510 WRITE UP: Dataset Project

IMPORTANT LINKS:
Hosted dataset: https://huggingface.co/datasets/TMVishnu/watch-market-gnn
Dataset code: https://github.com/calicartels/watch-market-gnn-code
Interactive EDA: https://incomparable-torrone-ccda90.netlify.app/

I) Description of problem/motivation you are solving with this dataset

The pre-owned luxury watch market presents a unique ecosystem where traditional recommendation systems fall significantly short. The secondary market operates under complex dynamics that simple price-based or feature-based recommendations cannot adequately capture.

I’ve tried listing most of the dynamics that I’ve taken into consideration while building this dataset:

- **Condition-Based Value Dynamics:** In the pre-owned sector, a watch's condition significantly influences its value and comparative position. Ex, A traditional recommendation engines might suggest a Rolex Submariner from the 1990s to a potential buyer but fail to recognize that an Excellent condition Omega Seamaster from the same era might represent a more relevant recommendation based on value proposition and market position.

- **Temporal Price Behaviors:** Watch values in the secondary market don't follow linear depreciation patterns. Certain models appreciate over time, while others depreciate, and these patterns vary significantly across brands and model lines. For example, certain vintage Patek Philippe models consistently appreciate, while similar-aged pieces from other luxury brands might maintain steady values or even depreciate.

- **Inter-Model Value Relationships:** Complex value relationships exist between different models that transcend simple brand hierarchies. For example, a discontinued steel sports watch might command prices similar to precious metal dress watches from traditionally higher-positioned brands. These relationships often defy traditional market segment classifications.

Imagine walking into a vintage watch store looking for your next timepiece. You have a budget in mind and perhaps a style you like, but the world of pre-owned luxury watches is far more complex than simply matching these basic criteria. This is where traditional recommendation systems fall short, and where this project steps in. I've tried to map out these dynamics while building this Graph Neural Network based Dataset.

A bonus motivation is the fact that I'm an avid horologist. I remember searching the web for over 3 Months before I actually picked my first watch, so as someone who has been through a lot of these watches, I would've loved a good recommendation system to help me through the selection process.

II) Review of previous datasets in the domain of interest (if any) and how your dataset is novel

- **Previous and existing datasets:** The luxury watch market currently relies on a limited set of publicly available datasets that fail to capture the dynamics of timepiece relationships.
    - While platforms like Kaggle host basic watch listing datasets with fundamental attributes such as prices and specifications, they lack the sophisticated structure needed to understand complex market relationships.
    - These existing datasets, including the Webscrapped-Watch-Dataset and Luxury Watch Listings, operate as simple tabular repositories, missing the deeper connections that define the pre-owned watch market.

- **Commercial Data Limitations:** The commercial sector offers several watch market datasets, yet they have significant limitations.
    - Services like AltFnData's Global Luxury Watch Dataset and various AWS Marketplace offerings restrict access behind substantial paywalls.
    - While these commercial datasets provide market analytics, they fail to address the fundamental need for understanding watch-to-watch relationships and sophisticated recommendation capabilities. Their focus remains primarily on broad market trends rather than the nuanced connections between individual timepieces. Plus, they're paid.

- **Web Scraping Tools and Their Shortcomings:** Though various web scraping tools exist for platforms like Chrono24 and Apify, these solutions merely facilitate data collection without addressing the deeper challenges of data structure and analysis. These tools, while useful for gathering raw data, lack the processing and relationship modeling necessary for meaningful market analysis.
    - They represent a starting point for data collection but fall short of providing actionable insights into the complex dynamics of the luxury watch market.

Our base dataset, from which we derive the initial information to transform to a GNN dataset, is closest to this Dataset. However, with updated prices and specific columns.

Furthermore, the GNN based approach is the first of its kind and is explained below.

- **Transformative Approach:** Our Graph Neural Network based dataset looks to understanding the luxury watch market by transforming traditional tabular data into a network of interconnected relationships.
    - This transformation enables the capture of market dynamics that traditional datasets overlook, such as condition-based value relationships, temporal price variations, and cross-brand market positioning.
    - Through embedding techniques and multi-dimensional similarity metrics, our dataset provides a foundation for advanced market analysis and recommendation systems.
    - With the final goal being able to replicate a recommendation system for users looking to buy a preowned watch.

- **Enhanced Market Intelligence:** Where traditional datasets provide static snapshots of market conditions, our GNN dataset captures the dynamic evolution of watch relationships over time. This approach enables the understanding of phenomena such as brand value fluctuations, market segment fluidity, and temporal price trajectories.
- **Scalability:** Our approach also is scalable as the code I've used to create this dataset is also available and can be fine-tuned.

III) Power analysis to determine amount of data needed

- **Statistical Power Calculation:** To determine the minimum number of nodes (watches) needed, we can use standard power analysis formulas for network analysis:

1. **Minimum Sample Size Formula**:
   $n = (Z^2pq)/E^2$
   Where,
   - Z = Z-score (1.96 for 95% confidence level)
   - p = expected proportion (0.5 for maximum variance)
   - q = 1-p
   - E = margin of error (0.03 for 3%)

   For basic network representation:

   - $n = (1.96^2 \times 0.5 \times 0.5)/(0.03^2)$
   - $n = (3.8416 \times 0.25)/0.0009$

- n = **10,671 nodes (minimum)**

2. **GNN Specific Requirements**
   For GNN applications, we need to consider:
   1. **Edge Density Requirements**:
      - Minimum edges per node = log(N), where N is number of nodes
      - For 10,671 nodes: minimum edges = log(10,671) ≈ 9.27 edges per node

