Stats 15 - Homework 2

Haojie Liu

Homework text and questions: Copyright Miles Chen. Do not post, share, or distribute without permission.

Academic Integrity Statement

By modifying this statement, I, Haojie Liu, declare that all of the work in this assignment is my own original work. At no time did I look at the code of other students nor did I search for code solutions online. I understand that plagiarism on any single part of this assignment will result in a 0 for the entire assignment and that I will be referred to the dean of students.

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Part 1: Textbook, Section 4.4, Exercise 1

The subset of the `babynames` data has been recreated for you here:

```
year <- c(2003, 1999, 2010, 1989, 1989, 1928, 1981, 1981, 1920, 1941)
sex <- c("M", "F", "F", "F", "F", "M", "F", "F", "F", "M")
name <- c("Bilal","Terria","Naziyah","Shawana","Jessi","Tillman","Leslee",
          "Sherise","Marquerite","Lorraine")
n <- as.integer(c(146, 23, 45, 41, 210, 43, 83, 27, 26, 24))
prop <- c(0.0000695, 0.0000118, 0.0000230, 0.0000206, 0.000105 , 0.0000377,
          0.0000464, 0.0000151, 0.0000209, 0.0000191)
baby_subset <- tibble(year, sex, name, n, prop)
print(baby_subset)
```

```
## # A tibble: 10 x 5
##     year sex   name           n      prop
##    <dbl> <chr> <chr>      <int>     <dbl>
##  1  2003 M     Bilal        146 0.0000695
##  2  1999 F     Terria        23 0.0000118
##  3  2010 F     Naziyah       45 0.000023
##  4  1989 F     Shawana       41 0.0000206
##  5  1989 F     Jessi        210 0.000105
##  6  1928 M     Tillman       43 0.0000377
##  7  1981 F     Leslee        83 0.0000464
##  8  1981 F     Sherise       27 0.0000151
##  9  1920 F     Marquerite    26 0.0000209
## 10  1941 M     Lorraine      24 0.0000191
```

Execute the appropriate `dplyr` commands that will recreate the output you see in the textbook.

a.

```
baby_subset %>%
  select(year,sex,name, n) %>%
  filter(n %in% 41:83)
```

```
## # A tibble: 4 x 4
##    year sex   name        n
##   <dbl> <chr> <chr>   <int>
## 1  2010 F     Naziyah    45
## 2  1989 F     Shawana    41
## 3  1928 M     Tillman    43
## 4  1981 F     Leslee     83
```

b.

```
baby_subset %>%
  filter(nchar(name) == 6)
```

```
## # A tibble: 2 x 5
##    year sex   name       n      prop
##   <dbl> <chr> <chr>  <int>     <dbl>
## 1  1999 F     Terria    23 0.0000118
## 2  1981 F     Leslee    83 0.0000464
```

c.

```
baby_subset %>%
  filter(year == 1989) %>%
  mutate(total=n/prop)
```

```
## # A tibble: 2 x 6
##    year sex   name        n      prop    total
##   <dbl> <chr> <chr>   <int>     <dbl>    <dbl>
## 1  1989 F     Shawana    41 0.0000206 1990291.
## 2  1989 F     Jessi     210 0.000105  2000000
```

d.

```
baby_subset %>%
  group_by(year) %>%
  summarize(
    total = sum(n)
  ) %>%
  select(year, total) %>%
  arrange(year)
```

```
## # A tibble: 8 x 2
##    year total
##   <dbl> <int>
## 1  1920    26
## 2  1928    43
## 3  1941    24
## 4  1981   110
## 5  1989   251
## 6  1999    23
## 7  2003   146
## 8  2010    45
```

Part 2: Textbook, Section 4.4, Exercise 3

The problem with the pipeline is that there will be no column called “am” after summarize the avg_mpg by mean(mpg). Therefore, filter() should placed before summarize() to filter the data.

