STAT 7350 Assignment 1

Jingyu Wang 7701969

March 21, 2019

1. Identify a question

2. Interact with the data

Question 1: Draw histogram of data for overall invasion threat, total invasion cost ,GDP proportion for 124 countries and 140 species of invasion threat.

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0       v purrr   0.2.5
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.3.1
## v readr   1.3.1       v forcats 0.3.0

## -- Conflicts ------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(gridExtra)
```

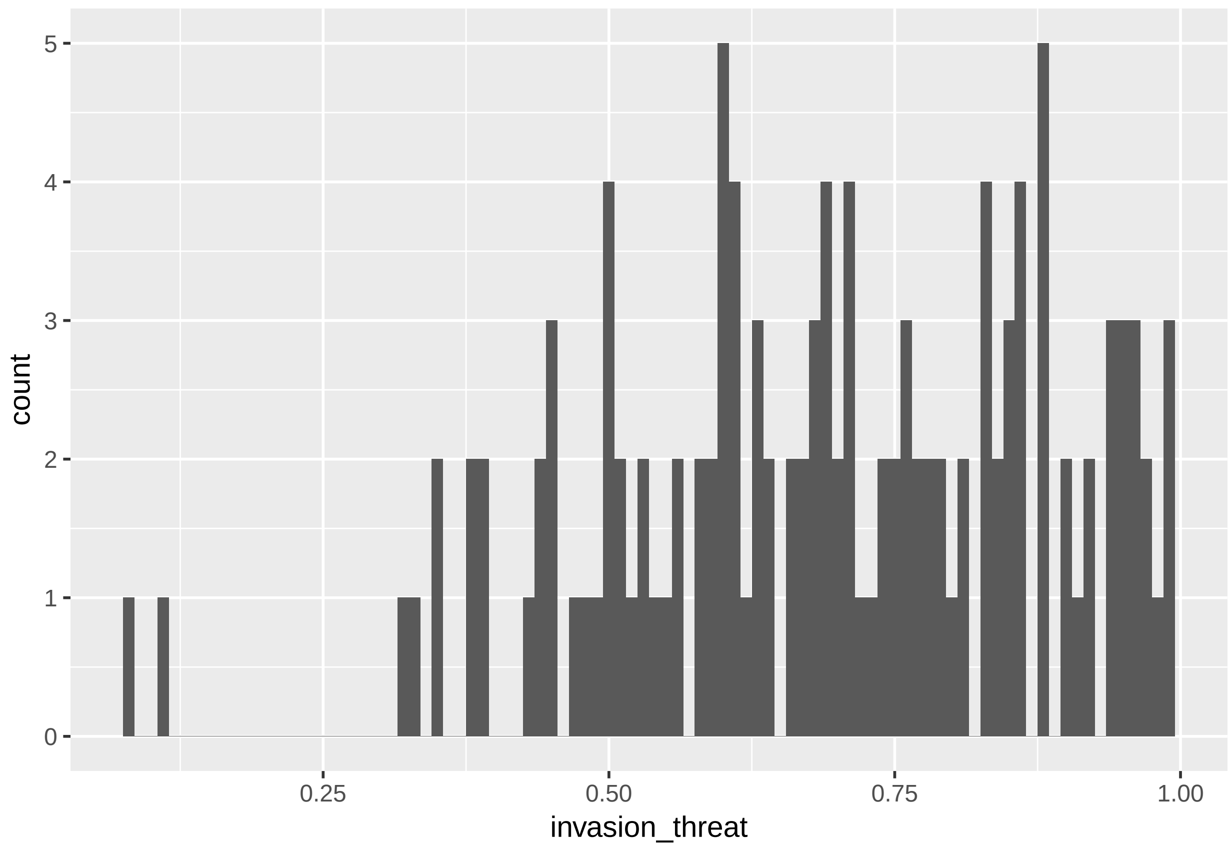
```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
# Invasion threat for 124 countries
inv.threat<-read_csv("/Users/summe/Downloads/Stat-7350-class/Assignment 1/Data/table_1.csv")
```

```
## Parsed with column specification:
## cols(
##   rank = col_double(),
##   country = col_character(),
##   invasion_threat = col_double()
## )
```

```r
ggplot(inv.threat,aes(invasion_threat)) +
  geom_histogram(binwidth = 0.01)
```



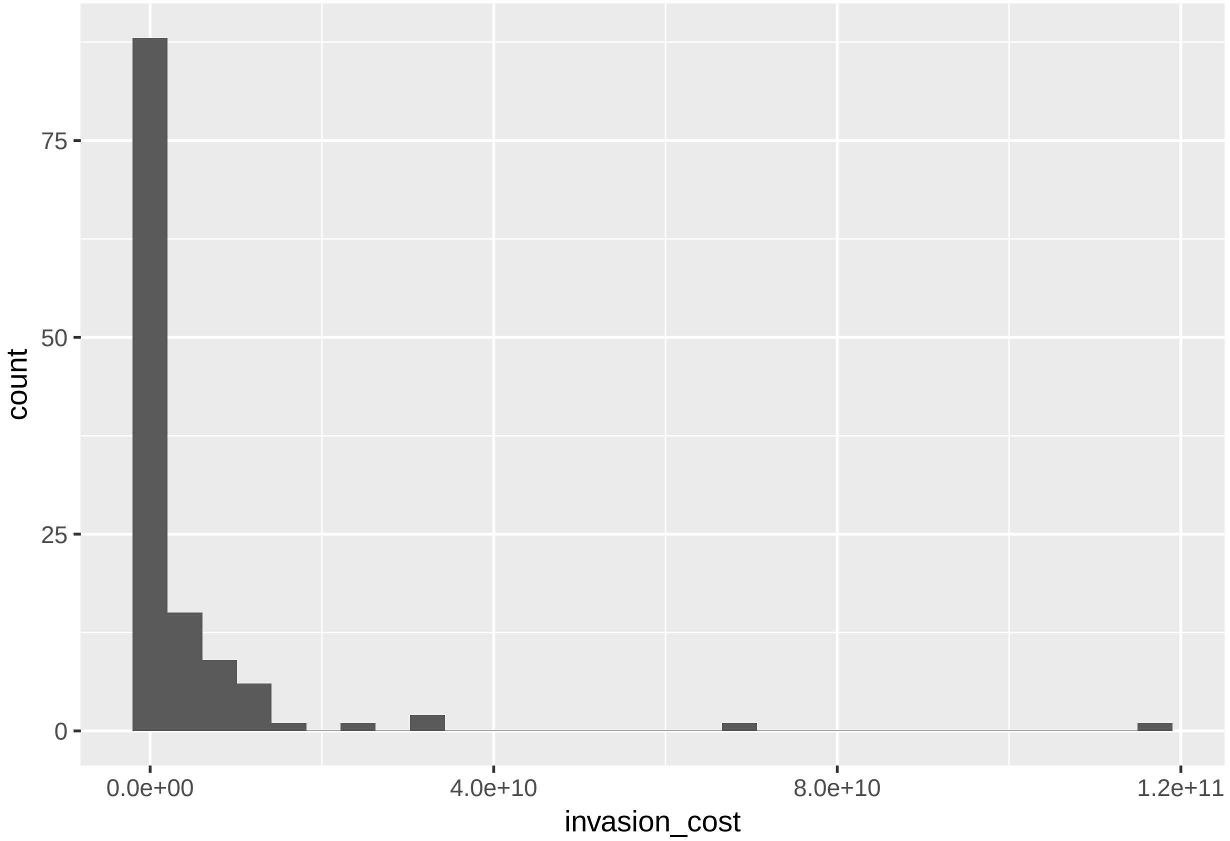
To see the histgram of invasion threat for 124 countries, each bar displays the range of probability and counts the number of countries in that range as well.

