Econometrics HW1

Tianyu Gao

2022-09-10

1.

a)

```
library(knitr)
```

$$\Pr(X = 0, Y = 0) = \Pr(X = 0|Y = 0) \times \Pr(Y = 0) = .75 \times .05 = .0375$$
$$\Pr(X = -10, Y = 0) = \Pr(Y = 0) - \Pr(X = 10, Y = 0) - \Pr(X = 0, Y = 0) = .75 - .4875 - .0375 = .225$$
$$\Pr(X = -10) = \Pr(X = -10, Y = 1) + \Pr(X = -10, Y = 0) = .225 + .175 = .4$$
$$\Pr(X = 10) = 1 - \Pr(X = 0) - \Pr(X = -10) = 1 - .4 - .05 = .55$$
$$\Pr(X = 10, Y = 1) = \Pr(X = 10) - \Pr(X = 10, Y = 0) = .55 - .4875 = .0625$$
$$\Pr(X = 0, Y = 1) = \Pr(X = 0) - \Pr(X = 0, Y = 0) = .05 - .0375 = .0125$$
$$\Pr(Y = 1) = 1 - P(Y = 0) = .25$$

The table is filled as follows:

```
q1table = data.frame(col1 = c(.225, .0375, .4875, .75), col2 = c(.175, .0125, .0625, .2
5), col3 = c(.4, .05, .55, 0/0), row.names = c('X = -10', 'X = 0', 'X = 10', 'Pr(X)'))
colnames(q1table) = c('Y = 0', 'Y = 1', 'Pr(Y)')
kable(q1table)
```

	Y = 0	Y = 1	Pr(Y)
X = -10	0.2250	0.1750	0.40
X = 0	0.0375	0.0125	0.05
X = 10	0.4875	0.0625	0.55
Pr(X)	0.7500	0.2500	NaN

b)

$$E(X) = \sum_{i=1}^{3} Pr(X = x_i)x_i$$
$$= .4 \times (-10) + .05 \times 0 + .55 \times 10$$
$$= 1.5$$
$$E(Y) = \sum_{i=1}^{2} Pr(Y = y_i)y_i$$
$$= .75 \times 0 + .25 \times 1$$
$$= .25$$

$$Var(X) = \sum_{i=1}^{3} Pr(X = x_i)(x_i - E(X))^2$$
$$= .4 \times (1.5 + 10)^2 + .05 \times 1.5^2 + .55 \times 8.5^2$$
$$= 92.75$$
$$Var(Y) = \sum_{i=1}^{2} Pr(Y = y_i)(y_i - E(Y))^2 = .75 \times .25^2 + .25 \times .75^2 = .1875$$

c)

When she does not study:

$$E(X|Y = 0) = \sum_{i=1}^{3} Pr(X = x_i|Y = 0)x_i = \frac{4.875 - 2.25}{.75} = 3.5$$

When she studies:

$$E(X|Y = 1) = \sum_{i=1}^{3} Pr(X = x_i|Y = 0)x_i = \frac{0.625 - 1.75}{.25} = -4.5$$

2

a)

$$\sigma = 100, \mu = 500$$
$$P(Score > 750) = P(\frac{Score - 500}{100} > \frac{750 - 500}{100}) = 1 - \Phi(2.5) \approx .006$$

```
print(1 - pnorm(2.5))
```

```
## [1] 0.006209665
```

$$P(\text{Score} > 600) = P(\frac{\text{Score} - 500}{100} > \frac{600 - 500}{100}) = 1 - \Phi(1) \approx 0.16$$

```
print(1 - pnorm(1))
```

```
## [1] 0.1586553
```

$$P(420 < \text{Score} < 530) = P(\frac{420 - 500}{100} < \frac{\text{Score} - 500}{100} < \frac{530 - 500}{100})$$
$$= \Phi(0.3) - \Phi(-0.8) \approx 0.41$$

```
pnorm(0.3) - pnorm(-0.8)
```

```
## [1] 0.406056
```

$$P(\text{Score} < 480) = P(\frac{\text{Score} - 500}{100} < \frac{480 - 500}{100}) = \Phi(-0.2) \approx 0.42$$

```
pnorm(-0.2)
```

```
## [1] 0.4207403
```

$$P(\text{Score} > 530) = P(\frac{\text{Score} - 500}{100} > \frac{530 - 500}{100}) = 1 - \Phi(0.3) \approx 0.38$$

```
1 - pnorm(0.3)
```

```
## [1] 0.3820886
```

b)

$$X : \text{Verbal score}$$
$$Y : \text{Math score}$$
$$\begin{cases} X \sim N(500, 100^2) \\ Y \sim N(500, 100^2) \\ X \perp\!\!\!\perp Y \end{cases} \Rightarrow X + Y \sim N(1000, 20000)$$
$$Var(X + Y) = 20000$$
$$E(X + Y) = 1000$$

c)

$$X: \text{Verbal score}$$
$$Y: \text{Math score}$$
$$E(X+Y) = E(X) + E(Y) = 1000$$
$$Var(X+Y) =$$
$$Var(X) + Var(Y) + 2\varrho_{XY}\sigma_X\sigma_Y = 35000$$

d)

$$E(\bar{Y}) = E(\frac{1}{25}\sum_{i=1}^{25} Y_i) = \frac{1}{25}\sum_{i=1}^{25} E(Y_i) = 500$$

$$Var(\bar{Y}) = Var(\frac{1}{25}\sum_{i=1}^{25} Y_i)$$

$$= \frac{1}{625} \times \sum_{i=1}^{25} Var(Y_i) = 400$$

$$P(\bar{Y} > 530) = P(\frac{\bar{Y} - 500}{20} > \frac{530 - 500}{20}) = 1 - \Phi(1.5) \approx .07$$

```
1 - pnorm(1.5)
```

```
## [1] 0.0668072
```

The variance of the mean is much smaller than the mean, and this is because when we calculate the population mean, we only care about an individual in the population, however, when we talk about the sample mean, we care about some samples assumed to be identical, such a change leads to the decrease of the variance, and when the variance is smaller, the observations are more close to their center, which means fewer outliners will appear. Thus, the probability we get an extreme observation is lower.

3

a)

For two-sided confident interval, when confidence level is 0.95, the critical value is 1.96,

```
qnorm(0.975)
```

```
## [1] 1.959964
```

So the 95% confidence interval is

$$[38644.86 - 1.96 \times \frac{7541.40}{\sqrt{108}}, 38644.86 + 1.96 \times \frac{7541.40}{\sqrt{108}}] = [37222.54, 40067.18]$$

```
38644.86 - 1.96 * 7541.4 / sqrt(108)
```

```
## [1] 37222.54
```

```
38644.86 + 1.96 * 7541.4 / sqrt(108)
```

```
## [1] 40067.18
```

b)

Such difference does not indicate discrimination in the job market against psychology majors, because psychology students might prefer jobs with better work-life balance but lower salary, while the employers have the same criteria about the two majors.

c)

$$\because \mathrm{n_{B+}} > 30, \mathrm{n_B} > 30$$
$$\therefore \mathrm{Y_{B+}} - \mathrm{Y_B} \sim \mathrm{N}(\bar{\mathrm{Y}}_{\mathrm{B+}} - \bar{\mathrm{Y}}_{\mathrm{B}}, \frac{\mathrm{S_{B+}}}{\sqrt{\mathrm{n_{B+}}}} + \frac{\mathrm{S_B}}{\sqrt{\mathrm{n_B}}}) = \mathrm{N}(2831.92, 1966.62)$$

```
39915.25 - 37083.33
```

```
## [1] 2831.92
```

```
sqrt(8330.21 / 59 + 6174.86 / 49)
```

```
## [1] 16.34648
```

$$\mathrm{T_0} : \bar{\mathrm{Y}}_{\mathrm{B+}} - \bar{\mathrm{Y}}_{\mathrm{B}} = 0$$
$$\mathrm{T_A} : \bar{\mathrm{Y}}_{\mathrm{B+}} - \bar{\mathrm{Y}}_{\mathrm{B}} \neq 0$$
$$\mathrm{p - value} = \Phi(-|\mathrm{t}^{\mathrm{act}}|) = .07$$

```
pnorm(-2831.92 / 1966.62)
```

```
## [1] 0.07493462
```

Therefore we fail reject the null hypothesis that the two starting salaries are in the same population. The results might not hold across years because the sample size would change, which means the student t-distribution will change.

