NYC Newborn Analysis 2011-2019

Nathan Lam, Hengyuan Liu, Timothy Tu
Yuki Urata, Shun Yao, Kahyun Jo

September 13, 2024

1 Abstract

The population of baby names has names come and go, and some have evolved. We wanted to see if we could predict sex based on ethnicity and name characteristics to see how these influences shift a name's sex. The dataset we used was publicly available data from the city of New York containing the top 75 most popular names for each sex and ethnic category for each year from 2011-2019. Through our analysis on the dataset in regard to sex, we saw that there was a high correlation between name endings with a vowel and sex, where female names were 10 times more likely to end with a vowel. Other factors, such as ethnicity, the amount of syllables in the name, the length of the name, and whether the name started with a vowel had statistically significant but small correlation with sex.

2 Introduction

2.1 Research Question and Dataset

For our research question, we would like to see if given certain features such as the ethnicity, name characteristics, and the year a baby was born, would we be able to predict the sex of a newborn. To be clear, we do not believe these factors cause the sex, but rather we would like to analyze the association of these factors with the sex of a newborn. Alternatively, we can say we are predicting a name's associated sex. The dataset we used was from the city of New York (NYC), provided by the Department of Health and Mental Hygiene (DoHMH). The dataset contains over 57,000 rows across 6 columns. 3 of these columns contain integer values, while the other 3 contain strings. The first column is *Year of birth*, which contains the year of birth for the row, ranging from 2011 to 2019. The next 3 columns are all strings, which in order are *Gender* containing the sex of the baby at birth (male or female), *Ethnicity* referring to the ethnicity of the biological mother, and *Child's First Name* as the name of the child. The ethnicity is split among 4 major categories, which included asian and pacific islander, black, hispanic, and white. The other two columns are *Count*, or the number of babies born with the specific name in the year and ethnicity, and *Rank* (the frequency of

name in descending order). According to the DoHMH, this data was "collected through civil birth registration", and is ordered by year and rank [4].

2.2 Data Preparation

When examining the data, we took a few steps in order to ensure that the data was properly inserted across the dataset. The first step was to drop duplicate rows, as there were multiple entries that were repeated, especially for the years 2011-2014. Within *Child's First Name*, we noticed that some names were lowercase, while some were all uppercase, so we converted the column to be all uppercase. We also removed any special characters that may have been present, which was just one apostrophe in one name. For the ethnicity column, we observed that certain ethnicity were shortened in the dataset, so we transformed any shortened versions into either "ASIAN AND PACIFIC ISLANDER", "BLACK NON HISPANIC", "HISPANIC", or "WHITE NON HISPANIC". We also renamed *Year of birth* to *Year*. Finally we also transformed the column containing the sex of the baby to a binary (1 for yes and 0 for no) to *is_female*.

In addition to cleaning the data, we also created a few columns based on features of the baby's name to see how they may help us predict sex based on ethnicity. These columns include *name_length*, *num_syllable*, *is_starting_vowel*, *is_ending_vowel*, and *is_unisex*. *name_length* and *num_syllables* are integers containing the amount of characters and syllables the name has, respectively. We used code that we found on stack overflow in order to determine the number of syllables [3]. The other 3 columns are all binary variables (using 1 for yes and 0 for no) which state whether the name starts with a vowel, whether the name ends with a vowel, and whether the name was found to be used for both male and female newborns, respectively.

3 EDA

3.1 Response - Sex

The primary response variable is the newly created *is_female*, indicating the sex of the baby (1 for yes and 0 for no). One unique note to point out is while there are more unique female entries within the table, there are overall more males that were born from this dataset. Another thing to note is that although there are 1974 different names when considering cases of both female unique names and male unique names, there are only 1938 unique names overall, as some names are unisex.

Sex	Total Unique Names (in Sex) From 2011 to 2019	Total Births
FEMALE	1077	266242
MALE	897	339862

Table 1: Summary of Unique Baby Names by Sex and Total Births

3.2 Predictors

3.2.1 Length and Number of Syllables of Name

Examining both *name_length* and *num_syllables*, we see that they are somewhat similar. This does make sense, as longer names are most likely to have more syllables. The proportions do appear more female heavy when the name is longer. In addition, most names tend to be between 4 to 7 numbers long, with an average of around 2 syllables.

