# String Distances

Allen Rahrooh

14 April 2020

## Contents

## Task I

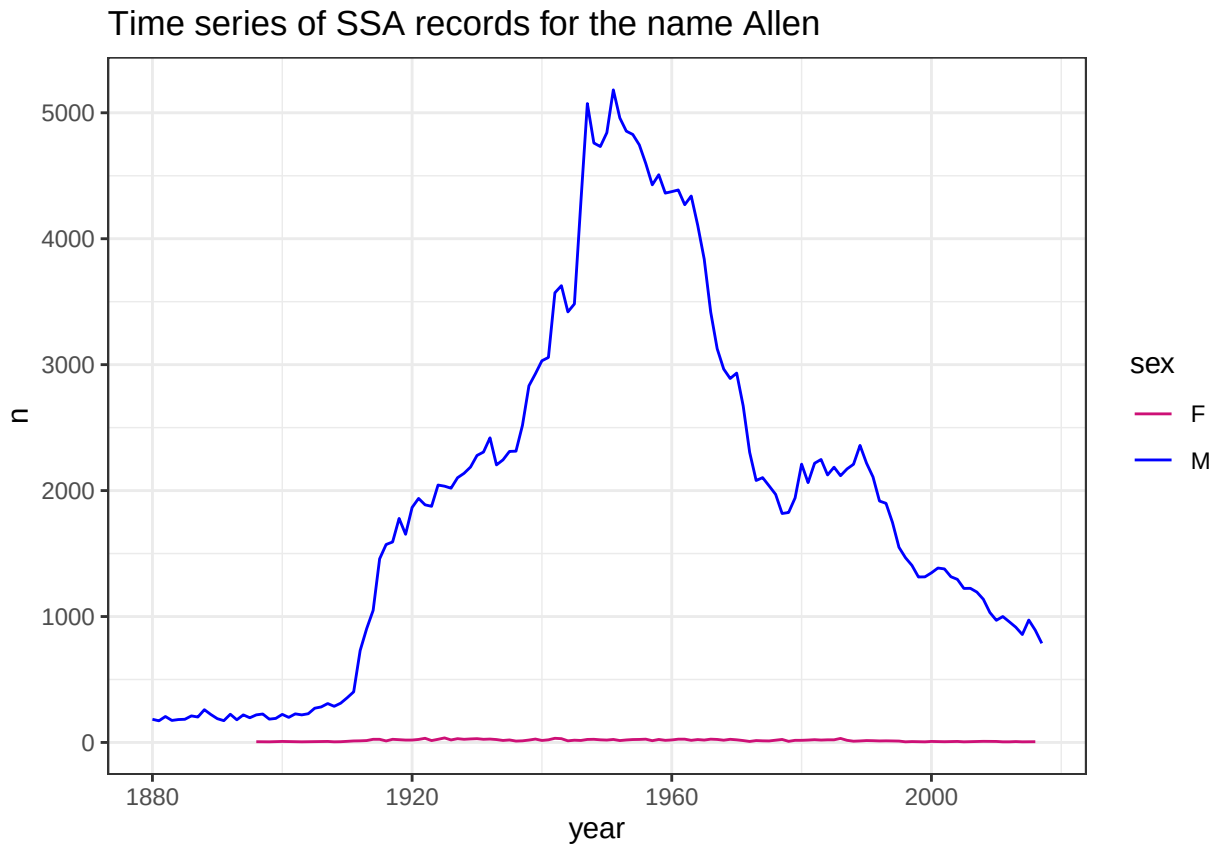
Find the ten names in the babynames::babynames data set that are the most similar to your first name 2. Plot the names as times series by year. Put the string distance used in the title of the plot. (9 plots, 90 points)

```
library(ggplot2)
library(babynames)

common_name <- as.data.frame(babynames)

Name_Allen <- common_name[common_name$name == "Allen",]

ggplot(Name_Allen) +
  aes(x = year,y = n,group = sex,color = sex) +
  geom_line() +
  theme_bw() +
  ggtitle("Time series of SSA records for the name Allen ") +
  scale_color_manual(values = c("deeppink3","blue"))
```




```
distance_name <- sort(unique(babynames$name))

distance_method <- c(
  "lcs", "lv", "osa", "dl",
  "qgram", "jw", "jaccard", "cosine", "soundex"
```

```
)


distance_Allen <- sapply(X = distance_method,
                         FUN = function(x)
  {
  stringdist::stringdist(a = "Allen",  b = distance_name, method = x)
  }
  )

rownames(distance_Allen) <- distance_name
colnames(distance_Allen) <- distance_method
head(distance_Allen)
```

```
##           lcs lv osa dl qgram        jw   jaccard    cosine soundex
## Aaban       6  3   3  3     6 0.4000000 0.6666667 0.7142857       1
## Aabha       8  4   4  4     8 0.5333333 0.8571429 0.8571429       1
## Aabid       8  4   4  4     8 0.5333333 0.8750000 0.8309691       1
## Aabir       8  4   4  4     8 0.5333333 0.8750000 0.8309691       1
## Aabriella   8  7   7  7     6 0.4592593 0.6250000 0.3710291       1
## Aada        7  4   4  4     7 0.5166667 0.8333333 0.8456967       1
```

```
round(cor(distance_Allen),3)
```

```
##           lcs    lv   osa    dl qgram    jw jaccard cosine soundex
## lcs     1.000 0.837 0.837 0.838 0.913 0.529   0.649  0.616   0.138
## lv      0.837 1.000 0.999 0.997 0.737 0.263   0.342  0.301   0.110
## osa     0.837 0.999 1.000 0.998 0.739 0.263   0.344  0.303   0.110
## dl      0.838 0.997 0.998 1.000 0.743 0.265   0.348  0.306   0.111
## qgram   0.913 0.737 0.739 0.743 1.000 0.556   0.800  0.766   0.125
## jw      0.529 0.263 0.263 0.265 0.556 1.000   0.720  0.696   0.094
## jaccard 0.649 0.342 0.344 0.348 0.800 0.720   1.000  0.903   0.172
## cosine  0.616 0.301 0.303 0.306 0.766 0.696   0.903  1.000   0.118
## soundex 0.138 0.110 0.110 0.111 0.125 0.094   0.172  0.118   1.000
```

