The most popular museum _Analysis+WorkingProcess

An Introduction to the Data

This data set is collected by annecool37 from TripAdvisor. As an museum fun I am always interested in how I could select museum from internet in a more effect way, or in the other world, what are the effective factors for select a good museum?

Clean up my Data

In order to work with this data, we need to clean and categorized the data set first.

step 1: Delete duplicate data

step 2: Sort through data to check erro

After first clean up, there are nearly 1,000 museums to work with. In order to work easily, I decied to categorize museum by country and continet.

step 3: Country classification

step 4: Continet Classification

With this clean data set, I start my general analysis

```
#Import the data
library(ggplot2)

museum <- read_csv(file = "data/museum.csv")
```

```
## Warning: Missing column names filled in: 'X23' [23], 'X24' [24], 'X25' [25],
## 'X26' [26]
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   Me = col_double(),
##   Count = col_double(),
##   Mcount = col_double(),
##   FeatureCount = col_double(),
##   longtitude = col_double(),
##   latitude = col_double(),
##   Rank = col_double(),
##   Rating = col_double(),
##   ReviewCount = col_number(),
##   TotalThingsToDo = col_number(),
##   X23 = col_logical(),
##   X24 = col_logical(),
##   X25 = col_logical(),
##   X26 = col_logical(),
##   ColorCode = col_logical()
## )
```

```
## See spec(...) for full column specifications.
```

```r
summary(museum)
```

```
##        Me          Address              Count        Continent
##  Min.   :1.000   Length:871         Min.   :  7.0   Length:871
##  1st Qu.:1.000   Class :character   1st Qu.:311.0   Class :character
##  Median :1.000   Mode  :character   Median :311.0   Mode  :character
##  Mean   :1.493                      Mean   :309.4
##  3rd Qu.:2.000                      3rd Qu.:397.0
##  Max.   :2.000                      Max.   :397.0
##  NA's   :798
##     sov_a3            Country              Mcount       Description
##  Length:871         Length:871         Min.   :  1.00   Length:871
##  Class :character   Class :character   1st Qu.:  8.00   Class :character
##  Mode  :character   Mode  :character   Median : 17.00   Mode  :character
##                                        Mean   : 26.19
##                                        3rd Qu.: 37.00
##                                        Max.   :283.00
##                                        NA's   :282
##   FeatureCount        Fee              longtitude           latitude
##  Min.   : 0.000   Length:871         Min.   :-157.9583   Min.   :-68.93
##  1st Qu.: 0.000   Class :character   1st Qu.: -79.3974   1st Qu.: 35.03
##  Median : 0.000   Mode  :character   Median :  -0.1355   Median : 41.39
##  Mean   : 1.505                      Mean   : -13.5078   Mean   : 38.05
##  3rd Qu.: 2.000                      3rd Qu.:  16.3600   3rd Qu.: 50.09
##  Max.   :27.000                      Max.   : 176.2603   Max.   : 69.65
##
##  LengthOfVisit       MuseumName          PhoneNum             Rank
##  Length:871         Length:871         Length:871         Min.   :  1
##  Class :character   Class :character   Class :character   1st Qu.:  2
##  Mode  :character   Mode  :character   Mode  :character   Median :  7
##                                                           Mean   : 17
##                                                           3rd Qu.: 20
##                                                           Max.   :397
##
##      Rating       ReviewCount      TotalThingsToDo  MuseumLoctation
##  Min.   :2.500   Min.   :   46.0   Min.   :   1.0   Length:871
##  1st Qu.:4.500   1st Qu.:  502.5   1st Qu.:  96.5   Class :character
##  Median :4.500   Median :  925.0   Median : 213.0   Mode  :character
##  Mean   :4.424   Mean   : 2347.0   Mean   : 330.0
##  3rd Qu.:4.500   3rd Qu.: 2018.5   3rd Qu.: 423.0
##  Max.   :5.000   Max.   :63114.0   Max.   :2279.0
##
##  MuseumTopic           Color              X23             X24
##  Length:871         Length:871         Mode:logical   Mode:logical
##  Class :character   Class :character   NA's:871       NA's:871
##  Mode  :character   Mode  :character
##
##
##
##
##    X25             X26           ColorCode
##  Mode:logical   Mode:logical   Mode:logical
##  NA's:871       NA's:871       NA's:871
##
```

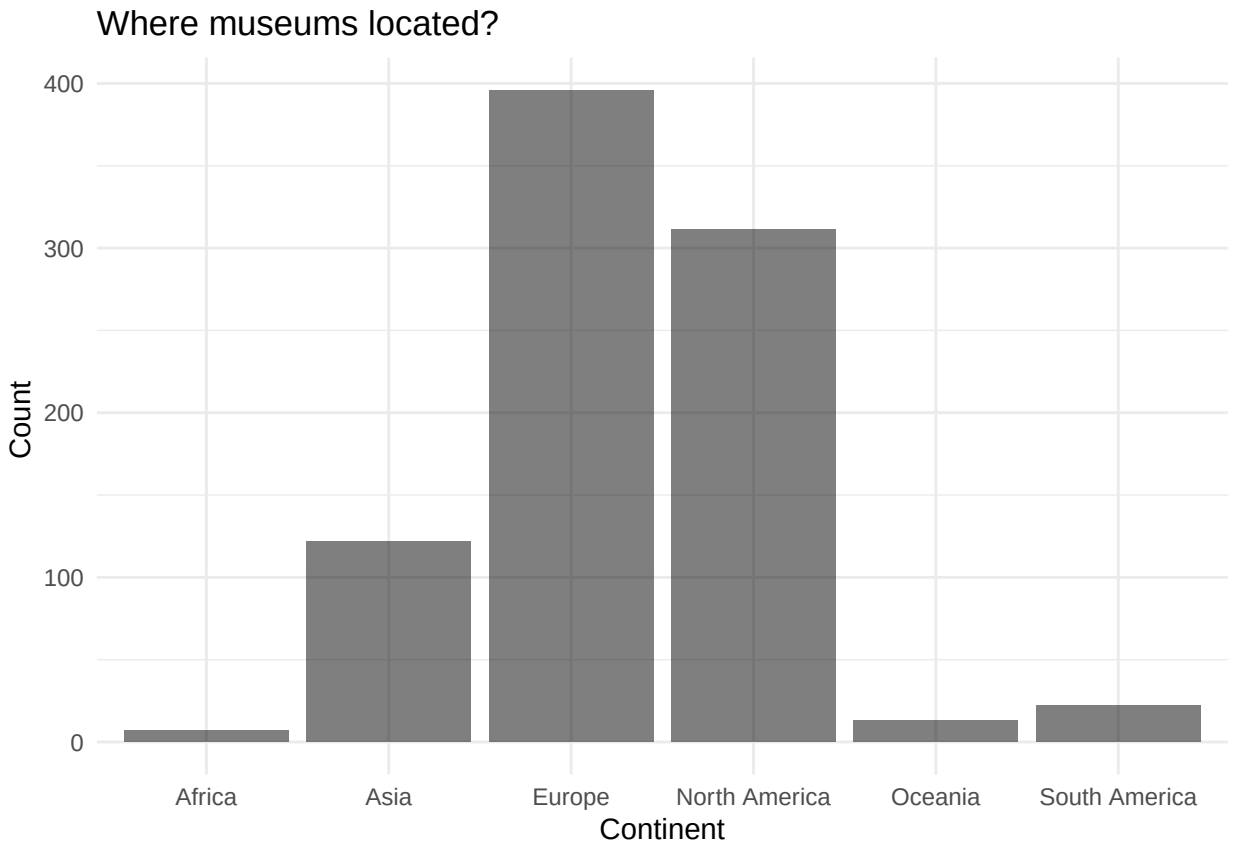
```
##
##
##
##
```

```
# group museums by continent

Eu <- museum  %>%
  filter(grepl ("Europe", Continent))


As <- museum  %>%
  filter(grepl ("Asia", Continent))

Af <- museum  %>%
  filter(Continent == "Africa")

Na<- museum  %>%
  filter(grepl ("North America", Continent))

Sa<- museum  %>%
  filter(grepl ("South America", Continent))

Oc<- museum  %>%
  filter(grepl ("Oceania", Continent))
```

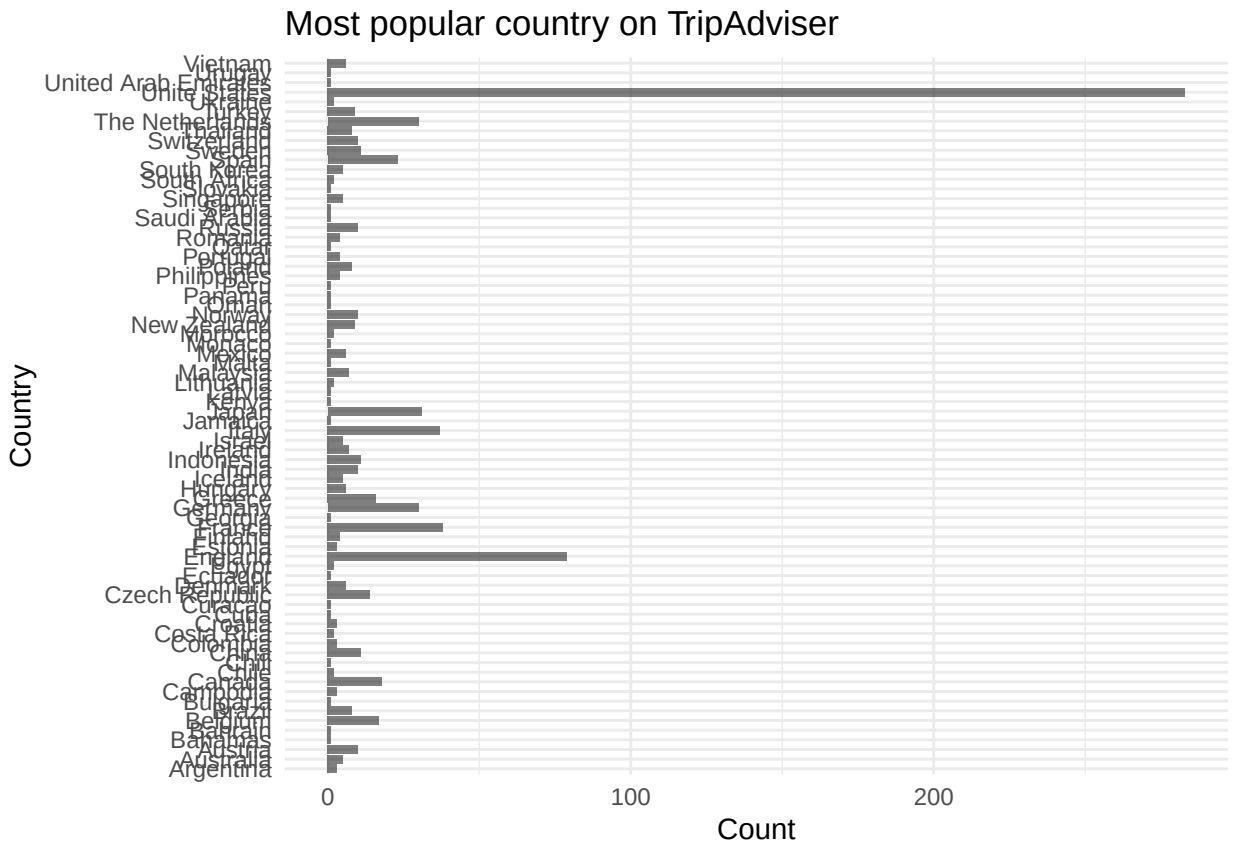
Does number of museums refelct the quality of museums?

My first step of analysis focus on continet and country. Does users on Tripadviser has a perference in location? If so, the count of museums might effect the review of the museums.

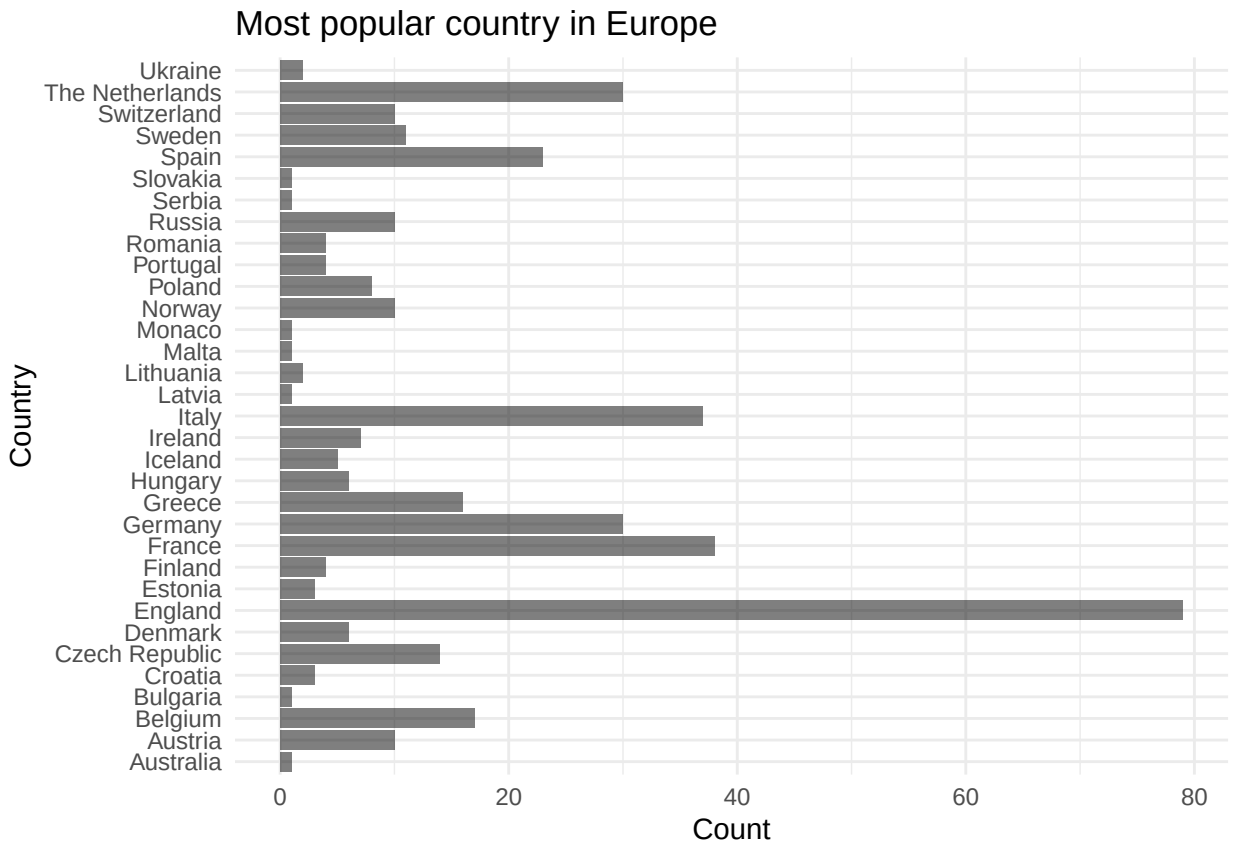
```
library(ggplot2)

# museums count by continent

ggplot(museum, aes(x=Continent)) +
  geom_bar(fill="black", alpha = 0.5 )+
  labs(x="Continent", y="Count ",title="Where museums located?") +
  theme_minimal()
```



