Lab_11

Fantastic Four

First three letters of names

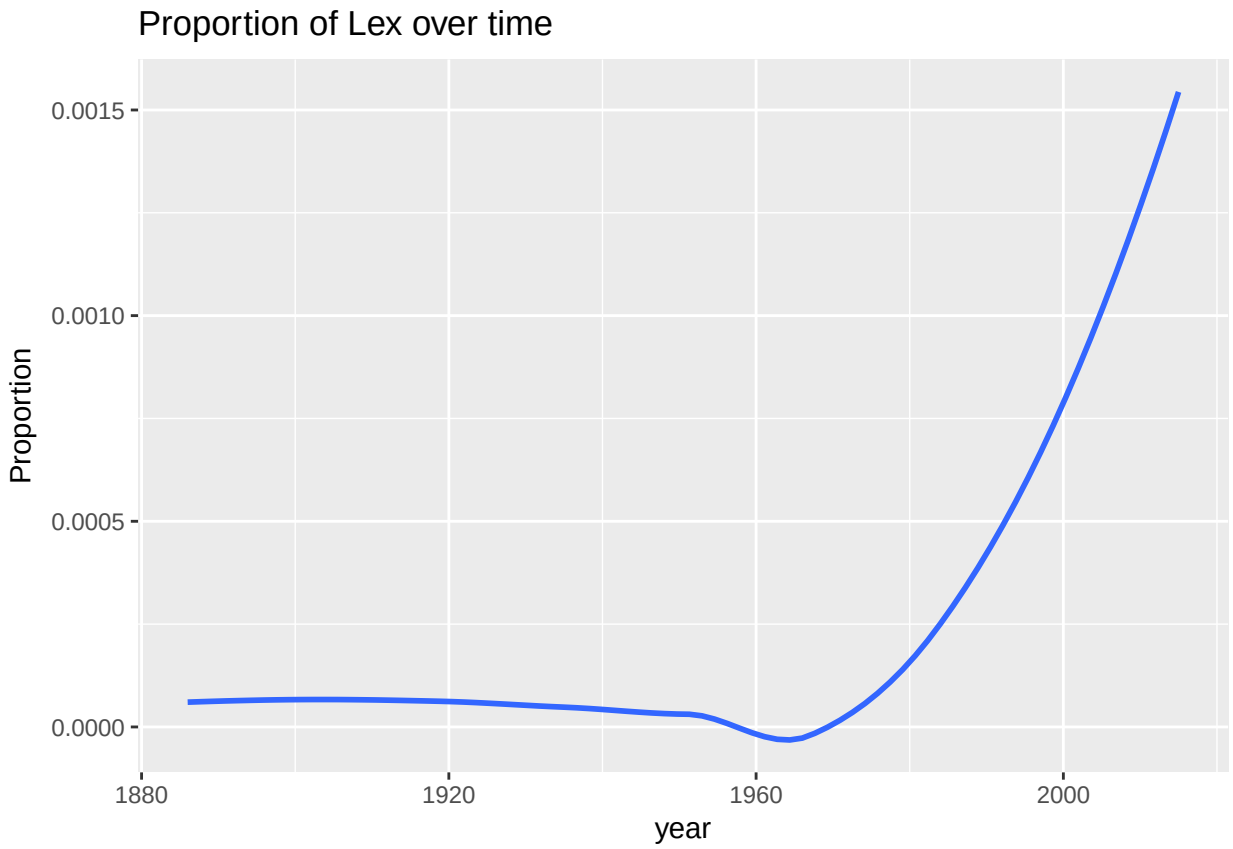
```
##          nn
## 1 0.5833333
```

```
##          nn
## 1 0.4336553
```

