KU LEUVEN

FACULTY OF SCIENCE

From Rembrandt to Van Gogh

An Analysis of Artwork Metadata Using Information Visualization and Machine Learning

Aona Wang

Supervisor: Prof. Dr. Frederik Truyen

Co-supervisor: Prof. Dr. Katrien Verbert

Mentor: Ivania Nadine Donoso Guzmán

Thesis presented in fulfillment of the requirements for the degree of Master of Science in Digital Humanities

Academic year 2022-2023

Preface

This journey began with Vincent van Gogh, one of my favorite artists, and his painting "Starry Night Over the Rhone" at the Musée d'Orsay. The painting's vitality enchanted me as I stood there, taking it all in, realizing I wanted to explore every aspect of it.

Then, I started wondering what can we discover from artworks like Van Gogh's paintings beyond the aesthetic value. Can we get something interesting that human eyes cannot perceive? That became my first thought, which is to analyze and visualize colors in the painting. As I delved into the metadata of Cultural Heritage Institutions, I gradually incorporated other relevant dimensions into the analysis.

It was a challenging task, as I had to narrow down the focus and find the appropriate techniques for each step based on a vague theme and abundant data source. I tried different data sources, tested with multiple techniques, and explored the works of different artists. Throughout the process, I encountered different forms of data and learned how to handle them effectively. As the exploration enables me to draw a link between Rembrandt and Van Gogh, I gradually find the pattern of retrieving insight from data by assembling tools, methods, and procedures.

Here I would like to express my gratitude to Professor Frederik Truyen, my thesis supervisor, who supports my ideas and encourages my exploration. His suggestion on Europeana and Rijksmuseum gave me original inspiration. Further, I extend my gratitude to Professor Katrien Verbert for being the co-promotor of my thesis. Professor Verbert's course on HTML and CSS equipped me with the necessary skills to produce the outcome, incorporating all my findings. I would also like to thank my mentor Ivania Nadine Donoso Guzman for offering me guidance on my technical problems, giving constructive feedback and valuable help to further my analysis. Lastly, I want to thank my family and my cat Tiger for giving me remote support that cheer me up during the whole process.

Summary

The trend of Big Data affects not only industry and business sectors but also the humanities field. With the advancement of technologies and the emphasis on data management, large amounts of data are created, transformed, and stored by various means. Hence, the problem lies in the data quality instead of quantity. Cultural Heritage Institutions (CHIs) set up standards for data management and promote data openness for the public so the end users can access the data for re-creation.

For general users, finding the appropriate data from various sources and transforming the data into meaningful insight is a complex task. The data extraction process requires an in-depth understanding of digital collections.

With a curiosity about the metadata situation in CHIs of European Unions, I examined the data of Rijksmuseum (the Dutch National Museum), which has been recognized as a good digitalization practice. I began the pattern-revealing based on museum metadata of the representative figure of the Dutch Golden Age—Rembrandt van Rijn. As I noticed that the data of artists whose artworks are mainly stored in other institutions (e.g. Vincent van Gogh) is incomplete, I retrieved the data from other sources to establish a new complete dataset. By connecting the artists via similar parameters, I then completed storytelling featuring the color pattern and link between them in the form of a web page.

In general, the thesis aims at the process of extracting artwork data from Rijksmuseum and supplementing data from other sources that are publicly accessible to uncover valuable insights, patterns, and connections within the digitized collections. The project involves using Python, Html, CSS, JavaScript, Machine Learning Models (CLIP), and other analytic tools with a focus on turning data into direct presentation while enriching public insight into understanding digital artworks.

Specifically, the thesis approaches the insight from two representative figures of different artistic styles and centuries in Dutch art history, summarising the main features and patterns based on extracted metadata and exploring methods for better storytelling.

Keywords: Artwork Metadata, Data Analysis, Machine Learning Model

List of Abbreviations

API	Application Programming Interface
CHI	Cultural Heritage Institution
CLIP	Contrastive Language Image Pre-training
CSV	Comma-Separated Values
CSS	Cascading Style Sheets
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation

Table of Content

1. Introduction

Apart from transforming physical artwork into digital format for restoration and long-term preservation, the digitalization of artwork is a strategic approach that requires cultural institutions to interpret, deduce art collections, create interactive platforms for publication, and initial new forms of representation. With vast types of artwork covering texts, images, and videos, the selection of techniques differs accordingly. Museums and other CHIs, the hubs hosting large sets of artwork collections, are creating their databases for data management and initializing online platforms to attract tourists from far away. This trend leads to the situation that not only the visibility of artworks is enhanced but also the public gradually gets access to the data behind artworks for artwork-based scientific analysis.

Unlike traditional analysis featuring the artwork itself, the analysis of data behind objects (namely metadata) expands artwork parameters, recognizes the common trend and patterns behind large datasets by assembling specific themes. Meanwhile, new forms of numeric data are created by data gathering. Assuming the analysis of a particular artwork unveils the feature of pigments and materials of this work, assembling hundreds of artworks turns the non-numeric data into numeric features that allow for further exploration of patterns and correlations.

The digitalization of artworks adds data either recognized by professionals or technologies, such as author name, size, material, into artwork, and the quality of metadata is constantly enhancing with higher standards and requirements set based on metadata standardization. As more attributes are included in the metadata, analyzing artwork metadata becomes an interdisciplinary attempt that works with multiple technical methods catering to the digital era.

2. Methodology

2.1 Literature Review

2.1.1 Overview of Rembrandt and Van Gogh

The study of Rembrandt van Rijn and Vincent van Gogh, the two astonishing figures in art history, is constantly emerging. Scholars approach each of them from the perspective of major artworks, characters, artistic styles, and pigment detection using new techniques. For Rembrandt, the analysis of his paintings is more than the artist himself due to the paucity of documentary information about Rembrandt's life combined with the abundance of pictorial material.[1] In this case, the audience tends to empathize with the painter through his artworks. In his Self-Portrait, he looks out at the audience directly with his eyes staring steadily and heavily. As Émile Michel remarked that he was even more expressive about the forcefulness of Rembrandt's gaze through the heavy wrinkles that had come to age his face so prematurely[2]. It's hard to criticize the romantic perspective people view paintings, but a restoration work done in 1992, which involved removing discolored varnish from Rembrandt's artwork, revealed a vibrant array of pinks and flesh tones on his face and enhanced the depiction of his emotional state[3]. Consequently, it can be misleading when viewers attempt to analyze the painting solely through the lens of the artist's life experiences, as there's a tendency to project a specific kind of sadness onto the works of artists who led difficult lives. In the meantime, Van Gogh's private life attracts no less attention from the public than his artworks. As an artist whose life has been haunted by mental illness and left the world with stories and letters, curiosity and exploration add a psychological point of view to the analysis of Van Gogh. Apart from the study of his illness and its connection with the works he creates, the comparison between the artwork and its underlying meaning, such as artwork representation based on different cities[4], still life from Heidegger and Van Gogh[5], is frequently seen in the analysis of Van Gogh. It is worth mentioning that Van Gogh and Paul Gauguin are linked closely in the pursuit of Impressionism and the clash between them.

As Rembrandt belongs to the Baroque Style of the Dutch Golden Age while Van Gogh is a Post-Impressionism artist, the difference in time and artistic style leads to the little discussion of the connection between the two artists. However, scholars approached the two artists from the perspective of art valuation using the Marshall-Forrest model.[6] The qualitative model evaluates the art value from five factors relating to the artist, intermediary, product, market demand, and purchaser, suggesting no clear causality between preliminary factors and fine arts valuations.

---

1 Schama, S., & van Rijn, R. H. (1999). Rembrandt's eyes (p. 747). New York: Alfred A. Knopf.
2 Émile Michel, "L'Exposition Rembrandt à Amsterdam," Gazette des BeauxArts 20 (December 1898): 478–480.
3 van Rijn, R. H., van de Wetering, E., Groen, K., Broekhoff, P., Franken, M., & Binkhorst, L. P. (1986). Self-portrait. Art, Architecture and Engineering Library.
4 Pickvance, R. (1986). Van Gogh in Saint-Rémy and Auvers. Metropolitan Museum of Art.
5 Schapiro, M. (1968). The Still Life as a Personal Object—A Note on Heidegger and van Gogh. In The reach of mind: Essays in memory of Kurt Goldstein (pp. 203-209). Berlin, Heidelberg: Springer Berlin Heidelberg.
6 Desborde, R., & Marshall, K. P. (2015). Rembrandt Versus Van Gogh: A Qualitative Contrast Study Applying a Visual Arts Valutation Model.

2.1.2 Scientific Analysis of Paintings

The divide between science and the humanities has historically limited the scientific study of paintings. It took a long time for CHIs to add the correct data for its collections due to the complexity of recognizing paintings at a time when feature recognition of paintings largely relies on expertise. The history of systematic scientific analysis of artworks dates back to the first half of the 20th century. Institutions used chemical analysis to recognize painting pigment and X-radiography to detect hidden layers.[7] Most attempts were used for setting identification for the artworks and are still on the go in archaeology and art history. The main application areas of artwork processing are digital version creation, image diagnostics, and virtual restoration.[8]

One major difference between paintings and their digital formats is that the colors on paintings changed due to light exposure, humidity, dust, or aging. In the 17th century, the pigment choice was limited but artists like Rembrandt knew exactly how to manipulate color contrast, texture, and translucency of the paint to achieve his intended painterly effects and pictorial illusion.[9] In the late works, Rembrandt used lead white paint in underlayers to build up the impasto.[10] Another unique feature of Rembrandt's late painting technique is the extensive use of coarse smalt, a blue cobalt glass.[11] He often mixed smalt with lakes, earths and black pigments, not only for its color but also for its drying properties and to give texture and translucency to the paint.[12] For Van Gogh's Field with Irises near Arles, his use of the very light-sensitive red paints cochineal and eosin, and chrome yellow add a yellowed gloss to the painting.[13] A survey of van Gogh's works from 1886 to 1890 indicated that paintings with a yellow dominance were numerous, episodic, and multi-regional.[14] While he openly acknowledged the impact of his illness and substance misuse on his art, there exists no direct evidence demonstrating a physiological or chemical alteration in his color perception. Despite this, the most viable explanatory framework for the prevalence of yellow tones in his paintings remains his artistic preference.[15] Meanwhile, scholars are interested in his utilization of complementary colors as an effort to imitate and enhance the portrayal of natural scenes.[16] The heightened incorporation of complementary colors is particularly noticeable during the later phase of Van Gogh's artistic career.

[7] Gavrilov, D., Maev, R. G., & Almond, D. P. (2014). A review of imaging methods in analysis of works of art: Thermographic imaging method in art analysis. Canadian Journal of Physics, 92(4), 341-364.

[8] Barni, Mauro & Pelagotti, Anna & Piva, Alessandro. (2005). Image processing for the analysis and conservation of paintings: Opportunities and challenges. Signal Processing Magazine, IEEE. 22. 141 - 144. 10.1109/MSP.2005.1511835

[9] van Loon, A., Noble, P., Krekeler, A., Van der Snickt, G., Janssens, K., Abe, Y., ... & Dik, J. (2017). Artificial orpiment, a new pigment in Rembrandt's palette. Heritage Science, 5(1), 1-13.

[10] Groen, K. (1997). Investigation of the use of the binding medium by Rembrandt. Zeitschrift fur Kunsttechnologie und Konservierung, 2(2), 208-211.

[11] Roy, A. (2012). Studying Rembrandt's techniques at the National Gallery, London. Technè Special Issue Rembrandt Approches Scientifiques et Restaurations, 35, 6-13.

[12] van Loon, A., Noble, P., Krekeler, A., Van der Snickt, G., Janssens, K., Abe, Y., ... & Dik, J. (2017). Artificial orpiment, a new pigment in Rembrandt's palette. Heritage Science, 5(1), 1-13

[13] Geldof, M., Proaño Gaibor, A. N., Ligterink, F., Hendriks, E., & Kirchner, E. (2018). Reconstructing Van Gogh's palette to determine the optical characteristics of his paints. Heritage Science, 6(1), 1-20. https://doi.org/10.1186/s40494-018-0181-6

[14] Arnold, W. N., & Loftus, L. S. (1991). Xanthopsia and van Gogh's yellow palette. Eye, 5(5), 503-510.

