MGT 6203 Team #64 Final Report

Introduction and Problem Statement

As a result of the COVID-19 pandemic, the global economy has faced many unique challenges over the past few years. With the current economic headwinds being faced by the world (supply chain issues, inflation, etc.), many experts are concerned that the direct consequence of the pandemic is a future recession. However, this is not the first recession the world has faced, and every recession presents a different set of challenges and learning opportunities. Governments will look at diverse ways to mitigate pending economic distresses. There are various factors that contribute to a recession that can affect recovery and can be critical to create policies or strategies to handle the current economic crisis.

As determined by (Kose, 2020), there have been 4 global recessions over the last 60 years (1974, 1981, 1990 and 2007), each lasting one year. The paper defines a global recession as a period of negative global economic growth as well as broad weakness in key indicators of global economic activity. It used a weighted average of GDP and purchase-power-parity around the globe to classify periods of global recession statistically and judicially. Our analysis will focus on these four periods. Each year following the global recession is defined as the recovery year where macroeconomic and financial activities rebound.

The goal of this analysis is to determine which factors from previous recessions will aid us in reducing the negative economic impact of a current global recession. By analyzing various economic, social and political impacts on various countries we can determine the predictors that have the greatest impact on a country, we can use these to determine how it would impact a recession. Directly, the response variables will be GDP Growth Rate (%), GDP and Stability.

We will build on the progress report data to find the best models and predictors to use in our conclusion. Based on the progress report, we can already eliminate many predictors that do not have enough data or will not be useful in prediction. We will use these observations to determine the results on GDP and stability.

Data Gathering

Based on the testing we did in the previous stages of this project, for this approach we chose to use the following predictors for Models 1-4.

- FIW (Democratic Indicator) – Aggregate Rating of political rights, civil liberties and freedom of expression in 13 categories. [2013-2022]
- Vaccine Rate - % of children aged 12-23 months [1960-2020]
- Life Expectancy – at birth [1960-2020]
- Population Growth Rate – Annual % [1960-2020]
- Internet Usage - % of population [1990-2020]
- Political Stability Estimate [1996-2020]
- CPIA Transparency, Accountability and Corruption Index – Rating of 1-6 with 1 being low and 6 being high [1995 - 2020]
- Trade – total merchandise imports annually in millions [1980-2019]
- Unemployment Rates – Total % of labor force unemployed [1991-2021]

- Inflation Rates – Annual Inflation Rate % [1960-2021]
- Interest Rates – Real Interest Rates % [1960-2021]

The remaining indicators that were previously identified were removed due to lack of data. For example, murder had almost all missing data points for most years. Migration rates had missing data for the years we were interested in.

Data Cleaning

Imputation and Country Selection:

For each predictor, the data was cleaned to ensure it was easier to use for analysis purposes.

The life expectancy dataset was used as the reference data set as it had the least missing data.

- The columns were renamed to remain consistent across all the predictors.
- Only data from the recession years 1973, 1974, 1975, 1980, 1981, 1982, 1989, 1990, 1991, 207, 2008 and 2009 were kept.
- The countries that had less than 3 'NA' values across all the years were kept and the rest were removed. This left 198 countries.
- A new predictor called developed was created. Based on some online research, countries that were developed were given a value of 1 and 0 otherwise.

For the rest of the predictors:

- The columns were renamed
- Only data from the recession years 1973, 1974, 1975, 1980, 1981, 1982, 1989, 1990, 1991, 2007, 2008 and 2009 were kept.
- The dataset was merged with life expectancy so only those countries were kept in all the datasets.
- The NA values were imputed with average values from the average values of the developed or non-developed countries. For example, if Aruba was missing data, its value was taken to be the average of the un-developed countries in that same year.

The following predictors were also added to the datasets for analysis:

- A year factor was added to each of the data frames with 1, 2, 3 demonstrating before the recession, during the recession and after the recession
- A difference column was added to the recovery years in the year datasets. This contained the difference between the GDP, population, interest rates and inflation rates from before the recession and after. If the values increased, a 1 was assigned to the country indicating a positive sign from the recession, otherwise a 0 was assigned.

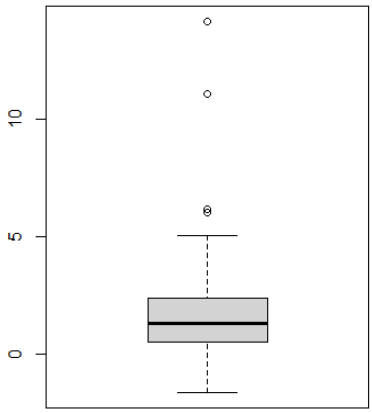
Outliers:

Outliers were removed. A boxplot was done on each of the datasets to visually inspect the data, then a Grubbs test was done to analytically determine the outliers. If the outlier was significant, it was removed. However, since the data had so much random variation due to imputation, not all the outliers could be removed without greatly decreasing the size of the dataset. For this reason, the Grubbs test

and boxplot were used in conjunction with a scatterplot to determine which datapoints were most impactful.

Below, you can see the outlier removal for the 2009 data.

