ABP – PROYECTO E INFORME:

TECNICATURA SUPERIOR EN CIENCIAS DE DATOS E INTELIGENCIA ARTIFICIAL

ESPACIO CURRICULAR: ANALISTA DE DATOS I

PROF. UGARTE MARCOS // PERATTA NAHUEL

AMERICAN THINGS – ANALISIS Y PROCESAMIENTO DE DATOS, PARA SELECCIÓN DE ESTRATEGIAS DE VENTAS.

Introducción y Desarrollo:

Grupo N° 6

AÑO 2025

Integrantes:

Guaraz Emanuel - https://github.com/JEmanuelG

Direni Carlos - https://github.com/Cdireni1

Ryser Lucas - https://github.com/lucasryser6

Lobo Bruno - https://github.com/lobosouza

Allende Olmedo Nicolás - https://github.com/AllendeNicolas

REPOSITORIO: https://github.com/JEmanuelG/Analista-de-datos-I

Lugar de presentación: ISPC (Instituto Superior Politécnico de Córdoba)

INTRODUCCIÓN:

1. **NOMBRE DEL PROYECTO**:

American Things – análisis y procesamiento de datos, para selección de estrategias de ventas.

Repositorio del Proyecto: https://github.com/AllendeNicolas/Analista-de-datos-I/tree/main

https://github.com/JEmanuelG/Analista-de-datos-I

2. **TIPO DE PROYECTO**:

El proyecto tiene un enfoque ***De Investigación y tecnológico***, donde se propone el análisis y procesamiento de datos, relacionados a las ventas más actuales de la empresa, obtenidos en un período de tiempo, con el fin de desarrollar y aplicar, tecnologías y métodos destinados a poner en conocimiento al establecimiento, de las categorías de productos, que han tenido un alza o una baja en ventas, y a partir de ello, abordar las estrategias necesarias para mejorar los diferentes sectores y generar un mejor rendimiento.

Estratégico y de Marketing, ya que apuntamos a que la propuesta del proyecto sea, el puntapié inicial, de una serie de toma de decisiones de marketing y publicidad, que apunten a generar, una mejora en las ventas de cada una de sus categorías, y reforzar aquellas, cuyas ventas son prometedoras.

De Investigación y Procesamiento, mediante el análisis a aplicación de software, utilizados en la materia Procesamiento de datos, y las técnicas en exploración y estadísticas, conocer y explotar de manera eficiente, todos los datos brindados, realizando una limpieza y puesta a punto del "Dataset", y utilizar cada uno de esos aportes, para seleccionar las mejores estrategias de venta, y determinar las decisiones más acertadas, relacionadas con el marketing de cada categoría de dicho establecimiento.

3. **ESPACIO CURRICULAR**:

Analista de Datos I – Profesor/es: Marcos Ugarte y Nahuel Pratta.

4**. EJES TEMÁTICOS/RED DE CONCEPTOS**:

En el contexto del módulo de Análisis de Datos, el presente proyecto responde a la necesidad de la empresa American Things de llevar a cabo análisis estratégicos eficaces. Este trabajo se centra en las etapas del procesamiento de datos, con el propósito de asegurar la calidad de los datos a analizar, así como en las técnicas de exploración de datos y en los métodos estadísticos empleados para obtener resultados precisos y confiables. El proyecto integra los contenidos de las materias Procesamiento de Datos y Estadística y Exploración de Datos I bajo el siguiente enfoque metodológico:

1. Identificación y Preparación del Conjunto de Datos:

- Aplicación de técnicas de limpieza
- Normalización de los datos
- Transformación en dataset utilizando herramientas como Python (pandas) y/o SQL, para garantizar la calidad y coherencia de los datos.

2. Definición de Variables y Exploración:

- Clasificación de variables
- Organización de los datos
- Presentación de los datos en tablas y gráficos
- Cálculo de medidas numéricas

3. Detección de Tendencias y Segmentación Estratégica

- Interpretación de la información obtenida.

5**. PROBLEMÁTICAS/NECESIDADES**:

American Things, un poli rubro, dedicado a la ventas de de productos en general, se ha visto en la necesidad de organizar la información alojada en sus base de datos, respecto a las ventas realizadas por esta, en un determinado periodo de tiempo. Su Data Set, presenta anomalías respecto de los datos que han sido cargados en ella, ya sea por un error externo, falta de previsión al ingresar datos a la base por sujetos no pertenecientes a la empresa, o internos, por no tomar los cuidados y las diligencias necesarias al momento de trabajar con la información y los datos proporcionados en la carga del Dataset, por personas encargadas pertenecientes al mismo establecimiento. La información incompleta o mal dispuesta de la base de datos, no permite a American Things, realizar análisis estratégicos eficientes, para tomar decisiones

efectivas, sobre las distintas categorías de productos que ofrecen, y no les permite conocer con precisión, cuales son los sectores que deben atender y mejorar, para lograr los resultados de ventas deseados, de manera general y específica en la empresa.

6. **FUNDAMENTACIÓN**:

La elección del problema, al cual está orientado el proyecto, se centra en la necesidad persistente de una constante actualización, mantenimiento y análisis, de las diversas bases de datos que una empresa pueda tener.

En un entorno empresarial altamente competitivo, el manejo eficiente de los datos es un pilar fundamental para la toma de decisiones estratégicas.

Impacto de una base de datos inconsistente:

Las inconsistencias en la base de datos generan problemas que repercuten en varios aspectos clave:

- **Dificultad en la segmentación de clientes:** Sin datos precisos, las empresas no puede identificar patrones de compra ni definir estrategias de marketing efectivas.
- **Imposibilidad de análisis de ventas preciso:** La empresa no puede determinar con exactitud qué productos tienen mayor demanda en determinados períodos.
- **Pérdida de oportunidades de optimización:** La toma de decisiones basada en datos defectuosos puede llevar a inversiones erróneas o falta de atención a segmentos estratégicos.
- **Riesgos operativos:** Información incorrecta puede generar problemas logísticos, afectando la gestión de inventario y el abastecimiento oportuno de productos.

