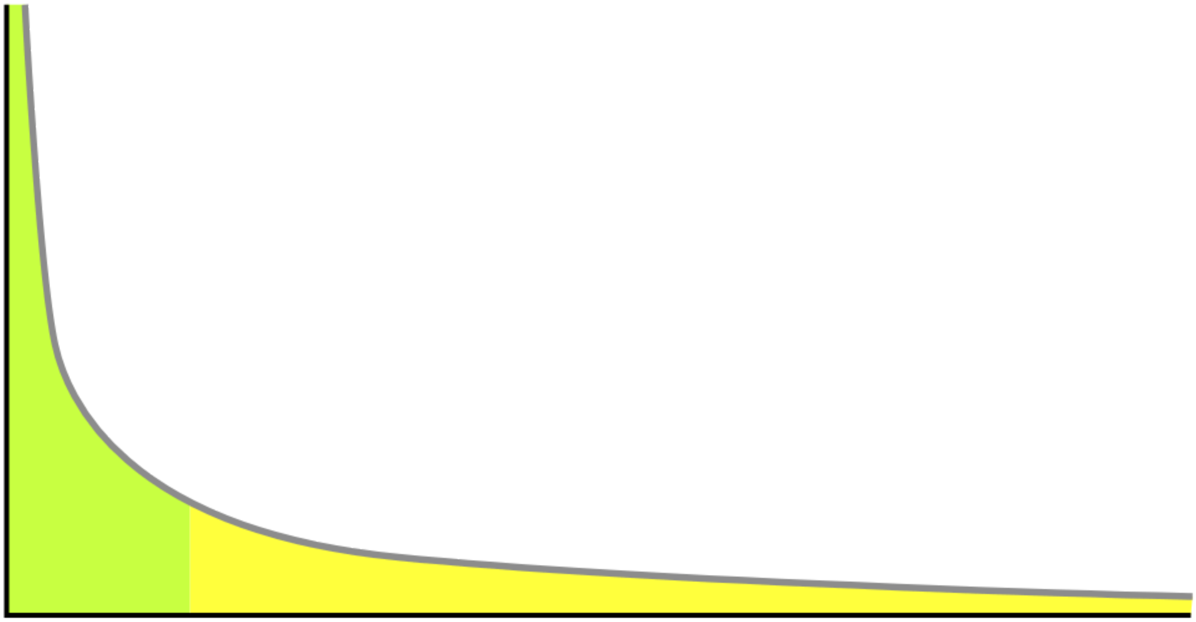
Identifying The Power Function Model in Common Data



Project by Vladimir Filtov, Abdullah Alrubian, Kristian Don

For Mathematical Modelling II

October 2024

Table of Contents

Case 1: Cities Populations in a Country

General Introduction and History

In 1949, George Zipf, a prominent linguist from Harvard University known for his work in statistical linguistics, discovered an intriguing pattern in urban populations. Zipf observed that if you rank cities and towns in any given country according to their population, the relationship between rank and population size follows a power law distribution. This empirical observation, now known as Zipfs' Law, suggests that the second-largest city is roughly half the size of the largest, the third-largest is about a third the size of the largest, and so on.

Zipfs' contributions extended beyond urban studies, as he also made significant advancements in understanding the distribution of word frequencies in languages.

Our source of the population data was worldpopulationreview.com, which means that in this project the countries, their population and the city count are all taken from the website. For that reason, the maximum city count per country is 400.

The Formula

b is the estimate for the population of the largest city
a is how fast the population declines ($a =- 1$ according to Zipf's law)

$y = bx^{a}$ is the initial power law equation

$ln(y) = ln(b) + ln(x^{a})$ we take the ln of the equation

$ln(y) = ln(b) + a * ln(x)$ transform it

$Y = B + aX$ get a linear version of it

We transformed $y = bx\hat{}a$, which is not linear, into $Y = B + aX$, which is linear. This allows us to use the following normal equations:

$$a \sum_{k=1}^{n} X_i^2 + B \sum_{k=1}^{n} X_i = \sum_{k=1}^{n} X_i Y_i$$

$$a \sum_{k=1}^{n} X_i + nB = \sum_{k=1}^{n} Y_i$$

These equations are from simple linear regression and are derived using the least squares method. They allow us to find a and b($b = e^B$) which are used to construct the model function.

