CIA World Factbook

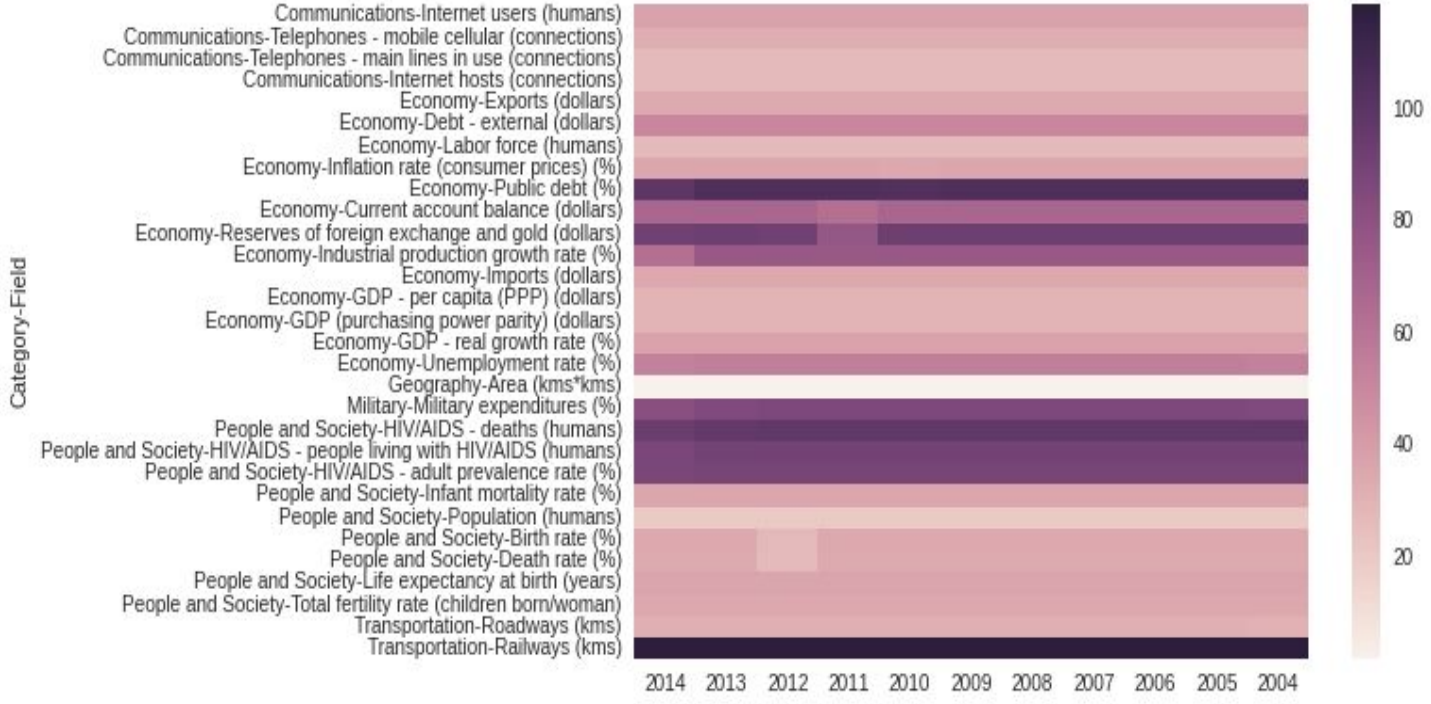
A Data Analysis Tour

Noel Carrascal

November 30, 2015

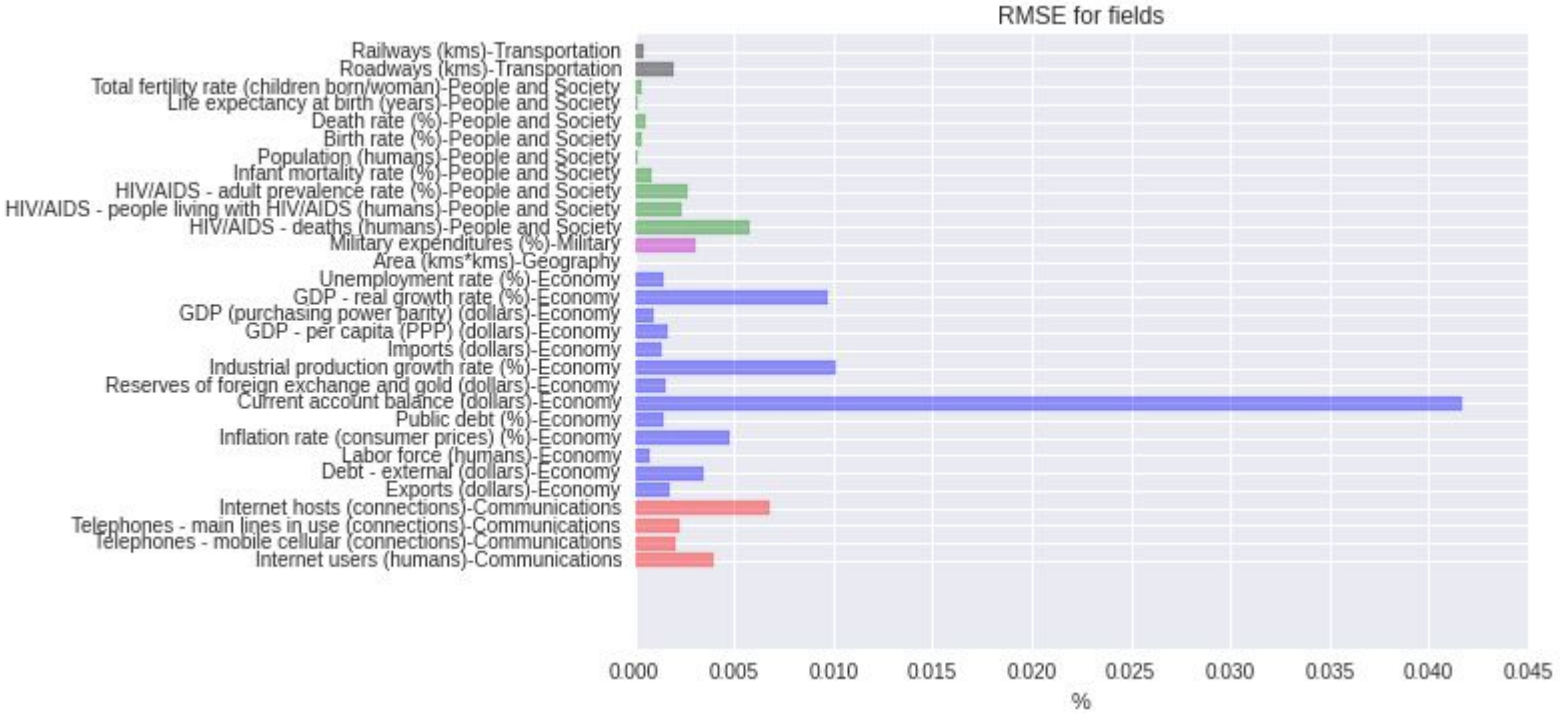
Introduction

- The CIA factbook contains information about countries around the world.
- For this project, I started with 30 fields of information on 261 countries.
- Fields are grouped in six categories: Economy, People and Society, Geography, Transportation, Military, Communications.
- The data contains a lot of null values in all fields. Not all countries collect statistics on all the fields.
- The CIA is also bad at collecting information. Some null values turned out to be ZERO! (i.e. Railways in Malta.)



Imputation: Missing values are added using predictive models.

