Assignment 1

Scott White

March 14, 2019

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyve
## v ggplot2 3.1.0       v purrr   0.3.0
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
## Warning: package 'tibble' was built under R version 3.5.2
## Warning: package 'readr' was built under R version 3.5.2
## Warning: package 'purrr' was built under R version 3.5.2
## Warning: package 'dplyr' was built under R version 3.5.2
## Warning: package 'stringr' was built under R version 3.5.2
## Warning: package 'forcats' was built under R version 3.5.2
## -- Conflicts ------------------------------------------------------------------------- tidyverse_co
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.5.3
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
t1 = read_csv("../data/table_1.csv")
```

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

```
t2 = read_csv("../data/table_2.csv")
```

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

```
t3 = read_csv("../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()
## )
```

```
t4 = read_csv("../data/table_4.csv")
```

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

```
t6 = read_csv("../data/table_6.csv")
```

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

```
t1;t2;t3;t4;t6
```

```
## # A tibble: 124 x 3
##     rank country       invasion_threat
##    <dbl> <chr>                   <dbl>
##  1     1 Mongolia                0.992
##  2     2 Guinea-Bissau           0.990
##  3     3 Nepal                   0.986
##  4     4 Bangladesh              0.980
##  5     5 Cambodia                0.969
##  6     6 Denmark                 0.966
##  7     7 Albania                 0.963
##  8     8 Chile                   0.961
##  9     9 Mauritius               0.960
## 10    10 Vietnam                 0.954
## # ... with 114 more rows
```

```
## # A tibble: 124 x 3
##    country   invasion_cost  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 114 more rows
```

```
## # A tibble: 124 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 Guinea         978000000  3380000000          0.289     3
##  4 Guinea         114000000   513000000          0.223     4
##  5 Mozambique    1218000000  6423000000          0.190     5
##  6 Madagascar    1074000000  5842000000          0.184     6
##  7 Cambodia      1121000000  6487000000          0.173     7
##  8 Nepal         1411000000  8411000000          0.168     8
##  9 Laos           508000000  3134000000          0.162     9
## 10 Ethiopia      2312000000 14344000000          0.161    10
## # ... with 114 more rows
```

```
## # A tibble: 124 x 3
##    country invasion_cost  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 114 more rows
```

```
## # A tibble: 140 x 3
##    species                     max_impact_percent  rank
##    <chr>                                    <dbl> <dbl>
##  1 Apiognomonia veneta                        0       1
##  2 Atherigona miliaceae                       0.3     2
##  3 Cryptophlebia illepida                     4       3
##  4 Conogethes punctiferalis                   5       4
##  5 Dysaphis plantaginea                       5.2     5
##  6 Bathycoelia thalassina                     9       6
##  7 Amrasca biguttula biguttula                9.2     7
##  8 Apiosporina morbosa                       10       8
##  9 Argyrotaenia citrana                      10       9
## 10 Ascochyta sorghi                          10      10
## # ... with 130 more rows
```

Since tables 1, 2, and 3 deal with threatened countries, instead of dealing with three tibbles, lets rename some of the columns so we can combine them and keep the different rank information separate.

```
# t1 <- rename(t1, overall_rank = rank)
# t2 <- rename(t2, total_cost_rank = rank)
# t3 <- rename(t3, prop_gdp_rank = rank)
```

```
# threatened <- t1 %>% full_join(t2, by = "country") %>%
#   full_join(t3, by = "country")

# Save the combined data
# write_csv(threatened, path = "../threatened.csv")

threatened <- read_csv("../threatened.csv")
```

```
## Parsed with column specification:
## cols(
##   overall_rank = col_double(),
##   country = col_character(),
##   invasion_threat = col_double(),
##   invasion_cost.x = col_double(),
##   total_cost_rank = col_double(),
##   invasion_cost.y = col_double(),
##   gdp_mean = col_double(),
##   gdp_proportion = col_double(),
##   prop_gdp_rank = col_double()
## )
```

After combining the data, it's noted that the invasion_cost variable is similar from two data sets (though not the same, why?). So we can remove one of these columns to simplify our tibble.

```
threatened <- threatened %>% select(-invasion_cost.y)

# Rename the column we keep
threatened <- threatened %>% rename(invasion_cost = invasion_cost.x)
```

The following shows that the original data does not contain only 124 countries, but many more than that.

