Analyzing Trends on the Characteristics of Notable People on Wikipedia*

Time Period and Occupation Strongly Associated with Individual Prominence

Parth Samant

December 14, 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 occupation and time period were most strongly associated with prominence, with subregion and gender being less significant. This highlights how prominence between already notable people selects for different characteristics.

Table of contents

1 Introduction **1**

2 Data **2**
2.1 Overview . . . 2
2.2 Measurement . . . 3
2.3 Data Cleaning and Variables . . . 3
2.4 Analysis of Variables . . . 5
2.4.1 Subregion . . . 5
2.4.2 time_period/years_since_birth variables . . . 6
2.4.3 Time Period and Subregion . . . 6
2.4.4 Gender, Occupation, and their Interaction . . . 7

3 Model **9**
3.1 Overview . . . 9

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

3.2 Model set-up . . . 9
3.3 Checking Model Assumptions and Validation of Final Model . . . 10
3.3.1 Validation of Final Model . . . 12
3.4 Limitations . . . 13

4 Results **13**
4.1 modelsummary() results . . . 13
4.2 Prominence and Gender . . . 15
4.3 Prominence and Time Period . . . 15
4.4 Prominence and Subregion . . . 16
4.5 Prominence and Occupation . . . 17
4.6 Overall Model Predictive Ability . . . 17

5 Discussion **20**
5.1 Notable Relationship between Time Period and Prominence . . . 20
5.2 Geographic Subregion and Prominence . . . 20
5.3 Similar Prominence Between Notable Men and Women . . . 21
5.4 Prominence by Occupation/Domain of Influence . . . 21
5.5 Discussion of modelsummary() results . . . 22
5.6 Weaknesses . . . 22
5.6.1 Results Only Relevant for Those Already Considered "Notable" . . . 22
5.6.2 Unequal Prevalence of Characteristics Affects Model Prediction . . . 22
5.6.3 Limited Prediction Ability . . . 23

Appendix **24**

A Model details **24**
A.1 Bayesian Information Criterion (BIC) . . . 24
A.2 Variance Inflation Factor (VIF) . . . 25
A.3 RMSE Comparison . . . 25
A.4 Full modelsummary() Results . . . 26

B Exploration of Notable People Dataset and Methodology **26**
B.1 Introduction . . . 26
B.2 Data Source and Collection . . . 26
B.2.1 Data Verification . . . 28
B.2.2 Variables of the dataset . . . 28
B.2.3 Sampling of the Dataset . . . 29
B.3 Strengths of Dataset . . . 29
B.4 Limitations of Dataset . . . 30
B.5 Potential Enhancements for Dataset . . . 30

References **32**

1 Introduction

Wikipedia is an online encyclopedia has been one of the most popular references of information globally. Containing milllions of articles on the platform, and thus acting somewhat as an archive for human information. This encyclopedia provides information on a diverse range of topics from history to entertainment. Within this vast encyclopedia, there is also a collection of biographical information on "notable people" who have ever lived in the past few thousand years. 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 geographical subregion, gender, occupation, and time period (at birth). By analysing these factors, this paper aims to uncover potential bias in characteristics of notable people and contribute to a deeper understanding of factors connected to prominence on Wikipedia.

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.

The findings of this paper showed that an individuals time period and occupation/field have a large effect on an individuals prominence. Those who were born in earlier time periods (such as before 500 BCE) were associated large increases in prominence. Additionally, those known for being part of a noble class, an important family, or for creative endeavors that shape culture (such as musicians, actors, or painters) also tend to have higher overall prominence.Subregion was somewhat associated with prominence, but with a large amount of variability. Notable individuals from Eastern Asia tend to have high levels of prominence, whereas those from Central America and Southern Africa have less. It was also found that within those who are already prominent, gender tends to have a somewhat insignificant effect.

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.

The remainder of this paper is structured as follows. Section 2 mentions features of the dataset, Section 3 introduces, explains, and justifies the constructed model, Section 4 summarizes the results of the paper, and Section 5 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 help of the packages `tidyverse`(Wickham et al. 2019), `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` (Kuhn and Wickham 2020), 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 B.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 B, 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, filtering out rows with missing information, and then randomly selecting 20,000 notable individuals. The variables selected are mentioned later on in this section.

Since this paper focuses on predicting historical prominence, I initially chose variables which I thought could potentially 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 prominence/notability ranking. A higher percentile indicates they were a more notable person.

Original 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 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 occupation (or possibly field/title) that the individual is known for.
  - Variables starting with "culture" refer to those who are known for producing immaterial goods
    - the 'culture-core' title refers to occupations that tend to be more artistic and creative in nature. These may include artists, writers, painters, or musicians (among others). Their contributions are often essential at shaping cultural values and aesthetics.
    - The 'culture-periphery' title refers to occupations that are still artistic/creative, but tend to include more commercial and applied work. These may include journalists, models, architects, or designers.
  - Other titles such as "family" and "nobility" highlight that they are predominanly known for being associated with an influential family or possessing a hereditary title in an aristocracy.

- titles such as "Worker/Business (small)" refer to miscellaneous jobs that are not covered by other variables. These may include those who are well-known farmers, librarians, booksellers, etc.
    - This is different than "Other", which mostly includes those known for negative notorious acts (such as a criminal or serial killer).
- a "Missing" occupation indicates that it was unable to be retrieved by the researchers.

2.4 Analysis of Variables

2.4.1 Subregion

Table 1 contains information on the different subregions as well as the proportion of notable people for each subregion.

Table 1: Proportion of Notable People On Wikipedia by Geographical Subregion

subregion	Proportion
Western Europe	0.4436961
Northern America	0.1545487
Southern Europe	0.0899025
South America	0.0611766
Northern Europe	0.0587785
Eastern Europe	0.0486760
Eastern Asia	0.0355630
Oceania Western World	0.0246441
South Asia incl. Indian Peninsula	0.0184193
Western Asia (Middle East Caucasus)	0.0157661
Central America	0.0108169
SouthEast Asia	0.0086739
West Africa	0.0059697
Caribbean	0.0049492
Southern Africa	0.0048982
East Africa	0.0044390
North Africa	0.0040818
Central Africa	0.0024491
Central Asia	0.0021430
Oceania not Aus Nze	0.0004082

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 100 times the number of notable people than West Africa.

