Gordon McIntire

Time-Series Analysis of Video Games Sales:
Final Report

Problem Statement

According to a report by game and esports data firm Newzoo (via VentureBeat), the market is expected to drop in 2021 for the first time since the firm began tracking in 2012. Of course, this is following the unprecedented circumstances brought on by the Covid-19 epidemic. People hunkered down and played video games at a level higher than ever before, so the next year could be seen as the market readjusting itself. In this project, I will build models to examine/predict the volatility as well as forecast video game sales.

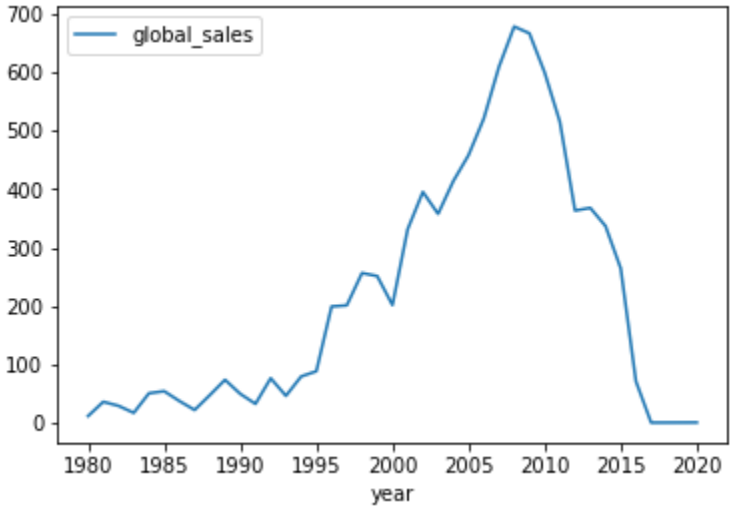
Data Cleaning/Transforming

The dataset I'm working with contains a list of video games with sales greater than 100,000 copies. It was generated by a scrape, based on BeautifulSoup using Python, of vgchartz.com.

There are 16,598 records, each containing the fields:

- Rank - Ranking of overall sales
- Name - The games name
- Platform - Platform of the games release (i.e. PC,PS4, etc.)
- Year - Year of the game's release
- Genre - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)
- Global_Sales - Total worldwide sales

I started by dropping all null values. I then converted the year column to a datetime in order for it to work correctly in a time series, and combined it with global sales. Here is a simple plot of the data:

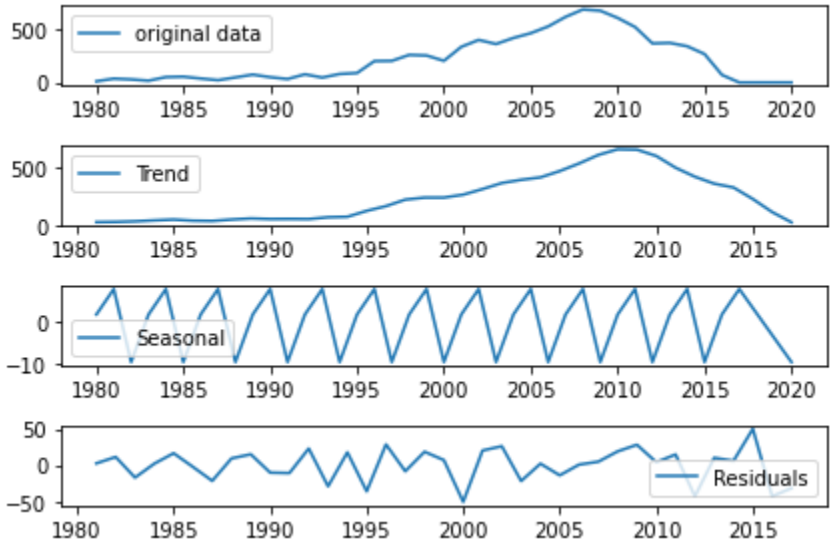


One can spot a steep upwards trend, over the past 25 years especially, followed by a similarly steep dropoff.

Modeling

1.Decomposition

To understand the yearly global sales data better, I used the decompose function, breaking the time series data into trend, seasonality, and noise.



2. KPSS Test for Stationarity

As is essential with Time Series data, I tested whether it was stationary using a KPSS test. The resulting p-value was over 0.05, which meant the data was stationary and suitable for the next step.

3. ARIMA Model

Once I knew the data was stationary, I then wanted to use an ARIMA model to forecast, which required finding the best parameters. To do this I made functions to find the MSE of a single ARIMA model and try different models with different parameters. Eventually I obtained the optimum model ARIMA model for my data.

