Stat 7350 - Assignment1

Xuejing Jiang

March 17, 2019

Contents

1. Interesting question:

Are the top ranked source countries are the countries more threatened by the invasive species?

Paini, Sheppard, Cook and all (2016) said that "Exactly one-half (10) of the countries ranked in the top 20 source countries were also ranked in the top 20 for threatened countries." I saw a potential positvie correlation between the invasion costs(link to the literature).

In this paper, they outlined an important measure, invasion cost, to quantify the economic cost of the invasive species. Specifically, invasin costs are calcuted for both threatened countries and source countries. For the threatened countries, the toal invasion cost for each country, TIC_t, was calculated by summing up the cost associated with all invasive species' impact on domestic crops (see equation [5,6]). While the total invasion cost from each source country ,TIC_s, was calculated by summing up the cost of source country's invasive species impacted on the crops in the threatened countries (see equation [8,9]).

Here, I am interested in the association between the invasion cost applied on threatened countries and the invasion cost that source countries can impose on other countries. In order to do so, I have followed a workflow of data cleaning , data visualization and possible interpretations of the plots as shown in the following sessions.

2. Interact with the data

2.1 Load in the data

```
# loading packages -------------------------------------------------------
suppressMessages(library("here"))
suppressMessages(library("tidyverse"))
suppressMessages(library("gridExtra"))
suppressMessages(library("sqldf"))
suppressMessages(library("Hmisc")) # for using %nin%
suppressMessages(library("skimr"))
suppressMessages(library("ggExtra")) # plot marginal histogram, density or boxplots
suppressMessages(library("scales")) # plot marginal histogram, density or boxplots
```

```
# import data ------------------------------------------------------------
setwd("..")
A1_wd <- getwd()
A1_wd
```

```
## [1] "C:/Users/ZhangYang/OneDrive/Academic Life/Topics_Community_Health_Science/STAT_7350_Stat_analys
```

```
datadir <- paste(A1_wd, "data", sep = "/")
datadir
```

```
## [1] "C:/Users/ZhangYang/OneDrive/Academic Life/Topics_Community_Health_Science/STAT_7350_Stat_analys
```

```
table1 <- read_csv(paste(datadir, "table_1.csv", sep = "/"))  #sorted by invasion_threat
table2 <- read_csv(paste(datadir, "table_2.csv", sep = "/"))  %>%  #sorted by invasion_cost
    rename(invasion_cost_threatCountry = invasion_cost)
table3 <- read_csv(paste(datadir, "table_3.csv", sep = "/"))  #sorted by invasion_gdp_proportion
table4 <- read_csv(paste(datadir, "table_4.csv", sep = "/"))  %>% #sorted by invasion_cost (source coun
    rename(invasion_cost_sourceCountry = invasion_cost)
invasive_species <- read_csv(paste(datadir, "table_6.csv", sep = "/")) #invasive species and impact per
africa_species <- read_csv(paste(datadir, "africa_species.csv", sep = "/"))
```

2.2 Clean up the datasets

Check if there are any duplicates in the datasets and remove all the duplicated records.

table_1:

```
# clean duplicated country in datasets -----------------------------------
(table1_count <- table1 %>%
    count(country) %>%
    filter(n>1) %>% as.tibble())
```

```
## # A tibble: 0 x 2
## # ... with 2 variables: country <chr>, n <int>
```

```
# no duplicates in table1
rm(table1_count)
```

table_2:

```
(table2_count <- table2 %>%
    count(country) %>%
```

```r
    filter(n>1) %>% as.tibble())
```

```
## # A tibble: 1 x 2
##   country     n
##   <chr>   <int>
## 1 Guinea      2
```

```r
(table2_dup <- table2[table2$country == table2_count$country,]) # or replace == by %in%
```

```
## # A tibble: 2 x 3
##   country invasion_cost_threatCountry  rank
##   <chr>                         <dbl> <dbl>
## 1 Guinea                    977500000    60
## 2 Guinea                    114300000   107
```

```r
# 1 duplicated country: Guinea
(table2_nodup <- distinct(table2, country, .keep_all=T)) # keep 1 copy of Guinea
```

```
## # A tibble: 123 x 3
##    country   invasion_cost_threatCountry  rank
##    <chr>                           <dbl> <dbl>
##  1 China                    117290000000     1
##  2 USA                       70381000000     2
##  3 Brazil                    33760000000     3
##  4 India                     33065000000     4
##  5 Japan                     23490000000     5
##  6 Korea                     14349000000     6
##  7 Turkey                    13267000000     7
##  8 Argentina                 13204000000     8
##  9 France                    12532000000     9
## 10 Mexico                    11277000000    10
## # ... with 113 more rows
```

```r
(table2_nodup <- table2[!table2$country == table2_count$country,]) # remove both copy of Guinea
```

```
## # A tibble: 122 x 3
##    country   invasion_cost_threatCountry  rank
##    <chr>                           <dbl> <dbl>
##  1 China                    117290000000     1
##  2 USA                       70381000000     2
##  3 Brazil                    33760000000     3
##  4 India                     33065000000     4
##  5 Japan                     23490000000     5
##  6 Korea                     14349000000     6
##  7 Turkey                    13267000000     7
##  8 Argentina                 13204000000     8
##  9 France                    12532000000     9
## 10 Mexico                    11277000000    10
## # ... with 112 more rows
```

```r
rm(table2, table2_count, table2_dup)
```

table_3:

```r
(table3_count <- table3 %>%
    count(country) %>%
    filter(n>1) %>% as.tibble())
```

```
## # A tibble: 1 x 2
##   country     n
##   <chr>   <int>
## 1 Guinea      2
```

```r
(table3_dup <- table3[table3$country == table3_count$country,]) # or replace == by %in%
```

```
## # A tibble: 2 x 5
##   country invasion_cost   gdp_mean gdp_proportion  rank
##   <chr>           <dbl>      <dbl>          <dbl> <dbl>
## 1 Guinea      978000000 3380000000          0.289     3
## 2 Guinea      114000000  513000000          0.223     4
```

```r
# 1 duplicated country: Guinea
(table3_nodup <- table3[table3$country != table3_count$country,])
```

```
## # A tibble: 122 x 5
##    country    invasion_cost    gdp_mean gdp_proportion  rank
##    <chr>              <dbl>       <dbl>          <dbl> <dbl>
##  1 Malawi        1071000000  3000000000          0.357     1
##  2 Burundi        398000000  1121000000          0.355     2
##  3 Mozambique    1218000000  6423000000          0.190     5
##  4 Madagascar    1074000000  5842000000          0.184     6
##  5 Cambodia      1121000000  6487000000          0.173     7
##  6 Nepal         1411000000  8411000000          0.168     8
##  7 Laos           508000000  3134000000          0.162     9
##  8 Ethiopia      2312000000 14344000000          0.161    10
##  9 Vietnam       7490000000 55702000000          0.134    11
## 10 Moldova        388000000  3130000000          0.124    12
## # ... with 112 more rows
```

```r
rm(table3, table3_count, table3_dup)
```