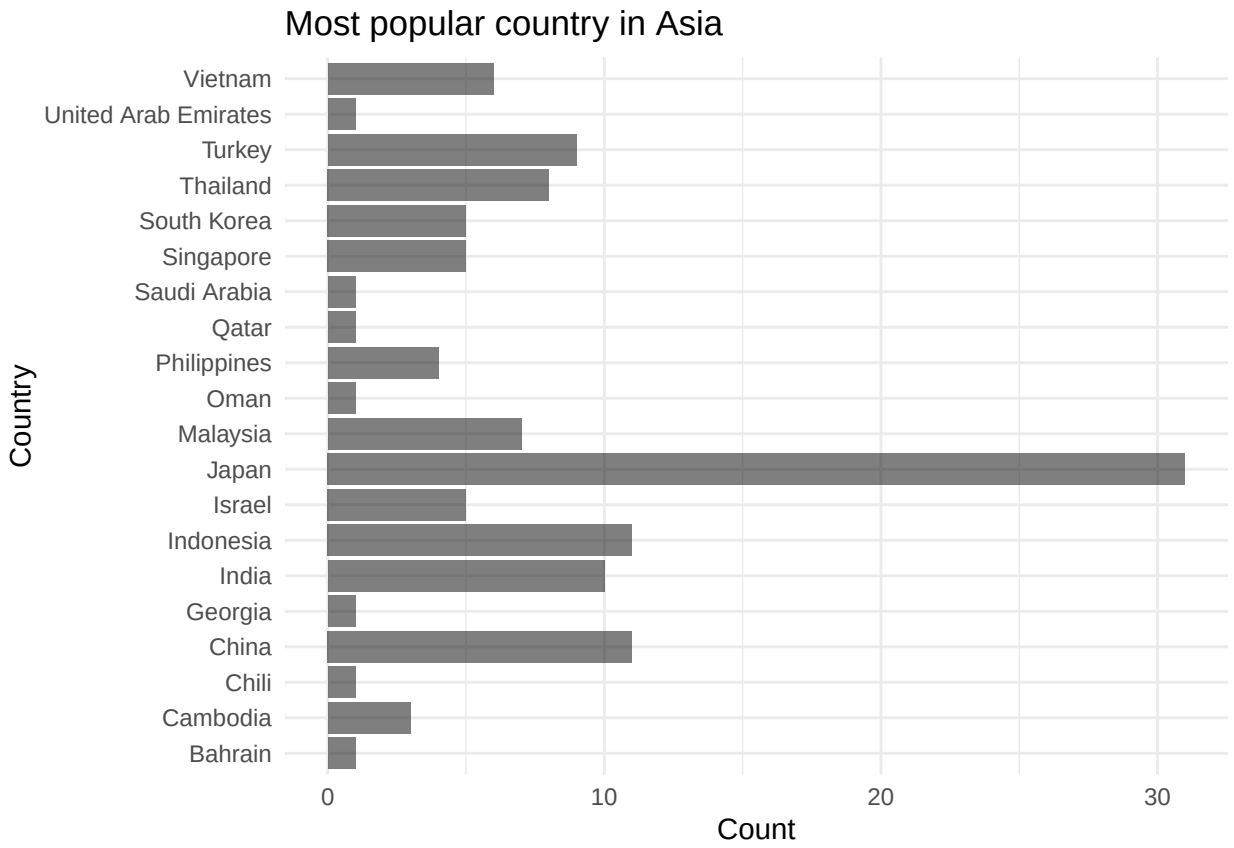

```
#museums count by country

ggplot(museum, aes(y=Country)) +
  geom_bar(position="stack", fill="black", alpha = 0.5)+
  labs(x="Count", y="Country",title="Most popular country on TripAdviser") +
  theme_minimal()
```



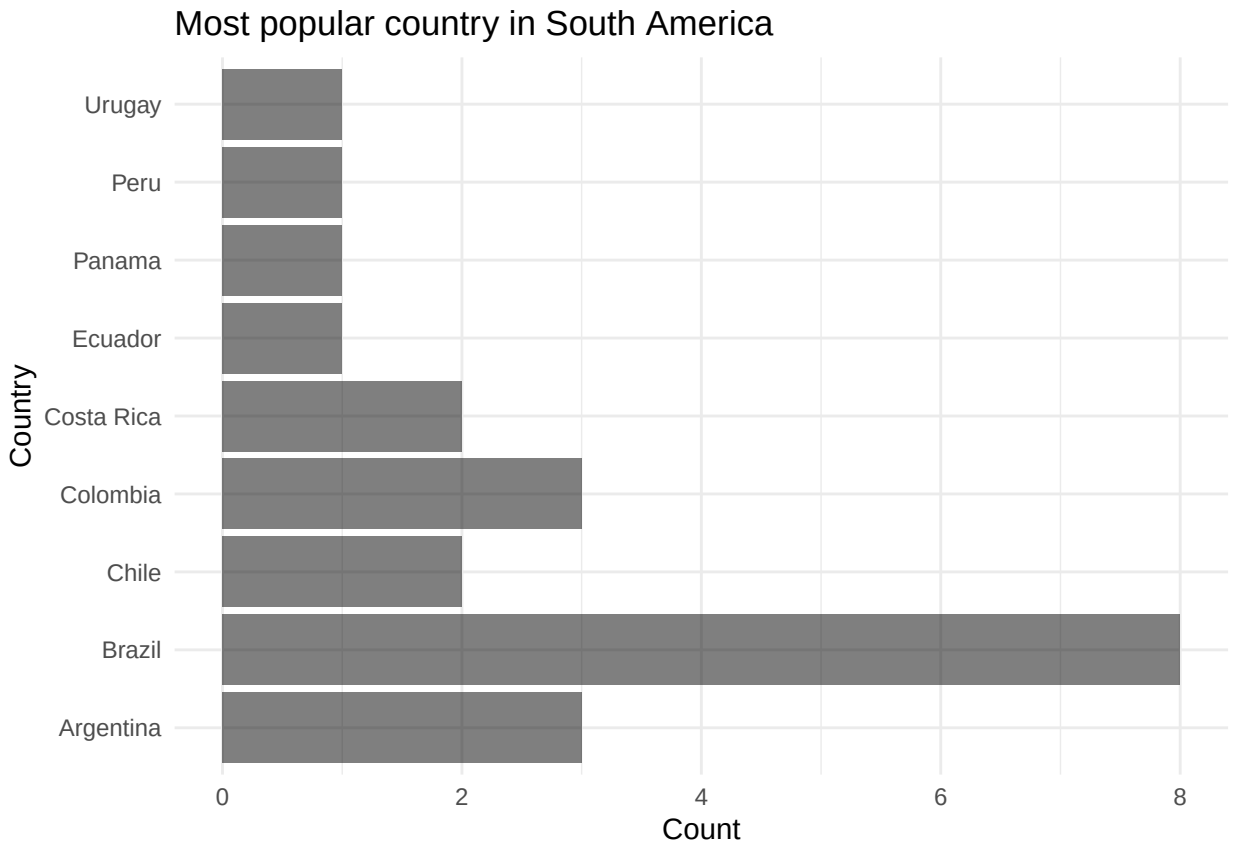

```
#country+continent

ggplot(Eu, aes(y=Country)) +
  geom_bar(position="stack", fill="black", alpha = 0.5)+
  labs(x="Count", y="Country",title="Most popular country in Europe") +
  theme_minimal()
```



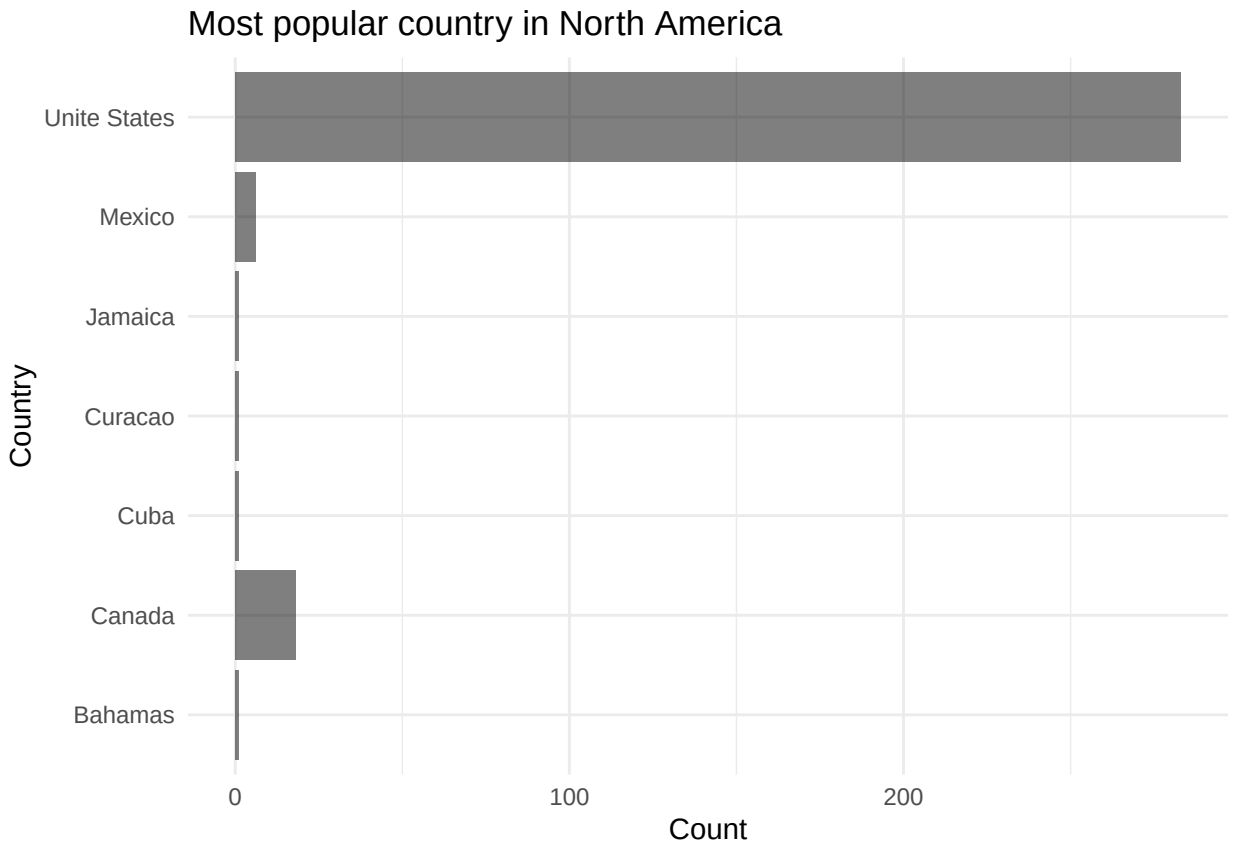

```
ggplot(As, aes(y=Country)) +
  geom_bar(position="stack", fill="black", alpha = 0.5)+
  labs(x="Count", y="Country",title="Most popular country in Asia") +
  theme_minimal()
```



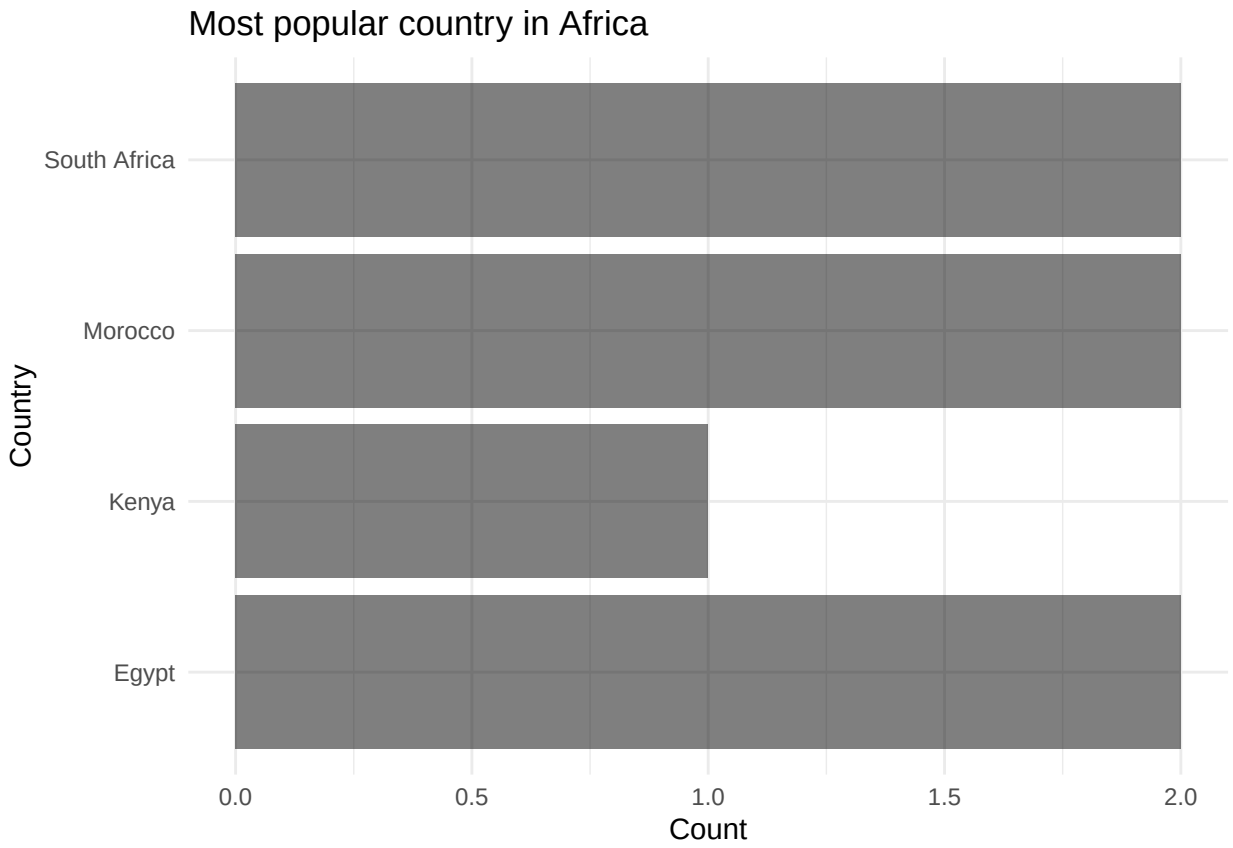

```
ggplot(Sa, aes(y=Country)) +
  geom_bar(position="stack", fill="black", alpha = 0.5)+
  labs(x="Count", y="Country",title="Most popular country in South America") +
  theme_minimal()
```