[15] Arnold, W. N., & Loftus, L. S. (1991). Xanthopsia and van Gogh's yellow palette. Eye, 5(5), 503-510.

[16] Hulsker, J., & Gogh, V. V. (1996). The New Complete Van Gogh: Paintings, Drawings, Sketches: Revised and enlarged edition of the catalogue raisonné of the works of Vincent van Gogh. (No Title).

As computer storage capacity has grown and a larger pool of technical experts has become engaged, there is now a heightened emphasis on gathering and manipulating data of artworks for purposes such as storing, transmitting, and examining. The metadata of paintings, namely the information identifying the relevant features such as title, date, medium and additional information such as copyright, has been enriched as a consequence. Other than achieving documentation and conservation objectives, painting metadata allows further possibilities. The metadata-enriching process facilitates the digitalization and the digital reproduction of paintings. In the collections of Rijksmuseum, high-quality digital images are produced and colors are emerged as a feature that enables accurate measurement of the color pattern of paintings. To reduce the so-called “Yellow Milkmaid Syndrome”[17]——the proliferation of low-quality images online, the Rijksmuseum licensed their highest-quality images and released them online for free. This soon leads to the extensive download of Rijksmuseum images. The high-quality image of Vermeer’s Milkmaid, for example, is downloaded two to three thousand times a month. For the public, Rijksmuseum is a reliable source, and the release of a Creative Commons CC0 means these images can easily be found on other platforms.[18]

Due to the vast number of distinct pixel values present in each image, surpassing what the human eye can perceive, numerous digital collections adopt a global palette approach to manage the overwhelming magnitude of color data. The global palette is a smaller, defined set of colors applied across the collection. Colors in the extracted image palettes can then be mapped or aggregated to similar colors in the global palette. The Rijksmuseum uses a 32-color global palette which reduces the color variations and the difficulty of searching via color.[19]

By studying the impact of the artwork metadata features, as well as manually-engineered visual features, such as naturalness, brightness, and contrast, scholars can achieve content-based artwork recommendations of physical paintings[20]. As the metadata keeps upgrading and enriching, CHIs began to explore the added value of data sharing to initialize the Linked Open Data and thus enhance the metadata of their own digital collections[21].

2.1.3 Computational Approach of Painting Representation

As Albert Einstein said: “If we trace out what we behold and experience through the language of logic we are doing science; if we show it in forms whose interrelationships are not accessible to our conscious thought but are intuitively recognized as meaningful, we are

[17] https://pro.europeana.eu/post/the-yellow-milkmaid-syndrome-paintings-with-identity-problems

[18] https://medium.com/made-with-creative-commons/rijksmuseum-2f8660f9c8dd

[19] Hinchcliffe, G., & Whitelaw, M. (2015). Colouring digital collections: Challenges and opportunities for the use of colour metadata in cultural collections. Museums and the Web Asia.

[20] Messina, P., Dominguez, V., Parra, D., Trattner, C., & Soto, A. (2019). Content-based artwork recommendation: integrating painting metadata with neural and manually-engineered visual features. User Modeling and User-Adapted Interaction, 29(2), 251-290.

[21] Dragoni, M., Cabrio, E., Tonelli, S., Villata, S. (2016). Enriching a Small Artwork Collection Through Semantic Linking. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S., Lange, C. (eds) The Semantic Web. Latest Advances and New Domains. ESWC 2016. Lecture Notes in Computer Science(), vol 9678. Springer, Cham. https://doi.org/10.1007/978-3-319-34129-3_44

doing art."[22] Information visualization of artworks can be regarded as an inter-discipline between art and science. The growth in Cultural Heritage data and the development of computational techniques motivated information visualization, which helps make sense of massive data collections and offers practical means to interact with these data[23]. Meanwhile, Increasing efforts are being made to improve the visual representation of cultural collections to add in-depth interaction between the collection and the public.

Cultural artworks, such as paintings, are investigated using quantitative techniques. In this context, the artwork itself serves as the subject of analysis, while its distinctive characteristics represent the metadata. Tthe artwork and its features constitute the dataset employed for analytical purposes. Statistical techniques (graphing, correlation, regression) are used to study the possible relationship between variables. Data science wants to use many features in the hope that together they contain the correct information for automating recognition, classification, or another cognitive task.[24]

Machine learning and computer vision are also highly linked to digitized paintings. AI-generated art attracts public attention, computational tools contribute to the feature recognition of artworks and image classification. By training on an extensive collection of digitized artworks, N. van Noord et al. detect the artist-specific visual feature to determine the artist eventually[25]. The method of classifying painting styles by extracting various features from paintings leads to image classification[26]. Lombardi[27] extracted the features by light, line, and color and classified the art paintings by k-nearest neighbor. Visualization and analysis of classification performance were conducted through unsupervised learning, hierarchical clustering, and self-organizing maps. Zujovic et al.[28] proposed the classification method for five genres: abstract expressionism, cubism, impressionism, pop art, and realism. These attempts bring new perspectives into the exploration of paintings regarding the artistic styles, genre, and time to uncover hidden patterns.

In 2021, OpenAi initialized its Contrastive Language–Image Pre-training (CLIP) neural network, which brings new insight into analyzing artworks in the humanities. CLIP is pre-trained on various images with a wide variety of natural language supervision on the internet[29]. The multi-model effectively connects text with images with high accuracy to achieve image clustering based on its text and image encoders. In this paper, I will work with the process of exploring metadata from CHIs, work with the CLIP model for artwork clustering and eventually generate visualization and representation featuring artistic patterns.

---

[22] Peitgen, H. O., & Richter, P. H. (1986). The beauty of fractals: images of complex dynamical systems. Springer Science & Business Media.

[23] Windhager, F., Federico, P., Schreder, G., Glinka, K., Dörk, M., Miksch, S., & Mayr, E. (2018). Visualization of cultural heritage collection data: State of the art and future challenges. IEEE transactions on visualization and computer graphics, 25(6), 2311-2330.

[24] Manovich, L. (2015). Data science and digital art history. International journal for digital art history, (1).

[25] Van Noord, N., Hendriks, E., & Postma, E. (2015). Toward discovery of the artist's style: Learning to recognize artists by their artworks. IEEE Signal Processing Magazine, 32(4), 46-54.

[26] Lee, S. G., & Cha, E. Y. (2016). Style classification and visualization of art painting's genre using self-organizing maps. Human-centric Computing and Information Sciences, 6, 1-11.

[27] Lombardi TE (2005) The classification of style in fine-art painting. PhD thesis, Pace University, New York

[28] Zujovic, J., Gandy, L., Friedman, S., Pardo, B., & Pappas, T. N. (2009, October). Classifying paintings by artistic genre: An analysis of features & classifiers. In 2009 IEEE International Workshop on Multimedia Signal Processing (pp. 1-5). IEEE.

[29] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. ArXiv. /abs/2103.00020

2.2 Methodology

2.2.1 Choice of Data Source

With a 650,000 art historical objects collection, Rijksmuseum provides high-quality data to meet FAIR Guiding Principles for Findability, Accessibility, Interoperability, and Reusability. In 2011, the museum released many high-resolution photographs and the corresponding metadata[30]. The object metadata can be accessed through its two APIs (Collection API and Collection Detail API)and a download file. Data retrieval requires the API key that can be applied via the Rijkstudio account.

2.2.2 Quantitative and Qualitative Approach

Quantitative methods are mainly used for statistical and mathematical analysis of data. Considering the type of data I am encountering, apart from the number of artworks, the hex code and object description are not numerical data that lead to direct quantitative analysis. However, the percentage of each color represented, the frequency of colors, and the summarized number of places and years can be considered as numbers for further exploration. The process requires one more step to recognize numerical data from non-numerical information or to add a mathematical function that converts non-numerical data into numerical.

The color attribute in the metadata is in hex code format, a six-digit, three-byte hexadecimal number used in computing applications to represent colors. The bytes represent the color's red, green, and blue components[31]. The bytes alone are not considered numerical data, but they can be converted to the value of red, green, and blue, with each value in the range of 0 to 255. For example, hex code #737c84 can be considered 115, 124, and 132 as their red, green, and blue value. This way, the three colors' degrees can be further explored.

In contrast, qualitative methods involve the analysis of information other than numbers. For the data source I encounter, location, title, material, and description are data that cannot be transformed into numbers. The image is also data that is hard to analyze with mathematical means. Then, text-analysis and image-detection techniques are efficient methods to enhance the analysis and supplement the exploration of more dimensions.

2.2.3 Information Visualization Principle

In visualizations, color is considered to set the tone and express underlying visual display. Some colors even have implicit psychological associations. However, when color is viewed as an attribute for analysis in exploring the color patterns among artworks, the choice of color in visualization is the direct display of a particular color instead of giving it a hidden message by selecting favored colors.

---

30 https://www.rijksmuseum.nl/en/research/conduct-research/data/overview

31 https://en.wikipedia.org/wiki/Web_colors

Under this circumstance, the visualization of artwork metadata is guided by the Gestalt Principle, focusing on Simplicity and Similarity[32]. The visualization follows the rule of "less is more" which minimizes the color choice of other attributes. The outcome of the analysis tries to convey the feature of all attributes (including time, place) as a whole. Thus, with color attributes displayed by its hex code, other attributes are organized with less color involved to avoid conflicts of the same color.

2.3 Research Question

The thesis aims to explore the situation of metadata in CHIs with Rijksmuseum as the primary data source. I started from the data that normal users can access in the hope of unveiling meaningful insight into exploring the color patterns that boost public understanding of artworks.

With this objective, the project would generate clear visualizations and a user-friendly interface as the outcome. Thus, the process would involve different steps that extract complex metadata into structured data for exploration, analyze the data for gathering information and summarize the findings as storytelling to the public.

Based on the above steps, the first research question is what is the situation of metadata in the Rijksmuseum, and are there any limitations in the metadata that may affect the data exploration? This question facilitates a general understanding of the museum data.

The second research question is about the pattern-finding process. What are the key patterns and relationships found in the metadata of artists' paintings, explicitly focusing on the representative artists—Rembrandt van Rijn and Vincent van Gogh? Answering this question would produce in-depth findings of artistic differences and color patterns.

The third research question focuses on conveying the message to the public. How can these patterns be effectively expressed using different techniques? The use of varied technical tools will be further explained in the following text.

3. Dataset

3.1 Data Source

Many CHIs provide APIs that allow users to explore data from their database. Finding appropriate data sources is a process of constant testing. For my analysis, the evaluation of an appropriate data source is determined mainly by the attributes available. Are these attributes sufficient to explore meaningful patterns, and are the APIs allow large-scale extraction to form the dataset? Namely, the more attributes found in a particular artwork, the

[32] Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & Von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. Psychological bulletin, 138(6), 1172.

more diverse feature of an object retrieved so that the exploration can cover multiple dimensions.

