Code

Baby Names of Ireland Exploratory Analysis

Shashank Sanjee Venkata Chalapathi (Graph 1, Graph 2, Plot_ly, Report Generation)
Azhar Shaikh (Data Manipulation, Graph 4, Graph 5, Cosmetic changes)
Cora Avril Leahy (Linear Models, Prediction, Graph 3, Graph 6)

Introduction

Problem Statement:

The objective of this project is to explore the Babynames dataset which helps to answer some interesting questions like which names are popular in the year 2009 to 2018, most popular unisex names and many more.

Insights:

- Some very interesting insights were obtained from this analysis-
  - Number of newborn babies is reducing year by year. Count of total number births tells us that decreasing the number of babies.
  - Ratio between most favourite and least favourite names is almost 5:1. This is explained in the below graphs.
  - On average, ~38 boys have the same name and the count is ~140 for girls.

Data Preparation

Data Import

This dataset is imported from the Central Statistics Office (CSO) Ireland website. The dataset contains Babynames in Ireland for 10 years with a total of 20,450 unique names which named for 6,15,468 babies. More information regarding the dataset can be found here (https://www.cso.ie/en/aboutus/whoweare/)

Data Manipulation

We took a total of 10 years of data i.e. from 2009 to 2018 and merged the files by year-wise, creating a new column called Year. The following code has been used for merging the code.

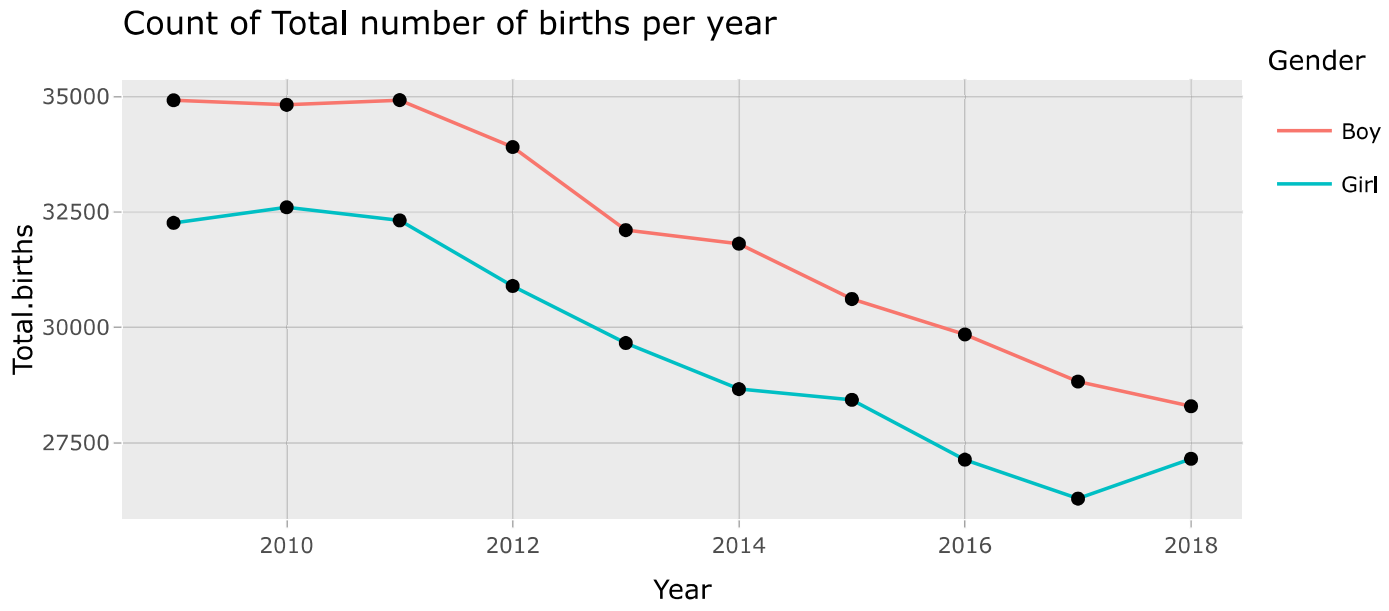
Code

Exploratory Data Analysis Description

1. What is the trend of Number of Births for the given data set?

The following graph shows the total number of babies born for both boys (Red) and girls (Blue) in each year from 2009-2018.

