Refugee-Induced Healthcare Expenditure Modeling:

Applied Statistics Capstone
Technical Report
Spring 2024

Table of Contents

Abstract

This capstone project examines the impact of refugee inflows on healthcare expenditures in Turkey, Germany, and Canada. Utilizing separate datasets for each country, the study employs Vector Autoregression (VAR) models for Germany and Canada, and Weighted Linear Regression (WLR) for Turkey, to analyze how refugee populations influence healthcare spending within their host countries. Utilizing both SAS Studio and R, models with significant predictors and high adjusted R-squares were created and validated. Results vary across countries; healthcare expenditure and city-specific GDPs are not consistently associated with refugee count. The project aims to inform both national and international policymakers involved in health and migration by identifying key determinants of healthcare costs, thus assisting in resource allocation and economic planning for countries facing significant refugee influxes.

Project Overview

Motivation

The initial motivation for our capstone project stems from a global concern: the rapidly growing number of refugees and asylum-seekers, and the impact these groups can have on host countries, particularly on their healthcare systems. Global conflicts, climate change, and political instability, among other things, have exacerbated ongoing refugee crises and contributed to a surge in displaced people. Additionally, refugees and migrants often face poorer health outcomes than host populations and must contend with additional cultural and linguistic barriers, diminishing the quality of healthcare they receive (1). Thus, we were motivated to understand the financial and operational strains that these vulnerable populations may impose on national healthcare infrastructure to ensure more equitable systems. By better quantifying the impact of refugees on healthcare systems, healthcare actors such as policymakers, healthcare providers, and non-governmental organizations can tailor their services to meet the unique needs of these populations, ensuring that both refugee populations and native residents receive necessary healthcare support.

Original Research Question

Our original research question aimed to quantitatively assess whether refugees and asylum-seekers place a burden on host countries' healthcare systems, exploring not only the economic impact but also non-economic dimensions such as healthcare accessibility, quality of care, and infrastructure strain. During data collection, we experienced difficulty finding comprehensive and uniform data across various countries that could accurately capture all the dimensions of our initial research question. While sourcing refugee data and healthcare expenditures for individual countries was trivial, we found it challenging to quantify healthcare burdens given the lack of international data standards and scarcity of relevant data. As a result, we chose to use predictors like the number of hospital beds or vaccination rates within each country to indirectly assess the efficacy of their healthcare systems. This data was often hard to verify or not collected across many countries, making it unsuitable for broadly modeling refugee effects on healthcare systems.

Another significant problem was the variations in the healthcare systems across different countries. Inherent differences in healthcare frameworks, geography, and society as a whole significantly affect both the availability of data and how refugees interact with these systems. Given these factors, we decided that a wide-range analysis would not accurately capture the nuances within each country's healthcare system, leading us to tailor our research question to more specific contexts.

Final Research Question

Considering the lack of reliable, standardized data, we opted to shift our research to a case-by-case approach, focusing on specific countries using the datasets and analytical techniques most appropriate for each context. Although potentially limiting the generalizability of our findings, this approach enables us to conduct a more detailed and precise analysis of how refugee flow impacts healthcare expenditure, which ultimately is more important for local policy makers.

The revised research question, "How does refugee flow impact healthcare expenditures in cities or countries, taking into account the total number of refugees or refugees per capita, total number of hospital beds available, GDP per capita, age distribution, and/or vaccination rate?" reflects a more focused inquiry. This shift was influenced by the availability of more detailed and reliable city-level or time-series data for certain countries, which allowed for a more granular analysis of the interactions between refugee populations and specific healthcare system metrics.

With this in mind, our project focuses on 3 countries: Turkey, Germany, and Canada. Turkey was chosen as a target of our analysis due to its proximity to a large humanitarian crisis and its uniquely open refugee policy that has seen the intake of the most refugees worldwide (2). In contrast, Germany and Canada employ a more regulated, bureaucratic approach, allowing for analysis on both ends of the spectrum and across continents. While the healthcare frameworks may vary across these countries, all three aim to ensure nationwide access to healthcare via a mix of public and private providers. By evaluating these countries in particular, we are able to investigate the impact of different scales of refugee inflow across diverse healthcare landscapes.

Data

Our case-by-case approach necessitates separate analyses for each country. Therefore, three individual datasets were compiled. These datasets contain information on each country's healthcare expenditure, refugee flows, and other relevant markers such as GDP or health indicators such as the number of hospital beds in the region. The data descriptions for each country can be seen below and the data dictionaries can be found in the appendix.

Germany

Germany's dataset is a time series spanning from 1970 to 2022 (52 years). This dataset has 3 predictors: the country's overall healthcare expenditure (in millions of euros), the refugee population, and the total GDP (in millions of euros). Since data was collected annually, the time series has 53 observations.

We obtained Germany's GDP and health expenditure data from the Organization for Economic Co-operation and Development, an international organization with rigorous data standards across its 38 member countries. The remaining refugee data was retrieved from the United Nations High Commissioner for Refugees website, which tracks refugee data and trends around the world.

Canada

The Canada dataset is another time series that ranges from 1976 to 2022 (46 years). This dataset combines Canada's annual GDP, healthcare expenditures, the number of hospital beds per capita, and the number of total immigrants and asylum seekers. The GDP and healthcare expenditure are measured in billions of USD and there are a total of 6 variables in this dataset with 47 observations.

The Canada dataset was sourced from a conglomeration of different sources, both from the national government and reliable non-governmental organizations. Data was obtained from several sources, including Statistics Canada (Canada's national statistical agency), the World Bank (a worldwide institution), the Canadian Institute for Health Information, and a reliable third-party website called MacroTrends, which aggregates and presents statistical information from other reliable sources. These websites were chosen due to their proven reliability and the range of the data, which spanned from 1976 to 2022. This made the data very easy to combine into one large dataset.

Turkey

The Turkey dataset is not a time series dataset but consolidates information from each of Turkey's 81 major cities in 2021 or 2022. Each row in the dataset corresponds to a city, with the variables of total number of refugees, vaccination rates, public hospitals, total hospitals, Syrian refugees as a percentage of the city, and GDP per capita (USD) residents in the city. As nearly all refugees in Turkey are from Syria, this specific refugee population is utilized. These are the five variables in this dataset, with 81 total observations. This dataset does not include a city-level healthcare expenditure component due to the unavailability of this data; instead, it utilizes GDP per capita to act as this metric. GDP per capita can provide insight into the overall economic health of a city, so we will use this variable to analyze the economic effect of refugees.

This data was collected from three main sources: the Refugees and Asylum Seekers Assistance and Solidarity Association of Turkey, which provides integrated social services to refugees and the local populations; the Turkish Ministry of Health, which is the government ministry responsible for proposing and executing the government policy on health; and the

Turkish Statistical Institute, which is the Turkish government agency commissioned for producing official statistics on Turkey. All three data sources were chosen due to their reputation as accurate and reliable data depositories, as well as the associated ease of downloading data from their websites.

