sta313 A1

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
movie <- read.csv("toronto-movies.csv", header = T, row.names = 1)
```

```
#data cleaning
#make variable runtime numeric(remove "min")
movie$runtime_in_min <- rep(NA, nrow(movie))
for (i in 1:nrow(movie)) {
  movie$runtime_in_min[i] <- unlist(strsplit(movie$runtime[i], split = " "))[1]
}
movie$runtime_in_min <- as.numeric(movie$runtime_in_min)



#count the number of languages that the movie translated to
movie$language_number <- rep(0, nrow(movie))
for (i in 1:nrow(movie)) {
  movie$language_number[i] <- length(unlist(strsplit(movie$language[i],
                                                     split = ", ")))
}
unique(movie$language_number)
```

```
## [1] 1 3 5 2 4
```

```
movie$languages_number <- as.factor(movie$language_number)

#classified the number of languages for a movie that is more than one as a movie that is translating to
movie$multi_languages <- rep("No", nrow(movie))
for (i in 1:nrow(movie)) {
  if ( movie$language_number[i] > 1) {
  movie$multi_languages[i] <- "Yes"
  }
}
movie$multi_languages <- factor(movie$multi_languages,
                                levels = c("Yes","No"))
movie$metascore[movie$metascore=="N/A"] <- NA
```

```r
#classified by the year
movie$released_recent <- rep("No", nrow(movie))
for (i in 1:nrow(movie)) {
  if ( movie$year[i] > 1999) {
  movie$released_recent[i] <- "Yes"
  }
}
movie$released_recent <- factor(movie$released_recent,
                                levels = c("Yes","No"))

#created a new tibble that only contains specific variables without any missing values
movie_toronto <- movie %>%
  select(imdb_rating, runtime_in_min, multi_languages, metascore, released_recent, year) %>%
  filter(!is.na(imdb_rating)&!
           is.na(runtime_in_min)&!
           is.na(multi_languages)&!
           is.na(metascore)&!is.na(released_recent))


movie_toronto$imdb_rating <- as.numeric(movie_toronto$imdb_rating)
movie_toronto$metascore <- as.numeric(movie_toronto$metascore)

movie_toronto %>%
  na.omit()
```

```
##     imdb_rating runtime_in_min multi_languages metascore released_recent year
## 1           6.9            102              No        45              No 1987
## 2           3.8             95              No        12             Yes 2002
## 3           5.3            111              No        36             Yes 2004
## 4           5.6            102              No        39             Yes 2001
## 5           7.6            101             Yes        64             Yes 2000
## 6           6.3            115             Yes        62             Yes 2002
## 7           6.3            109             Yes        54             Yes 2005
## 8           6.7            110              No        58             Yes 2005
## 9           7.5            110              No        88             Yes 2006
## 10          6.7            123              No        59             Yes 2013
## 11          6.1             91              No        31              No 1998
## 12          6.4             89             Yes        16              No 1995
## 13          7.2             98              No        65              No 1974
## 14          6.7            117             Yes        52             Yes 2002
## 15          5.1            107              No        17             Yes 2000
## 16          5.9            105             Yes        35             Yes 2002
## 17          7.8            108             Yes        44              No 1999
## 18          5.4             79              No        54              No 1997
## 19          7.9            120              No        72             Yes 2002
## 20          7.0            110             Yes        74             Yes 2007
## 21          4.8             93              No        24             Yes 2009
## 22          5.4             89              No        48              No 1998
## 23          7.7            134             Yes        87             Yes 2005
## 24          5.5            104             Yes        40             Yes 2003
## 25          6.4            102             Yes        49             Yes 2014
## 26          7.3            114              No        88             Yes 2005
## 27          4.0             82              No        19             Yes 2003
## 28          5.8            105              No        44             Yes 2001
```

