# Competitor Packet


Presented by **Citadel** and **Citadel Securities**
In Partnership with **CorrelationOne**

# Table of Contents

Day-Of Schedule……3

Technical Assessment Rubric……4

Problem Statement ……...5 – 10

Data Tables Schema ……..11 – 18

# Day-Of Schedule

NOTE: Subject to change.

| | |
|---|---|
| *9:00am – 12:00pm* | **Live Technical Assessment**<br>***5 minutes of overview + 40 minutes of Q&A/dialogue for each team by one pair of technical assessors*** |
| *12:00pm – 1:00pm* | **Technical Assessment Aggregation**<br>***Technical assessment ranks from remote and live sessions are aggregated and discussed*** |
| *1:00pm – 3:00pm* | **Official Judging Session**<br>***5-person panel ranks and selects the top 3 teams*** |
| *3:00pm – 3:30pm* | **Introduction of Judges** |
| *3:30pm – 4:30pm* | **Winners Announcement**<br>***Top 3 teams recognized*** |

# Technical Assessment Rubric

NOTE: There is no explicit points system across the criteria on this rubric – submissions will be evaluated holistically based on the below categories.

| Category | Scoring Criteria |
|---|---|
| **Non-Technical Executive Summary** | **Insightfulness of Conclusions**<br>• Chose relevant question and provided clear motivations for their decision<br>• Exhibited precise, nuanced conclusions as opposed to blanket (over)generalizations |
| **Technical Exposition** | **Wrangling & Cleaning Process**<br>• Conducted proper data quality control, such as handling missing values, outliers, errors, and non-normalized fields<br>• When appropriate, made clever transformations of different data fields to better join/use them together, justifying any unconventional tactics used<br>• Discusses any feature engineering performed and motivations for it |
| | **Investigative Depth**<br>• Conducted a multi-step EDA process with proper visualizations at each step, explaining why they used the visualizations they did and how the results informed and/or motivated their subsequent decisions<br>• Generated, tested, and interpreted the results of informed hypotheses<br>• Synthesized the partial results from individual analyses well |
| | **Analytical & Modeling Rigor**<br>• Reflects proper quantitative methodology as well as more qualitative components like outlier, residual, and mediator/instrumental variable analysis<br>• Assumptions and choices are reasoned, paying particular attention to the feature selection process<br>• For models, analyzed performance and discussed shortcomings<br>• For visualizations/statistical tests, discussed motivations behind the particular ones built and what they illuminate |
| | **Creativity & Miscellaneous**<br>• Generally will not make or break, but if team exhibits unusual creativity (e.g. bringing in an unexpected outside dataset, drawing a rigorous analogy to a different context) or an “X-Factor” not covered by the above, this can push them over the edge |

# Problem Statement

Welcome to the 2019 Citadel – Correlation One Data Open! This document explains the topic of the Data Open, important details about the datasets you'll be using, and guidance on how to submit your results.

## Background

Humans have evolved into a highly intellectual species capable of solving much of the world's profound mysteries and challenges. However, many other living organisms have followed suit and have evolved with us such as viruses, bacteria, fungi and parasites.

In the mid 1300s, a small specimen – a bacterium not visible to the naked eye (*Yersinia pestis*) eradicated roughly one-third of the European population and reduced the total world population by an estimated 100 million in the span of 5 years. Historical records indicate that this organism originated in Asia but rapidly spread through Europe by European trade merchants upon their arrival from Asia. Animals such as rats acted as the perfect home for the bacteria, enabling its ability to reproduce and infect other species. The Black Plague is a perfect example of an infectious disease which affected the global society in whole. Specifically, economies were disrupted, populations were diminished and civil unrest rose to an all-time high. Indeed, humans eventually determined the causes and developed antibacterial drugs to combat such infectious diseases from reoccurring.

The world has joined together in unified organizations such as the World Health Organization (WHO) to monitor and track the rise of infectious diseases through the globe and prevent any potential outbreaks. Over the course of the 20$^{th}$ century, much of the world's deadliest infectious diseases were eradicated from the developed world, and in the 21$^{st}$ century, the emerging world has begun to follow suit.

