Lauren Courtney

Data Visualization

March, 2020

**Project 3 – Interactive Visualization Using Tableau**

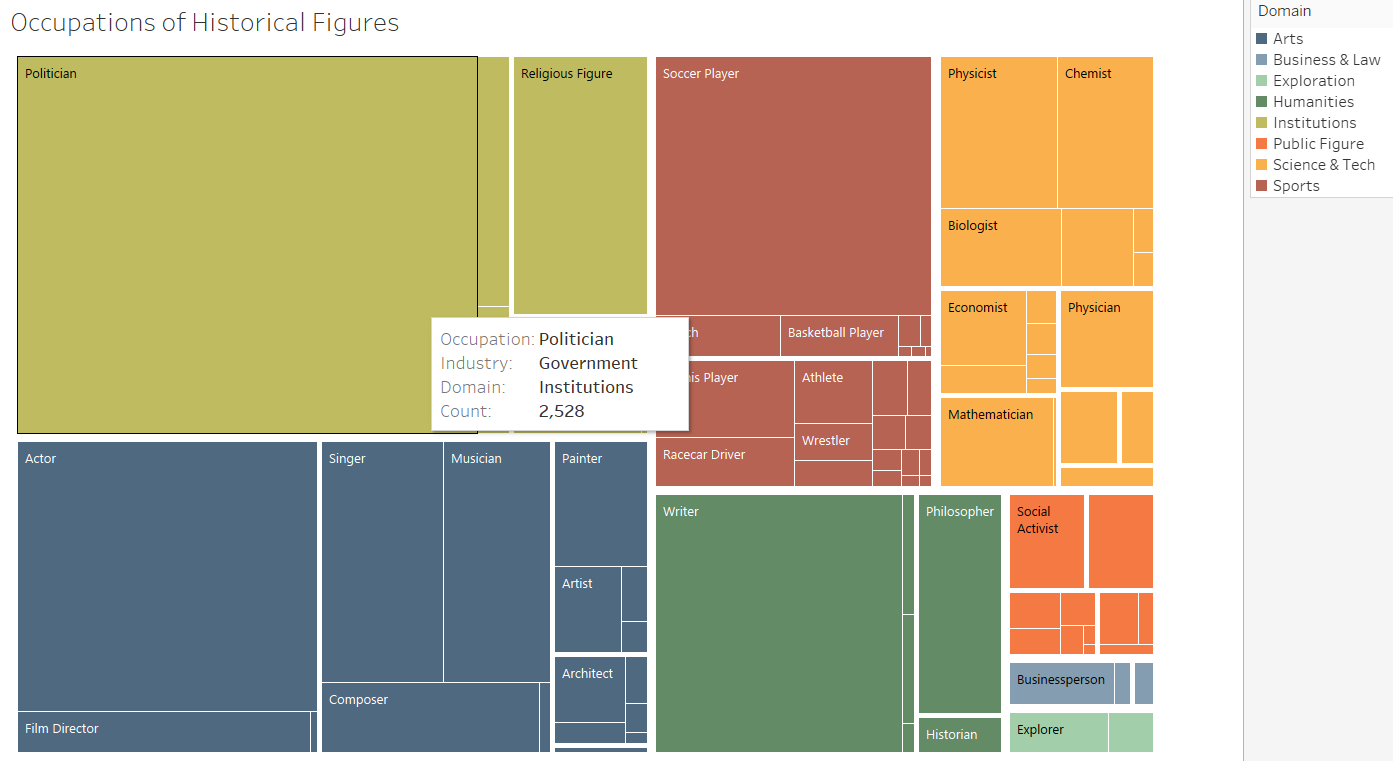
This dashboard covers the Pantheon Project dataset, a collection of historical figures and their attributes. All figures are given a popularity score based on their biographies in Wikipedia [1]. The historical figures in this dataset cover a wide range of industries, and include those born between 3500 BCE and 2005. This dataset has 17 columns and 11,341 unique figures. These figures were deemed as “historical” based on a combination of Wikipedia page views and the number of foreign language Wikipedia pages that exist on the figure. Attributes for each figure include their birth year, country in which they were active, and their occupation. This data offers an interesting insight into what it takes for a person to be considered historical, and the dataset lends itself to many interesting questions.

