Modeling Virtual Economies

Drew Kitik

Special Project Paper

Submitted in Fulfillment of the

Requirements for the Degree of

Master of Science in Data Science

Department of Mathematical Statistics

Central Connecticut State University

New Britain, Connecticut

Fall 2025

Capstone Advisor: Gurbakhshash Singh, PhD

Capstone Committee: Daniel T. Larose, PhD

Department of Mathematical Sciences

ABSTRACT

The project analyzes the virtual economy of *Old School RuneScape*, a popular *MMORPG* (massively multiplayer online role-playing game) as an academic laboratory using a custom data-scraping pipeline built from the official JAGEX & OSRS WIKI API documentation. Two complementary datasets were constructed: a 180-day macro series and a 90-day feature-rich high-frequency series spanning February 11th to August 10th, 2025.

Item-level price dynamics were modeled using a Factor-Augmented Vector Autoregression (FAVAR) framework alongside two recurrent neural network architectures, Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). Despite the common expectation of neural networks outperforming linear models, the VAR slightly outperformed both deep-learning models on mean-squared error (VAR MSE = 0.022 vs. 0.025 for LSTM and 0.028 for GRU). Adding trading volume as a feature provided only marginal improvements. Profit-based back-tests demonstrated that LSTM produced a modest positive cumulative return (+2.3 PnL units), while GRU effectively yielded zero.

Unsupervised learning via PCA and *k*-means clustering revealed strong latent market regime structure in the Rich daily data, with well-separated clusters (silhouette scores 0.80–0.88) and softer, overlapping item-level behavioral groups in the macro dataset. Qualitative clustering patterns appeared consistent with broad item behavior, though not

formally validated as economic sectors as no true sector labels exist to formally evaluate cluster purity.

Structural-break analysis using a Bai–Perron style test with mean-shift procedure (via *ruptures* library) identified statistically detectable regime shifts that occurred near major content updates such as “Yama: Master of Pacts” and “Summer Sweep-Up” patches. These findings provide evidence of event-driven shocks influencing item valuations in ways analogous to policy or news shocks in real financial markets.

Collectively, the results highlight how quantitative methods routinely used in real financial analysis; factor models, time-series forecasting, clustering and structural-break detection, can be effectively applied to a virtual game economy. The OSRS market exhibits qualities that reflect latent structures mirrored in real markets and demonstrate its value as a controlled laboratory for studying economic behavior.

Table of Contents

INTRODUCTION

In the last five years, demand and interest for specific investment equities such as stocks, gold, and cryptocurrencies (Niedens, 2025; Dubey, 2025) have surged to all-time highs, driven by several economic factors. Inflationary pressures, shifts in global trade policy and broader financial uncertainty have heightened attention in both the public and academic sectors regarding how these equity markets respond to changing economic conditions. The growing need for a better understanding of real-world equity price action makes the study of virtual economies an invaluable parallel. The most thriving virtual economies belong to what is known as massively multiplayer online roleplaying games

(*MMORPGs*). MMORPGs feature player-driven marketplaces where thousands of participants interact, trade, and compete within a closed but highly dynamic system, offering measurable data on supply, demand, and player behavior.

The MMORPG *Old School RuneScape* also known as *OSRS* (Jagex Ltd, 2013) features a centralized trading hub for its market known as the "Grand Exchange" (*GE*). The GE facilities trade of in-game items similarly to a stock market, using a pricing system that matches buyers and sellers through an exact bid/ask mechanism. Players trade in-game currency, gold coins (*GP*), in exchange for items that act as resources, equipment, supplies and cosmetics just like a real market. The trading hub also reflects real-world equities trading rules and restrictions with several implemented policies from the developers. This includes a 2% tax on items sold over a certain price (*100,000 GP*) and daily limits on low-value items as preventive measures towards monopolies and price manipulation. Additionally, players can purchase *membership bonds* normally sold through the company's website (*8.99$ per bond*) which can be redeemed for 14 days of premium membership or traded to other players in exchange for in-game gold (Old School RuneScape Wiki Community [OSRS Wiki], n.d.-a). This aspect gives players the option to offset the cost of premium membership fees by purchasing an item purely with in-game virtual currency. Since bonds are directly tied to a real-world price, the GP-to-bond ratio provides a natural exchange rate *($0.62 USD per million GP as of 18-Oct-25)*, making it possible to assign USD valuations to in-game items (OSRS Wiki, n.d.-a).

Just as traders in a real economy, players engage in speculation, arbitrage, hoarding, respond to scarcity, price & demand shocks/jumps and change trading behaviors based on upcoming changes through game updates and patches (like changes in policy or law). This virtual economy provides a ripe environment for academic study as it is a fully contained and observable market with explicit rules and transparent transaction data with few confounding variables or uncertainties. This makes the "*Grand Exchange*" of OSRS an ideal laboratory for investigating price formation, volatility, and the role of trading volume in driving both short and long-term economic outcomes.

The relevance of this study extends beyond evaluating item price movements. It seeks to deepen understanding of broader economic and technological concepts. Insights from modeling OSRS's virtual economy can shed light on consumer behavior and trading patterns in real markets, including those for virtual goods and decentralized exchanges such as cryptocurrency. As the boundary between physical and digital economies continues to blur, modeling a well-established virtual economy like OSRS provides both methodological contributions and practical implications for understanding equity markets.

Statement of Purpose

The purpose of this analysis is to model both prediction and explanation aspects of price dynamics for items traded within the OSRS Grand Exchange. The data were obtained using JAGEX's official API endpoints, accessed through RuneScape Wiki's community-documented resources, to develop an API data scraping tool in Python (RuneScape Wiki [RS Wiki], n.d.). Components and utility functions from the documented API were integrated into a unified pipeline to collect the data for this project. Details on the specific variables utilized in the tool can be found in 16.

Two complementary datasets were constructed. The first is a macro dataset containing 180 days of *daily* price data at 24-hour intervals, representing the maximum non-cached horizon available from the API (RS Wiki, n.d.). The second dataset is a feature-rich dataset consisting of *90** days of data aggregated from 6-hour intervals. This dataset includes high/low prices and corresponding high/low trade volumes. This enables more granular analysis of short-term volatility and price movements. For comparability, 6-hour intervals were aggregated (excluding entries with missing volumes) into 24-hour intervals as a rich daily dataset to match the macro dataset.

*(*Some outliers lie outside of this range due to liquidity and API fragmentation—Data range was extended slightly to cover low-frequency items while staying within the project's scope of 180 days)*.

The datasets were analyzed using a combination of econometric and machine learning models, including Factor-Augmented Vector Autoregression (FAVAR), recurrent neural network models (LSTM/GRU), Principal Component Analysis (PCA), clustering methods and Bai-Perron style structural break models.

Research Questions

The following research questions, formalized from the objectives derived in Phase 1, guide the primary tasks for modeling and evaluation in this study:

1) Which multivariate time-series model, linear-based Vector Autoregression (VAR) or neural networks, Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), more effectively forecasts item-level price movements?

2) Can unsupervised learning methods such as PCA and *k*-Means clustering reveal distinct behavioral “sectors” of items that share similar trading characteristics?
3) Do major announcements or content updates (introduction of new in-game content) create measurable structural breaks or price shocks in the market?

Literature Review

Virtual economies have been used as empirical laboratories for economic study due to their clearly defined marketplace rules, high transparency, observability, and availability of large datasets. Publications studying these virtual economies have closely followed the explosion of popularity in massive multiplayer online games (MMORPGs) since the early 2000s. One of the earliest and most influential contributions to this field, *On Virtual Economies* (Castronova, 2002) used a popular MMORPG *EverQuest* as a case study. Castronova argued that virtual economies should not be dismissed as trivial extensions of games but instead function as real economies where players display rational market behavior. His work quantified the real-world value of virtual goods by comparing in-game exchange rates with U.S. dollars, and highlighted phenomena such as inflation, labor value, and policy shocks within digital spaces. This established the legitimacy of virtual worlds as “natural experiments” for economists, where controlled rules and abundant data allow for testing theories of price formation, incentives, and welfare in ways difficult to achieve in real world economies.

Building on this foundation, Nazir and Lui (*A Brief History of Virtual Economy*, 2016) provided a comprehensive conceptual overview of virtual economies. The paper reviewed academic journals, articles and media publications from 2001 to 2015 for tracing developments across different games, social platforms, services and media publications. The core contribution of this publication was a form of taxonomy or framework for what makes up a virtual economy; marketplace design, exchange mechanisms and transaction modalities (Nazir et al., 2016). The publication utilized three large virtual economies as the basis for creating a generalized virtual economic framework and outlined specific intersections of the virtual economy to real-world currency activity. Citations sourced by the publication include the virtual economy of *Second Life* which contained virtual real estate and goods tied to real dollar evaluations and pricing. Additionally, *World of Warcraft (WOW)* experienced non-sanctioned "black market" third-party sites selling in-game gold, often facilitated through PayPal. These types of exchanges are commonly against the terms of service and are known as *real-money trading (RMT),* a common account bannable offense. While not the first review of its kind, this publication consolidated prior work into a coherent framework, clarifying how digital markets form, operate, and evolve when exposed to real-world currency flows.

Other publications focused on specific MMORPGs have demonstrated proven relevancy for using virtual economies for academic analysis. In *EVE Online*, researchers used several indicators such as the global peace/terrorist indexes, unemployment and exchange rates to quantify how relevant real world socioeconomic factors affected the virtual economy grouped by country (Belaza et. al, 2020). The results of the publication indicated that activities and player behavior (aggressiveness, economic risk tolerance, etc.*)* in virtual economies correlated and reflected real-world socio-economic activity.

These trends were also reflected in price shocks of virtual goods compared to actual economic supply scarcities or surpluses per country. This is an important aspect for analyzing the action within an economy as participants perceived feelings of scarcity or abundance can highly influence trading habits such as risk tolerance/reluctance and purchasing habits. This paper demonstrates how economic activity within a "simple game economy" can be the product of complex factors outside the game itself based on real-world phenomena acting on the player.

A publication on Korean MMORPG Aion Online (Chun et al., 2018) studied its virtual economy not only on economic activity and participation but the relationship between player social engagement, total playtime and wealth accumulation. The publication contains extensive modeling that reveals a distinct economic stratification of its player base population. Players with playtime well above the mean for the overall population consistently occupied what they classified as "upper class" in economic status. The researchers found that the upper-class players had higher playtimes, had either extremely low or extremely high social interaction, accumulated the most wealth and were most associated with RMT activity. The upper-class players had an RMT participation rate of over 50% while the overall RMT activity across the entire game (via an in-game exchange called 'the agency') was just over 10% (Chun et al., 2018). This finding of participants in a virtual economy mirrored real-world socioeconomic structure in which those with greater resources or greater time investment usually translate into higher economic returns. Additionally, it shows that the virtual economic activity within the game was not completely sealed off from external, real-world economic forces. Instead, they clearly demonstrated that the upper-class players actively participated with illicit practices (RMT) to either continue their edge within the game or profit off "black market" activities such as selling or laundering in-game gold. The results of the paper demonstrated how the economy of the game was not something "trivial" but rather a complex ecosystem subjected to social hierarchies, structural imbalances and other complexities that are observed in real economies.

A smaller paper from Cornell University (Hogan-Hennessy et al., 2022) studied OSRS specifically in which they examined the impact of two forms of market interventions introduced by the developer Jagex Studios after hiring an actual economist. The market interventions introduced including the GE tax at 1% (currently 2% in 2025) with the proceeds of in-game GP used to make an item "sink" by buying certain goods from players then removing them from circulation. Both mechanisms were introduced to combat inflationary pressures, stabilize prices, and reduce volatility. The analysis and modeling explored how the GE tax reduced high-frequency speculative trading (like equities scalping) while the item sinks affected the supply side of the economy for certain rare goods that were too common and flooding the market. This publication is specifically important as it highlights how a game developer company sought out to hire a real economist to apply market policy change to a virtual market with methods used in real economies. This literature not only validates as to why OSRS is a fertile ground for academic research but also provides a clear description of its market structure and background for topics that are discussed specifically in this paper.

While some of the studies referenced within this literature review are much larger in scale than this paper, they laid groundwork and interest for focused analysis and issues for specific virtual economies in an academic setting. For instance, Castronova’s work pointed toward inflation and policy controls, while game specific studies such as Chun et al. and Belaza et al. highlighted complex player behaviors, activity, and modalities in virtual economic environments. Together, these studies establish the academic legitimacy of virtual economies, map their evolution, highlight external influences such as RMT and policy interventions, and motivate the need for detailed modeling of OSRS’s Grand Exchange. This project seeks to extend the contribution of studying virtual economies by focusing on the price data of OSRS’s Grand Exchange through a dual dataset analysis approach at different levels of granularity to provide the most robust results for analysis.

Statement of Need

Most prior research focused on specific virtual economies have either attempted to bridge economic macro concepts to them, studied structure through market mechanics or player base, or the influence of external factors on the economy such as policy changes or real-world currency trading activity. Relatively few have attempted a ‘deep dive’ approach to directly map out supply and demand dynamics by tracking all tradable price item data, trends and relationships through econometric and machine learning techniques. This leaves a significant gap for applying methods such as VAR, LSTM networks, and clustering to better understand structural dynamics, predictive signals, and explaining market shocks and price changes. Addressing this gap is important for better rounding out the current academic understanding of the underlying functions of virtual economies.

This project demonstrates the adaptability and robustness of using econometric and machine learning techniques outside of the normal domain of traditional financial datasets to capture trends, lagged co-movements and volatility patterns in nontraditional context. Insights found within this paper can be used to the advantage of game developers to leverage insights on how to potentially improve market stability, better detect anomalies of transaction data indicating potential botting or real-world trading activity, and design better economic policies in virtual environments. Additionally, as digital economies and virtual assets continue to grow in popularity such as cryptocurrencies or in-game trading platforms, the lessons drawn from OSRS can inform thinking about real-world policy, virtual asset management, and the behavior of speculative markets.

In short, this project addresses an underexplored but increasingly relevant domain, where the intersection of econometrics, data science, and digital systems creates opportunities for both novel academic findings and practical applications.

PROJECT PROCEDURES

This analysis was performed following the Data Science Methodology (DSM), as proposed in *Data Science Using Python and R* (Larose & Larose, 2019). DSM is an adaptation of the Cross Industry Standard Practice for Data Mining (CRISP-DM), which is the most widely used analytics process standard.[1] It is a methodology that helps the analyst keep track of which phase is being performed (Larose & Larose, 2015). The methodology consists of seven phases:

[1] https://www.forbes.com/sites/metabrown/2015/07/29/what-it-needs-to-know-about-the-data-mining-process/#55190c21515f

1. Problem Understanding Phase
2. Data Preparation Phase
3. Exploratory Data Analysis (EDA) Phase
4. Setup Phase
5. Modeling Phase
6. Evaluation Phase
7. Deployment Phase

Each phase is an integral part of the analysis and are explained in detail as the phase is performed.

Phase 1: Problem Understanding Phase (Problem Definition)

This essential phase defines the framework in which the questions of this study are developed into analytical objectives and actionable data-science problems. The phase consists of two stages: (1) a clearly articulated project objective or question and (2) translating it into a well-defined data science problem. The primary objective of this project is to model and interpret the dynamics of Old School RuneScape's (OSRS) Grand Exchange, a player-driven virtual economy. The analysis focuses on three dimensions of market behavior: forecasting, classification, and event/news impact. The project seeks to achieve these market behaviors through the following objectives translated into data science problems:

1. *Investigate how prices change over time with multivariate forecasting, and which models perform the best.*

This objective is a time-series problem that involves modeling market dynamics through multivariate forecasting. Econometric methods such as Vector Autoregression

(VAR) and neural sequence models (LSTM/GRU) can be used to forecast item prices while integrating different explanatory features such as volume. Both models were evaluated using accuracy metrics such as RMSE and MAE to assess how well they capture item interactions and shock propagation through the economy. In addition to accuracy metrics, a profitability back test assessment was conducted to explore whether improved accuracy translates to potential profitability within in-game market contexts. Granger causality tests were used to identify directional influences between items or their groupings.

2. *Identify sectors of similar item behavior through clustering.*

This objective applies dimensionality reduction and clustering methods to classify items with similar price, volume or volatility patterns into behavioral "sectors." Unsupervised techniques were used, including Principal Component Analysis (PCA), and *k*-means clustering to extract latent factors that drive market changes and then cluster them into groupings of similar behavior. While PCA serves to reduce and group the predictors (price, volume, volatility) into principal components, clustering operates on the records themselves to group items with similar market behavior. Cluster quality was evaluated using several metrics, such as silhouette scores, to quantify the cohesion and separation of the resultant groupings.

3. *Analyze structural breaks and quantify the impact of events.*

The final objective is an event-based impact analysis to measure the effects of major content announcements or implementations (e.g., patch notes, announcements, updates). A Bai–Perron style linear structural break test (implemented via ruptures using an L2 mean-shift cost function) was applied to detect statistically significant shifts in each series. Event markers were then overlaid to assess whether detected breaks aligned with the largest announcements during the macro timeframe, illustrating how external interventions reshape virtual market structure.

Phase 2: Data Preparation Phase

Phase 2 focuses on preparing and structuring the data required for subsequent exploratory and modeling analysis. The emphasis of this phase is dataset construction, variable definition, and generation of features for modeling and analysis. Below are examples of variables from Jagex's official API endpoints, accessed through RuneScape Wiki(s) community-documented resources (RS Wiki, n.d.; OSRS Wiki n.d.).

- */mapping*: Returns metadata for each tradable item that exists in the game. This dataset includes the item ID, name, membership requirement, GE buy limits, high/low alchemy values (can be treated as in-game price floors) and others. This endpoint was used to build the item reference table and enrich price data with key contextual data and information.
- */graph/{item_id}*.json: This endpoint provides the entire available (180 days) historical *daily* price data (24-hour intervals) for individual items in JSON format. This was used to create the macro dataset which captures the long-term price trends across the thousands of tradable items within the game.
- */timeseries*: Used to select more granular time series with exact time intervals of 5 minutes, 1 hour, 6 hours or 24 hours, up to a maximum of 365 data points (excluding 24 hours) per any interval choice. This dataset includes high- and low-price quotes for the selected timeframe and matching high/low volume when selecting 6-hour timeframes or lower. This endpoint enabled the rich dataset to

allow *volume* to act as an optional feature for the neural network models (GRU/LTSM) to determine if it improves performance.

These endpoints were utilized to construct a data pipeline and API scraping tool that produced the datasets in this project. The macro dataset contained 180 days of *daily* average prices and the rich data consisting of up to ~90 days of data with high and low prices and volumes across 6-hour intervals. These rich data observations were aggregated into daily averages, excluding any days missing either high or low volume to maintain modeling validity. Some inconsistencies were observed for low liquidity or infrequently traded items. This caused their available data range to extend beyond the intended 180-day scope (matching the API limit), in some cases reaching as far back as 2021 (see LIMITATIONS).

A diagnostic was performed to determine the proportion of items within and outside of the project's range of 180 days from February 11th, 2025, to August 10th. Of these 4,281 observations, 87.4% of the item data population (3,642 items) were fully within this range, while 12.6% (523 items) had some data farther back than February 11th. Within the in-scope subset (3,642 records, 87%) the Rich Daily dataset (intended ~90-day window) showed that 78.3% (3,239) of items had first trades within the expected period of May 12th to August 10th, while 21.7% (382) began earlier. This behavior indicates that while trading activity aligns with the desired analysis window, the API introduces fragmentation based on item liquidity causing the left-tail outliers in Figure 1. The dotted lines within Figure 1 represent the macro data boundaries that represent the project's scope of data considered for modeling and analysis for this project.

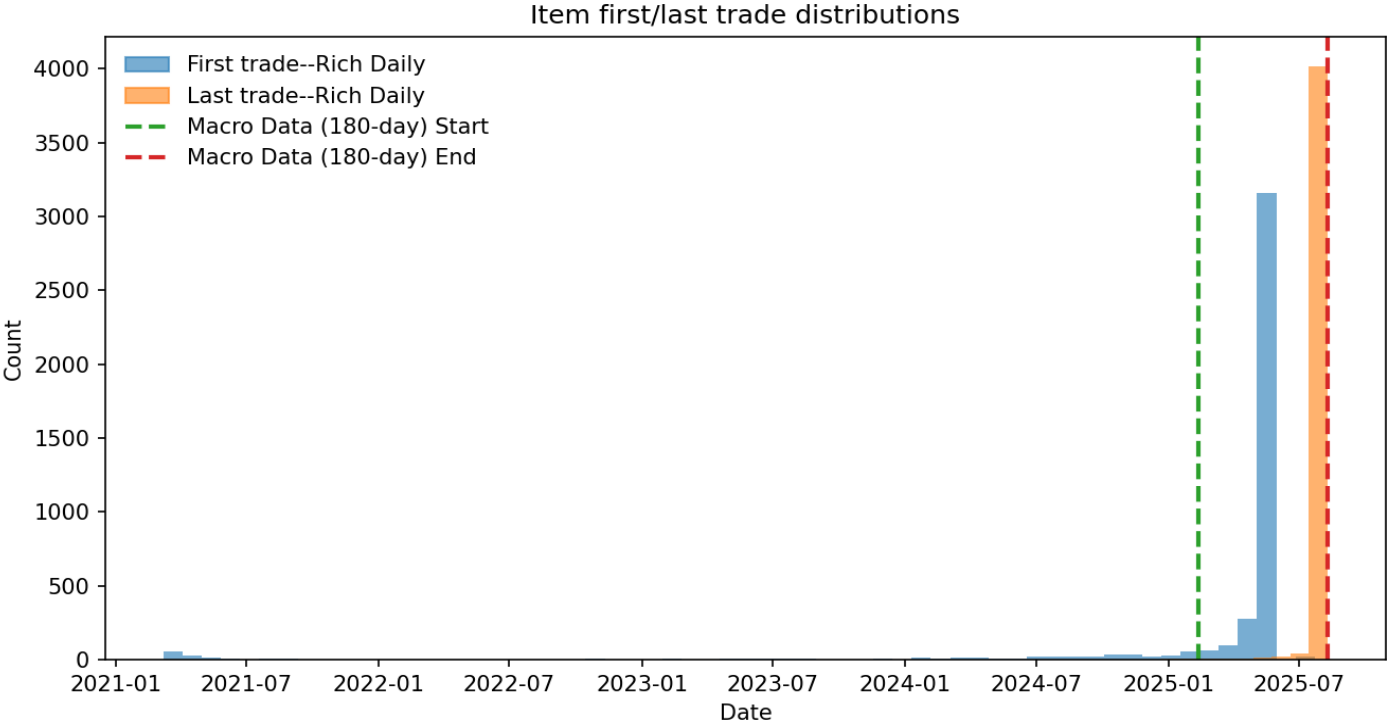


Figure 1 — First record of item trading data within Rich Daily dataset

Variable	Data Type	Description of variable
item_id	Integer	Unique identifier for the item within the database
timestamp	Date Time	Date (or 6-hour mark) when price or volume was recorded.
avg_price	Integer	Average traded price at the timestamp.
volume	Integer	The total quantity of item traded at the timestamp.
name	String	In-game name of the item (from /mapping)
members	Boolean	Determines if item requires premium membership.
highalch	Integer	In-game value for the output of using the spell 'high-alchemy' on that specific item. Acts as a price floor and used for economic reference.
price_changes	Float	Calculated percent change across 30/90/180 days for macro trend modeling.

