Analyzing Trends on the Characteristics of Notable People on Wikipedia*

[BLAH, BLAH,BLAH] characteristics are associated with increased prominence

Parth Samant

November 30, 2024

Wikipedia is an online encyclopedia that contains information on millions of diverse subjects, arguably serving as a repository for human knowledge. In this paper, I analyzed a dataset of notable people on Wikipedia from 3500 BCE to 2018 AD to identify whether certain characteristics (such as age and gender) were correlated with prominence. The analysis found that [WHAT DID YOU FIND???]. The findings highlight….

Table of contents

*Code and data are available at: https://github.com/samantparth/Wikipedia-Historical-Prominence-Trends.

1 Introduction

Overview paragraph

Wikipedia is an online encyclopedia that contains information on millions of diverse subjects, arguably serving as a repository for human knowledge. Within this vast encyclopedia, there is also a collection of biographical information on "notable people" who have lived spanning millennia. Thus, one research study (Laouenan et al. 2022) compiled a dataset of notable people on Wikipedia from 3500 BCE to 2018.

Using this dataset, I constructed a linear model to identify whether certain characteristics were associated with increased prominence. Some of the factors focused on include trends in geographical subregion, gender, occupation, and the number of years since the individuals birth. By analysing these factors, this paper aims to uncover potential bias in online biographical representation and contribute to a deeper understanding of factors connected to online prominence.

Estimand paragraph

The estimand (or what is being estimated) is the prominence of their Wikipedia biography based on percentile. "Prominence" of a biography is determined using multiple metrics, such as the average amount of views per year, the total word count, and the number of Wikipedia editions. This constructed percentile variable is also a transformation of the variable `ranking_visib_5criteria` present in the original dataset. Features of this variable are elaborated on in Section 2.2.

Results paragraph

Why it matters paragraph

Understanding which factors contribute to prominence on Wikipedia is essential for many reasons. Wikipedia is used as a crucial tool for obtaining information for millions of people worldwide, having the ability to shape perceptions of prominent individuals. Unequal representation of prominent people can perpetuate existing biases about which characteristics are seen as "ideal" for prominence. Furthermore, these biases can discourage some individuals from pursuing fields where prominent people sharing their characteristics are under represented. Thus, analysing these dynamics can help with addressing broader societal biases and striving for a more comprehensive view of individual prominence.

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2 mentions features of the dataset, **?@sec-model** introduces and explains the constructed model, Section 6 summarizes the results of the paper, and Section 7 discusses the significance of these findings.

2 Data

2.1 Overview

All data analysis was done through the statistical programming language R (R Core Team 2023) with the packages `tidyverse`(Wickham et al. 2019), `rstanarm` (Goodrich et al. 2022), `modelsummary` (Arel-Bundock 2022), `arrow`(Richardson et al. 2024), `readr`(Wickham, Hester, and Bryan 2024), `httr`(Wickham 2023), `R.utils`(Bengtsson 2023), `knitr`(Xie 2024), `tidymodels` and `ggplot2` (Wickham 2016).

2.2 Measurement

Wikipedia is an online encyclopedia that provides information on numerous subjects, including the lives of those who are relatively well-known. Using Wikipedia data, one research study (Laouenan et al. 2022) aimed to build a cross-verified database of 'notable people' who have ever lived. Individuals with a Wikipedia article are considered "notable", as the overwhelming majority of people who have ever lived do not have one. This data was obtained using the Wikidata universe (which provides data on Wikipedia), where they used the "instance of humans" category to select for a sample of notable individuals. Additionally, this dataset of notable people is very large (with hundreds of thousands of entries), reflecting the large amount of notable people that are written about on Wikipedia.

The data was also verified by the researchers through numerous ways. One way was the cross-verification of information that used 7 different versions of Wikipedia and Wikidata to make sure that the versions are consistent. Manual checks of validity were also used by hiring teams from a diverse set of countries (France, the UAE, and India) to compare the information on their database to that on Wikipedia/Wikidata.

Based on the information provided in those Wikipedia articles, the study was also able to identify numerous characteristics of these people - reflected by the 49 variables in the dataset. A more in-depth list of all these variables can be found in the appendix (Section C.2.2). Some of these factors of interest for this analysis include ones geographic origin, occupation, and age.

Furthermore, the researchers constructed many variables including `ranking_visib_5criteria`, which ranks the prominence of an individual on Wikipedia. The ranking is dependent on these 5 metrics :

- the number of different editions;
- the number of non-missing items for birth date, gender, and domain of influence;
- the total number of words for their article;
- average yearly number of viewers from 2015 to 2018;
- number of external links (such as sources and references) from Wikidata.

More on the measurement of the dataset can be found in Section C, which provides an in-depth exploration on the methodology of the dataset, as well as its limitations.

2.3 Data Cleaning and Variables

This dataset was cleaned by first mutating some variables, selecting/renaming variables of interest, and filtering out rows with missing information. The variables selected are mentioned later on in this section.

Since this paper focuses on predicting historical prominence, I chose variables which I thought could be associated with the outcome variable (`percentile_rank`).

Outcome Variable:

- **percentile_rank**: a transformation of the variable `ranking_visib_5criteria` (as mentioned in Section 2.2). This indicates the percentile associated with the notability ranking. A higher percentile indicates they were a more notable person.

Predictor Variables:

- **subregion**: the UN subregion corresponding to where they were from. This is simply a renaming of the variable `un_subregion`.
- **years_since_birth**: the number number of years that have passed since their birth. If they are alive, this is simply their age.
- **time_period**: The time period that they were born in.
- **gender**: The reported gender of the individual (either male or female).
- **occupation**: The primary field/occupation that the individual is known for.