```
setdiff(t1$country, t2$country)
```

```
##  [1] "Guinea-Bissau"          "El Salvador"
##  [3] "Czech Republic"         "Bosnia and Herzegovina"
##  [5] "South Africa"           "Dominican Republic"
##  [7] "Korea Republic of"      "United Kingdom"
##  [9] "Burkina Faso"           "Sri Lanka"
## [11] "New Zealand"            "Georgia (Republic)"
## [13] "Congo (Republic of)"    "Cape Verde"
## [15] "Trinidad and Tobago"    "Equatorial Guinea"
## [17] "Saudi Arabia"           "Costa Rica"
## [19] "Russian Federation"
```

Question 1

Is there a consistent relationship between the three ranks?

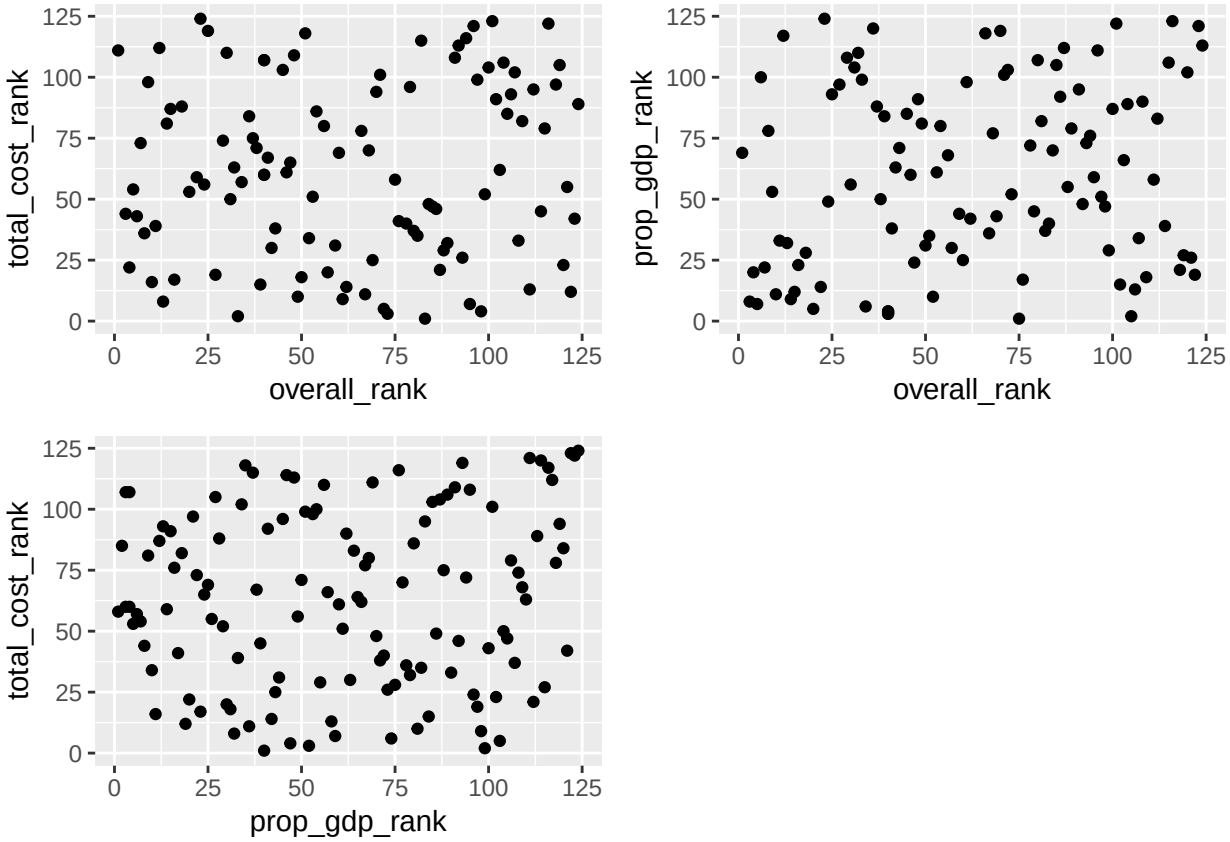
```
r1 <- threatened %>% ggplot(aes(x = overall_rank, y = total_cost_rank)) +
  geom_point()

r2 <- threatened %>% ggplot(aes(x = overall_rank, y = prop_gdp_rank)) +
  geom_point()

r3 <- threatened %>% ggplot(aes(x = prop_gdp_rank, y = total_cost_rank)) +
  geom_point()
```

