Data-driven insights for a fashion startup: Helping parents choose names & supporting business growth

Introduction

A newly launched fashion startup aims to help families select names for their newborns while offering personalized clothing. Using a U.S. baby names dataset spanning from 1880 to 2014, I believe that valuable insights—capable of guiding a startup's strategic moves or assisting a parent in finding the right name—can be extracted from the data. The goal is to uncover meaningful patterns in U.S. baby name trends that not only highlight promising new markets for the startup but also lay the groundwork for a personalized recommendation system to help parents choose a name that feels just right—and then offer personalized clothes fot the baby.

I see the data serving two purposes: fueling the startup's expansion and providing a thoughtful resource for families in search of the perfect name. This report details the journey of how I arrived at these insights and what they could mean for the road ahead.

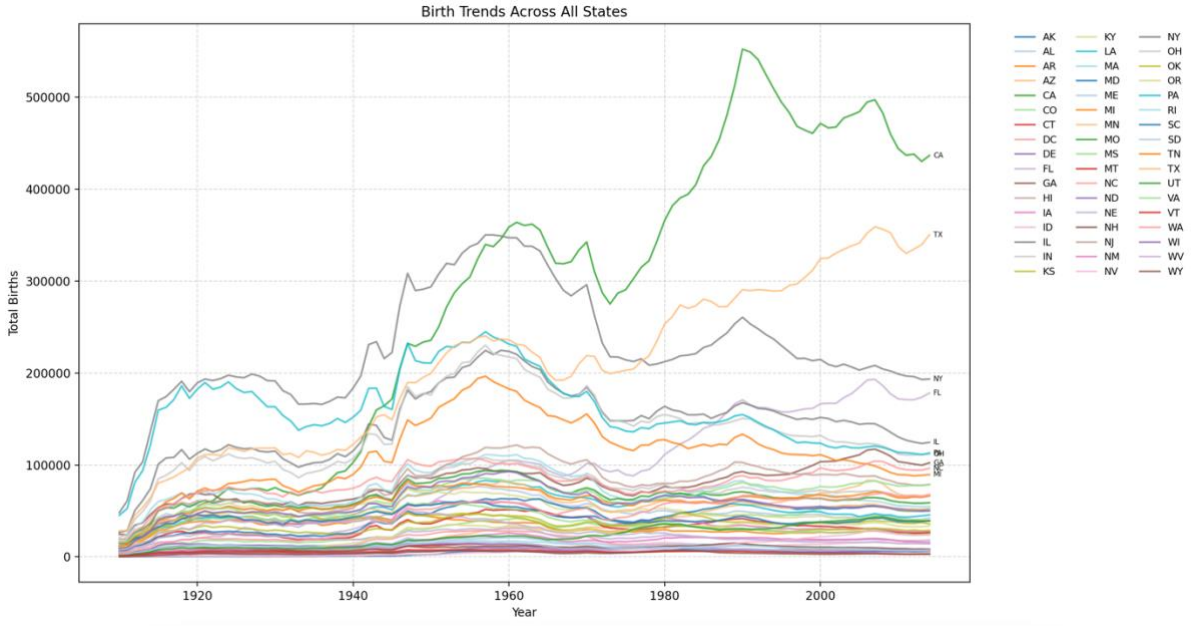
Below, I outline data findings that can help guide the startup's growth and expansion by pinpointing high-potential states and uncovering naming trends that could drive effective marketing.

Data exploration

Birth trends over time by state

Analysis

To understand market potential, I analyzed birth trends across different states using historical data. The first plot illustrates total baby births from 1880 onward.



