Exploring Data on the Olympics

After having to wait an extra year to watch the Summer Olympics, I got more excited than ever to watch the Olympics this year. I forgot how fascinating it was to see the athletes' performance until when I first watched the events again this year. It was also satisfying to see all these athletes from teams I support win all these medals. In fact, this is the first year I was actually interested in the awards as much as I was interested in the events. What sparked my interest even further was their showing the top countries in medal count this year. Now, with data on the Olympics medal available, I can explore more to see what other interesting things I can find.

1 Preparation

Before we get started, we need to import the packages we need:

```
library(tidyverse);
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

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

```
library(readr);
library(dplyr);
library(ggplot2)
```

Let's also define this function to make our work a little easier:

```
simplify_strings <- function(s){
  s <- str_to_lower(s);
  s <- str_trim(s);
  s <- str_replace_all(s,"[^a-z]+","_")
  s
}
```

This is to help make the names of the columns of the data more consistent so we can code more fluently. Note that to run this code, you need to have the relevant csv files saved to RStudio's working directory.

2 Importing and Processing Our Data

Of course, to work with our data, we first need to import it:

```
Athlete_Data <- read.csv("source_data/athlete_events.csv");
```

```
GDP_Data <- read.csv("source_data/gdp_csv.csv");

Pollution_Data <- read.csv("source_data/death-rates-from-air-pollution.csv");
```

Now, for the processing. Let's first simplify the names of our columns:

```
names(Athlete_Data) <- simplify_strings(names(Athlete_Data));
names(GDP_Data) <- simplify_strings(names(GDP_Data));
names(Pollution_Data) <- simplify_strings(names(Pollution_Data));
```

Now, since we are only interested in the awards in the Summer Olympics, let's filter Athlete_Data to show only the medalists in Summer. I renamed column noc to country_code so that it is consistent with the GDP_Data table.

```
Medal_Data<-Athlete_Data %>% filter(season=="Summer" & !is.na(medal)) %>% rename(country_code=noc)
```

Since it is more intuitive, let's rename the value column of GDP_Data to gdp. We will also rename the code column in Pollution_Data to country_code to make the column names consistent across different datasets.

```
GDP_Data <- GDP_Data %>% rename(gdp=value);
Pollution_Data <- Pollution_Data %>% rename(country_code=code)
```

The names of the columns in Pollution_Data are too long, so let's rename the columns. We also will only need the country codes, the year and the overall death rate, so we can remove some columns from this table.

```
Pollution_Data <- Pollution_Data %>% rename(death_rate=deaths_air_pollution_sex_both_age_age_standardiz
```

3 Data Exploration

Now, let's see if we can combine the data and manipulate it in ways and see what we can find.

3.1 GDP

Let's combine our data of Olympic awards with GDP data to see what we will find.

We wish to look at the association between a country's GDP and the number of medals a country gets. To do that, we first need to find the number of medals each country gets each year:

```
MedalCount <- Medal_Data %>% group_by(country_code, year) %>% summarise(medal_count=n());
```

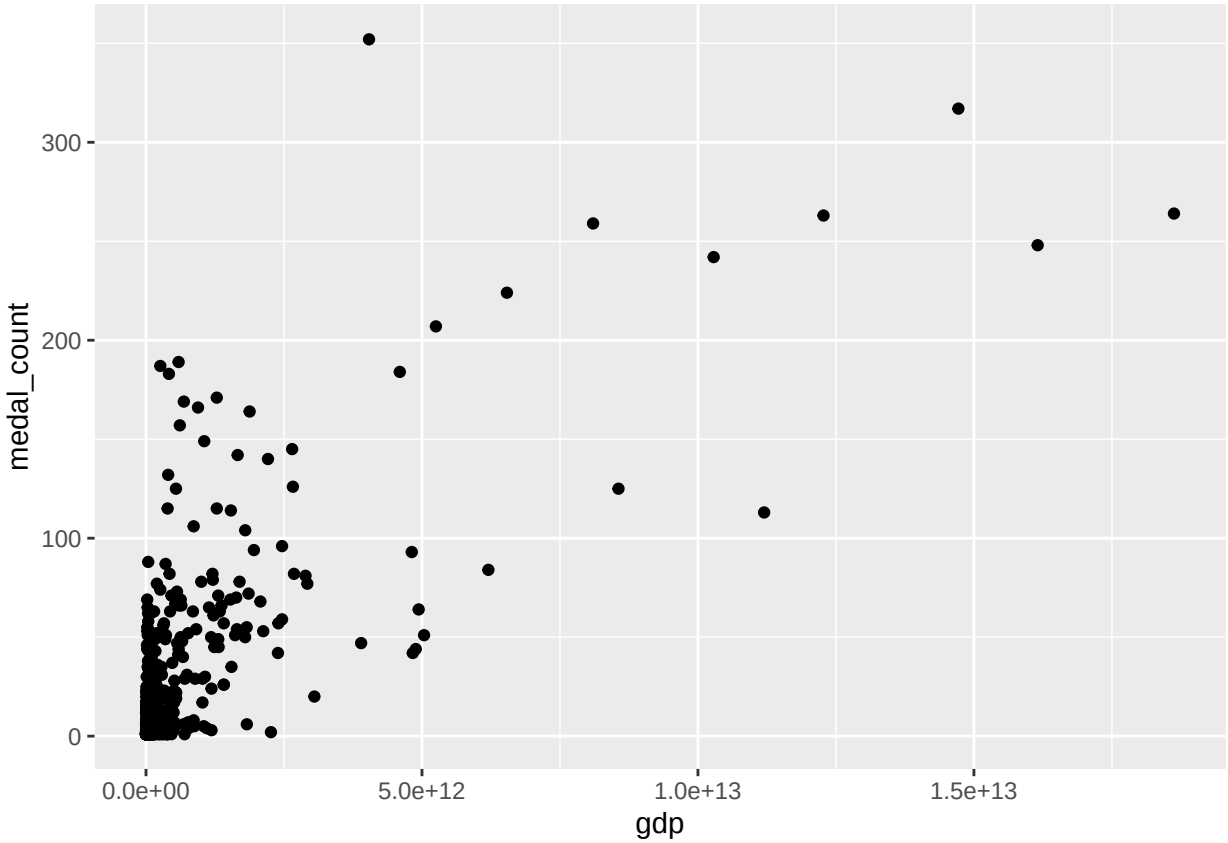
```
## `summarise()` has grouped output by 'country_code'. You can override using the `.groups` argument.
```

Now, we can combine it with the GDP Data by using the inner_join function:

```
GDP_vs_MedalCount <- MedalCount %>% inner_join(GDP_Data, by=c("country_code", "year"));
```

And with that, we can simply graph the relationship between GDP and number of medals won:

```
ggplot(GDP_vs_MedalCount, aes(x=gdp, y=medal_count)) + geom_point();
```



There appears to be a positive association between GDP and number of medals won. We can measure the strength of association by finding the Pearson correlation coefficient:

```
cor.test(GDP_vs_MedalCount$gdp, GDP_vs_MedalCount$medal_count)
```

```
##
##  Pearson's product-moment correlation
##
## data:  GDP_vs_MedalCount$gdp and GDP_vs_MedalCount$medal_count
## t = 24.28, df = 577, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6681004 0.7489829
## sample estimates:
##       cor
## 0.7108842
```

This indicates that the correlation is fairly strong. Let's see what happens if we adjust medal count by how many athletes from that country attended the Olympics that year.

We can find the number of athletes from each country attends the Olympics for each year by collapsing the Athletes_Data dataset:

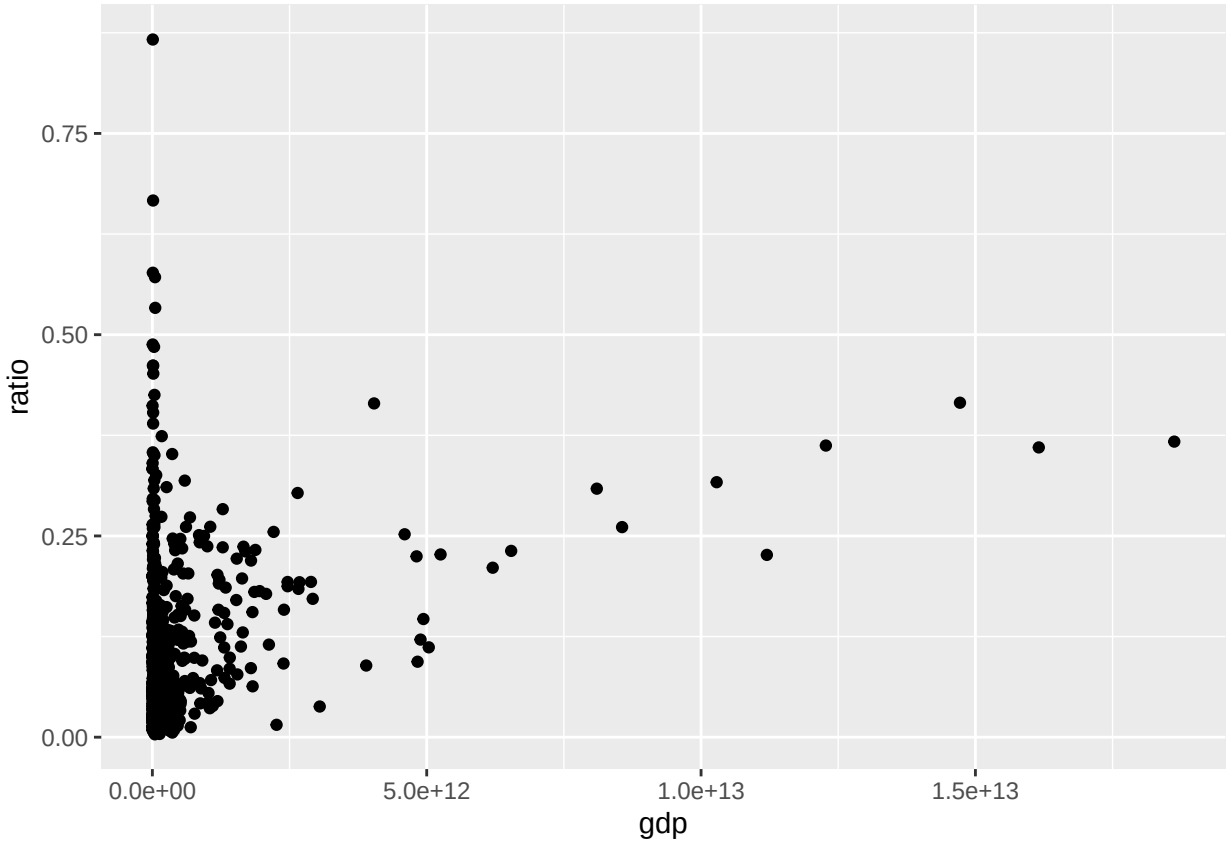
```
Athlete_Count<-Athlete_Data %>% group_by(year, noc) %>% summarise(athlete_count=n()) %>% rename(country
```

