Fuente: <https://www.netacad.com/courses/data-analytics-essentials?courseLang=en-US>

# Tipos de análisis

Descriptivo: ¿Que paso?

Predictivo: ¿Que podría pasar?

Prescriptivo: ¿Que debemos hacer a continuación?

Diagnóstico: ¿Porque sucedió esto?

## Análisis descriptivo

Este se usa para observar un set de datos y sus características, con el cual se puede realizar reportes sobre eventos pasados. Este tipo de análisis resume una gran cantidad de datos los cuales sirven para responder las preguntas “como” y “cuando”. También puede ser utilizado para responder preguntas específicas, para problemas específicos.

El proceso de análisis descriptivo es:

1. Realizar una pregunta sobre datos históricos que respondan una pregunta especifica.
2. Identificar los datos requeridos para responder dicha pregunta.
3. Reunir y preparar los datos.
4. Analizar los datos.

**Análisis Predictivo**

Este se usa para predecir eventos futuros en base a información en tiempo real o histórica en base a las probabilidades, también se utiliza para identificar futuras tendencias. El análisis predictivo usa modelos de simulación y pronósticos para sugerir que podría pasar, guiando para tener metas realistas, planeación efectiva, expectativas de manejo de resultado y evitar riesgos. Esta información puede empoderar a ejecutivos y gerentes para tomar decisiones y estrategias proactivas en base a hechos.

El proceso de análisis predictivo es:

1. ¿Responder a preguntas referente a eventos futuros, como “Podemos predecir cuanto se venderá X producto el próximo año?”.
2. Reunir y preparar datos.
3. Desarrollar un modelo para análisis predictivo.
4. Aplicar el modelo a los datos ya preparados.
5. Revisar el modelo y presentar resultados.

**Análisis Prescriptivo**

El análisis prescriptivo se basa en el análisis descriptivo y predictivo al recomendar cursos de acción que generarán el mayor beneficio para la organización. En resumen, el análisis prescriptivo dice qué se debe hacer en una situación determinada. Ayuda a ejecutivos, gerentes y empleados a tomar las mejores decisiones basadas en los datos disponibles.

Un buen ejemplo de análisis prescriptivo es el campo de las aplicaciones de direcciones y mapas basadas en GPS. Estas aplicaciones brindan opciones de ruta a un destino según el volumen de tráfico, las condiciones de la carretera y la velocidad máxima. Luego puede prescribir la mejor ruta en función de objetivos definidos por el usuario, como la distancia más corta o el tiempo más rápido.

**Análisis Diagnostico**

Esta mejora el análisis descriptivo, profundizando en el análisis e intenta descubrir las causas de un evento.

El proceso de análisis diagnóstico es:

1. Identificar anomalías o inconsistencias de un set de datos.
2. Reunir los datos relacionados con estas anomalías.
3. Usar técnicas de estadística para descubrir relaciones y tendencias que puedan explicar estas anomalías.
4. Presentar las posibles causas.

Un ejemplo de esto es analizar el motivo del porque usuarios dan de baja su suscripción a Netflix, para determinar la razón más común de estas bajas.

# El proceso de analizar datos

A continuación, se verá una representación gráfica de 6 pasos para el análisis de datos, algunos de estos pasos pueden llegar a repetirse durante el proceso.