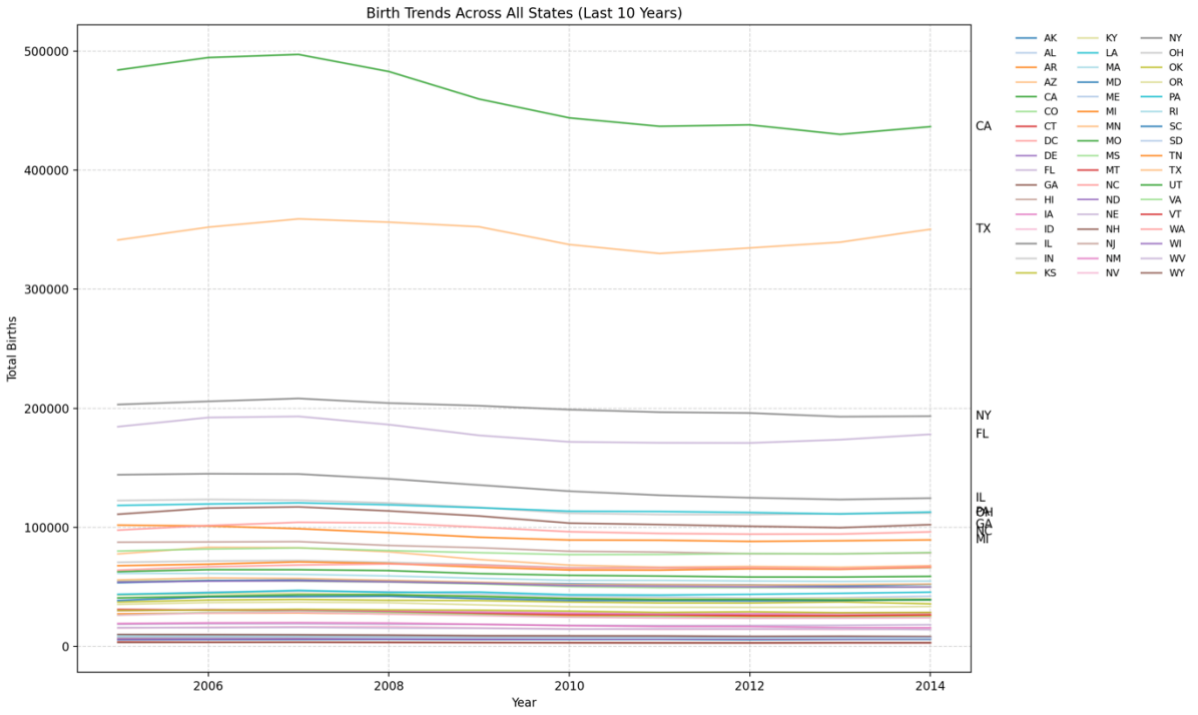
Findings

- The overall birth trend has shown continuous growth over time, aligning with general population increases.
- Certain states stand out in terms of newborn numbers: **California (CA), Texas (TX), and New York (NY)**, followed by **Illinois (IL), Pennsylvania (PA), and Ohio (OH).**
- Although these insights confirm high birth rates in key regions, the long historical scope (1880 onward) makes them less actionable for current business decisions.

Birth trends in the last 10 years

Analysis

To get a more relevant perspective (since a chart spanning from 1880-2014 gives us too long a view and the overall trend might not be meaningful for the present moment), I focused on analyzing birth trends from the most recent decade. The second plot shows the newborn counts by state for the last 10 years.



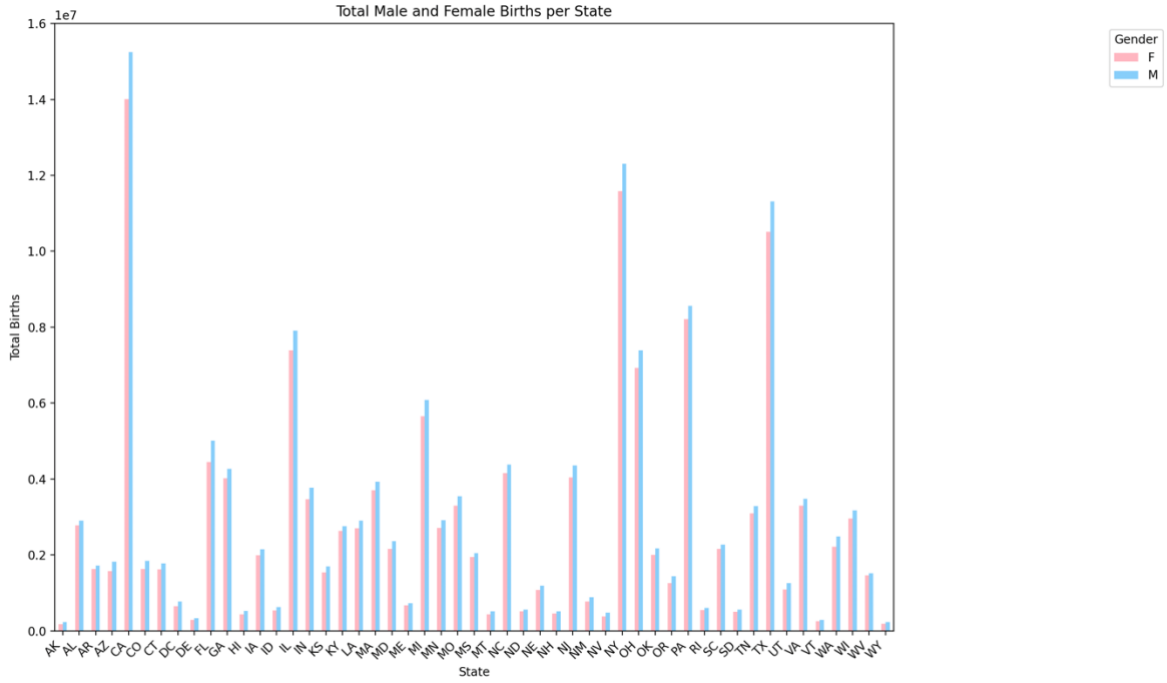
Findings

- In recent years, the number of newborns has remained relatively stable.
- Notably, **California (CA) shows a decreasing trend**, whereas other leading states maintain consistent birth rates.
- This stability suggests a reliable market for baby-related businesses, making high-birth states attractive locations for launching and expanding the brand's personalized clothing line.

