script.R

aurenferguson

Tue Dec 27 15:07:04 2016

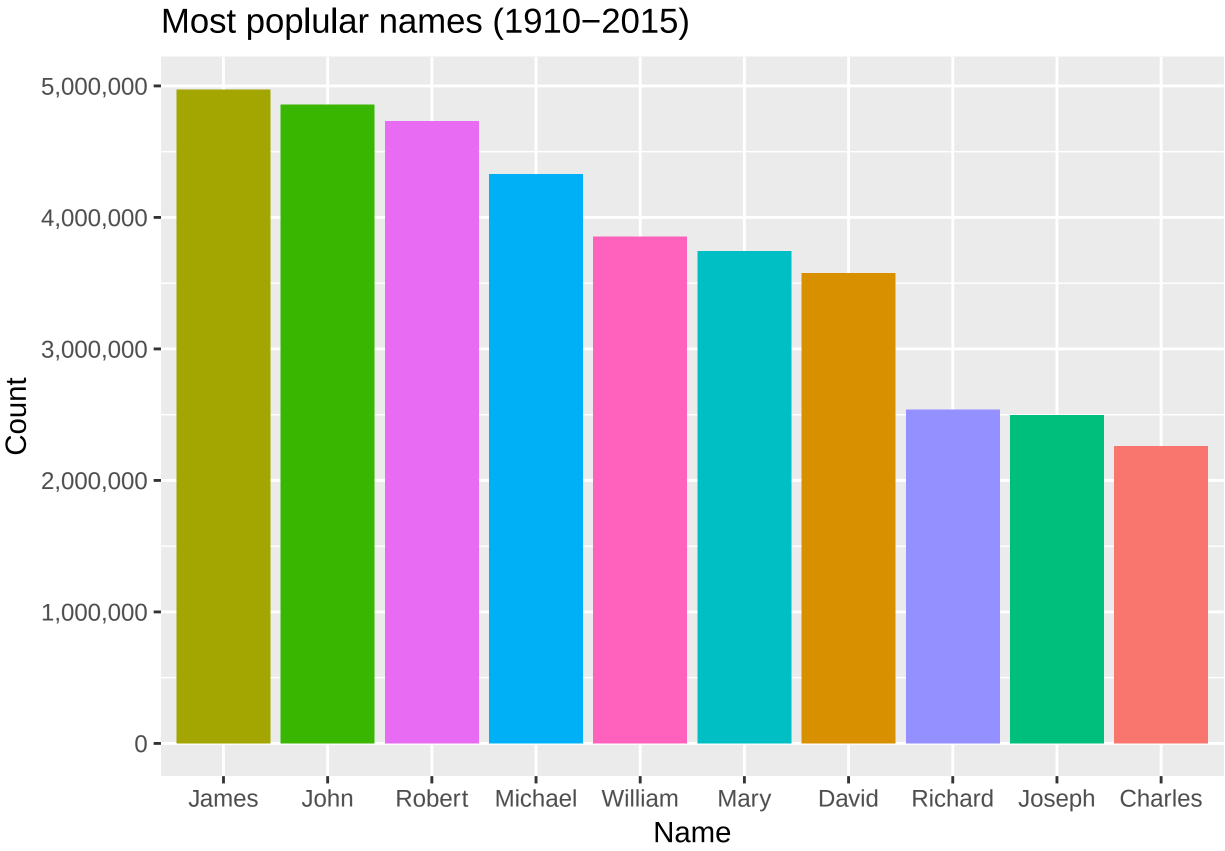
```r
# Exploring US baby names 1910-2015
# Author: Auren Ferguson
# Date : December 2016


# Preamble ---------------------------------------------------------------
library(data.table)
library(dplyr)
```

```
## -------------------------------------------------------------------------
## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!
## -------------------------------------------------------------------------
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:data.table':
##
##     between, first, last
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

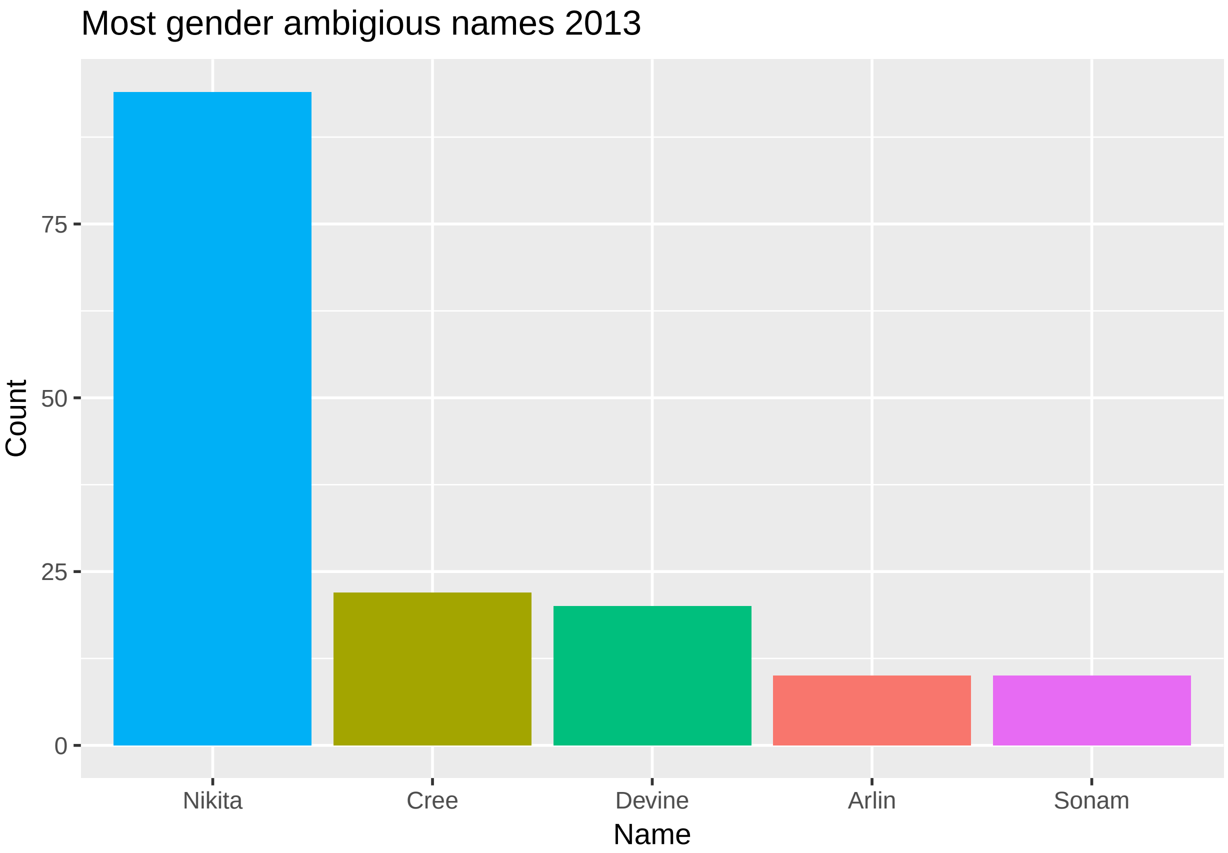
```r
library(ggplot2)
library(scales)