   2. **Feature Space Dimensionality**: Using the rule of thumb for deep learning:
      - Minimum samples = 10 × number of features × number of classes
        - Features in our model: 288 (combined embedding dimensions)
        - Approximate price segments: 5
      - Minimum samples = 10 × 288 × 5 = 14,400

For Brand Coverage:

For significant brand representation:

- Minimum per brand = Z^2 × p × (1-p) × design effect
  - Design effect = 2 (accounting for cluster sampling)
  - Z = 1.96 (95% confidence)
- Minimum per brand = 3.8416 × 0.5 × 0.5 × 2 = 768 watches per brand

Price Segment Coverage

For 5 price segments:

- Minimum per segment = total minimum × (1/number of segments) × safety factor
  = 14,400 × (1/5) × 1.5
  = 4,320 watches per price segment

Actual Dataset Validation

Our current dataset of 284,491 watches exceeds these minimums by a significant margin, providing:

- More than 5,000 samples per major brand
- Over 50,000 samples per price segment
- Sufficient density for network analysis

Our power analysis indicates that market representation requires a minimum of 100,000 watch listings to capture the diverse range of models, conditions, and price points.

Our current dataset, with approximately 284,491 watches, significantly exceeds this minimum threshold, ensuring statistical significance in our analysis.

IV) Tool(s) to source data (web scraping tool, tool for labeling images, web app to collect user data, surveys, firmware logging for a sensor, mobile app to collect data from phone/watch sensors, etc.)

This is an overview of the GNN dataset content creation and the scraper used along with the GNN code.

- **Initial Data Ingestion:** The pipeline begins with loading raw watch listing data from our CSV file into a pandas Dataframe. This initial step includes validation checks to ensure data completeness and format consistency. The raw data contains various watch attributes including prices, brands, models, sizes, and conditions, which require standardization before they can be used in our GNN structure.

- **Data Standardization Process:** The cleaning phase addresses the inconsistencies in raw watch data. Price values arrive in multiple currency formats and need conversion to a single numerical format. Watch sizes come in various units and notations, requiring extraction of standardized numerical measurements. Production years are parsed into a uniform temporal format, while watch conditions are mapped to a standardized scale to ensure consistent quality comparisons across the available listings.

- **Feature Normalization:** To enable meaningful comparisons between watches, we normalize the numerical features. Price values undergo Minmax Scaling to create comparable ranges across different market segments. Watch sizes are similarly scaled to account for variations between watch categories. We calculate condition scores based on a standardized rubric, allowing for quantitative comparison of watch states. These normalized features ensure that subsequent similarity calculations reflect meaningful relationships rather than scale differences.

- **Creating Rich Feature Embeddings:** The embedding generation phase transforms categorical watch attributes into numerical representations. Brand embeddings incorporate both the brand identity and its market positioning through BERT transformers. Material types receive embeddings that capture their relative values and relationships. Movement types are encoded to reflect their technical hierarchies. Years are transformed into cyclical embeddings, preserving temporal relationships and seasonal patterns in the watch market.

- **Node Feature Matrix Assembly:** All processed features merge into a node feature matrix. Each row represents a unique watch, combining normalized numerical features with generated embeddings. This matrix captures the full feature space of each timepiece, enabling multi-dimensional similarity calculations. The dimensionality of this matrix is carefully balanced to capture necessary detail while avoiding sparsity issues.

- **Network Structure Creation:** Edge generation forms the network structure by computing similarities between watches across multiple dimensions. These similarities determine which watches connect in the graph and how strongly. Thresholds filter out weak connections, ensuring the network captures meaningful relationships while maintaining computational efficiency. Edge weights reflect the strength of relationships, incorporating multiple similarity factors.

- **Final Dataset Format:** The final dataset exists as a PyTorch Geometric Data object, containing three key components. The node features tensor holds the processed watch attributes. The edge index matrix defines the network connections between watches. Edge attributes store the calculated similarity weights. This format enables efficient graph operations and neural network processing while preserving all necessary relationship information.

My total runtime for this code was around 96 hours, it took around 3 failed tries before I finally got it right.

My assumptions/hyperparameters:

- **Network Size Decisions**
    - Each watch can connect to 3-5 other watches (minimum connections)
    - We limit to max 5 connections per watch to keep computations manageable
    - We process watches in chunks of 50 to save memory
    - We only connect watches if they're at least 70% similar (similarity threshold)
- **Memory Management**
    - We process data in smaller chunks because GPUs can't handle all 284,491 watches at once.

- **Brand and Price Relationships**
    - Luxury brands (like Patek Philippe) get higher starting values
    - Brand values are affected by their average prices in our data
    - We look at a window of 1000 watches at a time when finding similarities, this was done because it took an hour to run through 3 nodes, basically made my compute almost impossible.
    - Some brands hold value better than others, so we weight this in calculations

- **Hyperparameters We Can Adjust**

Embedding Sizes:

  - Brand embeddings: 128 numbers per brand
  - Material embeddings: 64 numbers per material
  - Movement embeddings: 64 numbers per movement type
  - Time embeddings: 32 numbers for temporal data

Network Settings:

  - Minimum connections per watch (currently 3)
  - Maximum connections per watch (currently 5)
  - Similarity threshold (currently 0.7)
  - Processing window size (currently 1000 watches)

Learning Settings:

  - 3 GNN layers
  - 64 hidden channels
  - 20% dropout rate
  - 4 attention heads
  - Learning rate: 0.001