Boxplot Before:



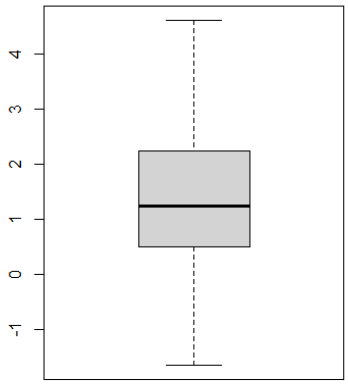
Grubbs Test Before:

```
> grubbs.test(Rec_2009$Population,type = 10, opposite = FALSE, two.sided=TRUE)

        Grubbs test for one outlier

data:  Rec_2009$Population
G = 7.39154, U = 0.72126, p-value = 2.638e-13
alternative hypothesis: highest value 14.14392358 is an outlier
```

Boxplot After:



Grubbs Test After:

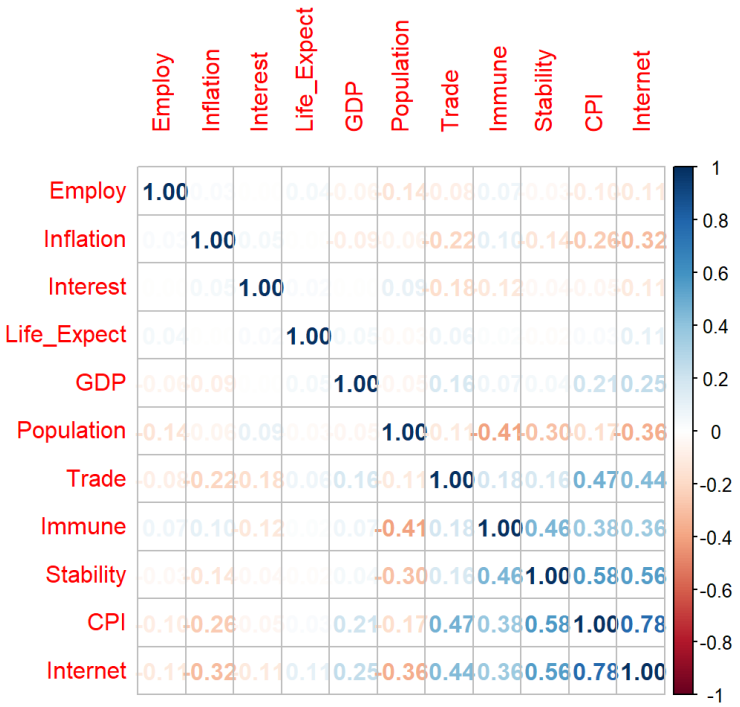
```
> grubbs.test(Rec_2009$Population,type = 10, opposite = FALSE, two.sided=TRUE)

        Grubbs test for one outlier

data:  Rec_2009$Population
G = 2.77134, U = 0.95979, p-value = 0.9982
alternative hypothesis: highest value 4.625126705 is an outlier
```

Correlation:

After outliers were removed, correlation matrices were created for all the data to determine if any should be removed. Below is the correlation matrix for 2009. Only CPI and Internet are moderately correlated. Since they are not strongly correlated, all the predictors are kept in the model.



After the data was cleaned, it left the following data frames for analysis:

a. Year Data:
    1. data_1973, data_1974, data_1975: life expectancy, GDP, population, trade, developed status
    2. data_1980, data_1981, data_1982: life expectancy, GDP, population, trade, developed status, immunization rate
    3. data_1989, data_1990, data_1991: life expectancy, GDP, population, trade, developed status, immunization rate, inflation rate, interest rate
    4. data_2007, data_2008, data_2009: life expectancy, GDP, population, trade, developed status, immunization rate, inflation rate, interest rate, political stability, CPI index, internet usage, unemployment rate
b. Recession Data:
    1. Rec_1: data from 1973-1975
    2. Rec_2: data from 1980-1982
    3. Rec_3: data from 1989-1991
    4. Rec_4: data from 2007-2009
c. data_all: years 1973-2009 with common predictors - life expectancy, GDP, population, trade, developed status

Modelling

We tried various models and modelling techniques to determine the best model to use for future predictions.

Model 1: Linear Regression with no Transformations

Figure 1: Using GDP as a response, this is an example from 1973. Population and trade were significant. Population negatively impacted GDP and Trade positively impacted GDP. However, the R2 value was very low so this model does not have predictive power. 1974, 1975, 1980, 1981 and 1982 showed similar results. 1989-2009 has no significant predictors so it is possible the addition of predictors decreased the predictive power of the model.

Figure 2: This is an example from 2009. Trade, vaccination rates and CPI were significant. Trade was negatively related to Stability. Vaccine and CPI were positively related. The R2 is 0.41. This is better than the model using GDP. 2007 and 2008 had similar results.

```
Call:
lm(formula = GDP ~ ., data = Rec_1973)

Residuals:
       Min         1Q     Median         3Q        Max
-2.733e+10 -6.884e+09  5.327e+08  6.511e+09  1.724e+10

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.257e+10  3.399e+09   3.697 0.000295 ***
Life_Expect -3.179e+07  5.426e+07  -0.586 0.558767
Population  -9.525e+08  4.243e+08  -2.245 0.026071 *
Trade        1.285e+06  3.867e+05   3.322 0.001097 **
Developed1  -2.480e+09  2.434e+09  -1.019 0.309592
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.484e+09 on 168 degrees of freedom
Multiple R-squared:  0.1128,    Adjusted R-squared:  0.09171
F-statistic: 5.342 on 4 and 168 DF,  p-value: 0.0004496
```

