John Alino

CSC 532

jalino2 - 670196995

May 3, 2020

**Predicting Economic Freedom of Countries**

**Abstract**

The project I’m going to be working on is to determine which variables of each country determines their economic freedom. There are only two categorical variables, and the rest are numerical. There are quite a few missing variables in a lot of the rows, and I have to determine if there’s enough correlation predicted in each one to conclude whether or not I should keep each one for the predictions later on. For the categorical variables, I used one-hot embedding to get them ready for the neural models. Other predictive methods I used include lasso/ridge linear regression, random forest, and GBT.

**Problem definition and project goals**

The main goal of the project is to predict what level of economic freedom each country will have depending on its features.

The dataset I’m using for this project can be found in this link1: <https://www.kaggle.com/gsutters/economic-freedom>. It is an index published in Economic Freedom of the World by the Fraser institute. This dataset measures the degree to which the policies and institutions of countries are supportive of economic freedom. There are five main areas to which the features of this dataset focus on: Size of Government, Legal System and Property Rights, Sound Money, Freedom to Trade Internationally, and Regulation. All the features are numerical except for the identifiers of each country.

With the data, I will first explore through the general summary and statistics. This will give me a broad idea of what I’m working with and to determine what steps I will need to take to clean up the data further. What I will do to prepare the data for the prediction models is to handle the missing variables, scale the numerical variables, and embed the categorical variables.

