Individual Delivery

26/12/2020 to 10/01/2021

—

María Cagigas
Data Science
The Bridge

Visión General

El proyecto se basa en el análisis de los Best Sellers de Amazon post-confinamiento, concretamente en la electrónica y en los videojuegos. También se incluye un análisis más exhaustivo del mercado de videojuegos en su venta física desde 1980 hasta 2016.

Objetivos

B

Especificaciones

Para poder sacar los objetivos y sacar el máximo partido a la entrega, se concreta todo lo necesario para hacerlo posible:

Software

Visual Studio Code

Power Point

Adobe Acrobat Reader DC

Hardware

i7 de 10th GEN con 16GB de RAM

Requisitos

Python 3.9.0

Librerías: Pandas, NumPy, Matplotlib, Seaborn, Math, Requests

Tipografías Power Point: Calibri (Cuerpo), Segoe UI Semilight, Segoe UI Black

Pasos

I. Investigación del contexto

La investigación previa se basó en leer artículos sobre los objetos más vendidos post-confinamiento en general y sobre la evolución del mercado de videojuegos en estos días.

Por ejemplo, para el primer caso se analizaron artículos sobre cómo es el consumidor después del confinamiento, a qué factores le da más importancia o cómo ha cambiado su estilo de vida.

Fuentes:

1. https://www.lavanguardia.com/economia/20200622/481892371839/consumo-proximidad-comprador-covid-coronavirus-espana-espana.html
2. https://pinkermoda.com/splio-consumidor-post-confinamiento/
3. https://elpais.com/sociedad/2020-06-12/el-consumidor-tras-el-coronavirus-mas-compras-por-internet-y-menos-ropa.html

Por otro lado, en cuanto a la venta de videojuegos vemos como el mercado físico baja en pro de la venta online.

Fuentes:

1. https://elordenmundial.com/mapas/evolucion-mercado-videojuegos/
2. https://es.statista.com/estadisticas/472651/prevision-de-valor-de-ventas-de-juegos-de-consola-en-espana/
3. https://www.qualitydevs.com/2019/04/03/evolucion-videojuegos-espana/

II. Obtener los datos

Los datos sobre el dataset de Amazon se han obtenido de Kaggle.com y los datos sobre el mercado de venta física de videojuegos son de data.world.

III. Data Wrangling

Los datos se obtuvieron en formato csv, posteriormente se pasaron a DataFrame de forma que, en ambos dataset se han seguido los siguientes pasos:

1. Se crea una función en el archivo mining_data_tb.py que incluya doble línea "//" en la lectura del path.

2. Se importa esa función desde utils.mining_data_tb.

3. Se lee el csv con pandas, con la ruta del path ya cambiada y se incluye la separación que tiene el csv, en ambos casos una coma ",".

IV. Data Mining

Para limpiar los datos en el dataset de Amazon, se han seguido estos pasos:

1. **Comprobación de valores NaN:** si hay, cuántos hay y en qué columnas están. (Resultado: solo hay NaN en la columna de precio)
2. **Comprobación de duplicados:** si hay y, en ese caso, eliminarlos.
    a) Se limpia la columna de precio: se pasa a string para quitar el símbolo de $.
    b) Se separa en dos columnas (high_price y low_price) y se pasa a float.
    c) Se reemplaza por cero los valores NaN en la columna low_price
    d) Se ordena por low_price y se eliminan duplicados cogiendo el precio más alto. (Resultado: de 2 millones de filas pasamos a 1 millón).
3. **Comprobación de valores NaN de nuevo:**
    a) Se itera sobre la columna "price" para saber dónde hay valores NaN y reemplazar los ceros en la columna low_price por NaN donde corresponda.
    b) Volvemos a contar valores NaN. (Resultado: es un 0,38% del total)
4. **Comprobación de precios:** para saber si podemos rellenar los datos de los valores NaN.
    a) Se saca la lista de los links de la columna "links" para comprobar que hay precios y, en ese caso, hacer web scraping.
    b) Se copian varios links al azar y se comprueba si tienen precios pegándolos en el buscador. (Resultado: 1 de 10 tiene precio)

c) Dado que la posibilidad de rellenar los datos es muy baja y el porcentaje de valores null es un 0.38% (solo en la columna de precio) se descarta hacer web scraping.

5. **Se limpian y se eliminan columnas:**
   a) Se separa la columna "category" por sub categorías.
   b) Se vuelve a separar la columna creada en sub categorías.
   c) Se limpia la columna rating y se hace una nueva columna con solo la puntuación y se pasa a float.
   d) Se eliminan las columnas innecesarias.
   e) Se ordena el DataFrame.
6. **Se limpia la columna de Main_Category:** las categorías están duplicadas ya que algunas tienen un espacio previo y al agruparlas se agrupan en diferentes celdas.