4

a)

$$H_0 : E(D_i) = 0$$
$$H_A : \mu \neq 0$$

b)

```
Wage_before = c(8.3, 9.4, 9, 10.5, 11.4, 8.75, 10, 9.5, 10.8, 12.55, 12, 8.65, 7.75, 11.
25, 12.65)
mean(Wage_before)
```

```
## [1] 10.16667
```

```
Wage_after = c(9.25, 9, 9.25, 10, 12, 9.5, 10.25, 9.5, 11.5, 13.1, 11.5, 9, 7.75, 11.5,
13)
mean(Wage_after)
```

```
## [1] 10.40667
```

```
d = Wage_after - Wage_before
```

```
mean(d)
```

```
## [1] 0.24
```

```
sqrt(var(d))/sqrt(15)
```

```
## [1] 0.1164147
```

$$p - \text{value} = 2 \times t_{14}(-\frac{.24}{.12}) = .06$$

$$ we fail to reject at confidence level of $5% $ and at confidence level of 1%

```
pt(-2, 14)
```

```
## [1] 0.03264398
```

d)

```
qt(.975, 14)
```

```
## [1] 2.144787
```

$$t_{14}(0.975) = 2.14$$

The 95% interval is $[.24 - .12 \times 2.14, .24 + .12 \times 2.14]$, or equivalently, $[-.02, .50]$

5

a)

```
Olympics = read.csv("/Users/kevintsukuyo/Documents/Course Files/2022F/Applied Econometri
cs/HW1/Olympics_HW.csv")
```

```
summary(Olympics$medals)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   0.000   1.751   0.000  37.000
```

```
summary(Olympics$athletes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    0.00    2.00   18.17   13.00  230.00
```

```
summary(Olympics$GDP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  0.0110  0.1212  0.3849  1.1691  1.5127 14.5230     156
```

b)

```
b5 = subset(Olympics, select = c('country', 'year', 'medals', 'athletes', 'GDP'))[c(0:
5), ]
b5
```

```
##   country year medals athletes    GDP
## 1 Albania 1980      0        0     NA
## 2 Albania 1984      0        0 0.0641
## 3 Albania 1988      0        0 0.0637
## 4 Albania 1992      0        0 0.0206
## 5 Albania 1994      0        0 0.0587
```

c)

```
c5 = data.frame(matrix(ncol = 2, nrow = length(unique(Olympics$year))))
colnames(c5) = c('Years', 'Numbers')
c5$Years = unique(Olympics$year)
for (i in unique(Olympics$year)){
 c5[c5$Years == i, ]$Numbers =  dim(subset(Olympics, year == i))[1]
}
c5
```

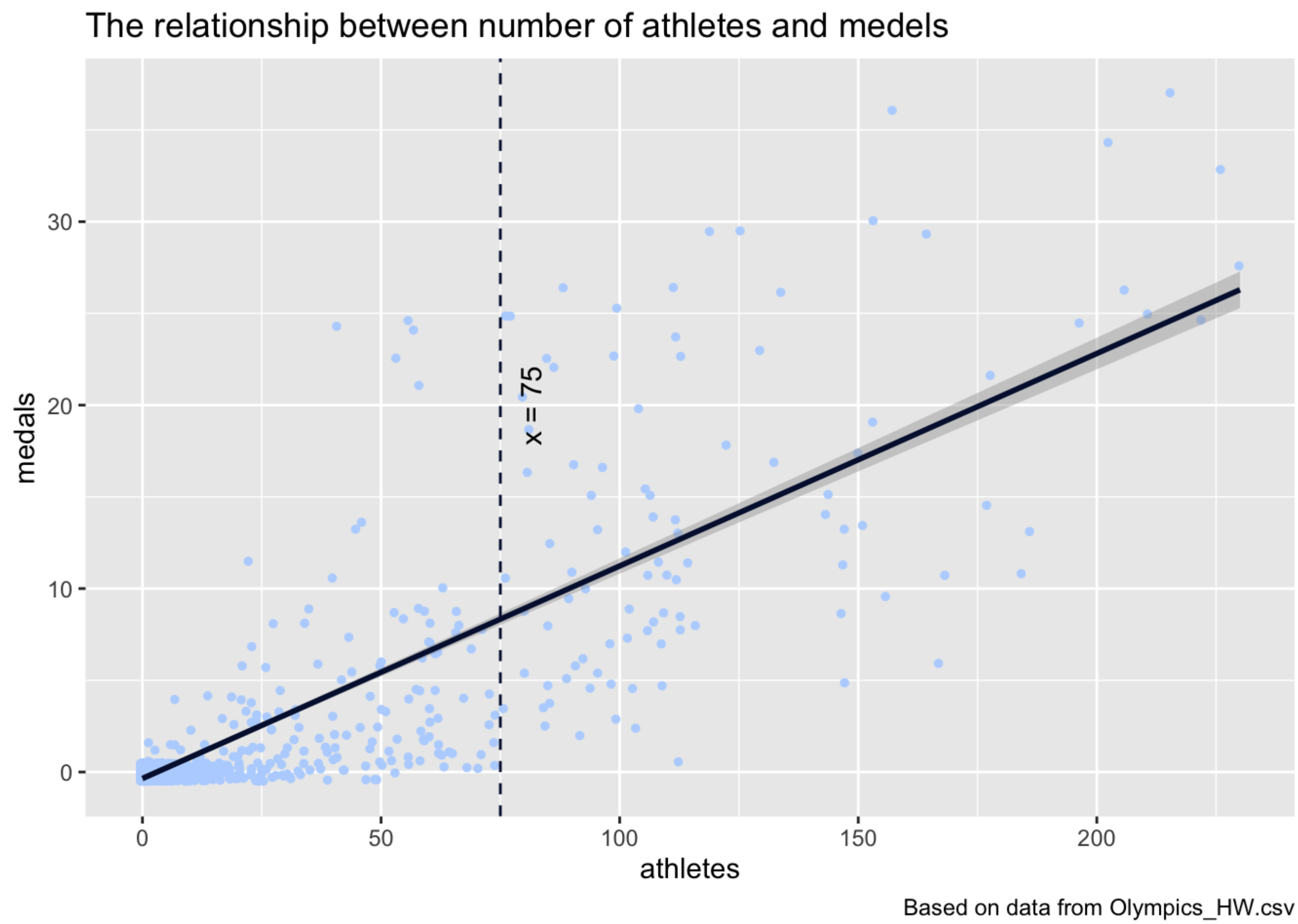
```
##    Years Numbers
## 1   1980     117
## 2   1984     117
## 3   1988     117
## 4   1992     113
## 5   1994     110
## 6   1998     110
## 7   2002     110
## 8   2006     110
## 9   2010     109
## 10  2014     109
```

d)

```
library(ggplot2)
```

```
ggplot(data = Olympics, aes(x = athletes, y = medals))+
  geom_point(size = 1, color = '#b8d5ff', position = position_jitter(height = .5))+
  labs(title = "The relationship between number of athletes and medels", caption = "Base
d on data from Olympics_HW.csv")+
  geom_smooth(method = lm, color = '#05133d')+
  geom_vline(xintercept = 75, color = '#05133d', linetype = 'dashed')+
  annotate(geom = "text", label = "x = 75", x = 75, y = 20, vjust = 2, angle = 90)
```

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




```
ggsave('plot.png')
```