En Conclusión, el manejo eficiente de la base de datos es un factor determinante para la competitividad de American Things en el mercado. La depuración y control de la información permiten generar estrategias basadas en datos confiables, optimizar la gestión de ventas y potenciar la toma de decisiones estratégicas. Implementando medidas correctivas y preventivas, la empresa puede garantizar un sistema de información robusto, con datos precisos y útiles para su crecimiento sostenido.

7. **VISIÓN DEL PROYECTO**:

OBJETIVO GENERAL:

- Ordenamiento, Limpieza, Exploración y Análisis de datos, para mitigar problemas y asegurar la fiabilidad del Dataset, para implementación de medidas estratégicas.

OBJETIVOS ESPECÍFICOS:

1) Analizar costos fijos y gastos variables. (Aquellos que puedan surgir por la ejecución de la iniciativa de innovación y gestión de organización de la base de datos).

2) Seleccionar los medios necesarios, software y hardware, de aplicación. (Aquellos medios que nos permitan realizar la limpieza y organización del Dataset, sin dañar y cuidado los datos existentes).

3) Desarrollar un entorno seguro de ejecución. (Que se tomen todas las precauciones y medidas de seguridad, para evitar abusos, estafas, y manipulación ilegal de los datos).

4) Validación y depuración periódica del Data Set. (Mediante técnicas de limpieza de datos para corregir registros incorrectos o duplicados).

5) Protocolos estrictos de ingreso de información. (Definir procesos estandarizados que eviten errores humanos en la carga de datos).

6) Capacitación del personal**.** (Asegurar que quienes manipulan la base de datos comprendan la importancia de la calidad de la información y sigan buenas prácticas de gestión).

7) Control de acceso y permisos. (Limitar el ingreso de información solo a personal autorizado para prevenir manipulaciones externas no controladas).

8) Implementación de herramientas de gestión automatizadas. (Utilizar software avanzado para verificar la integridad y coherencia de los datos, evitando errores manuales).

METAS:

1) Implementar el Dataset, actualizado, limpio y ordenado, para una explotación de datos más eficiente, y así aplicar una mejor calidad de información, a las nuevas decisiones estratégicas de mercado. (Utilización del Dataset. Aplicando los nuevos resultados del análisis de la base de datos, más fiables y óptimos).

8. **SELECCIÓN DE ACCIONES:**

Objetivo general	***Acciones***	***Habilidades o capacidades a***

		lograr
Ordenamiento, Limpieza, Exploración y Análisis de datos.	Mitigar problemas y asegurar la fiabilidad del Dataset, para implementar las medidas estratégicas pertinentes.	Gestionar y aplicar, medidas y técnicas, tendientes a la limpieza y conservación de los datos, y asegurar la fiabilidad del Dataset.
Objetivos específicos	***Acciones***	***Habilidades o capacidades a lograr***
Analizar costos fijos y gastos variables.	Desglosar e identificar, de manera específica, cada gasto y costo, que pueda suscitarse, en la modificación y actualización del Dataset, por el uso de Software y tecnologías destinadas a tal fin.	Determinar con exactitud, cuales son aquellos gastos y costos a considerar, para el desarrollo, y gestión de la Base de datos, y posteriormente, de su aplicación.
Seleccionar los medios necesarios, software y hardware de aplicación.	Conocer y aplicar, aquellos medios que nos permitan realizar la limpieza y organización del Dataset, sin dañar y cuidando los datos existentes.	Elegir los métodos, técnicas, software y demás tecnologías, que brinden una manipulación de los datos, de una manera, fluida y sencilla
Desarrollar un entorno seguro de ejecución.	Tomar todas las precauciones y medidas de seguridad, para evitar abusos, estafas, y manipulación ilegal de los datos	Determinar y obtener las autorizaciones necesarias, para la manipulación del Dataset, cuidando de proteger, todos aquellos datos que sean sensibles, o que requieran una atención especial de protección.
Validación y depuración periódica del Data Set.	Seleccionar técnicas de limpieza de datos para corregir registros incorrectos o duplicados.	Asegurar la limpieza del Dataset, rellenar datos faltantes, eliminar datos duplicados, recuperar datos que se hayan perdido y modificar aquellos mal ingresados, entre otras acciones.
Protocolos estrictos de ingreso de información.	Definir procesos estandarizados que eviten errores humanos en la carga de datos.	Crear un protocolo, que asegure la carga de datos de manera eficiente, evitando, en lo posible, la mayor cantidad de errores posibles.

Capacitación del personal**.**	Asegurar que quienes manipulan la base de datos, comprendan la importancia de la calidad de la información y sigan buenas prácticas de gestión.	Brindar la información necesaria, al personal autorizado, sobre los protocolos de ingresos de datos.
Control de acceso y permisos.	Limitar el ingreso de información solo a personal autorizado para prevenir manipulaciones externas no controladas.	Seleccionar al personal, encargado de realizar, la carga y manipulación de los datos, de cada Dataset de la empresa.
Implementación de herramientas de gestión automatizadas.	Utilizar software avanzado para verificar la integridad y coherencia de los datos, evitando errores manuales.	Aplicar software, que sirva de apoyo y guía, a aquellos que estén autorizados a gestionar la Base Datos, para una eficaz manipulación de esta.