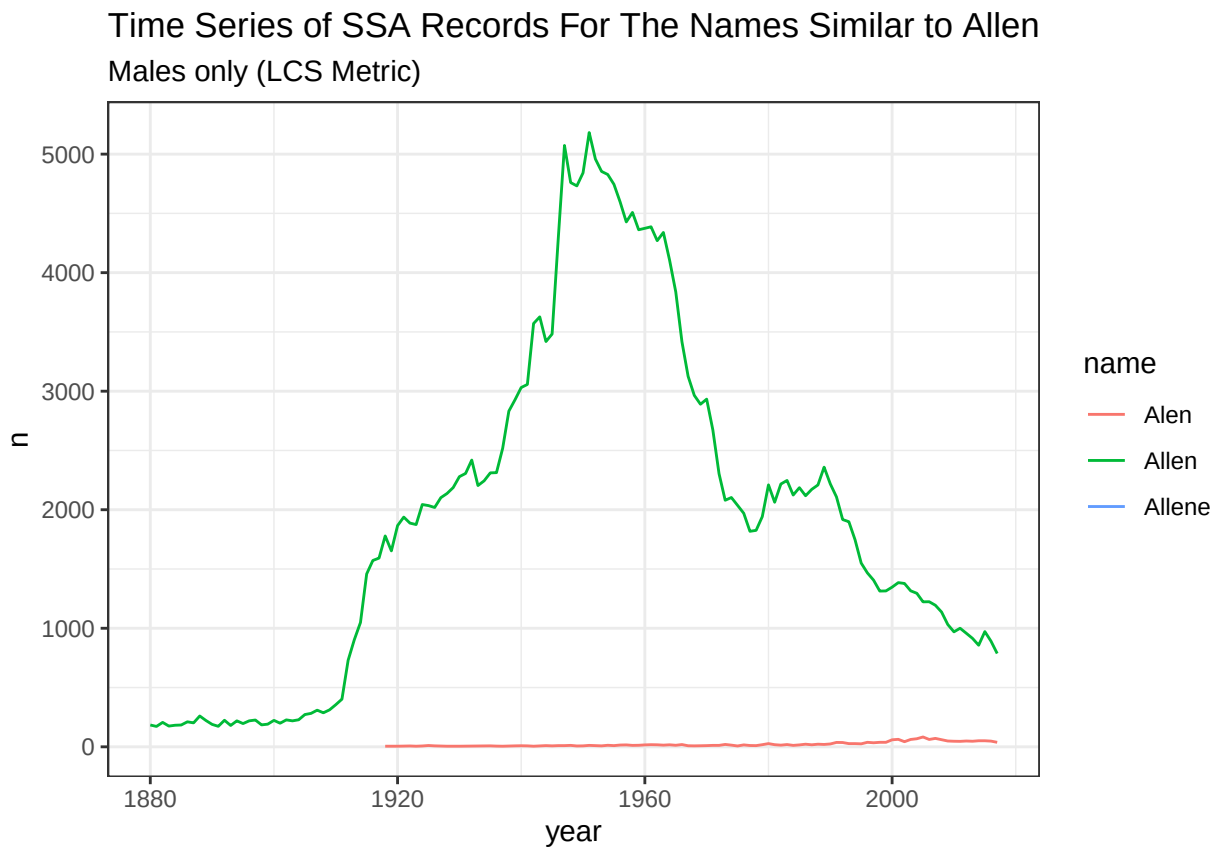
## LCS Metric

```
distance_name.lcs <- distance_name[order(distance_Allen[,"lcs"])]

common_name.2 <- common_name[common_name$name %in% distance_name.lcs[1:10],]

common_name.2<- common_name.2[common_name.2$sex == "M",]


ggplot(common_name.2) +
  aes(x = year,y = n,group = name,color = name) +
  geom_line() +
  theme_bw() +
  ggtitle("Time Series of SSA Records For The Names Similar to Allen","Males only (LCS Metric)")
```