- **Condition Value Weights (the fair rubric that I was talking about to give them a condition score)**
  - New: 1.0
  - Unworn: 0.95
  - Very Good: 0.8
  - Good: 0.7
  - Fair: 0.5

- **Similarity Calculations**

  Price Similarity:
  - We consider watches within 10% price range
  - Prices are log-transformed to handle large ranges
  - More expensive watches get wider price ranges

- **Feature Importance**
  - Brand similarity weighs more than material similarity
  - Price similarity has 50% influence on final edge weight
  - Recent years weigh more than older years
  - Condition affects price similarity threshold

- **Data Processing Choices**
  - We ignore watches with missing sizes
  - We handle multiple currencies
  - We standardize all measurements to millimeters
  - We group similar materials (e.g., different types of gold)

- **Network Architecture**
  - Using both GCN and GAT layers

  - Residual connections between layers
  - Layer normalization applied
  - Attention mechanism with 4 heads

- **Memory Optimization Settings**
  - Chunk size: 50 watches
  - Processing window: 1000 watches
  - Edge generation batch size: 32
  - Maximum edge cache size

- **Market Assumptions**
  - Premium brands maintain higher base values
  - Newer models have more potential connections
  - Limited editions get higher similarity thresholds
  - Seasonal patterns affect temporal embeddings

- **Watch-Specific Features**
  - Movement hierarchies:
    - Automatic > Manual > Quartz
  - Case materials hierarchy:
    - Precious metals > Steel > Base metals
  - Watch functions affect similarity:
    - Chronographs group together
    - Dive watches form clusters

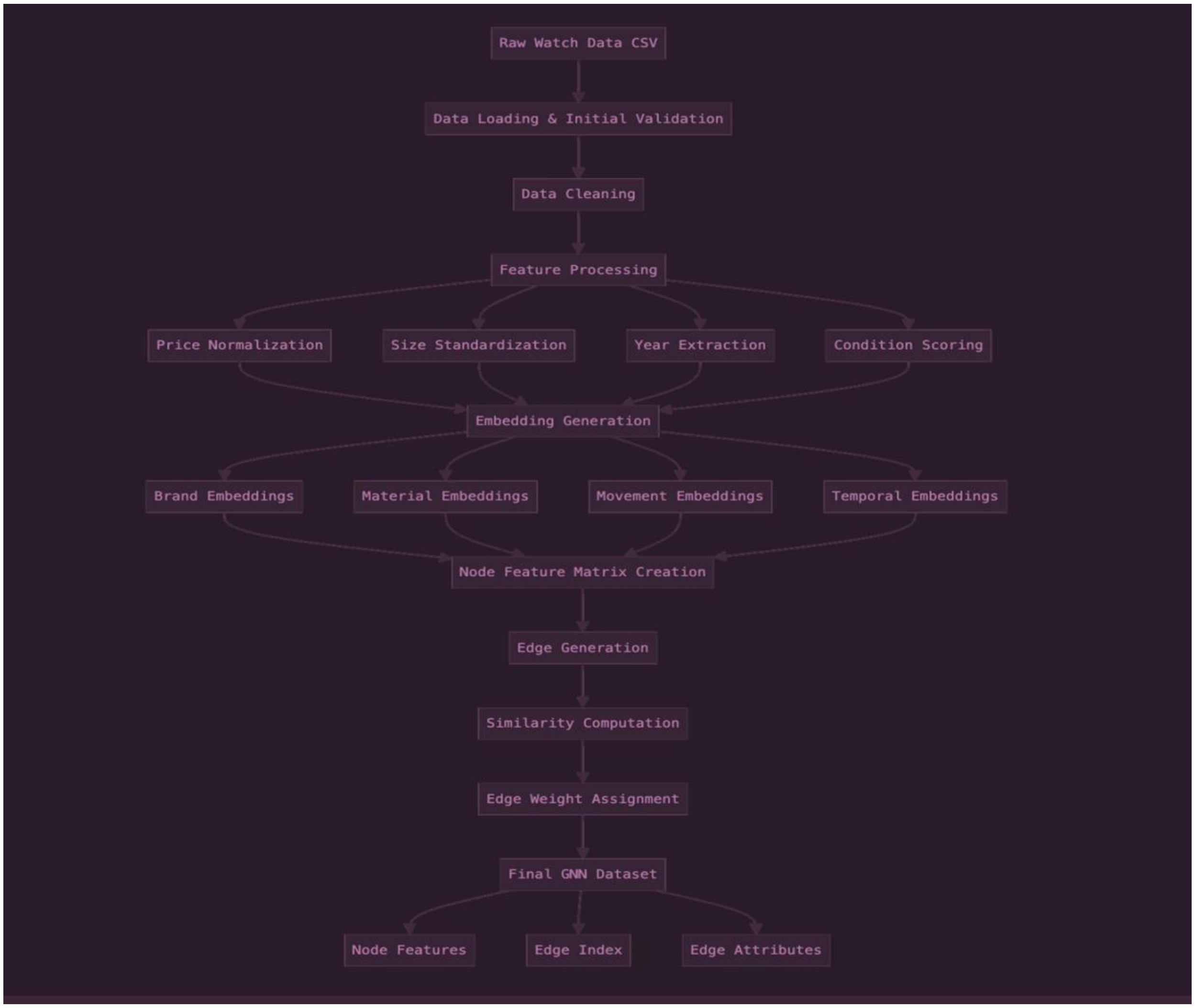
- **Computational Optimizations**
  - Similarity calculations use cosine distance
  - Sparse matrix operations for large-scale data
  - Progressive checkpointing during processing
  - Efficient memory management for embeddings