# Reading data -----------------------------------------------------------
setwd("~/Documents/Exploring_baby_names")

# filelist = list.files(path = "/Users/aurenferguson/Downloads/namesbystate", pattern = ".TXT")
#
# datalist = lapply(filelist, function(x)read.csv(x, header=F))
#
# #assuming the same header/columns for all files
# baby_names = bind_rows(datalist)
#
# # converting to tbl
# baby_names <- as.tbl(baby_names)
#
# # renaming columns
# baby_names <- dplyr::rename(.data = baby_names, state = V1, sex = V2, year = V3, name = V4, amount =
#
# fwrite(baby_names, file = "baby_names_all_states.csv")
# Importing combined file
baby_names <- as.tbl(fread(input = "baby_names_all_states.csv"))
```

```
##
Read 86.4% of 5743017 rows
Read 5743017 rows and 5 (of 5) columns from 0.105 GB file in 00:00:03
```

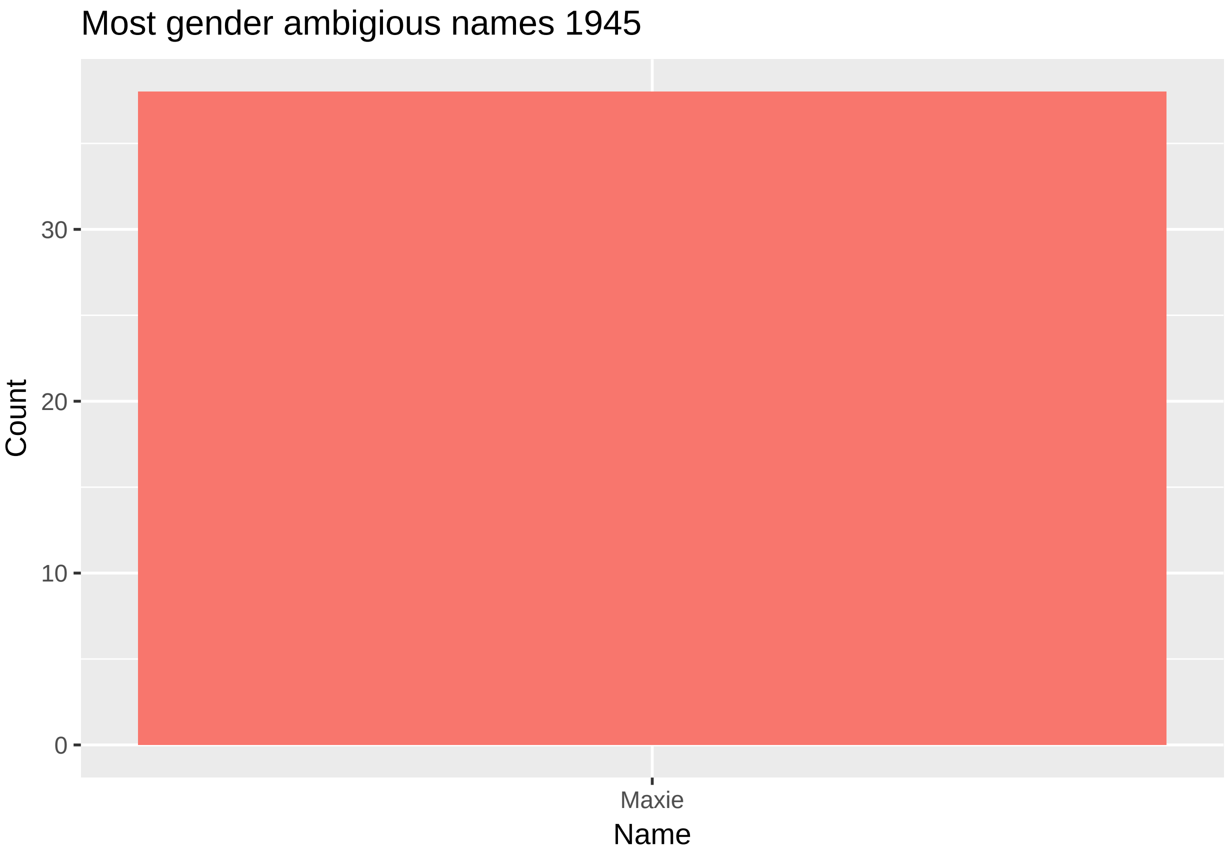
```
# Most popular baby name of all time --------------------------------------
popular_names_all_time <- function(df){
  df <- df %>% group_by(name) %>% summarise(total = sum(amount)) %>%
    arrange(desc(total)) %>% head(10)
  ggplot(data = df, aes(x = reorder(name, -total), y = total, fill = name)) +
    geom_bar(stat = "identity") +
    scale_y_continuous(labels = comma) +
    guides(fill = FALSE) +
    ylab("Count") +
    xlab("Name") +
    labs(title = "Most poplular names (1910-2015)")
}