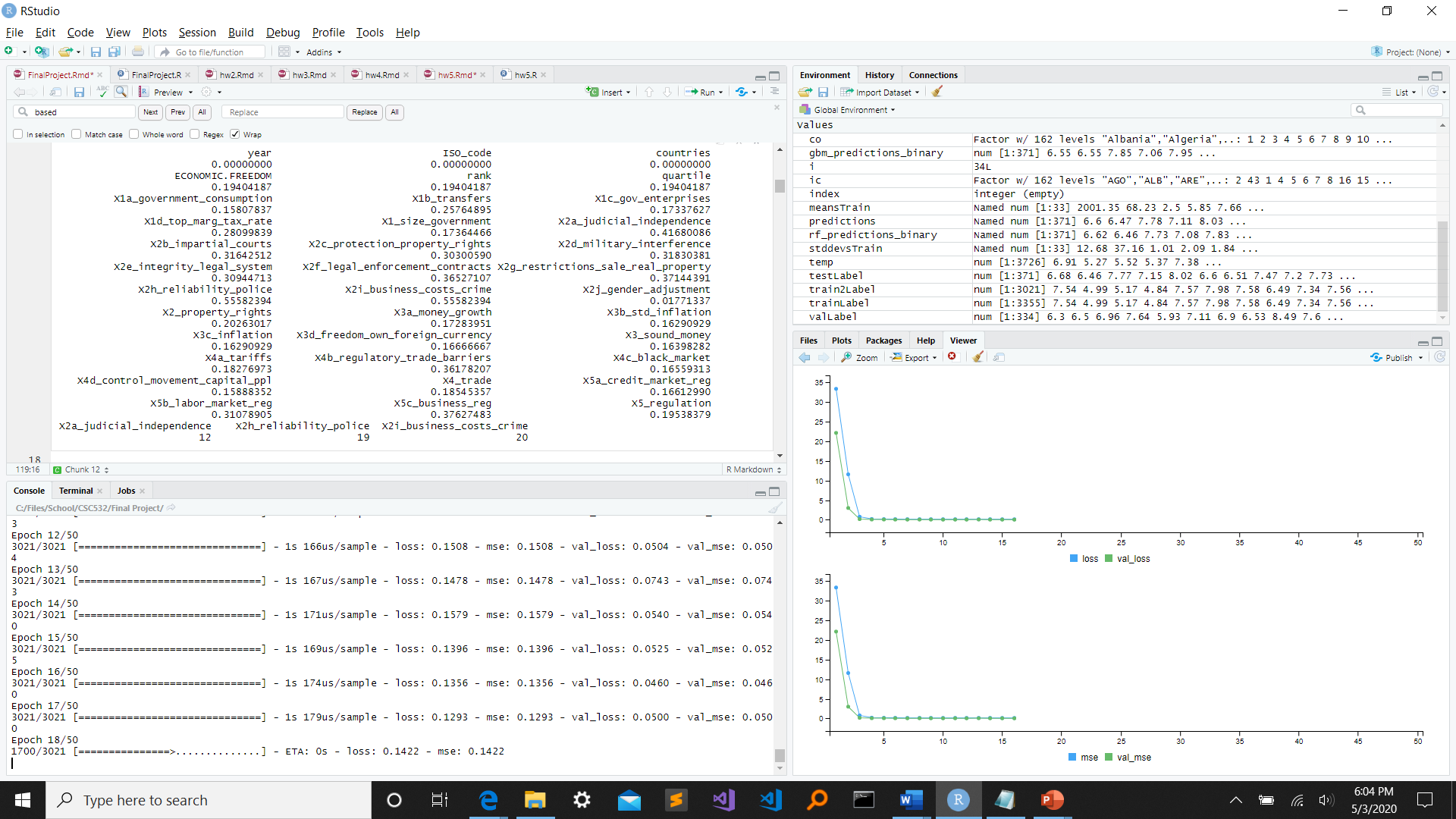
**Related work**

There haven’t been too many other works out there that are addressing the same problem, but I did find one that tackled trying to determine the top 11 factors to predict the country’s economic freedom. It is also written in R.

The author of that R notebook also started off looking through the dataset features and examining which ones have missing variables2. The data preparation phase involved three main points: changing the data types, renaming the columns, and removing the numerical values in the column names. After they did all the data analysis, they concluded that these are the top 11 variables for predicting the economic freedom of a country: freedom own foreign currency, judicial independence, credit market regulations, inflation, government enterprises, control of movement of capital and people, tariffs, labor market regulations, marginalized tax rate, government consumption, and black market.