```
# Invasion cost for 124 countries
tot.cost<-read_csv("/Users/summe/Downloads/Stat-7350-class/Assignment 1/Data/table_2.csv")
```

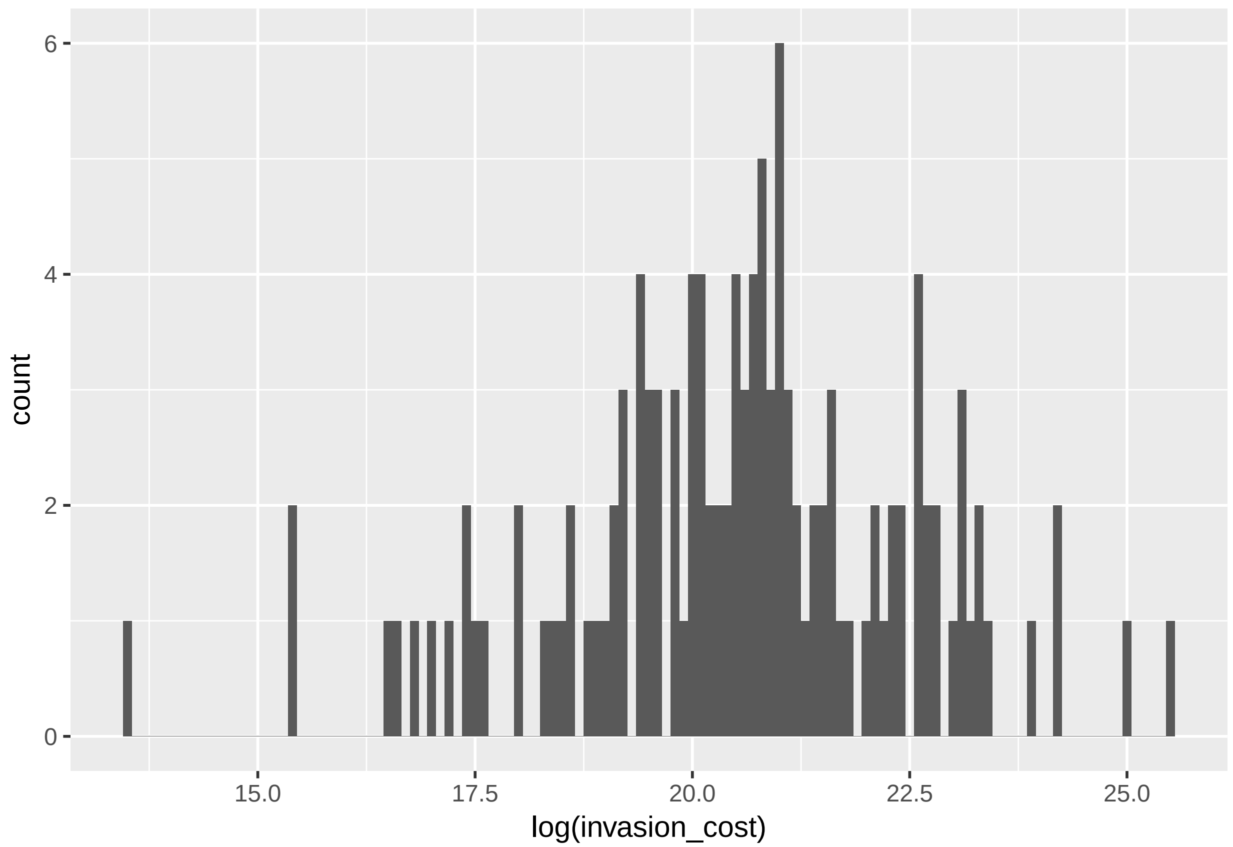
```
## Parsed with column specification:
## cols(
##   country = col_character(),
##   invasion_cost = col_double(),
##   rank = col_double(),
##   inv_cost_millions = col_double()
## )
```