Part 3: Textbook, Section 4.4, Exercise 4

```
library(Lahman)
Teams %>%
  tail()
```

```
##      yearID lgID teamID franchID divID Rank   G Ghome   W   L DivWin WCWin
## 2980   2021   NL    SFN      SFG     W    1 162    81 107  55      Y     N
## 2981   2021   NL    SLN      STL     C    2 162    81  90  72      N     Y
## 2982   2021   AL    TBA      TBD     E    1 162    81 100  62      Y     N
## 2983   2021   AL    TEX      TEX     W    5 162    81  60 102      N     N
## 2984   2021   AL    TOR      TOR     E    4 162    80  91  71      N     N
## 2985   2021   NL    WAS      WSN     E    5 162    81  65  97      N     N
##      LgWin WSWin   R   AB    H X2B X3B  HR  BB   SO  SB CS HBP SF  RA  ER  ERA
## 2980     N     N 804 5462 1360 271  25 241 602 1461  66 14  64 30 594 524 3.24
## 2981     N     N 706 5351 1303 261  22 198 478 1341  89 22  86 44 672 626 3.98
## 2982     N     N 857 5507 1336 288  36 222 585 1542  88 42  72 41 651 593 3.67
## 2983     N     N 625 5405 1254 225  24 167 433 1381 106 29  58 31 815 758 4.79
## 2984     N     N 846 5476 1455 285  13 262 496 1218  81 20  51 35 663 610 3.91
## 2985     N     N 724 5385 1388 272  20 182 573 1303  56 26  84 31 820 743 4.80
##      CG SHO SV IPouts   HA HRA BBA  SOA  E  DP    FP                 name
## 2980  2  18 56   4365 1254 151 416 1425 80 122 0.986 San Francisco Giants
## 2981  3  15 50   4251 1234 152 608 1225 84 137 0.986  St. Louis Cardinals
## 2982  1  13 42   4367 1264 184 436 1478 80 130 0.986       Tampa Bay Rays
## 2983  0   3 31   4273 1402 232 513 1239 83 146 0.986        Texas Rangers
## 2984  1  14 34   4216 