MoMA Mia!

Info-247 Final Project Report – Spring 2017

Anna Cho, Malavika Srinivasan, Selenne Berthely Garcia

Project Goals

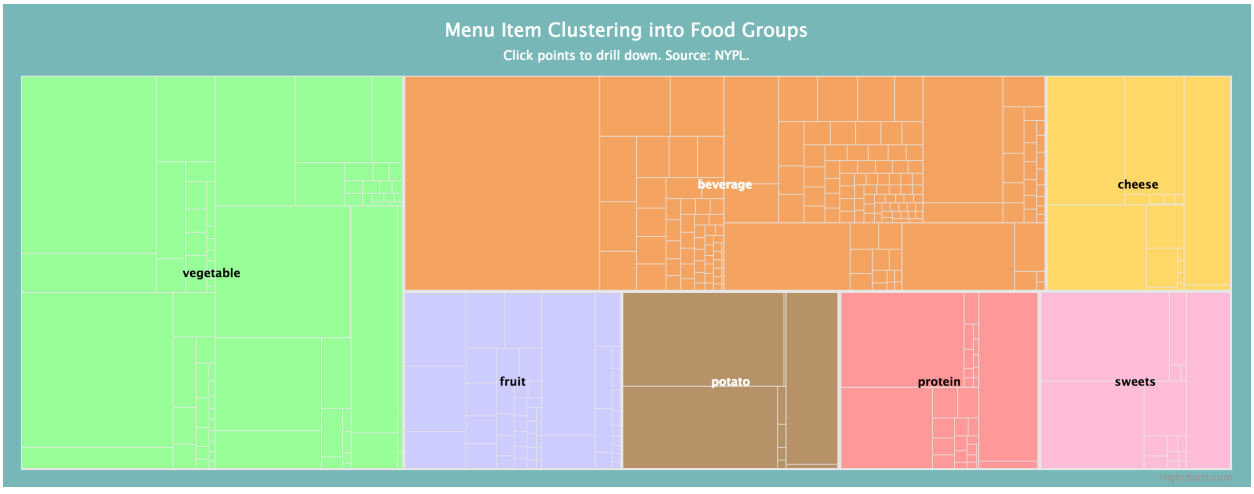
In our exploratory data analysis of the MoMA collection, one of our main goals was to help people understand what "Modern Art" is, namely what it looks like and the different formats it is exhibited in. To accomplish this goal, we implemented features that allow users to interact with the artwork in the collection. The sunburst chart allows users to explore the different categories of artwork the MoMA divides its artwork into and the different mediums that make up those departments. We also have an interactive stacked bar chart of colors that can be found in the MoMA collection. Users can see actual MoMA artwork using these colors. Our word cloud shows the prevalence of the art movements by region represented in the MoMA collection. Through these charts, users can obtain a more solid understanding of what artwork is part of the Museum of Modern Art.

Our second goal in visualizing the MoMA collection was to help people better understand where the MoMA collection originated from and how it arrived at the MoMA. We incorporated a timeline of MoMA acquisitions broken down by the method of acquisition (e.g. through gift or through purchase). We created a map that reveals how many pieces of artwork come from which countries. We also had a chord map displaying the collaborations between artists of different nationalities that produced the artwork. These pieces of our user interface help users develop a clear story of how artwork was created and how the MoMA acquired its collection.

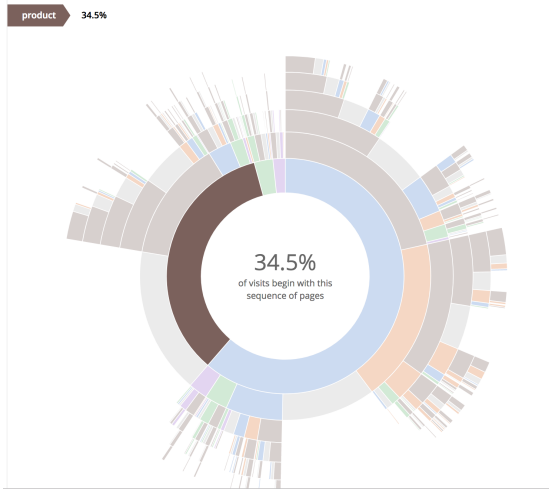
Related Work

Anna

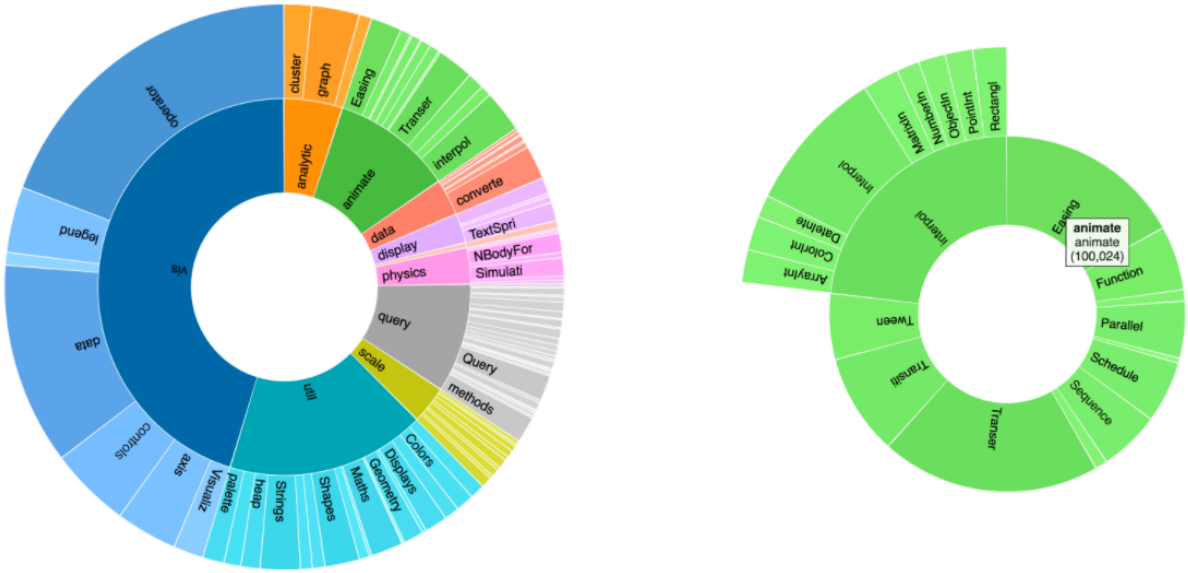
The first piece of related work that helped me design the zoomable sunburst was from "Menu Journeys" a final project for this class from 2015. In this project, the students used a zoomable treemap that clustered menu items by food groups. This was a useful and intuitive way to understand the breakdown of the food found in menus around the world.



After realizing that the MoMA dataset was much too complex and granular to effectively use a zoomable treemap like the one above, I found a sunburst example that displayed similar hierarchical information. However, our team found it difficult and tedious to explore the small arcs in the outer-most layers of the sunburst chart.



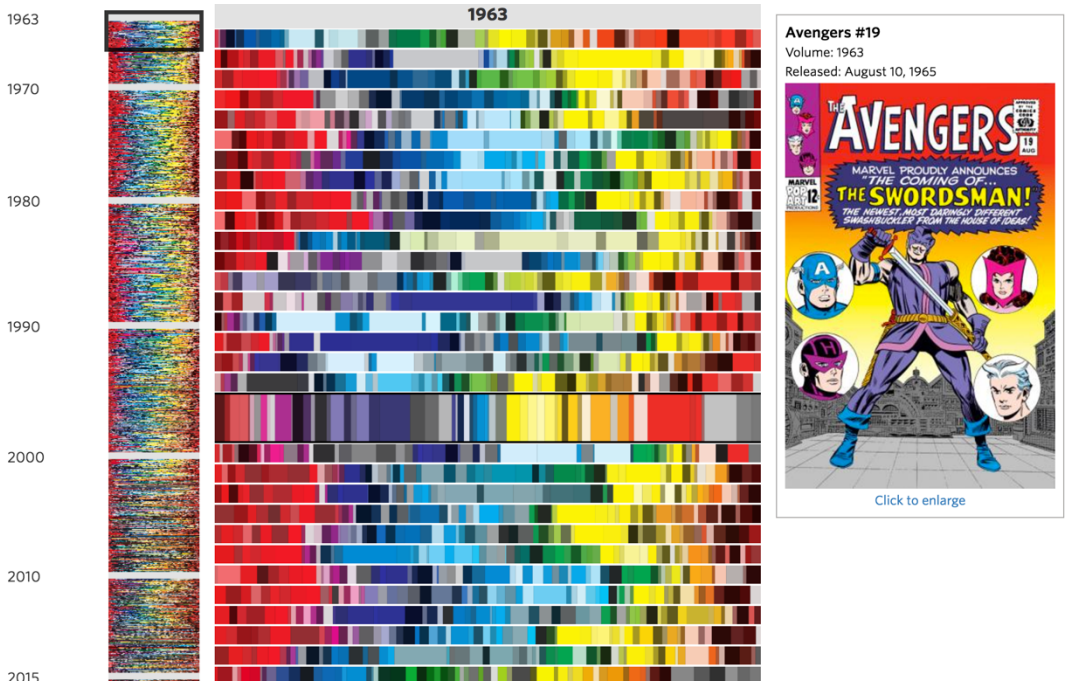
The third related work that I used and modeled my chart after was a zoomable sunburst diagram. This chart maintains a hierarchical view of the data but allows users to zoom into each layer of the diagram. This alleviated the difficulty of exploring the small outer arcs of the chart.



