Visualising “Pantheon” Data-set on Tableau

Pantheon Dataset

The Pantheon dataset contains information of globally famous people. The attributes included in the dataset correspond to information about country of origin, date of Birth, and the popularity of the person (HPI). Table 1.1 shows the variables categorised according to their variable type.

CATEGORICAL	• Occupation, Domain, Industry (Nominal) • Gender (Nominal) • Continent Name, Country (Nominal)
Quantitative	• Year of Birth (discrete) • HPI (continuous) • En Curid :Unique identifier(Discrete)

Table 1: Categorising the variables of interest

Due to the nature of the columns provided, many of the variables are correlated within each other. For example, county location, city state, continent, lat, long all represent information corresponding to the geographic location of the famous people. Similarly, The HPI index is calculated using a combination of Lstar, total page views (page views in non –eng +page views in eng). Hence, the attributes that are visualised in the dataset are selected using the question at hand.

Initial exploration of the dataset

The dataset provided is relatively large with 11341 observations and 24 variables.

The initial exploration is therefore conducted by splitting into two groups with Top 20 selected from their high HPI scores: “Very famous”, “Not Famous”.

For the purpose of this report, we look to answer:

How has occupation and gender of “globally famous” people changed over time? Is there an observable trend?

The initial sketches are therefore focused to answer this question at hand.

Sketches 2-4 in the attached sketches.pdf were initial sketches that were explored quickly but violated both the *expressiveness and effectiveness principle*. For example, Sketch 2: indicates some sort of natural ordering to the occupation, however the area in each category is used to represent the count of number of people in their respective occupation.

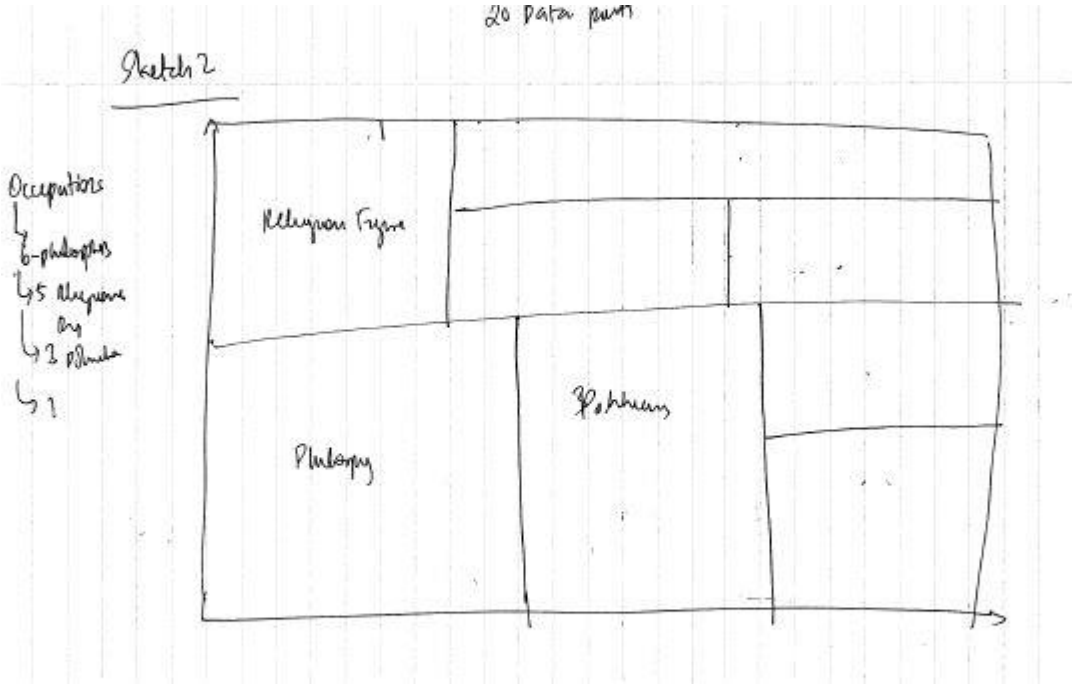


Figure 1: Sketch2: Visualising occupations of very famous people (sample size: 20)

From the initial sketching process, the below sketch accommodates both the expressiveness and effectiveness principle.

