Exploring the Pantheon Dataset

Kevin Le - kl2921

May 1, 2018

1 Introduction

For my project, I decided to investigate the Pantheon project dataset from Kaggle. This is a dataset curated by MIT researchers that contains information about historical figures documented in Wikipedia. The questions I explored were as follows:

1. How likely is it that a historical figure born before the end of the Middle Ages is more popular than a historical figure from the past two centuries?
2. What is the strength of the relationship between historical popularity index (HPI) and the number of article languages for Europe and North America? How well can HPI an article languages predict being a historical figure from Europe or North America?
3. What occupations of popular historical figures are most frequent in different continents?

My interest in studying this is on a social and cultural level. I believe historical figures (or what we consider historical figures) and their popularity are indicative of what different societies value. Furthermore, Wikipedia is directly curated by, and accessible to, the public, further making it a solid source for gaining insights about historical figures as a reflection of society.

2 Data Set

As mentioned, I used the Pantheon project dataset from MIT, which contains information about historical figures from around the world documented in Wikipedia with associated popularity indices.

To clean the data set, I removed the 'state' column (as that only appears to the United States). I then removed all remaining rows that carried a 'nan' or empty value. Finally, birth years were being read in as strings, so I applied a function that converted the birth years to integers.

After cleaning, the data set was about 10,000 lines and was a large sample that was only moderately computationally heavy, so I decided to use all of it.

The data set had the following column labels: article_id, full_name, sex, birth_year, city, country, continent, latitude, longitude, occupation, industry, domain, article_languages, page_views, average_views, and historical_popularity_index.

To be less verbose, I will sometimes refer to historical popularity index as HPI.

3 How likely is it that a historical figure born before the end of the Middle Ages is more popular than a historical figure from the past two centuries?

3.1 Method 1: Hypothesis Testing

The first method was a hypothesis test on the difference of the means of the historical popularity index between historical figures born before the end of the Middle Ages and of historical figures born in the past two centuries.

- Null hypothesis: The mean historical popularity index of historical figures from the past two centuries does not significantly deviate from the mean historical popularity index of historical figures born before the end of the Middle Ages.

- Alternative hypothesis: The mean historical popularity index of historical figures from the past two centuries does significantly deviate from the mean historical popularity index of historical figures born before the end of the Middle Ages.

I ran 10000 simulations to find an empirical P-value and to graph the distribution of mean differences. The following are the results:

- Observed Statistic: -4.19691660794
- Empirical P: 0.0

Graph: The bulk of the simiulated distribution is centered around 0. Under the null hypothesis,

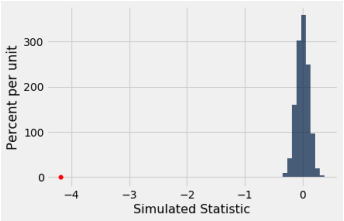


Figure 1: Hypothesis test results on difference of means

the historical popularity index of historical figures born in the past two centuries is a random sample from the all the indices, i.e. the same set as historical figures born before the end of the Middle Ages. Since both are random samples from the set of historical popularity indices under the null hypothesis, the two sets should have about equal historical popularity indices around 0 and their difference should be around 0.

The observed value of the test statistic is far to the left from the center of the distribution. Additionally, the empirical P-value is 0.0, indicating that the observed statistic will always tend towards the direction of the alternative hypothesis. So,

- We reject the null hypothesis and conclude that the difference between the average historical popularity indices of historical figures born in the last two centuries and historical figures born before 0 CE are too large to be a result of chance variation alone.
- Because the observed statistic is far to the left, this indicates that, on average, the historical popularity indices of historical figures born before 0 CE are higher than that of historical figures born in the last two centuries. In other words, historical figures born before 0 CE are more popular on Wikipedia than historical figures born in the last two centuries.

3.2 Method 2: Bootstrapping Confidence Intervals

To corroborate the above, I compared the two-sided 95% confidence intervals for the mean historical popularity index of figures born before the end of the Middle Ages and figures born in the last two centuries.