Forma

Descripción generada automáticamente

* **Formular la pregunta:** el proceso de análisis siempre comienza con una pregunta que debe responderse. Algunas preguntas son simples, como "¿qué color de bicicleta es el más popular entre nuestros compradores más jóvenes?". Otras son muy complejas, como "¿por qué ciertos tipos de células cancerosas presentan resistencia a los tratamientos de radiación?".
* **Obtención de los datos:** este paso implica el proceso de localizar y obtener datos que sean relevantes para la pregunta y luego determinar si hay suficientes datos para completar el análisis.
* **Investigación de los datos:** los datos se presentan en muchas formas y provienen de distintas fuentes. Este paso implica determinar si los datos están completos y contienen la información relevante para el análisis.
* **Preparación de los datos:** este paso puede implicar muchas tareas para transformar los datos a un formato adecuado para las herramientas que se utilizarán para analizarlos y presentarlos. Este proceso a veces se denomina “limpieza” de los datos, porque puede haber registros en blanco o errores obvios en el conjunto de datos.
* **Análisis de datos:** el análisis es el proceso de identificar patrones, correlaciones y relaciones contenidas en un conjunto o conjuntos de datos para extraer inferencias y conclusiones. A menudo, el análisis se basa en técnicas estadísticas y herramientas de software como hojas de cálculo y aplicaciones de visualización.
* **Presentación de los resultados:** este suele ser el último paso para los analistas de datos. Es el proceso de comunicar los resultados a los responsables de la toma de decisiones. Esto se puede hacer en forma de informe, representaciones gráficas o una combinación de ambos. En ocasiones, también se le pide al analista de datos que recomiende acciones en función de los resultados.

# Observaciones, variables y valores

Variables: Una variable en el contexto del análisis es cualquier tipo de dato cuyo valor se pueda medir, modificar o controlar escenarios teóricos.

Registros: Los registros de los valores, patrones y ocurrencias de un conjunto de variables son observaciones. El valor o conjunto de valores de una observación especifica se denominan puntos de datos. Cada observación puede considerarse y representarse como un registro en una base de datos o una fila en una hoja de cálculo de Excel. La colección de observaciones constituye el conjunto de datos para su análisis.

Las observaciones suelen tener un propósito y las variables incluidas dependerán de su relevancia para ese propósito. Por ejemplo, si ha perdido a su mascota y ha pedido a otras personas que lo ayuden a buscarla, solo un pequeño conjunto de variables (las características del perro) son relevantes para sus observaciones. Estas características pueden ser:

* **¿Qué tipo de animal es tu mascota?** Es un perro.
* **¿Qué tipo de perro?** Es un Schnauzer.
* **¿De qué color es tu schnauzer?** Es gris.
* **¿Qué tamaño tiene el Schnauzer?** Es un Schnauzer de tamaño mediano.
* **¿Cuánto pesa el Schnauzer?** Pesa 15 kg.

Tabla

Descripción generada automáticamente

Todas las características de un objeto son variables, estas pueden tener múltiples valores. Debido a que el propósito de sus observaciones es buscar a su perro perdido, las observaciones que no cumplen con los criterios requeridos se descartan.

**Tipos de variables**

Cuando buscamos patrones significativos en los datos, con frecuencia buscamos relaciones entre variables. Todas las variables se pueden clasificar según la característica que se esté estudiando.

Las variables serán categóricas o numéricas.

**Las variables categóricas** indican la pertenencia a un grupo determinado y tienen un valor cualitativo discreto o específico. Se clasifican a su vez en dos tipos:

* **Nominales**: Son variables que constan de dos o más categorías discretas cuyo valor se asigna en función de la identidad del objeto. Algunos ejemplos son el género, el color de los ojos o el tipo de animal.
* **Ordinal**: Son variables que constan de dos o más categorías en las que el orden es importante en el valor. Algunos ejemplos son la posición en la clase de los estudiantes (1.º, 2.º, 3.º) o las escalas de encuestas de satisfacción (insatisfecho, neutral, satisfecho).

**Las variables numéricas** son valores cuantitativos:

* **Continuas**: Son variables cuantitativas que se pueden medir a lo largo de un continuo o rango de valores. Hay dos tipos de variables continuas: Las variables de intervalo pueden tener cualquier valor dentro del rango de valores, como por ejemplo la temperatura o el tiempo; Las variables de razón son variables de intervalo especiales en las que un valor de cero (0) puede significar que no existe dicha variable, como por ejemplo los ingresos o el volumen de ventas.
* **Discretas**: Este tipo de variables continuas son cuantitativas, pero tienen un valor específico dentro de un conjunto finito de valores. Algunos ejemplos son la cantidad de sensores activados en una red o la cantidad de automóviles en un estacionamiento.