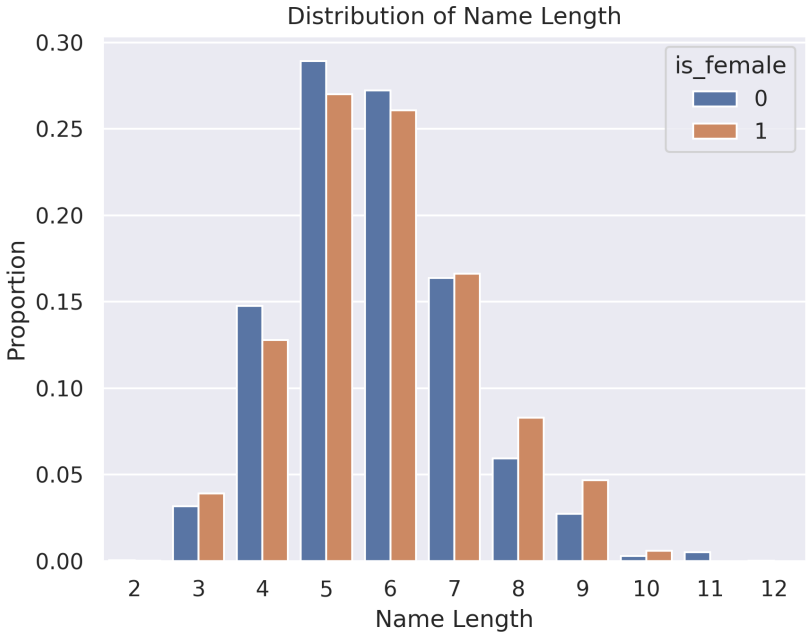

Figure 1: Histogram of Proportion of Sex by Name Length

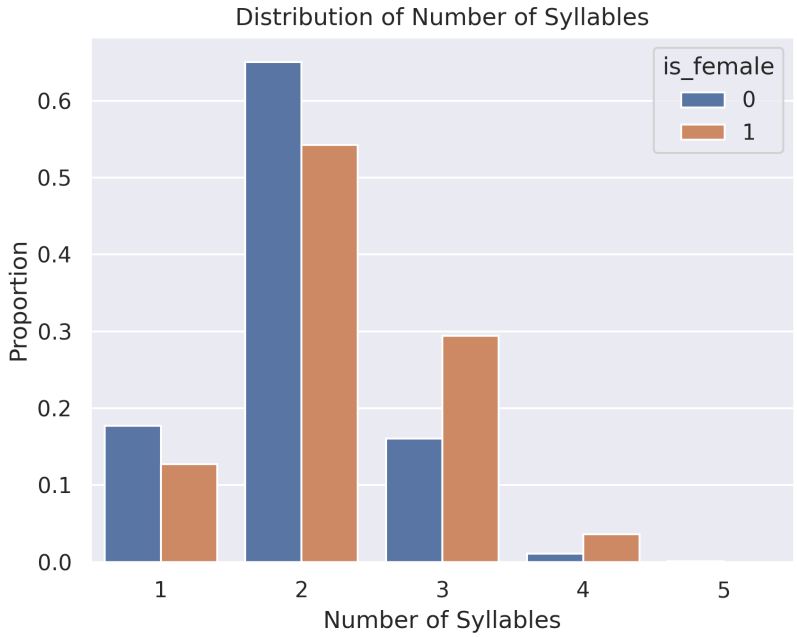

Figure 2: Histogram of Proportion of Sex by Number of Syllables in Name

3.2.2 Starting or Ending with a Vowel

The proportion of names that start with a vowel for both females and males seem pretty similar. However, we do see a noticeable difference for names ending with a vowel, where a large proportion of female names end with a vowel compared to only about a quarter of male names ending with a vowel. This intuitively does make sense, as some traditionally masculine names can be made more feminine by adding a vowel at the end, such as the change from Christian to Christina.

