Homework 1 of Data Visualization

Chutian Zhou

2/16/2018

0. Housekeeping Stuff

First of all, I read two csv files ("dictionary.csv" and "winter.csv") and then combine them together. I have also done some data cleaning work. The codes are as such:

```
#Load packages
library(ggplot2)
library(dplyr)

#Read tables
w<-read.csv("winter.csv")
d<-read.csv("dictionary.csv")

#Merge datraframes
final<-merge(w,d,by.x="Country",by.y="Code")
final<-final[,-1]
final<-final%>%arrange(Year)
final<-final%>%group_by(Sport,Discipline)
```

All the work in this assignment are based on this `final` dataframe (of course, there will be some meticulous modifications with regard to this dataframe when doing each task), which is the ultimate output of the codes above.

1. Medals Counts over Time

In the first task, I only put emphasis on the top 10 countries that have participated in the Winter Olympic Games most frequently. To get my subjects, the following codes are run:

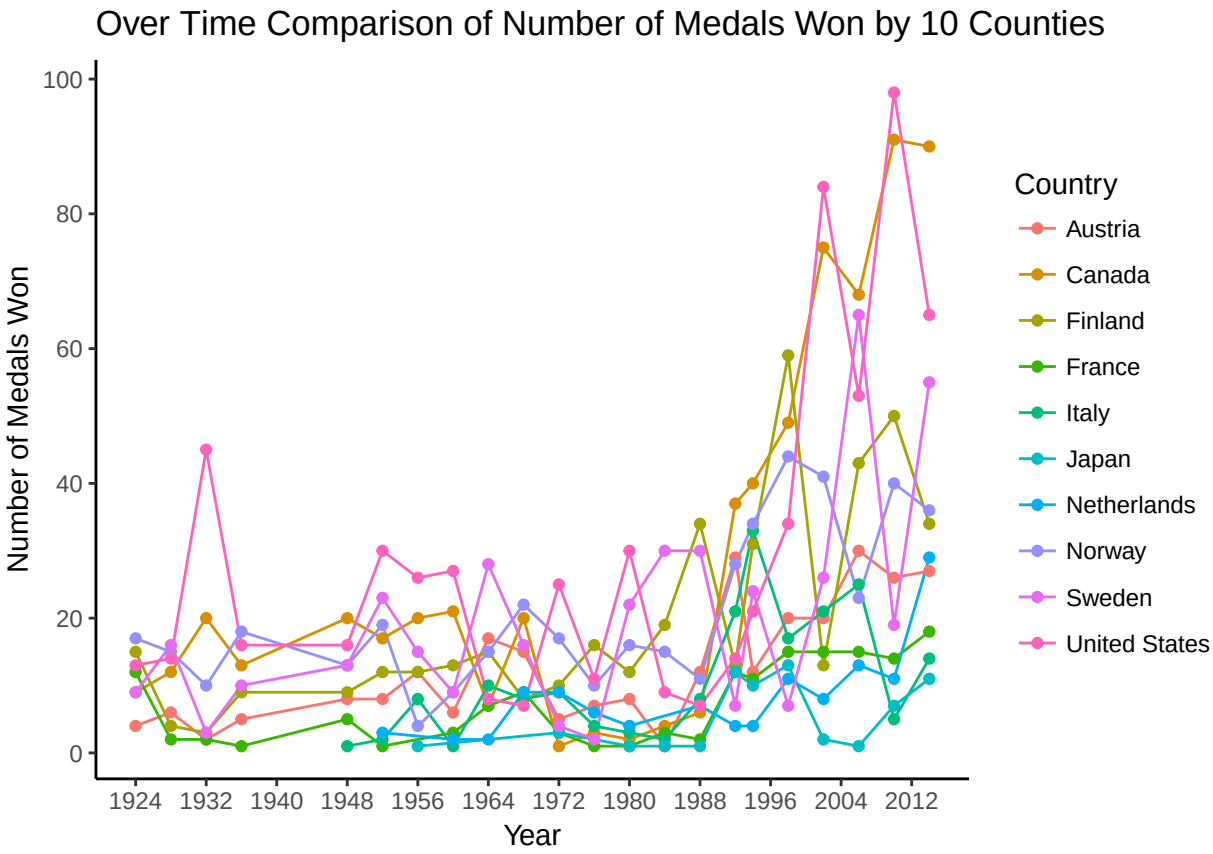
```
year_country<-unique(final[c("Year","Country")])
year_country<-year_country%>%group_by(Country)
```