popular_names_all_time(baby_names)
```



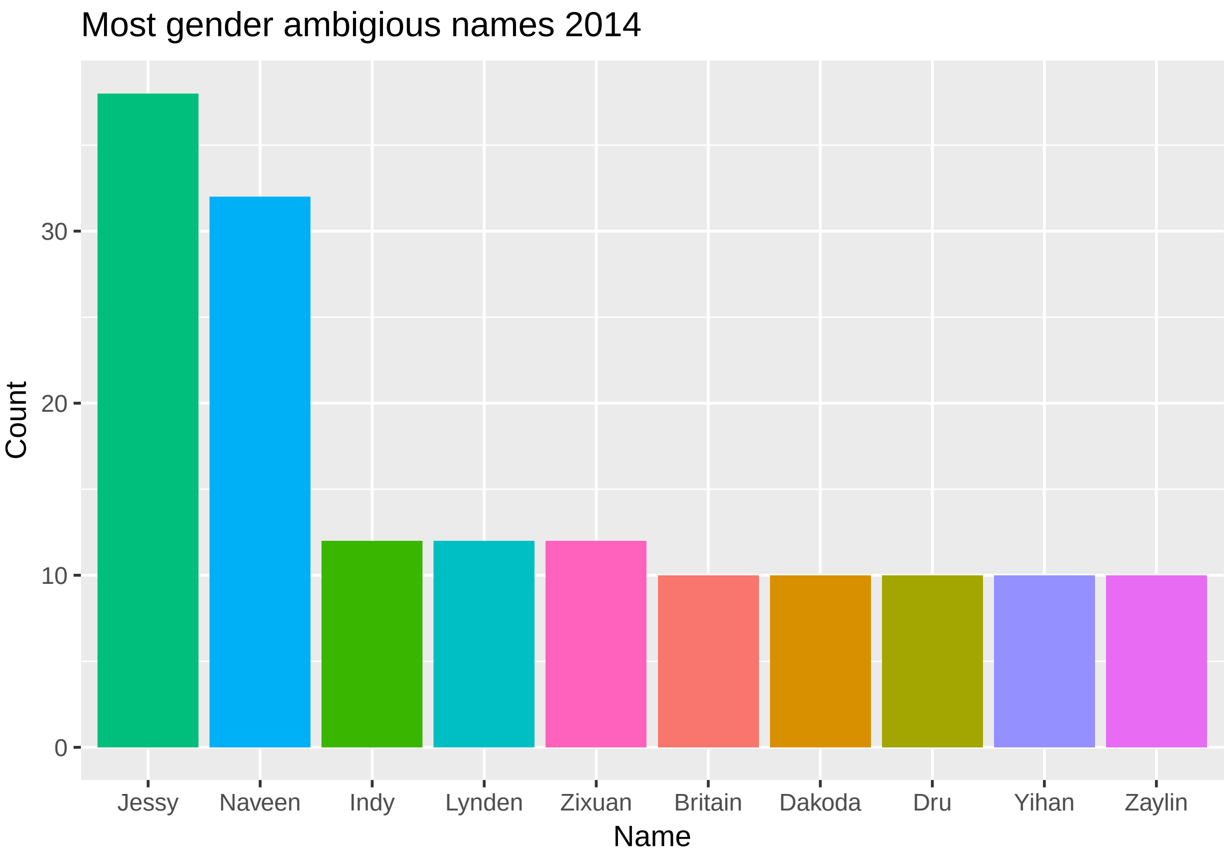
```
# What is the most gender ambiguous name in 2013? 1945? -------------------

Gender_ambig_name <- function(df, yr){

  # Filters for the year, aggregates to name, sex level and sums amount
  df <- df %>% filter(year == yr) %>% group_by(name, sex) %>%
    summarise(total = sum(amount)) %>% arrange(desc(total))

  # Keeps all duplicates, i.e. male and female of same name
  df <- df[duplicated(df$name) | duplicated(df$name, fromLast=TRUE), ]

  # Another dataframe that goes to name level,
  #therefore the difference between total is due to amount of male and female
  df_a <- df %>% group_by(name) %>% summarise(total_amt = sum(total)) %>%
    arrange(desc(total_amt)) %>% rename(total_male_female = total_amt)

  # Joins the 2 dataframes, allowing a ratio of male female to be calculated
  df <- left_join(df, df_a, by = "name")

  # Male/Female calculation ratio of total for each name
  df$ratio <- df$total / df$total_male_female

  # selects ratio of 0.5 and removes duplicates for clarity.
  # 0.5 corresponds to a name being exactly half male and half female
  df <- df %>% filter(ratio == 0.5) %>% distinct(name,.keep_all = T)

