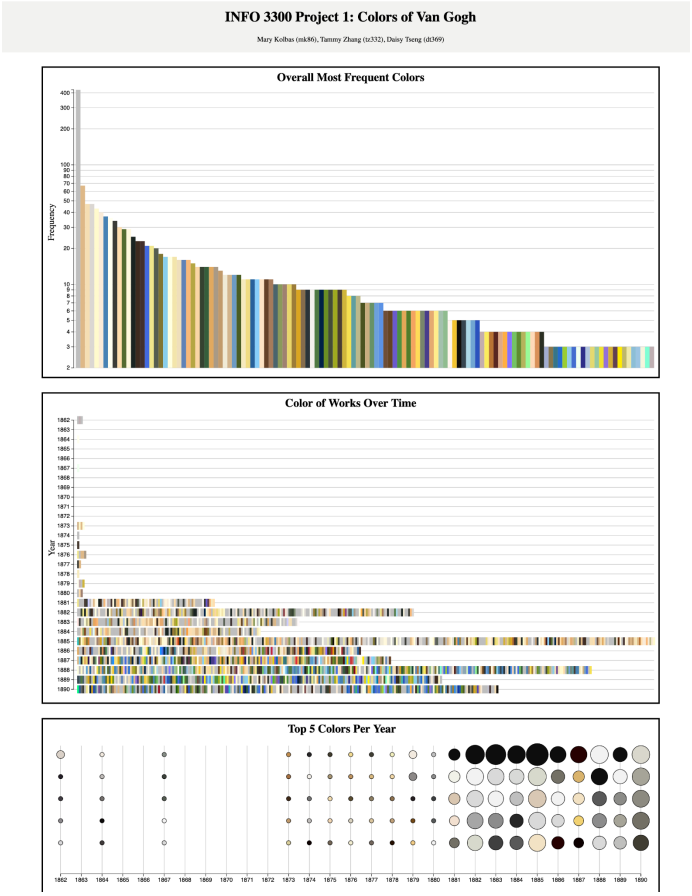
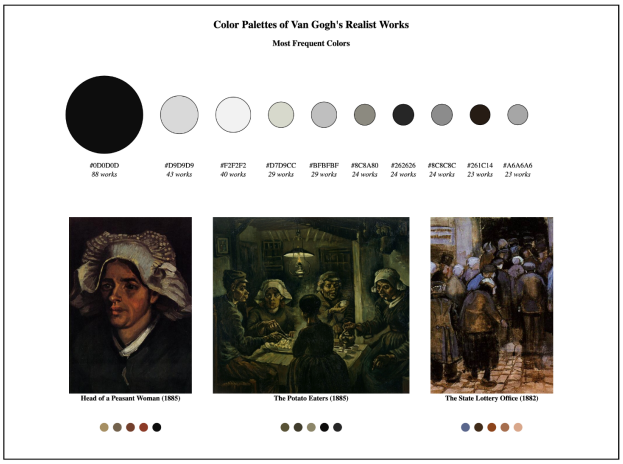
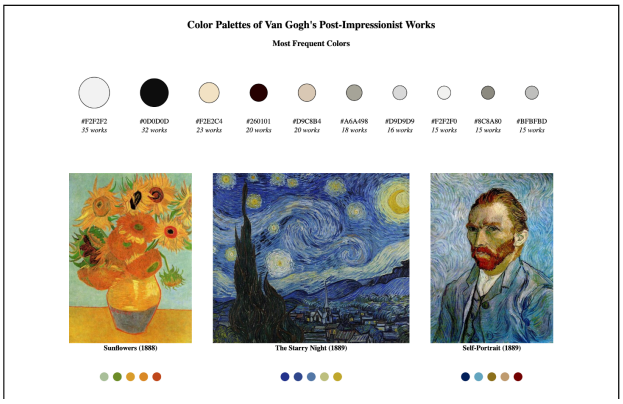
INFO 3300 - Data Driven Web Applications
Mary Kolbas (mck86), Tammy Zhang (tz332), Daisy Tseng (dt369)
11 October 2022

# Project 1 Written Report

## Final Visualization Screenshots:




## Data Description:

We got the "Colors of Van Gogh" dataset from Kaggle (source).