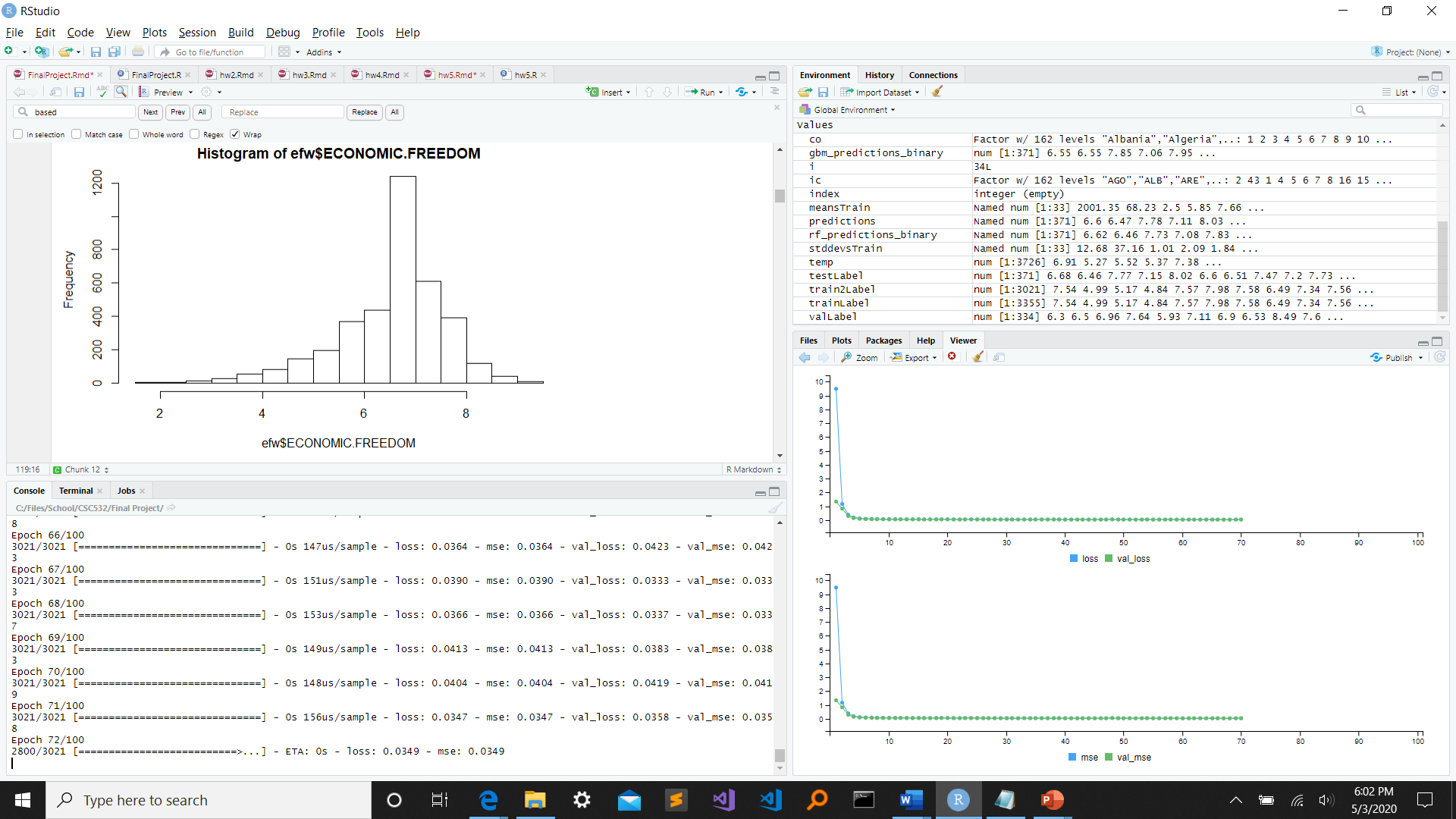
**Data exploration and preprocessing**

First, I determined which columns contained the most missing variables. Among those variables, I wanted to first confirm whether the ones that are over 40% missing are correlated enough with the economic freedom to justify deleting them from the data frame. After using the cross tables for judicial independence, reliability police, and business costs crime, I found that the p value for all three of them are low enough to retain them.

