The Seed in the Machine: Will a More Connected World Support a More Sustainable World?

Rochelle J. March | May 2017

Motivation: Clarifying the link between ICT and sustainable development

The Problem

It is well-understood that there is a need for increased global sustainable development.

Information communications technology (ICT), which includes technologies such as mobile phones, broadband and the Internet, can deliver communications and therefore solutions at an unprecedented speed and scale. ICT may be a key enabler for sustainable development, particularly in its ability to:

» Increase **access** to information
» **Connect** people and organizations to one another
» Improve **efficiencies** in resource, labor and market productivity

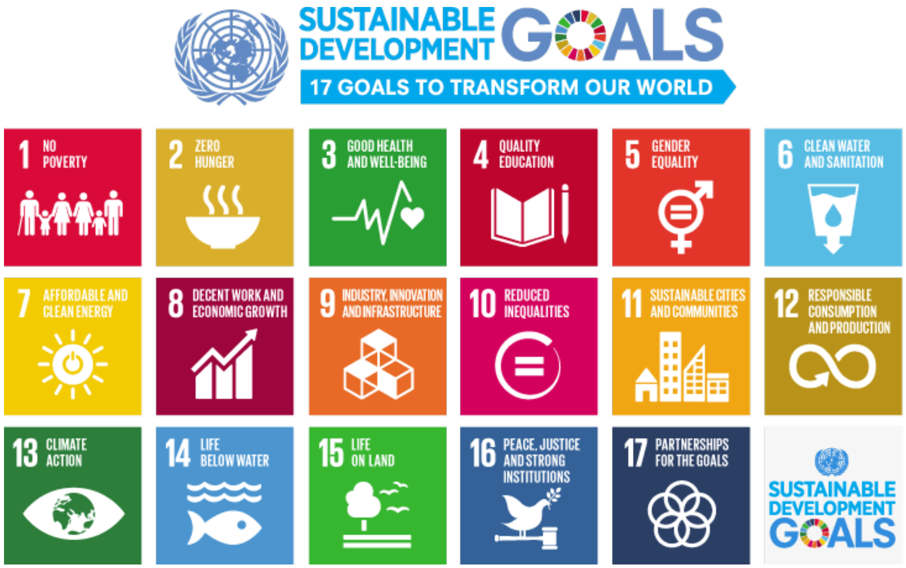
What sustainable development goals could most improve from leveraging information communications technologies?

Background: The Sustainable Development Goals

History of the Sustainable Development Goals (SDGs)

At the UN Sustainable Development Summit in September 2015, over 150 world leaders agreed upon a new sustainable development agenda. For the first time, these world leaders also included business leaders, who now must clarify what commitments their companies will make to support the Goals.

The Goals offer an ambitious and transformational vision for the future, with a target of 2030. Over the next fifteen years, these Goals will help mobilize efforts to end all forms of poverty, fight inequality, and tackle climate change, among other aims.


SUSTAINABLE DEVELOPMENT GOALS
17 GOALS TO TRANSFORM OUR WORLD
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

Goals of the Project: Explore and Practice

1. **Explore the relationship between ICT and sustainable development through data**
   - Build off existing research and efforts by the UN, ITU and technology companies (e.g., Huawei, Ericsson, Intel, ARM, etc.)
2. **Explore what are the areas to prioritize ICT investment to support further sustainable**
   - If there is a positive relationship between ICT development and sustainable development, determine what areas could see more immediate and greater results from increased ICT investment.
3. **Practice using supervised and unsupervised machine learning algorithms to determine correlations and clusters**
   - Determine relationship between ICT and sustainable development using linear regression and regularization (Lasso and Ridge)
   - Determine if there are country clusters in terms of ICT and certain types of sustainable development