The top 10 countries that have participated in the Games most frequently are Austria (22 times), Canada (22 times), Finland (22 times), Norway (22 times), Sweden (22 times), United States (22 times), France (21 times), Italy (18 times), Netherlands (15 times), and Japan (12 times).

To draw the plot of over time comparison, a new data frame called `TotalMedals` is generted (see codes below). Note that the `data` call in the `ggplot` function takes this `TotalMedals` data frame.

```
#Top 10's medals
top10<-final%>%filter(Country%in%c("Austria","Canada",
                                   "Finland","Norway","Sweden","United States",
                                   "France","Italy","Netherlands","Japan"))
TotalMedals<-top10%>%group_by(Year,Country)%>%summarise(count=n())
```

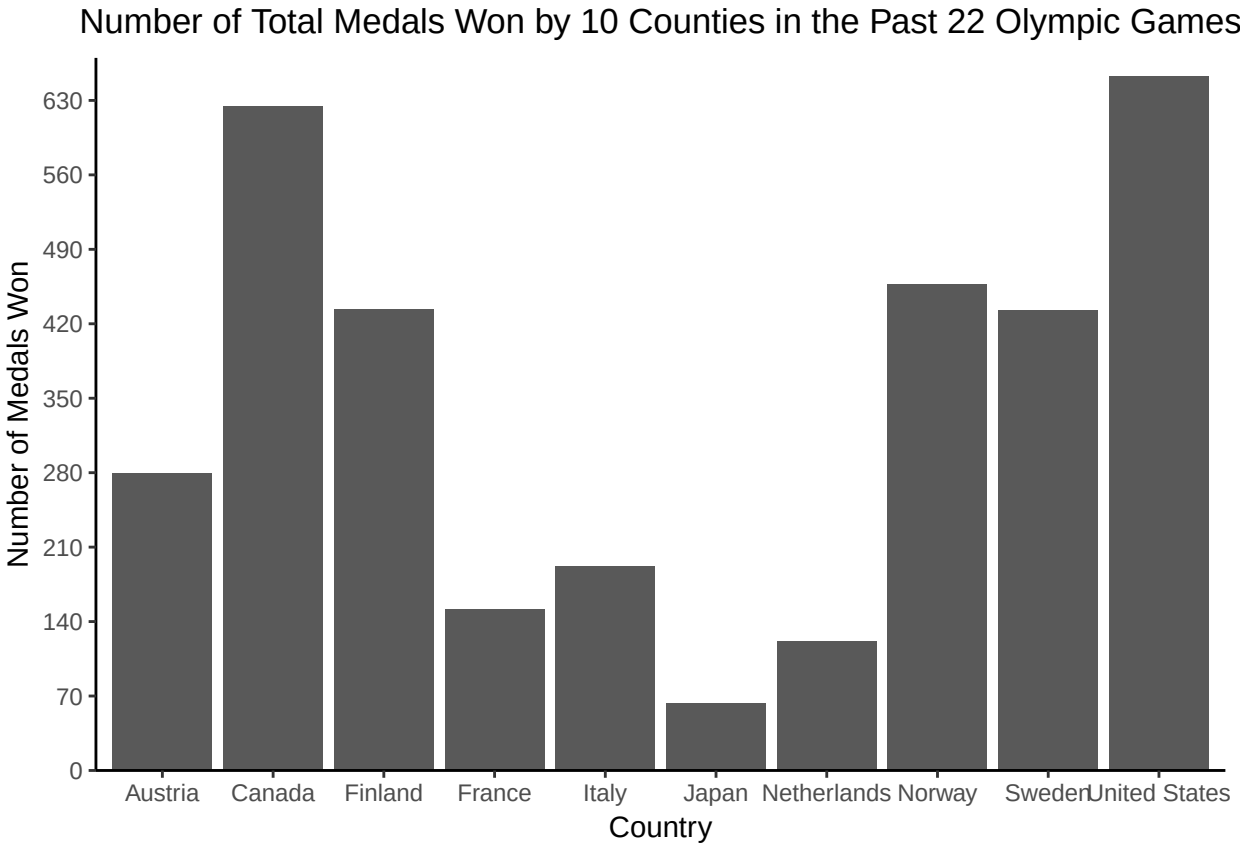
The over time comparison plot is shown below.