9. **CRONOGRAMA:**

CRONOGRAMA	***MES 0*** ***Semana*** ***01-02***	***MES 0*** ***Semana*** ***03-04***	***MES 1*** ***Semana*** ***01-02***	***MES 1*** ***Semana*** ***03-04***
Ordenamiento, Limpieza, Exploración y Análisis de datos	Diagnóstico del Dataset actual Selección y limpieza inicial de datos	Normalización y transformación de datos Exploración y primeros análisis	Análisis de tendencias Generación de reportes	Revisión final de análisis Preparación de informe de estrategias

Selección de herramientas	Relevamiento de herramientas necesarias Evaluación de alternativas	Adquisición e instalación de software y hardware	Capacitación en el uso de herramientas seleccionadas	Implementación de control y gestión
Desarrollo de un entorno seguro de ejecución	Diagnóstico de seguridad actual. Recomendaciones iniciales.	Diseño de protocolos de acceso y permisos	Capacitación en protocolos de seguridad	Aplicación y auditoría de medidas de seguridad
Validación y depuración periódica del Dataset	Primera depuración de datos (detección de errores)	Segunda depuración y validación cruzada	Corrección de inconsistencias finales	Validación definitiva del Dataset
Protocolos estrictos de ingreso de información	Diseño preliminar de protocolos	Revisión y ajuste de protocolos con feedback interno	Capacitación del personal	Implementación formal de protocolos
Capacitación del personal	Planificación de capacitaciones	Ejecución de capacitaciones técnicas	Ejecución de capacitaciones sobre buenas prácticas	Seguimiento y evolución del personal capacitado
Control de acceso y permisos	Identificación de roles críticos	Definición de niveles de acceso	Asignación de permisos	Revisión de acceso y ajustes

Implementación de herramientas de gestión automatizadas	Selección de herramientas de gestión	Implementación inicial	Integración con base de datos	Optimización y ajustes finales

Presupuesto con gastos fijos y variables.

Al mes 5 de 2025 - tipo de cambio $1200 ARS por dólar oficial.

Son los gastos que se van a mantener durante el proceso que dure el proyecto o que se contratarán por única vez.

COSTOS FIJOS

Concepto	Detalle	Estimado USD	Estimado $ (pesos argentinos)
Comunicación Interna	Plataforma de comunicación Suscripción anual a herramienta tipo Slack o Microsoft Teams (10 usuarios)	100	120.000
Software de análisis de datos	Python Python (con librerías Pandas, NumPy, Matplotlib) y Excel, Google Sheet.	15.00 Excel / Python Licencia Libre	18.000
Servicios en la nube para almacenamiento	Google o Dropbox (2 TB)	9.99/mes	12.000
Hardware	Notebook de gama media (i5, 16GB RAM, SSD)	1.200	1.440.000
Capacitación de personal	curso básico de limpieza de datos	50	60.000 x curso
Software de gestión y visualización de datos	Licencia Power BI Pro	120/año	144.000

GASTOS VARIABLES

Concepto	Detalle	Estimado USD	Estimado $
Hs. consultoría técnica	Consultor de datos freelance (20 hs)	200.00	240.000
Servicios de seguridad informática	Instalación	50.00	60.000
Revisión y auditoría externa	Validación de la limpieza final del dataset.	100.00	120.000
Eventos de capacitación interna	Reuniones, charlas de protocolos de carga de datos.	30.00 (catering)	36.000
Publicidad	Diseño Gráfico, Flyers Digitales, Campañas de email	500	600.000

10. **PRODUCTO FINAL:**

Se procede a elaborar un software de análisis y procesamiento de datos, relacionado con las ventas de una empresa, recibiendo como entrada un dataset con información bruta, la cual sería procesada mediante una limpieza y normalización de los datos, utilizando fundamentalmente la librería Pandas del lenguaje Python, entre otras como NumPy y Matplotlib, dando como resultado información sobre:

- Altas y bajas en las ventas de determinados rubros o categorías en distintos rangos de fechas.

- Promedios de valoraciones de productos y categorías.

- Un análisis de las reseñas de a acuerdo a las ventas por categorías, para fortalecer estrategias de publicidad, para aquellas que no cumplan con las expectativas del cliente.

- Un análisis entre costos y precios de venta de lo productos, para desarrollar estrategias y promociones, para generar un alza en las ventas de aquellas categorías menos redituables, o cuyo grado de utilidad, no es beneficioso para la empresa.

- Una Guía de análisis que considere las ventas por ciudad donde se encuentre cada sucursal, para generar informes de ventas, sobre las categorías más vendidas, el tipo de cliente que prolifera en esa zona, y si se trata de clientes Normales o Registrados. Para

desarrollar en cada sucursal, estrategias de retención y atracción de clientes, para atenciones personalizadas de acuerdo a sus consumos habituales.

Todos los resultados solicitados por el propietario de la aplicación, serán mostrados en pantallas, con visualizaciones interactivas. Cada consulta será analizada con un gestor de IA, que brinda información en tiempo real, hará sugerencias, y dará avisos de alertas cuando algún dato dentro del Data Set, haya sido ingresado de manera inadecuada, o exista algún error en el mismo. También dará alertas cuando la actualización de los datos de sector de ventas, presente alguna caída en alguna de sus sucursales, y dará avisos sobre qué categorías no están cumpliendo con las expectativas deseadas.

El proyecto se lleva a cabo gracias a la participación y acompañamiento del espacio curricular "Analista de Datos", donde adquirimos los conocimientos y herramientas para poder llevar a cabo dicho proyecto.

11. **BIBLIOGRAFÍA:**

Triola, M. F. (2009). *Estadística*. Pearson Educación

Mckinney, W. (2017). *Python Para Análisis De Datos*. Anaya Multimedia.

12. SEGUNDO CUATRIMESTRE:

Dealdata, es el resultado de la unificación de tres grupos diferentes, respecto de los equipos formados en el primer cuatrimestre. Por lo cual decidimos trabajar y afianzarnos como equipo, utilizando una Base de Datos diferente, ya que nuestros trabajos anteriores eran muy diferentes, y por cuestiones de incompatibilidad, no podíamos continuar con alguno de los Data Set anteriores. Por esto decidimos trabajar con un **Data Set de Ventas de Amazon**, el cual obtuvimos de la página de Kaggel (https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset), el cual contaba con los requisitos específicos de poder visualizar en el mismo, variables preferentemente cuantitativas para de esta manera poder realizar los análisis estadísticos pertinentes.

Nuestro primer paso fue la **normalización del data set**, eliminando datos erróneos o contradictorios, y reduciendo los Outliers, de la mejor manera posible. En el Colab de Google, se realizó un análisis de datos, aplicando diferentes métodos y técnicas centradas en comprender ciertos aspectos de los productos de Amazon.