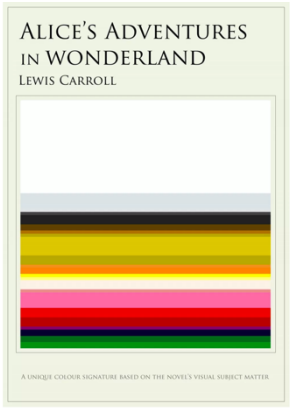
Malavika

To seek inspiration for color based visualizations, we looked at Flowing Data's visualizations on color. Of these, the most inspiring that we drew from for our visualization here are listed below -

- 50 Years of 'Avengers' Comic Book Covers Through Color - The Wall Street Journal created an amazing visualization that explore the covers of the comic book Avengers by color. What we really liked about this visualization was how the colors were neatly sorted by hue and provided immense visual appeal and also the interactivity of moving through the years and being able to see the exact comic book cover where the colors came from.

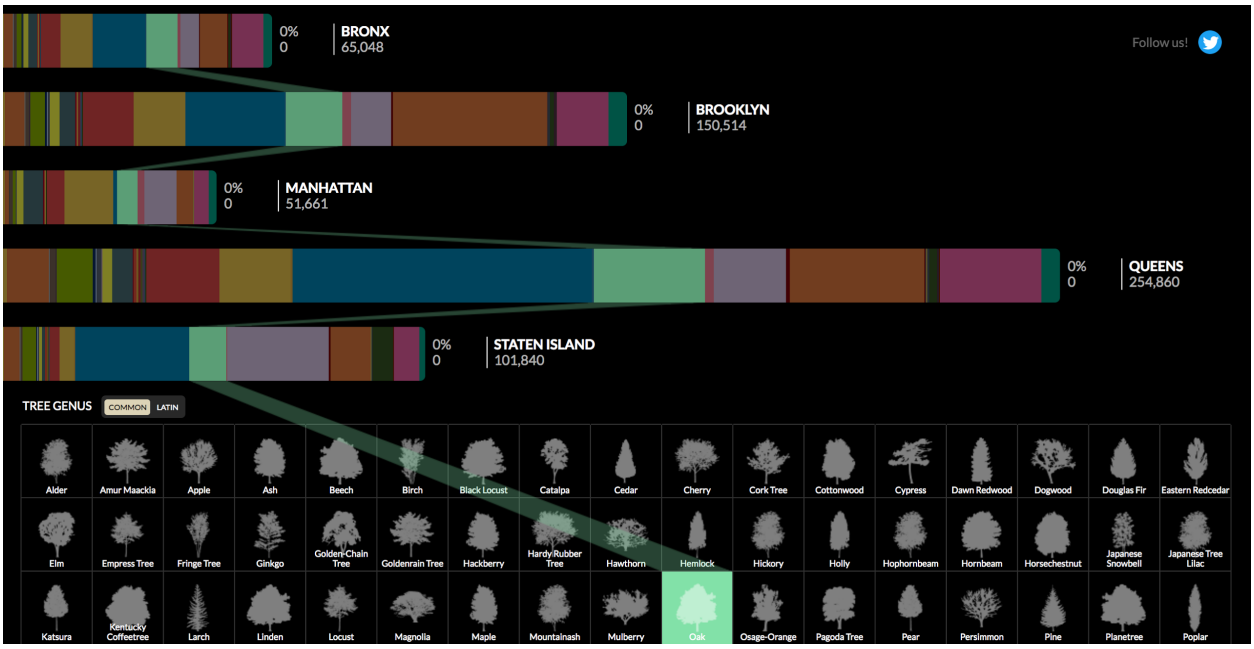


- Color signatures for classic novels - Jaz Parkinson's simple static visualizations had a unique approach to color by visualizing *mentions* of colors in classic novels. From this, we gained another perspective into the arrangement of colors along a vertical strip, the heights determined by the volume of mentions.



- Color Me Curious by Dina Bseiso and Natalie Goldberg - Dina and Natalie's Infoviz project, Spring 2016. We were greatly inspired by Dina and Natelie's use of sketches and color quantization to explain fashion. We believed that art too is very visceral like Fashion is.

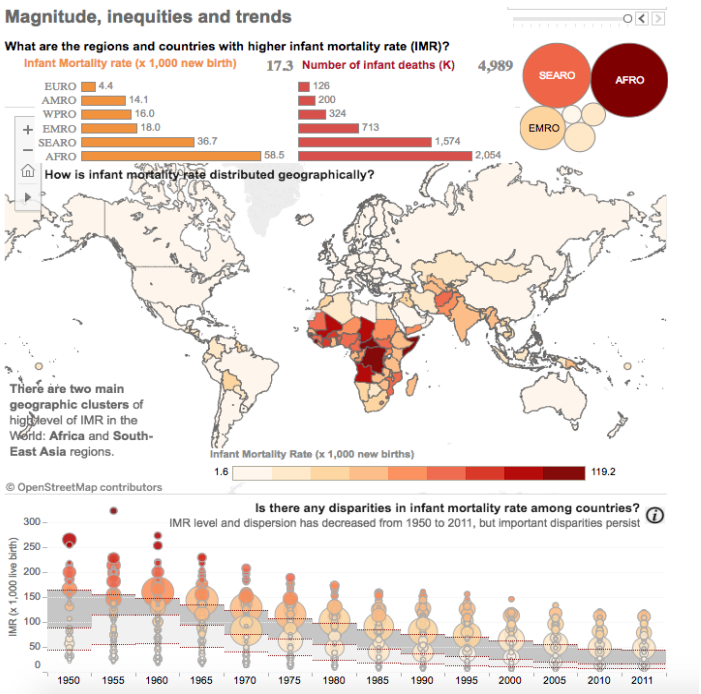
While the aforementioned visualizations gave us great ideas on how to visualize color data, our biggest inspiration of design and function of the final visualization came from Cloudred's interactive visualization of NYC's trees. We decided to split the depiction of colors by department to better understand the distribution of colors in the artworks of each department and also link the colors across departments for comparison and to display examples. The linked stacked bars in this example provided a very elegant way to do this, in our opinion. One drawback of the NYC trees visualization was that colors used did not have any visual significance except as a distinguishing factor by type. In our visualization, the colors of the bars represented the actual colors being visualized which made them very effective.



