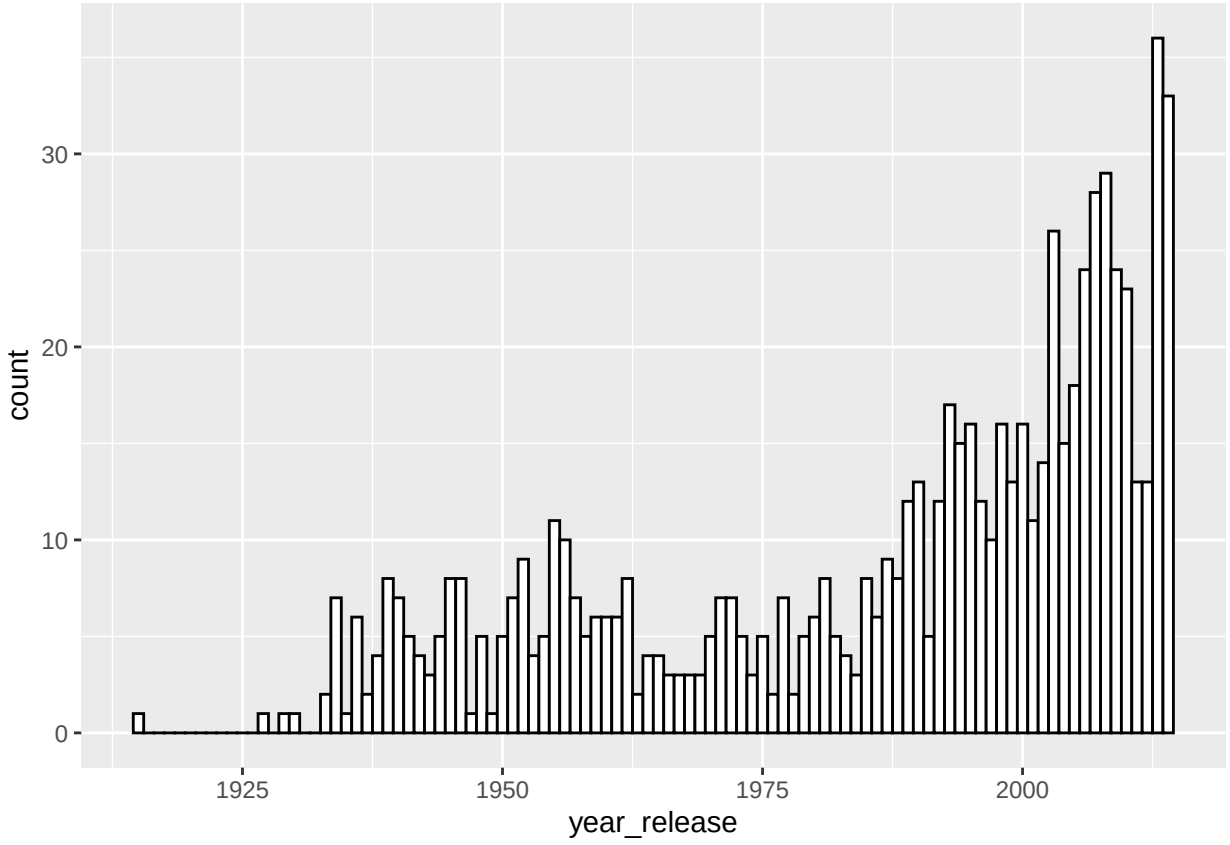
Assignment 2 MATH 208 (Question 1)

Prsa Yadollahi
260869949
MATH 208 - Assignment 2

a)

