National Personalities and Politics

Zachary Kleiman

December 9, 2025

Contents

1 Abstract

In this study, we set out to investigate whether mean personality traits of a country in a given year can be used to predict the state of civil and economic freedom in that country. To do this we train a model using datasets for both country's personalities (as measured by the Big Five) from a dataset found on Kaggle and scores for civil and economic freedom generated by the CATO institute. This

model demonstrates there is a very strong correlation between the two (with it being much stronger for some freedoms than others).

2 Introduction

> Before and besides social scientists, national character has been considered by informed observers from Tacitus to Madame de Stael.[1] – Dean Peabody

Personality is one of the defining features of a person. A culture or a group is defined by the personalities of its constituent members. This personality is a reflection of the essence or 'the soul' of a person or of the group or culture. Individual and group personality shape all interactions with other parties, how one relates to themselves, (of groups) how intra-group matters should be handled, etc. Because of this, since time immemorial people have spoken about personality traits (whether they viewed them as inherited or learned is immaterial for purposes of this study)[2][3] and how they should be handled.

In modern times, along with many other fields, the study of personality has turned scientific. There are many personality tests that are widely deployed. One of the most popular in business circles are MBTI types which are based on Jungian theory. The most scientifically reliable are either clinical tools or based on the Big Five trait model.[4] These Big 5 traits are Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Unlike other personality tests like the MBTI that assign types each of these is given a 0-100 score in the Big Five model. A dataset of over a million such responses from 2016-2018 was uploaded by Bojan Tunguz on Kaggle. The dataset covers a broad cross-section of countries.

The other discipline included in this interdisciplinary study is political science. Specifically, we are studying how cultural statistics about personality impact the level of freedom – both civic and economic – in a country. There are two notions of freedom: positive and negative liberty, as famously expressed by Isaiah Berlin.[5] Briefly: positive liberty refers to the freedom to pursue the good (whether obstructed by internal or external constraint) while negative liberty refers to freedom from constraint. Libertarianism is interested in the maximization of negative liberty. The CATO institute is one of the leading Libertarian think-tanks in the United States. Every year, they publish the Human Freedom Index ranking the state of negative liberty (civic and economic) in every country on earth across several metrics. The Index runs 2 years behind so eg. the 2023 issue covers the state of the world in 2021. These scores are used as reference scores for prediction based on personality traits. Due to running 2 years behind the reports from 2016-2018 were used.

[1] Dean Peabody, National Characteristics (Cambridge: Cambridge University Press, 1985), 5.

[2] "Is is noteworthy that the English words connoting personal uniqueness, 'personality' and 'individuality', come from Latin rather than Greek." Ancient Philosophy Volume 11, Issue 1, Spring 1991 Paula Reiner Pages 67-84 fn. 1 https://doi.org/10.5840/ancientphil199111135

[3] Interestingly there appears to be a lot of disagreement about this, "Roger Brown (1965) cites the study of Buchanan and Cantril (1953) where subjects were asked whether their own national characteristics were mainly inborn or due to the way they had been brought up. The belief that their national characteristics were mainly inborn was expressed by 59% of the West Germans, 39% of the British, and 15% of the Americans." Dean Peabody, National Characteristics (Cambridge: Cambridge University Press, 1985), 16.

[4] "Any personality test can be fun and intriguing. But from a scientific perspective, tools such as the Big Five Inventory (and others based on the five-factor model) and those used by psychological scientists, such as the MMPI, are likely to provide the most reliable and valid results." Sussex Publishers, LLC, "Personality Tests," Psychology Today, n.d., https://www.psychologytoday.com/us/basics/personality/personality-tests#: :text=Any%20personality%20test%20can%20be,most%20reliable%20and%20valid%20results.

[5] Isaiah Berlin, "Two Concepts of Liberty," Four Essays On Liberty, (Oxford, England: Oxford University Press, 1969), p. 118-172.

The 2020 report included a dataset that incorporated all historical data from 2008 on preventing time having to be dedicated to data entry and this provided dataset will be used in this study. It is believed we will find strong connections between national level personality traits and these freedoms. This makes sense theoretically. For instance, openness to experience makes someone (and in aggregate the nation) more willing to listen to the other side resulting in more free speech and a freer civil society.