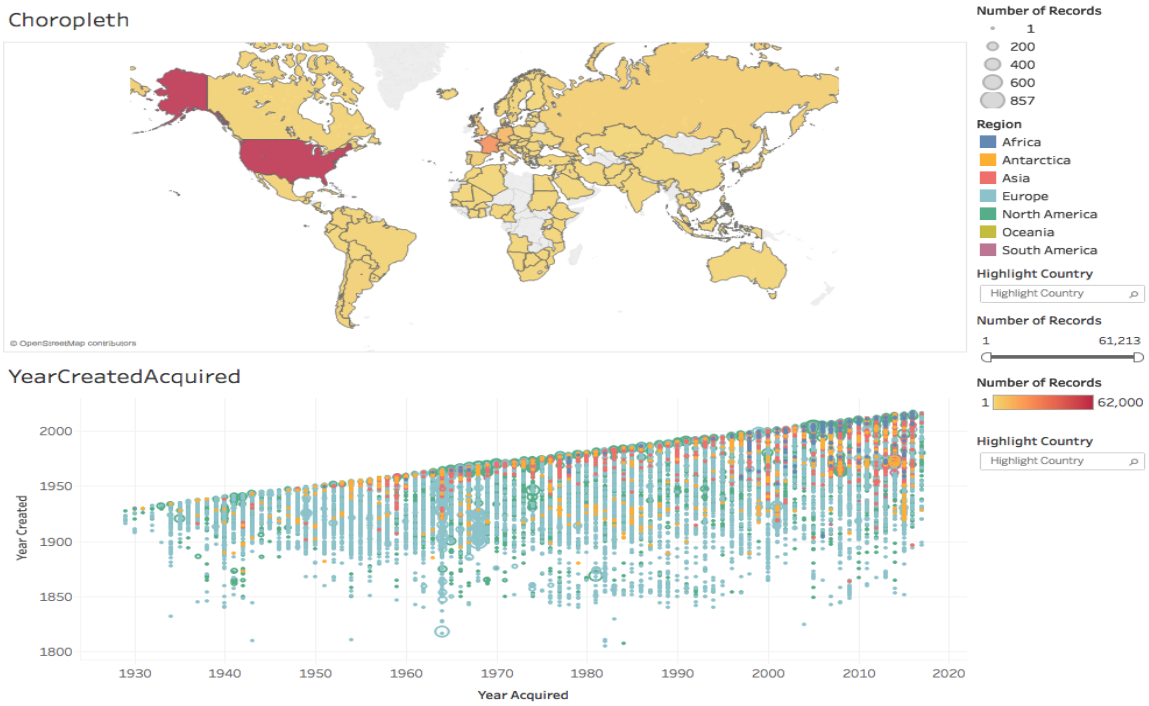
A lot of back-end python work went into creating the final visualization which involved processing approx. 60,000 images, performing k-means clustering on each of them to extract the prominent colors from them and then finding the inter-image clusters by department. For these, I used the following link for references along the way.
Color quantization using k- means - http://lmcaraig.com/color-quantization-using-k-means/

Selenne

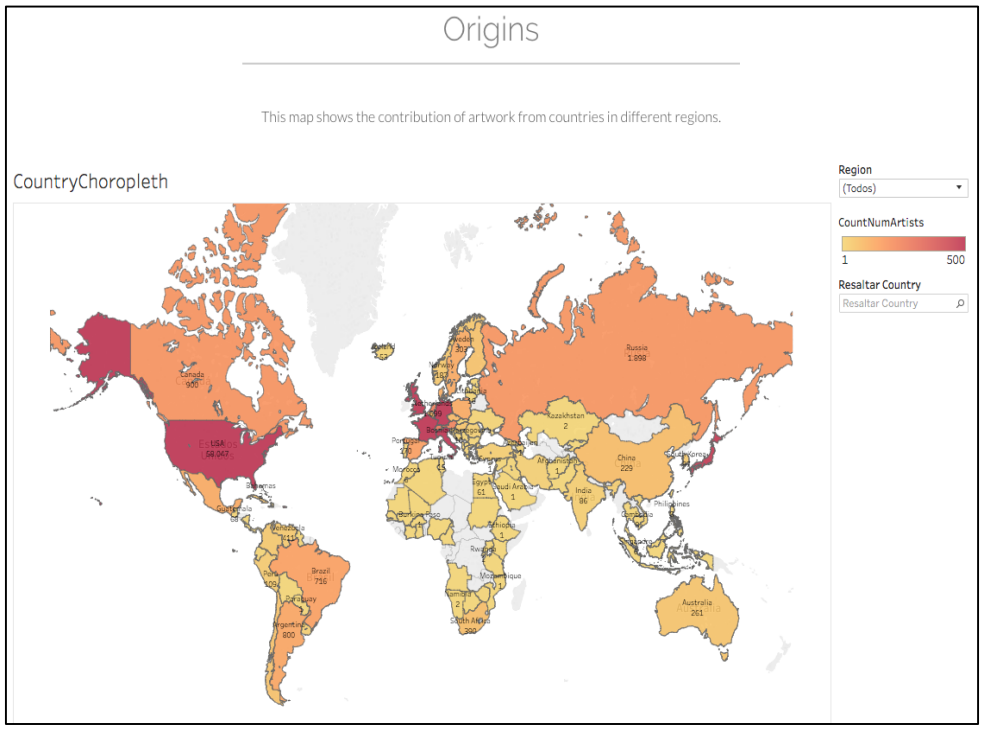
As previously mentioned, the dataset is amazingly big. 145,000 rows of artwork, and almost 65,000 different artists from different parts of the world. We found very interesting visualizing the number of artworks per country, then we looked for inspiration and the first piece of related work that we found was a Tableau dashboard representing the global overview of the magnitude, disparities and trend of infant mortality in the world from 1950 to 2011.



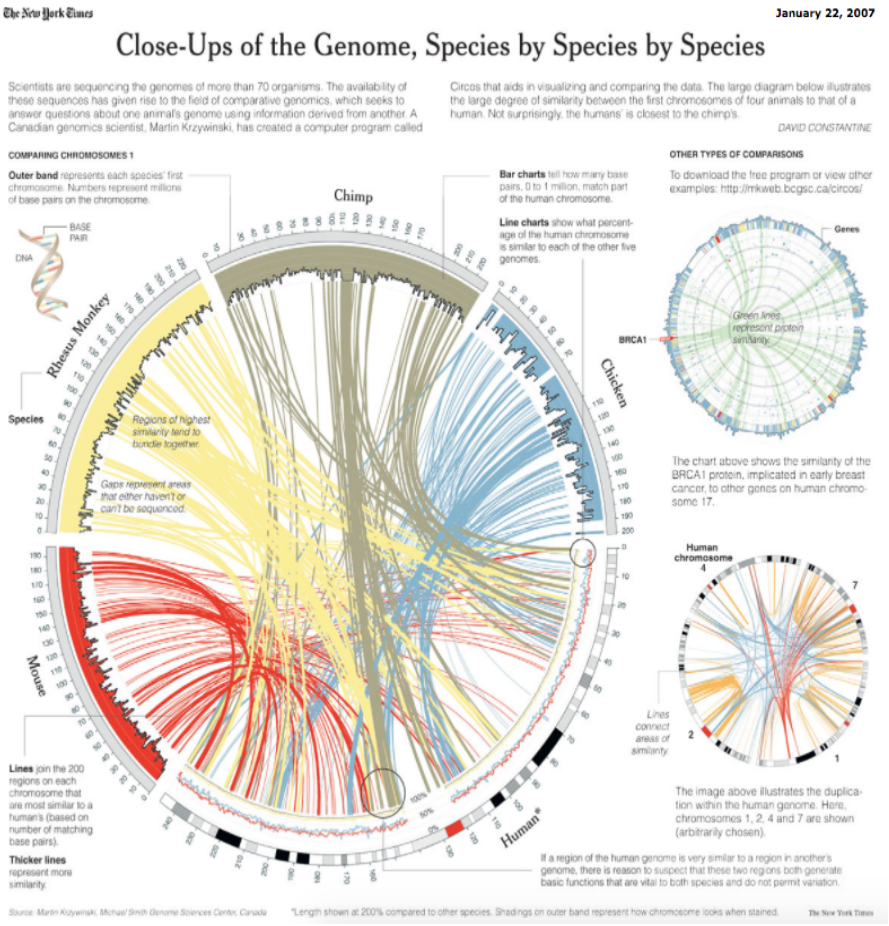
We felt that we could build a dashboard where we could show the countries per region and below how the collection has been acquired through the years, but the information is so granular that we felt that it was going to be difficult to convey a simple message since the filtering was not helping to show an specific trend.



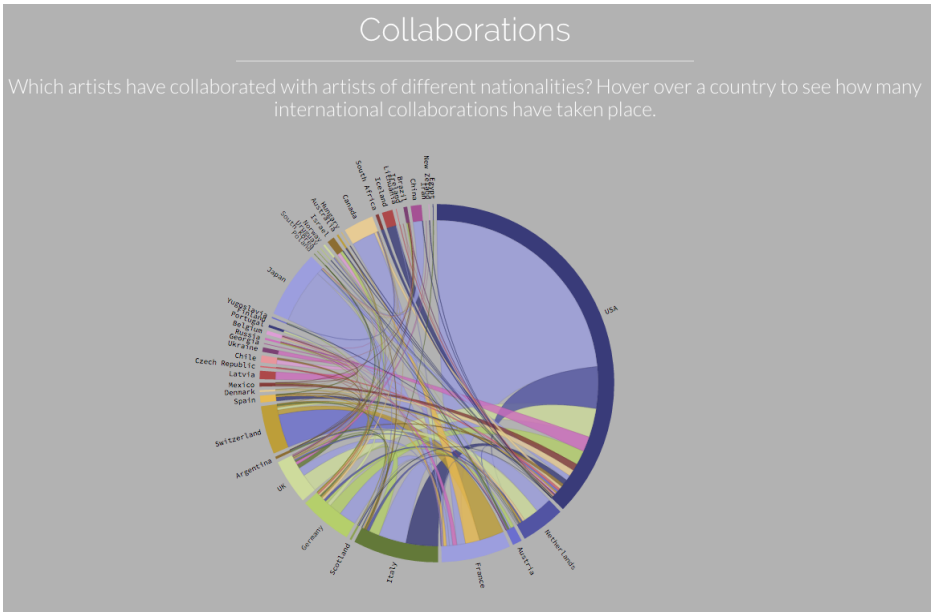
We finally felt that showing a map that could easily let the user see by color how each country has contributed to the collection was the best option.