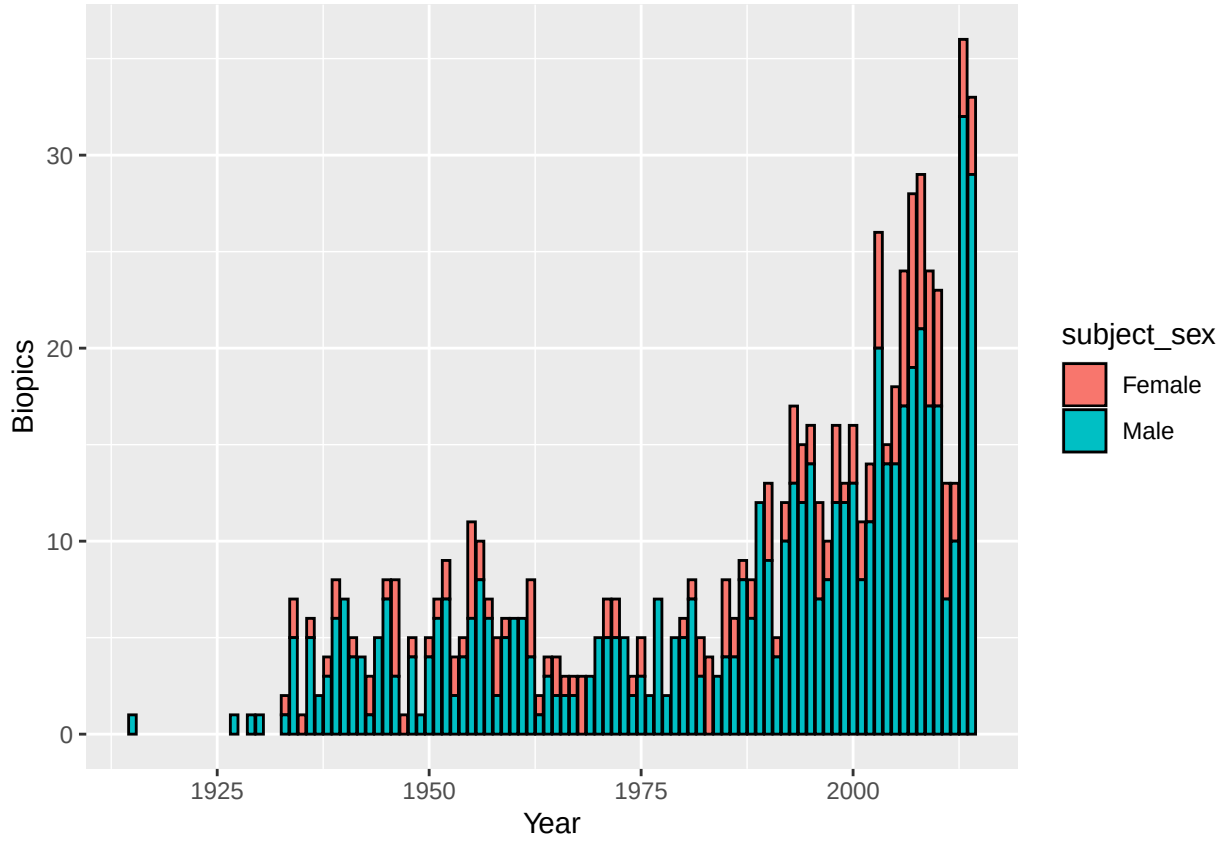
```
ggplot(data=biopics, aes(x = year_release)) +
  geom_histogram(aes(),color="black", fill = "white",
                 position = "identity", bins = 100)
```



From the plot we can deduct that in general the total number of biopics released per year has increased over time based on the timeline above. b)

```
ggplot(data=biopics, aes(x = year_release, group=subject_sex, fill = subject_sex)) +
  geom_bar(bins = 100, col="black")+ylab("Biopics")+xlab("Year")
```

```
## Warning: Ignoring unknown parameters: bins
```

