ARIMA forecasting in the collectibles assets market

Aurelien Giuglaris-Michael
Department of Advanced Computing Sciences
Faculty of Science and Engineering
Maastricht University
Maastricht, The Netherlands

***Abstract*—As trading between peers in the collectibles market becomes facilitated through technology, the data generated increases. New indices emerge in the space aggregating data from different assets to track the market. It is thus most interesting for actors in the space to have a forecast on this new kind of time series data to anticipate the future. This paper evaluates univariate and multivariate ARIMA models on a wine, watch, and art index showing promising results with a 90% accuracy improvement from the worst to the best-tuned ARIMA models. However, ARIMA models cannot predict sudden spikes (fat-tails) with a higher probability than the normal distribution suggests. As such due to the complexity of the data, the best ARIMA models cannot perform better than the arithmetic mean μ in their categories. This highlights the predictive power of simple models on complex time series in the collectibles market and the potential loss in accuracy and reliability that comes with complex models for in and out-of-sample forecasts.**

***Index Terms*—ARIMA models, forecasting, time series, collectibles, financial economics**

I. Introduction

This paper places itself in the collectibles asset market. Investments in those assets are becoming more accessible through technology. It is thus easier to trade between peers, and as a result, the amount of data available in the space has grown.

This paper addresses the need for actors in the space to have a global forecast for some of the most popular collectibles investments (watches, art, wine). This was not possible previously due to the accessibility and amount of data. Additionally, it is a need that cannot be addressed by examining or forecasting the prices of collectibles at auction which is what most of the forecasting work in the space has focused on so far.

The purpose of this paper is to be able to forecast prices of traditional collectible asset markets such as watches, wines, and art pieces. This forecasting relies on two strong assumptions: it assumes in its computations that there will be no black swans (major unforeseen and unpredictable events that have economic impacts: global wars, global natural catastrophes, global pandemics) in the coming years. Additionally, past performance must influence future performance.

This paper also innovates by fitting and examining the limitations of the predictive power in traditional ARIMA models to a new kind of time series data that has become more accessible in recent years: collectible asset indices. Indices are benchmarks that track the evolution of a basket of assets. In the context of this paper, those are baskets of wines, art paintings, and watches.

The data consists of a watch, art, and wine index retrieved online, partly through web scraping:

- The wine index, or Liv-Ex 100 Index has been retrieved from the Bloomberg terminal. According to Liv-Ex it is: "the industry-leading benchmark for monitoring fine wine prices. It represents the price movement of 100 of the most sought-after fine wines on the secondary market ", with monthly data price points going back up to July 2001 in GBP.
- The art index or All Art Index Family developed by Art Market Research (AMR): "records sales at auction every month and uses the price data to determine an average value for all artists in AMR's database. This value is called the Artist Price (AP) and these APs are combined to produce the All Art chart ", with monthly data going back up to 1978 in GBP.
- The watch index has been built for the need of this paper, by retrieving auction sale prices of five of the most popular collectible watches on the market according to Watchcharts in their overall watch market index consisting of 60 watches from 10 different luxury brands. The watches selected thus represent the top 5 watches of the index and account for 45.4% of the index combined. With weights proportional to their original weighting in the original Watchcharts index, this new watch index in EUR has monthly price points going back to November 2006. Since the sales auction prices

Table I: Components of Custom Watch Index

Name:	Weight
Patek Philipe 5711/1A	0.17
Rolex DateJust 126334	0.22
Rolex Daytona 116500	0.28
Rolex Daytona 116520	0.16
Rolex Daytona 116508	0.17

This thesis was prepared in partial fulfilment of the requirements for the Degree of Bachelor of Science in Data Science and Artificial Intelligence, Maastricht University. Supervisor(s): [Philippe Dreesen] are not exactly monthly, prices were duplicated if there was a gap in time between two auction price sales for a single watch before applying the weighted sum to make the index.

To obtain a reliable forecast and examine the predictive power of ARIMA models on this collectible asset index data, the paper [1] will answer the following research questions:

- RQ1: What are the strongest correlated variables with indices of each asset class?
- RQ2: What are the ARIMA models' univariate and multivariate performance across different asset classes in the collectibles market?

II. Related Work

Classical regression is often insufficient for explaining all of the interesting dynamics of a time series (Shumway et al., 2017). Instead, a linear relationship explained by the correlation between lagged values of time series data at time t gives rise to the auto-regressive and moving average models. Adding nonstationary models to the mix leads to the auto-regressive integrated moving average (ARIMA) model popularized in the landmark work by Box and Jenkins (Shumway et al., 2017). Time series data is everywhere, and ARIMA models have many applications such as predicting the Gross Domestic Product GDP (Yang et al., 2016), forecasting the Covid-19 virus behavior (Hernandez-Matamoros et al., 2020) or the energy consumption in Turkey (Ozturk and Ozturk, 2018). This comes from the fact that ARIMA models are quite agile, being able to factor in exogenous variables with ARIMAX and seasonal cycles that repeat (winter, summer, etc) with SARIMA being a non-linear model. However, Petrică et al., 2016 suggest that fat-tails for skewed plots (large losses or gains coming at a higher probability than the normal distribution would suggest) and volatility clustering, cannot be captured by integrated ARMA models, hence their limitation when being applied to financial and monetary economics with fat tails being decisive for a relevant analysis according to Christoffersen, 2011. Additionally, Commandeur and Koopman, 2007 argues that a problem arises because in "the economic and social fields, real series are never stationary [with mean 0 and variance 1] however much differencing is done".

Most of the forecasting work done in the space forecasts the potential price of collectibles at auction, such as Aubry et al., 2019 who use "a popular machine-learning technique—neural networks—to develop a price prediction algorithm based on both non-visual and visual artwork characteristics". Pires, 2020 uses "a data mining approach to predict the selling price of collectibles cars at auction and determine which vehicle's characteristics influence this value ". Or even, Penasse, 2014 who uses a "Mixed Data Sampling (MIDAS) modeling approach to forecast year-end art prices, using higher frequency variables related to the stock and bond markets and to art market sentiment ". Thus either forecasting collectible prices at auction or individual, specific short-term art price forecasts.

III. Methods

A. General Pre-processing

When comparing two economic variables, it is easier to quantify absolute changes when adjusting that data for inflation. Mendez-Carbajo, 2023 from the Federal Reserve Bank mentions that: "You can compare real values today with real values from the past because the general price level is held constant". The data was adjusted for inflation using historical CPI data and the following:

$$\text{current value} = \text{original value} \cdot \frac{\text{CPI}_{current}}{\text{CPI}_{past}} \tag{1}$$

Original value being the original dollar amount at time t in the past. $\text{CPI}_{current}$ is the latest CPI value. CPI_{past} is the past CPI rate at time t. More information on the CPI index is included in Appendix A. For a visualization of the final index data adjusted from inflation, please refer to Appendix B.