My exploration of data sources starts with Europe's digital cultural collection(Europeana). Europeana provides multiple APIs for different purposes. The APIs I mainly tested are Search API and Entities API. The Search API provides a way to search for metadata records and media on the Europeana repository. The Entities API allows retrieving named entities from Europeana's Entity Collection[33]. The above APIs can be used to search for a particular artist or a specific artwork and returns descriptive information (id, title, time, country, type, description, provider), structural data (link), and administrative data (rights). The first step of working with API is to retrieve the API Key from Europeana and add it to the parameters. A simple search query returns limited results.

```
# Search query related to "Johannes Vermeer"
response=requests.get("https://api.europeana.eu/record/v2/search.json",
params={"wskey":"API_key",  "query":"Johannes Vermeer"})
```

Similarly, the Rijksmuseum provides two APIs (Collection API and Collection Details API) that return general information about objects and detailed information of a specific object respectively. By modifying the parameters, a search function can be achieved based on different themes. For instance, here I search the artwork created by Rembrandt. The output reminds me of the artwork collection of Rembrandt in the Rijksmuseum is 1554 (1532 have image, 18 are on display), followed by general descriptions.

```
# Input
principalMaker="Rembrandt+van+Rijn"
url="https://www.rijksmuseum.nl/api/nl/collection?key="+Api_key+"&principalMaker="+principalMa
ker
response = requests.get(url).json()
```

```
# Output
{'elapsedMilliseconds': 0,
 'count': 1554,
 'countFacets': {'hasimage': 1532, 'ondisplay': 18},
 'artObjects': [{'links': {'self': 'http://www.rijksmuseum.nl/api/nl/collection/SK-C-5',
    'web': 'http://www.rijksmuseum.nl/nl/collectie/SK-C-5'},
   'id': 'nl-SK-C-5',
   'objectNumber': 'SK-C-5',
   'title': 'De Nachtwacht',
   'hasImage': True,
   'principalOrFirstMaker': 'Rembrandt van Rijn',
   'longTitle': 'De Nachtwacht, Rembrandt van Rijn, 1642',
   'showImage': True,
   'permitDownload': True,
   'webImage': {'guid': 'bbd1fae8-4023-4859-8ed1-d38616aec96c',
    'offsetPercentageX': 0,
    'offsetPercentageY': 1,
    'width': 5656,
```

33 https://pro.europeana.eu/page/apis

```
    'height': 4704,
    'url':
'https://lh3.googleusercontent.com/SsEIJWka3_cYRXXSE8VD3XNOgtOxoZhqW1uB6UFj78eg8gq3G4jAqL4Z_5K
wA12aD7Leqp27F653aBkYkRBkEQyeKxfaZPyDx0O8CzWg=s0'},
   'headerImage': {'guid': '29a2a516-f1d2-4713-9cbd-7a4458026057',
    'offsetPercentageX': 0,
    'offsetPercentageY': 0,
    'width': 1920,
    'height': 460,
    'url':'https://lh3.googleusercontent.com/nAHNYM604vhPa1hbE_hBJw55JI01-
ls0zCwHwvDEES38PpGyGHMgG_vaigVWSuK8GFkaxfn2Dmevw_Nmujn5dKW3jItgS6QaSI8VIsiH=s0'},
   'productionPlaces': ['Amsterdam']},
…
```


For my analysis, it is crucial to have the color parameter to explore the color patterns of artworks. Although Europeana API provides useful data, the lack of relevant color information in its metadata is the main reason that I select Rijksmuseum API eventually. Here, the color attribute is included in the output of the Collection Details API. The specific object is essential to retrieve detailed information about an object from the data source. The "colors" attribute enhances the exploration of colors in the targeted paintings.

```
# Get details of object BK-1973-158;
response=requests.get("https://www.rijksmuseum.nl/api/en/collection/BK-1973-158?key=API_key")
```


```
# Output
{'artObject': {'acquisition': {'creditLine': None,
                          'date': '1973-01-01T00:00:00',
                          'method': 'purchase'},
             'artistRole': None,
             'associations': [],
             'catRefRPK': [],
             'classification': {'events': [],
                             'iconClassDescription': [],
                             'iconClassIdentifier': [],
                             'motifs': [],
                             'objectNumbers': ['BK-1973-158'],
                             'people': [],
                             'periods': [],
                             'places': []},
             'colors': [{'hex': '#F1F0ED', 'percentage': 4},
                       {'hex': ' #393630', 'percentage': 0},
                       {'hex': ' #AEA18A', 'percentage': 15},
                       {'hex': ' #5A564C', 'percentage': 0},
                       ...,
             'colorsWithNormalization': [{'normalizedHex': '#F6ECF3',
                                      'originalHex': '#F1F0ED'},
                                     {'normalizedHex': '#000000',
                                      'originalHex': ' #393630'},
                                     {'normalizedHex': '#E0CC91',
                                      'originalHex': ' #AEA18A'},
                                     ...,
```

```
            'copyrightHolder': 'erven Jan Eisenloeffel',
            'dating': {'period': 20,
                       'presentingDate': '1903',
                       'sortingDate': 1903,
                       'yearEarly': 1903,
                       'yearLate': 1903},
            'description': 'Het servies bestaat uit een theepot (A) met een '
                           'rechte tuit, een scharnierend met riet omwondem '
                           'hengsel en een los deksel, een suikerpot (B) '
...
```

3.2 Data Extraction

To get my Rembrandt dataset with detailed descriptive information, I connected the Collection API and Collection Details API via the unique "object number". To do so, I set the artist's name (involvedMaker) as a variable to build a small pandas dataframe including the object number from all his works. Then I used the unique object number of each artwork to retrieve its detailed information covering the title, url, maker, width, height, colors, description, production places, acquisition_method, acquisition_date, materials, and physical medium.

One limitation of Rijksmuseum's API is that the largest number of results per page is 100. For the 1554 results of Rembrandt's work, the data needed to be extracted page by page.

The process can be written as follows:

```
# Get items from one page, then loop over pages(1-16);
for page_num in range(1, 17):
url=f"https://www.rijksmuseum.nl/api/nl/collection?key={api_key}&principalMaker={principalMake
r}&chronologic&ps=100&p={page_num}"
response = requests.get(url).json()
# Process each art object in the current page, get unique objectNumber;
    for artobject in response['artObjects']:
     if artobject is not None:
        objectNumber = artobject.get('objectNumber')
        title = artobject.get('title')
# Then use objectNumber as the primary key to extract detailed information;
url_detail="https://www.rijksmuseum.nl/api/nl/collection/"+objectNumber+"?key="+api_key
        response_detail = requests.get(url_detail).json()
        colors = response_detail['artObject']['colors']
        description = response_detail['artObject']['description']
        materials = response_detail['artObject']['materials']
        ...
# Append all data needed into a pandas dataframe;
        df_r=df_r.append({'objectNumber':objectNumber,
         'title':title,'colors':colors,'description':description,
         ...}, ignore_index=True)
      else:
        continue
```

Thus, my dataset of Rembrandt (df_r) is formed with sufficient descriptive information for exploration.

	object_id	objectNumber	title	maker	width	height	acquisition_date	acquisition_method	colors	colors_WithNormalization	...
0	nl-RP-P-OB-375	RP-P-OB-375	De Pers	Rembrandt van Rijn	2273	3000	1816-01-01	overdracht van beheer	[{'percentage': 66, 'hex': '#E4E3DB'}, {'perce...	[{'originalHex': '#E4E3DB', 'normalizedHex': '...	...
2	nl-RP-P-OB-440	RP-P-OB-440	Klein grijs landschap	Rembrandt van Rijn	3000	1533	None	onbekend	[{'percentage': 24, 'hex': '#ACA892'}, {'perce...	[{'originalHex': '#ACA892', 'normalizedHex': '...	...
3	nl-RP-P-OB-12.411	RP-P-OB-12.411	Drie hoofden van vrouwen, één slapend	Ignace-Joseph de Claussin, Rembrandt van Rijn	1778	2500	2006-01-01	onbekend	[{'percentage': 69, 'hex': '#DDD9CB'}, {'perce...	[{'originalHex': '#DDD9CB', 'normalizedHex': '...	...
5	nl-RP-P-OB-415	RP-P-OB-415	Een blinde draailierspeler met familie ontvang...	Rembrandt van Rijn	None	None	1816-01-01	overdracht van beheer	[{'percentage': 39, 'hex': '#E5E2DA'}, {'perce...	[{'originalHex': '#E5E2DA', 'normalizedHex': '...	...

Image 1. Sample Dataframe of Rembrandt

3.3 Data Enrichment

3.3.1 Descriptive Information

Apart from the dataset of Rembrandt, I applied the same method to extract the dataset of Van Gogh. However, as I retrieved the data correctly from the data source, I found only 11 works of Van Gogh in the digital collection of Rijksmuseum. The limited amount of work would be insufficient to find the correct pattern. To supplement the data, I resorted to Wikipedia data, where information regarding title, date, current location, created place, and medium are given on the webpage. Instead of accessing APIs from Wikipedia, I used web scraping (beautiful soup library) to get relevant information. The code looks like this:

```
import pandas as pd
import requests
from bs4 import BeautifulSoup # library to parse HTML texts
# Parse data from the html into a beautifulsoup object
soup = BeautifulSoup(response.text, 'html.parser')
tables = soup.find_all('table', {'class':"wikitable"})
# There are 6 tables with paintings created in different places. Get all tables and append
them together;
tables_paintings = []for i in range(1, 7):
    table = pd.read_html(str(tables))[i]
    tables_paintings.append(table)
# concatenate the tables together
df_v = pd.concat(tables_paintings)
# reset the index
df_v.reset_index(inplace=True, drop=True)
```

Thus, I got my Van Gogh dataset (df_v) as follows:

	Image Title	Date	Current location	Created in	Medium, Dimensions	.mw-parser-output .tooltip-dotted{border-bottom:1px dotted;cursor:help}No.	No.	Catalogue No.
0	Still Life with Cabbage and Clogs	November-December 1881	Van Gogh Museum, Amsterdam	The Hague	Oil on paper on panel 34 x 55 cm	F 1 JH 81	NaN	NaN
1	Beach at Scheveningen in Calm Weather	August 1882	Minnesota Marine Art Museum, Winona	The Hague	35.5 x 49.5 cm	F 2 JH 173	NaN	NaN
2	Dunes	August 1882	Private collection	The Hague	Oil on panel 36 x 58.5 cm	F 2a JH 176	NaN	NaN
3	Women Mending Nets in the Dunes	August 1882	Private collection	The Hague	Oil on paper on panel 42 x 62.5 cm	F 7 JH 178	NaN	NaN
4	A Girl in the Street, Two Coaches in the Backg...	August 1882	Villa Flora, Winterthur	The Hague	Oil on canvas on panel 42 x 53 cm	F 13 JH 179	NaN	NaN
...	...	...	...	...	...	...	...	...
863	The Fields	July 1890	Private collection	Auvers-sur-Oise	50 x 65cm	NaN	NaN	F 761 JH 2120
864	Haystacks under a Rainy Sky	July 1890	Kröller-Müller Museum, Otterlo	Auvers-sur-Oise	64 x 52.5cm	NaN	NaN	F 563 JH 2121
865	View of Auvers with Church (van Gogh) [Wikidata]	July 1890	Rhode Island School of Design Museum, Providence	Auvers-sur-Oise	34 x 42cm	NaN	NaN	F 800 JH 2122
866	Wheat Fields with Auvers in the Background	July 1890	Private collection	Auvers-sur-Oise	43 x 50cm	NaN	NaN	F 801 JH 2123
867	Sheaves of Wheat	July 1890	Dallas Museum of Art	Auvers-sur-Oise	50.5 x 101cm	NaN	NaN	F 771 JH 2125

868 rows × 8 columns

Image 2. Sample Dataframe of Van Gogh

Since I only want to analyze the oil paintings of Van Gogh, other drafts and watercolor paintings are excluded from the dataset. Here I got basic descriptive information about 868 works by Van Gogh, as he created about 2,100 artworks, including around 860 oil paintings. The summarized amount of 868 oil paintings aligns with his actual work amount.

3.3.2 Color Attribute

As we can see from the above image, the information extracted from Wikipedia does not contain the color attribute of each painting. To unify the color attribute similarly to the Rijksmuseum metadata, I need to retrieve the hex code and its corresponding percentage.

This process involves two color concepts: RGB and HEX. RGB is a color system using three primary colors: red, green, and blue, to display colors onscreen. HEX, which stands for Hexadecimal, is a six-digit combination of letters and numbers. The first two numbers represent red, the middle two represent green, and the last two represent blue[34]. The process of completing color attribute requires two libraries: extcolors and rgb2hex. The extcolors library extracts RGB color from an image, and the rgb2hex library converts the RGB value to hex color code.

To upload images for extraction, I formed an image dataset of Van Gogh's paintings on my local file, then used the extcolors library in Python to get the top 10 colors and calculate their proportions. The input of the local file path instead of the image URL is because the access of hundreds of images exceeds the Wikipedia rate limit. Thus, I downloaded the image files, extracted color data from images, and converted the result into a pandas dataframe.

The first step is to include relevant libraries to retrieve hex code. Some high-resolution images are too large to process, so the thumbnail versions are created for particular cases. Thus, by allowing truncated images, thumbnail files could be processed. However, loading truncated images could lead to incomplete image data and inaccuracy in color detection. I consider this a limit of my data enrichment process.