Code

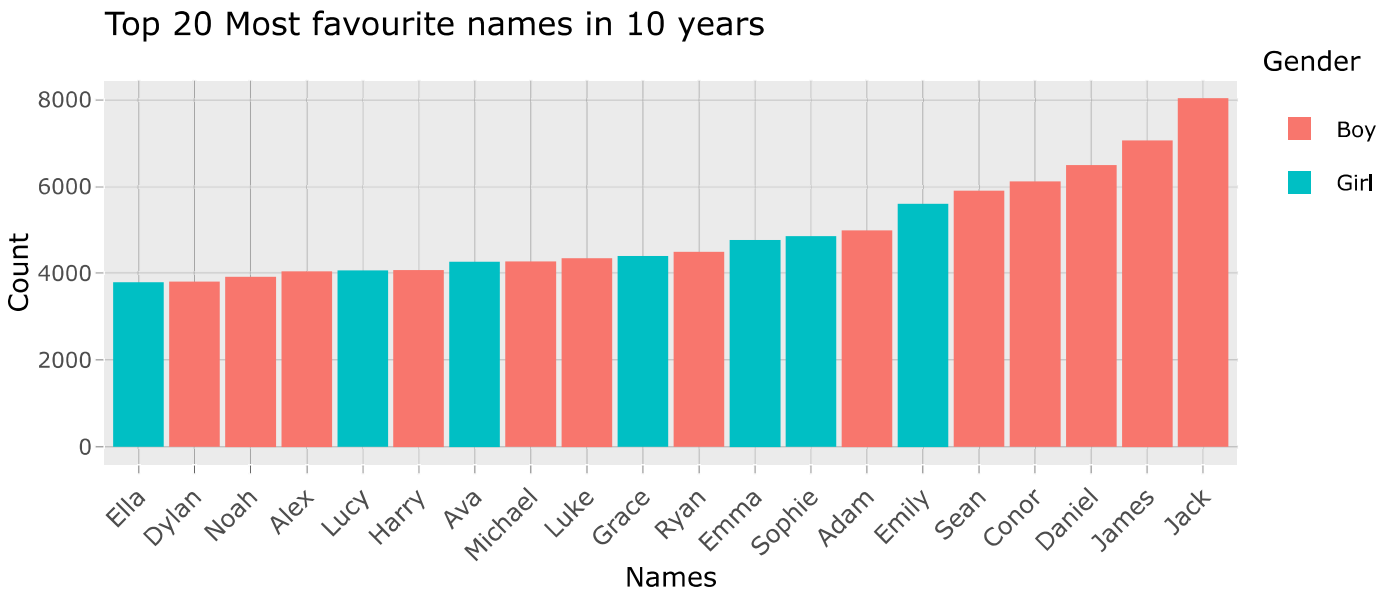


It can be observed that the number of babies born in 2009 to 2011 was almost the same with a little variation but after 2011 there is a steep decline in both boys and girls' birth rate.

2. Which names were popular during the year 2009-2018?

This bar plot shows the count of most popular names given to babies from the year 2009-2018.