A table describing each feature in the dataset is featured on the last page. The features used in the visualizations include those describing the figure’s profession - three identifiers are used, domain, industry, and occupation. Domain is the broadest, with only 8 values, and industry and occupation are respectively more specific, which allows for binning of historical figures based upon the fields in which they were active. Country, state, and city are also listed for figures for which this data is known. The locational data is based upon the city in which the figure was primarily active, not necessarily the place in which they were born. A third feature used in visualization was sex; the only options are male and female, none of the historical figures are considered to be non-binary in this dataset. Additionally, each figure is given a historical popularity index (HPI). This is a popularity score used to rank the historical figures.

Five analytical questions that could be asked of the data include:

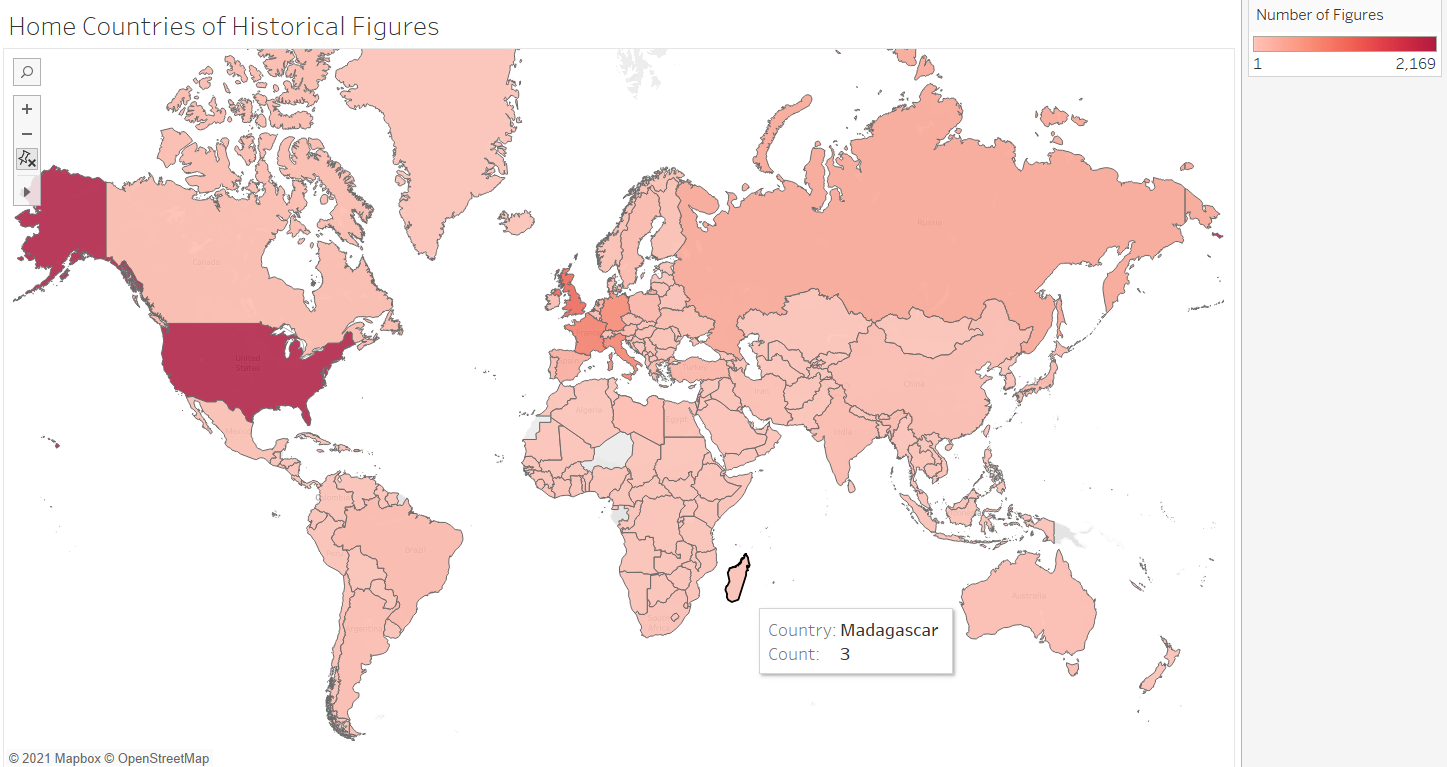
* 1. What parts of the world have produced the most historical figures?
  2. Were there more historical figures of one era than another?
  3. Which occupations are historical figures most likely to have? How have those occupations changed over time?
  4. Is there Western bias in the dataset? i.e. Does this dataset truly reflect the historical figures of the entire world?
  5. What factors do the most popular historical figures have in common? Is there a particular occupation or country that lends itself to more popularity?