En segundo lugar, se exploró la relación entre la cantidad de calificaciones recibidas por un producto y su calificación promedio. Se utilizaron diferentes **métodos de correlación** para ver si los productos más populares tendían a tener mejores calificaciones.

Luego, se aplicó una técnica llamada **ANOVA** para comparar las calificaciones promedio de los productos entre diferentes categorías. Esto ayudó a identificar si el tipo de producto influye en qué tan bien es calificado por los usuarios.

Finalmente, se desarrolló un modelo de **regresión logística multivariable**. El objetivo de este modelo fue predecir la probabilidad de que un producto sea considerado “altamente calificado”, basándose en características como su precio, el descuento ofrecido y la cantidad de calificaciones. Se evaluó qué tan bien el modelo podía hacer estas predicciones. Puede analizar, visualizar el desarrollo y las conclusiones obtenidas del trabajo, en el Repositorio del Grupo, y en la carpeta ABP del mismo.

ANALISTA DE DATOS I

INFORME PROYECTO ABP

AMERICAN THINGS: MARKETING Y ESTRATEGIAS DE VENTAS

Prof. Ugarte marcos /
Peratta Nahuel

Año 2025

Inst. Superior Politécnico de Córdoba

Índice:

1) Introducción:

Somos Data Humans, una empresa dedicada al desarrollo de software y a la gestión de bases datos, radicada en la Ciudad de Córdoba capital, comprometida con brindar a las empresas y la comunidad en general, soluciones tecnológicas para una eficiente utilización de los recursos informáticos y una mejor calidad de vida.

En esta oportunidad, realizamos la presentación del proyecto final del espacio curricular **Analista de Datos I**, mediante este informe, donde se detallan los pasos seguidos para la consecución de los objetivos tanto generales, como específicos y metas, planteadas en las acciones llevadas a cabo en el diagrama del proyecto.

Atento al problema y la necesidad planteados por la empresa **American Things**, un poli rubro, dedicado a la ventas de de productos en general, se ha visto en la necesidad de organizar la información alojada en sus base de datos, respecto a las ventas realizadas por esta, en un determinado periodo de tiempo. Su Data Set, presenta anomalías respecto de los datos que han sido cargados en ella, ya sea por un error externo, falta de previsión al ingresar datos a la base por sujetos no pertenecientes a la empresa, o internos, por no tomar los cuidados y las diligencias necesarias al momento de trabajar con la información y los datos proporcionados en la carga del Dataset, por personas encargadas pertenecientes al mismo establecimiento; el equipo se propuso analizar con eficiencia, el Data Set entregado por la empresa, y proceder a la depuración y optimización del mismo, con el fin de otorgarle al cliente, una base sólida para la toma de decisiones y estrategias de Marketing.

Como consecuencia, el equipo ofrece llevar adelante las acciones planteadas en el desarrollo del proyecto, mediante la utilización de software Python como lenguaje principal, y un entorno de programación denominado Júpiter (mediante un Notebook), junto a otras librerías como Pandas, Numpy y Matplotlib, con el fin de realizar la limpieza y organización de los datos, que se muestren como **ausentes, en formato incorrecto, erróneos o duplicados.**

Entre otras acciones a llevar a cabo, con el fin de proteger la integridad del DataSet, y la base de datos de nuestro cliente, se propone desarrollar un entorno seguro de ejecución, con el fin de cuidar y proteger la información sensible que puedan estar presentes en el mismo. También se propone el desarrollo de un protocolo de depuración periódica, y un protocolo estricto de ingreso de datos e información, para la resolución de contingencias, y un mejor aprovechamiento de la información obtenida. Y en última instancia, proponer la selección y capacitación estricta, del personal de la empresa, encargada del ingreso de datos, desde el ámbito interno, propio de la empresa.

A continuación, procederemos a mostrar los resultados obtenidos del análisis y depuración del Dataset, como así también los protocolos, y estrategias de ventas, que puede asumir la empresa para el desarrollo de sus campañas de Marketing y Publicidad.

2) Selección de herramientas – software:

Para la ejecución del proyecto, y proceder con el análisis y limpieza del Data Set brindado por la empresa, el equipo decidió utilizar las siguientes herramientas de software, considerando aspectos de las mismas como, **disponibilidad** (programas que están al alcance de cualquier usuario, por ser de código abierto), **completitud** (por ser entornos que ofrecen varias herramientas de análisis y visualizaciones, en un mismo paquete) y por su grado de **economicidad** (por tratarse de software de código abierto, y de acceso gratuito, procurando un presupuesto más económico para la empresa).

Las herramientas seleccionadas son:

Entorno **Jupyter:** es un proyecto de código abierto que proporciona herramientas y servicios para la computación interactiva, especialmente para la ciencia de datos y la visualización de datos.

Jupyter – Notebook: es una aplicación web que permite crear y compartir documentos interactivos con código, texto, visualizaciones y otros resultados. Ideal para trabajar con lenguaje de programación Python.

Lenguaje de programación **Python:** Entorno de programación de código abierto, caracterizado por su versatilidad, y utilizado en análisis y ciencias de datos, y aprendizaje automático.

Librería Numpy: perteneciente a Python, biblioteca fundamental en la computación científica y datos. Ofrece funciones matemáticas de alto nivel y operaciones con matrices.

Librería Matplotlib: biblioteca de visualización de datos en Python que se utiliza para crear gráficos, histogramas, diagramas de barras, gráficos de dispersión, entre otros. Ofrece una gran flexibilidad y control sobre las visualizaciones.

Librería Pandas: Permite trabajar con datos tabulares (en filas y columnas) y ofrece herramientas para realizar tareas como la limpieza, transformación, exploración y análisis de datos.

Con las herramientas antes mencionadas, se llevara a cabo en análisis integral del Data Set, y sus resultados, nos servirán para brindar una respuesta más clara y eficiente a las necesidades, elaboración de estrategias, y mejor entorno para la toma de decisiones por parte de la empresa.