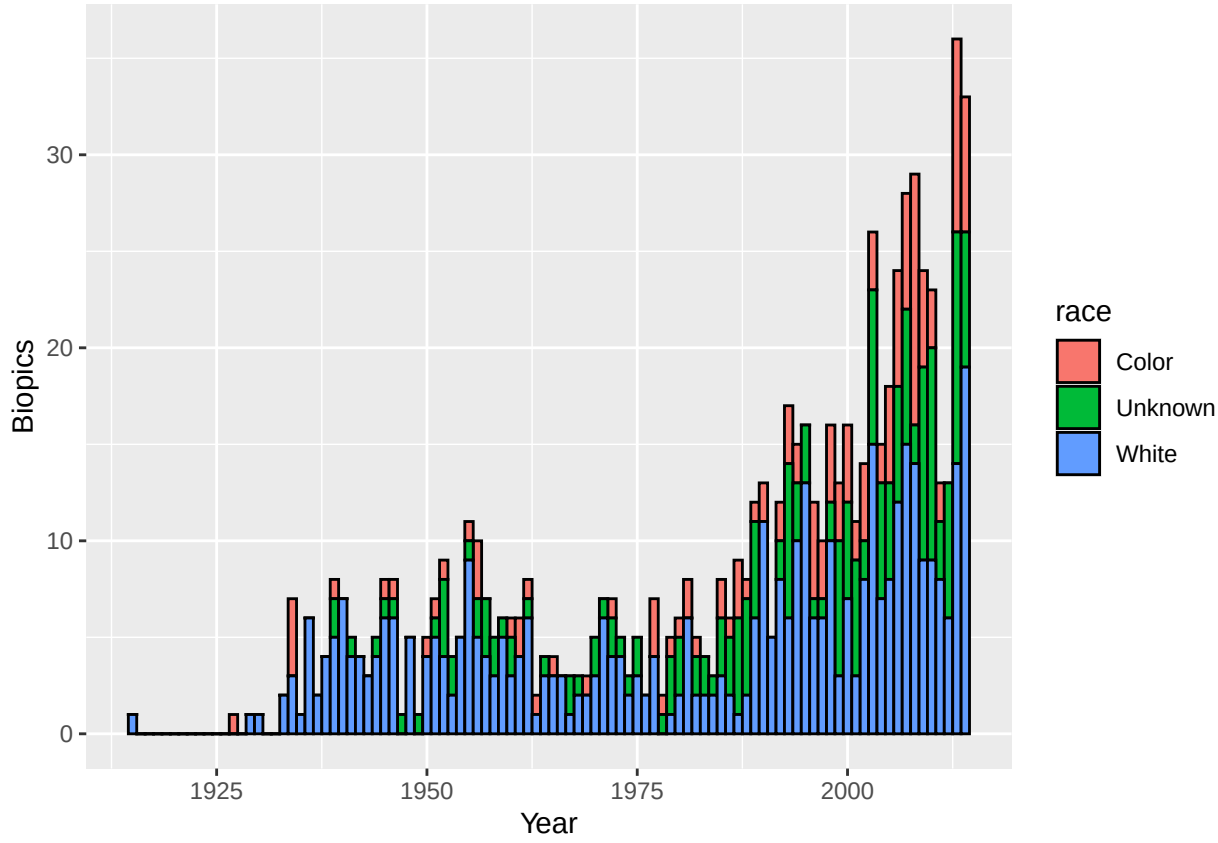


c)

```
biopics_by_race <- biopics %>%
  mutate(race=case_when(
      subject_race=="White" ~ "White",
      race_known=="Known"~ "Color",
      TRUE ~ "Unknown"
       )
    )
biopics_by_race[,"race"]
```

