Machine Learning - Day 1

Exploratory analysis

Mariachiara Fortuna

Libraries

```
library(dplyr)
```

```
##
## 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
```

```
library(ggplot2)
library(here)
```

```
## here() starts at /home/chiara/dev/machineLearningClass
```

DATA IMPORT

```
dc <-  file.path(here(), "data", "dc-wikia-data.csv") %>%
  read.csv(na.strings = "")
```

DATA EXPLORATION

Dataset structure

```
str(dc)
```

```
## 'data.frame':    6896 obs. of  13 variables:
##  $ page_id        : int  1422 23387 1458 1659 1576 1448 1486 1451 71760 1380 ...
##  $ name           : Factor w/ 6896 levels "3g4 (New Earth)",..: 593 6007 2487 2996 5278 6772 378 628
##  $ urlslug        : Factor w/ 6896 levels "\\/wiki\\/3g4_(New_Earth)",..: 597 6006 2488 2997 5277 67
##  $ ID             : Factor w/ 3 levels "Identity Unknown",..: 3 3 3 2 3 2 2 3 2 3 ...
##  $ ALIGN          : Factor w/ 4 levels "Bad Characters",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ EYE            : Factor w/ 17 levels "Amber Eyes","Auburn Hair",..: 4 4 5 5 4 4 4 4 4 4 ...
##  $ HAIR           : Factor w/ 17 levels "Black Hair","Blond Hair",..: 1 1 4 17 1 1 2 1 2 2 ...
##  $ SEX            : Factor w/ 4 levels "Female Characters",..: 3 3 3 3 3 1 3 3 1 3 ...
##  $ GSM            : Factor w/ 2 levels "Bisexual Characters",..: NA NA NA NA NA NA NA NA NA NA ...
##  $ ALIVE          : Factor w/ 2 levels "Deceased Characters",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ APPEARANCES    : int  3093 2496 1565 1316 1237 1231 1121 1095 1075 1028 ...
```

```
##  $ FIRST.APPEARANCE: Factor w/ 774 levels "1935, October",..: 15 455 156 461 20 33 39 486 261 129 ..
##  $ YEAR            : int  1939 1986 1959 1987 1940 1941 1941 1989 1969 1956 ...
```

Summary information

```
summary(dc)
```

```
##     page_id                                name
##  Min.   :  1380   3g4 (New Earth)          :   1
##  1st Qu.: 44106   500-ZQ (New Earth)       :   1
##  Median :141267   Aa (New Earth)           :   1
##  Mean   :147441   Aarden (New Earth)       :   1
##  3rd Qu.:213203   Aaron Babcock (New Earth):   1
##  Max.   :404010   Aaron Cash (New Earth)   :   1
##                   (Other)                  :6890
##                                     urlslug                      ID
##  \\/wiki\\/3g4_(New_Earth)                :   1   Identity Unknown:   9
##  \\/wiki\\/500-ZQ_(New_Earth)             :   1   Public Identity :2466
##  \\/wiki\\/A%27Hwiirdh-Paan%27A_(New_Earth):   1   Secret Identity :2408
##  \\/wiki\\/A%27monn_A%27mokk_(New_Earth)  :   1   NA's            :2013
##  \\/wiki\\/A%27morr_(New_Earth)           :   1
##  \\/wiki\\/Aa_(New_Earth)                 :   1
##  (Other)                                  :6890
##                  ALIGN             EYE               HAIR
##  Bad Characters    :2895   Blue Eyes :1102   Black Hair:1574
##  Good Characters   :2832   Brown Eyes: 879   Brown Hair:1148
##  Neutral Characters: 565   Black Eyes: 412   Blond Hair: 744
##  Reformed Criminals:   3   Green Eyes: 291   Red Hair  : 461
##  NA's              : 601   Red Eyes  : 208   White Hair: 346
##                            (Other)   : 376   (Other)   : 349
##                            NA's      :3628   NA's      :2274
##                      SEX                         GSM
##  Female Characters     :1967   Bisexual Characters  :  10
##  Genderless Characters :  20   Homosexual Characters:  54
##  Male Characters       :4783   NA's                 :6832
##  Transgender Characters:   1
##  NA's                  : 125
##
##
##                  ALIVE       APPEARANCES            FIRST.APPEARANCE
##  Deceased Characters:1693   Min.   :   1.00   2010, December:  78
##  Living Characters  :5200   1st Qu.:   2.00   2006, June    :  48
##  NA's               :   3   Median :   6.00   1989, January :  45
##                             Mean   :  23.63   2009, October :  44
##                             3rd Qu.:  15.00   1988, March   :  40
##                             Max.   :3093.00   (Other)       :6572
##                             NA's   :355       NA's          :  69
##       YEAR
##  Min.   :1935
##  1st Qu.:1983
##  Median :1992
##  Mean   :1990
##  3rd Qu.:2003
```

```
##  Max.   :2013
##  NA's   :69
```

Simple stats

Range

```
min(dc$YEAR)
```

```
## [1] NA
```

```
min(dc$YEAR, na.rm = T)
```

```
## [1] 1935
```

```
max(dc$YEAR, na.rm = T)
```

```
## [1] 2013
```

```
range(dc$YEAR, na.rm = T)
```

```
## [1] 1935 2013
```

Mean and Sd

```
mean(dc$APPEARANCES, na.rm = T)
```

```
## [1] 23.62513
```

```
sd(dc$APPEARANCES, na.rm = T)
```

```
## [1] 87.37851
```

Frequency tables

Univariate

```
table(dc$EYE)
```

```
##
##         Amber Eyes        Auburn Hair         Black Eyes
##                  5                  7                412
##          Blue Eyes         Brown Eyes          Gold Eyes
##               1102                879                  9
##         Green Eyes          Grey Eyes         Hazel Eyes
##                291                 40                 23
##        Orange Eyes Photocellular Eyes          Pink Eyes
##                 10                 48                  6
##        Purple Eyes           Red Eyes        Violet Eyes
##                 14                208                 12
##         White Eyes        Yellow Eyes
##                116                 86
```

Bivariate

```
ID_SEX_freq  <- table(dc$ID, dc$SEX)

ID_SEX_freq
```

```
##
##                    Female Characters Genderless Characters Male Characters
##   Identity Unknown                 0                      0               9
##   Public Identity                765                     11            1662
##   Secret Identity                625                      5            1751
##
##                    Transgender Characters
##   Identity Unknown                      0
##   Public Identity                       0
##   Secret Identity                       0
```

```
margin.table(ID_SEX_freq, margin = 1)
```

```
##
## Identity Unknown  Public Identity  Secret Identity
##                9             2438             2381
```

```
prop.table(ID_SEX_freq)
```

```
##
##                    Female Characters Genderless Characters Male Characters
##   Identity Unknown       0.000000000           0.000000000     0.001864126
##   Public Identity        0.158450704           0.002278376     0.344241922
##   Secret Identity        0.129453190           0.001035626     0.362676056
##
##                    Transgender Characters
##   Identity Unknown            0.000000000
##   Public Identity             0.000000000
##   Secret Identity             0.000000000
```

```
prop.table(ID_SEX_freq, margin = 2)
```

```
##
##                    Female Characters Genderless Characters Male Characters
##   Identity Unknown       0.000000000           0.000000000     0.002630041
##   Public Identity        0.550359712           0.687500000     0.485680888
##   Secret Identity        0.449640288           0.312500000     0.511689071
##
##                    Transgender Characters
##   Identity Unknown
##   Public Identity
##   Secret Identity
```