1257 209 473 1468 90 122 0.984    Toronto Blue Jays
## 2985  1   8 36   4183 1364 247 548 1346 96 116 0.983 Washington Nationals
##                   park attendance BPF PPF teamIDBR teamIDlahman45 teamIDretro
## 2980       Oracle Park    1679484  98  97      SFG            SFN         SFN
## 2981 Busch Stadium III    2102530  92  92      STL            SLN         SLN
## 2982   Tropicana Field     761072  92  91      TBR            TBA         TBA
## 2983  Globe Life Field    2110258  99 101      TEX            TEX         TEX
## 2984      Sahlen Field     805901 102 101      TOR            TOR         TOR
## 2985    Nationals Park    1465543  95  96      WSN            MON         WAS
```

```
Teams %>%
  mutate( X1B = H-X2B-X3B-HR) %>%
  summarize(
    BA = H/AB,
    SLG = ( X1B + 2*X2B + 3*X3B + 4*HR)/AB,
    yearID = yearID,
    teamID = teamID
  ) %>%
  ggplot(aes(
    x= yearID,
    y= SLG,
    color = teamID
    ))+geom_point(size=2)+geom_smooth(se =FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

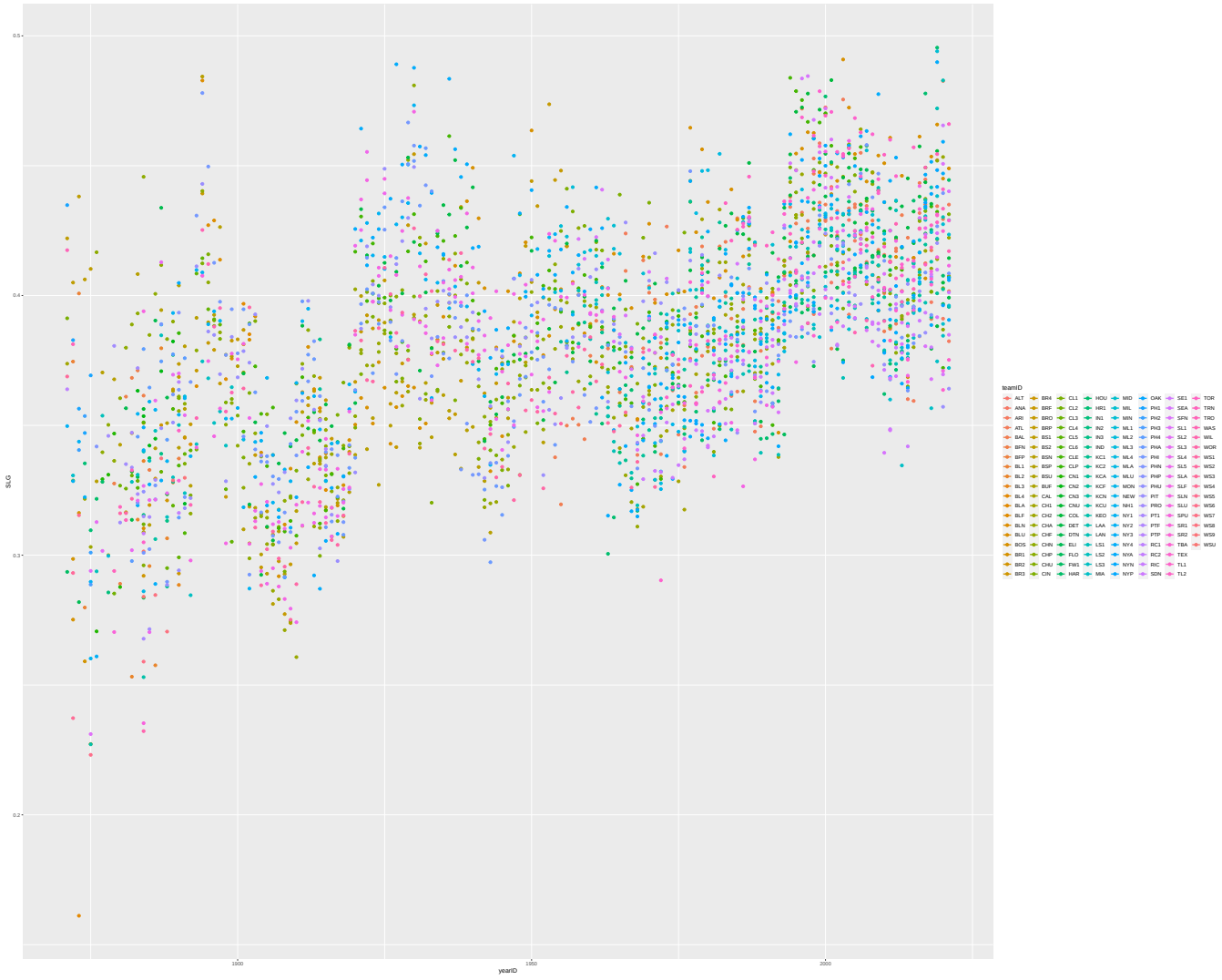
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 1872
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1.01
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1.0201

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : at 1890

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : radius 2.5e-05

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 1890

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 0.005

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : at 1891

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : radius 2.5e-05