Data: Gathering and cleanup

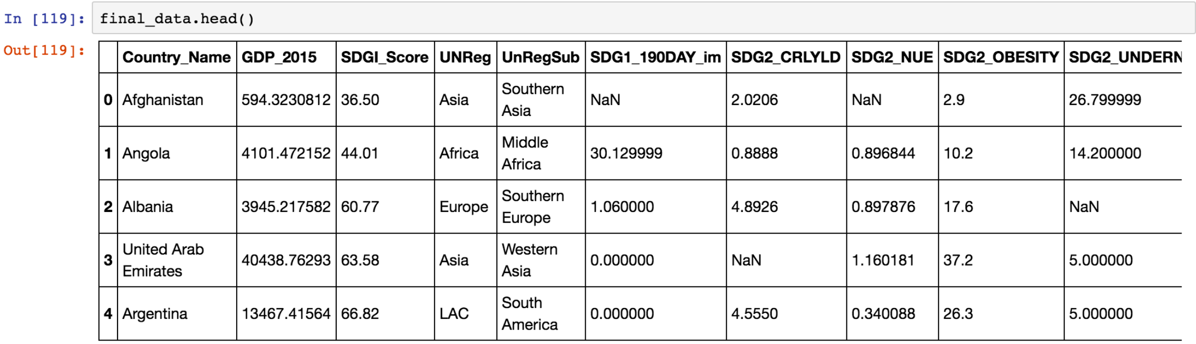
ICT Development Index 2016
#ITUdata

SDG INDEX & DASHBOARDS

1. Gather the data

- Sustainable development data: http://www.sdgindex.org/download/#data
- ICT data: http://www.itu.int/net4/ITU-D/idi/2016/
- GDP data: http://data.worldbank.org/indicator/NY.GDP.PCAP.CD

2. Merge using pandas merge and an outer join



```
In [119]: final_data.head()
Out[119]:
```

	Country_Name	GDP_2015	SDGI_Score	UNReg	UnRegSub	SDG1_190DAY_im	SDG2_CRLYLD	SDG2_NUE	SDG2_OBESITY	SDG2_UNDERN
0	Afghanistan	594.3230812	36.50	Asia	Southern Asia	NaN	2.0206	NaN	2.9	26.799999
1	Angola	4101.472152	44.01	Africa	Middle Africa	30.129999	0.8888	0.896844	10.2	14.200000
2	Albania	3945.217582	60.77	Europe	Southern Europe	1.060000	4.8926	0.897876	17.6	NaN
3	United Arab Emirates	40438.76293	63.58	Asia	Western Asia	0.000000	NaN	1.160181	37.2	5.000000
4	Argentina	13467.41564	66.82	LAC	South America	0.000000	4.5550	0.340088	26.3	5.000000

3. Inspect for NaNs and fill or drop them

- Fill with mean of column: filled_final_data = final_data.fillna(final_data.mean())
- Drop remaining nulls: final_final_data = filled_final_data.dropna(subset=['GDP_2015'])

4. Create dummies for regional data that is categorical

- UNReg_dummies = pd.get_dummies(final_final_data.UNReg, prefix='UNReg').iloc[:, 1:]final_final_data = pd.concat([final_final_data, UNReg_dummies], axis=1)