Table 1 — Key Variables in the item timeseries datasets

Discussion of Variables

The two main variables used for time series modeling and exploratory analysis are *avg_price* and *volume*. The *avg_price* variable captures the market price over fixed intervals and was log-transformed to normalize scale and reduce skew. The *volume* variable is important to clustering tasks in this analysis as high activity items allow widely used consumables or popular items to be differentiated from rare collectibles or highly sought 'chase' items. Item metadata such as *members* and *highalch* were not directly used within this modeling but can offer important contextual information when analyzing specific items.

Item ID	Item Name	Membership	Item Type[2]
11840	Dragon Boots	Yes	Combat Equipment
25386	Trailblazer relic hunter (T3) armour Set	Yes	Cosmetic (League Event)
20014	3rd Age Pickaxe	Yes	Ultra-Rare Clue Reward

[2] 'Type' is not official classification provided by the API but provided for the reader's context.

9075	Astral Rune	Yes	High-Volume Consumable
379	Lobster	No	Consumable (Food)

Table 2 — Item Category Examples by item_id

Features and Derived Variables

This section outlines key variables and features used within the project. Each feature was used to enhance interpretability, model performance, and ensure comparability across item-level price and volume data.

- Log Returns:

  Price series were converted to natural log returns to allow better correlation analysis across items considering volatility/variance across vastly different prices as shown below:

  $$rt = ln(Pt) - ln(Pt - 1)$$

  This is especially important since items within the game can range from a couple of thousand GP to millions, and over one billion GP for certain items. This transformation approximates percentage-based-changes and stabilizes variance for multivariate modeling and time-series analysis.

- Pivoted Log Returns Matrix:

  Log-returns were reshaped into a wide-format matrix structure meaning that all dimensionality-reduction and forecasting steps were performed on log-return values and not the raw price levels. Thus, each item's time series corresponds to a column in this matrix where dates and columns represent individual items. This format ensures consistent alignment across the date range within the dataset for use in PCA, clustering and VAR modeling.

- Top Filtered:

  To reduce noise from items with low liquidity or erratic patterns, exploratory and modeling analysis were limited to the most actively traded items. The feature is a list of the top 150 items ranked by overall trading volume. When mentioned in the project, the variable amount stated always starts with the number 1 highest volume item, going down the list up to the maximum of 150 for that specific analysis. This intentional bias improves robustness and reduces overfitting risks by minimizing interpolation and volatility distortions that low-volume items can produce. This feature selection approach was used primarily for Objective 1 tasks to prevent skewing model accuracy and potential profit calculations.

- Standardization

  All model inputs, including returns, volumes and rolling features were standardized with *z*-scores based on the training subsets. Standardization ensures comparability among features on different scales and prevents items with influential points or outliers causing model skew.

- Timeseries Alignment and Missing Data

  Each item's price and volume series was aligned to continuous daily frequency. Missing values were interpolated using time-based methods, and when necessary, filled the training-set means. This procedure was utilized to preserve a continuous time series while avoiding lookahead bias during model fitting.

- PCA Scores & Loadings:

  Principal component scores were extracted from standardized log-returns to capture shared market-wide variance across items. Loadings were used to identify major contributors to each component, supporting the Factor-Augmented VAR modeling (Objective 1, *rich_daily* dataset) and clustering analysis (Objective 2, *macro* dataset).

(Note: The number of retained components, K, was later tuned during model specification to balance variance coverage and lag feasibility in the VAR modeling phase.)

- Clustering Labels:

  K-means clustering was applied to PCA derived factors to identify groups of items exhibiting similar dynamics over time for Project Objective 2. Cluster assignments were later mapped to item names to enable visualization and interpretation of emergent "market sector" behaviors.

- Neural Network Feature Configurations

  Two features were utilized in configuration for the neural network models.

  1. : *price_only* : accounted for just standardized daily log-return of item data for its model training.
  2. *price_plus*_volume: Used a three-dimensional approach utilizing volume-related features to allow the neural network models the opportunity to see if the added feature assisted in capturing short-term demand shocks and longer-term trading momentum effects.
     - Daily log-returns (*ret)*
     - Log-transformed daily trade volume (*log_vol)*
     - Seven-day rolling average volumes (*vol_7)*

All together these features were utilized to support all modeling stages including factor extraction for VAR, clustering for behavioral grouping and neural network forecasting. These features were engineered in this project to ensure that the results produced are the most robust and methodological sound possible.

Phase 3: Exploratory Data Analysis (EDA)

The EDA phase is performed to obtain a basic understanding using graphical analysis and descriptive statistics (Larose & Larose, 2015). The exploratory data analysis section includes discussion of the variables outlined in phase 2 and possible relationships between them. The main objectives of this EDA include demonstrating distributions of price, volume, and volatility through histograms plots, correlation heatmaps and some example use of PCA.

The data in Table 3 shows a data summary from the 24-hour (averaged) 6-hour timestamp data from the rich dataset. The dataset demonstrates the heavy-tailed and highly varied structure of Old-School RuneScape's virtual economy. The lowest of worthless items trades at valuation of a single 'GP' alongside ultra rare items that hit the 32-bit (signed) integer limit for how much GP the game allows you to carry (known as in-game as "*max cash*"). Additionally, some items have virtually no liquidity with volumes as low as 1 (*NaN/0 excluded*) while others are so commonly traded, they reach into hundreds of millions in volume.

Metric	Daily Price	Daily Volume	6H Price	6H Volume
Count	357,770	357,770	1,133,432	1,505,156
Mean	8,515,610.58	476,947.85	6,285,004.84	114,035.51
Std. Dev.	92,392,868.53	5,172,294.29	68,512,648.88	1,354,082.64
Min	1.00	2.00	1.00	1.00
25th Percentile	413.38	62.00	360.00	9.00

Median (50%)	3,044.69	535.00	2,793.00	89.00
75th Percentile	43,500.00	11,539.00	40,870.62	1,954.25
Max	2,147,483,647.00	467,296,406.00	2,147,483,647.00	160,512,267.00

Table 3 — Price summary of 6H rich dataset with percentage quantiles

For the daily data, the average price exceeded over 8.5 million GP, but the median was only equivalent to 3,000 GP. This indicates that there are a small number of luxury or discounted items that dominate the upper tail. The standard deviation of a whopping 92 million GP reinforces this imbalance. Volume behavior follows a similar trend where the typical item trades around 600 units a day; the mean exceeds 470,000 and the maximum an absurd 476 million units traded.

On the more granular 6-hour scale, the number of observations is roughly triple but with a similar overall distribution. Median prices dropped slightly down to 2,800 GP, and median volume per interval declined sharply at around ~89 units. This difference, including the increased number of observations, is expected with the fragmentation of the daily trading activity. These results confirm that OSRS's virtual economy operates across an extremely varied spectrum of liquidity and valuation. This type of finding motivates investigation in the use of log returns, PCA and clustering to isolate systematic patterns from scale-driven noise.

To explore the distributional properties of item prices and trading volumes across the Grand Exchange, a series of log-binned histograms were constructed using both the daily and 6-hour rich data.

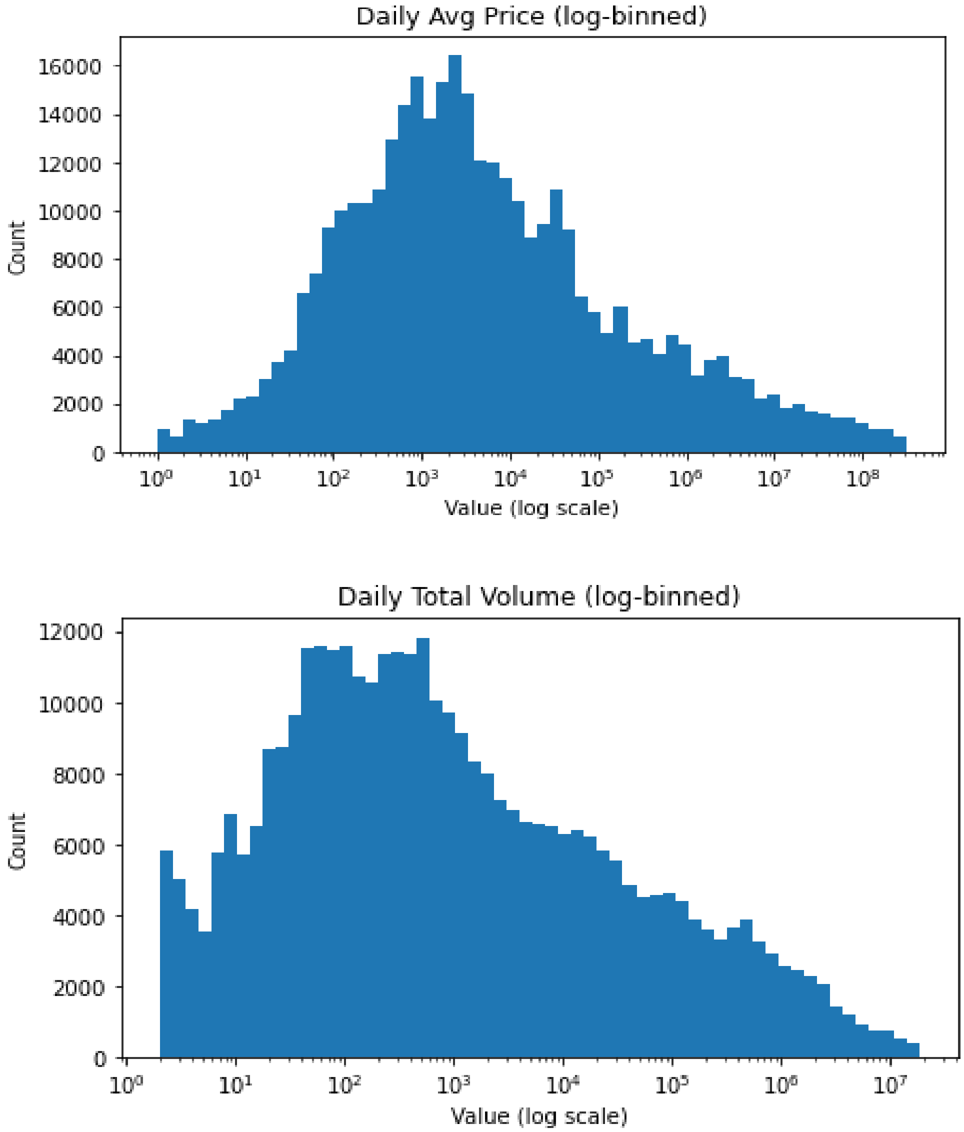


Figure 2 — Daily Volume and Price Histograms

These distributions (Figure 2) were found to be highly skewed and long tailed with extreme lows and highs for volume and price as mentioned in the data summary discussion with Table 3.

Observation of price and volume at the more granular level of the 6H dataset showed similar right-skewed plots with missing bins and sharp vertical bands (Figure 3). Despite log-transformation, artifacts were still present, indicating missing records and offered no meaningful patterns in either Figure 2 or Figure 3.

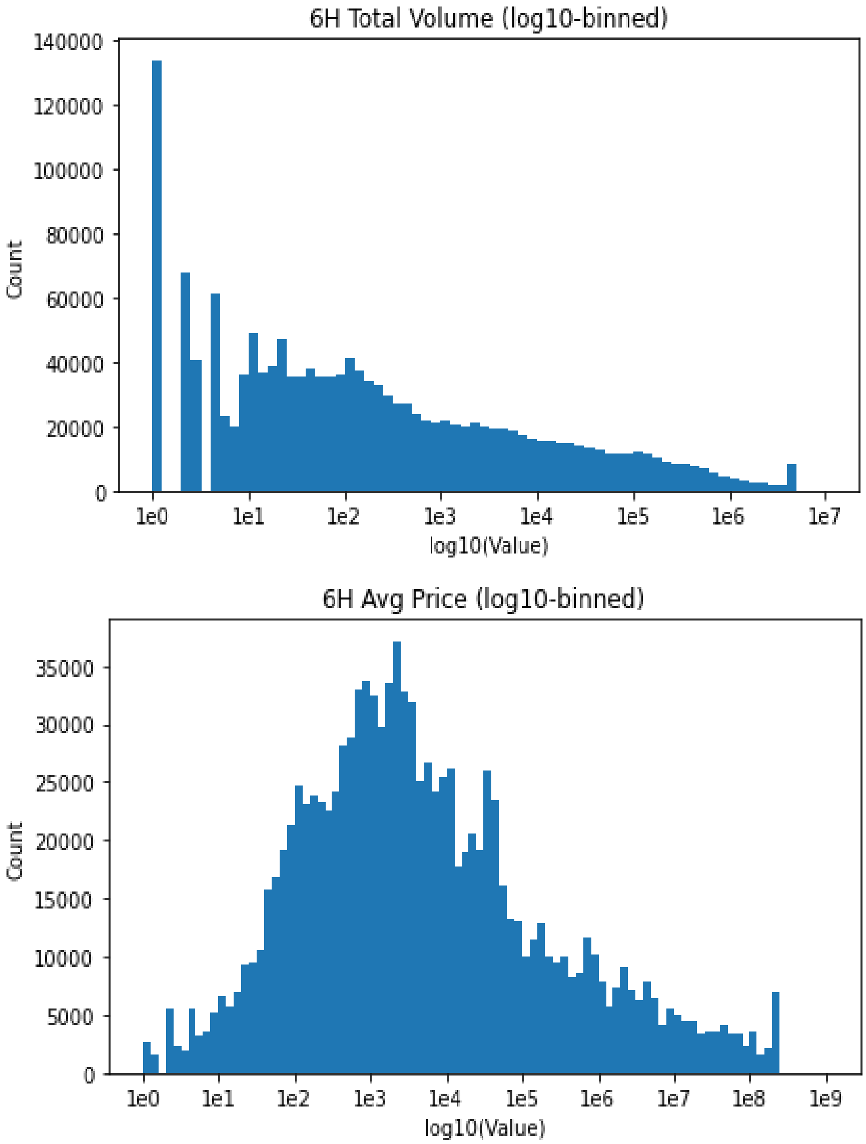


Figure 3 — 6-hour Volume and Price Histograms

Volatility analysis was performed by calculating log returns (log (P_t / P_{t-1}) per item and visualizing their distribution. Initially plotting exhibited extreme tails reaching beyond ±10, in which quantile-based trimming was applied (0.1% to 99.9%) which normalized the *x*-axis to a much more reasonable range for observation. Figure 4 demonstrates that the 6-hour volatility distribution binning seems to be consistent with high-frequency trading noise, while the daily dataset binning show a much smoother, broader dispersion.

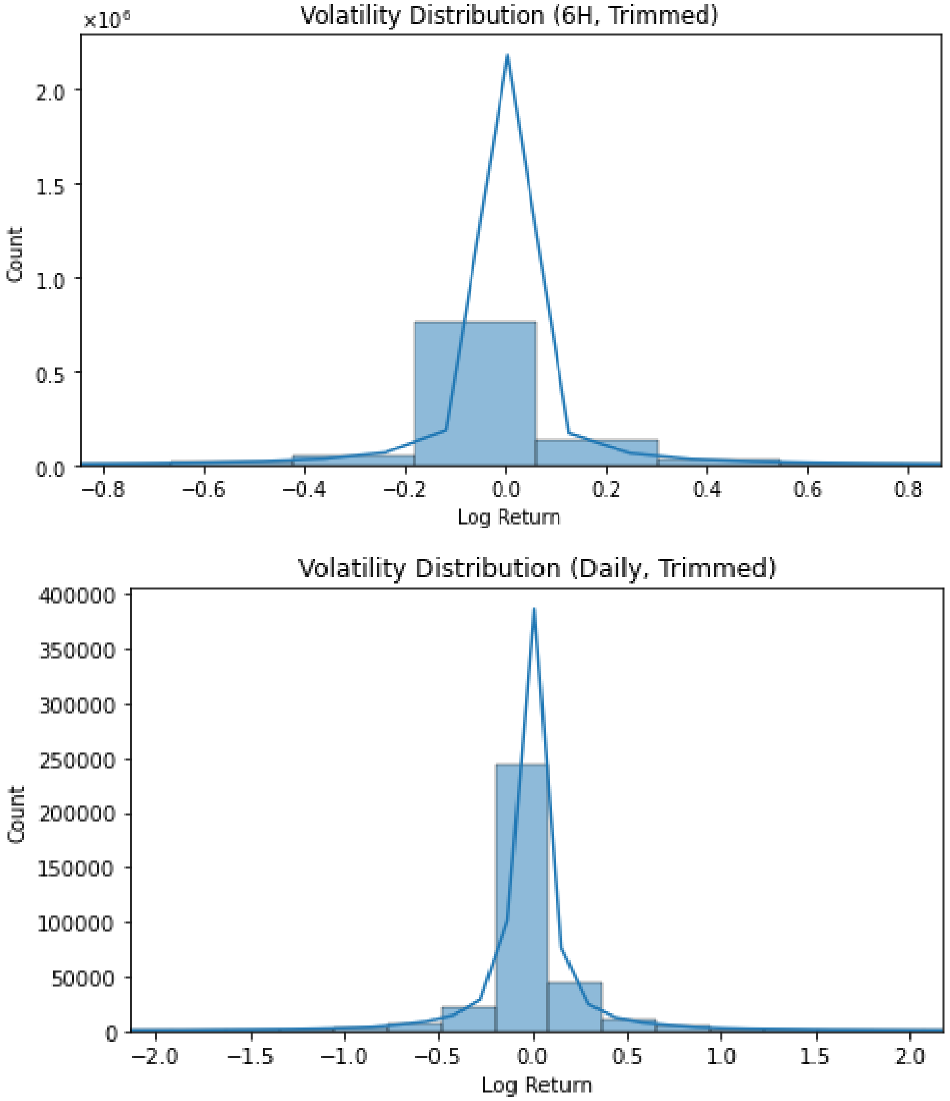


Figure 4 — Trimmed Volatility Distributions

A final overlay of the two volatility distributions normalized by density can be seen below in Figure 5 which highlights some differences in the volatility plots. The 6-hour dataset additionally contains many more intervals than the daily entries causing the 'spike' to be much larger, accounting for the large difference in the *y*-axis scale for the plots. Additionally, the log returns of the 6-hour dataset are smaller, so the same total area (1.0) is crammed together with more entries in a smaller space near 0 than the daily plot as well.

