**Project for Data Visualization (ME8135)**

# "Analytics and Data Visualization of baby name in New York City"

**Shailendra Khadka Yadav**

## 1. Introduction:

Visualization is the graphical presentation of information and data, with the goal of providing the viewer with a qualitative understanding of the information contents. It is also the process of transforming objects, concepts, and numbers into a form that is visible to the human eyes.

Data visualization is an effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization approaches.

The steps in data visualization are acquired, parse, filter, mine, represent, refine and interact [2]. Data can be collected from many resources such as books, files and digital documents. This is the beginning and fundamental step of data visualization. It is necessary to restructure the collected data. This structure will make it easier to know convey to others what data format, tags, names, and indices are about. However not all data is useful. For example, if one is focusing on the data of a specific period, remove the data of other periods. Data visualization is to help viewers seek for insights that may not be gained from raw data or statistics. Thus mining step helps get basic understanding of the data that is significant for the whole process. After mining, various visual models available are used to represent data using suitable type. The so represented data can be polished according to some basic color and graphic design theory. As a part of interact, one can add methods for manipulating the data or controlling what features are visible [2].

The rapid evolution of technology providing Internet access almost to all the developed parts of the world intensify the internationalization in our changing society. Consequently people being influenced by the same sources tend to make decisions about names in a similar way. Thus many of the platforms are being used for finding appropriate names. There are two primary portals that someone is able to explore baby names, also with the help of visualization tools via some graphs and heat maps. One can find them under the names of "Name Voyager" (http://www.babynamewizard.com/voyager) and "Name Trends" (http://nametrends.net/) [1].

This study includes Analytics and Data Visualization of baby name and birth in New York City, which describes the different visualization on the basis of baby name, county, and sex. As part of

this study, the study focuses on explaining the most popular baby name, birth analysis based on year, county and sex on the different attributes. The data source is taken from https://health.data.ny.gov/Health/Baby-Names-Beginning-2007/jxy9-yhdk [3].

Based on the above dataset information is inquired, sorted, grouped, and visualized for preferred analysis.

## 2. Problem Statement

It is increasingly observed nowadays the globalization phenomenon even to the naming patterns, so that children are more and more citizens of the world. We firmly believe a culture is mirrored in naming practices and changes in society are reflected in the names chosen by the parents. According to the observations from everyday life as well as from different surveys or articles, there is a considerable speculation as the parents realize that their offspring will be part of a more wide - ranging network and not only of their specific home town, which affect the name choices about them.

Consequently people being influenced by the same sources tend to make decisions about names in a similar way, often following their favorite celebrities' standards. Thinking about English as the first internationally spoken language, it is found the comparison, about the most common baby names as a reasonable one.

The goal is to study the similarities and differences, concerning the naming tendency.

Exploring the naming analogies or differences between counties, in an easy and comprehensive way. It is also inspired all these young parents who would like to get some ideas in naming their children.

## 3. Dataset Descriptions and Data Source:

The input dataset consists of 145,570 rows representing newborn children's names in the New York State Baby Names, aggregated by name, gender, year and state or county where the mother resided as stated on a New York State or New York City (NYC) birth certificate. The frequency of the Baby Name is listed if there are 5 or more of the same baby name in a county outside of NYC, or 10 or more of the same baby name in a NYC borough.

The major attributes mentioned in the dataset are

- Year (numeric) : In which Year the baby was born

- First.Name (string) : First Name of the baby
- County (string) : Name of the state
- Sex (string) : Gender of the baby
- Count (numeric) : Count the baby of that particular first name

The data source is available in following website:

https://health.data.ny.gov/Health/Baby-Names-Beginning-2007/jxy9-yhdk

## 4. Technologies and Issues

Implementation of analytics and visualization for this project is done using R. The following R packages are needed to visualize the above dataset in different graphical presentation:

- ggplot2
- Ploty
- RColorBrewe
- Wordcloud

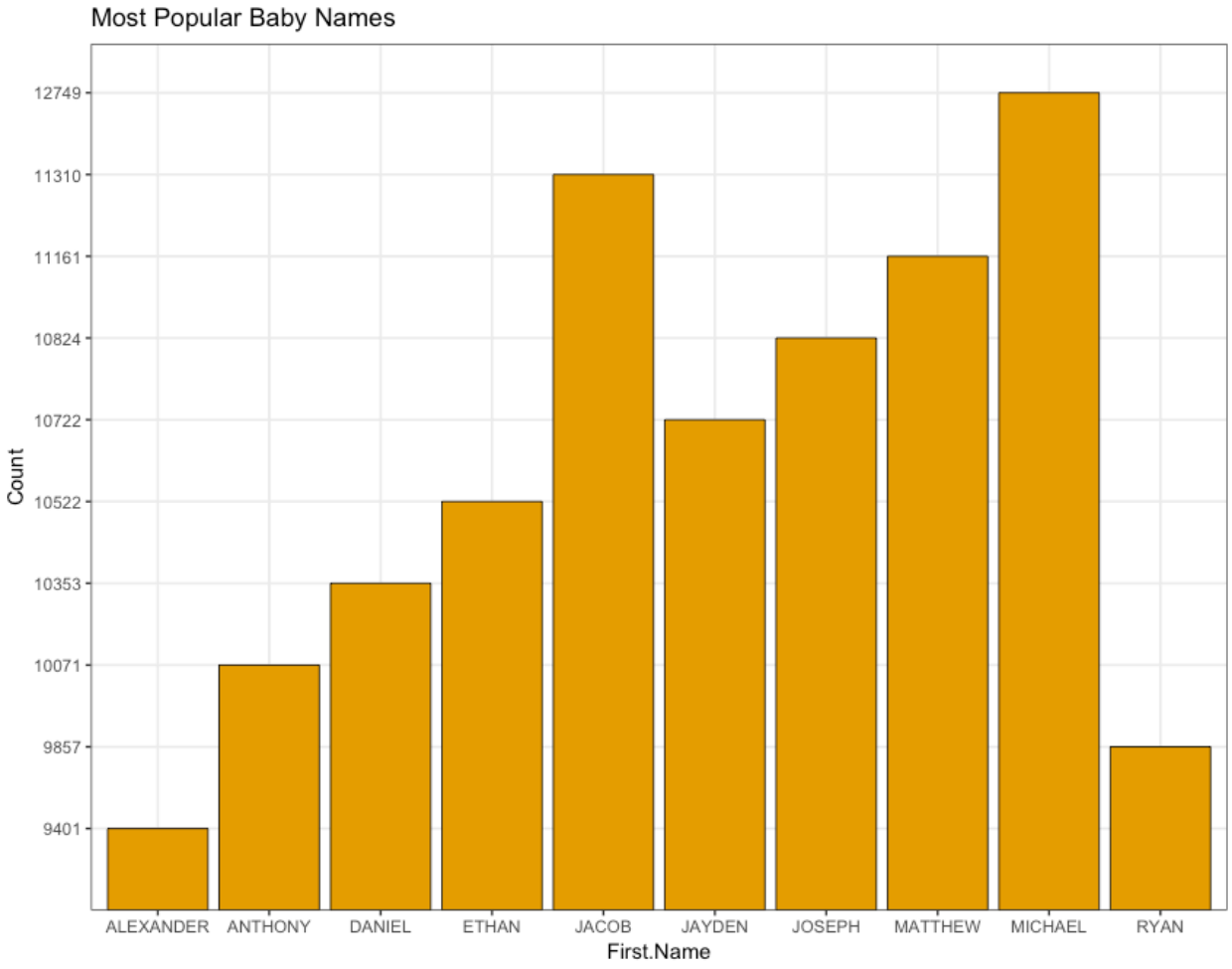
**Pre-processing done to the data**

Since the input dataset consists of 145,570 rows and 5 variables, the same County name is recorded as uppercase and lower case. For example kings and KINGS seems like two different county. So, the analysis is done after pre-processing to the data by converting all County name into uppercase.