table_4:

```r
(table4_count <- table4 %>%
    count(country) %>%
    filter(n>1) %>% as.tibble())
```

```
## # A tibble: 1 x 2
##   country     n
##   <chr>   <int>
## 1 Guinea      2
```

```r
(table4_dup <- table4[table4$country == table4_count$country,]) # or replace == by %in%
```

```
## # A tibble: 2 x 3
##   country invasion_cost_sourceCountry  rank
##   <chr>                         <dbl> <dbl>
## 1 Guinea                     47400000    97
## 2 Guinea                      1800000   122
```

```r
# 1 duplicated country: Guinea
(table4_nodup <- table4[table4$country != table4_count$country,])
```

```
## # A tibble: 122 x 3
```

```
##    country invasion_cost_sourceCountry  rank
##    <chr>                         <dbl> <dbl>
##  1 China                  222590000000     1
##  2 USA                    181730000000     2
##  3 Japan                  120750000000     3
##  4 Germany                 85864000000     4
##  5 Italy                   44228000000     5
##  6 France                  38159000000     6
##  7 Korea                   37620000000     7
##  8 India                   36913000000     8
##  9 Russian                 34336000000     9
## 10 United                  25670000000    10
## # ... with 112 more rows
```

```r
rm(table4, table4_count, table4_dup)
```

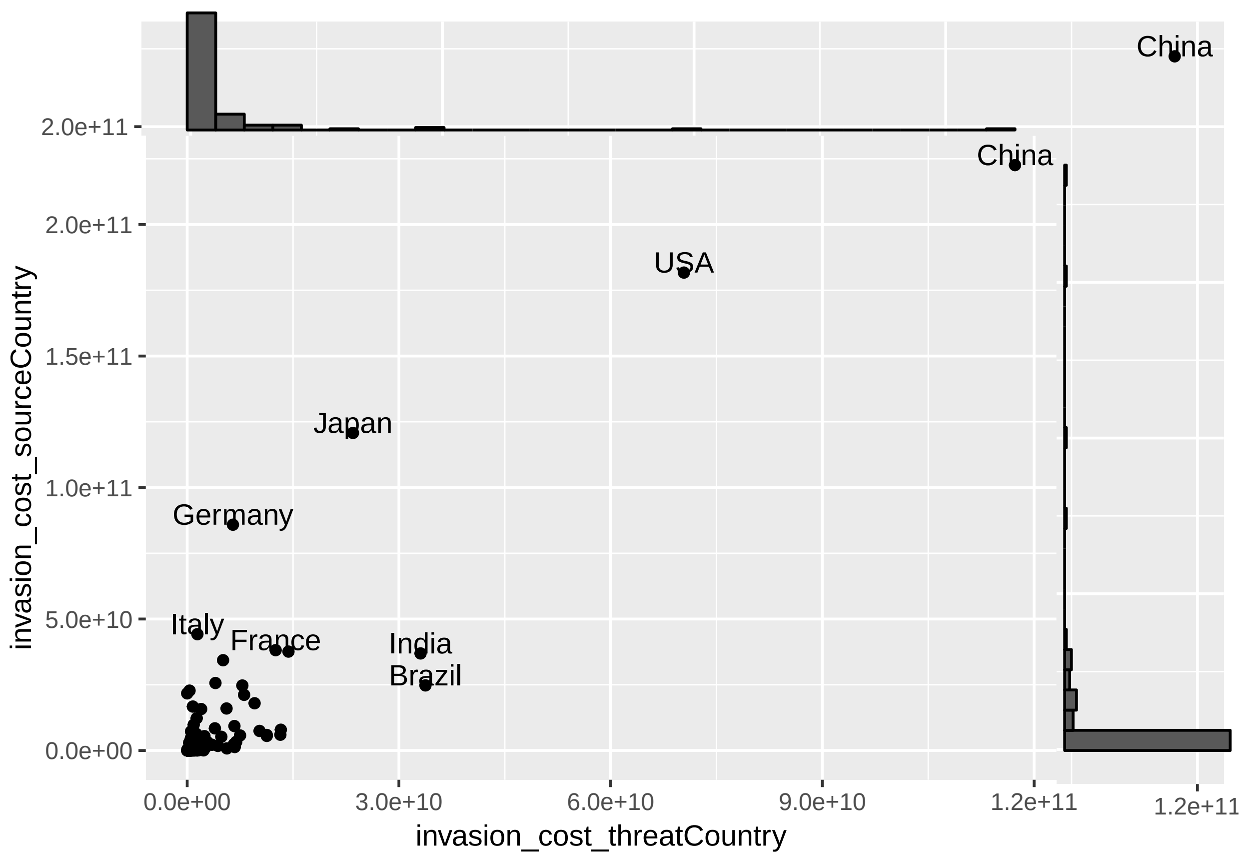
2.3 Merge the four tables together based on the same country IDs:

```r
# join 4 tables together (full join) --------------------------------------
table00 <- table1 %>%
    full_join(table2_nodup, by="country") %>%
    full_join(table3_nodup, by="country") %>%
    full_join(table4_nodup, by="country")
table00_count <- table00 %>%
    count(country) %>%
    filter(n>1)
# there is no dup country
rm(table00_count)


# clean up unwanted ranks
#   and re-define the units of invasion costs (threatened and source)
table01 <- select(table00, -starts_with("rank")) %>%
    mutate(ICt_million = invasion_cost_threatCountry/(10^6),
           ICs_million = invasion_cost_sourceCountry/(10^6))
# rm(table1, table2_dup, table3_nodup, table4_nodup)
```

