STRV Test Project

Matej Kosiba

March 2025

1 Introduction

The goal of the project is to help a newly launched fashion startup. Their business model is to help families choose names for their newborns and then sell personalized clothing featuring those names. The two key objectives are:

- Help parents make a data-driven decision when selecting a name
- Sell personalized clothing featuring those names

For this task, I have to use the US Baby Names Kaggle repository, which contains two datasets. The `NationalNames` dataset and the `StateNames` dataset. Each containing statistics of how many babies with a certain name were born for a given year over some time. The `NationalNames` dataset has these entries summed over all USA states, while the `StateNames` dataset has separate entries per name, gender and year for each of the USA states.

2 Exploring the datasets

First of all, I checked the structure of each of the two datasets. I was asking questions like: How many data points are there, how many columns, what do they represent?

The `NationalNames` dataset has ~ 1.8 million data points and five columns, Id, Name, Year, Gender, and Count. One data point (one row) in the dataset thus represents the number of newborns with a certain name and gender for a specific year. The year span of the entries is from 1880 up to 2014. Tab. 1 shows the first 5 entries of the `NationalNames` dataset.

Table 1: First 5 rows of the `NationalNames` dataset. Each row indicates the number of newborns (count) that were born for the reported year with the specific name and gender.

Id	Name	Year	Gender	Count
1	Mary	1880	F	7065
2	Anna	1880	F	2604
3	Emma	1880	F	2003
4	Elizabeth	1880	F	1939
5	Minnie	1880	F	1746

The second dataset, called `StateNames`, has ~ 5.6 million data points (rows) and 6 columns. The columns in the dataset are Id, Name, Year, Gender, State, and Count. It is essentially built by the same principles as the `NationalNames` dataset, with one change, the data points are not aggregates from all states on the national USA level as in the

`NationalNames` dataset, but instead, the `StateNames` dataset specifies an additional column identifying the state for which the count of the newborns in a specific year is recorded. One entry in the dataset thus represents the number (count) of newborns with a specific name and gender born in a certain year in a certain state of the USA. The entries in the dataset span years from 1910 up to 2014. Tab. 2 shows the first 5 entries of the `StateNames` dataset.

Table 2: First 5 rows of the `StateNames` dataset. Each row indicates the number of newborns (count) that were born for the reported year with the specific name and gender in the specific USA state.

Id	Name	Year	Gender	State	Count
1	Mary	1910	F	AK	14
2	Annie	1910	F	AK	12
3	Anna	1910	F	AK	10
4	Margaret	1910	F	AK	8
5	Helen	1910	F	AK	7

Next, the assignment of the project contained a few questions I should address to explore the dataset. Each of the questions is addresses in the following subsections.

2.1 How did the name Ida change period-over-period nationally?

To address this question, I collected statistics on how many babies with the name Ida were born for each year in the `NationalNames` dataset. Fig. 1 shows these statistics plotted as a histogram. We can see that the name Ida is currently loosing its popularity.

It experienced two popularity spikes. The first between 1880 and 1900 with an approximate yearly occurrence of ~ 2000 newborns, later its popularity slightly decreased to ~ 1900 newborns a year following a significant increase between the years 1910 and 1930 with a yearly average of ~ 3500 newborns named Ida with the record of 4451 newborns named Ida in 1918. Since 1930, the name Ida consistently decreases in popularity to the present average of ~ 95 newborns named Ida a year since 2000.

2.2 How did the name Ida change period-over-period in California?

Similarly to the previous question, I collected all occurrences of the Ida name for a newborn baby over all years in the `StateNames` dataset and selected the California state. Fig. 2 shows this distribution plotted as a histogram where the horizontal axis represents years and the vertical axis represents the number of newborns with name Ida in California.