  # Visualising results
  ggplot(data = df, aes(x = reorder(name, -total_male_female), y = total_male_female, fill = name)) +
    geom_bar(stat = "identity") +
    scale_y_continuous(labels = comma) +
    guides(fill = FALSE) +
    ylab("Count") +
    xlab("Name") +
    ggtitle(label = paste("Most gender ambigious names", yr))

}

Gender_ambig_name(baby_names, yr = 2013)
```



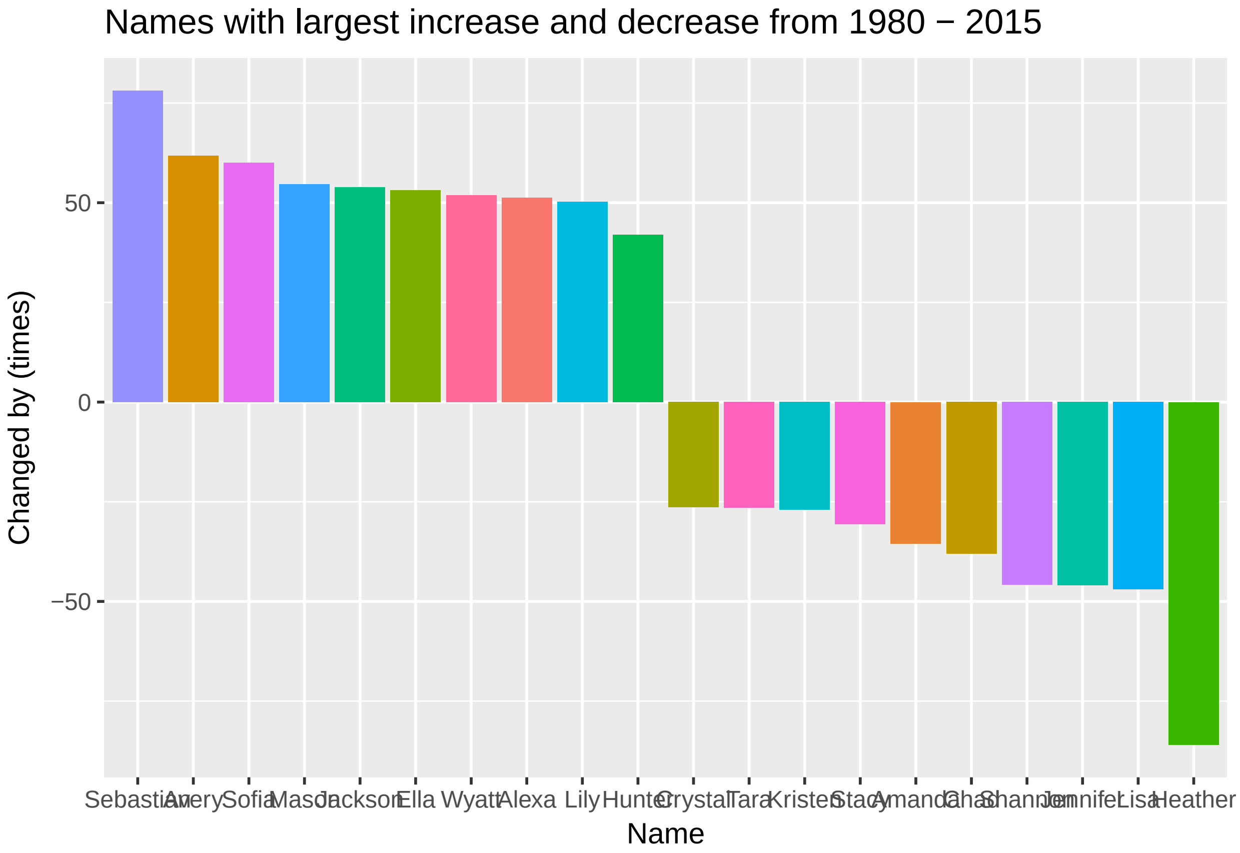
```