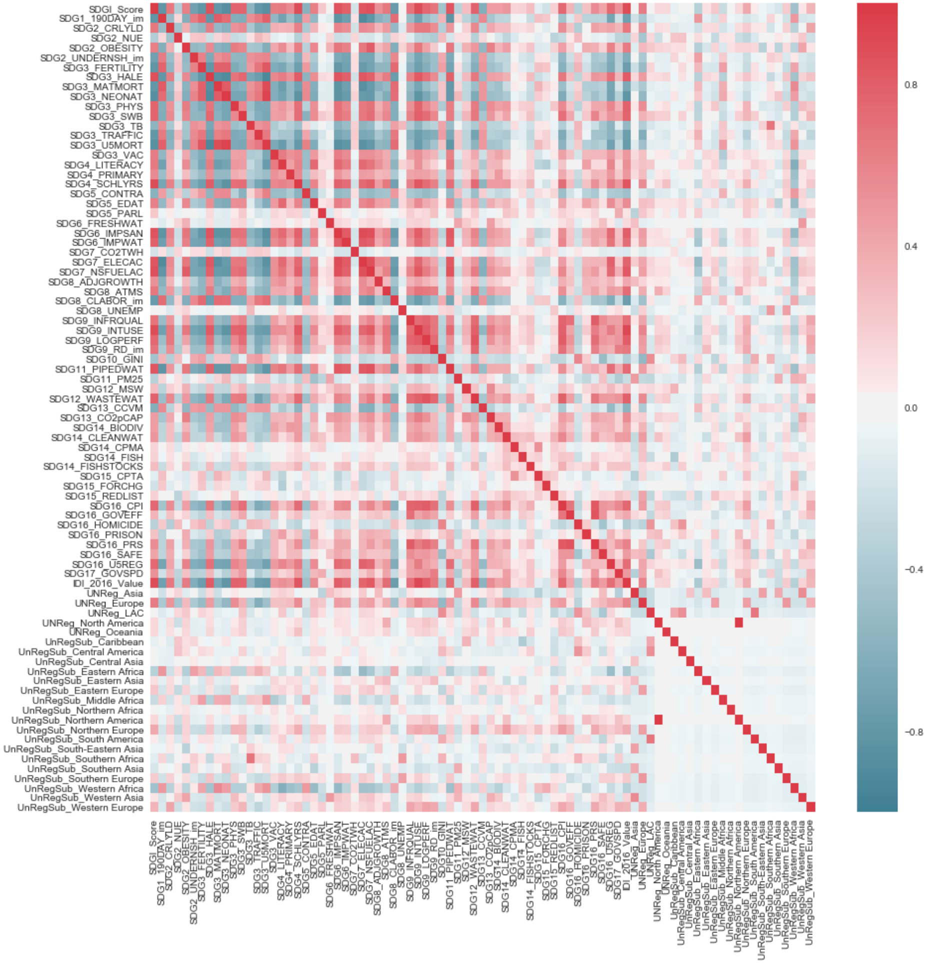
Linear Regression: Steps

1. **Data gathering and cleanup**
   - Merge three different datasets to get data for ICT, sustainable development and GDP
   - Deal with NaNs (or null values)
   - Create dummy variables for categorical variables (e.g., UNReg, UNSubReg)
   - After test and train, normalize/scale data using StandardScaler
2. **Data inspection and visualization**
   - Correlation matrix
3. **Linear regression**
   - Run linear regression using Scikit-Learn for dependent variable "IDI_Value_2016" which is a measure of ICT development of a country
   - Test and train model
   - Isolate relevant features using Lasso and Ridge regression and find best coefficients via GridSearch
4. **Robustness & model selection**
   - Compare R^2 scores of different regression algorithms to determine the preferred model
5. **Analysis of results**

Data: Regression visualization

1. Correlation Matrix

- Areas with dark blue and dark red have high correlation >0.8 or <-0.8
- Too many variables, going to try to find the relevant ones through Lasso and Ridge



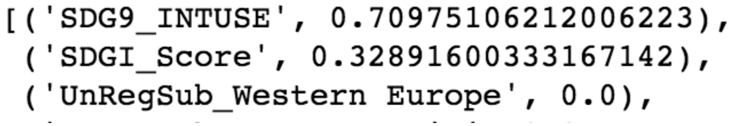
Results: Linear Regression

1. **Linear regression**
   - High R^2 = LR R2: 0.966786435801
   - But too many variables
2. **Lasso**
   - Lasso regression is a regularization that drops non-important features
   - R^2 = Lasso R2: 0.58355930028
   - Relevant coefficients are only two features now
3. **Ridge**
   - Ridge will not drop non-relevant features but will shrink them near to zero
   - R^2 = Ridge R2: 0.667475255468
   - Relevant coefficients are now more ~7

Analysis

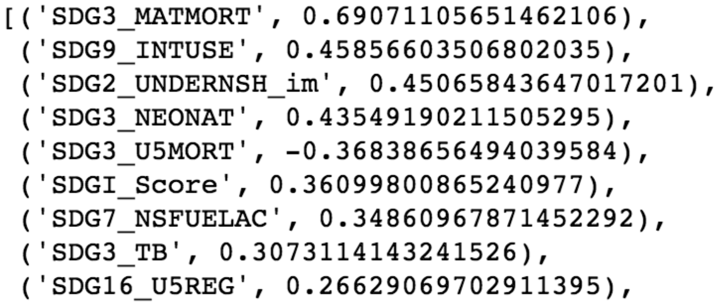
- Most of the high coefficients are health related (SDG 3: Good Health and Well-being), but are positively correlated except for mortality under 5 years
- Second most highly correlated is Internet use, which may be collinear

Lasso



```
[('SDG9_INTUSE', 0.7097510621200622 3),
 ('SDGI_Score', 0.3289160033316714 2),
 ('UnRegSub_Western Europe', 0.0),
```

