Citadel East Regional Datathon 2021

Owen Mo
Carnegie Mellon University
changhem@andrew.cmu.edu

Qiyun Chen
Carnegie Mellon University
qiyunc@andrew.cmu.edu

Cindy Zeng
Carnegie Mellon University
siqiz@andrew.cmu.edu

1 Executive Summary

1.1 Problem Statement

We are interested in investigating what factors may affect the percentage of people who smoke in a country. We mainly look at three categories of potential factors: governmental regulations, geographical traits, socioeconomic status. Furthermore, we are interested in building models to predict the percentage of people who smoke in a country. With the combined results, we hope to provide insights on whether certain smoking-related policies are effective, and what features policy-makers should pay attention when they're trying to estimate and control the percentage of people who smoke in a country.

1.2 Key Conclusions

When we naively look at governmental regulations and tobacco usage across the world, we make the counter-intuitive observation that more regulations is correlated with a higher tobacco usage, which can be simply explained by the increased willingness of a country to impose regulations when tobacco usage is high.

If we look at the tobacco usage in each region individually, we find that the effectiveness of government interventions vary widely depending on the region of the world. As a result, it is crucial for key actors to consider the cultural and socioeconomic realities of their region when creating policies.

Over the years from 2000 to 2018, within each geographical group, there is a decreasing trend in the mean percentage of smokers. Country-wise, most country show decreasing trends, while a few actually show increasing trends.

Additionally, the mean percentage of smokers at different geographical locations tend to differ from each other. For example, in 2018, out of all the geographical locations, South-East Asia has the highest mean percentage of smokers, while Africa has the lowest mean percentage of smokers. The differences in different geographical parts of the world indicate that countries who are geographically closer may have a closer percentage of smokers. We suspect that the countries closer to each other may share similar culture as regards to smoking. For example, South-East Asian countries might have a heavier smoking culture than other parts of the world.

Finally, we build several models to predict a country's percentage of smokers given its current socioeconomics and geographical features. We choose Decision Tree Regressor Model as our final model, and we found that the most important features selected by the model are geographical features (which part of the world a country belongs to), population of a country, and happiness related features. We suggest policy makers pay extra attention to those features when inspecting ways to control a country's percentage of smokers.

2 Technical Exposition

2.1 Exploratory Data Analysis

We started our analysis by conducting exploratory data analysis on the datasets.

2.1.1 Governmental Regulation Datasets

We investigate the effect of governmental regulations, namely excise tax rate on cigarettes and ban on cigarette advertisement, on the tobacco usage in different countries. The information on interventions were extracted from the `stop_smoking` dataset. In order to measure the effectiveness of the measures, we make use of the information on the percentage of population who consumes tobacco products in the `tobacco_use_ww` dataset.

In order to quantify the effectiveness of governmental measures, we can look at the percentage of the population who consumes tobacco and compare this quantity among countries with different levels of regulation. However, directly comparing these values may not yield insightful results since this doesn't take into account other factors, such as culture and decline in supply, which may affect each country in different manners. As a result, we propose 2 ways of normalizing this data: time normalization and geospatial normalization.

Time normalization is done by taking the tobacco usage at a certain year and subtracting the value from a previous year. In our case, we take the most recent data from 2018 and subtract the baseline value from 2000, which is the oldest data available.

Another factor which can affect tobacco usage is culture, which is related to the geospatial location of a country. To mitigate the influence of this factor, geospatial normalization uses the average tobacco usage from an area as the baseline and subtract the tobacco usage of each country by its area's baseline value. This way, we can compare the smoking rate of a country relative to nearby countries.

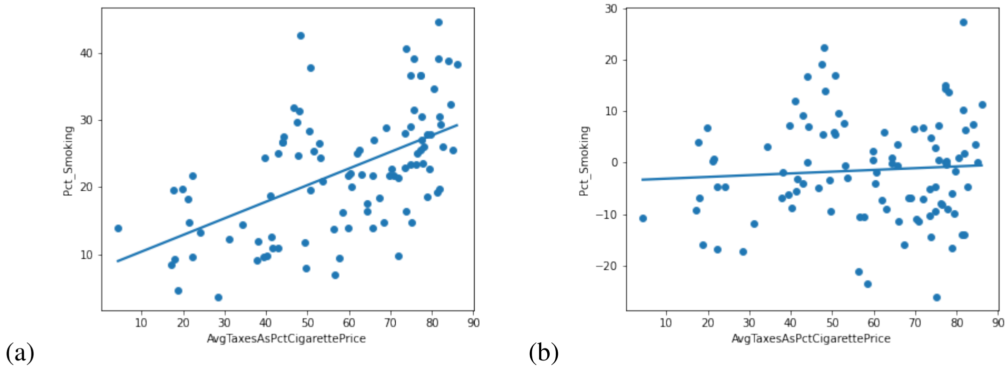


Figure 1: Tobacco usage and average tax rate on cigarettes. Each point represents a country. The tobacco usage data was taken from 2018 and the tax rate was taken from 2014. **(b)**: Unnormalized tobacco usage. **(b)** Tobacco usage normalized by area.

Observation 1: Higher tax rates on cigarettes is correlated with a higher tobacco usage.

In Figure 1a, we see that when we don't normalize tobacco usage, there is a positive correlation between the cigarette tax rate and tobacco usage ($r = 0.55$, p-value$= 4.1 \times 10^{-9}$). While this may be a counter-intuitive finding, it can be explained by the hypothesis that countries with higher smoking rate are more likely to implement more drastic measures. There might also be other confounding variables which are at play.