2.4 Analysis of Variables

2.4.1 Subregion

?@tbl-subregion_counts contains information on the different subregions as well as the proportion of notable people for each subregion.

subregion	Proportion
Western Europe	0.4497572
Northern America	0.1549926
Southern Europe	0.0887510
South America	0.0598836
Northern Europe	0.0560671

subregion	Proportion
Eastern Europe	0.0465828
Eastern Asia	0.0344298
Oceania Western World	0.0246580
South Asia incl. Indian Peninsula	0.0186597
Western Asia (Middle East Caucasus)	0.0169979
Central America	0.0110171
SouthEast Asia	0.0101119
West Africa	0.0058536
Southern Africa	0.0048194
East Africa	0.0047216
Caribbean	0.0044518
North Africa	0.0042328
Central Africa	0.0020314
Central Asia	0.0015172
Oceania not Aus Nze	0.0004634

Proportion of Notable People On Wikipedia by Geographical Subregion

Interestingly, subregions that are a part of the Western world tend to be over-represented in terms of notable people. In fact, the difference is so stark that the most well-represented region (Western Europe) has nearly 100x the representation of notable people than West Africa.

There could be many explanations for this, such as how Western nations tend to have more global cultural dominance (and thus more 'notable people').

2.4.2 time_period/years_since_birth

These are two variables that are directly associated with each other, since the time period of ones life is directly a result of the number of years since their birth.

However, a notable feature of `years_since_birth` (and thus `time_period`) is the distribution as shown in Figure 1.

From Figure 1, we can see that the overwhelming majority of documented notable people on Wikipedia tend to be born in the past 500 or so years. This makes sense, as retrieving information on individuals that lived a longer time ago is more difficult. Future plots and analyses will thus use the logarithm of this variable to have a more detailed view on its impact.

Figure 1 also has consequences for the `time_period` variable. There is an extreme lack of representation for those born in earlier time periods (especially before 500 AD), likely because of much documentation of notable people being lost from time.

Figure 1: Distribution of Notable Individuals by Years Since Birth

2.4.3 Time Period and Subregion

Different regions had different levels of cultural relevance/dominance depending on the time period. Because of this fact, it may be possible that the prevalence of subregions can heavily depend on the number of years since their birth.

In Figure 2, we can see this relationship quite clearly. Note that the proportions are out of the top 5 most popular subregions instead of every subregion. This was done to ensure readability.

From Figure 2, we can confirm the relationship between the time period and subregion - that the prevalency of notable individuals heavily changes depending on the time period in question. The most notable example is the decrease in prevalence for Southern Europe/East Asia and the subsequent rise of Northern America (i.e., the USA and Canada) and Western Europe.

As mentioned above, this confirms the phenomenon of how different subregions have different levels of influence depending on the time period. This graph especially highlights the rise of notable people in Western Europe as well as its descendant countries.

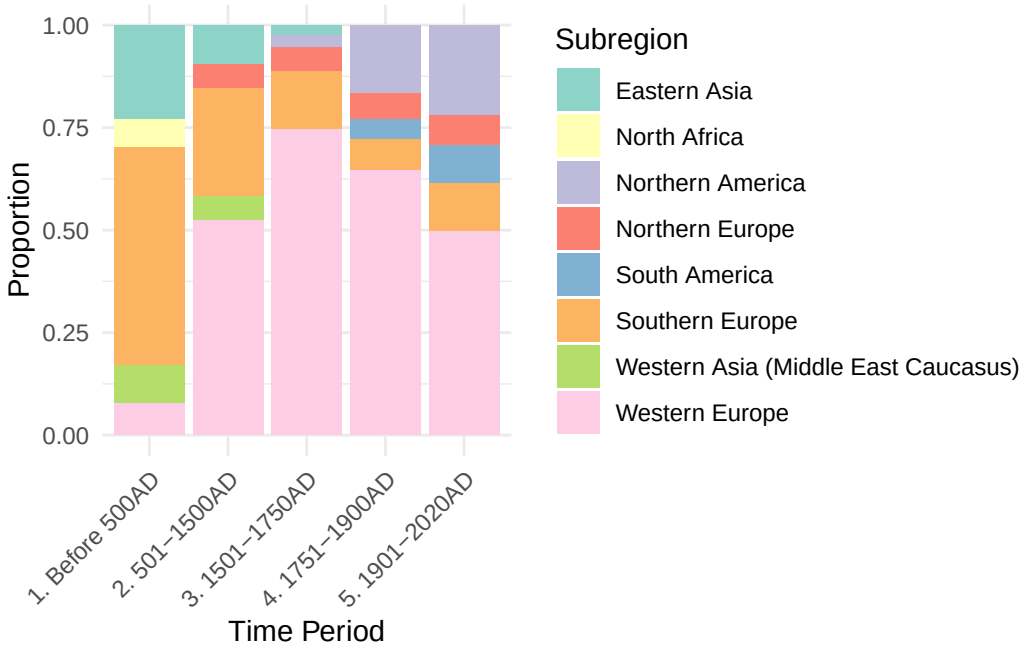


Figure 2: Top 5 Subregions of Notable People by Time Period. The graph illustrates the large variability of prominent subregions through time.

2.4.4 Gender and Occupation

The remaining variables, gender and occupation, may also show some sort of trend (since ones field of work is often quite gendered). This relationship is shown in **?@tbl-gender_occupation**, which shows ones occupation (or more specifically, what they are known for).

occupation	proportion_female	proportion_male
Family	0.31	0.69
Culture-core	0.24	0.76
Nobility	0.22	0.78
Culture-periphery	0.22	0.78
Other	0.21	0.79
Missing	0.15	0.85
Sports/Games	0.14	0.86
Worker/Business (small)	0.13	0.87
Politics	0.11	0.89
Academia	0.10	0.90
Corporate/Executive/Business (large)	0.06	0.94

occupation	proportion_female	proportion_male
Administration/Law	0.06	0.94
Religious	0.04	0.96
Explorer/Inventor/Developer	0.03	0.97
Military	0.02	0.98

Occupations of Notable People by Gender, Showing High Proportion of Males

?@tbl-gender_occupation shows that those of nobility (often by being born in a high social rank) tend to have a higher proportion of females, where other occupations - such as the military - tend to have a higher proportion of males.