- Observed statistic: -4.196916
- 95% Confidence Interval of HPI mean difference is from -4.192829 to -4.184548

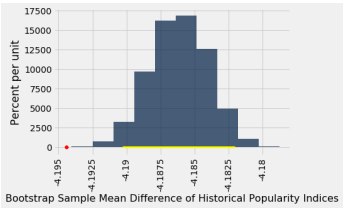
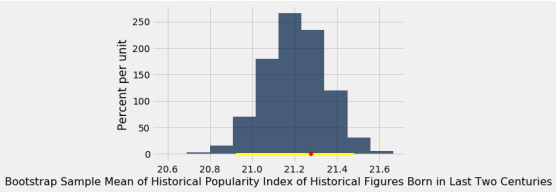


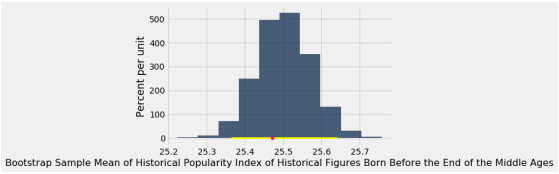
Figure 2: 95% Confidence interval of HPI mean difference

It appeared that the observed statistic did not fall within the 95% confidence interval of the mean difference between the average historical popularity index of historical figures born before the end of the Middle Ages and historical figures born in the last two centuries. This questions the results of the hypothesis test.

I decided to dig further, and look at each time period separately.



(a) 95% CI mean, past two centuries



(b) 95% CI mean, before end of Middle Ages

Figure 3: 95% individual confidence interval means

- Average Historical Popularity Index of figures born before the end of the Middle Ages: 25.471484
- 95% confidence interval of mean popularity of historical figures born before the end of the Middle Ages is from 25.364319 to 25.640782
- Average Historical Popularity Index of figures born in the last two centuries: 21.274567
- 95% confidence interval of mean popularity of historical figures born in last two centuries is from 20.921949 to 21.480005

While the observed difference of means statistic did not fall into the 95% confidence interval, the two means that were used to calculate the observed statistic did fall in a 95% confidence interval individually. Next, I decided to overlay the two histograms.

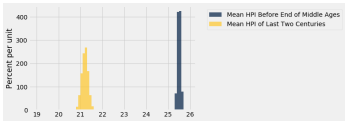


Figure 4: Overlaid distributions

In the histogram above, it can be seen that there is a zero percent chance of the historical popularity indices of historical figures born before the end of the Middle Ages and historical figures born in the last two centuries intersecting, on average. This is consistent with the observed empirical P-value of 1.0. The histogram of mean HPI before the end of the Middle Ages is to the right of the mean HPI of the last two centuries, thus indicating more specifically that the HPI of historical figures born before the end of Middle Ages will always be greater than the HPI of historical figures born in the last two centuries, on average.

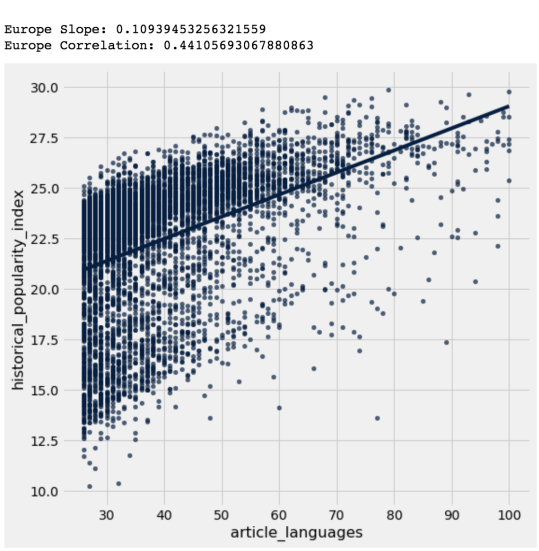
Therefore, although the observed statistic does not fall in a 95% confidence interval, the disparities of the individual mean HPIs, as shown above, further support the alternative hypothesis and further provide evidence that the mean HPI of historical figures before the end of the Middle Ages is greater than the mean HPI of historical figures born in the last two centuries.