```
ggplot(tot.cost,aes(invasion_cost)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(tot.cost,aes(log(invasion_cost))) + geom_histogram(binwidth = 0.1)
```



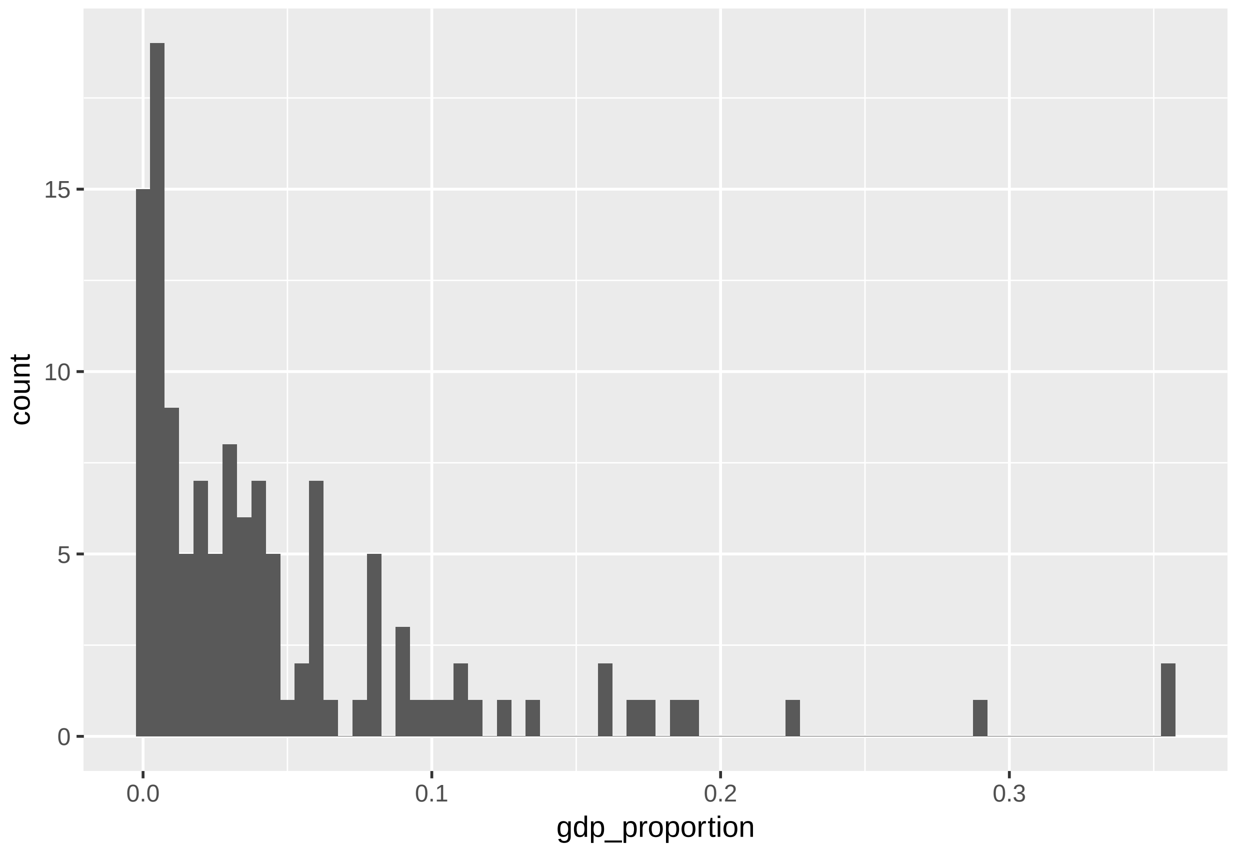
```
# The following plot is hard to show in the PDF file
#ggplot(tot.cost,aes(log(invasion_cost),fill = country)) + geom_histogram()
```

First histgram did not show data very well, the reasoon is that the range of invasion cost is too wide to perform. Thus, log transformation will be used for cost data and draw the secong plot. In the second plot, it counts the number of countries for the corresponding range of invasion cost.

```
# GDP proportion for 124 countries
GDP<-read_csv("/Users/summe/Downloads/Stat-7350-class/Assignment 1/Data/table_3.csv")
```

```
## Parsed with column specification:
## cols(
##   country = col_character(),
##   invasion_cost = col_double(),
##   gdp_mean = col_double(),
##   gdp_proportion = col_double(),
##   rank = col_double()
## )
```

```
ggplot(GDP,aes(gdp_proportion)) + geom_histogram(binwidth = 0.005)
```



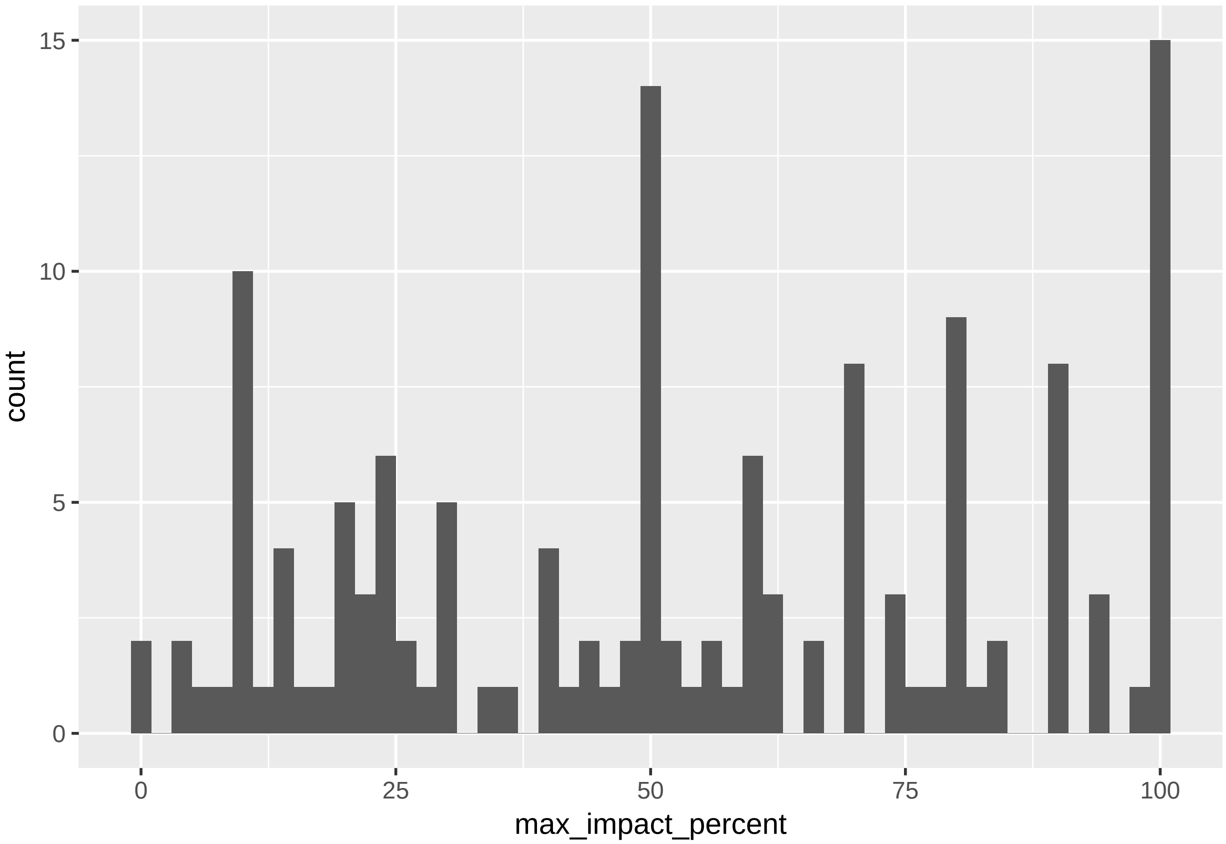
```
#The following plot is also gard to show in PDF file
#ggplot(GDP,aes(gdp_proportion, fill = country)) + geom_histogram(binwidth = 0.005)
```

Same kinds of histgram is to display the GDP proportion for 124 countries.