(**Variables categóricas:**

Imagina que tienes una caja con diferentes tipos de frutas. Las variables categóricas son como etiquetas que le pones a cada fruta para decir a qué grupo pertenece. Por ejemplo:

* **Nominales:** Son como etiquetas sin un orden específico. Podrías poner etiquetas como "manzana", "pera", "banana". No importa el orden en el que escribas las etiquetas, cada fruta sigue siendo la misma.
* **Ordinales:** Son como etiquetas que tienen un orden. Por ejemplo, en una carrera, puedes tener etiquetas como "primer lugar", "segundo lugar", "tercer lugar". Aquí el orden sí importa, porque te indica la posición de cada corredor.

**Variables numéricas:**

Estas variables son como números que miden algo.

* **Continuas:** Son como una regla que puede medir cualquier valor dentro de un rango. Por ejemplo, la temperatura puede ser 20.5 grados, 20.6 grados, etc. No hay límites entre los valores posibles.
  + **Intervalo:** La temperatura es un buen ejemplo de variable de intervalo. Puedes tener valores negativos (bajo cero) y no hay un punto de inicio absoluto (cero grados no significa ausencia de temperatura).
  + **Razón:** Las ventas son un ejemplo de variable de razón. Si tienes 0 ventas, significa que no vendiste nada, es decir, hay una ausencia total de ventas.
* **Discretas:** Son como números que cuentan cosas enteras. Por ejemplo, no puedes tener medio auto en un estacionamiento, solo puedes tener 1 auto, 2 autos, etc.

)

Interfaz de usuario gráfica, Gráfico

Descripción generada automáticamente con confianza media

¿Por qué es importante saber qué tipos de variables hay en su conjunto de datos?

Algunos tipos de análisis y visualización de datos están diseñados para funcionar con determinados tipos de datos. La forma en que elija presentar los resultados del análisis dependerá del tipo de variables utilizadas en los datos. Algunos tipos de variables se prestan mejor a los gráficos de barras, mientras que otros pueden permitir un mayor examen y descubrimiento mediante un diagrama de dispersión. En la figura anterior se pueden ver ejemplos de algunos de los tipos de gráficos sugeridos que representan los diferentes tipos de variables.

# Fuentes de datos

Para realizar la selección de datos relevantes, es necesario encontrar fuentes confiables para los mismos, estos datos pueden estar ya disponibles en un set en una sola fuente o puede ser necesario recopilar datos de diferentes fuentes las cuales se complementen para responder a las preguntas planteadas y las necesidades del análisis del proyecto en cuestión.

Por Ej. Si queremos analizar una película la cual esta basada en un libro, además de obtener los datos relevantes a la película, será necesario obtener datos del libro como la cantidad de ventas, el género, el año del libro, entre otros datos relacionados.

Algunas preguntas que debes plantearte al seleccionar una fuente de datos:

1. ¿Qué puntos de datos son necesarios para informar su análisis?
2. ¿Ya tengo acceso a estos datos o debo buscar un conjunto de datos de otra fuente?
3. ¿Dónde están fuentes confiables y verificables de estos datos?
4. ¿Con qué frecuencia se recopilan y actualizan los datos relevantes?
5. ¿Cómo se licencian los datos para su uso? ¿Tiene algún coste?
6. ¿Los datos están en un formato que puedo usar o convertir para usar con mis herramientas?

**Datos estáticos y en tiempo real (Streaming)**

Existen dos tipos de datos con los que trabajan los analistas: datos estáticos y datos en tiempo real. Los datos que se reciben y almacenan antes de realizar un análisis de estos se consideran datos estáticos. Cuando cada evento se procesa y analiza a medida que se recibe y los resultados posteriores se utilizan o almacenan, los datos se denominan datos en tiempo real.

Ej. Si analizamos las reseñas realizadas en un sitio web, podemos guardarlas en una base de datos para analizarlas de forma posterior a esto le llamamos datos estáticos, en cambio si deseamos analizar a medida que se realizan nuevos comentarios a estos los llamamos en tiempo real.

**Tipos y formatos de datos**

Para poder realizar un análisis comúnmente debemos obtener datos de diferentes fuentes, estas fuentes pueden tener datos formateados diferentes entre sí, por ejemplo, el de una fecha en formato texto que deberemos transformarlo en formato número.

