Van Gogh in the Age of Computers

AUTHORS REDACTED

Summary of Questions and Results

For this project, our research questions were:

1. How did the colors and styles Van Gogh used change over time?
    a. We found that the color gray was used the most (361 times) in 1885; the color black shadows was used the most (6 times) in 1890; the color snow was used the most (12 times) in 1885 and 1889; the color seashell was used the most (8 times) in 1890; the color ivory was used the most (31 times) in 1885; and the color honeydew was used the most (1 time) in 1867, 1883, and 1890.
    b. We found that the styles Van Gogh used the most were realist and post-impressionist. The most realist paintings (312 works) were created in 1885, the most post-impressionist paintings (273 works) were created in 1888, the most neo-impressionist painting (34 works) were created in 1887, the most japonist paintings (13 works) were created in 1888, and the most cloisonnist paintings (5 works) were created in 1888.
2. What colors were used most in each genre?
    a. We found that *sketch and study works* used mostly grays and creams; *animal paintings* used mostly grays, deep yellows, greens, and browns; *still lifes* used mostly grays and browns; *landscape paintings* used mostly greens with some grays, reds, and yellows; *genre paintings* used mostly grays and muted colors; *cityscapes* used mostly grays, browns, and creams; *portraits* used mostly grays, browns, and creams; *flower paintings* used mostly vivid reds, yellows, and green; and finally, *self-portraits* used mostly reds, yellows, and creams.
3. Can we create an accurate model to predict the style of a painting based on data such as the colors it contains and the year it was painted? And according to our model, what is the most important feature for determining the style of a painting?
    a. We found that an accurate model for predicting the style of a painting has the max depth 24 and a prediction accuracy of approximately 0.97 on test data. The most important feature for determining the style of a Van Gogh painting is the year it was painted.
4. What topics did Van Gogh paint about most?
    a. We found that the topic Van Gogh painted about most frequently was women, followed by men, portraits, landscapes, boats, still lifes, flowers, female nudes, rivers, and trees.

Motivation

Vincent Van Gogh has long been a prominent figure in the art world, well-known for works like his sunflowers and *The Starry Night*. His influence has only grown since his spike of popularity in the twentieth century, and while the post-Impressionist has been hailed as an innovator and master of various techniques, a comparatively small amount of digital analysis has ever been done to identify trends and patterns across his works. Thus, our primary motivation for this project has been to conduct this analysis to gain a better understanding of how he evolved throughout his career and discern relationships between different elements of his artwork.

We are also motivated to pursue this project to enhance the archival and restoration process of paintings. Given one of unknown origins, we could predict preliminary identifying information such as the style of the painting by training a machine learning model with the same data used for digital analysis. Furthermore, through interpreting our model, we can help restorationists determine what elements are most important for accurate classification. In other words, this project is the next step towards transforming the preservation of artwork to be more accurate and efficient.

Dataset

We are using the “Colors and Van Gogh” dataset found at https://www.kaggle.com/pointblanc/colors-of-van-gogh and made by user Konstantinos, and can be downloaded at that link by clicking the “Download (1GB)” button on the upper right side of the screen, below the header.

This dataset consists of three main files:

- df, containing the name, colors (in hex code format, selected using Adobe Color), year, genre, and style for 1,931 of Van Gogh’s works, all web-scraped from WikiArt, along with a link to a JPG of the painting on WikiArt
- df_reduced, containing the same information as df but with the colors in color name format (i.e. “Goldenrod”)
- color_space, containing the name of a color and the color’s RGB values

Of the three, most of our work will be done with df_reduced, though we will also use df in some questions so we can use the colors we’re mentioning in our visualization. For example, if we had a bar in a bar graph for the color “Goldenrod”, we would graph that bar with the hex code “#D9B573” so viewers can easily see what the dataset meant by “Goldenrod”.

We will also be using data that we gather from the Met Museum API, the documentation for which can be accessed at this link: https://metmuseum.github.io/ .

Method

Research Question 1:

To answer the question, *How did the colors and styles Van Gogh use change over time?*, we ended up:

1. Processing df_reduced and df so each color and hex code for each painting had its own row, then joined the two datasets together.
2. Creating a list of the unique colors in the processed data.
3. Looping through each color in the list, and for each color:
    a. Finding the unique years in the processed data.
    b. Filtering the processed data to just that color.
    c. Merging the unique years and filtered dataset.
    d. Counting the number of times the color was used each year.
    e. Filling in missing values in the Color column with the color.
    f. Converting the Year column into datetime format.
    g. Finding the first hex code corresponding to the color.
    h. Creating a time series using **a new library, bokeh**, where years are on the x-axis, the count for that color is on the y-axis, and the line is colored using the hex code corresponding to the color represented.
4. Adding the time series for each color to a single figure that viewers can scroll through, allowing them to compare graphs to one another and hover over each line to see the color represented, its count, and the year.
5. Repeating steps 2-3 to count and graph the number of times each style was used each year and stacking the time series for each style.

These computations and graphs will allow users to see how the colors and styles Van Gogh used changed over the years which answers our first research question.

Research Question 2:

To answer the question, *What colors were used most in each genre?*, we ended up:

1. Creating a list of the unique genres in the df_reduced file with more than 15 paintings (so as not to mislead users by implying Van Gogh frequently painted genres that he did not, and not try to draw conclusions based on very little data).
2. Looping through each genre in the list, and for each genre:
    1. Filtering the df_reduced and df files to just that genre.
    2. Joining the two filtered datasets together.
    3. Counting the number of each color.
    4. Selecting the top 10 colors by count.

    5. Creating a bar graph using **a new library, bokeh**, where the x-axis is the color name, and the bar height is the count for that color, and each bar is colored using the hex code corresponding to the color it represents.
3. Adding the bar graphs for each genre to a single figure that viewers can scroll through, allowing them to compare graphs to one another and hover over bars to see their counts.

These computations and graphs allow users to see the top ten colors for each genre, showing what colors were used most in each genre and answering our second research question.

Research Question 3:

To answer the question, *Can we create an accurate model to predict the style of a painting based on data such as the colors it contains and the year it was painted?* and *According to our model, what is the most important feature for determining the style of a painting?*, we ended up:

1. Creating a DecisionTreeClassifier model with the Style column as labels and the Name, Color, Year, and Genre columns as features.
2. Splitting data into 10% testing data, 10% validation data, and 80% training data.
3. Looping through possible max depth levels, and for each level:
    a. Training the DecisionTreeClassifer on the training data.
    b. Creating validation predictions and testing those predictions' accuracy.
    c. Creating test predictions and testing those predictions' accuracy.
4. Comparing accuracies to determine the max depth level that maximizes validation prediction accuracy to limit overfitting on the test data.
5. Creating a list of the feature names and feature importances of the features used to train the model with the chosen max depth and sorting the list by importance.
6. Creating a dataframe with columns Feature containing feature names, Importance containing feature importances, and Color containing hex codes from a palette.
7. Creating a bar graph with the **new library, bokeh**, where the x-axis has feature names, the bar heights represent feature importances, each bar is colored using the corresponding hex code in the Color column, and viewers can hover over bars to see the importances represented.