Summary of the model:

ARIMA Model Results

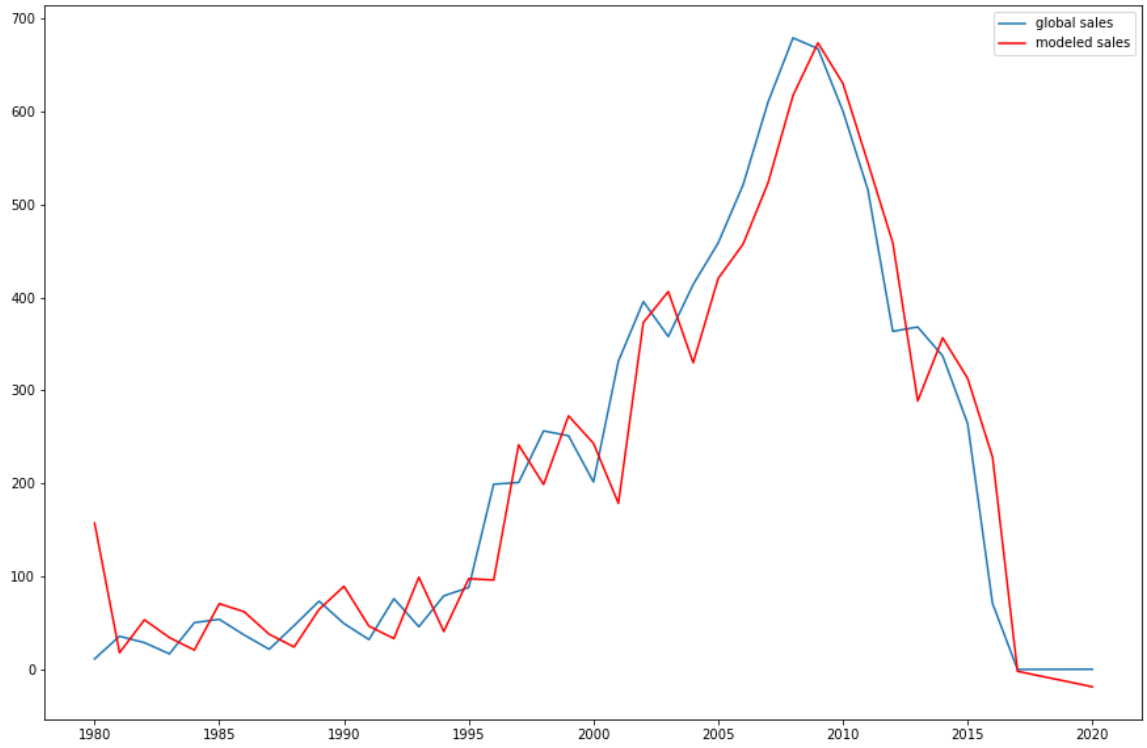
Dep. Variable:	D.#CigSales	**No. Observations:**	143
Model:	ARIMA(2, 1, 1)	**Log Likelihood**	1263.116
Method:	css-mle	**S.D. of innovations**	0.000
Date:	Wed, 28 Jul 2021	**AIC**	-2516.233
Time:	15:17:09	**BIC**	-2501.418
Sample:	02-01-1949	**HQIC**	-2510.213
	- 12-01-1960		

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.624e-06	5.06e-07	5.184	0.000	1.63e-06	3.62e-06
ar.L1.D.#CigSales	0.4681	0.156	3.003	0.003	0.163	0.774
ar.L2.D.#CigSales	-0.2640	0.109	-2.413	0.016	-0.478	-0.050
ma.L1.D.#CigSales	-0.8693	nan	nan	nan	nan	nan

