MSCS 264: Homework #13

Due Tues Nov 20 at 11:59 PM

You should submit a knitted pdf file on Moodle, but be sure to show all of your R code, in addition to your output, plots, and written responses.

Web scraping

1. Read in the table of data found at https://en.wikipedia.org/wiki/List_of_United_States_cities_by_crime_rate and create a plot showing violent crime rate (total violent crime) vs. property crime rate (total property crime). Identify outlier cities (those with "extreme" values for `VCrate` and/or `PCrate`) by feeding a data set of outliers into `geom_label_repel()`.

Hints:

- after reading in the table using `html_table()`, create a data frame with just the columns you want, using a command such as: `crimes3 <- as.data.frame(crimes2)[,c(LIST OF COLUMN NUMBERS)]`. Otherwise, R gets confused since it appears as if several columns all have the same column name.
- then, turn `crimes3` into a tibble with `as.tibble(crimes3)` and do necessary tidying: get rid of unneeded rows, parse columns into proper format, etc.

```r
wiki <- read_html("https://en.wikipedia.org/wiki/List_of_United_States_cities_by_crime_rate")
wiki1 <- html_nodes(wiki, css = "table")
html_table(wiki1, header = TRUE, fill = TRUE)
```

```
## [[1]]
## [1]
## [2] This article needs to be updated. Please update this article to reflect recent events or newly a
## <0 rows> (or 0-length row.names)
##
## [[2]]
##                  State                    City Population Violent Crime
## 1                State                    City Population         Total
## 2           New Mexico             Albuquerque    559,721         965.8
## 3           California                 Anaheim    349,471         363.7
## 4               Alaska               Anchorage    301,239        1070.9
## 5                Texas               Arlington    387,565         502.1
## 6              Georgia                 Atlanta    464,710        1119.6
## 7             Colorado                  Aurora    360,237         460.8
## 8                Texas                  Austin    938,728         372.5
## 9           California             Bakersfield    373,887         484.1
## 10            Maryland               Baltimore    621,252        1535.9
## 11       Massachusetts                  Boston    665,258         706.8
## 12            New York                 Buffalo    258,096        1118.6
## 13             Arizona                Chandler    258,875         189.3
## 14      North Carolina Charlotte-Mecklenburg    877,817         677.6
## 15            Illinois                 Chicago  2,728,695         903.8
## 16          California             Chula Vista    265,215         265.8
## 17                Ohio              Cincinnati    298,478         925.0
## 18                Ohio              Cleveland*    388,655        1334.3
## 19            Colorado        Colorado Springs    452,410         438.3
## 20                Ohio                Columbus    860,090         546.3
## 21               Texas          Corpus Christi    324,326         645.0
## 22               Texas                  Dallas  1,301,977         694.2
```

```
## 23          Colorado                     Denver   682,418         673.9
## 24          Michigan                    Detroit   673,225        1759.6
## 25    North Carolina                     Durham   257,911         847.2
## 26             Texas                    El Paso   686,077         366.6
## 27           Indiana                 Fort Wayne   259,712         378.9
## 28             Texas                 Fort Worth   829,731         525.4
## 29        California                     Fresno   520,837         551.2
## 30    North Carolina                 Greensboro   285,950         597.0
## 31            Nevada                  Henderson   282,554         168.5
## 32            Hawaii                   Honolulu   999,307         243.9
## 33             Texas                    Houston 2,275,221         966.7
## 34           Indiana               Indianapolis   863,675        1288.0
## 35        California                     Irvine   258,198          55.8
## 36           Florida               Jacksonville   867,258         648.3
## 37        New Jersey               Jersey City   265,159         521.6
## 38          Missouri                Kansas City   473,373        1417.3
## 39             Texas                     Laredo   256,280         379.3
## 40            Nevada                  Las Vegas 1,562,134         920.7
## 41          Kentucky                  Lexington   314,077         332.4
## 42          Nebraska                    Lincoln   276,585         370.6
## 43        California                 Long Beach   476,318         580.7
## 44        California                Los Angeles 3,962,726         634.8
## 45          Kentucky           Louisville Metro   680,550         631.8
## 46         Tennessee                    Memphis   657,936        1740.1
## 47           Arizona                       Mesa   471,034         418.7
## 48           Florida                      Miami   437,969        1021.3
## 49         Wisconsin                  Milwaukee   600,400        1596.1
## 50         Minnesota                Minneapolis   413,479        1062.9
## 51           Alabama                    Mobile2   250,346         610.8
## 52         Tennessee Nashville Metropolitan   658,029        1101.0
## 53         Louisiana                New Orleans   393,447         949.6
## 54          New York                   New York 8,537,673         585.8
## 55        New Jersey                    Newark*   279,110        1077.7
## 56        California                    Oakland   419,481        