**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.

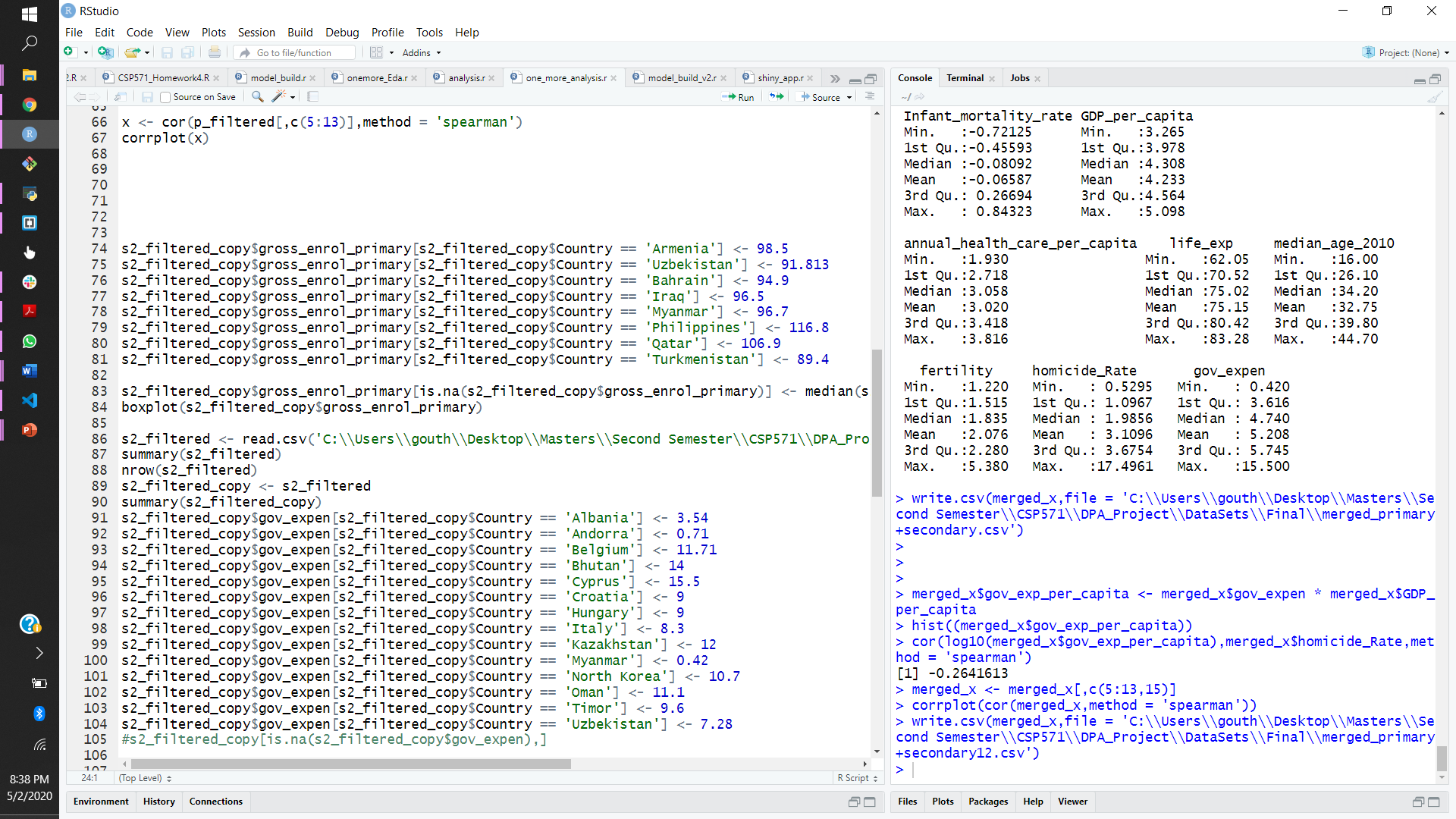
**Data Cleaning and Exploratory Data Analysis:**

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.