Interestingly, even the most female-dominated occupation (nobility) is still roughly 2/3rds men. This reflects a clear bias in people that are considered 'notable': men are very much over-represented in every occupation (and in this dataset, as a result).

?@tbl-gender reinforces the phenomenon of notable people on Wikipedia being overwhelmingly male. In fact, the proportion of males is roughly 85%.

Count	female_count	male_count	proportion_female	proportion_male
511476	76322	435154	0.1492191	0.8507809

Proportion of Notable People by Gender

3 *Make sure to include graph on occupation*

4 Model

4.1 Overview

The ultimate goal the modelling strategy is to estimate the prominence of individuals on Wikipedia (by percentile). The model incorporates interaction terms as well as validation techniques to ensure model robustness and accuracy.

Additional background details and diagnostics are included in the Appendix Section B.

4.2 Model set-up

The model is expressed by the equation:

$$\begin{aligned}\text{prominence percentile} = \beta_0 + \beta_1 \cdot \text{occupation} + \beta_2 \cdot \text{subregion} + \beta_3 \cdot \text{time period} \quad (1)\\ + \beta_4 \cdot \log(\text{age}) + \beta_5 \cdot \text{gender} + \beta_6 \cdot (\text{time period} \times \log(\text{age})) \quad (2)\\ + \beta_7 \cdot (\text{occupation} \cdot \text{gender}) + \epsilon \quad (3)\end{aligned}$$

with the variables representing the following:

- β_0 : the intercept term when all other predictors are set to zero.
- β_1 : the effect of ones occupation (relative to one with an occupation of an Academic) on prominence percentile ranking.
- β_2 : the effect of ones geographical subregion (relative to the Caribbean) on prominence percentile ranking.
- β_3 : the effect of ones time-period (relative to those born before 500AD) on prominence percentile ranking
- β_4 : the logarithmic effect of age (`years_since_birth`).
- β_5 : The effect of ones gender (relative to female) on prominence percentile ranking
- β_6 : the interaction between the individuals time period (at birth) and age
- β_7 : the interaction between ones occupation and their gender
- ϵ : residual term, or the variation in percentile_ranking not due to the predictors.

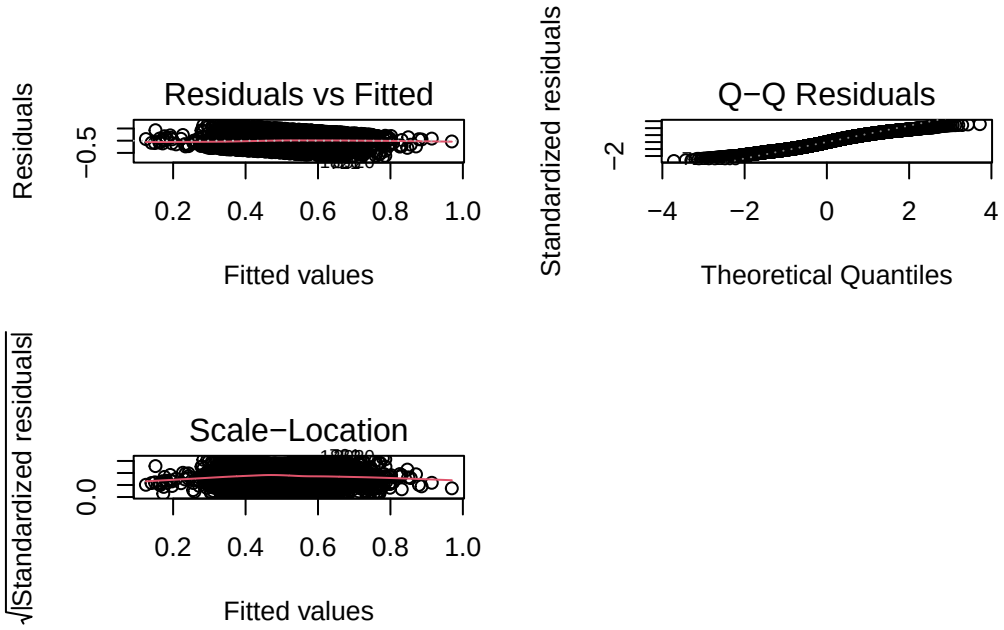
The model was ran in R (R Core Team 2023) using the base `lm()` function.

Many of the modelling decisions are reflective of the findings from the Section 2.4 section. These include:

- the decision to apply `log()` to the age, as the distribution of this variable followed an exponential distribution (thus transforming this variable aids with identifying these effects)
- Additionally, the decision to add an interaction term between the individual's age and time period is informed by how time period is a categorical representation of ones age (thus affecting one another)
- The decision to apply an interaction between gender and occupation. It was shown that gender ratios depend heavily on the occupation, making them strongly associated with one another.

4.3 Model Assumptions

Linear regression relies on assumptions that should be verified before drawing inferences on the results of the model. These assumptions include linearity, homoscedasticity (constant variance), and normally distributed error (residuals). Figure 3 provides graphs that checks these assumptions.



(a) Residuals vs Fitted Plot: Shows Roughly Linearity and Constant Variance

Figure 3: Compilation of Graphs Verifying Assumptions of Linear Regression

5 ELABORATE MORE ON LINEAR REGRESSION GRAPHS (SENTENCE or two for each graph)

5.1 Model Validation

5.1.0.1 Bayesian information Criterion

The Bayesian information criterion (BIC) is a metric used for model selection among a finite set of models. It is based on the likelihood function and penalises datasets that may be too large. This validation technique helps minimize overfitting while retaining the goodness of fit. (https://ieeexplore.ieee.org/document/1311138)

This statistical metric was used as an aid in determining which predictor variables to remove and which variables to include in the final model. More details on how this method was used is found in the appendix (Section B.1).