```
# 140 species of invasion threat
spec<-read_csv("/Users/summe/Downloads/Stat-7350-class/Assignment 1/Data/table_6.csv")
```

```
## Parsed with column specification:
## cols(
##   species = col_character(),
##   max_impact_percent = col_double(),
##   rank = col_double()
## )
```

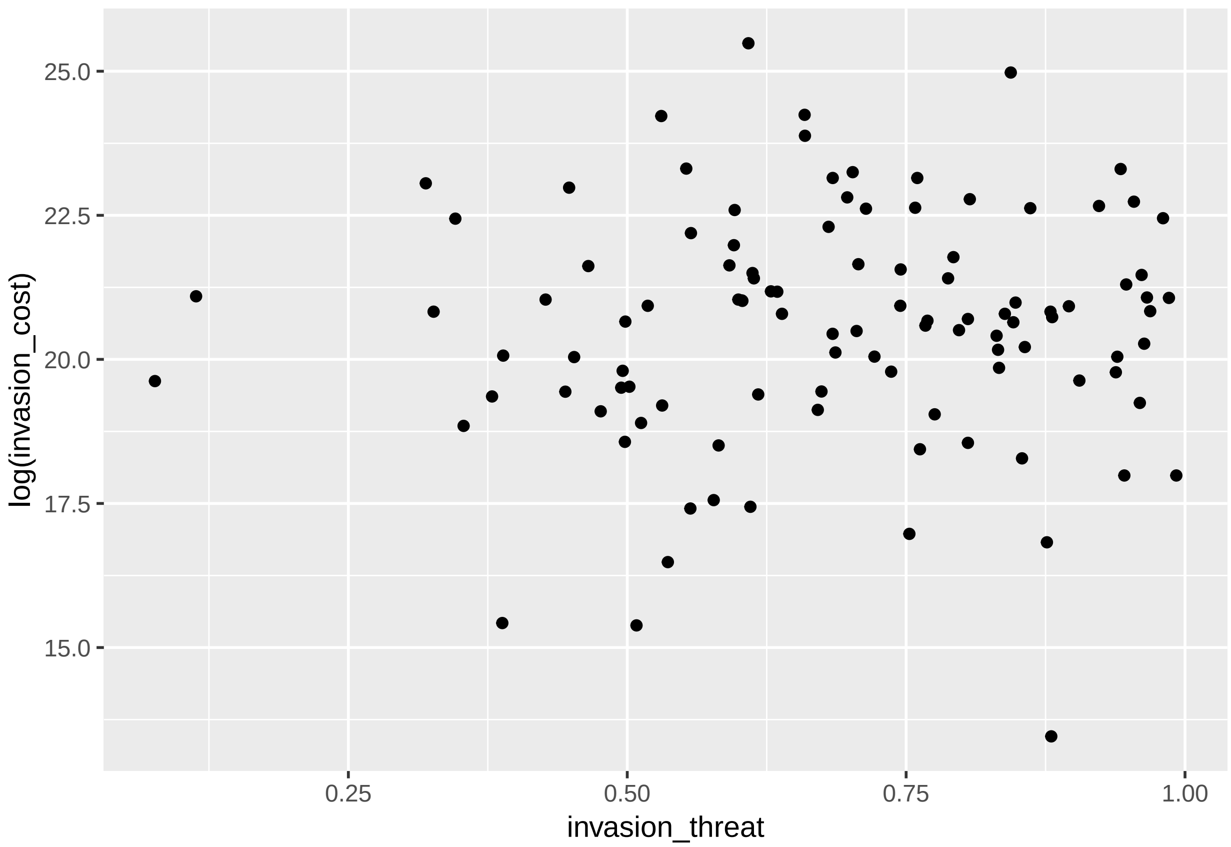
```
ggplot(spec,aes(max_impact_percent)) + geom_histogram(binwidth = 2)
```



This histgram shows 140 species of invasion threat and their maximum percentage which affects one of crops. It did not express what kinds of invasion in each country, so I do not know how to link this data with others.

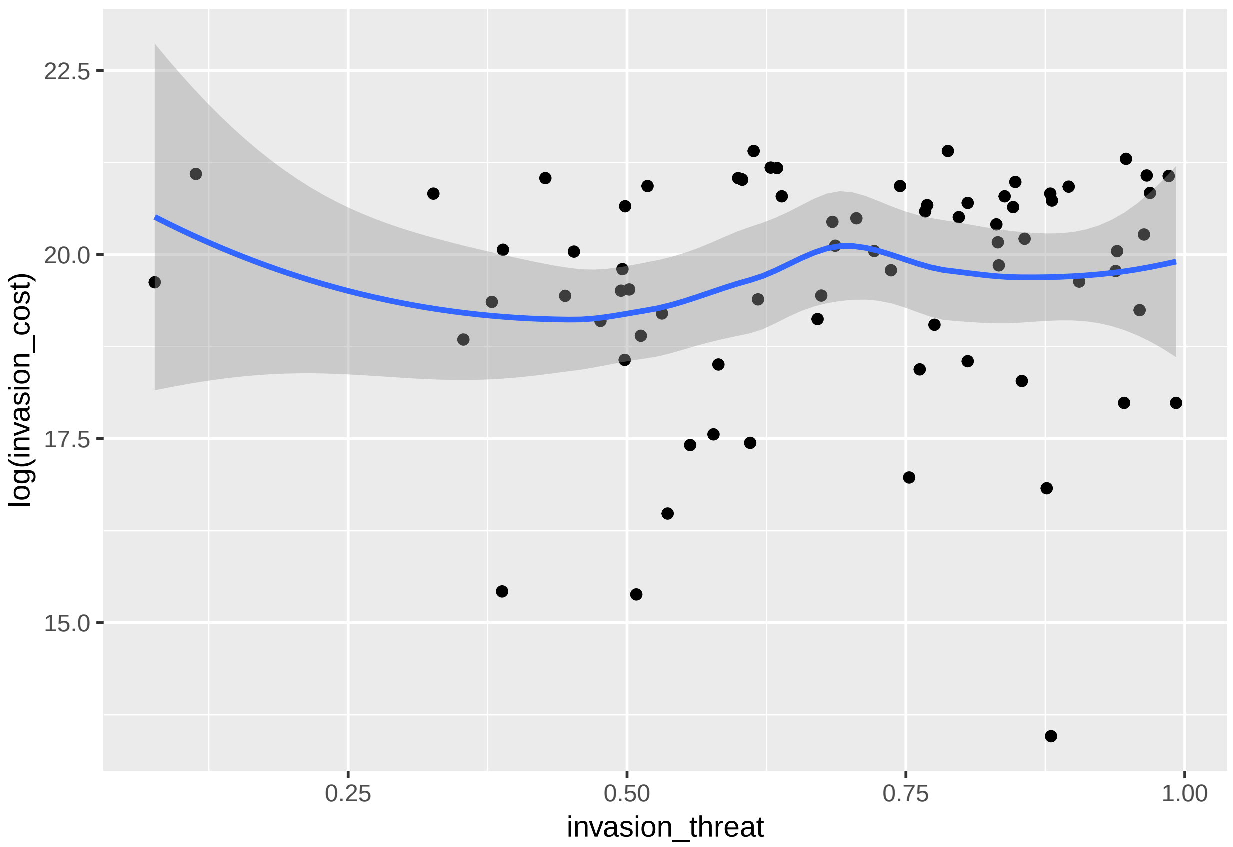
Question 2: Try to find the relationship between invasion threat and its cost.