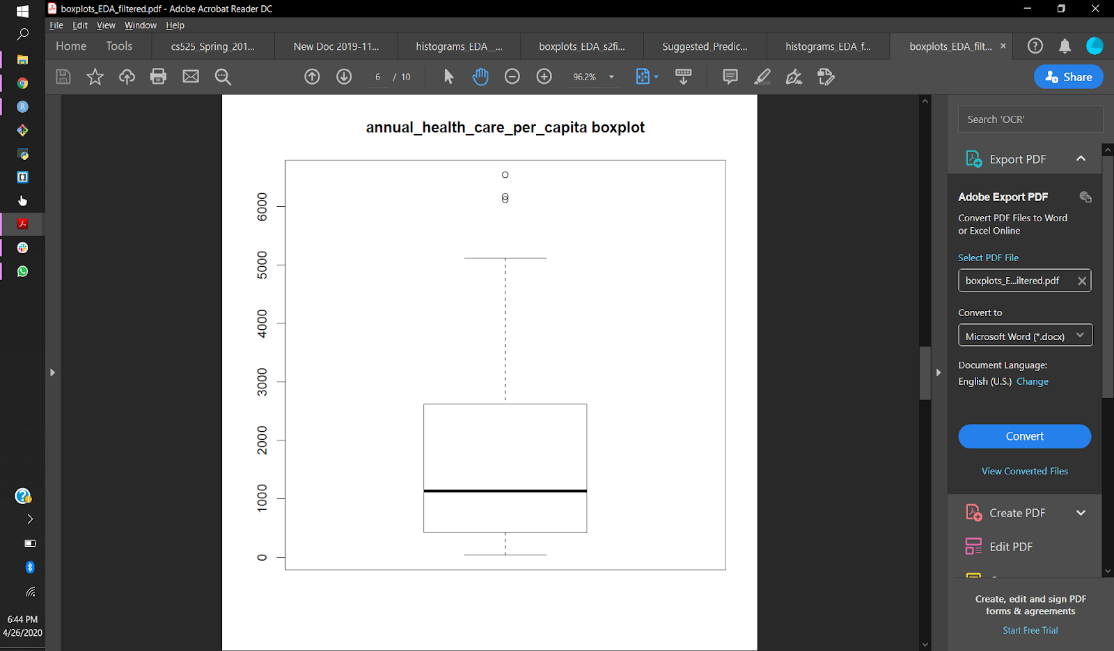
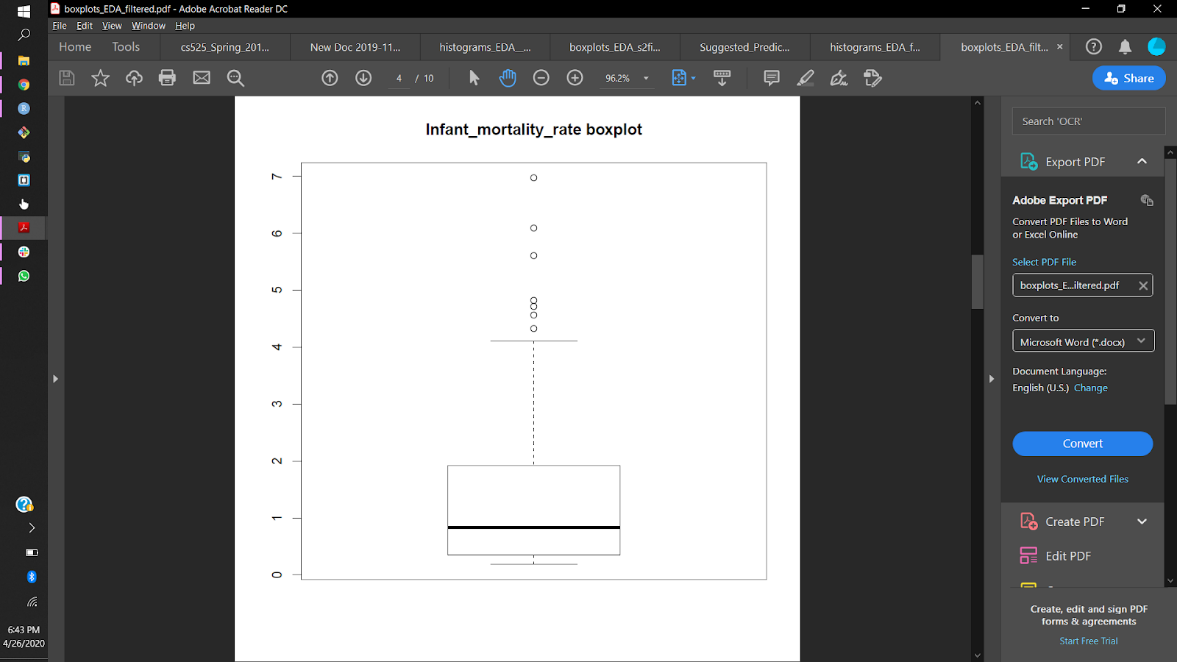
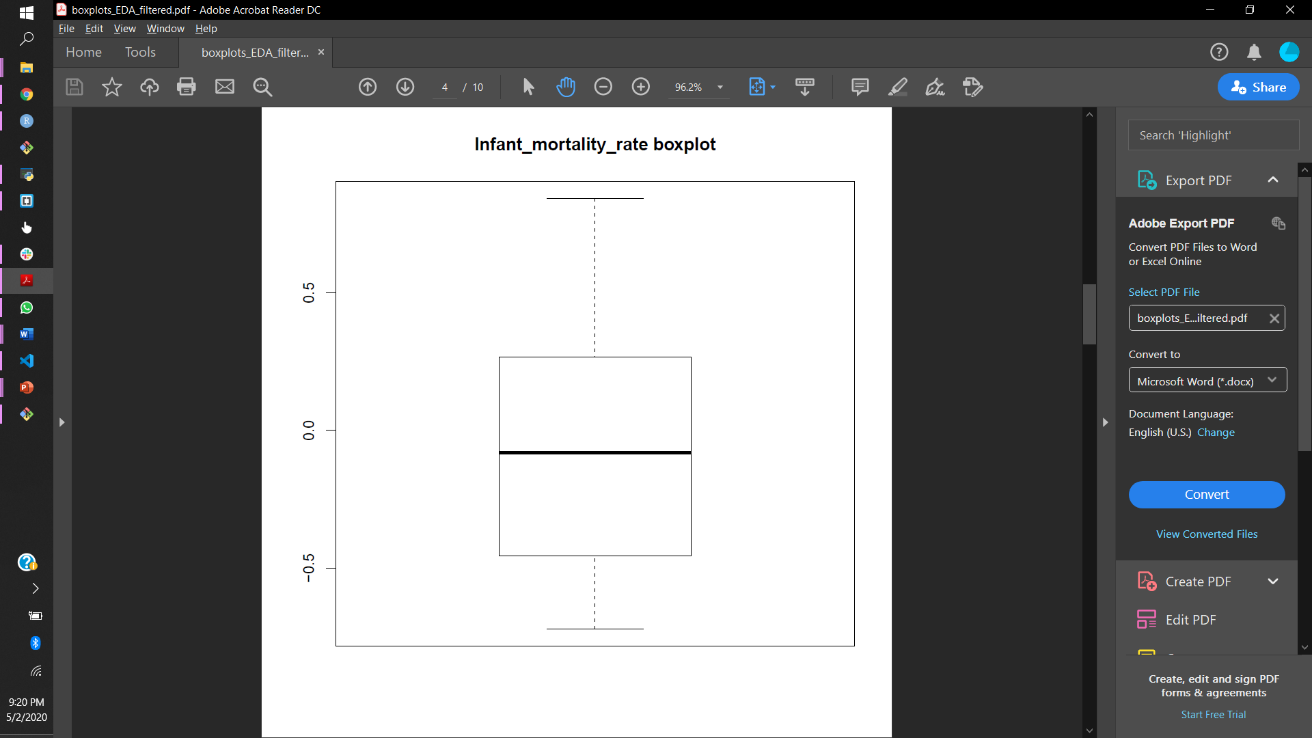
The below is the example for the recover missing values for the secondary dataset indicator **Government Expenditure on education per-capita.** Similarly, we did for all possible indicators and tried to maintain originality in data without any mean/median imputations.