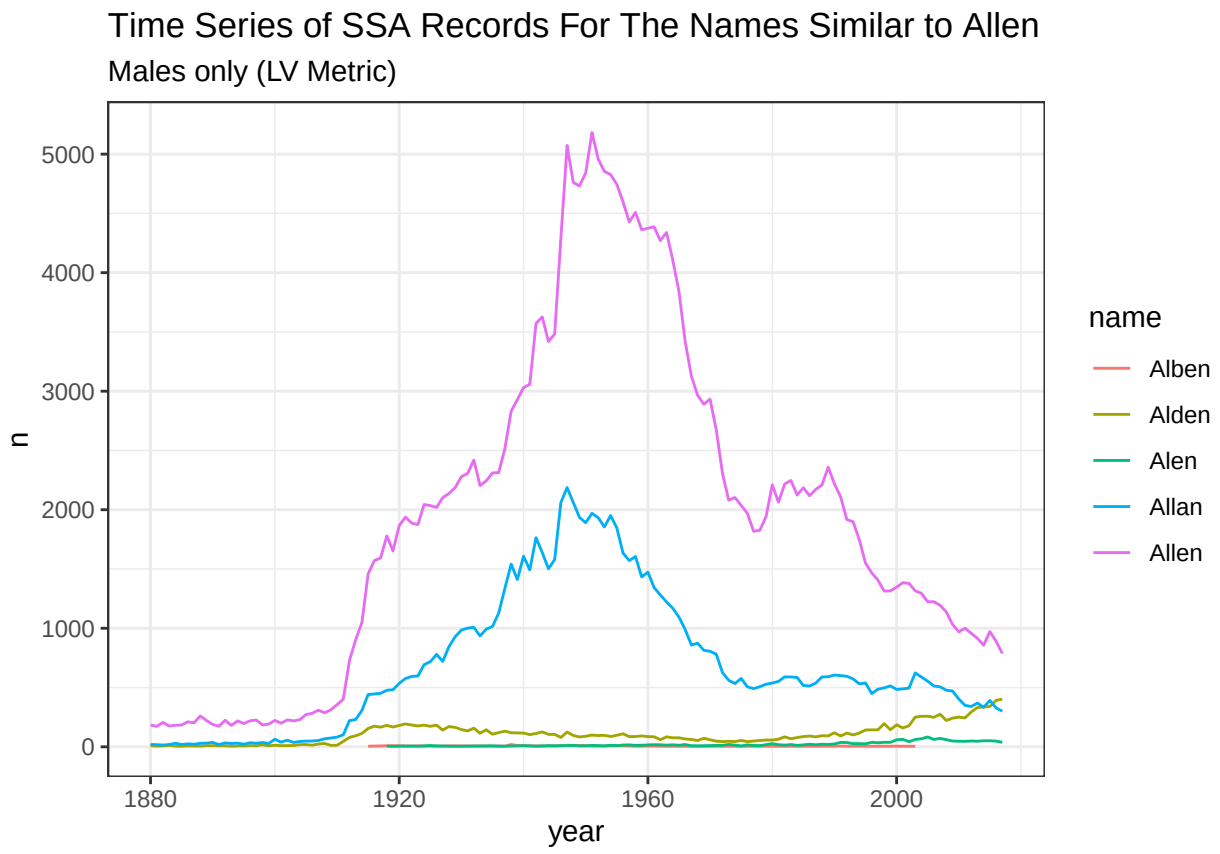

## LV Metric

```
distance_name.lv <- distance_name[order(distance_Allen[,"lv"])]

common_name.2 <- common_name[common_name$name %in% distance_name.lv[1:10],]

common_name.2<- common_name.2[common_name.2$sex == "M",]


ggplot(common_name.2) +
  aes(x = year,y = n,group = name,color = name) +
  geom_line() +
  theme_bw() +
  ggtitle("Time Series of SSA Records For The Names Similar to Allen","Males only (LV Metric)")
```