5.1.0.2 Root Mean Squared Error

Analysis of the root mean squared error (RMSE) can provide a clue as to how accurate the model's estimations are. Root mean squared error measures the average difference between values predicted by the model and the actual values. Using the `modelsummary()` function (as shown in Table 4), the model obtained an RMSE of 0.26, which corresponds to an average error of 26 percentile ranks.

5.1.0.3 Out-of-sample testing

Further validation of the model, including the validation of linear regression assumptions, is done through out-of-sample testing. Training data on different models was used to generate predictive abilities for the model. Then, based on their predictions of the testing data, comparisons of the RMSE of different models indicated that the chosen model tended to perform better (a more accurate predictive model would tend to have a lower RMSE). More details on the different models tested and their comparisons can be found in Section B.2.

5.2 Limitations

A notable limitation of this model is that it is theoretically possible for the model to estimate a negative percentile or a percentile above 100. However, since these estimations would be physically impossible, this suggests potentially inaccurate results for more extreme percentile estimates.

Another limitation is the relative lack of data for some underrepresented groups. Section 2.4.1, for example, illustrates the discrepency for certain subregions in the dataset. A comparative lack of data for those certain subregions are likely associated with less predictive certainty from the model. In other words, the models predictions of prominence for people in different time periods/contexts is likely less accurate.

A final limitation of this model is that predicting prominence using a dataset on Wikipedia relies heavily on the informational accuracy of the original dataset. Incorrect information on the characteristics results in an inaccurate model.

Table 4

	(1)
time_period3. 1501-1750AD	−0.898
	(0.350)
(Intercept)	0.837
	(0.346)
occupationMissing	−0.259
	(0.012)
occupationPolitics	−0.151
	(0.004)
occupationCorporate/Executive/Business (large)	−0.151
	(0.011)
occupationSports/Games	−0.149
	(0.004)
occupationAdministration/Law	−0.135
	(0.008)
subregionCentral America	−0.133
	(0.007)
Num.Obs.	511 476
R2	0.137
R2 Adj.	0.137
AIC	95 133.0
BIC	95 790.6
Log.Lik.	−47 507.519
RMSE	0.27

Output of modelsummary() for linear model, highlighting 6 of the most significant predictor variables. “Significance” was determined by co-efficients that have a p-value less than 0.05 that also large effect on prominence

6 Results

Our results are summarized in Table 4.

the `modelsummary()` function from the `modelsummary` package (Arel-Bundock (2022)) provides a summary for the findings of the linear model, and is shown in Table 4 to highlight the top 6 most significant characteristics that affect Wikipedia prominence. Aside from the time period, the most significant coefficients of the model were occupation-related.

