

# TECNICATURA SUPERIOR EN CIENCIA DE DATOS E IA

ENTREGA FINAL

BLOQUE: ANALISTA DE DATOS

## INTEGRANTES

BENDEZU Ana Valentina

DNI 35530141

-----

BLASICHE Andrés

DNI 31405881

-----

CABRERA Marcos Rodrigo

DNI 31667009

-----

PALOMEQUE Jonathan  
Manuel

DNI 36144579

-----

PALOMEQUE Dalila Macarena

DNI 39733230

-----

TEJEDA Romina Soledad

DNI 34501801

# Índice

<b>INTRODUCCIÓN</b>	<b>1</b>
<b>DESARROLLO</b>	<b>1</b>
Objetivos generales	1
Objetivos específicos	2
Descripción de los datos	2
Exploración de datos	3
Limpieza	3
Análisis y visualización	3
<b>CONCLUSIONES</b>	<b>5</b>
<b>BIBLIOGRAFÍA</b>	<b>5</b>

---

## INTRODUCCIÓN

---

El presente proyecto integra los aprendizajes adquiridos en el módulo Analista de datos, que está compuesto por las materias Estadística y exploración de datos, Ciencia de datos, y Procesamiento de datos.

Durante este primer cuatrimestre se profundizó en la estadística descriptiva, los entornos de trabajo para la ciencia de datos, y las diferentes librerías de Python para llevar a cabo un análisis de los datos recogidos en un dataset.

Para llevar a cabo un análisis exhaustivo de los datos se seleccionó un dataset de registro de ventas, en el cual se detalla las ventas realizadas de una determinada empresa las cuales se encuentran diferenciadas por países, período de tiempo, tipo de comercio, tipo de productos, como así también el precio de venta, costos y la utilidad bruta.

Este dataset fue elegido por su amplia cantidad de registros como así también por contener tanto datos cualitativos como cuantitativos, lo que nos permite realizar diferentes análisis a fin de poner en práctica lo aprendido en la carrera.

---

## DESARROLLO

---

### Objetivos generales

- Realizar un análisis de los datos de ventas presentados por la Empresa.

### Objetivos específicos

- Aplicar métodos de python para limpieza del dataset.
- Realizar un análisis estadístico de los datos.
- Aportar a una visualización de los datos mediante gráficos.

### Descripción de los datos

Para el presente proyecto elegimos un dataset que detalla las ventas realizadas por la empresa multinacional “Cuchufrito”, fabricante de artículos varios tales como calzado, vestimenta, accesorios, entre otros.

Muestra los datos distribuidos entre las siguientes columnas:

**País:** Refiere al país donde fue realizada la venta. Entre ellos se encuentran la mayoría de países de América, y algunos de Asia y Europa.

Tipo de dato: String. ['Argentina' 'Bolivia' 'Brazil' 'Belize' 'Canada' 'Chile' 'China' 'Colombia' 'Costa Rica' 'Germany' 'Dominican Republic' 'Ecuador' 'Guatemala' 'Honduras'

'Italy' 'Kazakhstan' 'Mexico' 'Nicaragua' 'Panama' 'Peru' 'Puerto Rico' 'Paraguay' 'El Salvador' 'United States' 'Uruguay' 'Venezuela' 'Bermuda' 'Aruba']

**Método de orden:** Refiere al medio a través del cual se concretó la venta del producto.

Tipo de dato: String. ['Representante' 'Mail' 'Teléfono' 'Website' 'Especial' 'MercadoLibre']

**Tipo de comercio:** Refiere a las características del comercio donde se efectuó la venta.

Tipo de dato: String. ['Mayorista' 'De proximidad' 'Grandes Almacenes' 'Especializado' 'Hipermercado']

**Línea:** Hace referencia al rubro al que pertenecen un grupo de productos.

Tipo de dato: String. ['Mascotas' 'Infantil' 'Hombre' 'Adultos' 'Mujer' 'Unisex']

**Tipo de producto:** Dentro de la línea, hay una subdivisión que agrupa a determinado tipo de productos.

Tipo de dato: String. ['Juguetes' 'Calzado' 'Vestimenta' 'Lentes' 'Joyas' 'Accesorios']

**Producto:** Refiere a cada producto específico.

Tipo de dato: String

**Año:** Es el año en que se realizó la venta,

Tipo de dato: numérico (entero o Date(Year)) [2019 2020 2021]