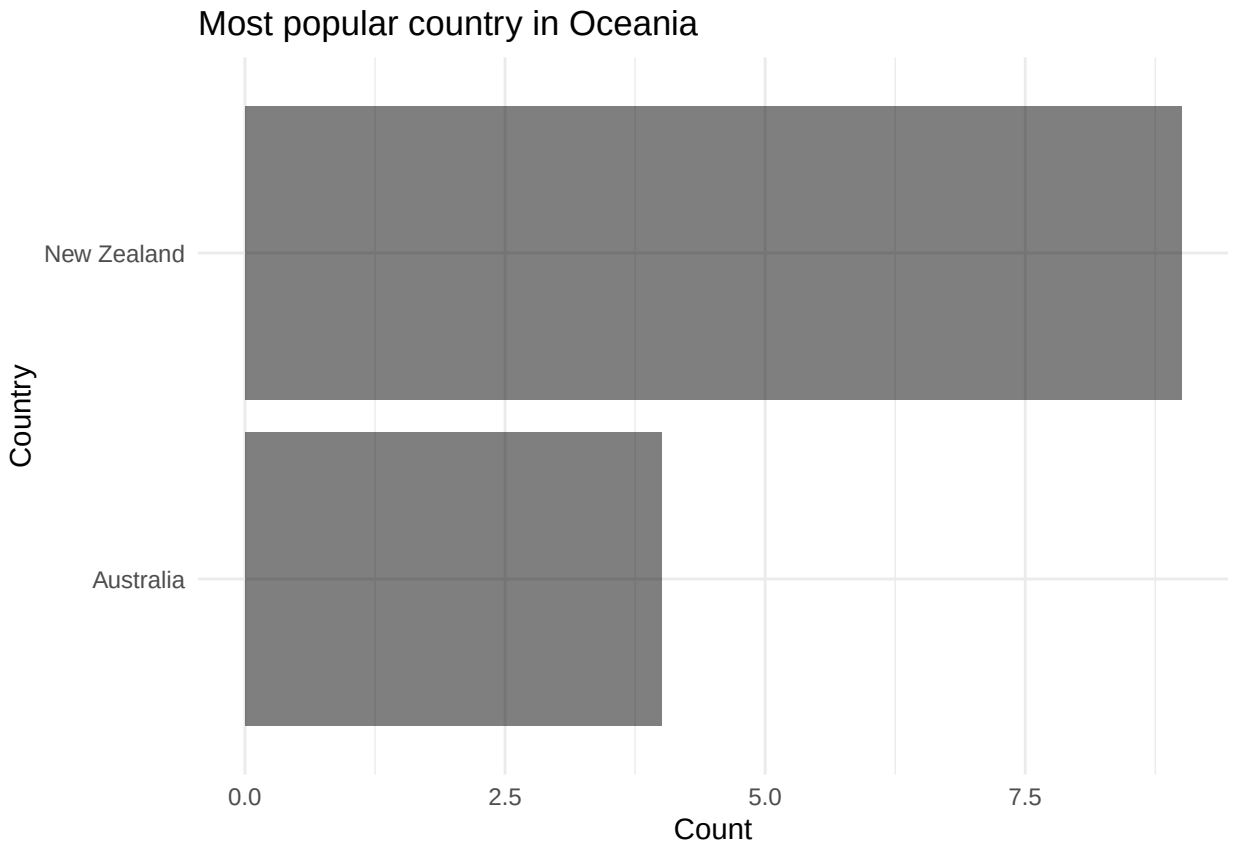

```
ggplot(Na, aes(y=Country)) +
  geom_bar(position="stack", fill="black", alpha = 0.5)+
  labs(x="Count", y="Country",title="Most popular country in North America") +
  theme_minimal()
```




```
ggplot(Af, aes(y=Country)) +
  geom_bar(position="stack", fill="black", alpha = 0.5)+
  labs(x="Count", y="Country",title="Most popular country in Africa") +
  theme_minimal()
```




```
ggplot(Oc, aes(y=Country)) +
  geom_bar(position="stack", fill="black", alpha = 0.5)+
  labs(x="Count", y="Country",title="Most popular country in Oceania") +
  theme_minimal()
```

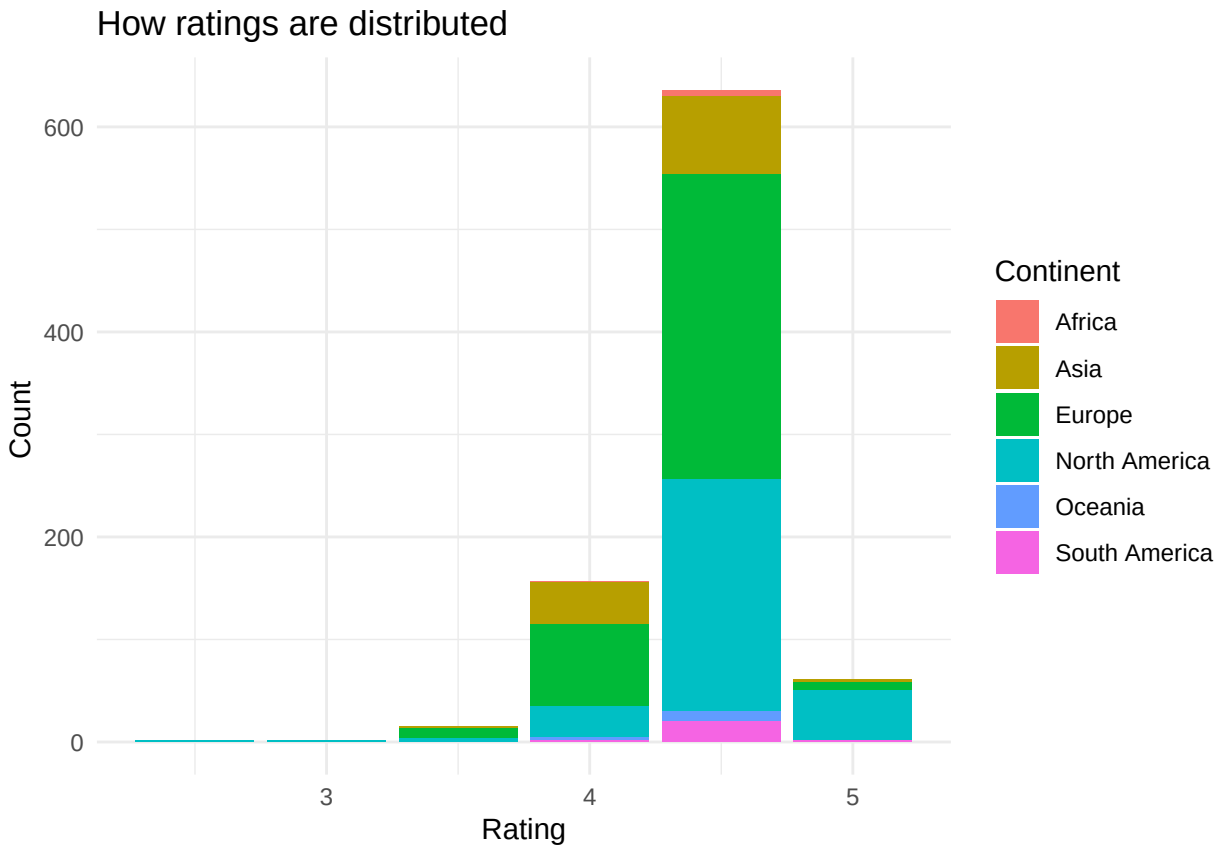


Clearly, Europe and North America have the most rated museums on TripAdviser. Also worth notice that United States have almost 400% more museums on the list.But it does not means museum in United States are in generally better than others. TripAdviser is an America company which means a huge number of users are American or English Speaker, it does not directly reflet the quality of the museums.

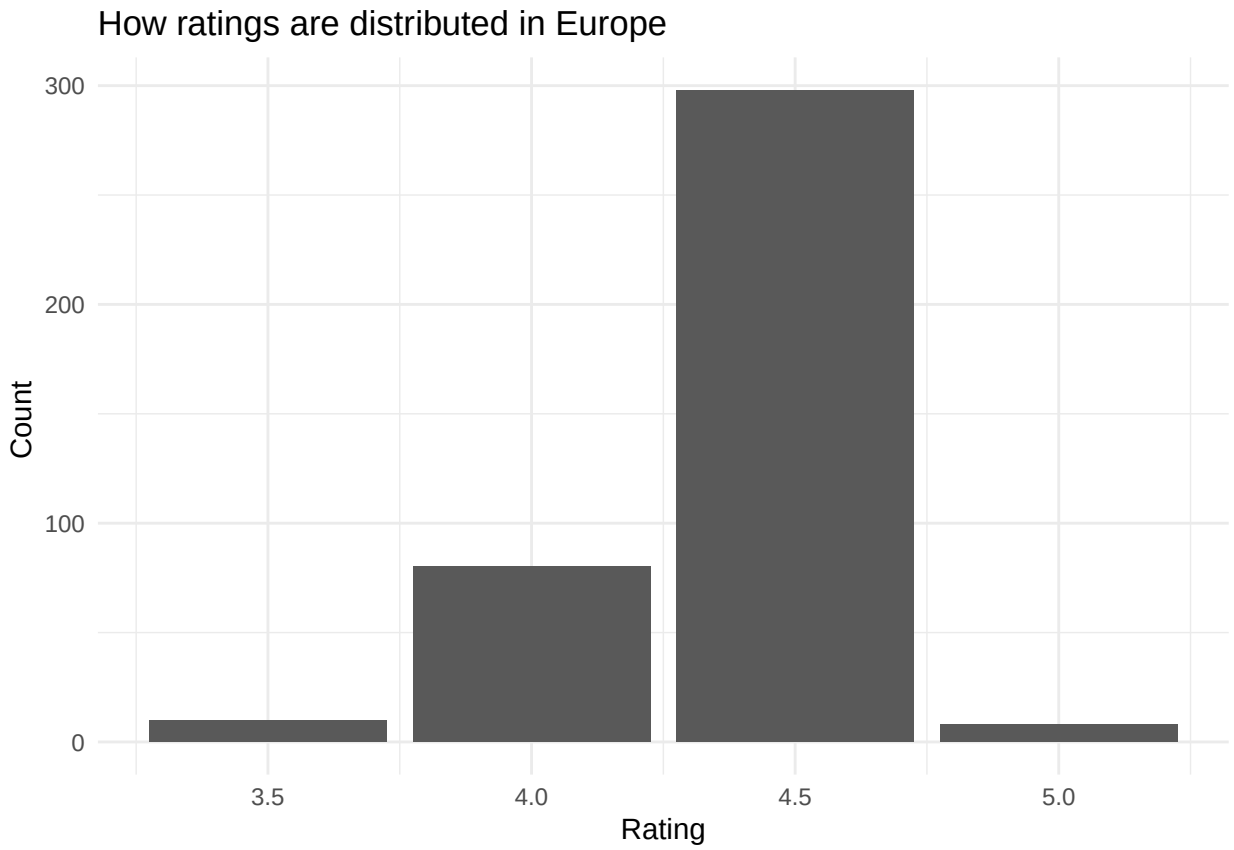
Does rating of museums refelct the quality of museums?

How rating distrubution looks like by geo?

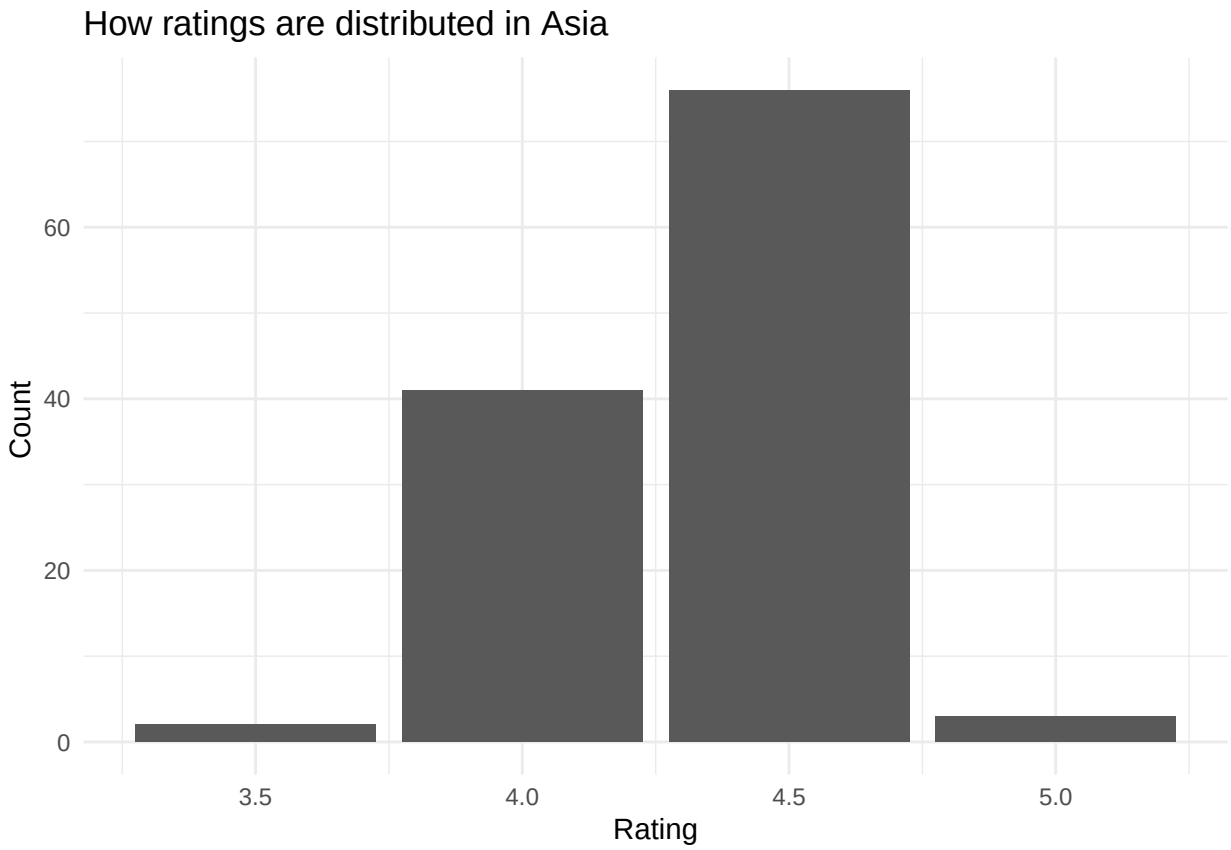
```
#museum rating stack chart by Continent

ggplot(museum, aes(x=Rating, fill=Continent)) +
  geom_bar(position="stack")+
  labs(x="Rating", y="Count ",title="How ratings are distributed") +
  theme_minimal()
```