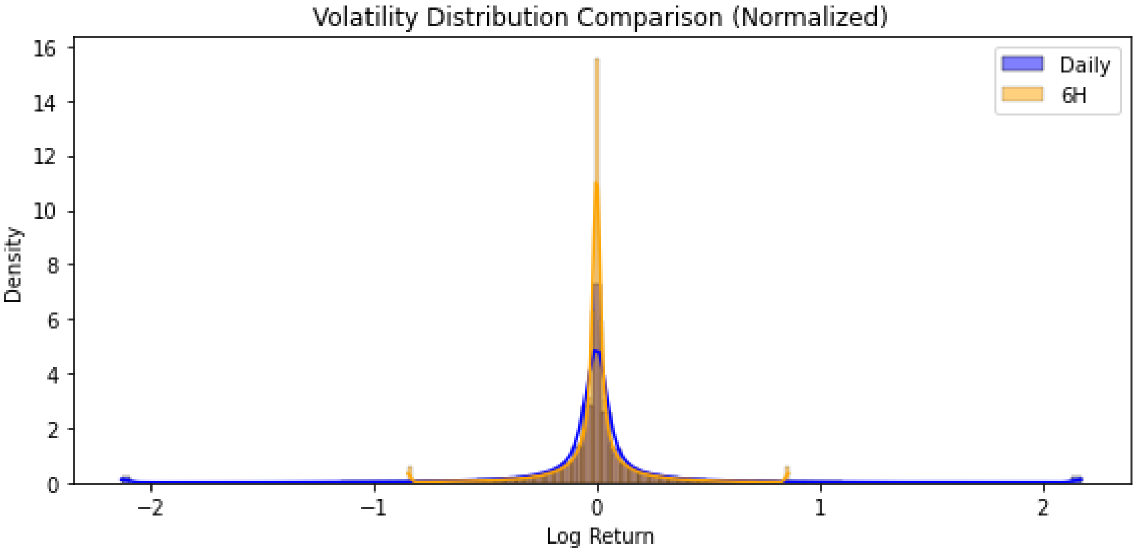
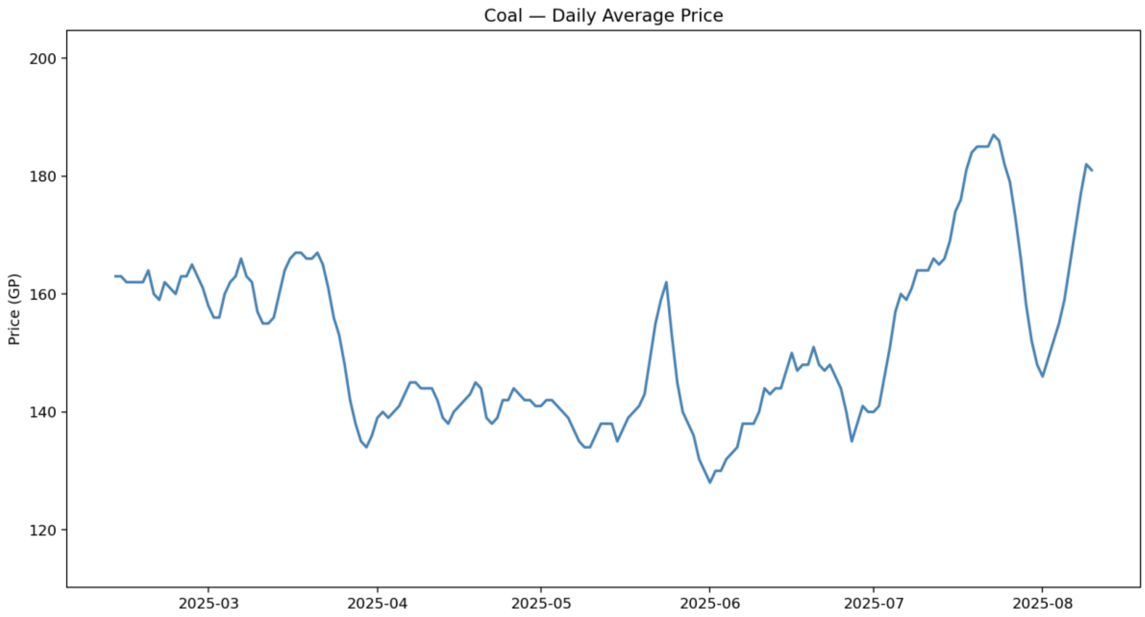


Figure 5 — Overlayed Trimmed Volatility

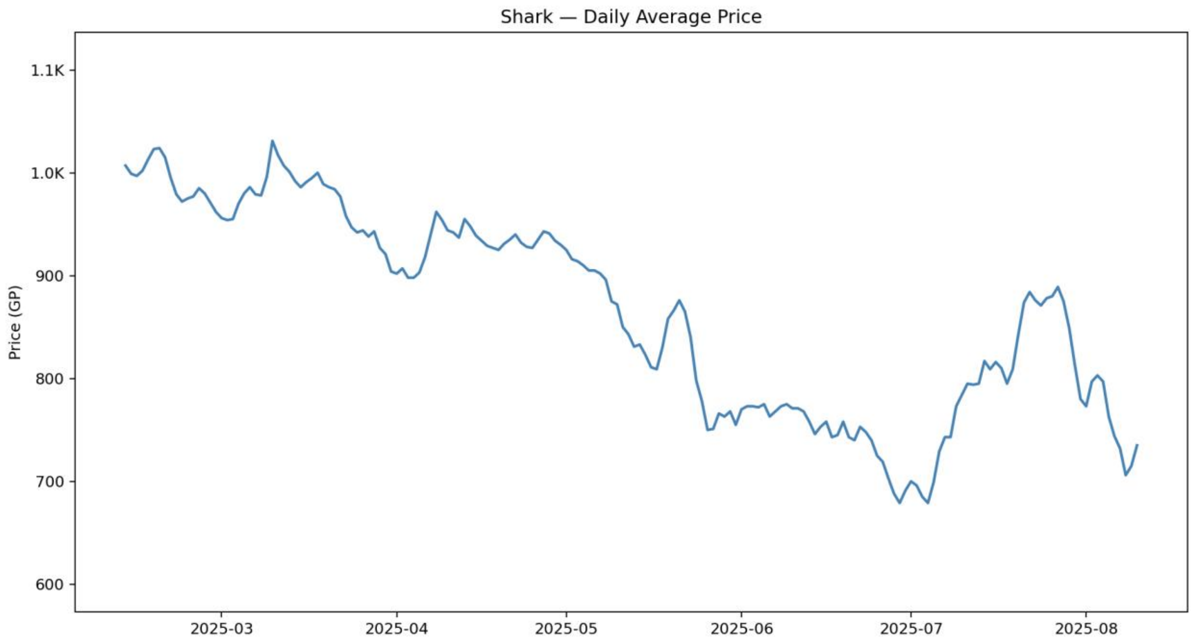
Missingness diagnostics showed that while the median item-day in the 6-hour dataset had 0.00% missing intervals, approximately 11.8% of item-days were missing more than half of their expected 6-hour slices. These severely fragmented sequences came from items that do not trade consistently throughout the day, resulting in patch-driven bursts or long periods of inactivity. Reconstructing these gaps required heavy interpolation, which introduced artificial volatility and destabilized PCA loadings, making factor estimation unreliable and reducing the effective sample length for rolling

VAR or neural network (LSTM/GRU) forecasts. In contrast, the daily dataset exhibited 0% missingness, providing a clean and uniform time grid. For these reasons, the 6-hour data was retained only for exploration and excluded from the formal modeling pipeline (see LIMITATIONS).

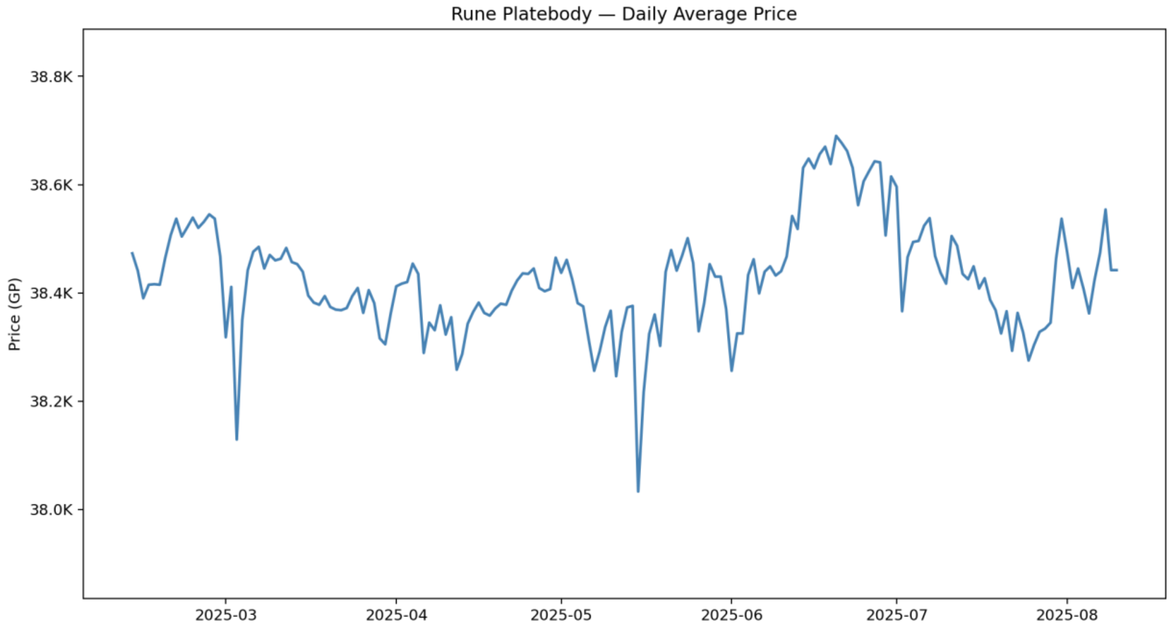
With volume and volatility assessments complete, time-series price action was observed for some key example items. *Figure 6* below demonstrates utilized data from the 180-day macro dataset with daily time points of 24 hours. Item prices on the *y*-axis are modified to reflect in-game game nominations of large sums (*K=thousands, M=millions, B=billions*).



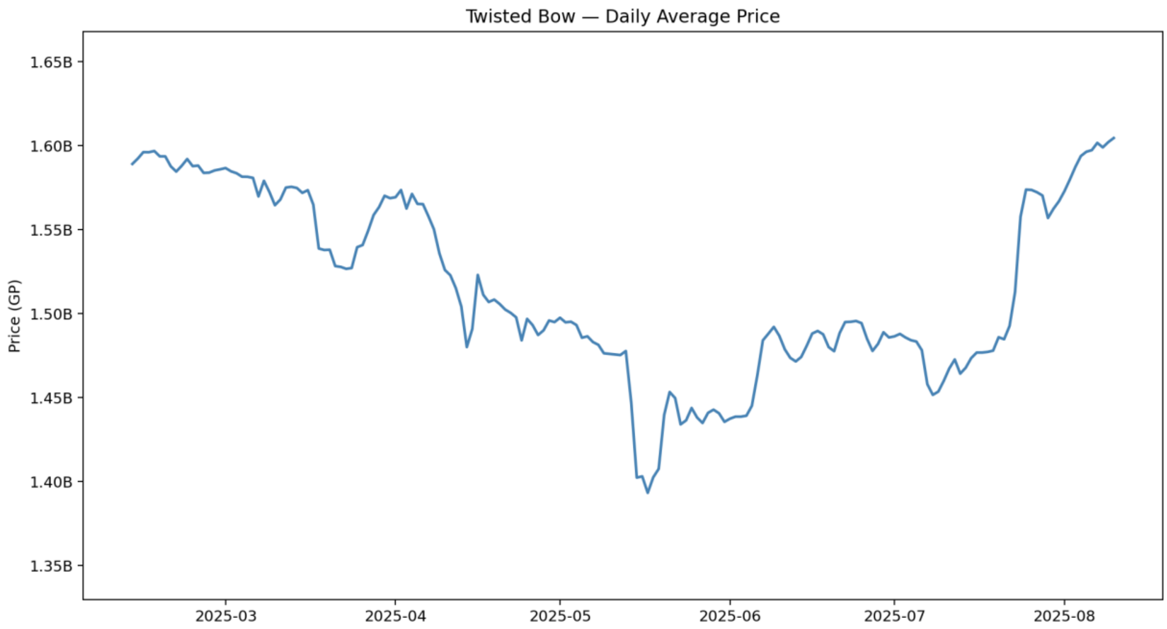
Coal: an intermediate ore obtained as a 'drop' reward from certain in-game activities or from the Mining gathering skill, required to smelt most metals from ore to usable bars.



Shark, a popular high healing food item obtained as a 'drop' reward from certain in-game activities or processed from Raw Shark through the Cooking player skill.



Rune Platebody, the strongest armor available for free-to-play players. Obtained from certain in-game activities or produced from Rune Bar with the player Smithing skill.



Twisted Bow, a 'mega rare' item considered the strongest ranged weapon overall that can only be obtained by one of the three 'raid' style (group play) activities in the game. Using the Oct-18th conversion rate ($0.63 USD per million GP), a Twisted Bow bought off market from scratch at 1.6B GP valuation would cost about 121 OSRS bonds (at 8.99$ USD per bond). …an approximate cost of 1088$ USD!

Figure 6 — Price time series of key example items (macro dataset)

Overall, Figure 6 illustrates distinct market behaviors across item categories within the OSRS economy. Commodity items such as *Coal* exhibit moderate volatility with cyclical dips and recoveries most likely tied to player demand. Consumables like *Shark* show a gradual decline consistent with supply saturation or reduced demand. *Rune platebody* maintains an exceptionally stable price because its value sits near the in-game *high alchemy* price floor, indicating that Jagex's control mechanism design within the game anchor its market value while allowing for minor arbitrage. In contrast, high-tier rare items such as *Twisted Bow* show pronounced swings and momentum cycle most likely driven by speculation or the supply due to being a rare reward for very difficult in-game content (*Chambers of Xeric* Raid). Together, these time-based patterns highlight the multi-layered structure of the OSRS virtual economy consisting of stable foundational goods, elastic consumables and highly volatile prestige assets. These provide essential context for the results discussion of time-series modeling in Phase 6: Evaluation Phase.

Pearson correlations (pairwise) were computed on the same-day movements of the daily log returns from the *Top Filtered* subset (30 items). The results of the top co-moving item pairs are reported below in Table 4.

Item ID 1	Item ID 2	Correlation	Item Name 1	Item Name 2
1877	8928	0.990	Ugthanki & onion	Hat eyepatch
1879	23336	0.983	Ugthanki & tomato	3rd age druidic robe top
1877	23124	0.952	Ugthanki & onion	Gilded dragonhide set
22239	6061	0.939	Dragon kiteshield ornament kit	Bronze bolts (p+)
1877	30430	0.921	Ugthanki & onion	Raging echoes banner
5656	7451	0.871	Steel knife(p+)	Cleaver
5656	28783	0.837	Steel knife(p+)	Trailblazer reloaded relic hunter (t3) armour set
5629	7070	0.764	Iron dart(p+)	Minced meat
5655	5617	0.744	Iron knife(p+)	Iron arrow(p+)
5655	28783	0.679	Iron knife(p+)	Trailblazer reloaded relic hunter (t3) armour set
5656	30422	0.611	Steel knife(p+)	Raging echoes top (t3)
6061	13038	0.593	Bronze bolts (p+)	Gilded armour set (sk)
1877	7451	0.588	Ugthanki & onion	Cleaver
5655	1877	0.549	Iron knife(p+)	Ugthanki & onion
5656	12437	0.501	Steel knife(p+)	3rd age cloak

Table 4: Top 15 most correlated item co-movers with Top Filtered subset

The correlation results contain surprising linkages between low-visibility items and ultra-rare, high-value drops. Specifically, intermediate food items such as *Ugthanki & Onion/Tomato* showed strong price correlations with clue scroll rewards like the *Gilded Dragonhide set* and ultra rare *3rd age druidic top*, as well as time-limited league items (*Raging Echos Banner*). Equally puzzling is the presence of weak, low-level poisoned ranged weapons in Iron/Steel Knives (p+) experiencing similar volatility to ultra rare clue scrolls rewards. Despite high correlation scores and using *Top Filter* items, these results are most likely noise or potentially spurious correlations rather than meaningful intra-item relationships. These results demonstrate that there must be some lagged dependencies as the Pearson pairwise methods were not capturing any significant same-day intra-item relationships.

To investigate if items had meaningful serial or lagged dependencies in their price movement, an Autocorrelation Function (ACF) was used. The analysis was performed on the daily log returns results of the top 25 items from *Top Filtered*. Unlike Pearson's correlations, which observes same-day relationships, the ACF observes an item's past changes over a variable lag length. In this heatmap observation, a lag amount of 20 was chosen which in the context of this project represents 24-hour periods per lag. The ACF heatmap (*Figure 7)* generally shows that there is very little autocorrelation across all lags.

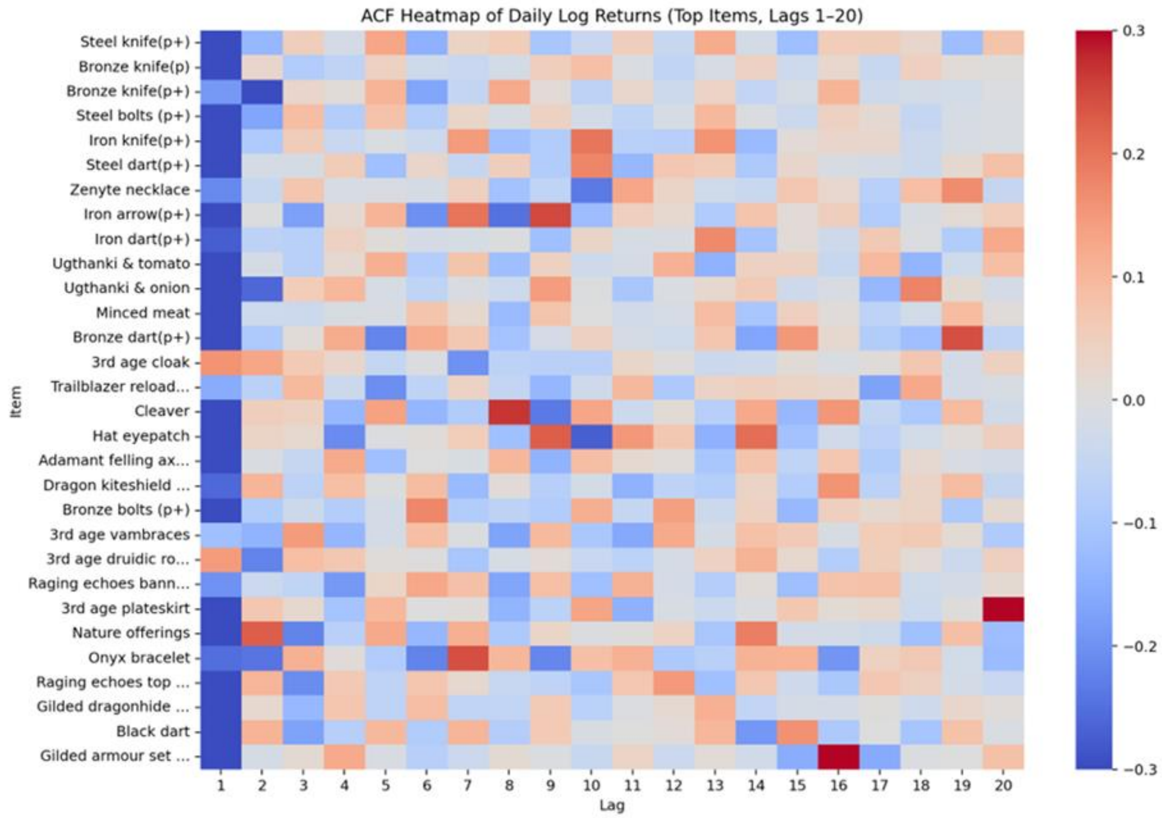


Figure 7 — ACF Heatmap of Daily Log Returns at Lag=20 (Twenty 24-hour periods)

Most values are typically clustering near zero and frequently change to slightly positive or slightly negative. The pattern strongly resembles weak-form efficient or random market structure, in which it is hard for past returns to provide quality indicators for future price movement. Additionally, the first few lags (1–3) show a small negative autocorrelation ‘wave’ for all items but taper off and do not persist with the later lags. This would indicate that whatever price action happened in the first few lags was not cyclical as there was no mean reversion back to all similarly negative correlations. Another observation is that there are no strong sustained positive autocorrelations occurring for most of the items. There are a few lags for specific items that go strongly positive (deep red) but tend to reverse very quickly and are not persistent. These act as supporting evidence to the first observation, in which there does not seem to be any strong indicators for ‘build ups’ or ‘ramping ups’ of items to higher positive or negative correlations (smooth transitions instead of flipping). The heatmap also lacks any vertical stripes other than the previously observed initial negative correlation ‘wave’ in the first lag. There does not seem to be consistent temporal effects affecting the entire group of items, and that autocorrelations are only appearing in isolated cases instead of systematic ones.

To investigate the latent structure in price movements, Principal Component Analysis (PCA) was applied to a matrix of daily average prices with rows as dates and columns as item IDs and *z*-score normalization for the top 150 most frequently traded items.

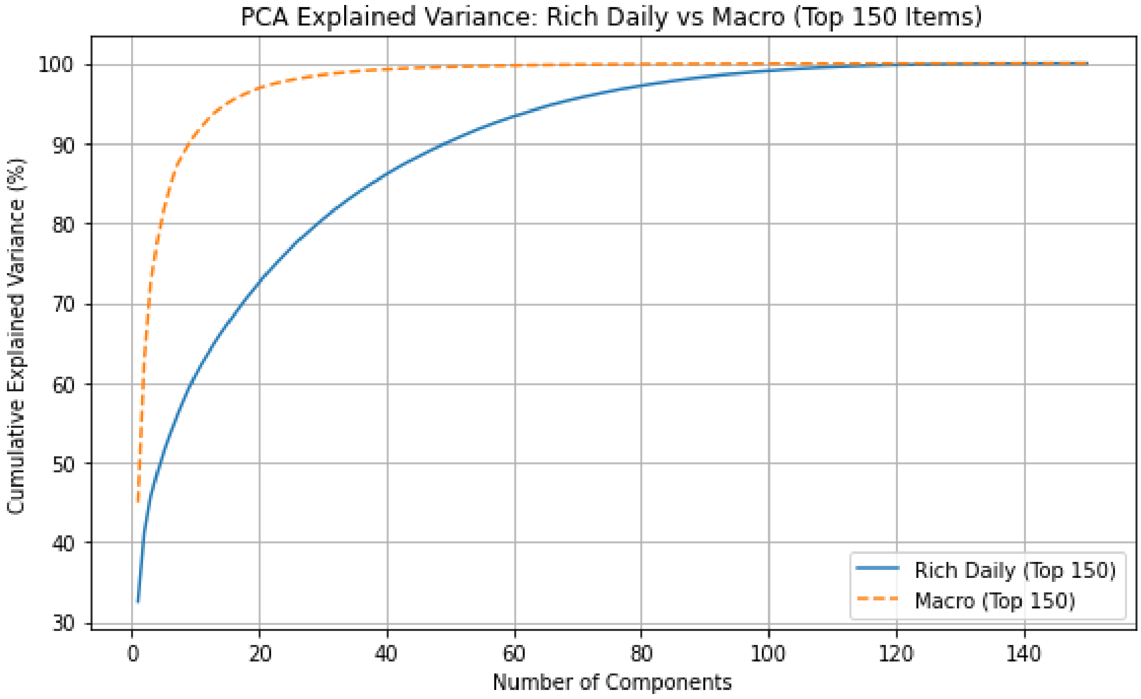


Figure 8 — Variance Explanation vs Cumulative Principal Components Comparison

The PCA comparison between the rich daily and macro datasets shows that both exhibit strong dimensional compressibility, but with meaningful differences in structure. The macro dataset reaches 90% explained variance within roughly 15–20 components, while the richer, longer-horizon dataset requires closer to 50 components to achieve the same threshold. This suggests that short-window daily prices contain a more concentrated set of dominant market factors, whereas the rich dataset reflects additional noise, temporal drift, or more diverse item-specific behaviors. Together, these results indicate that a relatively compact factor structure underlies OSRS item prices—supporting the feasibility of later clustering and factor-based modeling—in both datasets, with the macro window showing the clearest latent structure.

Phase 4: Setup Phase

The modeling performed in this paper was done in a Python 3.10 environment with TensorFlow 3.12.1 to support the neural network models (GRU/LSTM). The datasets utilized the *rich_daily* and *macro* datasets to explore the project objectives, which covered the internal trading data of the GE from February 11th, 2025, to August 10th, 2025. Following DSM methodology (Larose & Larose, 2019), the *rich_daily* data consisting of rich feature data were split into training, validation and test set partitions for model evaluation and comparison. While the nominal 90-day window for this dataset lies from May 12th to August 10th, the actual calendar span of the data partition totals 102 days from May 1st to August 10th. This slight extension was intentionally done to have sufficient training data length and to maintain coverage of low-frequency items whose first trade occurred marginally beyond the idealized 90-day mark (see LIMITATIONS). The data partitions were split according to the following scheme including end-dates:

1. Training: May 1st to July 25th (86 days)
2. Validation: July 26th to August 1st (7 days)
3. Test: August 2nd to August 10th (9 days)

Phase 5: Modeling Phase

Vector Autoregression models (VAR) models treat each variable as a function of its own values and past values of all other variables observed, shown below as an operation.