**Cuatrimestre:** Período del año en el que se realizó la venta.

Tipo de dato: String ['C1 2019' 'C2 2019' 'C3 2019' 'C4 2019' 'C1 2020' 'C2 2020' 'C3 2020' 'C4 2020' 'C1 2021' 'C2 2021' 'C3 2021' 'C4 2021']

**Ventas:** Se refiere al ingreso bruto por ventas,

Tipo de dato: numérico (decimal o flotante)

**Cantidad:** número de productos vendidos.

Tipo de dato: numérico (entero)

**Utilidad bruta (porcentaje):** Es la utilidad bruta expresada en porcentaje sobre el total de ventas (ganancia bruta).

Tipo de dato: numérico (decimal o flotante)

**Utilidad bruta:** Se refiere a la ganancia que se obtiene de la venta de un producto luego de restarle los costos asociados al mismo.

Tipo de dato: numérico (decimal o flotante)

**Costos:** Gastos asociados a la venta del producto.

Tipo de dato: numérico (decimal o flotante)

Se realiza el cálculo de medidas de tendencia central: media y mediana de las columnas numéricas y medidas de variación: desviación estándar, para identificar la tendencia de los datos y su variación.

## Exploración de datos

Mediante las librerías de Python podemos leer, explorar, limpiar y analizar los datos de un dataset. Se exploran las columnas, el total de los registros, los tipos de datos, para conocer con qué datos contamos al momento de realizar los análisis pertinentes.

## Limpieza<sup>1</sup>

Es importante mencionar que los datos están en crudo por lo cual, requieren de un importante trabajo de limpieza para obtener datos de calidad que permitan el posterior trabajo de análisis.

Se modificó el tipo de dato de las columnas numéricas que estaban como cadena de caracteres. Para ello, anteriormente fue necesario eliminar el formato pesos y cambiar símbolos (',' y '.').

Posteriormente se procedió a analizar las filas con datos vacíos o no registrados. En el contexto de nuestros datos y la cantidad que tenemos, decidimos eliminar los registros que contienen datos 'Nan' ya que la ausencia de estos hace que la trazabilidad de cualquier registro sea muy difícil y al momento de realizar medidas descriptivas podrían distorsionar los datos de manera grosera. Además de eso 65/12000 registros nos da 0,0054. Es decir ni 1% de los datos afectados.

También se buscaron registros duplicados, los cuales se eliminaron para evitar caer en redundancias.

Se aplicó un orden al dataset según criterio de las columnas: País, año y cuatrimestre.

## Análisis y visualización

Se realiza un análisis comparativo de la utilidad bruta anual de cada país mediante la utilización de tres **gráficos de barras**. Allí es posible identificar cómo se comporta la tendencia en cada país a grandes rasgos.

Por otra parte, se incluye un **diagrama de caja y bigotes** para conocer la dispersión de la utilidad bruta por país. Así también la concentración, puntos atípicos y como éstos se alejan de la media, todo esto discriminado por países para un mejor control.

También es posible visualizar mediante un **gráfico de torta** el porcentaje de utilidad bruta de cada tipo de producto. Se identifica que el tipo de producto con mayor utilidad bruta es el calzado con un porcentaje de 19,5%, seguido de lentes con el 18,1%. Y el tipo de producto con menor utilidad bruta, con un 12,4% son las joyas.

---

<sup>1</sup> Cabe aclarar que si bien accedimos a un dataset en crudo, realizamos una instancia de 'ensuciar' los datos a propósito para llevar a cabo más tareas de limpieza poniendo en ejecución diferentes métodos de Python.

---

## CONCLUSIONES

---

Para finalizar este proyecto buscamos identificar los procesos aprendidos en el bloque de Analista de Datos. En primera instancia realizamos una recopilación de datos mediante un dataset propuesto por una empresa. Aplicamos métodos para limpiar y ordenar dicha información, como paso crucial para obtener resultados de mayor calidad. Exploramos y describimos la información contenida para, posteriormente, poder llevar a cabo un análisis de las variables que se presentan. Esto permite identificar patrones, tendencias y relaciones en la información que lleva a tomar decisiones informadas respecto a tal organización.

---

## BIBLIOGRAFÍA

---

Material de estudio de la Tecnicatura Superior en Ciencia de Datos e Inteligencia Artificial ISPC. Módulo Analista de Datos. Cohorte 2023.

Triolla, M. (2004). Estadística. Ed. Pearson Educación. México.

<https://datascienceparichay.com/python-for-data-science/data-structures-python/>

<https://python-graph-gallery.com/>

<https://www.w3schools.com/python>