We can immediately see some similarities with the distribution of the name Ida on the National level. It is consistently decreasing in popularity in California, same as on the national level, and there are also two spikes in its popularity. However, the first spike at the national level is not present at the state level because of the difference in the recorded time for these two datasets. The first spike on the national level occurred before the recorded time for the state level. The second spike in popularity of Ida on the national level is consistent with the first popularity spike in California, both peaking around 1920. On the state level, there is also a small spike of popularity between years ~ 1945 and 1964, which is not visible on the national level.

The popularity spike of the name Ida in California between the years 1915 and 1925 had an approximate yearly occurrence of ~ 75 newborns, the second popularity spike between the years ~ 1945 and 1964 averaged only ~ 50 newborns. Since 1965, the name Ida consistently decreases in popularity to the present ~ 15 newborns named Ida a year in California.

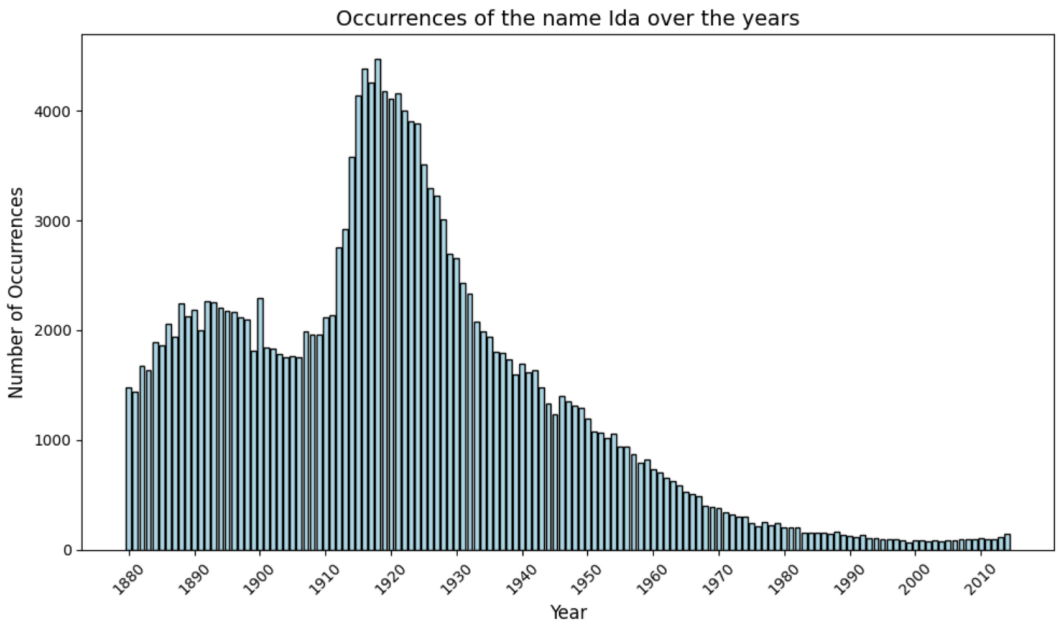

Figure 1: Distribution of the number of babies born with the name Ida on national level since 1880 to 2014.