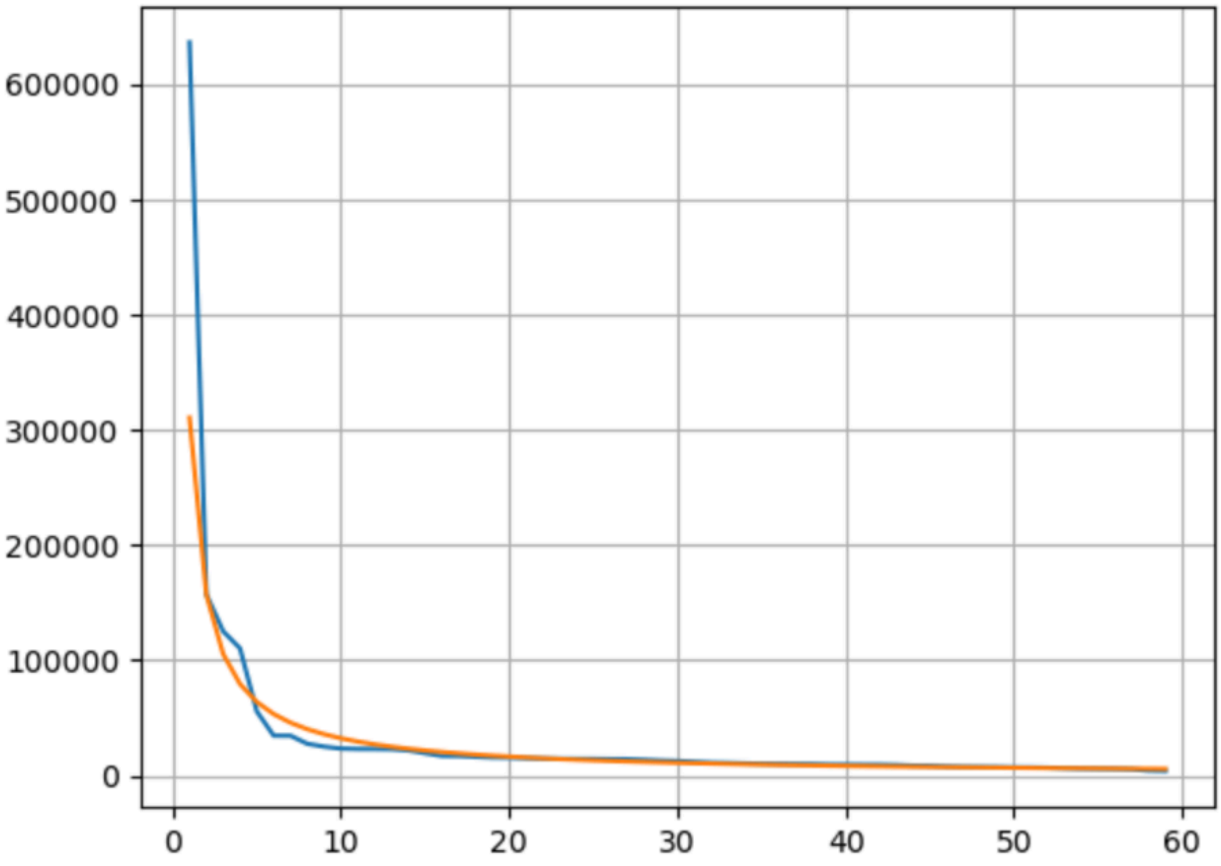
Moldova Example

Using the method explained earlier we can make a study on Moldova, there will be a marginal error which is generated by comparing the actual data with the generated model equation.

Calculation Steps

1. **Array of Values:** the file was imported from csv into a list and then was put in an (x,y) array, where the x is the rank, and y is the actual population of the city.
2. **Take The Ln:** the log of the data was taken, in which the first x value was 1 instead 0 because ln(0) is non-existent.
3. **B, a, b:** using the equations above we found B, a, and b to later use them in the model.
4. **Model Array:** the purpose of this step is to see if the data fits the model.
5. **Errors:** the importance of this step is to determine the reliability of the model array.
6. **Plotting:** the data is represented in a graph in which it shows the model generated using the data and the actual data.

Results



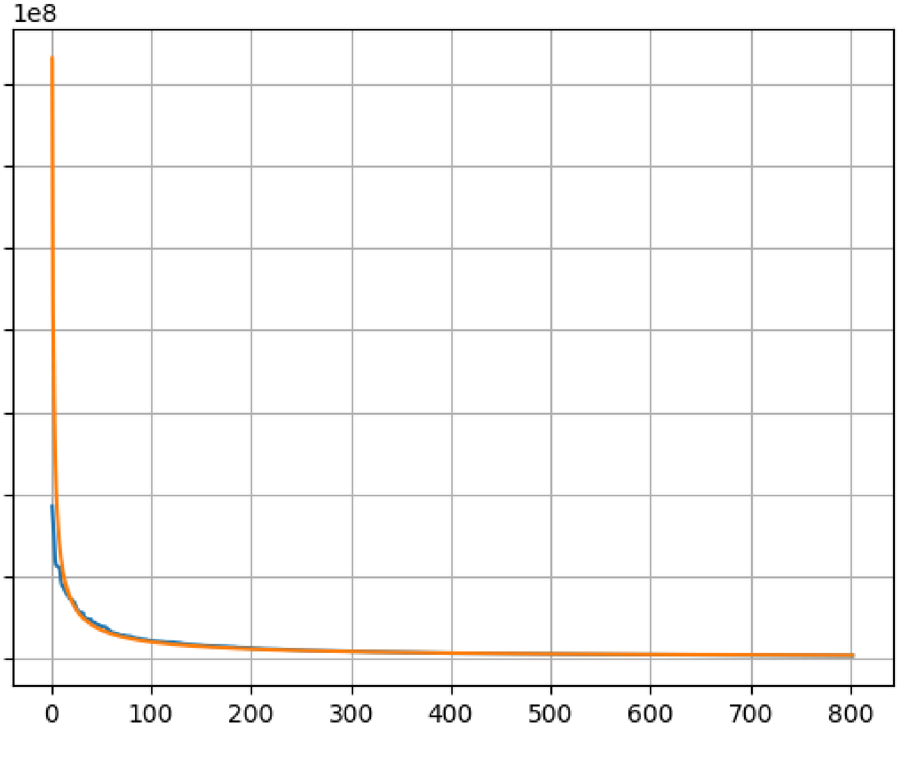
In the graph the blue line is the actual growth, and the orange line is the generated model. There is a clear difference between the estimate and the actual data. In this case the **B**, **a**, **b**, and **marginal error**, were calculated as (**12.64**, **-0.98**, **310329**, **47.93%**). While the estimate of the population growth is close to (-1) which is what Zipf law states, the marginal error is (47.93%) which shows that the formula can be accurate in certain aspects of a countrys' growth, and inaccurate in others.