Data Limitations

The unique characteristics and availability of data in our 3 countries of interest posed several challenges when it came to data collection. Our original intention was to create data tables based on each country's provinces/subregions, using those units as individual observations to inform a country-level healthcare expenditure conclusion. However, several obstacles prevented us from doing so. In particular, Germany and Canada have a very small number of administrative regions, 16 and 13, respectively, which quickly made the lack of observations an evident problem. The alternative was to obtain smaller, city-level data within each of those regions, which proved difficult to collect due to the scarcity of available datasets. In contrast, we found that Turkey had far more readily available city-level data than Canada and Germany combined, so our main concern was instead the overall lack of reliability. This could be due to a combination of the sheer magnitude of the refugee crisis there and data fragmentation in which data might be scattered across various governmental departments, international organizations, and NGOs. Given these findings, we chose to switch our strategy for Canada and Germany, opting for a time series analysis at the national level.

Although the time series approach enabled us to add many more observations to our datasets models, a lack of provincial data for Germany and Canada unfortunately means that our analysis is constrained to the national scale. We would have liked to use provincial data, as this would allow us to not only determine how much money should be relegated to the national healthcare system but also give deeper insights into the allocation of those resources. Therefore, we would potentially be able to address more local, region-specific healthcare needs in addition to our national-level conclusions.

Additionally, while Turkish city-level data was readily available for each of our predictors, obtaining the corresponding healthcare expenditures in 2021 for each city was challenging. For this reason, as mentioned above, we utilized each city's GDP per capita, an alternative measure of economic strength. Ideally, we would have liked to use city-level healthcare expenditure, as it is an economic factor less subject to national/global trends (i.e. inflation) when compared to GDP per capita.

All of this being said, one final limitation that may hinder the accuracy of our analysis is the range of our data. Our data only goes up to 2022, which represents a period notably marked by the unprecedented economic disruption and strain on healthcare systems due to the COVID-19 pandemic. The pandemic increased volatility in both healthcare expenditures and refugee inflows due to the urgent health crisis and an increase in border closures and heightened travel restrictions. Thus, our models may not be as effective in predicting current healthcare expenditures, which have since returned to more typical figures.

Analysis

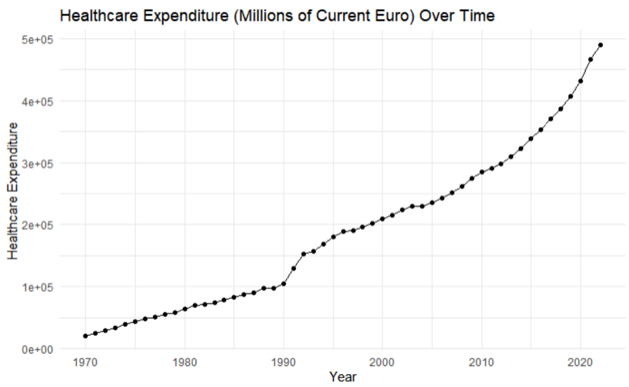
The time series data for both Canada and Germany was analyzed via a vector autoregression (VAR) model. This approach is appropriate for multivariate time series analysis, as VAR models each variable as a function of the lagged versions of itself and the other variables in the system. Thus, we were able to forecast healthcare expenditures while accounting for the effects of our various predictors, such as refugee flow, GDP, etc. Also, VAR allows us to see how the variables affect each other within the model and identify significant predictors, which can help to estimate the effect sizes of each of our explanatory variables. When determining the best method of time-series analysis, we wanted to select an approach that could effectively incorporate multiple predictors and account for any stationarity issues we encountered. Because of this, we initially also considered fitting an ARIMAX model, which is essentially an autoregressive integrated moving average – or ARIMA – model, but with exogenous variables (in this case, our predictors). Since the exogenous variables used do not necessarily need to be stationary, this type of model would help to account for any seasonality in our data or cases where we were unable to obtain a stationary series after differencing or other transformations. That being said, we were successful in stationary transformation during prior EDA, so we decided to forgo this approach in favor of VAR modeling, which is easier to interpret and better shows the linear interdependencies across each of our time series.

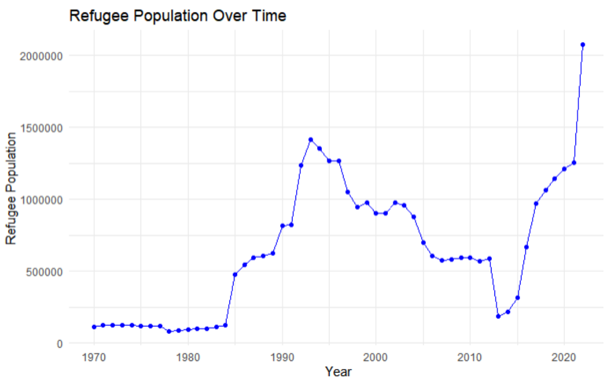
Germany

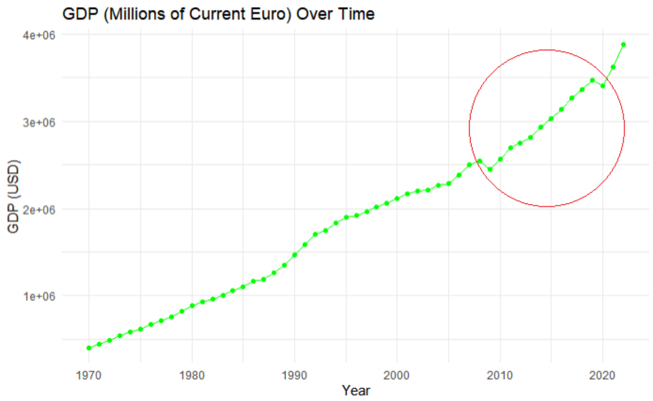
1. Data Cleaning and EDA

Prior to model creation preliminary data cleaning and EDA was performed. Data cleaning primarily consisted of aggregating the relevant predictors from each dataset, combining the year of the observation. There were no missing years in our period of observation (1970-2022) and all variables were transformed to numeric values before analysis.

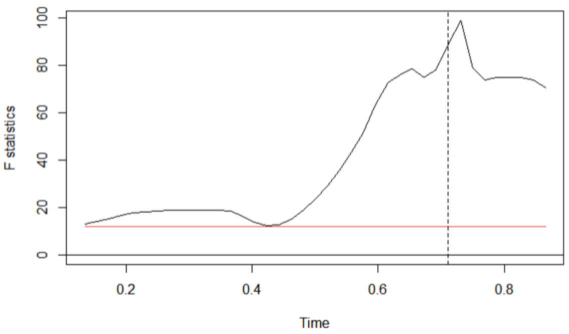
EDA was performed on each of the predictors in our finalized dataset and consisted of plotting time series graphs to visualize potential trends, or seasonality that might influence our analysis. The time series graphs for each variable can be seen below:

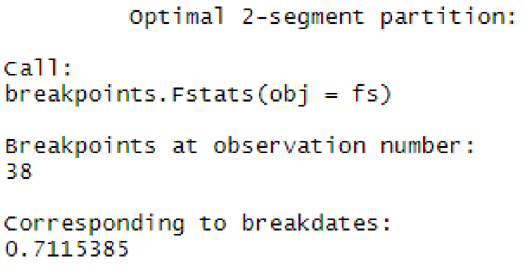






We observed a positive trend for both the healthcare expenditure and GDP over time in Germany. Both variables seem to be increasing at a relatively constant rate, but a potential structural break in the GPP from 2008-2019 (circled) was of significant concern, warranting further analysis. The variable for the refugee population was observed to have a net increase over time, with periods of notable stability and volatility. Since structural breaks can interfere with the stability of our estimations, we performed a Chow test to identify potential breakpoints.