## 4. Preferred Analysis and Visualizations

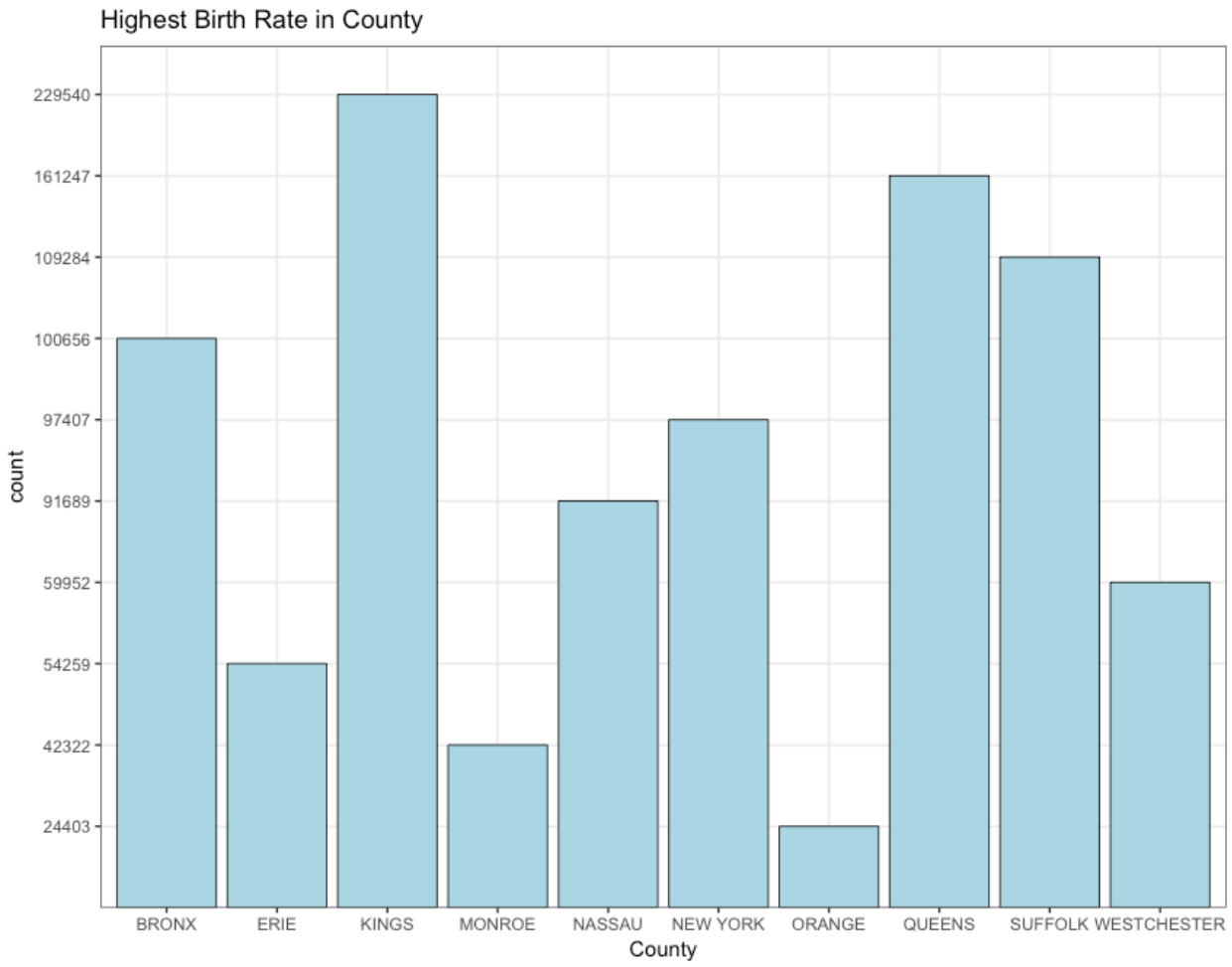
Different analysis is done to analyze the above dataset by different attributes. The major expected visualization are as follows:

- **Most popular first names for baby (Male and Female)**




The above Bar graph is shown that the most popular first names for baby (Male and Female) is “Micheal”. This name is repeated for 12749 times.