1442.5
## 57          Oklahoma              Oklahoma City   630,621         765.6
## 58          Nebraska                      Omaha   452,252         515.0
## 59           Florida                    Orlando   268,438         940.6
## 60      Pennsylvania               Philadelphia 1,567,810        1029.0
## 61           Arizona                    Phoenix 1,559,744         593.8
## 62      Pennsylvania                 Pittsburgh   306,870         706.2
## 63             Texas                      Plano   282,968         153.0
## 64            Oregon                  Portland*   615,672         472.8
## 65    North Carolina                  Raleigh**   428,993         392.3
## 66        California                  Riverside   323,064         446.0
## 67        California                 Sacramento   489,717         737.4
## 68             Texas                San Antonio 1,463,586         587.2
## 69        California                  San Diego 1,400,467         398.6
## 70        California              San Francisco   863,782         776.8
## 71        California                   San Jose 1,031,458         329.6
## 72        California                  Santa Ana   337,304         482.1
## 73        Washington                    Seattle   683,700         598.7
## 74          Missouri                  St. Louis   317,095        1817.1
## 75         Minnesota                   St. Paul   300,721         703.3
## 76           Florida             St. Petersburg   255,821         741.9
```

```
## 77           California                Stockton    304,890         1352.0
## 78              Florida                   Tampa    364,383          630.7
## 79                 Ohio                  Toledo    279,552         1128.9
## 80              Arizona                  Tucson    529,675          655.5
## 81             Oklahoma                   Tulsa    401,520          903.6
## 82             Virginia          Virginia Beach    452,797          138.3
## 83 District Of Columbia              Washington    672,228         1202.6
## 84               Kansas                 Wichita    389,824          984.8
##                                 Violent Crime Violent Crime Violent Crime
## 1  Murder and\nNonnegligent Manslaughter Violent Crime          Rape
## 2                                         7.7          72.2         301.2
## 3                                         5.2          36.9         125.6
## 4                                         8.6         171.6         206.1
## 5                                         2.1          53.7         136.5
## 6                                        20.2          36.6         429.3
## 7                                         6.7          97.7         124.1
## 8                                         2.5          51.9          99.0
## 9                                         5.9          19.0         175.2
## 10                                       57.8          46.2         694.2
## 11                                        5.7          36.1         233.1
## 12                                       15.9          67.0         400.2
## 13                                        0.4          30.5          44.0
## 14                                        6.9          24.6         221.8
## 15                                       23.8          52.5         353.6
## 16                                        2.3          23.8          90.1
## 17                                       22.1          79.1         423.1
## 18                                       16.2         124.0         769.3
## 19                                        5.5          75.4          83.1
## 20                                        9.1          95.1         264.2
## 21                                        5.2          86.9         121.5
## 22                                       10.4          60.1         320.8
## 23                                        7.8          80.3         180.2
## 24                                       43.4          78.7         513.5
## 25                                       13.2          28.7         286.1
## 26                                        2.5          46.9          59.8
## 27                                        9.6          37.3         171.7
## 28                                        6.7          62.2         118.2
## 29                                        7.5          32.1         194.3
## 30                                        9.1          25.2         185.3
## 31                                        1.4          34.7          63.7
## 32                                        1.5          31.8          89.7
## 33                                       13.3          43.3         451.7
## 34                                       17.1          78.4         440.2
## 35                                        0.8          10.5          22.1
## 36                                       11.2          54.3         161.2
## 37                                       10.2          36.6         207.0
## 38                                       23.0          77.3         359.8
## 39                                        3.1          51.9          63.2
## 40                                        8.1          70.9         320.7
## 41                                        4.8          51.9         162.7
## 42                                        0.4          69.8          77.4
## 43                                        7.6          37.2         221.3
## 44                                        7.1          55.7         225.9
## 45                                       11.9          30.1         227.0
```