DATA MANIPULATION

Create a new column: Mutate

```
max_year <- max(dc$YEAR, na.rm =T)

dc <-  dc %>%
  mutate(active_years = max_year - YEAR) %>%
  arrange(desc(active_years))
```

Recode factor

```
dc <- dc %>%
  mutate(
    sex = case_when(SEX == "Female Characters" ~ "F",
                    SEX == "Male Characters" ~ "M",
                    TRUE ~ "other")
    )
```

Subsetting data (rows and columns)

```
the_bad <- dc %>%
  filter(ALIGN == "Bad Characters") %>%
  select(name, ID, SEX, ALIVE, YEAR, APPEARANCES, FIRST.APPEARANCE, active_years)
```

More complex summary stats

```
the_bad %>%
  group_by(SEX) %>%
  summarize(n = n(),
            avg_appearance = mean(APPEARANCES, na.rm = T),
            avg_active_year = mean(active_years, na.rm = T))
```

```
## # A tibble: 5 x 4
##                       SEX     n avg_appearance avg_active_year
##                    <fctr> <int>          <dbl>           <dbl>
## 1      Female Characters   597       9.753521        18.18624
## 2  Genderless Characters    11       8.300000        18.09091
## 3        Male Characters  2223      10.911398        22.24366
## 4 Transgender Characters     1       4.000000         4.00000
## 5                    <NA>    63       7.175439        19.50794
```

Percentage

```
dc %>%
  group_by(SEX, ALIVE) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n))
```

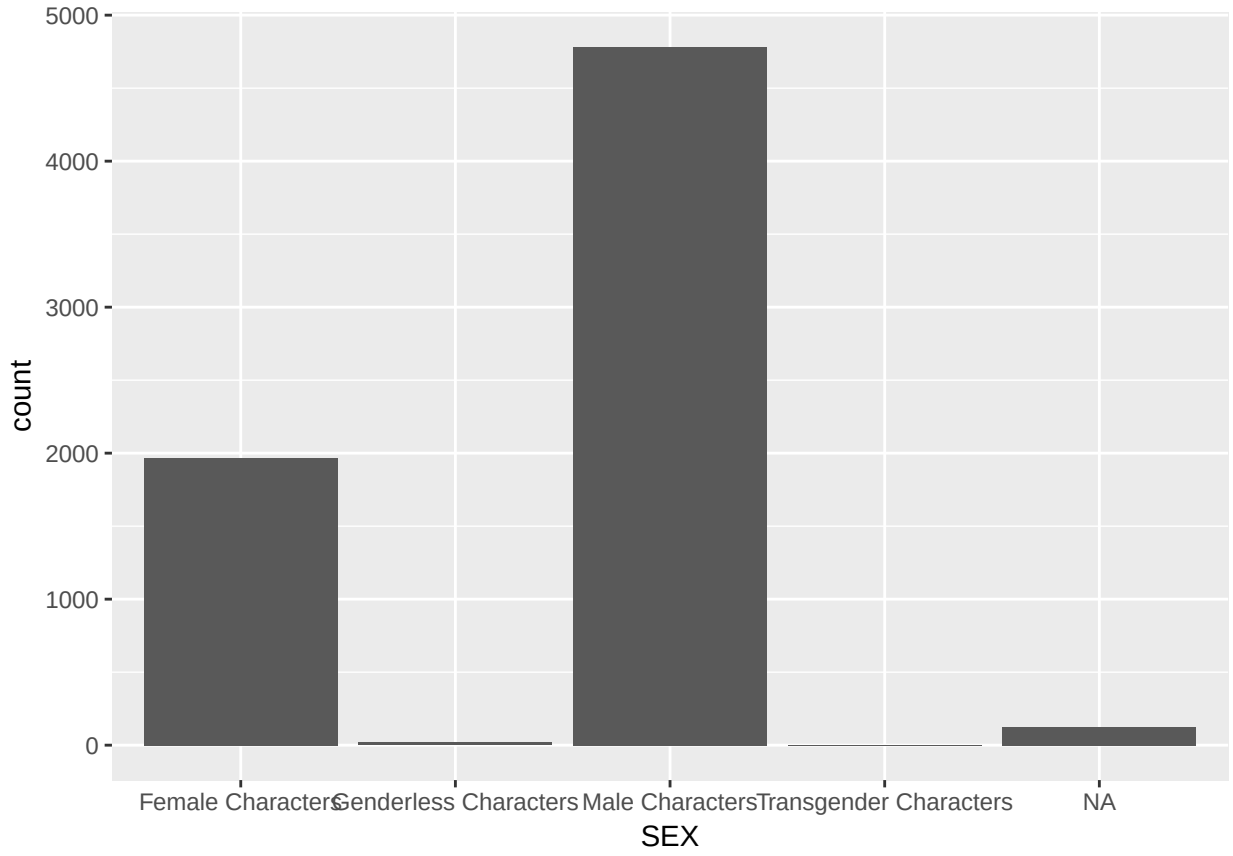
```
## # A tibble: 12 x 4
## # Groups:   SEX [5]
##                       SEX               ALIVE     n         freq
##                    <fctr>              <fctr> <int>        <dbl>
##  1      Female Characters Deceased Characters   392 0.1992882562
```

```
##  2      Female Characters   Living Characters  1574 0.8002033554
##  3      Female Characters                <NA>     1 0.0005083884
##  4  Genderless Characters Deceased Characters     5 0.2500000000
##  5  Genderless Characters   Living Characters    15 0.7500000000
##  6        Male Characters Deceased Characters  1271 0.2657328037
##  7        Male Characters   Living Characters  3511 0.7340581225
##  8        Male Characters                <NA>     1 0.0002090738
##  9 Transgender Characters Deceased Characters     1 1.0000000000
## 10                  <NA> Deceased Characters    24 0.1920000000
## 11                  <NA>   Living Characters   100 0.8000000000
## 12                  <NA>                <NA>     1 0.0080000000
```