```
grid.arrange(r1, r2, r3, ncol = 2)
```

```
## Warning: Removed 37 rows containing missing values (geom_point).
```

```
## Warning: Removed 37 rows containing missing values (geom_point).
```

```
## Warning: Removed 19 rows containing missing values (geom_point).
```



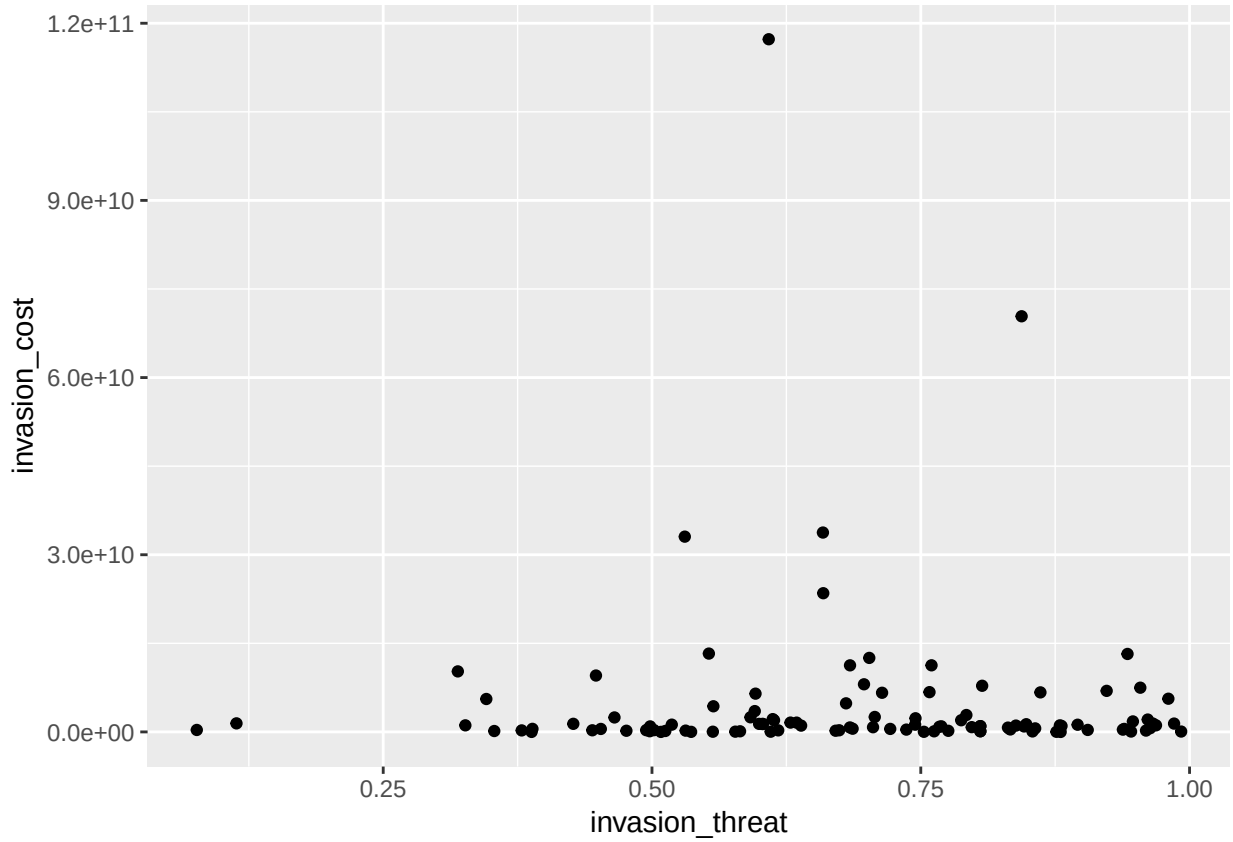
Based on the rough plots above it's safe to say there isn't any relationship between the three different rankings.

Question 2

Are those countries that have a higher level of invasion threat more likely to have a higher or lower cost?

```
threatened %>% ggplot(aes(x = invasion_threat, y = invasion_cost)) +
  geom_point()
```

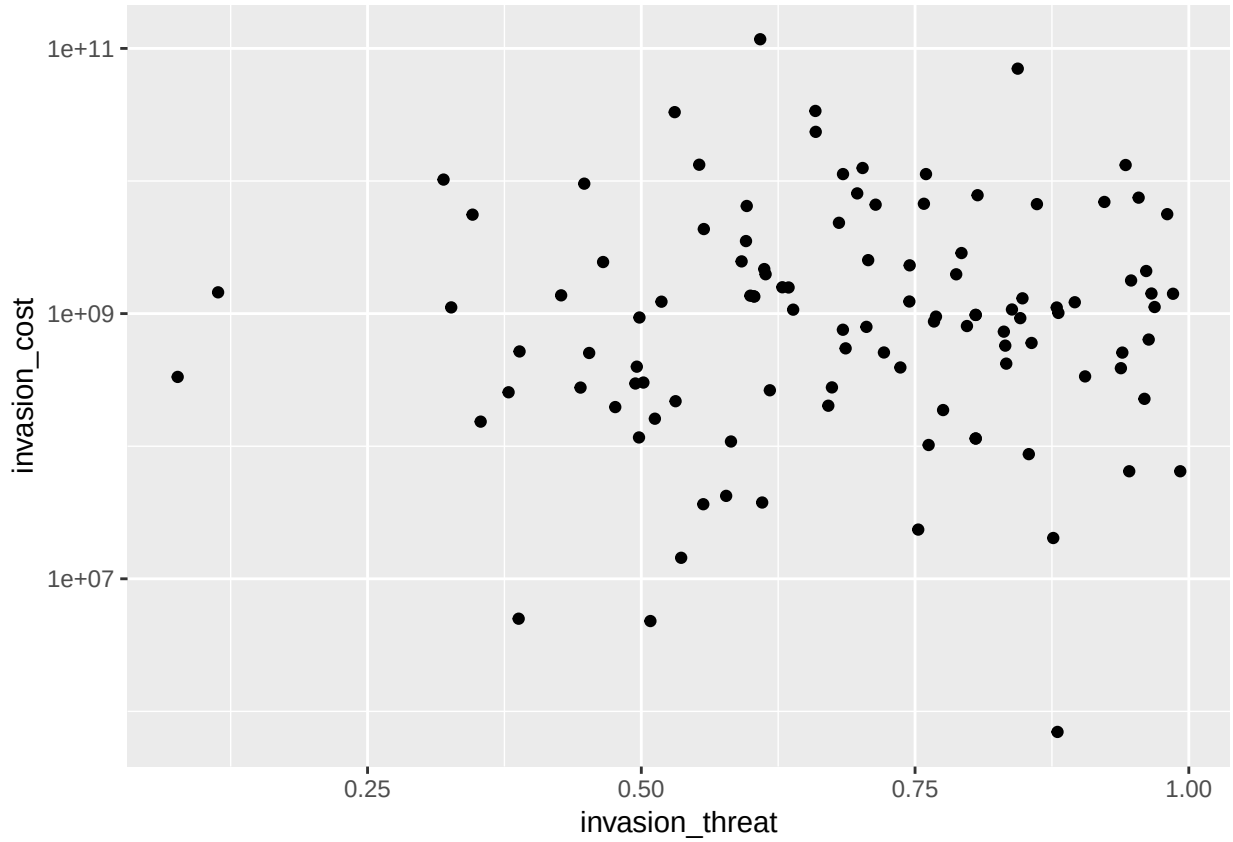
```
## Warning: Removed 37 rows containing missing values (geom_point).
```



It doesn't appear as if there's a relationship on a linear scale. What about a log scale?