In order to remove some other factors such as culture and geolocation, we perform area normalization on tobacco usage. After normalizing the data (Figure 1b), the correlation between tobacco usage and cigarette tax rate disappears ($r = 0.07$, p-value$= 0.50$). This data seems to suggest that cigarette tax rate has little to no effect on tobacco consumption worldwide. However, we had a hypothesis these variables may have different interactions for each region, which are lost when looking at the larger picture. Therefore, we also looked at the relationship between tax rate and tobacco usage in each region.

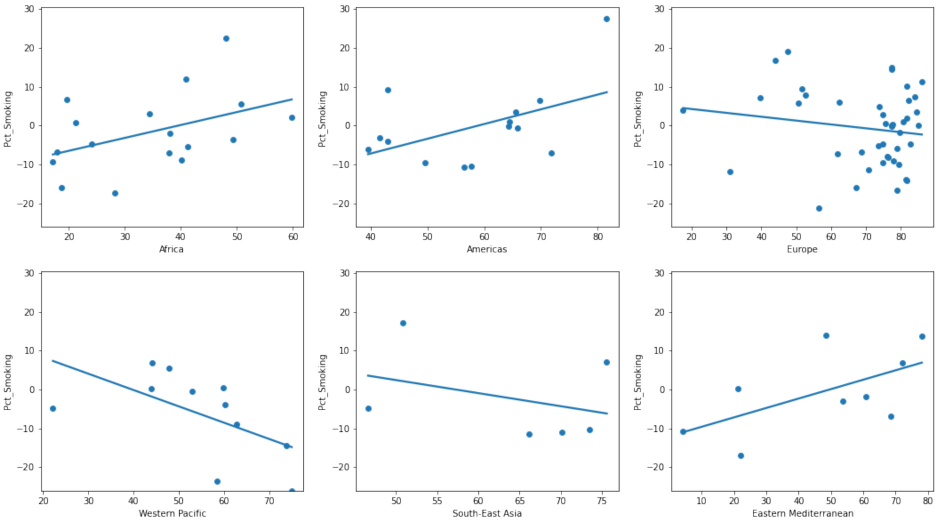


Figure 2: Area normalized tobacco usage and cigarette tax rate for each major region.

Observation 2: Government interventions have different effects in different areas of the world.

In figure 2, we see that if we look at each major region individually, we can observe a correlation between tobacco usage and cigarette tax rates. In addition, the relationship between the two variables are different depending on the region, as we hypothesized. Specifically, in Western Pacific and South-East Asia, there is a negative correlation as opposed to a positive correlation found in the rest of the world. In figure 3, we can make a similar observation that the effect of advertisement ban vary across regions.

This is an interesting finding that government interventions seem to have different effects depending on the region. We hypothesize that this may have to do with culture, willingness to listen to the government and poverty rates in different regions.

2.1.2 Geographical Traits Datasets

We are interested in the percentage of the whole population aged 15 years and over who currently use any tobacco product, thus we looked at the *Value* column in the *tobacco_use_ww* dataset for which *Gender* is equal to *Both Sexes*.

We first want to investigate the overall trend of percentage of smokers, thus we plot the mean of percentage of smokers per geographical group across the years.

Observation 3: Within each geographical group, there is a decreasing trend in the mean percentage of smokers (Figure 4).

Next, we want to look more closely at each country's trend of percentage of smokers per country, so we plot *Value* against *Year* per country.

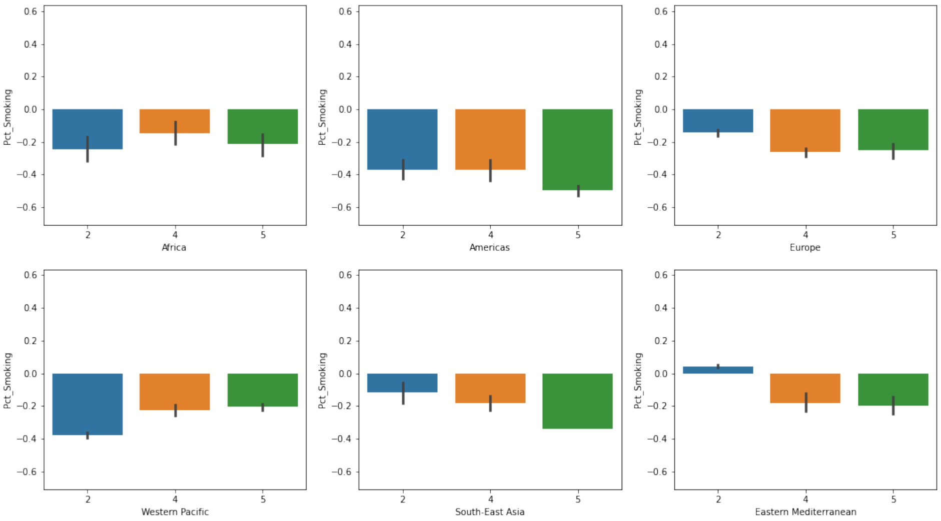


Figure 3: Time normalized tobacco usage and enforcement of ban on tobacco advertisement for each major region.

