San José State University

Department of Applied Data Science

DATA 230 Data Visualization
Instructor: Guannan Liu

Group Project

Historical Figures Decoded: A Data-Driven Exploration of Influence and Popularity

Group 9

Group Members:

Tanmay Singh - 016944265

Naga Shreya Chilumukuru - 017475315

Sowmya Neela - 017418219

Sarah Elsa Mathew - 017431700

Contents

Abstract

In this project, our objective is to conduct an in-depth analysis of a dataset containing information about Historical Figures. We will focus on crucial attributes such as their birthplace, occupation, age, place of death, the number of citations in various languages, gender, and their Historical Popularity Index. Our goal is to predict how these key features influence the Historical Popularity Index using different regression algorithms. We will then compare the performance of these machine learning models and determine which one provides the most accurate predictions for the data. Additionally, we plan to create interactive dashboards that will enable us to gain a deeper understanding of our dataset and extract valuable insights from it.

Below is the dataset that is used for this purpose:

https://pantheon.world/explore/rankings?show=people&years=-3501

Introduction

This project, "Historical Figures Decoded: A Data-Driven Exploration of Influence and Popularity" embarks on a detailed examination to demystify the enduring impact and acclaim of personalities throughout history. Leveraging a robust dataset rich with details of notable individuals, this aims to decode the many factors that contribute to a person's lasting fame and historical influence.

Combining vast historical records with modern data analysis tools, this project intends to shed light on the intricate web of historical acclaim and the enduring legacies of these personalities. Along with this, an additional goal is to extract meaningful patterns from visualizations about the historical figures and their impressive recognition.

Dataset

Our research utilizes the expansive Pantheon dataset, which is a collection of 88,860 biographies sourced from more than 25 language editions of Wikipedia. This collection goes beyond mere biographical entries; it precisely aligns each figure with a distinct cultural domain, place of origin, and time period, thus offering a detailed lens for understanding the complex narrative of historical cultural achievements.

Drawn from the collaborative and diverse databases of Freebase and Wikipedia, this dataset presents us with globally celebrated figures, reflecting a broad array of cultural outputs. It includes intricate demographic details, links to Wikipedia IDs, and connections to Wikipedia in various languages. It is further enriched by detailed monthly page view statistics, painting a

picture of the public's interest from January 2008 to December 2013 across all language editions of Wikipedia.

The dataset's integrity is bolstered by the corroboration of its indicators, especially the Historical Popularity Index (HPI), with independent standards of success in different cultural areas, sports included.

Limitations of Dataset

•Use of Wikipedia as a Data Source.

The dataset reflects the biases of Wikipedia's contributors, who predominantly speak English, hail from Western countries, are male, and often have a higher education and technical background. By including data from Wikipedia in various languages, the dataset mitigates the English-centric view, offering a more global perspective. For instance, the presence of O.J. Simpson as the sole American Football Player underscores the broader fame that surpasses his sporting achievements, highlighting the reduced English-language bias.

•Assignment of Locations Based on Place of Birth.

Geographic locations in the dataset are based on birthplaces according to current political borders, with location data standardized through geocoding APIs and manual checks. However, this approach doesn't account for individuals who gained worldwide recognition after moving countries. The birthplace-centric data might not reflect the true impact of migratory athletes or artists who found fame far from their birthplaces, which can only be gleaned from detailed historical accounts.

•Use of Biographies as Proxies for Historical Information.

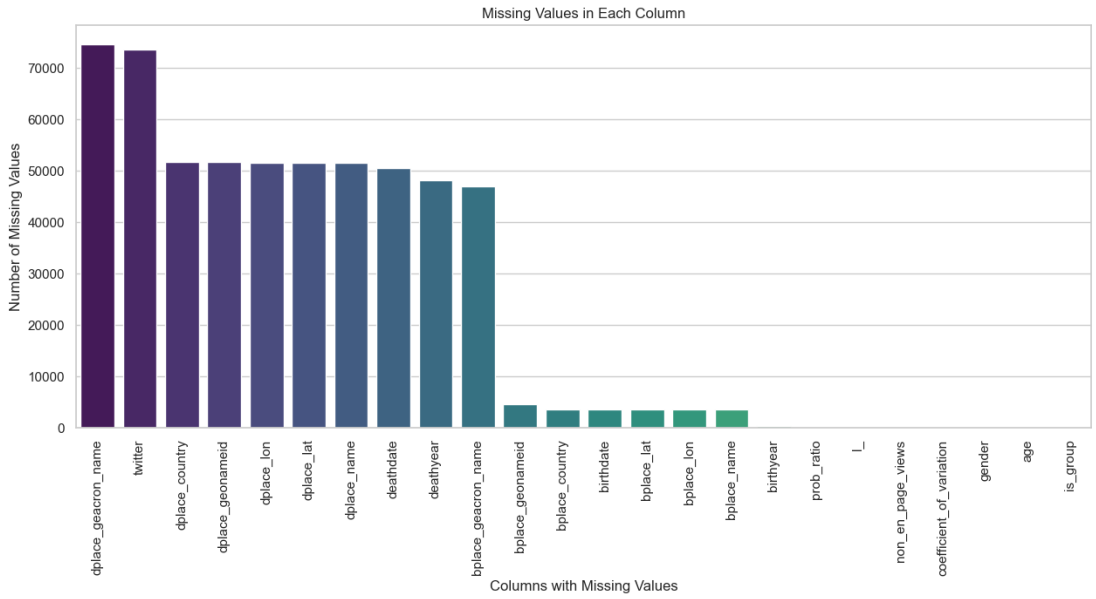
Biographies serve as a means to gather historical data, linking individuals to linguistic groups, geographic regions, occupations, and historical eras. They provide a wide historical canvas, not only including individuals with significant works or creations but also those who have been documented for their role in notable historical events.

•Other Technical Limitations.

The dataset is subject to the fluctuating nature of Wikipedia and other online sources, leading to reproducibility issues. Tools used in the dataset's creation, such as the Yahoo Placemaker API for geocoding, have been discontinued. Additionally, the pending retirement of Freebase and its uncertain data transfer to Wikidata raises questions about the future accessibility of some of the dataset's information sources.

Dataprocess :

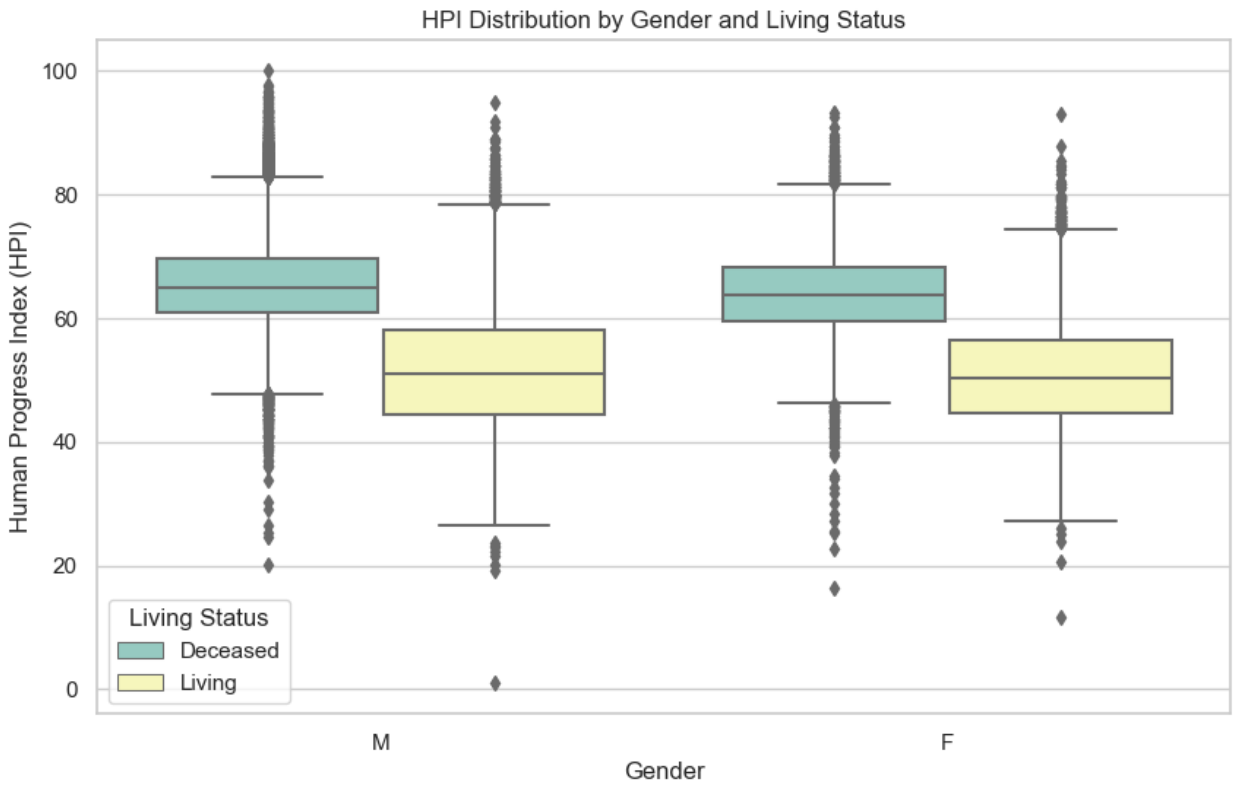
Exploratory Data Analysis



The bar graph highlights the prevalence of missing data across different fields within the dataset. Here's a summary of the findings:

- A noticeable number of entries are incomplete in columns related to birth and death locations, suggesting a lack of recorded information for many individuals' birthplaces and death details.
- The `deathyear` column also shows many missing entries, implying that for a significant group of individuals, the death year hasn't been captured or isn't relevant, which might be due to them being alive or due to outdated records.
- A lack of data in the `twitter` and `bplace_geacron_name` columns indicates that not every individual has an associated Twitter account or that the historical name of their birthplace isn't always noted.
- Essential attributes like `age`, `gender`, and `is_group` are largely complete, indicating reliable documentation for these details in the dataset.
- Missing figures in the `non_en_page_views` and `coefficient_of_variation` columns could potentially impact any analysis centered on online visibility metrics or their consistency.