[34] https://en.wikipedia.org/wiki/RGB_color_model

```
import os
from PIL import Image
import extcolors
import pandas as pd
from PIL import ImageFile
# ignore truncated images;
ImageFile.LOAD_TRUNCATED_IMAGES = True
# function to convert RGB to HEX code
def rgb2hex(r, g, b):
    return '#{:02x}{:02x}{:02x}'.format(r, g, b)
# function to extract color data and convert to dataframe
def color_to_df(input):
    colors_pre_list = str(input).replace('([(','').split(', (')[0:-1]
    df_rgb = [i.split('), ')[0] + ')' for i in colors_pre_list]
    df_percent = [i.split('), ')[1].replace(')','') for i in colors_pre_list]
    # Convert RGB to HEX code
    df_color_up = [rgb2hex(int(i.split(", ")[0].replace("(","")),
                      int(i.split(", ")[1]),
                      int(i.split(", ")[2].replace(")",""))) for i in df_rgb]

    df = pd.DataFrame(zip(df_color_up, df_percent), columns = ['c_code','occurrence'])
    return df
# Specify directory containing images
dir_path = '/local file path'
```

The second step is to loop through each image in the input directory and resize the output images to 900 pixels. The tolerance parameter determines the color distance threshold for grouping similar colors, and the limit parameter specifies the maximum number of colors to extract. Here I set the limit to 10 for the top 10 most frequent colors. Since the function returns the color occurrence, a calculation is added to count the percentage using occurrence/occurrence.sum().

```
# Create an empty list to store color dataframes
df_list = []
# Loop through each file in directory
for filename in os.listdir(dir_path):
    # Check if file is an image
    if filename.endswith('.jpg') or filename.endswith('.jpeg') or filename.endswith('.png'):
        # Open image
        img_path = os.path.join(dir_path, filename)
        img = Image.open(img_path)
        # Resize image
        output_width = 900
        wpercent = (output_width/float(img.size[0]))
        hsize = int((float(img.size[1])*float(wpercent)))
        img = img.resize((output_width,hsize), Image.ANTIALIAS)
        # Extract color data and count percentage
        colors_x = extcolors.extract_from_path(img_path, tolerance=12, limit=10)
        df_color = color_to_df(colors_x)
```

```
        df_color['percentage'] =
(df_color['occurrence'].astype('float')*100/df_color['occurrence'].astype('float').sum()).roun
d(2)
        df_color['filename'] = filename
        # Add color dataframe to list
        df_list.append(df_color)
# assemble all color dataframes into one
df_Paris = pd.concat(df_list, ignore_index=True)
```

3.4 Data Cleaning

Data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database[35]. An overview of the dataset extracted from the metadata shows that there are different forms of inaccuracy, such as duplicates, null values (NA or Null), and inconsistent values. To make the data accessible, the process of data cleaning is necessary.

3.4.1 Handle Null Values

Check Null values in the dataset. The results show that 1120 rows in the description are empty. The Rijksmuseum generates the description as the official explanation of its collections. The standard method of handling null values would be to remove empty values or enrich them from other reliable resources. However, description is not a key parameter in analyzing a single painting but could help measure the overall content or feature of paintings. Here, empty values in the description are kept.

```
#Input: check NA values;
df_r.isna().sum()

#Output
objectNumber             0
title                    0
colors                   0
presentingDate           0
productionPlaces         0
iconClassDescription     0
description           1120
dtype: int64
```

However, among the 1508 works of Rembrandt, not all of them contain values. Multiple rows are blank but are not detected by .isna(). As this feature officially originates from the Rijksmuseum, any other techniques extracting the color composition from artworks would lead to different color recognition outputs. Thus, I removed the empty values to maintain consistency instead of supplementing them with extracted results. The results show there are 1246 rows containing color values.

[35] Wu, S. (2013), "A review on coarse warranty data and analysis" (PDF), Reliability Engineering and System, 114: 1–11, doi:10.1016/j.ress.2012.12.021

```
# Drop rows where there are no value in the color columns;
import numpy as np
df_r['colors'].replace('', np.nan, inplace=True)
df_r = df_r.dropna(axis=0, subset=['colors'])
```

3.4.2 Modify Wrong Values

I noticed two wrong values in the Van Gogh dataset by counting values in the Place Created column: “38.8 x 31.3 cm” , “43.2 x 34.2 cm”. This error also appears on the Wikipedia Page[36].

```
# Input:check wrong value;
data_v['Place Created'].value_counts()

# Output
Paris            226
Nuenen           192
Arles            186
Saint-Rémy        143
Auvers-sur-Oise    77
The Hague         27
Antwerp            7
Nieuw-Amsterdam     3
Drenthe            3
Scheveningen        1
38.8 x 31.3 cm      1
43.2 x 34.2 cm      1
Amsterdam          1
Name: Place Created, dtype: int64
```

As I locate the wrong values, I found out the entire row, other than just the Place Created column, needs to be corrected. So I replaced this row with the correct information.

```
#locate wrong values, 137;
data_v.loc[(data_VVG['Place Created'] == '38.8 x 31.3 cm')]
#replace wrong values;
data_v.loc[137,['Date','Place Stored','Place Created']] = ['1885','Van Gogh Museum, Amsterdam','Nuenen']
```

3.5 Data Manipulation

The describe() method returns the data type, and the primary data type in the dataframe is object and int64. The data manipulation process handles both types and values that need unification. As seen below, the Date columns have multiple formats: Year, Month-Year, and Year Period. All in int64 type.

```
1887                  130
```

[36] https://en.wikipedia.org/wiki/List_of_works_by_Vincent_van_Gogh

```
1886                   94
June 1890               38
September 1889           24
July 1890               24
                     ...
March-April 1888          1
April-May 1888           14
July 1888              1
Autumn 1888              1
November-December 1881     1
Name: Date, Length: 88, dtype: int64
```

Since the analysis aims to explore artworks on a yearly basis instead of focusing on the precise date, I eventually converted the different formats into 'Year' using re(regular expression). I changed the data type from int64 to DateTime.

```
# Use re to extract 4digit from the date;
import re
def get_year(date):
   year = re.findall(r'\d{4}', date)
   if year:
      return int(year[0])
   else:
      return None
data_v['Date'] = data_v['Date'].apply(get_year)
```

3.6 Limitation

As I tried to gather relevant data and enrich the dataset with available methods, there were limitations mainly regarding the data availability, data quality, and different techniques during the processing phase.

Metadata insufficiency: including empty values and lack of completeness. Data insufficiency is a common phenomenon in retrieving metadata from CHIs. Rijksmuseum provides diverse attributes that cover the hex code of paintings and the corresponding percentage. However, the data size shrinks as there are empty values in multiple rows and non-reliable resources as a reference for the data enrichment. Besides, the empty values of 1120 rows would primarily affect the accuracy of feature detection if I want to explore the main feature of Rembrandt's paintings based on the description column.

The varied data source led to different analysis techniques. Based on the different statuses in the data source of Rijksmuseum and Wikipedia, selecting different data extraction and manipulation techniques is unavoidable. Although comparing similar attributes between the two datasets of Rembrandt and Van Gogh would be more objective, it is hard to unify all attributes based on varied data sources. Meanwhile, I do not want to neglect the unique attribute of either side that contains interesting patterns.

The different color extraction algorithms and different image resolutions affect the analysis of color patterns. For the two datasets I have, the color attribute of Rembrandt originates from the Rijksmuseum, while the color attribute of Van Gogh is extracted via Python libraries(extcolors). The different methods lead to different color recognition, and the input of different image resolutions affects the color extraction output. To minimize the impact, I downloaded the original file from Wikipedia and only resized extra large images to thumbnail versions so Python could process them.

Outdated data for download. Rijksmuseum provides access to its complete digital collections in CSV format but is not updated yearly. Thus, this data for download is outdated, and I still need to extract it manually to retrieve the latest version.

The different languages appear in the metadata. The mix of languages appears in both datasets of Rembrandt and Van Gogh. As a Dutch museum, Rijksmuseum allows the audience to explore varied languages. However, as for its metadata, some attributes (title, description) only provide a Dutch version. It would vastly increase the consistency of the analysis and representation to translate Dutch into English or unify the languages into one. In this project, analyzing metadata emphasizes finding meaningful insight from the data rather than spending much time on the language format. A non-Dutch-speaking person would take excessive time and effort to do the translation work. Thus, I chose to keep both languages.

4. Analysis and Visualization

In the previous steps, I extracted two datasets of Rembrandt and Van Gogh respectively, and one dataset of Rijksmuseum metadata from the 17th to 19th century. The three datasets have been cleaned and manipulated for further exploration.

Rembrandt and Van Gogh, as two renowned artists in Dutch art history, represent two different artistic styles (Baroque and Post-Impressionism). Their artworks are summarised via similar attributes——Year of Creation, Place of Creation, and Color to understand the different color compositions and artwork creation patterns regarding year and place. Meanwhile, I hope to find out if there are possible links between the two artists.

4.1 Van Gogh's Dataset

It is a dataset of 868 Van Gogh's oil paintings with attributes including title, date, place stored, and place created. Each artwork is linked by its title to the corresponding hex code and percentage. Based on the attributes, I approached the analysis from two aspects: the color palette and the link between the artwork and its location.

4.1.1 Color Palette

I compared the color distribution of a single painting in two pie charts. The first chart is made from the limited metadata of Van Gogh in the Rijksmuseum. In contrast, the second graph is extracted via the extcolors library with the data source of Wikipedia.

In the first pie chart, it is easy to find that the main tone (#52686D) matches the background color of Van Gogh's Self-Portrait. However, the hex color is more general and vague than the second graph. As we can see in Self-Portrait, the different lightness of blue is not detected in the first graph but clarified in the second graph. Nevertheless, the color detection of the same painting differs between the two techniques.

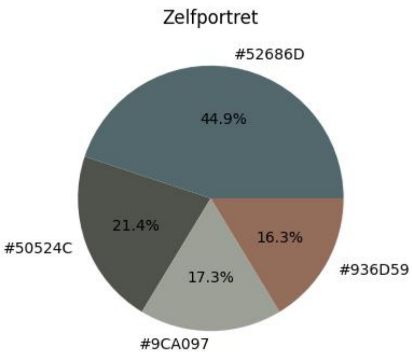


Figure 1. Color Composition (Rijksmuseum)

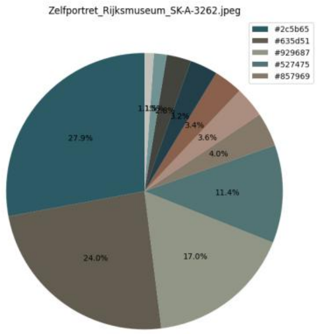


Figure 2. Color Composition (Wikipedia)

Figure 3. Van Gogh's Self-Portrait

The extcolors library, better at displaying the proportion of color of a painting, generates more color layers. It can recognize similar colors with different saturation and lightness.

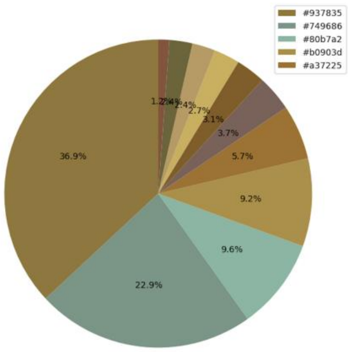


Figure 4. Color Composition of Sunflower


Figure 5. Van Gogh's Sunflower

4.1.2 Artwork and Location

Van Gogh traveled to multiple areas searching for inspiration for his artworks. To analyze how different places affect his creation, I counted the number of works he created and grouped the outcome by places.

As shown in the bar chart, Van Gogh created more works in Paris than in other cities, but he barely drew anything in Amsterdam and Scheveningen. An interactive map allows people to zoom in on the exact locations to explore which city is his inspirational hub. I included the

Folium Library and the Folium heatmap to make this map. I added geo-location (longitude, latitude) of relevant cities so the city location is marked on the graph and the amount of work is displayed via the depth of color.