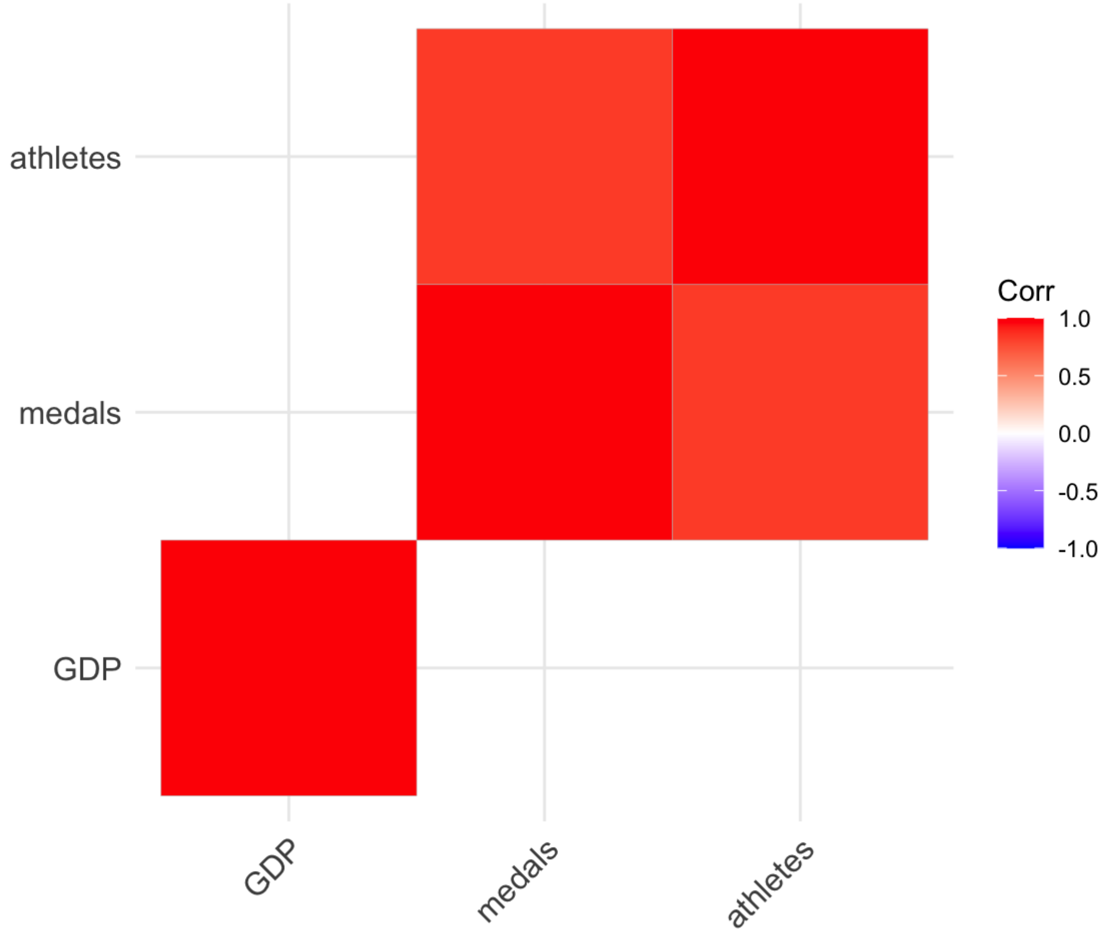
```
## Saving 7 x 5 in image
## `geom_smooth()` using formula = 'y ~ x'
```

The number of athletes and medals are positively correlated, which means there tend to be more medal winners if there are more athletes in a group. For some small athletes group with participants less than approximately 75 there are no medal winners, as the athlete group larger the medal-athlete relationship tends to be more linear.

e)

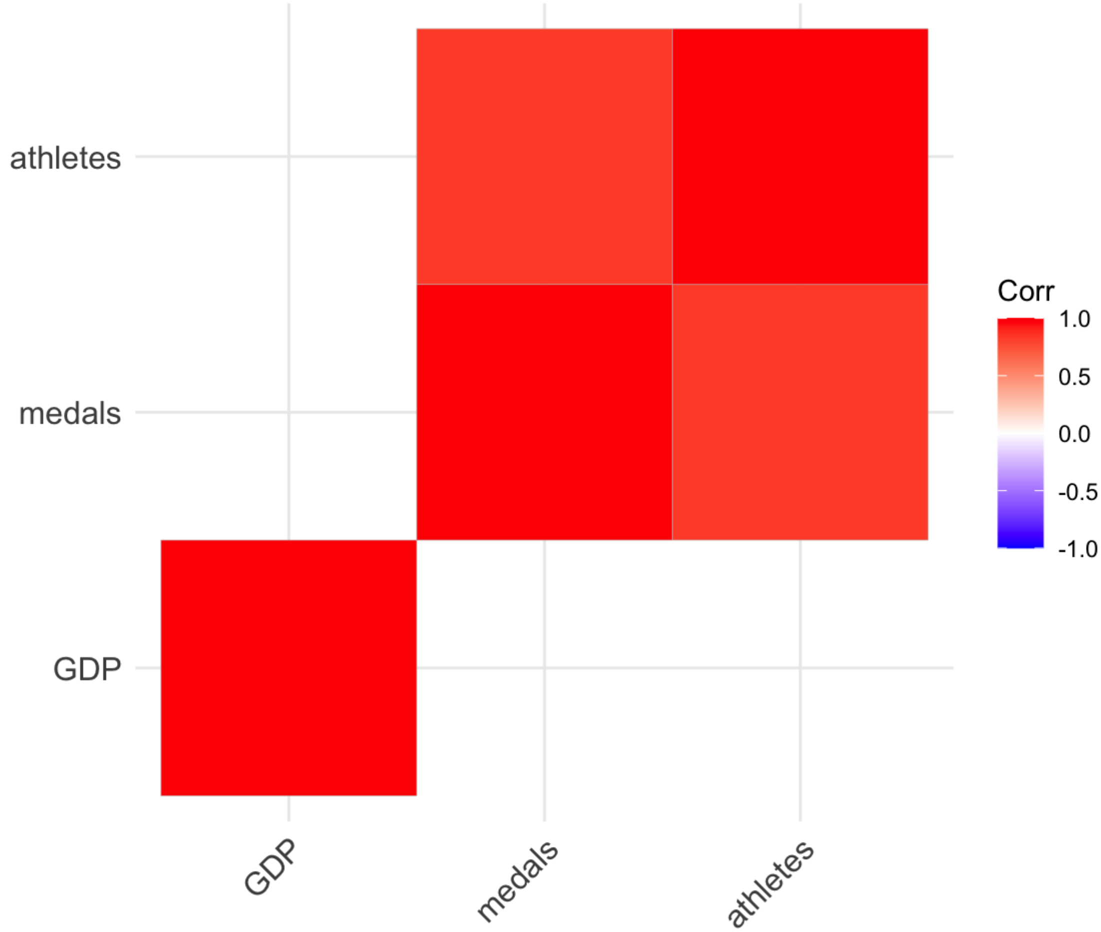
In the last session we find that medals and athletes are positively correlated. It could be explained by athletes are equally good at all events so coutries sending more athletes win more medals. However, it could be true that countries more developed tend to send more athletes and the athletes in the developed countries are trained better so they perform well.

```
library(ggcorrplot)
GDP_athletes_medals = subset(Olympics, select = c('GDP', 'medals', 'athletes'))
ggcorrplot(cor(GDP_athletes_medals))
```



It seems there is no correlation among GDP and other variables, but after observing the dataset I suspect the units are quite different among three variables. So I performed normalization on the variables.

```
GDP_athletes_medals$GDP = scale(GDP_athletes_medals$GDP)
GDP_athletes_medals$athletes = scale(GDP_athletes_medals$athletes)
GDP_athletes_medals$medals = scale(GDP_athletes_medals$medals)
ggcorrplot(cor(GDP_athletes_medals))
```



Yet, the correlation is not significant.