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \varepsilon_t$$

where y_t is an $N \times 1$ vector of variables (in this case—item returns), A_i represents the coefficient matrices $(N \times N)$, p is the lag order, and ε_t is a vector that represents white-noise residuals. For this project, *N* is the number of tradable items and *p* represents the number of lagged 24-hour intervals. The total number of parameters, given by $N^2 \times p$, scales quadratically with the number of variables *N* and linearly with lag order p. For example, with a "modest" selection of 100 items and a lag selection of p = 5 (five days of lagged correlation), would result in 50,000 parameters being considered within the model—completely unstable for modeling. Using all 3800+ items would result in about 14 million parameters even with a single lag. This dimensionality makes a direct VAR model infeasible for this project.

To mitigate this, dimensionality reduction was implemented prior to VAR modeling. Data were standardized to *z*-scores and reduced via Principal Component Analysis (PCA) that extract latent "factors" that explain the market-wide covariance patterns. After this transformation, the VAR model becomes a *Factor-Augmented VAR* (FAVAR), where the vector is PCA factor scores instead of raw item returns. Two FAVAR specifications were tested: a VAR using the top 65 PCA factors with one lag (VAR(65,1)) and a VAR using 41 PCA factors with two lags (VAR(41,2)).

The number of retained factors *K* was determined by the cumulative variance explained by PCA. From analysis of *Top K* performed in the EDA phase, *K*=65 (≈90% variance explained) served as the *breadth* specification, while a reduced model using *K*=41 factors (≈80% variance explained) was used as the *depth* specification. Lag values *p* were selected and validated using the constraint function below:

$$T - p > Kp \rightarrow p < \frac{T}{K+1}$$

where *T* is the available time series length, K is the number of PCA factors and p is the proposed lag order. Both models were trained on the same TRAIN data range (~86 days), trained on one-step rolling forecasts and then tested on the validation and test periods.

Model performance was quantified using RMSE (root mean square error) and MSE (mean square error) for both factor forecasts and reconstructed item-level forecasts that were compared to neural network models (See Phase 6: Evaluation Phase). The *depth* specification (VAR(41,2)) demonstrated substantially higher errors for validation and test sets, scoring worse in both factor tests and (MAE=10.98, RSME =15.57) and item spaces (MAE=0.150, RMSE =0.314). These results suggest some overfitting may have occurred and ability to generalize was reduced (Table 5). Conversely, the *breadth* model (VAR(65,1)) exhibited significantly lower errors across all metrics, indicating that a single daily lag captures most of the short-term autocorrelation occurring in market-wide movements. Based on these results, the final FAVAR configuration used for comparisons to the non-linear neural networks was VAR(65,1), which provides higher stability, lower forecast errors and stronger out-of-sample performance (Phase 6: Evaluation Phase).

Mode	*K* (factors)	Lag *p*	MAE (Factors)	RMSE (Factors)	MAE (Items)	RMSE (Items)
Breadth	65	1	2.23	2.85	0.084	0.133
Depth	41	2	10.98	15.57	0.150	0.314

Table 5. Breadth vs. Depth VAR – Validation Set Performance

A neural network model "consists of a *layered, feed-forward, completely connected* network of artificial neurons, or nodes." (Larose & Larose, 2015, p. 681). Unlike the factor-based linear forecasting approach of FAVAR, these models act like an artificial neuron with several layers that allow for non-linear problem solving and vary in structure depending on the model type. These types of models were generally made for classifications problem solving but can be applied for a variety of modeling tasks. Within this project, they are a contrast to the linear problem solving of VAR to analyze timeseries dependencies in item-level price movements.

In this project two neural network models (*NN*) were utilized, Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM). Both models are designed to "learn" from nonlinear dynamics and capture patterns that linear models like VAR cannot pick up on. The LSTM networks use a system of *input, output* and *forget* gates that regulate how information flows (pick up on useful trends, forget bad ones) while observing a time series. This allows the network to pick up and track long-term dependencies that persist within the complex, delayed relationships found in financial style data. The GRU networks by contrast combines the functionality of the *input* and *forget* gates into an *update gate* and uses fewer parameters overall. This allows for faster training times but still retaining learning power to still effectively model time series patterns.

The configuration of the neural networks used in this study utilized a training data range of 30 days of "look-back" windows for daily returns to predict the next-day return for each item. For a given time *t,* the input tensor contained the previous 30 days of data, resulting in sequences in the shape of $30\ x\ n$ features per item. Two configurations were used for both models consisting of *price_only* which used daily-log returns and *price_plus_volume* which utilized additional volume features (see "Neural Network Feature Configurations" in Phase 2: Data Preparation Phase). This design choice allowed for a secondary objective within Objective 1 to determine if augmenting price data with volume-based features provided meaningful predictive improvement.

Both *NN* models were implemented in TensorFlow/Keras with identical topologies except for recurrent cell type. The architecture was trained using an Adam optimizer (learning rate of 0.001) and Mean Squared Error (MSE) for loss, Mean Absolute Error (MAE) as the primary evaluation metric and patience levels set to 10 epochs to prevent overfitting.

The training and validation splits for the neural network models were kept the same as the VAR model for comparability (see Phase 4: Setup Phase). From the *Top-Selected* feature, the top 50 most-liquid items were individually modeled, trained and validated. Across all items considered, roughly ~7500 training sequences of 30-day inputs and 450 test sequences were produced per feature considered. All series were standardized within the training sets only, to prevent leaking into the validation and test sets during scaling.

Phase 6: Evaluation Phase

The results of the main project objectives are discussed in this model evaluation phase and consist of the following:

1) Objective 1: Comparison of Feature-Augmented Vector Autoregression (FAVAR) and Neural Network models on accuracy and other metrics.
2) Objective 2: Observations of clustering of Principal Component’s behaviors and potential formations of “economic sectors”.

3) Objective 3: Observations of potential structural breaks and event-aligned shocks to item price data (from major content release or updates) through Bai-Perron tests and other metrics.

The models were assessed through two accuracy metrics, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Both metrics evaluate accuracy by measuring the average magnitude of errors between the predicted value generated in the model vs the actual value from the test set. MAE averages the absolute difference between the pred and test values, in which it treats all errors the same (more robust for outliers). RMSE squares the error before averaging the errors then takes the square root, which penalizes larger errors more heavily (sensitive to outliers). The error metric equations can be seen below:

- Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N} |y_t - \hat{y}_t|$$

- Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N} \left([y_t - \hat{y}_t] - [\overline{y_t - \hat{y}_t}]\right)^2}$$

The results from these two-accuracy metrics act as the main discriminator for determining if virtual economic data from Old-School RuneScape is better suited toward nonlinear neural networks or more traditional multivariate statistical methods (in a time-series based analysis).

Objective 1 Discussion

After model training, the FAVAR and neural network models were evaluated using the held-out test data from *Top Filtered*, the 50 most liquid items. This was done to minimize noise from low liquidity items, preventing interpolation or fragmentated price histories which cause issues during model training. This filtering step ensures that performance difference between models reflect actual modeling ability rather than external forces acting upon them. The held-out test data period covered nine days, from August 2nd to August 10th. Two neural network configurations were examined: a *price_only* setup directly comparable to the VAR baseline, and a configuration using additional features "*volume"* (log-transformed daily trade volume) and "*vol7"* (seven-day rolling average volume feature). The second neural network configuration was considered to determine if an additional feature (volume) would improve the performance from the *"price_only"* baseline. Forecast quality was measured using mean absolute error (MAE) and root-mean-square error (RMSE) on these test segments.

Overall, results indicate that while the VAR model achieved the lowest median forecast error, both neural architectures performed comparably. Differences were small in magnitude, implying that, for stable and liquid markets, traditional linear forecasting remains highly competitive with nonlinear deep-learning methods.

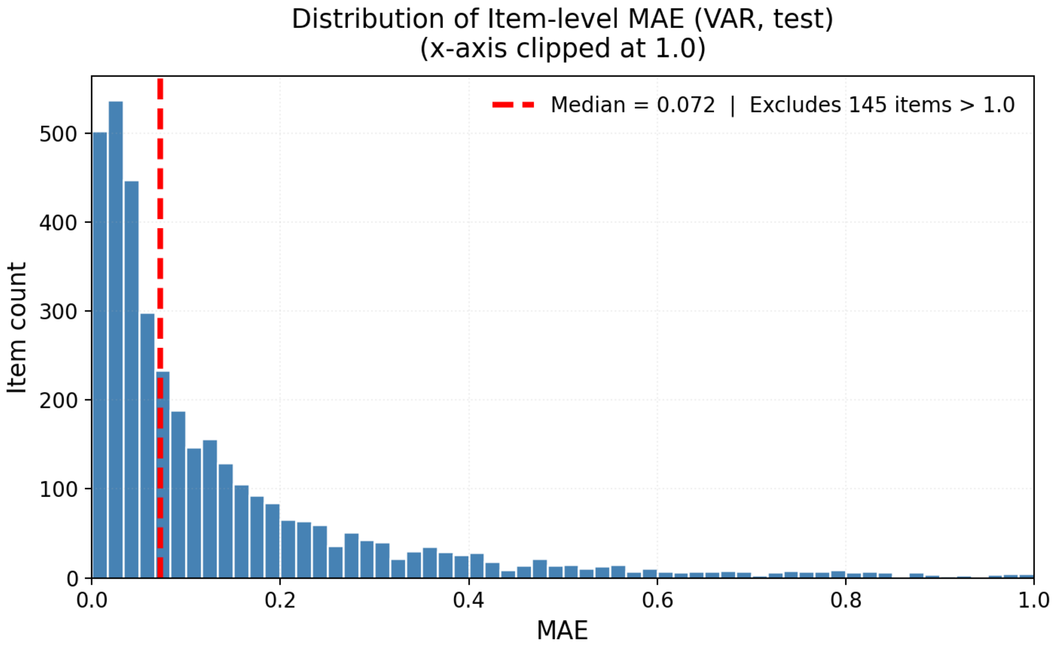


Figure 9 — Distribution of Item-Level MAE in FAVAR

The distribution of VAR forecast errors provides a reference for typical predictive difficulty across all items. As shown in Figure 9, most series exhibited very low test-set errors (median ≈ 0.072 MAE), with a right-skewed tail extending toward fewer stable items. The x-axis was clipped at 1.0 MAE, excluding 145 items above that threshold to emphasize the central distribution. This pattern reflects a market dominated by mean-reverting, stationary behavior with a small subset of volatile items contributing disproportionately to overall error variance. Those tail cases likely correspond to newly introduced items or episodic supply shocks. These phenomena are examined later under structural-break analysis in objective three.

Median test-set performance across the top 50 liquid items is summarized in Figure 10 with median MAE values of 0.022, 0.025 & 0.028 for VAR/LSTM/GRU respectively. Additionally, the neural network models were compared with *price + volume + vol7* features added (Figure 10). LSTM and GRU improved slightly to 0.024 MAE and 0.027

respectively. These differences are small at +0.01 MAE which indicates only marginal gains from additional feature dimensionality.

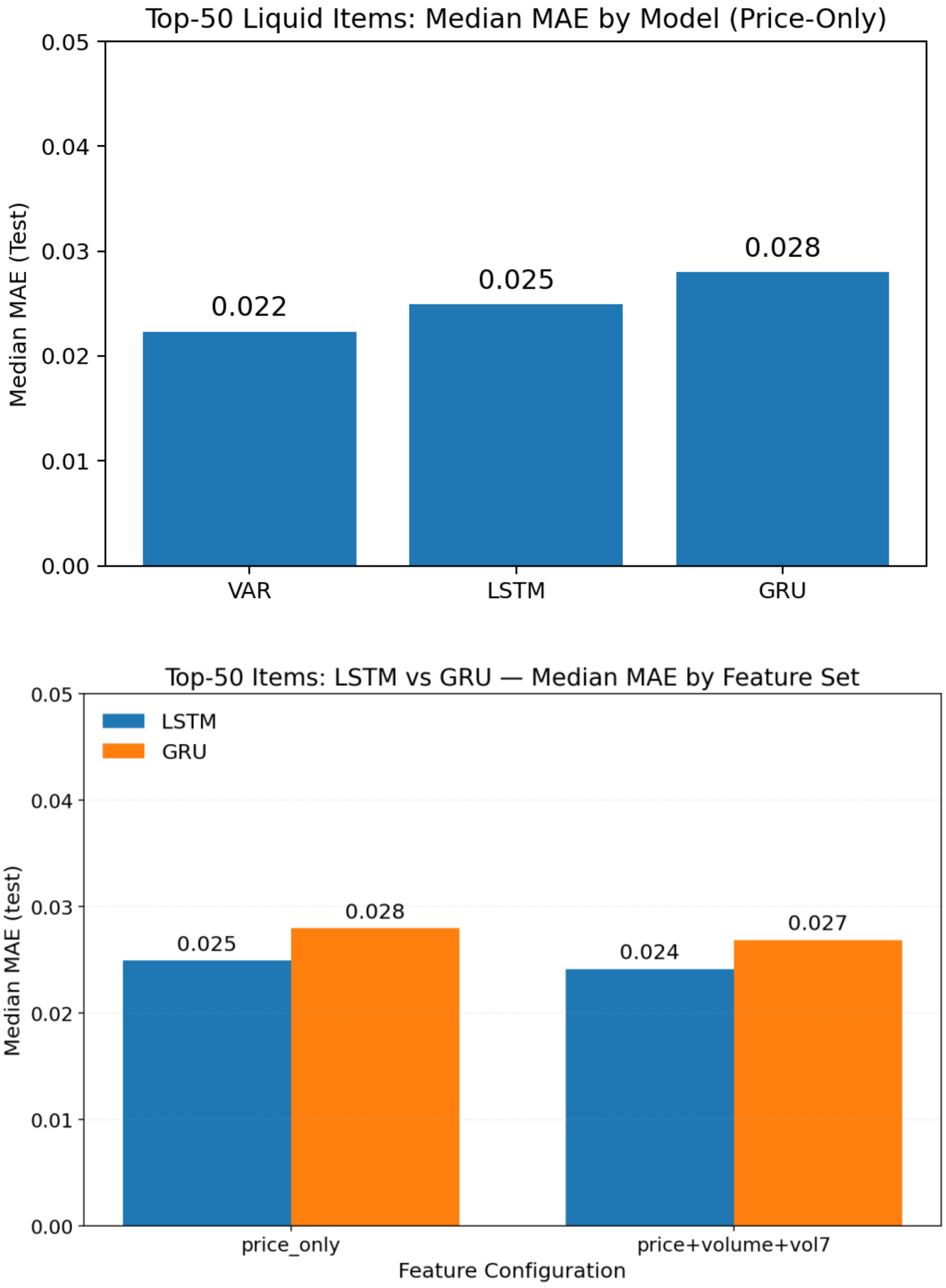


Figure 10 — Top-50 Median MAE by Model and Feature Sets

In short-horizon, high-liquidity forecasting, the VAR model's linear lag structure captured the dominant autocorrelations effectively. Neural models, though designed for nonlinear sequence dependencies, did not produce large advantages under these conditions, suggesting that daily OSRS prices for liquid items remain largely linear and stationary in behavior.

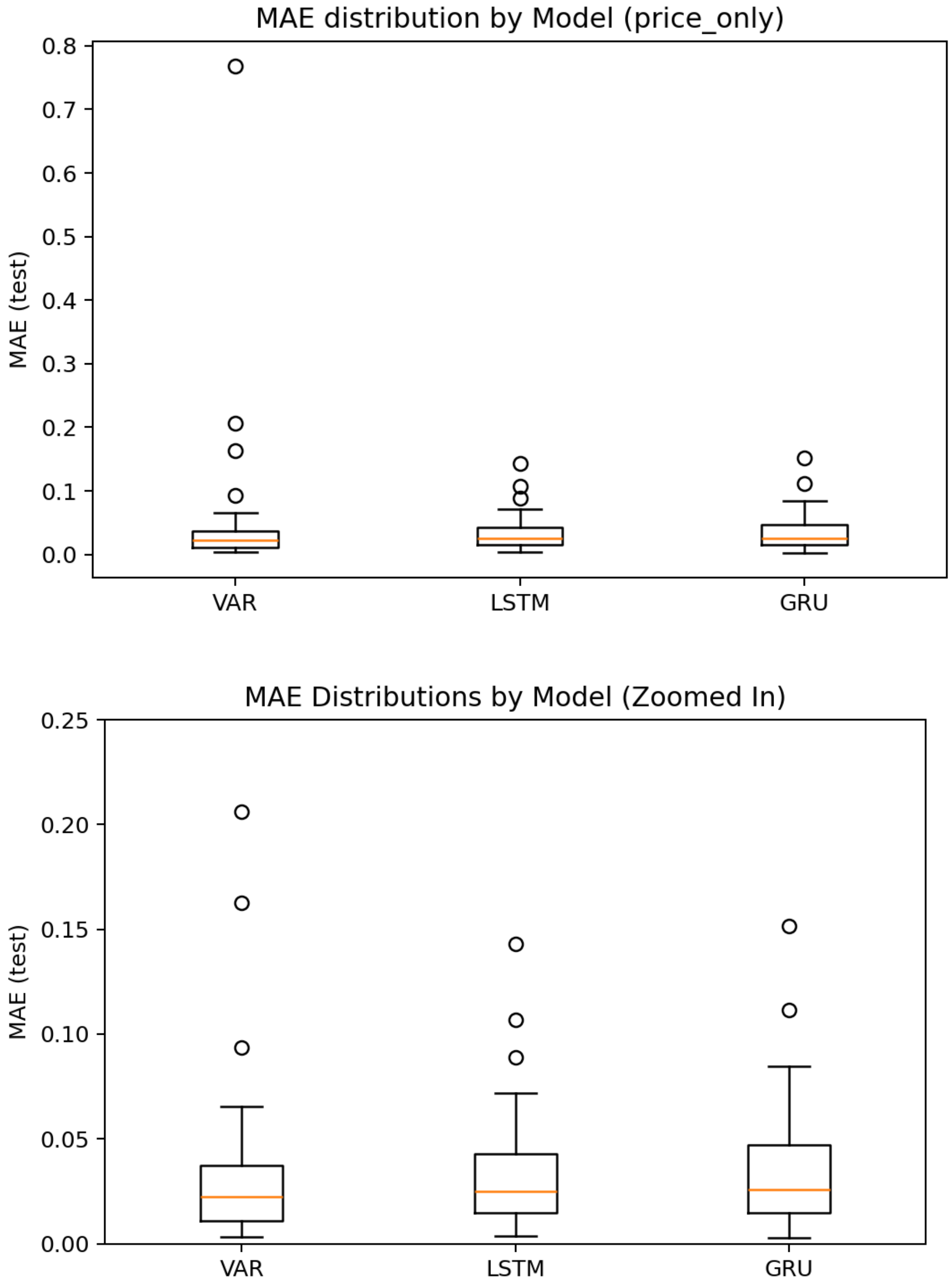


Figure 11 — Boxplots of MAE Distributions by Model

The dispersion of forecast errors offers further insight into model consistency. Figure 11 shows that the VAR model yielded the narrowest interquartile range, denoting stable predictive accuracy across items. Both LSTM and GRU presented slightly wider spreads and more outliers, implying greater sensitivity to temporary volatility or regime changes within the training horizon. Despite these differences, all three models maintained median errors below 0.03, confirming broad performance parity. The results reinforce that average accuracy alone is not the key differentiator; rather, neural models exhibit higher variance linked to the dynamic complexity of certain items.

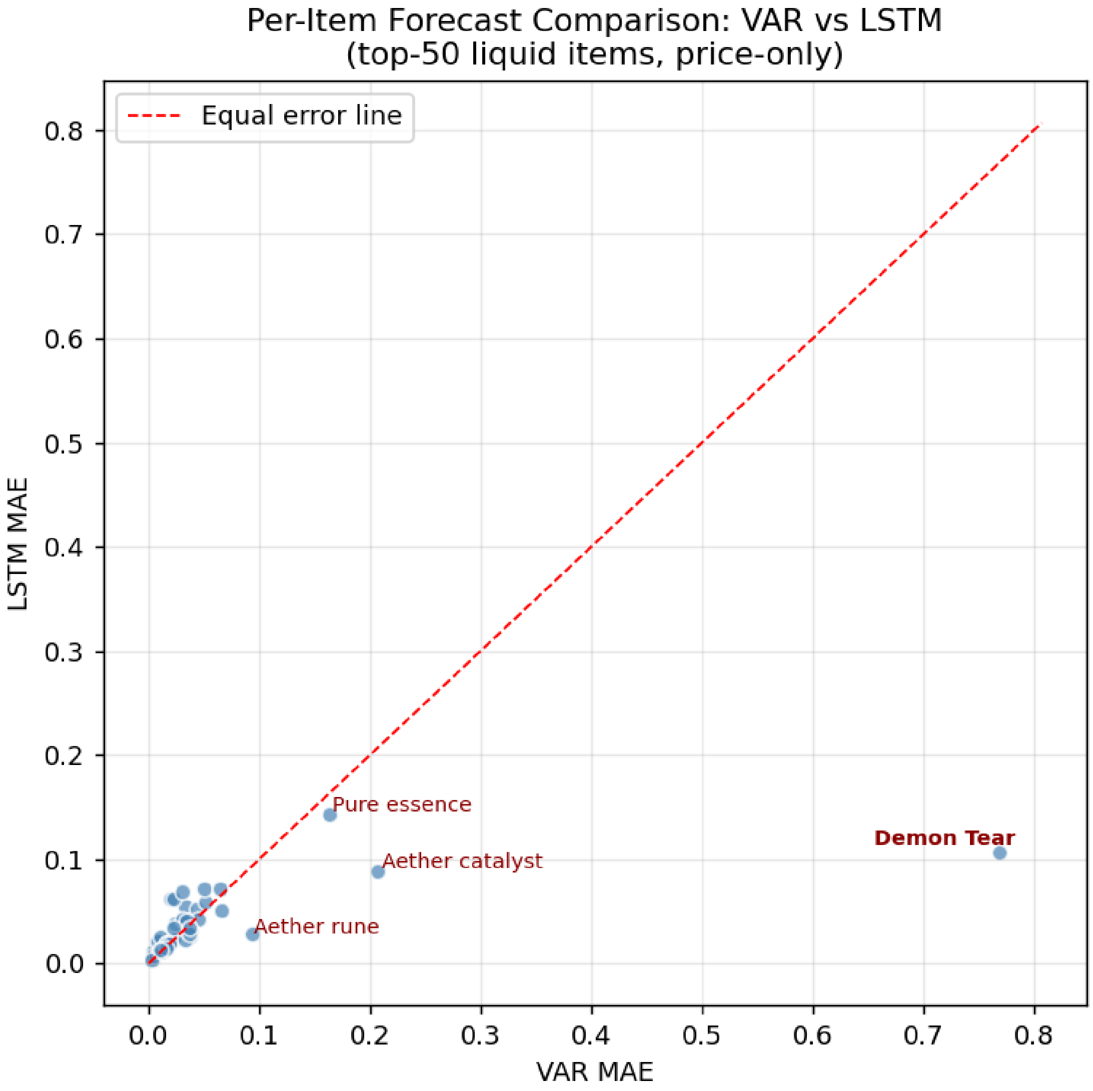


Figure 12— VAR vs LSTM Per-Item MAE Scatterplot

To visualize how individual items behave under the two different types of models, a per-item MAE plot from the VAR and LSTM models were compared directly (Figure

12). Each point represents one item, with the red diagonal line denoting equal errors across models. Most items lie closely along the red diagonal indicating similar predictive performance between both VAR and LSTM. A handful of outliers sit far away from majority of the items, including Demon Tear, Pure Essence, and Aether Catalyst. These deviations reflect sudden regime shifts such as low-GP items (Pure Essence) experiencing high volatility, or content-driven demand jumps or supply floods (Aether Catalyst/Demon Tear). When markets transition rapidly, recurrent networks may overfit the most recent movement while the VAR may lag the slope after price shocks. Items in this fringe category become important diagnostic markets that set the stage as valuable diagnostic context for later structural-break analyses.