DATA VISUALIZATION

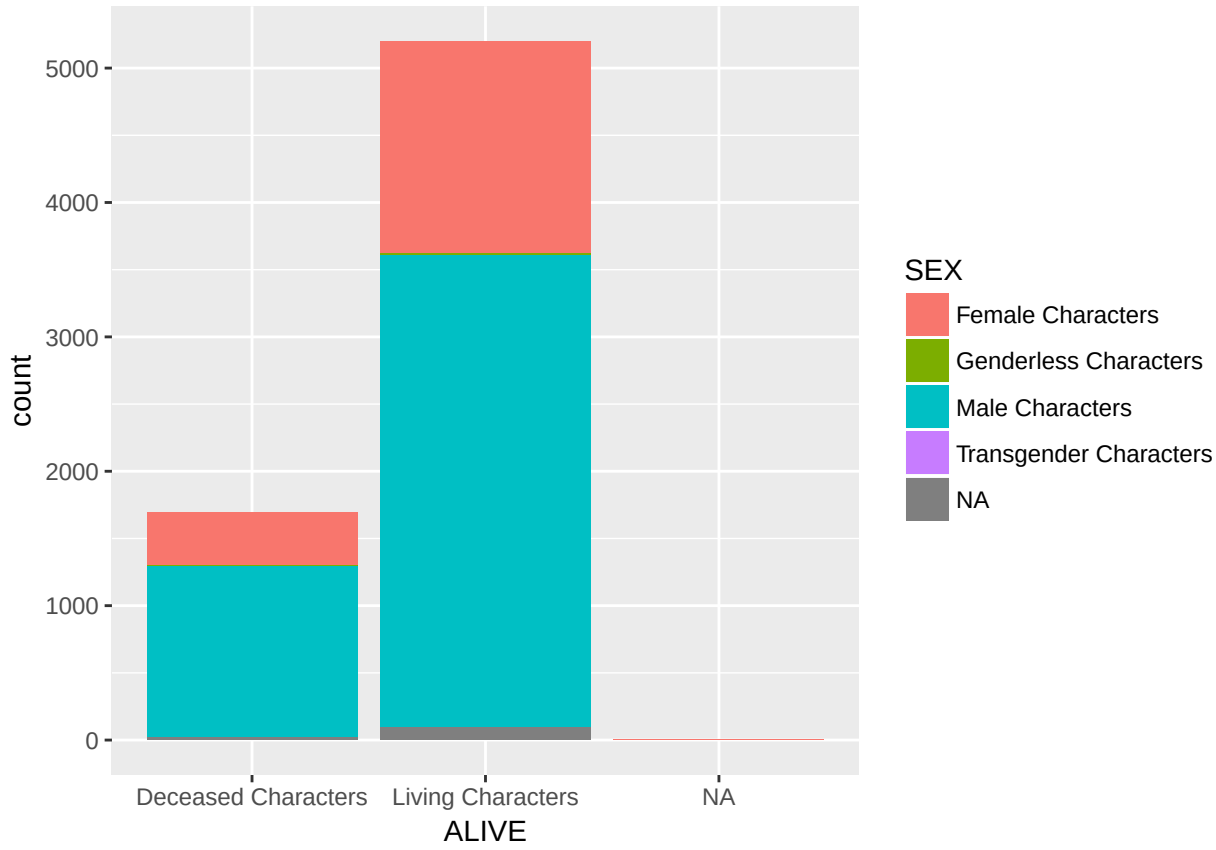
Barplot

```
ggplot(data=dc, aes(x = SEX)) +
  geom_bar()
```



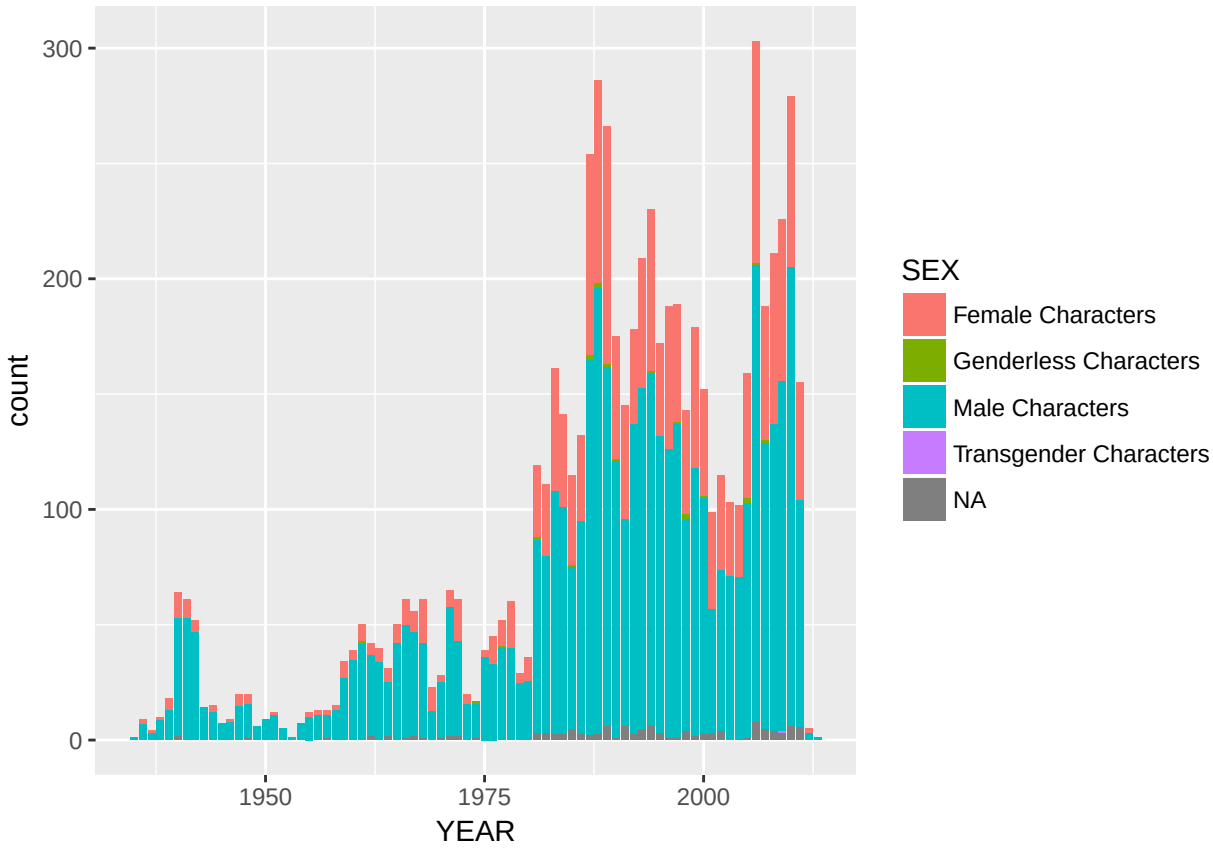
Barplot with grouping variable

```
ggplot(data=dc, aes(x = ALIVE, fill = SEX)) +
  geom_bar()
```




```
ggplot(data=dc, aes(x = YEAR, fill = SEX)) +
  geom_bar()
```

```
## Warning: Removed 69 rows containing non-finite values (stat_count).
```