Additionally, as mentioned in the introduction, these index prices are expressed in different currencies: GBP and EUR, however for an analysis, and especially when comparing them to exogenous variables later on, it is necessary to express them in USD. Thus daily USD/GBP and USD/EUR rates were extracted and each daily rate inside a month was averaged to reflect a monthly USD/EUR and USD/GBP rate. Then for each monthly price point in the indices, the corresponding monthly rate according to its date was applied to convert the wine and art index from GBP to USD and the watch index from EUR to USD. Taking the appropriate historical rate, instead of the most recent one to convert our indices allows us to keep the time-dependence intact. This is crucial when adjusting for inflation later on.

B. ARIMA Traditional Strategy

The first model used in this problem is a standard Auto-Regressive Integrated Moving Average ARIMA(p, d, q) model.

$$y_t = c + \phi_1 y_{d,t-1} + ... + \phi_p y_{d,t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} \tag{2}$$

ϕ and θ are the p, q coefficients of the auto-regressive and moving average models respectively. All coefficients are estimated through maximum likelihood estimation. The goal is to get an optimal set of p, q coefficients that minimize the difference between the observed data and values predicted by the model. $y_{d,t-p}$ are the previous differenced observed values from $1, 2, ..., p$ past values. ϵ_{t-q} are the forecast residuals from $1, 2, ..., q$ past errors. c captures the constant upward or downward evolving trend of the data if there is one, and is

[1] The github repository of this paper is available here.

referred to as the drift of the model. The ARIMA model is tuned according to the following methodology. To find the d parameter, the data is checked for stationarity: a necessary assumption when using an ARIMA model. If stationarity is not obtained the data is differenced with n being the order of differentiation.

$$\frac{d^n y_t}{dx^n} = \frac{d^{n-1} y_t}{dx^{n-1}} - \frac{d^{n-1} y_{t-1}}{dx^{n-1}} \tag{3}$$

As such the d parameter of the ARIMA model is set to 1 as it becomes stationary after a first-order difference.

$$y'_t = y_t - y_{t-1} \tag{4}$$

Stationarity is checked beforehand and confirmed by both the Augmented Dickey-Fuller (ADF) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) statistical tests. Both test if the time series contains a unit root. If a time series does not contain a unit root, then it is stationary.

The p, q parameter candidates are initially found by looking at the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots of the data. The ACF plot calculates the correlation (Pearson correlation coefficient $p \in [-1, 1]$) in a time series between current observations y_t and past observations $y_{t-1}, y_{t-2}, ..., y_1$. The PACF plot also computes the same correlation as the ACF. However, the effect of intermediate lags on the correlation between y_t and y_k is removed with $t > k$. Hyndman and Athanasopoulos, 2018 suggests the following: “The data may follow an ARIMA(p,d,0) model if the ACF and PACF plots of the differenced data show the following patterns: the ACF is exponentially decaying or sinusoidal; there is a significant spike at lag p in the PACF, but none beyond lag p. The data may follow an ARIMA(0,d,q) model if the ACF and PACF plots of the differenced data show the following patterns: the PACF is exponentially decaying or sinusoidal; there is a significant spike at lag q in the ACF, but none beyond lag q ”. In practice, this technique yields good results for both $p, q > 0$. These initial candidate solutions are brute-forced and the best one is chosen by evaluating the three metrics to minimize: Mean Absolute Error (MAE), Root Mean-Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). We apply the Box-Jenkins methodology to improve the model's initial parameters further. This constitutes plotting the training residuals of the model obtained during the fit and applying the Ljung-Box statistical test to see if they can be considered white noise. Finally, the ACF and PACF plots of the residuals can confirm the results of the Ljung-Box statistical test. This model diagnostic process is repeated iteratively until better parameters are found. For a closer look at the data's ACF and PACF plots, please refer to Appendix C.

C. SARIMA

The second model used in this paper is a SARIMA$(p, d, q)(P, D, Q, m)$ model which builds up on the previous ARIMA model by integrating a seasonal component (P, D, Q, m). It is worth noting that, unlike ARIMA, SARIMA is not a linear model.

$$\begin{aligned} y_t = c + \phi_1 y_{d,t-1} + ... + \phi_p y_{d,t-p} + \Phi_1 y_{D,t-m} \\ + ... + \Phi_P y_{D,t-Pm} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \\ \Theta_1 \epsilon_{t-m} + ... + \Theta_Q \epsilon_{t-Qm} \end{aligned} \tag{5}$$

Φ and Θ represent the P, Q coefficients of the seasonal auto-regressive and moving average models respectively. $y_{D,t-Pm}$ represents the P past values with period m and differenced with seasonal order D. ϵ_{t-Qm} represents the Q past errors with period m. This model was tuned following the same methodology as a standard ARIMA model. Indeed by performing stationarity tests (ADF and KPSS) on the seasonal component, extracted from the initial time series data, we can set our D seasonal differencing parameter. D is set to 0 for all our models as the ADF and KPSS statistical tests show that the seasonal component is stationary without any differencing required. Then by plotting the ACF and PACF of the seasonal component one can determine candidate solutions (P, D, Q, m), and validate them using the Box-Jenkins methodology and split cross-validation accuracy errors. Here m constitutes the period of the seasonal component, it can be inspected and determined in its ACF and oftentimes corresponds to the sampling of our data (7 for daily, 12 for monthly ... etc). ACF shows a repeating pattern of 6 for the art index, as such SARIMA models tuned on art data have a period of $m = 6$. All other SARIMA models tuned on wine and watch have a period of $m = 12$.

D. (S)ARIMAX

Up until now, (S)ARIMA's model capabilities only allowed for univariate analysis and forecast of our time series data. This was done by using lagged values of the time series itself as regressors. The third model used in this paper: (S)ARIMAX allows one to factor in exogenous variables to help improve the forecast accuracy by examining if the predicted variable can be explained by an external variable. This multivariate model has the same parameters as ARIMA or SARIMA if seasonality is directly taken into account, and as such is tuned in the same way as explained previously for our (S)ARIMA models.

$$\begin{aligned} y_t = c + \phi_1 y_{d,t-1} + ... + \phi_p y_{d,t-p} + \beta X_t \\ + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} \end{aligned} \tag{6}$$

$$
\begin{aligned}
y_t = c + \phi_1 y_{d,t-1} + \ldots + \phi_p y_{d,t-p} + \Phi_1 y_{D,t-m} \\
+ \ldots + \Phi_P y_{D,t-Pm} + \beta X_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} + \\
\Theta_1 \epsilon_{t-m} + \ldots + \Theta_Q \epsilon_{t-Qm} \quad (7)
\end{aligned}
$$

The first equation is the mathematical representation of an ARIMAX model, and the second is a SARIMAX model. Both models have the same equation as their univariate representations, except for the added βX_t term. β is a learned coefficient over the training data through maximum likelihood estimation. If multiple exogenous variables are added then β becomes a vector. X_t represents the value of the exogenous variable at time t. Thus both the predicted variable and exogenous variable are in a one-to-one correspondence and must have the same length. Including exogenous variables when appropriately selected can increase the predictive power and thus the overall forecast accuracy, however, while (S)ARIMAX's strength lies in its validation accuracy on a test set, it includes further bias when used for forecasting. The reason is that to forecast our predicted variable, we need the same values for our exogenous variable since both are in a one-to-one correspondence. As such to forecast the predicted variable, the selected exogenous variable was first forecasted using a rolling-window mean technique which will be explained later. All model accuracies, as well as the selection process of the exogenous variables based on their correlation with the asset classes, will be detailed in the next section.