To illustrate model behavior, in Figures 12.1 to 12.4 selected items from the evaluation set were examined qualitatively with the FAVAR model. All MAE/RMSE values reported in this study were computed on the log-return (factor) space, while the plotted price overlays reflect the back-projected forecasts transformed back into GP levels. As a result, even small errors in the return space (forecasted PCA) can translate into visually larger deviations in GP units once compounded over several forecast steps.

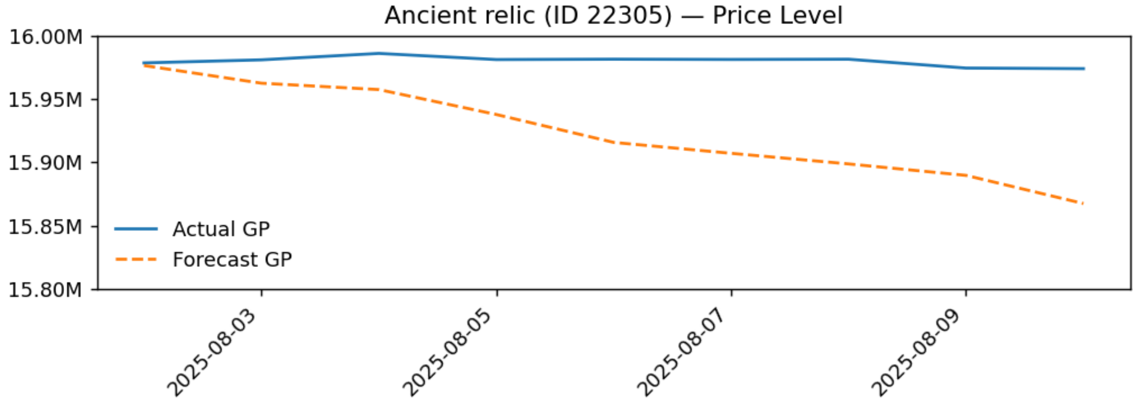


Figure 12.1 —Forecast Overlay: Ancient Relic (ID:22305) MAE = 0.001, RMSE=0.001

The Ancient Relic (ID:22305) is a high-value, low-drop rate item that is exclusive to a dangerous subsection of the map in a '*Player vs Player*' zone known in-game as the '*Wilderness'*. This item has a guaranteed GP value redeemable by a roaming non-player character within the same zone it drops, providing an in-game mechanism that keeps its market price consistent. The item with its stable, liquid and mechanically bound price with minimal noise unsurprisingly achieved extremely low error in the entire dataset across all models (≈0.001 MAE). While the GP deviation in Figure 12.1 (about 150k GP) seems like a large difference, it corresponds to less than a 1% difference on an item just under 16M GP. Because evaluation is performed on log-returns, the MAE of 0.001 reflects an average per-step error of approximately 0.1%. Over a multi-step forecasting window, these small return errors accumulate, producing a visually larger difference on the raw price plot while still representing small relative error. This item, along with others such as *Dragon Plate-legs* (ID:4087) and *Mystic Robe Top* (ID:4091), exhibits similar stable behavior because these items are ideal targets for in-game mechanics that create price floors (*High-Alchemy spell converts items to GP)* had the lowest errors across all models.

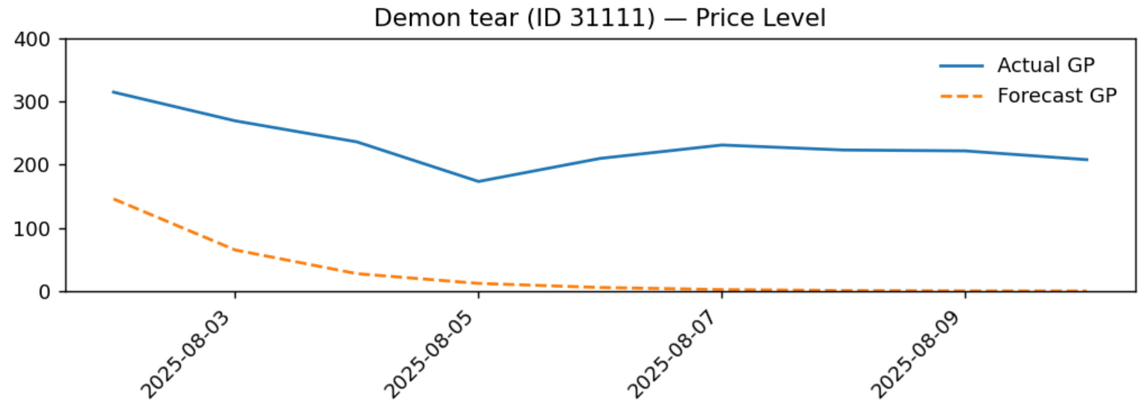


Figure 12.2 —Forecast Overlay: Demon Tear (ID:22952) MAE = 0.769, RMSE =0.774

Demon Tear (ID 22952) represents the clearest example of a forecasting failure under both models, but for reasons that extend beyond the shock itself. The item was introduced with the *Doom of Mokhaiotl* boss and immediately surged in demand due to it interacting with two of its major chase drops/rewards, *Avernic Trends* (ID 31088) and *Eye of Ayak* (ID 31115). Initially scarce, expensive and in high demand, the item became chased for high profit opportunity either through farming on a massive scale via engaging the boss or through the passive gathering player skill method. This caused a substantial supply influx shortly after release, as well as the initial demand spike collapsing as the narrow buyer demand elasticity. Price collapsed shortly after release, which resembles a rapid speculative cycle followed by a supply-driven correction.

Analysis under Objective 3 revealed a critical issue tied to the item's data availability. 'Demon Tear' item, along with other items released with the *Doom of Mokhaiotl* boss, only begin appearing in the rich datasets on July 23rd. This was discovered through the macro dataset in which items introduced from this content had price history backfilled by the API with median placeholder values. According to our training/validation/test split for the VAR model, that means the item only contributes three days of true training data, and roughly 10 usable points when factoring the validation data before the forecast/test data horizon starts. This left the VAR model with extraordinarily slim price history for predicting future behavior items specific to this content (Demon Tear', Eye of Ayak', Avernic Trends' etc.). 'Demon Tear' became a substantial outlier from this group as the model was effectively asked to predict a major supply driven crash with almost no prior regime information, stable history to estimate factor loadings, or meaningful embedding in the PCA factor structure. Therefore, in

Figure 12.2 the model experiences large errors in the prediction window. These findings mean that Demon Tear ‘outlier’ is a structural artifact of data availability that slipped through data processing and discovered in later analysis in Objective 3.

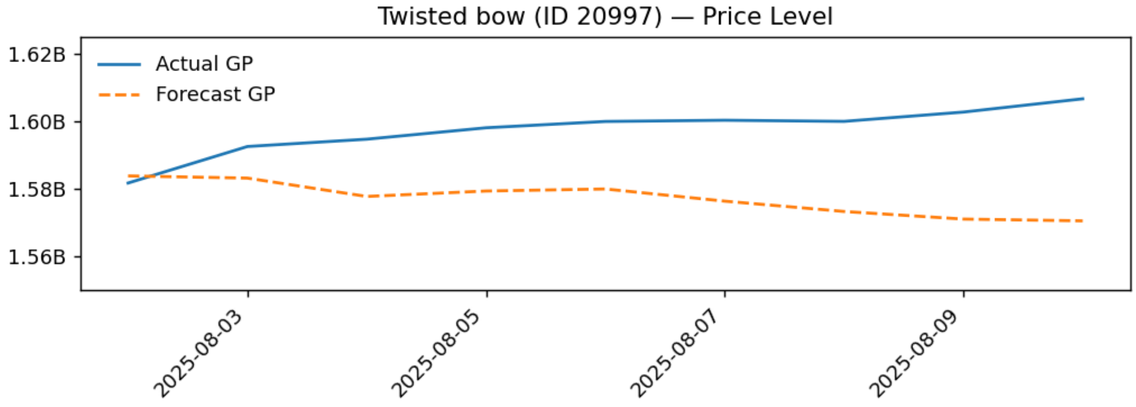


Figure 12.3 —Forecast Overlay: Twisted Bow (ID 20997) MAE = 0.003, RMSE = 0.003

Twisted Bow (ID 20997) is one of the most expensive and most traded high end ‘Mega-Rare’ items in OSRS with a price in the range of 1.5B or higher. Despite occasional long-term cyclical macro trends as shown previously in Figure 6, its short day-to-day volatility is low. Since the evaluation is performed on log-returns, the MAE of 0.003 corresponds to average daily direction error of 0.3%; because the item’s basis is extremely high, even a small percentage error produces large GP swings as demonstrated in the equation below:

$$1.6B \times 0.003 \approx 4.8M\ GP\ per\ day$$

Since the testing window is multi-step (recursive forecasting), these small return errors can accumulate producing a visible divergence as in the raw price plot for Figure 12.3:

$$\approx 20M\ GP = \frac{20M}{1.60B} \approx 1.25\% \text{ difference (actual vs forecast)}$$

This behavior mirrors what was seen in Figure 12.1, in which the GP difference seems large, but the relative error remains very small.

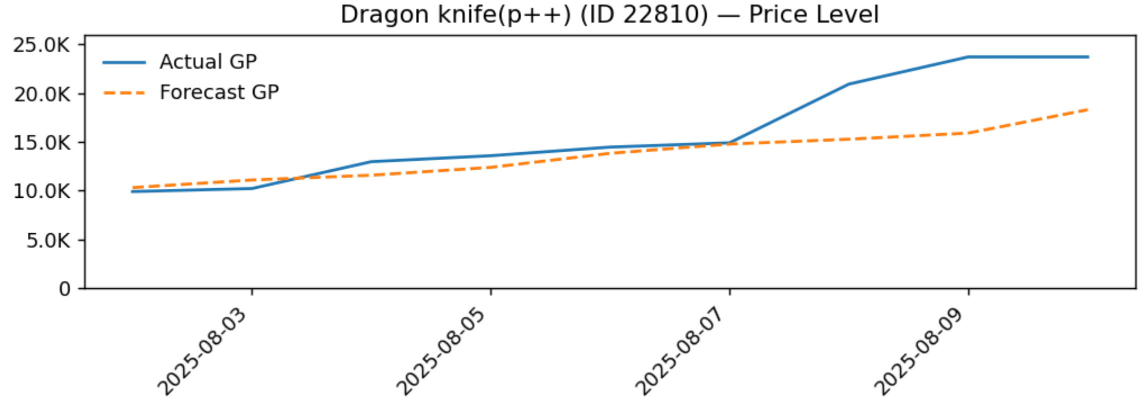


Figure 12.4—Forecast Overlay:Dragon Knife(p++) (ID:22810),MAE=0.1,RMSE=0.186

Dragon knives (especially this poisoned "++" variant) are the best-in-slot special attack weapon for the *Doom of Mokhaiotl* boss content. Following the *Varlamore Final Dawn* update in which the boss was released, demand and price increased sharply as players discovered their extreme effectiveness against a certain phase within the fight.

The VAR model successfully captures the directional upward trend but underestimates the magnitude, lagging the steepest increases. This middle-case scenario demonstrates that VAR can pick up on directional momentum but struggles with sharper valuation changes such as breaks. This transitional behavior seen in this case helps bridge the gap between the FAVAR's model performance between the more stable items (Ancient Relic, Twisted Bow) vs highly unstable items with data availability issues and fragmentation (Demon Tear). Several items exhibited price dynamics that standard linear forecasting had issue capturing. An example being items that were those affected by patch releases, meta shifts, or sudden supply influxes. These cases sparked interest in to systematically examining these types of items in in Objective 3, where structural breaks and update-driven shocks are analyzed in depth.

To assess the structural drivers of these price forecasts, two additional diagnostics were examined to ensure that the VAR model results are robust. The relationships between liquidity and forecast error in Figure 13 as well as auto correctional tests on the VAR factor residuals within Table 6 and Figure 14.

The first diagnostic examines whether items with higher trading activity are easier for the model to predict. In Figure 13, (Forecast Error vs Liquidity) the plots log-MAE on the *y*-axis against log-average daily volume. The overall trend shows a negative relationship with extremely low-volume items show MAE values spanning 10^{-3} to 10^{0} often with large spikes. In contrast, the high-volume items (10^{5}–10^{7}+ trades/day) cluster tightly around MAE $\approx 10^{-2}$ zone. Overall, the figure shows a fan-shaped pattern indicates heteroskedastic predictability with liquid items have stable, low-error behavior while illiquid items show high variance in forecast error. This behavior is common in real-world financial markets, where thinly traded assets (such as ‘microcaps’ stocks usually below $250M market cap) exhibit more volatile price behavior due to sparse order flow and larger bid–ask spreads. For the context of this project feature OSRS’s virtual economy, this reinforces why liquidity screening and utilizing the variable *Top Filtered* (Top 50 items here) was an important modeling decision for downstream fairness in comparing model performance.

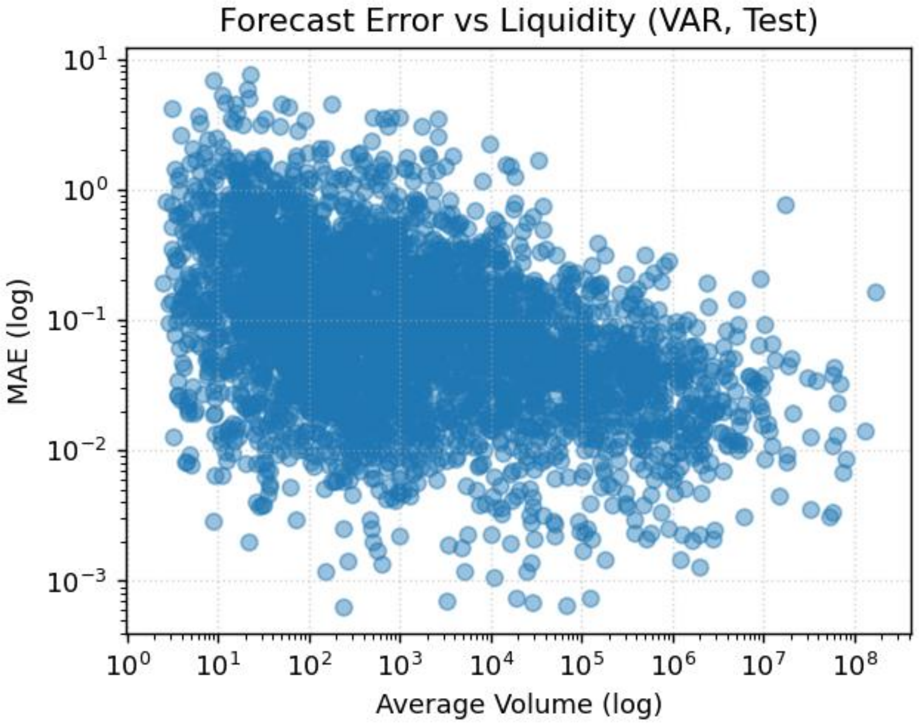


Figure 13 — Forecast error vs Liquidity with FAVAR model.

The context for the second diagnostic is observing if autocorrelation or seasonal cycles are present after fitting (training) the FAVAR model. The diagnostic used to quantify autocorrelation was Ljung-Box tests that were applied to the residuals of the first six principal components. The tests are generally used to evaluate if the extracted factors are behaving like white noise or if they are explaining market-wide variances. The result of this diagnostic can be seen below in Table 6 with the lag-10 *p*-values above the *α* (*p(10) = 0.05*) do not have enough evidence to suggest autocorrelation.

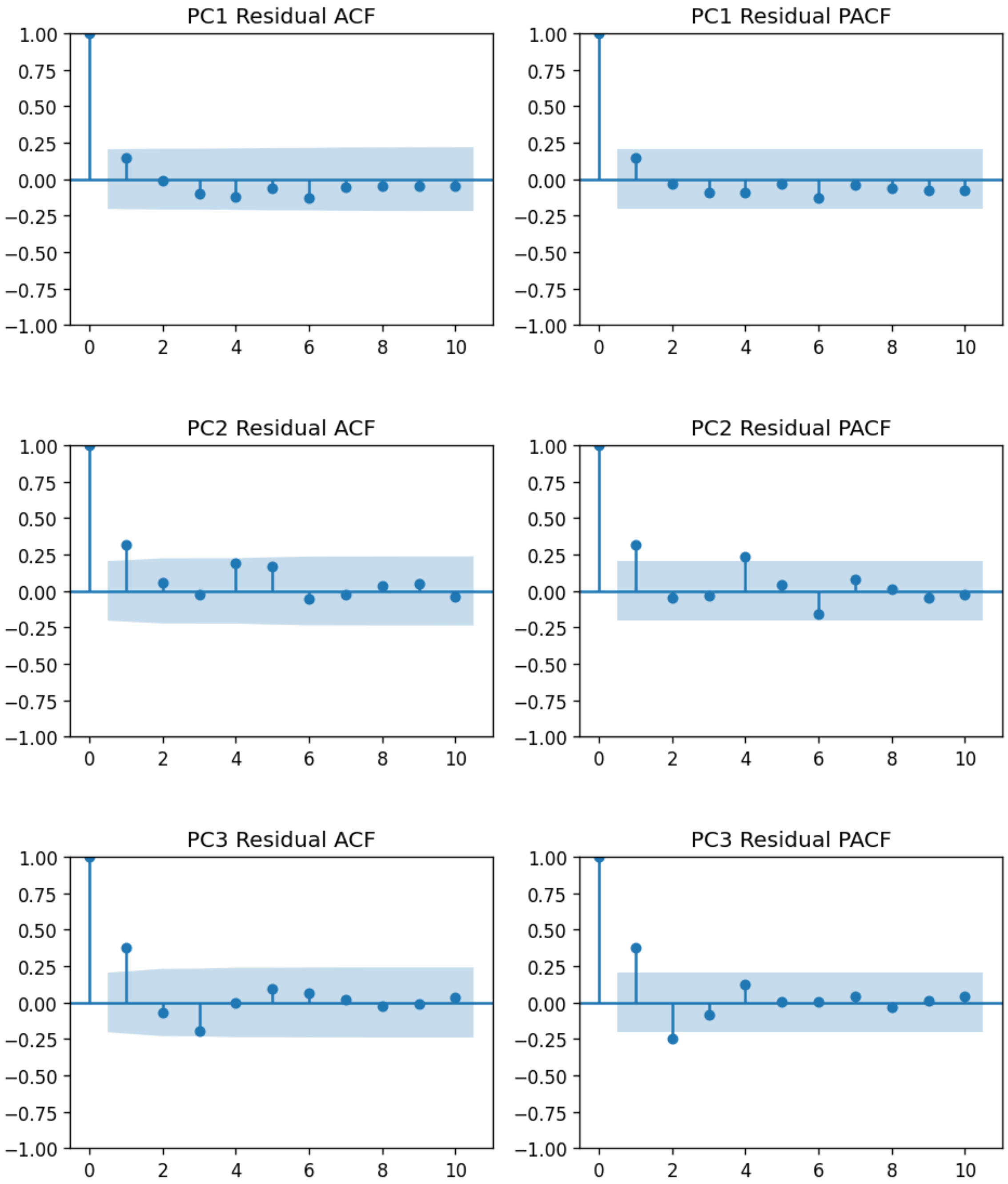
PC #	Ljung–Box Results (p(10) value > 0.05)	Result and Assessment
PC1	*p(10)* = 0.681	*Pass:* No autocorrelation
PC2	*p(10)* = 0.074	*Pass:* Borderline result
PC3	*p(10)* = 0.043	*Failure:* Weak autocorrelation
PC4	*p(10)* = 0.157	*Pass:* No autocorrelation

PC5	*p(10)* = 0.008	*Failure:* Significant autocorrelation
PC6	*p(10)* = 0.001	*Failure:* Significant autocorrelation

Table 6 —Principal Component Ljung Box Analysis—White Noise Assessment

For these results from Table 6, since the test is interpreting residuals rather than raw data, a failure indicates the VAR model did not fully remove autocorrelation present in that component. The leading market factor, PC1, captures most of the economy-wide variation and is well modeled by the VAR. PCs 2–4 exhibit mild or borderline autocorrelation but remain close to white-noise behavior. PCs 5 and 6 fail the Ljung–Box test, indicating meaningful of lag structure and serial dependence remain in these components that the VAR cannot fully absorb/capture. This pattern is not unexpected: in high-dimensional economic systems, later principal components often capture noise, thin-market artifacts, or idiosyncratic item-level shocks.

Because this is a virtual game economy, the presence of light "cyclical" patterns in some residuals (see Figure 14) is also reasonable. This may be due to weekly trading habits, patch-day effects, and daily GE buy-limit mechanics can all induce seasonality not fully captured by the factor structure. Since the VAR forecasts are primarily driven by the first few dominant components, these results indicate that the FAVAR model remains statistically adequate despite autocorrelation in the weaker PCs. Future work could explore higher-order VAR(*p*) specifications, dynamic factor models, or shrinkage-based covariance regularization to better capture these lower-variance components.