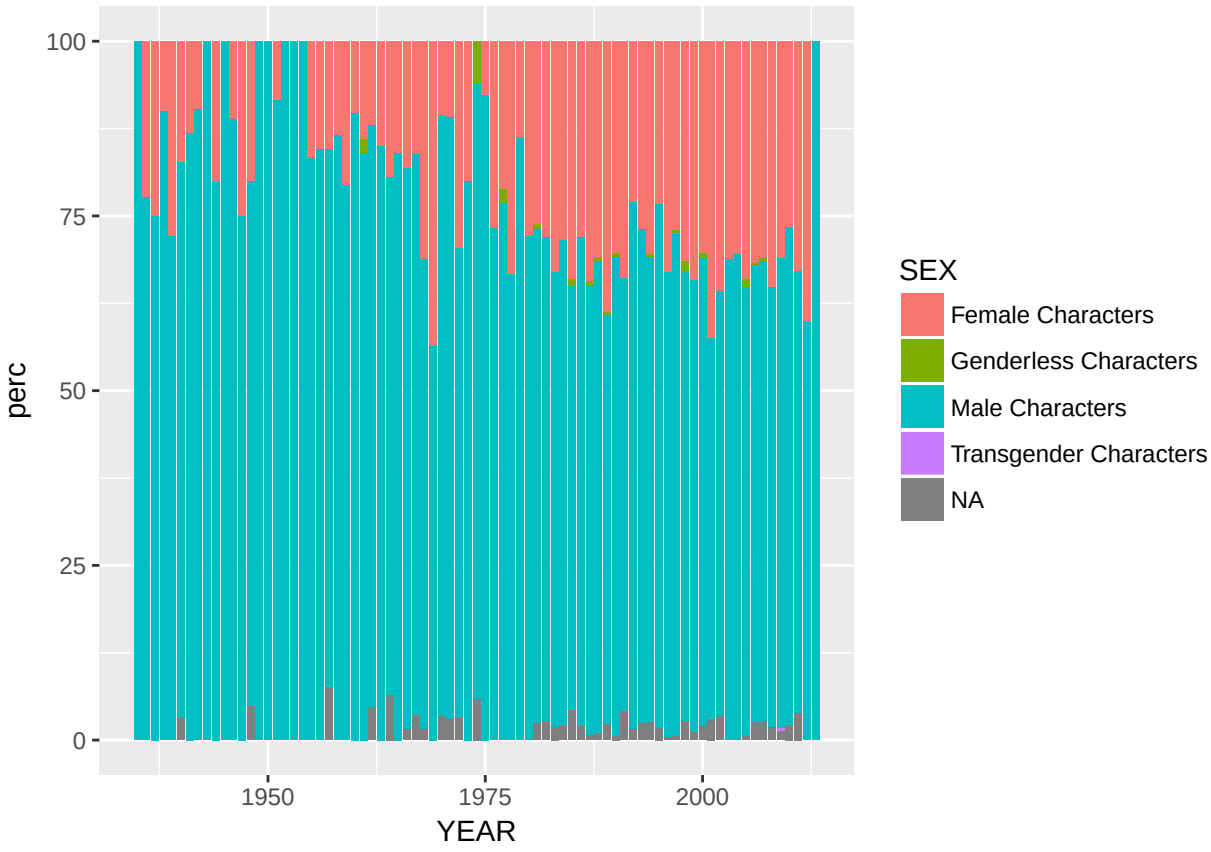
Percent values

```
YEAR_SEX_tab <- dc %>%
  group_by(YEAR, SEX) %>%
  summarize(n = n())

margin <- dc %>%
  group_by(YEAR) %>%
  summarize(tot = n())

perc_ys <- YEAR_SEX_tab %>%
  inner_join(margin) %>%
  mutate(perc = (n/tot)*100)
```

```
## Joining, by = "YEAR"
```

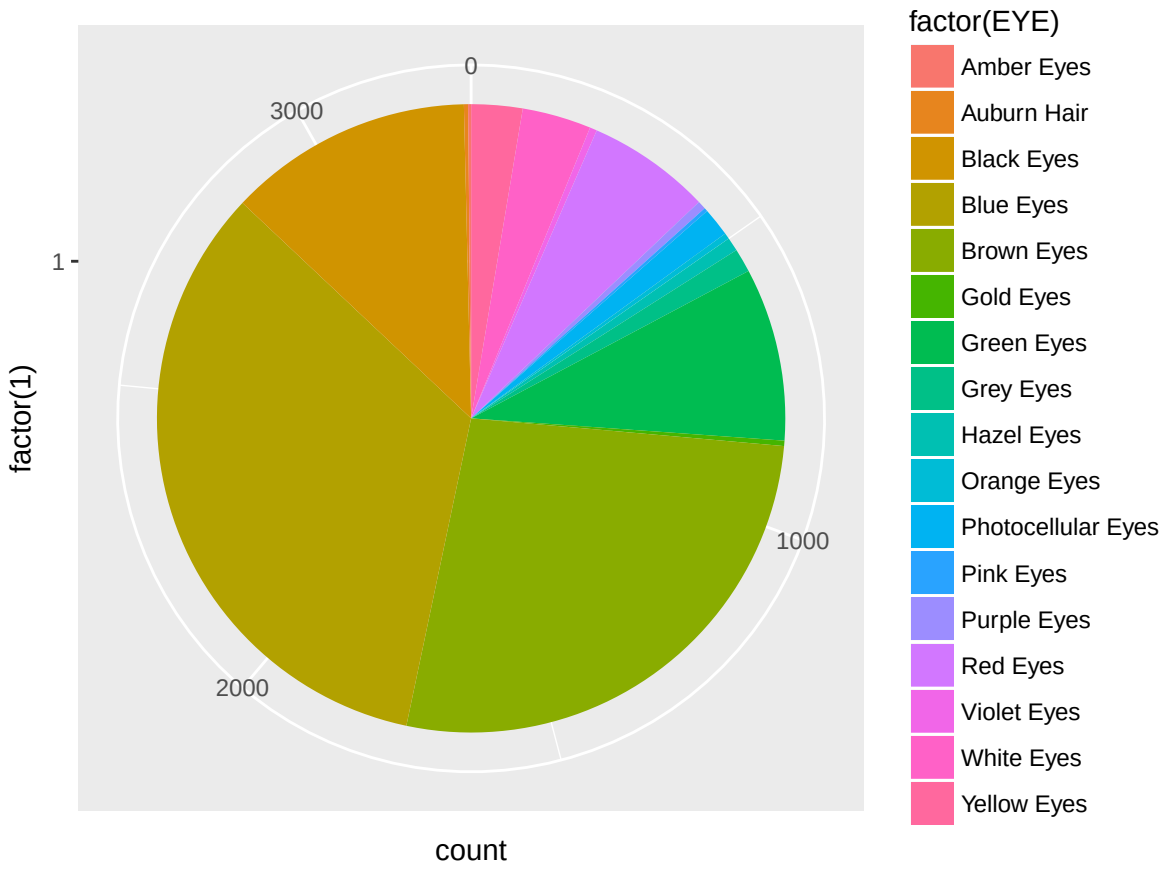
```
ggplot(data=perc_ys, aes(x = YEAR, y = perc, fill = SEX)) +
  geom_bar(stat = "identity")
```

```
## Warning: Removed 3 rows containing missing values (position_stack).
```