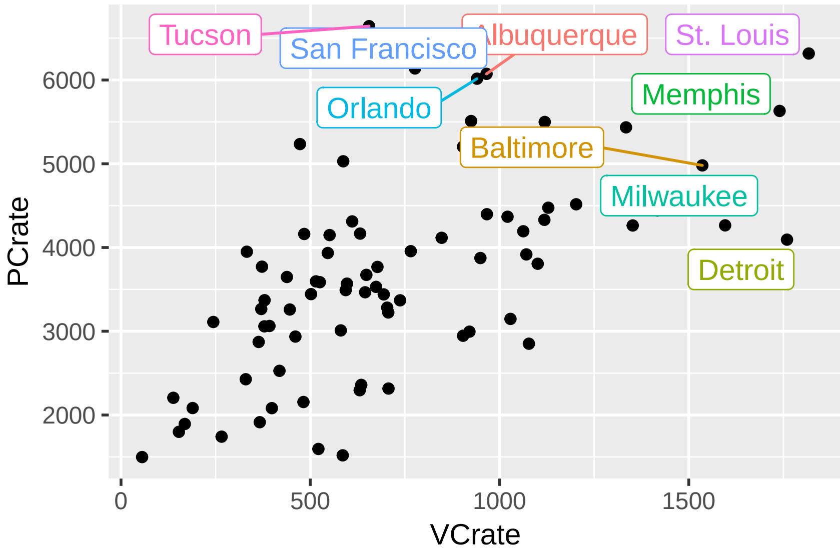
```
## 46                                  20.5          80.6         475.9
## 47                                   3.4          51.2          86.6
## 48                                  17.1          18.3         383.8
## 49                                  24.2          72.6         624.4
## 50                                  11.4          98.4         458.5
## 51                                   9.6          46.3         160.6
## 52                                  10.9          77.0         280.7
## 53                                  41.7         104.0         380.5
## 54                                   3.4          14.0         198.2
## 55                                  33.3          17.6         688.6
## 56                                  20.3          67.9         784.3
## 57                                  11.6          76.1         189.0
## 58                                  10.6          38.5         144.8
## 59                                  11.9          67.8         194.5
## 60                                  17.9          84.3         431.5
## 61                                   7.2          65.1         193.6
## 62                                  18.6          26.7         279.6
## 63                                   1.4          32.2          41.0
## 64                                   4.2          42.6         137.6
## 65                                   2.8          18.4         141.0
## 66                                   3.1          42.4         161.0
## 67                                   8.8          21.4         239.7
## 68                                   6.4          71.7         135.7
## 69                                   2.6          40.4          98.4
## 70                                   6.1          39.8         417.9
## 71                                   2.9          36.4         110.5
## 72                                   3.6          45.7         155.1
## 73                                   3.4          21.1         224.1
## 74                                  59.8          82.9         564.5
## 75                                   5.3          67.8         237.4
## 76                                   5.5          53.2         224.0
## 77                                  16.1          44.3         375.2
## 78                                   9.3          21.1         184.1
## 79                                   8.6          81.6         322.7
## 80                                   5.9          79.7         199.9
## 81                                  13.7          90.9         212.7
## 82                                   4.2          22.7          59.6
## 83                                  24.1          73.5         506.4
## 84                                   6.9          89.5         188.0
##    Violent Crime     Property Crime Property Crime Property Crime
## 1        Robbery Aggravated Assault          Total       Burglary
## 2          584.8             6073.2         1071.2         4076.7
## 3          196.0             2872.3          422.4         1972.4
## 4          684.5             3917.5          559.4         2975.0
## 5          309.9             3443.6          559.9         2657.1
## 6          633.5             5499.3         1028.8         3549.1
## 7          232.3             2936.7          467.2         2119.4
## 8          219.2             3771.0          532.6         2990.0
## 9          284.0             4161.4         1036.9         2484.2
## 10         740.1             4980.4         1248.6         2842.3
## 11         431.9             2316.1          354.3         1769.5
## 12         635.4             4330.2         1076.0         2875.3
## 13         114.3             2083.2          297.4         1686.9
## 14         424.2             3767.9          769.5         2744.4
```

```