E. ARIMA Decomposition forecasting recombination strategy

The traditional way of forecasting would be to give the preprocessed data to the ARIMA model as input as it already accounts directly for trend and indirectly for seasonality. However, it is also possible to seasonally decompose the time series into several components: the trend, the seasonality, and the residuals, the stochastic component left after accounting for the two previous components. Since those represent the original time series data we are trying to forecast, it is possible to forecast the trend, seasonality, and residuals individually, and recombine the predicted values by summing them to get the overall forecast in the original time series scale. To inspect the seasonal decomposition of the watch and art index, please refer to Appendix E. Here is how each component is forecasted individually. The residuals are forecasted using a tuned ARIMA(p, d, q) process using both the ACF and PACF plots as well as the Box-Jenkins methodology described before. The seasonality is simply replicated, as it is a period of values (usually of 12 for monthly data) that repeats itself throughout the data. The trend is forecasted using, the mean of all observations on the test set (last 20% of the data). Another possible choice is a rolling-window mean method which includes forecasted points in the mean computations over time. The second method captures evolving trends better, while allowing the output of the forecast to be less static, and is thus given more variability. However, it may also introduce additional bias if the included forecasted values over time are inaccurate in the overall mean calculation.

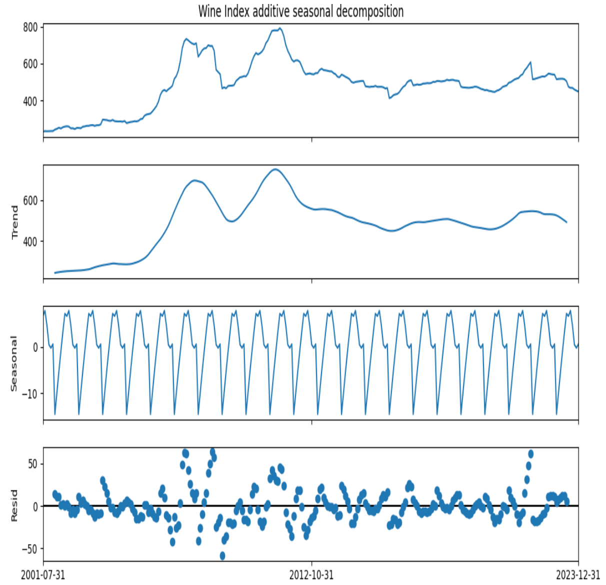

Figure 1: Seasonal decomposition of wine index into trend, seasonality, and residuals

This decomposition forecasting recombination strategy has the aim of improving accuracy by decomposing the time series. ARIMA is well suited to capture the autocorrelation left in the residuals, representing the data's stochastic component. Residuals can become difficult to predict when blended back into the original time series, hence why decomposition is essential.

F. ARIMA Rolling Window Strategy

A third forecasting and training strategy to improve accuracy using ARIMA was implemented under a common "rolling-window" technique. The optimal (p, d, q) model parameters are first determined and validated using the Box-Jenkins methodology and the ACF and PACF plots, as explained previously. The model then trains regularly on the training data but only up to a specified length: the window size. Finally, the model forecasts the next points, but only up to the specified window size length. In the next iteration, the forecasted points are added to the training data, and the model retrains and forecasts new additional points up to the window size, this process repeats until all summed-up window sizes correspond to the final forecasting length.

Retraining the model every time makes it easier to capture evolving trends, and factor in recent data. The goal is to progressively discard very early data points that have a minimal impact and predictive power on our more recent forecasted values. By doing so, the aim is to improve the forecast accuracy further.

IV. Research Results

A. Correlated Variables

This first research determines which exogenous variable is most correlated with each index, and thus the most appropriate variable to include in the multivariate (S)ARIMAX model. To that end, very common and popular variables representing the most important aspects of the US economy have been retrieved:

- The CPI index whose purpose is to measure the inflation of the US dollar.
- Crude oil is considered the number 1 commodity in the world, being the primary source of energy production today in 2024.
- Gold prices which still is the main hedge against inflation and has been already for several centuries.
- S&P 500 index containing the top 500 US companies by market capitalization.
- S&P US national home price index which tracks changes in value of single-family real estate across the US. In simpler terms, it measures how much home prices vary in the US.
- US dollar (USDX) index which measures the value of the US dollar relative to a basket of six major foreign currencies (Euro, Japanese Yen, British Pound, Canadian dollar, Swedish Krona and Swiss Franc).

A covariance calculation as well as a Pearson correlation coefficient are computed between the external variable examined and all three asset indices. Prior to this, all variables are log-transformed to stabilize the variance and get more accurate results. Furthermore, the significance of each correlation coefficient is examined using a statistical t-test to see if the correlation coefficient $p \in [-1, 1]$ indeed tells us the direction of the relationship between both variables (Simion et al., 2015). 1 being a perfectly positive relationship and -1 a perfectly negative relationship. S&P US National home price index has been abbreviated to NHPI for easier formatting purposes and results have been rounded at 10^{-4}:

Table II: Wine Correlation Results

Name:	Gold	S&P 500	CPI	Oil	NHPI	**USDX**
Cov	0.1205	0.0239	0.0246	0.0791	0.0228	**-0.0461**
p	0.8213	0.2080	0.5605	0.6661	0.2988	**-0.8716**

Table III: Watch Correlation Results

Name:	Gold	S&P 500	CPI	Oil	**NHPI**	USDX
Cov	0.0091	0.0478	0.0161	0.0039	**0.0444**	-0.0067
p	0.1778	0.4844	0.5823	0.0415	**0.7159**	-0.4212

Table IV: Art Correlation Results

Name:	Gold	S&P 500	**CPI**	Oil	NHPI	USDX
Cov	0.1747	0.7569	**0.4324**	0.0742	0.3111	-0.4959
p	0.3345	0.9065	**0.9671**	0.1371	0.9007	-0.9592

Examining the results for the wine index, it is possible to see that gold with a correlation coefficient $p = 0.8213$, and the US dollar index with $p = -0.8716$, stand out from the rest of the variables. With a p-value extracted from the student t-test $p_{value} < 0.05$, both have a significant correlation with the wine index. That means that as the wine index increases positively, gold also increases in the same direction, while the US dollar index decreases. Looking at the covariance results: $|$Cov(Wine,Gold)$| > |$Cov(Wine,USDX)$|$ and seeing how close the covariance values are, both variables are almost equally correlated with the wine index. Thus it is ultimately the cross-validation accuracy that indicates which variable is suitable for our model. With all three Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) error metrics slightly inferior: $54.8510 < 58.2188, 63.2706 < 68.5528, 8.8703 < 9.4446$ the US dollar index is the most appropriate exogenous variable for the wine multivariate model.