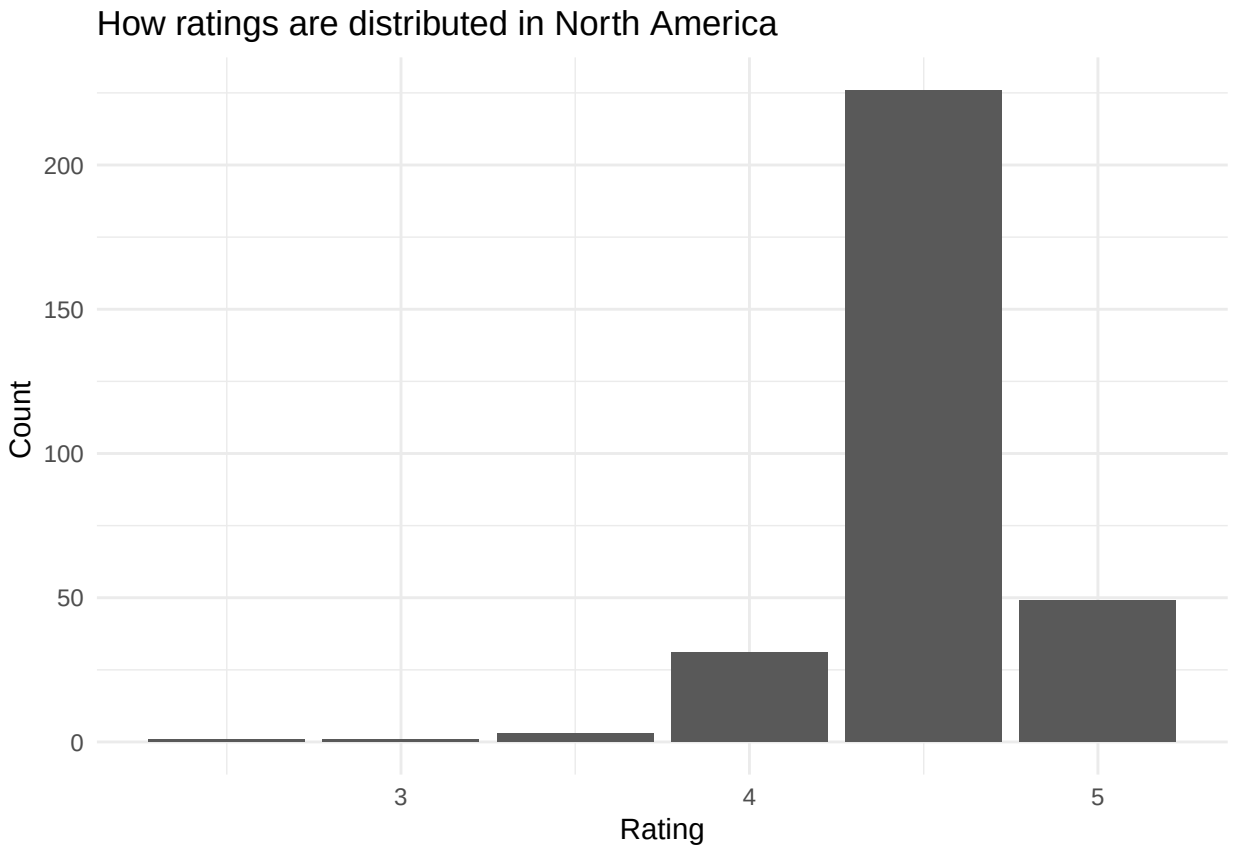

```
#museum rating stack chart by Country

ggplot(Eu, aes(x=Rating)) +
  geom_bar(position="stack")+
  labs(x="Rating", y="Count ",title="How ratings are distributed in Europe") +
  theme_minimal()
```



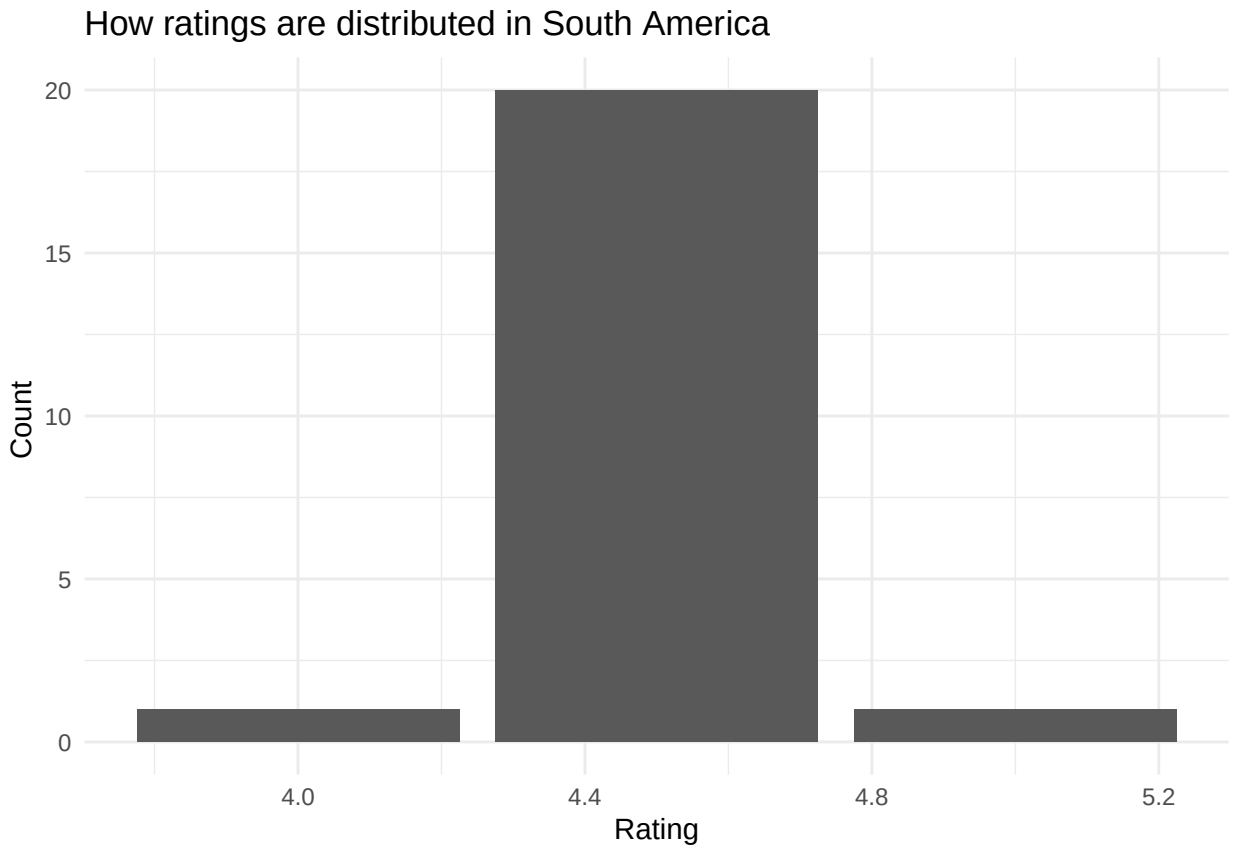

```
ggplot(As, aes(x=Rating)) +
  geom_bar(position="stack")+
  labs(x="Rating", y="Count ",title="How ratings are distributed in Asia") +
  theme_minimal()
```



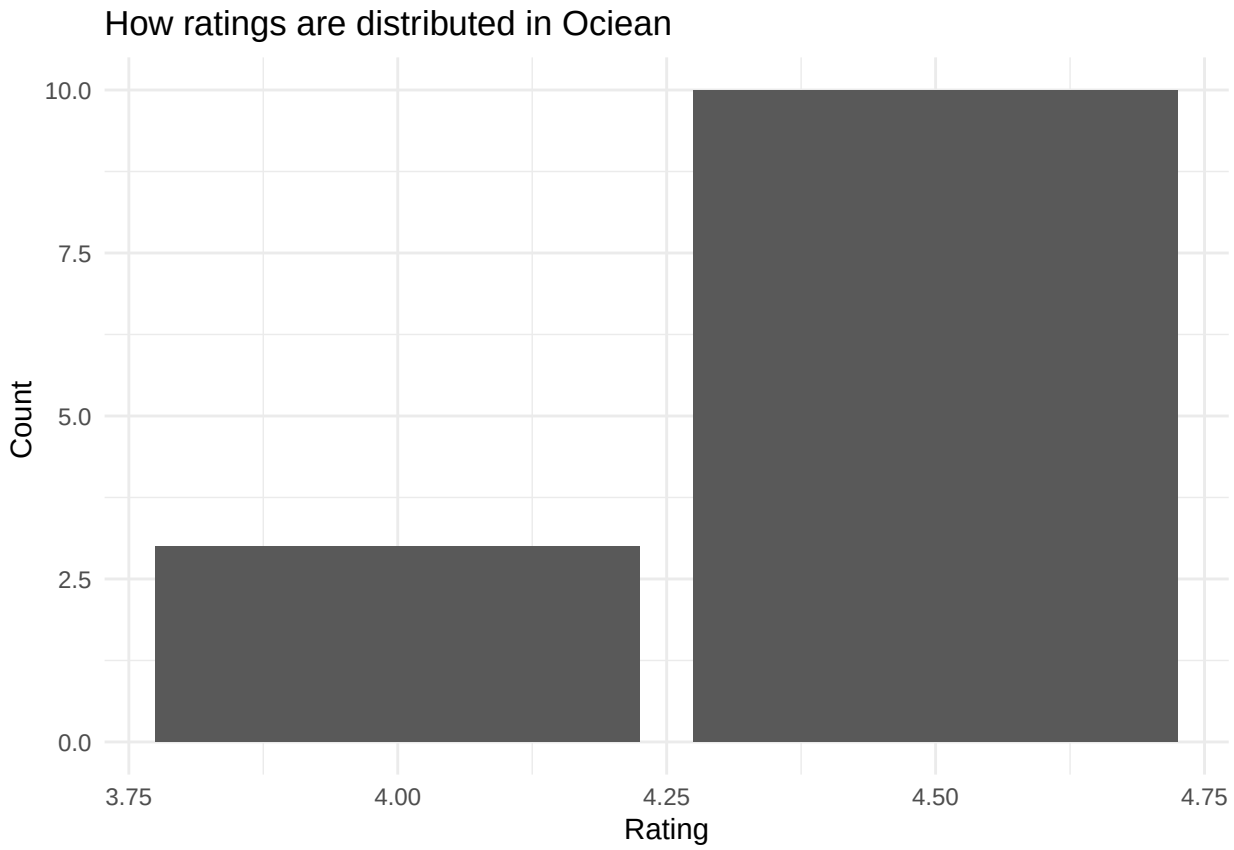

```
ggplot(Na, aes(x=Rating)) +
  geom_bar(position="stack")+
  labs(x="Rating", y="Count ",title="How ratings are distributed in North America") +
  theme_minimal()
```



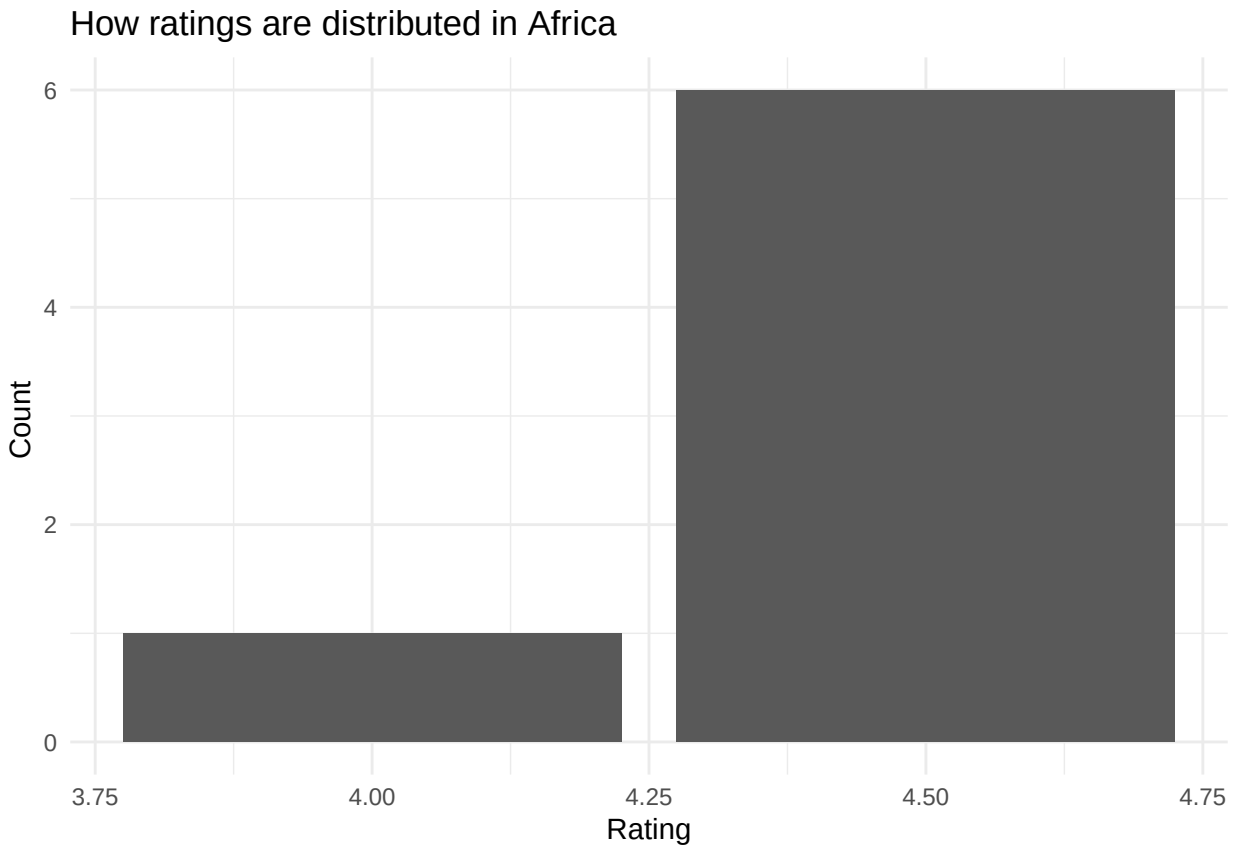

```
ggplot(Sa, aes(x=Rating)) +
  geom_bar(position="stack")+
  labs(x="Rating", y="Count ",title="How ratings are distributed in South America") +
  theme_minimal()
```




```
ggplot(Oc, aes(x=Rating)) +
  geom_bar(position="stack")+
  labs(x="Rating", y="Count ",title="How ratings are distributed in Ociean") +
  theme_minimal()
```




```
ggplot(Af, aes(x=Rating)) +
  geom_bar(position="stack")+
  labs(x="Rating", y="Count ",title="How ratings are distributed in Africa") +
  theme_minimal()
```



Clearly most museums are rate between 4 to 4.5, no matter which country and continent it located. There are many museums has the same rate, it also has no direct connection with the quality of the museums.