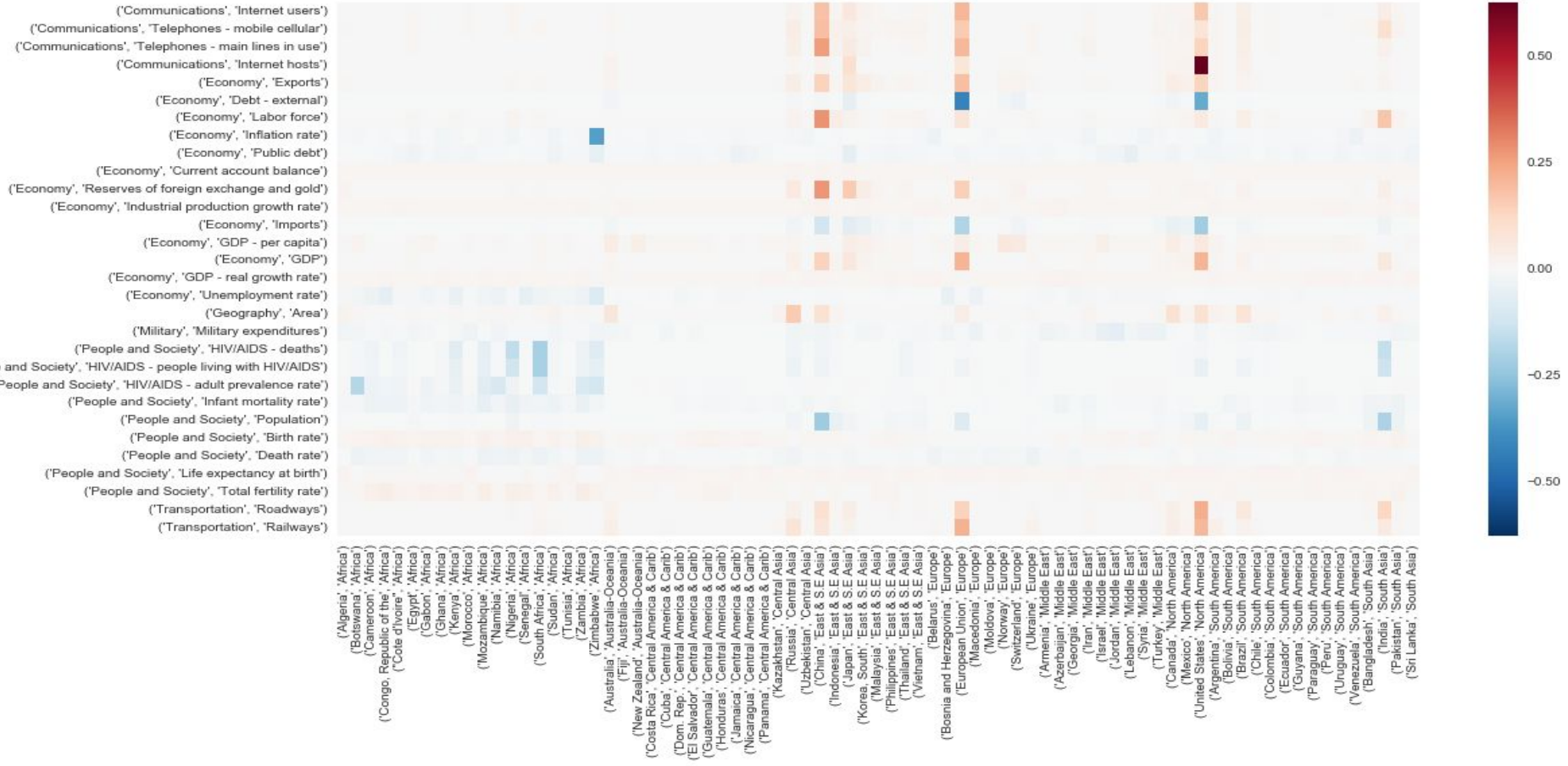
A linear model was used to fill in missing values. Using data up to 2013 to predict values on 2014, the following are the RMSE errors compared to known target values.