Figure 1

```
Call:
lm(formula = Stability ~ ., data = Rec_2009)

Residuals:
     Min       1Q   Median       3Q      Max
-2.16052 -0.37198  0.04922  0.42469  1.92636

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.822e+00  6.876e-01  -2.649  0.00887 **
Employ       3.122e-03  1.009e-02   0.309  0.75749
Inflation   -1.843e-02  1.647e-02  -1.118  0.26506
Interest    -9.831e-03  9.063e-03  -1.085  0.27964
Life_Expect -5.738e-03  6.541e-03  -0.877  0.38164
GDP         -6.683e-14  5.157e-14  -1.296  0.19683
Population  -6.765e-02  6.172e-02  -1.096  0.27470
Trade       -2.919e-06  1.067e-06  -2.735  0.00694 **
Immune       1.577e-02  5.275e-03   2.989  0.00323 **
CPI          2.439e-01  4.899e-02   4.979 1.63e-06 ***
Internet     3.866e-03  4.075e-03   0.949  0.34426
Developed1   2.297e-01  2.257e-01   1.018  0.31032
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7355 on 161 degrees of freedom
Multiple R-squared:  0.4498,    Adjusted R-squared:  0.4123
F-statistic: 11.97 on 11 and 161 DF,  p-value: 3.166e-16
```

Figure 2

Based on the above, linear regression with no transformations is not a good model to be used on this data. It is likely because the relationships in this data are not linear.

Model 2: Linear Regression with Transformations

Figure 3: This is a sample from 1973. The minimum value of Population was added to the log to avoid 0 values. Only log(Trade) was significant and it was positively related to log(GDP). The R2 was only 0.1757. In 1974, 1980 and 1989 log(Population) was also significant. In 1990, log(interest) became significant.

Figure 4: This is a sample from 2009. log(Employ), log(Trade), log(Stability), log(CPI) and log(Internet) are significant. Log(stability) and log(employ) are negatively related to GDP. The R2 value is 0.6432. This is better than Model 1.

```
Call:
lm(formula = log(GDP) ~ log(Life_Expect) + log(Population - b +
    1) + log(Trade) + Developed, data = Rec_1973)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4468 -0.9334  0.3057  0.8847  3.4850

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            21.60446    2.67212   8.085 1.17e-13 ***
log(Life_Expect)       -0.29944    0.60110  -0.498    0.619
log(Population - b + 1) -0.37320    0.51595  -0.723    0.470
log(Trade)              0.42498    0.07256   5.857 2.42e-08 ***
Developed1             -0.62204    0.50225  -1.238    0.217
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.529 on 168 degrees of freedom
Multiple R-squared:  0.1949,    Adjusted R-squared:  0.1757
F-statistic: 10.17 on 4 and 168 DF,  p-value: 2.146e-07
```

Figure 3

```
Call:
lm(formula = log(GDP) ~ log(Employ) + log(Inflation - a + 1) +
    log(Interest - c + 1) + log(Life_Expect) + log(Population -
    b + 1) + log(Trade) + log(Immune) + log(Stability - d + 1) +
    log(CPI) + log(Internet), data = Rec_2009)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4360 -0.6483  0.0422  0.8334  4.5994

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            19.05416    4.46685   4.266 3.38e-05 ***
log(Employ)            -0.30015    0.13873  -2.164  0.03196 *
log(Inflation - a + 1)  0.26954    0.28613   0.942  0.34759
log(Interest - c + 1)  -0.53084    0.32046  -1.656  0.09956 .
log(Life_Expect)        0.45408    0.70887   0.641  0.52271
log(Population - b + 1)  0.17137    0.38087   0.450  0.65334
log(Trade)              0.59643    0.05266  11.326  < 2e-16 ***
log(Immune)             0.15851    0.67550   0.235  0.81477
log(Stability - d + 1) -2.31312    0.48914  -4.729 4.88e-06 ***
log(CPI)                0.86550    0.33794   2.561  0.01135 *
log(Internet)           0.34862    0.11437   3.048  0.00269 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.286 on 162 degrees of freedom
Multiple R-squared:  0.6639,    Adjusted R-squared:  0.6432
F-statistic:    32 on 10 and 162 DF,  p-value: < 2.2e-16
```

Figure 4

Based on the above, trade, stability, CPI, internet, and unemployment may be valuable predictors.

Model 3: Linear Regression with Validation and Transformations

Now that the log transformations have been determined to be more successful than without, we can try using validation to further improve the model.

Figure 5: The sample below is from 1989. log(Population) and log(Trade) are significant and the R2 is 0.5271 for 1989. Population is negatively related and trade is positively related. This is simlar to results from model 1. Only trade was significant for 1973, 1975, 1980, 1981 and 1982. In 1990, developing country status was significant.

Figure 6: log(Trade), log(Stability), log(CPI) and log(Internet) are significant and the R2 is 0.6637. Trade, and internet are positively related. Population is negatively related.

```
Call:
lm(formula = log(GDP) ~ log(Interest - c + 1) + log(Life_Expect) +
    log(Population - b + 1) + log(Trade) + Developed, data = train_1989)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0729 -0.6173  0.1638  0.7427  5.2654

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            26.10017    5.17603   5.043 1.38e-06 ***
log(Interest - c + 1)  -0.93383    0.89350  -1.045  0.29774
log(Life_Expect)       -0.39644    0.90683  -0.437  0.66266
log(Population - b + 1) -1.95902    0.68550  -2.858  0.00491 **
log(Trade)              0.76558    0.06478  11.819  < 2e-16 ***
Developed1              1.00986    0.56540   1.786  0.07622 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.545 on 142 degrees of freedom
Multiple R-squared:  0.5432,    Adjusted R-squared:  0.5271
F-statistic: 33.78 on 5 and 142 DF,  p-value: < 2.2e-16
```

