Charles Fruitman Final Project

Project: US Baby Names

Charles Fruitman: 100% contribution

For this project, I plan to conduct an analysis of new-born baby names in the United States from the years 1880 to 2014 and give indications about demographic and cultural trends based on my findings. The dataset that I will be using is here:

https://www.kaggle.com/kaggle/us-baby-names?select=StateNames.csv (https://www.kaggle.com/kaggle/us-baby-names?select=StateNames.csv)

I will ask questions that can be studied through this dataset, which should be useful to people with interests in fields such as history, demographics, linguistics, etc. These questions include:

1. Throughout the 20th century, did names of certain ethnic origins increase in number as these peoples immigrated to the US, and then decline as these groups assimilated into the mainstream culture? To investigate this, I will use aggregated groups of names of different origins, such as Nordic, French, and Italian.
2. Are there noticeable changes in naming patterns that can be attributed to historical and cultural events? Does the name of the President affect the number of babies given that name? How about the names of celebrities? Did the number of babies named Adolf decrease during and after World War II?
3. Are there names whose variations in use over time are due to gender? Are there names that have gone from being generally a boys' name to a girls' name or vice versa?
4. Finally, is it possible for names that are similar phonetically to displace one another? For example, as the name Justin has increased in popularity, has there been a simultaneous stagnation or decline in the number of babies named Jason?

My goal will be to visualize the data I use to show whether my hypotheses are correct. While it is unlikely that I will be able to investigate all these ideas in thorough detail, I believe that it will be possible to find some evidence to support the above ideas, and that the investigation will uncover more interesting trends along the way.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.4
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
setwd("C:/Users/charl/Math 2820L")
NationalNames <- read.csv("NationalNames.csv")
```

These three groups of names, of Nordic, Italian, and French origin respectively, will be used for the first part of the investigation. I continued to add names to each group until it consisted of at least 2000 total occurrences from the NationalNames data set.

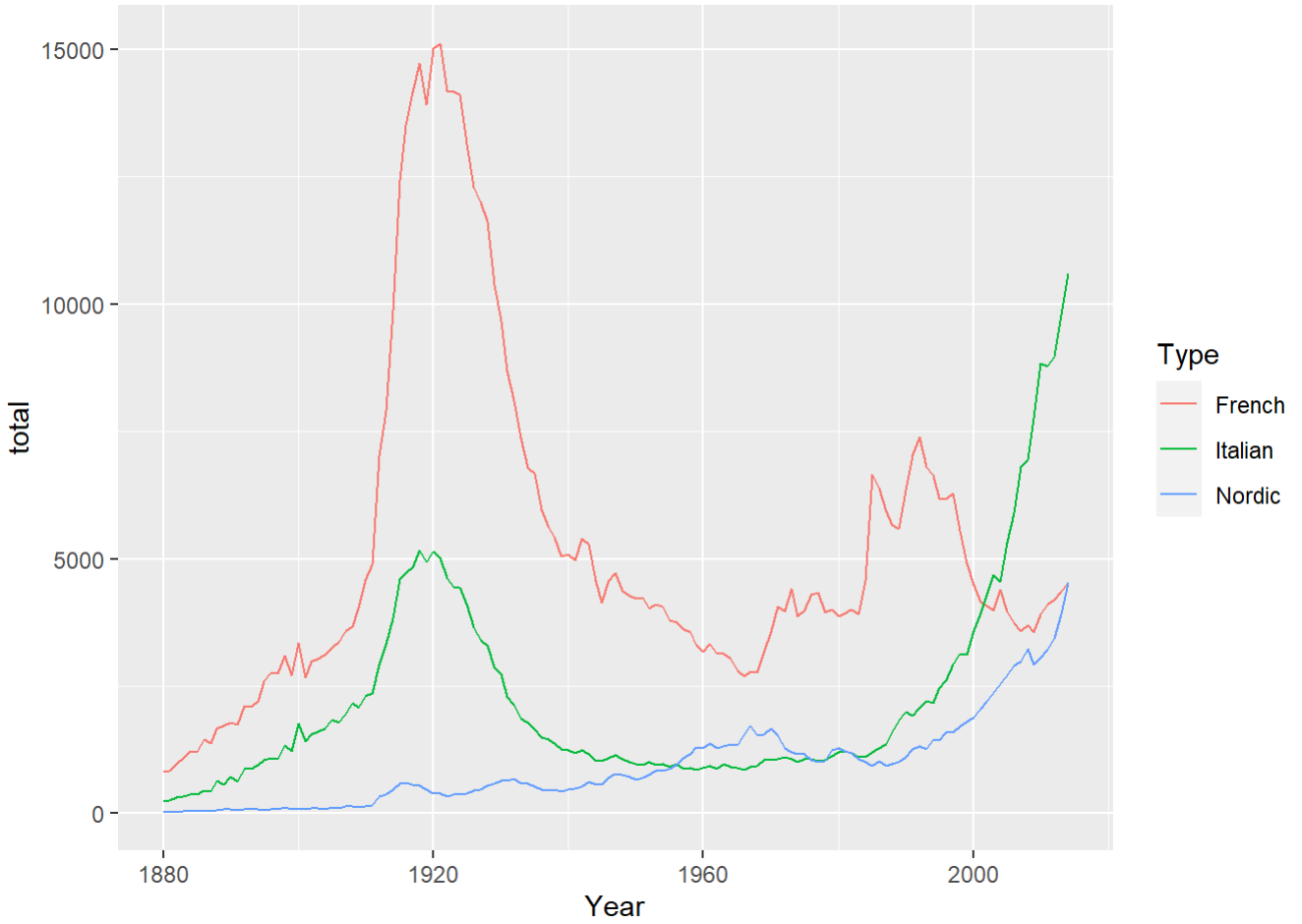
```
NordicNames <- NationalNames %>% filter(Name == "Greta" | Name == "Olaf" | Name == "Sven" | Name
== "Gunnar" | Name == "Freya" | Name == "Freja" | Name == "Freyja" | Name == "Nils" | Name == "A
rne" | Name == "Astrid" | Name == "Odin" | Name == "Jens" | Name == "Ingrid" | Name == "Lars" |
 Name == "Harald" | Name == "Thor" | Name == "Linnea" | Name == "Inga" | Name == "Bjorn" | Name
 == "Leif" | Name == "Soren" | Name == "Magnus")
NordicNames <- NordicNames %>% group_by(Year) %>% summarize(total = sum(Count))

ItalianNames <- NationalNames %>% filter (Name == "Giuseppe" | Name == "Luigi" | Name == "Romeo"
| Name == "Aldo" | Name == "Flavia" | Name == "Giovanni" | Name == "Viola" | Name == "Giancarlo"
| Name == "Allegra" | Name == "Vincenza" | Name == "Vincenzo" | Name == "Lorenza" | Name == "Nic
ola" | Name == "Antonella" | Name == "Filomena" | Name == "Arabella" | Name == "Valentina")
ItalianNames <- ItalianNames %>% group_by(Year) %>% summarize(total = sum(Count))