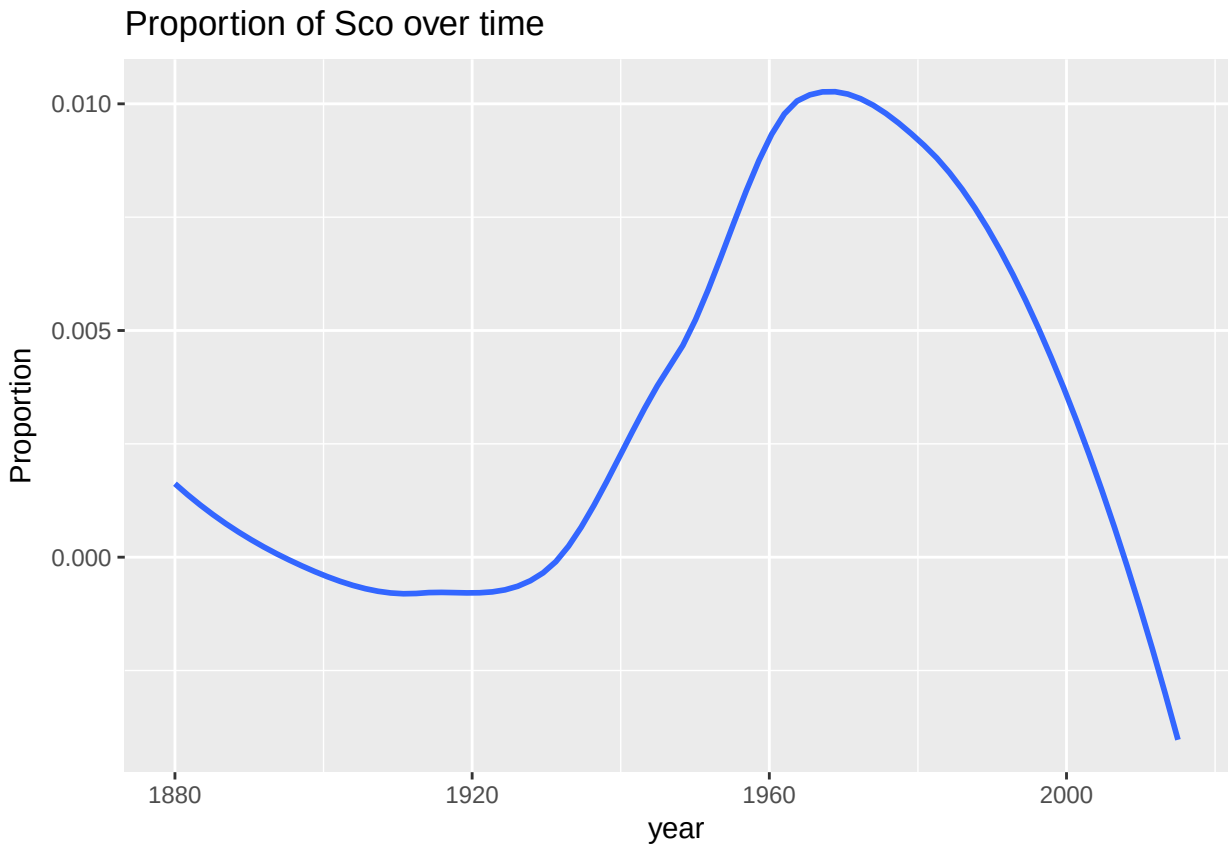
```
##          nn
## 1 0.1835393
```

```
##          nn
## 1 0.1513542
```

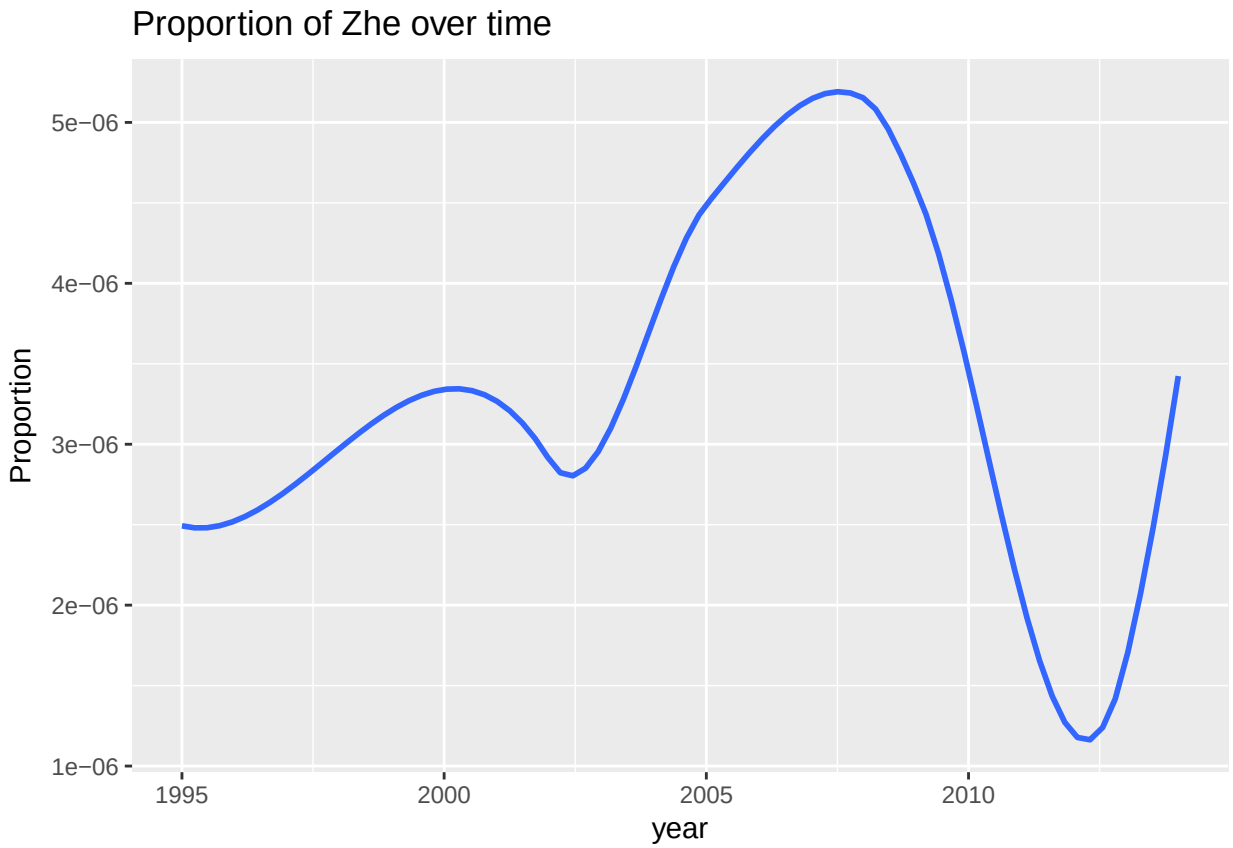
```
## `geom_smooth()` using method = 'loess'
```