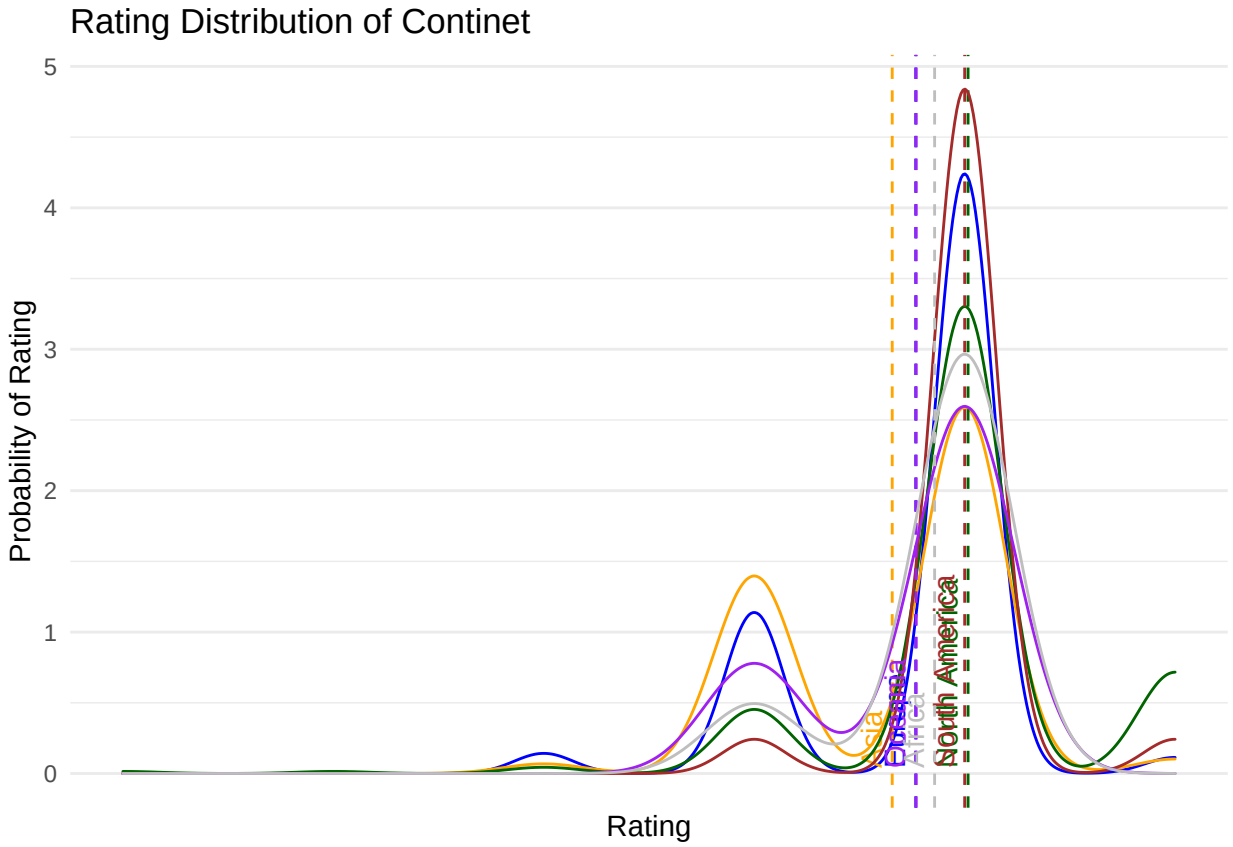
However, I find something very interesting, seems museums in North America are generally higher rated than others (most are rate as 4.5). But it does not means museums in North America are better then the rest of the world.

Does North America rated heigher than other places? (FAILED!!!)

I start to concluated mean, median for my data set.

```
ggplot()+

  ggtitle("Rating Distribution of Continet")+
  geom_density(data=Eu,aes(x=Rating), color="blue") +
  geom_vline( xintercept = mean(Eu$Rating), linetype = "dashed", color = "blue") +
  geom_text(aes(x= mean(Eu$Rating), y=.04), label= "Europe",color = "blue", angle=90, vjust=-0.5, hjust=


  geom_density(data=As,aes(x=Rating), color="orange") +
  geom_vline( xintercept = mean(As$Rating), linetype = "dashed", color = "orange") +
  geom_text(aes(x= mean(As$Rating), y=.04), label= "Asia",color = "orange", angle=90, vjust=-0.5, hjust=


  geom_density(data=Na,aes(x=Rating), color="darkgreen") +
  geom_vline( xintercept = mean(Na$Rating), linetype = "dashed", color = "darkgreen") +
  geom_text(aes(x= mean(Na$Rating), y=.04), label= "North America", color = "darkgreen",angle=90, vjust=
```

```
  geom_density(data=Sa,aes(x=Rating), color="brown") +
  geom_vline( xintercept = mean(Sa$Rating), linetype = "dashed", color = "brown") +
  geom_text(aes(x= mean(Sa$Rating), y=.04), label= "South America", color = "brown",angle=90, vjust=-0.5

  geom_density(data=Oc,aes(x=Rating), color="purple") +
  geom_vline( xintercept = mean(Oc$Rating), linetype = "dashed", color = "purple") +
  geom_text(aes(x= mean(Oc$Rating), y=.04), label= "Oceania", color = "purple",angle=90, vjust=-0.5, hju

  geom_density(data=Af,aes(x=Rating), color="gray") +
  geom_vline( xintercept = mean(Af$Rating), linetype = "dashed", color = "gray") +
  geom_text(aes(x= mean(Af$Rating), y=.04), label= "Africa", color = "gray",angle=90, vjust=-0.5, hjust=

  labs(x="Rating", y=" Probability of Rating ") +
  theme_minimal()+
  scale_x_continuous(breaks = NULL)
```

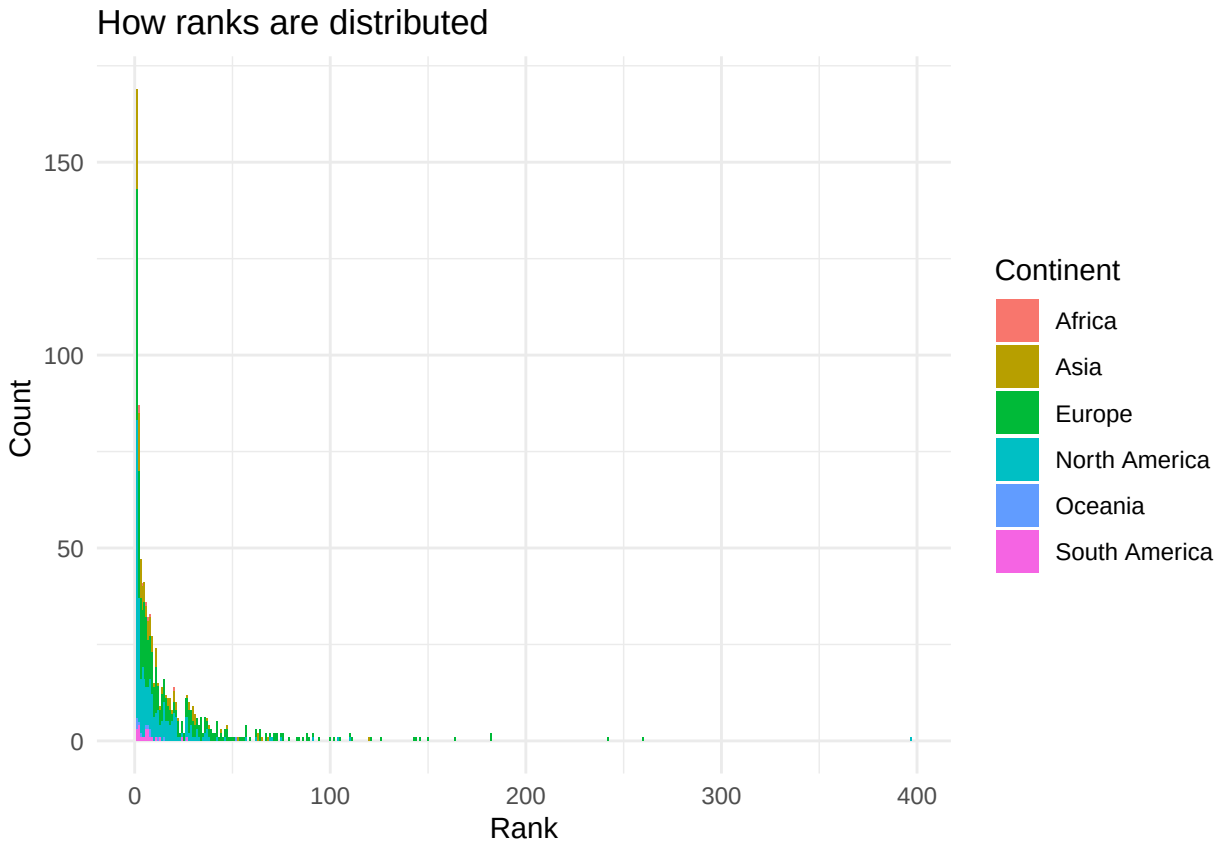


However, rating number is very different compare to other number data set, so the distrabuion does not really show the probability of ratings.

Does rank of museums refelct the quality of museums?

```
library(ggplot2)

#museum rank stack chart by Continent
```

```
ggplot(museum, aes(x=Rank, fill=Continent)) +
  geom_bar(position="stack")+
  labs(x="Rank", y="Count ",title="How ranks are distributed") +
  theme_minimal()
```



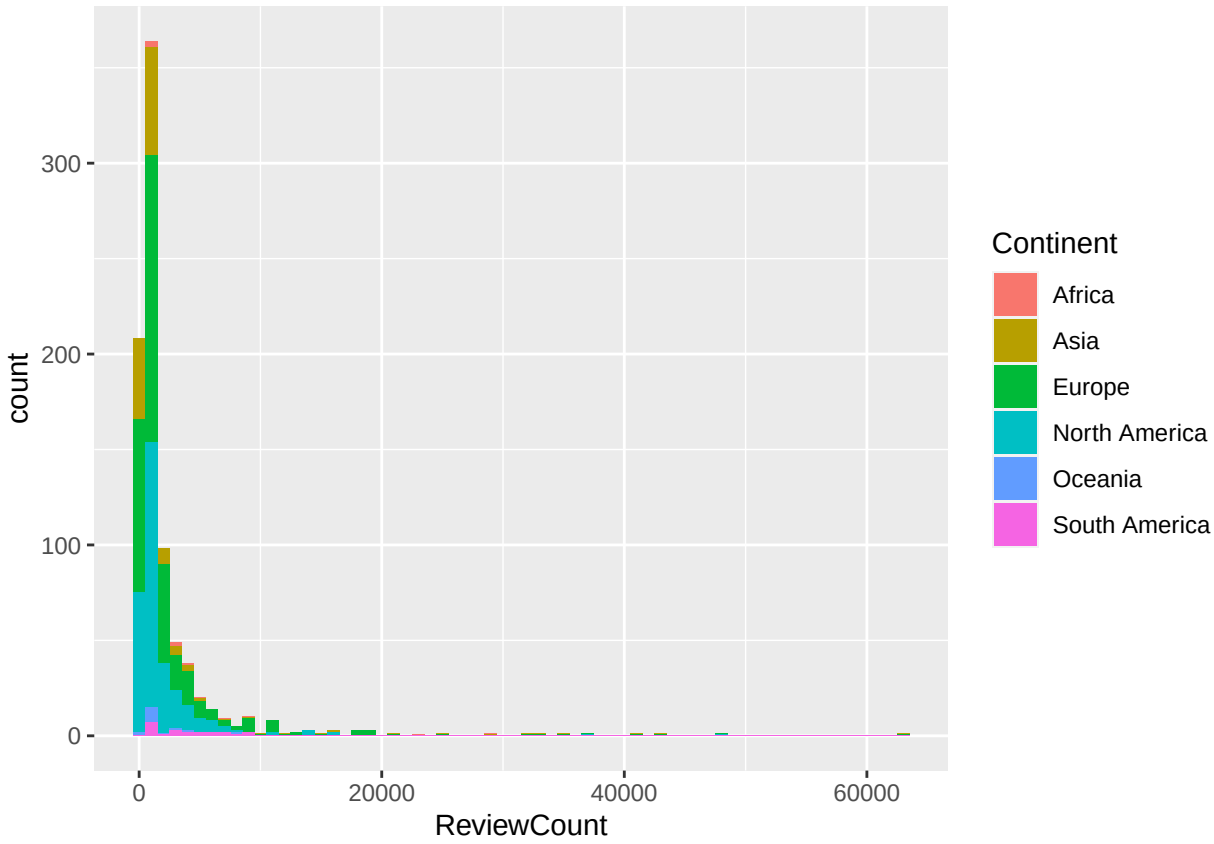
In fact, there can be several museums have same rank number.Morover, there are huge difference in museum numbers in each continent. In this case, the visualization can not represent the posibility of better museums. We shall start looking at the review count. The more reviews a museum has, more people have been there.

From this visualization we know that most museums are ranking in 100 range, which means they are ranked by their country. If a country has more museums then others, it is possible many good museums are not rank as the top.

Does museum review count reflect the quality of the museums?