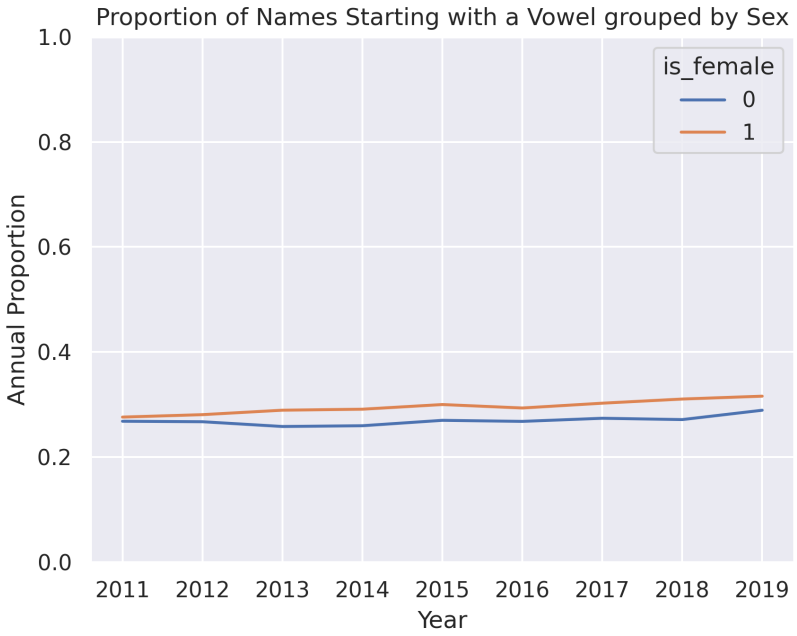

Figure 3: Proportion of Names Starting with a Vowel grouped by Sex

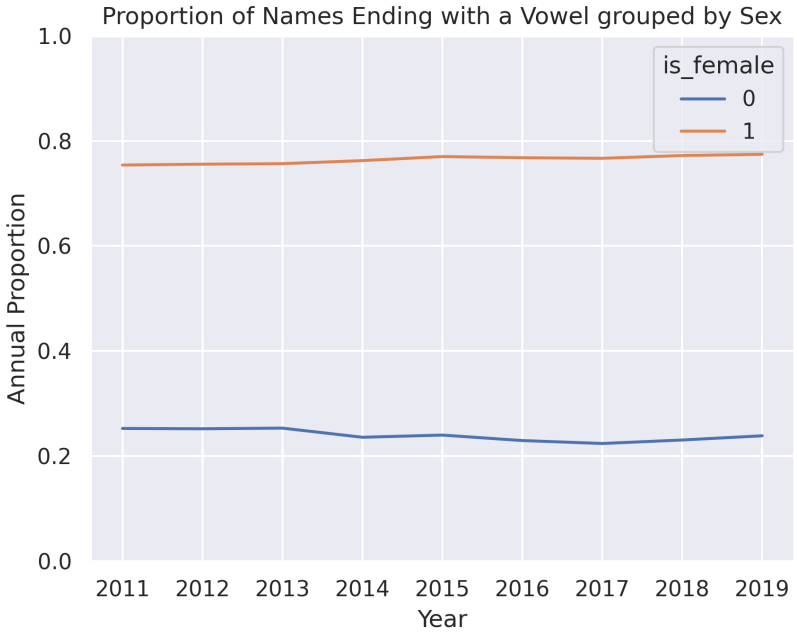

Figure 4: Proportion of Names Ending with a Vowel grouped by Sex

3.2.3 Unisex Names

Unisex names do not make up a large proportion of our data, with only 36 names occurring in both female and male newborns. However, we noticed that out of the newborns with unisex names, there is a larger proportion of males. This could be due to unisex names initially starting as male before shifting to be used as female. For example, Charlie is a name that has recently shifted to female.