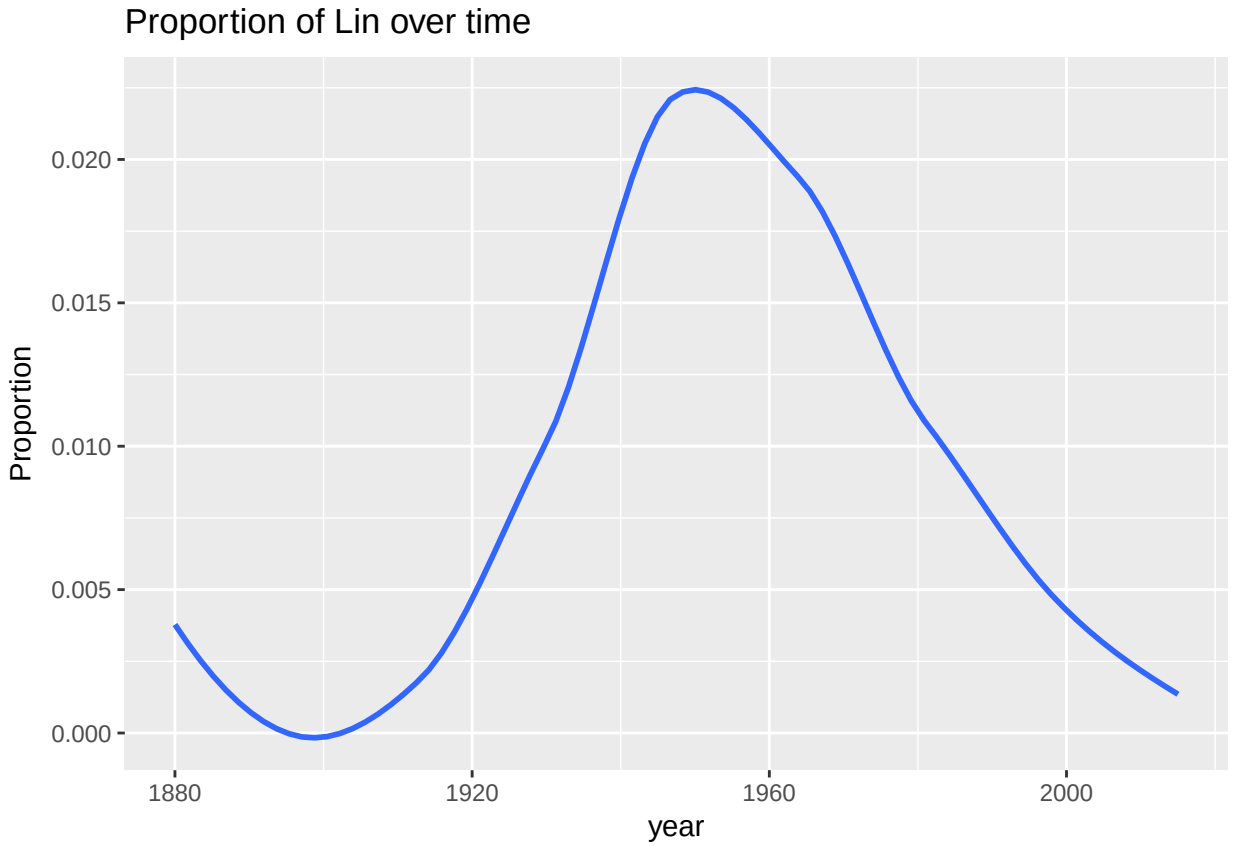

```
## `geom_smooth()` using method = 'loess'
```




```
## `geom_smooth()` using method = 'loess'
```




```
## `geom_smooth()` using method = 'loess'
```



The person who has the highest ratio for their name is Zhenlong, with 58%. Scott came in 2nd with 43%, Lexie is next with 18%, and Lindsay is last with 15%.