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 2.5e-05

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : zero-width neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : zero-width neighborhood. make span bigger

## Warning: Computation failed in `stat_smooth()`:
## NA/NaN/Inf in foreign function call (arg 5)
```




```
Teams %>%
  mutate( X1B = H-X2B-X3B-HR) %>%
  summarize(
    BA = H/AB,
    SLG = ( X1B + 2*X2B + 3*X3B + 4*HR)/AB,
    yearID = yearID,
    teamID = teamID
  ) %>%
  ggplot(aes(
    x= yearID,
    y= BA,
    color = teamID
    ))+geom_point(size =2)+geom_smooth(se =FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

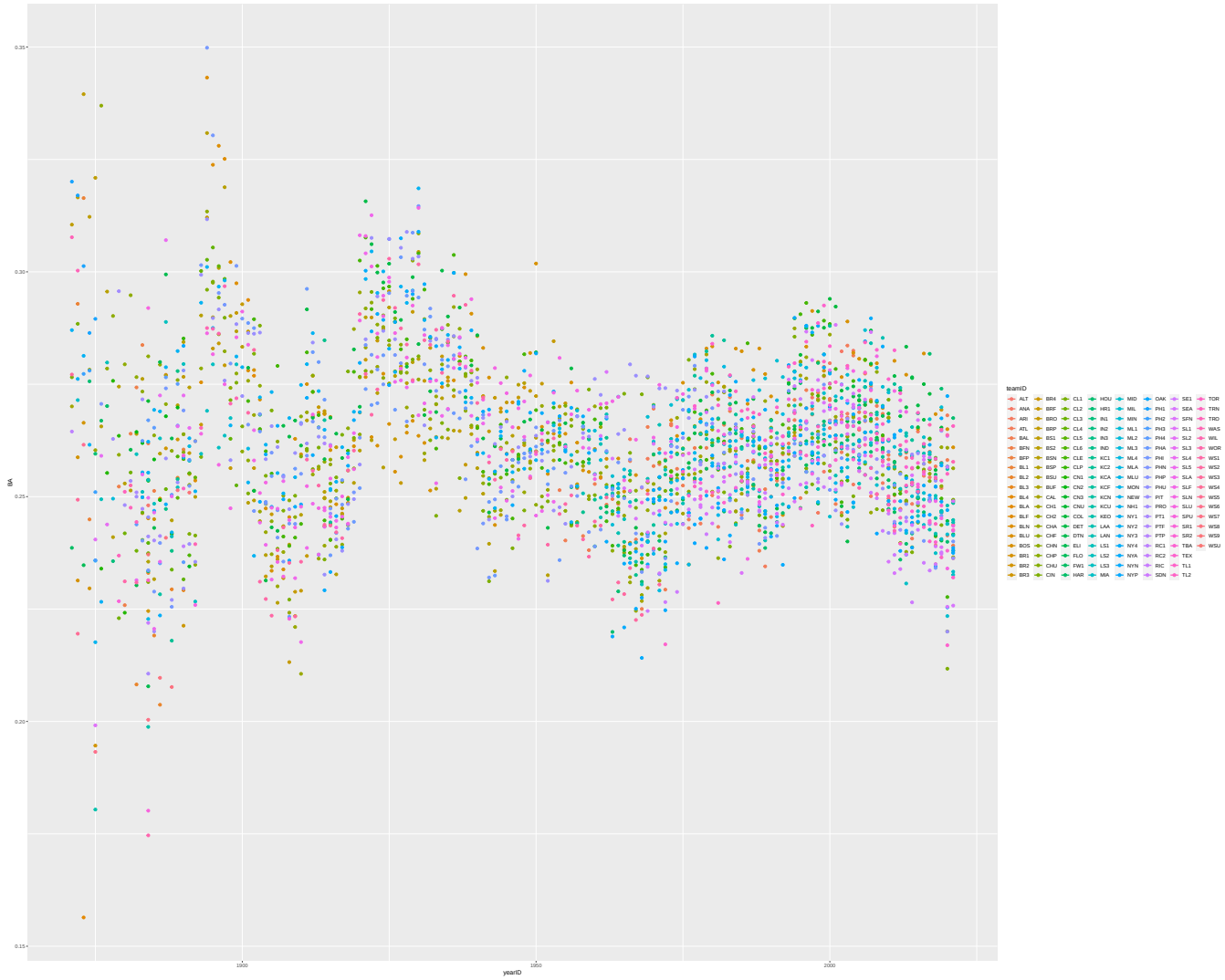
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 1872
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1.01
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1.0201

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : at 1890

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : radius 2.5e-05

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 1890

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 0.005

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : at 1891

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : radius 2.5e-05

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 2.5e-05

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : zero-width neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : zero-width neighborhood. make span bigger

## Warning: Computation failed in `stat_smooth()`:
## NA/NaN/Inf in foreign function call (arg 5)
```