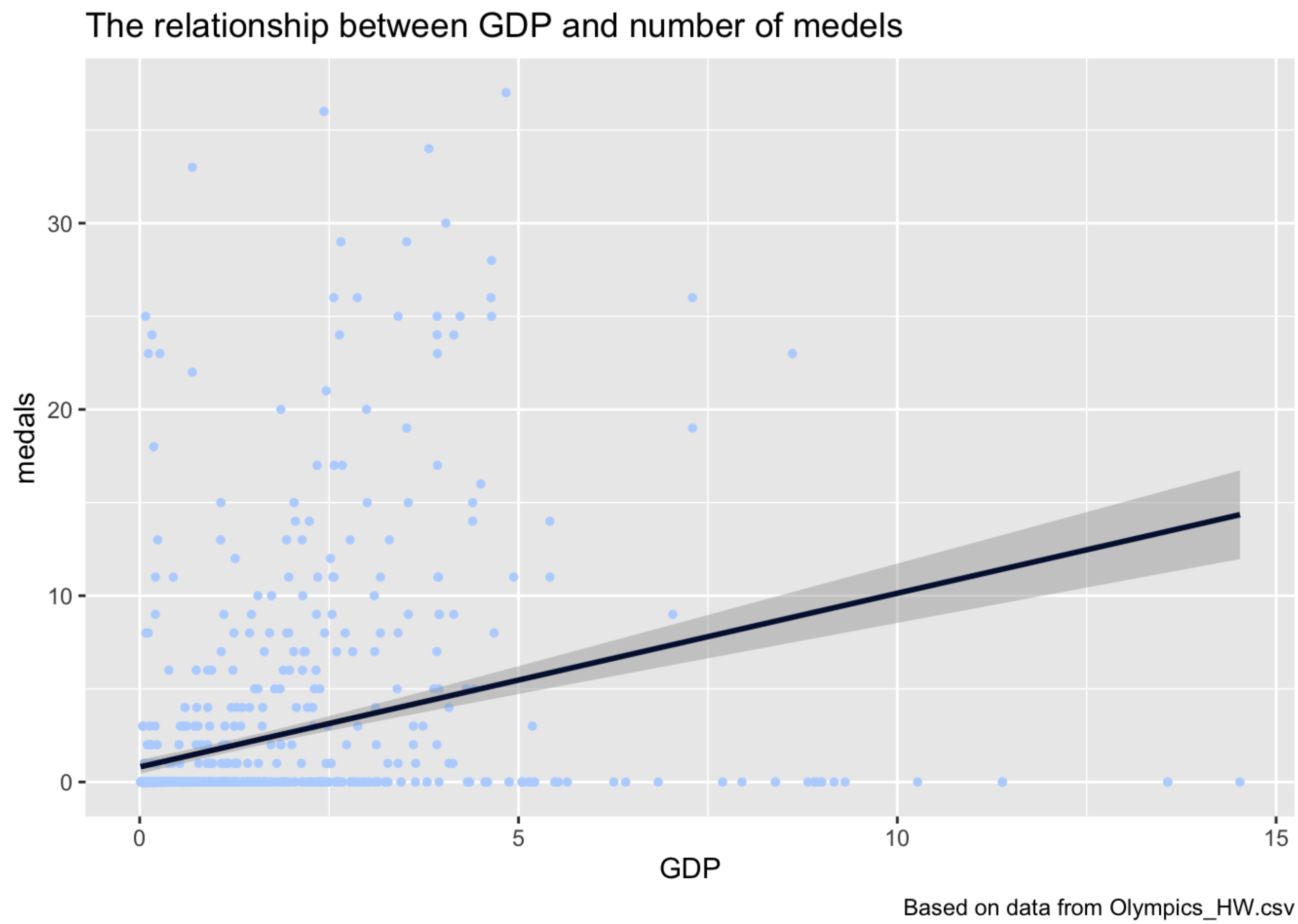
f)

```
ggplot(data = Olympics, aes(x = GDP, y = medals), na.rm = TRUE)+
  geom_point(size = 1, color = '#b8d5ff')+
  labs(title = "The relationship between GDP and number of medels", caption = "Based on
data from Olympics_HW.csv")+
  geom_smooth(method = lm, color = '#05133d')
```

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

```
## Warning: Removed 156 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 156 rows containing missing values (`geom_point()`).
```