Pie chart

```
eye_tab <- dc %>%
  filter(EYE != "") %>%
  group_by(EYE) %>%
  summarize(count = n())

 ggplot(data=eye_tab,
        aes(x=factor(1), y = count, fill = factor(EYE))) +
   geom_bar(width = 1, stat = "identity") +
   coord_polar(theta = "y")
```



Matrix plot

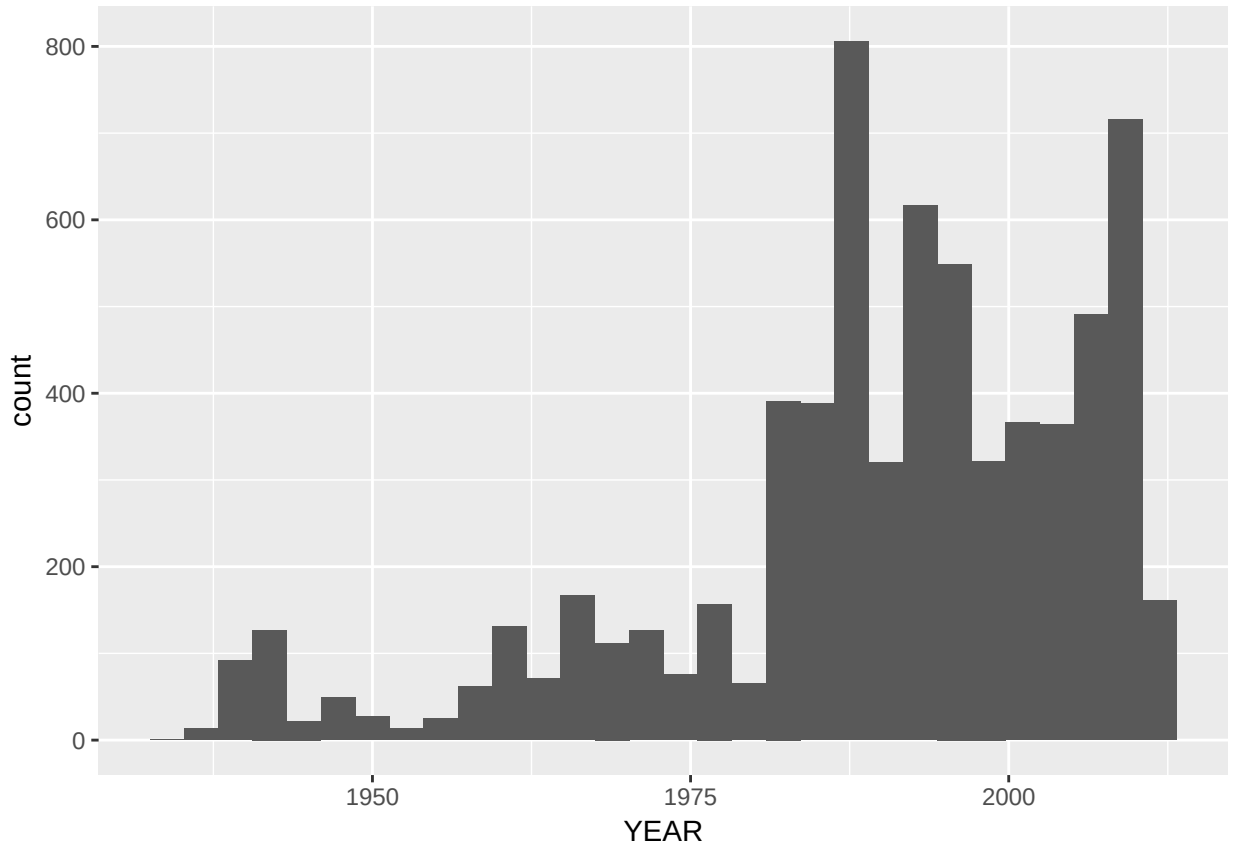
```
ggplot(data = dc, aes (x = EYE, y = HAIR)) +
  geom_count() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Histogram

```
ggplot(data = dc, aes(x = YEAR)) +
  geom_histogram()
```

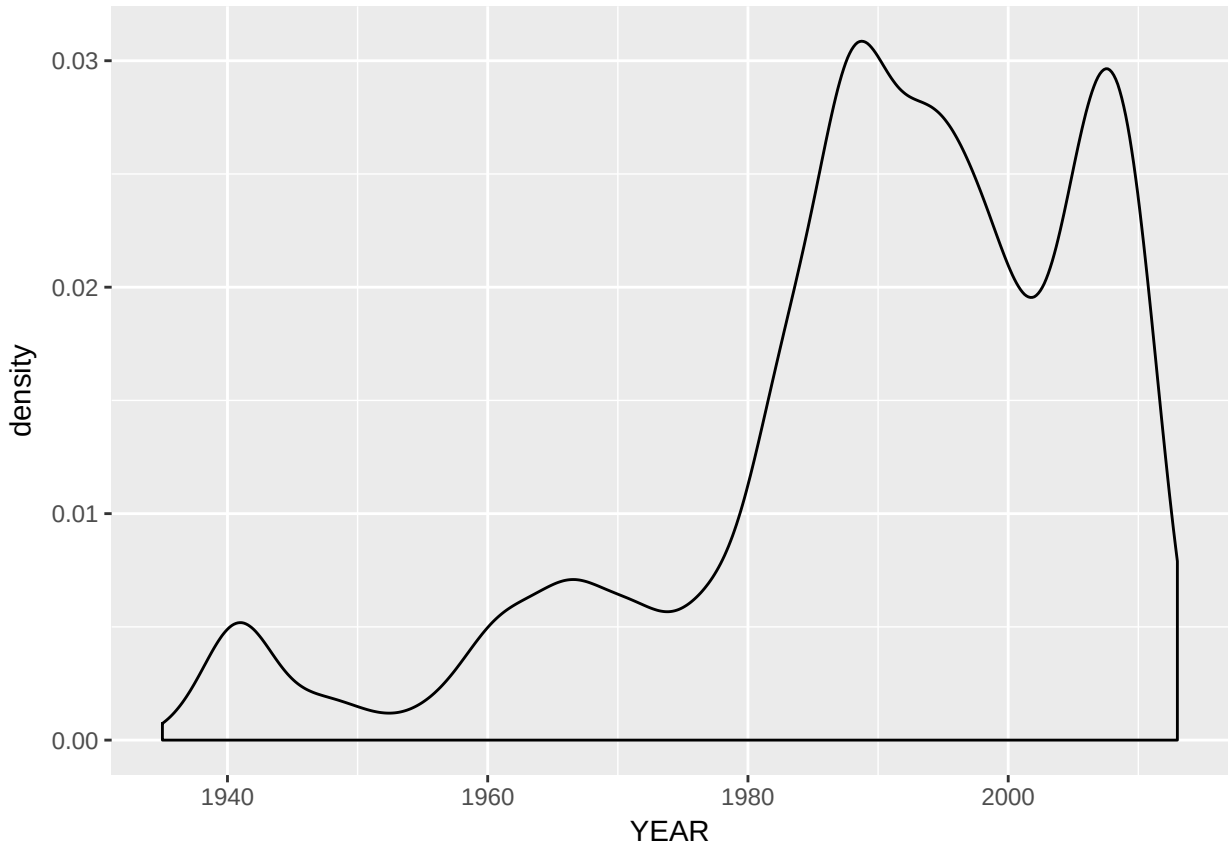
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 69 rows containing non-finite values (stat_bin).
```




```
ggplot(data = dc, aes(x = YEAR)) +
  geom_density()
```