Examining the results for the watch index, it is possible to see that only the S&P national home price index (NHPI) stands out with the highest correlation coefficient $p = 0.7159$. Similarly, it also has one of the highest covariance with Cov(Watch, NHPI) $= 0.0444$. Additionally, with a p-value retrieved from the statistical t-test, $p_{value} = 0.0 < 0.05$ the correlation is significant and tells us the direction. Having a positive correlation, this once again signifies that as the NHPI increases positively, so does the watch index. Thus the S&P national home price index is the most appropriate exogenous variable for the watch multivariate model.

Examining the results for the art index, it is possible to see that four variables stand out with high correlation coefficients: S&P 500 with $p = 0.9065$, CPI with $p = 0.9671$, NHPI with $p = 0.9007$ and USDX with $p = -0.9592$. Additionally, with all p-values $p_{value} < 0.05$, all have a significant correlation which tells us the direction of the relationship. Looking at covariance, it's the S&P 500 which has the highest value: Cov(Art,S&P 500) $= 0.7569$ however the cross-validation accuracy error is lowest when integrating CPI as our exogenous variable, looking at the MAPE rounded at 10^{-2}: $\text{CPI} < \text{SP500} < \text{NHPI} < \text{USDX}$. With $\text{CPI} = 11.81, \text{SP500} = 16.81, \text{NHPI} = 21.24$ and $\text{USDX} = 26.35$. Thus the CPI is the most appropriate exogenous variable for the art multivariate model.

B. Univariate and multivariate model performance

This second research assesses the accuracy of all optimal (S)ARIMA(X) models. For each asset class, models are compared to see which one exhibits the best accuracy.

The data was divided into an 80/20 split for training and testing the models on all three asset classes: wine, watches, and art using 4-split-cross validation. The 4 splits are respectively 50, 65, 80, and 100 percent of the data and then the 80/20 ratio within those splits (for train and test) is applied. This allows the model to be evaluated on more than one test set and reduces overfitting or performance likely due to chance. Additionally, it makes sure the temporal dependence of the time series is kept by always having the test set after the training set and discarding the traditional shuffling seen in regular cross-validation. The Mean-Absolute Error (MAE), Root-Mean-Squared Error (RMSE), and Mean-Absolute Percentage Error (MAPE) were used to evaluate the models on the test set. Finally, two baselines were created the first one being only the first data point of the test set or in other words, a naive random walk model and the second being the mean of all data points in the test set.

Tables V, VI, and VII constitute the results of all models for each asset class using the traditional ARIMA forecasting strategy mentioned previously which applies to both uni and multivariate models, rounded at 10^{-2}. Naive Random Walk has been abbreviated to NRW for easier formatting purposes:

Table V: Wine Models Strategy 1 Accuracy Results

Name:	Model:	MAE	RMSE	MAPE
ARIMA	**(3,1,3)**	**45.83**	**53.95**	**8.36**
SARIMA	(3,1,3)*(3,0,6,12)	57.59	70.57	9.82
ARIMAX	(3,1,3)	54.85	63.27	8.87
SARIMAX	(3,1,3)*(3,0,6,12)	46.16	54.64	8.22
NRW	c	48.14	58.04	8.77
Mean	μ	38.24	44.43	6.83

Table VI: Watch Models Strategy 1 Accuracy Results

Name:	Model:	MAE	RMSE	MAPE
ARIMA	(2,1,3)	2867.32	3556.40	11.26
SARIMA	(2,1,3)*(1,0,3,12)	2436.01	3141.69	9.80
ARIMAX	(2,1,3)	2196.66	2806.28	9.18
SARIMAX	**(2,1,3)*(1,0,3,12)**	**2153.33**	**2624.06**	**9.01**
NRW	c	3079.21	3774.65	11.79
Mean	μ	1856.43	2195.21	7.82

Table VII: Art Models Strategy 1 Accuracy Results

Name:	Model:	MAE	RMSE	MAPE
ARIMA	(6,1,8)	7344.03	9773.55	15.03
SARIMA	**(4,1,2)*(5,0,6,6)**	**4602.23**	**6601.81**	**9.81**
ARIMAX	(6,1,8)	5636.39	7434.17	11.81
SARIMAX	(4,1,2)*(5,0,6,6)	6137.25	8410.04	12.02
NRW	c	9641.75	12213.92	18.33
Mean	μ	6445.44	8148.34	14.29

Examining Table V, it is possible to see the mean μ stands out from the rest of the models by having the lowest evaluation metrics. The best (S)ARIMA(X) models perform better than the naive random walk by a small margin namely: ARIMA$(3,1,3)$ and SARIMAX$(3,1,3) * (3,0,6,12)$. ARIMA is the best one with slightly lower evaluation metrics MAE $45.83 < 46.16$ and RMSE $53.95 < 54.64$. A strong mean μ accuracy suggests that using a straight line is better to fit the data, this can also be seen in the wine index plot in appendix B. ARIMA$(3,1,3)$ has slightly better evaluation metrics even though it is less complex, as it can approximate a straight line better than its seasonal and exogenous variants. Although the difference in accuracy between the latter is almost negligible.

Inspecting Table VI, it is possible to see the mean μ still has the lowest evaluation metrics. All (S)ARIMA(X) models perform better than the naive random walk, with SARIMAX$(2,1,3) * (1,0,3,12)$ being the best one amongst them. The simple ARIMA$(2,1,3)$ model, even though it is better at approximating a straight line than its seasonal and exogenous variants has the worst accuracy (excluding the naive random walk). When looking at the watch index plot in Appendix B, a skewed fat-tail is present towards the end of the plot. This makes it difficult for any (S)ARIMA(X) model to accurately predict this sudden increase and decrease based on the previous data. This also explains why the mean μ still performs the best on this type of time series data when compared to (S)ARIMA(X) models.

Evaluating table VII, the model with the highest accuracy is the SARIMA$(4,1,2) * (5,0,6,6)$ model. Looking at the art index plot, even though skewed, the absence of fat-tails and the gradual increase make it easier for (S)ARIMA(X) models to accurately predict future data based on past data. This also explains why this model is better than its baselines. Additionally, when evaluating the seasonal decomposition of the art index, the scale of the decomposed seasonal signal compared to the scale of the original time series indicates an underlying seasonal pattern. In comparison, the small scale of the seasonal signals of the wine and watch index qualifies them as pure noise when compared to the scale of their original time series. This also explains why SARIMA yields the highest accuracy on this art data, as it directly accounts for seasonality.