The categorical attributes uses *Identity channels*: “Occupation” is coded using spatial region, while the “Gender” attribute is encoded using Colour Hue (Green-Male, Blue-Female for the below sketch only). The quantities variable: birth year is encoded using *magnitude channels* via position on a common scale as seen in Figure 2 below.

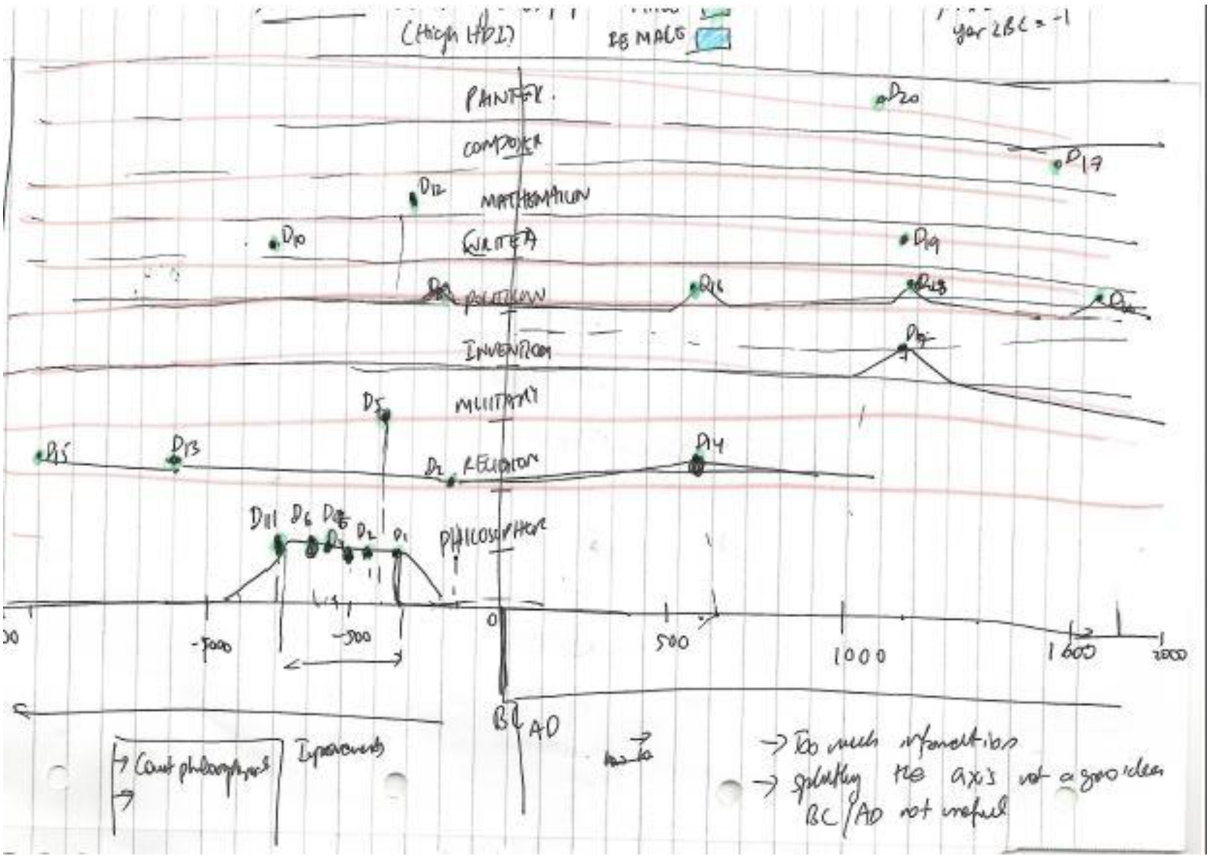


Figure 2: Final Sketch: Most popular Occupations of top 20 globally famous people

Implementation in Tableau

The above sketching process raises a few questions :

1. Is there a logical way to group different time periods to produce a succint visualisation?
2. Can all the 88 different occupations visualised , if not how to choose the most relevant ones?
3. Do certain countries produce people that are more famous and did the influence of these countries change over time?
4. Who are these influential people (characterised by their HPI)?