```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

Now, let's combine it with our medal count data and graph the relation between GDP and the number of medals a country won per athlete from that country. We will define a ratio column to mean the number of

medals won per athlete.

```
GDP_vs_Ratio <- GDP_vs_MedalCount %>% inner_join(Athlete_Count, by=c("year", "country_code")) %>% mutat

ggplot(GDP_vs_Ratio, aes(x=gdp, y=ratio)) + geom_point()
```



Correlation is apparently positive, but less clear from this plot. Let's see what happens if we were to find the correlation:

```
cor.test(GDP_vs_Ratio$gdp, GDP_vs_Ratio$ratio)
```

```
##
##  Pearson's product-moment correlation
##
## data:  GDP_vs_Ratio$gdp and GDP_vs_Ratio$ratio
## t = 7.0115, df = 577, p-value = 6.616e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2033588 0.3536104
## sample estimates:
##       cor
## 0.2801999
```

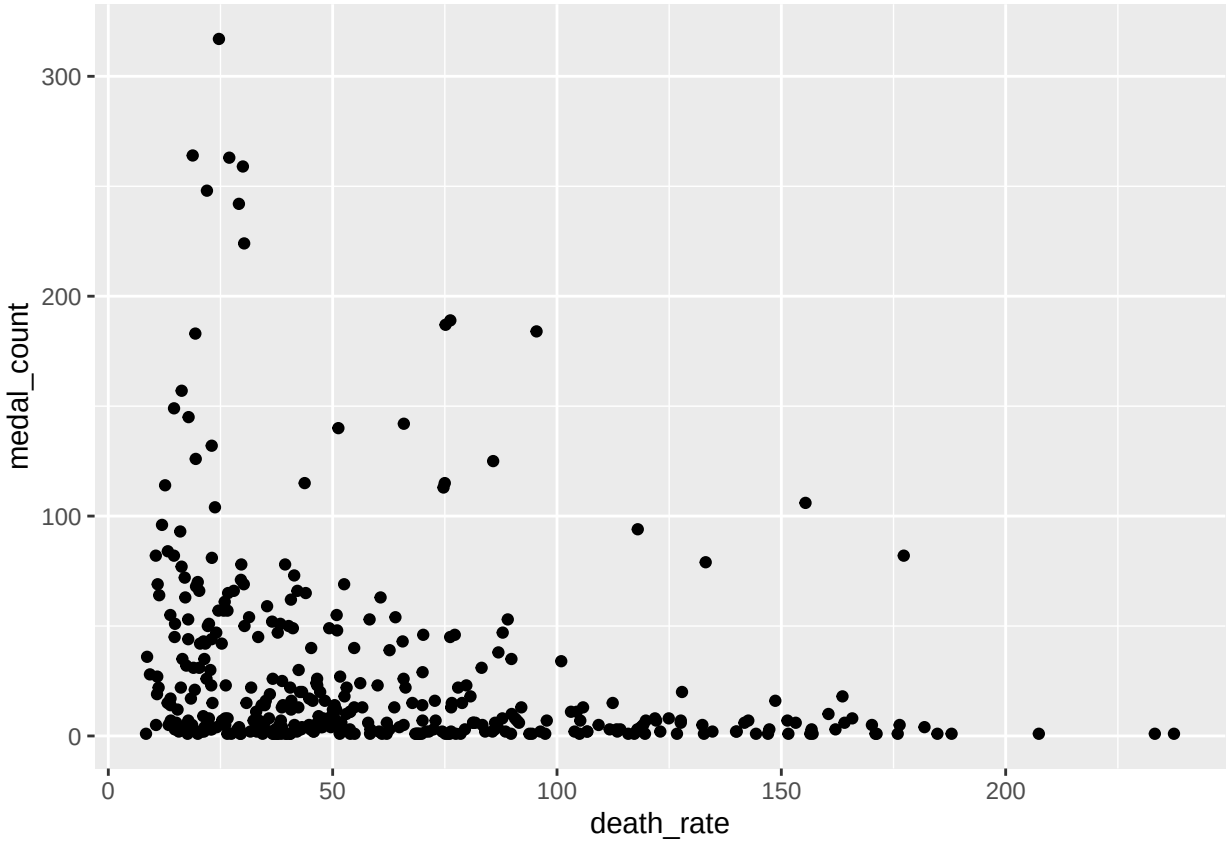
As you can see here, this is a weaker correlation. This suggests that there is a positive association between GDP and number of medals won, but number of athletes is a possible confounder in this relation. This makes sense because both would be positively associated with population.