```
threatened %>% ggplot(aes(x = invasion_threat, y = invasion_cost)) +
  geom_point() +
  scale_y_log10()
```

```
## Warning: Removed 37 rows containing missing values (geom_point).
```



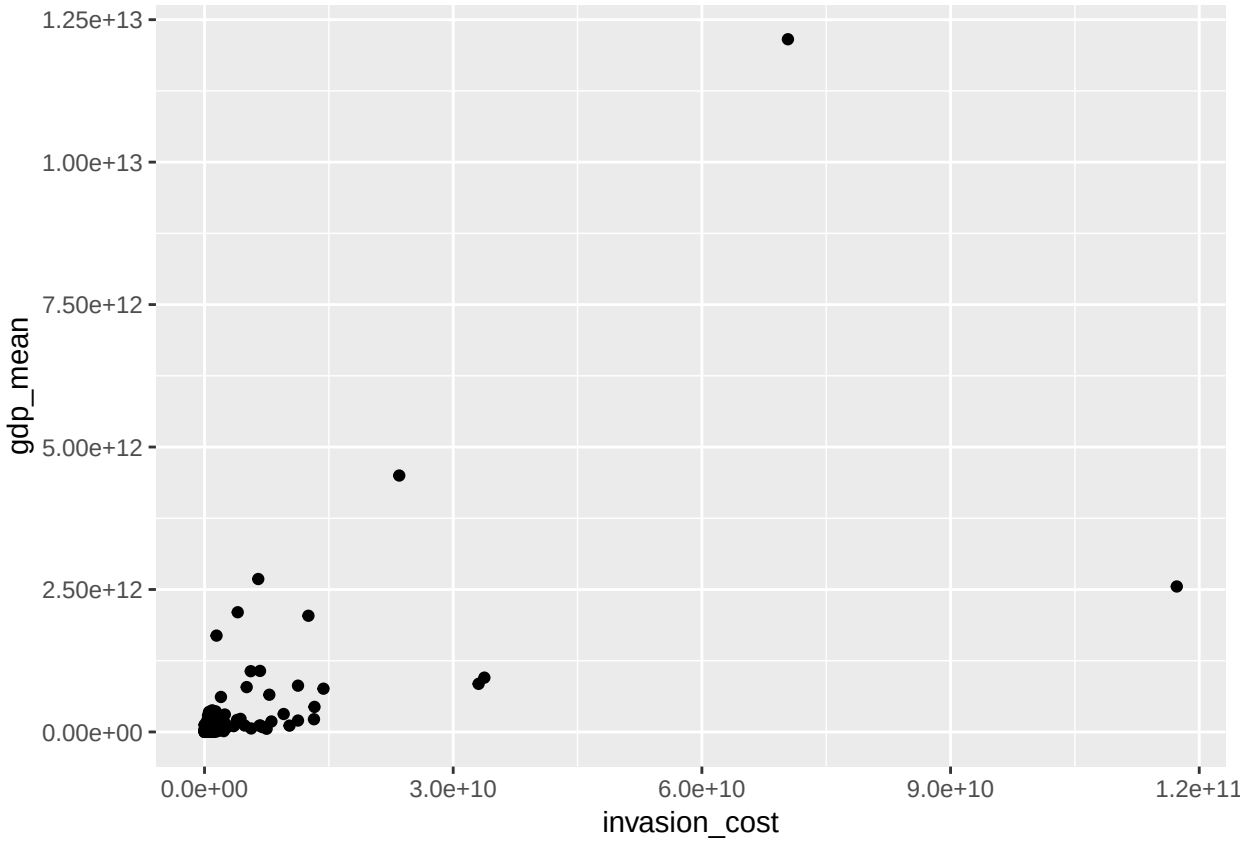
This doesn't appear to show much of a general pattern either, but it does give me a hint that log10 transformation may be quite useful to get points more appropriately spaced.

Question 3

Is there a relationship between invasion cost and the mean gdp of a country? It would make sense that a country with a larger mean GDP would have more chance of succumbing to larger financial costs if an invasive species were to spread throughout it.

