Worldwide Contraceptive Use

By Lara Haase

The Problem:

What variables affect the rate of contraceptive use worldwide, and which factors of well-being are most impacted by contraceptive use?

Potential Clients:

The World Bank, Human Rights Watch, Amnesty International, FIDH: Worldwide Human Rights Movement, International Women's Rights Action Watch, or other international human or women's rights organizations.

The Data:

The World Bank Gender Statistics:
https://data.worldbank.org/data-catalog/gender-statistics

Other Potential Datasets:

-United Nations Gender Statistics: https://genderstats.un.org/
-World Bank Education Statistics
-United Nations Gender Info:
https://www.quandl.com/api/v3/datasets/UGEN/GENC_5001.csv?api_key=522GtZ-iXb4LDM4xEEax
-Contraceptive Use- World- (also available by country):
https://www.quandl.com/api/v3/datasets/UGEN/CNTR_5001.csv?api_key=522GtZ-iXb4LDM4xEEax
-Abortion Rate (also available by country)
https://www.quandl.com/data/UGEN/ABRT_5001-Abortion-Rate-World
-Johnston's Archive (historical abortion stats for countries) **-** 130 data sets

Data Wrangling:

The World Bank Gender Statistics data set was first imported into a Pandas dataframe. Columns in the raw data set are the years the datapoints were recorded, so the dataframe was melted in order to stack the years into rows. Then NaN values were dropped from the data frame, to eliminate unnecessary rows created for years where no data is listed.

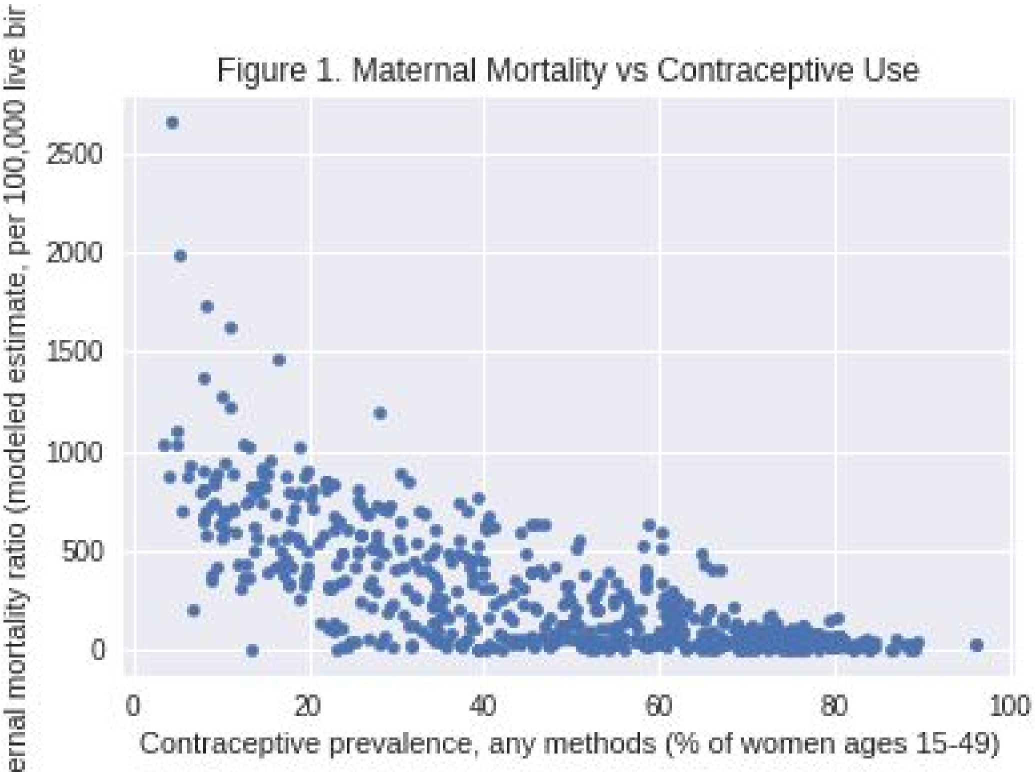
The rows of the raw data contain not only which country the data point is from, but also which feature (or variable) the point represents. Since the goal is to analyze the variables by how they correlate with country and year, the variables needed to be listed as columns. So the features were unstacked to list each measured variable as its own column, with the rows representing the country and year the value was collected or measured.

The data set has a large number of missing data points. Years since 1999 appear to have the highest not-null counts, so the data frame was restricted to years greater than 1999 and less than 2015.

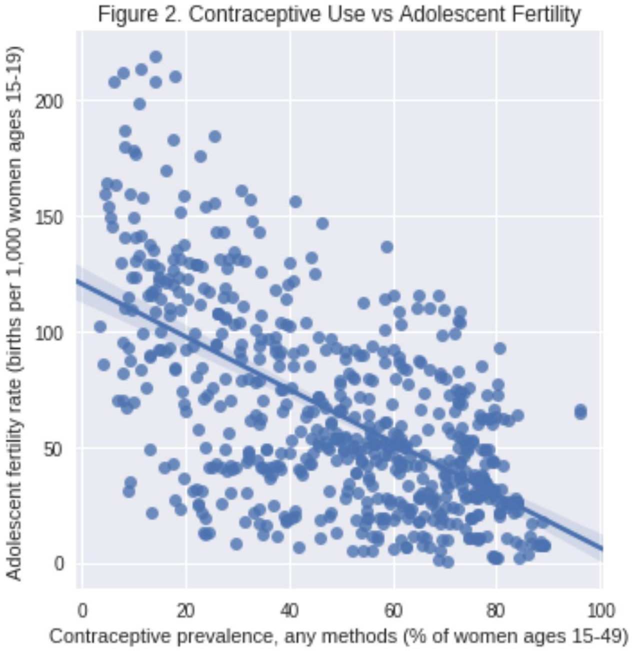
Initially any remaining missing values were left as NaN because replacing them with 0 would significantly affect the analysis. There are nearly 700 features contained in the dataset. This is an overwhelming amount of information to try to discover something interesting, but there was also concern for overlooking potentially interesting insights. Through the analysis, smaller data frames were made to focus on only a few features, rather than all the available data in the set. Until the final variables were selected to be studied, NaNs were dealt with differently depending on each analytical situation.

Exploratory Findings:
(The Spearman's R test was used to calculate all correlation coefficients.)