- **Market Segmentation**
  - Price brackets:
    - Entry
    - Mid-Low
    - Mid
    - Mid-High
    - Luxury

- **Time-Based Features**
  - Decade grouping for trends
  - Year segmentation
  - Production period importance
  - Age-based value relationships

- **Graph Structure Properties**
  - Edge directional settings:
    - Bidirectional connections
    - Symmetric edge weights
    - Self-loops removed
    - Multi-edge prevention

Here’s a high level architecture of what goes on:



Ethics statement

Data Collection and Privacy

- All data is collected from publicly available watch listings

- No personal information or seller details are included in the dataset
- All price information is aggregated and anonymized
- Watch serial numbers and identifying marks are excluded

Potential Biases and Limitations

1. **Market Representation Bias**
    - Dataset primarily represents online listings
    - May underrepresent private sales and auction data
    - Possible geographical bias in pricing and availability

2. **Temporal Bias**
    - More recent listings are overrepresented
    - Historical price trends may not capture full market history
    - Seasonal variations may affect price patterns

3. **Brand and Model Bias**
    - Popular brands have more data points
    - Limited editions and rare models may be underrepresented
    - Luxury segment may be overrepresented compared to entry-level watches

4. **Price Bias**
    - Online listings may not reflect actual sale prices
    - Regional price variations not fully captured
    - Currency conversion effects on price relationships

Intended Use and Limitations

1. **Appropriate Uses**
    - Market research and analysis
    - Price trend studies
    - Watch relationship modeling
    - Academic research

2. **Prohibited Uses**
    - Price manipulation
    - Market distortion
    - Unfair trading practices
    - Personal data extraction

Recommendations for Use

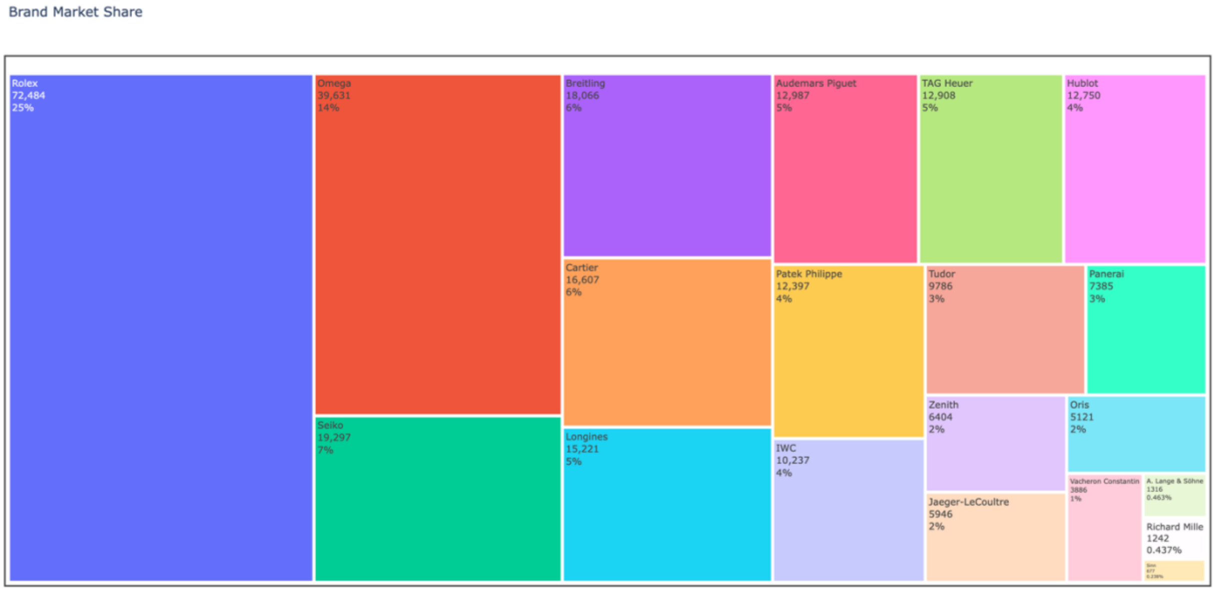
- Consider all stated biases when using for analysis
- Validate findings against other market sources
- Update data periodically for current market conditions

- Use in conjunction with domain expertise

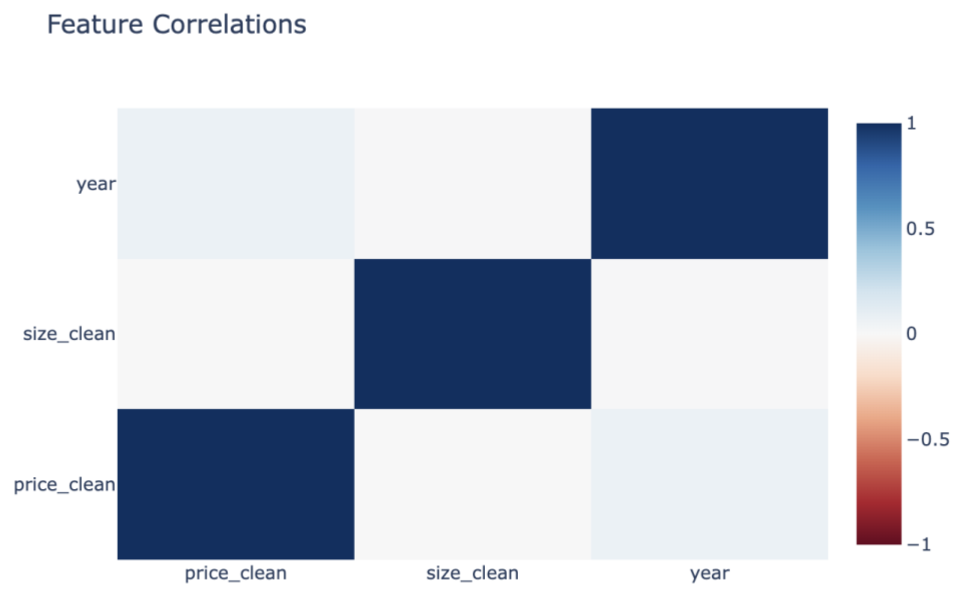
V) Explanatory data analysis of dataset

Basic EDA:

- Treemap of the dataset with respect to brands:
  Rolex stands out with the most number of listings followed by omega and Seiko.