```
## 29          5.0          89          No          28          No 1990
## 30          7.0          97          No          54         Yes 2007
## 31          5.5          94          No          34         Yes 2005
## 32          7.1         113         Yes          81         Yes 2002
## 33          6.3          96          No          48         Yes 2009
## 34          7.9          94          No          77          No 1983
## 35          8.0         144          No          69         Yes 2005
## 36          5.9         130          No          33         Yes 2013
## 37          6.6          98          No          49          No 1982
## 38          6.7          94          No          80         Yes 2005
## 39          5.9         104         Yes          12          No 1988
## 40          5.0         118          No          37         Yes 2003
## 41          4.7          89          No          33         Yes 2004
## 42          6.1         110         Yes          56          No 1999
## 43          7.7         112         Yes          66         Yes 2004
## 44          6.8          97         Yes          56          No 1999
## 45          7.2          90         Yes          61          No 1997
## 46          5.6         105          No          52         Yes 2005
## 47          7.3         101          No          59         Yes 2004
## 48          7.3         116          No          86          No 1988
## 49          6.4         109          No          38         Yes 2002
## 50          4.9          95         Yes          25          No 1994
## 51          6.9          95          No          33          No 1999
## 52          6.2          80         Yes          44         Yes 2010
## 53          6.1          94          No          65          No 1999
## 54          6.9          83          No          65         Yes 2013
## 55          6.3         113         Yes          38         Yes 2001
## 56          5.6         110          No          40          No 1991
## 57          6.8         112          No          73          No 1998
## 58          5.5          87          No          32         Yes 2001
## 59          4.9          99          No          26         Yes 2000
## 60          4.6         116         Yes          29         Yes 2001
## 61          6.9          91          No          61         Yes 2013
## 62          6.4          97          No          52         Yes 2018
## 63          6.8          97          No          68          No 1999
## 64          5.6         101          No          39         Yes 2001
## 65          7.0         103         Yes          71          No 1994
## 66          5.7         106          No          40         Yes 2005
## 67          6.2         104          No          56         Yes 2005
## 68          7.3         136         Yes          62         Yes 2000
## 69          5.5         111          No          43         Yes 2007
## 70          7.0         119         Yes          57         Yes 2008
## 71          7.6          96          No          79          No 1986
## 72          6.1         109         Yes          37          No 1997
## 73          6.8         109          No          49         Yes 2005
## 74          5.7          97          No          37         Yes 2003
## 75          7.3         118          No          67         Yes 2000
## 76          4.7         101          No          45         Yes 2002
## 77          5.4         117         Yes          45         Yes 2005
## 78          6.8         108          No          70         Yes 2000
## 79          2.3         104          No          14         Yes 2001
## 80          4.8         102          No          24         Yes 2004
## 81          8.3         126          No          70          No 1997
## 82          5.9          90         Yes          31         Yes 2000
```