Conclusions

In conclusion, Zipf law highly depends on the population of the country and the stability of its growth, this can result in either an accurate or inaccurate description of how fast the population declines, and the estimated population of the largest city. The external factors affecting the growth of each country are hard if not impossible to measure, in Zipf law it shows the estimate in accordance with the history of the country, which means that if the country has a fluctuating growth Zipf law would not accurately describe it, and the opposite is true.

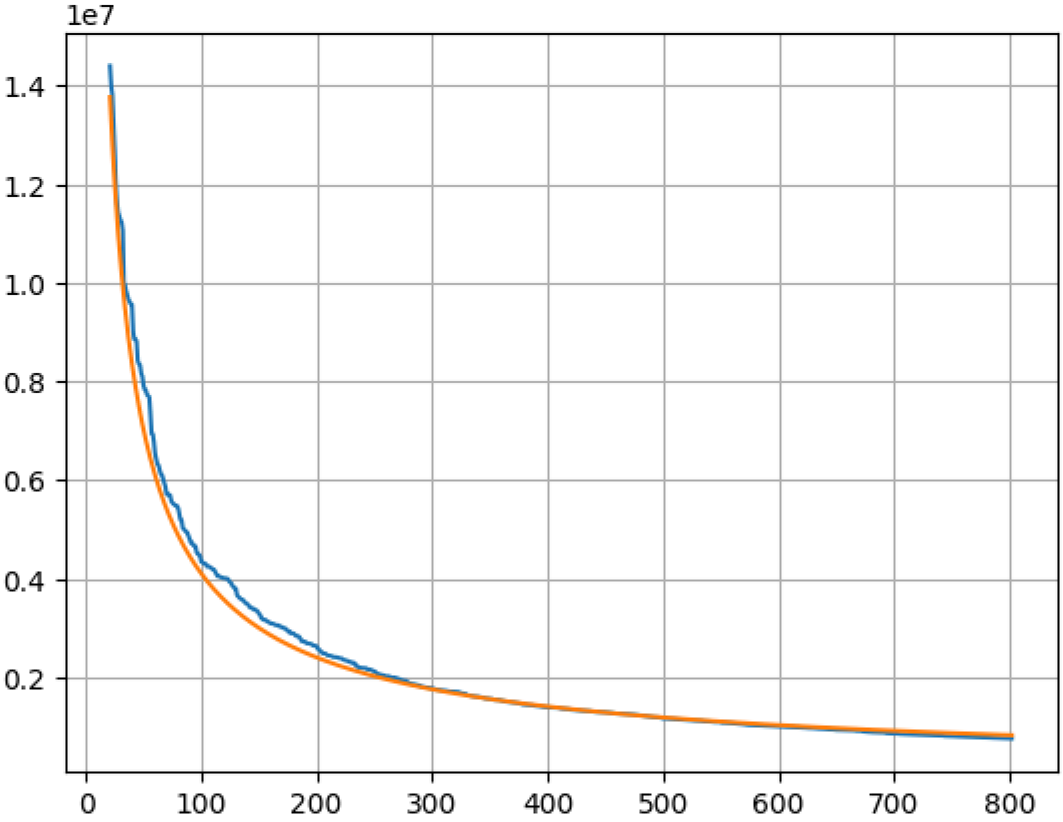
Case 2: Top Cities Populations

Results



58.47% error excluding tokyo

101.27% starting from the first one



Excluding the first 20

Conclusions

To conclude, we interestingly found that if the first 20 cities are excluded, the error found by the model is closer to the real error. By excluding just the first city the relative error decreased by 42.8%. The more cities we excluded the closer to the real error we got, up until about the first 20. With all the cities included the relative error reached a soaring 101.27%. This is the explanation behind our reasoning for excluding them from the graph and the calculations in this case. Based on the graph, it's evident that our model is very accurate, with the final relative error at a shallow 7.92%.