Next, turn to the total medal count. To draw this figure, a new data frame called `TotalMedals2` is generated (see codes below). Similarly as above, the `data` call in the `ggplot` function takes the `TotalMedals2` data frame.

```
TotalMedals2<-aggregate(count~Country,data=TotalMedals,sum)
```

The total medal count plot is presented below.



At first glance, I would recommend the figure of total medal count to the editor, given that it is not as complex as the over time comparison plot, which contains too many lines. Nevertheless, since the `plotly` package can make the latter interactive and much easier to read (by moving mouse to a certain point, the reader can see the detailed data), I would suggest using the over time comparison plot, but only after the `plotly` package is introduced.

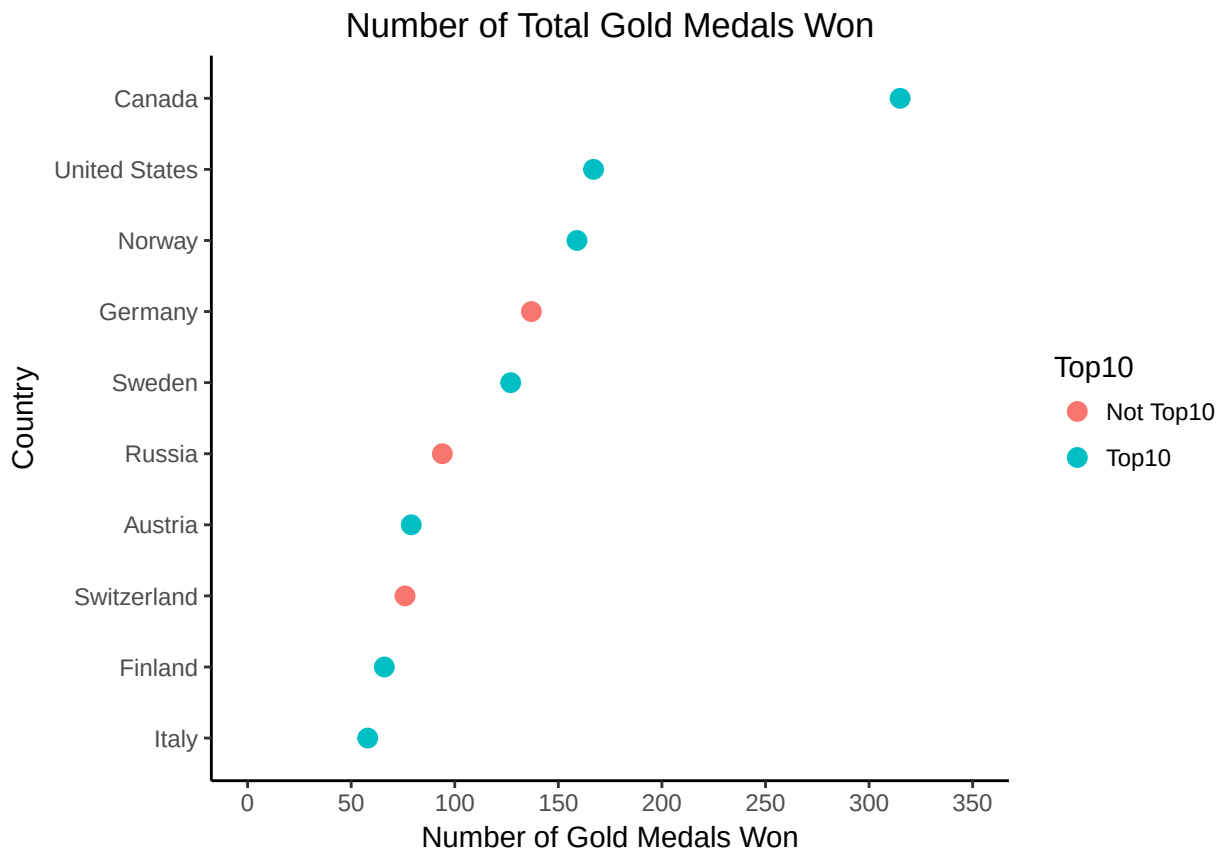
2. Medal Counts Adjusted by Population, GDP