Gender_ambig_name(baby_names, yr = 1945)
```



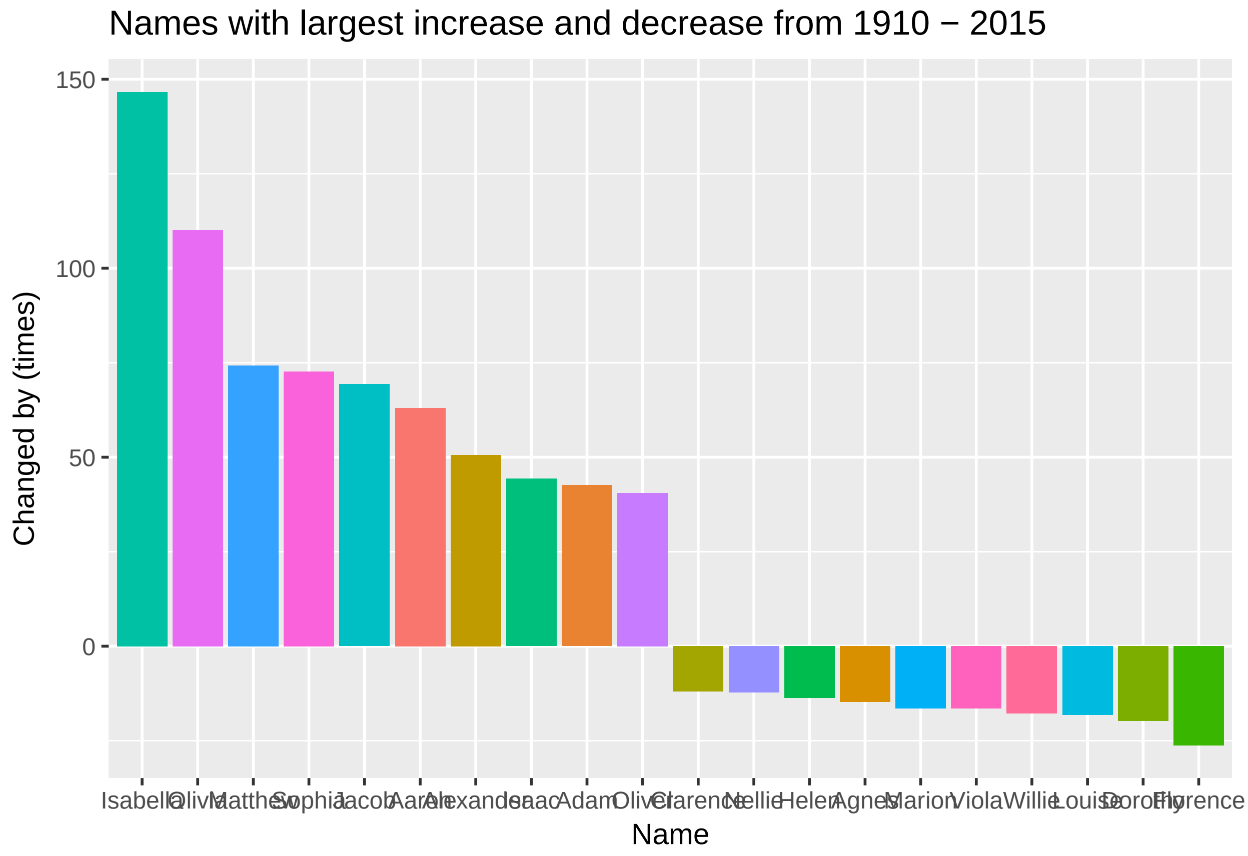
```
Gender_ambig_name(baby_names, yr = 2014)
```



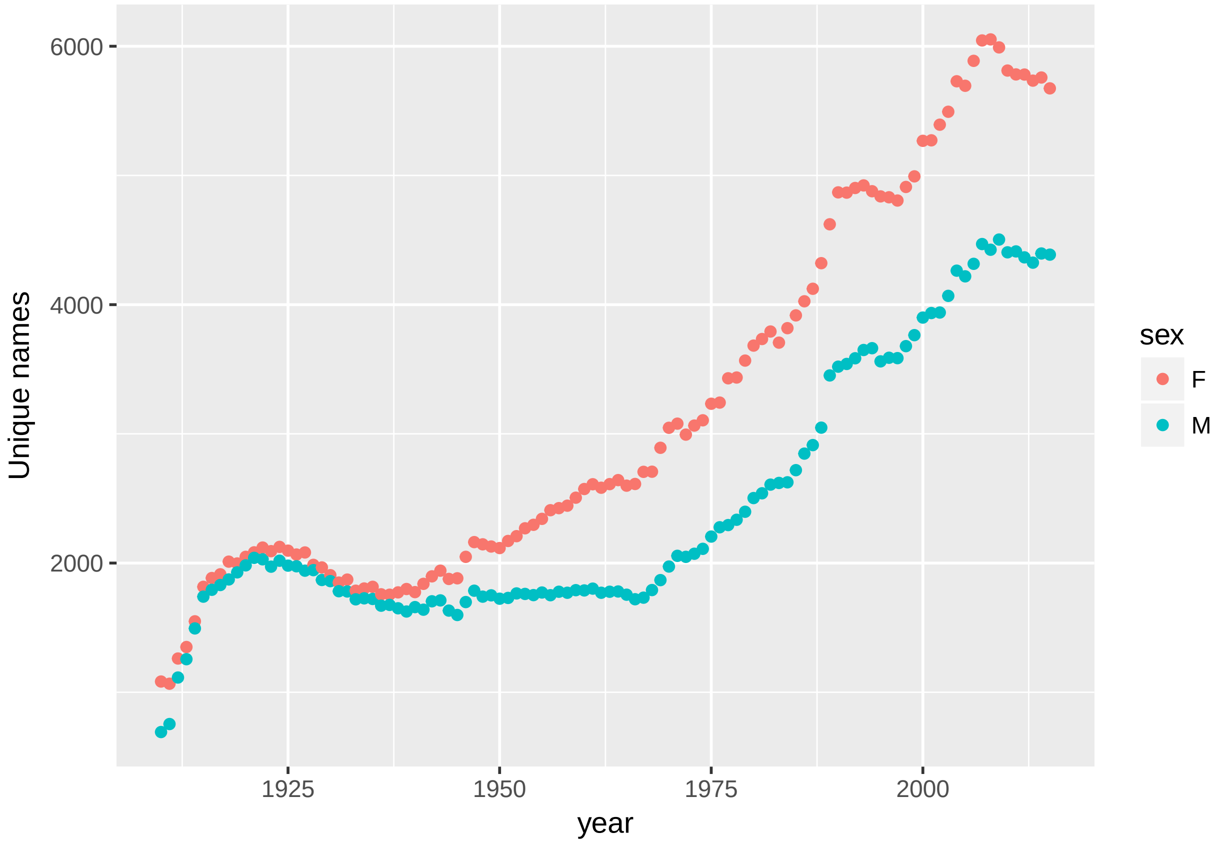
```r
# find the name that has had the largest percentage increase in popularity since 1980 --------

change_popularity_time <- function(df, yr_1 = 1980, yr_2 = 2015){

  if(yr_1 < yr_2){
    a <- df %>% filter(year == yr_1) %>% group_by(name) %>%
      summarise(total = sum(amount))

