Contenido

[VENTA DE VIDEO JUEGOS 2](#_Toc73107042)

[Fase 1. definición de necesidades del cliente (Comprensión del negocio) 2](#_Toc73107043)

[Fase 2. Estudio y comprensión de los datos 2](#_Toc73107044)

[Fase 3. Análisis de los datos y selección de características 3](#_Toc73107045)

[**Grafico de Torta por marca de consola** 5](#_Toc73107046)

[**Grafico de histogramas por región de ventas** 6](#_Toc73107047)

[**Gráfico de violín por marca de consola** 7](#_Toc73107048)

[**Gráfico de dispersión** 7](#_Toc73107049)

[**Gráfico de tendencia** 8](#_Toc73107050)

[**Gráfico de bigotes o boxplot** 9](#_Toc73107051)

[**Correlación** 9](#_Toc73107052)

[**Gráfico de líneas – histórico de ventas de video juegos por plataforma o consola** 11](#_Toc73107053)

[**Relación de la economía con la venta de video juegos en el mundo** 12](#_Toc73107054)

[Fase 4. Modelado 14](#_Toc73107055)

[**Regresión Lineal** 14](#_Toc73107056)

[**Random Forest Regressor - Bosques Aleatorios para Regresión** 15](#_Toc73107057)

[Fase 5. Evaluación 15](#_Toc73107058)

[Fase 6. Despliegue (Puesta en producción) 16](#_Toc73107059)

VENTA DE VIDEO JUEGOS

Fase 1. definición de necesidades del cliente (Comprensión del negocio)

El mercado de los video juegos ha venido ganando protagonismo en los últimos años, hay una gran variedad de marcas creadores y publicadoras de video juegos, sin embargo, el mercado de las consolas o plataformas es muy oligopólico, ya que unas pocas marcas representan casi la totalidad de la torta en cuanto consolas.

Se desea observar el porcentaje de ventas de video juegos por marca.

Se quiere conocer la correlación de los datos, con el fin de determinar si hay colinealidad entre variables y si hay variables predictoras interesantes.

Se desea observar el comportamiento de los video juegos lanzados en los diferentes años por plataforma y por marca.

Fase 2. Estudio y comprensión de los datos

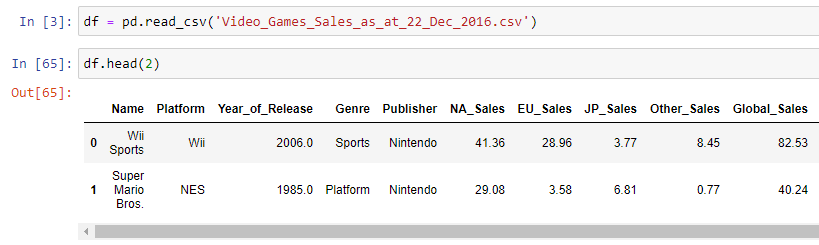
El dataset `Video\_Games\_Sales\_as\_at\_22\_Dec\_2016.csv` está compuesto por 18 columnas y 16.719 registros o filas, consiste en video juegos lanzados desde 1980 hasta el año 2020, con su fecha de lanzamiento, ventas en cantidad en diferentes regiones y a nivel mundial, entre otros. A continuación, se detallan las variables y su descripción.

|  |  |
| --- | --- |
| **Variable** | **Descripción** |
| Name | Nombre del video juego |
| Platform | Consola en la que funciona el video juego |
| Year\_of\_Release | Año de lanzamiento |
| Genre | Género o categoría del video juego |
| Publisher | Marca que publica el video juego |
| NA\_Sales | Cantidad de ventas en EEUU |
| EU\_Sales | Cantidad de ventas en Europa |
| JP\_Sales | Cantidad de ventas en Japón |
| Other\_Sales | Cantidad de ventas en otros países |
| Global\_Sales | Cantidad de ventas a nivel global |
| Critic\_Score | Puntaje o calificación otorgada al video juego por parte de la critica |
| Critic\_Count | Conteo de los críticos que calificaron el video juego |
| User\_Score | Puntaje o calificación otorgada al video juego por parte de los usuarios |
| User\_Count | Conteo de los usuarios que calificaron el video juego |
| Developer | Marca creadora del video juego |
| Rating | Clasificación del video juego |

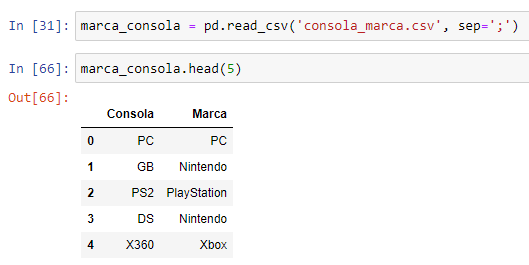
También se elabora un dataset adicional compuesto por la `Platform` o consola y la marca de esta consola, esto con el fin de agrupar por marca de consola y analizar por esta variable.