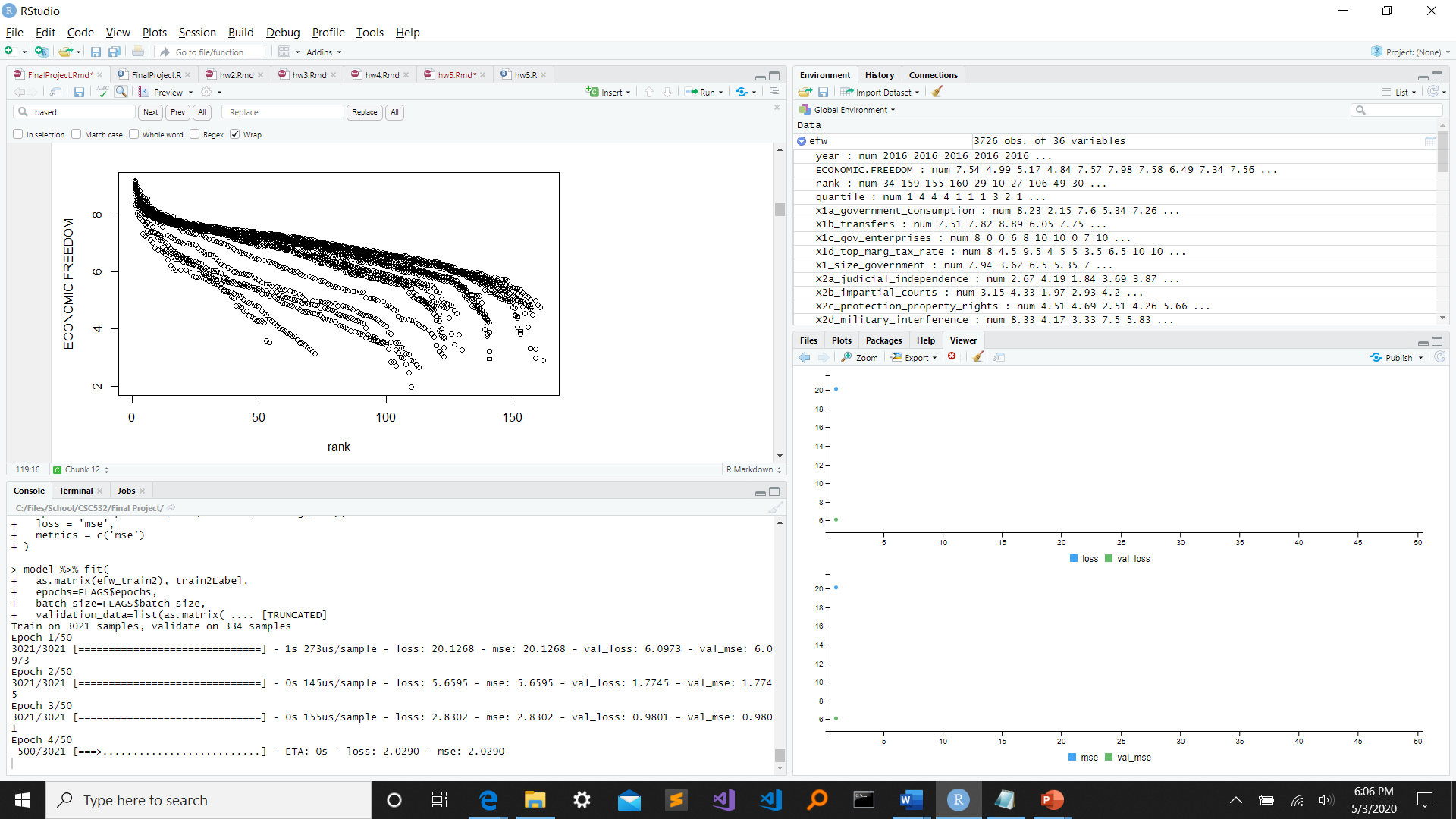


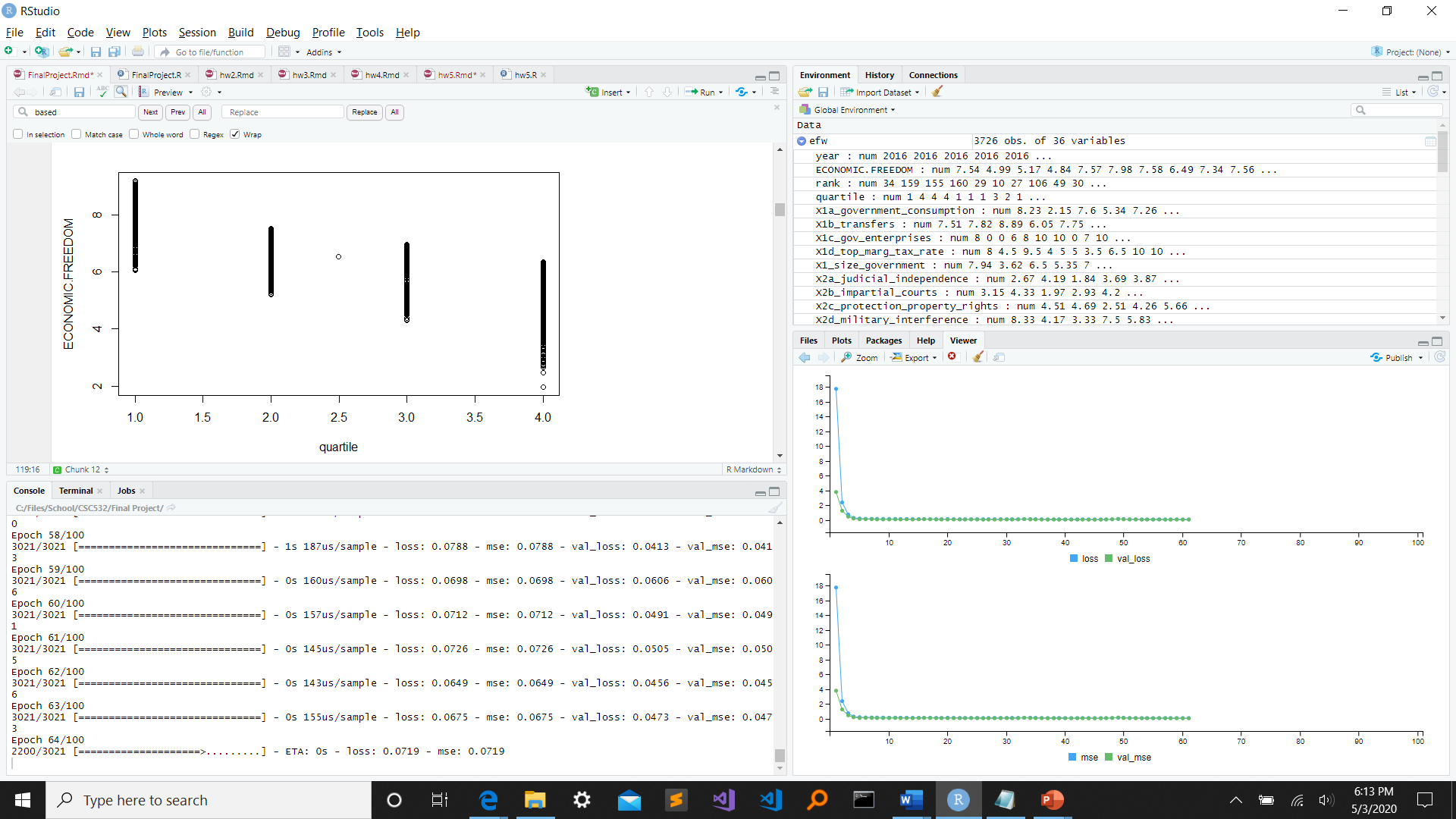
The next step is to handle the missing variables. The categorical columns were all good, so I didn’t have to do anything for them. All the features that needed inputting were numerical, so what I did was iterate through each column in the dataset and replaced rows that contained NA with the mean of the column.