A possible explanation of this is that prominent historical figures that were born before the end of the Middle Ages must be very prominent to still carry relevance today. Meanwhile, even if the last two centuries have more historical figures, most are likely less prominent than say, Aristotle.

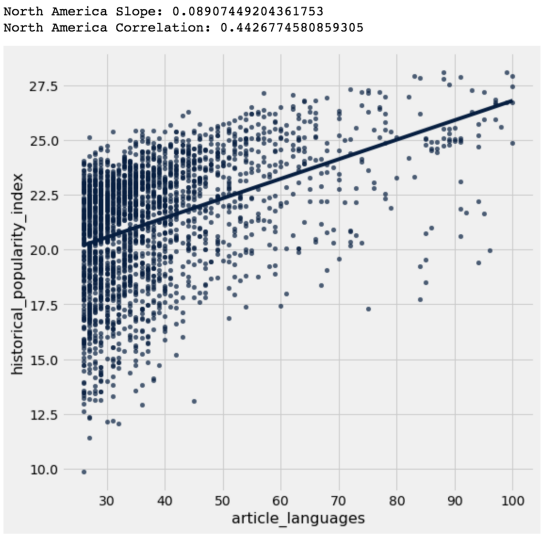
4 Q2. What is the strength of the relationship between HPI and the number of article languages for Europe and North America? How well can HPI an article languages predict being a historical figure being from Europe or North America?

4.1 Method 1: Linear Regression

To test the strength of the relationship between HPI and article languages, I first did a linear regression test.



(a) Europe HPI vs. Article Language (Linear)



(b) North America HPI vs. Article Languages (Linear)

Figure 5: Europe and North American linear regression plots

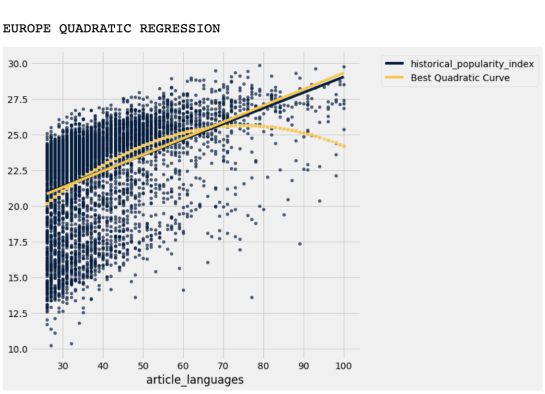
To make the above graphs, I used the fit_line parameter for scatter plots and I filtered out data above 100 article languages, as almost all the points fell below 100 languages and I felt the higher data could skew the results. I chose to focus on Europe and North America because they are both large datasets of similar size. Additionally, concerning article languages, Europe is much more multilingual than North America.

In both the European and North American case, there appears to be a moderate positive correlation between the historical popularity index of a figure and the number of article languages it is written in. The correlation values are both about 0.45 and the slopes are positive.

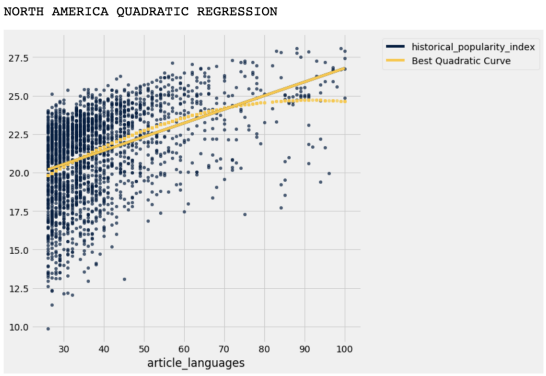
Interestingly, after cleaning the data the slopes for both plots became steeper, yet the correlation values for Europe went down slightly (it was initially 0.47). This means the outliers were skewing the data towards a more negative yet apparently stronger correlation for Europe, but weaker correlation for North America.