There could be many explanations for the over representation of Western countrise, 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 variables

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. These were both originally included to highlight the influence of not only the time period, but the more specific year in which an individual as born.

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

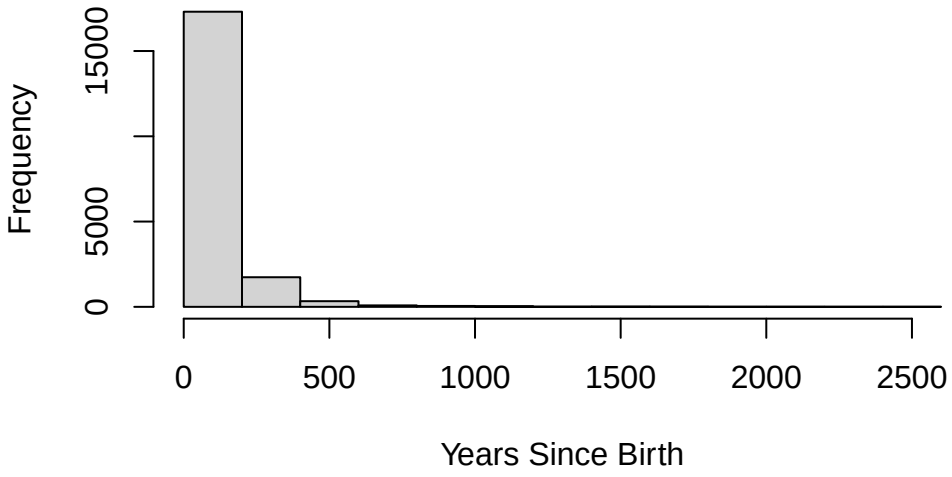


Figure 1: Distribution of Notable Individuals by Years Since Birth

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.

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.

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 and show overall trends.

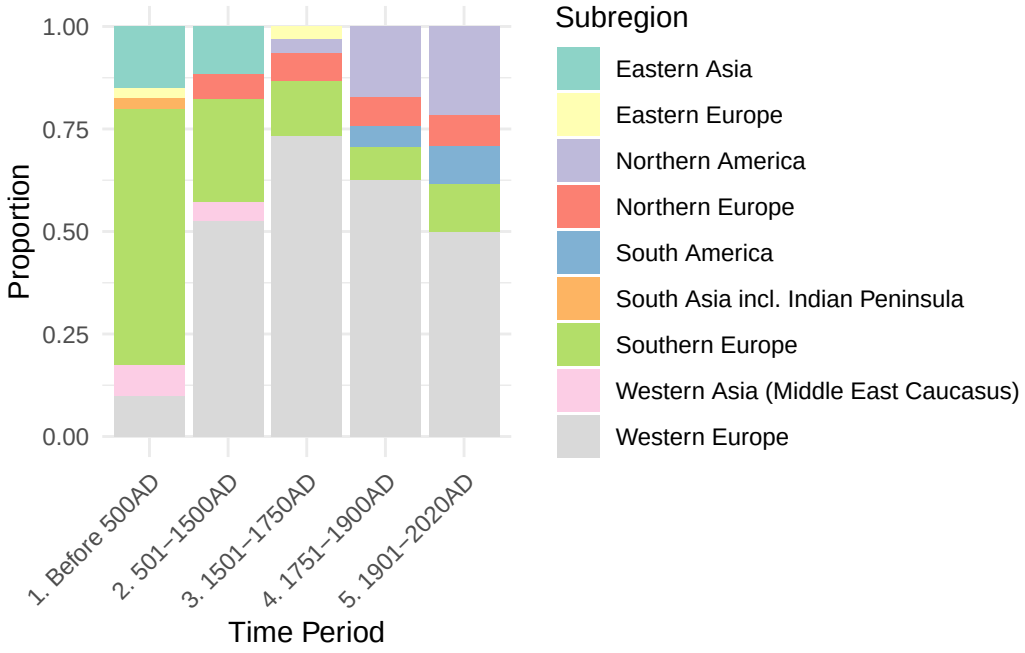


Figure 2: Top 5 Subregions of Notable People by Time Period. The graph illustrates the large variability of prominent subregions through time.

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.

2.4.4 Gender, Occupation, and their Interaction

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 Table 2, which shows ones occupation (or more specifically, what they are known for).

Table 2: Occupations of Notable People by Gender, Showing High Proportion of Males

occupation	proportion_female	proportion_male
Family	0.30	0.70
Culture-core	0.23	0.77
Nobility	0.23	0.77
Culture-periphery	0.23	0.77
Other	0.20	0.80
Worker/Business (small)	0.16	0.84
Sports/Games	0.14	0.86
Missing	0.11	0.89
Politics	0.11	0.89
Academia	0.09	0.91
Corporate/Executive/Business (large)	0.08	0.92
Administration/Law	0.06	0.94
Explorer/Inventor/Developer	0.04	0.96
Religious	0.04	0.96
Military	0.03	0.97

Table 2 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' (the family title, which means they are notable for being a part of an important family) 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).

Table 3 reinforces the phenomenon of notable people on Wikipedia being overwhelmingly male. In fact, the proportion of males is roughly 85%.

Table 3: Proportion of Notable People by Gender, Showing Male Over Representation

Count	female_count	male_count	proportion_female	proportion_male
19599	2905	16694	0.1482218	0.8517782

Furthermore, Figure 3 provides a visualization of the most common occupations in the dataset.