Let's limited our dataset to the top 50 ranked museums.

```
museum100 <- filter(museum,Rank <= 100)
ggplot( museum100, aes(x=ReviewCount, fill=Continent)) +
    geom_histogram( binwidth=1000)
```



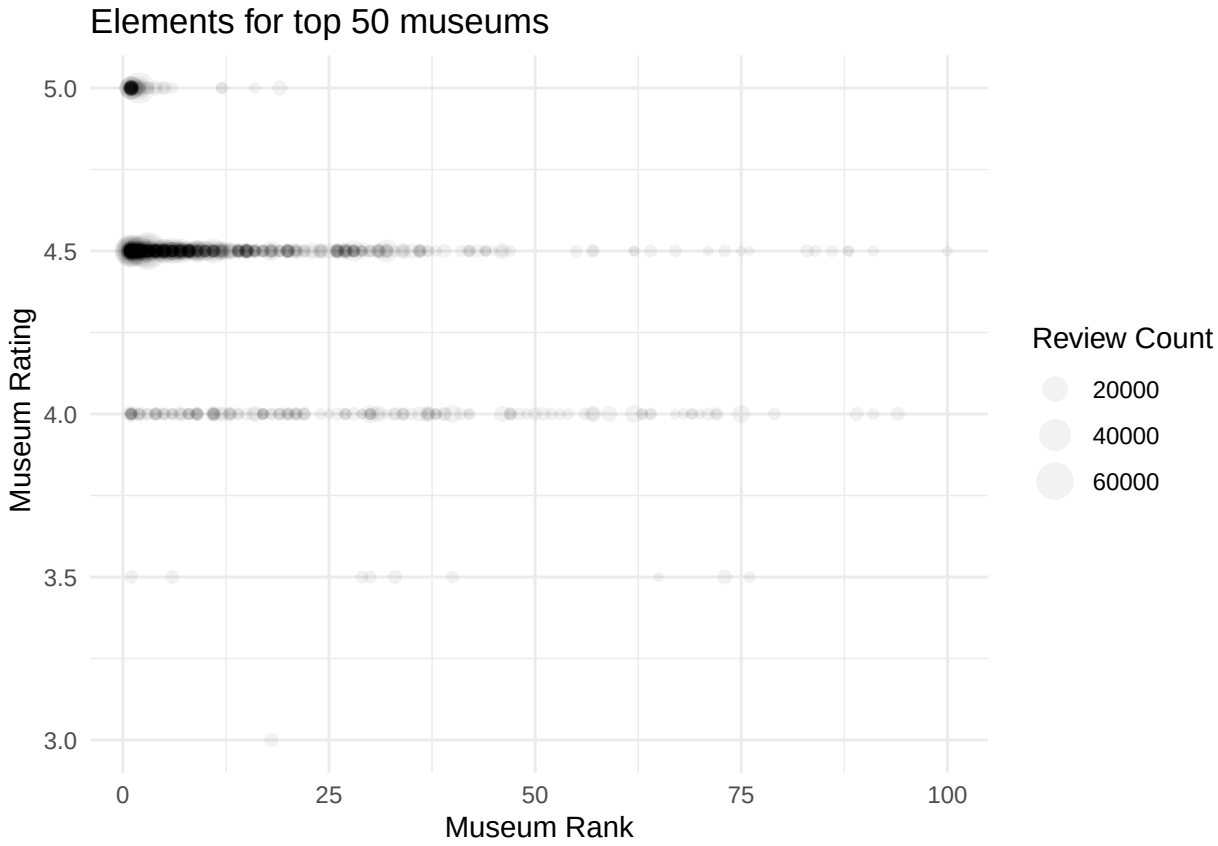
Clearly most museums have under 20,000 reviews.

The combination

Since there are no 1 elements can directly show the quality of the museums, I start to combin different elements in order to find a rang that represent museum from different perspective.

##Finding relationship between ranking and rate

```
museum100<- filter(museum, Rank <=100)
ggplot(museum100, aes(x=Rank, y=Rating,size=ReviewCount)) +
    geom_point(alpha=0.05) +labs( x = "Museum Rank", y=" Museum Rating", size = "Review Count",title =
  theme_minimal()
```



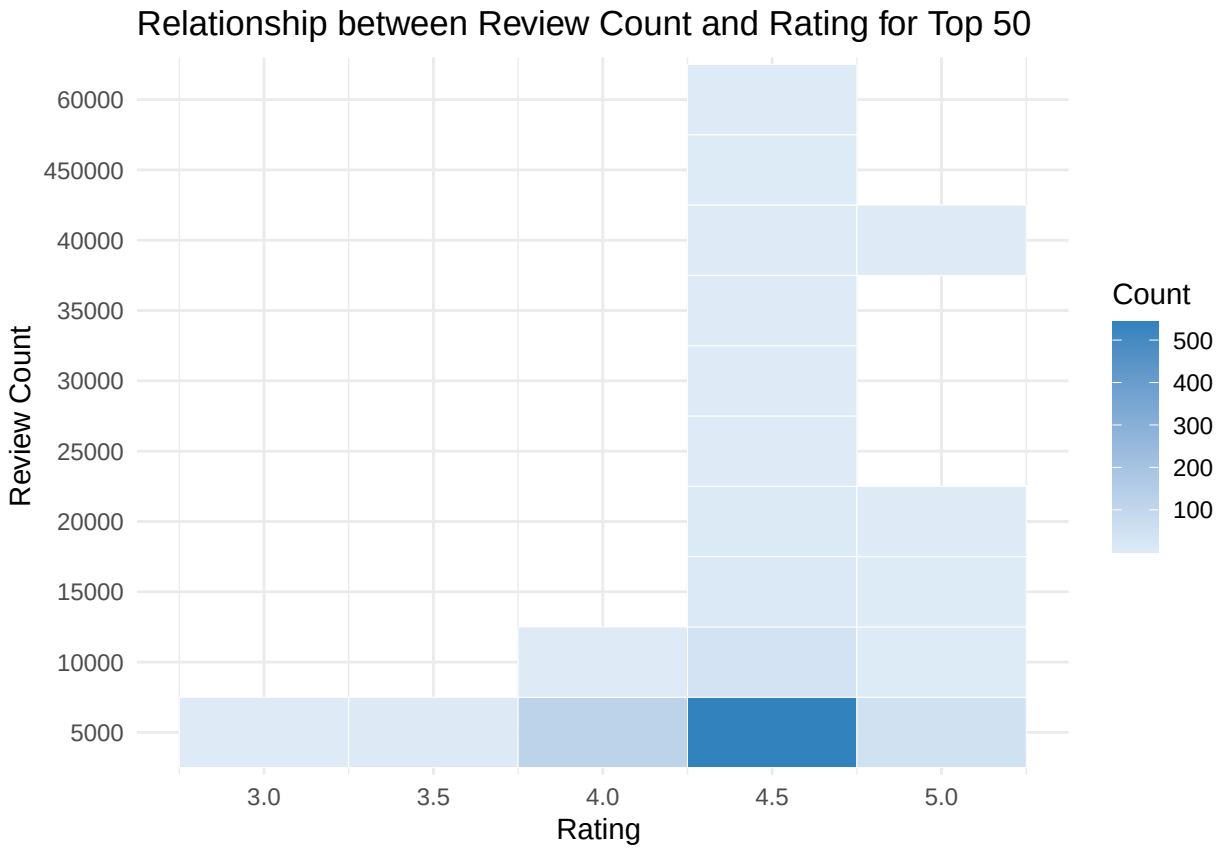
From this visualization, it seems museum with heigher rank, also has heigher rating and review count. In order to make a better choice in museum selection, we are looking for the darkest, and biggest area on this map :rank in top 50.

Finding the range

```
museum50 <- filter(museum100, Rank <=50)
counts <- museum50 %>% mutate(group = cut(ReviewCount, breaks = c(0,5000,10000,15000,20000,25000,30000,3
counts <- counts %>% group_by(group, Rating) %>% summarize(total = n())
```

```
## `summarise()` regrouping output by 'group' (override with `.groups` argument)
```

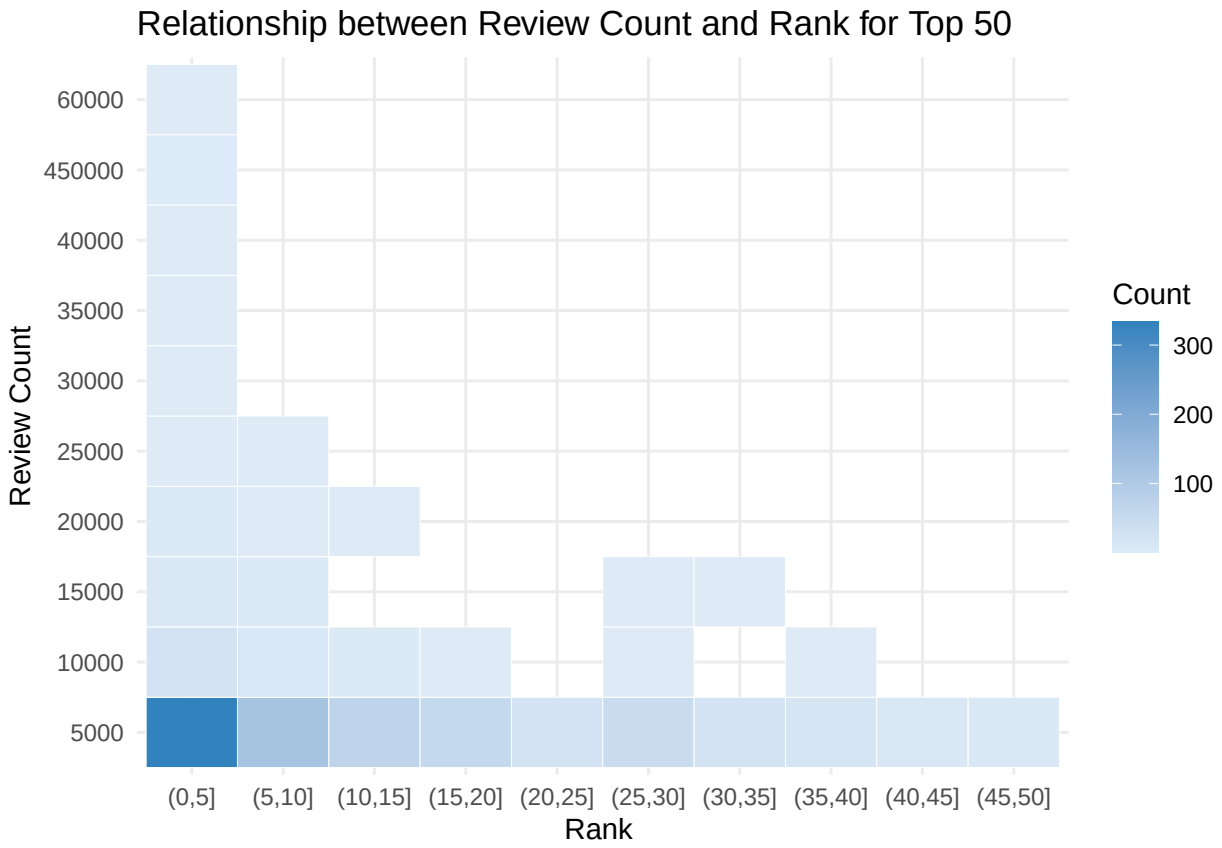
```
ggplot(counts) +
  geom_tile(mapping=aes(x=Rating, y= group, fill=total), color="#FFFFFF") +
  scale_fill_gradient(low="#deebf7", high="#3182bd") +
  theme(panel.grid.major = element_blank(), axis.text = element_text(size = 10)) +
  labs(y = "Review Count", x = "Rating ", fill="Count", title = "Relationship between Review Count and R
  theme_minimal()
```




```
counts2 <- museum50 %>%
  mutate(group2 = cut(Rank, breaks = c(0,5,10,15,20,25,30,35,40,45,50))) %>% mutate(group = cut(ReviewC
counts2 <- counts2 %>% group_by(group, group2) %>% summarize(total = n())
```

```
## `summarise()` regrouping output by 'group' (override with `.groups` argument)
```