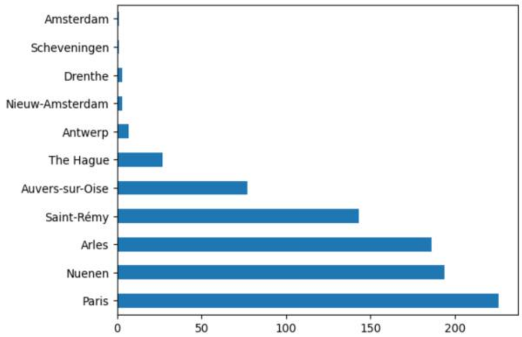


Figure 6. Bar chart shows the number of works in different locations

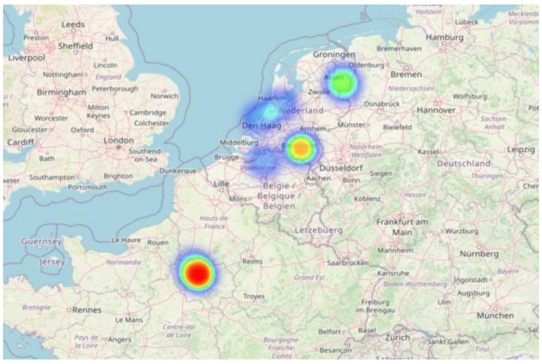


Figure 7. The color depth represents the work amount

4.2 Rembrandt’s Dataset

4.2.1 Color Palette

The sufficient metadata in the dataset enables me to find Rembrandt’s color palette and the exact color composition of any of his works. Although his paintings are famous for the dark and gloomy tone, he used light yellow and grey most frequently.

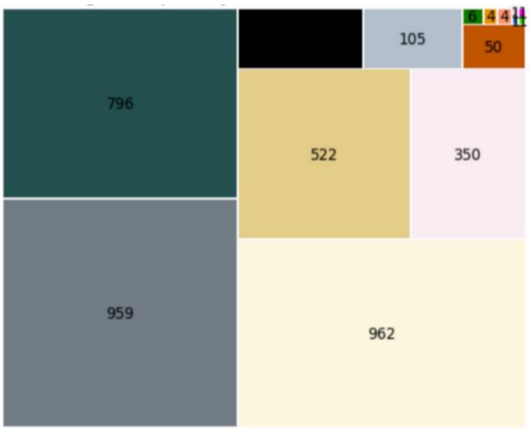


Figure 8. High frequency Colors of Rembrandt

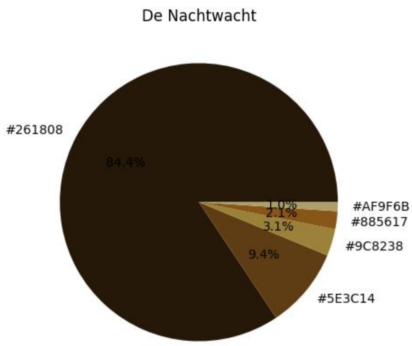


Figure 9. Color Composition of the Night Watch

4.2.2 Artwork and Date

Since Rembrandt made most of his paintings in the Netherlands, his location stayed the same, but the works he created have spread throughout his life. Learning how Rembrandt’s works or color choices changed would be interesting. The two graphs below display the relation between Rembrandt’s paintings and different periods showing his change of color preference and work amount.

The line graph tells the trend of Rembrandt’s work that his creation skyrocketed around 1630 and gradually vanished after 1660. The Largest Hex refers to the dominant color occupying the largest proportion in a painting. As I list the dominant color on the coordinate, there is no significant division between light and dark tones during his period. However, the overall tone is relatively gloomy and dark.

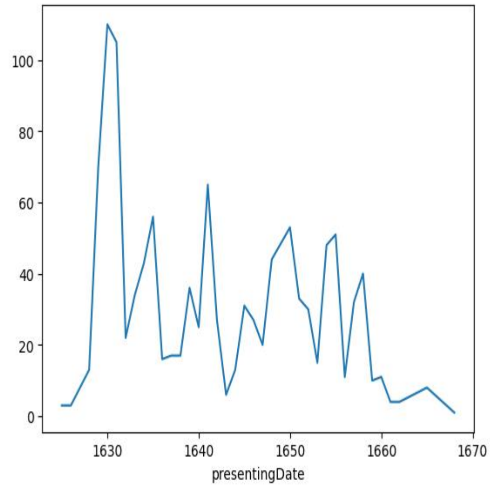


Figure 10. The Change of Work Amount Over Time

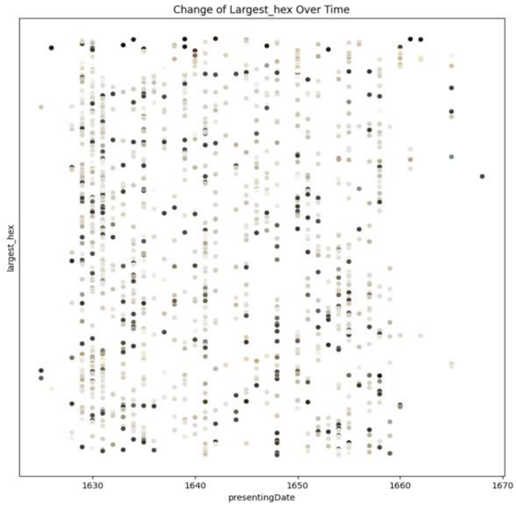


Figure 11. The Change of Main Tone Over Time

4.3 Other Links

Is there a link between color choice and the main theme? To find out, I made a word cloud of descriptions and titles with the help of the Matplotlib library. Apart from the conjunction and articles, Rembrandt paints children, beggars, old men, and self-portraits. He also draws religious themes: Christ, Matthew, Mark, and Luke. Meanwhile, Van Gogh is quite into painting peasants, women, self-portraits, still life, and landscape.

Figure 12. The Word Cloud of Rembrandt

Figure 13. The Word Cloud of Van Gogh

4.3 Metadata Overview

Apart from a particular artist, the exploration of metadata in the same century would be a good reference for people to get an overview of the status and understand the context.

The two artists come from the 17th and 19th century respectively. Thus, I summarized the metadata stored during this period and concluded via three main parameters (location, color, and artists). Here is the sample data from the 17th century.


```
'facets': [{'facets': [{'key': 'Jan Luyken', 'value': 2795},
   {'key': 'Rembrandt van Rijn', 'value': 1350},
```

```
   {'key': 'Wenceslaus Hollar', 'value': 1116},
   {'key': 'Romeyn de Hooghe', 'value': 1108},
   {'key': 'Sébastien Leclerc (I)', 'value': 1023},
   {'key': 'Stefano della Bella', 'value': 1021},
   {'key': 'Antonio Tempesta', 'value': 958},
{'key': 'Jacques Callot', 'value': 943},
...
 'name': 'principalMaker',

{'facets': [{'key': 'Amsterdam', 'value': 23308},
   {'key': 'Parijs', 'value': 10001},
   {'key': 'Nederland', 'value': 9318},
   {'key': 'Antwerpen (stad)', 'value': 7733},
   {'key': 'Noordelijke Nederlanden', 'value': 7547},
   {'key': 'Frankrijk', 'value': 7262},
   {'key': 'Italië', 'value': 4925},
   {'key': 'Rome', 'value': 4608},
...
'name': 'place',

{'facets': [{'key': ' #737C84', 'value': 20962},
   {'key': '#FBF6E1', 'value': 17082},
   {'key': ' #2F4F4F', 'value': 15608},
   {'key': ' #E0CC91', 'value': 13583},
   {'key': ' #FBF6E1', 'value': 7087},
   {'key': ' #000000', 'value': 6433},
{'key': '#737C84', 'value': 4298},
...
'name': 'normalized32Colors.hex',
…
```

Rembrandt contributes to the second most significant work (1350 works), following Jan Luyken (2796 works) in Rijksmuseum. Meanwhile, the hex code #737C84 (Deep Shale) appears most frequently in the 17th century.

4.4 Machine Learning Method

Apart from the attributes provided by metadata, the image itself contains meaningful information for people to explore. What objects are included in a painting? What is the primary color of Van Gogh's artworks? We detect features from images and paintings with human eyes, and our brains reflect on the exact information. Likewise, the CLIP model, pre-trained by a large amount of text and image set, can predict the texts of the given paintings. The CLIP model can be loaded via SentenceTransformer[37]—a Python framework for state-of-the-art sentence, text, and image embeddings. This framework can compute text embeddings for more than 100 languages, enabling semantic understanding.

[37] https://www.sbert.net/

4.4.1 Image Clustering

The Van Gogh dataset and the Rembrandt dataset contain 868 images of Van Gogh and 1532 images of Rembrandt respectively. To encode the images, I loaded the CLIP model and encoded the image path to the model.[38][39]

```
# Incorporate SentenceTransformer to Load the CLIP model;
model = SentenceTransformer('clip-ViT-B-32')

emb_art = model.encode([Image.open(path) for path in img_paths], batch_size=128,
convert_to_tensor=True, show_progress_bar=True)
```

To define the image cluster community, I set the smallest size to 3 and the largest to 500. Thus, any paintings with similar features within this range would be clustered.

```
def community_art(embeddings, threshold, min_community_size=3, init_max_size=500):
    # Compute cosine similarity scores;
    cos_scores = util.cos_sim(embeddings, embeddings)

    # Set k to the minimum size;
    top_k_values, _ = cos_scores.topk(k=min_community_size, largest=True)

    # Filter for rows >= min_threshold;
    communities_ext = []
    for i in range(len(top_k_values)):
        if top_k_values[i][-1] >= threshold:
            new_cluster = []

            # Most relavent topk value;
            top_val_large, top_idx_large = cos_scores[i].topk(k=init_max_size, largest=True)
            top_idx_large = top_idx_large.tolist()
            top_val_large = top_val_large.tolist()

            if top_val_large[-1] < threshold:
                for idx, val in zip(top_idx_large, top_val_large):
                    if val < threshold:
                        break

                    new_cluster.append(idx)
            else:
                # Iterate;
                for idx, val in enumerate(cos_scores[i].tolist()):
                    if val >= threshold:
                        new_cluster.append(idx)
            communities_ext.append(new_cluster)

    # Maximum clusters appear first;
```

[38] Github, 2023 (https://github.com/Mayurji/Explore-Libraries/blob/main/NLPLibraries/Imag %20Clustering%20Using%20 Sentence%20 Transformer.ipynb). Last consulted on the 9th of June 2023

[39] https://github.com/ivaniadg/poststarts/blob/main/notebooks/clip_clustering_search.ipynb

```
    communities_ext = sorted(communities_ext, key=lambda x: len(x), reverse=True)

    # Remove duplicates;
    communities_unique = []
    extracted_ids = set()

    for community in communities_ext:
        add_cluster = True
        for idx in community:
            if idx in extracted_ids:
                add_cluster = False
                break

        if add_cluster:
            communities_unique.append(community)
            for idx in community:
                extracted_ids.add(idx)
return communities_unique
```

The output clarifies that there are 33 clusters in the Van Gogh dataset and 92 clusters in the Rembrandt dataset, and the cluster size ranges from 3 to 64 images.

Image Clusters		
Number	**Van Gogh**	**Rembrandt**
Images	868	1532
Clusters	33	92

Table 1. Image Clusters

The CLIP model here is used to compare features in both oil paintings of the two artists. As I noticed the Rembrandt dataset from the Rijksmuseum(Figure 14) mainly covers Rembrandt's sketches rather than his oil paintings.

Figure 14. Sample Clusters of Rembrandt(Rijksmuseum)

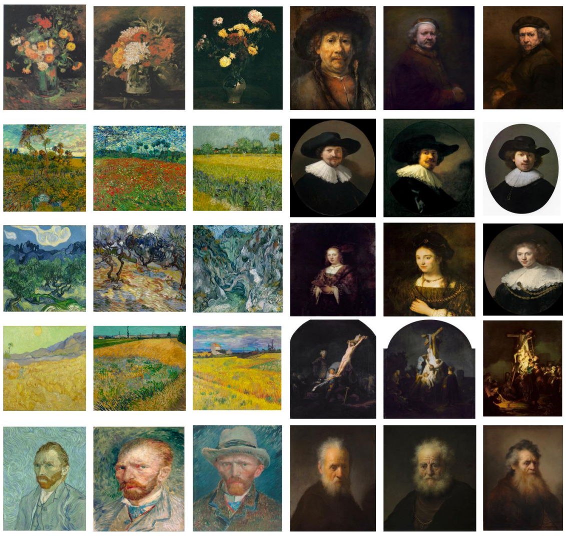


Figure 15. Sample Clusters Classified by CLIP Model

To compare among oil paintings, I then resorted to Wikipedia and retrieved a new Rembrandt dataset with 346 oil paintings, 5 clusters. As we can see, Figure 15 shows three sample images of the Top5 largest clusters of both artists. The left three columns are clusters of Van Gogh and the right three columns are clusters of Rembrandt. Table 2 summarizes the cluster size and theme. It is clear that Van Gogh's choice of colors is relatively colorful and his preference for natural scenes is evident compared to Rembrandt's emphasis on portraits.