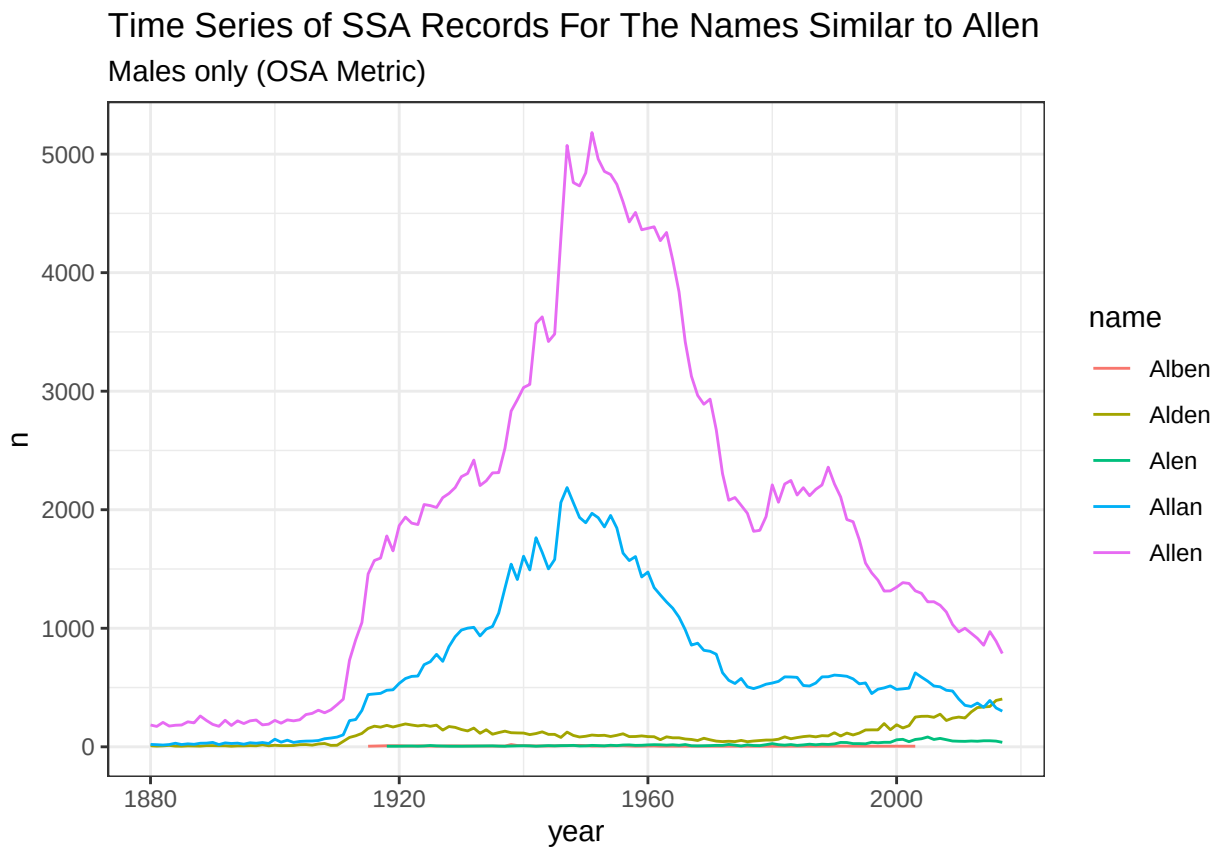

## OSA Metric

```
distance_name.osa <- distance_name[order(distance_Allen[,"osa"])]

common_name.2 <- common_name[common_name$name %in% distance_name.osa[1:10],]

common_name.2<- common_name.2[common_name.2$sex == "M",]


ggplot(common_name.2) +
  aes(x = year,y = n,group = name,color = name) +
  geom_line() +
  theme_bw() +
  ggtitle("Time Series of SSA Records For The Names Similar to Allen","Males only (OSA Metric)")
```



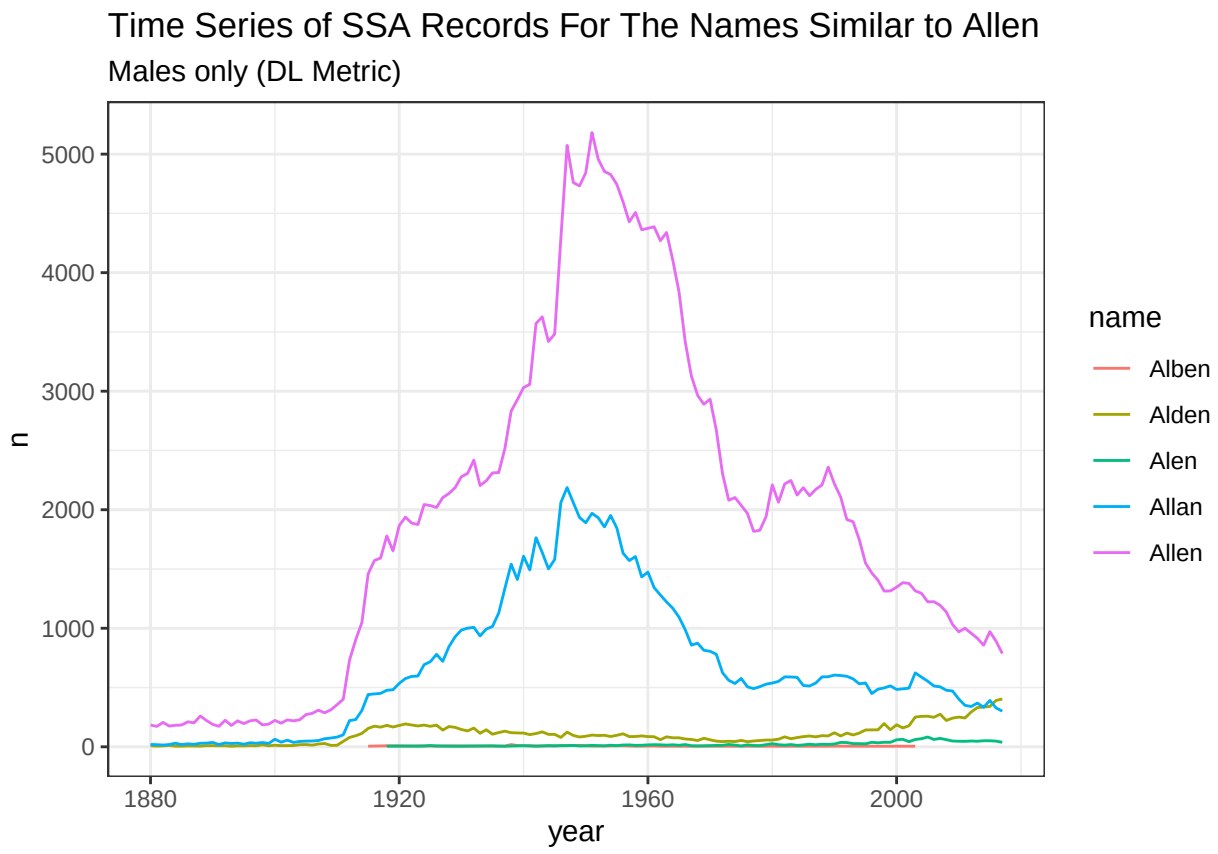

## DL Metric

```
distance_name.dl <- distance_name[order(distance_Allen[,"dl"])]

common_name.2 <- common_name[common_name$name %in% distance_name.dl[1:10],]

common_name.2<- common_name.2[common_name.2$sex == "M",]


ggplot(common_name.2) +
  aes(x = year,y = n,group = name,color = name) +
  geom_line() +
  theme_bw() +
  ggtitle("Time Series of SSA Records For The Names Similar to Allen","Males only (DL Metric)")
```