4.2 Method 2: Quadratic Regression

I was not sure if a linear regression was the best way to model the data as the correlation was only moderate, so I decided to compare with quadratic regressions.



(a) Europe HPI vs. Article Language (Quadratic)



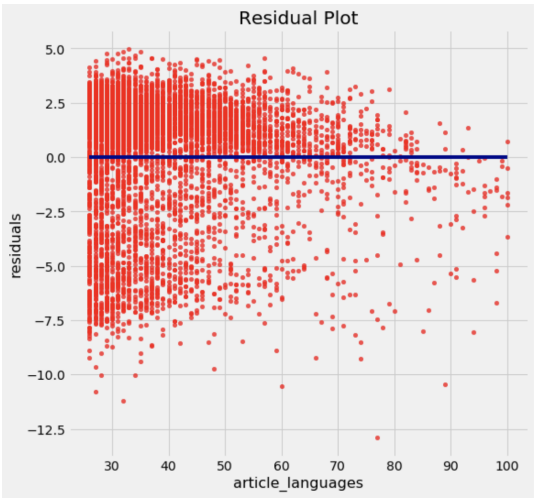
(b) North America HPI vs. Article Languages (Quadratic)

Figure 6: Europe and North American quadratic regression plots

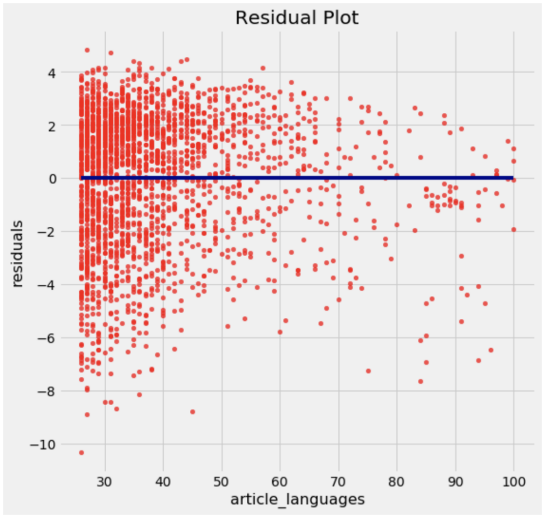
For North America, the quadratic regression roughly matches the linear regression until towards the end. On the other hand, for Europe, the quadratic regression is much more significant and appears to be a better fit than the linear regression.

4.3 Method 3: Residuals

To better visualize the results of the quadratic and linear regressions, I plotted the residuals of each continent.



(a) Europe linear regression residual plot



(b) North America linear regression residual plot

Figure 7: Europe and North American residual plots

For Europe:

- The residuals are approximately distributed asymmetrically above and below 0, with much more residuals above 0. This indicates that the linear regression was not a reasonable method of estimation and that the quadratic regression may be a better fit.
- As mentioned, most of the residuals are above 0, meaning that there is often a higher popularity index than expected for a given amount of article languages
- The quadratic curves downwards after about 70 article languages

For North America:

- The residuals are approximately distributed symmetrically above and below 0, thus indicating that the linear regression was a reasonable method of estimation. It is not a perfect symmetry, but it does support the quadratic and linear regression being roughly similar.

A potential explanation for this is that Europe is much more multi-lingual than North America. Therefore, a historical figure can be more popular with fewer languages. Furthermore, European historical figures are more familiar to Europeans and will be more likely to be searched by multi-lingual Europeans. On the other hand, most North Americans are unilingual, possibly resulting in a more incremental change in popularity as the number of article languages increase.

4.4 Method 4: Classification (k-Nearest Neighbors)

Next, I wanted to flip how I was using the data by seeing how well historical popularity index and article languages could predict what continent the historical figure came from, again focusing on North America and Europe to do a binary classification.