To gauge a country's success, I simply consider the number of gold medals won.

The plot is built upon the data frame called `goldonlycount`. See the chunk below for the coding of `goldonlycount`.

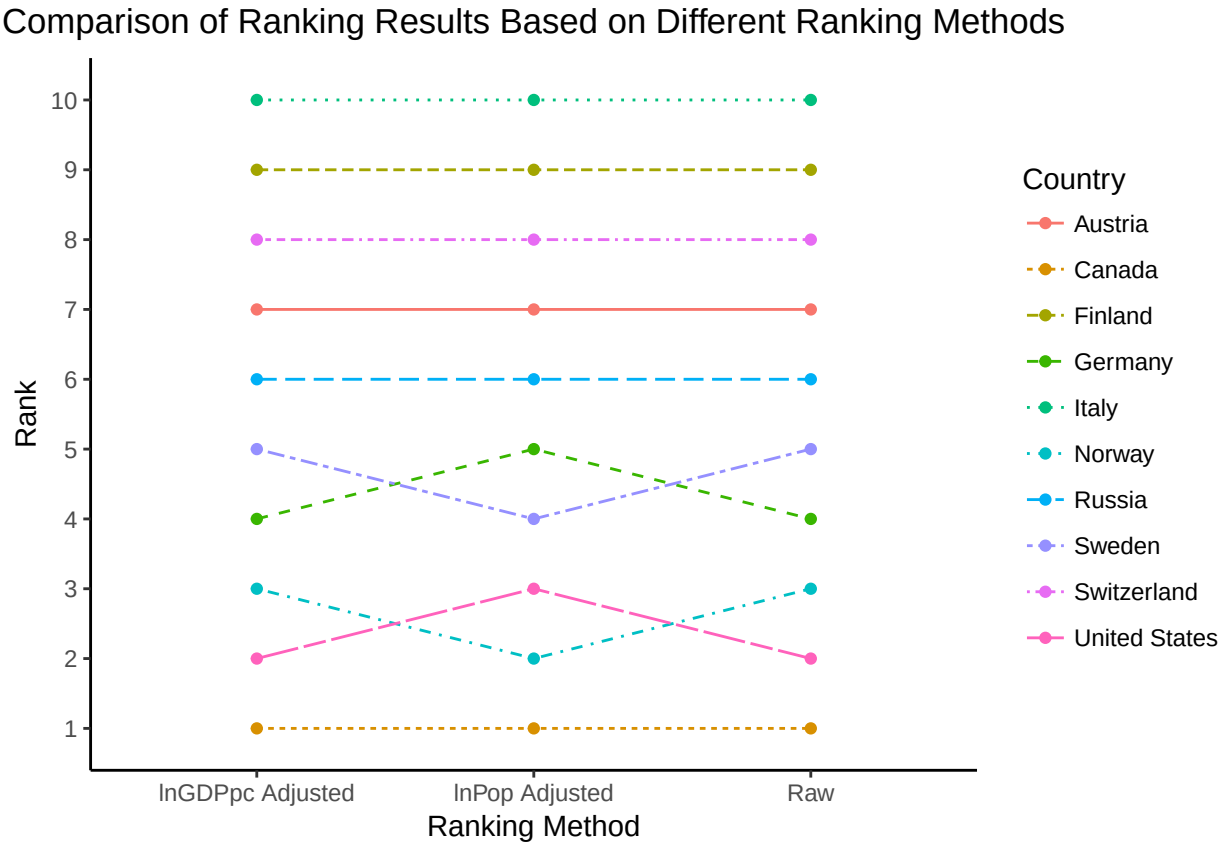
```
goldonly<-final%>%filter(Medal=="Gold")%>%group_by(Year,Country)%>%summarise(count=n())
goldonlycount<-aggregate(count~Country,data=goldonly,sum)
goldonlycount<-goldonlycount%>%arrange(desc(count))
goldonlycount<-goldonlycount[c(1:10),]

#Generate new variable (top 10 dummy)
goldonlycount=within(goldonlycount,{
  Top10=ifelse(Country%in%c("Austria","Canada","Finland",
                            "Norway","Sweden","United States",
                            "France","Italy","Netherlands","Japan"),
               "Top10", "Not Top10")})
```