This computation and graph will show users what features the model considered most when attempting to determine the style of a painting, answering the "what was the most important feature" question. We can also provide the prediction accuracy for testing data as a way to answer the "can we create an accurate model" question.

Research Question 4:

To answer the question, *What topics did Van Gogh paint about the most?*, we ended up:

1. Querying the Met Museum API using a **new library, requests**, for a list of all the Van Gogh paintings in the museum in order to use this **messy data**.

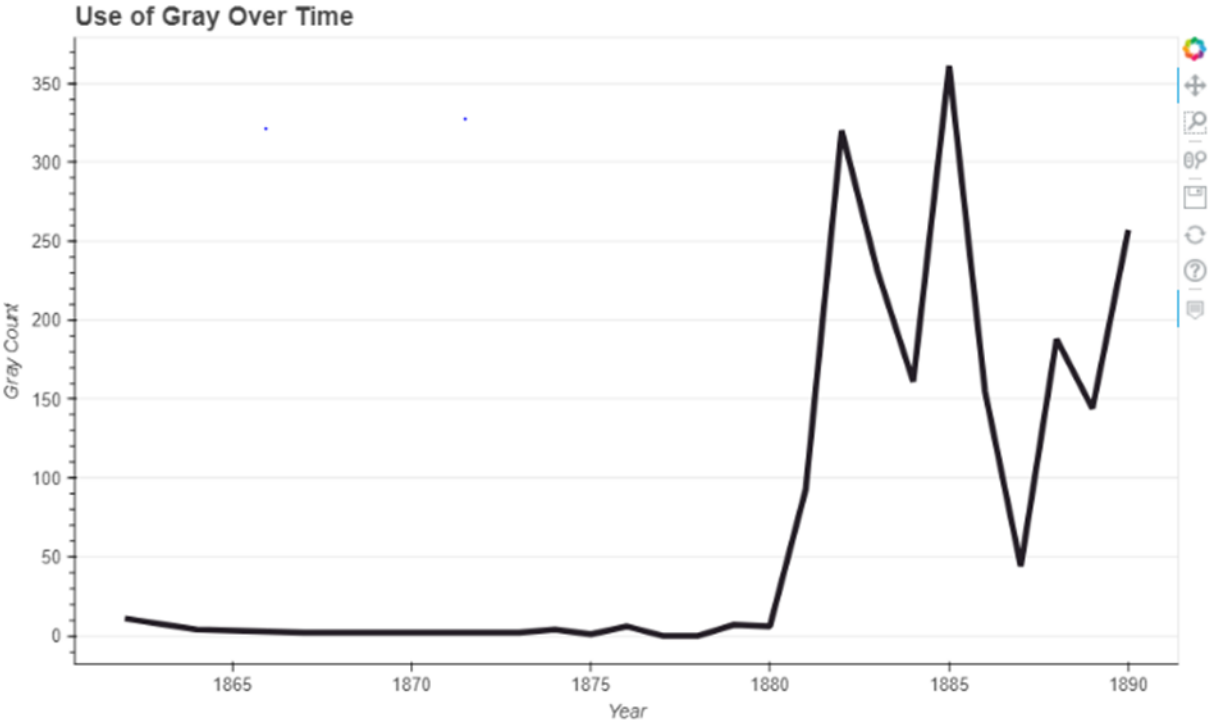
2. Looping through each item of the list, taking the objectID it returns, and querying the API again with that id to get back JSON for the painting.
3. Accessing the tags key for each painting, and for each tag the painting has, either adding the tag to a dictionary with a count of 1 if it is not already in the dictionary, or increasing the count for that tag by 1 if it is. If the tags key is empty, the dictionary is unchanged.
4. Converting the dictionary to a pandas dataframe with columns ‘Topic’ and ‘Count’.
5. Sorting the dataframe by ‘Count’ and returning the ten topics with the highest counts, in reverse order.
6. Creating a bar graph with the **new library, bokeh**, to show the number of times Van Gogh painted each tagged item or topic for the top 10 items or topics he painted about most.

This computation shows users the top 10 most common tags that the Met Museum assigned to their Van Gogh paintings, and since tags represent topics and subjects (an example set of tags might be “Portraits”, “Women”, and “Flowers”), the top 10 most common tags will show us the topics Van Gogh painted the most. While we can only look at this tagging data for the Met Museum, we can likely assume that the museum’s 229 Van Gogh paintings are a random and thus representative sample of all of Van Gogh’s works, so our findings based on these 229 paintings can be generalized across all of Van Gogh’s paintings.

Results

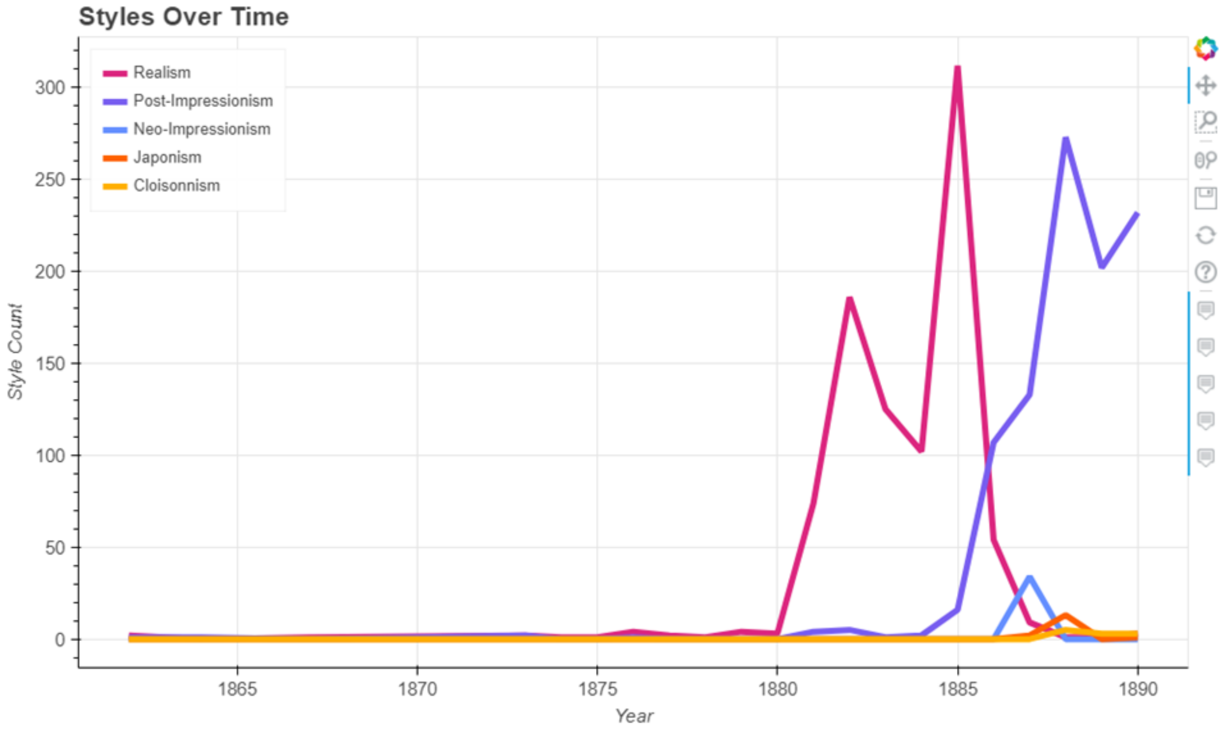
Research Question 1:

To study color usage over time, we only generated 6 graphs for 6 different colors because creating a time series for every color in the dataset would make testing inefficient. All 6 graphs can be found in graphs/q1-1.html, but running our code should also open them in a browser. Since every graph contains the main components of a time series produced by our code for any color in the dataset, we will only discuss the time series representing the uses of gray over time.