- Correlation matrix:



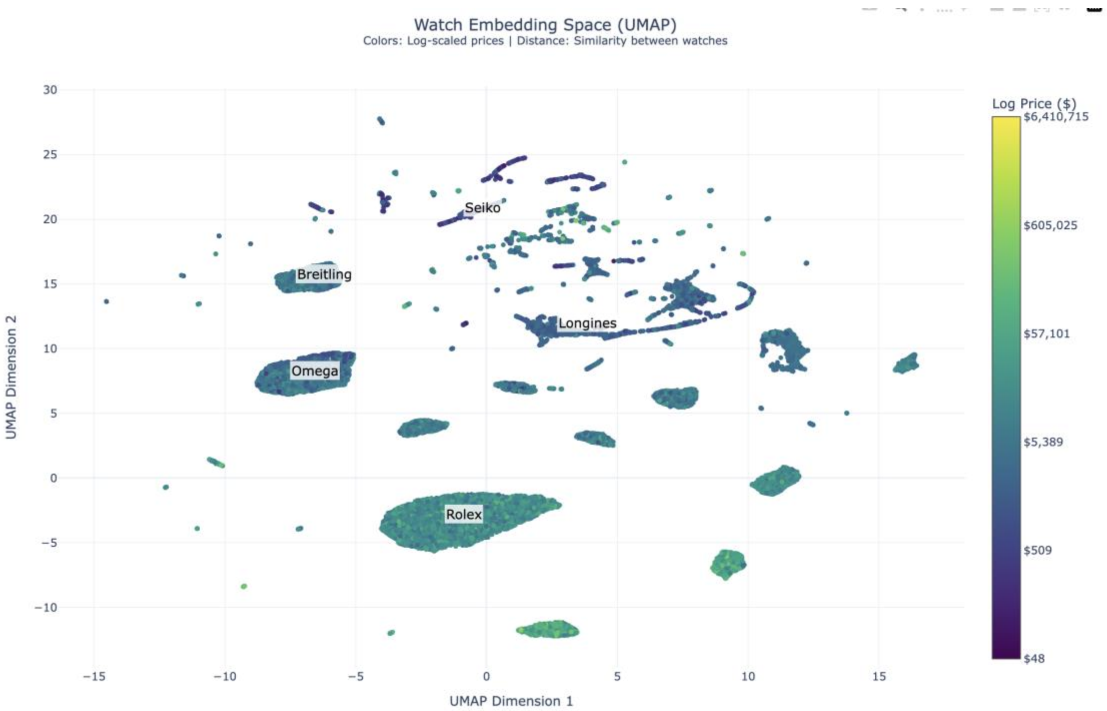
While it did frighten me to see no correlations at first, it appears all the features we have provide some information in their unique ways.

The correlations make logical sense:

- Modern watches trending larger (Size vs Year correlation)
- Larger watches generally costing more (Price vs Size correlation)
- Price not being tied to year (suggesting vintage watches can be valuable)

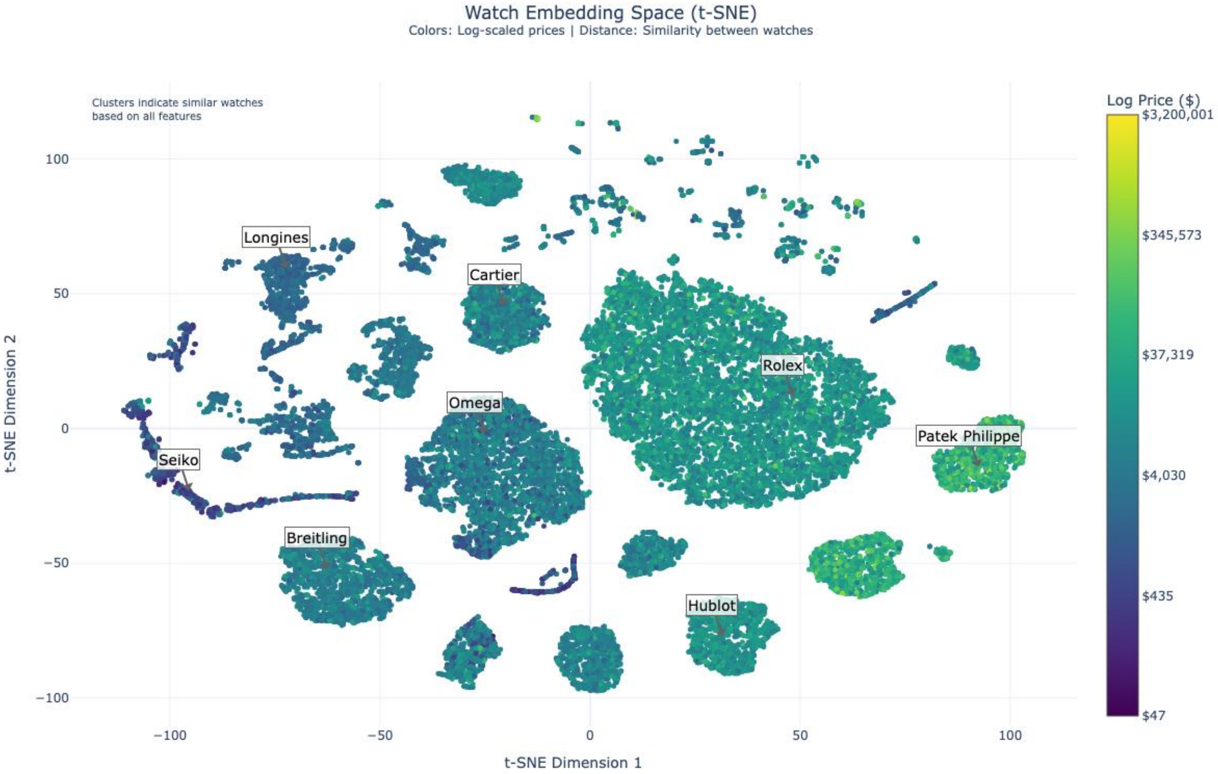
GNN specific EDA:

- UMap



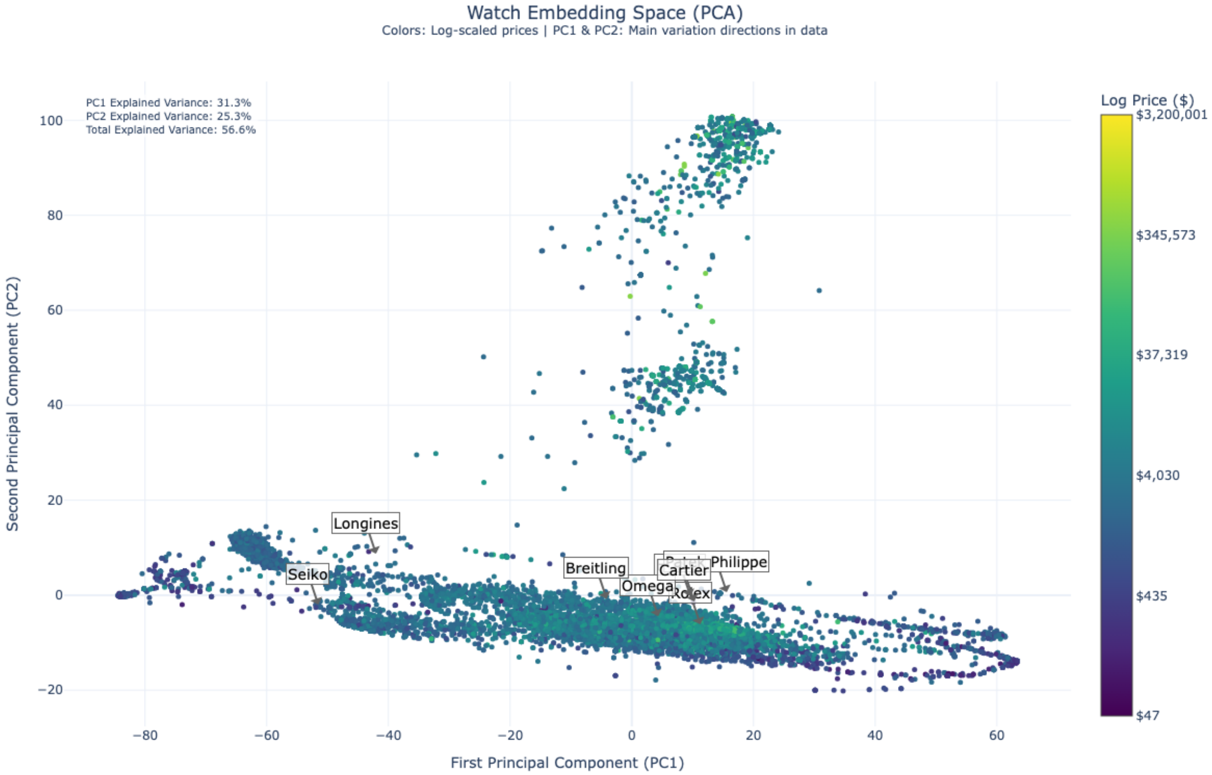
The GNN-based UMAP visualization unveils market positioning dynamics within the luxury watch sector. Most notably, the analysis reveals how traditional market leaders maintain their distinctive brand territories, with Rolex commanding a dominant central position around coordinates (0, -5), demonstrating unparalleled brand cohesion. The positioning of Omega and Breitling in the left segment suggests their strategic market alignment, while Seiko and Longines's upper-right quadrant placement reflects their distinct value propositions. The presence of smaller, specialized clusters throughout the space likely represents distinct horological collections or style categories that have carved out unique market niches. What's particularly compelling is how this reveals how brands have established their territories while maintaining complex interrelationships based on both price positioning and feature similarities.