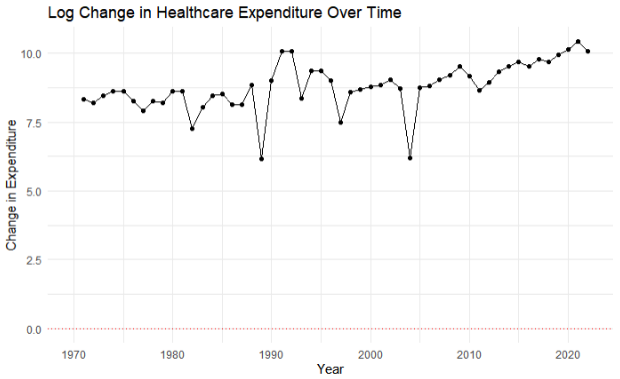


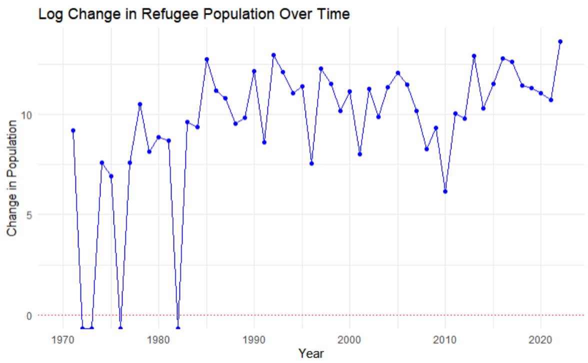
```
         Optimal 2-segment partition:

Call:
breakpoints.Fstats(obj = fs)

Breakpoints at observation number:
38

Corresponding to breakdates:
0.7115385
```

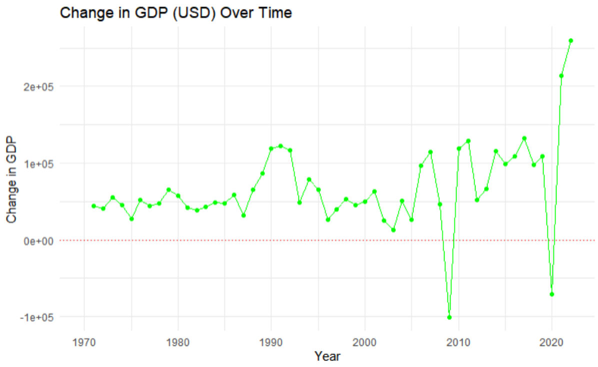
The optimal breakpoint was found to occur at observation 38, which corresponds to the year 2008. This was not very surprising, as 2008 saw the most severe financial crisis since the Great Depression, affecting countries worldwide. Although this result was concerning, we successfully identified a significant structural change, which we accounted for in model building and analysis.

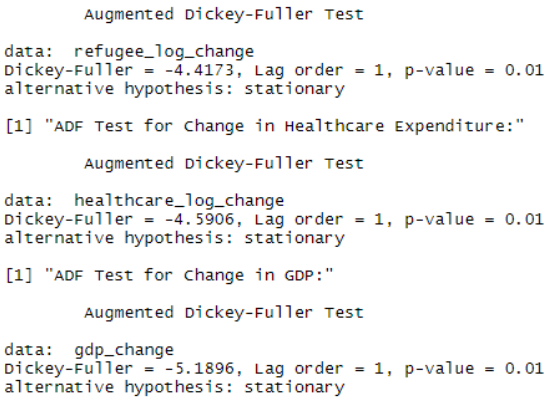
2. Model Assumptions

The core assumption of VAR analysis is that all time series are stationary, meaning that the mean and variance are constant over time. Achieving stationarity in time series analysis is crucial, as it allows us to be sure we are creating reliable estimates and not fitting any trends or seasonality in the data. We were able to achieve stationarity across all of our variables by taking the first differences of each and also applying a log transformation if needed.







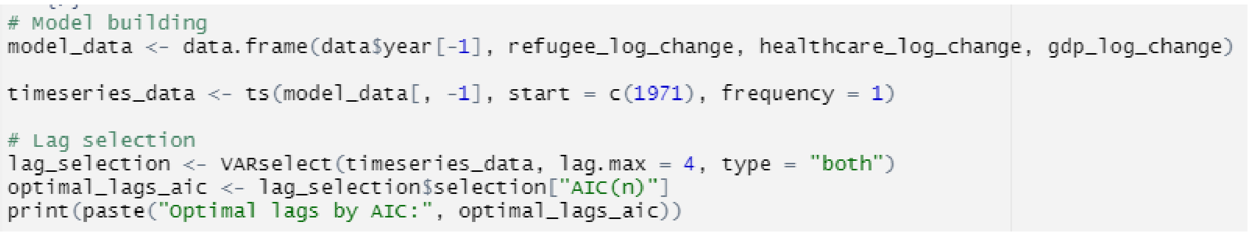


```
        Augmented Dickey-Fuller Test

data:  refugee_log_change
Dickey-Fuller = -4.4173, Lag order = 1, p-value = 0.01
alternative hypothesis: stationary

[1] "ADF Test for Change in Healthcare Expenditure:"

        Augmented Dickey-Fuller Test

data:  healthcare_log_change
Dickey-Fuller = -4.5906, Lag order = 1, p-value = 0.01
alternative hypothesis: stationary

[1] "ADF Test for Change in GDP:"

        Augmented Dickey-Fuller Test

data:  gdp_change
Dickey-Fuller = -5.1896, Lag order = 1, p-value = 0.01
alternative hypothesis: stationary
```

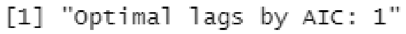
The above plots show each of our transformed time series. The ADF tests for each variable were significant at lag 1, indicating that they are stationary. Note that the first difference was sufficient to achieve stationarity with GDP and the log difference was taken for the refugee and healthcare expenditure variables. Since the data was collected annually, we chose to use a low lag order of 1, as seasonality would not be an issue. Thus, our stationarity assumptions were met and we proceeded to model building.

3. Model Building

The VAR model for Germany was built using 3 time series: the log change in refugee population, the log change in healthcare expenditure, and the change in GDP. We performed lag selection via AIC, which determined the optimal number of lags in our model to be 1, which was consistent with our stationarity lag level.



```
# Model building
model_data <- data.frame(data$year[-1], refugee_log_change, healthcare_log_change, gdp_log_change)

timeseries_data <- ts(model_data[, -1], start = c(1971), frequency = 1)

# Lag selection
lag_selection <- VARselect(timeseries_data, lag.max = 4, type = "both")
optimal_lags_aic <- lag_selection$selection["AIC(n)"]
print(paste("Optimal lags by AIC:", optimal_lags_aic))
```

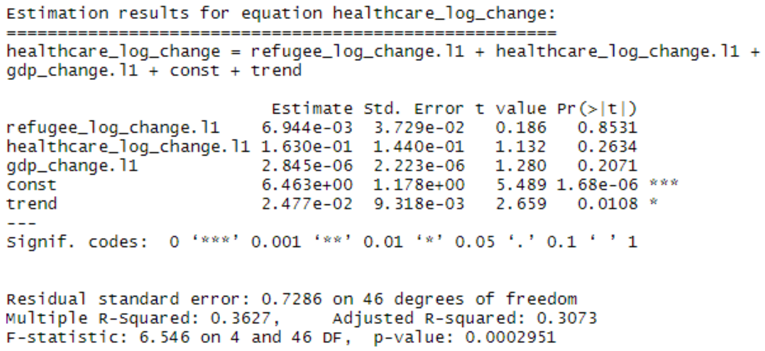


```
[1] "Optimal lags by AIC: 1"
```

The optimal number of lags was also reinforced by ACF and PACF plots of the response variable, which were both significant at lag 1, but not subsequent lags. Knowing this, we built an initial lag-1 VAR model using the 3 time series .