In essence, the visualization points to significant informational voids, particularly in areas detailing personal histories and digital footprints. Addressing these missing elements is crucial for comprehensive analysis.



The boxplot graphically represents the Human Progress Index (HPI) distributions across genders, differentiated by living status, using the provided dataset. Here are the key observations:

1. Gender Distribution: The median HPI is quite similar for both genders. The female distribution is slightly more concentrated around the median as evidenced by a tighter interquartile range (IQR), suggesting less variability in HPI scores for females compared to males.

2. Impact of Living Status: The median HPI for individuals who are still living is lower than for those who are deceased across both genders. This might indicate that figures from the past, who have all passed away, generally have higher HPIs, likely because their legacy has had more time to establish.

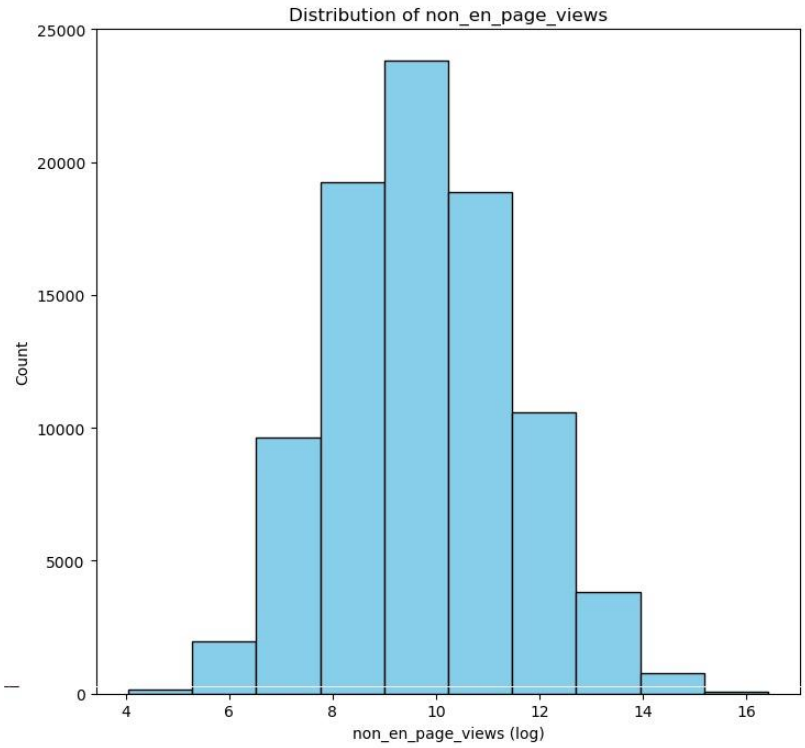
3. Presence of Outliers: There's a significant presence of outliers in all categories, particularly among deceased males. These outliers are individuals whose HPI scores are exceptionally higher or lower than the majority.

4. Variability in Scores: Deceased individuals show a wider range of HPI scores, with deceased males exhibiting the most variability. This could reflect the diverse levels of historical recognition these individuals have received.

5. Historical Perspective: The elevated HPIs seen among the deceased could be indicative of a historical perspective bias, where historical figures, by virtue of their lasting impact and extensive documentation, tend to receive higher recognition.

In relation to the dataset, these observations imply that the HPI is influenced by not just the individual's contributions or societal impact, but also by the duration and documentation of their

legacy, with historical personalities often having a more substantial influence on their HPI scores.
This analysis offers valuable comparative insights into the variation of HPI across gender lines and living status. It underscores the importance of considering how time and historical documentation can affect the HPI, particularly for those who have had longer to cement their legacies.



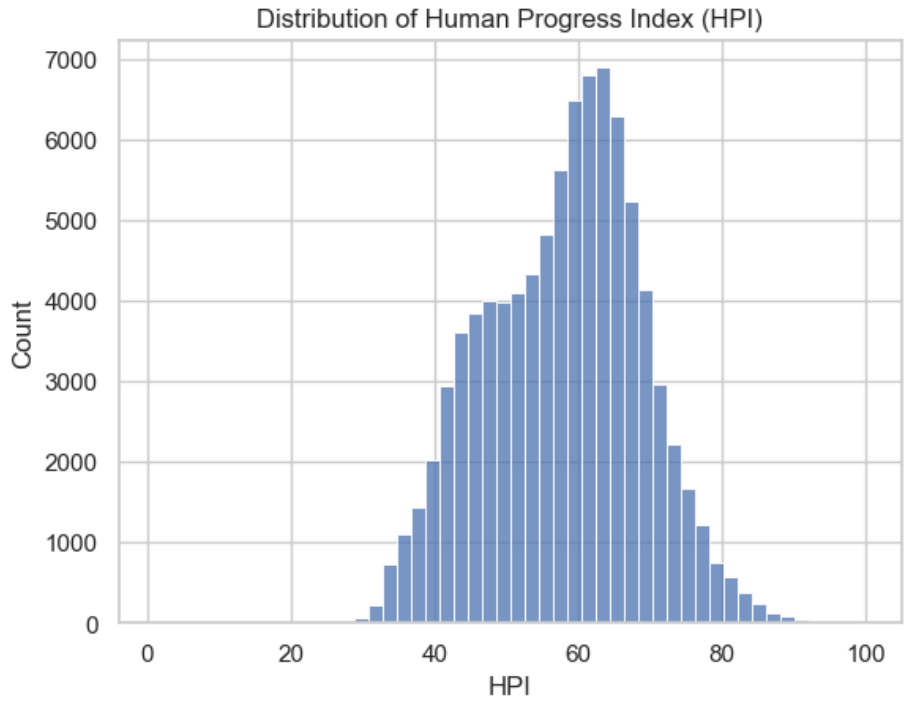
The histogram based on the dataset indicates the spread of non-English page views for the individuals profiled. Here's a connection between the visualization and the dataset:

1. Range of Visibility: The dataset spans a wide array of personalities, which is echoed in the diverse number of page views. Most individuals have relatively few non-English page views, as shown by the dense clustering at the lower end of the histogram.

2. Global Recognition: A handful of individuals stand out with exceptionally high page views. This likely points to their widespread recognition or influence across non-English speaking regions.
3. Distribution Characteristics: The distribution's skew to the right suggests that a small fraction of the dataset's individuals have a disproportionate impact on the higher end of the page view spectrum.

4. Noteworthy Outliers: The far-right outliers could be individuals of significant historical importance or current global fame, who naturally attract more international page views.

5. Insights from the Dataset: The dataset's rich information on individuals' lives and achievements is mirrored in the page views data. Non-English page views serve as a proxy for international interest and can highlight the global footprint of these individuals.
In essence, the non-English page view distribution provides insight into the international recognition of the individuals in the dataset. Most have limited international exposure, with a select few reaching high levels of global prominence.



The histogram illustrates the Human Progress Index (HPI) scores distribution from the dataset. Key observations include:

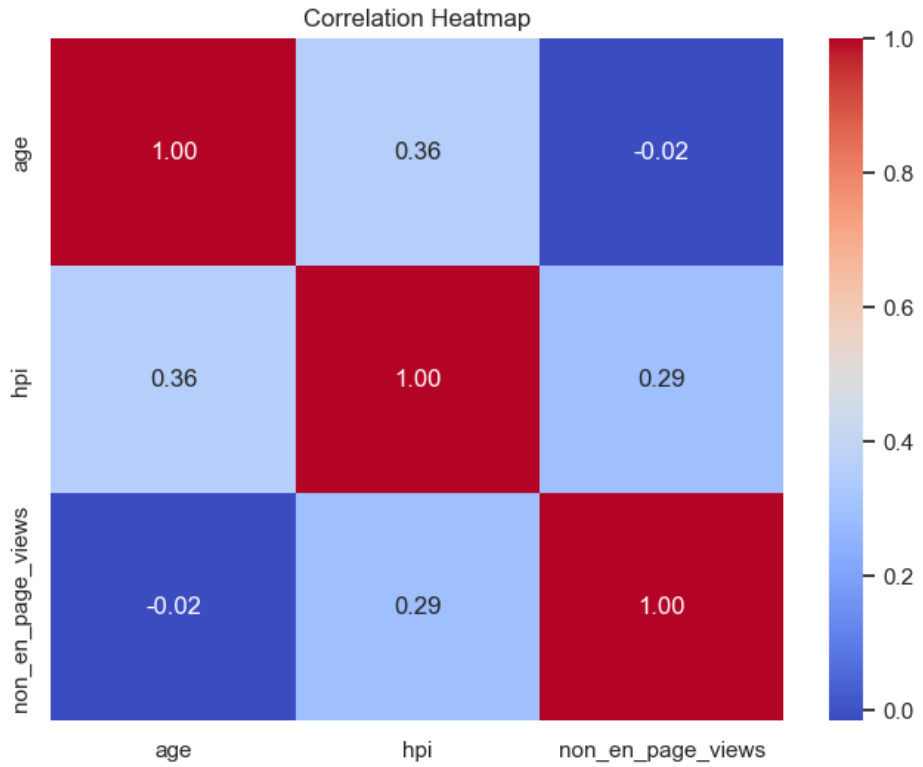
1. Concentration of HPI Scores: The HPI scores cluster around a median value, with most scores falling in the 50-60 range. This indicates that the dataset's individuals generally share a similar level of historical significance.

2. Score Range: The HPI scores show a fair amount of spread, suggesting diverse levels of historical recognition among the individuals in the dataset, although within a common range.