One important takeaway from these graphs is when the color was used the most and least. By looking at the interactive version of this graph, we learn that gray was used 361 times in 1885 and 0 times in 1877 and 1878 by hovering over the line. Interestingly, since gray was the first color in the dataset, we also know that the color was used in the earliest Van Gogh work in the dataset. One potential explanation for this is that oil painters are often taught to produce monochrome paintings first to learn about lighting and those paintings would include gray. Furthermore, we might interpret the 361 uses of gray in 1885 as an increase in the percentage of gray in his works because he created more sketches that year (and sketches use pencil, creating shades of gray). However, this spike in gray could also be influenced by an increase in the number of works produced that year. To validate this interpretation, we could conduct a future study to represent the most frequently used colors each year and their percentage out of all the colors used that year in a stacked bar graph.

To study the number of paintings in various styles over time, we generated a stacked time series with each line representing a style in the dataset and its use. We chose not to create a separate time series for each style because the dataset only had 5 styles total and they are easier to compare on the same graph. This graph can be found in graphs/q1-2.html, but should also open automatically when our code is run.



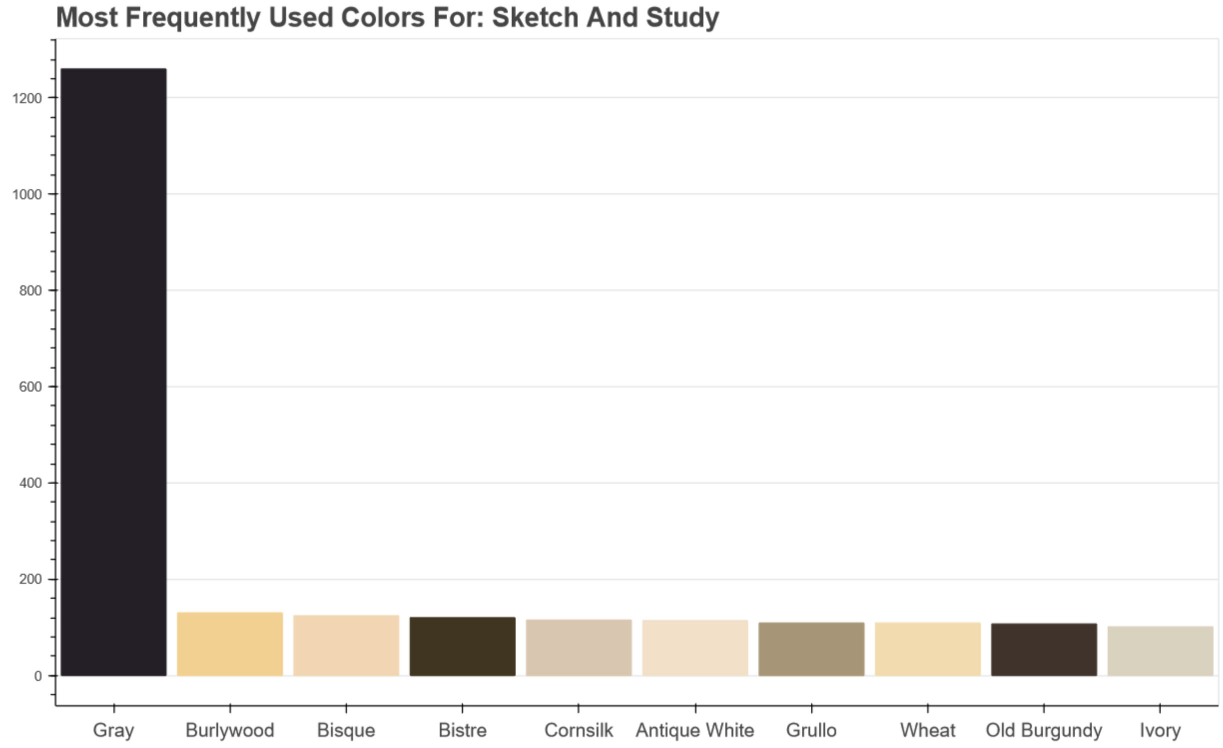
This graph demonstrates that Van Gogh was primarily a realist and post-impressionist painter, though he tried at least 3 other styles according to the dataset. By looking at the interactive version of this graph, we can see that he created the most realist paintings in 1885 and the most post-impressionist paintings in 1888 (312 and 273 pieces respectively). This is surprising because Van Gogh is known for being a post-impressionist painter, but our graph suggests that he produced the same if not more realist paintings overall. However, we can also see a distinct shift from realism to post-impressionism later in his career as an artist, which could explain why he is more closely associated with this style today.

The intersection of the lines for realism and post-impressionism in 1886 is representative of the shift from realism to post-impression as his main style. In that year, Van Gogh produced 54 realist paintings and 107 post-impressionist paintings. *The Starry Night,* his most famous post-impressionist piece, was created in 1889 when a total of 202 post-impressionist paintings were produced. This implies that Van Gogh needed at least 2 years with post-impressionism as his main style to achieve the level of mastery necessary for producing this piece. The fact that *The Starry Night* was painted near the end of his career could also explain why it is better remembered than some of his other works.