Ridge



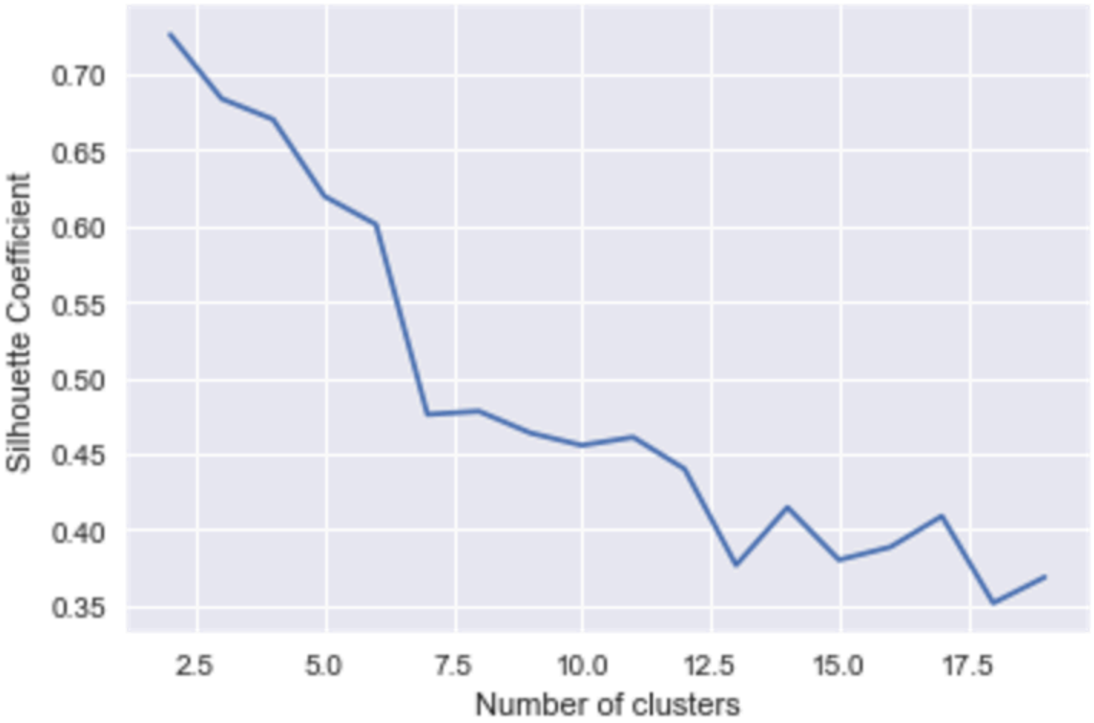
```
[('SDG3_MATMORT', 0.6907110565146210 6),
 ('SDG9_INTUSE', 0.4585660350680203 5),
 ('SDG2_UNDERNSH_im', 0.4506584364701720 1),
 ('SDG3_NEONAT', 0.4354919021150529 5),
 ('SDG3_U5MORT', -0.3683865649403958 4),
 ('SDGI_Score', 0.3609980086524097 7),
 ('SDG7_NSFUELAC', 0.3486096787145229 2),
 ('SDG3_TB', 0.307311414324152 6),
 ('SDG16_U5REG', 0.2662906970291139 5),
```

Cluster Analysis: Steps

1. **Same data gathering and clean-up**
2. **Data inspection and visualization**
   - Scatter plots
3. **Clustering**
   - Run clustering algorithm
   - Methods tests: K-means, DBSCAN
   - Use Silhouette Coefficient to determine number of clusters
4. **Robustness & model selection**
   - Compare silhouette scores of different clustering algorithms to determine the preferred model
5. **Analysis of results**

Data: Clusters visualization

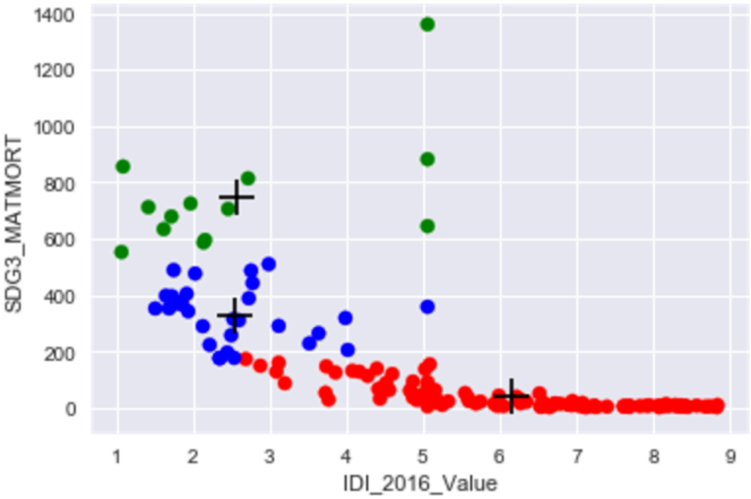
- Clusters determined by silhouette score
  – 3 clusters = 0.68399092238790271