3. Lack of Extreme Scores: The absence of significant outliers on either end of the HPI spectrum implies a dataset mostly devoid of individuals with exceptionally high or low levels of historical significance.

4. Distribution Shape: The bell-shaped curve of the distribution suggests normality, aligning with what might be anticipated from large datasets, as per the central limit theorem.

Connecting these points to the dataset indicates a spectrum of historical significance among the individuals, predominantly moderate, without extreme deviations. This overview is beneficial for those analyzing the dataset for historical impact trends, providing an understanding of how recognition is distributed among the people featured.



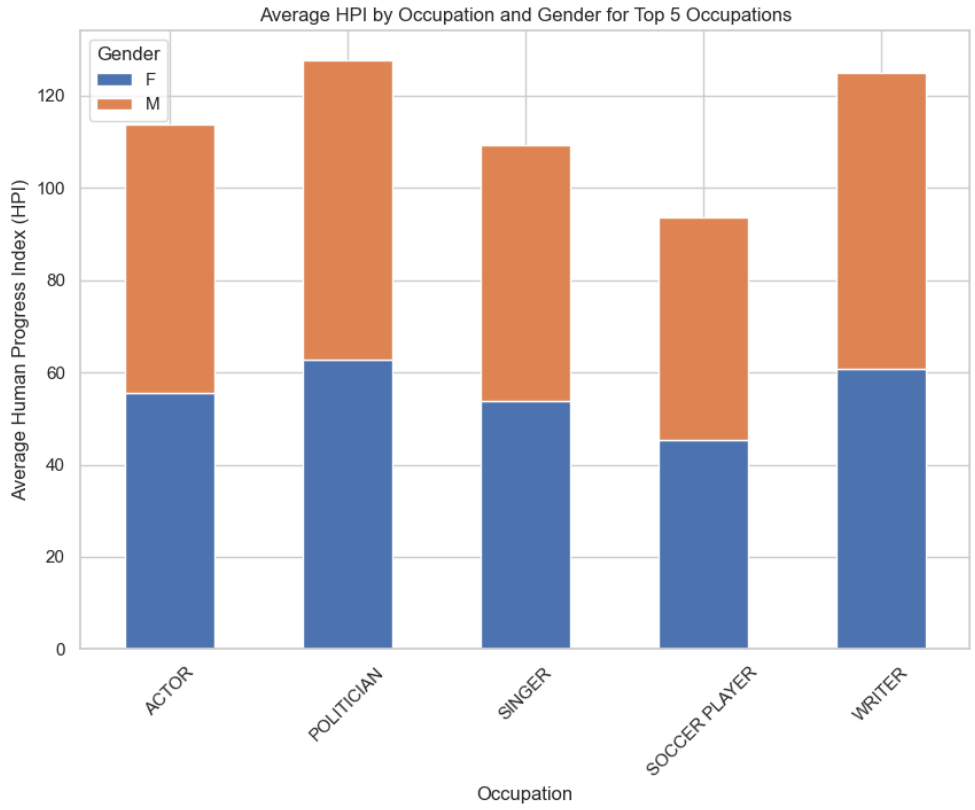
The heatmap provides an analysis of the relationships among age, Human Progress Index (HPI), and non-English page views within the dataset. Here's a summarized interpretation:

1. Link Between Age and HPI: There is a moderately positive correlation (0.36) between an individual's age and their HPI. This could indicate that individuals tend to achieve a higher HPI as they age, possibly due to accumulating notable works, recognition, or lifetime achievements.

2. Age Versus Non-English Page Views: A negligible negative correlation (-0.02) exists between age and non-English page views, suggesting no significant direct link between how old an individual is and the volume of non-English page views they attract.

3. HPI's Relationship with Non-English Page Views: A slight positive correlation (0.29) is observed between HPI and non-English page views, hinting that those with higher HPI scores may also enjoy a modestly increased presence in non-English online viewership, which could signal a broader international influence or recognition.

The heatmap serves as an efficient method to discern the magnitude and direction of variable interconnections. The findings indicate that age has some bearing on HPI, but its influence on non-English online presence is minimal. Conversely, the HPI's association with non-English page views, though not strong, could be attributed to factors like the extent of a person's impact or their global notoriety. The low correlation between non-English page views and both age and HPI suggests that other elements, possibly outside the dataset, may play a more significant role in affecting an individual's online popularity across different languages.



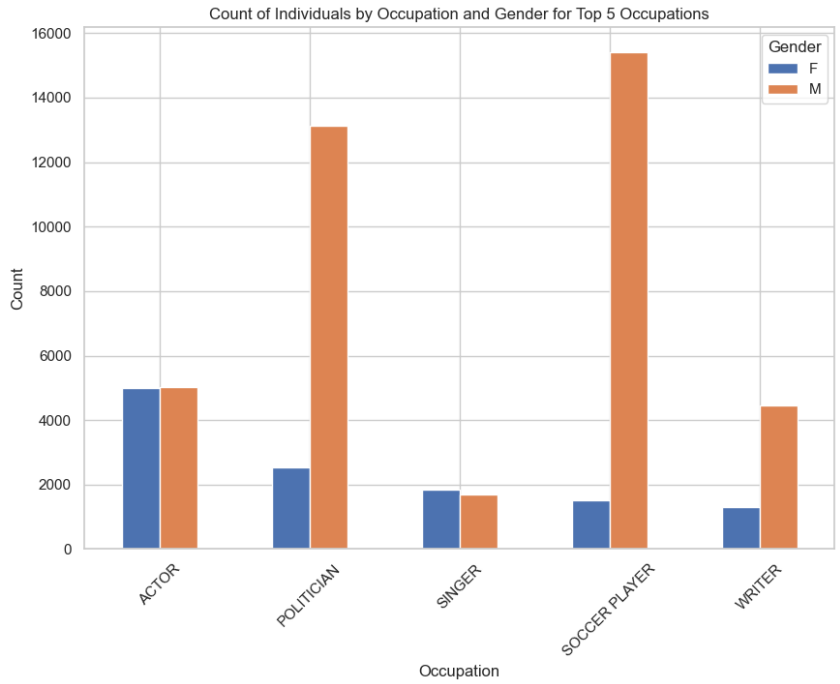
The stacked bar chart demonstrates the mean Human Progress Index (HPI) across five prominent occupations, segmented by gender, according to the dataset. Here's an analysis of the chart:

1. Gender Distribution: Each occupation is broken down into segments that show the average HPI for men and women. The representation is mixed, with men generally showing a higher average HPI in these roles.

2. HPI by Profession: Actors and writers stand out with elevated average HPIs, suggesting that these careers might be more historically or socially influential, as per the data collected.

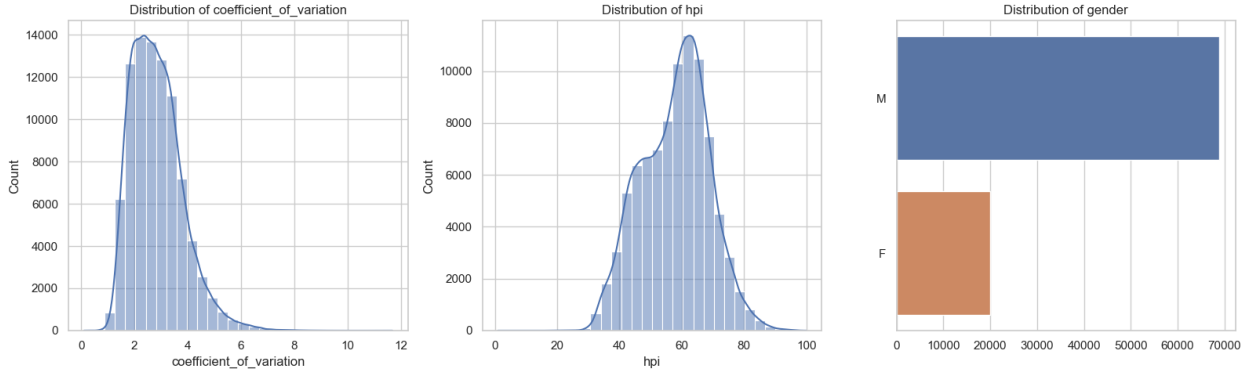
3. Gender Comparison: When comparing genders within each occupation, it is evident that men have a generally higher HPI, possibly reflecting historical biases or greater societal recognition of men's roles.

4. Cultural and Historical Context: The variation in HPI between genders across different occupations could be indicative of long-standing societal norms and the ways in which historical achievements have been recorded, often amplifying men's contributions in these fields.
The chart provides a basis for more in-depth investigations into the gender-based differences in HPI and how societal roles and historical recognition are portrayed in the dataset.



The side-by-side bar chart displays the number of people across the five most common occupations in the dataset, categorized by gender. Here's a breakdown of the observations:

- Prevalence of Males: Across all the leading occupations featured in the dataset, male counts exceed female counts, highlighting a gender skew in the dataset's documented professions.
- Occupational Presence: Occupations such as 'Politician' and 'Singer' are highly populated, with 'Politician' having the largest gender gap.
- Balanced Representation: Among the occupations, 'Actor' has a more balanced gender ratio, yet it still displays a tendency towards a higher number of males.
This graphical comparison points to a potential need for a gender-focused analysis of the dataset and hints at possible historical or societal influences that have led to the predominance of males in these occupational categories. It serves as a basis for delving deeper into the root causes of gender imbalance observed in the professions documented.



1. Coefficient of Variation Distribution: This histogram illustrates that the coefficient of variation generally clusters at lower figures, peaking at approximately 2.5. This pattern suggests that page view variability among most of the individuals in the dataset does not exhibit wide disparities, indicating more uniformity rather than erratic variations.