Figure 5

```
Call:
lm(formula = log(GDP) ~ log(Employ) + log(Inflation - a + 1) +
    log(Interest - c + 1) + log(Life_Expect) + log(Population -
    b + 1) + log(Trade) + log(Immune) + log(Stability - d + 1) +
    log(CPI) + log(Internet), data = train_2009)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9300 -0.5440 -0.0076  0.8267  4.4461

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            22.09783    5.60282   3.944 0.000132 ***
log(Employ)            -0.27591    0.15562  -1.773 0.078624 .
log(Inflation - a + 1)  0.30830    0.31128   0.990 0.323859
log(Interest - c + 1)  -0.46027    0.31688  -1.452 0.148829
log(Life_Expect)       -0.04571    0.75074  -0.061 0.951548
log(Population - b + 1)  0.15111    0.39955   0.378 0.705911
log(Trade)              0.61300    0.05728  10.701  < 2e-16 ***
log(Immune)            -0.22915    0.93922  -0.244 0.807639
log(Stability - d + 1) -2.16221    0.56475  -3.829 0.000201 ***
log(CPI)                0.85450    0.37465   2.281 0.024228 *
log(Internet)           0.40984    0.12309   3.330 0.001138 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.241 on 127 degrees of freedom
Multiple R-squared:  0.6887,    Adjusted R-squared:  0.6642
F-statistic:  28.1 on 10 and 127 DF,  p-value: < 2.2e-16
```

Figure 6

Based on the above trade and population may be valuable predictors.

Model 4: Linear Regression with Cross Validation and Transformations

Given the results from models 1-4, we decided only to use 2009 data for this model.

Figure7: For 2009, with GDP as response, log(Internet), log(Trade), log(Stability), log(CPI) and log(Employ) are significant and the R2 is 0.6432. Internet, trade and CPI are positively related while Stability and unemployment are negatively related.

Figure 8: log(CPI), log(Trade), log(Immune) are significant and the R2 is 0.4355. Only trade is negatively related while the rest are positively related.

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4360 -0.6483  0.0422  0.8334  4.5994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.05416    4.46685   4.266 3.38e-05 ***
Internet     0.34862    0.11437   3.048  0.00269 **
Inflation    0.26954    0.28613   0.942  0.34759
Population   0.17137    0.38087   0.450  0.65334
Interest    -0.53084    0.32046  -1.656  0.09956 .
Life_Expect  0.45408    0.70887   0.641  0.52271
Trade        0.59643    0.05266  11.326  < 2e-16 ***
Immune       0.15851    0.67550   0.235  0.81477
Stability   -2.31312    0.48914  -4.729 4.88e-06 ***
Employ      -0.30015    0.13873  -2.164  0.03196 *
CPI          0.86550    0.33794   2.561  0.01135 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.286 on 162 degrees of freedom
Multiple R-squared:  0.6639,    Adjusted R-squared:  0.6432
F-statistic:    32 on 10 and 162 DF,  p-value: < 2.2e-16
```

Figure 7

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66050 -0.10513  0.02575  0.13927  0.63859

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.625869   0.970938   0.645 0.520338
Population  -0.082371   0.080863  -1.019 0.310289
Interest    -0.024095   0.148775  -0.162 0.871598
Employ       0.001677   0.026027   0.064 0.948727
Life_Expect -0.124157   0.135117  -0.919 0.359887
Trade       -0.037900   0.009871  -3.840 0.000193 ***
Immune       0.292701   0.120206   2.435 0.016270 *
CPI          0.318505   0.055570   5.732 6.77e-08 ***
Internet     0.030888   0.023106   1.337 0.183652
Inflation    0.022187   0.054792   0.405 0.686199
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2175 on 128 degrees of freedom
Multiple R-squared:  0.4726,    Adjusted R-squared:  0.4355
F-statistic: 12.74 on 9 and 128 DF,  p-value: 2.591e-14
```

Figure 8

Based on the above trade, stability and CPI may be valuable predictors.

Model 5: Linear Regression with Trade Data

The goal of this model was to build upon the linear regression models that we created for our progress report where we were examining the relationship between each country's value of merchandise exports and their GDP growth rate during each recession time frame. However, this time we separated the countries into lists by population size to see if countries with larger populations would have similar or different trends to countries with smaller populations. Additionally, we ran these models twice: the first time against the data that we collected from the recession years and the second time against the data that we collected from the recession recovery years. Last, we added inflation rates as a second predictor variable. In summary, the output that we obtained from these models provided us with a better high-level understanding of trade on a country's performance during recessions and in the recovery phases, specifically by examining the effect of inflation rates and values of merchandise exports on the annual GDP growth rates of our countries by population size.

```
Call:
lm(formula = GDP_Rate ~ Inflation_Rate + Merchandise_Exports,
    data = recession[recession$Country %in% c("China", "India"),
        ])

Residuals:
     10      15      36      41      62      67      88      93
-1.7724 -2.1627  1.0593  2.3835 -0.1442  1.8647  0.1559 -1.3840

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.066e+00  1.393e+00   2.919   0.0331 *
Inflation_Rate     -4.402e-02  1.461e-01  -0.301   0.7753
Merchandise_Exports 4.035e-06  1.533e-06   2.633   0.0464 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.003 on 5 degrees of freedom
Multiple R-squared:  0.5845,    Adjusted R-squared:  0.4183
F-statistic: 3.517 on 2 and 5 DF,  p-value: 0.1113
```