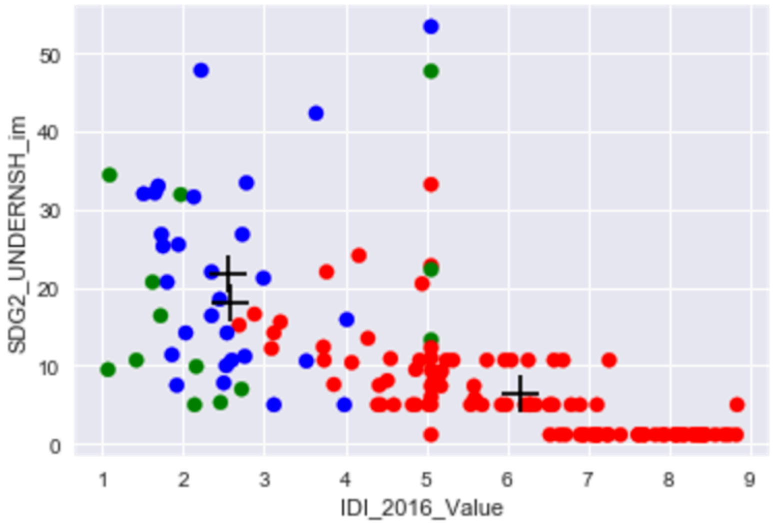


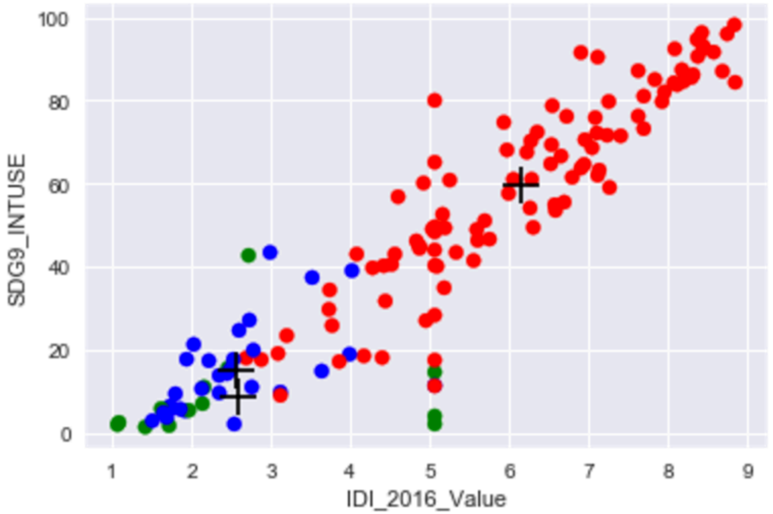
Data: Clusters visualization

1. Scatter plot

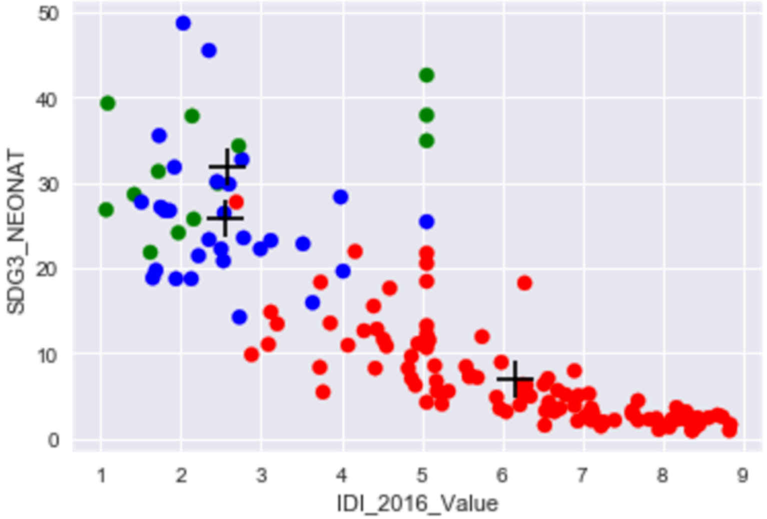
- Variables chosen from Lasso
- Correlation matrix (next time)
- Clusters determined by silhouette score