In 2016, the WHO indicated that the top 10 leading causes of the 56.9 million deaths worldwide were responsible for over half of such deaths. In fact, the percentage share of deaths due to the top 10 causes had increased from the year 2000, and furthermore many of these causes were disease-related. Although many of those were non-infectious diseases, such as heart disease, cancer, and stroke, a fair number were infectious, contagious diseases. The latter are of particular interest because unlike the former, which have complex scientific causes with long histories, they can often be mitigated through appropriate social and economic policy. It remains to be seen if the world will produce and adopt sensible such policies to further reduce the negative impact of such diseases.

## Your Task

Your goal is to analyze the provided datasets – and possibly in combination with external data – and explore hypotheses concerning factors driving the spread of infectious diseases. After

reaching your conclusions, you should strive to develop a proposal around what policymakers should do to best mitigate the negative effects of infectious diseases.

**You are asked to pose your own question and answer it using the available datasets in the available time**. What is important is both the creativity of your question and the quality of your data analysis. **You need not be comprehensive; depth of insight will be rewarded over breadth of the question posed.**

Submissions may be predictive, using machine learning and/or time series analysis to predict future disease spread trends. Submissions may also be illuminating, through use of data visualizations or through sound statistical tests.

Consider reading the following articles to expand your breadth of knowledge. You are strongly encouraged to conduct your own industry research to draw further inspiration.

- An Introduction to Infectious Disease – Provides surface-level grasp of infectious disease pathogens, where they come from, how we recover from them, and their effects on the human population
- Measles cases in Europe tripled last year, WHO says - Gives context on impact in real life and provides links to external data
- Disease experts reveal their biggest worries about the next pandemic – Contains interesting prompts in bold that are answers to scientists' "biggest worry about the next pandemic"

## Datasets

The provided datasets are stored in the "Datathon Materials" folder on Google Drive and are spread across nine tables. Your team should only use the tables that are relevant to your chosen question/topic. The raw data sources are noted; however, we encourage you to use our tables since they have been organized and cleaned to "play nice" with each other.

***country_codes***
Table which maps each country to a code which will be used throughout the other data tables.
*284 rows & 2 columns.* Size: ~0.1MB.

***mortality***
Information about the number of deaths from varying causes of death by country and age of victims from 1988-2017. https://icd.who.int/browse10/2016/en#/ provides exhaustive coverage on the causes of death.
*~1 million rows & 37 columns*. Size: ~200MB. Source: WHO Health Statistics and Information Systems.

***connectivity***

Information about prevalence of Internet and mobile phone usage per country from 2000 – 2017.
*792 rows & 21 columns.* Size: ~0.2MB. Source: The World Bank.

***consumption***
Information about food & services consumption patterns per country in 2010 .
*~160,000 rows & 9 columns.* Size: ~20MB. Source: Global Consumption Database.

***health_indicators***
Information about various health indicators per country from 1969 – 2017.
*12,936 rows & 246 columns.* Size: ~17MB. Source: WHO Global Health Observatory.

***immunization***
Information about percentage immunization rates of infants against various diseases per country from 1980 – 2017.
*849 rows & 41 columns.* Size: ~0.2MB. Source: WHO Global Health Observatory.

***influenza_activity***
Information about influenza infections per week per country from 2000 – 2018.
*75,719 rows & 22 columns.* Size: ~9MB. Source: WHO FluNet.

***migration***
Information about bilateral migration patterns every decade from 1960 – 2010.
*~160,000 rows & 11 columns.* Size: ~5MB. Source: The World Bank.

***sanitation***
Information about access to sanitation services per country from 2000 – 2015.
*1,152 rows & 20 columns.* Size: ~0.2MB. Source: WHO Global Health Observatory.

***water_quality***
Information about access to clean water & general water quality per country from 2000 – 2015.
*3,960 rows & 19 columns.* Size: ~0.7MB. The World Bank.

## Additional Datasets

We encourage you to explore additional data from the WHO Global Health Repository, both to generate ideas for interesting questions to ask as well as to support your reasoning for whatever question you choose to answer. The data can be found here: http://apps.who.int/gho/data/node.home.