2.4 Preliminary plots

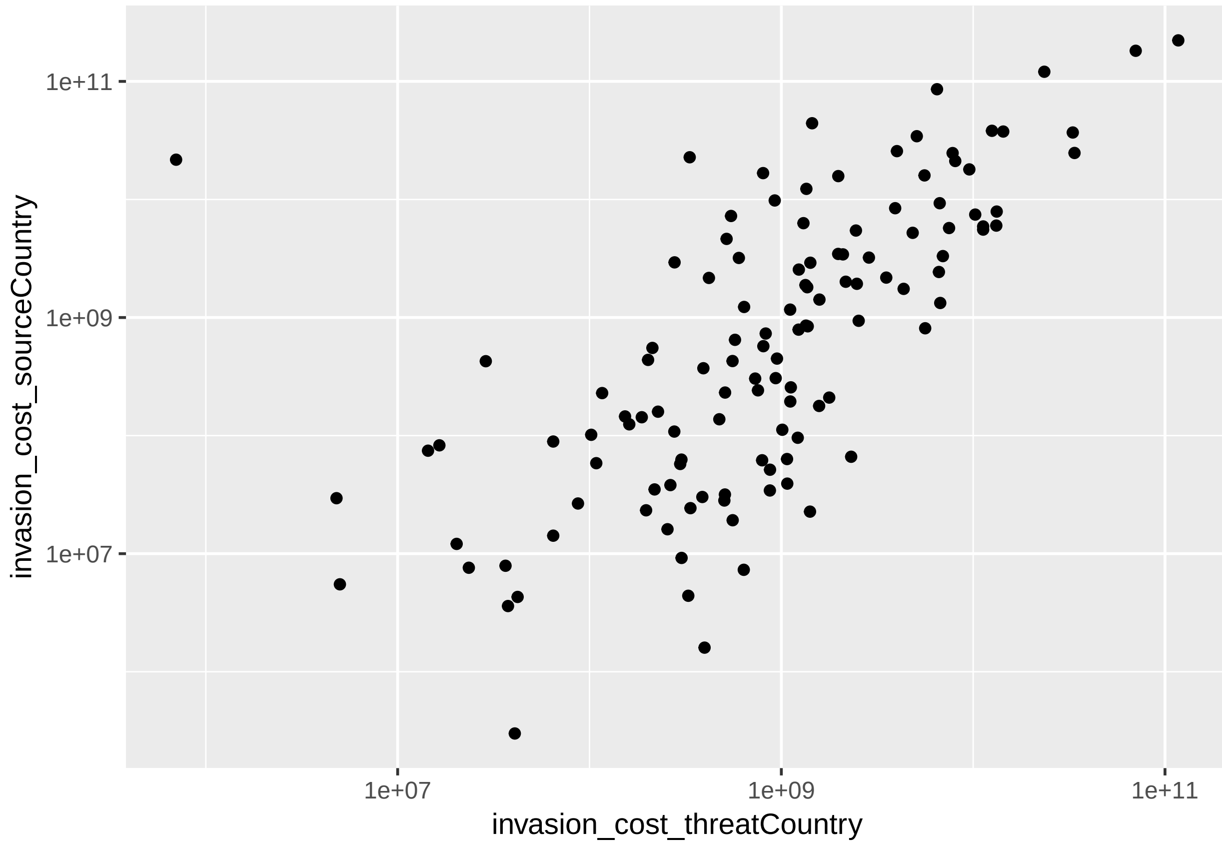
p1: Base plot

```
# base plot
(p1 <- table01
    %>% ggplot(aes(x=invasion_cost_threatCountry, y=invasion_cost_sourceCountry))
    +   geom_point()
    +   geom_text(data = subset(table01, ICs_million>34000 | ICt_million>14000 ), aes(label = country),

)
ggMarginal(p1, type="histogram")
```



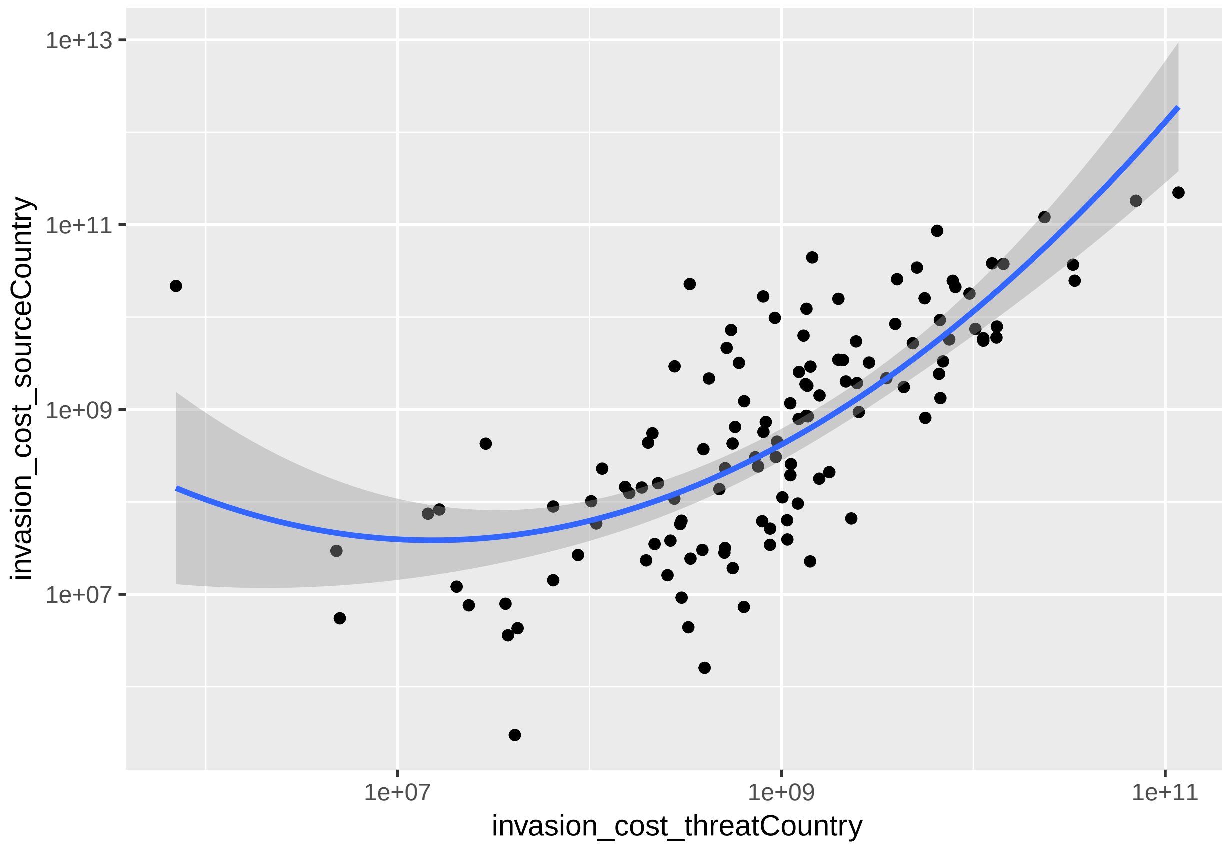
p2: Scale x and y axes for better visualization of the data

```
# scale x and y axes and make the graph more presentable, but note that the two outliers (China, USA) i
(p2 <- table01
    %>% ggplot(aes(x=invasion_cost_threatCountry, y=invasion_cost_sourceCountry))
    +   geom_point()
    +   scale_x_log10()+scale_y_log10()
)
```