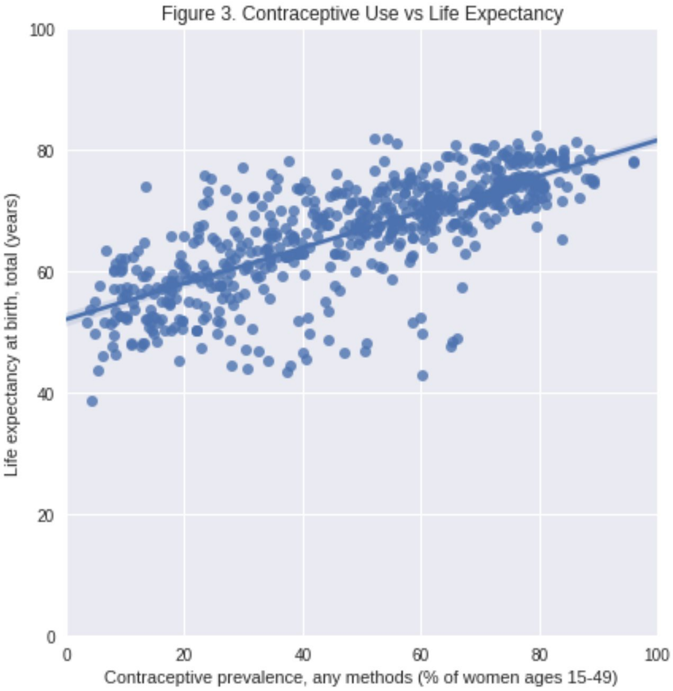
Worldwide, there is a correlation of -0.72 between Contraceptive Prevalence (any method) and Maternal Mortality Ratios, as visualized in **Figure 1.** This may be caused by affluence or culture, but the two topics seems as though they could very well be causational.



A negative correlation of -0.56 was also found between Contraceptive Prevalence and Adolescent Fertility Rate, which can be seen in **Figure 2**. More controlled research and experimentation would be necessary to determine concretely whether these relationships are causational.



Another strong correlation was between Contraceptive Prevalence and Life Expectancy with a coefficient of 0.77, which is plotted in **Figure 3.** The correlation may also be due simply to affluence, but it is noteworthy either way.



Maternal Mortality is extremely prevalent in Sub-Saharan Africa. **Figure 4** shows the 20 countries with the highest Maternal Mortality ratios, all of which are African countries.

Figure 4: Countries with the Highest rate of Maternal Deaths

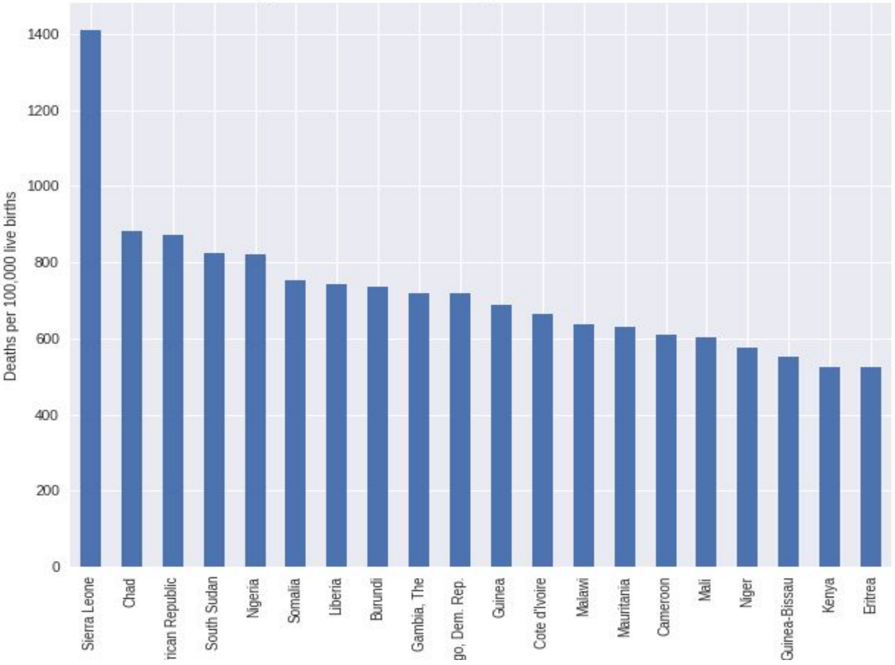
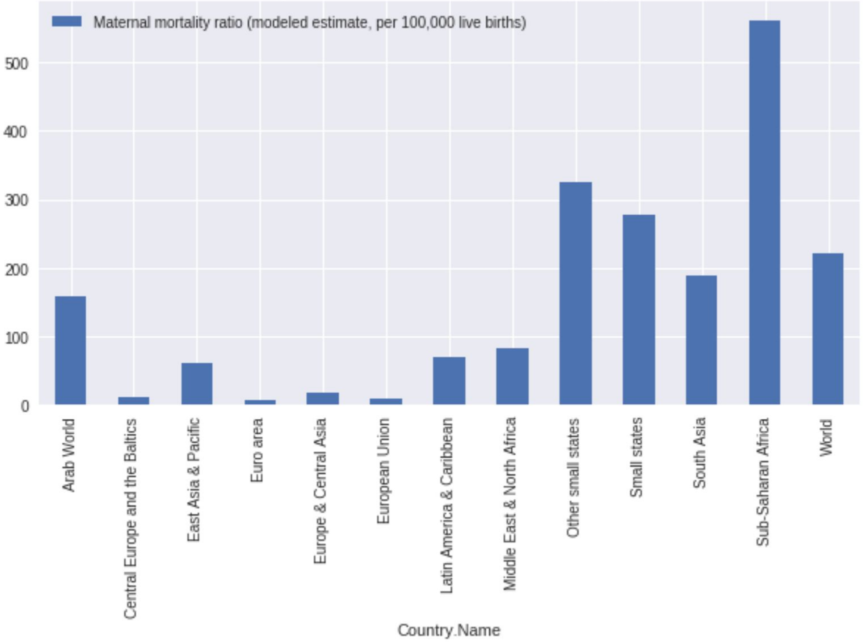
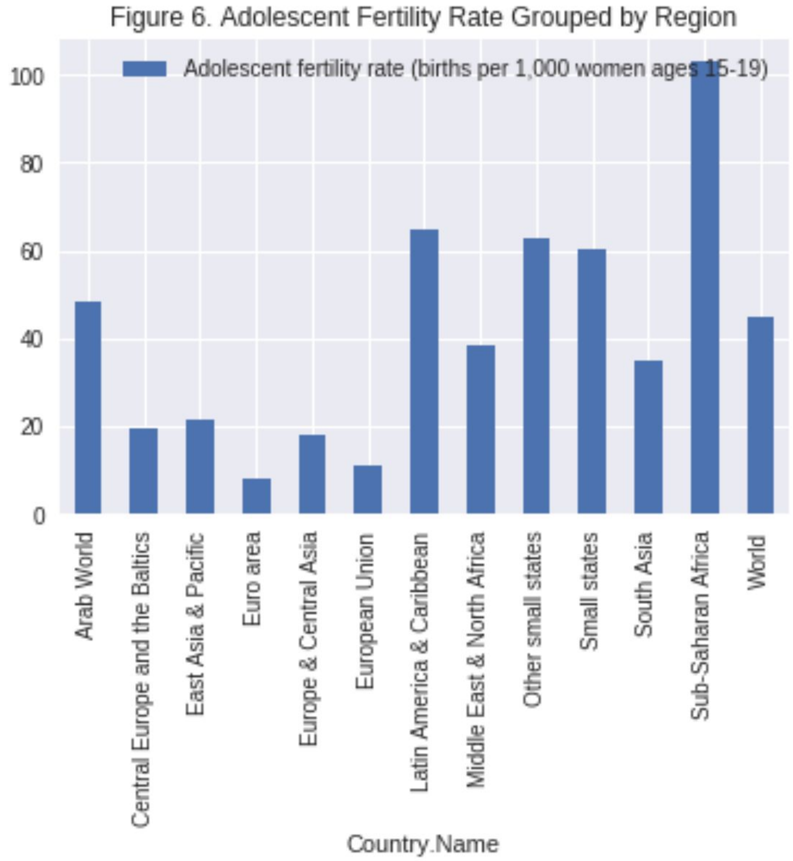