There are some missing values and in the GDP data, let us explore the data.

```
GDP_missing = Olympics[is.na(Olympics$GDP), ]
GDP_missing
```

```
##        ID year host                   country temp precipitation elevation
## 1       1 1980    0                   Albania 53.6         143.0      2764
## 21      3 1980    0            American Samoa 84.2         164.1       964
## 22      3 1984    0            American Samoa 84.2         164.1       964
## 23      3 1988    0            American Samoa 84.2         164.1       964
## 24      3 1992    0            American Samoa 84.2         164.1       964
## 25      3 1994    0            American Samoa 84.2         164.1       964
## 26      3 1998    0            American Samoa 84.2         164.1       964
## 27      3 2002    0            American Samoa 84.2         164.1       964
## 28      3 2006    0            American Samoa 84.2         164.1       964
## 29      3 2010    0            American Samoa 84.2         164.1       964
## 30      3 2014    0            American Samoa 84.2         164.1       964
## 39      4 2010    0                   Andorra 42.3          51.0      2946
## 51      6 1980    0                   Armenia 33.1          22.0      4090
## 52      6 1984    0                   Armenia 33.1          22.0      4090
## 53      6 1988    0                   Armenia 33.1          22.0      4090
## 81      9 1980    0                Azerbaijan 43.9          21.0      4485
## 82      9 1984    0                Azerbaijan 43.9          21.0      4485
## 83      9 1988    0                Azerbaijan 43.9          21.0      4485
## 91     10 1980    0                   Belarus 28.2          45.0       346
## 92     10 1984    0                   Belarus 28.2          45.0       346
## 93     10 1988    0                   Belarus 28.2          45.0       346
## 131    14 1980    0    Bosnia and Herzegovina 38.7          68.0      2386
## 132    14 1984    0    Bosnia and Herzegovina 38.7          68.0      2386
## 133    14 1988    0    Bosnia and Herzegovina 38.7          68.0      2386
## 134    14 1992    0    Bosnia and Herzegovina 38.7          68.0      2386
## 160    16 2014    0    British Virgin Islands 79.0          74.2       521
## 191    20 1980    0            Cayman Islands 79.0          71.0        43
## 192    20 1984    0            Cayman Islands 79.0          71.0        43
## 193    20 1988    0            Cayman Islands 79.0          71.0        43
## 194    20 1992    0            Cayman Islands 79.0          71.0        43
## 195    20 1994    0            Cayman Islands 79.0          71.0        43
## 196    20 1998    0            Cayman Islands 79.0          71.0        43
## 197    20 2002    0            Cayman Islands 79.0          71.0        43
## 198    20 2006    0            Cayman Islands 79.0          71.0        43
## 199    20 2010    0            Cayman Islands 79.0          71.0        43
## 200    20 2014    0            Cayman Islands 79.0          71.0        43
## 230    23 2014    0           Chinese Taipei 66.4          83.2      3952
## 251    26 1980    0                  Croatia 37.6          48.6      1831
## 252    26 1984    0                  Croatia 37.6          48.6      1831
## 253    26 1988    0                  Croatia 37.6          48.6      1831
## 271    28 1980    0            Czech Republic 32.7          23.5      1602
## 272    28 1984    0            Czech Republic 32.7          23.5      1602
## 273    28 1988    0            Czech Republic 32.7          23.5      1602
## 281    29 1980    0            Czechoslovakia 32.7          23.5      1602
## 282    29 1984    0            Czechoslovakia 32.7          23.5      1602
## 283    29 1988    0            Czechoslovakia 32.7          23.5      1602
## 318    34 1980    0                  Estonia 29.8          49.0       318
## 319    34 1984    0                  Estonia 29.8          49.0       318
## 320    34 1988    0                  Estonia 29.8          49.0       318
## 321    34 1992    0                  Estonia 29.8          49.0       318
## 328    35 1980    0                 Ethiopia 70.0         280.0      4533
```

```
## 368   39 1980    0                      Georgia 42.6          20.0      5201
## 369   39 1984    0                      Georgia 42.6          20.0      5201
## 370   39 1988    0                      Georgia 42.6          20.0      5201
## 418   44 1980    0                         Guam 86.3         100.6       406
## 419   44 1984    0                         Guam 86.3         100.6       406
## 420   44 1988    0                         Guam 86.3         100.6       406
## 477   51 1992    0                         Iran 43.0          63.1      5671
## 534   57 1980    0                   Kazakhstan 33.3          34.0      6995
## 535   57 1984    0                   Kazakhstan 33.3          34.0      6995
## 536   57 1988    0                   Kazakhstan 33.3          34.0      6995
## 554   59 1980    0                   Kyrgyzstan 37.8          26.0      7439
## 555   59 1984    0                   Kyrgyzstan 37.8          26.0      7439
## 556   59 1988    0                   Kyrgyzstan 37.8          26.0      7439
## 564   60 1980    0                       Latvia 27.9          33.7       312
## 565   60 1984    0                       Latvia 27.9          33.7       312
## 574   61 1980    0                      Lebanon 63.0         190.9      3088
## 575   61 1984    0                      Lebanon 63.0         190.9      3088
## 592   62 2010    0                Liechtenstein 39.7          41.0      2599
## 594   63 1980    0                    Lithuania 25.7          38.9       294
## 595   63 1984    0                    Lithuania 25.7          38.9       294
## 596   63 1988    0                    Lithuania 25.7          38.9       294
## 614   65 1980    0                    Macedonia 40.1          30.0      2764
## 615   65 1984    0                    Macedonia 40.1          30.0      2764
## 616   65 1988    0                    Macedonia 40.1          30.0      2764
## 654   69 1980    0                      Moldova 33.6          36.0       430
## 655   69 1984    0                      Moldova 33.6          36.0       430
## 656   69 1988    0                      Moldova 33.6          36.0       430
## 674   71 1980    0                     Mongolia  3.9           1.1      4374
## 684   72 1980    0                   Montenegro 49.1         192.0      2522
## 685   72 1984    0                   Montenegro 49.1         192.0      2522
## 686   72 1988    0                   Montenegro 49.1         192.0      2522
## 687   72 1992    0                   Montenegro 49.1         192.0      2522
## 688   72 1994    0                   Montenegro 49.1         192.0      2522
## 689   72 1998    0                   Montenegro 49.1         192.0      2522
## 738   78 1980    0                  North Korea 30.6          12.2      2744
## 739   78 1984    0                  North Korea 30.6          12.2      2744
## 740   78 1988    0                  North Korea 30.6          12.2      2744
## 741   78 1992    0                  North Korea 30.6          12.2      2744
## 742   78 1994    0                  North Korea 30.6          12.2      2744
## 743   78 1998    0                  North Korea 30.6          12.2      2744
## 744   78 2002    0                  North Korea 30.6          12.2      2744
## 745   78 2006    0                  North Korea 30.6          12.2      2744
## 746   78 2010    0                  North Korea 30.6          12.2      2744
## 747   78 2014    0                  North Korea 30.6          12.2      2744
## 798   84 1980    0                       Poland 32.2          21.0      2499
## 799   84 1984    0                       Poland 32.2          21.0      2499
## 828   87 1980    0                      Romania 34.7          40.0      2544
## 829   87 1984    0                      Romania 34.7          40.0      2544
## 838   88 1980    0                       Russia 25.0          52.0      5633
## 839   88 1984    0                       Russia 25.0          52.0      5633
## 840   88 1988    0                       Russia 25.0          52.0      5633
## 848   89 1980    0                   San Marino 45.5          59.0       755
```

```
## 849   89 1984    0                  San Marino 45.5          59.0       755
## 850   89 1988    0                  San Marino 45.5          59.0       755
## 851   89 1992    0                  San Marino 45.5          59.0       755
## 852   89 1994    0                  San Marino 45.5          59.0       755
## 853   89 1998    0                  San Marino 45.5          59.0       755
## 856   89 2010    0                  San Marino 45.5          59.0       755
## 868   91 1980    0                      Serbia 40.3          46.9      2169
## 869   91 1984    0                      Serbia 40.3          46.9      2169
## 870   91 1988    0                      Serbia 40.3          46.9      2169
## 871   91 1992    0                      Serbia 40.3          46.9      2169
## 872   91 1994    0                      Serbia 40.3          46.9      2169
## 878   92 1980    0       Serbia and Montenegro   NA            NA        NA
## 879   92 1984    0       Serbia and Montenegro   NA            NA        NA
## 880   92 1988    0       Serbia and Montenegro   NA            NA        NA
## 881   92 1992    0       Serbia and Montenegro   NA            NA        NA
## 882   92 1994    0       Serbia and Montenegro   NA            NA        NA
## 886   93 1980    0                    Slovakia 36.9          39.0      2655
## 896   94 1980    0                    Slovenia 37.4          71.0      2864
## 897   94 1984    0                    Slovenia 37.4          71.0      2864
## 898   94 1988    0                    Slovenia 37.4          71.0      2864
## 926   97 1980    0                Soviet Union   NA            NA        NA
## 927   97 1984    0                Soviet Union   NA            NA        NA
## 928   97 1988    0                Soviet Union   NA            NA        NA
## 969  102 1980    0                  Tajikistan 48.9          66.3      7495
## 970  102 1984    0                  Tajikistan 48.9          66.3      7495
## 971  102 1988    0                  Tajikistan 48.9          66.3      7495
## 989  104 1980    0                 Timor-Leste 75.9          44.9      2963
## 990  104 1984    0                 Timor-Leste 75.9          44.9      2963
## 991  104 1988    0                 Timor-Leste 75.9          44.9      2963
## 992  104 1992    0                 Timor-Leste 75.9          44.9      2963
## 993  104 1994    0                 Timor-Leste 75.9          44.9      2963
## 994  104 1998    0                 Timor-Leste 75.9          44.9      2963
## 1043 109 1994    0           US Virgin Islands 86.0          48.0       474
## 1044 109 1998    0           US Virgin Islands 86.0          48.0       474
## 1045 109 2002    0           US Virgin Islands 86.0          48.0       474
## 1046 109 2006    0           US Virgin Islands 86.0          48.0       474
## 1047 109 2010    0           US Virgin Islands 86.0          48.0       474
## 1048 109 2014    0           US Virgin Islands 86.0          48.0       474
## 1049 110 1980    0                     Ukraine 30.4          36.0      2061
## 1050 110 1984    0                     Ukraine 30.4          36.0      2061
## 1059 111 1992    0 Unified Team (Former Soviet)   NA            NA        NA
## 1080 114 1980    0                  Uzbekistan 43.2          57.8      4301
## 1081 114 1984    0                  Uzbekistan 43.2          57.8      4301
## 1082 114 1988    0                  Uzbekistan 43.2          57.8      4301
## 1100 116 1980    0                West Germany   NA            NA        NA
## 1101 116 1984    0                West Germany   NA            NA        NA
## 1102 116 1988    0                West Germany   NA            NA        NA
## 1103 117 1980    0                  Yugoslavia 38.7          68.0      2386
## 1104 117 1984    1                  Yugoslavia 38.7          68.0      2386
## 1105 117 1988    0                  Yugoslavia 38.7          68.0      2386
## 1110 117 2006    0                  Yugoslavia 38.7          68.0      2386
## 1111 117 2010    0                  Yugoslavia 38.7          68.0      2386
```