3) Análisis del Data Set:

a) Descripción y características:

El Data Set otorgado por la empresa, presenta anomalías en los datos cargados, las cuales podrían haber sido causadas por:

- **Errores externos**, como la manipulación de la base por parte de personas ajenas a la empresa.
- **Falta de previsión al ingresar información** sin validaciones adecuadas.
- **Errores internos**, producto de una manipulación descuidada por parte de personal perteneciente a la organización.

La base **contiene datos incompletos y/o maldispuestos**, lo que impide que American Things realice análisis estratégicos eficaces. Como consecuencia, no pueden tomar decisiones efectivas sobre:

- Las diferentes categorías de productos que ofrecen.
- La identificación de sectores críticos a mejorar.
- La optimización de sus estrategias de ventas, tanto a nivel general como específico.

Descripción de las Columnas, ordenamiento y organización de los datos:

A continuación presentamos un cuadro, en el cual se detallan la descripción de cada dato que almacena cada columna, su carácter, descripción y, los posibles valores que pude contener cada una de ellas.

Columna	**Tipo de dato**	**Descripción**	**Posibles valores**
Factura	Texto(str)	Código único de la factura de la compra.	Formato: XXX-XX-XXXX
Sucursal	Texto(str)	Sucursal donde se realizó la compra.	• "A" • "B" • "C"
Ciudad	Texto(str)	Ciudad en la que se encuentra la sucursal.	• "Nueva York" • “Houston” • "Chicago"
Tipo	Texto(str)	Tipo de cliente: si es miembro o no.	• "Member" • "Normal"

Genero	Texto(str)	Género del cliente.	• "Male" • "Female"
Categoria	Texto(str)	Categoría del producto o servicio adquirido.	• "Health and beauty" • "Electronics accesories" • Sports and travel • Fashion accesories • Food and beverages • Home and lifestyle
Costo	Numérico (float)	Costo de los productos vendidos (Cost of Goods Sold).	Valor numérico decimal
Ventas	Numérico (float)	Monto total de la venta.	Valor numérico decimal
Fecha	Fecha (str/date)	Fecha en la que se realizó la transacción.	Formato: MM/DD/AAAA

b) Análisis en entorno Júpiter – Notebook, descripción:

Se realizó el análisis y limpieza de los datos en el entorno Jupyter – Notebook, utilizando la librería Pandas. Se procedió a analizar cada una de las columnas tanto **numéricas** (la cantidad de valores válidos sobre el total, media y mediana, desviación estándar, valores ausentes y mal registrados o erróneos), como las columnas **categóricas** (la cantidad de registros para cada categoría, registros válidos sobre el total, ausentes y mal registrados), y luego de acondicionar, y depurar el Data Set, obtuvimos los siguientes resultados:

Columnas Numéricas:

El Data Set analizado, contiene las siguientes columnas con valores numéricos, “Costo”, “Ventas” y “Resnia”, el análisis muestra además, valores válidos sobre el total (en principio el Data set contenía 400 registros, ahora sólo muestra 395 válidos), la media, mediana, su desviación estándar, la cantidad de valores ausentes, y la cantidad de valores ausentes y erróneos.

Se muestra una captura de los resultados de la limpieza del data set, realizada en código.

- Columna “Costo”:

```
Análisis de la columna numérica: Costo
Valores válidos sobre el total: 395 / 395
Media: 315.64
Mediana: 280.62
Desviación estándar: 224.84
Valores ausentes: 0
Valores mal registrados o erróneos: 0
```

- Columna “Ventas”:

```
Análisis de la columna numérica: Ventas
Valores válidos sobre el total: 395 / 395
Media: 327.8
Mediana: 284.19
Desviación estándar: 233.88
Valores ausentes: 0
Valores mal registrados o erróneos: 0
```

- Columna “Resenia”:

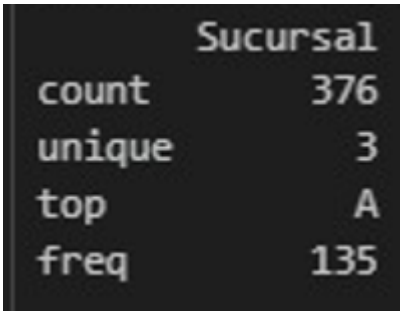
```
Análisis de la columna numérica: Resenia
Valores válidos sobre el total: 395 / 395
Media: 7.01
Mediana: 7.0
Desviación estándar: 1.71
Valores ausentes: 0
Valores mal registrados o erróneos: 0
```

Columnas Categóricas:

El Data Set analizado, contiene las siguientes columnas con los registros categóricos, “Sucursal”, “Ciudad”, “Tipo”, “Genero”, “Categoría” y “Método de Pago”. El análisis muestra además, la cantidad de registros válidos de cada columna, su frecuencia y el registro Top (el que más veces se muestra).

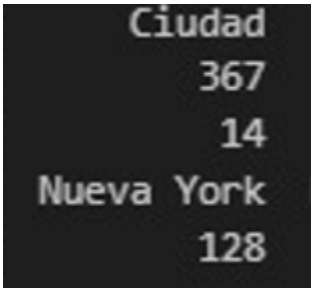
Se muestra una captura de los resultados de la limpieza del data set, realizada en código.

- Columna “Sucursal”:



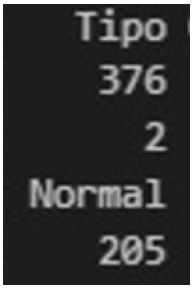
	Sucursal
count	376
unique	3
top	A
freq	135

- Columna “Ciudad”:



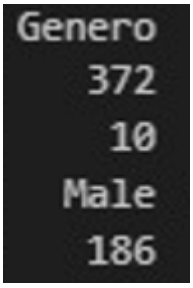
Ciudad	
	367
	14
Nueva York	
	128

- Columna “Tipo”:

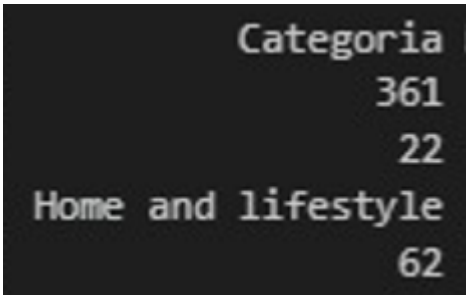


Tipo	
	376
	2
Normal	
	205

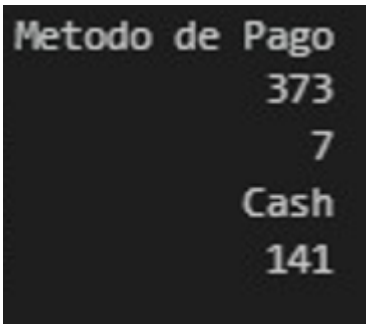
- Columna “Genero”:



- Columna “Categoría”:



- Columna “Método de Pago”:



Con la limpieza y organización de los datos, y con el análisis de cada una de sus columnas, nos encontramos en condiciones, de formular las primeras aproximaciones y deducciones, para elaborar las estrategias pertinentes, en cuanto a ventas, análisis de costos, y reseñas obtenidas para cada sucursal.