• Which County have highest birth of child?



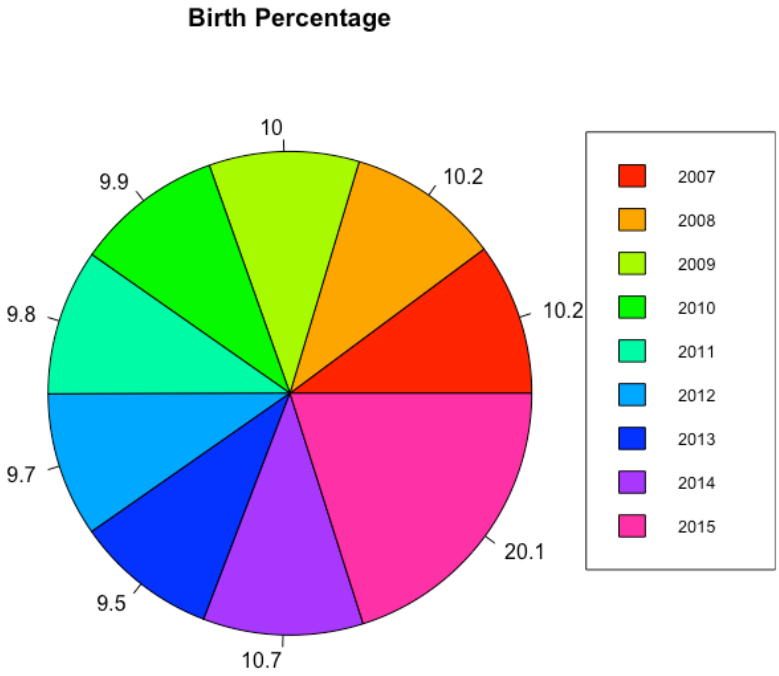

The above Bar graph is shown that the County have highest birth of child is “Kings”. This name is has  229540 counts.


The visualization of Wordcloud is used to show the most popular baby county name as KINGS.