In the graph for the proportion of Lex over time, it is very small and constant for a while until a little after 1960, then it increases rapidly for the rest of the plot. A name starting with Lex was still pretty uncommon in 1998, with a proportion around 0.0005.

In the graph for Sco, it starts out low at 1880, decreases then starts increasing around 1930. The popularity of Sco hits a peak around 1970 then decreases for the rest of the plot. In 1997, the proportion of Sco is around 0.005, but is decreasing for later years.

In the graph for Zhe, there is no data available for the years before 1995 but the graph osciallates for the 20 years. Around 1997 the proportion for Zhe was still very uncommon, around 0.000003 on an increasing trend.

In the graph for Lin it starts out decreasing in 1880 then increases around 1900's. The popularity of Lin hits a peak around 1950 then is decreasing for the rest of the plot. Around 1998, Lin had a proportion around 0.005.

Ariel and Rachel regexs

There were 2 different versions of the name "Ariel" in 1973, 9 in 1988, and 11 in 1990.

```
#how many versions of the name "Ariel" were there in 1973; 1988; 1990?
ariel1973 <- filter(babynames,str_detect(babynames$name,"^Ar[iey].l+[^a]?$"),year==1973,sex=="F")
ariel1988 <- filter(babynames,str_detect(babynames$name,"^Ar[iey].l+[^a]?$"),year==1988,sex=="F")
ariel1990 <- filter(babynames,str_detect(babynames$name,"^Ar[iey].l+[^a]?$"),year==1990,sex=="F")
count(ariel1973)
```

```
## # A tibble: 1 x 1
```

```
##      nn
##   <int>
## 1     2
```

```
count(ariel1988)
```

```
## # A tibble: 1 x 1
##      nn
##   <int>
## 1     9
```

```
count(ariel1990)
```

```
## # A tibble: 1 x 1
##      nn
##   <int>
## 1    11
```

There were 5 different ways to spell "Rachel" in 1973, 7 in 1988, and 6 in 1990.

```
#how many versions of the name "Rachel" were there in 1973; 1988; 1990?
rachel1973 <- filter(babynames,str_detect(babynames$name,"^Rach[:lower:]*l$"),year==1973,sex=="F")
rachel1988 <- filter(babynames,str_detect(babynames$name,"^Rach[:lower:]*l$"),year==1988,sex=="F")
rachel1990 <- filter(babynames,str_detect(babynames$name,"^Rach[:lower:]*l$"),year==1990,sex=="F")
count(rachel1973)
```

```
## # A tibble: 1 x 1
##      nn
##   <int>
## 1     5
```

```
count(rachel1988)
```

```
## # A tibble: 1 x 1
##      nn
##   <int>
## 1     7
```

```
count(rachel1990)
```

```
## # A tibble: 1 x 1
##      nn
##   <int>
## 1     6
```

The chance that a girl born in 1973 would be named Rachel or Ariel (or versions of) is 0.4735%. For 1988 and 1990, the chances are 1.0577% and 1.1198%, respectively.

```
#What are the chances a girl born in 1973 would be named either Rachel or Ariel (including various vers
sum(ariel1973$prop) + sum(rachel1973$prop)
```

```
## [1] 0.004735488
```

```
sum(ariel1988$prop) + sum(rachel1988$prop)
```

```
## [1] 0.01057712
```

```
sum(ariel1990$prop) + sum(rachel1990$prop)
```

```
## [1] 0.01198373
```

"The Little Mermaid" Effect

```
# Did *The Little Mermaid* cause more baby girls to be named Ariel?
library(babynames)
a = filter(babynames,str_detect(babynames$name,"Ar[iy]+.l+[^a]?$"),year==1988,sex=="F") %>% count(wt=pr
b = filter(babynames,str_detect(babynames$name,"Ar[iy]+.l+[^a]?$"),year==1990,sex=="F") %>% count(wt=pr
Ariel.diff = b - a
print(Ariel.diff)
```