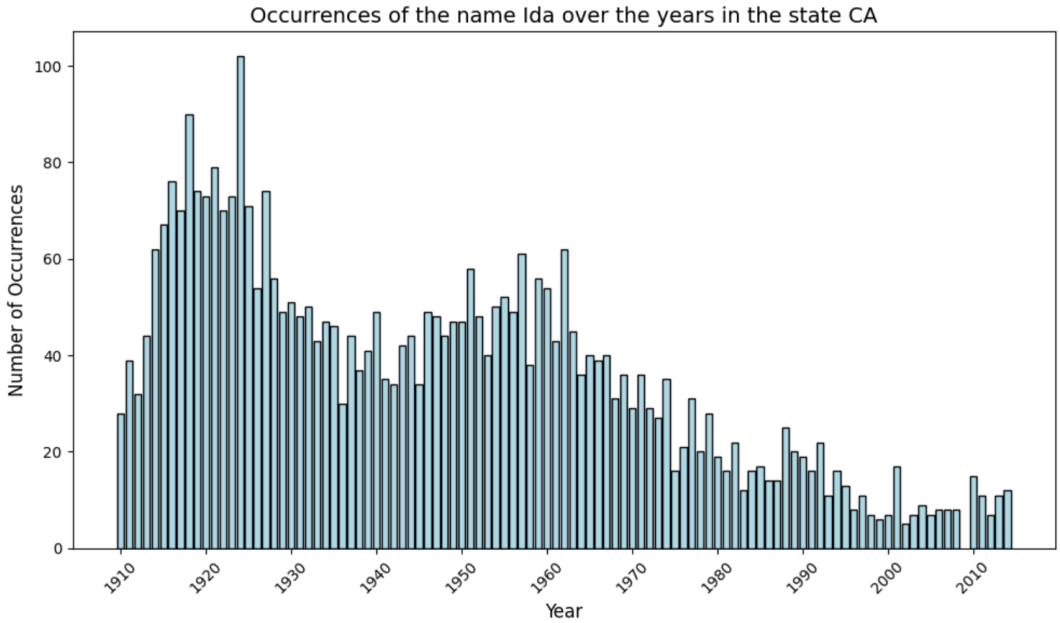

Figure 2: Distribution of the number of babies born with the name Ida in California since 1910 to 2014.

2.3 What name is the most unisex?

To address this question, I first need to define the 'unisexness' or lets say the 'unisex score'. This is not a completely trivial task. The simple approach of this problem would be to calculate statistics on the number of male and female newborns for every name and then select only the names that have the exact proportion. This might look as the best approach to get the names with the most unisex score, however, considering the business aspect of the startup's project, this might not be ideal at all. I anticipated most of the names which have same proportions of male and female newborns (unisex score = 0) being very rare names with extremely low occurrence, which, after exploration of the dataset, was exactly the case. And the problem with this is, that by definition, if a name is very rare, it is not a popular baby name for the customers, thus it would have little value for the startup.

To solve this problem, I introduced a definition of the unisex score, U, as:

$$U = \frac{min(F, M)}{max(F, M)},$$

where M is the number of male newborns with the name and F is the number of female newborns with the name. The min(M, F) takes the minimum value of the two quantities and the max(M, F) takes the maximum. A ratio of these numbers, the unisex score U is a number between 0 and 1, with 1 being exactly balanced over the genders and 0 being attributed solely to one of the genders. For example, the most occurring name with U = 1 is name Tkai with 144 occurrences, 72 male and 72 female, while if I allow for minimum U = 0.9 the name with the largest count will be Riley with 169099 occurrences, which is definitely more useful information for the startup's needs.

For the startup needs, the function that retrieves a random most unisex name for the user has a parameter `minimum_unisex_score`, which defines a minimum value of U as a threshold for generating a random most unisex name for the user. Allowing some margin in U is necessary because selecting only names with perfect U = 1 would return only extremely rare names.

Introducing the threshold on U solves the issue with very rare names only partially. There is a second parameter that is needed, a thresholding parameter that will filter the least popular names, `minimum_name_count`, as the startup customers will find little value in these [1].

Setting the `minimum_unisex_score` = 0.8 and `minimum_name_count` = 10000, I found the most Unisex name to be Riley with 169099 total occurrences and U $\simeq$ 0.93. Compared to the most frequent name on the national level, James with 5129096 total occurrences, Riley is still a quite infrequent name, occurring $\sim 30\times$ less frequently compared to James.

2.4 Which names are common nationally but rare at the state level?

Addressing this question requires defining what it means for a name to be common nationally and also what it means for a name to be rare at the state level.

First, I started by finding the most common name at the national level. This is the name with the largest total occurrence across all genders and years. Now, to calculate how common are the names, I introduce a new parameter, `Relative_Commonness_National` which defines relative commonness of each name with respect to the most common name at the national level. It is calculated as a ratio between the total count of newborns with a certain name to the total count of newborns with the most popular name on the national level. This parameter has a float value in a range between 0 and 1, where 1 is the most common name and the lowest value is attributed to the least common name.