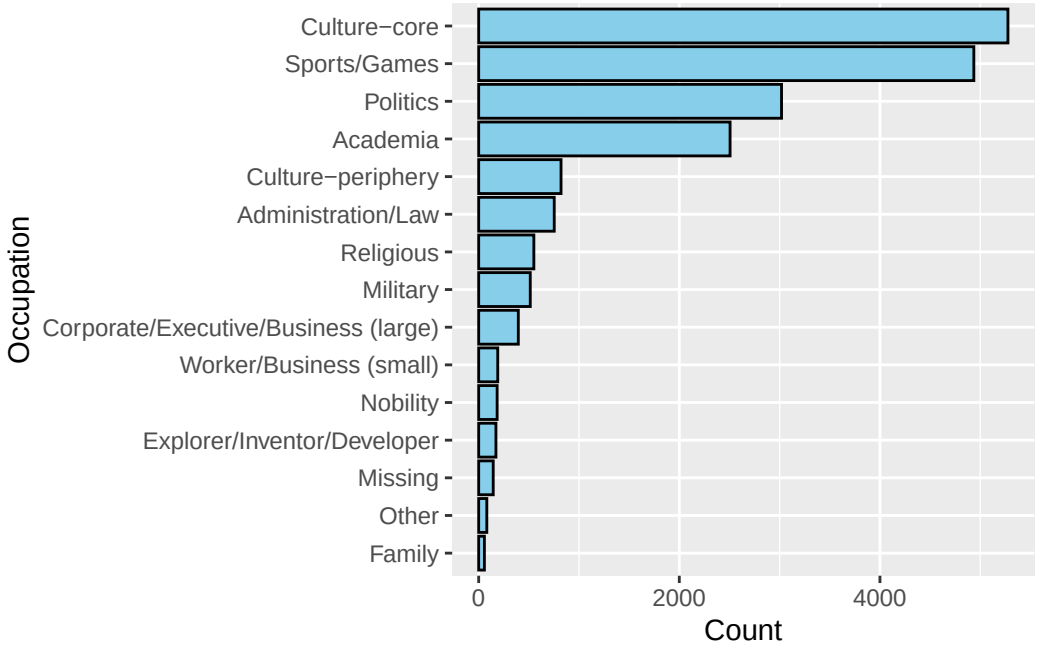


Figure 3: Number of Notable People by Occupation, Highlighting an Unequal Distribution

Evidently, some occupations tend to be very over-represented, such as Sports/Games, Culture-core (refer to Section 2.4), or Politics. This highlights that notable people on Wikipedia are most often associated with only a handful of occupations/categories.

3 Model

3.1 Overview

The ultimate goal the modelling strategy is to identify which characteristics are most associated with prominence. The model was split up into testing and training sets with the help of the `tidymodels` (Kuhn and Wickham 2020) package.

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

3.2 Model set-up

The model is expressed by the equation:

$$\text{prominence percentile} = \beta_0 + \beta_1 \cdot \text{occupation} + \beta_2 \cdot \text{subregion} + \beta_3 \cdot \text{time period} \quad (1)$$
$$+ \beta_4 \cdot \text{gender} + \epsilon \quad (2)$$

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 effect of ones gender on prominence percentile ranking
- ϵ : residual term, or the variation in percentile_ranking not due to the predictors.

The model was ran in the R programming language (R Core Team 2023) using the base `lm()` function, with the help of the `MLmetrics` and `broom` packages for model evaluation (shown in Section A of the appendix).

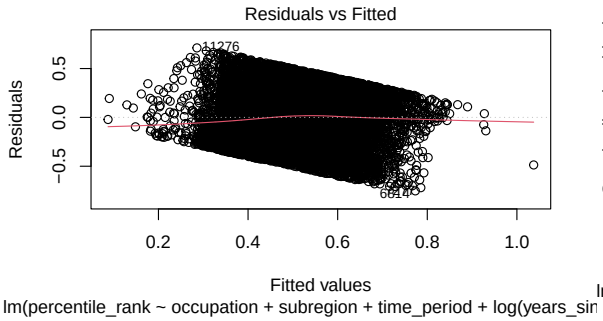
Many of the initial modelling decisions are reflective of the findings from the data section (in Section 2.4). However, these were later removed after performing various statistical tests. The initial decisions included:

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

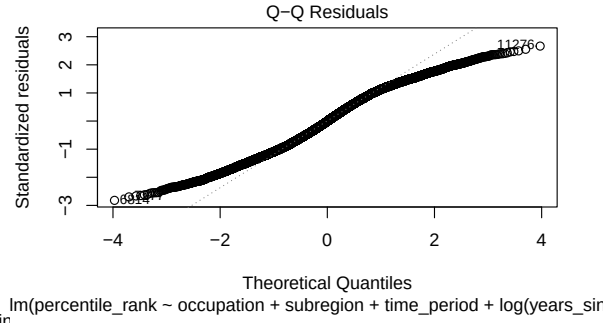
The equation of the full model, which includes these initial decisions, is shown in Section A.

3.3 Checking Model Assumptions and Validation of Final Model

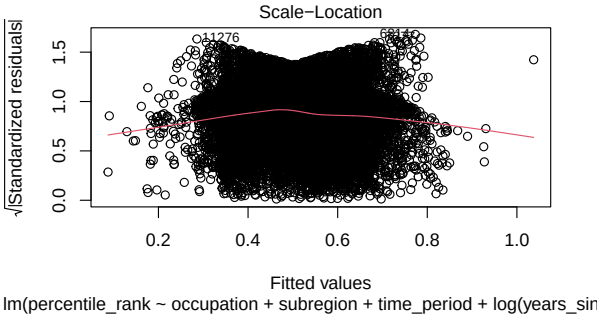
Linear regression relies on assumptions that should be verified before drawing inferences on the results of the model. These assumptions include a linear relationship between the dependent variable and independent variables, homoscedasticity (constant variance of the residual), normally distributed standard error (residuals), and no multicollinearity. Figure 4 provides graphs that checks the first three assumptions, with checks for multicollinearity being done in Section A.2.