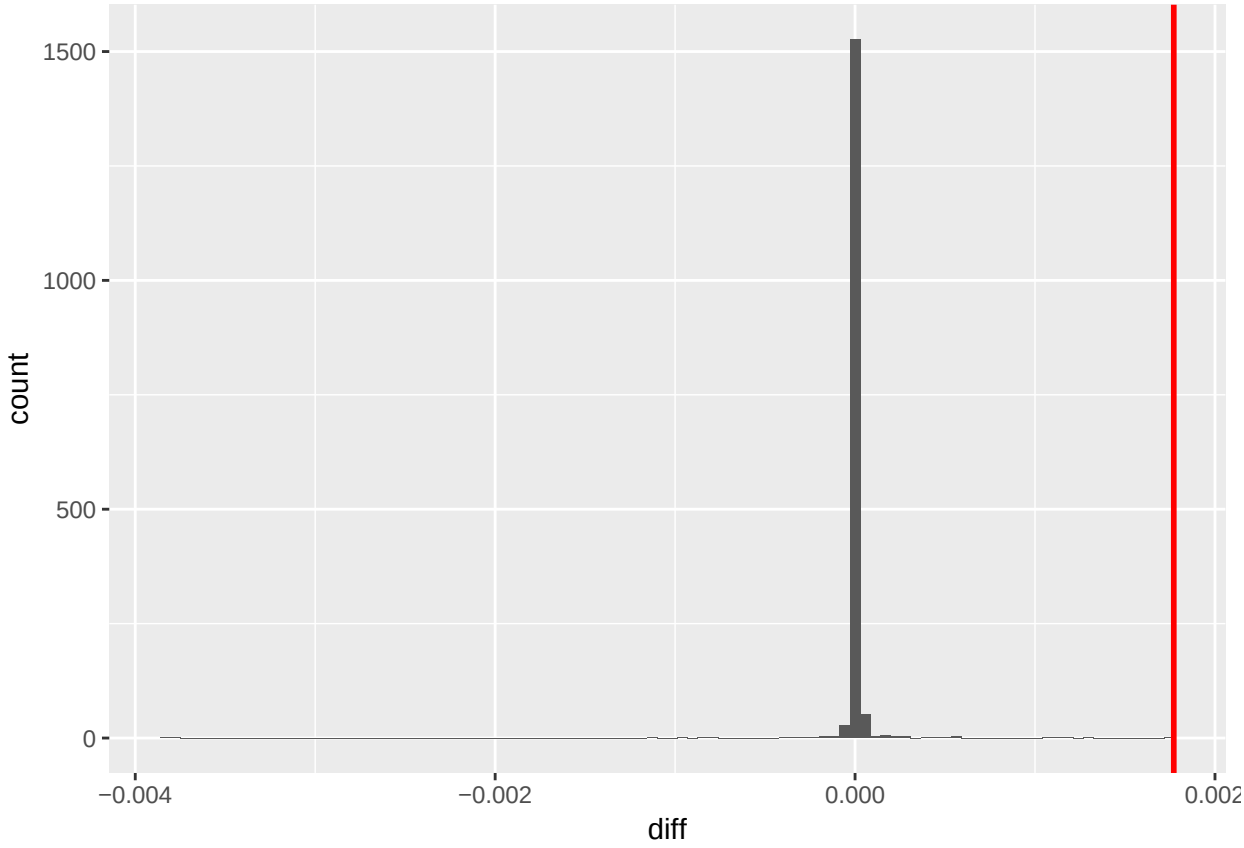
```
##           nn
## 1 0.00177116
```

```
# Deciding on list of Vowel girl names
vowel_names88 = filter(babynames, str_detect(babynames$name, "^[AEIOUY]"), year == 1988, sex == "F")

vowel_names90 = filter(babynames, str_detect(babynames$name, "^[AEIOUY]"), year == 1990, sex == "F")

Vowel_girls <- inner_join(vowel_names88, vowel_names90, by="name") %>% mutate(diff=prop.y-prop.x) %>% s

ggplot(Vowel_girls) +
  geom_histogram(aes(x=diff),bins=100) +
  geom_vline(aes(xintercept=Ariel.diff),color="red",lwd=1)
```



From this plot, we can see that the change in proportion of female "Ariel" baby names from 1988 to 1990 is bigger than the changes in other female names starting with a vowel over that time period.

```
# What percentile change is Ariel's change?
filter(Vowel_girls,diff<Ariel.diff) %>% arrange(desc(diff))
```

```
## # A tibble: 1,651 x 4
##    name         prop.x   prop.y     diff
##    <chr>         <dbl>    <dbl>    <dbl>
##  1 Alyssa     0.00371  0.00548  0.00177
##  2 Ariel      0.000473 0.00176  0.00128
##  3 Emily      0.00825  0.00943  0.00117
##  4 Olivia     0.00111  0.00225  0.00114
##  5 Alexis     0.00151  0.00256  0.00105
##  6 Abigail    0.00123  0.00181  0.000576
##  7 Alexandria 0.00118  0.00175  0.000567
##  8 Alexa      0.000478 0.00104  0.000560
##  9 Alexandra  0.00362  0.00414  0.000518
## 10 Arielle    0.000340 0.000764 0.000424
## # ... with 1,641 more rows
```

```
filter(Vowel_girls,diff<Ariel.diff) %>% count() / Vowel_girls %>% count()
```

```
##   n
## 1 1
```

```
 # 100 percentile
```

Alyssa has the biggest positive change from 1988 to 1990. Ariel has the second biggest change of all girl names. No baby girl name with other version of Ariels had a bigger difference in proportions than "Ariel".