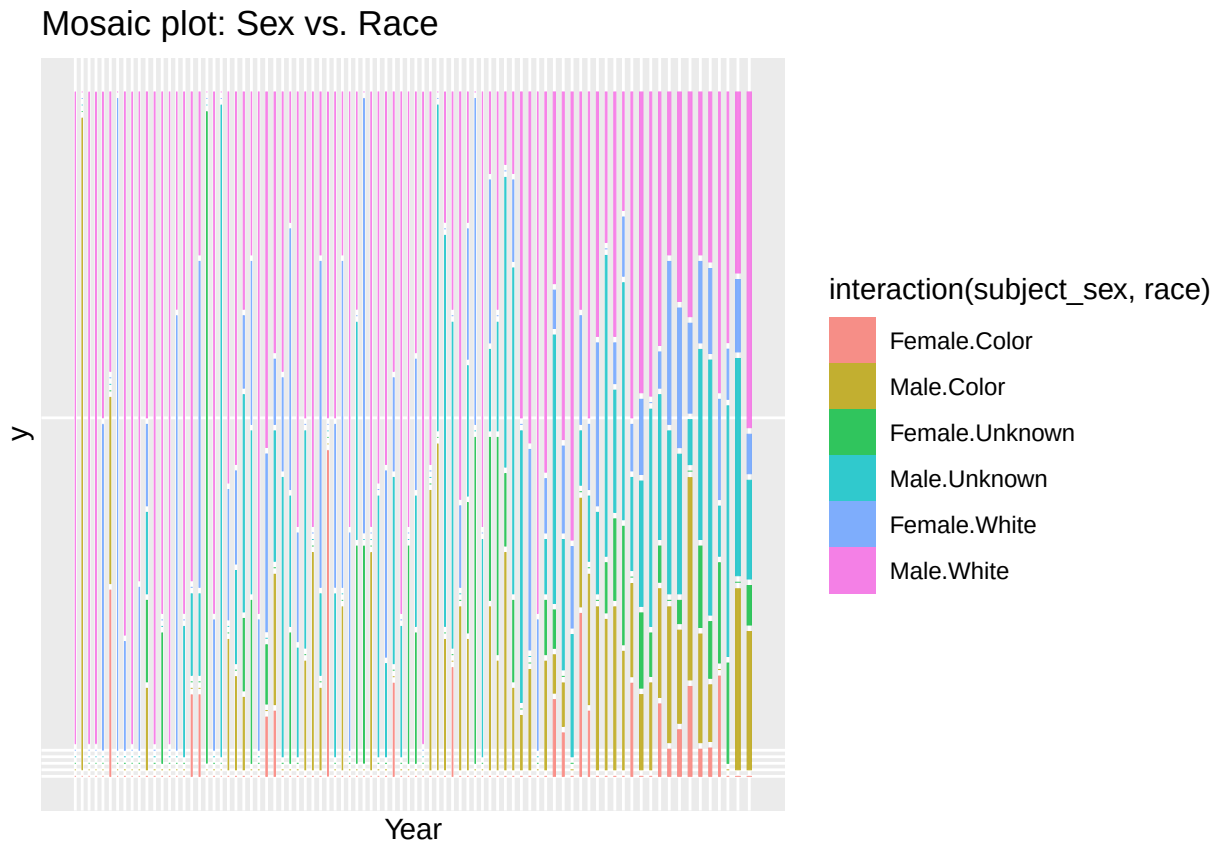
```
## # A tibble: 761 x 1
##    race
##    <chr>
##  1 Unknown
##  2 Color
##  3 Unknown
##  4 White
##  5 Unknown
##  6 Color
##  7 White
##  8 Color
##  9 Unknown
## 10 Unknown
## # ... with 751 more rows
```

```
ggplot(biopics_by_race,aes(x=year_release,group=race,fill=race)) +
  geom_histogram(bins=100,col="black")+ylab("Biopics")+xlab("Year")
```



d)

```
ggplot(data = biopics_by_race) +
 geom_mosaic(aes(x=product(year_release), fill=interaction(subject_sex, race), na.rm=TRUE)) +
 labs(title='Mosaic plot: Sex vs. Race', x="Year")
```



Based on the graph above. Females who are people of color are the most underrepresented

e)

```
new_biopic =
  as_tibble(biopics) %>%
  group_by(year_release) %>%
  mutate(count=n()) %>%
  mutate(
  subject_race =
    ifelse(
      is.na(subject_race),"Unknown",
    ifelse(
      subject_race=="White",subject_race,"Colour"
      )
    )
  ) %>%
  select(
    year_release,
    subject_race,
    subject_sex,
    count
    )

new_biopic_summary = new_biopic %>%
  group_by(
```