```
threatened %>%
  ggplot(aes(x = invasion_cost, y = gdp_mean)) +
  geom_point()
```

```
## Warning: Removed 19 rows containing missing values (geom_point).
```

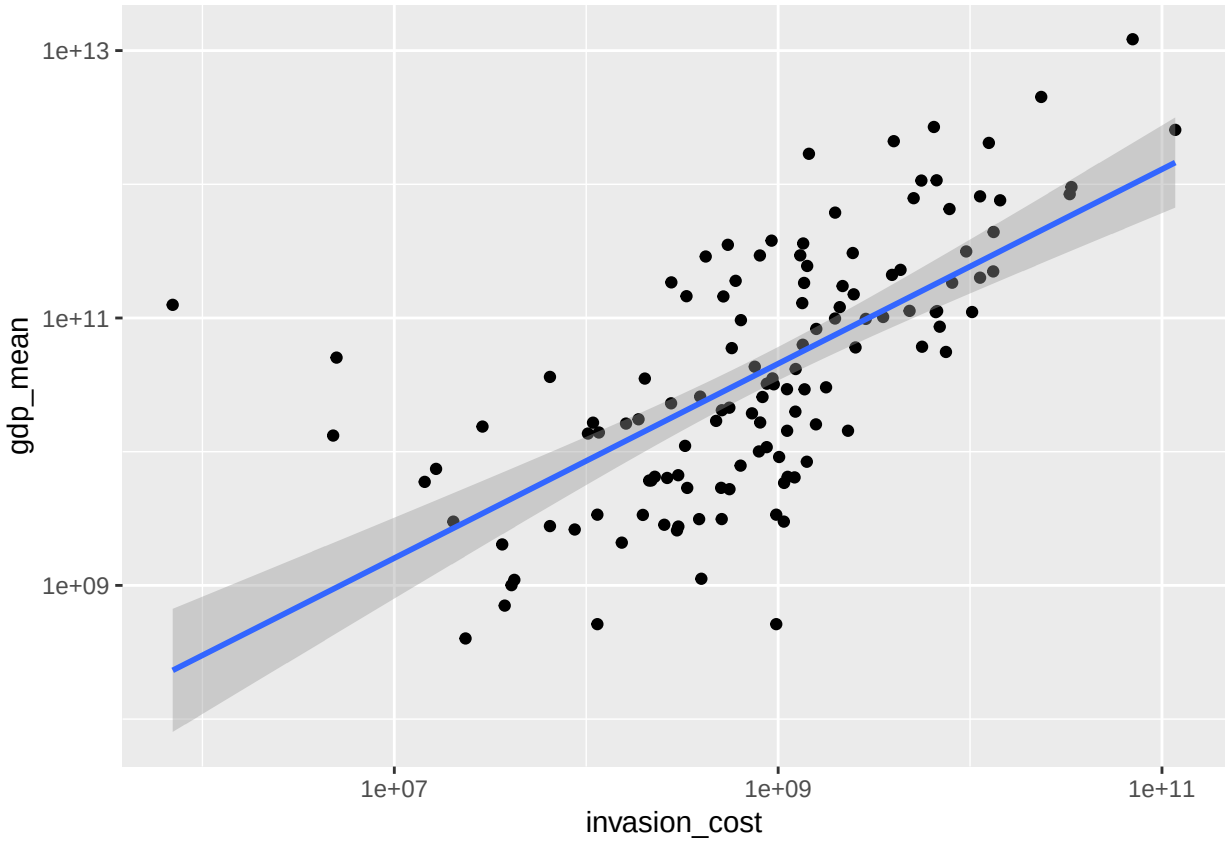


A few points are throwing interpretation of most points off, so lets change the scale.

```
threatened %>%
  ggplot(aes(x = invasion_cost, y = gdp_mean)) +
  geom_point() +
  scale_x_log10() +
  scale_y_log10() +
  geom_smooth(method = "lm")
```