After we cleaned the dataset, we could finally get the country from where each artist comes from. We also found in the dataset, that each Artwork not was always related to only one artist, for a lot of them, there were a lot of collaborations between artist for the same artwork. Thus we decided that we wanted to build and the second piece of related work we used as inspiration is a Chord Graph developed by the New York Times to present genome sequences between different animals and how are they related .



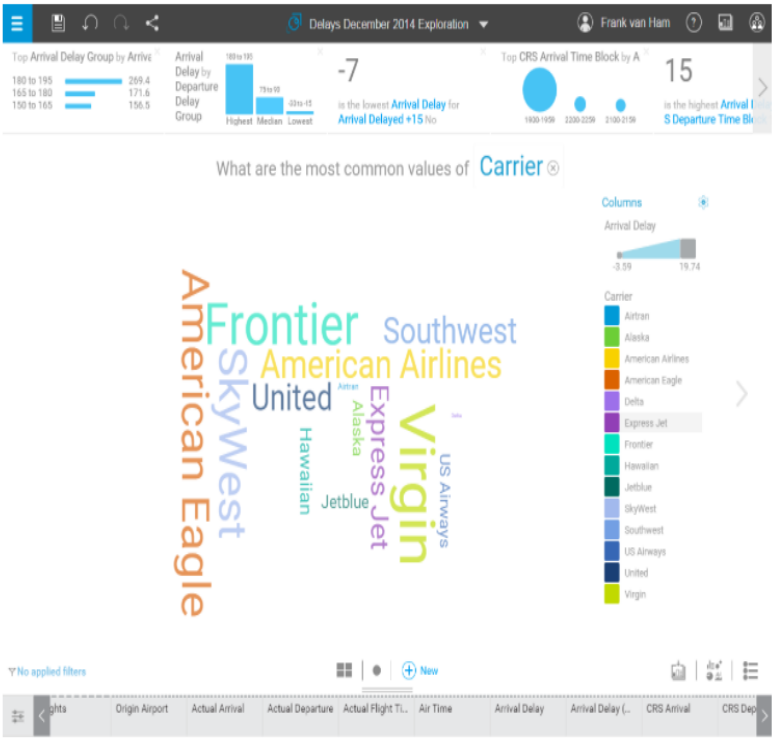
Then, we got determined to clean our dataset and separate the artist peer artwork and relate them in pairs, so we would be able to show how different countries have collaborated between each other. Our final result was this and we enjoyed the learning experience while doing it.



The third related work we used as inspiration was a nice way to show the Movements to which each artist belong and develop their work. We didn't have that information in the dataset, so we have to build our own dataset from a website we found that relate artists with the movement they belonged. (http://www.artcyclopedia.com/)

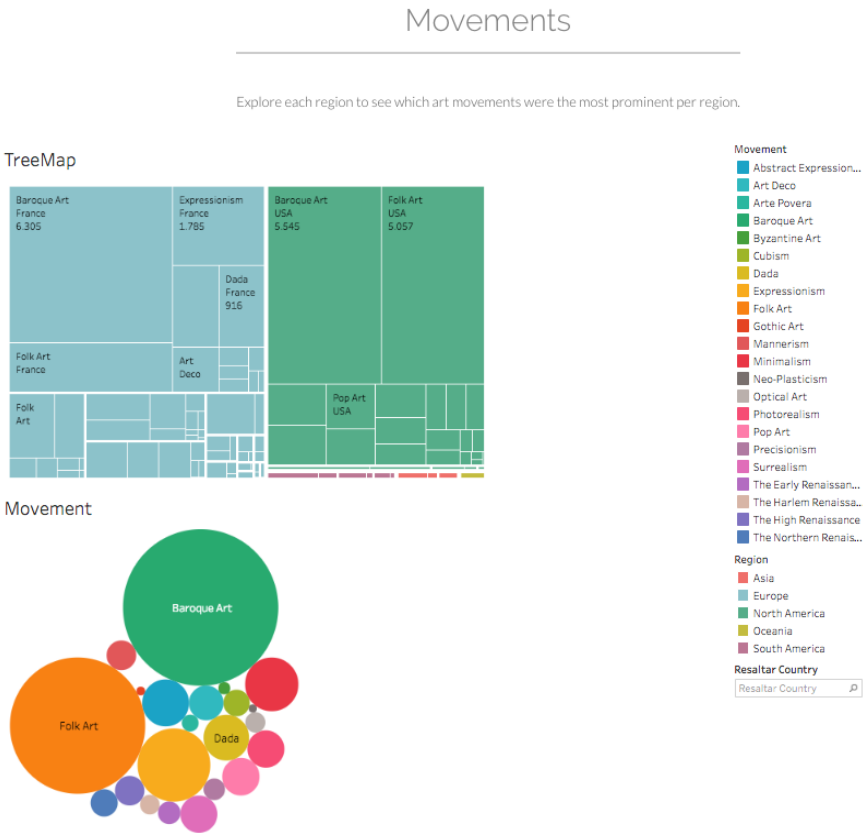
Our first design was a Wordcloud, which we liked a lot since we could filter by region and the tooltip was showing the number of artworks and the origin, but, it seem to be difficult to read and convey the idea of magnitude.

The idea was taken originally from an visualization of airlines with the highest average arrival delay.



Our first result was this Tableau Wordcloud.

After a lot of discussion between something we aesthetically liked, and something that was useful at all, we decided that the best way to present information regarding movements and their location, was not a map, but a bubble chart that interacted with a tree map, so the user will practically select from different art movements and regions and get the idea of distribution and how relevant have this movements be to the collection.