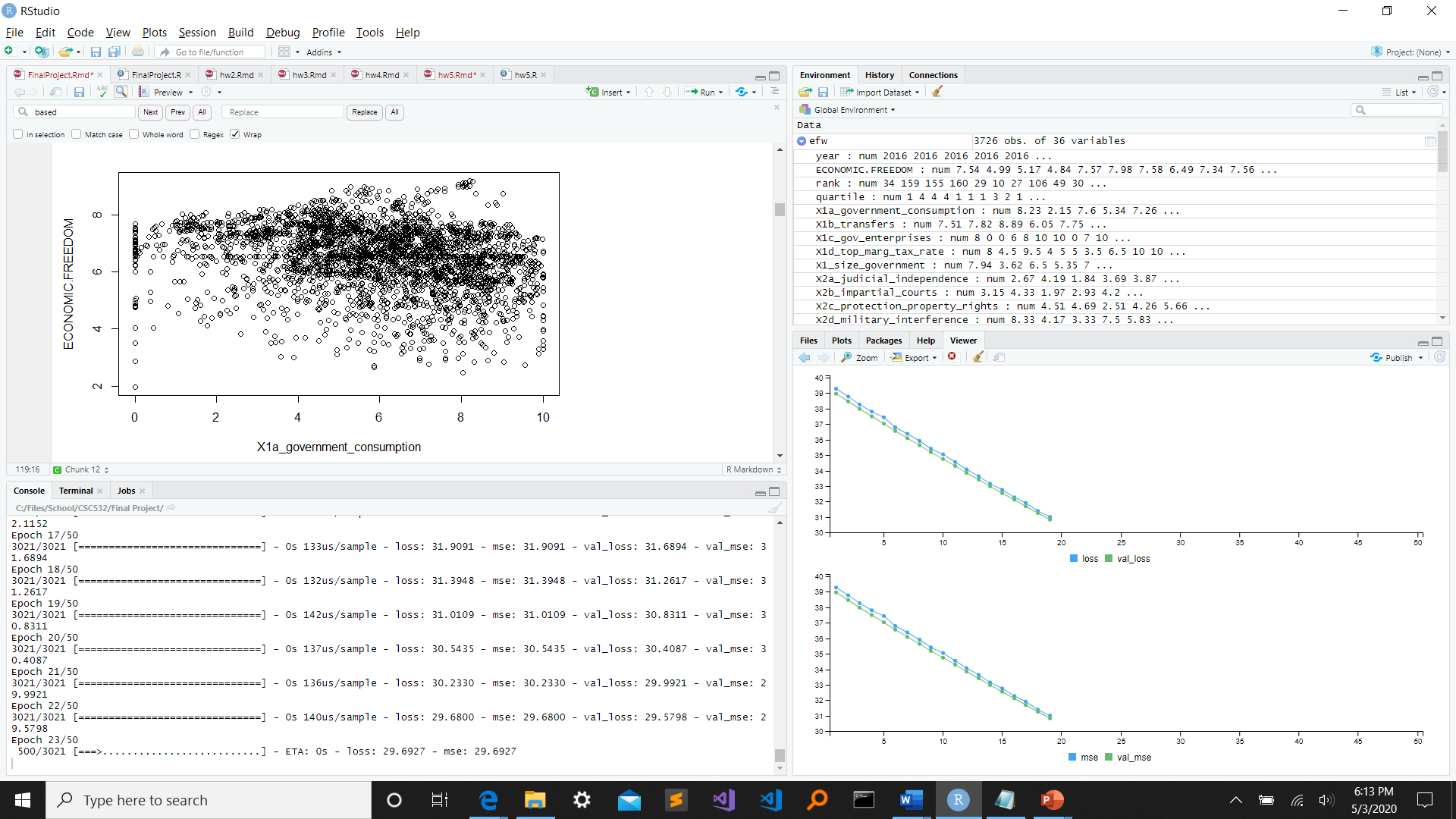
In order to get a good idea of what the economic freedom is throughout the world, I also created a histogram to get a visual of where most countries stack up. Based on the results, it looks like most countries are considered to have decent economic freedom as the graph is left-skewed, suggesting an above average score for most countries.