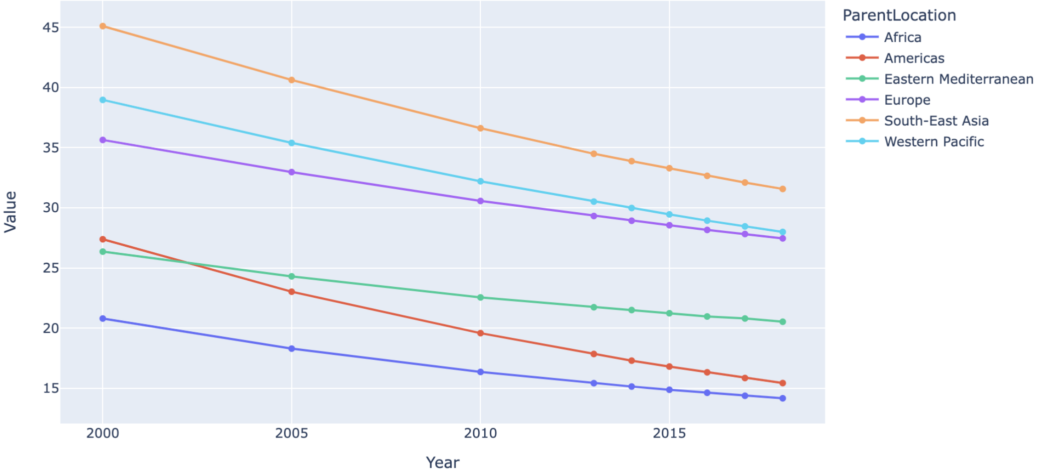
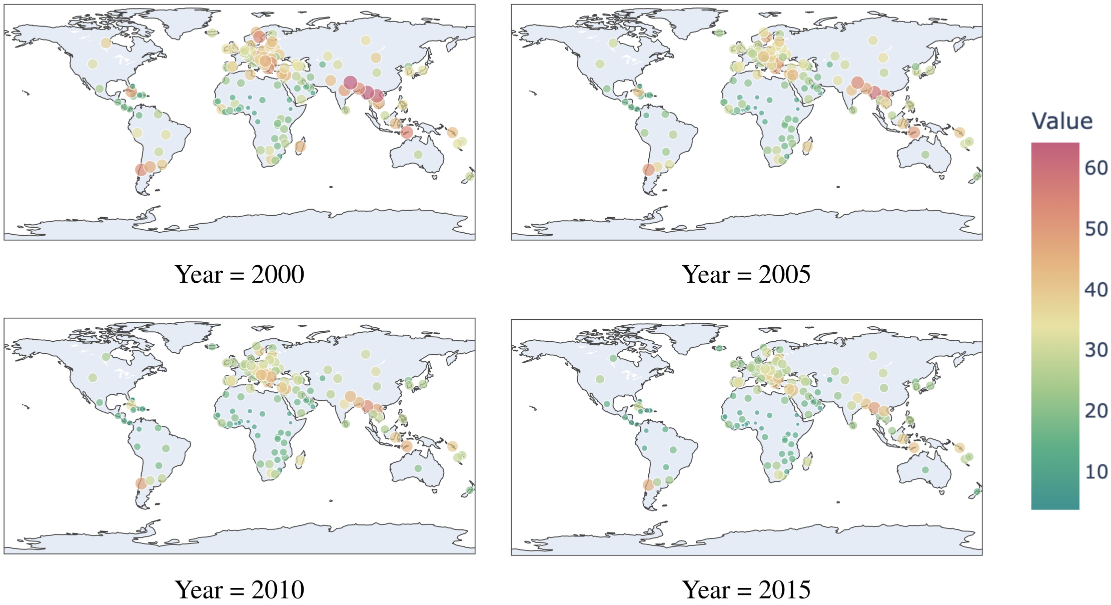


Figure 4: Mean Percentage of Smokers Per GeoLocation

Observation 4: We saw from that although most countries show a decrease in their percentage of smokers, however, a few actually show increase trends over the years (The plot is not shown in the paper since there're too many countries, but included in code).

To provide a better illustration than mere line plot, we created an interactive map that shows each country's trend of percentage of smokers across the years. Each data point on the world map represents the centroid location of a country. The size and the color of the point represents the scale of the percentage of smokers in a given year. Value represents the percent out of 100 of the smokers in a particular country. We included maps below for the year of 2000, 2005, 2010, and 2015.

We notice a decreasing trend over the years, we also notice that originally in the Year 2000, countries located in different geographical tend to show similar percentage of smokers (Figure 5). For example, the percentage of smokers at African countries appear to be relatively low compared to the worldwide



Year = 2000

Year = 2005

Year = 2010

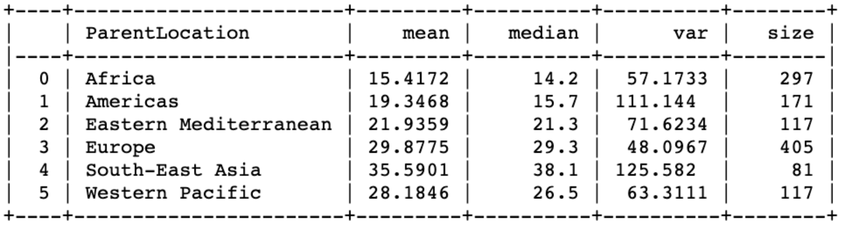
Year = 2015

Figure 5: Worldwide Percentage of Smokers Per Country

percentage of smokers, while Asia countries tend to have more percentage of smokers to start with in 2000.