To answer these questions, five visualizations were created. There was redundancy among them, and it turned out that a single dashboard using only three was able to tackle all five of the questions. The first visualization is a tree map of domains, industries, and occupations:



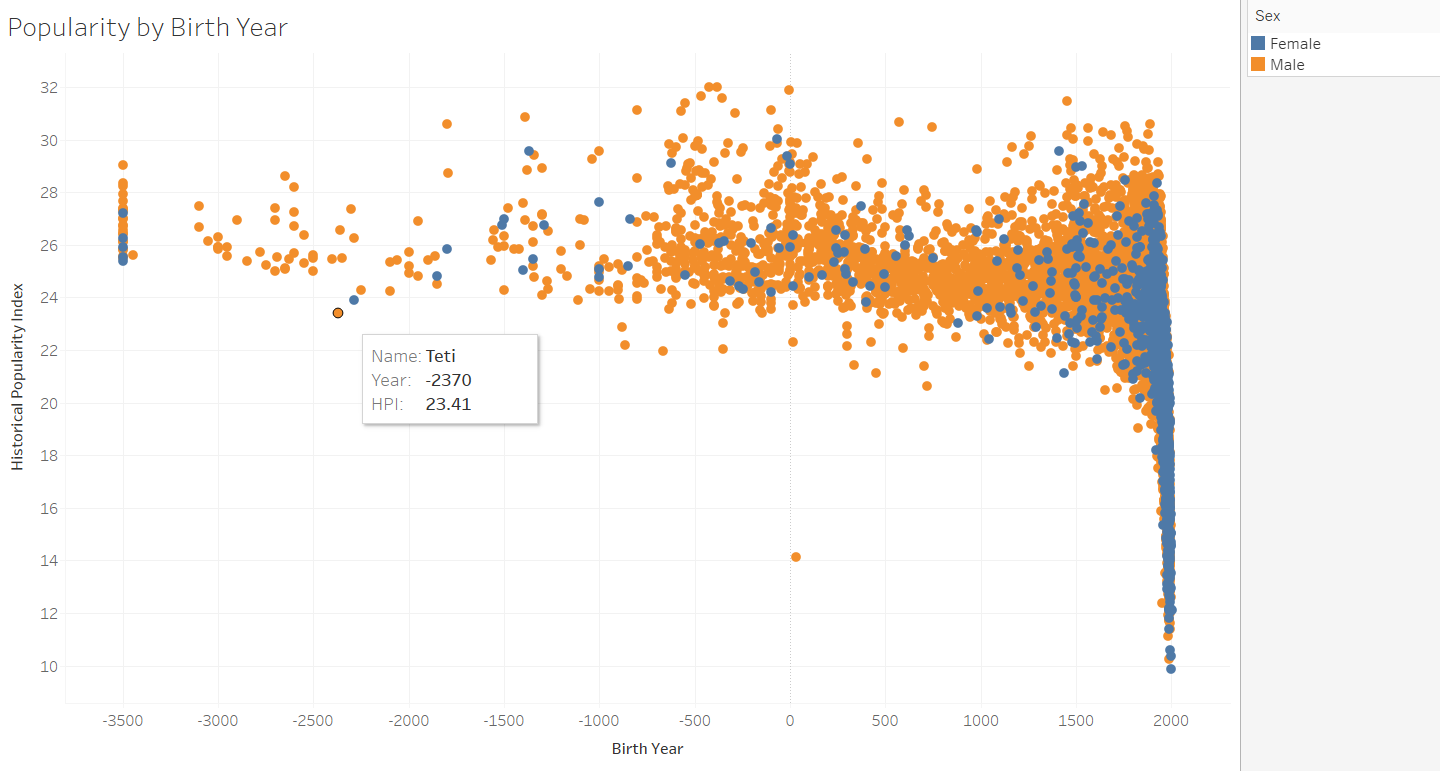
Each domain is represented with a different color. Within the domains further segmentation occurs based on industry, distinguished with thick white barriers. Finally, each occupation is represented by its own square. Hovering over any one square brings up a tool tip which lists the count of historical figures who have that occupation. The size of each box is indicative of this count, so I felt displaying the count in each square was unnecessary, but provided the information in the tool tip for users who wanted more details. Tableau automatically chooses whether the font color should be black or white based on the background color, and I had difficulty finding a color scheme where all the font colors were the same, so this is an area where there is room for improvement. This visualization answers the question of which occupations historical figures are most likely to have. It becomes immediately clear that most of them are Politicians. When combined in the dashboard, we will be able to answer the second part of this question, about whether the occupations change over time.