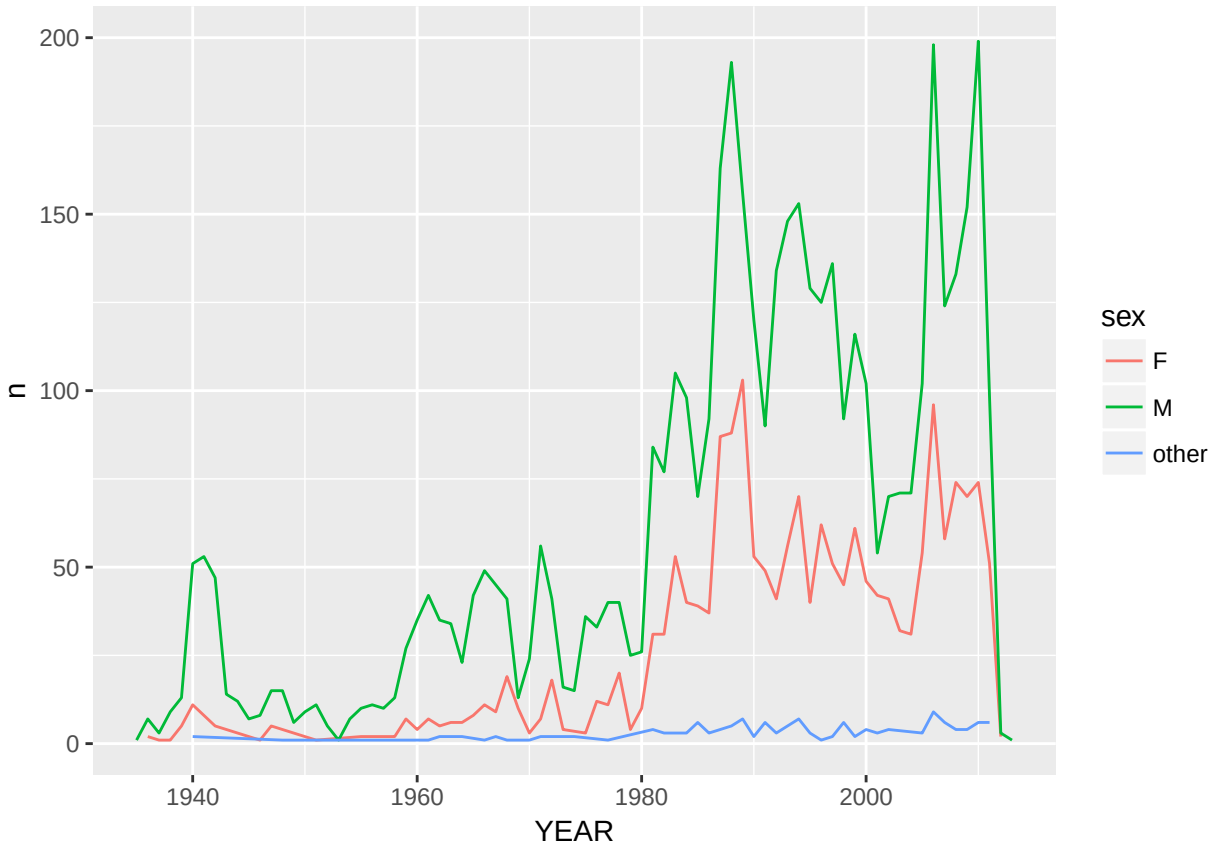
```
## Warning: Removed 69 rows containing non-finite values (stat_density).
```



Line plot

```
year_sex <- dc %>%
  group_by(YEAR, sex) %>%
  summarise(n = n())

ggplot(data = year_sex, aes(x = YEAR, y = n, color = sex)) +
  geom_line()
```