3.2 Pollution

Now, let's see if there is an association between air pollution death rates and number of medals won. If pollution death rates in a country are higher, then pollution is probably worse there. So, the athletes from that country are not expected to perform as well. Thus, our hypothesis is that there is a negative association between pollution death rate and number of medals won.

Let's combine the data and graph the association:

```
Pollution_vs_MedalCount <- MedalCount %>% inner_join(Pollution_Data, by=c("country_code", "year"));

ggplot(Pollution_vs_MedalCount, aes(x=death_rate, y=medal_count)) +geom_point()
```



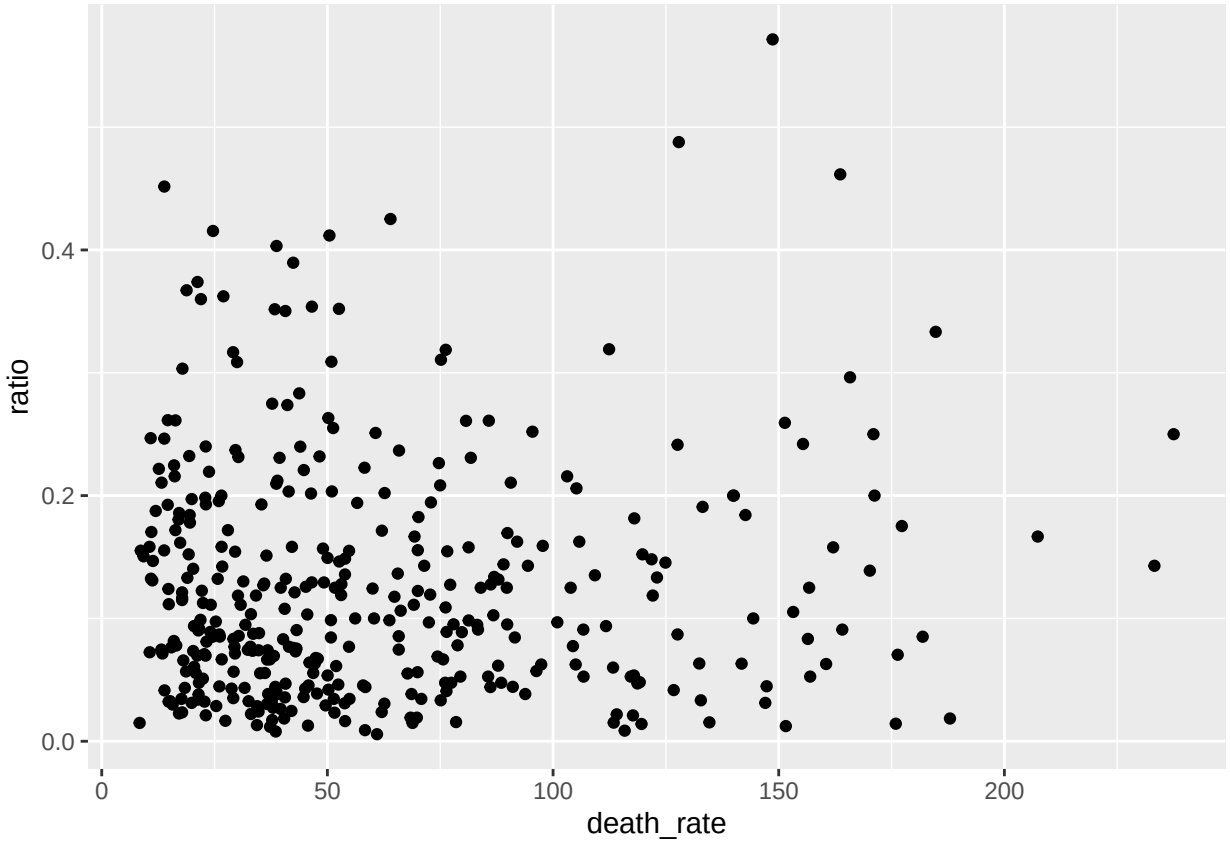
There is no clear correlation shown in this graph. Here is the Pearson correlation coefficient:

```
cor.test(Pollution_vs_MedalCount$death_rate, Pollution_vs_MedalCount$medal_count)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Pollution_vs_MedalCount$death_rate and Pollution_vs_MedalCount$medal_count
## t = -4.7636, df = 377, p-value = 2.718e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3310597 -0.1409190
## sample estimates:
##       cor
## -0.2382713
```