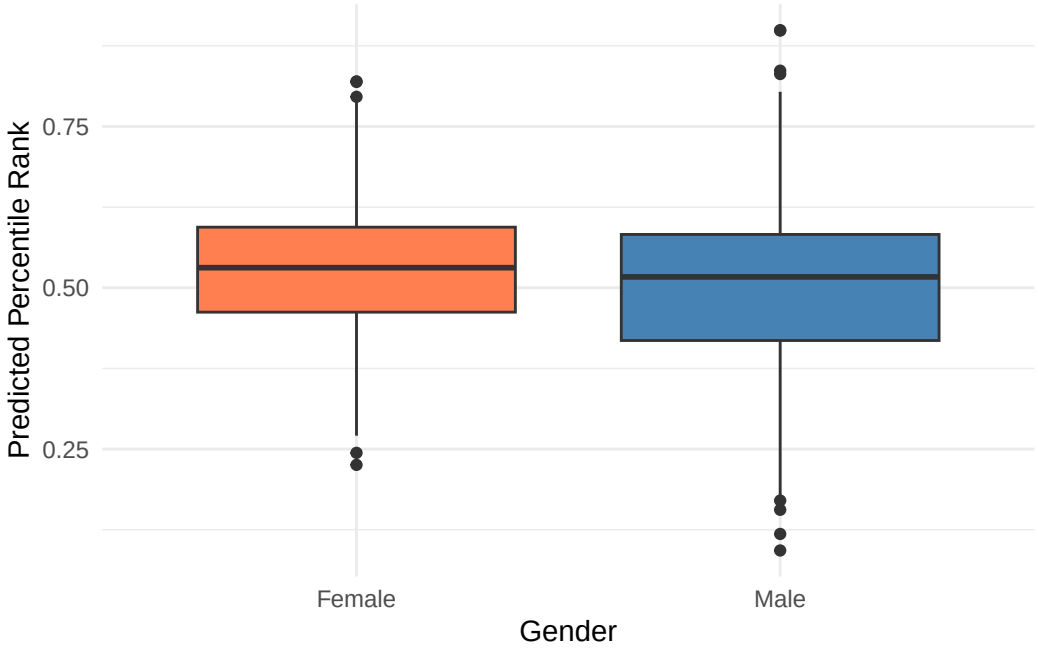


Figure 4: Boxplot of Model Predictions for Prominence by Gender

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
                                        (Intercept)
                                       0.8368625543
                          occupationAdministration/Law
                                      -0.1348132346
occupationCorporate/Executive/Business (large)
                                      -0.1510825854
                                occupationCulture-core
                                       0.0266402535
```

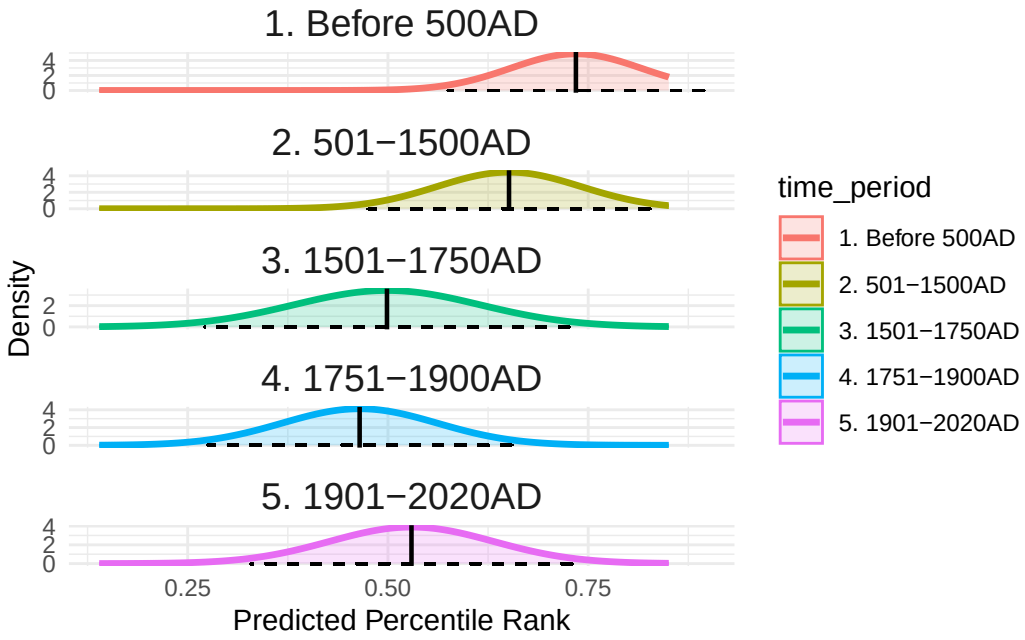


Figure 5: Model-Predicted Distributions of Percentile Rank by Time Period. The graph illustrates a much higher prominence for those born before 500 AD

```
occupationCulture-periphery
                -0.0783644799
occupationExplorer/Inventor/Developer
                -0.1218682029
          occupationFamily
                -0.0294693611
        occupationMilitary
                -0.0715211046
         occupationMissing
                -0.2588165237
        occupationNobility
                 0.0726272126
           occupationOther
                -0.0865664602
        occupationPolitics
                -0.1512125260
       occupationReligious
                -0.0703046397
occupationSports/Games
                -0.1489584962
```

occupationWorker/Business (small)
-0.1260862685
subregionCentral Africa
-0.0094135720
subregionCentral America
-0.1333431173
subregionCentral Asia
0.0038106858
subregionEast Africa
-0.0930998962
subregionEastern Asia
0.1053899814
subregionEastern Europe
0.0348917538
subregionNorth Africa
0.0263860859
subregionNorthern America
0.0343432383
subregionNorthern Europe
-0.0563574510
subregionOceania not Aus Nze
-0.0050365940
subregionOceania Western World
-0.1084800686
subregionSouth America
-0.1142211858
subregionSouth Asia incl. Indian Peninsula
-0.0673893884
subregionSouthEast Asia
-0.0247855040
subregionSouthern Africa
-0.0799185244
subregionSouthern Europe
-0.0199370342
subregionWest Africa
-0.0123327633
subregionWestern Asia (Middle East Caucasus)
0.0251748765
subregionWestern Europe
-0.0556311112
time_period2. 501-1500AD
-0.3912546840
time_period3. 1501-1750AD

```
                                                -0.8982717568
                                    time_period4. 1751-1900AD
                                                -0.4792717377
                                    time_period5. 1901-2020AD
                                                 0.2058102327
                                       log(years_since_birth)
                                                -0.0018340550
                                                   genderMale
                                                 0.0117577612
             time_period2. 501-1500AD:log(years_since_birth)
                                                 0.0384912995
           time_period3. 1501-1750AD:log(years_since_birth)
                                                 0.1158263559
           time_period4. 1751-1900AD:log(years_since_birth)
                                                 0.0441520033
           time_period5. 1901-2020AD:log(years_since_birth)
                                                -0.0988423743
                    occupationAdministration/Law:genderMale
                                                -0.0197602399
occupationCorporate/Executive/Business (large):genderMale
                                                -0.0114298919
                          occupationCulture-core:genderMale
                                                -0.0009929844
                       occupationCulture-periphery:genderMale
                                                -0.0001198824
           occupationExplorer/Inventor/Developer:genderMale
                                                 0.0450723732
                               occupationFamily:genderMale
                                                -0.1171893012
                             occupationMilitary:genderMale
                                                -0.0871412362
                              occupationMissing:genderMale
                                                -0.0723521617
                             occupationNobility:genderMale
                                                -0.1207374389
                                occupationOther:genderMale
                                                -0.0051224833
                             occupationPpolitics:genderMale
                                                -0.0268618423
                           occupationReligious:genderMale
                                                 0.0278447971
                          occupationSports/Games:genderMale
                                                 0.0451199794
```

```
occupationWorker/Business (small):genderMale
                                -0.0542905740