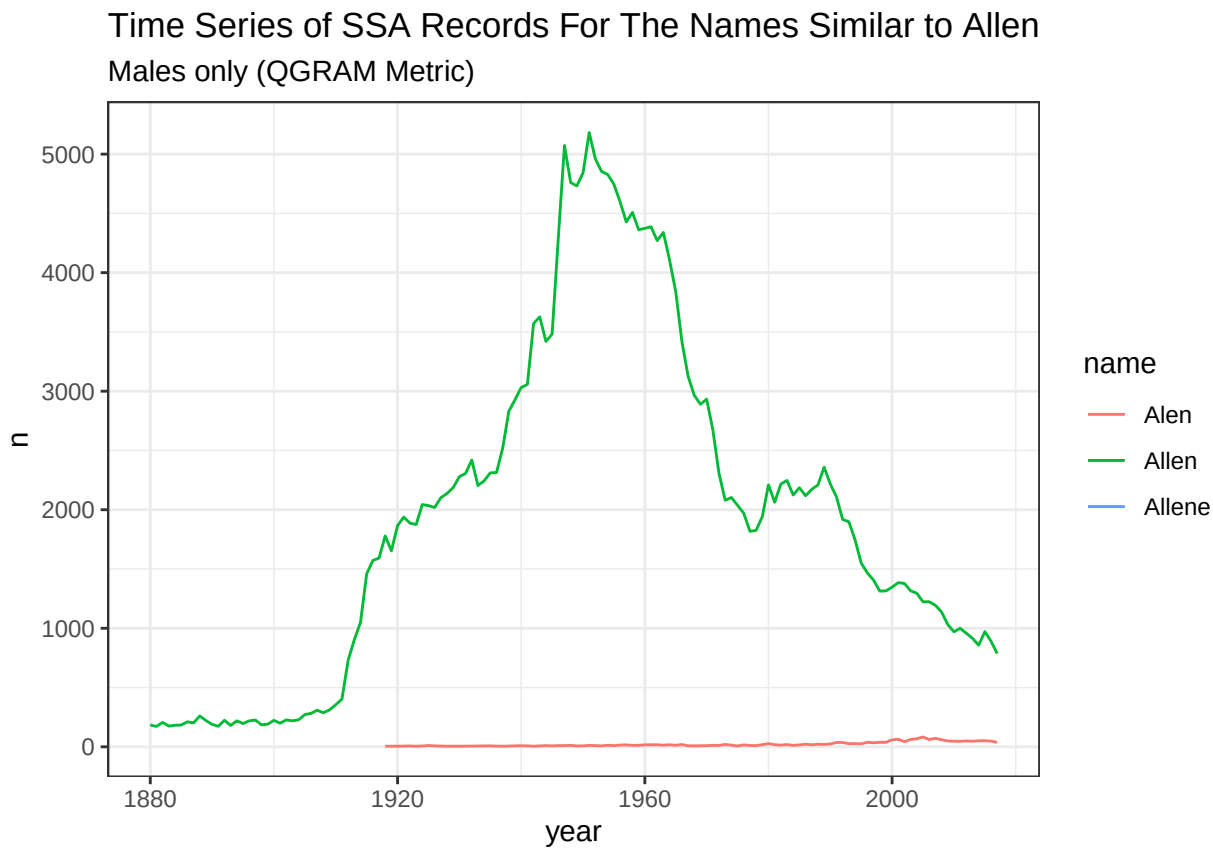

## QGRAM Metric

```
distance_name.qgram <- distance_name[order(distance_Allen[,"qgram"])]

common_name.2 <- common_name[common_name$name %in% distance_name.qgram[1:10],]

common_name.2<- common_name.2[common_name.2$sex == "M",]


ggplot(common_name.2) +
  aes(x = year,y = n,group = name,color = name) +
  geom_line() +
  theme_bw() +
  ggtitle("Time Series of SSA Records For The Names Similar to Allen","Males only (QGRAM Metric)")
```