The World Bank's AidFlows application visualizes how much development aid is provided and received around the world. Users can select individual donors (providing the aid) and beneficiaries (receiving the aid) to track the sources and uses of aid funding. A glossary of the

terms can be found here: http://www.aidflows.org/glossary.pdf, and the API to access the data can be found here: http://www.aidflows.org/api/.

Additionally, you are welcome to scour the Web for custom datasets to supplement your analysis. All additional data used should be public and should not exceed 2GB unzipped (consult Correlation One's technical product team via your team's dedicated Slack channel if you believe your idea is worthy of an exception).

## Other Materials

We will provide you the schema for each of the data tables in another packet.

## Submissions: Content

Submissions should have three components:

1. Report – this should have two main sections:
    a. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what is their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged if they help explain your thoughts.
    b. Technical Exposition – What was your methodology/approach towards answering the questions? Describe your data manipulation and exploration process, as well as your analytical and modeling steps. Again, use of visualizations is highly encouraged when appropriate.
    c. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you MUST include it or your entire submission will be discarded.**
2. High-Level Presentation – this should satisfy three requirements:
    a. Be in **Microsoft PowerPoint format**
    b. Be a **(very) condensed, non-technical version** of your technical report
    c. Include the topic question, key insights, results, and any data visualizations that you think are helpful to communicate and support your points

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your technical report without your team there to explain it; therefore, **your submission must "speak for itself"**. Please ensure that your main findings are clear and that any visualizations are functionally labeled.

## Submissions: Evaluation

The competition will have multiple rounds of evaluation. The most important component of this evaluation will be your Report, which will be judged as follows:

- **Non-Technical Executive Summary**
  - *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose it? Are your conclusions precise and nuanced, as opposed to blanket (over)generalizations?
- **Technical Exposition**
  - *Wrangling & Cleaning Process.* Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform? Please describe your process in detail within your Report.
  - *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypotheses tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?
  - *Analytical & Modeling Rigor.* What assumptions and choices did you make, and what was your justification for them? How did you perform feature selection? If you built models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular ones you built, and what do they tell you?

Additionally, you will have a 45-minute Q&A session with a pair of judges on the day of the actual event. These judges can ask anything they would like pertaining to your report; you will be evaluated based on the comprehensiveness and rigor of your responses.

## Submissions: Format

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

**However, please also include the source file used to generate your report.** For example, if you submit a PDF with math-type, equations, or symbols, please include your raw LaTeX source file.

Code should be submitted in a single zipped collection of files separate from your report.

Your team will be sent a Google Form at the beginning of the competition; you will use this form to upload and send in your submission content. **Submissions MUST be received by 11:59 PM on March 31st, 2019. Any submission received after that time will NOT be evaluated by the judges**.

## Tips & Recommendations

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: http://jupyter.org/install.html. Jupyter Notebook is an interactive, real-time development environment that eliminates many pain points of the standard "terminal + text editor" environment, and is compatible with both Python and R.

We also recommend that your team not try to learn new tools if possible; instead, leverage your existing skills to extract as much insight from the data as you can.

Finally, **we STRONGLY encourage you to start typing up your final submission AT LEAST eight hours before the submission deadline**. In the past, many teams have spent a lot of time conducting great analyses, only to realize that they left almost no time for actually writing up and presenting their results. **This cannot be stressed enough – quality data analysis that is incomplete or poorly presented will NOT win one of the top prizes**.

| Tips for Success | Try to Avoid |
|---|---|
| **1.** Focus on hypothesis testing when brainstorming your research question | **1.** Do not try to exhaust all different models you know just to yield an ideal cross validation accuracy |
| **2.** Spend at least 8 hours on your report to ensure strong communication through visualizations and writing | **2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so make sure those conditions are sufficient |
| **3.** Engage in proper causal analysis. Just because your model passes standard cross-validation checks it does not demonstrate (or even suggest) causality | **3.** Do not pick research statements and blindly stick to it trying to get it to work. Often times, further data exploration will show that it's not even true or worthwhile |

## Ask for Help

Correlation One's technical product team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move your analysis forward.