```
## 1112 117 2014    0                    Yugoslavia 38.7         68.0      2386
##      gold silver bronze population GDP participate medals athletes time
## 1       0      0      0   2.734776  NA           0      0        0    1
## 21      0      0      0   0.032456  NA           0      0        0    1
## 22      0      0      0   0.037687  NA           0      0        0    2
## 23      0      0      0   0.044049  NA           0      0        0    3
## 24      0      0      0   0.049597  NA           0      0        0    4
## 25      0      0      0   0.051807  NA           1      0        2    5
## 26      0      0      0   0.055899  NA           0      0        0    6
## 27      0      0      0   0.058729  NA           0      0        0    7
## 28      0      0      0   0.058652  NA           0      0        0    8
## 29      0      0      0   0.055636  NA           0      0        0    9
## 30      0      0      0   0.055128  NA           0      0        0   10
## 39      0      0      0   0.077907  NA           1      0        6    9
## 51      0      0      0   3.096298  NA           0      0        0    1
## 52      0      0      0   3.287588  NA           0      0        0    2
## 53      0      0      0   3.510439  NA           0      0        0    3
## 81      0      0      0   6.163990  NA           0      0        0    1
## 82      0      0      0   6.568857  NA           0      0        0    2
## 83      0      0      0   6.994139  NA           0      0        0    3
## 91      0      0      0   9.643000  NA           0      0        0    1
## 92      0      0      0   9.910000  NA           0      0        0    2
## 93      0      0      0  10.140000  NA           0      0        0    3
## 131     0      0      0   4.099903  NA           0      0        0    1
## 132     0      0      0   4.263393  NA           0      0        0    2
## 133     0      0      0   4.564265  NA           0      0        0    3
## 134     0      0      0   4.143068  NA           0      0        0    4
## 160     0      0      0         NA  NA           1      0        1   10
## 191     0      0      0   0.016164  NA           0      0        0    1
## 192     0      0      0   0.018543  NA           0      0        0    2
## 193     0      0      0   0.022539  NA           0      0        0    3
## 194     0      0      0   0.027402  NA           0      0        0    4
## 195     0      0      0   0.030055  NA           0      0        0    5
## 196     0      0      0   0.037742  NA           0      0        0    6
## 197     0      0      0   0.044742  NA           0      0        0    7
## 198     0      0      0   0.050026  NA           0      0        0    8
## 199     0      0      0   0.055509  NA           1      0        1    9
## 200     0      0      0   0.057570  NA           1      0        1   10
## 230     0      0      0         NA  NA           1      0        3   10
## 251     0      0      0   4.588000  NA           0      0        0    1
## 252     0      0      0   4.680000  NA           0      0        0    2
## 253     0      0      0   4.757000  NA           0      0        0    3
## 271     0      0      0  10.304193  NA           0      0        0    1
## 272     0      0      0  10.330213  NA           0      0        0    2
## 273     0      0      0  10.355276  NA           0      0        0    3
## 281     0      0      1  10.304193  NA           1      1       41    1
## 282     0      2      4  10.330213  NA           1      6       50    2
## 283     0      1      2  10.355276  NA           1      3       62    3
## 318     0      0      0   1.477219  NA           0      0        0    1
## 319     0      0      0   1.518617  NA           0      0        0    2
## 320     0      0      0   1.561900  NA           0      0        0    3
## 321     0      0      0   1.533091  NA           1      0       19    4
```

```
## 328      0      0      0  35.241209  NA           0      0       0    1
## 368      0      0      0   4.467700  NA           0      0       0    1
## 369      0      0      0   4.622200  NA           0      0       0    2
## 370      0      0      0   4.790700  NA           0      0       0    3
## 418      0      0      0   0.104131  NA           0      0       0    1
## 419      0      0      0   0.113961  NA           0      0       0    2
## 420      0      0      0   0.124677  NA           1      0       1    3
## 477      0      0      0  58.307457  NA           0      0       0    4
## 534      0      0      0  14.898332  NA           0      0       0    1
## 535      0      0      0  15.599590  NA           0      0       0    2
## 536      0      0      0  16.130832  NA           0      0       0    3
## 554      0      0      0   3.617400  NA           0      0       0    1
## 555      0      0      0   3.916400  NA           0      0       0    2
## 556      0      0      0   4.218400  NA           0      0       0    3
## 564      0      0      0   2.511701  NA           0      0       0    1
## 565      0      0      0   2.562047  NA           0      0       0    2
## 574      0      0      0   2.605294  NA           1      0       3    1
## 575      0      0      0   2.667229  NA           1      0       4    2
## 592      0      0      0   0.036120  NA           1      0       7    9
## 594      0      0      0   3.413202  NA           0      0       0    1
## 595      0      0      0   3.514205  NA           0      0       0    2
## 596      0      0      0   3.655049  NA           0      0       0    3
## 614      0      0      0   1.895727  NA           0      0       0    1
## 615      0      0      0   1.939913  NA           0      0       0    2
## 616      0      0      0   1.995338  NA           0      0       0    3
## 654      0      0      0   3.397000  NA           0      0       0    1
## 655      0      0      0   3.536000  NA           0      0       0    2
## 656      0      0      0   3.659000  NA           0      0       0    3
## 674      0      0      0   1.689621  NA           1      0       3    1
## 684      0      0      0   0.579088  NA           0      0       0    1
## 685      0      0      0   0.591105  NA           0      0       0    2
## 686      0      0      0   0.607814  NA           0      0       0    3
## 687      0      0      0   0.616630  NA           0      0       0    4
## 688      0      0      0   0.615521  NA           0      0       0    5
## 689      0      0      0   0.611947  NA           0      0       0    6
## 738      0      0      0  17.372172  NA           0      0       0    1
## 739      0      0      0  18.488002  NA           1      0       6    2
## 740      0      0      0  19.610518  NA           1      0       6    3
## 741      0      0      1  20.838090  NA           1      1      20    4
## 742      0      0      0  21.478552  NA           0      0       0    5
## 743      0      0      0  22.444993  NA           1      0       8    6
## 744      0      0      0  23.248059  NA           0      0       0    7
## 745      0      0      0  23.969917  NA           1      0       6    8
## 746      0      0      0  24.500520  NA           1      0       2    9
## 747      0      0      0  24.763188  NA           0      0       0   10
## 798      0      0      0  35.574150  NA           1      0      30    1
## 799      0      0      0  36.904134  NA           1      0      30    2
## 828      0      0      0  22.242653  NA           1      0      35    1
## 829      0      0      0  22.655940  NA           1      0      19    2
## 838      0      0      0 139.010000  NA           0      0       0    1
## 839      0      0      0 142.745000  NA           0      0       0    2
## 840      0      0      0 146.857000  NA           0      0       0    3
```

```
## 848      0      0      0   0.021397  NA            0      0        0    1
## 849      0      0      0   0.022514  NA            1      0        3    2
## 850      0      0      0   0.023571  NA            1      0        5    3
## 851      0      0      0   0.024766  NA            1      0        3    4
## 852      0      0      0   0.025402  NA            1      0        3    5
## 853      0      0      0   0.026321  NA            0      0        0    6
## 856      0      0      0   0.030861  NA            1      0        1    9
## 868      0      0      0   0.000000  NA            0      0        0    1
## 869      0      0      0   0.000000  NA            0      0        0    2
## 870      0      0      0   0.000000  NA            0      0        0    3
## 871      0      0      0   7.646424  NA            0      0        0    4
## 872      0      0      0   7.734639  NA            0      0        0    5
## 878      0      0      0   0.000000  NA            0      0        0    1
## 879      0      0      0   0.000000  NA            0      0        0    2
## 880      0      0      0   0.000000  NA            0      0        0    3
## 881      0      0      0   7.646424  NA            0      0        0    4
## 882      0      0      0   7.734639  NA            0      0        0    5
## 886      0      0      0   4.979815  NA            0      0        0    1
## 896      0      0      0   1.901315  NA            0      0        0    1
## 897      0      0      0   1.932154  NA            0      0        0    2
## 898      0      0      0   1.995196  NA            0      0        0    3
## 926     10      6      6   4.701961  NA            1     22       86    1
## 927      6     10      9   5.315479  NA            1     25       99    2
## 928     11      9      9   5.738763  NA            1     29      119    3
## 969      0      0      0   3.917642  NA            0      0        0    1
## 970      0      0      0   4.400743  NA            0      0        0    2
## 971      0      0      0   5.008827  NA            0      0        0    3
## 989      0      0      0   0.580730  NA            0      0        0    1
## 990      0      0      0   0.638938  NA            0      0        0    2
## 991      0      0      0   0.711794  NA            0      0        0    3
## 992      0      0      0   0.804803  NA            0      0        0    4
## 993      0      0      0   0.853716  NA            0      0        0    5
## 994      0      0      0   0.853069  NA            0      0        0    6
## 1043     0      0      0   0.107317  NA            1      0        8    5
## 1044     0      0      0   0.108535  NA            1      0        7    6
## 1045     0      0      0   0.108208  NA            1      0        8    7
## 1046     0      0      0   0.107700  NA            1      0        1    8
## 1047     0      0      0   0.106267  NA            0      0        0    9
## 1048     0      0      0   0.105275  NA            1      0        1   10
## 1049     0      0      0  49.973757  NA            0      0        0    1
## 1050     0      0      0  50.754000  NA            0      0        0    2
## 1059     9      6      8         NA  NA            1     23      129    4
## 1080     0      0      0  16.026812  NA            0      0        0    1
## 1081     0      0      0  17.778956  NA            0      0        0    2
## 1082     0      0      0  19.606739  NA            0      0        0    3
## 1100     0      2      3   0.000000  NA            1      5       80    1
## 1101     2      1      1   0.000000  NA            1      4       84    2
## 1102     2      4      2   0.000000  NA            1      8      116    3
## 1103     0      0      0   7.906977  NA            1      0       15    1
## 1104     0      1      0   9.276622  NA            1      1       71    2
## 1105     0      2      1  10.815614  NA            1      3       23    3
## 1110     0      0      0         NA  NA            0      0        0    8
```

```
## 1111    0      0      0  22.763008  NA          0      0        0    9
## 1112    0      0      0         NA  NA          0      0        0   10
```

```
summary(GDP_missing$medals)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0000  0.0000  0.8397  0.0000 29.0000
```