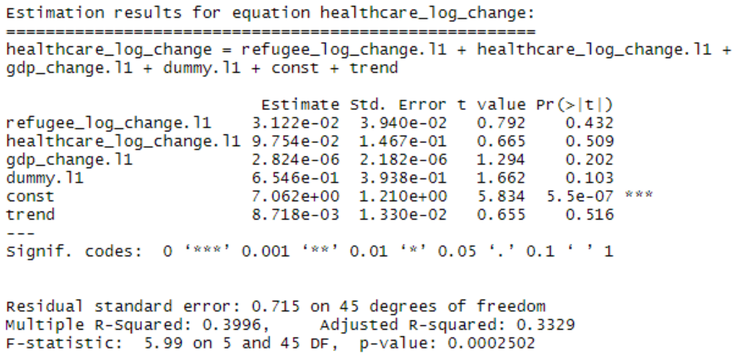
4. Model Validation and Selection

The initial VAR model was as follows:



```
Estimation results for equation healthcare_log_change:
======================================================
healthcare_log_change = refugee_log_change.l1 + healthcare_log_change.l1 +
gdp_change.l1 + const + trend

                         Estimate Std. Error t value Pr(>|t|)
refugee_log_change.l1    6.944e-03  3.729e-02   0.186   0.8531
healthcare_log_change.l1 1.630e-01  1.440e-01   1.132   0.2634
gdp_change.l1            2.845e-06  2.223e-06   1.280   0.2071
const                    6.463e+00  1.178e+00   5.489 1.68e-06 ***
trend                    2.477e-02  9.318e-03   2.659   0.0108 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.7286 on 46 degrees of freedom
Multiple R-Squared: 0.3627,     Adjusted R-squared: 0.3073
F-statistic: 6.546 on 4 and 46 DF,  p-value: 0.0002951
```

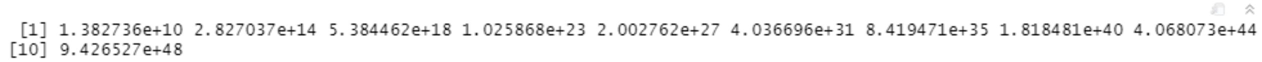
VAR gives linear regressions across each time series used. Since we are only interested in predicting healthcare expenditure, we only needed the section with the relevant response variable. We were able to create a significant model predicting the log healthcare change with an adjusted R-squared of 0.3073, which indicates a moderate to weak fit to our data. We also found that the refugee/gdp variables were not significant predictors of the response, but that there was a significant constant term. This result was worse than we expected, which we hypothesized was due to the structural break in GDP, which would have also interfered with our estimations due to its recency. Thus, we decided to fit a model using a dummy variable to represent 2 partitions in the data. The dummy variable had a value of 0 pre-2008 and 1 from 2008 onward. The model incorporating our dummy variable is as follows:



```
Estimation results for equation healthcare_log_change:
======================================================
healthcare_log_change = refugee_log_change.l1 + healthcare_log_change.l1 +
gdp_change.l1 + dummy.l1 + const + trend

                         Estimate Std. Error t value Pr(>|t|)
refugee_log_change.l1    3.122e-02  3.940e-02   0.792    0.432
healthcare_log_change.l1 9.754e-02  1.467e-01   0.665    0.509
gdp_change.l1            2.824e-06  2.182e-06   1.294    0.202
dummy.l1                 6.546e-01  3.938e-01   1.662    0.103
const                    7.062e+00  1.210e+00   5.834  5.5e-07 ***
trend                    8.718e-03  1.330e-02   0.655    0.516
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.715 on 45 degrees of freedom
Multiple R-Squared: 0.3996,     Adjusted R-squared: 0.3329
F-statistic:  5.99 on 5 and 45 DF,  p-value: 0.0002502
```

The adjusted R of our new model was 0.3329, representing an increase of 0.0255, which was not as significant a difference as expected. Thus, we decided to move forward with forecasting using our simpler initial model, as to avoid potentially overfitting to the additional variable.

The predicted values for Germany’s healthcare expenditures are as follows:



```
 [1] 1.382736e+10 2.827037e+14 5.384462e+18 1.025868e+23 2.002762e+27 4.036696e+31 8.419471e+35 1.818481e+40 4.068073e+44
[10] 9.426527e+48
```

Canada

1. Data Cleaning and EDA

Prior to model creation preliminary data cleaning and EDA was performed. Data cleaning primarily consisted of aggregating the relevant predictors from each dataset, combining the year of the observation. There were no missing years in our period of observation (1976-2022) and all variables were transformed to numeric values before analysis. Variables that initially had longer names were renamed to make the dataset more concise and easy to work with. To potentially provide more insight into our analysis, the combined variable “Total_Exp/GDP” was created to show the healthcare expenditure as a percentage of the GDP that year, but we ultimately did not incorporate it in our model.

EDA was performed on each of the predictors in our finalized dataset and consisted of plotting time series graphs to visualize potential trends or seasonality that might influence our analysis. The time series graphs for each variable can be seen below:

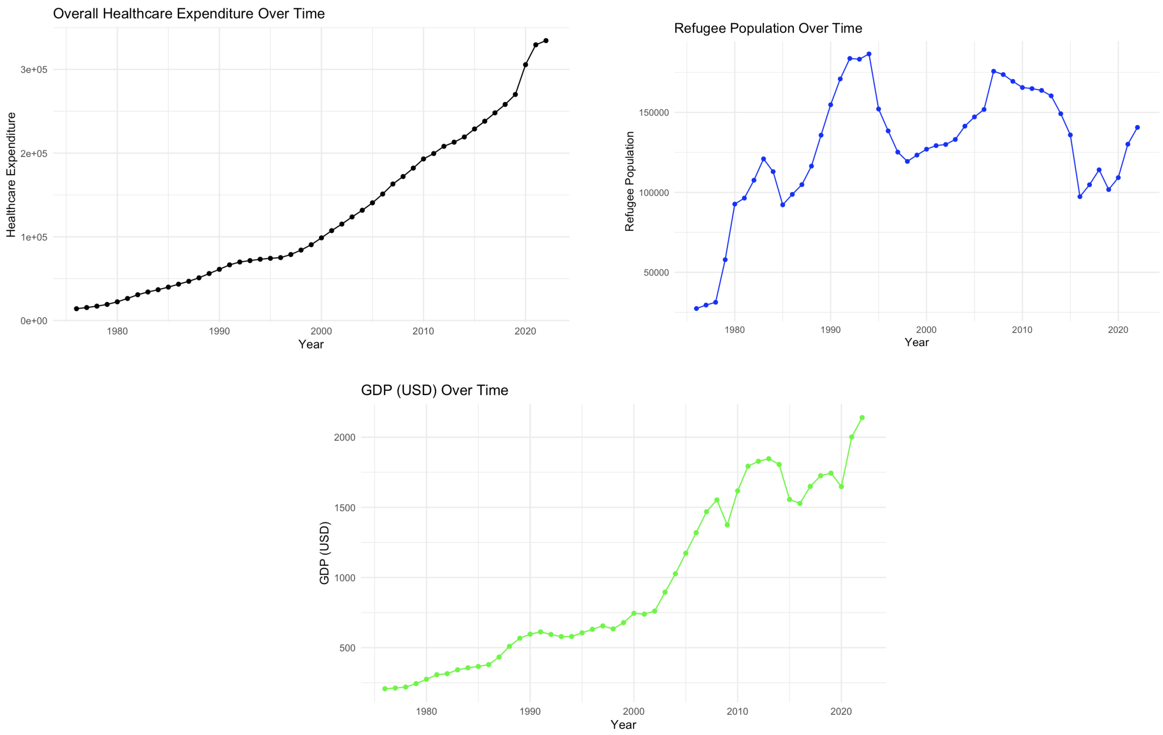
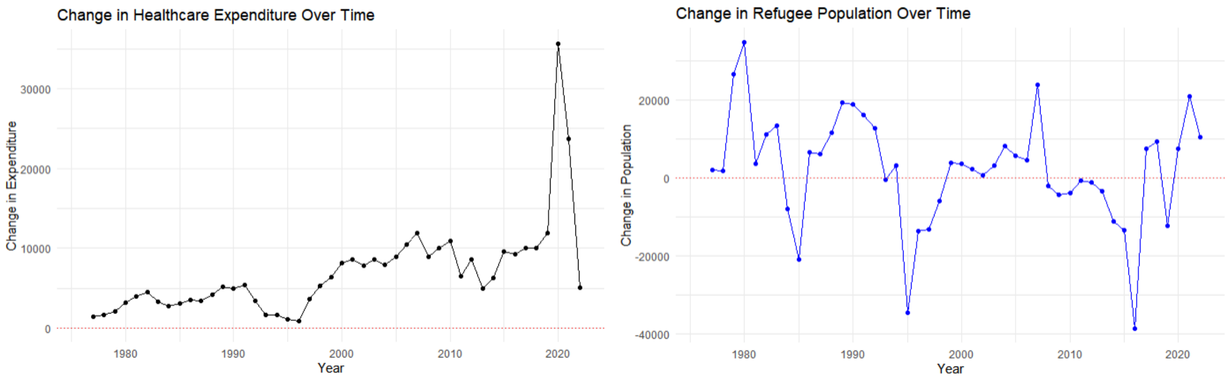
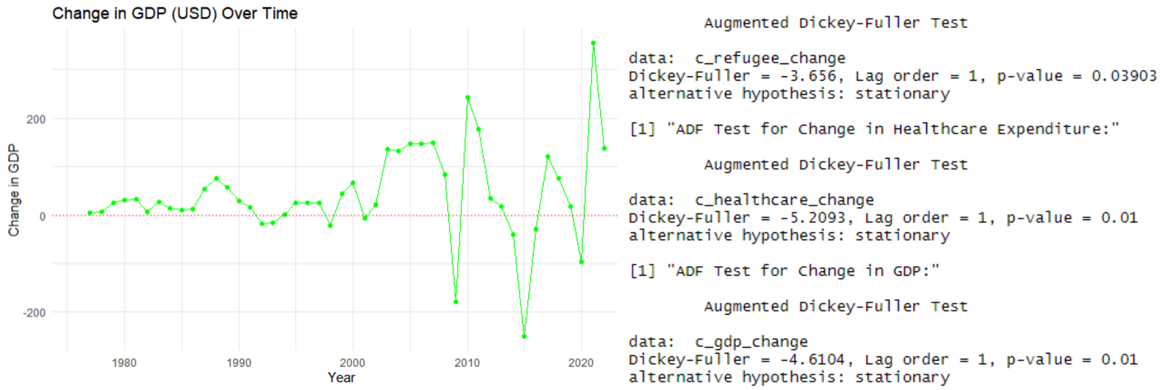


Figure 1. Time-Series Plots for Canada

Similar to our EDA for Germany, we observed a positive trend for both healthcare expenditure and GDP over time in Canada. Both variables seem to be increasing at a relatively constant rate, with GDP increasing in volatility post-2008. The variable for the refugee population was observed to rise and lower significantly over time, reaching its peak in 1994. Overall, we did not see any signs of seasonality so we moved on to our stationarity assumptions.

2. Model Assumptions





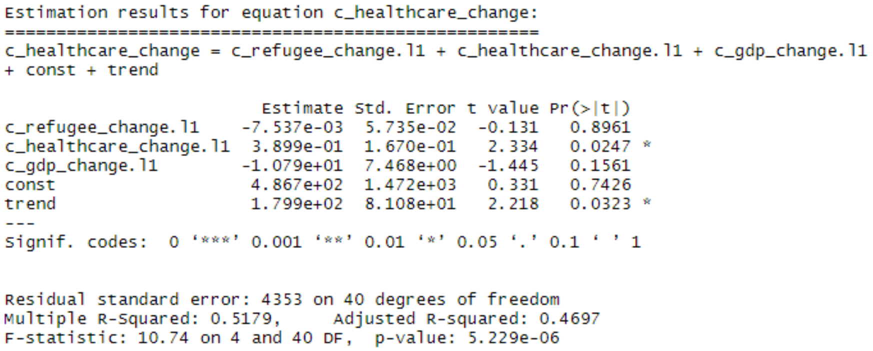
```
        Augmented Dickey-Fuller Test

data:  c_refugee_change
Dickey-Fuller = -3.656, Lag order = 1, p-value = 0.03903
alternative hypothesis: stationary

[1] "ADF Test for Change in Healthcare Expenditure:"

        Augmented Dickey-Fuller Test

data:  c_healthcare_change
Dickey-Fuller = -5.2093, Lag order = 1, p-value = 0.01
alternative hypothesis: stationary

[1] "ADF Test for Change in GDP:"

        Augmented Dickey-Fuller Test

data:  c_gdp_change
Dickey-Fuller = -4.6104, Lag order = 1, p-value = 0.01
alternative hypothesis: stationary
```

Figure 2. Canada Assumptions

In order to meet the stationarity assumption used by our VAR analysis, we took the first differences of each of our 3 variables, which was sufficient to achieve lag-1 stationarity. Since all ADF tests were significant, we proceeded to build our initial VAR model with the 3 time series.