# Data Tables Schema

## country_codes

Table which maps each country to a code which will be used throughout the other data tables. *284 rows & 2 columns.* Size: ~0.1MB.

| Field | Type | Description |
|---|---|---|
| **country_code** | STRING | Three letter code of country |
| **country_name** | STRING | Name of the country |

## mortality

Information about the number of deaths from varying causes of death by country and age of victims from 1988-2017. https://icd.who.int/browse10 provides exhaustive coverage on the causes of death. The provided dataset is a curated subset of the source dataset, so certain causes of death cannot be found in mortality.csv.
*~1 million rows & 37 columns*. Size: ~200MB. Source: WHO Health Statistics and Information Systems.

| Field | Type | Description |
|---|---|---|
| **year** | INTEGER | Year of death observed |
| **country_code** | STRING | 3 letter abbreviation for country |
| **cause_label** | STRING | ICD10 is a medical classification list. ***cause_label*** contains meta-data for ICD10 and is one of: 'ICD10 header category', 'ICD10 3 char detail', 'ICD10 4 char detail' Please consult WHO ICD10 index on breakdowns of categories https://icd.who.int/browse10 |
| **cause_code** | STRING | ICD10 code indicating a cause of death. 'ICD10 header category' = {1000…1103} 'ICD10 3-char detail' = {A00…Z99} 'ICD10 4-char detail' = {A000…Z999} |
| **cause_description** | STRING | Cause of death corresponding to the ***cause_code*** |
| **age_format** | INTEGER | Indicates which age-group format the mortality values follow. See **mortality_age_format** table for details. |
| **i_age_format** | INTEGER | Indicates which age-format the infant mortality values follow. See **mortality_age_format** table for details. |
| **deaths1** | FLOAT | Deaths at all ages |
| **deaths2** | FLOAT | Deaths at age 0 year |
| **deaths3** | FLOAT | Deaths at age 1 year |
| **deaths4** | FLOAT | Deaths at age 2 years |

| | | |
|---|---|---|
| **deaths5** | FLOAT | Deaths at age 3 years |
| **deaths6** | FLOAT | Deaths at age 4 years |
| **deaths7** | FLOAT | Deaths at age 5-9 years |
| **deaths8** | FLOAT | Deaths at age 10-14 years |
| **deaths9** | FLOAT | Deaths at age 15-19 years |
| **deaths10** | FLOAT | Deaths at age 20-24 years |
| **deaths11** | FLOAT | Deaths at age 25-29 years |
| **deaths12** | FLOAT | Deaths at age 30-34 years |
| **deaths13** | FLOAT | Deaths at age 35-39 years |
| **deaths14** | FLOAT | Deaths at age 40-44 years |
| **deaths15** | FLOAT | Deaths at age 45-49 years |
| **deaths16** | FLOAT | Deaths at age 50-54 years |
| **deaths17** | FLOAT | Deaths at age 55-59 years |
| **deaths18** | FLOAT | Deaths at age 60-64 years |
| **deaths19** | FLOAT | Deaths at age 65-69 years |
| **deaths20** | FLOAT | Deaths at age 70-74 years |
| **deaths21** | FLOAT | Deaths at age 75-79 years |
| **deaths22** | FLOAT | Deaths at age 80-84 years |
| **deaths23** | FLOAT | Deaths at age 85-89 years |
| **deaths24** | FLOAT | Deaths at age 90-94 years |
| **deaths25** | FLOAT | Deaths at age 95 years and above |
| **deaths26** | FLOAT | Deaths at age unspecified |
| **i_deaths1** | FLOAT | Infant deaths at age 0 day |
| **i_deaths2** | FLOAT | Infant deaths at age 1-6 days |
| **i_deaths3** | FLOAT | Infant deaths at age 7-27 days |
| **i_deaths4** | FLOAT | Infant deaths at age 28-364 days |

## mortality_age_format

Observations not using format code 0 combines certain age groups. In such cases, the data is recorded under the youngest of the combined age groups; the unused part is left blank. For example, when ***age_format*** = 2, mortality between the ages 1-4 are recorded under deaths at age 1.