To answer the question about which parts of the world produced the most historical figures, a map was an obvious choice.

****

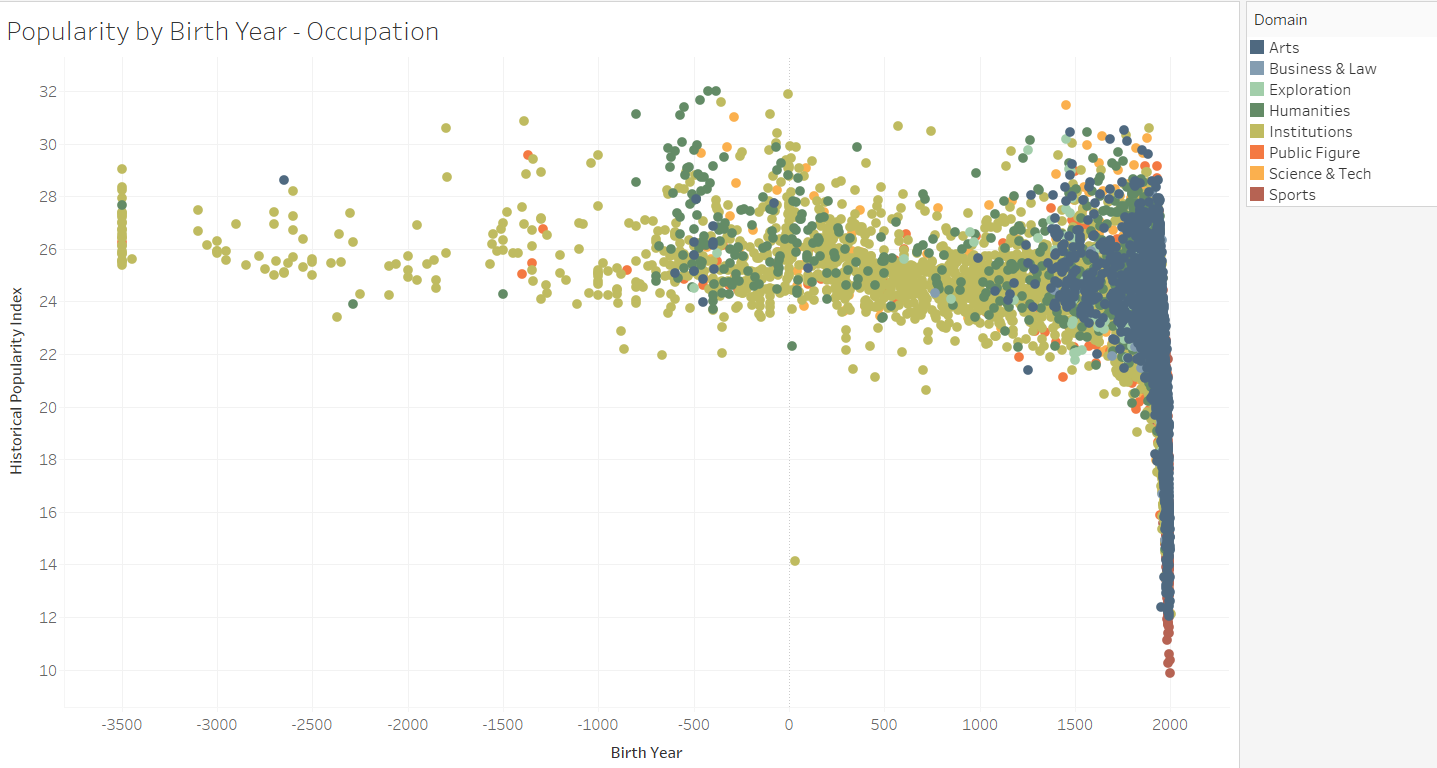
This map shows countries of the world shaded based on how many historical figures were attributed to that country. The color scale is from light pink to a dark maroon, chosen because of the continuous data. A downside to the color scale is that it is difficult to see small distinctions: many countries have less than 10 figures, while the United States has over 2000. Thus, it is difficult to purely visually answer questions about specific countries, such as: “Are there more historical figures from China or India?”. However, the map is interactive, and hovering over a country gives you a count, so the user has the ability to answer those questions for themselves. Immediately answered by the map is the question of which country has the most historical figures. The map also lends data to the question of Western bias: a large majority of historical figures are from Western countries. A counterpoint to the idea of bias is that most countries of the world are represented with at least one historical figure. Answering the question of bias definitively would require more investigation, but this map is a good starting point.