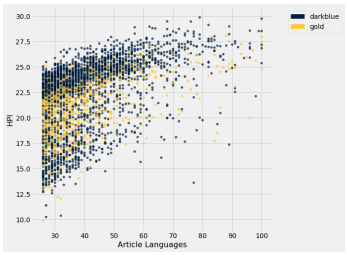


Figure 8: Scatter plot of training data

Gold dots represent North America, and dark blue dots represent Europe. As seen, North American figures appear to be a middle band between European figures.

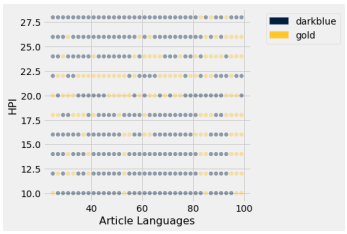


Figure 9: Test grid

This reveals some mixing between North America and Europe, there is no clear distinction (i.e. everything is one region or another), but it is patchy and predominantly European.

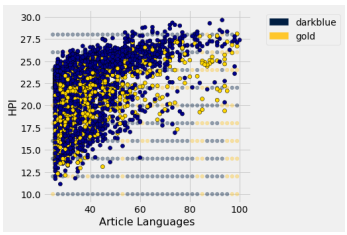


Figure 10: Overlaying test data on test grid

The visual data is still very cluttered and it is hard to tell what is going on, so I decided to defer to a quantitative approach by simulating the accuracy of the data 100 times. To do this, I sampled 1000 data points from both the training and test set (otherwise the computations would take too long to simulate).

After simulating 100 times, I found an accuracy of 73.3%. This indicates that historical popularity index and article languages is a strong predictor of whether a historical figure is from North

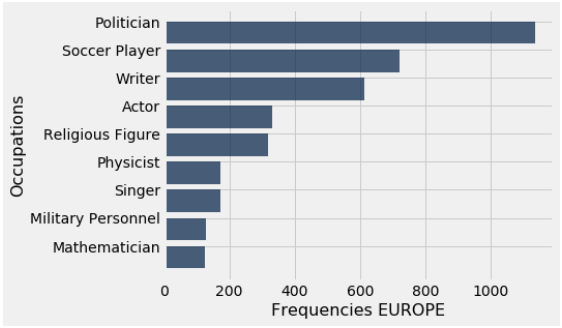
America or Europe. Going back to the differences in the regression of each, it is clear that the relationship between historical popularity index and article languages for the two continents are distinct from one another. In connection to the residuals, it is likely that these differences are mostly attributed to the number of article languages because the historical popularity indices of both North America and Europe are in a small range and appear to be similarly distributed (HPI of 10 to 30), whereas the article language range is much larger for both and skewed more towards the left for Europe.

5 Q3. What occupations of popular historical figures are most frequent in different continents?

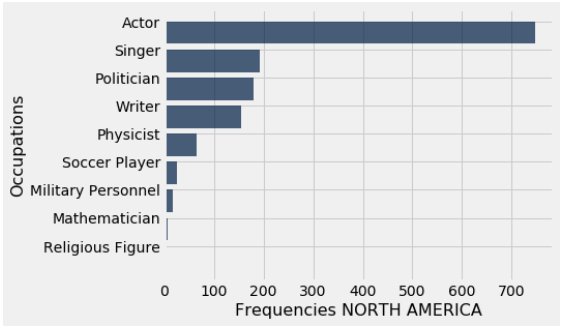
The occupations I initially looked at were: politician, military personnel, religious figure, soccer player, physicist, mathematician, writer, actor, and singer. The Pantheon project details about 88 different occupations, which would be too much for data visualizations. I selected a set of occupations that I felt demonstrated a variety of occupations both technical and non-technical and that I felt most historical figures fell under. Additionally, by selecting historical figures, I could create a standardized set of occupations to compare continents with. One note is that Australia and Antarctica were not in the dataset to begin with.