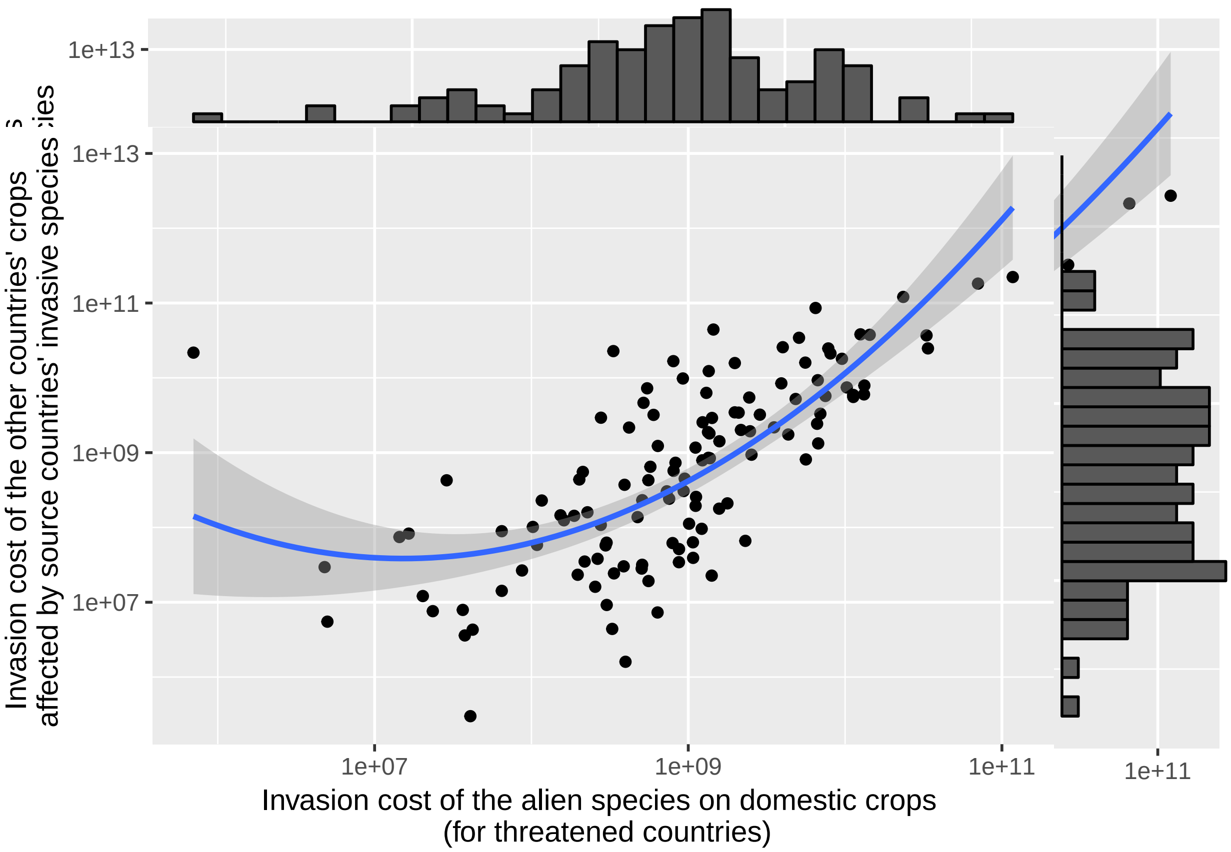
p3: Add a smooth curve to identify the association

```
# add a smooth curve to indicate the positive association
(p3 <- table01
    %>% ggplot(aes(x=invasion_cost_threatCountry, y=invasion_cost_sourceCountry))
    +   geom_point()
    +   scale_x_log10()+scale_y_log10()
    +   geom_smooth(span=10)
    # +   geom_label()
)
```



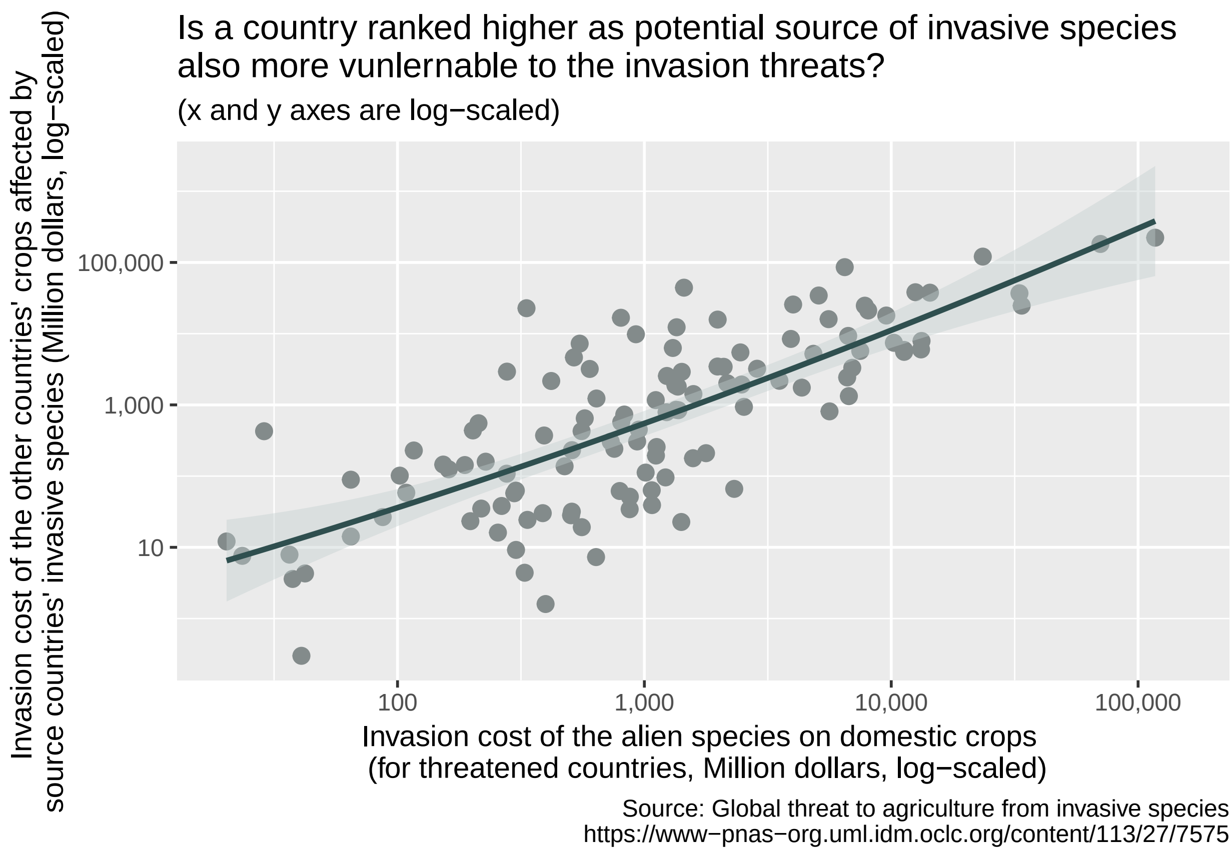
p4: Add marginal histograms

```
# modify the plot for better presentation;
#   add marginal histograms (https://www.r-graph-gallery.com/277-marginal-histogram-for-ggplot2/)
(p4 <- table01
    %>% ggplot(aes(x=invasion_cost_threatCountry, y=invasion_cost_sourceCountry))
    +   geom_point()
    +   scale_x_log10()+scale_y_log10()
    +   geom_smooth(span=10)
    +   xlab("Invasion cost of the alien species on domestic crops \n (for threatened countries)")
    +   ylab("Invasion cost of the other countries' crops \n affected by source countries' invasive spec
    # +   geom_label()
)
(p4 <- ggMarginal(p4, type="histogram"))
```