Based on this observation, we venture further to test if the geographical traits of a country may bring predictive power in regards to value, specifically we investigate the *ParentLocation* feature in our dataset.

We first perform preliminary analysis on the different groups for their mean, median, variance, and group sizes for a given year (here we included the results for the year 2018). We find that we have very different group sizes across groups.



```
+----+-----------------------+---------+----------+----------+--------+
|    | ParentLocation        |    mean |   median |      var |   size |
|----+-----------------------+---------+----------+----------+--------|
|  0 | Africa                | 15.4172 |     14.2 |  57.1733 |    297 |
|  1 | Americas              | 19.3468 |     15.7 | 111.144  |    171 |
|  2 | Eastern Mediterranean | 21.9359 |     21.3 |  71.6234 |    117 |
|  3 | Europe                | 29.8775 |     29.3 |  48.0967 |    405 |
|  4 | South-East Asia       | 35.5901 |     38.1 | 125.582  |     81 |
|  5 | Western Pacific       | 28.1846 |     26.5 |  63.3111 |    117 |
+----+-----------------------+---------+----------+----------+--------+
```

Figure 6: Histogram of *Value*

We suspect that the data across groups might have different variances. In order to validate our guess, we first draw out the boxplot and swarmplot of *Value* per *ParentLocation*, and we do observe traits of violations to the Homogeneity of Variance. We further prove this by running the Levene's test upon the data. We receive a p-value of $4.84e - 09 < 0.05$, thus we reject the Null Hypothesis that assumes equal variance across groups, and conclude that we have different variances across our groups.

Thus, in order to test our hypothesis, we choose the Welch's Analysis of Variance (ANOVA) test. We do not choose the Classic ANOVA test here because the Homogeneity of Variance assumption in the Classic ANOVA test is violated. We choose the Welch's ANOVA test because it is an alternative to the Classical ANOVA that could also help us find out if the differences between groups of data are statistically significant, given that the data may violate the Homogeneity assumption. We define the

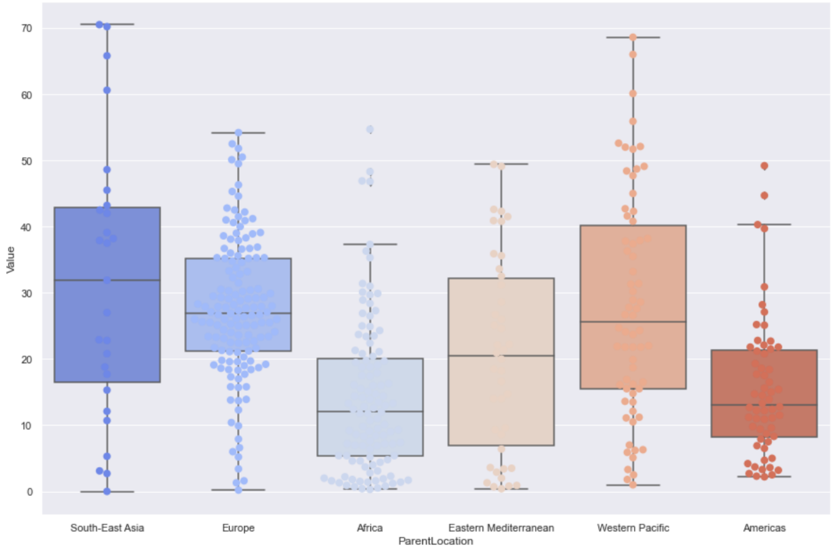


Figure 7: Boxplot and Swarmplot of Value Per *ParentLocation*

following hypotheses:

$$\begin{cases} \text{Null Hypothesis} & \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 \\ \text{Alternative Hypothesis} & \text{Not all } \mu_i s \text{ are the same.} \end{cases} \tag{1}$$

Each μ_i represents the mean of value per *ParentLocation* group. Recall that we have 6 unique *ParentLocation* group: *Africa*, *America*, *Eastern Mediterranean*, *Europe*, *South-East Asia*, and *Western Pacific*.

We first inspect Welch's ANOVA assumptions. The first assumption is that the population from which we draw the samples should be normally distributed. In order to test this, we plot the histogram of *Value* worldwide, and see that it is approximately normally distributed (Figure 8).

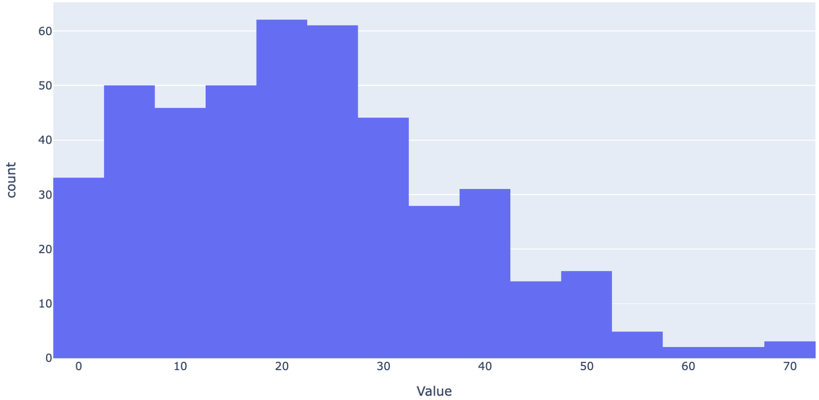


Figure 8: Histogram of *Value*

Next, we test for Welch's ANOVA's second assumption: independence of data. We assume that we have independent data since the data are related to the statistics from different countries.