Gender ratio of newborns

Analysis

Next, I wondered: are there noticeable differences in the number of baby girls and boys born? This could be important for shaping the startup's marketing strategy. I looked at the ratio of male to female newborns in different states, and the bar chart below shows the gender distribution



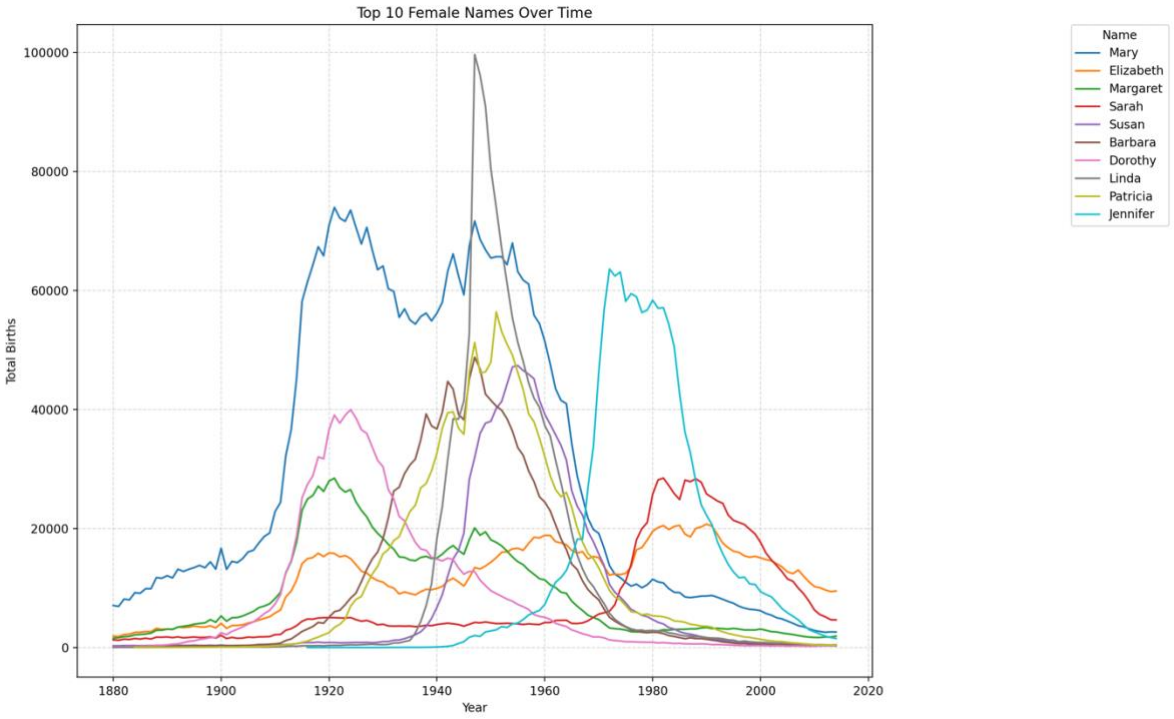
Findings

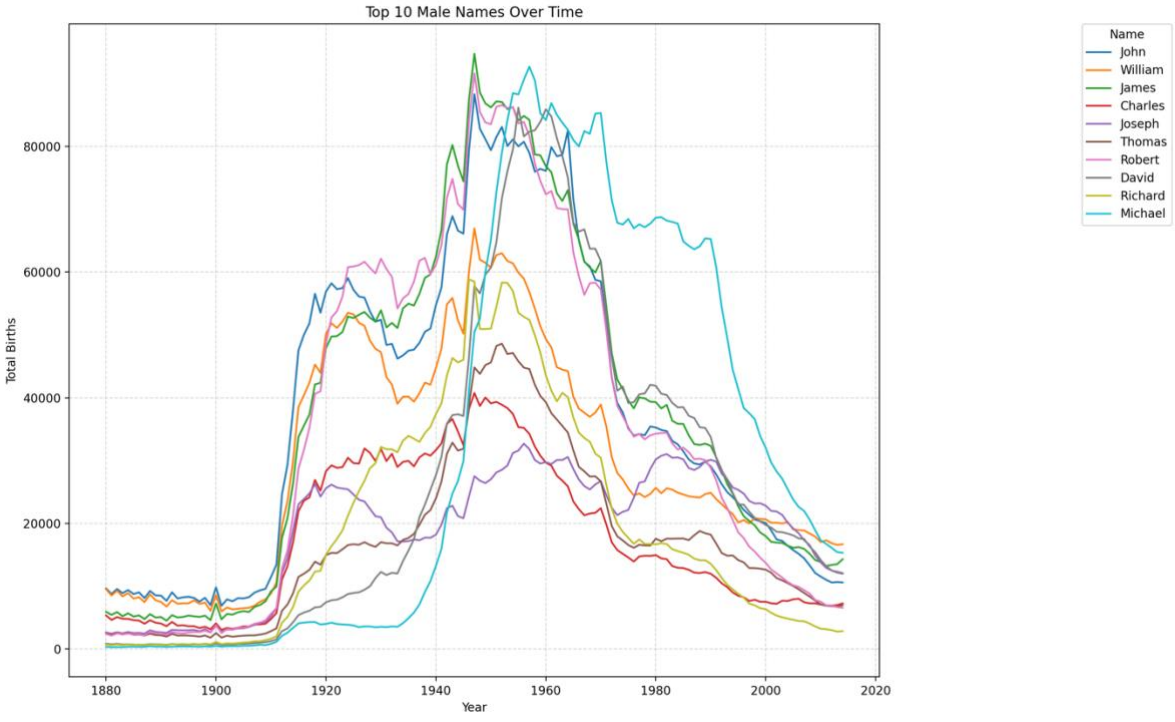
- States with the highest birth counts remain consistent across the board.
- There is a slight predominance of baby boys over baby girls.
- **Business Implication:** This ratio can inform inventory planning and marketing strategies, ensuring the startup's product line is balanced and aligned with the demographic distribution.

Most popular baby names over time

Analysis

To better understand naming trends, I analyzed the top 10 most popular names for boys and girls based on their total occurrences in the dataset. Looking at their popularity over time, it became evident that many of these names had their peak years in the past, and their usage has since declined.