Para limpiar los datos en el dataset de videojuegos, se han seguido estos pasos:

1. **Comprobación de valores NaN:** si hay, cuántos hay y en qué columnas están. (Resultado: solo hay NaN en las columnas "Year" y "Publisher", un 1,98% del total) Se descarta también completar los datos ya que los valores null es un porcentaje mínimo.
2. **Comprobación de duplicados:** si hay y, en ese caso, eliminarlos. (Resultado: no hay duplicados)
3. **Se limpian columnas:** el dataset es hasta 2017 y hay un dato que es en 2020, es necesario cambiarlo por su dato real.

V. Visualización

Dependiendo de la pregunta, se ha utilizado una gráfica u otra. Principalmente, los modelos elegidos son histogramas, mapa de correlación, pie chart y diagrama de línea.

Las librerías de visualización utilizadas para ello son:

- Matplotlib
  - Histograma
  - Pie chart
  - Diagrama de línea
- Seaborn
  - Mapa de correlación

VI. Otros

VI.I Responder preguntas nivel C

- **¿Ha sido posible demostrar la hipótesis? ¿Por qué?**
  La hipótesis ha quedado refutada ya que la hipótesis principal era: "La electrónica es la categoría más vendida en Amazon". Los datos muestran que la categoría más vendida es deporte y aire libre.
  La hipótesis secundaria: "Los videojuegos son la subcategoría de electrónica más vendida en Amazon". También ha quedado refutada ya que los videojuegos están en una categoría aparte, es decir, no pertenecen a la electrónica como subconjunto, sino que tienen su propia categoría y esta no es de las más vendidas ya que está en el puesto 20.

- **¿Cuáles son las conclusiones?**
  1. La electrónica no es la categoría más vendida, sino **deporte y aire libre**.
  2. Parece **importante para la venta** que un producto tenga muchas **reseñas**.
  3. Los **videojuegos** no pertenecen a la categoría Electrónica en Amazon y su venta tampoco es de las más grandes.
  4. Las **5 plataformas** más vendidas abarcan más del **50% de la venta** física de videojuegos.
  5. A pesar de que el **videojuego más vendido** en todo el mundo es de la Wii, esta no es la plataforma que más videojuegos vende.
  6. La **evolución** de la venta de **videojuegos** en forma física tiene una tendencia descendente en los últimos años.

- **¿Qué cambiarías si tuvieras que hacer otro proyecto EDA?**
  1. Hacer **web scraping**, para sacar los valores de los null.
  2. Hacer **nuevas funciones** para poder sintetizar más el código.
  3. Añadir **nuevos datos**, sobre la venta online de videojuegos.

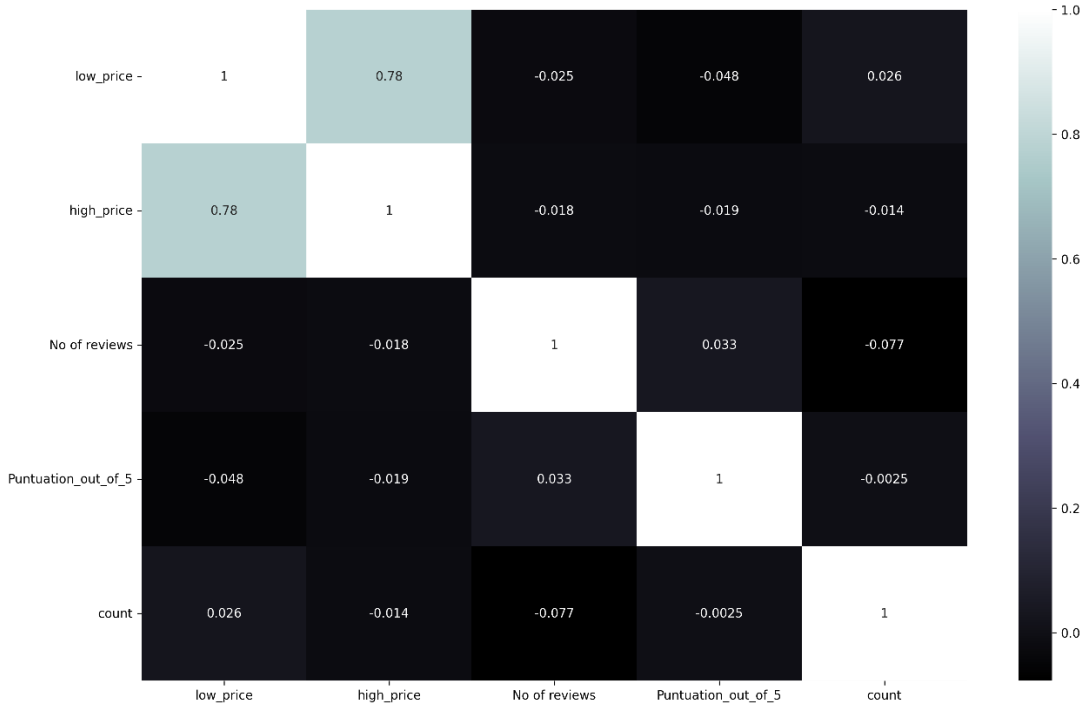
- **¿Qué has aprendido con este proyecto?**
  He aprendido a gestionar el tiempo y considerar qué es más importante a la hora de realizar un proyecto, como por ejemplo, valorar si es imprescindible o no sacar los valores null en base al tiempo vs resultado.
  También a utilizar librerías de visualización como matplotlib o seaborn, importar funciones y guardar las gráficas en una carpeta concreta, muy útil de cara al futuro.