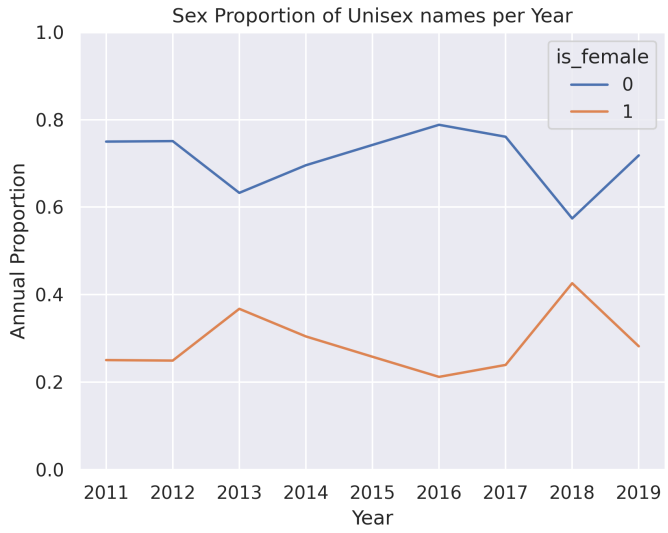

Figure 5: Proportion of Strictly Unisex Names by Sex

3.2.4 Ethnicity

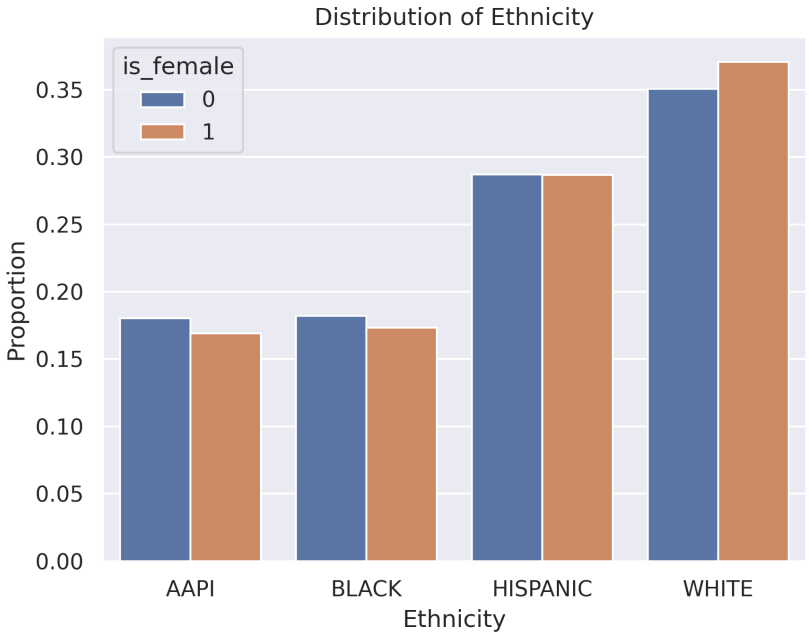

Figure 6: Bar Chart Showing the Distribution of Names associated with Sex among Ethnicity

The bar chart above shows the distribution of names across the ethnicities for each sex. From the bar chart above, we can see that our largest ethnic group is White, followed by Hispanic, and then a similar amount of both Asian and Pacific Islander and Black. There appears to be a minimal amount of variation between the proportion of male and female names.