Figure 5 shows Maternal Mortality Ratios by region, and confirms that Sub-Saharan Africa has the highest Maternal Mortality ratio, by far.

Figure 5. Maternal Mortality Grouped by Region



Sub-Saharan Africa also has the highest rate of Adolescent Fertility, as seen in **Figure 6:**



It is also clear that the distribution of Female Legislators, Senior Officials and Managers is not equal to males, as the mean percent of female legislators is well below 50%, somewhere around 30%, evident in **Figure 7.**

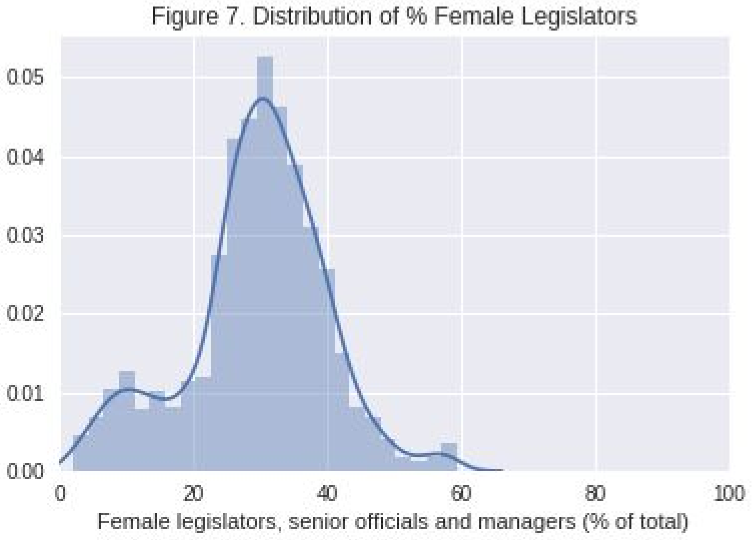
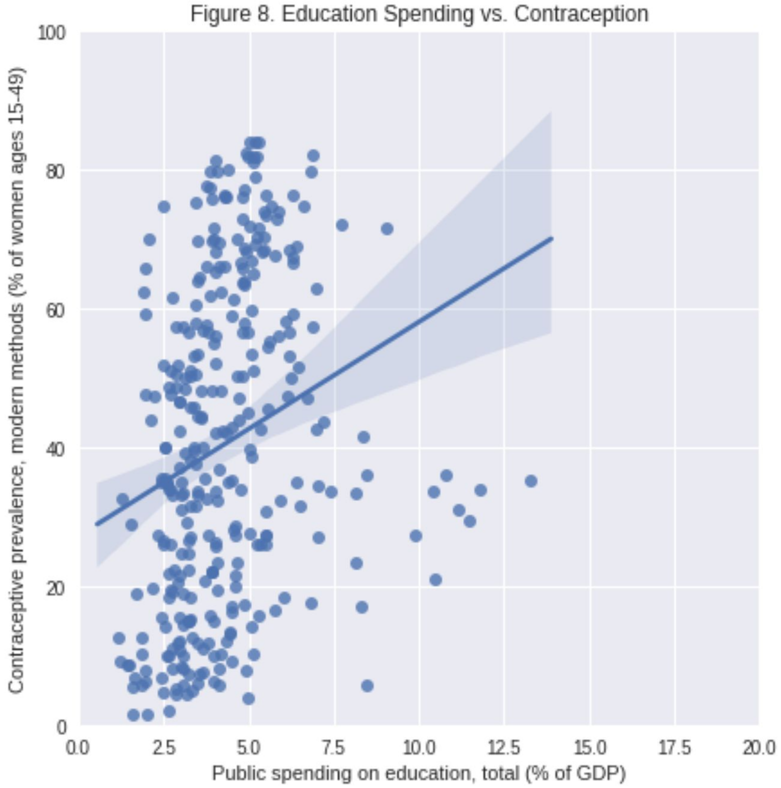
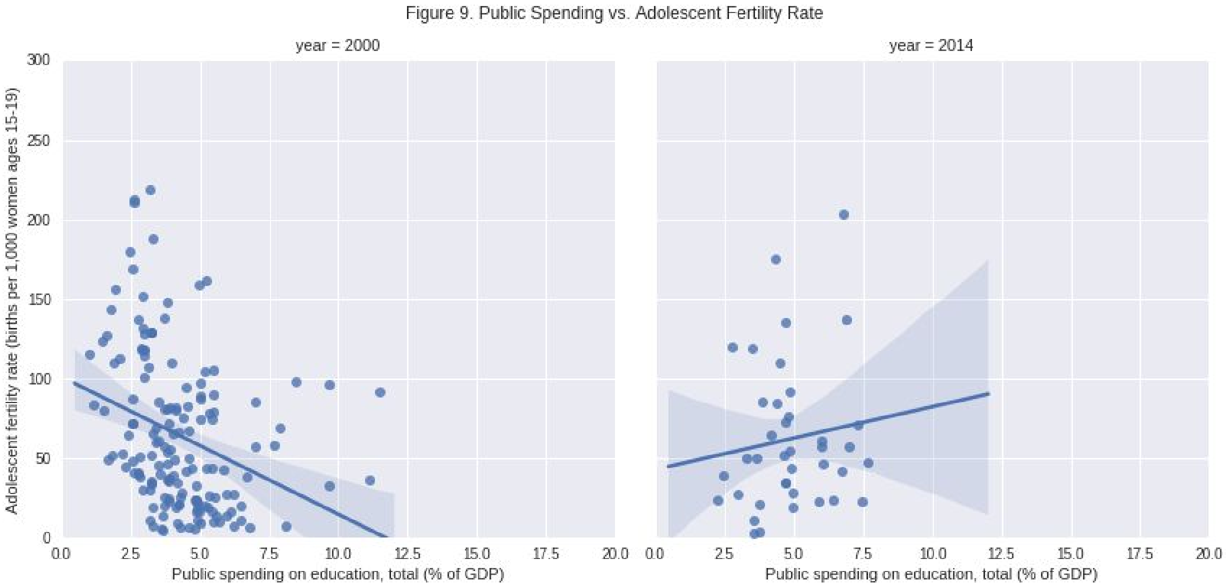