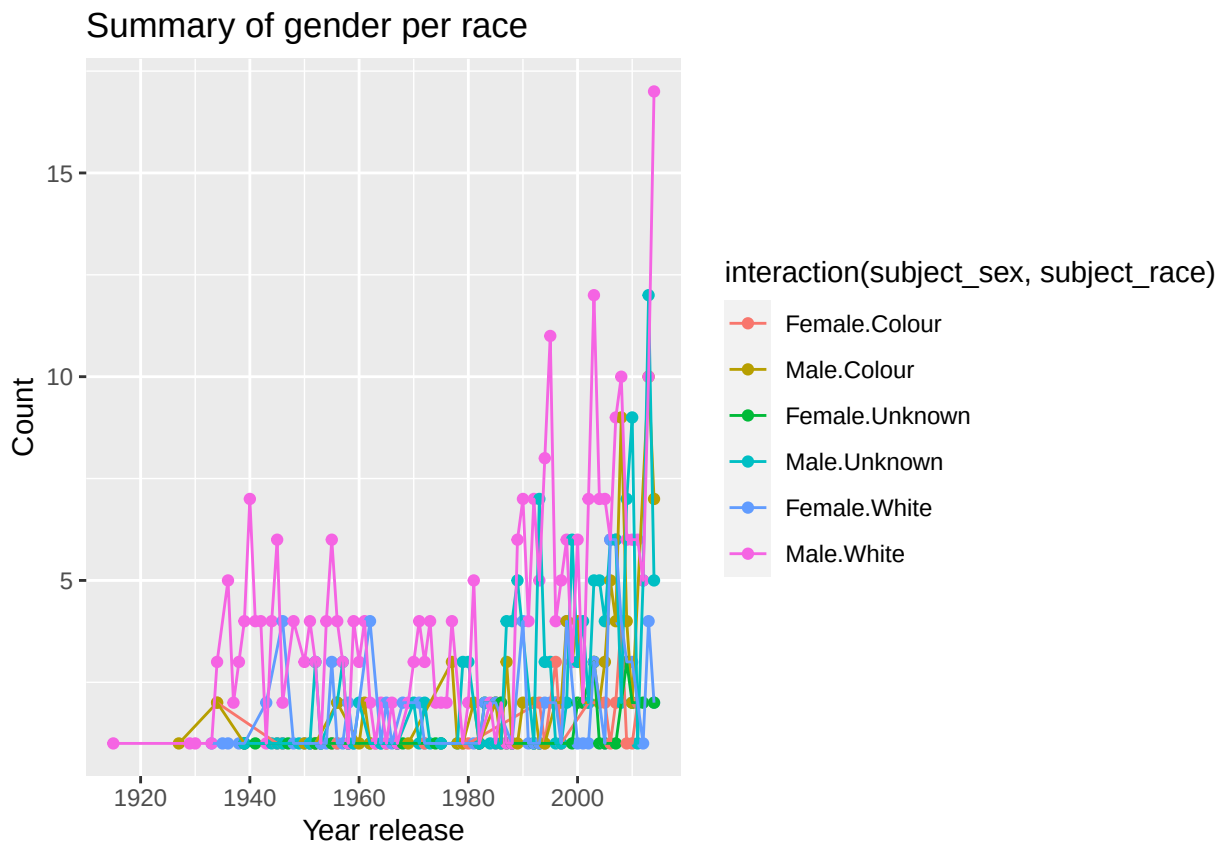
```
    year_release,
    subject_sex,
    subject_race
  ) %>%
  mutate(
    number = n(),
    proportion=number/count
    )

new_biopic_summary = new_biopic_summary[
  order(
    new_biopic_summary$year_release
    ),
  ] %>%
  unique(.) %>%
  select(
    year_release,
    subject_sex,
    subject_race,
    number,
    proportion
  )
new_biopic_summary
```

```
## # A tibble: 281 x 5
## # Groups:   year_release, subject_sex, subject_race [281]
##    year_release subject_sex subject_race number proportion
##           <int> <chr>       <chr>         <int>      <dbl>
##  1         1915 Male        White             1      1
##  2         1927 Male        Colour            1      1
##  3         1929 Male        White             1      1
##  4         1930 Male        White             1      1
##  5         1933 Female      White             1      0.5
##  6         1933 Male        White             1      0.5
##  7         1934 Female      Colour            2      0.286
##  8         1934 Male        White             3      0.429
##  9         1934 Male        Colour            2      0.286
## 10         1935 Female      White             1      1
## # ... with 271 more rows
```

f) i)

```
ggplot(
  data = new_biopic_summary,
  aes(
    x=year_release,
    y=number,
    fill=interaction(
      subject_sex,
      subject_race
      )
    )
  ) +
```