[1] The startup can be offered an additional feature for their app that will pick only very rare names, as there might be a smaller portion on the market with this interest.

Next, I do the same thing per state for the `StateNames` dataset, introducing parameter `Relative_Commonness_State`.

In the next step, I merge this information into one larger data file for easier comparison and I introduce two new parameters that will help me finding the names that are very common nationally but rather rare at the state level. These are `threshold_national`, which is used to filter only names that have `Relative_Commonness_National` larger than this threshold value, and `threshold_state` that is used to filter names on the state level which have smaller `Relative_Commonness_State` than this threshold value.

Tab. 3 shows the 6 names that are most popular on the national level while the least popular at the state level. This is retrievew for thresholds `threshold_national` = 0.75 and `threshold_state` = 0.65. More names, which will have smaller discrepancy between their commonness on national and state level can be found by decreasing the `threshold_national` and/or increasing the `threshold_state` parameters.

Table 3: Names that are common on national level but rare at the state level for `threshold_national` = 0.75 and `threshold_state` = 0.65. Id is the placement of the name based on its lowest relative comonness at the state level, Name is the newborn's name, RCN stands for `Relative_Commonness_National`, RCS stands for `Relative_Commonness_State` and State_{lc} stands for the state at which the name has the least relative comonness among the states.

Id	Name	RCN	RCS	State_{lc}
1	Mary	0.81	0.24	NV
2	Michael	0.84	0.31	MS
3	William	0.79	0.49	NM
4	Robert	0.94	0.51	SC
5	John	0.99	0.55	TN
6	James	1.00	0.63	RI

3 Tasks for Presentation

In this section, I document my approach for answering the 3 recommended tasks for the presentation part.

3.1 Find the TOP 10 trending names

For this task I created a function that collects information on the total count (across all genders) per year for all newborn names in the last X years. Next, for each name I fit this distribution with a linear regression to get a slope of the trend describing the change over time in the past X years for the occurrence of a particular baby name. Fig. 3 shows a fit to this distribution for the name Liam that scored with the steepest slope when analyzing the distribution of baby name counts in the last 5 years. Tab. 4 shows the most trending names over the last 5 years.

The TOP 10 trending names change significantly depending on the number of years for which this analysis is performed, the `years_to_analyze` parameter in my function. Tab. 5 shows the distribution of the TOP 10 trending names over the last 3 years. Fig. 4 shows the fit and statistics of the most trending name over the last 3 years.

I added the `Recent_Counts` column just for the business needs of the startup to keep information on how popular these names are, which would be certainly useful.