2. HPI Distribution: The graph of the Human Progress Index shows a generally normal distribution with a slight lean towards higher values. The bulk of individuals have HPI values in the moderate range of 50-60, implying a standard level of historical recognition. Nonetheless, a select few have significantly higher HPIs, pointing to a notable historical presence.

3. Gender Distribution: The bar graph displays a pronounced gender gap within the data, with male individuals markedly outnumbering females. This could be indicative of historical and societal tendencies to highlight male achievements over female ones, particularly in historical narratives and records.

In relation to the dataset, these visualizations collectively furnish insights into the distribution of page view consistency, historical prominence as gauged by HPI, and the distribution of gender across the documented individuals. These insights suggest areas for subsequent detailed exploration, such as the factors influencing page view consistency or the historical reasons for the observed gender disparity.

Data Cleaning

1. Preprocessing

• Checking for duplicates - There can be two individuals with the same name. In cases as such, their ID would be different. This way we remove the duplicates from the dataset.

• Some columns have most of the values as null eg Twitter id. For visualization, we observed null values in the Occupation column, which were filtered out.
• Columns that were dependent on other columns were removed for ML analysis but were kept for data visualizations.

```
cnt_nulls = pantheon.isnull().sum()
cnt_nulls
```

[66] ✓ 0.1s

```
id                      0
wd_id                   0
wp_id                   0
slug                    0
name                    0
occupation              0
prob_ratio             59
gender                 16
twitter             73587
alive                   0
l                       0
hpi_raw                 0
bplace_name          3503
bplace_lat           3506
bplace_lon           3506
bplace_geonameid     4641
bplace_country       3648
birthdate            3636
birthyear             430
dplace_name         51561
dplace_lat          51567
dplace_lon          51567
dplace_geonameid    51672
```

These were the columns that were dropped.

```
col_drop = ['id' , 'wd_id' , 'wp_id','slug','name','twitter', 'hpi_raw',
            'bplace_name' ,  'bplace_lat'  , 'bplace_lon',  'bplace_geonameid' ,  'birthdate', 'l',
            'is_group','bplace_geacron_name','dplace_geacron_name',
            'bplace_name','dplace_name','dplace_lat','dplace_lon','dplace_geonameid','deathdate','deathyear','dplace_country']
```

✓ 0.0s Python

```
pantheon = pantheon.drop(col_drop, axis = 1)
```

Then we divided our dataset into features and target and changed the datatype of columns that were marked as mixed.

```
hpi = pantheon['hpi']
features = pantheon.drop('hpi', axis = 1)
```
✓ 0.0s

```
features['occupation'] = features['occupation'].astype('string')
features['alive'] = features['alive'].astype('int')
```
✓ 0.0s

2. Scaling

All the numeric columns need to be standardized so that columns with larger values don't get higher importance in our model, so all numeric columns are converted between 0 to 1.

```
numberic_column =  features.select_dtypes(include=np.number).columns
numberic_column
```
[25] ✓ 0.0s

```
Index(['prob_ratio', 'alive', 'birthyear', 'l_', 'age', 'non_en_page_views',
       'coefficient_of_variation'],
      dtype='object')
```