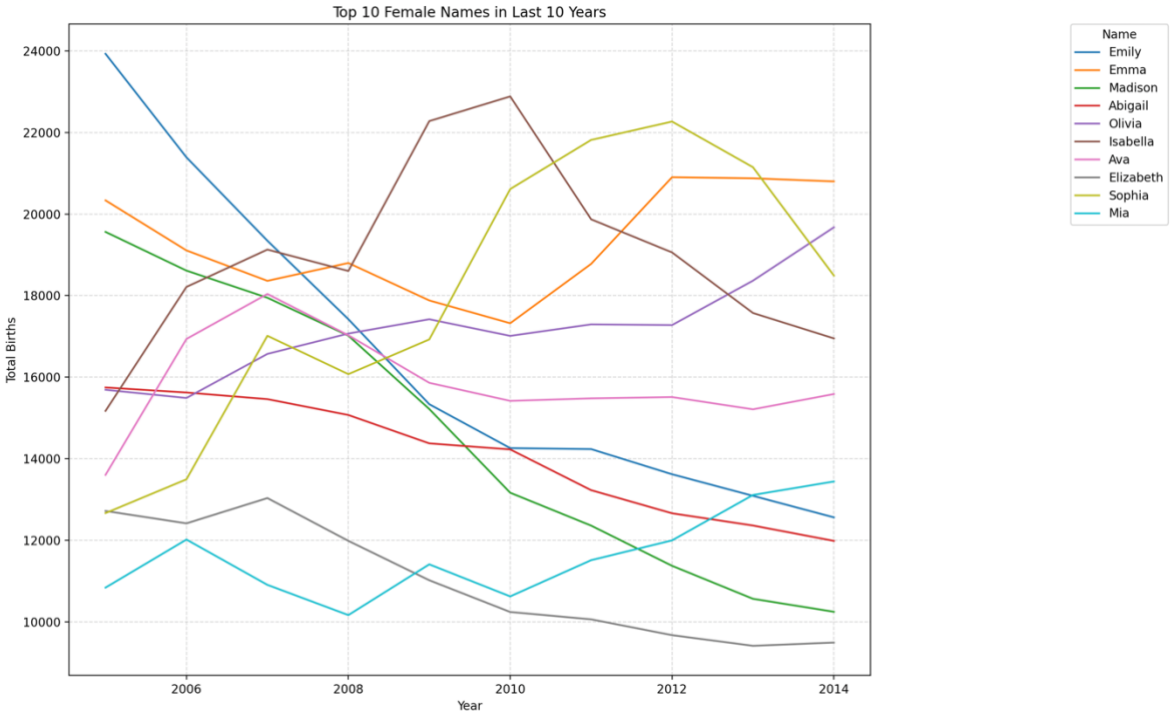
Findings

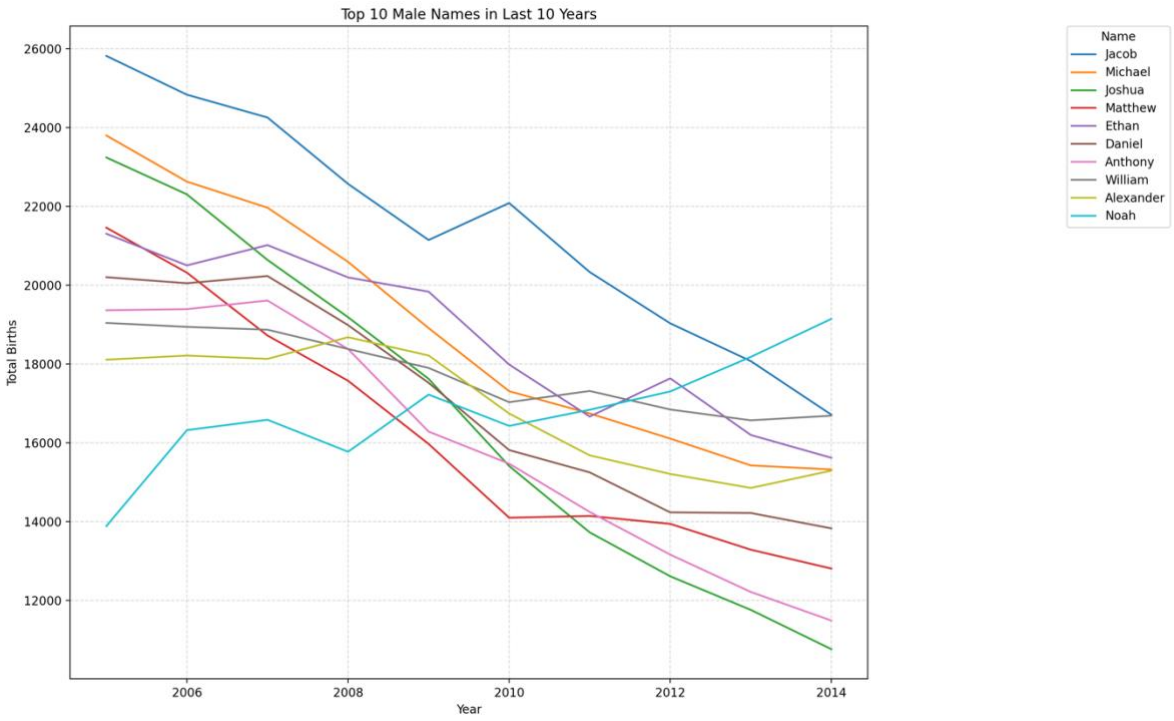
- Many historically popular names reached their peak in the past but have become less common recently, indicating shifting naming preferences.
- This suggests that relying solely on historically top-ranked names may not align with current trends.
- **Business Implication:** By understanding these trends, the startup can focus on names that are currently resonating with modern parents when designing personalized products.

Most popular names in the last 10 years

Analysis

Looking at all-time popularity can be misleading, since many names reached their peak years ago. To get a better sense of current trends, I focused on the top 10 names for boys and girls from the last decade. While most of these names show a downward trend, one exception clearly stood out.





Findings

- Most top names in the last decade have seen a decline in usage over time.
- However, a few names have remained stable, indicating consistency in parental preferences.
- **Noteworthy observation:** The male name Noah is the only one that showed a continuous increase in popularity.
- **Business implication:** This insight allows the startup to align its product offerings with modern naming trends by prioritizing currently trending names such as Noah, ensuring the brand remains relevant and appealing.

Grouping states based on naming preferences

Analysis

I was curious if we could group states by identifying similarities in the names parents choose. Given that the dataset used for clustering contains categorical variables (this will be explained later), K-Modes clustering is a more suitable choice than algorithms designed for continuous data—for example, K-Means, which was the first algorithm that came to mind.

My approach involved several steps:

1. **Creating a ratio dataset**

   I first created a dataset where, for every name, I computed the ratio of its usage in a particular state relative to its national usage. This ratio serves as a measure to determine if a name is notably more popular in certain states compared to its overall national popularity

```
Id_state,Name,Year,Gender,State,state_count,Id_national,national_count,ratio
1,Mary,1910,F,AK,14,88944,22848,0.0006127450980392157
2,Annie,1910,F,AK,12,88962,3519,0.0034100596760443308
3,Anna,1910,F,AK,10,88949,6436,0.0015537600994406464
4,Margaret,1910,F,AK,8,88946,8226,0.0009725261366399222
5,Helen,1910,F,AK,7,88945,10479,0.0006680026720106881
6,Elsie,1910,F,AK,6,88987,2141,0.0028024287716020553
7,Lucy,1910,F,AK,6,89018,1283,0.004676539360872954
8,Dorothy,1910,F,AK,5,88947,7318,0.000683246788740093
9,Mary,1911,F,AK,12,93573,24390,0.0004920049200492004
10,Margaret,1911,F,AK,7,93575,9279,0.0007543916370298524
11,Ruth,1911,F,AK,7,93577,8003,0.0008746719980007497
12,Annie,1911,F,AK,6,93596,3298,0.0018192844147968466
13,Elizabeth,1911,F,AK,6,93579,6298,0.000952683391552874
14,Helen,1911,F,AK,6,93574,11802,0.0005083884087442806
15,Mary,1912,F,AK,9,98440,32303,0.0002786118936321704
16,Elsie,1912,F,AK,8,98481,2895,0.002763385146804836
17,Agnes,1912,F,AK,7,98479,2945,0.0023769100169779285
18,Anna,1912,F,AK,7,98446,8586,0.00081528068949452 6
19,Helen,1912,F,AK,7,98441,16133,0.0004338932622574846 4
20,Louise,1912,F,AK,7,98456,5116,0.0013682564503518374
21,Jean,1912,F,AK,6,98503,2056,0.0029182879377431907
```

2. **Identifying significant states**

Next, I generated a second dataset containing all names along with their "significant states". For each name, a state is flagged as significant if its ratio is greater than a specified factor (currently set to 2) times the name's overall average ratio. Only names with a sufficient spread in their ratios (i.e., the difference between the maximum and minimum ratios exceeds a threshold) are included.

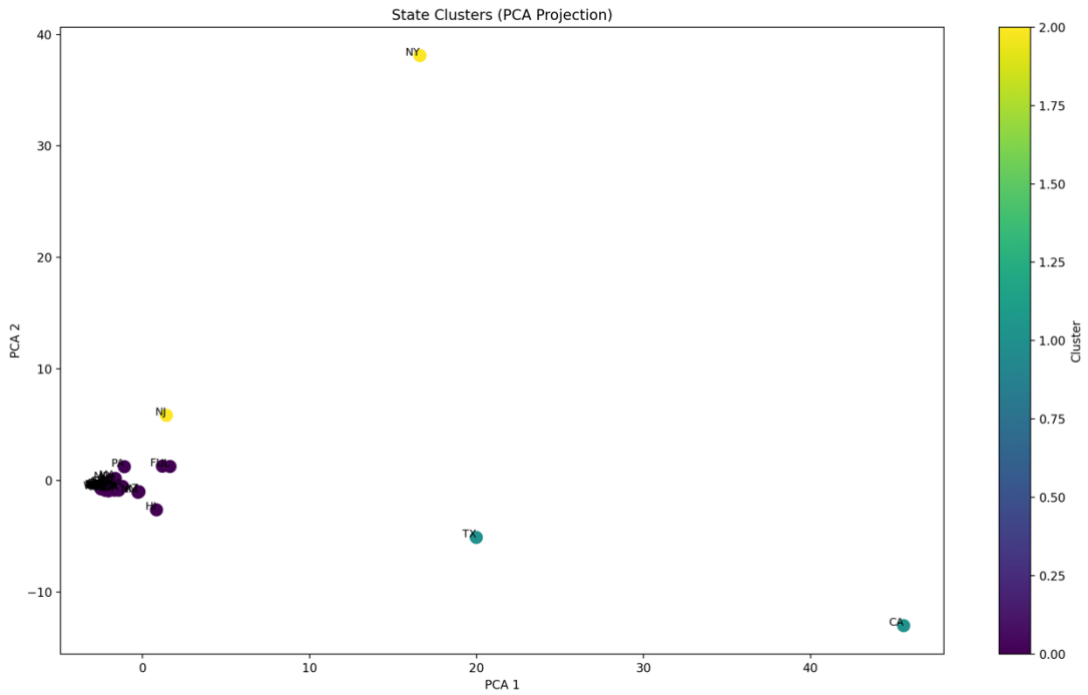
```
Name,State,ratio
Aaban,NY,0.4017857142857143
Aadan,CA,0.4013504611330698
Aadan,TX,0.22727272727272727
Aadarsh,IL,0.3125
Aaden,AL,0.025721956537175 09
Aaden,AR,0.014238993115296313
Aaden,AZ,0.023532564243158884
Aaden,CA,0.13922410933053844
Aaden,CO,0.02283530620710138
Aaden,CT,0.009998842948055829
Aaden,DE,0.003946329913180742
Aaden,FL,0.06402624392746122
```

3. **Clustering States**

I then used this dataset for clustering, with the goal of grouping states that exhibit similar patterns in parental naming preferences. This clustering could eventually support tailored marketing strategies—for example, if a store performs well in one state, a similar market in the same cluster could be a promising expansion target.