Finally, we run Welch's test on our dataset, and we have a p-value of $3.527168e - 17 < 0.05$, thus we reject the null hypothesis and conclude that not all $\mu_i s$ are the same. The statistical difference in distribution of *Value* across groups indicates that *ParentLocation* may predictive power when our target variable is *Value*.

Table 1: External Datasets Summary

Type	Name	Source
Economy	Consumer Price Index	World Bank
	Employment Rate for Age 15+	World Bank
	Tax on Goods and Service	World Bank
	National Income Per Capita	World Bank
	Exchange Rate	World Bank
Social	Happiness Index	World Happiness Report
	Secondary School Enrollment Rate	World Bank
Environment	Agricultural Land	World Bank
	Greenhouse Gas Emissions	World Bank
Demographics	Population	World Bank

Observation 5: The differences between some of the means of the percentage of smokers amongst the different *ParentLocation*s are statistically significant.

2.1.3 Socioeconomic Status Datasets

In order to reflect geological differences, inspired from the report from Australian Department of Health (1), we selected several indicators from other data source. All external datasets are briefly summarized in Table 1.

From (1), the financial status, usually influenced by the economy of a country, is an enabler for smoking, so Consumer Price Index, Employment Rate, Tax, National Income Per Capita, and Exchange Rate are considered. Besides, smoking is often seen as a kind of entertainment, so we selected several indicators in World Happiness Report 2021 (2) as one of the external data source to measure people's mental state. Another social factor that may influence smoking is education. By (1), many people learn to smoke because of peer examples. We assume some people develop smoking habits during teenage years when they are supposed to be in secondary schools, so secondary school enrollment rate is also another social factor we include. Futhermore, the dynamics of a market are not only impacted by consumers, but also producers as well. Therefore, environmental factors are integrated in our dataset: for example, agricultural land may influence the production of tobacco and thus raising tobacco's price, and similarly, the emission of greenhouse gas may accelerate climate change, which may indirectly affect the price of tobacco leaves. Population is also included for data normalization purpose. Except for happiness indicators, the source of most data comes from World Bank(`https://www.worldbank.org/en/home`), one of the world's largest institution. The raw annual data from World Bank includes indicators from 1960 to 2020 for 266 different countries, and the data from World Happiness Report include happiness features from year 2008 to 2020 for 166 countries.

2.2 Model Building

2.2.1 Workflow Summary

Our workflow are as follows: we first perform preprocessing and select features based on correlation with target variable and model assumptions. Next, we build different models for our prediction task. We start with Linear Regression and then LASSO/Ridge Linear Regression Model. We also experimented with constructing features that represent yearly changes. Then, we proceed with Tree-Based Regressor Model. In the end, we compare the model performances and interpret the results.

2.2.2 Data Preprocessing

We first one-hot encode the *ParentLocation* categorical variables, and we selected the data from 2013-2018 for further model building and feature construction. We apply cube root transformation for some of the right-skewed data. Next, we apply the MinMax Scaler to all our numerical data, tranforming them to be values in between [0, 1], and fill in the NA numerical values with the KNN Imputer. Next, we consolidate a few highly-correlated raw features via PCA. More details and reasoning behind those procedures are shown in the next section.

2.2.3 Feature Selection

We further inspect the relationship between each of our explanatory variables with our target variable *Value*. We plot a scatter plot along with a line of best fit between each explanatory variable and the target variable *Value*. We observe that the variables *cpi, exchange, gdp, green, income* and *pop* tend to be highly right skewed. Thus we decide to apply cube root transformation on them. The reason we chose cube root over log or square root transformation is that there are also negative values in our original data. We plot the scatter plot along with the line of best fit with the transformed variables and remaining explanatory variables (Figure 9). We observe that some variables tend to have a weak correlation with *Value*, such as cube-root transformed *cpi*, cube-root transformed *exchange*, cube-root transformed *green*, and *Generosity*, so we decide not to use those features as they may not bring too much predictive power, especially in linear models.