PC1 Residual ACF
PC1 Residual PACF
PC2 Residual ACF
PC2 Residual PACF
PC3 Residual ACF
PC3 Residual PACF
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
2
4
6
8
10

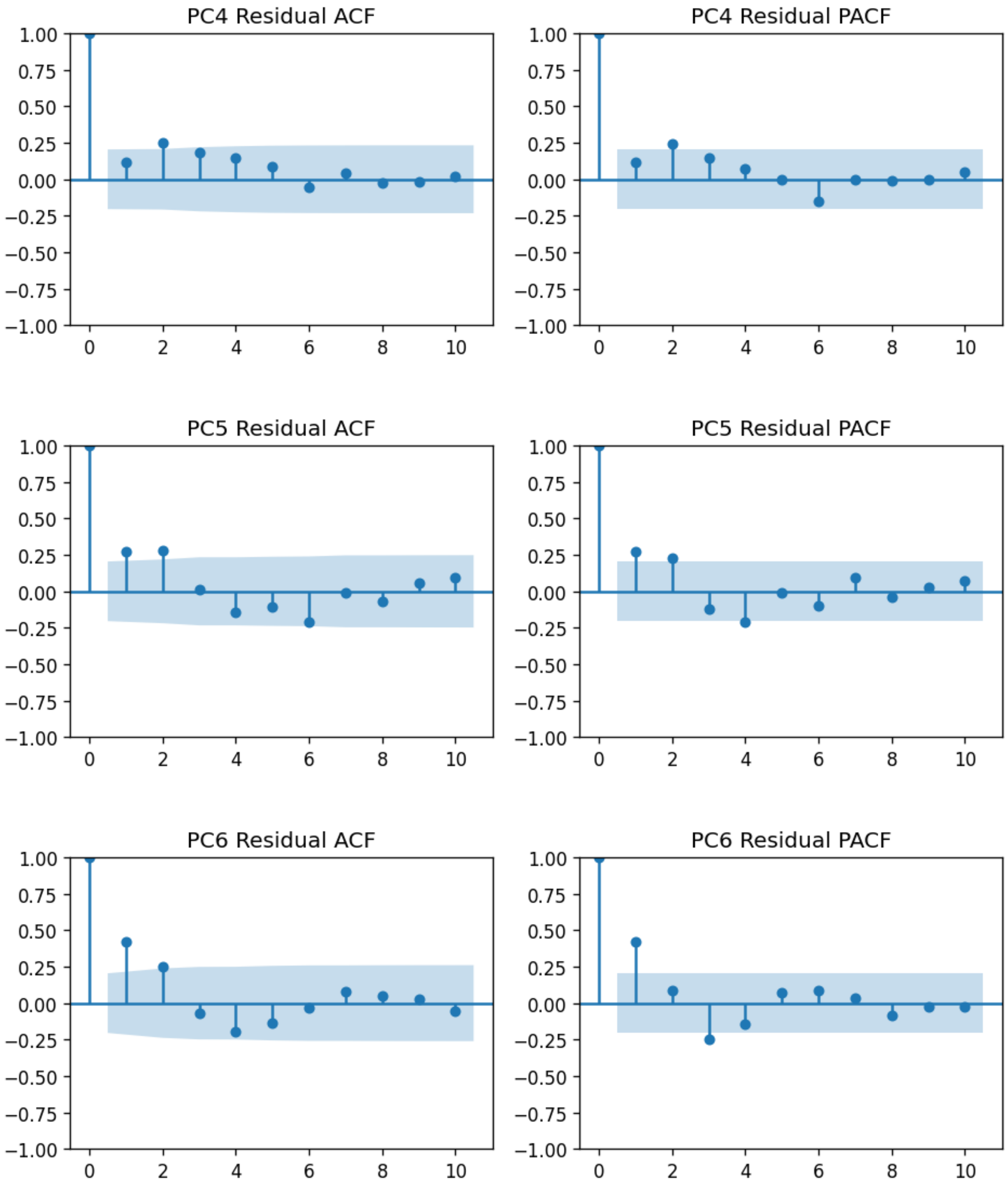


Figure 14- Autocorrelation Principal Component Visualizations

After testing the residuals of the VAR model, a Granger causality check was performed on the first four principal components (PC1–PC4). Granger causality is a standard test used in time-series analysis to evaluate if past values of one series can predict another (beyond what it predicts about itself). In the context of this project, it serves as another way of understanding if certain PCA-derived factors in the OSRS economy are either "leading" or "lagging" each other. This mirrors how real-world financial markets show directional swings in certain sectors from either macro, sentiment or liquidity driven events/factors.

The test consists of creating 'mini' VAR modes (1~2 lags) between each pair of factors (PC1 to PC4) selected for analysis from the training set. This consists of a 'restricted' and 'unrestricted' format in which the restricted uses the factor's own lagged values, while the unrestricted uses both lagged values from both PCs in the pair. After this a likelihood-ratio test is performed that determines if the unrestricted model significantly improves the predictive accuracy from the restricted results with a *p*-value test. The test boils down to determining if one factor between a pair of factors (Principal Components here) helps predict the other. From the results below (Table 7) the results in rows are the predictors and the columns are the predicted series (not symmetrical).

Granger Causality Test (PC1–PC4, lag=1~2)				
	PC1	PC2	PC3	PC4
PC1		0.029	0.243	0.002
PC2	0.000		0.000	0.005
PC3	0.000	0.000		0.436
PC4	0.188	0.000	0.715	

Table 7: Granger Causality Test (p < 0.05 significant Granger-casual influence)

From the results, we have eight factor comparisons of Principal components that pass the test ($p < 0.05$), with four failing records. Out of these passing results, the second principal component PC2 seems to be the strongest as all its p-value test results are almost zero when acting as a predictor. This would suggest that PC2 captures a leading component of market structure that represents broad liquidity trends or other sentiments such as risk-taking speculation cycles that feed into other areas of the economy. PC1 does not help predicting the PC3 series well but when flipped, PC3 scores a significant Granger causality result of zero, meaning it can help predict PC1 very well. This means that there is strong bidirectional interaction going on between the two Principal Components. This may be due to several factors such as co-movement between highly traded commodities, capturing mid-tier markets or groups of assets responding jointly to overlapping macro shocks.

Overall PC4 shows limited causal influence, with only one significant Granger Causality score of zero for explaining PC2, which is not impressive because it is the most robust PC of the group. It has two insignificant results for PC4–PC1 and PC4–PC3 (p = 0.188 & p = 0.715) meaning it has a hard time predicting these PCs. Additionally, when the pairing is flipped (PC3–PC4) it means that PC3 does not help explain PC4 either. However, this aligns with expectations as the lower PCs typically capture idiosyncratic noise or niche sub-markets rather than dominant market drivers.

Tracing back to the Ljung-Box results, PC1 and PC4 showed residuals consistent with near-white noise, while PC5–6 exhibited stronger autocorrelation. The Granger results reinforce this picture in which PC1 behaves as a large, stable macro factor and PC2 acts as a main predictive driver. The higher PCs start to weaken in their signaling for time-based indicators which is expected as the overall percentage of explanation left for PCs to solve lowers.

Because FAVAR forecasting uses only the leading K components ($K = 65$ in the breadth specification), understanding these causal interactions helps justify the model design. The significant predictive structure among PCs (especially PC2 → others) suggests the latent factor system contains real, exploitable time dependence. This supports the use of a VAR(1) specification for factor dynamics, since most meaningful predictability manifests at low lags. The lack of strong causality for PC4 and above is consistent with the expectation that higher components represent noise, seasonal quirks, or isolated item clusters.

Overall, the Granger causality analysis shows that the leading principal components are not independent white-noise signals; instead, they contain directional information that the FAVAR model is structured to capture. This strengthens the case that factor-based VAR modeling is appropriate for the OSRS economy and reinforces the statistical adequacy of the modeling pipeline.

While MAE and RMSE values act as potent discussion points for model performance, the domain of this type of data is ultimately in predicting where prices may go in the future like a stock-equity or cryptocurrency market. To evaluate how forecast differences might translate into actionable economic value, each model's predictions were converted into daily trading signals and passed through a simplified profit-and-loss (*PnL*) simulation. This mirrors common practices (back testing) of equity and crypto forecasting, where directional predictions are translated into cumulative return curves. *PnL* simulation was applied only to GRU/LSTM because they forecast item-level prices directly, which allows for meaningful translation into buy/selling decisions. The VAR forecasts PCA factors, and back-projected factor reconstructions which are not suitable for a realistic buy/sell simulation. Given the project's scope and emphasis on item-level model forecasting, the neural network models were chosen as the appropriate candidates for *PnL* simulation/evaluation.

For each high-value item, both the LSTM and GRU generated a one-step-ahead forecast for the next day's price. A directional trading rule was then defined:

- If forecast(*t*+1) > actual(t) → go long for that day
- If forecast(*t*+1) < actual(t) → go short for that day

Please note—it is technically not possible to 'short' (bet on downward equity movement) in the RuneScape economy and was substituted with a substitution to properly calculate and reward PnL predictions of negative returns in the models.

(See LIMITATIONS: Profit and Loss Simulation)

This forecast directional prediction creates a binary signal $s_t \in \{-1, +1\}$:

$$s_t = \mathrm{sign}(\hat{P}_{t+1} - P_t)$$

This assumption mimics a trader who bets on predicted direction rather than predicted magnitude. The realized daily return of an item itself within this exercise are represented by the following:

$$r_t = \frac{P_{t+1} - P_t}{P_t}$$

This generates simple market return calculations and does not interact with anything outside of price data (log-returns for example). Therefore, the models trading return per day is calculated by the following equation in which correct predictions (either direction) gives a positive PnL score and incorrect gives negative PnL:

$$\mathrm{PnL}_t = s_t \cdot r_t$$

The individual scores are then aggregated by a simple cumulative equation for each trade.

$$CumulativePnL(t) = \sum_{\tau=1}^{t} PnL_\tau$$

In *Figure 13,* the cumulative return curves show how each model's direction accuracy compounds over time. This type of modeling, back-testing, is as previously mentioned, a common type of time series analysis used in financial markets.

Items utilized in this analysis include "*God Wars Dungeon*" equipment such as God-swords (different variants), Armadyl crossbow, Bandos Armor, mixed items mid-range items like Abyssal whip, Toxic Blowpipe, and some ultra valuables like Elysian spirit shield and Tumeken's Shadow. Unlike the *Top Filtered* liquidity subset used for model evaluation (50 items), this portfolio reflects a mixed, high-end but realistic basket of premium assets with clear directional price trends and sufficient daily volume to avoid thin-market distortions.

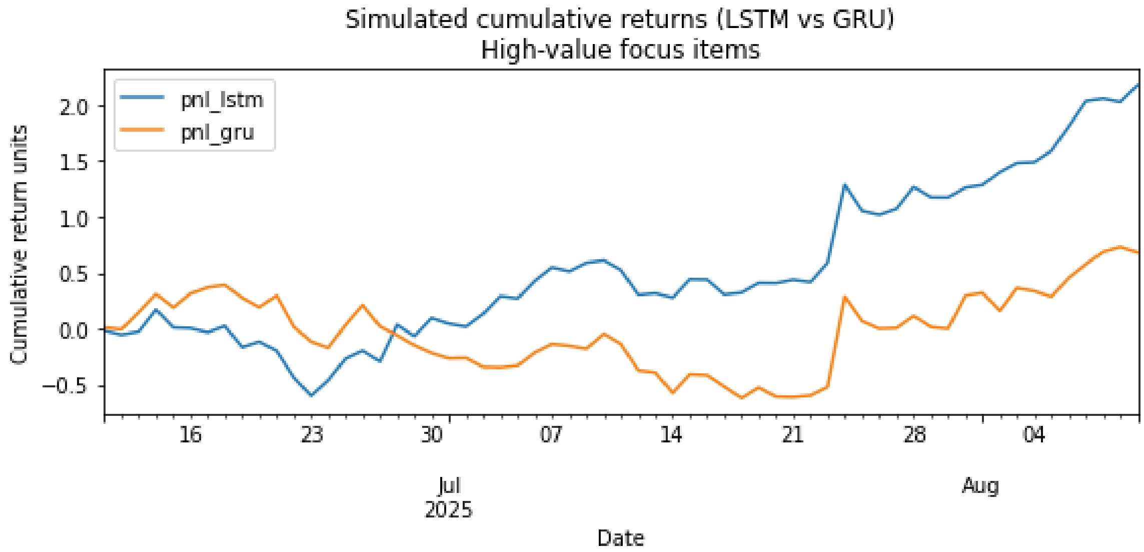


Figure 15 — Simulated Cumulative Returns for LSTM vs GRU on High-Value Items

From the results within *Figure 15,* the LSTM model exhibits a clear and sustained profitability trend within the evaluation period with an ending cumulative return of roughly +2.3 return units. While the model performs well over time, it also shows realistic movements and volatility seen with trading equities, with downward or standstill days. In contrast, the GRU model fluctuates around zero for most of the window, with early gains gradually eroded by a persistent sequence of weak or incorrect signals. This results in an almost zero unit return by the end of the evaluation period. These results suggest that although both models have comparable MAE/RMSE scores, the LSTM model translates its predictions into substantially more stable and profitable trading signals over time. This is an important distinction and justification for additional metrics testing beyond just accuracy/error statistics. The LSTM model has a clear advantage over the GRU model neural network in an applied setting of forecasting item movements within a virtual economy.

To translate these returns units into actual values, one return unit corresponds to a 1× multiple of starting capital which can be seen in the equation below:

$$\text{Growth Factor} = e^{\text{TotalPnL}}$$

Thus, if a trader began with 10 million GP, the LSTM's return at ~2.3 units would represent an approximate 23 million GP position after trading netting about 13M GP. In contrast, GRU model's return of ~0 units would result in a small net loss with trading fees applied. With trading fees calculated at a 2% total tax cost across all trades, net profit for LSTM would be roughly 12.74M GP. Using the October 18, 2025 conversion rate at 0.63$/1M GP, this return would result in an approximately profit of $1.73 USD.

To complement the portfolio-level cumulative return curves, a secondary item-by-item analysis was conducted using total PnL values for each individual asset in Table 8. This table measures how often and how strongly each model generated profitable directional predictions for specific items rather than aggregating all items into a single trading sequence. Therefore, these values should not be confused with the cumulative, unified pooled trading strategy scheme seen in Figure 15.

Item	TotalPnL LSTM	TotalPnL GRU
Zamorak godsword	**0.634**	0.129
Bandos godsword	0.443	**0.541**
Saradomin godsword	**0.373**	0.292
Abyssal whip	**0.369**	0.286
Armadyl crossbow	**0.303**	0.001
Toxic blowpipe (empty)	**0.222**	0.150
Elysian spirit shield	**0.212**	**–0.063**
Tumeken's Shadow (uncharged)	**0.209**	**–0.016**
Bandos tassets	**0.181**	0.018
Bandos chestplate	**0.170**	0.096

Table 8 —Per-Item Cumulative Returns for LSTM vs GRU on High-Value Items

The LSTM model captured strong positive return signals for weapons such as the Zamorak godsword (+0.63 units) and Armadyl crossbow. Conversely the GRU model occasionally outperformed the LSTM at least once with the Bandos godsword by a margin of +0.1 units. GRU had the only two negative returns on only the “ultra” and “mega-rare” items within the subset with *Elysian spirit shield* and *Tumeken’s Shadow*. This does not seem to be an effect of the very high price points of these items as LSTM’s returns are quite solid on these two exact items at roughly +0.21 units each. This dual perspective highlights that while both models can perform well selectively, the LSTM demonstrates far stronger consistency when signals are aggregated into an applied trading-style.

Objective 2 Discussion

To identify groups of items that move similarly, Principal Component Analysis (PCA) was applied to the top 150 most traded items, followed by *K*-means clustering on the resulting factor scores. The PCA score plots show that most of the variation in item behavior falls along a clear curved shape in the PC1–PC2 space, which suggests that a small number of underlying forces explain most cross-item price movement.

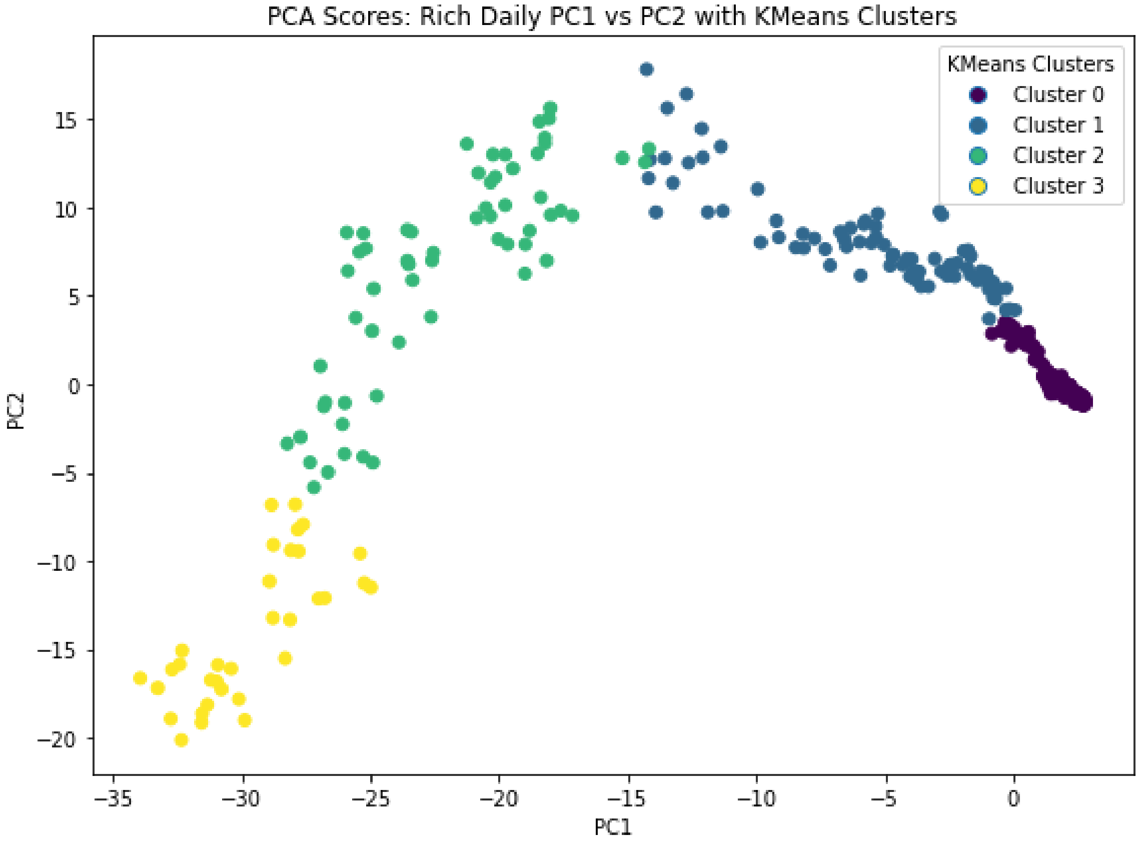


Figure 16: Rich Daily PC1 vs PC2 Comparison

Coloring the scores by PC3 shows a smooth gradient across this curve in Figure 17, which indicates that item behavior changes gradually along these factors rather than forming sharp or isolated regimes.

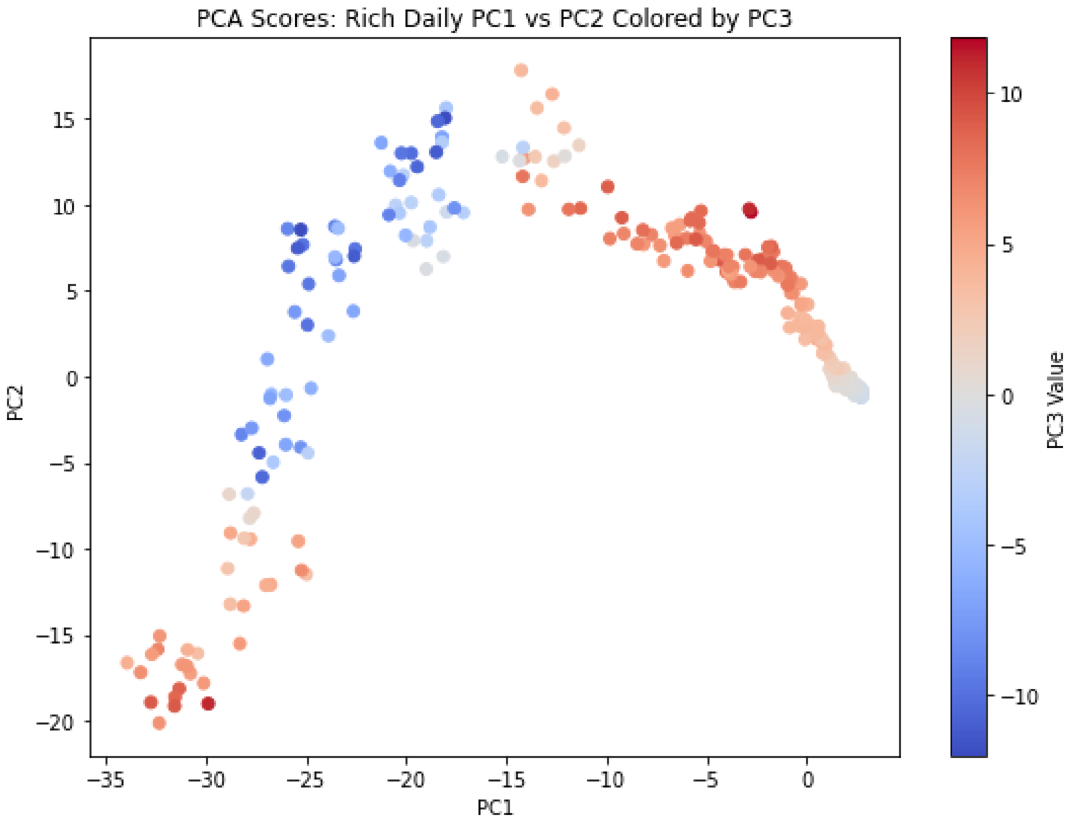


Figure 17: Rich Daily PC1 vs PC2 Comparison with PC3 color overlay

Applying *K*-means clustering to the PCA scores produced several clearly separated groups of items. In the PC1 vs PC2 (*Figure 16)* and PC2 vs PC3 (Figure 17) projections of the *Rich Daily* dataset produced four to five clusters emerging with meaningful separation. This indicates that items inside each cluster share similar price-movement or volatility patterns. These clusters broadly align with intuitive item types, such as stable consumables, skilling materials, high-volatility rare drops, and event-driven league items. Importantly, the same general grouping structure also appears when repeating the analysis with the macro dataset (using loadings), suggesting that these behavioral sectors are stable across different sampling resolutions.