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



(b) QQ-plot of Standardised Residuals: Shows Roughly Normally Distributed Standardised Error (Residuals)



(c) Scale-location plot which shows homoscedasticity (Approximately Constant Variance)

Figure 4: Compilation of Graphs Verifying Assumptions of Linear Regression

Figure 4 provides a series of graphs that can help assess the few assumptions mentioned above. These plots show:

- linearity: the residuals vs fitted plot shows roughly uncorrelated residuals (evidenced by the red line) However, there is a slight decrease in fitted values closer to 0 and a slight

increase in fitted values close to 1. This is possibly a consequence of how percentiles are not entirely normally distributed. Although the residuals appear to decrease as the fitted values increase, the trend line shows that the average value of the residual remains relatively constant.

- Homoscedasticity (constant variance): the scale-location plot evalutes the homoscedasticity of the model. The red line (corresponding to the averages of the fitted values based on standardised residuals) is roughly horizontal.
- Normally distributed standardised residuals: The QQ-plot shows that the standardised residuals are roughly normally-distributed, with a somewhat heavy tail for more extreme values.

Furthermore, checks for multicollinearity were done to ensure that proper significance was identified for coefficients. As mentioned before, the specific checks are found in Section A.2. This motivated the removal of the `years_since_birth` term - contributing to less multicollinearity and more accurate conclusions about coefficient significance, especially related to time period.

3.3.1 Validation of Final Model

A series of statistical tools and techniques were used to aid in the ultimate goal of this model, which was to identify which characteristics correlate the most to individual prominence on Wikipedia. These tools were used in conjunction with tests of linear regression assumptions and analysis of the Variance Inflation Factor (VIF).

3.3.1.1 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 over fitting while retaining the goodness of fit (Stoica and Selen (2004)).

This statistical metric was used as an aid in determining which predictor variables to remove and which variables to include in the final model. The results of the BIC on the full model encouraged removal of the interaction variable between gender and occupation. More details on how this method was used is found in the appendix (Section A.1).

3.3.1.2 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 0.27, which corresponds to an average error of 27 percentile ranks.

3.3.1.3 Out-of-sample testing

Further validation of the model was done through out-of-sample testing. Training data on different models was used to generate the predictive abilities of different models. Then, based on their predictions of the testing data, comparisons of the RMSE of different models indicated that the chosen model, with less complexity, tended to perform roughly as well as the full model. More details on the different models tested and their comparisons can be found in Section A.3.

3.4 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. This is also prevalent in other characteristics, such as occupation. This limitation of the model also has consequences for the results in Section 4.

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.

4 Results

4.1 modelsummary() results

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 most significant characteristics that affect Wikipedia prominence. Aside from the time period, the most significant coefficients of the model were occupation-related.

Evidently, terms that are related to occupation and time period tend to the most correlated with Wikipedia prominence. Those involved in politics, big business, as well as those without a listed occupation are associated with the largest difference in prominence. Additionally,

Table 4

	(1)
occupationMissing	−0.339
	p = <0.001
occupationPolitics	−0.176
	p = <0.001
occupationCorporate/Executive/Business (large)	−0.176
	p = <0.001
occupationAdministration/Law	−0.169
	p = <0.001
occupationWorker/Business (small)	−0.163
	p = <0.001
subregionSouthern Africa	−0.150
	p = 0.001
subregionCentral America	−0.149
	p = <0.001
occupationMilitary	−0.147
	p = <0.001
genderMale	0.030
	p = <0.001
Num.Obs.	14 000
R2	0.139
R2 Adj.	0.136
AIC	2802.9
BIC	3142.5
Log.Lik.	−1356.433
F	52.259
RMSE	0.27

Output of modelsummary() for linear model, highlighting the most significant predictor variables as well as the effect of gender. “Significance” was determined by co-efficients that have a p-value less than 0.025 and a large effect on prominence.

time periods that are more modern tend to also experience a larger decrease in prominence (compared to individuals born before 500 BCE), which is explored with more detail in Section 4.3.

4.2 Prominence and Gender

Those who are male also experienced a small (and not statistically significant) boost in notability among those who are already notable. Thus, to explore this trend of slightly higher male prominence, Figure 5 provides a view of the model's predictions for prominence divided by gender.

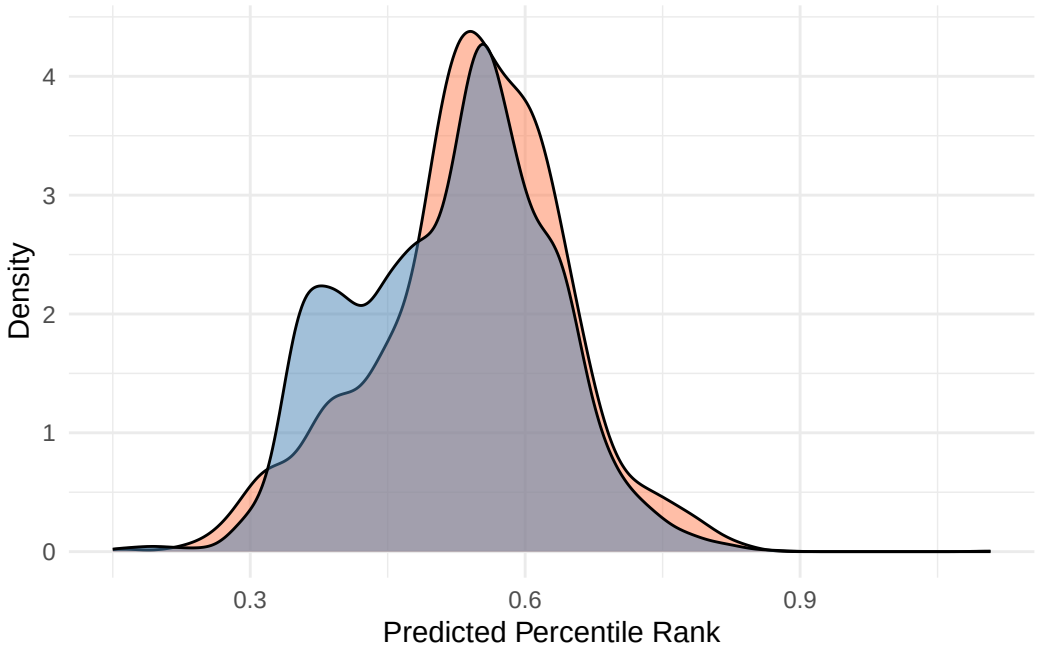


Figure 5: Densities of Gender-based Prominence Predictions

The distribution of predicted prominence by gender is roughly equal, with women being very slightly higher. Despite men experiencing slightly higher prominence (as shown in Table 4), the mean prediction of prominence favours women as shown in Table 5.