                      subregionCentral Africa
                                 -0.009413572
                    subregionCentral America
                                 -0.133343117
                       subregionCentral Asia
                                  0.003810686
                        subregionEast Africa
                                 -0.093099896
                       subregionEastern Asia
                                  0.105389981
                     subregionEastern Europe
                                  0.034891754
                       subregionNorth Africa
                                  0.026386086
                  subregionNorthern America
                                  0.034343238
                     subregionNorthern Europe
                                 -0.056357451
                subregionOceania not Aus Nze
                                 -0.005036594
              subregionOceania Western World
                                 -0.108480069
                      subregionSouth America
                                 -0.114221186
  subregionSouth Asia incl. Indian Peninsula
                                 -0.067389388
                     subregionSouthEast Asia
                                 -0.024785504
                    subregionSouthern Africa
                                 -0.079918524
                    subregionSouthern Europe
                                 -0.019937034
                       subregionWest Africa
                                 -0.012332763
subregionWestern Asia (Middle East Caucasus)
                                  0.025174876
                     subregionWestern Europe
                                 -0.055631111
```

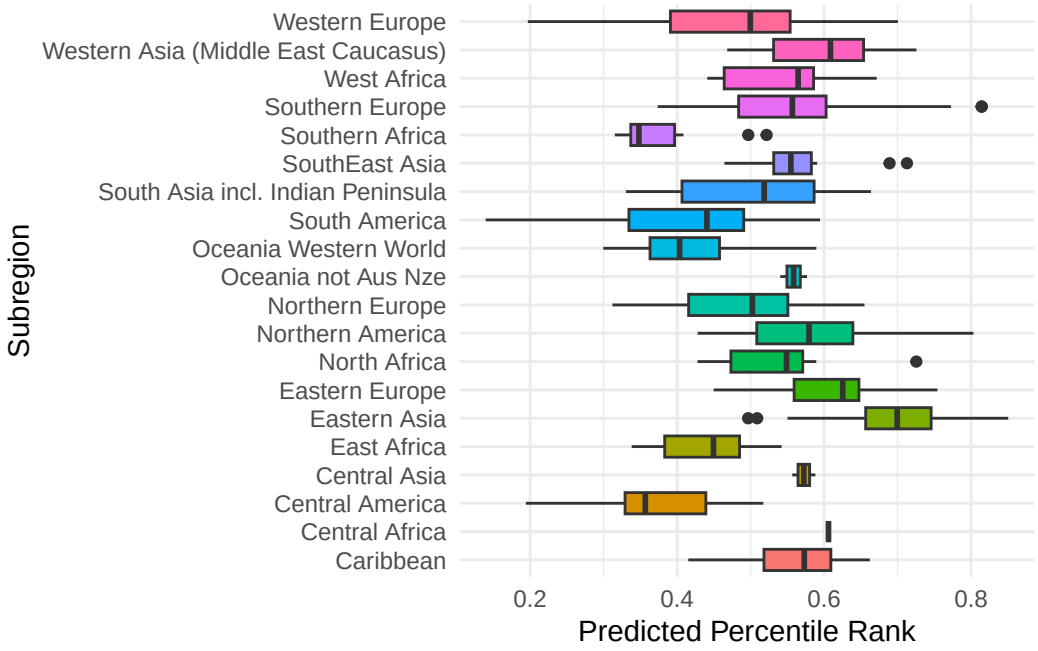


Figure 6: Boxplot of Model's Predicted Prominence of Individuals Based on Geographical Subregion

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

Figure 7 highlights the trend between predicted prominence of the model compared to the actual prominence. The shadow of the blue line indicates the 95% confidence interval of predictions from the linear model. This illustrates a weak, but existing, relationship between the prediction given by the model and the actual percentile rank.

7 Discussion

(MAKE THIS SECTION ~100 LINES)

7.1 Notable Relationship between Time Period and Prominence

Interestingly, time period tends to have quite a strong effect on prominence that is more significant than other characteristics that were evaluated for (such as subregion or gender). Notable individuals that were born before 500AD displaying higher prominence makes intuitive sense, as time may have resulted in many records of otherwise notable people not being

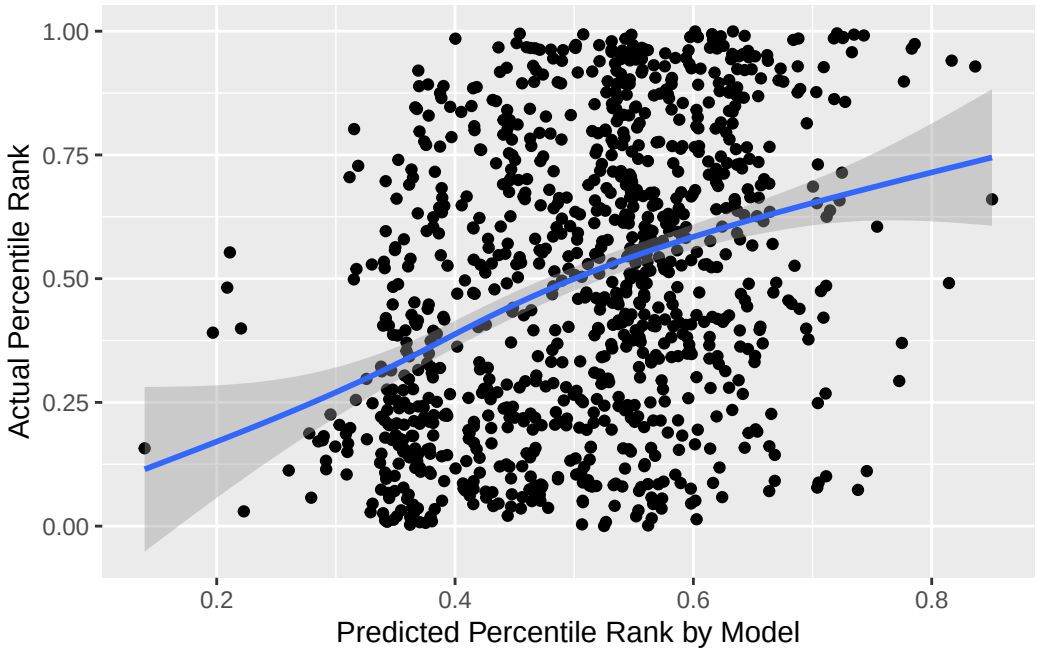


Figure 7: Graph comparing the model's predicted percentile of prominence vs. the actual percentile. Testing data (i.e., data the model was not trained with) was used to more accurately evaluate prediction ability. 1000 random datapoints were selected to highlight the overall results.

included on Wikipedia. It is possible that those who were significantly influential or notable thousands of years ago tended to have more information being written about them, making their biographical records last longer.

There is also a notable increase in prominence for those born between 501-1500AD, which might also be explained by the same phenomenon as those born before 500 AD. Interestingly, time periods after 1500 AD do not display this same trend in prominence, with those born between 1901-2020 AD having a higher mean prominence than those in 1501 AD.

Future studies could specifically focus on the relationship between individual prominence and time periods, including the reason for those in time periods after 1500 AD not showing notable mean differences in prominence.

7.2 Geographic Subregion and Prominence

https://data.worldbank.org/indicator/IT.NET.USER.ZS

Using data from the World Bank, it is evident that countries that tend to have a larger prominence of internet users tend to be wealthier countries (with many of them being part of the western world). This was shown previously in Section 2.4, which suggested a significant overrepresentation of those from Europe and Northern America.

Interestingly, the trend of western individuals being more notable does *not* continue when looking at already notable people (as shown by the lower predicted prominence of those from Oceania deemed to be part of the Western world). Even though the model underestimates the spread of the data, it is evident that those from Eastern Asia are an outlier in terms of prominence.

7.3 Similar Prominence Between Notable Men and Women

some studies, including the one by (Banaji and Greenwald 1995), highlighted that unconscious gender stereotyping was present and that individuals assigned a higher assignment of fame to males than females. Another meta-analysis of gender inequality in the workplace by (Stamarski and Son Hing 2015) highlights sexism where women tend to receive less promotions and less leadership roles, among other factors. Women receiving less leadership and more lower-status roles in the workplace could easily translate into biases in societal prominence.

However, the findings of this study do not entirely confirm previous studies between the relationship of gender and fame (which is heavily associated with prominence). In fact, the results relating to gender (shown in Figure 4) highlight that notable women tend to be slightly more prominent than notable men. However, it also suggests that there is a large amount of variation in prominence between notable men and women, indicating that gender is likely not a significant characteristic in predicting prominence from those who are already notable.

It also may be possible that the underrepresentation of women in this dataset means that the women who are already notable require a higher threshold of prominence to be on the dataset in the first place.Future studies can be conducted to explore the relationship between already notable people and how gender may (or may not) be associated with prominence.

7.4 Discussion of modelsummary() results

The results from the `modelsummary()`, which summarize the model, are shown in Table 4. Unsurprisingly, those that do not have their occupations mentioned (`occupationMissing` in `modelsummary()`) tend to be much less prominent overall. It is quite intuitive that a lack of information on an individual correlates to lower overall prominence.

Interestingly, some occupations such as being involved in politics; being an important person in a large business; and being associated with sports or games also tend to have a notable effect on percentiles of prominence. The results suggest that occupation plays quite an important role in determining prominence, at least compared to the other predictor variables. Additionally, those born in the time period from 1501-1700 AD and those from Central America tend to have a lower prominence. Future analyses could be done to explore whether there is genuinely a strong effect of these certain characteristics on prominence, as this also may indicate a potential flaw in the model.

7.5 Weaknesses

7.5.1 Results only relevant for those already considered "notable"

analyses only takes into account people who are ALREADY PROMINENT. @sec-analysis-of-variables identifies characteristics of prominent people, and the appearance of bias in favour of certain characteristics (such as being a man or from Western Europe) is already reflected. However, since this study measures difference in prominence between those who are **already** notable, the linear model does not tend to highlight those biases as strongly.

7.5.2 Unequal Distribution of Characteristics

Not only has the world population grown exponentially in the past few hundred years, but records of notable people can be lost due to time. This can make predictions of prominence heavily biased towards individuals who live (or have lived) in more recent time periods. It may be difficult to extend the finding of prominence onto individuals with less represented characteristics (including being born before 500AD).

7.5.3 Limited Prediction Ability

From @fig-percentile_prediction, we can see the predictive ability of the model leaves a lot to be desired. It suggests that given the specific predictors of the model (related to gender, subregion, age, and broad categorization of occupation), there is still a lot of variability of prominence that is not explained by these predictors. This is further evident by the RMSE and R^2 values shown in @tbl-modelresults. The R^2 value of 0.138 suggests that these predictors only explain 14% of the variance in `percentile_rank`, which is quite low. This suggests that an individuals prominence on Wikipedia is largely explained by other factors not included in the model (including ones that are not measured in the original dataset).

We can also see that the model tends to put most of its predictions near the 50th-percentile mark. A more ideal model would have a more spread out prediction of percentile. This results in the model underestimating the overall spread/variance of the data, making some differences appear more stark than others. This phenomenon is also evident in figures that portray prediction results for given characteristics (such as @fig-subregion_prediction, @fig-time_period_prediction, and @fig-gender_prediction). This means that the model largely underestimates the overall spread/variance of the data, resulting in characteristic-related differences appearing more significant than they actually are.

although the predictive ability of this model leaves room for improvement, the model is still quite useful for analysing overall trends of prominence for specific characteristics.

‘

Appendix

A Additional data details

B Model details

B.1 Bayesian Information Criterion (BIC)

B.2 RMSE Comparison

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