### JW Metric

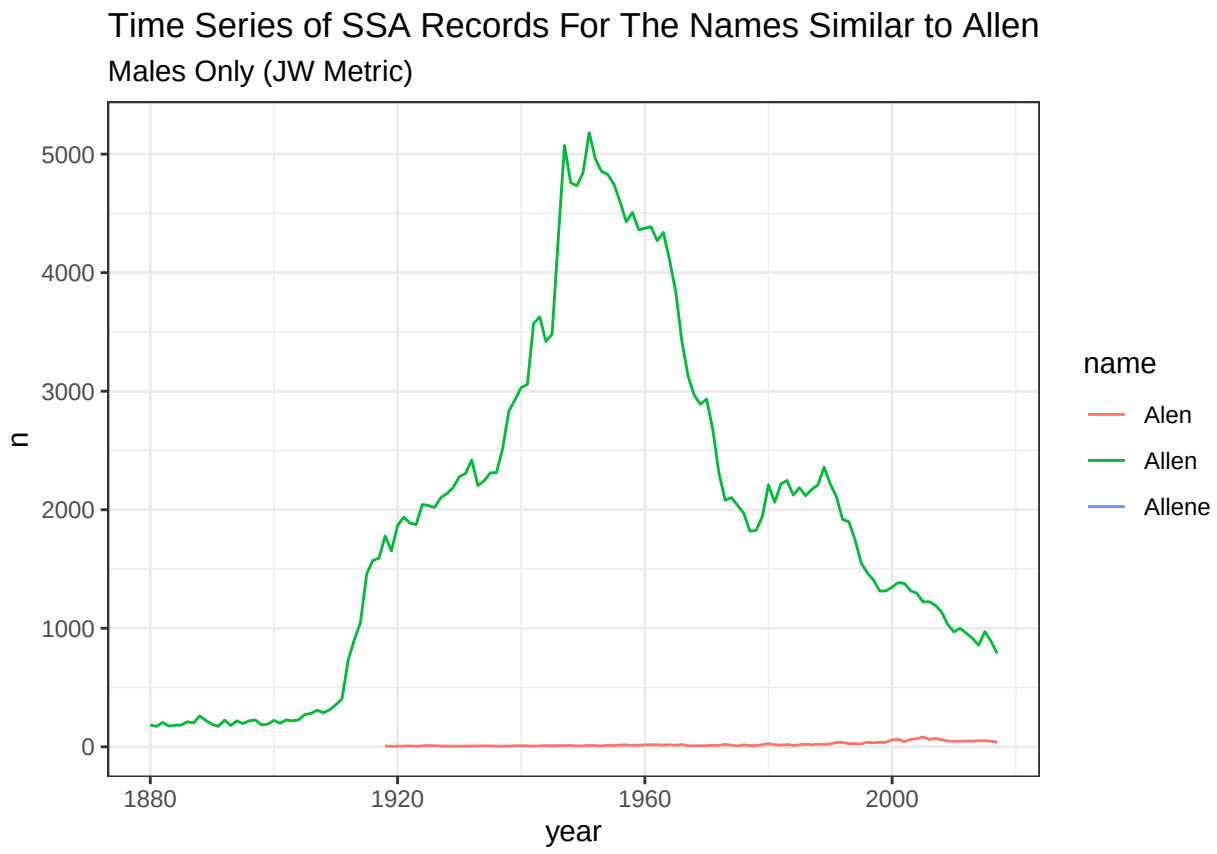
```
#first ten names most similiar to Allen

distance_name.jw <- distance_name[order(distance_Allen[,"jw"])]

common_name.2 <- common_name[common_name$name %in% distance_name.jw[1:10],]

common_name.2<- common_name.2[common_name.2$sex == "M",]


ggplot(common_name.2) +
  aes(x = year,y = n,group = name,color = name) +
  geom_line() +
  theme_bw() +
  ggtitle("Time Series of SSA Records For The Names Similar to Allen","Males Only (JW Metric)")
```