| **Format code for *age_format*** | **0** | **1** | **2** | **3** | **4** | **6** | **7** | **9** |
|---|---|---|---|---|---|---|---|---|
| Lower bound on age of death | All ages | All ages | All ages | All ages | All ages | All ages | All ages | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 1 | 1 | 1 | 1-4 | 1 | 1-4 | 1-4 | 1-4 | |
| 2 | 2 | 2 | | 2 | | | | |
| 3 | 3 | 3 | | 3 | | | | |
| 4 | 4 | 4 | | 4 | | | | |
| 5 | 5-9 | 5-9 | 5-9 | 5-9 | 5-9 | 5-9 | 5-14 | |
| 10 | 10-14 | 10-14 | 10-14 | 10-14 | 10-14 | 10-14 | | |
| 15 | 15-19 | 15-19 | 15-19 | 15-19 | 15-19 | 15-19 | 15-24 | |
| 20 | 20-24 | 20-24 | 20-24 | 20-24 | 20-24 | 20-24 | | |

N O A

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 25 | 25-29 | 25-29 | 25-29 | 25-29 | 25-29 | 25-29 | 25-34 | AGE DISTRIBUTION |
| 30 | 30-34 | 30-34 | 30-34 | 30-34 | 30-34 | 30-34 | | |
| 35 | 35-39 | 35-39 | 35-39 | 35-39 | 35-39 | 35-39 | 35-44 | |
| 40 | 40-44 | 40-44 | 40-44 | 40-44 | 40-44 | 40-44 | | |
| 45 | 45-49 | 45-49 | 45-49 | 45-49 | 45-49 | 45-49 | 45-54 | |
| 50 | 50-54 | 50-54 | 50-54 | 50-54 | 50-54 | 50-54 | | |
| 55 | 55-59 | 55-59 | 55-59 | 55-59 | 55-59 | 55-59 | 55-64 | |
| 60 | 60-64 | 60-64 | 60-64 | 60-64 | 60-64 | 60-64 | | |
| 65 | 65-69 | 65-69 | 65-69 | 65-69 | 65-69 | 65 &+ | 65-74 | |
| 70 | 70-74 | 70-74 | 70-74 | 70-74 | 70-74 | | | |
| 75 | 75-79 | 75-79 | 75-79 | 75 &+ | 75 &+ | | 75 &+ | |
| 80 | 80-84 | 80-84 | 80-84 | | | | | |
| 85 | 85-89 | 85 &+ | 85 &+ | | | | | |
| 90 | 90-94 | | | | | | | |
| 95 | 95 &+ | | | | | | | |

| Format code for i_age_format | 1 | 2 | 8 |
|---|---|---|---|
| Day of death for infants | < 1 year old | < 1 year old | < 1 year old |
| 0 days | 0 days | 0-6 days | 0-365 days |
| 1-6 days | 1-6 days | | |
| 7-27 days | 7-27 days | 7-27 days | |
| 28-365 days | 28-365 days | 28-365 days | |

## connectivity

Information about the prevalence of the Internet and mobile phone usage per country from 2000 – 2017.
*792 rows & 20 columns.* Size: ~0.2MB. Source: The World Bank.

| Field | Type | Description |
|---|---|---|
| **country_code** | STRING | Three letter code of country |
| **statistic** | STRING | 15 unique metrics measured and documented regarding water quality |
| **year [18 total]** | FLOAT | Value of ***statistic*** measured in year YYYY |

## consumption

Information about food & services consumption patterns per country in 2010.
*~160,000 rows & 9 columns.* Size: ~20MB. Source: Global Consumption Database.

| Field | Type | Description |
|---|---|---|
| **country_code** | STRING | Three letter code of country which this row corresponds to |
| **area** | STRING | Area of country which this row corresponds to - national, rural, or urban |