```
{# {r} # plot(relevance_model, "trace") #  # plot(relevance_model, "rhat")
```

C Exploration of Notable People Dataset and Methodology

C.1 Introduction

This section of this appendix contains a more thorough analysis and exploration onto the dataset that was used for this analysis. The dataset, *A cross-verified database of notable people, 3500BC-2018AD* (Laouenan et al. (2022)), contains a quite thorough verification and collection to ensure accurate data. This is especially important because of the nature of online encyclopedias and the sheer amount of individuals that were analysed.

C.2 Data Source and Collection

The database integrates information from both Wikipedia and Wikidata. The extracted data from Wikipedia comes from seven popular language editions: English, French, German, Spanish, Portuguese, and Swedish. Many notable people documented (30%) thus come from the 6 non-English Wikipedia editions that were examined. Information on individuals is verified using different language editions, ensuring more reliable data.

Data is also extracted from Wikidata, which contains structured data linked to a Wikipedia page. The sample of individuals chosen was defined by this data (The researchers then merged the information on both websites to avoid dealing with duplicates and ensure data reliability.

C.2.1 Data Verification

Further verification was performed to ensure that the collected data was accurate. In addition to the reliability checks mentioned in Section C.2, the researchers employed a series of algorithms, including ones that serve to:

- **decipher humans vs. non-humans** - a large number of entries were incorrectly identified as humans in Wikidata, which may include fictional characters or even music bands. The researchers used a list of expressions, including "list of", "duos", "bands," etc. that can identify faux-individuals. These entries were then removed from the dataset.
- **eliminate duplicate biographies (deduplication)** - some individuals have more than one Wikipedia biography. The algorithm used nearly a dozen different methods to ensure duplicates are not present. One example is aggressive standardization of names/titles, allowing easier identification of duplicates.

C.2.2 Variables of the dataset