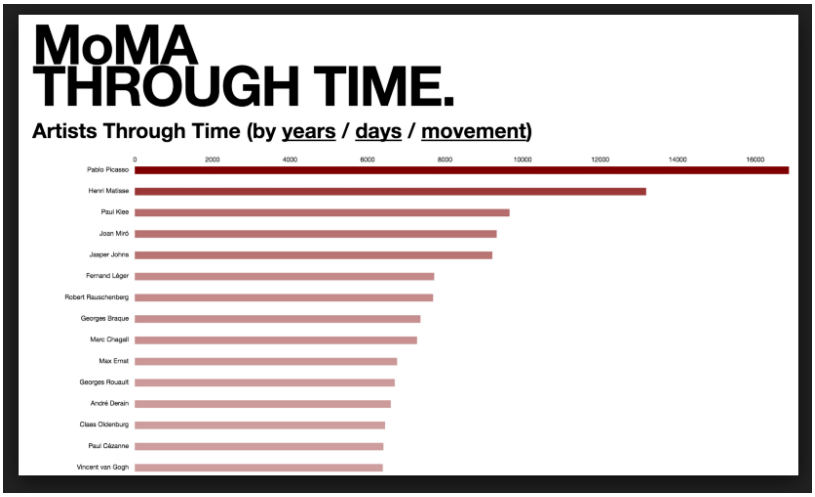


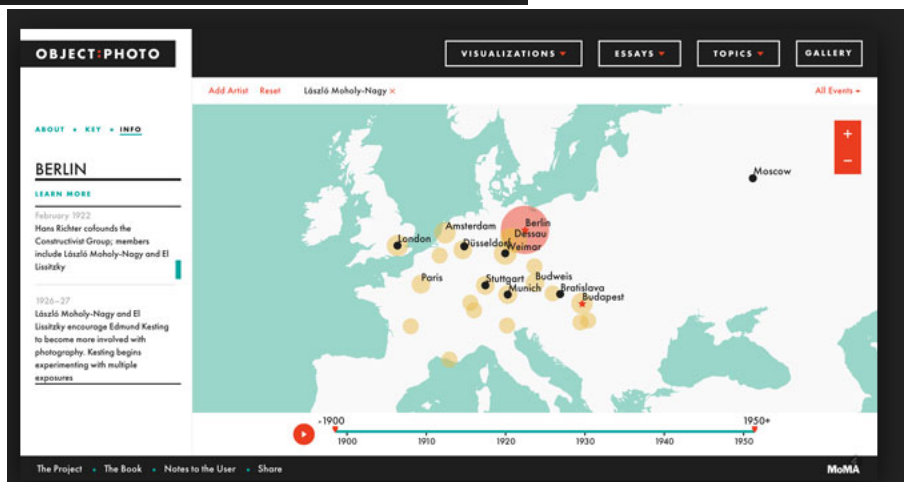
Steps to accomplish the goal

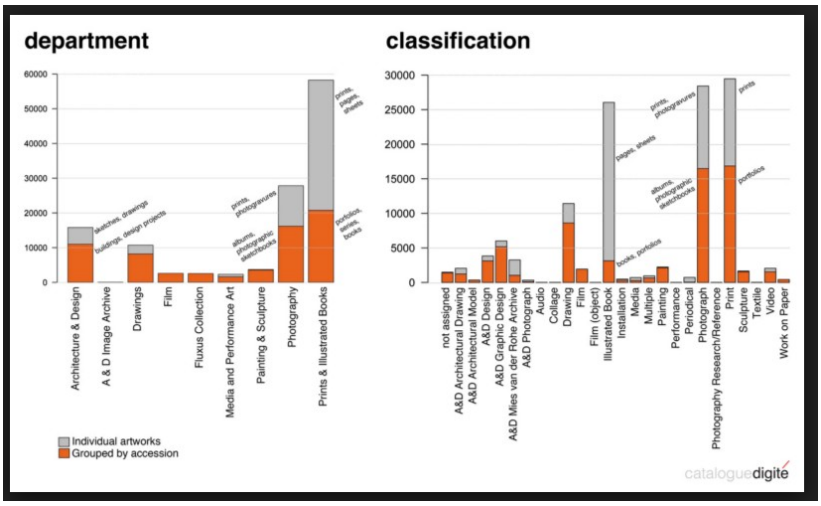
The process we followed to achieve the final result was:

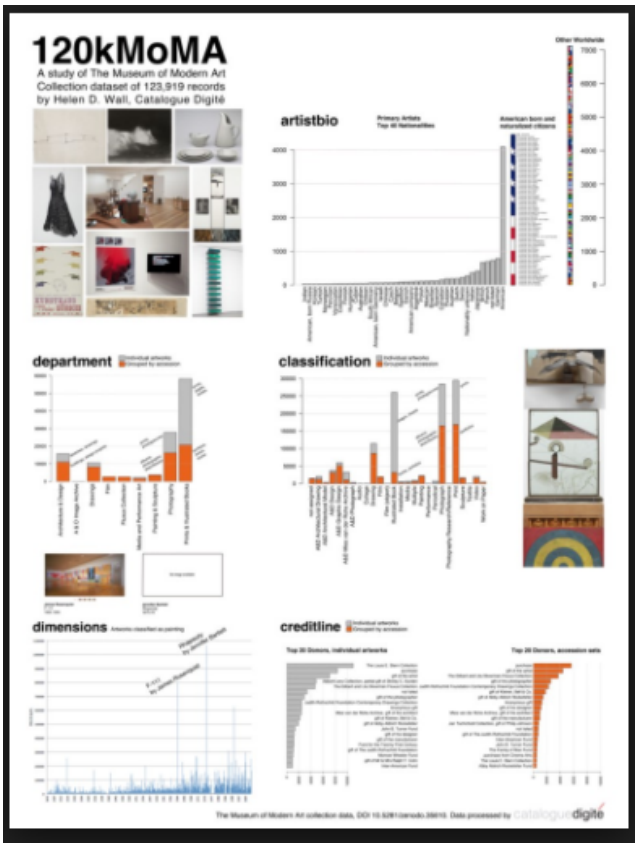
1) Get the datasets: For this process we searched over Kaggle (https://www.kaggle.com/momanyc/museum-collection) and we found a couple of datasets, but they we not updated and some columns were not separated. We kept looking for another datasets and we found a more complete one on github. (https://github.com/MuseumofModernArt/collection)

2) Inspiration: Find some inspiration and how the MoMA collection has been presented in some newspapers and websites over internet. But We always have something in mind, connecting the sections, so they could tell a story.
Some initial visualizations we took in consideration to get our minds started are shown below.

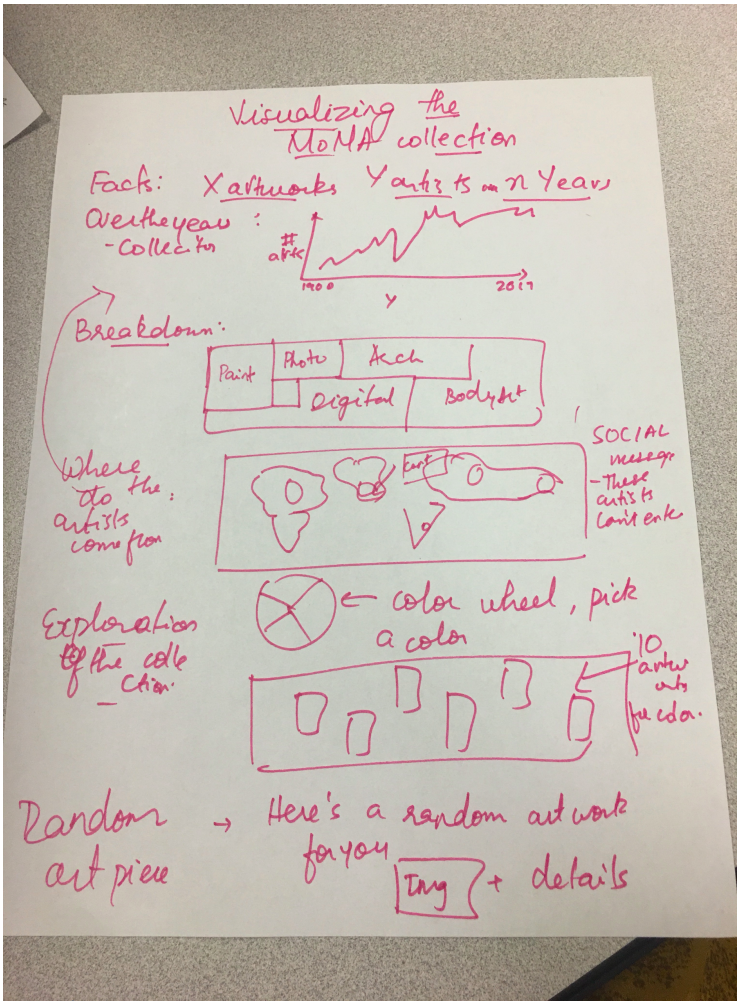








3) Website design: We did a brainstorm to come up with the first structure for our website, what we wanted to develop and what we wanted to convey:



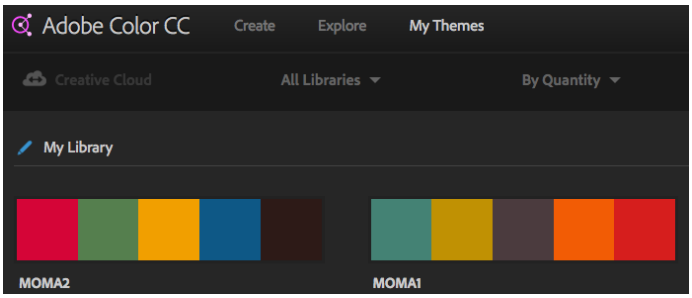
4) Dataset cleaning: This tasks was one of the most crucial and I believe it didn't stop through the whole development of the project. Every time we felt we have a solid dataset to present the information, we came up with another need that put us back to the need of cleaning or create a less granular dataset.

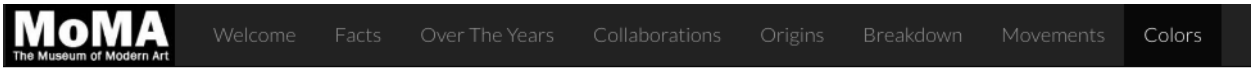
Even though the dataset was clean and very complete, some columns needed a lot of work to be useful for our purposes. For example the Department or Medium used in the Sunburst chart, are fields that we created based on filtering the information and creating less detailed fields that could let us show the diversity, but just until some point of granularity.