3. Model Building

The initial VAR model was as follows:



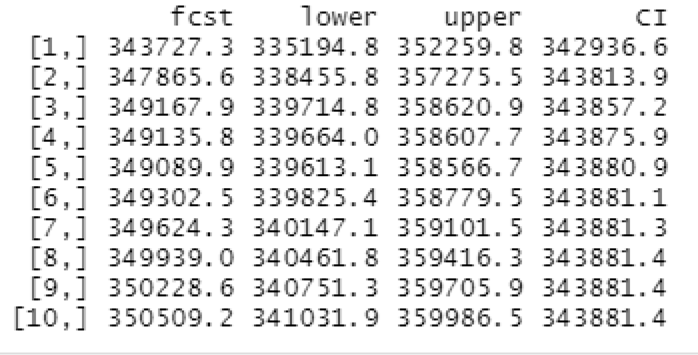
```
Estimation results for equation c_healthcare_change:
====================================================
c_healthcare_change = c_refugee_change.l1 + c_healthcare_change.l1 + c_gdp_change.l1
+ const + trend

                         Estimate Std. Error t value Pr(>|t|)
c_refugee_change.l1    -7.537e-03  5.735e-02  -0.131   0.8961
c_healthcare_change.l1  3.899e-01  1.670e-01   2.334   0.0247 *
c_gdp_change.l1        -1.079e+01  7.468e+00  -1.445   0.1561
const                   4.867e+02  1.472e+03   0.331   0.7426
trend                   1.799e+02  8.108e+01   2.218   0.0323 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 4353 on 40 degrees of freedom
Multiple R-Squared: 0.5179,     Adjusted R-squared: 0.4697
F-statistic: 10.74 on 4 and 40 DF,  p-value: 5.229e-06
```

Figure 3. Canada Model-Building

We were able to create a significant model predicting the change in healthcare expenditure with an adjusted R-squared of 0.4697, which indicates a moderate fit to our data. We also found past values of the change in healthcare to be a significant predictor of future changes in healthcare. This result looked more promising, as the adjusted R-squared was higher compared to Germany's VAR model and also because we were able to identify a significant predictor. Thus, we decided to move on to forecasting with our initial VAR model.

4. Model Validation and Selection

To validate our model, we first forecasted the future change in healthcare expenditures for the next 10 years. Since Canadian healthcare expenditure data for 2023 and onward has yet to be released, our goal was to compare these values with currently forecasted figures in order to assess the accuracy of our model. To obtain these predictions, we had to add the forecasted changes in healthcare expenditure to our latest observation (from 2022). The predicted values are visualized in figure 4 down below.



```
          fcst    lower    upper       CI
 [1,] 343727.3 335194.8 352259.8 342936.6
 [2,] 347865.6 338455.8 357275.5 343813.9
 [3,] 349167.9 339714.8 358620.9 343857.2
 [4,] 349135.8 339664.0 358607.7 343875.9
 [5,] 349089.9 339613.1 358566.7 343880.9
 [6,] 349302.5 339825.4 358779.5 343881.1
 [7,] 349624.3 340147.1 359101.5 343881.3
 [8,] 349939.0 340461.8 359416.3 343881.4
 [9,] 350228.6 340751.3 359705.9 343881.4
[10,] 350509.2 341031.9 359986.5 343881.4
```

Figure 4. Forecasted Healthcare Expenditure for Canada (2023 onward)

Compared to current estimates by the Canadian government, our model predicts the same healthcare expenditure for 2023, which was ~344 billion canadian dollars.

Turkey

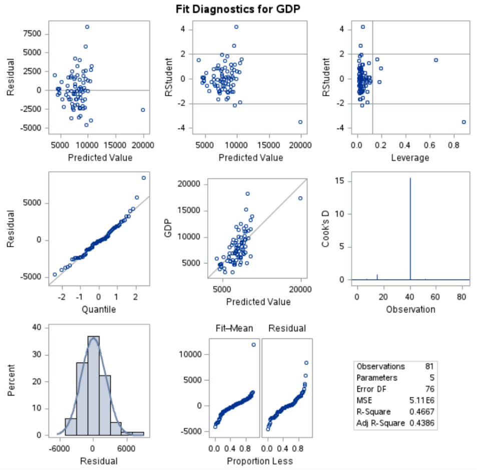
1. Data Cleaning

Before creating the multiple linear regression (MLR) model from the Turkey dataset, initial data cleaning was conducted. The refugee population numbers were initially divided by 1,000 in the downloaded dataset. To address this, all population numbers were multiplied by

1,000 (for example, 531.996 was modified to become 531,996). Additionally, the pre-cleaned data set had both city-level refugee sub-population count and city-level total population count. We created and utilized a new variable "Refugee Percentage" instead of the refugee sub-population count in our analysis. This variable was calculated for each city by dividing a city's refugee sub-population by the total city population. The total city population count variable, however, was kept in the dataset, as discussed below.

2. Assumptions

After data cleaning, we examined the MLR assumptions of linearity, independence, homoscedasticity, and normality of residuals. Based on Figure 3, we saw that without using the total city population count with weighted least squares, the assumptions were met. When using the total population count in the WLS model, we also noted that all assumptions were met based on the plots.



Observations	81
Parameters	5
Error DF	76
MSE	5.11E6
R-Square	0.4667
Adj R-Square	0.4386

Figure 5. Turkey Assumption Plots without WLS

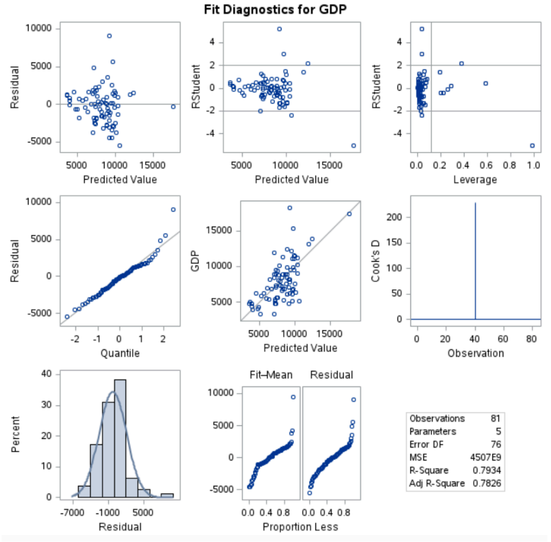
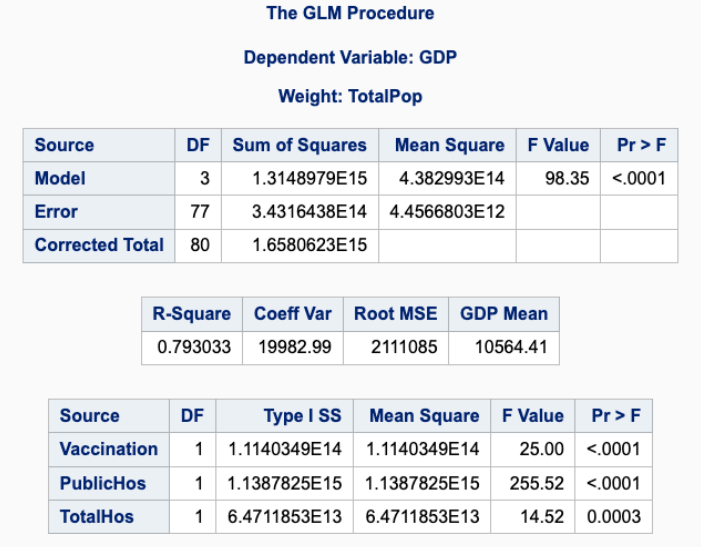


Figure 6. Turkey Assumption Plots with WLS

3. Model Creation

In the MLR model, GDP per capita was utilized as the response variable, while vaccination rate, number of public hospitals, total hospital count, and refugee percentage were predictor variables. The non-WLS MLR model had an R-square of 0.455828 while the WLS MLR model had an R-square of 0.793423. In the final stage of the Turkey analysis, in which we tested the model, only the higher-performing WLS MLR model was utilized. As refugee percentage was not a significant predictor of GDP per capita, with an associated p-value of 0.7057, this predictor was removed from the model, resulting in the final WLS MLR model with an R-square of 0.793033, below.