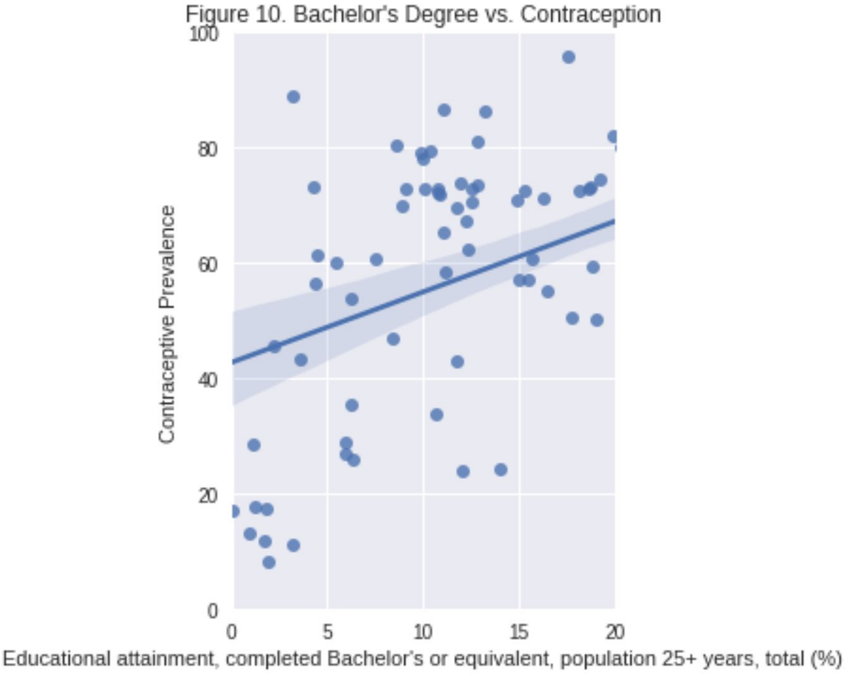
Figure 8 shows that Public Education Spending plotted against Contraceptive Prevalence at first does not appear to have a strong correlation(0.29):



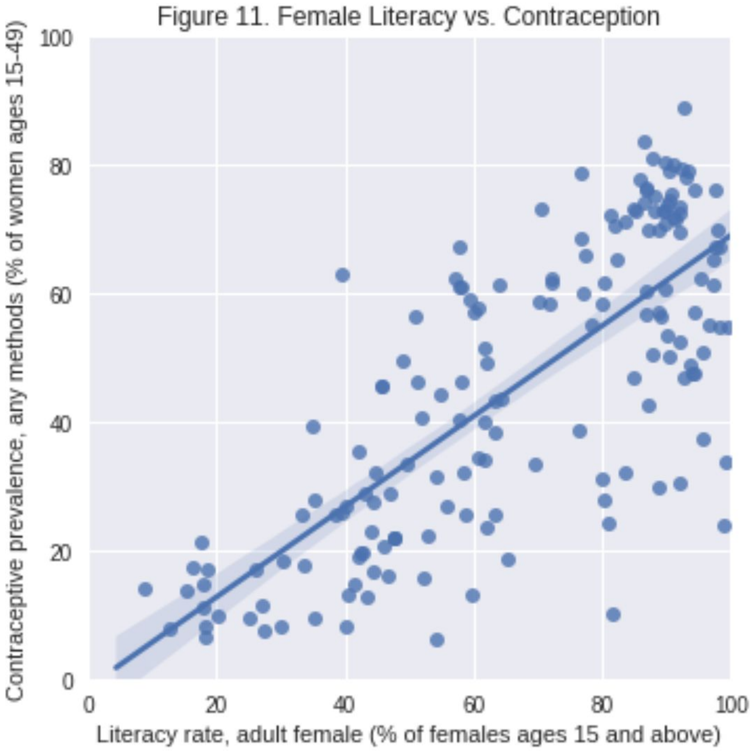
But when isolating 2000 vs 2014, as in **Figure 9** the correlation between Public Education Spending and Contraceptive Prevalence changes from negative in 2000 to positive in 2014. So there is either no correlation between the two features, or there are outside mitigating factors that changed over the decades.



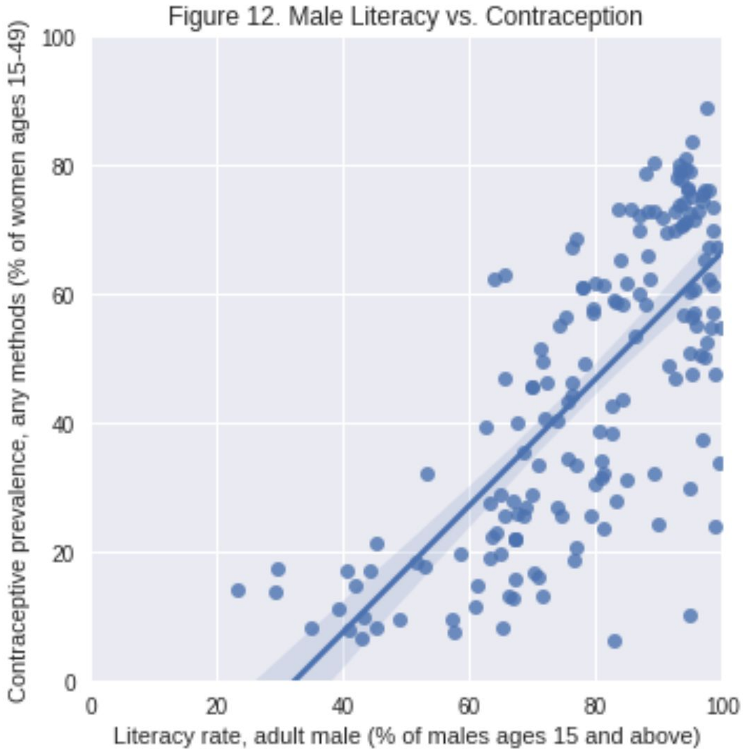
There is a slight positive correlation of 0.4388 between Total Bachelor's Degree Achievement and Contraceptive Prevalence as seen in **Figure 10**, but Female Bachelor's Degree Achievement vs Contraceptive Prevalence has a slightly stronger coefficient of 0.467.