We find that more than 75% countries with a missing GDP have no medal. Also, there are some outliers with $\mathrm{GDP} > 10$ but gaining few medals, let's watch the data.

```
GDP_valid = Olympics[!is.na(Olympics$GDP), ]
GDP_outlier = GDP_valid[GDP_valid$GDP > 4 & GDP_valid$medals < 2, ]
```

```
dim(GDP_outlier)
```

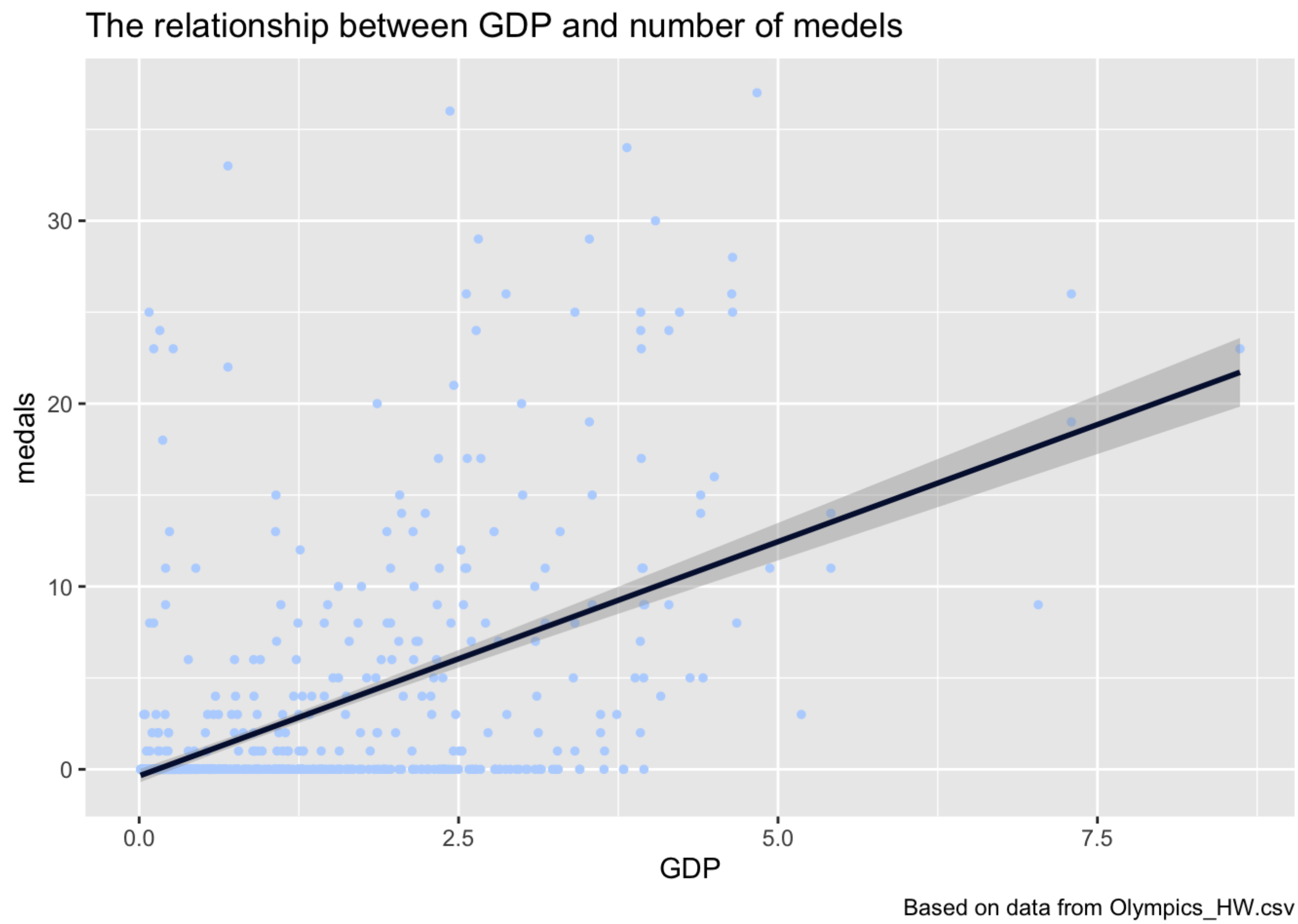
```
## [1] 39 16
```

There are 39 observations out of 966 valid observations (GDP captured) being outliers, which can be droped. This is because in some small countries with pretty low or no tax, some big international firms set headquarters there to avoid high tax which leads to a ridiculous high GDP.

```
GDP_trimmed = GDP_valid[!(GDP_valid$GDP > 4 & GDP_valid$medals <2), ]
```

```
ggplot(data = GDP_trimmed, aes(x = GDP, y = medals), na.rm = TRUE)+
  geom_point(size = 1, color = '#b8d5ff')+
  labs(title = "The relationship between GDP and number of medels", caption = "Based on data from Olympics_HW.csv")+
  geom_smooth(method = lm, color = '#05133d')
```

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



After trimming, our model performs better when GDP gets larger. As a result, GDP data are positively correlated to the number of medals. It infers that athletes in countries more developed are trained better and perform better in competitions.

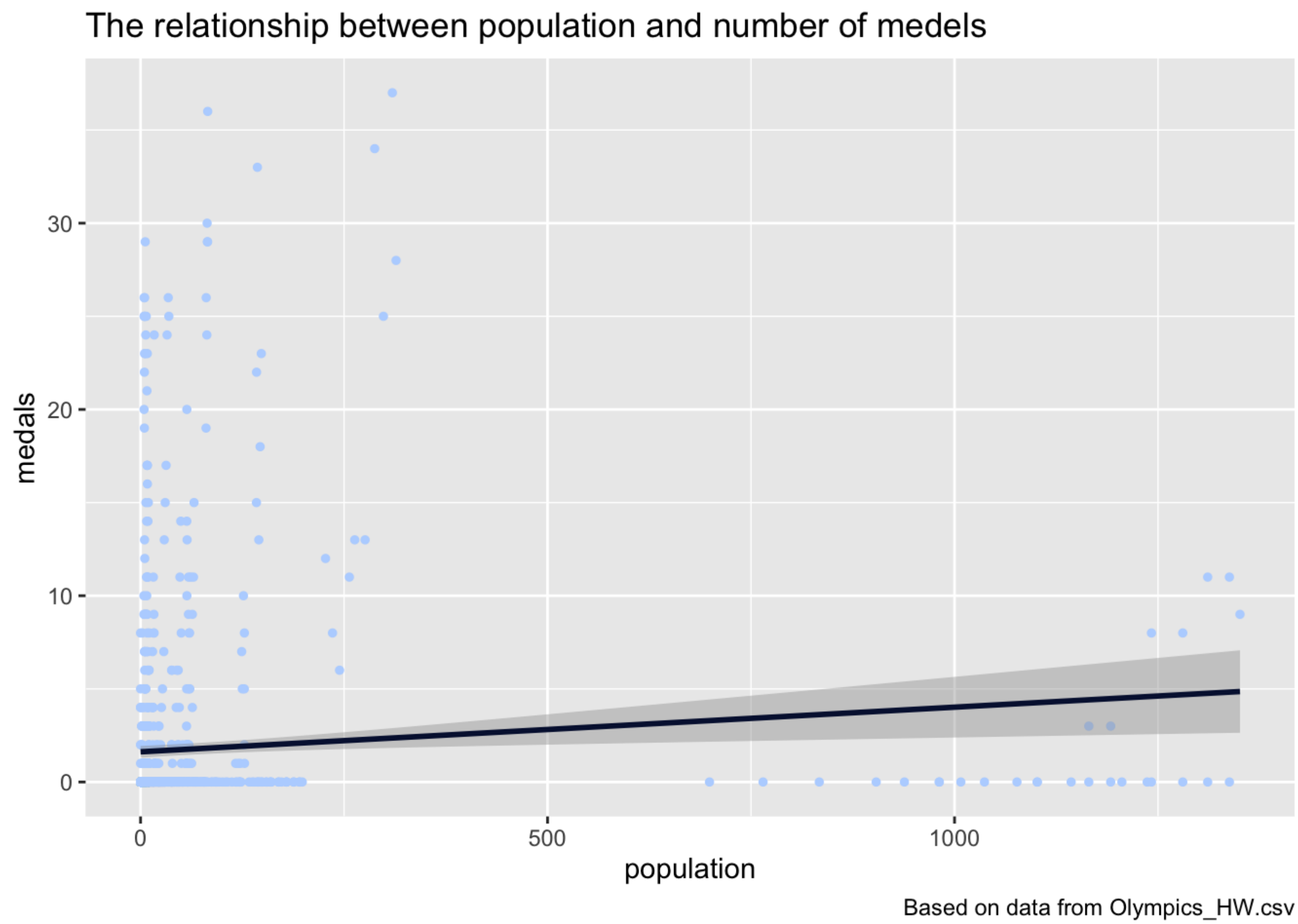
g)

```
ggplot(data = Olympics, aes(x = population, y = medals), na.rm = TRUE)+
  geom_point(size = 1, color = '#b8d5ff')+
  labs(title = "The relationship between population and number of medels", caption = "Ba
sed on data from Olympics_HW.csv")+
  geom_smooth(method = lm, color = '#05133d')
```

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

```
## Warning: Removed 9 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 9 rows containing missing values (`geom_point()`).
```