```
## Warning: Removed 19 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 19 rows containing missing values (geom_point).
```

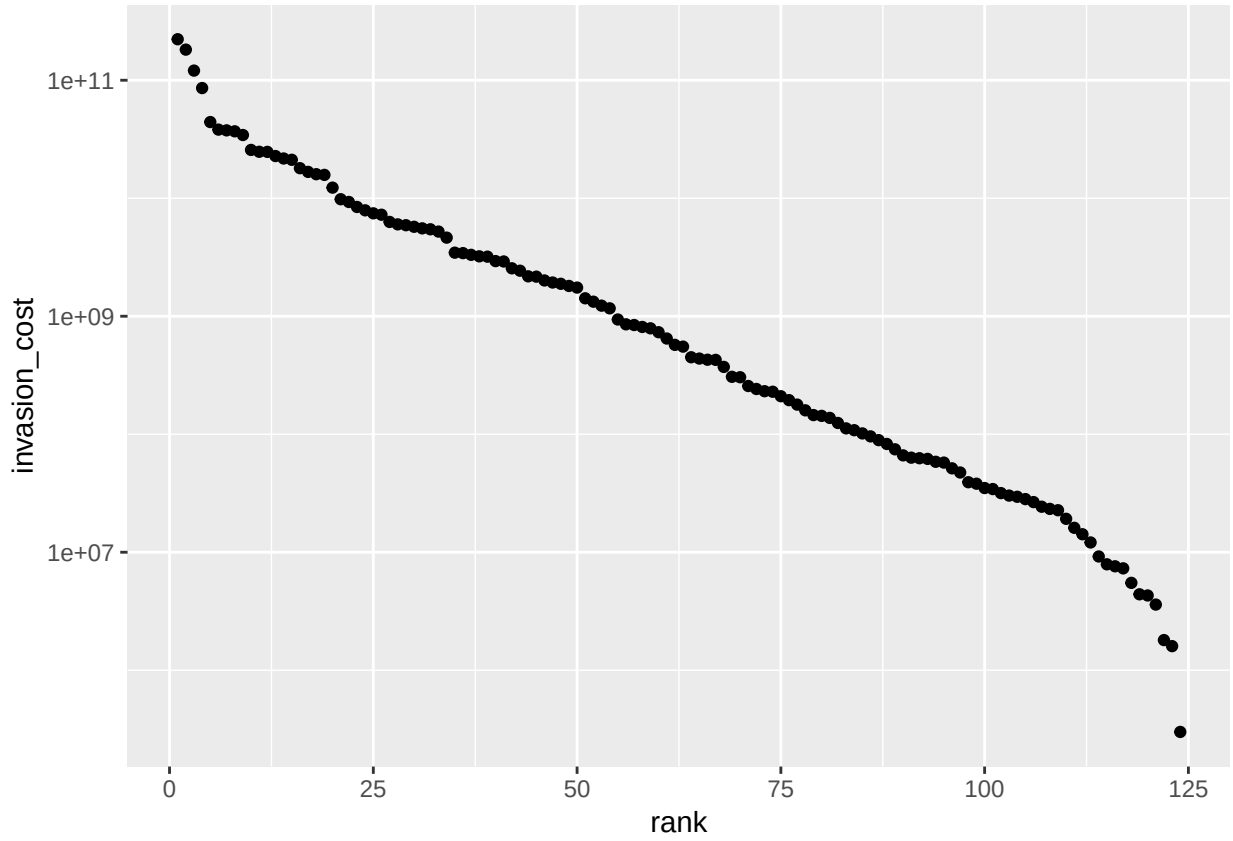


It appears there is a bit of a relationship between invasion cost and mean GDP when both are compared on a log10 scale. This could be something to be investigated a bit more later.

Question

Investigating the source country data set

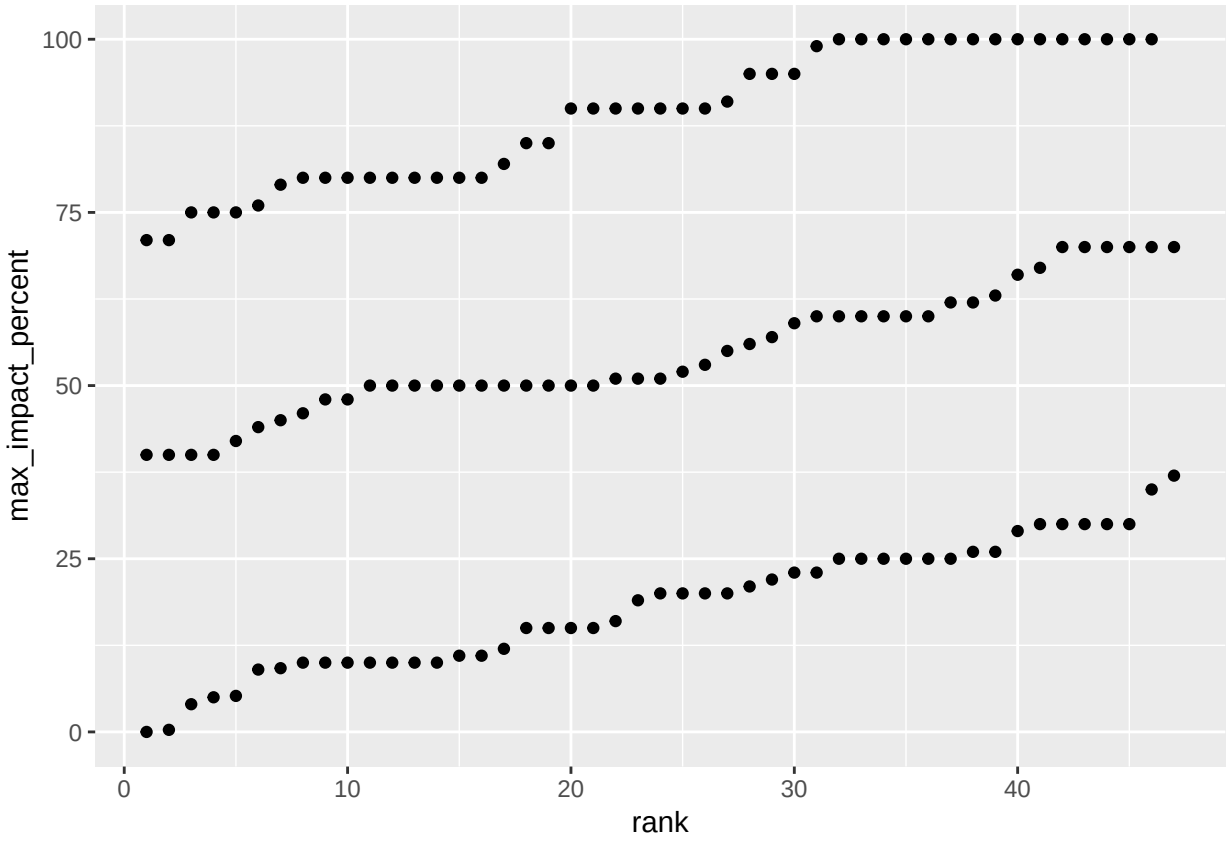
```
t4 %>%
  ggplot(aes(x = rank, y = invasion_cost)) +
  geom_point() +
  scale_y_log10()
```



Question

Investigating the species data set.