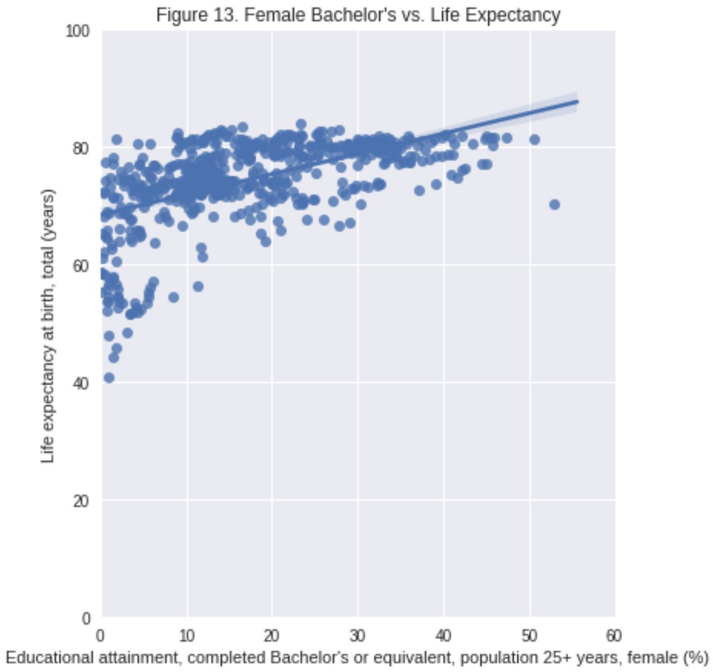
A much stronger correlation coefficient of 0.7128 was found between Female Literacy and Contraceptive Prevalence, seen in **Figure 11:**



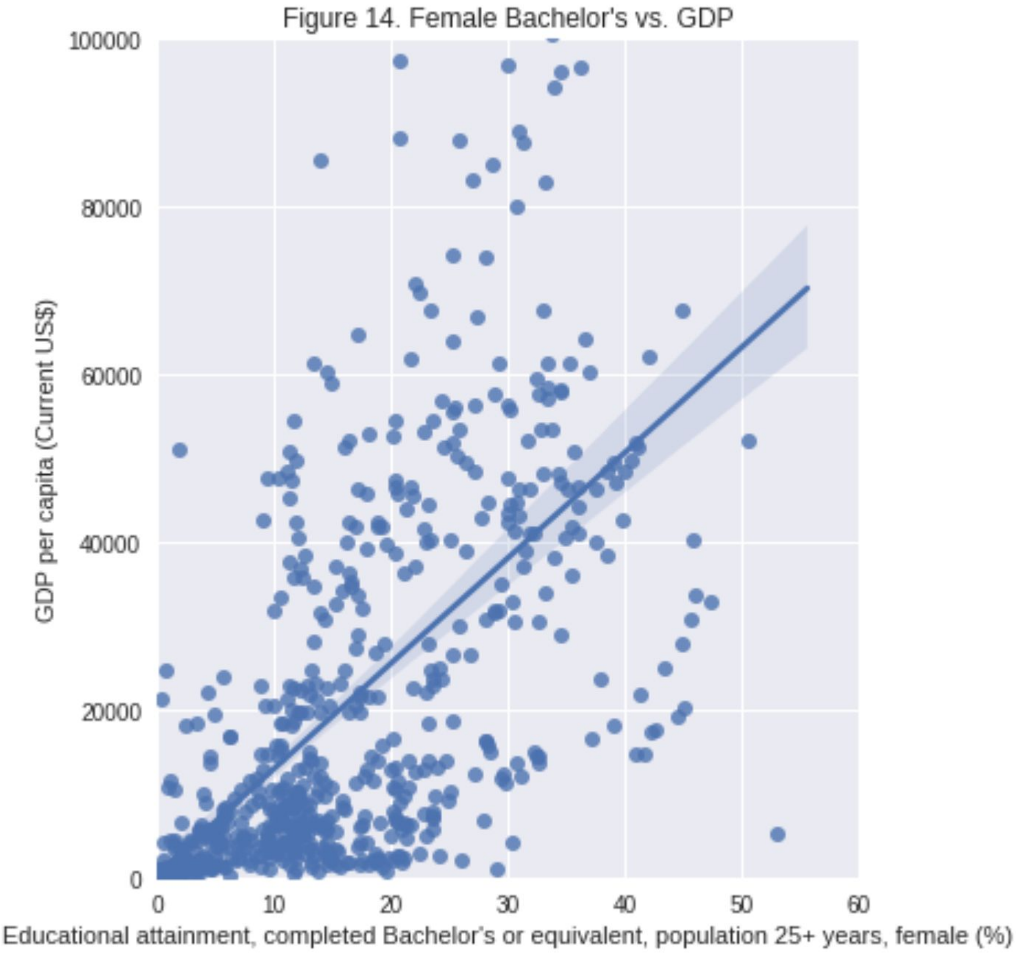
As well as 0.7122 for Male Literacy, seen in **Figure 12:**



There's a slight correlation of 0.467 between Female Bachelor's Degree Attainment and Life Expectancy, visualized in **Figure 13:**



But **Figure 14** shows a much stronger correlation of 0.66777 between Female Bachelor's Degree Attainment and GDP:



There are very clear correlations between Contraceptive Prevalence and the Decision Maker of a (married) Woman's Health Care. There is a positive correlation of 0.6178 when a wife makes her own decisions (see **Figure 15**), and a negative correlation of -0.68 when the husband makes these decisions (see **Figure 16**).