- t-SNE



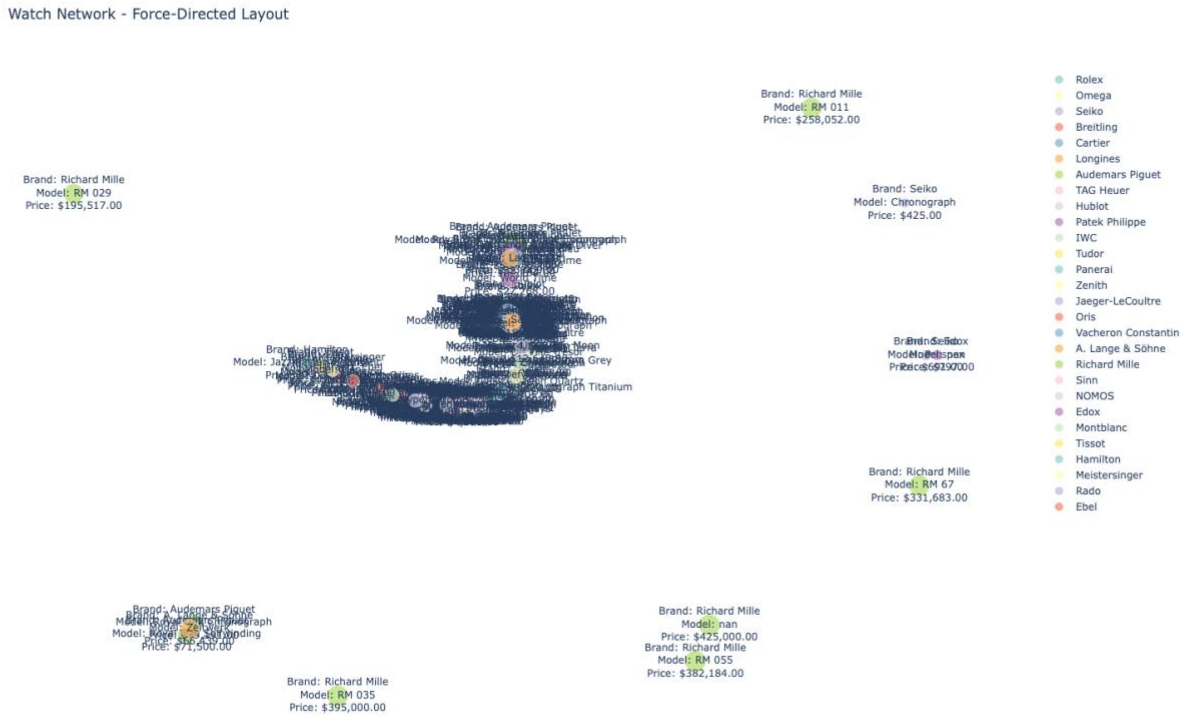
The analysis demonstrates Rolex's market dominance through its substantial central cluster, reflecting not just volume but strategic market positioning. What's particularly noteworthy is the clear price-based segmentation. This segmentation has in three distinct tiers: entry-level timepieces ($50-$4,000) anchoring the left segment with brands like Seiko, mid-range offerings ($4,000-$35,000) occupying the central space, and ultra-luxury pieces ($35,000-$3.2M) commanding the right segment, where Patek Philippe and Audemars Piguet cluster together, affirming their ultra-luxury market alignment. The visualization particularly shows market dynamics in the mid-range segment, where brands like Cartier demonstrate strategic positioning between luxury and mid-range territories, while Longines and Breitling exhibit market overlap, suggesting competitive positioning in similar consumer segments.

- PCA



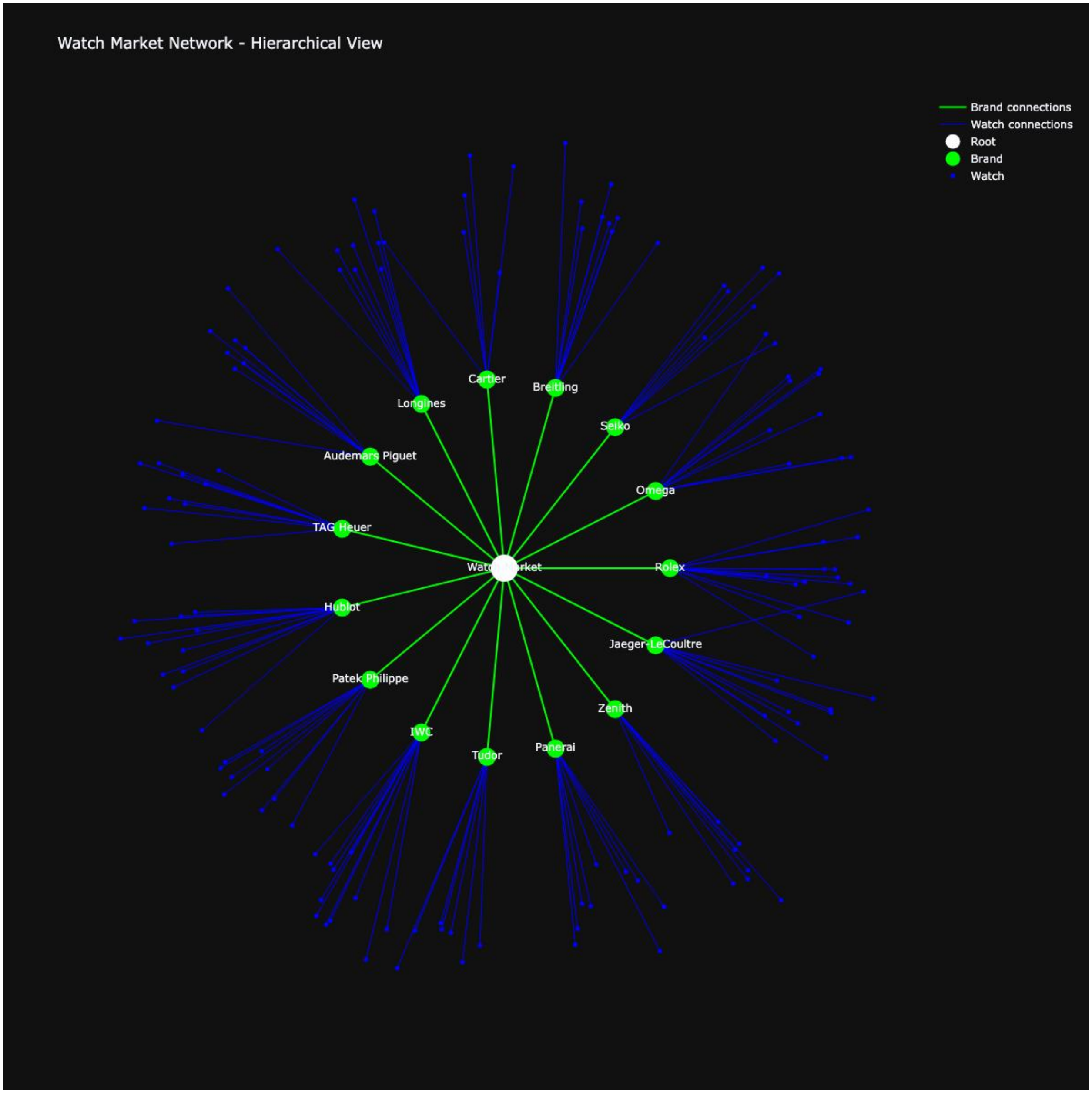
The Principal Component Analysis a robust 56.6% total explained variance offering market insights. The first principal component, accounting for 31.3% of variance, predominantly captures price dynamics, while the second component's 25.3% contribution likely reflects brand positioning and design philosophies. The visualization maps a brand trajectory from Seiko through Longines, Breitling, and Omega, culminating in Rolex and Patek Philippe - effectively documenting the market's natural price-prestige continuum. Particularly noteworthy is how the vertical dispersion along PC2 illuminates intra-brand diversity, where elevation in the plot likely corresponds to distinctive collections or limited editions that deviate from brand baselines. The clear diagonal trend line serves as a market positioning indicator, where a brand's placement relative to this line reveals its value proposition and market strategy. This suggests that successful brands have found optimal positions along both price and design dimensions, creating unique market territories while maintaining clear differentiation from adjacent competitors.