Los tipos de datos son relevantes al momento de realizar cálculos y comparaciones.

***Los tipos de datos más comunes son:***

|  |  |
| --- | --- |
| **Tipo de dato** | **Descripción** |
| String | Datos en formato texto |
| Integer | Números sin decimales |
| Floating point (float) | Números con decimales |
| Date & Time | Datos fecha y hora |
| Boolean | Tipo de dato que solo indica que es True o False |

El tipo de dato Date & Time, es un tipo de dato especial ya que dependiendo de como se maneje el dato este puede tener diferentes formatos regionales.

***Formatos comunes Date & Time:***

|  |  |  |
| --- | --- | --- |
| **Estandarización** | **Representación** | **Ejemplo** |
| ISO 8601 | AAAA-MM-DDThh:mm:mm+|-ss:mm | 2023-01-20T15:20:09 |
| RFC 1123 | DIA DD-LUN-AAAA hh:mm:ss GMT|UT | Viernes 20 de enero 2023 20:20:27 GMT |
| Americano | MM/DD/AAAA hh:mm:ss AM|PM | 20/01/2023 15:20:09 |

# Datos estructurados

Son aquellos datos que mantienen una estructura definida dentro de un archivo o base de datos, siendo fácilmente manipulados y analizados.

Algunas de las características de los datos estructurados incluyen:

* Hay una estructura bien definida y organizada.
* Se puede almacenar en tablas, generalmente dentro de columnas verticales y filas horizontales.
* Se documenta el contenido y el formato de los datos.
* Está organizado en archivos, registros y campos.
* Se puede buscar, ordenar y consultar.
* Los controles de entrada pueden reducir la posibilidad de datos no válidos.



**Archivos CSV**

Son archivos de datos estructurados delimitados por comas u otro símbolo de tabulación, haciendo que el carácter utilizado indique que existe una nueva línea para separar filas y son un archivo de texto plano sin formato estandarizado.



Cada fila también puede denominarse registro. Los archivos CSV se utilizan habitualmente para importar y exportar datos a hojas de cálculo y bases de datos tradicionales, pero también se pueden utilizar como entrada para programas analíticos y herramientas de visualización.

JSON (JavaScript Object Notation) y XML (Extensible Markup Language) también son tipos de archivos de texto sin formato estandarizados y comunes que se utilizan a menudo para representar registros de datos.

**Tipos de archivos estructurados**

* + **Bases de datos relacionales:** Es una colección de tablas las cuales se relacionan unas con otras a través de claves primarias y foráneas.
  + **Logs:** Es archivo histórico de transacciones, errores, inconsistencias, entre otros de un evento ocurrido en una máquina, los archivos de logs son considerados datos estructurados ya que estos se generan con un formato estándar.
  + **Hojas de cálculo**: Archivos estructurados por filas y columnas, los cuales pueden ser relacionados entre si siendo similar a una base de datos estructurada.
  + **Sensor Readings:** Es un formato estandarizado el cual puede variar según el fabricante, estos se separan por un delimitador y pueden estar separados por marcas de tiempo (time-stamp).
  + **Transaccional records:** Estos pueden estar almacenados en diferentes formatos, dependiendo del tipo de transacción y recurso. Algunas transacciones pueden ser generadas manualmente a través de formularios y otras generadas por un dispositivo.

# Datos NO estructurados

Los datos no estructurados carecen de organización, son datos sin procesar y no poseen un esquema predefinido. Los datos no estructurados pueden ser archivos PDF, PPT, revistas, documentos, webs, entre otros. En sí, el texto puede ser buscado, pero no está organizado en un formato predefinido.

Tanto los datos estructurados como los no estructurados son valiosos para las personas, las organizaciones, las industrias y los gobiernos. Es importante que las organizaciones tomen todos los formatos de datos y determinen formas de formatearlos para que se puedan gestionar y analizar.