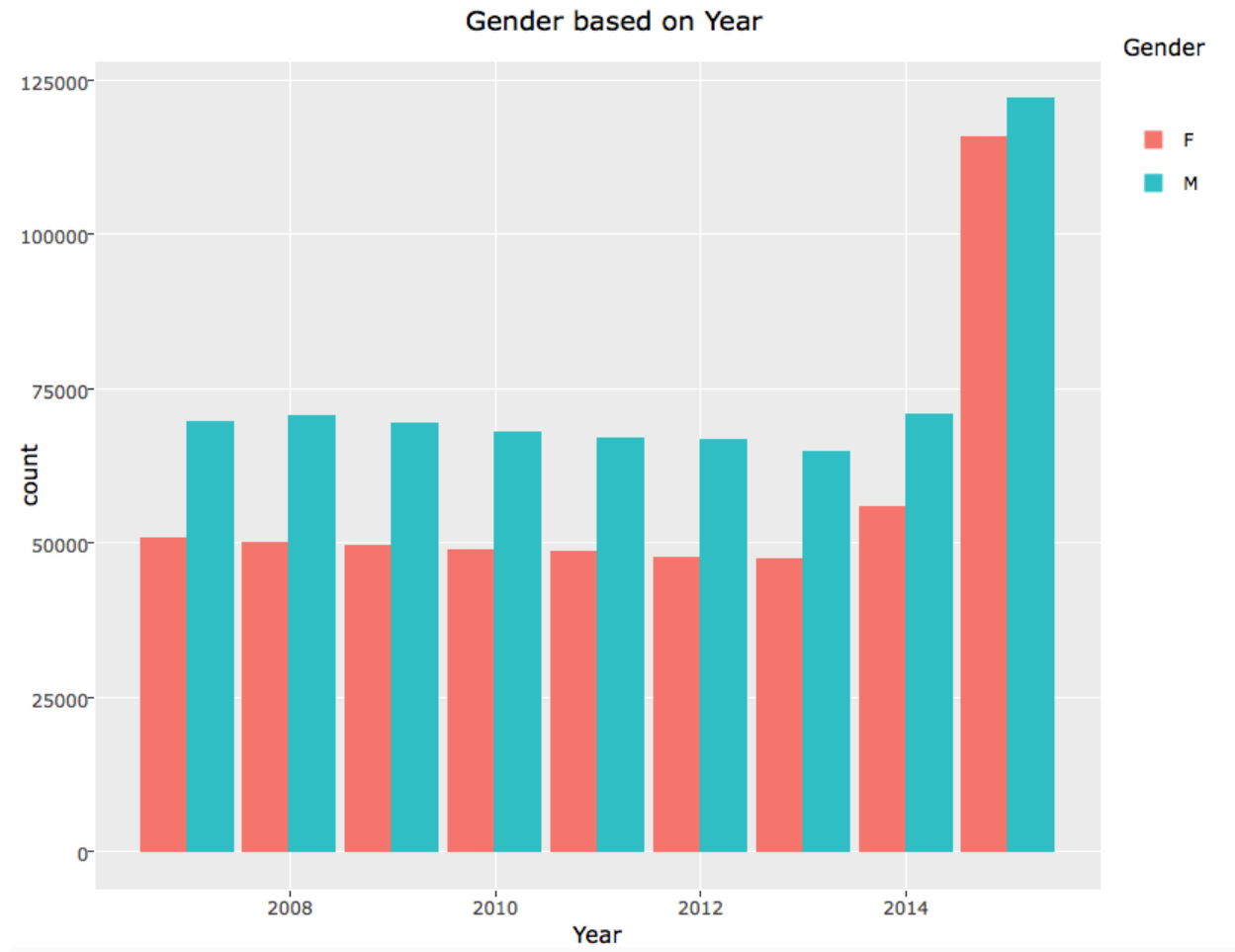
**• Year wise birth percentage of baby**

Here, the below pie chart shows the year wise birth percentage of baby.

The highest birth percentage rate (21.1%) is in year 2015. Second highest is in 2014 (10.7%) and the lowest birth percentage rate is in year 2013(9.5%).