Size and Theme of Major Clusters(Top 5)				
Cluster	**Van Gogh**		**Rembrandt**	
	Size	**Theme**	**Size**	**Theme**
1st	17	Flower/Vase	4	Self-Portrait/Rembrant
2nd	12	Flower Field	4	Portrait/Male
3rd	12	Tree/Curve	3	Portrait/Female
4th	9	Wheat Field	3	Bible/Jesus
5th	9	Self-Portrait/Van Gogh	3	Portrait/Old Man

Table 2. Top5 Clusters of Both Artists

4.4.2 Semantic Search

Due to the different languages that may associate with images, the semantic search used the multilingual CLIP model. Once the dataset is encoded in the model, define the search function with two variables(query and k). Query stands for the search keyword, while k is the amount of output that returns top k results.[40]

```
# To encode text and images, a multilingual CLIP model is needed;
model_multi = SentenceTransformer('clip-ViT-B-32-multilingual-v1')
# Define the search function with the input of query and k value;
def search(query, k):
    # Encode the query;
query_multi=model_multi.encode([query],convert_to_tensor=True,show_progress_bar=False)
    # Computes the cosine-similarity between the query embedding and image embeddings, returns
the topk most relavent results;
    hits = util.semantic_search(query_multi, emb_art, top_k=k)[0]
    display(query)
    for hit in hits:
        print(img_names[hit['corpus_id']])
        display(IPImage(os.path.join(img_dir, img_names[hit['corpus_id']]), width=200))
```

The Query can be set to an object or a color tone. In Figure 16, the cluster contains the paintings with different query keywords of the two artists. It provides easy access for people to explore and compare colors and artistic styles of varied artists using the same parameter. It is clear that the "sunflower", as a vital feature in Van Gogh's works, is not seen in Rembrandt's paintings.

[40] https://github.com/ivaniadg/poststarts/blob/main/notebooks/clip_clustering_search.ipynb

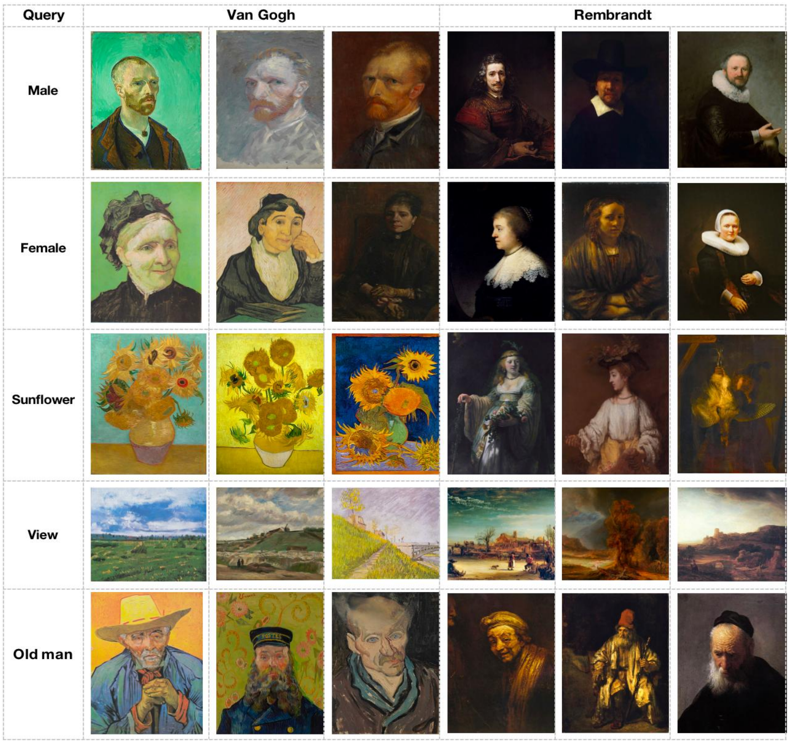


Figure 16. Search Output of Different Queries

The semantic search enhanced the image detection that facilitated the exploration of different features in artworks. It also enables comparison between artworks under a particular theme. For instance, the audience learns what other types of flowers Van Gogh painted other than his world-renowned “Sunflower” and exactly how many versions of sunflowers he created among his works.

5. Representation and Storytelling

5.1 Structure

In the previous analysis, I explored varied parameters (color, time, location) among the two artists to get their color preferences, pattern, and composition of specific paintings and the link between artworks, time, and locations. I have extracted the metadata from Rijksmuseum from the 17th to 19th century as the context information and processed the CLIP model for image clustering and main feature search of paintings.

The problem becomes how to link these findings in a way that triggers people's interest so they can understand the key message of the two artists. I realized that Rembrandt was considered an enormous inspiration for Van Gogh when he visited the Rijksmuseum in Amsterdam and even painted some of Rembrandt's works. Based on this link, why can't I complete the storytelling in the form of an online exhibition that connects the two artists, compare their different color preference, display famous paintings and present the patterns I found?

Thus, I made a website under the theme of "From Rembrandt to Van Gogh". The main structure of the website contains four main aspects: Color Palette, Galley, Patterns, and Data Source.

The color palette displays the preferable colors of the artists that allow people to get the different artistic styles and the link between color choice and the main theme of the artworks. The gallery represents example artworks and the color composition behind them for the appreciation of artworks. On the Patterns page, I summarized some critical patterns through my analysis and add the metadata dashboard of Rijksmuseum as the background information on the Data Souce page.

Image 3. The Front Page of the Website

5.2 Storytelling

5.2.1 Color Palette

The color palette page contains a Palette widget and the largest clusters(Top5) recognized by the CLIP model. In his paintings, Van Gogh uses low saturation colors to focus on still life, woman, self-portrait, and landscapes. The cluster and color palette logically fit with each other. The chosen color(hex code) can be copied by clicking on the color palette.

Image 4. The Page of Van Gogh's Palette

5.2.2 Gallery

Rembrandt's gallery is filled with brownish and grey colors, while Van Gogh's palette is vivid. By clicking on either image, a full view of the artwork or a tree map showing the color proportions of this painting will appear for people to appreciate the artwork while acquiring the dominant color.

Image 4. The Page of Rembrandt's Gallery

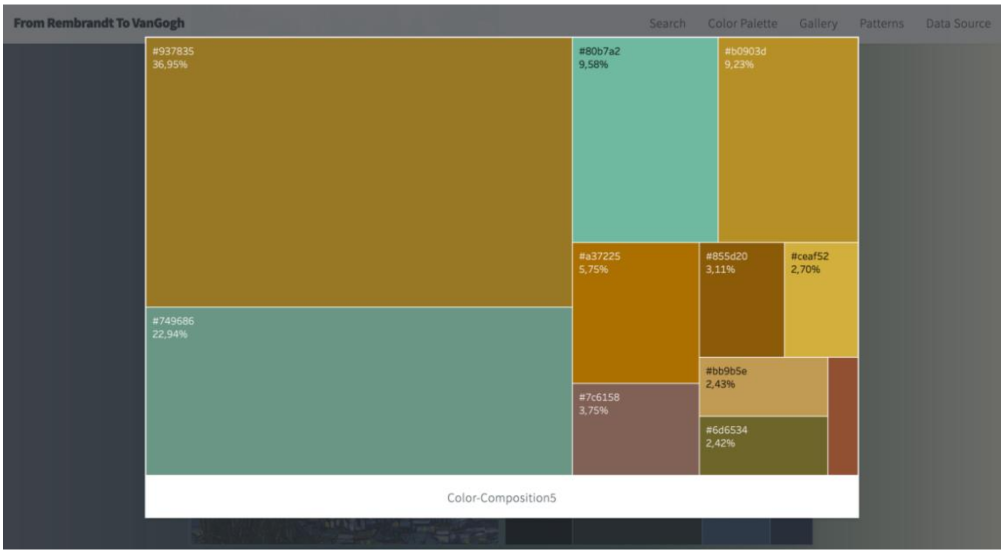


Image 5. The Page of Color Composition

5.2.3 Patterns

While most of Rembrandt's works were created in the Netherlands, Van Gogh traveled to various places to learn new techniques and find new inspirations. How did his selection of places affect his paintings? How did his color choice change over time and place?

The Patterns page reveals answers to these questions with a Multi-layer timeline and two interactive maps. The timeline(direction: from left to right) recorded Van Gogh's footprint as his art journey started from the Hague in 1881 and gradually moved towards the south to Paris and Arles. Among his Golden Decade of creation, his most extended stay was in Paris, and so is his most significant work. He spent almost equal time in the Netherlands and France. However, Paris was unique to him. Paris brought vibrant colors to his paintings, whereas most of his works were in dark tones before Paris. Vivid colors appear more in his paintings afterward, including his stay in Arles, Saint-Rémy, and Auvers-sur-Oise.

Image 6. Sample Page of the Patterns

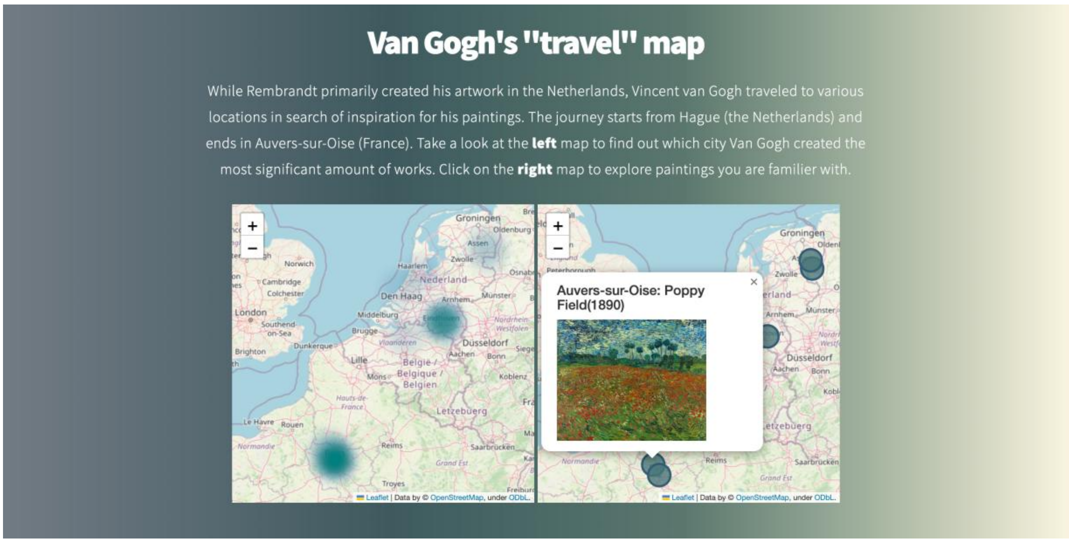


Image 7. Sample Page of the Patterns

To interact more with Van Gogh's trip, try to explore the following two maps. The left map shows his work amount by the different depths of color, while the right map lists one of his famous paintings created in a particular place. The audience clicks on the circle to explore the connection between his artwork and the specific location.

5.2.4 Metadata Overview

The metadata provides a selection parameter where users can choose among three different centuries(17th to 19th century). The information contains three dimensions (location, artist, and color) for people to get the background information and understand the context of Rembrandt or Van Gogh's time.

