Project Report

Researching & Qantifying the relationship between Modern art & the prices by distinguishing the difference of Historical art/ Post War

Introduction

Project Overview

This project explores a dataset of artworks, focusing on analysing trends, pricing, and movements within the art market. The dataset was sourced from Kaggle- https://www.kaggle.com/datasets/flkuhm/art-price-dataset/data, and it contains details such as artist names, the year that the artwork was created, the condition of the artwork whether it was signed or not or initialised, and the prices it was sold for. The goal of my project was to clean, analyse, and visualise the data to extract insights based on mainly the prices of the art work and the time periods of the pieces.This Dataset could also help me to understand what artist or styles have lost or gained value over time. That would be beneficial in helping private galleries or even colledtors understand what may be their best market.

Objectives

- Clean and preprocess the dataset for consistency and accuracy.
- Explore and visualise trends in the dataset, such as price distribution, popularity of the artists, and art movement influences.
- Although this is to find a representation of the artwork sold globally, its also to gain insights in how Sothebys has conducted their sales and their influences.
- Provide and demonstrate an analysis using Python coding.

Methodology

Data Sources The dataset used in this project is: original (oldartDataset.csv) and recent one artDataset2.csv, which contains information about various artworks sold by the Sothebys art gallery. The data has columns such as: * Title: The name of the artwork. * Artist: The artist of the artwork. * Year created: The year in which the artwork was made. * Time period: Post war, modern, contemporary art, etc. * Price: The selling price of the artwork. * Art movement: The artistic movement associated with the work. * Condition: The state of preservation of the artwork. * Signed: Has the artwork been signed or not? Columns that I have added, include: Age of artwork(years) and I replaced Price with Price(USD)

Data Cleaning & Processing

The dataset that I will be using is the 'Art price Dataset'found on Kaggle.com made by FI Kuhm. My source, the link to this Dataset is:https://www.kaggle.com/datasets/flkuhm/art-price-dataset/data

It was made in 03/19/2022 @misc{fl_kuhm_2022, title={Art Price Dataset}, url={https://www.kaggle.com/ds/2558650}, DOI={10.34740/KAGGLE/DS/2558650}, publisher={Kaggle}, author={Fl.Kuhm}, year={2022}}

How the data was collected and it's limitations/constraints

There is limited information on this dataset, it was made using data from the Sotheby's, art auction website: https://www.sothebys.com/en/ by webscraping. The data being taken in 2022 could mean that the data is not an accurate represention currently. ### The number of records: 754 entries #### The number of features: 9, they include; unnamed(0) price, artist, title, yearCreation, signed, condition, period and movement

Handling missing values: Replaced empty fields and NaN/nan values with pd.NA, then dropped rows with critical missing data or replacing ti with 'unknown' to keep the data more consistent by keeping rows that I can afford to keep.

Reformatted columns such as year created and Price(USD) to ensure they contain only numeric values, This way they could be sorted easily, such as ascending the numbers

Conversion of data types: Transformed price (USD) into a integer for analysis, such as above.

I've also created new columns such as age of artwork (years) to help in trend analysis. I've done reaserch here where i can use the age of the pieces to determine any relationships between the ages and prices.

Exploring my data analysis

My explorations and questions

Price analysis: What's the average price of artworks by different artists, periods, or movements?

Trends over time: How has the value of artworks changed over different periods or years?

Condition correlation: Do artworks in "excellent" condition tend to have higher prices?

Movement popularity: Which artistic movements are most represented in the dataset?

Artist comparisons: Who are the most expensive or prolific artists in the collection?

Time periods: How many artworks fall into each artistic period, and how does that correlate with price?

Signed vs. unsigned: Do signed works command a premium?