Figure 15. Wife as Health Care Decision Maker

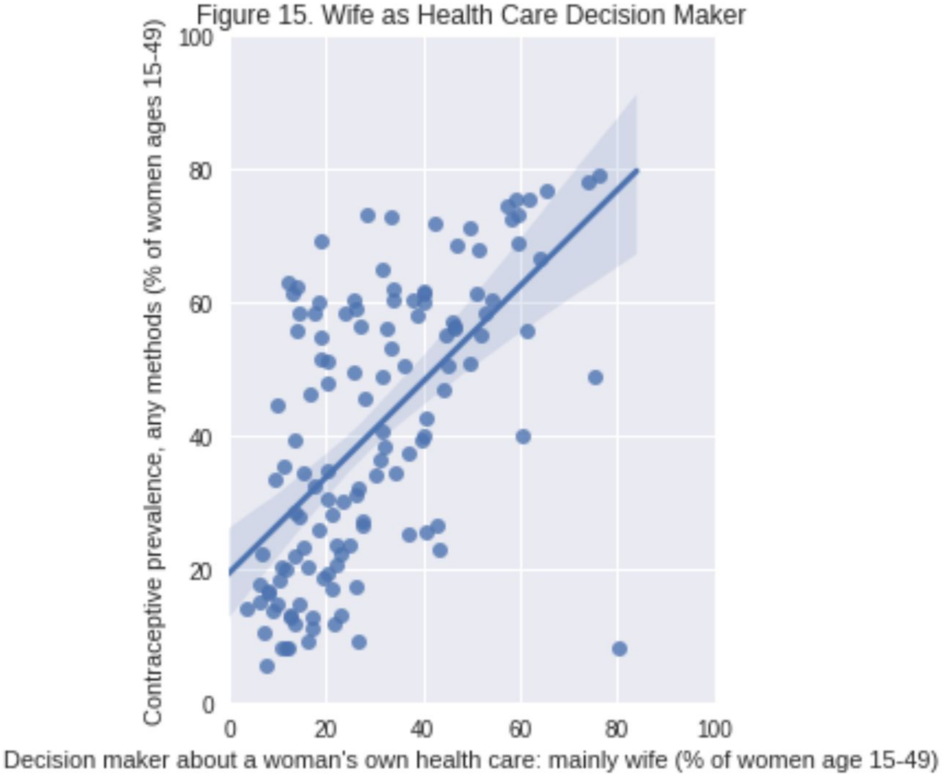
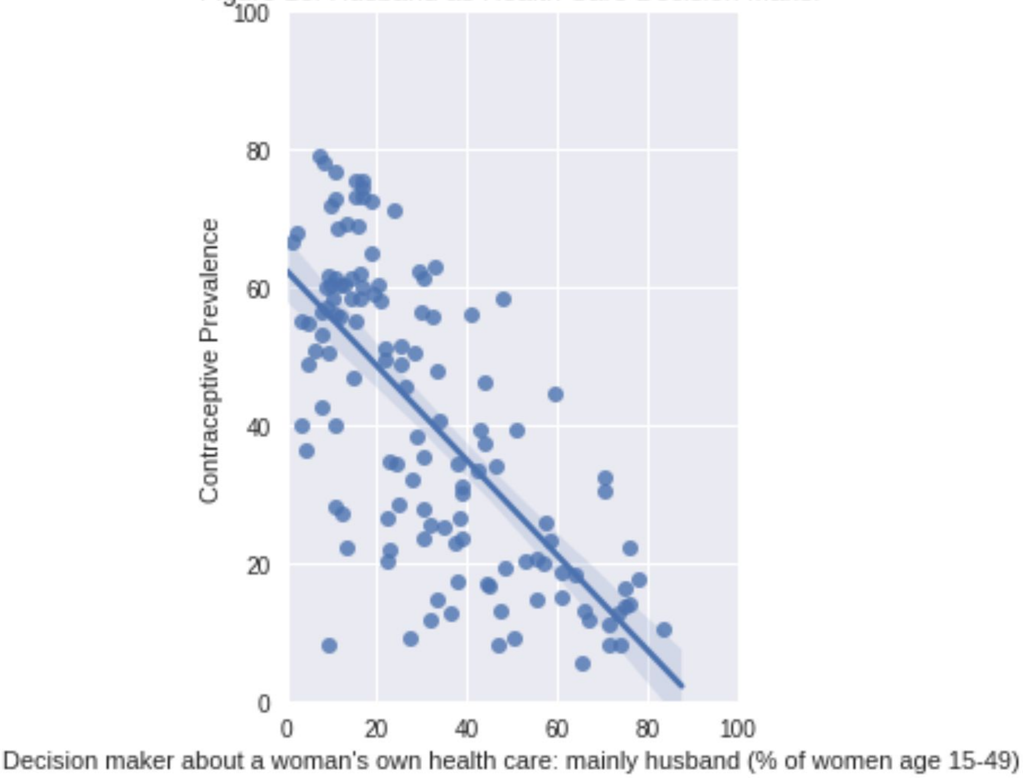


Figure 16. Husband as Health Care Decision Maker



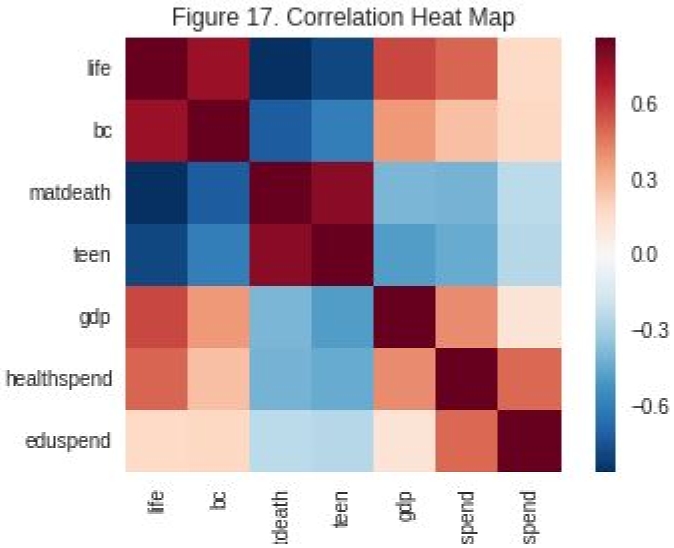
On that note, Maternal Mortality has a correlation coefficient of 0.65 with the husband as the Decision Maker, but when the wife is the Decision Maker the correlation coefficient with Maternal Mortality is -0.47.

Inferential Statistical Analysis:

The following variables were selected for model building (definitions and explanations of each variable are in the Appendix):

- Contraceptive Prevalence, any methods (% of women ages 15-49): relabeled as 'bc'
- Life Expectancy at birth, total (years): relabeled as 'life'
- Maternal Mortality ratio (modeled estimate, per 100,000 live births): relabeled as 'matdeath'
- Adolescent Fertility rate (births per 1,000 women ages 15-19): relabeled as 'teen'
- GDP per capita (Current US$): relabeled as 'gdp'
- Health expenditure, public (% of GDP): relabeled as 'healthspend'
- Public (Government) Spending on Education, total (% of GDP): relabeled as 'eduspend'

A Heat Map (**Figure 17**) was created to give a visual idea of how the selected variables correlate with each other.



The SGDRegressor model from Scikit-Learn was chosen. The dataframe was then scaled, and a Dependent Variable array and an Independent Variable matrix were created to test the model's accuracy. A smaller dataframe was created with only the Contraceptive Prevalence and Life Expectancy features. An SGDRegressor was fit to that data and scored 0.55.

Machine Learning- Model Building:

The seven chosen variables were grouped by whether they seem to be determining Contraceptive Prevalence or are affected by it. In the "Effect" group, assumed to be Dependent variables in relationship to Contraceptive Prevalence, there were very few missing data points, so the NaN values were simply dropped. In the "Cause" group, assumed to be Independent variables determining Contraceptive Prevalence, there were too many missing data points to simply drop the NaN values. The "Fancy Impute" package was used to interpolate the data in the "Cause" group using the "MICE" algorithm. A few more values were then tweaked after interpolation to fit with expected values.