3 Dataset Summary and Cleaning

3.1 Description of Datasets

3.1.1 Kaggle Dataset

The dataset contains for each entry, the answer to each question in the quiz and what category it fell under (eg. ext_1 standing for the first question in the extraversion category). For each question it has an _E category after that question (eg. ext_1_E) which stands for how much time was spent on that question in milliseconds. It also contains the timestamp for when the survey was started in the dateload column, screen width (screenw), screen height (screenh), introelapse which is the time in seconds spent on the landing / intro page, testelapse which is the time in seconds spent on the page with the survey questions. You also have endelapse which is the time in seconds spent on the finalization page (where the user was asked to indicate if they has answered accurately and their answers could be stored and used for research, this dataset only includes users who answered "Yes" to this question, users were free to answer no and could still view their results either way). The IPC is the the number of records from the user's IP address in the dataset. The user's country was determined by technical information and was not asked as a question. The other two columns are approximate latitudes and longitudes. The dataset consists of 1,015,342 rows. There is invalid data that needs to be cleaned out as documented below (though this is not itself documented and on Kaggle it's described as clean data despite ¿10% of columns containing provably invalid data, see below for details.) The Kaggle dataset covers every region and post-cleaning contains 273 qualify country-year matches.

3.1.2 CATO Human Freedom Index Dataset

This dataset contains scores from 2008-2018 (thereby including our needed 2016-2018 period) on a broad array of freedoms. The freedoms they score are as follows (the data also includes some raw information like the top marginal tax rate and some sub-sub categories eg. dividing homosexuality into male and female homosexuality but these are the general subcategories):

Table 1: Kinds of Freedom

Kinds of Freedom	Subcategories
Rule of Law	Procedural Justice Civil Justice Criminal Justice
Security and Safety	Homicide Disapperances, Conflicts, and Terrorism

Continued on next page

Table 1 continued from previous page

Kinds of Freedom	Subcategories
	Women's Security & Safety
Movement	Freedom of Foreign Movement Freedom of Domestic Movement Women's Freedom of Movement
Religion	Freedom of Religion Religious Organization Suppression Harassment and Physical Hostilities Legal and Regulatory Restrictions
Association Assembly and Civil Society)	Civil Society Entry and Exit Assembly Freedom to Form Political Parties Opposition Parties Autonomy Civil Society Repression
Expression and Information	Press Killed Press jailed Media Freedom Access to Cable and Satellite Access to Foreign Newspapers State Control over Internet Access
Relationship	Same-Sex Relationships Divorce Legal Gender
Size of Government	Government Consumption Transfer and Subsidies Government Enterprises Top Marginal Tax Rate
Legal System and Property Rights	Judicial Independence Impartial Courts Protection of Property Rights Military Interference Integrity of the Legal System Legal Enforcement of Contracts Regulatory Restrictions Reliability of Police
Sound Money	Money Growth Standard Deviation of Inflation Inflation: Most Recent Year Freedom to Own Foreign Currency/ Bank Accounts
Freedom to Trade Internationally	Tariffs Regulatory Trade Barriers

Continued on next page

Table 1 continued from previous page

Kinds of Freedom	Subcategories
	Black-Market Exchange Rates
	Movement of Capital and People
Regulation	Credit Market Regulations
	Labor Market Regulations
	Business Regulation

As can be seen from the above table, the dataset is very comprehensive and covers a broad range of conceivable freedoms. Due to system and time constraints the model is only tuned to the meta-categories of eg. Regulation and not to specific sub-measures.

3.2 Dataset Cleaning

Our goals here are as follows:

1. Eliminate data that is completely irrelevant and beyond the scope of the study.
2. Eliminate invalid data from the Kaggle dataset which was not professionally prepared and is raw.
3. Impute missing values for the CATO Human Freedom Index Dataset, when that is not possible remove those rows.
4. Combine these 2 cleaned up datasets into one pandas dataframe for studying.

3.2.1 Step 1:Remove rows with country 'NONE' or with 0 as an answer to any question from the Kaggle dataset

'None' means the country could not be determined meaning the data is unusable. Additionally only values from 1-5 are valid answers to the personality questions so any 0 necessarily represents corrupt, incomplete or otherwise unusable data. Any row with either of those must be dropped.

3.2.2 Step 2: Remove any rows from the Kaggle dataset with values other than 1-5

After doing Step 1 it was realized that there are rows with values above 5 that are not caught by step 1 so we decided to additionally enforce the 1-5 constraint directly and eliminate any rows that don't fit that requirement.

3.2.3 Step 3: Remove Macau

Macau is not rated by the CATO institute so despite having a sufficient sample size (¿100 for each year data is sampled) it could not be used and had to be dropped.

3.2.4 Step 4: Drop Inheritance Rights: Widows and Inheritance Rights: Daughters from the CATO dataset

These columns were used in old reports well before 2018-2020. They're simply preserved for historical reasons. They are irrelevant for our study.