Research Question 2:

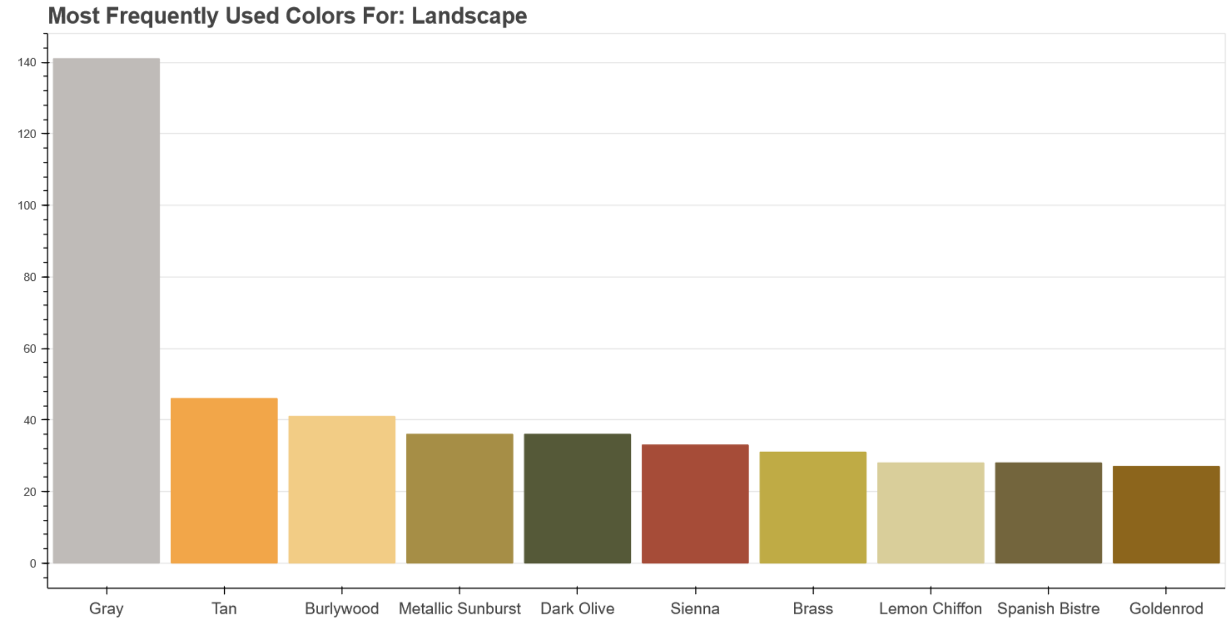
For this question, we looked at all the genres that had data on at least 15 paintings (a total of 9 genres) and created a bar graph showing the top 10 colors and their counts. The full 9 graphs can be found in graphs/q2.html in our code, but should also open automatically after running our code. For the sake of brevity, we will examine only the ‘sketch and study’ and ‘landscape’ genres because they had the most interesting results.

As shown in the bar graph below, we found that for sketches and studies, the most frequently used color was gray, being used 1,260 times. This is surprising because it’s used 1,130 times more than the next most frequent color - Burlywood, at 130 times. This can be explained by the way the data was collected. The original Kaggle dataset extracted 5 colors from each painting using Adobe Color. Those colors were stored as hex codes, but to make the data more readable, the dataset author also translated those hex codes to color names, many of which overlap. For example, the hex codes '#241E26', '#3F3A40', '#8B888C', '#D8D7D9', and '#BEBABF' (among many others) were all encoded as ‘Gray’. Therefore, all different shades of gray were marked as the same color and totaled together. This phenomenon is visible in the other genres as well, but is especially prevalent in sketch and study, likely because most sketches are done in pencil and there is a large volume of these works in the dataset. One explanation for this volume is the relative speed and simplicity to create a sketch or study, compared to the oil paintings of the other genres. The creams and whites also visible in the top 10 (burlywood, bisque, cornsilk, antique white, wheat, and ivory) are likely the color of the paper, which tends to yellow as it ages.



If we look at the graph for landscapes, we can see a somewhat similar story to the graph for sketch and study, where the most frequent color was gray, occurring 141 times. However, we see a greater variety in color afterwards, including greens, browns, some yellow, and one red color - Sienna. This may be surprising in that Van Gogh’s arguably most famous work, *The*

Starry Night, is a landscape, and yet the predominant colors of that work (blues, with some black and bright yellow) are mostly absent from the most frequent colors for this genre. This tells us that *The Starry Night* was unique among his other works of the same genre, and suggests that the work being more unique may have resulted in that work being more famous, potentially because its distinctiveness caused it to be better remembered.

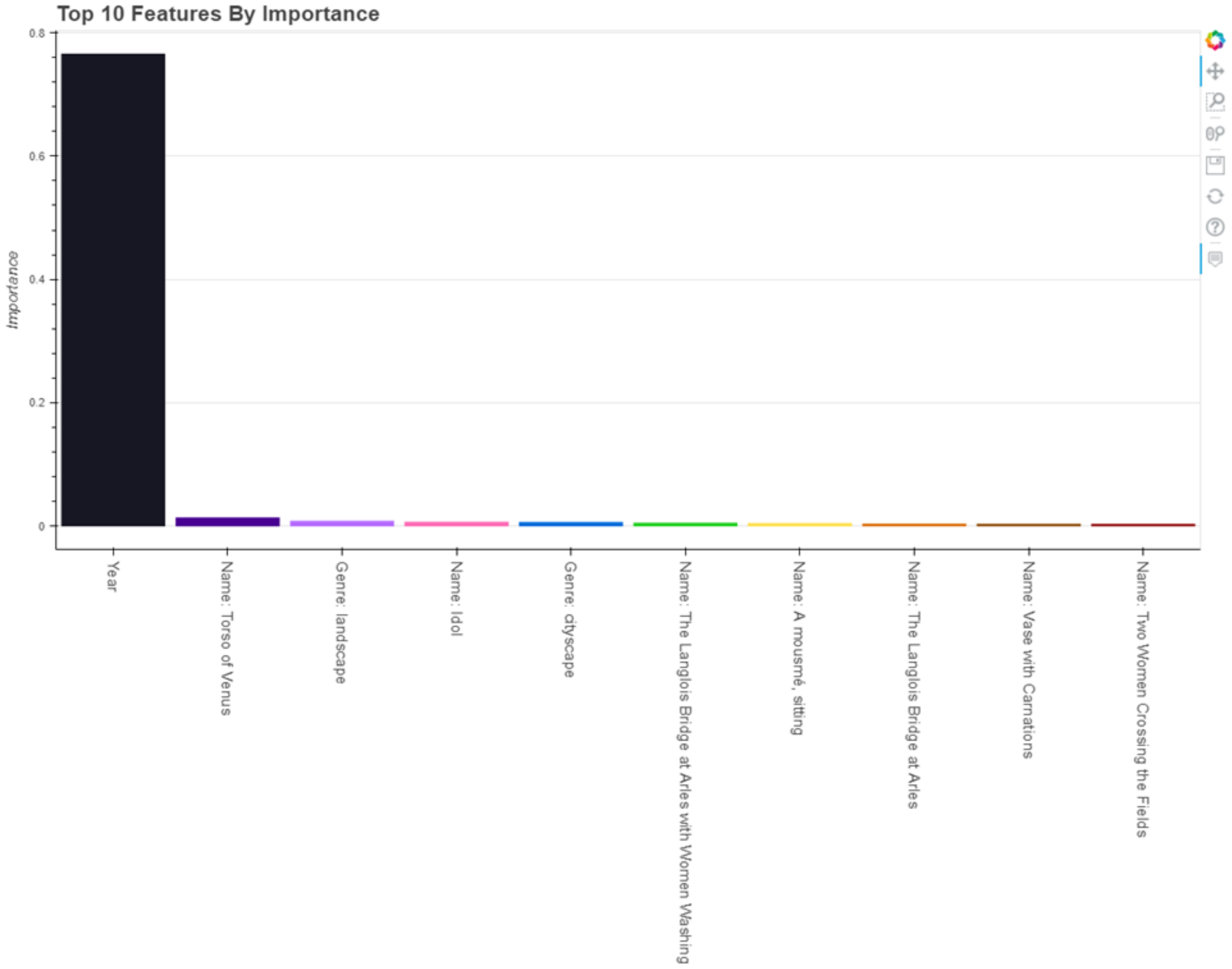


By looking at the most frequently used colors per genre, we can get a better understanding of how many works of each genre Van Gogh completed, what a "typical" work for each genre may look like, and, as our analysis of the landscape genre results pointed out, what may cause certain works to be more popular or famous. This could be interesting for artists, art critics, art historians, and even psychologists to examine in more detail.