## 15         480.2           2946.3        482.0         2089.7
## 16         149.7           1741.6        267.7         1166.2
## 17         400.7           5510.0       1478.5         3642.8
## 18         424.8           5434.4       1787.7         2659.2
## 19         274.3           3648.0        533.6         2732.7
## 20         177.9           3934.3        851.6         2715.6
## 21         431.4           3465.6        674.6         2632.8
## 22         302.8           3440.2        854.2         2002.8
## 23         405.6           3529.9        697.8         2192.5
## 24        1123.5           4093.6       1161.6         2157.2
## 25         519.2           4115.8       1234.5         2644.7
## 26         257.4           1914.2        206.8         1591.1
## 27         160.2           3058.4        569.9         2360.7
## 28         338.2           3585.7        723.7         2589.9
## 29         317.4           4148.3        850.4         2723.3
## 30         377.3           3568.5        828.5         2551.5
## 31          68.7           1893.1        479.9         1225.6
## 32         120.9           3110.7        428.7         2294.6
## 33         458.3           4397.5        872.8         2928.7
## 34         752.3           4790.8       1283.5         2929.5
## 35          22.5           1498.1        202.6         1217.7
## 36         421.6           3673.0        701.3         2704.6
## 37         267.8           1594.9        368.1          998.6
## 38         957.2           4441.3       1029.0         2587.6
## 39         261.0           3370.9        405.8         2843.8
## 40         521.0           2995.3        952.3         1537.0
## 41         113.0           3949.7        797.6         2834.0
## 42         223.1           3265.9        480.1         2655.2
## 43         314.7           3010.0        649.6         1766.3
## 44         346.0           2359.6        407.8         1544.2
## 45         362.8           4166.0        922.3         2769.8
## 46        1163.2           5630.8       1561.2         3655.2
## 47         277.5           2527.4        471.1         1881.2
## 48         602.1           4367.4        709.9         3132.9
## 49         874.9           4264.2        912.9         2122.1
## 50         494.6           4193.9        859.8         2918.9
## 51         394.3           4311.6        882.0         3186.4
## 52         732.3           3805.8        779.9         2805.8
## 53         423.4           3874.2        736.6         2497.9
## 54         357.2           1518.7        164.9         1267.4
## 55         338.2           2851.2        622.0         1365.1
## 56         570.0           5856.8        794.6         3539.1
## 57         488.9           3956.1        923.4         2573.3
## 58         321.1           3595.6        477.6         2555.7
## 59         666.4           6015.5       1267.0         4309.0
## 60         495.3           3147.4        515.6         2310.7
## 61         327.8           3491.3        820.5         2198.3
## 62         381.3           3224.5        715.9         2312.7
## 63          78.5           1799.1        260.1         1442.9
## 64         288.5           5234.8        673.4         4013.0
## 65         230.1           3063.0        735.9         2162.7
## 66         239.6           3259.7        505.5         2211.3
## 67         467.4           3369.5        758.2         2014.4
## 68         373.4           5029.5        794.8         3812.8
```

```
## 69          257.1              2082.0          366.2          1351.9
## 70          312.9              6138.0          600.4          4737.1
## 71          179.8              2427.1          474.7          1273.7
## 72          277.8              2155.3          269.5          1332.3
## 73          350.2              5522.0         1122.9          3831.9
## 74         1110.4              6316.1         1325.2          3998.8
## 75          392.7              3282.1          706.6          1994.2
## 76          459.3              5622.7          906.5          4120.9
## 77          916.4              4263.2          948.2          2662.9
## 78          416.0              2295.9          503.0          1629.3
## 79          716.1              4475.0         1476.3          2676.8
## 80          370.0              6642.8          691.7          5586.8
## 81          586.3              5203.2         1372.8          3170.2
## 82           51.7              2205.6          211.1          1898.4
## 83          598.6              4516.2          442.0          3599.1
## 84          700.3              5041.2          892.7          3623.9
##    Property Crime              Arson1
## 1   