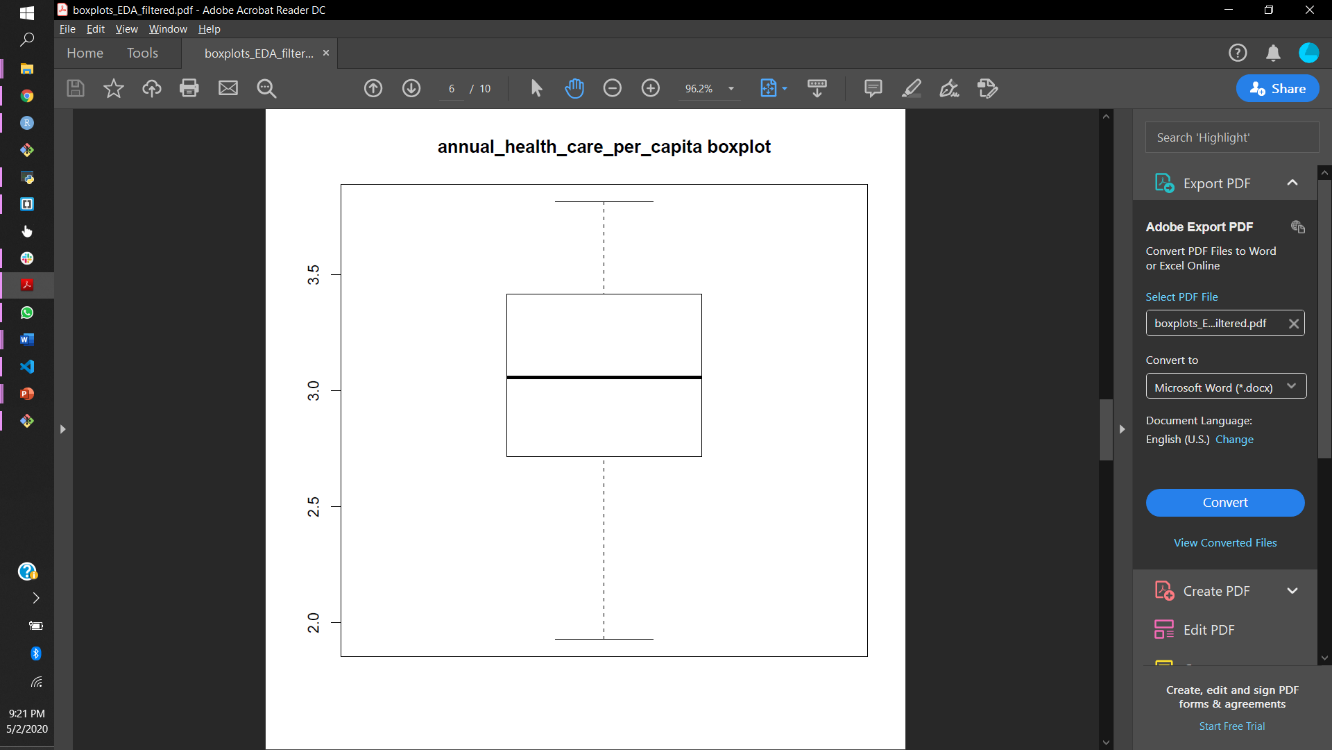
* **Mean/Median Imputation:** whenever we find there are outliers after we plot boxplots once recover missing values step is done, we tried to impute missing values by median. If there is no outliers then mean of the indicator is used to replace the missing values. This step is followed for fertility rate, annual health care per capita, GDP per capita and some other indicators which have missing values.