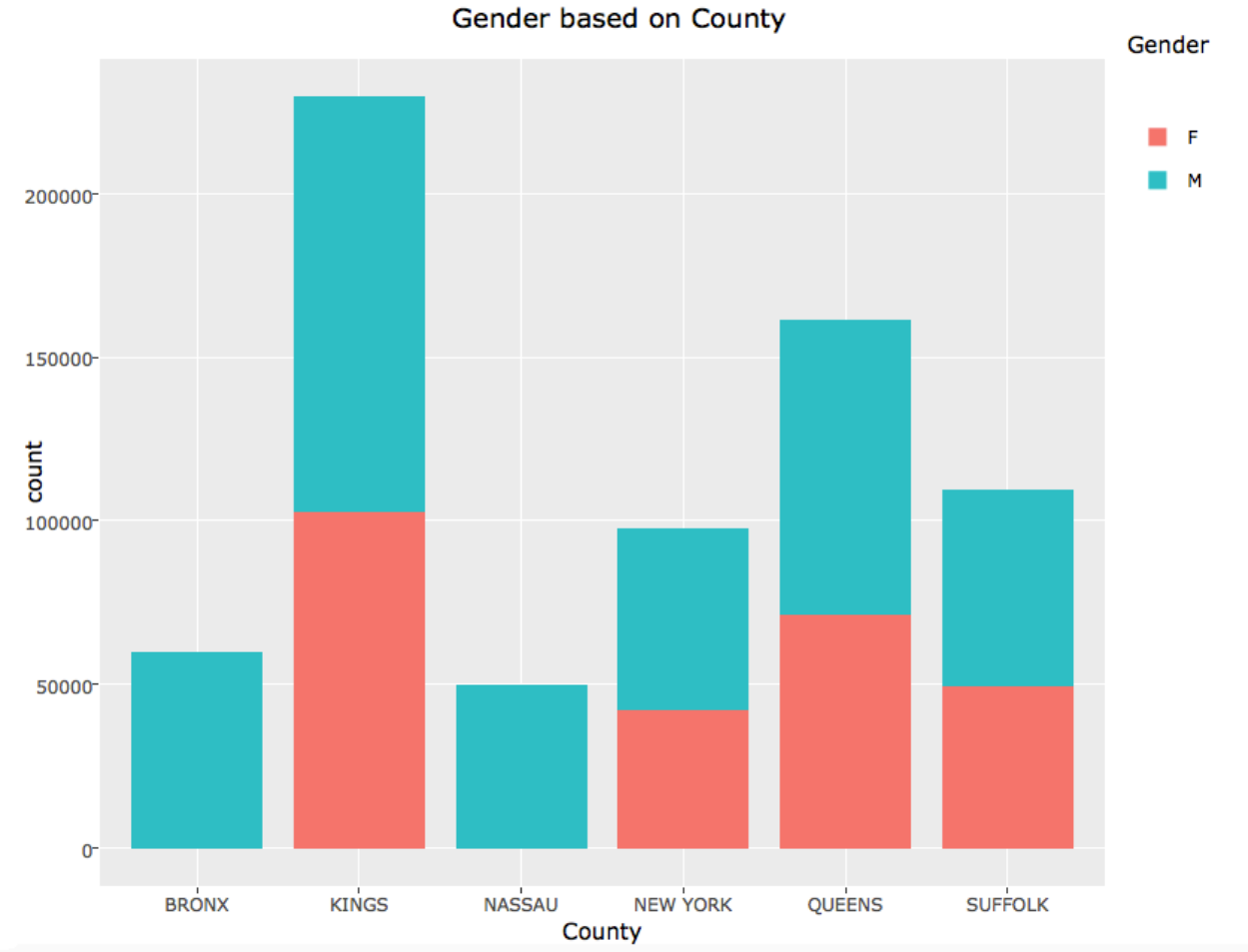


• **Gender wise baby birth analysis based on Year.**

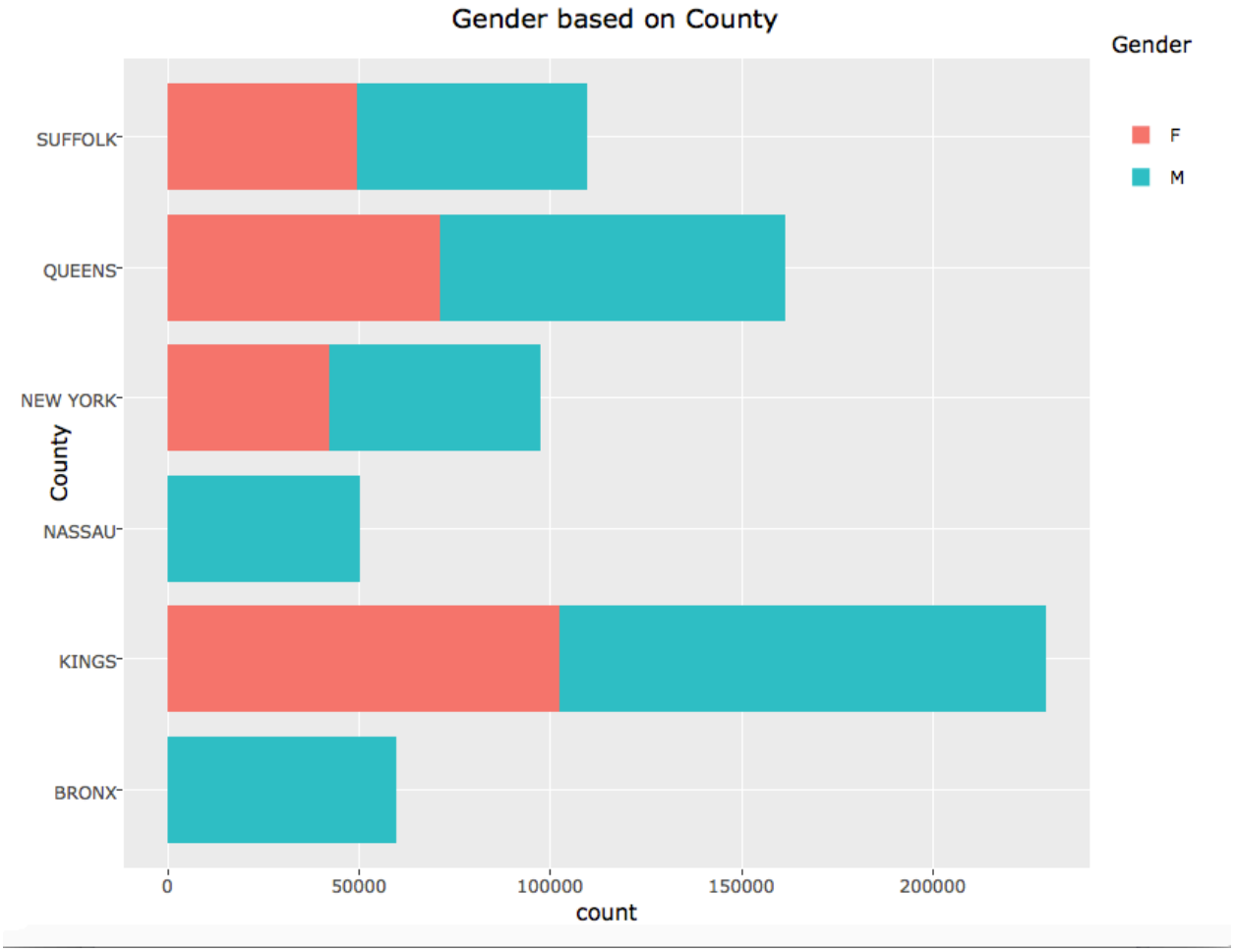



The above Bar graph is shown that the Gender wise baby birth analysis based on Year. It is shown that in year 2015 has the highest.

- **Gender wise baby birth analysis based on County.**




The above Bar graph is shown that the gender wise baby birth analysis based on county. It is clearly seen that KINGS has more birth rate than compare to other county.




Bar graph to show the gender wise baby birth analysis based on county highest baby birth is in Kings County.

## 5. Results and Analysis

Based on the implementation and observations on the data sets, the data results were sorted, grouped and visualized for interpretations. For the data set, the most popular names found were Micheal. Accordingly, the baby birth rate during 2007 to 2015, it was found that in the year 2015 had highest baby birthrate and in the year 2013 had the lowest baby birthrate on the various counties from the given data set. Among them the Kings County found to have largest number of babies and the figurative results for this are 229540.

## 6. Lessons Learned

This project has enriched my knowledge on R programming to prepare different visualization. I have learned how to create different variables in a single graph and how to solve various problems in coding.

I have worked on wordcount visualization, which is really exciting for me to get the more idea about visualization.

## References

[1] Magdalena Pöhl, Tsafou Kyriakoula and Michael Oppermann, Baby Names Explorer – Final Report

[2] https://www.dashingd3js.com/the-data-visualization-process

[3] https://health.data.ny.gov/Health/Baby-Names-Beginning-2007/jxy9-yhdk.