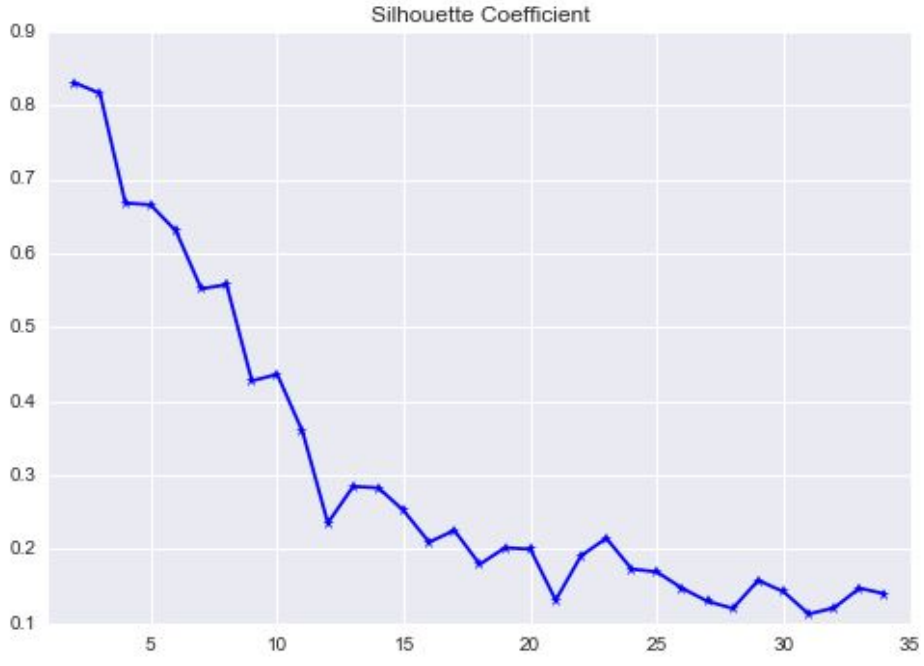
Data Regularization and Normalization

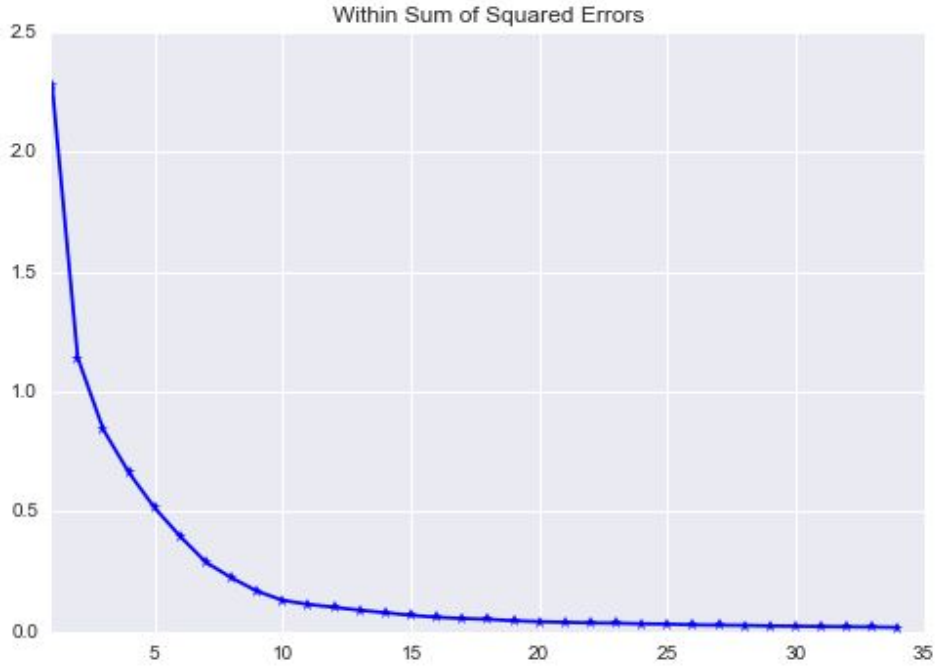
- Fields are given in different units, and the scales spawn several orders of magnitude.
- I used unit-based normalization Z = X−min(X) / max(X)−min(X) to make all values positive (from 0 to 1) for each column-field(X).
- Then I made all columns add up to 1: N = Z/sum(Z).
- How to interpret these data: In this way I rescale the columns to be a value from 0 to 1, similar to a percentiles. For example, United States GDP (PPP) is 0.206, or 20.6% of the world's GDP, and it has 60.21% of internet hosts in the world, and 0.823% of the total fertility rate of the world (Working and spending too much time on the computer?)
- Some field/columns were multiplied by -1 to reflect that is not something good for the wellbeing of people. GDP is positive, HIV deaths,negative; Exports, positive; imports, negative.