The 1st question is answered by splitting the data into 3 categories : Pre-Industrial Revoultion (<1712) , Industrial Revolution (1712-1942), 1943-2005 (Post-Industrial revolution) . By such categorsiation we are able to explore if these time periods had an influence on the occupations of famous people and their gender.

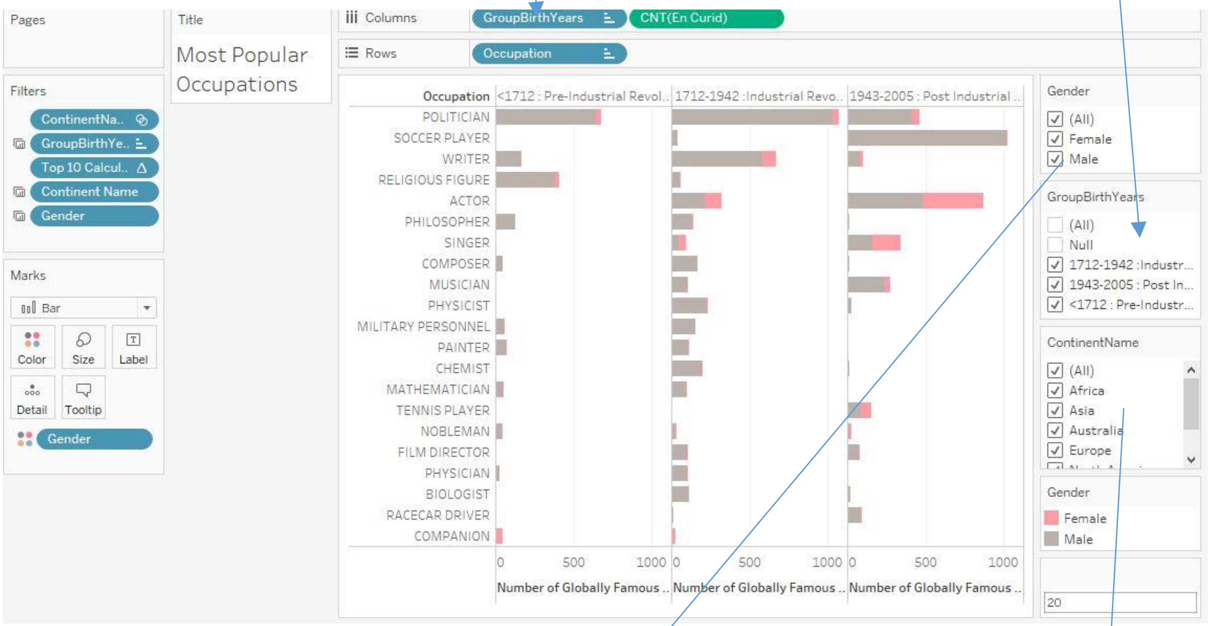
The 2nd question can be answered using Top "N" categories (or filtering on a condition), and producing visual representations according to user input.

Views 2 and 3 attempts to answer Q3 &4.

View1

Calculated Field to group different time periods, advanced users can simply change the groupings in "calculated field" to get an updated view on a different time

Filer according to times



Gender
☑ (All)
☑ Female
☑ Male

GroupBirthYears
□ (All)
□ Null
☑ 1712-1942 :Industr...
☑ 1943-2005 : Post In...
☑ <1712 : Pre-Industr...

ContinentName
☑ (All)
☑ Africa
☑ Asia
☑ Australia
☑ Europe

20

Figure 3: View 1: Top 20 occupations across the 3 -time periods discussed

Keep/ exclude Male –Female

Filter on Different continents

View1 implemented in tableau 10.2   is an adaption of the final sketch (Fig2), taking into consideration questions 1&2 discussed above.

The categorical attributes as discussed are coded using identity channels such as spatial position and colour hue, while the count of the id is using the magnitude channel (*Length of the bar*) on the x-axis.