Code

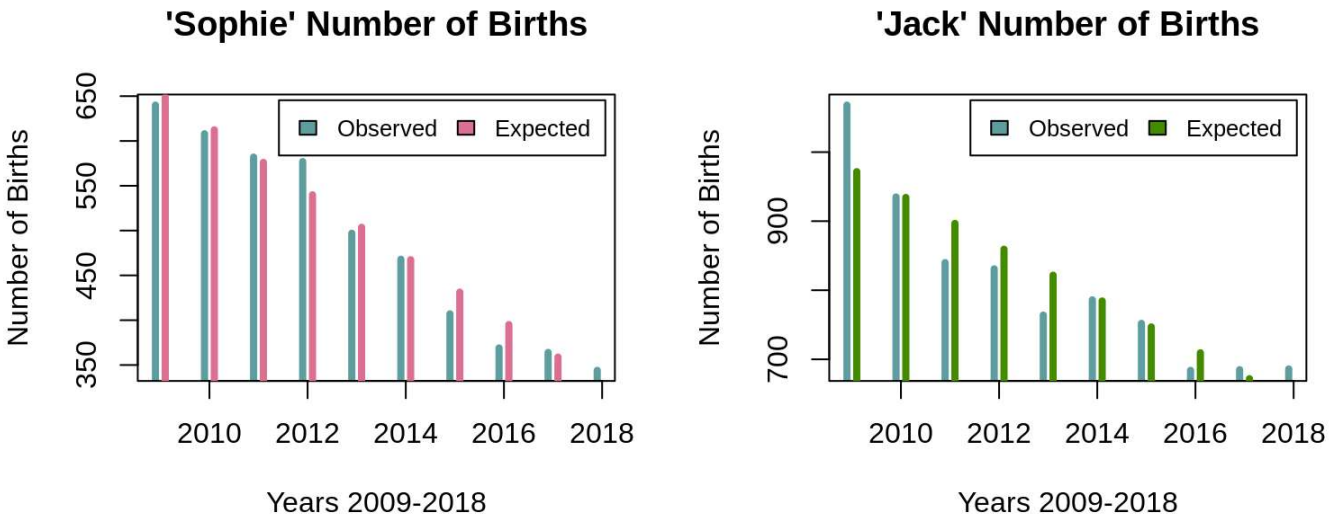


It is visible that "Jack" and "Emily" names were the most favourite names in boys and girls respectively. It can also be observed that in the top 20 names there are more boys(13) than girls(7).

3. Is linear regression modelling capable of producing reasonable predictions, using Jack & Sophie as variables?

For our analysis, we created subset groups for Sophie and Jack for over 10 years. Both names have consistently appeared within the top 5 ranking names for our chosen period. Baby Names are not known to follow a normal distribution but with the large volume of data and timespan, it will be interesting to see if we can adapt the theory to these chosen variables.

Code



Comparing the 'fs' and 'fj' graphs, a fitted linear regression line for both Sophie and Jack respectively. Sophie appears to be more fitted to the blue linear line than Jack, this is unexpected considering Jack has consistently maintained a no.1 Ranking for 9 out of the past 10 years. Sophie's ranking has gradually decreased over the years but the number of births has not fluctuated as much compared to Jack. Despite the ranking, Jack has gone from 1068 births in 2009 to 686 in 2018 while Sophie went from 640 to 344. The ranking appears to not play a significant role in the expected number of births.

Using these fitted lines, we generated graphs comparing the expected number of births to the observed number of births ('sophieExpect' and 'jackExpect'). We tested these results under a 95% confidence interval (CI) under 'sophieCI' and 'jackCI' and added the findings to their tables. Under a 95% CI, the predicted expected number of births are reasonable values e.g. in 2014, 648 Sophie was born with an expected

value of 467.5, this value falls within the CI range (418.3 to 516.6).