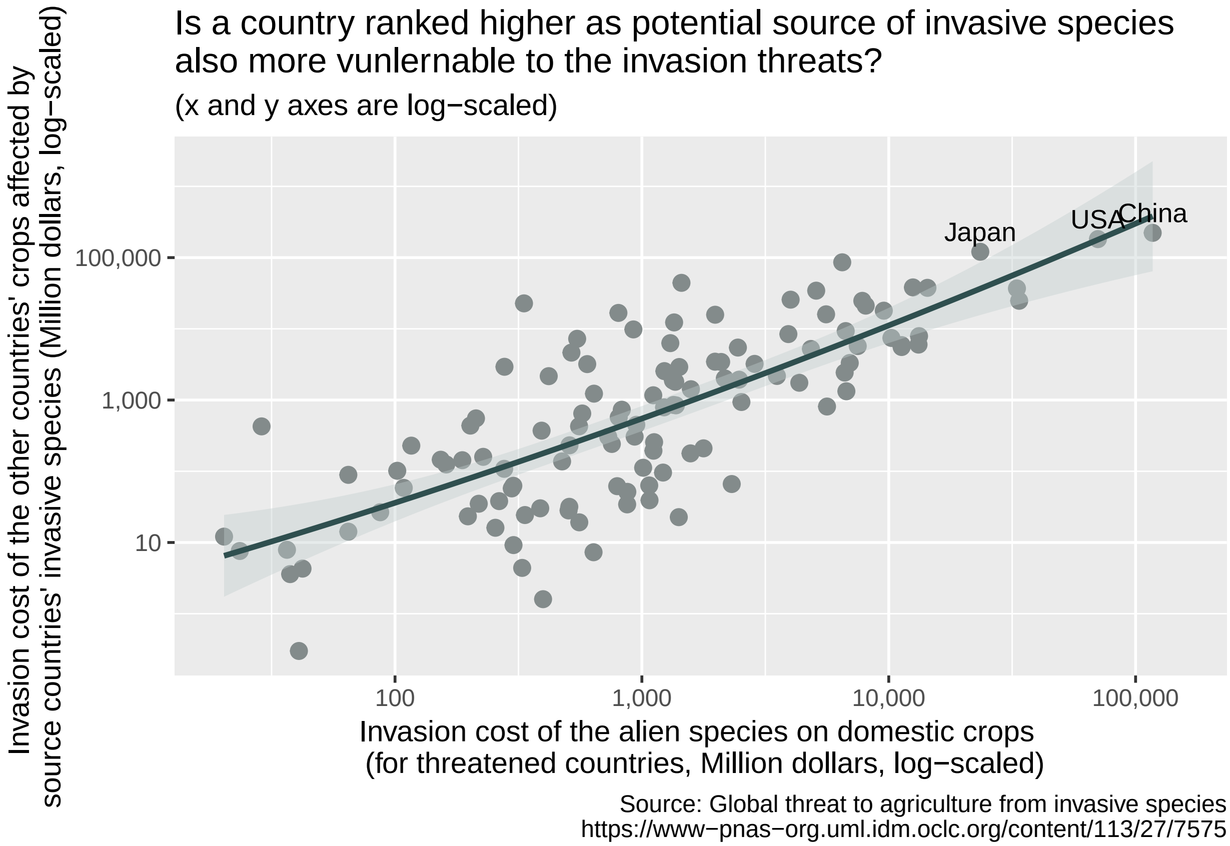


p5: Further modification on the plot

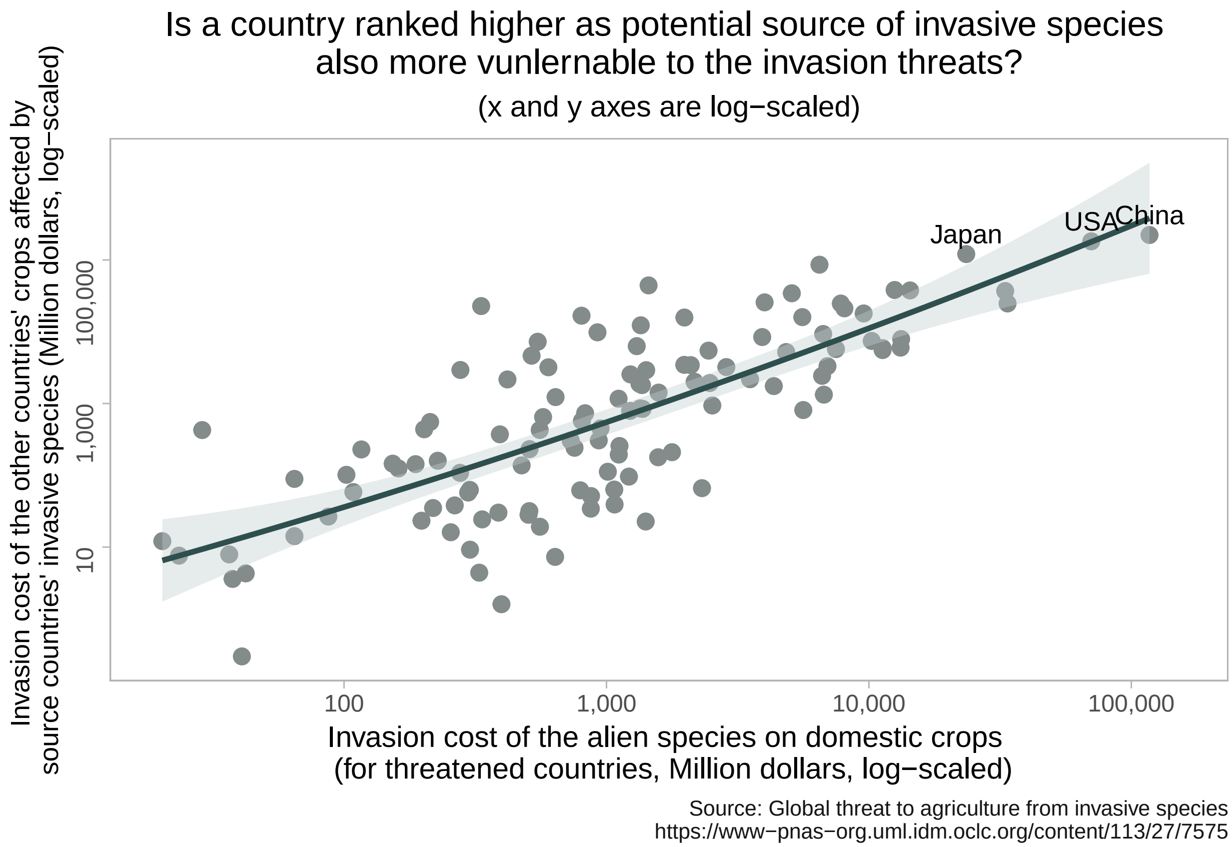
```
# p6 - log-scaled plot (final plot)
(p5 <- table01
    # change units into million dollars:
    %>% ggplot(aes(x=ICt_million, y=ICs_million))
    +   geom_point(colour="azure4", size=2.5)
    +   scale_x_log10(limits=c(20 , 1.5*10^5), labels=comma)+scale_y_log10(labels=comma) # labels=comma
    +   xlab("Invasion cost of the alien species on domestic crops \n (for threatened countries, Millio
    +   ylab("Invasion cost of the other countries' crops affected by \nsource countries' invasive spec
    +   labs(title="Is a country ranked higher as potential source of invasive species \nalso more vunle
             subtitle = "(x and y axes are log-scaled)",
             caption = "Source: Global threat to agriculture from invasive species
                       https://www-pnas-org.uml.idm.oclc.org/content/113/27/7575")
    +   geom_smooth(span=10, fill="azure3", colour="darkslategray")
)
```



```r
# label important points:
(p5 <- p5
    # +   geom_text(data = subset(table01, ICt_million>70000), aes(label = country), vjust = 0, nudge_y