**Fuentes de datos no estructurados**

Existen muchas fuentes de datos no estructurados en Internet, en forma de texto, imágenes, vídeos y audio. La mayoría de estos datos no están estructurados, lo que significa que no es fácil clasificarlos en una base de datos sin algún tipo de procesamiento. Los datos no estructurados se pueden extraer y procesar de varias formas:

* + Bases NoSql y Repositorios: Este tipo de repositorios se utilizan para almacenar datos en tiempo real.
  + Web Scraping: Esto corresponde a un método de extracción automática de datos en un sitio web a través de su archivo HTML, el web scraping usualmente se usa con crawlers los cuales buscan la información en diferentes sitios hasta encontrar el indicado y extraer su información.
  + API: Muchos sitios web proporcionan acceso a través de API’s para recopilar información de ellos, estas API suelen ser Restful y se puede acceder a ellas a través de peticiones HTTP y se obtiene de estos archivos JSON.

# Preparación de los datos

**Procesos ETL y ELT (Extract, Transform, Load)**

ETL y ELT son dos versiones del mismo proceso para mover datos a través de una canalización. Contienen los mismos pasos, pero en diferente orden para distintos casos de uso.

**Extraer, transformar y cargar (ETL)** es un proceso que consiste en recopilar datos de esta variedad de fuentes, transformarlos y luego cargarlos en una base de datos. Los datos de una empresa pueden encontrarse en documentos de Word, hojas de cálculo, texto sin formato, presentaciones de PowerPoint, correos electrónicos y archivos PDF. Estos datos pueden almacenarse en una variedad de formatos diferentes, lo que dificulta su combinación y análisis, por lo que la transformación se produce antes de la carga.

**Etapas ETL**

**Extraer:** En este paso se localizan y recopilan datos de diversas fuentes para convertirlos a un único formato para su análisis. Los datos pueden extraerse de una base de datos relacional, NoSQL, archivos planos, archivos XML u otros formatos.

**Transformar**: Los datos deben transformarse antes de poder cargarse en un almacén de datos para su análisis. El paso de transformación utiliza reglas para transformar los datos de origen al tipo de datos necesarios para la base de datos de destino.

El paso de transformación también requiere varias tareas adicionales. Algunas de estas tareas son unir datos de varias fuentes, agregar, ordenar, determinar nuevos valores que se calculan a partir de datos agregados y, luego, aplicar reglas de validación.

Algunos ejemplos de limpieza de datos son la eliminación de registros en blanco y la estandarización de formatos como fecha, hora y ubicación.

**Cargar**: Los datos transformados se cargan luego en la base de datos para realizar consultas. El proceso de carga real varía ampliamente, según los tipos de datos de origen, el tipo de base de datos de destino y el tipo de consulta que se va a realizar.

Durante el paso de carga, se aplican las reglas que se han definido en el esquema de la base de datos. Estas reglas comprueban las características necesarias, como la singularidad y la coherencia de los datos o que los campos obligatorios no estén vacíos. Estas reglas ayudan a garantizar que la carga y cualquier consulta posterior de los datos se realice correctamente.

# Excel

**Funciones básicas**

=SUMA(A1:A10): Suma un rango de celdas.

=PROMEDIO(A1:A10): Promedia un rango de valores numéricos.

=CONTAR(A1:A10): Realiza un conteo de elementos dentro de un rango.

=MIN(A1:A10): Devuelve el valor más bajo de un rango.

=MAX(A1:A10): Devuelve el valor más alto de un rango.

=MEDIANA(A1:A10): Devuelve el valor medio de un rango.

=CONTAR.BLANCO(A1:A10): Cuenta las celdas en blanco de un rango.

**Importar datos**

CSV: Los archivos en formatos CSV, son archivos delimitados por comas el cual es un formato de texto multiplataforma.

**CSV en Excel:**

**Abrir archivo:**

Archivos > Abrir > Seleccionar el archivo en nuestro dispositivo > Asistente de importación>

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Separarlos por coma >

Tabla

Descripción generada automáticamente

Formatear Columnas > Finalizar.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Cargándose el archivo por filas y columnas.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

**Cargar como datos a una hoja:**

Datos > De texto/CSV > Seleccionar el archivo en nuestro dispositivo > Importar > Se abre cuadro de datos > Cargar.  
  
Tabla

Descripción generada automáticamente

Cargándose una tabla.

Captura de pantalla de computadora

Descripción generada automáticamente