5.1 Method 1: Bar Graph Visualization of Selected Occupations

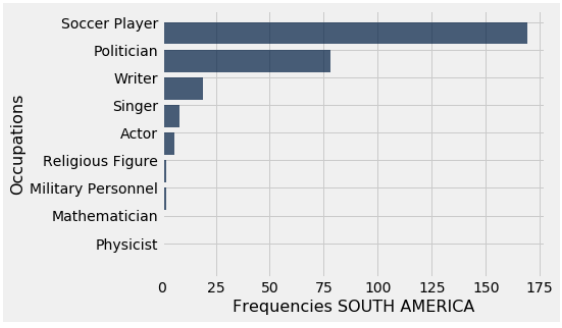
The first way I decided to visualize the data was through a categorical bar chart.



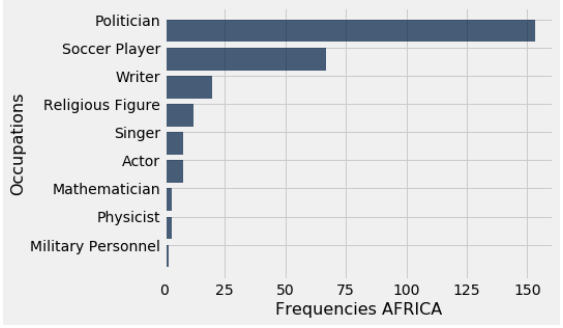
(a) Europe occupations bar chart



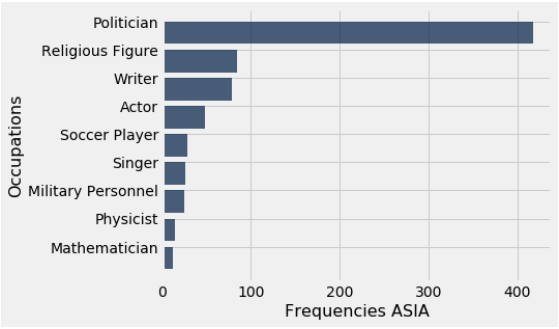
(b) North America occupations bar chart



(c) South America occupations bar chart



(d) Africa occupations bar chart



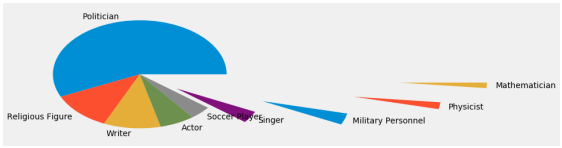
(e) Asia occupations bar chart

Figure 11: Bar charts of occupation by continent

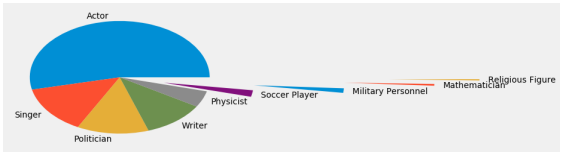
The bar charts give an idea of what occupations historical figures of each continent tended to have. However, I after constructing the bar charts, I felt it was not the most effective way of presenting the data. It showed the frequencies of each occupation on each continent, however there are some continents that have more historical figures in general than others (e.x. North America has more politicians than Africa, but it it is Africa's most frequent occupation and only North America's third most frequent occupation).

5.2 Method 2.1: Pie Chart Visualization of Selected Occupations

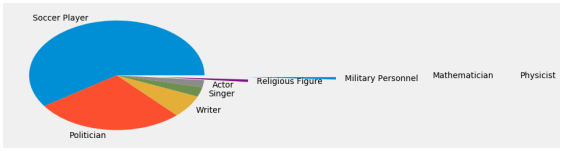
I decided to use a pie chart to demonstrate each occupation as parts of a whole to better visualize the relative frequency of occupations by continent.