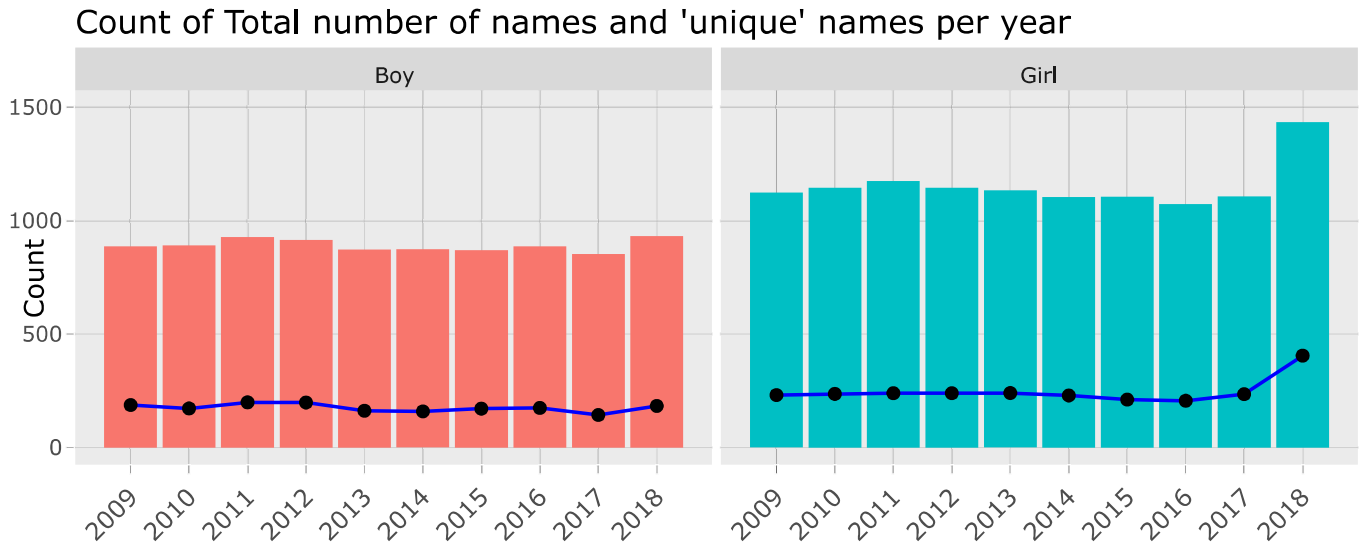
These fitted lines were also used to generate furtire prediction over the next 3 years, 'sophiePre' and 'jackPre'. For 2019, the expected total number of Sophies and Jacks to be born are 286 and 597. Trends are difficult to predict far into the future so it will be interesting to compare our 2019 with the next updated release of the CSO 'Baby Names of Ireland'.

To summarise our findings we created 2 graphs entitled ''Sophie' Number of Births' and ''Jack' number of Births'. These graphs tell us that as seen earlier when fitting the linear regression line, that the observed frequency for Sophie matches well except for the years 2015 & 2016. As expected for Jack, there is more disparity between the observed and expected frequencies, all while still being considered reasonable values.

4. What is the unique number of names in the last 10 years and how many of them were the least favourite?

The above bar plot represents the total number of names and unique(minimum frequency of 3) names in a particular year.