The multiple regressor algorithms available in Scikit Learn were imported and a "run_models" function was created to quickly fit each regressor to the data, score the model, and compare results visually. Four of the regressors returned usable models, so a function was created for each of these four algorithms (SVM Regressor, KNN Regressor, Gradient Boosting Regressor and Random Forest Regressor) to run each algorithm with Kfolds and GridSearch Cross Validation, then fit the model with the best scoring parameters to the data.

The first data subset analyzed looked at Contraceptive Prevalence as the Dependent Variable, and the Independent Variables assessed were GDP, Healthcare Spending, and Education Spending. The best model found was a Random Forest model with the following parameters (non-default parameters are highlighted):

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=4,
      max_features='auto', max_leaf_nodes=None,
      min_impurity_split=1e-07, min_samples_leaf=1,
      min_samples_split=2, min_weight_fraction_leaf=0.0,
                n_estimators=150, n_jobs=-1, oob_score=False,
                      random_state=None,verbose=0, warm_start=False)
```

Which resulted in a mean squared error of 326.85 with the training data, and a mean squared error of 323.2377 when evaluated with the testing data.

For the rest of the models Contraceptive Prevalence was set as the Independent variable. When paired with Life Expectancy as the only Dependent variable, the best resulting model was a K Nearest Neighbors Regressor with the following parameters:

```
KNeighborsRegressor(algorithm='auto', leaf_size=30, metric='minkowski',
     metric_params=None, n_jobs=1, n_neighbors=40, p=2,
     weights='uniform')
```

Which resulted in a mean squared error of -191.66 with the training data and 222.83 with the testing data.

Contraceptive Prevalence as the Independent variable with Maternal Mortality Ratio as the Dependent variable resulted in an SVR (Support Vector Machine Regressor) as the best model with parameters:

```
SVR(C=100, cache_size=200, coef0=0.0, degree=3, epsilon=13, gamma=5e-07,
 kernel='rbf', max_iter=-1, shrinking=True, tol=0.001, verbose=False)
```

Which resulted in a mean squared error of -223.4890 with the training data and 220.41 with the test data.

And finally, the best model with Adolescent Fertility rate as Dependent and Contraceptive Prevalence as the Independent variable was a Random Forest regressor with parameters:

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=4,
        max_features='auto', max_leaf_nodes=None,
        min_impurity_split=1e-07, min_samples_leaf=1,
        min_samples_split=2, min_weight_fraction_leaf=0.0,
        n_estimators=900, n_jobs=1, oob_score=False, random_state=None,
        verbose=0, warm_start=False)
```

Which had a mean squared error of -333.4272 with the training data and an mean squared error of 316.5440 with the test data.

Conclusion:

In this study, the people found to be most at risk are adolescent girls in Africa, as Sub-Saharan African countries suffer the highest Maternal Mortality Ratios in the world and also have the highest rate of Adolescent Fertility. Analyses could be run on the region as a whole or the individual countries in order to establish whether Maternal Mortality is most closely related to health, social, or economic issues.

The analyses in this project suggest that contraception is an integral part of the quality of life for women across the world. Since there are such strong correlations between Contraceptive Prevalence and both Maternal Mortality and Adolescent Fertility, one way to combat high rates in these areas may be to increase access to contraception as well as appropriate education on the various kinds of contraception available. Increased Contraceptive Prevalence may also help increase Life Expectancy, as these two factors are highly correlated as well.

The unequal distribution of Female Legislators, Senior Officials and Managers vs their male counterparts just scratches the surface. As female lawmakers are much more likely to make policy that positively affects women, it is hypothesized that certain economic, education and health factors for women may correlate with the percentage of female legislators and leaders. More analyses need to be run to establish if there is any actual effect.

The most compelling correlations found were between the Decision Maker of a Woman's Health Care and Contraception Prevalence. It is striking to see the high positive correlation between contraceptive use and the Wife as the decision maker shift to a high negative correlation when the Husband is the decision maker. Autonomy over health decisions has a huge impact on quality of life and successful outcomes.

The relationship between education and Contraceptive Prevalence is a bit more complicated to tease out. There are many ways to measure educations levels and some appear to have significantly stronger correlations with Contraceptive Prevalence than others. Examining the features over time could give much deeper insight to what effects education has on contraception and vice versa.

There is a strong correlation between Female Literacy and Contraceptive Prevalence. This suggests that one of the more effective ways to combat poverty and overpopulation issues is to teach women and girls how to read. Once literacy is established, a person is empowered to make choices that are vital in their lives, and teach themselves, without having to rely on information from others.

Overall, further analysis is needed to uncover more subtle relationships that might help direct the development of policies to help improve the situations of people most at risk.

Appendix:

INFORMATION EXTRACTED FROM series.csv c

Contraceptive prevalence, any methods (% of women ages 15-49): relabeled as 'bc'
Long.definition: Contraceptive prevalence rate is the percentage of women who are practicing, or whose sexual partners are practicing, any form of contraception. It is usually measured for women ages 15-49 who are married or in union.
General.comments: Relevance to gender indicator: it is an indicator of women's empowerment and are related to several Millennium Development Goals such as maternal health, HIV/AIDS, and gender equality.
Source: UNICEF's State of the World's Children and Childinfo, United Nations Population Division's World Contraceptive Use, household surveys including Demographic and Health Surveys and Multiple Indicator Cluster Surveys.
Statistical.concept.and.methodology: Contraceptive prevalence reflects all methods - ineffective traditional methods as well as highly effective modern methods. Contraceptive prevalence rates are obtained mainly from household surveys, including Demographic and Health Surveys, Multiple Indicator Cluster Surveys, and contraceptive prevalence surveys. Unmarried women are often excluded from such surveys, which may bias the estimates.
Development Relevance: Reproductive health is a state of physical and mental well-being in relation to the reproductive system and its functions and processes. Means of achieving reproductive health include education and services during pregnancy and childbirth, safe and effective contraception, and prevention and treatment of sexually transmitted diseases. Complications of pregnancy and childbirth are the leading cause of death and disability among women of reproductive age in developing countries.

Life expectancy at birth, total (years): relabeled as ‘life’

Short.Definition: Life expectancy at birth indicates the average number of years a newborn infant would live.

Long.definition: Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.

Source: Derived from male and female life expectancy at birth from sources such as: (1) United Nations Population Division. World Population Prospects, (2) Census reports and other statistical publications from national statistical offices, (3) Eurostat: Demographic Statistics, (4) United Nations Statistical Division. Population and Vital Statistics Reprot (various years), (5) U.S. Census Bureau: International Database, and (6) Secretariat of the Pacific Community: Statistics and Demography Programme.

Statistical.concept.and.methodology: Life expectancy at birth used here is the average number of years a newborn is expected to live if mortality patterns at the time of its birth remain constant in the future. It reflects the overall mortality level of a population, and summarizes the mortality pattern that prevails across all age groups in a given year. It is calculated in a period life table which reflects a snapshot of a mortality pattern of a population at a given time. It therefore does not reflect actual mortality patterns that a person actually goes through during his/her life, which can be calculated in a cohort life table.