Part 4: Textbook, Section 4.4, Exercise 7

```
Teams %>%
  mutate( X1B = H-X2B-X3B-HR) %>%
  summarize(
    teamID = teamID,
    SLG = ( X1B + 2*X2B + 3*X3B + 4*HR)/AB
  ) %>%
  arrange(desc(SLG)) %>%
  head(5)
```

```
##   teamID       SLG
## 1    HOU 0.4954570
## 2    MIN 0.4940684
## 3    BOS 0.4908996
## 4    NYA 0.4898800
## 5    NYA 0.4890593
```

```
Teams %>%
  filter(yearID >= 1969) %>%
  mutate( X1B = H-X2B-X3B-HR) %>%
  summarize(
    teamID = teamID,
    SLG = ( X1B + 2*X2B + 3*X3B + 4*HR)/AB
```

```
  ) %>%
  arrange(desc(SLG)) %>%
  head(5)
```

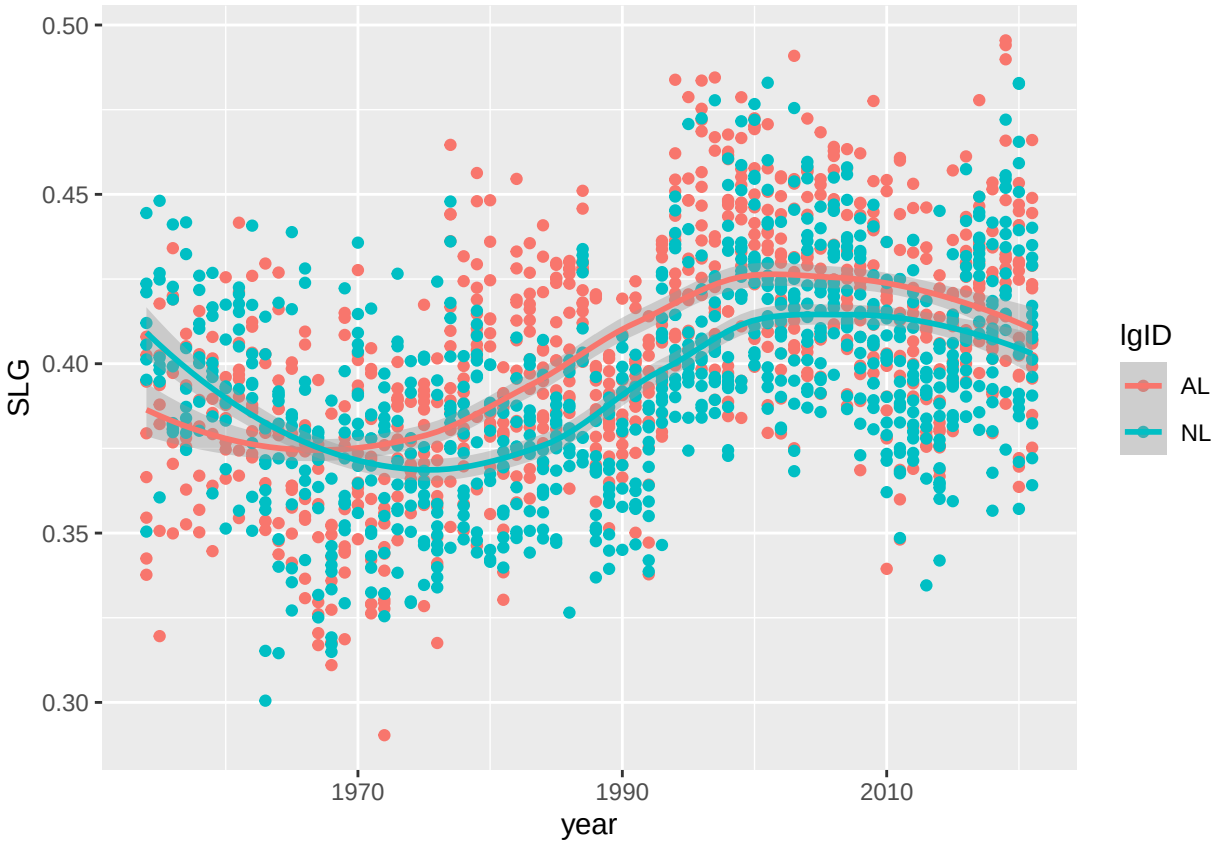
```
##   teamID       SLG
## 1    HOU 0.4954570
## 2    MIN 0.4940684
## 3    BOS 0.4908996
## 4    NYA 0.4898800
## 5    SEA 0.4845030
```