The third visualization is a scatter plot with the historical popularity index (HPI) on the y-axis and birth year on the x-axis.



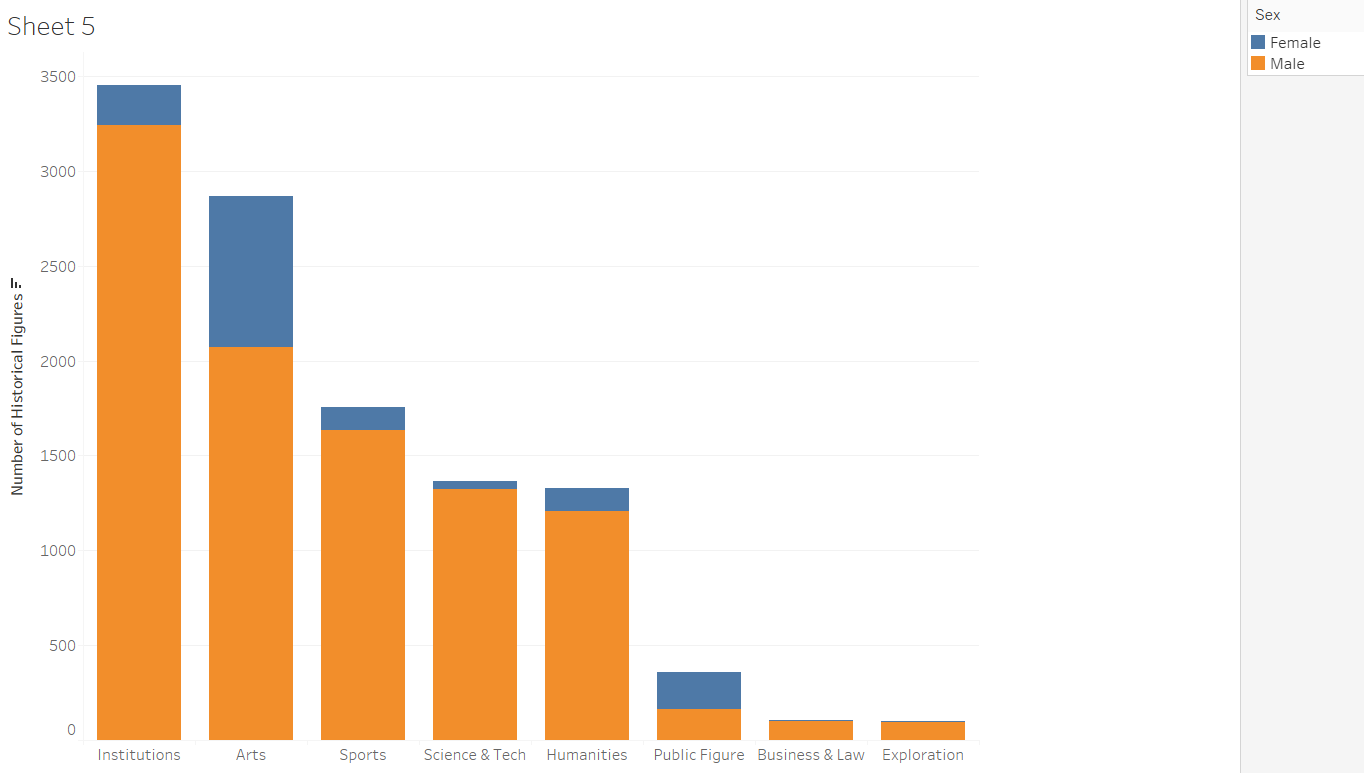
I wanted the viewer to be able to access the names of the historical figures if they so chose, and a scatter plot was the most obvious way to do this. As there are over 10,000 figures, the scatter plot does have cluttered and overlapping points. The user is able to zoom in which helps, but this problem is better solved in the dashboard when filters are added to get down to a more manageable amount of points. The dots are colored based on sex using contrasting colors, which gives an immediate view of how few female historical figures there were until quite recently. Hovering over a point shows the historical figure’s name, birth year, and HPI. This visualization answers the question of whether there are more historical figures from one era than another. The sheer density of the scatter plot in the modern era shows that there are far more historical figures born in this time range. Also interesting is how much less popular younger figures tend to be.