Another tool we have to use for the dataset cleaning was pandas. We have the need to separate the artist that collaborated in each of the artworks, where there were more than one artists for one artwork. Here is a snippet of the code.

```python
# Split 'ConstituentID' column
# http://stackoverflow.com/questions/12680754/split-pandas-dataframe-string-entry-to-separate-rows

def tidy_split(df, column, sep='|', keep=False):
    """
    Split the values of a column and expand so the new DataFrame has one split
    value per row. Filters rows where the column is missing.

    Params
    ------
    df : pandas.DataFrame
        dataframe with the column to split and expand
    column : str
        the column to split and expand
    sep : str
        the string used to split the column's values
    keep : bool
        whether to retain the presplit value as it's own row

    Returns
    -------
    pandas.DataFrame
        Returns a dataframe with the same columns as `df`.
    """
    indexes = list()
    new_values = list()
    df = df.dropna(subset=[column])
    for i, presplit in enumerate(df[column].astype(str)):
        values = presplit.split(sep)
        if keep and len(values) > 1:
            indexes.append(i)
            new_values.append(presplit)
        for value in values:
            indexes.append(i)
            new_values.append(value)
    new_df = df.iloc[indexes, :].copy()
    new_df[column] = new_values
    return new_df
```

5) Visualizations: Finally, we built our D3 diagrams, Tableau charts, and Malavika focused on developing the wonderful tool for visualizing the artwork selecting from the most prominent colors in the collection.
6) Website building: We got some initial experiments, then deciding that we were going to be using a bootstrap based website. We decided which pallete we were going to be using and finally decided the structure and navigation for the website.



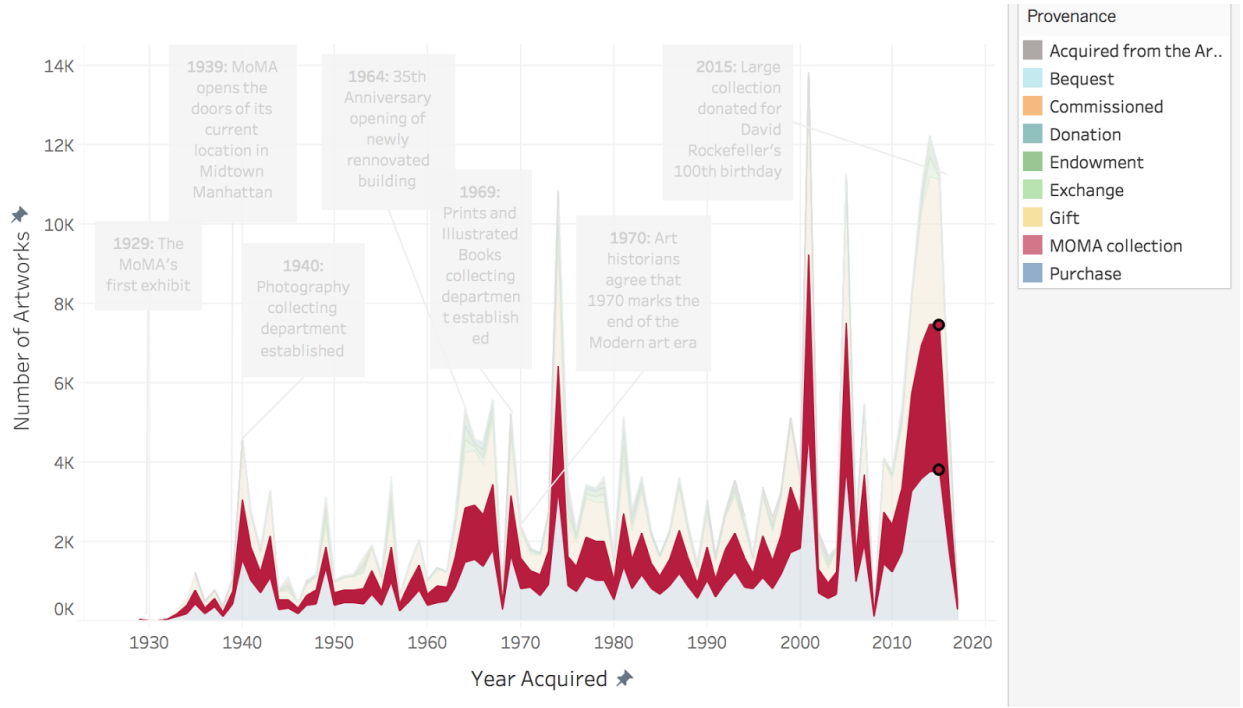


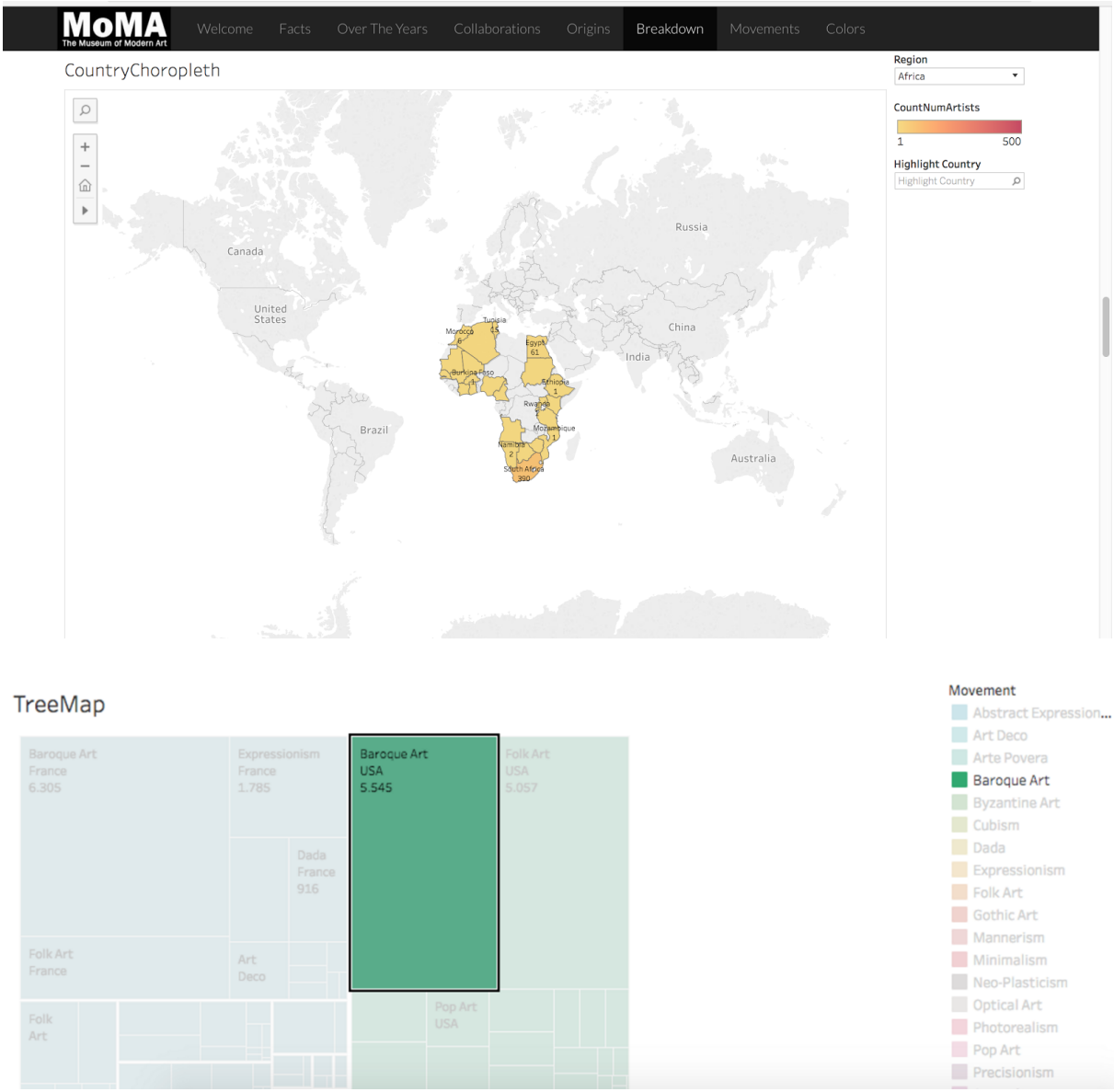
Data

All of our data came from the MoMA itself. The MoMA publishes its database of artwork and artists represented within its collection on its public Github account. We used two of the datasets within their repository: 'Artists.csv' and 'Artwork.csv'. The datasets were both very detailed and contained several features. From 'Artwork.csv', we used information about what categories the MoMA classifies each piece of artwork into (e.g. mediums and art movements), how and when the MoMA acquired the artwork, and the thumbnail image URLs for each image. From 'Artists.csv', we used information about the artists' nationalities. We also merged the two files together to determine which artists collaborated with artists of different nationalities. We also obtained external data for two purposes – for annotation of the over-the-years visualization, we gathered qualitative data about some milestones in the history of MoMA to try and explain the spikes in the trend. We also found data about art movements and merged this with the artists data to associate them to the movements for our visualization. There was a great deal of exploratory data analysis made possible by the richness of these two datasets.