Fase 3. Análisis de los datos y selección de características

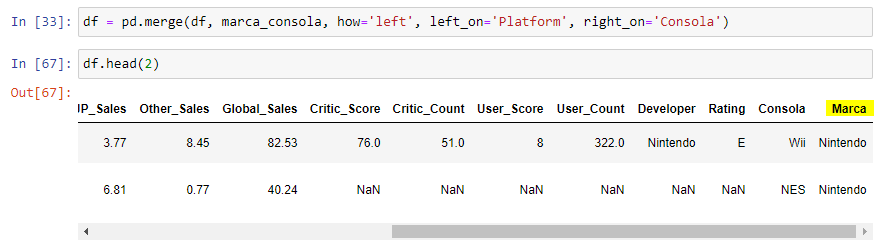
Se carga el archivo de la facturacion por video juegos y se lleva a la variable ‘df’



Se carga un archivo armado manualmente el cual tiene la plataforma o consola y la marca a la que pertenece. Para este caso se utiliza el parámetro ‘sep’ para indicar el delimitador.

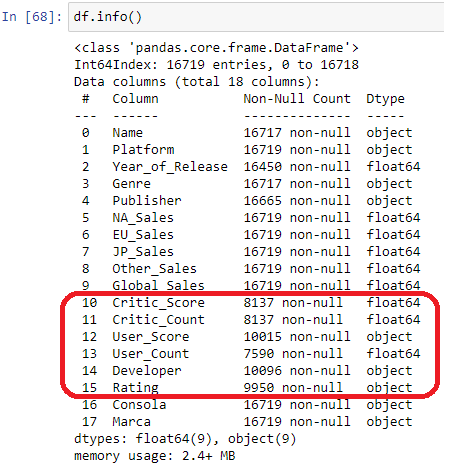


Se utiliza la función ‘merge’ de la librería pandas, para cruzar dos tablas o dataframes, se indica las llaves de cruce y el tipo de cruce, en este caso ‘left’, el cual consiste en que la tabla de la izquierda se conserva y se traen los datos que cruzan de la tabla de la derecha.

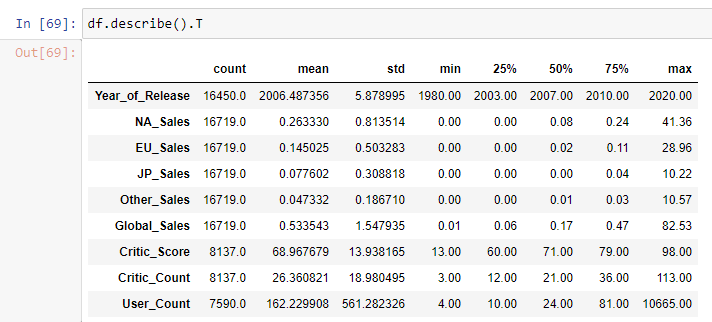


Se ejecuta el método que muestra el tipo de dato que tiene cada columna y cuantos registros ‘no nulos’ tiene. Esto con el fin de determinar que columnas se van a tener en cuenta para los análisis.

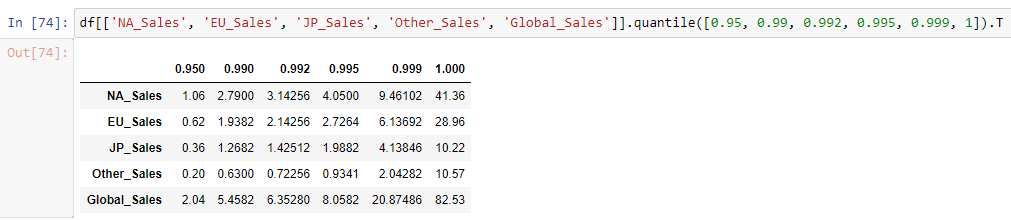
En rojo se marcan las columnas que poseen una gran cantidad de valores nulos.