Our Names

```
NAMES<-babynames::babynames
Lindsay98<-filter(NAMES, str_detect(NAMES$name, "^L[i|y]nd?s[a|e]?y"), year==1998)
Lee96<-filter(NAMES, str_detect(NAMES$name, "^L[e|y|i]e?$"), year==1996)
Lexie98<-filter(NAMES, str_detect(NAMES$name, "Lee?xx?i?e?[e|y|i]$"), year==1998)
Scott97<-filter(NAMES, str_detect(NAMES$name, "S[c|k]c?ott?$"), year==1997)

Lindsay15<-filter(NAMES, str_detect(NAMES$name, "^L[i|y]nds[a|e]?y"), year==2015)
Lee15<-filter(NAMES, str_detect(NAMES$name, "^L[e|y|i]e?$"), year==2015)
Lexie15<-filter(NAMES, str_detect(NAMES$name, "Lee?xx?i?e?[e|y|i]$"), year==2015)
Scott15<-filter(NAMES, str_detect(NAMES$name, "S[c|k]c?ott?$"), year==2015)

Lindsay79<-filter(NAMES, str_detect(NAMES$name, "^L[i|y]nds[a|e]?y"), year==1979)
Lee79<-filter(NAMES, str_detect(NAMES$name, "^L[e|y|i]e?$"), year==1979)
Lexie79<-filter(NAMES, str_detect(NAMES$name, "Lee?xx?i?e?[e|y|i]$"), year==1979)
Scott79<-filter(NAMES, str_detect(NAMES$name, "S[c|k]c?ott?$"), year==1979)

changePropLindsay<-inner_join(Lindsay15, Lindsay98, by="name")%>%
  mutate(newProp=prop.x-prop.y)%>%
  count(wt=newProp)
changePropLee<-inner_join(Lee15, Lee96, by="name")%>%
  mutate(newProp=prop.x-prop.y)%>%
  count(wt=newProp)
changePropLexie<-inner_join(Lexie98, Lexie15, by="name")%>%
```

```
  mutate(newProp=prop.x-prop.y)%>%
  count(wt=newProp)
changePropScott<-inner_join(Scott97, Scott15, by="name")%>%
  mutate(newProp=prop.x-prop.y)%>%
  count(wt=newProp)
changePropLindsay2<-inner_join(Lindsay79, Lindsay98, by="name")%>%
  mutate(newProp=prop.x-prop.y)%>%
  count(wt=newProp)
changePropLee2<-inner_join(Lee79, Lee96, by="name")%>%
  mutate(newProp=prop.x-prop.y)%>%
  count(wt=newProp)
changePropLexie2<-inner_join(Lexie98, Lexie79, by="name")%>%
  mutate(newProp=prop.x-prop.y)%>%
  count(wt=newProp)
changePropScott2<-inner_join(Scott97, Scott79, by="name")%>%
  mutate(newProp=prop.x-prop.y)%>%
  count(wt=newProp)
```

```
all15<-filter(NAMES, year==2015)
all79<-filter(NAMES, year==1979)
all<-filter(NAMES, year==1997)

changeProp15<-inner_join(all, all15, by="name")%>%
  mutate(diff=prop.x-prop.y)%>%
  count(wt=diff)
changeProp79<-inner_join(all, all79, by="name")%>%
  mutate(diff=prop.x-prop.y)%>%
  count(wt=diff)

message("Change in proportions to 2015: \n", "Lee:", changePropLee, "\n", "Lexie: ", changePropLexie, "
```

```
## Change in proportions to 2015:
## Lee:-0.000427223010988068
## Lexie: -0.0001269814341058
## Lindsay: -0.00502124503443007
## Scott: 0.00104451414741837
```

```
message("Relative Change as percent (same order as above): \n", 100*changePropLee/changeProp15, "\n", 1
```

```
## Relative Change as percent (same order as above):
## -0.722151654301705
## -0.214641651658757
## -8.48760557134216
## 1.76558284572546
```

```
message("Change in proportions since 1997: \n", "Lee:", changePropLee2, "\n", "Lexie: ", changePropLexi
```

```
## Change in proportions since 1997:
## Lee:0.00224152002456179
## Lexie: 0.00061517628817823
## Lindsay: 0.00174224567720317
## Scott: -0.0129780259628391
```

```
message("Relative Change as percent (same order as above): \n", 100*changePropLee2/changeProp15, "\n",
```