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
features = pd.DataFrame(data = features)
features[numberic_column] = scaler.fit_transform(features[numberic_column])
features
```

3. One-Hot Encoding

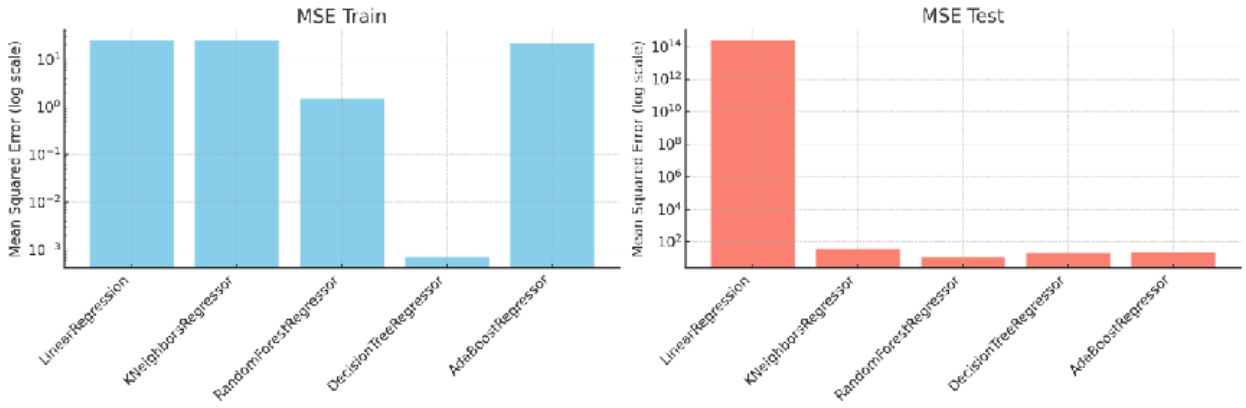
For non-numeric fields, we employ one-hot encoding to transform categorical data into a format that the model can comprehend to enhance the performance of future model training.

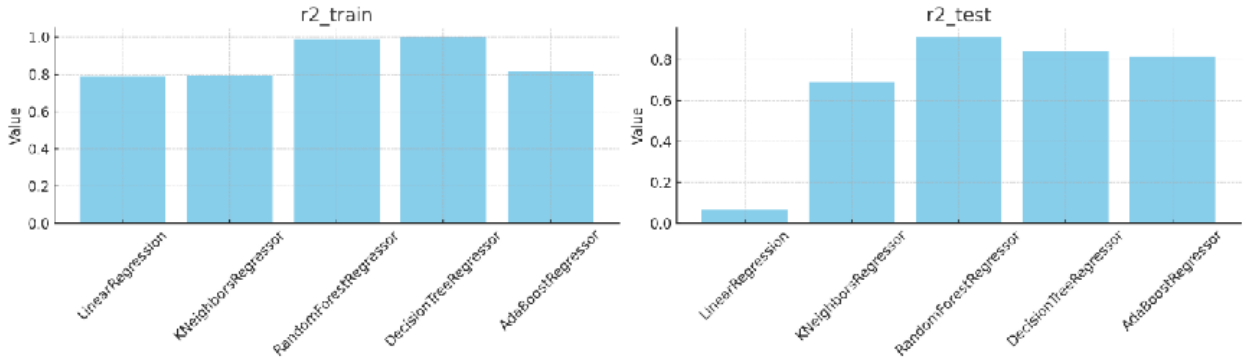
```
    #one hot endcoding
    features = pd.get_dummies(features)
    features = pd.get_dummies(features,columns=['occupation'])
    features['alive'].replace({'False': 0 , 'True' : 1}, inplace = True)
    features
[ ]
```

Proposed method

Our topic can be concluded as a regression problem. The target variable ranges from 0 to 100. We tried different regression models to see which models performed best with our dataset.

- Linear Regression
- K-Nearest Neighbors Regressor
- Random Forest Regressor
- Decision Tree Regressor
- AdaBoost Regressor





The comparison plots provide a nuanced look at the efficacy of various regression models, evaluated using mean squared error (MSE) and R-squared (R^2) metrics for both training and testing datasets. Here's a rephrased summary:

- Training Data MSE (mse_train): Excluding the DecisionTreeRegressor and RandomForestRegressor, which have notably distinct scores, the remaining models show comparable MSE values for training data. The DecisionTreeRegressor's score is exceptionally low, suggesting a potential overfit.

- Testing Data MSE (mse_test): The RandomForestRegressor outperforms other models with the lowest MSE on testing data, indicating robustness to new data. In contrast, the linear regression model's excessively high MSE suggests underperformance or possible issues in model training or data handling.

- Training Data R^2 (r2_train): The DecisionTreeRegressor achieves nearly perfect R^2 scores, closely followed by the RandomForestRegressor, both indicating a strong fit to the training data. However, this could raise concerns about overfitting, especially for the DecisionTreeRegressor.

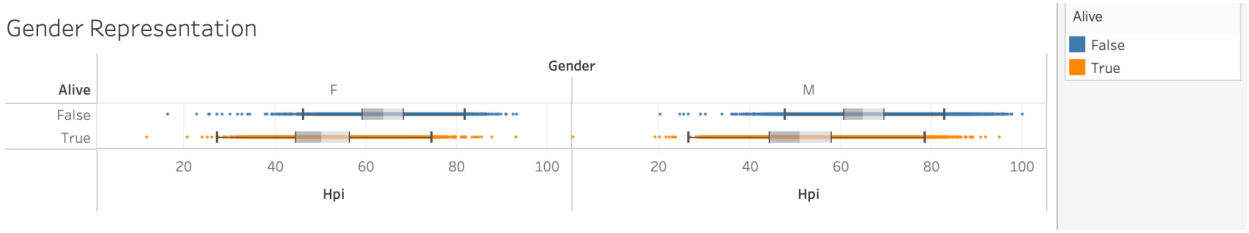
- Testing Data R^2 (r2_test): Again, the RandomForestRegressor leads with the highest R^2 score on testing data, signifying accurate predictions. The notably low and negative R^2 score for the LinearRegression model implies ineffective predictions, possibly worse than a basic mean-based forecast.