Case 3: Cities For Each Country

Collecting And Processing the data

Using the Selenium web scraping package in Python we first got a list of all the country names off the population website and then iterated through the URLs to fetch the city populations for each country. The result was a table with the first column storing country names and the rest storing population values for each city in the corresponding country from highest to lowest.

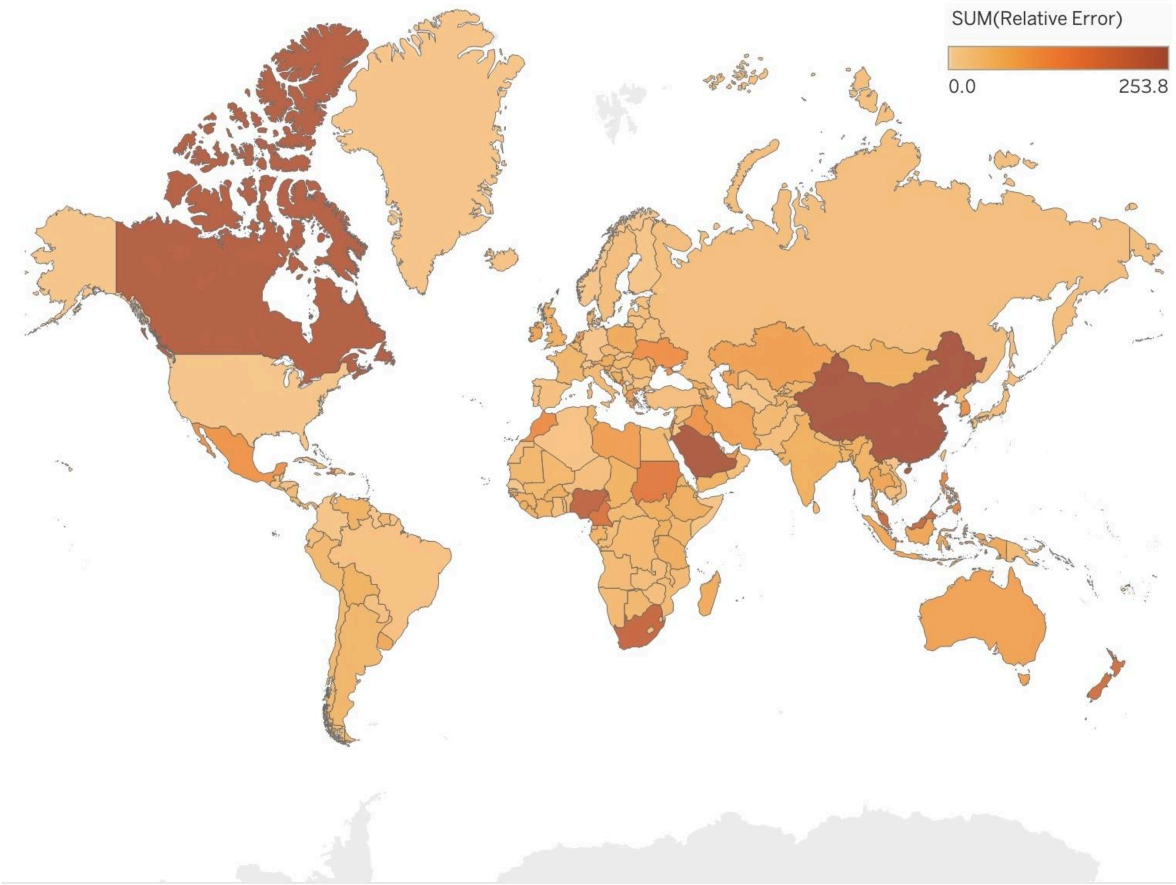
After that, using Pandas to transfer the data between Python and the device and Numpy to do the processing, we went through the calculation steps for each country as explained in detail in Case 1 to find B, a, b and the relative error. As an extra measure we also included the region and city count columns. To plot any particular country's model and real data arrays we developed a second table.

We did not include one city countries such as:

1. saint martin
2. american samoa
3. bermuda
4. guernsey
5. united states virgin islands
6. wallis and futuna
7. jersey
8. falkland islands
9. montserrat
10. gibraltar
11. curacao
12. turks and caicos islands
13. niue
14. singapore
15. french polynesia
16. british virgin islands
17. aruba
18. new caledonia
19. saint pierre and miquelon
20. guam
21. cook islands
22. faroe islands
23. northern mariana islands
24. saint barthelemy
25. tokelau
26. macau
27. isle of man
28. vatican city
29. cayman islands

Because if we have countries with only one city,
then their only x value is 1,
so their log(x value) is 0,
so their sum of log x values is 0,
so B is something divided by zero,
so B is infinity/undefined.
Therefore, we cannot include countries with only one city.