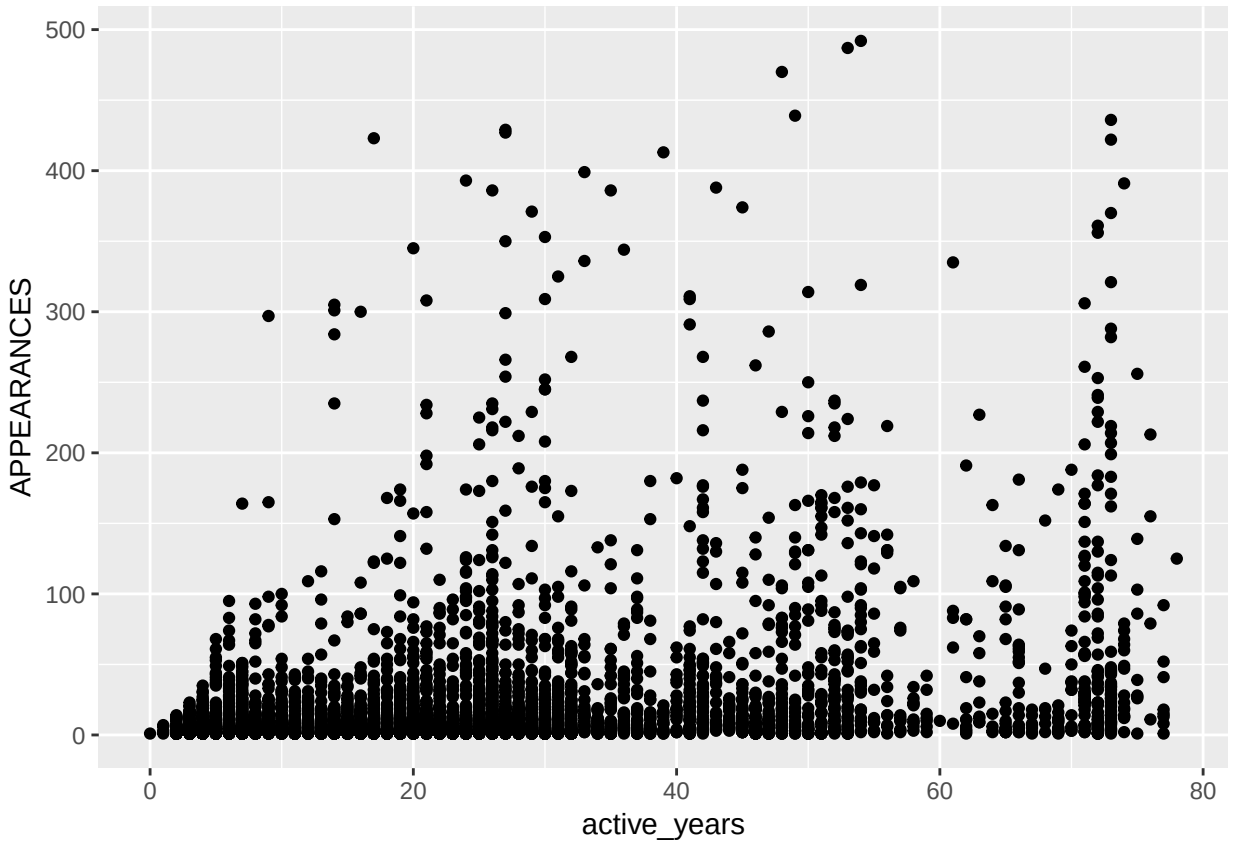
```
## Warning: Removed 3 rows containing missing values (geom_path).
```



Scatterplot

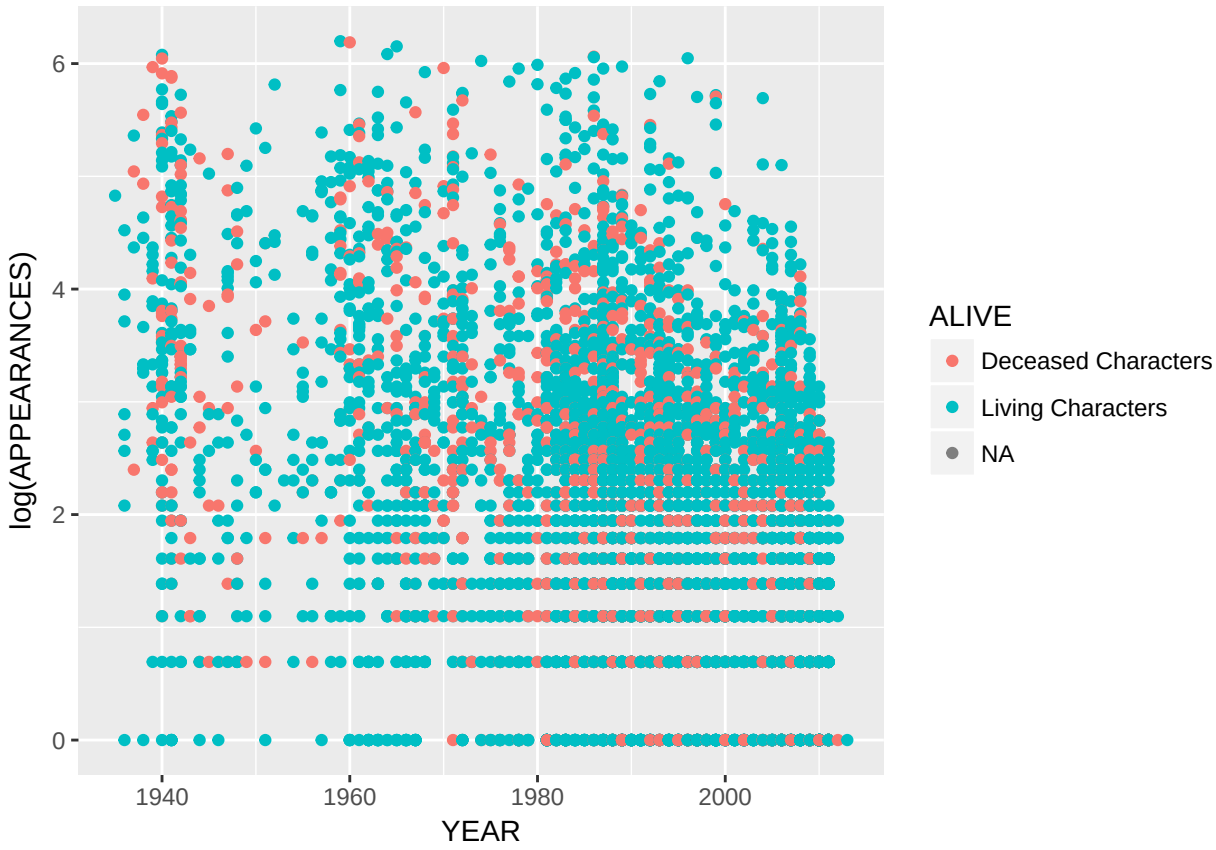
```
ggplot(data = dc %>% filter(APPEARANCES<500),
       aes(x = active_years, y = APPEARANCES)) +
  geom_point()
```

```
## Warning: Removed 60 rows containing missing values (geom_point).
```