As expected, it is a negative correlation, but it is weak. Let's see if we can get a better correlation by adjusting for the number of athletes:

```
Pollution_vs_Ratio <- Pollution_vs_MedalCount %>% inner_join(Athlete_Count, by=c("year", "country_code"

ggplot(Pollution_vs_Ratio, aes(x=death_rate, y=ratio)) + geom_point();
```




```
cor.test(Pollution_vs_Ratio$death_rate, Pollution_vs_Ratio$ratio)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Pollution_vs_Ratio$death_rate and Pollution_vs_Ratio$ratio
## t = 0.88145, df = 377, p-value = 0.3786
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.05563866  0.14542039
## sample estimates:
##        cor
## 0.04535012
```

This actually made the correlation weaker and positive. We can conclude from this that there is no strong correlation between the death rate due to air pollution and the athletes' performance.

3.3 Outside Events

Let's see how the trend in a country's medal count change in response to certain events.

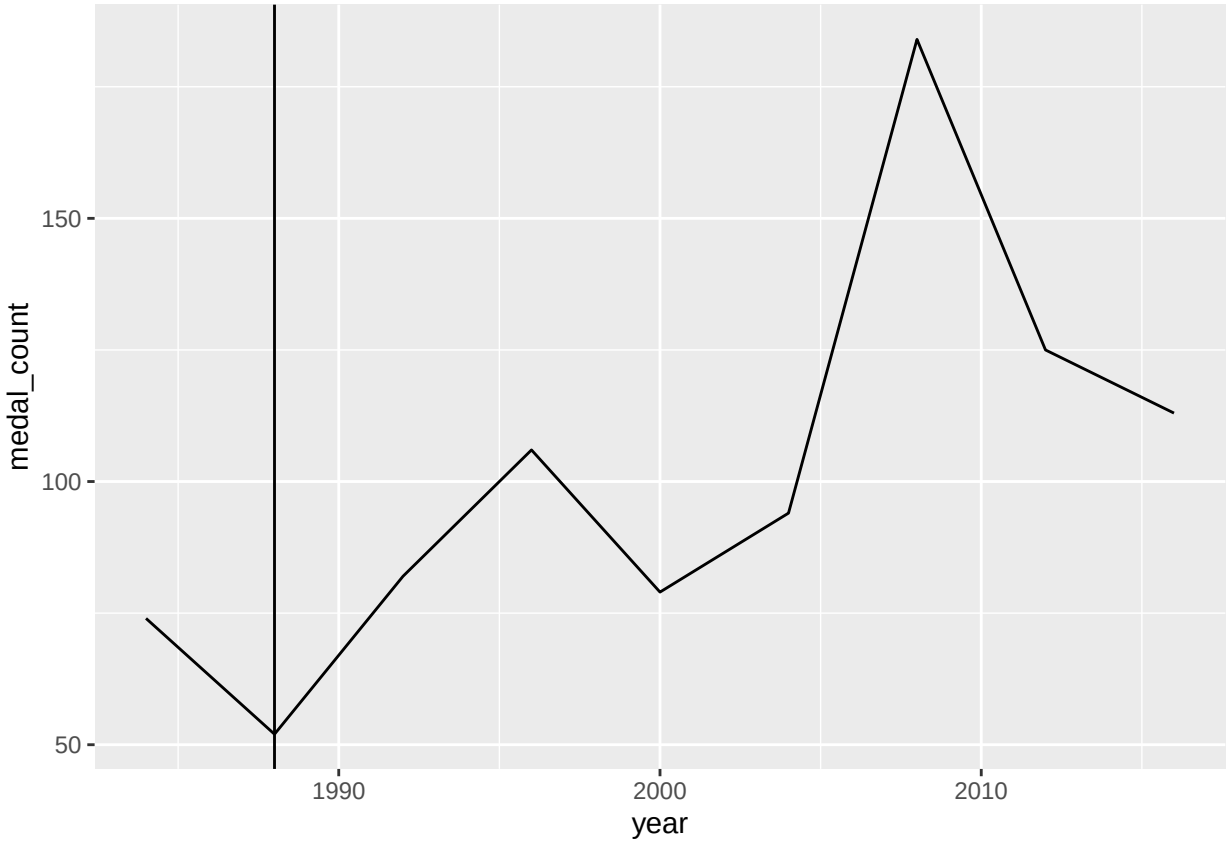
First of all, China is a country known to be good at ping-pong. Let's see how China's medal count changed after ping-pong was added to the Olympics.