Overall, the RandomForestRegressor emerges as the most balanced model, with the right mix of complexity and predictability, avoiding the overfitting seen in the DecisionTreeRegressor. The KNeighborsRegressor and AdaBoostRegressor demonstrate moderate but consistent performance. The linear regression model's poor results suggest it may not be the right choice for this dataset, warranting a review of model assumptions or data preprocessing steps.

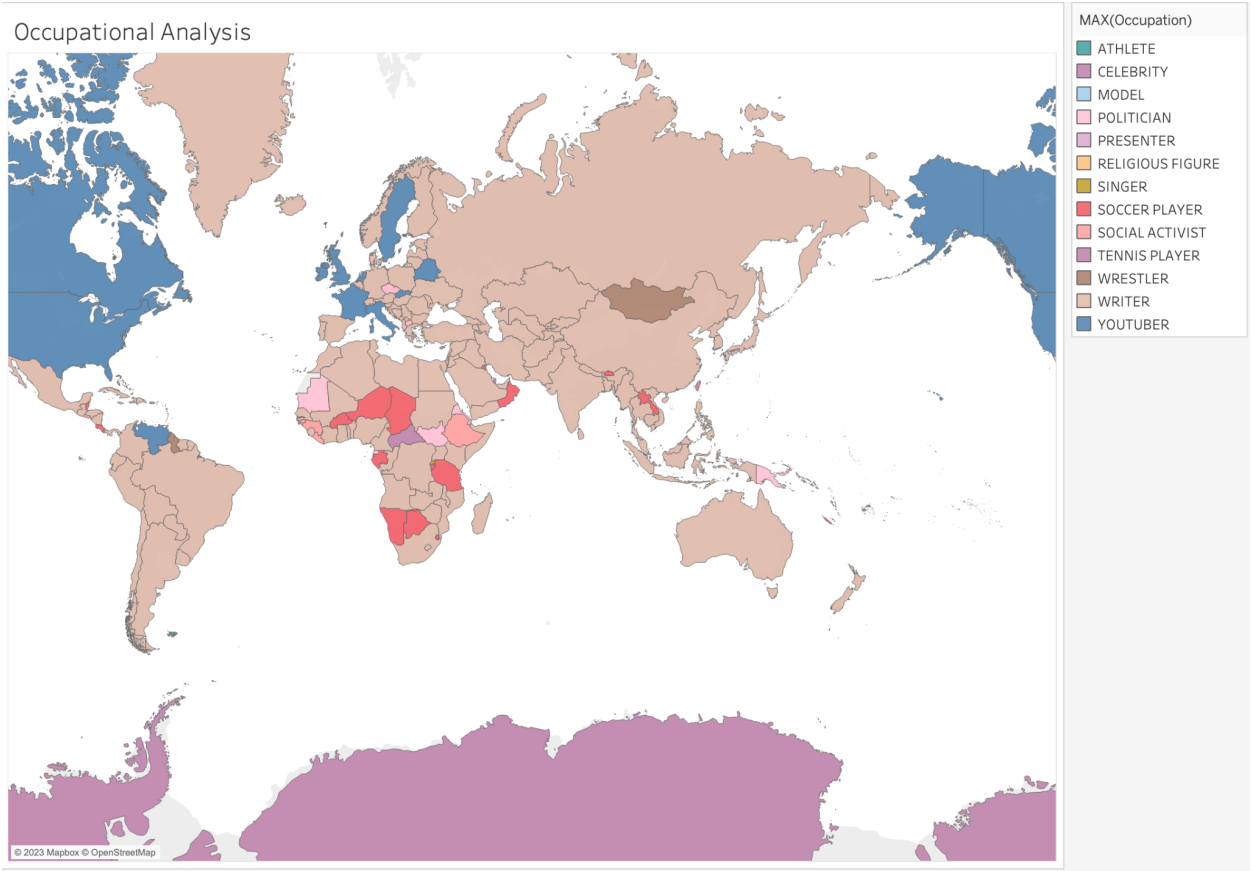
Visualization

Visualizations were done using Tableau to find insights into the dataset and understand the distribution of various factors like occupation, age and location.

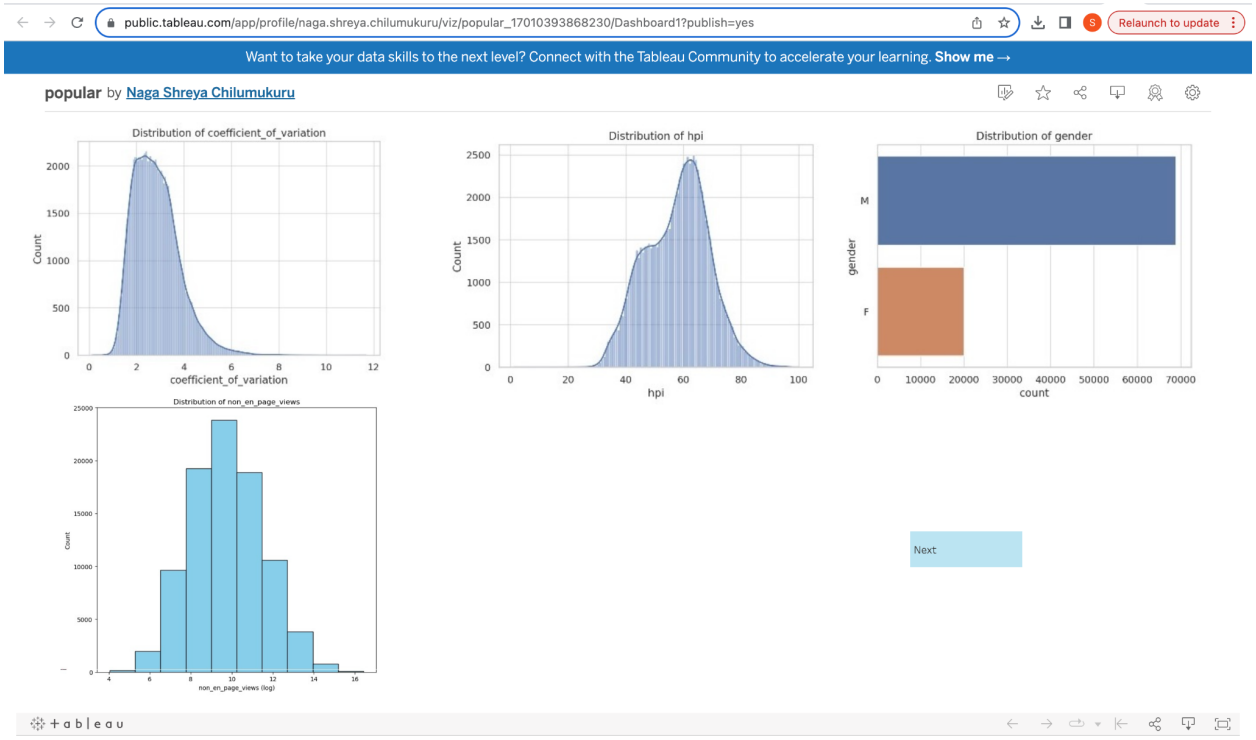
The Gender representation is mentioned in form of box plot taking the factors HPI and Alive (if the individual is alive mentioned as 'True' and if the individual is no more, then mentioned as 'False')



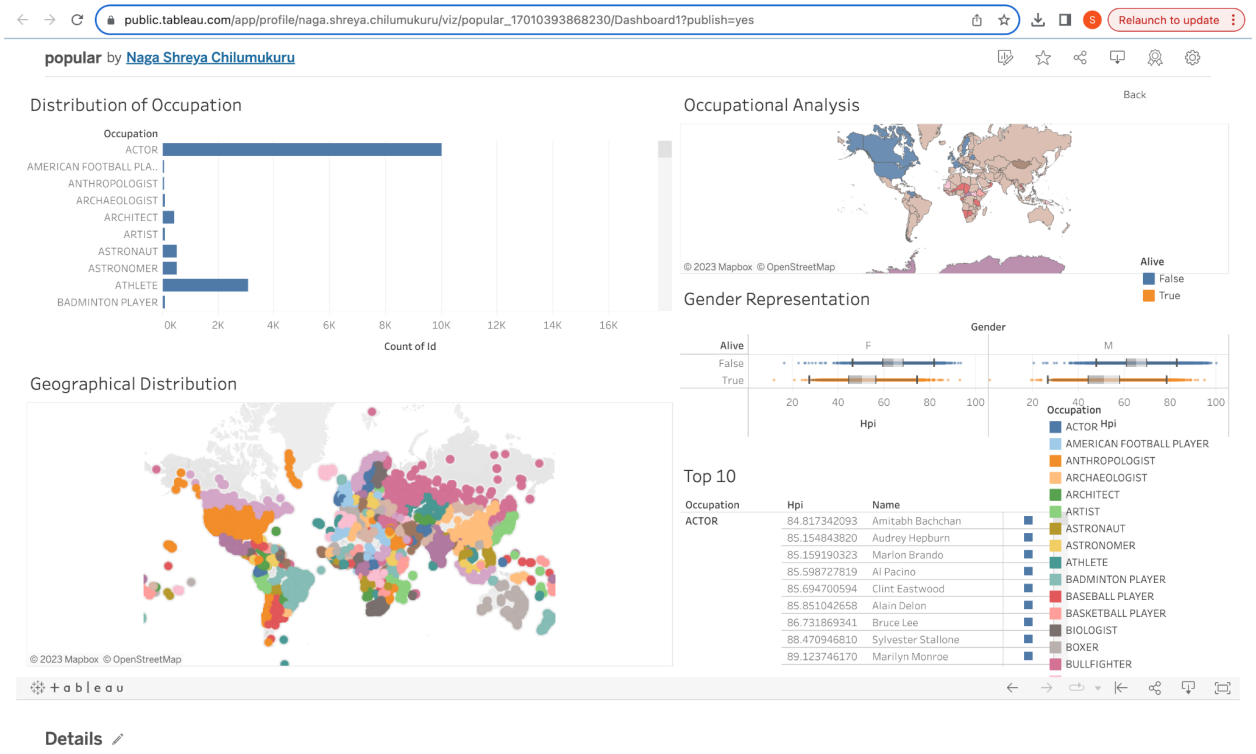
This represents the 'Occupation Analysis'. The various Occupations across the countries. Depending on the country, the occupations are color coded.



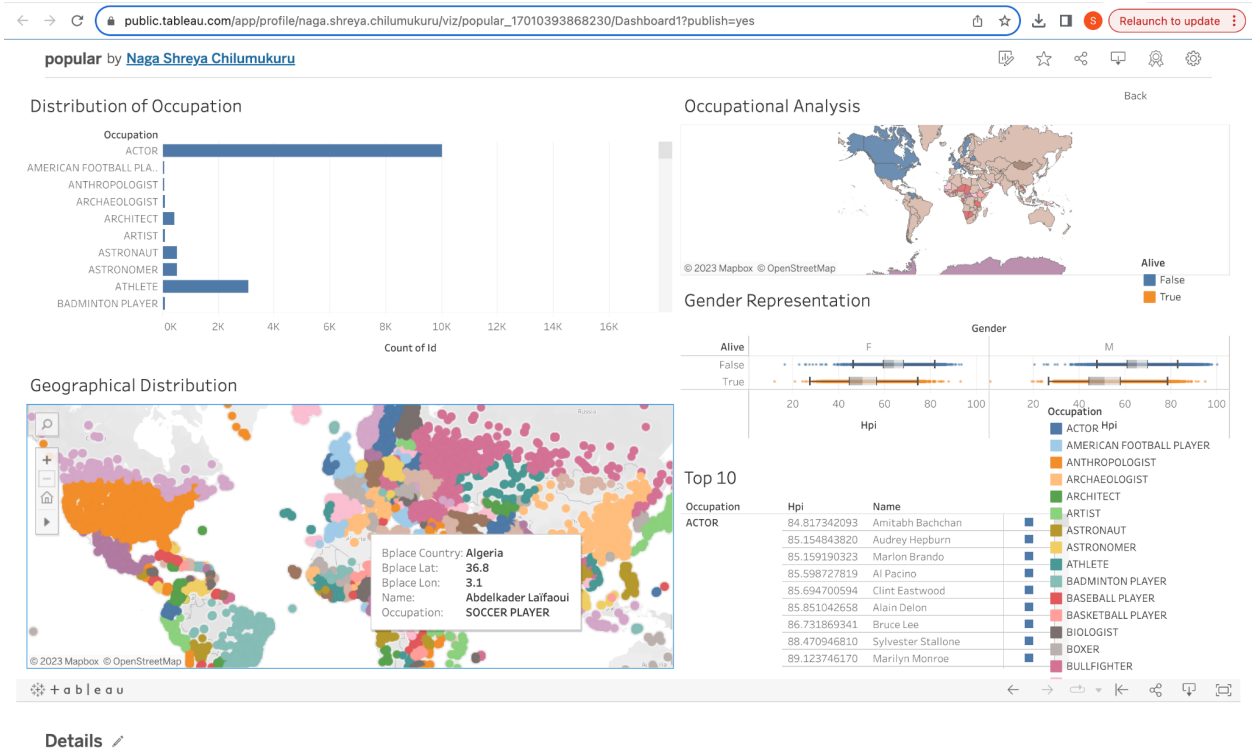
This dashboard shows the distribution between the variability of number of page views of the various Wikipedia editions, the histogram distribution of HPI with count, the count plot of gender of the dataset the and histogram distribution of non-english page views across the count.



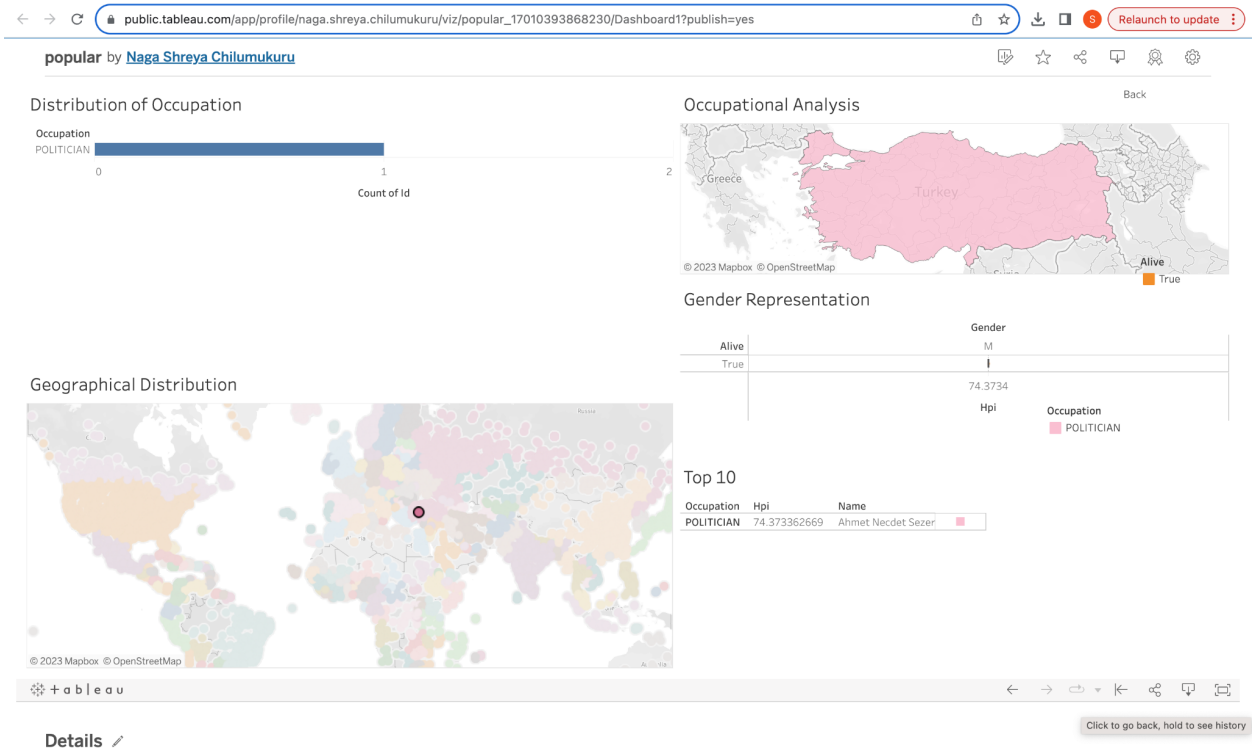
This is the dashboard that shows the Distribution of Occupations across the Geography, along with the gender distribution of the selected occupation. The Top 10 of the occupation is displayed, taken according to HPI.



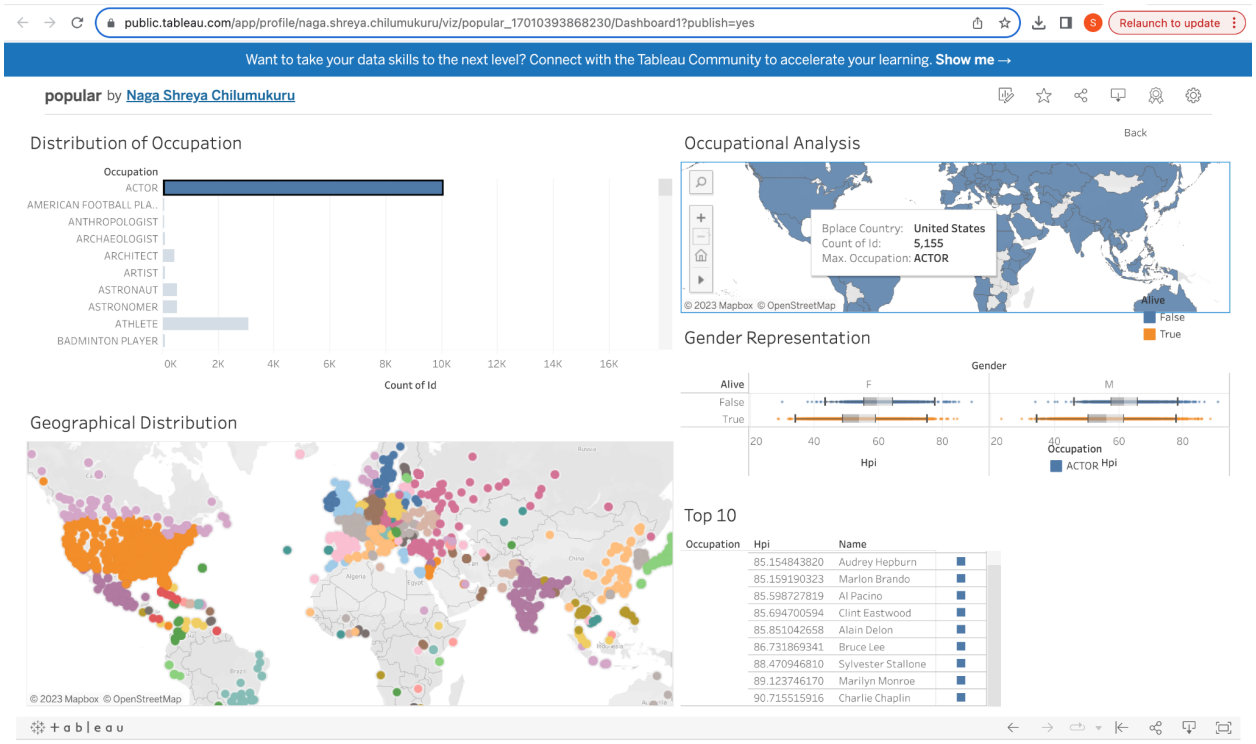
In the Geographical Distribution, we can select a datapoint. This datapoint represents individuals, color-coded based on the country. Hovering on the datapoint, it displays the Birthplace country, birthplace latitude, birthplace longitude, name and the occupation of that individual.



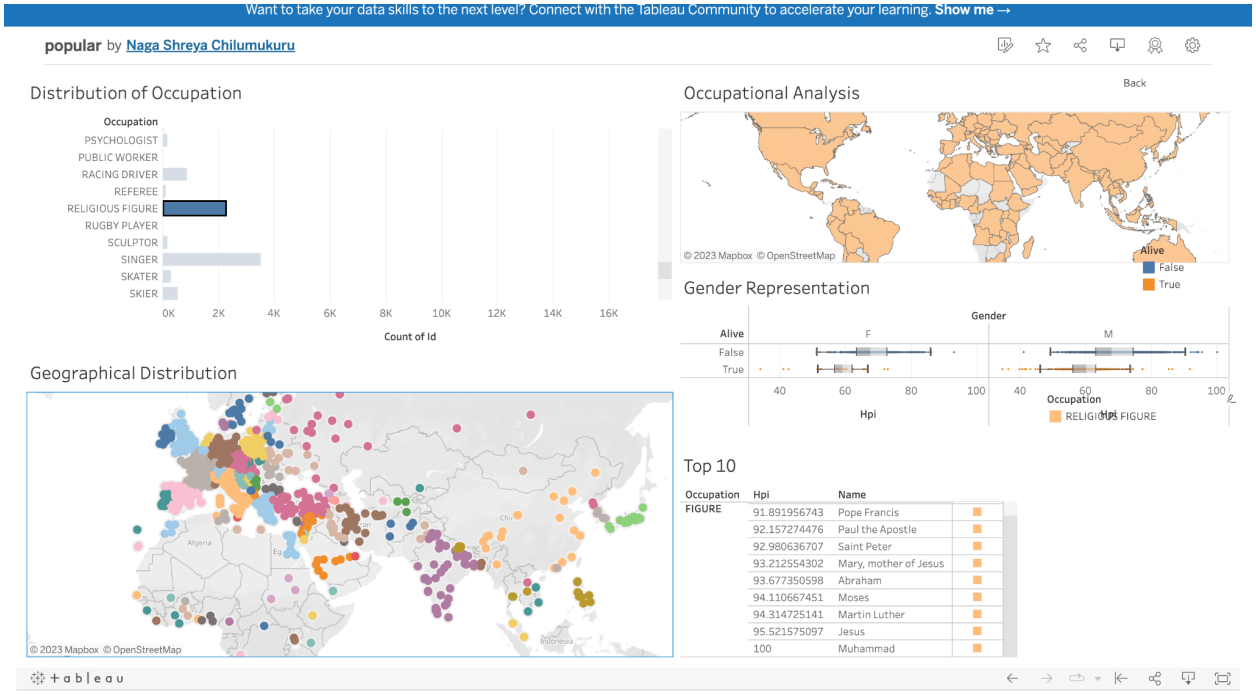
Upon selecting the data point, all the charts of the dashboard change accordingly. For eg, the data point selected on Geographical Distribution shows the individual is a Male Politician from Turkey with HPI - 74.373362669 and is alive.



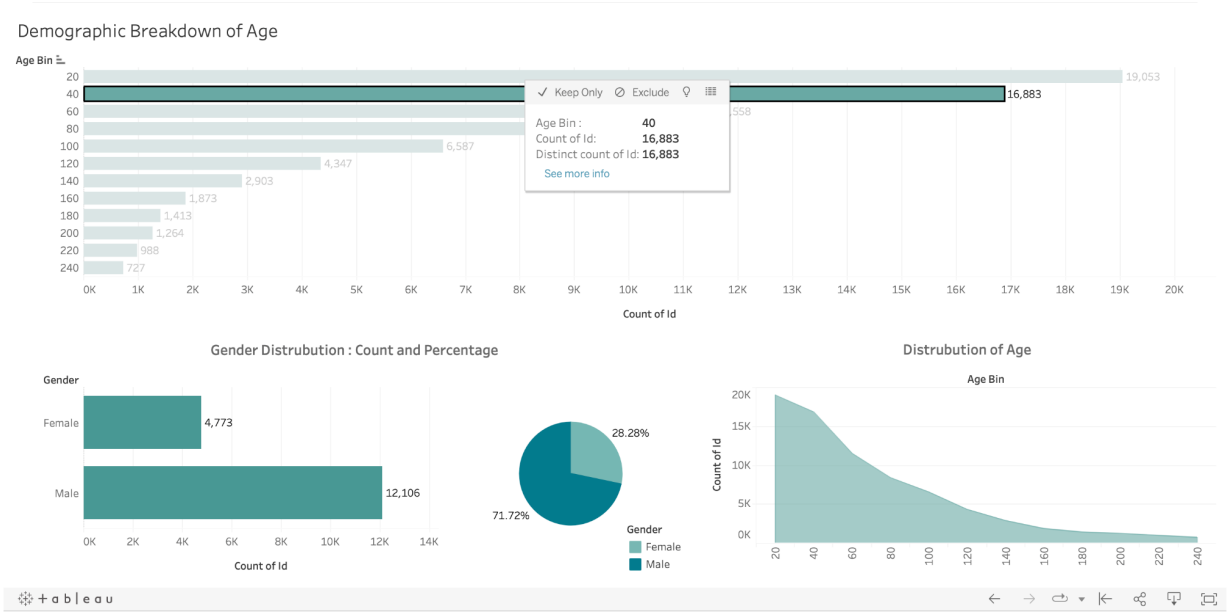