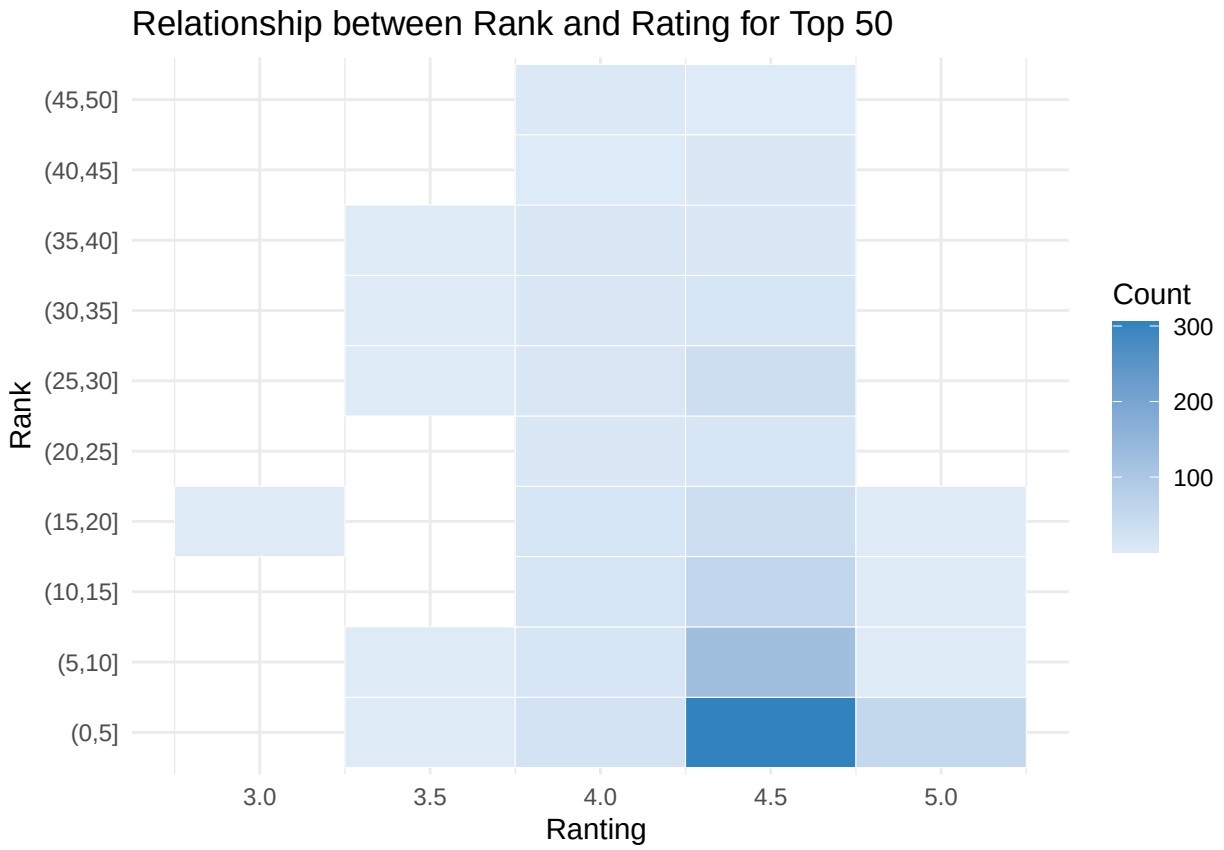
```
ggplot(counts2) +
  geom_tile(mapping=aes(x=group2, y= group, fill=total), color="#FFFFFF") +
  scale_fill_gradient(low="#deebf7", high="#3182bd") +
  theme(panel.grid.major = element_blank(), axis.text = element_text(size = 10)) +
  labs(y = "Review Count", x = "Rank ", fill="Count", title = "Relationship between Review Count and Ra
  theme_minimal()
```




```
counts3 <- museum50%>% mutate(group3 = cut(Rank, breaks = c(0,5,10,15,20,25,30,35,40,45,50)))
counts3 <- counts3 %>% group_by(Rating, group3) %>% summarize(total = n())
```

```
## `summarise()` regrouping output by 'Rating' (override with `.groups` argument)
```

```
ggplot(counts3) +
  geom_tile(mapping=aes(x=Rating, y= group3, fill=total), color="#FFFFFF") +
  scale_fill_gradient(low="#deebf7", high="#3182bd") +
  theme(panel.grid.major = element_blank(), axis.text = element_text(size = 10)) +
  labs(y = "Rank", x = "Ranting ", fill="Count", title = "Relationship between Rank and Rating for Top 5
  theme_minimal()
```




```
museum100
```

```
## # A tibble: 852 x 27
##       Me Address Count Continent sov_a3 Country Mcount Description FeatureCount
##    <dbl> <chr>   <dbl> <chr>     <chr>  <chr>    <dbl> <chr>               <dbl>
##  1    NA 180 Gr~   311 North Am~ USA    Unite ~    283 "The Natio~            5
##  2     2 Great ~   397 Europe    GBR    England     79 "A museum ~            6
##  3    NA Cromwe~   397 Europe    GBR    England     79 "A center ~            5
##  4    NA Cromwe~   397 Europe    GBR    England     79 "The world~            9
##  5    NA Clive ~   397 Europe    GBR    England     79 "The under~            6
##  6    NA Leeman~   397 Europe    GBR    England     79 "For a fan~            9
##  7    NA Chambe~   397 Europe    GBR    England     79 "Explore t~           14
##  8    NA Kelvin~   397 Europe    GBR    England     79 "Kelvingro~            5
##  9    NA Exhibi~   397 Europe    GBR    England     79 "This muse~            4
## 10    NA Liverp~   397 Europe    GBR    England     79 "The Museu~            3
## # ... with 842 more rows, and 18 more variables: Fee <chr>, longtitude <dbl>,
## #   latitude <dbl>, LengthOfVisit <chr>, MuseumName <chr>, PhoneNum <chr>,
## #   Rank <dbl>, Rating <dbl>, ReviewCount <dbl>, TotalThingsToDo <dbl>,
## #   MuseumLoctation <chr>, MuseumTopic <chr>, Color <chr>, X23 <lgl>,
## #   X24 <lgl>, X25 <lgl>, X26 <lgl>, ColorCode <lgl>
```

The hard part is, we do not know what really happened for the review under 5000.

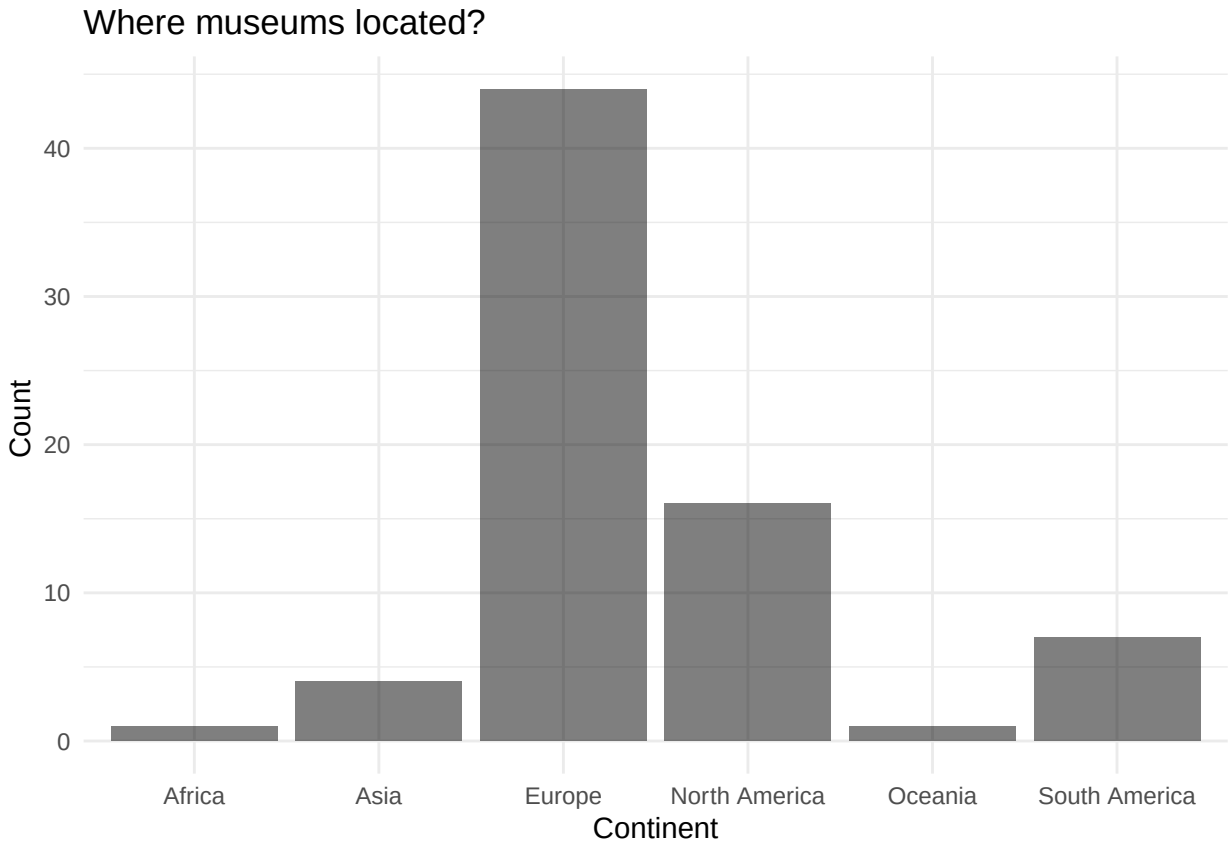
We could say, museums with over 5000 reviews, rating between 4 - 5 are generally recomannned by many people. Therefore, we could say: the heigher museum ranking and review counts are, the better museums might be, rating 4.5 is highly possible the best range.

```
museumList<- filter(museum50, Rating == 4.5)
museumList<- filter(museumList,ReviewCount >= 5000)
```

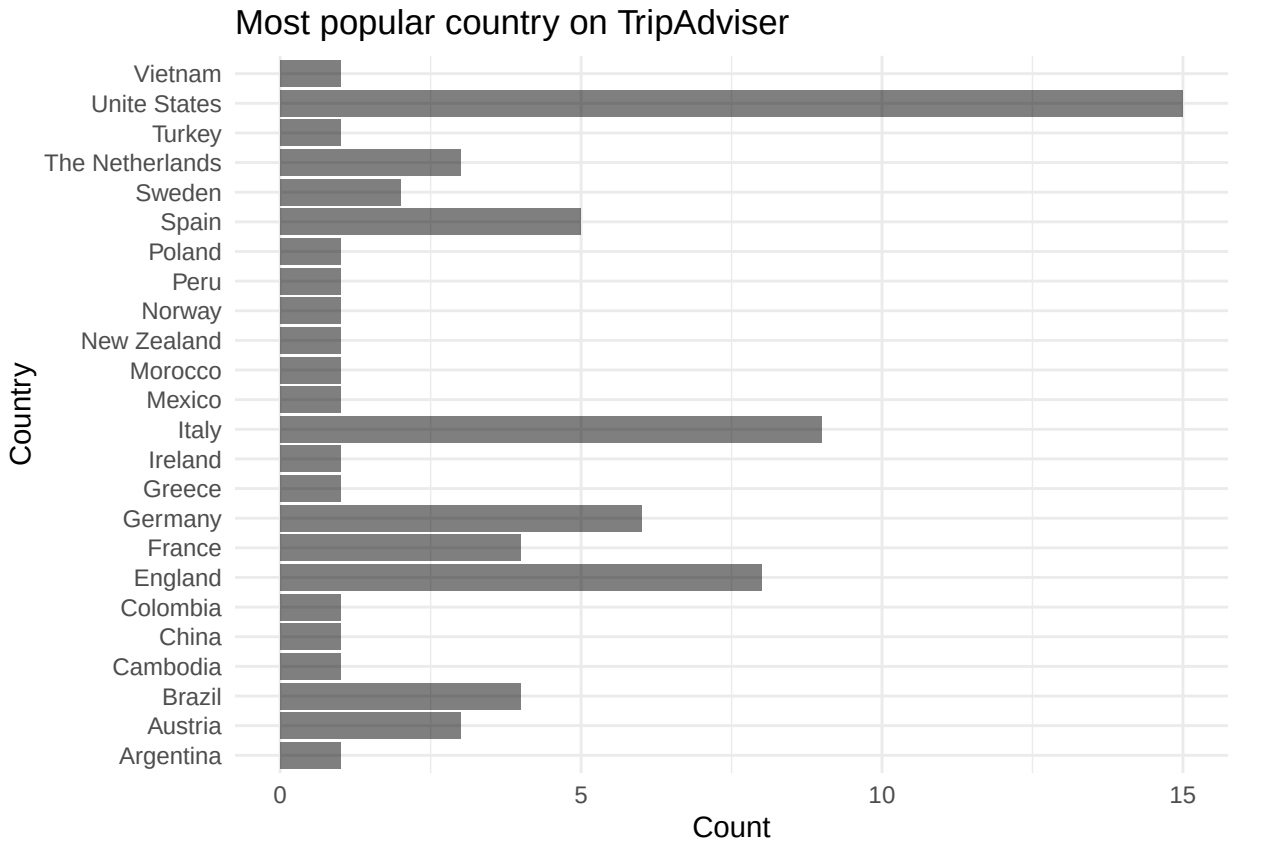
```
museumList
```

```
## # A tibble: 73 x 27
##       Me Address Count Continent sov_a3 Country Mcount Description FeatureCount
##    <dbl> <chr>   <dbl> <chr>     <chr>  <chr>    <dbl> <chr>               <dbl>
##  1    NA 180 Gr~   311 North Am~ USA    Unite ~    283 "The Natio~            5
##  2     2 Great ~   397 Europe    GBR    England     79 "A museum ~            6
##  3    NA Cromwe~   397 Europe    GBR    England     79 "A center ~            5
##  4    NA Cromwe~   397 Europe    GBR    England     79 "The world~            9
##  5    NA Clive ~   397 Europe    GBR    England     79 "The under~            6
##  6    NA Leeman~   397 Europe    GBR    England     79 "For a fan~            9
##  7    NA Chambe~   397 Europe    GBR    England     79 "Explore t~           14
##  8    NA Kelvin~   397 Europe    GBR    England     79 "Kelvingro~            5
##  9    NA Liverp~   397 Europe    GBR    England     79 "The Museu~            3
## 10     1 99 rue~   397 Europe    FRA    France      38 "Home to L~           13
## # ... with 63 more rows, and 18 more variables: Fee <chr>, longtitude <dbl>,
## #   latitude <dbl>, LengthOfVisit <chr>, MuseumName <chr>, PhoneNum <chr>,
## #   Rank <dbl>, Rating <dbl>, ReviewCount <dbl>, TotalThingsToDo <dbl>,
## #   MuseumLoctation <chr>, MuseumTopic <chr>, Color <chr>, X23 <lgl>,
## #   X24 <lgl>, X25 <lgl>, X26 <lgl>, ColorCode <lgl>
```

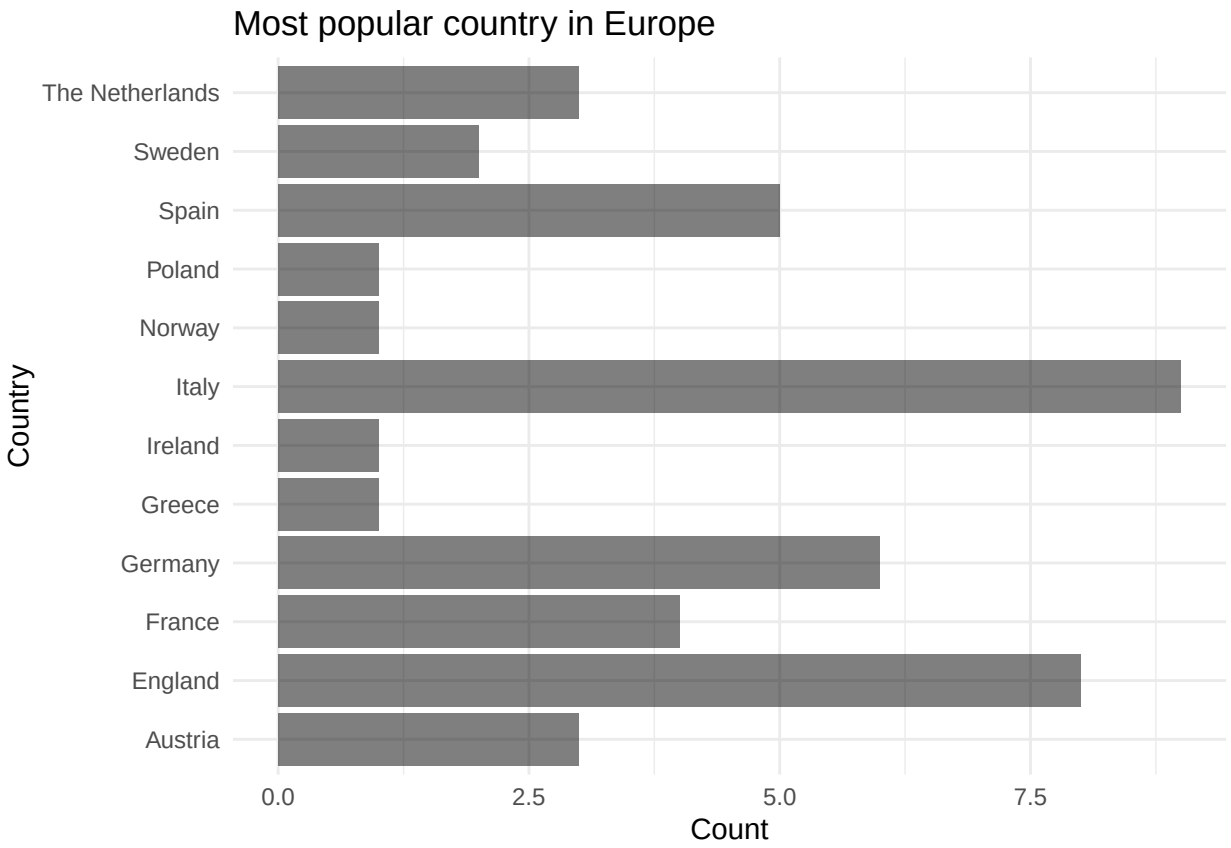
```
ggplot(museumList, aes(x=Continent)) +
  geom_bar(fill="black", alpha = 0.5 )+
  labs(x="Continent", y="Count ",title="Where museums located?") +
  theme_minimal()
```



```
ggplot(museumList, aes(y=Country)) +
  geom_bar(position="stack", fill="black", alpha = 0.5)+
  labs(x="Count", y="Country",title="Most popular country on TripAdviser") +
  theme_minimal()
```




```
Eu <- museumList  %>%
  filter(grepl ("Europe", Continent))

ggplot(Eu, aes(y=Country)) +
  geom_bar(position="stack", fill="black", alpha = 0.5)+
  labs(x="Count", y="Country",title="Most popular country in Europe") +
  theme_minimal()
```