3.2.5 Step 5: Round 1 of Imputation

The following sub-categories had missing values. These were imputed based on the larger category they were part of as follows:

Table 2: Round 1 Imputation

Category	Where Score Copied From
Access to Cable/Satellite	Expression & Information
Access to Foreign Newspapers	Expression & Information
Assembly	Association, Assembly, & Civil Society
Disappearances	Disapperances, Conflicts, and Terrorism
Divorce	Identity & Relationships
Domestic Movement	Movement
Foreign Movement	Movement
Harassment and Physical Hostilities	Religion
Inheritance Rights	Women's Security & Safety
Legal Gender	Identity & Relationships
Legal and Regulatory Restrictions	Religion
Organised Conflicts	Disapperances, Conflicts, and Terrorism
State Control over Internet Access	Expression & Information
Women's Movement	Movement

This imputed hundreds of missing values. Notably all the predicted values were not imputed. Only sub-categories were imputed.

3.2.6 Step 6: Drop Brunei, Belize and Bahamas

All of these countries have missing CATO data and cannot be reliably imputed because the substitutes are themselves absent also.

3.2.7 Step 7: Round 2 of Imputation

Upon merging there were missing values from the following categories also that required imputation. They were imputed as follows from the larger category the sub-categories were part of as was done in Round 1 of Imputation:

Table 3: Round 2 Imputation

Category	Where Score Copied From
Military interference in rule of law and politics	Legal System & Property Rights
Administrative requirements	Business regulations
Assembly	Association, Assembly, & Civil Society
Capital controls	Controls of the movement of capital and people
Centralized collective bargaining	Labor market regulations
Civil Justice	Rule of Law
Civil Society Entry and Exit	Association, Assembly, & Civil Society

Continued on next page

Table 3 continued from previous page

Category	Where Score Copied From
Civil Society Repression	Association, Assembly, & Civil Society
Compliance costs of importing and exporting	Regulatory trade barriers
Criminal Justice	Rule of Law
Extra payments/bribes/favoritism	Business regulations
Financial Openness	Controls of the movement of capital and people
Freedom of Religion	Religion
Freedom to Form Political Parties	Association, Assembly, & Civil Society
Government enterprises and investment	Size of Government
Harassment and Physical Hostilities	Religion
Hiring and firing regulations	Labor market regulations
Hiring regulations and minimum wage	Labor market regulations
Hours Regulations	Labor market regulations
Legal and Regulatory Restrictions	Religion
Mandated cost of worker dismissal	Labor market regulations
Non-tariff trade barriers	Regulatory trade barriers
Opposition Parties Autonomy	Association, Assembly, & Civil Society
Organised Conflicts	Disapperances, Conflicts, and Terrorism
Ownership of banks	Credit market regulations
Procedural Justice	Rule of Law
Reliability of police	Legal System & Property Rights
Religious Organization Suppression	Religion
State ownership of assets	Size of Government
Transfers and subsidies	Size of Government

3.2.8 Step 8: Drop some irrelevant columns

The following columns were deemed irrelevant to the study and were dropped as they're raw data of investments or tax information which are not expected to have any reliable correlation:

1. Revenue from trade taxes (% of trade sector)
2. Revenue from trade taxes (% of trade sector) DATA
3. Government enterprises and investment DATA
4. Transfers and subsidies DATA

At this point the data is now clean for model creation and can be joined with the Kaggle Dataset's computation of the mean for every country that has at least 100 entries for the given year based on an ISO2 to ISO3 map.

4 The Model

The model attempts to predict a country's scores for the following measures based on the CATO Human Freedom Index data:

1. Rule of Law

2. Security & Safety
3. Movement
4. Religion
5. Association
6. Assembly & Civil Society
7. Expression & Information
8. Identity & Relationships
9. Size of Government
10. Legal System & Property Rights
11. Sound Money
12. Ownership of banks
13. Regulation

Only these could be chosen due to time and hardware limitations as attempting to do all resulted in the computer running out of memory. The model itself is a MultiOutputRegressor on a StackingRegressor that uses five estimators: CatBoost, XGBoost, Ridge, KNearestNeighbors and MLPRegressor. It then uses Random Search to try to get the ideal values for each of these datapoints as they are not necessarily correlated. As this is very time-consuming already further optimizations were not added due to speed constraints. Metrics were then printed and KFold cross-validation is performed.

5 Results

The following metrics were taken for each of the categories on a run: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Max Error (MaxErr), the Coefficient of Determination (R2). Note, results may differ slightly on another run as RandomSearch was used making results non-deterministic.

The most important statistic here is the R2 value. In the social and behavioral science as this paper is, "R^2 between 0.10-0.30 is often acceptable, especially when predictors show statistical significance. Values from 0.30-0.50 are considered very good." (PromptLayer, 2025) Every single one was acceptable and many were very far above acceptable. Assuming values above 0.5 are considered excellent then 8(!) were excellent, 3 very good and 1 acceptable. This is overwhelmingly strong evidence of correlation and is as strong as could possibly be expected in the social sciences. This preliminary research constitutes solid grounds for further investigation, sampling and studying by providing solid statistical evidence of correlation. Results were not influenced by the imputation as none of the predicted values had any imputations.

