CS777 Final Project

Laleye. M. Stephan

10 – 14 – 24

**Project Report:**

The provided project focuses on applying a **K-means** clustering algorithm on dataset of **six attributes** and **over sixty-nine thousand rows**. The purpose of the **K-means** algorithm is to analyze a dataset of popular baby names given to newborns in New York City from the year 2011 to 2021. The primary goal of this analysis is to investigate signs of any trends found in the naming conventions across ethnic groups, to identify influential names that span multiple ethnicities, and to explore how naming trends differ between groups. Using PySpark, a scalable big-data platform that works well with large datasets, the implementation of the **K-means** clustering approach is to discover **patterns** and **insights** amongst the tuples recorded within the overall dataset. The dataset is first transformed into a suitable format for analysis and then processed using machine learning techniques to derive meaningful results. The dataset itself was derived from a popular site ( **Data.gov** ). The dataset referred to in this analysis is titled ***Popular Baby Names***.

The ***Popular Baby Names*** dataset used for this analysis contains six attributes: the child’s **first name**, the **ethnic group**, **rank in** name **popularity**, the **count** / **frequency** with which a particular name is encounter, the **gender** of a particular child, and the **birth year** of a particular child. The dataset itself, initially stored as an **Excel (.xlsx)** file, was converted into a **Parquet** file for optimized performance on Google Cloud and ease of scalability regarding the dataset. After the conversion between files, and after loading the **Parquet** file into a **PySpark** session, a preprocessing pipeline was implemented that included **encoding** categorical variables**, scaling numerical** variables, and assembling features required for **K-means** clustering. The **ethnicity** attribute for example, was **encoded** using **one-hot encoding** to ensure the **K-means** algorithm could interpret it as part of the input feature vector.

The **K-means** algorithm was then implemented with five clusters **( k=5 )** on the scaled dataset, where input features such as **ethnicity encoding**, **popularity rank**, and **count** all contributed to the clustering process of the data. The algorithm grouped baby names into clusters based on similarities it found within these input features. Once the **K-means** clustering model was **trained** on **eighty percent** of the dataset, the algorithm was then applied a **testing** portion of that the dataset that was **twenty percent** of the entire dataset. The **K-means** clustering model generated a **"prediction"** column that assigned each baby name to a specific cluster. Each cluster within the prediction column was designated as the origin ethnicity with respect to the baby’s name.

Once the clusters were generated by the algorith, further analysis was then conducted to explore any similarity patterns discovered within each cluster. The analysis was split into three primary outputs:

1. **Cluster Ethnicity Analysis**: The first step was to define the ethnicities represented within each cluster. By counting distinct ethnic groups per cluster, clusters that had more diverse ethnic representations were identified.
2. **Top Baby Names**: For the second output, the top 10 baby names within each cluster for both male and female names were identified. This allowed the analysis to consider which names were most prevalent within each ethnic group and cluster.
3. **Influential Names Across Clusters**: The third analysis focused on identifying names that appeared across multiple clusters. These influential names were important as they indicated names that transcended individual ethnic groups and were popular across different ethnic backgrounds.

Overall, the clustering results outputted from the **K-means** algortihm yielded somewhat insightful results as to the influence of names from ethnic group to the next. For example, as seen in the final output file, names such as "Savannah," "Armani," and "Faith" appeared across multiple ethnic groups. The final output file from this project’s script yielded **the top ten most influential names found across two or more clusters**, with each name dominating two different ehtnic clusters repsectively. This suggests that **certain names were popular across multiple ethnicities**, indicating a degree of cultural overlap in naming conventions. In clusters with high distinct ethnicities, names such as "Oscar," "Ethan," and "Christina" were found to be prevalent across multiple ethnic groups, showcasing the influence these names have across different cultural spheres.

The top names found within each cluster **highlighted the diversity** in naming preferences across ethnic groups. Specifically, names like "Ashley" and "Louis" were found within distinct clusters, reflecting how some names maintain cultural significance within specific ethnic groups.

Some unexpected results were in the **discrepancies** **between the second and third output files** from this project’s script and the dataset it reviewed. Regarding the second output file, the top ten most popular boy and the top ten most popular girl names for each ethnic cluster were recorded and printed out. The purpose of this output file was to serve as predictor for what names could potentially be the most influential across multiple clusters **( ethnicities )**. Interestingly, names like **Faith ( Hispanic Origin )** and **Oscar ( Caucasian Origin )** found to be the most prevalent across multiple clusters, were also found within the **top 10 most frequent** baby names, **from the second output file**, that listed their popularities within the **( Faith - Caucasian Origin )** and **( Oscar - Asian / Pacific Islander )** respectively. However, names like **Summer** and **Armani**, both listed as **( Hispanic Origin )**, and both included as part of the top ten **most** **frequent** baby names found across different ethnicites **provided by the third output file**, did not feature in the top ten most frequently used names within any of the ethnic categoires regarding the data that was provided within the return output file 2. These discrepancies suggest that some baby names, while they remain popular within one ethnicity, those same baby names may have little to no bearing in any other ethnicies listed by the original dataset.