View 2

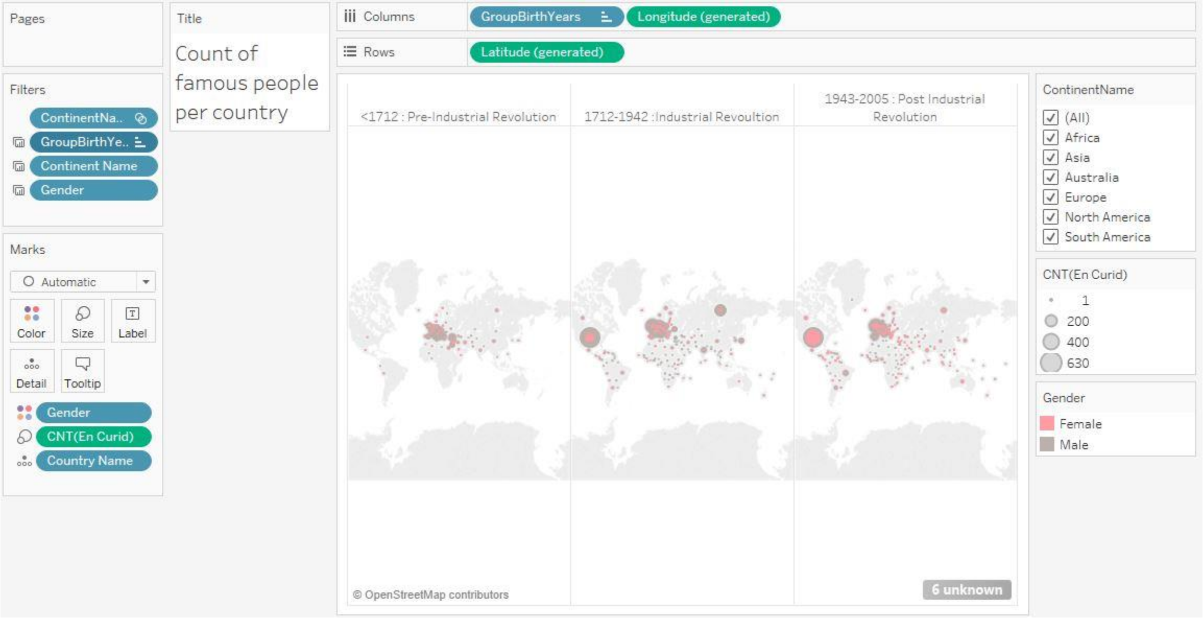
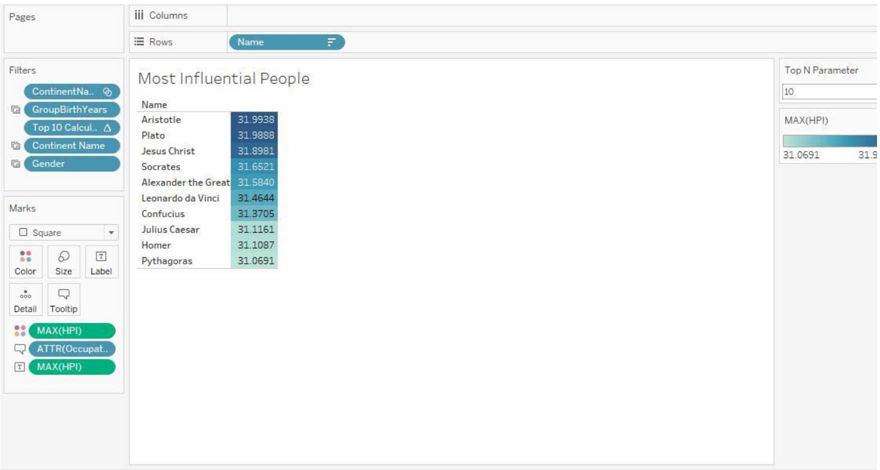


Figure 4: view 2: Counts of famous people per country

View 2, looks to explore, do certain countries produce people that are more famous and did the influence of these countries change over time. Here, the quantitative variable : count of famous people is coded using Area of the circle I.e. larger circles represent higher counts which is high on the effectiveness scale (Munzer, 2014) which is Fig 4 , shows United States producing more famous people through and post Industrial Revolution era.

View 3:

The final view shows the most influential people ranked according to their HPI .A table is chosen to show clarity of the different names as well as the use of colour saturation and position to indicate the size of HPI rating, thereby indicating the influence of the person.