A rolling window forecasting strategy has the aim of increasing model accuracy. To avoid bias, a realistic window size w situates itself $w \in [12, len(data)/3]$. $w < 12$ results in unrealistic accuracy as the test set becomes too small (shorter than a year) and thus easier to predict instead of reflecting better models. The same problem can be found with $w > len(data)/3$, the training data increases, and as such the testing data decreases, and with a smaller test set, all models have increased (biased) accuracy instead of being better with critical bias reached at $w > len(data)/2$. The following plots the wine univariate model's MAPE accuracy metric evolution with the window size increase in the appropriate interval:

The best window size for a rolling forecast is one that minimizes the MAPE error metric. In the above plot, a window size of 66 marks a sudden drop in the MAPE error metric to its lowest value. This is comparable with the early window values, except the model performs better than the naive random walk and gets closer to the mean's accuracy.

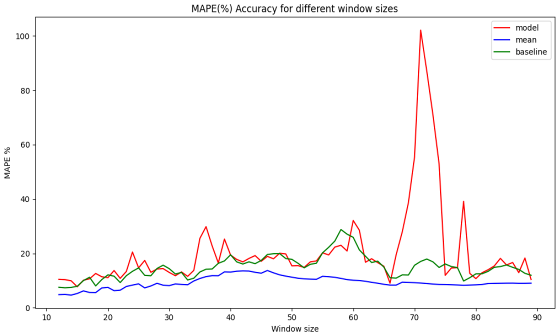

Figure 2: Wine MAPE in relation to window sizes

Knowing the optimal window size, we can compute strategy 3's accuracies; table XIII summarizes the model and baselines' accuracies for both strategy 2 and 3 rounded at 10^{-2}. Naive Random Walk has been abbreviated to NRW for easier formatting purposes:

Table VIII: Wine Models Strategy 2 and 3 Accuracy Results

Name:	Model:	Strategy	MAE	RMSE	MAPE
ARIMA	**(4,0,1)**	**2**	**38.39**	**43.36**	**6.77**
NRW	c	2	45.43	54.73	8.10
Mean	μ	2	37.65	43.23	6.65
ARIMA	(3,1,3)	3	53.65	65.09	9.03
NRW	c	3	65.41	79.24	11.06
Mean	μ	3	46.24	55.44	8.36

Examining table VIII, the best model amongst strategies 2 and 3 is the ARIMA$(4, 0, 1)$ model using the decomposition-forecasting-recombination strategy. Furthermore, this model has the lowest evaluation metrics compared to table V wine models. It is thus the most suitable to forecast the wine index. However, the mean from strategy 2 still has some slightly better accuracy metrics, highlighting the fact that it is very complicated for (S)ARIMA(X) models to perform better than a simple mean on this wine index.

The following plots the watch univariate model's MAPE accuracy metric evolution with the window size increase in the appropriate interval:

In the above plot, the window size that minimizes the MAPE error metric is clearly amongst the early values, as all MAPE errors from all models follow an increasing trend in relation to the window size. Although at window size 39, the model becomes better than both baselines, its accuracy is still worse than the window size of 16, as such the latter will be retained as the optimal window size.

Knowing the optimal window size, we can compute strategy 3's accuracies; table IX summarizes the model and baselines' accuracies for both strategy 2 and 3 rounded at 10^{-2}:

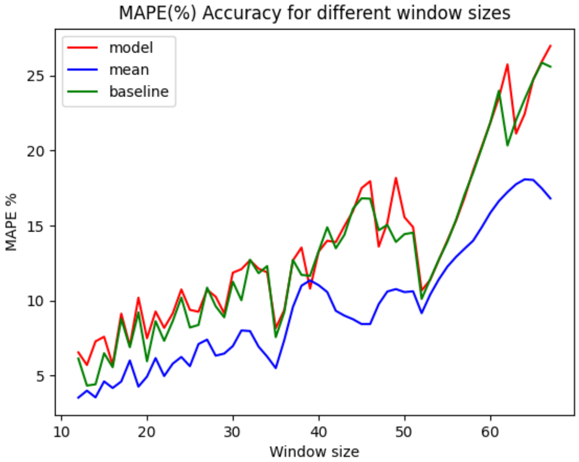

Figure 3: Watch MAPE in relation to window sizes

Table IX: Watch Models Strategy 2 and 3 Accuracy Results

Name:	Model:	Strategy	MAE	RMSE	MAPE
ARIMA	(2,0,0)	2	2001.69	2309.59	8.67
NRW	c	2	3501.12	4188.62	14.07
Mean	μ	2	1970.76	2265.66	8.48
ARIMA	**(2,1,3)**	**3**	**1180.82**	**1510.94**	**5.72**
NRW	c	3	1170.25	1552.19	5.56
Mean	μ	3	901.65	1072.50	4.18

Evaluating table IX, the model with the lowest evaluation metrics is ARIMA$(2, 1, 3)$ from strategy 3 with a window size $w = 16$. Additionally, it also performs better than any model using strategy 1 from Table VI. It is thus the most suitable model to forecast the watch index. However, this rolling-window ARIMA model cannot beat baselines in its strategy. This underlines once again the fact that a simple rolling window mean or rolling window naive random walk contains a lot of predictive power on this watch index.

The following plots the art univariate model's MAPE accuracy metric evolution with the window size increase in the appropriate interval:

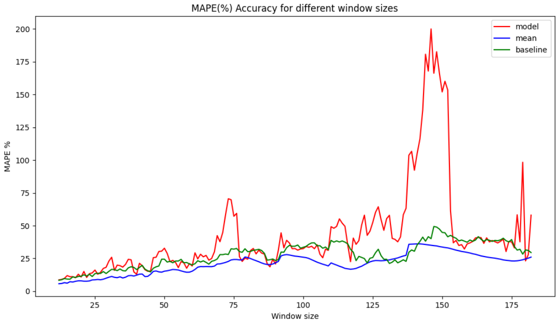

Figure 4: Art MAPE in relation to window sizes

The window size which minimizes the MAPE error metric in the above plot is amongst the early values. Although the increasing trend isn't as definite in comparison to the watch index, there is still a slight increase in the MAPE error metric in relation to the window size. Thus the appropriate window size is 12.