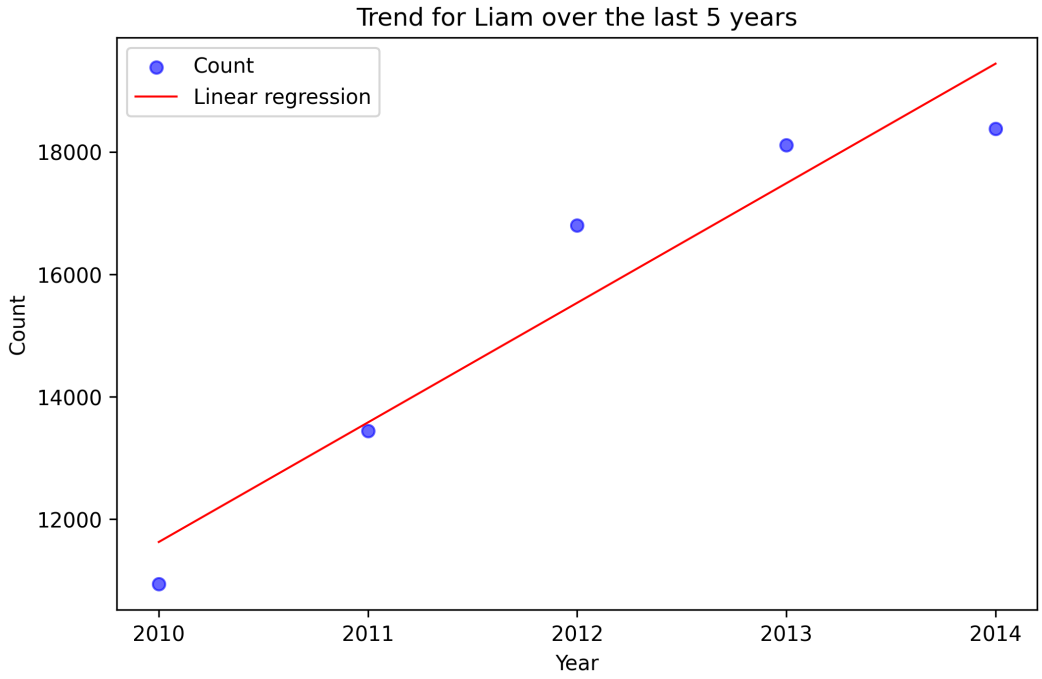

Figure 3: Count of newborns named Liam over the past 5 years fitted with linear regression.

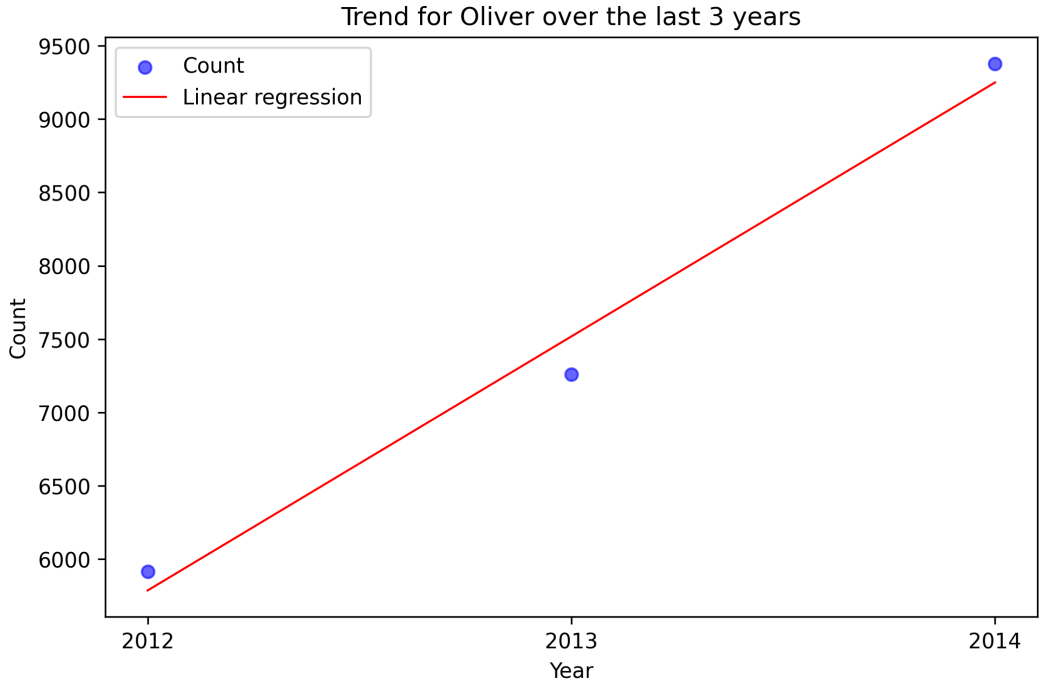

Figure 4: Count of newborns named Oliver over the past 3 years fitted with linear regression.

Table 4: Table of the TOP 10 trending names calculated for the period of the last **5 years**. The first column is the placement of the name, the second is the name, the third column shows the slope after linear regression, and the fourth column `Recent_Counts` shows the total counts of newborns in the given period with the given name.