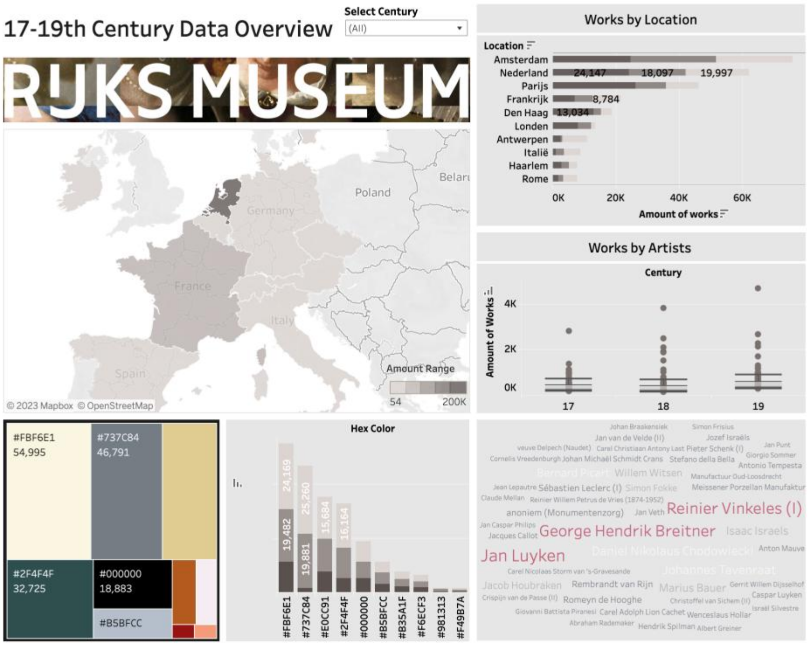


Image 8. The Screenshot of Metadata Overview

6. Conclusion

The exploration of attributes linking artists from different artistic styles transcends the limit of space and time, bringing the story alive to people. With the web interface, people could see for themselves as if viewing an online exhibition.

The representation of "From Rembrandt to Van Gogh" extracts valuable metadata, translates metadata into patterns, and converts patterns into complete storytelling. The exploration finds answers for understanding the current metadata status, getting the critical color patterns and possible links between the two representative artists---Rembrandt and Van Gogh, and eventually presenting the findings in a logical structure.

The result delves into connections between the two Dutch artists hailing from distinct centuries. They share a common theme of portraiture, with the notable distinction that Van Gogh's works feature sunflowers, a motif absent in Rembrandt's creations. Based on the color palette, Rembrandt's selection of color is dark and gloomy, which could be linked to his preference for drawing portraits and old people. At the same time, Van Gogh started with dark colors in his early paintings and shifted to more vivid color choices afterward, especially when he moved to Paris. Paris is the inspirational hub for Van Gogh, followed by Nuenen and Arles based on the amount of work.

As for the work amount, Rembrandt's artwork creation spread his entire life with a peak period around 1630, while Van Gogh's paintings were made during his golden decade from 1881 to 1890. Although the two artists are not from the same era, they produced a similar painting of an angel because Van Gogh got inspiration from Rembrandt's painting.

Reference

Arnold, W. N., & Loftus, L. S. (1991). Xanthopsia and van Gogh's yellow palette. Eye, 5(5), 503-510.

Barni, Mauro & Pelagotti, Anna & Piva, Alessandro. (2005). Image processing for the analysis and conservation of paintings: Opportunities and challenges. Signal Processing Magazine, IEEE. 22. 141 - 144. 10.1109/MSP.2005.1511835

Bobasheva, A., Gandon, F., & Precioso, F. (2022). Learning and Reasoning for Cultural Metadata Quality. Journal on Computing and Cultural Heritage, 15(3).

Callieri, M., Cignoni, P., Ganovelli, F., Impoco, G., Montani, C., Pingi, P., ... & Scopigno, R. (2004). Visualization and 3D data processing in the David restoration. IEEE computer graphics and applications, 24(2), 16-21.

Castellano, G., & Vessio, G. (2021). Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. Neural Computing and Applications, 33(19), 12263-12282.

Chang, M., Leggett, J. J., Furuta, R., Kerne, A., Williams, J. P., Burns, S. A., & Bias, R. G. (2004, June). Collection understanding. In Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries (pp. 334-342).

Desborde, R., & Marshall, K. P. (2015). Rembrandt Versus Van Gogh: A Qualitative Contrast Study Applying a Visual Arts Valutation Model.

Dragoni, M., Cabrio, E., Tonelli, S., & Villata, S. (2016). Enriching a small artwork collection through semantic linking. In The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29--June 2, 2016, Proceedings 13 (pp. 724-740). Springer International Publishing.

Émile Michel, "L'Exposition Rembrandt à Amsterdam," Gazette des BeauxArts 20 (December 1898): 478–480.

Gavrilov, D., Maev, R. G., & Almond, D. P. (2014). A review of imaging methods in analysis of works of art: Thermographic imaging method in art analysis. Canadian Journal of Physics, 92(4), 341-364.

Geldof, M., Proaño Gaibor, A. N., Ligterink, F., Hendriks, E., & Kirchner, E. (2018). Reconstructing Van Gogh's palette to determine the optical characteristics of his paints. Heritage Science, 6(1), 1-20. https://doi.org/10.1186/s40494-018-0181-6

Groen, K. (1997). Investigation of the use of the binding medium by Rembrandt. Zeitschrift fur Kunsttechnologie und Konservierung, 2(2), 208-211.

G. , & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. ArXiv. /abs/2103.00020

Hinchcliffe, G., & Whitelaw, M. (2015). Colouring digital collections: Challenges and opportunities for the use of colour metadata in cultural collections. Museums and the Web Asia.

Hulsker, J., & Gogh, V. V. (1996). The New Complete Van Gogh: Paintings, Drawings, Sketches: Revised and enlarged edition of the catalogue raisonné of the works of Vincent van Gogh. (No Title).

Kosch, H., Boszormenyi, L., Doller, M., Libsie, M., Schojer, P., & Kofler, A. (2005). The life cycle of multimedia metadata. IEEE MultiMedia, 12(1), 80-86.

Lee, S. G., & Cha, E. Y. (2016). Style classification and visualization of art painting's genre using self-organizing maps. Human-centric Computing and Information Sciences, 6, 1-11.

Li, Q. (2018). Data visualization as creative art practice. Visual Communication, 17(3), 299–312. https://doi.org/10.1177/1470357218768202

Lombardi TE (2005) The classification of style in fine-art painting. PhD thesis, Pace University, New York

Manovich, L. (2015). Data science and digital art history. International journal for digital art history, (1).

Messina, P., Dominguez, V., Parra, D., Trattner, C., & Soto, A. (2019). Content-based artwork recommendation: integrating painting metadata with neural and manually-engineered visual features. User Modeling and User-Adapted Interaction, 29(2), 251-290.

Peitgen, H. O., & Richter, P. H. (1986). The beauty of fractals: images of complex dynamical systems. Springer Science & Business Media.

Pickvance, R. (1986). Van Gogh in Saint-Rémy and Auvers. Metropolitan Museum of Art.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. ArXiv. /abs/2103.00020

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. ArXiv. /abs/2204.0612

Roy, A. (2012). Studying Rembrandt's techniques at the National Gallery, London. Technè Special Issue Rembrandt Approches Scientifiques et Restaurations, 35, 6-13.

Schama, S., & van Rijn, R. H. (1999). Rembrandt's eyes (p. 747). New York: Alfred A. Knopf.

Schapiro, M. (1968). The Still Life as a Personal Object—A Note on Heidegger and van Gogh. In The reach of mind: Essays in memory of Kurt Goldstein (pp. 203-209). Berlin, Heidelberg: Springer Berlin Heidelberg.

van Loon, A., Noble, P., Krekeler, A., Van der Snickt, G., Janssens, K., Abe, Y., ... & Dik, J. (2017). Artificial orpiment, a new pigment in Rembrandt’s palette. Heritage Science, 5(1), 1-13.

Van Noord, N., Hendriks, E., & Postma, E. (2015). Toward discovery of the artist's style: Learning to recognize artists by their artworks. IEEE Signal Processing Magazine, 32(4), 46-54.

van Rijn, R. H., van de Wetering, E., Groen, K., Broekhoff, P., Franken, M., & Binkhorst, L. P. (1986). Self-portrait. Art, Architecture and Engineering Library.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & Von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. Psychological bulletin, 138(6), 1172.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.

Windhager, F., Federico, P., Schreder, G., Glinka, K., Dörk, M., Miksch, S., & Mayr, E. (2018). Visualization of cultural heritage collection data: State of the art and future challenges. IEEE transactions on visualization and computer graphics, 25(6), 2311-2330.

Wu, S. (2013). A review on coarse warranty data and analysis. Reliability Engineering & System Safety, 114, 1-11.

Zujovic, J., Gandy, L., Friedman, S., Pardo, B., & Pappas, T. N. (2009, October). Classifying paintings by artistic genre: An analysis of features & classifiers. In 2009 IEEE International Workshop on Multimedia Signal Processing (pp. 1-5). IEEE.

Digital Resources:

CLIP, 2023 (https://github.com/openai/CLIP). Last consulted on the 8th of August 2023

Europeana, 2023 (https://pro.europeana.eu/post/the-yellow-milkmaid-syndrome-paintings-with-identity-problems). Last consulted on the 10th of August 2023

*Europeana API, 2023 (*https://pro.europeana.eu/page/apis). Last consulted on the 16th of May 2023

Github, 2023 (https://github.com/Mayurji/Explore-Libraries/blob/main/NLP Libraries/Image%20Clustering%20Using%20Sentence%20Transformer.ipynb). Last consulted on the 9th of June 2023

Github, 2023 (https://github.com/ivaniadg/poststarts/blob/main/notebooks/clip clustering search.ipynb). Last consulted on the 10th of August 2023

Medium, 2023 (https://medium.com/made-with-creative-commons/rijksmuseum-2f8660f9c8dd). Last consulted on the 7th of June 2023

Wikipedia, 2023 (https://en.wikipedia.org/wiki/RGB_color_model). Last consulted on the 7th of June 2023

Wikipedia, 2023 (https://en.wikipedia.org/wiki/List_of_works_by_Vincent_van_Gogh). Last consulted on the 7th of August 2023

Wikipedia, 2023 (https://en.wikipedia.org/wiki/List_of_paintings_by_Rembrandt). Last consulted on the 12th of August 2023

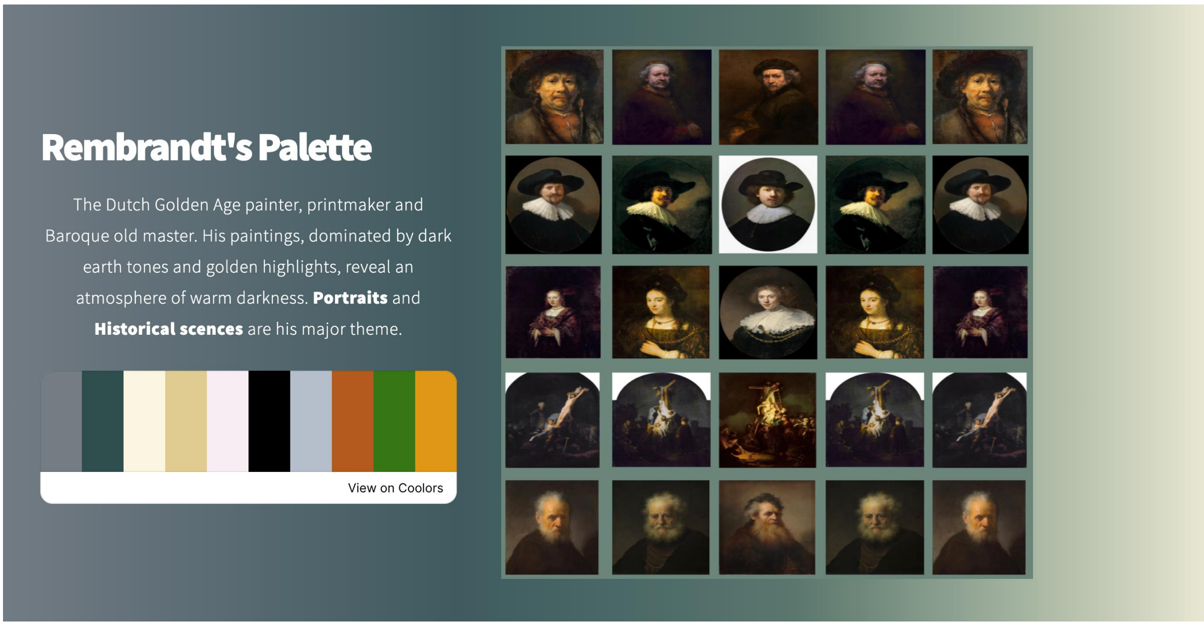
Rijksmuseum, 2023 (https://www.rijksmuseum.nl/en). Last consulted on the 7th of August 2023

Rijksmuseum API, 2023 (https://data.rijksmuseum.nl/object-metadata/api/). Last consulted on the 13th of June 2023