Se genera una descripción estadística de los datos



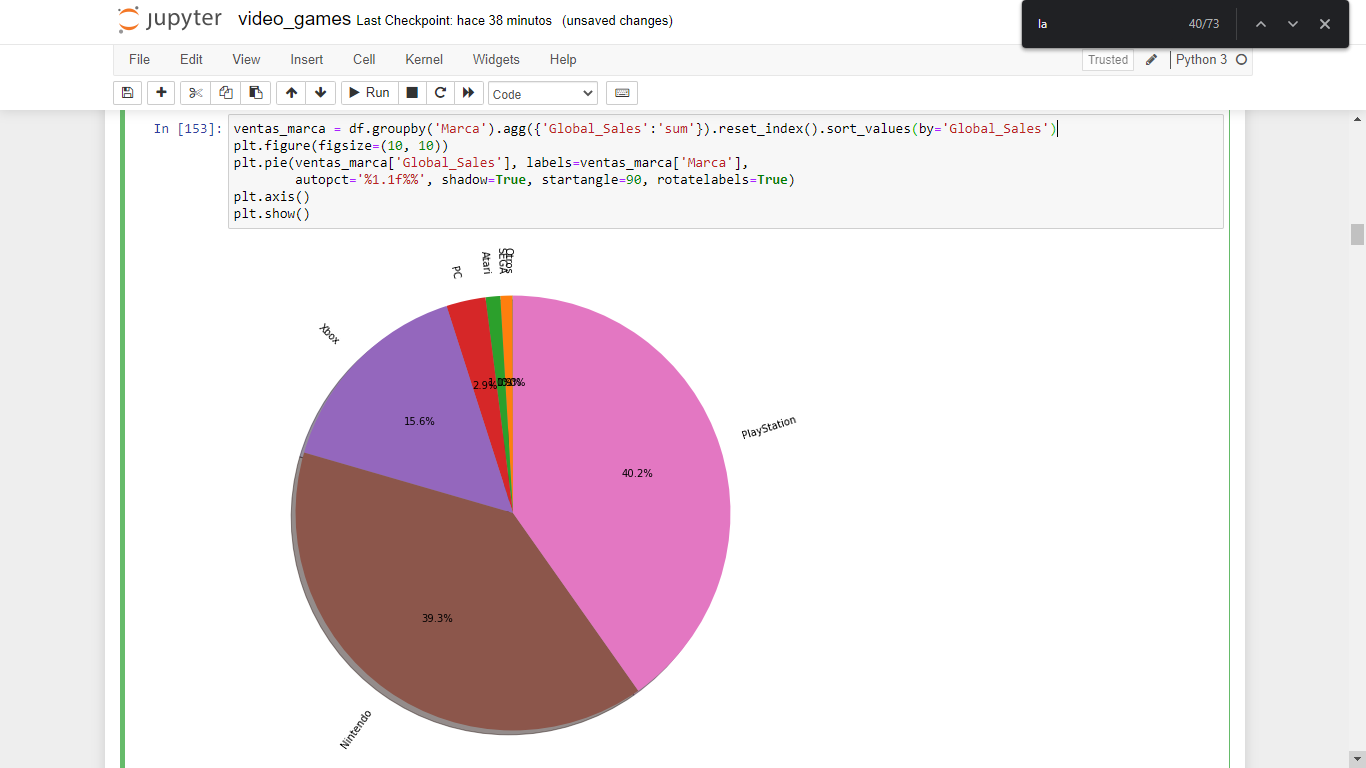
Se generan los cuantiles 0.95, 0.99, 0.992, 0.995, 0,999 y 1 con el fin de determinar si existen outliers que impidan la exploración estadística.



### **Grafico de Torta por marca de consola**

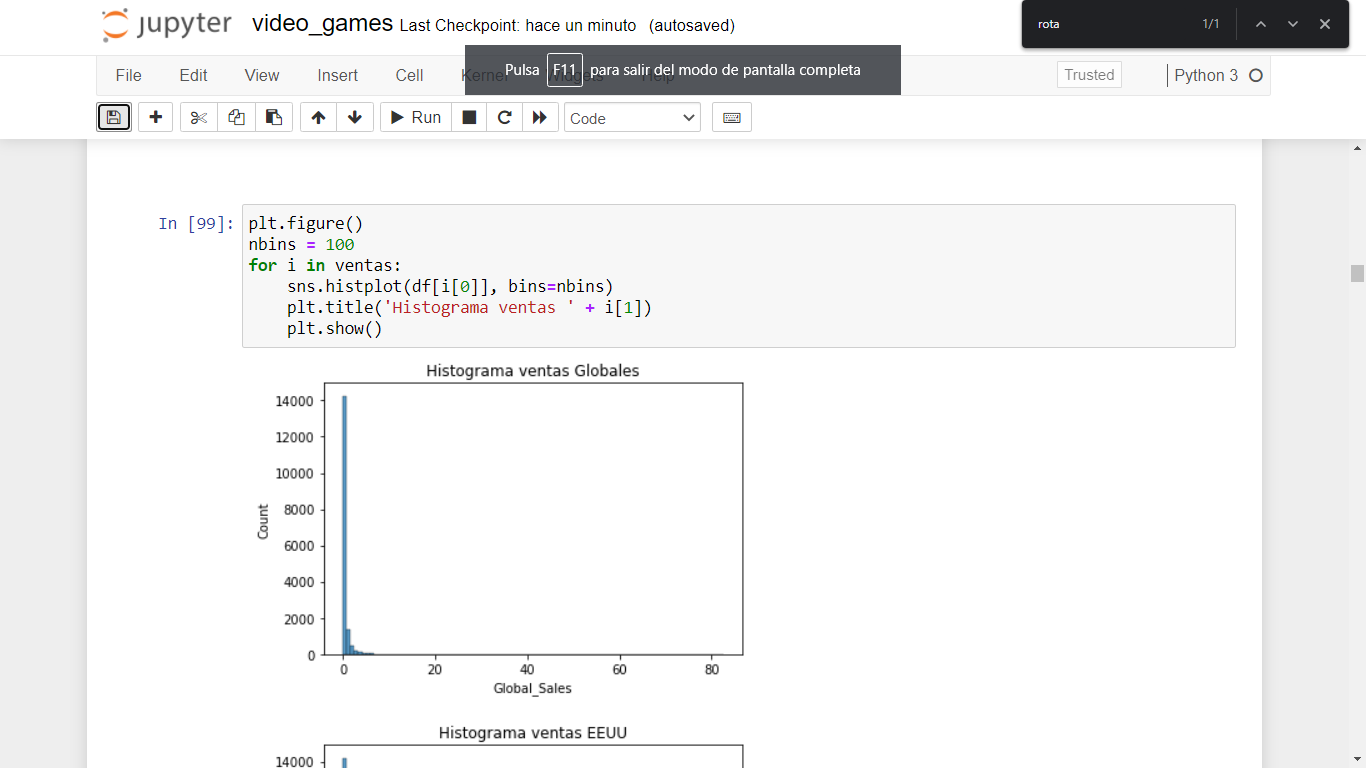
Se genera una agrupación por marca de plataforma y sumatoria de las ventas globales de video juegos, para esta agrupación se genera un gráfico de torta.