Code

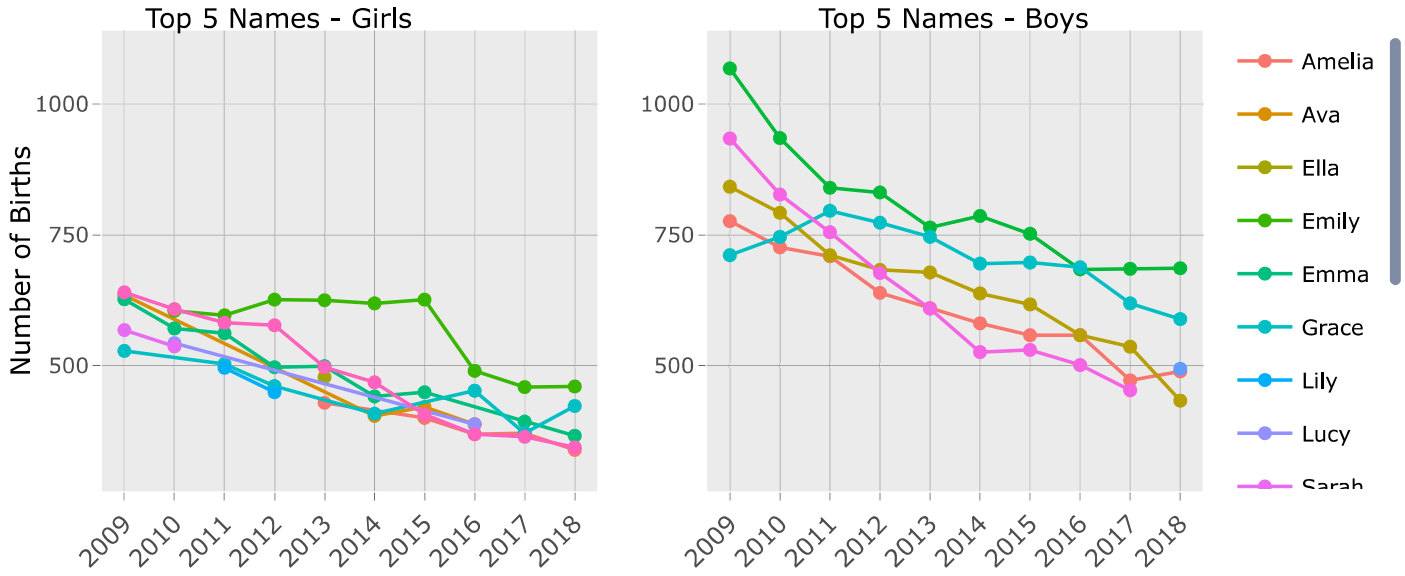


We can summarize that out of a sample of 100 names 20 were unique i.e. least chosen. It can also be observed that there is no variation in the number of baby names for boys but we can see an increase (20%) in the year 2018 for girls.

5. What observations can be made from the top 5 ranked names for Girls and Boys 2009-2018?

We created 2 subset tables entitled 'top5girls' and 'top5boys'. These tables contain the top 5 ranked names per girls and boys over 10 years. The 'top5girls' table contains 1 extra row due to Sophie and Amelia sharing the 5th rank of 2016.

Code



From the "Top 5 Names - Girls" graph we can observe an outlier 'Emily' which appeared within the top 5 in 2010. The trajectory of 'Emily' showed consistent values from 2012 to 2015 until the number of births fell sharply in 2016 to 490, a decrease of 136 births from the previous year. If we used 'Emily' as our variable in the above linear regression modeling, we would not be able to produce any reasonable predictions.

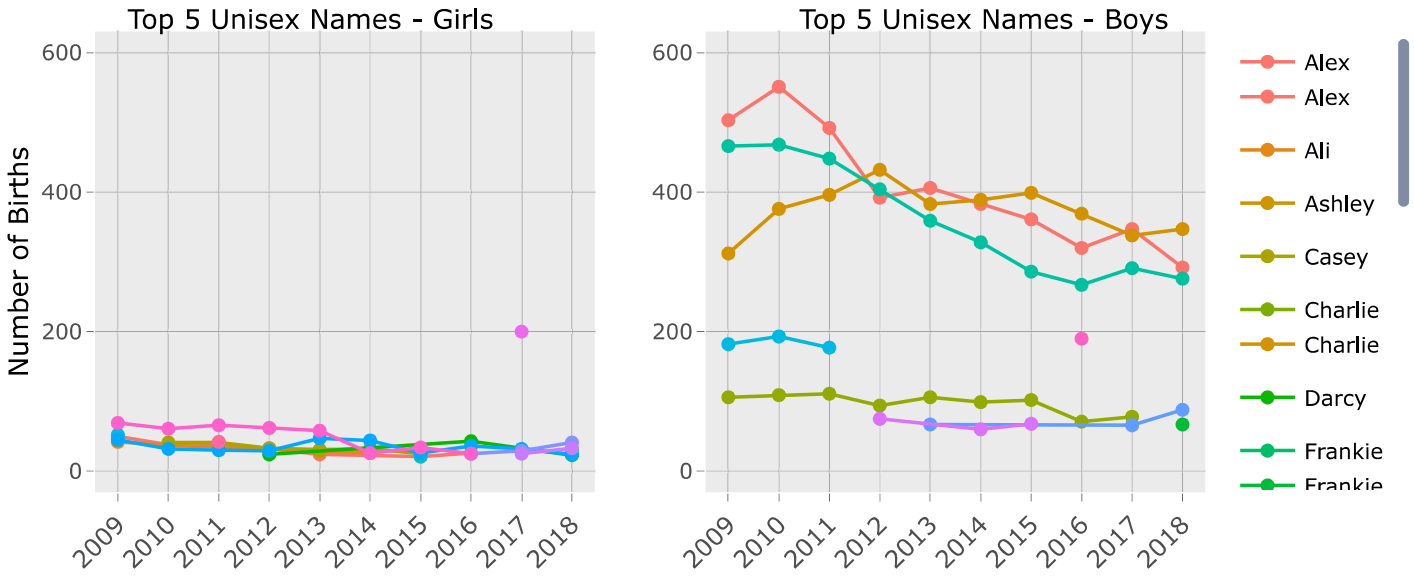
This highlights the unexpected effects outside circumstances such as pop culture, have on popular baby names and trends in general.