Key Findings

- The distribution of artwork prices varies significantly across different periods and movements.
- Certain artists have a higher frequency of sales and higher average prices, these artists include, cell 39: Russell Young - $226.600Alex Katz-$187.000Ed Ruscha-$154.500Cindy Sherman-$141.000Vija Celmins - $135.000Frank Schroeder - $120.000Gustave Blache III - $93.000Biff Elrod - $87.000Jasper Johns - $75.000Larry Bell - $71.500 Then I evaluated the movements that the top 10 artists are part of based on the pieces sold, cell 40: The top 3 artists are in Pop art, from this I drew the conclusion that pop art would sell better than other movements if sold by an gallery. In the next cell, I tried to find the percentages of the artists so that I can then put them into a pie chart, cell 41 & cell 43. I decided to find the top five movements and put that in a pie chart instead to see how this data would translate in the entire dataset. Turns out the most pieces sold were in Abstract pieces at 29%, followed by Realism and Expressionism at 18.5% at 24.5% found in cell 42. In cell 44, I was l looking at the top 10 artist by sales value in comparison to the top 10 most prolific artists by the number of artworks they have had represented by the Sothebys and sold. From cell 44 I have gathered that Russel Young is the top artists in both followed by Alex Katz and Ed Rucha, reinforcing that the top art movement sold or demanded is Pop art. Some artists have a higher total revenue despite selling fewer pieces, indicating a strong demand for their work.
- The condition of the artwork influences pricing trends.
- Older artworks generally fetch higher prices, though exceptions exist based on artist reputation and rarity. Visualisations

- Price trends over time: A line chart illustrating the fluctuation in average prices of artworks per decade.
- Artist popularity: A bar chart showing the top 10 artists based on total sales revenue.
- Art movements: A pie chart depicting the distribution of artworks across different movements.
- Artwork age vs. Price: A scatter plot examining the correlation between an artwork's age and its price.

Challenges & Learnings

Challenges Faced * Dealing with inconsistent data formats, especially in the year created column. * Handling missing and unknown values in crucial columns. * Ensuring the dataset was clean enough for meaningful visualisation and analysis. Lessons Learned * Data preprocessing is crucial to ensure reliable results in analysis. * Using pandas and seaborn effectively helped in extracting and presenting insights. * Structuring a project with reproducibility in mind makes future modifications easier.

Ethical Considerations

Biases & Limitations * Selection Bias: The dataset may not be fully representative of the global art market. * Missing Information: Some artworks lacked complete data, affecting overall analysis. * Potential Misinterpretations: Price variations may be influenced by external factors like auction house prestige and collector demand.

Conclusion & Future Work

Conclusion This project successfully cleaned, analysed, and visualised key trends in the art market. By identifying patterns in pricing and artist popularity, it offers valuable insights into how different factors impact the sale value of artworks. Future Work * Expanding the dataset with additional sources for a broader analysis. * Implementing machine learning models to predict artwork prices based on historical data. * Exploring external factors such as auction houses and geographic trends to refine insights.

Sources

- GitHub Repository: [Insert Link]
- Class Exercises & Homework: [Insert Links]
- Additional Data Files & Documentation: [Insert any additional resources]

Cite any papers, articles, or datasets referenced in your project. Link any essential documents or files here. Mention any sources, collaborators, or

tools that were helpful. https://github.com/peckham-daz/24-intro-to-data-science/tree/main/2025_01 https://matplotlib.org/stable/gallery/statistics/boxplot_colorhtml#sphx-glr-gallery-statistics-boxplot-color-py https://seaborn.pydata.org/tutorial/color_palettes.html https://proclusacademy.com/blog/customize_matplotlib_piechart/ https://www.kaggle.com/datasets/flkuhm price-dataset/data https://www.statology.org/pandas-remove-characters-from-string/ https://stackoverflow.com/questions/17468878/pandas-python-how-to-count-the-number-of-records-or-rows-in-a-dataframe https://www.geeksforgeeks.org/python-pandas-dataframe-replace/ https://saturncloud.io/blog/how-to-update-a-pandas-dataframe-row-with-new-values/#:~:text=This%20can%20be%20done%20using,the%20column%20nar https://www.w3schools.com/python/pandas/trypandas.asp?filename=demo_pandas_cleaning_dropna https://matplotlib.org/ipympl/ https://chatgpt.com https://www.geeksforgeeks.org/interactive-graphs-in-jupyter-notebook/ https://www.w3schools.com/python/pandas/ref_df_nunique.asphttps://proclus https://seaborn.pydata.org/tutorial/color_palettes.html https://matplotlib.org/stable/gallery/color/named_ https://matplotlib.org/stable/gallery/statistics/boxplot_color.html#sphx-glr-gallery-statistics-boxplot-color-py https://matplotlib.org/stable/gallery/lines_bars_and_markers/stackplot_ glr-gallery-lines-bars-and-markers-stackplot-demo-py I also got help from my collegue, a python coding tutor, Ben

Submission & Repository

File Format: This report is available in .pdf, .docx, .odt, and .md formats.

GitHub Repository: [Insert link to my project repository]