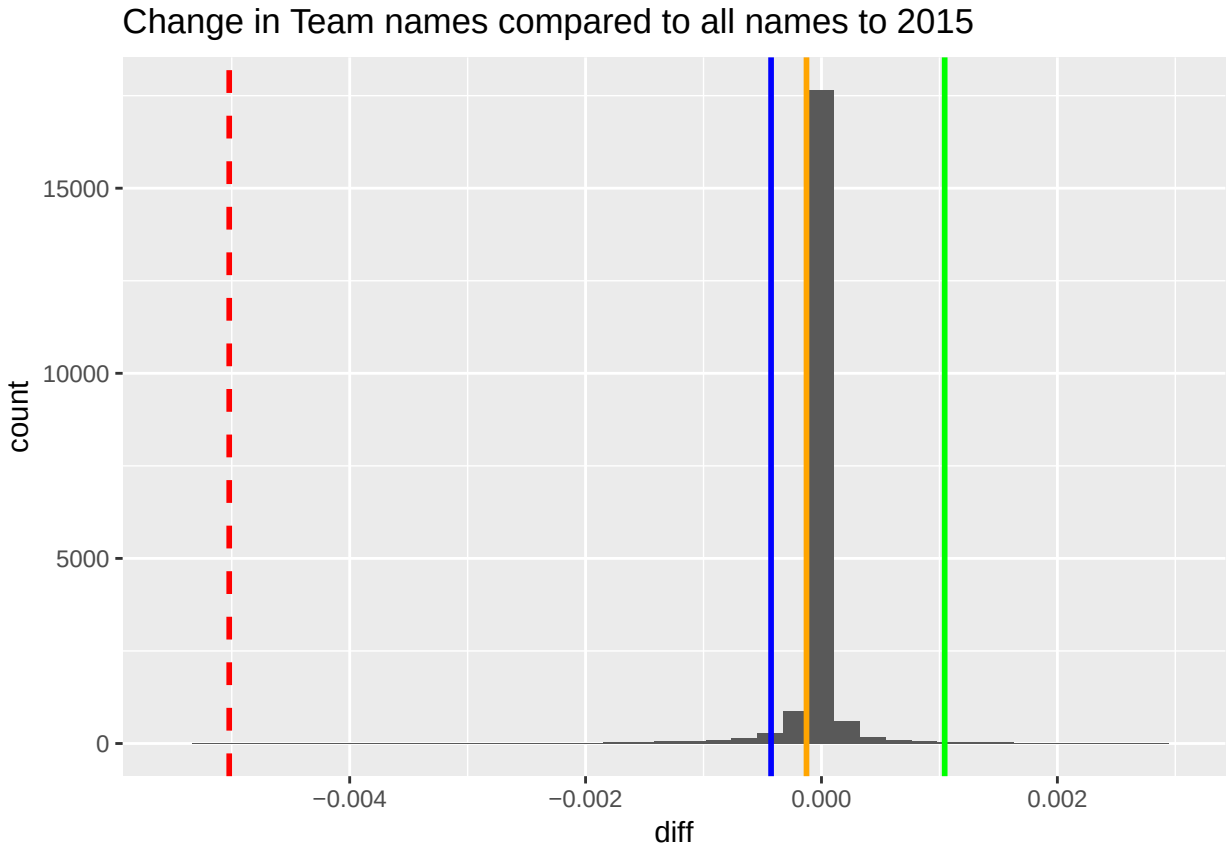
```
## Relative Change as percent (same order as above):
```

```
## 3.7889283869424
## 1.03985638125541
## 2.9449855593743
## -21.9372615181922
```

```
inner_join(all, all15, by="name")%>%
  mutate(diff=prop.x-prop.y)%>%
  select(name, prop.x, prop.y, diff)%>%
  ggplot()+ geom_histogram(mapping=aes(x=diff), bins=40) +xlim(-0.0055,0.003)+
  geom_vline(aes(xintercept=changePropLee), color="blue", lwd=1, show.legend=TRUE)+
  geom_vline(aes(xintercept=changePropLindsay), color="red", lwd=1, show.legend = TRUE, linetype="dashed
  geom_vline(aes(xintercept=changePropLexie), color="orange", lwd=1, show.legend = TRUE) +
  geom_vline(aes(xintercept=changePropScott), color="green", lwd=1, show.legend = TRUE) +
  ggtitle("Change in Team names compared to all names to 2015")
```

```
## Warning: Removed 172 rows containing non-finite values (stat_bin).
```