After completing missing values step, we still see there are outliers for the indicators. In the below images you can see there are outliers for infant mortality rate , annual health care per capita indicators. Not only for these two but this is the situation for some other indicators also. We found out this is happening due to the huge range of values are present for the specified indicator for example annual health per capita values are ranging from 0 to 6000 due to this, we are getting outliers. So, we have applied transformations on the data to overcome this problem.

**Transformations**:

* **Log or root transformations:** Whenever the data is right skewed , we can check this by plotting histogram for that indicator. If the data is right skewed we will first prefer to apply log transformation and see whether the data is normally distributed.
* **Square transformations:** If the data is left skewed we have preferred to apply square transformation and see whether the data is normally distributed. Below are the examples before and after the transformations.



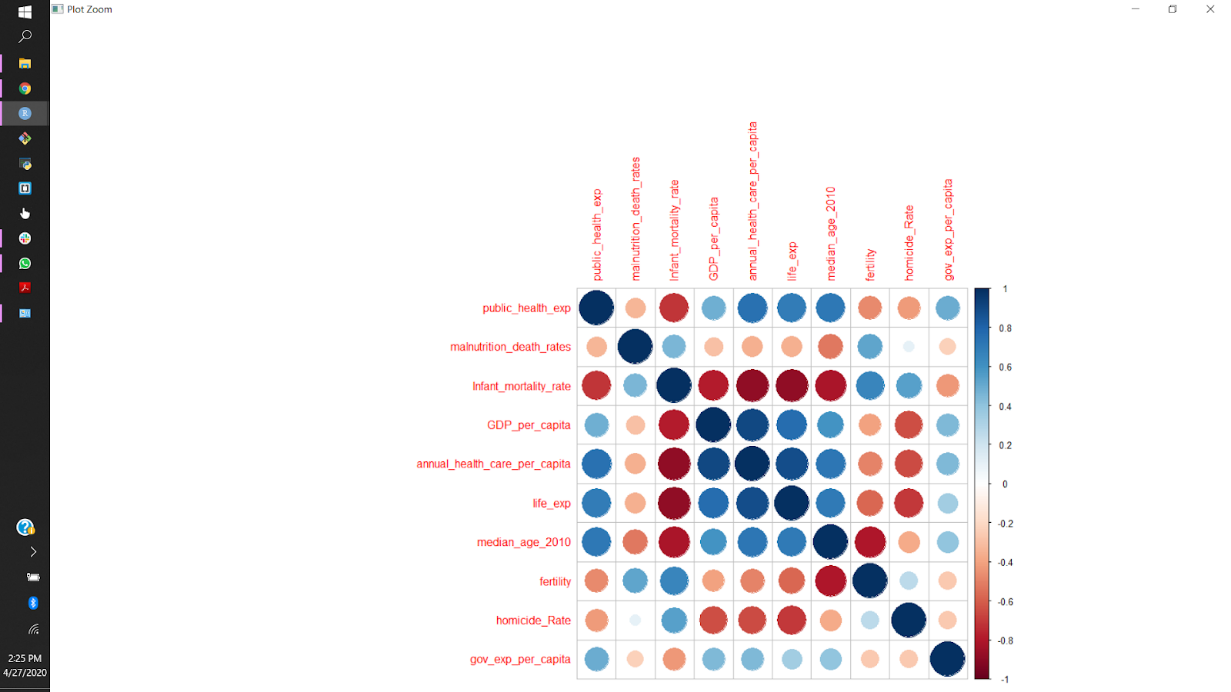
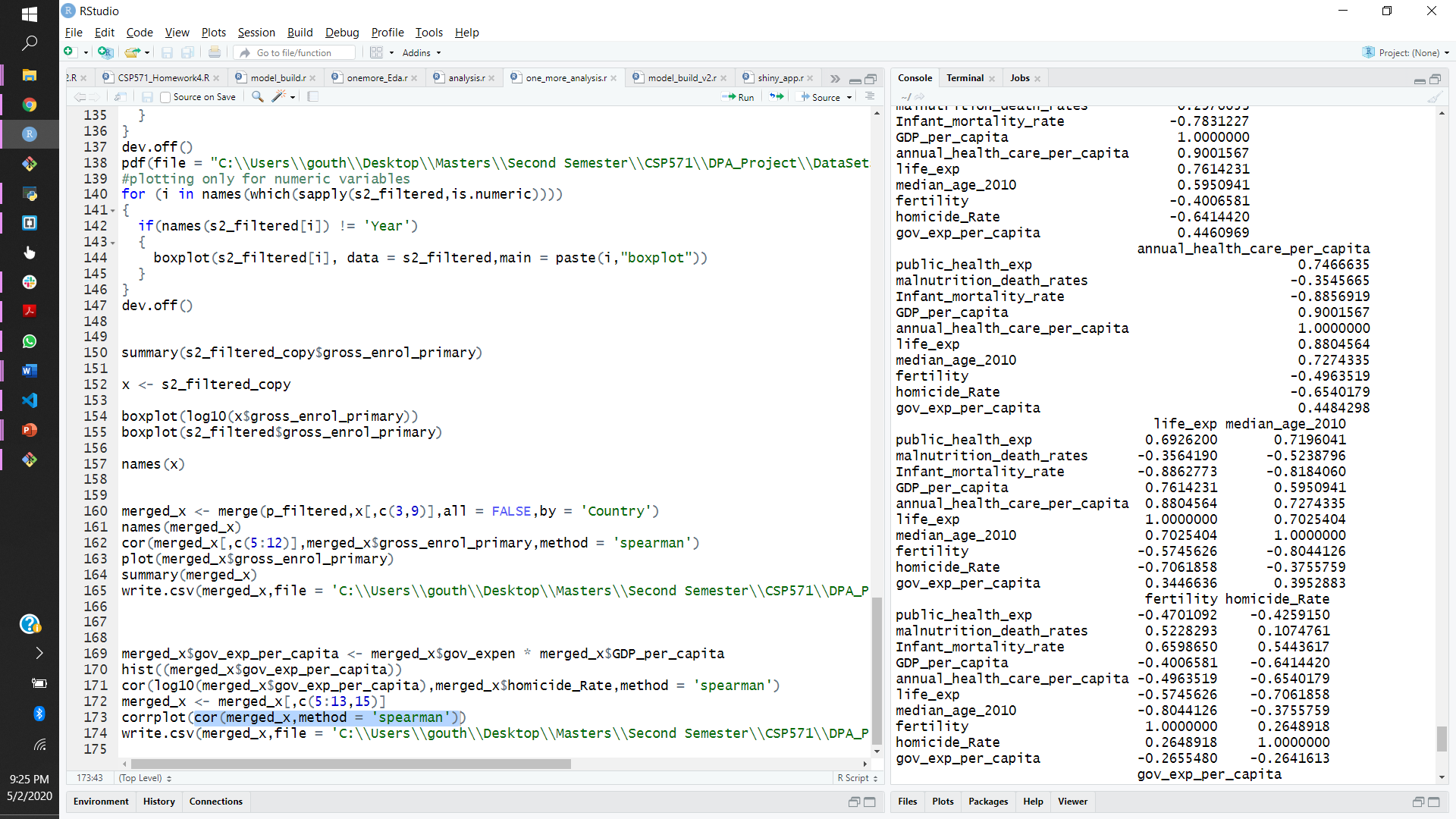


