Data Collection:

The data is collected from this source, <https://ourworldindata.org/>. In this website we can find whatever indicators we require for specific country and the information is ranging from 1950’s to 2017. We have choose primary datasets as country’s economic indicators. We have gathered the below indicators from the above mentioned website.

**Target Variable**:

**homicides-per-100000-people-per-year.csv (Homicide\_Rate)**: Homicides per 100,000 people year.

**Primary Dataset(Economic Indicators):**

1. **annual-healthcare-expenditure-per-capita.csv** : Total health expenditure is the sum of public and private health expenditures as a ratio of total population. Data are in international dollars converted using 2011 purchasing power parity (PPP) rates.
2. **gdp-per-capita-worldbank.csv:** GDP per capita adjusted for price changes over time (inflation) and price differences between countries – it is measured in international-$ in  
   2011 prices
3. **infant-mortality.csv:** The share of newborns who die before reaching one year of age.
4. **life-expectancy.csv:** Life expectancy at birth is defined as the average number of years that a newborn could expect to live if he or she were to pass through life subject to the age-specific mortality rates of a given period.
5. **malnutrition-death-rates.csv:** Deaths from protein-energy malnutrition per 100,000 people.
6. **median-age:** The median age divides the population in two parts of equal size: that is, there are as  
   many persons with ages above the median age as there are with ages below the  
   median ages
7. **size-poverty-gap-countries.csv:** The poverty gap is the amount of money that would be theoretically needed to lift the incomes of all people in extreme poverty up to the international poverty line of $1.90 a day. These estimates are expressed in international dollars using 2011 PPP conversion rates. This  
   means that figures account for differences in prices levels, as well as for inflation
8. **public-health-expenditure-share-GDP-OWID.csv:** Public health expenditure includes: recurrent and capital spending (central and local levels), external borrowing and grants (including donations from international agencies and NGOs), and social or compulsory insurance funds.

As secondary datasets, we chose one indicator women empowerment and education indicators, which are:

1. **fertility-rate-complete-gapminder.csv:** Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.
2. **Government Expenditure on education per-capita:** It is the product of GDP per capita and government expenditure on education as percentage of GDP per capita variables.

These are the economic, educational and women empowerment indicators used in our data analysis to find solution for homicide rate prediction.

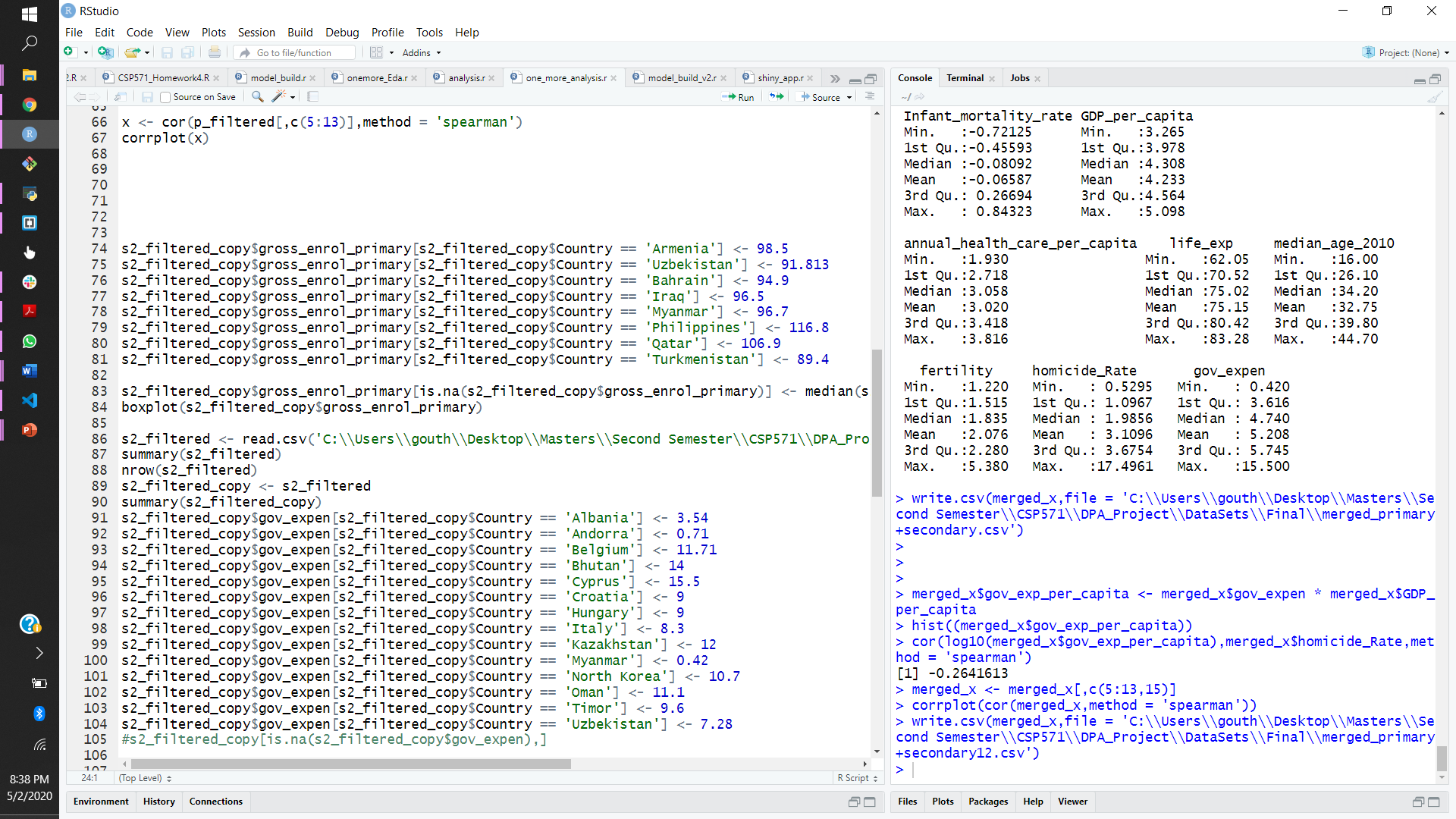
**Data Cleaning:**

As there are many indicators we only want to choose that indicators which are contributing to the target variable. This can be checked by getting correlation coefficient between the predictor indicators to the target. Before doing this we need to clean the data. To clean the data our first assumption is to choose only those indicators which are having less than (15%)missing values of total available for each indicator. We started looking for the years which there is less missing values and year 2012 has comparatively large data when compared with other years. So we started looking for the countries of year 2012, in that we found that the information for all countries is not available. So we filtered out the countries which are having less missing values and Europe, Asian countries are giving us consistent information of the indicators we choose. Now, we have performed data cleaning on this 87 countries (Europe + Asian) of year 2012.

After finalizing the data which we want to work, we have only choose those indicators whose missing values is less than threshold (15%), poverty gap indicator has more number of missing values so we have excluded that indicator and remaining every indicator has reasonable amount of missing values. Now we are left with 9 indicators(7 from primary,2 from secondary datasets).

**Filling Missing Values:**

* **Recover missing values:**  Using this another reliable website v <https://knoema.com/>. We tried to recover the missing values for the countries of year 2012 and manually entered them.

The below is the example for the recover missing values for the secondary dataset indicator

