Exploración y Visualización de Datos

- **Análisis descriptivo:**
  - Explorar distribuciones y medidas de tendencia central y dispersión.
  - Identificar correlaciones y posibles variables redundantes.
  - Detectar outliers que puedan sesgar el análisis.
- **Visualizaciones clave:**
  - Histogramas para ver forma y concentración de datos.
  - Scatterplots para relaciones bivariadas.
  - Boxplots para comparar distribuciones entre grupos.
  - Mapas de calor para visualizar patrones de correlación.
- **Iteración y validación:**
  - Formular hipótesis a partir de los hallazgos iniciales.
  - Ajustar segmentaciones o filtrar datos para profundizar el análisis.
  - Refinar las preguntas de negocio según lo observado.
- **Preparación para el modelado:**
  - Seleccionar las variables más prometedoras.
  - Definir transformaciones o nuevas variables derivadas.
  - Documentar insights y posibles riesgos antes de entrenar modelos.

Fase 3: Modelado Predictivo

Objetivo: Construir y validar modelos que generalicen correctamente y respalden decisiones basadas en datos.

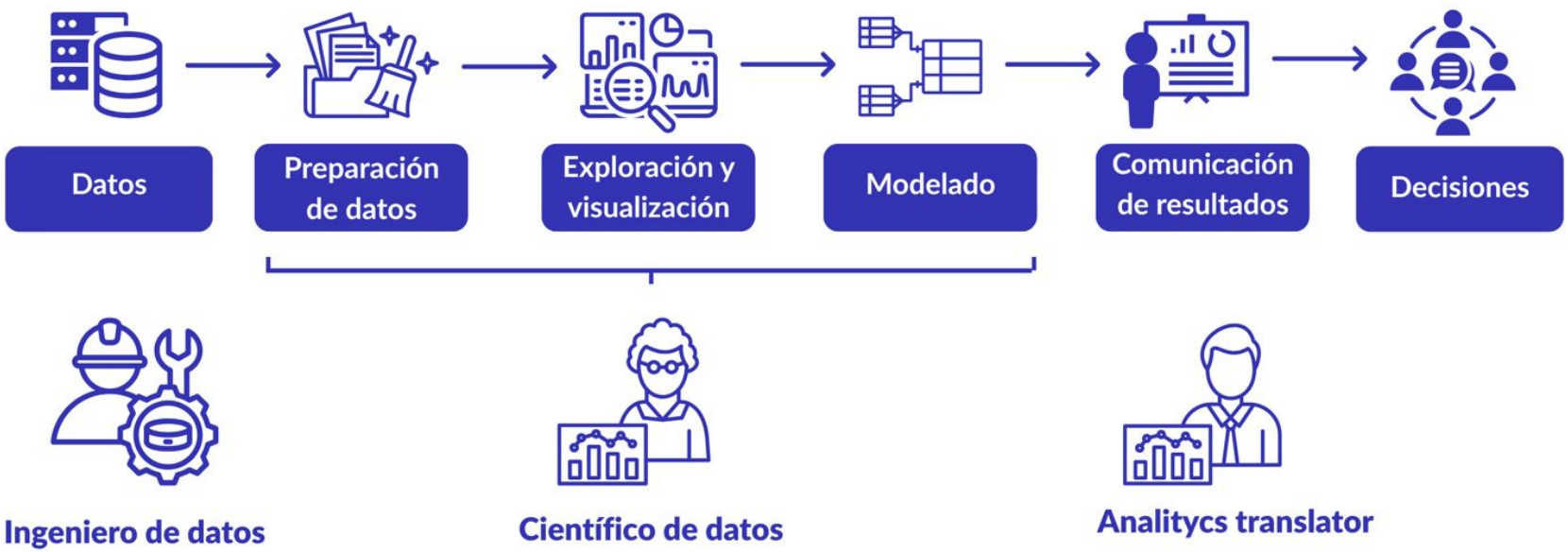
1. **Selección de algoritmos:** Clasificación, Regresión, Clustering, Series de tiempo
2. **Entrenamiento y validación:**
   - Dividir datos en conjuntos de entrenamiento y prueba.
   - Utilizar validación cruzada para estimar desempeño fuera de muestra.
   - Ajustar hiperparámetros y evitar sobreajuste (overfitting).
3. **Métricas de evaluación:**
   - Exactitud (accuracy), precisión (precision) y exhaustividad (recall) para clasificación.
   - Error cuadrático medio (RMSE) y error absoluto medio (MAE) para regresión.
   - Medidas de calidad de agrupamiento: Silhouette, Davies–Bouldin.
4. **Preparación para la implementación:**
   - Documentar supuestos y limitaciones del modelo.
   - Definir un plan de monitoreo y retraining en producción.

Fase 4: Análisis y Toma de Decisiones

Objetivo: Cerrar el ciclo analítico convirtiendo insights en acciones concretas de negocio.