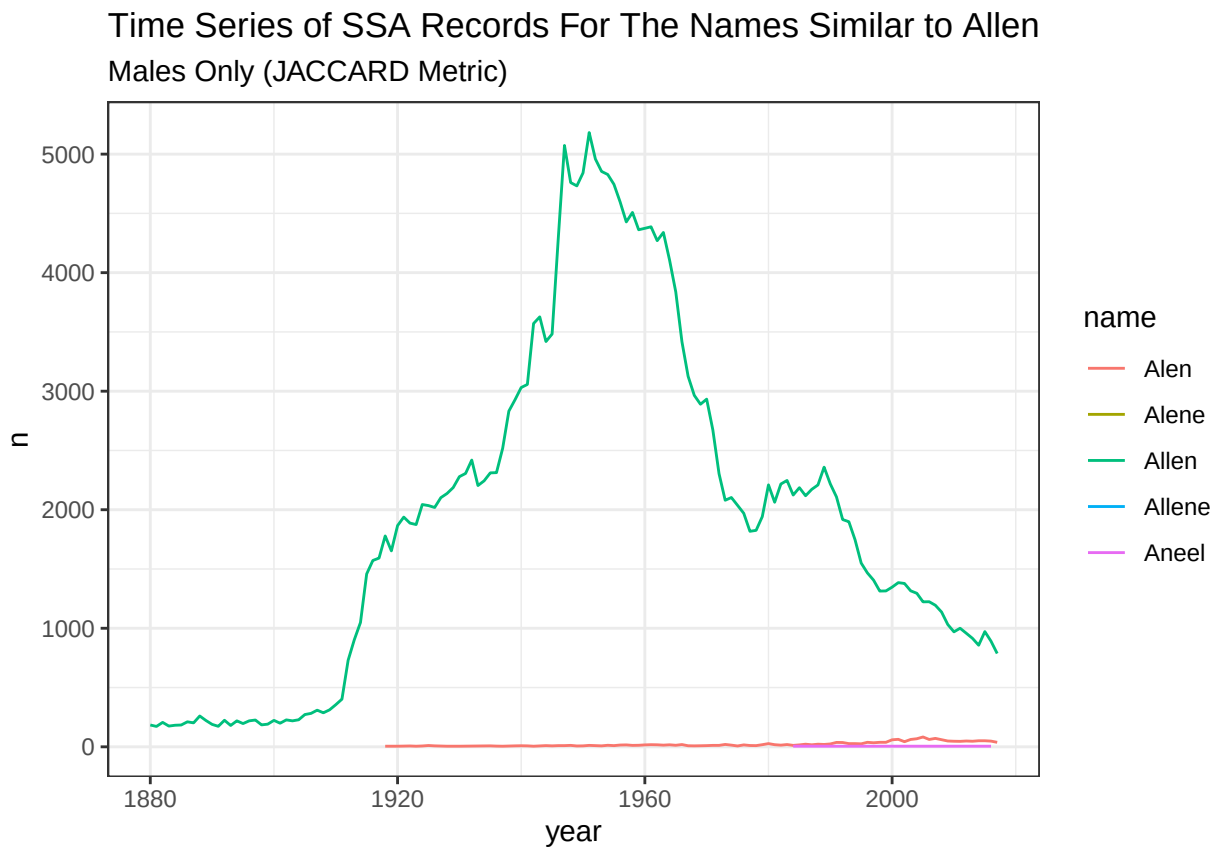

## JACCARD Metric

```
distance_name.jaccard <- distance_name[order(distance_Allen[,"jaccard"])]

common_name.2 <- common_name[common_name$name %in% distance_name.jaccard[1:10],]

common_name.2<- common_name.2[common_name.2$sex == "M",]


ggplot(common_name.2) +
  aes(x = year,y = n,group = name,color = name) +
  geom_line() +
  theme_bw() +
  ggtitle("Time Series of SSA Records For The Names Similar to Allen","Males Only (JACCARD Metric)")
```



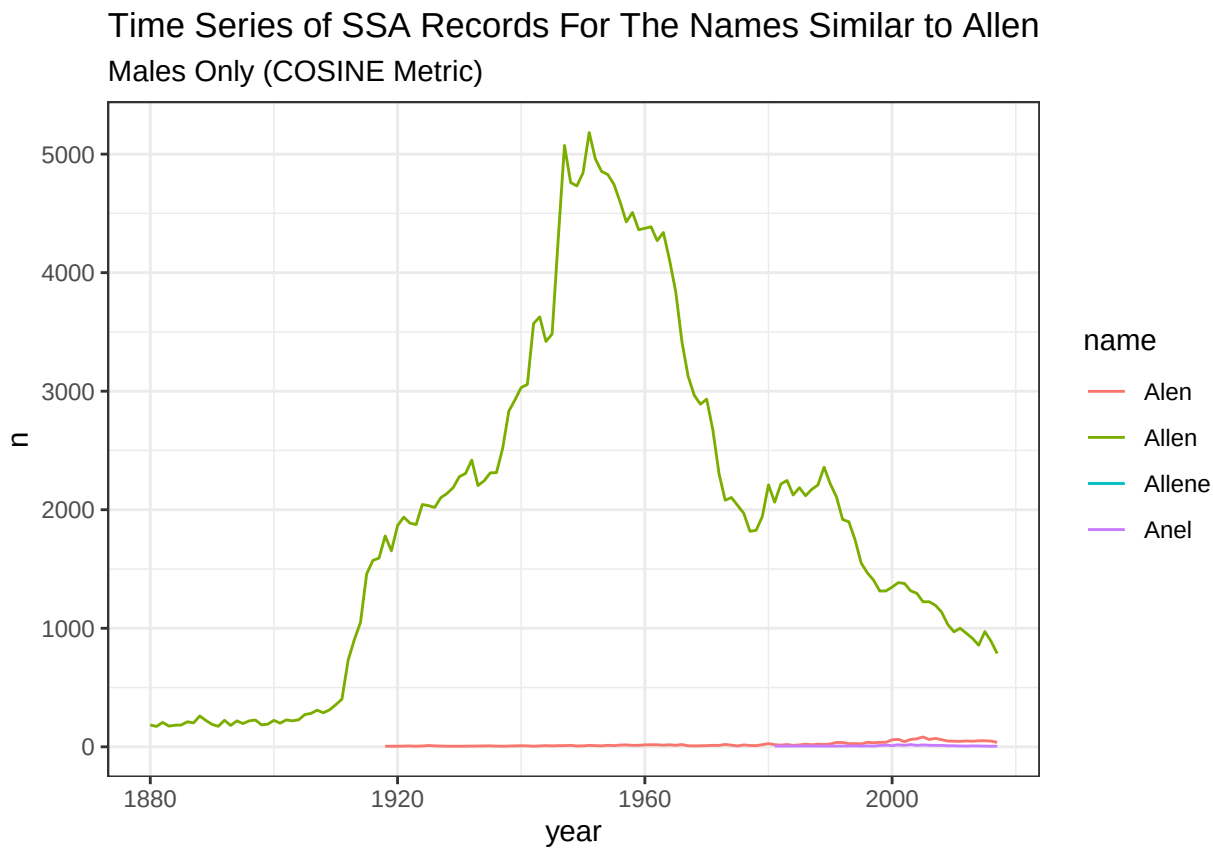

### COSINE Metric

```
distance_name.cosine<- distance_name[order(distance_Allen[,"cosine"])]

common_name.2 <- common_name[common_name$name %in% distance_name.cosine[1:10],]

common_name.2<- common_name.2[common_name.2$sex == "M",]


ggplot(common_name.2) +
  aes(x = year,y = n,group = name,color = name) +
  geom_line() +
  theme_bw() +
  ggtitle("Time Series of SSA Records For The Names Similar to Allen","Males Only (COSINE Metric)")
```