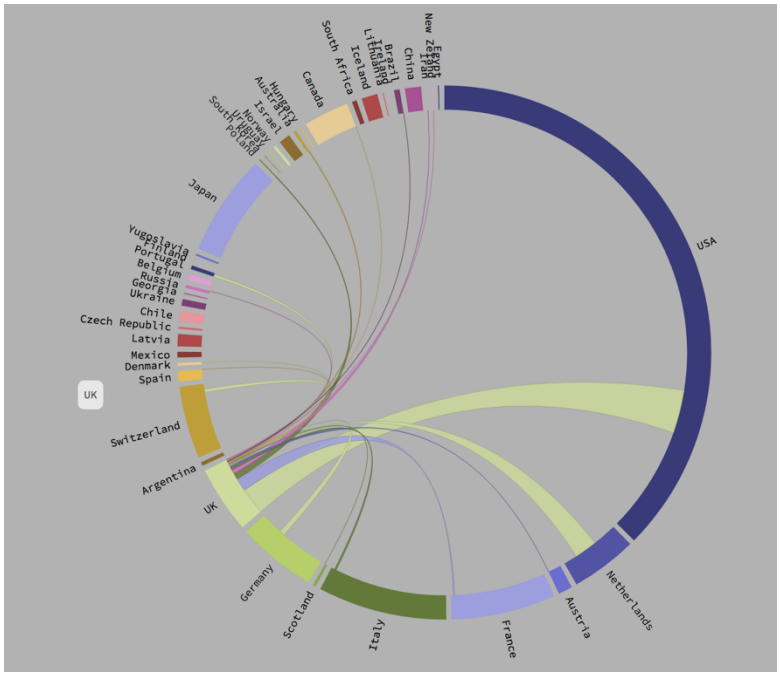
Tools Used

For each visualization we created, we used a variety of different tools. We used Tableau to create the timeline, map visualizations, and treemap. Tableau was a great choice for these visualizations because it has intuitive highlighting and filtering functionalities for our users to breakdown the MoMA collection by various categories (e.g. credit lines and countries):



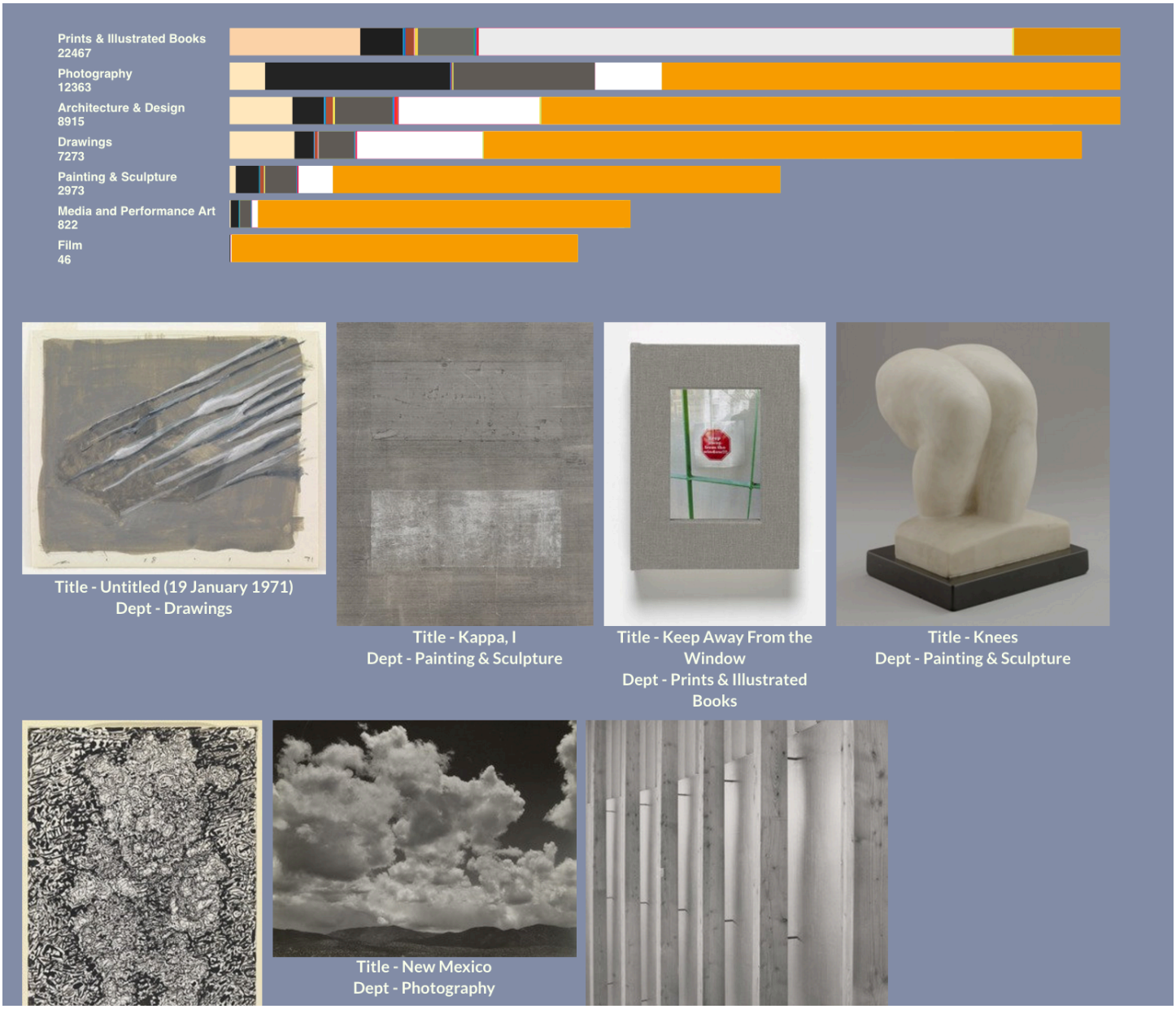


For the collaborations, sunburst chart, and color exploration visualizations, we used D3. The chord chart that was used in the collaborations visualization also has an interactive highlighting feature that updates for each country that the user mouses over. This makes the different connections between artists of different nationalities easy to compare and allows the user to observe the differences in collaboration for each country.



The sunburst diagram has a zooming-in animation feature that makes it clear that the user is getting a deeper look into the MoMA collection each time they click on an artwork category.

The exploration by color is also interactive and allows users to actually see the art that makes up the collection whenever they select a color from the stacked bar chart. It was built using d3 and required a lot of image processing and data preparation in the backend. The appendix has a detailed explanation of the procedure to create the data and the visualization.



The filtering, highlighting, and interactivity of our visualizations allow users to obtain a better understanding of the MoMA collection's history and characteristics.

To style the website that is hosting our data visualizations, we used a premade Bootstrap template. It provided us with clean fonts and a simple black and white color scheme that allowed our more colorful information visualizations to stand out.

Insights and Feedback – Malavika

Through our visualizations, we gained many interesting insights about the MoMA collection.
From the color section, we found that a lot of artwork had black, whites, greys and beige as the most prominent colors. This surprised us as we expected modern art to be more colorful and vibrant with more pinks, blues, greens and reds.

We let the users interact with our website on their own and asked them to think out loud their thoughts as they interacted with our visualizations. We also recorded any feedback that they had about the visualizations. Through this exercise, we gained many insights into the legibility and usability of our visualizations. Some of the feedback that we got and incorporated in our last iteration are listed below.

- Over the years – Many users were curious about the reasons behind the spikes that occurred in MoMa's acquisitions during certain years. To address this, we gathered qualitative data about the MoMa's acquisitions and annotated the graph with events that could've caused the spike
- Origins – In our first version of this visualization, we used bubbles to depict the volume of artworks that came from every country. However, we received feedback that this was not too intuitive, so made a chloropleth to better display the map.
- Movements – An interesting question we encountered for this visualization was to whether the use of words was the best way to express/visualize art and art movements. For instance, many people who are not art aficionados might not be familiar with art movements like "Baroque Art" or "Expressionism" or be familiar with typical examples of such paintings. We tried a different visualization to bridge this gap
- Overall, some users also did not clearly understand how to interact with certain visualizations or understand the meaning of certain visualizations from the descriptions that we had provided. We handled this by providing more detailed explanations and directions for interacting with the visualizations.