High mortality in young age groups significantly lowers the life expectancy at birth. But if a person survives his/her childhood of high mortality, he/she may live much longer. For example, in a population with a life expectancy at birth of 50, there may be few people dying at age 50. The life expectancy at birth may be low due to the high childhood mortality so that once a person survives his/her childhood, he/she may live much longer than 50 years.

Development.relevance: Mortality rates for different age groups (infants, children, and adults) and overall mortality indicators (life expectancy at birth or survival to a given age) are important indicators of health status in a country. Because data on the incidence and prevalence of diseases are frequently unavailable, mortality rates are often used to identify vulnerable populations. And they are among the indicators most frequently used to compare socioeconomic development across countries.

Maternal mortality ratio (modeled estimate, per 100,000 live births): relabeled as ‘matdeath'

Long.definition: Maternal mortality ratio is the number of women who die from pregnancy-related causes while pregnant or within 42 days of pregnancy termination per 100,000 live births. The data are estimated with a regression model using information on the proportion of maternal deaths among non-AIDS deaths in women ages 15-49, fertility, birth attendants, and GDP.

Limitations.and.exceptions: The methodology differs from that used for previous estimates, so data should not be compared historically. Maternal mortality ratios are generally of unknown reliability, as are many other cause-specific mortality indicators. The ratios cannot be assumed to provide an exact estimate of maternal mortality.

Notes.from.original.source: Estimates of maternal mortality are presented along with upper and lower limits of intervals (see footnote) designed to depict the uncertainty of estimates. The

intervals are the product of a detailed probabilistic evaluation of the uncertainty attributable to the various components of the estimation process. For estimates derived from the multilevel regression model, the components of uncertainty were divided into two groups: those reflected within the regression model (internal sources), and those due to assumptions or calculations that occur outside the model (external sources). Estimates of the total uncertainty reflect a combination of these various sources.
General.comments: Relevance to gender indicator: this indicator represents the risk associated with each pregnancy and is also a Millennium Development Goal Indicator for monitoring maternal health.
Source: WHO, UNICEF, UNFPA, World Bank Group, and the United Nations Population Division. Trends in Maternal Mortality: 1990 to 2015. Geneva, World Health Organization, 2015
Statistical.concept.and.methodology: Reproductive health is a state of physical and mental well-being in relation to the reproductive system and its functions and processes. Means of achieving reproductive health include education and services during pregnancy and childbirth, safe and effective contraception, and prevention and treatment of sexually transmitted diseases. Complications of pregnancy and childbirth are the leading cause of death and disability among women of reproductive age in developing countries.
Maternal mortality ratios are generally of unknown reliability, as are many other cause-specific mortality indicators. Household surveys such as Demographic and Health Surveys attempt to measure maternal mortality by asking respondents about survivorship of sisters. The main disadvantage of this method is that the estimates of maternal mortality that it produces pertain to 12 years or so before the survey, making them unsuitable for monitoring recent changes or observing the impact of interventions. In addition, measurement of maternal mortality is subject to many types of errors. Even in high-income countries with reliable vital registration systems, misclassification of maternal deaths has been found to lead to serious underestimation.
The modeled estimates are based on an exercise by the Maternal Mortality Estimation Inter-Agency Group (MMEIG) which consists of World Health Organization (WHO), United Nations Children's Fund (UNICEF), United Nations Population Fund (UNFPA), and World Bank, and include country-level time series data. For countries without complete registration data but with other types of data and for countries with no data, maternal mortality is estimated with a multilevel regression model using available national maternal mortality data and socioeconomic information, including fertility, birth attendants, and GDP.

Adolescent fertility rate (births per 1,000 women ages 15-19): relabeled as ‘teen'
Source: United Nations Population Division, World Population Prospects.
Statistical.concept.and.methodology: Reproductive health is a state of physical and mental well-being in relation to the reproductive system and its functions and processes. Means of achieving reproductive health include education and services during pregnancy and childbirth, safe and effective contraception, and prevention and treatment of sexually transmitted diseases. Complications of pregnancy and childbirth are the leading cause of death and disability among women of reproductive age in developing countries.
Adolescent fertility rates are based on data on registered live births from vital registration systems or, in the absence of such systems, from censuses or sample surveys. The estimated

rates are generally considered reliable measures of fertility in the recent past. Where no empirical information on age-specific fertility rates is available, a model is used to estimate the share of births to adolescents. For countries without vital registration systems fertility rates are generally based on extrapolations from trends observed in censuses or surveys from earlier years.

GDP per capita (Current US$): relabeled as 'gdp’

Long.definition: GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars.
Source: World Bank national accounts data, and OECD National Accounts data files.

Health expenditure, public (% of GDP): relabeled as 'healthspend'

Long.definition: Public health expenditure consists of recurrent and capital spending from government (central and local) budgets, external borrowings and grants (including donations from international agencies and nongovernmental organizations), and social (or compulsory) health insurance funds.
Notes.from.original.source: All the indicators refer to expenditures by financing agent except external resources which is a financing source. When the number is smaller than 0.05%, the percentage may appear as zero. In countries where the fiscal year begins in July, expenditure data have been allocated to the later calendar year (for example, 2008 data will cover the fiscal year 2007â€“08), unless otherwise stated for the country. For 2008, the use of yearly average exchange rates (compared to year-end exchange rates) may not fully represent the impact of the global financial crisis.
Source: World Health Organization Global Health Expenditure database (see http://apps.who.int/nha/database for the most recent updates).

'Public (Government) spending on education, total (% of GDP): relabeled as 'eduspend’

Long.definition: General government expenditure on education (current, capital, and transfers) is expressed as a percentage of GDP. It includes expenditure funded by transfers from international sources to government. General government usually refers to local, regional and central governments.
Source: United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute for Statistics.
Statistical.concept.and.methodology: Government expenditure on education, total (% of GDP) is calculated by dividing total government expenditure for all levels of education by the GDP, and multiplying by 100. Aggregate data are based on World Bank estimates.
Data on education are collected by the UNESCO Institute for Statistics from official responses to its annual education survey. All the data are mapped to the International Standard Classification of Education (ISCED) to ensure the comparability of education programs at the international

level. The current version was formally adopted by UNESCO Member States in 2011. GDP data come from the World Bank.

The reference years reflect the school year for which the data are presented. In some countries the school year spans two calendar years (for example, from September 2010 to June 2011); in these cases the reference year refers to the year in which the school year ended (2011 in the example).

Development.relevance: The percentage of government expenditure on education to GDP is useful to compare education expenditure between countries and/or over time in relation to the size of their economy; A high percentage to GDP suggests a high priority for education and a capacity of raising revenues for public spending. Note that government expenditure appears lower in some countries where the private sector and/or households have a large share in total funding for education.