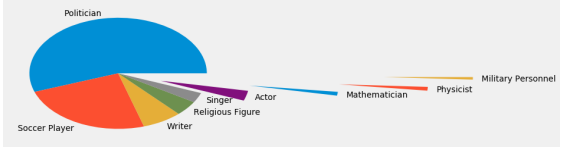
(a) Asia occupations pie chart



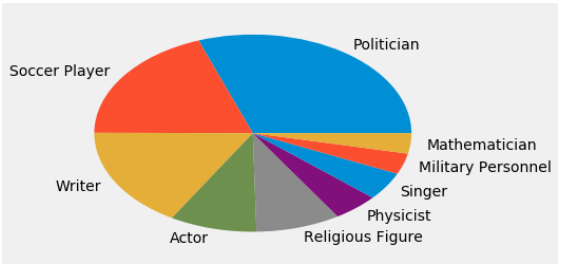
(b) North America occupations pie chart



(c) South America occupations pie chart



(d) Africa occupations bar chart



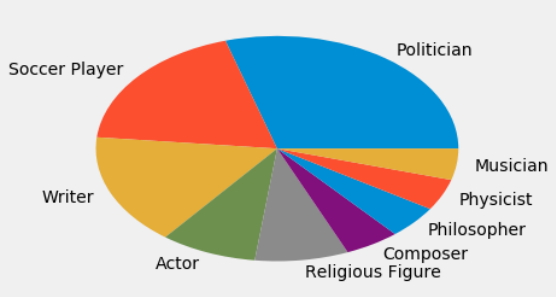
(e) Europe occupations pie chart

Figure 12: Pie charts of occupation by continent

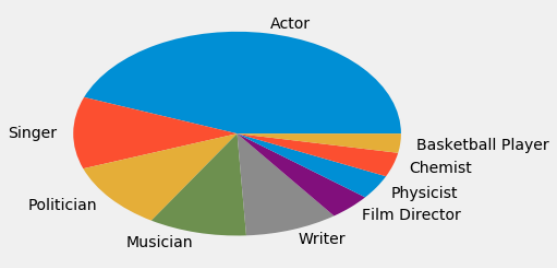
This visualization is much clearer. Of the selected occupations, Asian, African, and European historical figures are most frequently politicians, North American historical figures are most frequently actors, and South American historical figures are most frequently soccer players.

5.3 Method 2.2: Pie Chart Visualization of Top 9 Occupations By Continent

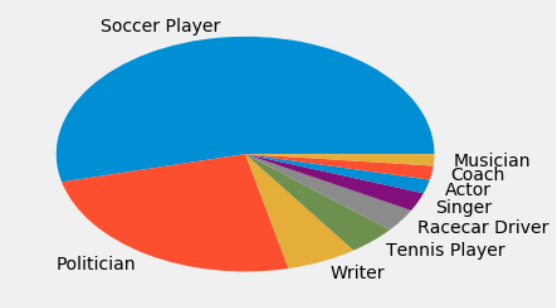
However, my results might be skewed since I standardized across a single set of occupations. I wanted to see what the top 9 occupations by continent were. I chose 9 since that was the size of my selected occupations set.



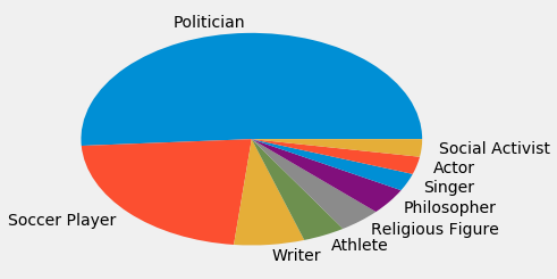
(a) Europe top 9 occupations pie chart



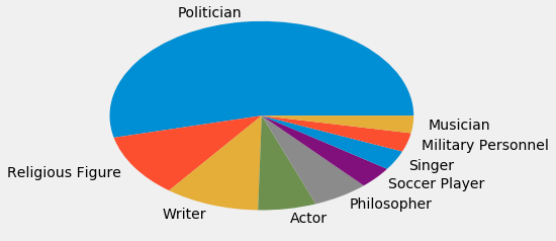
(b) North America top 9 occupations pie chart