FrenchNames <- NationalNames %>% filter (Name == "Marine" | Name == "Marion" | Name == "Guillaum
e" | Name == "Mathieu" | Name == "Jacques" | Name == "Francois" | Name == "Frederique" | Name ==
"Olivier" | Name == "Dominique" | Name == "Bertrand" | Name == "Genevieve" | Name == "Blanche" |
Name == "Aimee" | Name == "Celeste" | Name == "Gaston" | Name == "Eloise" | Name == "Laurent" |
 Name == "Girard")
FrenchNames <- FrenchNames %>% group_by(Year) %>% summarize(total = sum(Count))

Names <- data.frame(Year = c(NordicNames$Year, ItalianNames$Year, FrenchNames$Year), total = c(N
ordicNames$total, ItalianNames$total, FrenchNames$total), Type = c(rep("Nordic", 135), rep("Ital
ian", 135), rep("French", 135)))

ggplot(Names, aes(x = Year, y = total, color = Type)) + geom_line()
```

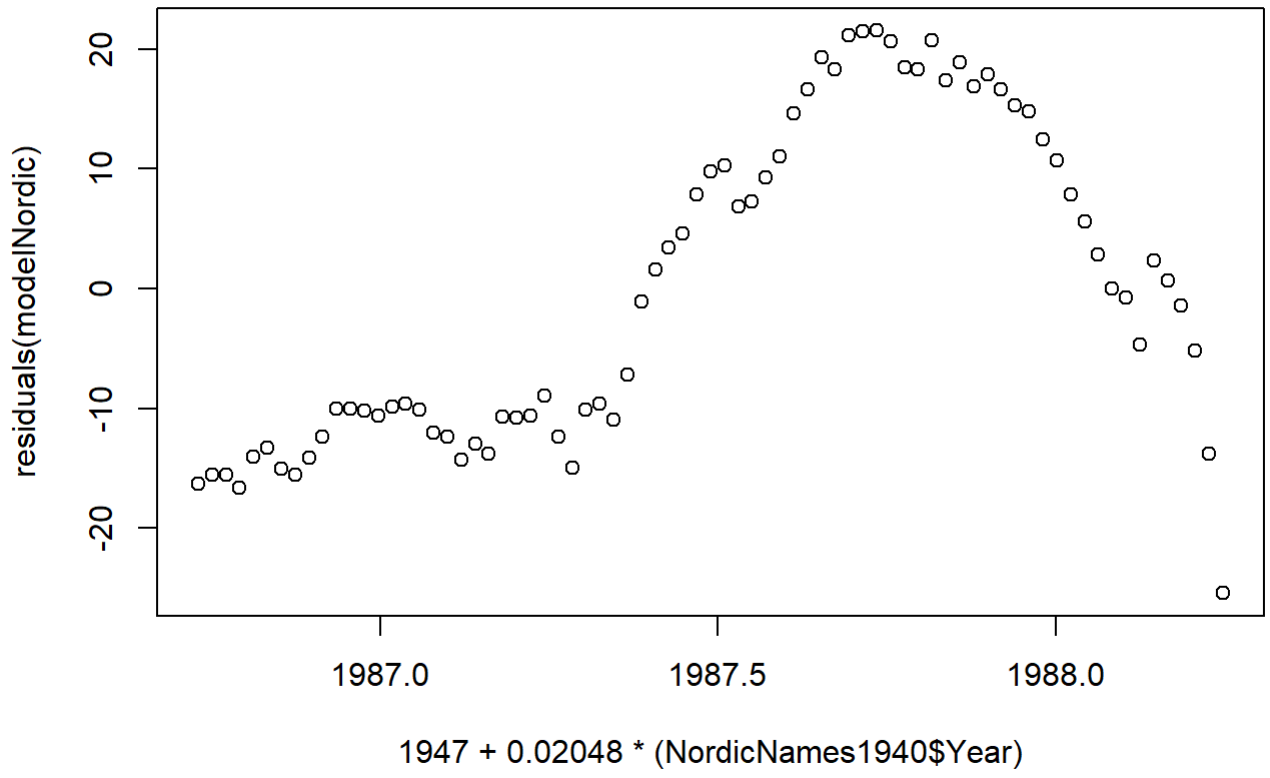


This graph shows that the first idea proposed in the introduction is at least partially true. Names of both French and Italian origin show a peak around the year 1920, at a time when immigration from France, Italy and French Canada had also reached a peak. They then show a steep decline afterwards as these groups of immigrants assimilated. However, what is interesting is that, starting around the 1990s, there seems to be a strong resurgence in the number of Italian names, to the point where by the 2010s they have become more popular than at any other time in history. Nordic names do not seem to fit the expected trend, so I next decided to investigate if there could be a linear model that explains the trend.

```
NordicNames1940 <- NordicNames %>% filter(Year > 1939)
modelNordic <- lm(Year ~ total, data = NordicNames1940)
summary(modelNordic)
```