Id	Name	Slope	Recent_Counts
1	Liam	1953.2	77663
2	Harper	1740.5	34159
3	Aria	1310.2	17195
4	Charlotte	1225.8	38589
5	Oliver	1127.3	32633
6	Jase	1074.5	10307
7	Jaxon	1042.0	30118
8	Jace	985.6	23904
9	Penelope	944.0	15282
10	Emma	902.8	98776

Table 5: Table of the TOP 10 trending names calculated for the period of the last **3 years**. The first column is the placement of the name, the second is the name, the third column shows the slope after linear regression, and the fourth column `Recent_Counts` shows the total counts of newborns in the given period with the given name.

Id	Name	Slope	Recent_Counts
1	Oliver	1731.5	22554
2	Aria	1331.5	14276
3	Jase	1300.0	9436
4	Charlotte	1291.0	26812
5	Penelope	1267.0	11867
6	Sebastian	1255.5	23518
7	Olivia	1200.5	55381
8	Harper	1148.0	26119
9	Sadie	1118.5	12058
10	Lincoln	989.0	11931

3.2 Find the Top 10 states with the most newborns

To find the Top 10 states with the most newborns I simply counted all newborns per state, sorted them by the largest value and found the top 10 states with the most newborn. These are shown in Tab. 6.

However, I would like to note that for the business needs of the startup, it might be more useful to find the most newborns in some recent history, e.g. last 5 or 10 years, which would help them more with deciding in which regions to target most of their marketing for their products.

3.3 Create a map showing the top names by region

For this task I firstly found which is the most popular name per state based on counting all newborns born with a certain name across all genders for the recorded time in the `StateNames` dataset. Fig. 5 shows a color-coded map of the USA based on the most occurring name in each state.

Table 6: Table of the TOP 10 states with the most newborns. The first column is the placement of the state, the second is the state's abbreviation, the third is the total count of the newborns in that state summed over the monitored period, since 1910 up to 2014.

Id	State	Count
1	CA	29252805
2	NY	23891045
3	TX	21820683
4	PA	16776664
5	IL	15304655
6	OH	14315490
7	MI	11734824
8	FL	9457005
9	NC	8537298
10	NJ	8398176

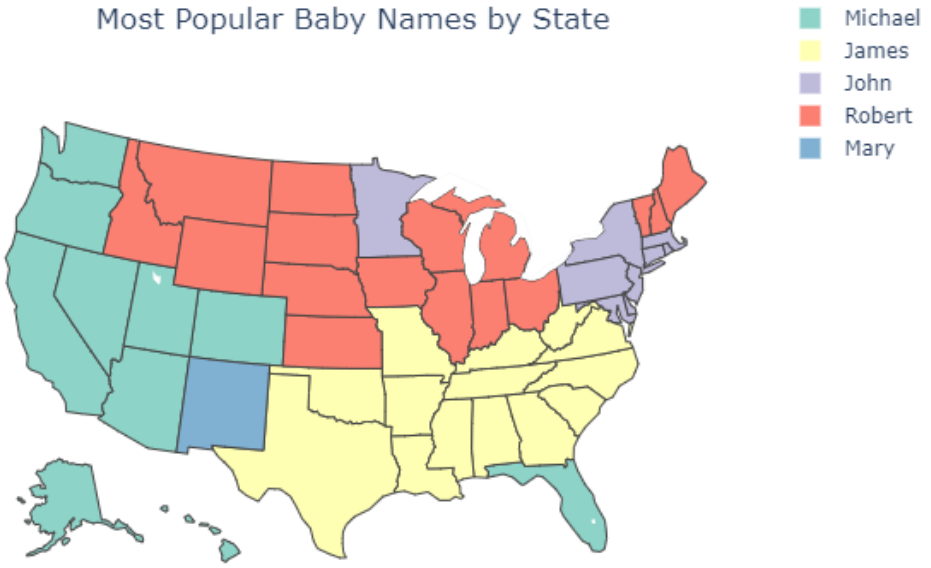

Figure 5: Color-coded map of the USA showing the most popular name per state from 1910 to 2014.