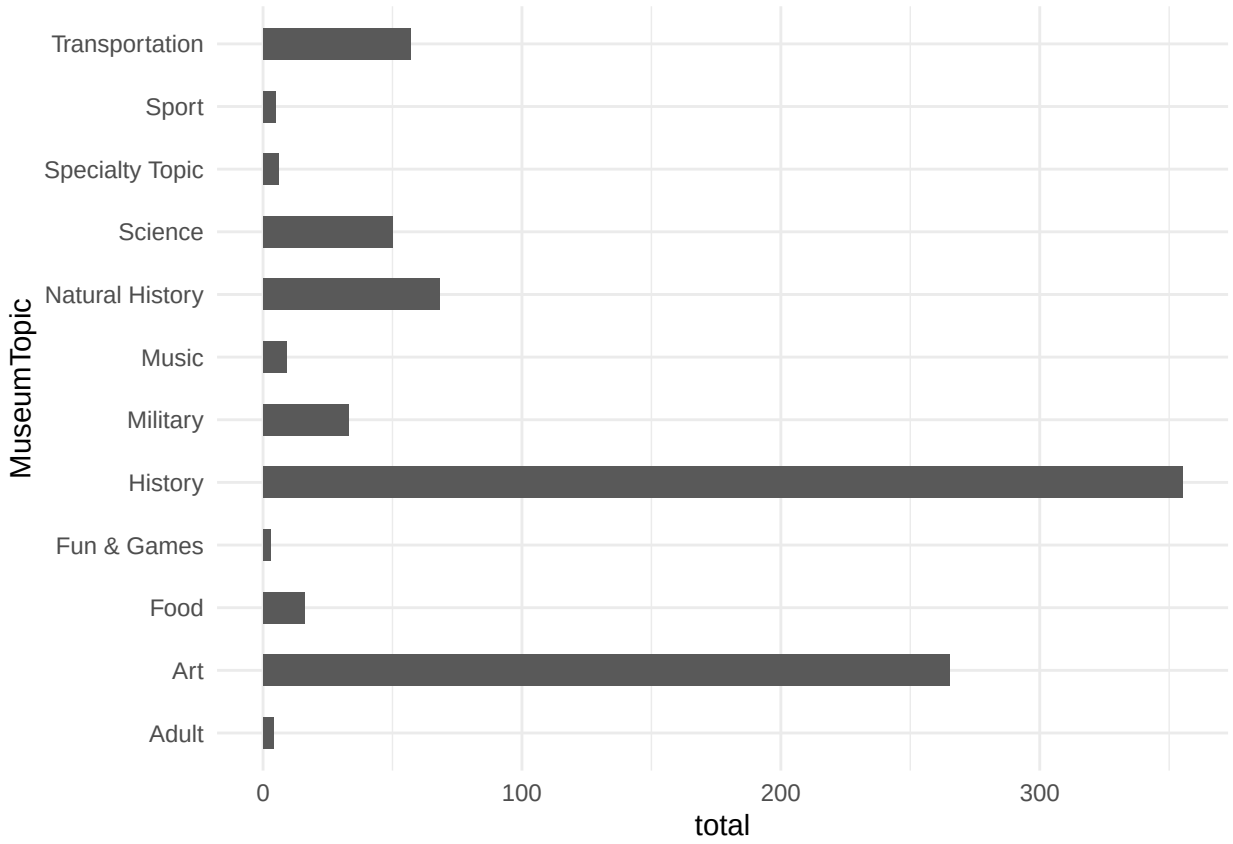
The advantages of Europe have become more obvious, and Italy has surpassed Britain to become the most popular country. The number of museums in Asian countries has decreased. This proves that most tourists choose Europe, North America and South America as destinations.

Other Visualization

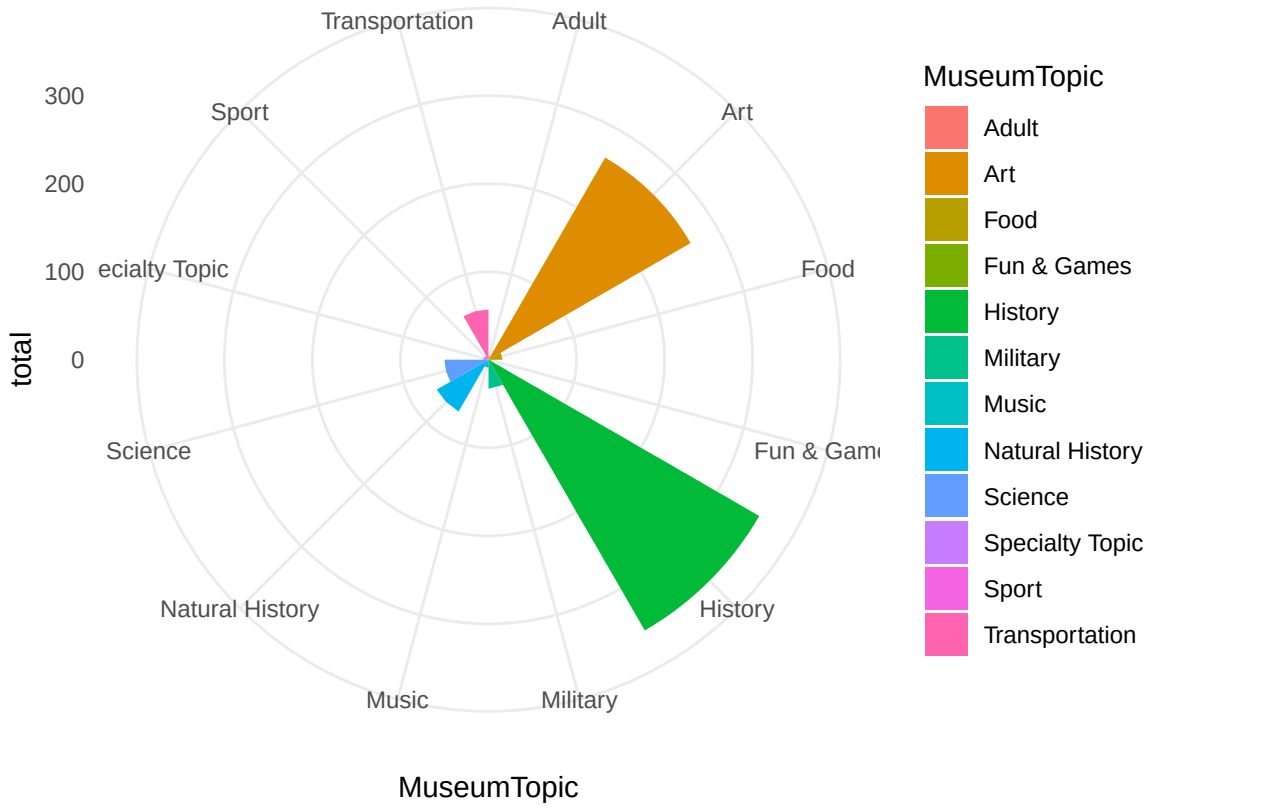
```
# Create a data frame for categories
# at each 'cut' value


MuseumTopic <- museum %>% group_by(MuseumTopic) %>% summarize(total = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

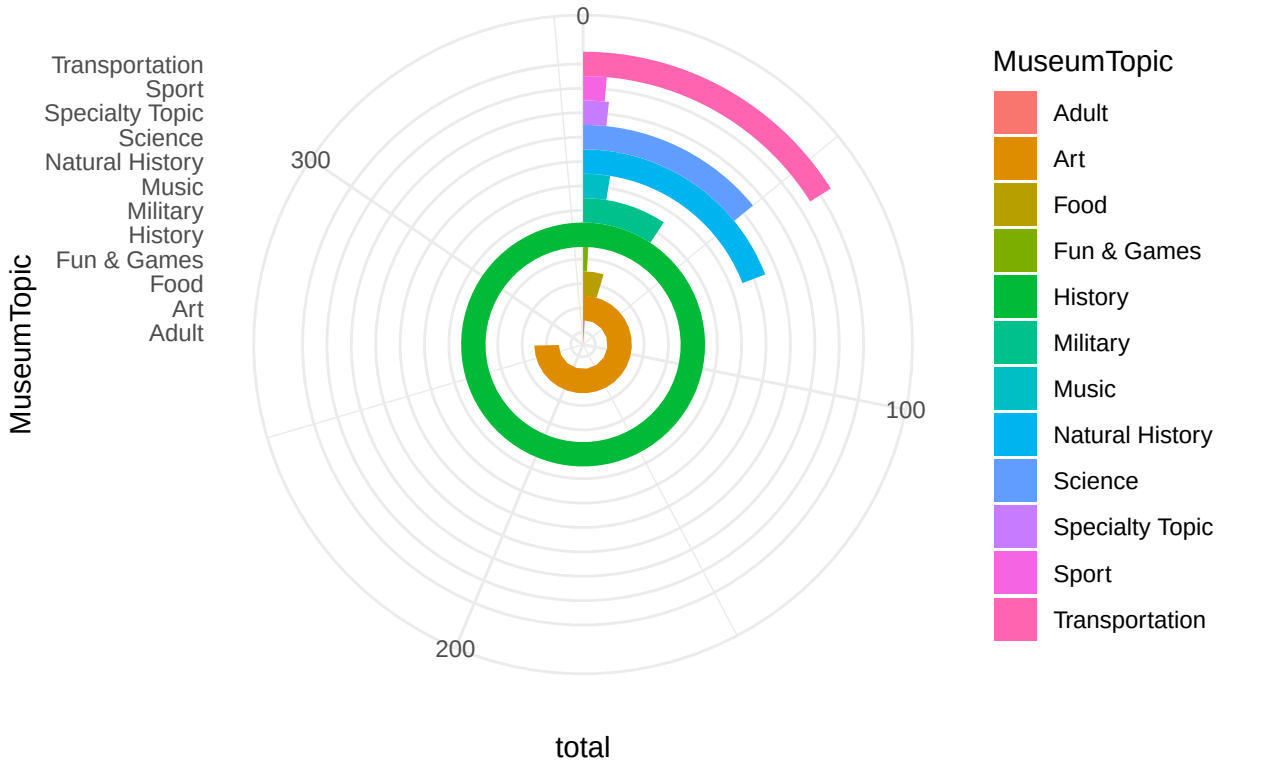
```
ggplot(MuseumTopic) +
  geom_bar(aes(x=total, y=MuseumTopic), stat="identity", width=0.5)+
  theme_minimal()
```




```
ggplot(MuseumTopic) +
  geom_bar(aes(x =total , y=MuseumTopic, fill = MuseumTopic), stat="identity", width=1) +
  coord_polar("y")+
  theme_minimal()
```

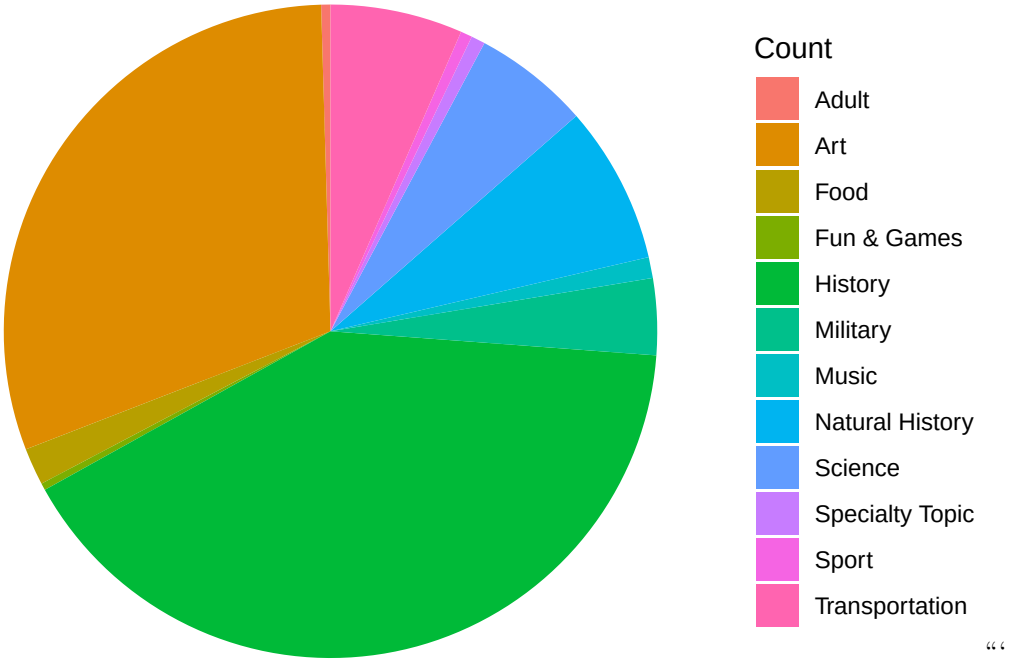



```
ggplot(MuseumTopic) +
  geom_bar(aes(x = MuseumTopic, y=total, fill = MuseumTopic), stat="identity", width=1) +
  coord_polar("y")+
  theme_minimal()
```



```
ggplot(MuseumTopic) +
  geom_bar(aes(x = "", y=total, fill = MuseumTopic), stat="identity", width=1) +
  coord_polar("y") +
  theme_void() +
  labs(fill="Count", title = "Most popular museum topics")
```

Most popular museum topics



“‘