Most Influential People

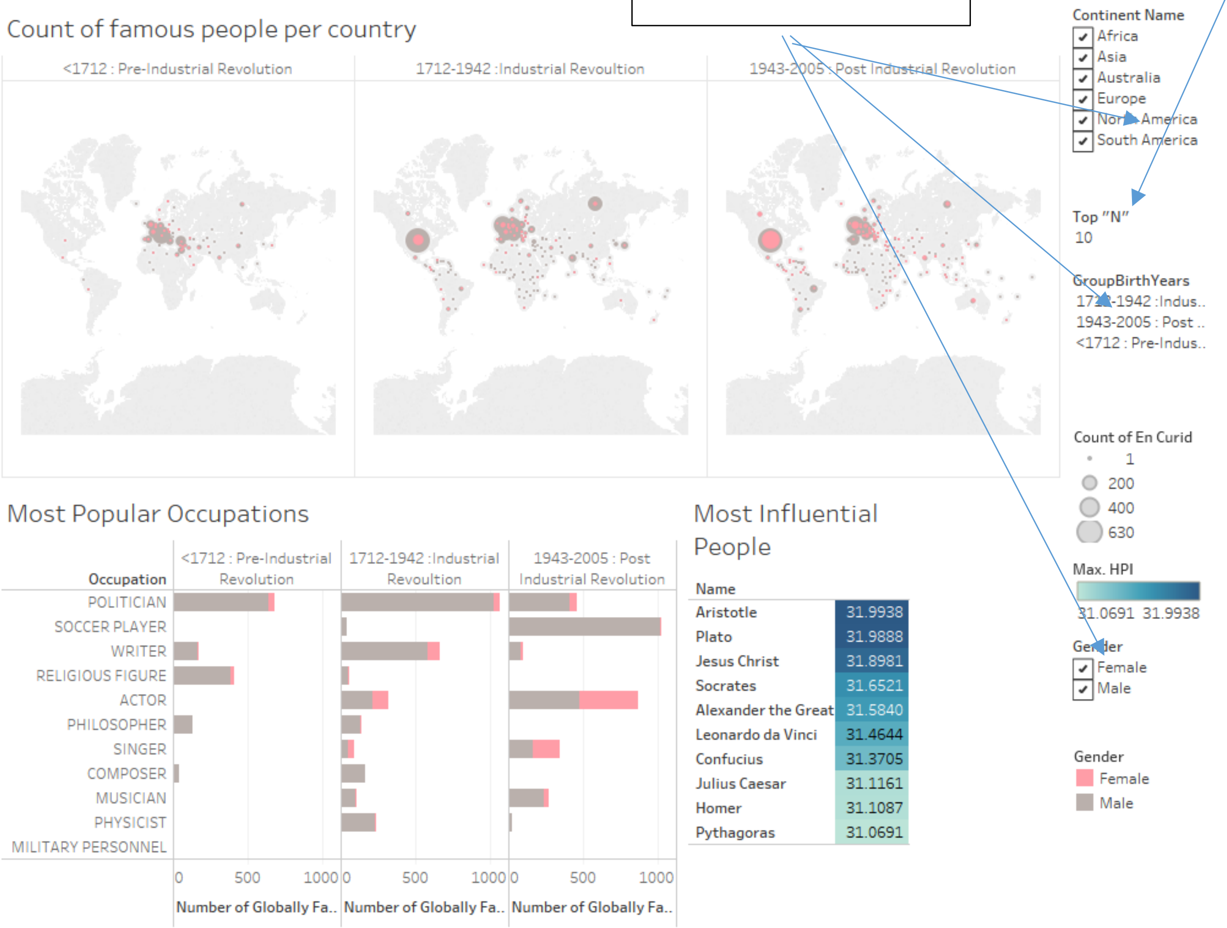
Name	
Aristotle	31.9938
Plato	31.9888
Jesus Christ	31.8981
Socrates	31.6521
Alexander the Great	31.5840
Leonardo da Vinci	31.4644
Confucius	31.3705
Julius Caesar	31.1161
Homer	31.1087
Pythagoras	31.0691

Figure 5:View3: 10 Most influential People across all eras (characterised by the HPI scores)

Allows user to choose Top "N" to show for most popular occupations and most influential peoples.