3.2.5 Overall Correlation between Columns

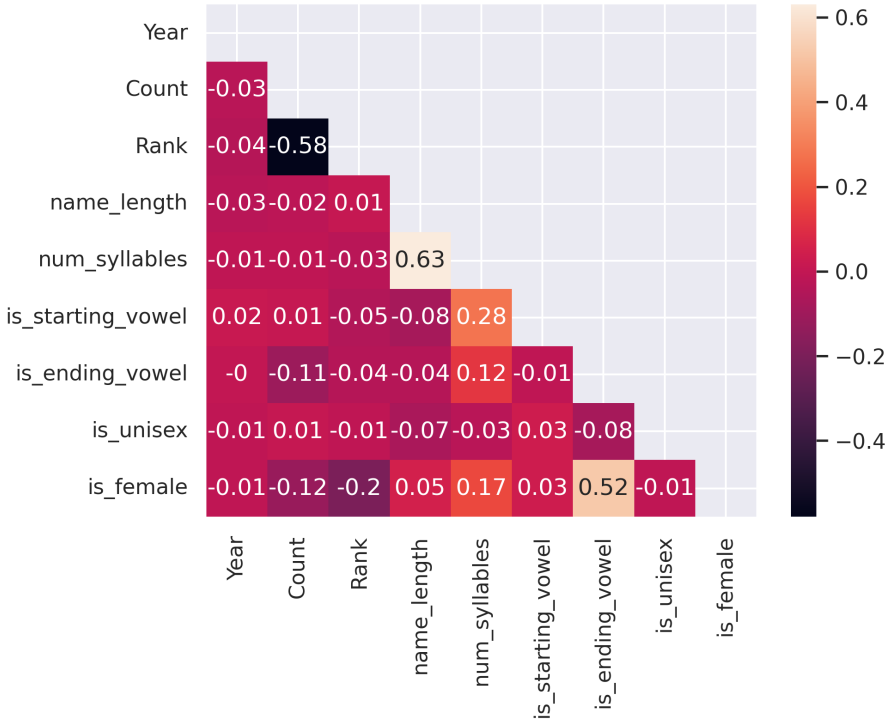

Figure 7: Heatmap showing correlation of columns

We also created a heatmap to show what columns or variables we may or may not expect to be related. The heatmap shows the correlation rounded to two decimals between two different columns and is color-coded based on how strong or weak the correlation is. As we stated earlier, there is some moderate correlation between *num_length* and *num_syllables*. We also see that *Count* is inversely correlated with *Rank*, which should occur, due to rank being based upon how popular a newborn name is.

3.3 Most Popular Names

In addition to looking at the predictors in relation to the response, we also examined some interesting trends, looking at some of the most popular names overall and among ethnicity and gender.

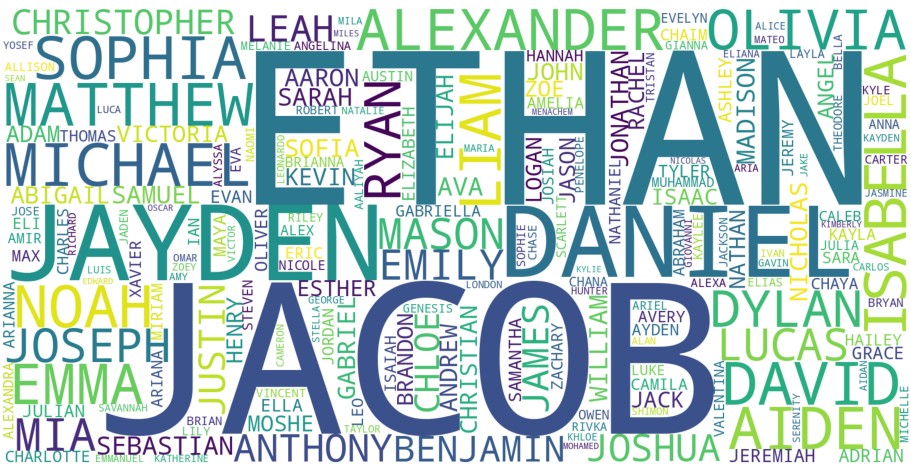

Figure 8: Word Cloud of the Most Popular Names Overall

	Child's First Name	Count	Proportion
0	ETHAN	5867	0.009680
1	JACOB	5649	0.009320
2	LIAM	5453	0.008997
3	JAYDEN	5210	0.008596
4	NOAH	5171	0.008532
5	SOPHIA	4814	0.007943
6	DANIEL	4753	0.007842
7	MATTHEW	4649	0.007670
8	DAVID	4617	0.007618
9	ISABELLA	4584	0.007563

Table 2: Table of the Most Popular Names Overall

The most popular name in NYC is Ethan, with almost 6,000 newborns donning the name. According to babycenter, Ethan was the top name across the US in 2009 - 2010, but has slightly fallen off, albeit remaining in the top 20 names for boys [1]. Earlier, we saw that there were more unique female name entries, but overall less females than males in terms of recorded births. This is somewhat reflected in the top 10 names, as 8 names are typically associated with males, Jayden being the only unisex name but still primarily given to males. Another thing to note is a majority of these names have religious connotations, in particular Hebrew biblical names such as Ethan, Jacob, and Noah.

3.3.1 The Most Popular Name: Ethan

Just for fun, we did explore certain names. The first name we examined that we found interesting was Ethan, as it was the most popular name.