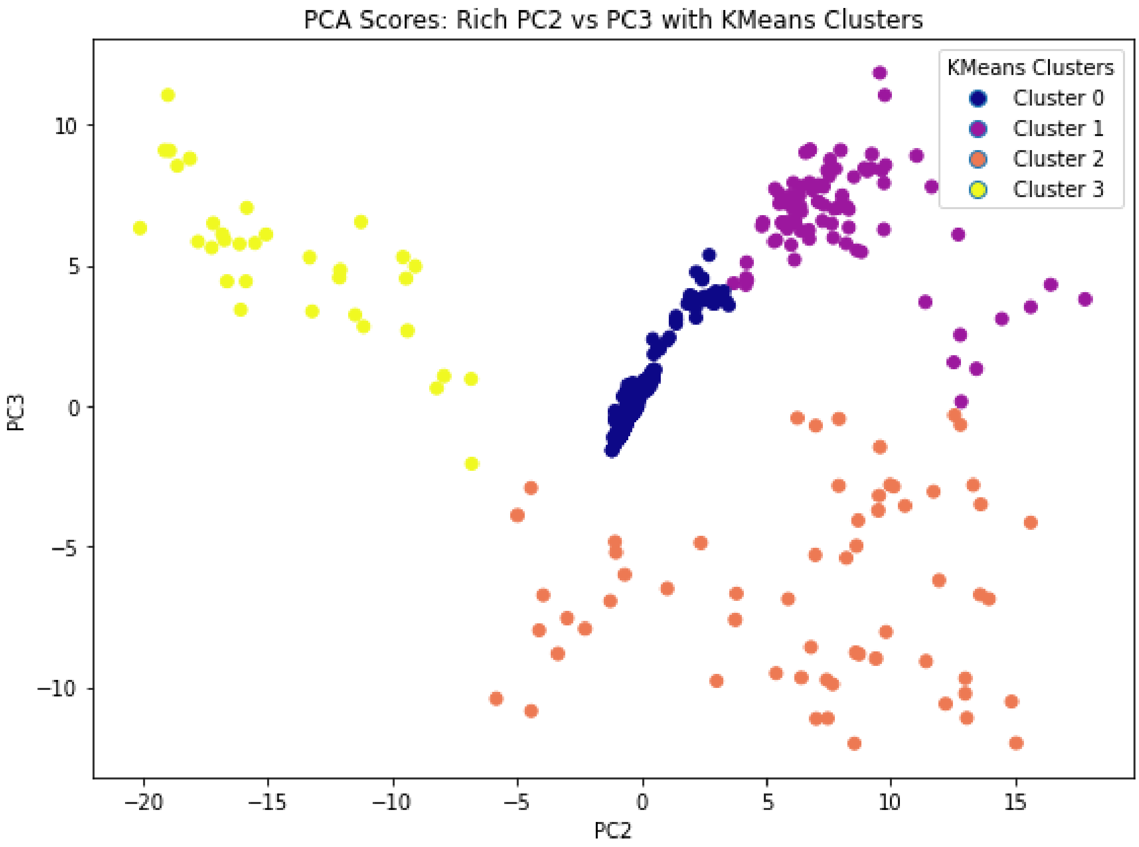


Figure 18: Rich Daily PC2 vs PC3 Comparison

The PCA loadings reveal several intuitive behavioral sectors. For interpretability, only a small subset of items is labeled in each plot, selected by taking the top and bottom scoring items along each principal component within the top 150 macro-frequency items. Highlighting roughly 10 items per direction ensures that the most influential contributors to each factor are shown without overwhelming the figure or obscuring the underlying cluster structure.

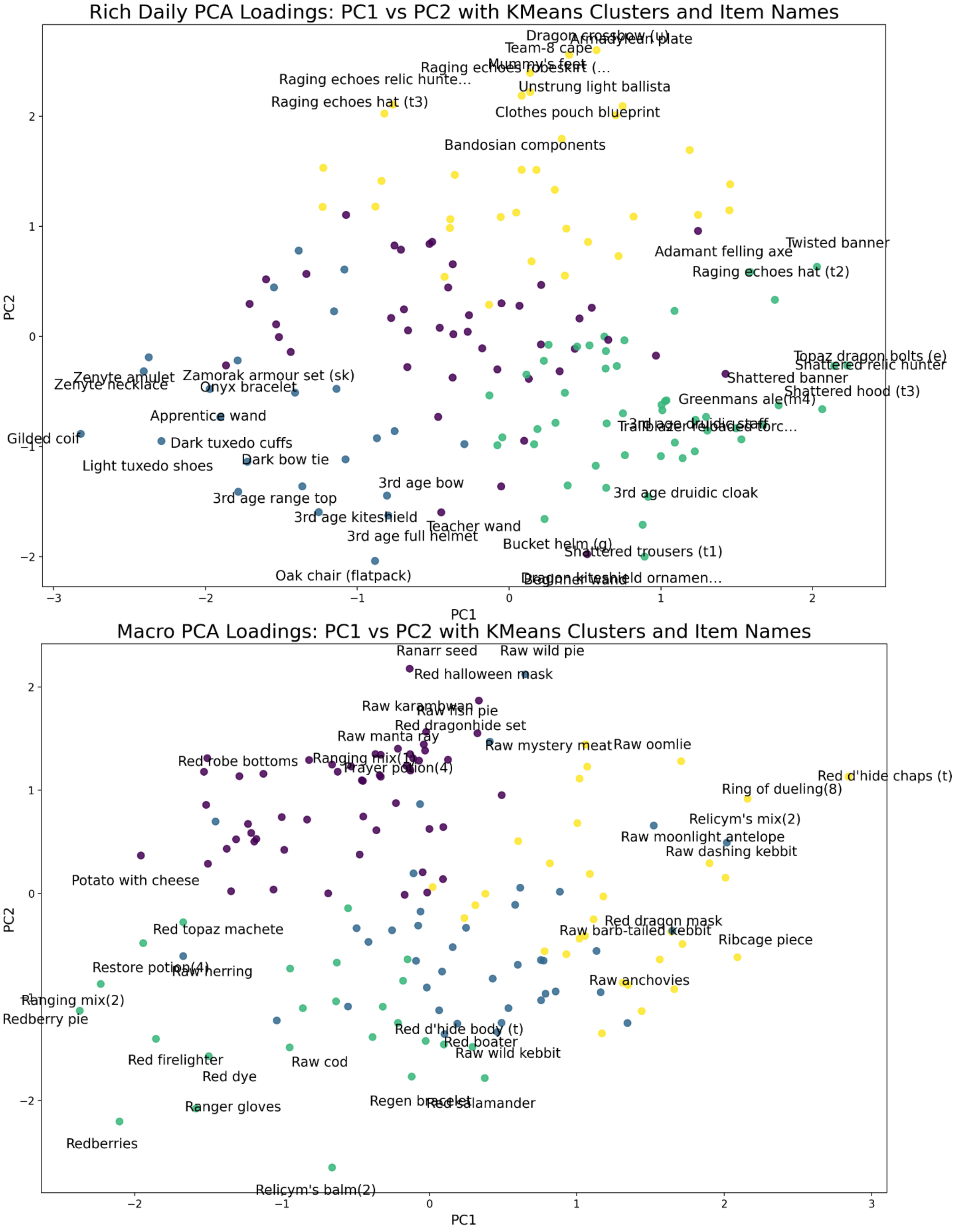


Figure 19: Rich Daily & Macro PC1 vs PC2 K-means Clusters with 'Standouts' Labels

In the rich-daily dataset plot (Figures 19, top) high-tier jewelry such as Zenyte amulets and Onyx jewelry form a stable high-value cluster, while third-age items group tightly in the lower PC2 region due to extremely low supply and large volatility spikes. Event-driven items from Leagues, such as Raging Echoes armor sets, appear in the high-PC1 and high-PC2 region, reflecting sharp but temporary activity spikes. In contrast, the macro dataset (Figure 19, bottom) produces broader long-run groupings, including a skilling-supplies sector dominated by raw food items or contain the phrase “raw” (~35% of entries), a potion and restoration cluster, and a group of hunters and Ranged materials tied to long-term training meta changes. This high concentration of similar naming conventions for different items, specifically by the prefix “raw”, reflects how OSRS naming conventions are standardized rather than some algorithmic bias. This also demonstrates that long-run market structure is heavily influenced by basic resource-chain items, especially food (healing items when cooked).

The clustering results in Objective 2 were generated using a hard *k*-means assignment, where each item was placed exactly into one behavioral sector. Initially, no misclassification analysis was possible because OSRS provides no predefined or “true” economic categories in which cluster assignment could be judged against. The clusters themselves were the first attempt at defining latent sectors purely from long-run price movements.

To approximate a validation step, a coarse set of sector labels (Equipment, Food/Potion, Resource, Other) was constructed from the macro dataset using item names and compared against cluster membership. For example, the “equipment” classification looked for Item Name terms representing weapons and armor such as “sword”, “shield”, "helmet". Similar logic was applied to the other cluster labels respective to their domains. This heuristic comparison does not constitute a true classification accuracy assessment but provides a qualitative diagnostic of cluster alignment. Additionally, these tags were not utilized during clustering and serve as post-hoc diagnostic only. The results of these cluster labels can be seen in Table 9 with the counts in the first table and the proportion by classification in the second table.

Counts				
Cluster	**Equipment**	**Food/Potion**	**Resource**	**Other**
0	7	17	13	17
1	7	7	14	12
2	3	4	6	12
3	6	4	14	7
Proportion within Cluster				
Cluster	**Equipment**	**Food/Potion**	**Resource**	**Other**
0	0.13	0.31	0.24	0.31
1	0.18	0.18	0.35	0.3
2	0.12	0.16	0.24	0.48
3	0.19	0.13	0.45	0.23

Table 9: Classification by count and proportion for cluster 0-3 for the Macro dataset.

Most clusters show mixed proportions, with many categories clustering around ~0.30, meaning roughly 30% of items in the cluster belong to each of those categories. This consistency reflects the limited expressiveness of the coarse label, with even proportion spread across the item categories indicating weak separability among clusters. However, Cluster 3 does show an enriched proportion of the "resource" tag at roughly 45%, suggesting that the cluster may favor items derived from skilling activities via raw resources exhibiting similar price movements. Cluster 0 and 1 show a much more heterogenous mix of resources, Food/Potion and 'Other' items. Cluster 2 shows the highest proportion of 'Other' items, which likely reflects items that didn't fit within the coarse labels system that was applied for each classification type. Additionally, in all clusters except Cluster 3, the ‘Other’ category exceeds 0.30, which aligns with the expectation that the heuristic tags used in this pseudo-validation check were not exhaustive. The results of this do not represent true classification accuracy but rather provide a qualitative check that at least one *K*-means sector (Cluster 3) captures a meaningful structure consistent with the in-game virtual economy.

Overall *k*-means cluster structure quality was determined by conducting silhouette score analysis for clustering performed on the rich daily PCA scores and macro PCA loadings. The rich daily scores were used to represent market regimes (days with similar system-wide price behavior) and macro loadings to represent item-level sectors based solely on long-run covariance structure. For each dataset, *K*-means was run with clusters $K = 2 - 10$ since the metric requires at least two clusters to compute a valid comparison. The silhouette score was calculated using the formula below:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

The silhouette score for each item is computed as the difference between its nearest-cluster and within-cluster distances, normalized by the larger of the two to yield a value between –1 and 1 indicating clustering quality.

Cluster	Rich Data Scores	Macro Data Loadings
K	Silhouette Score	Silhouette Score
2	0.876	0.284
3	0.873	0.287
4	0.866	0.311
5	0.850	0.332
6	0.837	0.320
7	0.800	0.326
8	0.804	0.319
9	0.804	0.324
10	0.564	0.311

Table 10: Silhouette scores of clusters 2 through 10

Silhouette analysis showed that the regime clustering within the rich daily PCA scores yields high-quality, well-separated clusters (0.80–0.88). Conversely, the macro PCA loadings exhibits lower but stable silhouette scores (0.28–0.33), which is expected for item-level covariance structures where many items share overlapping demand drivers and do not form sharply separated clusters. These results suggest that daily market regimes form a strong latent structure, whereas item sectors represent softer, economically meaningful groupings rather than sharply defined categories. This finding of 'softer' sectors was consistent with the coarse classification results, in which most of the clusters had relatively even spread with a median proportion of 0.30 among the four assigned categories.

A natural extension of this work would be to explore soft clustering approaches, such as Gaussian Mixture Models or fuzzy *c*-means, which provide probabilistic membership and may better capture items whose price behavior legitimately spans multiple latent economic activities

Objective 3 Discussion

Bai-Perron structural break framework is a widely used econometric method for detecting breakpoints in time-series data. The model estimates a linear regression and when the underlying data exhibits shifts in the mean, trend or autoregressive structure. The algorithm searches over potential partitions of the series and selects the configuration that reduces the total residual sum of squares using the Bayesian Information Criterion (BIC). This methodology lets the model detect if multiple breaks are present without needing pre-fed event dates or “hotspots” to look out for within the given date range for the data. This is a very appropriate model for data from a virtual economy as it is commonly used in economic or financial settings for seeking policy shocks, market crashes or structural shifts. In this project, the test was used to evaluate whether major game updates correspond to the occurrence of statistically identifiable breaks in price levels.

In this project, a Bai–Perron–style mean-shift break procedure was implemented using the ruptures library (PELT, L2 cost), which matches the Bai–Perron objective of identifying optimal segment partitions under a BIC-type penalty but uses a faster modern search algorithm. While conducting the first investigations into structural-break analysis, unexpected but important results emerged when plotting price levels of Yama & Doom items. Initially items related to the two new bosses seemed attractive candidates for event and structural break analysis due to dramatic price moments. However, the plots revealed that neither group contains sufficient real price history for valid structural-break testing. Items such as the Oathplate armor pieces (Helm, plate-body, plate-legs) and Demon tear were known to be newly introduced items, but the scale of which data were available to the project was below expectations. Items related to the 'Doom' boss were found to only start being tracked by via the GE within 3 days of the closing of the training window on July 23rd (Figure 20). Conversely 'Yama' related items had a larger window more closely representing the rich dataset with the first GE tracked data beginning on May 15th, covering most of the rich data's range at ~75 days (Figure 21). These missing data ranges as seen within these two plots, were passed through early Data Prep stages as N/A values were removed with the rich sets, and the macro set as seen below had a median value backfilled by the API which caused the discrepancies to go undetected at earlier steps.

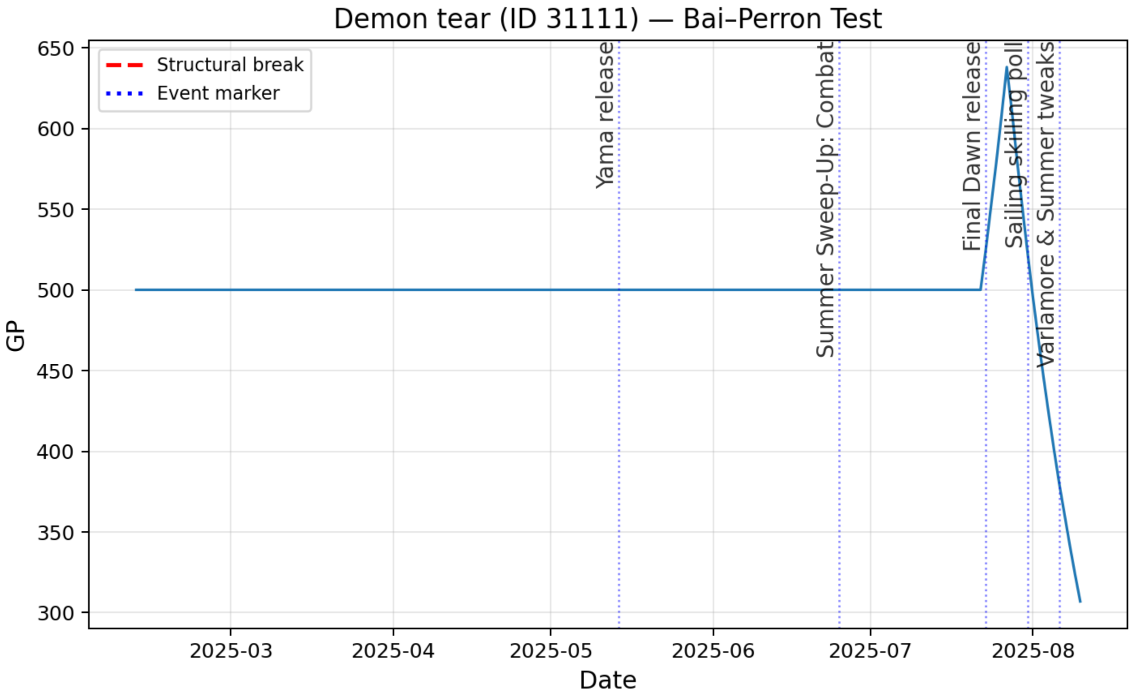


Figure 20: Demon Tear Bai-Perron Plot—insufficient available data.

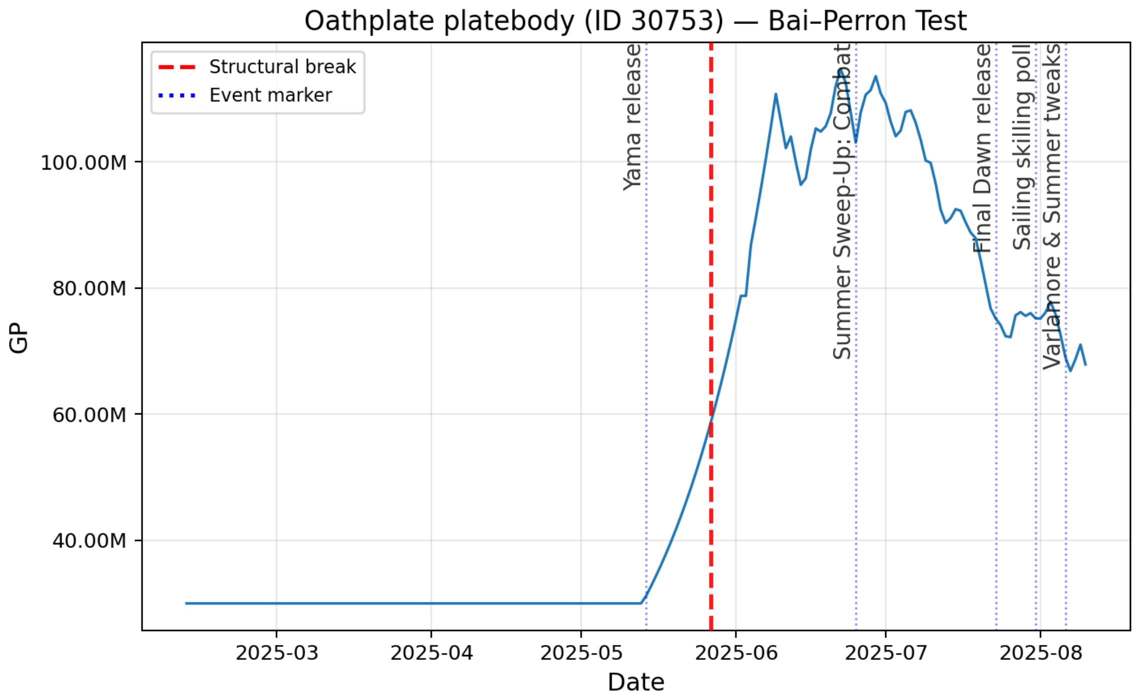


Figure 21: Oathplate plate-body Bai-Perron plot—false positive break.

While the ‘Yama’ items like “Oathplate plate-body’ seem to have a reasonable data window, especially when compared to ‘Doom’ items, both items are not suitable for analysis within this section. Bai-Perron break detection (main diagnostic in this objective discussion) assumes a broad continuous range of economical meaningful time series data in which these items violate that prerequisite. An example of this violation effecting the results is a false positive that occurs within the Bai-Perron test of Oathplate plate-body in Figure 21. The API backfilled data at 30M GP is incorrectly picked up by the Bai-Perron test as a “historic” stable price level that goes from the start of the data window to the actual release to the game. Therefore, when the price rapidly evaluates after releasing it gets picked up as a false positive, as the model thinks it was trading on a much larger timeframe than reality. These distortions, amplified by the limited windows of the

‘Doom’ items, explain why ‘Demon Tear’ surfaced as an extreme outlier within Objective 1. This finding confirmed the issue that was previously thought to be model misbehavior from a price shock but data fragmentation inherent to these items introduced so closely to the end of our dataset range.

For all these reasons, Yama and Doom items were excluded from structural break analysis within Objective 3 discussion. The final analysis instead exclusively focuses on items that have full coverage within the February 11th to August 10th range of the macro dataset and overall project. This decision allows for all observations to have a long enough market presence before the announcement date could affect them and long enough afterward to allow for recovery, stabilization, or reinforcement of the price shift. Items with long-standing histories, meaningful pre-event baselines, and enough length to allow the break tests to detect whether announcements, pre-release hype, or post-release supply shocks shifted their behavior.

After filtering out newly introduced items and confirming long-horizon price coverage, the Bai–Perron structural break test was applied to a set of long-standing assets with continuous price histories from February 11th to August 10th. Out of the 50 items evaluated, three produced statistically meaningful breakpoints (true positives): Soul Rune, Burning Claws and Granite Maul. These items all share two market-wide properties that make them strong candidates for event-driven shifts. They all have large active player markets (strong volumes of trade) and they interact directly with changes that occurred in the meta, combat balance and boss-related content within this period.