SentenceTransformers, 2023 (https://www.sbert.net/). Last consulted on the 7th of August 2023

Appendix

1. Website(Screenshots)



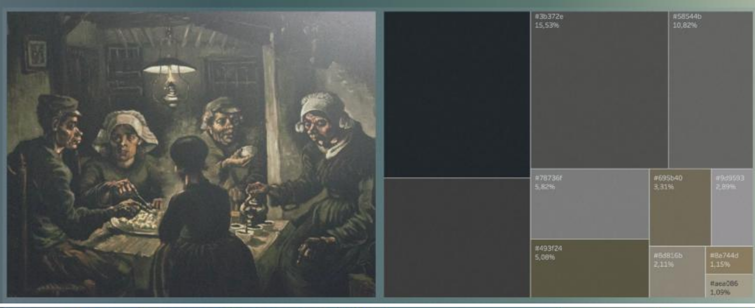

"Brownish and Gloomy"
He mastered a limited color range with earth tones, such as browns and ochres, creating a more subdued and tonal atmosphere in his paintings. His unique color choice is somehow affected by the focus on self-portrait, old man, beggar.
#291F11 56,70%
#7D5C39 17,50%
#957448 9,30%
#2F2718 15,30%
4,10%
#CEC9B6 4,10%
#755F42 3,10%
#BBACBE 3,10%
#4D4534 2,10%
#5E3C14 9,40%
#9C8238 3,10%
#AF9F6E

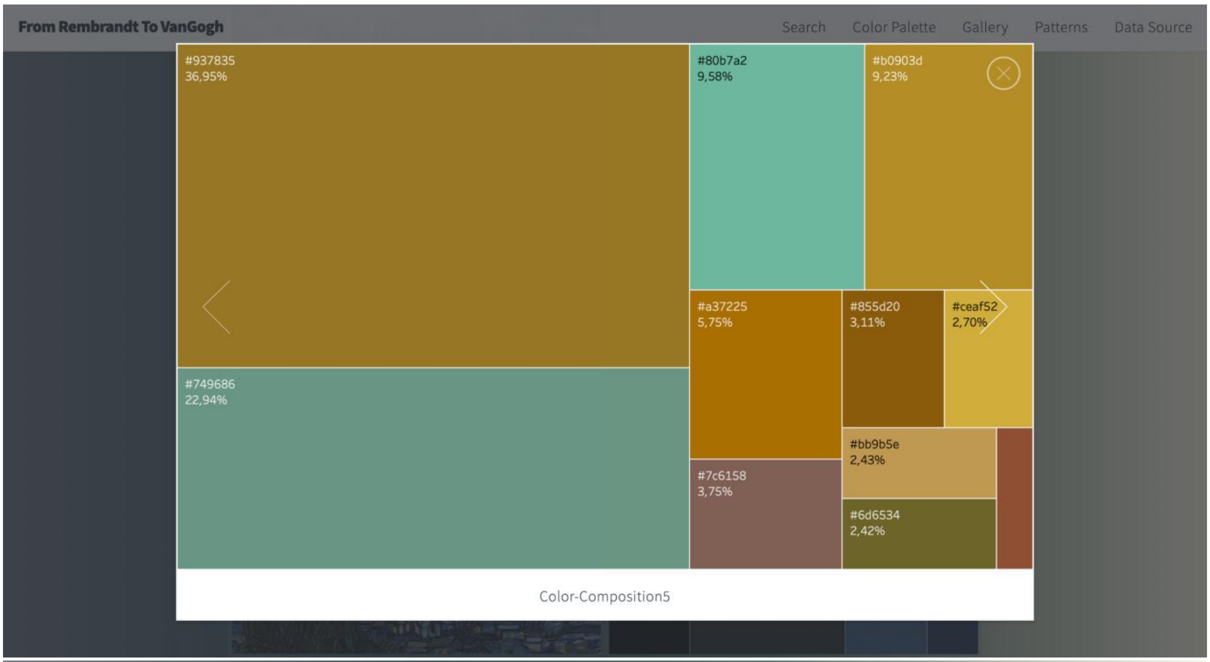
VanGogh's Palette

Dutch Post-Impressionist painter who's paintings are characterised by bold colours and dramatic, impulsive and expressive brushwork. He mainly focused on **Still life**, **Portrait**, **Landscape**.

"Vividly Evolve"

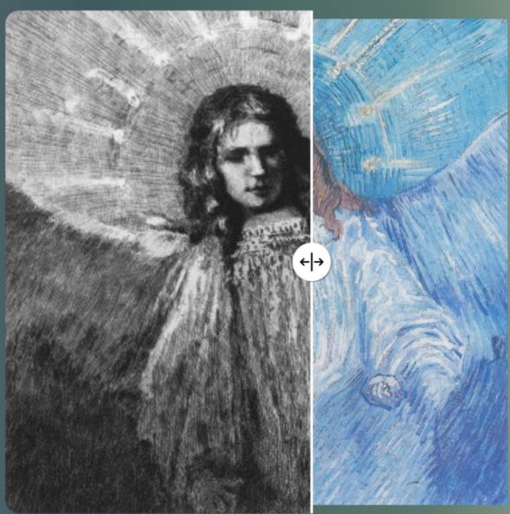
Renowned for his vibrant and expressive use of color, both dark colors and bright colors can be found in Van Gogh's artworks. His unique application of color, often in short brushstrokes or swirling patterns, created a sense of movement and energy. However, his early works are relatively darker when he lived in the Netherlands and Belgium. His color choices were influenced by the somber and muted tones of Dutch Old Masters and then evolve throughout his career.



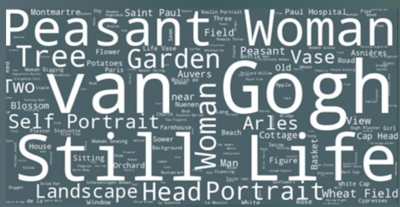


A painting that transcends time

The Angle Figures are painted by the two astonishing artists spanning two centuries.
(Left) Rembrandt van Rijn, 1655-1660;
(Right) Vincent van Gogh, 1889, Saint-Rémy.



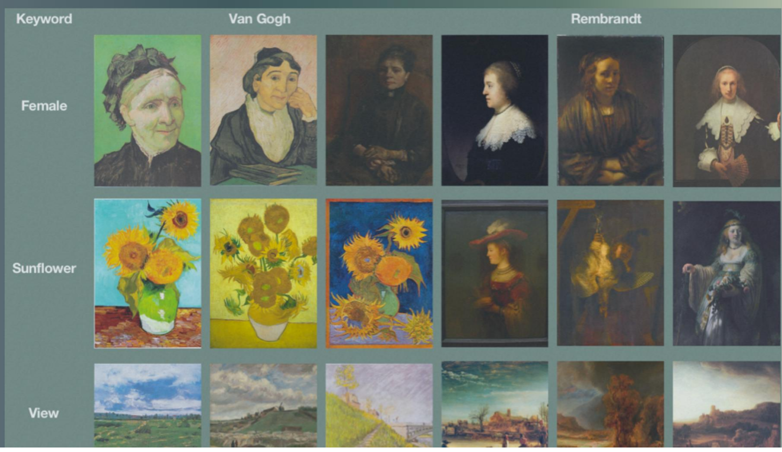
Theme in Common

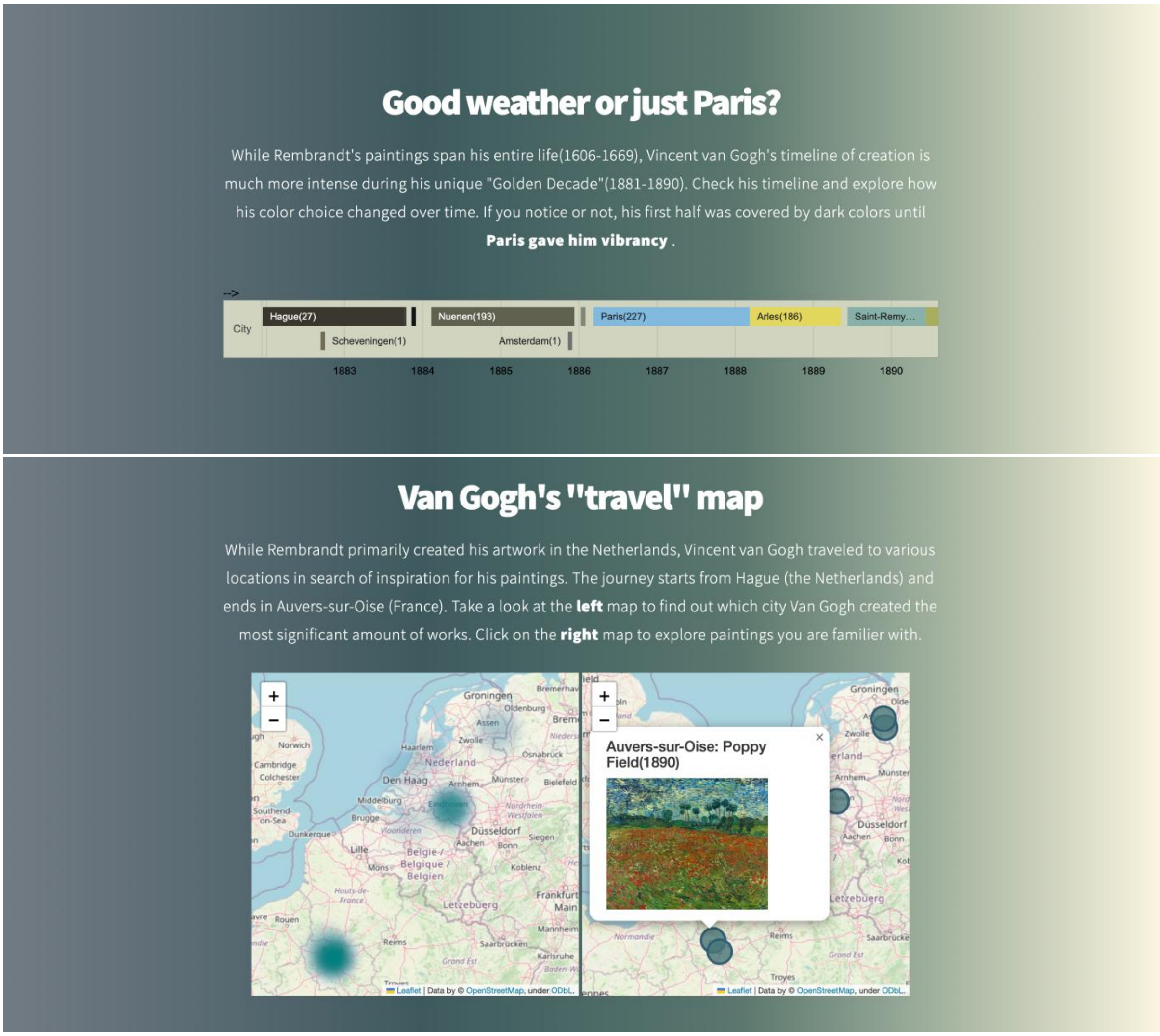


The two artists showcased their prowess in depicting distinct subjects. From their artwork descriptions and titles, it's evident that Rembrandt's primary focus lay in portraying elderly men, beggars, and creating self-portraits. In contrast, Van Gogh's artistry extended beyond landscapes, trees, and flowers, encompassing an adeptness at rendering portraits as well. Furthermore, both artists shared a common ground in their exploration of the theme of **"Portrait"**. Notably, historical scenes and religious contents also found a recurring presence within Rembrandt's canvases but rarely occur in Van Gogh's Paintings.

Keyword Search

By employing distinct search keywords, the CLIP model generates relevant images, illuminating the dissimilarities among artists. Through a comparative analysis of artworks sharing the same search terms, the contrasting artistic approaches between the creators become evident. Notably, Van Gogh's prominent theme of **"Sunflower"** is conspicuously absent in Rembrandt's oeuvre.





2. Code
   Github Code:https://github.com/effyvrversion/thesis
   Website URL:https://effyvrversion.github.io/Thesis-website/

AFDELING
Straat nr bus 0000
3000 LEUVEN, BELGIË
tel. + 32 16 00 00 00
fax + 32 16 00 00 00
www.kuleuven.be