Larceny-Theft Motor Vehicle Theft
## 2           925.3                15.9
## 3           477.6                 8.0
## 4           383.1                35.2
## 5           226.5                 7.5
## 6           921.4                10.8
## 7           350.0                17.8
## 8           248.3                 9.7
## 9           640.3                90.7
## 10          889.5                41.9
## 11          192.3                 N/A
## 12          378.9                67.0
## 13           98.9                17.4
## 14          253.9                24.6
## 15          374.6                19.6
## 16          307.7                10.2
## 17          388.6               147.4
## 18          987.5                78.2
## 19          381.7                27.9
## 20          367.2                47.8
## 21          158.2                18.2
## 22          583.3                26.8
## 23          639.6                16.4
## 24          774.8               125.1
## 25          236.5                12.0
## 26          116.3                 8.3
## 27          127.8                18.9
## 28          272.1                17.1
## 29          574.7                49.0
## 30          188.5                34.6
## 31          187.6                 7.1
## 32          387.4                23.1
## 33          596.0                29.5
## 34          577.9                 N/A
## 35           77.8                 3.1
## 36          267.0                 8.2
## 37          228.2                23.4
```

```
## 38          824.7                 41.8
## 39          121.4                 34.3
## 40          506.0                  8.6
## 41          318.1                 15.0
## 42          130.5                 32.9
## 43          594.1                 14.5
## 44          407.6                 28.5
## 45          473.9                  N/A
## 46          414.3                 45.9
## 47          175.1                 17.0
## 48          524.7                 20.8
## 49         1229.2                 37.1
## 50          415.3                 28.1
## 51          243.3                  N/A
## 52          220.1                  8.1
## 53          639.7                  9.1
## 54           86.4                  N/A
## 55          864.2                 14.0
## 56         1523.1                 42.7
## 57          459.4                 14.7
## 58          562.3                 17.9
## 59          439.6                 14.5
## 60          321.1                 21.0
## 61          472.5                  N/A
## 62          195.8                 56.0
## 63           96.1                  6.7
## 64          548.3                 27.0
## 65          164.3                 12.6
## 66          542.9                 22.0
## 67          596.9                 28.6
## 68          422.0                 17.8
## 69          363.9                 12.4
## 70          800.5                 31.5
## 71          678.7                  9.0
## 72          553.5                 11.0
## 73          567.2                 13.5
## 74          992.1                 70.3
## 75          581.3                 39.9
## 76          595.3                 25.0
## 77          652.0                 33.8
## 78          163.6                 14.8
## 79          321.9                  N/A
## 80          364.2                 22.1
## 81          660.2                 43.3
## 82           96.1                 21.0
## 83          475.1                  N/A
## 84          524.6                 33.1
##
## [[3]]
##                              vteUnited States Crime Rates By City Population
## 1 250,000 and Above\n100,000 to 250,000\n60,000 to 100,000\n40,000 to 60,000
##                              vteUnited States Crime Rates By City Population
## 1 250,000 and Above\n100,000 to 250,000\n60,000 to 100,000\n40,000 to 60,000
```

```
wiki2 <- html_table(wiki1, header = TRUE, fill = TRUE)[[2]]

crimes <- as.data.frame(wiki2)[-c(1),c(1, 2, 4, 9)]

crimes1 <- as.tibble(crimes) %>%
  mutate("VCrate" = as.double(`Violent Crime`),
  "PCrate" = as.double(`Property Crime`))
  outliers <- crimes1 %>%
  filter(PCrate >= 6000 | VCrate >= 1500)

crimes1 %>%
  ggplot(aes(x=VCrate, y=PCrate))+
  geom_point()+
  ggrepel::geom_label_repel(aes(label = City, colour = City), data = outliers,
  show.legend = FALSE)
```