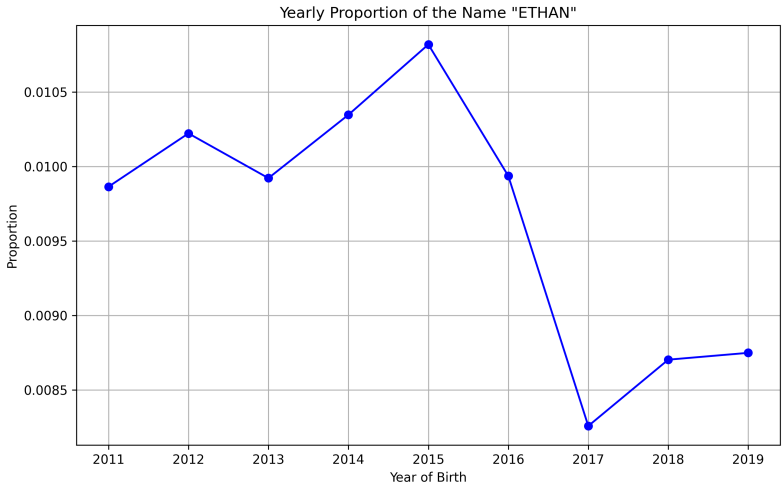

Figure 9: Proportion of Newborns Named Ethan

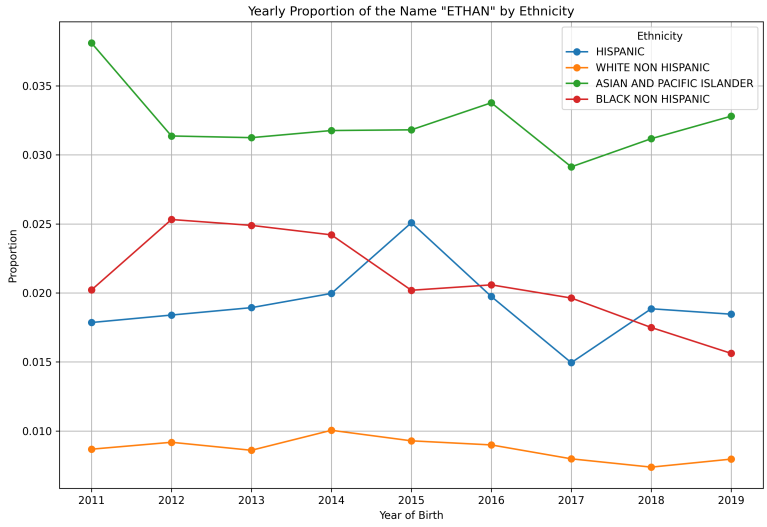

Figure 10: Proportion of Newborns Named Ethan by Ethnicity

In New York City, the proportion of newborns given the name Ethan reached a peak in around 2015, and has usually remained in the range around 1%. By ethnicity, it is most popular among Asian and Pacific Islanders.

4 Methodologies

4.1 Logistic Regression

To investigate whether we could predict sex based on ethnicity and name characteristics, we implemented a logistic regression model with the response variable *is_female*. For the predictor variables, we used *Year*, *Ethnicity*, *name_length*, *is_starting_vowel*, *is_ending_vowel*, and *is_unisex*. We are using a rather simple form of logistic regression, as we have the single response variable that is binary based on multiple categorical predictors. The dataset is split into training and test sets, with 30% of the data allocated for testing. After fitting the model, the coefficients of the model are extracted and transformed to log odds and odds ratios for interpretation.

Feature	**odds**	**p-value**
Intercept	1.000344	0.736
Year	0.99899	0.626
name_length	1.151292	0.000
is_starting_vowel	1.302457	0.000
is_ending_vowel	10.770005	0.000
is_unisex	1.091374	0.000
BLACK_NON_HISPANIC	1.039453	0.683
HISPANIC	0.982591	0.196
WHITE_NON_HISPANIC	0.973729	0.844

Table 3: Logistic Regression Coefficients and Odds Ratios

The table above showing the odds ratios and p-values exhibits that the features such as *name_length*, *is_starting_vowel*, *is_ending_vowel*, and *is_unisex* have positive coefficients, indicating higher odds of the name being associated with a female. The most important feature used in predicting the gender of a baby is *is_ending_vowel*. The names ending with a vowel have a significant positive effect, which in context we would conclude that . From the p-values, we can also see the year and ethnicities were not statistically significant.