Part 5: Textbook, Section 4.4, Exercise 8

a.

```
Teams %>%
  filter (yearID >= 1954) %>%
  group_by(lgID) %>%
  mutate( X1B = H-X2B-X3B-HR) %>%
  summarise(
    teamID = teamID,
    SLG = ( X1B + 2*X2B + 3*X3B + 4*HR)/AB,
    year = yearID
  ) %>%
  ggplot(aes(
    x = year,
    y = SLG,
    color = lgID
  ))+geom_point()+
  geom_smooth()
```

```
## `summarise()` has grouped output by 'lgID'. You can override using the
## `.groups` argument.
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



AL does have a higher SLG since 1954. According to the graph, most of the time, AL line is above NL line, therefore, it will has a higher mean compare to the others.

Part 6: Textbook, Section 5.5, Exercise 1

```
statenames <- tibble(names = state.name, twoletter = state.abb)
glimpse(statenames)
```

```
## Rows: 50
## Columns: 2
## $ names     <chr> "Alabama", "Alaska", "Arizona", "Arkansas", "California", "C~
## $ twoletter <chr> "AL", "AK", "AZ", "AR", "CA", "CO", "CT", "DE", "FL", "GA", ~
```

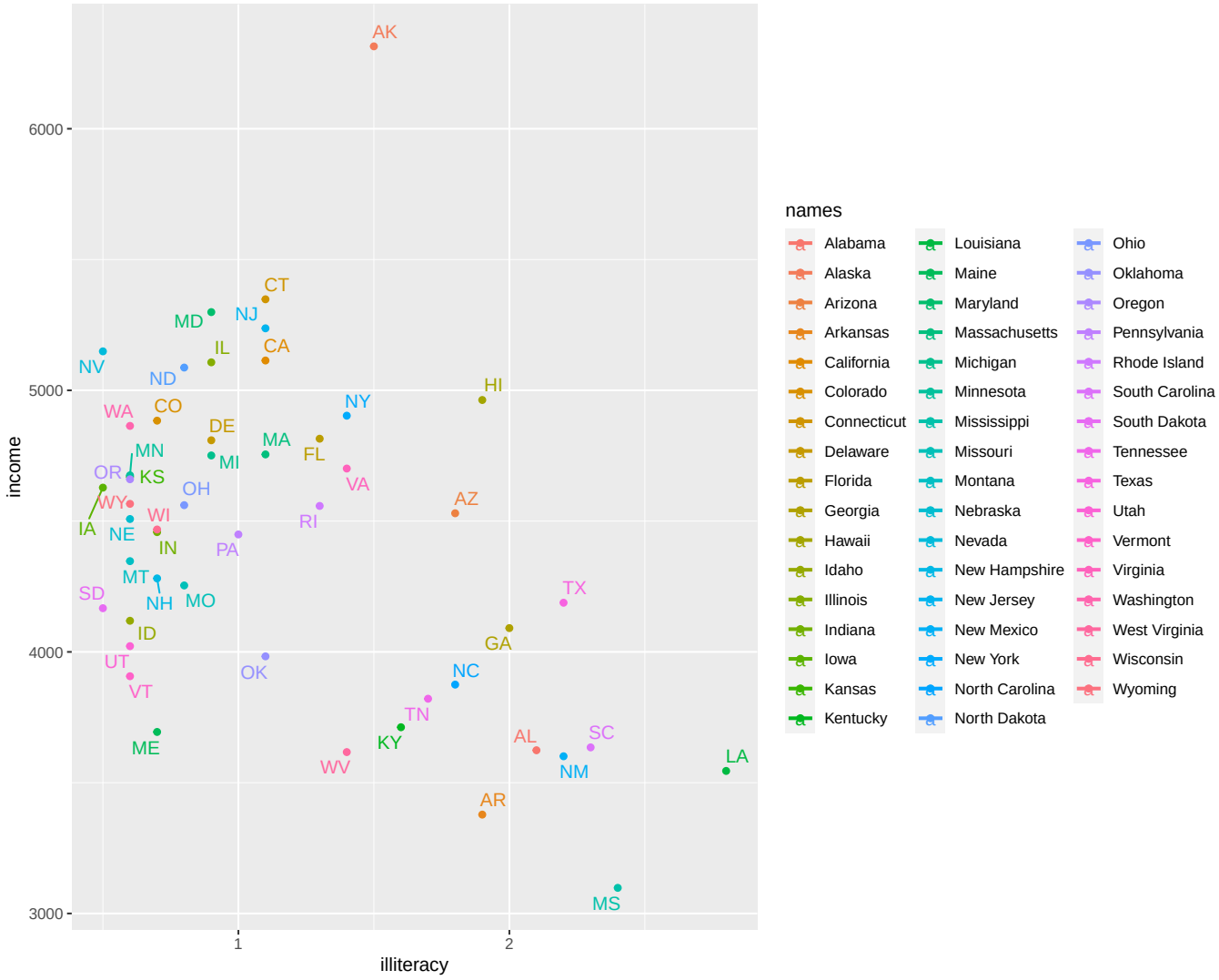
```
statedata <- tibble(
  names = state.name,
  income = state.x77[, 2],
  illiteracy = state.x77[, 3]
  )
glimpse(statedata)
```

```
## Rows: 50
## Columns: 3
## $ names      <chr> "Alabama", "Alaska", "Arizona", "Arkansas", "California", "~
## $ income     <dbl> 3624, 6315, 4530, 3378, 5114, 4884, 5348, 4809, 4815, 4091,~
## $ illiteracy <dbl> 2.1, 1.5, 1.8, 1.9, 1.1, 0.7, 1.1, 0.9, 1.3, 2.0, 1.9, 0.6,~
```

```
library(ggrepel)

statedata <- statedata %>%
  left_join(statenames, by = c("names" = "names"))

statedata %>%
```

```
  ggplot(aes(
    x = illiteracy,
    y = income,
    color = names
  ))+
  geom_point()+
  geom_text_repel(data = statedata, aes(label = twoletter))+
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Part 7: Textbook, Section 5.5, Exercise 2