The "raw" ranking of the countries' success is plotted below.



Note that "Not Top 10" means the country is not one of the top 10 countries that have competed in the Winter Olympic Games most frequently (for the list of top 10 countries, please refer to the paragraph under the first chunk in the first task). Intuitively, "Top 10" means the country is one of them. We can see that Canada outshines all its counterparts significantly in terms of the number of total gold medals won throughout the history.

Furthermore, I adjusted the raw ranking through dividing `count` by natural logged *GDP per capita* and *population* so that the number would not be extremely large.



Basically, the results did not change too much. Fluctuations in rankings can only be observed between the United States and Norway as well as Germany and Sweden. The reason is the variation in two newly introduced variables (*GDP per capita* and *population*) is not too large, especially *GDP per capita*. This explains the interesting phenomenon that the **lnGDPpc adjusted ranking** and the **raw ranking** are exactly the same.

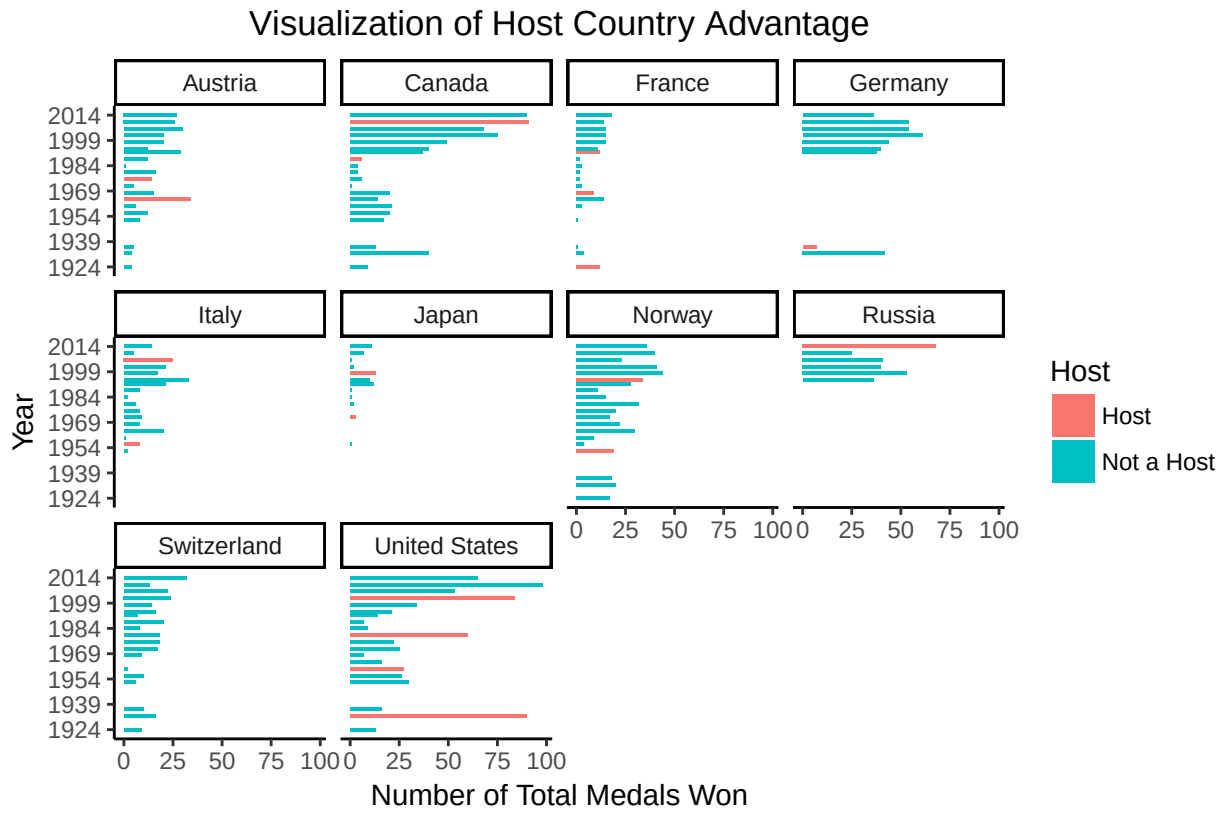
3. Host Country Advantage

The data frame preparation process for this task is much more tedious. The chunk below tracks how it goes. I merge the `hosts` data frame provided by Professor with the original `final` data frame. I have also generated a new dummy variable called *Host* to indicate whether the "host country advantage" occurs in such year. This is a critical step for later coloring the bar as a way to indicate if the country is the host in that year.