Table 5: Summary Statistics of Gender Prediction

gender	Mean Predicted Percentile	standard deviation
Female	0.5401590	0.1041685
Male	0.5190804	0.1064369

4.3 Prominence and Time Period

From the results of the `modelsummary()` function in Table 4, it is evident that time period tends to have a large effect on prominence. Thus, Figure 6 shows the effect of time period on model estimates with time periods categorised into five broad ranges.

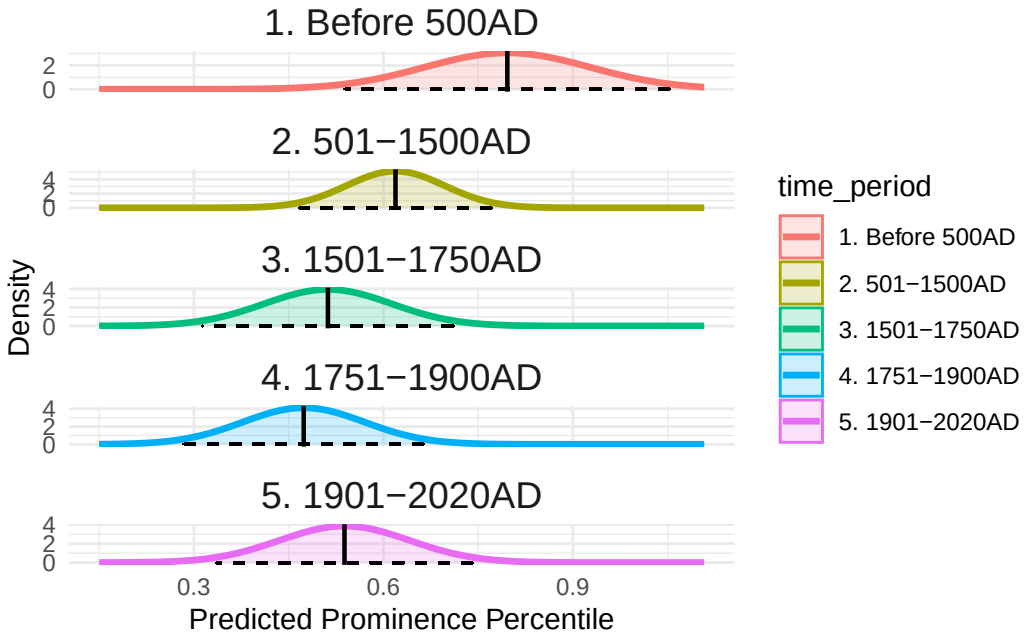


Figure 6: Model-Predicted Distributions of Percentile Rank by Time Period. Bars Shown On the Bottom Denote the Standard Deviation of the Prediction, With More Variability for Less Represented Time Periods (such as before 500 AD)

Figure 6 highlights the large variation in model predictions based on the time period the individual was born in. Although those born in earlier time periods tend to have significantly higher prominence, there were also a much lower amount of them.

4.4 Prominence and Subregion

The relationship between subregion and predicted prominence is explored in Figure 7.

Figure 7 indicates how much certain subregions are correlated with increased prominence percentiles, as well as 95% confidence interval bars that show the uncertainty associated with each coefficients effect. Notable people from Central America are associated with over a 10 percent reduction in percentile prominence (compared to the Caribbean, which is the reference). This is in stark contrast to those from Eastern Asia, with that subregion being associated with the largest increase in prominence.

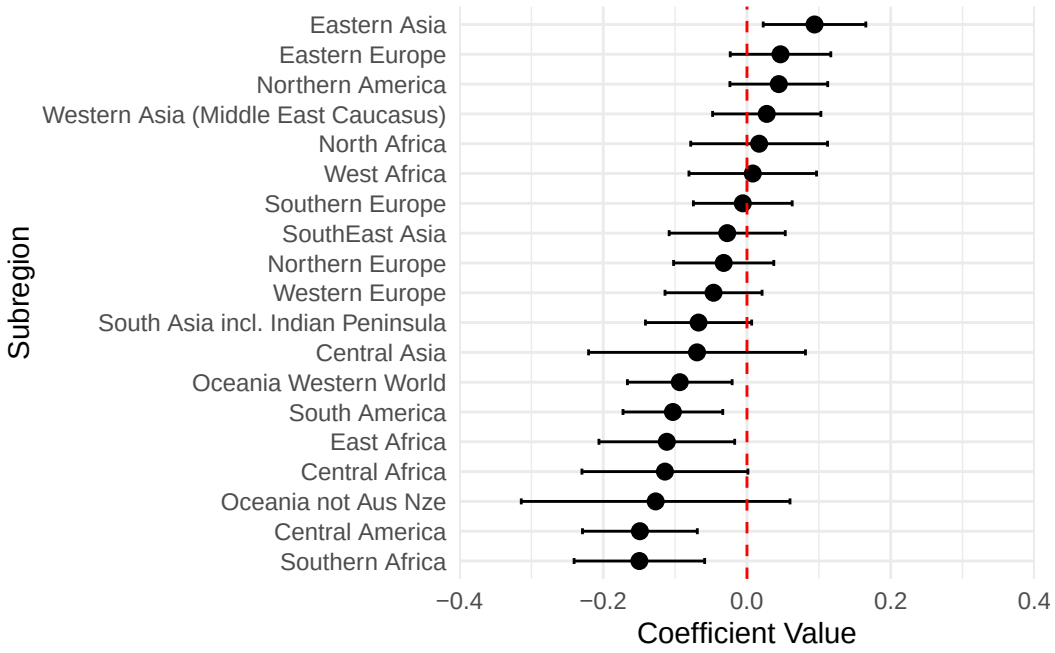


Figure 7: Barplot of Model's Coefficients Associated With Subregion With 95% Confidence Interval. Values Are In Comparison To Dummy Variable for Subregion (Carribbean)

4.5 Prominence and Occupation

Figure 8 compares the different effects that occupations (or domains of influence) may have on the prominence of an individual.