The GLM Procedure

Dependent Variable: GDP

Weight: TotalPop

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.3148979E15	4.382993E14	98.35	<.0001
Error	77	3.4316438E14	4.4566803E12		
Corrected Total	80	1.6580623E15			

R-Square	Coeff Var	Root MSE	GDP Mean
0.793033	19982.99	2111085	10564.41

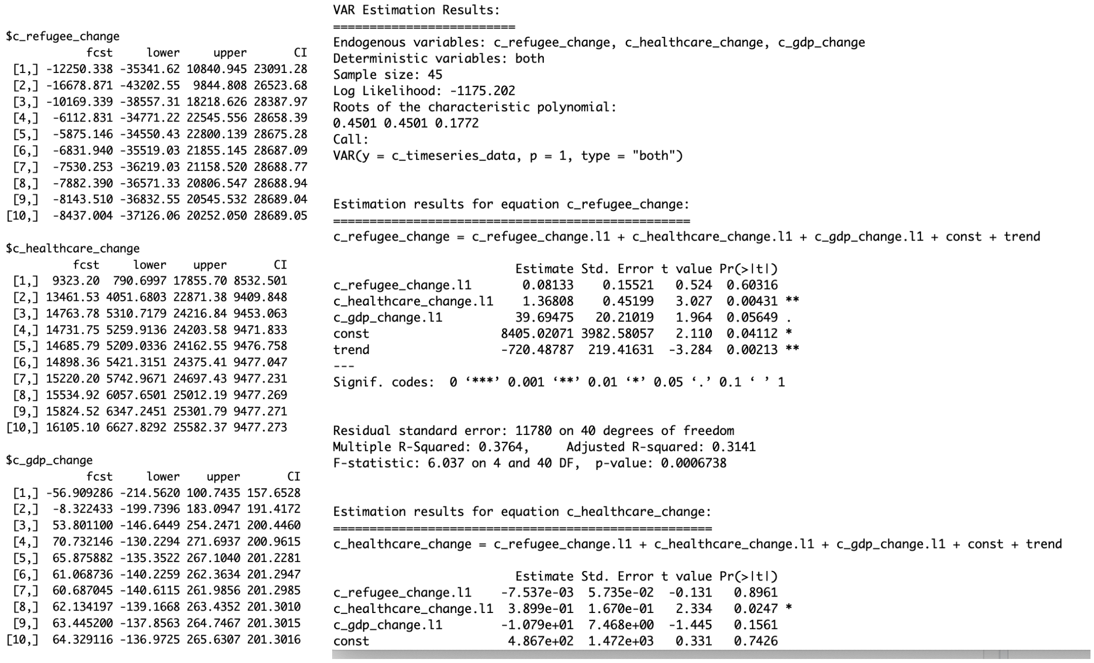
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Vaccination	1	1.1140349E14	1.1140349E14	25.00	<.0001
PublicHos	1	1.1387825E15	1.1387825E15	255.52	<.0001
TotalHos	1	6.4711853E13	6.4711853E13	14.52	0.0003

Figure 7. Turkey Three-Predictor WLS MLR Final Model

The WLS MLR is the most suitable approach for the Turkey dataset. MLR allowed us to simultaneously assess the influence of multiple predictor variables (vaccination rate, number of public hospitals, total hospital count, and refugee percentage) on a single outcome variable (GDP per capita). Additionally, WLS is designed to handle unequal variability by assigning weights to data points, giving less weight to data points with higher variance. This is particularly useful in our case because we expected cities with larger populations (ex. Istanbul, Ankara) to be more representative of the country as a whole compared to smaller cities which may not reliably capture the underlying relationship. The significant improvement in the R-square value when using WLS weighted on population (0.793423) compared to non-WLS MLR (0.455828) indicates that WLS provides a better fit of the model to the data.

Conclusion

In our analysis of Germany's economic and demographic data from 1970 to 2022, we conducted thorough data cleaning and exploratory data analysis to examine trends in healthcare expenditure, GDP, and refugee populations. We identified a significant structural shift in 2008, coinciding with the global financial crisis, which required adaptation in our modeling approach. Despite employing a Vector Autoregression (VAR) model and making adjustments for non-stationarities and structural breaks, our models only achieved a moderate-low fit. The initial model showed limited predictive power, leading us to attempt modifications, including the use of a dummy variable to account for changes post-2008. However, the improvement was minimal, indicating that our models might be omitting key influencing factors. This outcome suggests a need to revisit our analytical approach and consider additional or different variables to better understand the drivers of healthcare expenditure in Germany.



```
$c_refugee_change
            fcst     lower      upper       CI
 [1,] -12250.338 -35341.62 10840.945 23091.28
 [2,] -16678.871 -43202.55  9844.808 26523.68
 [3,] -10169.339 -38557.31 18218.626 28387.97
 [4,]  -6112.831 -34771.22 22545.556 28658.39
 [5,]  -5875.146 -34550.43 22800.139 28675.28
 [6,]  -6831.940 -35519.03 21855.145 28687.09
 [7,]  -7530.253 -36219.03 21158.520 28688.77
 [8,]  -7882.390 -36571.33 20806.547 28688.94
 [9,]  -8143.510 -36832.55 20545.532 28689.04
[10,]  -8437.004 -37126.06 20252.050 28689.05

$c_healthcare_change
          fcst     lower    upper       CI
 [1,]  9323.20  790.6997 17855.70 8532.501
 [2,] 13461.53 4051.6803 22871.38 9409.848
 [3,] 14763.78 5310.7179 24216.84 9453.063
 [4,] 14731.75 5259.9136 24203.58 9471.833
 [5,] 14685.79 5209.0336 24162.55 9476.758
 [6,] 14898.36 5421.3151 24375.41 9477.047
 [7,] 15220.20 5742.9671 24697.43 9477.231
 [8,] 15534.92 6057.6501 25012.19 9477.269
 [9,] 15824.52 6347.2451 25301.79 9477.271
[10,] 16105.10 6627.8292 25582.37 9477.273

$c_gdp_change
            fcst     lower    upper       CI
 [1,] -56.909286 -214.5620 100.7435 157.6528
 [2,]  -8.322433 -199.7396 183.0947 191.4172
 [3,]  53.801100 -146.6449 254.2471 200.4460
 [4,]  70.732146 -130.2294 271.6937 200.9615
 [5,]  65.875882 -135.3522 267.1040 201.2281
 [6,]  61.068736 -140.2259 262.3634 201.2947
 [7,]  60.687045 -140.6115 261.9856 201.2985
 [8,]  62.134197 -139.1668 263.4352 201.3010
 [9,]  63.445200 -137.8563 264.7467 201.3015
[10,]  64.329116 -136.9725 265.6307 201.3016
```

```
VAR Estimation Results:
=========================
Endogenous variables: c_refugee_change, c_healthcare_change, c_gdp_change
Deterministic variables: both
Sample size: 45
Log Likelihood: -1175.202
Roots of the characteristic polynomial:
0.4501 0.4501 0.1772
Call:
VAR(y = c_timeseries_data, p = 1, type = "both")


Estimation results for equation c_refugee_change:
================================================
c_refugee_change = c_refugee_change.l1 + c_healthcare_change.l1 + c_gdp_change.l1 + const + trend

                      Estimate Std. Error t value Pr(>|t|)
c_refugee_change.l1      0.08133    0.15521   0.524  0.60316
c_healthcare_change.l1   1.36808    0.45199   3.027  0.00431 **
c_gdp_change.l1         39.69475   20.21019   1.964  0.05649 .
const                 8405.02071 3982.58057   2.110  0.04112 *
trend                 -720.48787  219.41631  -3.284  0.00213 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 11780 on 40 degrees of freedom
Multiple R-Squared: 0.3764,     Adjusted R-squared: 0.3141
F-statistic: 6.037 on 4 and 40 DF,  p-value: 0.0006738


Estimation results for equation c_healthcare_change:
===================================================
c_healthcare_change = c_refugee_change.l1 + c_healthcare_change.l1 + c_gdp_change.l1 + const + trend

                        Estimate Std. Error t value Pr(>|t|)
c_refugee_change.l1    -7.537e-03  5.735e-02  -0.131   0.8961
c_healthcare_change.l1  3.899e-01  1.670e-01   2.334   0.0247 *
c_gdp_change.l1        -1.079e+01  7.468e+00  -1.445   0.1561
const                   4.867e+02  1.472e+03   0.331   0.7426
```