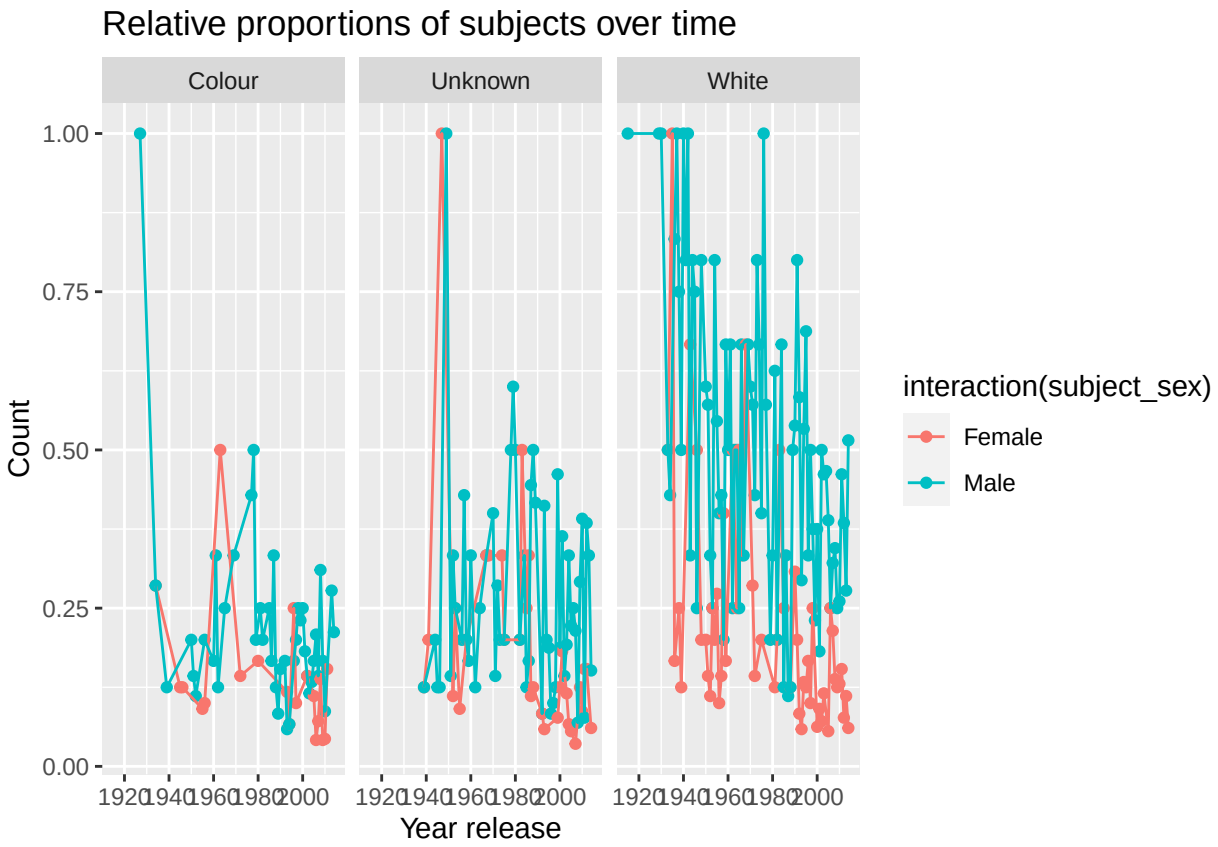
```
  geom_point(
    aes(
      color=interaction(
        subject_sex,
        subject_race
        )
      )
    ) +
  geom_line(
    aes(
      color=interaction(
        subject_sex,
        subject_race
        )
      )
    ) +
  labs(title="Summary of gender per race", x = "Year release", y = "Count")
```



ii)

```
ggplot(
  data = new_biopic_summary,
  aes(
    x=year_release,
    y=proportion,
```

```
    fill=interaction(
      subject_sex
      )
    )
  ) +
  geom_point(
    aes(
      color=interaction(
        subject_sex
        )
      )
    ) +
  geom_line(
    aes(
      color=interaction(
        subject_sex
        )
      )
    ) +
  labs(title="Relative proportions of subjects over time", x = "Year release", y = "Count") +
  facet_wrap(~interaction(subject_race))
```



From the plots drawn above, we can conclude that the problem of imbalance is getting better. We note that they are all converging.