```
##
## Call:
## lm(formula = Year ~ total, data = NordicNames1940)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.427 -11.522  -1.067  11.758  21.523
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.947e+03  3.115e+00  624.99   <2e-16 ***
## total       2.048e-02  1.827e-03   11.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.3 on 73 degrees of freedom
## Multiple R-squared:  0.6324, Adjusted R-squared:  0.6274
## F-statistic: 125.6 on 1 and 73 DF,  p-value: < 2.2e-16
```

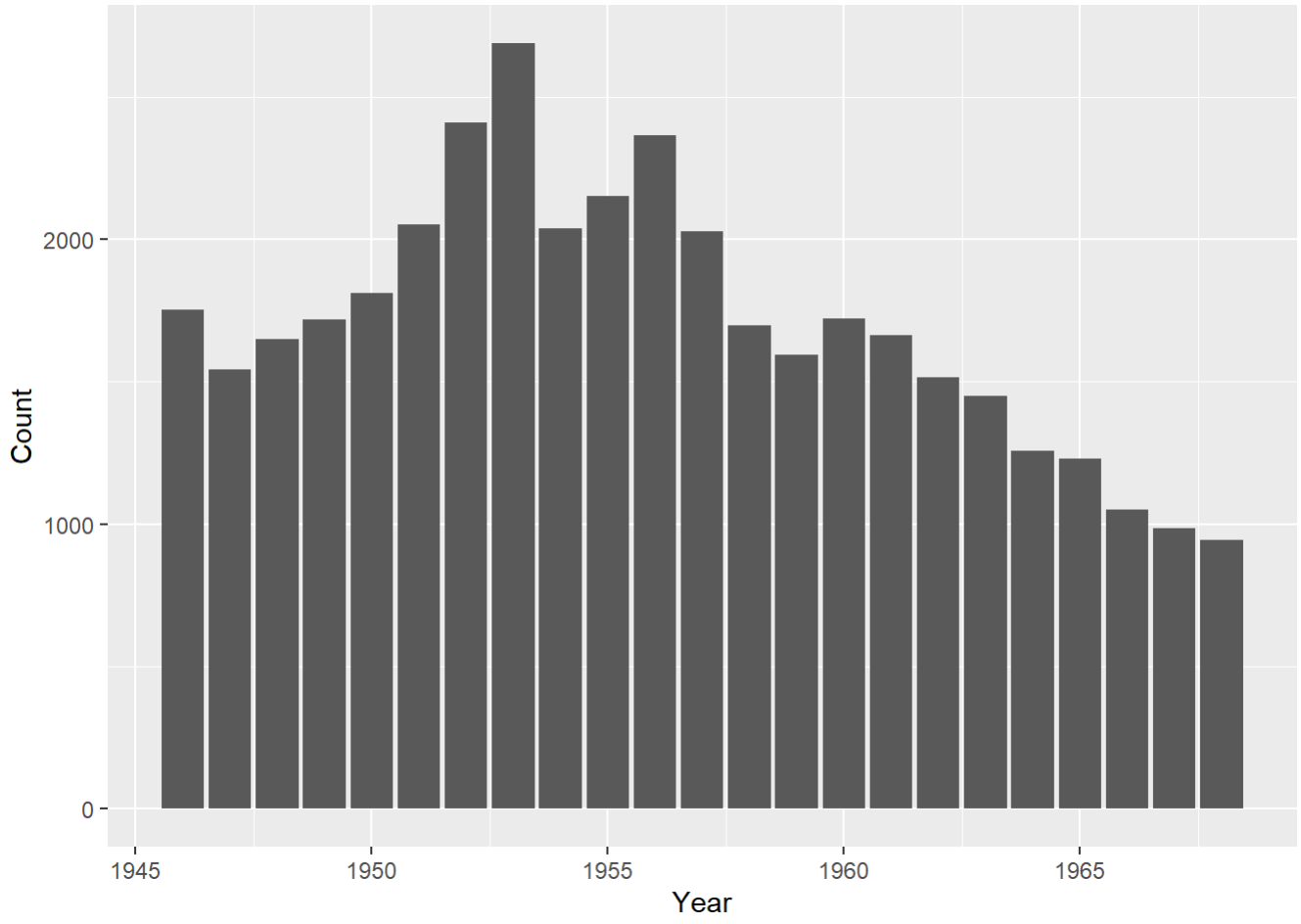
```
plot(1.947e+03 + 2.048e-02 * (NordicNames1940$Year), residuals(modelNordic))
```



As the summary and plot of fitted values against residuals shows, there is a clear pattern that the residuals are exclusively negative early on and then highly positive later before a steep decline again, forming an arc shape. It is therefore inappropriate to use this linear model to explain the trend in Nordic names.

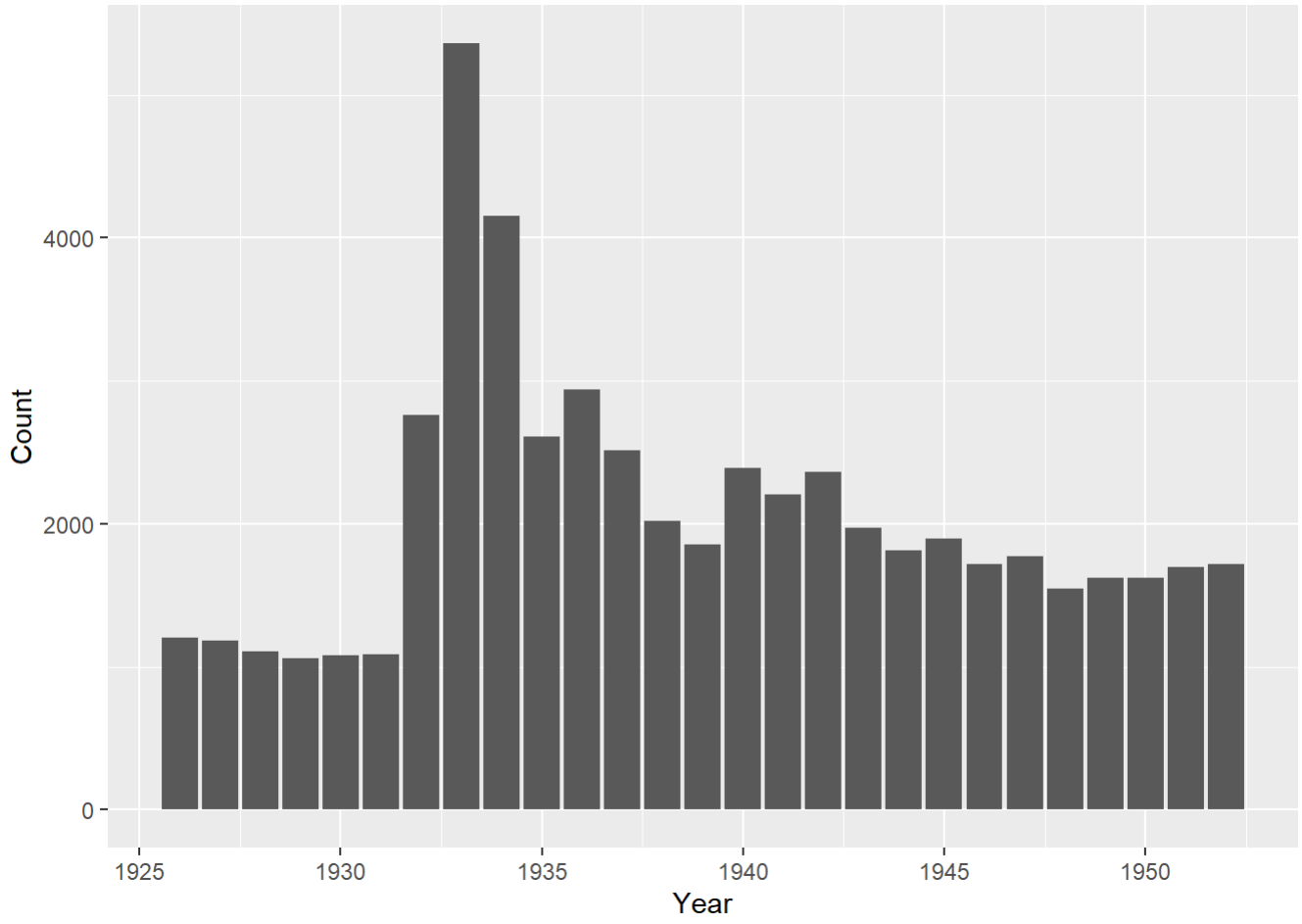
```
PresidentNames <- NationalNames %>% filter(Name == "Grover" | Name == "William" | Name == "Theod
ore" | Name == "Woodrow" | Name == "Warren" | Name == "Calvin" | Name == "Herbert" | Name == "Fr
anklin" | Name == "Harry" | Name == "Dwight" | Name == "John" | Name == "Lyndon" | Name == "Rich
ard" | Name == "Gerald" | Name == "Jimmy" | Name == "Ronald" | Name == "George" | Name == "Barac
k")
```