Combined Views – Final Dashboard

Filers linked between multiple sheets



Count of famous people per country

<1712 : Pre-Industrial Revolution
1712-1942 :Industrial Revoultion
1943-2005 : Post Industrial Revolution

Continent Name
☑ Africa
☑ Asia
☑ Australia
☑ Europe
☑ North America
☑ South America

Top "N"
10

GroupBirthYears
1712-1942 :Indus..
1943-2005 : Post ..
<1712 : Pre-Indus..

Count of En Curid
● 1
● 200
● 400
● 630

Max. HPI
31.0691 31.9938

Gender
☑ Female
☑ Male

Gender
🟥 Female
▨ Male

Most Popular Occupations

Occupation
<1712 : Pre-Industrial Revolution
1712-1942 :Industrial Revoultion
1943-2005 : Post Industrial Revolution
POLITICIAN
SOCCER PLAYER
WRITER
RELIGIOUS FIGURE
ACTOR
PHILOSOPHER
SINGER
COMPOSER
MUSICIAN
PHYSICIST
MILITARY PERSONNEL
0
500
1000
Number of Globally Fa..

Most Influential People

Name	
Aristotle	31.9938
Plato	31.9888
Jesus Christ	31.8981
Socrates	31.6521
Alexander the Great	31.5840
Leonardo da Vinci	31.4644
Confucius	31.3705
Julius Caesar	31.1161
Homer	31.1087
Pythagoras	31.0691

Figure 6: Final Dashboard combining multiple views

Insights from the visualisation

How has occupation and gender of globally famous people changed over time? (Fig 3)

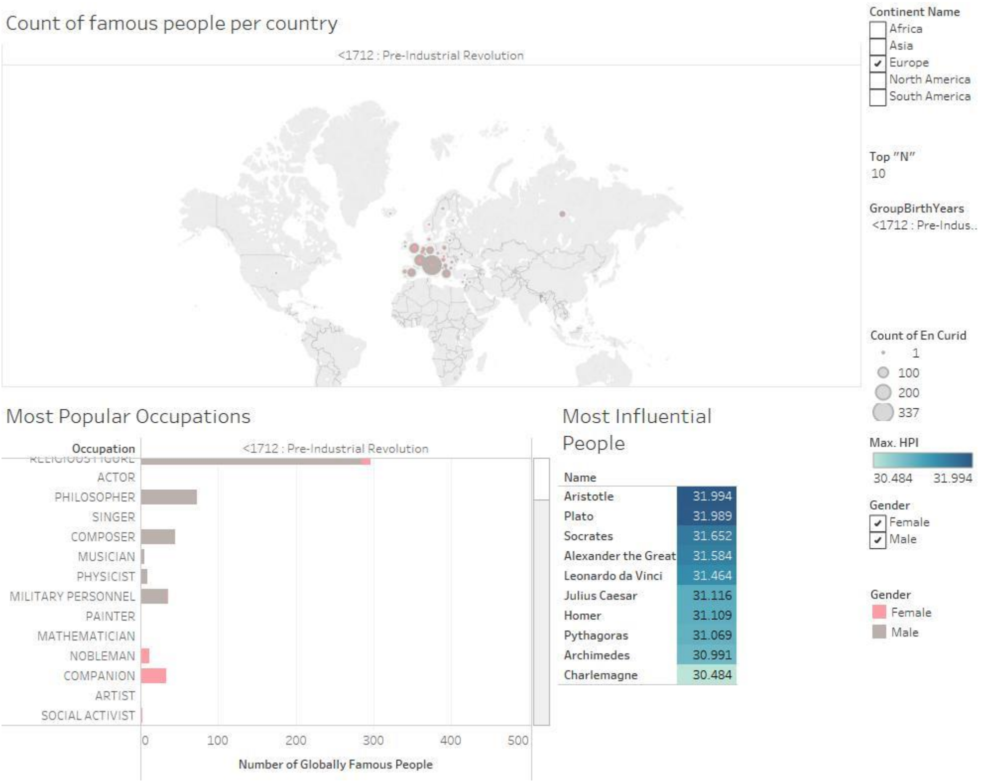
Politicians have remained almost similar across the three time periods we are interested in, but only having a slight decrease in 2005. While, there has been a large increase in soccer players, actors and singers post industrial revolution, which could be attributed to the development in TV and the internet. On the other hand, Religious figures and philosophers have reduced significantly throughout these ages.