    +   geom_text(data = subset(table01, ICs_million>100000), aes(label = country), vjust = 0, nudge_x=(
)
```



```r
# add theme
(p5 <- p5
    + theme_light()
    + theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(), #no gridline
            axis.title.y = element_text(size=10), # change the size of y axis label
            axis.text.y = element_text(angle = 90),
            plot.title = element_text(hjust =.5), # center plot title,
            plot.subtitle = element_text(hjust =.5),
            plot.caption = element_text(size=8, color = "gray8")) # change caption style
)
```



p6: Add the marginal histogram to finalize the plot

```
# add marginal histograms
(p6 <- ggMarginal(p5, type="histogram", fill="darkgray", colour="darkslategray", size=7))
```

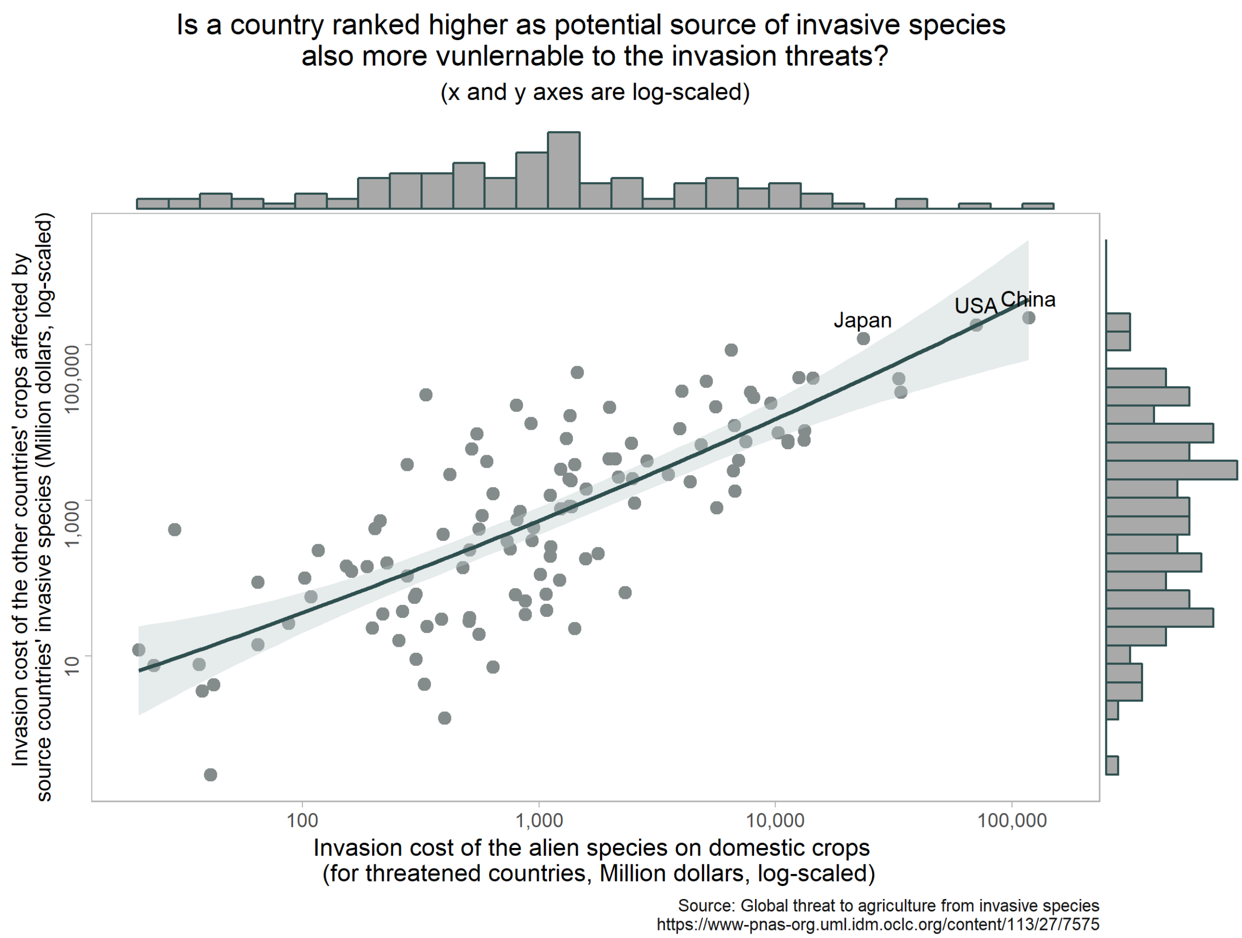


Figure 1:

Note that, this plot is saved on my computer by the ggsave function and then inserted in this document since the ggMarginal function does not plot nicely in R Markdown. As you may see, in p1 and p4, there are some points are plotted outside the axes due to the incompatibility of the ggMarginal() function.