Knowing the optimal window size, we can compute strategy 3's accuracies; table X summarizes the model and baselines' accuracies for both strategy 2 and 3 rounded at 10^{-2}:

Table X: Art Models Strategy 2 and 3 Accuracy Results

Name:	Model:	Strategy	MAE	RMSE	MAPE
ARIMA	(6,0,10)	2	6208.27	7953.83	14.59
NRW	c	2	6860.53	9463.54	15.62
Mean	μ	2	6015.96	7744.26	13.68
ARIMA	**(13,1,6)**	**3**	**2695.52**	**3303.00**	**7.65**
NRW	c	3	2819.95	3550.15	8.34
Mean	μ	3	1984.86	2483.66	5.72

Inspecting table X, the best performing model is ARIMA$(13, 1, 6)$ from strategy 3 using a window size $w = 12$. This model has lower evaluation metrics than all models from strategy 2 as well as strategy 1 from table VII. It is thus the most suitable to forecast the art index. Although it performs better than the naive random walk in its strategy it still has less predictive power than the mean from strategy 3.

C. Forecast

This research builds on previous results by applying the best model for each asset class and evaluating their forecasts over long periods, respectively being 10+ years depending on the length of the training data. Note that emphasis is put on long-term forecasts and their trends, as short-term forecasts are challenging, volatile, and oftentimes wrong due to market uncertainty even for professionals.

Here are the long-term (which already incorporate the short and medium-time spans) forecasts of the best model for each asset:

The overlap between the blue observed data and the red forecast is the test set, containing the last 20% of the data in the final split of the split cross-validation. All indices contain sudden spikes, called fat tails. As plotted in figures 5,7 with an emphasis on figure 6, (S)ARIMA(X) models have trouble identifying and forecasting fat-tails. A straight line signifies the models have trouble detecting repeating patterns (seasonality) in the data, as well as serial correlation and trend. As such, the models display straight lines in their forecasts. This is not necessarily a negative sign. It is not uncommon for simple techniques like mean or naive-random walks to outperform complex techniques like ARIMA models as shown by empirical evidence by Makridakis and Hibon, 2000, in the M3-Competition and by Green and Armstrong, 2015 from 92 comparisons in 32 papers. This is also validated by our collectible asset data, where the means in their respective strategies perform better than almost all the best (S)ARIMA(X) models. As such, having a nearly flat forecast yields the best reliability for our out-of-sample forecasts per the validation accuracy results.

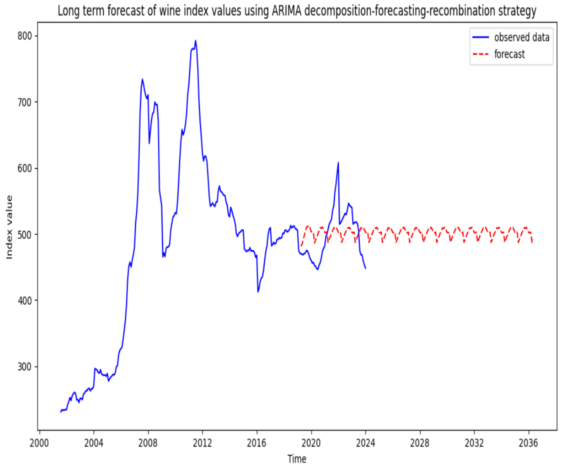

Figure 5: Wine ARIMA Decomposition-Recombination Forecast

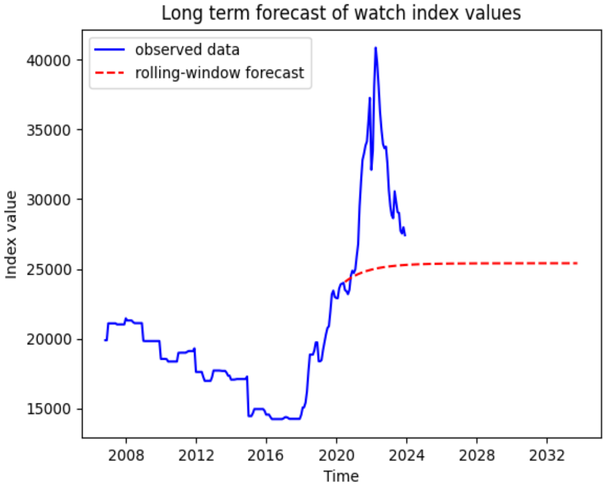

Figure 6: Watch ARIMA Rolling-Window Forecast

V. Conclusion and Future Work

The variables most correlated with each asset class are:

- the US dollar index (USDX) with the wine index.
- the S&P US national home price index (NHPI) with the watch index.
- the consumer price index (CPI) with the art index.

Research findings have demonstrated that these exogenous variables yield the best split cross-validation accuracy for the multivariate ARIMA models. However, multivariate ARIMA models do not yield the best results when compared with univariate ARIMA models.

The best (S)ARIMA(X) models for each asset class are:

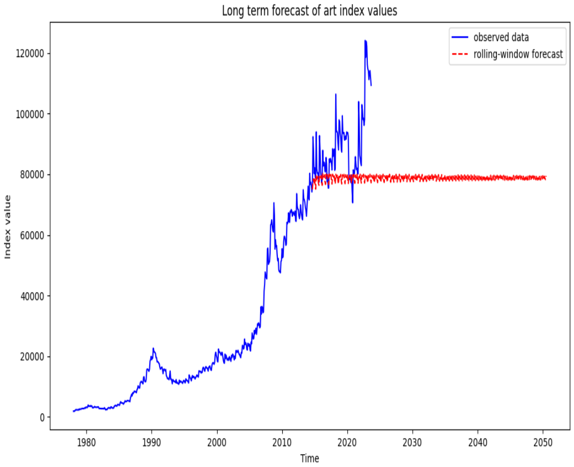

Figure 7: Art ARIMA Rolling-Window Forecast

- the ARIMA$(4, 0, 1)$ from the decomposition-forecasting-recombination strategy (strategy 2) with the wine index.
- the ARIMA$(2, 1, 3)$ from the rolling-window strategy (strategy 3) with the watch index.
- the ARIMA$(13, 1, 6)$ from the rolling-window strategy (strategy 3) with the art index.

Split cross-validation accuracy has shown that these ARIMA models yield the best results when compared to all ARIMA models from all three strategies. With a 90% improvement from the worst to the best performing ARIMA models for each asset class. Although the best ARIMA wine and art models outperform the naive random walk baseline by 16% and 8% respectively, they cannot get lower accuracies than their respective means. This highlights the predictive power of simple models on this complex time series data. ARIMA models are considered complex. As empirical evidence shows, complexity can increase forecast error by an average of 27% (Green & Armstrong, 2015). Furthermore, simple models like the arithmetic mean can increase understanding and reduce the likelihood of errors (Green & Armstrong, 2015). As such (S)ARIMA(X) outputs reflecting straight lines yield the best reliable forecast among all models in this paper considering the complexity of the data.

Further models could be tested to improve accuracy and forecast reliability in the collectibles assets market. Knowing simple techniques often yield the best results on complex time series data, exponential smoothing models should be explored. Additionally, although the complexity level increases, it is still worth examining the modeling of hybrid ARIMA models with deep learning. By using ARIMA to predict the linear component of the time series, deep-learning models to predict the non-linear component, and finally modeling the appropriate relationship between both components to get the final forecast.