| | | |
|---|---|---|
| **product** | STRING | The product which this row corresponds to |
| **value** | FLOAT | Numerical quantity for this row. Details of what this means will be provided by subsequent columns |
| **exchange_rate** | FLOAT | Exchange rate between local currency and US dollars, given in units of foreign currency per US dollar |
| **consumption_segment** | STRING | Consumption segment that this row corresponds to. Four levels of consumption are used to segment the market in each country: lowest, low, middle, and higher. They are based on global income distribution data, which rank the global population by income per capita. The lowest consumption segment corresponds to the bottom half of the global distribution, or the 50th percentile and below; the low consumption segment to the 51th–75th percentiles; the middle consumption segment to the 76th–90th percentiles; and the higher consumption segment to the 91st percentile and above. For more information, see: http://datatopics.worldbank.org/consumption/detail |
| **indicator** | STRING | The type of value displayed in the ***value*** column |
| **ppp_conversion_factor** | FLOAT | The implied exchange rate between local currency and US dollars, measured by how many local currency units would be needed to buy a basket of goods that one US dollar could buy. For more information, see: https://en.wikipedia.org/wiki/Purchasing_power_parity |
| **unit** | STRING | Units in which ***indicator*** is measured |

## health_indicators

Information about various health indicators per country from 1969 – 2017.
*12,936 rows & 246 columns.* Size: ~17MB. Source: WHO Global Health Observatory.

| Field | Type | Description |
|---|---|---|

| | | |
|---|---|---|
| **country_code** | STRING | Three letter code of country |
| **year** | STRING | Year of statistic or metric measured in format YYYY |
| ***Various metrics* [244 total]** | INTEGER, FLOAT | Values of metrics measured for the given year and country |

## immunization

Information about the percentage of immunization rates of infants against various diseases per country from 1980 – 2017.
*849 rows & 41 columns.* Size: ~0.2MB. Source: WHO Global Health Observatory.

| Field | Type | Description |
|---|---|---|
| **country_code** | STRING | Three letter code of country |
| **statistic** | STRING | Statistic of a particular vaccine or immunization metric (values are in percentage (%)) |
| **year [38 total]** | FLOAT | Year of data collected in format YYYY |

## influenza_activity

Information about influenza infections per week per country from 2000 – 2018.
*75,719 rows & 21 columns.* Size: ~9MB. Source: WHO FluNet.

| Field | Type | Description |
|---|---|---|
| **country_code** | STRING | Three letter code of country |
| **who_region** | STRING | Particular region as determined by WHO |
| **flu_region** | STRING | Particular region |
| **year** | INTEGER | Year of influenza infection in format YYYY |
| **week** | INTEGER | Week of influenza infection [0 – 52] |
| **start_date** | STRING | Start date of infection in format YYYY-MM-DD |
| **end_date** | STRING | End date of infection in format YYYY-MM-DD |
| ***various metrics* [14 total]** | INTEGER, FLOAT, STRING | Various metrics associated with influenza outbreak, such as the number of specimens received/processed, the number of total detected influenza viruses, etc. |

## migration

Information about bilateral migration patterns every decade from 1960 – 2010.
*160,776 rows & 8 columns.* Size: ~5MB. Source: The World Bank.

| Field | Type | Description |
|---|---|---|
| **country_orig_code** | STRING | Three letter code of origin country |

| | | |
|---|---|---|
| **country_dest_code** | STRING | Three letter code of destination country |
| **gender_code** | STRING | Three letter code of gender [FEM = female, MAL = male, TOT = total] |
| **year [5 total: 1960, 1970, 1980, 1990, 2000]** | FLOAT | Total individuals who migrated from the country of origin to the destination country in that particular decade. Here, each year corresponds to the decade starting in that year; e.g. 1960 means 1960 - 1969. |

## sanitation

Information about access to sanitation services per country from 2000 – 2015.
*1,152 rows & 19 columns.* Size: ~0.2MB. Source: WHO Global Health Observatory.

| Field | Type | Description |
|---|---|---|
| **country_code** | STRING | Three letter code of country |
| **category** | STRING | Region type (rural, urban or total) |
| **statistic** | STRING | Two unique metrics measured in percentage (%) |
| **year [16 total]** | FLOAT | Value of metric measured in year YYYYY |

## water_quality

Information about access to clean water and general water quality per country from 2000 – 2015.
*3,960 rows & 18 columns.* Size: ~0.7MB. Source: The World Bank.

| Field | Type | Description |
|---|---|---|
| **country_code** | STRING | Three letter code of country |
| **statistic** | STRING | 15 unique metrics measured and documented regarding water quality |
| **year [16 total]** | FLOAT | Value of metric measured in year YYYY |