(c) South America top 9 occupations pie chart



(d) Africa top 9 occupations bar chart



(e) Asia top 9 occupations pie chart

Figure 13: Pie charts of top 9 occupations by continent

The top 9 occupations and the selected occupations revealed the following:

- Most frequently, Europe, Asia, and Africa's historical figures are politicians.
- Most frequently, North America's historical figures are actors and artists.
- Most frequently, South America's historical figures are soccer players.
- Globally, soccer players are popular everywhere but North America
- No continent's occupations are frequently STEM figures. In fact, STEM occupations do not appear at all in South America, Africa, or Asia's top 9 occupations.

This represents a reflection on the environment of each continent. North America and especially the United States dominates popular media, hence the high amount of notable actors, singers, and musicians. Europe, Asia, and Africa have well-documented political histories that date very far back (in contrast, U.S. history is limited to about the past three to four centuries). This also explains why these three continents have higher proportions of religious figures than the Americas, whose long-term history is not as well documented, likely due to the impacts of colonialism.

With regards to South America, it demonstrates a large culture around and dominance of soccer. Conversely, soccer is not big sport in North America.

Interestingly, no continent has a large proportion of notable STEM figures. However, this is not to say historical STEM figures are not important, just not as frequent. This indicates who we most frequently consider notable as we look back on history: people who moved others through words, art, and athleticism.

6 Conclusion

In the end, each of three questions can be summarized as follows:

Question 1: How likely is it that a historical figure born before the end of the Middle Ages is more popular than a historical figure from the past two centuries?

- On average, historical figures born before the end of the Middle Ages are always more popular than historical figures born in the last two centuries. This may be because historical figures born before the end of the Middle Ages most be extremely notable to stand the test of time and be relevant today.

Question 2: What is the strength of the relationship between HPI and the number of article languages for Europe and North America? How well can HPI an article languages predict a historical figure being from Europe or North America?

- There is a moderate positive correlation (0.45 in both cases) between the historical popularity index and the number of article languages in both Europe and North America. However, Europe is better modeled by a quadratic regression. This means that given a European historical figure and a North American historical figure of equal popularity, a European figure will have less associated languages than the North American figure. This may be because Europe is more multilingual than North America.

- Using a k-nearest neighbors binary classification, HPI and article languages were able to predict whether or not a historical figure came from North America or Europe (given a dataset containing only those continents) about 73.3% of the time. This supports that Europe and North America have distinct relationships between historical popularity index and article languages. This may be more attributed to article languages than the historical popularity index, as discussed above.

Question 3: What occupations of popular historical figures are most frequent in different continents?

- In general, Europe, Asia, and African historical figures are most frequently politicians. North American historical figures are typically performers and artists. And South American historical figures are typically soccer players. Only Europe, Asia, and Africa had a significant amount of religious figures. No continent had a large frequency of STEM figures. All of this reveals Europe, Asia, and Africa's extensive political and religious history. However, it is important to note that in North America and South America, documentation of political and religious history may be more sparse due to the impacts of colonialism. North American historical figures frequency in performers and artists is indicative of North America's influence on popular culture around the world. And South America is dominant in the most popular sport in the world: soccer. But, across all five continents there is a general trend of art, humanities, and sports figures being considered historically notable more frequently than STEM figures.

Through all of these questions, one can gain an increased insight on how people are valued in history and how those values mirror different the places these historical figures come from. Question 1 revealed that, universally, people value the deep roots of the past. Question 2 showed how linguistic identity (in this case, multilingual vs. unilingual) impacts how popular a figure is – figures from multilingual places can have equal popularity from a unilingual place with fewer associated article languages. Question 3 was perhaps the most revealing, showing that society as a whole values the qualitative over the quantitative with regards to historical figures.