```
Dwight <- PresidentNames %>% filter(Name == "Dwight" & Gender == "M" & Year < 1969 & Year > 1945
)
ggplot(Dwight, aes(x = Year, y = Count)) + geom_col()
```



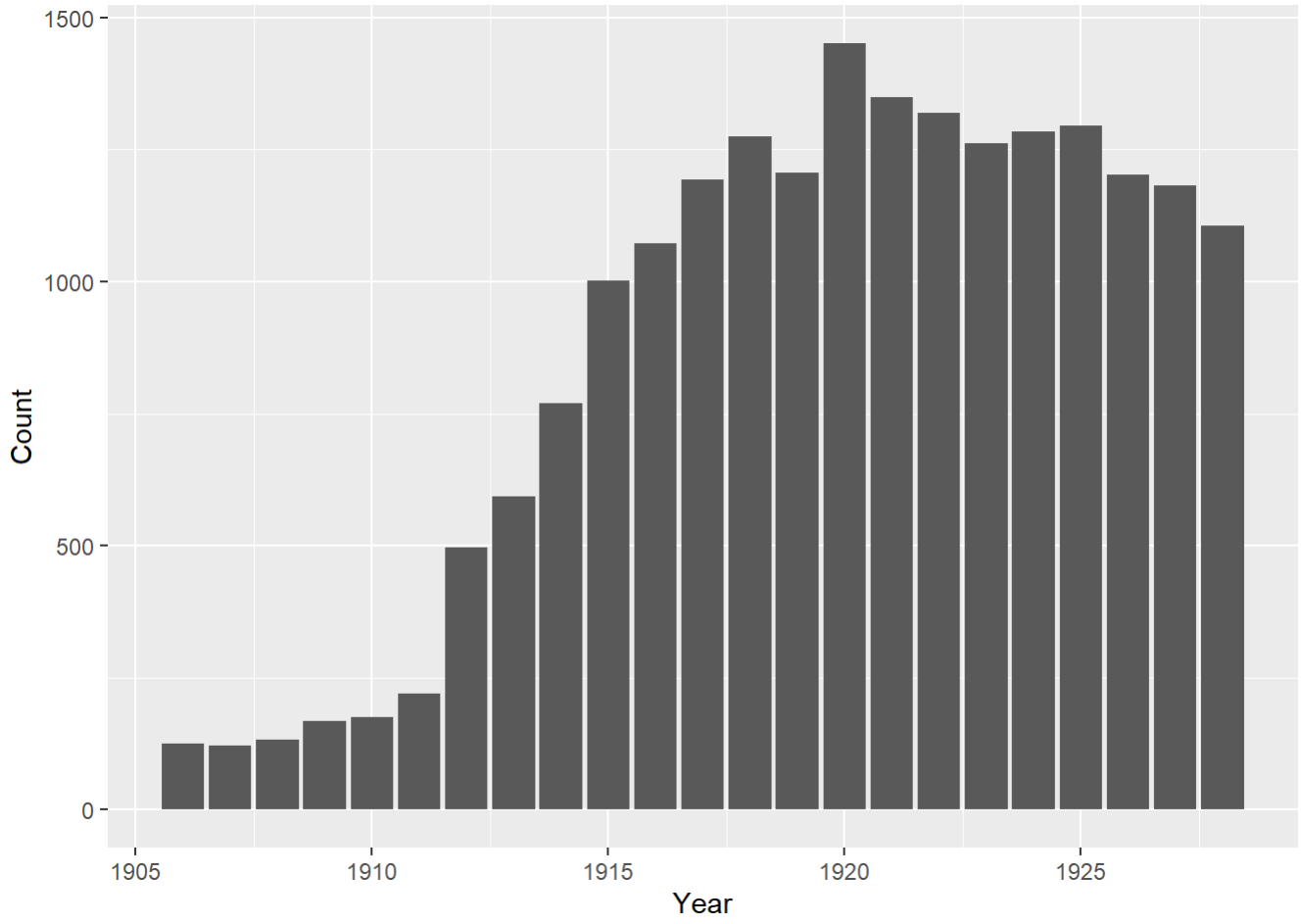
This graph shows that there was a slight increase in the number of boys named Dwight for the first few years after Eisenhower's presidency began in 1953, but that it declined soon after.

```
Franklin <- PresidentNames %>% filter(Name == "Franklin" & Gender == "M" & Year > 1925 & Year <
1953)
ggplot(Franklin, aes(x = Year, y = Count)) + geom_col()
```



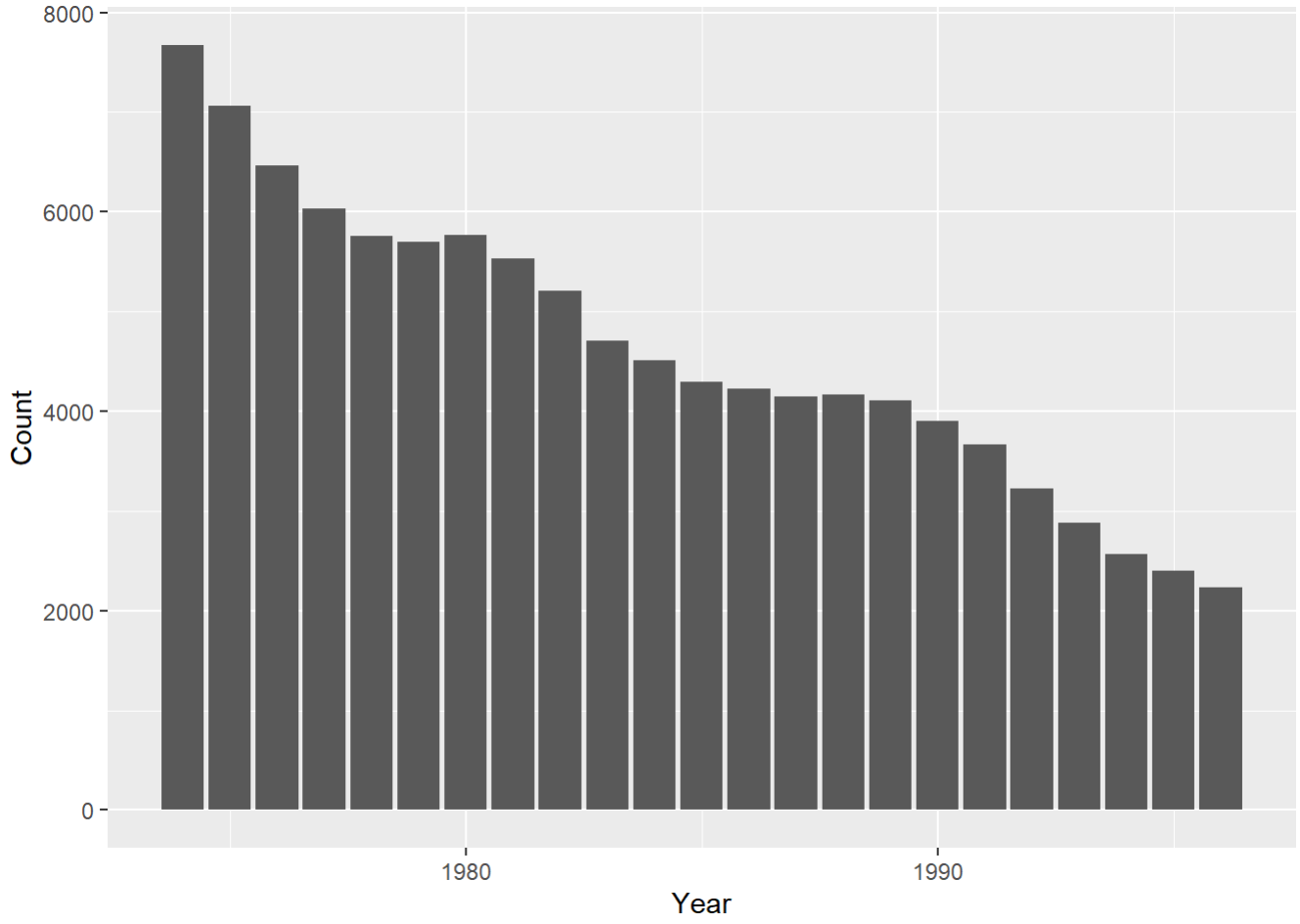
This graph shows a much sharper increase at the beginning of Roosevelt's presidency. This did come down soon after, but for the rest of his presidency, the name Franklin remained more prevalent than it had been beforehand.