```
inv.cost<-merge(inv.threat[,2:3],tot.cost[,1:2])
ggplot(data = inv.cost,
       mapping = aes(x = invasion_threat,
                     y = log(invasion_cost)))+
  geom_point()
```

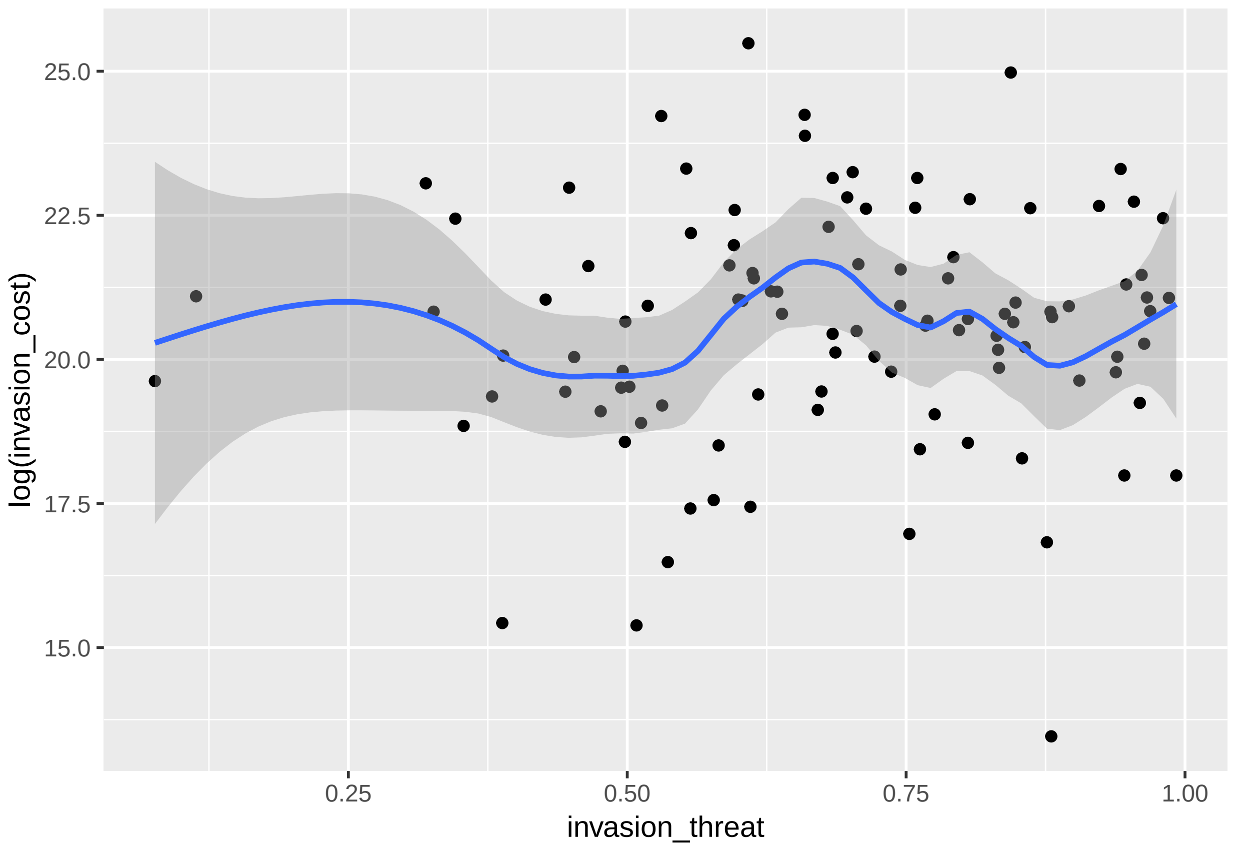