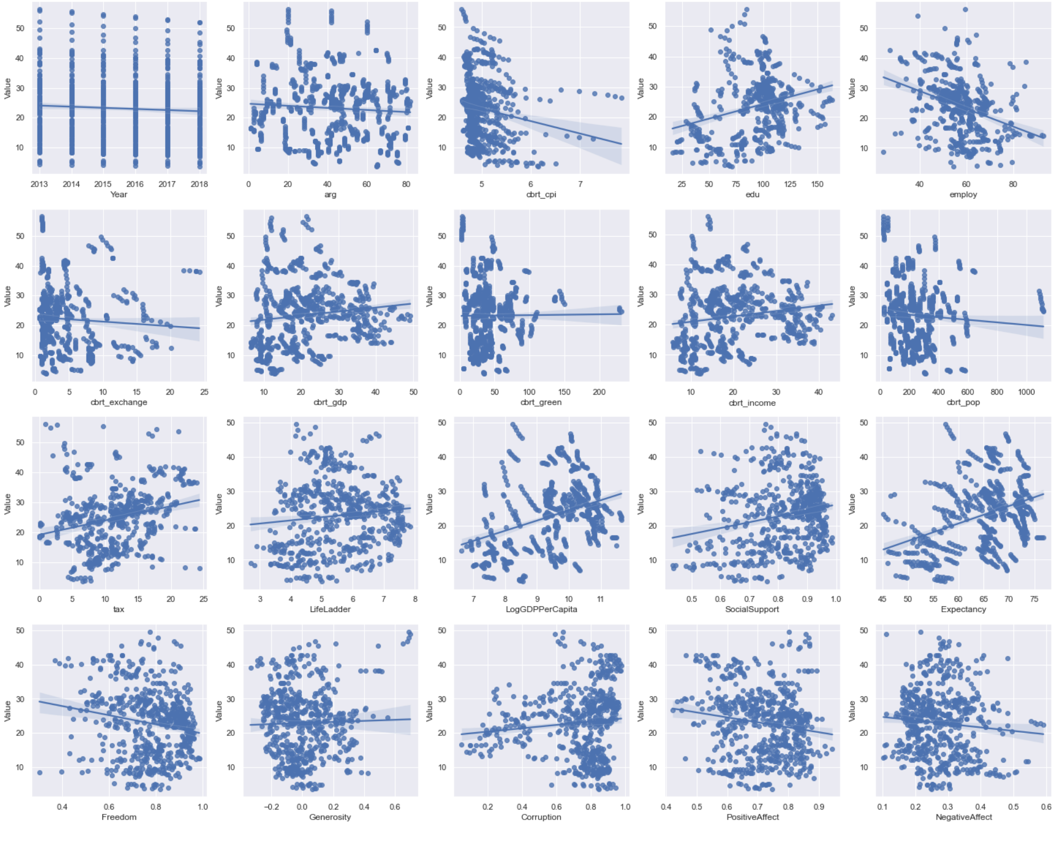


Figure 9: Scatterplot between *Value* and Explanatory Variables

We also want to check for multicollinearity issues in between the pairs of explanatory variables, thus we draw out a correlation plot (Figure 10). We observe that the variables related to the Happiness Index are highly correlated with one another, so we utilize PCA to consolidate *LifeLadder, SocialSupport, Expectancy, Freedom, PositiveAffect, NegativeAffect*. We did not choose to completely disregard any one variable, as we believe each variable within this group could bring different insights to the prediction, although their raw features may be correlated. We solve this by using PCA to formulate three PCA vectors from all the raw features.

Then, because we observe that different numerical columns tend to have very different ranges, thus, in order to make the model more accurate and robust, we perform MinMax Scaling to transform them to values $\in [0, 1]$.

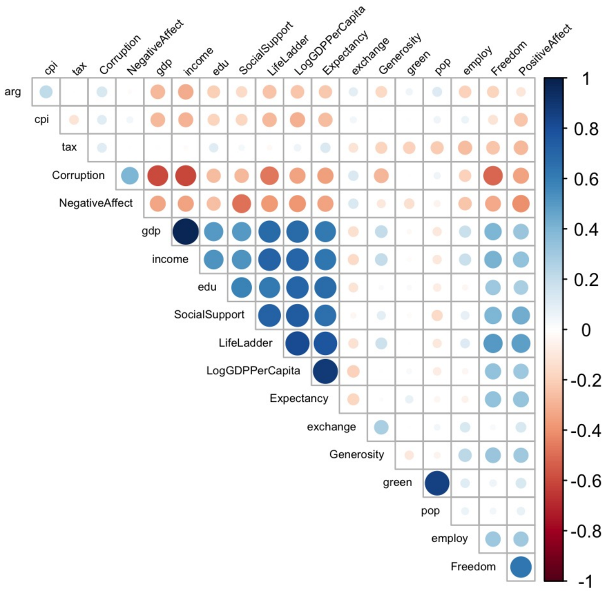


Figure 10: Correlation Plot in between Explanatory Variables

2.2.4 Multiple Linear Regression

Now that we have inspected all our explanatory and target variables, we want to proceed with building a model that could help us predict the *Value* of a country. Our model is formulated as follows:

$$Value = \beta_0 + \beta_1 * edu + \beta_2 * LogGDPPerCapita + \beta_3 * CubeRoot(income) + \beta_4 * pop + \beta_5 * tax$$

$$+\beta_6 * Africa + \beta_7 * Americas + \beta_8 * EasternMediterranean + \beta_9 * Europe$$

$$+\beta_{10} * SouthEastAsia + \beta_{11} * WesternPacific + \beta_{12} * PCA_1 + \beta_{13} * PCA_2 + \beta_{14} * PCA_3$$

Before start, we check for Linear Regression's assumptions. We have checked linearity between expectation of Y and explanatory variables via our previous scatterplots. We inspect each variable's homoscedasticity, and we decide not to use *Corruption* because its variance appear to increase with the increase of x. We assume independence of data points based on our data sources. We check that the data is approximately normally distributed by plotting QQ-plot for the numerical variables. We see that apart from some potential violations at the end, the variables approximately fall about a straight line, indicating normality of the distribution.

Given that the assumptions are met, we proceed with our modelling, we perform train test split and record the train accuracy and validation accuracy, and we repeat the experiment for 20 times to get the average performance. We evaluate our model by R2 score, which has the formula $R^2 = 1 - \frac{RSS}{TSS}$, where RSS is the sum of squares of residuals and TSS is the total sum of squares. R^2 measures how much variation of our target variable *Value* is explained by our explanatory variables. From the Multiple Linear Regression Model, we achieve a $R^2 = 0.4972$ for the training data and $R^2 = 0.4856$ for the validation data, which gives us a relatively weak model for our prediction purposes. We reflect upon our methodology, and we see two places for potential improvements. One is that we could potentially construct our original features to be relative features based on comparison with the previous year (explained in the later section). Another point is that we inspected the learned Linear Regression Model weights, and saw that the naive LR model gave too much weights to the *ParentLocation* feature (Table 2). Thus we iterate our model.