**Correlations:**

After cleaning the data, we tried to check the correlation coefficient with the finalized primary+ secondary datasets indicators with the target variable homicide rate and plotted the correlation plots.

We found that almost every indicator is having reasonably strong correaltions with the target variable. So we decided to check the model performace using these indicators.

So, we have decided to check the model performance using primary dataset and also considered to see whether including the secondary datasets will increase the model performance or not.

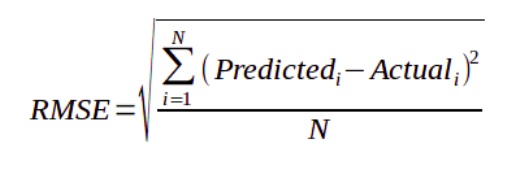
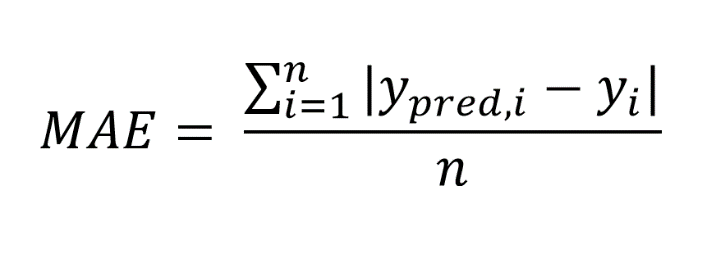
Below are the correaltion values with the target variable and correlation plot for all the indicators we are gonna use in the model building and evalutation.