Normalized Data

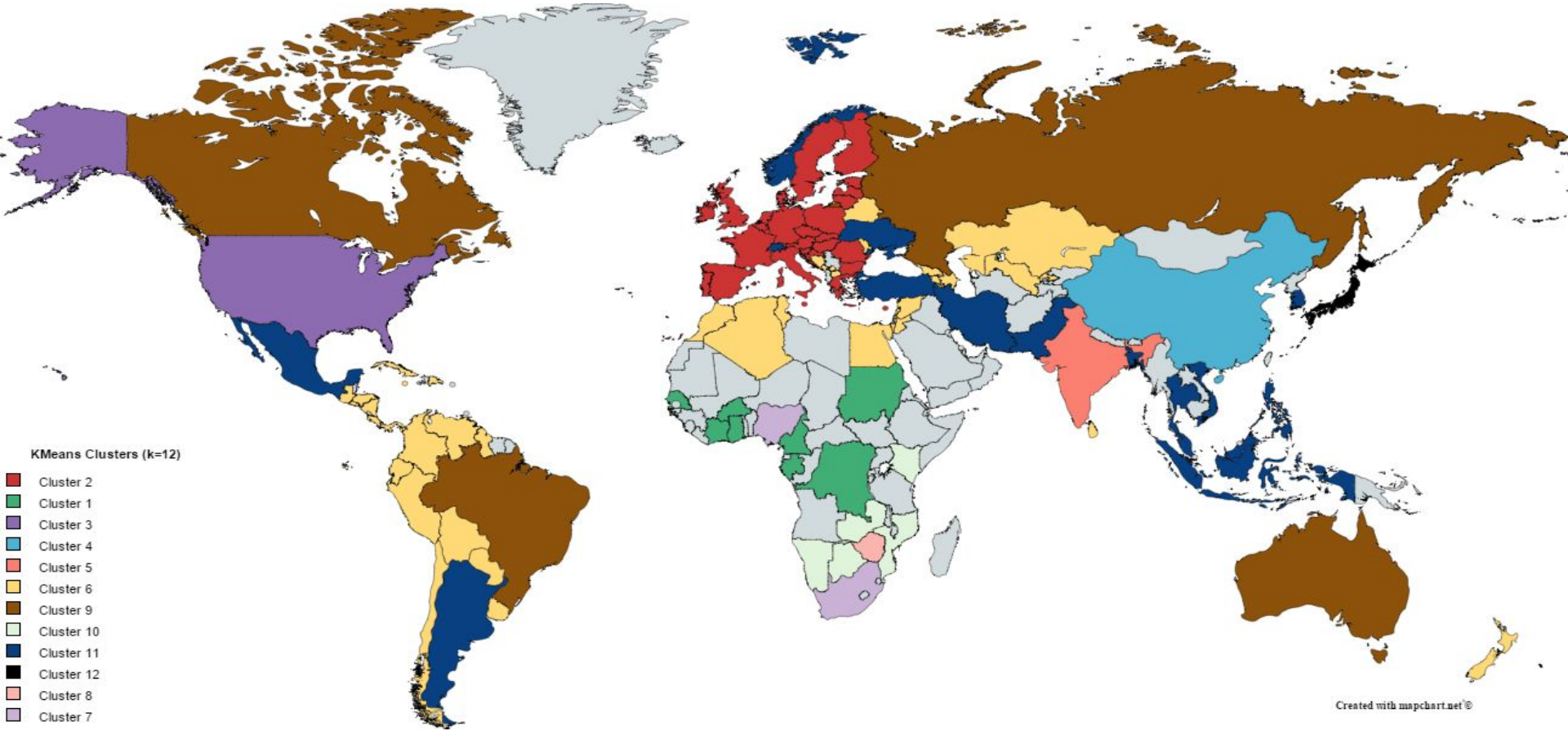


K-Means Clustering Results:





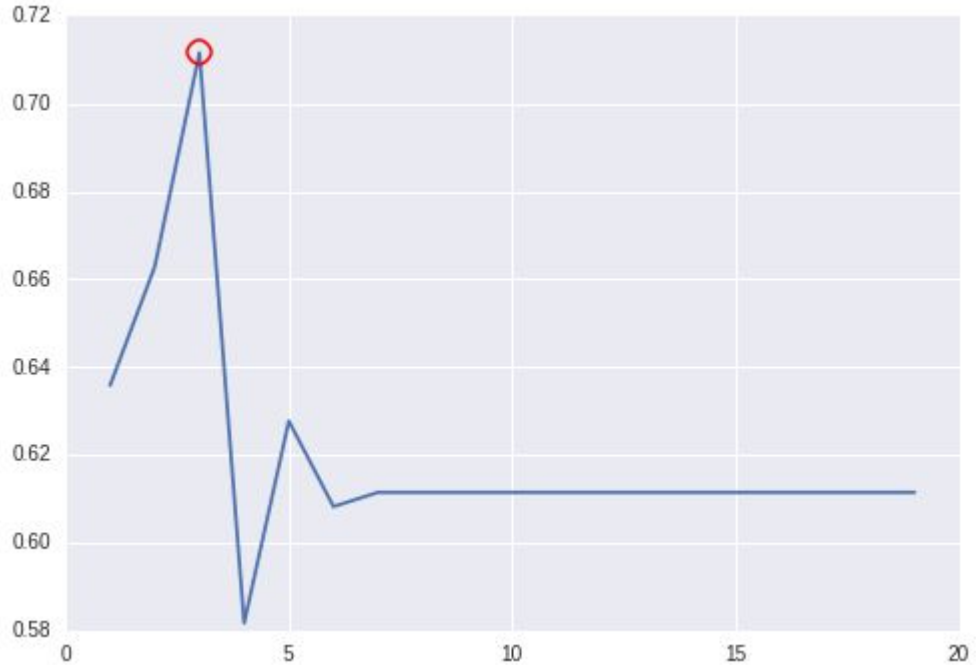
K-Means Clustering (K=12)