Figure 8 highlights a large variation related to prominence and occupation. Those with an occupation/domain of influence that was unknown/missing tended to have the lowest prominence (by far), whereas those identified as being part of "culture-core" experienced the highest. The large difference between occupations also highlights the importance of occupation on prominence.

4.6 Overall Model Predictive Ability

Furthermore, Figure 9 provides a visualization of the predictive ability of the model that was used. Note that the ultimate purpose of the model not to predict prominence accurately, but rather to highlight which characteristics were most associated with prominence among notable people.

Figure 9 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 the

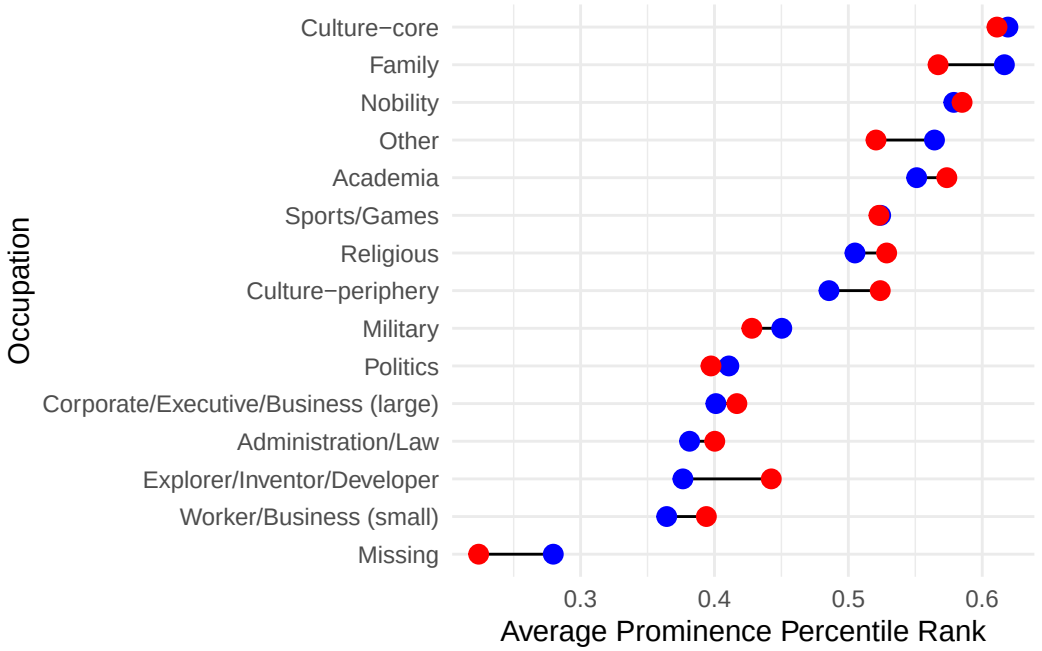


Figure 8: Comparison of Predicted (Red) vs Actual (Blue) Prominence Percentile Rank by Occupation. It illustrates the predictive ability of the model while also showing the effect of occupation on prominence.

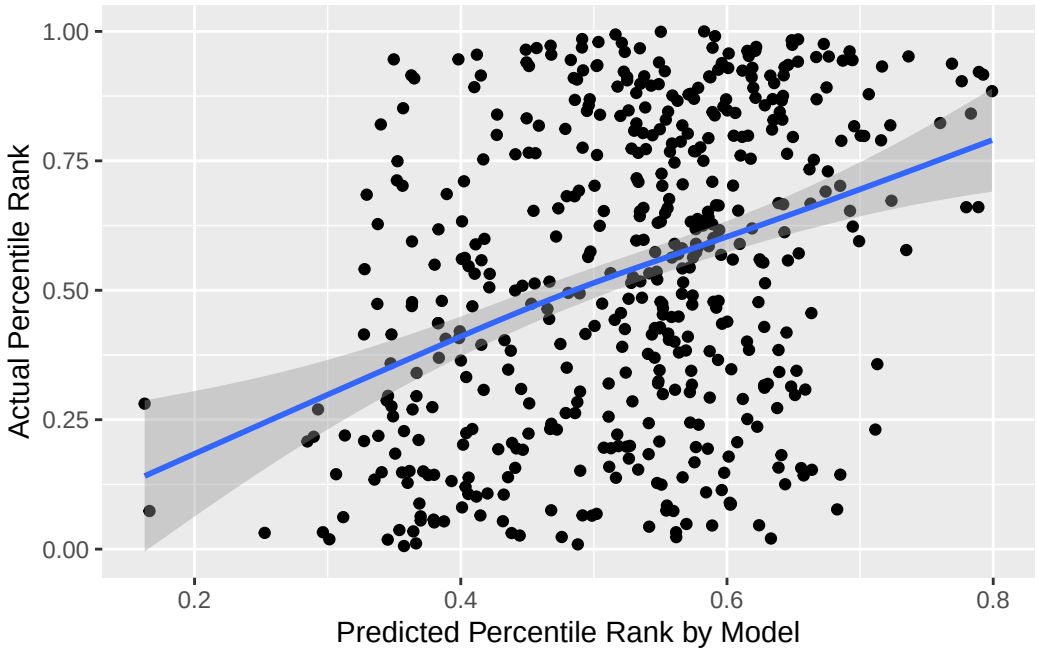


Figure 9: Graph comparing the model's predicted percentile of prominence vs. the actual percentile. Testing data (i.e., data the model was not trained with) was used to more accurately evaluate prediction ability. 500 random datapoints were selected to highlight the overall results.

mean predictions from the linear model. This illustrates a weak, but existing, relationship between the prediction given by the model and the actual percentile rank.