```
t6 %>% ggplot(aes(x = rank, y = max_impact_percent)) +
  geom_point()
```



The above plot is very interesting for the fact that the ranks are from 1-47, and all but 47 are repeated three times. This shows that the ranking of the species needs a bit more investigating.

Combining source and threatened data

```
# source_threatened <- t4 %>% rename(source_invasion_cost = invasion_cost,
#                                    source_rank = rank) %>% full_join(threatened, by = "country")

#write_csv(source_threatened, path = "../source_threatened.csv")

source_threatened <- read_csv("../source_threatened.csv")
```

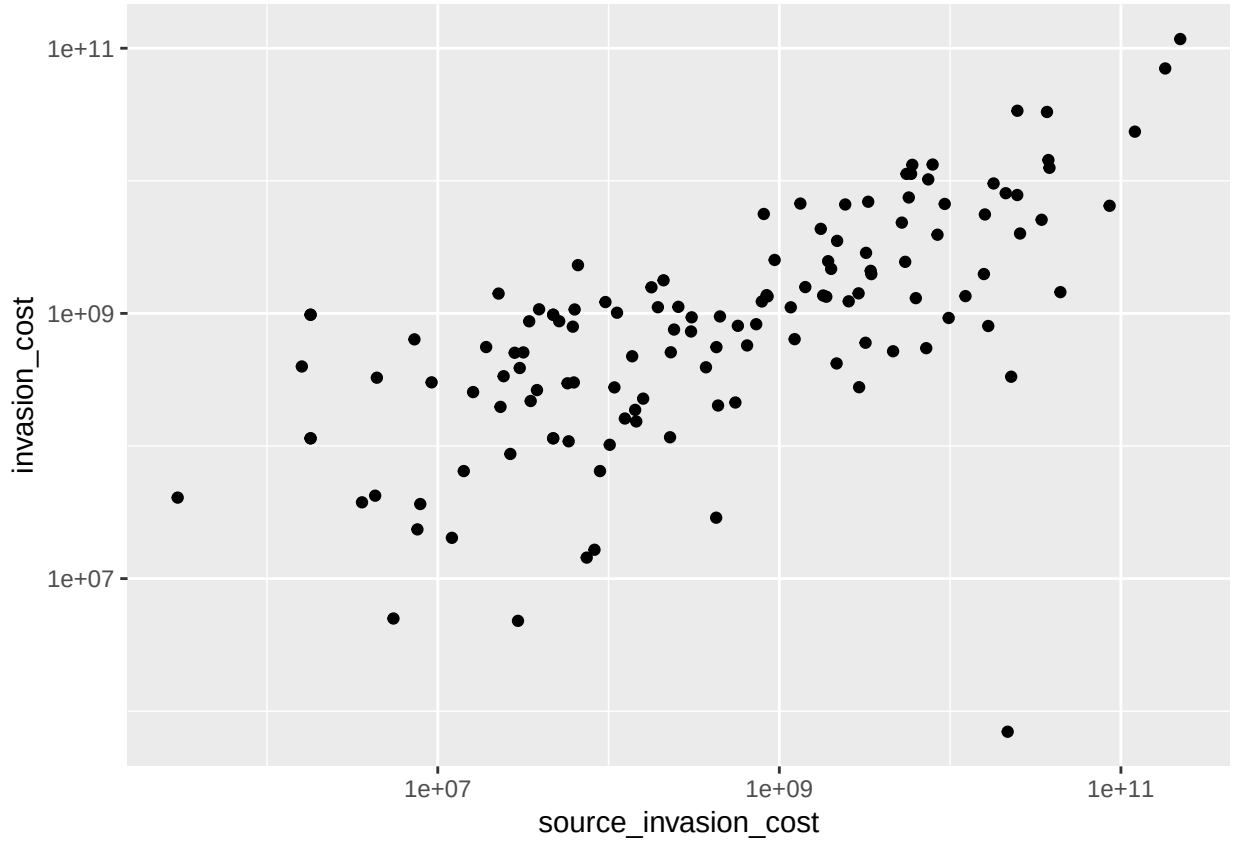
```
## Parsed with column specification:
## cols(
##   country = col_character(),
##   source_invasion_cost = col_double(),
##   source_rank = col_double(),
##   overall_rank = col_double(),
##   invasion_threat = col_double(),
##   invasion_cost = col_double(),
##   total_cost_rank = col_double(),
##   gdp_mean = col_double(),
##   gdp_proportion = col_double(),
##   prop_gdp_rank = col_double()
## )
```

Question

Is there a relationship between the total cost of a source country and it's cost as a threatened country?