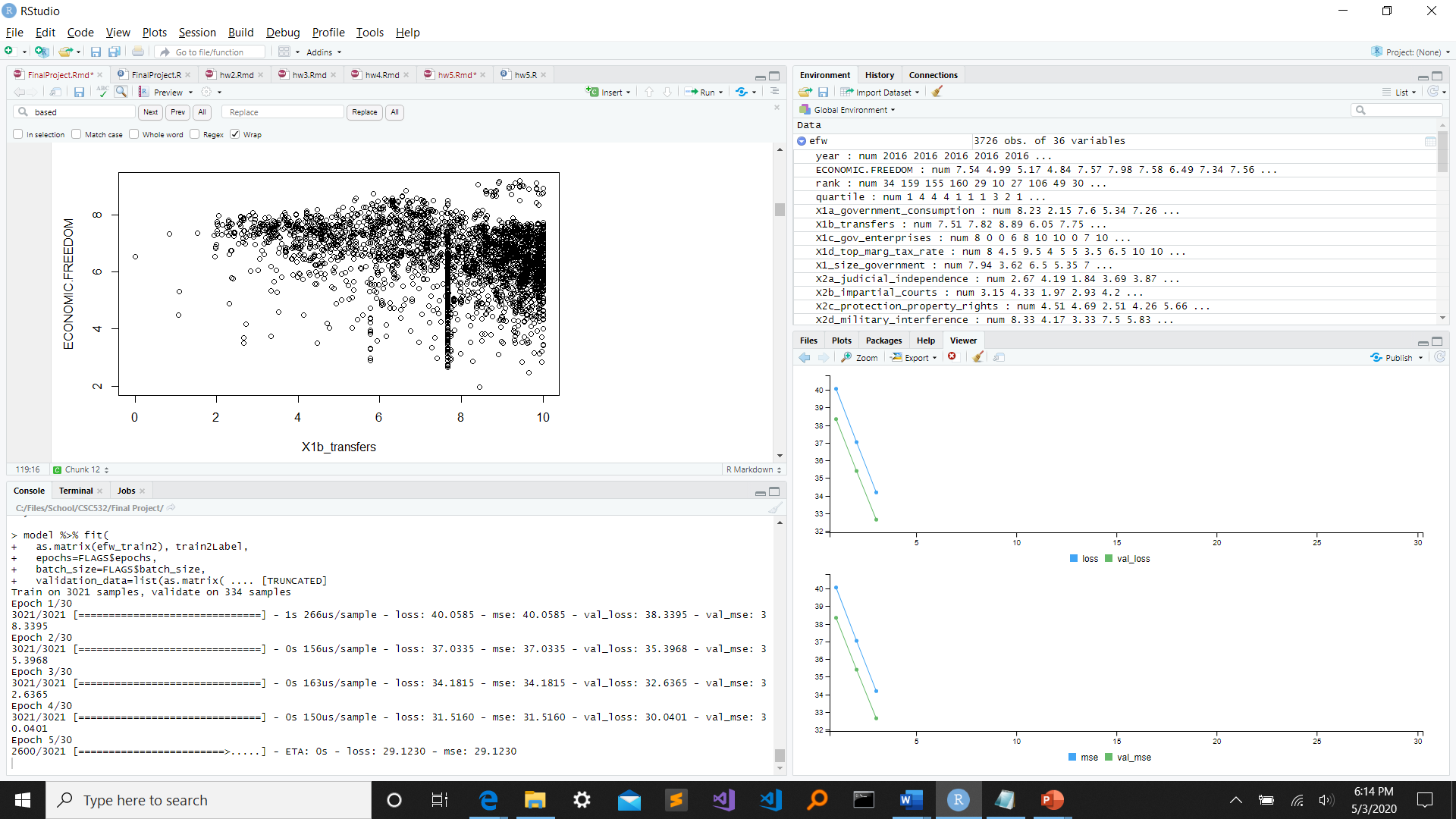


Another data exploration method I utilized were to create the scatterplots and correlation of each variable. Based on quick glance at the output, it looks like rank, quartile, government consumption, and transfers have the lowest correlation with the economic freedom.









**Data analysis and experimental results**

The models I used for the project after processing the data were lasso/ridge linear regression, random forest, GBT, and neural network models. When I compare all of the RMSE of each model, it looks like the neural network did the worst performance, while the rest did fairly similar to each other. The one that did the best slightly was the GBT model.

|  |  |
| --- | --- |
| **Model** | **RMSE** |
| Lasso linear regression | 0.2470609 |
| Ridge linear regression | 0.2685492 |
| Random forest | 0.2097298 |
| GBT | 0.2025671 |
| Neural network | 4.913784 |