Figure 9

```
Call:
lm(formula = GDP_Rate ~ Inflation_Rate + Merchandise_Exports,
    data = recession[recession$Country %in% c("United States",
        "Brazil", "Mexico", "Philippines"), ])

Residuals:
    Min      1Q  Median      3Q     Max
-7.7188 -0.6593 -0.0913  1.6201  4.8068

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.878e+00  9.575e-01   4.051  0.00137 **
Inflation_Rate     -3.078e-03  1.246e-03  -2.471  0.02809 *
Merchandise_Exports -3.417e-06  2.595e-06  -1.316  0.21077
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.208 on 13 degrees of freedom
Multiple R-squared:  0.3541,    Adjusted R-squared:  0.2548
F-statistic: 3.564 on 2 and 13 DF,  p-value: 0.05834
```

Figure 10

Looking at the output from the recession years-only data in Figure 9, for countries with populations over 1 billion people, inflation rates decreased and values of merchandise exports increased as GDP growth rates increased. We see in Figure 10 the same relationship for countries with populations between 100 million and 320 million people. During recovery years, our outputs indicate that the same trend occurred for countries with populations over 1 billion people (Figure 11). However, during recovery years for countries with populations between 100 million and 320 million people, we see that both inflation rates and the value of merchandise exports trended downward as GDP growth rates increased (Figure 12). We can see that the relationship between these variables appears to continue in the same direction as countries decrease in population size, as is noted in Figures 13 and 14 (output of the models for countries with less than 100 million people).

```
Call:
lm(formula = GDP_Rate ~ Inflation_Rate + Merchandise_Exports,
    data = recovery[recovery$Country %in% c("China", "India"),
        ])

Residuals:
      10       15       36       41       62       67       88       93
-0.796239 -0.606944 -0.007954 -1.714945  3.371728 -1.486884 -0.381790  1.623028

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.981e+00  1.031e+00   8.714  0.00033 ***
Inflation_Rate     -4.689e-01  1.426e-01  -3.288  0.02176 *
Merchandise_Exports  6.130e-07  1.913e-06   0.320  0.76156
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.015 on 5 degrees of freedom
Multiple R-squared:  0.7122,    Adjusted R-squared:  0.597
F-statistic: 6.185 on 2 and 5 DF,  p-value: 0.04445
```

Figure 11

```
Call:
lm(formula = GDP_Rate ~ Inflation_Rate + Merchandise_Exports,
    data = recession[recession$Country %in% c("Vietnam", "Germany",
        "France", "United Kingdom", "Italy", "Spain", "Argentina",
        "Canada", "Morocco", "Peru", "Afghanistan", "Angola",
        "Australia", "Ecuador", "Cambodia", "Cuba", "Israel",
        "Norway", "Barbados", "Aruba"), ])

Residuals:
    Min      1Q  Median      3Q     Max
-8.6313 -2.3432  0.0776  2.0047 16.0656

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.640e+00  5.911e-01   6.158 8.94e-08 ***
Inflation_Rate      -1.533e-03  6.128e-04  -2.502   0.0153 *
Merchandise_Exports -3.830e-06  2.292e-06  -1.671   0.1004
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.963 on 55 degrees of freedom
  (22 observations deleted due to missingness)
Multiple R-squared:  0.1327,    Adjusted R-squared:  0.1012
F-statistic: 4.209 on 2 and 55 DF,  p-value: 0.01992
```

Figure 13

```
Call:
lm(formula = GDP_Rate ~ Inflation_Rate + Merchandise_Exports,
    data = recovery[recovery$Country %in% c("United States",
        "Brazil", "Mexico", "Philippines"), ])

Residuals:
    Min      1Q  Median      3Q     Max
-5.9066 -1.7427 -0.1153  2.0410  3.7457

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.053e+00  9.204e-01   2.230   0.0440 *
Inflation_Rate      -2.302e-03  7.370e-03  -0.312   0.7597
Merchandise_Exports -6.198e-06  2.802e-06  -2.212   0.0455 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.857 on 13 degrees of freedom
Multiple R-squared:  0.2741,    Adjusted R-squared:  0.1625
F-statistic: 2.455 on 2 and 13 DF,  p-value: 0.1246
```

Figure 12

```
Call:
lm(formula = GDP_Rate ~ Inflation_Rate + Merchandise_Exports,
    data = recovery[recovery$Country %in% c("Vietnam", "Germany",
        "France", "United Kingdom", "Italy", "Spain", "Argentina",
        "Canada", "Morocco", "Peru", "Afghanistan", "Angola",
        "Australia", "Ecuador", "Cambodia", "Cuba", "Israel",
        "Norway", "Barbados", "Aruba"), ])

Residuals:
     Min       1Q   Median       3Q      Max
-13.7988  -2.1762  -0.6974   2.1842  19.2615

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.132e+00  8.100e-01   2.633   0.0109 *
Inflation_Rate      -1.130e-04  1.029e-02  -0.011   0.9913
Merchandise_Exports -9.125e-06  3.796e-06  -2.404   0.0196 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.083 on 56 degrees of freedom
  (21 observations deleted due to missingness)
Multiple R-squared:  0.09583,   Adjusted R-squared:  0.06354
F-statistic: 2.968 on 2 and 56 DF,  p-value: 0.05957
```

Figure 14

Model 6: Logistic Regression

Figure 15: Using the recession predictor as the response, there were no significant predictors in 1975 and the AUC was 0.3677. This means the model is not good at showing the variation in data.