Figure 8. Canada Analysis

The forecasted results of the Canada analysis show that healthcare expenditure is predicted to keep increasing at a higher and higher rate. The first four numbers of the healthcare change matrix suggest that healthcare will increase at an ever more rapid rate (9323, 13462, 14764, 14732), whereas the rate of change for Canada's GDP and refugee intake will actually decrease in the years to come. The adjusted R-squared value of 0.4697 indicates a moderate fit of data to the model. This means that for our model to be stronger, certain variables, such as refugee count, should be removed (as seen from the high p-value), and other variables be inserted into the model. Other factors that determine the healthcare expenditure were excluded, meaning that all the factors towards determining this forecast were not fully captured in the model. Despite this, comparing our results with the predicted outlook of the Canada Institute of Health Information (3) shows that the predictions are very similar to the official predictions by the government. If other significant factors are added, our model can be even more accurate than the model created by the Canadian government!

The results of the Turkey analysis demonstrate that factors such as the vaccination rate, the number of public hospitals, and the total hospital count significantly influence the GDP per capita, suggesting that these elements are integral to understanding the economic health of cities in Turkey. Notably, the refugee percentage did not significantly impact GDP per capita, indicating that the presence of refugees in a city does not necessarily correlate with lower economic performance, contrary to common assumptions. However, it's important to recognize the context of these results and the limitations of available data. The high R-square value in the model indicates a strong fit for the data analyzed, but this may not fully capture all factors at play or be generalizable beyond the cities and year studied.

By analyzing the models that we built for each country, one message truly stands out. Each model saw varying degrees of success and utilized different variables to yield the desired results. Through the use of different variables and data, the construction of three models shows how the unique characteristics of each country cannot be generalized into one overarching model. This highlights one main difficulty for statisticians when model building is conducted to create a general worldwide model. In this scenario, different issues and variables play a much heavier/lighter role in different countries. Variables such as hospital count and vaccination rate played a big role in the model for Turkey, whereas in the Canada and Germany models, they were not used at all (due to lack of data). These differences can be attributed to a lack of reliable data in certain countries, or due to the unique characteristics of how each country runs its healthcare systems. Therefore, if students or statisticians seek to build a model to measure a global variable, for healthcare or other purposes, they should instead look towards building individual models for each country, rather than one single worldwide model.

References

1. World report on the health of refugees and migrants. Geneva: World Health Organization; 2022. Licence: CC BY-NC-SA 3.0 IGO.
2. Refugee Statistics | UN Fact Sheet. https://www.unrefugees.org/refugee-facts/statistics/
3. https://www.cihi.ca/en/national-health-expenditure-trends-2023-snapshot

Appendix A: Data Dictionaries

Germany Data Dictionary

Variable	Variable Name	Measurement Unit	Allowed Values	Descriptions
Year	Year	N/A	N/A	The year of which the observation was taken
GDP	GDP	Numeric	Min: 0, Max: N/A	Total GDP of Germany in millions of Euro (current)

Number of Refugees	NumRefugee	Numeric	Min: 0, Max: N/A	Refugee population
Total Healthcare Expenditure	HealthcareExp	Numeric	Min: 0, Max: N/A	Total healthcare expenditure of Germany in millions of Euro (current)

Formatted based on the Open Science Framework's (OSF) "How to Make a Data Dictionary." (May 2023)

Canada Data Dictionary

Variable	Variable Name	Measurement Unit	Allowed Values	Descriptions
Year	Year	N/A	N/A	The year of which the observation was taken
GDP	GDP	Numeric	Min: 0, Max: N/A	Total GDP of Canada's economy, in billions of USD
Total Immigrants	Total_immigrants	Numeric	Min: 0, Max: N/A	Total immigrant population of Canada
Total Asylum Seekers	Asylum	Numeric	Min: 0, Max: N/A	Total Asylum seeking population of Canada
Total Healthcare Expenditure	Total_Exp	Numeric	Min: 0, Max: N/A	Total healthcare expenditure of Canada, in millions of USD
Total Healthcare Expenditure as a Percentage of GDP	Total_exp/GDP	Percentage	Min: 0, Max: 100	The measure of Canada's healthcare expenditure as a percentage of the GDP observed at the time.

Number of Hospital beds per Capita	Hospital_beds	Numeric	Min: 0, Max: N/A	The total number of private and public hospital beds in the country per 1000 people.

Formatted based on the Open Science Framework's (OSF) "How to Make a Data Dictionary." (May 2023)

Turkey Data Dictionary

Variable	Variable Name	Measurement Unit	Allowed Values	Descriptions
City name	City	N/A	N/A	Each city name corresponds to one of the 81 major cities included.
Vaccination rate	VacRate	Percentage	0.0-100.0	COVID-19 vaccination rate
Refugee Percentage	Refugees	Percentage	0.0-100.0	Percentage of the city population that identifies as a Syrian refugee/asylum-seeker.
Public Hospitals	PubHos	Numeric	Min: 0, Max: N/A	Number of public hospital beds in each city
Total Hospitals	TotalHos	Numeric	Min: 0, Max: N/A	Number of public and private hospital beds in each city
GDP Per Capita	GDP	Numeric	Min: 0, Max: N/A	A measure of each city's economic output (Gross Domestic Product) divided by its population.

Formatted based on the Open Science Framework's (OSF) "How to Make a Data Dictionary." (May 2023)