NOTA: Este análisis corresponde a la **Evidencia 2** – de la asignatura Analista de datos 1.

4) Capacitación de personal – Ingreso de datos -Automatización:

a) Protocolo de educación y entrenamiento del personal autorizado:

1. Principios Generales de Manipulación:

- **Precisión:** Los datos deben ingresarse exactamente como se reciben, sin interpretaciones personales.
- **Consistencia:** Mantener una estructura homogénea en formatos y convenciones (uso de mayúsculas, fechas, números decimales, etc.).
- **Integridad:** Completar todos los campos obligatorios y evitar valores nulos o incorrectos. Procurar para ello, un entrenamiento previo, y exhaustivo, para evitar errores.
- **Seguridad:** Acceder y manipular datos solo con las credenciales adecuadas y siguiendo políticas de protección. Sólo el personal autorizado, previa capacitación, podrá tener acceso al Data set, y podrá manipular los datos en el contenidos.

2. Estándares de Formato de Datos

- **Fechas:** Formato único (ej. `YYYY-MM-DD` o `DD/MM/YYYY` según estándares internos).
- **Números:** Uso de separadores decimales y miles estandarizados (`.` o `,` según contexto).
- **Texto:** Uso de nombres y términos oficiales, sin abreviaturas no autorizadas.
- **Categorías:** Aplicación correcta de clasificaciones y códigos predefinidos.
- **Id (identicadores únicos):** Uso de números en serie, para la identificación de categorías, transacciones, clientes, etc. Que sean únicos, para la individualización de los objetos a los que se refieran.

3. Procedimientos para la Carga de Datos

- **Validación previa:** Antes de ingresar información, verificar su fuente y formato.
- **Uso de plantillas:** Implementar hojas de carga predefinidas con validaciones automáticas.
- **Verificación cruzada:** Un segundo colaborador revisa la entrada de datos antes de su confirmación.
- **Auditoría periódica:** Revisar datasets con herramientas de análisis para detectar inconsistencias.

4. Capacitación y Responsabilidad

- **Sesiones de formación:** Entrenamientos regulares sobre protocolos de ingreso.
- **Documentación accesible:** Manual de referencia con ejemplos y mejores prácticas.
- **Supervisión y feedback:** Asignación de responsables para evaluar el cumplimiento de normas.
- **Registro de errores:** Sistema para documentar fallas y establecer mejoras continuas.

5. Prevención y Manejo de Errores

- **Identificación temprana:** Implementación de validaciones automáticas en los formularios de ingreso.
- **Corrección guiada:** Procedimientos claros para la rectificación de errores sin afectar integridad del dataset.
- **Control de versiones:** Registro de cambios y métodos de recuperación de datos previos.

b) Protocolo estricto de ingreso de datos y automatización:

1. Validación y Pre procesamiento de Datos

- **Reglas en Bases de Datos SQL:** Implementar restricciones CHECK, NOT NULL, DEFAULT y claves foráneas para asegurar integridad en la información.
- **Expresiones Regulares (Regex):** Utilizar scripts en Python, para validar formato de entradas (correos electrónicos, fechas o números).
- **Scripts de Pre procesamiento:** Automatizar limpieza de datos con Pandas en Python, asegurando que los registros cumplan con los estándares antes de ingresar a la base.

2. Interfaces y Formularios Inteligentes

- **Google Forms con Validaciones:** Configurar formularios con campos obligatorios y restricciones de datos para estandarizar la entrada.
- **Power Apps:** Diseñar interfaces personalizadas con validaciones embebidas para reducir la posibilidad de errores.
- **Tailwind y React en Frontend:** Crear componentes con restricciones y retroalimentación visual en tiempo real para guiar al usuario.

3. Automatización en la Carga de Datos

- **ETL (Extract, Transform, Load):** Implementar procesos con Apache NiFi, Talend o Power Automate para transformar datos antes de integrarlos en el dataset.
- **Batch Processing con Python:** Desarrollar scripts que validen y suban datos automáticamente en lotes, reduciendo intervención manual.
- **Webhooks:** Conectar plataformas de ingreso con la base de datos, para actualizaciones en tiempo real.

4. Monitoreo y Control de Errores

- **Dashboards en Power BI:** Analizar patrones de carga de datos y detectar anomalías en el ingreso.
- **Alertas con Power Automate:** Notificar a los responsables cuando se detectan inconsistencias o valores fuera de rango.

- **Logging y Auditoría:** Registrar cambios en bases de datos con triggers en SQL o logs detallados en aplicaciones backend.