5 Discussion

5.1 Notable Relationship between Time Period and Prominence

Interestingly, time period tends to have quite a strong effect on prominence that is more significant than some 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 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. This is captured in Figure 6, which highlights an exponentially lower amount of people that are from earlier time periods.

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.

Next steps for future studies could include exploring the amount of documentation of notable people through time and the distribution, exploring the relationship between the passage of time and the loss of documentation of individuals. It may be possible that time periods after the 1500's still retain most documentation of notable people, resulting in less differences in prominence and a larger sample.

5.2 Geographic Subregion and Prominence

Using data from the World Bank ((n.d.)), 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 finding was supported by the results in Section 2.4, which suggested a significant overrepresentation of those from Europe and Northern America.

Figure 7 provides an overview on various coefficients for subregion and how they may affect prominence. 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 (as mentioned later-on in Section 5.6 , it is evident that those from Eastern Asia overall have a higher prominence on Wikipedia.

Furthermore, the large range associated with the majority of subregions suggests that an individuals subregion, overall, does not tend to be too significant of a predictor variable. Nearly

every subregion, other than a few, contain a confidence interval that includes 0 (suggesting a notable chance that the difference may be negligible).

Larger ranges in confidence intervals also directly correlate to how much available data there is on the specific subregion, such as the large confidence interval range from Oceania (not including Australia/New Zealand). This comes as no surprise, as the dataset has the lowest proportion of notable individuals out of any subgroup (as shown by Section 2.4.1 in the data section).

5.3 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. The proportion of notable men compared to women, as shown in Section 2.4, reflects this bias - with men being very much overrepresented in this dataset.

However, when focusing exclusively on those who are already notable, the findings of this paper do not support previous studies regarding the relationship of gender on prominence. In fact, the results relating to gender (shown in Figure 5) highlight that notable women tend to be slightly more prominent than notable men - although not by much. 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. The insignificance of the `gender` variable is supported by the results in `modelsummary()` which indicate a very small positive difference in prominence with being a male, but with a very large p-value (suggesting that this difference could very well be due to chance).

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.

5.4 Prominence by Occupation/Domain of Influence

Prominence tends to be strongly associated with an individuals occupation, as shown by Figure 8 and Table 4. In fact, the difference based on occupation is so large that the groups with most prominence (such as "culture-core" or "family") are, on average, 30 percentile points more prominent than those with a category of "missing" (as shown in Figure 8). Unsurprisingly, those with "missing" occupations tend to be much less prominent overall. It is quite intuitive

that a lack of information on an individual correlates to lower overall prominence. Additionally, like other variables, `occupation` conserves the same trend of more popular categories having a more accurate model prediction.

Furthermore, the overall prominence based on occupation may correlate to the exclusivity of the respective occupation/group. For example, those belonging to 'nobility' or 'family' may tend to be more prominent as belonging to the noble class or belonging to a powerful family is inherently a very small and elite group of people. This could be compared to a less prominent occupation, such as "Worker/Business(small)", which often includes those who are farmers, librarians, or booksellers - a much less elite occupation.

5.5 Discussion of modelsummary() results

The results from the `modelsummary()`, which summarize the model, are shown in Table 4. As mentioned in Section 5.4, it makes intuitive sense that an occupation of "missing" would correspond to a very large overall decrease in prominence.

Interestingly, some occupations such as being involved in the military; being a politician; and being associated with administration/law 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, 1501-1750AD, and 1901-2020 AD tend to have a much lower amount of prominence than those born before 500 AD (as mentioned in Section 5.1).

5.6 Weaknesses

5.6.1 Results Only Relevant for Those Already Considered "Notable"

The correlations found in this paper only take into account people who are ALREADY already prominent. [Section 2.4 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. This is why, for example, being a male was not significantly correlated with prominence.

5.6.2 Unequal Prevalence of Characteristics Affects Model Prediction

Those with characteristics that are underrepresented in the dataset tend to have less accurate predictions of popularity from the model. This is mentioned in the model section (Section 3.4) but has downstream consequences for the results, especially when analysing less represented subregions or occupation.

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 the number of prominent individuals 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 (such as being born before 500AD).

This also extends to occupations, as highlighted in Figure 8. The difference between the prediction of the model and the actual percentile rank are heavily correlated with the distribution of occupations (as shown in Figure 3). Limited entries of certain occupations, such as "other", "missing", and "family" tend to show a less accurate model prediction.

5.6.3 Limited Prediction Ability

From Figure 9, 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 Table 4. The R^2 value of 0.121 suggests that these predictors only explain around 12% 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 Figure 7, Figure 6, and Figure 5). 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 Model details

A final model was intended to be chosen after performing numerous tests on a full model. The full model was as follows:

$$\begin{aligned}\text{prominence percentile} = \beta_0 + \beta_1 \cdot \text{occupation} + \beta_2 \cdot \text{subregion} + \beta_3 \cdot \text{time period} \quad (3)\\ + \beta_4 \cdot \log(\text{age}) + \beta_5 \cdot \text{gender} + \beta_6 \cdot (\text{time period} \times \log(\text{age})) \quad (4)\\ + \beta_7 \cdot (\text{occupation} \cdot \text{gender}) + \epsilon \quad (5)\end{aligned}$$

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 large amount of variables an interaction terms highlights the associations between different predictors, with their relationships being shown in Section 2.3. The ultimate motive for starting with a large model was to include the possibility of numerous different predictor variables being important for prediction. However, the final model was eventually chosen after a series of numerous statistical tests (including BIC, VIF, and RMSE).

A.1 Bayesian Information Criterion (BIC)

The method of BIC was an important tool used to remove unnecessary terms, and was the first step for obtaining the full model. This was preferred over a similar method called the Akaike information criterion, since the BIC is more harsh with an overly complex model (increasing the chance of possibly insignificant interaction terms being removed). BIC was performed with a backwards selection process (that goes from a more complex to less complex model) and removed the terms `gender` and `gender:occupation` (an interaction term). The

BIC helped with the decision to remove the gender/occupation interaction variable, but the `gender` variable was still retained to further analyse its significance in prominence.