Figure 16: Using the recession predictor as the response, unemployment, inflation, stability and the internet were related to recession success or failure. The AUC was 0.8144. This means the model is better than average and good at showing the variation in data. AIC is 181.593. This is lower than the linear regression model so it could be considered better.

```
Call:
glm(formula = Recession ~ Life_Expect + Population + Trade, family = "binomial",
    data = train_full)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6751   0.2549   0.2836   0.3100   0.5373

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.629e+00  2.360e+00   1.961   0.0499 *
Life_Expect -1.669e-02  3.808e-02  -0.438   0.6612
Population  -1.978e-01  1.322e-01  -1.496   0.1346
Trade       -1.041e-05  3.837e-05  -0.271   0.7862
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 57.319  on 157  degrees of freedom
Residual deviance: 55.071  on 154  degrees of freedom
AIC: 63.071

Number of Fisher Scoring iterations: 6
```

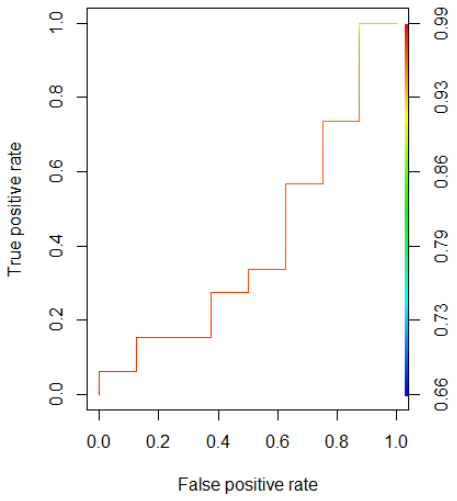


Figure 15

```
Call:
glm(formula = Recession ~ Life_Expect + Population + Employ +
    Inflation + Interest + Trade + Stability + CPI + Internet,
    family = "binomial", data = data_2009)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4989   0.1538   0.3517   0.6509   1.7603

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.257e+00  1.770e+00   2.405  0.01619 *
Life_Expect  1.511e-03  2.247e-02   0.067  0.94639
Population  -4.008e-02  1.156e-01  -0.347  0.72882
Employ      -1.176e-01  3.728e-02  -3.155  0.00161 **
Inflation   -1.093e-01  4.983e-02  -2.194  0.02822 *
Interest    -2.401e-02  2.119e-02  -1.133  0.25712
Trade       -1.129e-06  1.443e-06  -0.782  0.43396
Stability   -6.491e-01  3.263e-01  -1.989  0.04666 *
CPI          9.721e-02  1.580e-01   0.615  0.53835
Internet    -4.232e-02  1.353e-02  -3.127  0.00176 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 207.23  on 197  degrees of freedom
Residual deviance: 161.59  on 188  degrees of freedom
AIC: 181.59

Number of Fisher Scoring iterations: 5
```

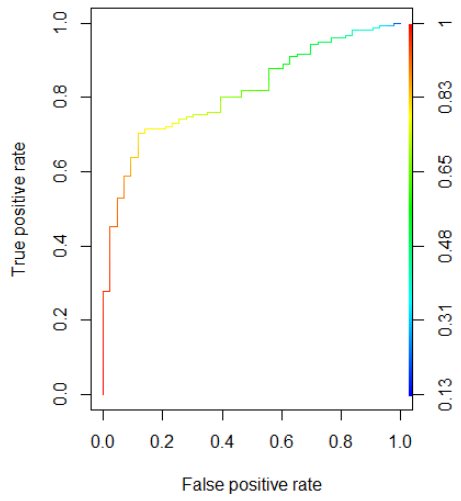


Figure 16

In summary, from models 1-5, stability, trade, population and unemployment predictors could be useful for future prediction. Logistic regression seems to be the best kind of model to use for this data as it is not linear.

Model 7: Logistic Regression using aggregate data

For this model, we used Aggregate data from all the Recession Years (1973, 1974, 1980, 1981, 1989, 1990, 2007, 2008) and the Recovery Years (1975, 1982, 1991, 2009). We combined some of the economic and social variables. A binary variable was added to indicate years with recessions and years with no recessions. The variables used were:

Response Variable	Predicting Variables
1. Y = Recession (Binary Variable)	1. GDP Growth Rate
	2. Inflation Rate
	3. Population Growth
	4. Fuel Exports
	5. Life Expectancy
	6. Merchandise Exports

We used data from the 26 previously selected countries:

> Afghanistan, Angola, Argentina, Aruba, Australia, Barbados, Brazil, Cambodia, Canada, China, Cuba, Ecuador, France, Germany, India, Israel, Italy, Mexico, Morocco, Norway, Peru, Philippines, Spain, United Kingdom, United States, Vietnam

70% of the data was used to train the model and 30% to test it. We used the MICE package to impute the missing values. Since all the variables were numeric, the imputation method used was Predictive Mean Matching with up to 40 iterations.