Do certain countries produce people that are more famous   and did the influence of these countries change over time and who are these people?

Fig4 shows United States producing more influential people –post industrial Revolution, while we see the influence of Europe of producing people that are more influential has stayed roughly the same across the three time-periods (visual –check). One can check if this is the case > while filtering on Europe > highlighting the data points >shows the sum of the count for each of the three times.

Adding Granularity on Continents:

Filter on" Continent Name": Europe (tick only Europe)>filter on "Group Birth views" <1712 Pre-Industrial Revolution: The most influential people from where the Philosophers (Aristotle, Plato etc.) but the most popular occupation were politicians followed by Religious figures.



Count of famous people per country

<1712 : Pre-Industrial Revolution

Most Popular Occupations

<1712 : Pre-Industrial Revolution

Most Influential People

Name	
Aristotle	31.994
Plato	31.989
Socrates	31.652
Alexander the Great	31.584
Leonardo da Vinci	31.464
Julius Caesar	31.116
Homer	31.109
Pythagoras	31.069
Archimedes	30.991
Charlemagne	30.484

Continent Name
□ Africa
□ Asia
☑ Europe
□ North America
□ South America

Top "N"
10

GroupBirthYears
<1712 : Pre-Indus..

Count of En Curid
● 1
● 100
● 200
● 337

Max. HPI
30.484   31.994

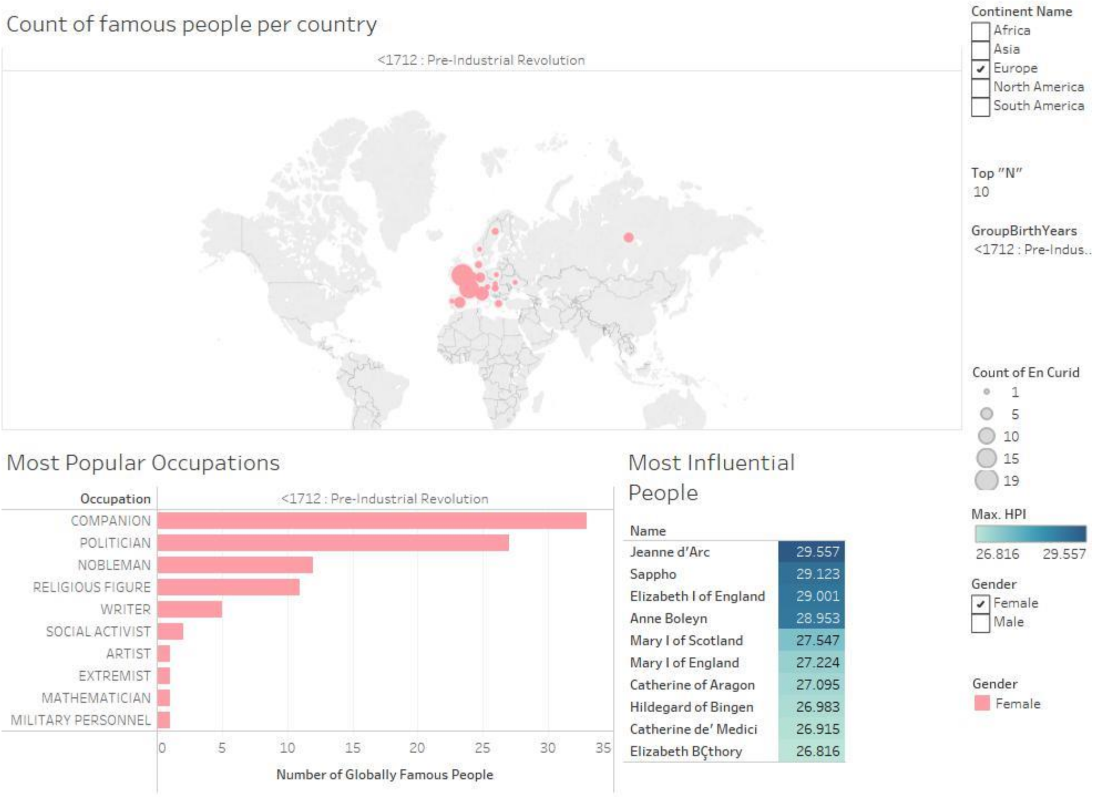
Gender
☑ Female
☑ Male

Gender
🟥 Female
▨ Male

Figure 7: Filtering on Continents (Adding granularity)