Se evidencia que la mayoría de las ventas de video juegos está concentrada en tres marcas de plataformas principales, PlayStation con 40.2%, Nintendo con un 39.3 % y Xbox con un 15.6%



### **Grafico de histogramas por región de ventas**

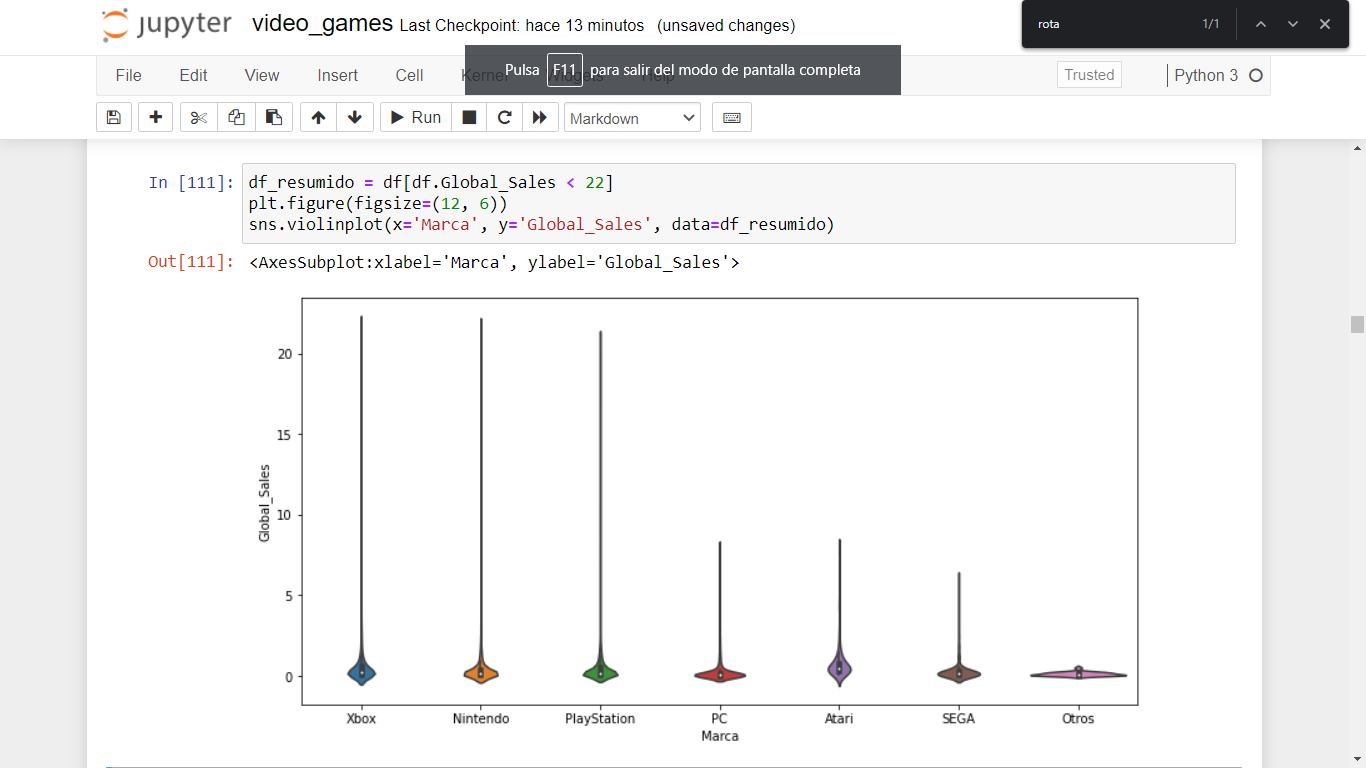
Utilizando un ciclo `for` se generan histogramas para las diferentes variables de ventas de video juegos, vemos que es asimétrico hacia la derecha.



### **Gráfico de violín por marca de consola**

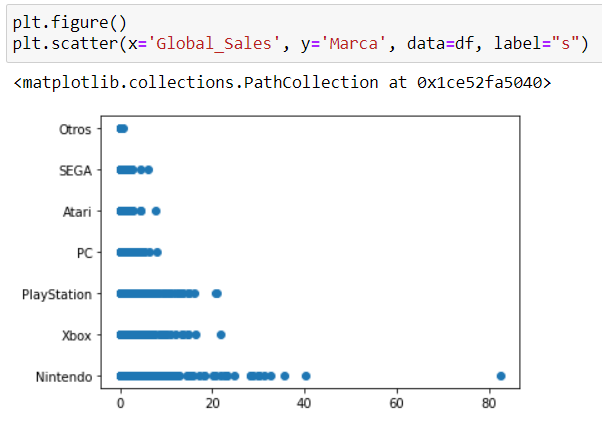
Se genera un gráfico de violín de ventas de video juegos por marca de consola, en este caso no se toman todos los valores, sino el 99.5% de ellos, con el fin de excluir los valores mayores y así poder visualizar correctamente.