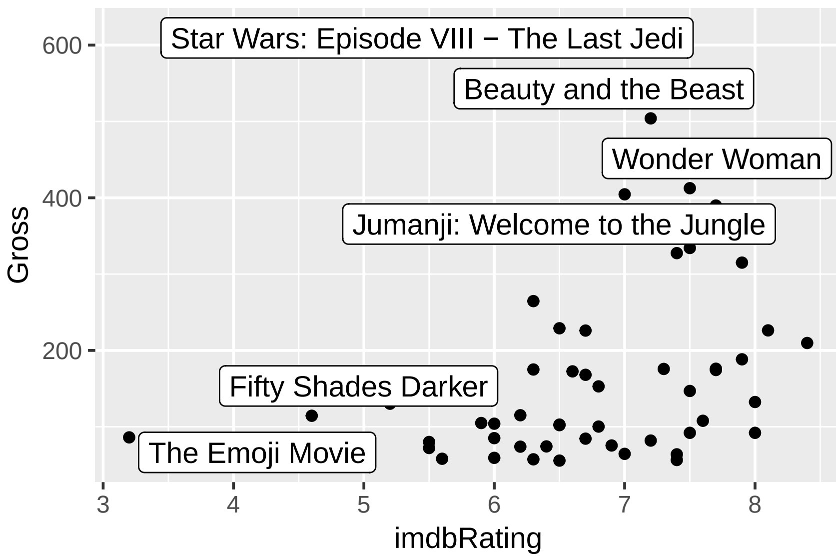


Test line for class

2. As we did in class, use the `rvest` package to pull off data from imdb's top grossing films released in 2017 at https://www.imdb.com/search/title?year=2017&title_type=feature&sort=boxoffice_gross_us,desc. Create a tibble that contains the title, gross, imdbscore, and metascore for the top 50 films. Then generate a scatterplot of one of the ratings vs. gross, labelling outliers as in Question 1 with the title of the movie.

```
top50 <- read_html("https://www.imdb.com/search/title?year=2017&title_type=feature&sort=boxoffice_gross_

imdbscore0 <- html_nodes(top50, ".ratings-imdb-rating strong")
imdbscore <- html_text(imdbscore0)

imdbtitle0 <- html_nodes(top50, ".lister-item-header a")
imdbtitle <- html_text(imdbtitle0)

imdbmetascore0 <- html_nodes(top50,".favorable")
imdbmetascore <- html_text(imdbmetascore0)

imdbgross0 <- html_nodes(top50, ".ghost~ .text-muted+ span")
imdbgross <- html_text(imdbgross0)
```

```
imdb <- tibble(Title = character(), Metascore = double(), imdbRating = double(), Gross = double())
for(i in 1:50){
imdb[i,1] <- imdbtitle[i]
imdb[i,2] <- parse_number(imdbmetascore)[i]
imdb[i,3] <- parse_number(imdbscore)[i]
imdb[i,4] <- parse_number(imdbgross)[i]
}

outliers <- imdb %>%
filter(imdbRating<5 | Gross >= 400)

ggplot(imdb, aes(imdbRating,Gross))+
  geom_point()+
  ggrepel::geom_label_repel(aes(label = Title), data = outliers,show.legend = FALSE)
```



3. 5 points if you push your Rmd file with HW13 solutions along with the knitted pdf file to your MSCS264-HW13 repository in your GitHub account. So that I can check, make your repository private (good practice when doing HW), but add me (username = proback) as a collaborator under Settings > Collaborators.