```
#These codes are provided by Professor
library(rvest)
library(stringr)
wiki_hosts<-read_html("https://en.wikipedia.org/wiki/Winter_Olympic_Games")
hosts<-html_table(html_nodes(wiki_hosts,"table")[[5]],fill=TRUE)
hosts<-hosts[-1,1:3]
hosts$city<-str_split_fixed(hosts$Host,n=2,",")[,1]
hosts$country<-str_split_fixed(hosts$Host,n=2,",")[,2]

#Cleaning hosts dataframe
hosts$city[hosts$city=="Garmisch-Partenkirchen"]<-"Garmisch Partenkirchen"
colnames(hosts)[5]<-"hostcountry"
hosts<-hosts[-c(5:6,26:27),-c(1:3)]
```

```
trim<-function(x)gsub("^\\s+|\\s+$","",x)
#hostcountry column has whitespaces. Have to trim them
hosts$hostcountry<-trim(hosts$hostcountry)

#Merge with original dataframe
final2<-merge(final,hosts,by.x="City",by.y="city")
final2<-final2%>%arrange(Year)

#Create variable "Host"
final2$Country<-as.character(final2$Country)
#Country variable in final2 is a factor instead of character
final2$Host=ifelse(final2$Country==final2$hostcountry,"Host","Not a Host")

#Create a new data frame
hostonly<-final2%>%filter(Country%in%c("France","Switzerland",
                                       "United States","Germany","Norway",
                                      "Italy","Austria","Japan",
                                      "Yugoslavia","Canada","Russia"))
hostonly2<-hostonly%>%group_by(Year,Country,Host)%>%summarise(count=n())
```

The visualization of host country advantage is put below.