References

Aubry, M., Kraeussl, R., Manso, G., Spaenjers, C., et al. (2019). Machines and masterpieces: Predicting prices in the art auction market. *Journal of Finance, Forthcoming*.

Christoffersen, P. (2011). *Elements of financial risk management*. Academic press.

Commandeur, J. J., & Koopman, S. J. (2007). *An introduction to state space time series analysis*. Oxford University Press, USA.

Dama, F., & Sinoquet, C. (2021). Time series analysis and modeling to forecast: A survey. *arXiv preprint arXiv:2104.00164*.

Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, *68*(8), 1678–1685.

Hernandez-Matamoros, A., Fujita, H., Hayashi, T., & Perez-Meana, H. (2020). Forecasting of COVID19 per regions using ARIMA models and polynomial functions. *Applied soft computing*, *96*, 106610.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Keene, O. N. (1995). The log transformation is special. *Statistics in medicine*, *14*(8), 811–819.

Makridakis, S., & Hibon, M. (2000). The m3-competition: Results, conclusions and implications. *International journal of forecasting*, *16*(4), 451–476.

Mendez-Carbajo, D. (2023). Adjusting for inflation. *Federal Reserve Bank of St. Louis Page One Economics®*.

Ozturk, S., & Ozturk, F. (2018). Forecasting energy consumption of Turkey by Arima model. *Journal of Asian Scientific Research*, *8*(2), 52.

Penasse, J. (2014). Real-time forecasts of auction prices. https://ssrn.com/abstract=2505653

Petrică, A.-C., Stancu, S., & Tindeche, A. (2016). Limitation of ARIMA models in financial and monetary economics. *Theoretical & Applied Economics*, *23*(4).

Pires, P. M. G. (2020). *A data-driven approach to predict the value and key features of collectible cars* [Master's thesis, Iscte, University Institute of Lisbon].

Shumway, R. H., Stoffer, D. S., Shumway, R. H., & Stoffer, D. S. (2017). ARIMA models. *Time series analysis and its applications: with R examples*, 75–163.

Simion, D., Stanciu, M., & Armăşelu, S. (2015). Correlation analysis between structure financial system and economic growth in Romania. *Procedia Economics and Finance*, *32*, 1332–1341.

Yang, B., Li, C., Li, M., Pan, K., & Wang, D. (2016). Application of ARIMA model in the prediction of the gross domestic product. *2016 6th international conference on mechatronics, computer and education informationization (MCEI 2016)*, 1258–1262.

Appendix A
What is the CPI Index ?

The Consumer Price Index (CPI) is a metric of inflation, tracking changes in the prices of a representative basket of goods and services over time. This basket reflects typical consumer purchases, including food, housing, transportation, and healthcare. By comparing CPI values from different periods, we can estimate how much more expensive it has become to live our daily lives. Inflation reduces the purchasing power of money, meaning a dollar today buys less than a dollar yesterday. To account for inflation and provide more accurate computations, financial data, wages, and even government benefits are often adjusted using the CPI. Published monthly by the U.S Bureau of Labor Statistics for the U.S. dollar, the CPI is a widely used tool by economists and policymakers to understand how the inflation of our fiat currencies impacts the world.

Appendix B
Visualizing each asset adjusted from inflation

Here are all three asset classes in USD and adjusted from inflation:

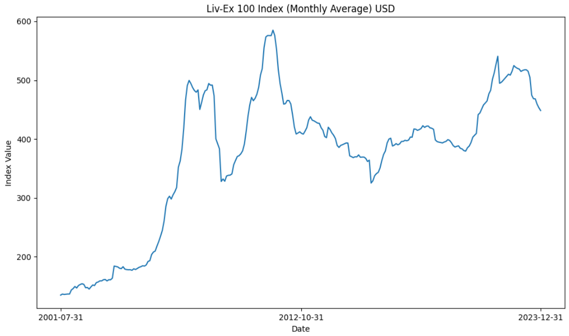

Figure 8: Wine Liv-Ex 100 Index adjusted from inflation

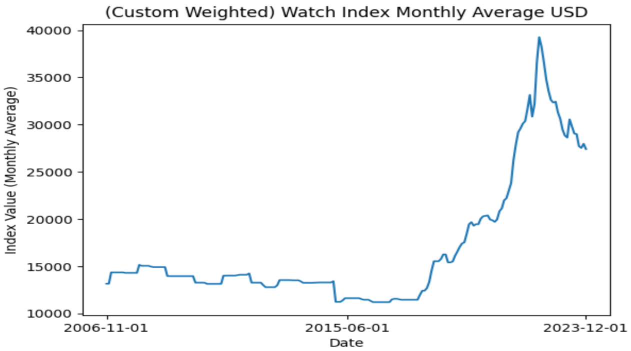

Figure 9: Watch Index adjusted from inflation

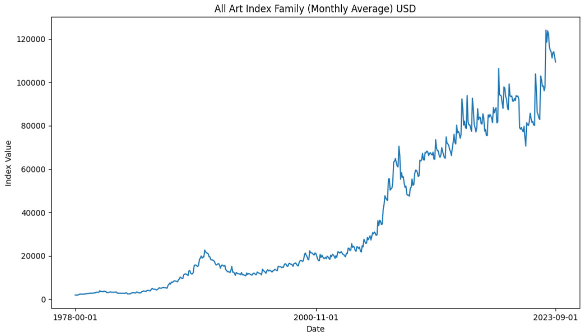

Figure 10: All Art Index adjusted from inflation

Appendix C
ACF and PACF plots of the data

Here are all Auto-Correlation and Partial Auto-Correlation Function plots of the data:

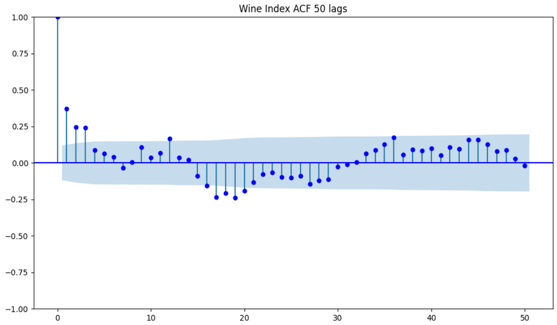

Figure 11: Wine ACF Lag 50

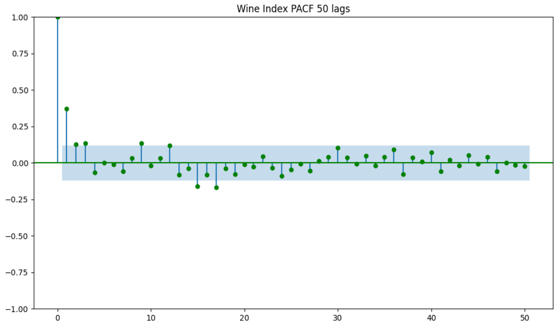

Figure 12: Wine PACF Lag 50

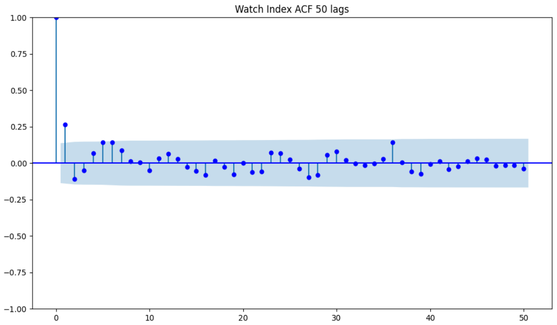

Figure 13: Watch ACF Lag 50

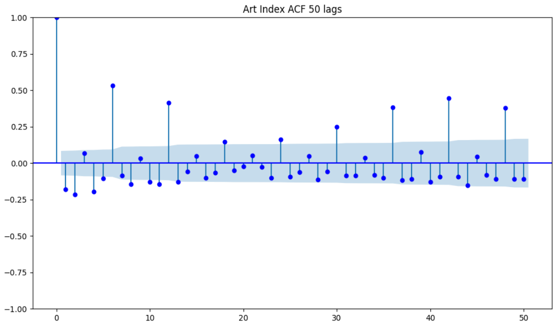

Figure 15: Art ACF Lag 50

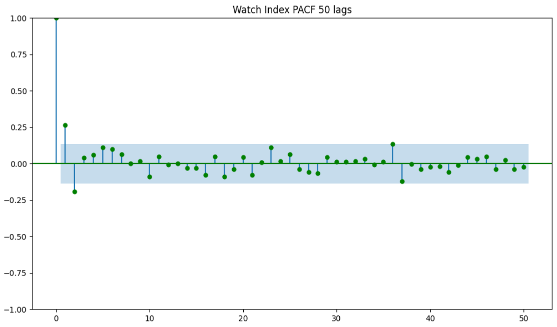

Figure 14: Watch PACF Lag 50

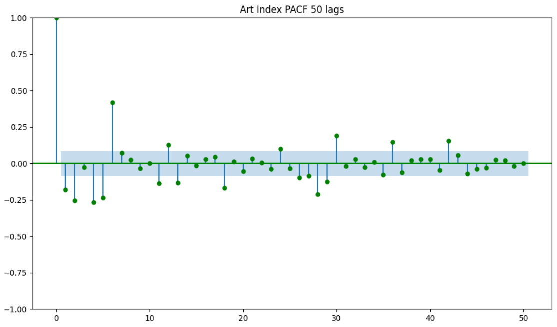

Figure 16: Art PACF Lag 50

Appendix D
Log-Transformation

Keene, 1995 mentions that: “the use of t-tests, analysis of variance, and analysis of covariance for continuous positive data on an interval scale is widespread. One of the easiest modifications to these simple parametric methods is the prior use of a log transformation.”If the mean increases with the standard deviation (heteroscedasticity), this is indicative of the need for a log transformation. Log transformation allows an analysis on a ratio scale (%) instead of the original scale.Keene, 1995 “takes the example of variables such as biochemical measurements typically show a skewed distribution,’ which can often be made symmetric using a log transformation. It has been argued that ’the theoretical justification for using this transformation for most scientific observations is probably better than that for using no transformation at all’ ”. Thus finally Keene, 1995 recommends that for “continuous positive data measured on an interval scale, a log-transformed analysis should frequently be preferred to an untransformed analysis ”. And indeed log-transforming the data before training the models stabilizes the variance and minimizes both the AIC (goodness of fit) and BIC (model complexity) criterion. However, it worsens the test set accuracy averages calculated with split cross-validation. This is likely because the data is complex and volatile, log-transforming the data results in less volatility in the model’s output: thereby indirectly reducing the accuracy. Additionally, the exponential transformation used to invert the scales back after forecasting amplifies the skewed trend picked up by the model during training even with the log-transformation applied: thus it increases the error by having an abnormal skewed trend in our models’ outputs. Finally, Dama and Sinoquet, 2021 suggests that the log-transformation is only appropriate for time series decomposition when the series exhibits a multiplicative trend and seasonality, which is not compatible with ARIMA which is an additive trend model only. As such the log-transformation was discarded in model training and testing.

APPENDIX E
WATCH AND ART INDEX SEASONAL DECOMPOSITION

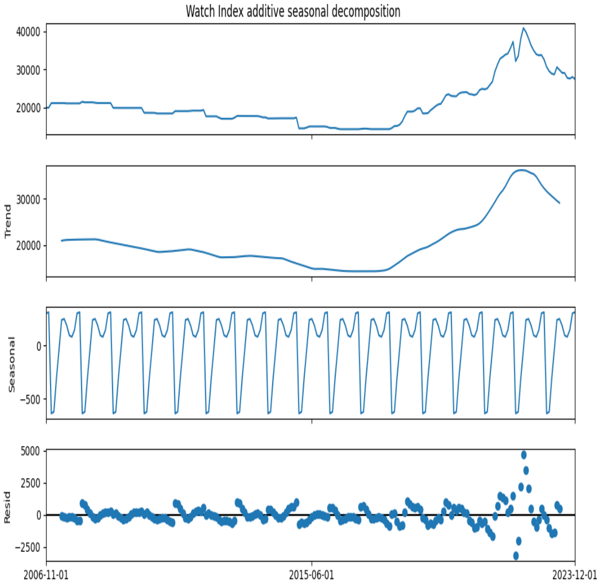

Figure 17: Seasonal decomposition of watch index into trend, seasonality, and residuals

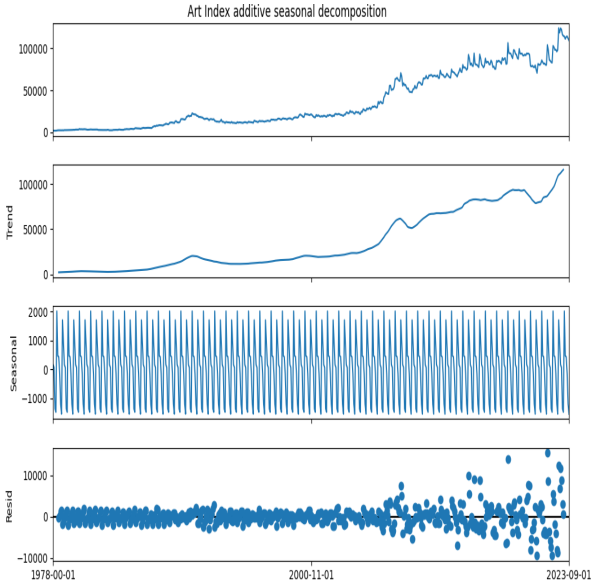

Figure 18: Seasonal decomposition of art index into trend, seasonality, and residuals