```
Warning in matrix(variable_names, ncol = 2, byrow = TRUE): data length [49] is
not a sub-multiple or multiple of the number of rows [25]
```

Table 5: Table of all variables present in the dataset

wikidata_code	birth
death	updated_death_date
approx_birth	approx_death
birth_min	birth_max
death_min	death_max
gender	level1_main_occ
name	un_subregion
birth_estimation	death_estimation
bigperiod_birth_graph_b	bigperiod_death_graph_b
curid	level2_main_occ
freq_main_occ	freq_second_occ
level2_second_occ	level3_main_occ
bigperiod_birth	bigperiod_death
wiki_readers_2015_2018	non_missing_score
total_count_words_b	number_wiki_editions
total_noccur_links_b	sum_visib_ln_5criteria
ranking_visib_5criteria	all_geography_groups
string_citizenship_raw_d	citizenship_1_b
citizenship_2_b	list_areas_of_rattach

Table 5: Table of all variables present in the dataset

area1_of_rattachment	area2_of_rattachment
list_wikipedia_editions	un_region
group_wikipedia_editions	bplo1
dplo1	bpla1
dpla1	pantheon_1
level3_all_occ	wikidata_code

The dataset includes a vast amount of variables (49) for each individual These variables display information relating to:

- Basic Demographic information: variables such as `name`, `birth`, `death`, `gender`, as well as geography-related variables such as `birthplace_name`, `deathplace_name`. Other variables, including `bplo1`, `dplo1`, `blpa1`, and `dpla1`, correspond to the longitude and latitude of these names (with `bp`, `dp` indicating birthplace/deathplace and `lo1`, `la1` indicating longitude/latitude).
- Domain of Influence: three sets of variables that are categorized hierarchically, corresponding to their predominant occupation or reason for prominence. `level1` variables correspond to one of 6 broad domains (e.g. culture, sports), `level2` variables correspond to fifteen subdomains (e.g., mathematician, politician), and `level3` variables corresponding to the specific occupation or domain of influence.
- Wikipedia-related Metrics: variables such as `number_wiki_editions` (number of different Wikipedia editions), `non_missing_score` (number of non-missing items from Wikipedia/Wikidata relating to demographic information), `total_count_words` (number of words in all biographies from Wikipedia), and `wiki_readers_2015_2018` (average yearly number of readers from 2015-2018)

  There are also some variables that were constructed by the researchers, including variables which measure prominence. Alternate rankings of prominence instead of `ranking_visib_5_criteria` include `sum_visib_ln_5criteria`, which is the sum of the log of the 5 variables plus one (the 5 variables used for this ranking are mentioned in Section 2.2).

C.2.3 Sampling of the Dataset

The researchers defined the sample of notable individuals by using the "instance of individuals" category present on Wikidata.

C.3 Strengths of Dataset

- Robust data-verification: the researchers employed a variety of algorithms and methods to make sure that the data was accurate. These checks are also quite essential, as encyclopedias on hundreds of thousands of individuals is bound to have potential errors.
- Comprehensive: The inherent goal of the study results in numerous different regions, time periods, and domains of influence being included. Thus, this large scope allows the dataset to be quite broad and applicable to many different situations.
- Diverse sources: The researchers obtained information on prominent individuals using 7 different Wikipedia language editions, which reduces Anglosphere bias and improves representation.
- Numerous amount of variables: The large amount of variables (49) not only allows the opportunity for many different types of analysis, but also highlights the large amount of information obtained for most individuals in the dataset.

C.4 Limitations of Dataset

C.5 Potential Enhancements for Measuring Prominence

C.5.1

C.5.2

References

Arel-Bundock. 2022. "Modelsummary: Data and Model Summaries in r." *Journal of Statistical Software*. https://doi.org/10.18637/jss.v103.i01.

Banaji, M. R., and A. G. Greenwald. 1995. "Implicit Gender Stereotyping in Judgments of Fame." *Journal of Personality and Social Psychology* 68 (2): 181–98. https://doi.org/10.1037//0022-3514.68.2.181.

Bengtsson, Henrik. 2023. *R.utils: Various Programming Utilities*. https://CRAN.R-project.org/package=R.utils.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Laouenan, Morgane, Palaash Bhargava, Jean-Benoît Eyméoud, Olivier Gergaud, Guillaume Plique, and Etienne Wasmer. 2022. "A Cross-Verified Database of Notable People, 3500BC-2018AD." *Scientific Data* 9 (1): 290. https://doi.org/10.1038/s41597-022-01369-4.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Stamarski, Cailin S., and Leanne S. Son Hing. 2015. "Gender Inequalities in the Workplace: The Effects of Organizational Structures, Processes, Practices, and Decision Makers' Sexism." *Frontiers in Psychology* 6 (September): 1400. https://doi.org/10.3389/fpsyg.2015.01400.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023. *Httr: Tools for Working with URLs and HTTP*. https://CRAN.R-project.org/package=httr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.