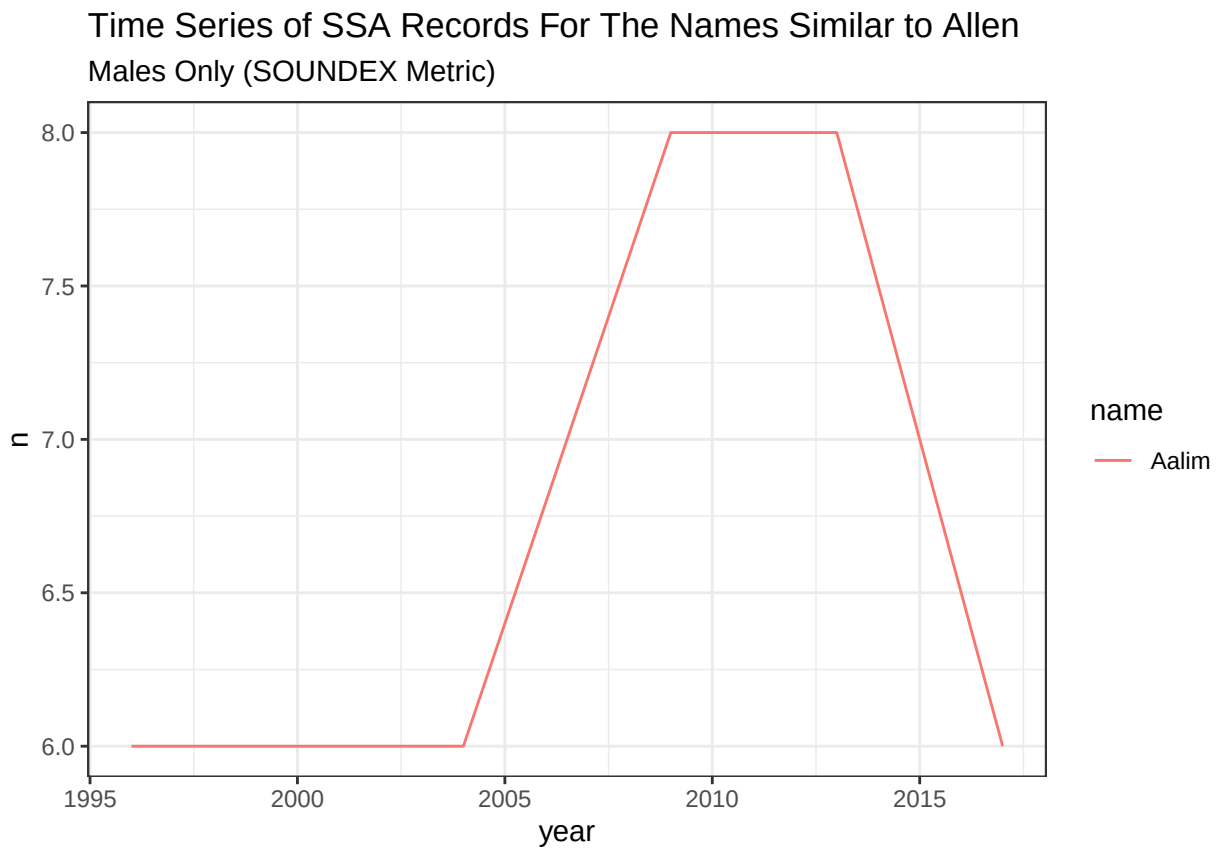

## SOUNDEX Metric

```
distance_name.soundex<- distance_name[order(distance_Allen[,"soundex"])]

common_name.2 <- common_name[common_name$name %in% distance_name.soundex[1:10],]

common_name.2<- common_name.2[common_name.2$sex == "M",]


ggplot(common_name.2) +
  aes(x = year,y = n,group = name,color = name) +
  geom_line() +
  theme_bw() +
  ggtitle("Time Series of SSA Records For The Names Similar to Allen","Males Only (SOUNDEX Metric)")
```




## Task II

Write a few sentences articulating the similarities and differences you notice about each metric. (10 points)

### Sentences

1. The soundex metric only had one common name being "Aalim".
2. The name "Allan" was the second most common for the DL, OSA, and LV metrics.
3. The name "Allen" was the most common among all the metrics besides soundex.
4. The name "Alen" was common among all the metrics besides soundex.
5. The name "Alben" was common among the lv, osa, and dl metrics.
6. The name "Allene" was common among the lcs, qgram, jw, jaccard, and cosine metrics.
7. The name "Alden" was common among lv, osa, and dl metrics.
8. The name "Aneel" was only common with the jaccard metric.
9. The name "Anel" was only common with the cosine metric.
10. The most accurate metrics are lv, osa, and dl both having "Allen" and "Allan" as the most common with my name.