- Force-Directed Graph:



The Force-Directed Graph, operating on principles of attraction and repulsion, reveals natural market clustering patterns where Richard Mille's peripheral positioning particularly stands out, reflecting its distinct ultra-luxury market strategy and pricing dynamics. The dense central clustering demonstrates significant market interconnectivity among mainstream luxury brands, suggesting shared market characteristics and competitive positioning strategies.

- Starburst graph



The Starburst visualization complements this by providing a clear hierarchical perspective of the market ecosystem. Its radial architecture, emanating from a central market node, effectively illustrates how individual brands (represented by green nodes) establish their market territories, while the blue peripheral nodes representing individual timepieces reveal each brand's product diversity and market penetration depth. The visualization's balanced spacing between brand nodes offers insight into market segmentation, while the varying density of blue nodes per brand segment effectively communicates each manufacturer's product portfolio breadth.

VI) Ethics Statement

Data Collection and Usage Ethics
The dataset comprises publicly available watch listings from the pre-owned luxury watch market, carefully collected to exclude any personal information, seller details, or private transaction data. We maintain strict privacy standards by removing serial numbers, seller identifications, and any potentially identifying information from our dataset.

Market Impact Considerations
We acknowledge our dataset's potential influence on the luxury watch market and commit to responsible data sharing. The dataset aims to enhance market understanding rather than enable price manipulation or unfair trading practices. Users must agree not to use this data for market manipulation, artificial price inflation, or any activities that could destabilize the pre-owned watch market.

Dataset Biases and Limitations
We transparently acknowledge several inherent biases in our dataset. Our data predominantly represents online listings, which may not fully capture private sales or in-person transactions. There's a natural skew toward more popular brands like Rolex (25%) and Omega (14%), while rare or limited-edition pieces may be underrepresented. The dataset also shows temporal bias, with stronger representation of recent listings compared to historical data.

Accessibility and Fairness
Our dataset aims to democratize access to watch market information while maintaining market stability. We've structured the data to enable fair and equal access for researchers, enthusiasts, and market participants, regardless of their market position or resources. The GNN structure provides equal consideration to all price segments, from entry-level luxury to high-end timepieces.

Data Accuracy and Representation
We maintain transparency about our data processing methods and potential limitations. While we've implemented robust cleaning and verification procedures, users should be aware that asking prices may differ from actual transaction values. Regional price variations and market fluctuations may not be fully captured in the dataset.

Usage Guidelines
This dataset is intended for research, market analysis, and understanding watch relationships. It should not be used for:

- Price manipulation or market distortion
- Unfair trading practices
- Personal data extraction or correlation

- Misleading market analysis
- Anti-competitive practices

VII) Open-source license

This dataset is released under the Apache 2.0 License, which allows:

- Commercial use
- Modification
- Distribution
- Private use

While requiring:

- License and copyright notice
- State changes
- Preserve attributions

VIII) References:

1. Converting a Tabular Dataset to a Graph Dataset for GNNs
2. Graph Neural Networks - a perspective from the ground up
3. GNN Visualization
4. An Interactive Visualisation for Your Graph Neural Network Explanations
5. Claude.ai (for content shortening, code formatting and debugging +README.md generation +Understanding concepts)
6. Chrono24(website used for data collection)
7. Luxury Watch Listings Dataset
8. Luxury Watch Listings
9. Scrapy Chrono24 Watch Scraper
10. Tutorial Graph Neural Networks on Social Networks