Table 2: Linear Model Coefficient Comparison

Feature	naive	lasso	ridge
edu	-5.457001541	-3.059784219	-4.955030867
employ	-8.303301096	-7.7523456	-7.900076791
LogGDPPerCapita	23.00467417	18.25057105	17.33957751
cbrt_income	-10.08771462	-7.791431335	-7.593392482
pop	-3.638464947	-2.191094247	-3.059846942
tax	2.073247596	2.467374893	1.919026692
Africa	2.75578E+13	-12.06358921	-9.171297383
Americas	2.75578E+13	-8.754987234	-5.86516594
Eastern Mediterranean	2.75578E+13	-6.59352319	-3.54633776
Europe	2.75578E+13	0.364947559	3.593127075
South-East Asia	2.75578E+13	6.776881065	9.921029424
Western Pacific	2.75578E+13	1.818706416	5.068644584
pca_1	7.838242674	7.134338475	6.400790136
pca_2	1.01068702	2.10837922	2.958515701
pca_3	-8.302756429	-8.169299018	-7.64245543

2.2.5 Improving Linear Regression Models

To solve the problem presented in the previous section, we use Lasso and Ridge to penalize the extremely large coefficients. Lasso usually works well when we have only a few very significant features, and ridge works better if there are many large similar coefficients. Given the experimental results from previous section, we saw high weights on *ParentLocation* so we expect Lasso might me more helpful. The coefficients of all linear models are shown in Table 2. The result validates our expectation. Training R^2 of the Lasso model is 0.4917 and testing R^2 is 0.5004; training R^2 of the Ridge model is 0.4985 and testing R^2 is 0.4743. Compared to naive LR, we can't see a huge improvement on the evaluation metrics R^2, but as expected, these two models provide a more balanced coefficients. Two new models don't emphasize much on only geological features, preserving the explanatory power of socioeconomic factors while maintaining a comparable accuralcy. From the coefficients of the regularized models, $LogGDPPerCapita, tax, Europe, South-EastAsia, WesternPacific, pca_1, pca_2$ are positively related to smoking usage, while others are negatively related. Although here the tax is the tax on all goods and service, instead of the tax on cigarette, this align with our previous observation that higher tax rate infers higher usage of tobacco.

Another way we tried for improvement is feature engineering. From Figure 5, we suspected there are some time-related features affecting smoking usage percentage. A general declining trend is shown, though different locations have inconsistent decreasing speed. However, unlike other feature, *year* is tricky to deal with. Setting it as a categorical variable is unfair since not all *year*s are completely independent, while it is also problematic to assign continuous values for *year*s by saying which *year* value is larger. This hypothesis is validated through p-values of *years* in a naive linear model. If we set *year* as a categorical variable, p-values are around 0.362 for each year, and if as a continuous variable, p-value is 0.051. Both are not significant enough to show non-zero relationship. To solve this problem, we engineered some new features to encode the time information in our old features.

$$X_{location\_prog,i} = (X_{location,i} - X_{location,i-1})/X_{location,i-1}$$

$$X_{world\_prog,i} = (\overline{X_{location,i}} - \overline{X_{location,i-1}})/\overline{X_{location,i-1}}$$

$$X_{relative\_prog,i} = X_{location\_prog,i} - X_{world\_prog,i}$$

$X_{location\_prog,i}$ is the progression rate for a continuous feature X in the original dataset in *year* $= i$. $X_{world\_prog,i}$ is the progression rate of the feature mean in *year* $= i$. Thus we created new features, $X_{relative\_prog,i}$, that reflect the relative development speed of a country compared to world average for all old features. We still use LASSO to incorporate these features to reduce the impact of collinearity between newly added features and the old ones.

From these experiment, we can conclude: [1] including time information directly is unhelpful to explain the smoking usage percentage. [2] Old features are sufficient. *year*s may overlap with the explanatory power of other features from external datasets. For example, population (*pop*) grows annually. However, compared with old features, new features seem to have limiting explanatory power. So we keep on building new models that can help to extract non-linearity relationship.

2.2.6 Tree-Based Models

The Linear Regression based model have stronger assumptions on the linearity between the explanatory and target variables, thus, we want to explore Tree-Based models for our problem setting, as Tree-Based models tend to be more robust to outliers, and may be able to capture more relationship between the variables compared to the linear model.

Since our target variable *Value* is a continuous variable, we proceed by using Decision Tree Regression. We perform 5-fold Cross Validation to choose the best hyperparameter for the depth of our Decision Tree Regressor. We choose depth $= 9$ to be our optimal k to avoid underfitting and overfitting, since the mean validation R squared seem to converge around depth $= 9$.