```
Woodrow <- PresidentNames %>% filter(Name == "Franklin" & Gender == "M" & Year > 1905 & Year < 1
929)
ggplot(Woodrow, aes(x = Year, y = Count)) + geom_col()
```



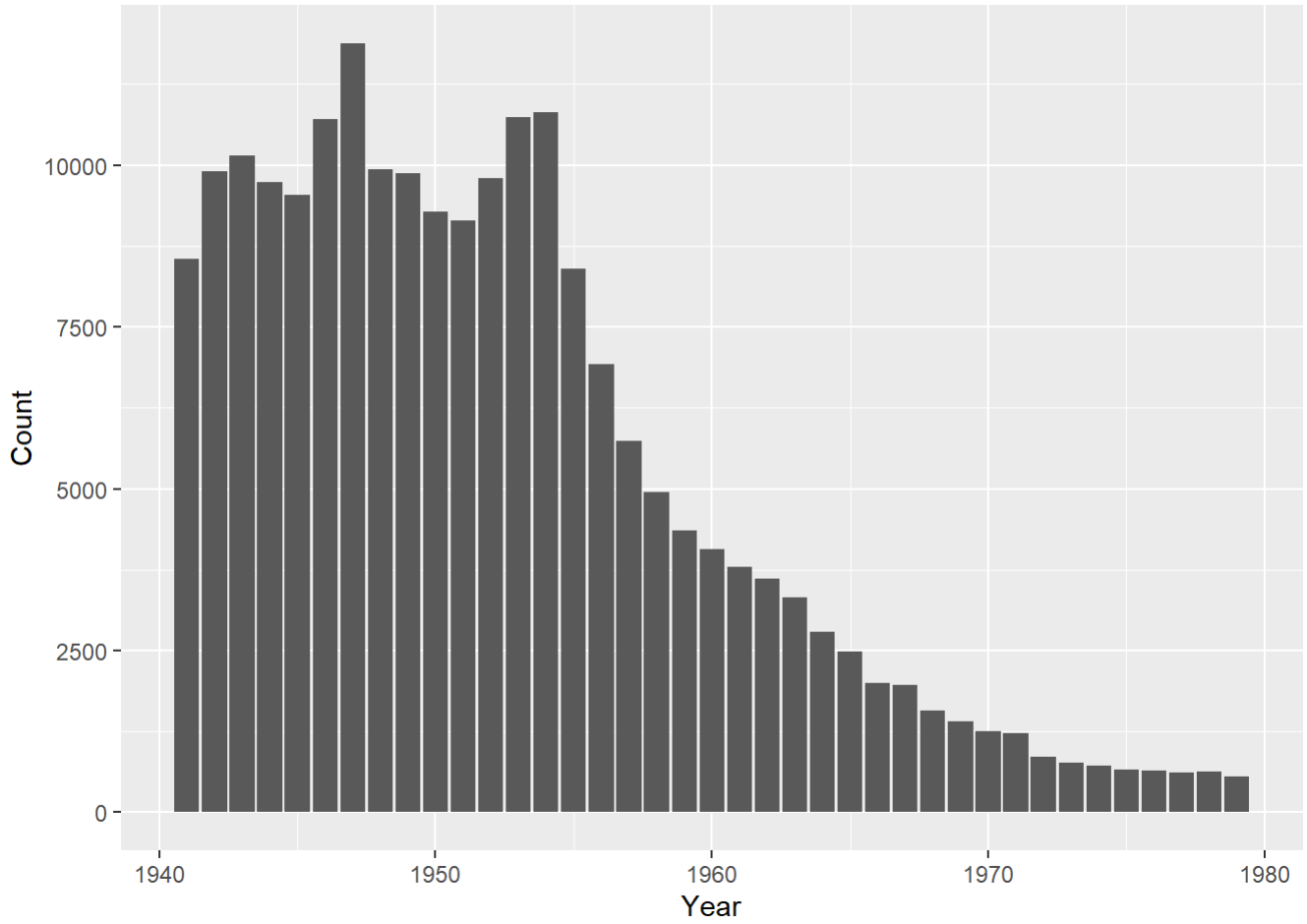
Woodrow Wilson’s presidency shows a steady increase in the number of boys given his name, peaking around the end of his presidency in the 1920s.