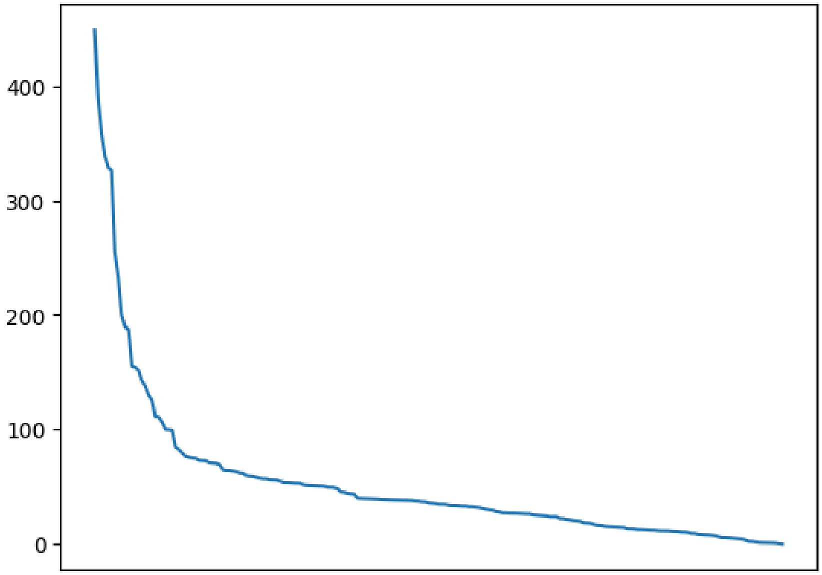
Error Map



Data Analysis

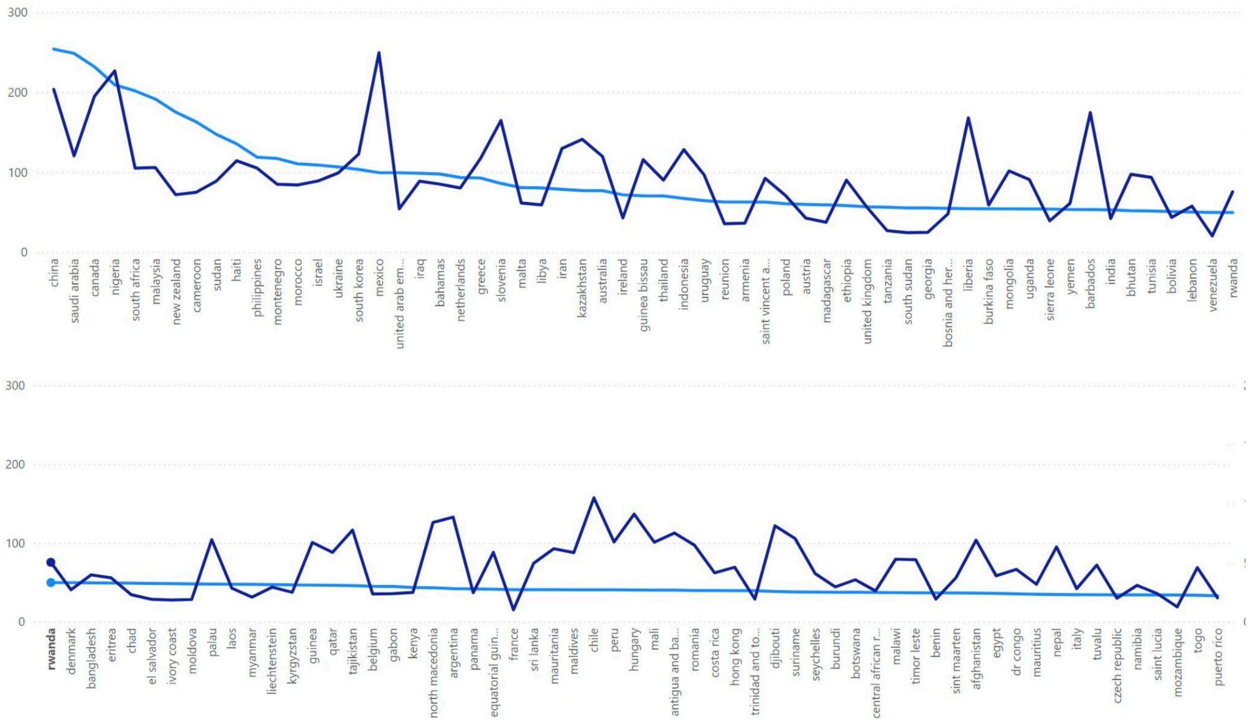
1. **Basic stats:** The average error for all countries is 47.3% and the variance is 36.6%. To get the relative variance we divided the variance by the mean.
2. **Correlation between error and city count:** The next thing we wanted to check was whether the number of cities in a country had an effect on the error. The results were not exactly what we were expecting, since there was no correlation between the two metrics, we plotted two line graphs with countries on the x axis and the error and city count on the y axis. Sorting the data from highest error to the lowest the city count was going up and down in no particular order. What was interesting though, was that the bottom 7 countries by error had 2 to 18 cities: the only legitimacy in the whole 205 countries. (This graph is included in the dashboard)

3. **Small b error:** the b value measures the estimate of the most populated city. To analyse how accurate the b value was for each country we took the percentage difference of the actual value and b. The average is 52.54% with the overall distribution looking something like this:



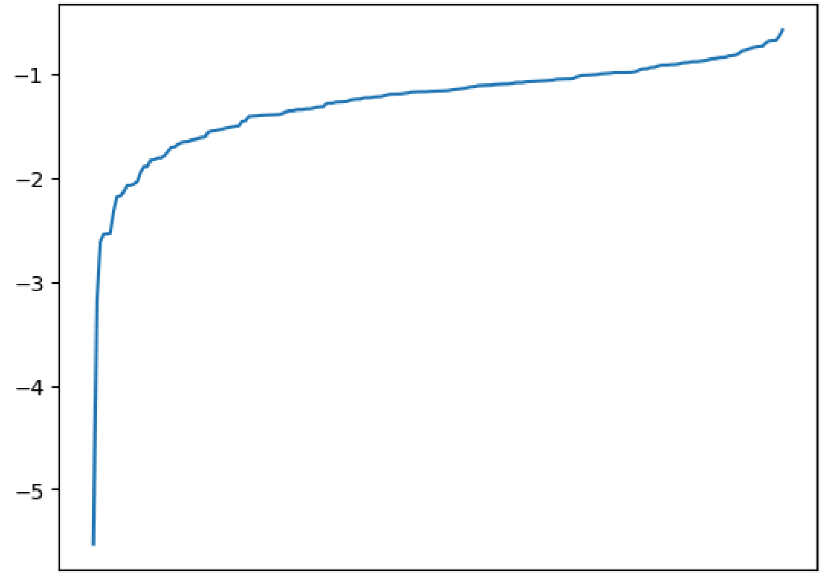
The top country was China with an error of 448%, then we can see a smooth decline after passing the 80-100 error mark. The data also shows that half of the countries have an error of 35% or less, with the lowest error belonging to Sao Tome And Principe and it is so small that Python registers it as zero.

Then why not start from the second largest city, since the error on the first one is so high? We compared the errors in normal data and data without the most populated city and here is the graph(half of all the countries)



It is visible that the capitalless dark blue line is usually higher than the normal data light blue line, which concludes that on average if we don't include the capital, then the error will only increase.

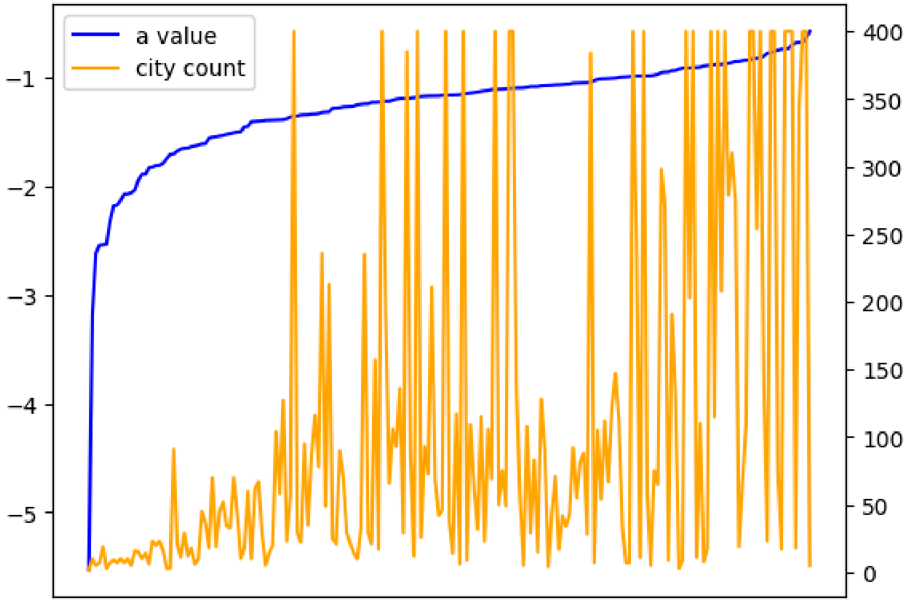
4. **Analysing the a value:** The countries with the largest (negative) a would be very capital-centric, having a large population in the capital (as a rule of thumb) and then a steep drop to the provincial centres. Here is the graph showing the a value from highest to lowest (negative):



Only 3 Countries have a value of less than -2.5 while the majority of the countries are in the [-2, -0.5] region. Zipf's law has the formula $P(n) \sim \frac{1}{n^{a}}$ where a is almost 1, which means that if we assume an error of 10%, then only 55 countries satisfy the formula. These countries have a city count range of around 20 to 100, so you can say they are "medium sized".