Research Question 3:

Our result answers the first part of this research question about creating an accurate model to predict the style of a Van Gogh painting. By calculating the prediction accuracy of our DecisionTreeClassifer model for validation data at different max depths, we determined that the best value to maximize accuracy is 24. With this, our model had a prediction accuracy of approximately 0.97 on test data. Our result suggests that we can create an accurate model to predict the style of a Van Gogh painting using data such as its name, genre, and the year it was painted. Given the data to train a model, the same could also be done to predict how the paintings of other artists should be classified.

To answer the second part of this research question (what feature was most important for a prediction), we generated a bar plot with a bar to represent each of the top 10 most important features for predicting the style of a Van Gogh painting. The height of each bar was the importance of the feature. The graph can be found in graphs/q3.html and should open automatically when our code is run. We only represented the top 10 most important features because features with lower importances are unlikely to help predict the classification of other artist's paintings.



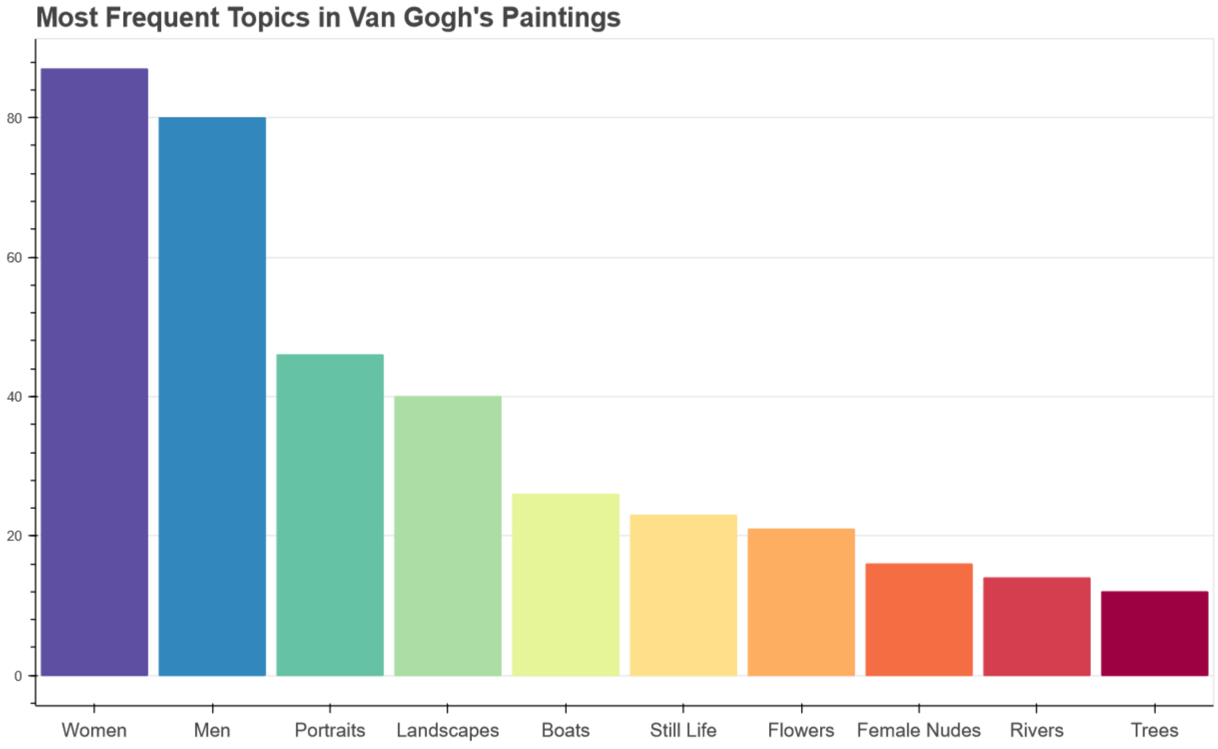
One interesting takeaway from this graph is that year is by far the most important feature for predicting style, with an importance of approximately 0.76. As stated in the motivation section, we hope to inform and enhance the process of archiving artwork, and this result may suggest that if any new Van Gogh paintings were discovered, archivists could use the year of its creation to help predict its style. The importance of year could be explained by our results for Research Question 1, where we discussed how each style varied over time. To potentially transform the overall process of categorizing artwork, we could conduct a future study where we compare our model with models that are trained using data about other artists' paintings.

Genre and name features are also shown in Top 10 Most Important Features, but the actual importances of these features are low in comparison to year. Therefore, archivists likely should not rely on these features to help predict the style of a newly discovered Van Gogh painting. The low importances of these features could be because paintings often have unique names, so a particular name is only useful for predicting the style of one painting. Furthermore, there are many more unique genres than years in the dataset, meaning a particular genre is also associated with fewer paintings.

Research Question 4:

As mentioned in Method, while we only looked at the 229 Van Gogh paintings that the Met Museum had data on, we can assume that those paintings are a random, and thus representative, sample of Van Gogh's works. Therefore, we can also assume the conclusions we draw from the Met Museum's data can be applied to the rest of his works.

From looking at the Met Museum data, we found that Van Gogh's most commonly painted topic was women, which was tagged on 87 of the 229 paintings, followed by men, at 80; portraits, at 46; landscapes, at 40; boats, at 26; still life, at 23; flowers, at 21; female nudes, at 16; rivers, at 14; and trees, at 12. These values are shown in the graph our program generated, below:



These results are surprising in that we were previously unaware of the variety of topics Van Gogh painted about. As previously mentioned, Van Gogh is often best remembered for his work, *The Starry Night,* but these results show that the majority of his works centered around people, which is shown by the most frequently used tags being 'Women', 'Men', and 'Portraits', with 'Female Nudes' close behind. One potential explanation for this could be that Van Gogh needed money to support himself and people were willing to pay more for paintings of themselves or their loved ones.

These results could also be interesting in that if we assume Van Gogh to be an “average” Post-Impressionist (as in, representative of the group), then we could assume that these results (or at least a similar pattern) would also be seen in the works of other Post-Impressionist painters. For example, we might expect them to also paint mainly about people, and following that, landscapes. To test this hypothesis, we could conduct a potential future study to compare these results to the results of other Post-Impressionist painters.