LINK TO DEMO:
http://people.ischool.berkeley.edu/~sberthely/MomaMia_Final/

LINK TO CODE:
https://github.com/sberthely/MomaMia.git

REFERENCES TO EXAMPLE CODE:
https://startbootstrap.com/template-categories/all/
https://bl.ocks.org/mbostock/4063269
https://bl.ocks.org/kerryrodden/477c1bfb081b783f80ad
http://jsfiddle.net/ZGVK3/
http://mbostock.github.io/d3/talk/20111116/bundle.html
http://bl.ocks.org/vgrocha/1580af34e56ee6224d33
https://bl.ocks.org/kerryrodden/7090426
http://www.adeveloperdiary.com/d3-js/create-stacked-bar-chart-using-d3-js/

http://lmcaraig.com/color-quantization-using-k-means/

REFERENCES TO INSPIRATION:

http://publichealthintelligence.org/content/global-overview-magnitude-disparities-and-trend-infant-mortality-world-1950-2011
http://www.nytimes.com/imagepages/2007/01/22/science/20070123_SCI_ILLO.html
https://www.ibm.com/communities/analytics/watson-analytics-blog/visualizing-network-data-to-illustrate-airline-delays/
http://people.ischool.berkeley.edu/~carlos/menujourneys/
https://www.jasondavies.com/coffee-wheel/
http://graphics.wsj.com/avengers/
https://flowingdata.com/2013/05/06/color-signatures-for-classic-novels/
http://groups.ischool.berkeley.edu/ColorMeCurious/
https://www.cloudred.com/labprojects/nyctrees/

DATASET LINKS:

https://github.com/MuseumofModernArt/collection

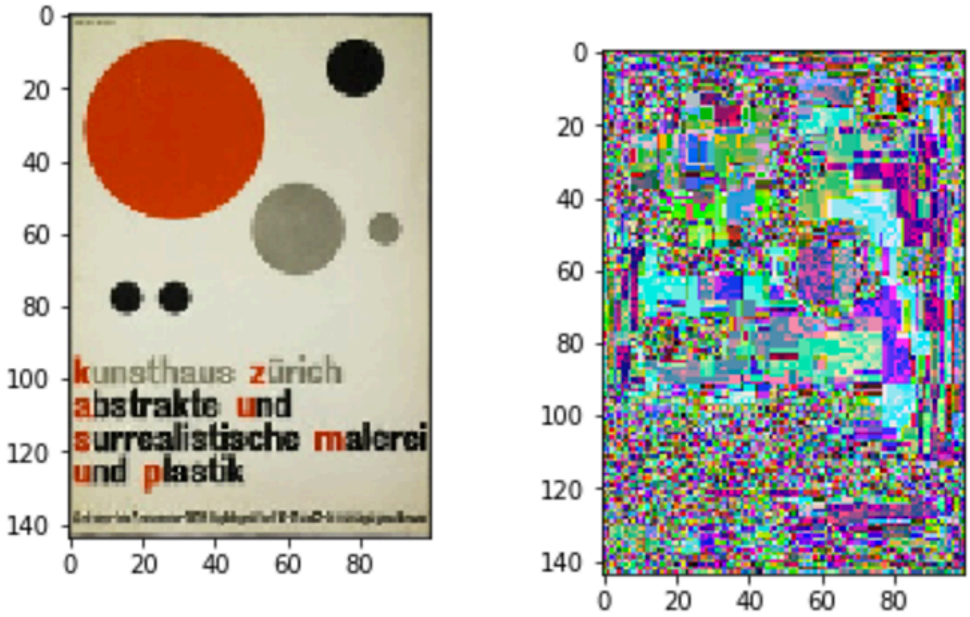
Division of Work

Visualization	Team Member Name(s)	Details
Data cleaning and Preparation	Selenne and Anna	Spent a LOT of time cleaning the data – handling missing values, in many cases, manually recoding the data to get it in the right format, created a single dataset after combining the artists, artworks and the movements data
Over The Years	Anna, Selenne	Experimented with many designs before settling on this one in Tableau. Manually gathered extensive amounts of interesting qualitative data for annotating the graph with historical events to explain the trends
Collaborations	Selenne	Backend – Created the data in the right format using pandas and Excel. Customized a library available for chord diagrams in d3 to suit our needs in d3
Origins	Selenne	Experimented with many prototypes of this viz in Tableau – iterativey incorporating feedback before settling in on the final viz. Did data scaling to make sure that some data points are not misrepresented.
Breakdown	Anna	Prepared the data in the required format for the d3 viz using javascript. Experimented with many different d3 libraries for creating the sunburst before deciding on the final one. Customized the library to fit our needs.
Movements	Selenne	Got data on Movements from external sources, integrated them with the data. Experimented with many different ways to present Movement information in Tableau.
Exploration by color	Malavika	Backend Python– Image Processing of ~60k images, K-Means Clustering, Data Preparation for the visualization in the right format Visualization – D3 + Javascript
Website	Anna and Selenne	Worked on building the website, integrating the visualizations into it as well as hosting it.

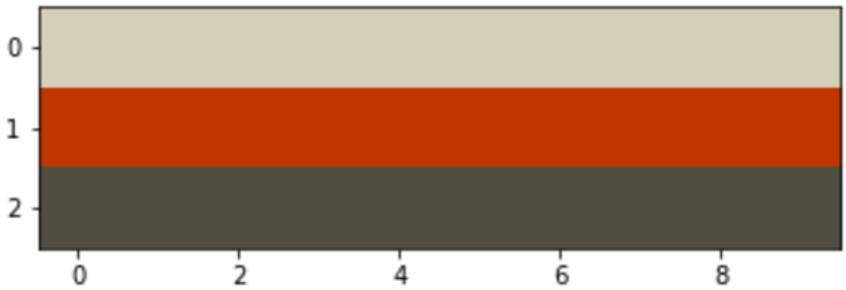
Appendix

Color visualization

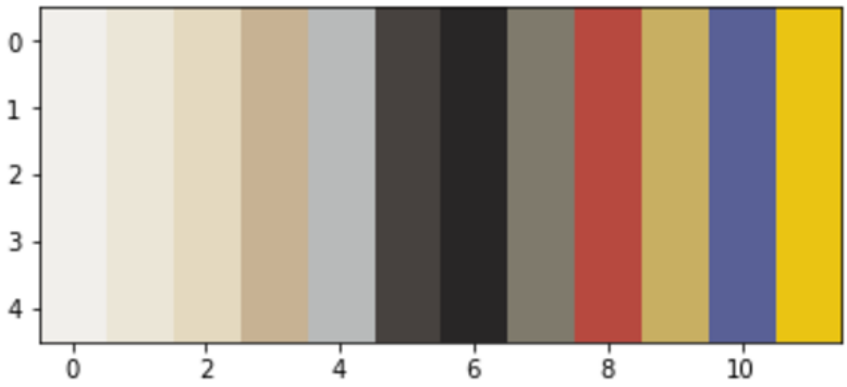
The first step in creating this visualization involved understanding the colors present in the images of the artworks. In the dataset provided by MoMA, this information was not present, but the URLs to the thumbnail images of the artwork for present for many pieces of art (nearly 40% of the data).
We set out to extract the prominent colors in the image using k-means clustering. For this, we read the image's pixel values in rgb format and then converted into the CIE-LAB format. This format was used as it encodes color in a way that is similar to how color is encoded according to the human perception of color and hence distances – similarities and differences between colors – are best explained by this color model.
Image in RGB and LAB format –



After converting the image into the LAB format, we performed k-means clustering on each image with 5 clusters to identify the top 5 prominent colors in the images.



This same process was repeated for nearly 60,000 images that had valid image URLS and their top 5 most prominent colors were stored. The next goal was to find the top colors at a department level. For this, we took the most prominent color of all images in that department and performed K-means clustering on these images.



While this was the story that the data was providing, we wanted to see if we can surface the blues and the greens etc. by mapping the prominent colors to a base set of colors. We experimented with using Kelly's colors of maximum contrast, but it was not as good as we would've expected. We tried with Crayola's 24 color set as the base and it gave us decent results as depicted in the visualization. It looks like there are a disproportionately large amounts of orange, but it could be that a lot of neutral/beige background are mapped to the orange as it is the closest to those in the color spectrum.

The stacked bar chart in the color explorer visualization was created in d3.