Given Data:

- df.csv: file holding the 1.) name, 2.) five principal colors used in hex code format, 3.) year painted, 4.) genre, 5.) style, and 6.) picture link for each of Van Gogh's works
- df_reduced.csv: file holding the same information as df.csv, but with the colors being binned into common name strings
    - e.g. '#241E26', '#3F3A40', '#8B888C', '#D8D7D9', '#BEBABF' becomes 'Gray', 'Gray', 'Gray', 'Gray', 'Gray', which reduces the number of unique values
- df_colorspace.csv: file holding each possible color common name and its corresponding rgb value

Processed Data:

- named_colors.csv: file holding binned color data, including the common name, rgb value, equivalent hex code value, and overall frequency
- long_merge_df_wbfy.csv: file holding the same data as named_colors.csv, but with associated year groupings included
- top5freq.csv: file holding the counts and hex codes for the top 5 most frequently used colors per year
- post_impressionism_colors.csv, realism_colors.csv: file holding the counts and hex codes for the top 10 most frequently used colors per style
- Other csv files: each file holds the data associated with one of the six displayed images of Van Gogh's work

Variables:

- Year: the year Van Gogh created the work (integer)
- Style: the art style used (either 'Post-impressionism' or 'Realism' for the purposes of this project, Van Gogh's two main preferred styles)
- Color: one of the five main colors used in the work (hex code in string format)
- Count: the number of works a color appears in
- Year_freq_counter: an index that counts from 0 to n date entries for each year

Reformatting and Filtering:

We did pre-processing in Jupyter Notebooks with pandas to convert the given df.csv data into a long dataframe, separating the 5 colors for each work for ease of plotting individual values and counting frequencies of color appearance. We grouped data by year and style to create easier-to-work-with csv files in JavaScript. We did not combine our dataset with any other datasets aside from the ones various csv files provided in this Kaggle dataset. We did consult several websites describing Van Gogh's most famous paintings to compile several for the first part of our visualization. We chose to focus primarily on data variables Colors, Year, and Style due to our purpose and story of showing his colors through the years and by style, though other variables (such as the name of the painting) helped us identify the specific data rows for the famous paintings we use in our visualization.

To work with binned colors, we did an inner join with df_reduced.csv and df_colorspace.csv so that each row had an rgb value corresponding to the color string. Using a function to convert rgb to hex, we then made a new column that held the equivalent hex code, which could then be interpreted directly by Javascript.

To generate frequency data, we used the pandas.DataFrame.value_counts() function on the column holding color data. After finding overall color counts, we sliced the dataframe to only include certain years or art styles to get a more specific look at how color use differed across those categories.

**Design Rationale:**

Visualization 1) Overall Most Frequent Colors:

Before delving into colors by year and style, we wanted an overview of the most frequent colors that appear in Van Gogh's paintings. The marks are rectangles and the channels are hue and vertical aligned positioning. We used the named_colors dataset where several hex codes are binned together. While the level of detail is a limitation and tradeoff, it enables the audience to get a more general overview and understanding of the top colors present in Van Gogh's paintings. We made this decision because plotting every unique hex code would have resulted in a very large visualization and made it difficult for the audience to draw any conclusions due to the sheer amount of colors presented.

We also utilized a logarithmic scale over a linear scale due to the large outlier of the gray color frequency. It would have been difficult to see all the rest of the colors due to the outlier. Thus, we made this tradeoff because we compared the two scales and decided that the final logarithmic scale visualization makes it clear that the gray color is most frequently used in his paintings, while still keeping the other colors visible. Lastly, in the data cleaning of the dataset file, we removed colors with a frequency of only 1 and 2 because the purpose of this specific chart is to provide the audience with an overview of the top colors present.

Visualization 2) Color of Works Over Time:

We created a visualization to display all the color data points in a bar chart over time, displaying how the colors used change throughout the years. The marks are rectangles and the channels are hue and horizontal aligned positioning. We used the df_reduced dataset, where certain hex codes are grouped together. This means that several different hex codes are all labeled "gray" and displayed in this visualization as the same shade of gray. Although this reduces the detail of this visualization, it makes it much more comprehensible than plotting every unique shade in hex code as is (which we tried, and is very difficult to draw conclusions from because some rows have over 1500 entries). By looking at this visualization, the viewer is able to see the difference between colors used prior to 1886 and after 1886, where after around this year there are more instances of blue and other lush colors.

Visualization 3) Top 5 Colors Per Year:

We created a visualization to display the top 5 colors from Van Gogh's paintings per year to see how his use of different colors changed over time. The marks are circles and the channels are hue and vertical aligned positioning. We limited the number of colors to be displayed by aggregating and counting the frequency of each color per year, based on hex code. We also wanted to use the size of each color circle to compare the frequency of the top 5 colors within the year, offering a comparison and clear visualization of whether colors were evenly used or if there were some dominating colors.