Findings

- When applying K-Modes clustering, I experimented with different numbers of clusters. Visual analysis of the clusters suggested that three clusters provided the most natural grouping:



- **Cluster 1: TX and CA** grouped together, indicating similar trends in parental name choices
- **Cluster 2: NY and NJ** formed a distinct group, reflecting their unique naming patterns
- **Cluster 3: All other states** clustered together, suggesting more homogenous or less distinctive naming trends across these regions

Room for improvement

One key area for further research is the method used to identify significant states. Currently, a state is flagged as significant if its ratio is greater than 2 × mean_ratio for that name. This fixed factor of 2 might not be optimal for capturing true deviations in state-level popularity.

Business implication

These insights can help inform targeted marketing strategies. For instance, if a business sees success in TX, they might consider expanding to other states within the same cluster to leverage similar consumer preferences (CA).

Data exploration conclusion

Drawing together these insights provides a strategic roadmap for the startup. By targeting markets with stable and high birth rates—particularly in states like Texas and New York—and by being mindful of evolving naming trends, the business can optimize both its product offerings and market positioning.

In addition, applying K-Modes clustering to group states by similarities in parental naming preferences has revealed distinct regional groupings. For example, states such as TX and CA tend to form one cluster, while NY and NJ emerge as another, with the remaining states forming a third cluster. These clusters suggest that tailored marketing strategies could be developed for each group, leveraging localized naming trends and consumer behaviors. A

balanced inventory that considers the slight predominance of baby boys, alongside a focus on names that are currently trending, will help tailor personalized products to meet customer demand.

Overall, integrating these data-driven findings—including both naming trends and clustering insights—into regular decision-making processes will support the startup's growth and help it maintain a competitive edge in the dynamic baby fashion industry.

Name recommendation systems

Recommendation system based on preferences

Idea

To help parents choose names that match their preferences, the startup could offer a personalized name recommendation system. The original datasets don't offer many features that a recommendation system could be based on. So, I thought about adding new characteristics that could improve the process. Drawing from common sense and conversations with people in general, I realized that even if parents aren't sure about a specific name, they often have preferences. For example, they may know whether they want a short or long name, something rare or common, recently popular or more traditional, or perhaps a unisex option. Newly added features are:

- **Name length**

Names are categorized into short, medium, or long using quantiles. Names in the first 33% by length are considered short, the 33–66% range is medium, and names above that threshold are classified as long.

- **Name rarity**

A name is deemed common if it falls within the top 10% of used names; all others are considered rare. This straightforward criterion is a solid starting point, though it could be refined further by incorporating more recent data and analyzing the overall frequency distribution.

- **Unisex**

A name is flagged as unisex if it appears for both males and females. This helps cater to parents who prefer gender-neutral options.

- **Popularity trend**

This attribute is computed over the last 10 years as follows:

- **Increasing:** If the name's usage is consistently rising
- **Decreasing:** If the usage is steadily falling
- **Varied:** If there is no consistent trend or insufficient data

Using these enriched features, the similarity between user preferences and name characteristics is calculated using cosine similarity. For example, consider the following user preferences:

```
user_preferences = {
    "length": "Long",       # Options: Short, Medium, Long
    "unisex": True,         # Boolean: True for unisex names, False
otherwise
    "gender": "F",          # Expected baby's gender: F or M
    "rarity": "Rare",       # Options: Rare or Common
    "popularity_trend": "Decreasing"  # Options: Increasing,
Decreasing, Varied
}
```

The system can generate top suggestions by running a function such as:

```
suggest_names_based_on_preferences(pd.read_csv('./data/NameSuggestio
nsDataset.csv'), user_preferences, top_n=10)
```

This returns the following top 10 name suggestions:

```
['Zechariah', 'Shaughnessy', 'Pressley', 'Theodore', 'Kearston',
'Johnelle', 'Chestine', 'Reginald', 'Donyelle', 'Marquette']
```

Room for improvement

While this recommendation system provides a solid baseline, there are several areas for future enhancement:

- **Name rarity**

  The current approach is very basic, using the top 10% of used names over all time. It could be refined by dynamically adjusting the threshold based on current trends.

- **Geographic specificity**

  The system currently relies on a national dataset. Incorporating state-level data could enable recommendations that are tailored to regional naming trends.

- **Dataset enrichment**

  Integrating additional datasets, such as demographic data for states, could allow for more sophisticated recommendations. For example, connecting names with the demographic characteristics of a region might yield insights into names that resonate with specific cultural or socio-economic groups.

Recommendation system based on phonetic similarity

Another approach for name recommendation leverages phonetic similarity instead of semantic similarity. Initially, I experimented with models like SBERT and word2vec. While SBERT (using the all-MiniLM-L6-v2 model) is fine-tuned for generating embeddings from sentences, it is less suitable for individual words like names. Similarly, word2vec excels at capturing context-based similarity (e.g., "queen" and "king" are similar), but this does not necessarily translate well when comparing names.