A.2 Variance Inflation Factor (VIF)

VIF is a technique that is used to help address the linear regression assumption of multicollinearity (that is, correlations between different predictor variables). More specifically, it shows how much variance of an estimated regression coefficient increases due to multicollinearity. Higher GVIF values indicate potential issues with multicollinearity that could obstruct how significant a characteristic truly is. With the help of the `car` package (Fox and Weisberg (2019)), the VIF table of the updated model (with `gender:occupation` removed) is shown in Table 6.

Table 6: GVIF Values (and thus multicollinearity) High in Time-Related Predictors

	GVIF	Df	GVIF^(1/(2*Df))
occupation	1.993139e+00	14	1.024938
subregion	1.323678e+00	19	1.007407
time_period	2.746747e+11	4	26.906197
log(years_since_birth)	5.683210e+03	1	75.387069
gender	1.098870e+00	1	1.048270
time_period:log(years_since_birth)	4.530047e+11	4	28.642626

The high GVIF shown by `time_period` and `log(years_since_birth)` is expected as they both relate to an individuals time of birth. However, the decision to remove `log(years_since_birth)` was to see more directly how time period influenced prominence. This aided with the discovery that the time period has quite a significant result on prominence, as shown in Table 4.

A.3 RMSE Comparison

After obtaining a reduced model, the root mean squared error (RMSE) between different models was used to verify that the reduced model still retained predictive abilities. RMSE calculations were performed with the help of the MLmetrics (Yan (2024))

Table 7: Similar RMSE highlights Relatively Similar Predictive Ability Between Reduced and Full Model

Model	Root Mean Square Error
RMSE for Reduced Model	0.2689250

Table 7: Similar RMSE highlights Relatively Similar Predictive Ability Between Reduced and Full Model

Model	Root Mean Square Error
RMSE for Full Model	0.2690102

A nearly identical root mean square error suggests that the reduced model is roughly as good as the full model when it comes to predicted vs. actual values. This comes at the benefit of having less multicollinearity and having a less complex model. Thus, the reduced model was ultimately chosen.

A.4 Full modelsummary() Results

Previous summaries of the final model focused more on a handful of significant variables. However, Table 8 provides each characteristic and its influence on prominence of notable people. The table was generated with the help of the `kableExtra` package (Zhu (2024)).

B Exploration of Notable People Dataset and Methodology

B.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.

B.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).

Table 8: Full Results of the Model, Including Characteristics Not Strongly Associated With Wikipedia Prominence

	(1)
(Intercept)	−1.720
	(2.903)
occupationAdministration/Law	−0.169
	(0.013)
occupationCorporate/Executive/Business (large)	−0.176
	(0.017)
occupationCulture-core	0.021
	(0.008)
occupationCulture-periphery	−0.064
	(0.013)
occupationExplorer/Inventor/Developer	−0.131
	(0.026)
occupationFamily	−0.093
	(0.042)
occupationMilitary	−0.147
	(0.015)
occupationMissing	−0.339
	(0.029)
occupationNobility	−0.056
	(0.026)
occupationOther	−0.084
	(0.034)
occupationPolitics	−0.176
	(0.009)
occupationReligious	−0.054
	(0.015)
occupationSports/Games	−0.120
	(0.009)
occupationWorker/Business (small)	−0.163
	(0.023)
subregionCentral Africa	−0.114
	(0.059)
subregionCentral America	−0.149
	(0.041)
subregionCentral Asia	−0.070
	(0.077)
subregionEast Africa	−0.112
	(0.048)

B.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 B.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.

B.2.2 Variables of the dataset

Table 9: 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
area1_of_rattachment	area2_of_rattachment
list_wikipedia_editions	un_region
group_wikipedia_editions	bplo1

Table 9: Table of all variables present in the dataset

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).

B.2.3 Sampling of the Dataset

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

B.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.

B.4 Limitations of Dataset

- Western-centric bias: the researchers focused specifically on 7 popular Wikipedia languages, which are predominantly spoken by those in the West. Additionally, 30% of individuals in the dataset come from English Wikipedia, which skews the data towards Western countries.
- Gender-bias: Strong evidence of unequal gender representation exists in terms of notable people on Wikipedia. Although this bias of Wikipedia is not a fault of the researchers, this gender bias results in comparatively less extrapolation of dataset trends compared to men.
- Bias towards those who have internet access: the nature of obtaining notable people through an online encyclopedia, such as Wikipedia, likely may not include notable people in communities that do not have the same level of internet access. This is especially a concern in countries that do not have widespread internet usage.
- Recency-bias: Individuals who are from more modern time periods benefit from much more documentation and representation.

B.5 Potential Enhancements for Dataset

Although Wikipedia can give an idea on the overall trends relating to prominence, the limitations of this dataset have consequences for predicting prominence more generally. There are, however, some ways this can potentially be mediated. Not only do these improvements improve the ability to make inferences on prominence, but they also provide a more robust dataset. This improvements may include:

- including data from historical records or repositories that are more focused on women to address their under representation. An intentional effort to select women could mitigate the bias against them.

- including non-European langauges in order to gain a more comprehensive dataset of notable people that exist. This could also include cooperating with historians or figures that are knowledgeable on notable figures (which may not be included on Wikipedia).
- more comprehensive measures of prominence that do not just take in Wikipedia-related metrics. This may include interviews with experts, as well as social media mentions (for more modern figures) to compare overall prominence.

References

n.d. *World Bank Open Data*. https://data.worldbank.org.

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.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://www.john-fox.ca/Companion/.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles*. https://www.tidymodels.org.

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.

Stoica, P., and Y. Selen. 2004. "Model-Order Selection: A Review of Information Criterion Rules." *IEEE Signal Processing Magazine* 21 (4): 36–47. https://doi.org/10.1109/MSP.2004.1311138.

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/.

Yan, Yachen. 2024. *MLmetrics: Machine Learning Evaluation Metrics*. https://CRAN.R-

project.org/package=MLmetrics.
Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.