Impact and Limitations

Our project has the potential to benefit art archivists and preservationists who could replicate our analyses to create representations of the patterns in the paintings of other artists and predict classifying information for new paintings, such as the style. Thus, our results and analyses are significant because of their potential to make the archival process more efficient. Our results should also be representative of any Van Gogh painting, so archivists could use our conclusions to describe Van Gogh’s work and make predictions about newly discovered works. However, our representations for different colors in Van Gogh’s paintings may be flawed in that we used one color name even when it corresponded to multiple hex codes. Our color data itself may also have bias because they were picked using Adobe Color, another algorithm which only selects 5 colors and may do so arbitrarily.

Given these considerations, our analyses might be unable to represent the specific hues used by monochromatic painters. Furthermore, our analyses are not applicable to artists of other mediums since their pieces may have completely different features from oil paintings, and we cannot assume that these alternative features could help accurately predict classifying information about a piece. Lastly, our conclusions about Van Gogh’s paintings should not be extrapolated to other oil painters because Van Gogh was a white male European artist from the last century and therefore not representative of all oil painters. By training and comparing models with data from a greater diversity of artists, we could potentially overcome this limitation and create more generalizable representations and models.

Challenge Goals

The first challenge goal we met is **machine learning**. We wanted to create and train a machine learning model to predict the style of each of Van Gogh’s paintings based on other information we have for each painting (such as the title, colors, year, and genre) in order to answer our third research question: “Can we create an accurate model to predict the style of a painting based on data such as the colors it contains and the year it was painted?” We also wanted to go deeper into machine learning to meet the challenge goal and answer the second part of our third research question: “What is the most important feature for determining the style of a painting?” At first, we intended to show the weights for each feature using the ELI5 library, but we learned that finding the weights for each feature failed to actually answer the second part of this research question. Instead, we ended up determining the best value for the hyperparameter max_depth for our model and finding the feature names and importances of its most important features. Since our project involved using machine learning and going more in-depth with it than we have in class, we believe our project meets the machine learning goal.

The second challenge goal we met is learning a **new library**. To do so, we used two libraries that we didn't learn in class - *requests* for querying the Met Museum API and *bokeh* to create our graphs and visualizations. Our original plan was to use ELI5 and plotly, but while writing the code for our project, we discovered we needed to use requests to work with the API and that we wanted to use bokeh for our graphs because it made it easier to color our visualizations with a given hex code, which we considered an important aspect of our visualizations. We also learned that ELI5 didn't return feature importances, so we decided not to use it. Since our project still involved learning and using two Python libraries not covered in class, we believe our project meets the new library goal.

Work Plan Evaluation

For this project, our planned tasks were:

1. **Get set up**: this means reading our df_reduced.csv as a pandas dataframe and importing all the necessary libraries.
    a. *Estimated time required*: less than 1 hour
    b. *Actual time required*: about 3 hours - we ended up deciding to develop locally, which took 1.5 hours to set up, and then also ended up joining two dataframes, which took a while to figure out, as well as importing more and different libraries than we expected, which also took some time to figure out and then find.
2. **Create the colors over time and style over time graphs**: this means following the steps defined in the methodology section for Research Question 1.
    a. *Estimated time required*: 4 hours
    b. *Actual time required*: 10 hours - actual time required was a lot longer than estimated time required because we spent a long time trying to process df_reduced so that each color in each list of colors was transformed into a row, then tried to write 1 method that produced the graphs representing color over time and style over time before writing 2 separate methods; also tried to display each color over time graph with a color matching the color it was representing when multiple hex codes were connected to the same color, which took a while to figure out.
3. **Create the colors per genre graph**: this means following the steps defined in the methodology section for Research Question 2.
    a. *Estimated time required*: 4 hours
    b. *Actual time required*: 10 hours - spent a long time first trying to see if we could make it interactive with plotly dropdowns and finding that didn't work, then seeing if we could display our graph with colors matching the colors they're representing, which meant learning how to connect our original dataframe with one involving hex codes, and then getting the hex codes to display correctly meant using bokeh instead of plotly, which took a while to learn.

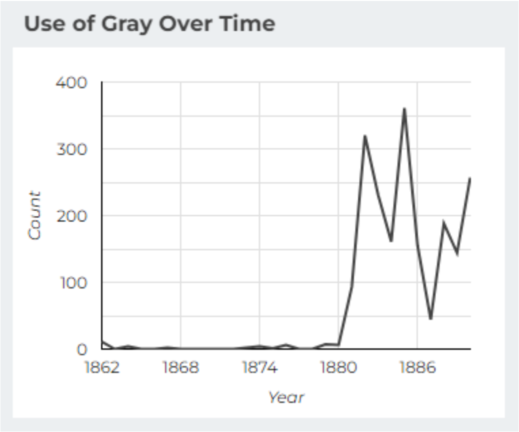
4. **Create, train, test, and interpret machine learning model**: this means following the steps defined in the methodology section for Research Question 3.
    a. *Estimated time required*: 6 hours
    b. *Actual time required*: 8 hours - underestimated actual time required because we did not expect to have trouble applying show_weights from the ELI5 library to the fitted model to show the weights for each feature; we spoke to our TA mentor who said that we could find the feature importances instead and create a bar graph with features on the x-axis and height representing importance.
5. **Create paintings per topic graph:** this means following the steps defined in the methodology section for Research Question 4 (including querying and processing the messy data from the Met Museum API).
    a. *Estimated time required:* 5 hours
    b. *Actual time required*: 9 hours - we decided to move from plotly to bokeh, which meant redoing the graph for this question, and my original function for querying the API stopped working for some reason at some point (we suspect a change on the API's side) which meant I had to rewrite it as well.
6. **Write report and slides**: this means editing sections we wrote here, copying graphs from our code into our report and slides, writing text explaining and analyzing those graphs, answering our research questions, and reviewing the document as a whole.
    a. *Estimated time required*: 6 hours
    b. *Actual time required*: 9 hours - some parts took longer than expected when writing, and review took a while due to the length of the document.