Further Filter on Gender

One can filter on to see who the popular females were and the most popular occupations:



Most Influential People

Name	
Jeanne d'Arc	29.557
Sappho	29.123
Elizabeth I of England	29.001
Anne Boleyn	28.953
Mary I of Scotland	27.547
Mary I of England	27.224
Catherine of Aragon	27.095
Hildegard of Bingen	26.983
Catherine de' Medici	26.915
Elizabeth BÇthory	26.816

Continent Name
□ Africa
□ Asia
☑ Europe
□ North America
□ South America

Top "N"
10

GroupBirthYears
<1712 : Pre-Indus..

Gender
☑ Female
□ Male

Figure 8: Filtering on Gender

Critical Discussion

Dataset errors

The dataset presented a few problems in terms of misspelled observations in names columns in particular. As an example : "AndrÇ¸ FrÇ¸dÇ¸ric Cournand" has the name replaced with certain non – English characters . In order to solve the problem, the excel file was loaded into R language, and small chunk of code, was written to work with only English alphabetic words, (included in the Appendix section), however this has not solved the problem completely as some names still contain non-alphabetic characters. For completeness, the returned dataset from R, which contains 9637 out 11341 observations, is included as mydata.csv.

Strengths and Limits of the visualisation

Strengths

- Easy visualisation of a large and complex data (no coding involved)
- Interactiivness: allows the user to filter on different attributes, and answer the question according to the user.
- Drag and Drop feature for attributes allows exploring many possible visualisations.

- Automatically updates and live connectivity with data t

Limitations

- Only able to produce visualisations according to the visualisation marks provided by the software.
- Computing time could be even slower if the dataset was larger.
- Poor manipulation or cleaning of data tools , in this case had to be done in another software
- Limits Creativity, as only can work with the combination of different attributes and marks available.

Conclusion and Future Work

In conclusion, Tableau allows quick and easy visualisation to explore large and complex datasets, all the while allowing user interactivity and live filtering of the data.

The dashboards created can be improved in several ways:

- Adding a both " Top" and "bottom "filter to the dashboard
- Instead of removing non-alphabetic rows, replacing them by scraping the data from the Wikipedia page (matching on their En curid).
- Dealing with errors in Misplaced geographic locations by correctly re-assignment
- Exploring Different eras of history: Pre-History (), Middle Ages, Early Modern Period, Modern era and subdividing them further.

References

Munzer, T. (2014). *Visualization Analysis & Design.* CRC Press.

Appendix

R-code for removing non-alphabetic characters:

```
data<-read.csv("C:/Users/robin/Dropbox/visualization/practical 2/famousPeople.csv")


names<-as.character(data$name)

summary(names)


x<-names

nameswithoutspaces<-gsub(" ","",x,fixed = TRUE)

index<-grep("^[[:alpha:]]*$", nameswithoutspaces)


allindex<-1:length(data$name)

notinindex<-!(index%in%allindex)

newdata<-data[index,]


write.csv(newdata, file = "MyData.csv")
```

Code for Tableau calculations:

For Top “N”:

```
if [Index]<=[Top N Parameter] THEN "Top N" ELSE IF [Index]>=SIZE()-[Top N
Parameter] THEN "Bottom n" end end
```

For Grouping Birth Years:

```
IF[Birthyear]>1942 AND [Birthyear]<=2005 THEN "1943-2005 : Post Industrial
Revolution"
ELSEIF [Birthyear]<=1942 AND [Birthyear]>=1712 THEN "1712-1942
:Industrial Revoultion"
ELSEIF [Birthyear]<1712 THEN "<1712 : Pre-Industrial Revolution "
END
```