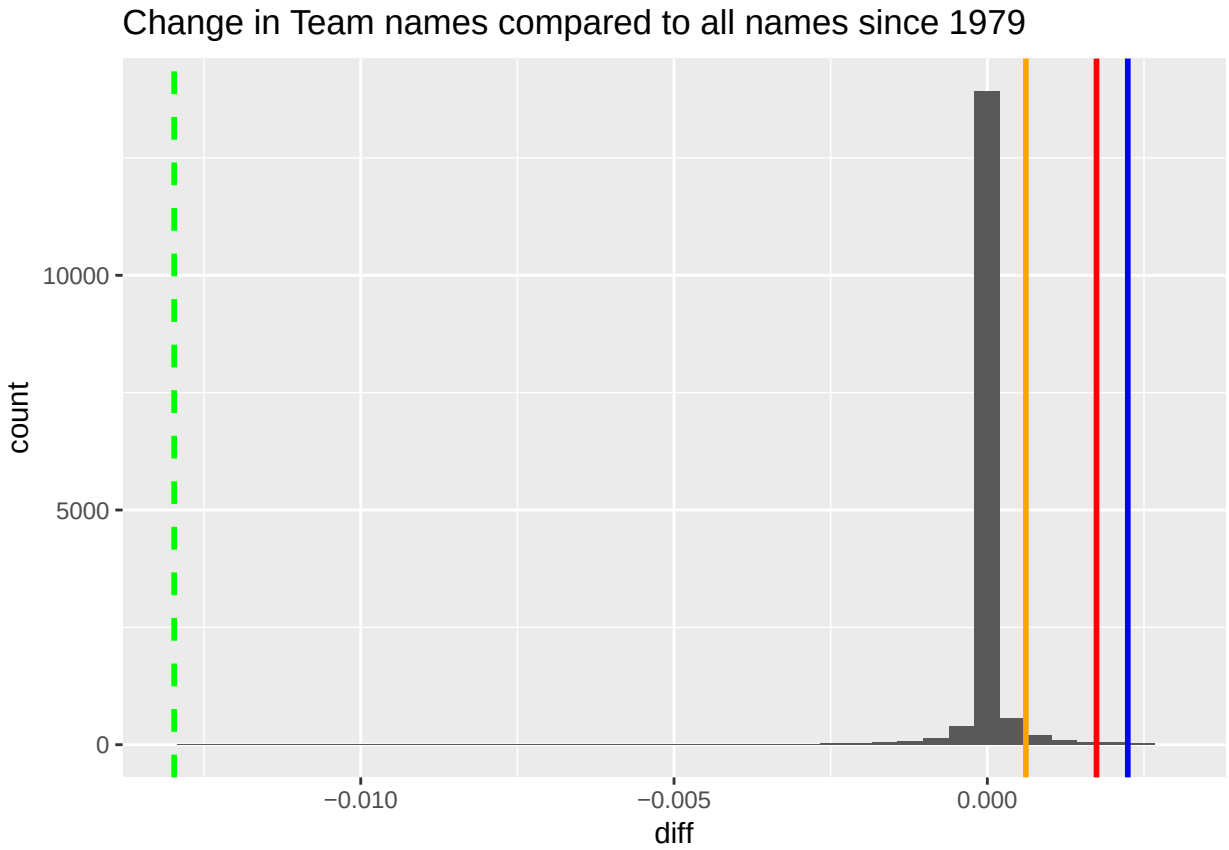
```
## Warning: Removed 1 rows containing missing values (geom_bar).
```




```
inner_join(all, all79, by="name")%>%
  mutate(diff=prop.x-prop.y)%>%
  select(name, prop.x, prop.y, diff)%>%
  ggplot()+ geom_histogram(mapping=aes(x=diff), bins=40) +xlim(-0.013,0.003)+
  geom_vline(aes(xintercept=changePropLee2), color="blue", lwd=1, show.legend=TRUE)+
  geom_vline(aes(xintercept=changePropLindsay2), color="red", lwd=1, show.legend = TRUE) +
  geom_vline(aes(xintercept=changePropLexie2), color="orange", lwd=1, show.legend = TRUE) +
  geom_vline(aes(xintercept=changePropScott2), color="green", lwd=1, show.legend = TRUE, linetype="dash
  ggtitle("Change in Team names compared to all names since 1979")
```

```
## Warning: Removed 168 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



The vertical lines in the graph correspond to the difference in name proportions for each team member, "Lee" is blue, "Lindsay" is red, "Lexie" is orange and "Scott" is green. The two graphs above illustrate the changes in teammember names from 1997-2015 and 1979-1997 in comparison to the changes in all names in the dataset over the same years. The dotted vertical lines signify the names which underwent significant changes in popularity over the range of years. Overall the name "Lexie" experienced the least amount of change in both ranges of years, as the differene in proportions is closer to zero. In contrast "Lindsay" and "Scott" demonstrate larger changes in popularity both before 1997 and after 1997. The name "Lee" conveys a mpderate increase in popularity from 1979-1997, and a slight decrease from 1997-2015. Since the bithyears span 1996-1998, to account for the variation of years, it was decided to average the birthyears and used 1997 as the comparison year. It was also decided to not filter by sex, and maintain all results for each name.

Contributions

- Lindsay: Team decided to use my code for the "your names section", using regex, ggplot, join, and dplyr functions.
- Lexie: The team decided to my plots the first three letters of name section.
- Li: Our team decided that "the Littile Mermaid Effect" part of mine was the best, so I added my write-up to that section.
- Scott: The group decided that my verion of the "Ariel and Rachel regex" section was the strongest, so I added that part to the lab. Additionally, I was responsible for knitting, submitting, and managing the git.