After creating this scatter plot, I thought it might be more interesting to color the points based on the domain the figure worked in rather than sex:



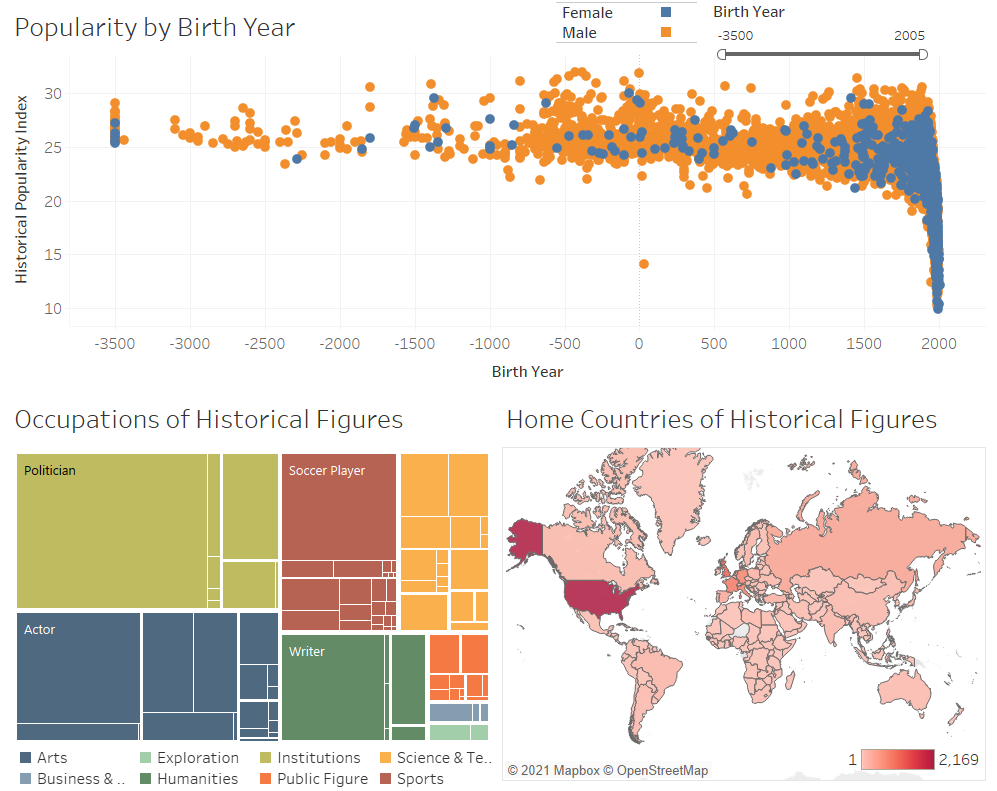
This version is able to clearly answer the question of how popular occupations have changed over time. Antiquity was full of famous historical politicians, and now we have many more artists. However, because of the volume of points and the overlap in modern birth years, I found this visualization to be too distracting.

I experimented with a very simple bar chart to show the number of figures per domains with a color distinction based on sex.