1. **Interpretación crítica**
   - Evaluar significancia estadística versus relevancia de negocio.
   - Identificar posibles sesgos o limitaciones en los resultados.
2. **Comunicación efectiva**
   - Diseñar reportes y dashboards con foco en la acción.
   - Contar una historia (storytelling) que conecte datos y decisión.
   - Adaptar mensaje a audiencia: ejecutivos, técnicos o clientes.
3. **Recomendaciones accionables**
   - Formular propuestas claras: "Qué hacer", "Cuándo" y "Cómo medirlo".
   - Priorizar iniciativas según impacto y factibilidad.
4. **Monitoreo y retroalimentación**
   - Definir métricas de seguimiento para evaluar la implementación.
   - Establecer ciclos de revisión y actualización de modelos.

Roles Profesionales en Analítica de Datos



Especialista en Preparación de Datos (Data Engineer)

- **Diseño y Mantenimiento de Pipelines:** - Estructura flujos ETL/ELT para extraer datos de fuentes internas y externas, transformarlos según estándares de calidad y cargarlos en repositorios de análisis.
- **Gestión de Infraestructura de Datos:** - Administra y optimiza data warehouses y data lakes, así como bases de datos SQL/NoSQL, garantizando rendimiento y escalabilidad.
- **Control de Calidad e Integridad:** - Implementa validaciones automáticas, manejo de duplicados y trazabilidad de cambios para asegurar datos limpios y confiables.
- **Orquestación y Monitoreo:** - Coordina herramientas de orquestación (Airflow, NiFi) y monitoreo en tiempo real para detectar fallos y automatizar recargas.
- **Habilidades Clave:** - Dominio de SQL y NoSQL, arquitecturas distribuidas (Hadoop, Spark), manejo de metadatos y estándares de documentación.

Especialista en Análisis y Exploración (Data Analyst)

- **Análisis Descriptivo:** - Calcula medidas de tendencia central (media, mediana), dispersión (rango, desviación estándar) y distribuciones para entender la estructura de los datos.
- **Exploración y Visualización:** - Genera gráficos clave (histogramas, boxplots, scatterplots, mapas de calor) para identificar patrones, outliers y relaciones.
- **Informes y Dashboards:** - Diseña reportes ejecutivos y dashboards interactivos que sintetizan hallazgos y facilitan la toma de decisiones.
- **Comunicación de Resultados:** - Presenta insights de forma clara a stakeholders, destacando implicaciones de negocio y próximos pasos.
- **Habilidades Clave:** - Herramientas de análisis: R o Python (bibliotecas de visualización y manipulación de datos). - Plataformas de reporting: Excel avanzado, Tableau o Power BI. - Competencias de storytelling y presentación.

Esp. en Modelado y Ciencia de Datos (Data Scientist)

- **Desarrollo de Modelos Avanzados:** Diseña y entrena modelos predictivos y prescriptivos utilizando técnicas de machine learning y estadística avanzada.
- **Validación y Optimización:** Compara algoritmos (clasificación, regresión, clustering) mediante métricas de desempeño y ajuste de hiperparámetros.
- **Diseño de Experimentos:** Planifica y ejecuta pruebas A/B y validaciones cruzadas para evaluar causalidad y robustez de los modelos.
- **Interpretación y Comunicación Técnica:** Traduce resultados complejos en insights comprensibles y recomendaciones para equipos de negocio.
- **Habilidades Clave:** Python/R avanzado, frameworks de ML (scikit-learn, TensorFlow, caret), estadística inferencial y generación de pipelines reproducibles.

Business Analyst / Data Translator

- **Recolección de Requisitos:** - Entiende necesidades de negocio y traduce objetivos en requisitos analíticos precisos.
- **Conexión Técnico–Negocio:** - Actúa como puente entre equipos de datos y stakeholders, garantizando alineación con la estrategia.
- **Traducción de Insights:** - Convierte resultados complejos en recomendaciones claras, priorizadas según impacto y factibilidad.
- **Storytelling de Datos:** - Estructura narrativas visuales que contextualizan hallazgos y respaldan decisiones estratégicas.
- **Elaboración de Material Ejecutivo:** - Diseña informes y dashboards adaptados a diferentes audiencias (ejecutivos, técnicos, operativos), resaltando hallazgos clave y riesgos.
- **Habilidades Clave:** - Visión de negocio, comunicación persuasiva, facilitación de workshops, herramientas de visualización (Power BI, Tableau) y gestión de proyectos analíticos.

Tarea en Parejas: Integrando Fases y Roles Analíticos

Tras revisar las **Fases del Proceso Analítico** y los **Roles en BA**, realicen esta actividad *en parejas* (30 min):

1. **Taller:**
   - Definan una *pregunta de negocio* que quieren responder.
   - **Desglose las Fases necesarias para responderla:**
     - ★ Por cada fase (Preparación, Exploración, Modelado, Resultados), describan brevemente las actividades clave.
   - **Asignación de Roles y Tareas:**
     - ★ Indiquen qué haría cada rol dentro del proyecto: Data Engineer, Data Analyst, Data Scientist, Data Translator.
2. **Entrega:**
   - Compilen un único PDF que contenga las siguientes secciones:
     1. Pregunta de negocio
     2. Fases del proceso analítico
     3. Roles y tareas asignadas
   - Recuerden incluir, al inicio del PDF, el nombre y código de cada estudiante; luego suban el archivo a Intu antes de que finalice el tiempo.