```
inv.cost %>%
  filter(invasion_cost<2e+09)%>%
  ggplot(aes(x = invasion_threat,y = log(invasion_cost)))+
  geom_point()+geom_smooth()
```

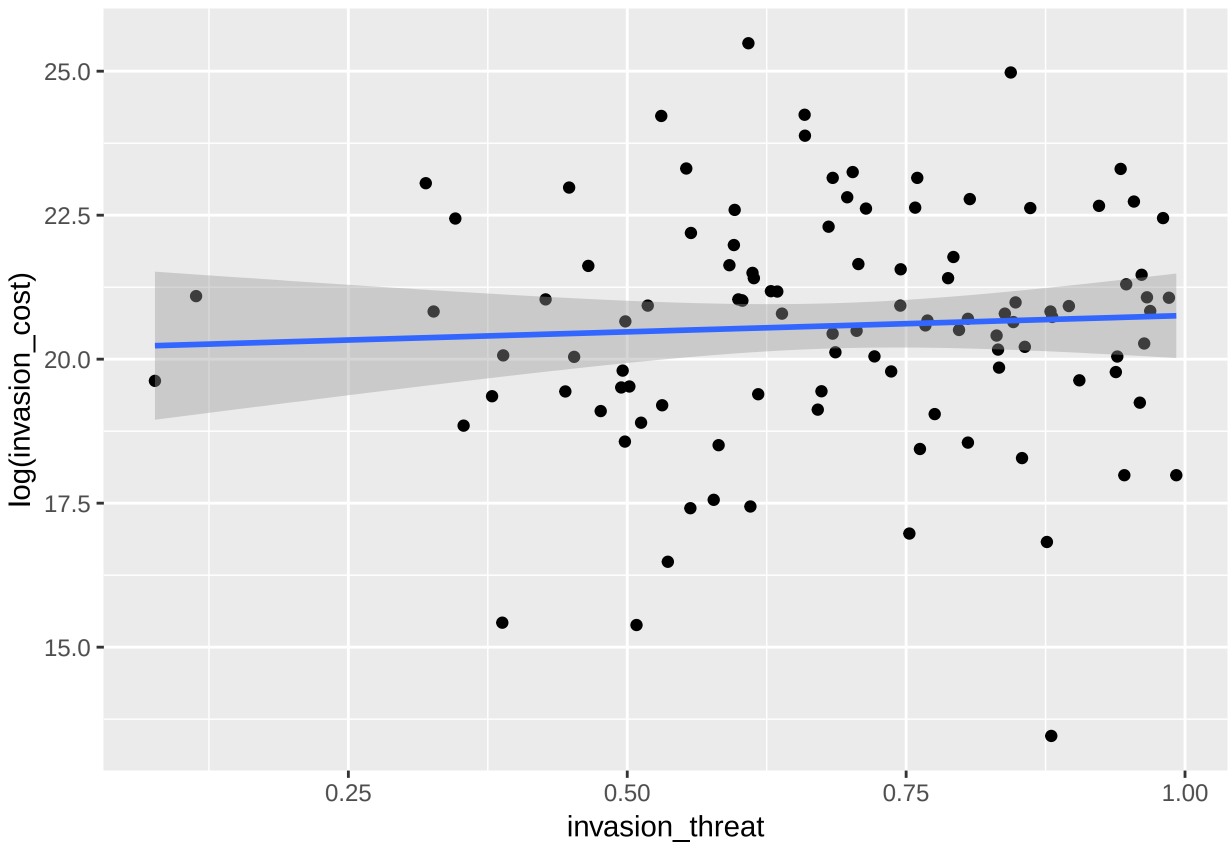
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
ggplot(inv.cost ,aes(x = invasion_threat,y = log(invasion_cost)))+
  geom_point()+geom_smooth(method = "loess", span = 0.4)
```



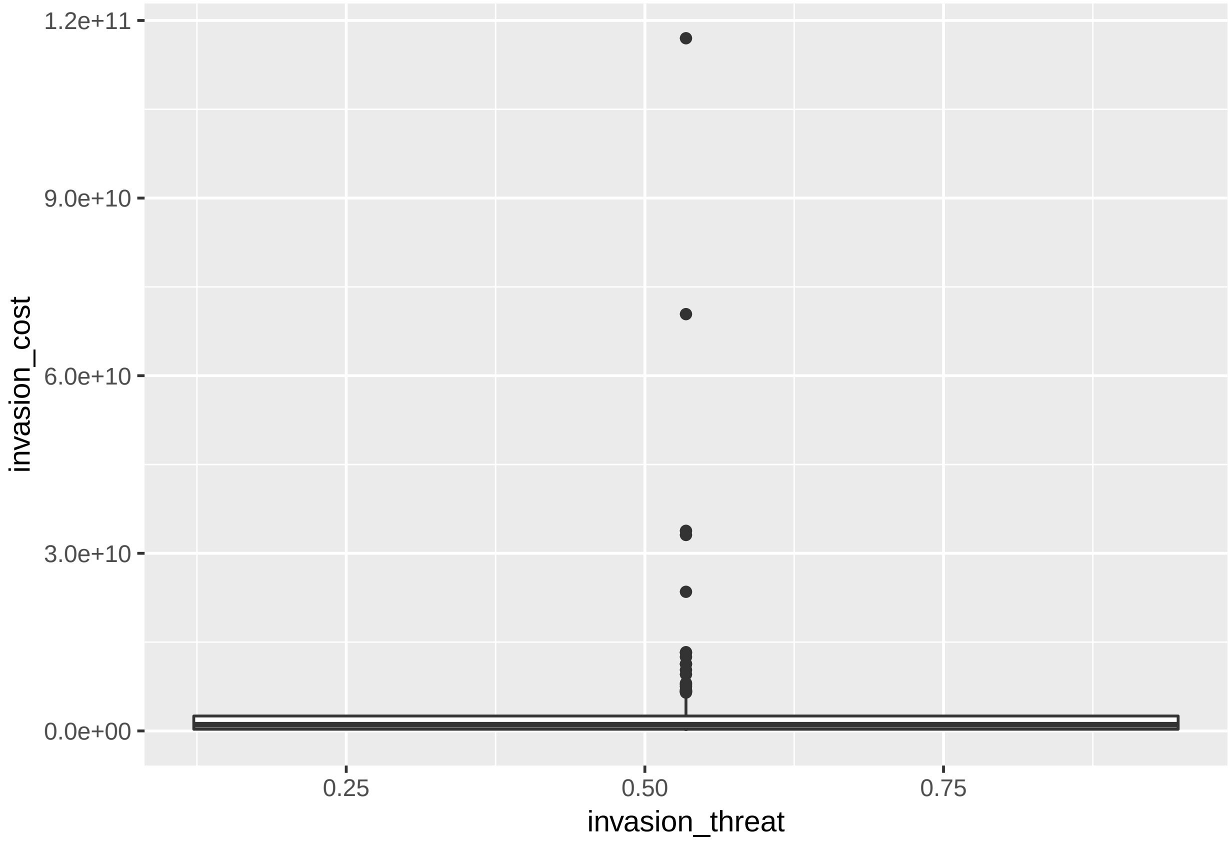
```
ggplot(inv.cost ,aes(x = invasion_threat,y = log(invasion_cost)))+
  geom_point()+geom_smooth(method = "lm")
```