5. Capacitación y Mejora Continua

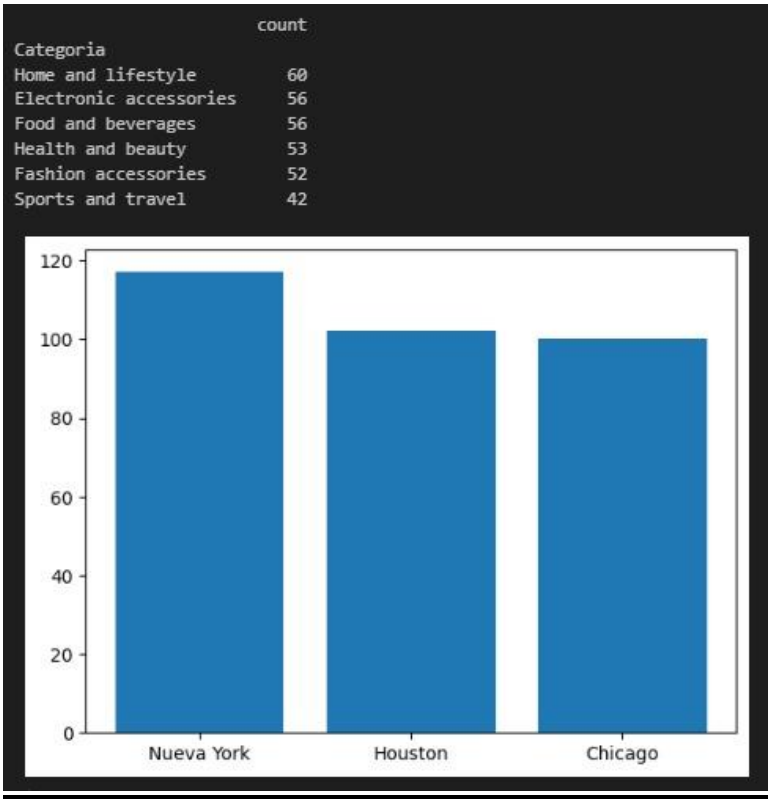
- **Simuladores de Carga:** Crear entornos de prueba donde los empleados puedan practicar sin afectar el dataset real.
- **Sistema de Feedback Automatizado:** Generar reportes de errores frecuentes y sugerir mejoras a los operadores.
- **Documentación Dinámica con Notion o Confluence:** Mantener un manual de procedimientos actualizado con ejemplos interactivos.

5) Evidencia 3- cálculos - gráficos y fundamentaciones

Gráficos en relación a los datos:

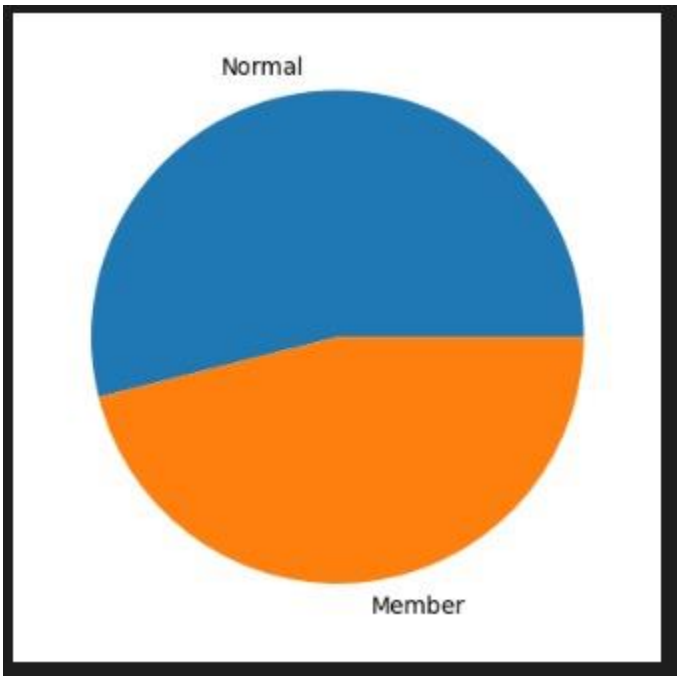
Ventas por ciudad:

Tiene el propósito de analizar las ventas realizadas por ciudad, donde se encuentra cada sucursal, para reforzar estrategias de marketing, en aquellas sucursales que menos ventas en general procesan.



Porcentaje de usuarios registrados y no registrados:

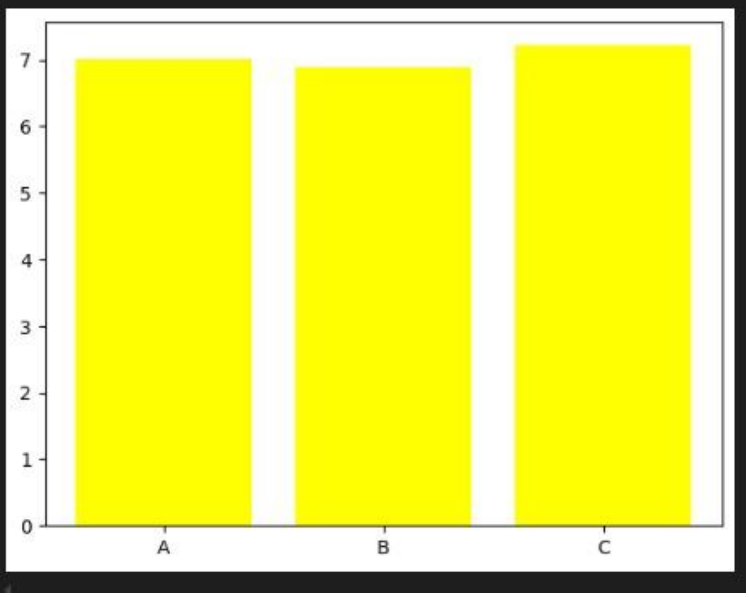
Tienen el propósito de analizar la cantidad de clientes que se registran en la base de datos y fomentar la registración de dichos clientes para obtener más metadatos y ofrecerles ofertas personalizadas.



Reseñas de sucursales:

Miden la popularidad de la sucursal, y la calidad de la atención propuesta al público.