VI.II Responder preguntas nivel B

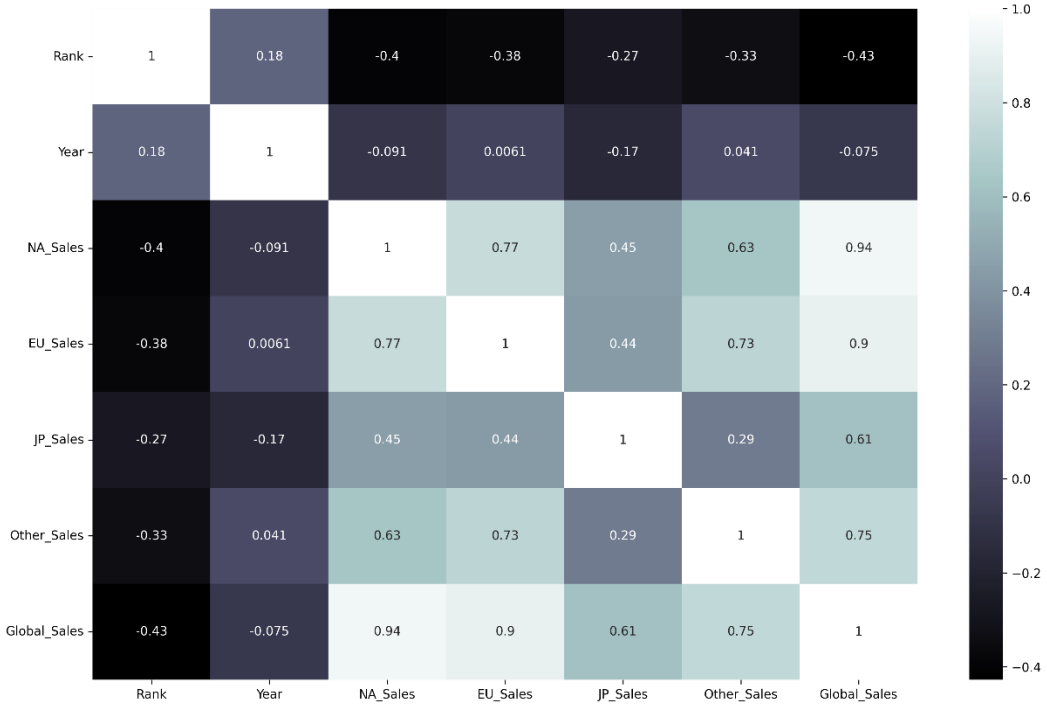
- **Mostrar cada histograma de cada columna usando bins = 5. ¿Cuántos rangos se han pintado?**

  Esto solo es posible en las columnas numéricas, dada esta casuística solo lo puedo hacer en una columna de Year del dataset de videojuegos. En esa gráfica se han pintado 8 rangos.

- **¿Cuáles son las columnas con más correlación? Mostrar la matriz de correlación.**



Esta es la matriz de correlación referente al dataset de Amazon. Las columnas con mayor correlación son low_price y high_price, pero este caso no es relevante para el estudio del proyecto.



Esta es la matriz de correlación del dataset de videojuegos. Las columnas con mayor correlación son Global_Sales con Na_Sales, EU_Sales, JP_Sales y Other_Sales. Esto tiene sentido ya que la suma de estas últimas es igual a la columna Global Sales.

También tiene relación de forma inversa la columna de Rank con todas las de ventas, ya que cuanto mayor sea la venta el ranking es más bajo (es decir, contando el primer puesto "1" como el más bajo).

- **Usar matplotlib para mostrar todas las gráficas. No directamente pandas.**
  Como he comentado anteriormente, las gráficas están hechas con matplotlib.