**Dataset #1**

1. **Initial Note**

If you have issues with running the code, please try using the Google Colab environment. Additionally, using this environment may require upgrading plotly. Please see the commented line at the top of the code (**python data cleaning + visualizations.py**) which does this upgrade.

1. **Introduction**

For this portion of task two, we cleaned the data file produced from our API code and made visualizations of certain variables in the dataset. The contents of this folder (the files) are as follows:

**art institute python updated API.py:** this file contains a very slightly modified version of our original API code. The only modification here was retaining the end\_date attribute column, which we deleted previously believing it to not be useful for future analyses. As we found we needed this attribute when we began the analysis process, we updated the API code to reflect this change and produced a new version of the dataset (please see **artworks.csv** and **artworks.xlsx**).

**artworks.csv**: contains the dataset we cleaned and used to generate visualizations.

**artworks.xlsx**: same as above except in .xlsx format.

**documentation.docx**: contains the documentation for the python part of the project (dataset #1).

**python data cleaning + visualizations.py**: contains the code for the cleaning of the dataset and visualizations produced.

1. **The Python Code: Cleaning + Visualizations**

The python file (**python data cleaning + visualizations.py**) is commented extensively to illustrate both the cleaning decisions made as well as the visualizations generated. Please see this file for more details.

1. **The Visualizations: Comments and Analysis** 
   1. **Treemap of # artworks by country** 
      1. This visualization reflects the number of artworks in the dataset belonging to each country. In this graphic, larger squares are allocated to the countries with a comparatively greater number of artworks in the dataset. Thus, the square sizes are proportional to the number of artworks a country has in the dataset. Looking at the graphic, we can see that the countries with the most artworks in the dataset (in order of greatest # of artworks to least) are Spain, the United States, France, and Italy. Please note that not all countries are represented in the treemap as not all countries have artworks in the dataset.
      2. Additionally, the user is able to hover over each country’s square to see the # of artworks the country has in the dataset.

Chart, treemap chart

Description automatically generated

* 1. **Pie chart of % of artworks by classification type**
     1. This visualization reflects the percent of artworks in the dataset that belong to each classification type. As we can see in the pie chart, nearly 17% of artworks in the dataset are etchings, about 13% are lithographs, about 9% are paintings and about 6% are typed as woodcut. These are the most frequent types of artworks in the dataset; the other types of artwork classifications have comparatively less frequency in the dataset.
     2. The user is able to hover over each “pie slice” to see the # of artworks the classification type has in the dataset.

Chart, pie chart

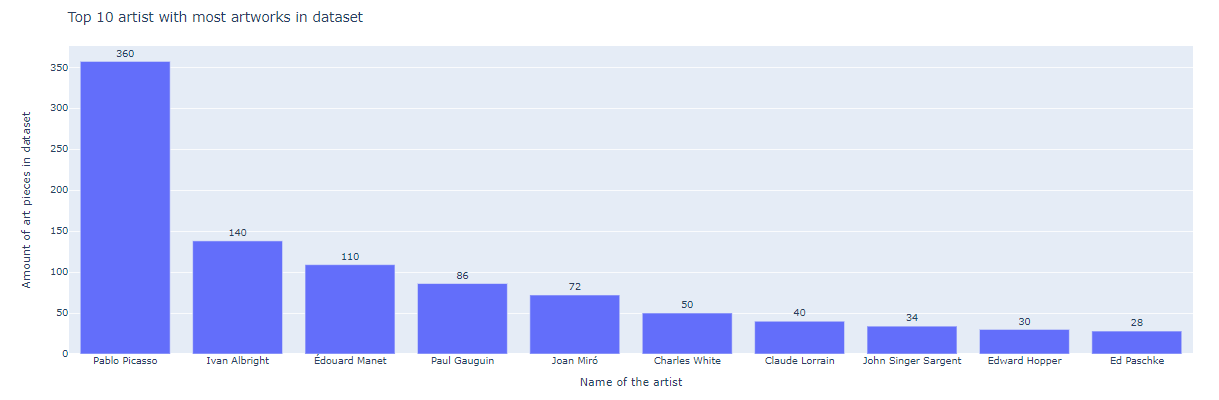
Description automatically generated

* 1. **Treemap of # of artworks by classification type**
     1. This visualization is very similar to the one above but takes the form of a treemap. The treemap form is a bit visually clearer as we are able to see other frequent classification types that disappear a bit (visually) in the pie chart presentation of the information.
     2. As with the previous treemap visual, the user is able to hover over each box to see the # of artworks which belong to a particular classification type in the dataset.

Chart, treemap chart

Description automatically generated

* 1. **Bar chart with top 10 artists with the highest # of artworks in the dataset**
     1. This visualization allows us to see the artists with the greatest number of artworks in the dataset. We only included the first 10 artists because otherwise the visualization would become too chaotic. As expected, we see that Pablo Picasso is the artist in the dataset with the most art pieces to his name.
     2. The user is able to hover over each bar to see the name of the artist and also the # of art pieces that specific artist made.



* 1. **% of artworks in the modern art department by country**
     1. This visualization reflects the distribution of the modern art segment per country. In other words, this pie chart allows us to interpret which countries hold the greatest presence in the dataset’s modern art department. We see that this art movement is dominated by Spain. In addition to that, we see that some other countries such as Belgium, Ireland, Jersey and France have a small contribution to this art movement.
     2. The user is able to hover over each “pie slice” to see the # of modern art pieces per country.

Chart, pie chart

Description automatically generated

* 1. **Artworks plotted by year artwork started being produced (x-axis) vs. time it took to complete the artwork in years (y-axis). In the point colors, we can see the countries with the biggest contribution to the dataset in terms of total # of artworks.**
     1. This visualization plots the amount of time taken to make an art piece by the time around when the artist started to make the art piece. In addition to that, we added the country of origin (point color) to each observation. In this way, we can see in which periods it took longer to make an art piece. Also, we can see which countries dominated certain periods (in the dataset) in terms of # of art pieces being made. We didn’t include all the countries, only the countries where a greater number of art pieces (comparatively) were made.

**Chart, scatter chart

Description automatically generated**