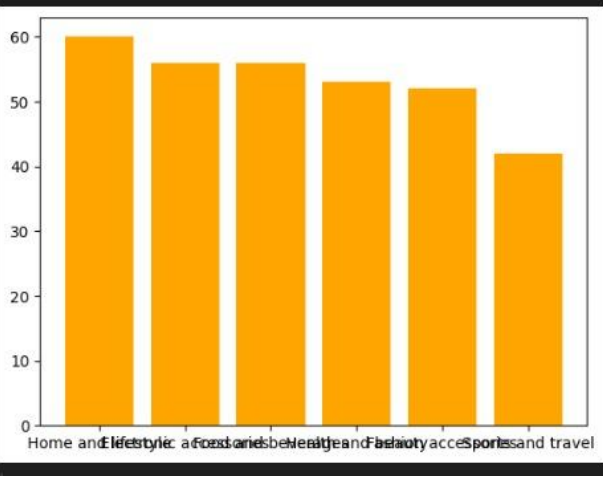
Sucursal	Resenia
A	7.002504
B	6.896929
C	7.211765



Ventas por categorías:

Refleja la cantidad de ventas por categorías producidas, analizar la categoría más vulnerable.

Categoria	count
Home and lifestyle	60
Electronic accessories	56
Food and beverages	56
Health and beauty	53
Fashion accessories	52
Sports and travel	42



Desarrollo:

Se procede a realizar un análisis estadístico descriptivo, en el cual se tiene en cuenta los datos según sean Simétricos o Asimétricos. Se pretende analizar con precisión, las Columnas Numéricas, de “Costo”, “Ventas” y “Resenia”, y su relación con las posibles estrategias de mercado que podríamos llegar elaborar.

Análisis basado en estadísticas descriptivas:

```
=== ESTADÍSTICAS DESCRIPTIVAS ===

Variables numéricas:
            Costo      Ventas     Resenia
count  376.000000  376.000000  376.000000
mean   315.701303  325.858803    6.979978
std    225.989003  232.839401    1.707989
min      2.000000   13.170000    4.000000
25%    142.157500  146.062500    5.600000
50%    279.930000  277.465000    6.996456
75%    440.970000  472.820000    8.400000
max    993.000000  944.620000   10.000000

Variables categóricas:
       Sucursal      Ciudad    Tipo Genero          Categoria Metodo de Pago
count       376         367     376    372                 361            373
unique        3          14       2     10                  22              7
top           A  Nueva York  Normal   Male  Home and lifestyle           Cash
freq        135         128     205    186                  62            141

=== SIMETRÍA DE VARIABLES NUMÉRICAS ===
Costo:
  Asimetría: 0.92
  Distribución moderadamente asimétrica
Ventas:
  Asimetría: 0.83
  Distribución moderadamente asimétrica
Resenia:
  Asimetría: 0.05
  Distribución aproximadamente simétrica
```

Estadísticas Descriptivas:

1. COSTO (*Asimetría 0.92*)

- Distribución confirmada: Media (315.70) > Mediana (279.93) → Asimetría positiva.
- Rango amplio: *Mín 2 - Máx 993* (variabilidad extrema)
- Dispersión:
  - Desviación estándar alta (225.99)
  - IQR amplio (*298.81* = Q3-Q1) Rango donde se concentra la mitad central de los datos. Elimina valores extremos.
- Outliers potenciales:
  - Valor máximo (993) está muy por encima de Q3 (440.97) (Límite superior = 440.97 + 1.5 × 298.81 = 889.19)
  - Posibles productos premium o errores de registro.

2. VENTAS (*Asimetría 0.83*)

- Patrón similar a costos: Media (325.86) > Mediana (277.47)
- Relación costo-ventas:
  - Mediana ventas (277.47) < Mediana costo (279.93) → Posible margen negativo en productos típicos.
  - Máximo ventas (944.62) < Máximo costo (993) → Los Productos más caros podrían no ser los más vendidos.
- Variabilidad: Desviación estándar (232.84) comparable a costo

3. RESEÑA (*Asimetría 0.05*)

- Distribución cercana a óptima: Media (6.98) ≈ Mediana (7.00)
- Rango controlado: Mín 4 - Máx 10 (sin reseñas extremadamente bajas)
- Centro estable: 50% de datos entre 5.0 y 8.4 (IQR = 3.4)
- Interpretación: Experiencia de cliente consistente

1. **Consideraciones de rentabilidad:**

- La mediana de ventas es MENOR que la mediana de costos → Posibles pérdidas en productos típicos.
- Necesidad **urgente** de análisis de rentabilidad por producto/categoría.

2. **Segmentación de productos:**

    - Productos de alto rendimiento **(estrella o premium**): 25% de productos con ventas > 472.82.
    - Productos de bajo rendimiento: 25% con ventas < 146.06 (requieren intervención)
    - Anexo: estos productos deben ser tenidos en cuenta para:
        - Stock
        - Promociones
        - Espacio en tiendas físicas/online
    - Oportunidades: Identificar por qué estos productos son exitosos (ej: calidad, categoría, ubicación).

3. **Análisis de categorías (variables categóricas):**

    - Ciudad: Nueva York domina (135 registros) → Posible sesgo geográfico
    - Categoría: "Home and lifestyle" lidera (186) → Motor del negocio
    - Método pago: "Cash" predominante (141) → Cultura de efectivo

Riesgos o alertas:

- Desbalance geográfico: 135/376 registros son de NY (35.9%)
- Problema de género: 205/367 registros son Male (55.9% - posible sesgo)
- Falta de datos: Método pago tiene 361/376 registros (4% faltantes)

NOTA: Este análisis corresponde a la **Evidencia 3** – de la asignatura Analista de datos 1.

CONCLUSIONES:

De acuerdo a los datos analizados, y las estadísticas generadas, hemos llegado a la conclusión, de que el Data Set, se encuentra en ordenado, depurado y en condiciones de ser utilizado con los fines previstos en los objetivos. Generar estrategias de Marketing, para mejorar la llegada a clientes de aquellas zonas donde se encuentren las sucursales, y también, para desarrollar la aplicación propuesta en el **Producto Final** del Proyecto ABP.

El Data Set, acondicionado, es útil ahora, pera tomar de manera más eficiente, las decisiones pertinentes, analizar el comportamiento de las sucursales, y la venta de las categorías de productos, atraer y dirigir estrategias personalizadas a los clientes registrados, y gestionar nuevas maniobras de marketing, para atraer a más público a las diferentes sucursales, entre otras estrategias referentes a los costos y promociones de productos, teniendo en cuenta también el género de las personas que visitan el local.