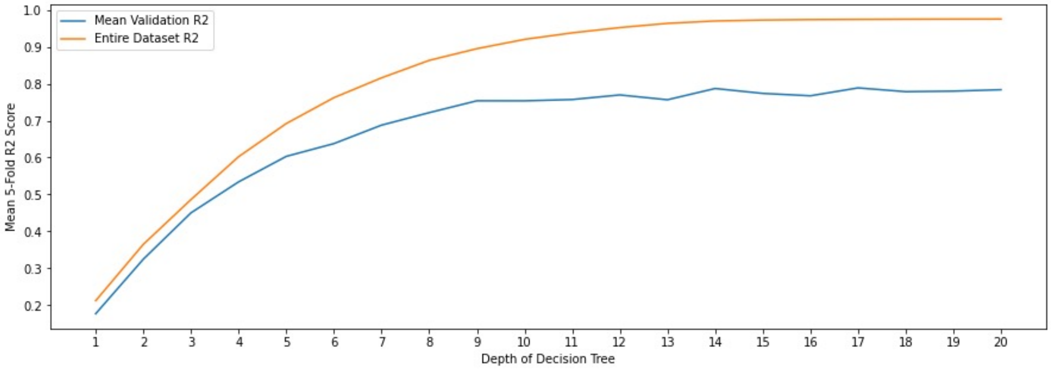


Figure 11: Accuracy v.s. Depth of Decision Tree Regressor

We have a resulting mean validation R2 score of 0.7536 and overall R2 score of 0.8946 with our Decision Tree Regressor Model of Depth 9. It seems that our Decision Tree Regressor Model outperforms the Linear Regression Based Models in the previous section, thus we choose Decision Tree to be our final model.

2.2.7 Features with Maximum Prediction Power

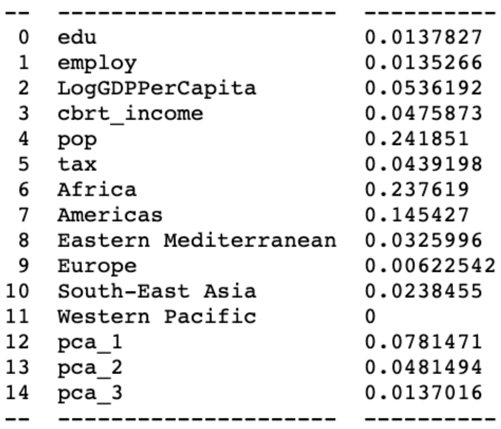
We are also interested in the features that contributed the most to our Decision Tree Regressor, thus we investigate each explanatory variable's feature importance (Figure 12).

It seems that the *ParentLocation* variables, population of a country, and happiness related features carry the most weight when predicting the percentage of people who smoke in a country.

2.3 Reflections

One thing we're concerned about our Decision Tree Regression Model is that it might potentially formulate a 1-1 correspondence between the features and our target value, given that our dataset is relatively small. If we have more time, we plan on performing pruning algorithms on our model to produce a more generalized model for unseen data.

Another thing we could look more into is the relative features we constructed. It seems that from our experiments, training a model only using the relative numerical features did not perform as well as using only the original features. However, we're interested in seeing the model performances by



```
--  --------------------  ----------
 0  edu                   0.0137827
 1  employ                0.0135266
 2  LogGDPPerCapita       0.0536192
 3  cbrt_income           0.0475873
 4  pop                   0.241851
 5  tax                   0.0439198
 6  Africa                0.237619
 7  Americas              0.145427
 8  Eastern Mediterranean  0.0325996
 9  Europe                0.00622542
10  South-East Asia       0.0238455
11  Western Pacific       0
12  pca_1                 0.0781471
13  pca_2                 0.0481494
14  pca_3                 0.0137016
--  --------------------  ----------
```

Figure 12: Decision Tree Feature Importance

potentially combining some of the original features and some of the relative features we constructed, in which case we need to check again for multicollinearity issues.

Thirdly, if the data sets are larger with more features, we are invested in constructing a deep learning model and investigating if there are more complicated relationships that could be captured. We could potentially create feature embedding layers to our data sets, and experiment with Multilayer Perceptron Layers and Convolutional Neural Network structures.

References

[1] Australian Government, Department of Health, “Factors influencing smoking levels among high smoking prevalence groups.”

[2] R. L. J. S. Helliwell, John F. and J.-E. D. Neve, “World happiness report 2021,” 2021.

Table 3: External feature definition

Name	Unit	Description
arg	% of land area	Area of arable land.
cpi	(2010=100)	Consumer Price Index. Its value is the cost of the current year's item over the cost of the base year's item multiplied by 100. The base is 2010. Reflects the relative change in cost throughout the years.
edu	% gross enrollment ratio	Secondary School enrollment Percentage in gross enrollment. The gross enrollment ratio is total enrollment, regardless of age, dividing the population of the age group in the level of education.
employ	% population	Employment rate. The proportion of a country's employed population for people aged 15+.
exchange	USD	Exchange rate. Units of a country's currency to 1 USD.
green	kt of CO2 equivalent	Green house gas emissions.
income	USD	National Income per capita.
pop	person	Population.
tax	% of value added	Taxes on goods and services.
LifeLadder	-	National average response to the question of life evaluations from a survey. 10 represents the best possible life and 0 is the worst.
LogGDPPerCapita	-	Log of GDP per capita.
SocialSupport	-	National average of 0/1 responses to the question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"
Expectancy	years	Healthy life expectancies at birth.
Freedom	-	National average of 0/1 responses to the question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
Generosity	-	Residual of regressing national average of 0/1 responses to the question "Have you donated money to a charity in the past month?" on GDP per capita.
Corruption	-	National average of the survey responses to two 0/1 questions: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?"
PositiveAffect	-	Average of three positive affect measures: happiness, laugh and enjoyment in the Gallup World Poll.
NegativeAffect	-	Average of three negative affect measures: worry, sadness and anger in the Gallup World Poll.