```
Ronald <- PresidentNames %>% filter(Name == "Ronald" & Gender == "M" & Year > 1973 & Year < 1997
)
ggplot(Ronald, aes(x = Year, y = Count)) + geom_col()
```



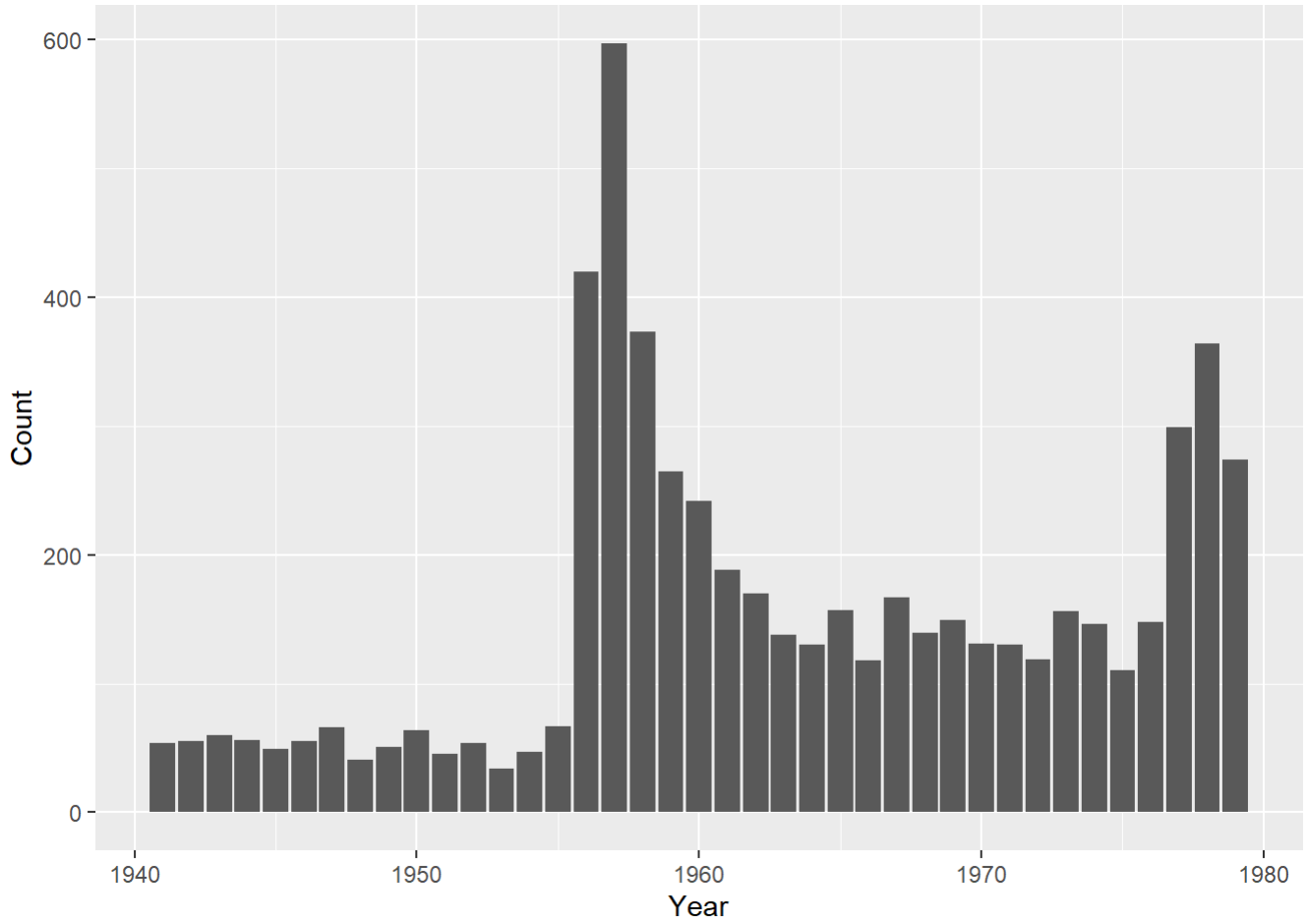
Ronald Reagan's presidency shows no increase in the number of boys given his name, in fact it steadily decreases throughout.

```
Marilyn <- NationalNames %>% filter(Name == "Marilyn" & Gender == "F" & Year > 1940 & Year < 198
0)
ggplot(Marilyn, aes(x = Year, y = Count)) + geom_col()
```



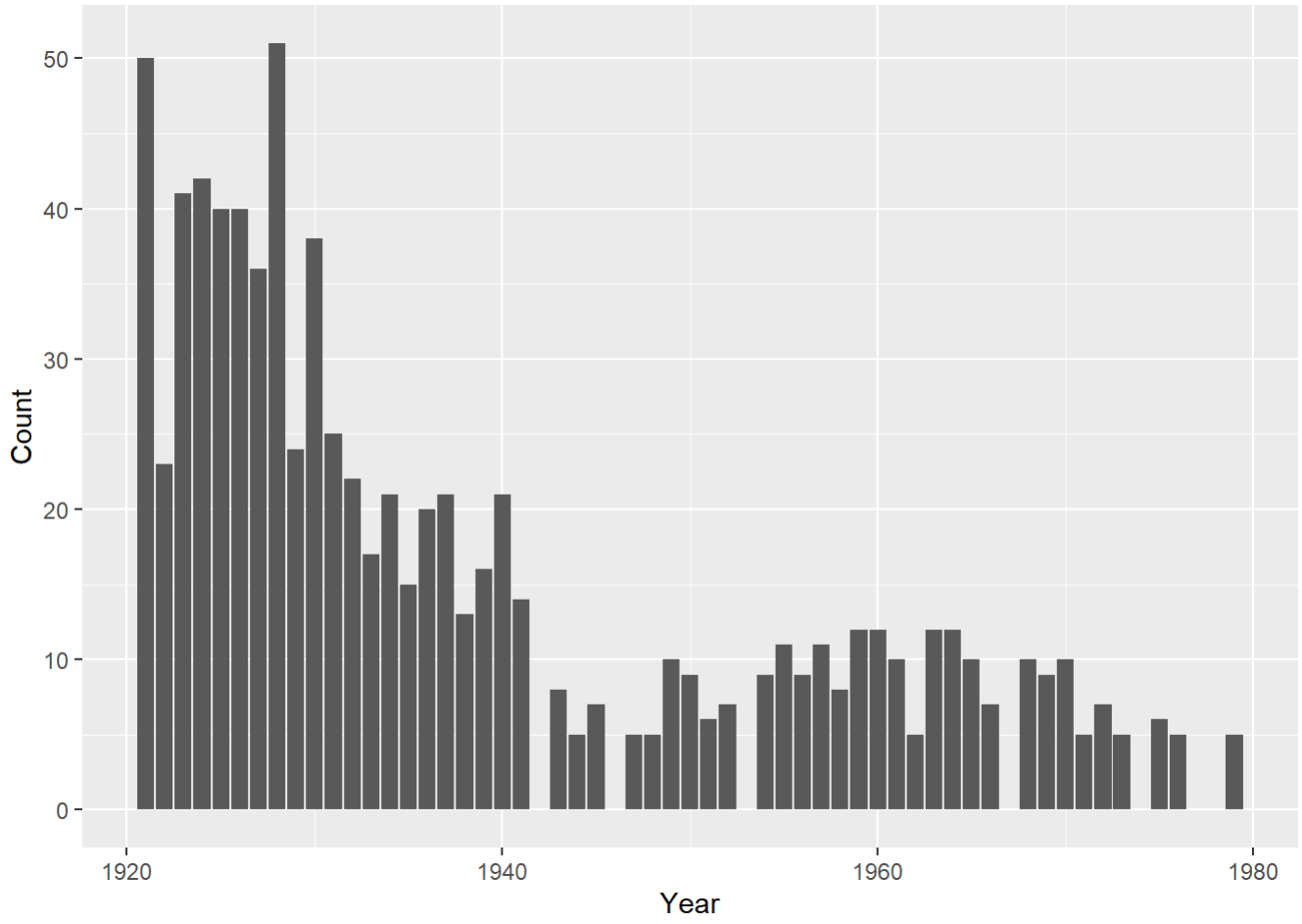
There was a sharp decrease in the number of girls named Marilyn soon after Marilyn Monroe's career began in the 1950s. By 1960, the number of girls being named Marilyn was less than half what it had been just 7 or 8 years earlier.