Let's first filter our MedalCount data table to include only awards from China:

```
MedalCount_China <- MedalCount %>% filter(country_code == "CHN")
```

We can now make a line graph that shows the trend China's medal count over the years. Since ping-pong was added to the Olympics in 1988, we will also add a vertical line to indicate it.

```
ggplot(MedalCount_China, aes(x=year, y=medal_count)) + geom_line() + geom_vline(xintercept=1988)
```

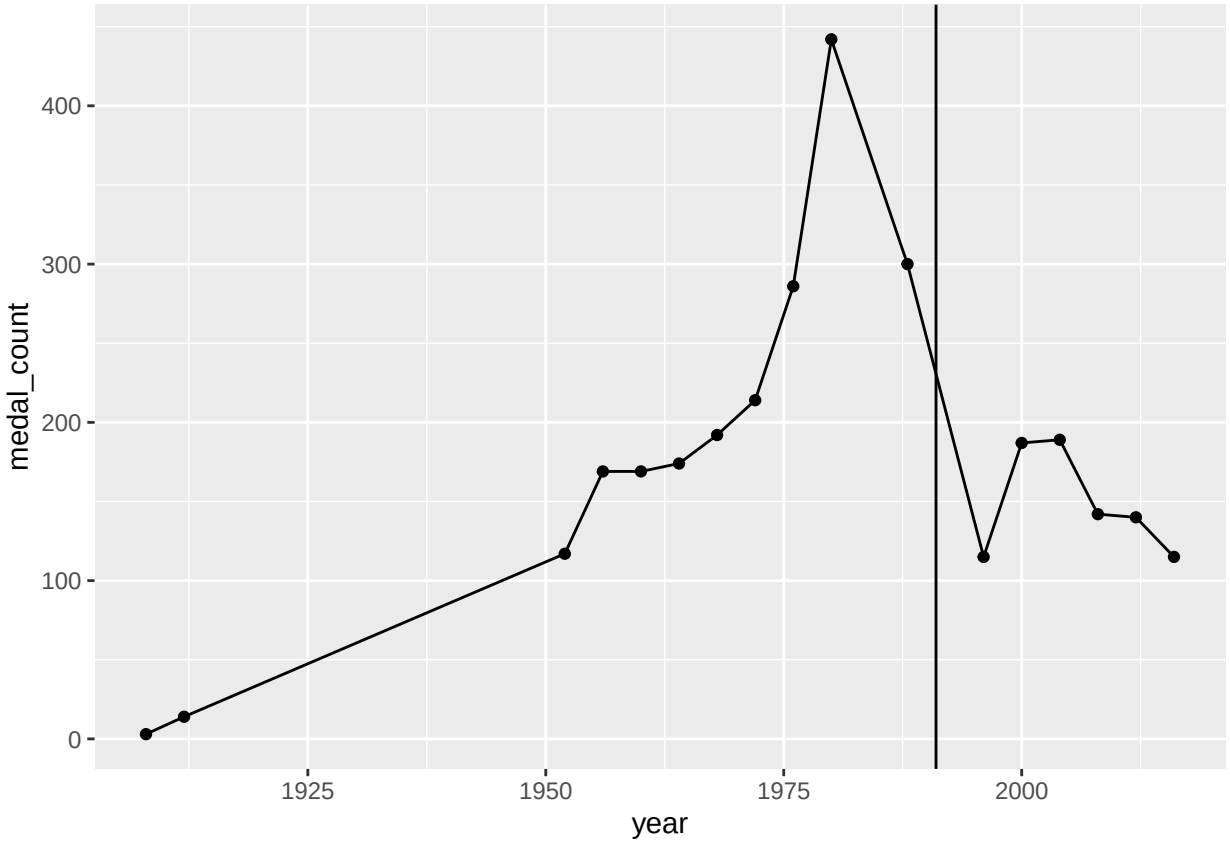


Note that the only year before 1984 that China competed in the Summer Olympics was 1952. In that year, China won 0 medals.

As you can see from this plot, China's medal count started going up after table tennis was added to the Olympics.