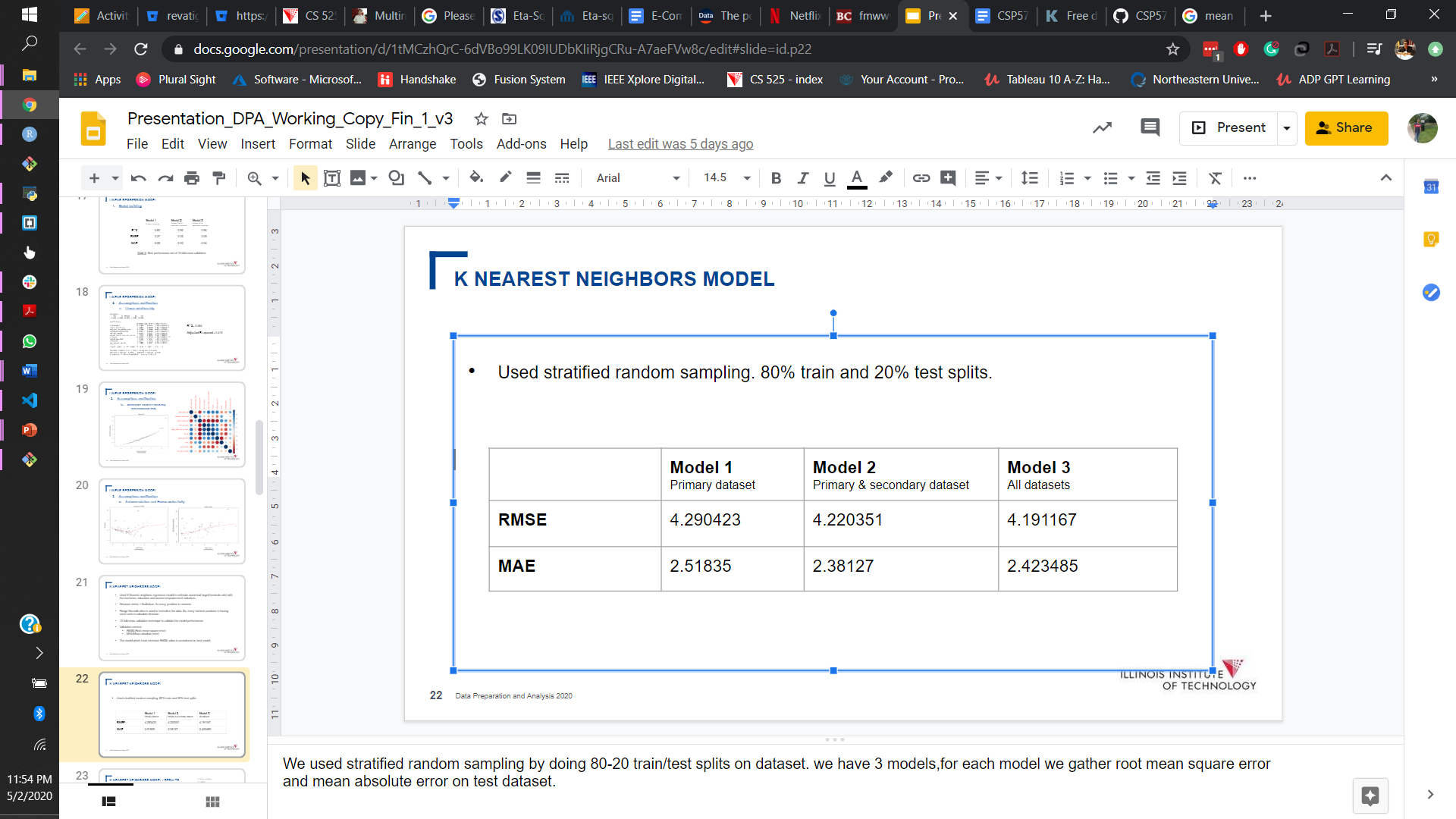
**K-Nearest Neighbors(KNN):**

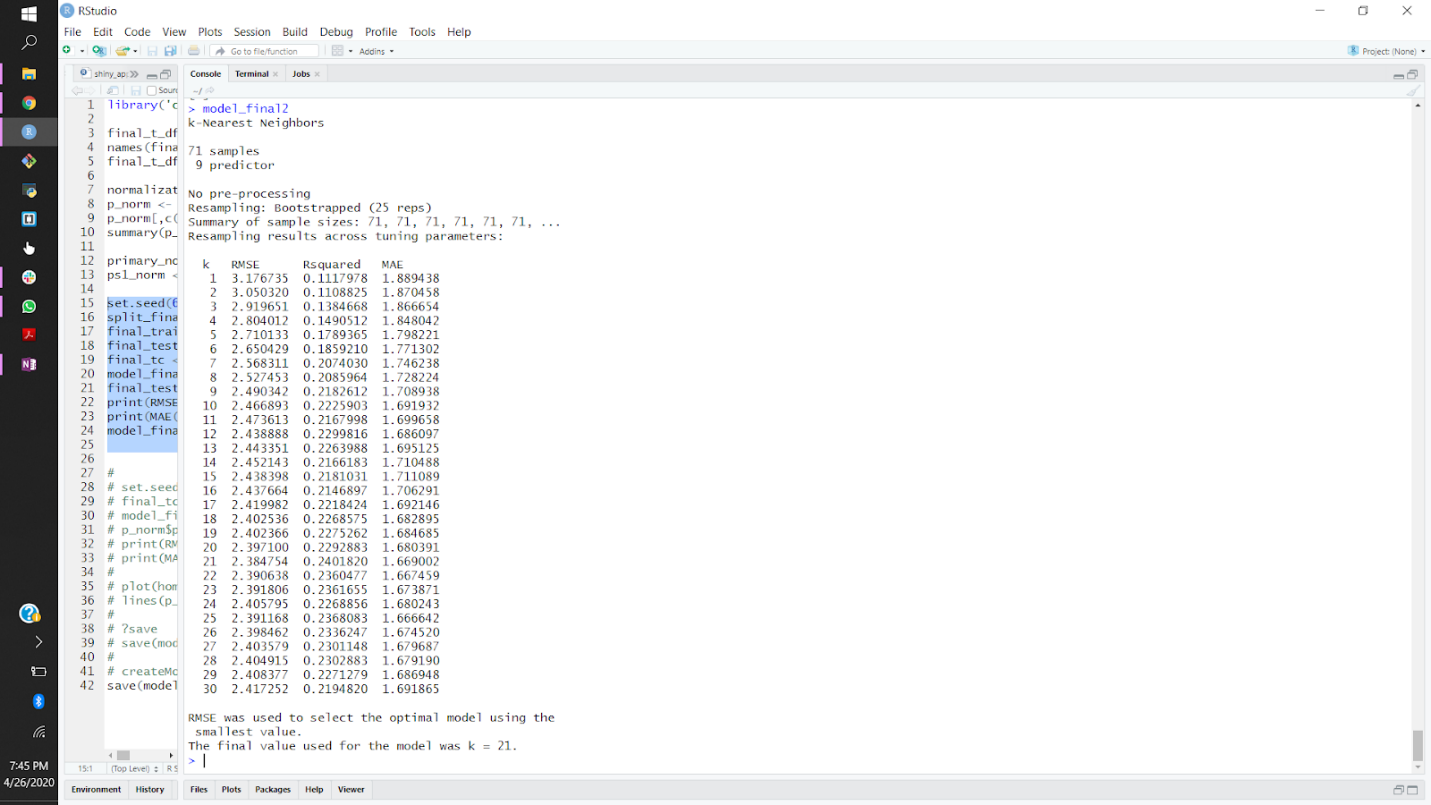
KNN is a technique in which predicted value for a given input will be the mean of the k -nearest neighbors for given input. Nearest neighbors are decided by calculating the distance between the observations. Of all distance metrics, we found that euclidean distance is performing better. KNN model we choose is regression version. As, we using distance metric to calculate the nearest neighbors the units of the observations might create problems.To overcome this issue, we have to normalize the data before training the model. So to normalize the data, we have applied range normalization.

Formula for Range normalization is . Where x will be the indicator/column of a dataframe.After applying this range normalization every column will be having values range of 0 to 1. Now, calculate distances will be giving correct results.

K – value is determined by repeating k for different values say 1 to 30 and choose the k value which has minimum error rate on the test dataset.In this way we will take care of initial steps before model building of KNN.

Now, we will split the dataset to train, test using stratified random sampling. Training dataset will be containing 80% and test is of 20%. On the test dataset we will measure the model performance using Root Mean Square Error(RMSE) and Mean Absolute Error(MAE).

We will have 3 models , model 1 trains only on primary dataset, model 2 primary + one secondary , model 3 is all indicators. The below are the RMSE, MAE values of the 3 models on the test datasets.

Best model is considered as the one which has lowest RMSE value. After looking above results we can clearly see that model 3 has best perfomance with low RMSE value of 4.19 of all other models. We have used 10-fold cross validation to train the model the results of the training model using 10-fold CV and best k value is provided . The K value of 21 has lowest RMSE value while training so k = 21 is the final k considered.

Resources:

<http://fmwww.bc.edu/repec/bocode/t/transint.html>