The accuracy and AUC (Area Under the Curve) metrics are utilized to assess the model's performance, with an accuracy of 76.37% and an AUC of 78.69%.

Figure 11: ROC curve for Logistic Regression

4.2 Decision Tree

In addition to performing a logistic regression, we thought it would also be interesting to see if we could create a decision tree in order to predict sex. A decision tree can be useful because it provides us a clear delineation on what factors are important in making a decision on sex by the model. We used a decision tree for predicting sex, or *is_female* based on the same predictors as the logistic regression along with *Year* and *Count*. The root node splits on whether the name ends with a vowel. If it doesn't, the majority class is male. Otherwise, the majority class is female.

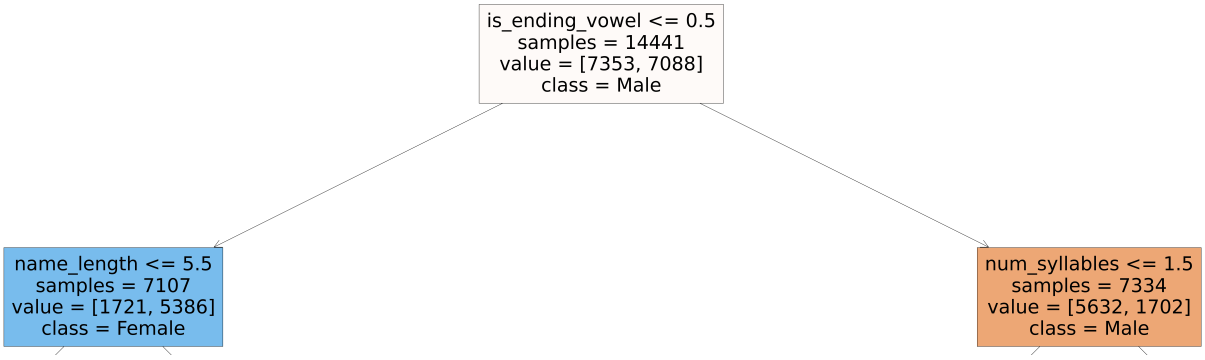

Figure 12: Decision Tree Predicting Sex

The decision tree up above has a maximum depth of 3, but has been truncated due to deeper nodes keeping the same decision regardless of other factors. The tree shows that

the most significant factor in sex prediction is whether the name ends with a vowel. The overall accuracy of the decision tree model is approximately 76%, indicating a relatively good performance in predicting sex.

5 Conclusions

5.1 Findings

From both the logistic regression as well as the decision tree, we were able to reach an accuracy of around 76% in predicting the sex of a newborn given certain factors. The most significant factor we observed in predicting sex for both models was whether the baby's name ended with a vowel or not, being the main determining factor in the root node for the decision tree and the factor with the largest odds ratio for the logistic regression model. While the decision tree was unable to use any other predictors than the name ending with a vowel, the log odds from logistic regression did point to certain factors such as longer name length and the name starting with a vowel being slightly more likely to be attributed to female names.

5.2 Weaknesses and Improvements

One of the limitations we faced within the dataset was the amount of predictors we could generate. If we spent more time, we possibly could have either found a dataset with more predictors we could use or brainstorm more ways we could find more predictors from the given dataset, although it would be difficult since there are certain pieces of information that people may want to keep private related to having a newborn. An alternative way we could have bolstered our research would have been utilizing another dataset that provided data regarding names and cultural relevance or societal trends. A dataset like this could provide us with information on why certain names spike in popularity or what events have cultural impact in affecting names given to newborns.

References

[1] Tahirah Blanding. Syllable count in python. Last accessed 30 May 2024.

[2] Richard Harris. Why are more baby boys born than girls. Last accessed 31 May 2024.

[3] Jeremy McGibbon. Syllable count in python, 2017. Last accessed 30 May 2024.

[4] City of New Yotk. Popular baby names, 2023. Last accessed 30 May 2024.