Now, let's see how Russia did in the Olympics during and after the Soviet Union. The Soviet Union collapsed in 1991.

```
MedalCount_USSR <- MedalCount %>% filter(country_code=="URS");
MedalCount_Russia <- MedalCount %>% filter(country_code=="RUS");
MedalCount_Russia_All <- rbind(MedalCount_USSR, MedalCount_Russia);
ggplot(MedalCount_Russia_All, aes(x=year, y=medal_count)) + geom_line() + geom_point() + geom_vline(xin
```



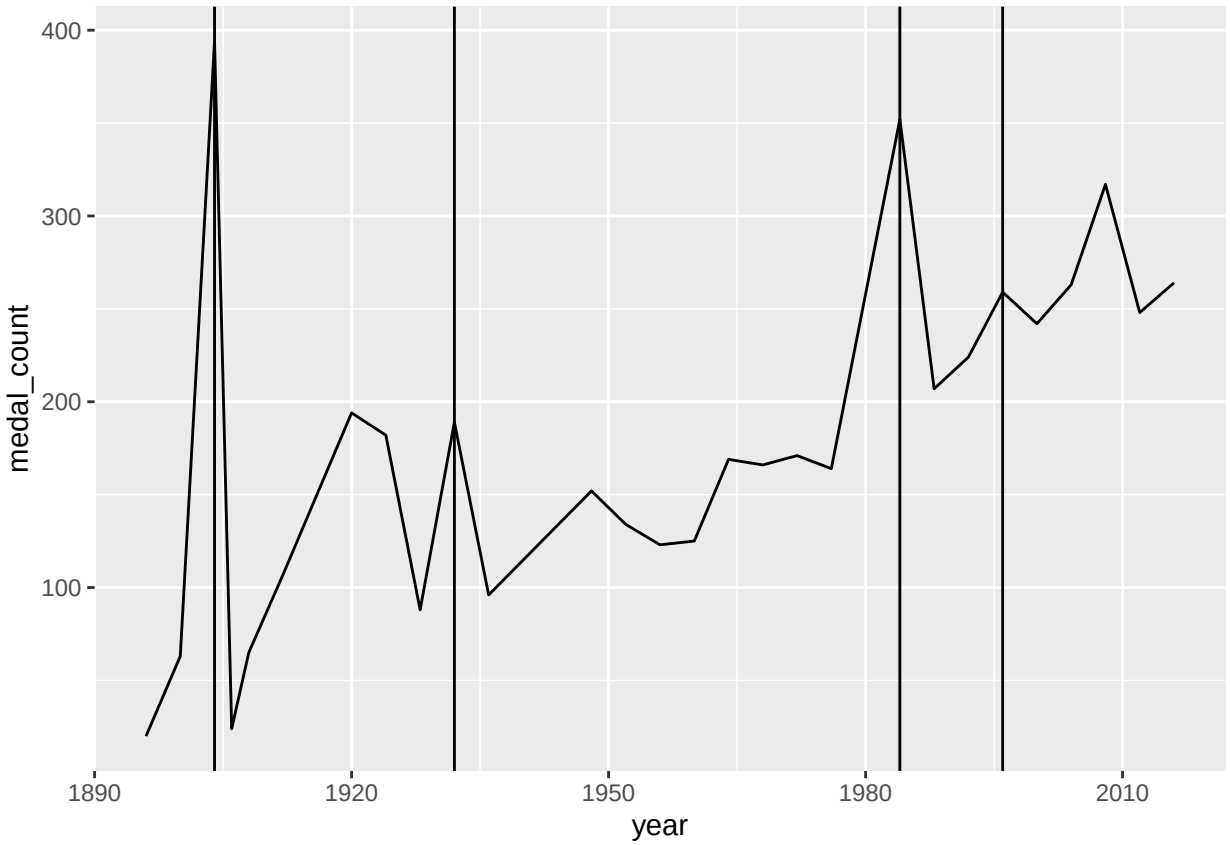
As you can see in this graph, Russia got fewer medals each year after the Soviet Union than it did during each of the last years of the Soviet Union.Note that the 1980 spike in medal count could be due to the Soviet Union's hosting, which we will talk about in the next section.

3.4 Hosting

Finally, let's see if we can observe an effect of a country's hosting on its medal count. We will look at USA, United Kingdom and Germany as they are among the top hosting countries in our data.

Let's see USA's trend in medal count over the years and mark the years USA hosted (1904, 1932, 1984 and 1996).

```
MedalCount_USA <- MedalCount %>% filter(country_code=="USA");

ggplot(MedalCount_USA, aes(x=year, y=medal_count)) + geom_line() + geom_vline(xintercept=c(1904, 1932,
```



Notice that whenever USA hosts the Summer Olympics, its medal count spikes. Let's see if this is also true for the United Kingdom and Germany.