```
## 83          4.8          112           No          24          Yes 2018
## 84          6.6          117           No          81          Yes 2007
## 85          6.7           82          Yes          16           No 1998
## 86          7.1           88          Yes          64          Yes 2004
## 87          7.7           95          Yes          85          Yes 2001
## 88          7.4           96           No          81          Yes 2005
## 89          6.5          126          Yes          62          Yes 2006
## 90          6.7           97           No          59          Yes 2004
## 91          5.4           94           No          37          Yes 2003
## 92          5.6          101           No          45          Yes 2003
## 93          6.4          116           No          45          Yes 2003
## 94          7.6          146           No          74           No 1999
## 95          6.0           98           No          55          Yes 2005
## 96          5.8           98          Yes          46          Yes 2003
## 97          6.2          115           No          48          Yes 2019
## 98          6.3           95           No          53           No 1999
## 99          6.7          112          Yes          61          Yes 2008
## 100         6.4          128          Yes          62          Yes 2005
## 101         7.3          135          Yes          69          Yes 2017
## 102         5.9          101           No          45           No 1995
## 103         4.4           91           No          25          Yes 2001
## 104         7.1          116          Yes          30          Yes 2002
## 105         5.7           96          Yes          33           No 1995
## 106         6.1           88          Yes          35          Yes 2008
## 107         6.7          138          Yes          58          Yes 2002
## 108         8.0          186           No          54          Yes 2003
## 109         6.3          128          Yes          60          Yes 2000
## 110         7.6          117           No          66          Yes 2010
## 111         6.5          103           No          41          Yes 2013
## 112         6.9           89           No          55           No 1996
## 113         6.5          101           No          63          Yes 2008
## 114         6.1           92           No          30          Yes 2001
## 115         5.2           84           No          22          Yes 2000
## 116         6.2           93          Yes          71          Yes 2005
## 117         6.8          121           No          44           No 1996
## 118         5.3           98           No          35          Yes 2000
## 119         3.8           87           No          24          Yes 2008
## 120         5.6           83           No          33          Yes 2005
## 121         6.2          115           No          39          Yes 2006
## 122         6.7          125           No          65           No 1999
## 123         5.4          100          Yes          31          Yes 2008
## 124         7.0           97          Yes          66          Yes 2004
## 125         6.0          105          Yes          55           No 1997
## 126         7.1          102          Yes          83           No 1987
## 127         6.2          105           No          35           No 1996
## 128         6.1          107           No          47           No 1997
## 129         6.5           95          Yes          62          Yes 2002
## 130         7.1          105          Yes          70          Yes 2002
## 131         5.0           91           No          33          Yes 2004
## 132         6.1           98           No          50          Yes 2008
## 133         4.5           84           No          29          Yes 2013
## 134         6.1           94           No          69          Yes 2016
## 135         6.9          123          Yes          42          Yes 2009
## 136         7.1          104           No          70          Yes 2003
```

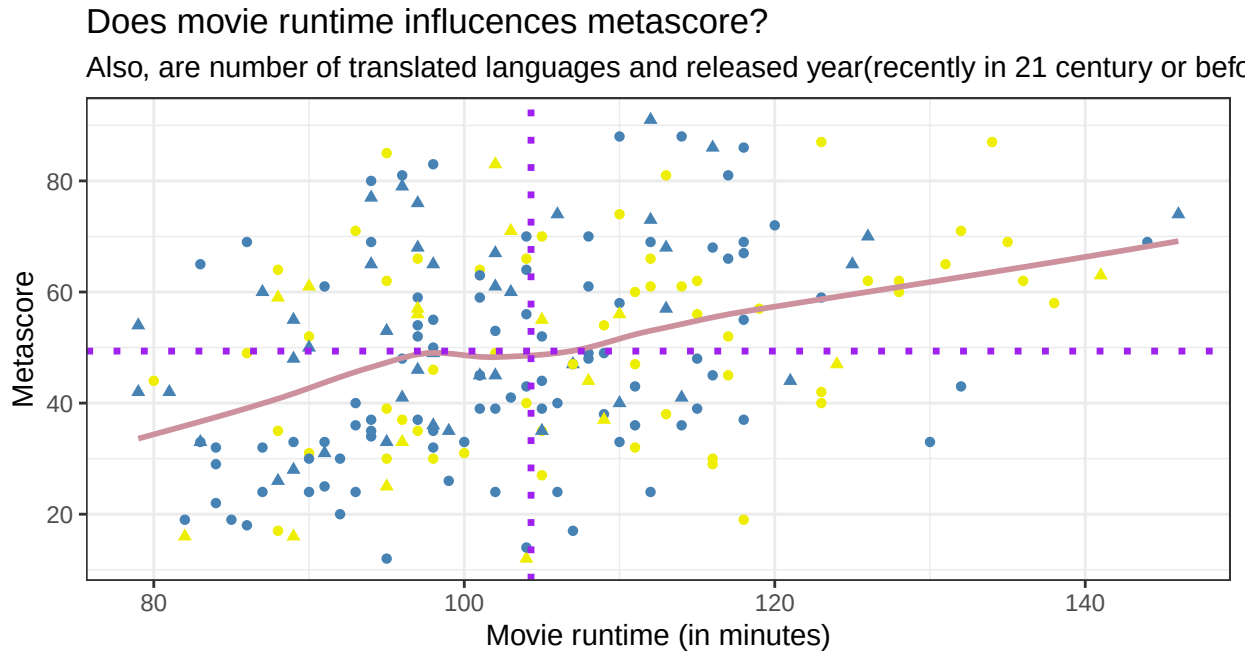
```
## 137  6.9  131  Yes  65  Yes 2013
## 138  5.6   95  Yes  30  Yes 2005
## 139  6.6   79   No  42   No 1994
## 140  5.5  105  Yes  27  Yes 2015
## 141  6.3  118  Yes  19  Yes 2019
## 142  6.7   96   No  41   No 1984
## 143  5.4   83   No  33   No 1986
## 144  5.0   88   No  26   No 1987
## 145  5.5  105   No  39  Yes 2014
## 146  5.9  111  Yes  47  Yes 2004
## 147  3.9   88  Yes  17  Yes 2008
## 148  6.0  124  Yes  47   No 1999
## 149  6.6  115  Yes  56  Yes 2003
## 150  7.0  111  Yes  60  Yes 2010
## 151  6.2  114   No  36  Yes 2012
## 152  6.6   98   No  32  Yes 2008
## 153  6.3  111  Yes  32  Yes 2010
## 154  6.7  100   No  33  Yes 2002
## 155  5.8   96  Yes  37  Yes 2010
## 156  6.2   94   No  35  Yes 2004
## 157  5.4   95  Yes  39  Yes 2012
## 158  6.5  132   No  43  Yes 2001
## 159  7.5  102   No  67   No 1987
## 160  6.2  102   No  53  Yes 2008
## 161  8.1  118   No  86  Yes 2015
## 162  5.5   97  Yes  35  Yes 2005
## 163  6.5   97  Yes  57   No 1994
## 164  5.6   90   No  24  Yes 2010
## 165  6.6   93   No  40  Yes 2005
## 166  6.2  108   No  48  Yes 2006
## 167  5.9   93   No  36  Yes 2007
## 168  5.8   92   No  20  Yes 2008
## 169  6.0   90   No  30  Yes 2009
## 170  6.8  103   No  60   No 1981
## 171  6.2  108   No  61  Yes 2019
## 172  7.5  112   No  69  Yes 2010
## 173  6.1  108   No  49  Yes 2006
## 174  6.9   90  Yes  52  Yes 2001
## 175  6.8   88  Yes  59   No 1985
## 176  7.1  132  Yes  71  Yes 2019
## 177  7.3  123  Yes  87  Yes 2017
## 178  6.7   86  Yes  49  Yes 2007
## 179  7.5  106   No  74   No 1978
## 180  6.9  114   No  41   No 1976
## 181  5.6  106   No  24  Yes 2000
## 182  6.8   98   No  83  Yes 2002
## 183  5.7  104  Yes  66  Yes 2009
## 184  6.7   90   No  50   No 1983
## 185  6.0  123  Yes  40  Yes 2016
## 186  5.2   81   No  42   No 1999
## 187  7.5  112   No  91   No 1997
## 188  6.6  118   No  55  Yes 2006
## 189  6.5  116   No  68  Yes 2011
## 190  7.3  118   No  69  Yes 2007
```