As mentioned above, KFold validation was performed and, as can be seen from the table below the predictive power is quite strong and these results are not an extreme outlier:

Even the means for each category is still well within the requirements for significance stipulated above.

	MAE	MSE	RMSE	MaxErr	R2
Rule of Law	0.5562	0.6213	0.7883	3.2584	0.7699
Security & Safety	0.6676	0.8846	0.9405	2.9884	0.6079
Movement	1.1714	2.6980	1.6426	5.5150	0.5709
Religion	0.6193	0.6630	0.8142	2.3833	0.7894
Association, Assembly & Civil Society	0.9369	1.8176	1.3482	4.3161	0.6691
Expression & Information	0.5185	0.5306	0.7284	2.1327	0.4984
Identity & Relationships	1.1747	2.2248	1.4916	3.4918	0.7699
Size of Government	0.7541	0.9179	0.9581	2.3417	0.1071
Legal System & Property Rights	0.4515	0.3514	0.5928	1.7316	0.7962
Sound Money	0.6142	0.6572	0.8107	2.9881	0.4283
Ownership of banks	1.7323	5.0839	2.2547	5.9717	0.3481
Regulation	0.4756	0.3957	0.6291	1.8340	0.6440

Table 4: Metrics for Each Category

Category	Scores from 5 folds	Mean of 5 scores
Rule of Law	0.56233244, 0.68012421, 0.76908571, 0.75707897, 0.68882375	0.6915
Security & Safety	0.61317492, 0.65410343, 0.7089709 , 0.61399038, 0.69777241	0.6576
Movement	0.59191506, 0.56686473, 0.57926578, 0.46320227, 0.5265503	0.5456
Religion	0.78770557,0.65882621, 0.72599501,0.58582493, 0.66637669	0.6849
Association, Assembly & Civil Society	0.59247607, 0.659207, 0.63650812, 0.6518365, 0.59355038	0.6267
Expression & Information	0.58723387, 0.58289108, 0.43260333, 0.60377456, 0.49254167	0.5398
Identity & Relationships	0.60431072, 0.70023063, 0.75062931,0.69626142, 0.68081377	0.6864
Size of Government	0.18554572, 0.358211, 0.39958934, 0.38156488, 0.43944153	0.3529
Legal System & Property Rights	0.63602977,0.73225037, 0.79553937,0.75925529, 0.77139422	0.7389
Sound Money	0.21402581,0.1698962, -0.12876422, 0.35795591, 0.44735392	0.2121
Ownership of banks	0.40925751,0.41654981, 0.38350327,0.38256657 ,0.46943092	0.4123
Regulation	0.67949446, 0.54657699 0.54283277, 0.6682811, 0.47775278	0.5830

Table 5: KFold Validation Scores

6 Limitations

Due to the limitations of the dataset national representativeness compared to those who'd be inclined to take personality quizzes cannot be ascertained definitively but the wildly differing results certainly indicates differences. The direction of causation can also not be ascertained from these results, merely the correlation. While the CATO Index is an ideologically loaded score it likely is an accurate measure of the degree of negative liberty present in an area in a country as it is very dear to Libertarians and they are likely to give an accurate measure of it.

7 Contact

The author's contact information can be found at his website: therootuser.dev

8 Citations

Amati, G. (2025, September 16). What is a Good R-Squared Value?. PromptLayer. https://blog.promptlayer.com/is-a-good-r-squared-value/

Peabody, D. (1985). National characteristics. Cambridge, MA: Cambridge University Press.

Isaiah Berlin, "Two Concepts of Liberty," Four Essays On Liberty, (Oxford, England: Oxford University Press, 1969), p. 118-172

Paula Reiner, Aristotle on Personality and some Implications for Friendship, Ancient Philosophy Volume 11, Issue 1, Spring 1991 Pages 67-84 fn. 1 https://doi.org/10.5840/ancientphil199111135

Sussex Publishers, LLC. "Personality Tests." Psychology Today. n.d. https://www.psychologytoday.com/us/ba tests#: :text=Any%20personality%20test%20can%20be,most%20reliable%20and%20valid%20results.

Bojan Tunguz, Big Five Personality Test (Kaggle, n.d.), https://www.kaggle.com/datasets/tunguz/big-five-personality-test

Vásquez, Ian, and Fred McMahon. Human Freedom Index 2020 Data. XLSX dataset. Cato Institute & Fraser Institute, 2020. https://www.cato.org/human-freedom-index/2020.