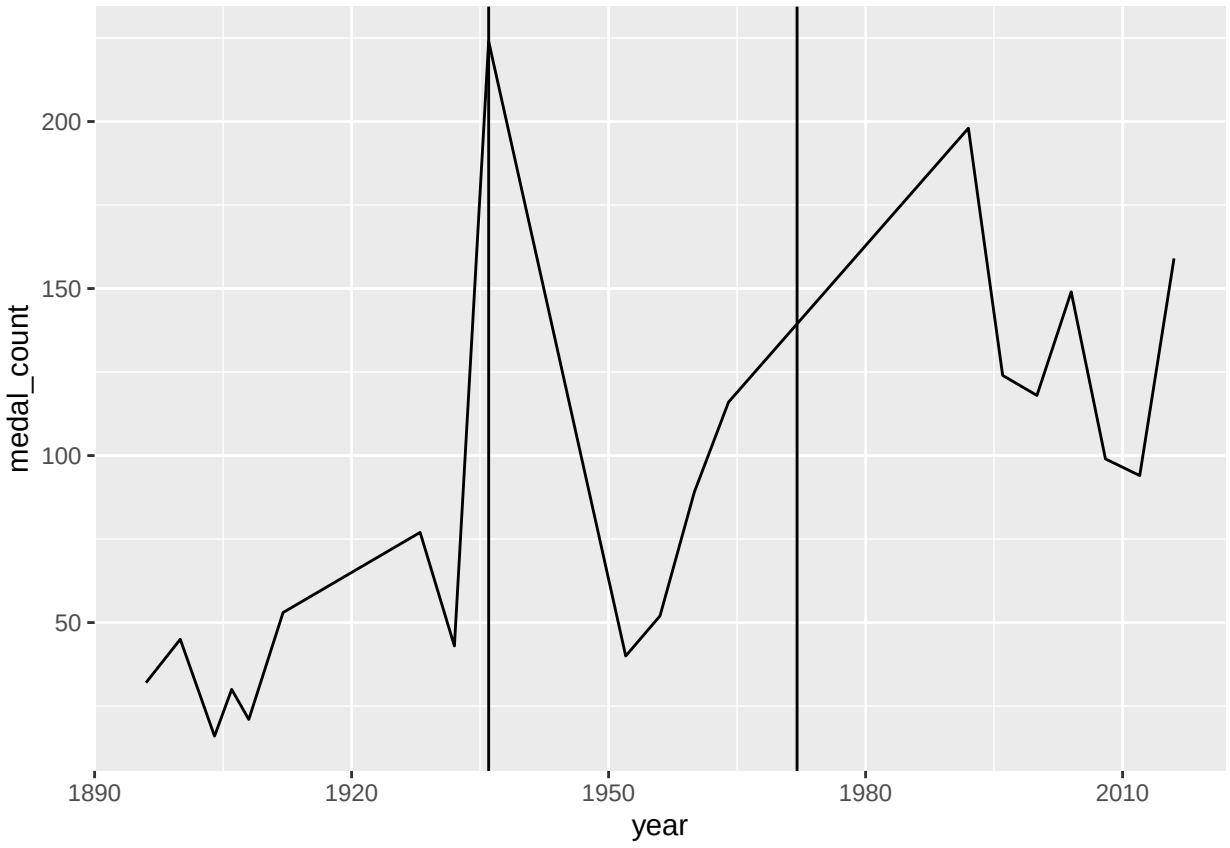
We will test this on the United Kingdom next. The Olympics were hosted in the United Kingdom in 1908, 1948 and 2012.

```
MedalCount_GBR <- MedalCount %>% filter(country_code=="GBR");

ggplot(MedalCount_GBR, aes(x=year, y=medal_count)) + geom_line() + geom_vline(xintercept=c(1908, 1948, 1
```

Again, we observed a spike in medal count during years the Summer Olympics were hosted in the United Kingdom.

Finally, we will test this on Germany. Germany hosted in 1936 and 1972.

```
MedalCount_GER <- MedalCount %>% filter(country_code=="GER");

ggplot(MedalCount_GER, aes(x=year, y=medal_count)) + geom_line() + geom_vline(xintercept=c(1936, 1972))
```



There is a major spike in 1936, but no spike in 1972.

Thus, if a country hosts the Olympic games, it tends to win more medals, although not always.

4 Conclusion

In conclusion, there is no noticeable association between pollution and medal count or medal count per athlete. There is an association between GDP and number of medals won, but number of athletes may be a confounder because there is no association between GDP and number of medals won per athlete. There also seems to be no correlation between pollution and medal count or medal count per athlete. Some events that occur outside of the Olympics could be associated with a change in medal count, like how the Soviet Union's collapse is associated with Russia's decrease in medal count. If a country hosts the Olympics, it is more likely to get a spike in medal count.

Although we had a few disappointing results, there are many interesting things that can be found if our data on the Olympic athletes and awards are processed or combined with other data and here, we have only scratched the surface of what is possible.