Se evidencia que las ventas de video juegos por marca de consola, muestra una gran densidad (frecuencia) en los valores cercanos bajos.



### **Gráfico de dispersión**

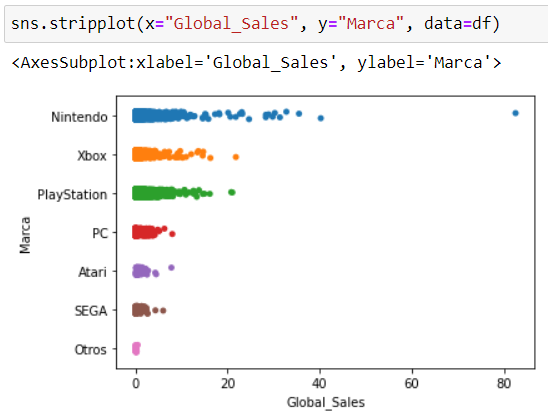
Se evidencia que en la marca de consolas `Nintendo` hay video juegos con grandes ventas, `PlayStation` y `Xbox` tienen un comportamiento muy similar, en estas dos marcas hay algunos video juegos que destacan en ventas sobre los demás.



### **Gráfico de tendencia**

Se genera un gráfico de tendencia de ventas globales de video juegos por marca de consola.

Se logra ver que para la marca `Nintendo`, hay video juegos con gran cantidad de ventas, pero sin tanta densidad en esas zonas, para `Xbox` y `PlayStation`, no hay tantos video juegos con ventas tan elevadas.



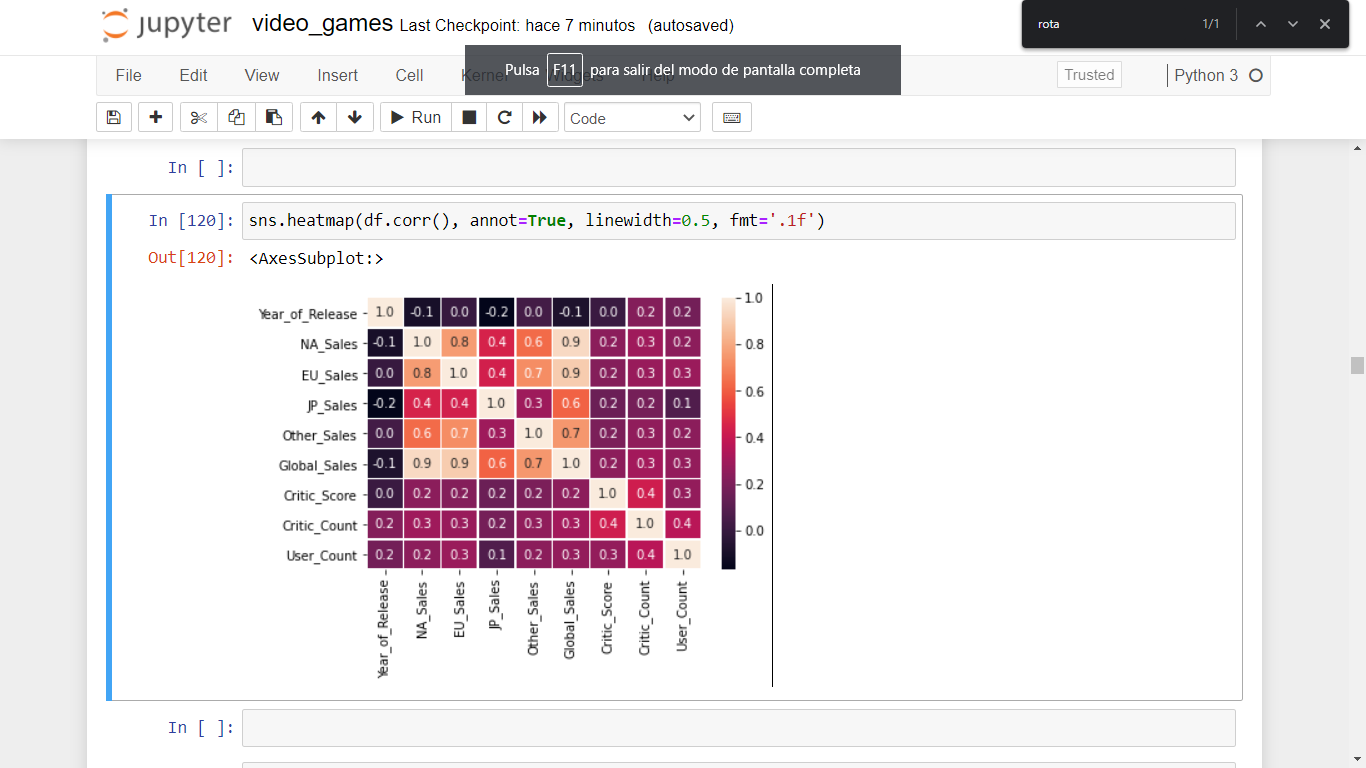
### **Gráfico de bigotes o boxplot**

En las 3 marcas principales de consolas, el promedio de ventas de video juegos es muy similar, según la gráfica, `Nintendo` y `PlayStation` tienen mayor número de video juegos con altas ventas.