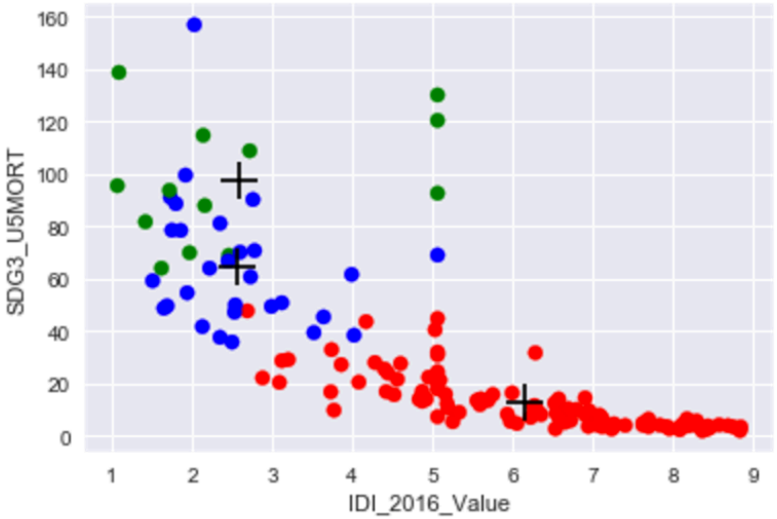


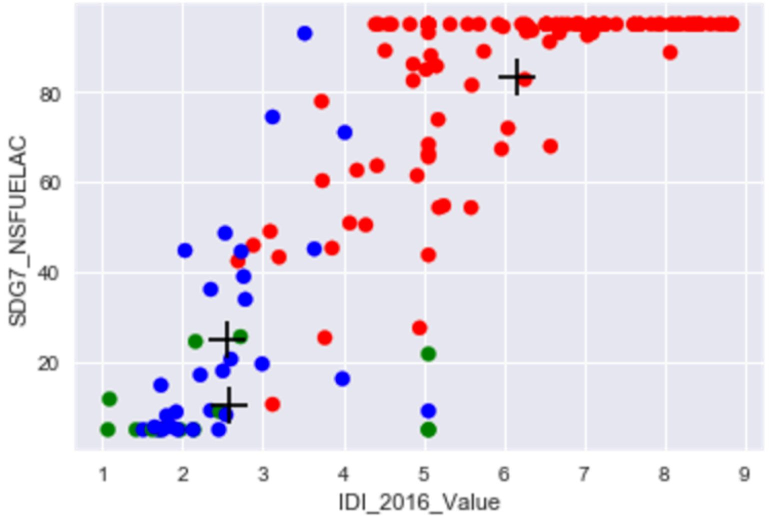




Data: Clusters visualization







Results: Cluster Analysis

1. **Clustering K-means**
   - Run clustering algorithm
   - Methods tests: K-means, DBSCAN
   - Use Silhouette Coefficient to determine number of clusters
2. **Clustering DBSCAN**
   - Compare silhouette scores of different clustering algorithms to determine the preferred model
3. **Silhouette scores**
   - Compare silhouette scores of different clustering algorithms to determine the preferred model
   - Cluster based on best silhouette coefficients
4. **Analysis of Results**
   - Merge back with data to test correlation (gdp, regions?)
   - What is the clustering telling me about the data?

Conclusion: Takeaways

What had the most impact on your work?

- Isolating certain variables using Lasso and Ridge
- Clustering to find similarities among countries

What can you confirm? What can you suggest? What is still to be determined?

- Health-related issues (maternal mortality, neonatal mortality, etc.) are more correlated with ICT development

Next Steps: Further Explorations

What should this project do moving forward?

- Test more models
- Try different, potentially more relevant, features

What would be the next two or three things you want to try? What impact might they have?

- Try correlation of clusters with other indicative data e.g., regions, income

What might your conclusions enable others to do?

- Focus on health-related issues as an area to tailor ICT solutions
- Refine metrics for sustainable development

Github

https://github.com/rochellemarch/project-2