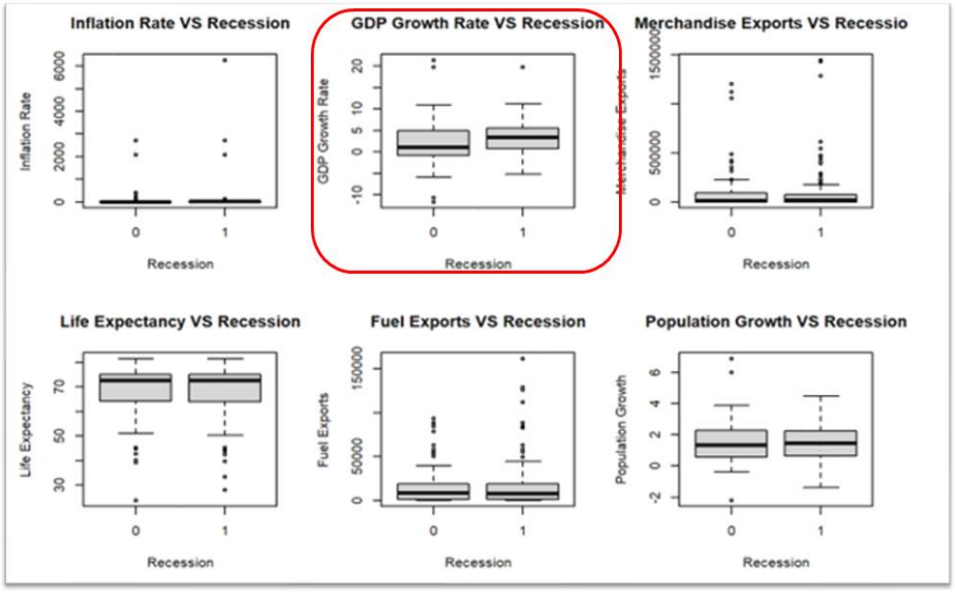


Figure 17

Figure 17 above shows the relationship between the predicting and the response variables. As we can see, there isn't a clear separation between Recession and Non-recession years for any of the variables

except for GDP Growth Rate. We can also see that most of the variables have some outliers which were removed using the Grubbs' Test.

After removing the outliers and fitting the logistic regression model, the output (Figure 18) points out that the only statistically significant coefficient is GDP Growth rate. This was expected after the box-plot analysis. The test for the overall regression shows that the p-value is large (0.14), therefore we can conclude that this model has no predictive power.

```
Call:
glm(formula = Recession ~ Life_expectancy + Population_Growth +
    ï..GDP_Rate + Fuel_Exports + Inflation_Rate + Merchandise_Exports,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6903  -1.0300  -0.3708   1.1599   1.8029

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         5.428e-01  3.274e+00   0.166   0.8683
Life_expectancy    -1.867e-02  4.232e-02  -0.441   0.6591
Population_Growth  -4.061e-03  3.549e-01  -0.011   0.9909
ï..GDP_Rate         1.941e-01  8.334e-02   2.329   0.0199 *
Fuel_Exports        5.329e-06  8.511e-06   0.626   0.5312
Inflation_Rate      1.246e-03  1.126e-03   1.107   0.2682
Merchandise_Exports 1.199e-06  1.051e-06   1.141   0.2538
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
      gstat
[1,] 9.479194 0.1483676
```

Figure 18

We can confirm this by running this model using the test data. In Figure 19 we can see the confusion matrix. The accuracy of this model is about 61% so the results weren't very good. The reason might be that for some of the variables over 25% of the values were missing and had to be imputed.

```
          Reference
Prediction  0  1
         0 22  2
         1 18  9

               Accuracy : 0.6078
                 95% CI : (0.4611, 0.7416)
    No Information Rate : 0.7843
    P-Value [Acc > NIR] : 0.9987476

                  Kappa : 0.2411

 Mcnemar's Test P-Value : 0.0007962
```

Figure 19

Model 8: Regression Tree

Usage of a tree model could allow recognition of patterns under different regimes. By dividing all the data into two datasets, Recession and Recovery, the goal of the tree model is to discover which thresholds of different predictors could act as differentiators between countries. Data is aggregated for all countries for all recession years and by predictors.

Prior to fitting a tree model, a straight linear regression model is fitted for both datasets. This model will provide a baseline accuracy to which we can compare the tree model. The outputs for each model are:

- Recession dataset: $R^2 = 0.28$
- Recovery dataset: $R^2 = 0.38$

```
Call:
lm(formula = GDP_Rate ~ ., data = recession)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3928 -1.9122  0.0053  1.8519  6.4071

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -8.74201    8.35130  -1.047   0.3009
Life_expectancy    0.09142    0.07218   1.267   0.2120
Murder            -0.14753    0.07791  -1.894   0.0649 .
Population_Growth  1.10370    0.57025   1.935   0.0594 .
Inflation_Rate     0.12988    0.05664   2.293   0.0267 *
Tot_Export         1.31128    3.36144   0.390   0.6983
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.954 on 44 degrees of freedom
Multiple R-squared:  0.2795,    Adjusted R-squared:  0.1976
F-statistic: 3.413 on 5 and 44 DF,  p-value: 0.01083
```

Figure 20

```
Call:
lm(formula = GDP_Rate ~ ., data = recovery)

Residuals:
     Min       1Q   Median       3Q      Max
-12.3418  -3.1461  -0.1542   2.6950  11.8931

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       14.5530819 10.2453830   1.420 0.162524
Life_expectancy   -0.1385636  0.1275447  -1.086 0.283222
Murder            -0.3080922  0.1386062  -2.223 0.031416 *
Population_Growth  2.7384900  0.7291839   3.756 0.000504 ***
Inflation_Rate     0.0002523  0.0101944   0.025 0.980370
Tot_Export        -0.3984818  0.5412961  -0.736 0.465539
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.391 on 44 degrees of freedom
Multiple R-squared:  0.3819,    Adjusted R-squared:  0.3117
F-statistic: 5.438 on 5 and 44 DF,  p-value: 0.0005581
```

Figure 21

The first tree model fitted uses the Recession years dataset and the tree is pruned. The tree model uses Inflation Rate and Murder Rate as differentiators. The R^2 for the Recession Regression Tree is approximately 0.41, a substantial jump compared to the baseline.