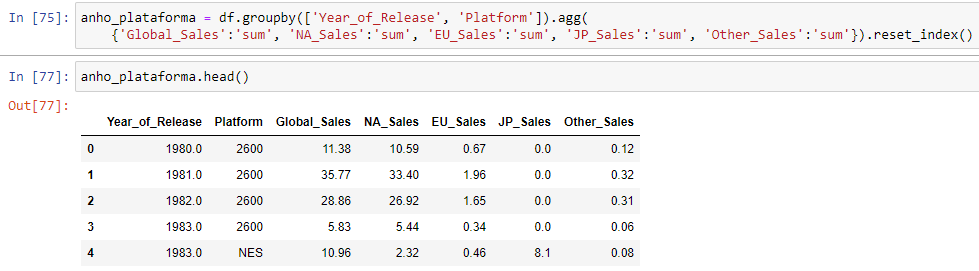


### **Correlación**

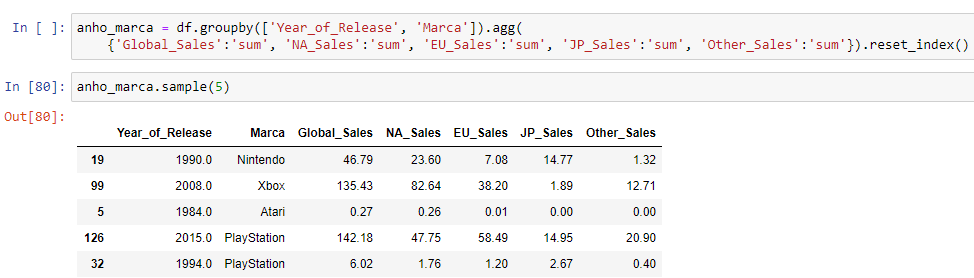
Se genera un mapa de calor con la correlación entre las variables numéricas, las diferentes variables de ventas están altamente correlacionadas entre sí, lo que muestra una colinealidad entre las variables, por tal razón no se podrían utilizar como predictores.



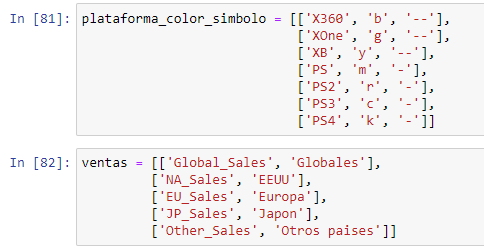
Se genera una agrupación por ‘año de lanzamiento’ y ‘plataforma’ y se totaliza por ‘ventas globales’, ‘ventas en EEUU’, ‘ventas en Europa’, ‘ventas en Japón’ y ‘otras ventas’.



También se genera una agrupación por ‘año de lanzamiento’ y ‘marca’ y se totaliza por las mismas variables del ejemplo anterior.

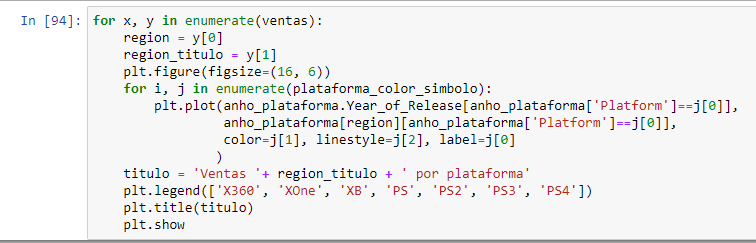


Se crean dos listas para la graficación, en la primer lista se definen las plataformas, el color y el tipo de linea, en la segunda lista se define el nombre de columna en el dataframe y el titulo que se va a graficar



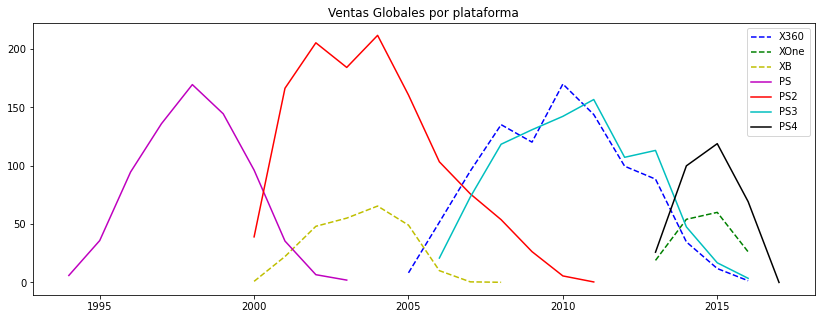
### **Gráfico de líneas – histórico de ventas de video juegos por plataforma o consola**

Se genera un ciclo `for`, con el fin de recorrer la lista que tiene los títulos de las columnas de ventas, después se declara otro ciclo `for` para graficar una línea por cada una de las plataformas, por último, se genera el título y las leyendas para cada uno de los gráficos.