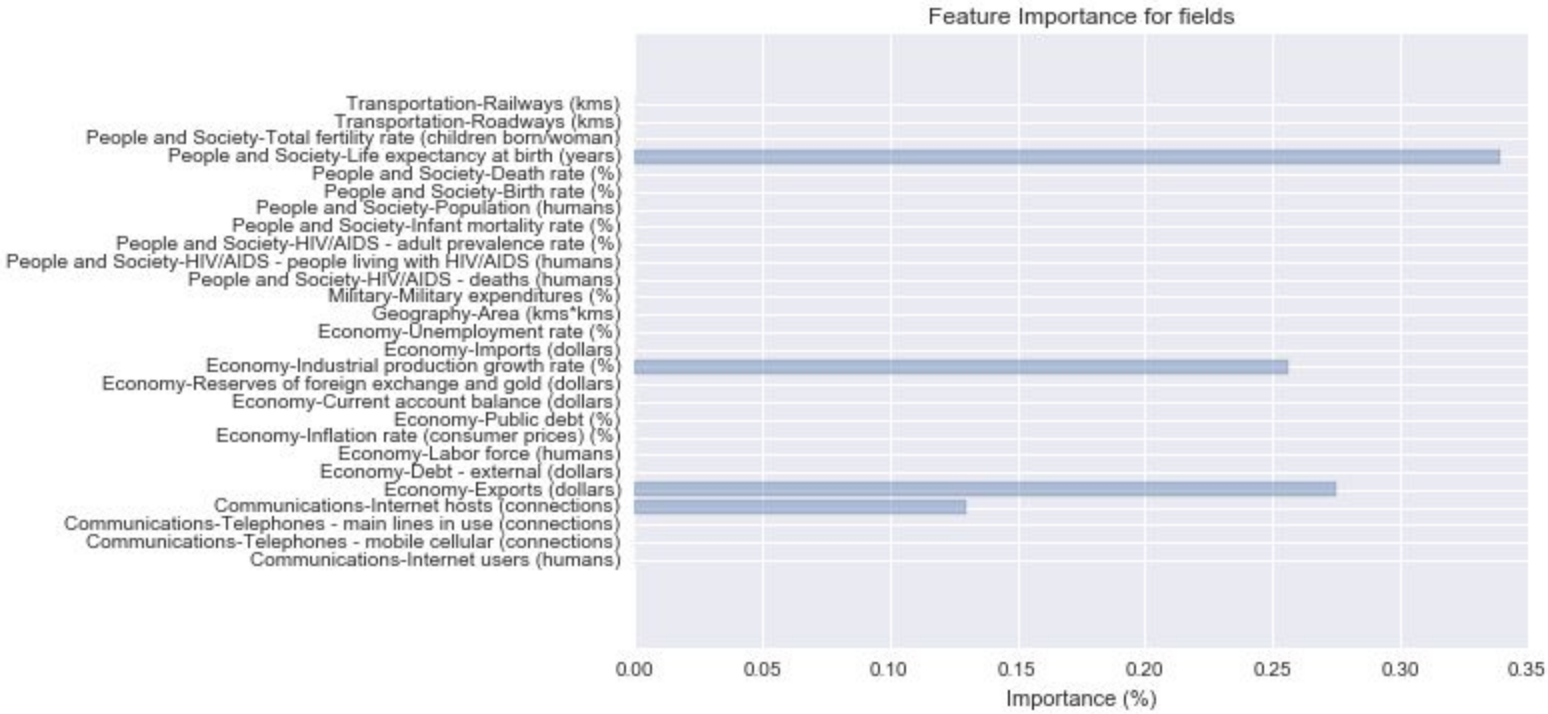
‘Predicting’ GDP

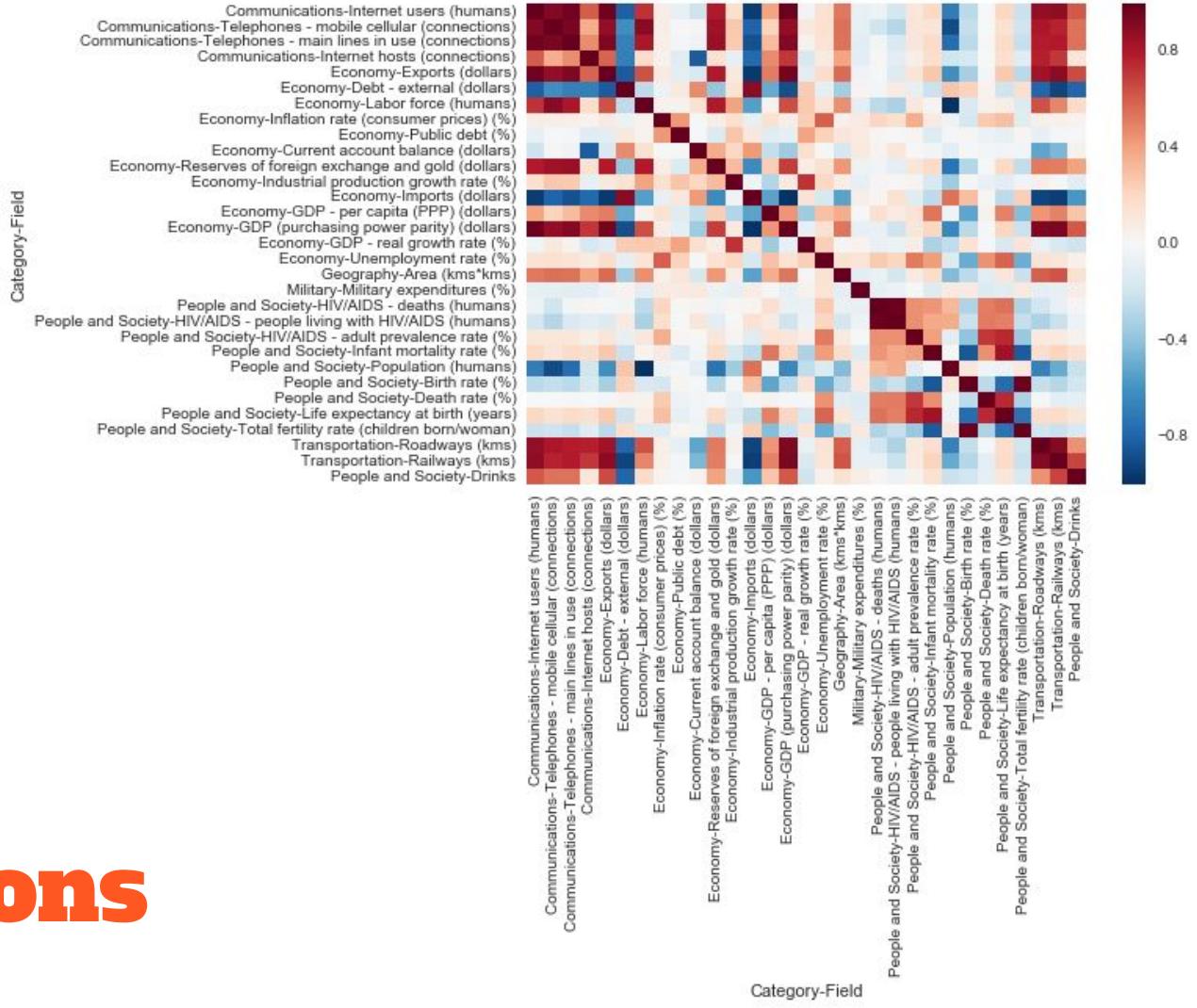
- The GPD Real Growth Rate was divided in two, the top and bottom 50% of the countries and assigned 1 (top) and 0 (bottom).
- Three fields directly related to GDP were removed.
- Decision trees, logistic regression and knn were used together with cross validation and grid search.

Logistic Regression	0.553
Knn (n=5)	0.521
Knn (n=12)	0.511
Decision tree (depth = 2)	0.663
Decision tree (depth = 3)	0.711
Decision tree (depth = 10)	0.611
Decision tree (depth = 3, entropy)	0.765



Feature Importance from Grid Search





Correlations

Thank You