```
source_threatened %>% ggplot(aes(x = source_invasion_cost, y = invasion_cost)) +
  geom_point() +
  scale_y_log10() + scale_x_log10()
```

```
## Warning: Removed 19 rows containing missing values (geom_point).
```

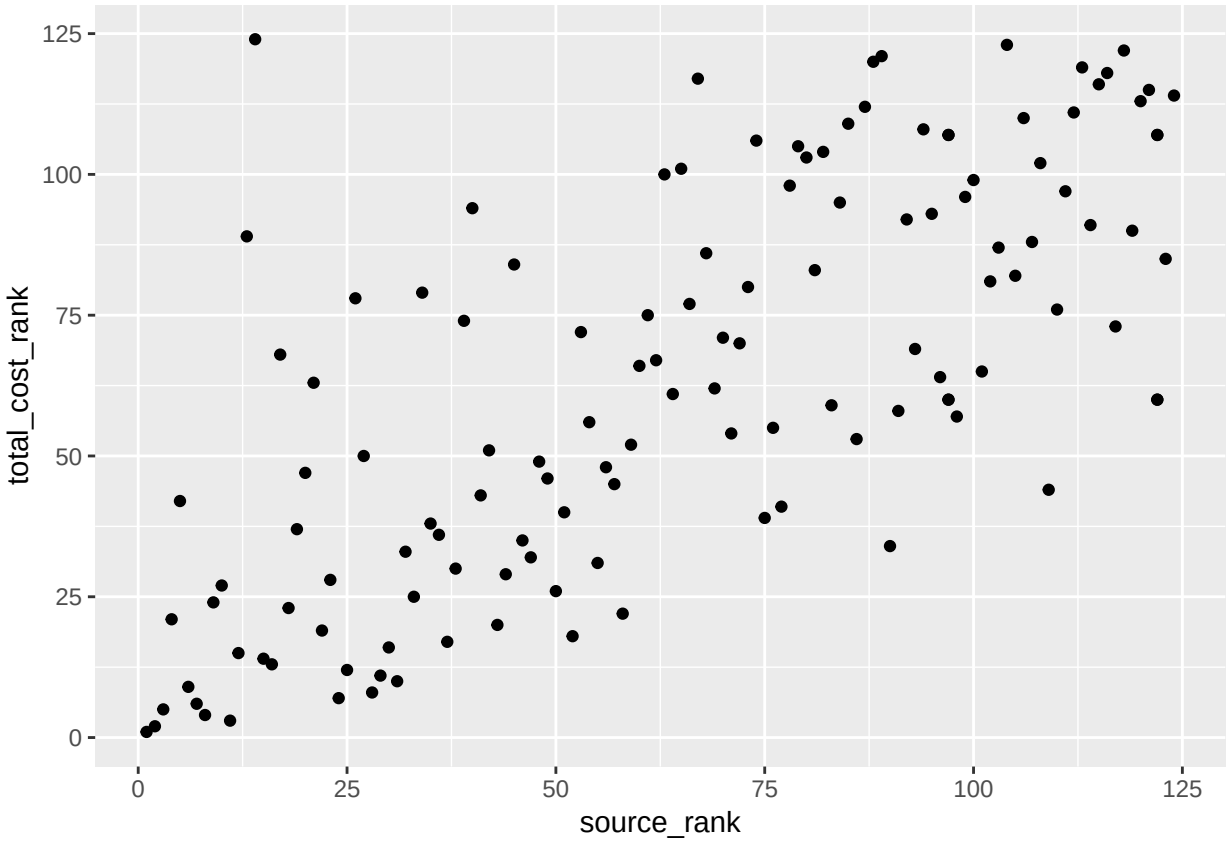


Question

Is there a relationship between the source rank and the total cost rank (should be similar as the previous grap I believe)

```
source_threatened %>%
  ggplot(aes(x = source_rank, y = total_cost_rank)) +
  geom_point()
```

```
## Warning: Removed 19 rows containing missing values (geom_point).
```



There does appear to be a similar pattern, though it is weaker.