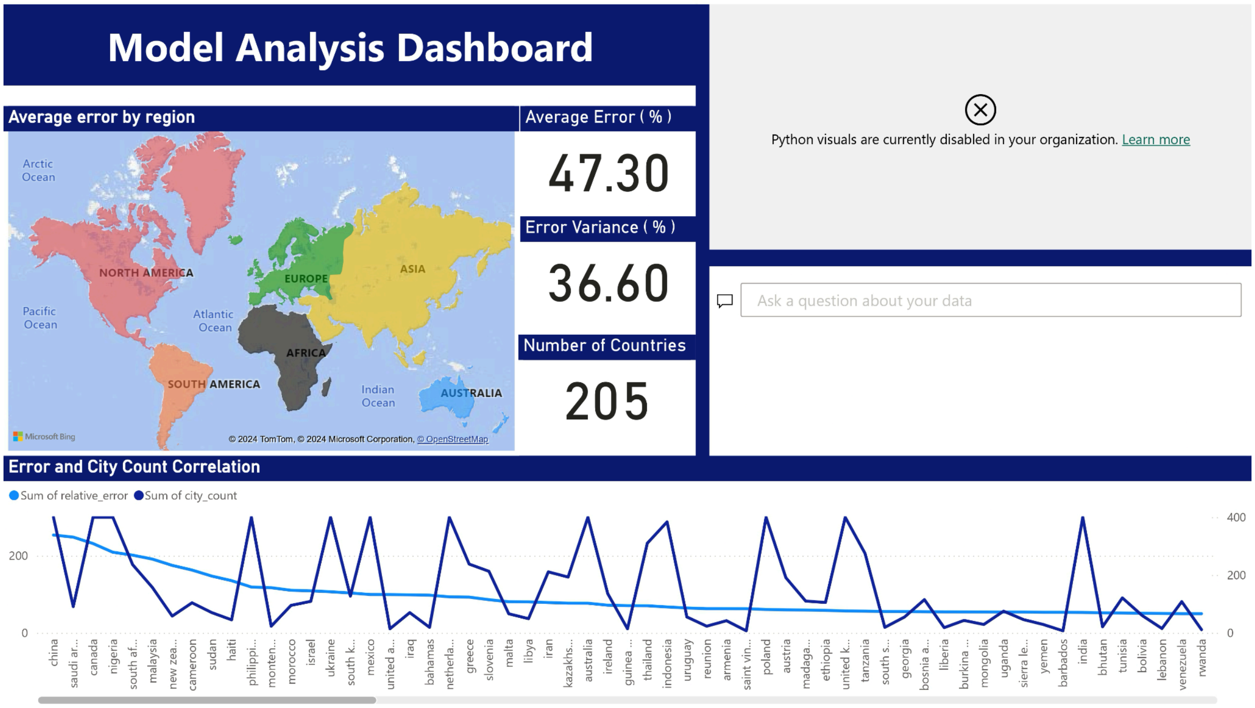
The countries with the largest (negative) a values are actually small countries with 5-20 cities. Meanwhile the countries with the smallest (negative) a values are large countries with the exception of Micronesia having a value of -0.58, the smallest in the world.

Correlation between the a value and city count:



Dashboard

Using Microsoft Power BI our team built a dashboard which reflects the key data in simple easy-to-use visualisations.



Model Analysis Dashboard

Average error by region

Average Error (%)

47.30

Error Variance (%)

36.60

Number of Countries

205

Python visuals are currently disabled in your organization. Learn more

Ask a question about your data

Error and City Count Correlation

The dashboard is fully interactive and includes a region map with the average errors per region, basic info like average error, variance and country count, a graph of city counts and errors and a search engine which uses AI to search info based on the data. An example of a query would be “countries with the least error in their region”. The output would be a value or a simple graph. And finally, on the top right is a Python visual, which plots the model function along the country’s data to compare how accurate the model is. Unfortunately, this type of visual is disabled in the TUD Microsoft365 tenant settings. Our team reached out and was unsuccessful with the request.

Online Dashboard view

In case of issues with accessing issues write to D23125154@mytudublin.ie

Conclusions

With an average error of 47.3% it seems the model is not generally a good fit, however it can vary depending on the country. There is no correlation between model error and city count of a country, except for the fact that the bottom 7 countries by error all have a small city count value. b is the estimate of the most populated city's population. Its error is also significant at 52.54%, however if we exclude the first city the error will generally increase. And finally, the a value shows how fast the populations decrease going from the first city to the least populated one. Once again, generally the countries don't follow Zipf's law and there is somewhat of a correlation between the city count and the a value.

It would be interesting to compare the data for different years to reveal the trends of population distribution inside countries. Another topic would be the trends of migration between countries and how they were affected by COVID-19 and numerous wars in recent years.

Since we found that the model is not a good fit, then why is it so famous? The rule was popularised because it fit the US data well at the time. Today, the US still has the smallest error out of the 400+ city countries. It should also be mentioned that Zipf's law is more popular in the context of language theory and word usage than it is in the context of city populations.

Case 4: Newborns Names

Introduction

The following task will determine if the power law can accurately describe the relationship between the name occurrences and their rank on newborn babies. We will start with boys then move on to girls. Afterwards, there will be a comparison between the boys and the girls in which four graphs assessing Marginal error, a, b, and B, then we conclude.

Results

At first the original data, estimate, and errors will demonstrate, afterwards 3 different cases will be analysed, where the 1st case will exclude the top name, then the 2nd case will exclude the 2 top names, then the last case where the 3 top names will be excluded. The purpose of these cases is to illustrate how the marginal error changes along with b, a, and B.