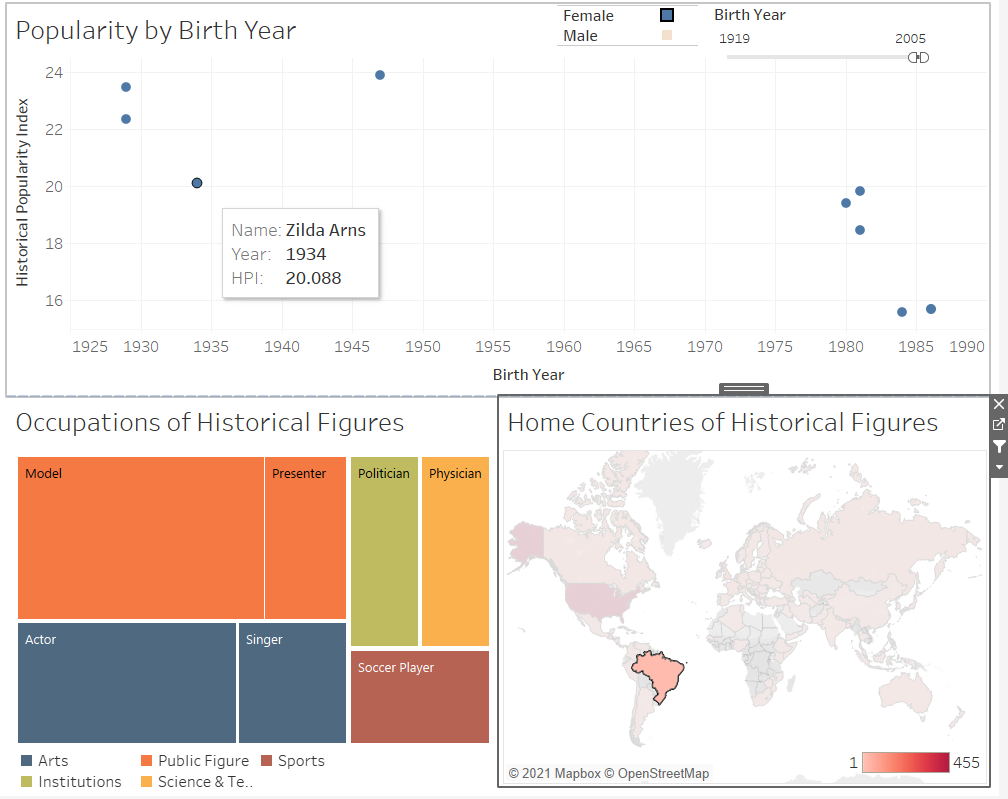
While this very plainly outlines that institutions are the most popular domain, the tree map also conveyed this data along with more specific information about industry and occupation, so the bar chart felt redundant.

The final comprehensive dashboard combines the first three visualizations - the tree map, the world map, and the scatter plot colored by sex. It can be accessed on Tableau Public [here](https://public.tableau.com/profile/lauren6941#!/vizhome/Project3_16147971034750/Dashboard1) [3].



This dashboard is able to showcase many of the features of the dataset: domain, industry, occupation, sex, name, birth year, and HPI. The three visualizations offer entirely different views of the data, so redundancy is avoided. All three visualizations act as a filter. If you click a country on the world map, the other two visualizations will adjust to only contain the figures from that country, which allows users to explore which occupations figures from a certain country have in the tree map, and narrows down the number of points in the scatter plot so they can investigate the names and birth years of those figures as well. The same happens with the tree map, if a user selects a specific occupation, the map and scatter plot adjust to only contain data of historical figures with that occupation. If a user selects a specific point in the scatter plot, the tree map will display just a single occupation, and only the country associated with that one figure will be colored. This allows users to view all the attributes of specific historical figures if they so choose.

Aside from the filters in each visualization, the dashboard uses two global filters. The first is Birth Year. The time slider allows a user to pick a specific era, and all three visualizations adjust to show only historical figures from that era. Interaction with this time slider allows a user to visualizae how occupations and locations of historical figures have changed over time. The second global filter is for sex, so users can see how occupations and locations vary based upon sex. This filter also eliminates the overlap in the scatter plot, giving users a better opportunity to narrow in on specific figures. The sex filter was created slightly differently than the standard procedure for making a global filter. The scatter plot came with a legend that demonstrated the colors for male and female, but legends in Tableau cannot act as global filters. A global filter cannot have boxes to indicate the colors of each label, so using a standard global filter also required use of a legend, meaning there were two places on a single dashboard that labeled the sexes. To avoid this, I dropped both the legend and global filter and created a sixth worksheet with a simple table for the sexes with their respective coloring. I added this as a fourth worksheet to the dashboard and made it a filter for the other worksheets, creating the same functionality as a global filter without the redundancy. Below shows some of these filters in action – this is looking for females born between 1919 and 2005 from Brazil. Each of the three visualizations has adjusted accordingly:



Let us now analyze how the dashboard answers our original questions. First, what parts of the world have produced the most historical figures? The map tells us quite plainly, and if we choose to filter by birth year, we can explore how the figures moved across the world over time. Secondly, were there more historical figures of one era than another? Without any filters, the scatter plot demonstrates this is true. What occupations are historical figures most likely to have? This is answered by the tree map, and if we use the time filter, we can also explore how those occupations have changed over time. Using the sex filter we can also explore how occupations vary between the sexes. As mentioned earlier, the map indicates there may be Western bias in the dataset. Using the time filter we see this is specifically apparent amongst figures of the last few centuries. Is this bias or is it an accurate representation of historical figures of the world? That would require more complex research. Finally, to answer what factors the most popular figures have in common, we can think about the dataset holistically. General trends in the dashboard give a good sense of what makes a figure popular: the most common figures of past eras are male politicians, whereas in the modern age they’re female singers from the US. A user may wonder, what makes the most popular of the historical figures more popular than others? This is not as simple to answer. It would require clicking on specific high popularity figures to find their occupation and birth countries, and comparing those with less popular figures. Perhaps the other scatter plot with domain would have made answering this question a bit simpler. However, the three visualizations chosen do the best job of answering our main analytic questions with the least amount of overlap among them. The dashboard illuminates many interesting features in the data that would otherwise be quite difficult to visualize.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Data Type** | **Min** | **Max** | **Avg** | **Most Frequent** | **# Missing** | **Interesting Features** | **Description** |
| **article\_id** | Nominal Number[2] | 307 | 3877829 | - | - | 0 | Unique | Unique identifier for the Wikipedia article of the historical figure’s biography |
| **full\_name** | Nominal | - | - | - | No value occurs more than twice | 0 | 13 sets of duplicates:  11321 unique | Full name of the historical figure |
| **sex** | Nominal | - | - | - | Male (9846) | 0 | 86.8% male | Sex of the historical figure – male and female are the only options |
| **birth\_year** | Quantitative | -3500 | 2005 | 1658.2 | 1981 (143) | 2 | Only 2 figures born after 2000, std = 690.62 | Year of birth (or best estimation of) |
| **city** | Nominal | - | - | - | Other (512), NYC (338) | 0 | 5096 unique cities | The city in which the figure was primarily active |
| **state** | Nominal | - | - | - | NY (441) | 9169 | 79 unique states / regions | The state (if applicable) in which the figure was primarily active |
| **country** | Nominal | - | - | - | United States (2169) | 33 blank, 435 Unknown | 195 unique countries | The country in which the figure was primarily active |
| **continent** | Nominal | - | - | - | Europe (6368) | 30 blank, 408 Unknown | No figures from Antarctica | The continent in which the figure was primarily active |
| **latitude** | Quantitative | -51.63 | 70.07 | 39.93 | 48.85670 (210) | 1047 | 4493 unique lats | The central latitude of the city, state, etc. where the figure was primarily active |
| **longitude** | Quantitative | -175.20 | 178.02 | -7.04 | 2.35080 (210) | 1047 | 4768 unique lons. Most popular lat,lon = (48.85, 2.35) which is Paris | The central longitude of the city, state, etc. where the figure was primarily active |
| **occupation** | Nominal | - | - | - | Politician (2530) | 0 | 88 unique | The historical figure’s specific occupation |
| **industry** | Nominal | - | - | - | Government (2705) | 0 | 27 unique | The industry (or topic area) the figure concentrated their work in |
| **domain** | Nominal | - | - | - | Institutions (3456) | 0 | 8 unique | The general area of contribution the figure is known for |
| **article\_**  **languages** | Quantitative | 26 | 214 | 40.77 | 26 (812) | 0 | Std = 17.47 | How many of the different language Wikipedias have an article on the figure |
| **page\_**  **views** | Quantitative | 1965 | 145,250,649 | 4,200,960 | No value occurs more than twice | 0 | Std = 7,677,474 | The estimated total page views for the figure across all Wikipedias |
| **average\_**  **views** | Quantitative | 49 | 1,515,232 | 89,413.60 | No value occurs more than 4 times | 0 | Std = 121,461.30 | The estimated average page views fr the figure per each Wikipedia edition article |
| **HPI** | Quantitative | 9.8794 | 31.9938 | 22.30 | No value occurs more than 3 times | 0 | Std = 3.35 | Historical Popularity Index - an index value measuring approximately how popular the article on this figure was, and hence, how well known this historical figure was.\* |

\*HPI is derived from age of the historical figure, concentration of page views among different languages, coefficient of variation in page views, and number of page views in languages other than English [1]

**References**

[1] https://www.kaggle.com/mit/pantheon-project

[2]https://en.wikipedia.org/wiki/Nominal\_number#:~:text=%22Nominal%20number%22%20can%20be%20broadly,no%20information%20other%20than%20identification%22.

[3] Dashboard link: <https://public.tableau.com/profile/lauren6941#!/vizhome/Project3_16147971034750/Dashboard1>