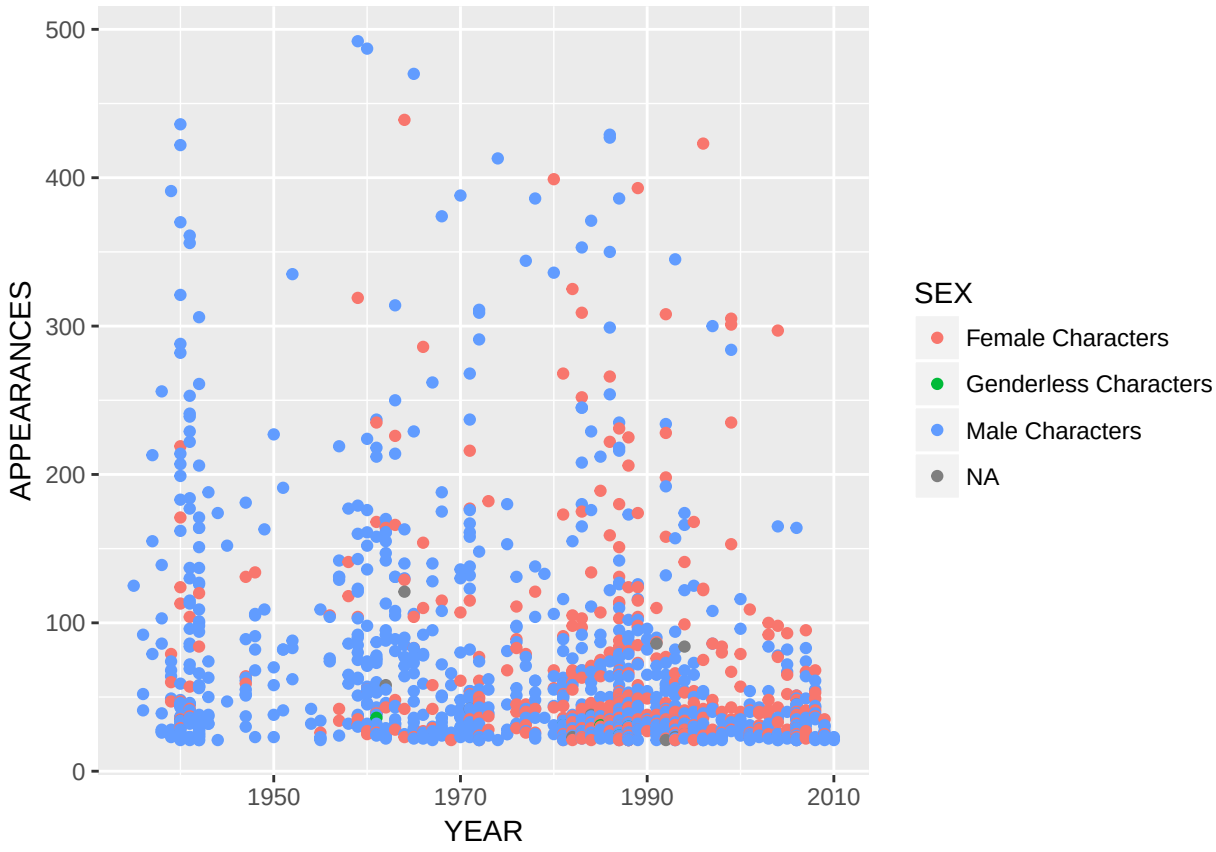

```
ggplot(data = dc %>% filter(APPEARANCES<500),
       aes(x = YEAR, y = log(APPEARANCES), color = ALIVE)) +
  geom_point()
```

```
## Warning: Removed 60 rows containing missing values (geom_point).
```



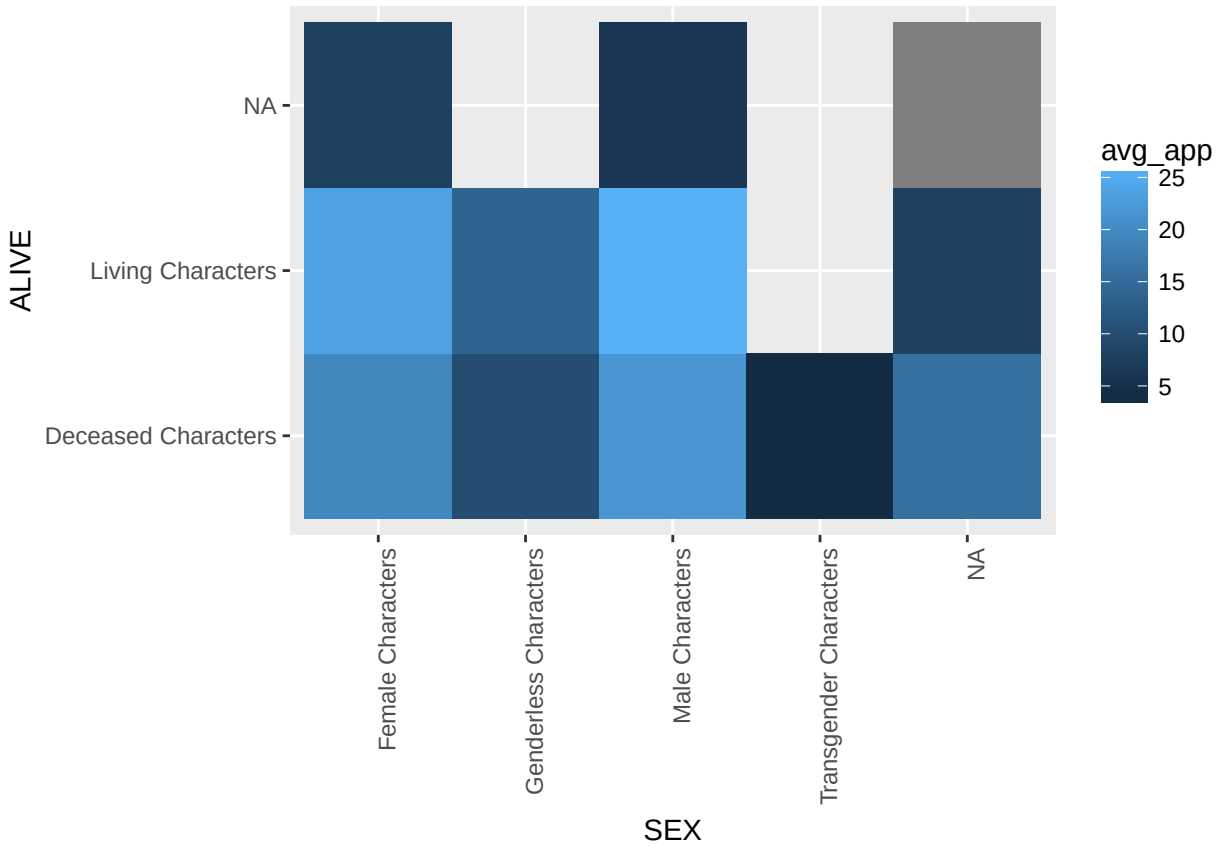

```
ggplot(data = dc %>% filter(APPEARANCES<500 & APPEARANCES>20),
       aes(x = YEAR, y = APPEARANCES, color = SEX)) +
  geom_point()
```

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



Carpet plot

```
sex_alive <- dc %>%
  group_by(SEX, ALIVE) %>%
  summarize(avg_app = mean(APPEARANCES, na.rm = T))

ggplot(data = sex_alive, aes(x = SEX, y = ALIVE)) +
  geom_tile(aes(fill = avg_app)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```


ALIVE
NA
Living Characters
Deceased Characters
Female Characters
Genderless Characters
Male Characters
Transgender Characters
NA
SEX
avg_app
25
20
15
10
5