Boys & Girls

Original:

Value	Boys	Girls
Marginal error	26.74%	31.65%
B	6.8	6.5
a	-0.55	-0.5
b	916.2	692.61

1st case:

Value	Boys	Girls
Marginal error	17.44%	20.16 %
B	6.92	6.68
a	-0.57	-0.54
b	1010.28	798.59

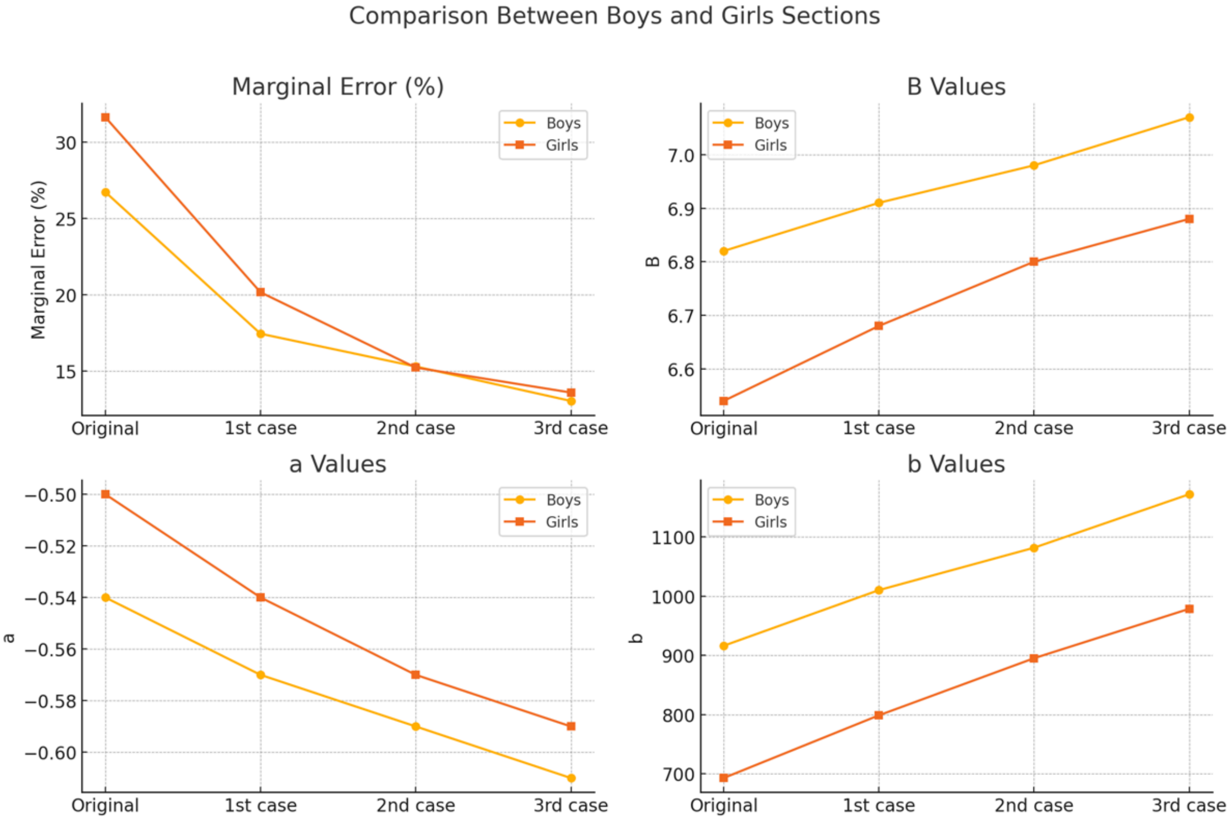
2nd case:

Value	Boys	Girls
Marginal error	15.3%	15.22%
B	6.99	6.8
a	-0.59	-0.57
b	1081.9	895.16

3rd case:

Value	Boys	Girls
Marginal error	13.03 %	13.6 %
B	7.067	6.9
a	-0.61	-0.6
b	1172.4	978.7

Boys and Girls comparison



Conclusions

In conclusion, both the results of girls and boys have similar grid behaviour. In the 1st case there is significant decrease in the marginal error, hence demonstrating that the highest occurred name in both girls and boys occurred significantly more than the other names, in the 2nd the girls do not decrease as much as the boys in which we deduce that the girls have a more distributed ratio where the 1st and 2nd most occurred names are excluded. For a value the boys are faster in which we can relate to the marginal error, the girls have better equally distributed occurrences. In b values, the gap between where the first value of both the girls and the boys is wider than all other graphs which means that the original values of the boys are significantly greater than that of the girls.For B which is the log estimate of the most occured names in both girls and boys the graph shows a stable increase.

Resources

- Population Data (2024)
- Zipf's Law
- Baby names in Ireland, 2023
- Software Used: Python (Selenium, NumPy, Pandas, MatPlotlib), Power BI, Tableau Public, MapleSoft.

Contributions

Vladimir Filatov	Case 3, coding and calculations
Abdullah Alrubian	Case 1, Case 4
Kristian Don	Case 2, General Introduction