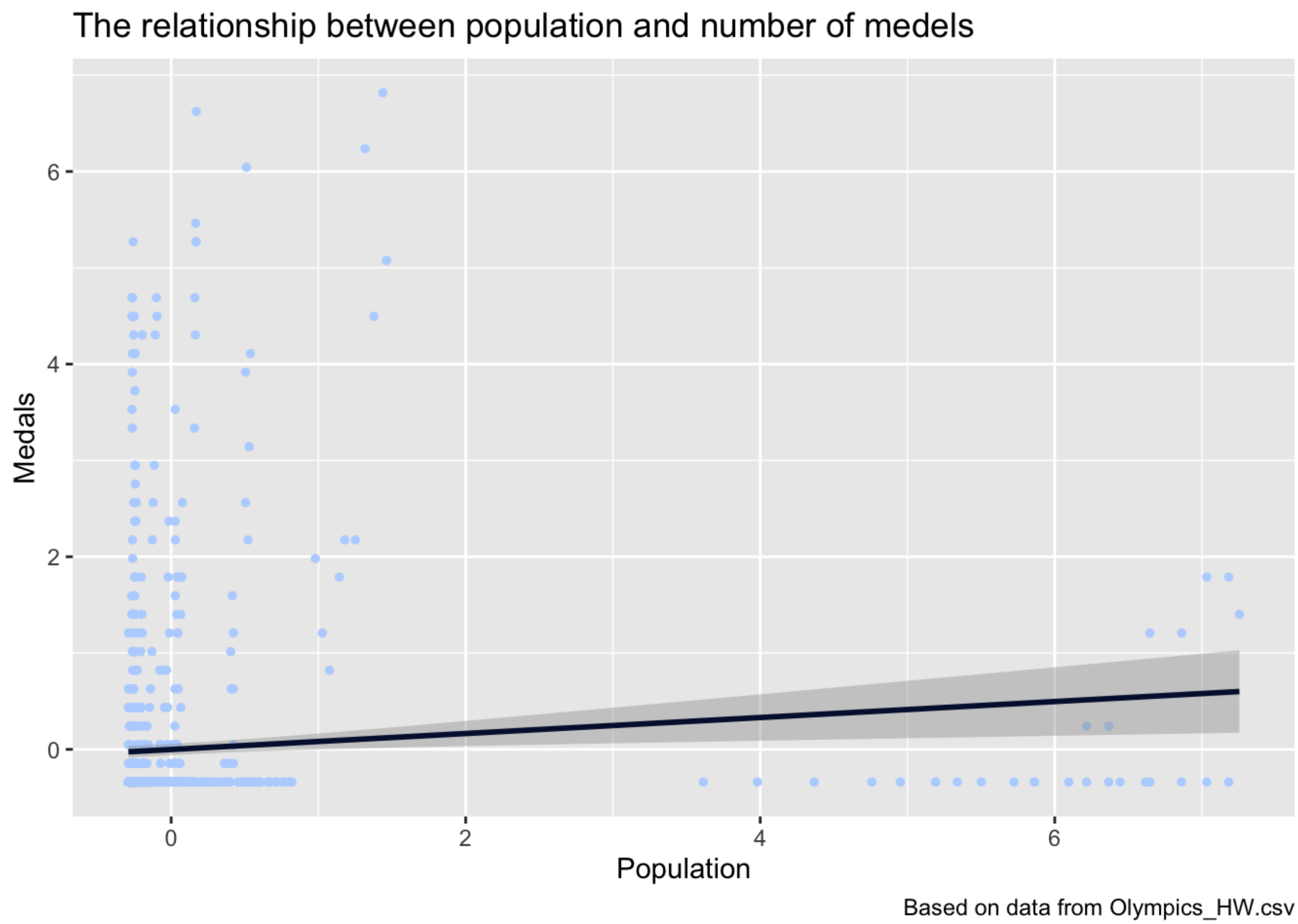
Intuitively, there could be some relationship between population and number of medals because when there is a large population there might be more people perform well on sports leading to a higher number in medals. However, because the measurements between medals and population differs a lot, the relationship is not significant.

```
g5 = as.data.frame(cbind(scale(Olympics$population), scale(Olympics$medals)))
colnames(g5) = c('Population', 'Medals')
ggplot(data = g5, aes(x = Population, y = Medals), na.rm = TRUE)+
  geom_point(size = 1, color = '#b8d5ff')+
  labs(title = "The relationship between population and number of medels", caption = "Ba
sed on data from Olympics_HW.csv")+
  geom_smooth(method = lm, color = '#05133d')
```

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

```
## Warning: Removed 9 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 9 rows containing missing values (`geom_point()`).
```



It can be found by the plot that population and methods are not linearly correlated. There are two groups mainly devided by population, for the group with lower population teams tend to earn more madels.

```
unique(Olympics[Olympics$medals >10 & Olympics$population < 500, 'country'])
```

```
##  [1] "Austria"       "Canada"        "East Germany"  "Finland"
##  [5] "France"        "Germany"       "Italy"         "Netherlands"
##  [9] "Norway"        "Russia"        "South Korea"   "Soviet Union"
## [13] "Sweden"        "Switzerland"   NA              "United States"
```

```
unique(Olympics[Olympics$medals >10 & Olympics$population > 500, 'country'])
```

```
## [1] "China" NA
```

The countries earn more medals in each cohort are all countries more developed.

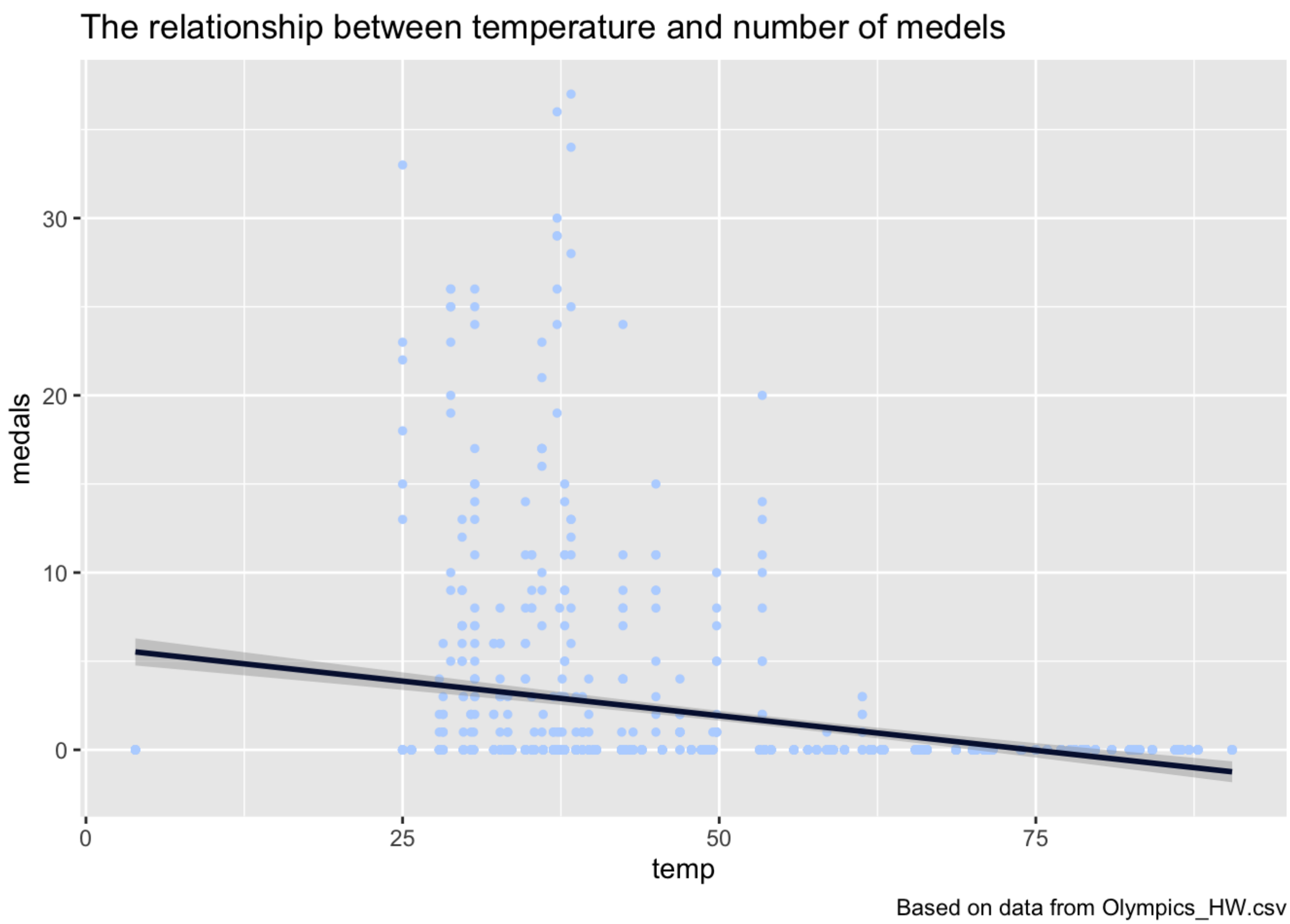
h)

```
ggplot(data = Olympics, aes(x = temp, y = medals), na.rm = TRUE)+
  geom_point(size = 1, color = '#b8d5ff')+
  labs(title = "The relationship between temperature and number of medels", caption = "B
ased on data from Olympics_HW.csv")+
  geom_smooth(method = lm, color = '#05133d')
```

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

```
## Warning: Removed 22 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 22 rows containing missing values (`geom_point()`).
```



For countries with highest temperature too high or too low, it is harder for them to win medals.