```
Elvis <- NationalNames %>% filter(Name == "Elvis" & Gender == "M" & Year > 1940 & Year < 1980)
ggplot(Elvis, aes(x = Year, y = Count)) + geom_col()
```



The name Elvis, on the other hand, skyrocketed when Elvis Presley's career began, also in the 1950s, increasing more than 10 fold in a period of two years. It then died down but made somewhat of a comeback in the late 70s.

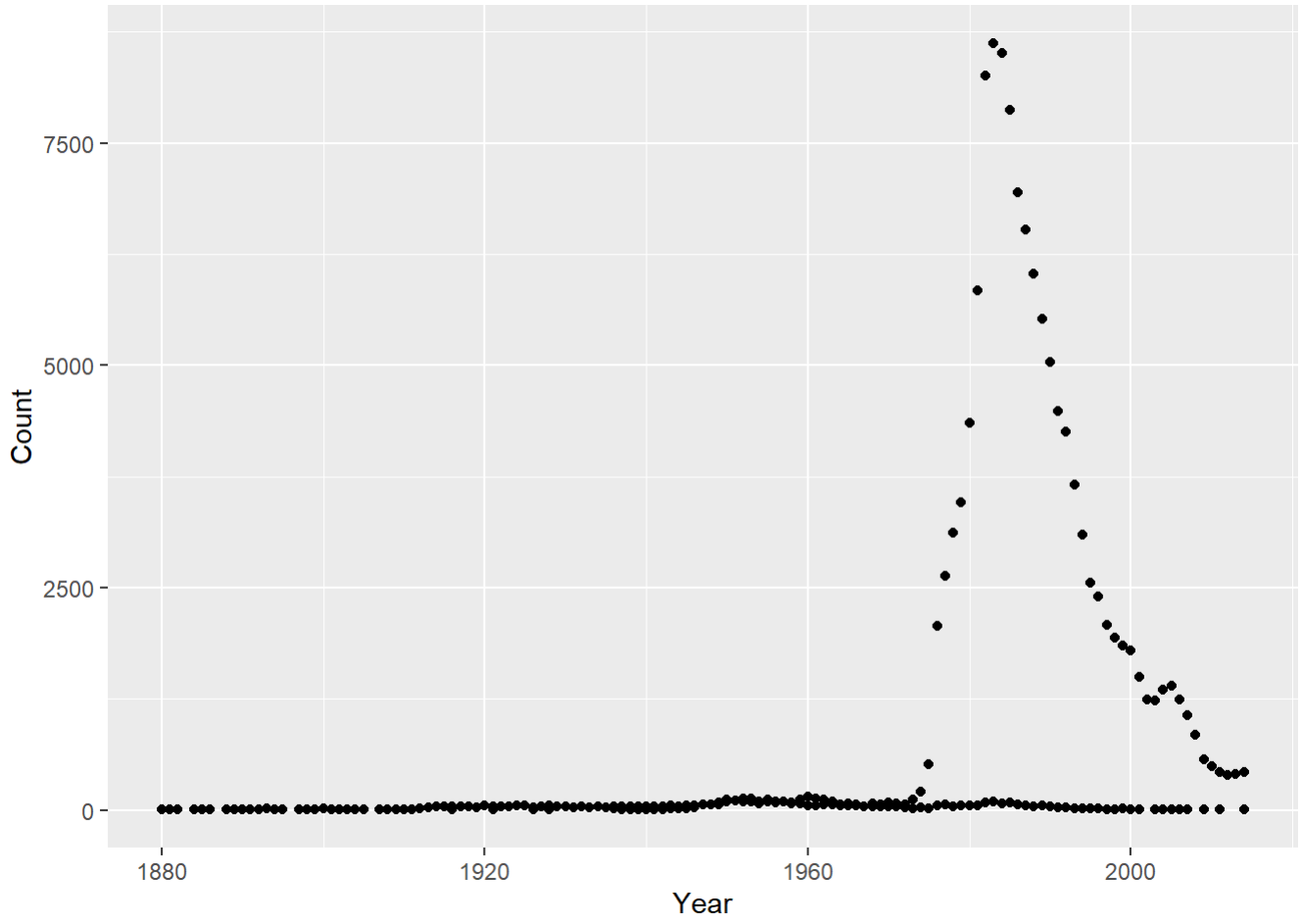
```
Adolf <- NationalNames %>% filter(Name == "Adolf" & Gender == "M" & Year > 1920 & Year < 1980)
ggplot(Adolf, aes(x = Year, y = Count)) + geom_col()
```



For obvious reasons, the name Adolf began to decrease in the 1930s and then even more so after 1940.

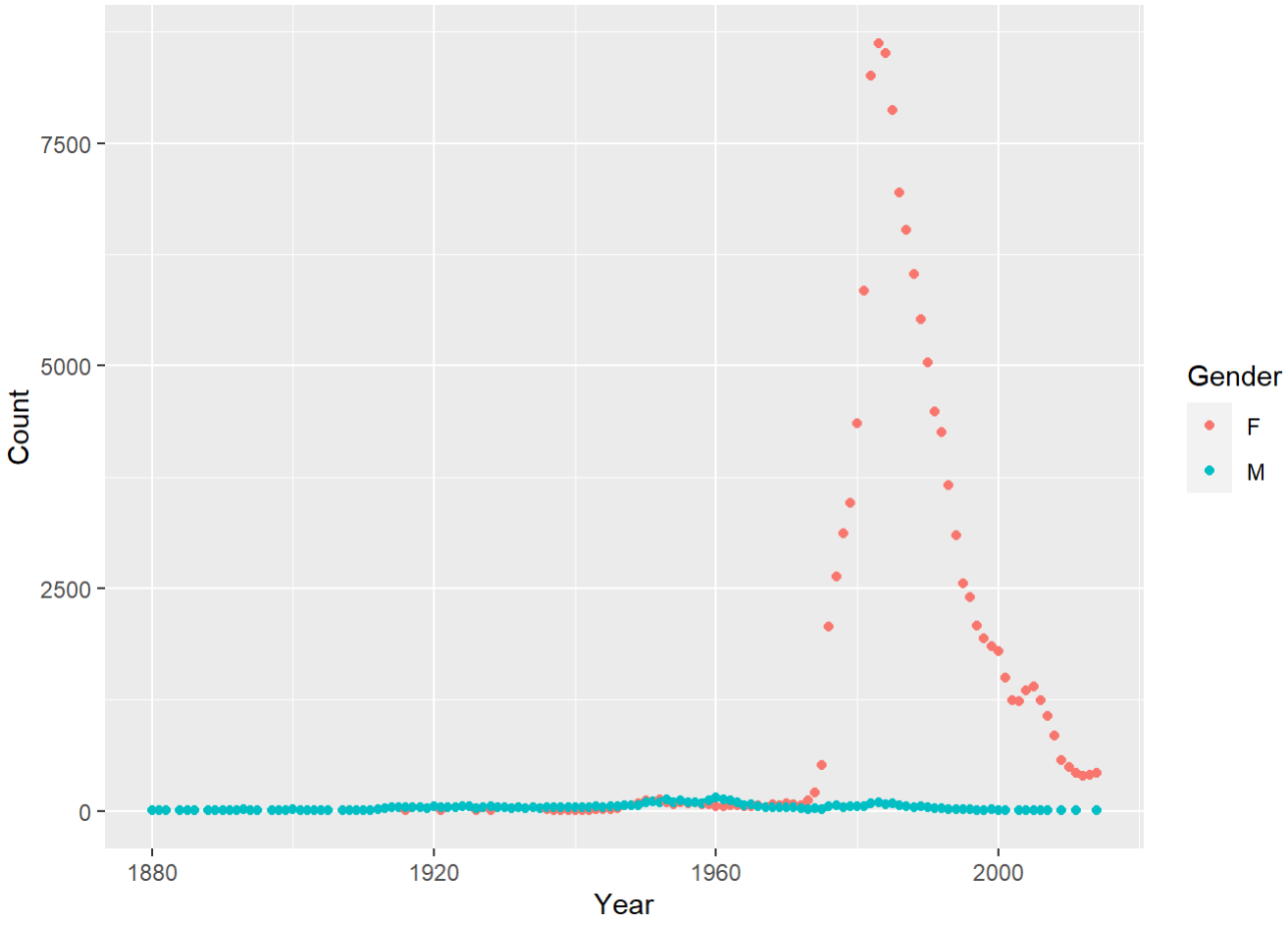
The following graphs will focus on briefly addressing the third and fourth ideas proposed in the introduction.