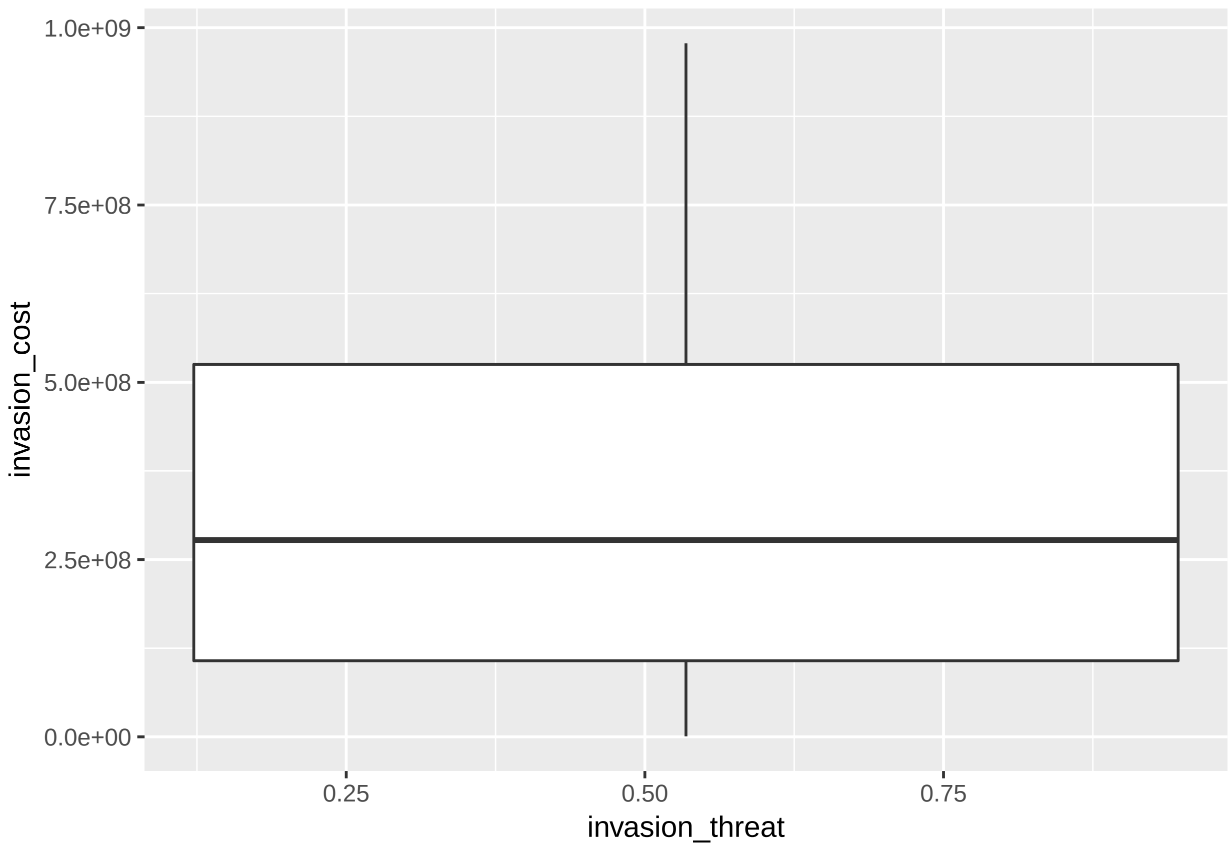
The invasion threat is the product of arrive likelihood and establishment likelihood. We guess there is relationship between invasion threat and its cost. Larger invasion threat may have larger probability of invasion cost in different country. In the forth plot, there is no obvioury increasing of cost when the invasion threat is getting larger. We only can say the larger probability of invasion threat have larger probability of cost. It is not enough to say the invasion cost must get larger. Thus, this should be the reason why there is only a little rising with increasing of our x axis.

Question 3: Boxplot of invasion cost and its threat.

```
ggplot(data = inv.cost,
       mapping = aes(x =invasion_threat, y =
                         invasion_cost)) +geom_boxplot()
```



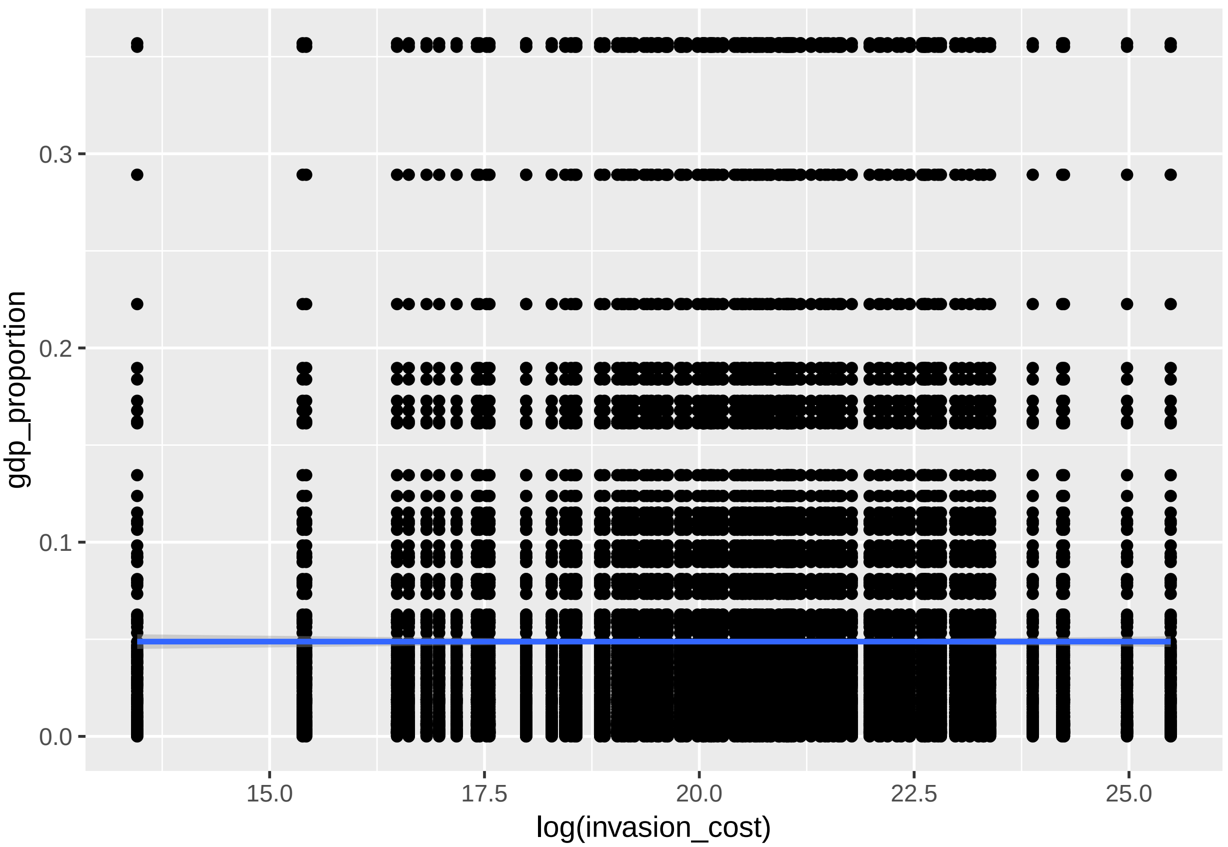
```r
inv.cost %>%
  filter(invasion_cost<1e+09)%>%
  ggplot(aes(x = invasion_threat,
             y = invasion_cost))+
  geom_boxplot()
```