3. Final plot

3.1 The finalized plot (p6) to address my question

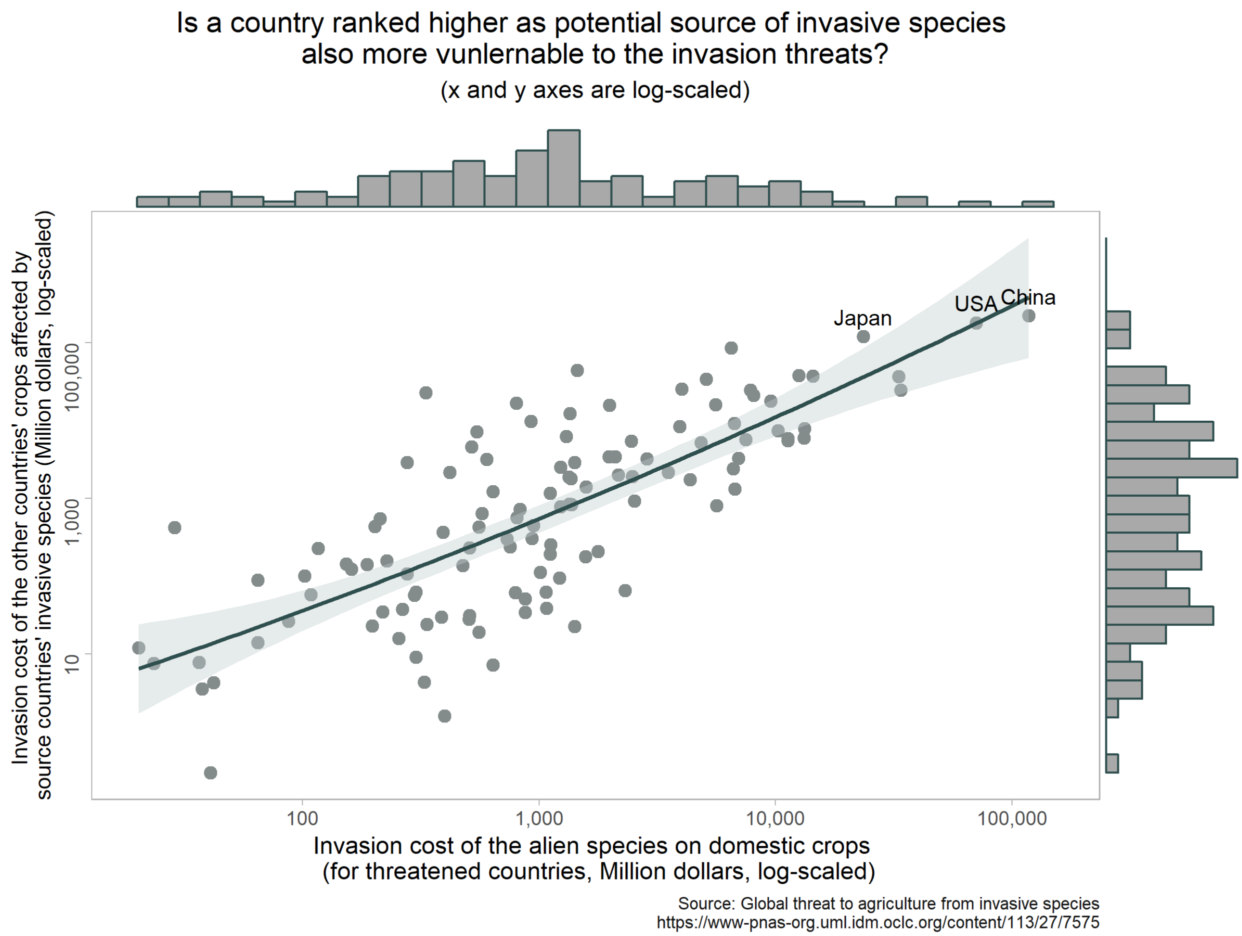


Figure 2:

3.2 The modified preliminary plot for comparison

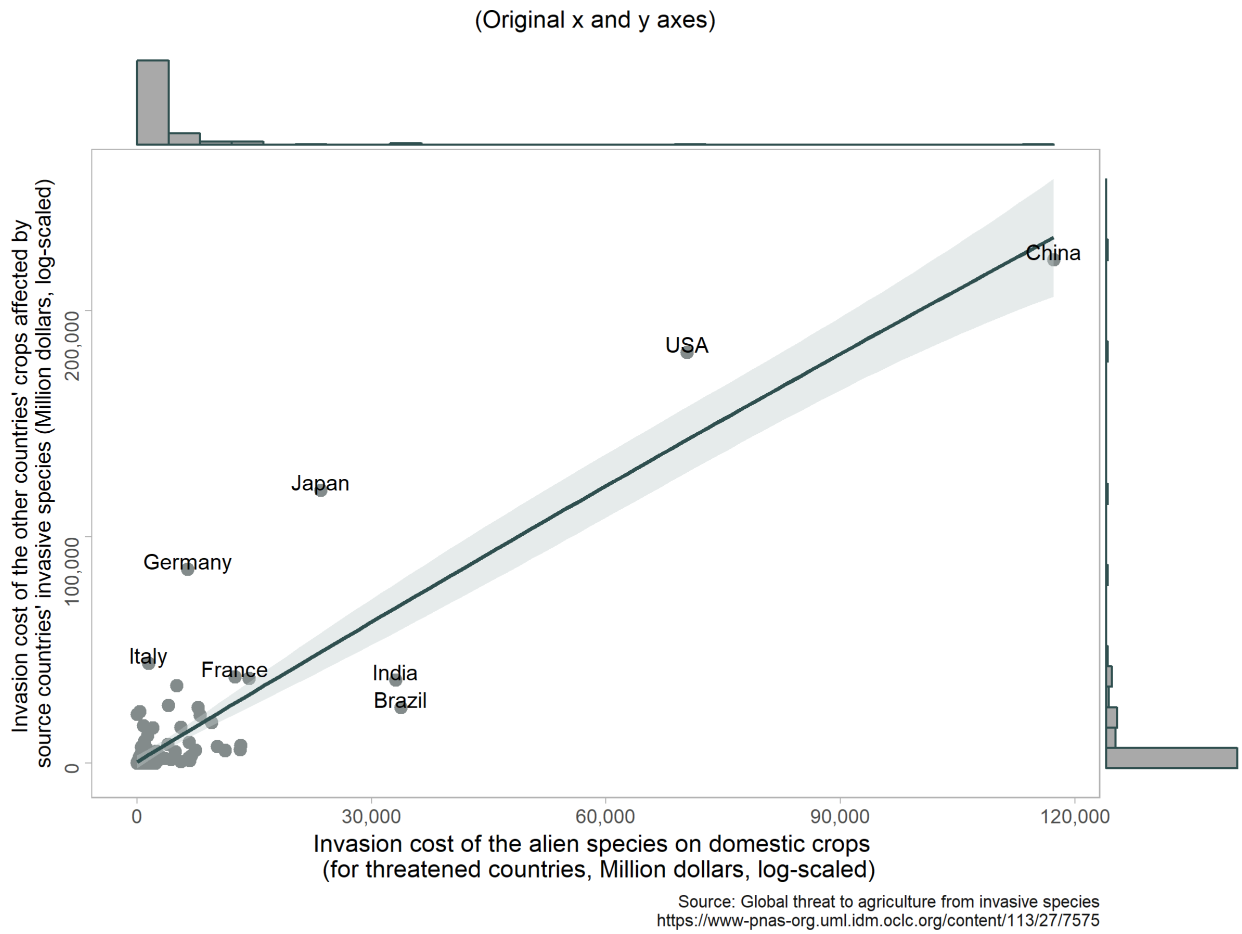


Figure 3:

3.3 Conclusion

From the finalized plot (p6), I can see that there is a relatively strong and positive association between the threat that source countries impose on the other countries and the invasion threat of this country received from foreign species invasion. Also, after scaling both axes logarithmly, the distributions of invasion cost from the source country and the invasion cost on the threatened country appear to be normal or at least somewhat sysmetric. However, by no means, the positive association implies a causal relationship, as the effect are most likely to be moderated by the trading amount of a country(Paini, Sheppard, Cook and all, 2016).

Though the final plot present the clear association, the preliminary and non-log-scaled plot give us a clear look of outlining points. In the modified preliminary plot, it is shown that China and USA are the two top threatening source countries, whose invastion cost being either a threatened or source country are way higher than the rest of the countries.

Codes to generate the modified preliminary plot for comparison

```
# add the un-logged plot for comparison

# modify the units
(p1 <- table01
    # change units into million dollars:
    %>% ggplot(aes(x=ICt_million, y=ICs_million))
    +   geom_point(colour="azure4", size=2.5)
    +   scale_x_continuous(labels=comma)+scale_y_continuous(labels=comma) # labels=comma: no to show th
    +   xlab("Invasion cost of the alien species on domestic crops \n (for threatened countries, Millio
    +   ylab("Invasion cost of the other countries' crops affected by \nsource countries' invasive spec
    +   labs(
             # title="Is a country ranked higher as potential source of invasive species \nalso more vu
             subtitle = "(Original x and y axes)",
             caption = "Source: Global threat to agriculture from invasive species
             https://www-pnas-org.uml.idm.oclc.org/content/113/27/7575")
    +   geom_smooth(span=10, fill="azure3", colour="darkslategray")
    )



# label important points:
(p1 <- p1
    # +   geom_text(data = subset(table01, ICt_million>70000), aes(label = country), vjust = 0, nudge_y
    +   geom_text(data = subset(table01, ICs_million>34000 | ICt_million>14000 ), aes(label = country),
)


# add theme
(p1 <- p1
    + theme_light()
    + theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(), #no gridline
            axis.title.y = element_text(size=10), # change the size of y axis label
            axis.text.y = element_text(angle = 90),
            plot.title = element_text(hjust =.5), # center plot title
            plot.subtitle = element_text(hjust =.5), # center plot title
            plot.caption = element_text(size=8, color = "gray8")) # change caption style
)
```

```r
# add marginal histograms
(p1 <- ggMarginal(p1, type="histogram", fill="darkgray", colour="darkslategray", size=7))
```

4. Save the final work

```r
ggsave("fig_output/A1_p6.png", p6, width = 8, height = 6)
```

To access the figure on the side, please go to the fig_output folder.

5. Potential implication

Interesting questions:

Q1. Which countries are the hosts of the most invasive species?

In order to answer to this question, extra country information of species needs to be provided in table 6.

Q2. Is the invasion threat of a country affected/drived by a country's trading amount?

To answer that, the trading amount of each country needed to be provided.

Q3. Can a country be better protected from invasion species (i.e. reduce the invasion cost) by having better agriculture inspection and imported food and animal product security check?

Extra information and measures on a country's agriculture inspection results and security level measures need to be provided for this purpose.

6. Disgarded work

Interests in finding out the spread of origin countries of the most invasive species

In table_6/invasive_species dataset, list of species and their maximum impact percentages are presented, and there are lists of countries and species in africa_species dataset. I was interested in finding out which countries are the hosts of the most invasive species (i.e. the species having the highest-ranked impact percentage). However, after joining the two tables together, little common species are present in both tables. In other words, I cannot identify which countries host the species in table 6. So, I cannot get a conclusion of which countries carries the most influentially invasive species.

```
# join species ------------------------------------------------------------
species <- invasive_species[invasive_species$species %in% africa_species$species, ]
species
```

```
## # A tibble: 1 x 3
##   species         max_impact_percent  rank
##   <chr>                        <dbl> <dbl>
## 1 Cinara cupressi                 12    17
```

```
species <- invasive_species %>%
    inner_join(africa_species, by="species")
#only 7 invasive species can be found in african speices dataset - this join is not usable
species
```

```
## # A tibble: 7 x 8
##   species max_impact_perc~  rank authority country kingdom environment_sys~
##   <chr>              <dbl> <dbl> <chr>     <chr>   <chr>   <chr>
## 1 Cinara~               12    17 (Buckton~ Libya   Animal~ host
## 2 Cinara~               12    17 (Buckton~ Morocco Animal~ host
## 3 Cinara~               12    17 (Buckton~ Rwanda  Animal~ host
## 4 Cinara~               12    17 (Buckton~ Ethiop~ Animal~ host
## 5 Cinara~               12    17 (Buckton~ Kenya   Animal~ host
## 6 Cinara~               12    17 (Buckton~ Uganda  Animal~ host
## 7 Cinara~               12    17 (Buckton~ Malawi  Animal~ host
## # ... with 1 more variable: origin <chr>
```

```r
summarise(species)
```

```
## # A tibble: 1 x 0
```