Despite some of the discrepancies found between naming inter and intra cluster naming popularities, and despite other valuable insights provided by the **K-means** clustering analysis, certain **inefficiencies** and **limitations** should be acknowledged when reviewing the results of the analysis. One **significant challenge** was the **reliance on categorical encoding for ethnicity**, which would not fully capture the cultural nuances of naming conventions. The one-hot encoding of ethnicities treated each group as mutually exclusive, potentially oversimplifying the relationships between different ethnic groups and the popularities of baby names from one group to the next. Furthermore, by implementing a **K-means** clustering algorithm, the analysis assumed that the data was distributed in a spherical manner and that the clusters were of equal size, which may not accurately reflect the true distribution of names across different ethnicities. Another limitation was that **K-means** algorithm did not automatically assign meaningful labels to each cluster, and as a result, defining the ethnicities for each cluster was based on post-processing steps rather than on inherent properties defined by the algorithm. Additionally, this analysis does not account for temporal changes in naming preferences or consider geographical variations within ethnic groups, which could influence naming trends. Considering these important factors may have **reduced the overall effectiveness of the analysis** in providing a complete picture of the cultural influences on baby naming conventions across New York City. And while the initial expectations of this analysis have been somewhat answered by the **K-means** algorithm’s end results, there still is a lot of ambiguity that leaves a more thorough analysis to be desired.

From this analysis, several definite conclusions can be drawn. Firstly, the clustering results reveal that certain names enjoy cross-cultural appeal, as evidenced by their prevalence in clusters with diverse ethnic representation. These influential names, such as **Savannah** and **Faith**, **suggest that societal factors,** such as **media** and **cultural trends**, may influence naming conventions across ethnic groups. Furthermore, the clustering approach allowed this analysis to identify patterns that might have not been visible from simple descriptive statistics, offering a more nuanced understanding of how names are shared across ethnic boundaries.

The purpose of identifying the top names within each cluster was done so to capture the uniqueness of each ethnic group’s naming conventions. For example, while **some names appear frequently across multiple ethnic groups**, other names, such as **( Dylan - Caucasian )**, appear to be strongly tied to specific cultural traditions. This showcases how names can be both markers of cultural identity and symbols of broader societal influences.

In conclusive thought, the implementation of the **K-means** clustering approach provided valuable insights into the naming conventions of different ethnic groups in New York City. The analysis itself highlighted names that spanned across cultural boundaries, illustrating how naming conventions can vary significantly between groups. However, certain **inefficiencies**, such as the **reliance on categorical encoding** for ethnicity and the **assumptions inherent in the K-means algorithm**, limited the accuracy of the results. The **one-hot encoding** of ethnicities **simplifies cultural relationships**, and **K-means assumes equal-sized, spherical clusters**, which **may not fully reflect the true diversity and distribution** of names across groups. Though despite these limitations, this analysis helps underscore the importance of machine learning techniques like **K-means clustering** in uncovering hidden patterns within datasets, making it a powerful tool for sociocultural analysis.

For future studies, incorporating better practices could improve the analysis and address some of these inefficiencies that plagued this analysis. Measures such as **expanding the number of clusters** might have captured a more **nuanced** rendering of **ethnic name** groupings outputted by the algorithm; while having the availability of additional features such as **birth year** would have allowed for an additional dimension of understanding the analysis in how naming trends evolve over time and differ with respect from ethnicity to the next. In addition, utilizing **alternative clustering algorithms** like Gaussian Mixture Models **(GMM)** or **hierarchical clustering** might have provided more flexible clusters and improved the interpretability of the output results. Additionally, including more sophisticated encoding methods for ethnicity, such as embeddings that capture the relationships between different groups, could have potentially led to a more accurate reflection of cultural overlap in naming conventions across the various ethnicities. Overall, while there was some success in the build-up and overall execution of this analysis, these enhancements could have furthered the analysis’ understanding of the cultural influences on any baby naming trends that lie beneath the recorded data, and as a result could have potentially provided richer insights into the different dynamics at play within the analyzed dataset.