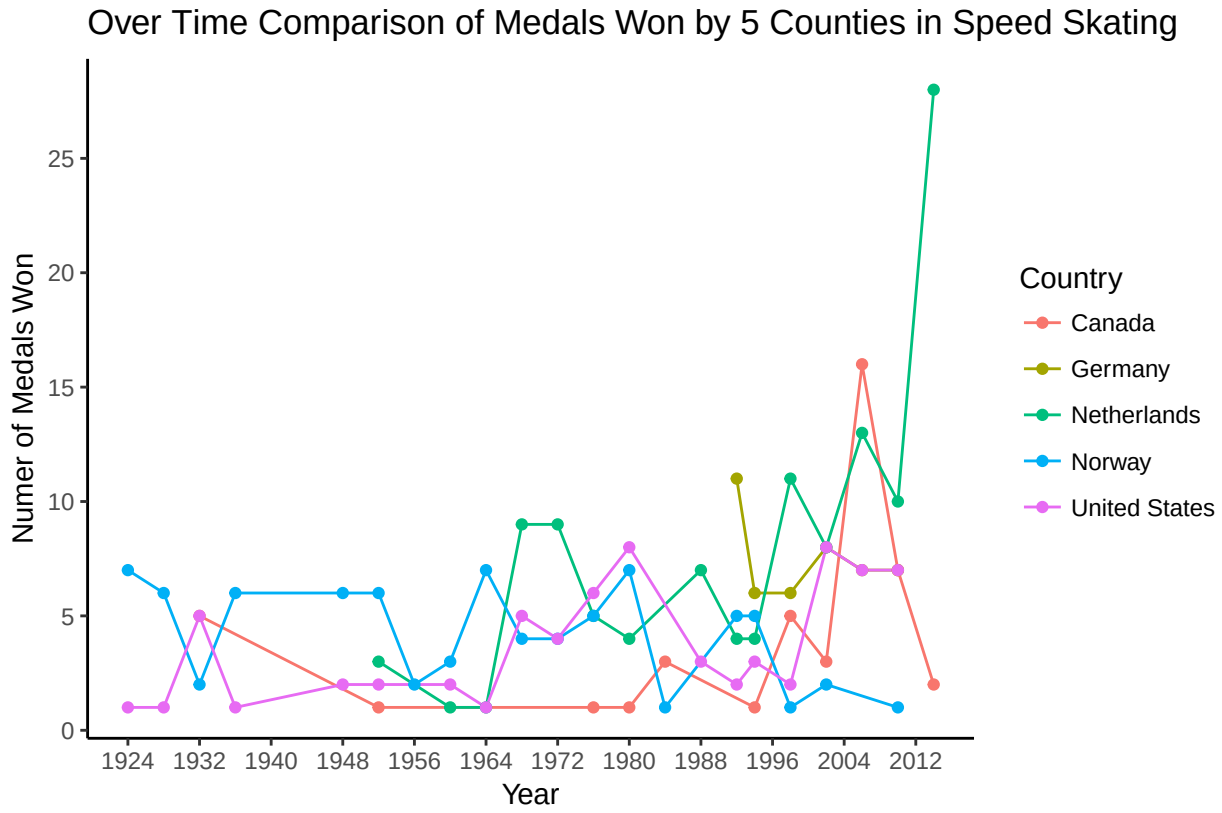
: These are countries that have hosted Winter Olympic Games throughout the history.

Keep in mind that the red color denotes the country is the host country in that specific year. Generally, we cannot sufficiently conclude the host country advantage prevails. Russia obviously has the advantage in the 2014 game, while the United States also enjoys it in 1932 and 1980. Strikingly, Switzerland does not have any advantages. In 1928 and 1948, it has not won any medals.

4. Country Success by Sport / Discipline / Event

In this task, I look at success in the discipline of speed skating at the country level. Five countries I picked used for comparison are Netherlands, Norway, United States, Canada and Germany. They are five most "successful" countries in that they have won most medals in this discipline throughout the history.