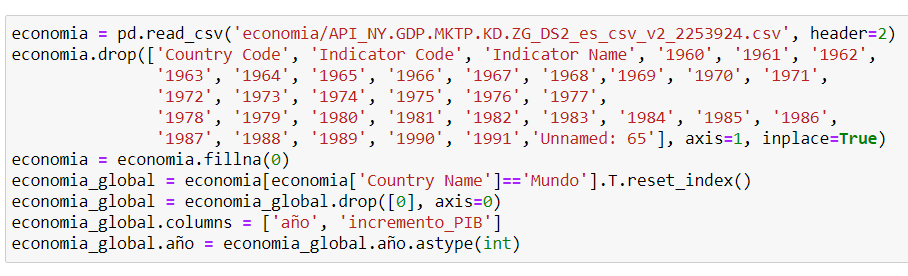
En el grafico podemos observar que cada que se lanza una nueva consola de la misma marca, el lanzamiento de video juegos de la consola anterior tiene un declive acelerado.

Las fechas de lanzamiento de los video juegos para ambas marcas PlayStation y Xbox por lo general son por las mismas fechas, lo cual denota una fuerte competencia.



### **Relación de la economía con la venta de video juegos en el mundo**

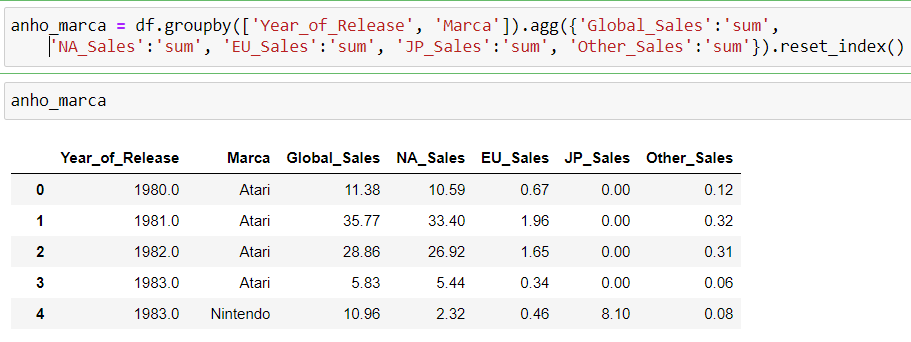
Se carga un archivo con el crecimiento del PIB (Producto interno bruto) por países, regiones y global.



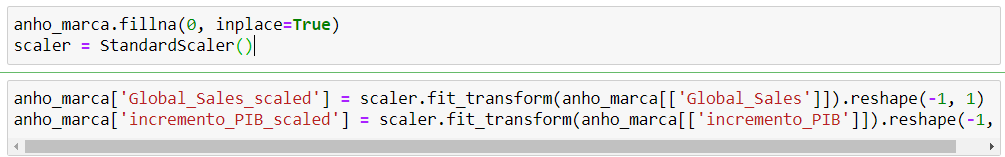
Se crea un gráfico de líneas, para mostrar el comportamiento del PIB a nivel mundial



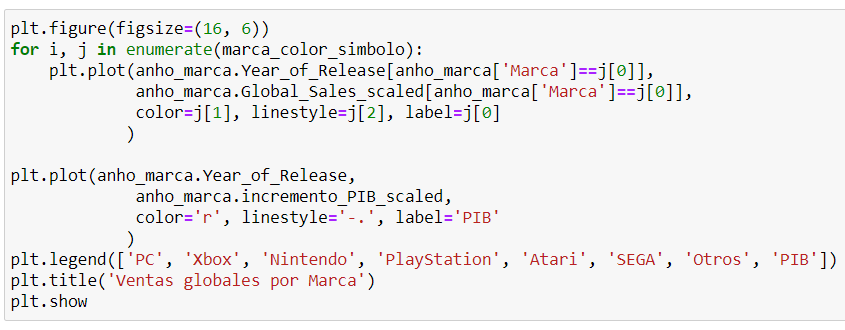
Se realiza una agrupación por marca y se sumariza por las diferentes variables de ventas

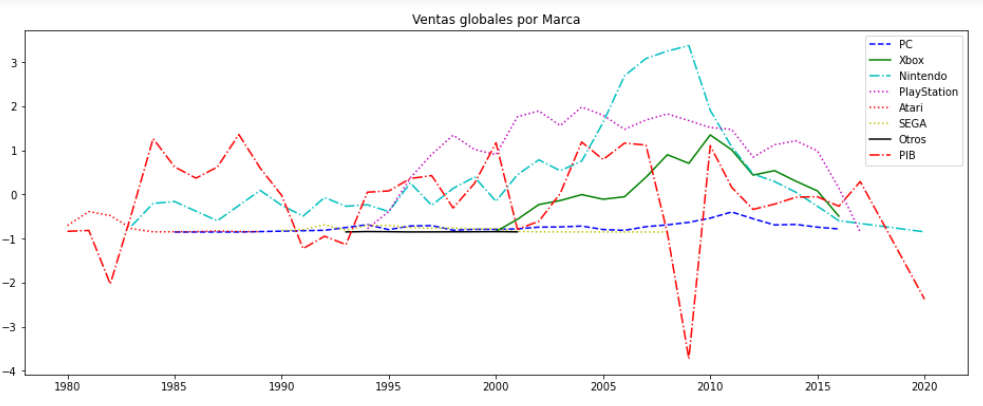


Se realiza el escalamiento de las variables ventas globales de video juegos y crecimiento año a año del PIB, con el fin de poder ver la comparativa visualmente.