We used a logarithmic scale to determine the radius of each circle based on the range of all frequencies in the dataset. This scale provides more visible details for smaller values and

between closely related numbers, while underestimating large frequencies. This is a tradeoff we made in order to make the comparison between circle colors and sizes more visible, while still preserving a scale that allows reasonable comparison across years. We also decided to outline the circles in a black border so that white/lightly colored circles and their size were visible to the reader.

Visualization 4 and 5) Color Palettes of Van Gogh's Realist Works (Most Frequent Colors) and Color Palettes of Van Gogh's Post-Impressionist Works (Most Frequent Colors):

Van Gogh's two primary art styles were Realism and Post-Impressionism; the number of works he created for each was roughly equal. To take a deeper look at the breakdown of Van Gogh's color use, we decided to display the top 10 colors used in each style. The marks are a circle for each color and the channels are hue and area size. Each circle's radius is the number of times the corresponding color appeared in works for that style. The colors are ordered left to right in order of frequency, following the reading pattern of English - most used colors appear as bigger circles, which decrease in size moving to the right.

A tradeoff of using circle areas to represent frequency is the ambiguity in comparing relative counts between different colors, which was a tradeoff we chose to make because these two visualizations are intended to be more simplistic and qualitative in nature - the hue is the most important component, and exact quantitative comparisons are not as relevant in this context.

For each style, we also provided three of Van Gogh's most famous works for that style underneath the visualization. This was intended to be an aesthetic supplement to emphasize the palette differences between Realism and Post-Impressionism, as well as encourage viewers to contrast Van Gogh's highly neutral palettes to the bright colors used in his most well-known works.

**The Story:**

As viewers scroll down, the visualizations move from general (Van Gogh's overall color use) to more specific (color use breakdowns over years and for different styles). Viewers can see through the first overall frequency chart that Van Gogh utilized the color gray most often in his works. He also favored neutral colors and blue shades over red and saturated tones.

After delving into colors more specifically over time in our second visualization, viewers can see through the use of varying hues that generally speaking, Van Gogh's paintings in earlier years used mostly neutral colors, while paintings in later years incorporated more vibrant colors.

When looking at the top 5 colors by year in the third visualization, it was surprising to find that despite adopting some brighter colors into his palette later on, Van Gogh remained relatively consistent in his overall most preferred colors - he tended to use neutral and dark colors the most in his work.

We then investigated style breakdowns in our last two visualizations. Although the size of Van Gogh's body of work for his earlier (Realist) and later (Post-Impressionist) periods are

roughly the same, there are notable differences in the colors most frequently used for each. Van Gogh's Realist works heavily incorporate darker, monochrome shades, in line with his often-used subject of the realities of poverty at the time (where he depicted peasants and the lifestyles of the lower class). The top 10 colors are overall represented by larger circles in Van Gogh's Realist palette than the top 10 colors of his Post-Impressionist palette, suggesting a heavier skew towards the most frequently used dark colors - and implying that in contrast, Van Gogh used a greater variety of colors in his Post-Impressionist works, which mostly depicted landscapes and still lifes.

There is also some interesting contrast to be found between our conclusion of Van Gogh's most common colors overall mostly involving neutral and gray hues and the general public's impression of Van Gogh's most famous works, which almost invariably include rich blues, bright yellows, and deep greens - colors associated with a select few of his Post-Impressionist works. This suggests that Van Gogh's most colorful Post-Impressionist works were the most likely to make an impact on the public, likely for aesthetic reasons - and also suggests that much of Van Gogh's works, like his Realist paintings, may go relatively unnoticed in comparison.

**Outline of Team Contributions:**

Mary:

- Cleaning/Processing data in Jupyter Notebook - 3 hours
- Colors by Year - 9 hours
    - Avg color by year (scrapped) - 3 hours
- Top 5 Colors per Year Chart - 7 hours
- Contribution to written report - 2 hours

Daisy:

- Experimentation with coding possible charts - 4 hours
    - Circles graph with palette (scrapped) - 1 hour
    - Emergent layouts bubbleplots attempts (scrapped) - 2 hours
- Overview of top frequent colors visualization - 6 hours
- Combine and clean all separate files, organize for final submission - 1 hour
- Layout cleaning of final page - 4 hours
- Written report - 2 hours

Tammy:

- Cleaning/Processing data in Jupyter Notebook - 5 hours
    - Created Python functions to generate csv files in Javascript-ready format - 4 hours
- Post-Impressionist / Realist color palettes and example pictures - 8 hours
- Layout cleaning - 4 hours
- Written report - 2 hours