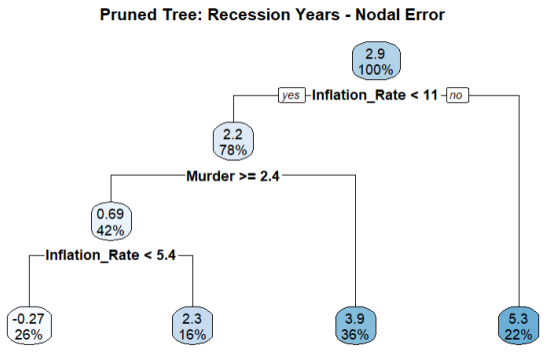


Figure 22

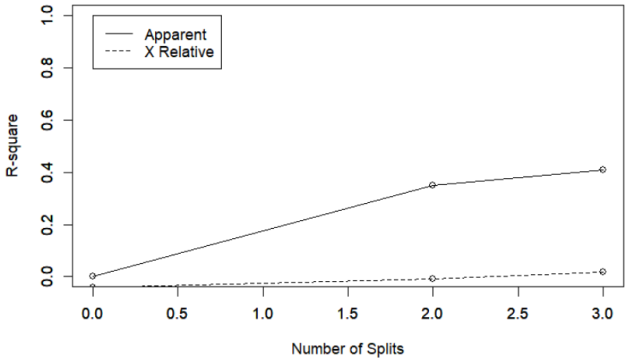


Figure 23

The second tree model was fitted for the Recovery years dataset. In contrast to the first model, this tree model uses Murder Rate and Population Growth Rate as differentiators. The R^2 for the Recovery Regression Tree is approximately 0.42, which is a slight improvement compared to the baseline.

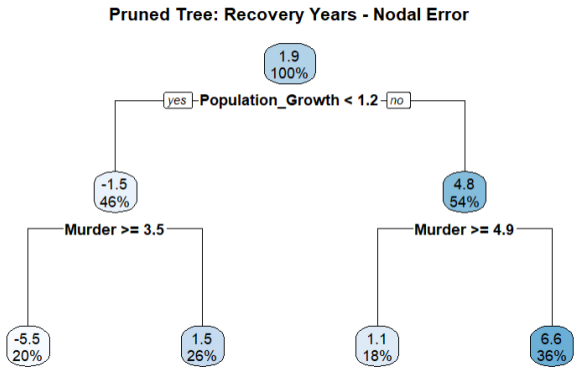


Figure 24

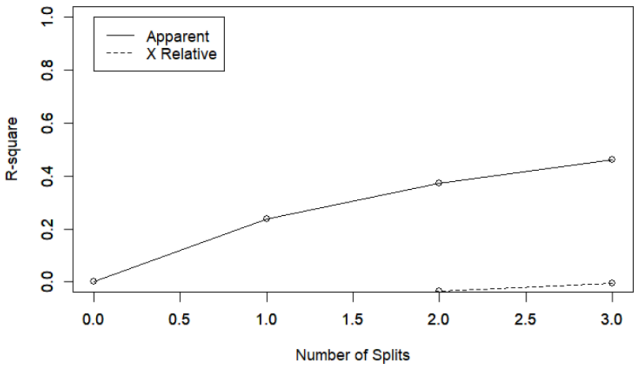


Figure 25

Comparing the trees to the baseline, it is interesting to note that Population Growth, Murder Rate, and Inflation Rate are deemed to be significant predictors that can determine the the response of a country to recession and its recovery after. While the R^2 value for the Recession dataset did indeed see an increas

Conclusion

The different models were definitively able to predict the likelihood of recovery from a recession or GDP Growth Rate with great degree of certainty. However, all models identified similar statistically significant predictors. Based on these models, the most useful predictors are:

- Trade
- Population Growth Rate
- GDP Growth Rate
- Inflation Rate
- Stability Rate

Some Important predictors that require more data are:

- Interest Rate
- Social Data (Crime, CPI Index)
- Unemployment rate

We would recommend that governments use these as a starting point to help limit the negative effects of a potential recession. However, there needs to be more work done on these datasets to get better results from them.

Future Improvements

Overall, the models did not perform as well as they could have. There could be a few reasons for this. The first is regarding the amount of missing data. There was a large quantity of missing data in various years for the recessions. The imputations will not be as accurate as the actual data. Also, there is fewer missing data for more recent years. If we had used more recent data, this problem may have been less prevalent.

The next reason is using more data analysis techniques. We might consider using different models and techniques such as PCA, and variable selection to better determine which predictors are valuable.

References

(n.d.). Retrieved from Our World in Data: https://ourworldindata.org/

Carlson, B. (1999, August). *Social Dimensions of Economic Development and Productivity: Inequality and Social Performance.* Retrieved from https://repositorio.cepal.org/bitstream/handle/11362/4659/S99117_en.pdf

Corruption Perceptions Index. (n.d.). Retrieved from Transparency International: https://www.transparency.org/en/cpi/2020

Fornari, F. (2010, October). *Predicting Recession Probabilities with Financial Variables Over Multiple Horizons.* Retrieved from European Central Bank: https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1255.pdf

Freedom in the World. (n.d.). Retrieved from Freedom House: https://freedomhouse.org/report/freedom-world

Kose, A. M. (2020, March). *Global Recessions.* Retrieved from https://documents1.worldbank.org/curated/en/185391583249079464/pdf/Global-Recessions.pdf

UIS.Stat. (n.d.). Retrieved from UNESCO Institute for Statistics: http://data.uis.unesco.org/

World Bank Open Data. (n.d.). Retrieved from The World Bank: https://data.worldbank.org/