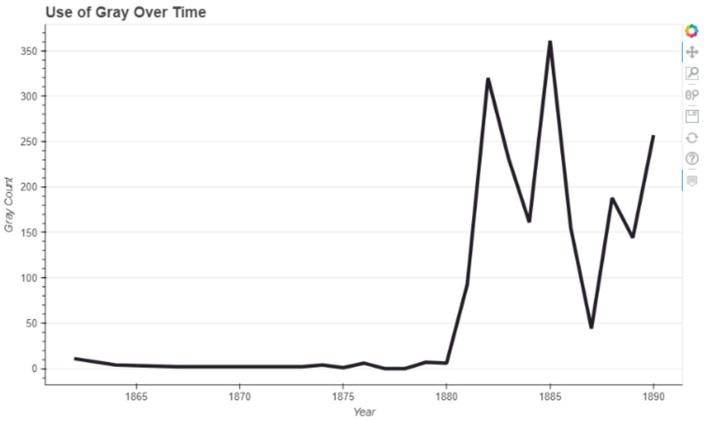
Testing

Research Question 1:

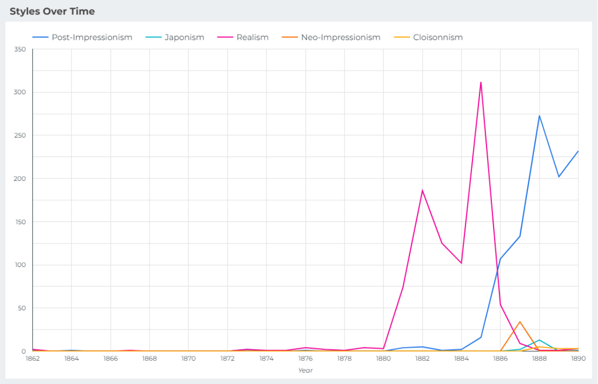
We wrote two graphing functions in response to this question, so our tests focused on making sure our graphs accurately represented our calculation for the number of times each color or style was used per year.

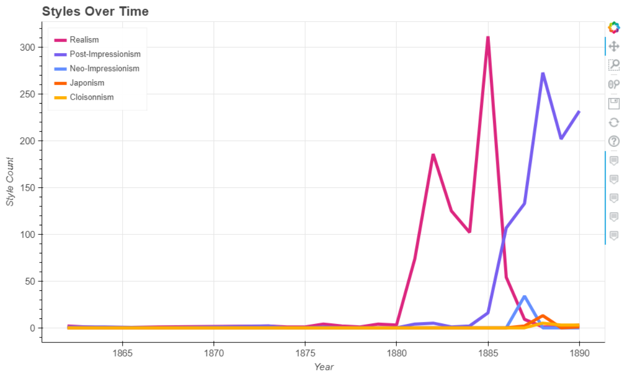
For each color, we saved the filtered dataset as a .csv file and uploaded it to Google Data Studio. We then created a time series with the Year column as the dimension and Count column as the metric. The full set of test graphs can be viewed here, but only the graphs representing the use of gray over time are discussed below.





To test our second function, I also saved the filtered dataset for each style as a .csv file and uploaded it to Google Data Studio. However, instead of creating a single graph for each style, I created a stacked time series by joining the files together. Similar to our first test, the Year column in the joined dataset was the dimension and the Count column for each style was the metric so that each style was represented by a different line on the graph. The interactive test graph can be viewed here.

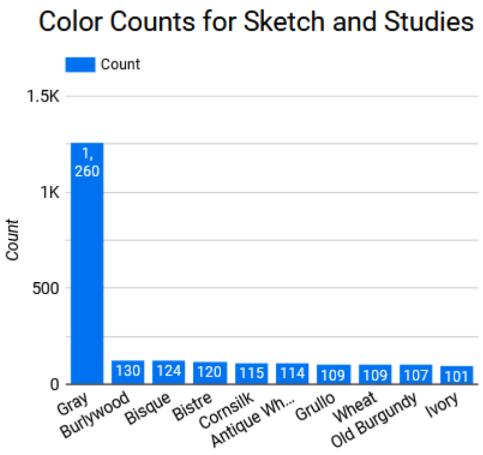


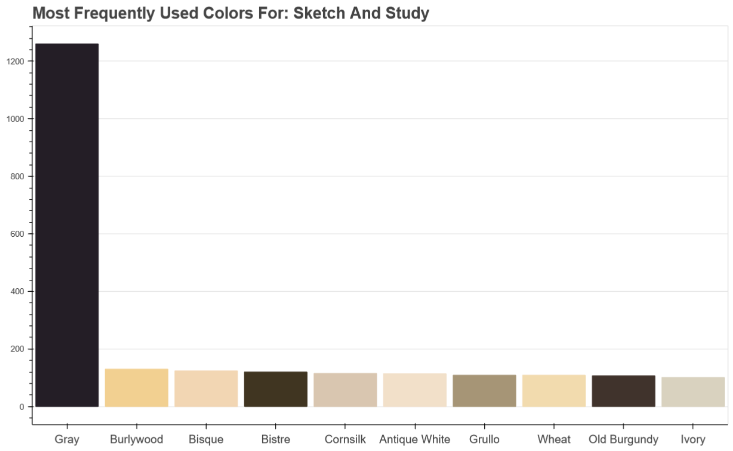


By comparing the graphs from Google Data Studio (on the left) and the graphs we generated (on the right), we can conclude that similar visualizations are produced by both methods, as the same counts correspond with each year when hovered over. On the x and y-axis of each set of visualizations, a similar range of years and counts are also represented, though the actual values written vary depending on the program. In other words, we can conclude our first and second function for answering this question can be trusted and correctly represent the count of each color and style because the same filtered datasets in a different program generate a nearly identical graph.

Research Question 2:

This question focused mainly on graphing, so our tests did as well - namely, we wanted to test that our graphs were accurate in terms of shape and counts. To do so, for each genre, we saved the filtered data for that genre, before counting. Then, we gave that data to Google Data Studio to graph. The full set of test graphs can be viewed here, but for the sake of brevity, we will only discuss the test graphs for one of the genres we discussed in our results section - sketch and study (shown below).





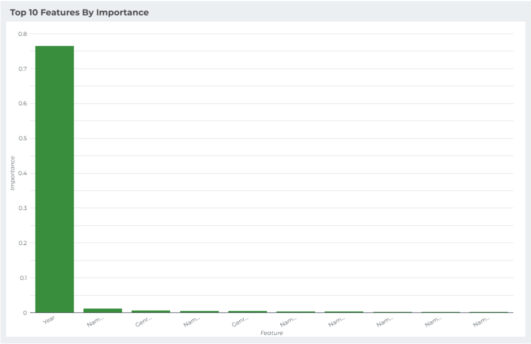
When comparing our graph for Sketch and Study from Google Data Studio (left) to the graph we generated (right), we can see that the overall shape is the same, with gray being the most frequent by far. We can also see that the list of top 10 colors (gray, burlywood, bisque, bistre, cornsilk, antique white, grullo, wheat, old burgundy, and ivory) is the same across both graphs, and in the same order. Finally, we can see that the counts are the same - gray is 1,260, burlywood is 130, etc. From this comparison, we can see that our function for this question is working correctly, as when we pass the data to a different program, it generates essentially the same graph. Therefore, we can trust our function to graph our top ten colors per genre correctly.