When occupation from the Distribution of Occupation is selected, we can see the Top 10 of that particular selected occupation, along with its gender distribution. We see that the United States has 5,155 number of actors. The actors are as data points in Geographical Distribution. Charlie Chaplin is the most popular actor with an HPI of 90.



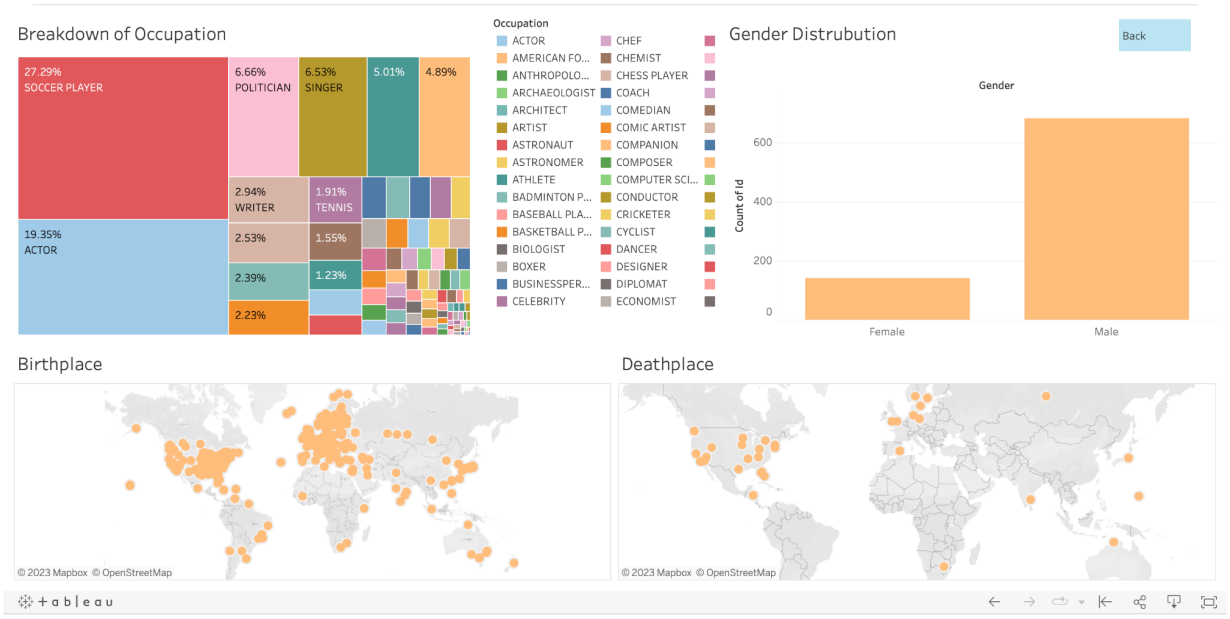
We see that ‘Muhammed’ is the most popular with HPI-100, who is a Religious Figure.



This dashboard shows the breakdown of age across the dataset along with information about the gender distribution in both count and percentage form. With the ages being split into bin sizes of 20, on clicking each bar, there is two fold options. Firstly, the gender distribution count chart and pie chart will be updated according to the respective bin. Secondly, one clicking ‘See More Info’, a second dashboard will be shown with occupation breakdown about that respective age bin.

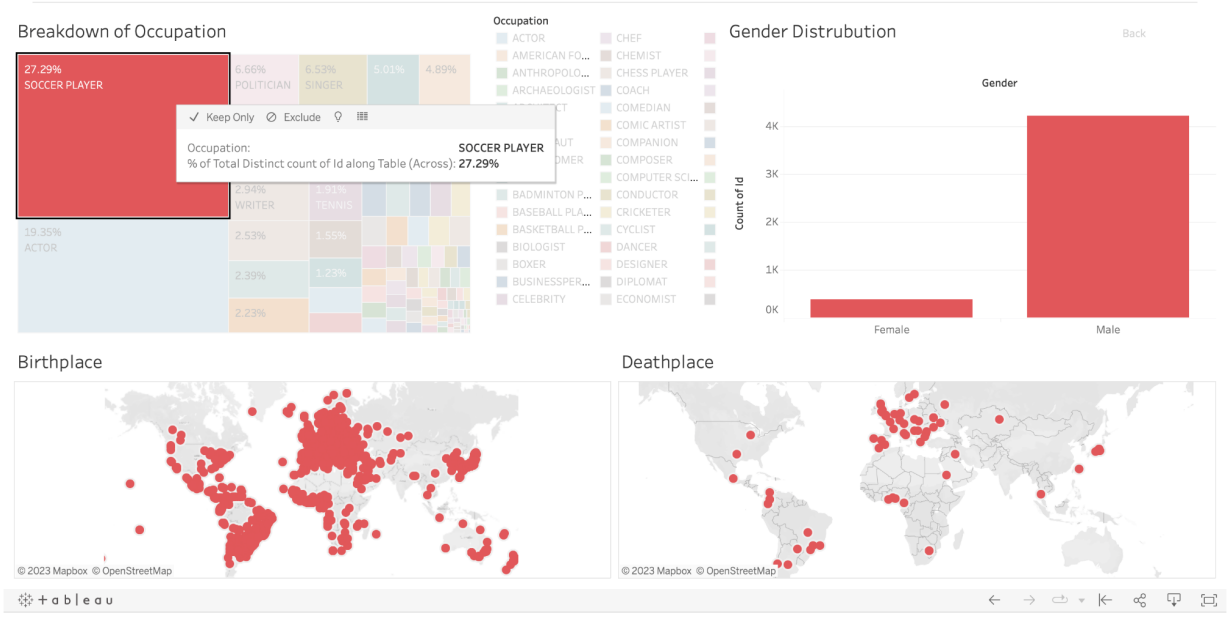


The second dashboard contains the breakdown of occupations shown in a tree map within the respective age bin, from this, the top percentages of each occupation can be inferred. For eg, in the screenshot below, from ages 40 to 60, politicians consist of the most at 27.29%, followed by actors with 19.35% and so on.

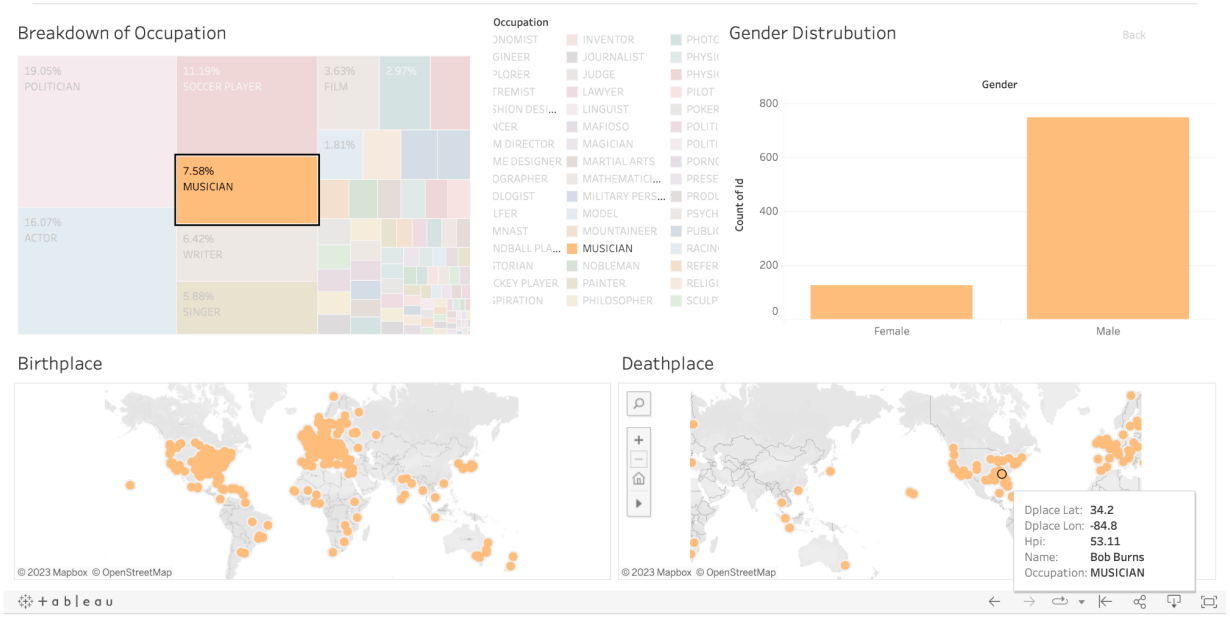


On clicking on the occupation in the tree map, the information of the individuals in the respective occupation will be displayed. The gender distribution chart with count of each will be shown on hovering along with two geographic maps. For example, on selecting soccer players,

the geographic distribution of the birthplaces of soccer players shows that the majority are clustered in South America and in Europe. Along with this, the geographic distribution of the location of places of death of the individuals are shown too.



A legend is also provided for all the different occupations and on clicking, the respective occupation will be highlighted in the other graphs. On the geographic maps, on hovering over them, information about each individual can be seen like HPI, name, and occupation.



Result

We optimized our Randomforest mode using RandomizedSearchCV and checked for different hyperparameters. After checking for different parameters we found that this parameter formed the best on the train and test dataset