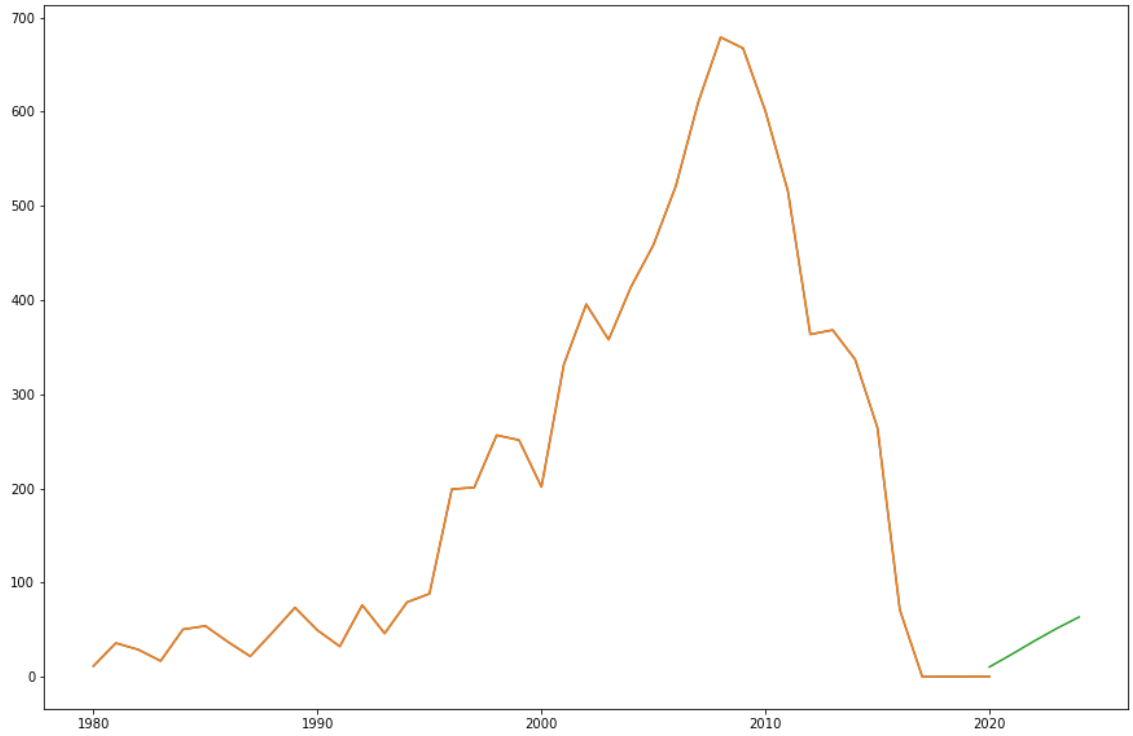
Roots

	Real	Imaginary	Modulus	Frequency
AR.1	0.8866	-1.7326j	1.9463	-0.1747
AR.2	0.8866	+1.7326j	1.9463	0.1747
MA.1	1.1504	+0.0000j	1.1504	0.0000

Original dataset plotted against my model:



Past sales including 5-year forecast, predicting a comeback in sales:



4. ARCH Model

Constant Mean - ARCH Model Results

Dep. Variable:	global_sales	**R-squared:**	0.000
Mean Model:	Constant Mean	**Adj. R-squared:**	0.000
Vol Model:	ARCH	**Log-Likelihood:**	-171.643
Distribution:	Normal	**AIC:**	355.286
Method:	Maximum Likelihood	**BIC:**	363.490
		No. Observations:	29
Date:	Fri, Jul 30 2021	**Df Residuals:**	28
Time:	09:04:20	**Df Model:**	1

Mean Model

	coef	std err	t	P>|t|	95.0% Conf. Int.
mu	45.4172	6.179	7.350	1.982e-13	[33.306, 57.528]

Volatility Model

	coef	std err	t	P>|t|	95.0% Conf. Int.
omega	477.0653	208.198	2.291	2.194e-02	[69.005,8.851e+02]
alpha[1]	1.0000	0.256	3.901	9.582e-05	[0.498, 1.502]
alpha[2]	3.7836e-13	0.162	2.331e-12	1.000	[-0.318, 0.318]
alpha[3]	1.9860e-12	0.141	1.410e-11	1.000	[-0.276, 0.276]
alpha[4]	1.8119e-12	1.030e-02	1.759e-10	1.000	[-2.019e-02,2.019e-02]

Covariance estimator: robust

Judging from the high Log-Likelihood value, this model is a relatively good fit. Combined with the fact that all of our mean and volatility coefficients were significant, this model should be able to reliably predict future variance.

5. GARCH Model

Constant Mean - GARCH Model Results

Dep. Variable:	global_sales	**R-squared:**	0.000
Mean Model:	Constant Mean	**Adj. R-squared:**	0.000
Vol Model:	GARCH	**Log-Likelihood:**	-182.947
Distribution:	Normal	**AIC:**	379.894
Method:	Maximum Likelihood	**BIC:**	389.465
		No. Observations:	29
Date:	Fri, Jul 30 2021	**Df Residuals:**	28
Time:	09:11:50	**Df Model:**	1

Mean Model

	coef	std err	t	P>|t|	95.0% Conf. Int.
mu	69.9772	15.926	4.394	1.114e-05	[38.762,1.012e+02]

Volatility Model

	coef	std err	t	P>|t|	95.0% Conf. Int.
omega	1.1193e+04	4359.510	2.568	1.024e-02	[2.649e+03,1.974e+04]
alpha[1]	0.9523	1.209	0.788	0.431	[-1.417, 3.322]
alpha[2]	7.7184e-14	1.713	4.506e-14	1.000	[-3.357, 3.357]
alpha[3]	8.7270e-14	1.614	5.407e-14	1.000	[-3.163, 3.163]
alpha[4]	1.0349e-13	1.161	8.917e-14	1.000	[-2.275, 2.275]
beta[1]	3.1899e-13	0.319	1.001e-12	1.000	[-0.624, 0.624]

The Log_Likelihood of the GARCH model is slightly less than the ARCH alternative, making it a worse fit.

Due to a relatively small amount of data, ARCH is the more appropriate model to use because of the complexity for which the GARCH model typically accounts.

Takeaways

The conclusions that I draw from my analysis of video game sales over the past 40 years are that while the market suffered a dip over the last decade, it will likely turn back to reflect positive growth. The data also suggests a significant level of volatility in future sales.

Future Research

For the next project, I would like to get a much bigger dataset of sales to work with, down to month/week/daily sales. Only having annual sales, I obviously was not able to examine seasonal trends, just the long-term growth and decay.

It would also be interesting to look at other media sales' performance over the last decade and see whether there is any correlation between sets of data.