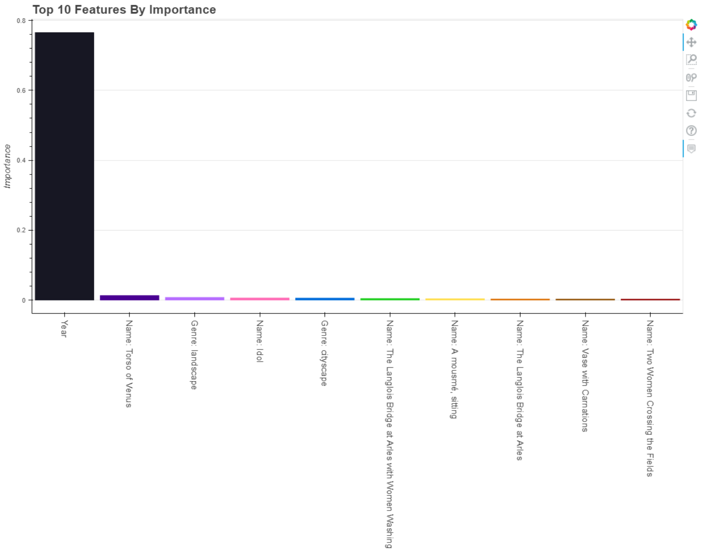
Research Question 3:

The first part of this function focused on creating an accurate machine learning model to predict the style of the painting, so we split our data into 10% testing data, 10% validation data, and 80% training data. We then trained our model on the training data, determined the best value for the hyperparameter max depth and calculated the prediction accuracy of the model with max depth set to this value on the test data. In other words, we tested the performance of our model at various max depths by comparing the prediction accuracy of the model on the validation data, to ensure the prediction accuracy of our model is not reduced by overfitting.

We responded to the second part of this question by first finding the feature names and importances of the features used to train the model along with the max depth that maximizes the prediction accuracy on validation data. We then wrote a function that takes in the feature names and importances as a list and sorts the list by feature importance before generating a bar graph with feature names on the x-axis and importance represented by height. Since we wrote a graphing function to answer this part of the question, we focused our test on making sure the graph accurately represented features and their corresponding importances.

Similar to question 2, we saved the filtered data for the top ten most important features as a .csv file and uploaded it to Google Data Studio. We then created a bar graph with the Feature column as the dimension and the Importance column as the metric. The interactive test graph can be viewed here.



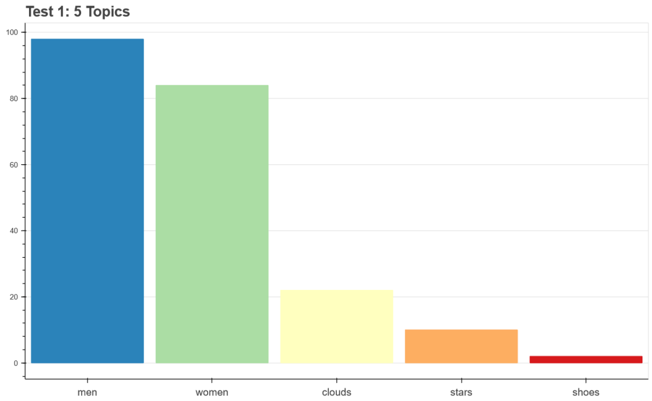


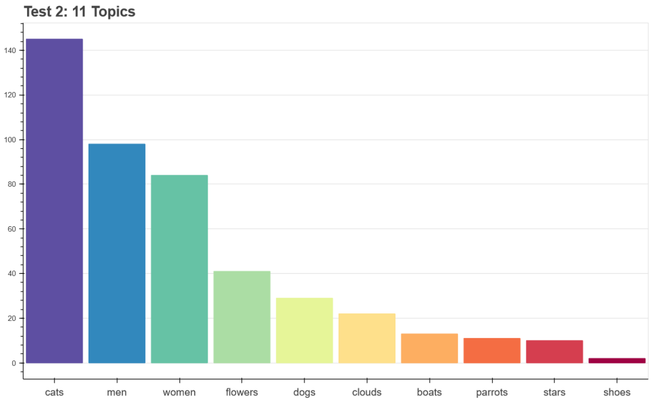
By comparing the graph from Google Data Studio (on the left) and the graph we generated (on the right), we can first identify that the visualizations are similar in shape and that features appear in the same order from most to least important on the x-axis. The horizontal grid lines on the first visualization and the tick marks on the second visualization also allow us to estimate the height of each bar and conclude that the heights of corresponding bars are approximately the same. The bar colors of the first graph are not based on the Color column of the filtered dataset because Google Data Studio is unable to graph bars in different colors and bar color is not relevant to answering the second part of this question. Based on these graphs, we can conclude our graphing function can be trusted to correctly represent the importance of each feature because the same filtered dataset in a different program generates a similar graph.

Research Question 4:

Like questions 1 and 2, this question mainly focused on graphing. However, the data comes in JSON format from an API, so our tests for it are different. Namely, we focused on looking at our graphing function because we can’t control the results the API gives us. We also looked at it in our code, rather than with an alternative software, because the results from querying the API (and thus the parameters for our graphing function) were in dictionary format. Since querying the API takes a while, we also created these tests to look at the formatting of our graph without waiting to query the API each time. Overall, it made more sense to do the tests in Visual Studio Code.

We created and included a testing function in main.py, test_most_frequent_topics(), that passes two smaller dictionaries of test data into our graphing function for question 4. The first dictionary tested that our function worked correctly if there were less than ten topics given to it by passing five topics with counts shuffled in random order, and the second dictionary tested that it correctly selected and ordered the top ten topics when it was passed more than ten topics. Both test dictionaries also let us test the formatting and hover tools for our graph. The graphs from these tests are below (with the first test dictionary on the left, and the second on the right), and the interactive versions can be found in the graphs folder, under q4_tests.





From our tests, we can see that our graphing function can take any number of topics, select the top 10 by counts, order them, graph them, and color them correctly. Thus, we can trust that no matter what results we get from our API, our graphs accurately represent the data we received.

Collaboration

In the course of learning new libraries, we consulted many online resources, including the original documentation for packages like bokeh, stackoverflow, and a few other websites. Specific instances are listed below:

For question 1:

- Figuring out how to transform each item in a list into a row - pandas documentation
- Figuring out how to add ‘Count’ as a column to a dataframe - stackoverflow
- Figuring out how to select the first hex code corresponding to a color - geeksforgeeks
- Learning how to drop a dataframe index column - datagy
- Learning how to make a time series with bokeh - bokeh documentation
- Learning how to make a stacked time series with bokeh - programming historian
- Learning how to create tooltips for a time series - bokeh documentation
- Learning how to change the font size of axis labels - stackoverflow
- Learning how to move the legend - bokeh documentation

For question 2:

- Same stackoverflow reference for the ‘Count’ column as question 1
- Same hex code reference as question 1

- Learning how to make a bar graph with bokeh - bokeh documentation
- Learning how to graph multiple bokeh figures with a for loop - stackoverflow
- Learning how to capitalize the first letter of every word in a string - thispointer.com
- Figuring out why my tooltips weren’t displaying correctly, and how to fix it - github

For question 3:

- Figuring out how to find the best max depth level and feature importances - github
- Same geeksforgreeks reference for selecting the first max depth corresponding to the highest validation accuracy as question 1
- Same bokeh graphing references as question 2

For question 4:

- Same bokeh graphing references as question 2
- Learning about palettes - bokeh documentation
- Learning how to query APIs - dataquest.io