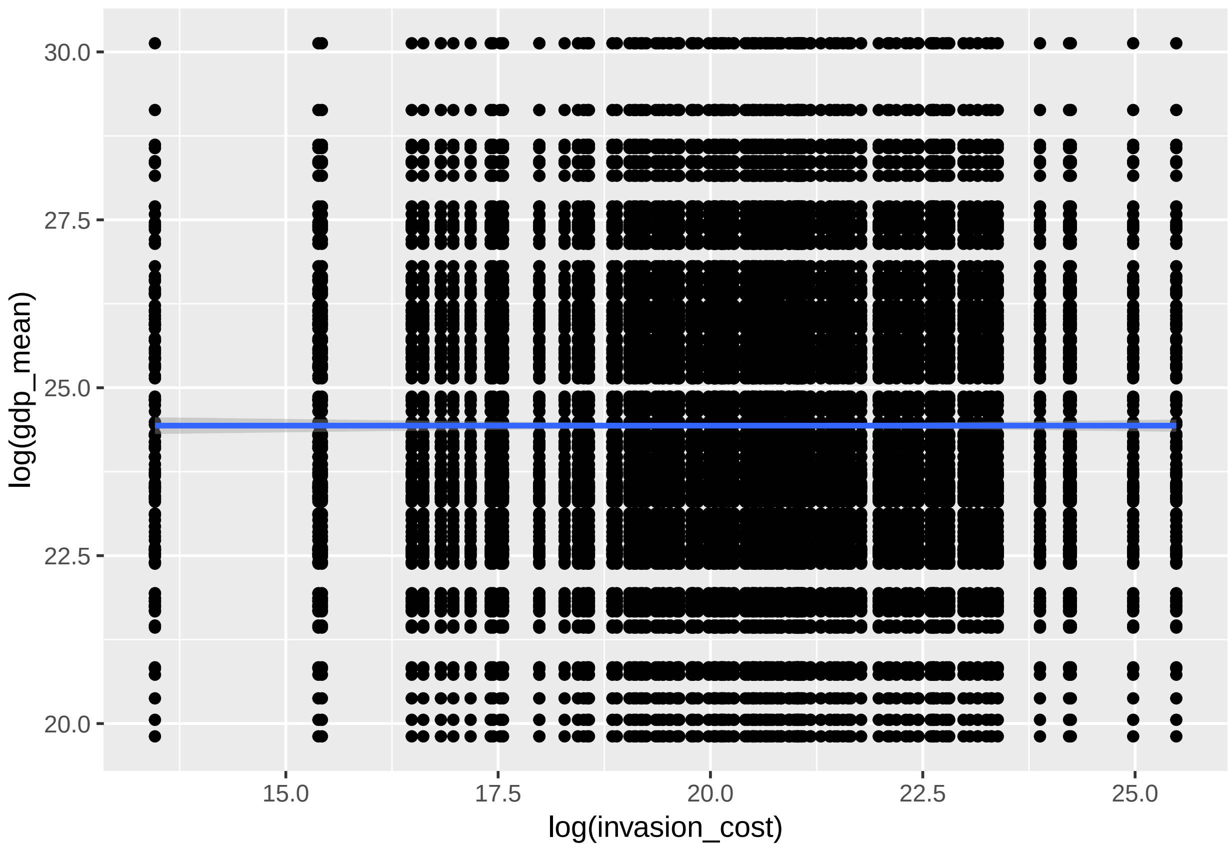
The boxplot did not work well in this case. It did not give much information except the median.

Question 4: What happens between invasion cost and gdp proportion in each countries.

```
cost.gdp<-merge(tot.cost[,1:2],GDP[,3:4])
ggplot(data = cost.gdp,
       mapping = aes(x = log(invasion_cost),
                     y = gdp_proportion)) +
  geom_point()+geom_smooth(method = "lm")
```



```
ggplot(data = cost.gdp,
       mapping = aes(x = log(invasion_cost),
                     y = log(gdp_mean))) +
  geom_point()+geom_smooth(method = "lm")
```



Both plot did not show any information for the relationship between invasion cost and gdp proportion or gdp mean. There is no relationship between them. The high invasion cost may not affect the gdp proportion, because the agriculture may be small part in that country. Even there is larger invasion cost, the gdp proportion still could small.

3. Redraw the most useful graph using principles of effective display

```
# This plot is too large to show in this file, so I save it in the Figures file as final plot.

# ggplot(inv.cost ,aes(x = invasion_threat,y = log(invasion_cost))+
#   geom_point(size = 2,
#              aes(color = country))+
#   geom_smooth(method = "lm")
```

Firstly, the menthod of “lm” was used *geom_smooth* function to give one more straight line. Then, I also use the different color in different countries to see which country in which area.

In new this plot, the straight line is more easlier to see the relationship between invasion cost and its threat, and the method “loess” si not obviously in this case. Also, we can distinguish the countries with different color. The single point displays the values of invasion threat and its cost for corresponding country.

4. Save the final graph

The larger probability of invasion threat have larger probability of cost. However, higher invasion threat probability may have higher invasion cost.

5. Identify follow-up quesions

1. If there is data of probability of invasion cost for each countrt, it would be easier to find the relationship between the invasion threat and its cost.
2. For table 6, if data indicate the invasion fron which country, it may connect with other data to get more information.