Part a.

```
Batting %>%
  group_by(playerID) %>%
  summarise(
    total_HR = sum(HR),
    total_SB = sum(SB)
  ) %>%
  right_join(People, by = c("playerID" = "playerID")) %>%
  filter(total_HR >= 300 ) %>%
  filter(total_SB >= 300) %>%
```

```
  select(playerID, nameFirst, nameLast, nameGiven, total_HR, total_SB)
```

```
## # A tibble: 8 x 6
##   playerID  nameFirst nameLast  nameGiven          total_HR total_SB
##   <chr>     <chr>     <chr>     <chr>                 <int>    <int>
## 1 beltrca01 Carlos    Beltran   Carlos Ivan             435      312
## 2 bondsba01 Barry     Bonds     Barry Lamar             762      514
## 3 bondsbo01 Bobby     Bonds     Bobby Lee               332      461
## 4 dawsoan01 Andre     Dawson    Andre Nolan             438      314
## 5 finlest01 Steve     Finley    Steven Allen            304      320
## 6 mayswi01  Willie    Mays      Willie Howard           660      338
## 7 rodrial01 Alex      Rodriguez Alexander Enmanuel      696      329
## 8 sandere02 Reggie    Sanders   Reginald Laverne        305      304
```

Part b.

```
Pitching %>%
  group_by(playerID) %>%
  summarise(
    total_W = sum(W),
    total_SO = sum(SO)
  ) %>%
  right_join(People, by = c("playerID" = "playerID")) %>%
  filter(total_W >= 300 ) %>%
  filter(total_SO >= 3000) %>%
  select(playerID, nameFirst, nameLast, nameGiven, total_W, total_SO)
```

```
## # A tibble: 10 x 6
##    playerID  nameFirst nameLast nameGiven       total_W total_SO
##    <chr>     <chr>     <chr>    <chr>             <int>    <int>
##  1 carltst01 Steve     Carlton  Steven Norman       329     4136
##  2 clemero02 Roger     Clemens  William Roger       354     4672
##  3 johnsra05 Randy     Johnson  Randall David       303     4875
##  4 johnswa01 Walter    Johnson  Walter Perry        417     3509
##  5 maddugr01 Greg      Maddux   Gregory Alan        355     3371
##  6 niekrph01 Phil      Niekro   Philip Henry        318     3342
##  7 perryga01 Gaylord   Perry    Gaylord Jackson     314     3534
##  8 ryanno01  Nolan     Ryan     Lynn Nolan          324     5714
##  9 seaveto01 Tom       Seaver   George Thomas       311     3640
## 10 suttodo01 Don       Sutton   Donald Howard       324     3574
```

Part c.

```
Batting %>%
  group_by(playerID, yearID) %>%
  summarize(
    TotalHR = sum(HR),
    BA = sum(H)/sum(AB
    )) %>%
  right_join(People, by = c("playerID" = "playerID")) %>%
  filter(TotalHR >= 50) %>%
  select( yearID, nameFirst, nameLast, nameGiven, TotalHR, BA) %>%
  arrange(BA)
```

```
## `summarise()` has grouped output by 'playerID'. You can override using the
## `.groups` argument.
## Adding missing grouping variables: `playerID`
```

```
## # A tibble: 46 x 7
## # Groups:   playerID [30]
##    playerID  yearID nameFirst nameLast nameGiven              TotalHR    BA
##    <chr>      <int> <chr>     <chr>    <chr>                    <int> <dbl>
```

```
##  1 alonspe01   2019 Pete      Alonso   Peter Morgan               53 0.260
##  2 bautijo02   2010 Jose      Bautista Jose Antonio               54 0.260
##  3 jonesan01   2005 Andruw    Jones    Andruw Rudolf              51 0.263
##  4 marisro01   1961 Roger     Maris    Roger Eugene               61 0.269
##  5 vaughgr01   1998 Greg      Vaughn   Gregory Lamont             50 0.272
##  6 mcgwima01   1997 Mark      McGwire  Mark David                 58 0.274
##  7 fieldce01   1990 Cecil     Fielder  Cecil Grant                51 0.277
##  8 mcgwima01   1999 Mark      McGwire  Mark David                 65 0.278
##  9 stantmi03   2017 Giancarlo Stanton  Giancarlo Cruz-Michael     59 0.281
## 10 judgeaa01   2017 Aaron     Judge    Aaron James                52 0.284
## # ... with 36 more rows
```

Pete Alonso in 2019 has the lowest batting average