Se genera una gráfica de línea para observar el comportamiento de las ventas globales de video juegos por marca de consola e incremento del PIB por año. Esto con el fin de buscar patrones entre ambas variables.





Fase 4. Modelado

Se borran los registros que no tienen los datos de críticas, con el fin de utilizar esas variables para utilizar un modelo predictivo de regresión. Se selecciona un modelo de regresión, ya que la variable a predecir es numérica continua.



Se separa el dataset en dos conjuntos de datos:

\* dataset `X` que corresponde a las variables predictoras, se excluyen las variables asociadas a ventas, debido a que existe una correlación superior al 80% con la variable a predecir de `Global\_Sales`, existe colinealidad.

\* dataset `y` que corresponde a la variable a predecir `Global\_Sales`

También se utiliza el método `get\_dummies` el cual sirve para convertir las variables categóricas a numéricas, este metodo realiza un `One-Hot Encode`.



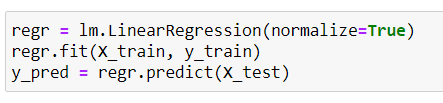
Se utiliza el método `train\_test\_split` el cual divide aleatoriamente los conjuntos de datos, con el fin de generar un conjunto de datos de entrenamiento y otro de pruebas.

En este caso se indicó que el conjunto de prueba es el 20% del total de los datos.



### **Regresión Lineal**

Se define el modelo de regresión lineal, posteriormente se entrena y por ultimo se predice la variable de salida para el conjunto de datos de prueba.



### **Random Forest Regressor - Bosques Aleatorios para Regresión**

Se crea un Bosque de árboles aleatorios, el cual es un método de ensamble bastante robusto, que combina varios árboles de decisión.

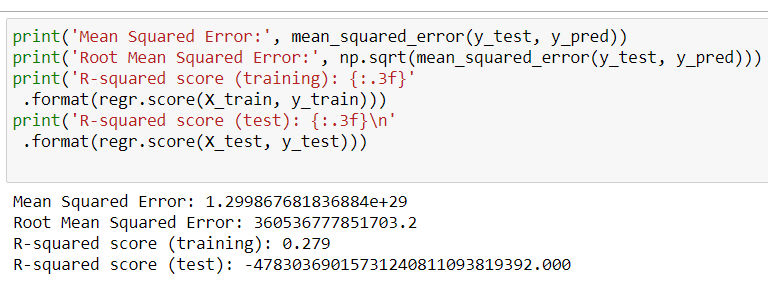
el algoritmo entrena cada modelo secuencialmente con todos los datos y, para cada nuevo modelo, se le da más peso a los datos que no fueron bien clasificados o cuyo error en regresión sea más alto.

Se complemento con el método `GridSearchCV` la sirve para buscar los mejores parámetros para el modelo, para cada parámetro se indica una lista con posibles valores, este método seleccionara los mejores.

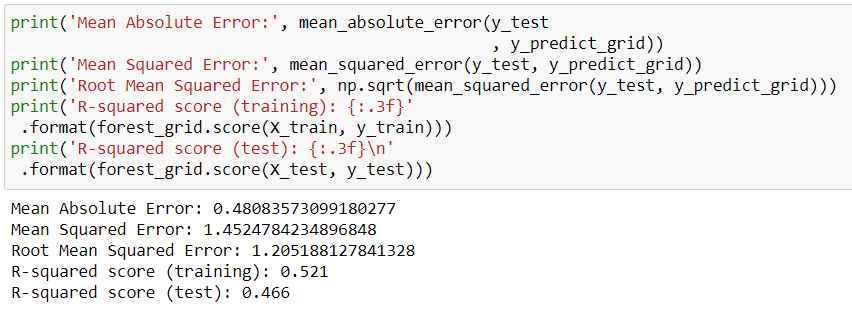


Fase 5. Evaluación

Se evalúa el modelo de regresión lineal, pero obtiene resultados muy bajos.



Se utilizan las mismas métricas utilizadas en la regresión lineal. se observa mejores resultados.  
Para el conjunto de datos de entrenamiento tenemos un resultado de 0.521 y para los datos de prueba tenemos 0.466



Se selecciona el método Random Forest Regressor, debido a que se obtienen mejores resultados, sin embargo, cabe destacar que es un modelo bastante robusto que es costoso computacionalmente.

Fase 6. Despliegue (Puesta en producción)

Como simulación de despliegue se montara el código a GitHub, para que cualquier persona lo pueda consultar y mirar los resultados.