Tarea Diagnóstica de R

Objetivo: Evaluar de manera inicial su nivel de manejo de R y familiaridad con el entorno de trabajo.

- **Contenido del Test:**
  - Creación y manipulación de vectores y data frames.
  - Operaciones básicas de filtrado y agregación.
  - Uso de funciones básicas (aritméticas, estadísticas y de análisis de estructuras).
- **Formato y Duración:**
  - Ejercicio práctico en línea.
  - Tiempo estimado: 15–20 minutos.
- **Puntuación:**
  - Entregar el test en cualquier estado otorga una nota de **5** en la actividad.
  - Quien no entregue recibe una nota de **1**.
- **Acceso al Test:** Disponible [aquí]
- **Plazo de Entrega:** Deberan subir a Intu el código de R en un archivo script (formato .R) antes del inicio de la próxima clase.

Entorno de Trabajo Instalado

Para trabajar sin contratiempos, asegúrense de contar con:

- **R (versión 4.0 o superior)**
- **RStudio**
- **Paquetes esenciales:**
  - `tidyverse`
  - `skimr`
- **Verificación del Entorno:**
  - Ejecute `R.version.string` en la consola para confirmar la versión de R.
  - Pruebe la instalación con `install.packages("tidyverse")` y cargue el paquete `library("tidyverse")`.
- **Guía de Instalación Completa:**
  - Disponible [aquí]

Nota: En las salas de cómputo de la Universidad, R, RStudio y los paquetes `tidyverse` y `skimr` ya están preinstalados.

Paquete datacienfi: Datos Públicos en R

Paquete en desarrollo de CIENFI para descargar, transformar y preparar automáticamente bases de datos públicas (DANE, Datos Abiertos, etc.), de forma estandarizada y reproducible.

Instalación:

```
## instalar devtools
install.packages("devtools")

## instalar datacienfi desde github
devtools::install_github("cienfi-icesi/datacienfi")

## cargar la librería datacienfi
library(datacienfi)

## Ejemplo: Descargar datos de pruebas Saber 11.
notas_saber <- get_notas()
```

Guía de Instalación Completa: Disponible [aquí]

Lectura obligatoria: Introducción a R

Revisen este recurso para consolidar los fundamentos de R antes de la próxima clase:

- **Entorno de Trabajo**
  - Diferencias entre consola y editor de scripts.
- **Objetos Básicos**
  - Tipos de datos atómicos: numérico, carácter y lógico.
  - Estructuras compuestas: vectores, factores, listas y data frames.
- **Operaciones Esenciales**
  - Asignación de variables y nombres de objetos.
  - Funciones integradas: cálculo de estadísticos y manipulación de datos.
  - Uso de la ayuda: `?` y `help()`.
- **Gestión de Paquetes**
  - Instalación y carga de librerías (`install.packages()`, `library()`).
- **Acceso al Material** Disponible [aquí]

Lectura obligatoria: Tidy Data

Esta guía práctica muestra cómo organizar y transformar tablas usando la gramática "tidy":

- **Acceso a columnas:** uso de la función `$` para crear y modificar variables.
- **Creación de variables:** `mutate()` para derivar nuevas columnas (operaciones, condicionales, `case_when()`).
- **Selección y filtrado:** `select()` y `filter()` para aislar columnas y filas según condiciones o patrones.
- **Reordenamiento:** `arrange()` para ordenar datos por valores de variable.
- **Combinación de tablas:** funciones `join()` (`left_join`, `inner_join`, etc.) para unir conjuntos de datos por claves.

Acceso al Material Disponible [aquí]

Lectura obligatoria: Guía de Prácticas Reproducibles

- **Estructura de Repositorio**
  - Carpeta `input/`: código fuente y datos sin procesar.
  - Carpeta `output/`: resultados generados (gráficos, tablas, reportes).
- **Mapeo Código–Datos**
  - Cada archivo en `output/` debe corresponder a un script en `input/`.
  - Uso de nombres claros y metadatos para rastrear procesos.
- **Documentación de Scripts**
  - Encabezado con autor, fecha y propósito.
  - Comentarios que expliquen cada bloque de código.
- **Reproducibilidad Técnica**
  - Fijar semillas (`set.seed()`) al inicio de cada script.
  - Control de versiones con Git/GitHub y archivo `.gitignore`.
- **Buenas Prácticas**
  - Uso de proyectos en RStudio y rutas relativas.
  - Gestión de paquetes.

Acceso al Material Disponible [aquí]

Preparación para la Próxima Clase

Para aprovechar al máximo la próxima sesión, por favor:

- **Tarea Diagnóstica de R**
  - Test de R: Disponible [aquí]
- **Entorno de Trabajo Instalado**
  - R (versión 4.0 o superior)
  - RStudio
  - Paquetes: `tidyverse` y `skimr`
- **Librería** `datacienfi`
  - Guía de instalación: Disponible [aquí]
- **Revisar Material Compartido**
  - Introducción a R: Disponible [aquí]
  - Tidy Data: Disponible [aquí]
  - Guía de Prácticas Reproducibles: Disponible [aquí]