From the “Top 5 Names - Boys” graph we can see a more uniform pattern. In comparison, all the names are showing a similar downward trend in the number of observations. Referring to the “Count of Total number of births per year” from **graph 1**, the numbers of baby girls and boys being born have been in decline, we can infer that this downward trend is possibly due to the decline in births.

6. Are unisex names playing a role in the decline of traditionally popular baby names?

For our final analysis, we are looking at the occurrences of ‘unisex names’, names that can be given to both girls and boys. We chose ‘unisex names’ due to interesting social changes within the last few years such as the increasing influence of gender equality.

Code



Both graphs show the top 5 ranking unisex names per year e.g. In 2010, in ranking order the top 5 names amongst boys were Alex, Jamie, Charlie, Jayden, and Dara. An overview of both graphs shows a surprising decline through the mid-section which leads to a plateau. This pattern is similar to the trend of the top 5 overall for girls and boys. Another immediate observation is the lack of overlapping in a given year between both genders, at most girls and boys share 2 names within their top 5.

Another interesting observation is that despite these names being considered unisex, there is a disproportionate number of boys given these names. ‘Alex’ appears consistently in both genders up until 2016 amongst girls. At its peak, it accounted for 551 boys and just 50 for girls (it last ranked within the top 5 among girls in 2016). ‘Frankie’ a recent entry in 2018 for both genders accounting for 31 and 67 births for girls and boys respectively. These results can partly be explained by the volume of selection available, historically girls have a wider variety of names to choose from, out of the 21 names graphed above, 16 appeared within the girls top 5 rankings. Therefore the spread amongst boys’ names will naturally be more contained amongst a select few names.

Conclusion

- Total births for boys (~19%) and girls (~16%) have steadily dropped from 2009 to 2018.
- Top 5 are boys in the “Top 20 Favorite Names”, followed by Emily, the most liked girl name.
- Using Sophie and Jack’s linear regression fit we can conclude that Sophie’s expected value is very close to the actual value. While there are variations between expected and actual values in Jack’s scenario.
- While for both Boys and Girls, the overall number of names and Unique (i.e. least chosen) names is somewhat consistent, there is a significant increase (~20%) in these two values for 2018.
- The no. of births in Top 5 category for both boys and girls have also declined gradually over the last 10 years. Sophie is the top-ranked girl in 2009 and 2010. Emily has been the top ranked girl for all the other years. In Boys, James is the top ranked in 2016, while Jack is the favorite name for the rest of 9 years.
- The boys’ unisex names have more variation than the girls’ unisex names. Compared to boys there are more girls to choose from.