    a <- a %>% mutate(pct_of_total_a = (total / sum(total)) * 100) %>%
      # select(name, pct_of_total_a) %>%
      arrange(desc(pct_of_total_a))

    b <- df %>% filter(year == yr_2) %>% group_by(name) %>%
      summarise(total = sum(amount))

    b <- b %>% mutate(pct_of_total_b = (total / sum(total)) * 100) %>%
      #select(name, pct_of_total_b) %>%
      arrange(desc(pct_of_total_b))

    c <- inner_join(a, b, by = "name")

    c <- filter(c, total.x >= 100 & total.y >= 100)

    c <- c %>% mutate(pct_change = ifelse(total.x < total.y, (total.y - total.x) / total.x,
                                          (total.y - total.x) / total.y)) %>%
      arrange(desc(pct_change))
    #%>% head(10)
    d <- c[1:10,]

    e <- c[(nrow(c) - 9): nrow(c),]

    f <- bind_rows(d,e)

    ggplot(data = f, aes(x = reorder(name, -pct_change), y = pct_change, fill = name)) +
      geom_bar(stat = "identity") +
      scale_y_continuous(labels = comma) +
      guides(fill = FALSE) +
      ylab("Changed by (times)") +
      xlab("Name") +
      ggtitle(label = paste("Names with largest increase and decrease from", yr_1, "-", yr_2))
  }else{
    print("yr_1 should be before yr_2, please swap their values")
  }
}