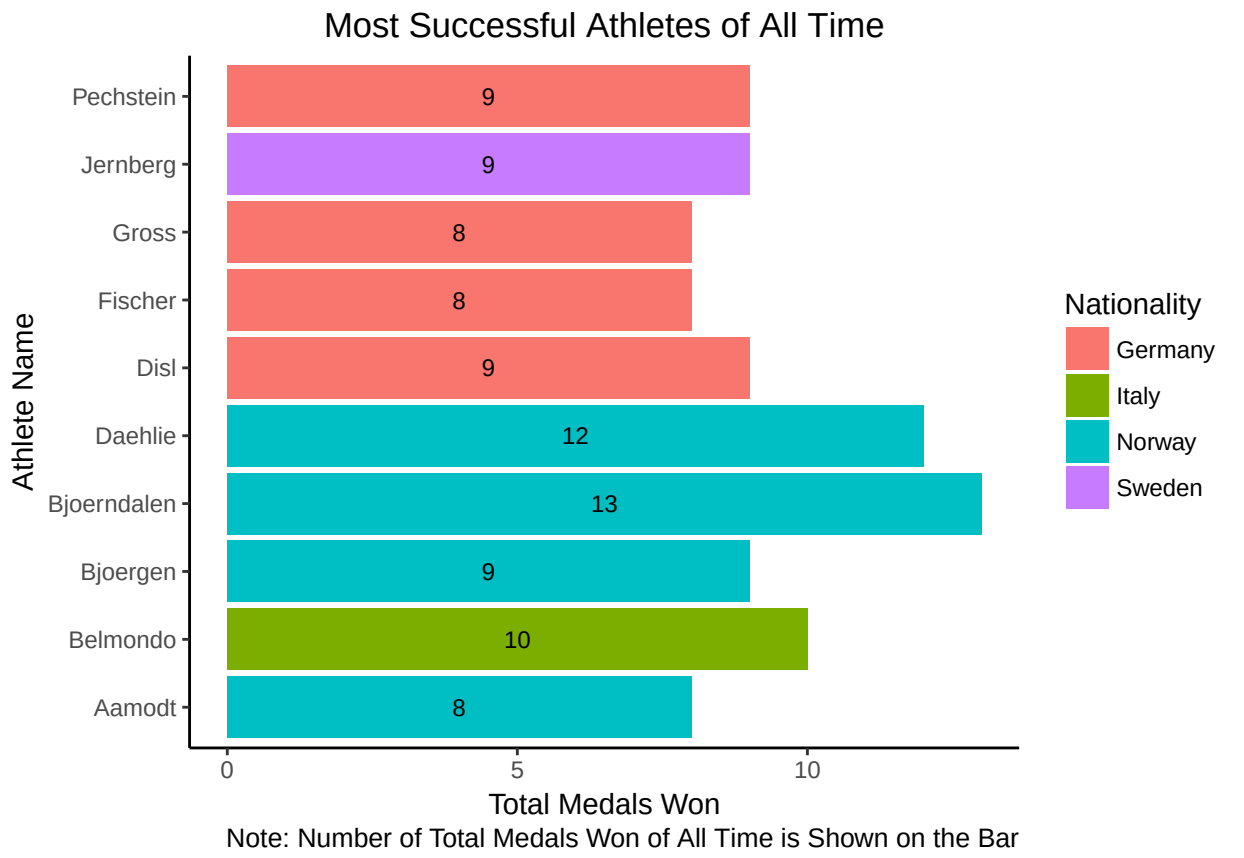
The line plot is put below. The most interesting pattern is undoubtedly the sudden jump of the number of medals won by Netherlands in 2014 Sochi Winter Olympic Games.



Note:These five countries are top five that have won most medals in speed skating.

5. Most Successful Athletes

In this task, I look at the name list of athletes who have won most medals across all 22 Games.



The player's nationalities are marked by different colors in the above bar plot. Generally, Germany and Norway are winners. Specifically, the most and the second successful athletes, Bjoerndalen and Daehile, are both Norwegians.

6. Make Two Plots Interactive

I applied interactivity to the figure of over time comparison of number of medals won by 10 countries (first figure in task 1) and the figure of over time comparison of medals won by 5 ountries in speed skating (figure in task 4). The reason is they contain too many lines, making graphs look complex. After the interactive widgets are added, the readabilty of two figures are strengthened.

7. Data Table

The data frame used for generating the data table is the one on speed skating, which also is used in task 4. Of course, I have revised a little bit.

In this data table, readers can filter for information they find meaningful or interesting with respect to the athlete's records of medal winning in speed skating.

```
library(DT)
speadskating<-ss[,-c(3:4,10:11)]
colnames(speadskating)[7]<-"Nationality"
dt<-speadskating%>%datatable(rownames = FALSE,
                          filter = list(position = "top"),
```

```
                          options = list(language = list(sSearch = "Filter:")))
dt
```