Instead, I found that **Soundex**—a phonetic algorithm—is better suited for this task because it groups names by how they sound. Soundex is designed to index names by their pronunciation, making it useful for finding names that are phonetically similar, even if they differ in spelling.

For example, consider the following usage:

```
name = "James"
desired_gender = "M"  # "M" or "F"; set to None to disable filtering
similar_names = get_similar_names(name, top_k=10,
gender=desired_gender)
```

This returns:

```
['James', 'Jonas', 'Junius', 'Jones', 'Junious', 'Jens', 'Janes',
'Junus', 'Johannes', 'Junuis']
```

Here, Soundex groups together names that sound similar to "James" based on their phonetic encoding. I believe this approach could be helpful for parents who have a name in mind but wish to explore phonetically similar alternatives—especially when they are concerned about factors like commonality or unwanted associations that they might have with the name on their mind.

Key insights

1. Market expansion, high-birth states, and similarities in naming preferences

Key states such as California, Texas, and New York stand out due to their high birth rates, making them prime targets for business expansion. While California shows a slight decline in births, the overall trend in Texas and New York remains strong, providing a reliable market base for the startup’s personalized clothing line.

Furthermore, by applying K-Modes clustering to identify similarities in parental naming preferences, we can group states that share common trends. For instance, Texas and California often appear together in one cluster, while New York and New Jersey emerge in another. This approach enables the startup to tailor its marketing strategies more precisely: if a product or campaign performs well in one state, there is a good chance of success in a similar (clustered) state.

2. Evolving baby naming trends and recommendation system

Baby name trends continue to shift, with many historically popular names gradually declining in usage, while others—like “Noah”—are on the rise. Staying attuned to these evolving trends is essential to keeping product offerings relevant.

To address this, a recommendation system can be implemented, combining:

- **Preference-based suggestions**: Factoring in elements such as name length, rarity, gender, and popularity trends

- **Phonetic similarity**: Identifying names that sound alike, even if spelled differently, which helps parents explore variations of a name they already like

3. Tailored marketing strategies and future growth

Drawing these insights together provides a strategic roadmap for the startup. Focusing on states with high birth rates and leveraging state-level clustering ensures more efficient market expansion. At the same time, the recommendation system’s integration of evolving name trends and phonetic similarity keeps product lines and name-based personalization relevant.

Overall, incorporating these data-driven findings—regarding both **where** (high-birth or similar states) and **what** (popular or phonetically matched names) to emphasize—will support the startup’s growth and help maintain a competitive edge in the baby fashion industry.

Evaluation of the approach

Strengths

- **Data-driven market targeting**

Focusing on high-birth states like California, Texas, and New York leverages available demographic data to identify promising markets. This enables more efficient resource allocation and targeted marketing efforts.

- **Clustering for regional similarity**

Using K-Modes clustering to group states based on parental naming preferences helps identify markets with shared consumer behavior patterns (for example, Texas with California, and New York with New Jersey). This allows for replicating successful campaigns across similar markets.

- **Integration of evolving trends**

Recognizing shifts in baby naming trends, such as the rising popularity of “Noah”, ensures that both the product offerings and the recommendation system remain relevant. By combining trends with phonetic similarities, the recommendation system provides a more nuanced selection of names.

- **Dual-purpose use of data**

The hypothesis outlines two clear applications for the dataset:

- **Business Expansion:** Identifying high-potential markets based on birth rates and naming clusters.

- **Customer Engagement:** Offering parents personalized, data-backed name recommendations that go beyond simple popularity metrics.

Areas for improvement

- **Data recency**

The baby names dataset spans from 1880 to 2014. While historical trends provide valuable context, integrating more recent data or continuously updating the dataset would improve the relevance of insights to current market conditions.

- **Assumption of correlation**

The strategy assumes that naming trends are directly linked to clothing purchasing behavior. Additional validation or qualitative research may be needed to confirm this connection and ensure that the marketing strategies are well-targeted.

- **Clustering limitations**

Clustering by naming preferences may not capture all cultural, socioeconomic, or regional differences. Even if states share similar naming trends, other consumer behavior factors could affect product uptake, suggesting a need for broader data analysis.

- **Recommendation system development**

While the envisioned recommendation system incorporates factors like preference-based suggestions and phonetic similarity, its effectiveness could be enhanced by integrating additional socio-demographic data. This would provide a more comprehensive understanding of the target market and help tailor suggestions even more accurately.

Overall Evaluation

The strategy is well-grounded in data and outlines a roadmap for leveraging historical insights for both strategic expansion and customer personalization. To fully realize its potential, it would be beneficial to:

- Update or supplement the dataset with more recent data
- Validate the connections between naming trends and clothing purchasing behavior through further research
- Incorporate broader socio-demographic information to enhance the recommendation system

With these adjustments, the approach could offer a strong competitive advantage in targeting and engaging the startup's niche market.