Items that produced no-breaks (most of our tested records at 46) generally have smooth adaptation to supply and demand pressures with no sudden price pivots. These serve as the control of the analysis and that the model is not overfitting on small changes or noise being misidentified as significant structural breaks. A great example to show this is the item *Tormented Synapse* (used to make the strongest demon-bane weaponry), which saw a significant price increase leading up to the "Yama" boss fight since its monster typing was demon. However, even with a large price increase and decrease, it did not reflect an actual structural break in the item's behavior (Figure 22)

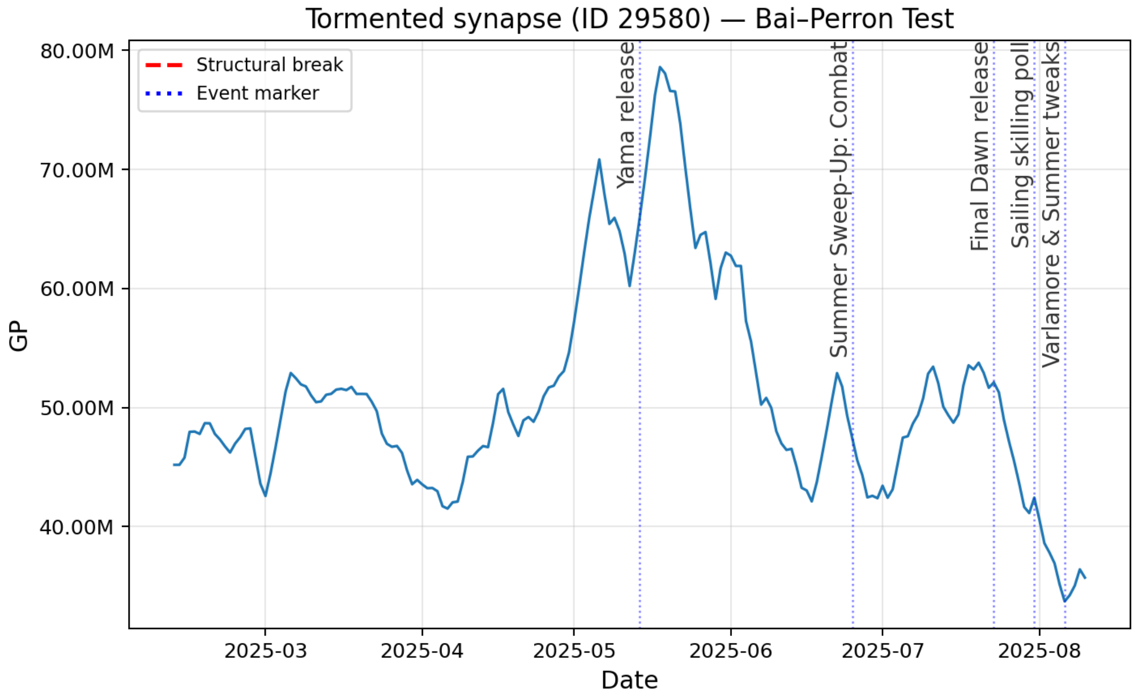


Figure 22: Tormented Synapse Bai-Perron Test plot—no break

The Soul Rune (*ID 566*) plot (Figure 23) shows one of the clearest structural breaks in the entire dataset. The detected breakpoint on May 27th aligns tightly with a surge in demand driven by the post-Yama meta utilizing "*Dark Demonbane*" and "*Mask of Darkness*" in combination to fight the boss or farm *Tormented Demons* for farming Tormented Synapses. The widespread adoption of these high-tier spell combinations caused Soul Runes to be required in very large quantities, with buy-side pressure accelerating sharply after players adapted to the new boss and other demon-type monster farming methods. Prices jumped from ~180 GP to over 350–450 GP in the weeks following this transition. The break here reflects an abrupt demand-side regime shift, where the rune moved from a stable low-variance commodity to a bottleneck input for an extremely popular bossing strategy.

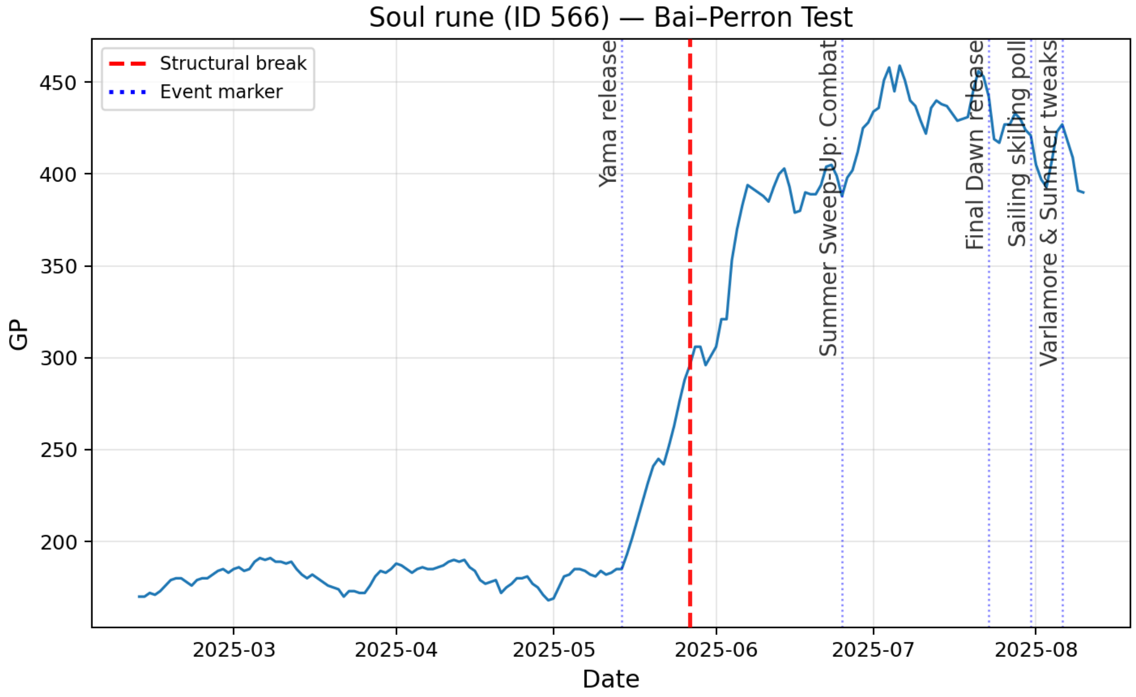


Figure 23: Soul Rune Bai-Perron Test plot with structural break

Burning Claws (29577**)** are a popular ‘special attack’ weapon that are popular for its high burst damage and lingering burn effect. exhibits a pronounced break on June 6th, closely following Yama release and preceding the early-June Summer Sweep-Up combat adjustments. While not tied directly to a single announcement, the timing corresponds to a shift in melee meta incentives and player opinions. The item shows a clean transition from a broad sideways band between February and early June into a steady decline afterward. Unlike speculative spikes, this pattern reflects a durable reevaluation of the item's future combat relevance. From a purely statistical standpoint, the break suggests that player expectations about weapon viability changed sharply during this patch window and remained consistent afterward. With additional context from a player perspective, the claws are demon-bane weapons and were speculated to be the optimal 'special-attack' weapon at the Yama boss fight. However, a unique and powerful melee strategy became popularized known as '*DONOFLY*' by OSRS content creators that once memorized offered a consistent, lower risk strategy specifically for the final phase of the fight. Implementation of this method reduced the desire for special attack weapons to quickly rush down the boss and may have contributed to changing player sentiment which was lined up with the location of the breakpoint for Burning claws plot.

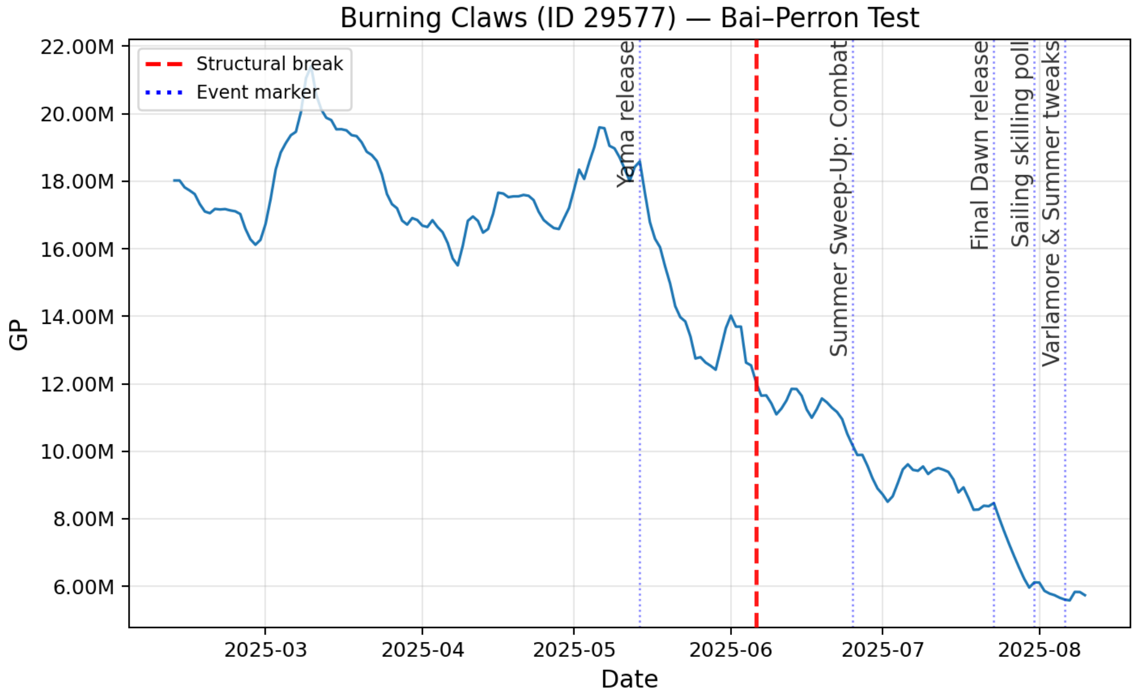


Figure 24: Burning Claws Bai-Perron Test plot with structural break

The Granite Maul (ID 4153) is a popular PvP special attack weapon that can unleash a flurry of quick blows and is commonly used as a 'KO' or knockout weapon. Its break on May 12th occurs immediately before the official Yama release window. The item shows a clear pattern: moderate volatility early in the series, followed by a sharp downward shift as the PvP environment changed and other factors. While the break occurs almost exactly at the same time as the Yama update, no changes were made to it or "Gargoyle", the *slayer monster* that drops them. It seems that the broader combat changes introduced in the Spring update window including accuracy tuning, damage distribution adjustments, and PvP balance tweaks were more likely responsible. Demand for this class of gimmick or burst-centric "knockout" melee weapons most likely dropped from these changes. The break reflects a structural move away from older PvP staples as the patch cycle directed players toward newer, more consistent meta tools. Additionally, the Gargoyle slayer monster that drop Granite Mauls as a reward are commonly targeted by botting accounts trying to farm GP and may have contributed to its drop as well.

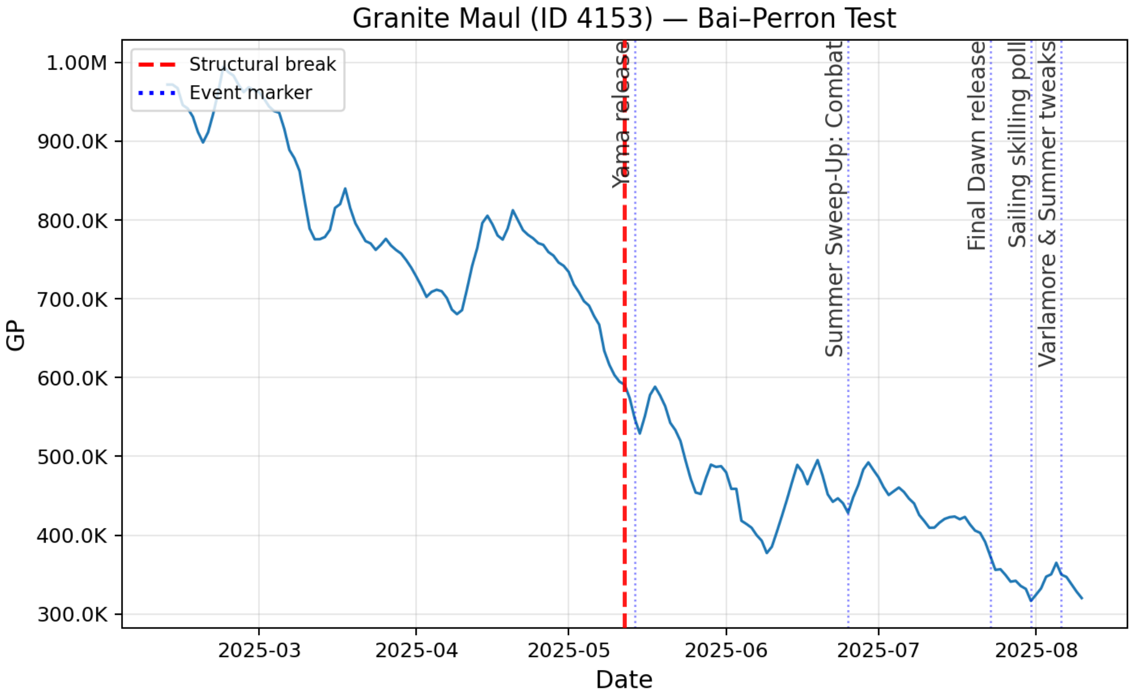


Figure 25: Granite Maul Bai-Perron Test plot with structural break

Across all items tested, the structural break points point toward a consistent pattern. The most meaningful shifts occurred when combat balance patches, meta-adjustments or new boss strategies redefined the expected value of long-standing weapon staples and resources. Additionally, three major conclusions can be made around the breaks witnessed in this Objective. Many breaks occurred and cluster around the Spring/Summer Combat updates, in which many were discovered overtime and were not always inherently obvious from the patch note headlines. The next conclusion was that long-standing, high liquidity items (Soul Rune) respond more cleanly to structural tests compared to other item types. This was a surprising find as the initial items screened for potential breaks were assumed/selected from the most volatile, most-meta items (mega-rares and super high-level expensive equipment), which ended up being an incorrect assumption. The last conclusion from this objective found was the most items showed no breaks, which reinforces that the Bai-Perron test is detecting specific, meaningful transitions rather than picking up on noise. Stability found within most weapons, runes and consumables indicates the economy absorbed many updates incrementally.

Phase 7: Deployment Phase

The deployment stage focuses on packaging the entire project into a reproducible and transparent form suitable for academic review and future extension. All cleaned data, code and excel file outputs used in the project are uploaded to a public GitHub repository (https://github.com/Drew-Kitik). This includes environmental files, data-preparation scripts, and Jupyter/Spyder-ready versions of each modeling phase, ensuring that any reviewer can reproduce every figure and table generated in the study. Comments within the code mention specific lines that require modification (working directory) so the end-user can run and reproduce the code as written. By consolidating the workflow into a documented, version-controlled platform, the project meets the core reproducibility standards expected in applied data science while also providing a foundation for future research on virtual economies.

From an applied perspective, the results of this study illustrate several key takeaways for OSRS economy design. The strong stability observed in most long-standing items, contrasted with sharp structural breaks around major content patches, mirrors the behavior seen in real-world markets such as crypto, commodities, and equities during policy or technology shocks. High-liquidity assets adapt smoothly, while items tied to meta shifts or supply shocks undergo abrupt repricing. These parallels reinforce the relevance of virtual economies as controlled microcosms for economic research. Virtual economies exhibit many of the same volatility regimes, behavioral reactions, and event-driven dynamics as real financial markets, but in a contained system with transparent rules. The study's findings could help inform game designers about the economic impact of updating cadence, supply creation mechanisms, and unintended volatility introduced by content releases.

Together, the findings demonstrate that virtual economies can mirror the structural, behavioral, and statistical properties of real financial systems, highlighting their utility as controlled environments for economic modeling and methodological experimentation.

LIMITATIONS

While this project is designed to provide meaningful insights into the structure and behavior of the Old School RuneScape economy, several limitations must be acknowledged:

Historical Data Availability and Item Coverage

The Grand Exchange API provides a maximum of 365 data points with a ‘limit’ of 180-day history depending on the time series selected to be retrieved (5m, 1h, 6h, and 24h from */timeseries* selection). While the */mapping* endpoint returns all tradeable item IDs from this, many items do not generate a complete time series. Ultra-low-liquidity items may go days or weeks without trading, and these inactive timepoints don’t toward the 365 record (entries not days) cap. At the time the data were pulled from the API, some items had data stretching back to 2021. This produces inconsistent sample lengths across items and violates the intended 180-day analysis using the largest available time series interval at 24 hours, despite the API’s nominal limit of 180 days of data. To standardize coverage, the project scope was restricted to February 11th–August 10th, 2025, ensuring that both macro and rich datasets operate within the same clean 180-day window and preventing inactive items from distorting the modeling sample. Therefore, this project cannot analyze long-term cycles, multi-year inflation, or decade-scale asset value due to these restrictions.

Structural Fragmentation in the 6-Hour Dataset

The 6-hour rich dataset, while initially promising for more granular data analysis, was found to be structurally incomplete. The 6-hour dataset, before being aggregated to *rich_daily*, was found to have 11.8% of item-days lacking more than 50% of their expected 6-hour intervals entries. This type of intraday fragmentation would require heavy interpolation at the sub-daily level. This would artificially reshape the return dynamics the FAVAR modeling is trying to pick up on by inflating cross-item variance, distorting covariance structure and reduces reliability of PCA factor extraction. The project scope was limited to match the API intended limitation of 180-days of data matching the macro datasets intended range and capping the rich dataset's fragmentation. Because both VAR and neural network models depend on consistent sampling intervals and reliable factor structure, the 6-hour dataset was deemed unfit for modeling. It was therefore used only in Phase 3: Exploratory Data Analysis (EDA), while all forecasting relied on the daily aggregated dataset (r*ich_ daily*), which provides stable, well-formed 24-hour observations.

Model Generalizability

Models such as VAR or LSTMs were tuned for RuneScape-specific economic dynamics and are biased toward the most liquid, frequently traded assets for optimal model performance. While parallels to real-world systems are possible, game mechanics (e.g., developer updates, bot bans, drop rates) impose exogenous shocks that have no direct analogue in real-world markets. Thus, price changes and shocks cannot be made directly comparable specifically for speculative changes of asset prices in external systems.

Profit and Loss Simulation

The closest equivalent to “shorting” within the OSRS economy within the game is selling an item you already own in anticipation of a price drop. Therefore, signals with a negative sign should be interpreted as selling (or avoiding buying) an item in anticipation of a price decline. The simulated profits represent avoided losses rather than realizable gains from borrowing and short selling. Additionally, since the profitability back-test is not a fully fleshed out trading simulator, the aggregated cumulative returns for each model are not directly translatable to actual GP. Instead, they are calculated on this basis:

$$e^{\text{TotalPnL}}$$

Examples:

A cumulative PnL of 0.50, would be a growth factor of or of $e^{0.5} \approx 1.65$x.

A cumulative PnL score of –0.30 would be to $e^{-0.3} \approx 0.74$x, meaning a 26% loss relative to starting capital. Since there is no starting pool ‘give’ to each model, the translated returns is a general indicator performance for profitability, not a robust model.

Computational and Storage Constraints

With over 4,000 tradable items, the dataset result from the code can easily approach the limits of common storage formats (Excel’s row limits). The creation of the ‘rich’ dataset encountered this issue and required an additional branching of code to fill out the remaining items of the catalog. This caused the dataset to be split across three different excel files, which slightly increases the complexity in downstream processing and coding analysis.

Behavioral and Exogenous Factors

Unlike real economies, RuneScape's economy is heavily affected by developer interventions (new quests, items, balancing changes) and by player-driven behaviors such as speculation via hoarding or cheating through botted accounts flooding the market with common, low-requirement resource gathering. These shocks are not always quantifiable in the dataset, limiting explanatory power when unexpected volatility occurs.

Clustering Classification Limitations

Because OSRS has no "true" item sectors, formal misclassification could not be computed. A coarse label set (Equipment, Food/Potions, Resource, Misc) was used only as a qualitative, post-hoc check and did not serve as an accurate benchmark. Additionally, hard *k*-means forces single-cluster membership, limiting its ability to capture items whose economic roles naturally overlap.

Summary of Limitations

By acknowledging these topics upfront, the project frames its contribution appropriately: as a scalable, robust framework for short- to medium-term economic modeling in virtual environments, with clear awareness of where the results may or may not generalize. These limitations do not undermine the validity of the study but instead clarify the scope within which meaningful insights can be drawn.

REFERENCES

Belaza, A. M., Ryckebusch, J., Schoors, K., Rocha, L. E. C., & Vandermarliere, B. (2020). On the connection between real-world circumstances and online player behaviour: The case of EVE Online. *PloS one*, *15*(10), e0240196. https://doi.org/10.1371/journal.pone.0240196

Castronova, E. (2002). On Virtual Economies. Game Studies, 3(2). Doi: 10.2139/ssrn.338500. https://www.researchgate.net/publication/4811957_On_Virtual_Economies

Chun, S., Choi, D., Han J., Kim, H.K., & Kwon, T. (2018). Unveiling a Socio-Economic System in a Virtual World: A Case Study of an MMORPG. *In Proceedings of the 2018 World Wide Web Conference (WWW '18) (pp. 1929–1938). International World Wide Web Conferences Steering Committee.* https://dl.acm.org/doi/10.1145/3178876.3186173

Dubey, P. (2025). Crypto Market Cap Hit Highest Levels Since 2021 in Q3: CoinGecko. *Coinspeaker*. https://www.coinspeaker.com/crypto-market-cap-highest-since-2021-q3/

Hogan-Hennessy, S., Xenopoulos, P., & Silva, C. (2022). Market Interventions in a Large-Scale Virtual Economy. https://arxiv.org/abs/2210.07970

Jagex Ltd. (2013). *Old School RuneScape [Video game]. Jagex Ltd. https://oldschool.runescape.com*

Larose, C. D., & Larose, D. T. (2019). *Data Science using Python and R*. Hoboken: Wiley.

Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics* (2nd ed.). Wiley.

Nazir, M. & Lui, C. (2016). A Brief History of Virtual Economy. Journal of Virtual Worlds Research. 9. 1-23. 10.4101/jvwr.v9i1.7179. https://www.researchgate.net/publication/303698171_A_Brief_History_of_Virtual_Economy

Niedens, L. (2025). *Gold prices continue to break records. How much higher can they climb? Investopedia.* https://www.investopedia.com/gold-prices-continue-to-break-records-how-much-higher-can-they-climb-11831908

Old-school RuneScape Wiki Community (n.d.-a). *Old school bond information [https://oldschool.runescape.wiki/w/Old_school_bond#cite_note-Bonds-1]*

Old School RuneScape Wiki Community. (n.d.-b). *OSRS item mapping and timeseries data schemas.* https://oldschool.runescape.wiki/w/RuneScape:Real-time_Prices

Old School RuneScape Wiki Community. (n.d.-b). *OSRS item mapping and timeseries data schemas*. https://oldschool.runescape.wiki/w/RuneScape:Real-time_Prices

RuneScape Wiki Community (n.d.) *Application Programming Interface and API information.* https://runescape.wiki/w/Application_programming_interface