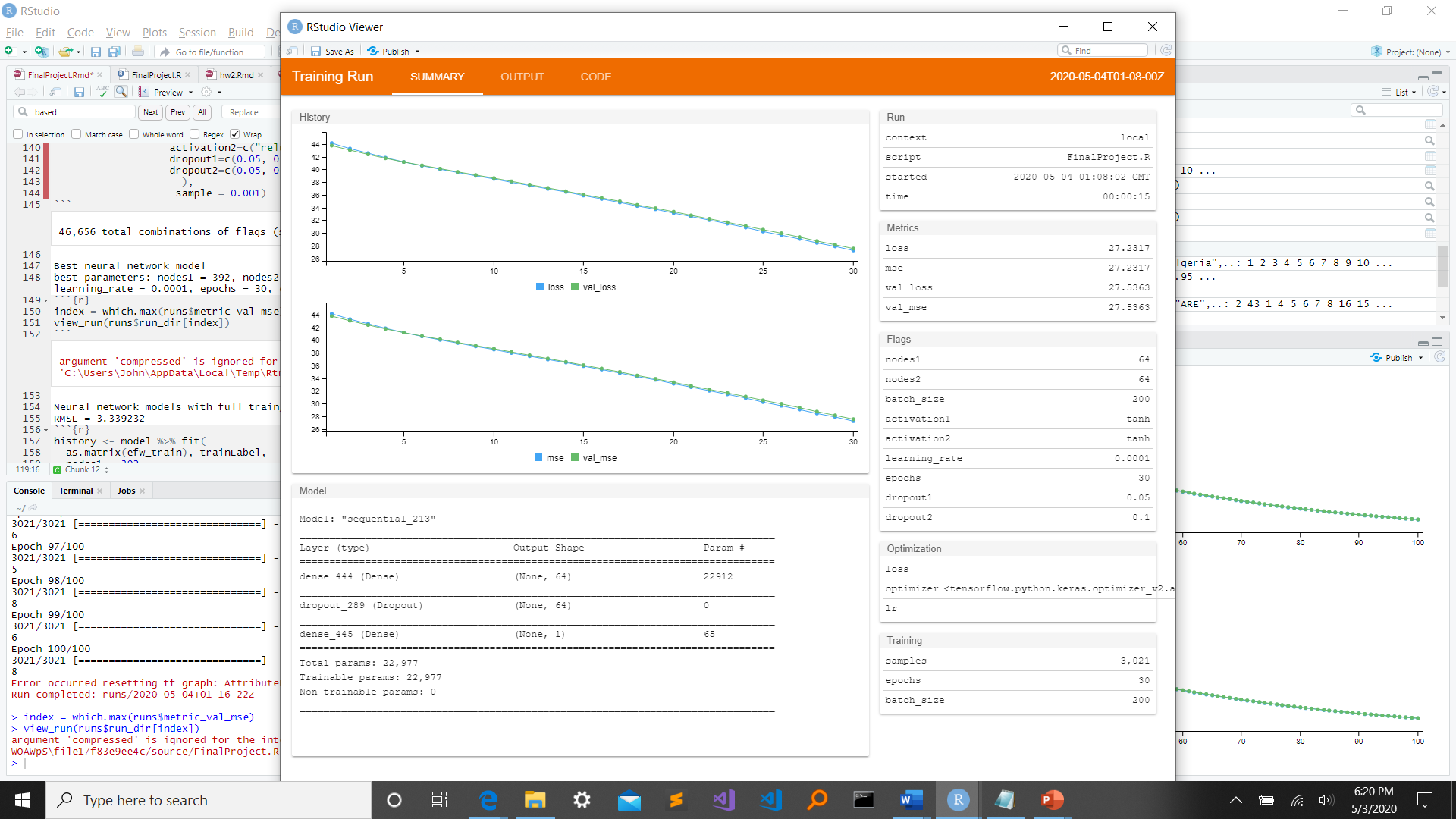
According to the lasso linear regression, these variables were considered the most correlated based on their coefficients.

|  |  |
| --- | --- |
| **Feature** | **Coefficient** |
| Government consumption | 9.917899e-04 |
| Sound money | 9.347317e-03 |
| Rank | 8.743194e-04 |
| Inflation | 5.543481e-02 |
| Money growth | 4.658241e-02 |

For the random forest model, these were the top variables.

|  |  |
| --- | --- |
| **Feature** | **Overall** |
| Regulation | 100 |
| Trade | 89.10 |
| Size government | 87.19 |
| Quartile | 86.33 |
| Control movement capital ppl | 79.36 |
| Rank | 79.25 |
| Sound money | 78.75 |

For the neural network models, I tried passing in a list of parameters and then after all the possible runs were executed, I searched for the most optimal parameter configuration and I got the following: nodes1 = 64, nodes2 = 64, batch\_size = 200, activation1 = “tanh”, activation2 = tanh, learning\_rate = 0.0001, epochs = 30, dropout1 = 0.05, dropout2 = 0.1



**Conclusion**

I was surprised to see that the RMSE of the basic lasso/ridge linear regression models, random forest, and GBT were all around the same level and somewhat low. I was unsure about whether I should’ve kept in some of the features, but even with keeping all of them, the output was rather consistent. The main part that needs improvement is the neural network portion. I’m still trying to figure out how to correctly implement the model and build it to more accurately predict just like the other models. This is where I can go further with research to see how others make their neural networks to improve mine. The dataset itself is interesting since it gives a concrete score for each country how good their economic freedom is. This can then go into the questions of what countries can do to improve their score and how much of the concept is subjective.

**References**

[1] <https://www.kaggle.com/gsutters/economic-freedom>

[2] <https://www.kaggle.com/farazrahman/economic-freedom-top-11-factors>