change_popularity_time(baby_names)
```



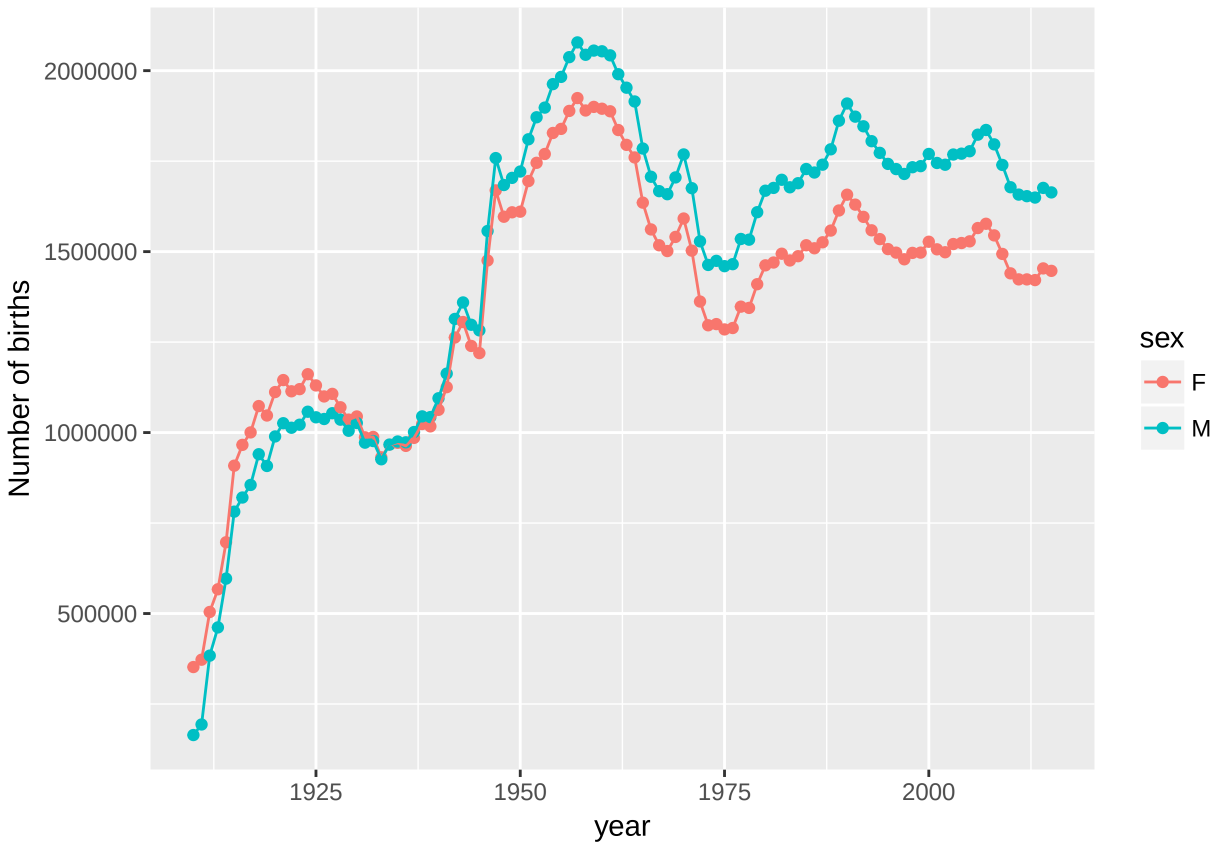
```
change_popularity_time(baby_names, yr_1 = 1910, yr_2 = 2015)
```



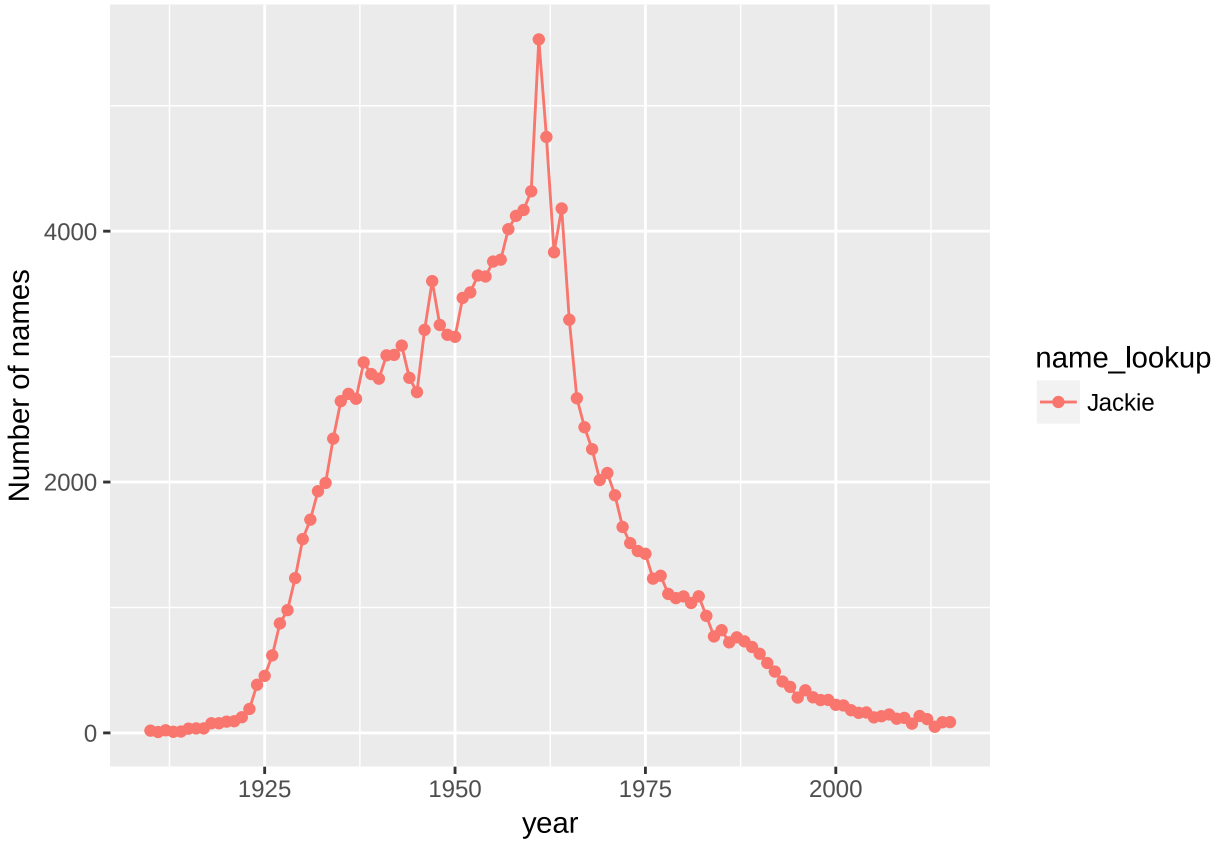
```
# name diversity ----------------------------------------------------------
name_diversity <- function(df){
  df <- df %>% group_by(year, sex) %>% summarise(total_unique = length(unique(name)))
  ggplot(data = df, aes(x = year, y = total_unique, color = sex)) +
    geom_point() +
    labs(y = "Unique names")
}
name_diversity(df = baby_names)
```



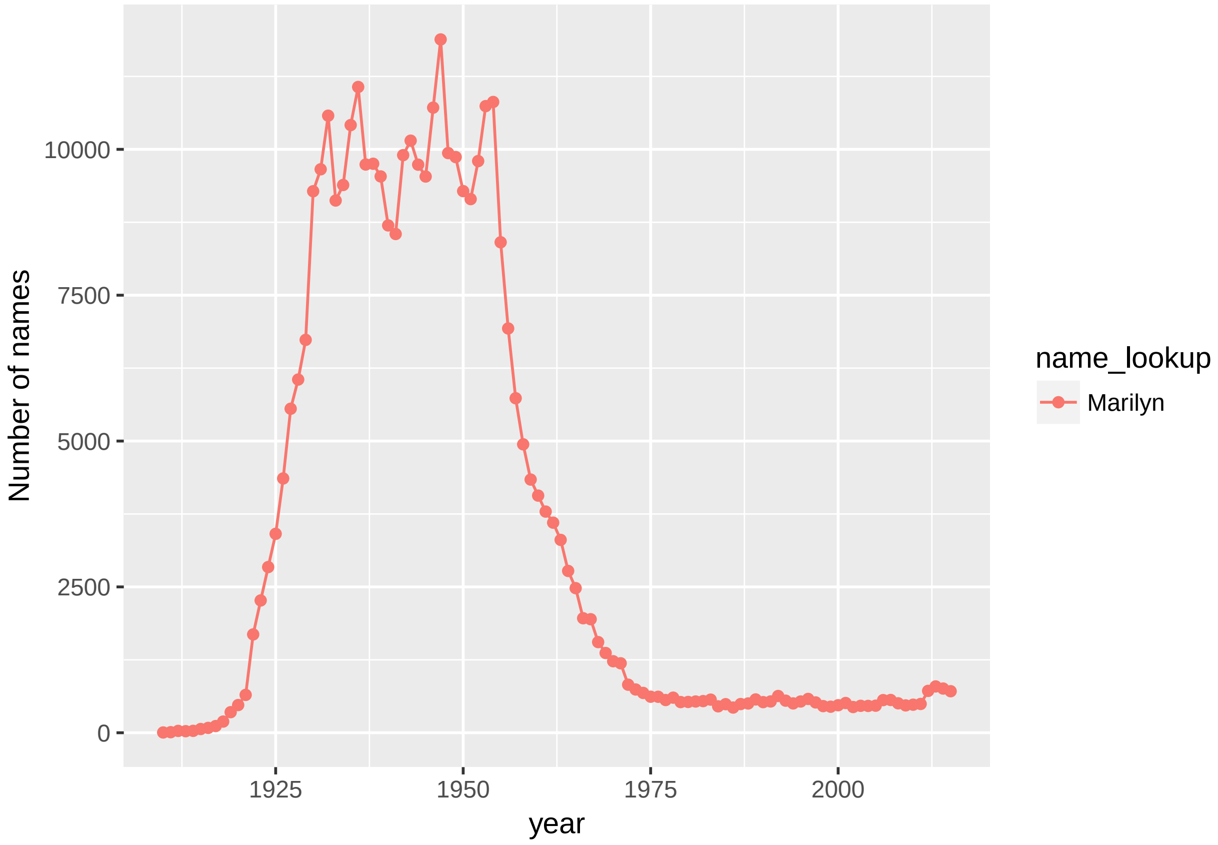
```
# How many births per year ------------------------------------------------
birth_numbers_per_year <- function(df){
  df  <- df %>% group_by(year, sex) %>% summarise(total_births = sum(amount))
  ggplot(df, aes(x = year, y = total_births, color = sex)) + geom_point() + geom_line() + labs(y = "Numb
}