```
clf_cv = joblib.load("random_forest_cv.pkl")

clf_cv.best_params_
```

```
{'n_estimators': 1400,
 'min_samples_split': 5,
 'min_samples_leaf': 1,
 'max_features': 'sqrt',
 'max_depth': 30,
 'bootstrap': True}
```

these were the score .

'mse_train': 1.4822596883071955
'mse_test': 10.879188449279965
'r2_train': 0.9875369326667329
'r2_test': 0.9086826888371029

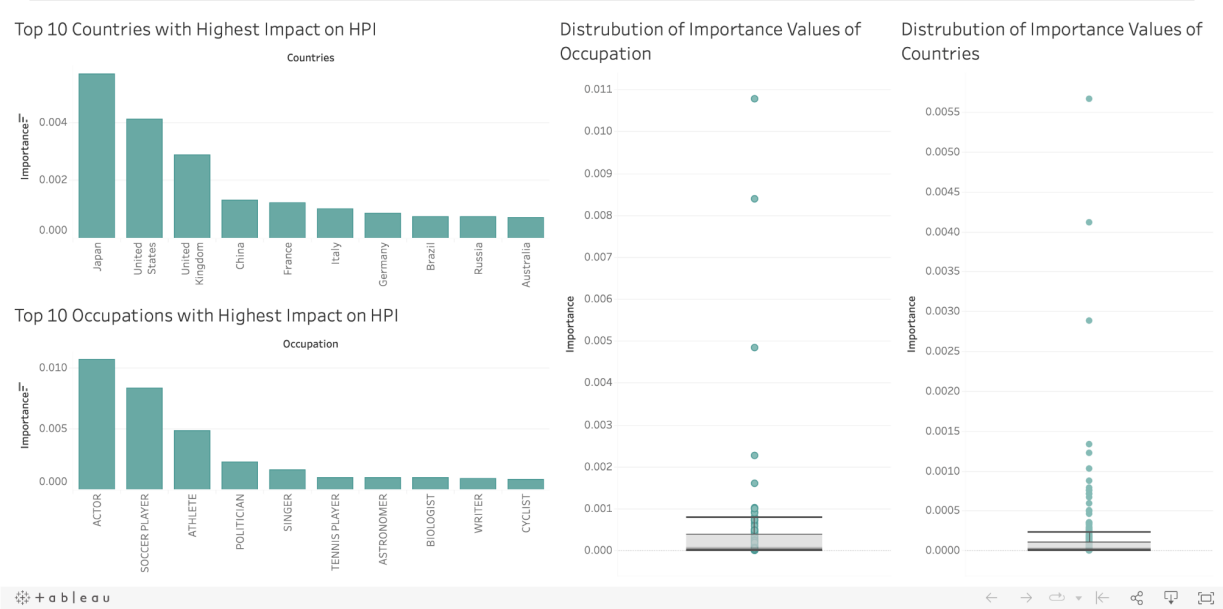
These were the top 20 most important features

```
imp.sort_values(by = 'importance')[-20:]
```

	importance
bplace_country_Italy	0.001033
bplace_country_France	0.001223
gender_F	0.001299
gender_M	0.001329
bplace_country_China	0.001331
occupation_SINGER	0.001613
occupation_POLITICIAN	0.002271
bplace_country_United Kingdom	0.002884
bplace_country_United States	0.004116
occupation_ATHLETE	0.004849
bplace_country_Japan	0.005668
occupation_SOCCER PLAYER	0.008385
occupation_ACTOR	0.010780
non_en_page_views	0.015758
coefficient_of_variation	0.026167
birthyear	0.028125
l_	0.102103
prob_ratio	0.152285
age	0.177869
alive	0.408792

Visualizations from Machine Learning Model

The insights from the regression model are shown in the dashboard below. The model showed that occupation and place of birth have a high importance value for HPI. The countries that have high importance values on HPI are shown to be Japan followed by the United States, and United Kingdom while the top occupations are actors followed by soccer players and athletes.



Discussion

Advantages

1. Exceptional Accuracy:

We achieved an exceptionally high accuracy in predicting HPI scores, scoring an impressive R-squared value of 0.90, giving us a high degree of confidence in our model's predictions.

2. Utilization of Ensemble Learning (Random Forest):

By employing ensemble learning, such as Random Forest, we construct a robust and dependable model by combining multiple weaker learners, leading to enhanced generalization and reduced overfitting.

3. Rapid Model Training:

Machine learning models like Random Forest exhibit swift training times, allowing for efficient model development.

4. Assessment of Feature Significance:

The RandomForest model offers insights into the importance of each feature, aiding us in comprehending the individual contributions of features towards a person's memorability.

Disadvantages

1. Lacks Individual Emphasis:
This model primarily considers general characteristics such as birthplace and occupation, which are insufficient to fully characterize an individual.

2. Influence of Contemporary Events:
There is a bias towards recent events, as people tend to assign greater significance to current circumstances. Consequently, our model identifies a person's current vitality as the most influential feature.

References

• Tableau Visualization links -
•https://public.tableau.com/app/profile/naga.shreya.chilumukuru/viz/popular_17010393868230/Dashboard1?publish=yes

•https://public.tableau.com/app/profile/sarah.mathew2948/viz/FamousPeopleDataset/Dashboard1?publish=yes

• Github link - https://github.com/Universe-89/Decoding_Fame

• https://www.youtube.com/watch?v=J-KgHlcYv8M

• Dataset - https://pantheon.world/explore/rankings?show=people&years=-3501,2023