```
Lindsay <- NationalNames %>% filter(Name == "Lindsay")
ggplot(Lindsay, aes(x = Year, y = Count)) + geom_point()
```



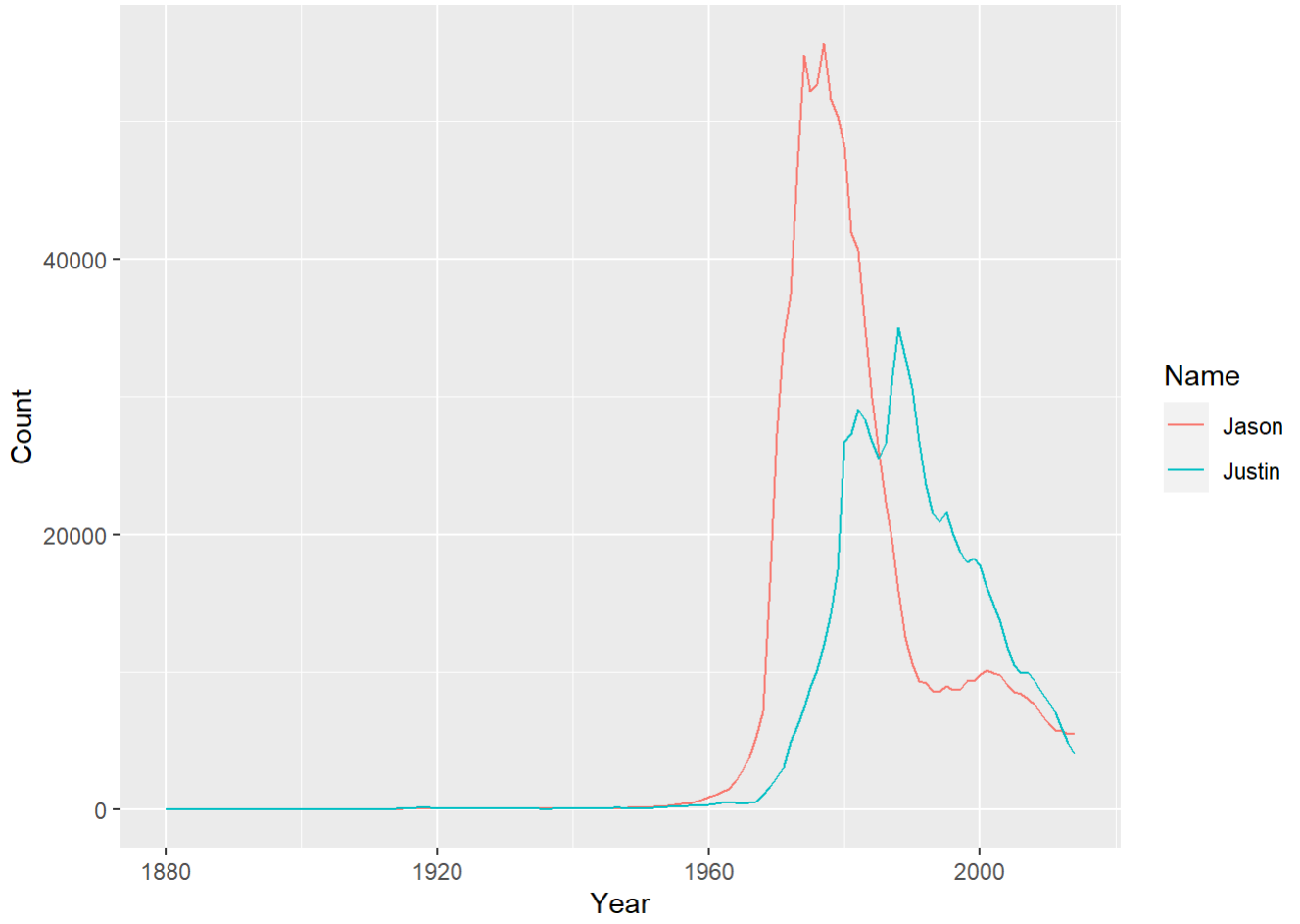
This graph shows wide range of results for the name Lindsay, especially from 1970 to 2000.

```
ggplot(Lindsay, aes(x = Year, y = Count, color = Gender)) + geom_point()
```



This graph shows that differences in naming trends based on gender provide a clear explanation for this high level of variation.

```
JustinJason <- NationalNames %>% filter(Name == "Justin" | Name == "Jason") %>% filter(Gender == "M")
ggplot(JustinJason, aes(x = Year, y = Count, color = Name)) + geom_line()
```



This graph shows that there is perhaps some merit to the idea that the name Jason was displaced by the name Justin. As the name Jason decreased in the 1980s, Justin continued to increase. However, by the 1990s both were decreasing.

In conclusion, this analysis has demonstrated the validity of several of the ideas proposed in the introduction. However, that is not to say that any hypotheses have been proven. While there was an increase in French and Italian names up until the 1920s, followed by a decrease, it may not be simply a matter of immigration, but also whether these types of names are fashionable at certain points in time, which could explain the rebound almost 100 years later. Names of key figures in certain generations can have effects on naming trends, but as the bar graphs show, this can often be erratic and unpredictable.

This method has many shortcomings. The groups of names of different origin should be larger and more varied to improve the accuracy and precision of the method. It is also difficult to make inferences from the results without other sources of knowledge. Further research would focus on the geographic distribution of names using the StateNames data set, as originally mentioned in my proposal. This proved to be difficult for me to complete in a timely manner due to the nature of the csv file, being too large for some functions to work properly and organized in a way that made it difficult to manipulate, as well as my own shortcomings and inexperience working with data.