```
## 191         4.1             85             No         19            Yes 2000
## 192         6.0            102             No         61             No 1987
## 193         6.0             98             No         36             No 1999
## 194         7.1            107             No         47            Yes 2009
## 195         7.1             97             No         46             No 1995
## 196         7.5            113             No         57             No 1990
## 197         7.0            114            Yes         61            Yes 2008
## 198         7.2            141            Yes         63             No 1994
## 199         4.2             86             No         18            Yes 2000
## 200         5.4             98            Yes         30            Yes 2002
## 201         6.4            113             No         68             No 1996
## 202         5.8             86             No         69            Yes 2002
## 203         5.6             99             No         35             No 1998
## 204         7.2             87             No         60             No 1983
## 205         7.2             97             No         76             No 1999
## 206         6.8            104             No         43            Yes 2012
## 207         5.2            110             No         33            Yes 2004
## 208         6.4            107            Yes         47            Yes 2005
## 209         6.1             84             No         32            Yes 2003
## 210         7.4            104             No         64            Yes 2000
```

```
movie_toronto %>%  filter(runtime_in_min < 170) %>%
  ggplot(aes(runtime_in_min, metascore)) +
  geom_point(aes(colour = multi_languages, shape = released_recent)) +
  geom_smooth(aes(runtime_in_min, metascore), se =FALSE,  colour = "pink3") +
  theme_bw() +
  scale_x_continuous(breaks=seq(0,180,20)) +
  scale_y_continuous(breaks = seq(0,100,20)) +
  scale_color_manual(name="Translated into multiple(more than 1) languages", values = c("yellow2", "ste

  geom_vline(xintercept = mean(movie_toronto$runtime_in_min, na.rm=T),
             linetype="dotted", color = "purple",
             size = 1.2) +
  geom_hline(yintercept = mean(movie_toronto$metascore,
                               na.rm=T),
             linetype="dotted",
             color = "purple",
             size = 1.2) +
  labs(x = "Movie runtime (in minutes)",
       y = "Metascore",
       title = "Does movie runtime influcences metascore?",
       subtitle = "Also, are number of translated languages and released year(recently in 21 century or
       caption = "Source of data: https://en.wikipedia.org/wiki/List_of_films_shot_in_Toronto \n
                  Correlation coefficient (r): 0.34(positive moderate correlation between runtime and m
       ) +
  theme(legend.position = "bottom")
```

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



Source of data: https://en.wikipedia.org/wiki/List_of_films_shot_in_Toronto

Correlation coefficient (r): 0.34(positive moderate correlation between runtime and metascore)

```
summary(movie_toronto)
```

```
##   imdb_rating    runtime_in_min  multi_languages   metascore
##  Min.   :2.300   Min.   : 79.0   Yes: 73         Min.   :12.00
##  1st Qu.:5.700   1st Qu.: 95.0   No :137         1st Qu.:35.00
##  Median :6.350   Median :102.0                   Median :48.00
##  Mean   :6.284   Mean   :104.3                   Mean   :49.36
##  3rd Qu.:6.900   3rd Qu.:112.0                   3rd Qu.:63.00
##  Max.   :8.300   Max.   :186.0                   Max.   :91.00
##  released_recent     year
##  Yes:150         Length:210
##  No : 60         Class :character
##                  Mode  :character
##
##
##
```

```
correlation <- cor(movie_toronto$runtime_in_min, movie_toronto$imdb_rating)
correlation
```

```
## [1] 0.4032022
```