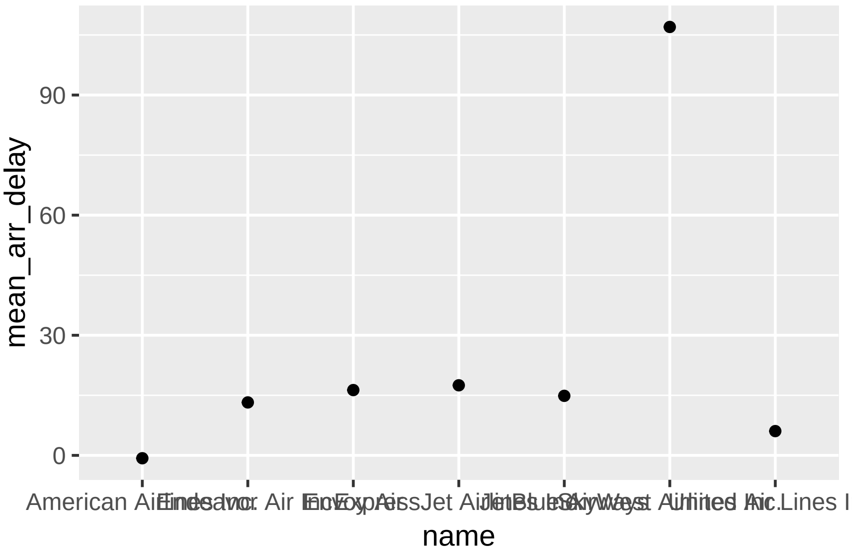
Factors

Read Chapter 15 on factors and attempt the following problems:

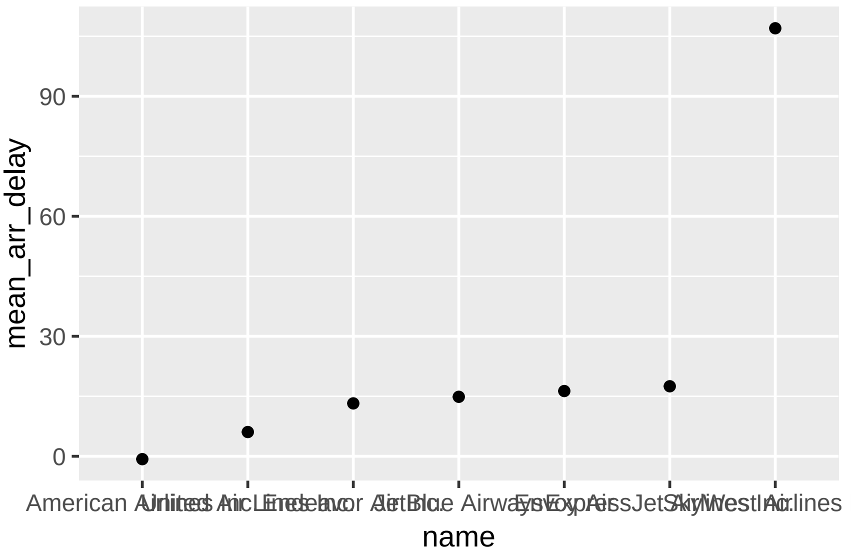
4. In the `nycflights13` data, just consider flights to O'Hare (dest==“ORD”), and summarize the mean arrival delay by carrier (actually use the entire name of the carrier after merging carrier names into `flights`). Then use `geom_point` to plot mean arrival delay vs. carrier - first without reordering carrier names, and second after reordering carrier names by mean arrival delay.

```
flights1 <- flights %>%
  full_join(airlines, by = "carrier")

ord <- filter(flights1, dest == "ORD")

ord %>%
  group_by(name)%>%
```

```
  summarise("mean_arr_delay" = mean(arr_delay, na.rm = TRUE)) %>%
  ggplot(aes(x=name, y=mean_arr_delay))+
    geom_point()
```



```
ord %>%
  group_by(name)%>%
  summarise("mean_arr_delay" = mean(arr_delay, na.rm = TRUE)) %>%
  mutate(name = fct_reorder(name, mean_arr_delay))%>%
  ggplot(aes(x=name, y=mean_arr_delay))+
    geom_point()
```



5. Again considering only flights to O'Hare, create a new factor variable which differentiates national carriers (American and United) from regional carriers (all others which fly to O'Hare). Then create a violin plot comparing arrival delays for all flights to O'Hare from those two groups (you might want to exclude arrival delays over a certain level).

```
flights1 %>%
filter(arr_delay <= 300, dest == "ORD") %>%
mutate(airline_type = ifelse(carrier == "UA" | carrier == "AA", "National", "Regional")) %>%
```

```r
ggplot(aes(x = airline_type, y= arr_delay))+
geom_violin()
```