birth_numbers_per_year(baby_names)
```



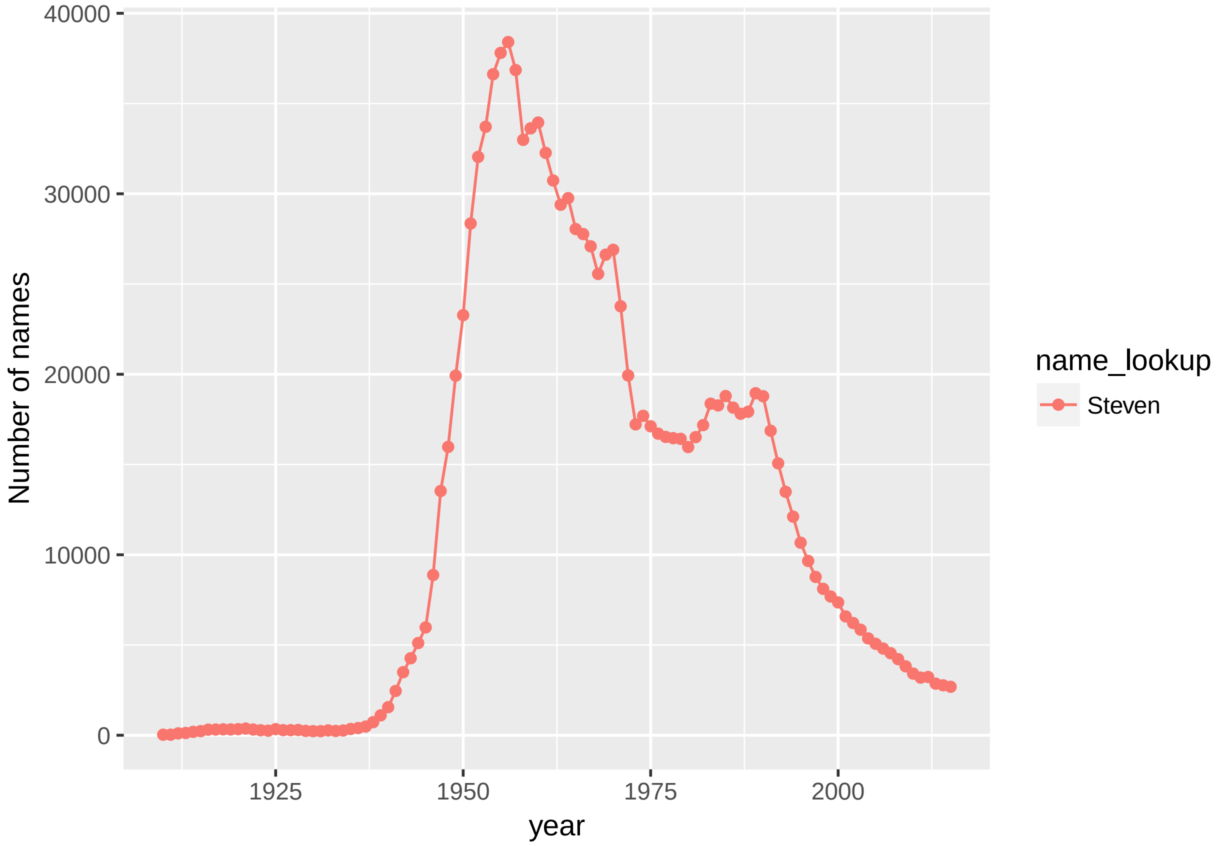
```
# Do people name their babies after famout people, using examples of Jackie (Kennedy) & Marilyn (Monroe
famous_name_check <- function(df, name_lookup = "Jackie"){
  df <- df %>% group_by(name, year) %>% summarise(total = sum(amount)) %>% filter(name == name_lookup)
  ggplot(data = df, aes(x = year, y = total, color = name_lookup)) +
    geom_point() + geom_line() + labs(y = "Number of names")
}

famous_name_check(baby_names, name_lookup = "Jackie")
```



```
famous_name_check(baby_names, name_lookup = "Marilyn")
```



```
famous_name_check(baby_names, name_lookup = "Steven")
```


40000
30000
20000
10000
0
Number of names
1925
1950
1975
2000
year
name_lookup
Steven